CUSTOMER SEGMENTATION CLASSIFICATION USING R

R – PROGRAMMING (MINI PROJECT) REPORT

Submitted in partial fulfillment of

BACHELOR OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE & DATA SCIENCE

by

1. P. NAMRATHA	**(21BQ1A5441)**
2. B. JHANSI	**(21BQ1A5404)**
3. D. ARADHANA	**(21BQ1A5409)**
4. CH. MOUNIKA	**(21BQ1A5407)**

UNDER THE GUIDANCE OF

K. RAJANI, M. Tech, (Ph. D), Assoc Prof

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY
(Affiliated to JNTU Kakinada, Approved by AICTE, New Delhi)
Accredited by NAAC with 'A' Grade, Accredited by NBA
An ISO 9001:2015 Certified Institution, Nambur:522508, Andhra Pradesh, India

DECLARATION

I hereby declare that this project report **CUSTOMER SEGMENTATION CLASSIFICATION** done by **P. Namratha**. I affirm that this report has been completed as a Mini Project in R Programming in **Vasireddy Venkatadri Institute Of Technology**. I further declare that the information presented in this report is authentic, and any external sources used for reference have been duly acknowledged through appropriate citations and references. The ideas, concepts, and conclusions presented herein are the product of my own research and analysis.

Signature Of The Candidate

(P.Namratha)

CERTIFICATE

This is to certify that the project entitled **CUSTOMER SEGMENTATION CLASSIFICATION** is the bona-fide work carried out by **P. Namratha (21BQ1A5441)** in partial fulfilment of the requirements for the award of the degree Bachelor of Technology in Artificial Intelligence & Data Science from **Vasireddy Venkatadri Institute Of Technology** during the year 2022-23 under supervision and guidance of (Mrs. K. Rajani).

Signature Of The H.O.D	**Signature Of The Guide**
(T. Sudhir)	**(K. Rajani)**

ACKNOWLEDGEMENTS

We greatly indebted to Dr. T. Sudhir, Professor & Head of the Department, for providing motivation and valuable suggestions at every stage of our course of study.

It is our pleasure to express our deep and sincere gratitude to our mini project guide Mrs. K. Rajani, M. Tech, (Ph. D), Assoc Professor, for extending her sincere and heartfelt guidance throughout this project work.

We extend our sincere thanks to all other Teaching and Non-Teaching Faculty members of the department for their cooperation and encouragement throughout our course of study.

We affectionately acknowledge our friends for their motivation and suggestions which helps us in successfully completing our course.

We have no words to acknowledge the warm affection, constant inspiration and encouragement that we received from our Parents and Family Members.

P. Namratha	21BQ1A5441
B. Jhansi	21BQ1A5404
D. Aradhana	21BQ1A5409
Ch. Mounika	21BQ1A5407

CUSTOMER SEGMENTATION CLASSIFICATION

ABSTRACT

Customer segmentation classification refers to the process of categorizing customers into distinct groups based on certain characteristics or behaviours. The goal of customer segmentation is to divide a large customer base into smaller, more homogeneous groups, allowing businesses to better understand and target their customers' needs, preferences, and behaviours.

Classification, in the context of customer segmentation, involves using machine learning algorithms or statistical techniques to assign customers to predefined segments or clusters. These algorithms use a set of input variables or features, such as demographic information, purchasing history, browsing behaviour, or psychographic data, to determine the appropriate segment for each customer.

The process typically involves the following steps:

1. Data collection: Gathering relevant customer data from various sources, such as transaction records, surveys, website analytics, or social media.

2. Data pre-processing: Cleaning and transforming the data to ensure consistency and usability. This may include handling missing values, standardizing variables, or normalizing data.

3. Feature selection: Identifying the most relevant features that can effectively differentiate customers and contribute to meaningful segmentation.

4. Model training: Applying classification algorithms, such as decision trees, k-means clustering, logistic regression, or support vector machines, to build a predictive model based on the selected features.

5. Model evaluation: Assessing the performance of the classification model by using appropriate metrics, such as accuracy, precision, recall, or F1 score, to measure how well the model predicts the segment membership of customers.

6. Segment assignment: Using the trained model to assign customers to specific segments based on their input features. Each customer is classified into the segment that best represents their characteristics or behaviours.

Once customer segmentation classification is complete, businesses can tailor their marketing strategies, product offerings, pricing, and communication channels to each segment's specific needs and preferences. This targeted approach can lead to more effective customer acquisition, retention, and overall customer satisfaction.

APPLICATIONS

Customer segmentation classification has various applications across industries. Here are some common applications:

1. Targeted Marketing: By segmenting customers based on their demographics, behaviours, and preferences, businesses can create personalized marketing campaigns that are tailored to specific customer segments. This allows for more effective targeting and messaging, resulting in higher conversion rates and improved return on investment (ROI).

2. Product Development: Customer segmentation can provide valuable insights into the needs, preferences, and pain points of different customer groups. Businesses can use this information to develop new products or enhance existing ones to better meet the requirements of specific segments, increasing customer satisfaction and loyalty.

3. Customer Retention: Segmentation can help identify customers who are at risk of churn or who have a higher likelihood of remaining loyal. By understanding the characteristics and behaviours of these segments, businesses can implement targeted retention strategies such as personalized offers, loyalty programs, or proactive customer support, reducing churn rates and improving customer retention.

4. Pricing Strategies: Different customer segments often have varying price sensitivities. By segmenting customers based on their willingness to pay and price preferences, businesses can develop pricing strategies that maximize revenue. This may involve offering different pricing tiers, discounts, or bundling options for specific customer segments.

