FACTOR ANALYSIS AND CLUSTERING CASE STUDY

University of North Texas

Marissa McKee
ADTA 5120
4/6/2020

Table of Contents

Executive Summary

The purpose of this report is to describe the findings of factor analysis and K means clustering analysis for three clients. Client 1 is a mobile phone company that wants to determine the different clusters of customers it has based on survey data. Client 2 is a university that would like to revise a questionnaire. Client 3 is a manager who would like the determine segments of university students in order to market to different groups.

The major points of this report are noted in the modeling section where the K means clustering model and factor analysis are explained. The results of clustering the data provided by client 1 is proposed as 5 segments of customers. The data provided by client 3 is segmented into 4 clusters based on family income and graduation rate. The data provided by client 2 is grouped by 7 factors that explains more than half the variance.

Business Understanding

The first client, client X, is a mobile phone company that wants to determine the different types of customer clusters it has based on their attributes of service. Client X has provided the results of a survey of 250 customers. The results of the survey are based off of 24 different satisfaction criteria on a scale of 1 to 10. 1 being the lowest and 10 being the highest.

The second client, client Y, is the University of North Texas. UNT wants to revise its Teaching of Statistics questionnaire which is based on Bland’s theory that good researchers should have four characteristics. The characteristics include a profound love of statistics, an enthusiasm for experimental design, a love of teaching, and a complete absence of interpersonal skills. UNT has provided the results of the questionnaire which was given to 239 research methods lecturers around the world.

The third and final client, client Z, is a manager in the education industry that would like to determine how his firm should segment colleges and universities into distinct marketing segments.

Data Understanding and Preparation

Client X

Client X supplied the results of a survey of 250 customers. The results of the survey are based on 24 different satisfaction criteria on a scale of 1 to 10. The 24 satisfaction criteria are listed below:

- QualityExp
- MeetNeedsExp
- GoWrongExp
- OverallSat
- Fulfilled
- IsIdeal
- ComplaintHandling
- BuyAgain
- SwitchForPrice
- Recommend
- Trusted
- Stable
- Responsible
- Concerned
- Innovative
- OverallQuality
- NetworkQuality
- CustomerService
- ServiceQuality
- RangeProdServ
- Reliability
- ClearInfo
- FairPrice
- GoodVal

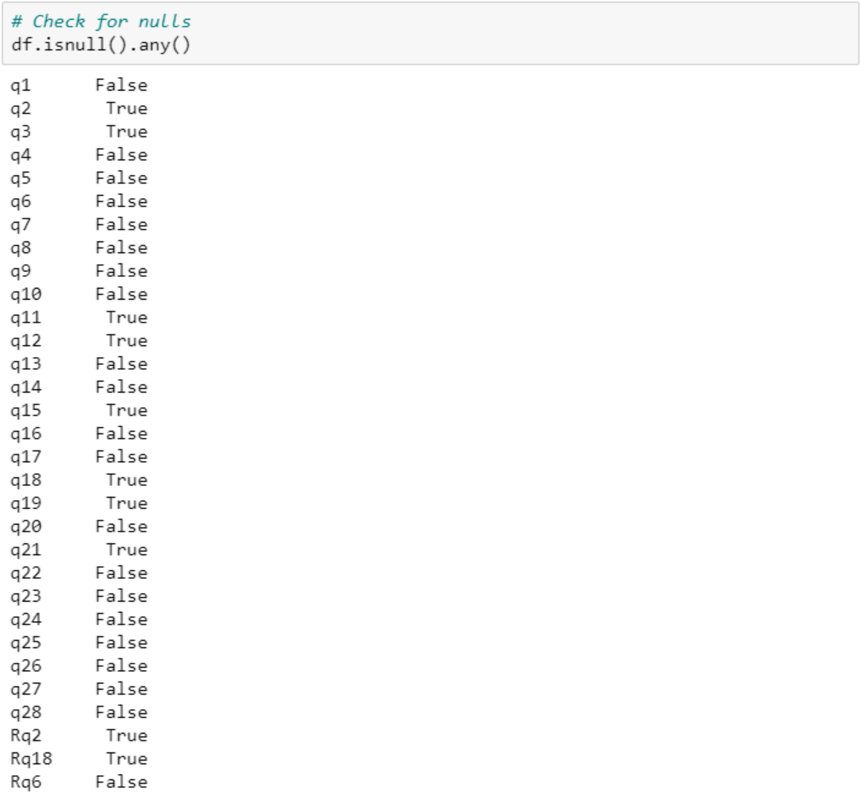
To clean the data provided, I used IBM SPSS and Jupyter Notebook. In Jupyter Notebook, I checked for nulls and found none of the columns had empty fields. In SPSS, I changed each column to a scale variable. Scale variables are a type of measurement numeric variable, like and integer or float data type in python.

Client Y

Client Y has provided the results of the questionnaire which was given to 239 research methods lecturers around the world. The columns of the questionnaire are below.

q1
q2
q3
q4
q5
q6
q7
q8
q9
q10
q11
q12
q13
q14
q15
q16
q17
q18
q19
q20
q21
q22
q23
q24
q25
q26
q27
q28
Rq2
Rq18
Rq6

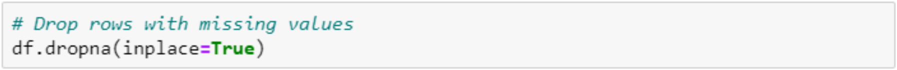
In Jupyter Notebook I checked for nulls and found many columns had empty fields.



```
# Check for nulls
df.isnull().any()
```

```
q1      False
q2       True
q3       True
q4      False
q5      False
q6      False
q7      False
q8      False
q9      False
q10     False
q11      True
q12      True
q13     False
q14     False
q15      True
q16     False
q17     False
q18      True
q19      True
q20     False
q21      True
q22     False
q23     False
q24     False
q25     False
q26     False
q27     False
q28     False
Rq2      True
Rq18     True
Rq6     False
```

In order to use factor analysis the null values need to be handled. The code below drops the rows where nulls exist.



```
# Drop rows with missing values
df.dropna(inplace=True)
```

Client Z

Client Z has provided college data to determine how his firm should segment colleges and universities into distinct marketing segments. The columns provided is below.

FamilyIncome

PctFtFaculty

Ownership

SATAVG

GradRate

FacultySalary

AdmissionRate

PCTFedLoan

UndergradPop

LoanCurrent3yr

DEBT

CostToAttend

PartTime_students

Income_8yr

Similar to client X I used IBM SPSS and Jupyter Notebook. In Jupyter Notebook for preprocessing the data. I checked for nulls and found none of the columns had empty fields. In SPSS, I changed each column to a scale variable.

Modeling

I used a K means clustering model for the analysis of the data provided by client X and client Z. For Client Y, I used factor analysis to analyze the data provided. The code and SPSS files can be found here on my GitHub account.

K Means Cluster Analysis Client X and Client Z

Clustering is an unsupervised learning technique used for pattern detection. Clustering is useful when you don't know exactly what you're looking for. It is most often used in EDA. Clustering is a good way to split the data before analyzing deeper into the segments that behave differently. There are several clustering algorithms like K means, agglomerative, and divisive to name a few. For client X and client Z, I used K means clustering to segment the data.

The K means clustering algorithm is an iterative clustering algorithm that partitions the data into K clusters. The centroid of each cluster is the arithmetic mean of all the data points that belong to that specific cluster. The centroid is initialized by shuffling the data and randomly selecting K data points. When the assignment of data points to clusters has stopped changing, then it's time to complete the iterating of the K means algorithm.

Performing K Means Clustering Analysis

Client X and client Z were analyzed similarly. For the K means clustering model I used SPSS. Clustering is a classification problem. In SPSS under classify I chose the K means clustering option. The scree plots created in Jupyter Notebook determined the number of clusters I chose for both clients. Refer to Figure 1 and Figure 17 below. The K means analysis was used to cluster client X and client Z with 3, 4, 5, and 6 clusters. Those results can be found in the appendix or here. For the most optimal results, 5 clusters were chosen for client X and 4 clusters were chosen for client Z.

Figure 1: Client X Clusters Scree Plot - K Means Analysis

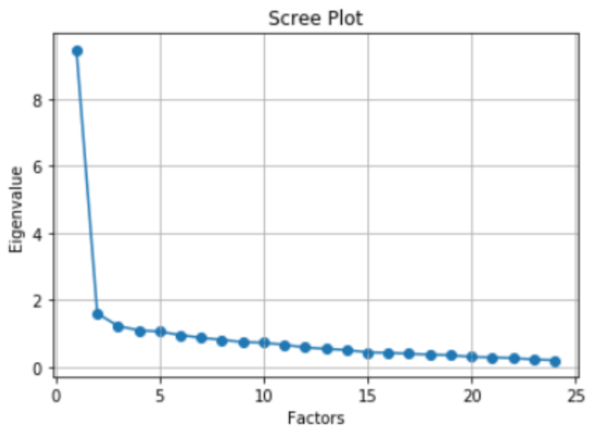
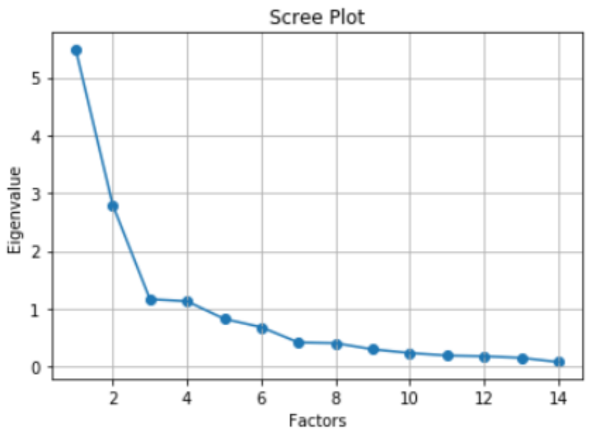


Figure 17: Client Z Clusters Scree Plot - K Means Analysis



The maximum iterations were set to 40 and the cluster membership, ANOVA table, distance from cluster center, and initial cluster center options were selected in SPSS to analyze. The K means algorithm iterated over 9 times for client X and 24 times for client Z.

Factor Analysis Client Y

Factor analysis is a linear statistics model. It’s used to explain the variance among the observed variables and condenses a set of the observed variables into the unobserved variables called factors. Observed variables are modeled as a linear combination of factors and error terms. Factor variables are associated with multiple observed variables who have common patterns of responses. Factor analysis helps in data interpretations by reducing the number of variables.

A factor is a latent variable which describes the association among the number of observed variables. The maximum number of factors are equal to a number of observed variables. Every factor explains a certain variance in observed variables.

There are four main assumptions when using factor analysis:

1. There are no outliers in the data
2. Sample size should be greater than the factor
3. There should not be perfect multicollinearity
4. There should not be homoscedasticity between the variables

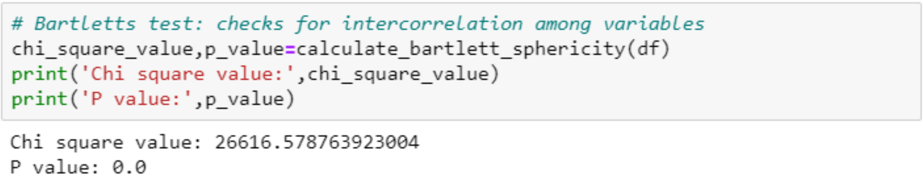
There are a couple types of factor analysis, exploratory and confirmatory. Exploratory factor analysis is when any observed variable is directly associated with a factor. Confirmatory factor analysis is when each factor is associated with a particular set of observed variables.

The primary objective of factor analysis is to reduce the number of observed variables and find unobservable variables. These unobserved variables help analysts to conclude the surveys. This conversion of the observed variables to unobserved variables can be achieved in two steps:

- Factor extraction: The number of factors for extraction selected using partitioning methods such as principal components analysis.
- Factor rotation: Rotation converts factors into uncorrelated factors. The main goal of this step is to improve the overall interpretability. There are several rotation methods that are available such as varimax, quartimax, and promax. For this assignment I’ll be using varimax.

Adequacy Test

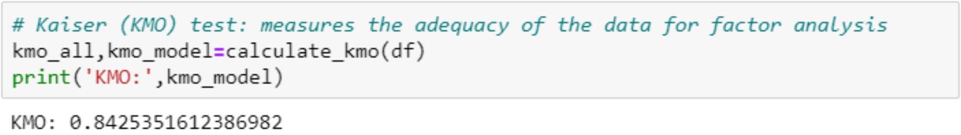
Before performing a factor analysis, the factorability of the data needs to be evaluated. To do this I'll use the Bartlett's test and the Kaiser Meyer Olkin (KMO) test. Bartlett's test of sphericity checks whether the observed variables intercorrelate using the observed correlation matrix against the identity matrix.



```
# Bartletts test: checks for intercorrelation among variables
chi_square_value,p_value=calculate_bartlett_sphericity(df)
print('Chi square value:',chi_square_value)
print('P value:',p_value)
```

```
Chi square value: 26616.578763923004
P value: 0.0
```

The p value is statistically significant, indicating that the observed correlation matrix is not and identity matrix. The KMO test measures the suitability of the data for factor analysis. It determines the adequacy for each observed variable for the complete model. KMO estimates the proportion of variance among the observed variables.



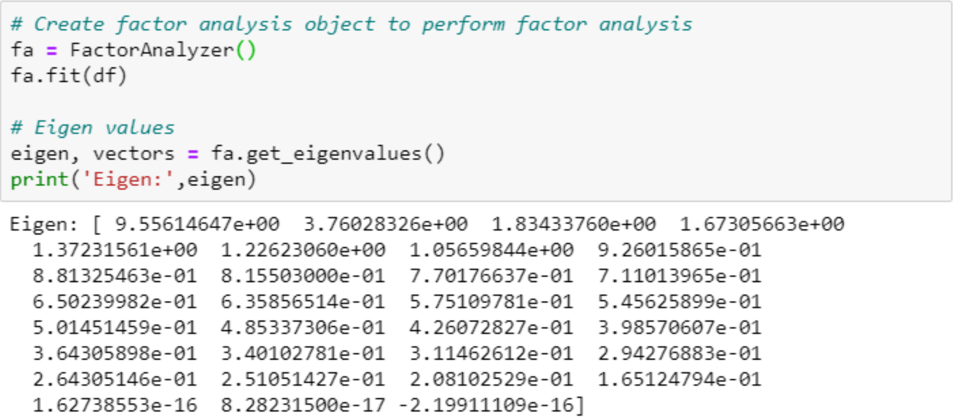
```
# Kaiser (KMO) test: measures the adequacy of the data for factor analysis
kmo_all,kmo_model=calculate_kmo(df)
print('KMO:',kmo_model)
```

```
KMO: 0.8425351612386982
```

Adequate KMO values range from 0 to 1. A KMO value of less than 0.6 is considered inadequate. The overall KMO for the data is 0.8425. This value indicates that a factor analysis for the client's data is appropriate.

Choosing the Number of Factors

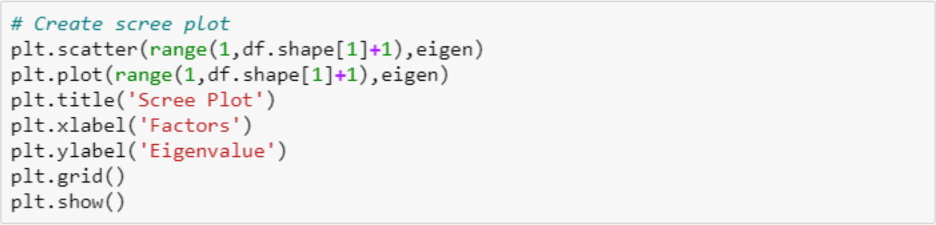
For choosing the number of factors I used the scree plot which is based off of eigenvalues. Eigenvalues represent the variance explained by each factor from the total variance. From the eigenvalues below, there are 7 factors that have eigenvalues greater than 1. The Kaiser criterion states that the number of factors to be extracted should be equal to the number of factors having an eigen value of 1 or greater than 1. This means we should have 7 factors.



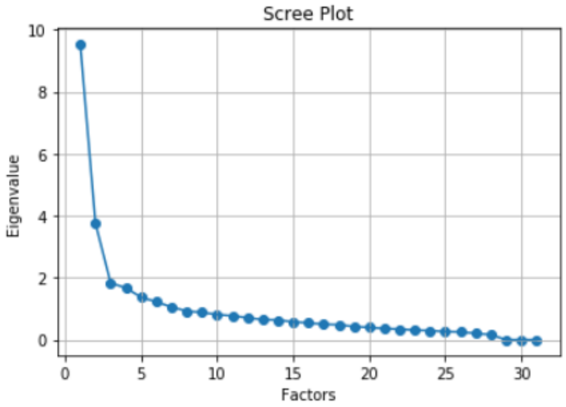
```
# Create factor analysis object to perform factor analysis
fa = FactorAnalyzer()
fa.fit(df)

# Eigen values
eigen, vectors = fa.get_eigenvalues()
print('Eigen:',eigen)
```

```
Eigen: [ 9.55614647e+00  3.76028326e+00  1.83433760e+00  1.67305663e+00
  1.37231561e+00  1.22623060e+00  1.05659844e+00  9.26015865e-01
  8.81325463e-01  8.15503000e-01  7.70176637e-01  7.11013965e-01
  6.50239982e-01  6.35856514e-01  5.75109781e-01  5.45625899e-01
  5.01451459e-01  4.85337306e-01  4.26072827e-01  3.98570607e-01
  3.64305898e-01  3.40102781e-01  3.11462612e-01  2.94276883e-01
  2.64305146e-01  2.51051427e-01  2.08102529e-01  1.65124794e-01
  1.62738553e-16  8.28231500e-17 -2.19911109e-16]
```

The scree plot is visual representation of the total variance associated with each factor. The steep slope shows the large factors. Gradually trailing off lower than 1 are the factors that don't matter as much. We want to find the elbow of the plot, where the eigenvalues are greater than 1. Again, that value is the 7th plot from the left representing 7 factors.

