MichaelBasta_FinalProject

Michael Basta

2022-12-12

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

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

```
library(ISLR)
library(flexclust)
```

```
## Loading required package: grid
## Loading required package: lattice
## Loading required package: modeltools
## Loading required package: stats4
```

```
MallCustomers <- read.csv("/Users/michaelbasta/Documents/Fundmentals of Machine Learning /Final/Mall Cu

paste("Data used is Mall Customers data consisting of 5 columns
      (CustomerId - Gender - Age - Annual Income in K - Spending Score 1-100)")
```

```
## [1] "Data used is Mall Customers data consisting of 5 columns \n      (CustomerId - Gender - Age - Ar
```

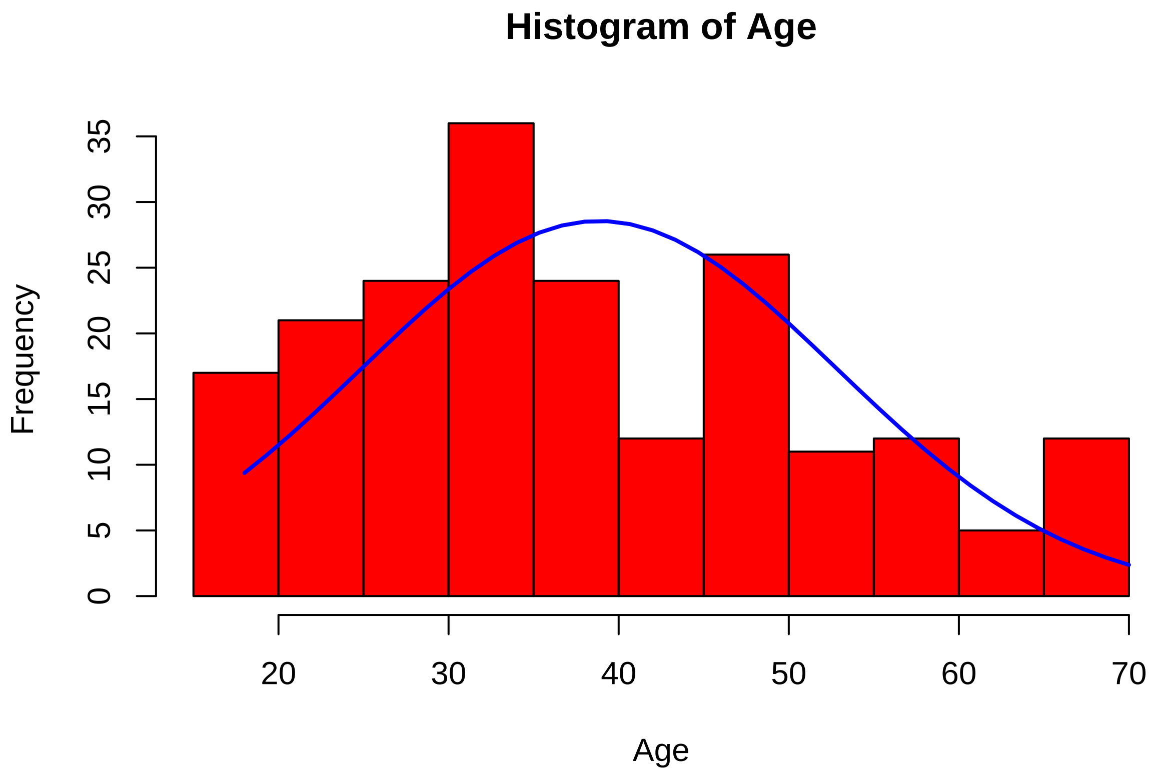
```
head(MallCustomers)
```

```
##   CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
## 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
```

```
paste("Histogram plot to show the destribution of Age in the Customers data")
```

```
## [1] "Histogram plot to show the destribution of Age in the Customers data"
```

