Final Project

Anuraag Vasal

12/10/2021

```
library(dplyr)

##
## Attaching package: 'dplyr'

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

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

library(ggplot2)
library(data.table)

## Warning: package 'data.table' was built under R version 4.1.2

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

library(ggcorrplot)

## Warning: package 'ggcorrplot' was built under R version 4.1.2

library(pastecs)

## Warning: package 'pastecs' was built under R version 4.1.2

##
## Attaching package: 'pastecs'

## The following objects are masked from 'package:data.table':
##
##     first, last

## The following objects are masked from 'package:dplyr':
##
##     first, last

library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.2

## corrplot 0.92 loaded

setwd("C:/Users/anura/Desktop/Machine Learning Final")
list.files()

## [1] "Final Project Source.txt"         "Final Project.Rmd"
## [3] "Retail Marketing.csv"             "Screenshot 2021-12-11 214133.jpg"

raw.data <- read.csv("Retail Marketing.csv")
str(raw.data)

## 'data.frame':    1000 obs. of  10 variables:
##  $ Age        : chr  "Old" "Middle" "Young" "Middle" ...
##  $ Gender     : chr  "Female" "Male" "Female" "Male" ...
##  $ OwnHome    : chr  "Own" "Rent" "Rent" "Own" ...
##  $ Married    : chr  "Single" "Single" "Single" "Married" ...
##  $ Location   : chr  "Far" "Close" "Close" "Close" ...
##  $ Salary     : int  47500 63600 13500 85600 68400 30400 48100 68400 51900
80700 ...
##  $ Children   : int  0 0 0 1 0 0 0 0 3 0 ...
##  $ History    : chr  "High" "High" "Low" "High" ...
##  $ Catalogs   : int  6 6 18 18 12 6 12 18 6 18 ...
##  $ AmountSpent: int  755 1318 296 2436 1304 495 782 1155 158 3034 ...

#There are 10 variables and 1000 records (before cleaning the data)
```

#PART I : Cleaning and orginazing the data

```
str(raw.data)

## 'data.frame':    1000 obs. of  10 variables:
##  $ Age        : chr  "Old" "Middle" "Young" "Middle" ...
##  $ Gender     : chr  "Female" "Male" "Female" "Male" ...
##  $ OwnHome    : chr  "Own" "Rent" "Rent" "Own" ...
##  $ Married    : chr  "Single" "Single" "Single" "Married" ...
##  $ Location   : chr  "Far" "Close" "Close" "Close" ...
##  $ Salary     : int  47500 63600 13500 85600 68400 30400 48100 68400 51900
80700 ...
##  $ Children   : int  0 0 0 1 0 0 0 0 3 0 ...
##  $ History    : chr  "High" "High" "Low" "High" ...
##  $ Catalogs   : int  6 6 18 18 12 6 12 18 6 18 ...
##  $ AmountSpent: int  755 1318 296 2436 1304 495 782 1155 158 3034 ...

table(is.na(raw.data$Age))

##
## FALSE
##  1000

table(is.na(raw.data$Gender))
```

```
##
## FALSE
##  1000
```

```
table(is.na(raw.data$OwnHome))
```

```
##
## FALSE
##  1000
```

```
table(is.na(raw.data$Married))
```

```
##
## FALSE
##  1000
```

```
table(is.na(raw.data$Location))
```

```
##
## FALSE
##  1000
```

```
table(is.na(raw.data$Salary))
```

```
##
## FALSE
##  1000
```

```
table(is.na(raw.data$Children))
```

```
##
## FALSE
##  1000
```

```
table(is.na(raw.data$History))
```

```
##
## FALSE  TRUE
##   697   303
```

```
table(is.na(raw.data$Catalogs))
```

```
##
## FALSE
##  1000
```

```
table(is.na(raw.data$AmountSpent))
```

```
##
## FALSE  TRUE
##   994     6
```

#Replacing NA in History with ‘Unknown’:

```
raw.data$History <- as.character(raw.data$History)
raw.data$History[is.na(raw.data$History)] <- 'Unknown'
raw.data$History <- factor(raw.data$History)
table((raw.data$History))

##
##    High     Low  Medium Unknown
##     255     230     212     303
```

#Removing the 6 NAs in no amount spent

```
retail.df <- raw.data[!is.na(raw.data$AmountSpent),]
```

#factorizing the variable Children

```
retail.df$Children <- factor(retail.df$Children)
View(retail.df %>%
       group_by(Catalogs) %>%
       summarise(mean_of_amount = mean(AmountSpent),numebr_of_appirances =
n()))
```

#By looking at the table above variable Catalogs is actually a factor variable where 6 is the 'low_end' products and 24 is 'high_end' products,where as 12 and 16 are the mid_range products.

#Changing the notation to more intuitive notation.

```
retail.df<- (retail.df %>%
       mutate(Catalog = ifelse (Catalogs ==6, 'low_end',
                                (ifelse(Catalogs == 12, "low_midrange",
                                        (ifelse(Catalogs == 18,
"high_midrange", "high_end")))))))
```