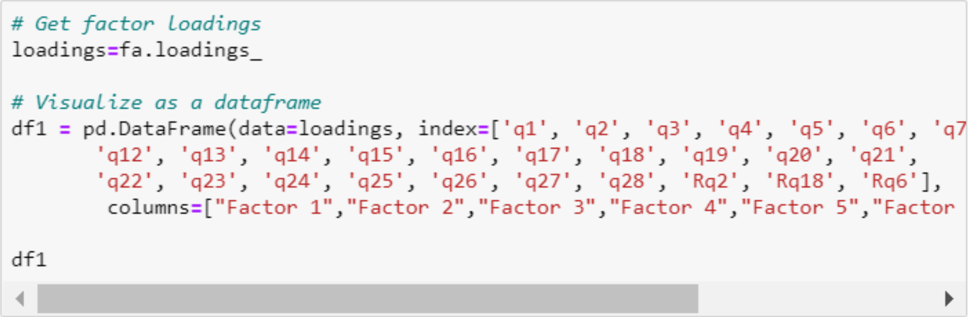


```
# Create scree plot
plt.scatter(range(1,df.shape[1]+1),eigen)
plt.plot(range(1,df.shape[1]+1),eigen)
plt.title('Scree Plot')
plt.xlabel('Factors')
plt.ylabel('Eigenvalue')
plt.grid()
plt.show()
```



Performing Factor Analysis

The rotation of the factors determines how you want to rotate your factors. Rotation maximizes the high loadings and minimizes the low item loadings by shifting the factors themselves. Rotation is a tool for better interpretation of factor analysis and can be classified as orthogonal or oblique. It redistributes the commonalities with a clear pattern of loadings. Commonalities are the sum of the squared loadings for each variable. It represents the common variance and ranges from 0-1. Values closer to 1 represents more variance. For the analysis of client Y's data I'm using a varimax rotation. The varimax rotational method is a type of orthogonal rotation where factors are independent from each other. The varimax rotation maximizes the squared loadings across variables.



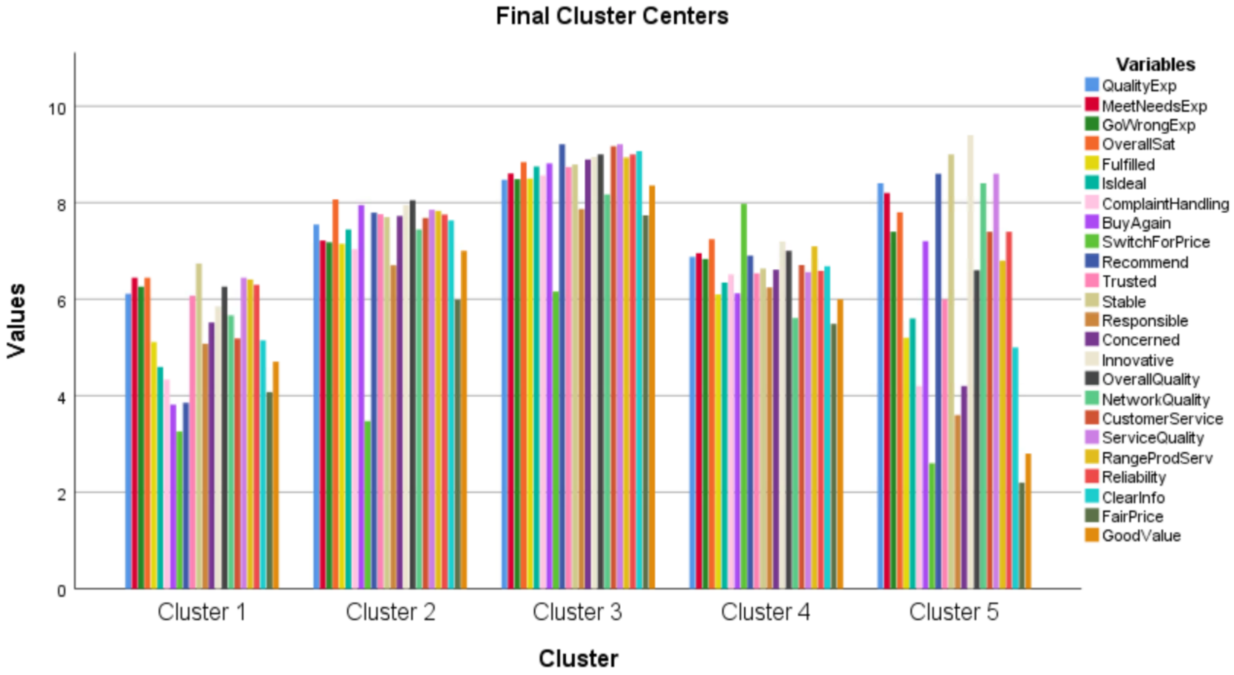
```
# Get factor loadings
loadings=fa.loadings_

# Visualize as a dataframe
df1 = pd.DataFrame(data=loadings, index=['q1', 'q2', 'q3', 'q4', 'q5', 'q6', 'q7
       'q12', 'q13', 'q14', 'q15', 'q16', 'q17', 'q18', 'q19', 'q20', 'q21',
       'q22', 'q23', 'q24', 'q25', 'q26', 'q27', 'q28', 'Rq2', 'Rq18', 'Rq6'],
        columns=["Factor 1","Factor 2","Factor 3","Factor 4","Factor 5","Factor

df1
```

Evaluation

Client X Evaluation

The final cluster centers for client X are depicted below in Figure 12. All 5 clusters are positive and this is because the data provided had survey takers select options from 1 to 10. Cluster 1 consistently has the lowest scores. This group would not recommend client X and is unhappy with their service and products. However cluster 1 won't switch for price either out of loyalty or necessity. Cluster 2 represents an average customer. They believe they're getting a fair price for the quality of services and products. Cluster 3 has the highest satisfaction ratings. These customers are happy with their service and are loyal to their brand. They will be buying more products or services in the future and believe they're getting a great deal. Cluster 4 represents another average customer however they have a higher tendency to switch for price than cluster 2. These customers may believe the services and products are too expensive and are not brand loyal. Cluster 5 is quite a mixed bag. The customers in this cluster believe the products are reliable, the customer service is good, and that the network quality is great. They don't believe they're getting a fair price or good value. These customers seem to have had bad experiences with the company in regards to the handling of complaints. Refer to Figure 33 in the Appendix. Cluster 5 has the smallest amount of cases. This may be a reason why the experience is less organized than the other clusters. This group is certainly annoyed with their products but has the potential to become happier with client X.

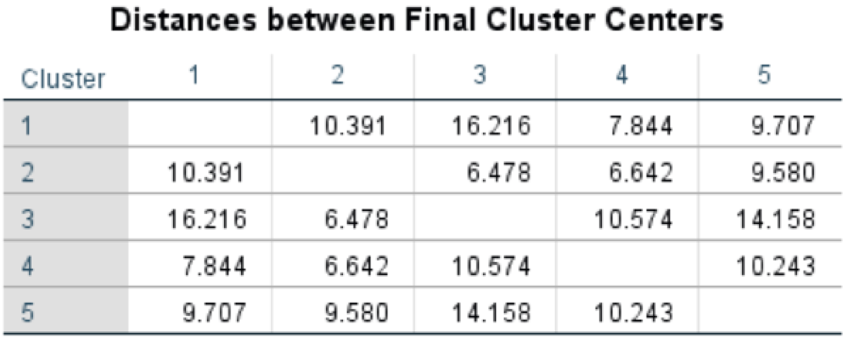
Figure 12: Client X 5 Clusters Bar Graph - K Means Analysis



Client X Evaluation

Refer to Figure 13 below. Another way to analyze the if the clusters are a good fit is to measure the distance between final clusters. Clustering is dependent on the distance between points. Cluster 1 is least similar with cluster 3. Cluster 5 is least similar with cluster 3. Cluster 3 has the highest satisfaction ratings among the 5 clusters while cluster 1 and 5 have the lowest satisfaction ratings among the 5 clusters.

Figure 13: Client X 5 Clusters Distance Between Cluster Centroids- K Means Analysis



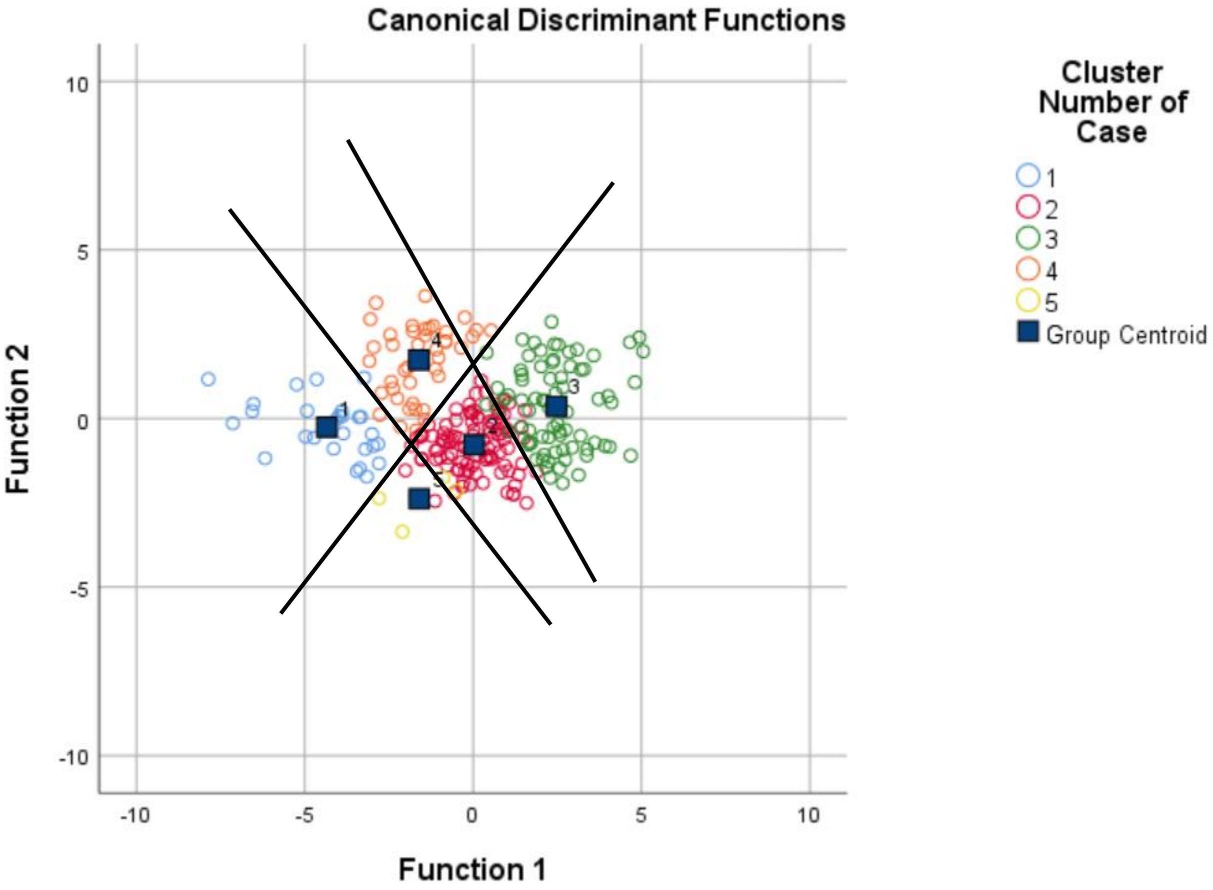
Distances between Final Cluster Centers

Cluster	1	2	3	4	5
1		10.391	16.216	7.844	9.707
2	10.391		6.478	6.642	9.580
3	16.216	6.478		10.574	14.158
4	7.844	6.642	10.574		10.243
5	9.707	9.580	14.158	10.243	

Client X Evaluation

Refer to Figure 15 below. The discriminant plot below is defined on the range of grouping variables from 1 to 5. Missing values are replaced with the mean, although there should not be any missing from the dataset. Each observation is plotted and the centroids of the different clusters are denoted by small squares. This graph represents the qualitative judgment of distance between the clusters. I've added lines to depict the distinct clusters. The graph shows little overlap among the 5 clusters.

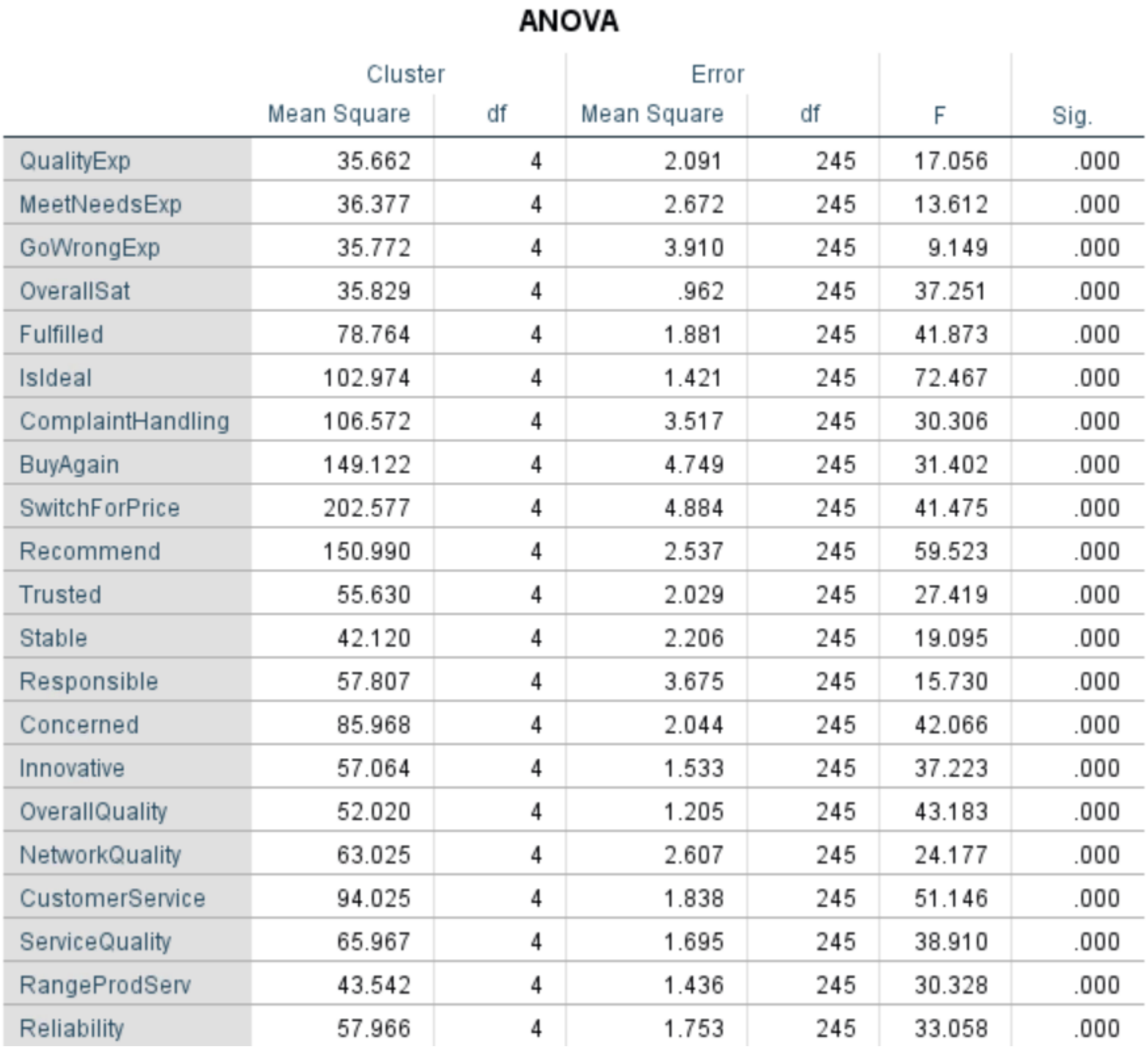
Figure 15: Client X 5 Clusters Discriminant Plot- K Means Analysis



Client X Evaluation

Refer to Figure 16 below. The ANOVA table displays that all the variables have a statistically significant impact in determining which cluster a variables is segmented into

Figure 16: Client X 5 Clusters ANOVA - K Means Analysis



ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
QualityExp	35.662	4	2.091	245	17.056	.000
MeetNeedsExp	36.377	4	2.672	245	13.612	.000
GoWrongExp	35.772	4	3.910	245	9.149	.000
OverallSat	35.829	4	.962	245	37.251	.000
Fulfilled	78.764	4	1.881	245	41.873	.000
IsIdeal	102.974	4	1.421	245	72.467	.000
ComplaintHandling	106.572	4	3.517	245	30.306	.000
BuyAgain	149.122	4	4.749	245	31.402	.000
SwitchForPrice	202.577	4	4.884	245	41.475	.000
Recommend	150.990	4	2.537	245	59.523	.000
Trusted	55.630	4	2.029	245	27.419	.000
Stable	42.120	4	2.206	245	19.095	.000
Responsible	57.807	4	3.675	245	15.730	.000
Concerned	85.968	4	2.044	245	42.066	.000
Innovative	57.064	4	1.533	245	37.223	.000
OverallQuality	52.020	4	1.205	245	43.183	.000
NetworkQuality	63.025	4	2.607	245	24.177	.000
CustomerService	94.025	4	1.838	245	51.146	.000
ServiceQuality	65.967	4	1.695	245	38.910	.000
RangeProdServ	43.542	4	1.436	245	30.328	.000
Reliability	57.966	4	1.753	245	33.058	.000

Client Y Evaluation

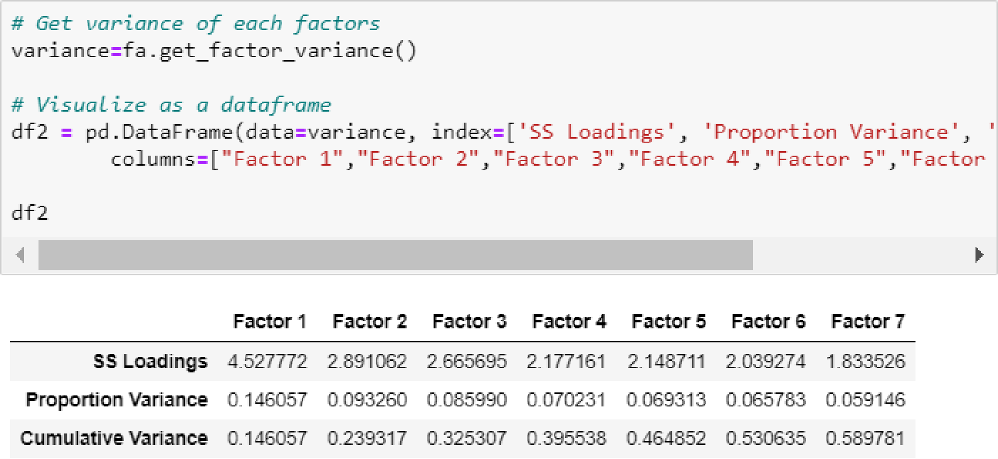
The results of the factor analysis are below in a heatmap.

