Customer Segmentation applying K-Means Clustering

Upasana Purohit

2023-05-13

Goal Of The Project:

Market segmentation is the process of dividing a larger market into smaller groups of consumers who have similar needs or characteristics. The goal of segmentation is to identify groups of consumers who are likely to respond to marketing efforts in a similar way, so that companies can tailor their marketing strategies to each group's specific needs and preferences.

Market segmentation can be based on a wide range of factors, such as demographic characteristics (e.g., age, gender, income), psychographic traits (e.g., personality, values, attitudes), behavioral patterns (e.g., buying habits, brand loyalty), or geographic location.

There are several benefits of market segmentation. By targeting specific segments of the market, companies can:Develop more focused marketing strategies that are tailored to each segment's needs and preferences,Improve the effectiveness of marketing campaigns by delivering more relevant messages to each segment,Increase customer satisfaction and loyalty by offering products and services that better meet their specific needs. Enhance profitability by avoiding the need to compete solely on price and instead offering products or services that command a premium due to their unique features or benefits.Overall, market segmentation is an important tool for companies to achieve more effective marketing and improve their bottom line.

```
# R packages that I have used for Data Analysis and Clustering.

library(dplyr)

library(ggplot2)

library(cluster)
```

Exploring Data

To explore the data, I will typically begin by loading it into R and examining its structure and summary statistics.

```
# Importing the "Mall_Customers.csv".
Customer_Malldata <- read.csv("/Users/upasanapurohit/Desktop/Mall_Customers.csv")
# Check the names of columns and structure of the dataset.
names(Customer_Malldata)
```

```
## [1] "CustomerID"             "Gender"                 "Age"
## [4] "Annual.Income..k.."     "Spending.Score..1.100."
```

```
str(Customer_Malldata)
```

```
## '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 ...
```

Renaming the variable, so that I can work with it easily. Then I will find a summary of the Data.

```
# Rename some columns names

Customer_Malldata <- rename(Customer_Malldata, Annual_Income =Annual.Income..k..,
                           Spending_Score = Spending.Score..1.100.)
# Summarise the data.

summary(Customer_Malldata)
```

```
##    CustomerID        Gender               Age        Annual_Income
##  Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00
##  1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50
##  Median :100.50   Mode  :character   Median :36.00   Median : 61.50
##  Mean   :100.50                      Mean   :38.85   Mean   : 60.56
##  3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00
##  Max.   :200.00                      Max.   :70.00   Max.   :137.00
##  Spending_Score
##  Min.   : 1.00
##  1st Qu.:34.75
##  Median :50.00
##  Mean   :50.20
##  3rd Qu.:73.00
##  Max.   :99.00
```

Descriptive Analyses.

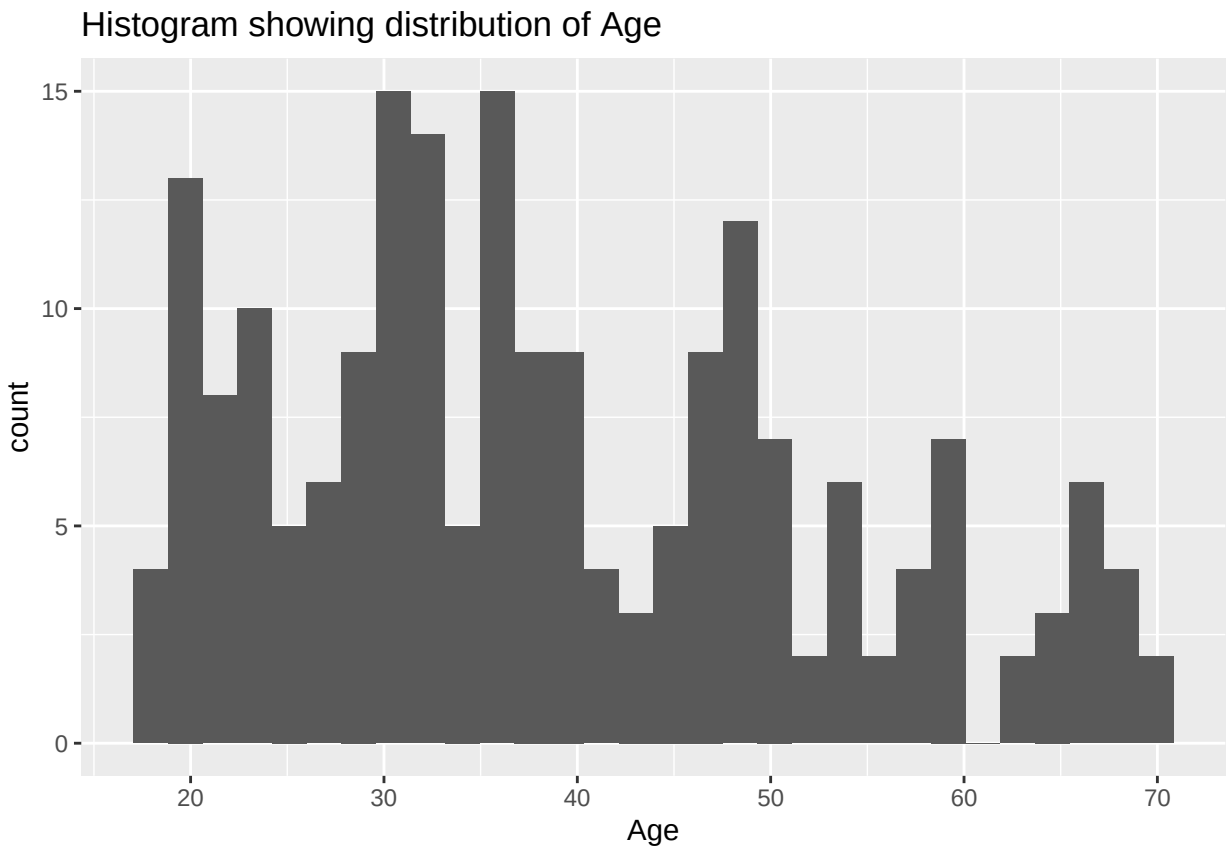
Descriptive analysis refers to a set of techniques used to summarize and describe the key features of a dataset, such as its central tendency, variability, and distribution. These techniques are used to understand the characteristics of the data and to identify patterns, trends, and outliers.

