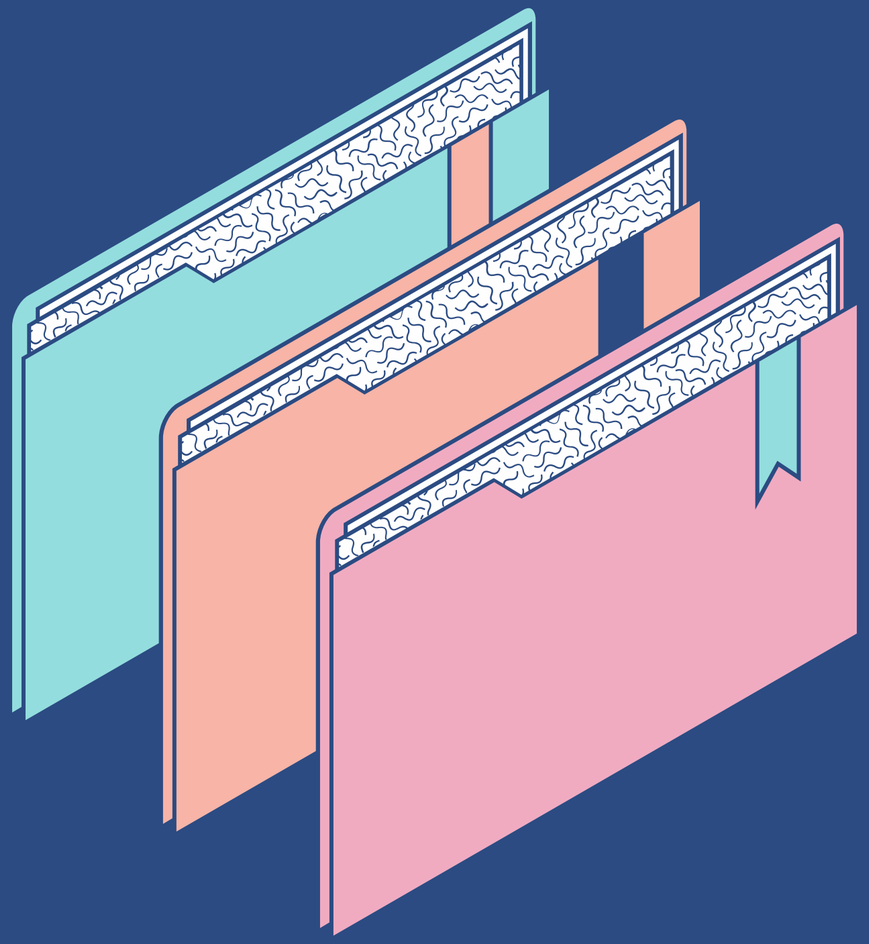
CUSTOMER SEGMENTATION

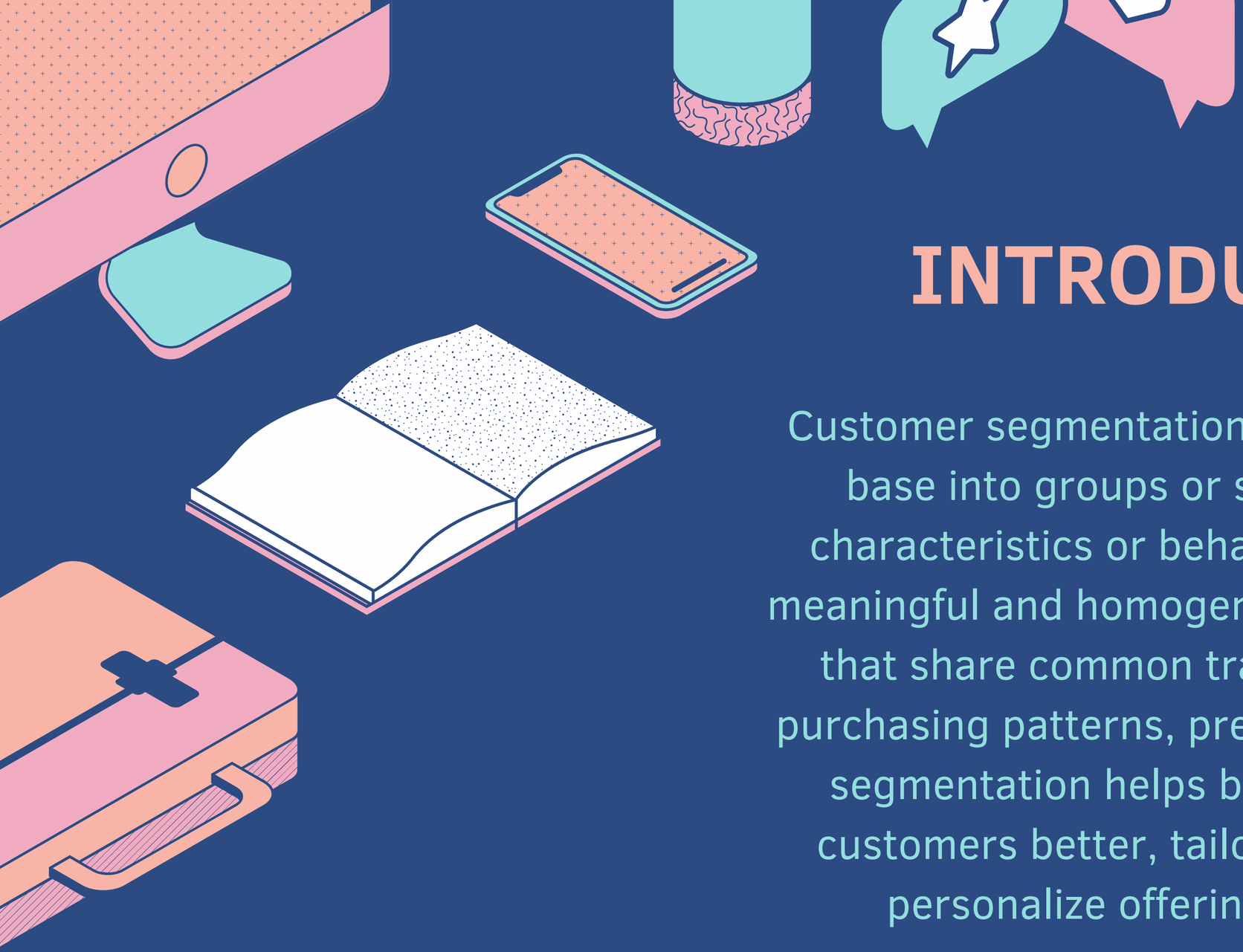
Group Members:

SWASTI NITYA (07001192022)
SURYANSHI (06901192022)
TANISHA (07101192022)



