DSC 423 Spring 2019
Group Project - Technical Report

Stratagem
I travelled 15 hours and All I got was a Futon?

Shweta Gujrathi, Brendan Foley, Cody Smith,
Andy Huang, Theresa Shen and Ying Kam Chiu

Table of Content

Abstract

The goals of the analysis, regarding AirBnB rentals in Melbourne, Australia, was (1) to determine which independent variables have the greatest significant impact that align to the price of a nightly stay at an AirBnB listing and (2) to provide a final model that has a high degree of confidence to predict future outcomes.

The publicly available dataset was obtained from Kaggle.com. The team initialized the analysis on the dataset during the pre-processing stage by exploring which of the variables could be categorized as a dependent variable, independent variable, or a variable not to be used in the analysis. At this stage we were also able to reclassify the 250+ postal codes into 5 regions that constitute Melbourne. Subsequently we randomized and divided the 22K observations from the full data set into blocks of 2.5K observations, of which each of the 6 team members received their own unique randomized dataset. During the exploratory stage we chose to use linear regression to explain the chosen response variable Y (Price). Transforming the Y variable was necessary in order to stabilize the variance. Each team member fitted the full model to arrive at a final model that had adjusted-R2 ranging between 54% - 66%.

Qualitative variables that proved to be a constant theme amongst the team were determined to be included in a successful model were host response time, region, and room type. The number of people the property accommodates, and the rating score were the most prevalent quantitative variable that was included in all final models. Variables such as if the owner was a superhost, the bed type listed, and cancelation policy were deemed not to be as relevant and were dropped from most final models.

In order to provide a model that has a higher level of confidence to determine the price an AirBnB rental, additional variables or variables with complete information are required to be included in further improving the final models presented.

Introduction

In recent years, the sharing economy has gained popularity in many industries, such as transportation (e.g. Uber, Lyft) and accommodation (e.g. Airbnb, HomeAway). The sharing economy can be defined as “peer-to-peer-based activity of obtaining, giving and sharing the access to goods and services, coordinated through community-based online services”. In the peer-to-peer (P2P) accommodation industries, Airbnb is a leader in the market, in which its business model enables hosts to offer their unoccupied properties or rooms for short-term rental. Since its establishment in 2008,

Airbnb's business has grown significantly and served more than 150 million guests through over 3 million listings in more than 190 countries in less than a decade.

Melbourne ranked as the 6th on the list of top ten cities for users globally in 2016 and has been one of the top ten cities since then. Not only limited to short-term rentals, Airbnb in Melbourne has entered the market competition of long-term rentals. According to Rawnsley and Schmahmann (2018), "The median number of nights hosted per year has increased from 42 nights to 66 nights per year. Of the listings that have hosted guests, over 35 percent of listings host guests for up to 30 nights per year. Approximately 27 per cent of listings host guests for more than 180 nights per year." Even though the article stated that the impact of Airbnb on the Melbourne housing markets appears minimal, it is still interesting to understand how Airbnb's pricing has enabled itself to extend to another new market.

Listing price is often considered to be one of the critical factors that impact consumer's choices of lodging. In this paper, our objective is to assess different listing prices per night of Airbnb listings in the city of Melbourne with a number of factors (independent variables), such as number of listings that the host has, the maximum number of guests that the listing accommodates, required amount of security deposit, required amount of cleaning fee, review scores rating, the average time that the host responses to a guest's inquiry, whether the host is a super host, room types, bed types, types of cancellation policy and locations.

More importantly, we intend to go beyond the data that we already have direct access to and explore how the mentioned independent variables play a role in the pricing decision. The dataset consists of 8-year of Airbnb listing history from 2010 to 2018 in Melbourne and was consolidated in December 2018. The population of the dataset has 22,895 observations and each of us is given 2,500 observations as sample to begin our individual analysis.

We conduct the analysis through linear regression and draw conclusions on final predictive models to predict future Airbnb listing prices per night in Melbourne. Since each of us begins with different samples, we may come up with different final models. It would be interesting to identify which independent variable(s) is/are significant that will be included in most of our models and how well our models are to predict future prices with unseen dataset.

Methodology

A. Andy

The dataset, Melbourne Airbnb Open Data, was acquired from website Kaggle: https://www.kaggle.com/tylerx/melbourne-airbnb-open-data#listings_dec18.csv. This analysis aims to

predict the price for Airbnb in Melbourne Australia and to see how each selected predictor interact with listing price. There was a total of 22,895 observations in our original dataset, each of the team member was assigned with 2,500 observations for individual analysis with a common goal.

For pre-processing stage, I identify the dependent variable as well as the independent variables. The variables are all classified as either quantitative variables or qualitative variables. At the beginning, the numbers of observations are precisely 2500. However, since there are some observations without the recorded "region" information, I decided to remove them since those data without the recorded "region" are usually lack of other recorded information, therefore are not sufficient for analysis.

At the data exploration stage, dummy variables are created for all qualitative variables in order to continue my analysis through SAS 9.4. In addition, an interaction term is also created to explore the relationship between two independent variables. At this stage, I build histogram and scatterplots of each independent variable versus the dependent variable to get a basic understanding of my dataset by data visualization.

The scatterplots are a good indicator of detecting outliers. Through data visualization, I identify the issue of my dataset, if there is any. For the histogram, if the histogram shows a skewed distribution, then applying transformation will be necessary. At the end of this stage, I will fit a regression model, and there are five indicators that I will utilize in order to fit a better optimal model. These five indicators are

- Parameter estimates by beta weights
- The p-value (smaller than 0.05 if significant) and the F-value
- High significance by t-test p-value
- High R2 and Adj-R2 values
- Low RMSE or MSE value

For the Parameter Estimates, it is necessary to examine the results of the t-test for the coefficients of each independent variable. The independent variable with the largest p-value will be removed since it has the least or no effect on the response variable. Moreover, the independent variable with the largest p-value will be removed one by one and rerun the regression whenever there is an independent valuable been deleted.

After the data exploration, I entered the data analysis stage. At the beginning of this stage, I computed the VIF value to identify the multicollinearity problem. If there are any predictors with the VIF value larger than ten than I will remove them one by one and rerun the regression. At this point, I have fitted my final model.

Next, I proceeded on checking outliers and influential points based on the studentized residuals and cook's D, to see if my final model can be further improved. I have removed all the observations if they are both an outlier and an influential point.

After removing the observations that were both an outlier and an influential point, I rerun the

model every time until there was no such kind of observations. At this stage, I kept the remaining outliers or influential points in my analysis, only removed the flagged observations which were the observations that are both outliers and influential points.

After finish removing outliers and influential points, I examined the model assumptions to see if the assumptions are satisfied. The assumptions include

- Linearity
- Constant variance
- Independence
- Normality

Furthermore, by generated the Standardized coefficients values, I have identified what the most critical predictors in my model are.

Next, I tested the model by applying Cross Validation. By splitting the dataset into Training and Testing set, I knew that my final model performs well. In addition, Stepwise selection method has also been applied at this stage.

I validated the final model by examining the ASE graph, ASE value for both training set and testing set, and Stepwise selection summary. At the end of my analysis, I performed two predictions based on my final model.

B. Theresa

Data was retrieved from Kaggle and the total observations is above 20,000. The goal is to have a
better understanding of Airbnb price and how does each variable influence price in general. The major stages include, data pre-processing, data exploration, and data analysis. The group have decided to remove several variables in the beginning as those variables adds no value to our analysis of listing price. We started off by grouping dataset into five different regions based on their zip code, and this is part of our data cleaning stage. After each group member has been randomly assigned 2500 observations, I removed those observations that has no value for regions.

Before entering the exploration stage, I performed some other data pre-processing steps, which
includes distinguishing between dependent variable and independent variables; classify independent variables as either qualitative or quantitative. Region is one of my independent variables, but due to the fact that some observations does not have region associated with it, I have to remove them before starting data exploration stage.

After data are cleaned and pre-processed. I started exploring the dataset. I build histograms,

scatterplots, data descriptive, and frequency table to visualize the data and to have a general understanding of the data and the variables. The figures give me important characteristics of data, such as, the central tendencies and the spread of the variables.

Frequency table is used to provide me with a big picture of the qualitative variables, and this provides me with valuable information on the number of dummy variables I have, and their ratio to each other. The scatterplots are used for each independent variable against the dependent variable. This two-dimensional data visualization technique provides information on the relationship between two variables and how are they correlated. Histogram gives the shape (distribution) and the spread of the data, which is extremely meaningful for understanding the data; it allows me to inspect for outliers, skewness, etc. The data descriptive is also needed as this provides information of the mean, median, and different quartiles of my data. Dummy variable will be created for variables that has been identified as qualitative variables. Interaction term will also be created if assessed to have value adding effect on my data analysis.

After I have a basic understanding of the data, I can further identify if there are any issues with the data. If there are a few observations with missing values, I will remove them for further data exploration. Afterwards, I will proceed with identifying outliers through examine histograms and scatterplots from above. I need to take into consideration that holiday season may affect the data, so outliers may exist. If the outliers do not seem to be too far away from the holiday season expectations, then I may keep those outliers for now. If outliers fall outside of my expectation range, then I will state how it will affect regression analysis and remove it respectively.

Data will be transformed if the histogram above show a skewed distribution of data. I will fit a regression model at the end of exploration stage. The key indicators that I will be using when determining a good regression model are F-Values, P-values, RMSE, R2, and adjusted R2. I will also check for significance of each variable; I will remove variables that is p-value greater than 0.05, and I will remove them one by one to make sure all variables are significant.

The next stage is analysis stage. I will identify multicollinearity by computing Pearson correlation and VIF statistics. If variable have VIF higher than 10, I will remove them one by one as it indicates that multicollinearity exists. I will once again re-run regression to double check that the issue has been resolved. Outliers and influential points will be checked next. Studentized residuals and cook's D table will be checked for outlier issues. I will remove observations that are both an outlier and an influential point.

Constant variance, independence, linearity, and normality will be verified next. I will look

through dependent variable against each predictor and see if assumptions can be satisfied. Further transformation would be needed if assumption is not satisfied. Then, I will split my dataset into training and testing at 75/25 using randomly selected seed value; using selection method to select the best variables to fit the final model. The indicators above, such as, RMSE and Adj R2, will be checked again. Afterwards, I will validate my final model using test set. If my testing has loser ASE than training, then the final model is validated. In addition, I will compute two sets of prediction for further data analysis. Prediction intervals give clear information on the data itself.

C. Shweta

This analysis is to predict the price for Airbnb in Melbourne Australia and the variables which influence the pricing. Since pricing is a numeric variable, linear regression is used for the analysis.

The methodology followed for the analysis is as follows –

- Data preprocessing (cleaning and renaming variables to make sure everything is in a format that SAS 9.4 can read)
- Data Exploration (Frequency tables, Histograms, Boxplots, Scatterplots)
- Data Analysis (Multicollinearity, outliers and influential points, residuals, transformation)
- Model Validation using the Train and Test method
- Predictions

The data was obtained from Kaggle and had more than 20K observations.
https://www.kaggle.com/tylerx/melbourne-airbnb-open-data#listings_dec18.csv.
After cleaning the data and deciding which variables to go ahead with, the data was divided into different datasets of 2500 observations for each group member.

Data Preprocessing

Data cleaning was done in excel. Initial sample consisted of 2500 observations out of which 469 observations were deleted due to missing data points, which brought the number of observations to 2031. Since the group had already worked together to combine regions based on city, suburb and zip code, deleted the City, Suburb and zip code columns. Deleted the column bed_type since only 8 observations had bed_type other than 'Real bed'. Deleted the observations where room_type was 'Shared room' since the number of observations for it were 15 (less than 30). Deleted the observations where response_time was 'a few days or more' since the number of observations for it were 15 (less than 30). The final dataset used for analysis consisted of 2001 observations with no missing data.

Shortened the names of the variables and qualitative data points for easier import to SAS and writing code. Here is the list of all variables and categories that was changed in excel and will be used moving forward in the final file used for the project-

Name in original file	Name in cleaned file
response_time	res_time
within an hour	anhour
within a few hours	fewhours
within a day	aday
Private room	P_room
Entire home/apt	home_apt
accommodates	acc

The final number of independent variables in the dataset is 10 which includes 5 qualitative variables (res_time, superhost, room_type, can_policy, and region) and 5 quantitative variables (total listings, acc, security_deposit, cleaning fee, review_score_rating).
The cleaned excel file was converted into csv to be imported into SAS.

Data Exploration

Data was imported into SAS using the infile statement. Dummy variables for quantitative variables were created at the same time. Interaction term sd*sh (superhost*security_deposit) was also created at the same time.

Frequency tables was created to check if the dummy variables have been coded correctly and for missing observations. Descriptive statistics, histograms and scatter plots were created to analyze the qualitative variables. Boxplots were created for the variables ‘superhost’ and ‘room_type’. Although the histograms showed skewness and the gplots didn’t seem exactly linear, residual analysis was done to check if the polynomial regression is needed. After analyzing the residuals, it was decided that the data is fairly linear with a lot of outliers and influential points. Pearson correlation

coefficient table was also checked for collinearity. None of the variables including the interaction term seemed to be collinear.

Data Analysis

At the data analysis stage, multicollinearity, outliers and influential points were checked. No multicollinearity was observed as per the vif value. Close to 100 observations were deleted in 3 rounds since the output was improving after each stage of removing influential points and outliers. Student residuals and cook's d tables were used to analyze outliers and influential points. Removal of outliers and influential points was stopped once the output didn't seem to improve.

Residual plots were again checked for assumptions after each round of removing outliers and influential points. The graphs got better with each stage. The normal probability plot was still showing a slight 'S' shape after the 4^{th} round of removing outliers and influential points. Therefore, a log transformation was done on the y-variable – 'Price'. All the assumptions for linearity, constant variance, normality and independence) were satisfied after the log transformation of the y-variable price.

Data Validation using train and test

The data was divided into training and test sets in a 75/25 ratio at random. Stepwise and cp model selection methods were used on the train set to arrive on a fit model. Since both the selection methods showed the same variables for the final model, decided to proceed with the stepwise selection. The fit model was then used on the test set to validate the accuracy of the model. CV R-square for the test model is less than 0.3 thus proving that the model is good for prediction.

Predictions

Predictions was done with 2 sets of random data points along with calculating the confidence and the prediction interval.

Retransformation

The transformed y-variable Price was retransformed to arrive on the model statement and to write the effect of each x-variable on price.

D. Cody

I chose to focus my efforts on AirBnB rentals from Melbourne, Australia. The publicly available dataset was obtained from Kaggle.com (https://www.kaggle.com/tylerx/melbourne-airbnb-open-data#listings_summary_dec18.csv). I initialized the analysis on the dataset during the pre-processing stage by exploring which of the variables could be categorized as a dependent variable, an

independent variable, or a variable not to be used in the analysis. We removed certain columns within the initial data file that didn't deem necessary or lacked complete data to draw a proper conclusion from (Square Ft, Total # Reviews, Reviews Per Month).

Also, at this stage it was determined that we would need to reclassify the 250+ postal codes and cluster them into 5 regions, as called on a similar AirBnB study, Perez-Sanchez VR, Serrano-Estrada L, Marti P, Mora-Garcia R-T. (2018), that constitute Melbourne, based on classifications from Bob (2019), to make the analysis have an easier geographic reference point to create dummy variables based on the newly created Region variable. The below list of variables 11 independent (6 qualitative and 6 quantitative) and 1 dependent quantitative variable (price) were chosen for the analysis. An interaction term between accommodates and security deposit was also created (acc*sd).

Variable	Description	Code/Values	Name
1	Host Response Time to Renter	0 = N/A 1 = Within an hour 2 = Within a few hours 3 = Within a day 4 = A few days or more	host_response_time
2	**Renter is Superhost**	**0 = T; 1 = F**	**host_is_superhost**
3	# of Listings Completed Since Listing on Site	count	host_total_listings_count
4	**Room Type**	**0 = Entire home/apt** **1 = Private Room** **2 = Shared Room**	**room_type**
5	Max # People Rental Accomidates	count	accommodates
6	**Bed Type**	**0 = Real bed** **1 = Couch** **2 = Pull-out Sofa** **3 = Futon** **4 = Airbed**	**bed_type**
7 (DV)	Price of Rental Per Night	AU$	price
8	**Security Deposit Per Stay**	**AU$**	**security_deposit**
9	Cleaning Fee Per Stay	AU$	cleaning_fee
10	**Avg Review Score Since Listing Property on Site**	**0 - 100**	**review_scores_rating**
11	Cancellation Policy	0 = Flexible; 1 = Strict; 2 = Moderate	cancellation_policy
12	**Region of Melbourn**	**0 = WM (Western Melbourne)** **1 = IM (Inner Melbourne)** **2 = SEM (South Eastern Melbourne)** **3 = NSM (Northern Suburbs Melbourne)** **4 = EM (Eastern Melbourne)**	**Region**
13	Interaction var between Accomidates and Security Deposit	multipicative product	acc_sd

Subsequently the data was randomized and divided from the original +22K observations from the full data set into blocks of 2.5K observations, of which each of the 6 team members received their

own unique randomized dataset. There were several formatting and missing data issues related to instances within the file. For missing data for security_deposit, cleaning_fee, and review_scores_rating, which may not have been required during the input stage, were given a value of zero. The host_response_time that contained blanks was reclassified as “N/A”. The cancellation_policy variable was revised to only having 3 levels instead of the original 6; anything that contained the word strict was reclassified to “strict”. Any listings that didn’t have a postal code were removed completely from the dataset.

During the exploratory stage we chose to use linear regression to explain the chosen response variable Y (Price). The initial exploration stages were done with the aid of histograms and boxplots. The initial price histogram (Fig D.1) illustrated it was positively skewed and probability plot (Fig D.2) showed an exponential shapes curve. In order to stabilize the variance, the transforming of the price variable was required. I tested several different methods of transformation, but log(Y) was the best fit when examining the residual plots (Fig D.3/Fig D.4).

Frequency tables were utilized to ensure that all the dummy variables were being correctly coded and represented in the dataset (Fig D.10). Boxplots for superhost (Fig D.5), room type (Fig D.6), cancellation policy (Fig D.7), bed type (Fig D.8), and response time (Fig D.9) were utilized to explore the interactions that the qualitative variables play against price. Looking at the gplots and the matrix scatterplot (Fig D.11), linearity didn’t appear to be present, but after analyzing the residuals it showed a decent number of outliers that required to be removed. The Pearson correlation table was ran and it confirmed that none of correlation values are above the 0.9 threshold, meaning there is no issue with multicollinearity. Also, the VIF statistics were all less than 10 The interaction term acc_sd (accommodation * security deposit) was rather high at .854, but was still acceptable to remain in the model due to being underneath the threshold.

Before the analysis section was able to get under way there was a need to take initial data set of 2,482 observations and remove the outliers and influential points where the observations that had a high studentized residual > |3|, Cook’s D > (4/n), and/or a hat h_{ii} value > 0.5 were flagged and removed from the dataset. This took several rounds to clean up before moving to fit the model, with the remain 2,415 observations that remained. A regression analysis was ran between each round and showed tremendous improvements from the original dataset. The normal probability plot was nearly a perfectly line at the end.

The dataset was divided into train/test with a 70/30 break with a random seed value. Stepwise (Fig D.13) and Forward (Fig D.14) methods were applied to the training dataset in order to fit the model. They both resulted in the same 11 significant variables, so it wasn’t necessary to compare the

different models side by side. The validation tests (Fig D.15/D.16) identified that the CV R^2 result (0.0294) was less than 0.3, meaning that the model is valid and can be used for new data added to the dataset to make predictions. A 5-Fold Cross Validation (Fig D.19/D.20) was also ran and the ASE (Train) > ASE (Test), proving that it is a valid model.

2 random sets of predictions (Fig D.20) were created and added to the dataset in order to test the final model, which provided confidence intervals and prediction intervals. Price (Y) variable was then retransform for the final model statement in order to calculate that each of the x-variables had in relation to price.

E. Brendan

The service and hotel industry impact the lives of the majority of working professionals, tourists and repeat business is critical towards long term stability and profits. Air Bnb is a disruptive technology company that allows people to rent out their homes and earn profit from their extra room capacity and time away. The ability to maximize the value for both the renter and tenant relies heavily on the location and the price and this analysis is to see how price is worked out.

Data secured through the Kaggle website provides all rentals within the city of Melbourne during the month of December in 2018. There are over 21000 different observation data points with up to 96 different variables. These observations are the basis from which a planned analysis of the price is derived from. At the end of this analysis, the expectation is that we should be able to predict the price of a listing in Melbourne based on specific key variables and use these insights to advice

In the pre-processing state, the dependent variable of price was identified as well as 11 quantitative and qualitative variables as being important. The bi-variate variables were first explored through histograms, sgscatter plot matrices and then reviewed for linearity, normality and if there are any missing data points. Boxplots were used to identify the variation of the qualitative variables as well as checked for outliers and missing data. Following the identification any potential interaction, variables were assembled and investigated for multicollinearity and influence on the dependent variable price

In order to better identify if the data points had a strong enough correlation, the missing data points were either deleted or replaced with 0 or the mean. This was to ensure that the system could read the missing data. Following the data replacement, a correlation proc was executed to identify if there are issues of multicollinearity which in turn went through a regression model to identify their influence.

A regression model fit was used in order to cleanly identify the independent variables and model's importance towards the dependent variable. Of particular value would be variables that a) pass the

<.05 p-value test, b) a high R2 and Adj-R2 score, c) high MSRE and SE score and d) a strong F-value and that the entire model passes the p-value test.

This information will identify if the variance is caught with the current variable with a high R2 and Adj-R2. The f-value strength is indicative of how strong a model it is and whether it captures the variation of the y-variable, the dependent variable. The individual p-value indicator tests in the parameter estimate box indicates if the individual variables impact the dependent variable in any measurable way, a score of less than .05 indicates an impact or relationship, while anything higher than a .05 alpha indicates low to no impact.

Following the identification of the models' important variables, they were then split into a test and training sets. These then through selection process of backwards regression and forwards regression which variables influenced the R2 and had an alpha of less than .05. Using this final model, I then predicted the possible prices based on the key variables.

F. Ying

In this paper, Melbourne, a city of Australia was chosen as the study site. The dataset of Melbourne Airbnb Open Data was obtained from website Kaggle (https://www.kaggle.com/tylerx/melbourne-airbnb-open-data#listings_dec18.csv), which provides Airbnb listing information from Airbnb.com. In accordance with the source, Melbourne was the 6th on the list of top ten cities for users globally in 2016 and has been one of the top cities for listings globally since then.

In the pre-processing stage, I identify the selected 12 predictors with a quantitative variable – listing price per night *(price)* as the dependent variable and 11 other independent variables, including quantitative: number of listings that the host has *(host_total_listings_count*), maximum number of guests that the listing accommodates *(accommodates)*, amount of security deposit *(security_deposit)*, amount of cleaning fee *(cleaning_fee)* and review scores rating *(review_scores_rating)* and qualitative variables: average time the host responses to a guest's inquiry: "N/A", "within an hour", "within a few hours", "within a day" or "within a few days' *(host_response_time),* whether the host is a super host: "t" or "f" (host_is_superhost), room type: entire apartment, private room, shared room *(room_type)*, what the type of the bed: real bed, sofa bed, futon, airbed, pull-out bed *(bed_type)*, type of cancellation policy; strict, moderate or flexible *(cancellation_policy)* and region.

Out of 22,895 observations, our team randomized the population in excel and each team member was assigned with a different sample of 2,500 observations. There are some data issues needed to be fix in excel before loading into SAS. Blanks for qualitative variables like security_deposit, cleaning_fee and review_scores_rating were replaced by zeros. For variable region, it is not an original variable from the dataset. With reference to Bob (2019), we divided Melbourne

into 5 regions: Inner Melbourne (IM), Northern Suburbs Melbourne (NSM), South Eastern Melbourne (SEM), Eastern Melbourne (EM) and Western Melbourne (WM) by using given zip codes and cities. In addition, there are 9 observations removed because they have no location-identifier and therefore there is no way to verify the data. For variable cancellation_policy, there are 4 layers of strict policy, such as “strict”, “strict_14_with_grace_period”, “super_strict_30” and “super_strict_60”. The 4 alike layers were consolidated into one layer as “strict”. The blanks for host_response_time were replaced with “N/A”, representing either the host did not or was never needed to respond. There is one blank for host_is_superhost, which was replaced with ‘f’. After all the data-preprocessing and data cleansing, there are 2491 observations left to load to SAS for further analysis.

In the data exploration stage, I took an overview into the dataset through data visualization. For univariate analysis, I built histograms and boxplots to explore quantitative variables. These figures give a general understanding about the central tendencies and the spreads of the variables. Transforming the variables will be necessary if the histogram shows a skewed distribution. With reference to Gut and Herrmann (2015), dummy variables are computed for qualitative variables: 4 dummy variables that indicate the average time the host responses to a guest’s inquiry between dHRT0 and dHRT4, a dummy variable that indicates whether the host is a super host between dHIS0 and dHIS1, 2 dummy variables that indicate room type between dRT0 and dRT2, 3 dummy variables that indicate bed types between dBT0 and dBT3, 2 dummy variables that indicate cancellation policy type between dCP0 and dCP2 and 4 dummy variables that indicate regions between dRG0 and dRG4. At the same time, an interaction variable for two selected independent variables dHRT1 and dHIS was built.

For bivariate analysis, scatterplots were built for each independent variable against the dependent variable or a scatterplot matrix to observe the patterns displayed and the relationship between all independent variables within a single matrix. Boxplots are also a good visualization tool to compare attributes of the qualitative variables.

At this phase, a basic understanding of the dataset was obtained to better identify data issues. Then, I proceeded to detect if there was more missing data. If there are only a few missing values and they appeared to be random, I may proceed with the deletion of these cases. If not, I may replace the values with the median, mean or mode.

After fixing data issues, I checked for outliers through data visualization, such as scatterplots, and histograms. I kept the outlier or verify if it fit the dataset. “Verified” meant that the data does not fall too far outside our expectations and can be reasonably assumed considering this is the holiday season. Otherwise, I would report how the outlier would change the regression line or analysis result

and remove it.

At the end of the exploration stage, I tried to fit a regression model. The criteria for an accurate model come from checking of assumptions and 5 indicators: (1) parameter estimates by beta weights, (2) high significance by t-test p-values (<0.05), (3) F-values and their p-values (<0.05) (4) low RMSE/ MSE and (5) a high R2 and Adj- R2 values. With reference to Christensen, we can test with the "error term, which is the residual that cannot be explained by the variables in the model". The assumptions are that "the error terms are independent from one another, identically distributed (constant variance), linear and normally distributed". If constant variance, independence and normality assumptions are violated, the regression line estimate are still unbiased but standard errors, confidence intervals and prediction intervals will be incorrect. In order to stabilize the variance, we have to transform the variables. If linearity assumption is not satisfied, a more complex regression would be used. The next step was to detect severe multicollinearity with a scatterplot matrix and a Pearson correlation matrix for each pair of the independent variables.

After fitting a full model, it entered the analysis stage to identify severe multicollinearity by computing VIF statistics. If the VIF is above 10, it indicates that the regression coefficient of the variable is poorly estimated. Then, I continue checking for outliers and influential points with the help from studentized residuals and cook's D. Observations would be flagged if they are classified as both outliers and influential points. I would remove observations one by one and then rerun regression for the remaining data to check for the next flagged observation. It would be a judgement call to either delete or keep the observation.
Following the outlier test, I checked to verify if the assumptions: constant variance, independence, linearity and normality, hold. If the residual plots of full model comparing the dependent variable against each independent variable test well, I could reasonably assume that the data observed has constant variance, independence, linearity and normality. Should they fail these tests, transformations will be invoked.

After transforming the variables, I then tested the performance of the full model by splitting the dataset into train and test sets. The train set was used to select the "best" variables to fit the final model by comparing two model selection methods, such as forward and backward methods. Checking of assumptions, the above mentioned 5 indicators of an accurate model and diagnostics for multicollinearity were performed again. Test performance was factored in to decide on the final model. The test performance trumps everything. If the test set has a lower RMSE, higher R^2 and adj-R^2 than the train set and Cross-Validated R^2 (CV-R^2) is smaller or equal to absolute 0.3, the final model is said to be a good model. Lastly, I used the validated final model to perform two predictions.

Analysis, Results and Findings

A. Andy

In the remaining 2478 observations that I got, I created dummy variables for all the qualitative variables that I have. For variable "host response time," I chose "N/A" as baseline since "N/A" keep repeating in the dataset. A total of 4 dummy variables were created since I got 5 levels: "N/A," "within an hour," "within a few hours," "within a day," and "a few days or more."

Similarly, the remaining dummy variables were created as following:
For variable "host is super host," only 1 dummy variable was created since there are only two levels, "t" and "f." In addition, "t" was selected to be the baseline.

For variable "region," the level "IM" was selected to be the baseline, and 4 dummy variables were created since there are a total of 5 levels: "IM," "WM," "EM," "SEM," and "NSM," for this qualitative variable.

For variable "room type," the level "Entire home/apt" was selected to be the baseline, and 2 dummy variables were created since there are a total of 3 levels: "Entire home/apt," "Private room," and "Shared room" for this qualitative variable.

For variable "bed type," the level "Real Bed" was selected to be the baseline, and 3 dummy variables were created since there are a total of 4 levels: "Real Bed," "Pull-out sofa," "Futon," and "Airbed" for this qualitative variable.

For variable "cancellation policy," the level "Strict" was selected to be the baseline, and 2 dummy variables were created since there are a total of 3 levels: "Strict" "moderate," and "flexible" for this qualitative variable.

Moreover, an Interaction term was created and added after all the aforementioned dummy variables to analysis the relationship between variable "cleaning fee" and variable "host is super host."

At this stage, I created a histogram to take a first look at the normal distribution of the current dataset, as figure A.1 shown. The normal distribution is the most widely known and used of all distribution, it helps identify the probability problems of my current dataset, if any.

From the histogram on this exploration stage, the curve is not symmetrical, which means that the

“mean” is not centered, and it signals a probability problem. The histogram on this stage is “positively skewed,” which means that it has skewed to the right, with its “mean” owns the largest value, followed by the “median,” and the “mode” holds the least value. From figure A.2, I can tell that the “mean” holds a value of 140.8975, the “median” holds a value of 109, and the mode holds a value of 100. In addition, most value have fall within the lower range.

The scatterplots of the dependent variable “price” response to all independent variables are also created to examine how much each independent variable is affected by the dependent variable, “price.” The scatterplot matrix is not useful in this analysis since the number of quantitative variables plus all the dummy variables created before are so many that the scatterplot matrix cannot show an obvious correlation.

For all the quantitative variables: “host total listing counts,” “accommodates,” “security deposit,” “cleaning fee,” “review scores rating,” most of the scatterplots show no perfect positive/negative correlation or even high or low positive/negative correlation. The points on all the scatterplots are gathered at the bottom left corner such as “price” versus “host total listing counts,” “security deposit,” and “cleaning fee,” gathered at the bottom right corner such as “price” versus “review scores rating,” or not explainable such as “price” versus “accommodates.” Figure A.3 to figure A.7 presented how the scatterplots in the current stage distributed.

However, from figure A.8, the scatterplot “price” versus the interaction term, “cleaning fee versus host is super host,” the association seems somehow close to linear, but still is not an obvious linear relationship, which means not preferable as all the other scatterplots mentioned earlier do.

With dummy variables, we cannot see a linear relationship on the scatterplots, since all points will be scattered along 0 and 1. I did not generate boxplot since I believe generating histogram and scatterplots are both appropriate ways and even better ways to visualize the distribution of the numerical data, especially for such large numbers of independent variables in my analysis.

At this stage, it is clear that the current dataset has serious normal distribution problems and applying a transformation to the response variable “price” to stabilize the variance is necessary for my analysis to continue. One of the most common transformation is Log(Y). I decided to fit the regression model on the transformed “price” only and examine the residual plots to see if the model assumptions are satisfied.

The first step after applying for a transformation is checking if the new variable, Log(Price), is created, and use the new “price,” – “ln_price,” for all the following analysis. Then I generated a

descriptive table to take a first look at my transformed dataset.

Next, I created the scatterplots of "ln_price" versus each independent variable to check that the transformed variable "ln_price" is linearly associated to the x-variables.

From the new scatterplot outputs, the scatterplots of "ln_price" versus "cleaning fee" and "ln_price" versus the interaction term, "cleaning fee versus host is superhost" have been improved after the transformation of the dependent variable, as figure A.9 and figure A.10 shown. Other scatterplots also improved but do not show much of the improvement. Again, with dummy variables, all points will be scattered along 0 and 1, so we cannot see a linear relationship on the scatterplots.

Figure A.11 shows the histogram at this stage again for the purpose of comparing to the previous histogram to see of the normal distribution has been improved. Comparing to the previous histogram, it is obvious that the new histogram is closer to a normal histogram, which means it shows no skew. The normal curve represents a perfectly symmetrical distribution, and this is the preferred result.

In addition, according to the output of the univariate procedure as figure A.12 shown, the line show on the graph appears to be linear, which is preferred as well.

A full model was then fitted to find the independent variables that have a significant effect on "ln_price." The Adj-R2 value at this stage is 0.5944, which means that there are 59.44% of the data are captured and explained by the model (figure A.13). However, it is still necessary to keep testing the model to see if the Adj-R2 value can be further improved.

After fitting the regression model, the independent variables "d_a_few_days_or_more" and "d_airbed" had been set to 0. When this happen, it means that the observations are not enough for run the model with the independent variable. Therefore, I removed the variable "d_a_few_days_or_more" at first, then rerun the regression model, then I removed the variable "d_airbed," and rerun the model again. After these two times removals, there is no value changed, the output is exactly the same as the previous output.

For the remaining 20 independent variables, I decided to delete the variable with the highest p-value one by one and rerun the model every time until all the predictors in my optimal model are significant.

The variable "d-moderate" was deleted since it holds the p-value of 0.9145, which is the highest insignificant p-value among all the predictors. After removed this insignificant predictor and rerun the

model, the Adj-R2 value has increased to 0.5947, which is preferred since the higher Adj-R2 value, the better. Also, the F-value has changed from 98.53 to 103.80, which is better than the previous model.

The variable “d_EM” was then be deleted since it holds the p-value of 0.7846, which is the highest insignificant p-value among all the remaining predictors. After removed this insignificant predictor and rerun the model, the Adj-R2 value has slightly increased to 0.5950 and the F-value has changed from 103.80 to 109.63, which is better than the previous model.

Then the variable “d_pullout-sofa” was deleted since it holds the p-value of 0.6983, which is the highest insignificant p-value among all the remaining predictors. After removed this insignificant predictor and rerun the model, the Adj-R2 value has again, slightly increased to 0.5953 and the F-value has changed from 109.63 to 116.15, which is better than the previous model.

The only interaction term - “cleaning_fee_d_host_is_superhost” then be removed since it holds the p-value of 0.6329, which is the highest insignificant p-value among all the remaining predictors. After removed this insignificant predictor and rerun the model, the Adj-R2 value has slightly increased to 0.5955 and the F-value has changed from 116.15 to 123.47, which is better than the previous model.

The variable - “d_futon” was then be deleted since it holds the p-value of 0.1707, which is the highest insignificant p-value among all the remaining predictors. What is different is that, after removed this insignificant predictor and rerun the model, the Adj-R2 value has slightly decreased to 0.5952. However, the F-value has changed from 123.47 to 131.49, which is better than the previous model.

The new Adj-R2 value of 0.5952 is almost identical as the previous Adj-R2 value of 0.5955, but the new F-value has significantly improved; therefore, I decided to ignore the slightly decrease in the value of Adj-R2 value and kept deleting the predictor with the highest insignificant p-value.

Next, the variable “host_total_listings_count” was deleted since it holds the p-value of 0.1819, which is the highest insignificant p-value among all the remaining predictors. After removed this insignificant predictor and rerun the model, the Adj-R2 value has slightly decreased to 0.5950 and the F-value has changed from 131.49 to 140.67, which is better than the previous model. Again, although the Adj-R2 value has slightly decreased, it was still almost identical with the previous Adj-R2 value. On the other hand, the new F-value has again drastically increased. Therefore, I decided to keep ignoring the slightly decreased in new Adj-R2 value unless there is a significantly decreased.

In addition, the predictor “d_host_is_superhost” was previously one of the insignificant

predictors, but after deleting variable “host_total_listings_count” then rerun the model, “d_host_is_superhost” has turned into a significant predictor, its p-value has changed to 0.0347, which is less than 0.05, therefore, it is significant.

The variable “d_within_a_day” was then be deleted since it holds the p-value of 0.1673, which is the highest insignificant p-value among all the remaining predictors. After removed this insignificant predictor and rerun the model, the Adj-R2 value has slightly decreased to 0.5947, but it was almost identical with the previous one. The F-value has changed from 140.67 to 151.24, which is significantly better than the previous model.

The variable “d_flexible” then been eliminated since it holds the p-value of 0.1129, which is the highest insignificant p-value among all the remaining predictors. After removed this insignificant predictor and rerun the model, the Adj-R2 value has slightly decreased to 0.5942, but it was almost identical with the previous one. The F-value has changed from 151.24 to 163.44, which is significantly better than the previous model.

The variable “review_scores_rating” then been deleted since it holds the p-value of 0.0933, which is the highest insignificant p-value among all the remaining predictors. After removed this insignificant predictor and rerun the model, the Adj-R2 value has slightly decreased from 0.5942 to 0.5695, but there is not difference between 0.59 and 0.56. The F-value has changed from 163.44 to 193.63, which is a huge improvement.

The two predictors – “security_deposit” and “d_within_a _few_hours” are two significant predictors originally; however, after deleting “review_scores_rating,” both of those have become insignificant predictors. However, due to the fact that F-value has significantly improved and Adj-R2 is not a significant difference, I decided to remove predictor “security_deposit” next and see what happens.

The variable “security_deposit” was then been removed since it holds the p-value of 0.2405, which is the highest insignificant p-value among all the remaining predictors. After removed this insignificant predictor and rerun the model, the Adj-R2 value has slightly increased from 0.5695 to 0.5814 and the F-value has changed from 193.63 to 261.16, which is a very good improvement.

Besides, now, all of my remaining 10 predictors are all significant and their VIF values are all less than 10, which means there is no multicollinearity problem in my model, as figure A.14 shown. I decided to check for assumptions to make sure the final model performs well. I utilized residual analysis to check model assumptions such as Constant Variance and Independence.

To check the assumption of Constant Variance and Independence, I checked Plot residuals versus Predicted Values (figure A.15). If the pattern of the spread shows a definite pattern, then there is a problem. The distribution shows that somehow points are randomly scattered inside a band centered around the horizontal line. As a result, the plots do show Constant Variance and Independence.

At this stage, I finally got my final model:

*ln_price = 4.62402 + 0.09557*accommodates + 0.00181*cleaning_fee - 0.16545*d_within_an _hour – 0.09442*d_within_a_few_hours + 0.06436* d_host_is_superhost - 0.31864*d_WM - 0.07842*d_SEM - 0.19151*d_NSM - 0.57445*d_private_room - 1.07743*d_shared_room + e*

where d_within_an_hour =1 when Host Response Time = within an hour

d_within_a_few_hours =1 when Host Response Time = within a few hours

d_host_is_superhost = 1 when Host Is Super Host = "t"

d_private_room = 1 when Room Type = "Private room,"

d_shared_room= 1 when Room Type = "Shared room,"

d_WM = 1 when Region = "WM,"

d_SEM = 1 when Region = "SEM,"

d_NSM = 1 when Region = "NSM"

Next, I proceed on checking the outliers and influential points. By removing the points that are both an outlier and an influential point, I rerun a total of 4 times until there is no point which is both an outlier and an influential point in my output, and there are a total of 2433 observation left, as figure A.16 shown. The number of observations that I removed as well as the remaining total number of observations that I had are as follow:

	Number of Obs removed	Remaining total number of Obs
First Time	23	2455
Second Time	15	2440
Third Time	6	2434
Fourth Time	1	2433

Now, the final model is satisfied with model assumptions such as Linearity, Constant Variance, Independence, and Normality. The distribution of "Plot residuals versus Predicted Value" in figure A.17 shows that points are randomly scattered inside a band centered around the horizontal line, so

the plots do show Constant Variance and Independence. To check the assumption of Normality, I examined the normal probability plot (figure A.18) of the residuals to see if the points lie close to the line. In figure A.18, the points show almost a straight line, so the plots do show Normality. As for figure A.19, the "Studentized versus Predicted" plot, the residual plot is not curvilinear, which means that it matches the assumptions of all Independence, Constant Variance, and Linearity.

After deleting a total of 45 flagged observations, my final model has significantly improved. By comparing figure A.14 and A.16, I found that the Adj-R2 value increased from 0.5814 to 0.6685, which means that at this stage, 66.85% of the variance in my dependent variable has been captured and explained. The F-value has also significantly increased, from 261.16 to 369.65. The new model equation after removing outliers and influential points is as follow

*ln_price = 4.55926 + 0.10452*accommodates + 0.00153*cleaning_fee - 0.13951*d_within_an _hour – 0.06649*d_within_a_few_hours + 0.06551* d_host_is_superhost - 0.29170*d_WM - 0.07655*d_SEM - 0.19426*d_NSM - 0.57727*d_private_room - 1.19174*d_shared_room + e*

where d_within_an_hour =1 when Host Response Time = within an hour

d_within_a_few_hours =1 when Host Response Time = within a few hours

d_host_is_superhost = 1 when Host Is Super Host = "t"

d_private_room = 1 when Room Type = "Private room,"

d_shared_room= 1 when Room Type = "Shared room,"

d_WM = 1 when Region = "WM,"

d_SEM = 1 when Region = "SEM,"

d_NSM = 1 when Region = "NSM"

The independent variables "accommodates," "cleaning_fee," and "d_host_is_superhost" are positively associated to "ln_price;" while "d_within_an _hour," "d_within_a_few_hours," "d_WM," "d_SEM," "d_NSM," "d_private_room," and "d_shared_room" are all negatively associated with response variable "ln_price." Besides, based on the results of the final model equation, now I can compute how each independent variable influence predicted price.

- Assuming all other variables being constant, for any additional guests that the property accommodates, price for a night is predicted to **increase** by 11.01%, calculated as (exp(0.10452) - 1) * 100 = 11.01.
- Assuming all other variables being constant, for any additional dollar amount increase in cleaning fee, price for a night is predicted to **increase** by 0.15%, calculated as (exp(0.00153) - 1) * 100 = 0.15.

- Assuming all other variables being constant, if host is a superhost, price for a night is predicted to **increase** by 6.77%, calculated as (exp(*0.06551*) - 1) * 100 = 6.77.
- Assuming all other variables being constant, if host response time is within an hour, price for a night is predicted to **decrease** by 14.97%, calculated as (exp(*0.13951*) - 1) * 100 = 14.97.
- Assuming all other variables being constant, if host response time is within a few hours, price for a night is predicted to **decrease** by 6.87%, calculated as (exp(*0.06649*) - 1) * 100 = 6.87.
- Assuming all other variables being constant, if the property is located at WM, price for a night is predicted to **decrease** by 3.87%, calculated as (exp(*0.29170*) - 1) * 100 = 33.87.
- Assuming all other variables being constant, if the property is located at SEM, price for a night is predicted to **decrease** by 7.95%, calculated as (exp(*0.07655*) - 1) * 100 = 7.95.
- Assuming all other variables being constant, if the property is located at NSM, price for a night is predicted to **decrease** by 21.44%, calculated as (exp(*0.19426*) - 1) * 100 = 21.44.
- Assuming all other variables being constant, if the property provided a private room, price for a night is predicted to **decrease** by 78.11%, calculated as (exp(*0.57727*) - 1) * 100 = 78.11.
- Assuming all other variables being constant, if the property provided a shared room, price for a night is predicted to **decrease** by 229.28%, calculated as (exp(*1.19174*) - 1) * 100 = 229.28.

At this stage, I would like to know what the most important predictors in my model are. From the absolute values of standardized coefficients, I identified that the independent variable "d_private_room" has the strongest influence on my dependent variable, "ln_price," since it has the highest absolute standardized estimate value of |-0.42812| compare to other predictors (figure A.20). The independent variable "accommodates" also has a strong influence on my dependent variable, since "accommodate" shows the second highest absolute value of standardized coefficient, 0.37650, among the rest of the predictors. In addition, a Pearson Correlation Coefficients table was generated as figure A.21 shown. There is no value larger than 0.9, so I do not have a multicollinearity problem.

Finally, I applied Cross Validation on my optimal model, A.22 shows all the outputs. The Average Square Error (ASE) graph shows that the red line (Test set) is below the blue line (Training set). In addition, within the "Stepwise Selection Summary," the 4 important indicators: "SBC," "ASE," "Test ASE," and "CV Press" all show dropping numbers, which means that during the selection process, the error has kept being reduced, and eventually be minimized. I also checked the value of ASE (Train) and ASE (Test). The value of ASE for my Training set is about 0.12, while the value of ASE for my Testing set is about 0.11. The fact that there is only 0.01 in the difference between 0.12 and 0.11 indicates that my model is good and Cross Validation is successfully conducted.

After Cross Validation, I could utilize my final model for prediction, as figure A.23 shown. For

instance, if there is an Airbnb which accommodates 4 people, requires AU$70 for cleaning fee, host response time is within an hour, the host is a superhost, located in WM, and provides private room, then how much will the price of this Airbnb be?

4.1386 with C.I. 4.0366, 4.2406

(exp(4.1386) - 1) * 100 = 6171.49

(exp(4.0366) - 1) * 100 = 5563.34

(exp(4.2406) - 1) * 100 = 6844.95

Since I have transformed the dependent variable "price" to Log(price) at the beginning, I have to retransform "ln_price" back to complete my predictions. After retransformation, the Airbnb price is as follow:

AU$ 6171.49 with C.I. AU$ 5563.34, AU$ 6844.95

Or, if there is an Airbnb which accommodates 6 people, requires AU$100 for cleaning fee, host response time is more than an hour, the host is not a superhost, located in NSM, and provides shared room, then how much will the price of this Airbnb be?

4.0148 with C.I. 3.8158, 4.2138

(exp(4.0148) - 1) * 100 = 5441.22

(exp(3.8158) - 1) * 100 = 4441.30

(exp(4.2138) - 1) * 100 = 6661.29

After retransformation, the Airbnb price is as follow:

AU$ 5441.22 with C.I. AU$ 4441.30, AU$ 6661.29

B. Theresa

I started my dataset with 2500 observations; it has 11 independent variables and 1 dependent variable. Of the 11 independent variables, 6 are qualitative variables and 5 are quantitative variables. There are five different regions, but due to the fact that there are 18 observations that does not have regions associated with it, and there is no way to determine which regions they actually belong to; therefore, I removed those observations during data pre-processing stage.

I decided to create dummy variables for all 6 qualitative variables as it will ease the analysis later on, which are host response time, host is super host, room type, bed type, cancellation policy, and region. Together, 16 dummy variables are creased. I've also created the interaction term between review score ratings and whether the host is superhost or not. This is to test if the review score ratings tend to be higher for superhost than those who are not a superhost, the assumption is that super host listings tend to have higher rating, and reputation, thus, they tend to have higher price point.

After data pre-processing; I started off with computing the descriptive, histogram,

scatterplots, and frequency table. The descriptive in appendix B.1 show that the mean is 145.56 and the median is 109, The mean is greater than the median, which means it is skewed to the right. Kurtosis is greater than 3, which indicates the dataset has heavier tails than a normal distribution. By looking at the histogram in appendix B.2, it is very clearly indicated that the data are not normally distributed, the majority of data are on the far left side of the histogram.

The scatterplot form B.4-B.24 also proves that there does not seem to be any relationship between price (dependent variable) and each independent variable. We can see from every single scatterplot that the correlation is very weak. The scatter plot for dummy variables that I have created earlier have points fall on either 0 or 1 as I have coded them; this makes these scatterplots less informative; as a result, I've decided to include the frequency table (appendix B.3) for all those qualitative variables that have dummy variables. The frequency table give me an idea of their ratio to each other, and how often are they occurring in the dataset.

At this stage, it is very clear to see that transformation of data is absolutely in need. I use log transformation to continue with my dataset. After transformation, I computed descriptive (appendix B.26), histogram (appendix B.25), and some scatterplots (appendix B.27). We can see the transformation is successful. The histogram (appendix B.25) now have normal distribution, and it is bell shaped and unimodal.

The Pearson correlation table (appendix B.25.1) shows that numroom2 has the highest correlation with log price; its correlation value is 0.649, followed by accommodates of 0.618. We can see from the descriptive table; mean is now 4.70 and median 4.69, which is very close. Kurtosis is less than 3, so there might be potential outliers. It is a lot better than the histogram before log transformation in appendix B.2.

The scatterplots show the correlation has improved. Despite those that has been coded with 0 and 1, other quantitative variables now have some sort of relationship with the dependent variable; we can see there is clearly linear correlation between price and cleaning fee. Also, from the scatterplot, there are points that fall outside of 3 standard deviations, and those might be potential outliers.

Now, I can first try to fit a regression model. We can see from the regression model (appendix B.28) that one of my dummy variables, numbed2 (air bed), have 0 DF; by going back to my frequency table, I found out numbed2 only have two observations; since it does not have enough/sufficient observations and that I have no way to get more observations, I decided to take out numbed2. I tried to fit another model after removing numbed2, the model is shown in appendix B.29.

The full model at this stage is as follows:
ln_price = 3.25398 – 0.00039*host_total_listings_count + 0.09661*accommodates + 0.00007884*security_deposit + 0.00127*cleaning_fee + 0.0027*review_scores_rating – 0.04617*numresponse1 + 0.12028*numresponse2 + 0. 495*numresponse3 + 0.26265*numresponse4 – 1.02072*numsuper + 0.4989*numroom1 + 1.0509*numroom2 – 0.01709*numbed1 + 0.15078*numbed3 – 0.00889*numcancellation1 – 0.00175*numcancellation2 – 0.03417*numregion1– 0.12836*numregion2 – 0.23448*numregion3 – 0.27136*numregion4 + 0.01113*numsuper_review_scores_rating

Where numresponse1 = 1 when host_response_time = ‘within an hour’
numresponse2 = 1 when host_response_time = ‘within a day’
numresponse3 = 1 when host_response_time = ‘within a few hours’
numresponse4 = 1 when host_response_time = ‘a few days or more’
numsuper = 1 when host_is_superhost = ‘t’
numroom1 = room_type = ‘Private room’
numroom2 = 1 when room_type = ‘Entire home/apt’
numbed1 = 1 when bed_type = ‘Real Bed’
numbed3 = 1 when bed_type = ‘Pull-out’
numcancellation1 = 1 when cancellation_policy = ‘strict’
numcancellation2 = 1 when cancellation_policy = ‘flexible’
numregion1 = 1 when Region = ‘IM’
numregion2= Region = ‘SEM’
numregion3 = 1 when Region = ‘NSM’
numregion4 = 1 when Region = ‘WM’
numsuper_review_scores_rating = numsuper*review_scores_rating

(Exp(0.00039)-1)*100 = 0.04
Every unit increase in host_total_listings_count, Price will decrease by 0.04%

(Exp(0.09661)-1)*100 = 10.14
Every unit increase in accommodates, Price will increase by 10.14%

(Exp(0.00007884)-1)*100 = 0.0079
Every 1% increase in security_deposit, Price will increase by 0.0079%

(Exp(0.00127)-1)*100 = 0.09
Every 1% increase in cleaning_fee, Price will increase by 0.09%

(Exp(0.0027)-1)*100 = 0.27
Every 1 point increase in review_scores_rating, Price will increase by 0.27%

(Exp(0.04617)-1)*100 = 4.73
Every change from numresponse1 (within an hour), Price will decrease by 4.73%

(Exp(0.12028)-1)*100 = 12.78
Every change from numresponse2 (within a day), the Price will increase by 12.78%

(Exp(0. 495)-1)*100 = 64.05
Every change from numresponse3 (within a few hours), the Price will increase by 64.05%

(Exp(0.26265)-1)*100 = 30.04
Every change from numresponse4 (a few days or more), the Price will increase by 30.04%

(Exp(1.02072)-1)*100 = 177.52
Every change from numsuper (superhost), Price will decrease by 177.52%

(Exp(0.4989)-1)*100 = 64.69
Every change from numroom1 (private room), the Price will increase by 64.69%

(Exp(1.0509)-1)*100 = 186.02
Every change from numroom2 (Entire home/apt), the Price will increase by 186.02%

(Exp(0.01709)-1)*100 = 1.72
Every change from numbed1 (real bed), the Price will decrease by 1.72%

(Exp(0.15078)-1)*100 = 16.27
Every change from numbed3 (pull-out), the Price will increase by 16.27%

(Exp(0.00889)-1)*100 = 0.89
Every change from numcancellation1 (strict), the Price will decrease by 0.89%

(Exp(0.00175)-1)*100 = 0.18
Every change from numcancellation2 (flexible), the Price will decrease by 0.18%

(Exp(0.03417)-1)*100 = 3.48
Every change from numregion1 (IM), the Price will decrease by 3.48%

(Exp(0.12836)-1)*100 = 13.70
Every change from numregion2 (SEM), Price will decrease by 13.70%

(Exp(0.23448)-1)*100 = 26.43
Every change from numregion3 (NSM), Price will decrease by 26.43%

(Exp(0.27136)-1)*100 = 31.17
Every change from numregion4 (WSM), Price will decrease by 31.17%

(Exp(0.01113)-1)*100 = 1.12
Every change of numsuper_review_scores_rating, Price will increase by 1.12%

At this point, F. value is 108.51, and p-value is less than 0.0001, which means we can reject the null hypothesis; there is at least one predictor that is significantly associated with price (dependent variable).

The value for R^2 is 0.6376 and the value for Adj-R^2 is 0.6318. R-Square of 0.6376 indicates 63.76% of the variation in price is explained by its relationship with its independent variables. Similar to R-Square, Adj R-Sq indicate 63.18% of the variation in price is explained by the regression line. RMSE is now 0.369. For the studentized residuals vs the predicted value (appendix B.29.1). Points are not randomly scattered around the horizontal zero line, and we can see a pattern, which look like a fan shape. At this point, all residual plots don't satisfies assumptions. The plot shows a pattern and looks like the spread is increasing. The studentized residuals vs other predictors does not satisfy constant of variance. When checking for independence, all residuals show a pattern at this stage, points are not randomly scattered. For linearity, we can see that scatterplot from above have a moderate to weak relationship; scatterplot shows a moderate linearity price and cleaning fee. The normal probability plot (appendix B.29.1) of the residuals has points lie around the line; it is not very straight, and more action can be taken to improve this. There are a number of variables that have p-value greater than 0.05. I removed them one by one and I rerun the regression model each time. The appendix B.30 shows current fitted model.

The model now has 11 independent variables after removing the ones that are non-significant. The adjusted R square is now 0.6301, and RMSE of 0.370; both indicators did not change much from before; and all variables have p-value less than 0.05. Next, I computed VIF statistics to check for

collinearity. The variance inflation column from appendix B.31 indicates numroom1 and numroom2 have VIF greater than 10, which indicated multicollinearity existed. As a result, I removed one variable at a time, and the end result is that all VIF are now below 10, which means the issue of collinearity has been resolved (appendix B.32).

After the full model, I will be removing outliers and influential points. Cook's D table has been computed. I decided to remove observations that are both influential point and outliers. The observations I deleted are shown in appendix B.33. I re-run the studentized residuals and cook's D table; and I further delete observations with outlier issues (appendix B.33). Appendix B.34 shows final model after influential points and outliers have been removed. F value has increased to 261.77; adj r sq increase from 0.6301 to 0.6680; and RMSE decreased to 0.336. We can see the final model have 10 variables and all of them are significant.

Next, I analyzed each predictor's influence on log price; as the Parameter Estimate table shows (appendix B.35), accommodates have the highest influence on log price as it has the highest absolute value of standardized estimate; the second highest is numroom2 (entire home/apt).

Studentized residual is up next; from the graphs (appendix B.36) we can see that in comparison to the previous studentized residual, the graphs show significant improvement. Points are more randomly scattered around the zero line. Points are more spread out for the residual plot vs all predictors. The plots of dummy variables still show points around 0, and this is again due to coding. The graph of studentized residual vs the predicted value show that constant of variance and independence is satisfied; we can see from the plot that points are randomly scattered around the horizontal zero line, and there is no pattern. The normal probability plot of the residuals is normal with points lie close around the line; we can see that it is almost a straight line, so it is normal.

The final model from appendix B.34:

ln_price = 3.52172 + 0.10353*accommodates + 0.00008683*security_deposit + 0.000876*cleaning_fee + 0.00455*review_scores_rating – 0.08426*numresponse1 + 0.57472* numroom2 – 0.08369*numregion2 – 0.17846*numregion3 – 0.21107*numregion4 + 0.00043*numsuper_review_scores_rating

Where numresponse1 = 1 when host_response_time = within an hour
numroom2 = 1 when room_type = Entire home/apt
numregion2= Region = SEM
numregion3 = 1 when Region = NSM

numregion4 = 1 when Region = WM

numsuper_review_scores_rating = numsuper*review_scores_rating

(Exp(0.10353)-1)*100 = 10.91
Every unit increase in accommodates, Price will increase by 10.91%

(Exp (0.00008683)-1)*100 = 0.008
Every 1% increase in security_deposit, Price will increase by 0.008%

(Exp (0. 0.000876)-1)*100 = 0.09
Every 1% increase in cleaning_fee, Price will increase by 0.09%

(Exp (0.00455)-1)*100 = 0.46
Every 1 point increase in review_scores_rating, Price will increase by 0.46%

(Exp (0.08426)-1)*100 = 8.79
Every change from numresponse1(within an hour), Price will decrease by 8.79%

(Exp (0.57472)-1)*100 = 77.66
Every change from numroom2 (Entire home/apt), Price will increase by 77.66%

(Exp (0.08369)-1)*100 = 8.73
Every change from numregion2 (SEM), Price will decrease by 8.73%

(Exp (0.17846)-1)*100 = 19.54
Every change from numregion3 (NSM), Price will decrease by 19.54%

(Exp (0.21107)-1)*100 = 23.50
Every change from numregion4 (WSM), Price will decrease by 23.50%

(Exp (0.00043)-1)*100 = 0.04
Every 1% increase in numsuper_review_scores_rating, Price will increase by 0.04%

Of all the variables, reply time of within an hour and regions all have negative association with log price; the greatest one been region WSM, when the listing is not located in WSM, log price will decrease by 23.50%. Therefore, it is confident to say that WSM have higher price than different regions. Also, from the model we can see that room type also play a significant factor in price. When

the room type is not entire home/apt, price will increase by 77.66%, which means entire home/apt have a relatively lower log price.