#Factorizing the new variable

```
retail.df$Catalog <- as.factor(retail.df$Catalog)

#And removing the old one:
retail.df$Catalogs <- NULL
str(retail.df)

## 'data.frame':    994 obs. of  10 variables:
##  $ Age        : chr  "Old" "Middle" "Young" "Middle" ...
##  $ Gender     : chr  "Female" "Male" "Female" "Male" ...
##  $ OwnHome    : chr  "Own" "Rent" "Rent" "Own" ...
##  $ Married    : chr  "Single" "Single" "Single" "Married" ...
##  $ Location   : chr  "Far" "Close" "Close" "Close" ...
##  $ Salary     : int  47500 63600 13500 85600 68400 30400 48100 68400 51900
80700 ...
##  $ Children   : Factor w/ 4 levels "0","1","2","3": 1 1 1 2 1 1 1 1 4 1
...
##  $ History    : Factor w/ 4 levels "High","Low","Medium",..: 1 1 2 1 1 2 3
```

```
1 2 4 ...
##  $ AmountSpent: int  755 1318 296 2436 1304 495 782 1155 158 3034 ...
##  $ Catalog    : Factor w/ 4 levels "high_end","high_midrange",..: 3 3 2 2
4 3 4 2 3 2 ...

#We are left with 10 variables, 8 of them are factors and 2 are integers
(salary + amount spent)
```

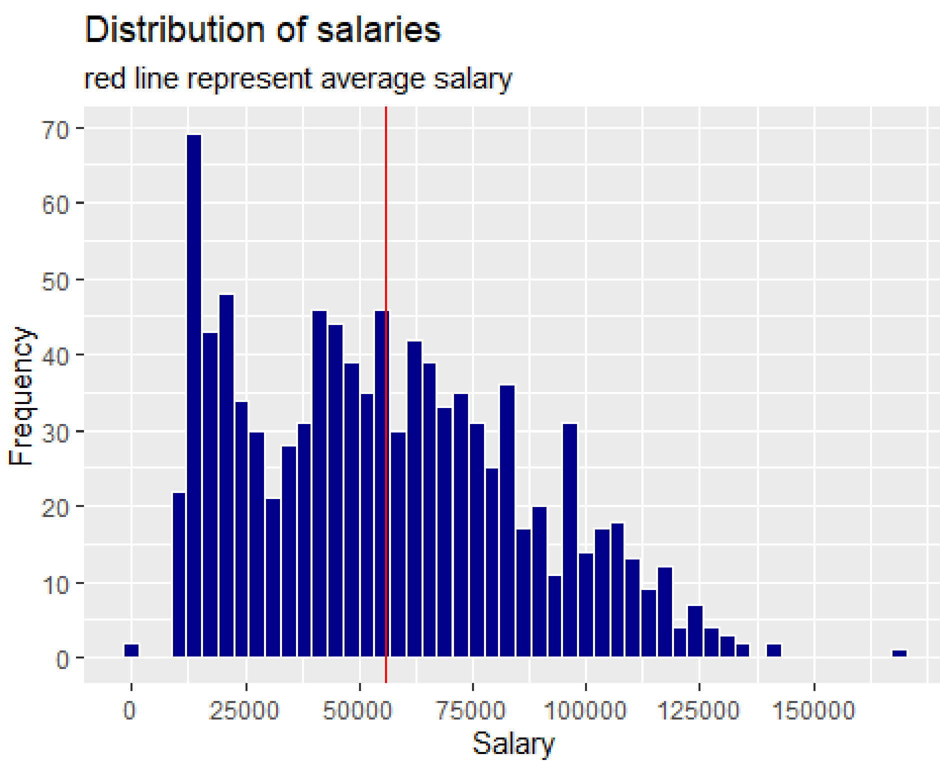
#PART II :Summary and Statistics

#Distribution of each categorical variables

```
lapply( retail.df %>%
          select(c("Age", "Gender", "OwnHome", "Married", "Location",
"Children", "History","Catalog"))
        ,table)

## $Age
##
## Middle    Old  Young
##    504    205    285
##
## $Gender
##
## Female   Male
##    501    493
##
## $OwnHome
##
##  Own Rent
##  514  480
##
## $Married
##
## Married  Single
##     500     494
##
## $Location
##
## Close   Far
##   706   288
##
## $Children
##
##   0   1   2   3
## 462 267 143 122
##
## $History
##
##    High     Low  Medium Unknown
##     254     229     211     300
```

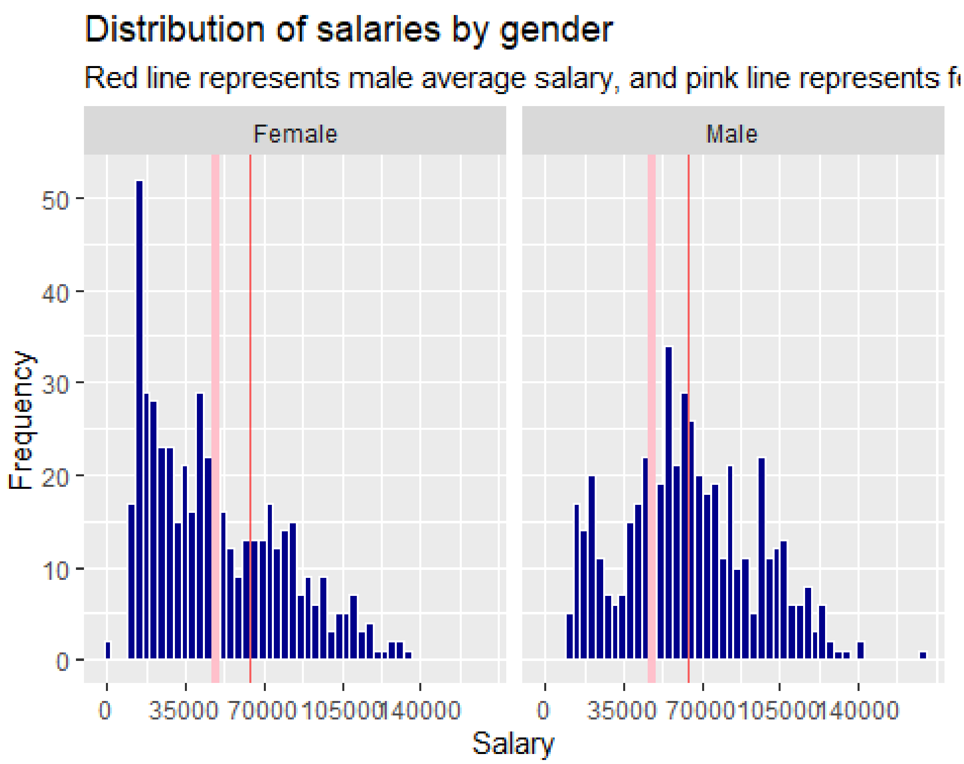
```
##
## $Catalog
##
##       high_end high_midrange       low_end  low_midrange
##            232           232           250           280
```

```
ggplot(data = retail.df, aes(x = Salary))+
  geom_histogram(bins = 50, colour = 'white', fill = 'darkblue')+
  scale_x_continuous(breaks = seq(0,150000,25000))+
  scale_y_continuous(breaks = seq(0,70,10))+
  xlab("Salary")+
  ylab("Frequency")+
  ggtitle("Distribution of salaries")+
  geom_vline(xintercept = mean(retail.df$Salary), color = 'red')+
  labs(subtitle  = 'red line represent average salary')
```

