```
* STEP1 - import data from file and create dummy variables;
proc import datafile="Airbnb.csv" out=Airbnb_import replace;
delimiter=',';
getnames=yes;
datarow=2;
run;

proc print;
run;

*Create dummy variables for host_response_time, host_is_superhost, region,
room_type, bed_type, and cancellation_policy;
data Airbnb_import;
set Airbnb_import;

Title "Airbnb Dataset With Recoded Variables";
** Create dummy variable for host_response_time;
** Base = N/A;

* Create dummy variable for "within an hour";
d_within_an_hour=1; *initializing dummy variables for "within an hour";
if (host_response_time='N/A')then d_within_an_hour=0;
if (host_response_time='within a few hours')then d_within_an_hour=0;
if (host_response_time='within a day')then d_within_an_hour=0;
if (host_response_time='a few days or more')then d_within_an_hour=0;

* Create dummy variable for "within a few hours";
d_within_a_few_hours=1; *initializing dummy variables for "within a few
hours";
if (host_response_time='N/A')then d_within_an_hour=0;
if (host_response_time='within an hour')then d_within_a_few_hours=0;
if (host_response_time='within a day')then d_within_a_few_hours=0;
if (host_response_time='a few days or more')then d_within_a_few_hours=0;

* Create dummy variable for "within a day";
d_within_a_day=1; *initializing dummy variables for "within a day";
if (host_response_time='N/A')then d_within_a_day=0;
if (host_response_time='within an hour')then d_within_a_day=0;
if (host_response_time='within a few hours')then d_within_a_day=0;
if (host_response_time='a few days or more')then d_within_a_day=0;

* Create dummy variable for "a few days or more";
d_a_few_days_or_more=1; *initializing dummy variables for "a few days or
more";
if (host_response_time='N/A')then d_a_few_days_or_more=0;
if (host_response_time='within an hour')then d_a_few_days_or_more=0;
if (host_response_time='within a few hours')then d_a_few_days_or_more=0;
if (host_response_time='within a day')then d_a_few_days_or_more=0;

** Create dummy variable for host_is_superhost;
** Base = t;

* Create dummy variable for t;
d_host_is_superhost=1; *initializing dummy variables for t;
if (host_is_superhost='f')then d_host_is_superhost=0;

** Create dummy variable for region;
** Base = IM;

* Create dummy variable for WM;
d_WM=1; *initializing dummy variables for WM;
if (region='IM')then d_WM=0;
if (region='EM')then d_WM=0;
if (region='SEM')then d_WM=0;
if (region='NSM')then d_WM=0;

* Create dummy variable for EM;
d_EM=1; *initializing dummy variables for EM;
if (region='IM')then d_EM=0;
if (region='WM')then d_EM=0;
if (region='SEM')then d_EM=0;
if (region='NSM')then d_EM=0;

* Create dummy variable for SEM;
d_SEM=1; *initializing dummy variables for SEM;
if (region='IM')then d_SEM=0;
if (region='WM')then d_SEM=0;
if (region='EM')then d_SEM=0;
if (region='NSM')then d_SEM=0;

* Create dummy variable for NSM;
d_NSM=1; *initializing dummy variables for NSM;
if (region='IM')then d_NSM=0;
if (region='WM')then d_NSM=0;
if (region='EM')then d_NSM=0;
if (region='SEM')then d_NSM=0;

** Create dummy variable for room_type;
** Base = Entire home/apt;

* Create dummy variable for Private room;
d_private_room=1; *initializing dummy variables for Private room;
if (room_type='Entire home/apt')then d_private_room=0;
if (room_type='Shared room')then d_private_room=0;

* Create dummy variable for Shared room;
d_shared_room=1; *initializing dummy variables for Shared room;
if (room_type='Entire home/apt')then d_shared_room=0;
if (room_type='Private room')then d_shared_room=0;

** Create dummy variable for bed_type;
** Base = Real Bed;

* Create dummy variable for Pull-out Sofa;
d_pullout_sofa=1; *initializing dummy variables for Pull-out sofa;
if (bed_type='Real Bed')then d_pullout_sofa=0;
if (bed_type='Futon')then d_pullout_sofa=0;
if (bed_type='Airbed')then d_pullout_sofa=0;

* Create dummy variable for Futon;
d_futon=1; *initializing dummy variables for Futon;
if (bed_type='Real Bed')then d_futon=0;
if (bed_type='Pull-out Sofa')then d_futon=0;
if (bed_type='Airbed')then d_futon=0;
```

```
* Create dummy variable for Airbed;
d_airbed=1; *initializing dummy variables for Airbed;
if (bed_type='Real Bed')then d_airbed=0;
if (bed_type='Pull-out Sofa')then d_airbed=0;
if (bed_type='Futon')then d_airbed=0;

** Create dummy variable for cancellation_policy;
** Base = Strict;

*Create dummy variable for moderate;
d_moderate=1; *initializing dummy variables for moderate;
if (cancellation_policy='Strict')then d_moderate=0;
if (cancellation_policy='flexible')then d_moderate=0;

*Create dummy variable for flexible;
d_flexible=1; *initializing dummy variables for flexible;
if (cancellation_policy='Strict')then d_flexible=0;
if (cancellation_policy='moderate')then d_flexible=0;

* Create interaction term;
cleaning_fee_d_host_is_superhost = cleaning_fee*d_host_is_superhost;
run;

* STEP-2 : Print dataset;
proc print data=Airbnb_import;
run;

Title "Histogram";
* data = Airbnb_dataset;

proc univariate normal;
var price;
* est - estimate the mean (mu)and s.d (sigma);
histogram / normal (mu=est sigma=est);
qqplot / normal(mu=est sigma=est);
Title "Distribution of price";
run;


* Create frequency table for the variable price;
Title "Frequency Table";
proc freq;
tables price;
run;

/* produces scatterplot matrix */