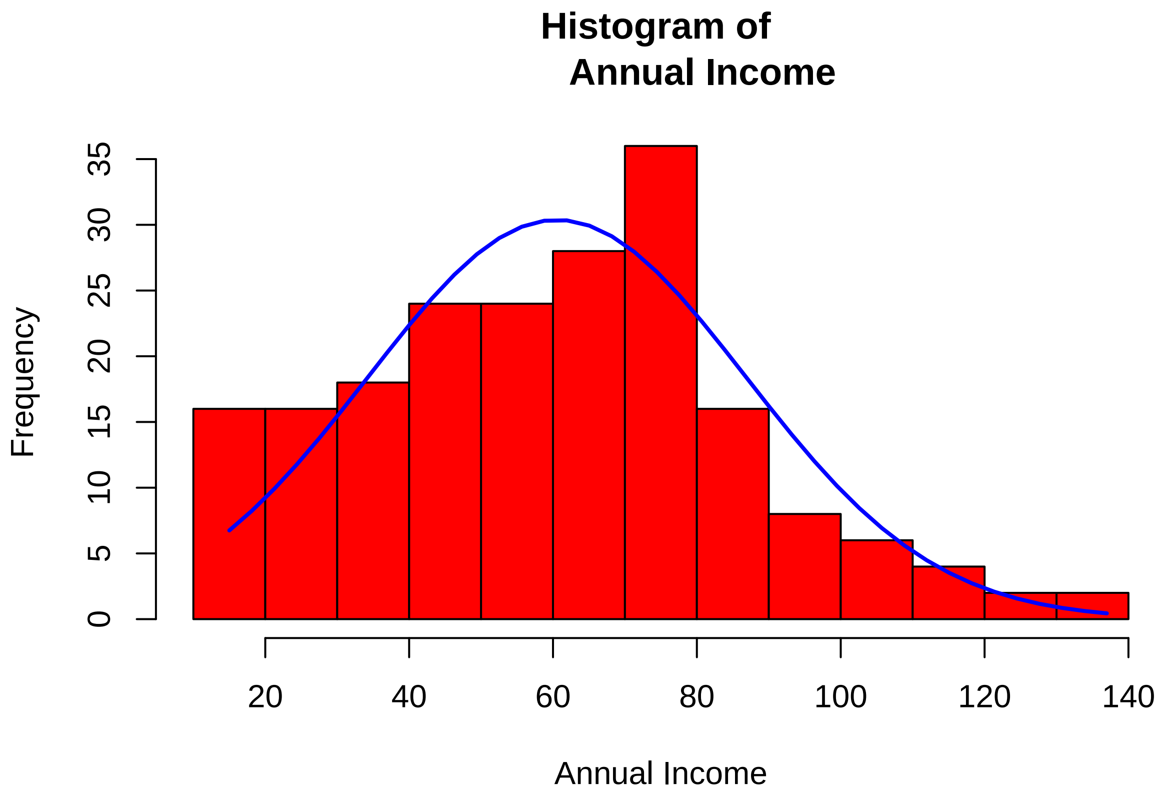
```
x <- MallCustomers$Age
h<-hist(x, main = "Histogram of Age", col = "red", xlab="Age")
xfit <- seq(min(x), max(x), length=40)
yfit <- dnorm(xfit, mean = mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



```
paste("Histogram plot to show the destribution of
      Annual Income in the Customers data")
```

```
## [1] "Histogram plot to show the destribution of \n      Annual Income in the Customers data"
```

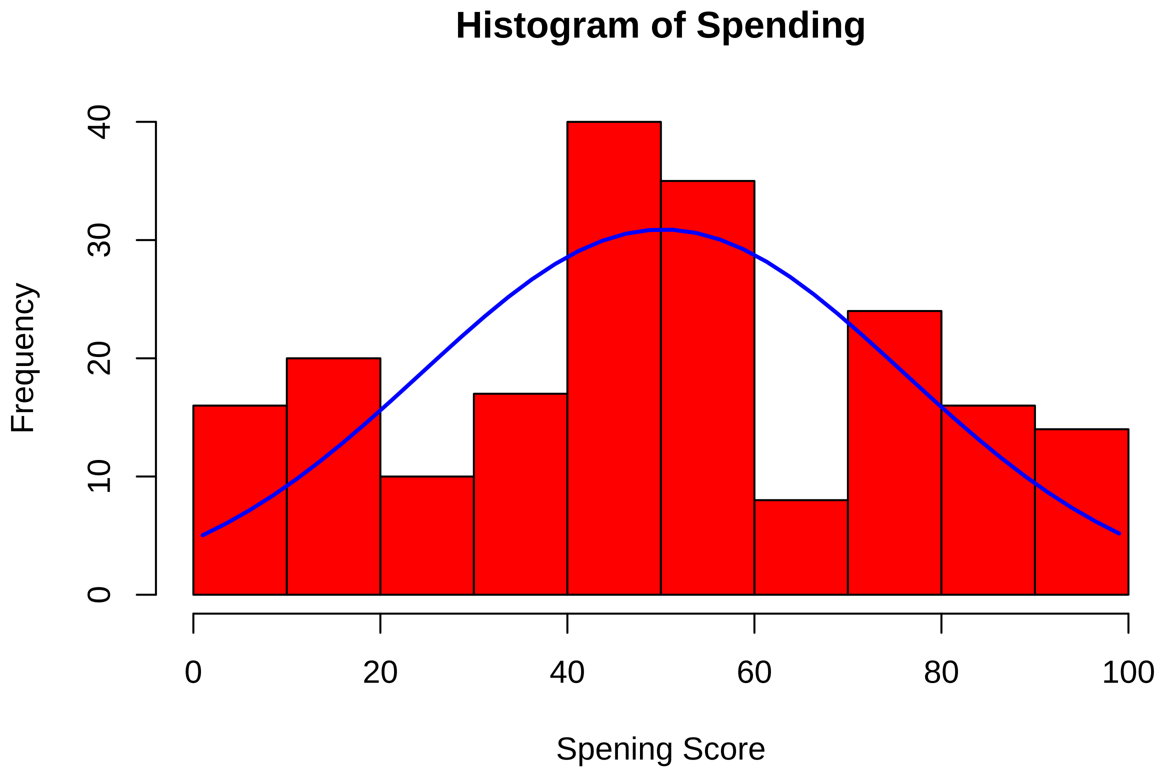
```
x <- MallCustomers$Annual.Income..k..
h<-hist(x, main = "Histogram of
        Annual Income", col = "red", xlab="Annual Income")
xfit <- seq(min(x), max(x), length=40)
yfit <- dnorm(xfit, mean = mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



```
paste("Histogram plot to show the destribution of
      Spending Score in the Customers data")
```

```
## [1] "Histogram plot to show the destribution of \n      Spending Score in the Customers data"
```

