



**Symbiosis Institute of Technology**

**PROJECT REPORT ON**

**Division of Customer Base in Shopping malls in targeted groups  
(Customer Segmentation)**

Submitted by:

**Anvita Gupta (18070124013)**

Under the Guidance of

**Dr.Preeti Mulay**

**Department of CS & IT**

**SYMBIOSIS INSTITUTE OF TECHNOLOGY, PUNE**

## **TABLE OF CONTENTS**

1. INTRODUCTION .....	1
1.1 Problem Statement .....	1
1.2 Purpose of this document.....	1
1.3 Background.....	1
1.4 Motivation or Significance of Work.....	2
1.5 Project Scope.....	2
1.6 System Purpose.....	2
1.6.1 Users.....	2
1.6.2 Need.....	3
1.7 Limitations.....	3
1.8 Overview of Document.....	3
2. Functional Requirements.....	4
2.1 High Priority.....	4
2.2 Medium Priority.....	5
2.3 Low Priority.....	5
3. Non Functional Requirements.....	5
3.1 Reliability.....	5
3.2 Usability.....	5
3.3 Performance.....	6
3.4 Supportability.....	6
4. Context Model.....	6
5. Use-Case.....	7

5.1 Use-Case model .....	7
5.2 Use-Case description.....	8
6. Entity Relationship Model.....	10
7. Sequence Diagram.....	10
8. Data Flow Diagram.....	11
8.1 Level-1 DFD.....	11
8.2 Level-2 DFD.....	12
9. Software Development Lifecycle Model (SDLC).....	13
10. Methodology used in the system.....	13
11. Project Size Estimation.....	14
12. Scheduling .....	15
13. Deliverables.....	15
14. Risk Management.....	15
14.1 Risks Involved.....	15
14.2 Risk Mitigation.....	16
15. Code Snippets.....	16
16. Screenshots of output.....	19
17. Testing details.....	29
18. Conclusion.....	42
19. References.....	43

# **1. INTRODUCTION:**

## **1.1 Problem Statement**

Division of customer base in shopping malls in groups to target potential customers (Customer Segmentation analysis).

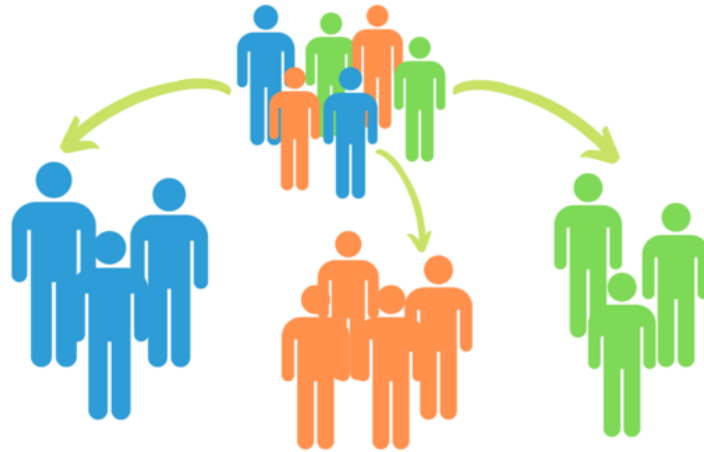
## **1.2 Purpose of Document**

This is a Requirements Specification document for a Project on Division of customer base in shopping malls in groups to target potential customers (Customer Segmentation analysis). The new system will upgrade the current businesses to provide customers and employees customized marketing of products. This document describes the scope, objectives and goal of the system. In addition to describing non-functional requirements, this document models the functional requirements with use cases, interaction diagrams, and class models. This document is intended to direct the design and implementation of the target system.

## **1.3 Background**

*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. To **create an irresistible product or service**, one needs to know who they are selling to. Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unsatisfied customer needs. Using the above, data companies can then outperform the competition by developing uniquely appealing products and services. We will be analyzing a dataset using K-means clustering technique in R language for this project. Through customer segmentation analysis, marketers can come up with the right messages, using the right words, to promote their products. Specifically, segmentation helps a company in the following ways:*

- 1) Improving the whole product and determining appropriate pricing.
- 2) Focusing on customized marketing messages.
- 3) Allowing the sales teams to pursue better opportunities instead of spending time on all the opportunities.
- 4) Prioritize new product development efforts.
- 5) Stabilizing customer base (Better customer relationships)
- 6) Getting higher quality revenues



### **1.4 Motivation or Significance of Work**

Coming from a business background I clearly understand the importance of customer segmentation to achieve better sales and profits. In these testing times the businesses at every scale are suffering huge losses and need to apply sales and marketing techniques in a better way in order to survive. Companies should reimagine their business model as they return to full speed and in order to do that they have to segment their customer base. Customer segmentation can be practiced by all businesses regardless of size or industry and whether they sell online or in person. It begins with gathering and analyzing data and ends with acting on the information gathered in a way that is appropriate and effective.

For example: A small business selling hand-made guitars, might decide to promote lower-priced products to younger guitarists and higher-priced premium guitars to older musicians based on segment knowledge that tells them that younger musicians have less disposable income than their older counterparts.

### **1.5 Project Scope**

The scope of this project is a RMD file using which the business analyst can enter the customer details in the software and then can analyze their spending scores, age, gender etc. so as to target their potential customers. The actual implementation of a user interface is not a part of this project.

### **1.6 System Purpose**

#### **1.6.1 Users**

Those who will primarily benefit from the system and those who will be affected by the new system include:

Customers:

Upon implementation of the new system, customers will be provided with customized products according to their liking easily and faster.

Product Owners:

Product owners will manufacture products according to customer's preferences and demand and in turn will receive huge profits.

Customer Service Department:

The new system should reduce the workload of Customer Service as customers will be happier and there are fewer chances of returns.

Marketing Department:

The marketing department will benefit a lot as they'll focus on marketing targeted groups of people for a specific product. Their time will be saved.

Information Technology Department:

This department will be responsible for implementing the new project and adding details in csv file.

### **1.6.2 Need**

This system is needed in order to provide customers with products according to their preferences and also provide the organization with huge profits. The sales and marketing team will not waste their time in searching for customer specific products. It can also be used in E-commerce applications. Ecommerce transactions are no longer a new thing. Many people shop with ecommerce and many companies use ecommerce to promote and to sell their products. Because of that, overloading information appears on the customers' side. Overloading information occurs when customers get too much information about a product then feel confused. **Personalization** will become a solution to overloading problem. In marketing, personalization technique can be used to get potential customers in a case to boost sales. The potential customer is obtained from customer segmentation or market segmentation.

### **1.7 Limitations**

The actual implementation of a GUI is not a part of this project. The data should be clear and concise so as to avoid irrelevant outputs.

### **1.8 Overview of Document**

The rest of this document gives the detailed specifications for the project. It is organized as follows:

- **Section 2: Functional Objectives:**  
Each objective gives a desired behavior for the system, a business justification, and a measure to determine if the final system has successfully met the objective. These objectives are organized by priority. In order for the new system to be considered successful, all high priority objectives must be met.
- **Section 3: Non-Functional Objectives:**  
This section is organized by category. Each objective specifies a technical requirement or constraint on the overall characteristics of the system. Each objective is measurable.

- Section 4: Context Model:  
This section gives a text description of the goal of the system, and a pictorial description of the scope of the system in a context diagram. Those entities outside the system that interact with the system are described.
- Section 5: Use Case Model:  
The specific behavioral requirements of the system are detailed in a series of use cases. Each use case accomplishes a business task and shows the interaction between the system and some outside actor. Each use case is described with both text and an interaction diagram. An interface prototype is also shown. The system use case diagram depicts the interactions between all use cases and system actors.
- Section 6: Class Model:  
A class is a collection of objects in the system that has the same data and behavior. All analysis classes and their relationships are shown on the class diagram.
- Section 7: Sequence Model:  
A sequence diagram is a type of interaction diagram because it describes how—and in what order—a group of objects works together.
- Section 8: Data Flow Diagram:  
Data flow diagrams are used to graphically represent the flow of data in a business information system. DFD describes the processes that are involved in a system to transfer data from the input to the file storage and reports generation.
- Section 9: Software development Lifecycle Model(SDLC):  
Software Development Life Cycle (SDLC) is a process used by the software industry to design, develop and test high quality software's.
- Section 10: Technical Details:  
This section contains all the technical aspects of the project.
- Section 11: Scheduling:  
Scheduling is the process of arranging, controlling and optimizing work and workloads in a production process or manufacturing process.
- Section 12: Deliverables:  
This section contains all the deliverable to the client.
- Section 13: Risk Management:  
There are risks involved in every project therefore this section contains all the risks and their mitigation.

## **2.FUNCTIONAL REQUIREMENTS:**

### **2.1 High Priority**

1. The project must be able to display accurate output in the form of graphs and clusters.
2. The user must be able to input different data after every few months in a csv file and the output should be accordingly changed.
3. The Output must be clear and fast for the user to understand easily.
4. It should reduce unnecessary marketing to uninterested group of customers for a particular product.

5. The system shall display information that is customized based on the user's input. This feature will improve service.
6. Better customer relations and stabilized customer base should be formed after implementing this system.
7. There should be no extra clusters and should be optimal.

## **2.2 Medium Priority**

1. The system shall provide data visualization on other factors also. The system must show the following graphs:
  - A) Gender
  - B) Age
  - C) Spending Scores
  - D) Annual Income
2. The system must summarize the given data which was given in the csv file.

## **2.3 Low Priority**

1. The system shall allow the user's last data to be stored for the next time he returns to the project. This will save the user x minutes per visit by not having to reenter already supplied data.
2. The system shall provide clusters with a particular customer's information. This information will allow marketing team to determine what information prompts a purchase and help target potential customers more effectively. This will increase annual revenue in additional sales.
3. The system shall translate every type of file into the system like csv, xls, zip etc. This will improve the service and reduce the work in converting file types.

## **3. NON- FUNCTIONAL REQUIREMENTS:**

### **3.1 Reliability**

- The system shall be completely operational all of the time.
- Down time after a failure shall not exceed 2 hours .

### **3.2 Usability**

- The I.T. team should be able to use the system in his job after 7 days of training.
- A user who already knows how to operate the system shall be able to view the clusters in 10 seconds.
- The number of clusters should not exceed 10.



### **3.3 Performance**

- The system should be able to support 1 user at a particular time.
- The mean time to view clusters after entering the data file shall not exceed 15 seconds.

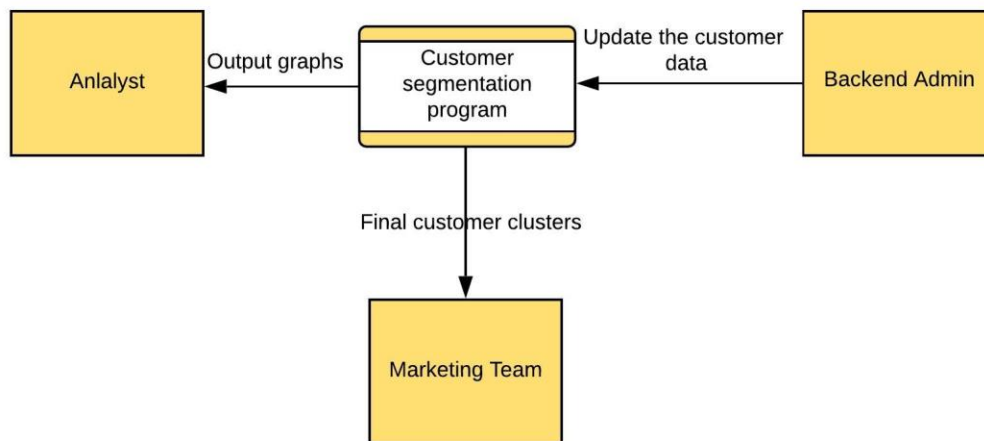
### **3.4 Supportability**

- The system should be able to accommodate new data entries without major reengineering.
- The system should run on any R studio version above 1.2.

## **4. CONTEXT MODEL:**

### **CUSTOMER SEGMENTATION PROJECT CONTEXT DIAGRAM**

Er.Anvita Gupta | December 3, 2020

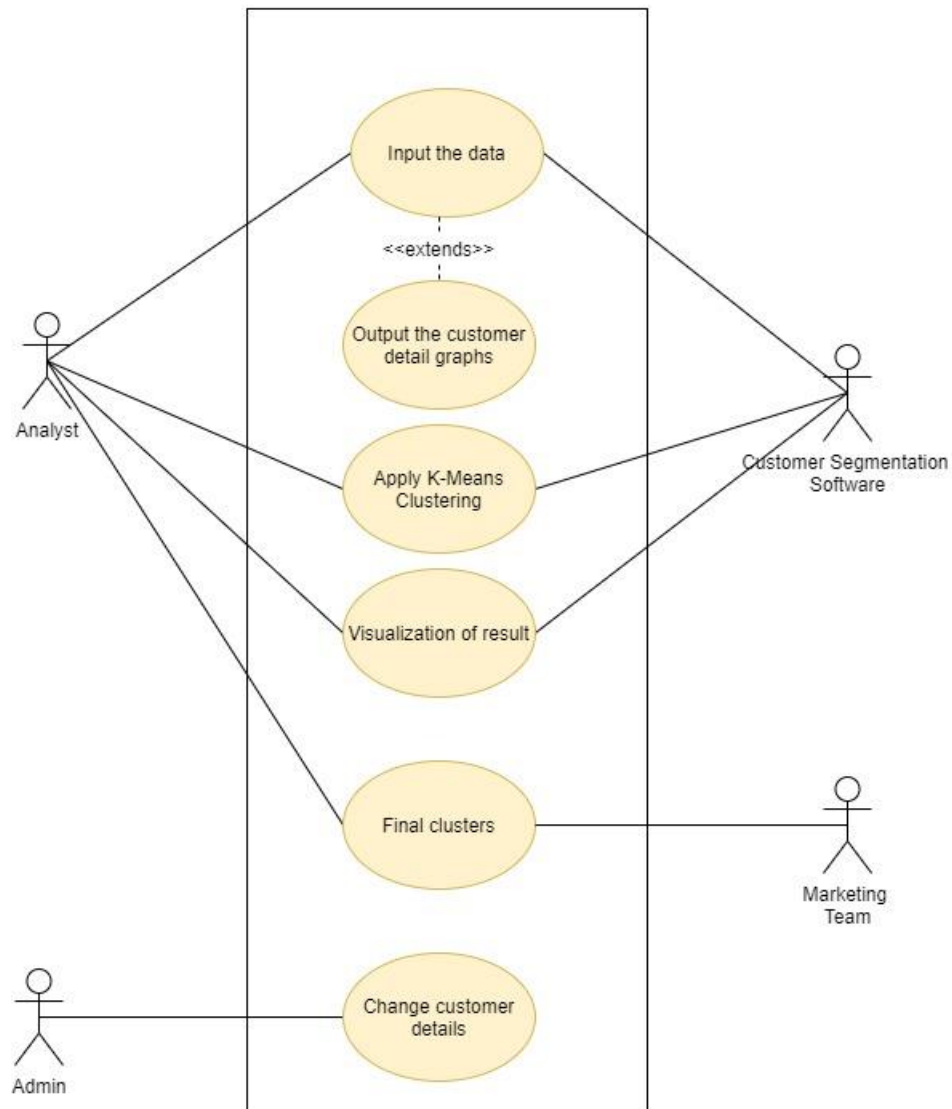


## **5. USE CASE MODEL:**

### **5.1 Use Case Diagram**

#### **USE CASE DIAGRAM**

Customer segmentation Project | By Er.Anvita Gupta | 27 October,2020



## **5.2 Use Case Description**

### **Input the data**

Use Case Name:	Input the data
Summary:	In order to add data and visualize the data the analyst adds the data into the software.
Basic Flow:	1. The use case starts when the analyst inputs the data into software.
Extension Points:	The data can be extended to output customer details graphs.
Preconditions:	The data is in csv format.
Post conditions:	None