AIM

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

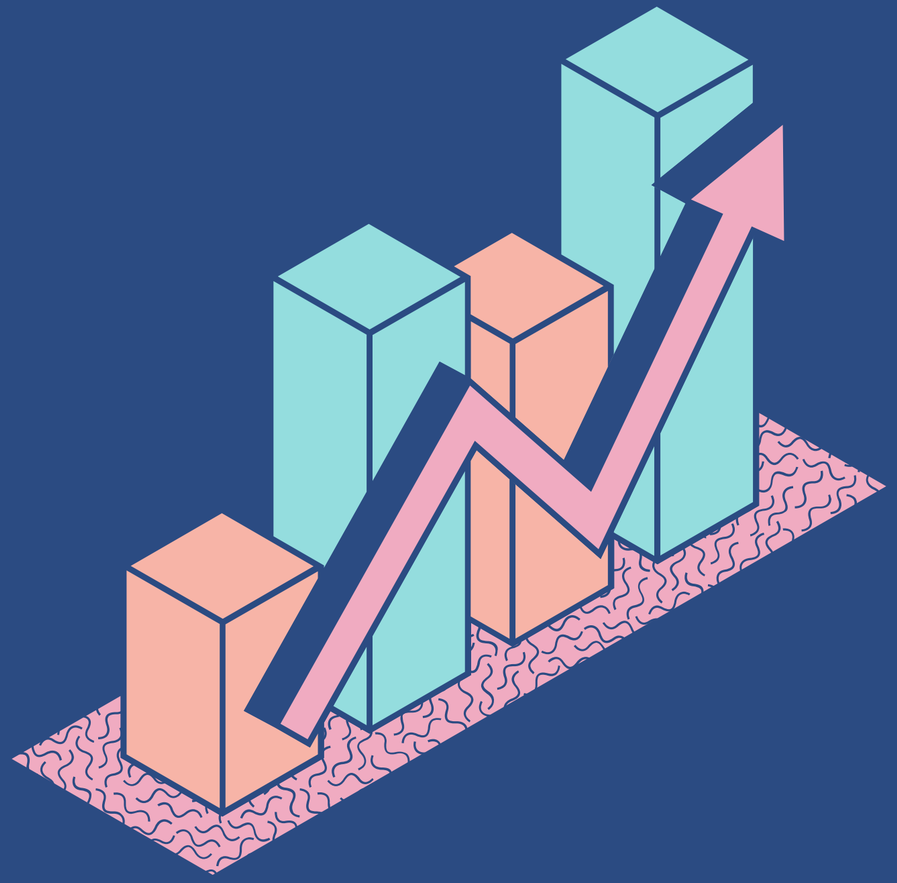


INTRODUCTION

Customer segmentation involves dividing a customer base into groups or segments based on similar characteristics or behaviors. The goal is to identify meaningful and homogeneous subgroups of customers that share common traits, such as demographics, purchasing patterns, preferences, or needs. Customer segmentation helps businesses understand their customers better, tailor marketing strategies, and personalize offerings to specific segments.

PROPOSED OUTCOME

The outcome of customer segmentation is to gain a comprehensive understanding of the customer base, identify actionable insights, and develop strategies to effectively target and serve different customer segments for improved business outcomes.



METHODOLOGY

DATA COLLECTION AND PREPARATION

Data Collection: The data is being collected from (https://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx) .

Data Preparation:

1.Load the dataset: Upload the "Mall_Customers.csv" file in working directory, use the read.csv() function to load the dataset in RStdio.

```
customer <- read_excel("C:/Users/ACER/Downloads/customer_segmentation.xls")
names(customer)
```

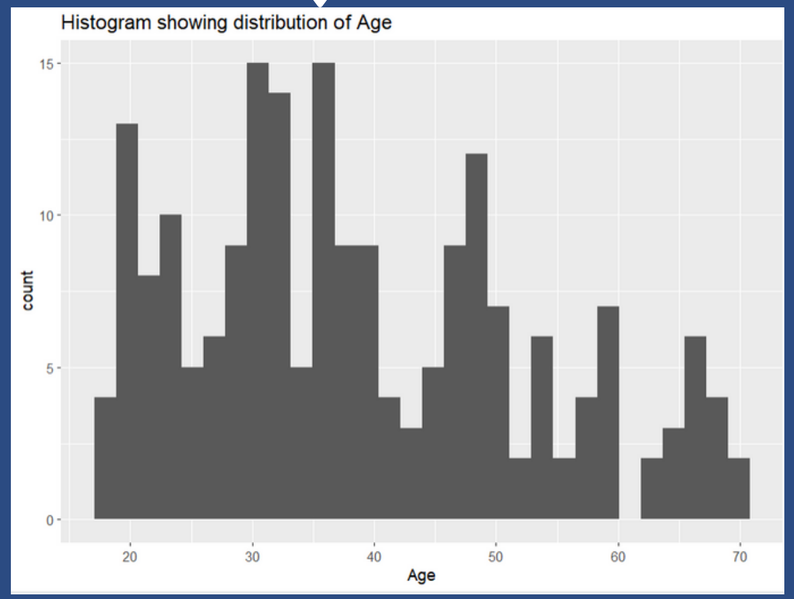
2.Explore the dataset:Use functions such as head(), summary(), and str() to gain an understanding of the dataset's structure, variables, and any missing values:

```
str(customer)
customer <- rename(customer, annual_income=Annual Income (k$) ,
                   spending_score=Spending Score (1-100))
summary(customer)
```