Descriptive analysis techniques include measures of central tendency (e.g., mean, median, mode), measures of variability (e.g., range, standard deviation, variance), and measures of distribution (e.g., skewness, kurtosis). Graphical methods, such as histograms, box plots, and scatter plots, are also commonly used to visualize the data and gain insights into its characteristics.

Descriptive analysis is often used as a first step in data analysis to explore the data and identify any potential issues or problems with the data that may need to be addressed before proceeding with further analysis. It can also be used to generate summary statistics and visualizations that can be used to communicate the key findings of the analysis to others.

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

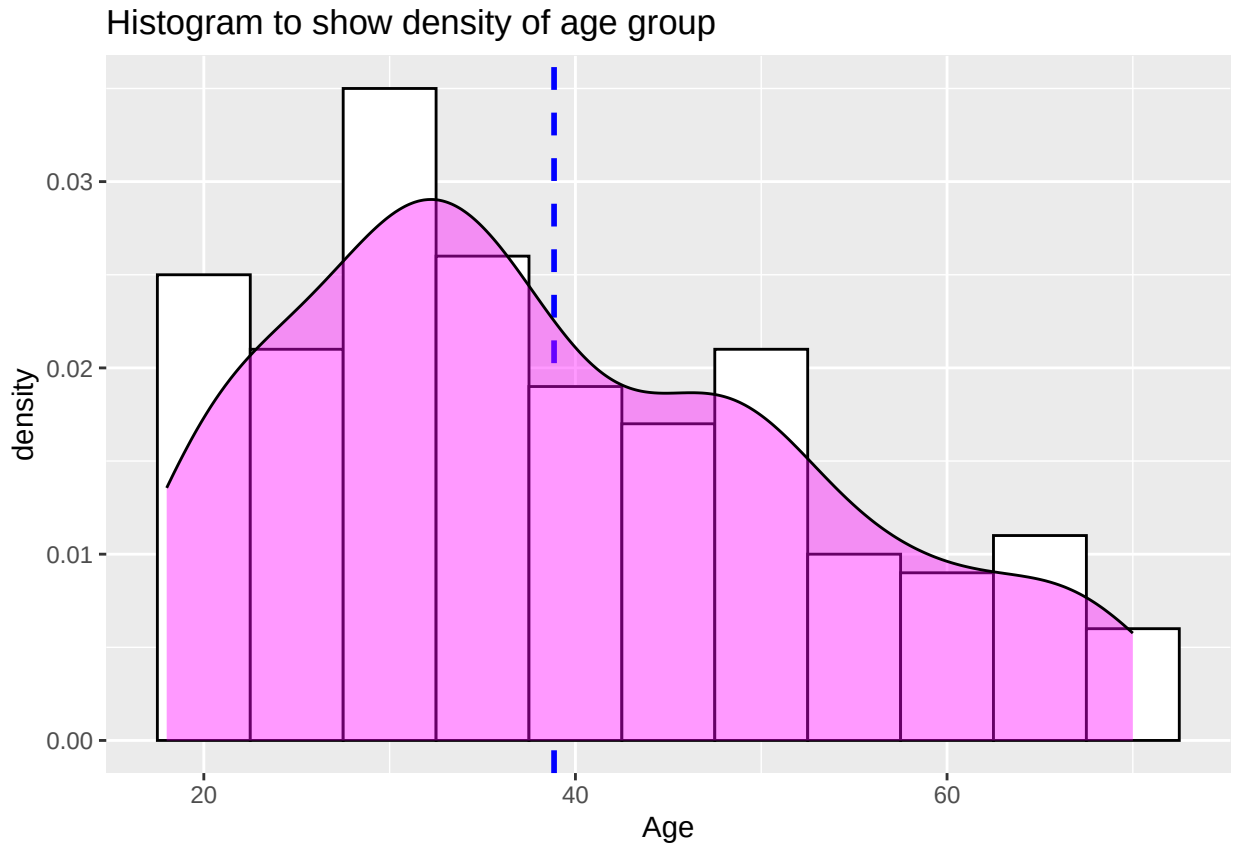


we will create a histogram with density plot to show the dispersion of mall customers based on age groups. we can see the distribution of mall customers based on their ages.The histogram shows that, the customers range from– those below their 20's, to those in their 70's. One way to simplify this data is to create a histogram of the customers based on age groups.This creates a neater visualisation of our data.

