MOHAMMED KHAN
DUE 12.18.2020
DSEG5113: DATA EXPLORATION
SEGMENTATION AND PROFILING PAPER
DR. ANDREW BANASIEWICZ

Executive Summary

Overview

As a marketing analysis manager for a telecommunications company, I have been tasked with developing customer segmentation that can support effective and economically sound customer retention efforts. The Customer Database dataset is comprised of various demographic, historical, behavioral and financial metrics. The objective here is to identify high value customers and make recommendations on how to retain their patronage for the benefit of the company. This exercise will attempt to produce clusters of the patient population and analyze targeting efforts which will yield the most business from the most lucrative customers.

Business Goals

This analysis will be viewed through the lens of the segmentation and profiling of the consumer population in this dataset. The objective is to identify distinct consumer cohorts so customer analysis can be done in an efficient manner. After identification the groups will be subjected to profiling, which will uncover shared trends and attitudes within groups. This segmentation and profiling will be taken into consideration to provide a snapshot for product and promotional targeting efforts. The bottom line will be to drive and maintain sales by identifying high value customers and prioritizing their loyalty to the company.

Data Description

The Customer Database dataset consists of 5000 observations with 59 attributes. To set this dataset up for segmentation, missing data must be imputed and categorical data must be converted to numerical values. There are also columns that should be feature engineered, either to provide more clarify and conciseness, or to transform data points for greater flexibility. Categorical variables were recoded with integers, starting with 0. For this segmentation exercise, this dataset is not scaled, due to the conflict that would arise when trying to translate scaled cluster results to real world applications. The recoded dataset contains 89 attributes, an increase from the 59 that were initially present.

Methodological Summary

RStudio IDE and R programming language are used in tandem for this exercise. There are a host of packages utilized for plotting, normalization, clustering, formatting, and data manipulation. There is no metadata or data dictionary provided with the dataset, which makes big picture context difficult to gauge, however there is enough information on each customer that this is not a debilitating concern for the exercise. Tabulation software is also utilized in tandem with the plots and tables written into the code.

Utmost importance was given to the high value customers. Once the clustering is established, the next step is to go back and profile only the high value consumers from the initial dataset. A column with cluster number is added to the data frame, giving an indication of which high value customers belonged to which clusters. From there, a summary of the updated groups was taken to give a clue as to which habits high value customers show, this time by cluster value.

Results

The full process, from data due diligence to segmentation and profiling, produced a recoded dataset, several plots for visualizing clustering and a multitude of tabular information to describe the relationships between the clusters and the variables themselves. In short, since there was no consensus between the silhouette and gap plots, the most reasonable values of 4 and 5 clusters were studied. The gap plot recommends a group of 4 clusters while the silhouette plot suggests a cluster counts of two, five and four, respectively. A commonality both clustering attempts shared is how the largest groups across both sets of clusters also displayed the highest value of *"NewsSubscriber."* This combination has the makings of a significant asset. It would be beneficial to target groups that show a willing to stay with the company and bring in revenue reliably, while also being subscribed to a news service. It displays loyalty and customer satisfaction to the company. This method of contact between company and consumer opens a plethora of possibilities for product and service promotion, especially when more powerful forms of social media are considered.

Recommendations

The most valuable takeaway from this exercise is the evidence of large existing cohort that has the capability to receive direct communication from the company. High value customers can fall into one of two groups. The first subset are customers that pay a large fee for the company's services each month, taking advantage of multiple services that are provided. The second group may only use a few of the services provided, however they have been within the

company's ecosystem long enough that they can be considered high value consumers. Ideally, a successful plan would be to cater to both groups. While we were not able to differentiate between these two groups, we can acknowledge that they are undoubtedly important to the company. The profiling method shows that a significantly sizable amount of high value consumers are also subscribed to the company's news service. Establishing a base group of consumers is made easier with this advantage. Customers are most likely subscribed to one telecommunications company at a time, being offered data, voice and equipment services. This gives us three ways to present promotions, primarily for loyalty programs. We can tell based on the results of the clustering that high value consumers are more likely to receive direct news from the company. Therefore, by sending out loyalty promotions, we can appeal to a large base of customers, since these customers already come under the umbrella of regular consumers. Referral programs will also be effective, this time the target group being friends and family of our high value customers. This method would promote the growth of the company with minimal effort because prospective buyers are more likely to trust people they know over corporations or ads they see on multimedia.