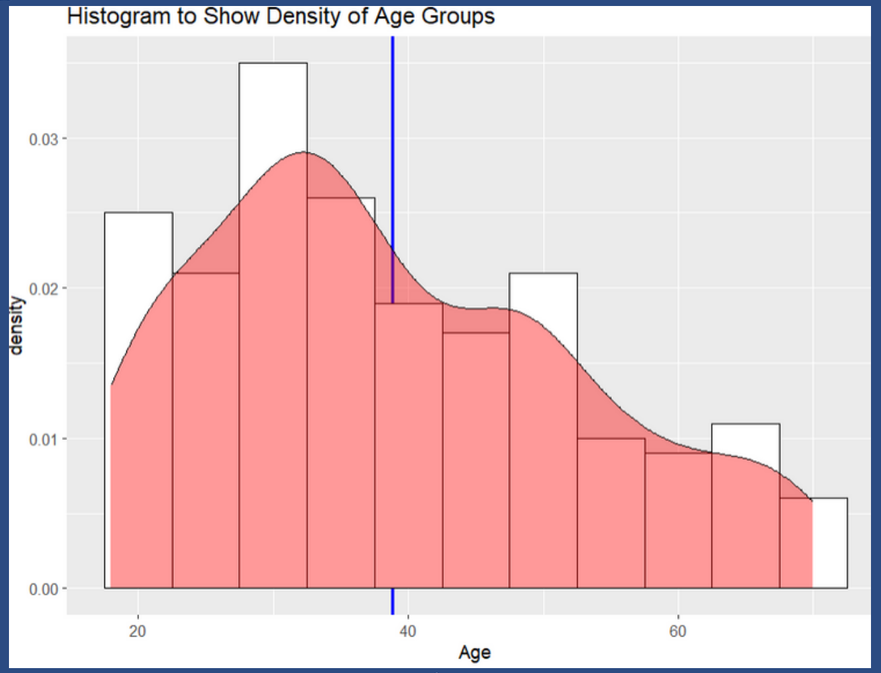
3. Data visualization: Use ggplot2 or other visualization libraries to explore relationships and patterns in the data.

By these histogram we conclude that the maximum customer ages are between 30-40 and the minimum is 18 whereas maximum age is 72.