The next step I performed was cross validation. Appendix B.37 show the result of splitting my dataset into testing and training. As the result shows, after 5-fold cross validation, training set has value of 0.116 and testing set has value of 0.103, test is lower than training. The adj r sq has a value of 0.6637 and RMSE of 0.34. The graph also shows that the test line is lower than the train line, which indicates the model is validated.

The validation model is shown below:

ln_price = 3.470059 + 0.10887*accommodates + 0.000086997*security_deposit + 0.0007*cleaning_fee + 0.005296*review_scores_rating – 0.083287*numresponse1 + 0.566806* numroom2 – 0.087951*numregion2 – 0.190821*numregion3 – 0.188125*numregion4

Where numresponse1 = 1 when host_response_time = within an hour
numroom2 = 1 when room_type = Entire home/apt
numregion2= Region = SEM
numregion3 = 1 when Region = NSM
numregion4 = 1 when Region = WM

(Exp (0.10887)-1)*100 = 11.50
Every unit increase in accommodates, Price will increase by 11.50%

(Exp(0.000086997)-1)*100 = 0.0087
Every 1% increase in security_deposit, Price will increase by 0.0087%

(Exp (0.0007)-1)*100 = 0.07
Every 1% increase in cleaning_fee, Price will increase by 0.07%

(Exp (0.005296)-1)*100 = 0.53
Every 1 point increase in review_scores_rating, Price will increase by 0.53%

(Exp (0.083287)-1)*100 = 8.69
Every change from numresponse1 (within an hour), Price will decrease by 8.69%

(Exp (0.566806)-1)*100 = 76.26

Every change from numroom2 (Entire home/apt), Price will increase by 76.26%

(Exp (0.087951)-1)*100 = 9.19
Every change from numregion2 (SEM), Price will decrease by 9.19%

(Exp (0.190821)-1)*100 = 21.02
Every change from numregion3 (NSM), Price will decrease by 21.02%

(Exp (0.188125)-1)*100 = 23.50
Every change from numregion4 (WSM), Price will decrease by 27.12%

The last step I perform is prediction intervals. I chose 2 sets of values as follows:

Accommodates	2	4
Security_deposit	550	600
Cleaning	100	110
Review_scores_rating	90	85
Numresponse1	1	0
Numroom2	1	1
Numregion2	1	0
Numregion3	0	1
Numregion4	0	0
Numsuper_review_scores_rating	0	0

The result shown in appendix B.38 indicate that the predicted log price for the first set of prediction is $4.6808, which when retransformed back give $10,685.63. 95% confidence interval is ($10,079.90, $11,328.55). The 95% prediction interval is ($5,461.76, $20,816.01). For the second prediction, price turn out to be $12,902.15, 95% confidence interval is ($11,978.35, $13,896.61). The 95% prediction interval is ($6,594.02, $25,154.77).

When an Airbnb listing that have an accommodation 2 people, security deposit of $550, cleaning fee $100, review scores rating of 90, and it is located in SEM (South Eastern Melbourne), the price is predicted to be $10,685.63. The second prediction reveals when an Airbnb listing that have an accommodation 4 people, security deposit of $600, cleaning fee $110, review scores rating of 85, and it is located in NSM (Northern Suburbs Melbourne), the price is predicted to be $12,902.15.

First set of prediction calculations:
(Exp(4.6808)-1)*100 = $10,685.63
(Exp(4.6230)-1)*100 = $10,079.90
(Exp(4.7387)-1)*100 = $11,328.55
(Exp(4.0185)-1)*100 = $5,461.76
(Exp(5.3431)-1)*100 = $20,816.01

Second set of prediction calculation:
(Exp(4.8677)-1)*100 = $12,902.15
(Exp(4.7940)-1)*100 = $11,978.35
(Exp(4.9414)-1)*100 = $13,896.61
(Exp(4.2038)-1)*100 = $6,594.02
(Exp(5.5316)-1)*100 = $25,154.77

C. Shweta

Data Preprocessing & Exploration

The dataset required renaming some of the variables as well as creating dummy variables. Here is a summary of the final variable names with descriptions and dummy variables used in the analysis –

Variable	Description	Dummy variables
res_time	Average time the host responses to a guest inquiry: "anhour", "fewhours", "aday".	For res_time, k=3 so k-1=2 dummy variables
		anhour is the baseline used

		d_res2=1 if res_time="fewhours", 0 for otherwise
		d_res3=1 if res_time="aday", 0 for otherwise
superhost	Whether the host is a superhost: true or false	d_superhost=1 if superhost =t (host is superhost), 0 if superhost=f (host is not a superhost)
total_listings	Number of listings the host has on AirBnB	
room_type	The property type of the listing: home_apt (entire home/apartment) and p_room (private room)	d_room=1, if room_type=home_apt, 0 if otherwise (room_type=p_room))
acc	Max number of guests that can be accommodated	
price	Price per night	
security_deposit	Security Deposit	
cleaning_fee	Cleaning Fee	
review_scores_rating		
can_policy	Cancellation policy (flexible, moderate or strict)	For can_policy, k=3 so k-1=2 dummy variables
		flexible is the baseline used
		d_can2=1 if can_policy="moderate", 0 for otherwise

		d_can3=1 if can_policy="strict", 0 for otherwise
Region	Region in which Airbnb is located	For res_time, k=5 so k-1=4 dummy variables
		IM is the baseline used
		d_reg2=1 if region="NSM", 0 for otherwise
		d_reg3=1 if region="SEM", 0 for otherwise
		d_reg4=1 if region="EM", 0 for otherwise
		d_reg5=1 if region="WM", 0 for otherwise

Created interaction term, sd_sh=d_superhost*security_deposit

The descriptive statistics (figure C.1) shows that the average number of total listings per host is around 16, but this number can be skewed because the maximum number of total listings is 276 which seems a little high. The inter quartile range is 1-13 and this is where most of the values for total listings fall.

The average number of guests accommodated by an AirBnB property is 4, with 1 being the minimum and 16 being the maximum.

The average price per night is AU$ 150, with AU$ ‘0’ being the minimum and AU$ 1501 as maximum. The minimum and maximum values maybe outliers and will need to be removed in the analysis stage since price cannot be AU$=0. The inter quartile range is AU$ 93- AU$ 169, therefore the maximum value of AU$ 1501 seems to be a bit extreme.

The average security deposit charged is AU$ 289 with the interquartile range of AU$ 141-AU$350. The minimum is AU$ 0, which is acceptable since not all Airbnb’s charge security deposit. The maximum security deposit charged is AU$ 5000, which looks like an outlier and may be needing to be removed later.

The average cleaning fee charged is AU$ 70 and the interquartile range is AU$ 35 – AU$ 95.

The minimum is AU$ 0 which is acceptable since some AirBnB's may not charge a separate cleaning fee but instead include it with the price. The maximum cleaning fee charged is AU$ 467, which is very far from the 75-quartile value of AU$ 95. It may be an outlier.

The review score rating range from 20-100, with the inter quartile range of 92-100. The minimum review score rating is 20, which may be an outlier since the 5th quartile value is 80. We can assume that most values for the review score rating are on a higher side.

The histogram for total_listings (Figure C.2), acc (Figure C.3), price (Figure C.4), security_deposit (Figure C.5) and cleaning fee (Figure C.6) are all left skewed. The histograms show outliers as also seen in descriptive statistics (Figure C.1). There may be a need for transformation. The histogram for the variable review_score_listing (Figure C.7) is right skewed.

The observations for hosts who are not superhosts is higher than those are (Figure C.8). The number of observations for region 'IM' is a lot higher than the rest of the regions (Figure C.8). The number of observations for can_policy 'strict' is also a hot higher than 'moderate' & 'flexible' (Figure C.8). The observations for room_type 'apt_home' are more than thrice for room_type 'p_room' (Figure C.8). The number of observations for res_time 'an hour' is more than 10 times higher than res_time 'aday' (Figure C.8).

The scatter plot matrix (Figure C.9) shows that the regression line for price against all the x-variables (quantitative) is fairly linear. There seems to be an outlier and influential point's issue. There doesn't seem to be any collinearity issues within the different quantitative x-variables. Checked for multicollinearity in the Pearson correlation coefficient table (Figure C.10) as well. There is no collinearity even amongst the interaction term sd_sh (d_superhost*security_deposit) with the x-variables, d_superhost & security_deposit, therefore the centering of the said x-variables was not needed.

To check whether a polynomial regression is required instead of the linear regression, linear regression was performed on the full model and residuals were checked. The full model (Figure C.11) shows adj r-square of 39.54% which is low. Residuals were also checked. The fit diagnostics for price (Figure C.12), shows that data is fairly linear and polynomial regression will not be needed. Although, will need to remove outliers and influential points. The residuals plots of x-variables (acc, security_deposit, cleaning_fee, review_score_rating) against the y-variable price are violating the assumptions of independence and constant variance. (Figure C.13, Figure C.14, Figure C.15 Figure C.16. Figure C.17). The normal probability plot (Figure C.18) shows a slight S-shape. Therefore, transformation is required.

Data Analysis

The full model (Figure C.11) is shown below. Adj R-square is 39.54% and RMSE is 91.372. P-value is <.0001, so we can reject the null hypothesis as at least 1 variable is significantly associated with the Y-variable.

Price (y-variable) = -133 – 0.07235*(total_listings) + 21.11167 *(acc) + 0.03725 *(security_deposit) +0.47361 *(cleaning_fee) + 1.51716 *(review_score_rating) + 26.91881 *(d_res2) +16.43878 *(d_res3) +5.72444 *(d_superhost) + 18.19029 *(d_room) -5.04343 *(d_can2) – 6.21231 *(d_can3) – 7.51898 *(d_reg2) -4.45586 *(d_reg3) +10.18351 *(d_reg4) -47.92809 *(d_reg5) + 0.00034658 *(sd_sh) + e

Where

d_res2=1 if res_time="fewhours", 0 for otherwise

d_res3=1 if res_time="aday", 0 for otherwise

d_superhost=1 if superhost =t (host is superhost), 0 if superhost=f (host is not a superhost)

d_room=1, if room_type=home_apt, 0 if otherwise (room_type=p_room)

d_can2=1 if can_policy="moderate", 0 for otherwise

d_can3=1 if can_policy="strict", 0 for otherwise

d_reg2=1 if region="NSM", 0 for otherwise

d_reg3=1 if region="SEM", 0 for otherwise

d_reg4=1 if region="EM", 0 for otherwise

d_reg5=1 if region="WM", 0 for otherwise

sd_sh=d_superhost*security_deposit (interaction term)

None of the vif value is above 10, thus there is no multicollinearity issue. Therefore, no need to remove any variables. The p-value for the interaction term is above 0.05, which means it is insignificant. Not removing it at this stage as it will be removed in the model selection process.

Studentized residuals and Cook's D output was checked for outliers and influential points. Observations that's shown as outliers and influential points were removed. After removing 25 outliers and influential points, the full model was run again (Figure C.19). Adj r-square increased to 43.84% and RMSE was decreased to 62.202 which proved that removing outliers and influential points was needed. P-value is still <.0001, so we can continue to reject the null hypothesis as at least 1 variable is significantly associated with the Y-variable. The Vif values are less than 10 thus showing there are no multicollinearity issues. Checked for residuals at this point, the graphs seem to be getting slightly better but still more outliers and influential points need to be removed. Again, Studentized residuals and Cook's D output was checked for outliers and influential points. Observations that's shown as outliers and influential points were removed.

After removing 24 observations, the full model was run again (Figure C.21). Adj R-square has now increased to 46.52% and RMSE has decreased to 52.789. P-value is still <.0001, so we can continue to reject the null hypothesis as at least 1 variable is significantly associated with the Y-variable. The Vif values are less than 10 thus showing there are no multicollinearity issues. Checked for residuals at this point, the graphs seem to be getting slightly better but still more outliers and influential points need to be removed. Again, studentized residuals and Cook's D output was checked for outliers and influential points.

After removing 21 observations, the full model was run again (Figure C.23). Adj R-square has now increased to 48.50% and RMSE has decreased to 47.859. P-value is still <.0001, so we can continue to reject the null hypothesis as at least 1 variable is significantly associated with the Y-variable. The p-value for the interaction term is still above 0.05, so it will be removed in the model selection process. The Vif values are less than 10 thus showing there are no multicollinearity issues. Studentized residuals and Cook's D output was checked for outliers and influential points. I removed some outliers and influential points again but adj r-square was decreasing therefore didn't proceed with more removal of observations. Checked for residuals at this point and the studentized residuals plots (Figure C.25, Figure C. 26, Figure C.27, Figure C.28, Figure C.29, Figure C.30) though seem to be better but the normal probability plot is slightly s shaped (Figure C.31). The histogram for Price (Figure C.32) is still slightly left skewed although less skewed than the initial histogram (Figure C.4). The scatterplot after removing all the influential points and outliers (Figure C.33) shows better linearity than the initial scatter plot (Figure C.13) Therefore, proceeded with log transformation of the y-variable – price.

After transforming the y-variable price, adj r-square of the full model increased to 58.77% and RMSE decreased to 0.32. (Figure C.34). P-value is still <.0001, so we can continue to reject the null hypothesis as at least 1 variable is significantly associated with the Y-variable. The p-value for the interaction term is still above 0.05, so it will be removed in the model selection process. The Vif values are less than 10 thus showing there are no multicollinearity issues. Studentized residuals and Cook's D output was checked for outliers and influential points. There weren't many outliers and influential points so did not remove any. There were 3 observations with missing values and therefore not used in the regression (Figure C.34). Since the log transformation needs the y variable not equal to 0 and some of the values for price was 0 which were not removed manually. Since it was just 3 observations, did not take any action. The assumptions for independence and constant variance seem to be fairly satisfied at this point (Figure C.36, Figure C.37, Figure C.38, Figure C.39, Figure C.40, Figure C.41). The assumption for normality also is satisfied (Figure C.42). The Histogram looks symmetrical i.e. unimodal (Figure C.43) and scatter plots (Figure C.44) looks fairly linear.

Note: In order to check if I can get better results with transforming the x-variables, I performed sqrt transformation on the quantitative x-variables (total_listings, acc, security_deposit, cleaning_fee and review_score_rating) and the y-variable price. Since there were values=0 for security_deposit and cleaning_fee, I could not proceed with doing log transformations if needed for both x and y variables. I ran a model with only transformed y-variable, sqrt_price, (Figure C.45), a model with transformed x-variables (sqrt_total_listings, sqrt_acc, sqrt_security_deposit, sqrt_cleaning_fee, and sqrt_review_score_rating) (Figure C.46) and a model with all the x and y variables transformed (Figure C.47). None of the models with the sqrt transformed showed better results than the log transformation of y-variable price. The residual plots and scatter plots also didn't show any significant difference. Therefore, decided to proceed with the model where y-variable price has undergone log transformation.

Model Training & Test

The data was divided into Train and Test Sets in 75:25 ratio and stepwise and adj r-square selection methods were used to arrive on the fitted model. Both the models showed the same 11 variables which were significant. Therefore, there was no need for model comparison. (Figure C.48, Figure C.49)

The fitted model (Figure C.50) is with 11 variables with adj r-square of 59.72% and a low RMSE of 0.31385. The P-value is less than .0001, F-value is 195.79. The vif value for all the 11 variables is below 10, thus proving that there are no multicollinearity issues. The standardized estimate shows that number of guests that can be accommodated at an AirBnB seem to be the most influential variable. Assumptions for independence and constant variance are satisfied (Figure C.52 – Figure C.56). The normal probability plot looks normal and linear (Figure C.57). Studentized residuals and cook's D graphs were checked for outliers and influential points but since there weren't many, did not remove.

The Final Fitted model equation based on train set is –

new_y=3.6422 + 0.0899 *(acc) + 0.0001 * (security_deposit) +0.0005 * (cleaning fee) + 0.0033 *(review_score_rating) + 0.0702 *(d_res2) + 0.1143 *(d_res3) + 0.0394 (d_superhost) + 0.5507 *(d_room) -0.216 *(d_reg2) – 0.0487 *(d_reg3) - 0.2994 *(d_reg5)

Where,

d_res2=1 if res_time="fewhours", 0 for otherwise

d_res3=1 if res_time="aday", 0 for otherwise

d_superhost=1 if superhost =t (host is superhost), 0 if superhost=f (host is not a superhost)

d_room=1, if room_type=home_apt, 0 if otherwise (room_type=p_room)

d_reg2=1 if region="NSM", 0 for otherwise

d_reg3=1 if region="SEM", 0 for otherwise

d_reg5=1 if region="WM", 0 for otherwise

Interpretation

The x-variable ‘acc’ (number of guests accommodated in an AirBnB) is positively associated with price. new_y (log of price) will change by 0.0899 for an additional guest accommodation assuming all the other factors are constant. Price will increase 9.4% for every additional guest.
Calculation: (exp (0.0899)-1)*100 = 9.4064

The x-variable ‘security_deposit’ is positively associated with price. new_y (log of price) will change by 0.0001 for an AU$ 1 increase in security deposit. Price will increase by 0.01% for every AU$ increase in security deposit.
Calculation: (exp (0.0001)-1)*100 = 0.01

The x-variable ‘cleaning fee’ is positively associated with price. new_y (log of price) will change by 0.0005 for an AU$ 1 increase in cleaning fee. Price will increase by 0.05% for every AU$ increase in cleaning fee.
Calculation: (exp (0.0005)-1)*100 = 0.05

The x-variable ‘review_score_rating’ is positively associated with price. new_y (log of price) will change by 0.0033 for every review score rating. Price will increase by 0.33% for every unit increase in review score rating.
Calculation: (exp (0.0033)-1)*100 = 0.33

new_y (log of price) will change by 0.0702 when response time is few hours as compared to the response time of an hour. Price will be 7.27% higher if the host’s response time is a few hours as compared to the host's response time of an hour.
Calculation: (exp (0.0702)-1)*100 = 7.27227

new_y (log of price) will change by 0.1143 when response time is a day as compared to the response time of an hour. Price will 12.109% higher if the host’s response time is few hours as compared to the host's response time of an hour.
Calculation: (exp (0. 0.1143)-1)*100 = 12.109

For the x-variable superhost, the price of the AirBnB if the host is a superhost will be 4.019% higher than if the host is not a superhost.
Calculation: (exp (0.0394)-1)*100 = 4.019

For the x-variable superhost, the price of the AirBnB for an entire home/apt will be 73.44% higher is than a private room.
Calculation: (exp (0. 5507)-1)*100 = 73.44

For the x-variable region, the price for an Airbnb in region NSM will be 19.43% lower than price for an Airbnb in IM. The price for an Airbnb in region SEM will be 4.753% lower than price for an Airbnb in IM. The price for an Airbnb in region WM will be 25.87% lower than price for an Airbnb in IM.

d_reg2 (region=NSM) - Calculation: (exp (-0. 216)-1)*100 = -19.43
d_reg3 (region=SEM) - Calculation: (exp (-0.0487)-1)*100 = -4.753
d_reg5 (region=WM) - Calculation: (exp (-0.2994)-1)*100 = -25.87

Computing the prediction on new values (Figure C.60) – Predictions was done for 2 new data observations.

1st data prediction: The scenario is the Airbnb accommodates 2 guests, charges AU$ 300 as security deposit, no cleaning fee, has review score rating of 96, host's response time is within an hour, host is a superhost and Airbnb is In the NSM region.

The model predicts that the price will be AU$ 6041 for the above condition with 95% confidence interval of (AU$ 5331, AU$ 6844) a 95% prediction interval of (AU$ 3036, AU$ 11926). This is a good model as the predicted value falls between the confidence interval. (Figure C.59)

2nd data prediction: The scenario is the Airbnb accommodates 4 guests, charges AU$ 500 as security deposit, AU$ 50 cleaning fees, has review score rating of 100, the host's response time is within a few hours, the host is not a superhost and Airbnb is in the IM region.

The model predicts that the price will be AU$ 15559 for the above condition with 95% confidence interval of (AU$ 13963, AU$ 17334) a 95% prediction interval of (AU$ 7917, AU$ 30485). This is a good model as the predicted value falls between the confidence interval. (Figure C.59)

D. Cody

Data Exploration

Following the data preprocessing stage to get to a full dataset that was free of errors and

included only the relevant information, the data exploration initiated to determine if the selected Y-variable or any of the x-variables would require any sort of transformation. Linear regression was chosen to explain the chosen response variable Y (Price) due to the infinite number of possible values that can be predicted. The initial exploration stages were done with the aid of histograms and boxplots. The initial price histogram (Fig D.1) illustrated it was positively skewed with the mean value of AU$149 and smaller median value of AU$114, with a range of AU$0-AU$2,699. The IQR (middle 50) is AU$71-AU$166, with a standard deviation of +/- AU$163.85. The probability plot (Fig D.2) shows an exponential shaped curve.

In order to stabilize the variance, the transforming of the price variable was required. I tested several different methods of transformation, but log(Y) was the best fit when examining the residual plots. The transformed logPrice histogram (Fig D.3) illustrated it is now appearing more normalized with a symmetrical distribution and shows a mean value of AU$4.72 and a nearly equal median value of AU$4.73, with a range of AU$2.48-AU$7.90. The IQR (middle 50) is AU$4.26-AU$5.11, with a standard deviation of +/- AU$.6958. However, looking at the probability plot (Fig D.4) it was clear that there was a slight 'U' shape in the data, likely due to outliers and influential points that will have to be dealt with later in the analysis.

Half of the variable in the analysis were qualitative and required the creating of dummy variables in order to be able to analyze their impact on the model. I also created an interaction variable from two of the independent variables that combined the max # of the people the rental could accommodate and the security deposit that is required to rent the property, calling the new variable acc_sd (accommodation * security deposit). According to iGMS (2018), 59% of AirBnB hosts require a security deposit that typically shouldn't exceed more than 20% of the total cost of the booking in order to protect themselves in the event that property damage occurs.

Frequency tables were utilized to ensure that all the dummy variables were being correctly coded and represented in the dataset (Fig D.10). Host_response_time showed that half of all the AirBnB property owners respond back to their potential guests within 1 hour. There was also a good majority (33% of the population) of the responses that were not tracked and were designated as an 'N/A' value. Looking at whether the AirBnB property owner was a super host (host_is_superhost) showed that the large majority, 76%, are not super hosts. The type of rental that people are choosing (room_type) shows that 63% of the rental properties are the entire house/apartment. The remainder were private rooms (35%) or shared rooms (2%). Bed types that were listed on with the rental was nearly all 'real beds', 99%. Cancellation policies ranged from strict (40%) to flexible (33%) to moderate (27%). From the 5 regions, Inner Melbourne had over half (53%) of the locations. Southeast was the second highest region with 22% of the listings. The remaining 3 regions all had 10% of less

of the total listings.

I utilized boxplots to explore the interactions that the qualitative variables play against price. The boxplot for price and superhost (Fig D.5) shows that hosts that are designated with that classification show a slightly smaller price IQR (AU$4.46-AU$5.19) , but their mean value is higher than hosts that are not super hosts (AU$4.84 vs. AU$4.69). The boxplot for price and room type (Fig D.6) shows that the mean (AU$5.06) and IQR (AU$4.69-AU$5.29) are well above renting a private room or shared room with a mean around AU$4 and an IQR AU$3.50-AU$4.20. The boxplot for cancellation policy (Fig D.7) shows that strict and moderate are nearly equal from a mean (AU$4.80) and IQR perspective (AU$4.4-AU$5.1). Hosts with a flexible cancellation policy have a lower mean price (AU$4.50) as well a much wider range in the IQR (AU$3.99-AU$5.00). The boxplot for price and response time (Fig D.9) shows that hosts that respond with an hour have a higher mean value (AU$4.80) and very narrow IQR (AU$4.45-AU$5.13) when compared against the other response times. Hosts that take a few days or more to answer had the lowest mean price (AU$4.59) and also the biggest IQR (AU$4.00-AU$5.26), with an overall whisker being smaller than all other response intervals.

During the bivariate analysis, a matrix scatterplot (Fig D.11) was used to compare the dependent variable of logPrice against the quantitative independent variables to determine if any relationship would be represented graphically. Dummy variables that were created for the qualitative variables were not included because they serve no purpose in the analysis since they are only comprised of 1's or 0's and you can't interpret a relationship from that. logPrice and accommodates show are fairly decent linear relationship, but it appears that after analyzing the residuals it showed a decent number of outliers that required to be removed. The Pearson correlation table was ran and it confirmed that none of correlation values are above the 0.9 threshold, meaning there is no issue with multicollinearity. The interaction term acc_sd (accommodation * security deposit) was rather high at .854 but was still acceptable to remain in the model due to being underneath the threshold.

Fitting the Model

Following the exploratory phase, the full model can now be created with the following output as the result:

logPrice = 4.54267 -0.00006422(host_total_listing_count) + .11341(accommodates) + 0.0000573(securirty_deposit) + 0.00014904(clearning_fee) - 0.00178(review_score_rating) - .08330(dResp1) + .02069(dResp2) - .0089(dResp3) + 0.12063 (dResp4) - 0.06004(dSuper) - 0.56145 (dRoom1) - 1.06659(dRoom2) - .73506(dBed1) - .24269(dBed2) - .07064(dBed3) - 0.62216(dBed4) - .0032(dCan1) - .00274(dCan2) + 0.23545(dReg1) + 0.14641(dReg2)

+ .04504(dReg3) + 0.15735(dReg4) + 0.00002664(acc_sd) + e

Where:

dResp1=(host_response_time="within an hour");

dResp2=(host_response_time="within a few hours");

dResp3=(host_response_time="within a day");

dResp4=(host_response_time="a few days or more");

dSuper=(host_is_superhost="f");

dRoom1=(room_type="Private room");

dRoom2=(room_type="Shared room");

dBed1=(bed_type="Couch");

dBed2=(bed_type="Pullout-Sofa");

dBed3=(bed_type="Futon");

dBed4=(bed_type="Airbed");

dCan1=(cancellation_policy="Strict");

dCan2=(cancellation_policy="moderate");

dReg1=(Region="IM");

dReg2=(Region="SEM");

dReg3=(Region="NSM");

dReg4=(Region="EM");

acc_sd=accommodates*security_deposit;

All variables in the full model have a VIF statistic < 10 and a tolerance above >.1, meaning there is no multicollinearity among the variables. The starting adjusted R^2 for the full model is .5511, which means that 55.11% of the variation in Y (price) can be explained by the model. Also, the model contains a low F-value of 133.43. the P-value is <.0001, so we can reject the null hypothesis, because at least 1 of the variables is significantly associated with the effect on price.

The initial data set included 2,482 observations. During the removal of the outliers and influential points, observations that had a high Studentized residual > |3|, Cook's D > (4/n), and/or a hat h_{ii} value > 0.5 were flagged and removed from the dataset. Following several rounds of removing the flagged influential points and outliers, the model's adjusted R^2 jumped up to .6409, which is an improvement of .0898 from the original dataset. The F-value also increased to 188.28 (+54.85). The RSME decreased to .38816 from .46623 in the original dataset (Fig D.10)

Training & Testing

The dataset was divided into training and test sets, with a 70 (training)/30 (test) breakout along with a random seed value. Stepwise (Fig D.13) and Forward (Fig D.14) methods were applied

to the training dataset in order to fit the model. They both resulted in the same 11 significant variables, so it wasn't necessary to compare the different models side by side. The resulting data showed and improved adjusted-R^2 of .6486, a lower RMSE .1498, p-value <.001, and a higher F-value of 281.70. The normal probability plot looks linear and normal. No other outliers or influential points were needed to be removed at this point.

The final fitted model (Fig D.20) equation base on the training set:
logPrice = 4.50773 + .11415(accommodates) - 0.00129(review_score_rating) - 0.07557(dResp1) + 0.17691 (dResp4) - 0.09559 (dSuper) - 0.56355 (dRoom1) - 1.10662 (dRoom2) + 0.19286(dReg1) + 0.11484(dReg2) + 0.13424(dReg4) + 0.00003923(acc_sd) + e
Where:
dResp1 = 'Within an hour';
dResp4 = 'A few days or more';
dSuper = 'f';
dRoom1 = 'Private room';
dRoom2 = 'Shared room';
dReg1 = 'IM' Inner Melbourne;
dReg2 = 'SEM' South Eastern Melbourne;
dReg4 = 'EM' Eastern Melbourne

The validation tests (Fig D.16) identified that the CV R^2 result (0.0294) was less than 0.3, meaning that the model is valid and can be used for new data added to the dataset to make predictions. A 5-Fold Cross Validation (Fig D.17/D.18) that utilized the stepwise approach was also ran and the ASE (Train) .15175 > ASE (Test) .15059, proving that it is a valid model.

Interpretation
The x-variable 'accommodates' is positively associated with price. logPrice will change by 0.11415 for each additional guest that is able to stay, assuming all other variables held constant, price will increase by 12.09% for every additional guest.

The x-variable 'review_score_rating' is negatively associated with price. logPrice will change by -0.00129 for every review score rating point, assuming all other variables held constant, price will decrease by .12892% for every unit increase in review score rating

logPrice will change by -.07557 when the response time is 'within an hour' as compared to the host's response time not being captured by the data. Price will be 7.28% lower than compared to a host time that wasn't recorded.

logPrice will change by .17691 when the response time is ‘a few days or more’ as compared to the host’s response time not being captured by the data. Price will be 17.69% higher than compared to a host time that wasn’t recorded.

The x-variable ‘superhost’ is negatively associated with price. logPrice will change by -0.09559 if the host IS NOT a designated super host, assuming all other variables held constant, price will decrease by 9.116 %.

logPrice will decrease by -.56355 when the room type is a private room as compared to an entire house/apt. Price will be 43.08% lower than compared to an entire house/apartment

logPrice will decrease by -1.106662 when the room type is a shared room as compared to an entire house/apt. Price will be 66.93% lower than compared to an entire house/apartment

The x-variable region, the price of the Inner Melbourne (IM) region will be 21.27% higher than price for a place in Western Melbourne (WM). The price of the South Eastern Melbourne (SEM) region will be 12.17% higher than price for a place in Western Melbourne (WM). The price of the Eastern Melbourne (EM) region will be 14.37% higher than price for a place in Western Melbourne (WM).

Computing New Prediction Values

Scenario #1: Accommodates 6 guests, has an overall review rating of 91%, is not designated a superhost, responds within an hour, the type of room is a private room, and it’s located in Inner Melbourne.

The model predicts the average nightly price will be AU$ 9,599 with a 95% CI value between AU$9,073 - AU$10,156 and 95% PI of AU$4,413 – AU$20,747. This is a good model as the predicted average nightly price falls within both the CI and PI ranges.

Scenario #2: 4 guests, overall rating of 100%, is designated a superhost, responds within a few days, room type is a shared room, and it’s located in South Eastern Melbourne.

The model predicts the average nightly price will be AU$ 5,741, with a 95% CI value between AU$4,771 - AU$6,906 and 95% PI of AU$2,566 – AU$12,699. This is a good model as the predicted average nightly price falls within both the CI and PI ranges.

E. Brendan

Choosing Initial Variables

The data first needed to be reviewed and cleaned up before it could be useful. The 96 variables needed to be slimmed down to be more manageable. Going through the different variables, some were eliminated for being text only and not measurable, some for being irrelevant or redundant such as the longitude and latitude of a location as well as the zip code. Others were eliminated as they could not be broken down into quantifiable variables or dummy variables. Once completed, the optimal number of variables for this study was 12. These variables were then imported and through infile assessed into SAS 9.4 and. A final variable, observations was eliminated as it did not offer any value when deciding on price.

The final set of variables chosen were host_response_time, host_is_superhost, host_total_listings_count, zip code, room_type, accommodates, bed_type, price, security_deposit, cleaning_fee, review_scores_rating, cancellation_policy and Region. Each of these went through a linearity test, through the sgsprocess matrix where their grouping, progression, movement, and distance from the x or y graph lines were reviewed. Then they were analyzed for range, mean, and their $1^{st}2^{nd}$and 3^{rd}quartile breakdown.

Linearity and Skewness

Noted here were that most of the variables were not linear. The best viable independent variable graphs, ones with a positive linear progression, were security deposit and cleaning fee, while accommodates, a list of how many possible spaces available was a close secondary variable for potential linearity. Ones with the least likely relationship with the DV were review scores rating which progressed in a negative progression from the opposite end of the spectrum and was heavily clumped together and host total listings. Region showed some possible relationship as well but needed to be broken down into dummy variables for a better analysis.

Histograms of the various quantitative variables indicated skewness, peaks and range and gave some insight into range. First observed was the DV, dependent variable and y-variable, price. Heavily skewed towards the left, the mode of observations was around $200 dollars with over 50% of the observations falling within that range. 40% of the observations fell below a $200 price towards $0, which did not seem reasonable and most likely indicated missing data. In order to better measure the v-variable going forward it was transformed into the square root of price as there were observations with the price of 0 which would not have been transformed into logarithmic form and been readable. (see e.1) The end result moved the bell curve inwards, spreading out the range, kept at one peak. This was then the new DY. (see e.2)

Accommodates, review score ratings, host total listings and security deposit were viewed as histograms as well. These three independent continuous variables indicated a right skew leaning. Accommodations was spread out over a very tight range, from 0 up to 15.5 and skewed to the left, meaning a lot of the variation was not captured and could affect the price. (see e.5) The post transformation spread the bell curve with multiple peaks but a much better spread (see e.6) see

Review score ratings is heavily skewed to the right with a median of 95 out a 100. The transformation of the variable did not resolve the skewness of the variable and created gaps between observation and so the variable was kept in the original form.

Both the cleaning fee variable and the accommodations variables were heavily skewed to the left and transformed via the square root function. The data observations thereafter were much uniform, had an overall single peak and were more spread out

The histogram for the original dependent variable showed a heavily skew towards the left with a single peak. The minimum was recorded at $0 and the maximum was $1200. No potential outliers were observed at this point. Post square root transformation the model is noted as being much more spread out and having one peak. Square root was chosen over a logarithmic transformation as there were $0 price observations noted which exclude logarithmic as an option.

Transformations and Dummy variables

Following the linearity tests for the other quantitative variables' accommodations, room type and home listing numbers were also transformed in order to achieve linearity. Each transformation achieved a better distribution of observations. While the distribution of the observations and the independent variable did achieve better visuals and a stronger model, the matrix of relationship still showed a weak linear correlation between the dependent variable and the independent transformed variables. (see e.3)

Qualitative variables were identified as host_response_time, host_is_superhost, zip code room_type, bed_type and cancelation policy. These non-continuous variables were could be of profound influence by type or listing and were broken down into individual dummy variables. The importance of each variable was measure against the DV, sqrt_price. (see.8)

The independent variable of region had three options and so two dummy variables were created. The baseline chosen was within a few days. The first dummy variable was not applicable, and the second dummy variable was within a week.

Response time had three viable options and so two dummy variables were created. The base line was chosen as entire home. The most occurring option of private room was assigned the dummy variable one and the other option of a shared room was given the second dummy variable.

Bed type had four different options where the base line was chosen as a real bed. This made up 70% of the options offered. The first dummy variable was assigned to Airbed, there was one option listed. The second dummy variable chosen was the futon and third assigned to the pullout couch.

Cancellation policy had three different options. The two dummy variables chosen were first the flexible option offered by the landlord and the second variable was assigned to the moderate option chosen by the landlord. The baseline chosen was the most occurring strict policy

Region was made up of over 100 different options which were broken down into 6 regions. The baseline was EM. The first variable chosen was N/A, not given. The second region variable was IM, the third region NSM, the fourth SEM and the final region chosen was WM

Interrelated Variables

Looking over the variables, potential synergies and interactive variables could potentially be teased out by connecting some of the data points to make a join variable, a combination of either a dummy variable and another variable or a combination of two independent variables. The best two key relationships identified were a combination of the security deposit/accommodation and host reviews and host total listings.

The first relationship is hypothesized as highly correlated as the chance of the related high number of rooms requiring some form of deposit could lend an argument for a price change. The new variable created and analyzed was sec_acc. It was injected into the portion of the code that defined dummy variables.

Another potential relationship of combined variables was the host total listings, a set of two dummy variables, and review scores ratings. The combination of the two should give insights into how price can be tied to both the availability of having multiple listings, either a few days, N/A or within a day. The two new variables created here were host1_rev and host2_rev. (see e.9)

These three different interrelated variables were tested first on their correlation

and then on the VIF of above the threshold**.** The three variables were highly correlated with other variables. Both host response times dummy variables correlated high with a score a .9 with the two different reviews score rating and host response time interactive variables. The accommodates and security variable also had a high, over .9. score with security deposit. These variables were then trimmed down and used in a regression model to see what the VIF. Here they did not pass the p-value scores. (see e.10, e.11)

Replacing or Deleting Missing Data

Following the regression testing of the model, the fit test was assessed and deemed low. (see e.7) An influence attachment was made in order to identify which observations might also be influential. Following a rigorous review, the NPP indicated that the model was not following normality and there was a curve visible rather than a straight line. (see e.13)

There was a total of 600 observations at this point throughout the model that were not being read. This was due to missing data for the variables at different point which severely hindered the analysis of the model. For missing accommodations, sqrt_acc, the mean was chosen as 0 possible locations at a location was not likely. For missing price data points, the mean was also chosen. For missing cleaning fees and security deposit the mean was also chosen as a 0 would severely hamper the data. The data was changed in order to better analyze the outcomes and make sure the information in the model contained as many observation points as possible.

Observations and Influence Points

1stWave there were 39 observations that were outside the standardized residual zero, both above the 3 standard deviations and below. They were analyzed each individually to see if there were naturally occurring and found to be causing too much influence on the linear progression of the model. The regression model was then again tested. (see e.12)

2ndWave found that there were 35 observations that were both outliers and influence points which were removed. They did not occur naturally either. The allowance of variance recognized through the R2 of the model increased by 5% from 49 to 54% and the fitness improved to 156.

3rdWave of observations found that there were 22 observations that were both influencing the data linearity and that were needing to be removed. The fitness increased to 168 and the R2 to 57.22.

4thWave had only 7 observations that needed to be changed. They increased the fitness to

168.98 and the r2 to 58.57. The following observations decreased the fitness and adj-r2 and r2 of the model.

Final Model Regression

The 5th model regression post removal of a final 5 different observations that were deemed influential at creating deviations in the linear aspects resulted in the final positive changes to the fit test score and the R2.

The final count was fitness of 181 and an r2 of .59 and there were now 2385 observations left of the original 2500. (e.14)

Test and Train Samples (e.15)

Once the model had no more influential points and observations that were altering the data, the information was split into both a training section and test section. The test section was to have 20% of the data and the training had 80%. This was done at random using the splitting code of samprate. The data selected for the test was given a 1 next to the observation going forward. The split is being done in order to best identify how strong the model work

New DV

A new DV was selected for this data set to better analyze the relationship between the various x variables and the y variable. The new name for this variable was New_Lnp. This new data set was then tested and analyzed.

Backward Validating

Post addition of the new variable New_Lnp, the backward model was broken down into key variables host total listings, sqrt acc security deposit sqrt clean, review scores rating, d_rty1, d_rty2, d_bty1, d_bty3, d_cpol2, d_reg2, d_reg3. The test set of this model was then used to predict what the value of the missing DY is expected to be through the p=yhat command. This command would give a predicted value for the new y variable based on the information in the test set. This only occurs where there is no current value for the DY, where there is missing data, in the non-selected set of the training section for the split data. A total of 478 observations where now part of the training set.

The difference between the observed and the predicted was analyzed by subtracting the

difference between what is observed and what was predicted in the above section of the test set. The difference was assigned to a new variable on the table d. Using the difference, we could then compute the predictive root error, mean and mean absolute error. The RMSE and MAE give a stronger indication of the predictive power and strength of this overall model in accounting for the variation. The results show a RMSE of 2.12951 for the root mean square error and a MAE of 1.64890 for the training set of data for the measurement of the accuracy of the continuous variables. This measured the possibility that there are errors occurring and being seen in the model itself. RMSE was larger than MAE so that was good.

The Pearson correlation indicated that the R2 test was .74459^2 and the R2 for the train was .6031. Subtracting the r2 train from the r2 test results in a difference of .049132.

Stepwise Validating

Post addition of the new variable New_Lnp, the selection model was broken down into key variables host total listings, sqrt acc security deposit, sqrt clean, review scores rating and the two dummy variables for host response time. The test set of this model was then used to predict what the value of the missing DY is expected to be through the p=yhat command. This command would give a predicted value for the new y variable based on the information in the test set. This only occurs where there is no current value for the DY, where there is missing data, in the non-selected set of the training section for the split data. A total of 478 observations where now part of the training set. (e.17)

The RMSE and MAE give a stronger indication of the predictive power and strength of this overall model in accounting for the variation. The results show a RMSE of 2.12916 for the root mean square error and a mae of 1.64892 for the training set of data for the measurement of the accuracy of the continuous variables. This measured the possibility that there are errors occurring and being seen in the model itself. RMSE was larger than MAE so that was good. The numbers here though were the same for both models.

A Pearson correlation of the coefficients chart for the DY variable and the y-hat prediction showed a yhat of .74459 for the training set for the stepwise.

The earlier Regression test had the train R-Square of .5935. This R2 test, the yhat^2 indicated a score of .5544. Subtracting the r2 train and the r2 test set brings a score of .048586

Selected Data results:

	Backward	Stepwise		
Train				
RMSE	2.01342	2.03802		step is lower
R2	0.6048	0.5935		same
ADJ-R2	0.6021	0.5923		step is lower
GOF	Yes	Yes		
Residuals	Yes	Yes		
TEST				
RMSE	2.12951	2.12916		step is lower
Mae	1.6489	1.64892		same
R2	0.553967604	0.554414268		higher
adjr2	Compute using formula	Compute using formula		
CV-R2	0.050832396	0.03908573		coeficient is lower for stepwise
		Winner Stepwise		

Having looked over the various different set of Data, the Training set of the Backward method had the lower RMSE, had a higher adj-r2 and the other conditions were the same. For the test set though, the RMSE was better for Stepwise, the R2 was higher and most importantly the coefficient of the variance was lower. This identified the stepwise method as being the better of the two as the test set is the more important of the two

Narrowing the Field

For the prediction section for this model, both models were used to see if the end results would differ and therefore give a substantially different price.

Prediction

Backward model predictions of the sqrt_price were made with an $800 dollar security, a rating of 90,, a super host rating of 1, meaning a good host, a d_host_r1 of 1, meaning land lords got back to questions from the renters of Not/applicable, the room that is shared, d_rty2 is active, a pull out sofa couch and moderate cancellation policy, is not located in the NSM region of Melbourne, the sqrt_accommodations were of 2.7, an indicator of around 5 rooms available in the listing and sqrt_clean of 12, meaning a fee of $144 dollars.

The expected price came to be 11.60. A price of $134.56 a night. The range of viable options would be between 9.96 and 13.23, of the sqrt_price variable. This means the price can be expected to be between $99 and $175 for this selection of options. (e.19)

Selection Stepwise model predictions were based on the total listings count of the host being 9, the security deposit of $300, a review scores rating of 90, there being a private room, d_rty1, no shared room, accommodations for 4 people, sqrt_acc of 2, and a cleaning fee of 11, $121. (e.19)

The predicted price came to be 10.9225, $119.30. The range of expected is between 10.368 and 11.4768, $107 and $132. (e.20).

Results-Best Model (b.21)

DY (sqrt_price) =3.557 + 2.5337(sqrt_acc) + .151(sqrt_clean)
((-.00651(host_total_listings_count) + 000924(security_deposit) + 0233(review_scores_rating) - 2.2936(d_rty1) - 4.0123 (d_rty2)) ^2

The results indicated that all else remaining constant, for every room introduced by the host, the price should go by 2.5, this means $6.25. For event 1 unit of cleaning fee the price will be expected to increase by the Australian dollar equivalent of .02. If it is a single bed ,d_rtly1, the price will decrease by $5.24 and if it is shared room, d_rty2 it is will decrease by 16. For every increase in the deposit the price is expected to increase by .00081. For every increase in the number of listings, the price will increase by a fraction of a percent

F. Ying

In application to my sample, I started with plotting a histogram for univariate analysis to predict Airbnb listing prices in the city of Melbourne, Australia against 11 other independent variables. Out of the 2,491 selected properties, the central tendency for predicted price was at AU$150.7, which is fairly low. Properties of an average amount AU$0 has indicated a very low in the pool, while some had much higher average price of AU$8,000. With the middle 50% of predicted price between AU$69 and AU$168 among all selected properties, the median of AU$112 is slightly closer to lower quartile. The spread of the distribution was very large with a range of AU$8,000. The mean of distribution was greater than the median, so the distribution is said to be right/positively skewed and unimodal (figure F.1). The distribution also had a peaked top with lighter tails. There is a potential outlier when predicted price was at AU$8,000, which is to the right of the graph. Skewed distribution will not generate a good predictive analysis. Therefore, predicted price must be transformed with log.

After applying log transformation, figure F.2 shows a much more symmetrical and normal distribution for Inprice (transformed predicted price). The mean of distribution of Inprice (AU$4.72) is now the same as the median. Properties with Inprice on an average amount of AU$2.64 are in the lowest, while some have a high amount of AU$8.99. With the middle 50% of Inprice between AU$4.23 and AU$5.12 among all selected properties, the median is about right in the middle of lower

and upper quartiles. The spread of the distribution is much smaller than predicted price before transformation, with a range of only AU$6.35. The distribution has a flat top/ heavy tails (Kurtosis<3) with potential outliers to the right of the graph as well.

Majority of the independent variables in this dataset are qualitative and therefore dummy variables are computed to better analyze the data. At the same time, an interaction variable for two selected independent variables dHRT1 and dHIS is built. According to Shatford (2018), to qualify as a super host, one should host a minimum of 10 stays in a year, respond to guests quickly and maintain a 90% response rate or higher, have at least 80% 5-star reviews, etc. Therefore, a super host should have a joint effect with the average time the host responses to a guest's inquiry is "within an hour" on Inprice. Consumers tend to choose Airbnb properties with high review scores and if the host is a super host. These properties are believed to be more pleasant and host is nice and reliable. Therefore, Inprice is believed to be higher. The newly created interaction term is called "HRT_superhost".

In the interest of length limitation, I only explore the dataset for two variables dRT1 and dRG4 with boxplots. Boxplots are built to evaluate how Inprices vary by dRT1 (Private room) and dRG4 (NSM). From figure F.3, it is obvious that the middle 50% of properties of entire home/apartment (dRT1=0) and properties of private room (dRT1=1) in the sample are quite different. Both types of properties seem to have low Inprices since the boxes were closer to the lower extremes. 75% (upper quartile) of selected properties of entire home/apartment is under AU$5.5 and the remainder 25% were up to AU$8.1. For properties of private room, the box is much lower than properties of entire home/apartment, with its 75% (upper quartile) lower than the lower quartile of properties of entire home/apartment. The 75% of properties of private room is about AU$4.2 and the remainder 25% are up to AU$9. Given much longer whisker for properties of private room, it is interpreted that it varies wider in Inprice from AU$2.5 to AU$9. Properties of entire home/apartment swings less from AU$2.8 to AU$8.1. The means for both room types were very close to their medians. Mean and median for properties of entire home/apartment overlapped, which indicates the distribution of Inprice should be normal and symmetric. Relatively, the mean is higher than the median so the distribution of Inprice for properties of private room is slightly skewed right.

From figure F.4, it is obvious that the middle 50% of properties in region IM (dRG1=0) and properties in NSM (dRG1=1) in sample are quite different. Both types of properties seemed to have low Inprices since the boxes were closer to the lower extremes. 75% (upper quartile) of selected properties in IM was under AU$5.5 and the remainder 25% are up to AU$9. For properties in NSM, the height of the box was longer than properties in IM, i.e. wider difference between its 1st and 3rd quartiles. The 75% of properties in NSM was about AU$5.5 and the remainder 25% were up to AU$7. Given much longer whisker for properties in IM, it was interpreted that it varied wider in predicted

Inprice from AU$2.8 to AU$9. Properties in NSM swung less from AU$2.8 to AU$7. The means for both room types were very close to their medians. Mean and median for properties of in IM overlapped, which indicates the distribution of Inprice should be normal and symmetric. Relatively, the mean is higher than the median so the distribution of Inprice for properties in NSM is slightly skewed right.

For bivariate analysis, a scatterplot matrix was built for each variable against the dependent variable Inprice to observe the patterns displayed and the relationship between all independent variables within a single matrix. Dummy variables are excluded from the matrix for two reasons: first, there were too many dummy variables in my sample. Inclusive of all dummy variables would make all plots squeeze together within a small graph and made it difficult to observe the pattern, Secondly, dummy variables are qualitative variables and their points would just scatter along 0 or 1. Their scatterplots are not appropriate or meaningful to check for association. For interaction variable HRT_superhost in figure F.5, it is a product of two dummy variables dHRT1 and dHIS and hence it is also not meaningful to check for association as well.

With reference to figure F.5, the scatterplot showed positive linear relationship between Inprice and quantitative variables host_total_listings_count, accommodates, security_deposit, cleaning_fee and review_scores_rating. For variable *accommodates*, its scatterplot indicated a fairly strong association because most points followed a clear form. There could be outliers at the top left corner. Variable cleaning_fee has a slightly weak association because most points followed some forms, but some spread toward the center. There could be outliers at the top left corner. Scatterplots for host_total_listings_count, security_deposit and review_scores_rating indicated that they had a very weak relationship with Inprice. Most points spread out the graphs.

At the end of data exploration stage, a full model is fitted through linear regression. The full model statement is as follows:

*Inprice = 4.628 – 0.0001*host_total_listings_count + 0.126*accommodates + 0.0001*security_deposit + 0.0005*cleaning_fee – 0.002*review_scores_rating – 0.038*dHRT1 + 0.04*dHRT2 – 0.013*dHRT3 + 0.153*dHRT4 + 0.011*dHIS – 0.548*dRT1 – 1.012*dRT2 – 0.162*dBT1 – 0.237*dBT3 – 0.206*dBT4 + 0.003*dCP1 + 0.001*dCP2 – 0.236*dRG1 – 0.053*dRG2 – 0.006*dRG3 – 0.268*dRG4 + 0.214*HRT_superhost + e*

where dHRT1=1 when host_response_time='within an hour',
dHRT2=1 when host_response_time='within a few hours',
dHRT3=1 when host_response_time='within a day',
dHRT4=1 when host_response_time='a few days or more'

dHIS = 1 when host_is_superhost = "t",
dRT1 = 1 when Room Type = "Private room",
dRT2 = 1 when Room Type = "Shared room",
dBT1=1 when bed_type = 'Futon',
dBT3=1 when bed_type = 'Airbed',
dB4=1 when bed_type = 'Couch'
dRG1 = 1 when Region = "NSM",
dRG2 = 1 when Region = "SEM",
dRG4 = 1 when Region = "WM",
HRT_superhost = dHRT1 dHIS*
(dBT2 (Bed type = pull-out sofa) is set to 0 by SAS because of not enough observations)

According to the parameters estimate in figure F.6, beta weights indicates significant effect on Inprice. The higher the beta weight, the more significant effect on Inprice. However, t-test is a better methodology to measure significance of variables than the ranking of beta. Using the t-test on each model parameter, variables accommodates, security_deposit, review_scores_rating, dRT1-2, dRG1-2 and dRG4 had significant influence on Inprice (p-values for t-test are smaller than 0.05). On the other hand, the p-values for the rest of the independent variables, such as host_total_listings_count, cleaning_fee, dHRT1-4, dHIS, dBT1-4, dCP1-2, dRG3 and interaction variables were larger than 0.05. Therefore, we cannot reject the hypothesis that these variables have no effect on Inprice, which should be removed from the model.

F-Test Hypotheses:
H_o: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$
H_a: At least one coefficient $\beta_j \neq 0$

F=131.84 with p-value smaller than 0.05 (at alpha=0.05). The null hypothesis of no association between Inprice and other variables is rejected. We accept the alternative hypothesis that at least one coefficient of the independent variables has significant effect on Inprice. F-test gives strong support to the all variables.

The coefficients of determination of R^2 (54.06%) and adj-R^2 (53.65%) represent the amount of variation in Inprice explained by the regression model. For this analysis, about 53% of the variable in Inprice is explained by the model. To check if the model is good, we should look at adj R^2 of 53.65% because it does not increase with the addition of an independent variable that does not improve the regression model.

Next, the full model is taken to fit diagnostics to analyze if the model violates the 4 model assumptions. Referring to figure F.7 residual plots for each variable, predicted value against inprice and figure F.8 normality graph. Residual plots for dummy variables and the interaction term are qualitative variable and their points are scattering along 0 or 1, which are not appropriate or meaningful for residual analysis and its assumptions. From figure F.7, the spread of the seven plots (host_total_listings_count, accommodates, security_deposit, cleaning_fee and review_scores_rating) are randomly scattered around the zero line, showing constant variance and independence. Even though there are some clusters in host_total_listings_count, , security_deposit and cleaning_fee towards these variable amounts are close to zero , these plots are considered to be normal because security_deposit and cleaning_fee usually will not be very expensive and within a small range and host_total_listings_count also will not be very high because the dataset only consists of 8-year data. For accommodates, it is normal as well because properties are typically small to accommodate less than 5 people. Review_scores_rating is normal because the rating is a subjective preference. The plot for predicted value is random and scattering along the zero line. The plots of host_total_listings_count, accommodates, security_deposit, cleaning_fee and review_scores_rating are linear because the pattern of the spread shows a straight line. Figure F.8 shows almost 45-degree line, which indicates the model is normally distributed and linear.

In figure F.9, severe multicollinearity will be detected if an independent variable has a correlation value more than absolute 0.9 with another for independent variable. In the model, the data does not seem to have this issue. After checking for multicollinearity, I check for outliers and influential points. In figure F.7, we could see some outliers in each independent variable residual plot, where outliers are beyond the +3/-3 bands. By referring to Studentized Residual and Cook's D (figure F.10) for Inprice, observations with arrowhead are indicated as both outliers and influential points, which are needed to removed first. After removing #413, regression is rerun again to check for the next observation to remove until there is no improvement by removing observations from the model. In total, I removed 5 observations from the model and record the changes in R^2 and adj-R^2 (figure F.11). It is realized that both R^2 (0.5569) and adj-R^2 (0.5529) are not high, meaning 55% of variables of predicted Inprice can be explained using the sample predictors. I stop removing observations because there is no further significant improvement to the model.

Following the outlier and influential point removal steps and checking for severe multicollinearity, the dataset is split into train set and test set with a ratio of 75% to 25% respectively. The train set is used to build the final model, while the test set is used to validate the model. The dependent variable is now called new_y of the train set. Figure F.12 shows the result after the split. The sample size of the train set now has 1865 observations. The train set is then taken to fit the final

model by comparing two model selection methods: forward and backward methods. According to figure F.13, forward method selected 14 variables to the final model, while backward method only selected 9 variables. Looking into the summary of forward selection, the partial R^2 becomes very small after the 9^{th} variable (dRG2) is added to the model. Comparing the selected variable from backward method and the first 9 variables from forward method, they are actually the same, i.e. accommodates, security_deposit, review_scores_rating, dHRT4, dRT1-2, dRG1-2, 4. The key to select the "best" variables/ model is to select fewer variables and a higher R^2. In this case, my final model factors in the variables that are suggested by backward selection method, except for security_deposit and review_scores_rating. I decided to exclude these two variables from my final model because the parameter estimates were so small that they are believed to have no effect on new_y. My final model is taken to regression again.

According to the parameters estimate (figure F.14), 6 variables are with p-values for t-test are smaller than 0.05, which means that variables accommodates, dHRT4, dRT1, dRT2, dRG1 and dRG4 have significant influence on new_y. However, dRG2 now becomes insignificant having p-value (0.0657) of t-test larger than 0.05 and it should be removed from the model. Then regression should be rerun. In figure F.15, all 6 variables are now with p-values for t-test are smaller than 0.05. Then I can conclude my final model statement after model selection as follows:

*new_y = 4.50 + 0.14*accommodates + 0.22*dHRT4 – 0.55*dRT1 – dRT2 – 0.22*dRG1 – 0.30*dRG4 + e*

where dHRT4= 1 when host_response_time = "a few days or more",

dRT1 = 1 when Room Type = "Private room",

dRT2 = 1 when Room Type = "Shared room",

dRG1 = 1 when Region = "NSM",

dRG4 = 1 when Region = "WM"

Since Airbnb listing prices was transformed with log in the beginning, I have to retransform independent variable new_y in order to interpret the final model statement. Variable accommodates is positively associated to new_y. Model shows that assuming all other variables constant, for any additional guests that the property accommodates, predicted price for a night increases by 15.02% computed as $100*(e^{0.14}-1) = 100*(1.1502-1) = 15.02$.

The two dummy variables for host response time show that expected price for a night varies depending on the host response time to guests' inquiries. Thus, on average, predicted price for a property whose host responds to inquiries within a few days or more is $100*(e^{0.22}-1)$ = 24.61% higher than predicted price for a night for a property whose host does not respond.

The parameter estimates for the two dummy variables for room type show that expected price for a night varies depending on the room types. Thus, on average, predicted price for a property with a private room is *(100*($e^{-0.55}$-1) = -42.31%)* 42.31% lower than predicted price for a night for a property with the entire house/ apartment; and predicted price for a property with a shared room is *(100*(e^{-1}-1) = -63.21%)* 63.21% lower than for a property with the entire house/ apartment.