### **Output the customer details graph**

Use Case Name:	Output the customer details graph
Summary:	In order to provide all the graphs of customer details
Basic Flow:	1. The use case starts when the analyst inputs the data into software. 2. Then the software outputs the graphs.
Extends Use cases:	Input the data
Preconditions:	The data is already fed in the software.
Post conditions:	None

### **Apply K-Means Clustering**

Use Case Name:	Apply K-Means Clustering
Summary:	In this step we apply K-Means Clustering method.
Basic Flow:	1. The software applies the K-means clustering algorithm to the data input.
Extension Points:	None
Preconditions:	The data inserted should be correct and accurate.
Post conditions:	None

### **Visualization of Results**

Use Case Name:	Visualization of results
Summary:	In order to visualize the clustering results using PCA.
Basic Flow:	<ol style="list-style-type: none"><li>1. The use case starts when the software has applied Kmeans Clustering.</li><li>2. The clustering results are printed as a summary using principal component analysis.</li></ol>
Extension Points:	None
Preconditions:	The Kmeans clustering should have already have been applied.
Post conditions:	None

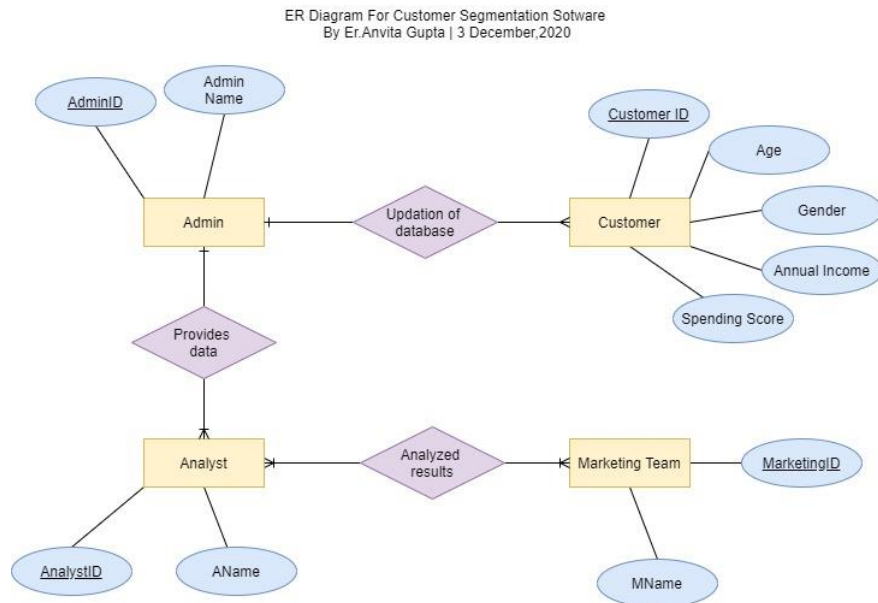
### **Final Clusters**

Use Case Name:	Final clusters
Summary:	In order to ouput the final customer clusters.
Basic Flow:	<ol style="list-style-type: none"><li>1. The use case starts when the user visualized the clusters and applied PCA.</li><li>2. The clusters have to be displayed combining the results.</li><li>3. Then the marketing team makes use of these clusters.</li></ol>
Extension Points:	None
Preconditions:	Kmeans has to be applied.
Post conditions:	None

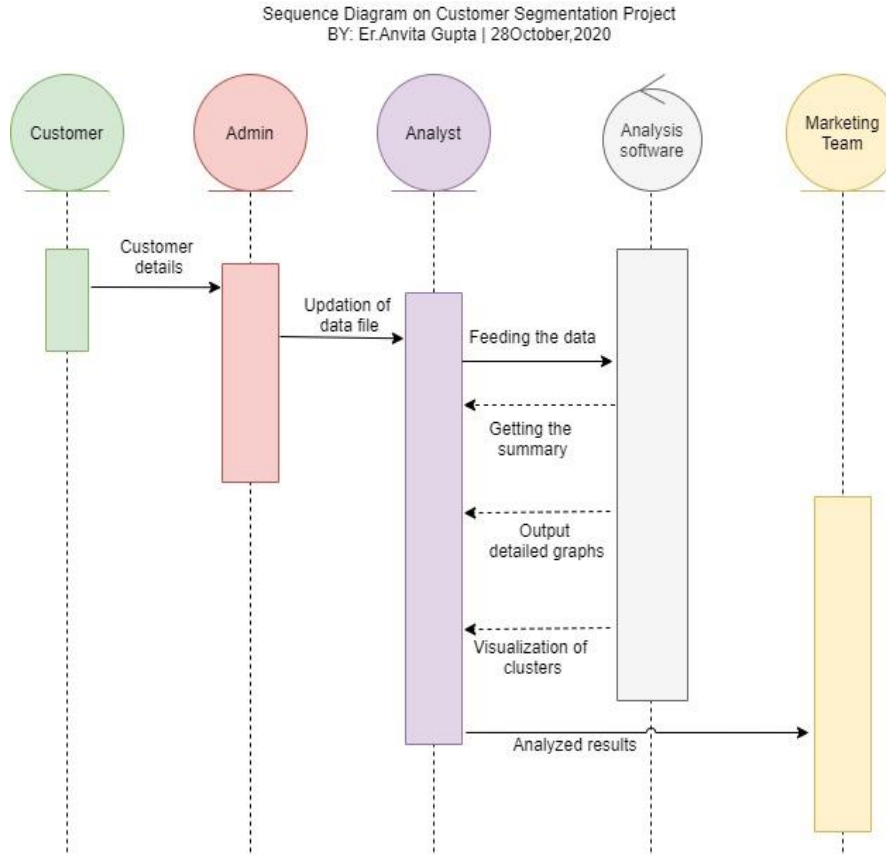
### **Change customer details**

Use Case Name:	Change customer details
Summary:	In order to change or update customer details in the dataset.
Basic Flow:	<ol style="list-style-type: none"><li>1. The use case starts when the admin changes or updates customer data in the dataset so as to get updated clusters.</li></ol>
Extension Points:	None
Preconditions:	None
Post conditions:	None

## 6.Entity Relationship Model

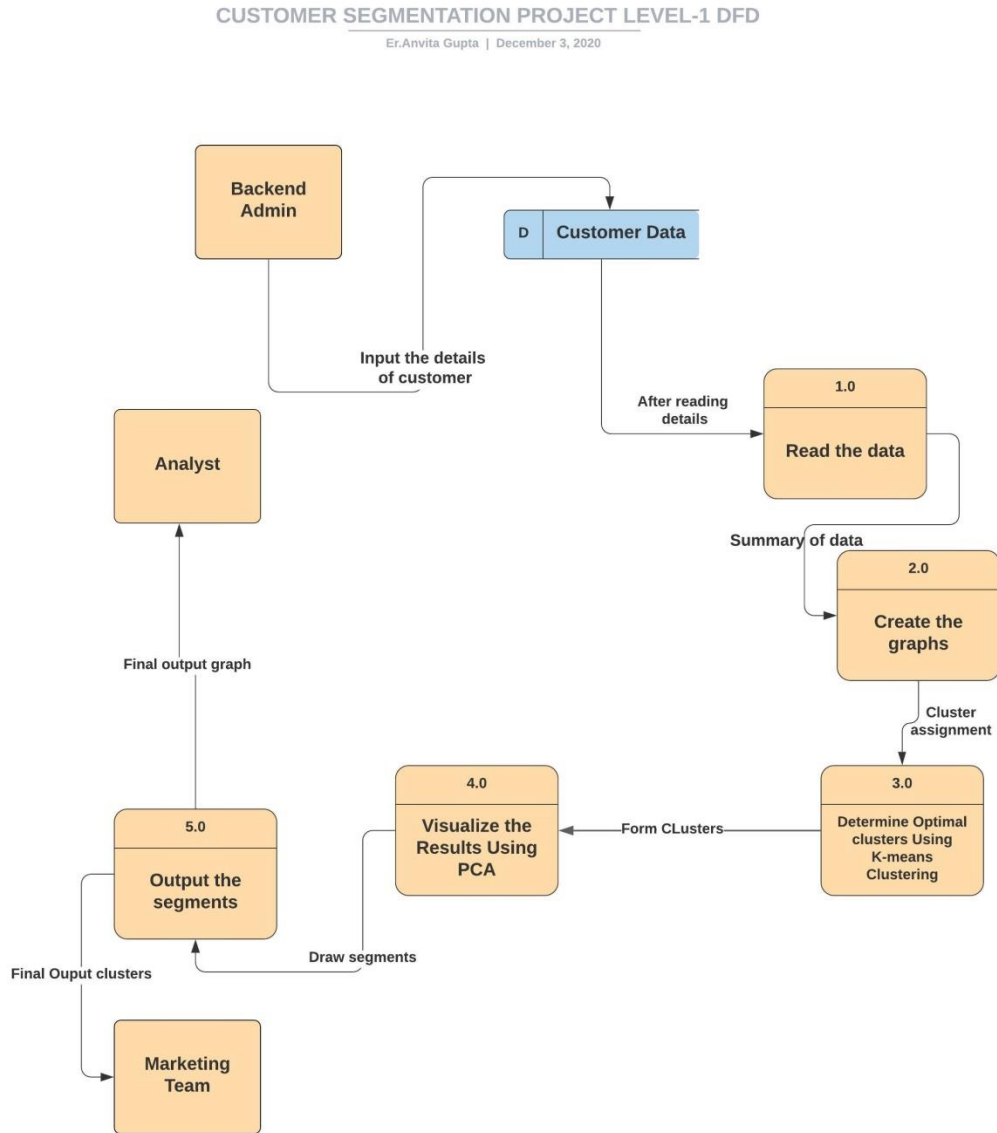


## 7.SEQUENCE DIAGRAM:

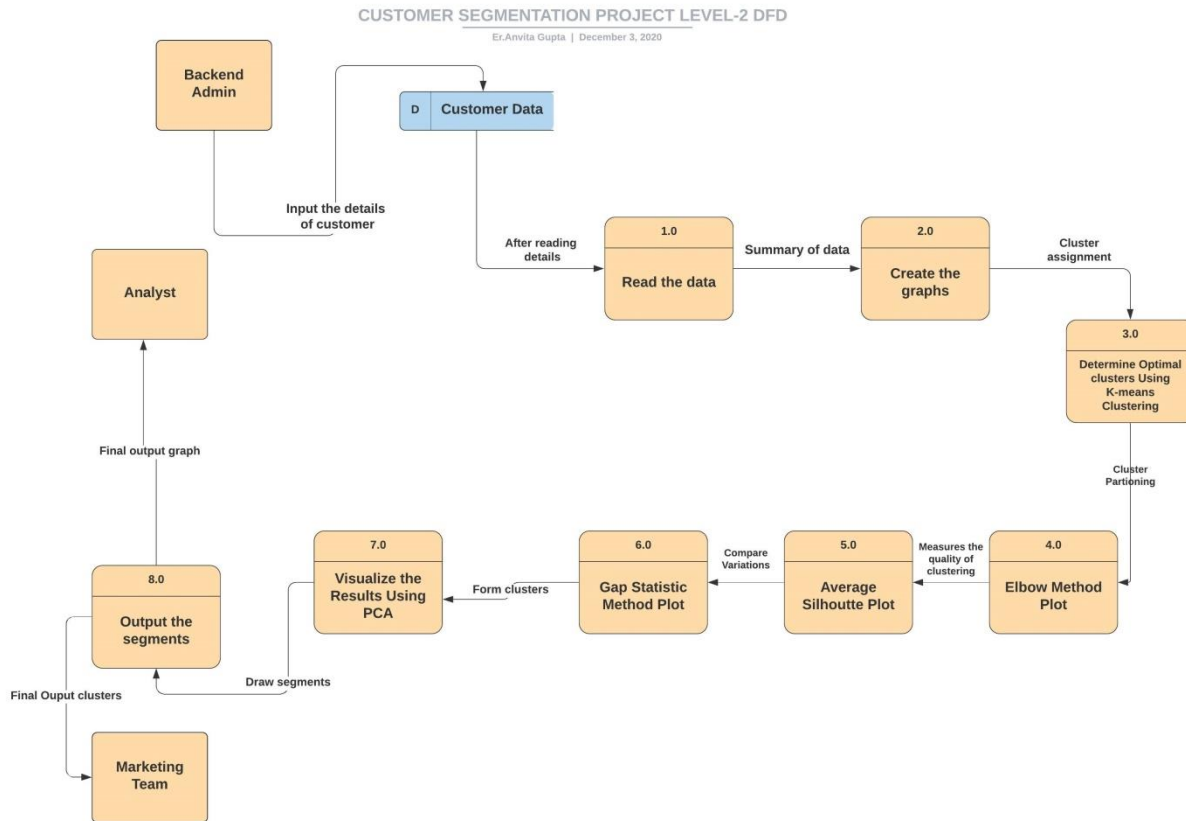


## 8.DATA FLOW DIAGRAM:

### 8.1 LEVEL-1 DFD

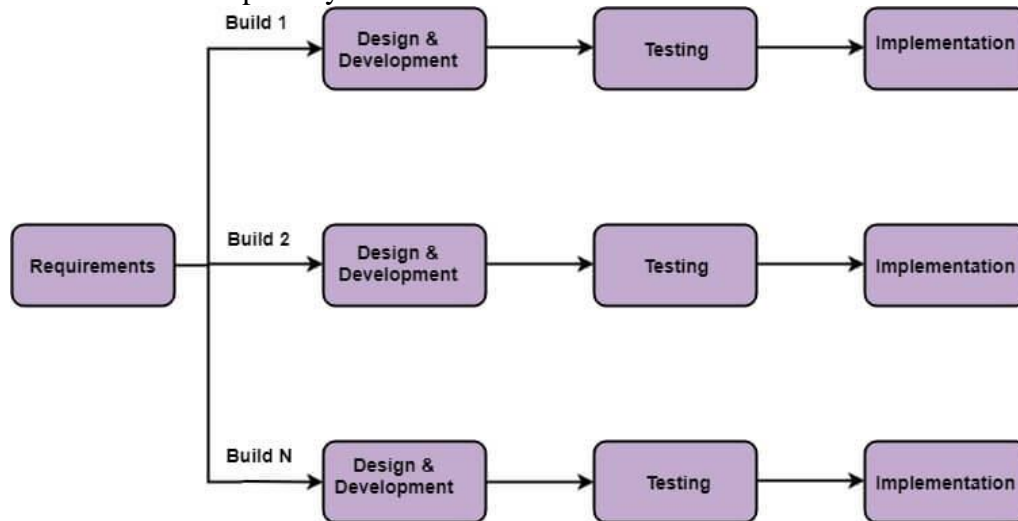


## 8.2 LEVEL-2 DFD



## **9. SOFTWARE DEVELOPMENT LIFECYCLE MODEL (SDLC):**

The SDLC model used for this project is Incremental Model. In this approach, each module goes through the requirements, design, implementation and testing phases. Every subsequent release of the module adds function to the previous release. The process continues until the complete system is achieved.



