UNIVERSITÀ
DEGLI STUDI
DI MILANO

# Statistical Learning, Deep Learning and Artificial Intelligence

**Instructor:** Respected Prof. Silvia Salini

**Student:** Vickysingh Chandansingh Baghel (964952) - Data Science and Economics

**Project:** Customer Segmentation using Unsupervised Statistical methods .

## Table of Contents:

## Abstract:

In this project, we will perform one of the most essential applications of machine learning, Customer Segmentation by using K-Means Clustering Algorithm. In this project, we will implement customer segmentation in R Studio. Whenever you need to find your best customer, customer segmentation is the ideal methodology. Then we will explore the data upon which we will be building our segmentation model. Also, in this Statistics Learning project, we will see the descriptive analysis of our data and then implement several versions of the K-means algorithm. Furthermore, through the data collected, we can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, we can strategize the marketing techniques more efficiently and minimize the possibility of risk to the investment.

Whenever you need to find your best customer, customer segmentation is the ideal methodology.

Key Factors :- Kmean Clustering Algorithm

# 1. Introduction:

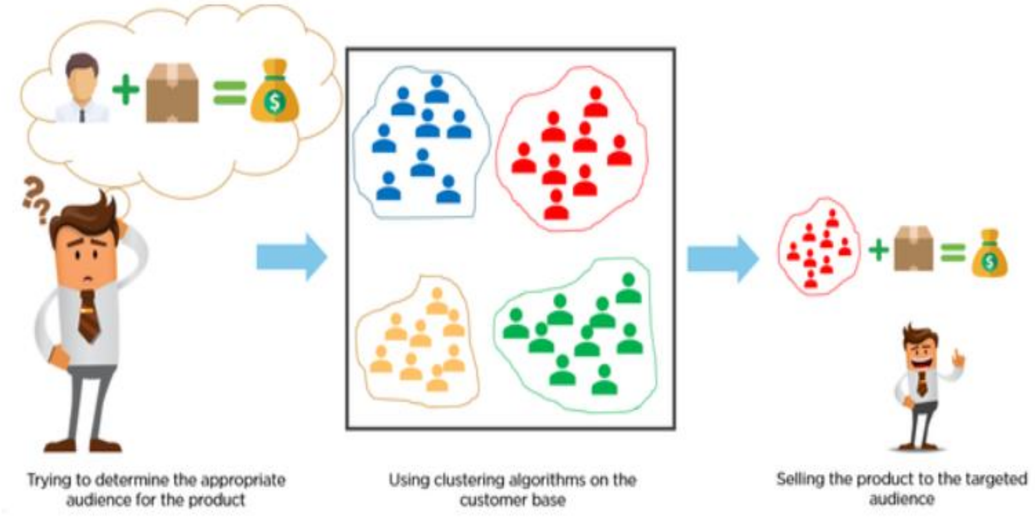
The purpose of this analysis is to uncover underlying patterns in the customer base, and to groups of customers accordingly, often known as market segmentation. In doing so, the marketing team can have a more targeted approach to reach consumers, and the mall can make more informed strategic decisions to increase profits.

Segmentation of customers is one of the business areas where clustering methods can be very useful. Thanks to this unsupervised learning technique, a company can adjust marketing strategy in more efficient manner and focus on those customers who will ensure the highest revenue.

Clustering technique is critically important step in data mining process. It is a multivariate procedure quite suitable for segmentation. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this Statistic learning project, we will make use of **K-means clustering** which is the essential algorithm for clustering unlabeled dataset. Before ahead in this project, learn what actually customer segmentation is.

## What is Customer Segmentation

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.


Trying to determine the appropriate audience for the product
Using clustering algorithms on the customer base
Selling the product to the targeted audience

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit.

This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments. In the first step of this data science project, we will perform data exploration. We will import the essential packages required for this role and then read our data. Finally, we will go through the input data to gain necessary insights about it.

## 2. Methodology

All Analysis was performed in R. Small tables (bucketed customer age, spending tiers, and cluster centers) were formatted with Excel.

- Exploratory Data Analysis and Descriptive Statistics
- Data Visualization
- Unsupervised Machine Learning (K-Means Clustering)
- The Elbow Method
- Silhouette Method
- Gap Statistic Method

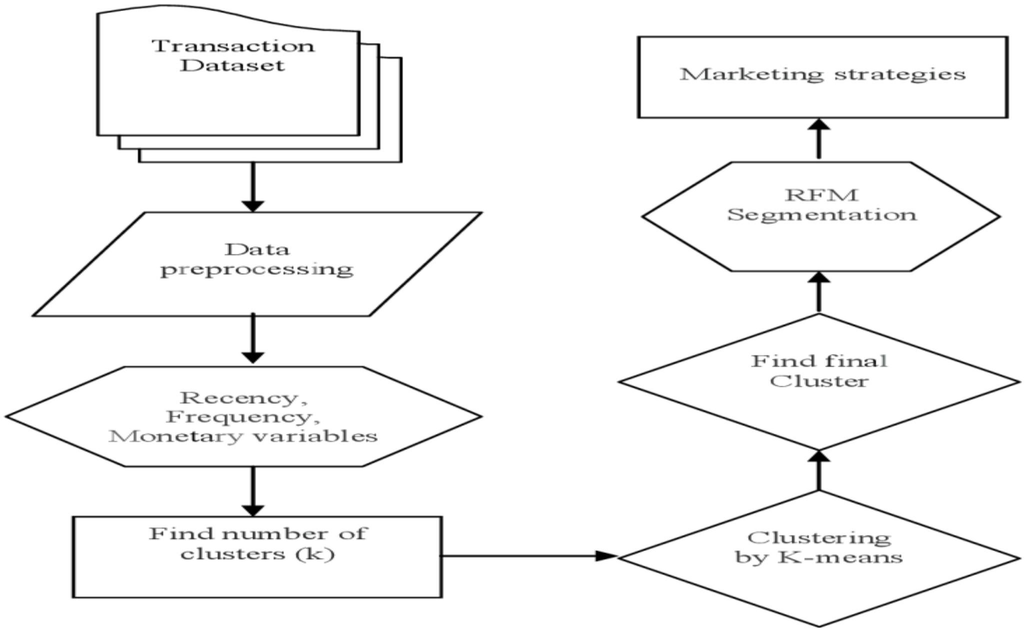


Fig 1. Framework for Customer Segmentation based on RFM model

### Reason to use Unsupervised Learning Algorithms

- Unlike Supervised Learning, Unsupervised Learning has only independent variables and no corresponding target variable. The data is unlabeled. The aim of unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

- We are going to examine a dataset that is about Mall Customers for segmentation. There is no any feature about label of customers. That is to say, we don’t have information about customer’s characteristics. We are going to try clustering clients through identifying similarities with machine learning algorithms. Segmentation of customers has a pretty significant position for companies in new marketing disciplines. Firms must reach to the right target audiences with right approaches because of costs.

## 3. Data Collection & Dataset Description:

The dataset is collected through Kaggle  The dataset is a customer database of a mall. It contains 200 observations & 5 varaibles with basic information such as

