Kimberly-Clark Case
Factor Analysis and Cluster Analysis

Andy Huang

Factor Analysis

1. **Factor analyze Diaper Dash brand ratings (Overall_Quality (Q4a), Brand_I_Trust (Q4b), and Brand_I_Recommend (Q4c)). Use principal components with varimax rotation. Interpret and explain the results.**

The purpose of the analysis has to be defined at the beginning. In this case, we will be using *principal component analysis* as the method of our analysis. In principle component analysis, the total variance within our 300 responses will be considered. In addition, principal component analysis allows us to determine the minimum number of factors that will account for the highest possible variance within our 300 responses.

The variables to be analyzed are *Overall_Quality*, *Brand_I_Trust*, and *Brand_I_Recommend*. In other words, the respondents' perspectives on Huggies's overall quality, the respondents' degree of trust toward Huggies, and the respondents' degree of recommendation for Huggies, will all be examined.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Overall Quality	8.10	2.020	293
Brand I Trust	8.23	2.192	293
Brand I Recommend	7.98	2.489	293

The Descriptive Analysis table shows that there are 7 missing values. As a result, we will be using 293 valid responses for further analysis. Although there are 7 missing values, we still have a sample size of 293 responses, so the sample size is large enough to satisfy the assumption of sampling adequacy.

Correlation Matrix

		Overall Quality	Brand I Trust	Brand I Recommend
Correlation	Overall Quality	1.000	.793	.731
	Brand I Trust	.793	1.000	.781
	Brand I Recommend	.731	.781	1.000

The Correlation Matrix table shows the simple correlations between all possible pairs of variables included in our analysis and give us a general idea about our variables. In our case, each of the pair has a correlation value above 0.7, which indicates that each of our variables has strong correlation with each other.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.750
Bartlett's Test of Sphericity	Approx. Chi-Square	587.640
	df	3
	Sig.	.000

Kaiser-Meyer-Olkin Measure of Sampling Adequacy, or KMO, is an optimal approach for us to examine the appropriateness of our analysis. KMO values varied from 0 to 1 – the closer to 1, the better. A KMO value of above 0.5 is to be commonly perceived as appropriate. On the other hand, a KMO value below 0.5 indicates that the analysis may not be appropriate.

In our case, the KMO value is 0.750, as shown in the above table. The KMO value of 0.750 implies that our analysis is appropriate and satisfy the assumption of sampling adequacy.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.750
Bartlett's Test of Sphericity	Approx. Chi-Square	587.640
	df	3
	Sig.	.000

Bartlett's Test of Sphericity is another critical indicator. Bartlett's Test of Sphericity is a test statistic for examining the hypothesis that the variables are uncorrelated in our analysis. In other words, we want to reject this hypothesis, since we want each our variable to be correlated with each other and we do not wish to see an identity correlation matrix. Ideally, we would like our variables – *Overall_Quality*,

Brand_I_Trust, and *Brand_I_Recommend* – to be reduced to a smaller number of components, so it is necessary to have adequate correlations between the variables.

In our case, the p-value is less than 0.05, even less than 0.001, which is statistically significant. Therefore, we can reject the hypothesis. In other words, *Overall_Quality*, *Brand_I_Trust*, and *Brand_I_Recommend* are **not** uncorrelated, and this is what we desired. Bartlett's Test of Sphericity reveals the fact that our data is suitable for data reduction since we have adequate correlations between the variables.

Anti-image Matrices

		Overall Quality	Brand I Trust	Brand I Recommend
Anti-image Covariance	Overall Quality	.339	-.162	-.102
	Brand I Trust	-.162	.284	-.154
	Brand I Recommend	-.102	-.154	.356
Anti-image Correlation	Overall Quality	.765[a]	-.521	-.294
	Brand I Trust	-.521	.710[a]	-.484
	Brand I Recommend	-.294	-.484	.781[a]

a. Measures of Sampling Adequacy(MSA)

The way to interpret Anti-image Correlation is identical to the way we interpret KMO values. Anti-image Correlation provides us the insights by performing KMO on **each variable** while the KMO value of 0.750 mentioned in the previous section came from performing KMO on the variables as a whole.

A value of above 0.5 means it is appropriate to perform factor analysis while a value below 0.5 implies not appropriate. We can see that KMO values of each variable are all above 0.7; therefore, we can conclude that it is appropriate to perform factor analysis on *Overall_Quality*, *Brand_I_Trust*, and *Brand_I_Recommend*.

Communalities

	Initial	Extraction
Overall Quality	1.000	.837
Brand I Trust	1.000	.873
Brand I Recommend	1.000	.828

Extraction Method: Principal Component Analysis.

Communalities provides us the proportion of each variable's variance that can be explained by the factors. Since we have three variables in this case, three Extractions are generated as the result.

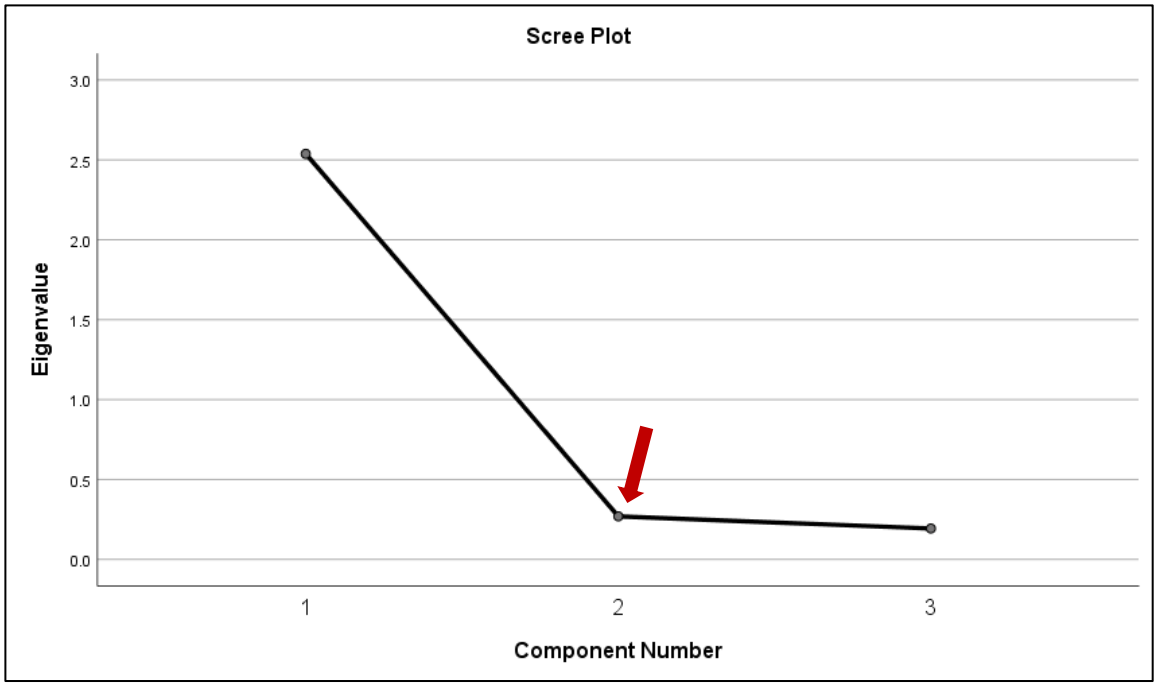
For instance, **83.7%** of *Overall Quality*'s variance can be explained by *Overall Quality*, *Brand I Trust*, and *Brand I Recommend* as a whole. Similarly, **87.3%** of *Brand I Trust*'s variance can be explained by the common factors while **82.8%** of *Brand I Recommend*'s variance can be explained by the common factors. The Extraction values – 0.837, 0.873, and 0.828 – are high, which means *Overall Quality*, *Brand I Trust*, and *Brand I Recommend* are well represented in the common factor space.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.538	84.587	84.587	2.538	84.587	84.587
2	.269	8.974	93.561			
3	.193	6.439	100.000			

Extraction Method: Principal Component Analysis.

