Myocard Porject

Francielle Mina

15/06/2021

Motivation

This project is a final assessment of the Data Science Professional Certificate. We were encouraged to find any dataset from UCI or Kaggel that we feel comfortable about it. For doing this, we are using the Myocardial infarction complications dataset from UCI repository machine learning. All the variables will explain along with the project.

1.0 Introduction

Until the last decade, world healthy has to change, including human behaviour and diseases. The actual world is fighting to prevent infections is the one possible strategy to deal with the increased hallmark of the current time. Prediction disease and chronic disease help the health system help to develop healthy ageing adults. Besides that, the "new world" already has to face many data generations, which can be suitable for data science. Data science has begun to provide a new set of tools that will leverage enhanced laboratory medicine and reinforce its value in a continuously transforming healthcare ecosystem (Gruson D, et al 2019). Using data science tools, it can work with a multidisciplinary field that uses scientific methods, process, algorithm and system to extracts insights from data (Peck, RW 2020). This current project is part of the Data Science Professional Certificate. To do this, we are applying techniques of data visualisation, exploration and linear regression and machine learning. According to this, we are using data from the ULC repository and predicted outcomes of myocardial infarction. We are going to see if the time of hospitalisation predicts new myocardial infarction. The time variable in this project counts how many time the patients were hospitalised.

Myocardial infarction (MI), commonly known as "heart attack", is irreversible damage caused by prolonged ischemia and hypoxia. That means the heart can receive so much blood (ischemia), or the heart doesn't have blood and oxygen (hypoxia). This a serious medical emergency, meaning the supply of blood to the heart is suddenly blocked, usually by a blood clot. Our heart constantly needs blood and oxygen for a healthy life. When the body stopped or increase the blood flux, can happen a heart attack or MI.

In recent years MI has become one of the most severe diseases in several countries. Despite the treatment, some risk factors can predict MI, increasing the chances of first or second MI, such as age, hypertension, diabetes, and smoking (Fatemeh Kiani, 2016). With advanced medicine, it's easy to find blood tests for helping the diagnostic or prevention of MI. Different of levels enzymes can indicate how good the body and heart are. These enzymes, creatinine phosphokinase, serum creatinine, serum sodium and platelets, are usually inside the cells of your heart. When those cells or heart are injured, these enzymes spread out into your bloodstream. Measuring the levels of these enzymes is a good sign to know how the heart is. Following the World Health Organisation, 85% of deaths cardiovascular diseases caused by MI. It's essential to know to predict this disease to avoiding the patient's deaths.

2.0 Objective

This current project is part final assessment Data Science Professional certificate. The main objective is to predict new outcomes of MI in a patient by time. The variable time is how many times the patient admitted to the University of Leicester hospitals. We are trying to predict if the time of hospitalisation and other health issues can predict new MI.

3.0 Methods and Exploratory Data Analysis (EDA)

For this project, we are using several packages from CRAN to assist our analysis. All the packages will be load along with the development of the project. First of all, we downloaded the dataset from the website and built a linear regression model. After that, split the data into train and test and built a model for machine learning.

3.1 Explore dataset.

First of all, explore and analyse the data set. It's essential to understand how the data are structured, characteristics for better knowledge.

```r
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.4     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: caret

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: data.table

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

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

## The following object is masked from 'package:purrr':
##
##     transpose
```

```
if(!require(dplyr)) install.packages("dplyr", repos = "http://cran.us.r-project.org")
if(!require(RColorBrewer)) install.packages("RColorBrewer", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: RColorBrewer
```

```
if(!require(broom)) install.packages("broom", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: broom
```

```
library(tidyverse)
library(dplyr)
library(caret)
library(broom)
library(RColorBrewer)

read_uci <- read.csv("https://archive.ics.uci.edu/ml/machine-learning-databases/00519/heart_failure_cli


heart_disease <- read_uci
```

3.1.2 Analysing the Missing value

Dowloaded the dataset, we can analysing if there is missing values.

```
missing_value <- table(is.na(heart_disease))
missing_value
```

```
##
## FALSE
##  3887
```

We observe there is no missig_value on data set.