1. Customer ID

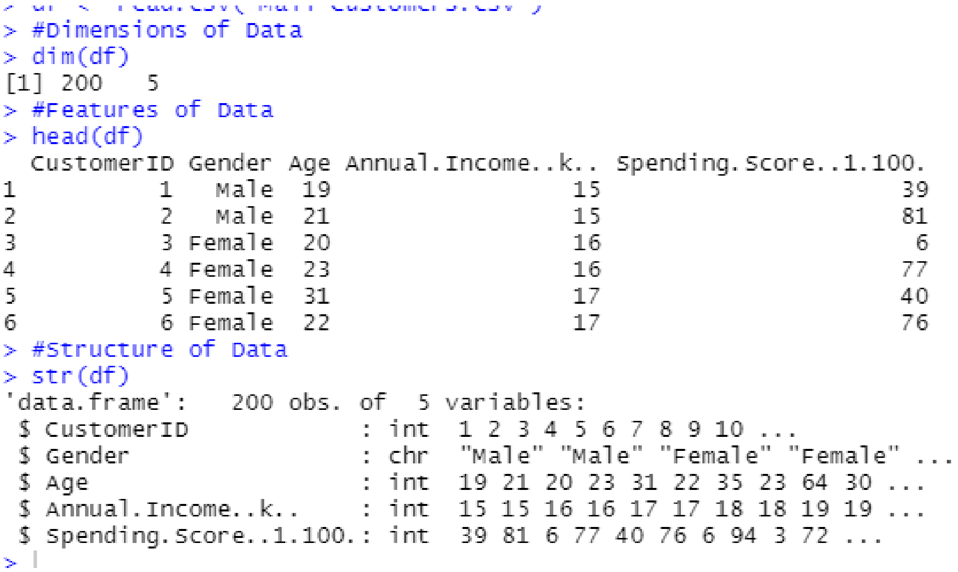
2. Customer Gender

3. Customer Age

4. Annual Income of the customer (in Thousand Dollars)

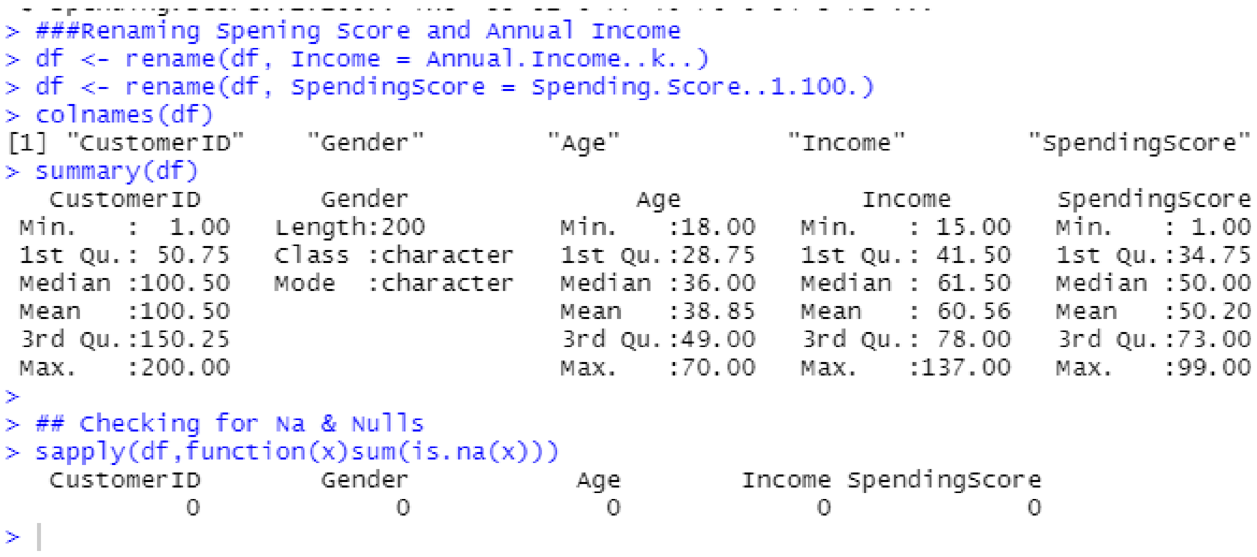
5. Spending score of the customer (based on customer behaviour and spending nature)

### 3.1 Dataset Insights:



```
> df <- read.csv("Mall Customers.csv")
> #Dimensions of Data
> dim(df)
[1] 200   5
> #Features of Data
> head(df)
  CustomerID Gender 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
> #Structure of Data
> str(df)
'data.frame':	200 obs. of  5 variables:
 $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : 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 ...
> 
```

As we see above we have 5 variables and 200obs. 4 variables are integer type & 1 is categorical. change names of last two variables, in order to have easier access in the next analysis.

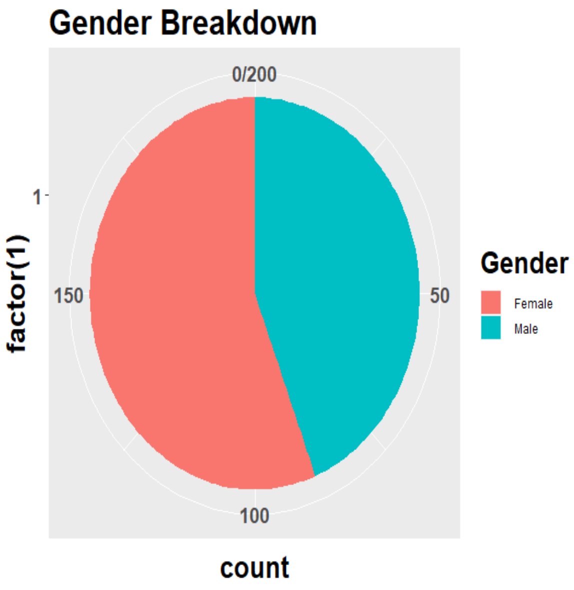


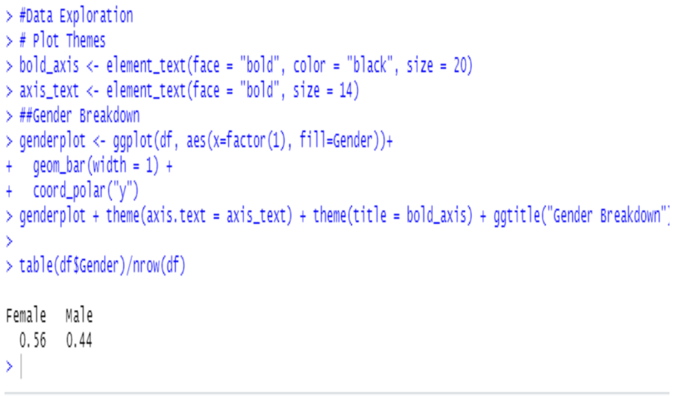
```
> ###Renaming Spening Score and Annual Income
> df <- rename(df, Income = Annual.Income..k..)
> df <- rename(df, SpendingScore = Spending.Score..1.100.)
> colnames(df)
[1] "CustomerID"    "Gender"        "Age"           "Income"        "SpendingScore"
> summary(df)
   CustomerID        Gender               Age            Income        SpendingScore
 Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00   Min.   : 1.00
 1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50   1st Qu.:34.75
 Median :100.50   Mode  :character   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
>
> ## Checking for Na & Nulls
> sapply(df,function(x)sum(is.na(x)))
   CustomerID        Gender           Age        Income SpendingScore
            0             0             0             0             0
> |
```

We can see from summary chat, The age of the customers varies from 18 to 70 and the annual income from 15 to 137 (given in thousands). The variable Spending Score is a score assigned by the mall based on the customer behavior and spending nature, where 99 is a maximum and 1 is a minimum value. In order to better understand the data. There is no Null values in our dataset.

## 4. Exploratory Data Analysis & Feature Engineering:

First, I wanted to understand the distribution of the variables gender wise with help of pie chat.

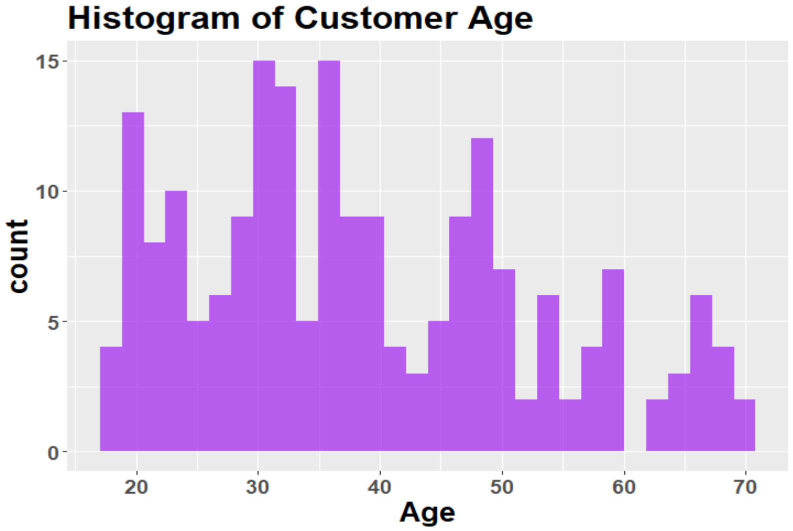




```
> #Data Exploration
> # Plot Themes
> bold_axis <- element_text(face = "bold", color = "black", size = 20)
> axis_text <- element_text(face = "bold", size = 14)
> ##Gender Breakdown
> genderplot <- ggplot(df, aes(x=factor(1), fill=Gender))+
+   geom_bar(width = 1) +
+   coord_polar("y")
> genderplot + theme(axis.text = axis_text) + theme(title = bold_axis) + ggtitle("Gender Breakdown")
>
> table(df$Gender)/nrow(df)

Female   Male
  0.56   0.44
> |
```

56% of customers are female, and 44% are male, a considerable difference.

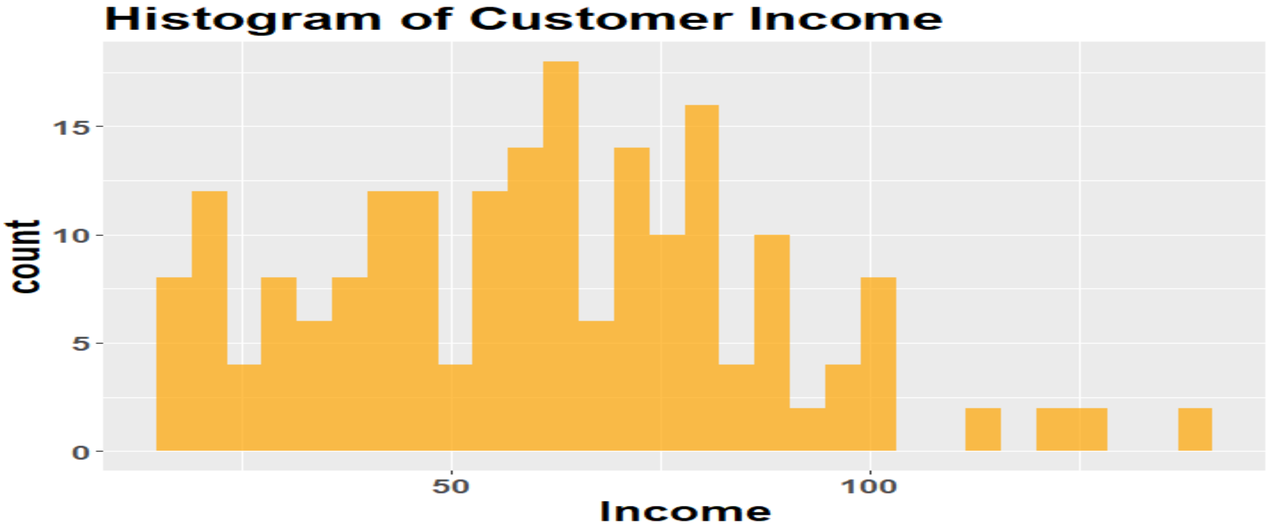
**Let us plot a histogram to view the distribution to plot the frequency of customer ages.**



| Age | Female | Male |
|---|---|---|
| < 30 | 29 | 26 |
| 30 – 39 | 37 | 24 |
| 40 - 49 | 24 | 15 |
| 50 - 70 | 22 | 23 |