Total Variance Explained shows that there is only one factor being retained. Note that since there is only one Component being extracted, the table of *Rotated Sums of Squared Loadings* is not available – the solution cannot be rotated. As the result, we do not have a *Rotated Component Matrix* available in this case. Through rotation, the factor matrix will be transformed into simpler one that is easier to interpret, which is why the rotation is desired. However, even though the rotation is not available, we can still interpret the unrotated factor matrix since there are only three variables in our case. Total Variance Explained shows that **84.587%** of variance are explained, which is very high and is optimal.



A scree plot presents to us the Eigenvalues against the number of factors in order of extraction. These values correspond to the first two column of Total Variance Explained table. The first factor will always capture the most variance, hence always has the highest Eigenvalue, which represents the total variance explained by each factor. The variance captured will become less and less in the successive factors, as we can see the line became almost a flat line from where the arrow points out. In this case, since the flat line starts at Component number 2 (red arrow), a number of one component is recommended.

Component Matrix[a]

	Component
	1
Brand I Trust	.934
Overall Quality	.915
Brand I Recommend	.910

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

Rotated Component Matrix[a]

a. Only one component was extracted. The solution cannot be rotated.

Since there is only one component been extracted, we could not compare between the components. For the same reason, the Rotated Component Matrix is not available. At this point, we could know for sure that there is only one factor in this case.

Reproduced Correlations				
		Overall Quality	Brand I Trust	Brand I Recommend
Reproduced Correlation	Overall Quality	.837[a]	.855	.832
	Brand I Trust	.855	.873[a]	.850
	Brand I Recommend	.832	.850	.828[a]
Residual[b]	Overall Quality		-.061	-.101
	Brand I Trust	-.061		-.069
	Brand I Recommend	-.101	-.069	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 3 (100.0%) nonredundant residuals with absolute values greater than 0.05.

The Reproduced communalities remains the same – 0.837, 0.873, and 0.828. In addition, Residuals are the difference between the observed correlations and the reproduced correlations. The observed correlations are presented above in the previous Correlation Matrix section.

As for the Reproduced Correlation, the values are all increased as presented below:

- The correlation value of the pair of *Overall Quality* and *Brand I Trust* increase from 0.793 to 0.855.
- The correlation value of the pair of *Overall Quality* and *Brand I Recommend* increase from 0.731 to 0.832.
- The correlation value of the pair of *Brand I Trust* and *Brand I Recommend* increase from 0.781 to 0.850.

Conclusion

According to Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO), we know that our data with 293 valid observations satisfied the assumption of sampling adequacy. In addition, the KMO measure for each individual variable shows the same result. According to Bartlett's test of sphericity, we are able to reject the hypothesis that the correlation matrix is an identity matrix, based on the p-value. Bartlett's test of sphericity reveals the fact that there are adequate correlations between *Overall Quality*, *Brand I Trust*, and *Brand I Recommend*, and our data is suitable for data reduction.

From Communalities, we get the insights that *Overall Quality*, *Brand I Trust*, and *Brand I Recommend* are all well represented in the common factor space. **83.7%** of

Overall Quality's variance can be explained by the common factors, **87.3%** of *Brand I Trust*'s variance can be explained by the common factors, and **82.8%** of the variance within *Brand I Recommend* can be explained by the common factors.

Since there is only one component generated, the varimax rotation is not available. However, we do get the insight that **84.587%** of variance are explained, according to Total Variance Explained. 84.587% is very high and is optimal. As another method to confirm the number of component, scree plot also reveals that there should be only one component by showing us a drastically decline in the Eigenvalues and a continued flat line from there.

At this point, we can conclude that our variables – *Overall Quality*, *Brand I Trust*, and *Brand I Recommend* – are all assigned to be within a single factor, and there are adequate correlations between them. In other words, *Overall Quality*, *Brand I Trust*, and *Brand I Recommend* have high similarity. The analysis is conducted appropriately based on the criteria mentioned above. The variance explained by our analysis are high, so the outcome is optimal.

2. **Factor analyze the message ratings (Info_New_Different (Q6), Info_Appropriate (Q7), Info_Believable (Q8), and Info_Understanding (Q9)). Use principal components with varimax rotation. Interpret and explain the results.**

Same with the previous question, we have to define the purpose of the analysis first. In this case, we will still be using *principal component analysis* as the method of our analysis. Principal component analysis allows us to determine the minimum number of factors that will account for the highest possible variance within our 300 responses.

The variables to be analyzed are *Info_New_Different*, *Info_Appropriate*, *Info_Believable*, and *Info_Understanding*. The respondents' perspectives on "the information on the message is new and different," "the information on the message is appropriate for their baby," "the information on the message is believable," and "the information on the message is understandable," will all be examined.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
Information is New & Different	3.14	1.270	299
Information is Appropriate	1.51	.796	299
Information is Believable	1.90	.765	299
Information is Understandable	1.21	.484	299

The Descriptive Analysis table shows that there is one value missing. As a result, we will be using 299 valid responses for further analysis. Although there is one missing value, we still have a sample size of 299 responses, so the sample size is large enough and satisfied the assumption of sampling adequacy.

Correlation Matrix

		Information is New & Different	Information is Appropriate	Information is Believable	Information is Understandable
Correlation	Information is New & Different	1.000	.200	.374	-.044
	Information is Appropriate	.200	1.000	.407	.331
	Information is Believable	.374	.407	1.000	.300
	Information is Understandable	-.044	.331	.300	1.000

The Correlation Matrix table shows the simple correlations between all possible pairs of variables included in our analysis and give us a general idea about our variables. In our case, *Information is Appropriate* and *Information is Believable* has the highest

positive correlation among all available pairs; however, the correlation value is only 0.407, which implies a not very strong positive correlation.
On the other hand, there is a negative correlation between *Information is New & Different* and *Information is Understandable*. If there is an additional increase in one variable, the other variable decrease.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.589
Bartlett's Test of Sphericity	Approx. Chi-Square	156.045
	df	6
	Sig.	.000

Kaiser-Meyer-Olkin Measure of Sampling Adequacy, or KMO, is an optimal approach for us to examine the appropriateness of our analysis. KMO values varied from 0 to 1 – the closer to 1, the better. A KMO value of above 0.5 is to be commonly perceived as appropriate. On the other hand, a KMO value below 0.5 indicates that the analysis may not be appropriate.

