

# Assignment1\_ML

```
chooseCRANmirror(graphics = getOption("menu.graphics"), ind = 79,  
                  local.only = FALSE)  
#install.packages("mars1")  
install.packages("dplyr")
```

```
## Installing package into 'C:/Users/ibeme/Documents/R/win-library/4.1'  
## (as 'lib' is unspecified)
```

```
## package 'dplyr' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'dplyr'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying C:  
## \Users\ibeme\Documents\R\win-library\4.1\00LOCK\dplyr\libs\x64\dplyr.dll to C:  
## \Users\ibeme\Documents\R\win-library\4.1\dplyr\libs\x64\dplyr.dll: Permission  
## denied
```

```
## Warning: restored 'dplyr'
```

```
##
```

```
## The downloaded binary packages are in
```

```
## C:\Users\ibeme\AppData\Local\Temp\RtmpiexK0x\downloaded_packages
```

```
#install.packages("eval")  
install.packages("Hmisc")
```

```
## Installing package into 'C:/Users/ibeme/Documents/R/win-library/4.1'  
## (as 'lib' is unspecified)
```

```
## package 'Hmisc' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'Hmisc'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying C:  
## \Users\ibeme\Documents\R\win-library\4.1\00LOCK\Hmisc\libs\x64\Hmisc.dll to C:  
## \Users\ibeme\Documents\R\win-library\4.1\Hmisc\libs\x64\Hmisc.dll: Permission  
## denied
```

```
## Warning: restored 'Hmisc'
```

```
##
```

```
## The downloaded binary packages are in
```

```
## C:\Users\ibeme\AppData\Local\Temp\RtmpiexK0x\downloaded_packages
```

```
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(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##   src, summarize

## The following objects are masked from 'package:base':
##
##   format.pval, units
```

```
#Heart Disease Dataset extracted from Kaggle
```

```
heart_data<-read.csv("heart.csv")
View(heart_data)
```

```
##### DESCRIPTIVE STATISTICS #####
```

```
describe(heart_data)
```

```
## heart_data
##
## 12 Variables      918 Observations
## -----
## Age
##      n missing distinct      Info      Mean      Gmd      .05      .10
```

```

##      918      0      50      0.999      53.51      10.71      37      40
##      .25      .50      .75      .90      .95
##      47      54      60      65      68
##
## lowest : 28 29 30 31 32, highest: 73 74 75 76 77
## -----
## Sex
##      n missing distinct
##      918      0      2
##
## Value      F      M
## Frequency  193  725
## Proportion 0.21 0.79
## -----
## ChestPainType
##      n missing distinct
##      918      0      4
##
## Value      ASY      ATA      NAP      TA
## Frequency  496    173    203    46
## Proportion 0.540 0.188 0.221 0.050
## -----
## RestingBP
##      n missing distinct      Info      Mean      Gmd      .05      .10
##      918      0      67      0.993      132.4      20.09      106      110
##      .25      .50      .75      .90      .95
##      120      130      140      160      160
##
## lowest : 0 80 92 94 95, highest: 180 185 190 192 200
## -----
## Cholesterol
##      n missing distinct      Info      Mean      Gmd      .05      .10
##      918      0      222      0.993      198.7      115.9      0.0      0.0
##      .25      .50      .75      .90      .95
##      173.2    222.5    267.0    305.0    331.3
##
## lowest : 0 85 100 110 113, highest: 491 518 529 564 603
## -----
## FastingBS
##      n missing distinct      Info      Sum      Mean      Gmd
##      918      0      2      0.536      214      0.2331      0.3579
##
## -----
## RestingECG
##      n missing distinct
##      918      0      3
##
## Value      LVH Normal      ST
## Frequency  188    552    178
## Proportion 0.205 0.601 0.194
## -----
## MaxHR
##      n missing distinct      Info      Mean      Gmd      .05      .10
##      918      0      119      1      136.8      29.03      96      103

```

```

##      .25      .50      .75      .90      .95
##      120      138      156      170      178
##
## lowest :  60  63  67  69  70, highest: 190 192 194 195 202
## -----
## ExerciseAngina
##      n missing distinct
##      918      0      2
##
## Value      N      Y
## Frequency  547  371
## Proportion 0.596 0.404
## -----
## Oldpeak
##      n missing distinct      Info      Mean      Gmd      .05      .10
##      918      0      53  0.934  0.8874  1.126  0.0  0.0
##      .25      .50      .75      .90      .95
##      0.0      0.6      1.5      2.3      3.0
##
## lowest : -2.6 -2.0 -1.5 -1.1 -1.0, highest:  4.2  4.4  5.0  5.6  6.2
## -----
## ST_Slope
##      n missing distinct
##      918      0      3
##
## Value      Down  Flat  Up
## Frequency    63  460  395
## Proportion 0.069 0.501 0.430
## -----
## HeartDisease
##      n missing distinct      Info      Sum      Mean      Gmd
##      918      0      2  0.741  508  0.5534  0.4948
##
## -----

