Indira Gandhi Delhi Technical University for Women

(Established by Govt. of Delhi vide Act 09 of 2012)

Kashmere Gate, Delhi - 110006

PROJECT REPORT

R Programming(B.Tech AIML)

Group Members: **Suryanshi(069), Swasti(070), Tanisha(071)**

Index

- ACKNOWLEDGMENT
- CERTIFICATE OF COMPLETION
- PROBLEM STATEMENT(AIM)
- INTRODUCTION
- OBJECTIVE(SCOPE)
- PLATFORM
- RESEARCH METHODOLOGY
    - □CUSTOMER SEGMENTATION (ANALYSIS)
    - □DATA COLLECTION AND PREPARATION
    - □DATA VISUALISATION
    - □K-MEANS CLUSTERING
- CODE
- GRAPHS AND OUTPUTS (TESTING)
- ACCURACY LEVEL
- AREA OF IMPROVEMENT
- PROJECTED OUTCOME
- INFERENCE
- REFERENCES

ACKNOWLEDGEMENT

I would like to thank all those who have contributed to the completion of this project and helped us with valuable suggestions for improvement. I am grateful to the members of my team who have collaborated with me on various aspects of this project . Their expertise, insights, and collaboration have enriched the quality of this work .I appreciate their hard work, commitment, and the countless hours they have devoted to this project. Furthermore ,I am extremely grateful to our Prof. Santanoo Pattnaik, Professor, Department of Information Technology, for providing me with the atmosphere for the creative work, guidance and encouragement.

Certificate Of Completion

This is to certify that Ms.Suryanshi , Ms.Swasti and Ms.Tanisha ,students of the programme B.Tech AI & ML, department of information technology , have successfully completed their project in R programming titled ‘CUSTOMER SEGMENTATION USING R’ under the able guidance of Prof. Santanoo Pattnaik .

Signature of project guide

PROBLEM STATEMENT

The aim of customer segmentation is to identify distinct groups or segments of customers based on their characteristics and behaviors. Customer segmentation allows businesses to understand their customer base better and tailor their marketing strategies to each segment's specific needs and preferences.

CUSTOMER SEGMENTATION CLASSIFICATION

Introduction

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, we will make use of k-mean Clustering which is the essential algorithm for clustering unlabelled datasets.

SCOPE

Whenever you need to find your best customer, customer segmentation is the ideal methodology. We will perform one of the most essential applications of machine learning – Customer Segmentation. In this project, we will implement customer segmentation in R.

PLATFORM : R Studio

R was specifically designed for statistical analysis, which makes it highly suitable for data science applications. Although the learning curve for programming with R can be steep, especially for people without prior programming experience, the tools now available for carrying out text analysis in R make it easy to perform powerful, cutting-edge text analytics using only a few simple commands. One of the keys to R’s explosive growth has been its densely populated collection of extension software libraries, known in R terminology as packages, supplied and maintained by R’s extensive user community. Each package extends the functionality of the base R language and core packages, and in addition to functions and data must include documentation and examples, often in the form of vignettes demonstrating the use of the package. The best known package repository, the Comprehensive R Archive Network (CRAN), currently has over 10,000 packages that are published.

Text analysis in particular has become well established in R. There is a vast collection of dedicated text processing and text analysis packages, from low-level string operations to advanced text modeling techniques such as fitting Latent Dirichlet Allocation models, R provides it all. One of the main advantages of performing text analysis in R is that it is often possible, and relatively easy, to switch between different packages or to combine them. Recent efforts among the R text analysis developers' community are designed to promote this interoperability to maximize flexibility and choice among users. As a result, learning the basics for text analysis in R provides access to a wide range of advanced text analysis features interoperability to maximize flexibility and choice among users. As a result, learning the basics for text analysis in R provides access to a wide range of advanced text analysis features.

PROJECT SPECIFICATION

□ R Studio version 1.2.5033

DATASET

□ Mall_Customers.csv

PACKAGES REQUIRED:

plotrix , dplyr , ggplot2 , cluster

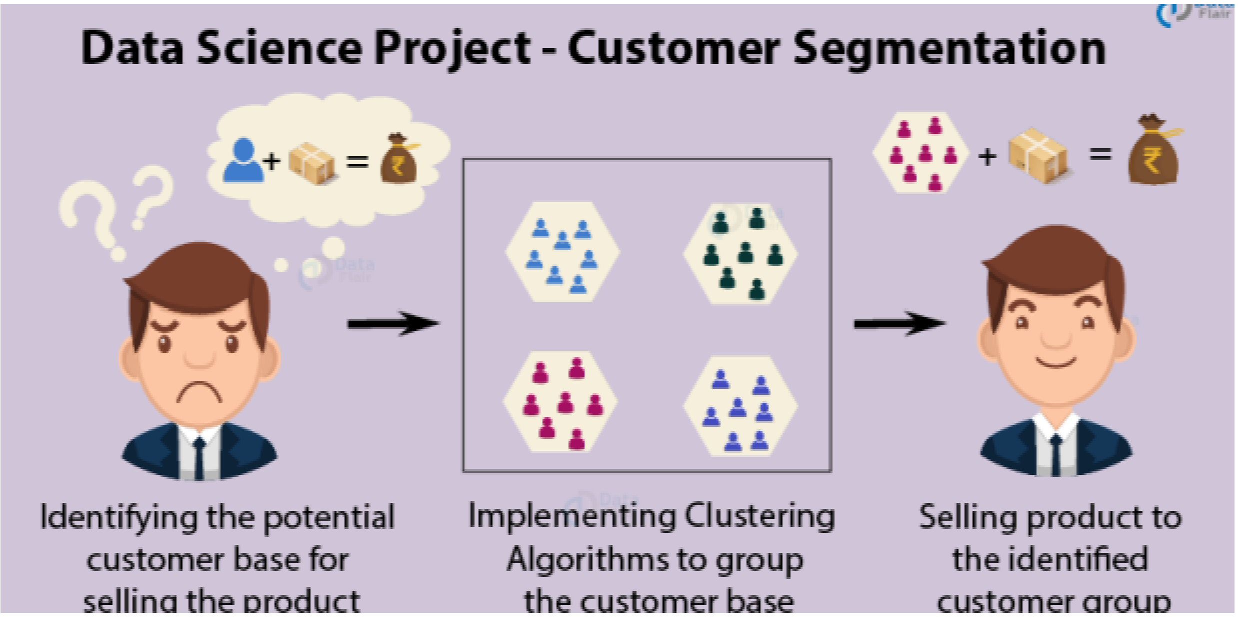
OBJECTIVE

To identify and gather relevant data sources, such as customer transaction data, product information, and customer demographics. Choosing suitable segmentation techniques based on the dataset and objectives. Common approaches include clustering algorithms like k-means, hierarchical clustering, or DBSCAN. Perform market basket analysis across segments to identify common patterns or associations that transcend individual segments. Develop a comprehensive report or presentation summarizing the project methodology, results, and actionable recommendations.

RESEARCH METHODOLOGY

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. 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.


Data Science Project - Customer Segmentation
Identifying the potential customer base for selling the product
Implementing Clustering Algorithms to group the customer base
Selling product to the identified customer group

IMPLEMENTATION

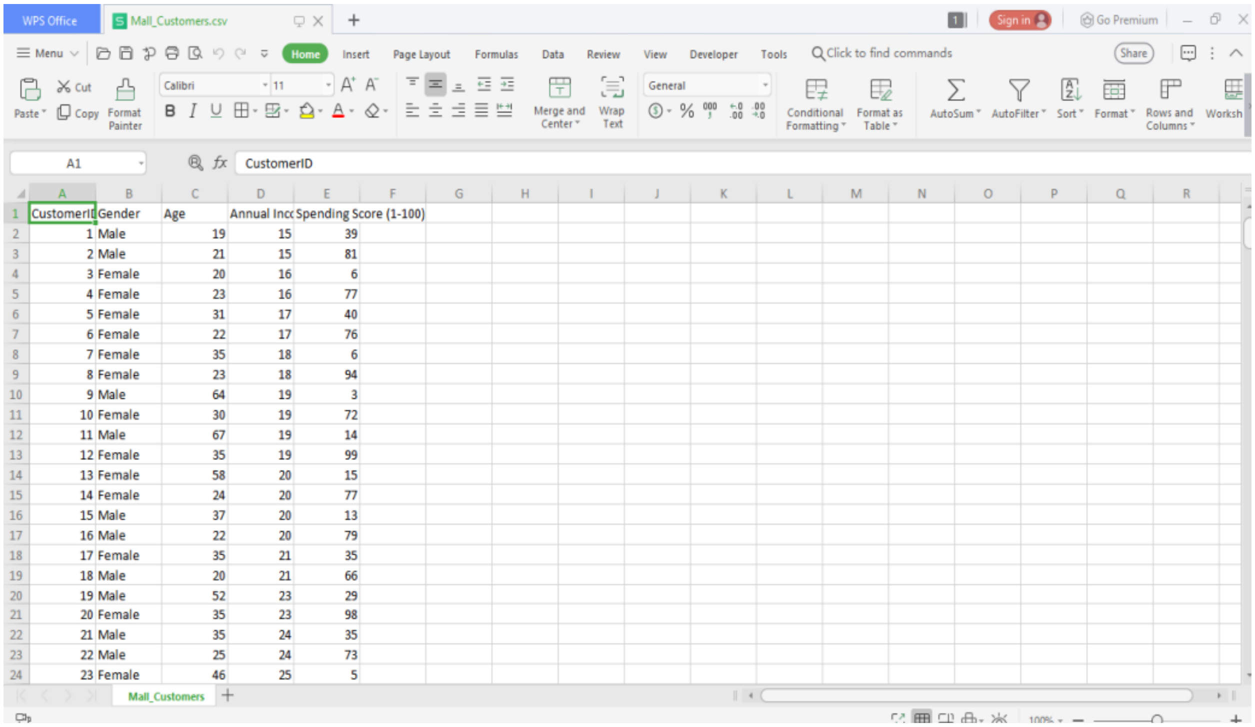
In the first step of this data science project, we will perform data exploration. We will import the essential packages required for this role and then read our data. Finally, we will go through the input data to gain necessary insights about it.

1. Data collection and preparation

Identify and gather relevant data sources, such as customer transaction data, product information, and customer demographics. Clean and pre-process the data, handling missing values, outliers, and transforming variables if necessary.

