CSP571: Data Preparation & Analysis

Project Report

on

Customer Segmentation

Submitted to: Professor Jawahar Panchal

Submitted By:

Albert Mandizha(A20493341), Carlos Muñoz Losa(A20521562), Chao Yang(A20511318), Talvinder Kour(A20489751), Yuanyuan Sun(A20487775)

Illinois Institute Of Technology, Chicago

Index

Index of figures

Abstract

In the extremely competitive environment in which the companies are involved it is essential to obtain an advantage over your rivals in today's fast-paced socioeconomic environment. It is also essential to categorize a company's customer base and analyze customer behavior. This primarily empowers one's marketing and strategy team to take action. Customer segmentation should be based on four types of dataset: demographic (age, gender, marital status, income, education), geographic (specific or global location like cities, countries based on the company's objective) , psychographic (qualities, class) and behavioral data (need, demand, usage and consumption). In our project, we choose a data set from Data Flair which is about the customers of a small mall. In this data set, there are some basic and necessary features. We have used 200 observations with 5 variables to do our proposal of customer segmentation. Businesses can target the prospective user base by using clustering techniques to find various client segments. In this project, we have used K-means clustering, a crucial algorithm for grouping unlabeled information, to investigate marketing-relevant factors including gender, age, interests, and other buying patterns. We have attempted to assess the model's performance by computing BSS/TSS ratio in order to assess K-Means clustering accuracy. The goal of the Elbow Criterion approach and the "silhouette" method is to identify the optimal clusters. Defaulty, we prefer cluster no. (k) at which the SSE rapidly decreases, but for clearer clusters on age presentation and better visual effects on three-dimensional diagrams, we need a model with better accuracy on relatively small data sets.

Keywords: Customer Segmentation, Machine Learning, K-means, cluster analysis

Introduction

Businesses may successfully and appropriately market their products to target efficiently different clients. Pre and Post covid pandemic, market has always targeted customers with augmented and evolved with most efficient marketing strategy, segmenting the audience in different subgroups, understanding and meeting their needs. Right audience in right location and providing right communication is the key element for any business to generate profit and revenue. Industries analyze target customers based on customer segmentation on different subgroups and tailor ad messages based on interest and needs of consumers who are first timers and focus on retaining customers. Further companies focus on the best marketing channel to influence and create needs amongst the customers. Businesses can target the prospective user base by using clustering techniques to find various client segments. Based on companies' service or product, advertising is required to understand the consumer behavior and purchasing pattern of customers, which might require traditional or digital marketing strategy.

- Determine opportunities for new or improved products or services
- Develop stronger client ties
- Enhance customer service and client retention; cross-sell and upsell additional goods and services.

Customer segmentation is the dividing of a market into distinct groups of customers who have common traits. Client segmentation can be an effective tool for locating unmet customer needs. Companies can then exceed the competition by creating distinctively appealing products and services using the data mentioned above. Below are the research questions which this project would give analysis in details as mentioned below:

- How can a company improve the customer experience?
- How can the company predict the behavior of its customers?
- How can the company increase client retention and loyalty?
- What recommendations are required to enhance the conversion?

Finally it is worth telling that all the code, which we have developed for our project, is available in a github repository https://github.com/greensure/CSP571 the final version is named "CSP571 Project R code-Ver6.Rmd" and it is possible to find the html version generated in RStudio.

Literature review

(Clustering Algorithms for Bank Customer Segmentation) The rapid development of data mining methods enables using large databases of customer data to extract knowledge, supporting the marketing decision process. As the ability to acquire new clients and retain existing customers is crucial, especially in the finance marketplace, the possibility of customer segmentation by obtaining information on unknown hidden patterns has a big significance (Zakrzewska and Murlewski, 2005). (Customer Segmentation Architecture Based on Clustering Techniques) Knowledge of consumer habits is essential for companies to keep customers satisfied and to provide them with personalized services. We present a data mining architecture based on clustering techniques to help experts segment customers based on their purchase behaviors (Lefait and Kechadi, 2010). To run a company or big customer service organizations, some basic information about your clients, such as Customer ID, age, gender, annual income, and spending score.

Data sources

Our data is stored and taken as in a Google Drive CSV file. We will use our entire dataset to study and analyze the customers. We have a 5 feature dataset with a population of 200 samples. The data summary is presented in the table below, presenting each feature, its type and a short description of its meaning.(Data Set)

Field Name	**Data type**	**Description**
CustomerID	String	A unique identifier for each customer
Gender	String	Describe the gender with value of Male/Female
Age	Int	Describe the age of the customer
Annual Income	Int	Describe the annual income of the current customer
Spending score	Int	Describe the spending score of the current customer achieved

Brief overview of the research methodology:

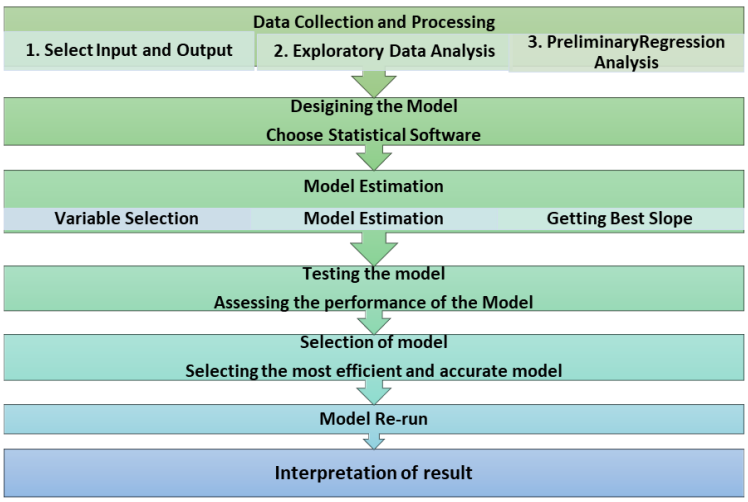

Figure 1. Methodology

1. Clean Environment & Load Libraries

The environment is cleaned from previous variables. We used different libraries in the code for different purposes: readr, corrplot, naniar, ggplot, plotrix, carex, cluster, factoextra, fpc, gridExtra and scatterplot3d etc.

2. Source, Collect and Load Data

2.1 Source and Collect Data

The goal of 'data sourcing and collection' is to find data that is relevant to solving the problem or supports an analytical solution of the stated objective.

This step involves reviewing existing data sources and finding out if it is necessary to collect new data. It may involve any number of tasks to get the data in-hand, such as querying databases, scraping data from data streams, submitting requests to other departments, or searching for third-party data sources. Here, we collect the data from third-party. We get the data set from Data Flair.

2.2 Load Data

We load the data and find the structure of the data set with 200 observations with 4 integer variables and one factor variable with str() function. We have used the summary() function to see the quartile ranges, mean and median values.

```
##    CustomerID        Gender         Age        Annual.Income    Spending.Score
##  Min.   :  1.00   Female:112   Min.   :18.00   Min.   : 15.00   Min.   : 1.00
##  1st Qu.: 50.75   Male  : 88   1st Qu.:28.75   1st Qu.: 41.50   1st Qu.:34.75
##  Median :100.50                Median :36.00   Median : 61.50   Median :50.00
##  Mean   :100.50                Mean   :38.85   Mean   : 60.56   Mean   :50.20
##  3rd Qu.:150.25                3rd Qu.:49.00   3rd Qu.: 78.00   3rd Qu.:73.00
##  Max.   :200.00                Max.   :70.00   Max.   :137.00   Max.   :99.00
```

3. Process and Clean Data

In this step, we will look for data errors, missing data, or extreme outliers etc.