Core Report

Overview

The purpose of this task is to communicate how this company can adapt practical techniques and actions that lead to customer retention beyond what pure business sense can provide. With the tools of segmentation, profiling, and data analysis, customer information can be organized and manipulated to discover underlying patterns that can be advantageous for business. The goal of doing better business remains static; what is being done here is merely viewing this situation through a new lens. Segmentation and clustering is the method we chose to use because it effectively identifies clusters of data points that are close to each other while also providing metrics of the relationships between separate clusters. This idea is how clusters can be judged based on how cohesive they are. The more cohesive the cluster, the more likely that a particular targeting strategy will be effective for that group of people. Ideally, there is a large cohort of customers that fit tightly into a series of characteristics who are also driving company revenue by a significant amount. Targeting this hypothetical group of consumers and molding promotional strategies that are appealing to this population would boost business greatly. Unfortunately, the existence of a group of this magnitude is not a given, which is why it is necessary to get creative and target multiple groups with multiple business strategies.

Business Goals

Whereas clustering and segmentation are used to place consumers into boxes, profiling provides hints on how and why a set of consumers have similar behavior. Understanding the mindset and situation of a company's customers is paramount. If a company is unable to meet the demands of its established customer base, it is inevitable that there will come along a competitor that will meet those demands and poach consumers. If it is a struggle just to retain existing customers, naturally attracting a new customer base becomes more difficult. Fundamentally, a loyal and firmly established population of 'regulars' is the foundation for business survival and growth. The most valuable customers are those that offer the most currency to a company, whether that be in the form of monetary value or attention. Attention cannot be quantified in an airtight manner, however there exist engagement metrics for digital forms of outreach which include social media.

Consolidating the attention of consumers is beneficial primarily to attract a new customer base using the outreach of an established consumer base. There is potential for explosive growth with this type of strategy, beyond just keeping an initial customer base intact. Harnessing and interpreting this information can pay massive dividends for a telecommunications company that is making its first foray into the analytics age.

Data Description

The curated “Customer_Dataset_File” provided consists of 5,000 rows and 59 columns. The objective of this exercise is to perform proper statistical analysis and segmentation on this dataset to unveil customer habits and behavior for targeted marketing. This dataset presents with a number of missing values, categorical type data and unrefined raw data. To prep the dataset for analysis, the missing data must be filled with realistic figures, the categorical data must be converted to numeric values, and further feature engineering must be done to increase the amount of valuable derived data.

The categorical data is converted to numeric values. This converted data goes through one of two methods of recoding. The values that are purely categorical within context (e.g Gender, Votes) are recoded with numbers starting with 0. For a variable with outputs of “No/Yes”, these outputs were recoded to 0 and 1, respectively. This recoding convention is consistent throughout all attributes with the “No/Yes” output. The attributes with more than 2 possible outputs (e.g JobCategory, CreditCard), were assigned numeric values starting from 0 in alphabetical order. There will be no arithmetic operations done on this type of categorical variable, because the context and meaning behind the numerical figures would be destroyed and become obsolete. On the other hand, the values that are classified as characters, but are interacted with numerically in the real world (e.g HHIncome, CarValue), can be manipulated with arithmetic operations to provide greater value in the form of derived data. After these variables are converted from character to numeric, the sum of a series of similar singular attributes can be computed to derive new attributes of cumulative data (voice, equipment and data). These cumulative values can be handled arithmetically.

After plotting histograms of each of the singular and cumulative attributes, it is clear that most of the plots are skewed to the right. In order to perform statistical analysis, the ideal type of distribution to work with is the normal distribution. Since most of these attributes are right skewed distributions, running the raw data through the logarithmic function yields a nearly normal distribution when plotted. The probability distribution and traits of the normal distribution are simpler to work with than the right skewed distributions of varying degrees.