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



Visualizing of Age Distribution

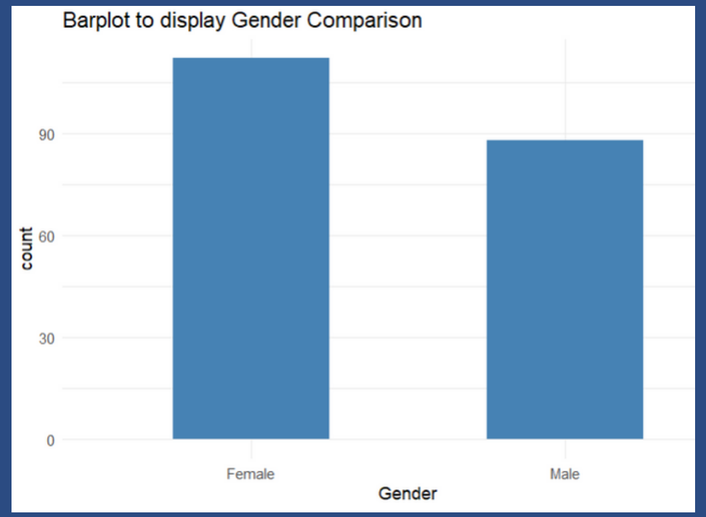



```
# Creating a histogram to show dispersion of mall customers based on age groups
ggplot(customer, aes(x = Age)) +
  geom_vline(aes(xintercept = mean(Age)), color = "blue",  #adding an intercept to indicate mean
             linetype = "dashed", size = 1.0) +
  geom_histogram(binwidth = 5, aes(y =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

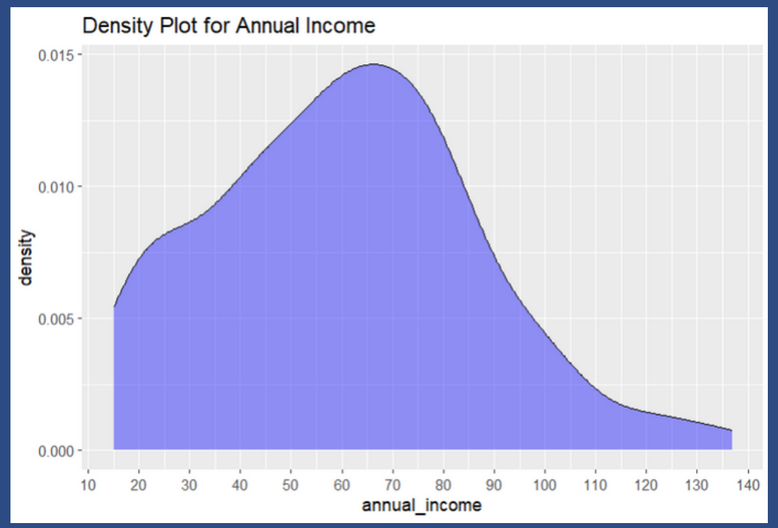
Customer Gender Visualization

From this barplot, we conclude that females customers are more than no. of males

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



Analysis of Annual Income of Customers

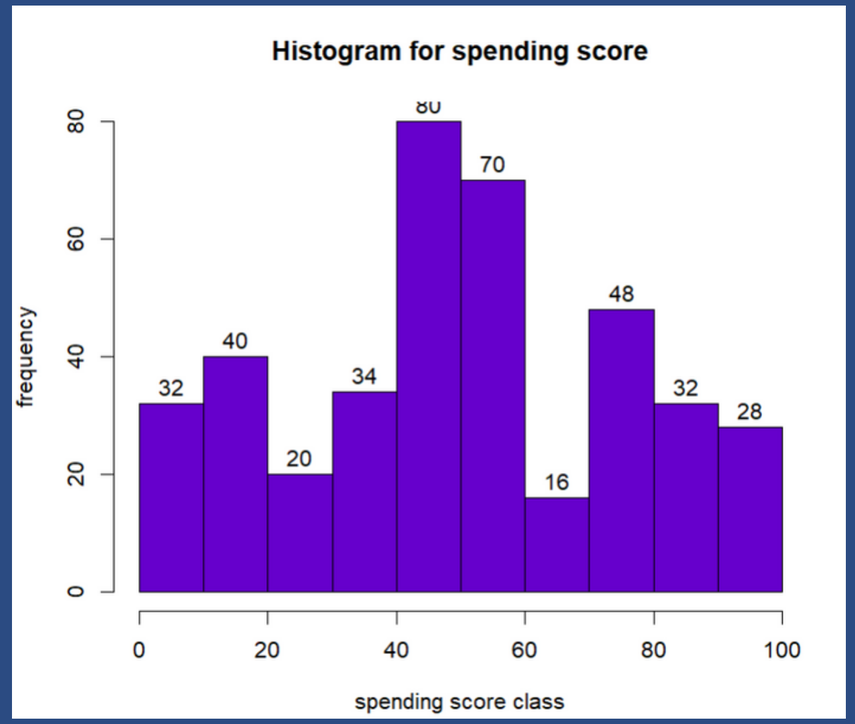



```
ggplot(customer, aes(x = annual_income)) +
  geom_density(alpha=0.4, fill="blue") +