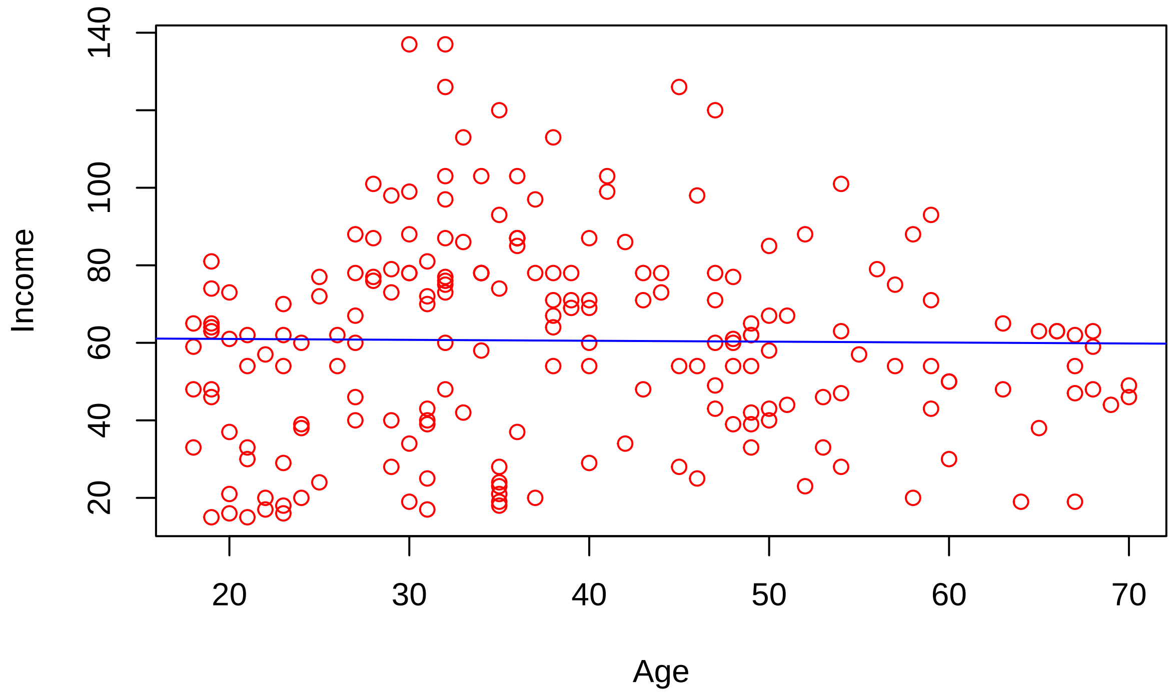
```
x <- MallCustomers$Spending.Score..1.100.
h<-hist(x, main = "Histogram of Spending", col = "red", xlab="Spening Score")
xfit <- seq(min(x), max(x), length=40)
yfit <- dnorm(xfit, mean = mean(x),sd=sd(x))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=2)
```



```
paste("Plot for Age Against Income to see corrolation")
```

```
## [1] "Plot for Age Against Income to see corrolation"
```