**Fig: Incremental Model**

This particular model was suitable because:

- Requirements of the system are clearly understood.
- Such methodology is more in use for product based companies
- It is flexible and less expensive to change requirements and scope.
- Errors are easy to be identified
- When high-risk features and goals are involved
- This model is less costly compared to others.

## **10. Methodology Used in the System:**

In this project, we will implement customer segmentation in R language. To find the best customer, customer segmentation is the ideal methodology. We will first explore the dataset upon which we will be building our segmentation model. We will also see the descriptive analysis (Graphs) of the data and then implement several versions of the K-means algorithm. The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. Finally, we will output the optimal customer clusters.

### **Recommended Hardware Requirements:**

- 8GB RAM
- 16 GB of free hard disk space



- 1 GHz or faster processor
- Windows 8 or higher required

### **Software Requirements:**

1. R Studio(R markdown)

Libraries used:

- Plotrix
- Purrr
- NbClust
- Factoextra
- R-table
- Ggplot2

2. MS-Excel to open Dataset in csv format

### **Online tools Used:**

1. Kaggle (for dataset)
2. Lucidchart (for diagrams)

## **11.PROJECT SIZE ESTIMATION:**

Estimation of the size of software is an essential part of Software Project Management. It helps the project manager to further predict the effort and time which will be needed to build the project. Various measures are used in project size estimation. I am using Lines of Code (LOC) here. Lines of code or LOC is the most popular and used metrics to estimate size. LOC determination is simple as well. LOC measures the project size in terms of number of lines of statements or instructions written in the source code. In this count, comments and headers are ignored. The size is estimated by comparing it with the existing systems of same kind.

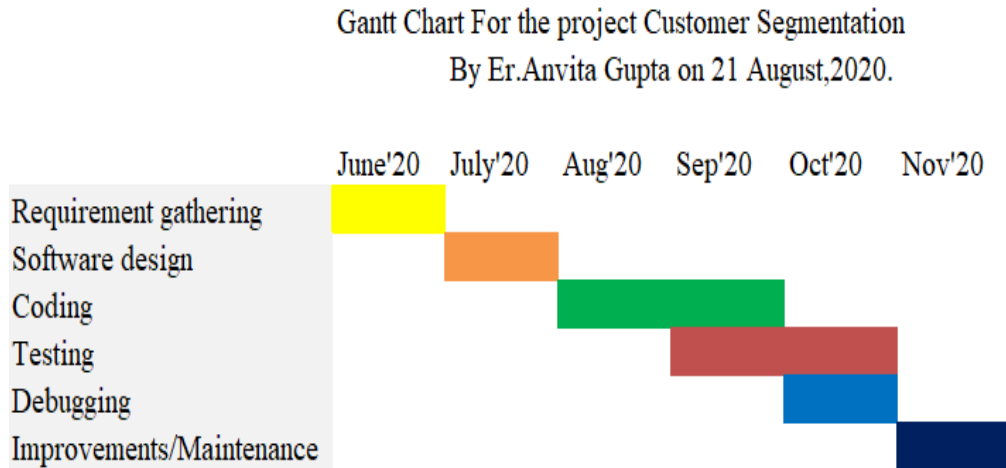
The total non-comment lines of code (NLOC) in this project are 124.

### **FUNCTIONS IN DATA FLOW DIAGRAM ALONG WITH THEIR LOC:**

1. DATA READING-5
2. DATA EXPLORATION-5
3. DATA VISUALIZATION-48

4. K-MEANS CLUSTERING-41
5. VISUALIZE USING PCA-16
6. FINAL CLUSTERS IN GRAPH-5

## **12.SCHEDULING:**



## **13.DELIVERABLES:**

I'll deliver the following during the course of development:

- Feature specification
- Product design
- Test plan
- Development document
- Source code

## **14.RISK MANAGEMENT:**

### **14.1 Risk Identification**

Following will be the risk involved in my project:

- 1) If the customer data is incorrect it might produce inaccurate results which might result in heavy losses.

2) The analyst might not use properly formatted file with null values which might cause an error or imprecisely read the outputs.

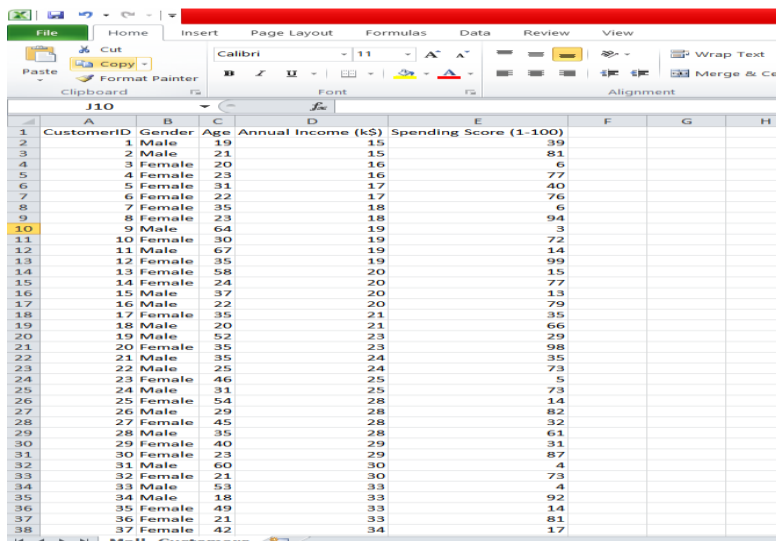
## **14.2 Risk Mitigation**

The admin needs to carefully enter customer data and recheck them. He should also search for any anomalies in the given data and remove them.

The analyst should prepare the data before starting the software and should properly analyze the outputs. Thus, I think that the amount of risks will be drastically reduced if the work is done carefully and precisely.

## **15.CODE SNIPPETS**

### **15.1Dataset:**



CustomerID	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
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

### **15.2 Code**

```

---
title: "Customer Segmentation"
author: "Anvita"
date: "7/24/2020"
output: html_document
---

```{r}
#import dataset
customer_data=read.csv("C:/Users/ASUS/Desktop/SEM V/SE/Project/Wall_Customers.csv")
str(customer_data)
```

```{r}
#print column names
names(customer_data)
```

```{r}
#display the first six rows of our dataset
head(customer_data)
#output the summary
summary(customer_data$Age)
```

```{r}
sd(customer_data$Age)
summary(customer_data$Annual.Income..k..)
sd(customer_data$Annual.Income..k..)
summary(customer_data$Age)
```

```{r}
sd(customer_data$Spending.Score..1.100.)
```

```{r}
#gender visualization using barplot
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

```{r}
install.packages("plotrix")
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
#gender visualization using pie-plot
library(plotrix)
pie3D(a,labels=lbs,
      main="Pie Chart Depicting Ratio of Female and Male")
```

```{r}
summary(customer_data$Age)
```

```{r}
#age visualization using histogram
hist(customer_data$Age,col="green",main="Histogram to Show Count of Age Class", xlab="Age Class", ylab="Frequency",labels=TRUE)

```

```

```{r}
#age visualization using boxplot
boxplot(customer_data$Age,
        col="#ff0066",
        main="Boxplot for Descriptive Analysis of Age")
```

```{r}
#annual income visualization using histogram
hist(customer_data$Annual.Income..k..,
      col="#660033",
      main="Histogram for Annual Income",
      xlab="Annual Income Class",
      ylab="Frequency",
      labels=TRUE)
```

```{r}
#annual income visualization using density plot
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")
```

```{r}
#spending score visualization using boxplot
summary(customer_data$Spending.Score..1.100.)
boxplot(customer_data$Spending.Score..1.100.,
        horizontal=TRUE,
        col="red",
        main="BoxPlot for Descriptive Analysis of Spending Score")
```

```{r}
#spending score visualization using histogram
hist(customer_data$Spending.Score..1.100.,
      main="Histogram for spending Score",
      xlab="Spending Score Class",
      ylab="Frequency",
      col="#6600cc",
      labels=TRUE)
```

```{r}
#finding optimal clusters
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 no of clusters vs intra-cluster sum of squares
plot(k.values, iss_values,
     type="b", pch = 19, frame = FALSE,

```

```

xlab="Number of clusters K",
ylab="Total intra-clusters sum of squares")
...
{r}
install.packages("gridExtra")
library(cluster)
library(gridExtra)
library(grid)
#Using average silhouette method
#Finding highest average which will be of optimal cluster
k2<-kmeans(customer_data[,3:5],2,iter.max=100,nstart=50,algorithm="Lloyd")
s2<-plot(silhouette(k2$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
k3<-kmeans(customer_data[,3:5],3,iter.max=100,nstart=50,algorithm="Lloyd")
s3<-plot(silhouette(k3$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
k4<-kmeans(customer_data[,3:5],4,iter.max=100,nstart=50,algorithm="Lloyd")
s4<-plot(silhouette(k4$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
k5<-kmeans(customer_data[,3:5],5,iter.max=100,nstart=50,algorithm="Lloyd")
s5<-plot(silhouette(k5$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
k6<-kmeans(customer_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
s6<-plot(silhouette(k6$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
k7<-kmeans(customer_data[,3:5],7,iter.max=100,nstart=50,algorithm="Lloyd")
s7<-plot(silhouette(k7$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
k8<-kmeans(customer_data[,3:5],8,iter.max=100,nstart=50,algorithm="Lloyd")
s8<-plot(silhouette(k8$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
k9<-kmeans(customer_data[,3:5],9,iter.max=100,nstart=50,algorithm="Lloyd")
s9<-plot(silhouette(k9$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
k10<-kmeans(customer_data[,3:5],10,iter.max=100,nstart=50,algorithm="Lloyd")
s10<-plot(silhouette(k10$cluster,dist(customer_data[,3:5],"euclidean")))
...
{r}
install.packages("NbClust")
install.packages("factoextra")
...
{r}
library(NbClust)
library(factoextra)

```

```

{r}
library(NbClust)
library(factoextra)
#Visualize the optimal number of clusters
fviz_nbclust(customer_data[,3:5], kmeans, method = "silhouette")
...
{r}
library(cluster)
set.seed(123)
#gap statistic method
stat_gap <- clusGap(customer_data[,3:5], FUN = kmeans, nstart = 25,K.max = 10, B = 50)
fviz_gap_stat(stat_gap)
...
{r}
#Optimal no of clusters will be 6
k6<-kmeans(customer_data[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
k6
...
{r}
#visualizing the clustering results using PCA
pcclust<-prcomp(customer_data[,3:5],scale=FALSE) #principal component analysis
summary(pcclust)
...
{r}
pcclust$rotation[,1:2]
...
{r}
set.seed(1)
ggplot(customer_data, aes(x =Annual.Income.k., y = Spending.Score..1.100.)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$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")
...
{r}
ggplot(customer_data, aes(x =Spending.Score..1.100., y =Age)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  scale_color_discrete(name="")
  breaks=c("1", "2", "3", "4", "5","6"),
  labels=c("Cluster 1", "Cluster 2", "Cluster 3", "Cluster 4", "Cluster 5","Cluster 6")) +
  ggtitle("Segments of Mall customers", subtitle = "Using K-means clustering")
...
{r}
#Final customer clusters
kcols=function(vec){cols=rainbow (length (unique (vec)))
return (cols[as.numeric(as.factor(vec))])}
digcluster<-k6$cluster; dignm<-as.character(digcluster); # K-means clusters
plot(pcclust$x[,1:2], col =kcols(digcluster),pch =19,xlab ="K-means",ylab="classes")
legend("bottomleft",unique(dignm),fill=unique(kcols(digcluster)))

```

## 16.SCREENSHOTS OF OUTPUT:

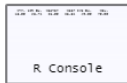
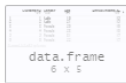
```
#import dataset
customer_data=read.csv("C:/Users/ASUS/Desktop/SEM V/SE/Project/Mall_Customers.csv")
str(customer_data)
```

```
'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 ...
```

```
##{r}
#print column names
names(customer_data)
```

```
[1] "customerID"      "Gender"          "Age"             "Annual.Income..k.." "Spending.Score..1.100."
```

```
##{r}
#display the first six rows of our dataset
head(customer_data)
#output the summary
summary(customer_data$Age)
```



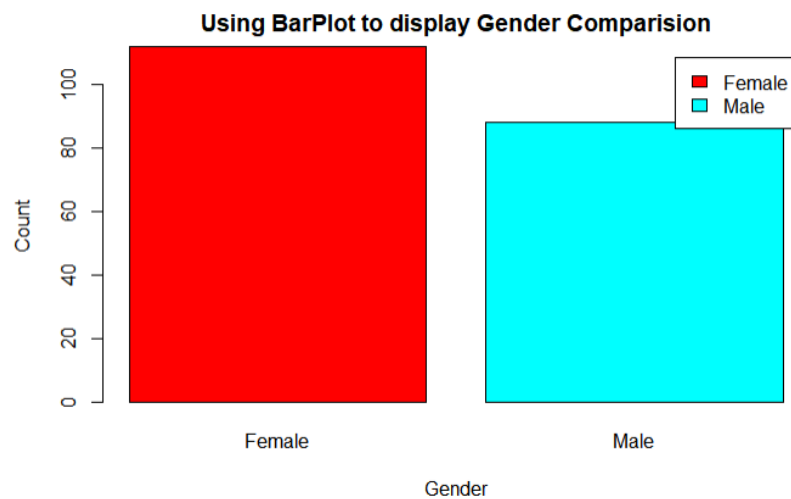
```
Min. 1st Qu. Median Mean 3rd Qu. Max.
18.00 28.75 36.00 38.85 49.00 70.00
```