While this piece of feature engineering is not very meaningful in its numerical form, the analysis that can be done with this type of distribution will allow for more meaningful inputs into the model. Adjusting for the skewness of the data also transforms the data to include a variety of data points, instead of the majority of the data points clustered together at one value (which is the fundamental difference between the normal distribution vs. any skewed distribution and why a normal distribution is preferred). Having both the raw and transformed versions of each given attribute also introduces versatility in data analysis. Having the transformed version allows the feasibility of methods such as t-tests and calculating z-scores, through the presence of constant variance. However, having the raw data on hand opens up convenience to perform real time arithmetic operations with true real world figures.

Most of the missing values in this dataset have been imputed with the corresponding median of an attribute. The purpose of choosing to impute the median instead of the mean is to discount outliers in the data. A sample of 5,000 is large (in the context that a t-test would most likely not be appropriate), so outliers are less likely to hold value in predictive modeling, as opposed to being more significant in smaller sample sizes. In hindsight, the MICE imputing technique would have been the most ideal. MICE can make the distinction between which variables are to be used as predictors for missing data and which variables should be imputed. Going through the attributes of this data and imputing the vast majority of the missing values with just a singular measure of central tendency might boost the strength of the relationships between variables, but overall decreases a model's ability to make nuanced calculations to output the most accurate predictive figures. Since outliers are less likely to significantly impact a large set of data, the MICE method might be the preferred method because it provides more variability without straying into the realm of producing outliers.

There are still 13 missing values left in this feature engineered dataset, all of which are present in the HomeOwner attribute. This attribute has been recoded from categorical to continuous, through the "No/Yes" to 0/1 series of transformations. Both inputs were well-represented, so choosing just one input to impute into the missing data did not feel reasonable. In a dataset of 5,000 observations, 13 inputs going one way or the other might not seem significant but the comparison should be made with context and the other variables within the dataset in mqaind. Imputing missing data not only alters the proportions of the inputs, but also affects relationships with other variables, which means any segmentation analysis will be affected as well. If a need arises to impute data into these 13 cells, it can be considered later on.

The final count of attributes in the revised dataset is 89, an increase from the original 59 columns.

Methodological Summary

This exercise is based on the Customer Database dataset. There is a lack of a data dictionary and metadata information, which would have been useful in gleaning further context and actionable insights with how to further implement the recommendations presented here. However, since the objective is to identify clusters that both make sense and are practical for profiling, it serves the purpose better without the bias of knowing which attributes are connected. A comprehensive data model may have been a useful tool to draw conclusions after segmentation, but that is an opportunity for further follow-up.

RStudio is the preferred IDE for this analysis, as it provides ample support and functionality to perform data due diligence, set up clustering and generating plots. R language is a compatible pairing with this IDE. The code itself could be more concise and elegant, as there are blocks of code with similar structure throughout which could have been replaced with defined functions using an input/output system.

In particular, the variables used for this clustering were *"TotalDebt", "EmployeeLength", "NewsSubscriber", and "highValueTenure." "highValueTenure"* is the mutated attribute that is the backbone of this exercise. A value of 0 was given to the observations that came in lower than the median of another mutated attribute, *"TotalTenure."* Conversely, a value of 1 was given to the observations that came in greater than the median of the *"TotalTenure"* column, which is calculated to be $2,809. Using this method is a straightforward way to split consumers based on low-value consumers and high-value consumers. The population for each group ended up being almost equally split, with there being 2,503 customers labeled as low-value and 2,497 customers labeled as high-value. Having a nearly identical sample size is beneficial since we do not have to address a small sample size and the disclaimers that come with it. So with the focus being on high value customers, the three aforementioned variables were chosen to be a part of this segmentation.

The next step requires a set up of the K-means segmentation algorithm. The initial set up was done in clusters of 5, 6 and 7. These were plotted using the **fviz_cluster** function, which is a part of the **factoextra** package (Plots A and B). This function is able to plot an entity of four variables into a two-dimensional plot. This particular plot is expressed in units of Principal Component Analysis (PCA). PCA is used to summarize the information in the multivariate data by reducing dimensionality while keeping the correlation of the data as high as possible.