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

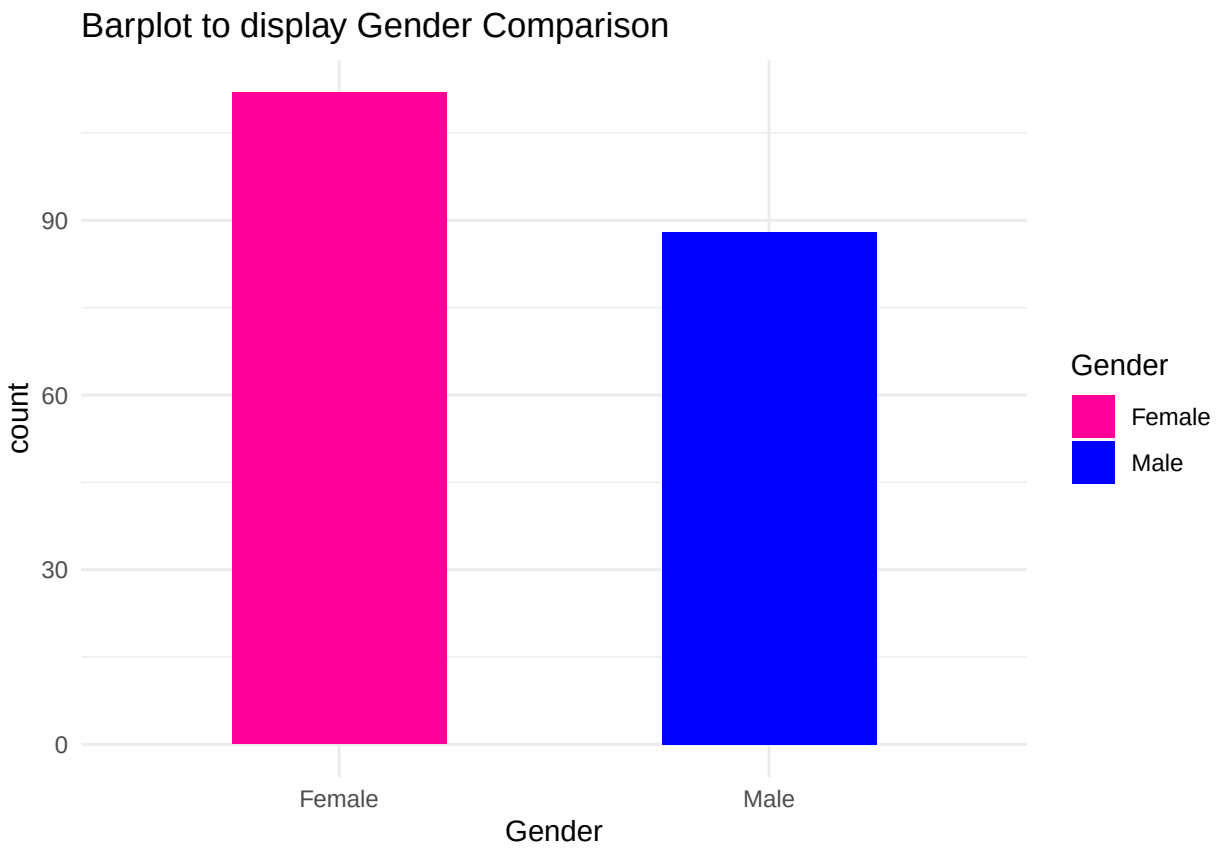
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