3.1.3 Exploratory Data Analysis (EDA)

This data collected in patients from Jan 2000 to Jan 2020 at the hospital University of Edx, USA. All the patients have a previous diagnostic of MI and health issue.

```
class(heart_disease)
```

```
## [1] "data.frame"
```

The class funtion tells us kind of R object we have. In our case class heart_disease is a data_frame.

```
str(heart_disease)
```

```
## 'data.frame':    299 obs. of  13 variables:
##  $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
##  $ anaemia                 : int  0 0 0 1 1 1 1 1 0 1 ...
##  $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
##  $ diabetes                : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
##  $ high_blood_pressure     : int  1 0 0 0 0 1 0 0 0 1 ...
##  $ platelets               : num  265000 263358 162000 210000 327000 ...
##  $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
##  $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
##  $ sex                     : int  1 1 1 1 0 1 1 1 0 1 ...
##  $ smoking                 : int  0 0 1 0 0 1 0 1 0 1 ...
##  $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
##  $ DEATH_EVENT             : int  1 1 1 1 1 1 1 1 1 1 ...
```

The str function gives us the structure for the heart_disease dataset. Following this, we can notice the dataset has 299 obs and 13 variables. Also, we can observe the name of the variables on the dataset.

```
dim(heart_disease)
```

```
## [1] 299  13
```

This function dim shows us how many rows and columns we have on dataset heart_disease. We can observe 299 rows and 13 variables.

```
head(heart_disease)
```

```
##   age anaemia creatinine_phosphokinase diabetes ejection_fraction
## 1  75       0                      582        0                20
## 2  55       0                     7861        0                38
## 3  65       0                      146        0                20
## 4  50       1                      111        0                20
## 5  65       1                      160        1                20
## 6  90       1                       47        0                40
##   high_blood_pressure platelets serum_creatinine serum_sodium sex smoking time
## 1                   1    265000              1.9          130   1       0    4
## 2                   0    263358              1.1          136   1       0    6
## 3                   0    162000              1.3          129   1       1    7
## 4                   0    210000              1.9          137   1       0    7
## 5                   0    327000              2.7          116   0       0    8
## 6                   1    204000              2.1          132   1       1    8
##   DEATH_EVENT
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

The function head shows us the top few rows or "head" of the dataset heart_disease.

Our project has the variable "sex", but we are assuming the 0 is woman and 1 for man. Following this, we can find the proportion for the "sex" in our case.

What's the proportion of women on dataset heart_disease?

```
mean(heart_disease$sex == 0)
```

```
## [1] 0.3511706
```

We can observe around 0.3511706 or 35% of heart_disease is women.

What's the proportion of men on dataset heart_disease?

```
mean(heart_disease$sex == 1)
```

```
## [1] 0.6488294
```

We can observe around 0.6488294 or 65% of heart_disease is men.

3.2 EDA and Visualisation

In this part of the project, we use the mutate function for better visualisation and add a new variable with gender, male or female.

```
myocard <- heart_disease %>% mutate(gender = ifelse(sex > 0, "male", "female"))
myocard
```