5. Customer Service and Support: Segmentation can assist in providing tailored customer service experiences. By categorizing customers into segments, businesses can anticipate the specific needs and preferences of each group, allowing for more personalized interactions, quicker issue resolution, and improved customer satisfaction.

6. Cross-Selling and Upselling: Segmenting customers based on their purchasing behaviour and preferences enables businesses to identify cross-selling and upselling opportunities. By recommending relevant products or services based on the specific segment's interests and past purchases, businesses can increase the average transaction value and drive additional revenue.

7. Channel Optimization: Understanding the preferred communication and purchasing channels of different customer segments can help optimize marketing and sales efforts. By targeting specific channels preferred by each segment, businesses can increase engagement, improve customer experience, and drive conversions.

These are just a few examples of how customer segmentation classification can be applied. The specific applications may vary depending on the industry, business goals, and available data.

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 dataset.

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

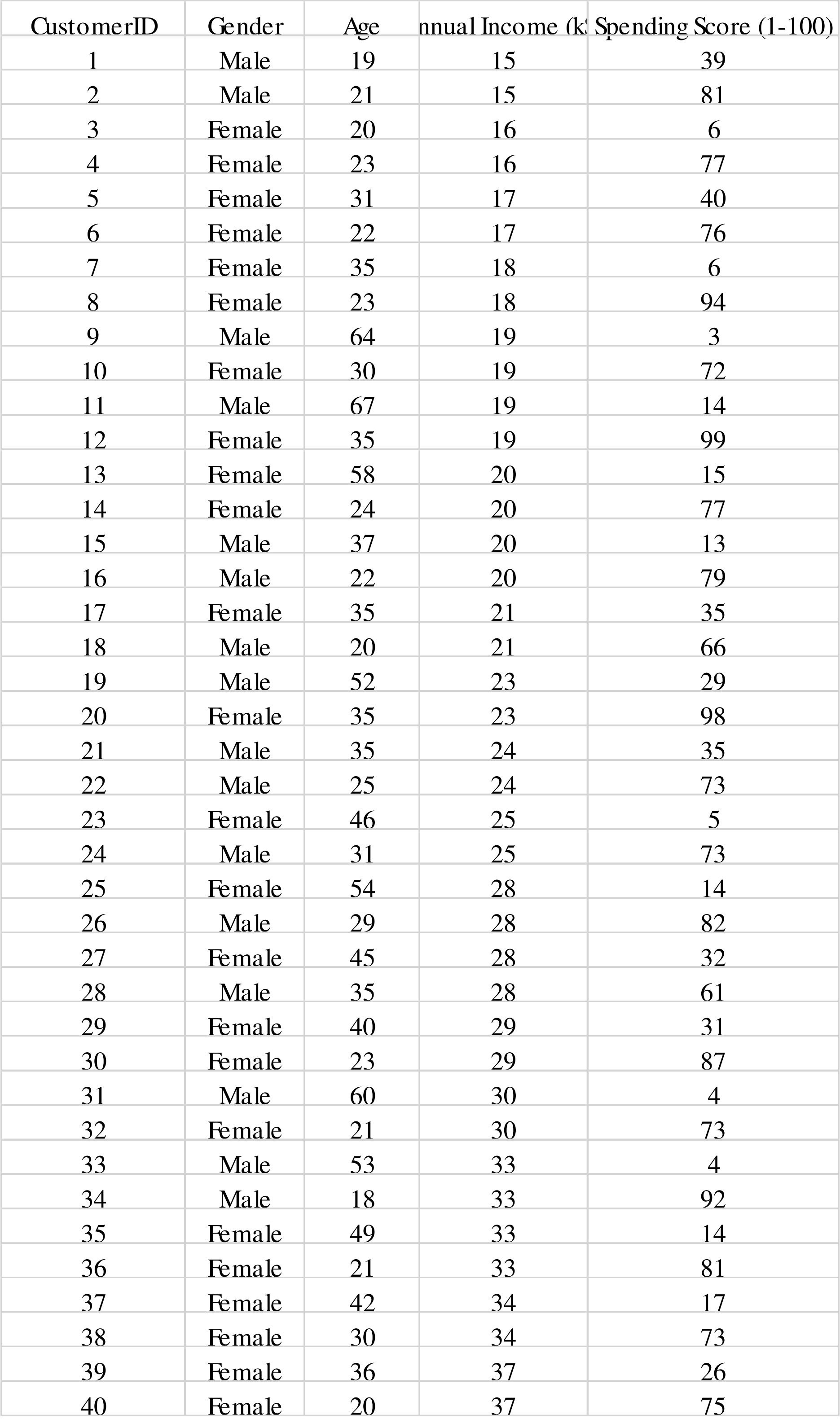
R PLATFORM: R STUDIO

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 modelling 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

DATA SET

Mall_Customers.csv



CustomerID	Gender	Age	nnual 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
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35
18	Male	20	21	66
19	Male	52	23	29
20	Female	35	23	98
21	Male	35	24	35
22	Male	25	24	73
23	Female	46	25	5
24	Male	31	25	73
25	Female	54	28	14
26	Male	29	28	82
27	Female	45	28	32
28	Male	35	28	61
29	Female	40	29	31
30	Female	23	29	87
31	Male	60	30	4
32	Female	21	30	73
33	Male	53	33	4
34	Male	18	33	92
35	Female	49	33	14
36	Female	21	33	81
37	Female	42	34	17
38	Female	30	34	73
39	Female	36	37	26
40	Female	20	37	75

This is a sample set from the original dataset of 400 entries

PACKAGES REQUIRED

- Plotrix

The Plotrix package in R is a collection of functions that extends the basic plotting capabilities of R. It provides a wide range of graphical tools and utilities to enhance data visualization and create more advanced plots. The package is designed to be flexible and customizable, allowing users to create complex plots with ease.