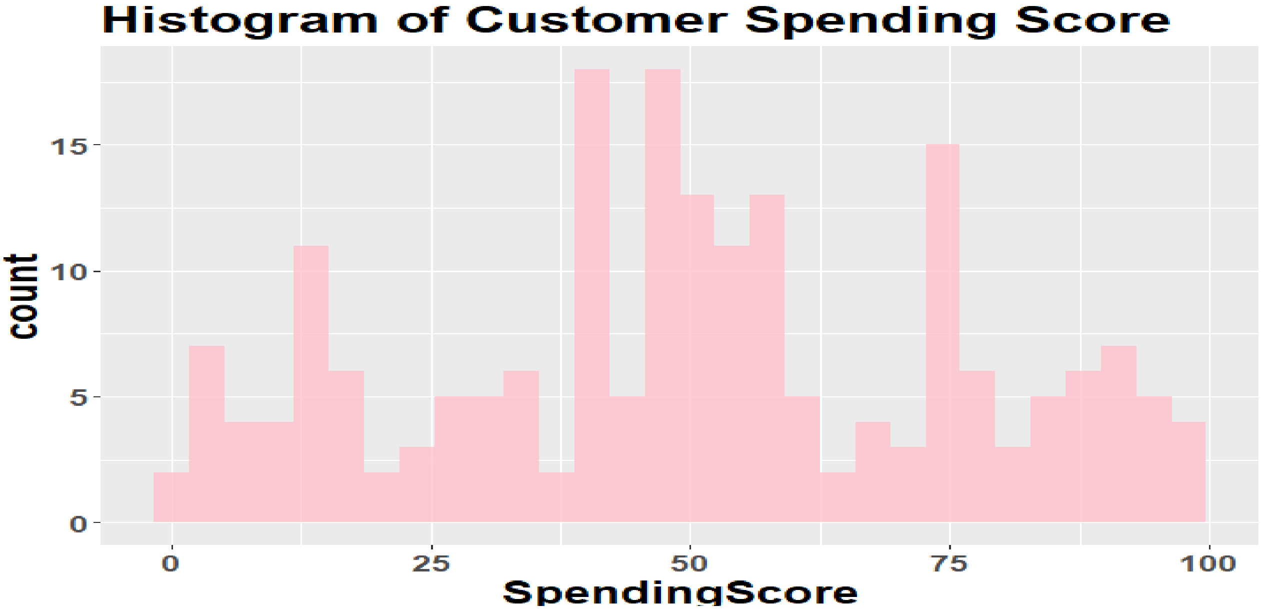
The most represented age group is between 30-39. Intuitively, this makes sense, as many of these individuals are likely parents who are buying for their young kids. The older group (55+) is not nearly as represented, and this can be due to the aging of their kids to an age when they can purchase for themselves (18+), as well as a change in lifestyle (retiring and moving). Otherwise, this could reflect the underlying demographic of the nearby towns.

**Let us plot a histogram to view the distribution to plot the frequency of customer income.**



The customers seem to be somewhat normally distributed with respect to income, with 60-90K representing the income of most customers. Very few individuals with incomes over 120K are purchasing at the mall, which may be indicative of the underlying population of the town, or that the wealthier individuals are consuming more from other outlets (other stores, or online). This breakdown is important from both a marketing and strategic standpoint. From a marketing point of view, are there behaviors generalizable across individuals of different income groups? An understanding of consumption of TV, online media, and other advertising mediums can be leveraged to reach individuals of specific income groups. We do know that individuals within the 60-90k income bracket are visiting this mall more often than other groups, so they would likely be receptive to in-person advertisements (storefront signs, in-store promotions, coupons, etc.) Additionally, without more knowledge of the cost of living in the area, it is somewhat safe to say much of the consumer base has considerable spending power.

**Let us plot a histogram to view the distribution to plot the frequency of customer income**



The distribution is somewhat normalized as well, with the spending scores between 40-60 appearing most frequently among the customers (and 63% of consumers within 1SD of the mean of 50.2). This distribution illustrates that there are a substantial amount of customers who are "somewhat frequent visitors", or "moderate purchasers" who likely have the means to become more loyal customers. Had the distribution been more concentrated on the ends of the spectrum (high concentrations of spending score < 30 and high concentration > 70), I believe it would be more concerning, as converting those low spending score customers into higher spending score customers would be difficult if the underlying driver to their spending behavior is their income. Given the current state, there should be two objectives. The first is to acquire new customers, and if the distribution of new customers is similar to that of the current one (assumes no improvement in targeted marketing), a substantial amount of them will have the spending power to be at least moderate purchasers. The second focus should be on shifting the current consumer base primarily from moderate spenders to loyal customers, and incorporating new incentives and/or a loyalty program should encourage this.

### Recap of EDA

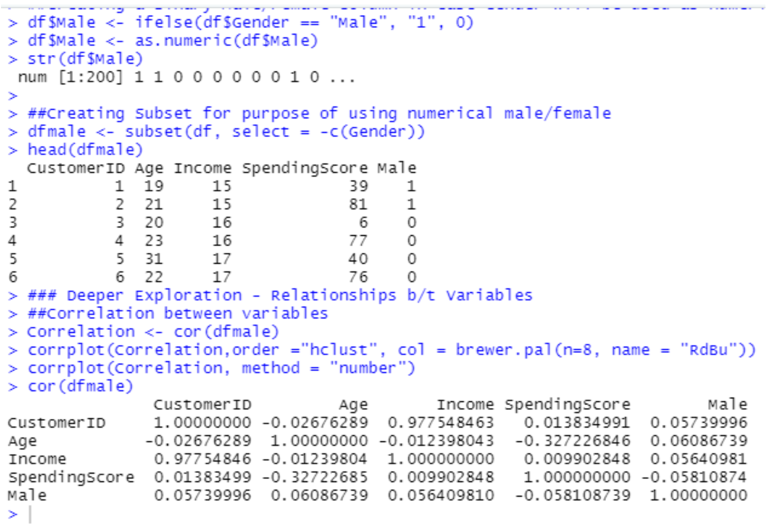
We have learned the following through the initial data exploration.

- Female skew (58% of all customers),
- Young-middle age demographic (77.5% of customers between age 20-49)
  - 7.5% between 20-29
  - 30.5% between 30-39
  - 9.5% between 40-49

- Most customers are middle to upper-middle class (60-90K income), mean = 60.56K
- Medium Spending score individuals make up the majority of the dataset
  - Mean of 50.2
  - SD = 25.8
  - 63% of customers within 1 SD (between 25-75)

## Further EDA

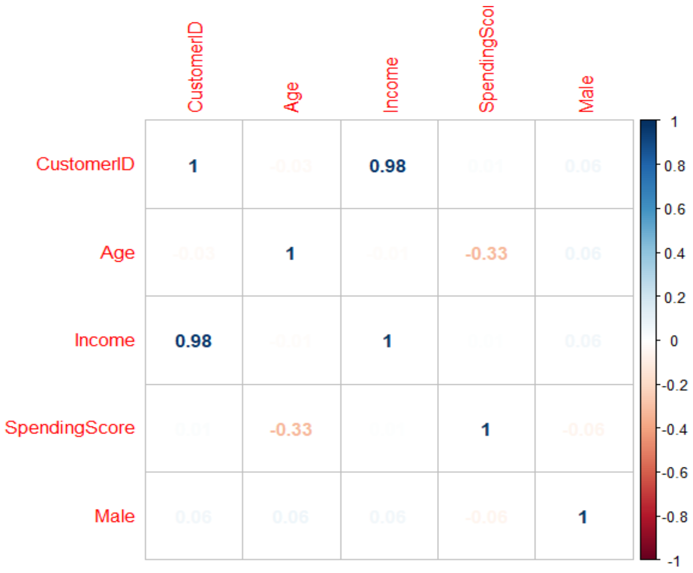
Next, I decided to examine the relationship between variables



```
> df$Male <- ifelse(df$Gender == "Male", "1", 0)
> df$Male <- as.numeric(df$Male)
> str(df$Male)
 num [1:200] 1 1 0 0 0 0 0 0 1 0 ...
>
> ##Creating Subset for purpose of using numerical male/female
> dfmale <- subset(df, select = -c(Gender))
> head(dfmale)
  CustomerID Age Income SpendingScore Male
1          1  19     15            39    1
2          2  21     15            81    1
3          3  20     16             6    0
4          4  23     16            77    0
5          5  31     17            40    0
6          6  22     17            76    0
> ### Deeper Exploration - Relationships b/t Variables
> ##Correlation between variables
> Correlation <- cor(dfmale)
> corrplot(Correlation,order ="hclust", col = brewer.pal(n=8, name = "RdBu"))
> corrplot(Correlation, method = "number")
> cor(dfmale)
               CustomerID         Age       Income SpendingScore        Male
CustomerID     1.00000000 -0.02676289  0.977548463   0.013834991  0.05739996
Age           -0.02676289  1.00000000 -0.012398043  -0.327226846  0.06086739
Income         0.97754846 -0.01239804  1.000000000   0.009902848  0.05640981
SpendingScore  0.01383499 -0.32722685  0.009902848   1.000000000 -0.05810874
Male           0.05739996  0.06086739  0.056409810  -0.058108739  1.00000000
> |
```

We convert categorical variables into binary 0,1 as you can see above . to give the mathematical weights then I did correlation.