```
##        age anaemia creatinine_phosphokinase diabetes ejection_fraction
## 1   75.000       0                      582        0                20
## 2   55.000       0                     7861        0                38
## 3   65.000       0                      146        0                20
## 4   50.000       1                      111        0                20
## 5   65.000       1                      160        1                20
## 6   90.000       1                       47        0                40
## 7   75.000       1                      246        0                15
## 8   60.000       1                      315        1                60
## 9   65.000       0                      157        0                65
## 10  80.000       1                      123        0                35
## 11  75.000       1                       81        0                38
## 12  62.000       0                      231        0                25
## 13  45.000       1                      981        0                30
## 14  50.000       1                      168        0                38
## 15  49.000       1                       80        0                30
## 16  82.000       1                      379        0                50
## 17  87.000       1                      149        0                38
## 18  45.000       0                      582        0                14
## 19  70.000       1                      125        0                25
## 20  48.000       1                      582        1                55
## 21  65.000       1                       52        0                25
## 22  65.000       1                      128        1                30
## 23  68.000       1                      220        0                35
## 24  53.000       0                       63        1                60
## 25  75.000       0                      582        1                30
## 26  80.000       0                      148        1                38
## 27  95.000       1                      112        0                40
## 28  70.000       0                      122        1                45
## 29  58.000       1                       60        0                38
## 30  82.000       0                       70        1                30
## 31  94.000       0                      582        1                 38
## 32  85.000       0                       23        0                 45
## 33  50.000       1                      249        1                 35
## 34  50.000       1                      159        1                 30
## 35  65.000       0                       94        1                 50
## 36  69.000       0                      582        1                 35
## 37  90.000       1                       60        1                 50
## 38  82.000       1                      855        1                 50
## 39  60.000       0                     2656        1                 30
## 40  60.000       0                      235        1                 38
## 41  70.000       0                      582        0                 20
## 42  50.000       0                      124        1                 30
## 43  70.000       0                      571        1                 45
## 44  72.000       0                      127        1                 50
## 45  60.000       1                      588        1                 60
## 46  50.000       0                      582        1                 38
## 47  51.000       0                     1380        0                 25
## 48  60.000       0                      582        1                 38
## 49  80.000       1                      553        0                 20
## 50  57.000       1                      129        0                 30
## 51  68.000       1                      577        0                 25
## 52  53.000       1                       91        0                 20
## 53  60.000       0                     3964        1                 62
## 54  70.000       1                       69        1                 50
## 55  60.000       1                      260        1                 38
## 56  95.000       1                      371        0                 30
## 57  70.000       1                       75        0                 35
## 58  60.000       1                      607        0                 40
## 59  49.000       0                      789        0                 20
## 60  72.000       0                      364        1                 20
## 61  45.000       0                     7702        1                 25
## 62  50.000       0                      318        0                 40
## 63  55.000       0                      109        0                 35
## 64  45.000       0                      582        0                 35
## 65  45.000       0                      582        0                 80
## 66  60.000       0                       68        0                 20
## 67  42.000       1                      250        1                 15
## 68  72.000       1                      110        0                 25
## 69  70.000       0                      161        0                 25
## 70  65.000       0                      113        1                 25
## 71  41.000       0                      148        0                 40
## 72  58.000       0                      582        1                 35
## 73  85.000       0                     5882        0                 35
## 74  65.000       0                      224        1                 50
## 75  69.000       0                      582        0                 20
## 76  60.000       1                       47        0                 20
## 77  70.000       0                       92        0                 60
## 78  42.000       0                      102        1                 40
## 79  75.000       1                      203        1                 38
## 80  55.000       0                      336        0                 45
## 81  70.000       0                       69        0                 40
## 82  67.000       0                      582        0                 50
## 83  60.000       1                       76        1                 25
## 84  79.000       1                       55        0                 50
## 85  59.000       1                      280        1                25
## 86  51.000       0                       78        0                50
## 87  55.000       0                       47        0                35
## 88  65.000       1                       68        1                60
## 89  44.000       0                       84        1                40
## 90  57.000       1                      115        0                25
## 91  70.000       0                       66        1                45
## 92  60.000       0                      897        1                45
## 93  42.000       0                      582        0                60
## 94  60.000       1                      154        0                25
## 95  58.000       0                      144        