In this case, the KMO value is 0.589, as shown in the above table. The KMO value of 0.589 is not a very high value, but it does above the 0.5 standard. Therefore, the KMO value of 0.589 still implies that our analysis is appropriate and satisfy the assumption of sampling adequacy.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.589
Bartlett's Test of Sphericity	Approx. Chi-Square	156.045
	df	6
	Sig.	.000

Bartlett's Test of Sphericity is a test statistic for examining the hypothesis that the variables are uncorrelated in our analysis. We would like our variables – *Info_New_Different*, *Info_Appropriate*, *Info_Believable*, and *Info_Understanding* – to be reduced to a smaller number of components, so it is necessary to have adequate correlations between the variables. In this case, the p-value is less than 0.05, even less than 0.001, which is statistically significant. Therefore, we can reject the hypothesis. In other words, *Info_New_Different*, *Info_Appropriate*, *Info_Believable*, and *Info_Understanding* are **not** uncorrelated, and this is what we desired. Bartlett's Test

of Sphericity reveals the fact that our data is suitable for data reduction since we have adequate correlations between the variables.

Anti-image Matrices

		Information is New & Different	Information is Appropriate	Information is Believable	Information is Understandable
Anti-image Covariance	Information is New & Different	.825	-.083	-.273	.162
	Information is Appropriate	-.083	.778	-.207	-.203
	Information is Believable	-.273	-.207	.700	-.187
	Information is Understandable	.162	-.203	-.187	.824
Anti-image Correlation	Information is New & Different	.505[a]	-.103	-.359	.196
	Information is Appropriate	-.103	.672[a]	-.280	-.254
	Information is Believable	-.359	-.280	.596[a]	-.246
	Information is Understandable	.196	-.254	-.246	.552[a]

a. Measures of Sampling Adequacy(MSA)

The way to interpret Anti-image Correlation is identical to the way we interpret KMO values. Anti-image Correlation provides us the insights by performing KMO on **each variable** while the KMO value of 0.589 mentioned in the previous section came from performing KMO on the variables as a whole.

A value of above 0.5 means it is appropriate to perform factor analysis while a value below 0.5 implies not appropriate. We can see that KMO values of each variable are all above 0.5; therefore, we can conclude that it is appropriate to perform our analysis on *Info_New_Different*, *Info_Appropriate*, *Info_Believable*, and *Info_Understanding*.

Communalities

	Initial	Extraction
Information is New & Different	1.000	.836
Information is Appropriate	1.000	.598
Information is Believable	1.000	.684
Information is Understandable	1.000	.775

Extraction Method: Principal Component Analysis.

Communalities provides us the proportion of each variable's variance that can be explained by the factors. Since we have four variables in this case, four Extractions are generated as the result.

For instance, **83.6%** of the variance within *Information is New & Different* can be explained by *Info_New_Different*, *Info_Appropriate*, *Info_Believable*, and *Info_Understanding* as a whole. Similarly, **59.8%** of the variance within *Info_Appropriate* can be explained by the common factors, **68.4%** of the variance

within *Info_Believable* can be explained by the common factors, and **77.5%** of the variance within *Info_Understanding* can be explained by the common factors. The Extraction values of our variables – 0.836, 0.598, 0.684, and 0.775 – imply that *Information is New & Different, Information is Appropriate, Information is Believable,* and *Information is Understandable* are well represented in the common factor space.

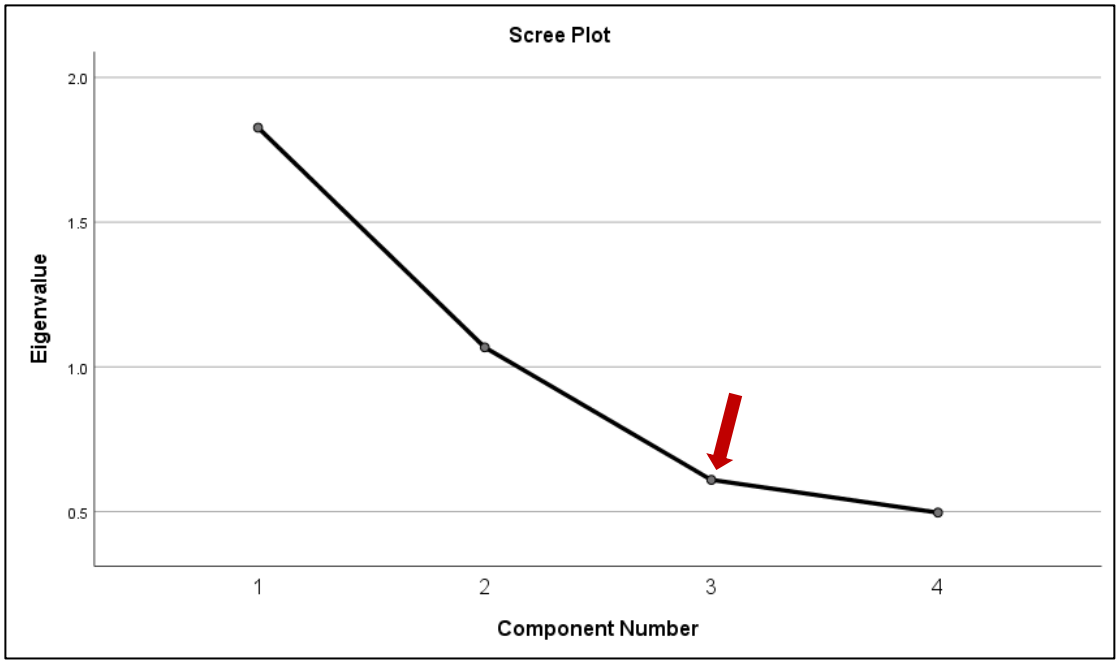
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.826	45.662	45.662	1.826	45.662	45.662	1.534	38.353	38.353
2	1.067	26.687	72.348	1.067	26.687	72.348	1.360	33.996	72.348
3	.609	15.237	87.585						
4	.497	12.415	100.000						

Extraction Method: Principal Component Analysis.

Total Variance Explained shows that there are two factors being retained this time. Unlike Question one, there are two Components being extracted, so the table of *Rotated Sums of Squared Loadings* is available – the solutions have been rotated, which is what we want.

Through rotation, the factor matrix will be transformed into simpler one that is easier to interpret, which is why the rotation is desired. The table of Total Variance Explained shows that **72.348%** of variance are explained, which is high and is optimal. In addition, Component 1 explains 38.353% of the variance, corresponding to the Eigenvalue of 1.826; Component 2 explains 33.996% of the variance, corresponding to the Eigenvalue of 1.067.



A scree plot presents to us the Eigenvalues against the number of factors in order of extraction. These values correspond to the first two column of Total Variance Explained table (the Eigenvalues of 1.826 and 1.067). The first factor will always capture the most variance, hence always has the highest Eigenvalue, which represents the total variance explained by each factor. The variance captured will become less and less in the successive factors, as we can see the line became almost a flat line from where the arrow points out. In this case, since the flat line starts at Component number 3 (red arrow), the number of two components are recommended.