READING EVENTS FROM MALL_CUSTOMERS.CSV:-

Before going to customer segmentation analysis, the first step is to read the data for performing analysis on. The data is saved in a dataset named Mall_Customers.csv. This dataset contains 400 records of various types of customers. The events saved in the dataset are unstructured. To perform analysis, reading of the data set is done using command “read.csv”.



fx CustomerID

	A	B	C	D	E
1	CustomerID	Gender	Age	Annual Inco	Spending Score (1-100)
2	1	Male	19	15	39
3	2	Male	21	15	81
4	3	Female	20	16	6
5	4	Female	23	16	77
6	5	Female	31	17	40
7	6	Female	22	17	76
8	7	Female	35	18	6
9	8	Female	23	18	94
10	9	Male	64	19	3
11	10	Female	30	19	72
12	11	Male	67	19	14
13	12	Female	35	19	99
14	13	Female	58	20	15
15	14	Female	24	20	77
16	15	Male	37	20	13
17	16	Male	22	20	79
18	17	Female	35	21	35
19	18	Male	20	21	66
20	19	Male	52	23	29
21	20	Female	35	23	98
22	21	Male	35	24	35
23	22	Male	25	24	73
24	23	Female	46	25	5

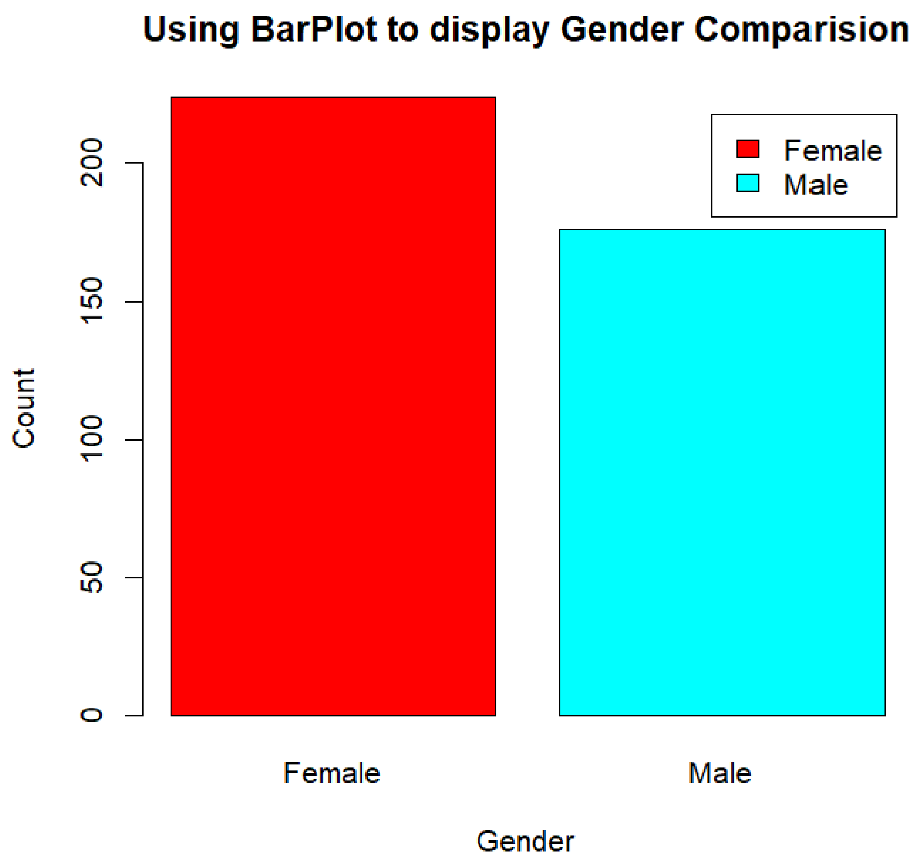
VARIABLES USED

DEPENDENT VARIABLES	In customer segmentation, the dependent variable could be a specific behavior or characteristic that you want to understand or predict. For example, it could be the likelihood of a customer making a purchase, their level of satisfaction, or their churn probability.No such variable is present in this project of customer segmentation.
INDEPENDENT VARIABLES	In a customer segmentation project, the independent variables are the features or attributes used to differentiate and group customers into distinct segments. These variables are typically selected based on their relevance and ability to capture meaningful differences among customers. The choice of independent variables depends on the specific goals of the segmentation project and the available data. Annual income and spending score are the independent variables in this project.

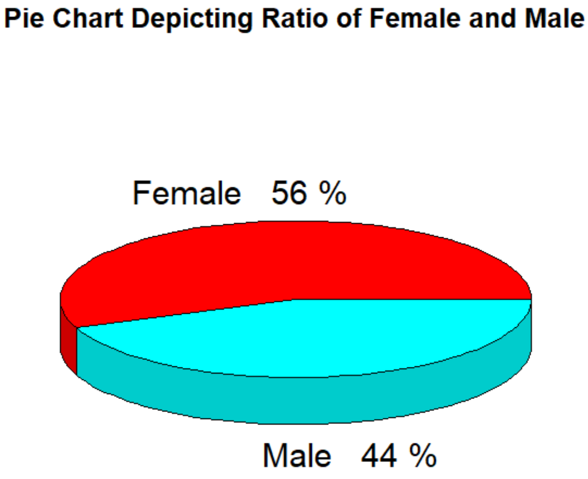
2. Data Visualisation

Customer Gender Visualization:

In this, we will create a barplot and a piechart to show the gender distribution across our customer_data dataset. A bar chart represents data in rectangular bars with length of the bar proportional to the value of the variable. R uses the function barplot() to create bar charts. R can draw both vertical and Horizontal bars in the bar chart. In the bar chart each of the bars can be given different colors.

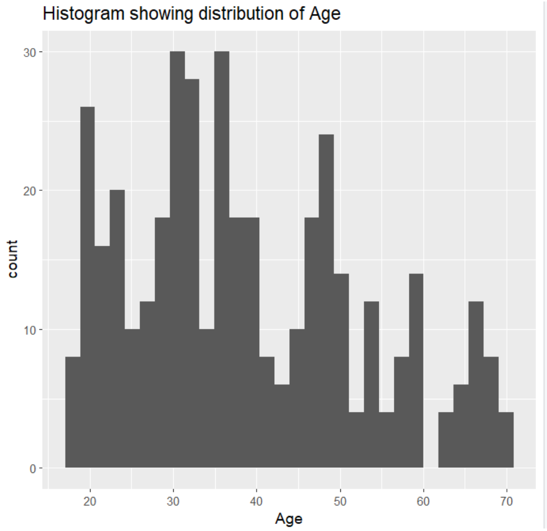


From the below graph, we conclude that the percentage of females is 56%, whereas the percentage of male in the customer dataset is 44%.



VISUALIZATION OF AGE DISTRIBUTION:

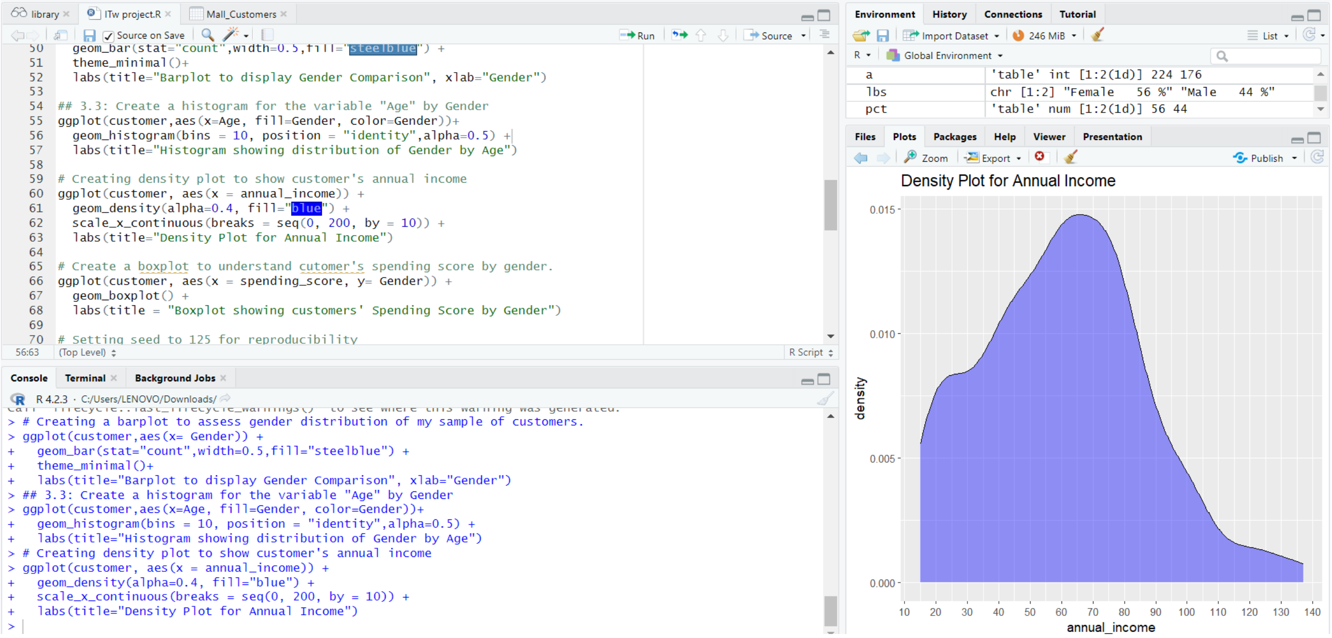
Let us plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking a summary of the Age variable.



From the above two visualizations, we conclude that the maximum customer ages are between 30 and 35. The minimum age of customers is 18, whereas, the maximum age is 70.

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 proceed to examine this data using a density plot .



```
  geom_bar(stat="count",width=0.5,fill="steelblue") +
  theme_minimal()+
  labs(title="Barplot to display Gender Comparison", xlab="Gender")

## 3.3: Create a histogram for the variable "Age" by Gender
ggplot(customer,aes(x=Age, fill=Gender, color=Gender))+
  geom_histogram(bins = 10, position = "identity",alpha=0.5) +
  labs(title="Histogram showing distribution of Gender by Age")

# Creating density plot to show customer's annual income
ggplot(customer, aes(x = annual_income)) +
  geom_density(alpha=0.4, fill="blue") +
  scale_x_continuous(breaks = seq(0, 200, by = 10)) +
  labs(title="Density Plot for Annual Income")

# Create a boxplot to understand cutomer's spending score by gender.
ggplot(customer, aes(x = spending_score, y= Gender)) +
  geom_boxplot() +
  labs(title = "Boxplot showing customers' Spending Score by Gender")

# Setting seed to 125 for reproducibility
```

56:63 (Top Level) R Script

```
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
> # Creating a barplot to assess gender distribution of my sample of customers.
> ggplot(customer,aes(x= Gender)) +
+   geom_bar(stat="count",width=0.5,fill="steelblue") +
+   theme_minimal()+
+   labs(title="Barplot to display Gender Comparison", xlab="Gender")
> ## 3.3: Create a histogram for the variable "Age" by Gender
> ggplot(customer,aes(x=Age, fill=Gender, color=Gender))+
+   geom_histogram(bins = 10, position = "identity",alpha=0.5) +
+   labs(title="Histogram showing distribution of Gender by Age")
> # Creating density plot to show customer's annual income
> ggplot(customer, aes(x = annual_income)) +
+   geom_density(alpha=0.4, fill="blue") +
+   scale_x_continuous(breaks = seq(0, 200, by = 10)) +
+   labs(title="Density Plot for Annual Income")
> 
```

```
a      'table' int [1:2(1d)] 224 176
lbs    chr [1:2] "Female   56 %" "Male   44 %"
pct    'table' num [1:2(1d)] 56 44
```

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

3. ALGORITHM

K MEANS ALGORITHM :

Choose suitable segmentation techniques based on your dataset and objectives. Common approaches include clustering algorithms like k-means, hierarchical clustering, or DBSCAN.