Here are some key features and functionalities of the Plotrix package:

Polar Plots
Bar Plots
Pie Charts
3D Plots
Plot Annotation
Color Management
Miscellaneous Functions

- Purr

The "purr" package in R is a powerful and popular package that provides a set of tools for working with and manipulating data in a concise and consistent manner. It is part of the "tidyverse" collection of packages, which are designed to enhance data manipulation, visualization, and analysis workflows.

Some key features and functions provided by the "purr" package include:

Mapping Functions
Modifying Functions
Combining Functions
Predicates

- Cluster

In R, the cluster package provides functions and algorithms for performing cluster analysis, which is a technique used to identify groups or clusters of similar objects or observations within a dataset. Cluster analysis is commonly used in various fields, including data mining, pattern recognition, and machine learning.

The cluster package offers several clustering methods and related functionalities, such as:

K-means clustering
Hierarchical clustering
Model-based clustering
Evaluation and visualization

- ## gridExtra

The gridExtra package in R is a powerful tool that provides functions for arranging multiple grid-based plots or graphical objects on a single page or within a single plotting region. It is commonly used in conjunction with other plotting packages like ggplot2 or lattice to create complex and customized layouts.

Here are some key features and functionalities of the gridExtra package:

Arranging Plots
Combining Plots
Customizing Layouts
Exporting Grid

- ## grid

The grid package is a fundamental graphics package in the R programming language. It provides a powerful system for creating and customizing high-quality plots and graphics. The grid package is part of the core R distribution, so you don't need to install any additional packages to use it.

Here are some key features and concepts associated with the grid package:
Grid Graphics System
Modular Structure
Hierarchical Layout
Fine-grained Control
Device Independence
Integration with Other Packages

- ## nbClust

The nbClust package in R is a useful tool for determining the optimal number of clusters in a dataset. It provides a set of functions and algorithms to perform cluster analysis and assess the number of clusters that best fit the data

- ## Factoextra

The factoextra package is a popular R package used for extracting and visualizing information from multivariate data analysis models, particularly those obtained from factor analysis, principal component analysis (PCA), and clustering algorithms. It provides a convenient set of functions to assist in interpreting the results of these analyses.

Here are some key features and functionalities of the factoextra package:
Visualizations
Factor analysis
Principal component analysis (PCA)
Clustering
Data extraction

- ggplot2

  ggplot2 is a widely used data visualization package in R, designed to create attractive and informative graphs and charts. It is built on the grammar of graphics concept, which allows users to express complex visualizations in a simple and consistent manner.

  Here are some key features and concepts related to ggplot2:

  Grammar of graphics
  Data structure
  Aesthetics
  Geometric objects
  Scales
  Themes
  Layering
  Faceting

- dplyr

  dplyr is a popular R package that provides a set of functions for data manipulation and transformation. It is designed to make working with data frames and data tables easier and more intuitive. The package follows a consistent and efficient grammar of data manipulation, allowing you to express complex data manipulations using a few simple verbs.

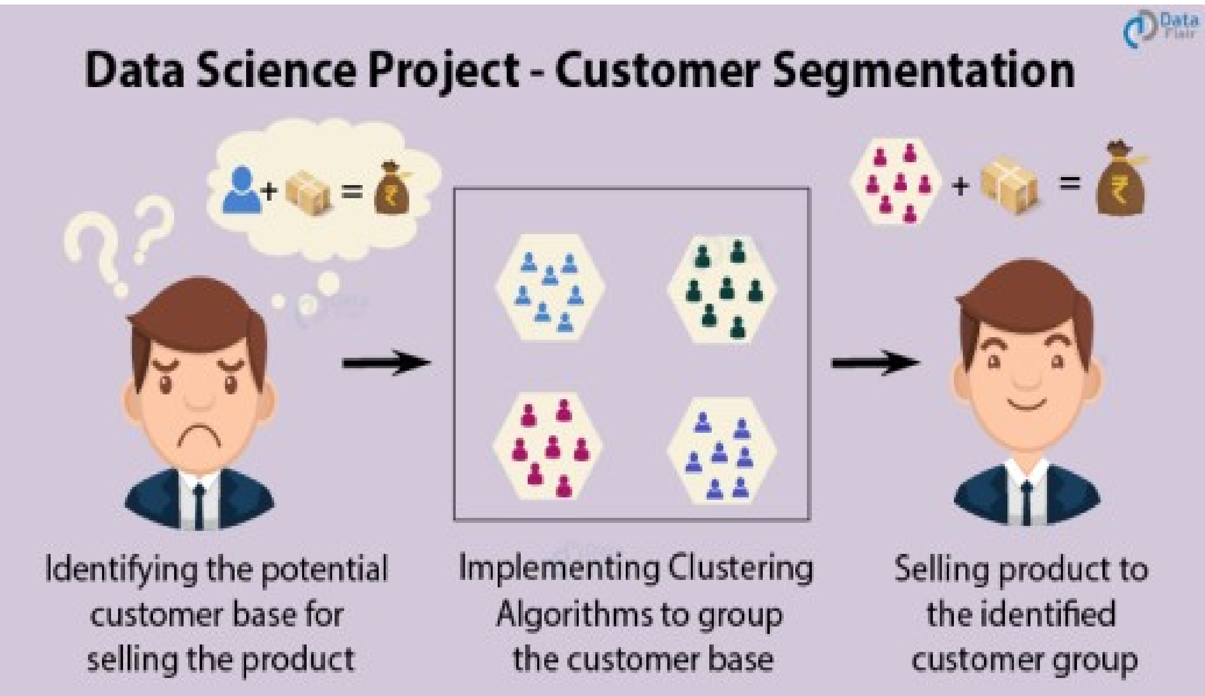
  Here are some key features and functions provided by dplyr:
  Selecting columns
  Filtering rows
  Mutating data
  Arranging data
  Summarizing data
  Grouping data
  Joining data frames

