ML Final Assignment

Abhinay Chiranjeeth Marneni

2022-12-16

Data Exploration:

```
#Reading the data
Data<- read.csv("C:/Users/abhin/OneDrive/Documents/Assigments Buss 1sem/ML/Mall_Customers.csv")
head(Data)
```

```
##   CustomerID Gender Age Annual_Income Spending_Score
## 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
```

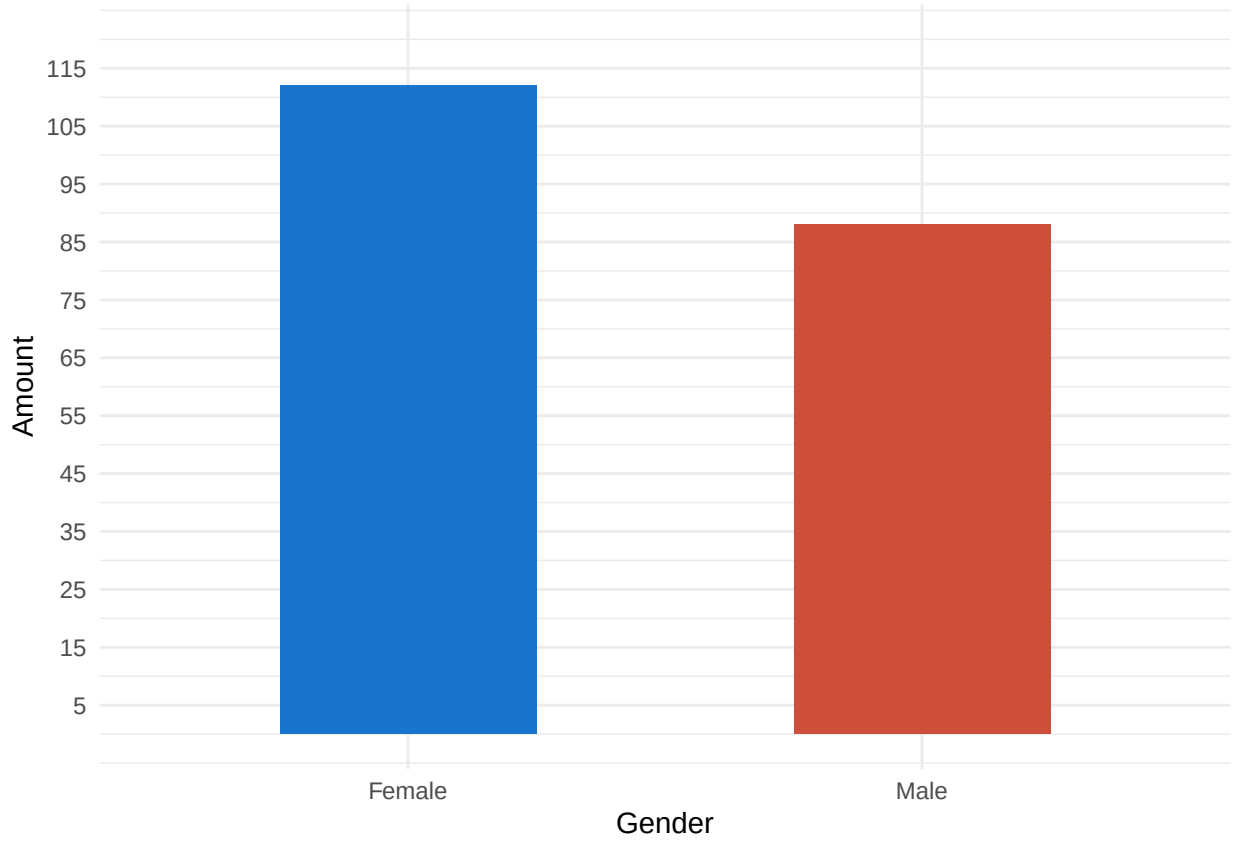
```
summary(Data)
```

```
##    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
```

```
#Converting Gender column into dummy variables

Male <- ifelse(Data$Gender=="Male" ,1,0)
Female <- ifelse(Data$Gender=="Female" ,1,0)

#Combining newly defined columns to the original dataset while removing CustomerID column.
New_data<-cbind(Data[,-c(1,2)],Male,Female)

#Normalizing the data:

Norm_data<- scale(New_data)
```