1                38
## 96  58.000       1                      133        0                60
## 97  63.000       1                      514        1                25
## 98  70.000       1                       59        0                60
## 99  60.000       1                      156        1                25
## 100 63.000       1                       61        1                40
## 101 65.000       1                      305        0                25
## 102 75.000       0                      582        0                45
## 103 80.000       0                      898        0                25
## 104 42.000       0                     5209        0                30
## 105 60.000       0                       53        0                50
## 106 72.000       1                      328        0                30
## 107 55.000       0                      748        0                45
## 108 45.000       1                     1876        1                35
## 109 63.000       0                      936        0                38
## 110 45.000       0                      292        1                35
## 111 85.000       0                      129        0                60
## 112 55.000       0                       60        0                35
## 113 50.000       0                      369        1                25
## 114 70.000       1                      143        0                60
## 115 60.000       1                      754        1                40
## 116 58.000       1                      400        0                40
## 117 60.000       1                       96        1                60
## 118 85.000       1                      102        0                60
## 119 65.000       1                      113        1                60
## 120 86.000       0                      582        0                38
## 121 60.000       1                      737        0                60
## 122 66.000       1                       68        1                38
## 123 60.000       0                       96        1                38
## 124 60.000       1                      582        0                30
## 125 60.000       0                      582        0                40
## 126 43.000       1                      358        0                50
## 127 46.000       0                      168        1                17
## 128 58.000       1                      200        1                60
## 129 61.000       0                      248        0                30
## 130 53.000       1                      270        1                35
## 131 53.000       1                     1808        0                60
## 132 60.000       1                     1082        1                45
## 133 46.000       0                      719        0                40
## 134 63.000       0                      193        0                60
## 135 81.000       0                     4540        0                35
## 136 75.000       0                      582        0                40
## 137 65.000       1                       59        1                60
## 138 68.000       1                      646        0                25
## 139 62.000       0                      281         1                35
## 140 50.000       0                     1548         0                30
## 141 80.000       0                      805         0                38
## 142 46.000       1                      291         0                35
## 143 50.000       0                      482         1                30
## 144 61.000       1                       84         0                40
## 145 72.000       1                      943         0                25
## 146 50.000       0                      185         0                30
## 147 52.000       0                      132         0                30
## 148 64.000       0                     1610         0                60
## 149 75.000       1                      582         0                30
## 150 60.000       0                     2261         0                35
## 151 72.000       0                      233         0                45
## 152 62.000       0                       30         1                60
## 153 50.000       0                      115         0                45
## 154 50.000       0                     1846         1                35
## 155 65.000       1                      335         0                35
## 156 60.000       1                      231         1                25
## 157 52.000       1                       58         0                35
## 158 50.000       0                      250         0                25
## 159 85.000       1                      910         0                50
## 160 59.000       1                      129         0                45
## 161 66.000       1                       72         0                40
## 162 45.000       1                      130         0                35
## 163 63.000       1                      582         0                40
## 164 50.000       1                     2334         1                35
## 165 45.000       0                     2442         1                30
## 166 80.000       0                      776         1                38
## 167 53.000       0                      196         0                60
## 168 59.000       0                       66         1                20
## 169 65.000       0                      582         1                40
## 170 70.000       0                      835         0                35
## 171 51.000       1                      582         1                35
## 172 52.000       0                     3966         0                40
## 173 70.000       1                      171         0                60
## 174 50.000       1                      115         0                20
## 175 65.000       0                      198         1                35
## 176 60.000       1                       95         0                60
## 177 69.000       0                     1419         0                40
## 178 49.000       1                       69         0                50
## 179 63.000       1                      122         1                60
## 180 55.000       0                      835         0                40
## 181 40.000       0                      478         1                30
## 182 59.000       1                      176         1                25
## 183 65.000       0                      395         1                25
## 184 75.000       0                       99         0                38
## 185 58.000       1                      145         0                25
## 186 60.667       1                      104         1                30
## 187 50.000       0                      582         0                50
## 188 60.000       0                     1896         1                25
## 189 60.667       1                      151         1                40
## 190 40.000       0                      244         0                45
## 191 80.000       