```
##{r}
sd(customer_data$Age)
summary(customer_data$Annual.Income..k..)
sd(customer_data$Annual.Income..k..)
summary(customer_data$Age)
```

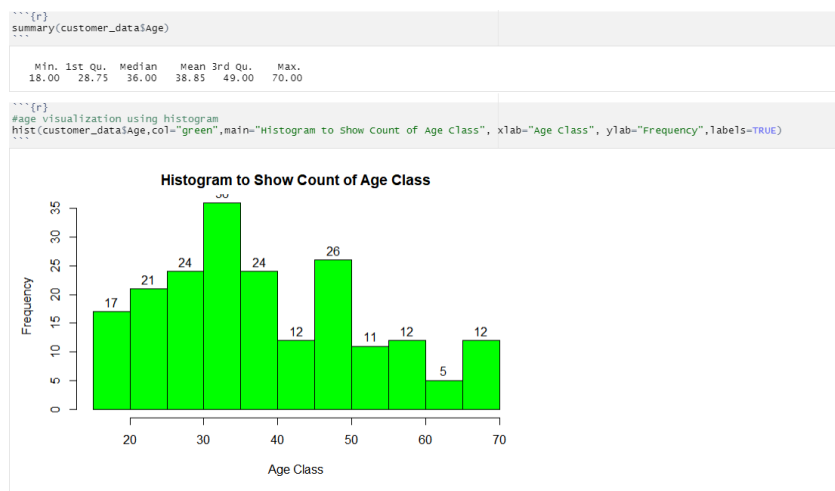
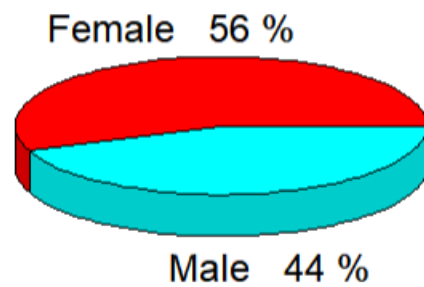
```
[1] 13.96901
Min. 1st Qu. Median Mean 3rd Qu. Max.
15.00 41.50 61.50 60.56 78.00 137.00
[1] 26.26472
Min. 1st Qu. Median Mean 3rd Qu. Max.
18.00 28.75 36.00 38.85 49.00 70.00
```