proc sgscatter;
Title "Scatterplot Matrix for price";
matrix price host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost;
run;

/* produces individual scatterplots */
```

```
proc gplot;
plot price*(host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost);
run;

** Create log price variable in a data step - Airbnb_import;
data Airbnb_import;
* Set command copies original Airbnb_import dataset;
set Airbnb_import;
ln_price=log(Price);

proc print;
run;

Title "Discriptive";
* data = Airbnb_import;

proc means min p25 p50 p75 max;
var ln_price host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost;
run;


/* Produces individual scatterplots */
proc gplot;
plot ln_price* (host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost);
run;

/* Produces scatterplot matrix */
proc sgscatter;
Title "Scatterplot Matrix for ln_price";
matrix ln_price host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost;
run;


Title "Histogram";
* data = Airbnb_import;

proc univariate normal;
var ln_price;
* est = estimate the mean (mu) and the s,d (sigma);
histogram / normal (mu=est sigma=est);
qqplot / normal (mu=est sigma=est);
```

```
Title "Distribution of ln_price";
run;


*Regression Model-1 for ln_price response variable;
proc reg;
*Regression Model-1: Full Model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost;

* Residiual plot : residuals vs x-variables;
plot student.*( host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost);
*Residual plot : residual vs pred. values;
plot student.*predicted.;
*Normal probability plot or QQ plot;
plot npp.*student.;
run;

proc reg;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost / vif tol;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
*Full model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_a_few_days_or_more d_host_is_superhost d_WM d_EM d_SEM d_NSM
d_private_room d_shared_room d_pullout_sofa d_futon d_airbed d_moderate
d_flexible cleaning_fee_d_host_is_superhost;
run;

* Drop a variable or column;
data Airbnb_import_new;
* set command copies original Airbnb_import dataset;
set Airbnb_import;
* remove variable d_a_few_days_or_more;
drop d_a_few_days_or_more;
run;

proc print data = Airbnb_import_new;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
```

```
proc reg corr;
* Reduced model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_host_is_superhost d_WM d_EM d_SEM d_NSM d_private_room
d_shared_room d_pullout_sofa d_futon d_airbed d_moderate d_flexible
cleaning_fee_d_host_is_superhost;
run;

* Drop a variable or column;
data Airbnb_import_new2;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new;
* remove variable d_airbed;
drop d_airbed;
run;

proc print data = Airbnb_import_new2;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_host_is_superhost d_WM d_EM d_SEM d_NSM d_private_room
d_shared_room d_pullout_sofa d_futon d_moderate d_flexible
cleaning_fee_d_host_is_superhost;
run;


* Drop a variable or column;
data Airbnb_import_new3;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new2;
* remove variable d_moderate;
drop d_moderate;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_host_is_superhost d_WM d_EM d_SEM d_NSM d_private_room
d_shared_room d_pullout_sofa d_futon d_flexible
cleaning_fee_d_host_is_superhost;
run;


* Drop a variable or column;
data Airbnb_import_new4;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new3;
* remove variable d_EM;
```

```
drop d_EM;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_host_is_superhost d_WM d_SEM d_NSM d_private_room
d_shared_room d_pullout_sofa d_futon d_flexible
cleaning_fee_d_host_is_superhost;
run;

* Drop a variable or column;
data Airbnb_import_new5;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new4;
* remove variable d_pullout_sofa;
drop d_pullout_sofa;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_host_is_superhost d_WM d_SEM d_NSM d_private_room
d_shared_room d_futon d_flexible cleaning_fee_d_host_is_superhost;
run;

* Drop a variable or column;
data Airbnb_import_new6;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new5;
* remove variable cleaning_fee_d_host_is_superhost;
drop cleaning_fee_d_host_is_superhost;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_host_is_superhost d_WM d_SEM d_NSM d_private_room
d_shared_room d_futon d_flexible;
run;

* Drop a variable or column;
data Airbnb_import_new7;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new6;
* remove variable d_futon;
drop d_futon;
run;
```

```
/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = host_total_listings_count accommodates security_deposit
cleaning_fee review_scores_rating d_within_an_hour d_within_a_few_hours
d_within_a_day d_host_is_superhost d_WM d_SEM d_NSM d_private_room
d_shared_room d_flexible;
run;

* Drop a variable or column;
data Airbnb_import_new8;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new7;
* remove variable host_total_listings_count;
drop host_total_listings_count;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = accommodates security_deposit cleaning_fee
review_scores_rating d_within_an_hour d_within_a_few_hours d_within_a_day
d_host_is_superhost d_WM d_SEM d_NSM d_private_room d_shared_room d_flexible;
run;

* Drop a variable or column;
data Airbnb_import_new9;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new8;