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 behavioural patterns play a crucial role in determining the company direction towards addressing the various segments.


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

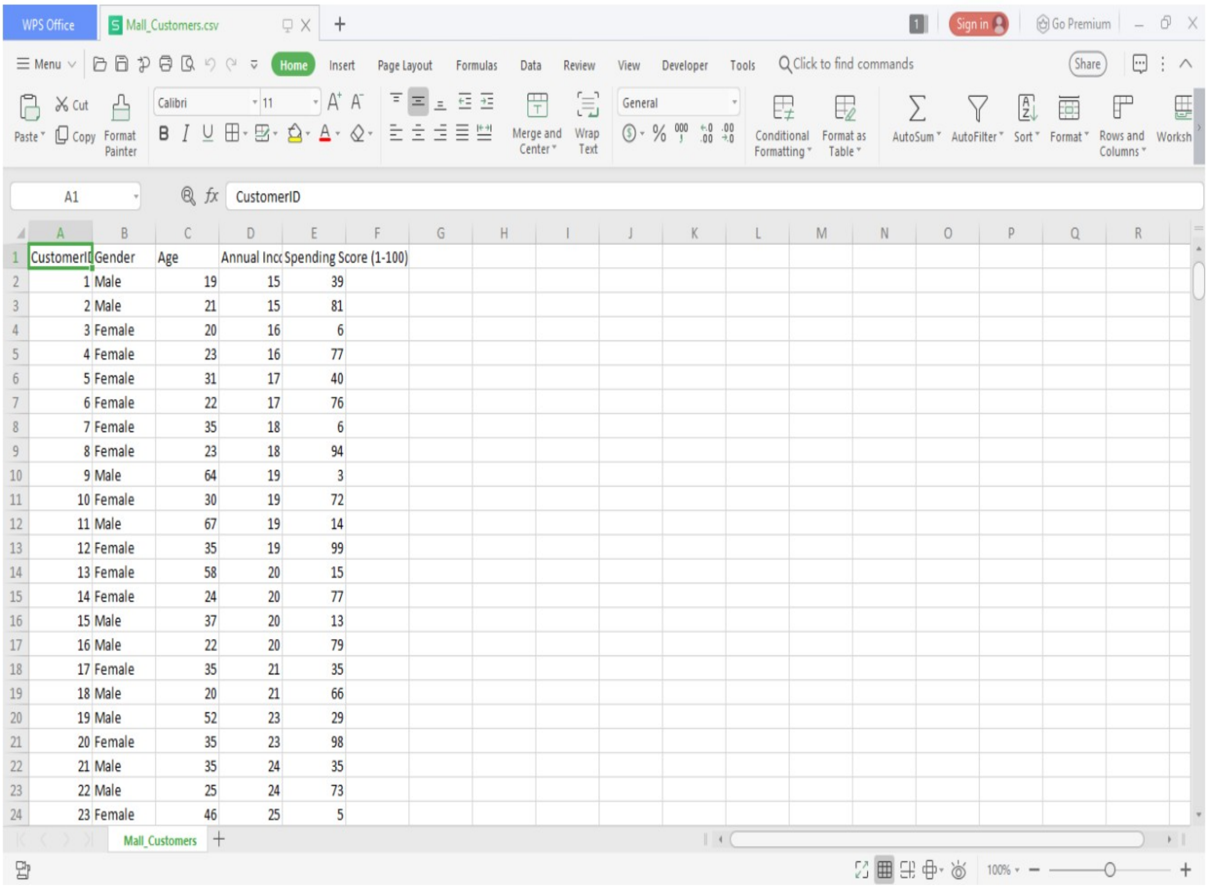
IMPLEMENTATION:

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.

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 dataset named as Mall_Customers.csv. This dataset contains 400 records of various type of customers. The events saved in dataset are unstructured. To perform analysis, reading of data set is done using command "read.csv".

```
customer_data=read.csv("C:/home/desktop/Mall_Customers.csv")
```



A1 | 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

Figure 1. Mall_Customers.csv

Customer Gender Visualization:

In this, we will create a bar plot and a pie chart 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 bar chart each of the bars can be given different colors