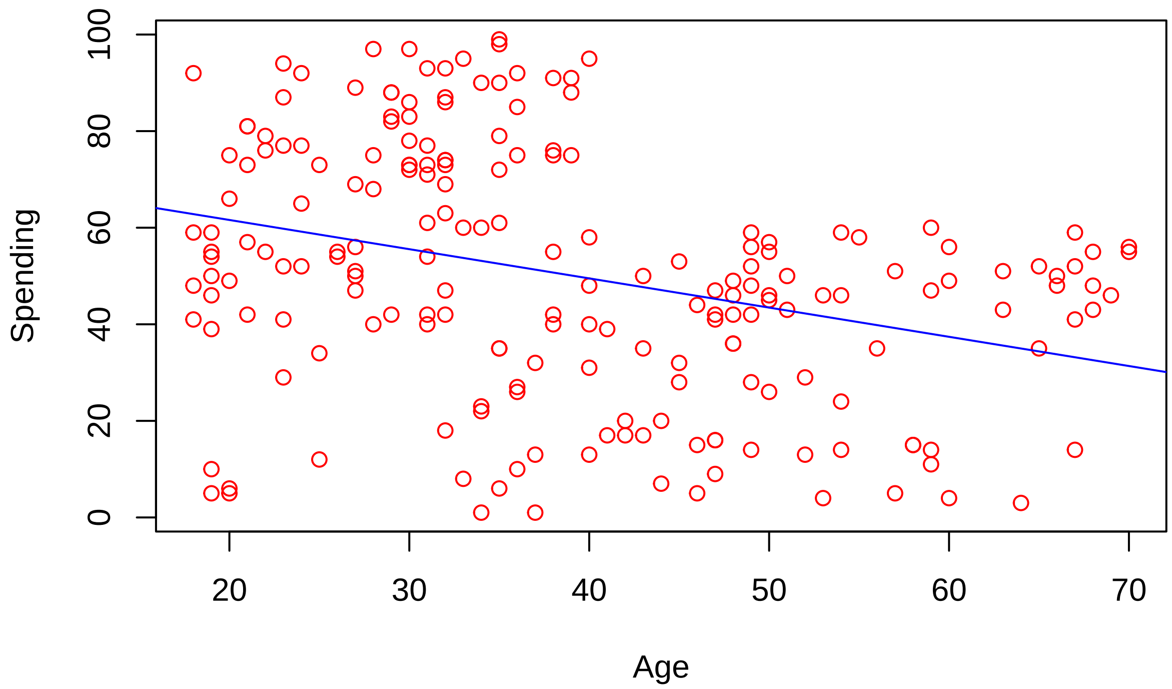
```
y <- MallCustomers$Annual.Income..k..
x <- MallCustomers$Age
plot(x, y, col="red", xlab = "Age", ylab = "Income")
abline(lm(y ~ x), col='blue')
```



```
paste("Plot for Age Against Spending Score to see corrolation")
```

```
## [1] "Plot for Age Against Spending Score to see corrolation"
```

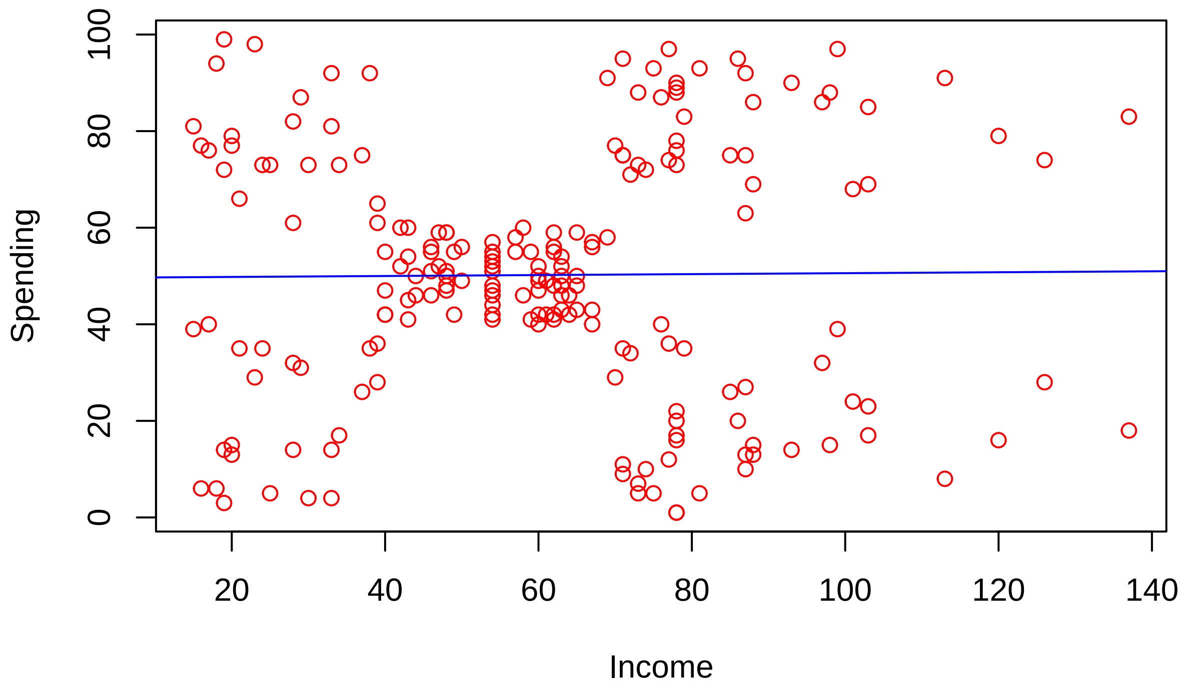
```
y <- MallCustomers$Spending.Score..1.100.
x <- MallCustomers$Age
plot(x, y, col="red", xlab = "Age", ylab = "Spending")
abline(lm(y ~ x), col='blue')
```



```
paste("Plot for Income Against Spending Score to see corrolation")
```

```
## [1] "Plot for Income Against Spending Score to see corrolation"
```

