ANA 515 Assignment 4 - Customer Segmentation using Machine Learning

Amanda Thiell

5/8/2022

1. Discuss the business problem/goal

## Customer Segmentation is one the most important applications of unsupervised learning and is an important practise of dividing customers base into individual groups that are similar. It is useful in customized marketing. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this exploration we will use mall customer information to segment the customers based on the age, gender, and interest to improve their individualized marketing and future shopping experiences.

1. identify where the dataset was retrieved from

## The dataset used for this project was sourced from the website kaggle. Kaggle is a platform utilized by data scientists and machine learning practitioners which allows users to find/publish data sets for public review. The data used for this project can be found at the following link: <https://www.kaggle.com/datasets/shwetabh123/mall-customers>

1. Identify the code that imported and saved your dataset in R

url <- "C:\\Users\\Amanda\\Downloads\\Mall\_Customers.csv"  
customer\_data <- read.csv(url)

1. Describe your data set

## Our data contains 200 rows of customer data and including 5 columns of variables. The column names are as follows: CustomerID, Genre, Age, Annual.Income..k.., Spending.Score..1.100.. In addition, below is a breakdown of the percentile breakdown of this dataset.

## CustomerID Genre Age Annual.Income..k..  
## Min. : 1.00 Length:200 Min. :18.00 Min. : 15.00   
## 1st Qu.: 50.75 Class :character 1st Qu.:28.75 1st Qu.: 41.50   
## Median :100.50 Mode :character Median :36.00 Median : 61.50   
## Mean :100.50 Mean :38.85 Mean : 60.56   
## 3rd Qu.:150.25 3rd Qu.:49.00 3rd Qu.: 78.00   
## Max. :200.00 Max. :70.00 Max. :137.00   
## Spending.Score..1.100.  
## Min. : 1.00   
## 1st Qu.:34.75   
## Median :50.00   
## Mean :50.20   
## 3rd Qu.:73.00   
## Max. :99.00

## The standard deviation (SD) of our variables Age, Income, and Spending are as follows:

## SD(age)= 13.9690073

## SD(income)= 26.2647212

## SD(spending)= 25.8235217

1. Discuss any data preparation, missing values and errors

## Prior to actually completing tests on a dataset, it is important to clean the dtata and reorganize as necessary. Due to the level of needed accuracy, it is crucial that any outliers/anomalies, mising values, or inaccurate data are removed/resolved. This ensures there is no skew to the results because of the incorrect data.

## In this specific sample, there are no missing values to be removed. To ensure there are no anomalies, the anomalize function can be used with additional review of the standard deviation of files and their mean value. After searching for incorrect dtaa, we want to see if any reformatting is necessary. One frequently seen example of this is date format - is the associated date in the same format as the others? Ex: MM/DD/YYYY vs MM/YY. Next, data preparation includes combining data. From my prior experience, this is most helpful when one data set does not contain all necessary information for review. Ex - adding a column, creating a subset, combining full datasets. Renaming of the dataset’s columns was not necessary in this case since they are pretty self explanatory, but renaming could be completed for ease of reference. For example, the column indicating “gender” is named “genre” ; this could be updated but naming is not an issue as we will correct during the visualization phase. These are all great ways to ensure the data is prepped properly prior to analysis.

1. Discuss the modeling

## In this machine learning project, we will make use of K-means clustering which is the essential algorithm for clustering unlabeled dataset. K-means clustering is an iterative algorithm that partitions a group of data containing n values into k subgroups. We then find patterns within this data which are present as k-clusters. Each of the n value belongs to the k cluster with the nearest mean.

1. produce and discuss the output (see below)
2. provide explanation with any visuals (see below)

## Data exploration

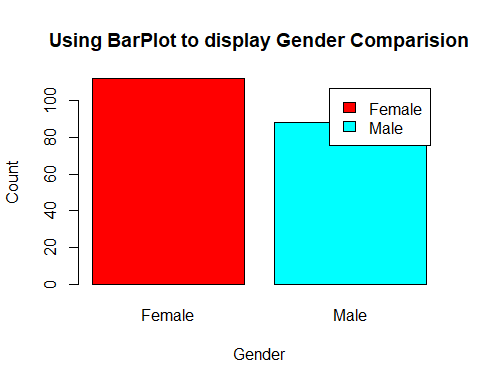
## 'data.frame': 200 obs. of 5 variables:  
## $ CustomerID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Genre : chr "Male" "Male" "Female" "Female" ...  
## $ Age : int 19 21 20 23 31 22 35 23 64 30 ...  
## $ Annual.Income..k.. : int 15 15 16 16 17 17 18 18 19 19 ...  
## $ Spending.Score..1.100.: int 39 81 6 77 40 76 6 94 3 72 ...