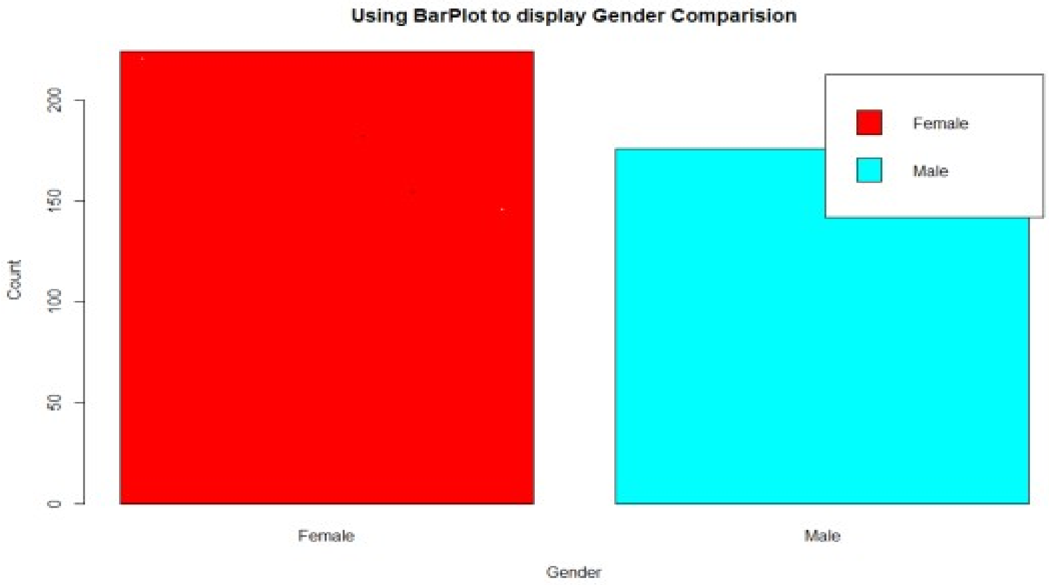


Figure - 2 Gender Comparison

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

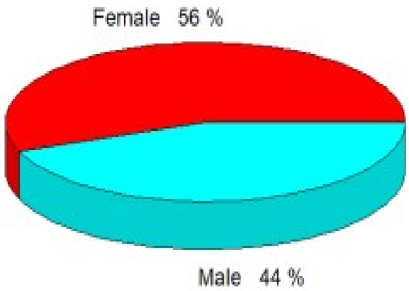


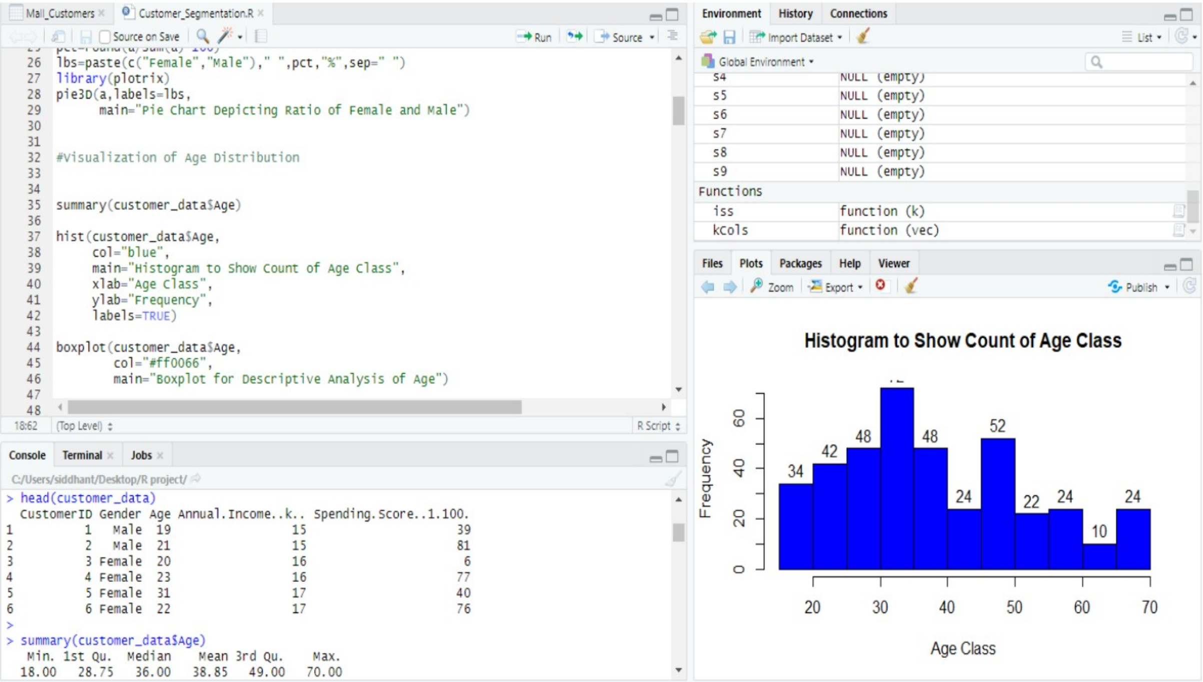
Figure - 3 Gender ratio

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

Code:

```
summary(customer_data$Age)

hist(customer_data$Age,

            col="blue",

             main="Histogram to Show Count of Age Class",

             xlab="Age Class",

             ylab="Frequency",

             labels=TRUE)
```



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


#Visualization of Age Distribution


summary(customer_data$Age)

hist(customer_data$Age,
     col="blue",
     main="Histogram to Show Count of Age Class",
     xlab="Age Class",
     ylab="Frequency",
     labels=TRUE)

boxplot(customer_data$Age,
        col="#ff0066",
        main="Boxplot for Descriptive Analysis of Age")
```