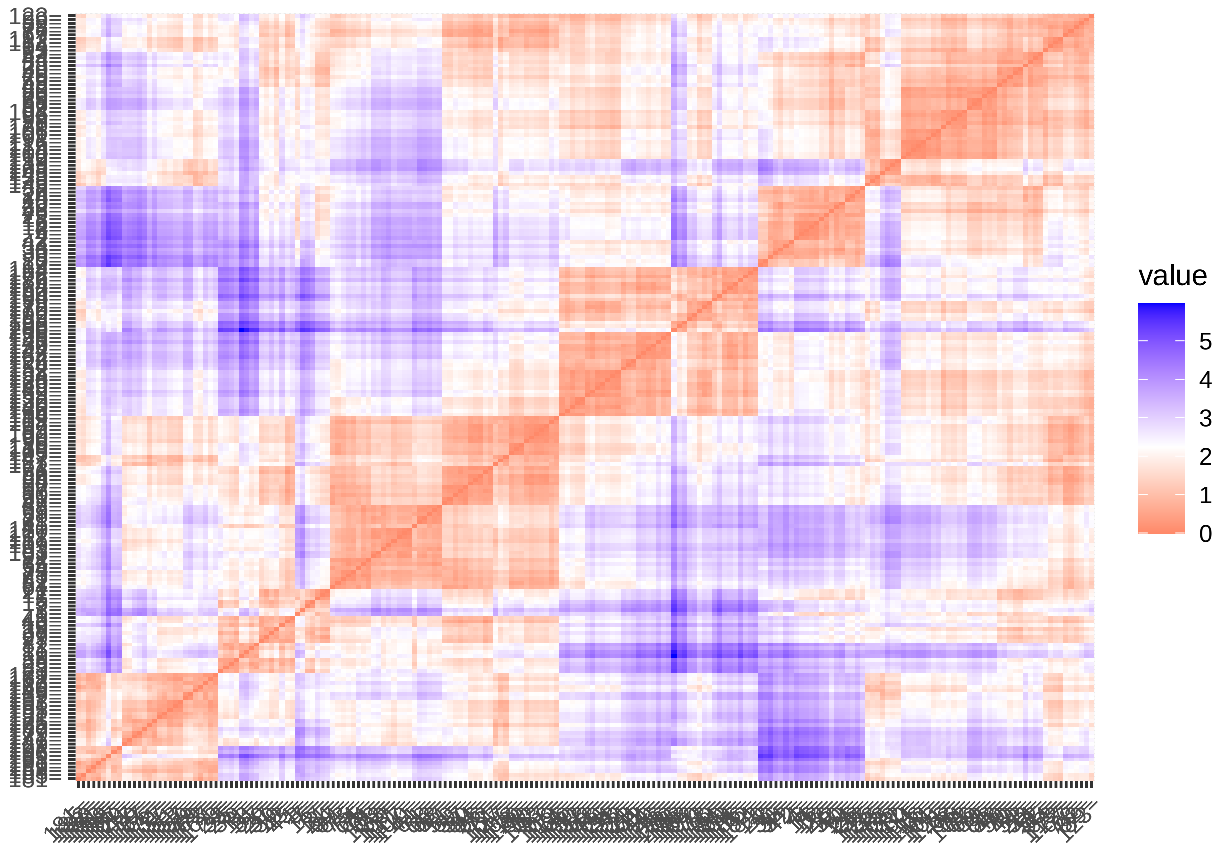
```
y <- MallCustomers$Spending.Score..1.100.
x <- MallCustomers$Annual.Income..k..
plot(x, y, col="red", xlab = "Income", ylab = "Spending")
abline(lm(y ~ x), col='blue')
```