Since there are three viable candidates for clustering, optimization is required to find which set of clusters fit the data best. This is done with the gap method and silhouette plot techniques

(Plots C and D). Based on the results of these techniques, an ideal number of clusters is given. If both techniques give the same result, this would give credence that there is a specific number of clusters that works best for this data. However, if the number of clusters given by both techniques are different, content would need to examined to determine what number of clusters would illustrate the segmentation best for our needs. The gap plot suggests a cluster number of four, while the silhouette plot suggests a cluster of five. Both are studied in this exercise. Comments are added prior to each block of code, with a brief explanation providing the thought process behind the code. The RStudio console served as a convenient output display for experimentation throughout the process, while the global environment stored all of the data frames, clusters and statistical tests in a clear manner. The window to display plots proved handy for cycling between iterations of plots quickly.

Results

The results of this exercise were a mixed bag. We were able to identify a pipeline through which we can appeal to established and prospective consumers alike, however categories such as *"TotalDebt"* and *"EmploymentLength"* were not able to contribute as much to the result. *"TotalDebt"* is defined as the sum of a consumer's credit debt along with other debt they have on file. A pattern shows that consumers with a higher total debt tend to also have a high employment length. It is uncertain what kinds of conclusions can be drawn from this observation, however the prevalence of higher debt suggests generous payment habits. The tables show a trend with the relative value of total debt and employment length being on a similar level. Regression analysis may show a significant relationship between this couple of variables, which could an avenue to look into. We may also be able to assume that a higher employment length would correlate with a higher age. This group may be where a sizable portion of senior citizens are clustered. Including an age metric to the clustering formula would confirm these suspicions, which in turn would inspire a more solid conclusion on how to appeal to this demographic.
Unfortunately, the aforementioned variables do not show a connection with the *"NewsSubscriber"* variable. We would generally prefer to target the third cluster in the 5 cluster exercise, as this group displays higher than average metrics in all three variables. The most promising component of this cluster is its size. A higher cluster population is excellent in both theoretical and practical terms. A greater sample size correlates with replicable and sustainable data, which can be leveraged into an important pillar for the company's revenue. A greater sample size also lends to practical benefits in terms of outreach and advertising.
For the four cluster group, we can follow a similar strategy as the one above by focusing on cluster number two. This cluster also has a high prevalence of total debt and employment length,

but outreach is low and is it not a sizable group. The size of cluster three in the 5 cluster groups is approximately five timer more than the size of cluster two in the 4 cluster group. A more reasonable group to target may be cluster one or cluster three in this set of clusters.

Recommendations

Forming educated inferences on customer behavior in tandem with this data will be able to provide valuable insights into how our consumers are thinking and behaving regarding the services we provide. The first subset of high value consumers likely pay a large fee for several services, but may have not been contracted to the company for long. The second subset is comprised of consumers that have been loyal to the company for several years, but may not take the liberty to use all of the services offered. No matter which group these customers fall into, they are defined as high value consumers in this exercise. Without having a method to distinguish the time each customer has been contracted to the company, it would be wise to take a broad approach to satisfy both types of consumers. Offering a loyalty incentive for the first subset of customers would be sensible, along with sending promotional offers for various services to the second subset of customers. There does not need to be a distinction in the advertising between these two groups because they fall under the general umbrella of being a high value consumer.

The third cluster in the five cluster group is made of 2,754 consumers (Table A). This is more than 50% of the total dataset that was initially collected. Priority number one is to appeal to this batch of consumers. This batch of customers also has a relatively high outreach measure, as quantified by the *"NewsSubscriber"* attribute. Almost 83% of this group is subscribed to the news service. Focusing on making competitive offers to this cohort will be a lucrative endeavor. It would be possible to take into account total debt and employment length when interacting with this group, since these relative measures for this group are also high. It may be worthwhile to collaborate with companies in other industries that appeal to this group of people, such as providing benefits for employee related achievements or a form of debt relief. Groups with a high or baseline debt score would be worth reaching out to, since we can assume due to their spending habits that they might be wiling to pay more for services as opposed to those groups with a low debt score.