K-means Algorithm While using the k-means clustering algorithm, the first step is to indicate the number of clusters (k) that we wish to produce in the final output. The algorithm starts by selecting k objects from the dataset randomly that will serve as the initial centers for our clusters. These selected objects are the cluster means, also known as 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. We refer to this step as “cluster assignment”. When the assignment is complete, the algorithm proceeds to calculate a 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, the objects undergo reassignment. This goes on repeatedly through several iterations until the cluster assignments stop altering. The clusters that are present in the current iteration are the same as the ones obtained in the previous iteration.

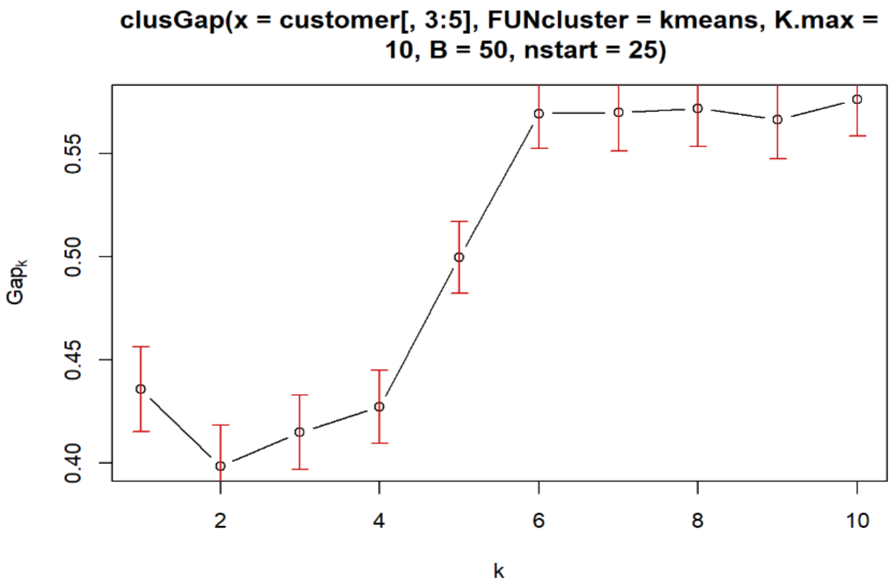
Summing up the K-means clustering –

- · We specify the number of clusters that we need to create.
- · The algorithm selects k objects at random from the dataset. This object is the initial cluster or mean. ·
- The closest centroid obtains the assignment of a new observation. We base this assignment on the Euclidean Distance between object and the centroid. ·
- k clusters in the data points update the centroid through calculation of the new mean values present in all the data points of the cluster.
- The kth cluster’s centroid has a length of p that contains means of all variables for observations in the k-th cluster
- We denote the number of variables with p. · Iterative minimization of the total within the sum of squares. Then through the iterative minimization of the total sum of the square, the assignment stops wavering when we achieve maximum iteration. The default value is 10 that the R software uses for the maximum iterations

CONDUCTING THE CLUSTER ANALYSIS:

1.Choose the number of clusters: We will use Gap statistics to determine the optimal number of clusters to segment the mall customers into.

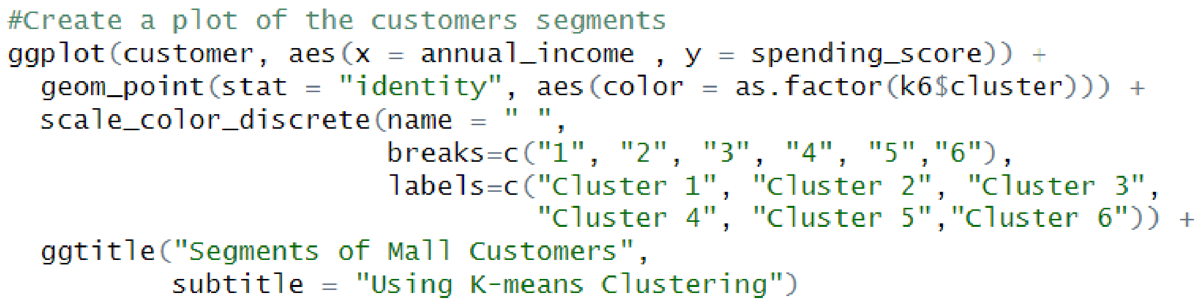
2.Creating the K-means Clustering model:



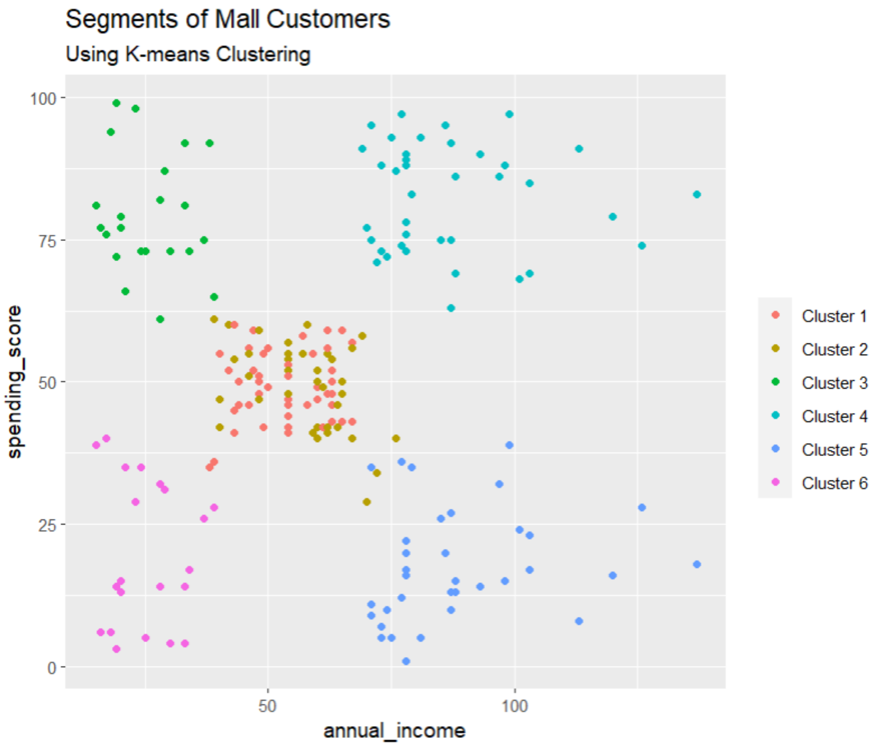
3. Showing the six means clustering: From the clustering model, it seems that two main components can explain up to 66% of the variability in the data. The results also show more details of the cluster, including the means of the customers' age, annual income, and spending score in each cluster.

4.Standardization: We will perform a Principal Component Analysis (PCA) to reduce the dimensionality of the data and capture the 2 most significant components of the data.

5.Plot of Customer Segments: Results from the PCA show that components 1 and 2 (PC1 and PC2) contribute the most variance to the data. The high correlation between PC1 and spending score (-0.786) and PC2 and annual income (-0.808) show that annual income and spending income are the 2 major components of the data.Finally, we will plot the customer segments based on results from the cluster analysis and PCA.



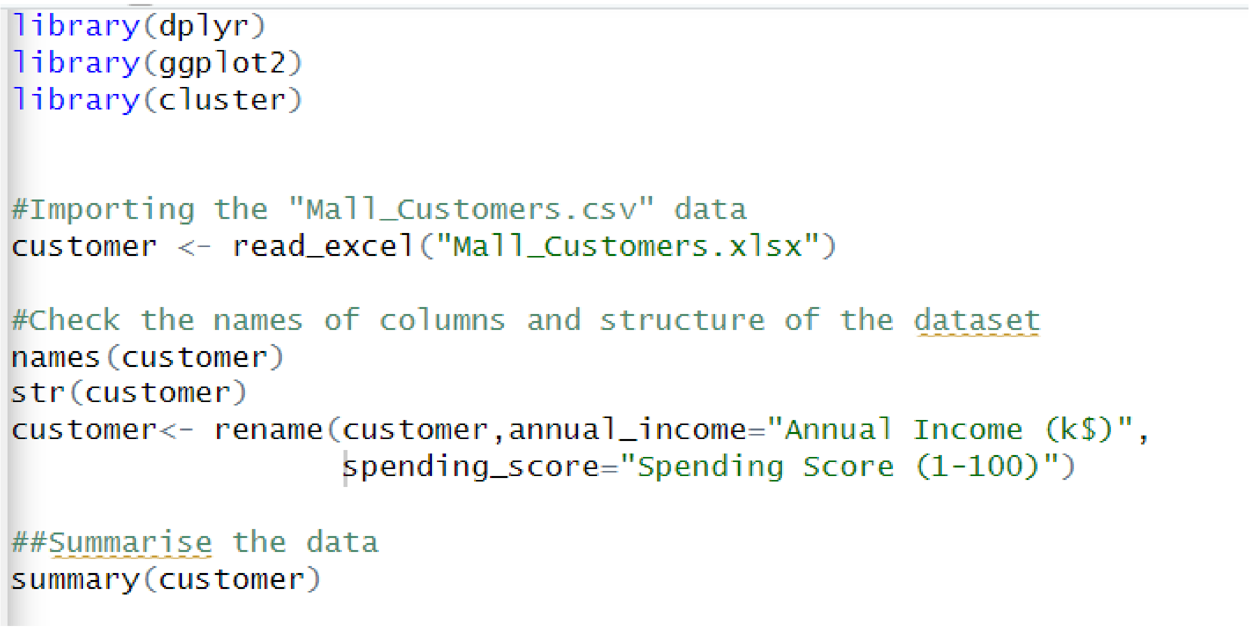
```
#Create a plot of the customers segments
ggplot(customer, aes(x = annual_income , y = spending_score)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  scale_color_discrete(name = " ",
                       breaks=c("1", "2", "3", "4", "5","6"),
                       labels=c("Cluster 1", "Cluster 2", "Cluster 3",
                                "Cluster 4", "Cluster 5","Cluster 6")) +
  ggtitle("Segments of Mall Customers",
          subtitle = "Using K-means Clustering")
```