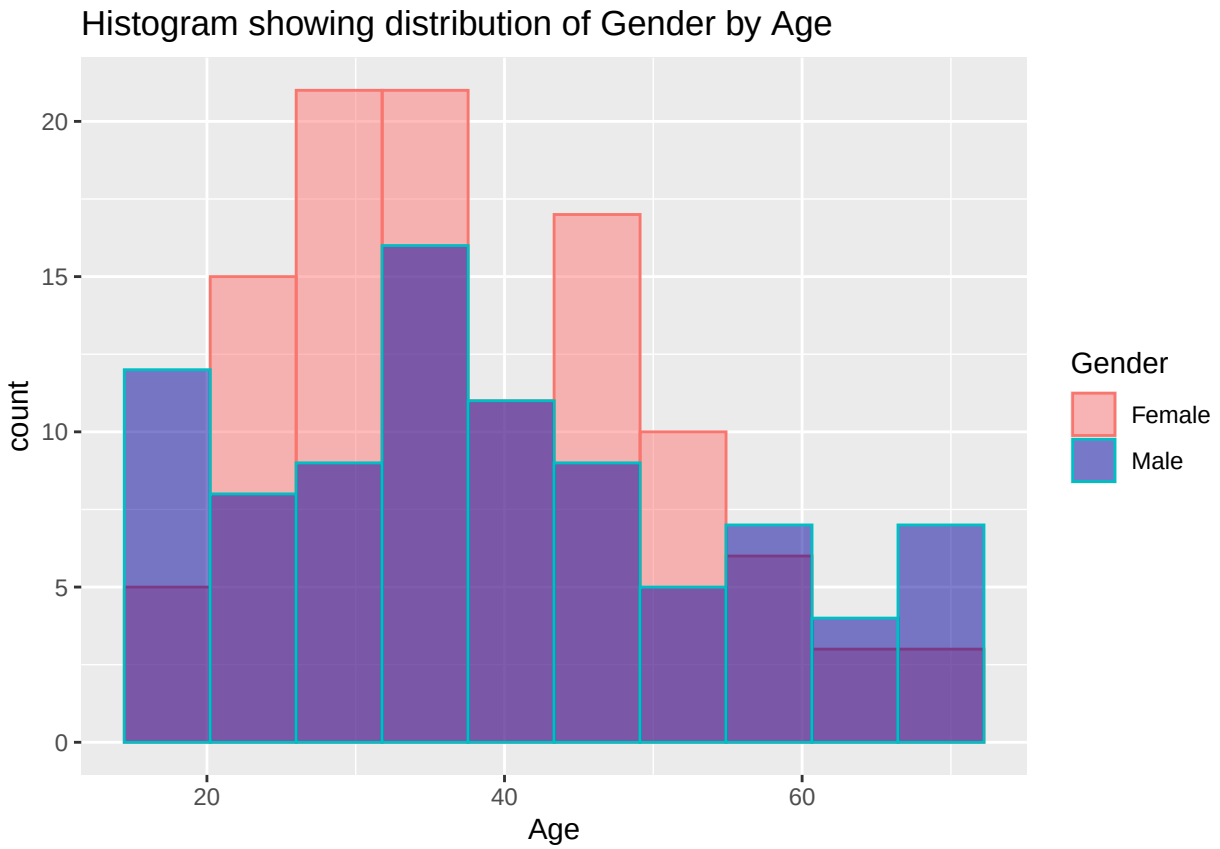
Now we will create gender distribution of the mall's customers base.

```
ggplot(Customer_Malldata,aes(x = Gender,fill = Gender))+
   geom_bar(stat = "count", width = 0.5) +
  scale_fill_manual(values = c("#ff0099", "#0000ff")) +
  theme_minimal() +
  labs(title = "Barplot to display Gender Comparison", xlab = "Gender")
```



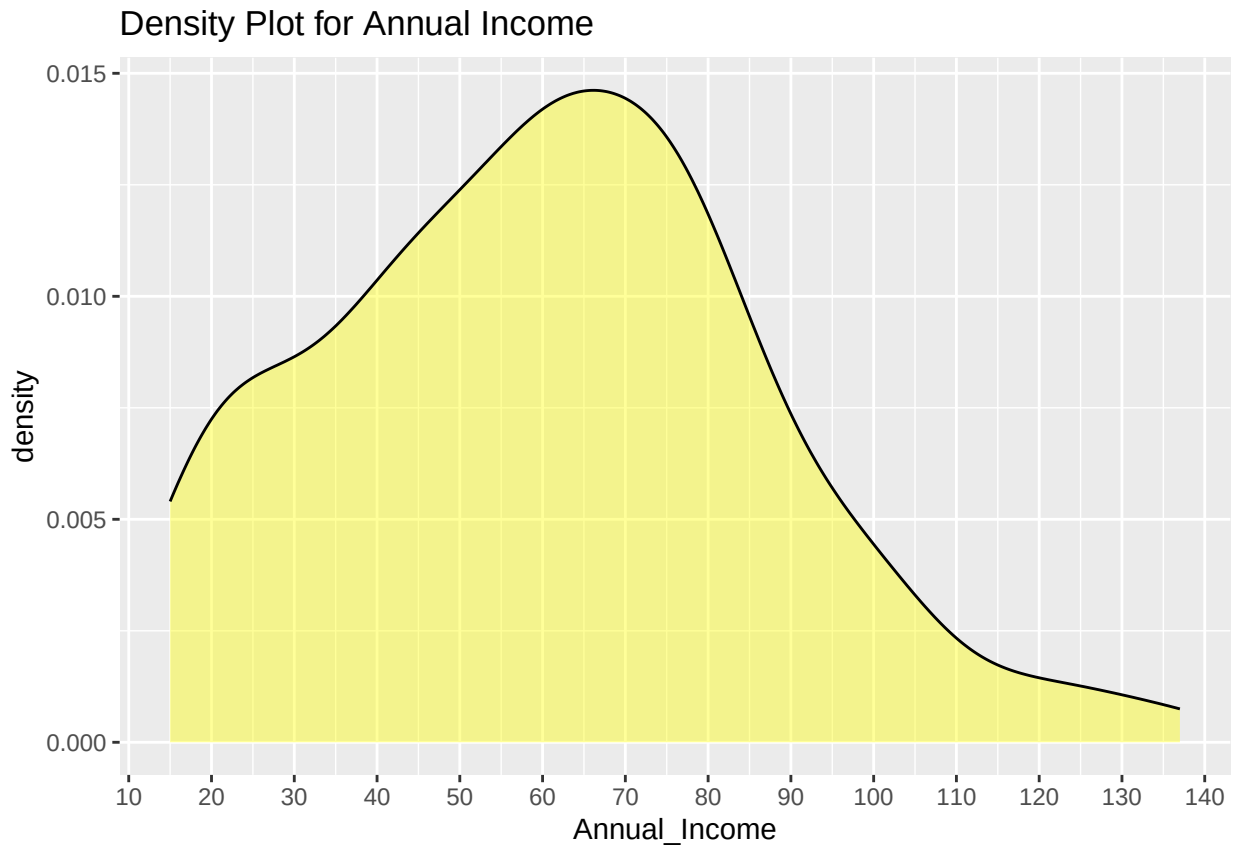
So, overall, this below code creates a histogram of the Age variable in the customer data frame, colored by the Gender variable. The bars are placed directly on the x-axis, and the transparency level of the bars is set to 0.5. The plot has a title indicating that it shows the distribution of Gender by Age.

```
## 3.3: Create a histogram for the variable "Age" by Gender
# Position = identiy is bar will be placed in the x axis.
ggplot(Customer_Malldata,aes(x=Age, fill=Gender, color=Gender))+
  scale_fill_manual(values=c("#ff7777","#00009f"))+
  geom_histogram(bins = 10, position = "identity",alpha=0.5) +
  labs(title="Histogram showing distribution of Gender by Age")
```