- Factor 1 has high factor loadings for q8, q13, q14, q16, q17, and q22.
- Factor 2 has high factor loadings for q19, q20, q25, q26, and q27.
- Factor 3 has high factor loadings for q3, q4, and q24.
- Factor 4 has high factor loadings for q2 and Rq2.
- Factor 5 has high factor loadings for q6 and Rq6.
- Factor 6 has high factor loadings for q18 and Rq18.
- Factor 7 has high factor loadings for q5, q23, and q28.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
q1	0.454876	-0.04494	0.497364	0.268611	0.07931	0.165094	0.223624
q2	0.145316	-0.02821	0.110149	0.929068	0.257803	0.162026	0.081091
q3	0.198855	0.369092	0.557367	0.040595	0.032559	-0.02767	0.04391
q4	0.165964	0.290389	0.558037	0.112586	0.105114	0.193621	0.072879
q5	0.254544	-0.00059	0.073997	0.141238	0.037248	0.062257	0.597914
q6	0.129107	-0.04641	0.081892	0.228787	0.952264	0.106164	0.036394
q7	0.230995	0.488324	0.33603	0.036574	-0.01732	-0.07486	0.009306
q8	0.595434	0.368429	0.362865	0.104425	0.051439	0.117461	0.09857
q9	0.476171	0.062111	0.456661	0.216881	0.150917	0.227111	0.180151
q10	0.3956	0.092963	0.107497	0.134775	0.238808	0.181561	0.128139
q11	0.374129	0.442203	0.278666	0.051581	0.001177	-0.05579	0.305632
q12	0.022692	0.111797	-0.05807	-0.08612	-0.01174	0.086549	0.292441
q13	0.559411	0.181691	0.344331	-0.02752	0.097756	0.082266	0.0698
q14	0.708584	0.139129	-0.10705	0.089237	0.058993	0.086973	0.293111
q15	0.422967	-0.00626	0.368343	0.130893	0.016198	0.005074	0.184952
q16	0.699399	0.28016	0.033142	0.004526	-0.03812	0.015473	0.145037
q17	0.7151	0.192433	0.273484	0.08804	0.120081	0.177788	0.054093
q18	0.207894	-0.06636	0.10503	0.166892	0.124524	0.913115	0.233058
q19	-0.02518	0.604951	0.019166	-0.08542	-0.04965	-0.04586	-0.1015
q20	0.128313	0.542353	0.180767	0.002292	0.051073	0.077742	0.06801
q21	0.492315	0.243677	0.426101	0.017792	0.026825	0.007239	0.058296
q22	0.778154	0.101851	0.313268	0.100257	0.113141	0.139875	0.09314
q23	0.094627	0.034686	0.198536	0.165852	0.082085	0.163332	0.568654
q24	0.18368	0.282064	0.636805	0.003142	0.045903	0.045559	0.15051
q25	0.171533	0.609519	0.124269	0.03201	0.016986	-0.00848	0.175596
q26	0.336316	0.57348	0.087951	-0.10176	-0.10399	-0.05681	0.189021
q27	0.165659	0.599348	0.168225	0.077403	-0.02305	-0.02162	0.408155
q28	0.15238	0.144484	0.185974	0.017562	0.015441	0.07902	0.549187
Rq2	-0.14532	0.028211	-0.11015	-0.92907	-0.2578	-0.16203	-0.08109
Rq18	-0.20789	0.066355	-0.10503	-0.16689	-0.12452	-0.91312	-0.23306
Rq6	-0.12911	0.046408	-0.08189	-0.22879	-0.95226	-0.10616	-0.03639

Client Y Evaluation

Evaluation of the reliability of the final factors is below. Factor 1 explains the largest proportion, 15% of the variance. The 7 factors explain 59% of the variance.



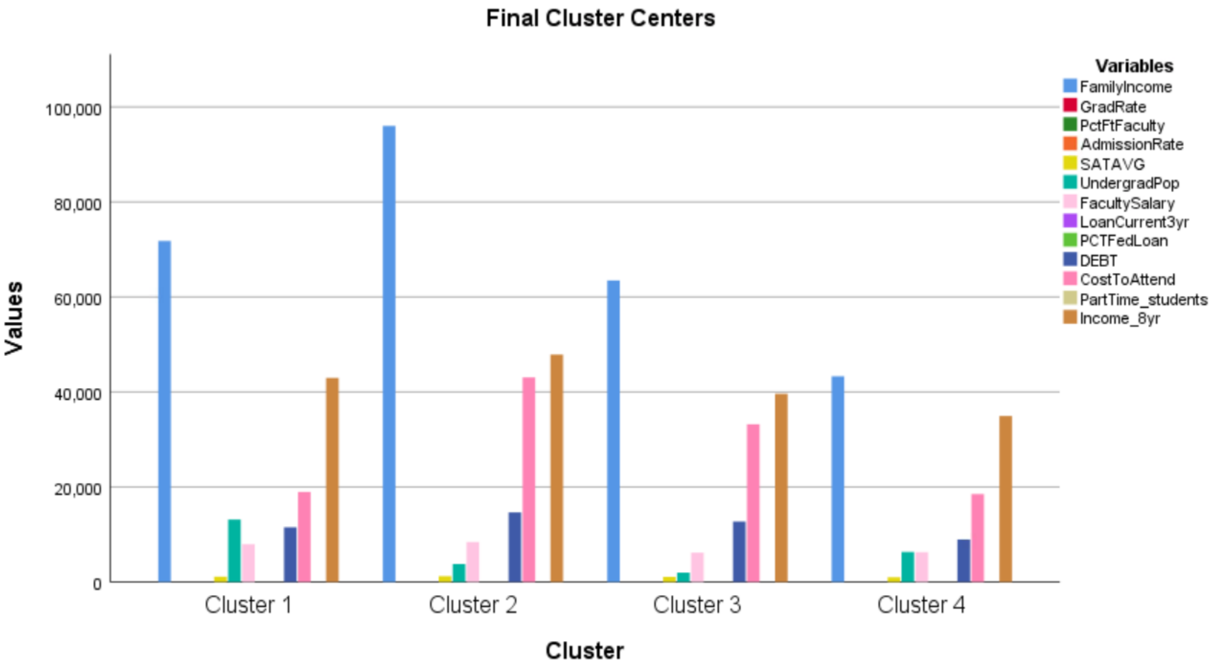
```
# Get variance of each factors
variance=fa.get_factor_variance()

# Visualize as a dataframe
df2 = pd.DataFrame(data=variance, index=['SS Loadings', 'Proportion Variance', '
        columns=["Factor 1","Factor 2","Factor 3","Factor 4","Factor 5","Factor

df2
```

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
SS Loadings	4.527772	2.891062	2.665695	2.177161	2.148711	2.039274	1.833526
Proportion Variance	0.146057	0.093260	0.085990	0.070231	0.069313	0.065783	0.059146
Cumulative Variance	0.146057	0.239317	0.325307	0.395538	0.464852	0.530635	0.589781

Client Z Evaluation

The final cluster centers for client X are depicted below in Figure 26 and Figure 35 in the Appendix. Cluster 1 could be categorized as having high family income placing them in the upper middle class with high graduation rates. This group attends a school where the cost is relatively low, possibly a public school. They have similar statistics to cluster 3 but will end up with less debt. Cluster 2 has the highest family income and graduation rate. This segment could be defined as wealthy and educated. This cluster attends highly coveted schools with low admission rates and high salaries for their staff. Most of these students need to take out a loan and will accrue the most debt. Cluster 3 has slightly lower family income and graduation rate than cluster 1 and a slightly higher family income and graduation rate than cluster 4. Cluster 3 could be segmented as lower middle class with an average graduation rate. The cost to attend is high, possibly meaning the students are attending a private school. Cluster 4 has the lowest family income, graduation rate, SAT average, and debt. This segment also has the highest part time student rate. This group could be considered as struggling to stay in school.

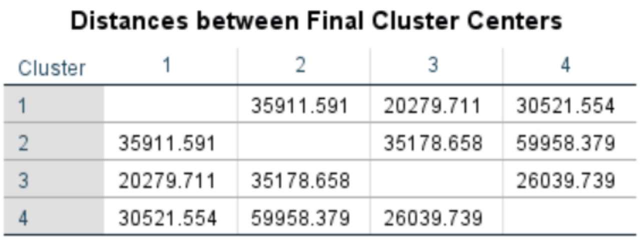
Figure 26: Client Z 4 Clusters Bar Graph - K Means Analysis



Client Z Evaluation

Refer to Figure 27 below. Another way to analyze the if the clusters are a good fit is to measure the distance between final clusters. Clustering is dependent on the distance between points. Cluster 1 is most similar to cluster 3 and least similar to cluster 2. Cluster 4 is least similar with cluster 2. As we saw earlier cluster 4 has the lowest graduation rate and income level while cluster 2 has the highest graduation rate and family income.

Figure 27: Client Z 4 Clusters Distance Between Cluster Centroids- K Means Analysis



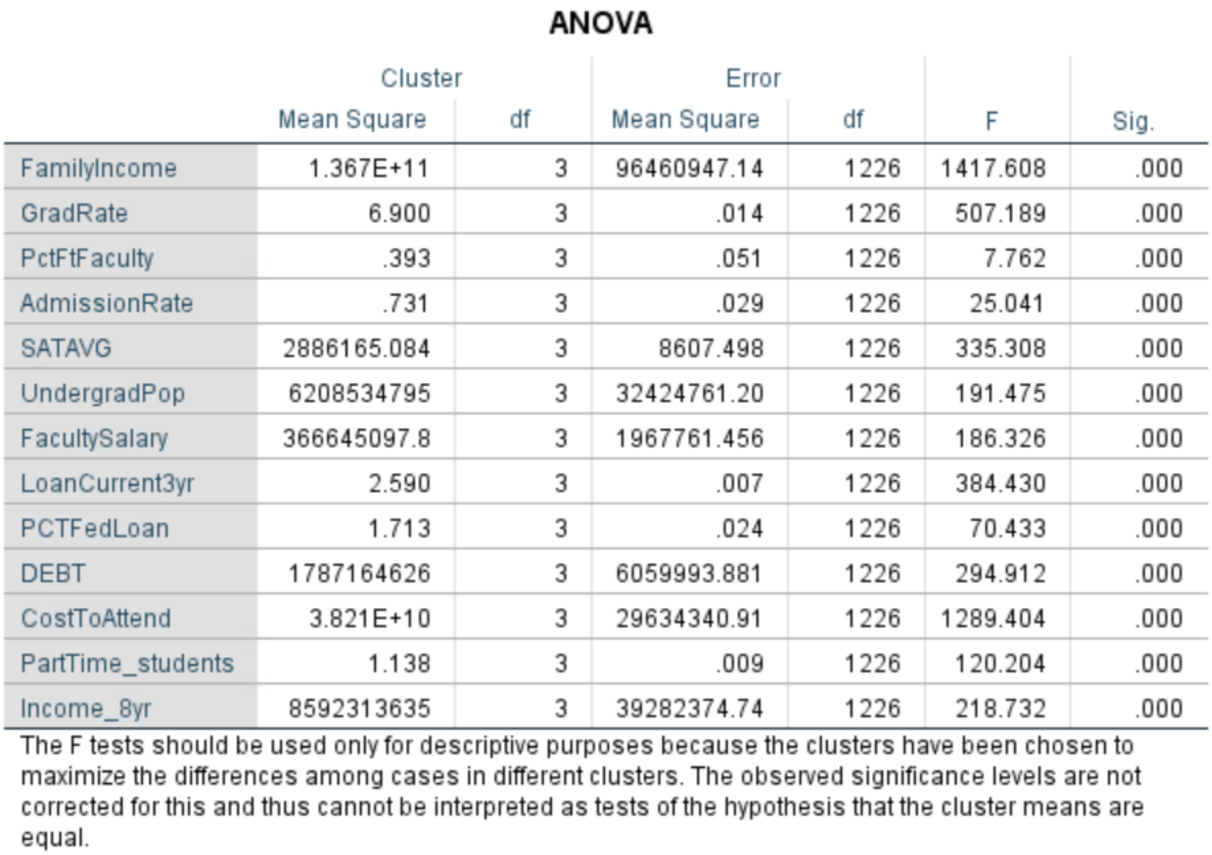
Distances between Final Cluster Centers

Cluster	1	2	3	4
1		35911.591	20279.711	30521.554
2	35911.591		35178.658	59958.379
3	20279.711	35178.658		26039.739
4	30521.554	59958.379	26039.739	

Client Z Evaluation

Refer to Figure 23 below. The ANOVA table displays that all the variables have a statistically significant impact in determining which cluster a variables is segmented into

Figure 23: Client Z 4 Clusters ANOVA - K Means Analysis



ANOVA

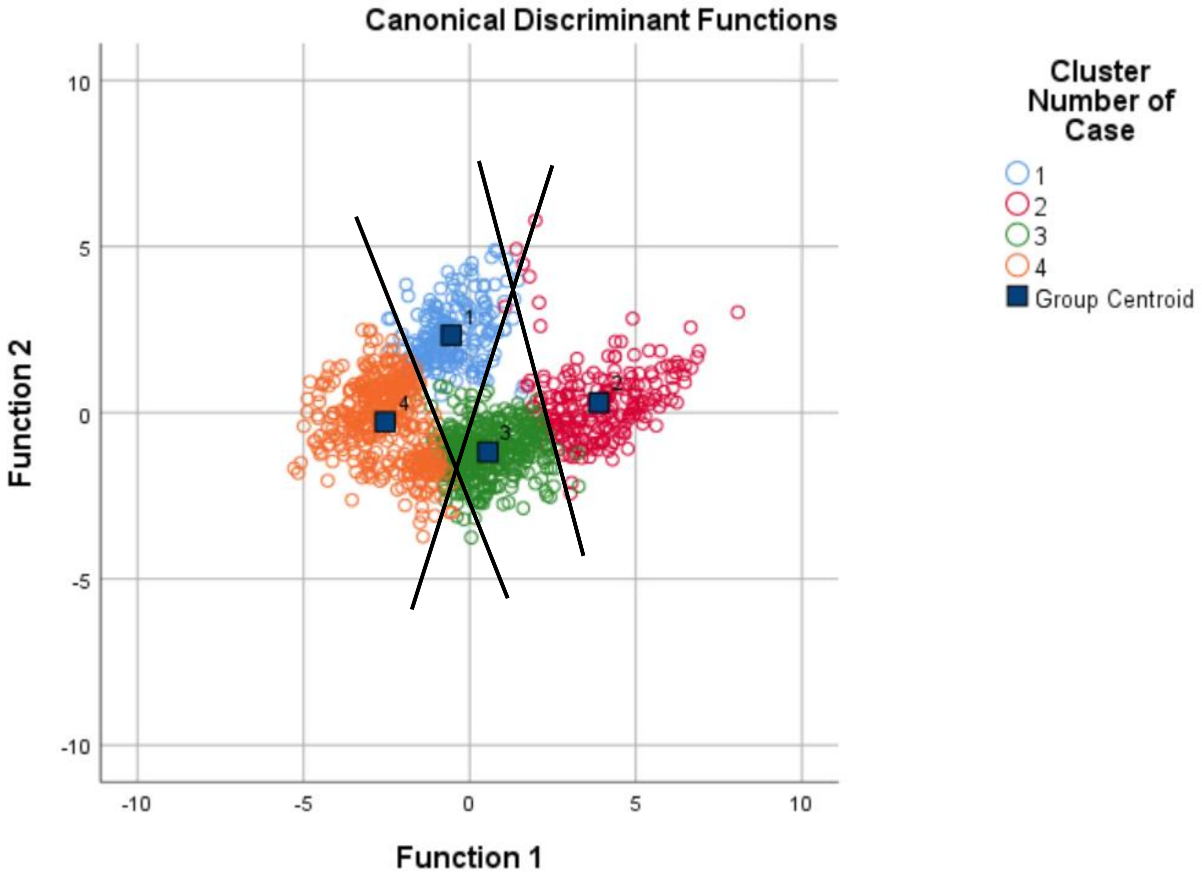
	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
FamilyIncome	1.367E+11	3	96460947.14	1226	1417.608	.000
GradRate	6.900	3	.014	1226	507.189	.000
PctFtFaculty	.393	3	.051	1226	7.762	.000
AdmissionRate	.731	3	.029	1226	25.041	.000
SATAVG	2886165.084	3	8607.498	1226	335.308	.000
UndergradPop	6208534795	3	32424761.20	1226	191.475	.000
FacultySalary	366645097.8	3	1967761.456	1226	186.326	.000
LoanCurrent3yr	2.590	3	.007	1226	384.430	.000
PCTFedLoan	1.713	3	.024	1226	70.433	.000
DEBT	1787164626	3	6059993.881	1226	294.912	.000
CostToAttend	3.821E+10	3	29634340.91	1226	1289.404	.000
PartTime_students	1.138	3	.009	1226	120.204	.000
Income_8yr	8592313635	3	39282374.74	1226	218.732	.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.

Client Z Evaluation

Refer to Figure 25 below. The discriminant plot below is defined on the range of grouping variables from 1 to 4. Missing values are replaced with the mean, although there should not be any missing from the dataset. Each observation is plotted and the centroids of the different clusters are denoted by small squares. This graph represents the qualitative judgment of distance between the clusters. I've added lines to depict the distinct clusters and it shows little overlap among the 4 clusters.

Figure 25: Client Z 4 Clusters Discriminant Plot- K Means Analysis



Conclusion

Client X

In summary, all of the variables were significant in determining the customer segments. There were five distinct clusters for the survey data provided. Cluster 1 were unhappy customers. Cluster 2 was categorized as average customers. Cluster 3 had the highest satisfaction rate. That segment could be identified as happy customers. Cluster 4 had another set of average customers but were not brand loyal. Cluster 5 had disappointed customers that could be won over with better handling of complaints.

Client Y

In conclusion, the 7 factors explain 59% of the variance. Factor 1 and factor 2 explained the largest proportion of the variance and had high loadings for a good portion of the questions in the Teaching of Statistics Questionnaire. Bland's theory that good researchers should have a profound love of statistics, an enthusiasm for experimental design, a love of teaching, and an absence of interpersonal skills is mostly justified by the results of the factor analysis.

Client Z

In review of the findings, all of the variables were significant in determining the segmentation of marketing groups for universities. There were four distinct clusters of the data provided. The college dataset reflects that wealthier students have higher graduation rates and greater debt. Less wealthy students have low test scores and low graduation rates. The middle class can be split into two groups. Both groups have similar statistics although one group goes to a more expensive university, perhaps a private school while the other segment goes to a more affordable school.