Similarly, for the two dummy variables for regions show that expected price for a night varies depending on the regions. Thus, on average, predicted price for a property in region NSM is *(100*($e^{-0.22}$-1) = -19.75%)* 19.75% lower than predicted price for a night for a property in IM; and predicted price in WM is *(100*($e^{0.3}$-1) = -25.92%)* 25.92% lower than for a property in IM.

One way to analyze the strongest or most influential variable is to refer to standardized estimate. Since the beta is normalized by standard deviation of the dependent variable, all coefficients will have the same until of measurement. Therefore, the values of the standardized coefficients can be compared. It indicates that dRT1 is the most influential variable as it has the highest absolute standardized coefficient of 0.55.

F-Test Hypotheses:

H_o: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

H_a: At least one coefficient $\beta_j \neq 0$

F=350.49 with p-value less than 0.05 (at alpha=0.05). The null hypothesis of no association between new_y and other variables is rejected. We accept the alternative hypothesis that at least one coefficient of the independent variables has significant effect on predicted new_y. F-test gives strong support to 6 variables.

The coefficients of determination of R^2 (53.12%) and adj-R^2 (52.97%) represent the amount of variation in new_y explained by the regression model. For this analysis, about 53% of the variable in Inprice is explained by the model. To check if the model is good, we should look at adj-R^2 of 52.97% because it does not increase with the addition of an independent variable that does not improve the regression model. To conclude, adj-R^2 of an average rate has indicated that this is an average model.

The next step is to analyze if model built by train set has satisfied the model assumptions. Again, we can ignore scatterplots for dummy variables. The predicted value plot does not show a strong form of pattern. Like full model assumption analysis, the spread of residual plot between new_y and accommodates are randomly scattered around the zero line and shows linearity because the pattern of the spread follows a straight line. The normality graph shows almost 45-degree line,

which indicates the model is normally distributed and linear. Then we should check if the train set has severe multicollinearity. In figure F.15, VIF of all variables are less than 10, which means severe multicollinearity does not seem to be a problem.

After checking for multicollinearity, I continue checking for outliers and influential points. Based on the residual plot of new_y for train set, we could see some outliers in each independent variable residual plot (figure F.16), where outliers are beyond the +3/-3 bands. By referring to Studentized Residual and Cook's D for new_y (figure F.18), observations with arrowhead are indicated as both outliers and influential points, which are needed to removed first. After removing #519, regression is rerun again to check for the next observation to remove until there is no improvement by removing observations from the model. In total, I removed 4 observations from the model and record the changes in R^2 and adj-R^2 (figure F.19). It is realized that both R^2 (0.5569) and adj-R^2 (0.5529) are not high, meaning 55% of variables of predicted new_y can be explained using the sample predictors. I stop removing observations because there is no further significant improvement to the model.

Following testing regression result of the train set, I then proceed to measure predictive performance of the final model using test set. By referring to figure F.20, it shows the model of the test set is a better case because CV-R^2 is 0.2 (less than 0.3), RMSE is smaller and both R^2 and adj-R^2 is 0.2 higher than train set. Therefore, the final model by train set is said to be validated.

After validating the final model, I run regression again to evaluate the model performance with 5 indicators, check if it satisfied model assumptions, has severe multicollinearity and more outliers or influential points needed to be removed. By referring to figure F.21, all independent variables are with p-values for t-test are smaller than 0.05. This means that variables accommodates, dHRT4, dRT1, dRT2, dRG1 and dRG4 have significant influence on new_y.

F-Test Hypotheses:

H_o: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

H_a: At least one coefficient $\beta_j \neq 0$

F=371.7 with p-value less than 0.05 (at alpha=0.05). The null hypothesis of no association between new_y and other variables is rejected. We accept the alternative hypothesis that at least one coefficient of the independent variables has significant effect on predicted new_y. F-test gives strong support to 6 variables. It still indicates that dRT1 is the most influential variable as it has the highest absolute standardized coefficient of 0.55.

The coefficient of determination of R2 (54.63%) and adj-R2 (54.48%) represent the amount

of variation in new_y explained by the regression model. For this analysis, about 54% of the variable in new_y is explained by the model. To check if the model is good, we should look at adj-R2 of 54.48% because it does not increase with the addition of an independent variable that does not improve the regression model. To conclude, adj-R2 of an average rate has indicated that this is an average model.

The next step is to analyze if the final model has satisfied the model assumptions. Again, we can ignore scatterplots for dummy variables. In figure F.22, the predicted value plot does not show a strong form of pattern. Similar to the full model, the spread of residual plot between new_y and accommodates are considered to be randomly scattered around the zero line and shows linearity because the pattern of the spread follows a straight line. The normality graph shows almost 45-degree line, which indicates the model is normally distributed and linear. Then we should check if the train set has severe multicollinearity. In figure F.21, VIF of all variables are less than 10, which means severe multicollinearity does not seem to be a problem.

There are outliers that are captured in the residual plot between predicted value and new_y, especially when there is a data point to the top left of the plot. I checked the source and this property offers a private room in a townhouse and the stay should be around March 2018, which is the summer in Australia. Therefore, this outlier could be explained by the holiday season. I decided not to remove this outlier since this only point cannot affect the regression line much. After removing 4 observations from train set, my final model is restated as follows:

*new_y = 4.50 + 0.14*accommodates + 0.22*dHRT4 – 0.55*dRT1 – 1.11dRT2 – 0.21*dRG1 – 0.31*dRG4 + e*

where dHRT4= 1 when host_response_time = "a few days or more",

dRT1 = 1 when Room Type = "Private room",

dRT2 = 1 when Room Type = "Shared room",

dRG1 = 1 when Region = "NSM",

dRG4 = 1 when Region = "WM"

Since Airbnb listing prices was transformed with log in the beginning, I have to retransform independent variable new_y in order to interpret the final model statement. Variable accommodates is positively associated to new_y. Model shows that assuming all other variables constant, for any additional guests that the property accommodates, predicted price for a night increases by 15.02% computed as $100*(e^{0.14}-1) = 100*(1.1502-1) = 15.02$.

The two dummy variables for host response time show that expected price for a night varies depending on the host response time to guests' inquiries. Thus, on average, predicted price for a

property whose host responds to inquiries within a few days or more is $100*(e^{0.22}-1) = 24.61\%$ higher than predicted price for a night for a property whose host does not respond.

The parameter estimates for the two dummy variables for room type show that expected price for a night varies depending on the room types. Thus, on average, predicted price for a property with a private room is *$(100*(e^{-0.55}-1) = -42.31\%)$* 42.31% lower than predicted price for a night for a property with the entire house/ apartment; and predicted price for a property with a shared room is *$(100*(e^{-1.11}-1) = -67.04\%)$* 67.04% lower than for a property with the entire house/ apartment.

Similarly, for the two dummy variables for regions show that expected price for a night varies depending on the regions. Thus, on average, predicted price for a property in region NSM is *$(100*(e^{-0.21}-1) = -18.94\%)$* 18.94% lower than predicted price for a night for a property in IM; and predicted price in WM is *$(100*(e^{-0.31}-1) = -26.66\%)$* 26.66% lower than for a property in IM.

Using the final model to predict the average Airbnb listing prices, the condition is where the property accommodates 5 guests, host responds to inquiries within a few days or more, the room type is with a private room and the property is in WM. The model predicts average price per night ($e^{4.55}$) = AU$9,363. The predicted average price is within the 95% confidence interval between AU$7,648 ($e^{4.35}$) and AU$11,458 ($e^{4.75}$). Therefore, the final model is said to be a good model.

Another condition is where the property accommodates 2 guests, host does not respond to inquiries, the room type is with entire house/ apartment and the property is in IM. The model predicts average price per night ($e^{3.45}$) = AU$3,050. The predicted average price is within the 95% confidence interval between AU$2,505 ($e^{3.26}$) and AU$3,747 ($e^{3.65}$). Therefore, the final model is said to be a good model.

Model Comparison

	R-square	Adj R-square	# of IV
Andy	0.6703	0.6685	10
Theresa	0.6700	0.6680	10
Shweta	0.6003	0.5972	11
Cody	0.6407	0.6391	11
Brendan	0.5444	0.5405	4

Ying	0.5463	0.5548	6

We have selected Andy and Theresa's models as the best models for recommendations as the R-square and Adj r-square is the highest and with fewer variables.

Future Work

The study investigated key factors that affect Airbnb listing price and our analysis are based on 22,895 observations. From the analysis we have so far, it seems like the price of Airbnb listings have many determinants, such as different room type, bed type, location where the property is located, whether the host is a super host or not, etc., all of them are price determinants. We know that price is a vital topic and important factor when considering which listings to choose from. However, our finding suggests that none of the predictors have a high correlation with price, all predictors we examined so far have some sort of relationship with price, but evidence is still needed to determine the factors affecting the price. In the future, it is necessary to know the top factor affecting Airbnb's lodging price; something that we did not have the opportunity to explore include property amenities and reputation, which is something that should be investigated for a better understanding of the price.

The Pearson correlation value between all independent variables and price are relatively low, with the highest value of 0.65, meaning it is only moderately correlated with price. The regression model that we have been using is able to capture some critical characteristics of price. Listings that are located in Western Melbourne (WM) have the highest price, followed by Northern Suburbs Melbourne (NSM), and then, South Eastern Melbourne (SEM). Listings that are located in Western Melbourne (WM) has the lowest price, which we could consider as not having a significant effect on price. In addition, our analysis found out that majority of Airbnb listings have real bed, therefore, analyzing the price with different bed type may be unnecessary.

There are certain limitations to our study. Missing data appeared repetitively which makes our analysis difficult to proceed. For example, for variable "region," we have five recorded levels: IM (Inner Melbourne), WM (Western Melbourne), EM (Eastern Melbourne), SEM (South Eastern Melbourne), and NSM (Northern Suburbs Melbourne), and our analysis have shown that, within these five levels, three of them have proved to have a significant negative impact on the response variable, "price." This shows that the place where the Airbnb located has a significant impact on price; however, there are many missing data within the variable "region," and we do not know that if our final result will be affected when all the missing data has been assigned to a specific region.

Although we already got plenty of predictors for the dependent variable, listing price; there

are still ways that we can explore in our future studies for a better understanding of listing price. For future study, we would like to add in more variables that the potential customers will likely to be interested in when searching for Airbnb. For instance, proximity to public transit or pets are allowed to bring. We believe getting such kind of data could benefit us by obtaining and analyzing customer preferences and better predicting the listing price of Airbnb in a given area.

References

A. Andy

Aindrila Ghosn, Mona Nashaat, James Miller, Shaikh Quader, Chad Marston. 12/04/2018.
A comprehensive review of tools for exploratory analysis of tabular industrial datasets
https://www.sciencedirect.com/science/article/pii/S2468502X18300561

Manfred Te Grotenhuis, Paula Thijs
Dummy variables and their interactions in regression analysis: examples from research on body mass index
https://arxiv.org/ftp/arxiv/papers/1511/1511.05728.pdf

Ankit Peshin, Sarang Gupta, Ankita Agrawal. 12/10/2018.
Exploratory Data Analysis and Visualization of Airbnb Dataset
http://www.columbia.edu/~sg3637/airbnb_final_analysis.html

B. Theresa

Austin, Z., Sutton, J. (2015). Qualitative Research: Data Collection, Analysis, and Management. *US National Library of Medicine National Institute of Health*, 68(3): 226–231.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4485510

Colubi, A.M., Kontoghiorghes, E.J. (n.d.). Computational Statistics & Data Analysis. *International Association of Statistical Computing,* SSN: 0167-9473.
https://www.journals.elsevier.com/computational-statistics-and-data-analysis

Simpson S. H. (2015). Creating a Data Analysis Plan: What to Consider When Choosing Statistics for a Study. *The Canadian journal of hospital pharmacy*, *68*(4), 311–317.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4552232

C. Shweta

Scott Shatford, April 2018, *What is Airbnb's Superhost Status Really Worth?* https://www.airdna.co/blog/airbnb_superhost_status

iGMS, May 2018, *Airbnb Security Deposit: To Charge or Not to Charge?* https://www.airgms.com/airbnb-security-deposit/

Teubner, T., Hawlitschek, F., & Dann, D. (2017). PRICE DETERMINANTS ON AIRBNB: HOW REPUTATION PAYS OFF IN THE SHARING ECONOMY. *Journal of Self-Governance and Management Economics, 5*(4), 53-80. Doi: http://dx.doi.org.ezproxy.depaul.edu/10.22381/JSME5420173

D. Cody

Torti, F., Perrotta, D., Riani, M., Cerioli, A. (2019) Assessing Trimming Methodologies for Clustering Linear Regression Data. *Advances in Data Analysis and Classification 13(227),* 227-257. https://link.springer.com/article/10.1007/s11634-018-0331-4

Perez-Sanchez VR, Serrano-Estrada L, Marti P, Mora-Garcia R-T. (2018) The What, Where, and Why of Airbnb Price Determinants. *Sustainability* 10(12):4596. https://www.mdpi.com/2071-1050/10/12/4596

iGMS, May 2018, *Airbnb Security Deposit: To Charge or Not to Charge?* https://www.airgms.com/airbnb-security-deposit/

E. Brendan

SAS Tutorials: Frequency Tables using PROC FREQ. 2019. Kent State University. https://libguides.library.kent.edu/SAS/Frequencies

SAS Institute Inc. 2019. Visited: June 1st, 2019 http://support.sas.com/documentation/cdl/en/lrcon/62955/HTML/default/viewer.htm#a000695119.htm

Sas Institute Inc. 2019. Visited May 25th. 2019 http://support.sas.com/documentation/cdl/en/lrcon/62955/HTML/default/viewer.htm#a002316433.htm

Biven, Josh. “The economic costs and benefits of Airbnb” Economic Policy Institute. 2019. https://www.epi.org/publication/the-economic-costs-and-benefits-of-airbnb-no-reason-for-local-policymakers-to-let-airbnb-bypass-tax-or-regulatory-obligations/

F. Ying

Terry Rawnsley and Laura Schmahmann. 2018. *What impact does Airbnb have on the Sydney and Melbourne housing markets?* https://www.sgsep.com.au/publications/what-impact-does-airbnb-have-sydney-and-melbourne-housing-markets

Dominik Gut and Philipp Herrmann. 2015. *Sharing Means Caring? Hosts' Price Reaction to Rating Visibility*

https://aisel.aisnet.org/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1053&context=ecis2015_rip

Scott Shatford. 2018. *What is Airbnb's Superhost Status Really Worth?* https://www.airdna.co/blog/airbnb_superhost_status

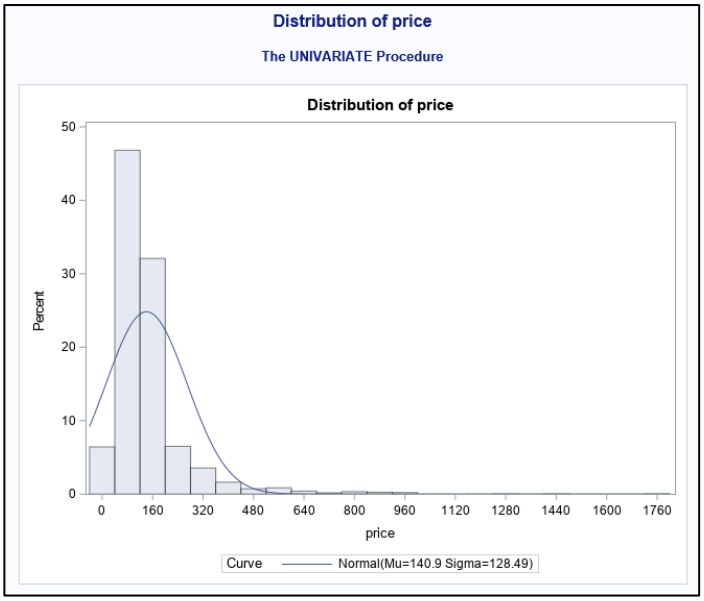
Leslie A. Christensen. *Introduction to Building a Linear Regression Model* The Goodyear Tire & Company, Akron Ohio

Andy Krause and Gideon Aschwanden. 2017. *A Census of Melbourne's Airbnb Market* https://cpb-ap-se2.wpmucdn.com/blogs.unimelb.edu.au/dist/b/193/files/2017/03/censusAnalysis-2hd2q7t.pdf

Bob. 2019. *Living in Melbourne* https://www.bobinoz.com/living-in-australia/melbourne

Appendix A – Andy

A.1 – Histogram



A.2 – Distribution of price

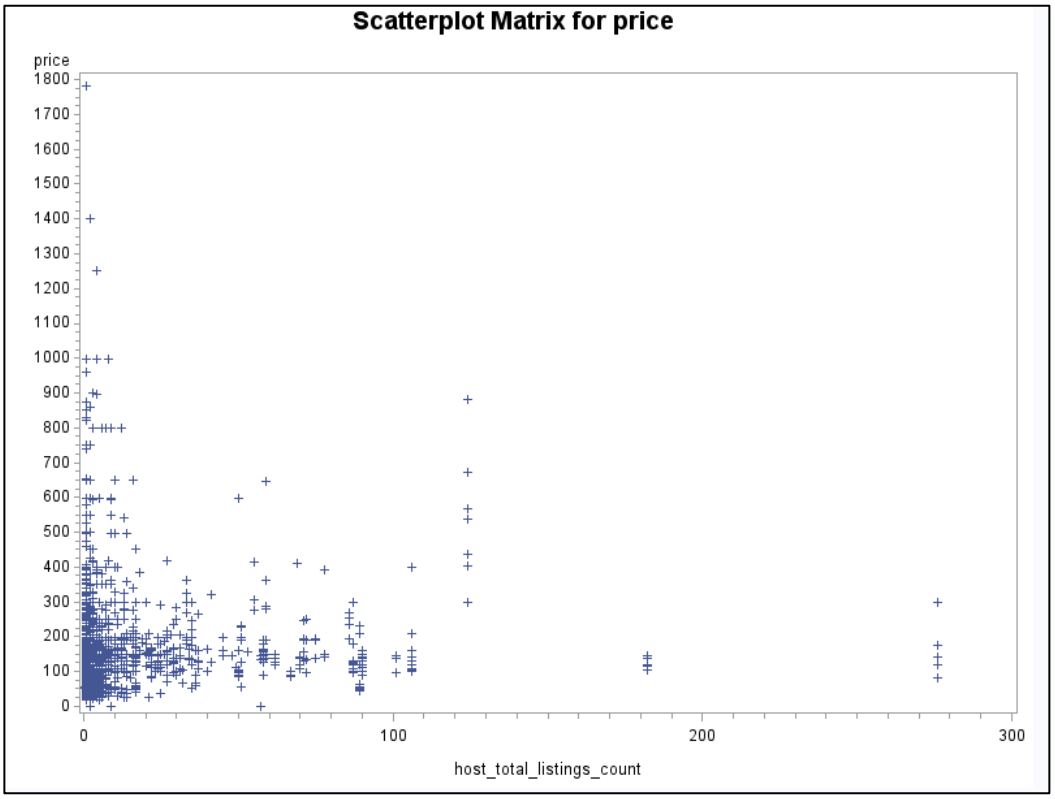
Distribution of price

The UNIVARIATE Procedure
Variable: price

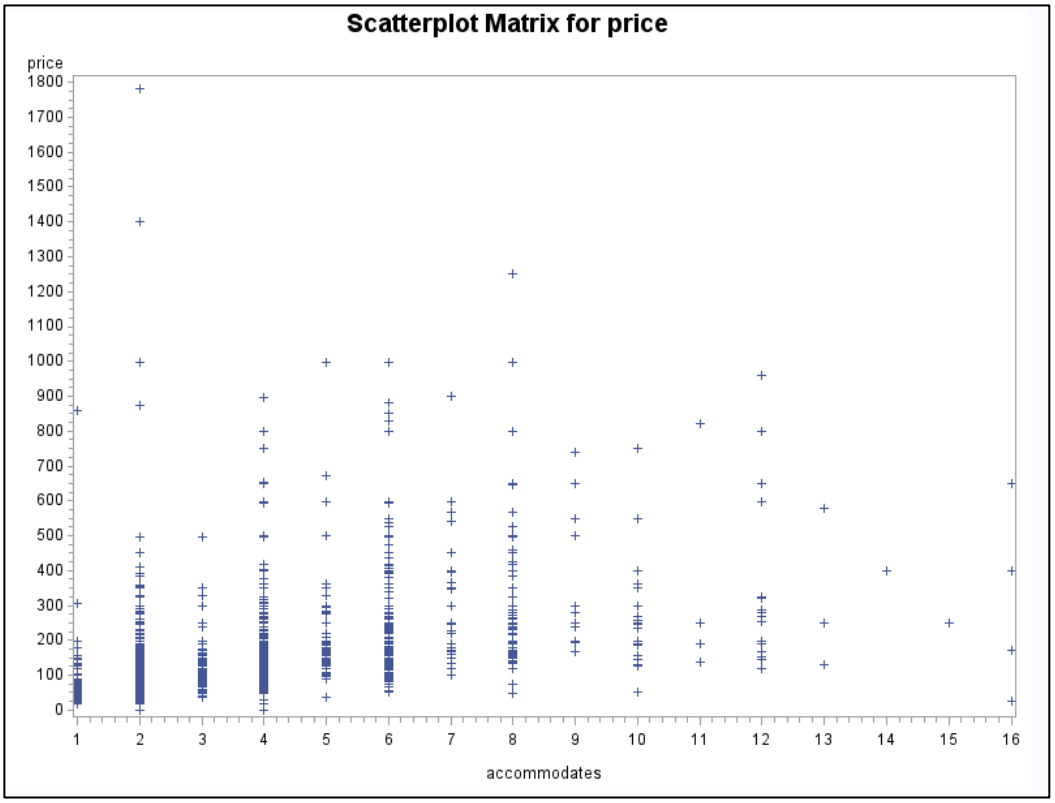
Moments			
N	2478	Sum Weights	2478
Mean	140.897498	Sum Observations	349144
Std Deviation	128.489617	Variance	16509.5817
Skewness	4.00614005	Kurtosis	26.4658237
Uncorrected SS	90087750	Corrected SS	40894234
Coeff Variation	91.1936827	Std Error Mean	2.58117461

Basic Statistical Measures			
Location		Variability	
Mean	140.8975	Std Deviation	128.48962
Median	109.0000	Variance	16510
Mode	100.0000	Range	1784
		Interquartile Range	94.00000

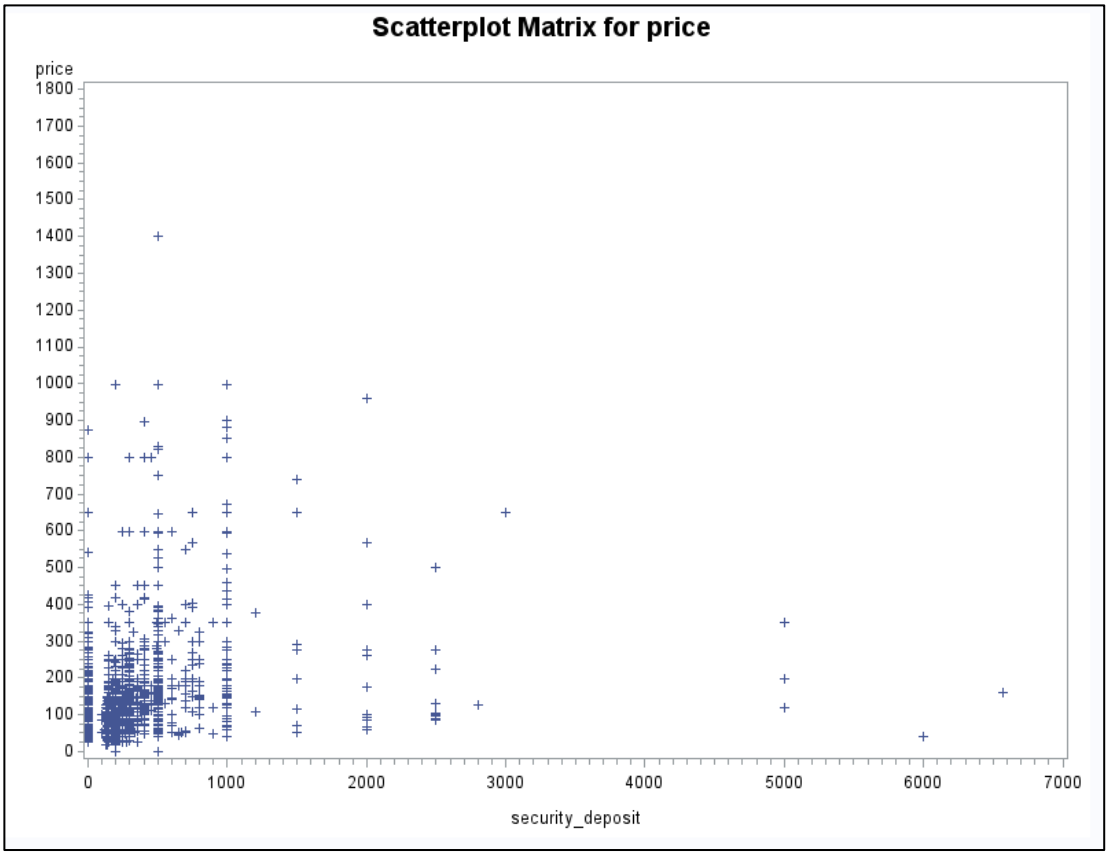
A.3 – Scatterplots (price versus host_total_listings_count)



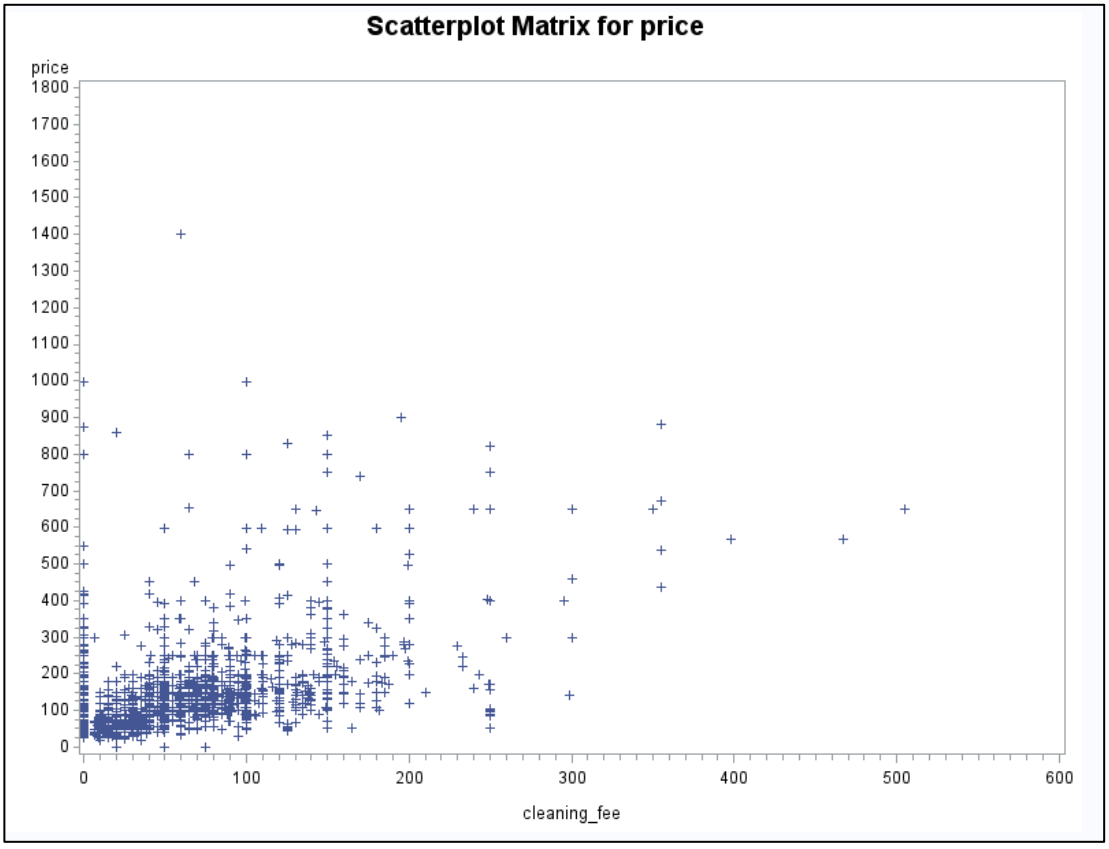
A.4 – Scatterplots (price versus accommodates)



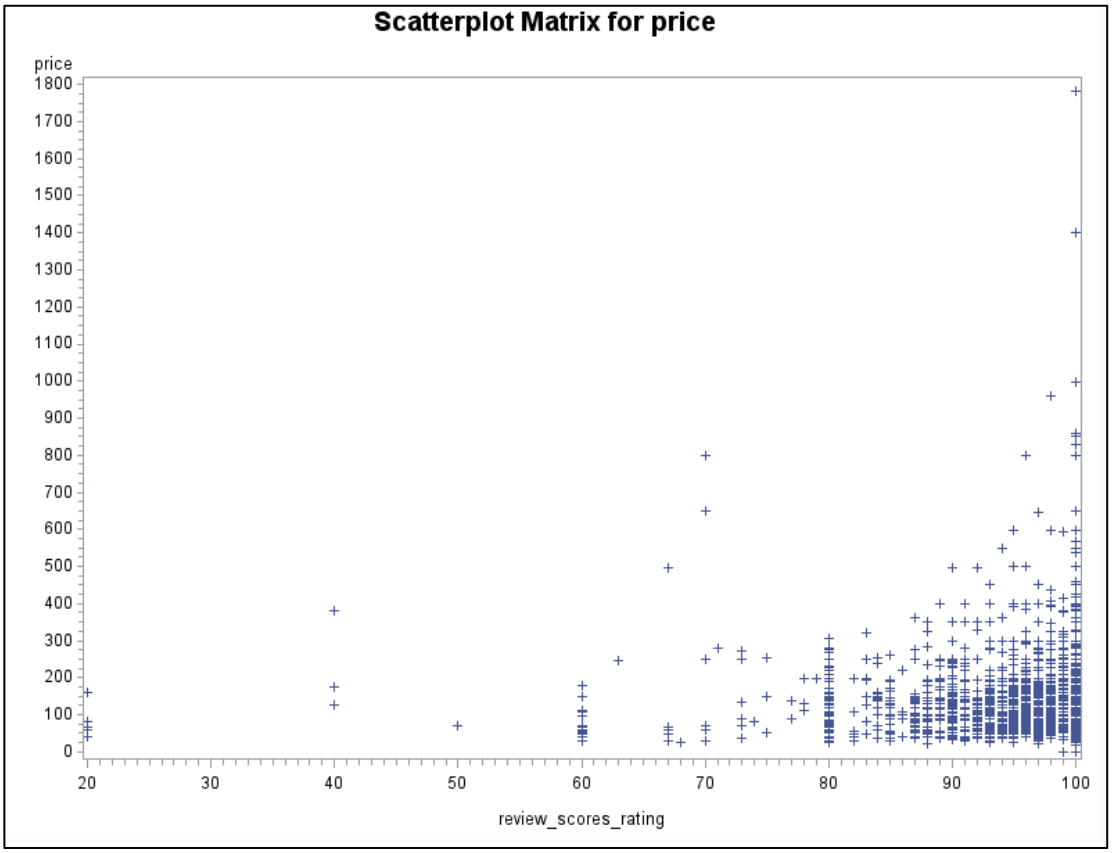
A.5 – Scatterplots (price versus security_deposit)



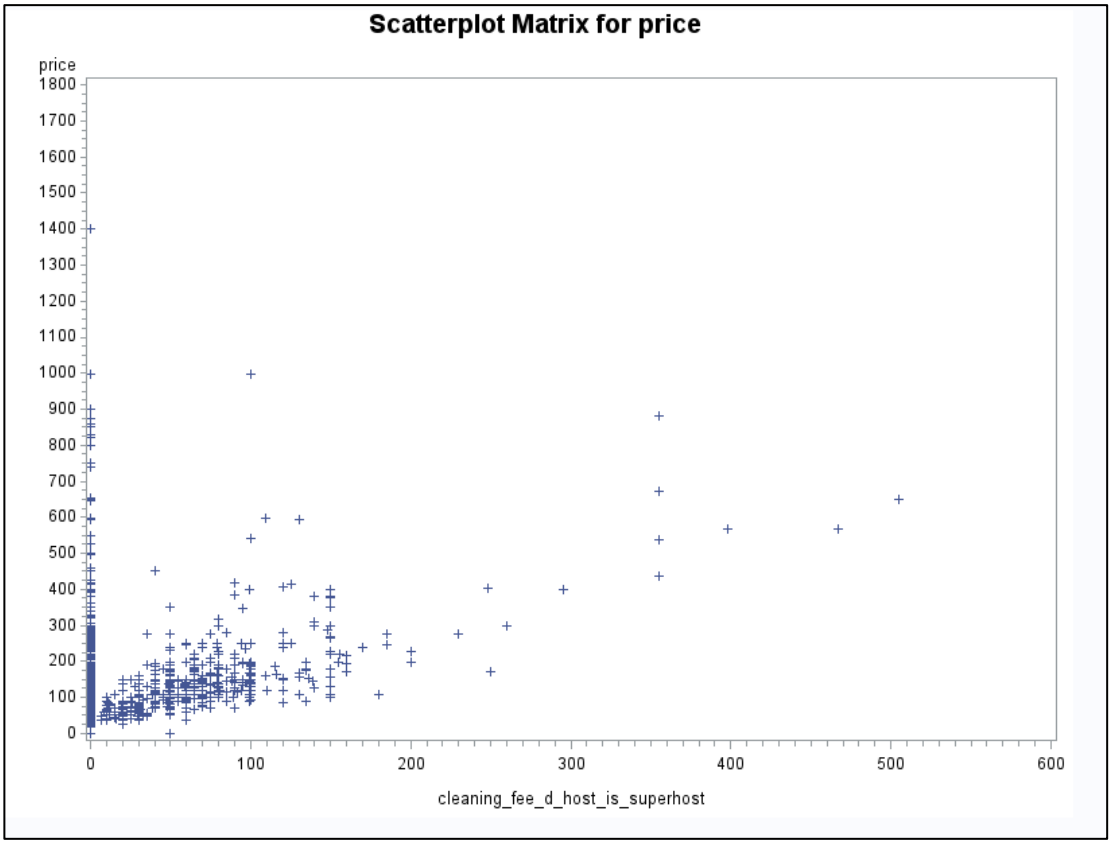
A.6 – Scatterplots (price versus cleaning_fee)



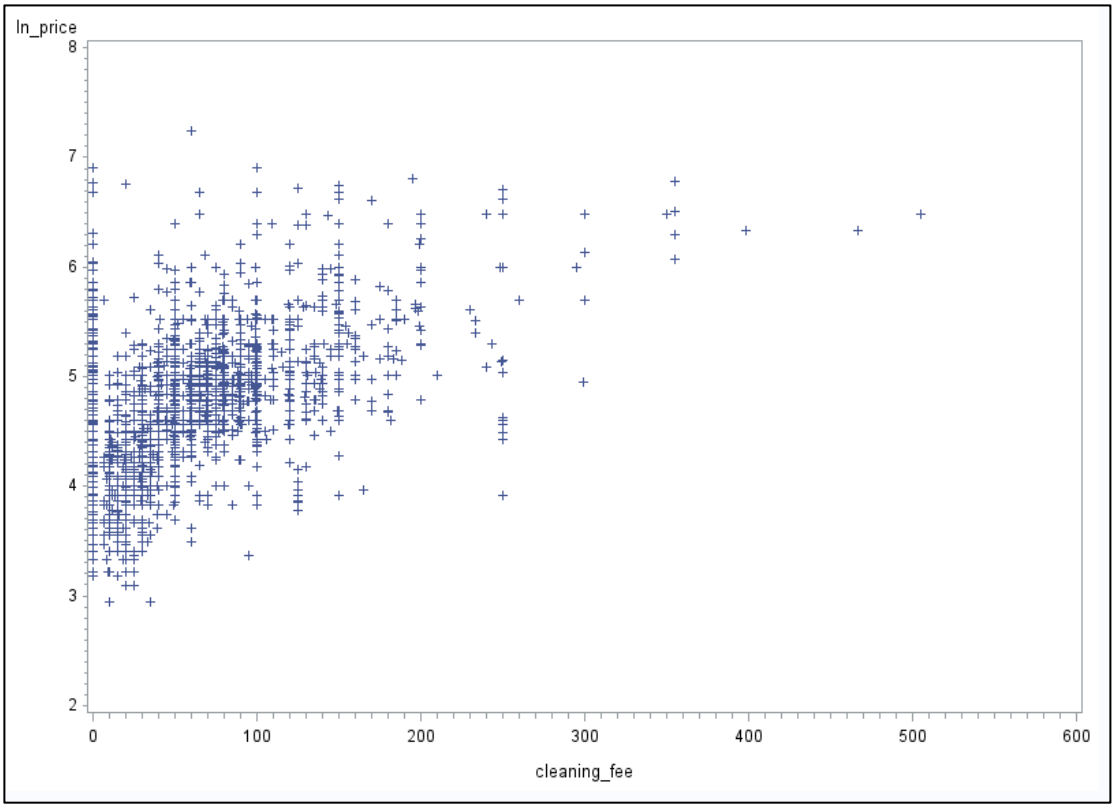
A.7 – Scatterplots (price versus review_scores_rating)



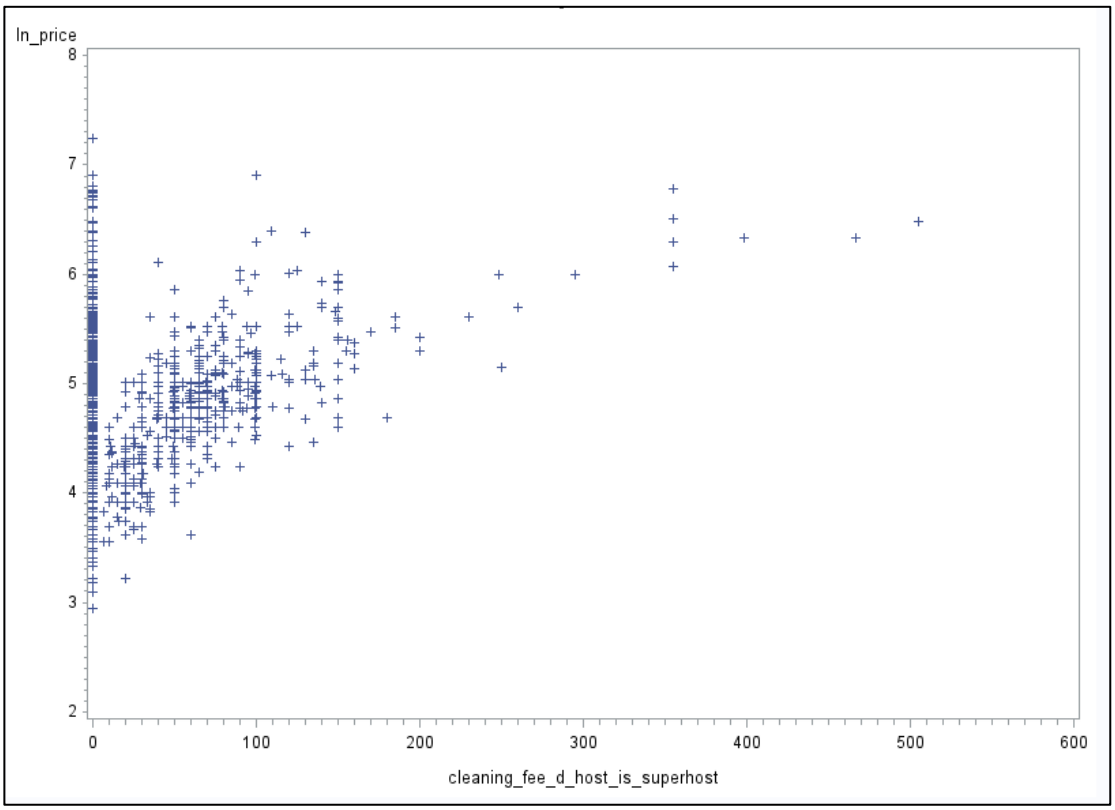
A.8 – Scatterplots (price versus cleaning_fee_d_host_is_superhost)



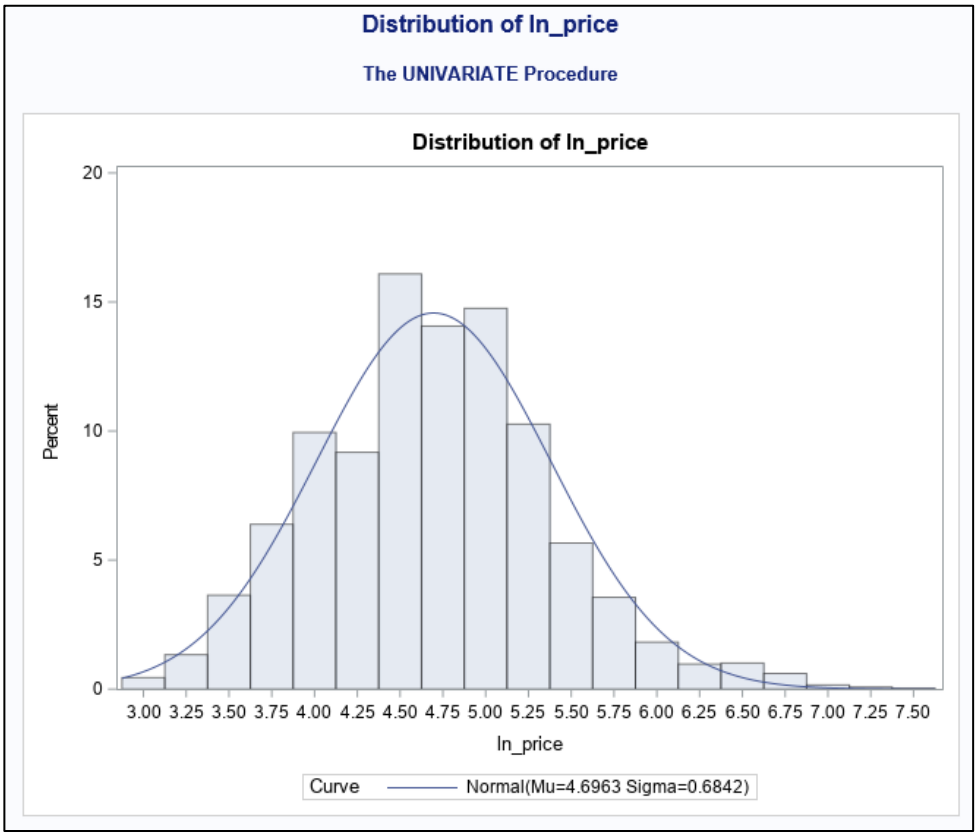
A.9 Scatterplot (ln_price versus cleaning_fee)



A.10 scatterplot (ln_price versus cleaning_fee_d_host_is_superhost)



A.11 Histogram



A.12 QQ plot

A.13 Full Model

The REG Procedure
Model: MODEL1
Dependent Variable: ln_price

Number of Observations Read	2478
Number of Observations Used	1332
Number of Observations with Missing Values	1146

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	281.43068	14.07153	98.53	<.0001
Error	1311	187.22647	0.14281		
Corrected Total	1331	468.65715			

Root MSE	0.37790	R-Square	0.6005
Dependent Mean	4.80035	Adj R-Sq	0.5944
Coeff Var	7.87244		

A.14 Final Model & VIF

The REG Procedure
Model: MODEL1
Dependent Variable: ln_price

Number of Observations Read	2478
Number of Observations Used	1874
Number of Observations with Missing Values	604

Analysis of Variance

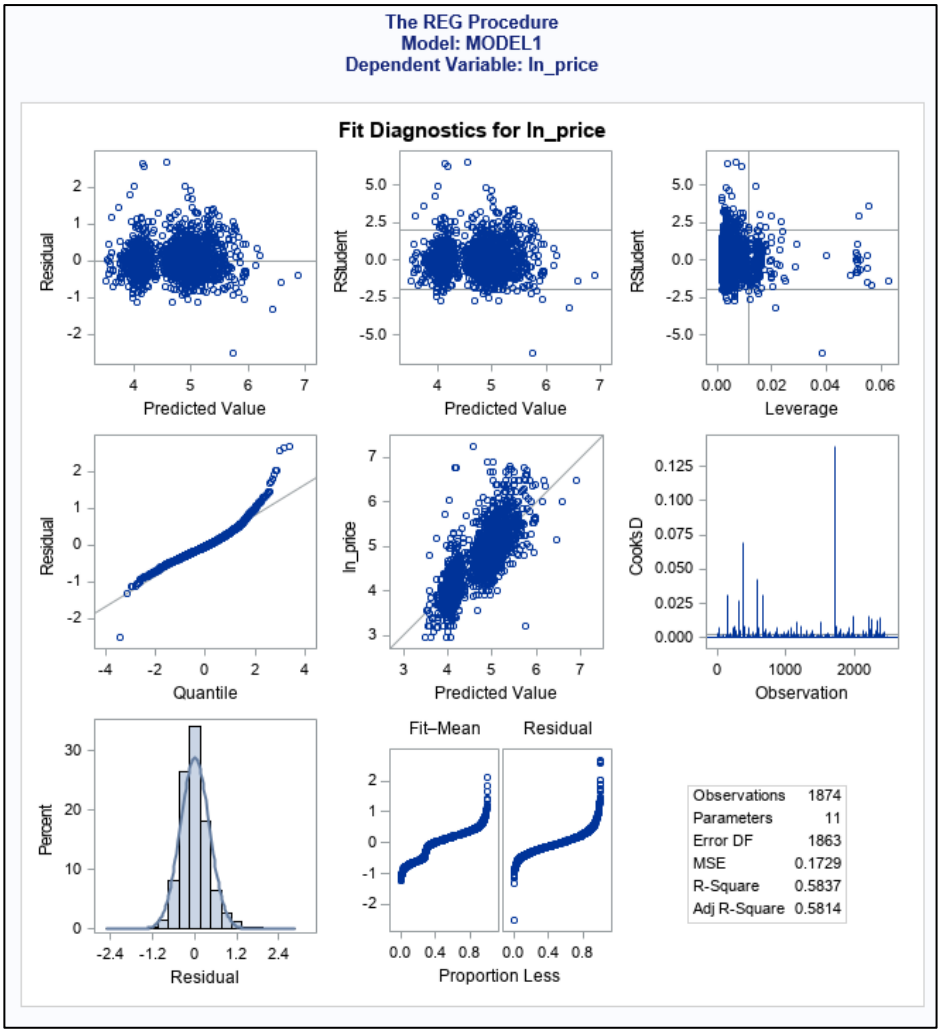
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	451.56720	45.15672	261.16	<.0001
Error	1863	322.12565	0.17291		
Corrected Total	1873	773.69285			

Root MSE	0.41582	R-Square	0.5837
Dependent Mean	4.75759	Adj R-Sq	0.5814
Coeff Var	8.74016		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	4.62402	0.04389	105.35	<.0001	.	0
accommodates	1	0.09557	0.00534	17.89	<.0001	0.65240	1.53279
cleaning_fee	1	0.00181	0.00021346	8.46	<.0001	0.67240	1.48721
d_within_an_hour	1	-0.16545	0.03743	-4.42	<.0001	0.27317	3.66071
d_within_a_few_hours	1	-0.09442	0.03857	-2.45	0.0145	0.28127	3.55535
d_host_is_superhost	1	0.06436	0.02185	2.95	0.0033	0.94622	1.05684
d_WM	1	-0.31864	0.04879	-6.53	<.0001	0.94861	1.05417
d_SEM	1	-0.07842	0.02422	-3.24	0.0012	0.92663	1.07917
d_NSM	1	-0.19151	0.03157	-6.07	<.0001	0.91158	1.09699
d_private_room	1	-0.57445	0.02589	-22.19	<.0001	0.68346	1.46313
d_shared_room	1	-1.07743	0.09530	-11.31	<.0001	0.96215	1.03933

A.15


The REG Procedure
Model: MODEL1
Dependent Variable: ln_price
Fit Diagnostics for ln_price
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
ln_price
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 1874
Parameters 11
Error DF 1863
MSE 0.1729
R-Square 0.5837
Adj R-Square 0.5814

A.16

The REG Procedure
Model: MODEL1
Dependent Variable: ln_price

Number of Observations Read	2433
Number of Observations Used	1829
Number of Observations with Missing Values	604

Analysis of Variance

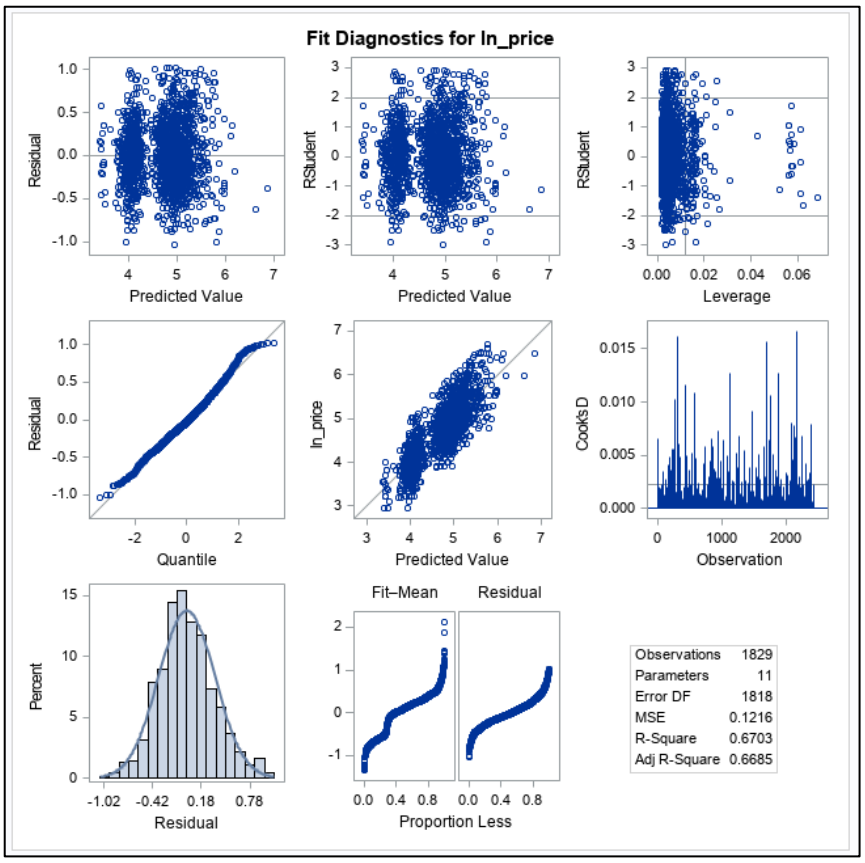
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	449.33034	44.93303	369.65	<.0001
Error	1818	220.98772	0.12156		
Corrected Total	1828	670.31807			

Root MSE	0.34865	R-Square	0.6703
Dependent Mean	4.72984	Adj R-Sq	0.6685
Coeff Var	7.37124		

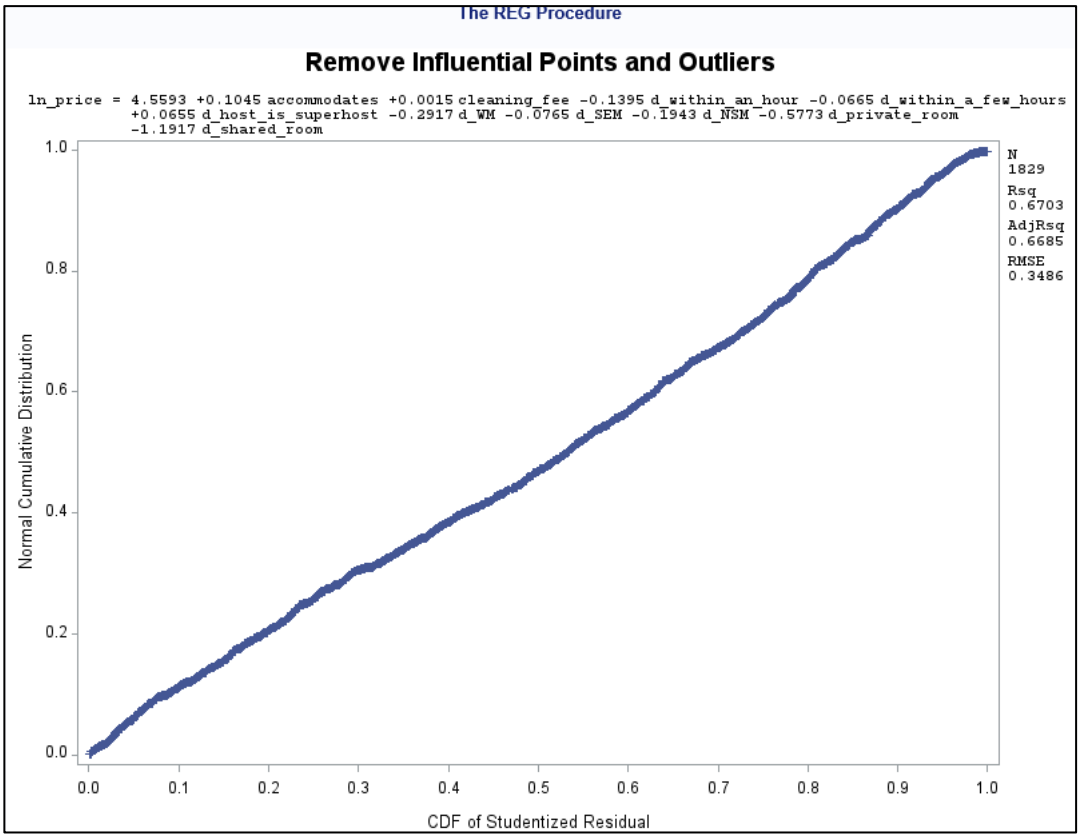
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.55926	0.03745	121.73	<.0001
accommodates	1	0.10452	0.00465	22.47	<.0001
cleaning_fee	1	0.00153	0.00018498	8.28	<.0001
d_within_an_hour	1	-0.13951	0.03187	-4.38	<.0001
d_within_a_few_hours	1	-0.06649	0.03286	-2.02	0.0432
d_host_is_superhost	1	0.06551	0.01849	3.54	0.0004
d_WM	1	-0.29170	0.04135	-7.05	<.0001
d_SEM	1	-0.07655	0.02055	-3.73	0.0002
d_NSM	1	-0.19426	0.02683	-7.24	<.0001
d_private_room	1	-0.57727	0.02199	-26.26	<.0001
d_shared_room	1	-1.19174	0.08401	-14.19	<.0001

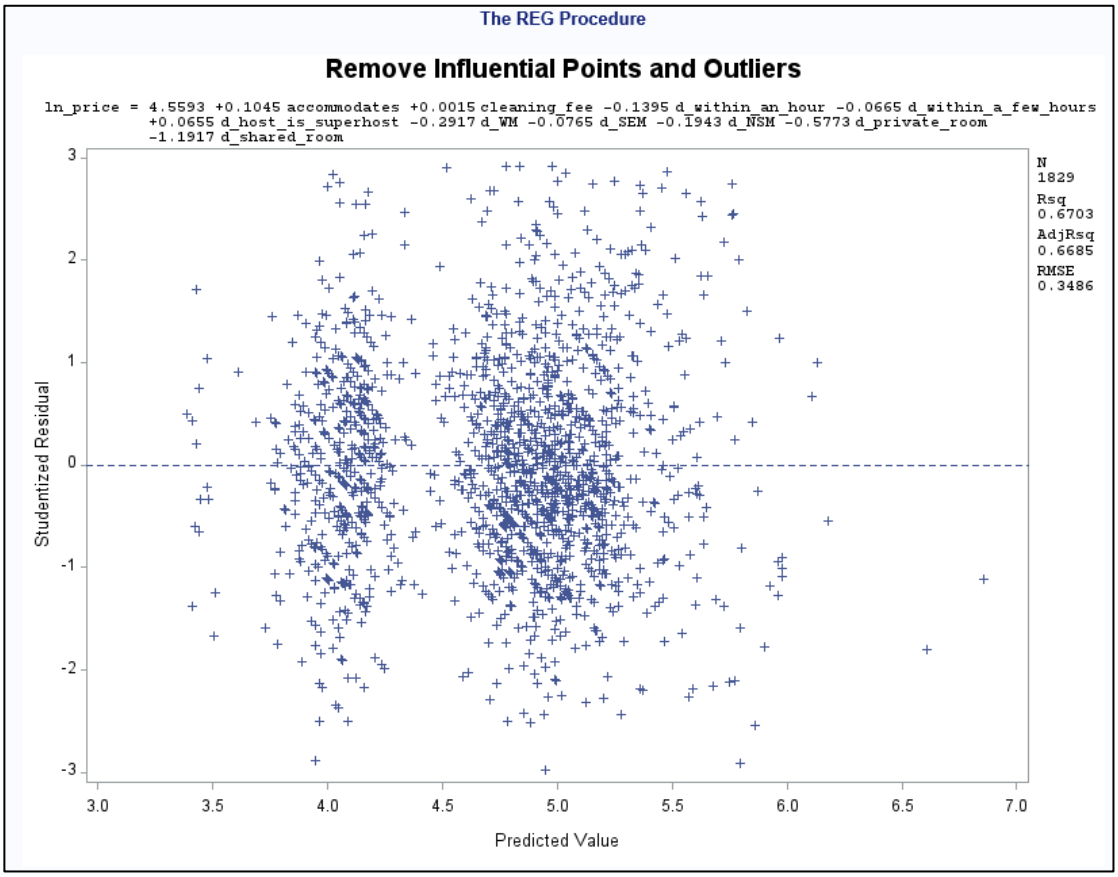
A.17



A.18 Normal Probability Plot



A.19 Studentized versus Predicted Values



A.20 Standardized Estimate

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	4.55926	0.03745	121.73	<.0001	0
accommodates	1	0.10452	0.00465	22.47	<.0001	0.37650
cleaning_fee	1	0.00153	0.00018498	8.28	<.0001	0.13643
d_within_an_hour	1	-0.13951	0.03187	-4.38	<.0001	-0.11306
d_within_a_few_hours	1	-0.06649	0.03286	-2.02	0.0432	-0.05152
d_host_is_superhost	1	0.06551	0.01849	3.54	0.0004	0.04898
d_WM	1	-0.29170	0.04135	-7.05	<.0001	-0.09736
d_SEM	1	-0.07655	0.02055	-3.73	0.0002	-0.05208
d_NSM	1	-0.19426	0.02683	-7.24	<.0001	-0.10208
d_private_room	1	-0.57727	0.02199	-26.26	<.0001	-0.42812
d_shared_room	1	-1.19174	0.08401	-14.19	<.0001	-0.19433