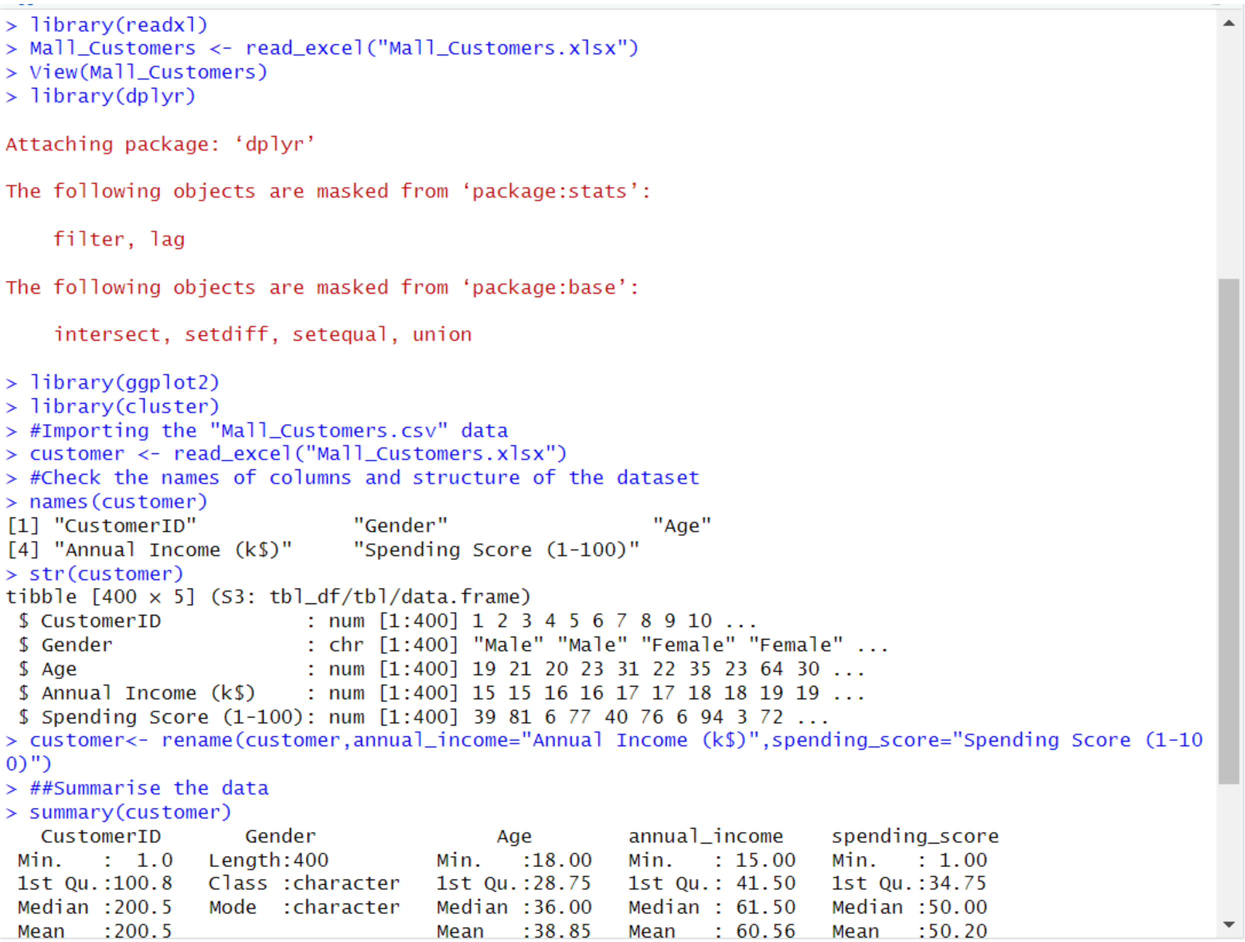
From the above visualization, we observe that there is a distribution of 6 clusters as follows –

- 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.

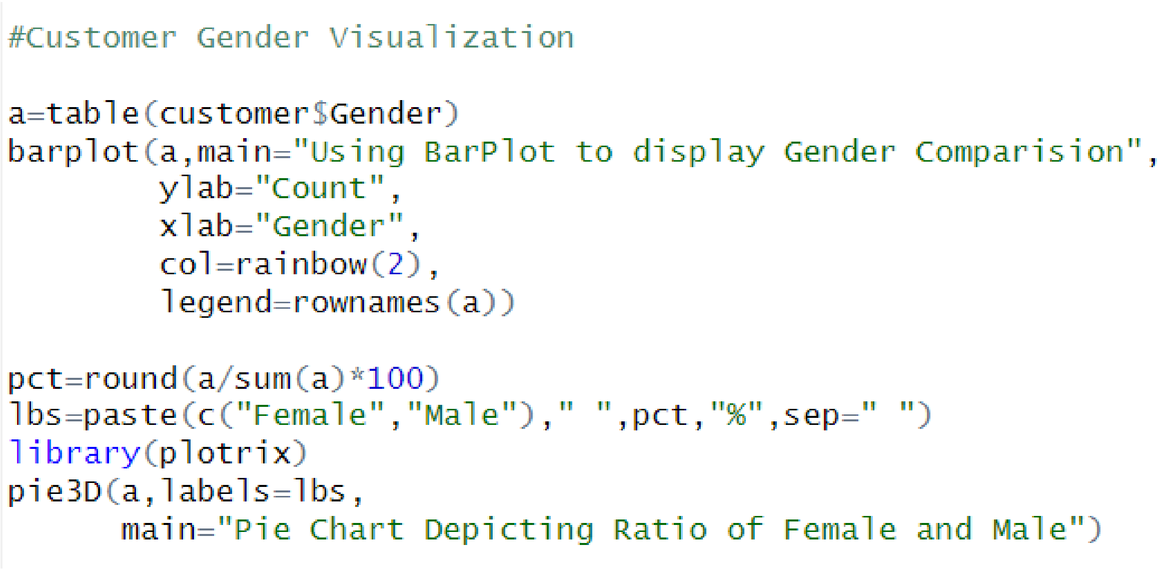
Source Code and its Outputs



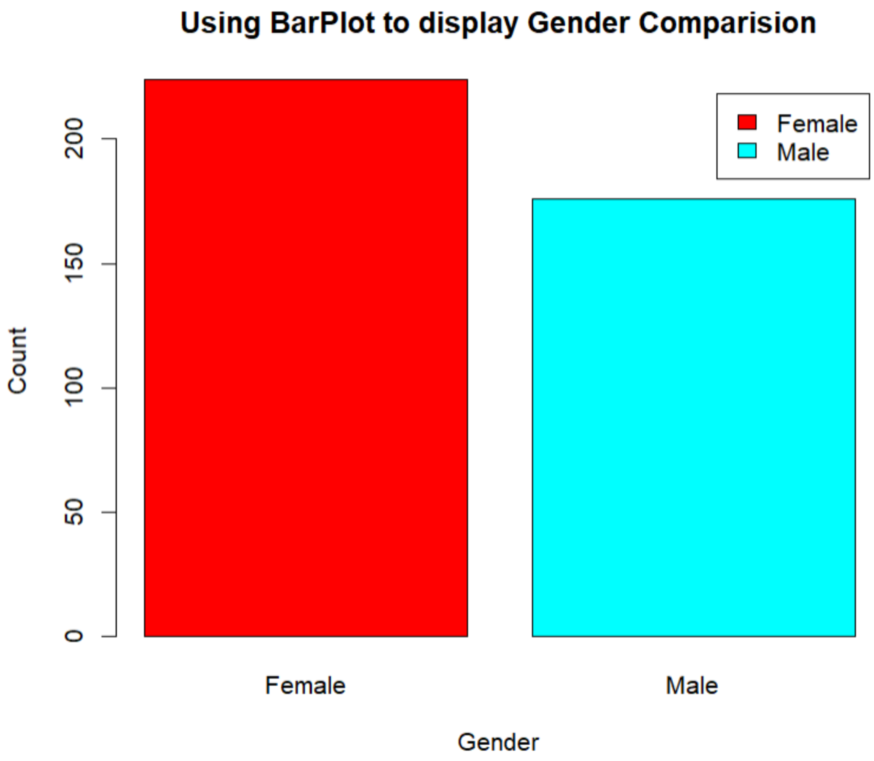
```r
library(dplyr)
library(ggplot2)
library(cluster)


#Importing the "Mall_Customers.csv" data
customer <- read_excel("Mall_Customers.xlsx")

#Check the names of columns and structure of the dataset
names(customer)
str(customer)
customer<- rename(customer,annual_income="Annual Income (k$)",
                  spending_score="Spending Score (1-100)")

##Summarise the data
summary(customer)
```



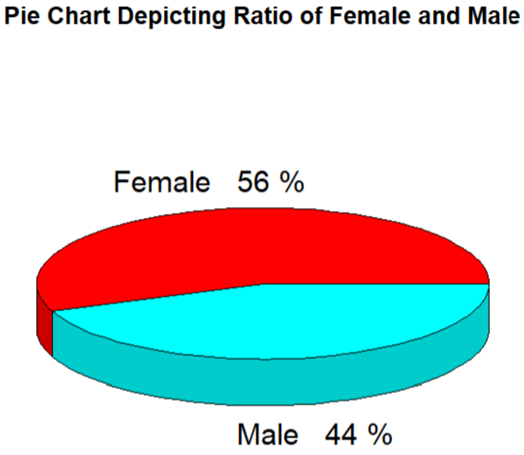
```
> library(readxl)
> Mall_Customers <- read_excel("Mall_Customers.xlsx")
> View(Mall_Customers)
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> library(ggplot2)
> library(cluster)
> #Importing the "Mall_Customers.csv" data
> customer <- read_excel("Mall_Customers.xlsx")
> #Check the names of columns and structure of the dataset
> names(customer)
[1] "CustomerID"             "Gender"                 "Age"
[4] "Annual Income (k$)"     "Spending Score (1-100)"
> str(customer)
tibble [400 x 5] (S3: tbl_df/tbl/data.frame)
 $ CustomerID            : num [1:400] 1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : chr [1:400] "Male" "Male" "Female" "Female" ...
 $ Age                   : num [1:400] 19 21 20 23 31 22 35 23 64 30 ...
 $ Annual Income (k$)    : num [1:400] 15 15 16 16 17 17 18 18 19 19 ...
 $ Spending Score (1-100): num [1:400] 39 81 6 77 40 76 6 94 3 72 ...
> customer<- rename(customer,annual_income="Annual Income (k$)",spending_score="Spending Score (1-10
0)")
> ##Summarise the data
> summary(customer)
   CustomerID        Gender               Age         annual_income    spending_score
 Min.   :  1.0   Length:400         Min.   :18.00   Min.   : 15.00   Min.   : 1.00
 1st Qu.:100.8   Class :character   1st Qu.:28.75   1st Qu.: 41.50   1st Qu.:34.75
 Median :200.5   Mode  :character   Median :36.00   Median : 61.50   Median :50.00
 Mean   :200.5                      Mean   :38.85   Mean   : 60.56   Mean   :50.20
```



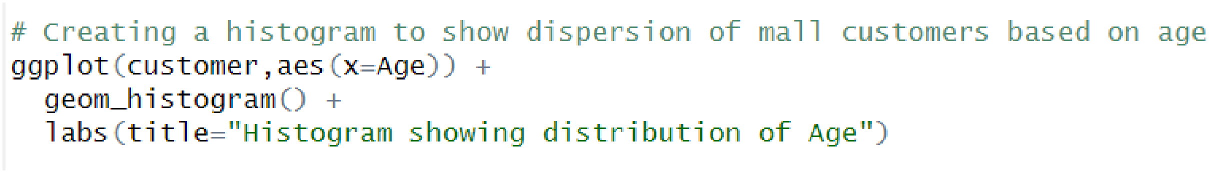
```
#Customer Gender Visualization

a=table(customer$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))

pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
library(plotrix)
pie3D(a,labels=lbs,
      main="Pie Chart Depicting Ratio of Female and Male")
```