The Given dataset contains a more majority of female clients compared to male customers.

```
library(ggplot2)
ggplot(Data, aes(x= Gender))+
  scale_y_continuous(limits= c(0,120), breaks = seq(from= 5, to= 115, by= 10))+
  scale_x_discrete(labels= c("Female", "Male"))+
  ylab("Amount")+
  theme_minimal()+
  geom_bar(fill= c("dodgerblue3", "tomato3"), width = 0.5)
```




```
# Ratio of Female to Male in a Pie Chart
Gender_Table <- table(Data$Gender)
Gender_Table
```

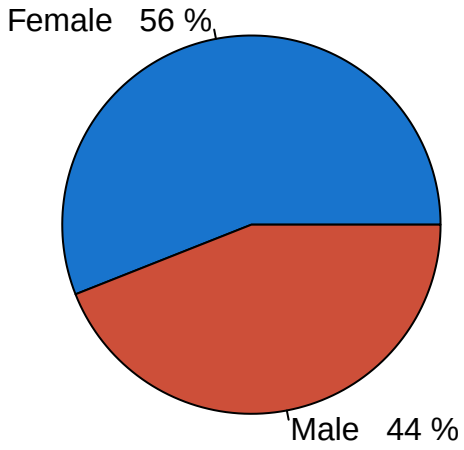
```
##
## Female   Male
##    112     88
```

```
Percent_Gender <- (table(Data$Gender)/sum(table(Data))) * 100
Percent_Gender
```

```
##
```

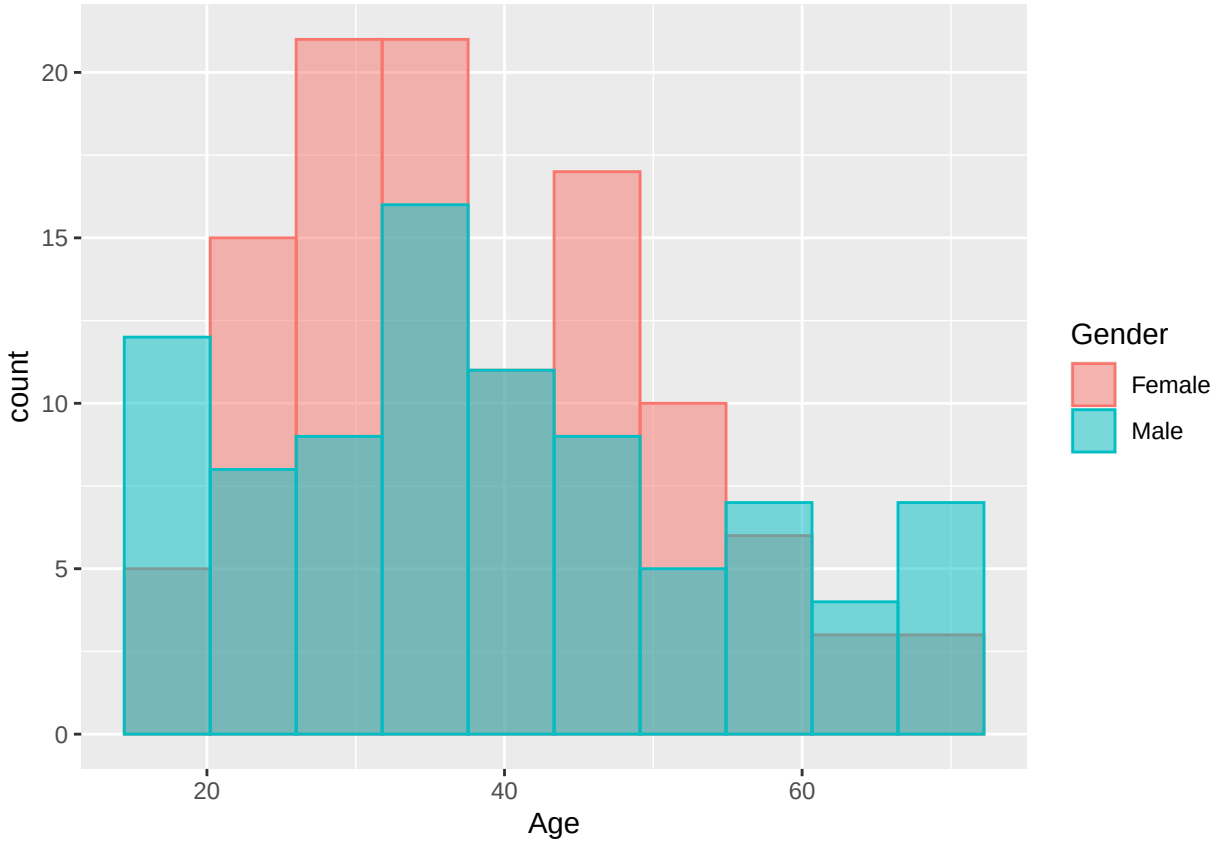
```
## Female   Male
##     56     44
```

```
Gender_labels <- paste(c("Female","Male"), " ", Percent_Gender, "%", sep = " ")
# Create pie chart
pie(x = Gender_Table, labels = Gender_labels, col = c("dodgerblue3", "tomato3"))
```