```

```
summary(heart_data)
```

```

##      Age      Sex      ChestPainType      RestingBP
## Min.   :28.00 Length:918      Length:918      Min.    : 0.0
## 1st Qu.:47.00 Class :character  Class :character  1st Qu.:120.0
## Median :54.00 Mode  :character  Mode  :character  Median :130.0
## Mean   :53.51                      Mean   :132.4
## 3rd Qu.:60.00                      3rd Qu.:140.0
## Max.   :77.00                      Max.   :200.0
## Cholesterol      FastingBS      RestingECG      MaxHR
## Min.    : 0.0 Min.    :0.0000 Length:918      Min.    : 60.0
## 1st Qu.:173.2 1st Qu.:0.0000 Class :character  1st Qu.:120.0
## Median :222.5 Median :0.0000 Mode  :character  Median :138.0
## Mean    :198.7 Mean    :0.2331                      Mean    :136.8
## 3rd Qu.:267.0 3rd Qu.:0.0000                      3rd Qu.:156.0
## Max.    :603.0 Max.    :1.0000                      Max.    :202.0
## ExerciseAngina      Oldpeak      ST_Slope      HeartDisease
## Length:918      Min.    :-2.6000 Length:918      Min.    :0.0000
## Class :character  1st Qu.: 0.0000 Class :character  1st Qu.:0.0000

```

```
## Mode :character Median : 0.6000 Mode :character Median :1.0000
## Mean : 0.8874 Mean :0.5534
## 3rd Qu.: 1.5000 3rd Qu.:1.0000
## Max. : 6.2000 Max. :1.0000
```

```
#summary of the heart diseases by age
d_age<- heart_data %>% group_by(Age)
dsumm_age<- summarise(d_age,heartdiseaseCount_by_Age=sum(HeartDisease==0))
View(dsumm_age)
dsumm_age
```

```
## # A tibble: 50 x 2
##   Age heartdiseaseCount_by_Age
##   <int> <int>
## 1 28 1
## 2 29 3
## 3 30 1
## 4 31 1
## 5 32 3
## 6 33 1
## 7 34 5
## 8 35 7
## 9 36 4
## 10 37 10
## # ... with 40 more rows
```

```
#Summary of the Heart Disease by gender
d_sex<- heart_data %>% group_by(Sex)
summ_sex<- summarise(d_sex,heartDiseaseCount_by_Gender=sum(HeartDisease==0))
View(summ_sex)
summ_sex
```

```
## # A tibble: 2 x 2
##   Sex heartDiseaseCount_by_Gender
##   <chr> <int>
## 1 F 143
## 2 M 267
```

```
##### DATA TRANSFORMATION #####
#USED MUTATE(), ARRANGE(),FILTER()
#Mutated a new column target with factors 1- No disease(N) and 0 - disease(Y)
```

```
heart_data<-mutate(heart_data,target=factor(heart_data$HeartDisease,levels=c(1,0),labels = c("N","Y")))
```

```
#Arrange the data by age
```

```
heart_data_sorted_by_age<-arrange(heart_data,Age)
View(heart_data_sorted_by_age)
head(heart_data_sorted_by_age)
```

```
## Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
```

```
## 1 28 M ATA 130 132 0 LVH 185
## 2 29 M ATA 120 243 0 Normal 160
## 3 29 M ATA 140 263 0 Normal 170
## 4 29 M ATA 130 204 0 LVH 202
## 5 30 F TA 170 237 0 ST 170
## 6 31 M ASY 120 270 0 Normal 153
## ExerciseAngina Oldpeak ST_Slope HeartDisease target
## 1 N 0.0 Up 0 Y
## 2 N 0.0 Up 0 Y
## 3 N 0.0 Up 0 Y
## 4 N 0.0 Up 0 Y
## 5 N 0.0 Up 0 Y
## 6 Y 1.5 Flat 1 N
```