Resources

https://github.com/marissamckee/Clustering_Factor_Analysis

Appendix

Figure 1: Client X Clusters Scree Plot - K Means Analysis

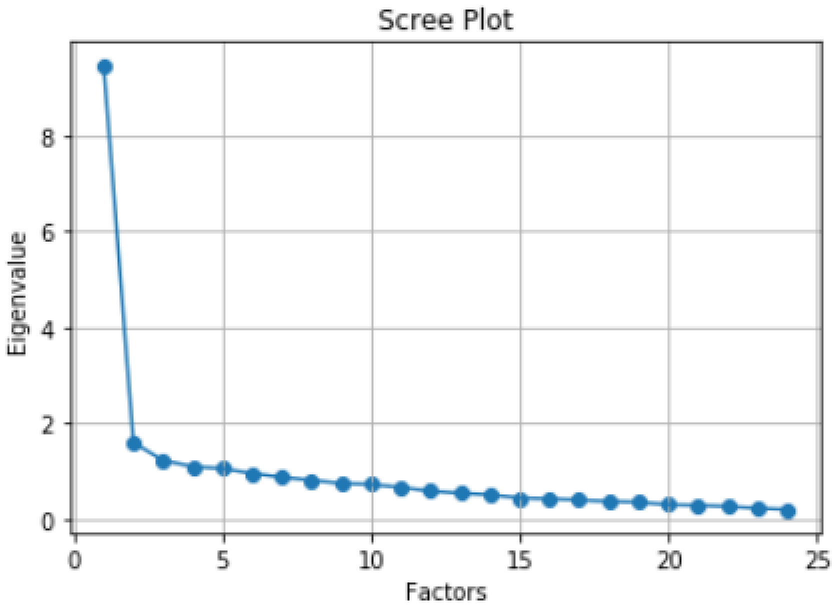


Figure 2: Client X 3 Clusters Bar Graph - K Means Analysis

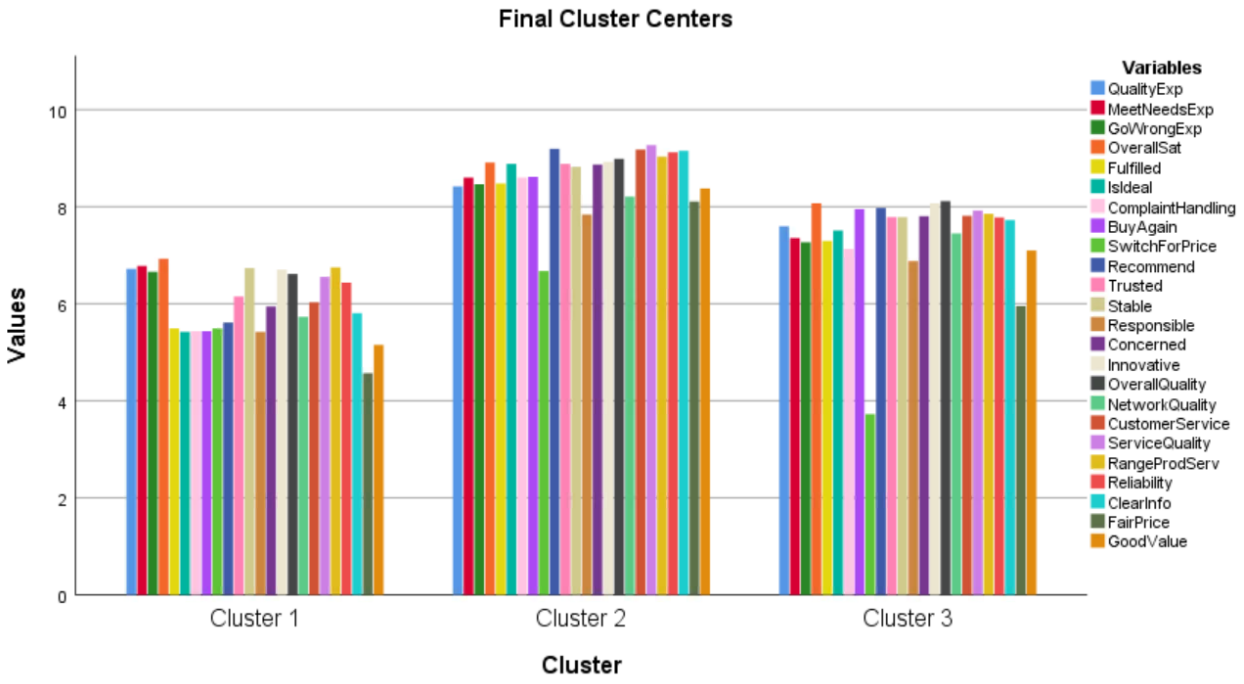
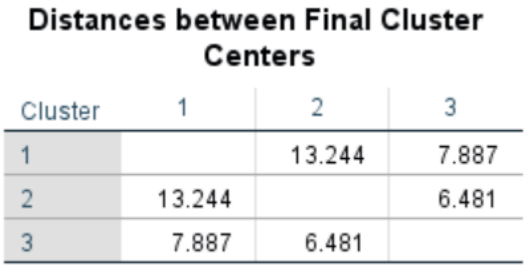


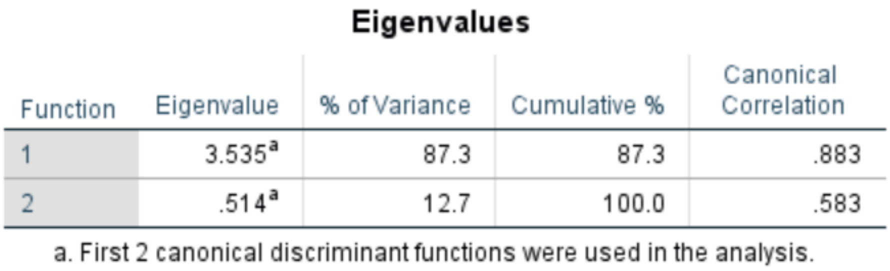
Figure 3: Client X 3 Clusters Distance Between Cluster Centroids- K Means Analysis



Distances between Final Cluster Centers

Cluster	1	2	3
1		13.244	7.887
2	13.244		6.481
3	7.887	6.481	

Figure 4: Client X 3 Clusters Eigenvalues- K Means Analysis



Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	3.535[a]	87.3	87.3	.883
2	.514[a]	12.7	100.0	.583

a. First 2 canonical discriminant functions were used in the analysis.

Figure 5: Client X 3 Clusters Discriminant Plot- K Means Analysis

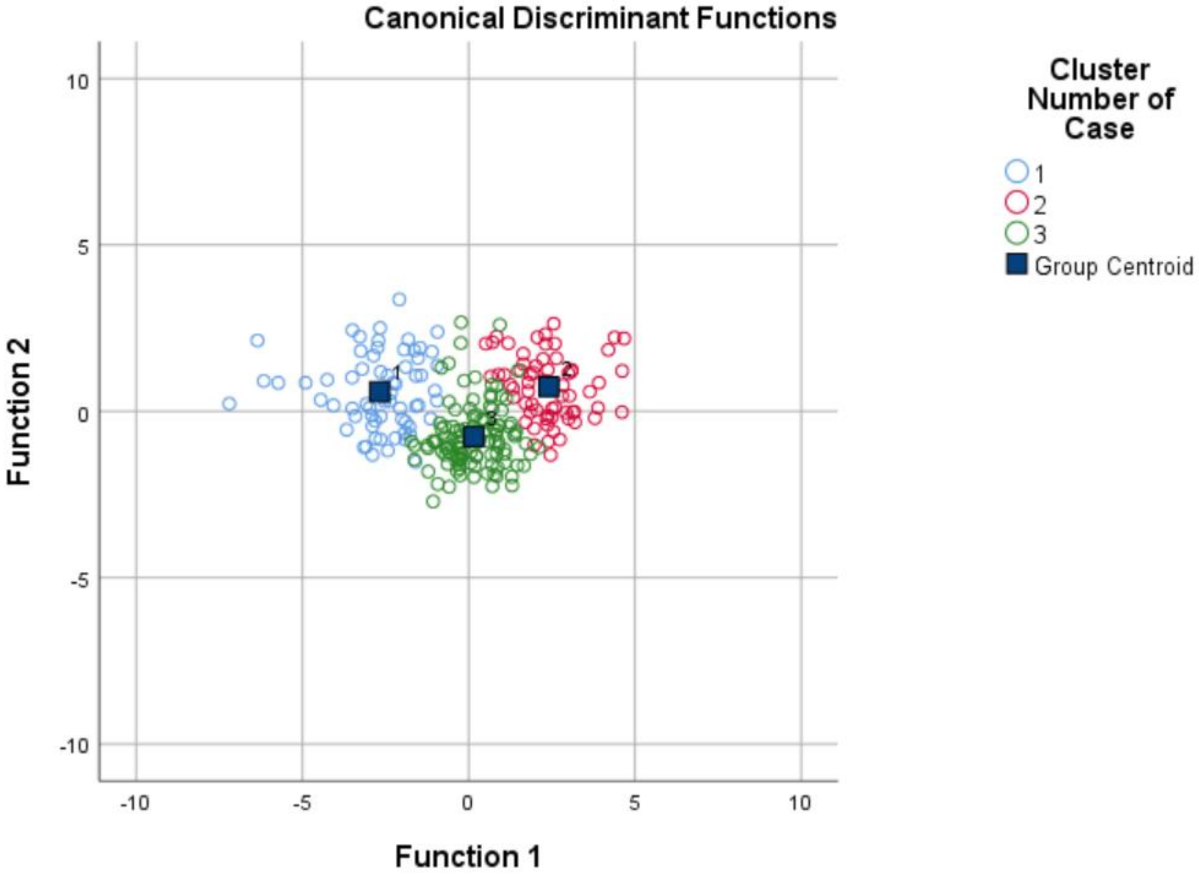
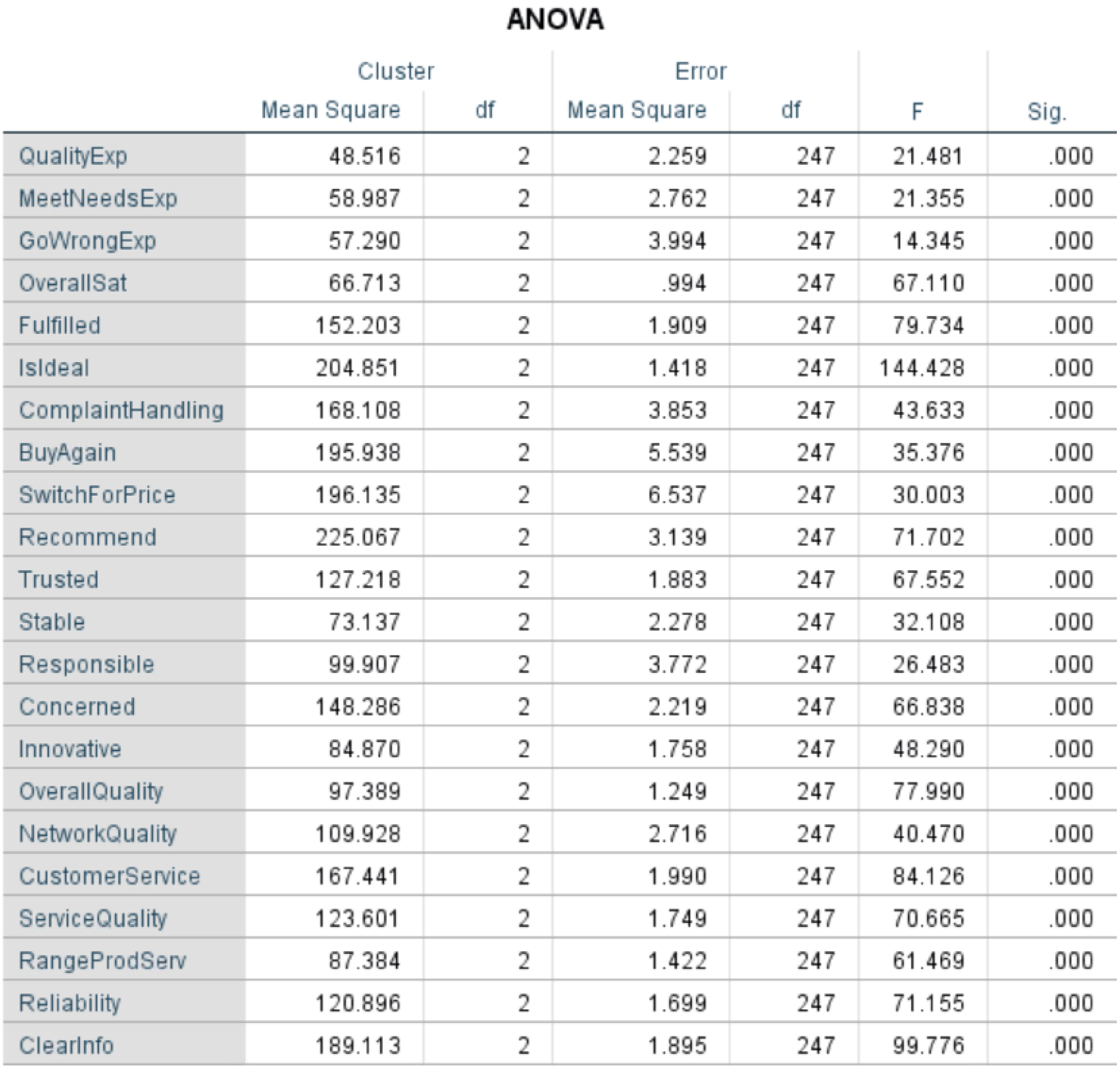


Figure 6: Client X 3 Clusters ANOVA - K Means Analysis



ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
QualityExp	48.516	2	2.259	247	21.481	.000
MeetNeedsExp	58.987	2	2.762	247	21.355	.000
GoWrongExp	57.290	2	3.994	247	14.345	.000
OverallSat	66.713	2	.994	247	67.110	.000
Fulfilled	152.203	2	1.909	247	79.734	.000
IsIdeal	204.851	2	1.418	247	144.428	.000
ComplaintHandling	168.108	2	3.853	247	43.633	.000
BuyAgain	195.938	2	5.539	247	35.376	.000
SwitchForPrice	196.135	2	6.537	247	30.003	.000
Recommend	225.067	2	3.139	247	71.702	.000
Trusted	127.218	2	1.883	247	67.552	.000
Stable	73.137	2	2.278	247	32.108	.000
Responsible	99.907	2	3.772	247	26.483	.000
Concerned	148.286	2	2.219	247	66.838	.000
Innovative	84.870	2	1.758	247	48.290	.000
OverallQuality	97.389	2	1.249	247	77.990	.000
NetworkQuality	109.928	2	2.716	247	40.470	.000
CustomerService	167.441	2	1.990	247	84.126	.000
ServiceQuality	123.601	2	1.749	247	70.665	.000
RangeProdServ	87.384	2	1.422	247	61.469	.000
Reliability	120.896	2	1.699	247	71.155	.000
ClearInfo	189.113	2	1.895	247	99.776	.000

Figure 7: Client X 4 Clusters Bar Graph - K Means Analysis

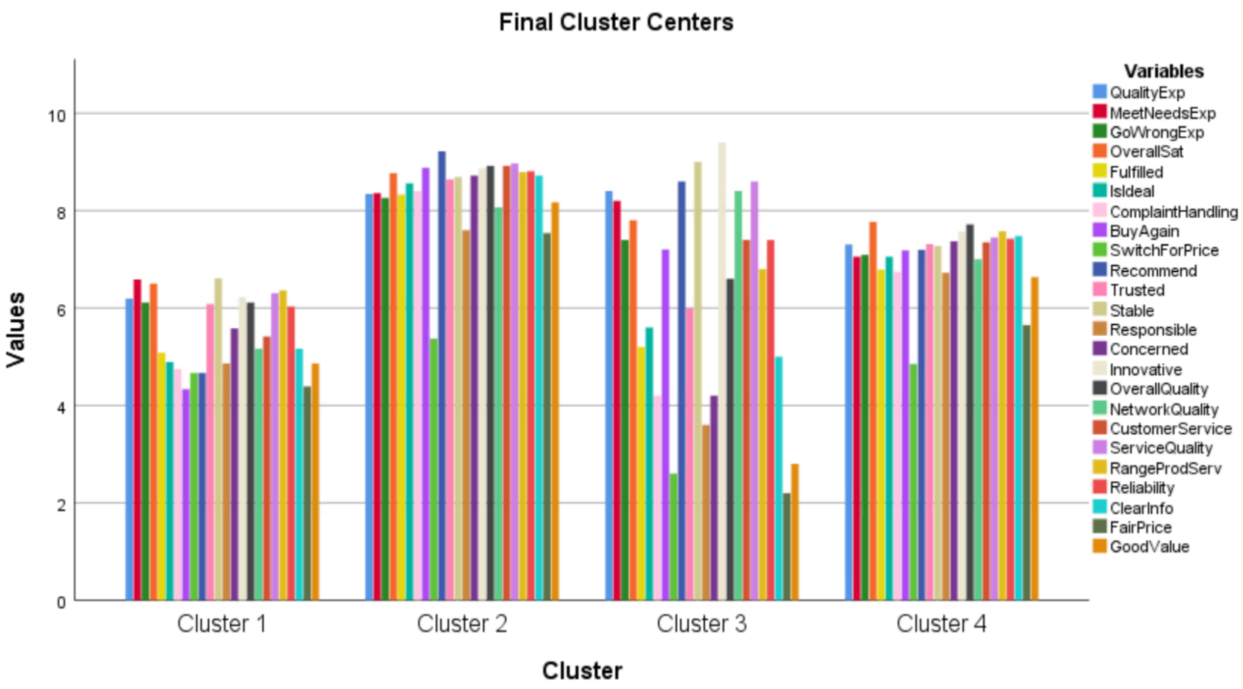
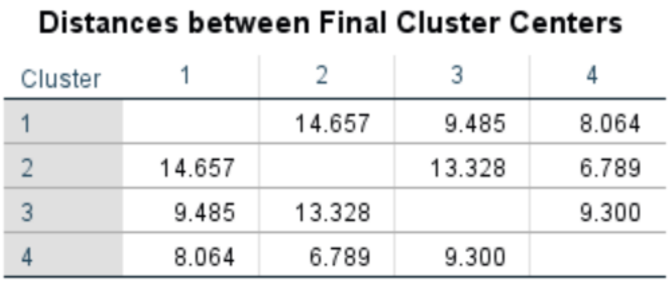


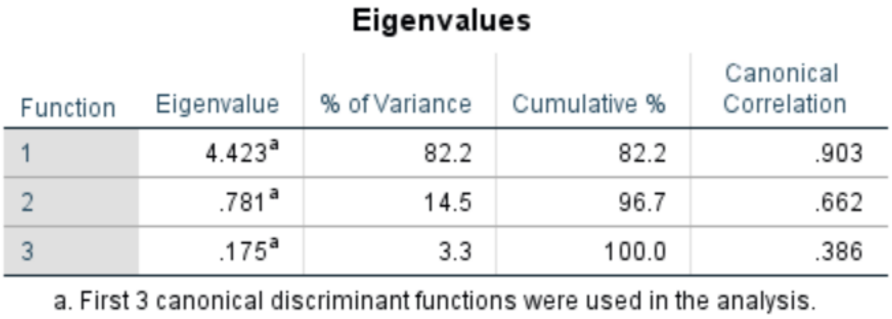
Figure 8: Client X 4 Clusters Distance Between Cluster Centroids- K Means Analysis



Distances between Final Cluster Centers

Cluster	1	2	3	4
1		14.657	9.485	8.064
2	14.657		13.328	6.789
3	9.485	13.328		9.300
4	8.064	6.789	9.300	

Figure 9: Client X 4 Clusters Eigenvalues- K Means Analysis



Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	4.423[a]	82.2	82.2	.903
2	.781[a]	14.5	96.7	.662
3	.175[a]	3.3	100.0	.386

a. First 3 canonical discriminant functions were used in the analysis.