In the group of four clusters, it would not be advised to pursue the same strategy as the one for the five cluster group. The close equivalent of this cluster in the four cluster group would cluster number two (Table B). It would not be worthwhile to pursue this group because it displays similar characteristics to cluster three in the five cluster group, so there is likely to be overlap, since both groups are taking data points from the same dataset. Another reason is that this cluster

is just not as promising as cluster three in the five cluster group because it has a low news subscription value. This limits outreach when we already have access to a group with a high outreach score. The best cluster to pursue in the group of four might be cluster number three. The main rationale for this is the high outreach score, similar to the best cluster in the group of five. Another potential positive is the low length of employment. It may be that this demographic is of a lower age than the rest, which would open up completely different campaign and promotional strategies. This group can be seen as an investment as opposed to a short term payoff, building up loyalty and goodwill with younger consumers to establish a future base of high value customers.

In short, cluster three in the group of five should be seen as strengthening the foundation of the company and firmly solidifying a loyal base. Cluster three in the group of four should be seen through a future oriented lens, as a method of growth and outreach. In-company promotions would be effective for the former group, while referral offers would be more ideal for the latter.

Appendix

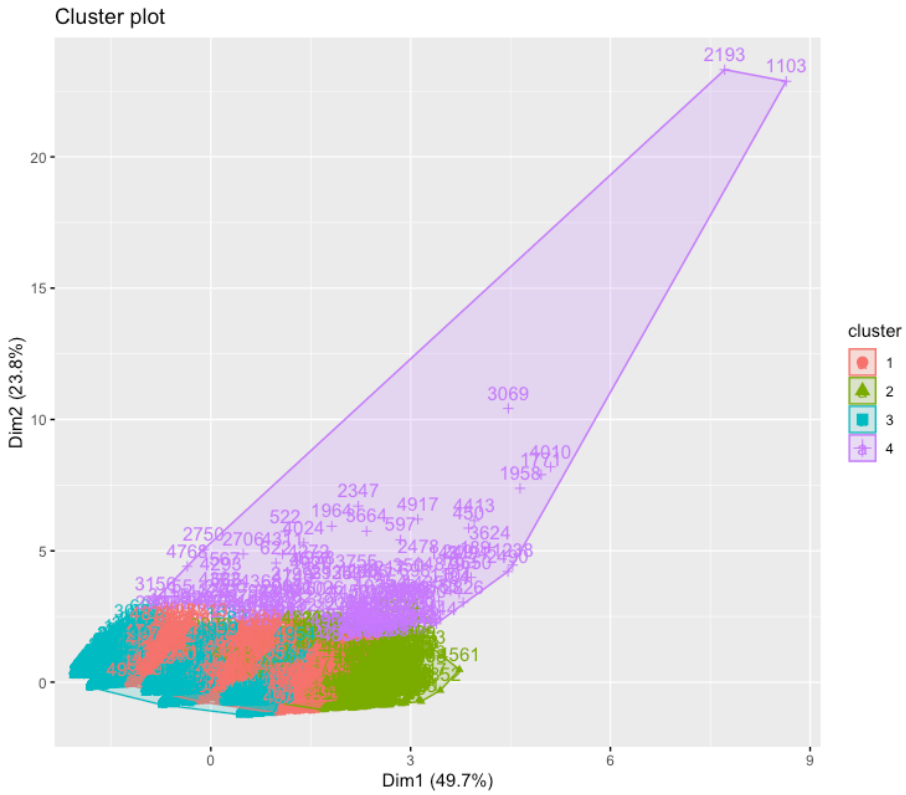
Table A

ClusterNumber	Size Rank	TotalDebt	EmploymentLength	NewsSubscriber
1	5	Baseline	Baseline	Baseline
2	2	Low	Baseline	Baseline
3	1	High	High	High
4	4	Low	Low	Low
5	3	Low	Low	Low