3.1 Explore missing value patterns

Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples (Kang, 2013). Many machine learning algorithms fail if the dataset contains missing values. To avoid ending up building a biased machine learning model which will lead to incorrect results if the missing values are not handled properly, we will check missing values and handle missing values in this step. The way we could handle missing values is using mode instead NA or NULL. According to the output below, we see data in the data resource is quite straightforward and clean with no missing values.

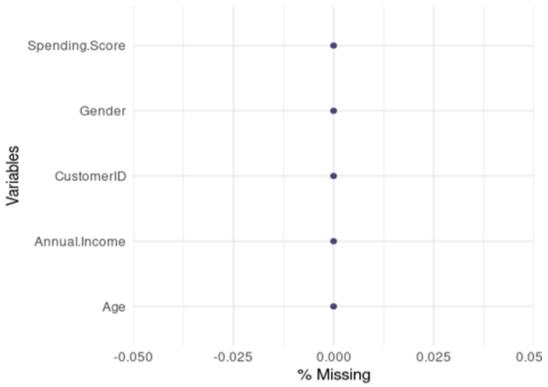

Figure 2. Missing Values

The number of NA values in this data set = 0. There are no NA values in the original data set. So we have a complete dataset and we can use all the samples.

3.2 Features Data Summarization

Basic statistical summary reports and charts can help reveal any serious issues or gaps in the data. Before we do feature engineering and more data processing, we firstly check the basic information in the original data set. In the output of the summary, we can see from the distribution of gender on the customer data set that the amount of female customers is larger than the amount of male customers. We see the annual income, age and spending score of customers varies widely. We would be interested the following problems:

1. Can we segment the spending power of customers based on some certain features?
2. Which features could impact customer segmentation significantly?

3. Can we narrow down the customer clusters based on certain features for better marketing strategy?

Based on the summary and something interesting here, we will do feature visualization to get more information in the following steps.

3.3 Data parsing (Set dummy variables)

Data parsing is converting data from one format to another. Widely used for data structuring, it is generally done to make the existing, often unstructured, unreadable data more comprehensible. We find 'Gender' is binary in the original data set. So, it's better to set the dummy variable on gender for further data processing and analysis. We assigned the value 1 to the male samples and 2 to the female samples.

CustomerID <int>	Gender <int>	Age <int>	Annual.Income <int>	Spending.Score <int>
1	2	19	15	39
2	2	21	15	81
3	1	20	16	6
4	1	23	16	77
5	1	31	17	40
6	1	22	17	76

Figure 3. Data parsing (Gender)

3.4 Checking & Treating Outliers

Dealing and Treating Outliers in data is always essential for accurate results. According to the boxplot below of the original data set, we find the original data set has outliers in Annual.Income feature. The outlier is out of the confidence interval. For extreme values like outliers on continuous features, there would be a huge gap on mean and median values. The outliers in our data could push up the mean and median values here. If we want to know the practical average or median on Annual.Income of the customers.

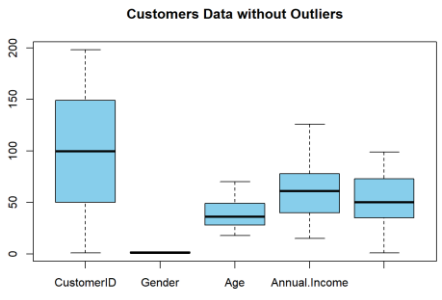

Figure 4. Boxplot with outliers

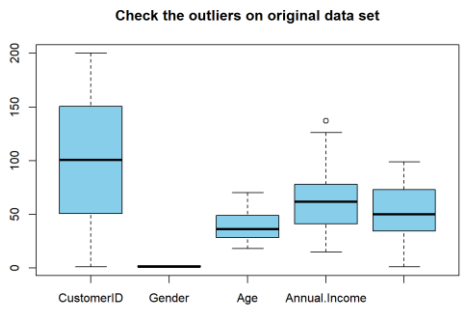

Figure 5. Boxplot without outliers

3.5 Data Correlation

In statistics, we're often interested in understanding the relationship between two variables. One way to quantify this relationship is to use the Pearson correlation coefficient, which is a measure of the linear association between two variables. It has a value between -1 and 1 where:

-1 indicates a perfectly negative linear correlation between two variables

0 indicates no linear correlation between two variables

1 indicates a perfectly positive linear correlation between two variables

According to the correlation matrix, we see the correlation values between two different variables are almost close to 0. The features of each other are lowly correlated. It indicates no linear correlation between two different features on the original data set.

Figure 6. Correlation Matrix

4. Data Transformation

4.1 Data Transforming (Remove ID)

Data transformation is a technique used to convert the raw data into a suitable format that efficiently eases data mining and retrieves strategic information. Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form. According to the detection in previous steps, we found customer ID cannot be considered as a feature because it has no meaning here since it only represents each of the customers to distinguish between them with no further meaning. It's better to remove it.

Gender <int>	Age <int>	Annual.Income <int>	Spending.Score <int>
2	19	15	39
2	21	15	81
1	20	16	6
1	23	16	77
1	31	17	40
1	22	17	76

Figure 7. Data Frame without Customer ID

4.2 Formating Final Data Frame

According to the display of the internal structure of the data set. All features in the data frame are integers. There's no need to format columns.

4.3 Normalizing the Data

Since clustering techniques use Euclidean Distance to form the cohorts, it will be wise e.g to scale the variables having heights in meters and weights in KGs before calculating the distance. In machine learning, feature scaling is useful in situations where a set of input features differs wildly in scale. If some of those features are thrown into a model, then the model will need to balance its scale while figuring out what to do. Drastically varying scale in input features can lead to numeric stability issues for the model training algorithm. In those situations, it's a good idea to standardize the features. Clustering algorithms such as K-means do need feature normalization before they are fed to the algorithm. We have scaled all the continuous features on the data frame which is for training models.

Gender <int>	Age <dbl>	Annual.Income <dbl>	Spending.Score <dbl>
2	-1.4210029	-1.734646	-0.4337131
2	-1.2778288	-1.734646	1.1927111
1	-1.3494159	-1.696572	-1.7116178
1	-1.1346547	-1.696572	1.0378135
1	-0.5619583	-1.658498	-0.3949887
1	-1.2062418	-1.658498	0.9990891

Figure 8.Data Frame with scaling

4.4 Univariate Distribution

4.4.1 Histogram

Now we plot the histogram of each feature, the y axis represents the number of samples in the dataset and the x axis the value that each sample has. The peaks represent the most common values. The most common age in the customers is around 30-35. Compared to the kids and the senior people, there are more middle-aged consumer groups in this mall. The

mall might need to sell more goods which are popular among middle-aged customers or try to sell products with the brands that are popular among middle-aged people. This could both satisfy customer demand and increase sales

The most common values on ‘Annual.Income’ feature are around 80. We don't know the income criteria in this area or this country. For the further analysis, we need assistance from domain experts. If ‘Annual.Income’ which is around 80 means high income. The mall might need to adjust sales strategy for high consumer groups. For example, the mall might adapt his prices to the prime members and offer extra discounts on their purchase.