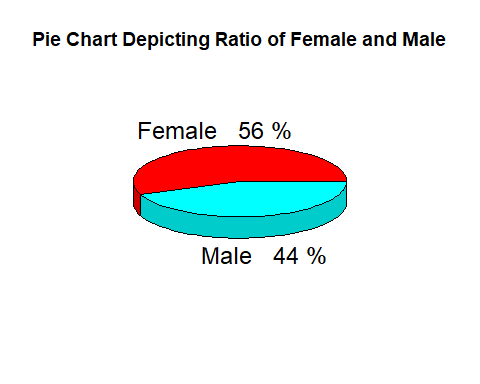
## [1] "CustomerID" "Genre" "Age"   
## [4] "Annual.Income..k.." "Spending.Score..1.100."

## CustomerID Genre Age Annual.Income..k.. Spending.Score..1.100.  
## 1 1 Male 19 15 39  
## 2 2 Male 21 15 81  
## 3 3 Female 20 16 6  
## 4 4 Female 23 16 77  
## 5 5 Female 31 17 40  
## 6 6 Female 22 17 76

## Customer Gender Visualization - bar graph and pie chart

## The following charts show the gender distribution across our customer\_data dataset



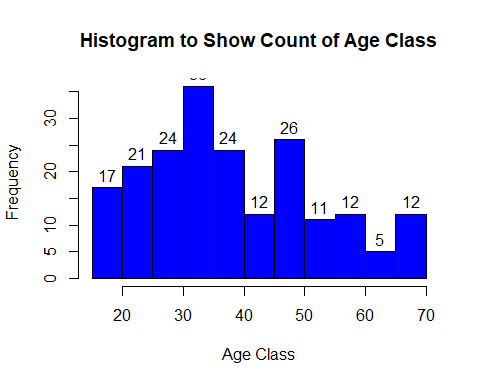


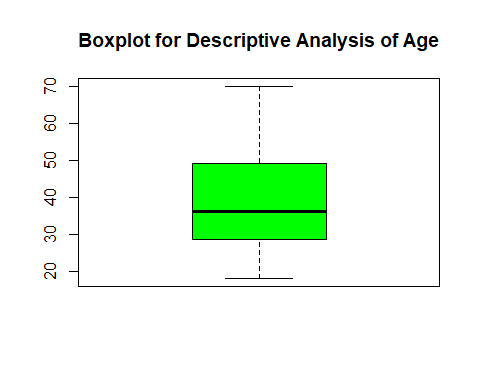
## From the above graph, we conclude that the percentage of females is 56%, whereas the percentage of male in the customer dataset is 44%.

## Visualization of Age Distribution

## Let us plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking summary of the Age variable.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 28.75 36.00 38.85 49.00 70.00



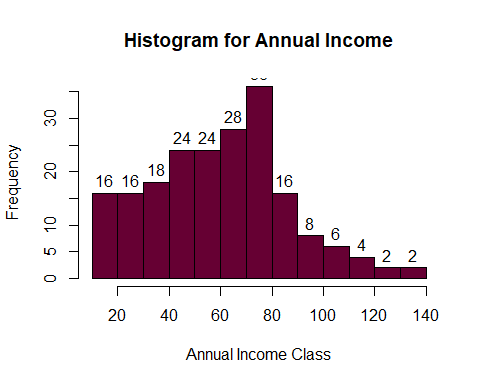


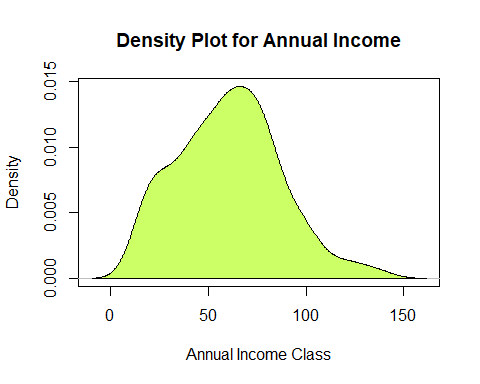
## From the above two visualizations, we conclude that the majority of customer ages fall between 30 and 35. The minimum age of customers is 18, whereas, the maximum age is 70. The box plot also shows the inner quartiles range between 30 to 50, with the least amount of customer ages falling in the range of 60 to 65.