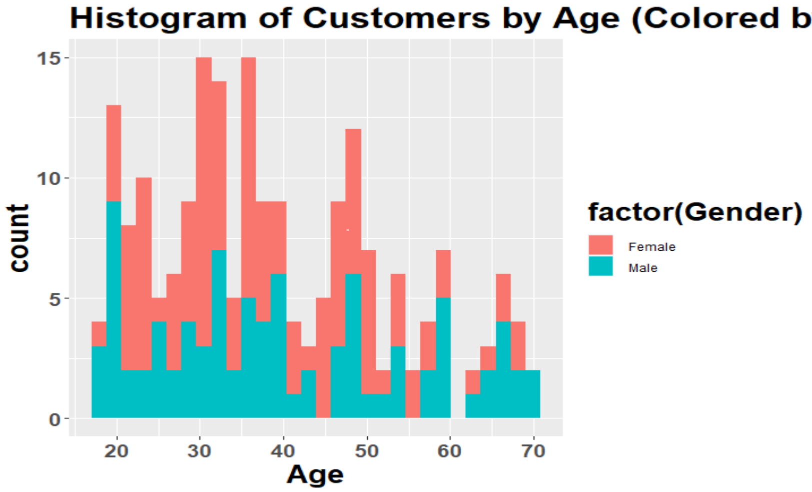
### Correlation Matrix



There is not much correlation between variables, though there is a slight inverse relationship between age and spending score (-.33), for reasons potentially highlighted before. Though strong correlation is not present in many of these variables, I still think it is important to explore their relationship, as trends may emerge within subsets of the data.

### Age and Gender

We know the population at large is female skewed, and that the most represented age group is 30-39, but what is the distribution of gender by age?

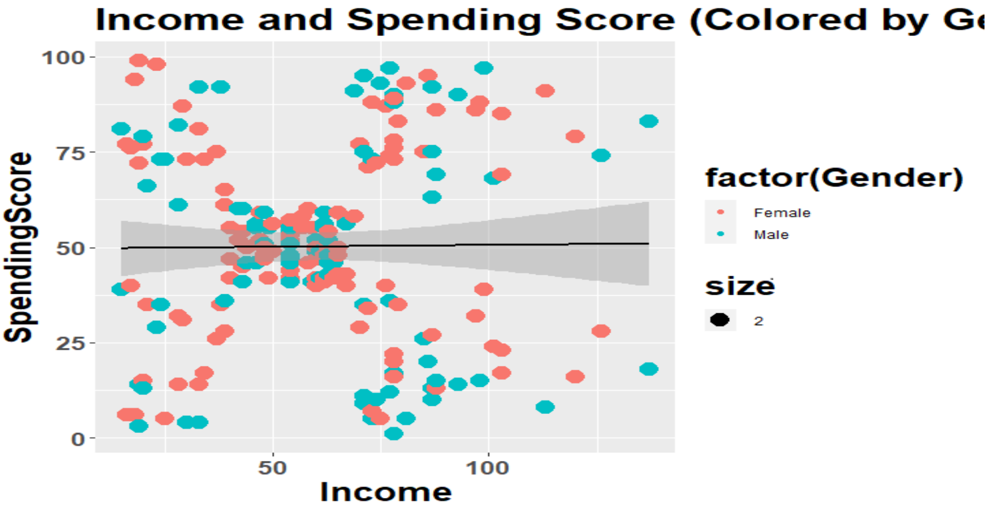


| Age | Female | Male |
|---|---|---|
| < 30 | 29 | 26 |
| 30 – 39 | 37 | 24 |
| 40 - 49 | 24 | 15 |
| 50 - 70 | 22 | 23 |
| Total | 112 | 88 |

We can see that the gender split is fairly equal among those in their twenties, or above 50. For people between 30-49, the distribution is skewed towards females (61% of customers aged 30-39 and 62% of customers aged 40-49 are female). This supports the idea that this group may comprise of mothers of young children.

### Income and Spending Score

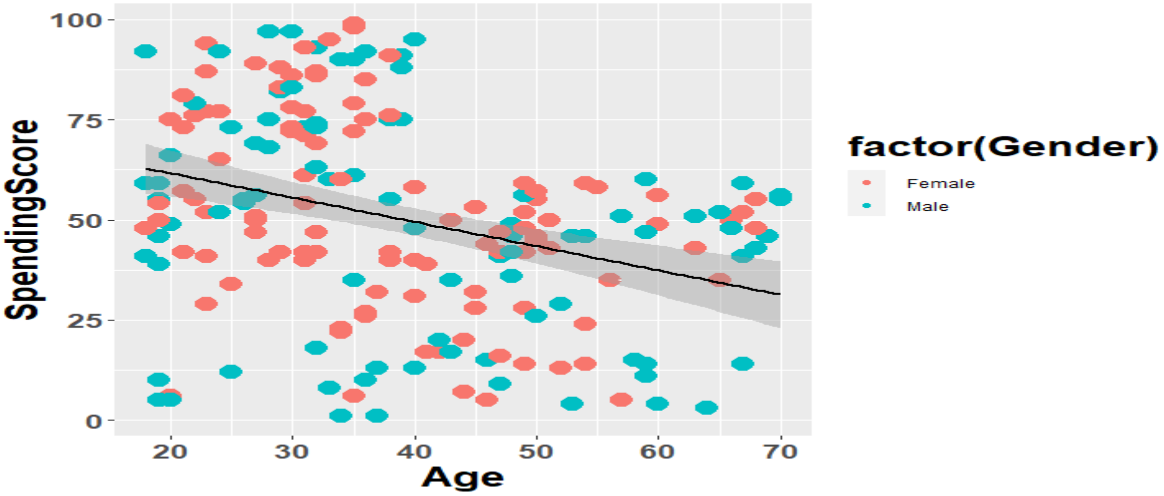
Is there a relationship between income and spending score (by gender)?



It is difficult to identify a clear relationship between income and spending score, however, clusters do appear to form within the data. It is difficult to interpret the relationship btween gender and spending score with this plot, so more testing should be done.

## Age and Spending Score

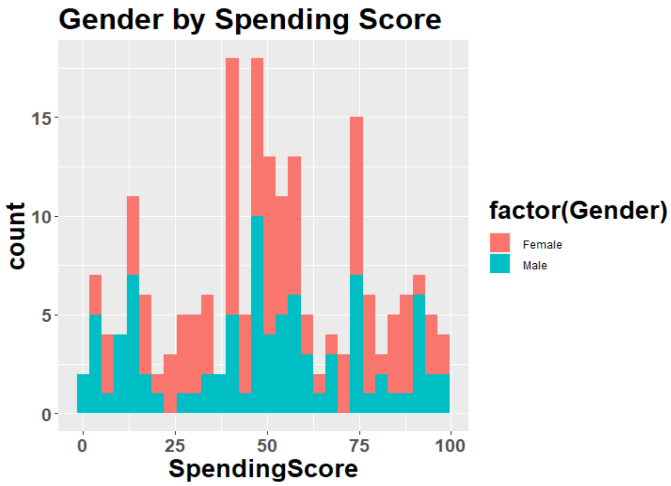
Is there a relationship between age and spending score? We know the two features have an inverse relationship at the aggregate level, but it is still important to understand the distribution.

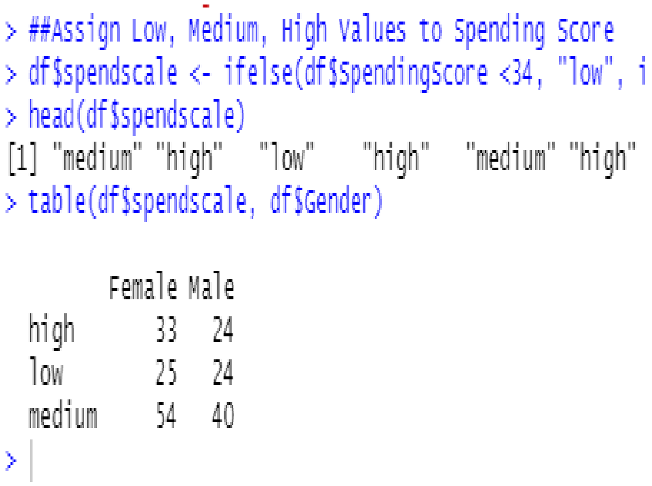


The inverse relationship is fairly clear. Interestingly, all of the highest spending individuals are between the ages 20 and 40. The decline in spending seems to drop off sharply for customers who are 40 years old, at which point the spending score seems to level off. Interestingly, almost all individuals over 65 years old are moderate spenders, and this is an insight that could not be communicated through simple correlation between the two variables. This age group is likely made up of grandparents who are shopping for their young kids.

## Gender and Spending Score

What is the breakdown of gender by spending score?





```
> ##Assign Low, Medium, High Values to Spending Score
> df$spendscale <- ifelse(df$SpendingScore <34, "low", i
> head(df$spendscale)
[1] "medium" "high"   "low"    "high"   "medium" "high"
> table(df$spendscale, df$Gender)

        Female Male
  high      33   24
  low       25   24
  medium    54   40
> 
```

It seems that females make up the majority of medium and high spend customers. With a bit of feature engineering, we can bin these spending scores into tiers (low-high) to communicate this. In doing so, we find that there is an increase in the proportion of females as the spending tier increases (albeit slight increase).

| Tier | Female | Male | Female % |
|---|---:|---:|---:|
| Low | 25 | 24 | 51% |
| Medium | 54 | 40 | 57% |
| High | 33 | 24 | 58% |

### Key Takeaways from EDA

We were able to dig deeper to understand relationship between variables

- Age 30-39 accounts for 30.5% of all customers, and 61% of this age group are women
- No clear relationship between income and spending score, though clusters seem to form
- Spending score is highest among those aged 20-39, and drops off at 40
    - Customers older than 65 are almost all moderate spenders
- Proportion of women increases with spending tier (made possible through feature engineering)

After the exploratory phase, we have already communicated multiple insights that are valuable for the mall. An understanding of the prevalence of specific characteristics within the customer base should be leveraged to tailor marketing and revenue strategies.