The most common spending score in this customer data is around 50. We are not clear about the local price level and the overall level of consumption over the years here. We also don’t know the spending score metric in recent 10 years in this city. For the further analysis, we need assistance from domain experts.

We will do data analysis in detail on the data visualization part later.

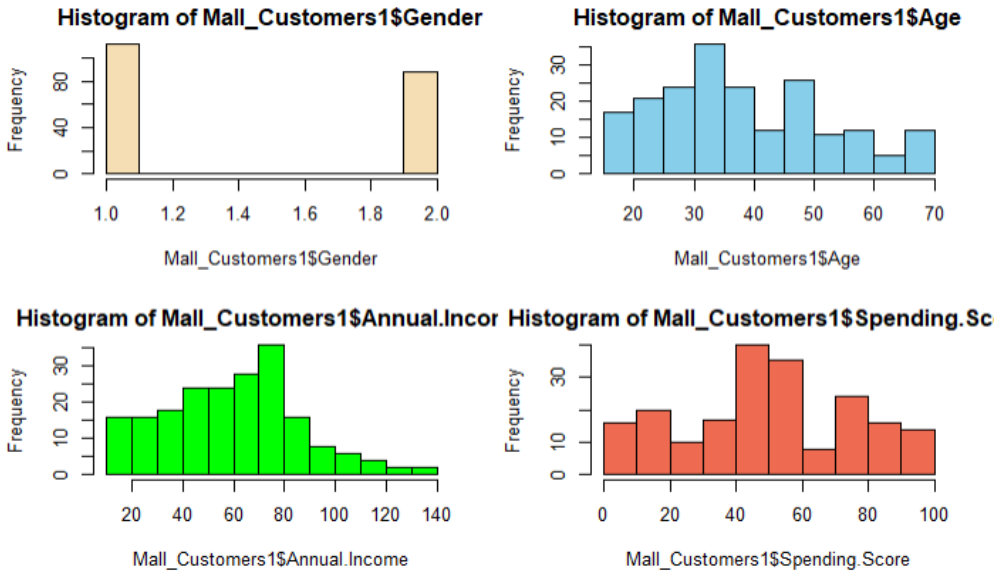

Figure 9. Histograms

Model Selection:

A common unsupervised learning problem is customer segmentation. Unsupervised algorithms draw conclusions from the data set without using known or labeled results, relying just on input vectors. Because no labels are given in the majority of unsupervised learning

approaches, it is difficult to compare model performance. Clustering like this Analysis, representation learning, and density estimation are typically the most common tasks in unsupervised learning. In each of these scenarios, we wish to uncover the underlying structure of the data without requiring labels to be given explicitly. Popular approaches include principal component analysis, autoencoders, and K-means clustering.

4.4.2 Kernel Density Plot

Density curves allow us to quickly see whether or not a graph is left skewed, right skewed, or has no skew. We see there are no features with their density that have no skew. Except for gender, the density plot of features which are 'Annual.Income', 'Spending.Score' are right skewed. It means the mean is greater than the median on features which are 'Annual.Income', 'Spending.Score'.

Looking in detail at these graphs we can see that the age and the Annual.Income features are skewed (the values are not equally distributed around the mean). Most customers' annual income is less than average. However, Spending.Score might follow a weird kind of normal distribution, its values are symmetrical around the mean value.

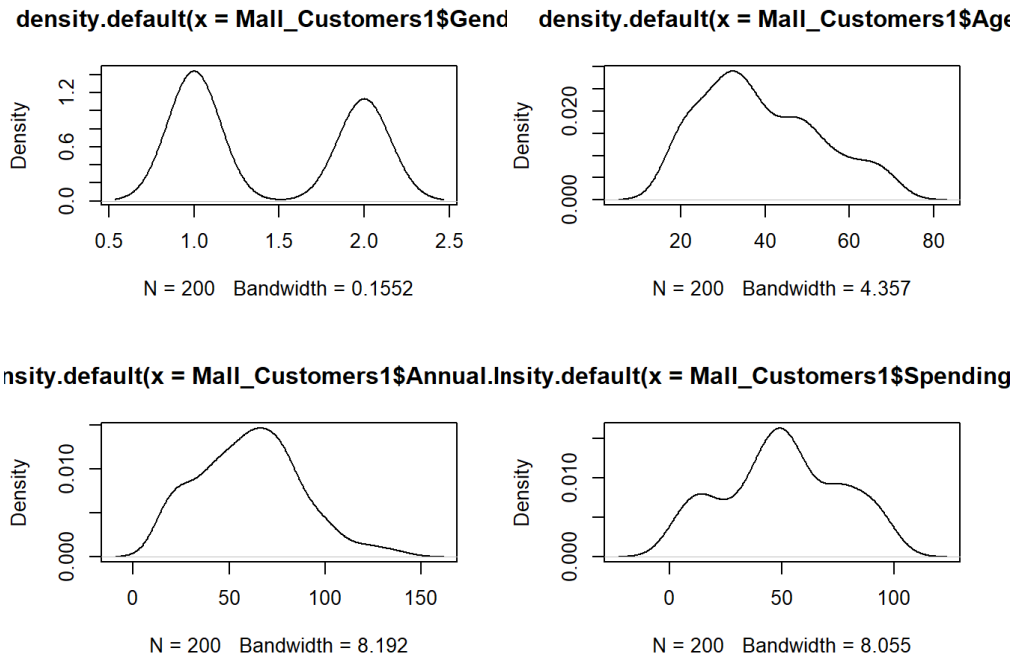

Figure 10. Density plots

4.4.3 Q-Q Plot

As we know that Q-Q Plot is just a visual check, not an air-tight proof, it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

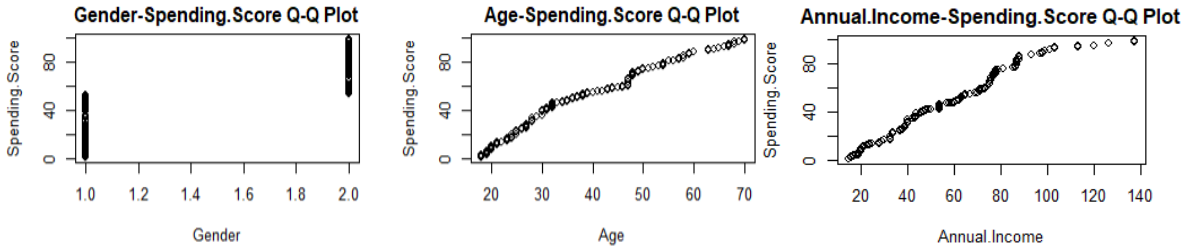

Figure 11. Q-Q plots

According to the output of the Q-Q Plot above, we see on (Spending.Score, age), (Spending.Score, Annual.Income) the points forming lines that are roughly straight. On (Spending.Score, gender), we clearly see the distribution is not normal. Non-normal distributions may lack symmetry, may have extreme values, or may have a flatter or steeper "dome" than a typical bell. There is nothing inherently wrong with non-normal data; some traits simply do not follow a bell curve. For example, data about coffee and alcohol consumption are rarely bell shaped. Based on the empirical experience, people who have higher Annual.Income more likely have higher Spending.Score. If we have time to figure out something among features, we would like to do linear regression on these features to see the relationships on (Spending.Score, age), (Spending.Score, Annual.Income), and compare it with the empirical distribution and normal distribution.