  scale_x_continuous(breaks = seq(0, 200, by = 10)) +
  labs(title="Density Plot for Annual Income")
```

From this graph, we conclude that minimum annual income of customers is 15 and maximum is 137and people earning an average income of 70 has the highest frequency count.

Visualization of Spending score

```
hist(customer$spending_score,
     main="HistoGram for Spending Score",
     xlab="Spending Score Class",
     ylab="Frequency",
     col="#6600cc",
     labels=TRUE)
```



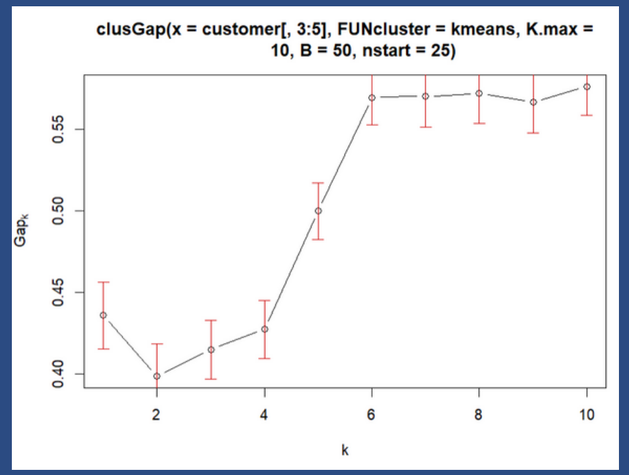
From the histogram, we conclude that customers between class 40 and 50 have the highest spending score among all the classes.

CONDUCTING THE CLUSTER ANALYSIS:

1.Choose the optimal clusters:

We use Gap statistics method to determine the optimal number of clusters .
The optimal no. of clusters taken is 6.