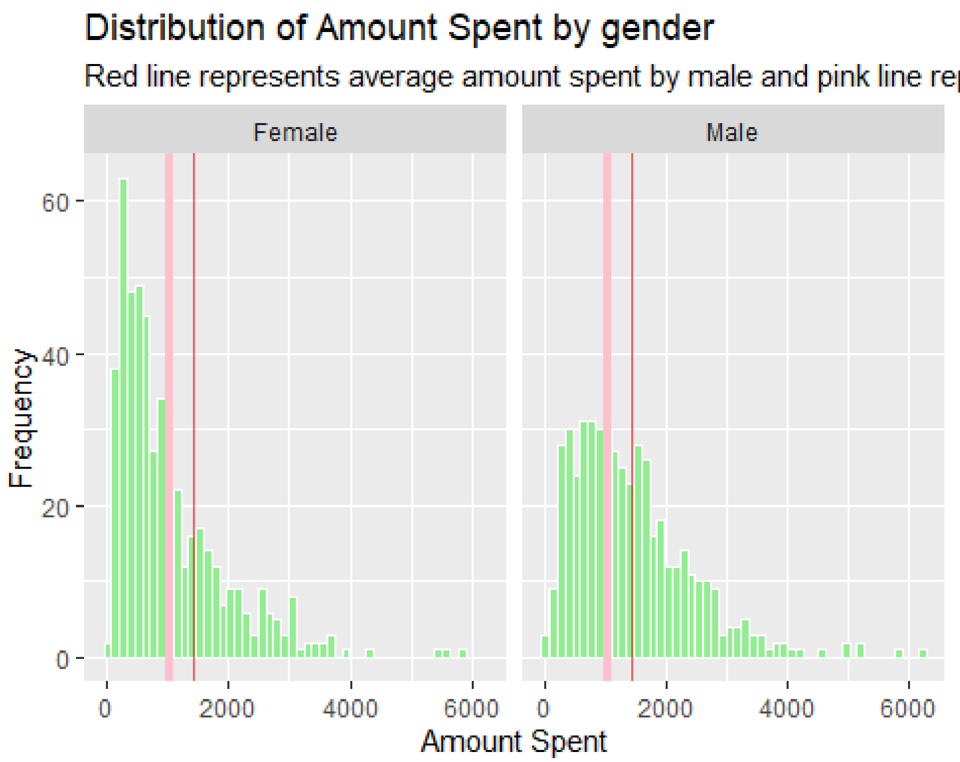



```
mean_salary_female <- mean(retail.df$Salary[retail.df$Gender =="Female"])
mean_salary_male <- mean(retail.df$Salary[retail.df$Gender =="Male"])


ggplot(data = retail.df, aes(x = Salary))+
  geom_histogram(bins = 50, colour = 'white', fill = 'darkblue')+
  scale_x_continuous(breaks = seq(0,150000,35000))+
  scale_y_continuous(breaks = seq(0,70,10))+
  xlab("Salary")+
  ylab("Frequency")+
  ggtitle("Distribution of salaries by gender")+
```

```
  geom_vline(xintercept = mean_salary_female, color = 'pink',size=1.5)+
  geom_vline(xintercept = mean_salary_male, color = 'red', alpha= 0.6)+
  labs(subtitle  = "Red line represents male average salary, and pink line
represents female average salary")+
  facet_wrap(~Gender)
```