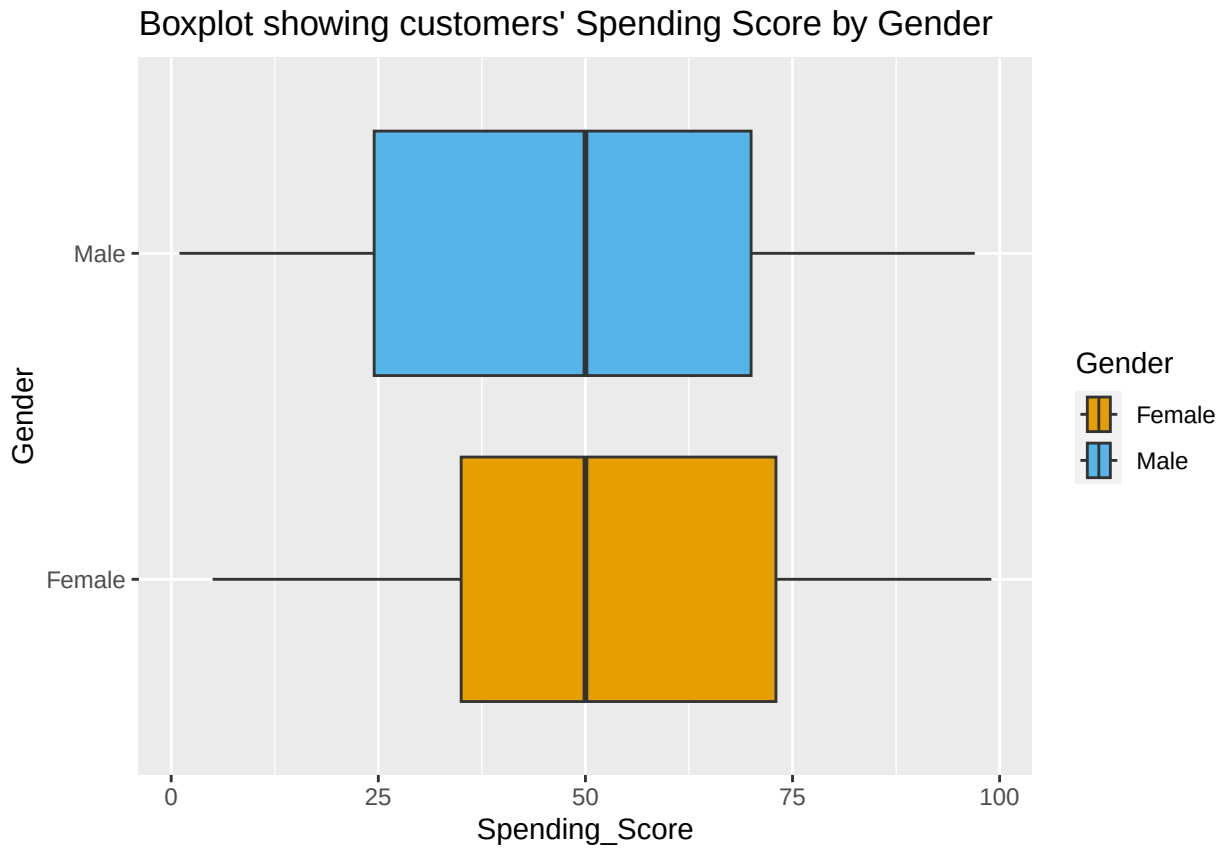
I can see the demographic information of the customers from the above analysis. ## Spending behaviour of the customers.

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



The density plot shows the distribution of the mall customer's annual income(in thousand). It seems that a majority of the customer base earns between 50,000 to around 80,000. Not a lot of customer base earns more than 100,000 annually.

```
# Create a box plot to understand customer spending score by gender.
ggplot(Customer_Malldata, aes(x = Spending_Score, y = Gender, fill = Gender)) +
  geom_boxplot() +
  scale_fill_manual(values = c("#E69F00", "#56B4E9")) +
  labs(title = "Boxplot showing customers' Spending Score by Gender")
```



From the boxplot, we can see that the median spending score for both males and females are equal. We can also see that more women have a spending score above the median (50), whereas men tend to have a spending score below the median.