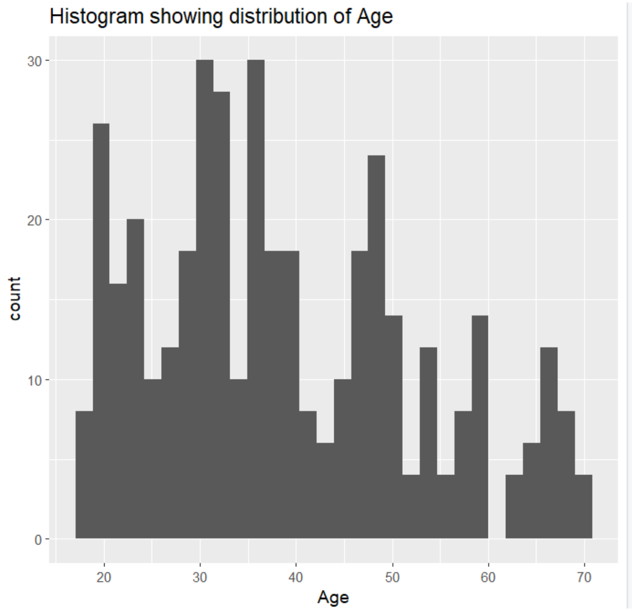
Using BarPlot to display Gender Comparision

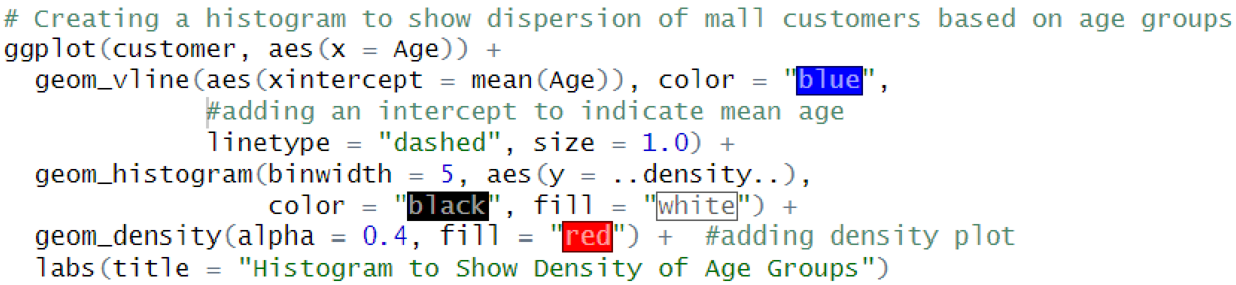


Pie Chart Depicting Ratio of Female and Male

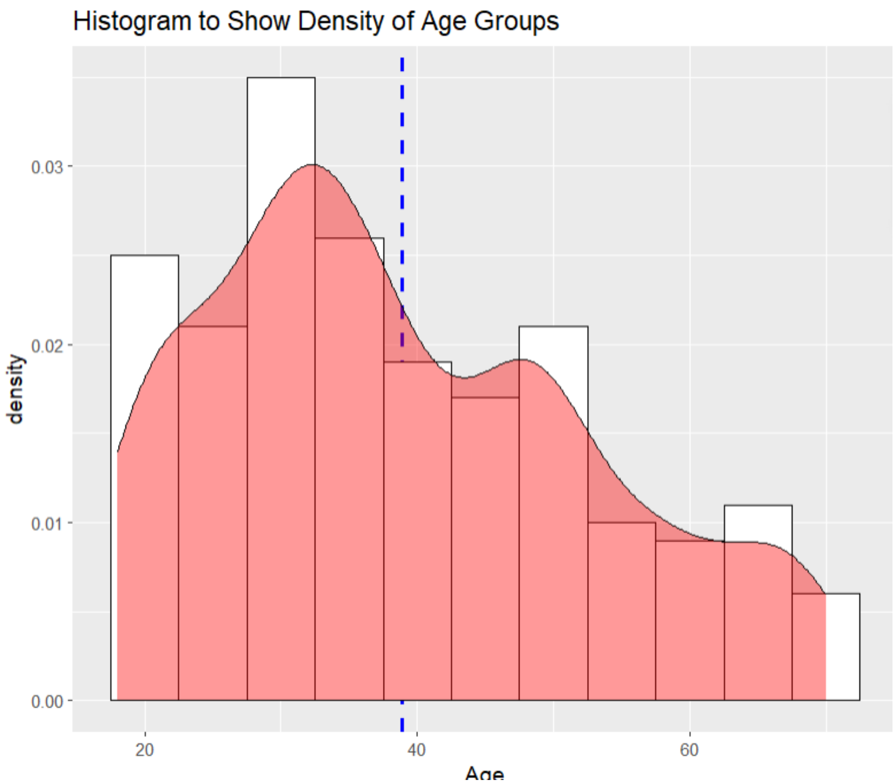


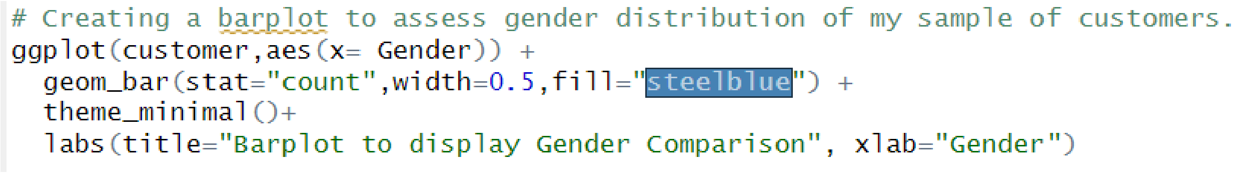
```
# Creating a histogram to show dispersion of mall customers based on age
ggplot(customer,aes(x=Age)) +
  geom_histogram() +
  labs(title="Histogram showing distribution of Age")
```



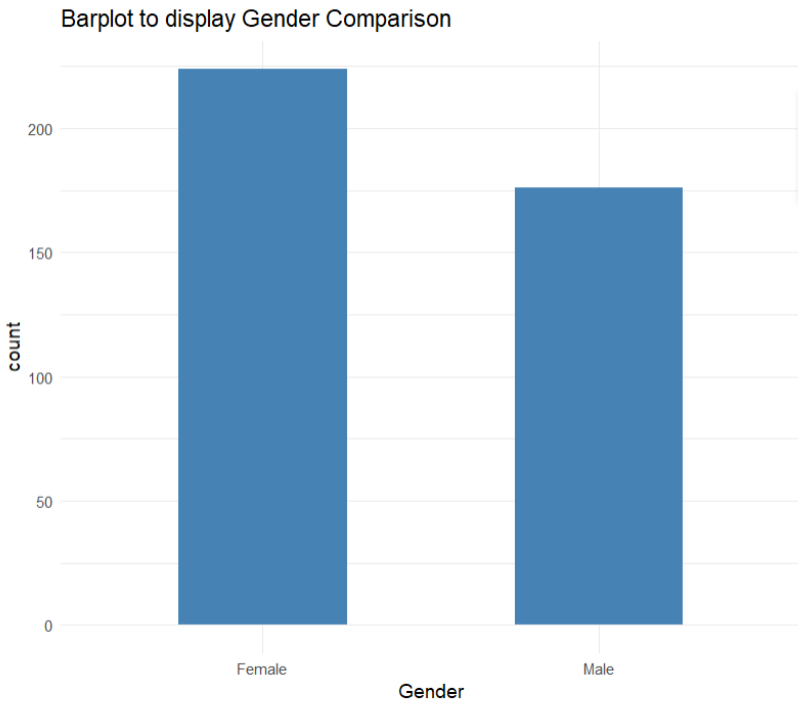


```
# Creating a histogram to show dispersion of mall customers based on age groups
ggplot(customer, aes(x = Age)) +
  geom_vline(aes(xintercept = mean(Age)), color = "blue",
             #adding an intercept to indicate mean age
             linetype = "dashed", size = 1.0) +
  geom_histogram(binwidth = 5, aes(y = ..density..),
                 color = "black", fill = "white") +
  geom_density(alpha = 0.4, fill = "red") +  #adding density plot
  labs(title = "Histogram to Show Density of Age Groups")
```