0                      582         1                35
## 192 64.000       1                       62         0                60
## 193 50.000       1                       121        1                40
## 194 73.000       1                       231        1                30
## 195 45.000       0                       582        0                20
## 196 77.000       1                       418        0                45
## 197 45.000       0                       582        1                38
## 198 65.000       0                       167        0                30
## 199 50.000       1                       582        1                20
## 200 60.000       0                      1211        1                35
## 201 63.000       1                      1767        0                45
## 202 45.000       0                       308        1                60
## 203 70.000       0                        97        0                60
## 204 60.000       0                        59        0                25
## 205 78.000       1                        64        0                40
## 206 50.000       1                       167        1                45
## 207 40.000       1                       101        0                40
## 208 85.000       0                       212        0                38
## 209 60.000       1                      2281        1                40
## 210 49.000       0                       972        1                35
## 211 70.000       0                       212        1                17
## 212 50.000       0                       582        0                62
## 213 78.000       0                       224        0                50
## 214 48.000       1                       131        1                30
## 215 65.000       1                       135        0                35
## 216 73.000       0                       582        0                35
## 217 70.000       0                      1202        0                50
## 218 54.000       1                       427        0                70
## 219 68.000       1                      1021        1                35
## 220 55.000       0                       582        1                35
## 221 73.000       0                       582        0                20
## 222 65.000       0                       118        0                50
## 223 42.000       1                        86        0                35
## 224 47.000       0                       582        0                25
## 225 58.000       0                       582        1                25
## 226 75.000       0                       675        1                60
## 227 58.000       1                        57        0                25
## 228 55.000       1                      2794        0                35
## 229 65.000       0                        56        0                25
## 230 72.000       0                       211        0                25
## 231 60.000       0                       166        0                30
## 232 70.000       0                        93        0                35
## 233 40.000       1                       129        0                35
## 234 53.000       1                       707        0                38
## 235 53.000       1                       582        0                45
## 236 77.000       1                       109        0                50
## 237 75.000       0                       119        0                50
## 238 70.000       0                       232        0                30
## 239 65.000       1                       720        1                40
## 240 55.000       1                       180        0                45
## 241 70.000       0                        81        1                35
## 242 65.000       0                       582        1                30
## 243 40.000       0                        90        0                35
## 244 73.000       1                      1185        0                40
## 245 54.000       0                       582        1                38
## 246 61.000       1                        80        1                38
## 247 55.000       0                     2017        0                25
## 248 64.000       0                      143        0                25
## 249 40.000       0                      624        0                35
## 250 53.000       0                      207        1                40
## 251 50.000       0                     2522        0                30
## 252 55.000       0                      572        1                35
## 253 50.000       0                      245        0                45
## 254 70.000       0                       88        1                35
## 255 53.000       1                      446        0                60
## 256 52.000       1                      191        1                30
## 257 65.000       0                      326        0                38
## 258 58.000       0                      132        1                38
## 259 45.000       1                       66        1                25
## 260 53.000       0                       56        0                50
## 261 55.000       0                       66        0                40
## 262 62.000       1                      655        0                40
## 263 65.000       1                      258        1                25
## 264 68.000       1                      157        1                60
## 265 61.000       0                      582        1                38
## 266 50.000       1                      298        0                35
## 267 55.000       0                     1199        0                20
## 268 56.000       1                      135        1                38
## 269 45.000       0                      582        1                38
## 270 40.000       0                      582        1                35
## 271 44.000       0                      582        1                30
## 272 51.000       0                      582        1                40
## 273 67.000       0                      213        0                38
## 274 42.000       0                       64        0                40
## 275 60.000       1                      257        1                30
## 276 45.000       0                      582        0                38
## 277 70.000       0                      618        0                35
## 278 70.000       0                      582        1                38
## 279 50.000       1                     1051        1                30
## 280 55.000       0                       84        1                38
## 281 70.000       0                     2695        1                40
## 282 70.000       0                      582        0                40
## 283 42.000       0                       64        0                30
## 284 65.000       0                     1688        0                38
## 285 50.000       1                       54        0                40
## 286 55.000       1                      170        1                40
## 287 