## 5. Preprocessing Data:

In simple words, pre-processing refers to the transformations applied to your data before feeding it to the algorithm. It involves further cleaning of data, data transformation, data scaling and many more things. For our data, we will deal with skewness and scale the numerical variables

Standardizing data is a good practice when clustering, as the range of values within each feature will influence how the cluster is formed, which is not usually desirable. Kmeans clustering uses Euclidean distance to measure the similarity between objects, so if a feature has a range much larger than another feature, it will dominate the other features in the clustering process.

$$Z = \frac{x - \mu}{\sigma}$$

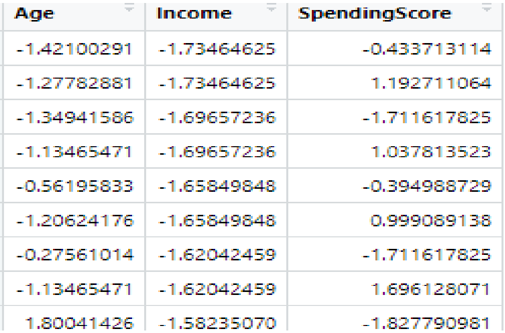
$Z$ = standard score

$x$ = observed value

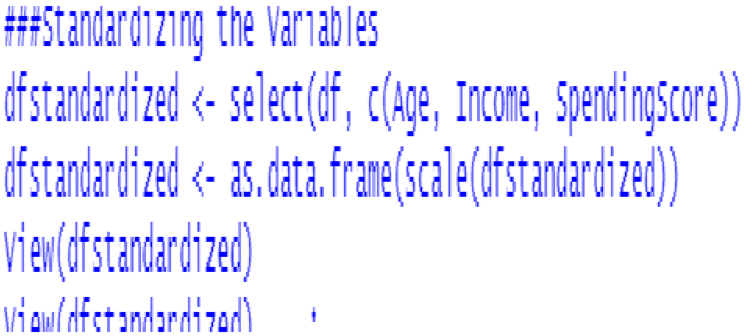
$\mu$ = mean of the sample

$\sigma$ = standard deviation of the sample

The model was initiated on the standardized data for clustering based on age, income, and spending score.



| Age | Income | SpendingScore |
|---|---|---|
| -1.42100291 | -1.73464625 | -0.433713114 |
| -1.27782881 | -1.73464625 | 1.192711064 |
| -1.34941586 | -1.69657236 | -1.711617825 |
| -1.13465471 | -1.69657236 | 1.037813523 |
| -0.56195833 | -1.65849848 | -0.394988729 |
| -1.20624176 | -1.65849848 | 0.999089138 |
| -0.27561014 | -1.62042459 | -1.711617825 |
| -1.13465471 | -1.62042459 | 1.696128071 |
| 1.80041426 | -1.58235070 | -1.827790981 |

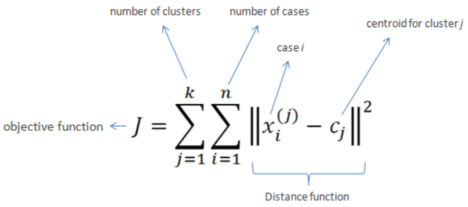


```
###Standardizing the Variables
dfstandardized <- select(df, c(Age, Income, SpendingScore))
dfstandardized <- as.data.frame(scale(dfstandardized))
View(dfstandardized)
View(dfstandardized)
```

# 6. Initial Model Building:

## K-Means Clustering:

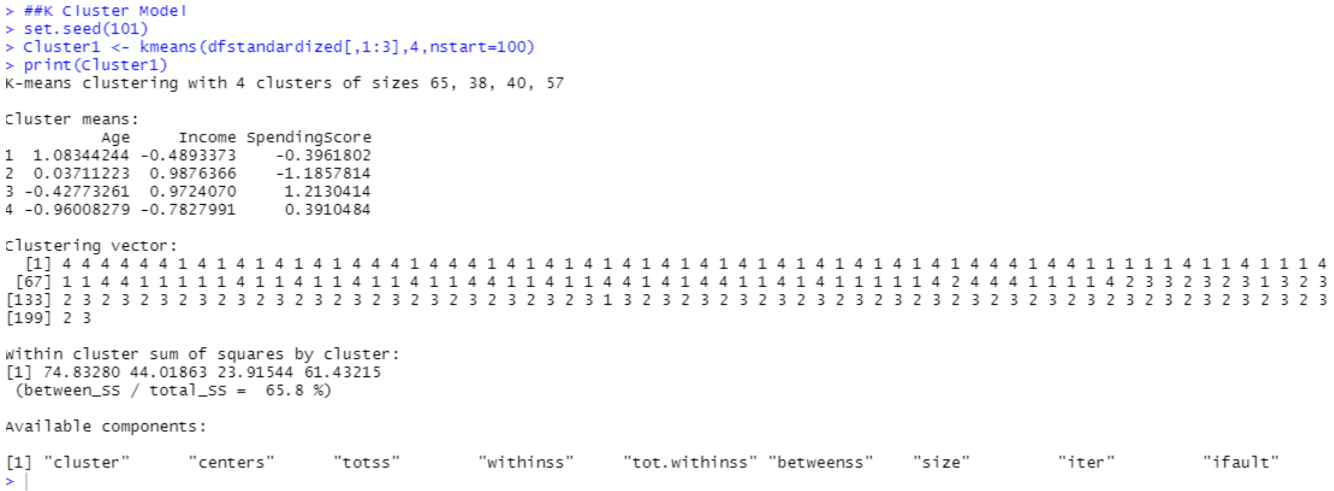
K-Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.
Formula



$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

I'll use k-means algorithm, a method which identifies k number of centroids (centers of a cluster) and allocates every single observation (data point) to the nearest cluster. In the previous part of my paper we found out that three variables: AnnualIncome, SpendingScore & Age are the ones that influence consumer behavior the most. Therefore the clusters will be generated only on the basis of these three variables. I initially perform kmeans clustering algorithm on standardized data we found following outcomes

Clustered data (K-Means) into 4 customer segments based on age, income, and spending score, facilitating targeted marketing to decrease customer acquisition cost and to increase customer retention.The goodness of cluster is 65.8%



```
> ##K Cluster Model
> set.seed(101)
> Cluster1 <- kmeans(dfstandardized[,1:3],4,nstart=100)
> print(Cluster1)
K-means clustering with 4 clusters of sizes 65, 38, 40, 57

Cluster means:
          Age     Income SpendingScore
1  1.08344244 -0.4893373    -0.3961802
2  0.03711223  0.9876366    -1.1857814
3 -0.42773261  0.9724070     1.2130414
4 -0.96008279 -0.7827991     0.3910484

Clustering vector:
  [1] 4 4 4 4 4 4 1 4 1 4 1 4 1 4 1 4 4 4 1 4 4 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 1 4 4 4 1 4 4 1 1 1 1 1 4 1 1 4 1 1 1 4
 [67] 1 1 4 4 1 1 1 1 1 4 1 1 4 1 1 4 1 1 4 1 1 4 4 1 1 4 1 1 4 4 1 4 1 4 4 1 1 4 1 4 1 1 1 1 1 4 2 4 4 4 1 1 1 1 4 2 3 3 2 3 2 3 1 3 2 3
[133] 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 1 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3 2 3
[199] 2 3

Within cluster sum of squares by cluster:
[1] 74.83280 44.01863 23.91544 61.43215
 (between_SS / total_SS =  65.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
> |
```

# 7. Tuning the Model:

When using Kmeans, the number of clusters (k), is a value to be set by the user. There are a few methods to determine the appropriate number of clusters, as shown below.

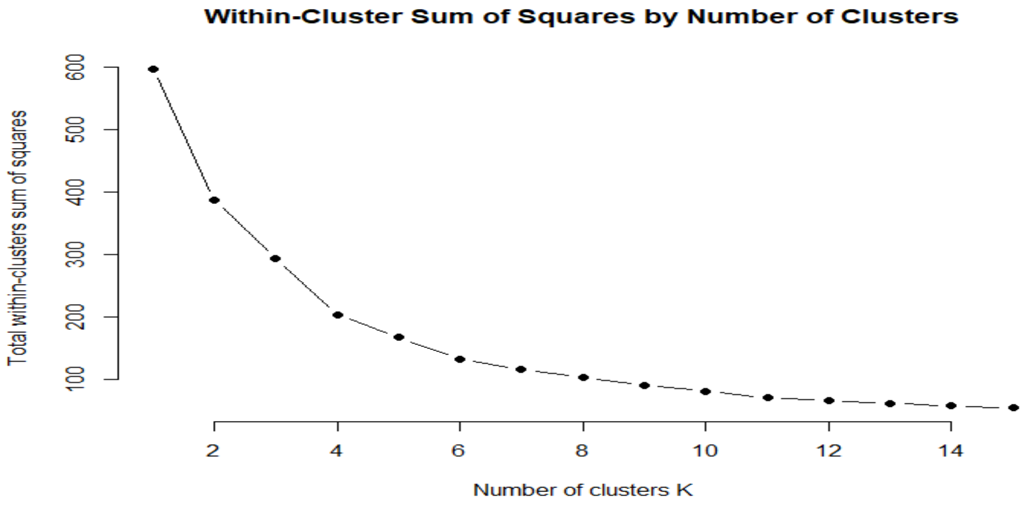
## 7.1 The Elbow Method:

The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.
When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the “elbow” (the point of inflection on the curve) is the best value of k. The “arm” can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.
We use the Elbow Method which uses Within Cluster Sum Of Squares (WCSS) against the the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is

given by the below formula.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

where Yi is centroid for observation Xi. The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.