## Analysis of annual income of customers

## In this section of the R project, we will create visualizations to analyze the annual income of the customers. We will plot a histogram and then we will proceed to examine this data using a density plot.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15.00 41.50 61.50 60.56 78.00 137.00



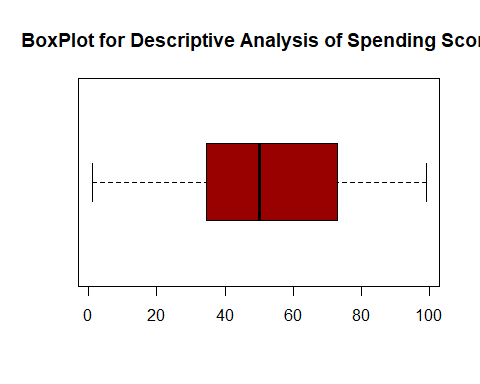


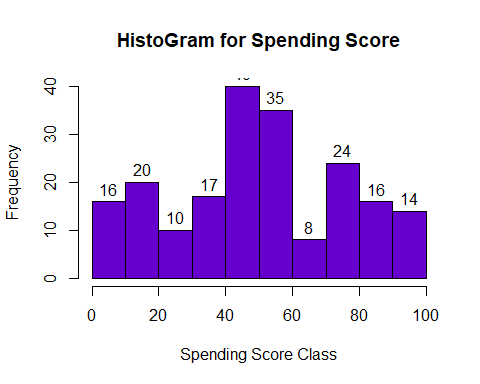
## From the above descriptive analysis, we conclude that the minimum annual income of the customers is 15 (in thousands) and the maximum income is 137 (in thousands). People earning an average income of 70k have the highest frequency count in our histogram distribution. The average salary of all the customers is 60.56k. In the Kernel Density Plot that we displayed above, we observe that the annual income has a normal distribution.

## Analyzing Spending Score of the Customers

## In this section of the project, we want to review how the customer spending score is affected by the other variables, and oranize the data accordingly for presentation. To do this, we can use a combination of a box plot and histogram for analysis of distribution and frequency.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 34.75 50.00 50.20 73.00 99.00





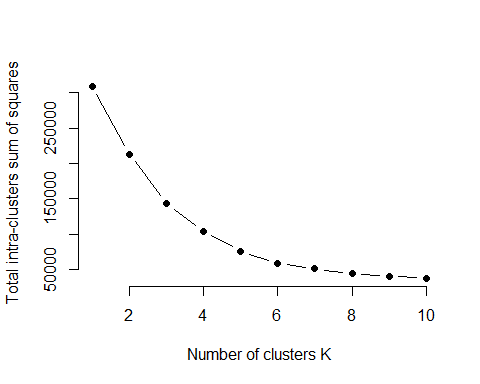
## The minimum spending score is 1, maximum is 99 and the average is 50.20. We can see Descriptive Analysis of Spending Score is that Min is 1, Max is 99 and avg. is 50.20. From the histogram, we conclude that customers between class 40 and 50 have the highest spending score among all the classes.

## K-means Algorithm

## While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) we are trying to produce. First, k objects from dataset are selected randomly that will serve as the initial centers for our clusters. These are the cluster means (aka centroids). Then, the remaining objects have an assignment of the closest centroid. This centroid is defined by the Euclidean Distance present between the object and the cluster mean (referred to as “cluster assignment”). When the assignment is complete, the algorithm proceeds to calculate new mean value of each cluster present in the data. After the recalculation of the centers, the observations are checked if they are closer to a different cluster. Using the updated cluster mean, then are reassigned as needed. This continues to occur until no updates are necessary through this process.