```
##{r}
sd(customer_data$Spending.Score..1.100.)
```

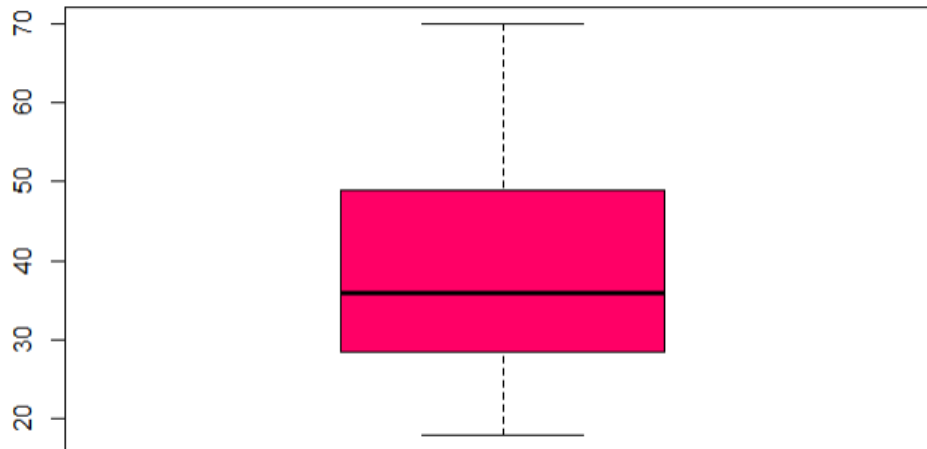
```
[1] 25.82352
```



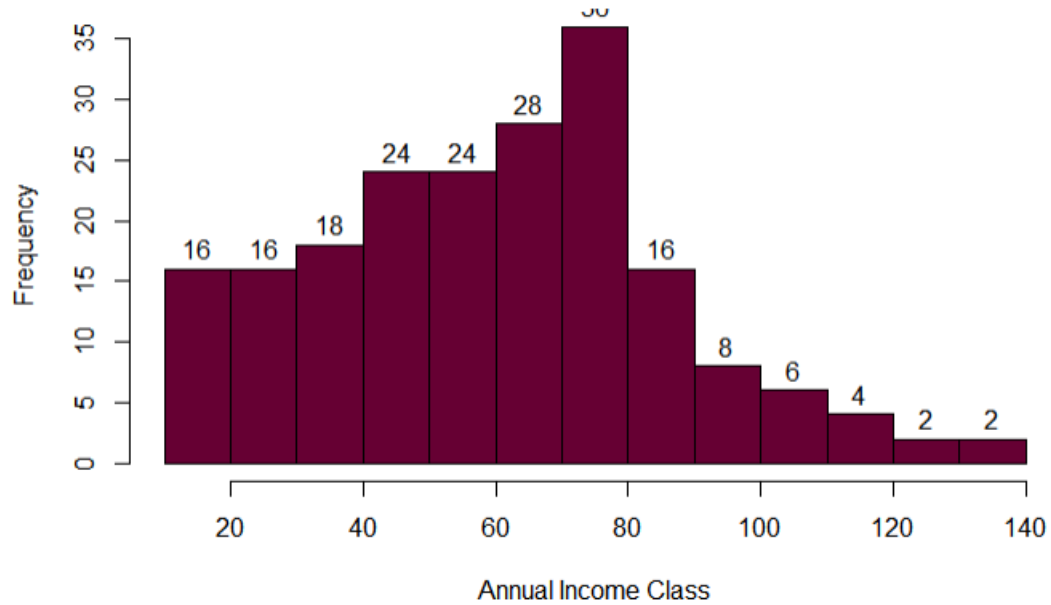
**Pie Chart Depicting Ratio of Female and Male**



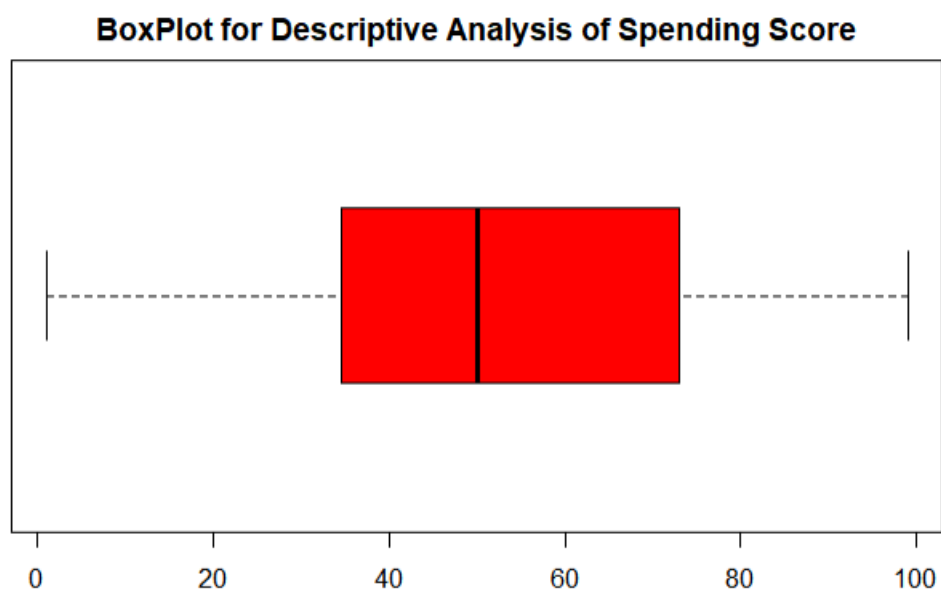
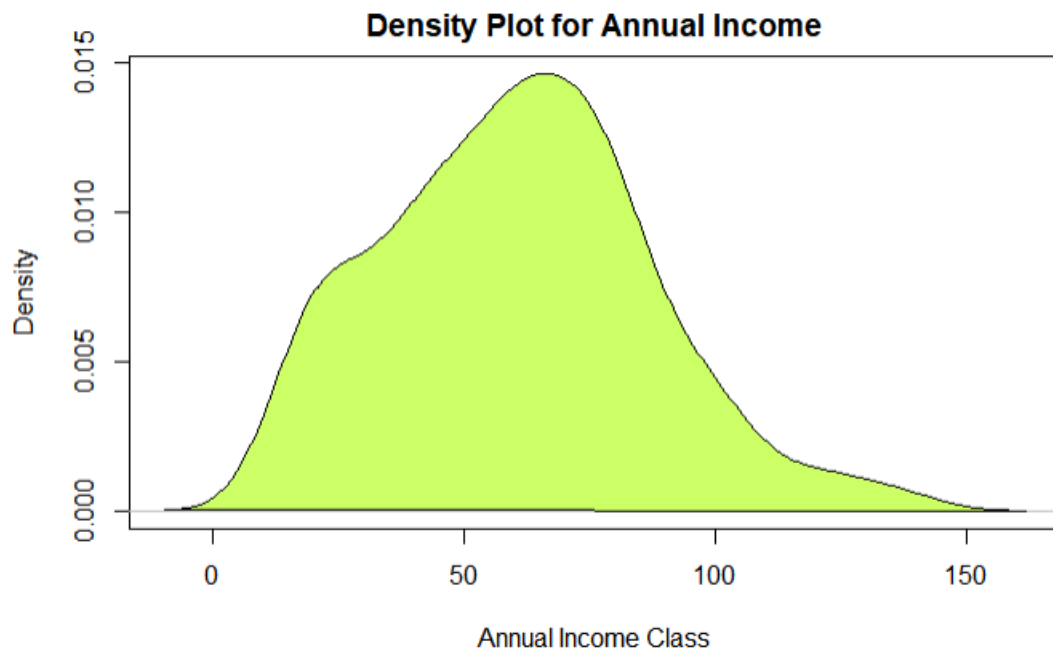
**Boxplot for Descriptive Analysis of Age**

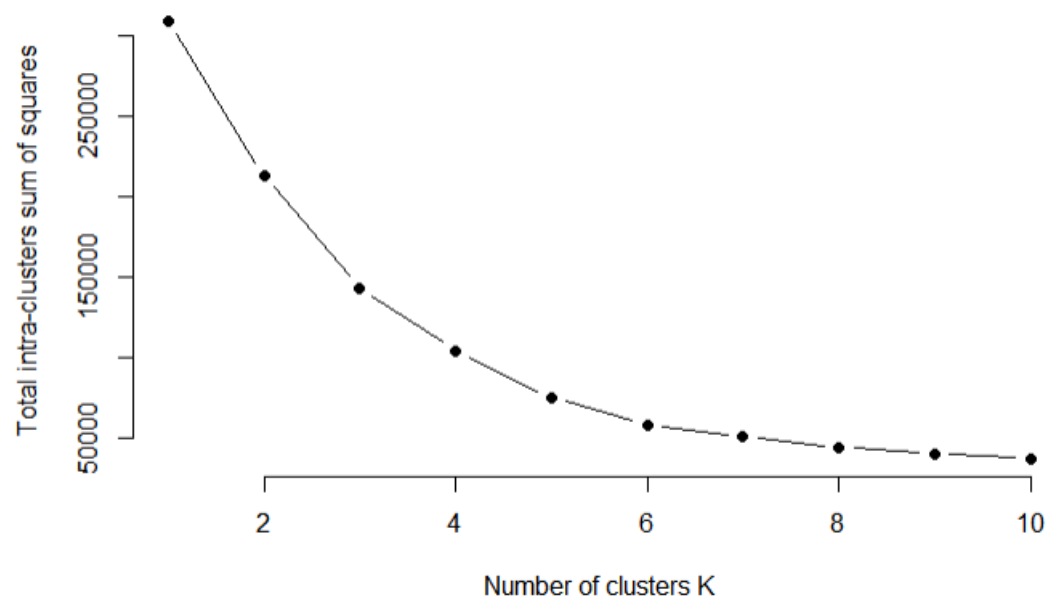
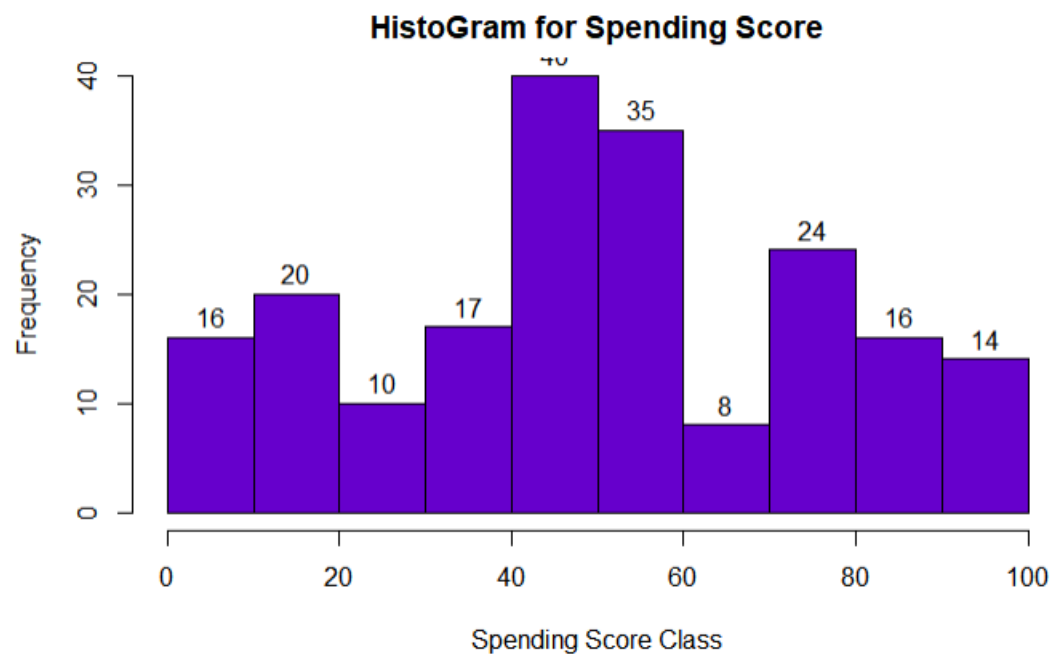


**Histogram for Annual Income**

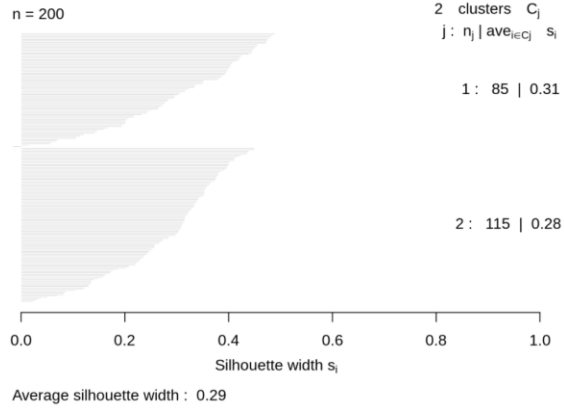




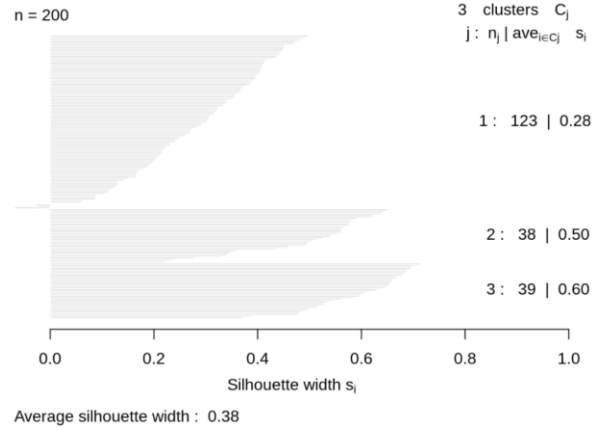




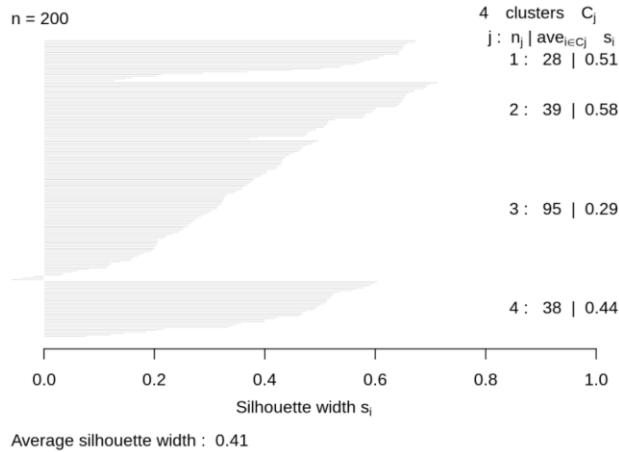
Silhouette plot of (x = k2\$cluster, dist = dist(customer\_data[, 3:5],



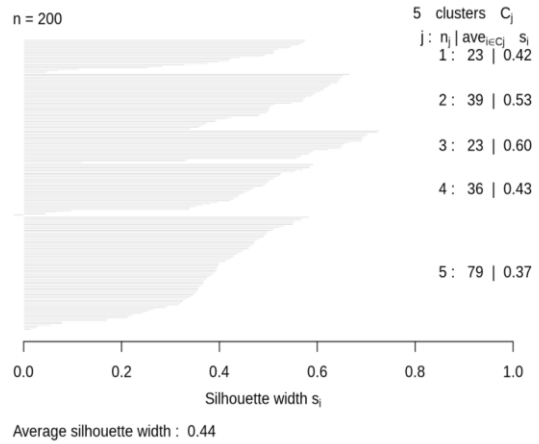
Silhouette plot of (x = k3\$cluster, dist = dist(customer\_data[, 3:5],



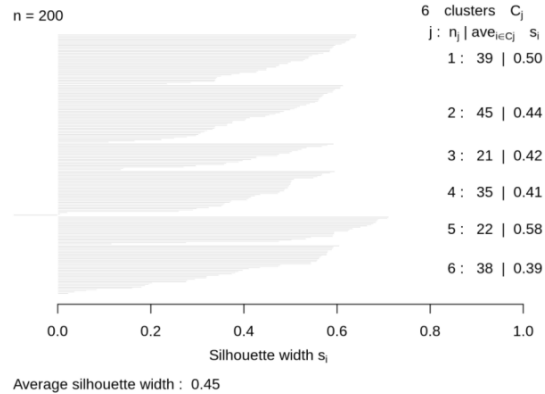
Silhouette plot of (x = k4\$cluster, dist = dist(customer\_data[, 3:5],



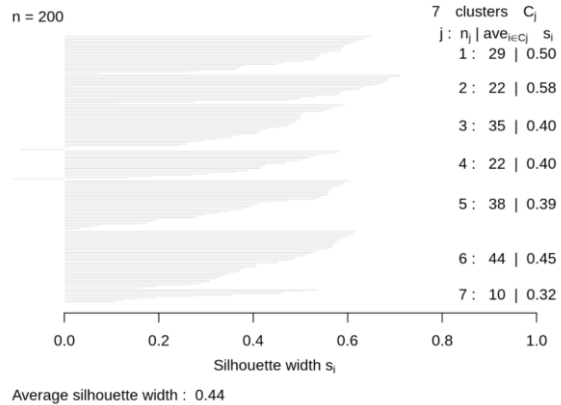
Silhouette plot of (x = k5\$cluster, dist = dist(customer\_data[, 3:5],



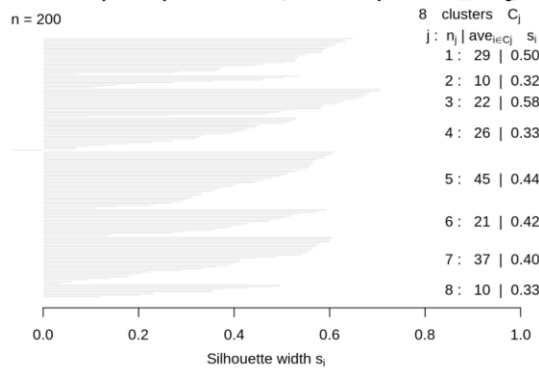
Silhouette plot of (x = k6\$cluster, dist = dist(customer\_data[, 3:5],



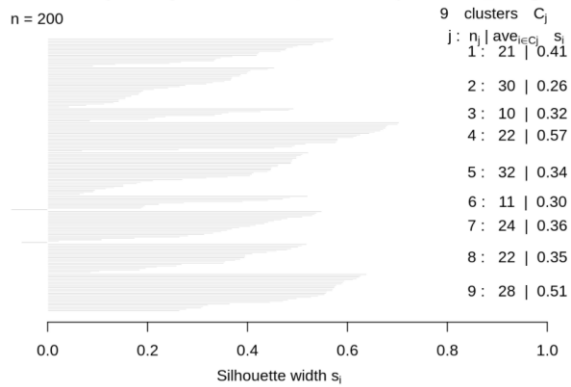
Silhouette plot of (x = k7\$cluster, dist = dist(customer\_data[, 3:5],



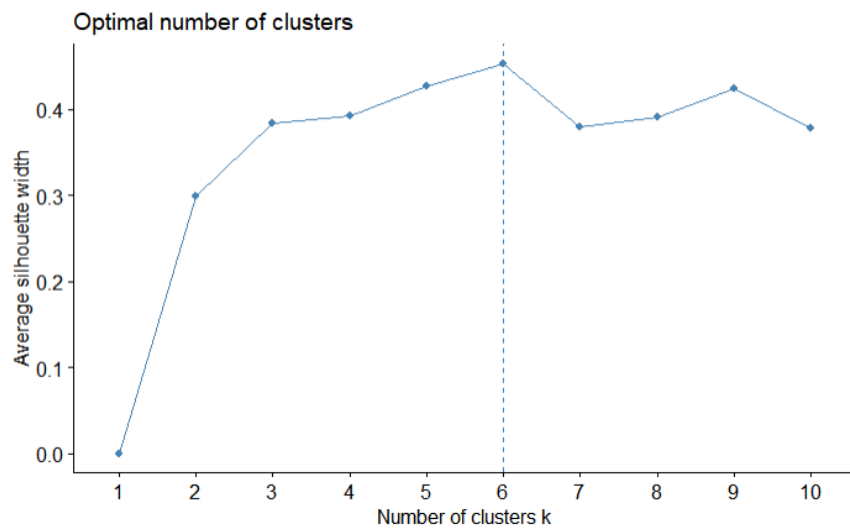
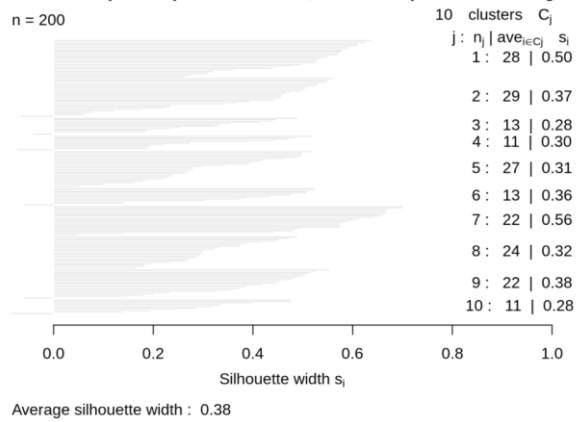
Silhouette plot of (x = k8\$cluster, dist = dist(customer\_data[, 3:5],

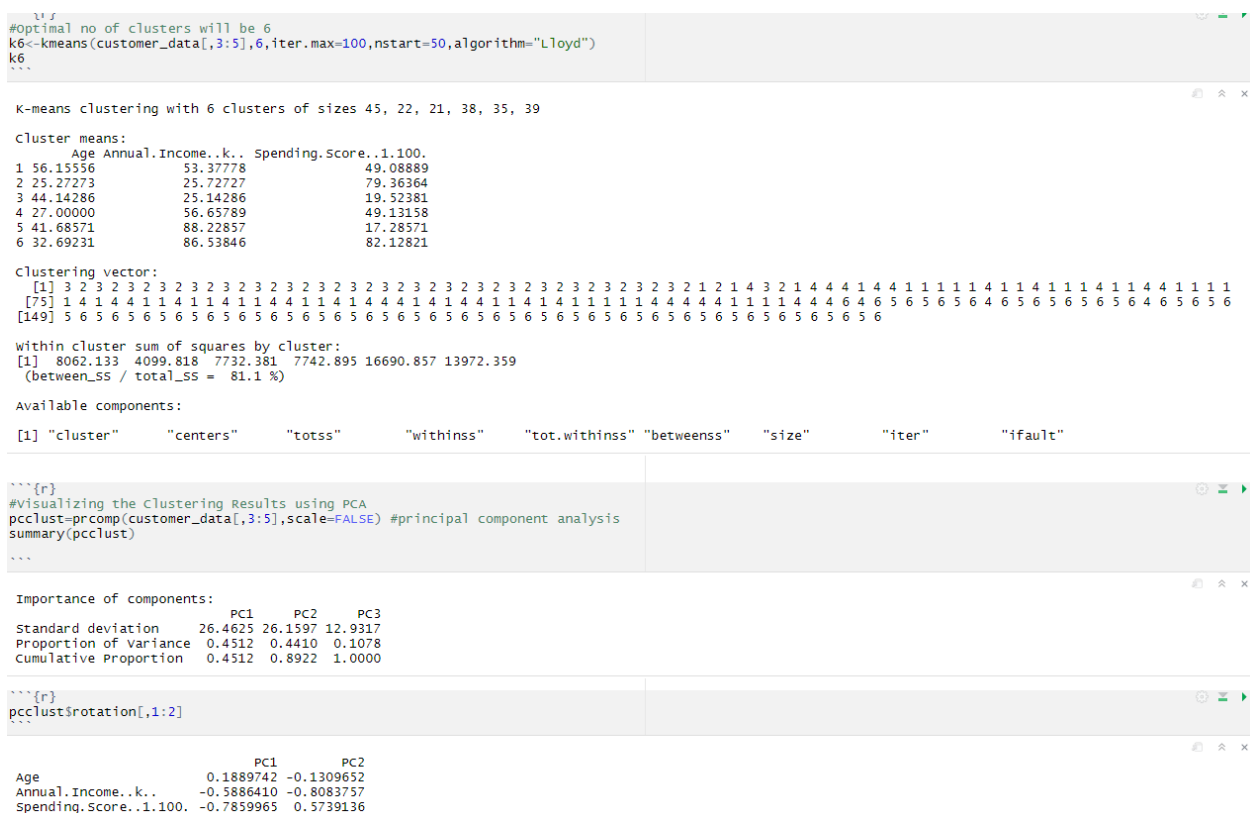


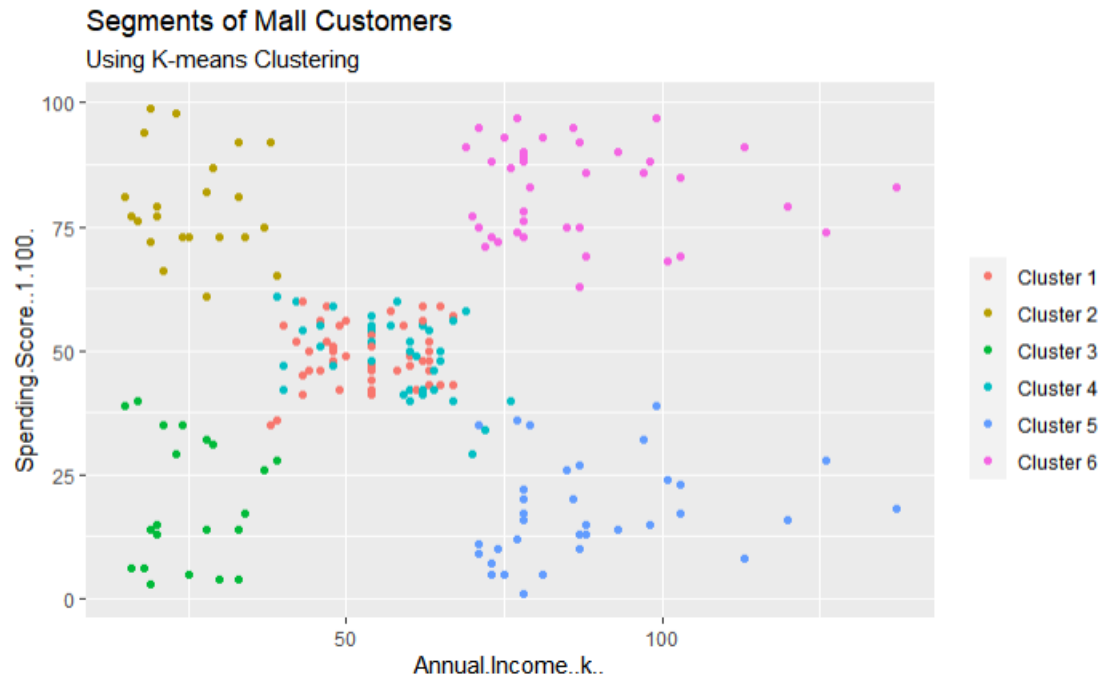
Silhouette plot of (x = k9\$cluster, dist = dist(customer\_data[, 3:5],



Silhouette plot of (x = k10\$cluster, dist = dist(customer\_data[, 3:5],







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

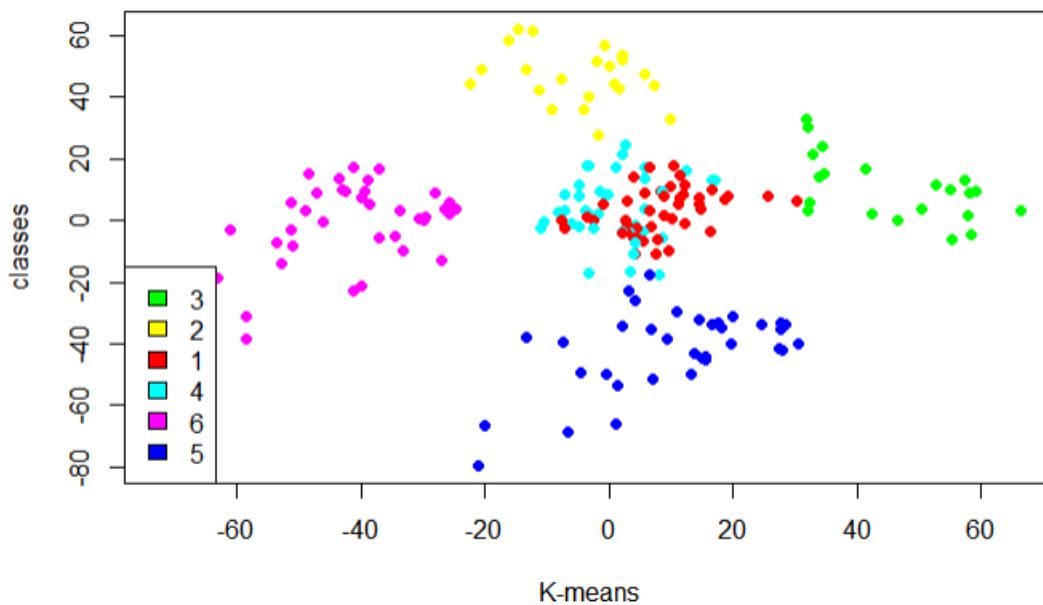
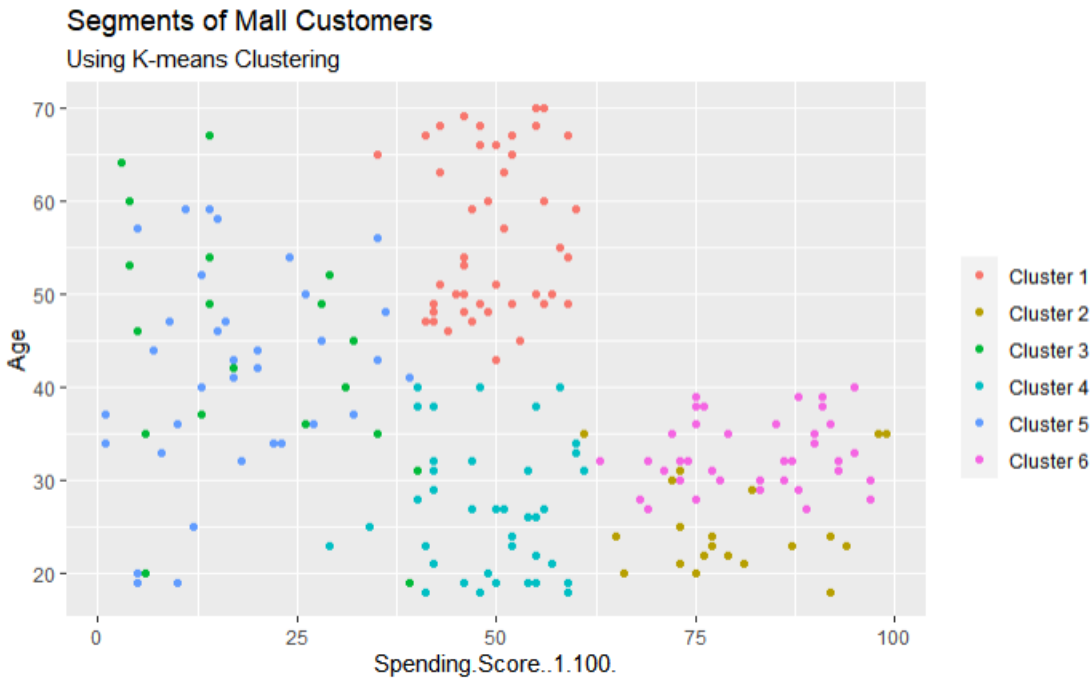
Cluster 1 and 4 – These clusters represent the customer\_data with the medium income salary as well as the medium annual spend of salary.

Cluster 6 – 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 income spend.

Cluster 5 – This cluster denotes a high annual income and low yearly spend.

Cluster 2 – This cluster represents a low annual income but its high yearly expenditure.



Final result:

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.

## **17.TESTING DETAILS**

Software Testing is a method to check whether the actual software product matches expected requirements and to ensure that software product is Defect free. It involves execution of software/system components using manual or automated tools to evaluate one or more properties of interest. The purpose of software testing is to identify errors, gaps or missing requirements in contrast to actual requirements. The project uses Program driven development. In this the program is written first then it's tested using various methods which are as follows:

1. Unit testing-This software testing approach is followed by the programmer to test the unit of the program. It helps developers to know whether the individual unit of the code is working properly or not. All the units have been tested and shown below.
2. Component Testing- It is mostly performed by developers after the completion of unit testing. Component Testing involves testing of multiple functionalities as a single code and its objective is to identify if any defect exists after connecting those multiple functionalities with each other. All the components have been tested and shown below and are working properly.
3. Positive Testing- Positive testing is a type of testing which is performed on a software application by providing the valid data sets as an input. It checks whether the software application behaves as expected with positive inputs or not. Positive testing is performed in order to check whether the software application does exactly what it is expected to do. The dataset is valid and tested in the program and works fine.
4. Validation Testing- Validation Testing ensures that the product actually meets the client's needs. It can also be defined as to demonstrate that the product fulfills its intended use when deployed on appropriate environment. The entire system is tested in the end together which gives the final clusters.
5. Negative Testing- Testers having the mindset of “attitude to break” and using Negative Testing they validate that if system or application breaks. A Negative Testing technique is performed using incorrect data, invalid data or input. It validates that if the system throws an error of invalid input and behaves as expected. The changes in the dataset have been made and tested in the program and it gives inaccurate outputs which have been shown.



The dataset is imported and also correctly displayed-