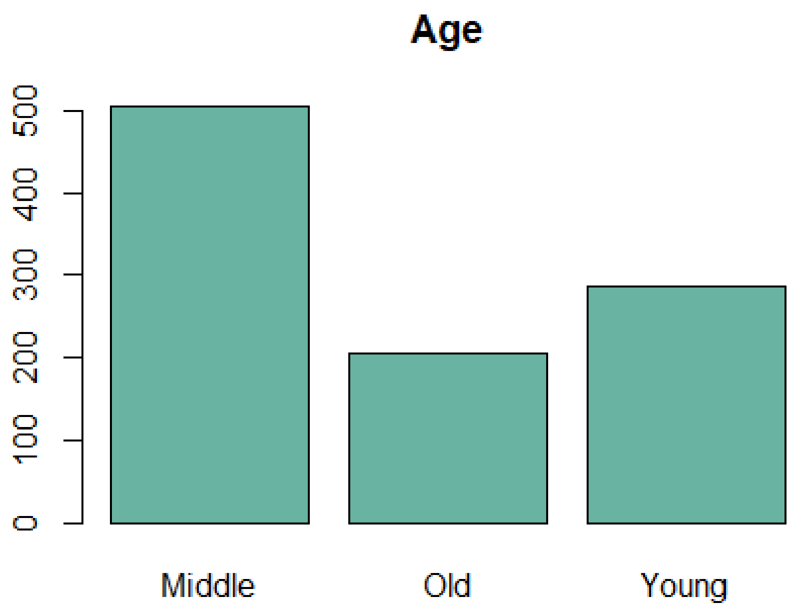

```
mean_AmountSpent_female <- mean(retail.df$AmountSpent[retail.df$Gender
=="Female"])
mean_AmountSpent_male <- mean(retail.df$AmountSpent[retail.df$Gender
=="Male"])

ggplot(data = retail.df, aes(x = AmountSpent))+
  geom_histogram(bins = 50, colour = 'white', fill = 'lightgreen')+
  scale_x_continuous()+
  scale_y_continuous()+
  xlab("Amount Spent")+
  ylab("Frequency")+
  ggtitle("Distribution of Amount Spent by gender")+
  labs(subtitle  = "Red line represents average amount spent by male and pink
line represents average amount spent by female")+
  facet_wrap(~Gender)+
  geom_vline(xintercept = mean_AmountSpent_female, color = 'pink',size=1.5)+
  geom_vline(xintercept = mean_AmountSpent_male, color = 'red', alpha= 0.6)
```



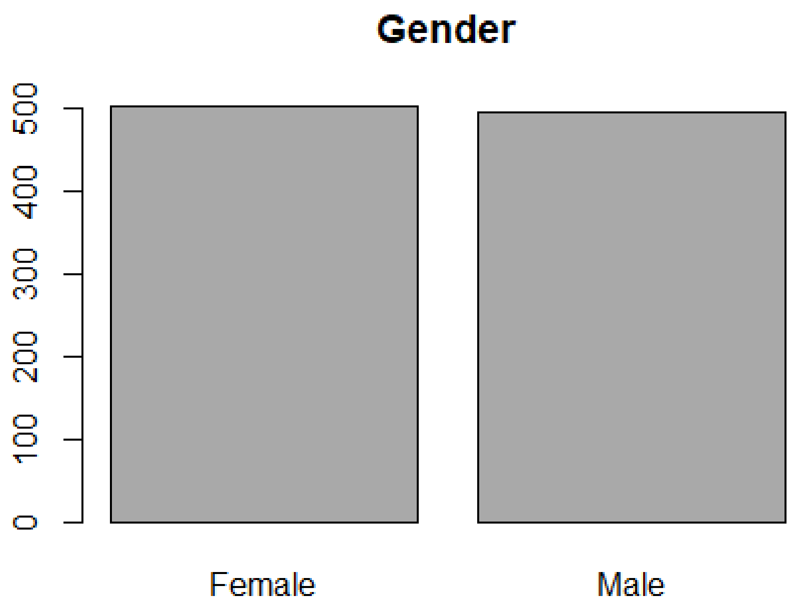

```r
#raw.data <- read.csv("Retail Marketing.csv")
raw.data <- raw.data[!is.na(raw.data$AmountSpent),]
head(raw.data$Age, 10)

##  [1] "Old"    "Middle" "Young"  "Middle" "Middle" "Young"  "Middle"
"Middle"
##  [9] "Middle" "Old"

cor.data <- raw.data
levels(raw.data$Age)

## NULL

cor.data$Age <- ifelse(cor.data$Age == 'Young', 0,
                       ifelse(cor.data$Age == 'Middle',1,2))

levels(raw.data$Gender)

## NULL

cor.data$Gender <- ifelse(cor.data$Gender == "Female", 0 ,1)
levels(raw.data$OwnHome)

## NULL

cor.data$OwnHome <- ifelse(cor.data$OwnHome == "Rent", 0 ,1)
levels(raw.data$Married)
```

```
## NULL

cor.data$Married <- ifelse(cor.data$Married == "Single", 0 ,1)
levels(raw.data$Location)

## NULL

cor.data$Location_close <- ifelse(cor.data$Location == "Far", 0 ,1)
cor.data$History<- NULL
cor.data$Location<- NULL

str(cor.data)

## 'data.frame':    994 obs. of  9 variables:
##  $ Age           : num  2 1 0 1 1 0 1 1 1 2 ...
##  $ Gender        : num  0 1 0 1 0 1 0 1 0 1 ...
##  $ OwnHome       : num  1 0 0 1 1 1 0 1 1 1 ...
##  $ Married       : num  0 0 0 1 0 1 0 0 1 1 ...
##  $ Salary        : int  47500 63600 13500 85600 68400 30400 48100 68400
51900 80700 ...
##  $ Children      : int  0 0 0 1 0 0 0 0 3 0 ...
##  $ Catalogs      : int  6 6 18 18 12 6 12 18 6 18 ...
##  $ AmountSpent   : int  755 1318 296 2436 1304 495 782 1155 158 3034 ...
##  $ Location_close: num  0 1 1 1 1 1 1 1 1 0 ...

cor.maxtrix<- cor(cor.data, method = "pearson", use = "complete.obs")
corrplot(cor.maxtrix)
```