Conducting the cluster analysis

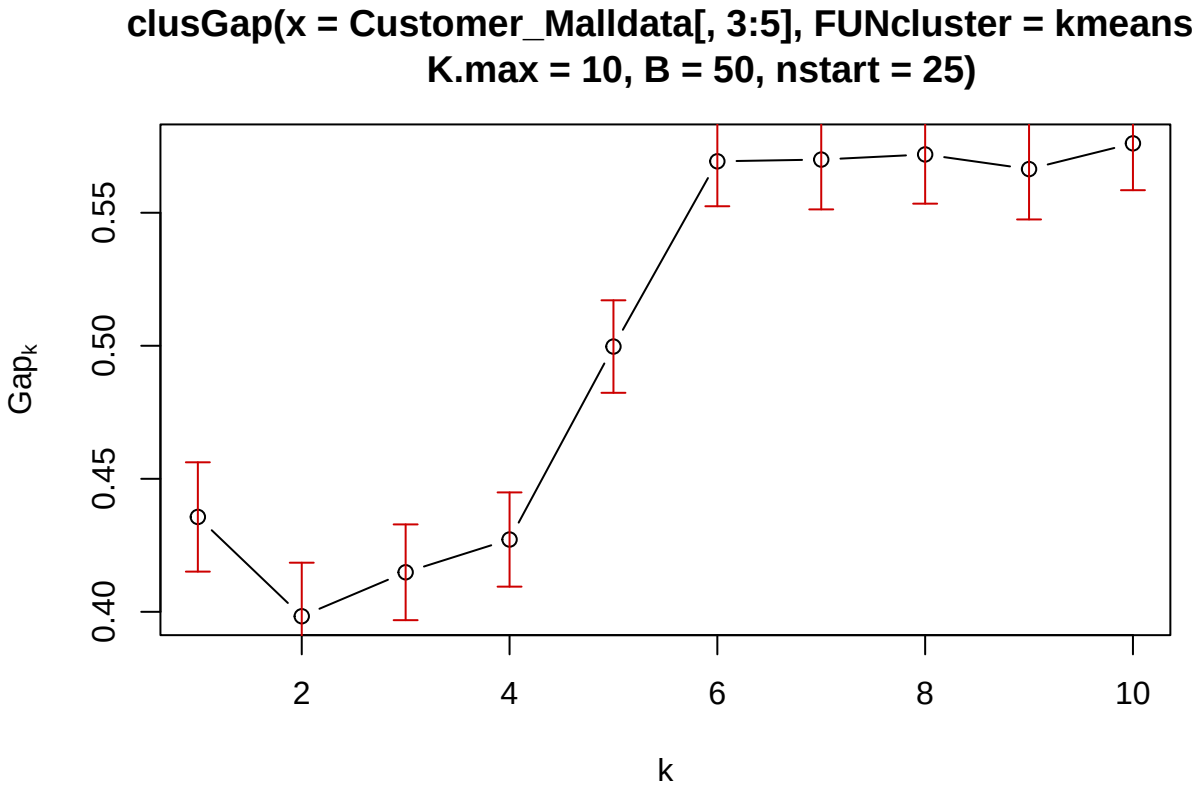
The general steps for conducting a cluster analysis using a k-means algorithm is as follows:

Choose the number of clusters "K" Select random K points that are going to be the centroids for each cluster Assign each data point to the nearest centroid, doing so will enable us to create "K" number of clusters Calculate a new centroid for each cluster Reassign each data point to the new closest centroid Go to step 4 and repeat Now, I will use Gap statistics to determine the optimal number of clusters to segment the mall customers into. A more detailed explanation on the use of gap-statistics in k-means clustering, can be found in this well-written article by Tim Löhr.

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

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

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



The plot above shows that, based on the gap statistic,6 is the optimal number of clusters to segment the mall customers into. Now, it's time to create the k means clustering for the data,