```
{r}
#import dataset
customer_data=read.csv("C:/Users/ASUS/Desktop/SEM V/SE/Project/Mall_Customers.csv")
str(customer_data)

'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 ...
```

The column names are correctly displayed-

```
{r}
#print column names
names(customer_data)

[1] "customerID"      "Gender"          "Age"             "Annual.Income..k.." "Spending.Score..1.100."
```

The summary and head are correctly displayed-

```
{r}
#display the first six rows of our dataset
head(customer_data)
#output the summary
summary(customer_data$Age)

data.frame
6 x 5

R Console

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

The summary and standard deviation are correctly displayed-

```
{r}
sd(customer_data$Age)
summary(customer_data$Annual.Income..k..)
sd(customer_data$Annual.Income..k..)
summary(customer_data$Age)

[1] 13.96901
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 15.00  41.50   61.50   60.56  78.00   137.00
[1] 26.26472
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 18.00  28.75   36.00   38.85  49.00   70.00

{r}
sd(customer_data$Spending.Score..1.100.)

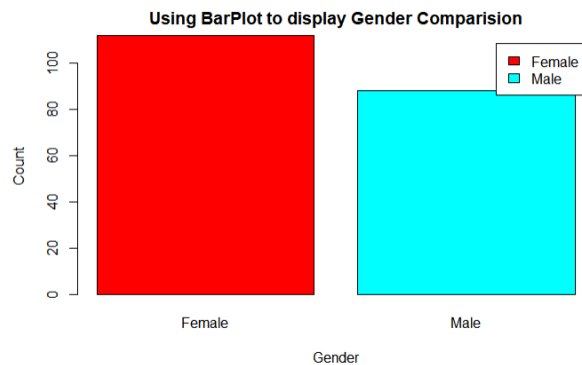
[1] 25.82352
```

The gender barplot and pieplot is plotted correctly and can be visualized-

```

[[[r]
#gender visualization using barplot
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparison",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
]]]

```



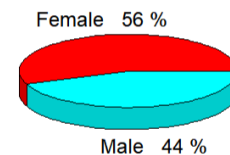
```

[[[r]
install.packages("plotrix")
pct=round(a/sum(a)*100)
lbs=paste(c("Female","Male")," ",pct,"%",sep=" ")
#gender visualization using pie-plot
library(plotrix)
pie3D(a,labels=lbs,
      main="Pie Chart Depicting Ratio of Female and Male")
]]]

```



**Pie Chart Depicting Ratio of Female and Male**

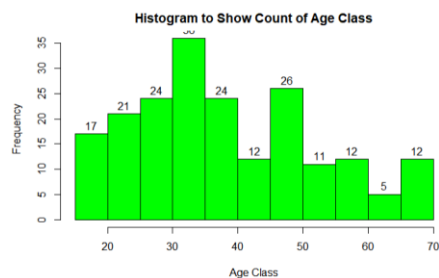


The age histogram and boxplot is plotted correctly and can be visualized-

```

[[[r]
#age visualization using histogram
hist(customer_data$Age,col="green",main="Histogram to show Count of Age Class", xlab="Age Class", ylab="Frequency",labels=TRUE)
]]]

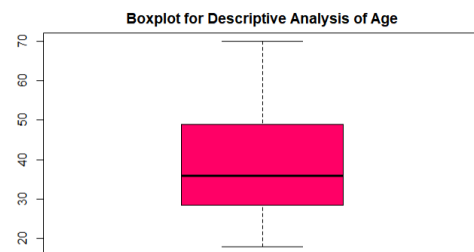
```



```

[[[r]
#age visualization using boxplot
boxplot(customer_data$Age,
        col="#ff0066",
        main="Boxplot for Descriptive Analysis of Age")
]]]

```

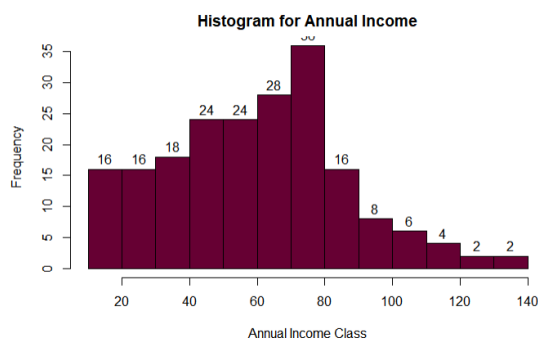


The annual income histogram and density plot is plotted correctly and can be visualized-

```

[[[r]
#annual income visualization using histogram
hist(customer_data$Annual.Income.k.,
      col="#660033",
      main="Histogram for Annual Income",
      xlab="Annual Income Class",
      ylab="Frequency",
      labels=TRUE)
]]]

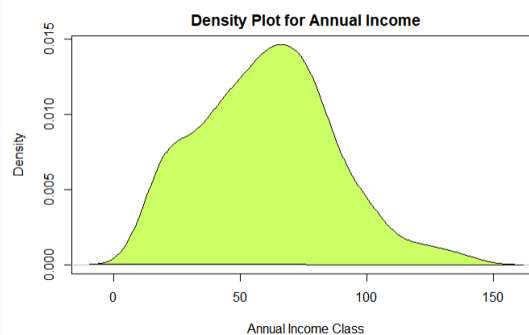
```



```

[[[r]
#annual income visualization using density plot
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")
]]]

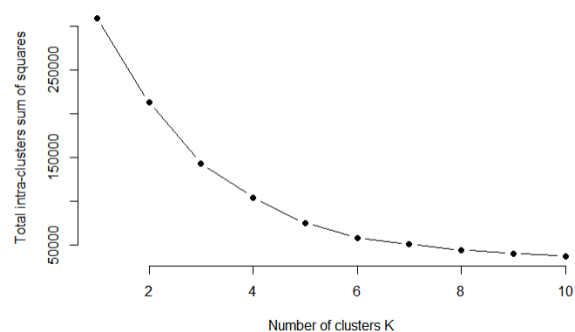
```



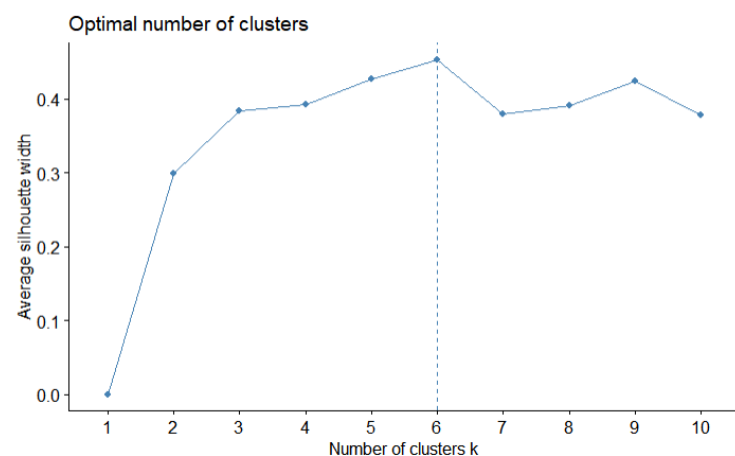
The spending score histogram and boxplot is plotted correctly and can be visualized-



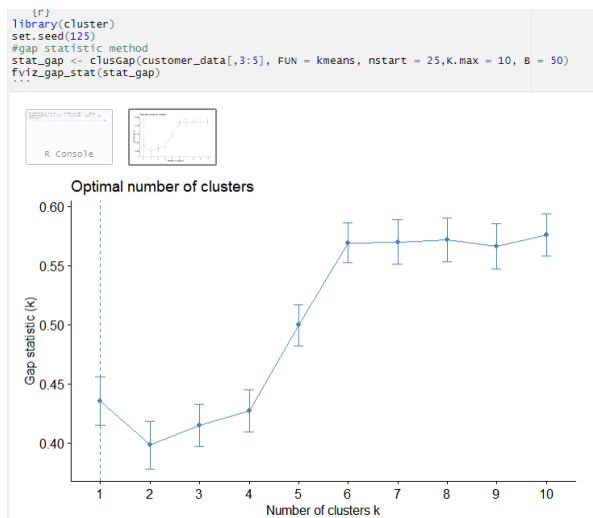
The total intra cluster sum of squares vs no of clusters graph using dataset comes out to be correct-