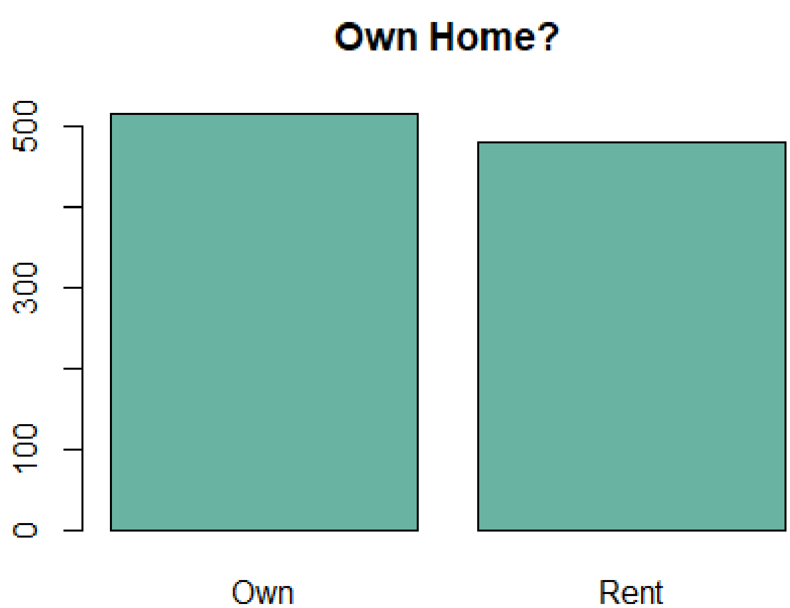
```r
library(ggplot2)
par(mfrow=c(1,1))
barplot(table(raw.data$Age), main="Age", col = "#69b3a2")
```



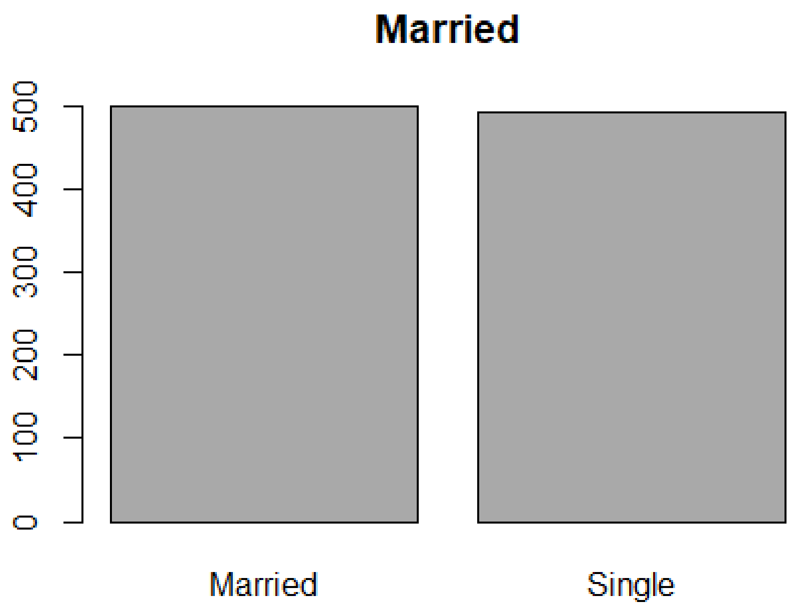

```r
barplot(table(raw.data$Gender), main="Gender", col = "#A9A9A9")
```



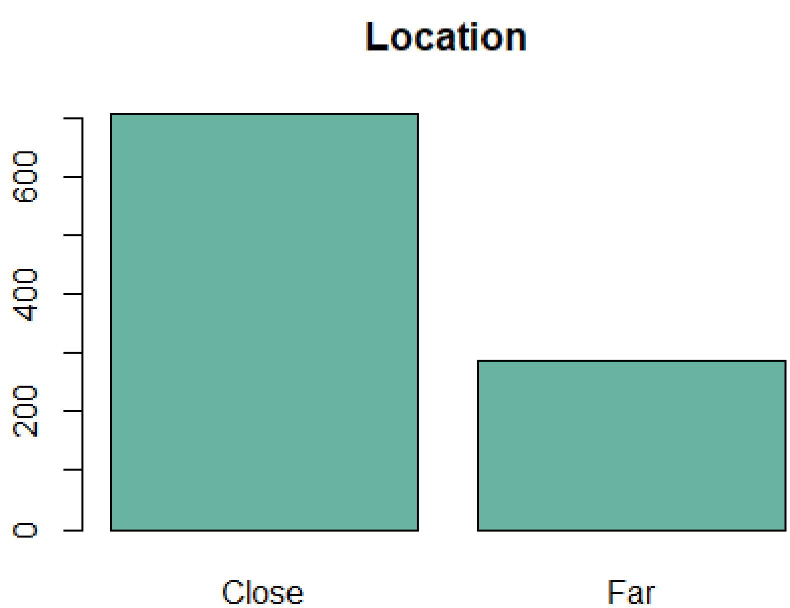

```
barplot(table(raw.data$OwnHome), main="Own Home?", col = "#69b3a2")
```



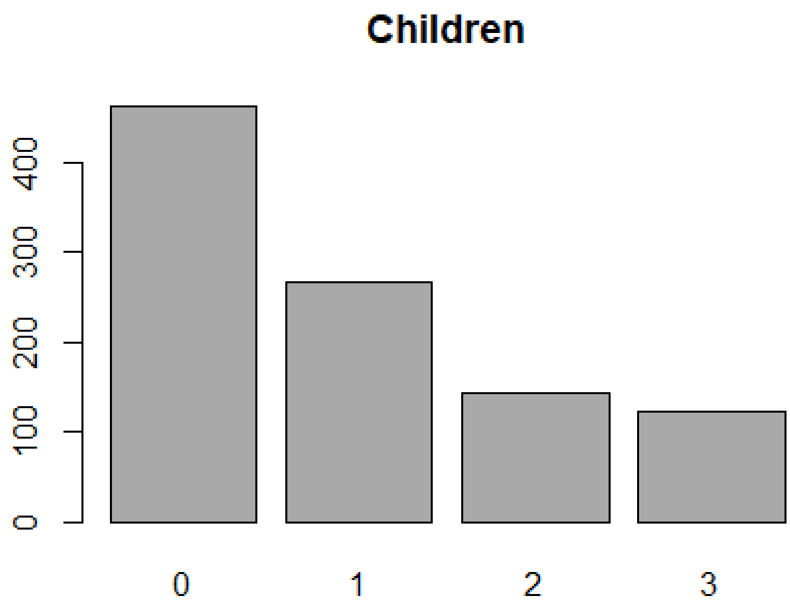

```
barplot(table(raw.data$Married), main="Married", col = "#A9A9A9")
```