Figure 10: Client X 4 Clusters Discriminant Plot- K Means Analysis

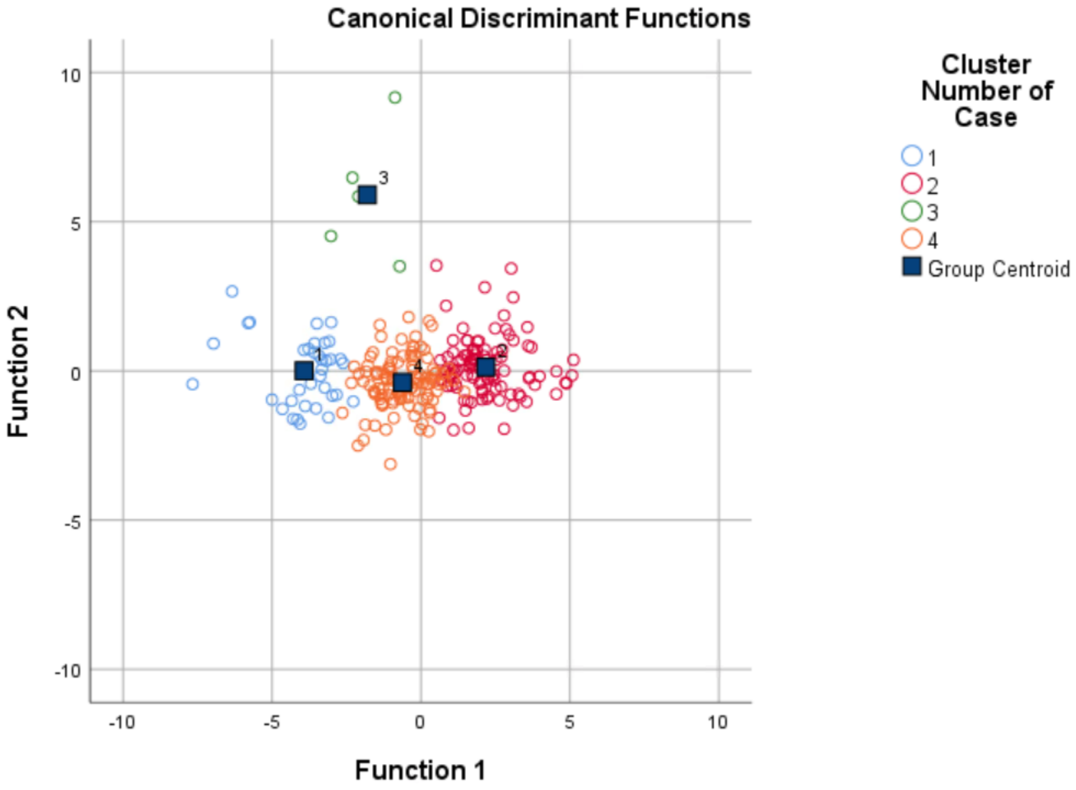
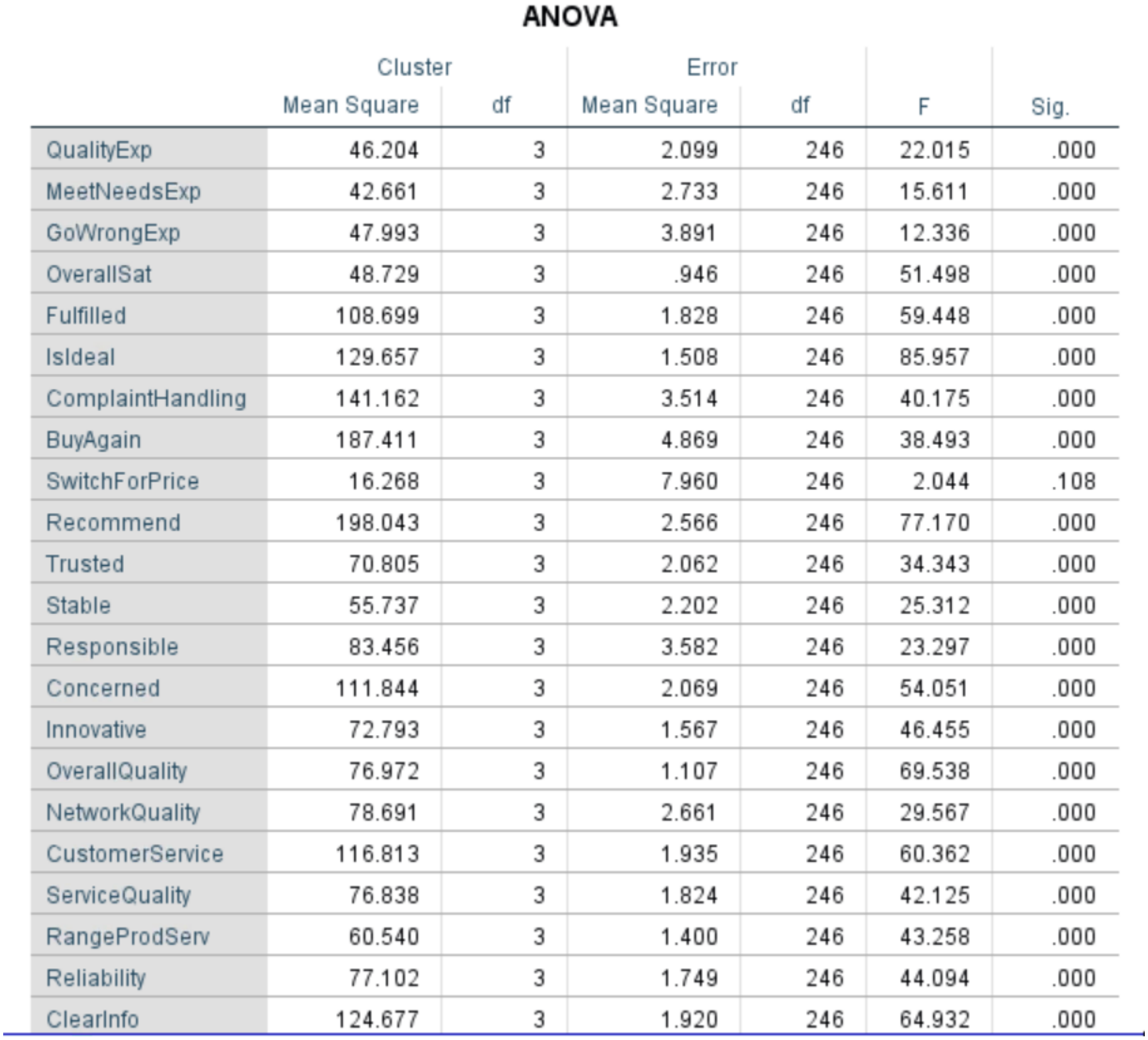


Figure 11: Client X 4 Clusters ANOVA - K Means Analysis



ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
QualityExp	46.204	3	2.099	246	22.015	.000
MeetNeedsExp	42.661	3	2.733	246	15.611	.000
GoWrongExp	47.993	3	3.891	246	12.336	.000
OverallSat	48.729	3	.946	246	51.498	.000
Fulfilled	108.699	3	1.828	246	59.448	.000
IsIdeal	129.657	3	1.508	246	85.957	.000
ComplaintHandling	141.162	3	3.514	246	40.175	.000
BuyAgain	187.411	3	4.869	246	38.493	.000
SwitchForPrice	16.268	3	7.960	246	2.044	.108
Recommend	198.043	3	2.566	246	77.170	.000
Trusted	70.805	3	2.062	246	34.343	.000
Stable	55.737	3	2.202	246	25.312	.000
Responsible	83.456	3	3.582	246	23.297	.000
Concerned	111.844	3	2.069	246	54.051	.000
Innovative	72.793	3	1.567	246	46.455	.000
OverallQuality	76.972	3	1.107	246	69.538	.000
NetworkQuality	78.691	3	2.661	246	29.567	.000
CustomerService	116.813	3	1.935	246	60.362	.000
ServiceQuality	76.838	3	1.824	246	42.125	.000
RangeProdServ	60.540	3	1.400	246	43.258	.000
Reliability	77.102	3	1.749	246	44.094	.000
ClearInfo	124.677	3	1.920	246	64.932	.000

Figure 12: Client X 5 Clusters Bar Graph - K Means Analysis

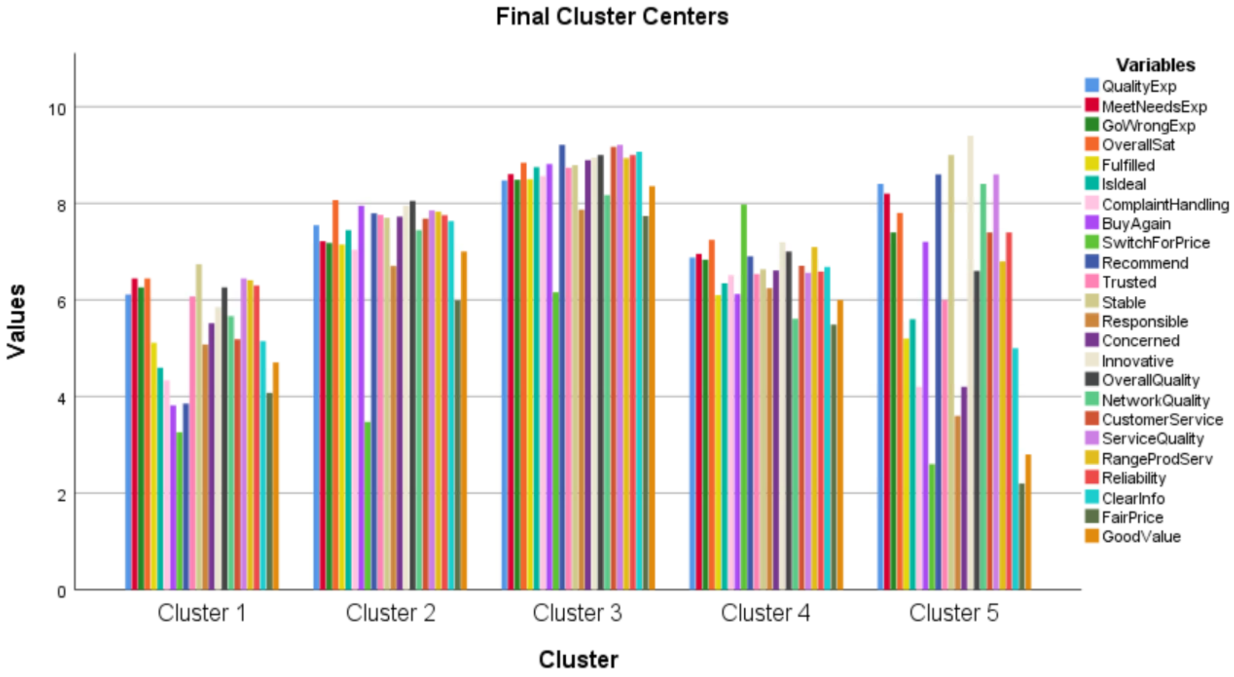
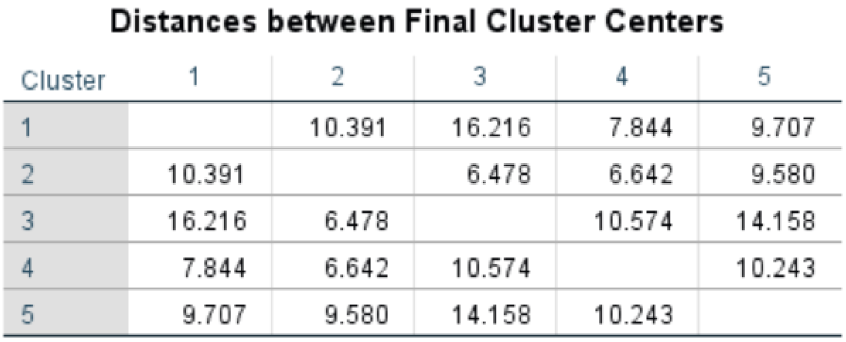


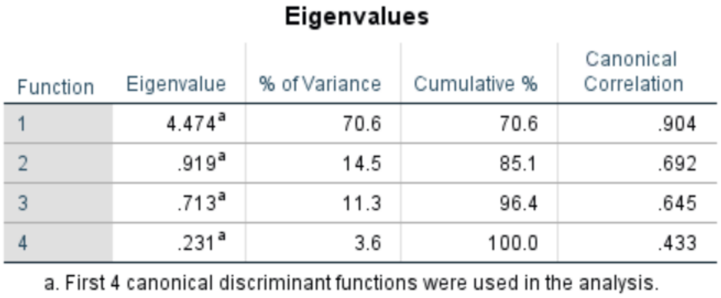
Figure 13: Client X 5 Clusters Distance Between Cluster Centroids- K Means Analysis



Distances between Final Cluster Centers

Cluster	1	2	3	4	5
1		10.391	16.216	7.844	9.707
2	10.391		6.478	6.642	9.580
3	16.216	6.478		10.574	14.158
4	7.844	6.642	10.574		10.243
5	9.707	9.580	14.158	10.243	

Figure 14: Client X 5 Clusters Eigenvalues- K Means Analysis



Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	4.474[a]	70.6	70.6	.904
2	.919[a]	14.5	85.1	.692
3	.713[a]	11.3	96.4	.645
4	.231[a]	3.6	100.0	.433

a. First 4 canonical discriminant functions were used in the analysis.

Figure 15: Client X 5 Clusters Discriminant Plot- K Means Analysis

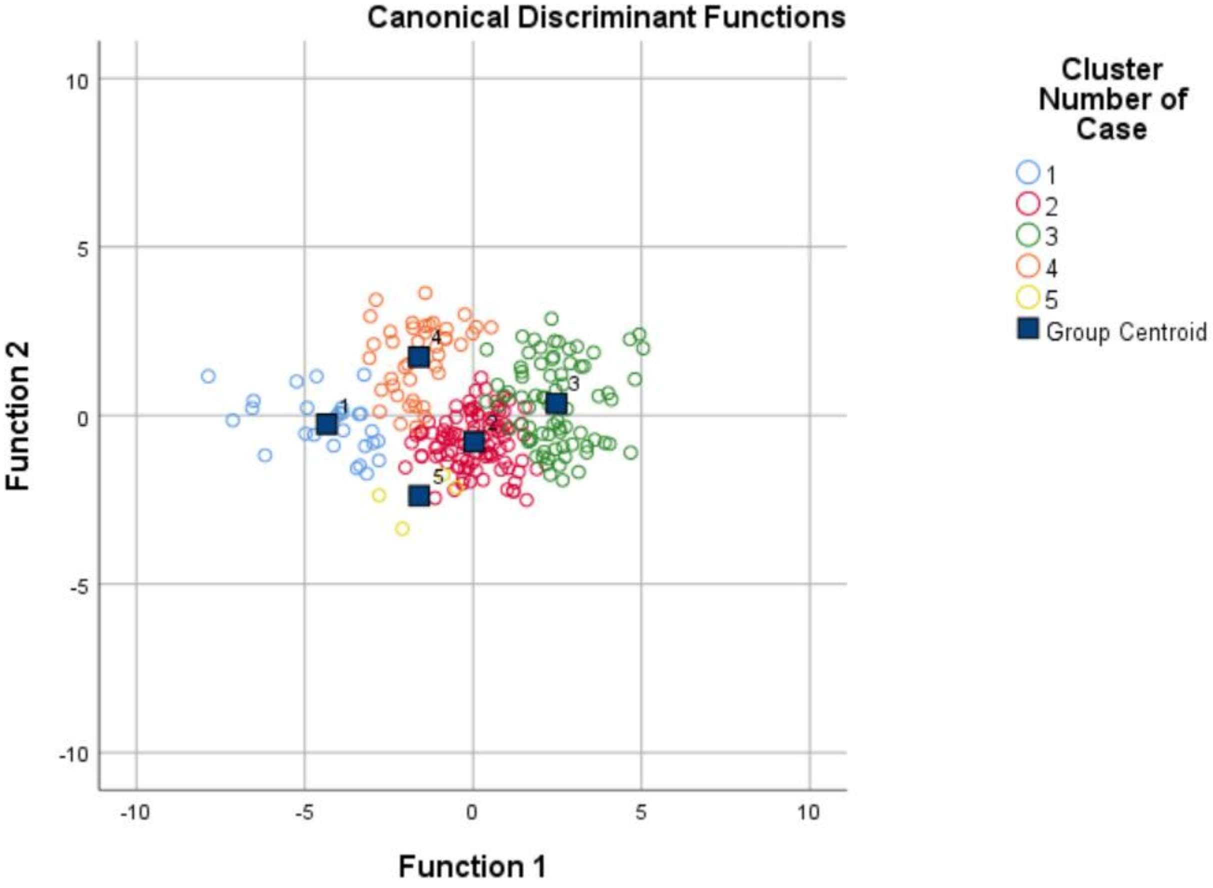
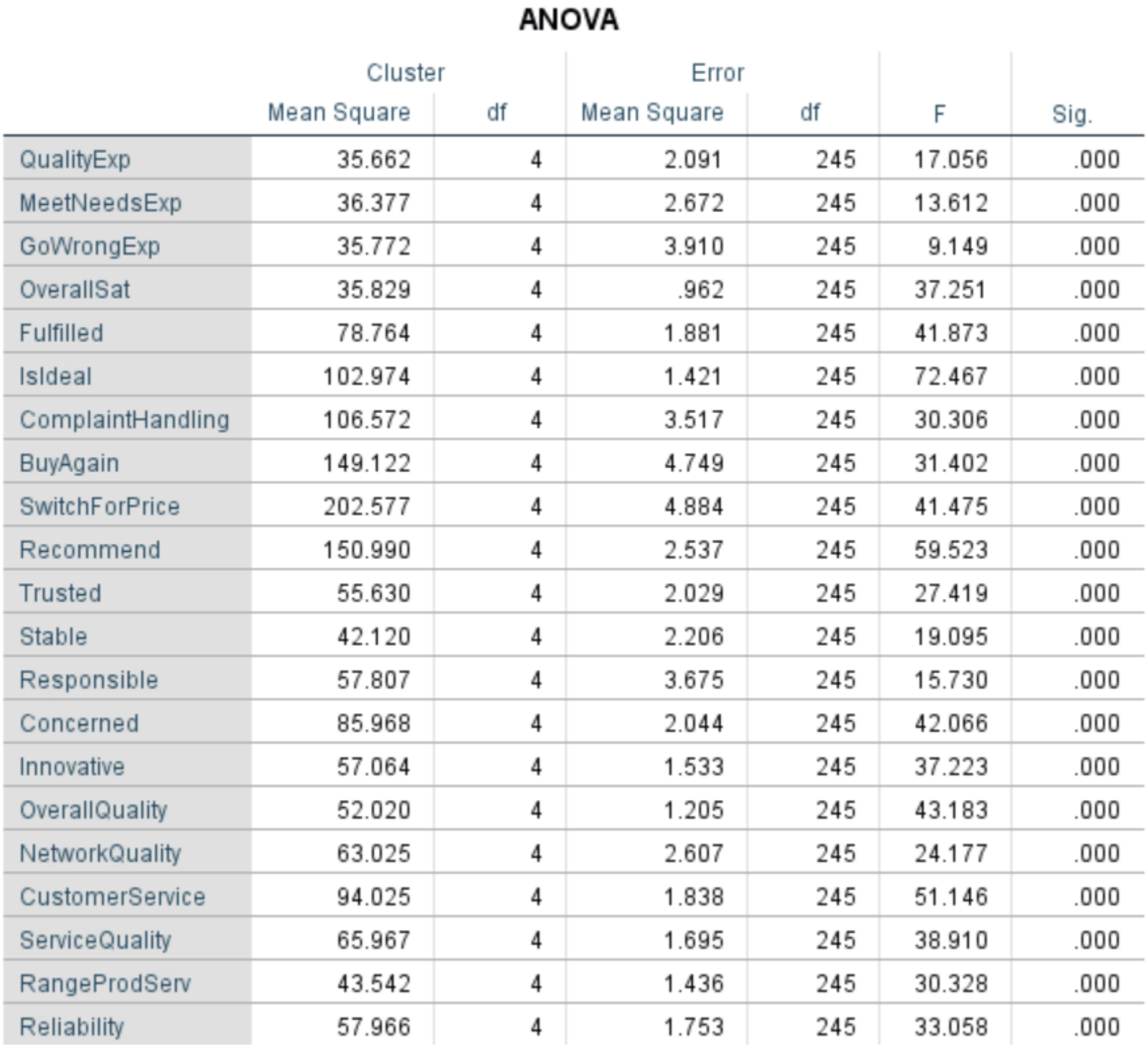