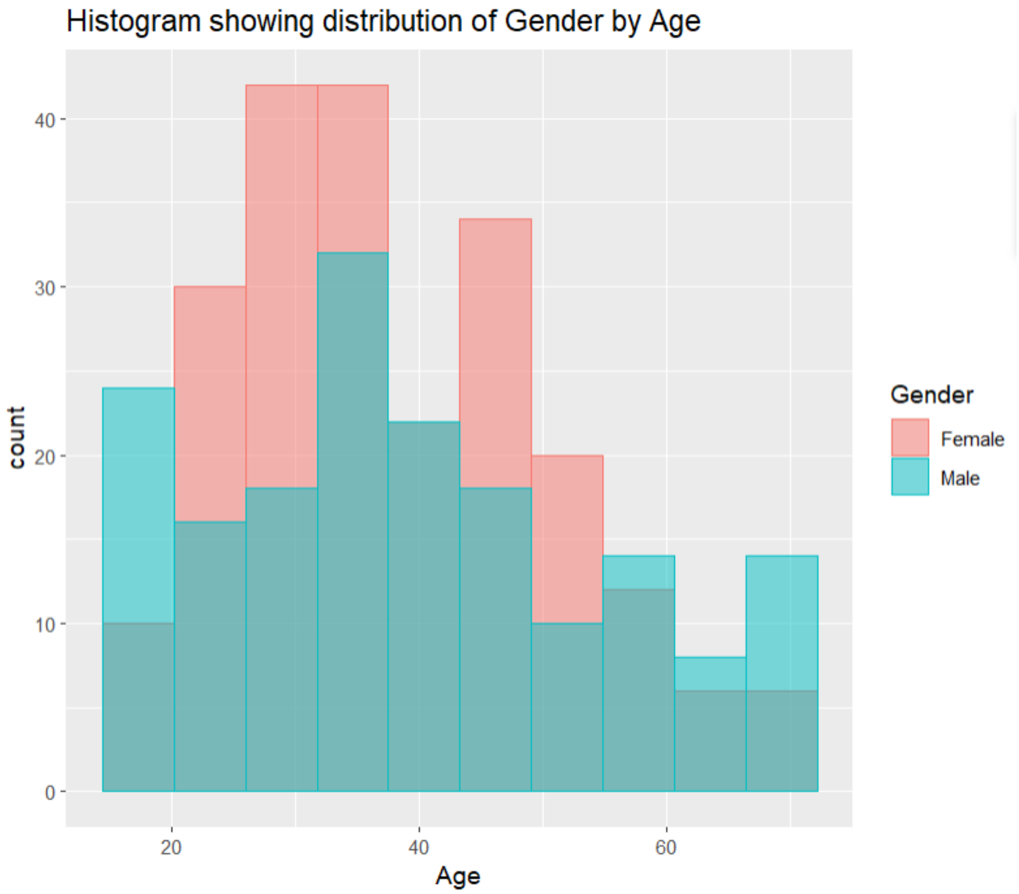


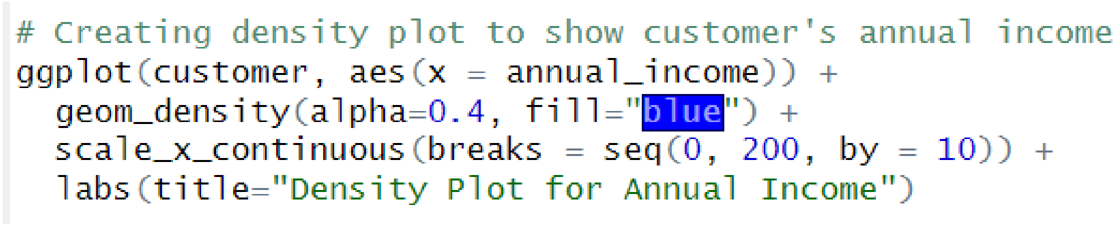


```
# Creating a barplot to assess gender distribution of my sample of customers.
ggplot(customer,aes(x= Gender)) +
  geom_bar(stat="count",width=0.5,fill="steelblue") +
  theme_minimal()+
  labs(title="Barplot to display Gender Comparison", xlab="Gender")
```

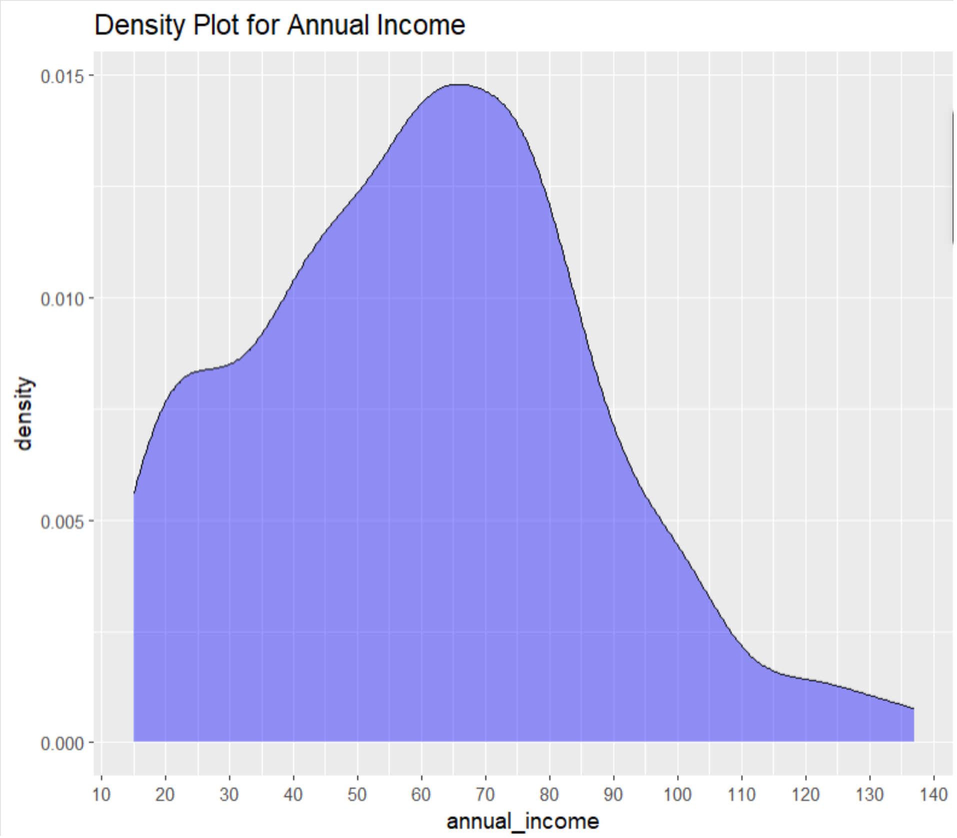


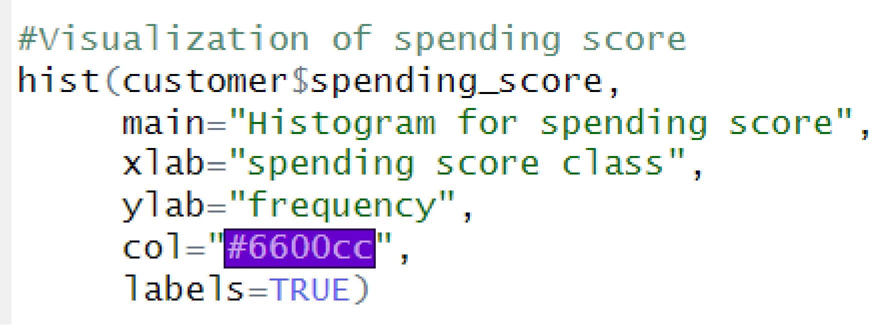
```
## 3.3: Create a histogram for the variable "Age" by Gender
ggplot(customer,aes(x=Age, fill=Gender, color=Gender))+
  geom_histogram(bins = 10, position = "identity",alpha=0.5) +
  labs(title="Histogram showing distribution of Gender by Age")
```





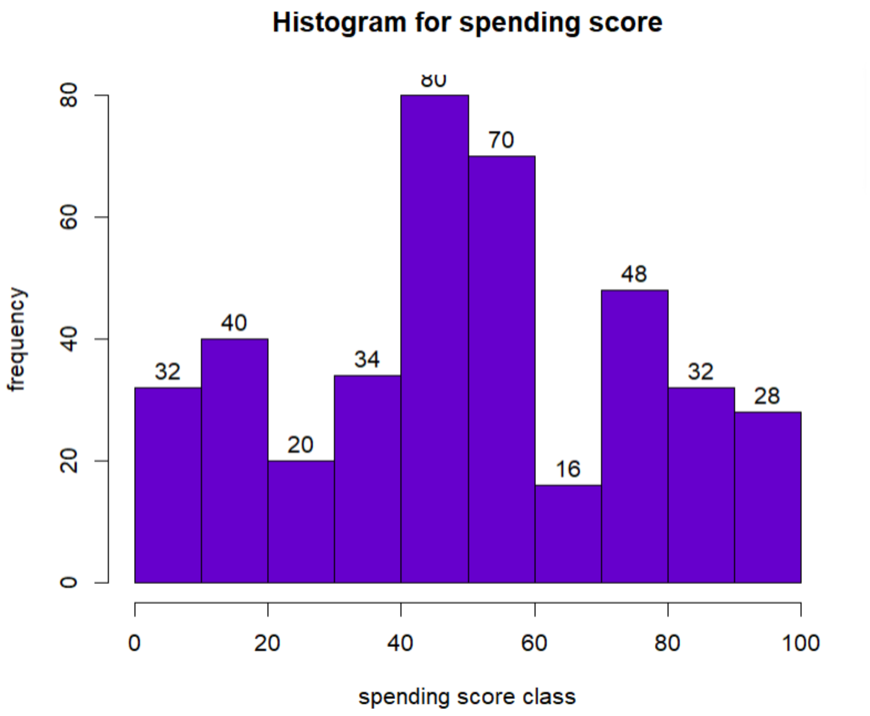
```
# Creating density plot to show customer's annual income
ggplot(customer, aes(x = annual_income)) +
  geom_density(alpha=0.4, fill="blue") +
  scale_x_continuous(breaks = seq(0, 200, by = 10)) +
  labs(title="Density Plot for Annual Income")
```

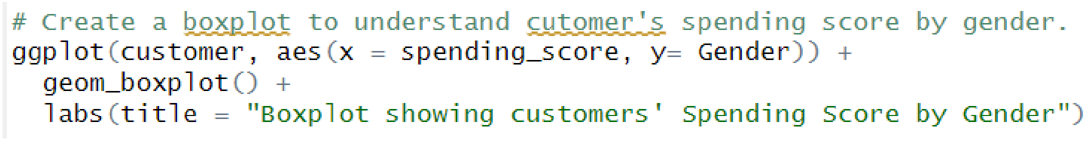




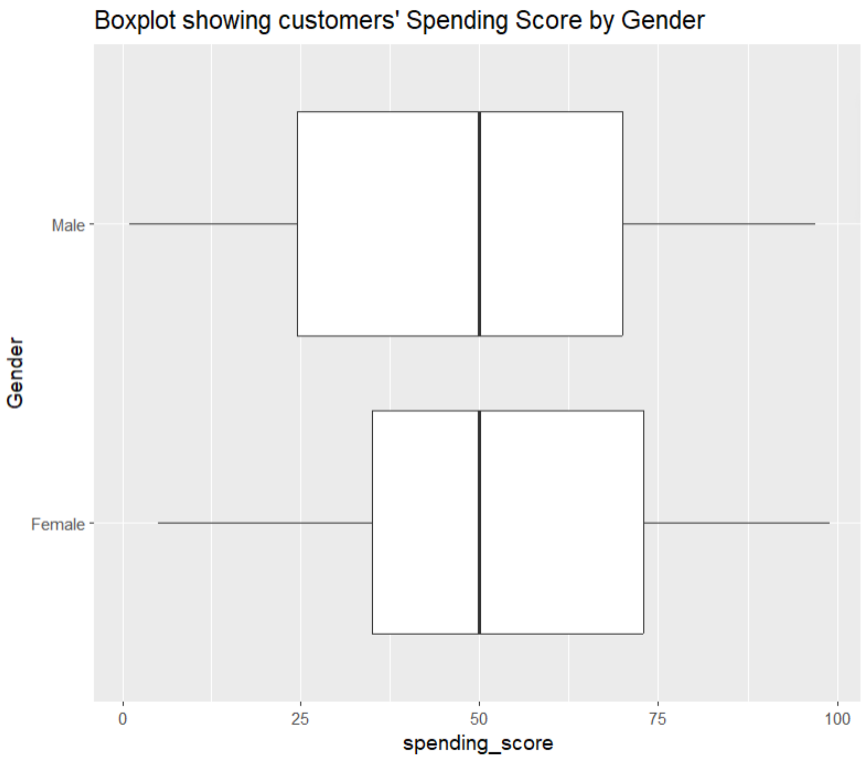
```
#Visualization of spending score
hist(customer$spending_score,
     main="Histogram for spending score",
     xlab="spending score class",
     ylab="frequency",
     col="#6600cc",
     labels=TRUE)
```