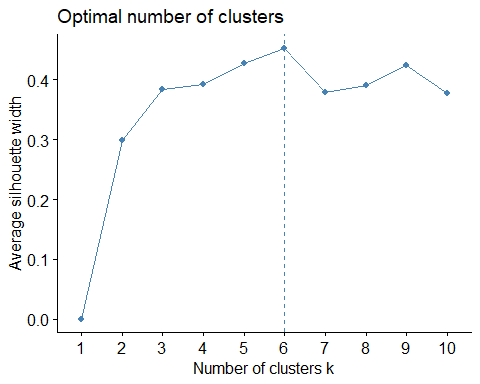
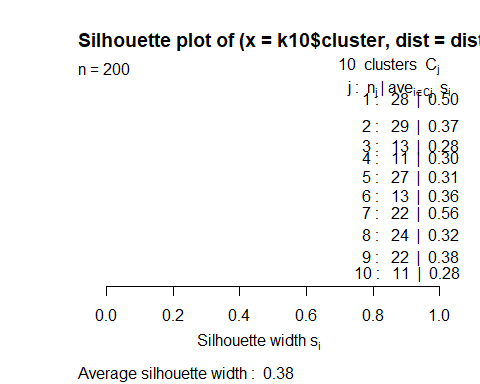
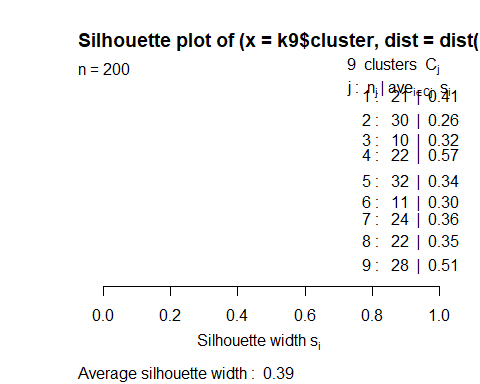
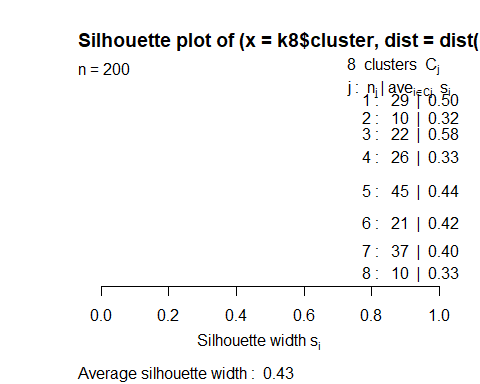
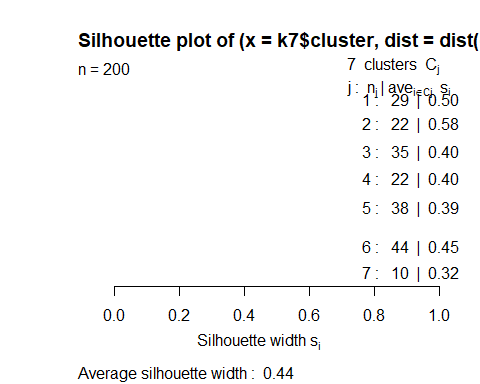
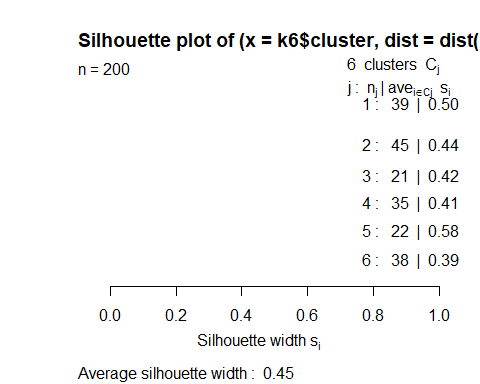
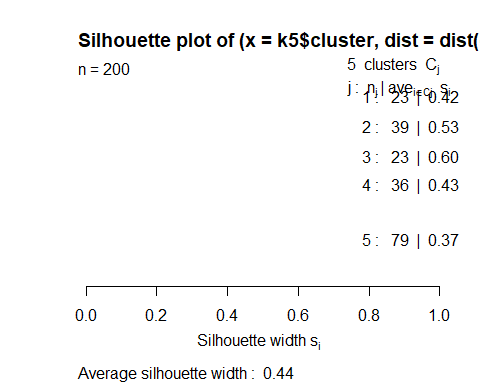