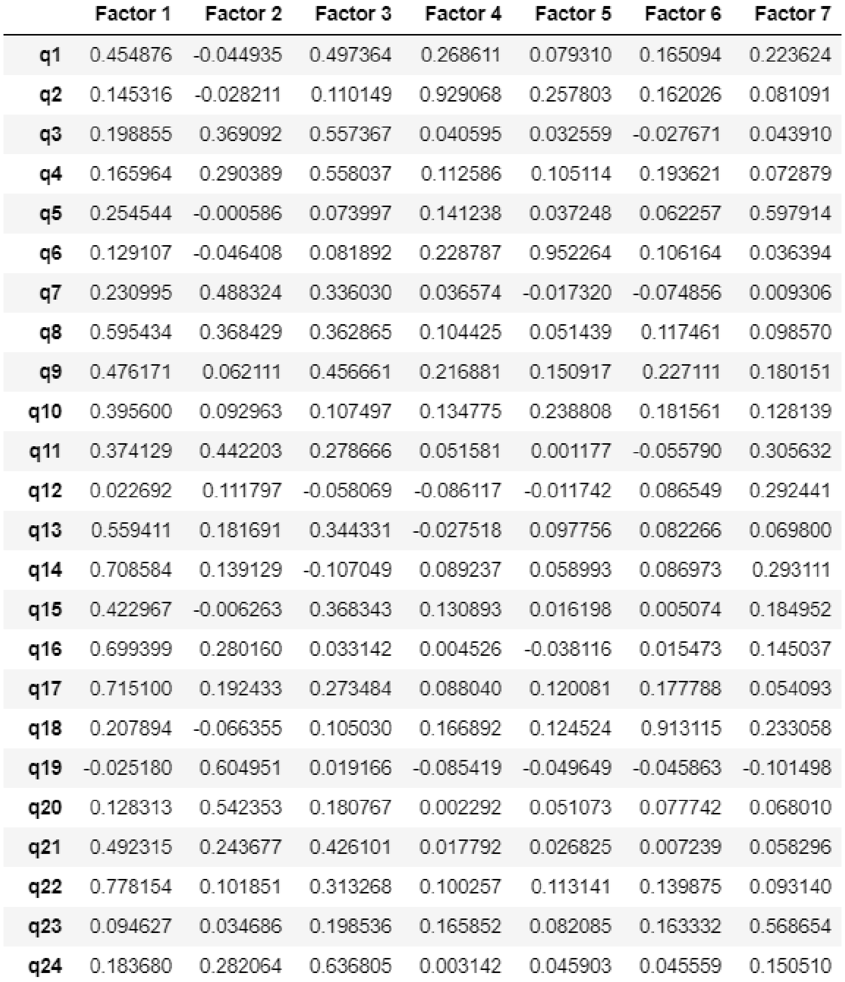
Figure 16: Client X 5 Clusters ANOVA - K Means Analysis



ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
QualityExp	35.662	4	2.091	245	17.056	.000
MeetNeedsExp	36.377	4	2.672	245	13.612	.000
GoWrongExp	35.772	4	3.910	245	9.149	.000
OverallSat	35.829	4	.962	245	37.251	.000
Fulfilled	78.764	4	1.881	245	41.873	.000
IsIdeal	102.974	4	1.421	245	72.467	.000
ComplaintHandling	106.572	4	3.517	245	30.306	.000
BuyAgain	149.122	4	4.749	245	31.402	.000
SwitchForPrice	202.577	4	4.884	245	41.475	.000
Recommend	150.990	4	2.537	245	59.523	.000
Trusted	55.630	4	2.029	245	27.419	.000
Stable	42.120	4	2.206	245	19.095	.000
Responsible	57.807	4	3.675	245	15.730	.000
Concerned	85.968	4	2.044	245	42.066	.000
Innovative	57.064	4	1.533	245	37.223	.000
OverallQuality	52.020	4	1.205	245	43.183	.000
NetworkQuality	63.025	4	2.607	245	24.177	.000
CustomerService	94.025	4	1.838	245	51.146	.000
ServiceQuality	65.967	4	1.695	245	38.910	.000
RangeProdServ	43.542	4	1.436	245	30.328	.000
Reliability	57.966	4	1.753	245	33.058	.000

Figure 16: Client Y Factor Analysis



	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
q1	0.454876	-0.044935	0.497364	0.268611	0.079310	0.165094	0.223624
q2	0.145316	-0.028211	0.110149	0.929068	0.257803	0.162026	0.081091
q3	0.198855	0.369092	0.557367	0.040595	0.032559	-0.027671	0.043910
q4	0.165964	0.290389	0.558037	0.112586	0.105114	0.193621	0.072879
q5	0.254544	-0.000586	0.073997	0.141238	0.037248	0.062257	0.597914
q6	0.129107	-0.046408	0.081892	0.228787	0.952264	0.106164	0.036394
q7	0.230995	0.488324	0.336030	0.036574	-0.017320	-0.074856	0.009306
q8	0.595434	0.368429	0.362865	0.104425	0.051439	0.117461	0.098570
q9	0.476171	0.062111	0.456661	0.216881	0.150917	0.227111	0.180151
q10	0.395600	0.092963	0.107497	0.134775	0.238808	0.181561	0.128139
q11	0.374129	0.442203	0.278666	0.051581	0.001177	-0.055790	0.305632
q12	0.022692	0.111797	-0.058069	-0.086117	-0.011742	0.086549	0.292441
q13	0.559411	0.181691	0.344331	-0.027518	0.097756	0.082266	0.069800
q14	0.708584	0.139129	-0.107049	0.089237	0.058993	0.086973	0.293111
q15	0.422967	-0.006263	0.368343	0.130893	0.016198	0.005074	0.184952
q16	0.699399	0.280160	0.033142	0.004526	-0.038116	0.015473	0.145037
q17	0.715100	0.192433	0.273484	0.088040	0.120081	0.177788	0.054093
q18	0.207894	-0.066355	0.105030	0.166892	0.124524	0.913115	0.233058
q19	-0.025180	0.604951	0.019166	-0.085419	-0.049649	-0.045863	-0.101498
q20	0.128313	0.542353	0.180767	0.002292	0.051073	0.077742	0.068010
q21	0.492315	0.243677	0.426101	0.017792	0.026825	0.007239	0.058296
q22	0.778154	0.101851	0.313268	0.100257	0.113141	0.139875	0.093140
q23	0.094627	0.034686	0.198536	0.165852	0.082085	0.163332	0.568654
q24	0.183680	0.282064	0.636805	0.003142	0.045903	0.045559	0.150510

Figure 17: Client Z Clusters Scree Plot - K Means Analysis

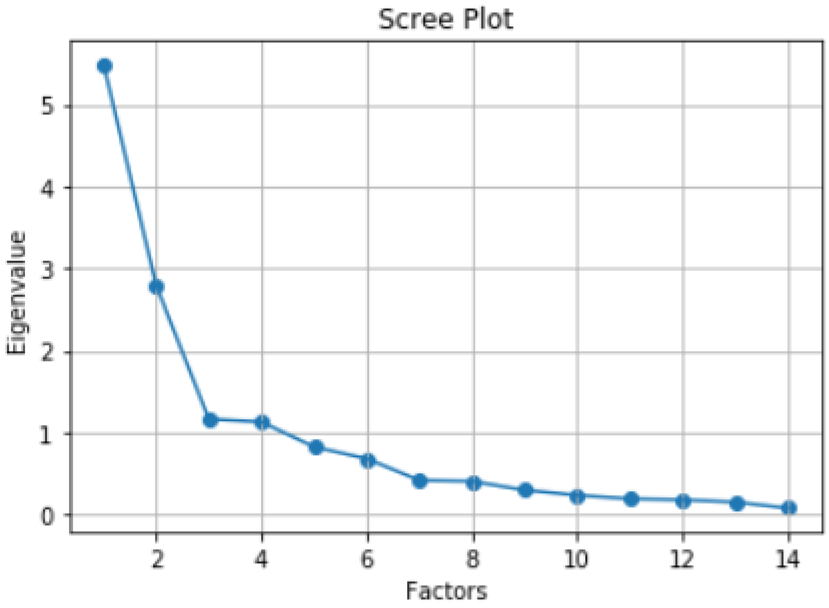
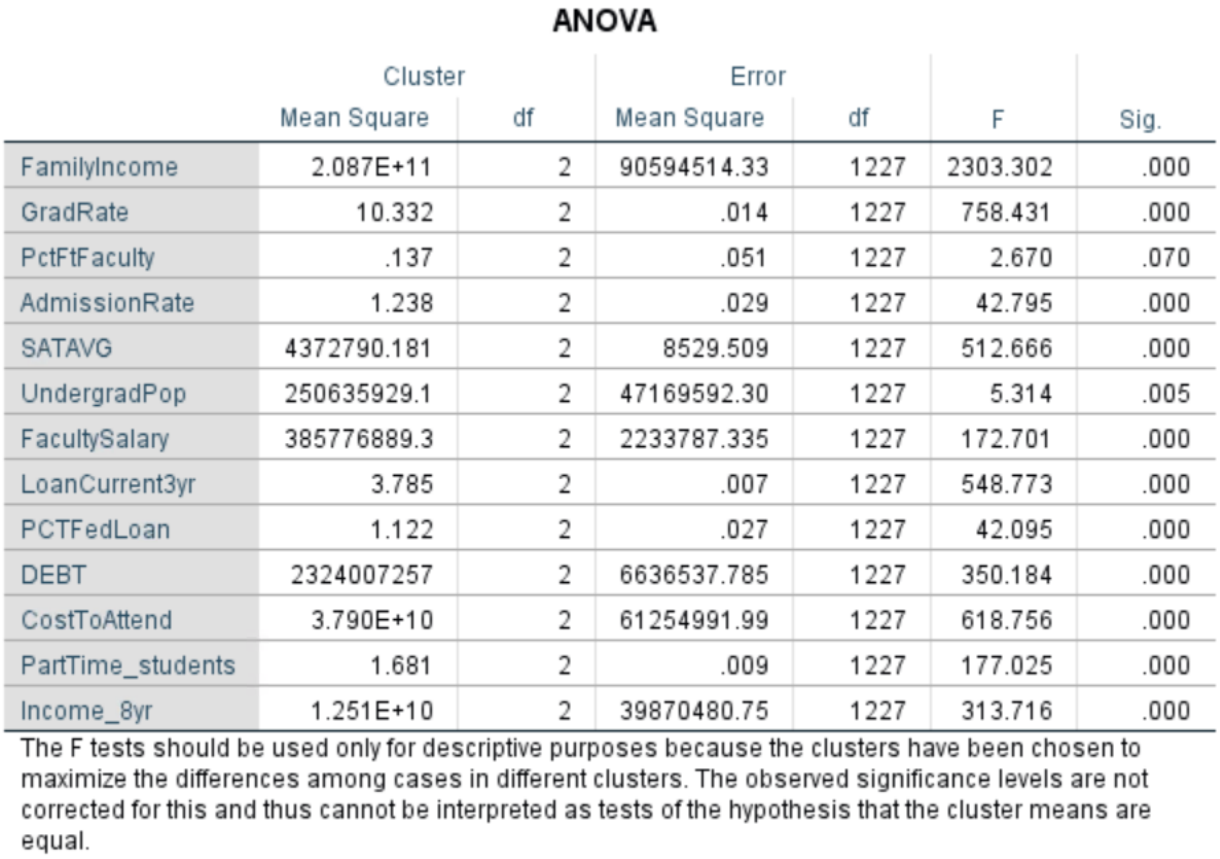


Figure 18: Client Z 3 Clusters ANOVA - K Means Analysis

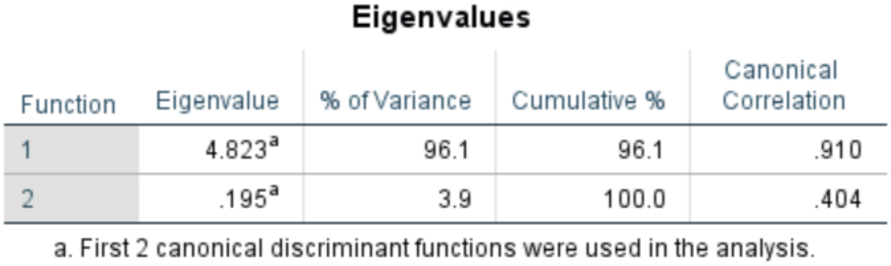


ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
FamilyIncome	2.087E+11	2	90594514.33	1227	2303.302	.000
GradRate	10.332	2	.014	1227	758.431	.000
PctFtFaculty	.137	2	.051	1227	2.670	.070
AdmissionRate	1.238	2	.029	1227	42.795	.000
SATAVG	4372790.181	2	8529.509	1227	512.666	.000
UndergradPop	250635929.1	2	47169592.30	1227	5.314	.005
FacultySalary	385776889.3	2	2233787.335	1227	172.701	.000
LoanCurrent3yr	3.785	2	.007	1227	548.773	.000
PCTFedLoan	1.122	2	.027	1227	42.095	.000
DEBT	2324007257	2	6636537.785	1227	350.184	.000
CostToAttend	3.790E+10	2	61254991.99	1227	618.756	.000
PartTime_students	1.681	2	.009	1227	177.025	.000
Income_8yr	1.251E+10	2	39870480.75	1227	313.716	.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.

Figure 19: Client Z 3 Clusters Eigenvalues- K Means Analysis



Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	4.823[a]	96.1	96.1	.910
2	.195[a]	3.9	100.0	.404

a. First 2 canonical discriminant functions were used in the analysis.