The plot of age range..

```
ggplot(Data, aes(x = Age, fill = Gender, colour = Gender)) +
  geom_histogram(bins = 10, position = "identity", alpha = 0.5)
```

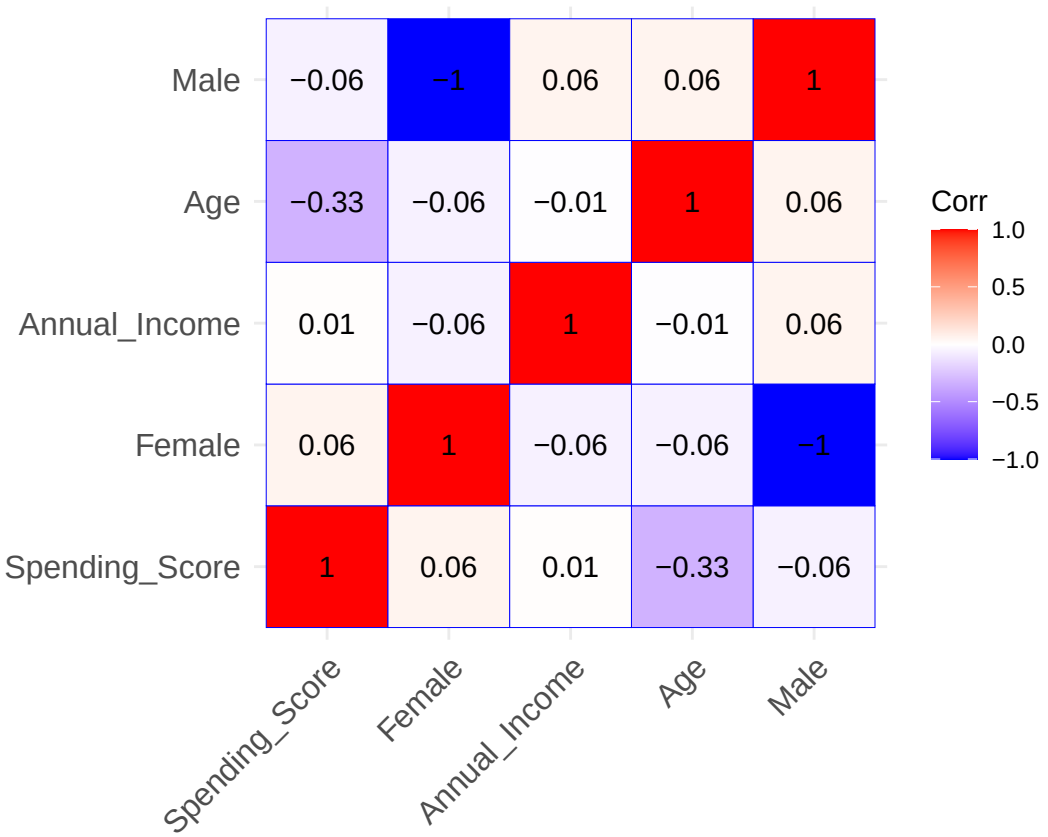


To check the major variables, I using the correlation matrix method

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.2.2
```

```
set.seed(650)
corr <- cor(Norm_data)
ggcorrplot(corr, outline.color = "blue", lab = TRUE, hc.order = TRUE, type = "full")
```



Implementation of K-means algorithm: ## The k-means algorithm can be used to manually adjust K numbers to observe how the dataset clusters. I've chosen k values at random from 2,3,4 and 5.