A.21 Pearson Correlation Coefficients

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	ln_price	accommodates	cleaning_fee	d_within_an_hour	d_within_a_few_hours	d_host_is_superhost	d_WM	d_SEM	d_NSM	d_private_room	d_shared_room
ln_price	1.00000 2429	0.63142 <.0001 2429	0.54084 <.0001 1829	0.13663 <.0001 2429	-0.12649 <.0001 2429	0.09639 <.0001 2429	-0.09993 <.0001 2429	-0.02027 0.3179 2429	-0.17714 <.0001 2429	-0.62991 <.0001 2429	-0.21500 <.0001 2429
accommodates	0.63142 <.0001 2429	1.00000 2433	0.51279 <.0001 1833	0.18809 <.0001 2433	-0.15865 <.0001 2433	0.04620 0.0227 2433	0.02670 0.1880 2433	-0.01802 0.3743 2433	-0.06292 0.0019 2433	-0.49188 <.0001 2433	-0.10921 <.0001 2433
cleaning_fee	0.54084 <.0001 1829	0.51279 <.0001 1833	1.00000 1833	0.10952 <.0001 1833	-0.12007 <.0001 1833	-0.00369 0.8745 1833	-0.03204 0.1704 1833	0.05606 0.0164 1833	-0.13129 <.0001 1833	-0.44285 <.0001 1833	-0.07318 0.0017 1833
d_within_an_hour	0.13663 <.0001 2429	0.18809 <.0001 2433	0.10952 <.0001 1833	1.00000 2433	-0.84409 <.0001 2433	0.26079 <.0001 2433	-0.00001 0.9995 2433	-0.08900 <.0001 2433	-0.12077 <.0001 2433	-0.23743 <.0001 2433	-0.02686 0.1854 2433
d_within_a_few_hours	-0.12649 <.0001 2429	-0.15865 <.0001 2433	-0.12007 <.0001 1833	-0.84409 <.0001 2433	1.00000 2433	-0.22047 <.0001 2433	0.00317 0.8758 2433	0.07929 <.0001 2433	0.10765 <.0001 2433	0.19445 <.0001 2433	0.00905 0.6555 2433
d_host_is_superhost	0.09639 <.0001 2429	0.04620 0.0227 2433	-0.00369 0.8745 1833	0.26079 <.0001 2433	-0.22047 <.0001 2433	1.00000 2433	-0.04062 0.0452 2433	-0.02554 0.2078 2433	-0.04405 0.0298 2433	-0.10691 <.0001 2433	-0.06458 0.0014 2433
d_WM	-0.09993 <.0001 2429	0.02670 0.1880 2433	-0.03204 0.1704 1833	-0.00001 0.9995 2433	0.00317 0.8758 2433	-0.04062 0.0452 2433	1.00000 2433	-0.13082 <.0001 2433	-0.08706 <.0001 2433	0.10064 <.0001 2433	0.02788 0.1691 2433
d_SEM	-0.02027 0.3179 2429	-0.01802 0.3743 2433	0.05606 0.0164 1833	-0.08900 <.0001 2433	0.07929 <.0001 2433	-0.02554 0.2078 2433	-0.13082 <.0001 2433	1.00000 2433	-0.20175 <.0001 2433	0.03680 0.0695 2433	-0.03929 0.0526 2433
d_NSM	-0.17714 <.0001 2429	-0.06292 0.0019 2433	-0.13129 <.0001 1833	-0.12077 <.0001 2433	0.10765 <.0001 2433	-0.04405 0.0298 2433	-0.08706 <.0001 2433	-0.20175 <.0001 2433	1.00000 2433	0.12146 <.0001 2433	-0.02651 0.1912 2433
d_private_room	-0.62991 <.0001 2429	-0.49188 <.0001 2433	-0.44285 <.0001 1833	-0.23743 <.0001 2433	0.19445 <.0001 2433	-0.10691 <.0001 2433	0.10064 <.0001 2433	0.03680 0.0695 2433	0.12146 <.0001 2433	1.00000 2433	-0.09540 <.0001 2433
d_shared_room	-0.21500 <.0001 2429	-0.10921 <.0001 2433	-0.07318 0.0017 1833	-0.02686 0.1854 2433	0.00905 0.6555 2433	-0.06458 0.0014 2433	0.02788 0.1691 2433	-0.03929 0.0526 2433	-0.02651 0.1912 2433	-0.09540 <.0001 2433	1.00000 2433

A.22 Cross Validation

5-fold crossvalidation + 25% Testing Set

The GLMSELECT Procedure

Data Set	WORK.AIRBNB_IMPORT_NEW16
Dependent Variable	ln_price
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	20446001

Number of Observations Read	2433
Number of Observations Used	1829
Number of Observations Used for Training	1371
Number of Observations Used for Testing	458

Dimensions	
Number of Effects	11
Number of Parameters	11

5-fold crossvalidation + 25% Testing Set

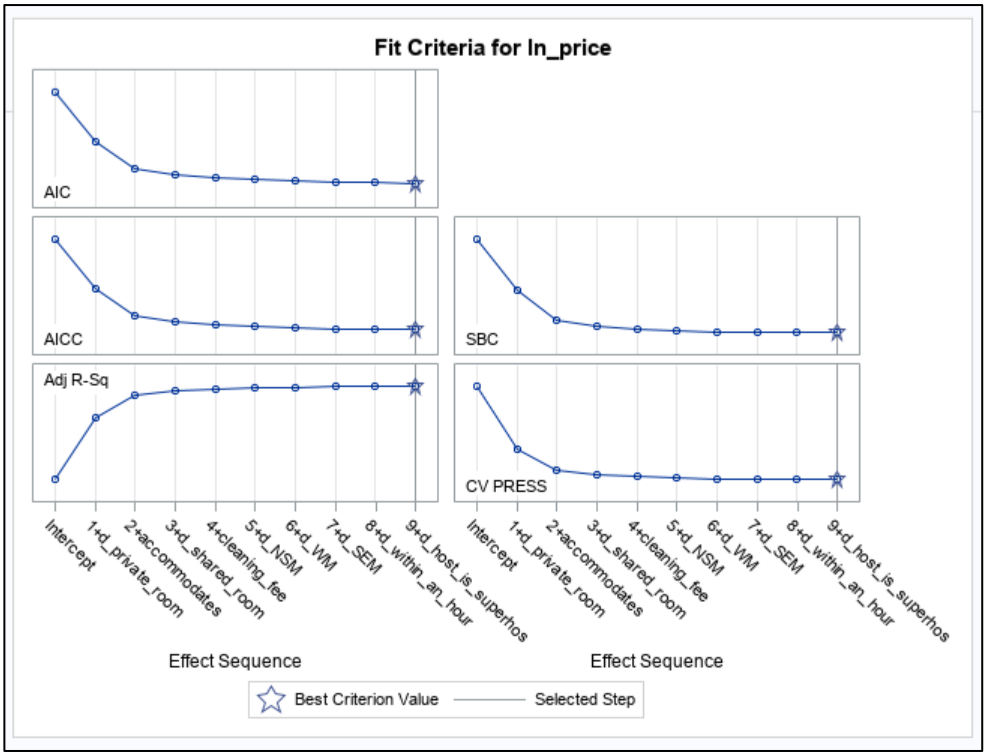
The GLMSELECT Procedure

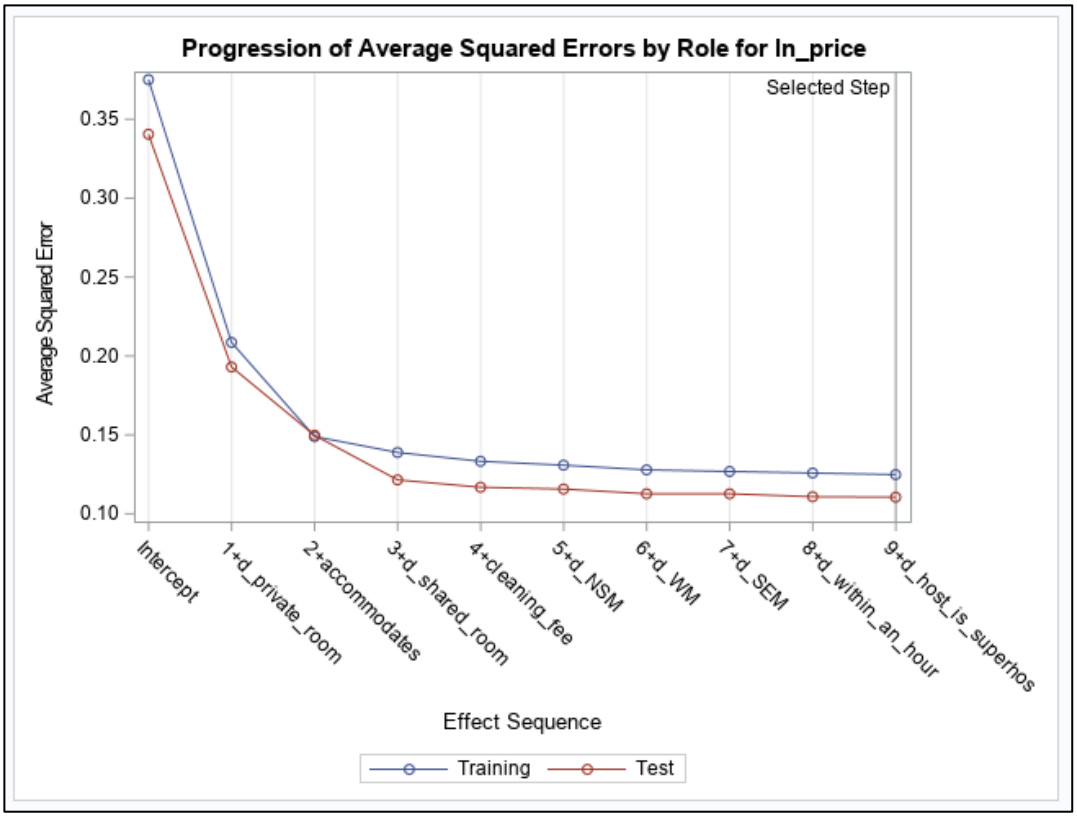
Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0	Intercept		1	-1336.8367	0.3752	0.3406	516.5884
1	d_private_room		2	-2133.9926	0.2087	0.1931	287.0427
2	accommodates		3	-2588.1686	0.1490	0.1498	206.2044
3	d_shared_room		4	-2677.2848	0.1389	0.1215	192.1855
4	cleaning_fee		5	-2726.3611	0.1333	0.1169	184.9125
5	d_NSM		6	-2744.7071	0.1309	0.1157	182.2233
6	d_WM		7	-2768.8995	0.1279	0.1128	178.1158
7	d_SEM		8	-2772.6758	0.1269	0.1127	177.1578
8	d_within_an_hour		9	-2776.1008	0.1259	0.1109	176.0205
9	d_host_is_superhost		10	-2779.8126*	0.1249	0.1106	174.6503*

* Optimal Value of Criterion

Selection stopped as adding or dropping any effect does not improve the selection criterion.





5-fold crossvalidation + 25% Testing Set

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 9).

Effects:	Intercept accommodates cleaning_fee d_within_an_hour d_host_is_superhost d_WM d_SEM d_NSM d_private_room d_shared_room

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	9	343.13773	38.12641	303.04
Error	1361	171.23366	0.12581	
Corrected Total	1370	514.37139		

Root MSE	0.35470
Dependent Mean	4.72530
R-Square	0.6671
Adj R-Sq	0.6649
AIC	-1459.04554
AICC	-1458.85128
SBC	-2779.81258
ASE (Train)	0.12490
ASE (Test)	0.11061
CV PRESS	174.65035

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1096	275	36.4812
2	1097	274	33.7798
3	1097	274	34.0360
4	1097	274	34.7124
5	1097	274	35.6410
Total			174.6503

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	4.481637	0.030119	148.80	4.48862	4.45811	4.48036	4.4797	4.50078
accommodates	1	0.109778	0.005396	20.34	0.10473	0.11904	0.10521	0.1119	0.10819
cleaning_fee	1	0.001555	0.000213	7.32	0.00158	0.00146	0.00178	0.0014	0.00155
d_within_an_hour	1	-0.079988	0.020507	-3.90	-0.07938	-0.08579	-0.07375	-0.0718	-0.09020
d_host_is_superhost	1	0.071503	0.021659	3.30	0.07345	0.06513	0.06946	0.0729	0.07633
d_WM	1	-0.286848	0.047272	-6.07	-0.25262	-0.30938	-0.29034	-0.2814	-0.29808
d_SEM	1	-0.087336	0.024479	-3.57	-0.06273	-0.08503	-0.09561	-0.1119	-0.08083
d_NSM	1	-0.203605	0.030987	-6.57	-0.17609	-0.24305	-0.20101	-0.1973	-0.20151
d_private_room	1	-0.555871	0.025445	-21.85	-0.58051	-0.53708	-0.54800	-0.5601	-0.55226
d_shared_room	1	-1.113228	0.109228	-10.19	-1.18208	-1.09575	-1.06222	-1.0643	-1.15209

A.23 Compute Predictions

The REG Procedure
Model: MODEL1
Dependent Variable: ln_price

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	4.1386	0.0520	4.0366	4.2406	3.3167	4.9605	.
2	.	4.0148	0.1015	3.8158	4.2138	3.1753	4.8542	.
3	5.70	5.3493	0.0239	5.3025	5.3962	4.5325	6.1662	0.3478

Appendix B – Theresa

B.1 - Descriptive

The MEANS Procedure

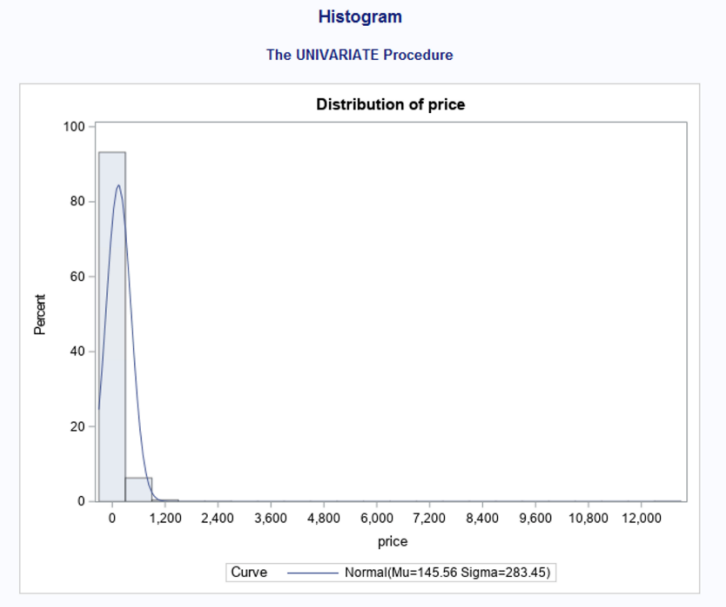
Analysis Variable : price									
Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum
145.5640612	283.4509102	5.6895375	134.4073298	156.7207927	0	71.0000000	109.0000000	165.0000000	12624.00

The UNIVARIATE Procedure
Variable: price

Moments			
N	2482	Sum Weights	2482
Mean	145.564061	Sum Observations	361290
Std Deviation	283.45091	Variance	80344.4185
Skewness	34.9849745	Kurtosis	1517.48796
Uncorrected SS	251925342	Corrected SS	199334502
Coeff Variation	194.725888	Std Error Mean	5.68953754

Basic Statistical Measures			
Location		Variability	
Mean	145.5641	Std Deviation	283.45091
Median	109.0000	Variance	80344
Mode	100.0000	Range	12624
		Interquartile Range	94.00000

B.2 - Histogram



B.3 – Frequency Table for Qualitative Variables

frenquency table

The FREQ Procedure

host_response_time	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N/A	790	31.84	790	31.84
a few days or more	46	1.85	836	33.70
within a day	169	6.81	1005	40.51
within a few hours	236	9.51	1241	50.02
within an hour	1240	49.98	2481	100.00
Frequency Missing = 1				

host_is_superhost	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	1873	75.49	1873	75.49
t	608	24.51	2481	100.00
Frequency Missing = 1				

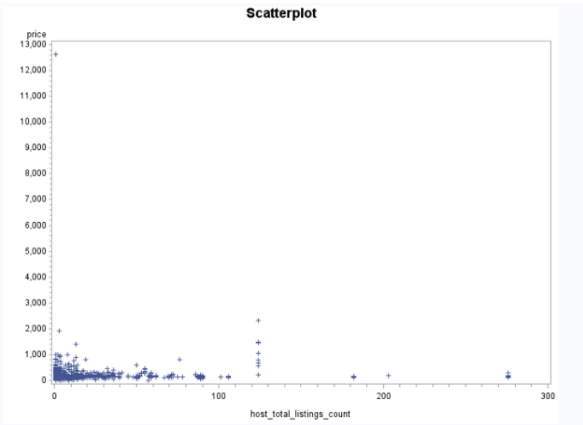
room_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Entire home/apt	1537	61.93	1537	61.93
Private room	896	36.10	2433	98.03
Shared room	49	1.97	2482	100.00

bed_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Airbed	2	0.08	2	0.08
Futon	4	0.16	6	0.24
Pull-out	13	0.52	19	0.77
Real Bed	2463	99.23	2482	100.00

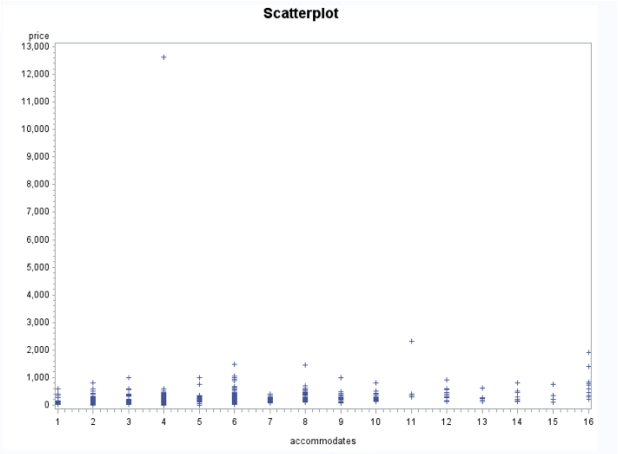
cancellation_policy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Strict	1038	41.82	1038	41.82
flexible	844	34.00	1882	75.83
moderate	600	24.17	2482	100.00

Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
EM	246	9.91	246	9.91
IM	1293	52.10	1539	62.01
NSM	272	10.96	1811	72.97
SEM	515	20.75	2326	93.71
WM	156	6.29	2482	100.00

B.4 – Scatterplot for host_total_listing_count.

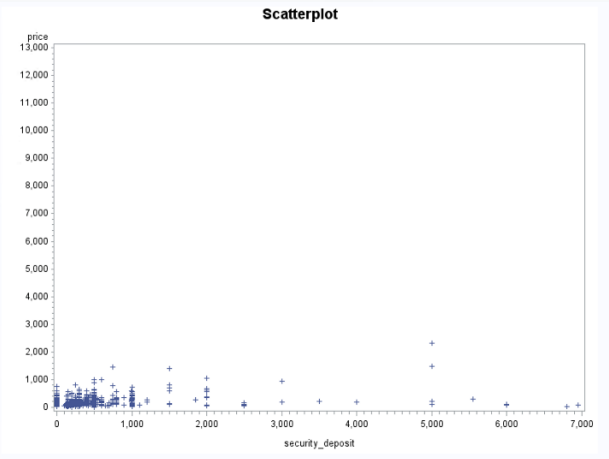


B.5 – Scatterplot for accomodates

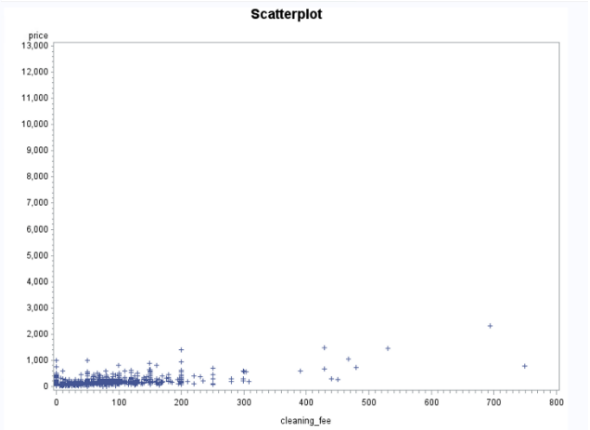


B.6 – Scatterplot for security_deposit

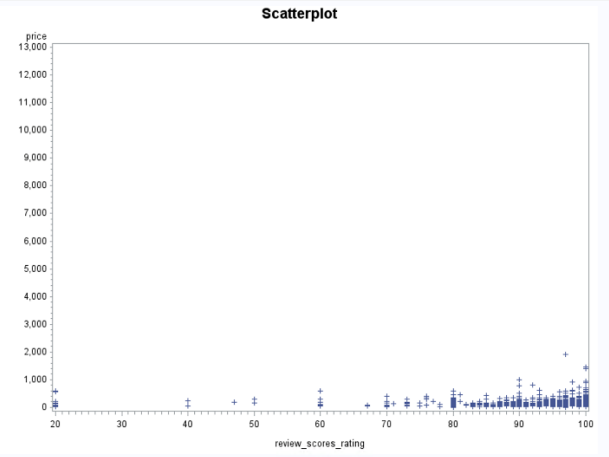
B.7 – Scatterplot for cleaning_fee



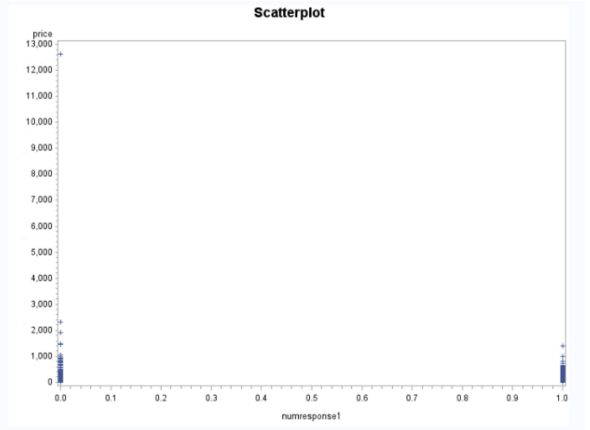
B.8 – Scatterplot for review_scores_rating.



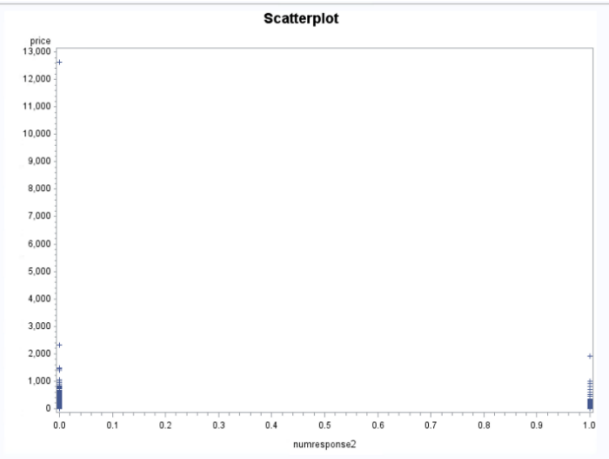
B.9 – Scatterplot for numresponse1

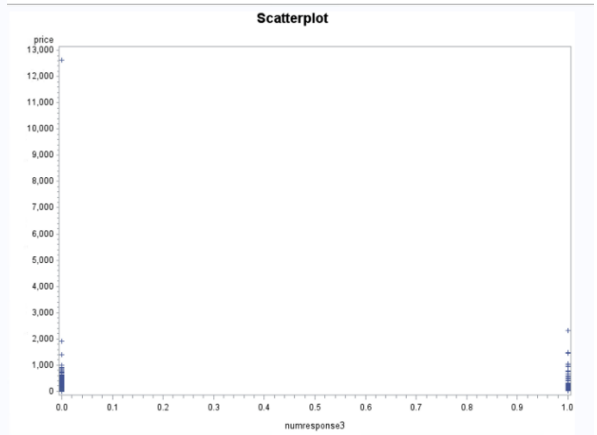


B.9.1 – Scatterplot for numresponse2

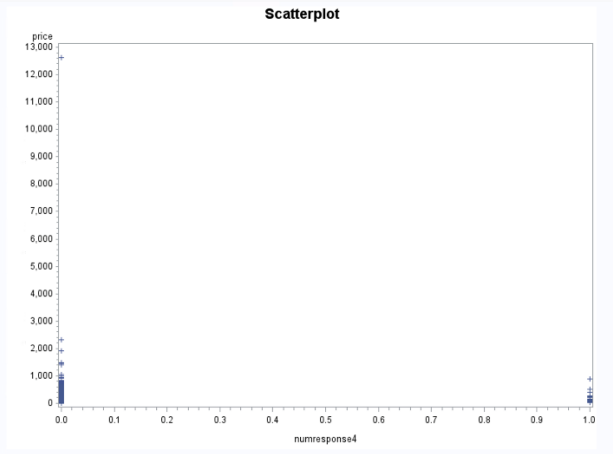


B.10 – Scatterplot for numresponse3

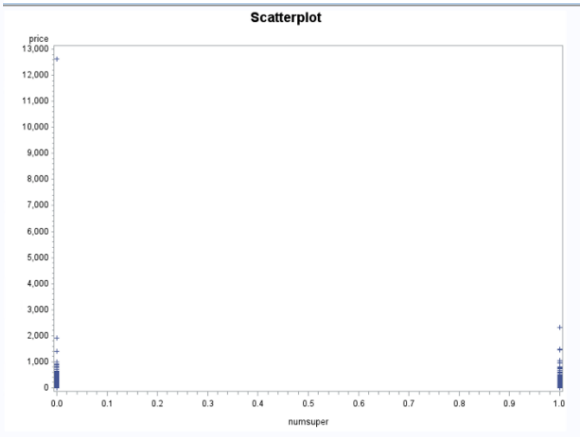




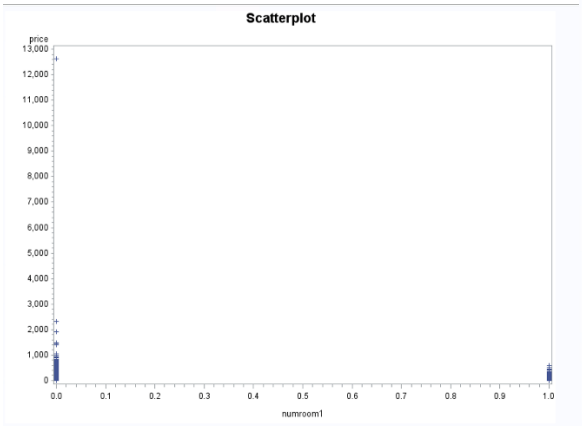
B.11 – Scatterplot for numresponse4



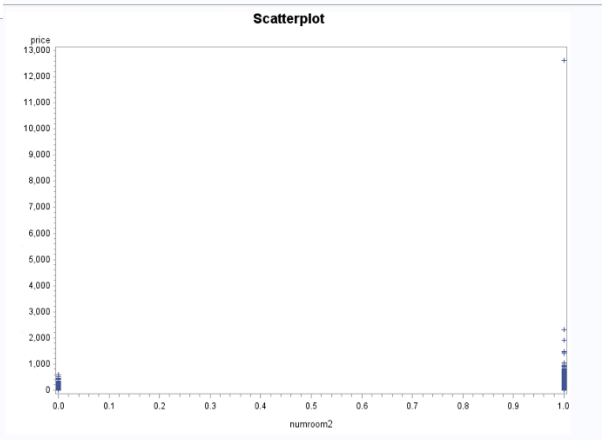
B.12 – Scatterplot for numsuper



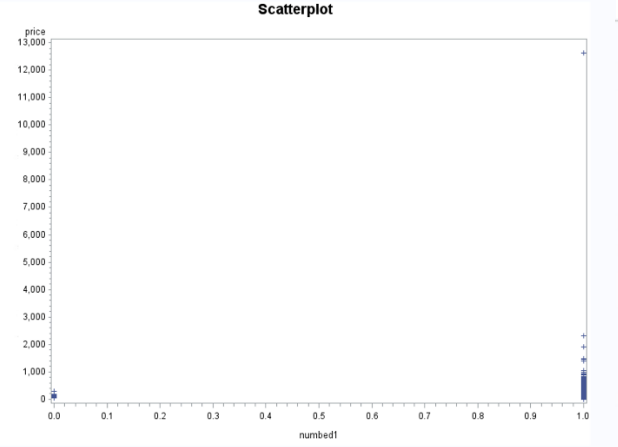
B.13 – Scatterplot for numroom1



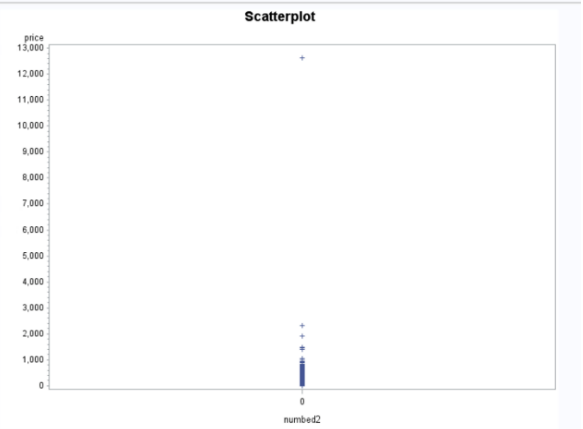
B.14 – Scatterplot for numroom2



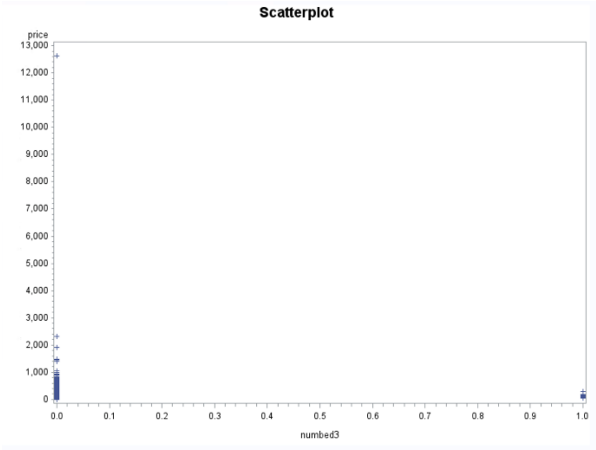
B.15 – Scatterplot for numbed1



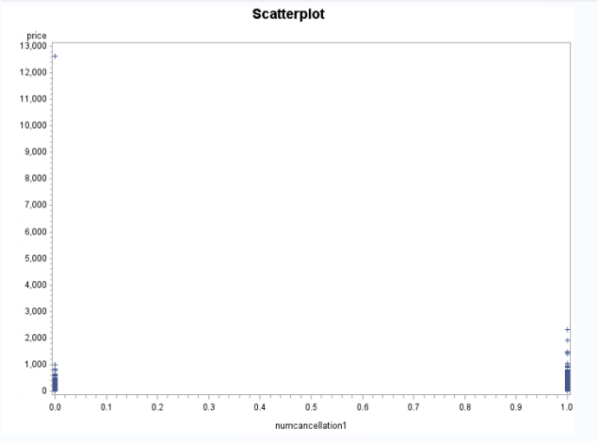
B.16 – Scatterplot for numbed2



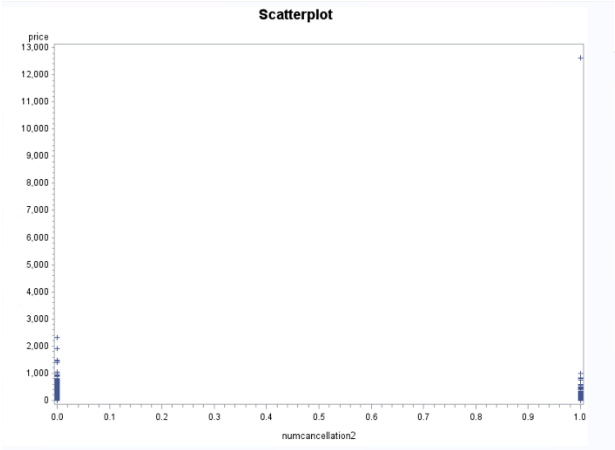
B.17 – Scatterplot for numbed3



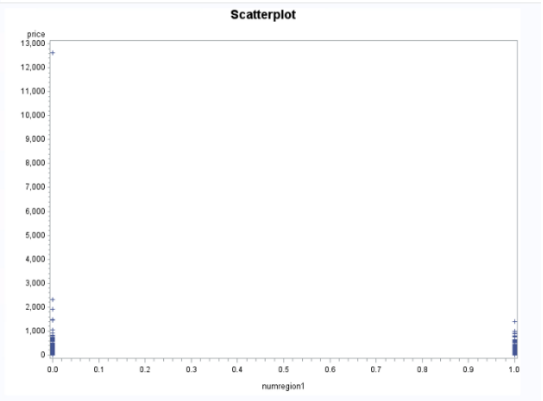
B.18 – Scatterplot for numcancellation1



B.19 – Scatterplot for numcancellation2

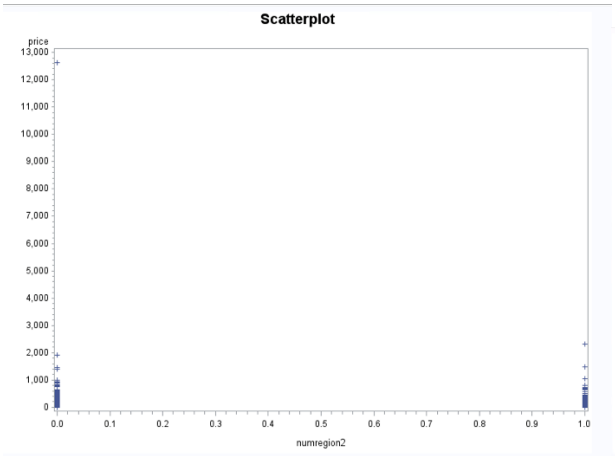


B.20 – Scatterplot for numregion1

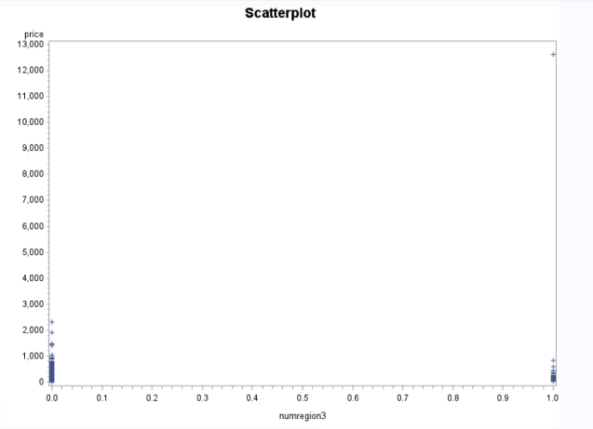


B.21 – Scatterplot for numregion2

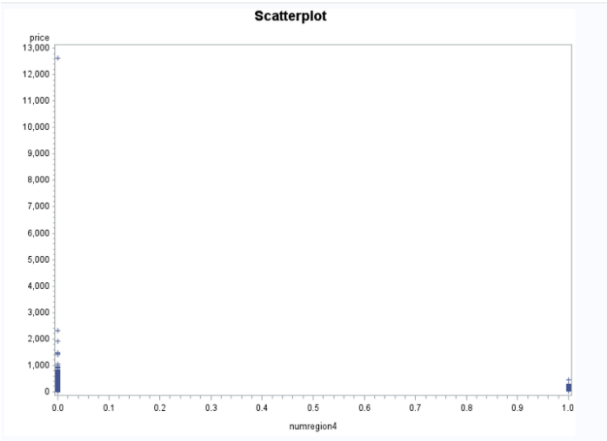
B.22 - Scatterplot for numregion3

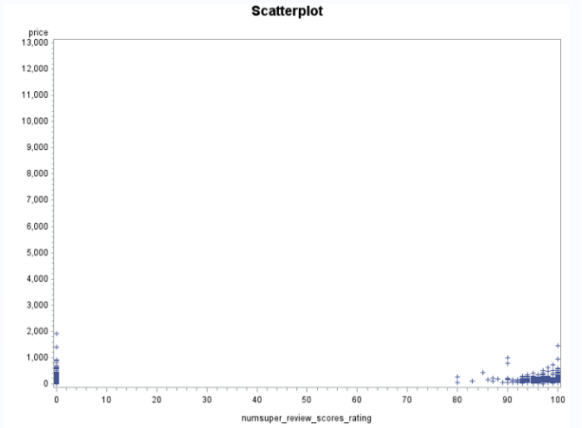


B.23 – Scatterplot for numregion4



B.24 – Scatterplot for interaction term





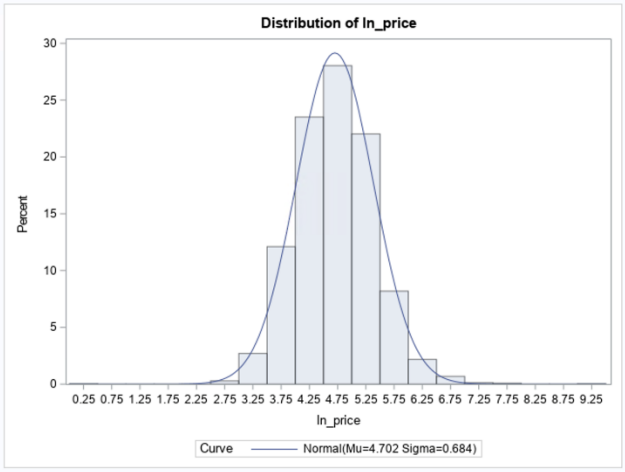
After log transformation

B.25 – Histogram for ln_price

The UNIVARIATE Procedure
Variable: ln_price

Moments			
N	2479	Sum Weights	2479
Mean	4.70197056	Sum Observations	11656.185
Std Deviation	0.68401323	Variance	0.4678741
Skewness	0.26760837	Kurtosis	2.07758937
Uncorrected SS	55966.4308	Corrected SS	1159.39201
Coeff Variation	14.5473737	Std Error Mean	0.01373809

Basic Statistical Measures			
Location		Variability	
Mean	4.701971	Std Deviation	0.68401
Median	4.691348	Variance	0.46787
Mode	4.605170	Range	9.44336
		Interquartile Range	0.84931



B.25.1 – Pearson Correlation

	ln_price
ln_price	1.00000 2479
host_total_listings_count	0.14528 <.0001 2478
accommodates	0.61842 <.0001 2479
security_deposit	0.19361 <.0001 1654
cleaning_fee	0.52546 <.0001 1865
review_scores_rating	0.03877 0.0952 1854
numresponse1	0.11645 <.0001 2479
numresponse2	-0.02684 0.1816 2479
numresponse3	0.02114 0.2926 2479
numresponse4	-0.00973 0.6283 2479
numsuper	0.11958 <.0001 2479
numroom1	-0.59407 <.0001 2479
numroom2	0.64924 <.0001 2479
numbed1	0.03989 0.0470 2479
numbed2	. . 2479
numbed3	-0.02207 0.2721 2479
numcancellation1	0.23340 <.0001 2479
numcancellation2	-0.22099 <.0001 2479
numregion1	0.16537 <.0001 2479
numregion2	-0.04127 0.0399 2479
numregion3	-0.13011 <.0001 2479
numregion4	-0.11415 <.0001 2479
numsuper_review_scores_rating	0.09666 <.0001 1854

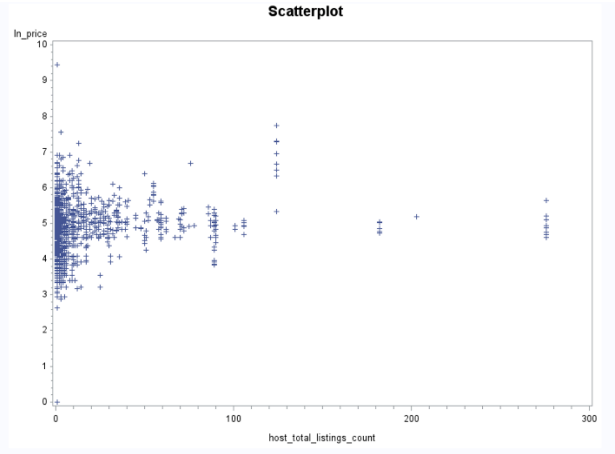
B.26 - Descriptives

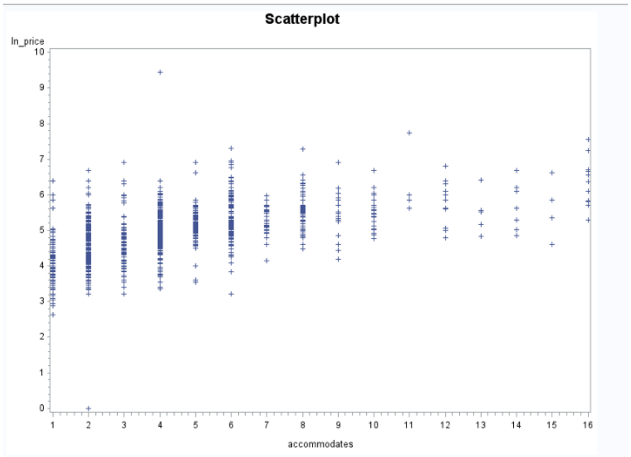
Descriptives for ln_price

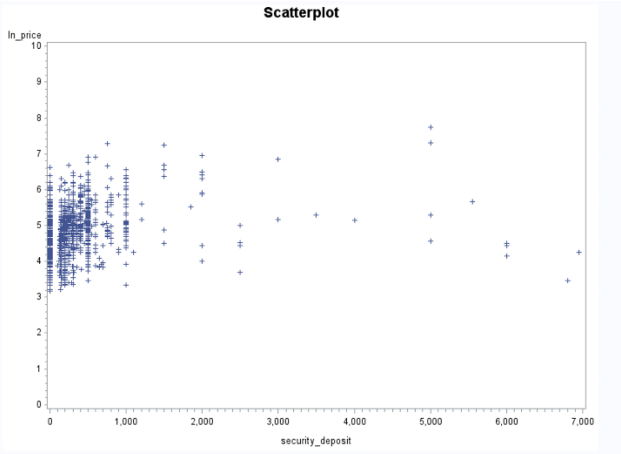
The MEANS Procedure

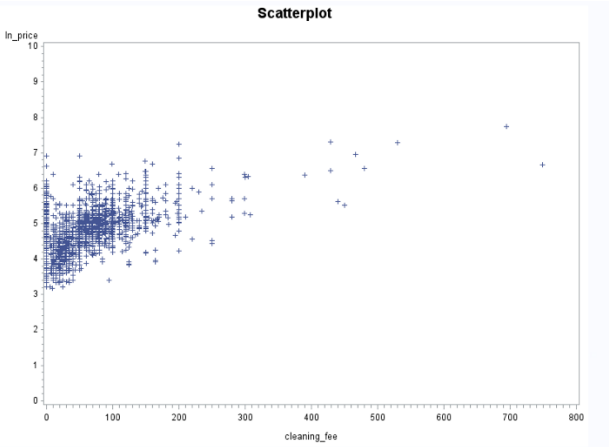
Analysis Variable : ln_price									
Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean	Minimum	25th Pctl	50th Pctl	75th Pctl	Maximum
4.7019706	0.6840132	0.0137381	4.6750312	4.7289099	0	4.2626799	4.6913479	5.1119878	9.4433550

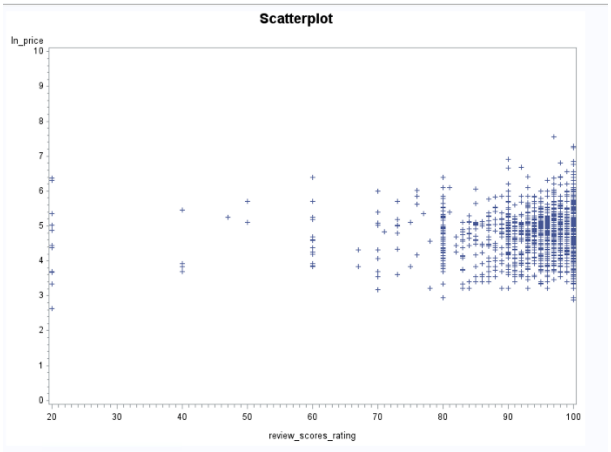
B.27 – Scatterplots after transformation

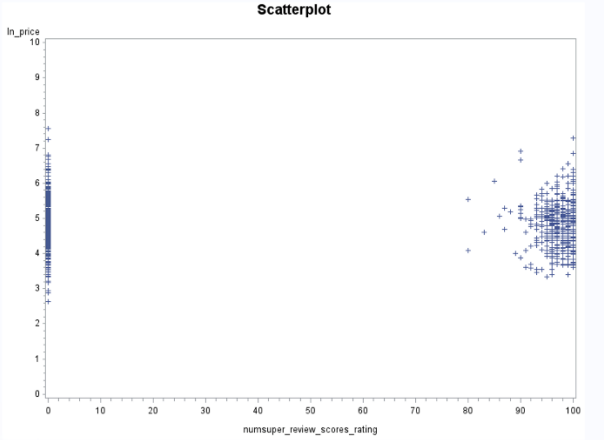



Scatterplot
ln_price
accommodates


Scatterplot
ln_price
security_deposit


Scatterplot
ln_price
cleaning_fee





B.28 – Regression Model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	21	310.85896	14.80281	108.51	<.0001
Error	1295	176.66522	0.13642		
Corrected Total	1316	487.52418			

Root MSE	0.36935	R-Square	0.6376
Dependent Mean	4.80200	Adj R-Sq	0.6318
Coeff Var	7.69163		

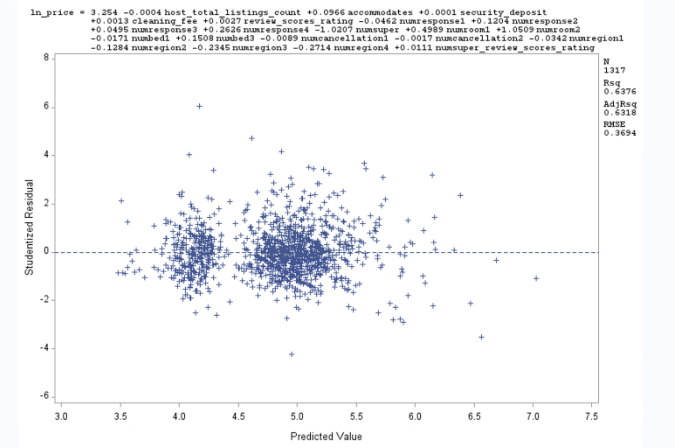
numbed2 =	0

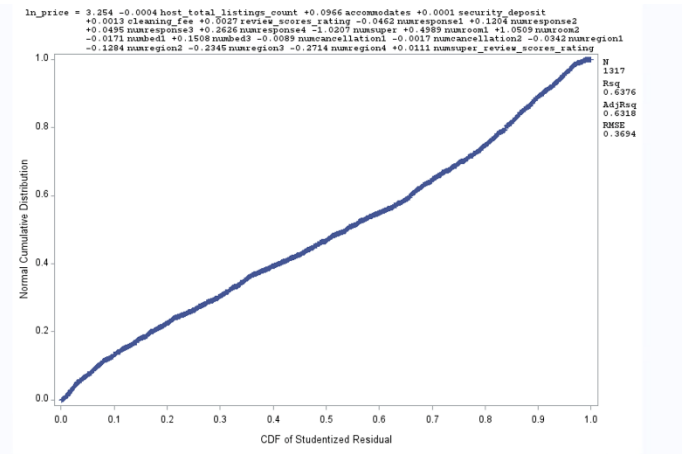
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.25398	0.25979	12.53	<.0001
host_total_listings_count	1	-0.00039000	0.00034150	-1.14	0.2537
accommodates	1	0.09661	0.00513	18.81	<.0001
security_deposit	1	0.00007884	0.00001965	4.01	<.0001
cleaning_fee	1	0.00127	0.00023601	5.38	<.0001
review_scores_rating	1	0.00270	0.00118	2.28	0.0227
numresponse1	1	-0.04617	0.02838	-1.63	0.1040
numresponse2	1	0.12038	0.04993	2.41	0.0160
numresponse3	1	0.04950	0.04017	1.23	0.2181
numresponse4	1	0.26265	0.13358	1.97	0.0495
numsuper	1	-1.02072	0.62649	-1.63	0.1035
numroom1	1	0.49890	0.09452	5.28	<.0001
numroom2	1	1.05090	0.09373	11.21	<.0001
numbed1	1	-0.01709	0.21495	-0.08	0.9366
numbed2	0	0	.	.	.
numbed3	1	0.15078	0.25802	0.58	0.5591
numcancellation1	1	-0.00889	0.02484	-0.36	0.7203
numcancellation2	1	-0.00175	0.03176	-0.05	0.9562
numregion1	1	-0.03417	0.03774	-0.91	0.3654
numregion2	1	-0.12836	0.04293	-2.99	0.0028
numregion3	1	-0.23448	0.04895	-4.79	<.0001
numregion4	1	-0.27136	0.05857	-4.63	<.0001
numsuper_review_scores_rating	1	0.01113	0.00647	1.72	0.0854

B.29 – Taking out numbed2 as it does not have enough observations

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.25398	0.25979	12.53	<.0001
host_total_listings_count	1	-0.00039000	0.00034150	-1.14	0.2537
accommodates	1	0.09661	0.00513	18.81	<.0001
security_deposit	1	0.00007884	0.00001965	4.01	<.0001
cleaning_fee	1	0.00127	0.00023601	5.38	<.0001
review_scores_rating	1	0.00270	0.00118	2.28	0.0227
numresponse1	1	-0.04617	0.02838	-1.63	0.1040
numresponse2	1	0.12038	0.04993	2.41	0.0160
numresponse3	1	0.04950	0.04017	1.23	0.2181
numresponse4	1	0.26265	0.13358	1.97	0.0495
numsuper	1	-1.02072	0.62649	-1.63	0.1035
numroom1	1	0.49890	0.09452	5.28	<.0001
numroom2	1	1.05090	0.09373	11.21	<.0001
numbed1	1	-0.01709	0.21495	-0.08	0.9366
numbed3	1	0.15078	0.25802	0.58	0.5591
numcancellation1	1	-0.00889	0.02484	-0.36	0.7203
numcancellation2	1	-0.00175	0.03176	-0.05	0.9562
numregion1	1	-0.03417	0.03774	-0.91	0.3654
numregion2	1	-0.12836	0.04293	-2.99	0.0028
numregion3	1	-0.23448	0.04895	-4.79	<.0001
numregion4	1	-0.27136	0.05857	-4.63	<.0001
numsuper_review_scores_rating	1	0.01113	0.00647	1.72	0.0854

B.29.1 – Studentized Residual





B.30 - model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	308.70470	28.06406	204.81	<.0001
Error	1305	178.81948	0.13703		
Corrected Total	1316	487.52418			

Root MSE	0.37017	R-Square	0.6332
Dependent Mean	4.80200	Adj R-Sq	0.6301
Coeff Var	7.70868		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.24808	0.14292	22.73	<.0001
accommodates	1	0.09803	0.00503	19.49	<.0001
security_deposit	1	0.00008109	0.00001959	4.14	<.0001
cleaning_fee	1	0.00120	0.00022662	5.28	<.0001
review_scores_rating	1	0.00285	0.00116	2.46	0.0140
numresponse1	1	-0.09156	0.02232	-4.10	<.0001
numroom1	1	0.47665	0.09292	5.13	<.0001
numroom2	1	1.02298	0.09178	11.15	<.0001
numregion2	1	-0.09198	0.02796	-3.29	0.0010
numregion3	1	-0.20255	0.03713	-5.45	<.0001
numregion4	1	-0.23800	0.04929	-4.83	<.0001
numsuper_review_scores_rating	1	0.00064631	0.00023367	2.77	0.0058

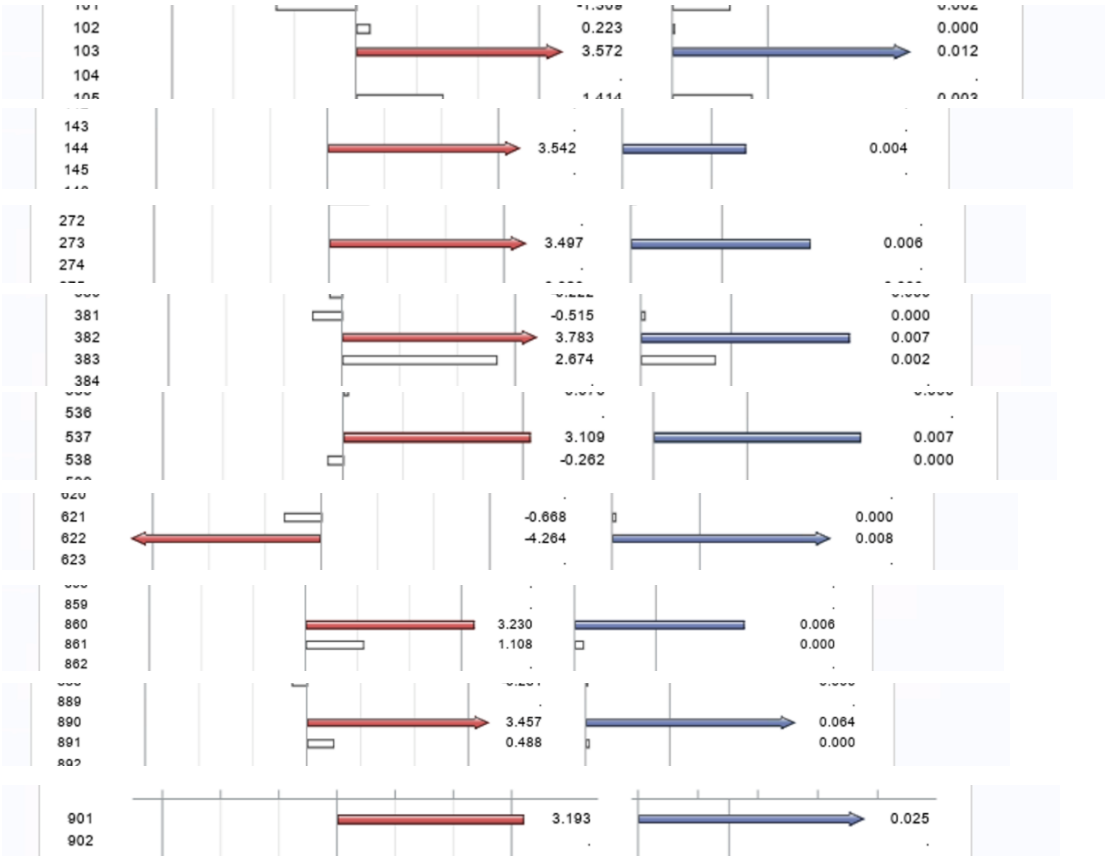
B.31 - Multicollinearity

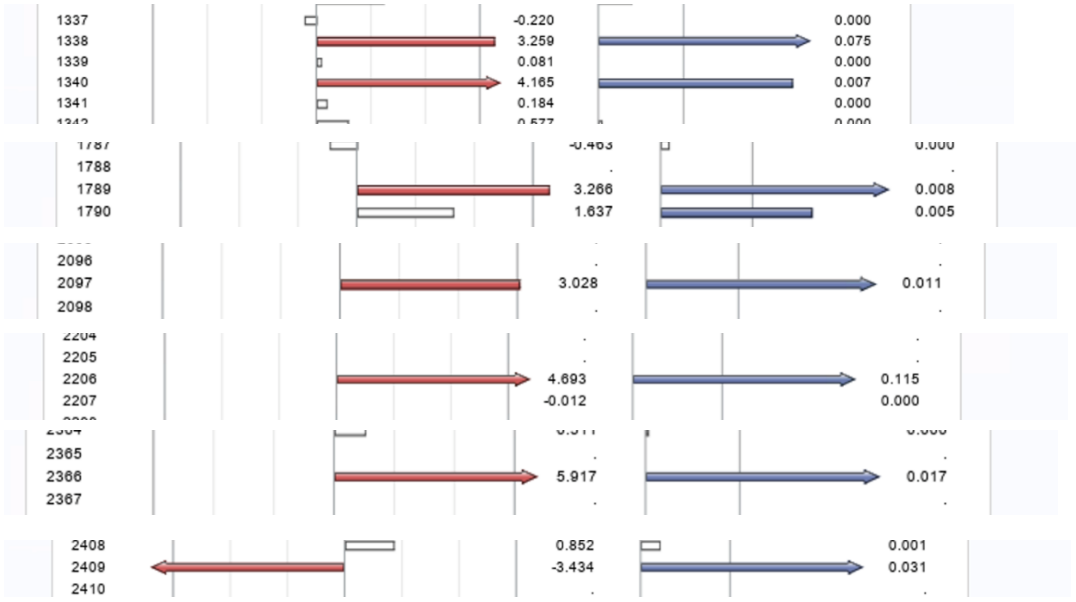
Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	3.24808	0.14292	22.73	<.0001	.	0
accommodates	1	0.09803	0.00503	19.49	<.0001	0.64269	1.55595
security_deposit	1	0.00008109	0.00001959	4.14	<.0001	0.93239	1.07251
cleaning_fee	1	0.00120	0.00022662	5.28	<.0001	0.60410	1.65536
review_scores_rating	1	0.00285	0.00116	2.46	0.0140	0.93477	1.06978
numresponse1	1	-0.09156	0.02232	-4.10	<.0001	0.91957	1.08746
numroom1	1	0.47665	0.09292	5.13	<.0001	0.06430	15.55256
numroom2	1	1.02298	0.09178	11.15	<.0001	0.06377	15.68193
numregion2	1	-0.09198	0.02796	-3.29	0.0010	0.94981	1.05285
numregion3	1	-0.20255	0.03713	-5.45	<.0001	0.94678	1.05621
numregion4	1	-0.23800	0.04929	-4.83	<.0001	0.95448	1.04769
numsuper_review_scores_rating	1	0.00064631	0.00023367	2.77	0.0058	0.90841	1.10083