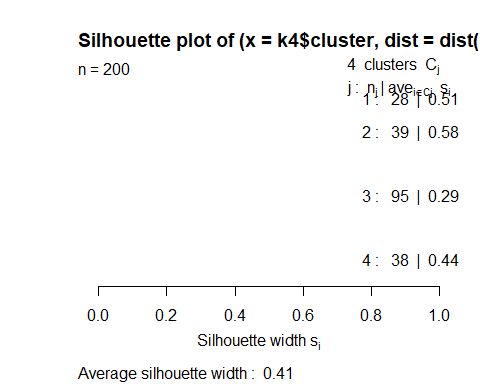
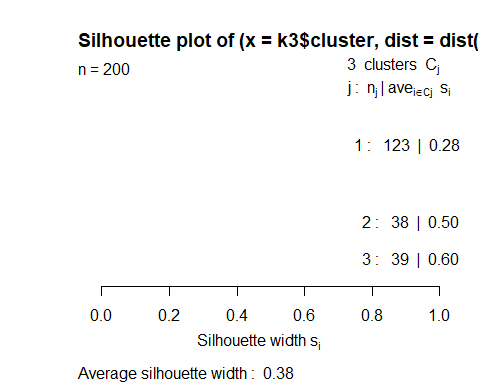
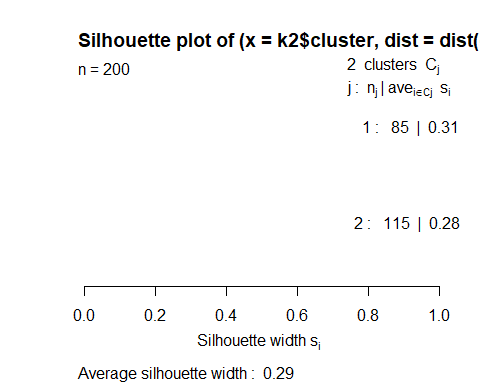
## To determining the optimal clusters, there are three popular methods – Elbow method, Silhouette method, and Gap statistic