```
set.seed(125)
stat_gap<-clusGap(customer[,3:5], FUN=kmeans, nstart=25, K.max = 10, B=50)
plot(stat_gap)
```



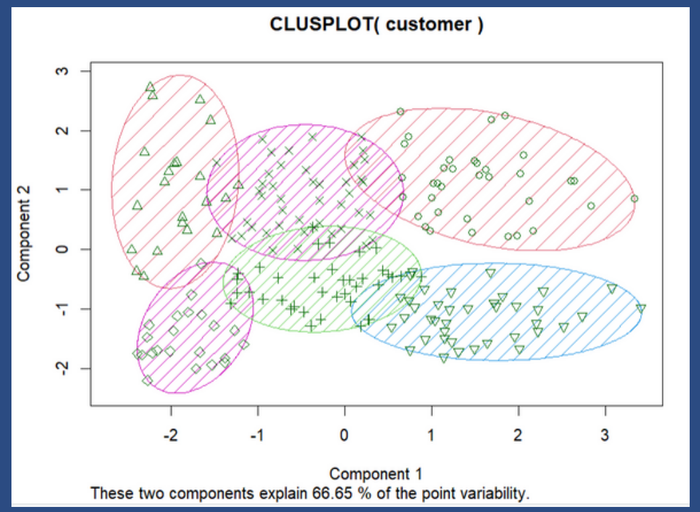
2.Creating the K-means Clustering model:

This model is created using 6 no. of clusters

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

```
K-means clustering with 6 clusters of sizes 35, 22, 38, 44, 22, 39

Cluster means:
       Age annual_income spending_score
1 41.68571       88.22857       17.28571
2 44.31818       25.77273       20.27273
3 27.00000       56.65789       49.13158
4 56.34091       53.70455       49.38636
5 25.27273       25.72727       79.36364
6 32.69231       86.53846       82.12821

Clustering vector:
  [1] 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 2 5 4 5 2 3 2
 [46] 5 4 3 3 3 4 3 3 4 4 4 4 4 3 4 4 3 4 4 4 3 4 4 3 3 4 4 4 4 4 3 4 3 3 4 4 3 4 4 3 4 4 3 3 4
 [91] 4 3 4 3 3 3 4 3 4 3 3 4 4 3 4 3 4 3 4 4 4 4 4 3 3 3 3 3 4 4 4 4 3 3 3 6 3 6 1 6 1 6 1 6 3 6 1
[136] 6 1 6 1 6 1 6 3 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6
[181] 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6 1 6

Within cluster sum of squares by cluster:
[1] 16690.857  8189.000  7742.895  7607.477  4099.818 13972.359
 (between_SS / total_SS =  81.1 %)

Available components:

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

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



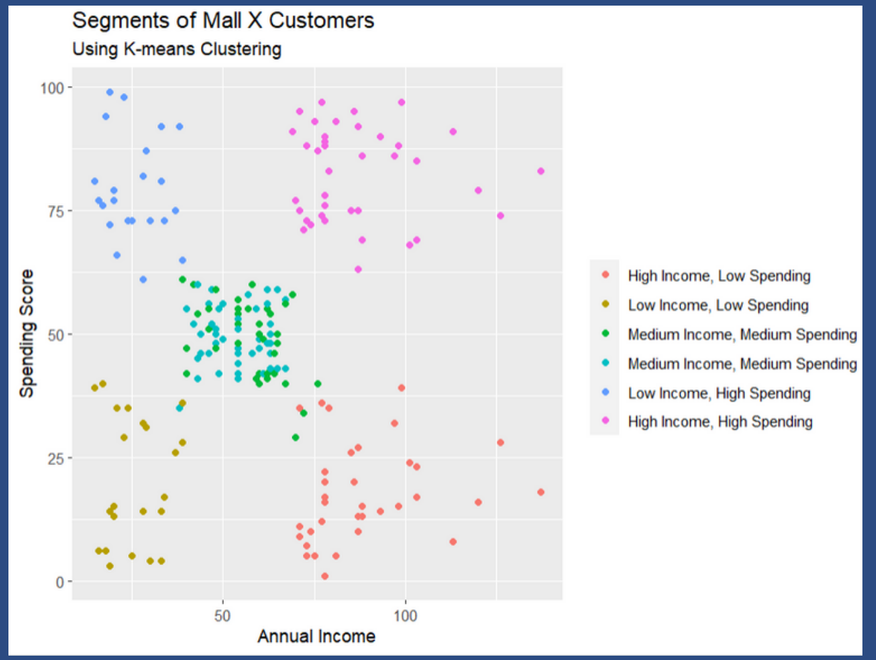

```
clusplot(customer, k6$cluster, color=TRUE, shade=TRUE, labels=0, lines=0)
```