B.32 – Resolving Multicollinearity Issue

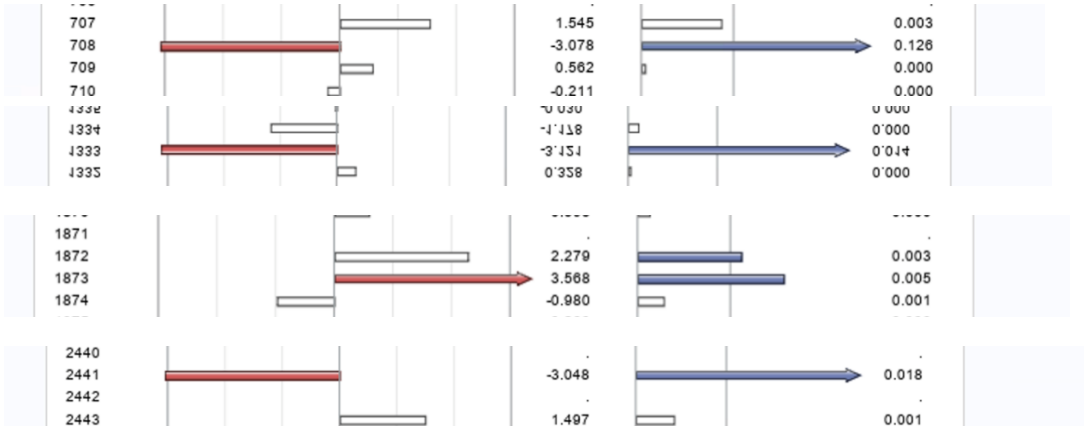
Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	3.71096	0.11190	33.16	<.0001	.	0
accommodates	1	0.09769	0.00508	19.24	<.0001	0.64280	1.55568
security_deposit	1	0.00007406	0.00001973	3.75	0.0002	0.93698	1.06726
cleaning_fee	1	0.00120	0.00022880	5.27	<.0001	0.60414	1.65526
review_scores_rating	1	0.00277	0.00117	2.37	0.0181	0.93495	1.06958
numresponse1	1	-0.09995	0.02248	-4.45	<.0001	0.92453	1.08163
numroom2	1	0.57368	0.02768	20.72	<.0001	0.71460	1.39938
numregion2	1	-0.08771	0.02821	-3.11	0.0019	0.95065	1.05191
numregion3	1	-0.19068	0.03742	-5.10	<.0001	0.95048	1.05210
numregion4	1	-0.22495	0.04970	-4.53	<.0001	0.95703	1.04490
numsuper_review_scores_rating	1	0.00068063	0.00023583	2.89	0.0040	0.90915	1.09992

B.33 – Influential Points and Outliers





Second Round of Removing Influential Points and Outliers



B.34 – Final Model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	296.06347	29.60635	261.77	<.0001
Error	1286	145.44918	0.11310		
Corrected Total	1296	441.51265			

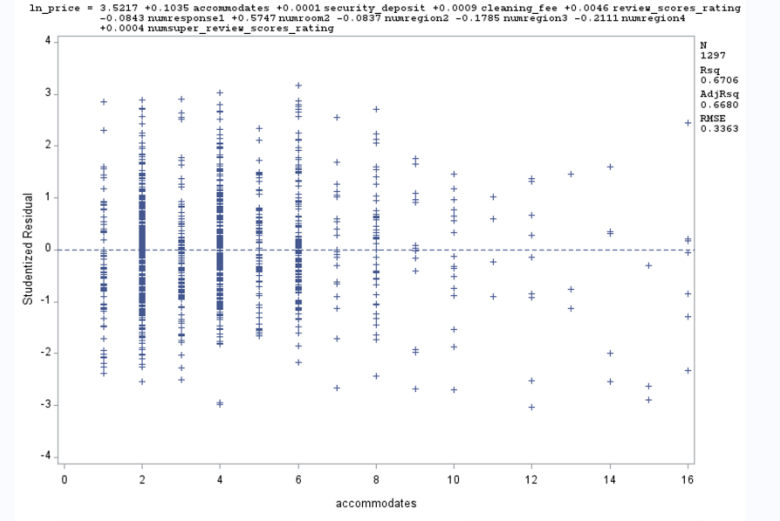
Root MSE	0.33631	R-Square	0.6706
Dependent Mean	4.78501	Adj R-Sq	0.6680
Coeff Var	7.02833		

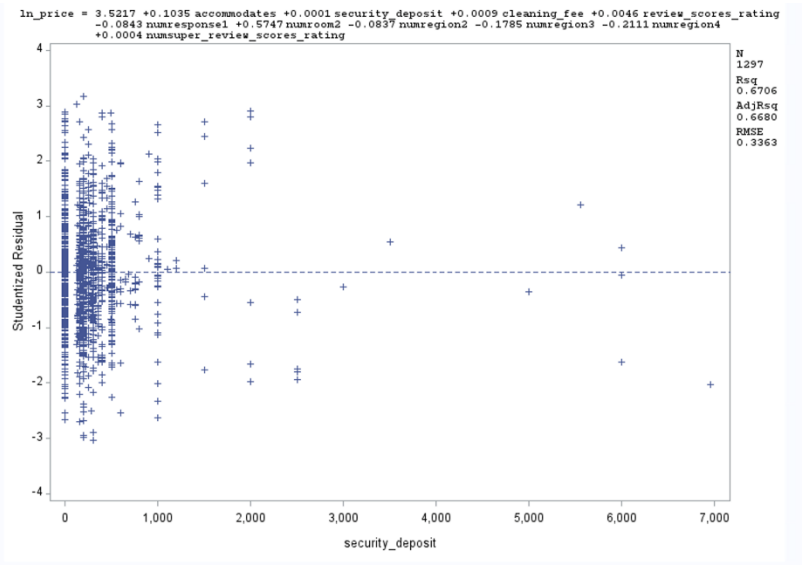
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.52172	0.10624	33.15	<.0001
accommodates	1	0.10353	0.00471	21.96	<.0001
security_deposit	1	0.00008683	0.00001917	4.53	<.0001
cleaning_fee	1	0.00087615	0.00021626	4.05	<.0001
review_scores_rating	1	0.00455	0.00111	4.09	<.0001
numresponse1	1	-0.08426	0.02047	-4.12	<.0001
numroom2	1	0.57472	0.02518	22.82	<.0001
numregion2	1	-0.08369	0.02555	-3.28	0.0011
numregion3	1	-0.17846	0.03373	-5.29	<.0001
numregion4	1	-0.21107	0.04479	-4.71	<.0001
numsuper_review_scores_rating	1	0.00043183	0.00021440	2.01	0.0442

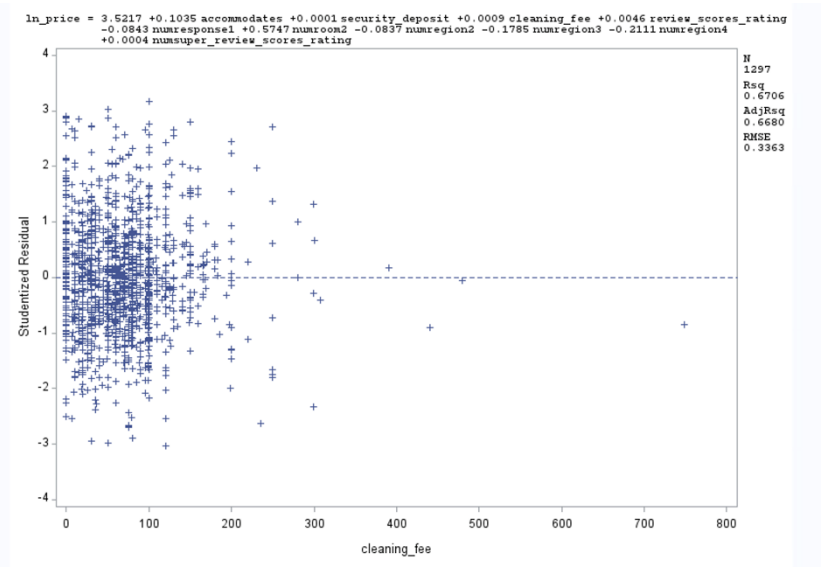
B.35 – Influence on price

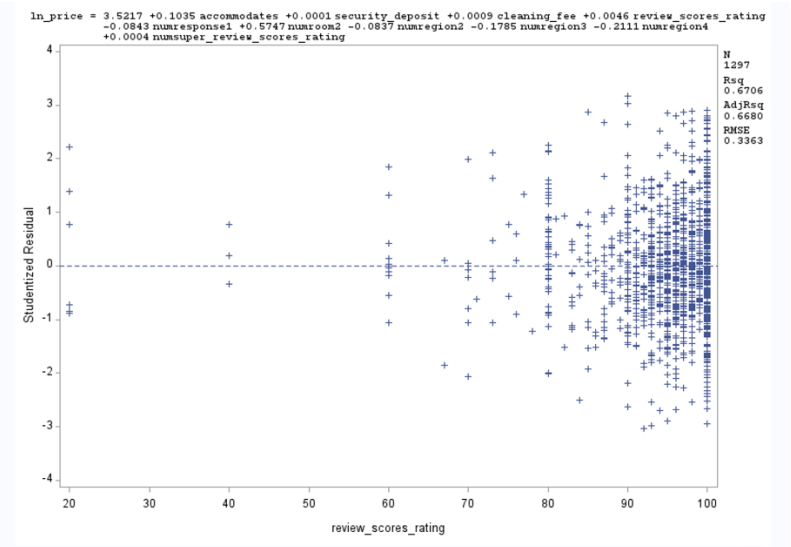
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	3.52172	0.10624	33.15	<.0001	0
accommodates	1	0.10353	0.00471	21.96	<.0001	0.44064
security_deposit	1	0.00008683	0.00001917	4.53	<.0001	0.07516
cleaning_fee	1	0.00087615	0.00021626	4.05	<.0001	0.08425
review_scores_rating	1	0.00455	0.00111	4.09	<.0001	0.06782
numresponse1	1	-0.08426	0.02047	-4.12	<.0001	-0.06871
numroom2	1	0.57472	0.02518	22.82	<.0001	0.43404
numregion2	1	-0.08369	0.02555	-3.28	0.0011	-0.05383
numregion3	1	-0.17846	0.03373	-5.29	<.0001	-0.08695
numregion4	1	-0.21107	0.04479	-4.71	<.0001	-0.07718
numsuper_review_scores_rating	1	0.00043183	0.00021440	2.01	0.0442	0.03386

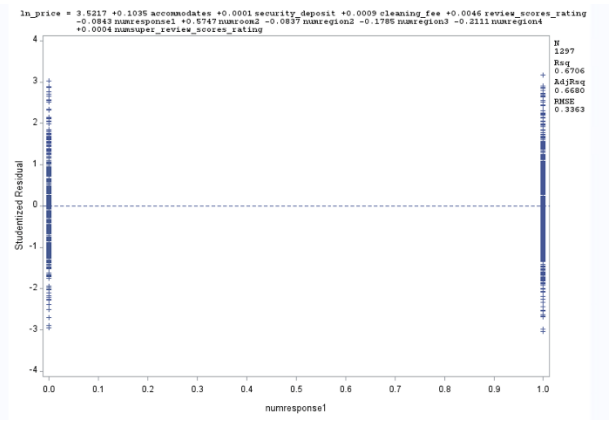
B.36 – Studentized Residual

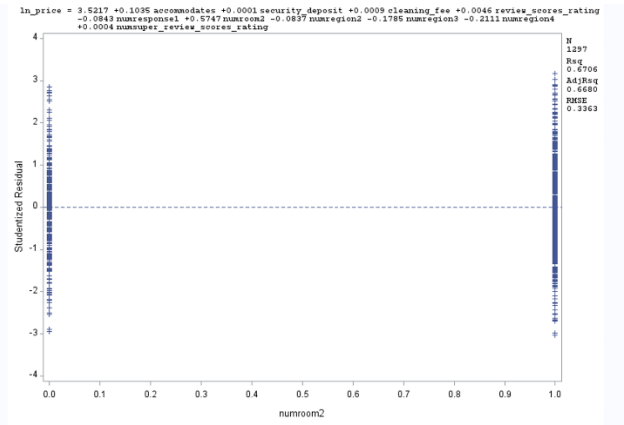


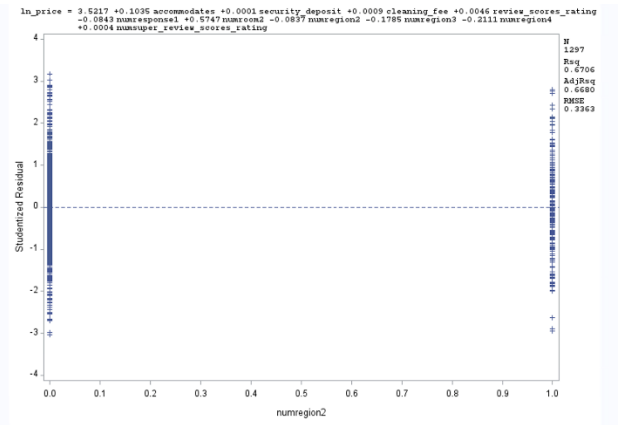

ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
security_deposit

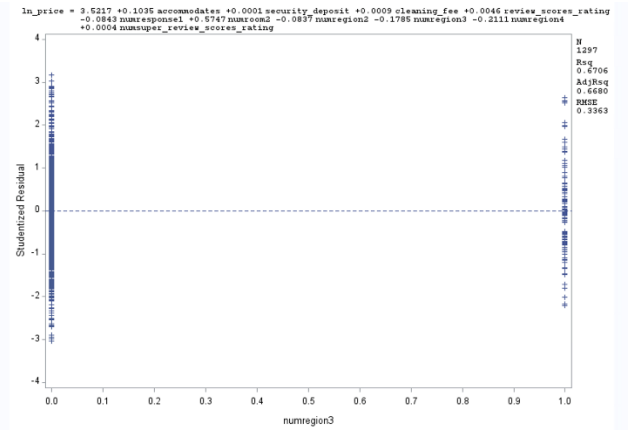

ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
cleaning_fee

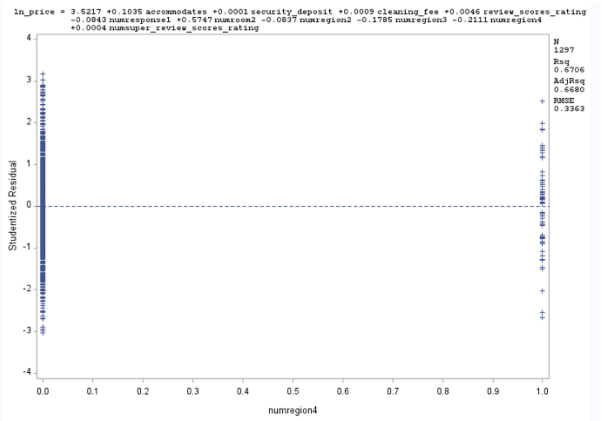

ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
review_scores_rating

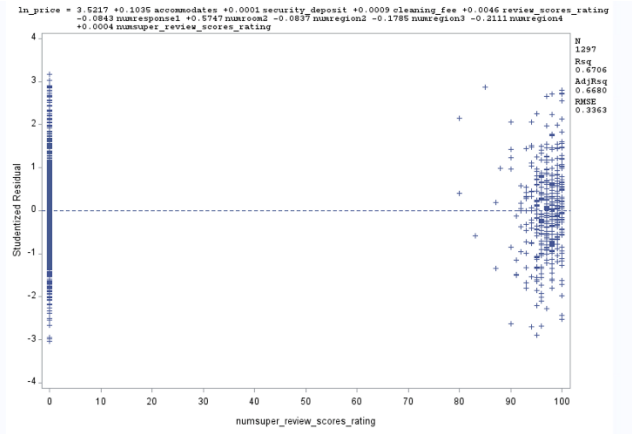

ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
numresponse1

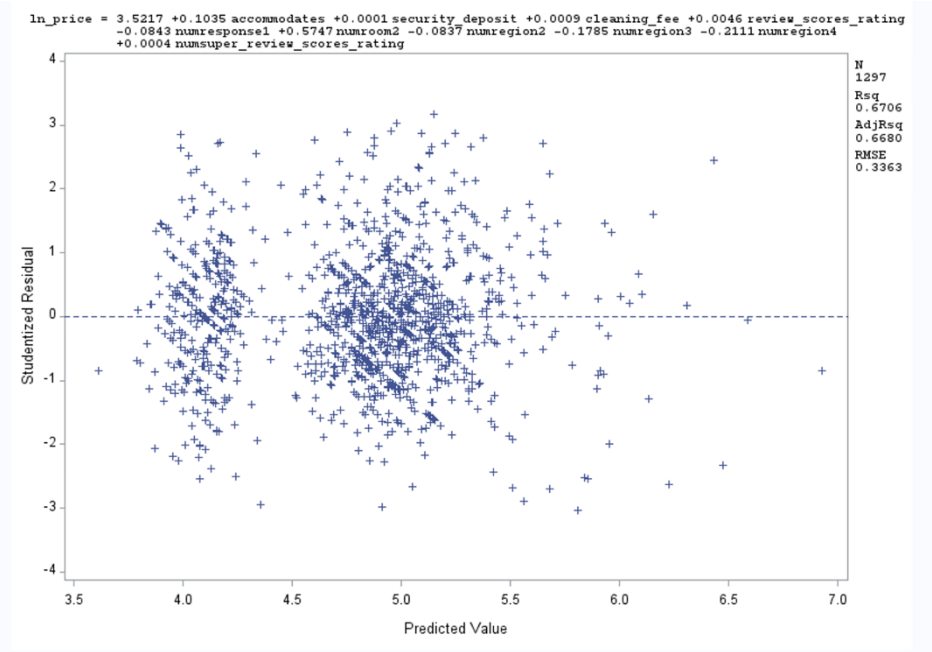

ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
numroom2

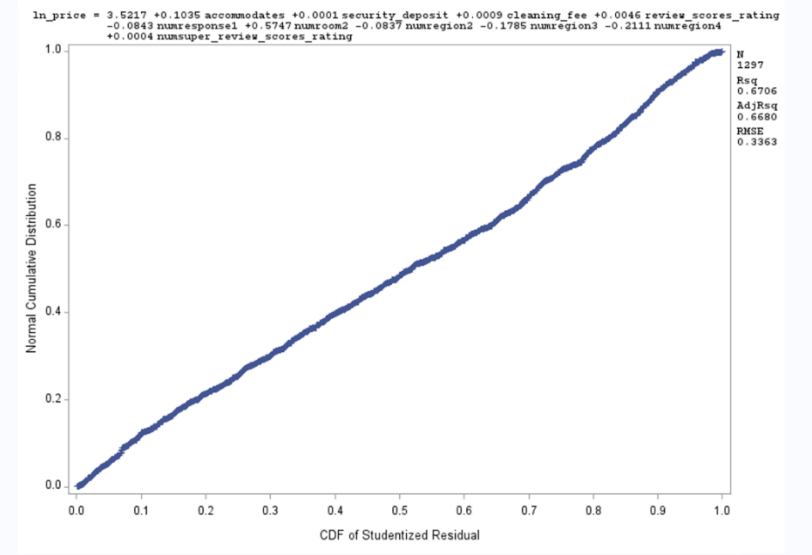

ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
numregion2


ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
numregion3


ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
numregion4


ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
numsuper_review_scores_rating


ln_price = 3.5217 +0.1035 accommodates +0.0001 security_deposit +0.0009 cleaning_fee +0.0046 review_scores_rating
-0.0843 numresponse1 +0.5747 numroom2 -0.0837 numregion2 -0.1785 numregion3 -0.2111 numregion4
+0.0004 numsuper_review_scores_rating
N
1297
Rsq
0.6706
AdjRsq
0.6680
RMSE
0.3363
Studentized Residual
Predicted Value



B.37 - Validation

The GLMSELECT Procedure

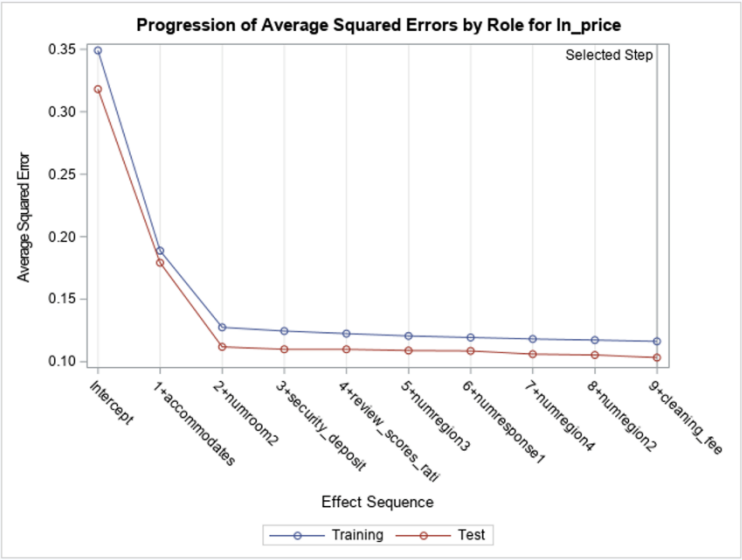
Data Set	WORK.AIRBNB3
Dependent Variable	ln_price
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None
Random Number Seed	338785001

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC	ASE	Test ASE	CV PRESS
0	Intercept		1	-982.2436	0.3492	0.3182	329.5222
1	accommodates		2	-1552.9857	0.1889	0.1792	178.1136
2	numroom2		3	-1915.5621	0.1275	0.1119	120.8806
3	security_deposit		4	-1930.7288	0.1245	0.1099	119.7798
4	review_scores_rating		5	-1939.4834	0.1225	0.1099	118.2201
5	numregion3		6	-1947.1860	0.1206	0.1089	116.6540
6	numresponse1		7	-1949.9639	0.1194	0.1086	115.6943
7	numregion4		8	-1952.6802	0.1182	0.1060	114.6113
8	numregion2		9	-1952.9177	0.1173	0.1054	114.0461
9	cleaning_fee		10	-1954.0582*	0.1163	0.1033	113.0107*

* Optimal Value of Criterion



Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	9	218.89257	24.32140	206.91
Error	930	109.31854	0.11755	
Corrected Total	939	328.21111		

Root MSE	0.34285
Dependent Mean	4.80003
R-Square	0.6669
Adj R-Sq	0.6637
AIC	-1060.51703
AICC	-1060.23255
SBC	-1954.05823
ASE (Train)	0.11630
ASE (Test)	0.10328
CV PRESS	113.01071

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	3.470059	0.117082	29.64	3.4247858	3.5126977	3.463534	3.414251	3.5368728
accommodates	1	0.108870	0.005535	19.67	0.1066105	0.1108244	0.112192	0.105674	0.1087871
security_deposit	1	0.000086997	0.000023251	3.74	0.0000853	0.0000919	0.000051	0.000146	0.0000736
cleaning_fee	1	0.000700	0.000249	2.82	0.0007382	0.0006611	0.000588	0.000660	0.0008437
review_scores_rating	1	0.005296	0.001213	4.37	0.0061245	0.0046426	0.005603	0.005614	0.0045186
numresponse1	1	-0.083287	0.024058	-3.46	-0.0779362	-0.0831122	-0.102185	-0.076525	-0.0773329
numroom2	1	0.566806	0.030492	18.59	0.5434787	0.5802999	0.559998	0.583771	0.5595401
numregion2	1	-0.087951	0.030507	-2.88	-0.1051110	-0.0866017	-0.105717	-0.074689	-0.0658938
numregion3	1	-0.190821	0.040970	-4.66	-0.2161524	-0.1760283	-0.204089	-0.190760	-0.1705298
numregion4	1	-0.188125	0.052602	-3.58	-0.1810813	-0.1849422	-0.188970	-0.178560	-0.2111526

B.38 - Prediction

The REG Procedure
Model: MODEL1
Dependent Variable: ln_price

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	4.6808	0.0295	4.6230	4.7387	4.0185	5.3431	.
2	.	4.8677	0.0376	4.7940	4.9414	4.2038	5.5316	.
3	3.91	.	.	.	.	.	.	.
4	4.50	.	.	.	.	.	.	.
5	4.74	4.9841	0.0165	4.9516	5.0165	4.3235	5.6446	-0.2479
6	3.91	.	.	.	.	.	.	.
7	4.67	.	.	.	.	.	.	.
8	4.93	4.8590	0.0474	4.7661	4.9520	4.1927	5.5253	0.0682
9	4.77	4.7602	0.0271	4.7071	4.8133	4.0983	5.4221	0.0105
10	5.39	4.8430	0.0178	4.8081	4.8778	4.1823	5.5036	0.5507
11	4.43	4.7722	0.0211	4.7308	4.8137	4.1111	5.4333	-0.3414
12	4.58	4.8199	0.0250	4.7708	4.8690	4.1583	5.4815	-0.2349

Appendix C– Shweta

Figure C.1 Descriptive Statistics

Descriptives

The MEANS Procedure

Variable	Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean	Minimum	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	Maximum
total_listings	16.5792104	36.6379427	0.8190446	14.9729405	18.1854803	0	1.0000000	1.0000000	1.0000000	3.0000000	13.0000000	276.0000000
acc	4.0074963	2.3584834	0.0527241	3.9040963	4.1108962	1.0000000	2.0000000	2.0000000	2.0000000	4.0000000	5.0000000	16.0000000
price	150.0964518	117.5139234	2.6270345	144.9444408	155.2484627	0	48.0000000	60.0000000	93.0000000	125.0000000	169.0000000	1501.00
security_deposit	289.0344828	357.6318838	7.9948936	273.3552907	304.7136749	0	0	0	141.0000000	200.0000000	350.0000000	5000.00
cleaning_fee	70.0084958	51.4387347	1.1499176	67.7533338	72.2636577	0	0	10.0000000	35.0000000	65.0000000	95.0000000	467.0000000
review_scores_rating	94.4217891	7.2665676	0.1624448	94.1032104	94.7403678	20.0000000	80.0000000	87.0000000	92.0000000	96.0000000	100.0000000	100.0000000

Figure C.2 Histogram for total_listings

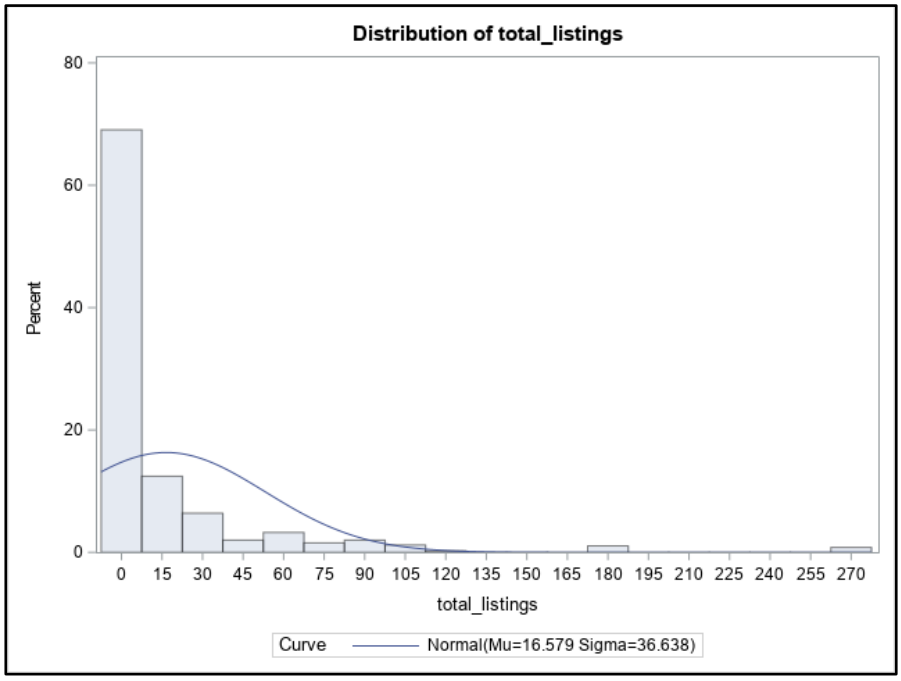


Figure C.3 Histogram for accomodates

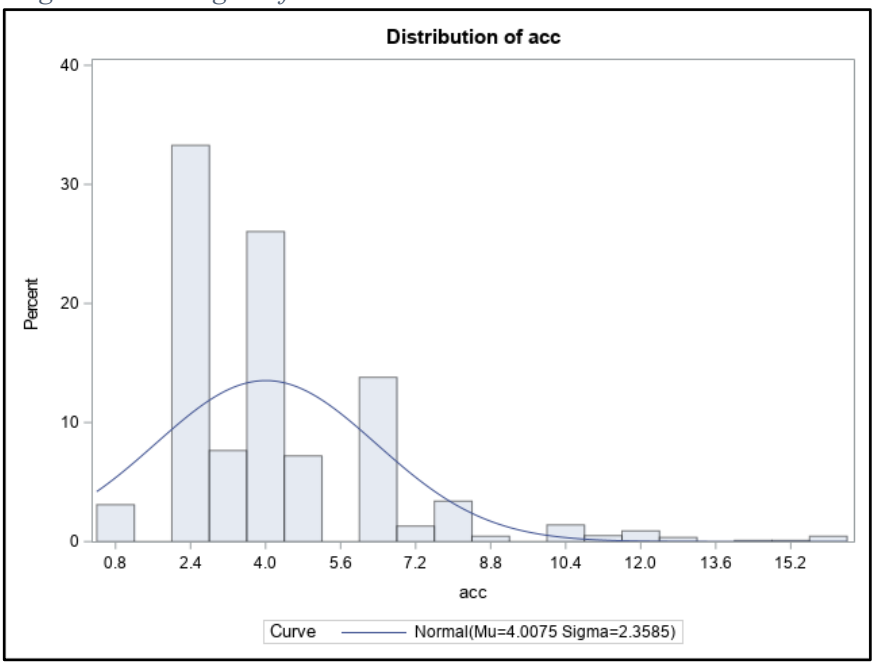


Figure C.4 Histogram for price

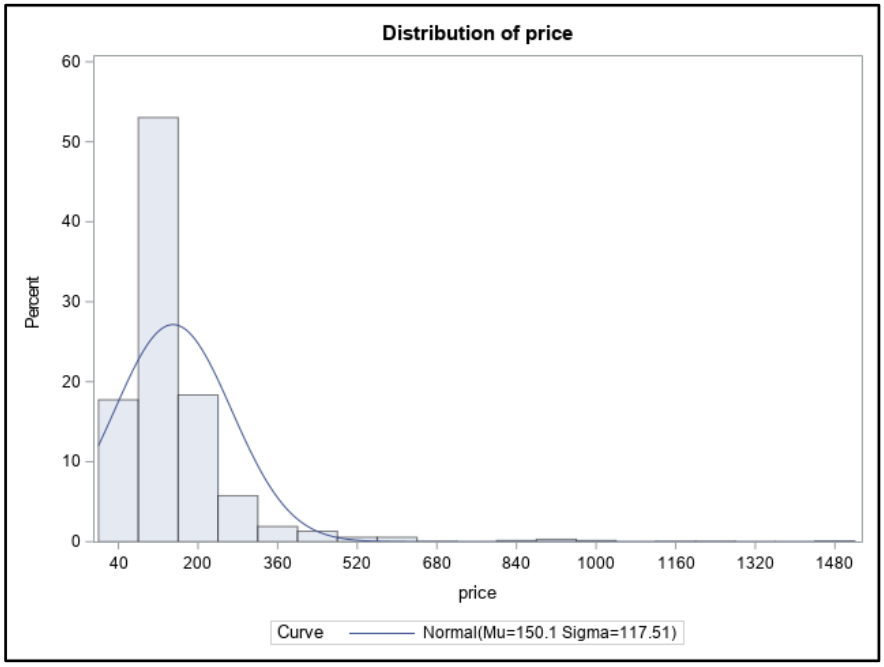


Figure C.4 Histogram for security deposit

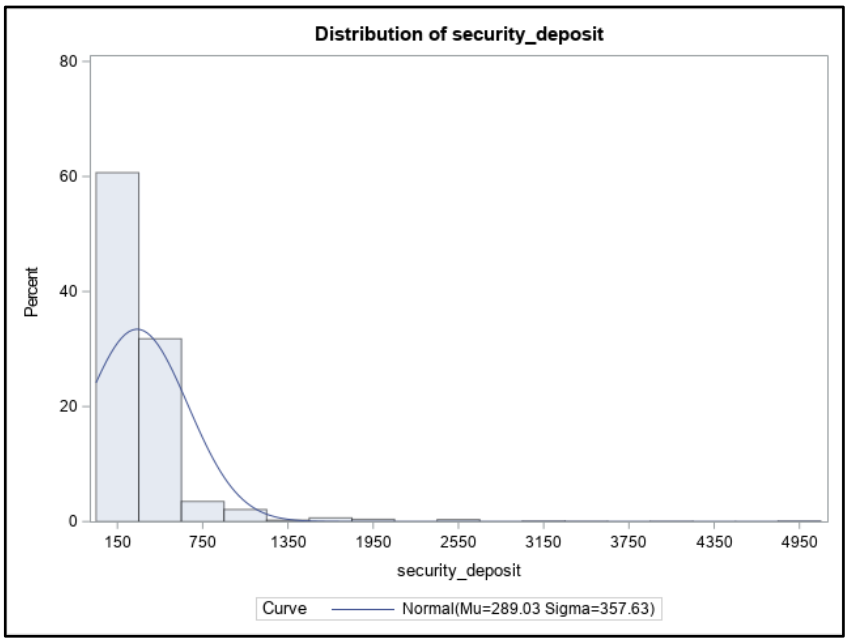


Figure C.6 Histogram for cleaning fee

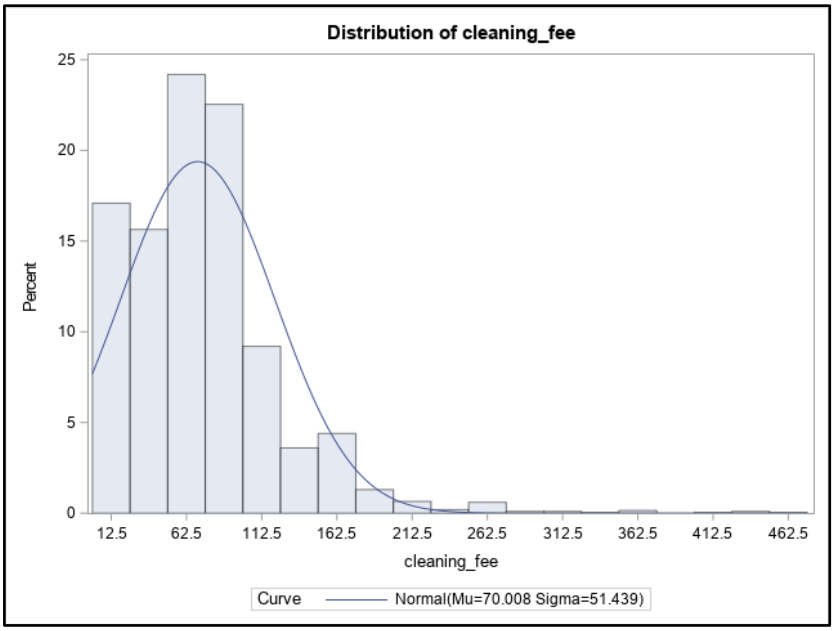


Figure C.7 Histogram for review score rating

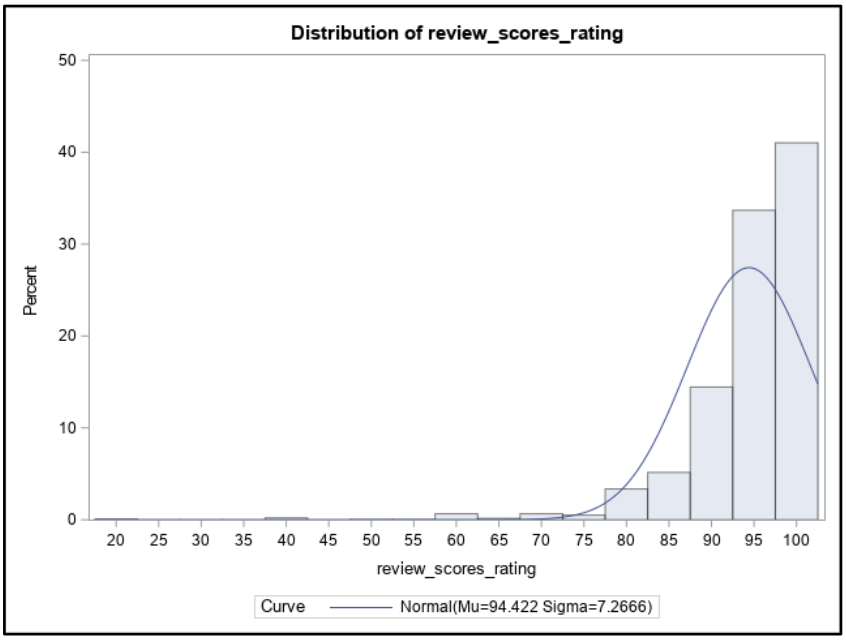


Figure C.8 Frequency Table for qualitative variables

Frequency Table

The FREQ Procedure

superhost	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	1273	63.62	1273	63.62
t	728	36.38	2001	100.00

region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
EM	167	8.35	167	8.35
IM	1236	61.77	1403	70.11
NSM	140	7.00	1543	77.11
SEM	358	17.89	1901	95.00
WM	100	5.00	2001	100.00

can_policy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Strict	1078	53.87	1078	53.87
flexible	302	15.09	1380	68.97
moderate	621	31.03	2001	100.00

room_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
P_room	429	21.44	429	21.44
home_apt	1572	78.56	2001	100.00

res_time	Frequency	Percent	Cumulative Frequency	Cumulative Percent
aday	140	7.00	140	7.00
anhour	1623	81.11	1763	88.11
fewhours	238	11.89	2001	100.00

Figure C.9 Scatter plot matrix

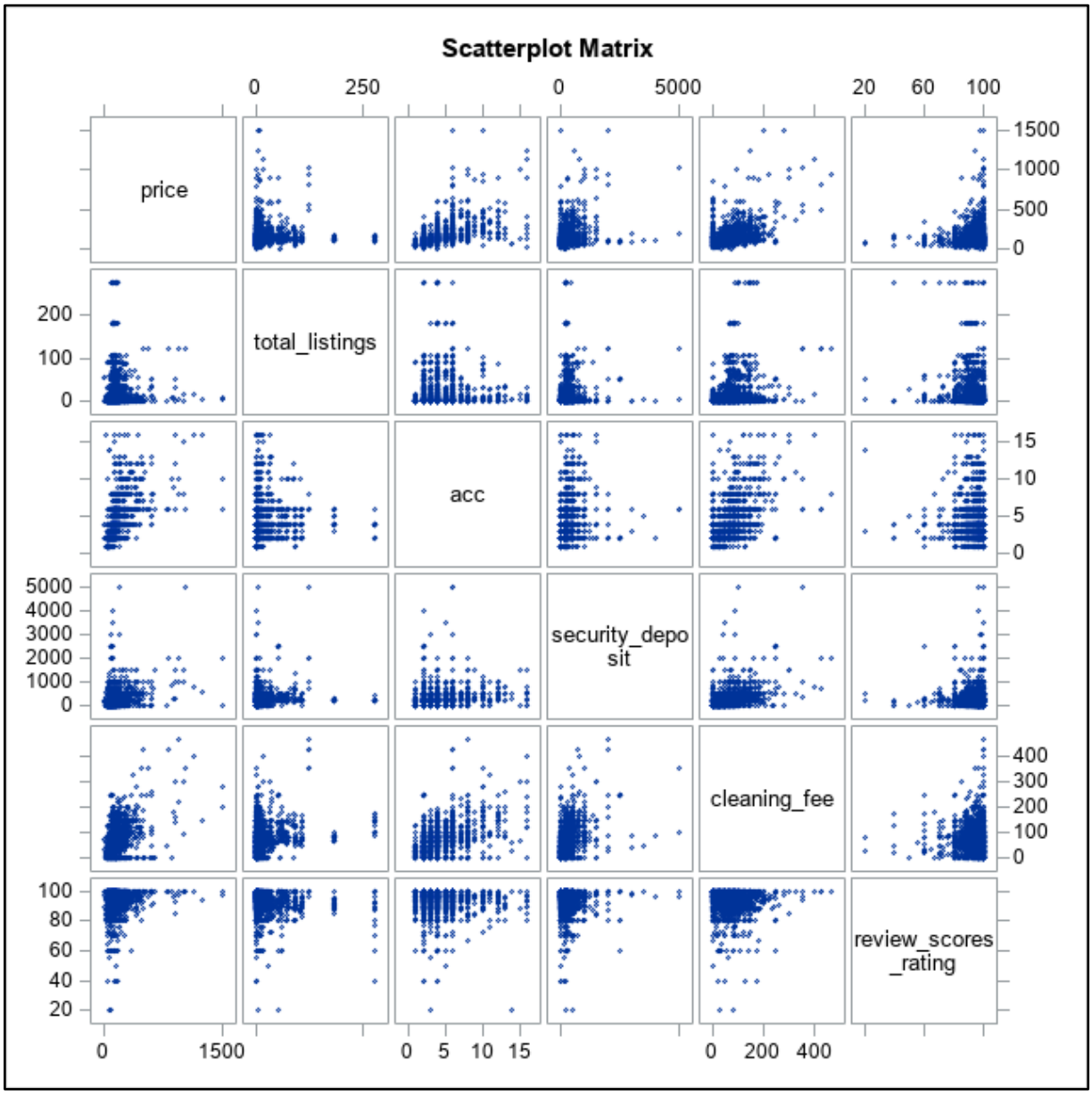


Figure C.10 Pearson Correlation Coefficient table

Pearson Correlation Coefficients, N = 2001
Prob > |r| under H0: Rho=0

	price	total_listings	acc	security_deposit	cleaning_fee	review_scores_rating	d_res2	d_res3	d_superhost	d_room	d_can2	d_can3	d_reg2	d_reg3	d_reg4	d_reg5	sd_sh
price	1.00000	0.04789 0.0322	0.54964 <.0001	0.28776 <.0001	0.47012 <.0001	0.07356 0.0010	0.02067 0.3554	-0.01538 0.4916	0.03900 0.0811	0.32830 <.0001	-0.03919 0.0797	0.10475 <.0001	-0.04215 0.0594	-0.00948 0.6716	0.06230 0.0053	-0.09241 <.0001	0.19026 <.0001
total_listings	0.04789 0.0322	1.00000	0.05906 0.0082	0.07524 0.0008	0.25102 <.0001	-0.16773 <.0001	-0.11105 <.0001	-0.07317 0.0011	-0.15208 <.0001	0.17000 <.0001	-0.14131 <.0001	0.23493 <.0001	-0.05162 0.0209	-0.04739 0.0340	-0.05736 0.0103	-0.08384 0.0002	-0.02212 0.3226
acc	0.54964 <.0001	0.05906 0.0082	1.00000	0.17865 <.0001	0.47537 <.0001	-0.05996 0.0073	-0.05747 0.0101	-0.07731 0.0005	-0.00769 0.7310	0.38073 <.0001	-0.07681 0.0006	0.15898 <.0001	-0.01998 0.3717	-0.03632 0.1043	0.11399 <.0001	0.04402 0.0490	0.10015 <.0001
security_deposit	0.28776 <.0001	0.07524 0.0008	0.17865 <.0001	1.00000	0.43799 <.0001	-0.01278 0.5678	-0.02408 0.2816	-0.00957 0.6687	0.00740 0.7408	0.16636 <.0001	-0.08187 0.0002	0.16830 <.0001	-0.03526 0.1148	0.00538 0.8099	-0.04803 0.0317	-0.05728 0.0104	0.64177 <.0001
cleaning_fee	0.47012 <.0001	0.25102 <.0001	0.47537 <.0001	0.43799 <.0001	1.00000	-0.04049 0.0701	-0.07004 0.0017	-0.03765 0.0923	-0.02327 0.2981	0.41648 <.0001	-0.07218 0.0012	0.19151 <.0001	-0.06523 0.0035	0.02386 0.2861	-0.08624 0.0001	-0.07416 0.0009	0.21861 <.0001
review_scores_rating	0.07356 0.0010	-0.16773 <.0001	-0.05996 0.0073	-0.01278 0.5678	-0.04049 0.0701	1.00000	0.02223 0.3203	0.03747 0.0938	0.28725 <.0001	0.00435 0.8458	0.12729 <.0001	-0.14251 <.0001	-0.01323 0.5543	0.00790 0.7241	0.03670 0.1007	0.01257 0.5741	0.14327 <.0001
d_res2	0.02067 0.3554	-0.11105 <.0001	-0.05747 0.0101	-0.02408 0.2816	-0.07004 0.0017	0.02223 0.3203	1.00000	-0.10078 <.0001	-0.05644 0.0116	-0.11275 <.0001	0.07721 0.0005	-0.09358 <.0001	0.00211 0.9249	0.05808 0.0094	0.00076 0.9727	0.05035 0.0243	-0.00486 0.8279
d_res3	-0.01538 0.4916	-0.07317 0.0011	-0.07731 0.0005	-0.00957 0.6687	-0.03765 0.0923	0.03747 0.0938	-0.10078 <.0001	1.00000	-0.11783 <.0001	-0.08585 0.0001	-0.01037 0.6430	-0.07240 0.0012	0.02461 0.2711	0.03043 0.1737	-0.00485 0.8285	0.00003 0.9989	-0.06114 0.0062
d_superhost	0.03900 0.0811	-0.15208 <.0001	-0.00769 0.7310	0.00740 0.7408	-0.02327 0.2981	0.28725 <.0001	-0.05644 0.0116	-0.11783 <.0001	1.00000	0.03057 0.1716	0.08099 0.0003	-0.05250 0.0188	-0.00381 0.8649	0.01830 0.4132	-0.01787 0.4244	0.00771 0.7302	0.47663 <.0001
d_room	0.32830 <.0001	0.17000 <.0001	0.38073 <.0001	0.16636 <.0001	0.41648 <.0001	0.00435 0.8458	-0.11275 <.0001	-0.08585 0.0001	0.03057 0.1716	1.00000	-0.06544 0.0034	0.19815 <.0001	-0.11927 <.0001	-0.02938 0.1889	-0.01407 0.5292	-0.15961 <.0001	0.08876 <.0001
d_can2	-0.03919 0.0797	-0.14131 <.0001	-0.07681 0.0006	-0.08187 0.0002	-0.07218 0.0012	0.12729 <.0001	0.07721 0.0005	-0.01037 0.6430	0.08099 0.0003	-0.06544 0.0034	1.00000	-0.72496 <.0001	0.00234 0.9168	0.05608 0.0121	0.05926 0.0080	0.00479 0.8306	-0.01784 0.4252
d_can3	0.10475 <.0001	0.23493 <.0001	0.15898 <.0001	0.16830 <.0001	0.19151 <.0001	-0.14251 <.0001	-0.09358 <.0001	-0.07240 0.0012	-0.05250 0.0188	0.19815 <.0001	-0.72496 <.0001	1.00000	-0.05668 0.0112	-0.07550 0.0007	-0.06513 0.0036	-0.07763 0.0005	0.05160 0.0210
d_reg2	-0.04215 0.0594	-0.05162 0.0209	-0.01998 0.3717	-0.03526 0.1148	-0.06523 0.0035	-0.01323 0.5543	0.00211 0.9249	0.02461 0.2711	-0.00381 0.8649	-0.11927 <.0001	0.00234 0.9168	-0.05668 0.0112	1.00000	-0.12803 <.0001	-0.08277 0.0002	-0.06291 0.0049	-0.04846 0.0302
d_reg3	-0.00948 0.6716	-0.04739 0.0340	-0.03632 0.1043	0.00538 0.8099	0.02386 0.2861	0.00790 0.7241	0.05808 0.0094	0.03043 0.1737	0.01830 0.4132	-0.02938 0.1889	0.05608 0.0121	-0.07550 0.0007	-0.12803 <.0001	1.00000	-0.14086 <.0001	-0.10706 <.0001	0.01655 0.4594
d_reg4	0.06230 0.0053	-0.05736 0.0103	0.11399 <.0001	-0.04803 0.0317	-0.08624 0.0001	0.03670 0.1007	0.00076 0.9727	-0.00485 0.8285	-0.01787 0.4244	-0.01407 0.5292	0.05926 0.0080	-0.06513 0.0036	-0.08277 0.0002	-0.14086 <.0001	1.00000	-0.06921 0.0020	-0.04955 0.0267
d_reg5	-0.09241 <.0001	-0.08384 0.0002	0.04402 0.0490	-0.05728 0.0104	-0.07416 0.0009	0.01257 0.5741	0.05035 0.0243	0.00003 0.9989	0.00771 0.7302	-0.15961 <.0001	0.00479 0.8306	-0.07763 0.0005	-0.06291 0.0049	-0.10706 <.0001	-0.06921 0.0020	1.00000	-0.02605 0.2441
sd_sh	0.19026 <.0001	-0.02212 0.3226	0.10015 <.0001	0.64177 <.0001	0.21861 <.0001	0.14327 <.0001	-0.00486 0.8279	-0.06114 0.0062	0.47663 <.0001	0.08876 <.0001	-0.01784 0.4252	0.05160 0.0210	-0.04846 0.0302	0.01655 0.4594	-0.04955 0.0267	-0.02605 0.2441	1.00000

Figure C.11 Full Regression Model

Full Regression Model_1 for Price

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	2001
Number of Observations Used	2001

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	11054870	690929	82.76	<.0001
Error	1984	16564174	8348.87823		
Corrected Total	2000	27619044			

Root MSE	91.37220	R-Square	0.4003
Dependent Mean	150.09645	Adj R-Sq	0.3954
Coeff Var	60.87565		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-133.46568	28.97092	-4.61	<.0001	0	0
total_listings	1	-0.07235	0.06092	-1.19	0.2351	-0.02256	1.19339
acc	1	21.11167	1.05339	20.04	<.0001	0.42371	1.47858
security_deposit	1	0.03725	0.00907	4.11	<.0001	0.11336	2.51879
cleaning_fee	1	0.47361	0.05347	8.86	<.0001	0.20731	1.81237
review_scores_rating	1	1.51716	0.29964	5.06	<.0001	0.09381	1.13573
d_res2	1	26.91881	6.48880	4.15	<.0001	0.07417	1.05751
d_res3	1	16.43878	8.25581	1.99	0.0466	0.03569	1.06297
d_superhost	1	5.72444	5.64746	1.01	0.3109	0.02344	1.76926
d_room	1	18.19029	5.85417	3.11	0.0019	0.06354	1.38346
d_can2	1	-5.04343	6.53505	-0.77	0.4404	-0.01986	2.19075
d_can3	1	-6.21231	6.34484	-0.98	0.3276	-0.02636	2.39766
d_reg2	1	-7.51898	8.29947	-0.91	0.3651	-0.01633	1.07424
d_reg3	1	-4.45586	5.54742	-0.80	0.4219	-0.01454	1.08349
d_reg4	1	10.18351	7.76171	1.31	0.1897	0.02397	1.10447
d_reg5	1	-47.92809	9.83581	-4.87	<.0001	-0.08889	1.10085
sd_sh	1	0.00034658	0.01158	0.03	0.9761	0.00087107	2.80427

Figure C.12 Fit Diagnostics for price

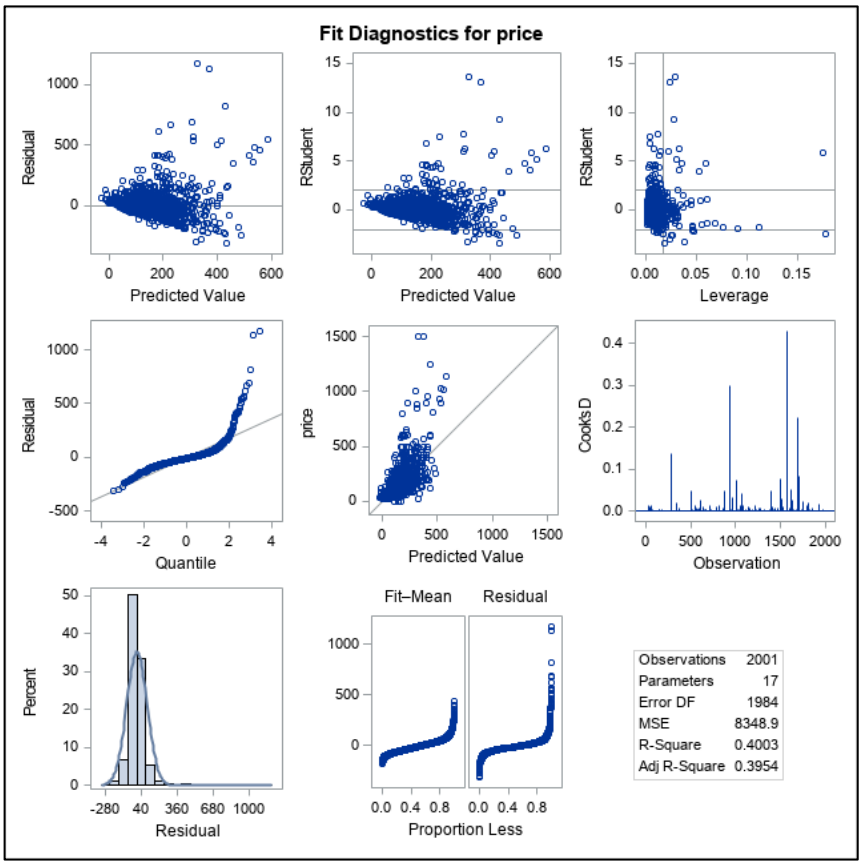


Figure C.13

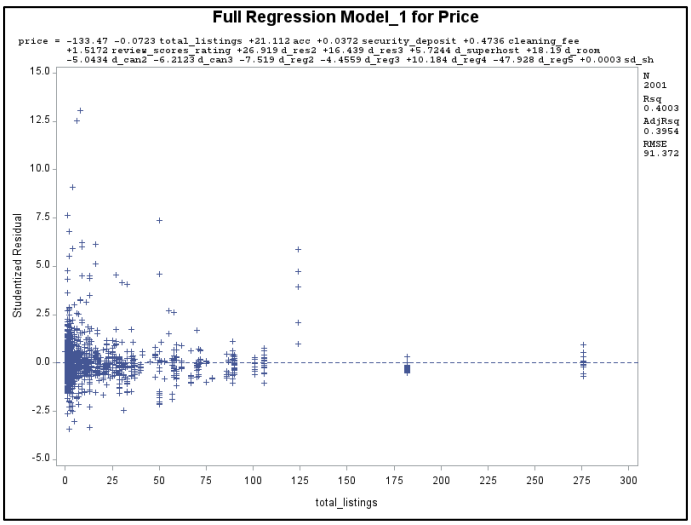


Figure C.14

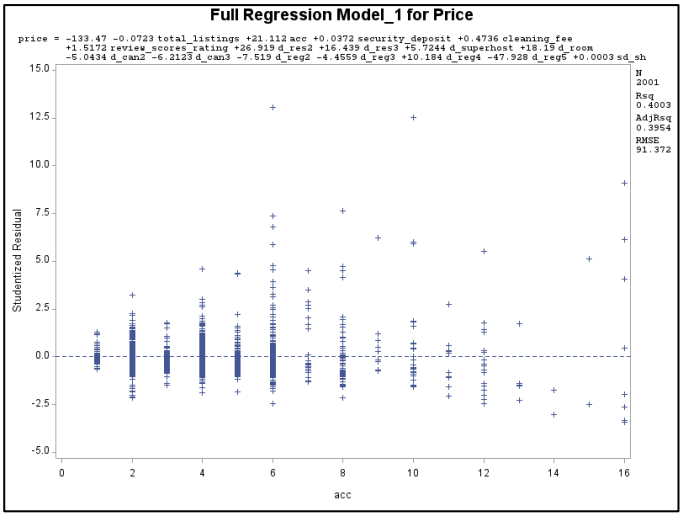


Figure C.15

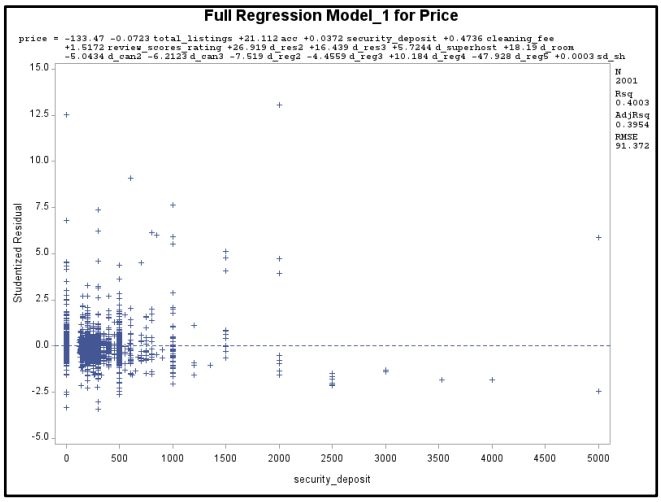


Figure C.16

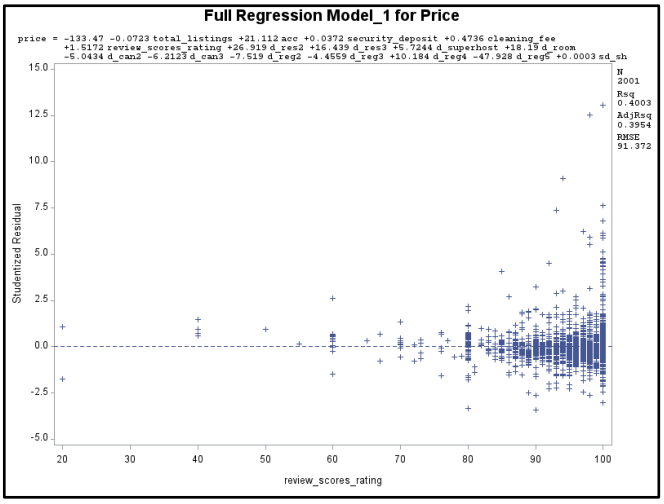


Figure C.17

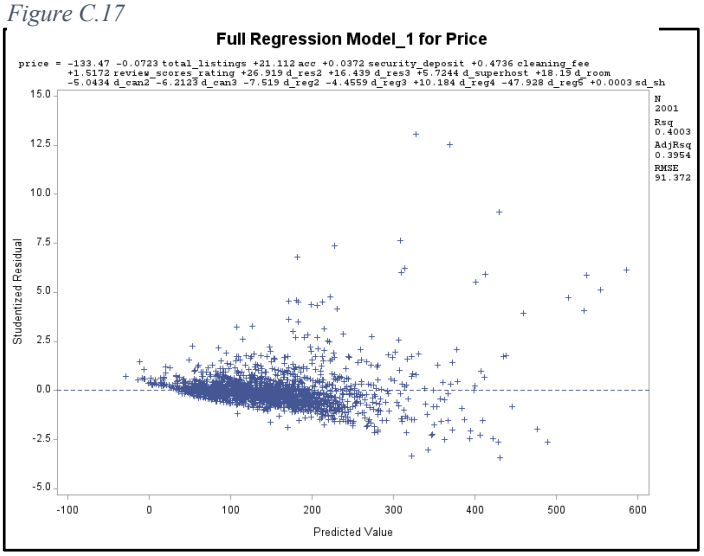


Figure C.18 Normal probability plot

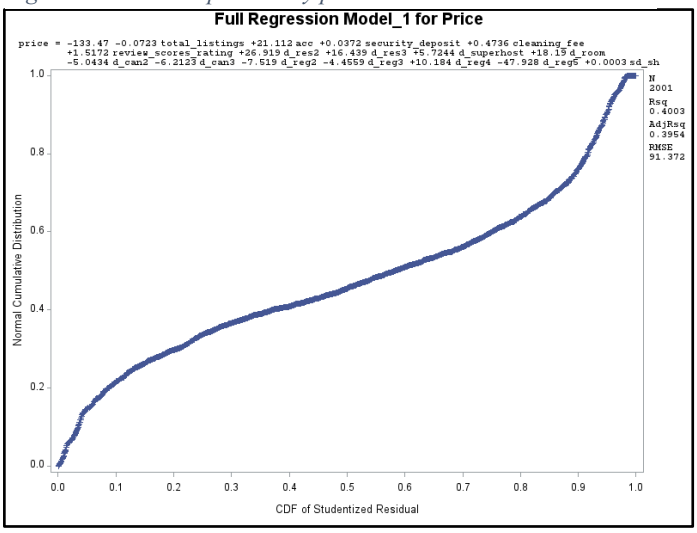


Figure C. 19 Full Regression Model after removing outliers' _1

Full Regression Model I for Price

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	1976
Number of Observations Used	1976