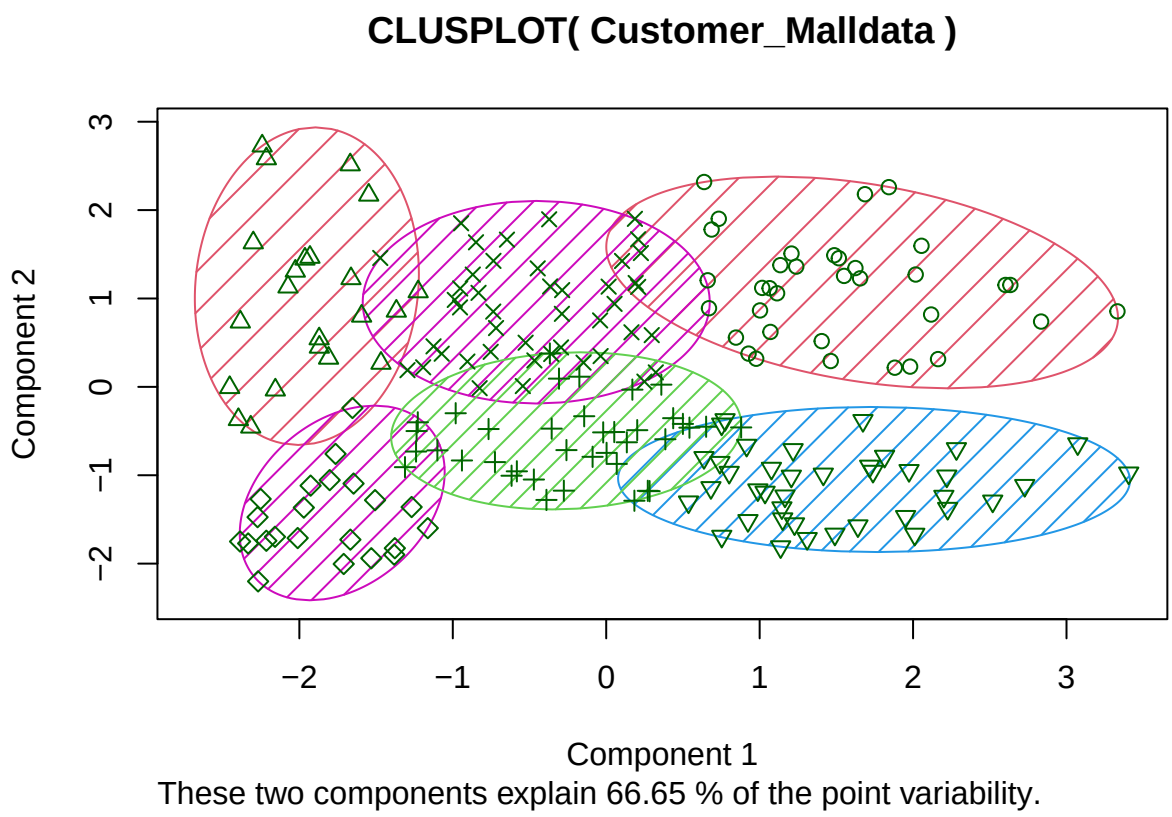
```
#Creating the customer clusters with KMeans
k6<-kmeans(Customer_Malldata[,3:5], 6, iter.max = 100, nstart=50,
           algorithm = "Lloyd")

#Printing the result
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
##  [38] 5 2 5 4 5 2 3 2 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
##  [75] 4 3 4 3 3 4 4 3 4 4 3 4 4 3 3 4 4 3 4 3 3 3 4 3 4 3 3 4 4 3 4 3 4 4 4 4 4
## [112] 3 3 3 3 3 4 4 4 4 3 3 3 6 3 6 1 6 1 6 1 6 3 6 1 6 1 6 1 6 3 6 1 6 1 6
## [149] 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
```

```
## [186] 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"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

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



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.

Next, I will perform a Principal Component Analysis (PCA) to reduce the dimensionality of the data and capture the 2 most significant components of the data. For more information on the PCA, this is a well-written write up to refer to.

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

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

```
## Importance of components:
##                             PC1     PC2     PC3
## Standard deviation     26.4625 26.1597 12.9317
## Proportion of Variance  0.4512  0.4410  0.1078
## Cumulative Proportion   0.4512  0.8922  1.0000
```

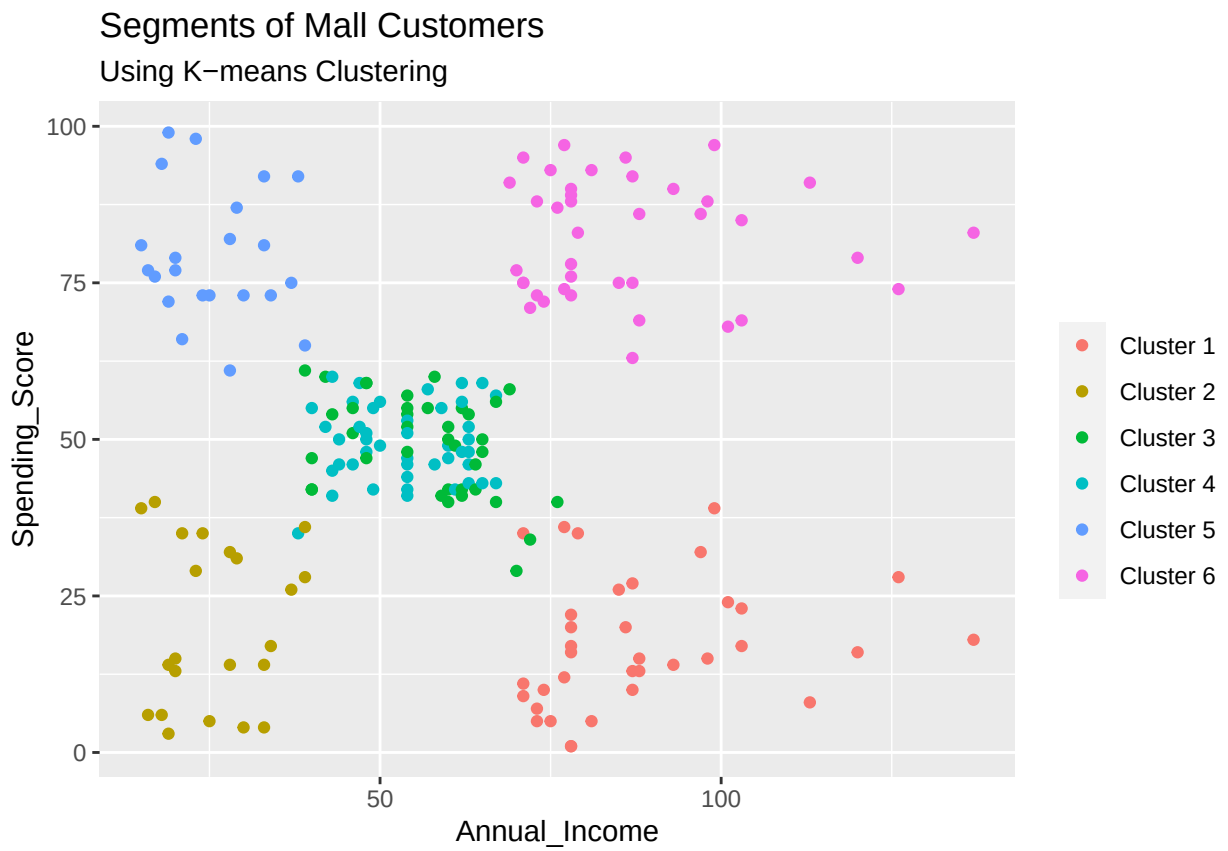
```
# Applying the PCA model on the data
pcclust$rotation[, 1:2]
```