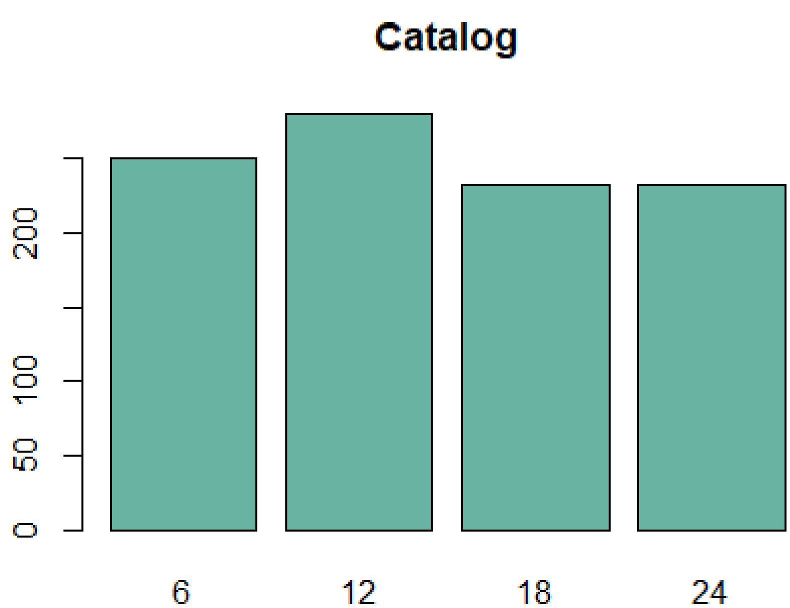

```
barplot(table(raw.data$Location), main="Location", col = "#69b3a2")
```



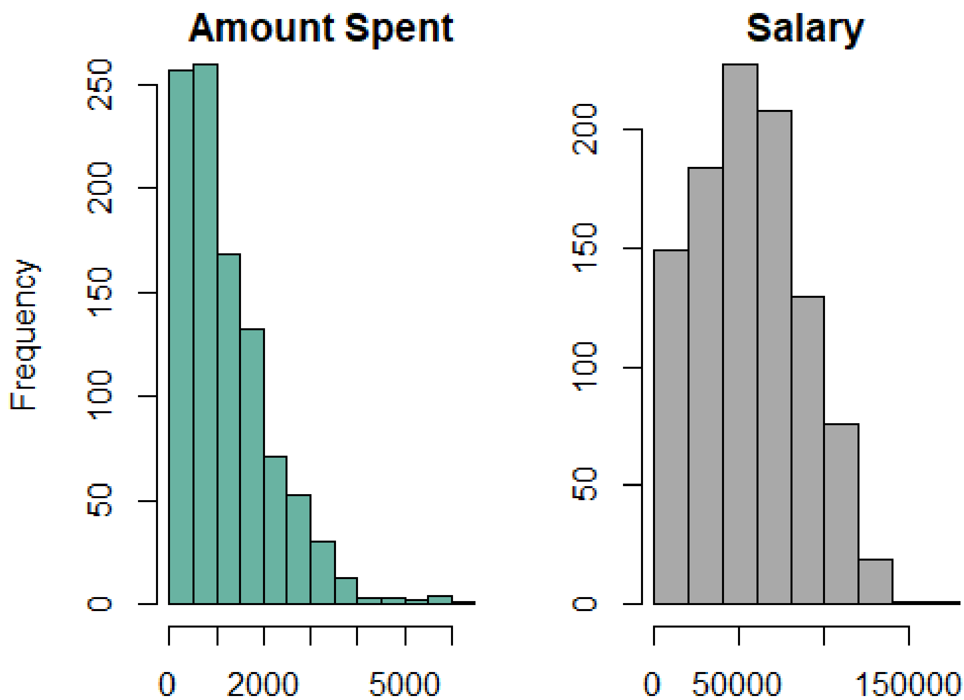

```
barplot(table(raw.data$Children), main="Children", col = "#A9A9A9")
```




```
barplot(table(raw.data$Catalog), main="Catalog", col = "#69b3a2")
```



```r
par(
  mfrow=c(1,2),
  mar=c(4,4,1,0)
)
hist((raw.data$AmountSpent), xlab="", main="Amount Spent", col = "#69b3a2")
hist((raw.data$Salary), xlab="", ylab="", main="Salary", col = "#A9A9A9")
```