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.2.2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.2.2
```

```
set.seed(650)
k2 <- kmeans(Norm_data, centers = 2)
k3 <- kmeans(Norm_data, centers = 3)
k4 <- kmeans(Norm_data, centers = 4)
k5 <- kmeans(Norm_data, centers = 5)
k2$size
```

```
## [1] 112  88
```

```
k3$size
```

```
## [1] 88 57 55
```

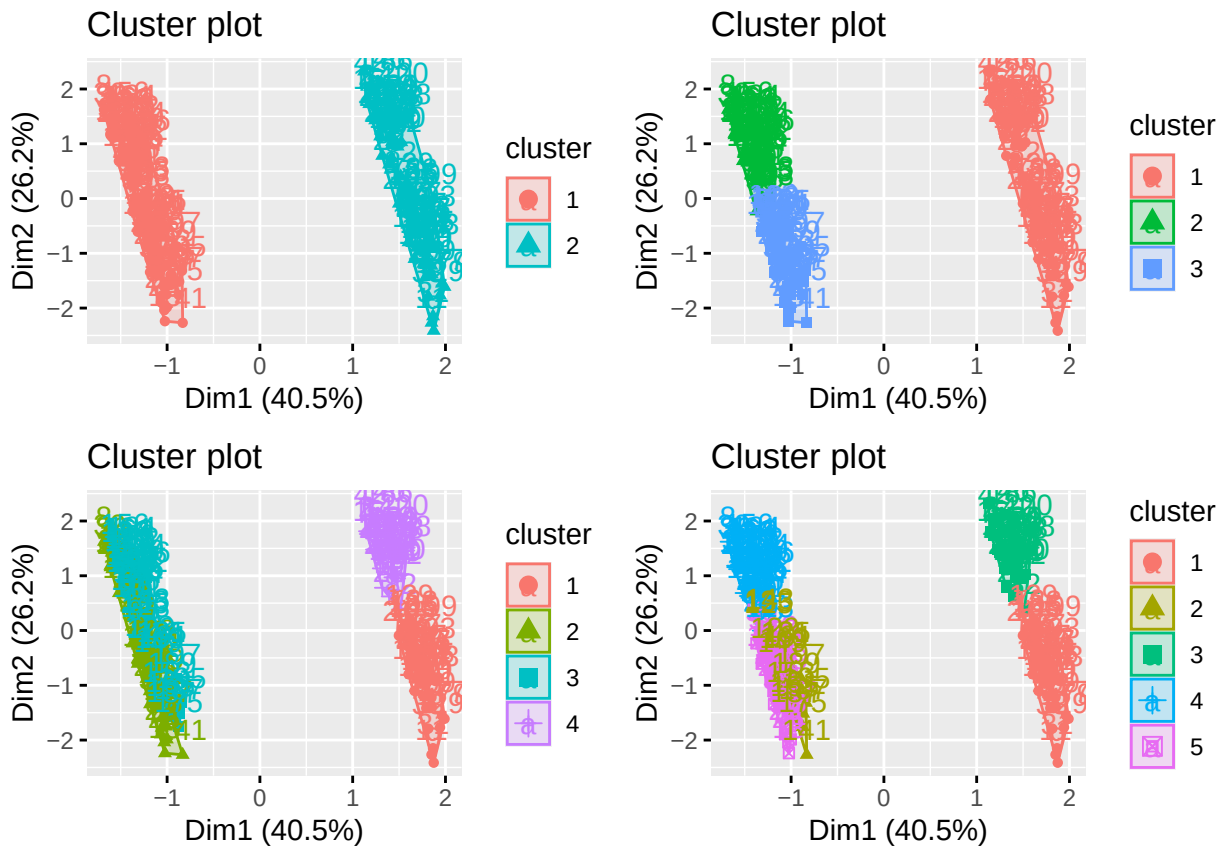
```
k4$size
```

```
## [1] 48 60 52 40
```

```
k5$size
```

```
## [1] 48 19 40 52 41
```

```
k21 <- fviz_cluster(k2, data =Norm_data)
k31 <- fviz_cluster(k3, data = Norm_data)
k41 <- fviz_cluster(k4, data = Norm_data)
k51 <- fviz_cluster(k5, data = Norm_data)
grid.arrange(k21,k31,k41,k51)
```