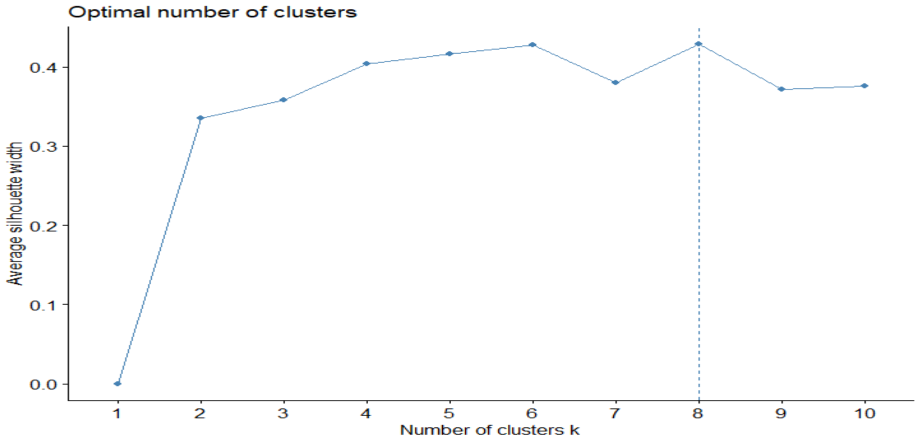


The within-cluster sum of squares is a measure of the variability of observations within each cluster. A smaller sum of squares creates a more "compact" cluster, which is usually preferred. The elbow method seeks to optimize the number of clusters by selecting the number of clusters that minimize the within-cluster sum of squares before diminishing returns takes place. As the number of clusters increases, the sum of squares will decrease, as the data is being broken into a larger number of "tighter" clusters. The key is to select the number of clusters at which point we can account for a substantial amount of variance within the data while also not overfitting, and the "elbow" forms at a point where the addition of a cluster results in a relatively minimal decrease in error. Given the above plot, one could say that clusters between 4 and 6 would be optimal, so I decided to use 6 clusters, as this number would provide the most compact clusters, and thus the most detailed information about the customers. Choosing to have fewer clusters would result in looser associations between datapoints, minimizing the amount of meaningful takeaways that can be deduced.

## 7.2 Average Silhouette Method:

The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

Silhouette coefficients (as these values are referred to as) near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.

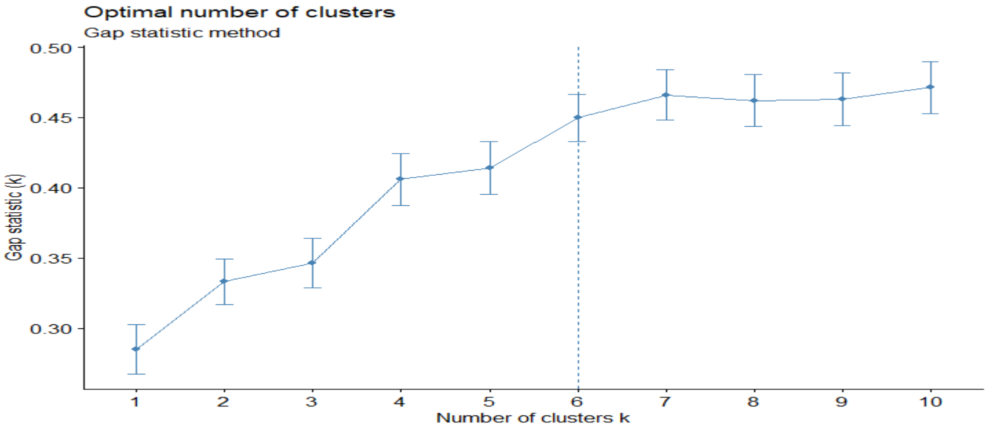


Silhouette analysis allows you to calculate how similar each observation is with the cluster it is assigned relative to other clusters. This metric ranges from -1 to 1 for each observation in your data and can be interpreted as a poor fit (-1), a loose fit that is borderline between clusters (0), and a great fit (1). Maximizing the silhouette metric is the goal, and should yield the optimal amount of clusters.

## 7.3 Gap Statistic Method

We can use this method to any of the clustering method like K-means, hierarchical clustering etc. 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. With the help of **Monte Carlo simulations**, one can produce the sample dataset. For each variable in the dataset, 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.

For computing the gap statistics method we can utilize the clusGap function for providing gap statistic as well as standard error for a given output.



The gap statistic compares the total intracluster variation for different values of k with their expected values under null reference distribution of the data (i.e. a distribution with no obvious clustering). The reference dataset is generated using Monte Carlo simulations of the sampling process. The goal is to maximize the gap-stat, though assuming that the plot is just going to continue to increase, of course, the results are less useful. Tibshirani suggests the 1-standard-error method:

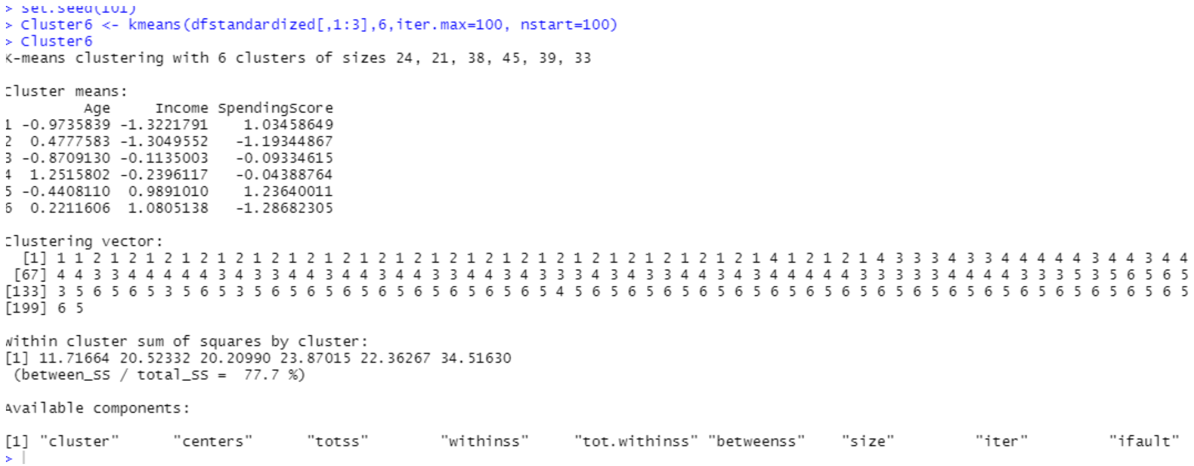
Choose the cluster size k^ to be the smallest k such that Gap(k)≥Gap(k+1)−sk+1.

Which translates to identifying the point at which the rate of increase of the gap statistic begins to "slow down".

Given all of the above, I felt most comftorable moving forward with 6 clusters. Because 2 methods indicate 6 clusters are best for your model.

## 8. Final Model:

After doing tunning technique we see that 6 clusters are best for your model so we implement our final model and we choose 6  as K_ number of clusters. We found below results.



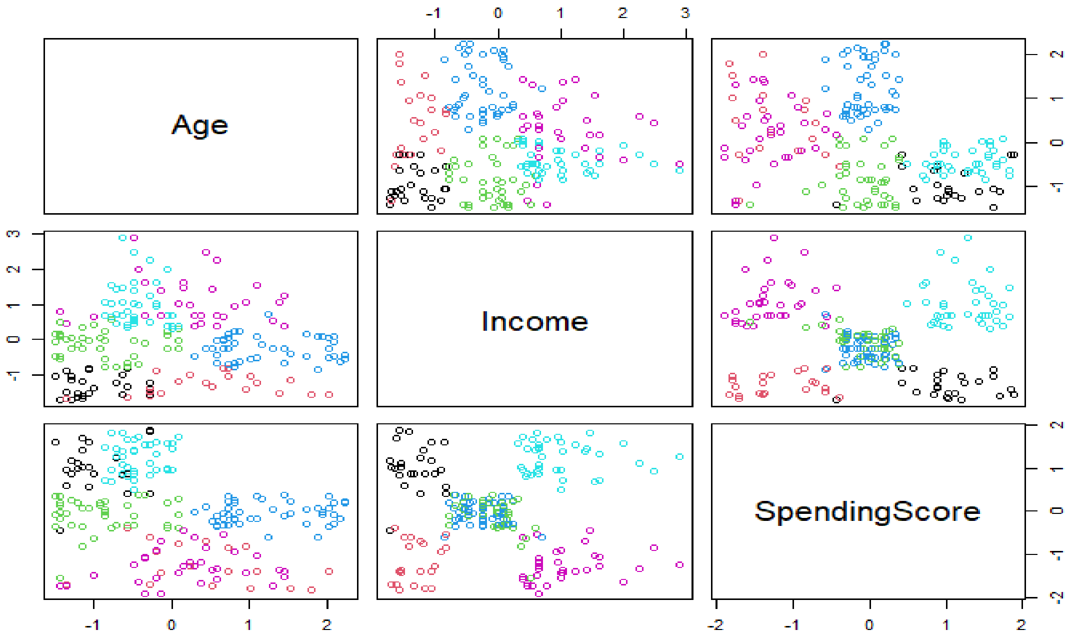
```
> set.seed(101)
> Cluster6 <- kmeans(dfstandardized[,1:3],6,iter.max=100, nstart=100)
> Cluster6
K-means clustering with 6 clusters of sizes 24, 21, 38, 45, 39, 33

Cluster means:
         Age     Income SpendingScore
1 -0.9735839 -1.3221791    1.03458649
2  0.4777583 -1.3049552   -1.19344867
3 -0.8709130 -0.1135003   -0.09334615
4  1.2515802 -0.2396117   -0.04388764
5 -0.4408110  0.9891010    1.23640011
6  0.2211606  1.0805138   -1.28682305

Clustering vector:
  [1] 1 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 4 1 2 1 2 1 4 3 3 3 4 3 3 4 4 4 4 4 3 4 4 3 4 4
 [67] 4 4 3 3 4 4 4 4 4 3 4 3 3 4 4 3 4 4 3 4 4 3 3 4 4 3 4 3 3 3 4 3 4 3 3 4 4 3 4 3 4 4 4 4 4 3 3 3 3 3 4 4 4 4 3 3 3 5 3 5 6 5 6 5
[133] 3 5 6 5 6 5 3 5 6 5 3 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 4 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5 6 5
[199] 6 5

Within cluster sum of squares by cluster:
[1] 11.71664 20.52332 20.20990 23.87015 22.36267 34.51630
 (between_SS / total_SS =  77.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
>
```

Now we can see the goodness of cluster is increase 77.7%

## 9. Matrix of Tuned Model:

This is a matrix of clusters built on two-variable combinations of age, income, and spending score.

