# TD2_Descrip_des_données.R

r2342438

2023-11-03

```
## Heart Attack Risk Prediction Dataset Generated by CHATGPT :  Sourav BANERJEE , kaggle
##https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

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
```

```
df<- read.csv("https://dl.dropboxusercontent.com/scl/fi/81s470lw7qksgii98zp5z/heart_attack_prediction_da
str(df)
```

```
## 'data.frame':    8763 obs. of  26 variables:
##  $ Patient.ID                     : chr  "BMW7812" "CZE1114" "BNI9906" "JLN3497" ...
##  $ Age                            : int  67 21 21 84 66 54 90 84 20 43 ...
##  $ Sex                            : chr  "Male" "Male" "Female" "Male" ...
##  $ Cholesterol                    : int  208 389 324 383 318 297 358 220 145 248 ...
##  $ Blood.Pressure                 : chr  "158/88" "165/93" "174/99" "163/100" ...
##  $ Heart.Rate                     : int  72 98 72 73 93 48 84 107 68 55 ...
##  $ Diabetes                       : int  0 1 1 1 1 1 0 0 1 0 ...
##  $ Family.History                 : int  0 1 0 1 1 1 0 0 0 1 ...
##  $ Smoking                        : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ Obesity                        : int  0 1 0 0 1 0 0 1 1 1 ...
##  $ Alcohol.Consumption            : int  0 1 0 1 0 1 1 1 0 1 ...
##  $ Exercise.Hours.Per.Week        : num  4.17 1.81 2.08 9.83 5.8 ...
##  $ Diet                           : chr  "Average" "Unhealthy" "Healthy" "Average" ...
##  $ Previous.Heart.Problems        : int  0 1 1 1 1 1 0 0 0 0 ...
##  $ Medication.Use                 : int  0 0 1 0 0 1 0 1 0 0 ...
##  $ Stress.Level                   : int  9 1 9 9 6 2 7 4 5 4 ...
##  $ Sedentary.Hours.Per.Day        : num  6.62 4.96 9.46 7.65 1.51 ...
##  $ Income                         : int  261404 285768 235282 125640 160555 241339 190450 122093 2508
##  $ BMI                            : num  31.3 27.2 28.2 36.5 21.8 ...
##  $ Triglycerides                  : int  286 235 587 378 231 795 284 370 790 232 ...
##  $ Physical.Activity.Days.Per.Week: int  0 1 4 3 1 5 4 6 7 7 ...
##  $ Sleep.Hours.Per.Day            : int  6 7 4 4 5 10 10 7 4 7 ...
##  $ Country                        : chr  "Argentina" "Canada" "France" "Canada" ...
##  $ Continent                      : chr  "South America" "North America" "Europe" "North America" ..
##  $ Hemisphere                     : chr  "Southern Hemisphere" "Northern Hemisphere" "Northern Hemis
##  $ Heart.Attack.Risk              : int  0 0 0 0 0 1 1 1 0 0 ...
```

```
######################################################################################################
# Les variables quantitatives : Age , Cholesterol | Heart Rate | Exercise Hours per Week

# Les variables qualitatives ; * Nominal: Catégorique Binaire ::  Medication use|Previous Heart Problem
#                              * Heart attack Risk*

######################################################################################################
## Conversion des variables binaire as.factor

col_convert <- c("Diabetes","Family.History","Smoking","Obesity","Alcohol.Consumption","Previous.Heart.I
for (col in col_convert) {
  df[[col]] <- as.factor(df[[col]])
}
str(df)
```