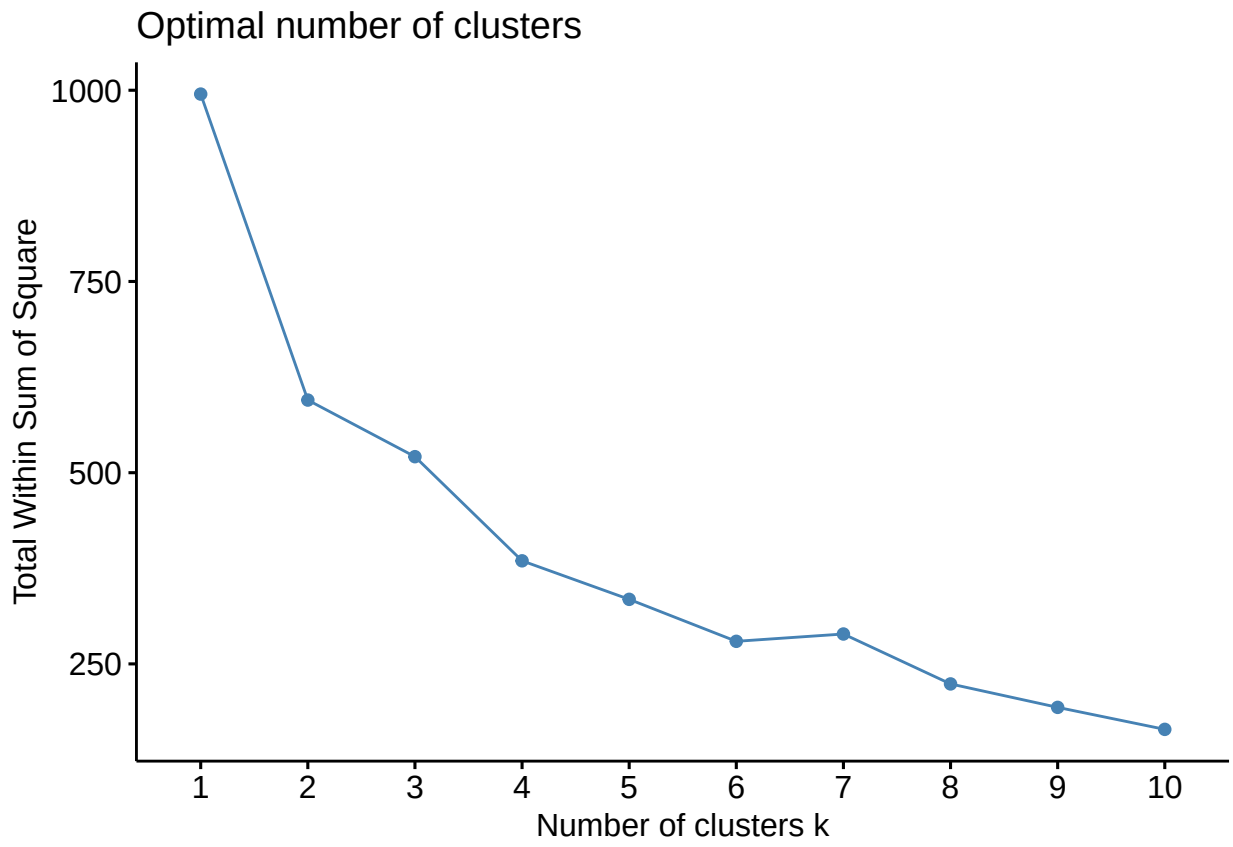



```
set.seed(650)
```

```
library(dplyr)
```