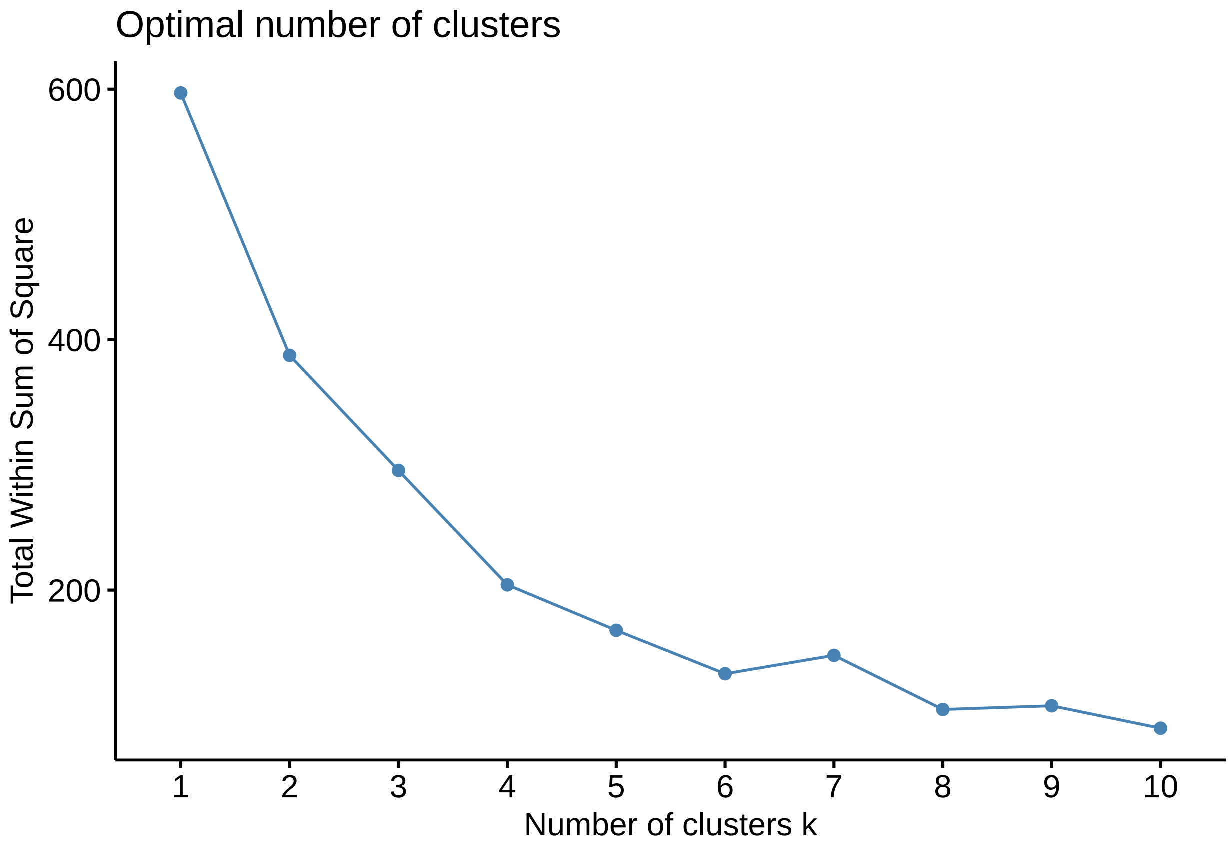
```
df <- MallCustomers[,3:5]
summary(df)
```

```
##       Age        Annual.Income..k.. Spending.Score..1.100.
##  Min.   :18.00   Min.   : 15.00     Min.   : 1.00
##  1st Qu.:28.75   1st Qu.: 41.50     1st Qu.:34.75
##  Median :36.00   Median : 61.50     Median :50.00
##  Mean   :38.85   Mean   : 60.56     Mean   :50.20
##  3rd Qu.:49.00   3rd Qu.: 78.00     3rd Qu.:73.00
##  Max.   :70.00   Max.   :137.00     Max.   :99.00
```