```
## 'data.frame':    8763 obs. of  26 variables:
##  $ Patient.ID                     : chr  "BMW7812" "CZE1114" "BNI9906" "JLN3497" ...
##  $ Age                            : int  67 21 21 84 66 54 90 84 20 43 ...
##  $ Sex                            : Factor w/ 2 levels "Female","Male": 2 2 1 2 2 1 2 2 2 1 ...
##  $ Cholesterol                    : int  208 389 324 383 318 297 358 220 145 248 ...
##  $ Blood.Pressure                 : chr  "158/88" "165/93" "174/99" "163/100" ...
##  $ Heart.Rate                     : int  72 98 72 73 93 48 84 107 68 55 ...
##  $ Diabetes                       : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 1 1 2 1 ...
##  $ Family.History                 : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 1 1 1 2 ...
##  $ Smoking                        : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 2 2 2 ...
##  $ Obesity                        : Factor w/ 2 levels "0","1": 1 2 1 1 2 1 1 2 2 2 ...
##  $ Alcohol.Consumption            : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 2 2 1 2 ...
##  $ Exercise.Hours.Per.Week        : num  4.17 1.81 2.08 9.83 5.8 ...
##  $ Diet                           : Factor w/ 3 levels "Average","Healthy",..: 1 3 2 1 3 3 2 1 1 3 .
##  $ Previous.Heart.Problems        : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 1 1 1 1 ...
##  $ Medication.Use                 : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 1 ...
##  $ Stress.Level                   : int  9 1 9 9 6 2 7 4 5 4 ...
##  $ Sedentary.Hours.Per.Day        : num  6.62 4.96 9.46 7.65 1.51 ...
##  $ Income                         : int  261404 285768 235282 125640 160555 241339 190450 122093 2508
##  $ BMI                            : num  31.3 27.2 28.2 36.5 21.8 ...
##  $ Triglycerides                  : int  286 235 587 378 231 795 284 370 790 232 ...
##  $ Physical.Activity.Days.Per.Week: int  0 1 4 3 1 5 4 6 7 7 ...
##  $ Sleep.Hours.Per.Day            : int  6 7 4 4 5 10 10 7 4 7 ...
##  $ Country                        : chr  "Argentina" "Canada" "France" "Canada" ...
##  $ Continent                      : chr  "South America" "North America" "Europe" "North America" ..
##  $ Hemisphere                     : chr  "Southern Hemisphere" "Northern Hemisphere" "Northern Hemis
##  $ Heart.Attack.Risk              : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 2 1 1 ...
```

```
## Création d'une nouvelle colonne avec 3 levels (Hypertension | Hypotension et Normal) pour la tension
#df$Blood.Pressure
df$Blood.Pressure <- sapply(df$Blood.Pressure, function(bp) {
  systolic <- as.numeric(gsub("/.*", "", bp))
  diastolic <- as.numeric(gsub(".*/", "", bp))

  if (systolic <= 120 && diastolic <= 80) {
    return("Normal")
  } else if (systolic > 140 || diastolic > 90) {