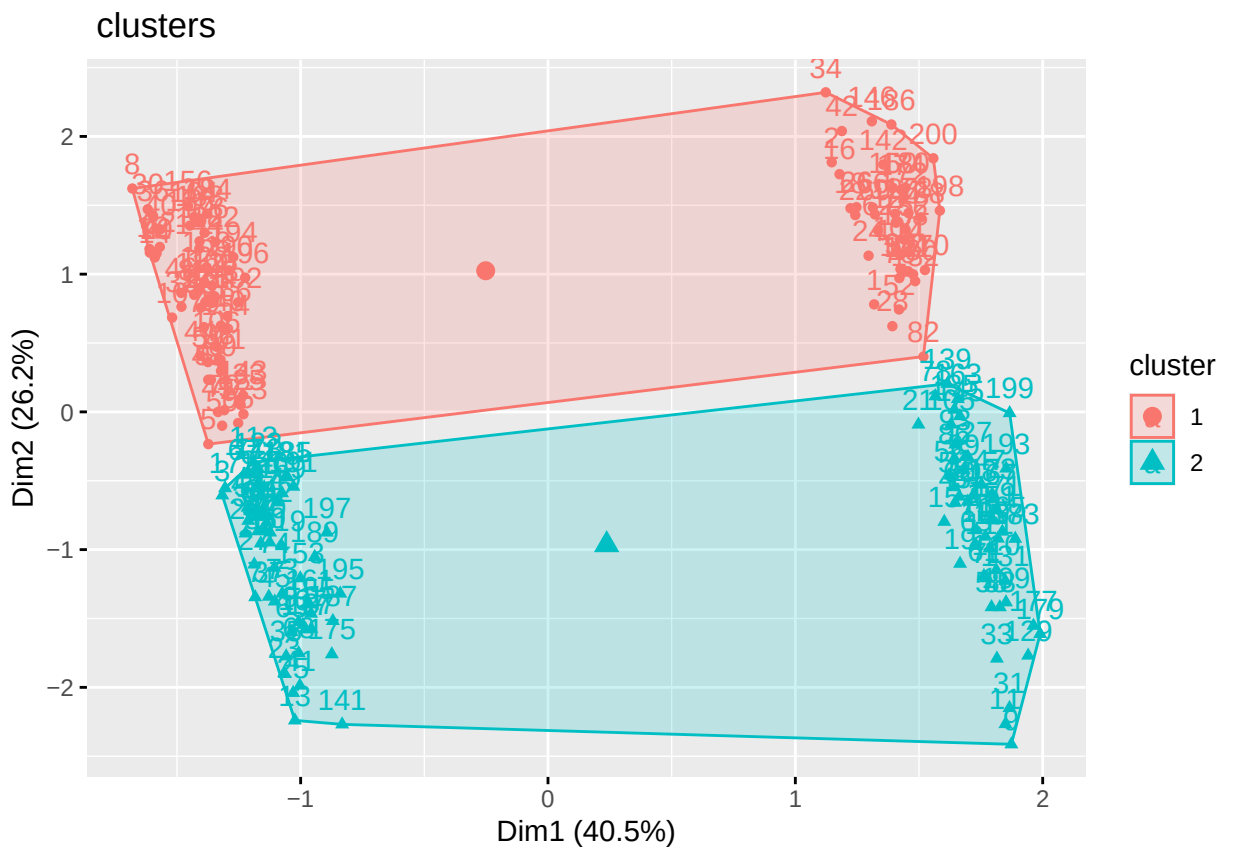
```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#Checking the optimal number of clusters to perform K-means algorithm
fviz_nbclust(Norm_data,kmeans,method="wss")
```