```
s4           NULL (empty)
s5           NULL (empty)
s6           NULL (empty)
s7           NULL (empty)
s8           NULL (empty)
s9           NULL (empty)
Functions
  iss        function (k)
  kCols      function (vec)
```

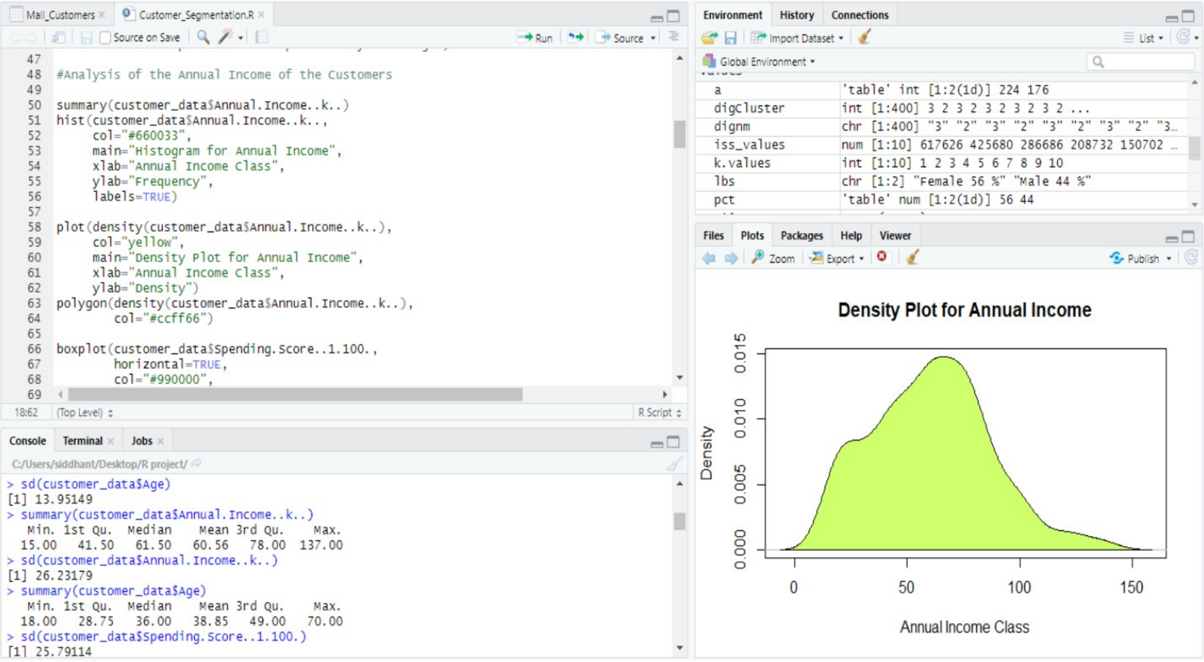
```
> head(customer_data)
  CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
1          1   Male  19                 15                     39
2          2   Male  21                 15                     81
3          3 Female  20                 16                      6
4          4 Female  23                 16                     77
5          5 Female  31                 17                     40
6          6 Female  22                 17                     76
>
> summary(customer_data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   28.75   36.00   38.85   49.00   70.00 
```

Figure- 4 Age Distribution

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 the Annual Income of the 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



```
#Analysis of the Annual Income of the Customers

summary(customer_data$Annual.Income..k..)
hist(customer_data$Annual.Income..k..,
     col="#660033",
     main="Histogram for Annual Income",
     xlab="Annual Income Class",
     ylab="Frequency",
     labels=TRUE)

plot(density(customer_data$Annual.Income..k..),
     col="yellow",
     main="Density Plot for Annual Income",
     xlab="Annual Income Class",
     ylab="Density")
polygon(density(customer_data$Annual.Income..k..),
        col="#ccff66")

boxplot(customer_data$Spending.Score..1.100.,
        horizontal=TRUE,
        col="#990000",
```

```
> sd(customer_data$Age)
[1] 13.95149
> summary(customer_data$Annual.Income..k..)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.00   41.50   61.50   60.56   78.00  137.00 
> sd(customer_data$Annual.Income..k..)
[1] 26.23179
> summary(customer_data$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   28.75   36.00   38.85   49.00   70.00 
> sd(customer_data$Spending.Score..1.100.)
[1] 25.79114
```

```
a             'table' int [1:2(1d)] 224 176
digCluster    int [1:400] 3 2 3 2 3 2 3 2 3 2 ...
dignm         chr [1:400] "3" "2" "3" "2" "3" "2" "3" "2" "3…
iss_values    num [1:10] 617626 425680 286686 208732 150702 …
k.values      int [1:10] 1 2 3 4 5 6 7 8 9 10
lbs           chr [1:2] "Female 56 %" "Male 44 %"
pct           'table' num [1:2(1d)] 56 44
```

Figure- 5 Annual Income

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.

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 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 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 stop wavering when we achieve maximum iteration. The default value is 10 that the R software uses for the maximum iterations.

we calculate the clustering algorithm for several values of k. This can be done by creating a variation within k from 1 to 10 clusters. We then calculate the total intra-cluster sum of square (iss). Then, we proceed to plot iss based on the number of k clusters. This plot denotes the appropriate number of clusters required in our model. In the plot, the location of a bend or a knee is the indication of the optimum number of clusters. Let us implement this in R as follows

Code:

```
library(purrr)
set.seed(123)
# function to calculate total intra-cluster sum of square
iss <- function(k) {
kmeans(customer_data[,3:5],k,iter.max=100,nstart=100,algorithm="Lloyd" )$tot.withinss
}
k.values <- 1:10
iss_values <- map_dbl(k.values, iss)
plot(k.values, iss_values,
        type="b", pch = 19, frame = FALSE,
        xlab="Number of clusters K",
        ylab="Total intra-clusters sum of squares")
```

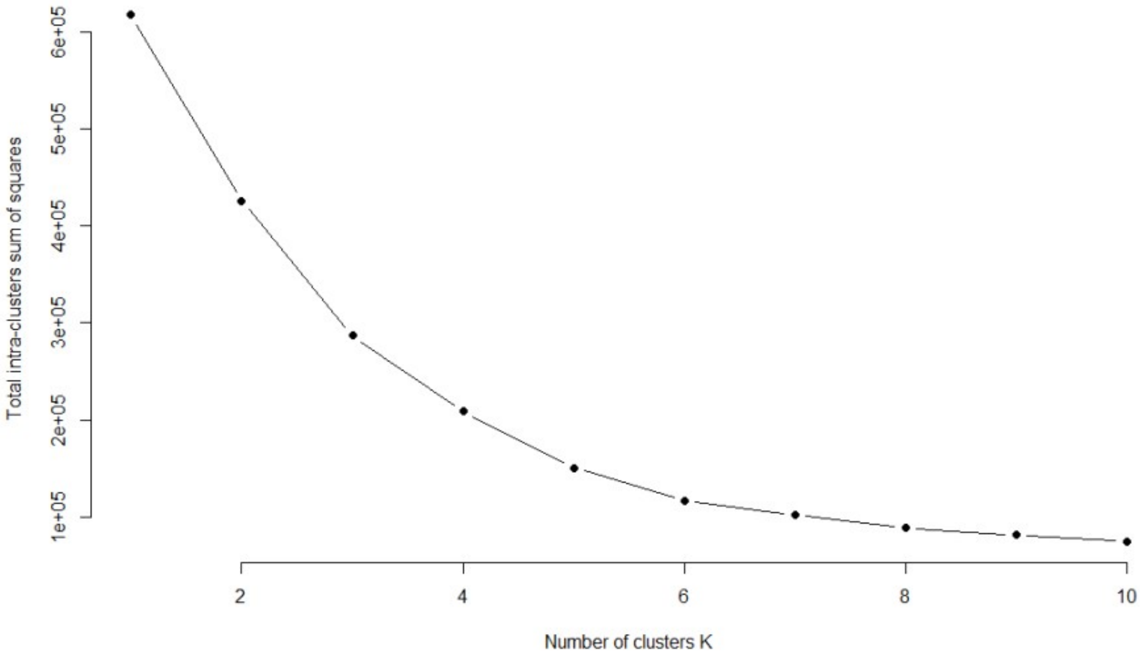


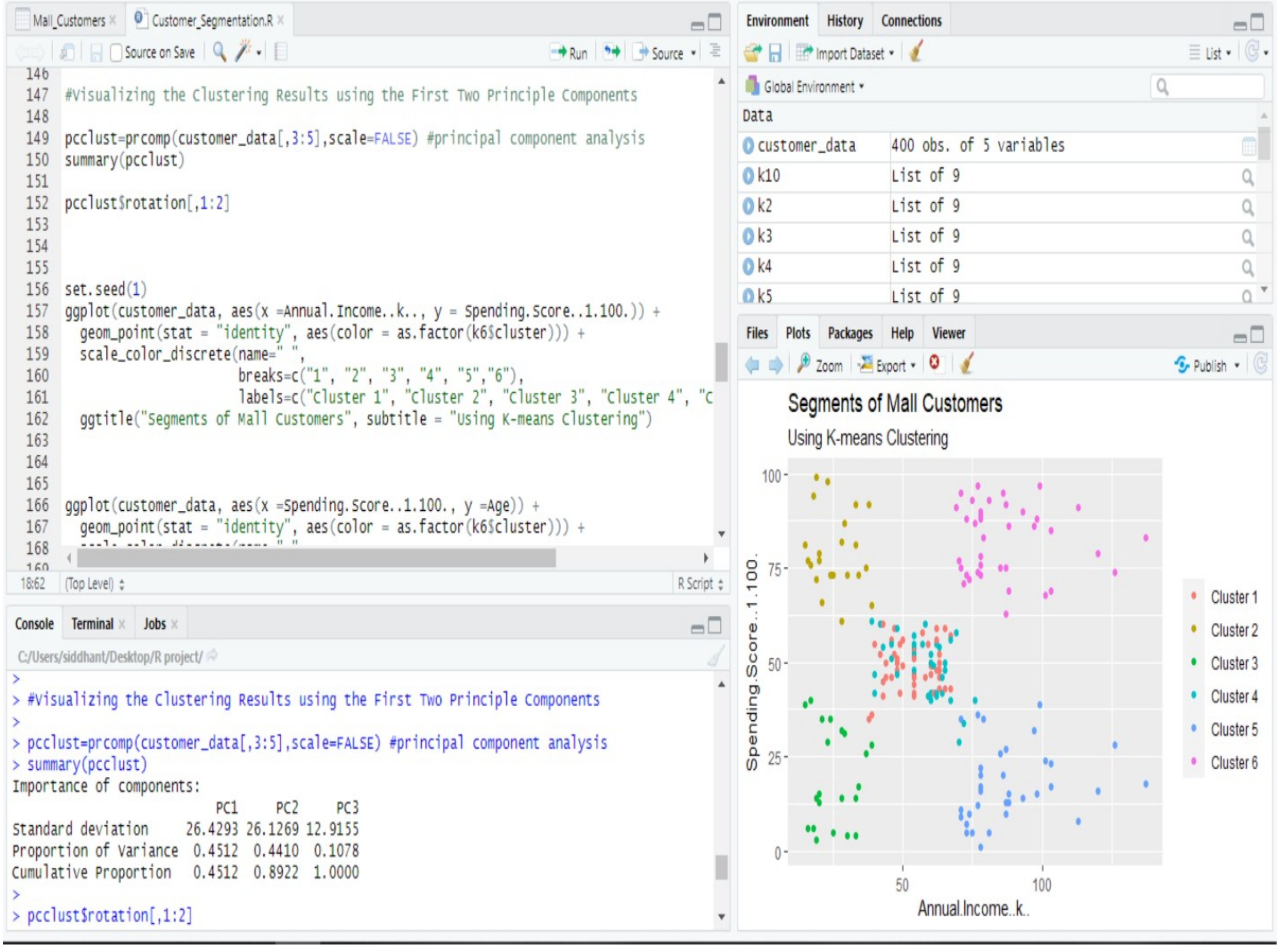
Figure - 6 Culsters

Visualizing the Clustering Results using the First Two Principle Components:

A line chart or line plot or line graph or curve chart is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments. It is a basic type of chart common in many fields. Used across many fields, this type of graph can be quite helpful in depicting the changes in values over time. We are going to use ggplot for depicting the line plot.

Code:

```
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")
```



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

pcclust=prcomp(customer_data[,3:5],scale=FALSE) #principal component analysis
summary(pcclust)

pcclust$rotation[,1:2]



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", "C
  ggtitle("Segments of Mall Customers", subtitle = "Using K-means Clustering")



ggplot(customer_data, aes(x =Spending.Score..1.100., y =Age)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
```

```
>
> #Visualizing the Clustering Results using the First Two Principle Components
>
> pcclust=prcomp(customer_data[,3:5],scale=FALSE) #principal component analysis
> 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
>
> pcclust$rotation[,1:2]
```

Data

customer_data	400 obs. of 5 variables
k10	List of 9
k2	List of 9
k3	List of 9
k4	List of 9
k5	List of 9

Figure-7 visualization

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

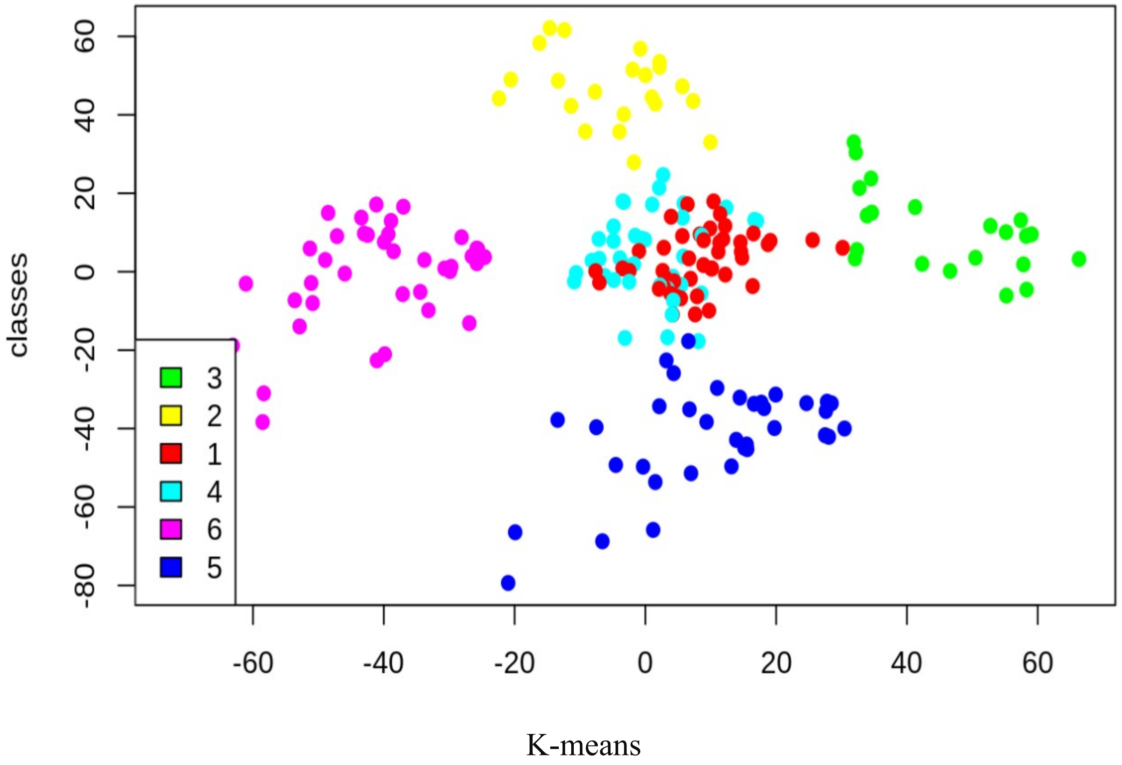


Figure 8 k-mean visualization

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.

With the help of clustering, we can understand the variables much better, prompting us to take careful decisions. With the identification of customers, companies can release products and services that target customers based on several parameters like income, age, spending patterns, etc. Furthermore, more complex patterns like product reviews are taken into consideration for better segmentation

CONCLUSION

In this data science 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 analysed and visualized the data and then proceeded to implement our algorithm. Hope you enjoyed this customer segmentation project of machine learning using R