5. Features Visualization and Analysis of Features

5.1 Visualization of Gender

We can see that Population of Females is slightly more than Male. In the Pie Chart, it is shown that Female is 56% and Male is 44% in the whole Population.

According to the output of Kernel Q-Q Plot, 'Gender' has no relationship with Spending.Score. For marketing analysis and insights, we would like to see and compare the population of Female and Male.

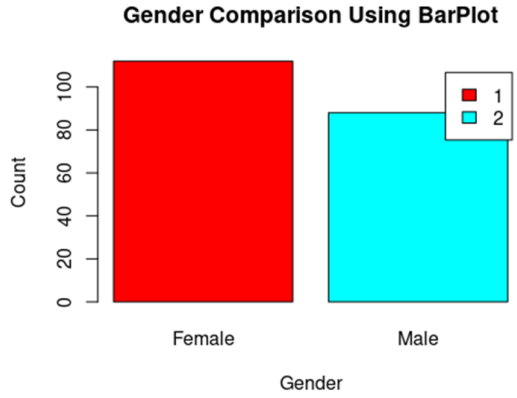

Figure 13. Gender Barplot

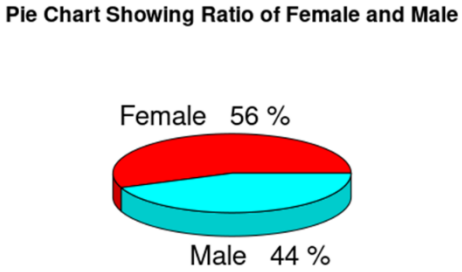

Figure 12. Gender Pie chart

5.2 Visualization/Analysis of Age

We can see that the maximum population is between 30 to 35 in the age group . We can also see Descriptive Analysis that Minimum age is 18, Maximum age is 70 and avg. age is 38.85.

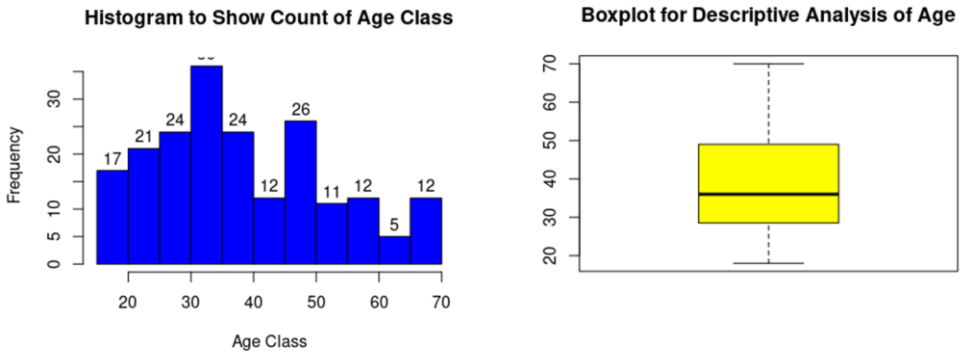

Figure 14. Age histogram and boxplot

Looking at the histogram established that we can have 5 segments of different ages which can be presented and targeted as follows

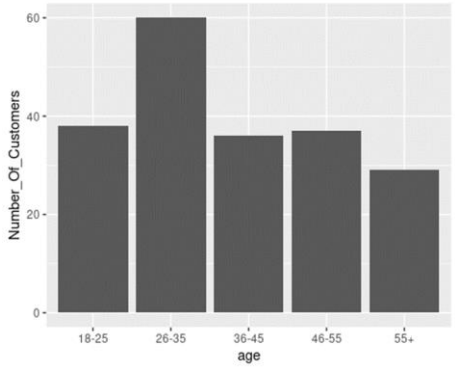

Figure 15. Age segmentation

The age distribution shows that most of our customers in the distribution are of the age range 26-35. This provides the business leaders with insights as to the age group to target for different marketing strategies.

The other age groups in our data present a more similar number of customers that are the age ranges of 18-25, 36-45, 46-55 and 55+. If the business decides to segment by age this figure represents groups that can derive the most value.A good example is a marketing team of a clothing company deciding on the taste of clothes to stock within their shop.

5.3 Analysis of Annual Income

We can see Descriptive Analysis that Minimum Annual Income is 15 and Maximum is 137 with an avg. annual income of 60.56 units. In the histogram, that Maximum Population has Annual Income between 70 to 80 units. We can see in the Kernel Density Plot of Annual Income, Annual Income is distributed almost normally.

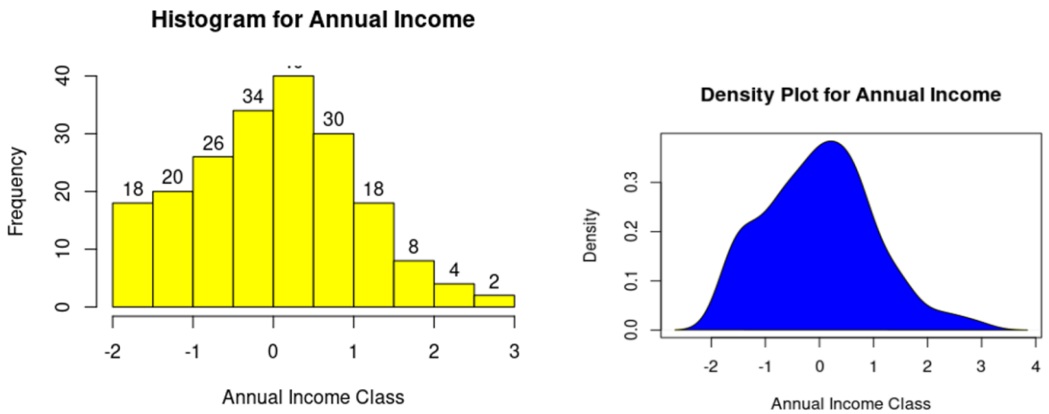

Figure 16. Annual.Income histogram and density plot

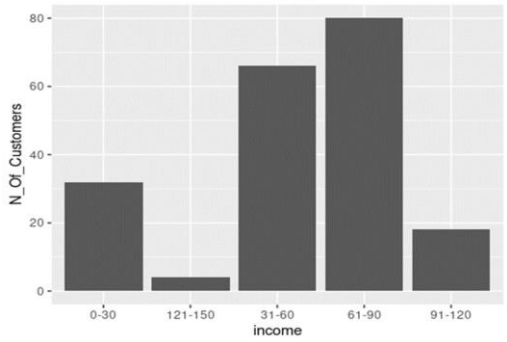

Figure 17. Annual.Income segmentation

The annual income follows a normal distribution as illustrated in the figure above. From our income data it can be noted that the number of customers visiting the mall have different levels of income with the 61000 to 90000 income range having the most number of customers.

From the bar plot it can be established that the customer base has an average annual income level between 61 000 to 90 000. As a business it's in the marketing team to understand these income levels so as to establish the target groups. For instance if our product is a high end product placing it at our shop at the mall will be ideal as the greater number of customers visiting the more have high income levels.

5.4 Analysis of Spending Score

We can see in the Descriptive Analysis of Spending Score that Min is 1, Max is 99 and avg. is 50.20. We can visualize Descriptive Analysis with BoxPlot. In Histogram that most people have a Spending Score between 40 and 50.