Based on the above plot, consdiering $k = 2$

```
#Implementing K-means algorithm
k2<-kmeans(Norm_data,centers=2)

#Visual representation of clusters
fviz_cluster(k2,New_data,main=" clusters")
```




```
#Assigning clusters to the original dataset
assigned_data<-cbind(New_data,k2$cluster)

#Finding mean within each column grouped by clusters for better interpretation
mean_k2 <- New_data %>% mutate(Cluster = k2$cluster) %>% group_by(Cluster) %>% summarise_all("mean")
head(mean_k2)
```

```
## # A tibble: 2 x 6
##   Cluster   Age Annual_Income Spending_Score  Male Female
##     <int> <dbl>         <dbl>          <dbl> <dbl>  <dbl>
## 1       1  28.4          60.6           69.3 0.412  0.588
## 2       2  48.7          60.5           32.2 0.466  0.534
```

Analysis of Clusters:

Cluster 1:

Customers with average age of 28 years have been grouped into first cluster. This group of customers has the maximum spending score of 69 percent, which indicates that young people are tend to spend more in malls than aged people. Among all the customers of this age group, females are 58% and males are 42%. This indicates that females spend the most at malls compared to men. Surprisingly, the average annual income of both the clusters is close to 60k dollars per annum.

Cluster 2:

This group consists of customers whose average age is 48 years. The spending score of this group of customers is very low compared to the first cluster. Even in this cluster, females take up a majority of the percentage.

The average percentage of females who spend at malls is 53% whereas males is 47%. The average income of this group of customers is also close to 60k dollars per annum.

Conclusion:

By analyzing the spending patterns of customers in a mall, we can conclude that the customers below the age of 30 are the one's who are spending the most. Moreover, In all age groups, percentage of Women spending in the mall is greater than men.

Therefore, in order to attract more customers to spend at malls, it would be advisable to have stores related to women like clothing and beauty products. This would add to the maximum profits at the mall.

Verifying if there is any correlation exists between spending score variable and other variables using simple linear regression for better interpretation.

```
Model1<-lm(New_data$Spending_Score~.,data=New_data)
summary(Model1)
```

```
##
## Call:
## lm(formula = New_data$Spending_Score ~ ., data = New_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.152 -17.903   1.567  18.066  46.535
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.930034   6.642253  11.130  < 2e-16 ***
## Age           -0.600371   0.124916  -4.806 3.06e-06 ***
## Annual_Income  0.007929   0.066420   0.119    0.905
## Male          -2.013234   3.511825  -0.573    0.567
## Female               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.57 on 196 degrees of freedom
## Multiple R-squared:  0.1086, Adjusted R-squared:  0.09496
## F-statistic:  7.96 on 3 and 196 DF,  p-value: 4.914e-05
```

As we can see, age holds a significant relationship with the spending score(as the P value is too small). As we have a negative slope, we can conclude that there is a negative correlation. R square value is 10.86% which indicates that there is a very less variability between spending score and all other variables.

Implementing a simple linear regression model on spending score and age variable based on the results of the above model.

```
Model2<- lm(New_data$Spending_Score~New_data$Age,data=New_data)
summary(Model2)
```

```
##
## Call:
## lm(formula = New_data$Spending_Score ~ New_data$Age, data = New_data)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -57.21 -17.53   2.02  18.10  46.47
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   73.7012     5.1238  14.384  < 2e-16 ***
## New_data$Age  -0.6049     0.1241  -4.873 2.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.46 on 198 degrees of freedom
## Multiple R-squared:  0.1071, Adjusted R-squared:  0.1026
## F-statistic: 23.74 on 1 and 198 DF,  p-value: 2.25e-06
```