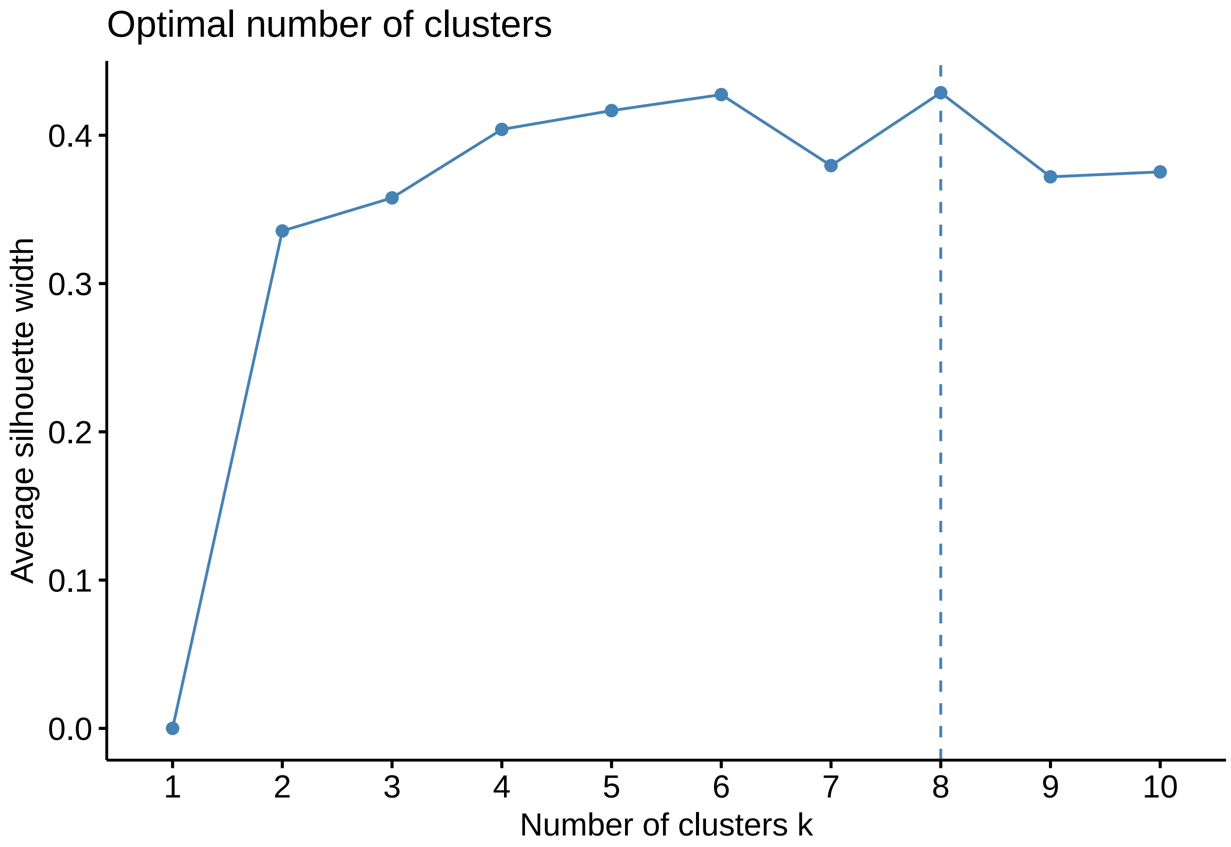
```
# It doesn't seem that there's a direct
# correlation between variables so
# Euclidean distance should be suitable in this case
# Also since it is scale dependant we had to scale it before applying
df <- scale(df)
distance <- get_dist(df)
fviz_dist(distance)
```