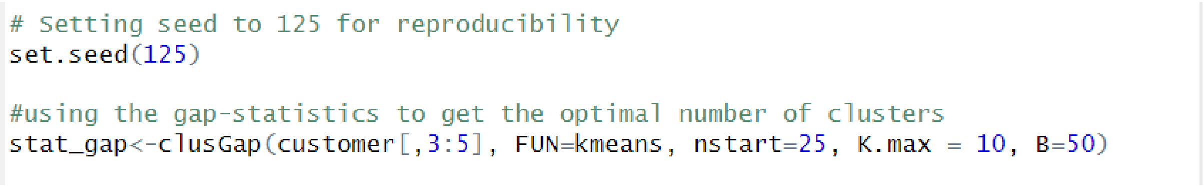
\



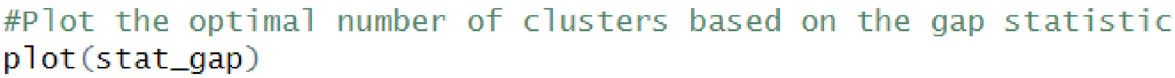


```
# Create a boxplot to understand cutomer's spending score by gender.
ggplot(customer, aes(x = spending_score, y= Gender)) +
  geom_boxplot() +
  labs(title = "Boxplot showing customers' Spending Score by Gender")
```

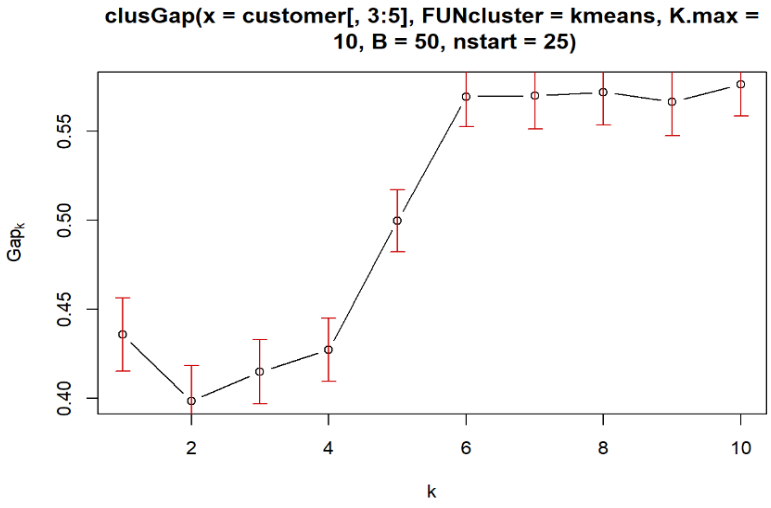


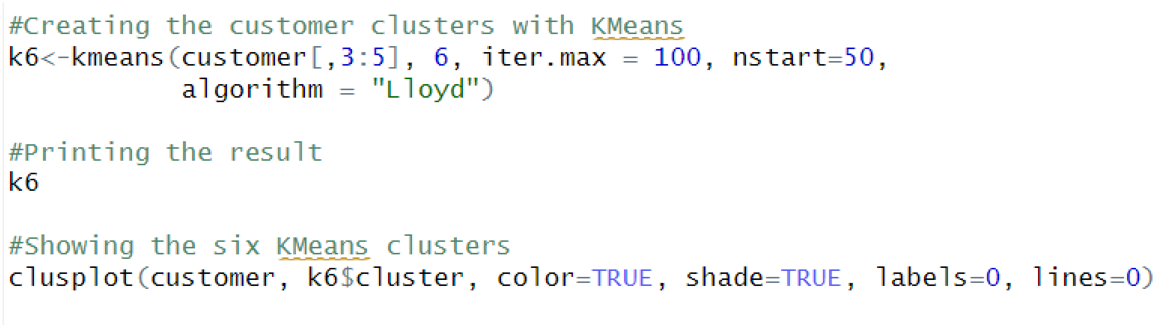


```
# Setting seed to 125 for reproducibility
set.seed(125)

#using the gap-statistics to get the optimal number of clusters
stat_gap<-clusGap(customer[,3:5], FUN=kmeans, nstart=25, K.max = 10, B=50)
```

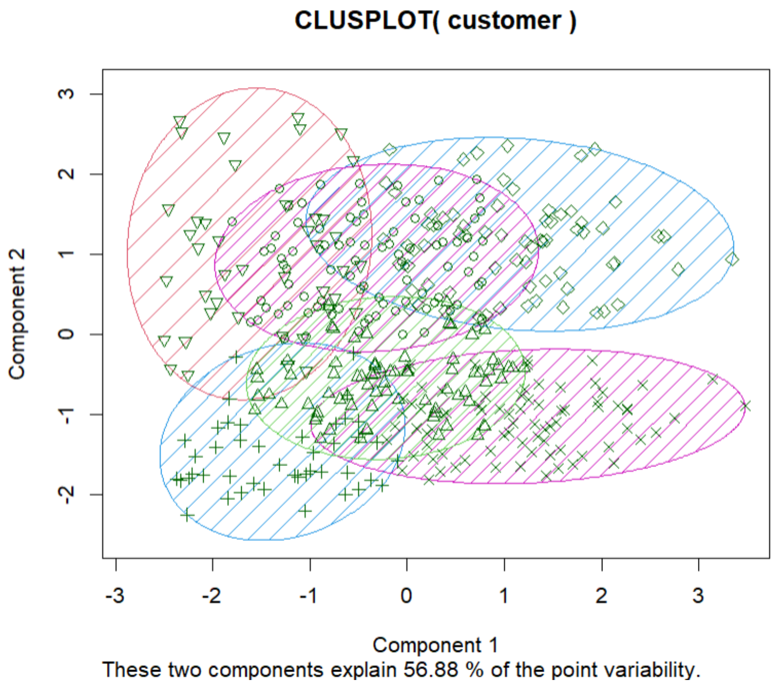


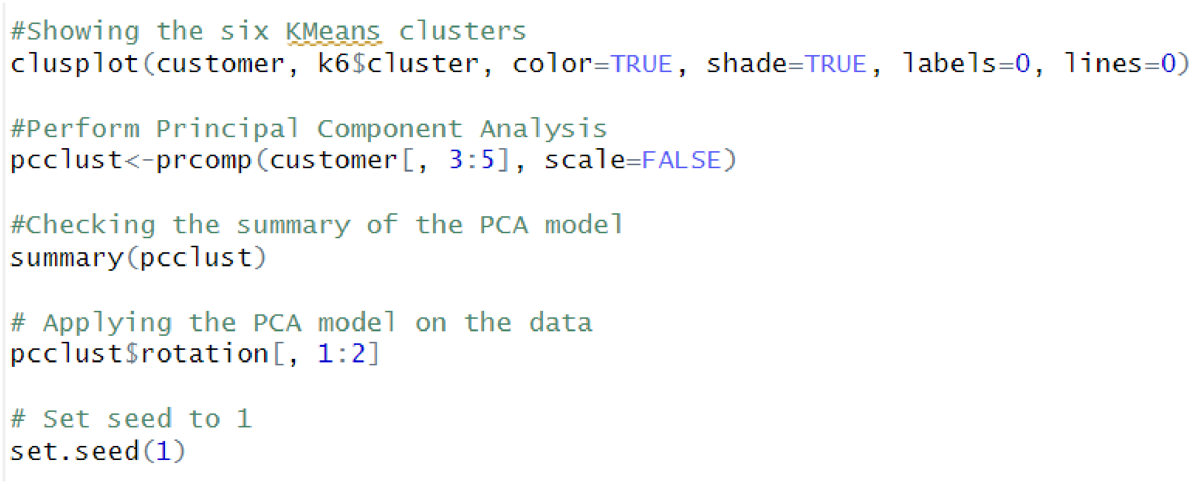
```
#Plot the optimal number of clusters based on the gap statistic
plot(stat_gap)
```



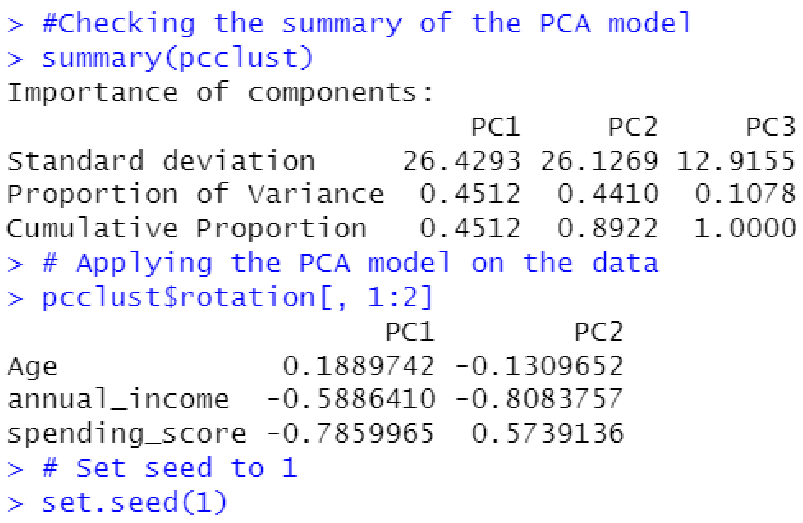


```
#Creating the customer clusters with KMeans
k6<-kmeans(customer[,3:5], 6, iter.max = 100, nstart=50,
           algorithm = "Lloyd")

#Printing the result
k6

#Showing the six KMeans clusters
clusplot(customer, k6$cluster, color=TRUE, shade=TRUE, labels=0, lines=0)
```

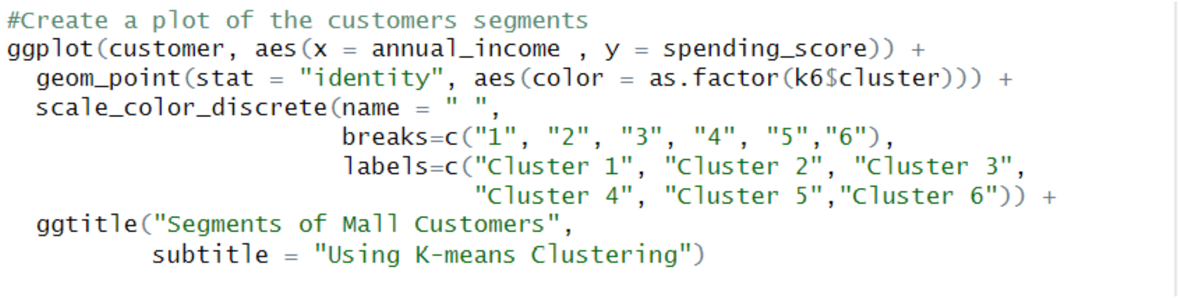




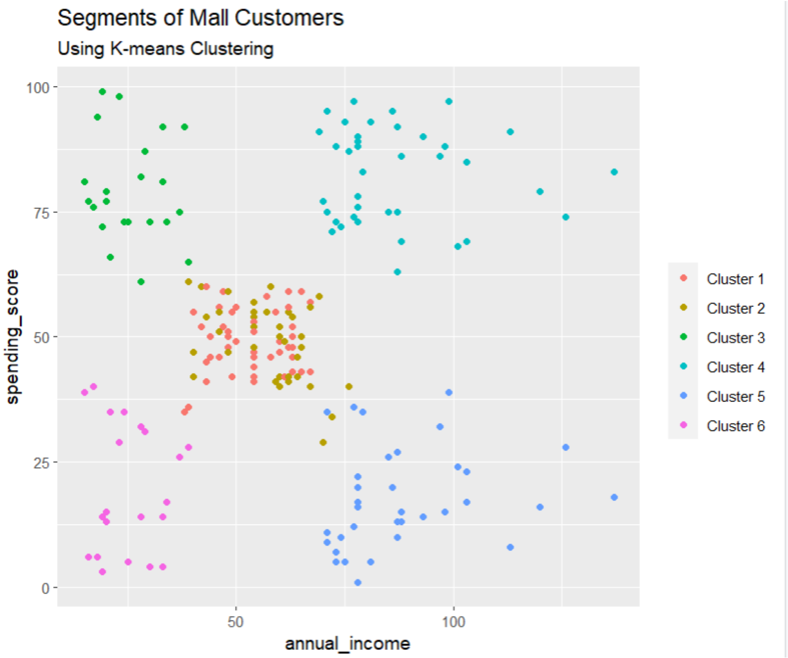
```
#Showing the six KMeans clusters
clusplot(customer, k6$cluster, color=TRUE, shade=TRUE, labels=0, lines=0)

#Perform Principal Component Analysis
pcclust<-prcomp(customer[, 3:5], scale=FALSE)

#Checking the summary of the PCA model
summary(pcclust)

# Applying the PCA model on the data
pcclust$rotation[, 1:2]

# Set seed to 1
set.seed(1)
```