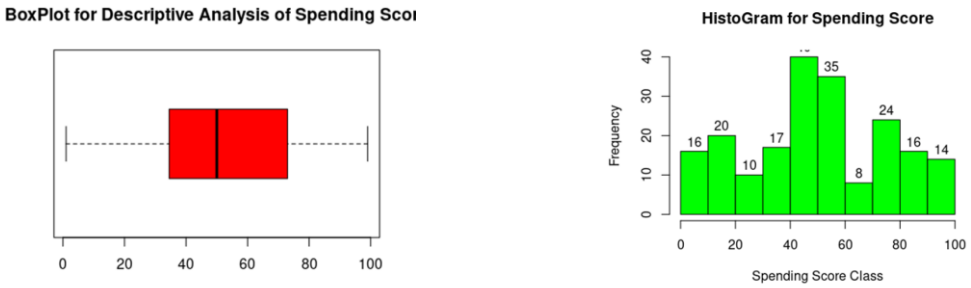

Figure 18.Spending.Score boxplot and histogram

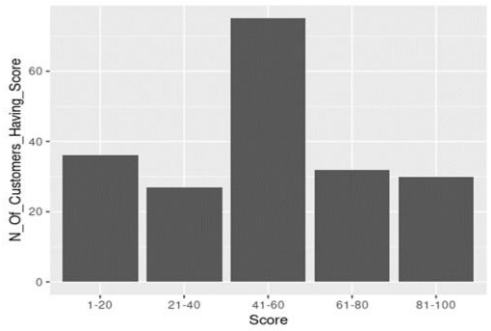

Figure 19. Spending.Score segmentation

From the bar plot we can see that the spending score of most of our customers is 41-50. Based on the sending score it enables the business to focus their marketing activities and campaigns on the customers that are most likely to respond favorably to their efforts. The customers in the range 1-20 and 21-40 provide market insights that as a business they need to embark more on penetrative and retention marketing strategies. These groups of customers are likely to be non-attractive to a cost centered business but rather a market share centered business.

6. Data Analysis Stage

For the data analysis we will use unsupervised learning algorithms such as Clustering. Clustering looks to find homogeneous subgroups among the observations. It refers to a very broad set of techniques for finding subgroups, or clusters in a data set. When we cluster the observations in a small data set of a mall, we seek to partition them into distinct groups so that the observations within each group are quite similar to each other, while observations in different groups are quite different from each other. Based on this characteristic, we will use K-means clustering to train our data set. K-means clustering is a simple and elegant approach for partitioning a data set into K distinct, non-overlapping clusters. In K-means clustering, we firstly need to seek to partition the observations into a pre-specified number of clusters. Hence, we first specify the desired number of clusters K as the optima via “silhouette” method and “elbow” method. The “silhouette” method and “elbow” method are visual checks. We can narrow down to find the optimal clusters. Since visual checks are not air-tight proofs, they are somewhat subjective. We use a measurement to check the model’s accuracy to choose the optimal one. The measurement we use is cluster means (between_SS/total_SS). It means combine to give the centroids (centers) of the clusters in the multivariate space defined by the input variables. Hence the set of means for cluster 1 that it shows are the coordinates of the centroid (center) for that cluster. They are computed as the mean of the values for each variable for those samples assigned to that cluster. During the pipelines of model fitting, we

are looking to find something interesting so that may answer the question: Are there any relationships among features since every feature is lowly correlated to each other?

6.1 Initial Exploration of Data Clusters

The first thing we have done is to explore the relations between the features to see if we can find any clues about what features will better segment the data points. In the plot we can see that the age feature does not clearly show any possible clusters, meanwhile annual income provides a set of possible groupings that we can study.

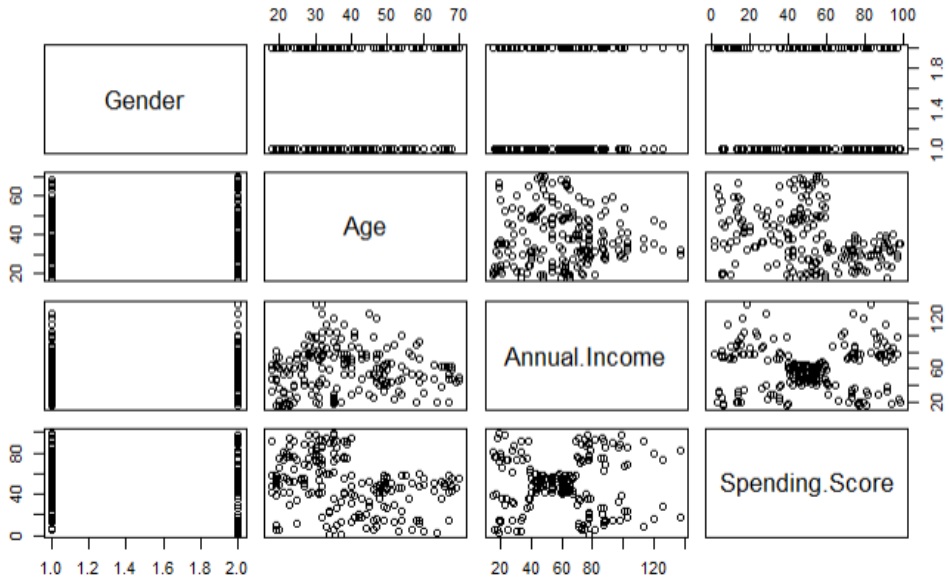

Figure 20. Data clusters plot

6.2 Choose the Optimal Number of Clusters

The first important decision that has to be made is how many clusters should we expect the model to split our points. It is a crucial decision because it will substantially affect our final conclusion. In the following points we will explain how we have implemented and interpreted the elbow and silhouette method.

6.2.1 The experiment on elbow method

Using the “wss” method, draw the screen plot for the optimal number of clusters. The location of a bend or a knee is the indication of the optimum number of clusters on the elbow method. We see the location of a bend could be k=3 or k=5 or k =6. However, the elbow method is somewhat subjective, different people may identify the elbow at different locations. Some may argue that k=3 or k=5 is the elbow, some may say K=6 is the elbow in the “wss” method plot. Moreover, the elbow may not always be apparent.

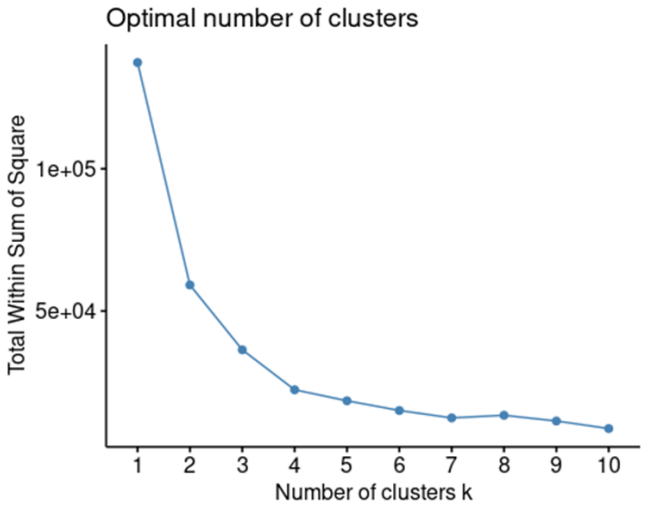

Figure 21. Elbow method number of clusters

6.2.2 The experiment on silhouette method

Looking at the resulting plot for the “wss” method we are not able to determine which is the optimal number of clusters that we may use. So we decided to use the “silhouette” method to draw the screen plot for the optimal number of clusters.