*# List of patients with heart disease by high cholesterol*

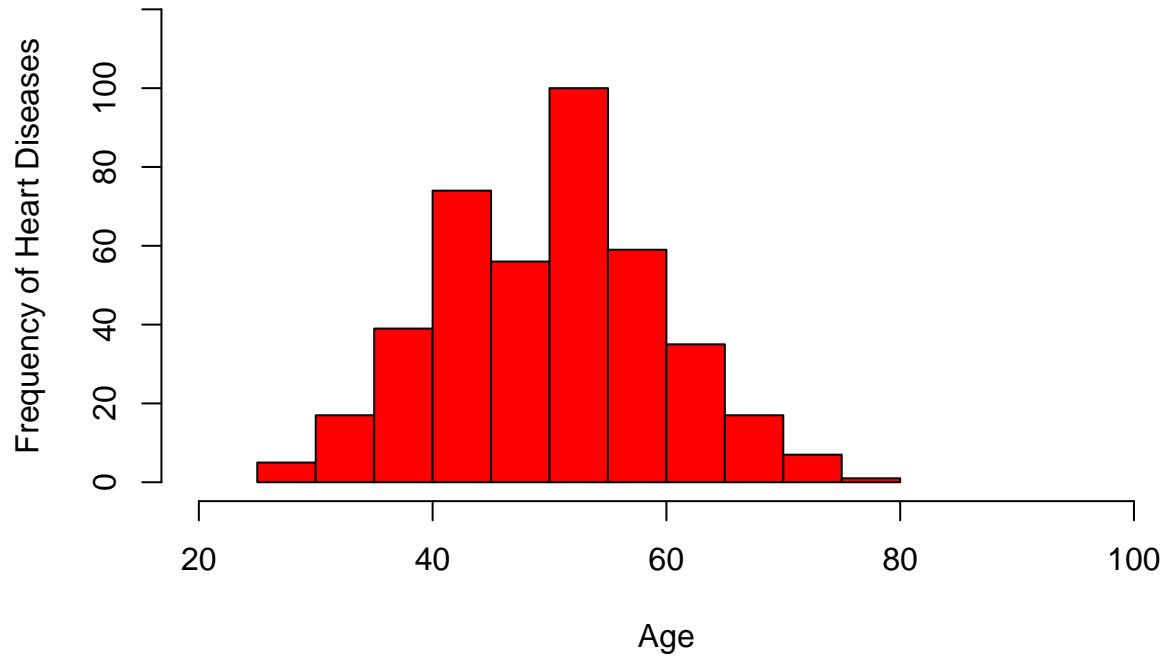
```
List_high_cholesterol<-filter(heart_data_sorted_by_age,heart_data_sorted_by_age$Cholesterol>200,heart_data)
View(List_high_cholesterol)
head(List_high_cholesterol)
```

```
## Age Sex ChestPainType RestingBP Cholesterol FastingBS RestingECG MaxHR
## 1 29 M ATA 120 243 0 Normal 160
## 2 29 M ATA 140 263 0 Normal 170
## 3 29 M ATA 130 204 0 LVH 202
## 4 30 F TA 170 237 0 ST 170
## 5 31 F ATA 100 219 0 ST 150
## 6 32 M ATA 125 254 0 Normal 155
## ExerciseAngina Oldpeak ST_Slope HeartDisease target
## 1 N 0 Up 0 Y
## 2 N 0 Up 0 Y
## 3 N 0 Up 0 Y
## 4 N 0 Up 0 Y
## 5 N 0 Up 0 Y
## 6 N 0 Up 0 Y
```

*#Heart Disease frequency as age increases*

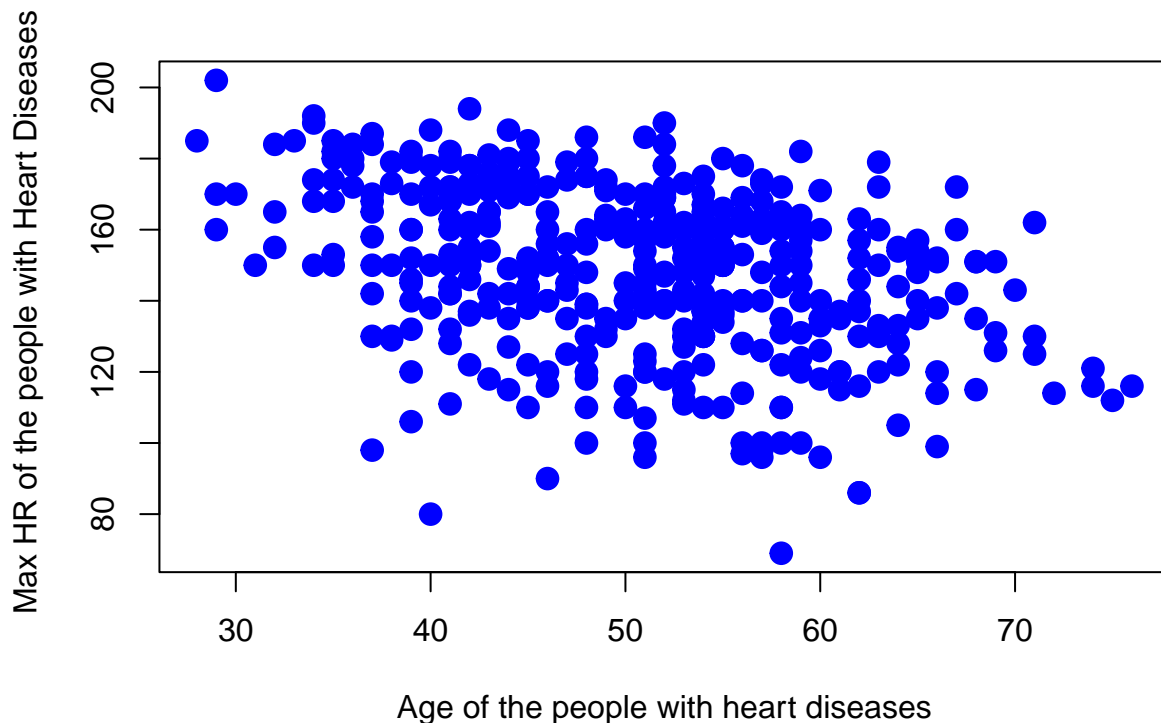
```
hist(heart_data$Age [heart_data$HeartDisease==0],
     xlim = c(20,100) ,#Limiting the scale on x
     ylim = c(0,120),
     breaks = 10,
     main="Heart Diseases Frequency by Age",
     xlab= "Age",
     ylab="Frequency of Heart Diseases",
     col="red")
```