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	6027622	376726	97.37	<.0001
Error	1959	7579548	3869.09022		
Corrected Total	1975	13607170			

Root MSE	62.20201	R-Square	0.4430
Dependent Mean	141.66296	Adj R-Sq	0.4384
Coeff Var	43.90845		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-50.05122	19.84642	-2.52	0.0118	0	0
total_listings	1	-0.07965	0.04167	-1.91	0.0561	-0.03519	1.19208
acc	1	15.60082	0.74653	20.90	<.0001	0.42570	1.45935
security_deposit	1	0.00363	0.00642	0.56	0.5722	0.01459	2.34543
cleaning_fee	1	0.35135	0.03813	9.21	<.0001	0.20676	1.77080
review_scores_rating	1	0.88503	0.20481	4.32	<.0001	0.07773	1.13806
d_res2	1	17.21046	4.46917	3.85	0.0001	0.06689	1.06099
d_res3	1	13.23659	5.65849	2.34	0.0194	0.04065	1.06226
d_superhost	1	0.86119	3.91713	0.22	0.8260	0.00499	1.81170
d_room	1	33.98072	4.03911	8.41	<.0001	0.16868	1.41381
d_can2	1	-6.75921	4.45573	-1.52	0.1294	-0.03771	2.17358
d_can3	1	-6.64406	4.32732	-1.54	0.1249	-0.03993	2.37853
d_reg2	1	-17.10270	5.70775	-3.00	0.0028	-0.05235	1.07359
d_reg3	1	-3.66278	3.79879	-0.96	0.3351	-0.01693	1.08380
d_reg4	1	9.01525	5.34776	1.69	0.0920	0.02980	1.09926
d_reg5	1	-37.83209	6.70533	-5.64	<.0001	-0.09993	1.10326
sd_sh	1	0.01408	0.00850	1.66	0.0977	0.04597	2.70777

Figure C. 20 Fit Diagnostics for price

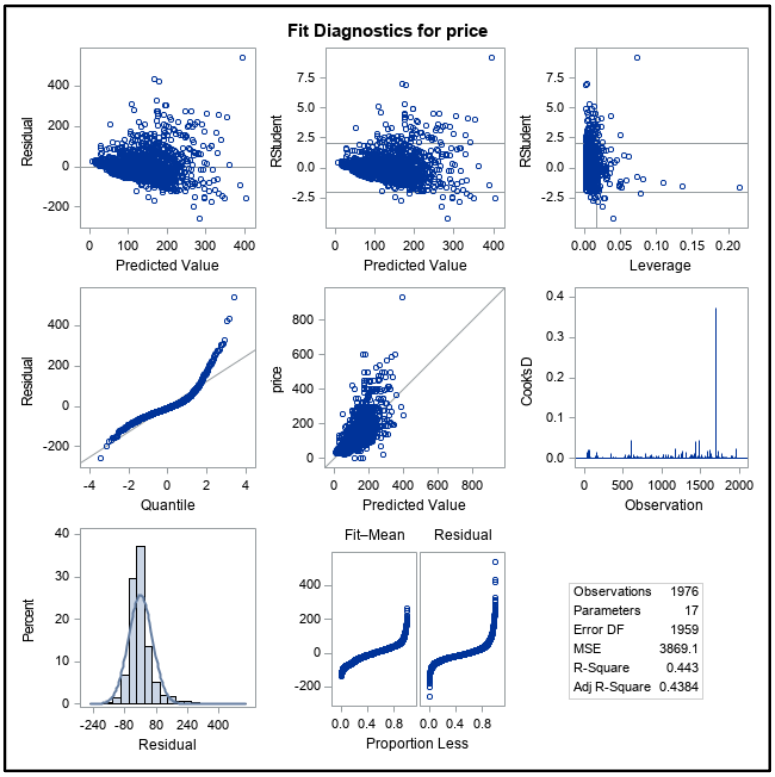


Figure C.21 Full Regression Model after removing outliers_2

Full Regression Full Model_II for Price

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	1952
Number of Observations Used	1952

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	4774831	298427	107.09	<.0001
Error	1935	5392289	2786.71277		
Corrected Total	1951	10167120			

Root MSE	52.78932	R-Square	0.4696
Dependent Mean	137.39857	Adj R-Sq	0.4652
Coeff Var	38.42058		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-30.64156	17.01174	-1.80	0.0718	0	0
total_listings	1	-0.09042	0.03557	-2.54	0.0111	-0.04596	1.19259
acc	1	14.78609	0.65244	22.66	<.0001	0.45389	1.46348
security_deposit	1	0.00866	0.00551	1.57	0.1158	0.03982	2.33683
cleaning_fee	1	0.20252	0.03419	5.92	<.0001	0.13130	1.79312
review_scores_rating	1	0.72882	0.17560	4.15	<.0001	0.07331	1.13837
d_res2	1	14.18934	3.83966	3.70	0.0002	0.06315	1.06533
d_res3	1	13.94521	4.82062	2.89	0.0039	0.04936	1.06226
d_superhost	1	3.33847	3.33990	1.00	0.3176	0.02225	1.80722
d_room	1	40.28475	3.46188	11.64	<.0001	0.23056	1.43224
d_can2	1	-5.03632	3.81255	-1.32	0.1867	-0.03233	2.18551
d_can3	1	-5.28649	3.70583	-1.43	0.1539	-0.03653	2.39254
d_reg2	1	-18.58560	4.86384	-3.82	0.0001	-0.06556	1.07409
d_reg3	1	-8.95865	3.25961	-2.75	0.0060	-0.04735	1.08291
d_reg4	1	4.76507	4.57567	1.04	0.2978	0.01806	1.09727
d_reg5	1	-36.15360	5.73310	-6.31	<.0001	-0.10939	1.09785
sd_sh	1	0.00158	0.00734	0.21	0.8299	0.00584	2.69668

Figure C. 22 Fit Diagnostics for Price

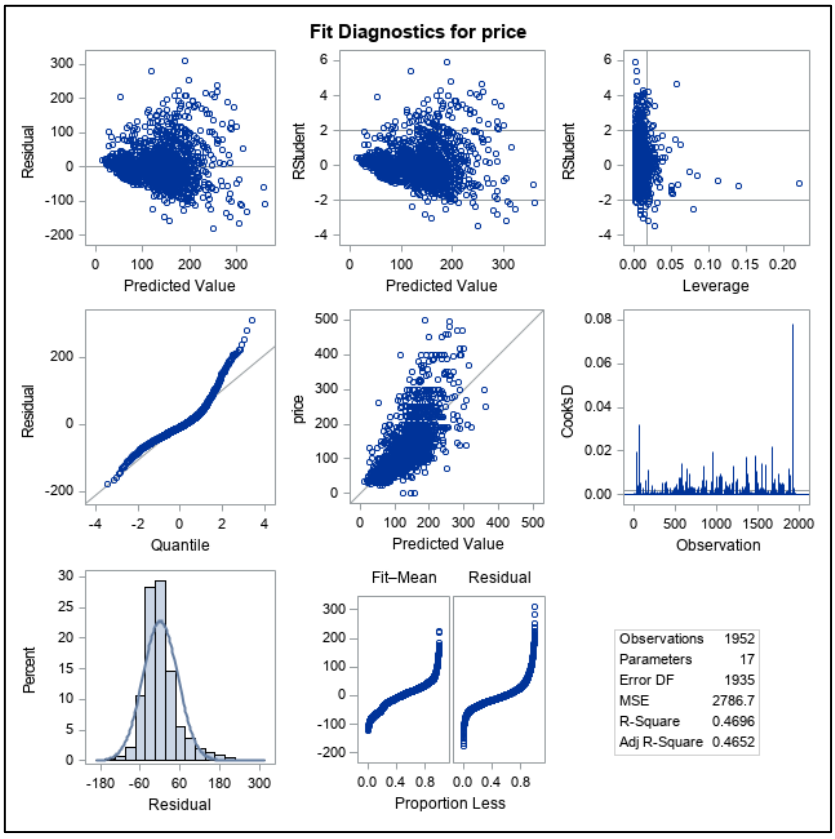


Figure C.23 Full Regression Model after removing outliers _3

Full Regression Full Model_III for Price

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	1931
Number of Observations Used	1931

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	4199256	262454	114.59	<.0001
Error	1914	4383910	2290.44418		
Corrected Total	1930	8583167			

Root MSE	47.85859	R-Square	0.4892
Dependent Mean	134.60228	Adj R-Sq	0.4850
Coeff Var	35.55555		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-25.85450	15.47121	-1.67	0.0949	0	0
total_listings	1	-0.07779	0.03234	-2.41	0.0162	-0.04291	1.19204
acc	1	14.22349	0.60885	23.36	<.0001	0.46491	1.48413
security_deposit	1	0.01080	0.00507	2.13	0.0334	0.05377	2.38936
cleaning_fee	1	0.15402	0.03233	4.76	<.0001	0.10591	1.85242
review_scores_rating	1	0.69667	0.15964	4.36	<.0001	0.07615	1.14109
d_res2	1	12.32655	3.51041	3.51	0.0005	0.05932	1.06948
d_res3	1	12.58393	4.40568	2.86	0.0043	0.04813	1.06406
d_superhost	1	3.28396	3.04045	1.08	0.2802	0.02368	1.80107
d_room	1	40.28611	3.16022	12.75	<.0001	0.25014	1.44282
d_can2	1	-3.32076	3.47903	-0.95	0.3399	-0.02308	2.19144
d_can3	1	-3.14946	3.38118	-0.93	0.3517	-0.02356	2.39709
d_reg2	1	-20.21221	4.43948	-4.55	<.0001	-0.07704	1.07304
d_reg3	1	-7.63644	2.96770	-2.57	0.0102	-0.04373	1.08215
d_reg4	1	1.56044	4.19081	0.37	0.7097	0.00636	1.09312
d_reg5	1	-36.22505	5.22350	-6.94	<.0001	-0.11868	1.09747
sd_sh	1	-0.00283	0.00672	-0.42	0.6738	-0.01132	2.70816

Figure C.24 Fit Diagnostics for price

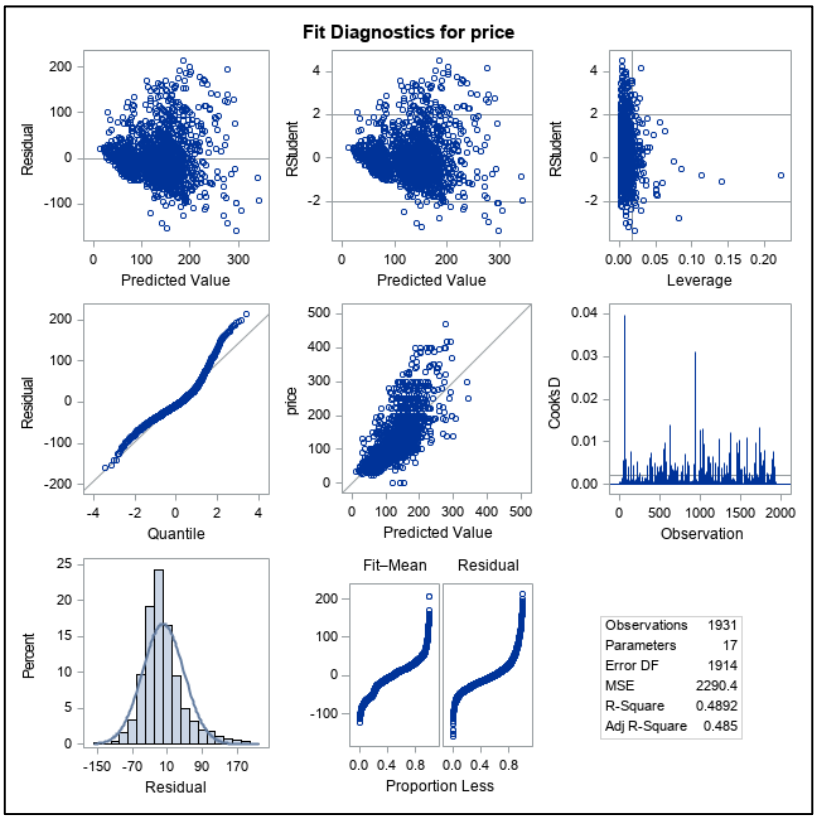


Figure C.25

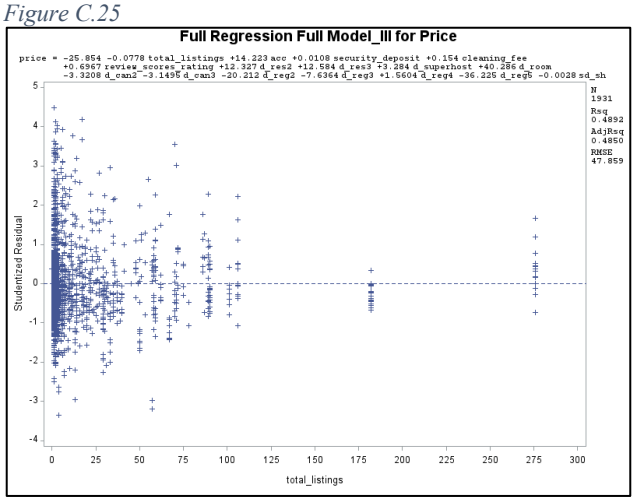


Figure C.26

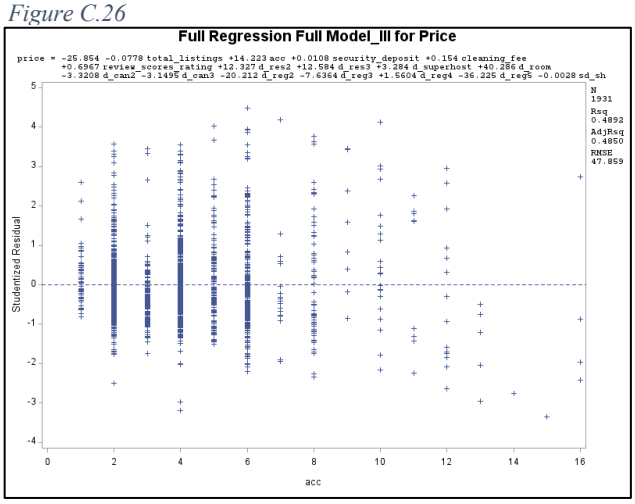


Figure C.27

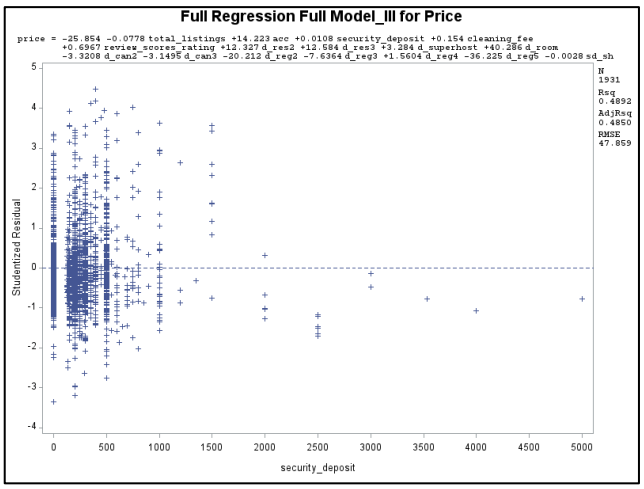


Figure C.28

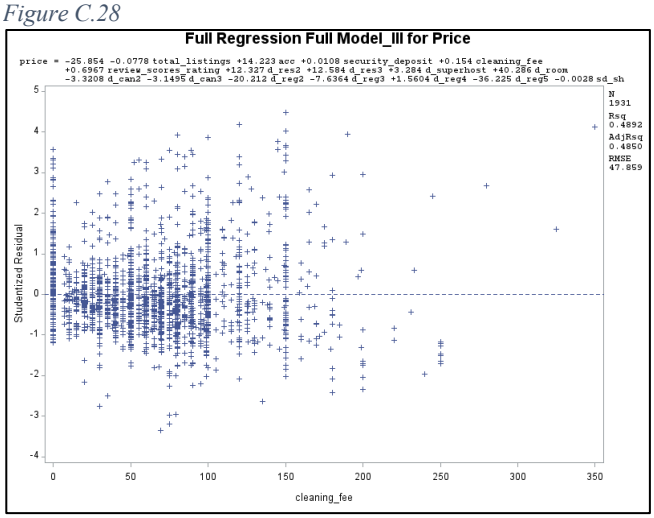


Figure C.29

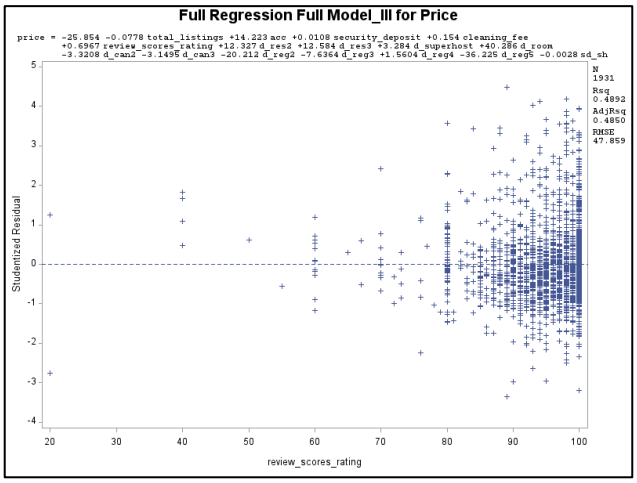


Figure C.30

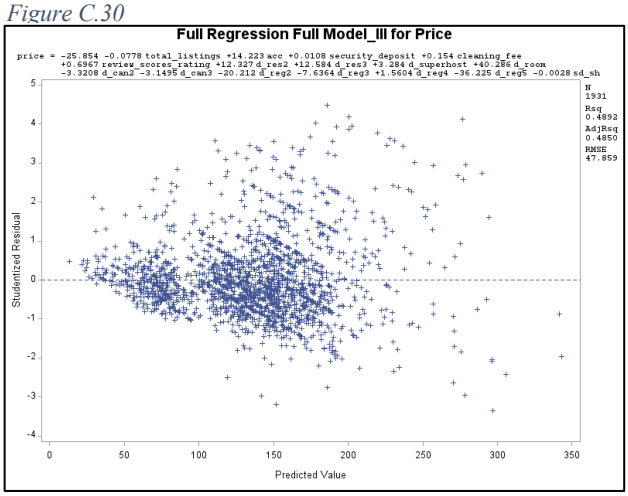


Figure C.31 Normal probability plot

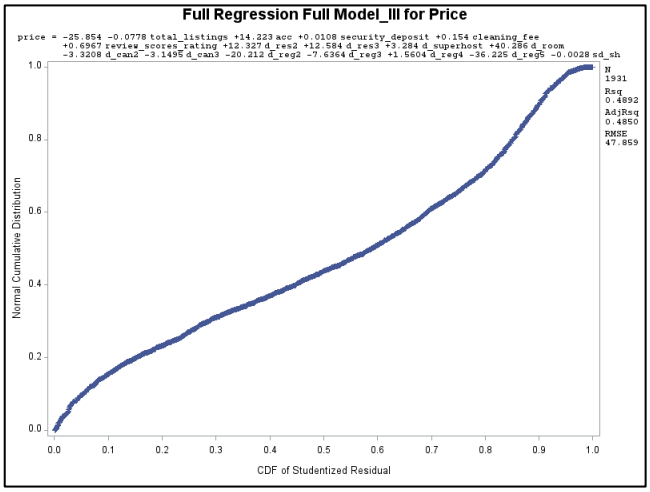


Figure C.32 Histogram for price

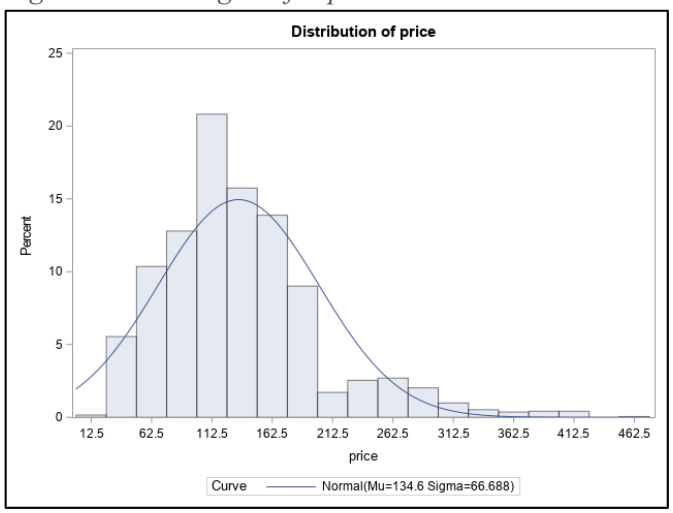


Figure C.33 Scatterplot matrix after removing the influential points & outliers

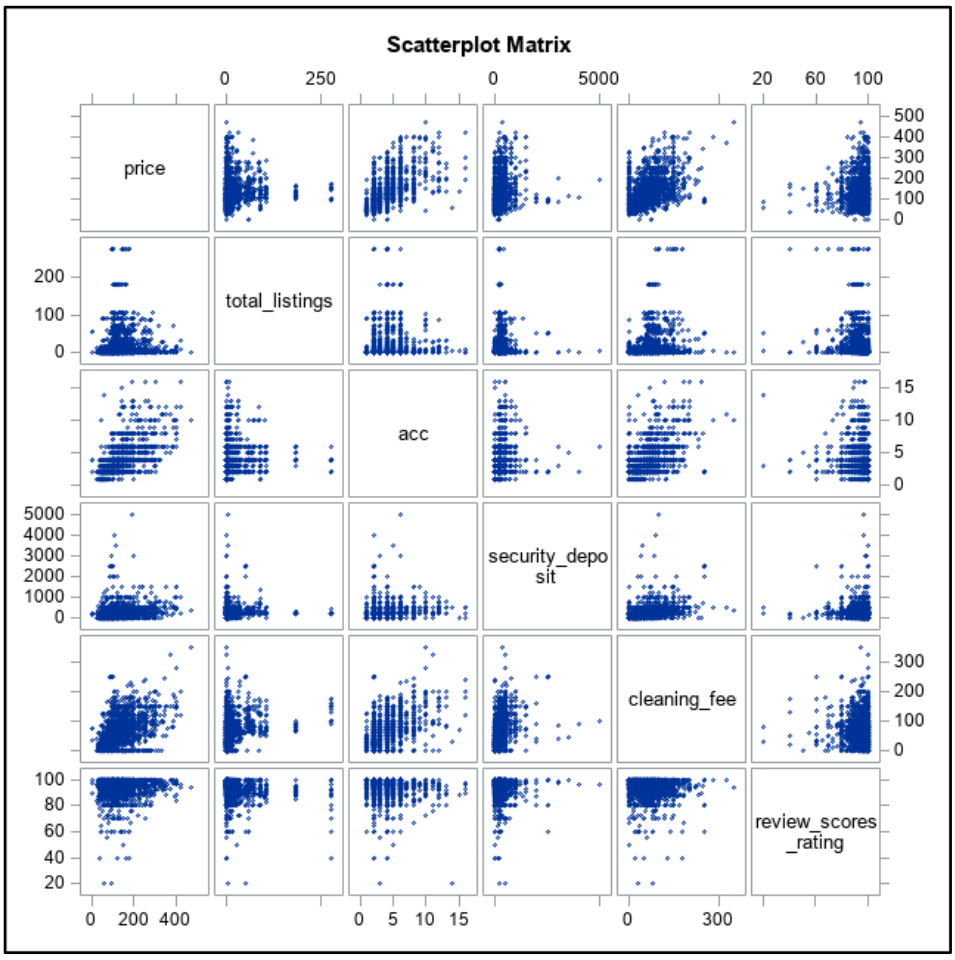


Figure C.34 Full Regression Model with log transformation of 'price'

Model with Transformed y-variable - ln_Price

The REG Procedure
Model: MODEL1
Dependent Variable: ln_Price

Number of Observations Read	1931
Number of Observations Used	1928
Number of Observations with Missing Values	3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	283.70727	17.73170	172.64	<.0001
Error	1911	196.27402	0.10271		
Corrected Total	1927	479.98129			

Root MSE	0.32048	R-Square	0.5911
Dependent Mean	4.78536	Adj R-Sq	0.5877
Coeff Var	6.69709		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	3.57177	0.10361	34.47	<.0001	0	0
total_listings	1	-0.00024268	0.00021666	-1.12	0.2628	-0.01789	1.19189
acc	1	0.08945	0.00408	21.92	<.0001	0.39089	1.48564
security_deposit	1	0.00007283	0.00003397	2.14	0.0322	0.04847	2.38939
cleaning_fee	1	0.00064690	0.00021654	2.99	0.0028	0.05948	1.85220
review_scores_rating	1	0.00454	0.00107	4.25	<.0001	0.06637	1.14104
d_res2	1	0.06135	0.02351	2.61	0.0091	0.03948	1.06941
d_res3	1	0.08413	0.02951	2.85	0.0044	0.04303	1.06417
d_superhost	1	0.03095	0.02037	1.52	0.1288	0.02982	1.80039
d_room	1	0.52319	0.02117	24.71	<.0001	0.43431	1.44371
d_can2	1	-0.00602	0.02333	-0.26	0.7966	-0.00559	2.19498
d_can3	1	-0.01266	0.02268	-0.56	0.5769	-0.01265	2.40105
d_reg2	1	-0.20190	0.02973	-6.79	<.0001	-0.10291	1.07296
d_reg3	1	-0.06806	0.01987	-3.42	0.0006	-0.05210	1.08182
d_reg4	1	-0.00412	0.02832	-0.15	0.8844	-0.00222	1.09361
d_reg5	1	-0.32889	0.03498	-9.40	<.0001	-0.14408	1.09749
sd_sh	1	-0.00002271	0.00004500	-0.50	0.6138	-0.01215	2.70824

Figure C.35 Fit Diagnostics for Price

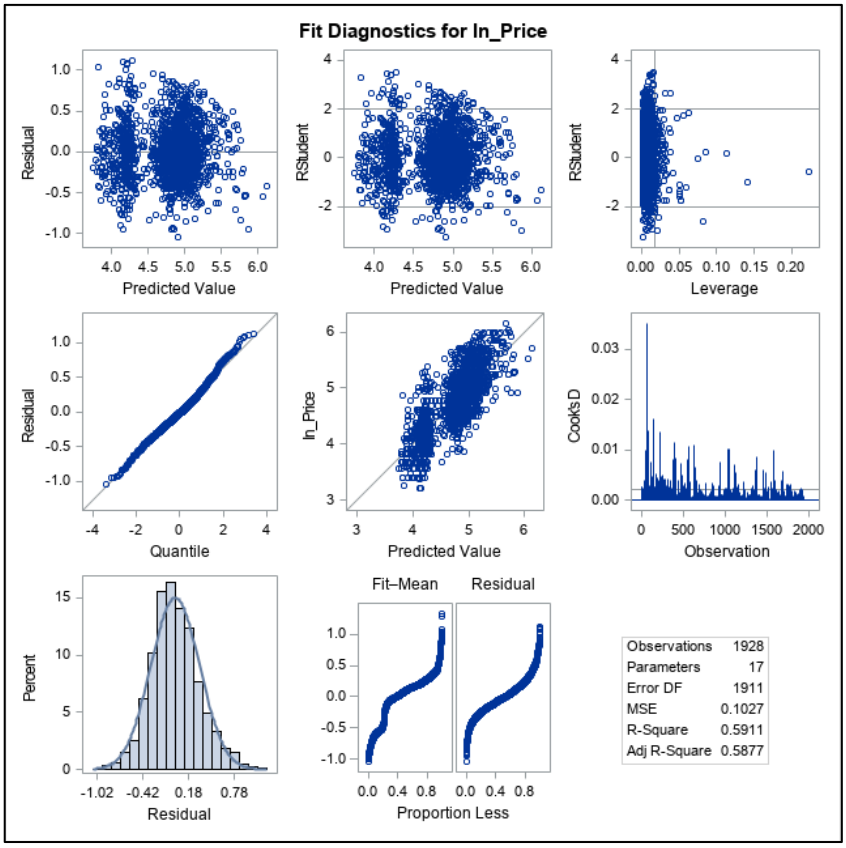


Figure C.36

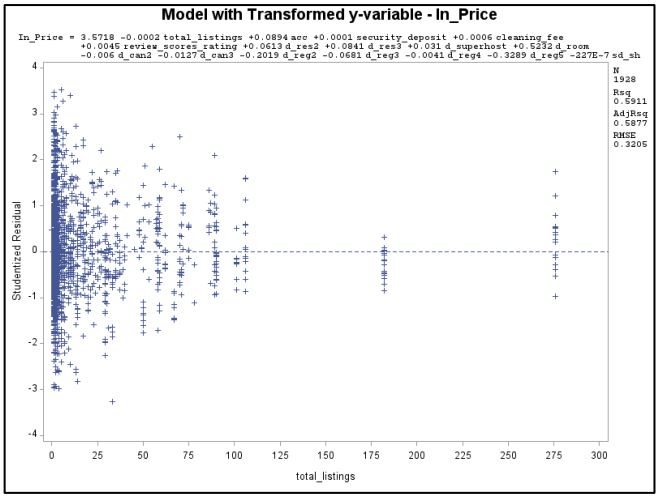


Figure C.37

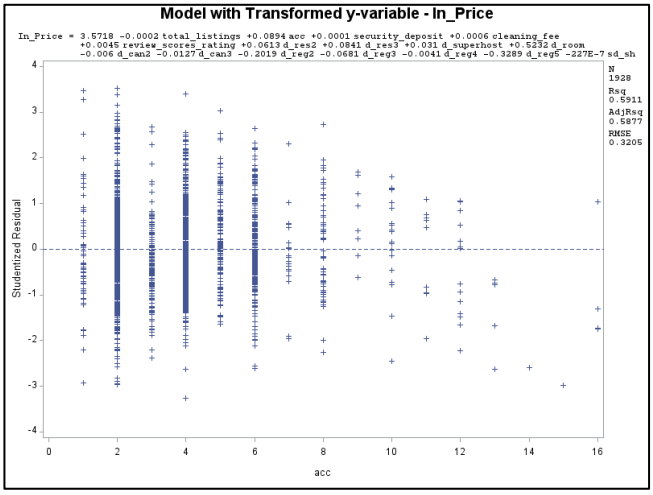


Figure C.38

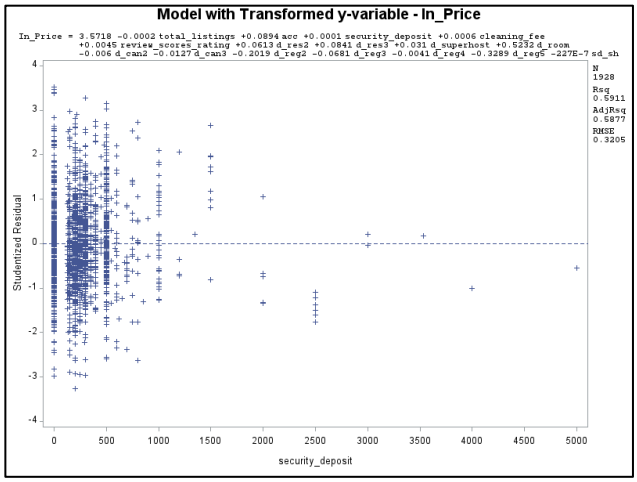


Figure C.39

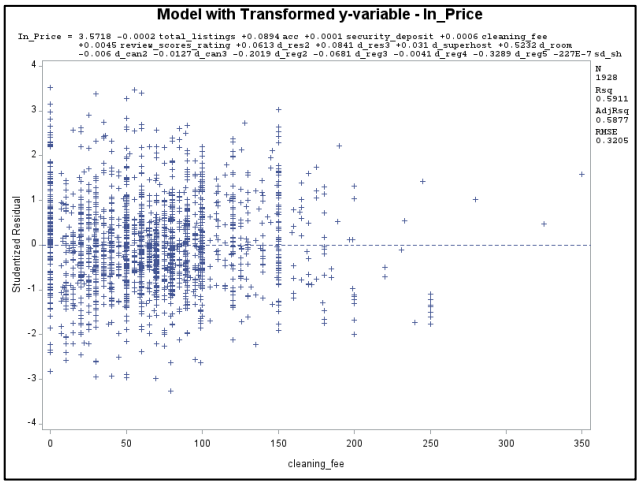


Figure C.40

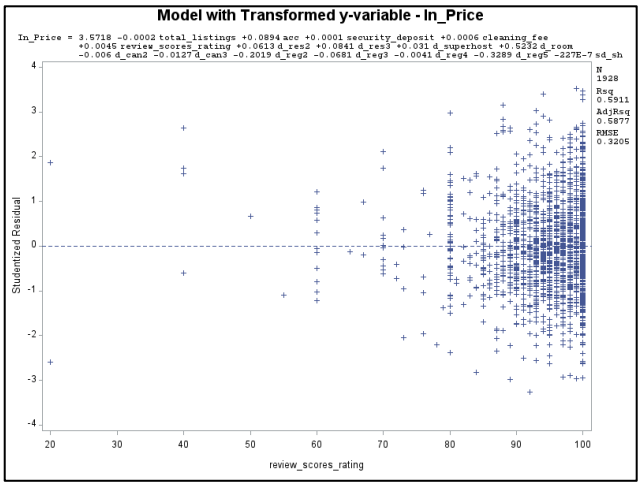


Figure C.41

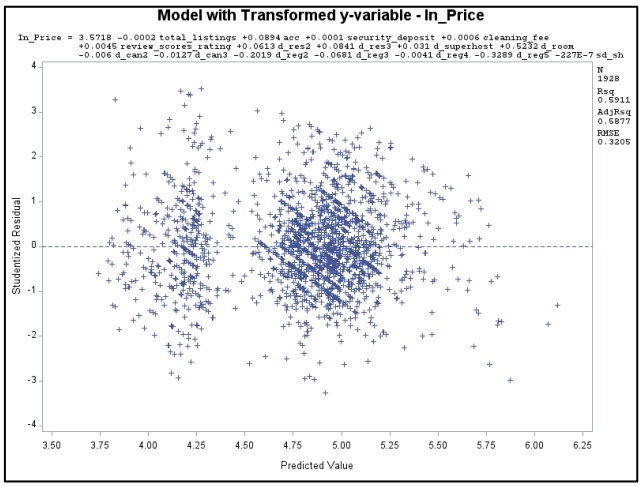


Figure C.42 NPP

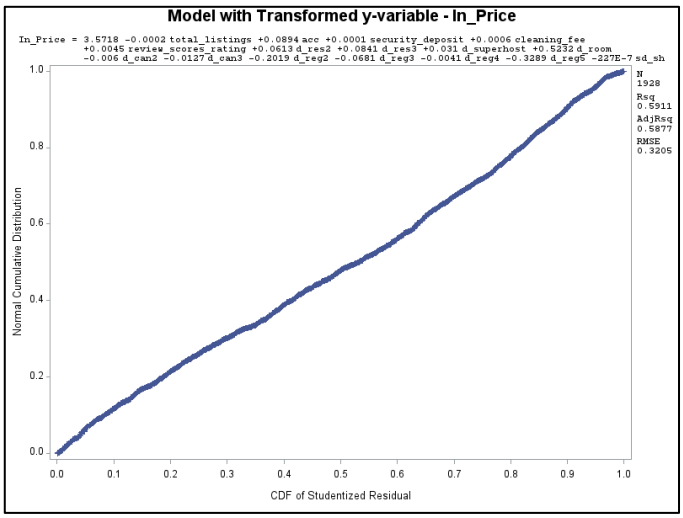


Figure C.43 Histogram for In_Price

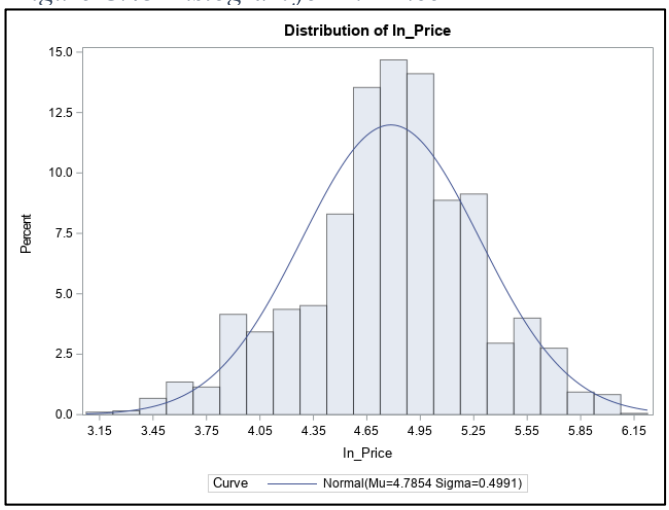


Figure C.44 Scatter plot matrix for In_price

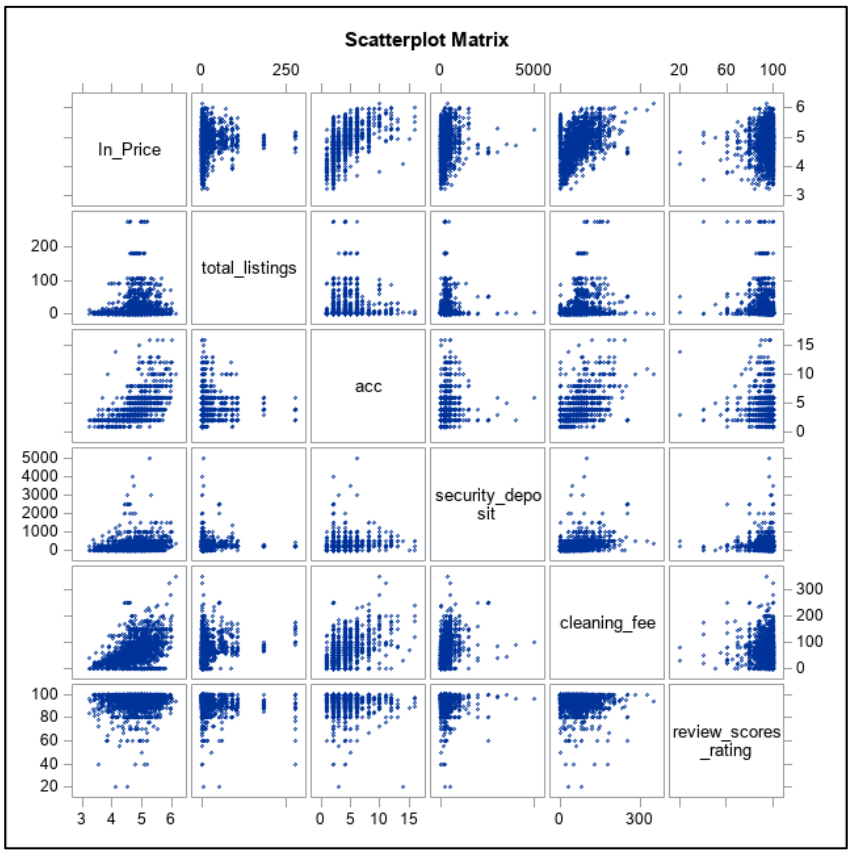


Figure C.45

Model with Transformed y-variable - sqrt_Price

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_Price

Number of Observations Read	1931
Number of Observations Used	1931

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	7996.69776	499.79361	136.81	<.0001
Error	1914	6992.04552	3.65311		
Corrected Total	1930	14989			

Root MSE	1.91131	R-Square	0.5335
Dependent Mean	11.26233	Adj R-Sq	0.5296
Coeff Var	16.97082		

Figure C.46

Model with Transformed x-variables only

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	1931
Number of Observations Used	1931

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	4103744	256484	109.59	<.0001
Error	1914	4479423	2340.34620		
Corrected Total	1930	8583167			

Root MSE	48.37712	R-Square	0.4781
Dependent Mean	134.60228	Adj R-Sq	0.4738
Coeff Var	35.94079		

Figure C.47

Model with Transformed y-variable Price and x variables

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_Price

Number of Observations Read	1931
Number of Observations Used	1931

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	16	7937.14571	496.07161	134.65	<.0001
Error	1914	7051.59757	3.68422		
Corrected Total	1930	14989			

Root MSE	1.91943	R-Square	0.5295
Dependent Mean	11.26233	Adj R-Sq	0.5256
Coeff Var	17.04294		

Figure C.48 Stepwise selection on training set

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	283.49211	25.77201	251.31	<.0001
Error	1916	196.48918	0.10255		
Corrected Total	1927	479.98129			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	3.54502	0.10016	128.47353	1252.77	<.0001
acc	0.08946	0.00398	51.90986	506.18	<.0001
security_deposit	0.00006140	0.00002406	0.66807	6.51	0.0108
cleaning_fee	0.00061519	0.00020689	0.90673	8.84	0.0030
review_scores_rating	0.00474	0.00105	2.07927	20.28	<.0001
d_res2	0.06518	0.02327	0.80487	7.85	0.0051
d_res3	0.08862	0.02924	0.94199	9.19	0.0025
d_superhost	0.02749	0.01610	0.29911	2.92	0.0878
d_room	0.52028	0.02098	63.04133	614.73	<.0001
d_reg2	-0.19833	0.02934	4.68532	45.69	<.0001
d_reg3	-0.06522	0.01948	1.14930	11.21	0.0008
d_reg5	-0.32403	0.03448	9.05627	88.31	<.0001

Figure C.49 Adj r-sq selection on training set

Number in Model	Adjusted R-Square	R-Square	Variables in Model
12	0.5884	0.5910	total_listings acc security_deposit cleaning_fee review_scores_rating d_res2 d_res3 d_superhost d_room d_reg2 d_reg3 d_reg5
11	0.5883	0.5906	acc security_deposit cleaning_fee review_scores_rating d_res2 d_res3 d_superhost d_room d_reg2 d_reg3 d_reg5
13	0.5882	0.5910	total_listings acc security_deposit cleaning_fee review_scores_rating d_res2 d_res3 d_superhost d_room d_can3 d_reg2 d_reg3 d_reg5
13	0.5882	0.5910	total_listings acc security_deposit cleaning_fee review_scores_rating d_res2 d_res3 d_superhost d_room d_reg2 d_reg3 d_reg5 sd_sh

Figure C.50 Fitted Model on Training Set

Final Model - Training set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	1931
Number of Observations Used	1446
Number of Observations with Missing Values	485

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	212.13755	19.28523	195.79	<.0001
Error	1434	141.24877	0.09850		
Corrected Total	1445	353.38632			

Root MSE	0.31385	R-Square	0.6003
Dependent Mean	4.78514	Adj R-Sq	0.5972
Coeff Var	6.55878		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	3.64221	0.11689	31.16	<.0001	0	.	0
acc	1	0.08993	0.00466	19.28	<.0001	0.38402	0.70272	1.42305
security_deposit	1	0.00006813	0.00002685	2.54	0.0113	0.04670	0.82260	1.21565
cleaning_fee	1	0.00052516	0.00023788	2.21	0.0274	0.04777	0.59525	1.67998
review_scores_rating	1	0.00332	0.00123	2.71	0.0068	0.04738	0.91109	1.09759
d_res2	1	0.07025	0.02656	2.64	0.0083	0.04542	0.94528	1.05789
d_res3	1	0.11427	0.03346	3.41	0.0007	0.05838	0.95381	1.04843
d_superhost	1	0.03937	0.01826	2.16	0.0313	0.03805	0.89464	1.11777
d_room	1	0.55067	0.02401	22.93	<.0001	0.45552	0.70656	1.41530
d_reg2	1	-0.21596	0.03349	-6.45	<.0001	-0.10981	0.96133	1.04023
d_reg3	1	-0.04873	0.02186	-2.23	0.0260	-0.03808	0.95513	1.04698
d_reg5	1	-0.29936	0.03988	-7.51	<.0001	-0.12908	0.94235	1.06118

Figure C.51 Fit Diagnostics for new_y on training set

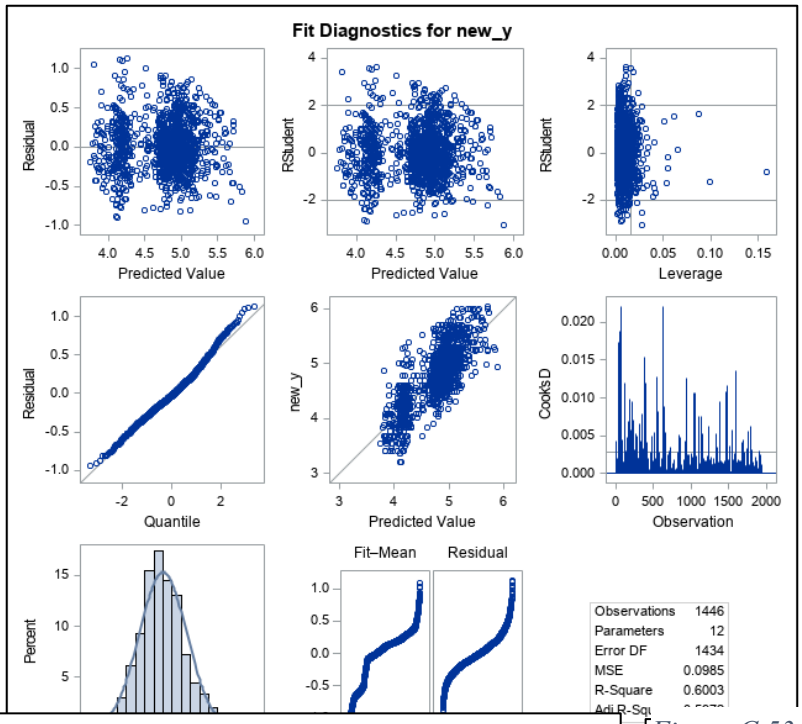


Figure C.52

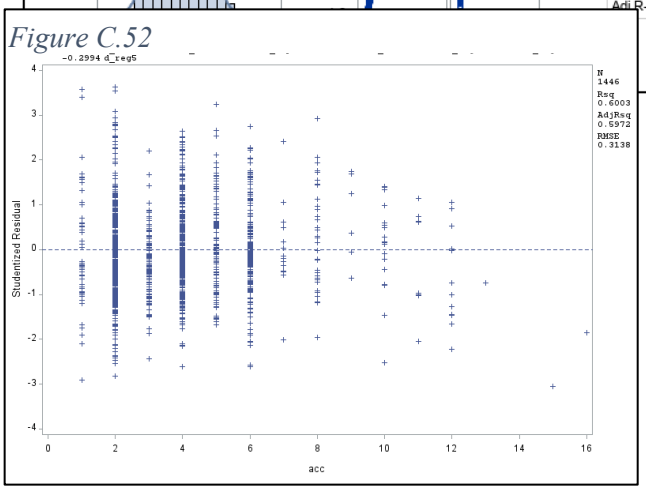


Figure C.53

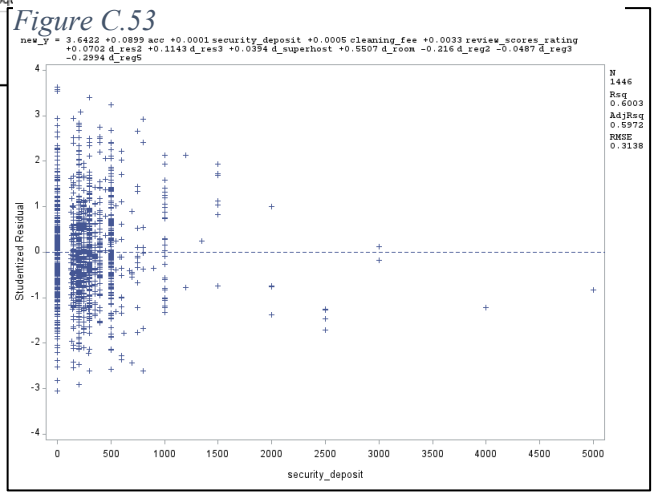


Figure C.54

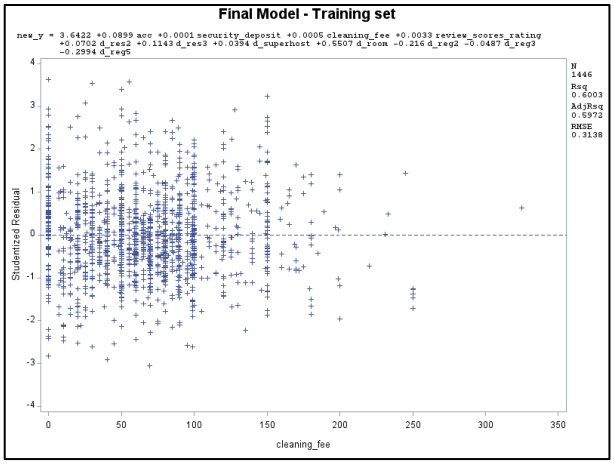


Figure C.55

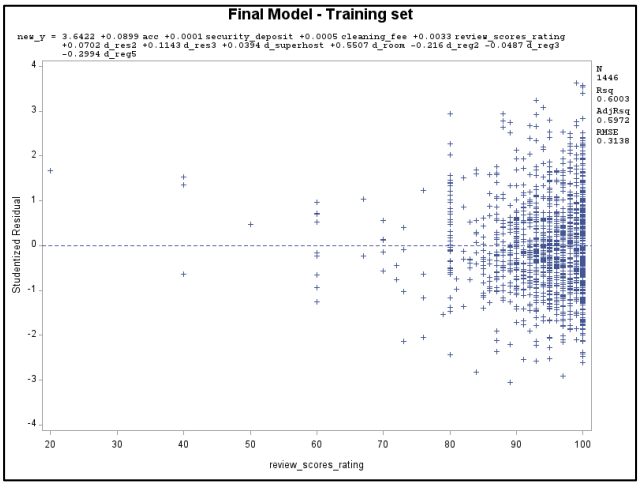


Figure C.56

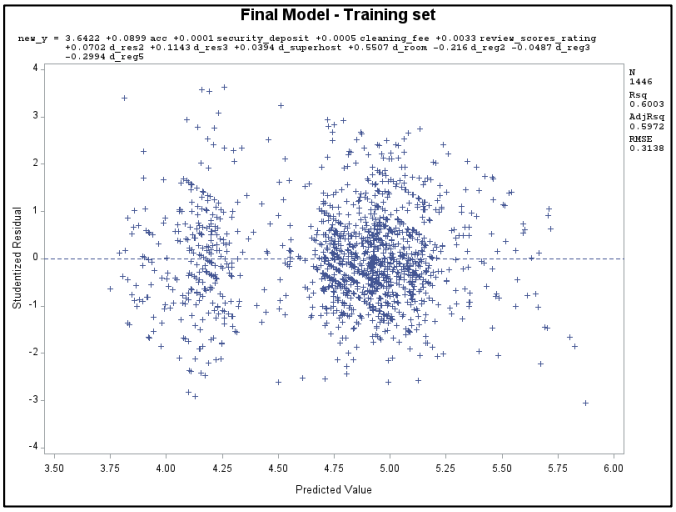


Figure C.57 NPP for training set

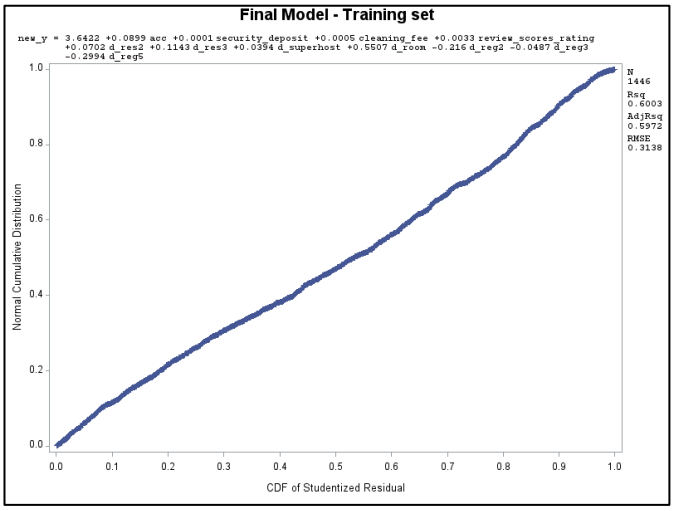


Figure C.58 Validation statistics for test set

Validation statistics for model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	485	0.34030	0.26636

Validation statistics for model

The CORR Procedure

2 Variables:	ln_Price yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ln_Price	482	4.78604	0.51302	2307	3.25810	6.15486	
yhat	485	4.76352	0.41592	2310	3.77769	6.11888	Predicted Value of new_y

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	ln_Price	yhat
ln_Price	1.00000	0.75090
		<.0001
	482	482
yhat	0.75090	1.00000
Predicted Value of new_y	<.0001	
	482	485

Figure C.59 Predictions

Compute predictions with fitted model with confidence & prediction interval

The REG Procedure
Model: MODEL1
Dependent Variable: ln_Price

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	4.1175	0.0625	3.9947	4.2404	3.4454	4.7897	.
2	.	5.0536	0.0547	4.9461	5.1610	4.3841	5.7231	.
3	4.01	4.1038	0.0654	3.9753	4.2324	3.4306	4.7771	-0.0965
4	4.38	4.0636	0.0699	3.9262	4.2010	3.3886	4.7386	0.3184
5	4.01	4.3741	0.0566	4.2630	4.4853	3.7040	5.0443	-0.3668
6	3.56	3.7813	0.0718	3.6403	3.9223	3.1056	4.4570	-0.2259

Figure C.60 adding new observations and joining with original dataset

compute predictions

Obs	acc	security_deposit	cleaning_fee	review_scores_rating	d_res2	d_res3	d_superhost	d_room	d_reg2	d_reg3	d_reg5
1	2	300	0	96	0	0	1	0	1	0	0
2	4	500	50	100	1	0	0	1	0	0	0

join new dataset with Airbnb datset

Obs	acc	security_deposit	cleaning_fee	review_scores_rating	d_res2	d_res3	d_superhost	d_room	d_reg2	d_reg3	d_reg5	Selected	res_time	superhost	total_listings	room_type	price	can_policy	region	d_can2	d_can3	d_reg4	sd_sh	ln_Price	new_y	yhat
1	2	300	0	96	0	0	1	0	1	0	0	.			.		.			.	.	.	.	.	.	.
2	4	500	50	100	1	0	0	1	0	0	0	.			.		.			.	.	.	.	.	.	.
3	2	700	0	70	0	0	0	0	0	0	0	0	anhour	f	1	P_room	55	Strict	IM	0	1	0	0	4.00733		4.10212
4	4	150	10	94	0	0	0	0	0	0	1	0	anhour	f	4	P_room	80	flexible	WM	0	0	0	0	4.38203		4.03006
5	3	150	35	84	1	0	0	0	0	0	0	0	fewhours	f	2	P_room	55	Strict	IM	0	1	0	0	4.00733		4.28968
6	3	0	0	93	0	0	0	0	0	0	1	0	anhour	f	2	P_room	35	flexible	WM	0	0	0	0	3.55535		3.78933