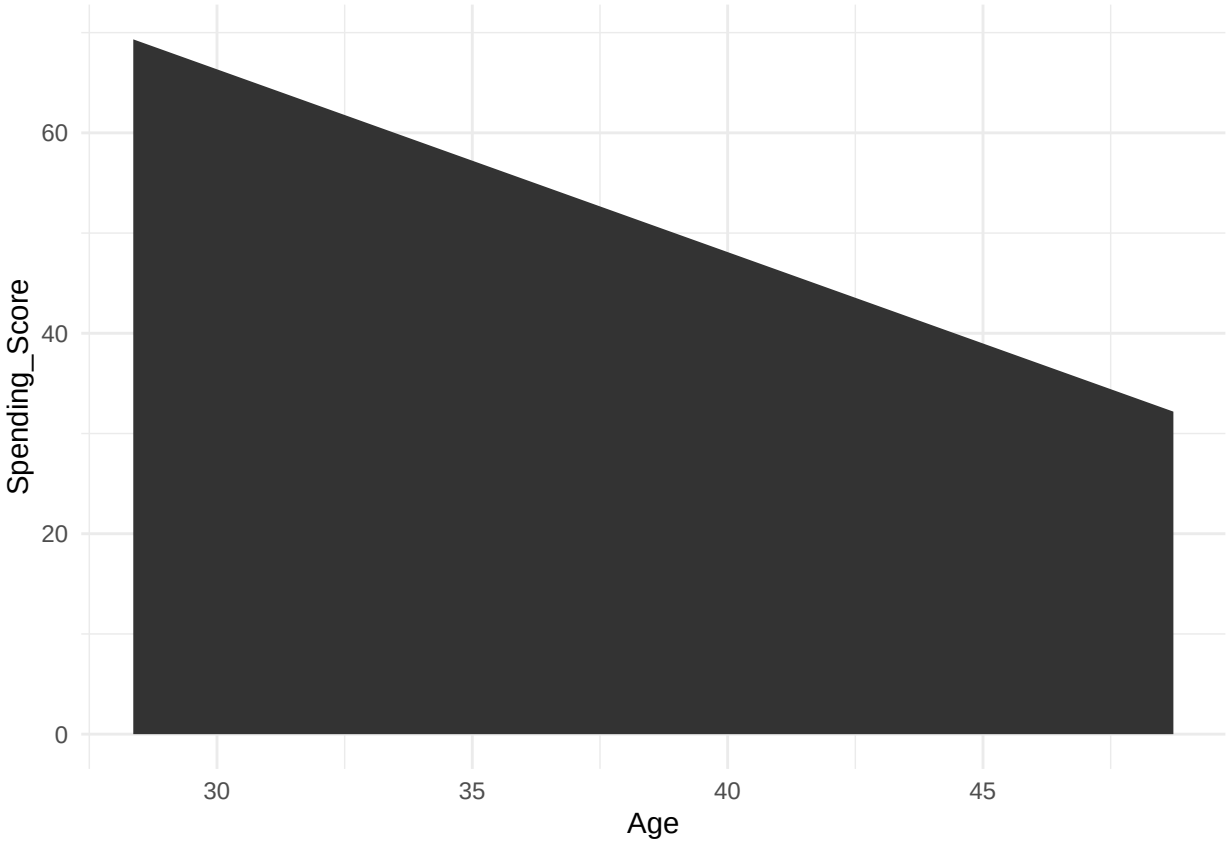
The results indicates the significant relationship between age and spending score. The P value and R square values are pretty close to the values observed with respect to the above model.

Visually representing the relationship between age and Spending score variable.

```
library(esquisse)
```

```
## Warning: package 'esquisse' was built under R version 4.2.2
```

```
library(ggplot2)

#Age Vs Spending Score
ggplot(mean_k2) +
 aes(x = Age, y = Spending_Score) +
 geom_area(size = 1.5) +
 theme_minimal()
```



The above graph indicates that as the age increases, the spending score of customers in the mall is decreasing.