## Elbow method



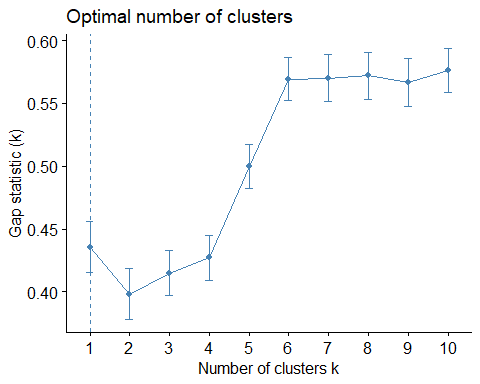
## Average Silhouette Method

## Using the average silhouette method, we can measure the quality of our clustering operation. We can then determine how well the data object fits within the cluster is the data object. If we obtain a high average silhouette width, it means that we have good clustering. The average silhouette method calculates the mean of silhouette observations for different k values. With the optimal number of k clusters, one can maximize the average silhouette over significant values for k clusters.



## Gap Statistic Method

## Using the gap statistic, one can compare the total intracluster variation for different values of k along with their expected values under the null reference distribution of data. For each variable, we can calculate the range between min(xi) and max (xj) through which we can produce values uniformly from interval lower bound to upper bound.



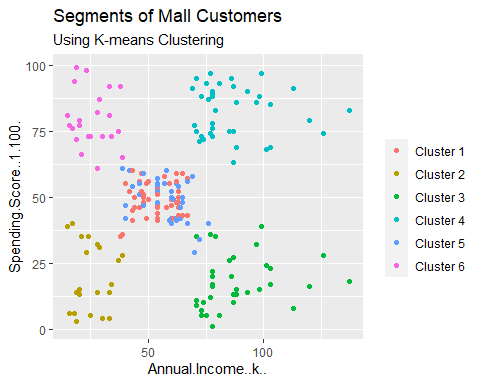
## K-means clustering with 6 clusters of sizes 45, 21, 35, 39, 38, 22  
##   
## Cluster means:  
## Age Annual.Income..k.. Spending.Score..1.100.  
## 1 56.15556 53.37778 49.08889  
## 2 44.14286 25.14286 19.52381  
## 3 41.68571 88.22857 17.28571  
## 4 32.69231 86.53846 82.12821  
## 5 27.00000 56.65789 49.13158  
## 6 25.27273 25.72727 79.36364  
##   
## Clustering vector:  
## [1] 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2  
## [38] 6 2 6 1 6 1 5 2 6 1 5 5 5 1 5 5 1 1 1 1 1 5 1 1 5 1 1 1 5 1 1 5 5 1 1 1 1  
## [75] 1 5 1 5 5 1 1 5 1 1 5 1 1 5 5 1 1 5 1 5 5 5 1 5 1 5 5 1 1 5 1 5 1 1 1 1 1  
## [112] 5 5 5 5 5 1 1 1 1 5 5 5 4 5 4 3 4 3 4 3 4 5 4 3 4 3 4 3 4 3 4 5 4 3 4 3 4  
## [149] 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3  
## [186] 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4  
##   
## Within cluster sum of squares by cluster:  
## [1] 8062.133 7732.381 16690.857 13972.359 7742.895 4099.818  
## (between\_SS / total\_SS = 81.1 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

## Visualizing the Clustering Results using the First Two Principle Components

## Importance of components:  
## PC1 PC2 PC3  
## Standard deviation 26.4625 26.1597 12.9317  
## Proportion of Variance 0.4512 0.4410 0.1078  
## Cumulative Proportion 0.4512 0.8922 1.0000

## PC1 PC2  
## Age 0.1889742 -0.1309652  
## Annual.Income..k.. -0.5886410 -0.8083757  
## Spending.Score..1.100. -0.7859965 0.5739136

set.seed(1)  
ggplot(customer\_data, aes(x =Annual.Income..k.., y = Spending.Score..1.100.)) +   
 geom\_point(stat = "identity", aes(color = as.factor(k6$cluster))) +  
 scale\_color\_discrete(name=" ",  
 breaks=c("1", "2", "3", "4", "5","6"),  
 labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5","Cluster 6")) +  
 ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")



## From the above visualization, we observe that there is a distribution of 6 clusters as follows –

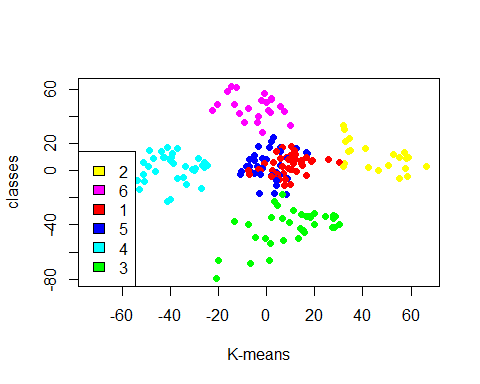
## Cluster 6 and 4 – These clusters represent the customer\_data with the medium income salary as well as the medium annual spend of salary.

## Cluster 1 – This cluster represents the customer\_data having a high annual income as well as a high annual spend.

## Cluster 3 – This cluster denotes the customer\_data with low annual income as well as low yearly spend of income.

## Cluster 2 – This cluster denotes a high annual income and low yearly spend.

## Cluster 5 – This cluster represents a low annual income but its high yearly expenditure.



## Cluster 4 and 1 – These two clusters consist of customers with medium PCA1 and medium PCA2 score.

## Cluster 6 – This cluster represents customers having a high PCA2 and a low PCA1.

## Cluster 5 – In this cluster, there are customers with a medium PCA1 and a low PCA2 score.

## Cluster 3 – This cluster comprises of customers with a high PCA1 income and a high PCA2.

## Cluster 2 – This comprises of customers with a high PCA2 and a medium annual spend of income.

## In conclusion, the colustering techniques reviewed above assist with organizing variables without prior classification. In this case, organizing the characteristics of variables for mall shoppers allows shows to market to these types of customers to improve and expand their sales.