CODY – APPENDIX D

Fig (D.1) Price Histogram

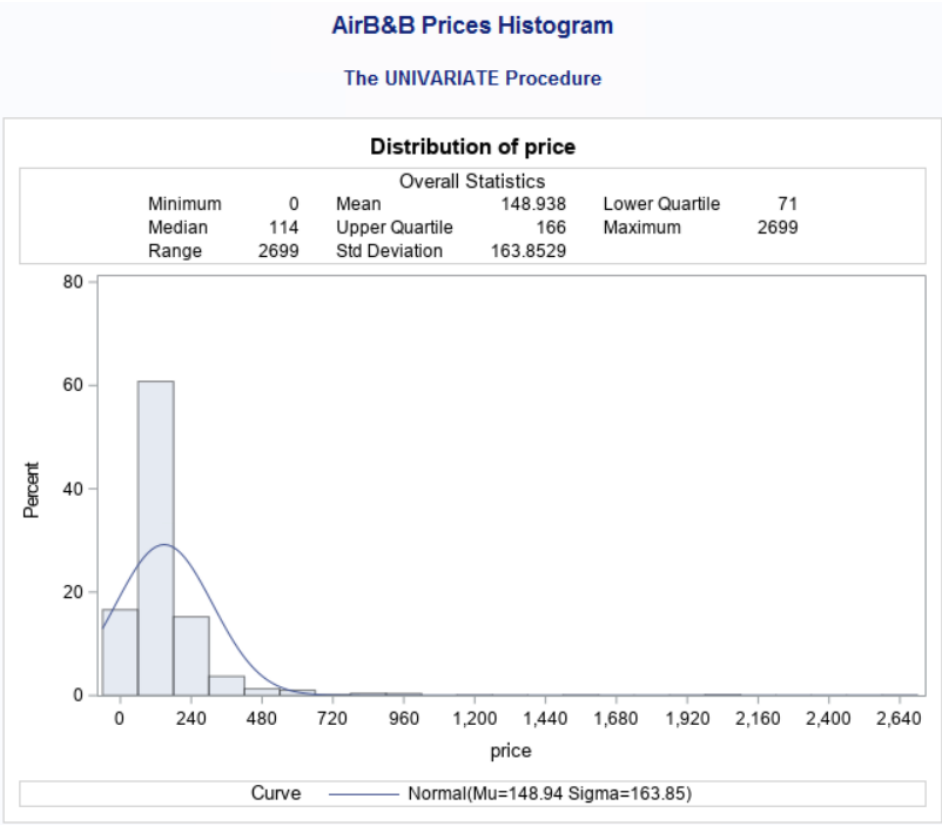


Fig (D.2)- Price Probability Plot

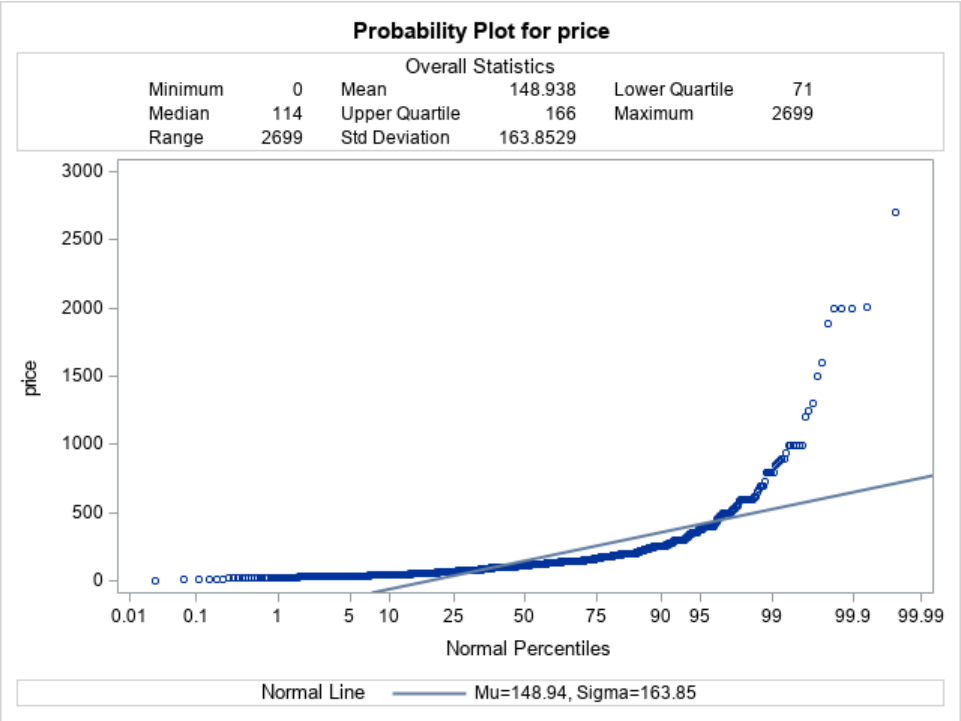


Fig (D.3)- logPrice Histogram

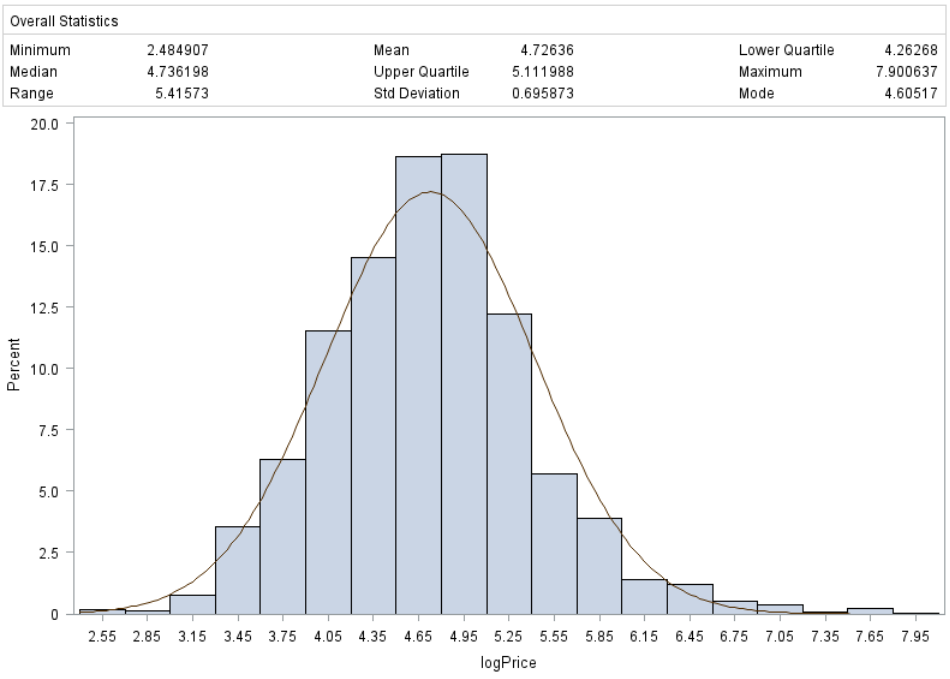


Fig (D.4) Full Model Regression Probability Plot for logPrice

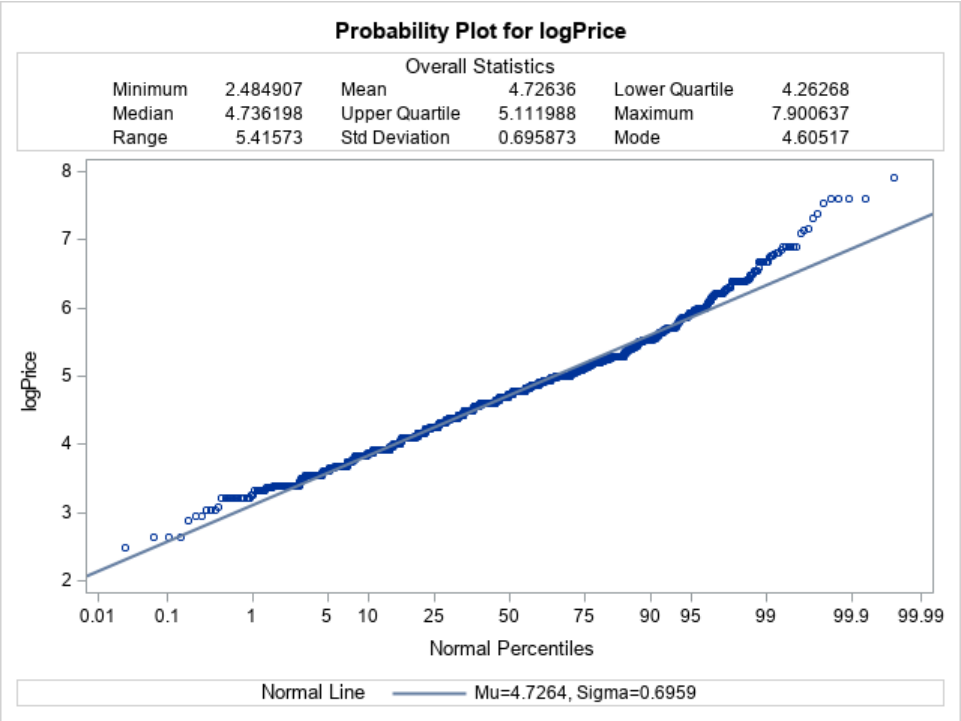


Fig (D.5) Price and SuperHost Boxplot

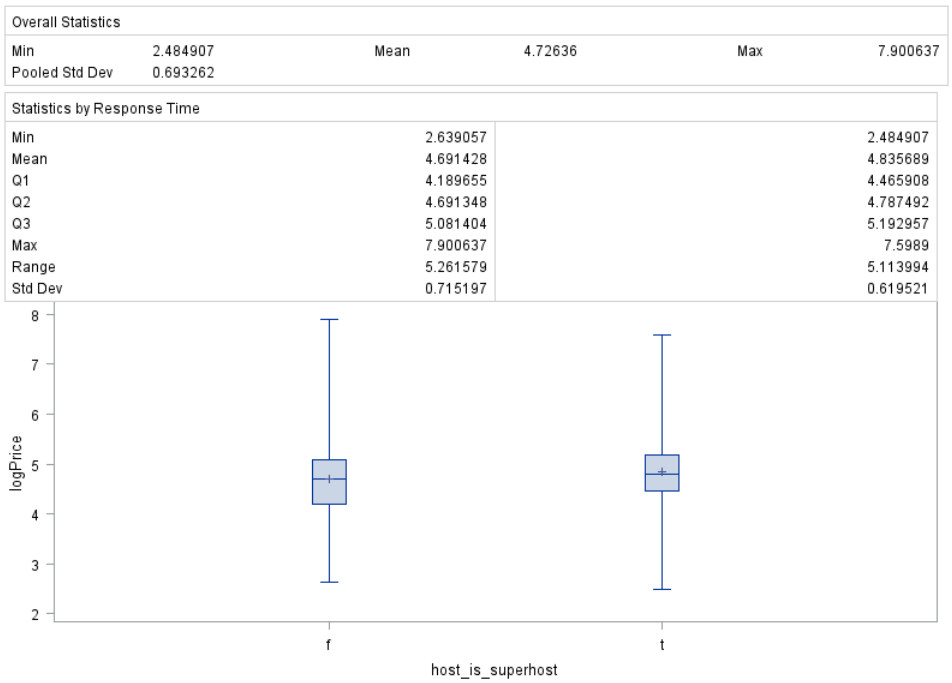


Fig (D.6) Price and Room Type Boxplot

Boxplots - Price and Room Type

Overall Statistics					
Min	2.484907	Mean	4.72636	Max	7.900637
Pooled Std Dev	0.540339				

Statistics by Response Time			
Min	3.688879	2.484907	2.639057
Mean	5.05771	4.181008	3.678503
Q1	4.691348	3.871201	3.332205
Q2	4.976734	4.127134	3.555348
Q3	5.293305	4.394373	4.094345
Max	7.900637	7.377759	5.293305
Range	4.211757	4.892852	2.654247
Std Dev	0.546898	0.526083	0.585169

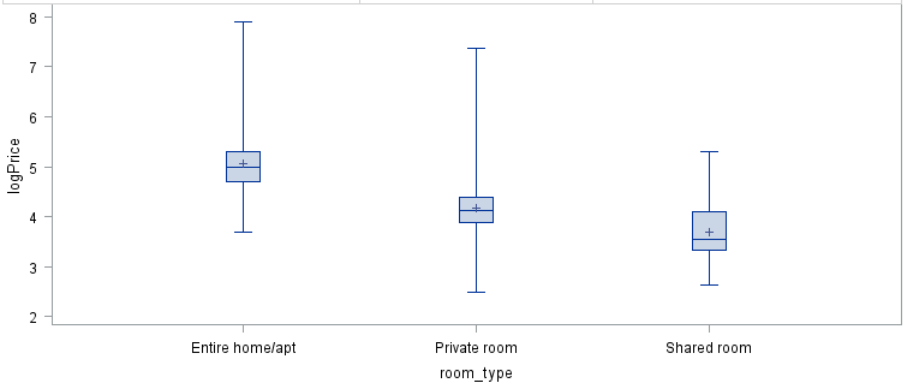


Fig (D.7) Price and Cancellation Boxplot

Boxplots - Price and Cancellation Policy

Overall Statistics					
Min	2.484907	Mean	4.72636	Max	7.900637
Pooled Std Dev	0.674579				

Statistics by Response Time			
Min	2.639057	2.484907	3.218876
Mean	4.908296	4.505144	4.725601
Q1	4.584967	3.988984	4.317488
Q2	4.867534	4.430817	4.744932
Q3	5.247024	5.010635	5.068904
Max	7.600902	7.900637	7.313887
Range	4.961845	5.41573	4.095011
Std Dev	0.664603	0.728191	0.61786

Fig (D.8) Price and Bed Type Boxplot

Boxplots - Price and Bed Type

Overall Statistics					
Min	2.484907	Mean	4.72636	Max	7.900637
Pooled Std Dev	0.693947				

Statistics by Response Time					
Min	3.401197	4.26268	3.871201	2.639057	2.484907
Mean	3.622856	4.26268	4.496016	3.999981	4.731001
Q1	3.401197	4.26268	3.912023	3.713275	4.26268
Q2	3.555348	4.26268	4.473187	4.075667	4.744932
Q3	3.912023	4.26268	5.0166	4.536687	5.123964
Max	3.912023	4.26268	5.293305	4.70953	7.900637
Range	0.510826	.	1.422104	2.070473	5.41573
Std Dev	0.262019	.	0.571189	0.681033	0.694543

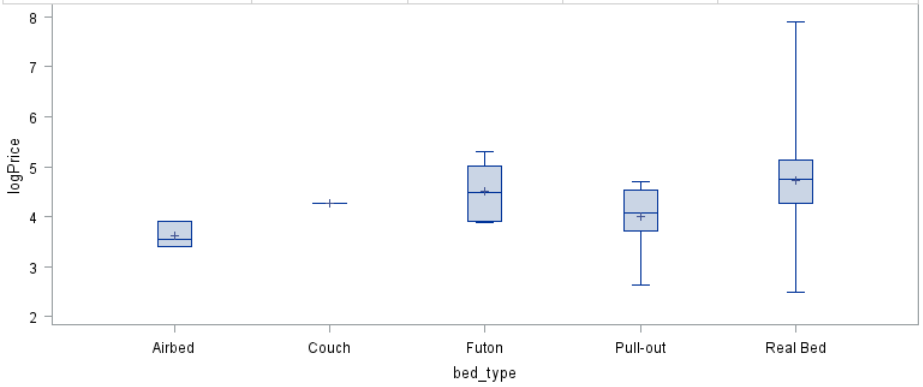


Fig (D.9) Price and Response Time Boxplot

Boxplots - Price and Response Time

Overall Statistics					
Min	2.484907	Mean	4.72636	Max	7.900637
Pooled Std Dev	0.691178				

Statistics by Response Time					
Min	2.639057	2.944439	3.044522	2.890372	2.484907
Mean	4.616003	4.59685	4.738897	4.719343	4.803317
Q1	4.094345	4.007333	4.174387	4.174387	4.454347
Q2	4.564348	4.459727	4.691348	4.60517	4.828314
Q3	5.068904	5.259526	5.192957	5.209352	5.129899
Max	7.900637	6.214608	7.600902	7.5989	7.600902
Range	5.261579	3.270169	4.55638	4.708529	5.115996
Std Dev	0.742665	0.827758	0.814855	0.759461	0.618164

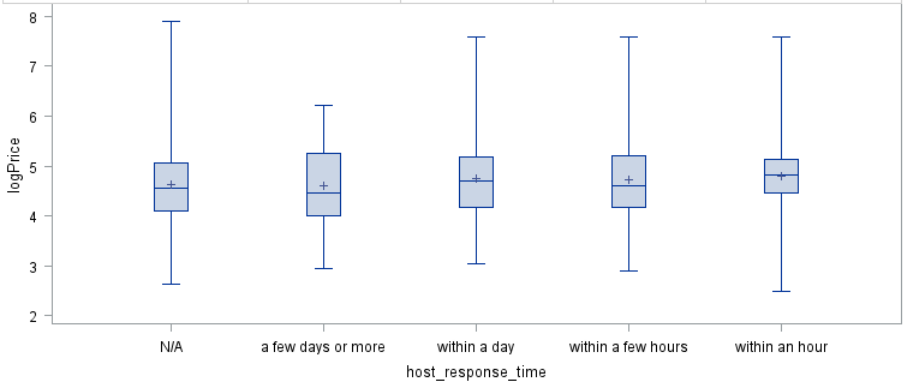


Fig (D.10) Frequency Tables

Frequency of qual var

The FREQ Procedure

host_response_time	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N/A	827	33.31	827	33.31
a few days or more	36	1.45	863	34.76
within a day	153	6.16	1016	40.92
within a few hours	224	9.02	1240	49.94
within an hour	1243	50.06	2483	100.00

host_is_superhost	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	1882	75.80	1882	75.80
t	601	24.20	2483	100.00

room_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Entire home/apt	1569	63.19	1569	63.19
Private room	872	35.12	2441	98.31
Shared room	42	1.69	2483	100.00

bed_type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Airbed	3	0.12	3	0.12
Couch	1	0.04	4	0.16
Futon	8	0.32	12	0.48
Pull-out	8	0.32	20	0.81
Real Bed	2463	99.19	2483	100.00

cancellation_policy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Strict	1002	40.35	1002	40.35
flexible	821	33.06	1823	73.42
moderate	660	26.58	2483	100.00

Region	Frequency	Percent	Cumulative Frequency	Cumulative Percent
EM	254	10.23	254	10.23
IM	1320	53.16	1574	63.39
NSM	240	9.67	1814	73.06
SEM	539	21.71	2353	94.76
WM	130	5.24	2483	100.00

Fig (D.11) Scatterplot Matrix

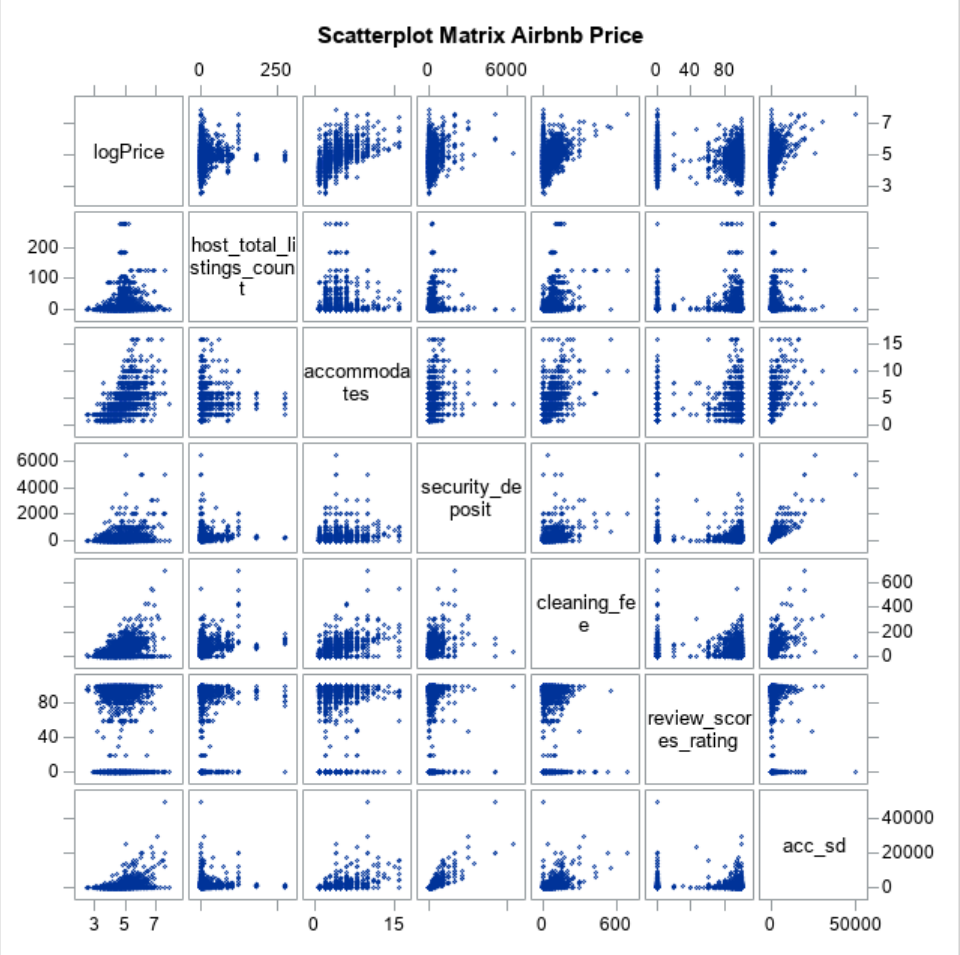


Fig (D.11) Full Regression Model

Regression Full

The REG Procedure
Model: MODEL1
Dependent Variable: logPrice

Number of Observations Read	2483
Number of Observations Used	2482
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	667.09748	29.00424	133.43	<.0001
Error	2458	534.29917	0.21737		
Corrected Total	2481	1201.39665			

Root MSE	0.46623	R-Square	0.5553
Dependent Mean	4.72636	Adj R-Sq	0.5511
Coeff Var	9.86449		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	4.54267	0.06071	74.82	<.0001	.	0
host_total_listings_count	1	-0.00006422	0.00037578	-0.17	0.8643	0.82221	1.21623
accommodates	1	0.11341	0.00653	17.38	<.0001	0.42513	2.35220
security_deposit	1	0.00005730	0.00005443	1.05	0.2925	0.20654	4.84164
cleaning_fee	1	0.00014904	0.00023380	0.64	0.5239	0.51458	1.94332
review_scores_rating	1	-0.00178	0.00024935	-7.14	<.0001	0.80081	1.24874
dResp1	1	-0.08330	0.02423	-3.44	0.0006	0.59667	1.67596
dResp2	1	0.02069	0.03630	0.57	0.5687	0.80947	1.23538
dResp3	1	-0.00890	0.04180	-0.21	0.8314	0.86673	1.15376
dResp4	1	0.12063	0.07960	1.52	0.1298	0.96696	1.03417
dSuper	1	-0.06004	0.02402	-2.50	0.0125	0.82722	1.20887
dRoom1	1	-0.56145	0.02560	-21.93	<.0001	0.58624	1.70580
dRoom2	1	-1.06659	0.07647	-13.95	<.0001	0.90026	1.11079
dBed1	1	-0.73506	0.46717	-1.57	0.1157	0.99640	1.00361
dBed2	1	-0.24269	0.16683	-1.45	0.1459	0.97937	1.02107
dBed3	1	-0.07064	0.16572	-0.43	0.6699	0.99258	1.00748
dBed4	1	-0.62216	0.27034	-2.30	0.0215	0.99262	1.00744
dCan1	1	-0.00320	0.02568	-0.12	0.9008	0.55191	1.81188
dCan2	1	-0.00274	0.02648	-0.10	0.9177	0.63973	1.56317
dReg1	1	0.23545	0.04413	5.34	<.0001	0.18060	5.53718
dReg2	1	0.14641	0.04612	3.17	0.0015	0.24215	4.12962
dReg3	1	0.04504	0.05118	0.88	0.3789	0.38272	2.61287
dReg4	1	0.15735	0.05071	3.10	0.0019	0.37206	2.68777
acc_sd	1	0.00002664	0.00000960	2.78	0.0055	0.17388	5.75104

Fig (D.13) Fitting Model - Stepwise Method

Variable dResp4 Entered: R-Square = 0.6486 and C(p) = 18.4818

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	464.30741	42.20976	281.70	<.0001
Error	1679	251.57680	0.14984		
Corrected Total	1690	715.88421			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	4.50773	0.04638	1415.61746	9447.70	<.0001
accommodates	0.11932	0.00565	66.76093	445.56	<.0001
review_scores_rating	-0.00142	0.00024321	5.11375	34.13	<.0001
dResp1	-0.06085	0.02081	1.28156	8.55	0.0035
dResp4	0.18641	0.08225	0.76968	5.14	0.0236
dSuper	-0.07978	0.02362	1.70972	11.41	0.0007
dRoom1	-0.53856	0.02411	74.76075	498.95	<.0001
dRoom2	-1.06721	0.07539	30.02290	200.37	<.0001
dReg1	0.18369	0.02798	6.45694	43.09	<.0001
dReg2	0.11390	0.03166	1.93877	12.94	0.0003
dReg4	0.14775	0.03862	2.19305	14.64	0.0001
acc_sd	0.00003902	0.00000461	10.74382	71.70	<.0001

Fig (D.14) Fitting Model – Forward Method

Variable dResp4 Entered: R-Square = 0.6486 and C(p) = 18.4818

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	464.30741	42.20976	281.70	<.0001
Error	1679	251.57680	0.14984		
Corrected Total	1690	715.88421			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	4.50773	0.04638	1415.61746	9447.70	<.0001
accommodates	0.11932	0.00565	66.76093	445.56	<.0001
review_scores_rating	-0.00142	0.00024321	5.11375	34.13	<.0001
dResp1	-0.06085	0.02081	1.28156	8.55	0.0035
dResp4	0.18641	0.08225	0.76968	5.14	0.0236
dSuper	-0.07978	0.02362	1.70972	11.41	0.0007
dRoom1	-0.53856	0.02411	74.76075	498.95	<.0001
dRoom2	-1.06721	0.07539	30.02290	200.37	<.0001
dReg1	0.18369	0.02798	6.45694	43.09	<.0001
dReg2	0.11390	0.03166	1.93877	12.94	0.0003
dReg4	0.14775	0.03862	2.19305	14.64	0.0001
acc_sd	0.00003902	0.00000461	10.74382	71.70	<.0001

Fig (D.15) Validation Test Set

Validation - Test Set

The REG Procedure
Model: MODEL1
Dependent Variable: train_y

Number of Observations Read	2416
Number of Observations Used	1691
Number of Observations with Missing Values	725

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	464.30741	42.20976	281.70	<.0001
Error	1679	251.57680	0.14984		
Corrected Total	1690	715.88421			

Root MSE	0.38709	R-Square	0.6486
Dependent Mean	4.69182	Adj R-Sq	0.6463
Coeff Var	8.25027		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.50773	0.04638	97.20	<.0001
accommodates	1	0.11932	0.00565	21.11	<.0001
review_scores_rating	1	-0.00142	0.00024321	-5.84	<.0001
dResp1	1	-0.06085	0.02081	-2.92	0.0035
dResp4	1	0.18641	0.08225	2.27	0.0236
dSuper	1	-0.07978	0.02362	-3.38	0.0007
dRoom1	1	-0.53856	0.02411	-22.34	<.0001
dRoom2	1	-1.06721	0.07539	-14.16	<.0001
dReg1	1	0.18369	0.02798	6.56	<.0001
dReg2	1	0.11390	0.03166	3.60	0.0003
dReg4	1	0.14775	0.03862	3.83	0.0001
acc_sd	1	0.00003902	0.00000461	8.47	<.0001

Fig (D.16) Validation Stats (11 predictors) – RMSE, MAE, and CV-R^2 for Test Set

Validation Stats for Model 1

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	725	0.39544	0.31172

Validation Stats for Model 1

The CORR Procedure

2 Variables:	logPrice yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
logPrice	724	4.70030	0.64072	3403	3.04452	6.90675	
yhat	725	4.70646	0.49586	3412	3.47998	6.64067	Predicted Value of train_y

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	logPrice	yhat
logPrice	1.00000	0.78691
		<.0001
	724	724
yhat	0.78691	1.00000
Predicted Value of train_y	<.0001	
	724	725

Fig (D.17) Training 5-Fold Cross-Validation

Effects:	Intercept accommodates review_scores_rating dResp1 dResp4 dSuper dRoom1 dRoom2 dReg1 dReg2 dReg4 acc_sd

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	11	648.87302	58.98846	389.58
Error	2403	363.85328	0.15142	
Corrected Total	2414	1012.72630		

Root MSE	0.38912
Dependent Mean	4.69436
R-Square	0.6407
Adj R-Sq	0.6391
AIC	-2129.87982
AICC	-2129.72822
SBC	-4477.40637
CV PRESS	369.28565

Cross Validation Details

Index	Observations		CV PRESS
	Fitted	Left Out	
1	1932	483	77.6604
2	1932	483	69.3710
3	1932	483	74.8324
4	1932	483	74.6781
5	1932	483	72.7438
Total			369.2857

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Cross Validation Estimates				
					1	2	3	4	5
Intercept	1	4.539395	0.039501	114.92	4.5337006	4.5254666	4.5374041	4.5702581	4.5306532
accommodates	1	0.114150	0.004850	23.54	0.1114412	0.1188234	0.1150007	0.1100877	0.1149294
review_scores_rating	1	-0.001292	0.000205	-6.29	-0.0012524	-0.0015333	-0.0013150	-0.0011793	-0.0011728
dResp1	1	-0.075573	0.017321	-4.36	-0.0747006	-0.0662827	-0.0795894	-0.0844970	-0.0732345
dResp4	1	0.176912	0.066072	2.68	0.0823292	0.1926341	0.2404602	0.2220765	0.1522651
dSuper	1	-0.095591	0.019820	-4.82	-0.0889082	-0.0879032	-0.0952697	-0.1005857	-0.1051297
dRoom1	1	-0.563554	0.020500	-27.49	-0.5608511	-0.5444110	-0.5765952	-0.5820285	-0.5545567
dRoom2	1	-1.106625	0.064349	-17.20	-1.1133849	-1.0466156	-1.0985503	-1.1562834	-1.1127354
dReg1	1	0.192864	0.024045	8.02	0.1894182	0.2048615	0.1961102	0.1813699	0.1936312
dReg2	1	0.114836	0.026946	4.26	0.1197253	0.1261279	0.1323375	0.0844453	0.1110399
dReg4	1	0.134237	0.032470	4.13	0.1280903	0.1555757	0.1396409	0.1240830	0.1244566
acc_sd	1	0.000039227	0.000004017	9.77	0.0000399	0.0000386	0.0000397	0.0000408	0.0000378

Fig (D.18) Testing 5-Fold Cross-Validation (Stepwise)

5-fold crossvalidation +30% testing set (stepwise)

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 10).

Effects:	Intercept accommodates review_scores_rating dResp1 dSuper dRoom1 dRoom2 dReg1 dReg2 dReg4 acc_sd

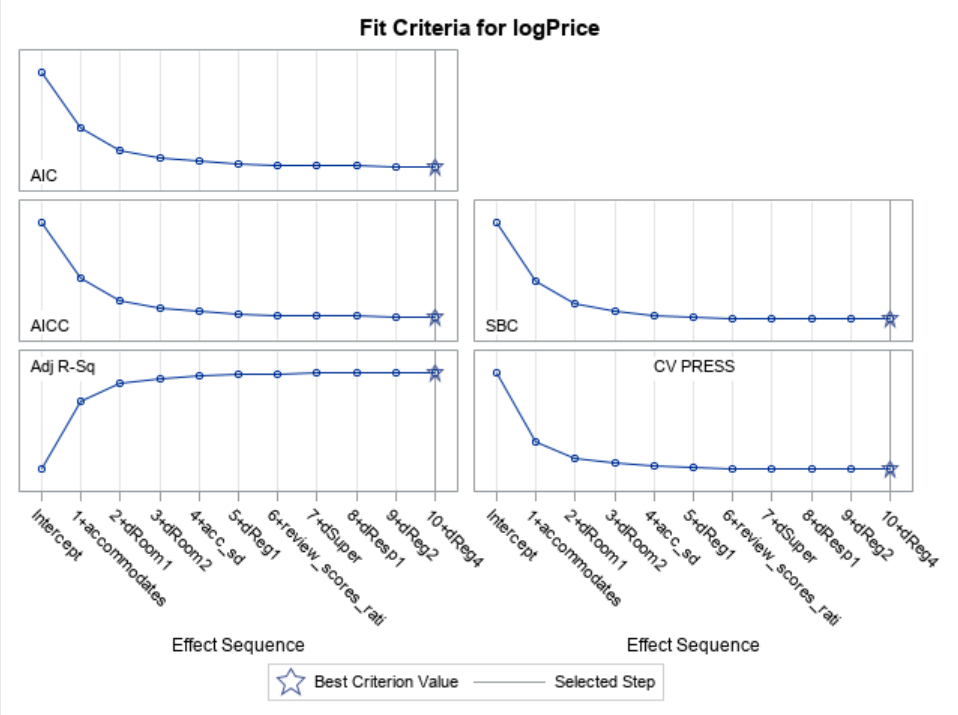
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	10	462.21651	46.22165	302.60
Error	1671	255.23856	0.15275	
Corrected Total	1681	717.45508		

Root MSE	0.39083
Dependent Mean	4.70032
R-Square	0.6442
Adj R-Sq	0.6421
AIC	-1465.47860
AICC	-1465.29167
SBC	-3089.77348
ASE (Train)	0.15175
ASE (Test)	0.15059
CV PRESS	259.80262

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1345	337	48.0827
2	1345	337	58.7788
3	1346	336	51.1116
4	1346	336	51.9041
5	1346	336	49.9255
Total			259.8026

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	4.516392	0.047831	94.42	4.5108952	4.4956879	4.4955401	4.5197050	4.558095
accommodates	1	0.117689	0.005860	20.08	0.1201250	0.1132220	0.1194328	0.1188950	0.116001
review_scores_rating	1	-0.001376	0.000247	-5.57	-0.0013724	-0.0013640	-0.0011708	-0.0014855	-0.001464
dResp1	1	-0.069307	0.020715	-3.35	-0.0781639	-0.0517864	-0.0668389	-0.0801920	-0.068739
dSuper	1	-0.087479	0.024082	-3.63	-0.0885320	-0.0927156	-0.0729199	-0.0829716	-0.099313
dRoom1	1	-0.556493	0.024497	-22.72	-0.5405823	-0.5338444	-0.5674323	-0.5739005	-0.565954
dRoom2	1	-1.116718	0.084649	-13.19	-1.0917578	-1.0724266	-1.1981157	-1.0374185	-1.174276
dReg1	1	0.217333	0.029011	7.49	0.2159521	0.2303758	0.2193353	0.2193863	0.201035
dReg2	1	0.124708	0.032037	3.89	0.1266269	0.1282109	0.1299211	0.1369978	0.102413
dReg4	1	0.115467	0.039244	2.94	0.1263585	0.1278674	0.1448658	0.1049279	0.076348
acc_sd	1	0.000037408	0.000004623	8.09	0.0000396	0.0000414	0.0000316	0.0000376	0.000038



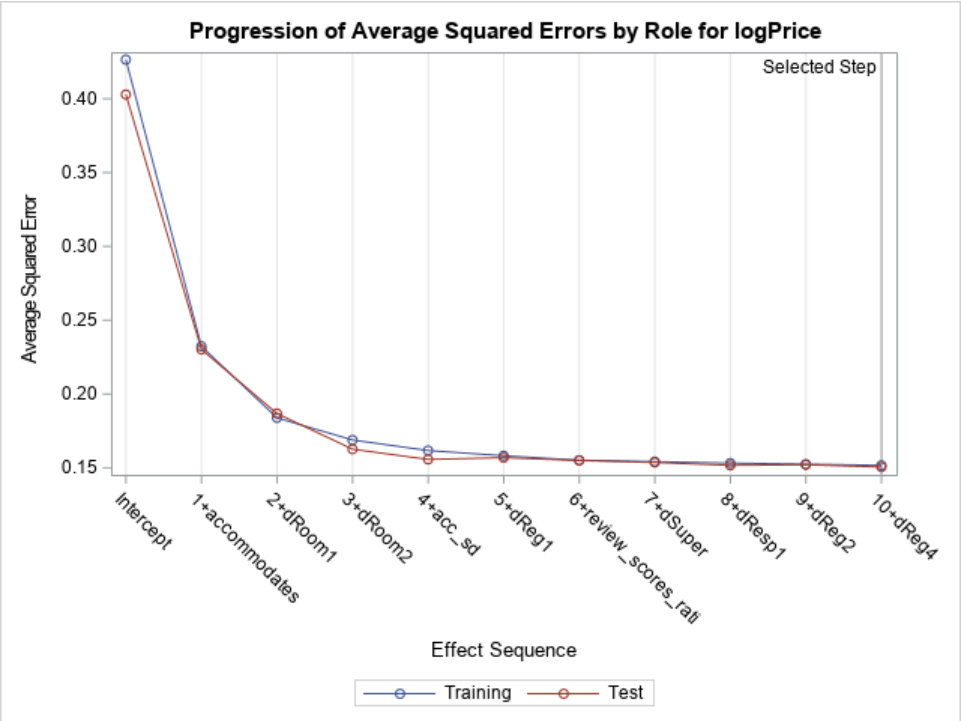

Progression of Average Squared Errors by Role for logPrice
Selected Step
Average Squared Error
0.40
0.35
0.30
0.25
0.20
0.15
Intercept
1+accommodates
2+dRoom1
3+dRoom2
4+acc_sd
5+dReg1
6+review_scores_rati
7+dSuper
8+dResp1
9+dReg2
10+dReg4
Effect Sequence
Training
Test

Fig (D.19) Final Model

Final Model

The REG Procedure
Model: MODEL1
Dependent Variable: logPrice

Number of Observations Read	2416
Number of Observations Used	2415
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	648.87302	58.98846	389.58	<.0001
Error	2403	363.85328	0.15142		
Corrected Total	2414	1012.72630			

Root MSE	0.38912	R-Square	0.6407
Dependent Mean	4.69436	Adj R-Sq	0.6391
Coeff Var	8.28914		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	4.53940	0.03950	114.92	<.0001	0	0
accommodates	1	0.11415	0.00485	23.54	<.0001	0.38116	1.75416
review_scores_rating	1	-0.00129	0.00020549	-6.29	<.0001	-0.08297	1.16508
dResp1	1	-0.07557	0.01732	-4.36	<.0001	-0.05835	1.19622
dResp4	1	0.17691	0.06607	2.68	0.0075	0.03311	1.02247
dSuper	1	-0.09559	0.01982	-4.82	<.0001	-0.06350	1.15940
dRoom1	1	-0.56355	0.02050	-27.49	<.0001	-0.41506	1.52468
dRoom2	1	-1.10662	0.06435	-17.20	<.0001	-0.21810	1.07577
dReg1	1	0.19286	0.02404	8.02	<.0001	0.14857	2.29455
dReg2	1	0.11484	0.02695	4.26	<.0001	0.07309	1.96754
dReg4	1	0.13424	0.03247	4.13	<.0001	0.06259	1.53286
acc_sd	1	0.00003923	0.00000402	9.77	<.0001	0.13795	1.33465

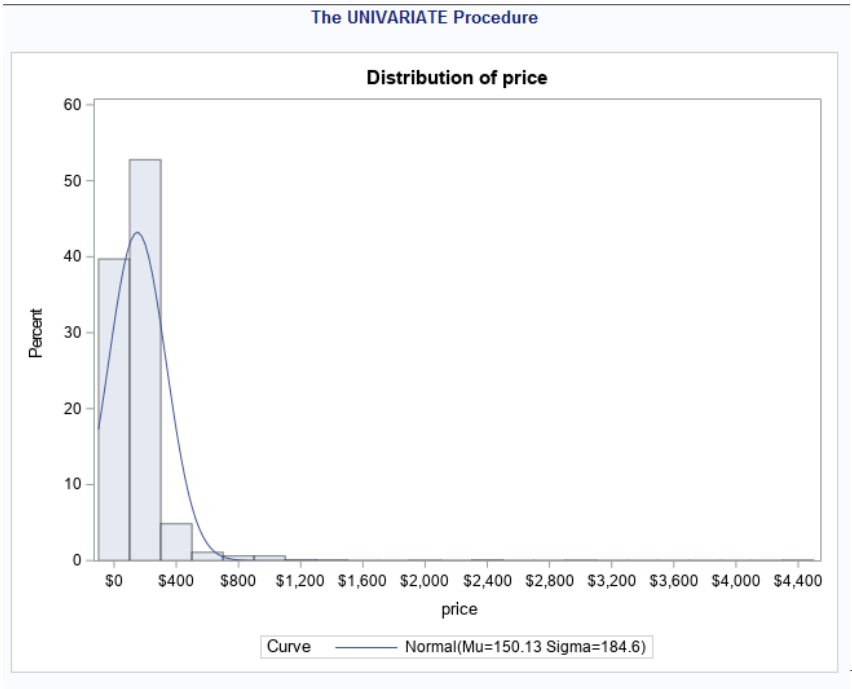
Fig (D.20) Predictions

Predictions for Final Model

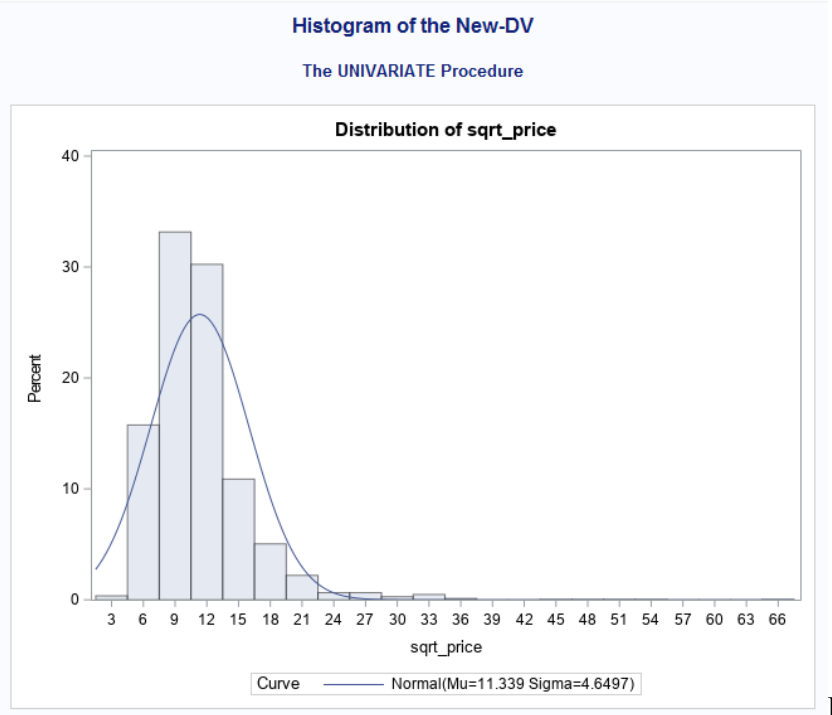
The REG Procedure
Model: MODEL1
Dependent Variable: logPrice

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	4.5747	0.0284	4.5189	4.6305	3.8096	5.3398	.
2	.	4.0676	0.0927	3.8859	4.2494	3.2832	4.8520	.
3	3.40	4.3225	0.0257	4.2722	4.3729	3.5578	5.0872	-0.9213
4	3.91	4.0052	0.0234	3.9594	4.0511	3.2408	4.7696	-0.0932

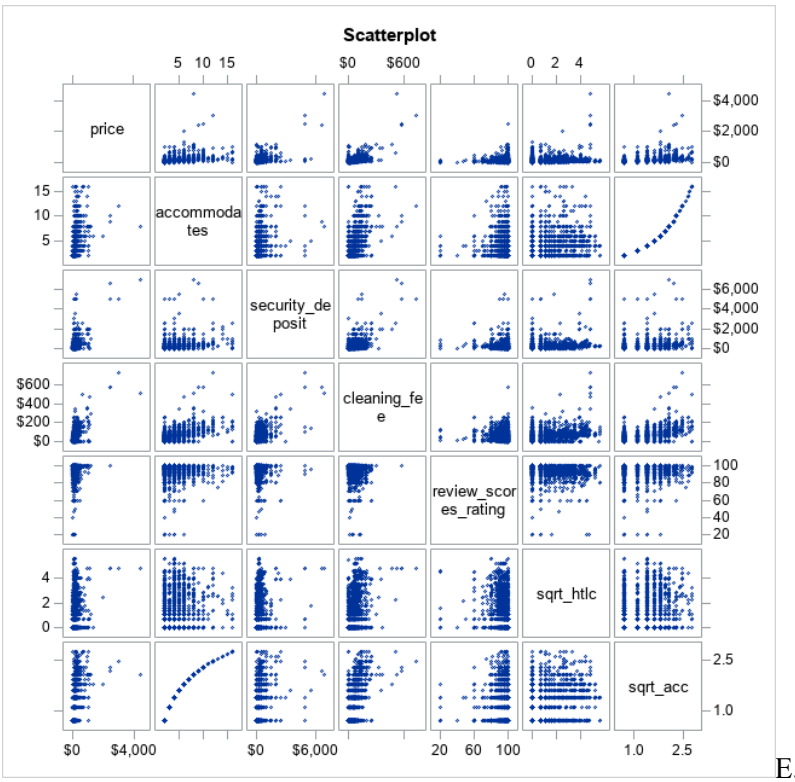
Appendix E - Brendan A. Foley



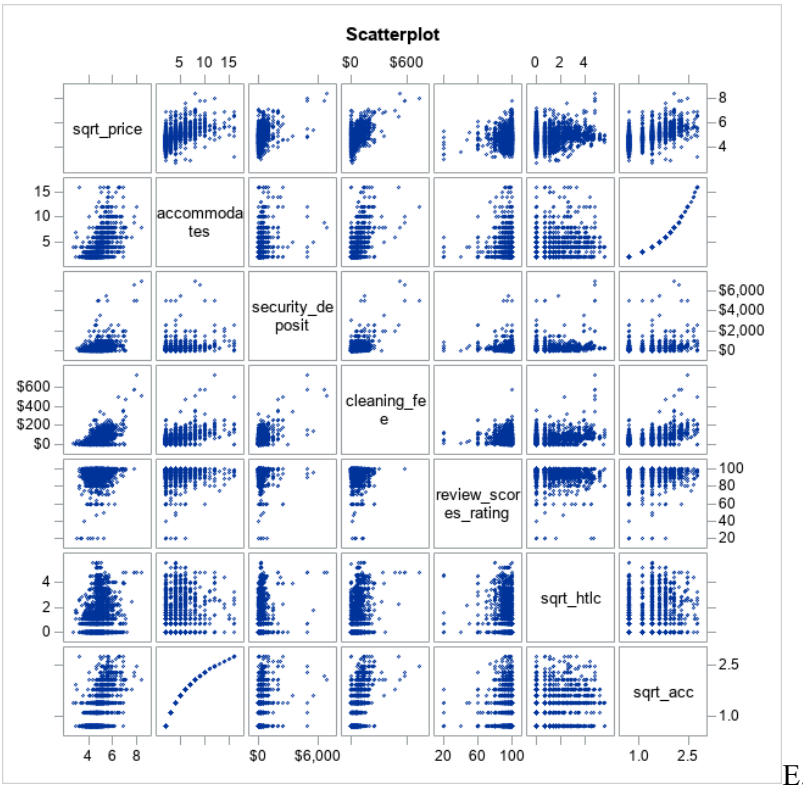
E.1.



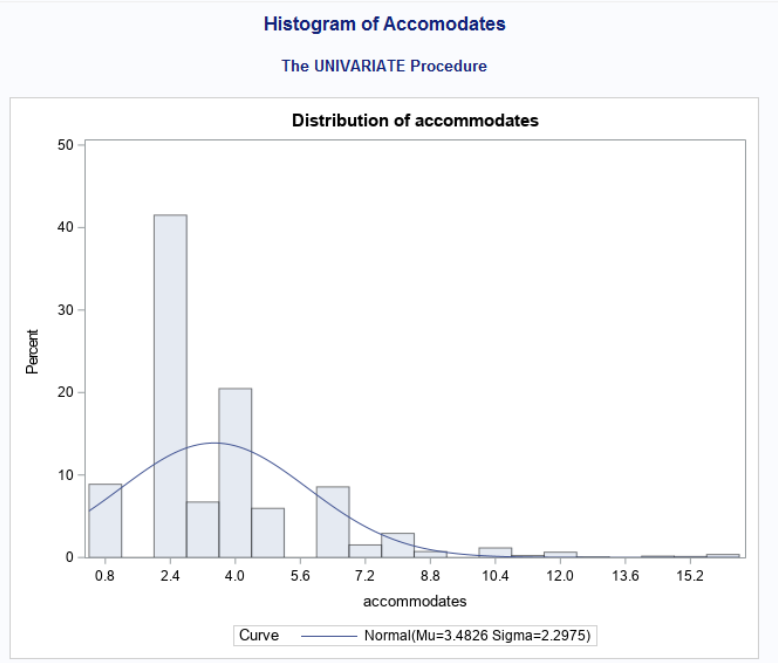
E.2.


Scatterplot
price
accommodates
security_deposit
cleaning_fee
review_scores_rating
sqrt_htlc
sqrt_acc

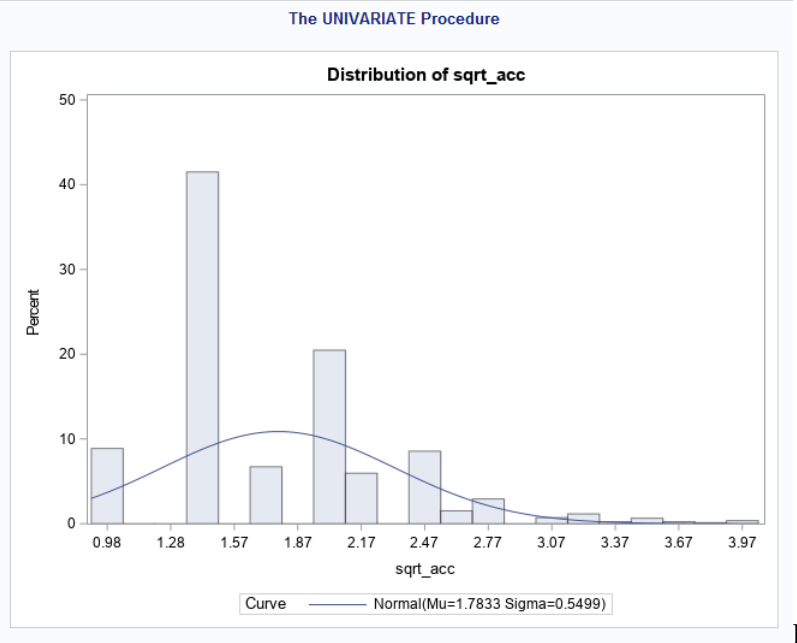
E.3.


Scatterplot
sqrt_price
accommodates
security_deposit
cleaning_fee
review_scores_rating
sqrt_htlc
sqrt_acc

E.4.