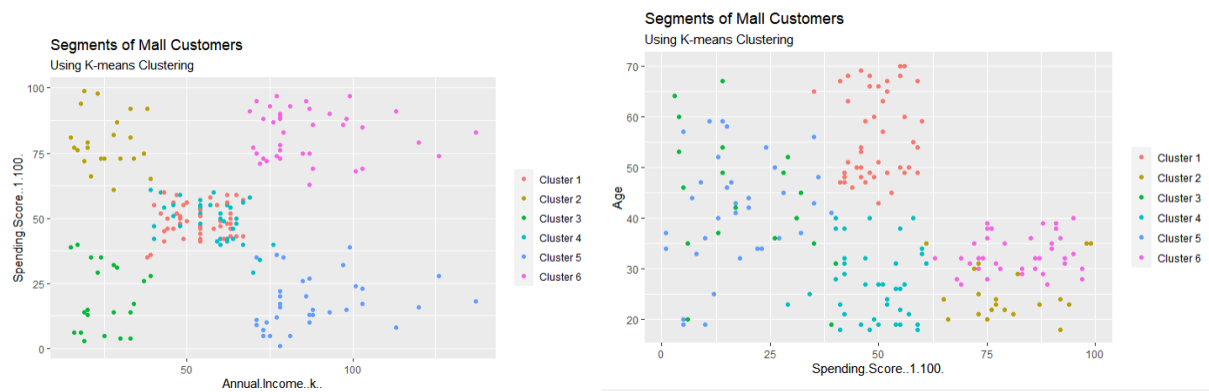
All the silhouette plots come out be accurate and correct and can given optimal clusters-



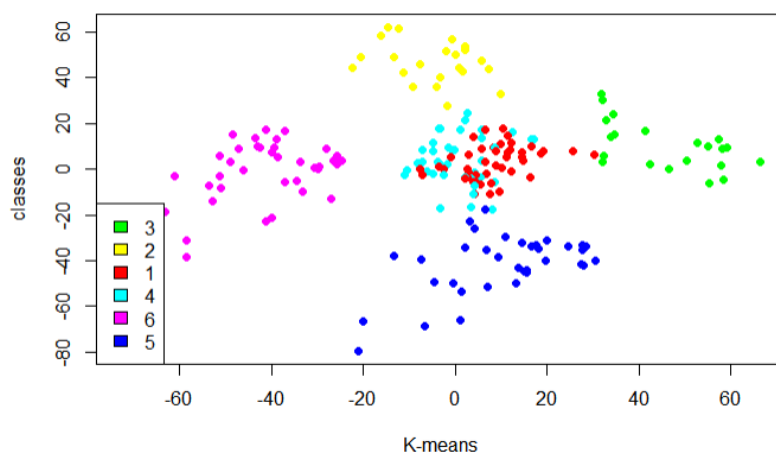
The gap statistic method works fine-



The clusters graph between spending score vs annual income and age vs spending score graph comes out to be readable-



The final clustering operation works fine and provide the correct output-



## TEST CASES:

TEST SCENARIO ID	TEST SCENARIO DESCRIPTION	TEST CASE ID	TEST CASE DESCRIPTION	PREREQUISITES	TEST STEPS	EXPECTED RESULTS	ACTUAL RESULT	STATUS (PASS/ FAIL)
1.	The input age of customers in the dataset should be correct.	TC_AGE_01	Enter valid and real customer age.	Input Customer Data	Check age graphs	All the graphs and plots should be precisely represented.	All the graphs and plots should be precisely represented.	PASS
		TC_AGE_02	Enter invalid and vague customer age like zero,thousands or in decimals.	Input Customer Data	Check age graphs	The graphs and plots should be inaccurate.	The graphs and plots should be inaccurate.	PASS

### TC\_AGE\_01

The Dataset :

CustomerID	Gender	Age	Annual
1	Male	19	
2	Male	21	
3	Female	20	
4	Female	23	
5	Female	31	
6	Female	22	
7	Female	35	
8	Female	23	
9	Male	64	
10	Female	30	
11	Male	67	
12	Female	35	
13	Female	58	
14	Female	24	
15	Male	37	
16	Male	22	
17	Female	35	
18	Male	20	
19	Male	52	
20	Female	35	
21	Male	35	
22	Male	25	

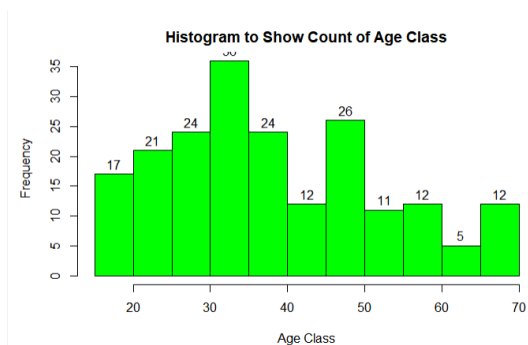
The dataset is valid.

### TC\_AGE\_02

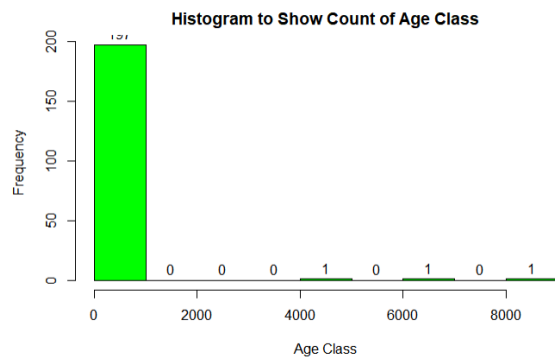
A	B	C	
CustomerID	Gender	Age	Annual Incc
1	Male	19	
2	Male	121	
3	Female	20	
4	Female	23	
5	Female	31	
6	Female	623	
7	Female	35	
8	Female	23	
9	Male	64	
10	Female	0	
11	Male	67	
12	Female	35	
13	Female	58	
14	Female	0.55	
15	Male	37	
16	Male	22	
17	Female	35	
18	Male	20	
19	Male	655	
20	Female	35	
21	Male	35	
22	Male	4567	
23	Female	46	

The age column is filled with invalid values.

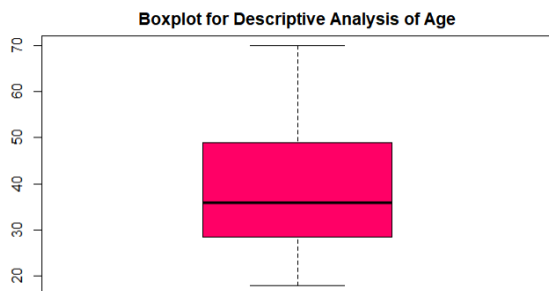
The variations in both the test cases are shown below:



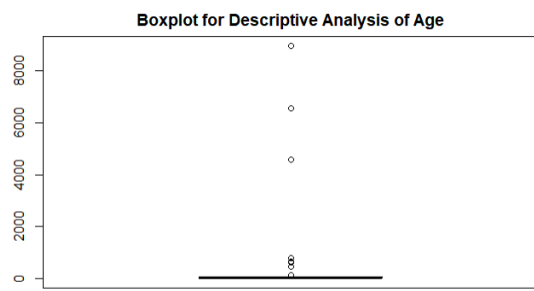
The histogram looks fine.



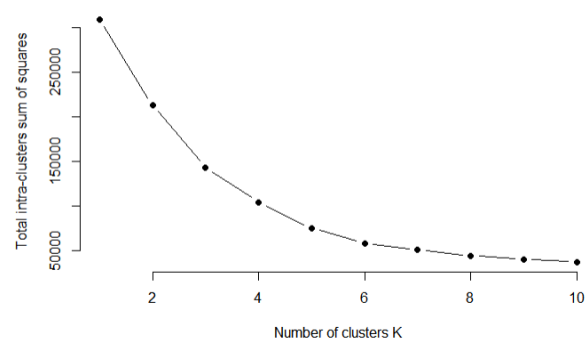
The histogram looks inaccurate.



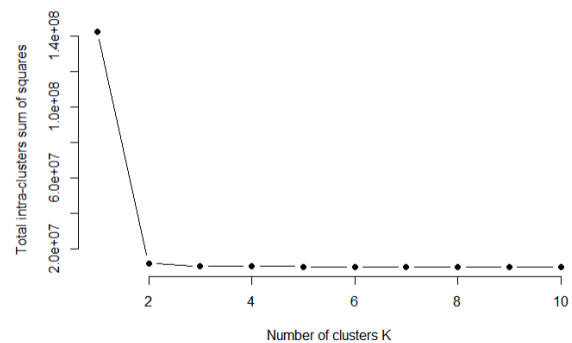
The boxplot looks fine.



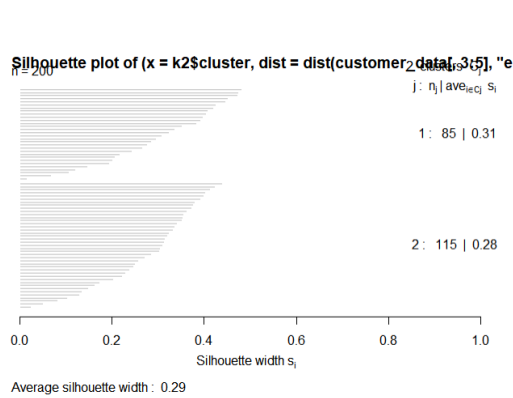
The boxplot looks inaccurate.



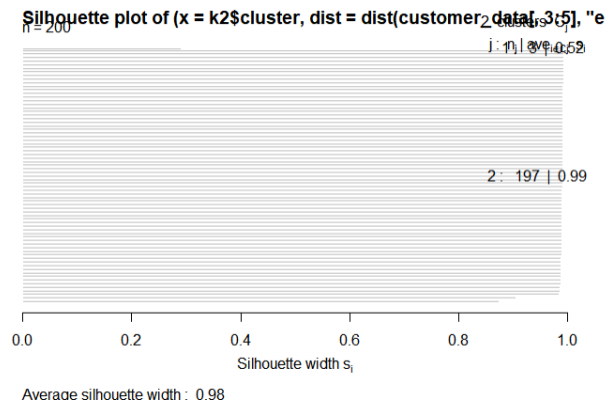
The graph looks fine.



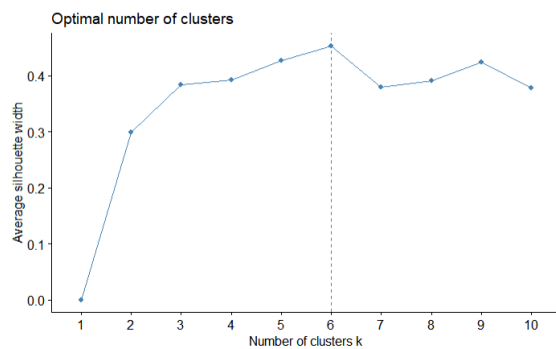
The graph comes out to be inaccurate.



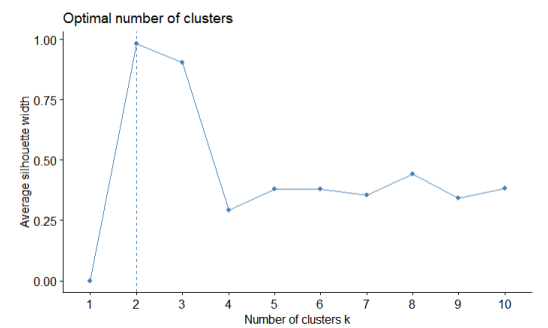
The silhouette plot looks fine.



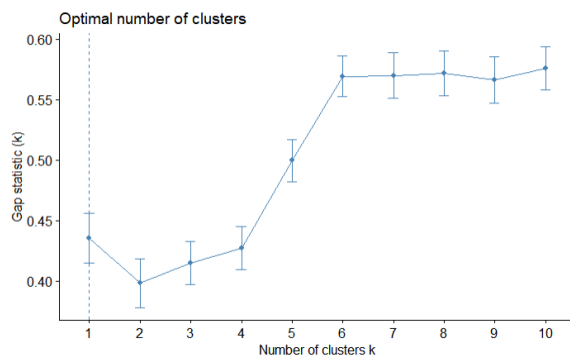
The silhouette plot looks inaccurate.



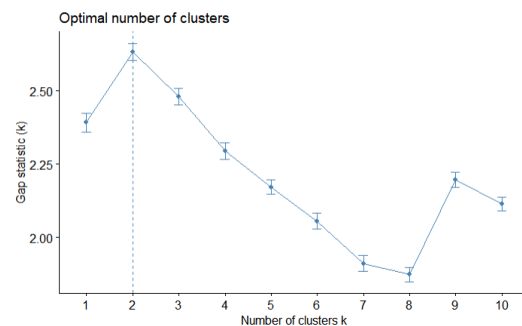
The silhouette plot graph looks fine.



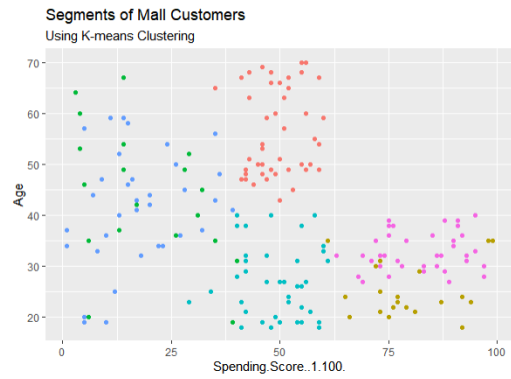
The silhouette plot graph looks inaccurate.



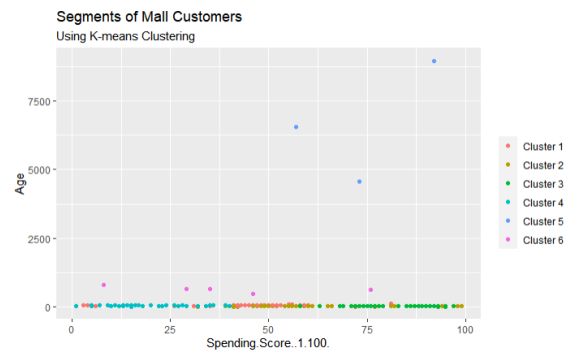
The gap statistic method looks fine.