Figure 20: Client Z 3 Clusters Discriminant Plot- K Means Analysis

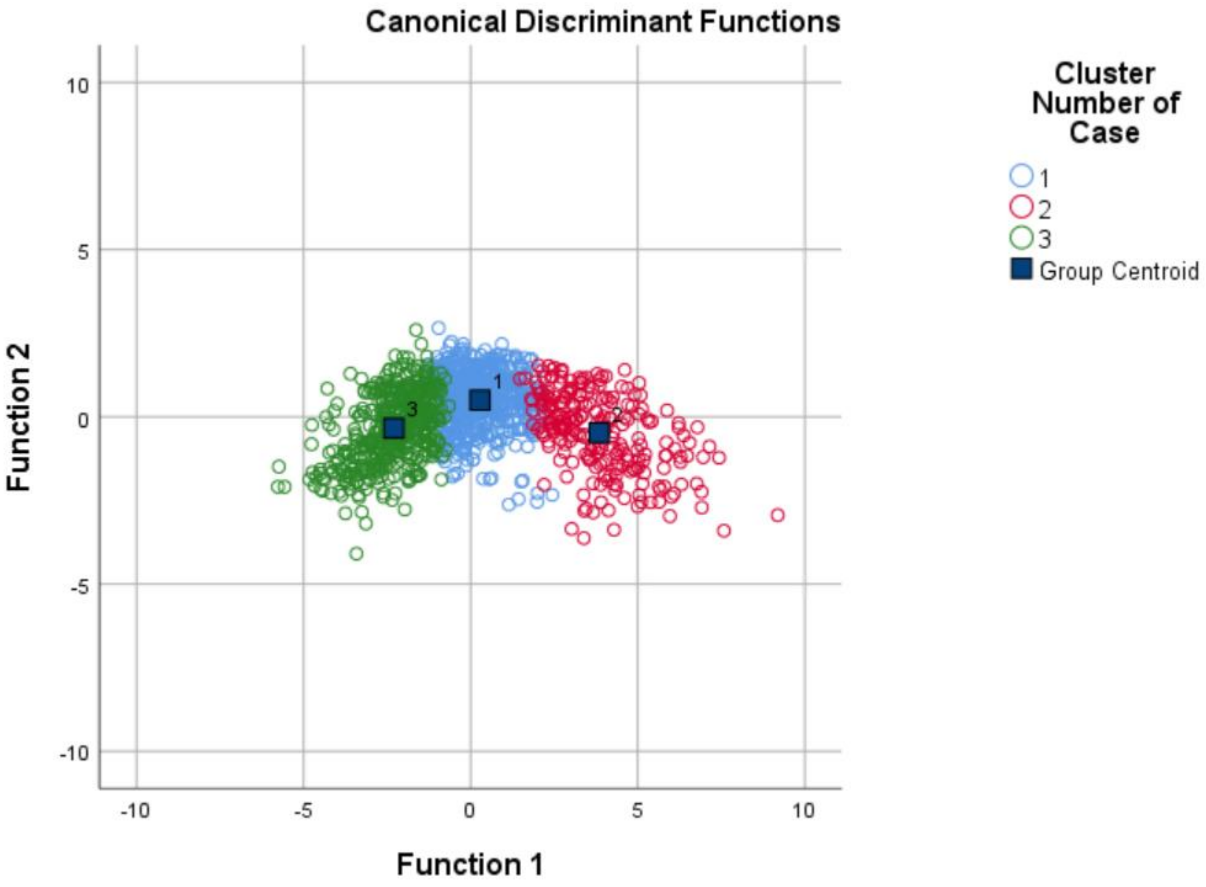


Figure 21: Client Z 3 Clusters Bar Graph - K Means Analysis

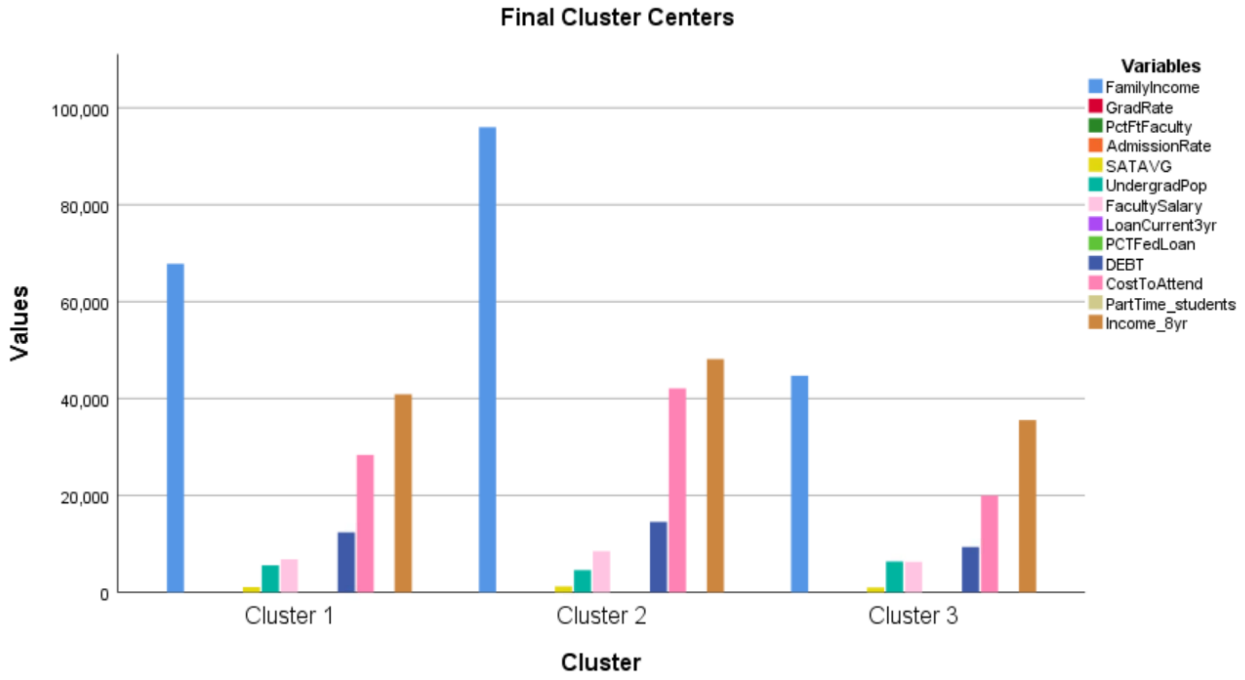
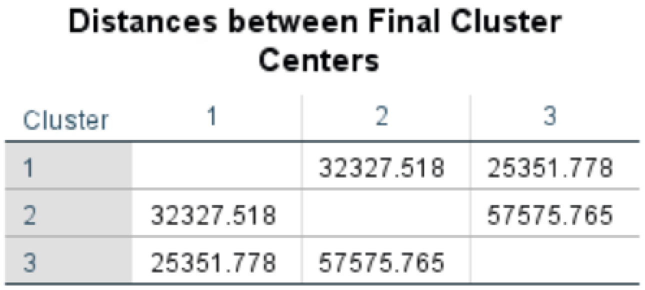


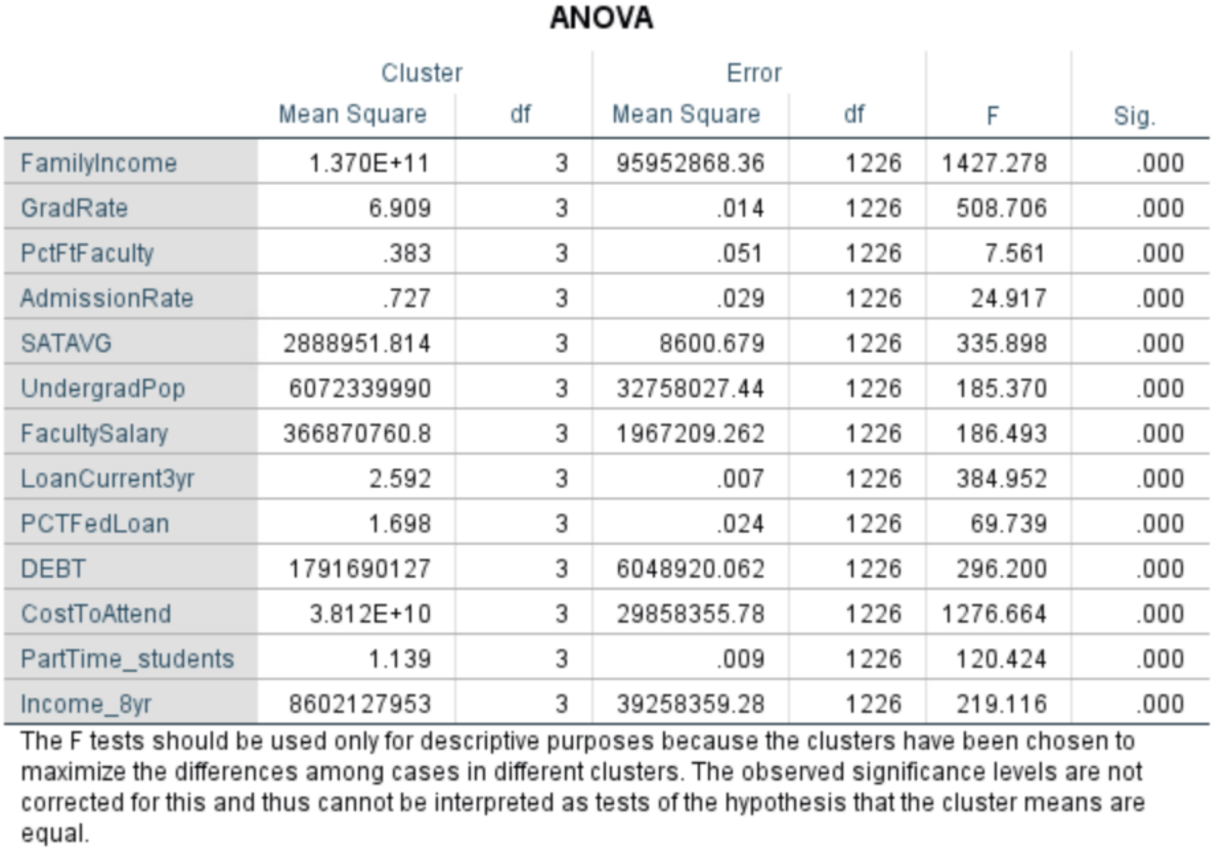
Figure 22: Client Z 3 Clusters Distance Between Cluster Centroids- K Means Analysis



Distances between Final Cluster Centers

Cluster	1	2	3
1		32327.518	25351.778
2	32327.518		57575.765
3	25351.778	57575.765	

Figure 23: Client Z 4 Clusters ANOVA - K Means Analysis

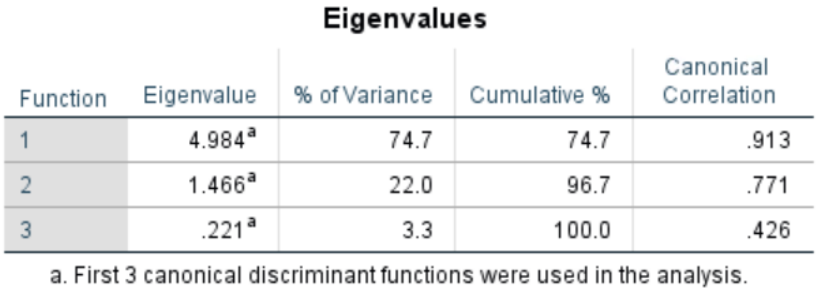


ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
FamilyIncome	1.370E+11	3	95952868.36	1226	1427.278	.000
GradRate	6.909	3	.014	1226	508.706	.000
PctFtFaculty	.383	3	.051	1226	7.561	.000
AdmissionRate	.727	3	.029	1226	24.917	.000
SATAVG	2888951.814	3	8600.679	1226	335.898	.000
UndergradPop	6072339990	3	32758027.44	1226	185.370	.000
FacultySalary	366870760.8	3	1967209.262	1226	186.493	.000
LoanCurrent3yr	2.592	3	.007	1226	384.952	.000
PCTFedLoan	1.698	3	.024	1226	69.739	.000
DEBT	1791690127	3	6048920.062	1226	296.200	.000
CostToAttend	3.812E+10	3	29858355.78	1226	1276.664	.000
PartTime_students	1.139	3	.009	1226	120.424	.000
Income_8yr	8602127953	3	39258359.28	1226	219.116	.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.

Figure 24: Client Z 4 Clusters Eigenvalues- K Means Analysis



Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	4.984[a]	74.7	74.7	.913
2	1.466[a]	22.0	96.7	.771
3	.221[a]	3.3	100.0	.426

a. First 3 canonical discriminant functions were used in the analysis.

Figure 25: Client Z 4 Clusters Discriminant Plot- K Means Analysis

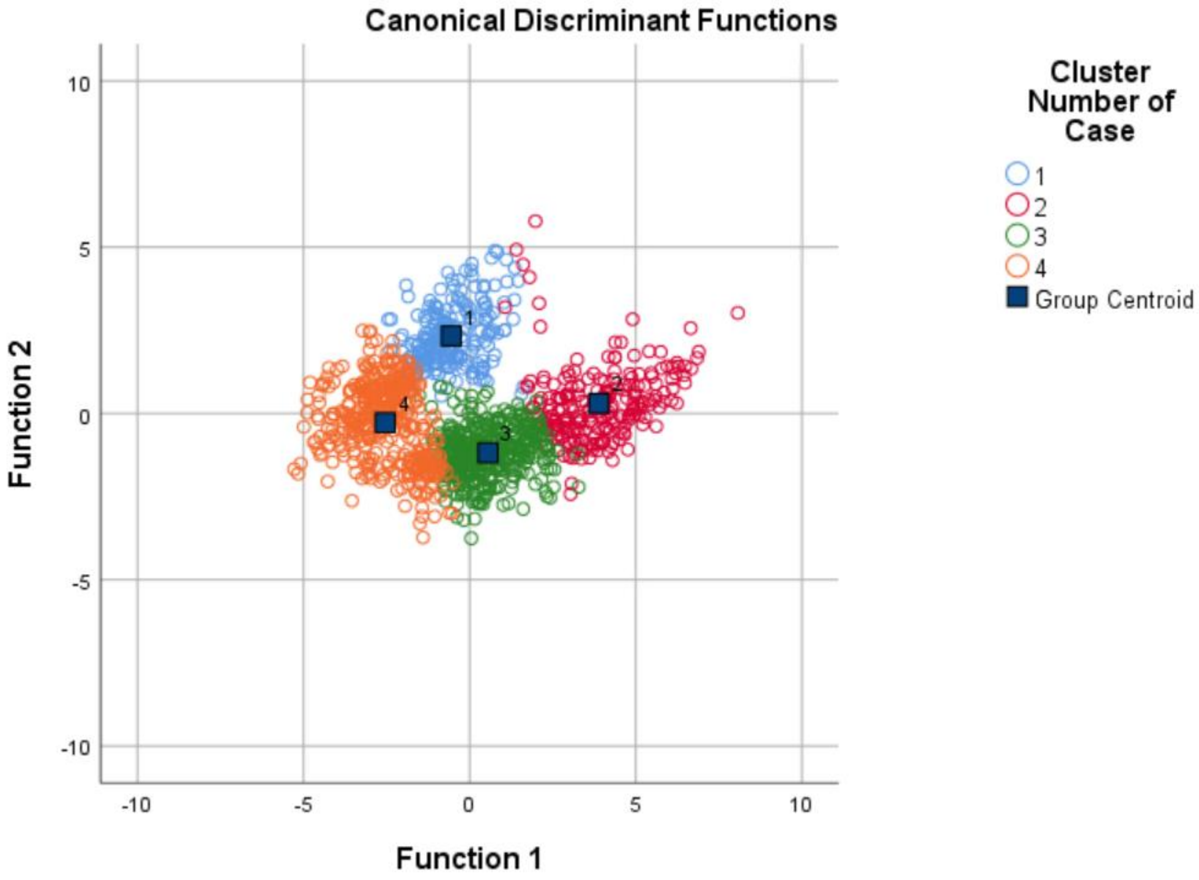


Figure 26: Client Z 4 Clusters Bar Graph - K Means Analysis

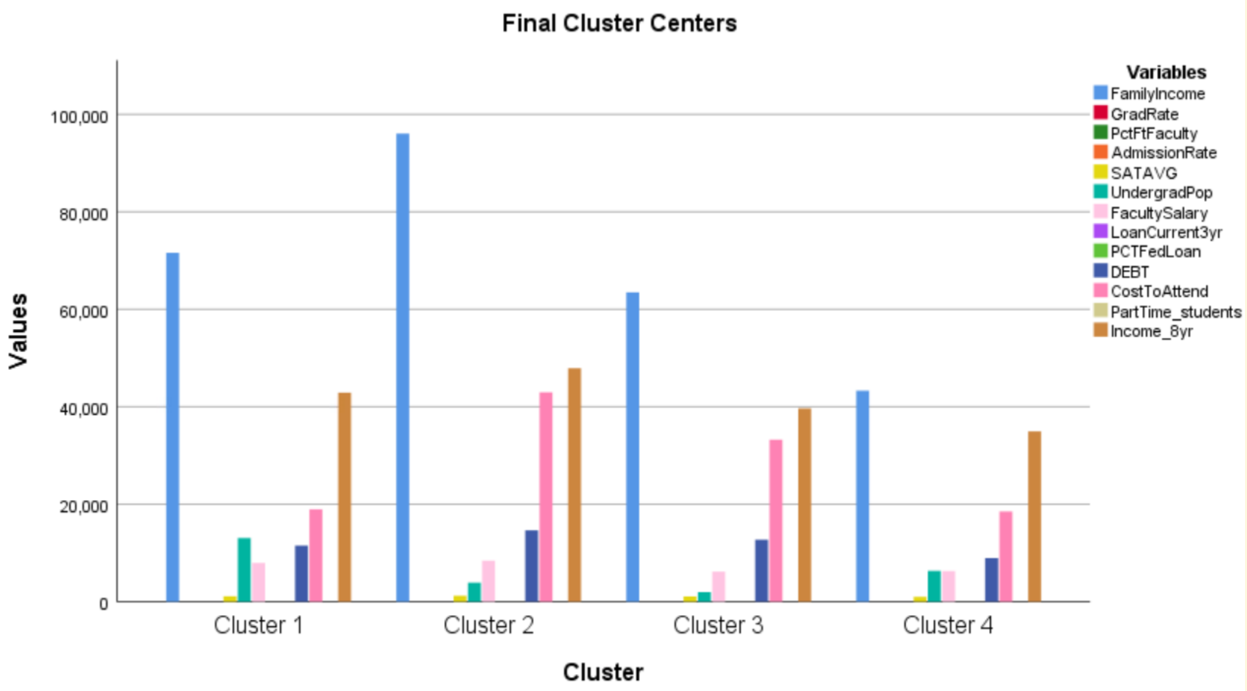
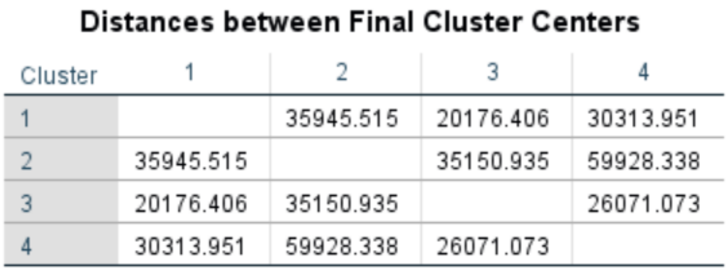