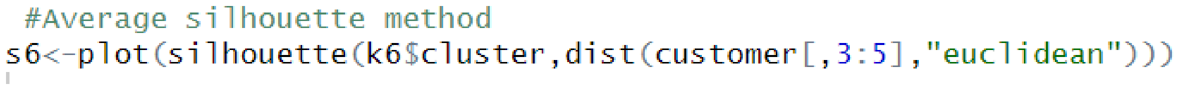


```
> #Checking the summary of the PCA model
> summary(pcclust)
Importance of components:
                           PC1     PC2     PC3
Standard deviation     26.4293 26.1269 12.9155
Proportion of Variance  0.4512  0.4410  0.1078
Cumulative Proportion   0.4512  0.8922  1.0000
> # Applying the PCA model on the data
> pcclust$rotation[, 1:2]
                      PC1        PC2
Age             0.1889742 -0.1309652
annual_income  -0.5886410 -0.8083757
spending_score -0.7859965  0.5739136
> # Set seed to 1
> set.seed(1)
```

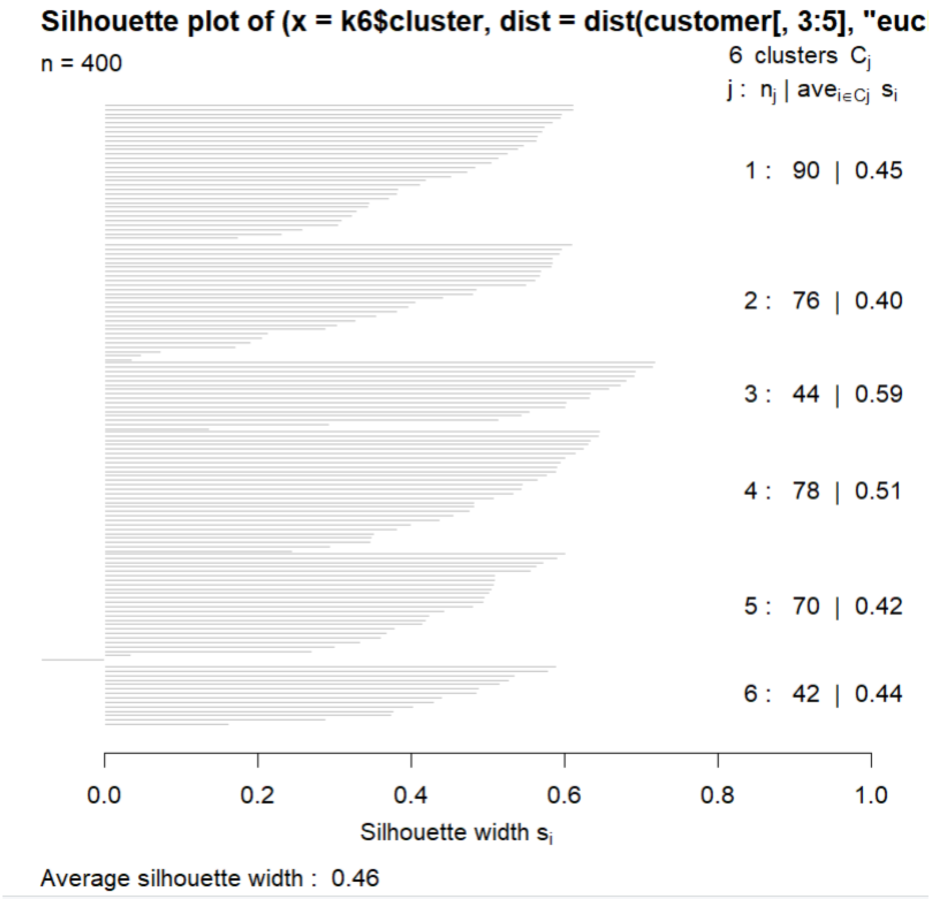


```
#Create a plot of the customers segments
ggplot(customer, aes(x = annual_income , y = spending_score)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  scale_color_discrete(name = " ",
                           breaks=c("1", "2", "3", "4", "5","6"),
                           labels=c("Cluster 1", "Cluster 2", "Cluster 3",
                                    "Cluster 4", "Cluster 5","Cluster 6")) +
  ggtitle("Segments of Mall Customers",
          subtitle = "Using K-means Clustering")
```





```
 #Average silhouette method
s6<-plot(silhouette(k6$cluster,dist(customer[,3:5],"euclidean")))
```



ACCURACY LEVEL

The accuracy of a data science project is not a universal metric that can be applied to all projects. The concept of accuracy depends on the specific task or problem being addressed in the project. Accuracy is typically used in the context of supervised learning tasks, where the goal is to predict a target variable based on input features. In such cases, accuracy refers to the proportion of correctly predicted instances (or the accuracy rate) out of the total number of instances in the dataset.

However, many data science projects involve tasks beyond simple classification or prediction. For example, customer segmentation, is an unsupervised learning task where the goal is to group similar customers together without specific target labels. In such cases, accuracy is not the appropriate metric for evaluating the quality of the segmentation since there are no predefined correct labels to compare against.

Average Silhouette Method

With the help of the average silhouette method, we can measure the quality of our clustering operation. With this, we can determine how well 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.

Using the silhouette function in the cluster package, we can compute the average silhouette width using the k mean function. Here, the optimal cluster will possess the highest average.

AREA OF IMPROVEMENT

Determining areas of improvement in a k-means clustering project can depend on various factors, such as the specific dataset, the objectives of the project, and the evaluation metrics used. However, here are some common areas to consider for potential improvement in k-means clustering:

1. Initialization:
The initial selection of cluster centroids can affect the final clustering results. The standard k-means algorithm randomly initializes centroids, which may lead to

suboptimal solutions. Exploring alternative initialization methods, such as k-means++, can improve the quality and stability of the clustering.

2. Outlier Handling:
K-means clustering is sensitive to outliers, as they can significantly impact the centroid calculation and clustering assignments. Preprocessing techniques, such as outlier detection and removal or using robust distance metrics, can help mitigate the influence of outliers on the clustering process.

3. Determining the Optimal Number of Clusters:
Choosing the appropriate number of clusters is critical. While domain knowledge and evaluation metrics like the elbow method and silhouette coefficient can provide guidance, exploring alternative methods, such as hierarchical clustering or density-based clustering algorithms, might help identify the optimal number of clusters.

4. Feature Selection and Scaling:
Consider the impact of feature selection and scaling on clustering results. Some features may have more significant contributions to the clustering process than others. Applying feature selection techniques or feature scaling methods, such as normalization or standardization, can enhance the performance and interpretability of the clustering results.

5. Evaluation Metrics:
Besides traditional metrics like within-cluster sum of squares (WCSS), explore other evaluation metrics specific to your problem domain. For example, silhouette analysis, Dunn index, or adjusted Rand index can provide additional insights into the quality and coherence of the clusters.

6. Alternative Clustering Algorithms:
K-means clustering is a popular method, but it may not always be the most suitable for all datasets or objectives. Exploring other clustering algorithms, such as DBSCAN, Agglomerative Clustering, or Gaussian Mixture Models, can offer different perspectives and potentially improve clustering results.

The areas of improvement may vary depending on the specifics of your project and dataset. It's crucial to analyze the data, understand the problem, and experiment with various techniques to identify the most effective improvements for your specific k-means clustering project.

PROJECTED OUTCOME

1. **Clear customer segments**: The project will identify distinct groups or clusters of customers based on their similarities in terms of demographics, behavior, preferences, or other relevant factors. These segments help in understanding the diversity within the customer base.
2. **Actionable insights**: Customer segmentation provides valuable insights into the different needs, preferences, and behaviors of each segment. This knowledge can be used to develop targeted marketing strategies, personalized recommendations, or tailored product offerings for each segment.
3. **Improved marketing effectiveness**: With well-defined customer segments, you can optimize your marketing efforts by delivering more targeted and relevant messages to each segment. This leads to improved customer engagement, response rates, and overall marketing effectiveness.

4. **Enhanced customer experience**: By understanding different customer segments, you can identify specific pain points or areas for improvement for each segment. This allows you to tailor your products, services, and customer support to meet the unique needs of each segment, ultimately enhancing the overall customer experience.
5. **Increased customer retention**: Customer segmentation helps in identifying high-value customer segments that are more likely to be loyal and have a higher lifetime value. By focusing on these segments and developing retention strategies specific to their needs, you can improve customer retention rates and customer loyalty.
6. **Targeted acquisition strategies**: Customer segmentation also helps in identifying potential target segments that are currently underserved or untapped. This enables you to develop targeted acquisition strategies to attract new customers from these segments, expanding your customer base and market reach.

   Business growth and profitability: Ultimately, the outcome of a customer segmentation project is to drive business growth and profitability. By understanding and catering to the unique needs of different customer segments, you can increase customer satisfaction, loyalty, and ultimately revenue and profitability

CONCLUSION

In this project, we went through the customer segmentation model. We developed this using a class of machine learning known as unsupervised learning. Specifically, we made use of a clustering algorithm called K-means clustering. We analyzed and visualized the data and then proceeded to implement our algorithm.Finding an optimal number of unique customer groups will help you understand how your customers differ, and help you give them exactly what they want. Customer segmentation improves customer

experience and boosts company revenue. That’s why segmentation is a must if you want to surpass your competitors and get more customers. Doing it with machine learning is definitely the right way to go .
Hope you enjoyed this customer segmentation project of machine learning using R.