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

```
pcclust<-prcomp(customer[, 3:5], scale=FALSE)
summary(pcclust)
pcclust$rotation[, 1:2]
```

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

```
ggplot(customer, aes(x = annual_income , y = spending_score)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$cluster))) +
  scale_color_discrete(name = " ",
  breaks=c("1", "2", "3", "4", "5","6"),
  labels=c("High Income, Low Spending", "Low Income, Low Spending", "Medium Income, Medium Spending",
  "Medium Income, Medium Spending", "Low Income, High Spending","High Income, High Spending")) +
  labs(x="Annual Income", y="Spending Score") +
  ggtitle("Segments of Mall X Customers",
          subtitle = "Using K-means Clustering")
```

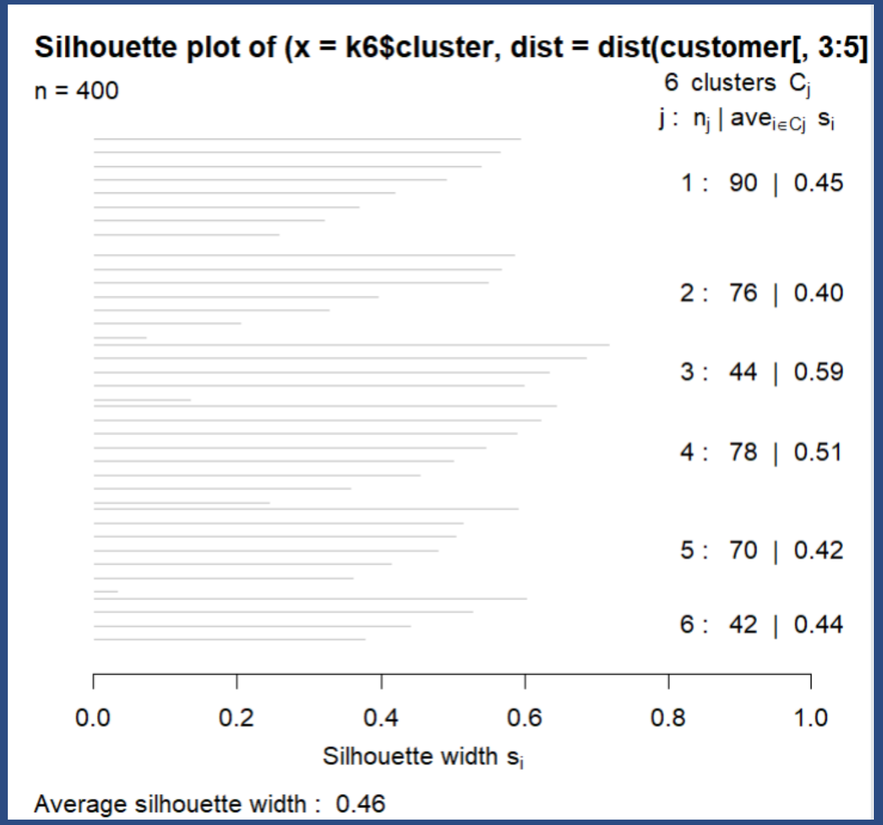


Analyze Quality of model

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

With the help of the average silhouette method, we can measure the quality of our clustering operation.If we obtain a high average silhouette width, it means that we have good clustering. The average silhouette method calculates the mean of silhouette observations for different k values.

In this model, the average silhouette width is 0.45 for k=6 which is higher than other for using different optimal no. of clusters.



Outcome

From this model, we observe that there are 6 types of customersfrom which we can get better understanding customer purchasing patterns and then use it for increasing sales are 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.

Thank you!