Component Matrix[a]

	Component	
	1	2
Information is Believable	.814	
Information is Appropriate	.754	
Information is New & Different	.520	.752
Information is Understandable	.570	-.671

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Since there are two Components been extracted this time, we are able to compare the strength between the Component 1 and Component 2. It is interesting that *Information is Understandable* holds a positive value in Component 1; however, *Information is Understandable* holds a negative value in Component 2.

Since the Component Matrix is for reference and general ideas, it is reasonable to focus on the Rotated Component Matrix for closer examination of our components. The Rotated Component Matrix is examined later in our analysis.

Reproduced Correlations

		Information is New & Different	Information is Appropriate	Information is Believable	Information is Understandable
Reproduced Correlation	Information is New & Different	.836[a]	.263	.535	-.208
	Information is Appropriate	.263	.598[a]	.588	.545
	Information is Believable	.535	.588	.684[a]	.364
	Information is Understandable	-.208	.545	.364	.775[a]
Residual[b]	Information is New & Different		-.064	-.161	.164
	Information is Appropriate	-.064		-.182	-.214
	Information is Believable	-.161	-.182		-.064
	Information is Understandable	.164	-.214	-.064	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 6 (100.0%) nonredundant residuals with absolute values greater than 0.05.

The Reproduced communalities remains the same – 0.836, 0.598, 0.684, and 0.775. In addition, Residuals are the difference between the observed correlations and the reproduced correlations. The observed correlations are presented above in the previous Correlation Matrix section.

As for the Reproduced Correlation, the values are all increased as presented below:

- The correlation value of the pair of *Information is New & Different* and *Information is Appropriate* **increase** from 0.200 to 0.263.
- The correlation value of the pair of *Information is New & Different* and *Information is Believable* **increase** from 0.374 to 0.535.
- The correlation value of the pair of *Information is New & Different* and *Information is Understandable* **decrease** from -0.044 to -0.208.
- The correlation value of the pair of *Information is Appropriate* and *Information is Believable* **increase** from 0.407 to 0.588.
- The correlation value of the pair of *Information is Appropriate* and *Information is Understandable* **increase** from 0.331 to 0.545.
- The correlation value of the pair of *Information is Believable* and *Information is Understandable* **increase** from 0.300 to 0.364.

From Reproduced Correlations, we can find that the positive pair of correlations are all further strengthened toward the positive end while the only negative correlation is also further strengthened, but to the negative end.

Rotated Component Matrix[a]

	Component	
	1	2
Information is Understandable	.863	
Information is Appropriate	.698	.334
Information is New & Different		.913
Information is Believable	.546	.621

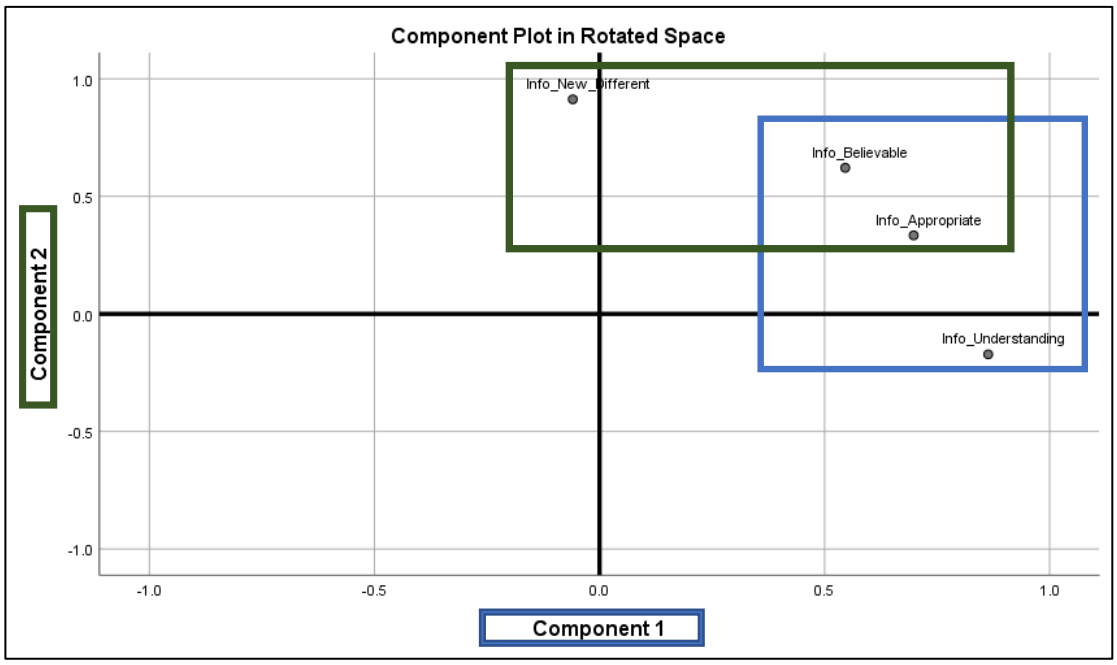
Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

We can see that Component 1 and Component 2 both explained *Information is Appropriate* and *Information is Believable*. Ideally, the situation that two components explain the same variable simultaneously should not exist, since we would not know for sure that the variance within *Information is Appropriate* and *Information is Believable* are explained by Component 1 or Component 2.

The below Component Plot in Rotated Space visualized our concerns. Component 1 is visualized by grouping them in the blue square while Component 2 is visualized by grouping them in the green square. As we can see, *Information is Appropriate* and *Information is Believable* are grouped within the overlapped area of both blue square and green square.



Conclusion

According to Correlation Matrix, we got a general understanding of the correlation between each pair of variables. The Correlations between each pair of variables are not very strong and there is a negative correlation between *Information is New & Different* and *Information is Understandable.*

Based on Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO), we found that our data with 299 valid observations satisfied the assumption of sampling adequacy. In addition, the KMO measure for each individual variable shows the same result. According to Bartlett's test of sphericity, we are able to reject the hypothesis that the correlation matrix is an identity matrix, based on the p-value. Bartlett's test of sphericity reveals the fact that there are adequate correlations between *Information is New & Different, Information is Appropriate, Information is Believable, and Information is Understandable,* hence our data is suitable for data reduction.

From Communalities, we get the insights that *Information is New & Different, Information is Appropriate, Information is Believable, and Information is Understandable* are all well represented in the common factor space. Among all four variables, the variance within I*nformation is New & Different* have been explained the most, there are 83.6% of the variance within *Information is New & Different* are explained. On the other hand, among all four variables, the variance within *Information is Appropriate* has been explained the least, but there are still 59.8% of the variance are explained, which is acceptable.

There are two components being retained this time. **72.348%** of the variance within *Information is New & Different, Information is Appropriate, Information is Believable, and Information is Understandable* are explained. 72.348% is high and is optimal. Component 1 explains 38.353% of the variance, corresponding to the Eigenvalue of 1.826; Component 2 explains 33.996% of the variance, corresponding to the Eigenvalue of 1.067. Furthermore, scree plot also implies that there should be two components been retained in our analysis by showing us a drastically decline in the Eigenvalues and a continued flat line from there.