```
##                        PC1        PC2
## Age              0.1889742 -0.1309652
## Annual_Income   -0.5886410 -0.8083757
## Spending_Score -0.7859965  0.5739136
```

Results from the PCA show that components 1 and 2 (PC1 and PC2) contribute t he 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.

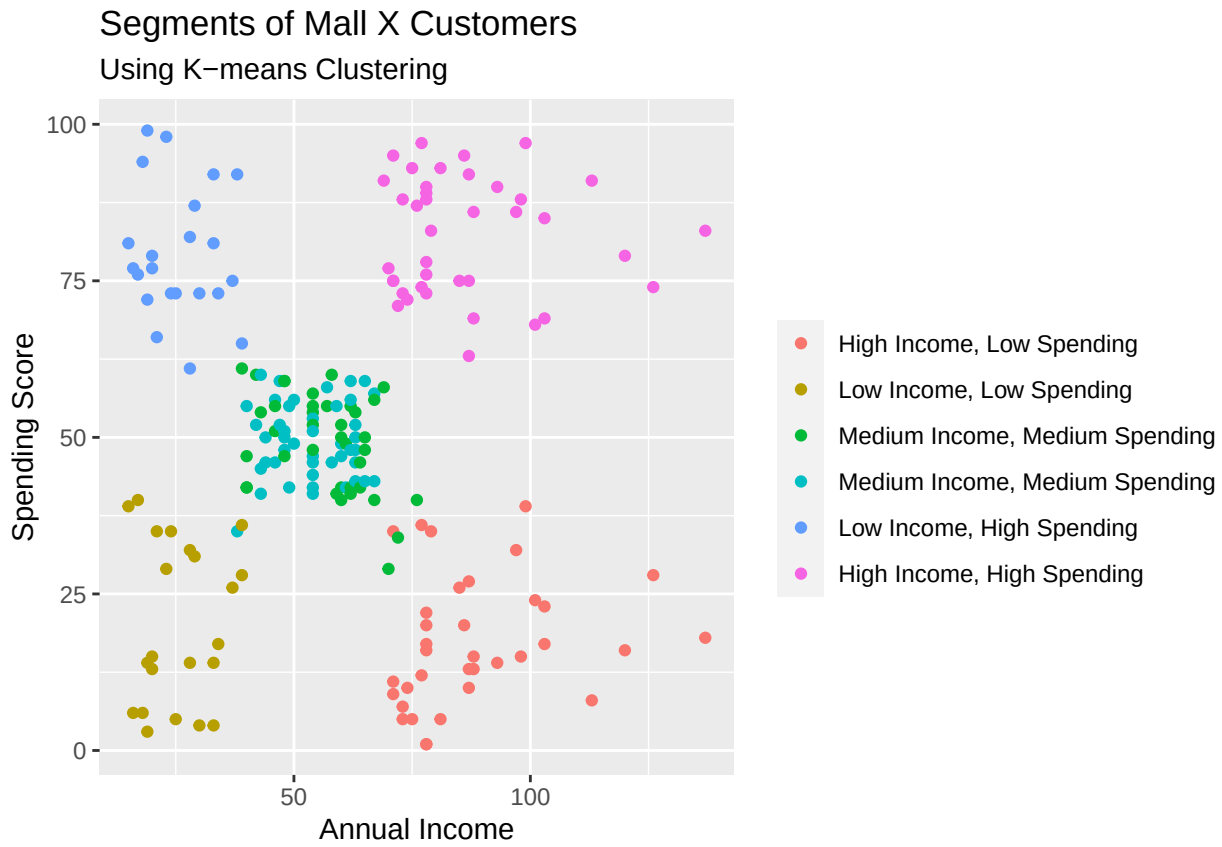
```
# Set seed to 1
set.seed(1)

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



I want to make this plot more consumable and formal for the use of external stakeholders. Based on the plot, we can easily classify each cluster by annual income and spending score.

```
#Create a plot of the customers segments
ggplot(Customer_Malldata, 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 Income, Medium Spending", "Low Income, High Spending","High Inc
  labs(x="Annual Income", y="Spending Score") +
  ggtitle("Segments of Mall X Customers",
          subtitle = "Using K-means Clustering")
```



Now I have a final plot that can be easily understood. The results show 6 distinct clusters of customers of Mall X. However, we can see that some overlapping in areas of the 3rd and 4th clusters (Medium Income, Medium Spending). Further analysis is needed to figure out why these two segments were separated in different clusters, although it might be indicative of a 3rd factor at play.