60.000       0                      253        0                35
## 288 45.000       0                      582        1                55
## 289 65.000       0                      892        1                35
## 290 90.000       1                      337        0                38
## 291 45.000       0                      615        1                55
## 292 60.000       0                      320        0                35
## 293 52.000       0                      190        1                38
## 294 63.000       1                      103        1                35
## 295 62.000       0                       61        1                38
## 296 55.000       0                     1820        0                38
## 297 45.000       0                     2060        1                60
## 298 45.000       0                     2413        0                38
## 299 50.000       0                      196        0                45
##     high_blood_pressure platelets serum_creatinine serum_sodium sex smoking
## 1                     1    265000             1.90          130   1       0
## 2                     0    263358             1.10          136   1       0
## 3                     0    162000             1.30          129   1       1
## 4                     0    210000             1.90          137   1       0
## 5                     0    327000             2.70          116   0       0
## 6                     1    204000             2.10          132   1       1
## 7                     0    127000             1.20          137   1       0
## 8                     0    454000             1.10          131   1       1
## 9                     0    263358             1.50          138   0       0
## 10                    1    388000             9.40          133   1       1
## 11                    1    368000             4.00          131   1       1
## 12                    1    253000             0.90          140   1       1
## 13                    0    136000             1.10          137   1       0
## 14                    1    276000             1.10          137   1       0
## 15                    1    427000             1.00          138   0       0
## 16                    0     47000             1.30          136   1       0
## 17                    0    262000             0.90          140   1       0
## 18                    0    166000             0.80          127   1       0
## 19                    1    237000             1.00          140   0       0
## 20                    0     87000             1.90          121   0       0
## 21                    1    276000             1.30          137   0       0
## 22                    1    297000             1.60          136   0       0
## 23                    1    289000             0.90          140   1       1
## 24                    0    368000             0.80          135   1       0
## 25                    1    263358             1.83          134   0       0
## 26                    0    149000             1.90          144   1       1
## 27                    1    196000             1.00          138   0       0
## 28                    1    284000             1.30          136   1       1
## 29                    0    153000             5.80          134   1       0
## 30                    0    200000             1.20          132   1       1
## 31                    1    263358             1.83          134   1       0
## 32                    0    360000             3.00          132   1       0
## 33                    1    319000             1.00          128   0       0
## 34                    0    302000             1.20          138   0       0
## 35                    1    188000             1.00          140   1       0
## 36                    0    228000             3.50          134   1       0
## 37                    0    226000             1.00          134   1       0
## 38                    1    321000             1.00          145   0       0
## 39                    0    305000             2.30          137   1       0
## 40                    0    329000             3.00          142   0       0
## 41                    1    263358             1.83          134   1       1
## 42                    1    153000             1.20          136   0       1
## 43                    1    185000             1.20          139   1       1
## 44                    1    218000             1.00          134   1       0
## 45                    0    194000             1.10          142   0       0
## 46                    0    310000             1.90          135   1       1
## 47                    1    271000             0.90          130   1       0
## 48                    1    451000             0.60          138   1       1
## 49                    1    140000             4.40          133   1       0
## 50                    0    395000             1.00          140   0       0
## 51                    1    166000             1.00          138   1       0
## 52                    1    418000             1.40          139   0       0
## 53                    0    263358             6.80          146   0       0
## 54                    1    351000             1.00          134   0       0
## 55                    0    255000             2.20          132   0       1
## 56                    0    461000             2.00          132   1       0
## 57                    0    223000             2.70          138   1       1
## 58                    0    216000             0.60          138   1       1
## 59                    1    319000             1.10          136   1       1
## 60                    1    254000             1.30          136   1       1
## 61                    1    390000             1.00          139   1       0
## 62                    1    216000             2.30          131   0       0
## 63                    0    254000             1.10          139   1       1
## 64                    0    385000             1.00          145   1       0
## 65                    0    263358             1.18          137   0       0
## 66                    0    119000             2.90          127   1       1
## 67                    0    213000             1.30          136   0       0
## 68                    0    274000             1.00          140   1       1
## 69                    0    244000             1.20          142   0       0
## 70                    0    497000             1.83          135   1       0
## 71                    0    374000             0.80          140   1       1
## 72                    0    122000             0.90          139   1       1
## 73                    0    243000             1.00          132   1       1
## 74                    0    149000             1.30          137   1       1
## 75                    0    266000             1.20          134   1       1
## 76                    0    204000             0.70          139   1       1
## 77                    1    317000             0.80          140   0       1
## 78                    