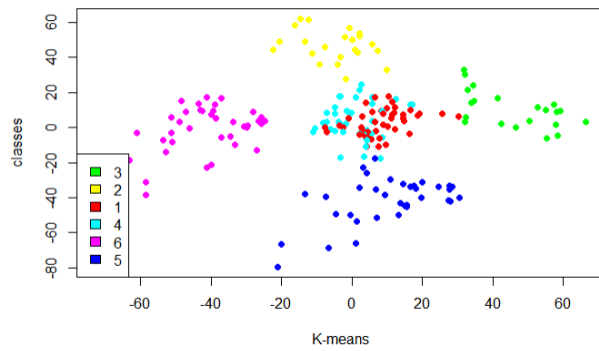
The gap statistic method looks inaccurate.



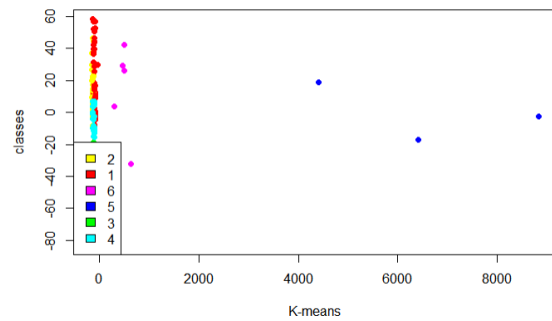
The graph is correct.



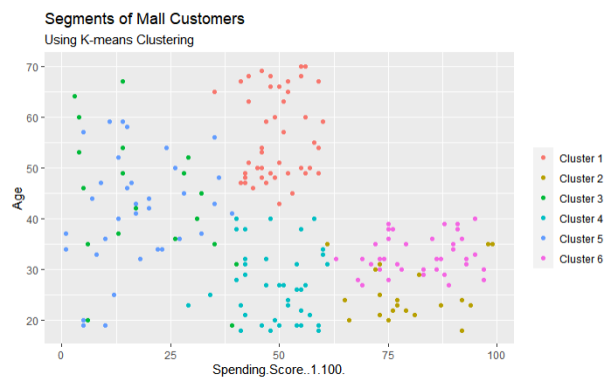
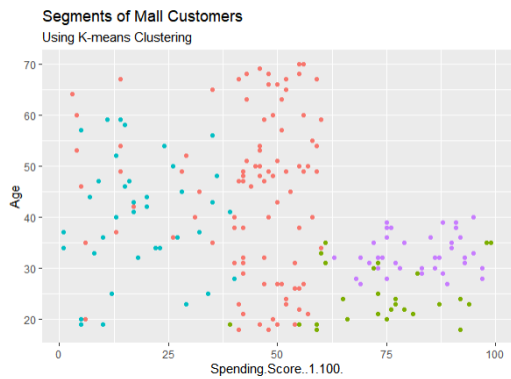
The graph is incorrect.



The graph is correct.



The graph is incorrect.





TEST SCENARIO ID	TEST SCENARIO DESCRIPTION	TEST CASE ID	TEST CASE DESCRIPTION	PRE REQUISITES	TEST STEPS	EXPECTED RESULTS	ACTUAL RESULT	STATUS (PASS/ FAIL)
2.	The summary of customer age should be adequately displayed	TC_SUMMARY_01	Enter valid and real customer age.	Input Customer Data	1.Check summary if it correct.	The summary should contain accurate values.	The summary should contain accurate values.	PASS
		TC_SUMMARY_02	Enter invalid and vague customer age like zero,thousands or in decimals.	Input Customer Data	1.Check summary if it incorrect.	The summary should contain inaccurate values.	The summary should contain inaccurate values.	PASS

TC\_SUMMARY\_01

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

TC\_SUMMARY\_02

Min. 1st Qu. Median Mean 3rd Qu. Max.  
0.00 28.00 36.00 152.69 49.25 8978.00

The max and min comes within age limits.

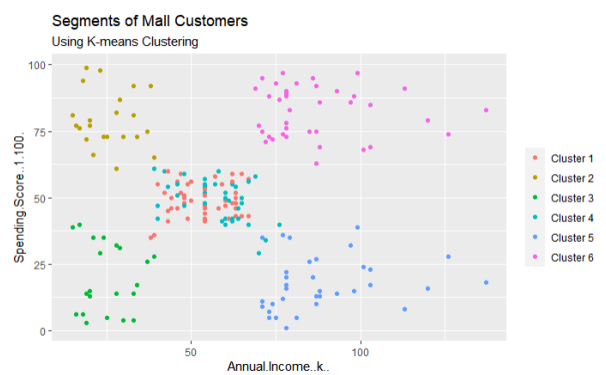
The max and min comes out of age limits.

TEST SCENARIO ID	TEST SCENARIO DESCRIPTION	TEST CASE ID	TEST CASE DESCRIPTION	PREREQUISITES	TEST STEPS	EXPECTED RESULTS	ACTUAL RESULT	STATUS (PASS/ FAIL)
3.	The number of clusters formed should be optimal	TC_CLUSTER_1	Optimal cluster detection from elbow graph and nbclust.	Kmeans clustering has to be applied	1.Chec k the graph 2.Look for the bent in graph	The final clusters are easy to comprehend	The final clusters are easy to comprehend	PASS
		TC_CLUSTER_2	Read an inaccurate	Kmeans clustering has	1.Chec k the	The final clusters are	The final clusters are	PASS

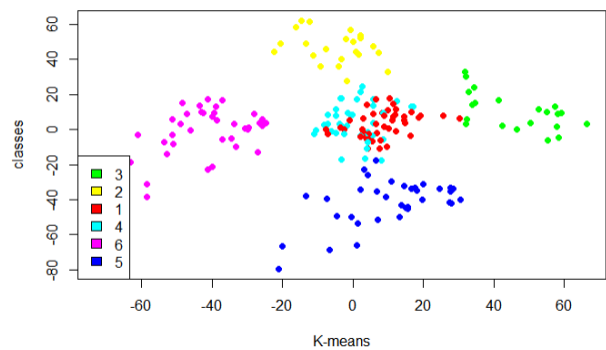
			value from the graph	to be applied	graph 2.Look for a wrong value of optimal clusters.	difficult to comprehend	difficult to comprehend	
--	--	--	----------------------	---------------	---	-------------------------	-------------------------	--

TC\_CLUSTER\_1

The optimal clusters taken here are 6.



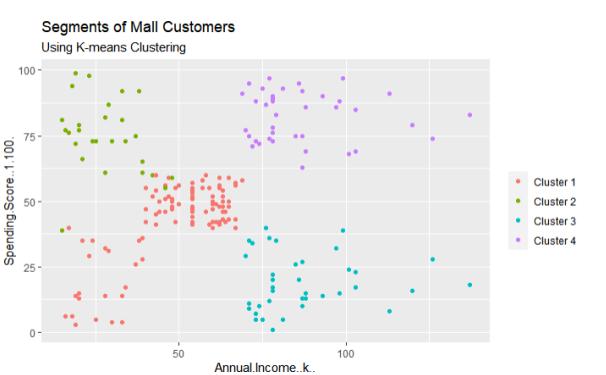
The graph is easy to comprehend.



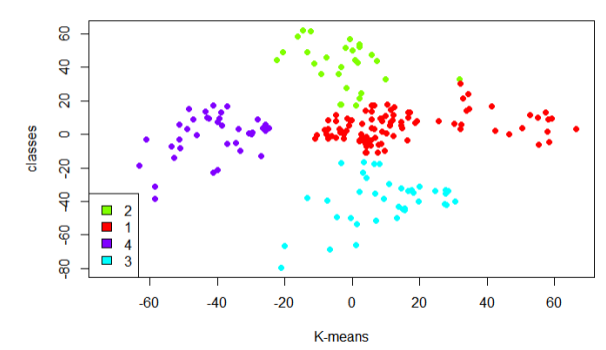
The graph is easy to comprehend.

TC\_CLUSTER\_2

The optimal clusters taken here are 4.



The graph is difficult to comprehend.

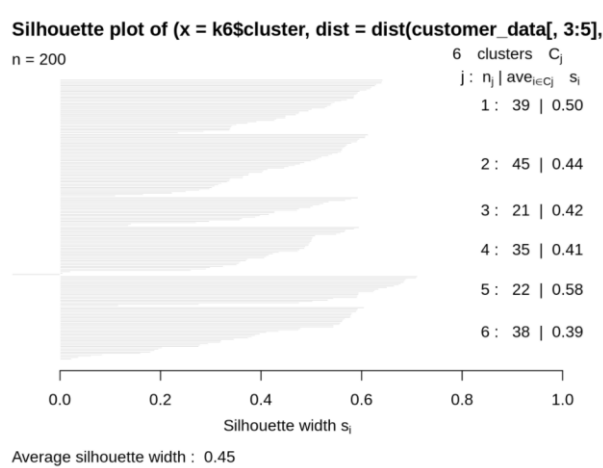


The graph is difficult to comprehend.

TEST SCENARIO ID	TEST SCENARIO DESCRIPTION	TEST CASE ID	TEST CASE DESCRIPTION	PREREQUISITES	TEST STEPS	EXPECTED RESULTS	ACTUAL RESULT	STATUS (PASS/ FAIL)
4.	Measure the quality of Clustering	TC_QUALITY_01	Determine how well within the cluster is the data object if <b>optimal cluster is chosen.</b>	Kmeans and Average Silhouette Method has to applied	1.Look at the graph and reading beneath it.	Obtain a high average silhouette width	Obtained a high average silhouette width	PASS
		TC_QUALITY_02	Determine how well outside the cluster is the data object if <b>optimal cluster is not chosen</b>	Kmeans and Average Silhouette Method has to applied	1.Look at the graph and reading beneath it.	Obtain a low average silhouette width	Obtained a low average silhouette width	PASS

TC\_QUALITY\_1

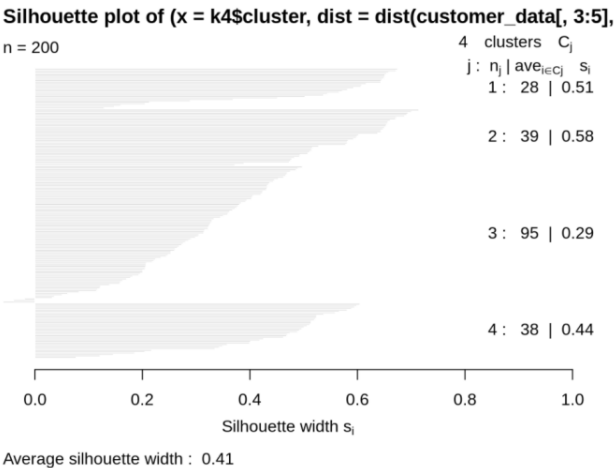
Optimal cluster-6



The avg width is 0.45.

TC\_QUALITY\_2

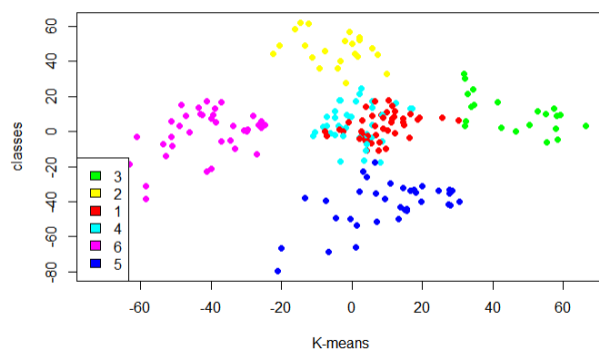
Optimal cluster-4



The avg. width is 0.41.

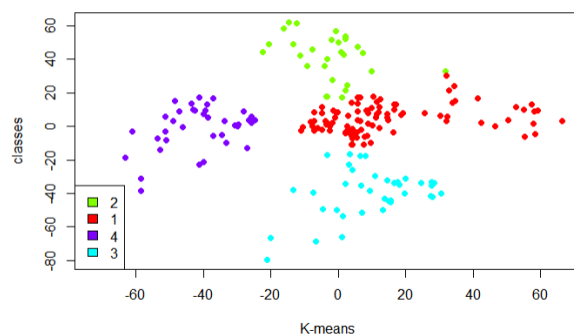
TEST SCENARIO ID	TEST SCENARIO DESCRIPTION	TEST CASE ID	TEST CASE DESCRIPTION	PREREQUISITES	TEST STEPS	EXPECTED RESULTS	ACTUAL RESULT	STATUS (PASS/FAIL)
5.	Visualize the clustering results	TC_VISUALIZE_01	Construct a correct ggplot showing segments if optimal cluster chosen.	Applied first two principal components Chosen optimal cluster	1.Check for the different segments and infer from them in the graph	Inference of segments comes out to be accurate	Inference of segments comes out to be accurate	PASS
		TC_VISUALIZE_02	Construct a incorrect ggplot showing incorrect segments if optimal cluster not chosen.	Applied first two principal components Optimal cluster is not chosen.	1.Check for the different segments and infer from them in the graph	Inference of segments comes out to be inaccurate	Inference of segments comes out to be inaccurate	PASS

TC\_VISUALIZE\_1



The segments are easy to be inferred.

TC\_VISUALIZE\_2



The segments are merged and are difficult to infer.

## **18.CONCLUSION:**

Customer Segmentation is a very aspect for any business. Segmentation allows businesses to make better use of their marketing budgets, gain a competitive edge over rival companies and, importantly, demonstrate a better knowledge of your customers' needs and wants. Breaking down a large customer base into more manageable pieces, makes it easier to identify the target audience and launch campaigns to the most relevant people, using the most relevant channel. During the process of grouping customers into clusters, one may find that you have identified a new market segment, which could in turn alter your marketing focus and strategy to fit. Once you have identified the key motivators for your customer, such as design or price or practical needs, you can brand your products appropriately. Using segmentation, marketers can identify groups that require extra attention and those that churn quick, along with customers with the highest potential value. It can also help with creating targeted strategies that capture your customers' attention and create positive, high-value experiences with your brands. Therefore Customer Segmentation is essential and a need in todays time for every business.

## **19.REFERENCES:**

1. For understanding format of SRS-“Dr.Preeti Mulay” - Unit3 lecture slides

2. For all the diagrams (DFD, Sequence diagram, Class diagram)-

[https://www.lucidchart.com/pages/examples/uml\\_diagram\\_tool](https://www.lucidchart.com/pages/examples/uml_diagram_tool)

3. For understanding of Requirements Document-

[http://web.cse.ohio-state.edu/~bair.41/616/Project/Example\\_Document/Req\\_Doc\\_Example.html](http://web.cse.ohio-state.edu/~bair.41/616/Project/Example_Document/Req_Doc_Example.html)

4. To understand writing style in SRS-

<https://www.geeksforgeeks.org/software-engineering-quality-characteristics-of-a-good-srs/>

5. For understanding of customer segmentation-

<https://www.imsmarketing.ie/business-strategy/the-importance-of-market-segmentation>

6.For understanding of Software Testing-

<https://www.softwaretestinghelp.com/types-of-software-testing/>