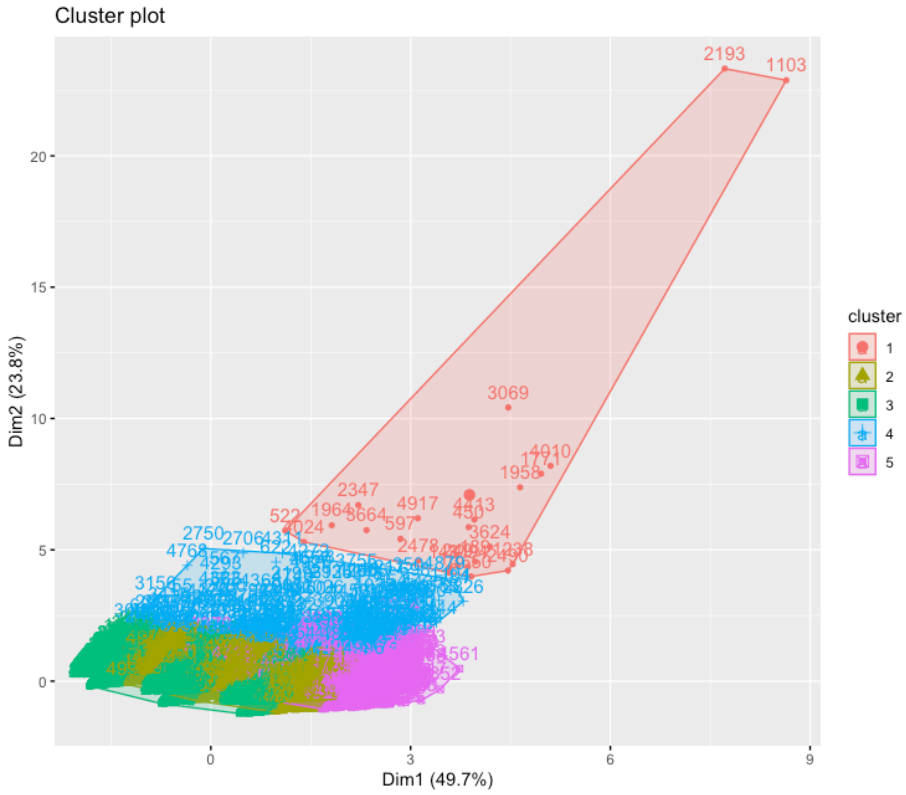
Table B

ClusterNumber	Size Rank	TotalDebt	EmploymentLength	NewsSubscriber
1	2	Baseline	Baseline	Baseline
2	3	High	High	Low
3	1	Low	Baseline	High
4	4	Low	Low	Baseline