    return("Hypertension")
  } else {
    return("Hypotension")
  }
}) %>% as.factor()

########################################################################################################
##############################################  Age  ###################################################
# paramètre de position
mean(df$Age)
```

```
## [1] 53.70798
```

```
median(df$Age)
```

```
## [1] 54
```

```
quantile(df$Age, probs = c(0.25, 0.5,0.75))
```

```
## 25% 50% 75%
##  35  54  72
```

```
# paramètre de dispertion

sd(df$Age)
```

```
## [1] 21.24951
```

```
max(df$Age)
```

```
## [1] 90
```

```
min(df$Age)
```

```
## [1] 18
```

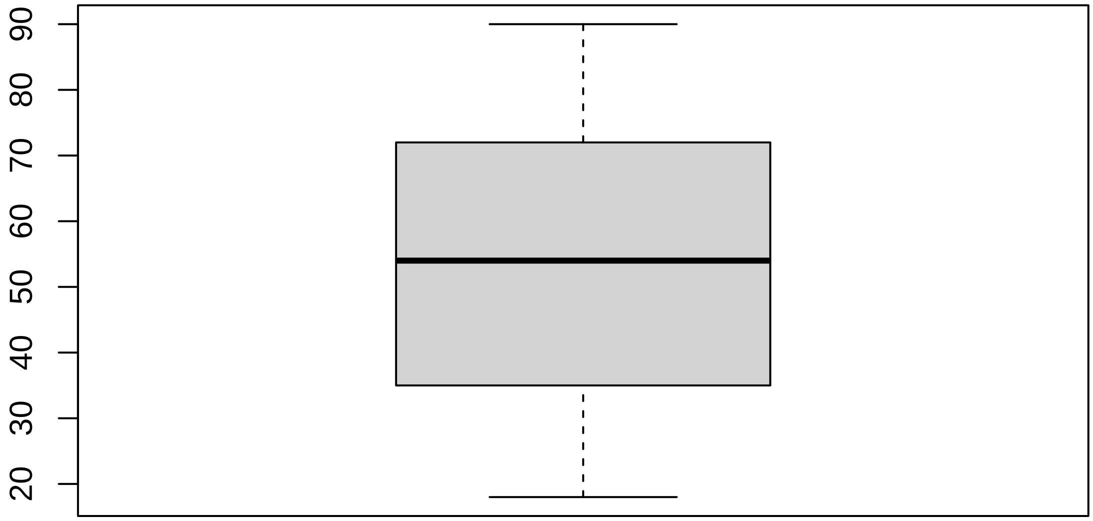
```
range(df$Age)
```

```
## [1] 18 90
```

```
IQR(df$Age)
```

```
## [1] 37
```

```
# Preview using box plot
boxplot(df$Age)
```



```
#commentaires :
#ensemble de données sur l'âge semble être relativement symétrique,
#avec un âge moyen de 53,71 ans, et une quantité modérée de variabilité
#(comme indiqué par l'écart type). La plage d'âges va de 18 à 90 ans,
#et les quartiles et l'IQR offrent des informations sur la distribution
#des âges dans votre ensemble de données.


##################################  BMI ##################################################
# paramètre de position
mean(df$BMI)
```

```
## [1] 28.89145
```

```
median(df$BMI)
```

```
## [1] 28.769
```

```
quantile(df$BMI, probs = c(0.25, 0.5,0.75))
```

```
##      25%      50%      75%
## 23.42299 28.76900 34.32459
```

```
# paramètre de dispertion

sd(df$BMI)
```

```
## [1] 6.319181
```

```
max(df$BMI)
```

```
## [1] 39.99721
```

```
min(df$BMI)
```

```
## [1] 18.00234
```

```
range(df$BMI)
```

```
## [1] 18.00234 39.99721
```

```
IQR(df$BMI)
```

```
## [1] 10.90161
```

```
# Preview using box plot
boxplot(df$BMI)



## commentaires :
#données sur l'IMC montre une distribution légèrement asymétrique avec une moyenne
#d'environ 28,89. La plage de l'IMC va de 18,00 à 39,99, et les quartiles
#et l'écart type fournissent des informations sur la distribution
#de l'IMC dans votre ensemble de données.

########################################################################################################
#                                                                                                      #
#                               VARIABLES QUALITATIVES                                                 #
#                                                                                                      #
########################################################################################################
###############################    Heart Attack Risk  ##################################################

table(df$Heart.Attack.Risk)