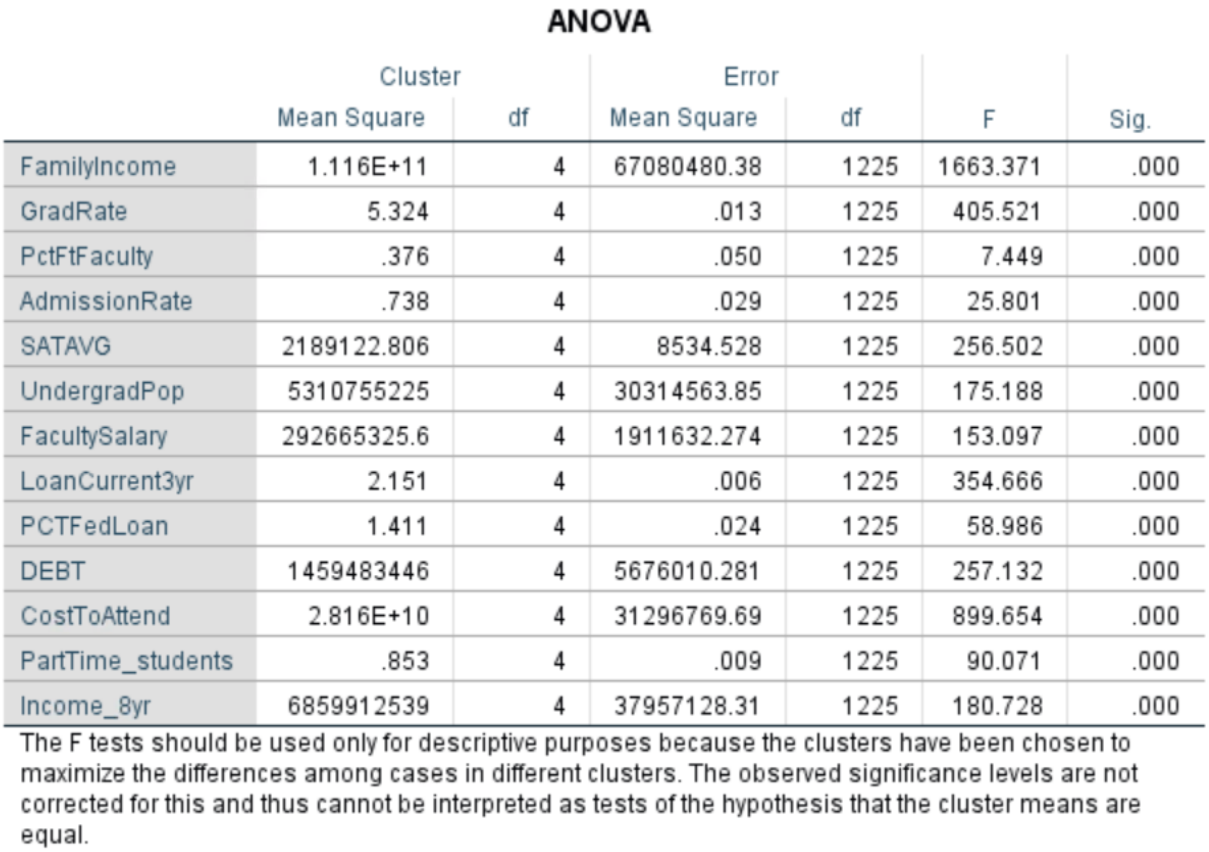
Figure 27: Client Z 4 Clusters Distance Between Cluster Centroids- K Means Analysis



Distances between Final Cluster Centers

Cluster	1	2	3	4
1		35945.515	20176.406	30313.951
2	35945.515		35150.935	59928.338
3	20176.406	35150.935		26071.073
4	30313.951	59928.338	26071.073	

Figure 28: Client Z 5 Clusters ANOVA - K Means Analysis

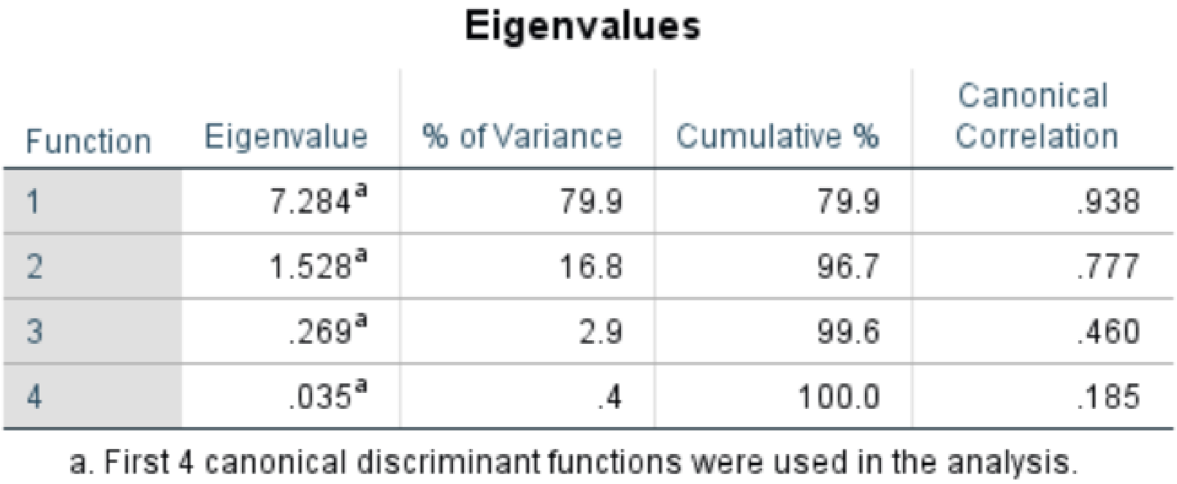


ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
FamilyIncome	1.116E+11	4	67080480.38	1225	1663.371	.000
GradRate	5.324	4	.013	1225	405.521	.000
PctFtFaculty	.376	4	.050	1225	7.449	.000
AdmissionRate	.738	4	.029	1225	25.801	.000
SATAVG	2189122.806	4	8534.528	1225	256.502	.000
UndergradPop	5310755225	4	30314563.85	1225	175.188	.000
FacultySalary	292665325.6	4	1911632.274	1225	153.097	.000
LoanCurrent3yr	2.151	4	.006	1225	354.666	.000
PCTFedLoan	1.411	4	.024	1225	58.986	.000
DEBT	1459483446	4	5676010.281	1225	257.132	.000
CostToAttend	2.816E+10	4	31296769.69	1225	899.654	.000
PartTime_students	.853	4	.009	1225	90.071	.000
Income_8yr	6859912539	4	37957128.31	1225	180.728	.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.

Figure 29: Client Z 5 Clusters Eigenvalues- K Means Analysis



Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	7.284[a]	79.9	79.9	.938
2	1.528[a]	16.8	96.7	.777
3	.269[a]	2.9	99.6	.460
4	.035[a]	.4	100.0	.185

a. First 4 canonical discriminant functions were used in the analysis.

Figure 30: Client Z 5 Clusters Discriminant Plot- K Means Analysis

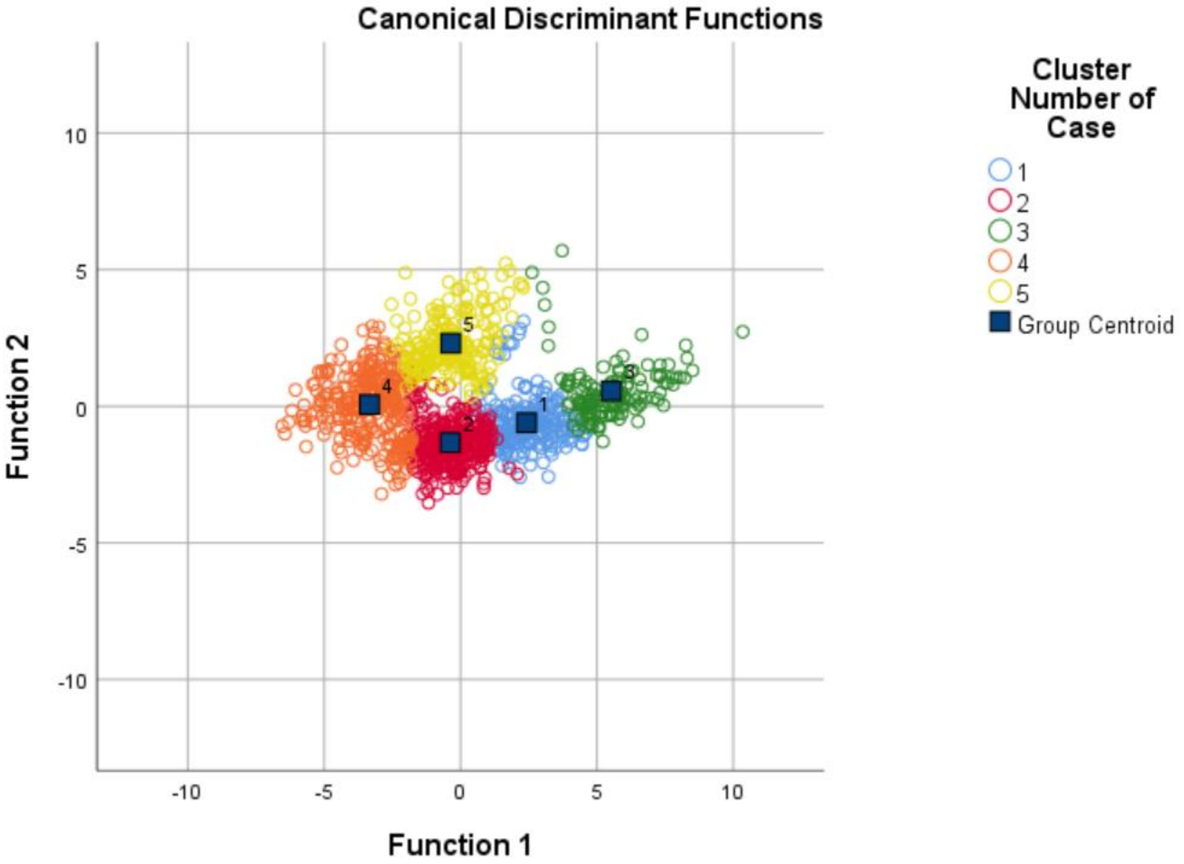


Figure 31: Client Z 5 Clusters Bar Graph - K Means Analysis

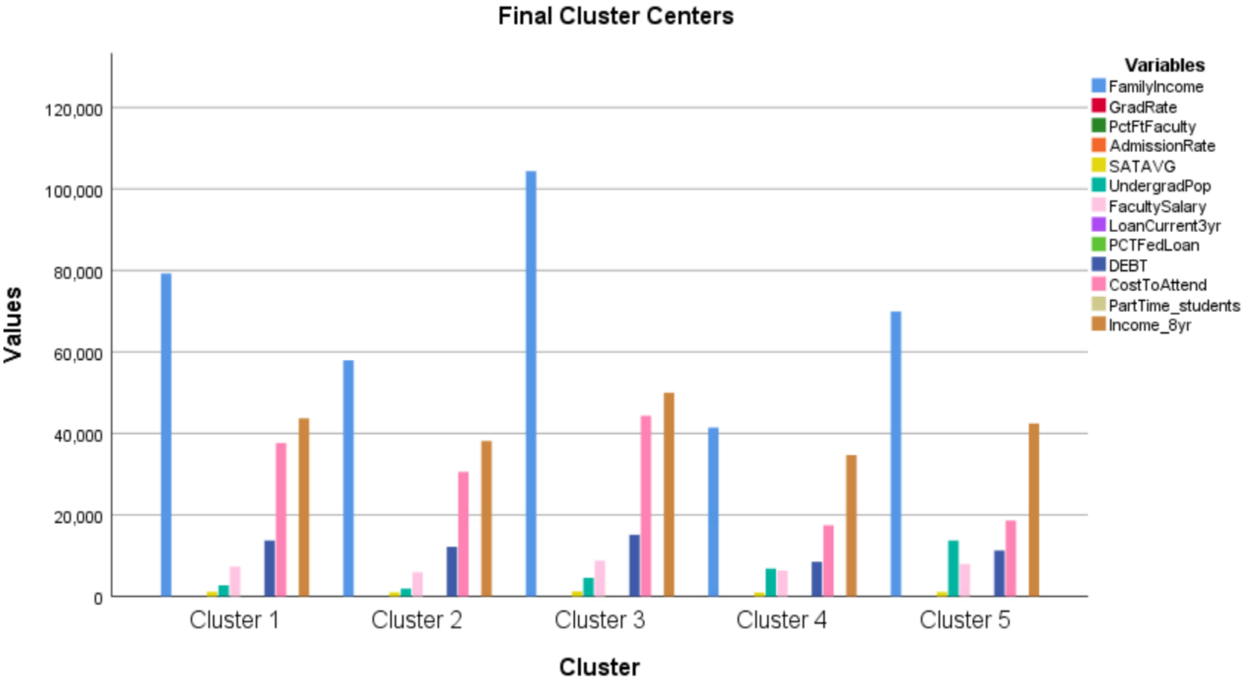
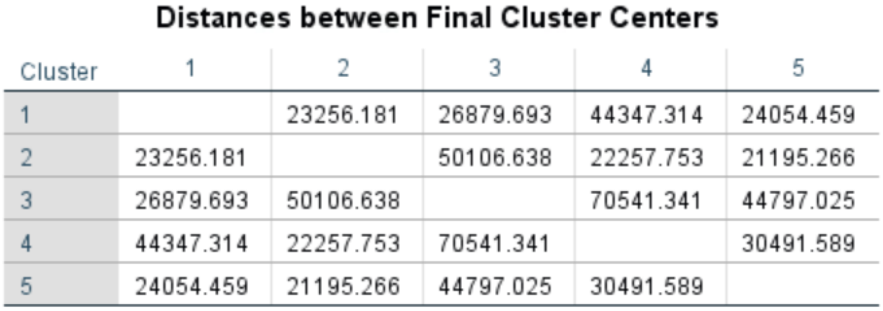


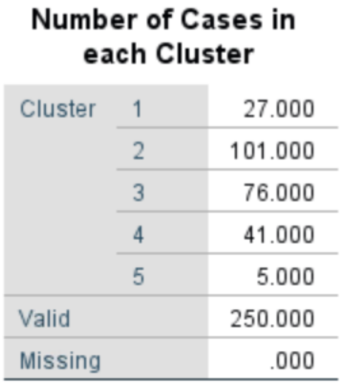
Figure 32: Client Z 5 Clusters Distance Between Cluster Centroids- K Means Analysis



Distances between Final Cluster Centers

Cluster	1	2	3	4	5
1		23256.181	26879.693	44347.314	24054.459
2	23256.181		50106.638	22257.753	21195.266
3	26879.693	50106.638		70541.341	44797.025
4	44347.314	22257.753	70541.341		30491.589
5	24054.459	21195.266	44797.025	30491.589	

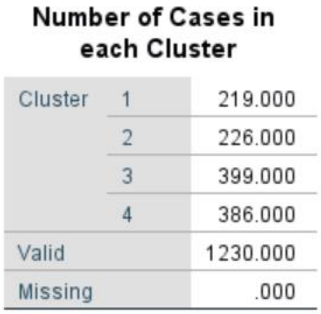
Figure 33: Client X 5 Clusters Number of Cases in Each Cluster- K Means Analysis



Number of Cases in each Cluster

Cluster	1	27.000
	2	101.000
	3	76.000
	4	41.000
	5	5.000
Valid		250.000
Missing		.000

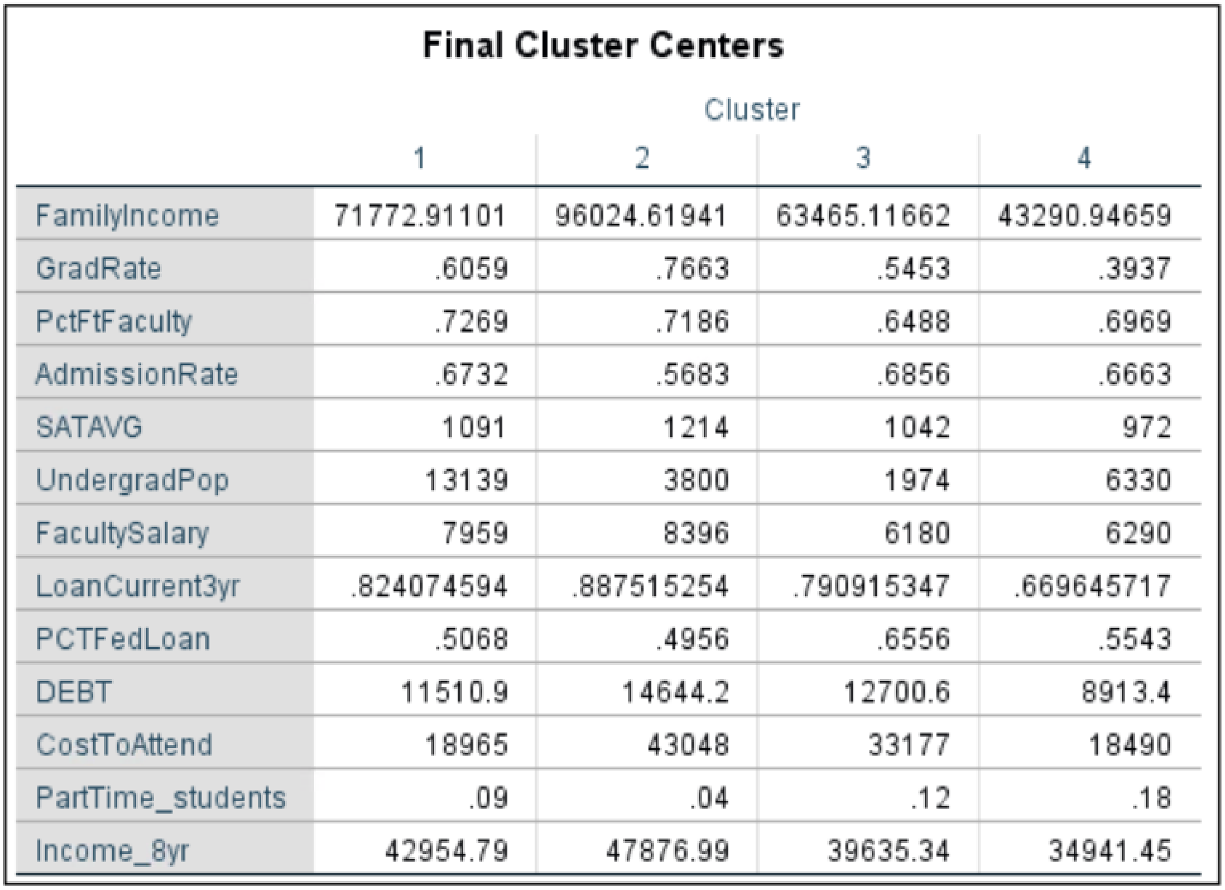
Figure 34: Client Z 4 Clusters Number of Cases in Each Cluster- K Means Analysis



Number of Cases in each Cluster

Cluster	1	219.000
	2	226.000
	3	399.000
	4	386.000
Valid		1230.000
Missing		.000

Figure 35 Client Z 4 Final Cluster Centers- K Means Analysis



Final Cluster Centers

	Cluster			
	1	2	3	4
FamilyIncome	71772.91101	96024.61941	63465.11662	43290.94659
GradRate	.6059	.7663	.5453	.3937
PctFtFaculty	.7269	.7186	.6488	.6969
AdmissionRate	.6732	.5683	.6856	.6663
SATAVG	1091	1214	1042	972
UndergradPop	13139	3800	1974	6330
FacultySalary	7959	8396	6180	6290
LoanCurrent3yr	.824074594	.887515254	.790915347	.669645717
PCTFedLoan	.5068	.4956	.6556	.5543
DEBT	11510.9	14644.2	12700.6	8913.4
CostToAttend	18965	43048	33177	18490
PartTime_students	.09	.04	.12	.18
Income_8yr	42954.79	47876.99	39635.34	34941.45