From Reproduced Correlations, we found that the positive pair of correlations are all further strengthened toward the positive end while the only negative correlation is also further strengthened, but to the negative end.

According to the Rotated Component matrix and the Component Plot in Rotated Space, we can see that an issue arose – Component 1 and Component 2 have both explained *Information is Appropriate* and *Information is Believable*. Ideally, the situation that two components explain the same variable simultaneously should not exist, since we would not know for sure that the variance within *Information is Appropriate* and *Information is Believable* are explained by Component 1 or Component 2.

Nevertheless, since we have only four variables in our analysis, we decided to not remove either component. There are two reasons for this decision:

- Total variance Explained table and Scree Plot are both well indicators with regard to how many components should we retain. Both Total variance Explained table and Scree Plot imply that there should be **two components** generated, rather than only one component.
- The fact that we used only four variables for our analysis imply that we should keep both components, since there are still only two components generated – if we remove one component, there will be only one component left. In order to make our analysis more meaningful, we decided to keep both components.

Component 1 contains *Information is Understandable, Information is Appropriate,* and *Information is Believable*. Component 1 explains 38.353% of the variance. Component 2 contains *Information is New & Different, Information is Appropriate,* and *Information is Believable.* Component 2 explains 33.996% of the variance.

Another insight from our analysis is that *Information is New & Different* and *Information is Understandable* **does not** have high similarity. The analysis is conducted appropriately based on the criteria mentioned above. The variance explained by our analysis are high, so the outcome has certain level of accuracy.

3. **Factor analyze the mailer ratings (High_Quality_Brand (Q10a), Info_is_Informative (Q10b), and Info_I_Want (Q10c)). Use principal components with varimax rotation. Interpret and explain the results.**

This time, the variables to be analyzed are *High_Quality_Brand*, *Info_is_Informative*, and *Info_I_Want*. In other words, the respondents' perspectives on whether or not the product is from a high quality brand, whether or not the mailer is informative, and whether or not the mailer has the information they want to know, will all be examined.

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
High Quality Brand	4.48	.774	299
Information is Informative	4.43	.771	299
Information I Want	4.33	.920	299

The Descriptive Analysis table shows that there are 1 missing value. As a result, we will be using 299 valid responses for further analysis. There is only 1 missing value, so the sample size is large enough to satisfy the assumption of sampling adequacy.

Correlation Matrix

		High Quality Brand	Information is Informative	Information I Want
Correlation	High Quality Brand	1.000	.508	.548
	Information is Informative	.508	1.000	.599
	Information I Want	.548	.599	1.000

The Correlation Matrix table gives us a general idea about our variables. In our case, each of the pair has a correlation value above 0.5, which means that each of our variables has adequate correlation with each other.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.699
Bartlett's Test of Sphericity	Approx. Chi-Square	259.435
	df	3
	Sig.	.000

As mentioned before, KMO values varied from 0 to 1 – the closer to 1, the better. A KMO value of above 0.5 is to be commonly perceived as appropriate. In this case, the KMO value is 0.699, which implies that our analysis is appropriate and satisfy the assumption of sampling adequacy.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.699
Bartlett's Test of Sphericity	Approx. Chi-Square	259.435
	df	3
	Sig.	.000

Bartlett's Test of Sphericity is a test statistic for examining the hypothesis that the variables are uncorrelated in our analysis. Ideally, we would like our variables – *High_Quality_Brand*, *Info_is_Informative*, and *Info_I_Want* – to be reduced to a smaller number of components; therefore, it is necessary to have adequate correlations between the variables.

In this case, the p-value is less than 0.05, even less than 0.001, which means it is statistically significant. Therefore, we can reject the hypothesis. *High_Quality_Brand*, *Info_is_Informative*, and *Info_I_Want* are **not** uncorrelated, and this is the outcome we want. Bartlett's Test of Sphericity reveals the fact that our data is suitable for data reduction since we have adequate correlations between the variables.

Anti-image Matrices

		High Quality Brand	Information is Informative	Information I Want
Anti-image Covariance	High Quality Brand	.649	-.167	-.213
	Information is Informative	-.167	.595	-.257
	Information I Want	-.213	-.257	.561
Anti-image Correlation	High Quality Brand	.739[a]	-.268	-.353
	Information is Informative	-.268	.696[a]	-.445
	Information I Want	-.353	-.445	.671[a]

a. Measures of Sampling Adequacy(MSA)

Anti-image Correlation provides us the insights by performing KMO on **each variable**. A value of above 0.5 means it is appropriate to perform further analysis while a value below 0.5 implies not appropriate. We can see that KMO values of each variable are all above 0.6. Therefore, we can conclude that it is appropriate to perform further analysis on *High_Quality_Brand*, *Info_is_Informative*, and *Info_I_Want*.

Communalities

	Initial	Extraction
High Quality Brand	1.000	.661
Information is Informative	1.000	.705
Information I Want	1.000	.738

Extraction Method: Principal Component Analysis.

Communalities provides us the proportion of each variable's variance that can be explained by the factors. Since we have three variables in this case, three Extractions are generated.

We can see that **66.1%** of *High Quality Brand*'s variance can be explained by *High Quality Brand*, *Information is Informative*, and *Information I Want* as a whole. Similarly, **70.5%** of *Information is Informative*'s variance can be explained by the common factors while **73.8%** of *Information I Want*'s variance can be explained by the common factors. The Extraction values – 0.661, 0.705, and 0.738 – are high, which means *High Quality Brand*, *Information is Informative*, and *Information I Want* are well represented in the common factor space.

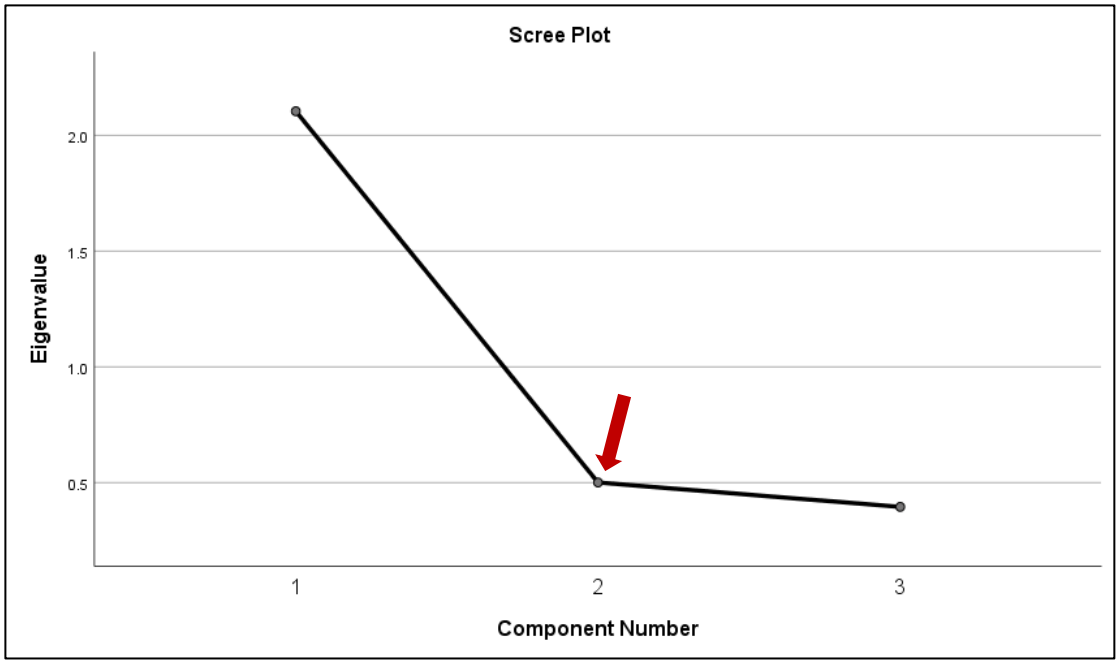
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.104	70.133	70.133	2.104	70.133	70.133
2	.501	16.687	86.821			
3	.395	13.179	100.000			