```

```
##
##    0    1
## 5624 3139
```

```
table(df$Heart.Attack.Risk) %>% prop.table()*100
```

```
##
##        0        1
## 64.17893 35.82107
```

```
#install.packages("ggplot")
#library(ggplot)
#ggplot(df, aes(x=Heart.Attack.Risk)) + geom_bar()


#######################################     Sex   ######################################################
table(df$Sex) #summary(df$Sex)
```

```
##
## Female   Male
##   2652   6111
```

```
prop.table(table(df$Sex))*100
```

```
##
##   Female     Male
## 30.26361 69.73639
```

```
#ggplot(df, aes(x = Sex)) + geom_bar()


#######################################     Blood pressure   ###########################################

table(df$Blood.Pressure) # summary(df$Blood.Pressure)
```

```
##
```

```
## Hypertension  Hypotension       Normal
##         5814         1684         1265
```

```
prop.table(table(df$Blood.Pressure))*100
```

```
##
## Hypertension  Hypotension       Normal
##     66.34714     19.21716     14.43570
```

```
#ggplot(df, aes(x = Blood.Pressure)) + geom_bar()

#######################################################################################################
#######################################################################################################
#                                                                                                     #
#                               Croisement de variable  QUALI | QUALI                                 #
#                                                                                                     #
#######################################################################################################
# SEX & DIABETES

prop.table(table(df$Heart.Attack.Risk, df$Blood.Pressure),1) ## ligne ie: proportion des gens avec Hyper
```

```
##
##     Hypertension Hypotension    Normal
##   0    0.6591394   0.1961238 0.1447368
##   1    0.6712329   0.1850908 0.1436763
```

```
prop.table(table(df$Heart.Attack.Risk, df$Blood.Pressure),2) ## colonne ie: proportion des gens avec "l
```

```
##
##     Hypertension Hypotension    Normal
##   0    0.6375989   0.6549881 0.6434783
##   1    0.3624011   0.3450119 0.3565217
```

```
prop.table(table(df$Heart.Attack.Risk, df$Blood.Pressure)) ## table ie : proportion des gens avec Hyper
```

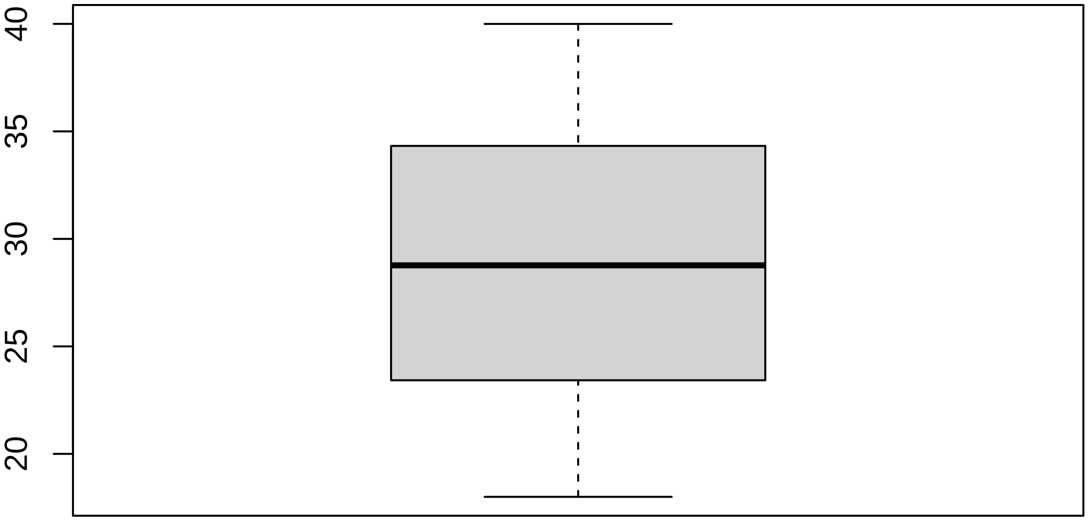
```
##
##     Hypertension Hypotension     Normal
##   0   0.42302864  0.12587014 0.09289056
##   1   0.24044277  0.06630149 0.05146639
```

```
#install.packages("gmodels")
library(gmodels)
CrossTable(df$Heart.Attack.Risk, df$Blood.Pressure, prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  8763
##
##
```

```