* remove variable d_within_a_day;
drop d_within_a_day;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = accommodates security_deposit cleaning_fee
review_scores_rating 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 d_flexible;
run;

* Drop a variable or column;
data Airbnb_import_new10;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new9;
* remove variable d_flexible;
drop d_flexible;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
```

```
model ln_price = accommodates security_deposit cleaning_fee
review_scores_rating 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;
run;

* Drop a variable or column;
data Airbnb_import_new11;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new10;
* remove variable review_scores_rating;
drop review_scores_rating;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model ln_price = accommodates security_deposit 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;
run;

* Drop a variable or column;
data Airbnb_import_new12;
* set command copies original Airbnb_import dataset;
set Airbnb_import_new11;
* remove variable security_deposit;
drop security_deposit;
run;

/* Regression analysis fitting a linear model
the "corr" option computes a correlation analysis*/
proc reg corr;
* Reduced model;
model 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;
run;

proc reg;
model 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 /vif tol;
run;

* Optimal model analysis with options r, influence, vif, stb;
Title "Optimal Model w/ options";
proc reg data=Airbnb_import_new12;
model 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 /vif r influence stb;
run;

* Regression model-optimal model;
model 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;
```

```
* Residual plot: residual vs x-variables;
plot student.*(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);
* Residual plot: residual vs pred. values;
plot student.*predicted.;
* Normal probability plot or QQ plot;
plot npp.*student.;
run;

* Remove influential points and outliers;
* Writing the new dataset after the deletion into Airbnb_import_new13;
Title "Remove Influential Points and Outliers";
data Airbnb_import_new13;
set Airbnb_import_new12;
if _n_ in (158, 258, 319, 373, 401, 520, 553, 661, 708, 867, 1086, 1107,
1120, 1122, 1180, 1225, 1415, 1718, 1719, 1762, 1953, 2260, 2337)then delete;
run;

proc print;
run;


* Rerun model without outlier and influential points - use new dataset;
proc reg data = Airbnb_import_new13;
model 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
      / influence r;
plot student.*( 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 predicted.);
plot npp.*student.;
run;


data Airbnb_import_new14; * write to a different dataset;
set Airbnb_import_new13;
if _n_ in (27, 223, 228, 318, 435, 520, 566, 657, 687, 929, 1719, 1886, 1968,
2200, 2396)then delete;
run;


* Rerun model without outlier and influential points - use new dataset;
proc reg data = Airbnb_import_new14;
model 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
      / influence r;
plot student.*(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 predicted.);
plot npp.*student.;
run;
```

```
data Airbnb_import_new15; * write to a different dataset;
set Airbnb_import_new14;
if _n_ in (372, 739, 1128, 1927, 1971, 2332)then delete;
run;


* Rerun model without outlier and influential points - use new dataset;
proc reg data = Airbnb_import_new15;
model 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
      / influence r;
plot student.*(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 predicted.);
plot npp.*student.;
run;


data Airbnb_import_new16; * write to a different dataset;
set Airbnb_import_new15;
if _n_ in (1365)then delete;
run;


* Rerun model without outlier and influential points - use new dataset;
proc reg data = Airbnb_import_new16;
model 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
      / influence r;
plot student.*(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 predicted.);
plot npp.*student.;
run;


/* Regression analysis with standardized coefficients*/
proc reg;
model 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
      /stb;
run;

/* Produces correlation matrix */
proc corr;
var 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;
run;



** 5-fold CV with separtae partition for Test Set (Selection method
Stepwise);
```

```
/* Apply 5-fold crossvalidation with Stepwise Selection and 25% of data
removed for testing; */
Title "5-fold crossvalidation + 25% Testing Set";
proc glmselect data=Airbnb_import_new16
plots=(asePlot Criteria);
* Partition defines a test set (25% of data)to validate model on
* new data;
partition fraction(test=0.25);
* selection=stepwise uses stepwise selection method;
* stop=cv: minimizes predictin residual sum of squares for
* variable selection;
model 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/
      selection=stepwise(stop=cv) cvmethod=split(5) cvDetails=all;
run;

** Computes predictions;
data pred;
input 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;
datalines;
. 4 70 1 0 1 1 0 0 1 0
. 6 100 0 1 0 0 0 1 0 1
;
data Airbnb_import_new17;
set pred Airbnb_import;
run;

proc reg;
model 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/p clm cli;
run;
```