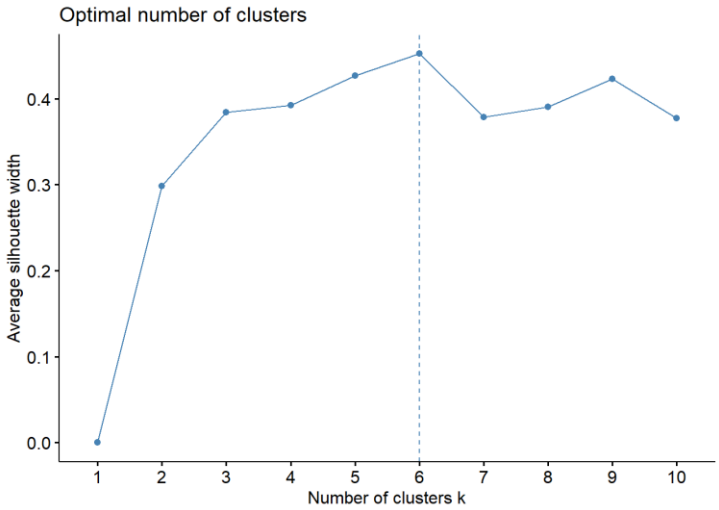

Figure 22. Silhouette method number of clusters

In the “silhouette” method plot, it seems there are two peaks. One of the silhouette coefficient peaks at K=3, and the other is at K=6. It’s a bit hard to determine the optimum K. The silhouette coefficient may provide a more objective means to determine the optimal number of clusters. It performs K-Means clustering over a range of k, finds the optimal K that produces the largest silhouette coefficient, and assigns data points to clusters based on the optimized K. But when there are more than one peaks, it’s a bit hard to choose the optimal K.

Experiment 1: We choose k=5 as the best number of clusters.

Experiment 2: We choose k=3 as the best number of clusters.

Experiment 3: We choose k=6 as the best number of clusters.

6.3 Model Fitting & Model Accuracy

6.3.1 Model Fitting Based on Experiments

In this part we will explain how we have done the comparison between the different clustering models. So we can visualize the plots and the metrics that each model provides. We used this method to evaluate the model we fit. Because we want to keep the WSS as small as possible (the distance between each point to the centroid of its cluster), therefore, theoretically, a high ratio of BSS to TSS is what we are looking for.

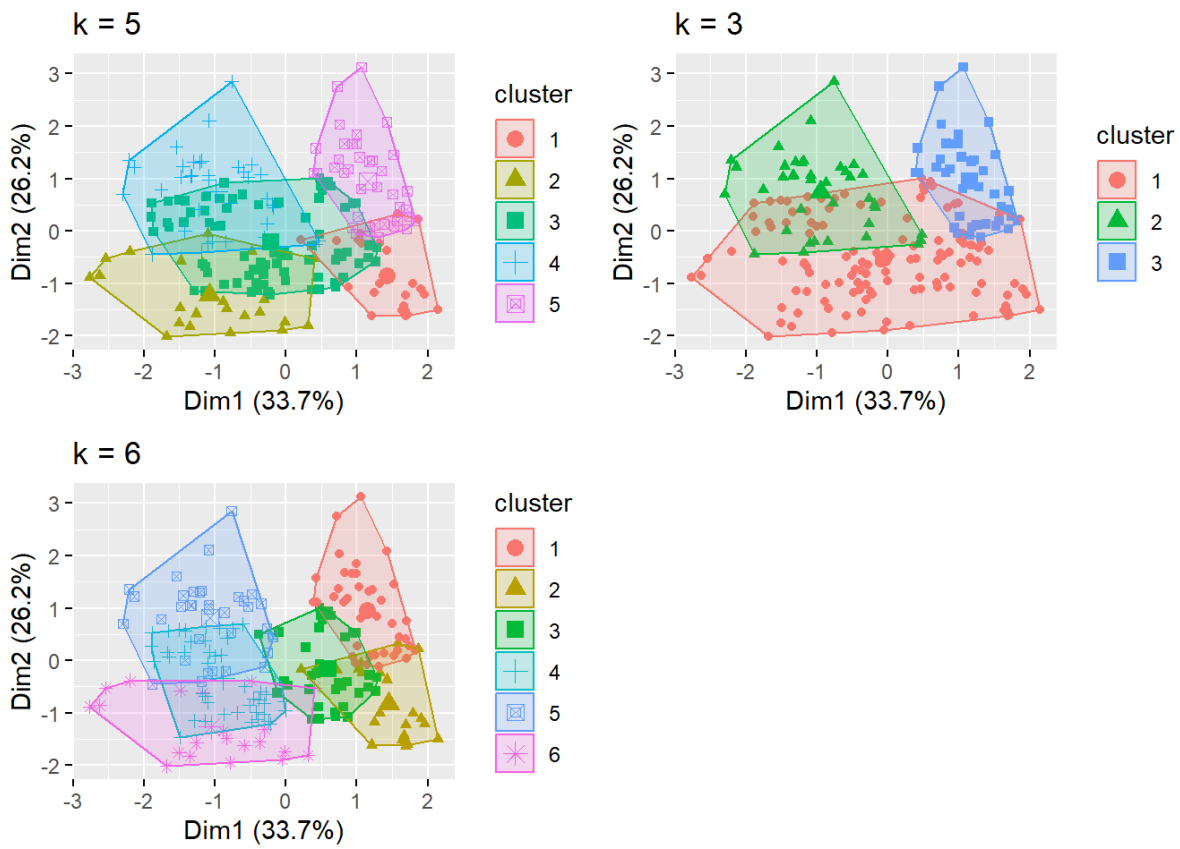

Figure 23. Clustering with k=5, k=3 and k=6

1. k = 5, the Between SS / Total SS is 0.756
2. k = 3, the Between SS / Total SS is 0.536
3. k = 6, the Between SS / Total SS is 0.811

When k=5, the Between SS / Total SS is 0.756. As we learned that Between SS / Total SS is between 0.0 and 1.0. The higher this ratio, the more variance is explained by the clusters.

Between SS / Total SS is 0.536 that indicates a bad fit. When k=3, We see the value on “Between SS / Total SS” is smaller than the value we calculated on experiment 1; Hence, k = 3 is not the optimal number of clusters. When k=6, We see the value on “Between SS / Total SS” is greater than the value we calculated on experiment 1; Hence, k = 6 is the optimal number of clusters. By running K-means clustering three times using 3 different initial, k = 6 is the optimal number of clusters.

6.3.2 Model Fitting on the Optimal Number of Cluster

Now we decided to fit the model using more features, so we included the age and fit another model. After that we plot the points drawing each point with the color of its cluster and we obtain a very visual representation of the classification. We can see that the points in the clusters are really close to each other but it is difficult to distinguish the boundaries. Using this visualization we can infer that each cluster corresponds to a different customer type. For example the green cluster represents customers with very low spending scores and there are not many examples so these customers are not worth targeting in a market strategy, meanwhile the yellow cluster represents middle aged customers with high annual incomes that have a relatively high spending score so that group could be potential customers worth targeting. This model performed not so well with a “Between SS / Total SS” value of 0.777.