##                      | df$Blood.Pressure 
## df$Heart.Attack.Risk | Hypertension |  Hypotension |       Normal |    Row Total | 
## ---------------------|--------------|--------------|--------------|--------------|
##                    0 |         3707 |         1103 |          814 |         5624 | 
##                      |        0.659 |        0.196 |        0.145 |        0.642 | 
##                      |        0.638 |        0.655 |        0.643 |              | 
##                      |        0.423 |        0.126 |        0.093 |              | 
## ---------------------|--------------|--------------|--------------|--------------|
##                    1 |         2107 |          581 |          451 |         3139 | 
##                      |        0.671 |        0.185 |        0.144 |        0.358 | 
##                      |        0.362 |        0.345 |        0.357 |              | 
##                      |        0.240 |        0.066 |        0.051 |              | 
## ---------------------|--------------|--------------|--------------|--------------|
##         Column Total |         5814 |         1684 |         1265 |         8763 | 
##                      |        0.663 |        0.192 |        0.144 |              | 
## ---------------------|--------------|--------------|--------------|--------------|
## 
## 
```

```
#install.packages("ggplot2")
library(ggplot2)
```



```
#ggplot(data = df, aes(x = Blood.Pressure, fill = Heart.Attack.Risk )) +
#geom_bar()




########################################################################################################
#                                                                                                      #
#                               Croisement de variable  QUALI | QUANTI                                #
#                                                                                                      #
########################################################################################################
# Age | Heart attack risk
######################################################

mean(df$Age[which(df$Heart.Attack.Risk==1)])
```

```
## [1] 53.89009
```

```r
mean(df$Age[which(df$Heart.Attack.Risk==0)])
```

```
## [1] 53.60633
```

```r
by(df$Age, df$Heart.Attack.Risk, mean)
```

```
## df$Heart.Attack.Risk: 0
## [1] 53.60633
## ------------------------------------------------------------
## df$Heart.Attack.Risk: 1
## [1] 53.89009
```

```r
# using package
#install.packages("doBy")
library(doBy)
```

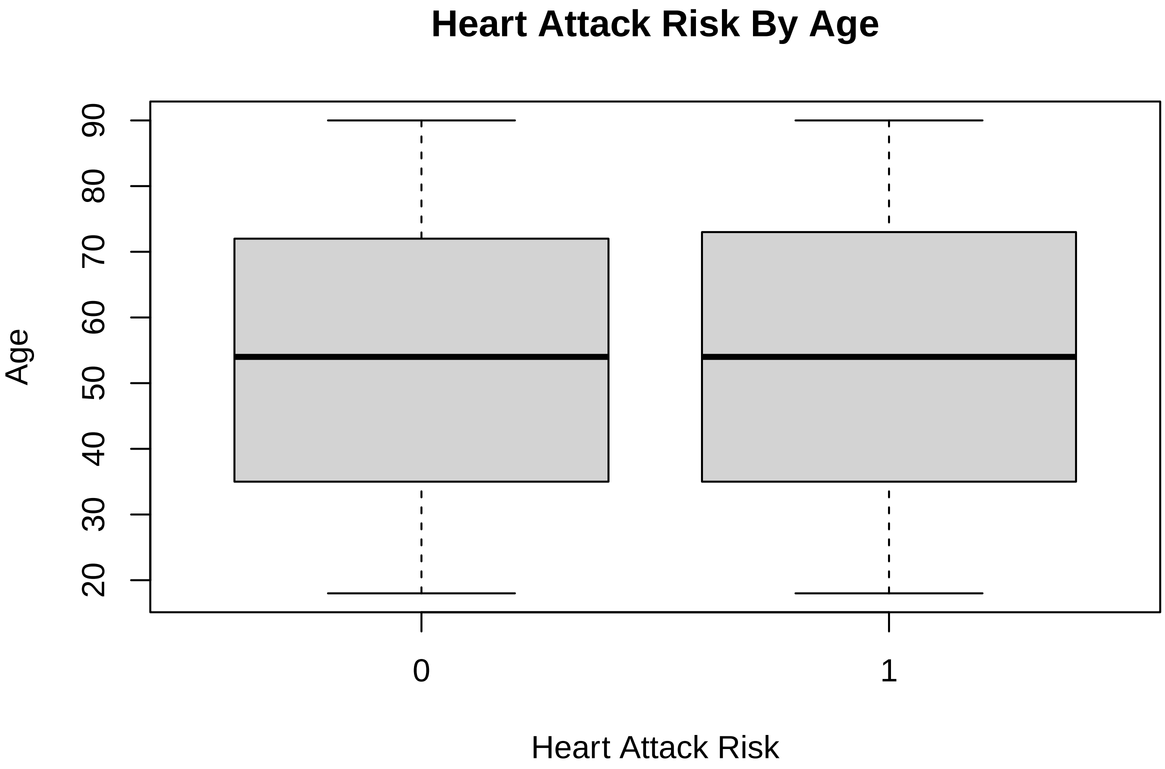
```
##
## Attaching package: 'doBy'
```

```
## The following object is masked from 'package:dplyr':
##
##     order_by
```

```r
summaryBy(Age ~ Heart.Attack.Risk, data = df)
```

```
##   Heart.Attack.Risk Age.mean
## 1                 0 53.60633
## 2                 1 53.89009
```

```r