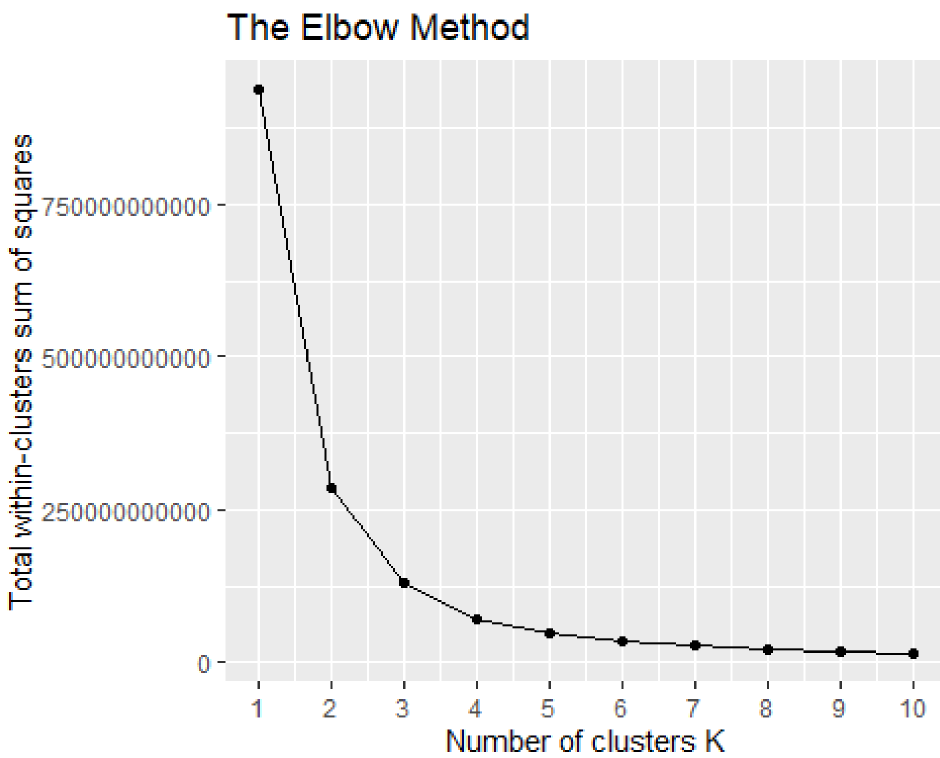

```r
library(cluster)

## Warning: package 'cluster' was built under R version 4.1.2

library(factoextra)

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

library(flexclust)

## Warning: package 'flexclust' was built under R version 4.1.2

## Loading required package: grid

## Loading required package: lattice

## Loading required package: modeltools

## Loading required package: stats4

library(fpc)
```

```
## Warning: package 'fpc' was built under R version 4.1.2

library(clustertend)
library(ClusterR)

## Warning: package 'ClusterR' was built under R version 4.1.2

## Loading required package: gtools

## Warning: package 'gtools' was built under R version 4.1.2

library(data.table)

retail.df <- raw.data[!is.na(raw.data$AmountSpent),]

clustering.df <- cor.data
dim(clustering.df)[2]

## [1] 9
```

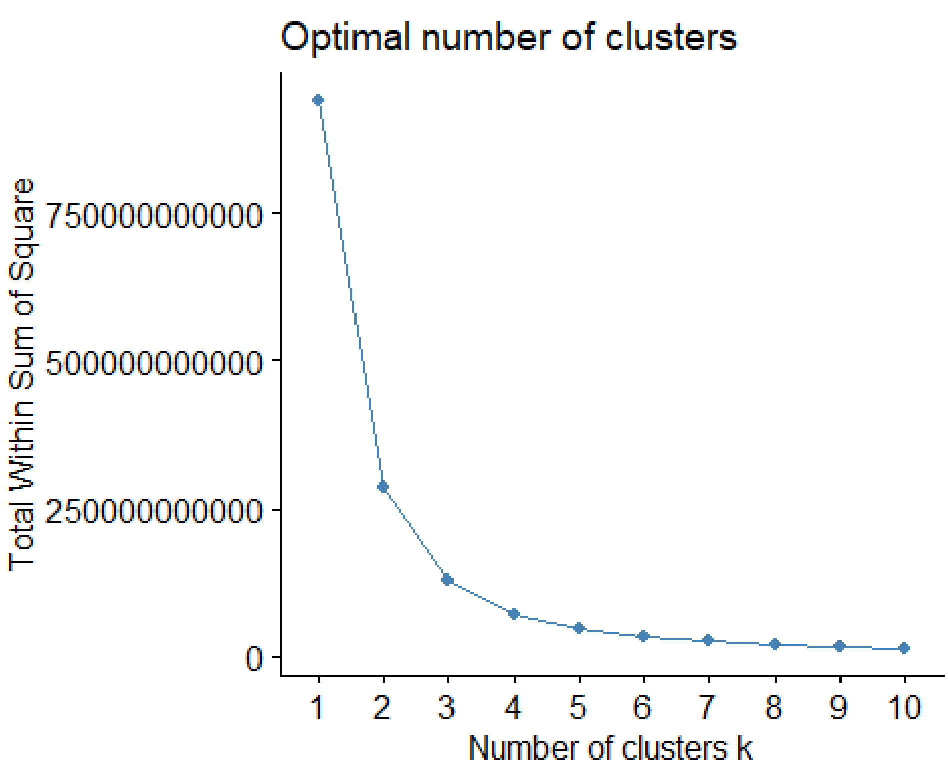
#Choosing optimal number of clusters

#Assuming the maximum K to cluster is 10:

```
k.max <- 10

#we will create a vector of the total within sum of squars, in order to
visulize it
wss <- sapply(1:k.max, function(k){kmeans(clustering.df, k,
nstart=50,iter.max = 1000 )$tot.withinss})

options("scipen"=999)
ggplot()+ aes(x = 1:k.max, y = wss) + geom_point() + geom_line()+
  labs(x = "Number of clusters K", y = "Total within-clusters sum of
squares")+
  scale_x_continuous(breaks = seq(0,10,1))+
  ggtitle("The Elbow Method")
```