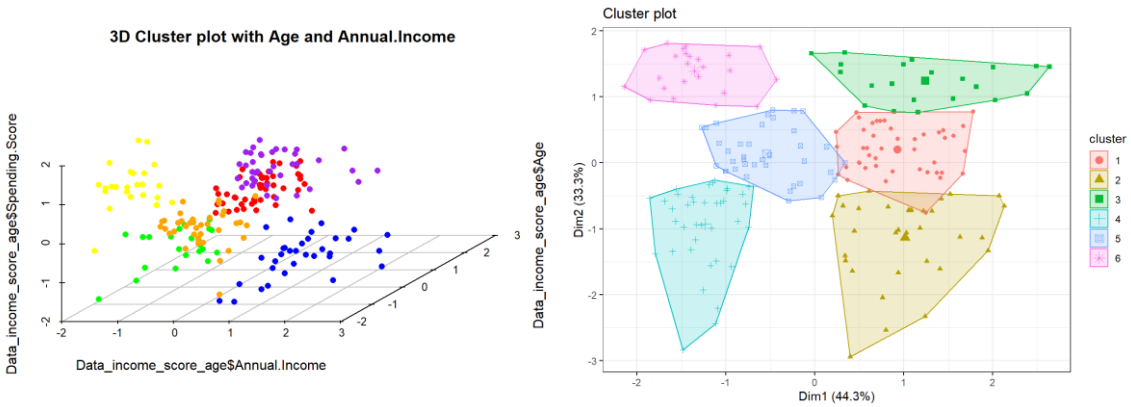

Figure 24. Clustering with Age-Annual.Income-Spending.Score

We also tried the model substituting the age by the gender. Plotting the scatter plot we can see that there are similar patterns in both genders and the clusters are well defined and the boundaries are very clear to see. So we can say that segmentation by gender could be a good idea. We can even say that we can make subgroups within the gender for example the most populated group is the female with around mean value for annual.income. It is very interesting to notice that the clusters are very similar in both genders with exception of the middle annual.income that in each gender those points have different color (belong to different clusters). So a strategy based on the number of sales might target that customers and the strategy must vary when targeting males and females. Also there are just 4 clusters that overlap in two dimensions so we can determine that the clusters are very well defined and the boundaries are very clear.

We can also see that presenting the data in two dimensions makes less clear the clusters because we are losing information and the clusters seem to overlap while in the 3D plot we can see that they don't. This model performed extremely well with a "Between SS / Total SS" value of 0.790.

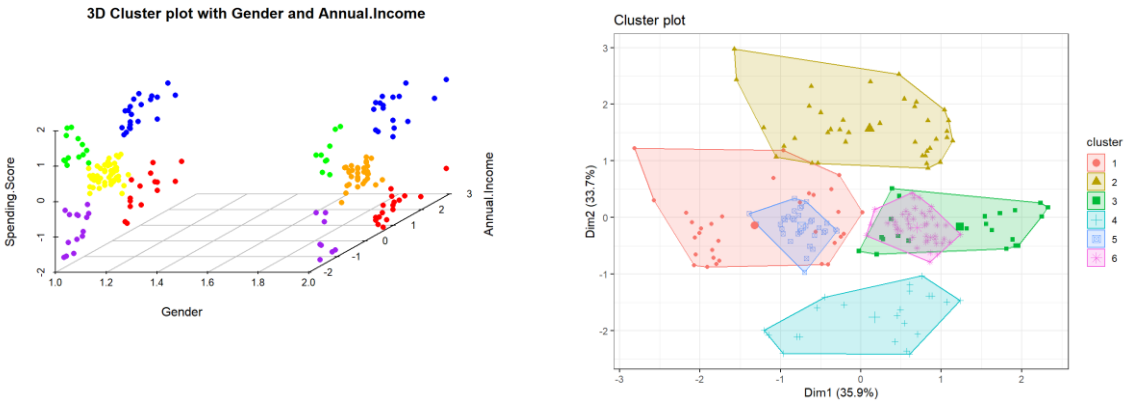

Figure 25. Clustering with Gender-Annual.Income-Spendig.Score

After doing this, we thought, what will happen if we do the clustering using all the available features? So we did it and we obtained the following plot. This plot is very chaotic and the clusters overlap. We think that they overlap because it is represented in two dimensions, so we are losing grades of freedom and the points do not necessarily overlap but in two dimensions they do. The same happened with the previous model. This model performed worse than the previous with a "Between SS / Total SS" value of 0.719.

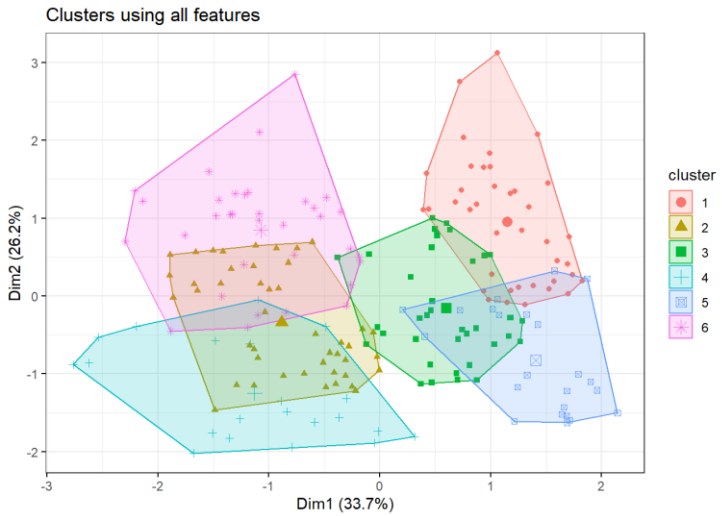

Figure 26. Clustering with all features

In the table below we summarize the results that we obtained for the different models.

Model	Experiment ID	#Clusters	BSS/TSS
K-means (Gender, Age,Annual.Income, Spending.score)	E1	3	0.756
	E2	5	0.536
	E3	6	0.777
K-means (Gender,Annual.Income,Spending.score)	E4	6	0.790
K-means (Age,Gender,Annual.Income,Spending.score)	E5	6	0.719

According to the table above, Experiments 1 to Experiments 3 help us find the optimal clusters on features excluding CustomerID. We do further model fitting(E4, E5) on the optimal clusters (k=6). We find model fitting on E4 that features are Gender, Annual.Income, Spending.score performs better than model fitting on E5 which features are Age, Gender, Annual.Income,

Spending.score. Hence, we reasoned the better and useful combination of features could be Gender, Annual.Income, Spending,score. Consider the mall interested in clustering customers based on these features that we applied in E4. The goal is to identify subgroups of similar customers. So that customers with each subgroup we trained can be shown advertisements or items that are particularly likely to interest them.

7. Discussion of Results:

In this analysis we explored the Mall customer data to try and observe important factors for the business and segment the customer base into insightful information that can be implemented for better customer relationship management.

- Customer segmentation is a good way to understand the behavior of different customers and plan a good marketing strategy accordingly.
- There isn't much difference between the spending score of women and men, which leads us to think that our behavior when it comes to shopping is pretty similar.
- Observing the clustering graphic, it can be clearly observed that the ones who spend more money in malls are young people. That is to say they are the main target when it comes to marketing, so doing deeper studies about what they are interested in may lead to higher profits.
- Although the young age group from 25 to 35 seem to be the ones spending the most, we can't forget there are more people we have to consider, like people who belong to the pink cluster, they are what we would commonly name after "middle class" and it seems to be the biggest cluster.
- Promoting discounts on some shops can be something of interest to those who don't actually spend a lot and they may end up spending more!
- One of the most interesting facts that we have found doing the project is that the models perform very different when adding or dropping features. The model with gender, annual.income and spending.score is better than the model with all the features.