boxplot(Age ~ Heart.Attack.Risk, data = df, xlab = "Heart Attack Risk", ylab = "Age", main = "Heart Att
```



```
# UN PLUS:
########################################################################################################
#                               Croisement de variable  Qaunti | Quanti                                #
#                                                                                                      #
#                                                                                                      #
########################################################################################################
cor(df$Age, df$Physical.Activity.Days.Per.Week)
```

```
## [1] 0.001383668
```

```
# Croisement de plusieurs variables Quantitatives en utilisant corrplot package
# Corelation Matrix
Num_data <- df[, c("Age","Cholesterol","Heart.Rate","Exercise.Hours.Per.Week","Sedentary.Hours.Per.Day"
                   "Income","BMI","Triglycerides","Physical.Activity.Days.Per.Week","Sleep.Hours.Per.Da
cor(Num_data)
```

```
##                                          Age   Cholesterol    Heart.Rate
## Age                              1.000000000 -9.107011e-03 -0.0038440129
## Cholesterol                     -0.009107011  1.000000e+00  0.0003149083
## Heart.Rate                      -0.003844013  3.149083e-04  1.0000000000
## Exercise.Hours.Per.Week          0.001205639  2.151714e-02  0.0082763293
## Sedentary.Hours.Per.Day          0.017280134  1.891449e-02 -0.0102320484
## Income                          -0.001732790  6.750208e-06  0.0048734774
## BMI                             -0.002611846  1.729187e-02  0.0052985748
## Triglycerides                    0.003414957 -5.453721e-03  0.0122436948
## Physical.Activity.Days.Per.Week  0.001383668  1.605594e-02  0.0008343817
## Sleep.Hours.Per.Day             -0.002184704  4.456229e-03  0.0018112469
##                                 Exercise.Hours.Per.Week Sedentary.Hours.Per.Day
## Age                                         0.001205639            1.728013e-02
## Cholesterol                                 0.021517136            1.891449e-02
## Heart.Rate                                  0.008276329           -1.023205e-02
## Exercise.Hours.Per.Week                     1.000000000            8.755601e-03
## Sedentary.Hours.Per.Day                     0.008755601            1.000000e+00
## Income                                     -0.023413847            3.510621e-03
## BMI                                         0.003776921           -2.356074e-05
## Triglycerides                               0.001716949           -5.784609e-03
## Physical.Activity.Days.Per.Week             0.007725186           -6.178012e-03
## Sleep.Hours.Per.Day                        -0.001245336            