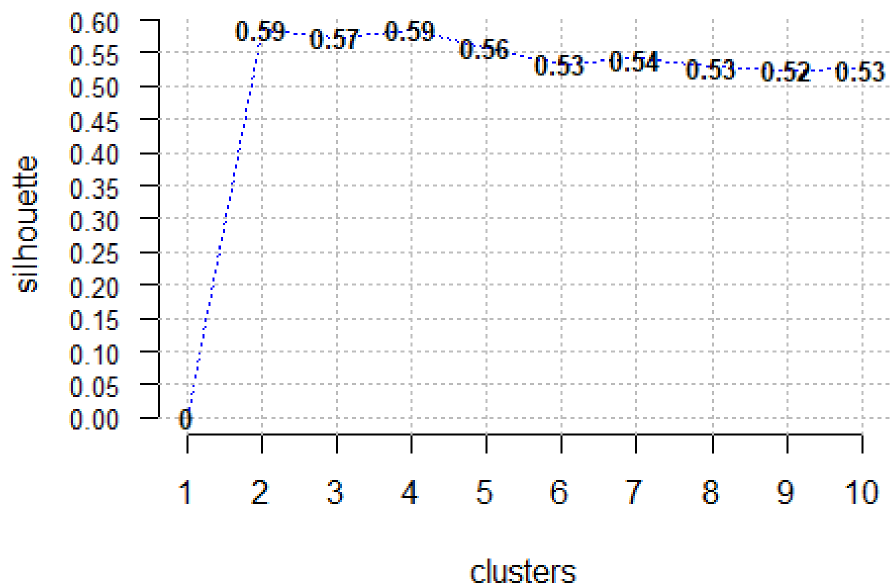



```
fviz_nbclust(clustering.df, FUN = kmeans,method = "wss" ,nstart = 50)
```



#Cannot determine optimal number of clusters, when looking at the Elbow Method, therefore using silhouette score.

```
opt.k.sil<- Optimal_Clusters_KMeans(clustering.df, max_clusters=10, plot_clusters=TRUE,
                                    criterion="silhouette")
```




```
#both 2 and 4 number of clusters generated a high silhouette score of 5.9
#combining that with the WSS output we can conclude that the optimal number of clusters would be 4.
```

#Clustering

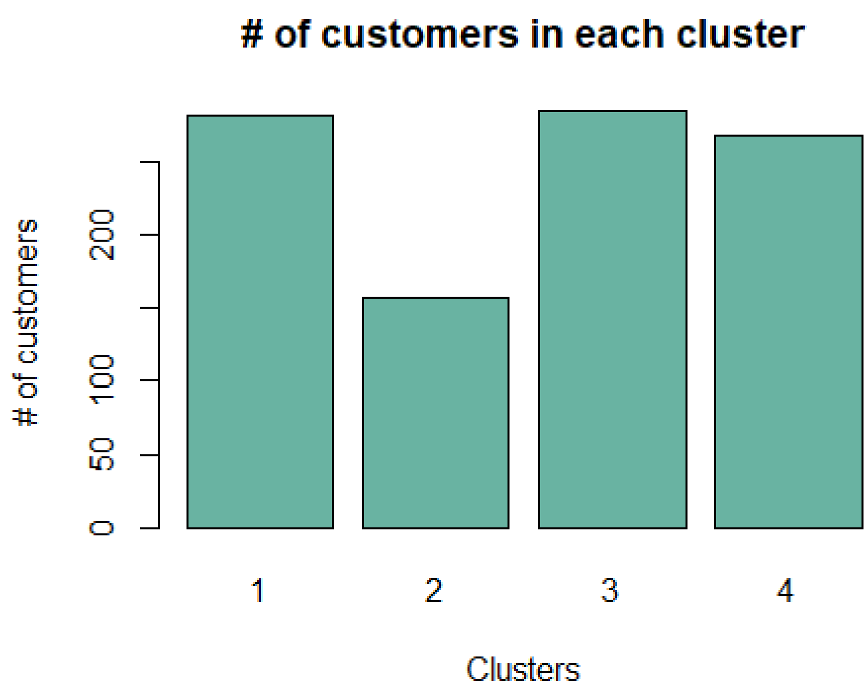
```
retail.df$History <- NULL
retail.df <- raw.data[!is.na(raw.data$AmountSpent),]
KMC <- kmeans(clustering.df,centers = 4,iter.max = 999, nstart=50)
retail.clustered <- (cbind(retail.df, cluster= KMC$cluster))
```

#Creating a new DF, consisted with the original DF with the cluster number for each observation.

```
table_of_cluster_distribution <- table(retail.clustered$cluster)
#   The result:
#   1   2   3   4
# 283 157 285 269
table_of_cluster_distribution
```

```
##
##   1   2   3   4
## 283 157 285 269
```

```r
barplot(table_of_cluster_distribution, xlab="Clusters",
        ylab="# of customers", main="# of customers in each cluster",
        col="#69b3a2")
```




```r
retail.clustered <- data.table(retail.clustered)
retail.clustered[, avg_AmountSpent_in_cluster :=
mean(AmountSpent),by=list(cluster)]
retail.clustered[, avg_SalarySpent_in_cluster :=
mean(Salary),by=list(cluster)]

retail.clustered  <-  retail.clustered[, c("Age", "Gender", "OwnHome",
"Married",
         "Location", "Children", "Catalogs", "Salary","AmountSpent",
         "avg_AmountSpent_in_cluster", "avg_SalarySpent_in_cluster",
"cluster" )]

cluster_1 <- retail.clustered[retail.clustered$cluster==1,]
cluster_2 <- retail.clustered[retail.clustered$cluster==2,]
cluster_3 <- retail.clustered[retail.clustered$cluster==3,]
cluster_4 <- retail.clustered[retail.clustered$cluster==4,]
```