7.1 Segmentation

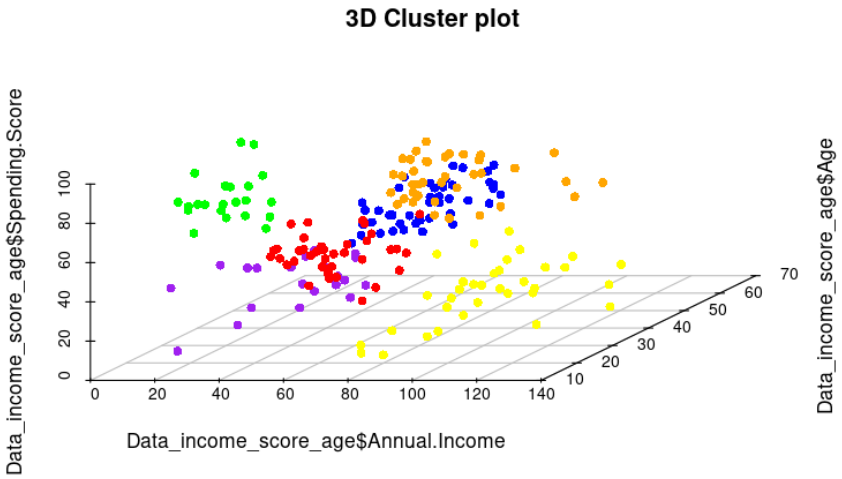

Figure 27. Clustering with unscaled data

From the figure 27 It can be established that the following clusters have different characteristics hence these become the segments a business can focus on given their characteristics.

1. Cluster 1 - Yellow- groups young people with moderate to low annual income who actually spend a lot.
2. Cluster 2 - The purple cluster groups reasonably young people with pretty decent salaries who spend a lot.
3. Cluster 3 - The pink cluster basically groups people of all ages whose salary isn't pretty high and their spending score is moderate.
4. Cluster 4 - The orange cluster groups people who actually have pretty good salaries and barely spend money, their age usually lies between thirty and sixty years.
5. Cluster 5 - The blue cluster groups whose salary is pretty low and don't spend much money in stores, they are people of all ages.

8. Conclusion and Recommendations

- In this analysis we explored the mall customers data to try and observe important factors for the business and segment the customer base into insightful information that can be implemented for better customer relationship management.

- Based on the findings from the demographic clusters established through the age, income levels, spending scores and gender of our customers the business can apply different marketing initiatives which are cluster centered.
- We have established six clusters with different characteristics based on age, income levels, spending habits and gender.
- The business can proceed to do inference and prediction on each cluster. For instance, the business can proceed to do linear regression on cluster one to predict sales likely to be made in the next financial year upon having applied their marketing strategy in the previous year.
- The K-Means model adopted in this project has no stable solution hence there is no global solution, hence an opportunity to conduct more clustering techniques like Gaussian Modelling.
- The concept of adopting different clustering techniques will help in optimizing the clusters.
- The next step for our project would be to gather more data about the products that are sell to each different customer segments and analyze how can we maximize the companies revenues.

9. Challenges

I. Overall Customer segmentation is the process of separation of customers into groups based on common characteristics or patterns so companies can market their products to each group effectively and significantly. We focused on segmenting our customer into a corresponding group so that K-means, an unsupervised learning, is very compatible here. Although this method works very well for this kind of problems it is only a mathematical approach, the challenge of interpreting those clusters and provide useful advice with this data is extremely important. We are also aware that explaining this concept and this results to people who are not familiar with the data analysis engineering is very difficult, and if we had to explain our research to economic background professionals, we should consider using a simplified approach.

II. Unsupervised learning has a lot of challenges of being difficult to test results and there is no reference for the training models. As a measure for a correct interpretation of the data does not exist, we can only expect the interpretation to be useful, which is subjective in nature.

III. The low dimensional data is easy to manage and that allow us to do a better interpretation of the data and the clusters, it allows us to represent the data in 3D plots

which makes the interpretation of the results easier because human understanding is improved if the results are visualized.

IV. We found our quantitative data like age, annual income and spending score are in normal distribution, there is no correlation among each other variable.

V. We found that the K-means algorithm is a lazy algorithm which means if we choose k=1, it will fit all datasets well, however it will lead to overfit and derived results have no practical means. To determine the optimal number of clusters, silhouette method can compute silhouette coefficients of each point that measures how much a point is like its own cluster compared to other clusters.

VI. We had a lot of discussion in whether we should perform feature scaling in our project, after research and discussion, we found that K-Means typically needs to have some form of normalization done on the datasets to work properly since it is sensitive to both the mean and variance of the datasets.

VII. Diversity in the project group was a mixture of knowledge sharing, coordination and enriched our understanding of the problem, however this diversity was challenging to coordinate the project and bring it up to the completion stage. We have to adapt to each other's workflow patterns, tools and interpretations. Also, we had to deal with different time zones to schedule the meetings and coordinate the development of the project.

VIII. We have to combine and make use of many different collaborative tools such as github, R studio, google docs, Microsoft project and communication tools such as whatsapp, google meet and gmail. We would like to remark the added difficulty of adapting to these platforms that not everybody is familiar with.

IX. At the end of the project we realized that the best model was the one using the gender and not the age feature. We did not have enough time to remake the segmentation for that model so we decided to leave as we had previously but we wanted to note that we are aware that we should have performed the customer segmentation for the best model.

10. References:

1. Scaling the data (Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: Principles and techniques for data scientists (First edition). O'Reilly.)

2. https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0
3. Data Parsing (https://r4ds.hadley.nz/parsing.html)
4. Normalizing the data (Feature Engineering for Machine Learning, Alice Zheng and Amanda Casari, 2018" P30-P35)
5. Density Plot (https://www.statology.org/density-curves/)
6. Density Plot(https://www.r-bloggers.com/2021/11/how-to-perform-univariate-analysis-in-r/)
7. QQ Plot (https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0)
8. Model fitting (https://uc-r.github.io/kmeans_clustering)
9. https://towardsdatascience.com/k-means-clustering-in-r-feb4a4740aa
10. Model Accuracy (https://towardsdatascience.com/k-means-clustering-in-r-feb4a4740aa)
11. Increasing dimensions (http://www.sthda.com/english/wiki/scatterplot3d-3d-graphics-r-software-and-data-visualization)
12. Increasing dimensions (https://www.datanovia.com/en/blog/k-means-clustering-visualization-in-r-step-by-step-guide/)
13. Scatterplot (http://www.sthda.com/english/wiki/scatterplot3d-3d-graphics-r-software-and-data-visualization)
14. Q-Q plot
15. Ford, C. (2015, Agust). *Understanding q-q plots* . Research Data Services + Sciences; University of Virginia Library.
16. https://data.library.virginia.edu/understanding-q-q-plots.
17. Explore missing value patterns
18. Kang, H. (2013). The prevention and handling of the missing data. Korean Journal of Anesthesiology, 64(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402
19. Research Methodology Flowchart (https://www.researchgate.net/figure/Diagram-of-machine-learning-and-experimental-design-framework_fig5_339252056)