Extraction Method: Principal Component Analysis.

Total Variance Explained shows that there is only one component being retained. Since there is only one component being extracted, the table of *Rotated Sums of Squared Loadings* is not available – the solution cannot be rotated. Through rotation, the factor matrix will be transformed into simpler one that is easier to interpret, which is why the rotation is desired. In addition, same with question one, we do not have a *Rotated Component Matrix* available this time.

However, even though the rotation is not available, we can still interpret the unrotated factor matrix since there are, again, only three variables in our case. Total Variance Explained shows that **70.133%** of variance are explained, which is high and is optimal.



We could use scree plot to check the number of component since a scree plot presents the Eigenvalues against the number of factors in order of extraction. What's worth mentioning, the Eigenvalues correspond to the first two column of Total Variance Explained table. The first factor will always capture the most variance, hence always has the highest Eigenvalue, which represents the total variance explained by each factor.

The variance captured will become less and less in the successive factors, as we can see the line became almost a flat line from where the arrow points out. In this case, since the flat line starts at Component number 2 (red arrow), a number of one component is recommended by scree plot.

Component Matrix[a]

	Component
	1
Information I Want	.859
Information is Informative	.840
High Quality Brand	.813
Extraction Method: Principal Component Analysis.	
a. 1 components extracted.	

Rotated Component Matrix[a]

a. Only one component was extracted. The solution cannot be rotated.

Since there is only one component, or factor, been extracted, we could not compare the components. For the same reason, the Rotated Component Matrix is not available. At this point, we could know for sure that there is only one factor in this case.

Reproduced Correlations				
		High Quality Brand	Information is Informative	Information I Want
Reproduced Correlation	High Quality Brand	.661[a]	.683	.698
	Information is Informative	.683	.705[a]	.722
	Information I Want	.698	.722	.738[a]
Residual[b]	High Quality Brand		-.175	-.150
	Information is Informative	-.175		-.123
	Information I Want	-.150	-.123	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities

b. Residuals are computed between observed and reproduced correlations. There are 3 (100.0%) nonredundant residuals with absolute values greater than 0.05.

The reproduced communalities remain the same – 0.661, 0.705, and 0.738. In addition, Residuals are the difference between the observed correlations and the reproduced correlations. For observed correlations, they are presented above in the previous Correlation Matrix section.

As for the Reproduced Correlation, the values are all increased as presented below:

- The correlation value of the pair of *High Quality Brand* and *Information is Informative* increase from 0.508 to 0.683.
- The correlation value of the pair of *High Quality Brand* and *Information I Want* increase from 0.548 to 0.698.
- The correlation value of the pair of *Information is Informative* and *Information I Want* increase from 0.599 to 0.722.

Conclusion

According to Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO), we know that our data with 299 valid observations satisfied the assumption of sampling adequacy. In addition, the KMO measure for each individual variable shows the same result. According to Bartlett's test of sphericity, we are able to reject the hypothesis that the correlation matrix is an identity matrix, based on the p-value. Bartlett's test of sphericity reveals the fact that there are adequate correlations between *High Quality Brand*, *Information is Informative*, and *Information I Want*, hence our data is suitable for data reduction.

From Communalities, we get the insights that *High Quality Brand*, *Information is Informative*, and *Information I Want* are all well represented in the common factor

space. **66.1%** of the variance within *High Quality Brand* can be explained by the common factors, **70.5%** of the variance within *Information is Informative* can be explained by the common factors, and **73.8%** of the variance within *Information I Want* can be explained by the common factors.

Since there is only one component generated, the varimax rotation is not available. However, we do get the insight that **70.133%** of the variance within *High Quality Brand*, *Information is Informative*, and *Information I Want* are explained. 70.133% is high and is optimal. Furthermore, scree plot also implies that there should be only one component in our analysis by showing us a drastically decline in the Eigenvalues and a continued flat line from there.

Finally, we can conclude that our variables - *High Quality Brand*, *Information is Informative*, and *Information I Want* – are all assigned to be within a single factor, and there are adequate correlations between them. In other words, *High Quality Brand*, *Information is Informative*, and *Information I Want* have high similarity. The analysis is conducted appropriately based on the criteria mentioned above. The variance explained by our analysis are high, so the outcome is optimal.

Cluster Analysis

1. **Cluster the respondents based on message rating (Info_New_Different (Q6), Info_Appropriate (Q7), Info_Believable (Q8), and Info_Understanding (Q9)). Interpret the results.**

The set of variables selected should generally describe the similarity between objects, and the objects are relevant to the marketing research problem. In this case, we will be analyzing the respondents' perspectives on the message rating – information is new and different, information is appropriate, information is believable, and information is understandable.

It is recommended to use both hierarchical clustering and nonhierarchical clustering to compare the results, in order to reach a better decision of clustering. Therefore, we will be using Ward's method as hierarchical clustering and K-Means clustering as nonhierarchical clustering. Finally, a Two-Step clustering will be performed to test the final results.

Hierarchical clustering
Ward's method

Case Processing Summary[a,b]

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
299	99.7	1	.3	300	100.0

a. Squared Euclidean Distance used

b. Ward Linkage

There is one missing value within our dataset. Therefore, we will be using 299 valid responses for further analysis. Our sample size with 299 responses are large enough to satisfy the assumption of sampling adequacy.