Histogram of Accomodates
The UNIVARIATE Procedure
Distribution of accommodates
Percent
50
40
30
20
10
0
0.8
2.4
4.0
5.6
7.2
8.8
10.4
12.0
13.6
15.2
accommodates
Curve
Normal(Mu=3.4826 Sigma=2.2975)

E.5


The UNIVARIATE Procedure
Distribution of sqrt_acc
Percent
50
40
30
20
10
0
0.98
1.28
1.57
1.87
2.17
2.47
2.77
3.07
3.37
3.67
3.97
sqrt_acc
Curve
Normal(Mu=1.7833 Sigma=0.5499)

E.6

Regression of All

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_price

Number of Observations Read	2501
Number of Observations Used	1329
Number of Observations with Missing Values	1172

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	10780	598.88670	85.75	<.0001
Error	1310	9148.79134	6.98381		
Corrected Total	1328	19929			

Root MSE	2.64269	R-Square	0.5409
Dependent Mean	11.61258	Adj R-Sq	0.5346
Coeff Var	22.75714		

E.7

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.94975	0.96797	2.01	0.0442
host_total_listings_count	1	-0.00563	0.00254	-2.22	0.0268
sqrt_acc	1	2.77022	0.16714	16.57	<.0001
security_deposit	1	0.00129	0.00021441	6.02	<.0001
cleaning_fee	1	0.01590	0.00201	7.91	<.0001
review_scores_rating	1	0.03896	0.00917	4.25	<.0001
num_h_super	1	-0.29779	0.16266	-1.83	0.0674
d_host_r1	1	0.39332	0.19590	2.01	0.0449
d_host_r2	1	0.33946	0.94513	0.36	0.7195
d_rty1	1	-1.76934	0.20685	-8.55	<.0001
d_rty2	1	-2.82449	0.83932	-3.37	0.0008
d_bty1	0	0	.	.	.
d_bty2	1	-0.95701	2.65427	-0.36	0.7185
d_bty3	1	-1.44517	1.87626	-0.77	0.4413
d_cpol1	1	-0.09132	0.22220	-0.41	0.6811
d_cpol2	1	-0.26585	0.16782	-1.58	0.1134
d_reg1	1	1.42742	1.08547	1.32	0.1887
d_reg2	1	0.47806	0.35612	1.34	0.1797
d_reg3	1	-0.18726	0.22772	-0.82	0.4110
d_reg4	1	-0.04849	0.18166	-0.27	0.7896

E.8

d_cpol2	d_reg1	d_reg2	d_reg3	d_reg4	host1_rev	host2_rev	sec_acc	
-0.03558 0.0753 2500	-0.01192 0.5512 2500	-0.00906 0.6507 2500	-0.01076 0.5908 2500	0.00430 0.8299 2500	-0.08861 <.0001 2500	-0.03511 0.0793 2500	0.52175 <.0001 2500	sqrt_price
-0.07679 0.0001 2501	0.00890 0.6566 2501	0.02616 0.1909 2501	-0.00965 0.6294 2501	-0.01823 0.3622 2501	-0.18772 <.0001 2501	-0.03300 0.0989 2501	0.16779 <.0001 2501	host_total_listings_count
-0.08583 <.0001 2501	0.01518 0.4481 2501	0.01858 0.3529 2501	0.00051 0.9799 2501	-0.02056 0.3041 2501	-0.26176 <.0001 2501	-0.04180 0.0366 2501	0.17692 <.0001 2501	sqrt_htlc
-0.00927 0.6432 2501	-0.01208 0.5461 2501	-0.00618 0.7574 2501	0.00434 0.8281 2501	0.01340 0.5029 2501	-0.19730 <.0001 2501	-0.08465 <.0001 2501	0.38947 <.0001 2501	sqrt_acc
-0.07153 0.0003 2501	-0.00438 0.8267 2501	-0.02369 0.2362 2501	0.00224 0.9107 2501	-0.01190 0.5521 2501	-0.01420 0.4779 2501	-0.00952 0.6341 2501	0.94326 <.0001 2501	security_deposit
-0.06425 0.0013 2501	-0.00984 0.6228 2501	-0.01435 0.4732 2501	-0.01234 0.5372 2501	-0.00640 0.7491 2501	-0.11645 <.0001 2501	-0.01175 0.5570 2501	0.49335 <.0001 2501	sqrt_clean
0.08824 <.0001 2501	0.01950 0.3295 2501	0.01334 0.5050 2501	-0.00698 0.7272 2501	-0.00897 0.6538 2501	0.04992 0.0125 2501	0.01033 0.6055 2501	-0.00686 0.7318 2501	review_scores_rating
-0.17751 <.0001 2501	-0.02085 0.2974 2501	0.01076 0.5906 2501	-0.01480 0.4594 2501	0.01733 0.3864 2501	0.25433 <.0001 2501	0.06502 0.0011 2501	-0.05532 0.0057 2501	num_h_super
-0.10129 <.0001 2501	-0.05126 0.0103 2501	0.00690 0.7300 2501	0.00922 0.6451 2501	-0.02045 0.3066 2501	0.99557 <.0001 2501	-0.07934 <.0001 2501	-0.06056 0.0024 2501	d_host_r1
-0.03048 0.1275 2501	0.03747 0.0610 2501	-0.00324 0.8713 2501	0.01027 0.6076 2501	-0.02832 0.1568 2501	-0.07986 <.0001 2501	0.98908 <.0001 2501	-0.02440 0.2226 2501	d_host_r2

Identifies the multicolinearity between the interaction variables. E.9

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.68882	1.17936	3.13	0.0018	0
host_total_listings_count	1	-0.01216	0.00752	-1.62	0.1060	8.78827
sqrt_acc	1	1.87236	0.18007	10.40	<.0001	2.34148
sqrt_htlc	1	0.08456	0.09131	0.93	0.3545	9.63696
security_deposit	1	-0.00377	0.00055949	-6.74	<.0001	12.63868
sqrt_clean	1	0.23646	0.03485	6.79	<.0001	1.71070
review_scores_rating	1	0.03084	0.01091	2.83	0.0048	1.68583
num_h_super	1	-0.18693	0.16550	-1.13	0.2588	1.18928
d_host_r1	1	0.62890	1.79135	0.35	0.7256	166.53450
d_host_r2	1	0.86731	3.93164	0.22	0.8254	49.58834
d_rty1	1	-2.31187	0.17319	-13.35	<.0001	1.64194
d_rty2	1	-3.65548	0.51559	-7.09	<.0001	1.12367
d_bty1	1	-1.67426	3.29593	-0.51	0.6115	1.03868
d_bty2	1	-0.12814	1.14951	-0.11	0.9112	1.00792
d_bty3	1	0.44795	1.23682	0.36	0.7172	1.02140
d_cpol1	1	-0.20182	0.17177	-1.17	0.2401	1.57040
d_cpol2	1	-0.36073	0.16999	-2.12	0.0339	1.33346
d_reg1	1	0.09255	0.88357	0.10	0.9166	1.03961
d_reg2	1	-0.02943	0.19375	-0.15	0.8793	2.24236
d_reg3	1	-0.24505	0.25449	-0.96	0.3357	1.62132
d_reg4	1	-0.02147	0.22064	-0.10	0.9225	1.97664
host1_rev	1	-0.00113	0.01895	-0.06	0.9526	166.50851
host2_rev	1	-0.00078862	0.04190	-0.02	0.9850	49.52331
sec_acc	1	0.00277	0.00022788	12.17	<.0001	14.35431

E.10

Interaction variables, decision to center thereafter

Regression of All

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_price

Number of Observations Read	2501
Number of Observations Used	2500
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	10264	1710.63391	97.45	<.0001
Error	2493	43764	17.55486		
Corrected Total	2499	54028			

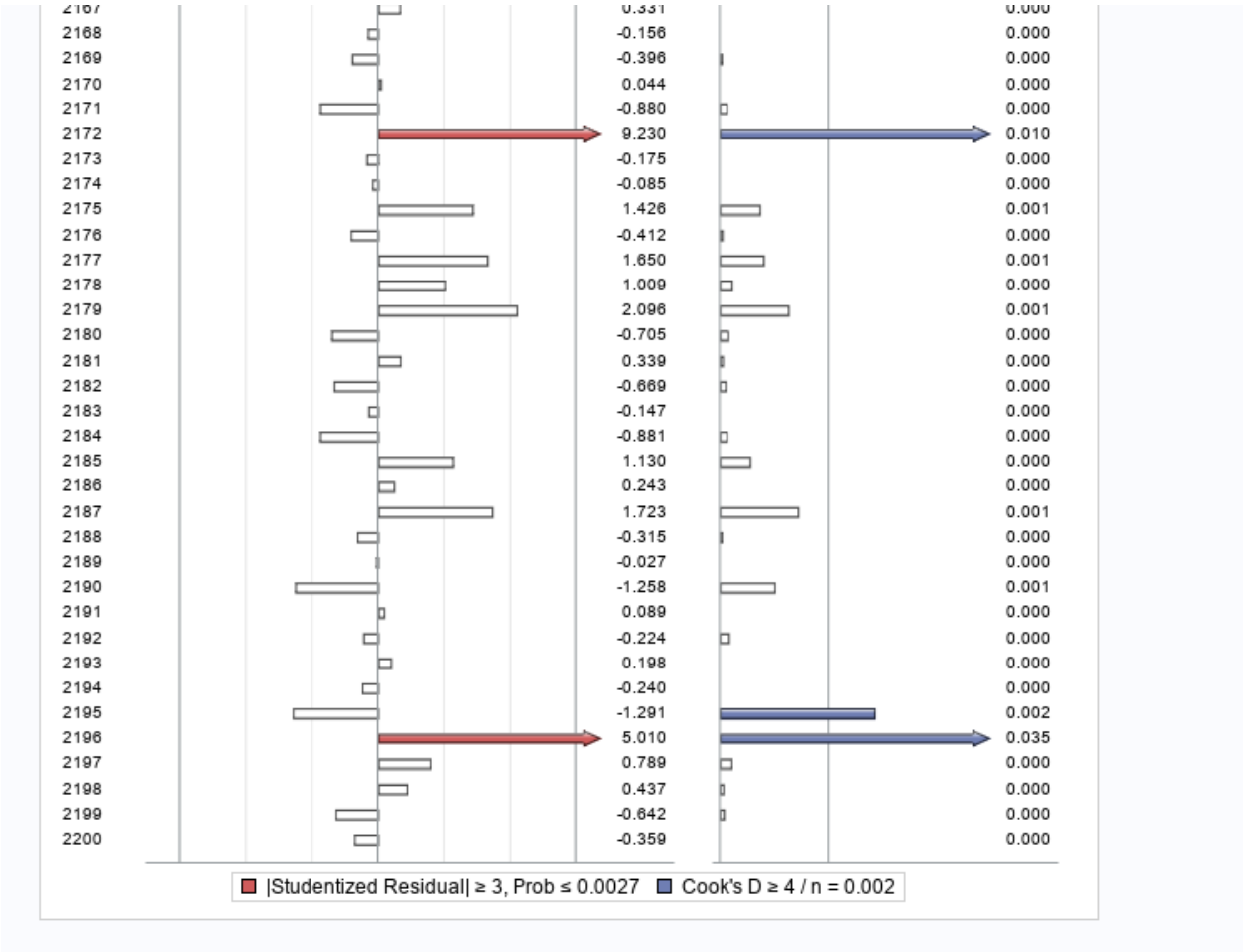
Root MSE	4.18985	R-Square	0.1900
Dependent Mean	11.33917	Adj R-Sq	0.1880
Coeff Var	36.95025		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	5.25696	4.22065	1.25	0.2131	0	0
security_deposit_c	1	-0.00256	0.00035587	-7.18	<.0001	-0.22597	3.04500
d_host1_rev_c	1	4.15792	1.91249	2.17	0.0298	0.41667	113.04258
d_host2_rev_c	1	5.98751	4.90900	1.22	0.2227	0.14918	46.03843
host1_rev_c	1	-0.03772	0.02161	-1.75	0.0810	-0.33454	113.03572
host2_rev_c	1	-0.33579	0.37202	-0.90	0.3668	-0.11039	46.03165
sec_acc_c	1	1.992756E-7	2.893326E-8	6.89	<.0001	0.21679	3.04921

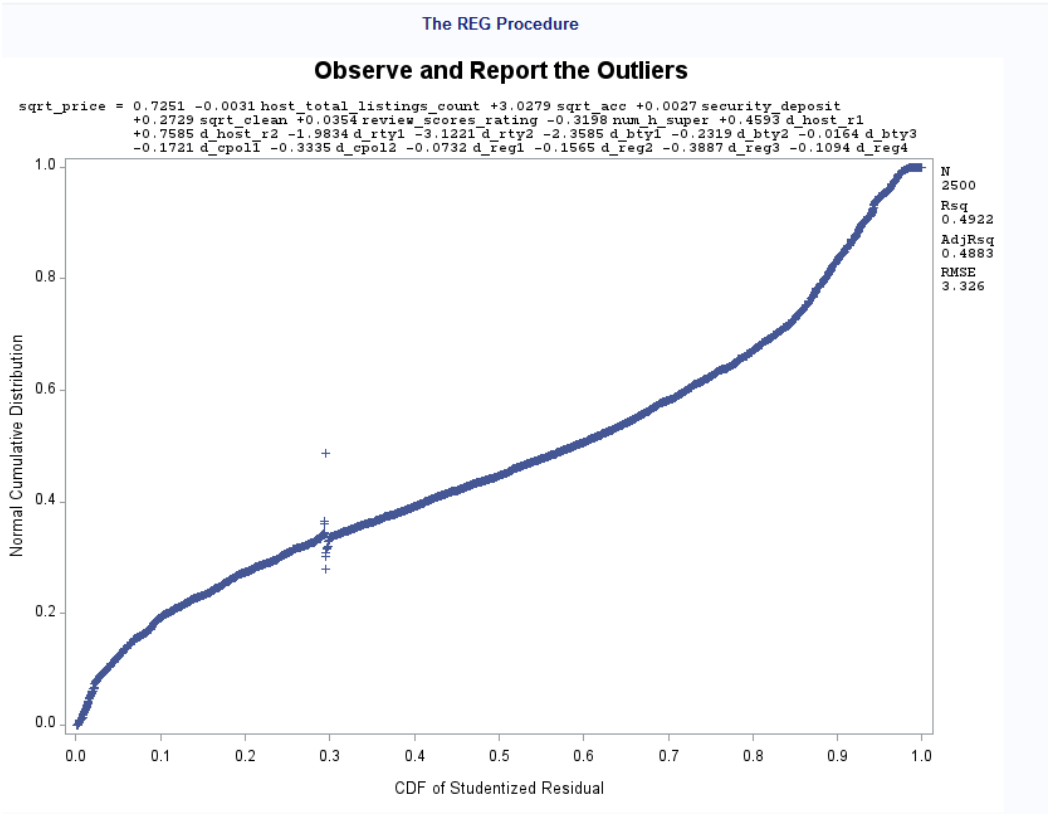
E.11

Post centering of the interaction variables. Too high a p-value for the model to continue with. Security deposit impacts the model more than the combined security deposit accommodations variable so it was discarded as well.



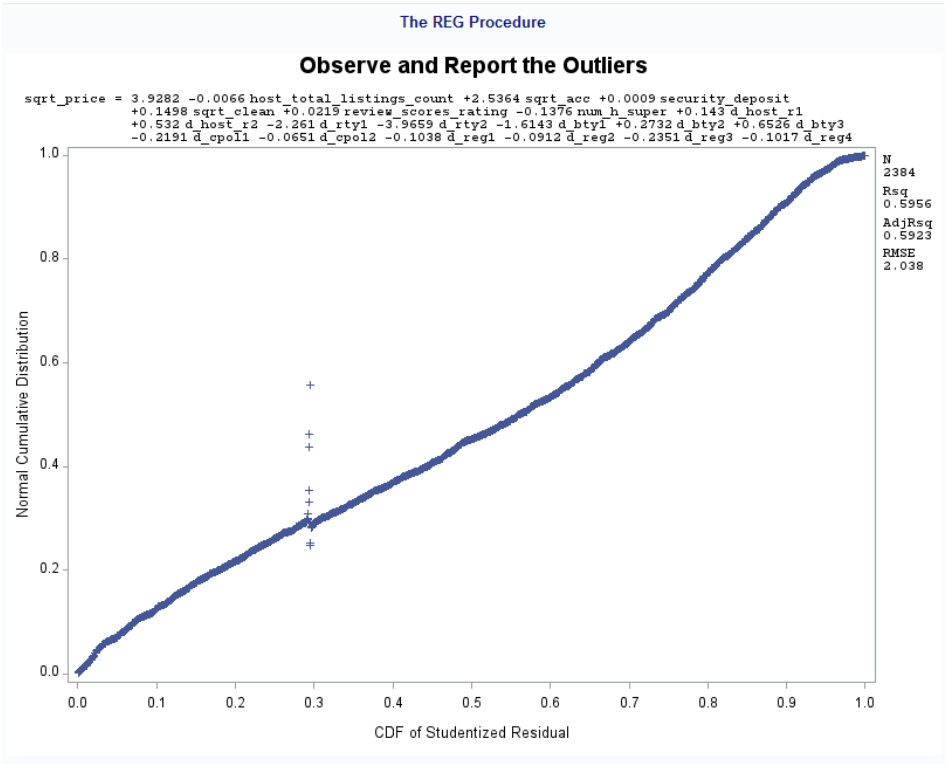
E.12

Decision to remove outliers based on the higher than 3 Cook's D and the higher than 3 Studentized Residual both occurring as seen above.



E.13

Finished model NPP



E.14

Future uses of this data may need to better eliminate outliers that do not coincide with influence

points.

Regression of All

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	MEL_BNB_NEW5
Random Number Seed	495857
Sampling Rate	0.8
Sample Size	1908
Selection Probability	0.8
Sampling Weight	0
Output Data Set	TEST_MELB_01

E.15

Breakdown of the test and train data set.

Root MSE	2.03802
Dependent Mean	10.69886
R-Square	0.5935
Adj R-Sq	0.5923
AIC	5788.70947
AICC	5788.78529
SBC	3448.92175
CV PRESS	9929.62229

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	1907	477	2006.6201
2	1907	477	2117.0156
3	1907	477	1911.2231
4	1907	477	1961.9334
5	1908	476	1932.8301
Total			9929.6223

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	3.557067	0.569182	6.25	3.865525	3.62486	3.825696	2.712810	3.652726
host_total_listings_	1	-0.006518	0.001712	-3.81	-0.007015	-0.00672	-0.007016	-0.005351	-0.006282
sqrt_acc	1	2.533749	0.099736	25.40	2.452284	2.56164	2.567726	2.595788	2.497616
security_deposit	1	0.000924	0.000162	5.71	0.000815	0.00100	0.000894	0.000979	0.000934
sqrt_clean	1	0.151009	0.022734	6.64	0.170770	0.13387	0.163452	0.149759	0.136858
review_scores_rating	1	0.023305	0.005456	4.27	0.020333	0.02312	0.019212	0.030637	0.024199
d_rty1	1	-2.293645	0.107388	-21.36	-2.328732	-2.31022	-2.278464	-2.223643	-2.323423
d_rty2	1	-4.012348	0.325300	-12.33	-4.050740	-3.91714	-4.077072	-3.998594	-4.032388

E.16

Stepwise Model

Validation-Test Set

The REG Procedure
Model: MODEL1
Dependent Variable: New_LnP

Number of Observations Read	2385
Number of Observations Used	1907
Number of Observations with Missing Values	478

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	11709	1672.75934	412.00	<.0001
Error	1899	7710.14732	4.06011		
Corrected Total	1906	19419			

Root MSE	2.01497	R-Square	0.6030
Dependent Mean	10.65930	Adj R-Sq	0.6015
Coeff Var	18.90341		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.83335	0.60629	6.32	<.0001
host_total_listings_count	1	-0.00648	0.00189	-3.43	0.0006
sqrt_acc	1	2.47180	0.10835	22.81	<.0001
security_deposit	1	0.00087251	0.00017452	5.00	<.0001
sqrt_clean	1	0.14883	0.02506	5.94	<.0001
review_scores_rating	1	0.02208	0.00577	3.82	0.0001
d_rty1	1	-2.38805	0.11819	-20.21	<.0001
d_rty2	1	-4.03544	0.35776	-11.28	<.0001

E.17

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	11746	903.50086	222.87	<.0001
Error	1893	7673.95152	4.05386		
Corrected Total	1906	19419			

Root MSE	2.01342	R-Square	0.6048
Dependent Mean	10.65930	Adj R-Sq	0.6021
Coeff Var	18.88885		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.06685	0.63955	6.36	<.0001
host_total_listings_count	1	-0.00641	0.00194	-3.31	0.0010
sqrt_acc	1	2.47496	0.10972	22.56	<.0001
security_deposit	1	0.00087959	0.00017502	5.03	<.0001
sqrt_clean	1	0.14751	0.02514	5.87	<.0001
review_scores_rating	1	0.02066	0.00589	3.51	0.0005
num_h_super	1	-0.16365	0.11796	-1.39	0.1655
d_host_r1	1	0.15575	0.10922	1.43	0.1540
d_host_r2	1	0.67673	0.40606	1.67	0.0958
d_cpol1	1	-0.20167	0.10908	-1.85	0.0646
d_rty1	1	-2.35872	0.12093	-19.51	<.0001
d_rty2	1	-3.98048	0.35944	-11.07	<.0001
d_bty3	1	0.17301	1.01598	0.17	0.8648
d_reg2	1	0.04462	0.09254	0.48	0.6298

E.18

Backward Variables

Compute the prediction

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_price

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	11.6023	0.8325	9.9698	13.2349	7.2776	15.9270	.
2	15.10	16.4316	0.1646	16.1088	16.7544	12.4138	20.4493	-1.3319
3	17.32	12.4898	0.1378	12.2196	12.7600	8.4759	16.5036	4.8307
4	11.83	9.3716	0.1321	9.1125	9.6308	5.3585	13.3848	2.4605
5	11.27	14.0912	0.1423	13.8120	14.3703	10.0767	18.1056	-2.8218
6	16.12	14.1943	0.1746	13.8519	14.5367	10.1750	18.2137	1.9302
7	5.92	7.9839	0.1363	7.7167	8.2511	3.9702	11.9975	-2.0678
8	9.80	12.4182	0.0993	12.2234	12.6130	8.4088	16.4277	-2.6203

Stepwise

Joining the Data set Predictions Test

Obs	Selected	host_total_listings_count	security_deposit	review_scores_rating	sqrt_price	d_rty1	d_rty2	sqrt_acc	sqrt_clean
1	1	9	$300.00	90.000	.	1	0	2.00000	11.0000
2	1	10	$800.00	94.100	15.0997	0	0	3.16228	12.8452
3	1	24	$250.00	96.000	17.3205	0	0	2.00000	7.7460
4	1	5	$0.00	94.100	11.8322	1	0	2.00000	7.0711
5	1	4	$400.00	88.000	11.2694	0	0	2.44949	12.6491
6	0	3	$500.00	80.000	16.1245	0	0	2.44949	14.1421
7	1	1	$326.98	90.000	5.9161	1	0	1.41421	3.8730
8	1	1	$300.00	96.000	9.7980	0	0	2.00000	8.3066
9	1	17	$500.00	98.000	13.1909	0	0	2.00000	11.4018
10	1	1	$326.98	100.000	6.7823	1	0	1.41421	8.0916
11	0	2	$2,000.00	99.000	19.4679	0	0	2.00000	12.2474
12	0	3	$300.00	80.000	14.1067	0	0	2.82843	10.0000
13	0	1	$500.00	93.000	12.9615	0	0	2.44949	10.0000

E.19

Compute the prediction

The REG Procedure
Model: MODEL1
Dependent Variable: sqrt_price

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	10.4977	0.1188	10.2647	10.7307	6.4395	14.5559	.
2	15.10	16.0326	0.1363	15.7653	16.3000	11.9723	20.0930	-0.9330
3	17.32	12.1935	0.0575	12.0807	12.3063	8.1404	16.2466	5.1270
4	11.83	9.8236	0.0919	9.6433	10.0039	5.7680	13.8792	2.0086
5	11.27	14.1811	0.1011	13.9829	14.3794	10.1247	18.2375	-2.9117
6	16.12	14.4373	0.1268	14.1886	14.6860	10.3781	18.4965	1.6872
7	5.92	7.7809	0.0934	7.5978	7.9640	3.7252	11.8366	-1.8648
8	9.80	12.2897	0.0545	12.1829	12.3965	8.2368	16.3427	-2.4918
9	13.19	12.8208	0.0808	12.6624	12.9791	8.7661	16.8754	0.3701
10	6.78	8.5047	0.0765	8.3548	8.6547	4.4504	12.5591	-1.7224
11	19.47	12.9659	0.0954	12.7788	13.1529	8.9100	17.0217	6.5021

E.20

Outcome

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	14410	2058.56780	495.62
Error	2376	9868.81078	4.15354	
Corrected Total	2383	24279		

Root MSE	2.03802
Dependent Mean	10.69886
R-Square	0.5935
Adj R-Sq	0.5923
AIC	5788.70947
AICC	5788.78529
SBC	3448.92175
CV PRESS	9969.34550

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	3.557067	0.569182	6.25
host_total_listings_	1	-0.006518	0.001712	-3.81
sqrt_acc	1	2.533749	0.099736	25.40
security_deposit	1	0.000924	0.000162	5.71
sqrt_clean	1	0.151009	0.022734	6.64
review_scores_rating	1	0.023305	0.005456	4.27
d_rty1	1	-2.293645	0.107388	-21.36
d_rty2	1	-4.012348	0.325300	-12.33

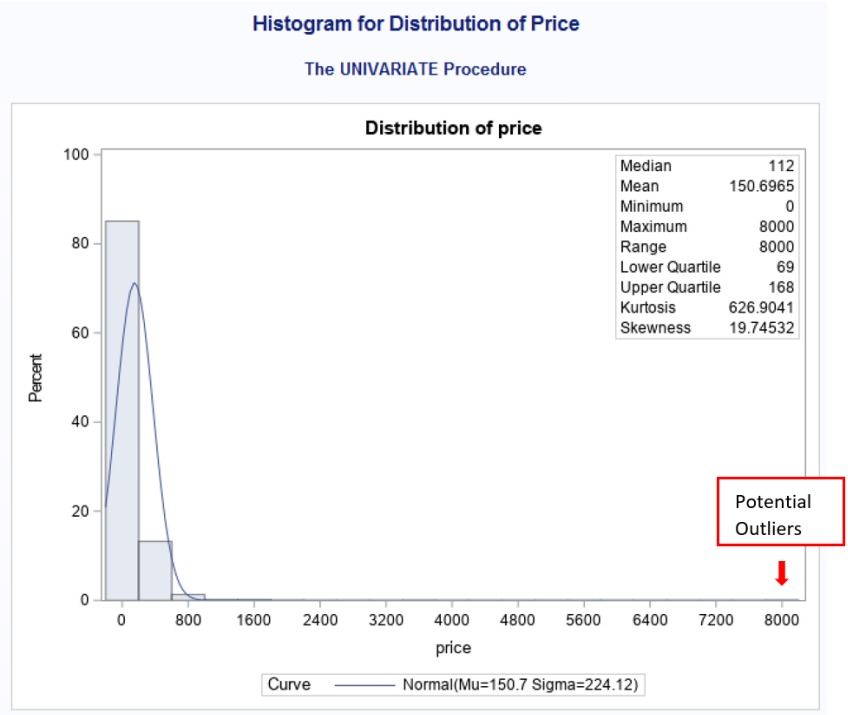
E.21

Root MSE	2.03802
Dependent Mean	10.69886
R-Square	0.5935
Adj R-Sq	0.5923
AIC	5788.70947
AICC	5788.78529
SBC	3448.92175
CV PRESS	9969.34550

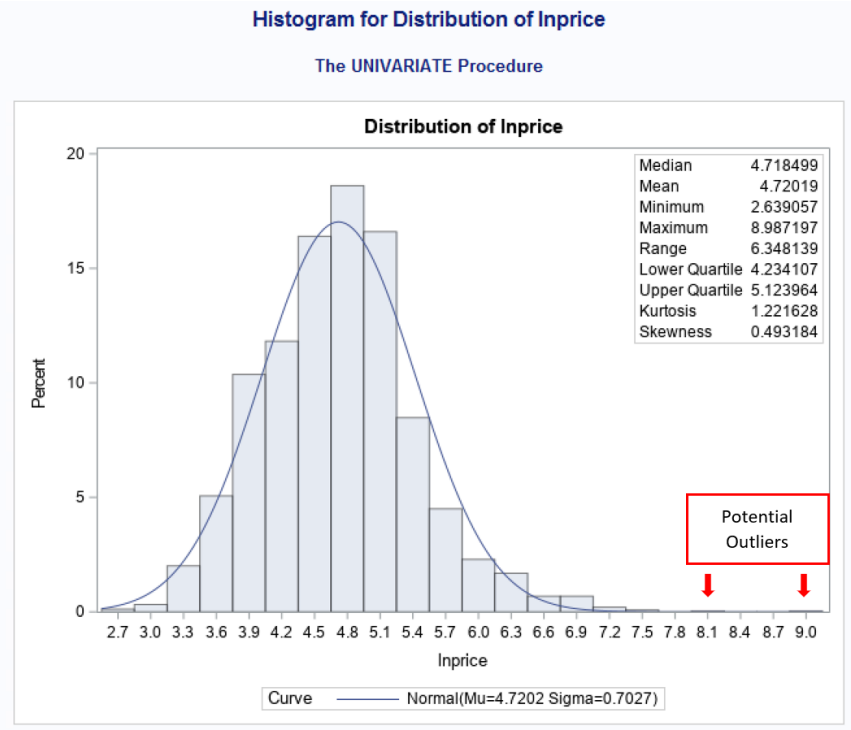
Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	3.557067	0.569182	6.25
host_total_listings_	1	-0.006518	0.001712	-3.81
sqrt_acc	1	2.533749	0.099736	25.40
security_deposit	1	0.000924	0.000162	5.71
sqrt_clean	1	0.151009	0.022734	6.64
review_scores_rating	1	0.023305	0.005456	4.27
d_rty1	1	-2.293645	0.107388	-21.36
d_rty2	1	-4.012348	0.325300	-12.33

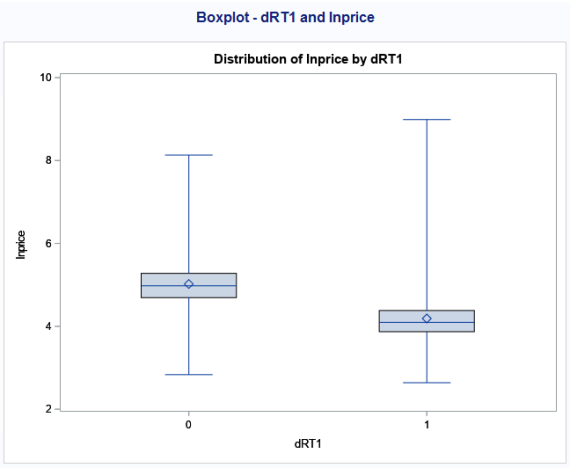
Appendix F : Ying



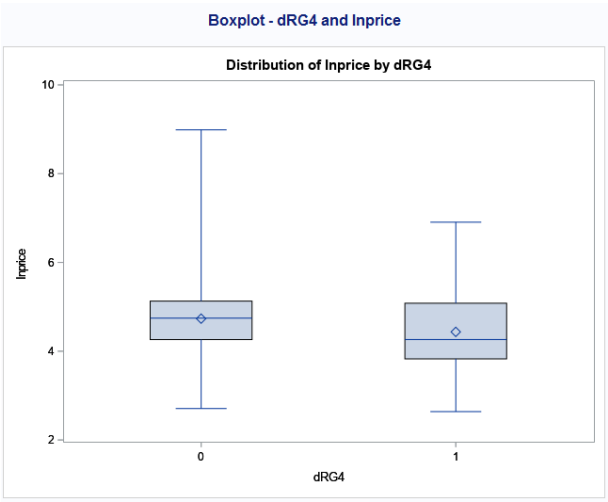
F.1 Histogram before log transformation



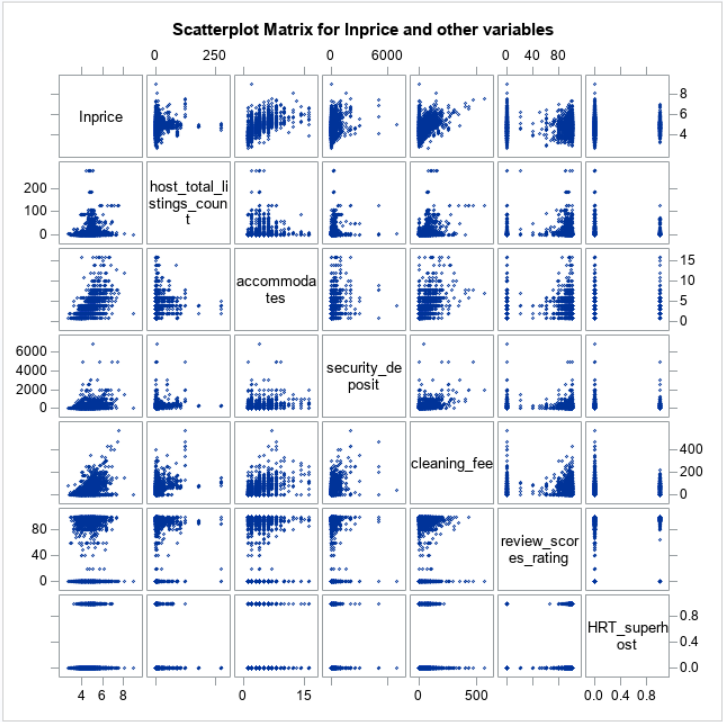
F.2 Histogram after log transformation



F.3 Boxplot – dRT1 and lnprice



F.4 Boxplot – dRG4 and lnprice



F.5 Scatterplot Mattrix for lnprice and other variables

Full Model

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	2491
Number of Observations Used	2488
Number of Observations with Missing Values	3

Analysis of Variance

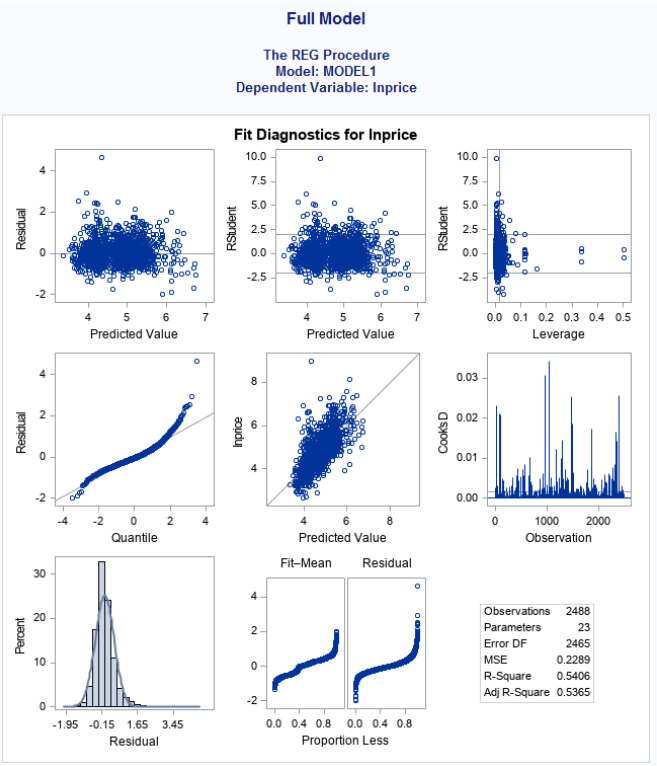
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	663.87800	30.17627	131.84	<.0001
Error	2465	564.18906	0.22888		
Corrected Total	2487	1228.06706			

Root MSE	0.47841	R-Square	0.5406
Dependent Mean	4.72019	Adj R-Sq	0.5365
Coeff Var	10.13548		

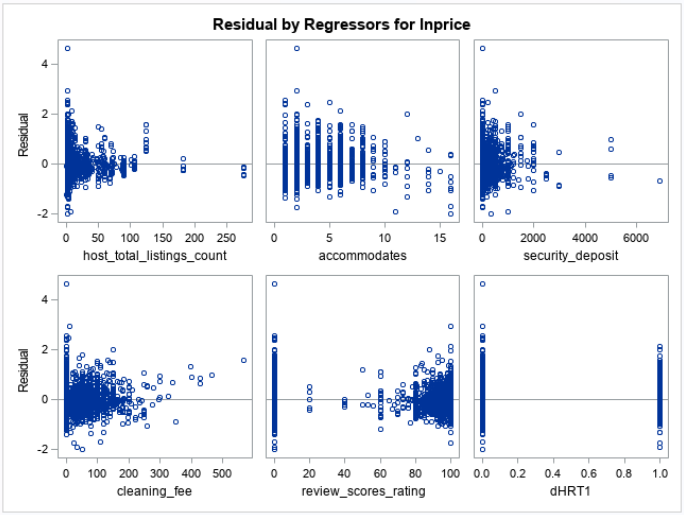
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.63506	0.03867	119.85	<.0001	0
host_total_listings_count	1	-0.00014450	0.00042592	-0.34	0.7344	1.25099
accommodates	1	0.12622	0.00561	22.49	<.0001	1.60069
security_deposit	1	0.00011804	0.00002803	4.21	<.0001	1.30212
cleaning_fee	1	0.00045401	0.00024233	1.87	0.0611	1.90107
review_scores_rating	1	-0.00171	0.00025461	-6.71	<.0001	1.24587
dHRT1	1	-0.03836	0.02641	-1.45	0.1465	1.89563
dHRT2	1	0.04049	0.03657	1.11	0.2684	1.30009
dHRT3	1	-0.01298	0.04204	-0.31	0.7575	1.14242
dHRT4	1	0.15292	0.08523	1.79	0.0729	1.03336
dHIS	1	0.01112	0.04547	0.24	0.8068	3.99421
dRT1	1	-0.54754	0.02498	-21.92	<.0001	1.56666
dRT2	1	-1.01180	0.07918	-12.78	<.0001	1.07808
dBT1	1	-0.16178	0.16064	-1.01	0.3140	1.01106
dBT2	0	0	.	.	.	.
dBT3	1	-0.23736	0.27667	-0.86	0.3910	1.00212
dBT4	1	-0.20630	0.33975	-0.61	0.5438	1.00784
dCP1	1	0.00271	0.02510	0.11	0.9140	1.31269
dCP2	1	0.00061958	0.02570	0.02	0.9808	1.59956
dRG1	1	-0.23584	0.03328	-7.09	<.0001	1.12264
dRG2	1	-0.05271	0.02487	-2.12	0.0342	1.15331
dRG3	1	0.00570	0.03380	0.17	0.8660	1.11864
dRG4	1	-0.26747	0.04568	-5.85	<.0001	1.09890
HRT_superhost	1	0.02139	0.05346	0.40	0.6891	4.47329

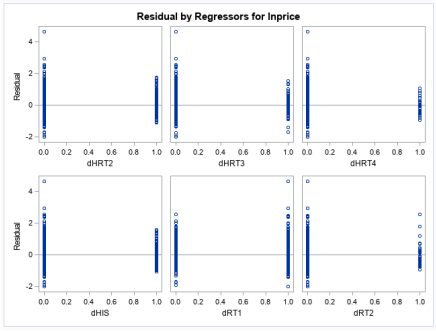
F.6 Full Model Regression

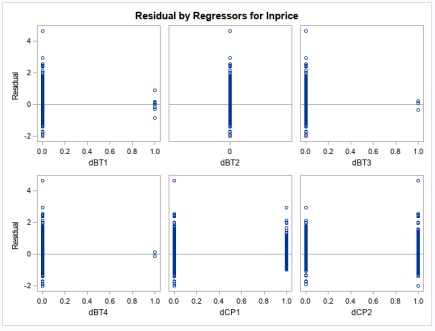


F.7 Residula plots for full model

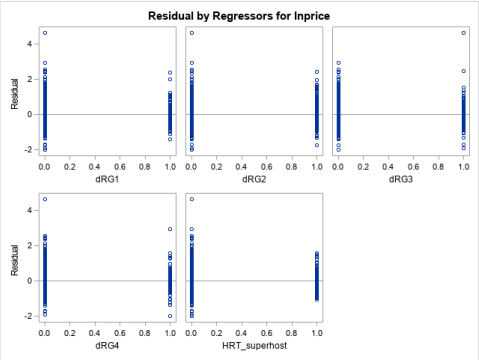


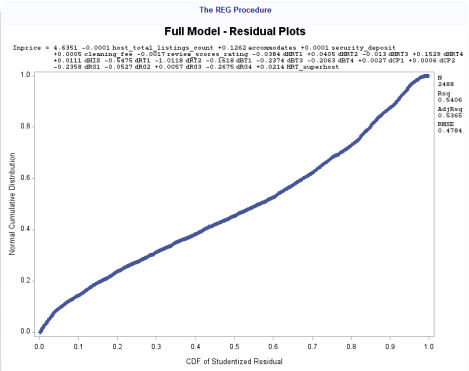
F.7 Residula plots for full model





F.7 Residula plots for full model



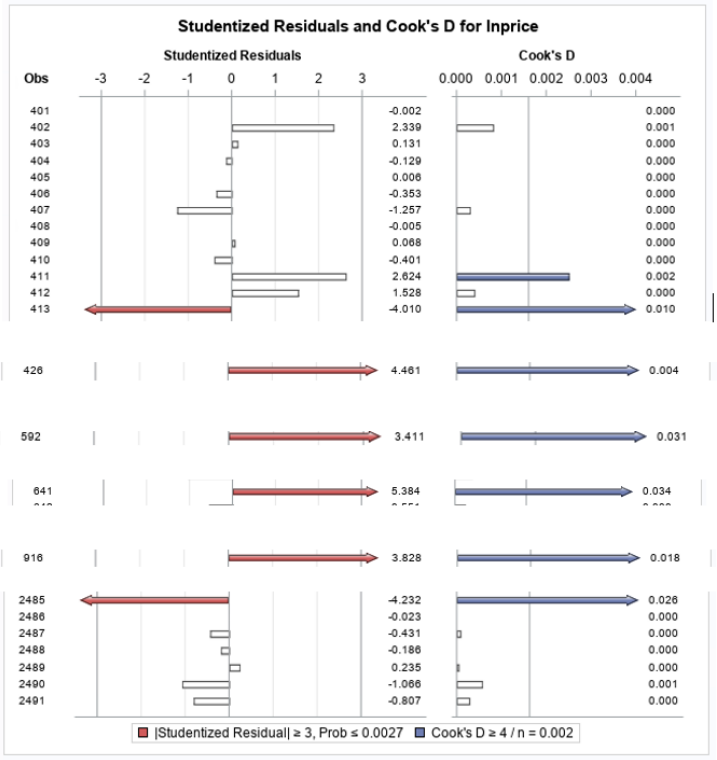


F.7 Residula plots for full model

F.8 Normality graph for full model

						Correlation					
Variable	host_total_listings_count	accommodates	security_deposit	cleaning_fee	review_scores_rating	dHRT1	dHRT2	dHRT3	dHRT4	dHIS	lnprice
host_total_listings_count	1.0000	0.1134	0.1597	0.3447	0.0133	0.2113	-0.0339	0.0162	-0.0305	-0.0169	0.1328
accommodates	0.1134	1.0000	0.2716	0.4888	0.0675	0.1709	0.0043	-0.0234	-0.0317	0.0784	0.6137
security_deposit	0.1597	0.2716	1.0000	0.4606	-0.0034	0.0677	0.0573	-0.0025	-0.0083	0.0658	0.2779
cleaning_fee	0.3447	0.4888	0.4606	1.0000	0.1049	0.1821	0.0640	-0.0182	-0.0267	0.0998	0.4331
review_scores_rating	0.0133	0.0675	-0.0034	0.1049	1.0000	0.2717	0.0446	-0.0867	-0.0602	0.2908	0.0071
dHRT1	0.2113	0.1709	0.0677	0.1821	0.2717	1.0000	-0.3368	-0.2642	-0.1176	0.2715	0.1188
dHRT2	-0.0339	0.0043	0.0573	0.0640	0.0446	-0.3368	1.0000	-0.0865	-0.0385	0.0348	0.0230
dHRT3	0.0162	-0.0234	-0.0025	-0.0182	-0.0867	-0.2642	-0.0865	1.0000	-0.0302	-0.0919	-0.0340
dHRT4	-0.0305	-0.0317	-0.0083	-0.0267	-0.0602	-0.1176	-0.0385	-0.0302	1.0000	-0.0552	0.0054
dHIS	-0.0169	0.0784	0.0658	0.0998	0.2908	0.2715	0.0348	-0.0919	-0.0552	1.0000	0.0700
dRT1	-0.1681	-0.4737	-0.1979	-0.4324	-0.1558	-0.2067	-0.0010	0.0612	0.0516	-0.1032	-0.5696
dRT2	-0.0233	-0.1203	-0.0495	-0.0946	-0.0970	-0.0658	0.0324	0.0191	-0.0148	-0.0625	-0.1841
dBT1	-0.0207	-0.0486	-0.0227	-0.0235	0.0199	0.0192	-0.0200	-0.0157	-0.0070	0.0305	-0.0532
dBT2	.	.	.	.	.	.	.	.	.	.	.
dBT3	-0.0110	-0.0066	-0.0022	-0.0075	-0.0031	0.0111	-0.0115	-0.0090	-0.0040	0.0084	-0.0244
dBT4	-0.0097	-0.0185	-0.0151	0.0027	-0.0148	-0.0288	0.0380	-0.0074	-0.0033	0.0181	-0.0028
dCP1	-0.0918	-0.0255	-0.0278	-0.0088	0.1853	0.0603	0.0312	-0.0144	0.0038	0.1119	-0.0065
dCP2	-0.2015	-0.2277	-0.2203	-0.3366	-0.2636	-0.2266	-0.0421	0.0106	0.0144	-0.1429	-0.2200
dRG1	-0.0568	-0.0426	-0.0277	-0.0906	-0.0189	-0.0589	-0.0075	0.0300	-0.0280	-0.0152	-0.1467
dRG2	-0.0756	-0.0515	-0.0258	-0.0147	-0.0565	-0.1154	0.0282	0.0130	0.0318	-0.0309	-0.0339
dRG3	-0.0268	0.0174	-0.0520	-0.1069	0.0006	-0.0088	0.0236	-0.0209	-0.0035	0.0205	-0.0017
dRG4	-0.0552	0.0499	-0.0518	-0.0551	-0.0357	0.0058	-0.0098	0.0220	-0.0109	-0.0102	-0.0937
HRT_superhost	-0.0231	0.0654	0.0373	0.0646	0.2475	0.4531	-0.1526	-0.1197	-0.0533	0.8384	0.0603
lnprice	0.1328	0.6137	0.2779	0.4331	0.0071	0.1188	0.0230	-0.0340	0.0054	0.0700	1.0000

F.9 Correlation for lnprice and other variables



F.10 Full Model Extract of Studentized Residual and Cook's D

Full Model

	n	Type	Current#	R Sq		adj R Sq	
	2491			0.5406		0.5365	
1st	2490	O & IF	413	0.5435	↑	0.5394	↑
2nd	2489	O & IF	2419	0.5489	↑	0.5448	↑
3rd	2488	O & IF	2009	0.5509	↑	0.5468	↑
4th	2487	O & IF	1586	0.5545	↑	0.5505	↑
5th	2486	O & IF	915	0.5569	↑	0.5529	↑

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	2486
Number of Observations Used	2483
Number of Observations with Missing Values	3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	677.65199	30.80236	140.51	<.0001
Error	2460	539.26826	0.21921		
Corrected Total	2482	1216.92026			

Root MSE	0.46820	R-Square	0.5569
Dependent Mean	4.71785	Adj R-Sq	0.5529
Coeff Var	9.92410		

F.11 Outliers and Influential Points removal record and final regression

Train and Test Sets for Airbnb Prices

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	PRICE_NEW
Random Number Seed	495857
Sampling Rate	0.75
Sample Size	1865
Selection Probability	0.750201
Sampling Weight	0
Output Data Set	XV_ALL

F.12 Splitting of Train set and Test set

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	accommodates	1	0.3807	0.3807	689.508	1144.03	<.0001
2	dRT1	2	0.1052	0.4859	258.574	380.63	<.0001
3	dRT2	3	0.0277	0.5136	146.461	105.99	<.0001
4	security_deposit	4	0.0102	0.5238	106.667	39.63	<.0001
5	review_scores_rating	5	0.0093	0.5331	70.4842	36.90	<.0001
6	dRG1	6	0.0080	0.5411	39.5577	32.36	<.0001
7	dRG4	7	0.0076	0.5487	10.1282	31.39	<.0001
8	dHRT4	8	0.0012	0.5499	7.1592	4.97	0.0259
9	dRG2	9	0.0012	0.5511	4.1404	5.03	0.0250
10	dCP2	10	0.0004	0.5516	4.4420	1.70	0.1919
11	dHIS	11	0.0004	0.5520	4.7289	1.72	0.1899
12	dHRT1	12	0.0003	0.5523	5.3595	1.38	0.2411
13	cleaning_fee	13	0.0003	0.5526	6.2282	1.14	0.2866
14	dBT1	14	0.0001	0.5527	7.6547	0.58	0.4481

Backward Elimination: Step 13

Variable dCP2 Removed: R-Square = 0.5511 and C(p) = 4.1404

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	506.60145	56.28905	252.80	<.0001
Error	1853	412.58747	0.22266		
Corrected Total	1862	919.18892			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	4.62538	0.03630	3615.20879	16236.5	<.0001
accommodates	0.13037	0.00601	104.85438	470.92	<.0001
security_deposit	0.00018269	0.00003109	7.68936	34.53	<.0001
review_scores_rating	-0.00171	0.00026584	9.16865	41.18	<.0001
dHRT4	0.20484	0.09009	1.15102	5.17	0.0231
dRT1	-0.55199	0.02700	93.07853	418.03	<.0001
dRT2	-1.04402	0.09388	27.53524	123.67	<.0001
dRG1	-0.23822	0.03671	9.37696	42.11	<.0001
dRG2	-0.06185	0.02756	1.12104	5.03	0.0250
dRG4	-0.30322	0.05180	7.63024	34.27	<.0001

F.13 Model Selections – Forward and Backward

Train Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2486
Number of Observations Used	1863
Number of Observations with Missing Values	623

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	489.04506	69.86358	301.29	<.0001
Error	1855	430.14387	0.23188		
Corrected Total	1862	919.18892			

Root MSE	0.48154	R-Square	0.5320
Dependent Mean	4.72104	Adj R-Sq	0.5303
Coeff Var	10.19993		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	4.50783	0.02966	151.97	<.0001	0	0
accommodates	1	0.13903	0.00600	23.18	<.0001	0.42610	1.33909
dHRT4	1	0.22308	0.09188	2.43	0.0153	0.03864	1.00404
dRT1	1	-0.54095	0.02714	-19.93	<.0001	-0.37041	1.36940
dRT2	1	-1.00905	0.09545	-10.57	<.0001	-0.17168	1.04549
dRG1	1	-0.23312	0.03745	-6.22	<.0001	-0.10206	1.06568
dRG2	1	-0.05168	0.02806	-1.84	0.0657	-0.03021	1.06647
dRG4	1	-0.31204	0.05269	-5.92	<.0001	-0.09675	1.05795

F.14 Regression for Train set

Train Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2486
Number of Observations Used	1863
Number of Observations with Missing Values	623

Analysis of Variance

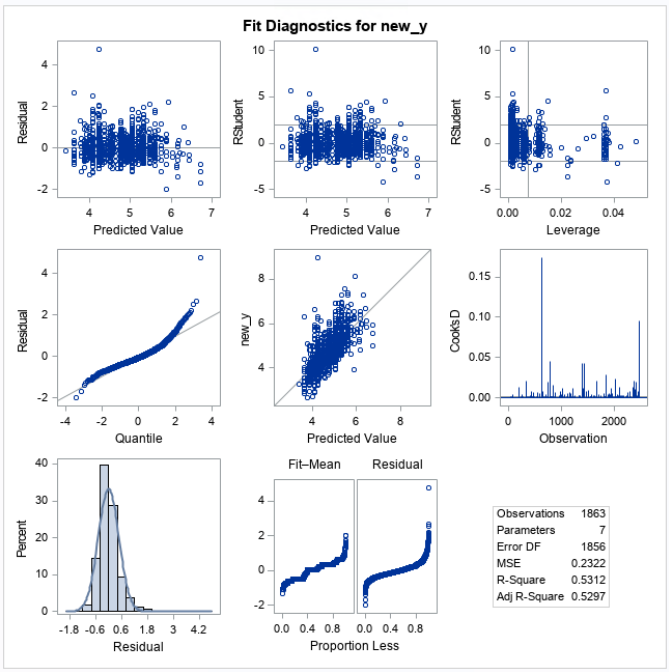
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	488.25853	81.37642	350.49	<.0001
Error	1856	430.93039	0.23218		
Corrected Total	1862	919.18892			

Root MSE	0.48185	R-Square	0.5312
Dependent Mean	4.72104	Adj R-Sq	0.5297
Coeff Var	10.20650		

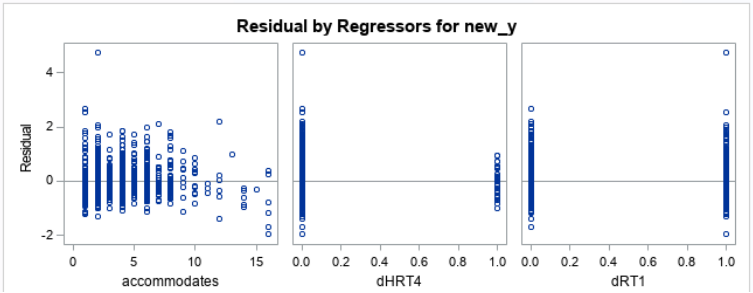
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	4.49647	0.02903	154.87	<.0001	0	0
accommodates	1	0.13895	0.00600	23.16	<.0001	0.42585	1.33902
dHRT4	1	0.21952	0.09192	2.39	0.0170	0.03802	1.00360
dRT1	1	-0.54555	0.02705	-20.17	<.0001	-0.37356	1.35782
dRT2	1	-1.00384	0.09547	-10.51	<.0001	-0.17080	1.04457
dRG1	1	-0.21915	0.03670	-5.97	<.0001	-0.09594	1.02197
dRG4	1	-0.29765	0.05214	-5.71	<.0001	-0.09228	1.03467

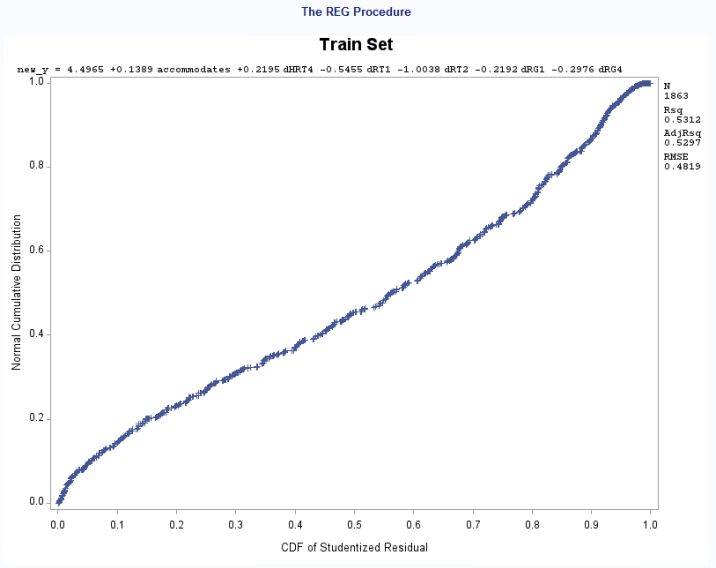
F.15 Regression for Train set (removed dRG2)



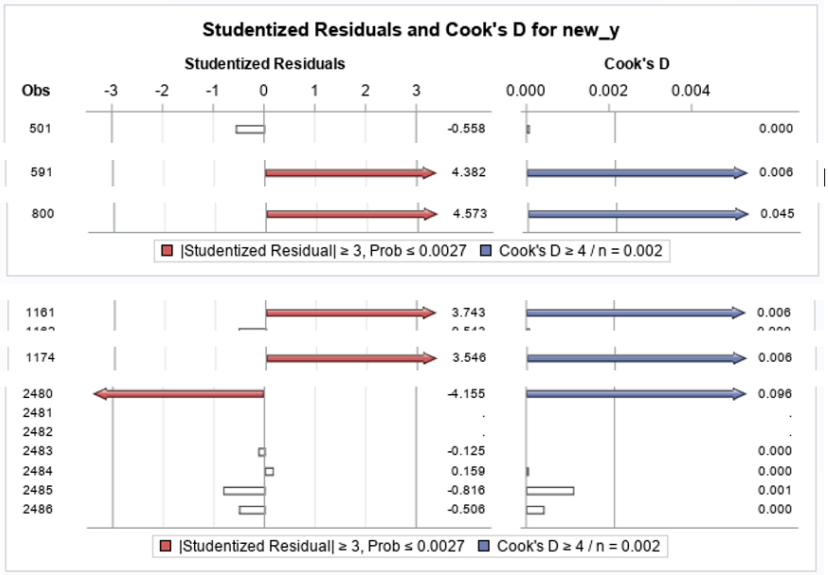
F.16 Residual Plots for Train set (removed dRG2)



F.16 Residual Plots for Train set (removed dRG2)



F.17 Final Model - Normality graph for Train set (removed dRG2)



F.18 Final Model Extract of Studentized Residual and Cook's D

	n	Type	Current#	R Sq		adj R Sq	
	2486			0.5312		0.5297	
1st	2485	O & IF	591	0.5319	↑	0.5304	↑
2nd	2484	O & IF	2261	0.5371	↑	0.5356	↑
3rd	2483	O & IF	2421	0.5395	↑	0.538	↑
4th	2482	O & IF	639	0.5463	↑	0.5448	↑

Train Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2482
Number of Observations Used	1859
Number of Observations with Missing Values	623

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	493.67119	82.27853	371.70	<.0001
Error	1852	409.95928	0.22136		
Corrected Total	1858	903.63047			

Root MSE	0.47049	R-Square	0.5463
Dependent Mean	4.71702	Adj R-Sq	0.5448
Coeff Var	9.97429		

F.19 Outliers and Influential Points removal record and final regression

Validation Statistics for Model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	623	0.46087	0.32980

Validation Statistics for Model

The CORR Procedure

2 Variables:	Inprice yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Inprice	620	4.70826	0.69344	2919	2.89037	7.23201	
yhat	623	4.72007	0.53068	2941	3.31455	6.70972	Predicted Value of new_y

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Inprice	yhat
Inprice	1.00000	0.74743
		<.0001
	620	620
yhat	0.74743	1.00000
Predicted Value of new_y	<.0001	
	620	623

Train	RMSE	0.47049
	R-square	0.5463
	Adj-R sq	0.5448
	GOF	OK
	Residuals	OK
Test	RMSE	0.46087
	MAE	0.3298
	R-square	0.74743
	Adj-R sq	0.7449699
	CV-R sq	0.20113

F.20 Test set vs Train set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2482
Number of Observations Used	1859
Number of Observations with Missing Values	623

Analysis of Variance

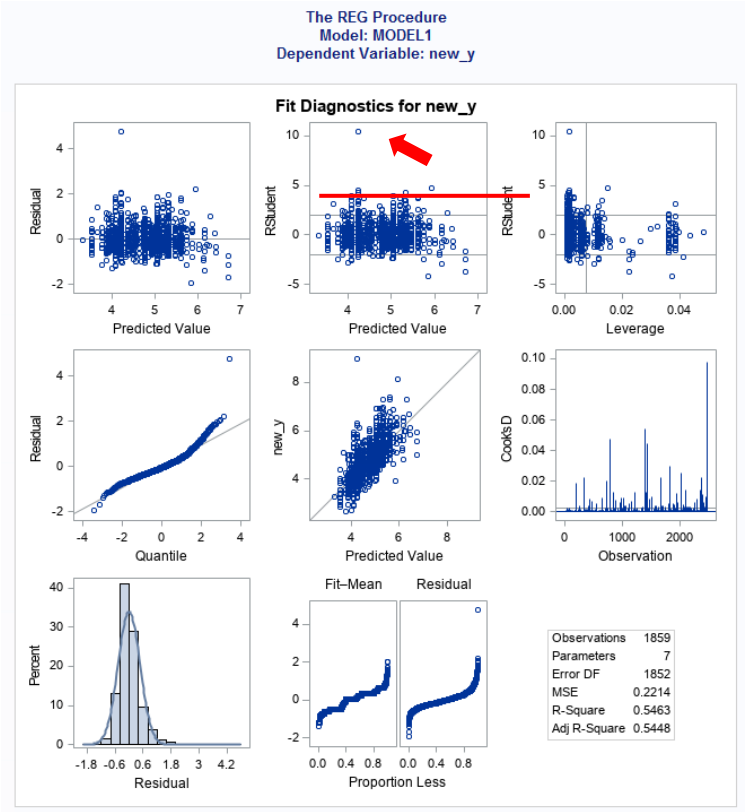
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	493.67119	82.27853	371.70	<.0001
Error	1852	409.95928	0.22136		
Corrected Total	1858	903.63047			

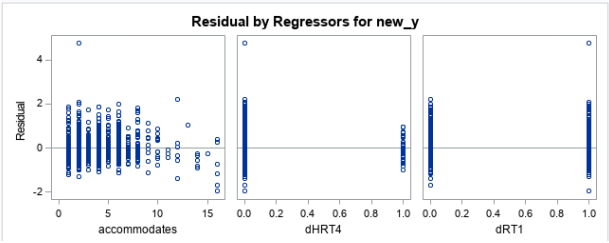
Root MSE	0.47049	R-Square	0.5463
Dependent Mean	4.71702	Adj R-Sq	0.5448
Coeff Var	9.97429		

Parameter Estimates

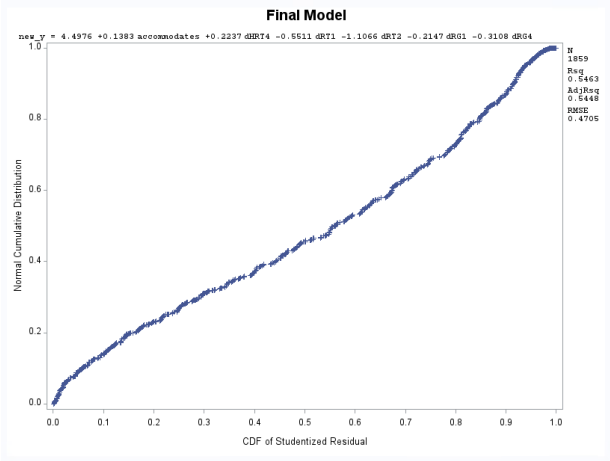
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	4.49757	0.02836	158.60	<.0001	0	0
accommodates	1	0.13826	0.00586	23.58	<.0001	0.42676	1.33724
dHRT4	1	0.22369	0.08975	2.49	0.0128	0.03908	1.00360
dRT1	1	-0.55114	0.02642	-20.86	<.0001	-0.38015	1.35593
dRT2	1	-1.10656	0.09489	-11.66	<.0001	-0.18638	1.04270
dRG1	1	-0.21472	0.03584	-5.99	<.0001	-0.09480	1.02192
dRG4	1	-0.31080	0.05115	-6.08	<.0001	-0.09668	1.03361

F.21 Final Model Regression





F.22 Final Model Residual Plots



F.23 Final Model Normality Graph

Prediction for Final model

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	4.5506	0.1023	4.3500	4.7512	3.6063	5.4949	.
2	.	3.4528	0.0986	3.2595	3.6461	2.5100	4.3956	.
3	5.01	5.0506	0.0144	5.0225	5.0788	4.1274	5.9738	-0.0400
4	4.58	5.0506	0.0144	5.0225	5.0788	4.1274	5.9738	-0.4656
5	5.13	5.0506	0.0144	5.0225	5.0788	4.1274	5.9738	0.0793

F.24 Final Model Predictions