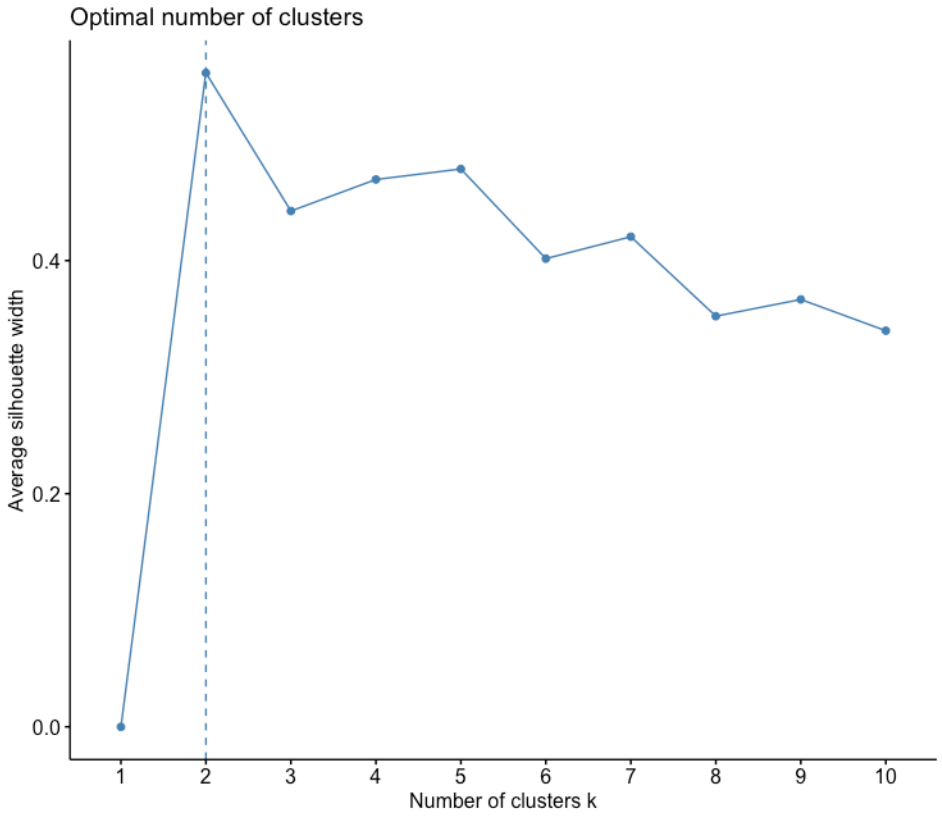
Plot A


Cluster plot
2193
1103
3069
4010
1958
2347
4917
4413
522
1964
3664
597
450
3624
4024
2750
2706
4768
2478
1561
Dim2 (23.8%)
Dim1 (49.7%)
0
5
10
15
20
0
3
6
9
cluster
1
2
3
4

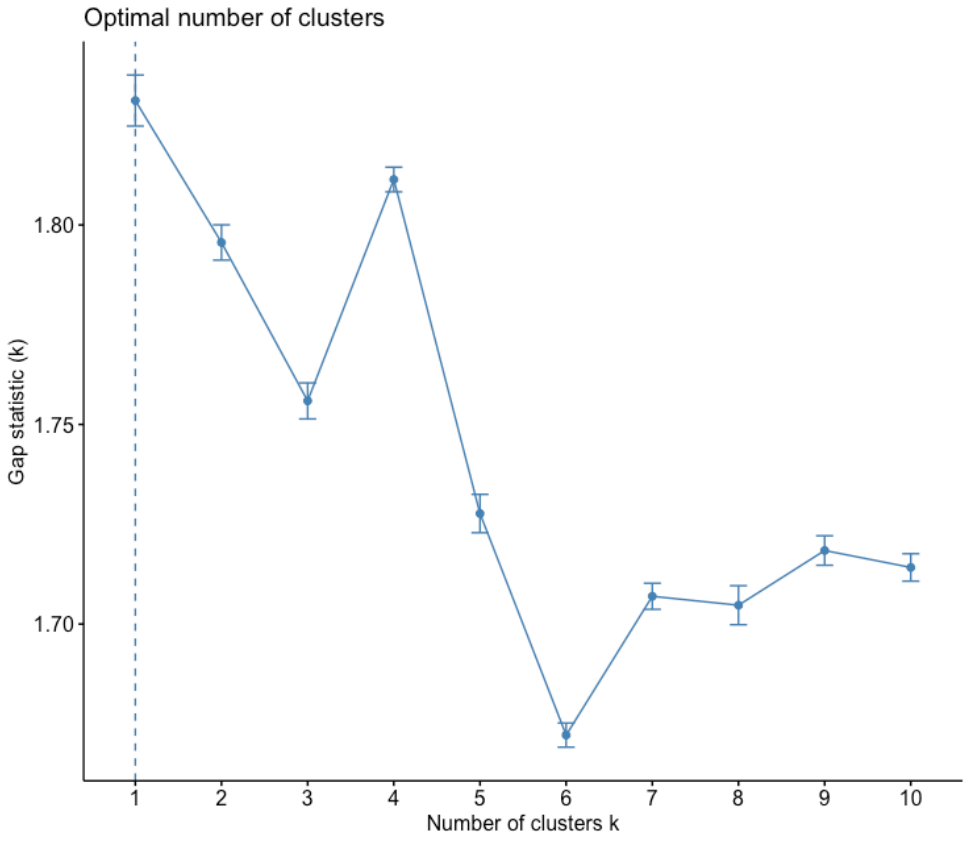
Plot B


Cluster plot
2193
1103
3069
4010
1958
2347
4917
4413
522
1964
3664
597
450
3624
4024
2750
2706
4768
2478
1561
Dim2 (23.8%)
Dim1 (49.7%)
0
5
10
15
20
0
3
6
9
cluster
1
2
3
4
5

Plot C


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Plot D


Optimal number of clusters
Gap statistic (k)
1.70
1.75
1.80
1
2
3
4
5
6
7
8
9
10
Number of clusters k