Agglomeration Schedule

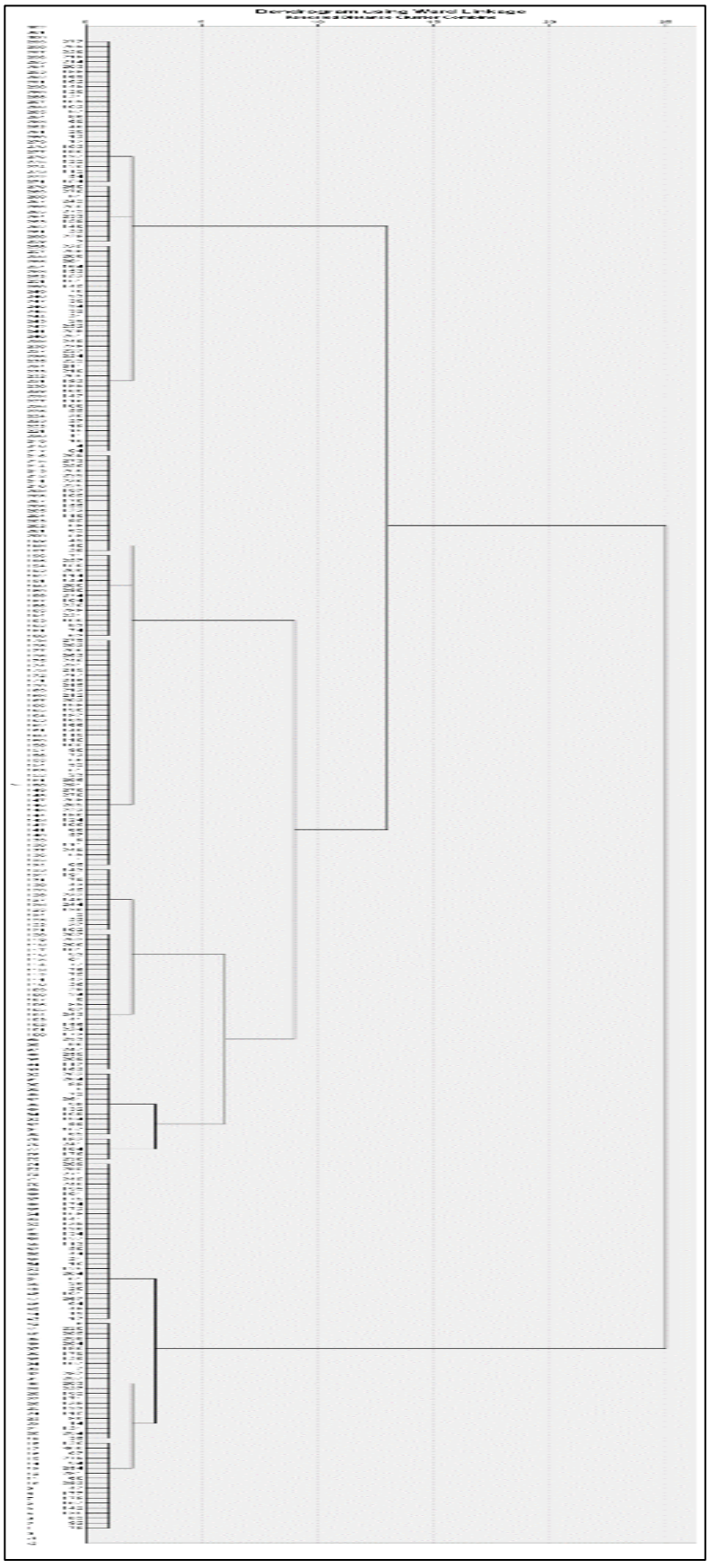
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	277	299	.000	0	0	21
2	285	298	.000	0	0	14
3	287	297	.000	0	0	12
4	235	296	.000	0	0	58
5	289	295	.000	0	0	10
6	60	293	.000	0	0	242
7	207	292	.000	0	0	81
8	286	291	.000	0	0	13
9	280	290	.000	0	0	18
10	271	289	.000	0	5	27
11	226	288	.000	0	0	63
12	283	287	.000	0	3	16
13	278	286	.000	0	8	20
14	284	285	.000	0	2	15
15	186	284	.000	0	14	101
16	237	283	.000	0	12	56
17	275	282	.000	0	0	23
18	274	280	.000	0	9	24
19	67	279	.000	0	0	241

290	1	38	216.841	287	210	296
291	3	12	236.839	289	279	297
292	6	47	257.300	278	282	294
293	5	176	284.829	280	286	295
294	6	16	313.321	292	283	298
295	2	5	377.561	288	293	296
296	1	2	478.266	290	295	297
297	1	3	624.920	296	291	298
298	1	6	913.458	297	294	0

An Agglomeration Schedule listed the processing stages as how the cluster combinations formed. More specifically, an agglomeration schedule provides us the information on the objects or cases being combined at each stage of the hierarchical process. For instance, the blue square indicates that response #277 and response #299 are clustered. From the red square, we can see that, at stage 10, response #271 is combined with Stage 5, where the response #289 is. The value “5” (red arrow) indicated that response #271 is combined with Stage 5. In addition, at the last stage, Stage 298, all the responses in our dataset are combined, as shown in the green square. By checking the Agglomeration Schedule, we can know that all the responses are properly combined and that we would be able to perform detail examination for the stages, if needed.

Cluster Membership			
Case	4 Clusters	3 Clusters	2 Clusters
1	1	1	1
2	2	1	1
3	3	2	1
4	3	2	1
5	2	1	1
6	4	3	2
7	4	3	2
8	2	1	1
9	1	1	1
10	3	2	1
11	3	2	1
12	3	2	1
13	3	2	1
14	4	3	2
15	3	2	1
16	4	3	2
17	4	3	2
18	3	2	1
19	4	3	2
20	4	3	2
21	4	3	2
22	2	1	1
23	1	1	1
24	2	1	1
25	3	2	1

Cluster Membership indicates the cluster to which each object or case belongs. For instance, if there are four clusters generated, all the cases (responses) will be assigned to the value of 1, 2, 3, or 4, representing each cluster. If there are three clusters generated, all the cases will be assigned to the value of 1, 2, or 3, representing each cluster. Likewise, if there are only two clusters created, all the cases will be assigned to the value of 1 or 2, representing each cluster. We have 299 cases in our dataset, but we only show cases 1 to 25 here, for simplicity.



The image presented here is called dendrogram, or tree graph. A dendrogram is a graphical device for displaying clustering results and it's a great way to visualize the number of clusters. Vertical lines represent clusters that are joined together. The distances at which clusters are joined are indicated by the position of the line on the scale. From our dendrogram, we could see that it seems like there are four big clusters formed.

Non-Hierarchical Clustering
(K-Means Clustering)

To assess the reliability and validity of the cluster analysis, we decided to use different methods of clustering and compare the results. K-means cluster is useful to quickly cluster large datasets. One difference between k-means clustering and hierarchical clustering is that the researcher would have to define the number of clusters in advance, when using k-means. In addition, k-means is useful to test different models with a different assumed number of clusters, this function of k-means turned out to be beneficial to us as our analysis went on.

Four Clusters

From Ward's Method, it seemed that there are four clusters generated, so we decided to define the number of clusters to be **four**.