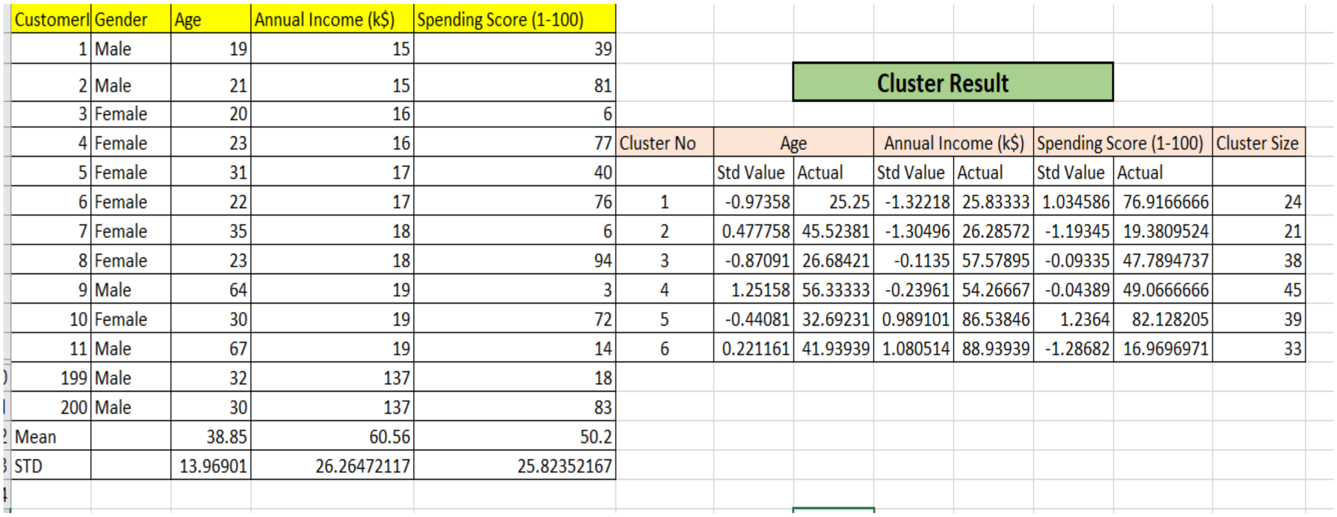
After tuning the model to have 6 clusters, the clusters explained 77.7% of the variance within the data



## 10. Results:

- Clustered data (K-Means) into 6 customer segments based on age, income, and spending score, facilitating targeted marketing to decrease customer acquisition cost and to increase customer retention
- Identified ideal segments to focus on given their potential increase in lifetime value
- Recommended initiatives such as creating/revamping a loyalty program, optimizing the marketing mix, tailored promotions, and enhancing the in-store experience as levers to drive traffic and increase revenue

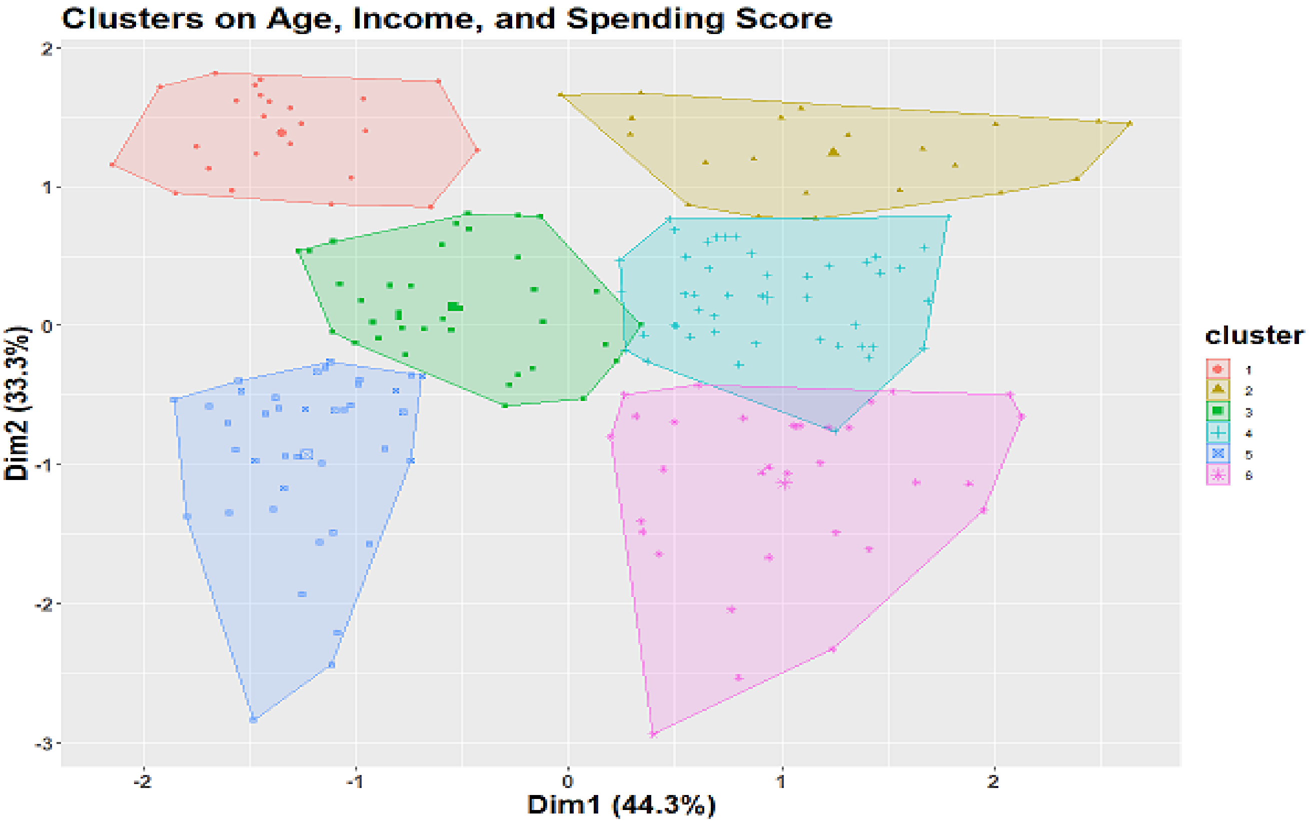
Formula : X = (Z*STDEV)+Mean



| CustomerI | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|
| 1 | Male | 19 | 15 | 39 |
| 2 | Male | 21 | 15 | 81 |
| 3 | Female | 20 | 16 | 6 |
| 4 | Female | 23 | 16 | 77 |
| 5 | Female | 31 | 17 | 40 |
| 6 | Female | 22 | 17 | 76 |
| 7 | Female | 35 | 18 | 6 |
| 8 | Female | 23 | 18 | 94 |
| 9 | Male | 64 | 19 | 3 |
| 10 | Female | 30 | 19 | 72 |
| 11 | Male | 67 | 19 | 14 |
| 199 | Male | 32 | 137 | 18 |
| 200 | Male | 30 | 137 | 83 |
| Mean |  | 38.85 | 60.56 | 50.2 |
| STD |  | 13.96901 | 26.26472117 | 25.82352167 |

Cluster Result

| Cluster No | Age |  | Annual Income (k$) |  | Spending Score (1-100) |  | Cluster Size |
|---|---|---|---|---|---|---|---|
|  | Std Value | Actual | Std Value | Actual | Std Value | Actual |  |
| 1 | -0.97358 | 25.25 | -1.32218 | 25.83333 | 1.034586 | 76.9166666 | 24 |
| 2 | 0.477758 | 45.52381 | -1.30496 | 26.28572 | -1.19345 | 19.3809524 | 21 |
| 3 | -0.87091 | 26.68421 | -0.1135 | 57.57895 | -0.09335 | 47.7894737 | 38 |
| 4 | 1.25158 | 56.33333 | -0.23961 | 54.26667 | -0.04389 | 49.0666666 | 45 |
| 5 | -0.44081 | 32.69231 | 0.989101 | 86.53846 | 1.2364 | 82.128205 | 39 |
| 6 | 0.221161 | 41.93939 | 1.080514 | 88.93939 | -1.28682 | 16.9696971 | 33 |



1. The first cluster consists of younger shoppers who have moderate levels of income and spending score. This cluster consists of people who can likely transition into more loyal customers given they are already frequent visitors who have a reasonable level of disposable income. In-store promotions and a revamped loyalty program could incentivize this group to become higher spenders. It helps that this cluster is young (mean age of 27), as reaching them shouldn’t be difficult considering the majority of customers are under 40, so current marketing tactics are likely effective.