4.792013e-03
##                                        Income           BMI Triglycerides
## Age                             -1.732790e-03 -2.611846e-03   0.003414957
## Cholesterol                      6.750208e-06  1.729187e-02  -0.005453721
## Heart.Rate                       4.873477e-03  5.298575e-03   0.012243695
## Exercise.Hours.Per.Week         -2.341385e-02  3.776921e-03   0.001716949
## Sedentary.Hours.Per.Day          3.510621e-03 -2.356074e-05  -0.005784609
## Income                           1.000000e+00  8.835838e-03   0.010738559
## BMI                              8.835838e-03  1.000000e+00  -0.005963607
## Triglycerides                    1.073856e-02 -5.963607e-03   1.000000000
## Physical.Activity.Days.Per.Week  1.302733e-04  8.110375e-03  -0.007556419
## Sleep.Hours.Per.Day             -6.598343e-03 -1.003041e-02  -0.029215971
##                                 Physical.Activity.Days.Per.Week
## Age                                                0.0013836679
## Cholesterol                                        0.0160559355
## Heart.Rate                                         0.0008343817
## Exercise.Hours.Per.Week                            0.0077251861
## Sedentary.Hours.Per.Day                           -0.0061780115
## Income                                          0.0001302733
## BMI                                             0.0081103748
## Triglycerides                                  -0.0075564192
## Physical.Activity.Days.Per.Week                 1.0000000000
## Sleep.Hours.Per.Day                             0.0140334379
##                                 Sleep.Hours.Per.Day
## Age                                    -0.002184704
## Cholesterol                             0.004456229
## Heart.Rate                              0.001811247
## Exercise.Hours.Per.Week                -0.001245336
## Sedentary.Hours.Per.Day                 0.004792013
## Income                                 -0.006598343
## BMI                                    -0.010030410
## Triglycerides                          -0.029215971
## Physical.Activity.Days.Per.Week         0.014033438
## Sleep.Hours.Per.Day                     1.000000000
```

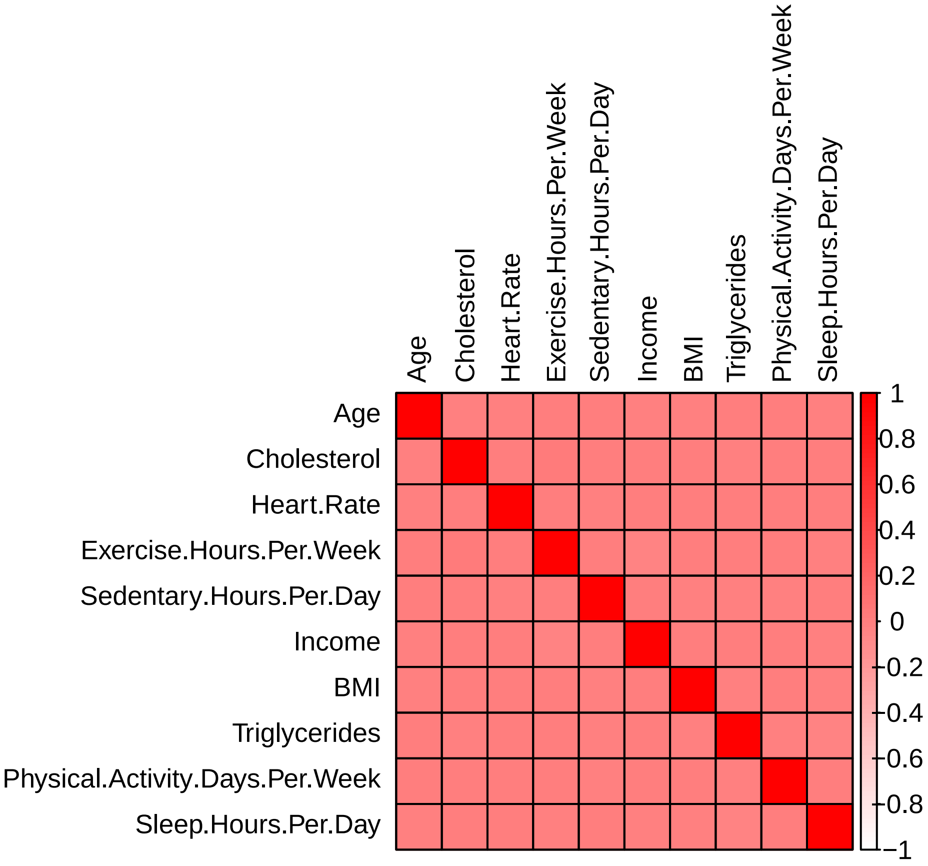
```
correlation_matrix <- cor(Num_data)


#install.packages("corrplot")

library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(
  correlation_matrix,
  method = "color",
  is.corr = TRUE,
  tl.col = "Black",
  col = colorRampPalette(c("white", "Red"))(100),
  tl.srt = 90,
  tl.cex = 0.8,
  addgrid.col = "Black"
)
```


Age
Cholesterol
Heart.Rate
Exercise.Hours.Per.Week
Sedentary.Hours.Per.Day
Income
BMI
Triglycerides
Physical.Activity.Days.Per.Week
Sleep.Hours.Per.Day
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1