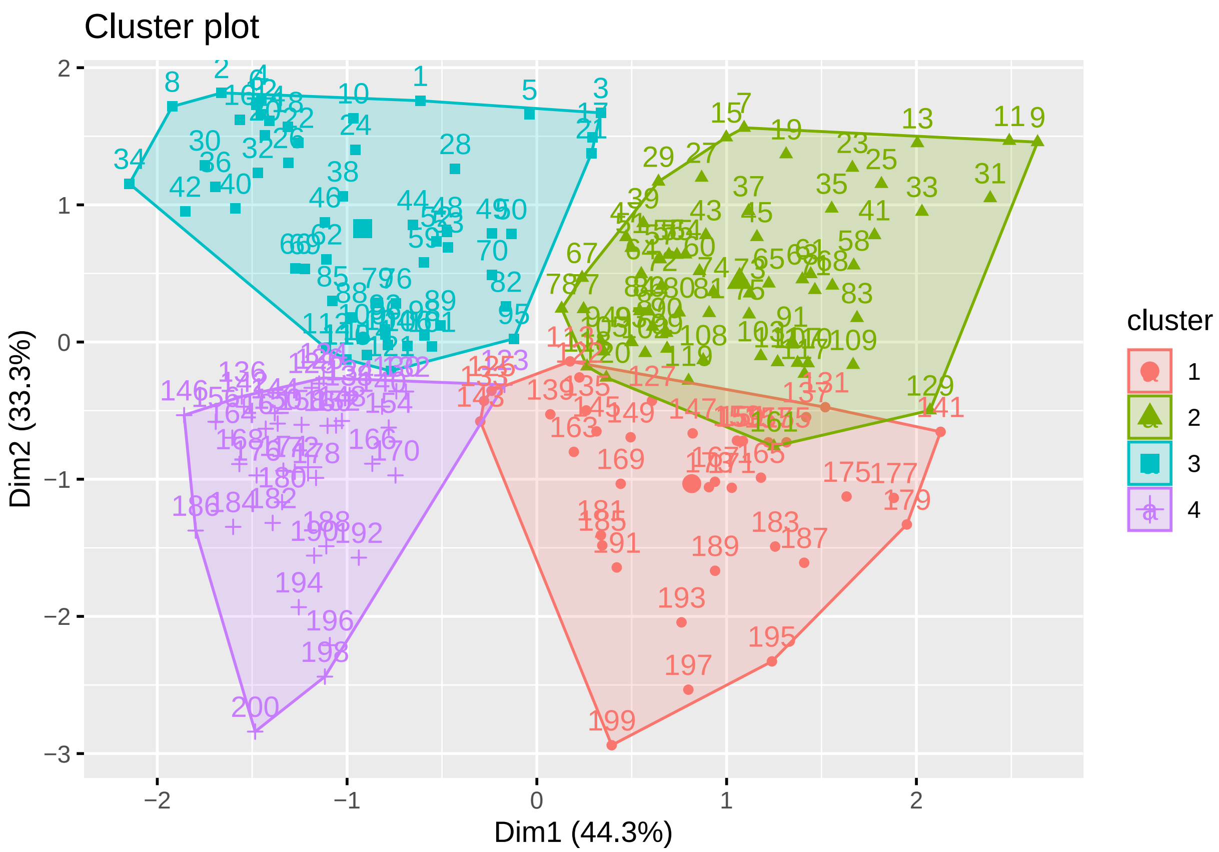
```r
fviz_nbclust(df, kmeans, method="wss")
```



```
fviz_nbclust(df, kmeans, method="silhouette")
```



```
# From the earlier step it shows that 4 clusters is the best choice
k4 <- kmeans(df, centers=4, nstart=25)
fviz_cluster(k4, data = df)
```



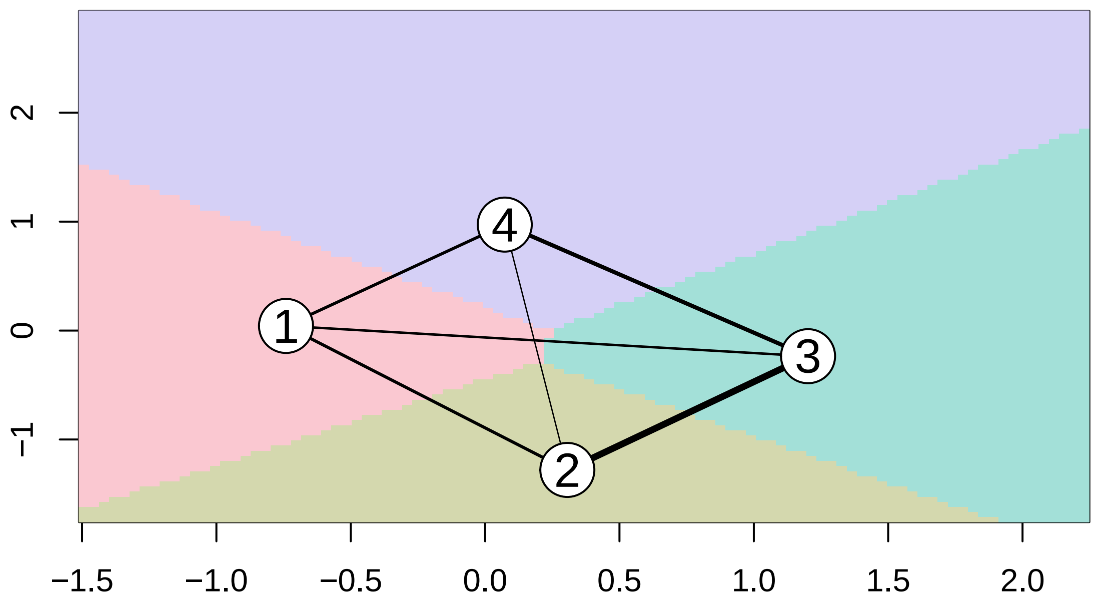
```
k4$size
```

```
## [1] 38 65 57 40
```

```
k4$centers
```

```
##            Age Annual.Income..k.. Spending.Score..1.100.
## 1  0.03711223          0.9876366             -1.1857814
## 2  1.08344244         -0.4893373             -0.3961802
## 3 -0.96008279         -0.7827991              0.3910484
## 4 -0.42773261          0.9724070              1.2130414
```

```
k4 = kcca(df, k=4, kccaFamily("kmeans"))
clusters_index = predict(k4)
image(k4)
```


1
2
3
4
2
1
0
−1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0