2. The sixth cluster presents another interesting opportunity. Accounting for nearly 17% of customers, this cluster represents high income, but low spending score individuals. This group certainly has the financial means to spend more within the mall, however, they likely consume through other channels (other store outlets, online shopping), so more effort needs to be put towards getting them into the mall. Enhancing the in-store experience should help create a more viable alternative to their current purchasing outlets. Stores with customer journeys that cannot be easily replicated online should take full advantage (such as fitting clothes in person, in-person demonstrations of technology, etc.) Additionally, implementing click and collect and other features to expedite the ordering process can provide the convenience that this segment seeks. Finally, a mean age of 42 communicates that this segment likely consists of parents of young children, so promotions and product offerings catered towards that demographic should be emphasized as well.

3. The fourth cluster is the largest, accounting for nearly 23% of customers. Similar to the third cluster, this group possesses a moderate spending score and income, so they likely possess the means to spend more within the mall. On the other hand, this demographic is older (mean of 56 years old), so the lifetime value of these customers may not be as high as the younger individuals in the same financial bracket. That coupled with the fact that a different marketing mix would likely be used to reach this group means it is not necessarily the best use of resources. Customer acquisition cost and expected lifetime value should be considerations here.

4. The fifth cluster is the wealthiest segment, and already the most loyal. Prioritizing this group might seem enticing, and may reap rewards, but given that this segment is already extremely loyal, targeting this group doesn't seem necessary to drive purchases. The resources can be better spent elsewhere to move the needle on customers who have the financial means to purchase more than they currently are, but who also need to be enticed with an additional value proposition.

5. The second and fisrt clusters are lower income individuals, and though the second cluster has a considerable spending score, these groups are the smallest segments and will have a tough time spending more money than they are already, so these groups are lower priority segments.

## 11. Conclusion:

Companies, Malls, super markets on Small Business Enterprises should carry out Market Basket Analysis for their business. This will enable companies to target specific groups of customers, a customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities. When a group of customers is sent personalized messages as part of a marketing mix that is designed around their needs, it's easier for companies to send those customers special offers meant to encourage them to buy more products. Customer segmentation can also improve customer service and assist in customer loyalty and retention. As a by-product of its personalized nature, marketing materials sent out using customer segmentation tend to be more valued and appreciated by the customer who receives them as opposed to impersonal brand messaging that doesn't acknowledge purchase history or any kind of customer relationship Finally with customer segmentation Companies will stay a step ahead of competitors in specific sections of the market and identify new products that exist or potential customers could be interested in or improving products to meet customer expectations.

# 12. Appendix

## Github Link :

https://github.com/vicky61992/Statistics_Unsupervised_Learning_project

```
getwd()
setwd("C:/Users/VSBAG/Desktop/DSE_Milan/3rd_sem_subject/ML&SL/SL_Project/My_Unsupervised_Learning")

###Packages
install.packages("ggplot2")
install.packages("plotly")
install.packages("ggthemes")
install.packages("corrplot")
install.packages("dplyr")
install.packages("caTools")
install.packages("RColorBrewer")
install.packages("cluster")
install.packages("factoextra")

library(ggplot2)
library(plotly)
library(ggthemes)
library(corrplot)
library(dplyr)
library(caTools)
library(RColorBrewer)
library(cluster)
library(factoextra)

##Loading in File
df <- read.csv("https://raw.githubusercontent.com/vicky61992/Statistics_Unsupervised_Learning_project/main/Mall%20Customers.csv")

#Dimensions of Data
dim(df)
#Features of Data
head(df)
#Structure of Data
```

```
str(df)


###Renaming Spening Score and Annual Income
df <- rename(df, Income = Annual.Income..k..)
df <- rename(df, SpendingScore = Spending.Score..1.100.)
colnames(df)
summary(df)

## Checking for Na & Nulls
sapply(df,function(x)sum(is.na(x)))

#Data Exploration
# Plot Themes
bold_axis <- element_text(face = "bold", color = "black", size = 20)
axis_text <- element_text(face = "bold", size = 14)

##Gender Breakdown
genderplot <- ggplot(df, aes(x=factor(1), fill=Gender))+
  geom_bar(width = 1) +
  coord_polar("y")
genderplot + theme(axis.text = axis_text) + theme(title = bold_axis) + ggtitle("Gender Breakdown")

table(df$Gender)/nrow(df)

## Plot Customers by Age
Plotage <- ggplot(df,aes(x=Age))
Plotage + geom_histogram(fill="purple", alpha = 0.7) + theme(axis.text = axis_text) + theme(title = bold_axis) + ggtitle("Histogram of Customer Age")

##Bucket Age by 10s
Twenties <- filter(df, Age <30)
table(Twenties$Gender)

Thirties <- filter(df, Age >= 30 & Age <= 39)
table(Thirties$Gender)

Fourties <- filter(df, Age >= 40 & Age <= 49)
table(Fourties$Gender)

(nrow(Twenties) + nrow(Thirties) + nrow(Fourties)) / nrow(df)

FiftiesPlus <- filter(df, Age >= 50)
table(FiftiesPlus$Gender)
table(FiftiesPlus$Income)

## Plot by Income
Plotincome <- ggplot(df, aes(x= Income))
Plotincome + geom_histogram(fill="orange", alpha = 0.7) + theme(axis.text = axis_text) + theme(title =
bold_axis) + ggtitle("Histogram of Customer Income")
mean(df$Income)

## Plot by Spending Score
Plotscore <- ggplot(df, aes(x = SpendingScore))
Plotscore + geom_histogram(fill="pink", alpha = 0.8) + theme(axis.text = axis_text) + theme(title = bold_axis) +
ggtitle("Histogram of Customer Spending Score")
mean(df$SpendingScore)
sd(df$SpendingScore)
OneSD_SpendingSCore <- filter(df, SpendingScore >= 24 & SpendingScore <= 76)
nrow(OneSD_SpendingSCore) / nrow(df)  ## % of people within 1SD

##Creating a Binary Male/Female Column in case Gender will be used as numerical value
df$Male <- ifelse(df$Gender == "Male", "1", 0)
df$Male <- as.numeric(df$Male)

str(df$Male)

##Creating Subset for purpose of using numerical male/female
dfmale <- subset(df, select = -c(Gender))
head(dfmale)


### Deeper Exploration - Relationships b/t Variables
##Correlation between variables
Correlation <- cor(dfmale)
corrplot(Correlation,order ="hclust", col = brewer.pal(n=8, name = "RdBu"))
corrplot(Correlation, method = "number")
cor(dfmale)

## Histogram of Age and Gender
Plotagegender <- ggplot(df, aes(x = Age))
Plotagegender + geom_histogram(aes(fill = factor(Gender))) + ggtitle("Histogram of Customers by Age (Colored
by Gender)") + theme(axis.text = axis_text) + theme(title = bold_axis)

table(Thirties$Gender)

##Scatter of Income and Spending Score, colored by gender
scatter <- ggplot(df, aes(x = Income, y = SpendingScore)) + geom_point(aes(size = 2, color = factor(Gender)))
scatter + geom_smooth(method = "lm", color = "black") + theme(axis.text = axis_text) + theme(title =
bold_axis) + ggtitle("Income and Spending Score (Colored by Gender)")

##Scatter of Age and Spending Score, colored by Gender
scatter2 <- ggplot(df, aes(x = Age, y = SpendingScore)) + ggtitle("Customer Age and Spending Score (Colored by
Gender)") + geom_point(aes(size = 2, color = factor(Gender)))
scatter2 +geom_smooth(method = "lm", color ="black") + theme(axis.text = axis_text) + theme(title =
bold_axis)
```

```
##histogram of Gender and Spending Score
Genderscore <- ggplot(df, aes(x = SpendingScore))
Genderscore + geom_histogram(aes(fill= factor(Gender))) + theme(axis.text = axis_text) + theme(title =
bold_axis) + ggtitle("Gender by Spending Score")

##Assign Low, Medium, High Values to Spending Score
df$spendscale <- ifelse(df$SpendingScore <34, "low", ifelse(df$SpendingScore >= 34 & df$SpendingScore <= 67,
"medium", "high"))
head(df$spendscale)

table(df$spendscale, df$Gender)

###Standardizing the Variables
dfstandardized <- select(df, c(Age, Income, SpendingScore))
dfstandardized <- as.data.frame(scale(dfstandardized))

##K Cluster Model
set.seed(101)
Cluster1 <- kmeans(dfstandardized[,1:3],4,nstart=100)
print(Cluster1)

##Tweaking the Model
#Computing WSS ### Elbow Method
k.max <- 15
wss <- sapply(1:k.max,
        function(k){kmeans(dfstandardized[,1:3],k, iter.max = 100, nstart = 100)$tot.withinss})
wss

#Plotting Elbow method
plot(1:k.max, wss,
    type = "b", pch = 19, frame = FALSE,
    xlab = "Number of clusters K",
    ylab = "Total within-clusters sum of squares",
    main = "Within-Cluster Sum of Squares by Number of Clusters")

##Aerage Silhouette Method

fviz_nbclust(dfstandardized, kmeans, method = "silhouette")

##Gap Statistic Method
set.seed(101)
fviz_nbclust(dfstandardized, kmeans, nstart = 25,  method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")

###Adjusting Kmeans Model
set.seed(101)
Cluster6 <- kmeans(dfstandardized[,1:3],6,iter.max=100, nstart=100)
Cluster6
```

```
plot(dfstandardized[,1:3], col=Cluster6$cluster)

Cluster6$centers

##Visualize the data, Clustering on Age, Income, and Spending Score
fviz_cluster(Cluster6, data = dfstandardized[,1:3], label = 0, main = "Clusters on Age, Income, and Spending Score") + theme(axis.text = axis_text) + theme(title = bold_axis)
```