## Heart Diseases Frequency by Age



*#Plot for Maximum heartrate by age in case of heart patients*

```
plot(heart_data$Age[heart_data$HeartDisease==0],heart_data$MaxHR[heart_data$HeartDisease==0],  
     pch=19, #solid circle  
     cex=1.5, #make 150% size  
     col="blue",  
     xlab="Age of the people with heart diseases",  
     ylab=" Max HR of the people with Heart Diseases")
```



## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

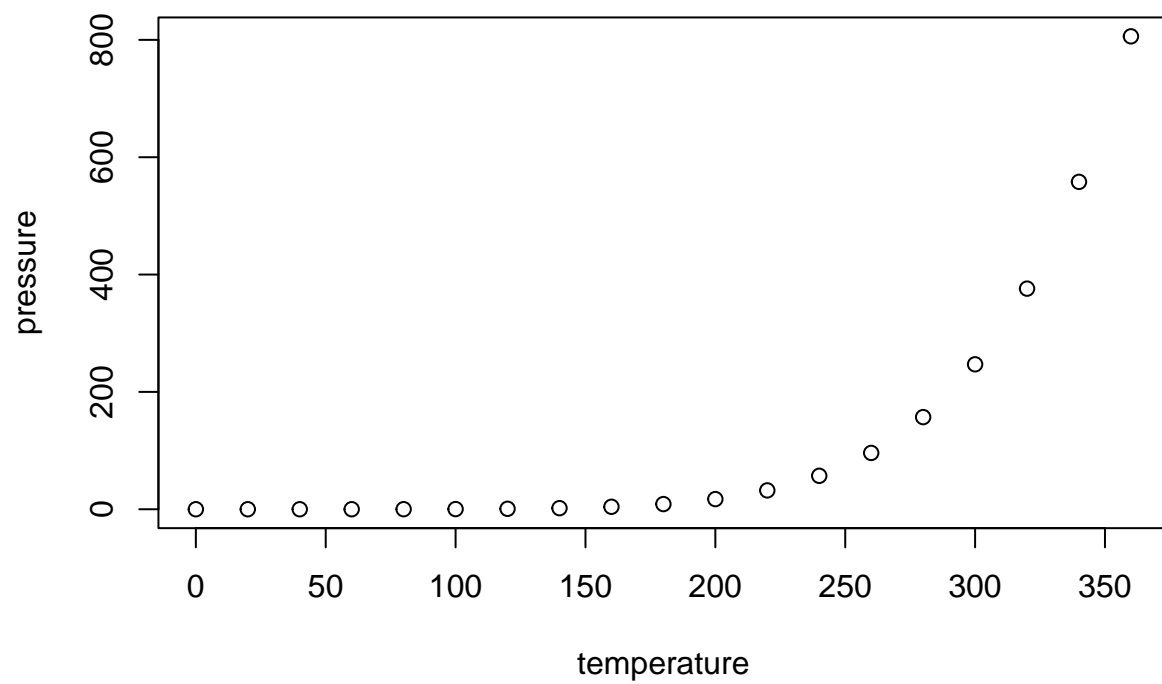
```
summary(cars)
```

```
##      speed      dist
##  Min.   : 4.0    Min.   :  2.00
##  1st Qu.:12.0    1st Qu.: 26.00
##  Median :15.0    Median : 36.00
##  Mean   :15.4    Mean    : 42.98
##  3rd Qu.:19.0    3rd Qu.: 56.00
##  Max.   :25.0    Max.    :120.00
```

## Including Plots

You can also embed plots, for example:





Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.