Distances between Final Cluster Centers

Cluster	1	2	3	4
1		4.698	2.909	3.631
2	4.698		2.039	2.504
3	2.909	2.039		1.973
4	3.631	2.504	1.973	

Final Cluster Centers

	Cluster			
	1	2	3	4
Information is New & Different	5	2	3	5
Information is Appropriate	4	1	3	1
Information is Believable	5	2	2	2
Information is Understandable	2	1	1	1

The Euclidean distances between the final cluster centers are presented above. The greater the distances between clusters, the greater the dissimilarities between clusters. Although the relationships between the clusters can also be intuited from the final cluster centers, it would become more difficult as the number of clusters and variables increases. According to the above table, we can conclude that Cluster 1 and 2 are most different, corresponding to the value of 4.698. On the other hand, Cluster 3 and 4 are most similar, corresponding to the value of 1.973.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Information is New & Different	119.497	3	.415	295	288.176	.000
Information is Appropriate	35.225	3	.282	295	125.113	.000
Information is Believable	20.639	3	.381	295	54.230	.000
Information is Understandable	3.686	3	.199	295	18.532	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

The ANOVA table provides us the insights that which variables contribute the most to our cluster results. Variables with large *F* -values provide the greatest separation between clusters. *Information is New & Different* has the highest F-values of 288.176, which indicate that *Information is New & Different* has the greatest separation between clusters. *Information is Believable* and *Information is Understandable* own a small F-values of 54.230 and 18.532, which indicate low separations between clusters, with *Information is Understandable* being the lowest.

Number of Cases in each Cluster

Cluster	1	5.000
	2	159.000
	3	45.000
	4	90.000
Valid		299.000
Missing		1.000

From the above table, we can see the number of cases within each cluster. Among all our 299 cases (responses), there are 159 cases been classified to Cluster 2, 90 cases been classified to Cluster 4, and 45 cases been classified to Cluster 3. However, there are only 5 cases been classified to Cluster 1, make the number of cases within number Cluster 1 significantly lower than other clusters.

In order to make our analysis more meaningful and have the number of cases in each cluster not being so drastically different, we decide to rerun k-means clustering, except this time we will predefine the number of clusters to be **three**.

K-Means Clustering

Three Clusters

Distances between Final Cluster Centers

Cluster	1	2	3
1		2.203	2.880
2	2.203		2.005
3	2.880	2.005	

Final Cluster Centers

	Cluster 1	Cluster 2	Cluster 3
Information is New & Different	2	4	4
Information is Appropriate	1	1	3
Information is Believable	2	2	3
Information is Understandable	1	1	2

The Euclidean distances between the final cluster centers are presented above. The greater the distances between clusters, the greater the dissimilarities between clusters. From the Final Cluster Centers, we can also sense the relationships between the clusters. “Cluster 1 and 2” & “Cluster 2 and 3” seem to be equally similar, each pair correspond to the value of 2.203 and 2.005. On the other hand, Cluster 1 and 3 became most different this time, corresponding to the value of 2.880, which is slightly larger than 2.203 and 2.005, but not too big a difference.

The respondents’ responses that have been classified to Cluster 3 generally seem to care on relatively high levels of all *Information is New & Different*, *Information is Appropriate*, *Information is Believable*, and *Information is understandable*.

The responses that have been classified to Cluster 1 and that of Cluster 2 are similar, but these two groups of respondents differ on their perspectives on *Information is New & Different*. Compare to Cluster 1, the responses within Cluster 2 emphasize more on *Information is New & Different.*

ANOVA

	Cluster Mean Square	Cluster df	Error Mean Square	Error df	F	Sig.
Information is New & Different	153.603	2	.587	296	261.889	.000
Information is Appropriate	51.407	2	.290	296	177.113	.000
Information is Believable	21.771	2	.441	296	49.328	.000
Information is Understandable	3.451	2	.212	296	16.259	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

The ANOVA table provides us the insights that which variables contribute the most to our cluster results. Variables with large *F* -values provide the greatest separation between clusters. *Information is New & Different* still has the highest F-values of 261.889 this time, which indicate that *Information is New & Different* has the greatest

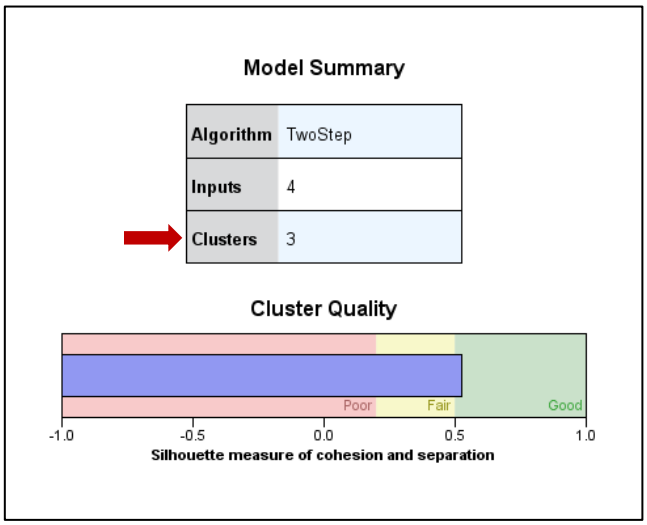
separation between clusters. *Information is Believable* and *Information is Understandable* still own a small F-values of 49.328 and 16.259, which indicate low separations between clusters. *Information is Understandable* has the lowest separation between clusters.

Number of Cases in each Cluster

Cluster	1	104.000
	2	155.000
	3	40.000
Valid		299.000
Missing		1.000

From the above table, we can see the number of cases within each cluster. Among all our 299 cases (responses), there are 155 cases been classified to Cluster 2, 104 cases been classified to Cluster 1, and 40 cases been classified to Cluster 3. Unlike the last time when we predefined the number of clusters to be four, there is no extreme small number of cases been classified to an individual cluster this time, which is optimal, and make our analysis more meaningful, since even the least number of cluster contains 40 cases in it.

Two-Step Cluster



Two-step cluster analysis is a combined method of hierarchical and non-hierarchical clustering. Two-step cluster identifies clusters by running pre-clustering first and then by running hierarchical methods. Two-step cluster is suitable for large dataset that

would take longer to run with hierarchical clustering methods, since Two-step cluster uses a quick cluster algorithm. Because of above reasons, Two-step cluster analysis is a great way to test our results about the number of clusters.

Two-step cluster analysis automatically suggested that the number of clusters should be **three**, which matches our previous result. Furthermore, the cluster quality seems decent, especially for our dataset with 299 responses. At this point, we can finally report that three clusters are being created.

Conclusion

Unlike the results when we produced four clusters, there are no significantly low number of cases been classified to any individual cluster when we generated **three** clusters by using k-means clustering. By running the Two-step cluster analysis at last, we have tested our result of having three clusters for our dataset based on our four variables.

Cluster 2 have the highest number of cases. Cluster 1 contains 104 cases (**34.78%**), Cluster 2 contains 155 cases (**51.83%**), and Cluster 3 contains 40 cases (**13.37%**).

As for the characteristics of each cluster, the respondents' responses that have been classified to Cluster 3 generally seem to care on relatively high levels of all *Information is New & Different*, *Information is Appropriate*, *Information is Believable*, and *Information is understandable*.

In addition, the responses that have been classified to Cluster 1 and that of Cluster 2 are similar, but these two groups of respondents differ on their perspectives on *Information is New & Different*. Compare to Cluster 1, the responses within Cluster 2 emphasize more on *Information is New & Different.*