0    237000             1.20          140   1       0
## 79                    1    283000             0.60          131   1       1
## 80                    1    324000             0.90          140   0       0
## 81                    0    293000             1.70          136   0       0
## 82                    0    263358             1.18          137   1       1
## 83                    0    196000             2.50          132   0       0
## 84                    1    172000             1.80          133   1       0
## 85                    1    302000             1.00          141   0       0
## 86                    0    406000             0.70          140   1       0
## 87                    1    173000             1.10          137   1       0
## 88                    1    304000             0.80          140   1       0
## 89                    1    235000             0.70          139   1       0
## 90                    1    181000             1.10          144   1       0
## 91                    0    249000             0.80          136   1       1
## 92                    0    297000             1.00          133   1       0
## 93                    0    263358             1.18          137   0       0
## 94                    0    210000             1.70          135   1       0
## 95                    1    327000             0.70          142   0       0
## 96                    1    219000             1.00          141   1       0
## 97                    1    254000             1.30          134   1       0
## 98                    0    255000             1.10          136   0       0
## 99                    1    318000             1.20          137   0       0
## 100                   0    221000             1.10          140   0       0
## 101                   0    298000             1.10          141   1       0
## 102                   1    263358             1.18          137   1       0
## 103                   0    149000             1.10          144   1       1
## 104                   0    226000             1.00          140   1       1
## 105                   1    286000             2.30          143   0       0
## 106                   1    621000             1.70          138   0       1
## 107                   0    263000             1.30          137   1       0
## 108                   0    226000             0.90          138   1       0
## 109                   0    304000              1.10          133   1       1
## 110                   0    850000              1.30          142   1       1
## 111                   0    306000              1.20          132   1       1
## 112                   0    228000              1.20          135   1       1
## 113                   0    252000              1.60          136   1       0
## 114                   0    351000              1.30          137   0       0
## 115                   1    328000              1.20          126   1       0
## 116                   0    164000              1.00          139   0       0
## 117                   1    271000              0.70          136   0       0
## 118                   0    507000              3.20          138   0       0
## 119                   1    203000              0.90          140   0       0
## 120                   0    263358              1.83          134   0       0
## 121                   1    210000              1.50          135   1       1
## 122                   1    162000              1.00          136   0       0
## 123                   0    228000              0.75          140   0       0
## 124                   1    127000              0.90          145   0       0
## 125                   0    217000              3.70          134   1       0
## 126                   0    237000              1.30          135   0       0
## 127                   1    271000              2.10          124   0       0
## 128                   0    300000              0.80          137   0       0
## 129                   1    267000              0.70          136   1       1
## 130                   0    227000              3.40          145   1       0
## 131                   1    249000              0.70          138   1       1
## 132                   0    250000              6.10          131   1       0
## 133                   1    263358              1.18          137   0       0
## 134                   1    295000              1.30          145   1       1
## 135                   0    231000              1.18          137   1       1
## 136                   0    263358              1.18          137   1       0
## 137                   0    172000              0.90          137   0       0
## 138                   0    305000              2.10          130   1       0
## 139                   0    221000              1.00          136   0       0
## 140                   1    211000              0.80          138   1       0
## 141                   0    263358              1.10          134   1       0
## 142                   0    348000              0.90          140   0       0
## 143                   0    329000              0.90          132   0       0
## 144                   1    229000              0.90          141   0       0
## 145                   1    338000              1.70          139   1       1
## 146                   0    266000              0.70          141   1       1
## 147                   0    218000              0.70          136   1       1
## 148                   0    242000              1.00          137   1       0
## 149                   0    225000              1.83          134   1       0
## 150                   1    228000              0.90          136   1       0
## 151                   1    235000              2.50          135   0       0
## 152                   1    244000              0.90          139   1       0
## 153                   1    184000              0.90          134   1       1
## 154                   0    263358              1.18          137   1       1
## 155                   1    235000              0.80          136   0       0
## 156                   0    194000              1.70          140   1       0
## 157                   0    277000              1.40          136   0       0
## 158                   0    262000              1.00          136   1       1
## 159                   0    235000              1.30          134   1       0
## 160                   1    362000              1.10          139   1       1
## 161                   1    242000              1.20          134   1       0
## 162                   0    174000              0.80          139   1       1
## 163                   0     448000            0.90          137    1       1
## 164                   0      75000            0.90          142    0       0
## 165                   0     334000            1.10          139    1       0
## 166                   1     192000            1.30          135    0       0
## 167                   0     220000            0.70          133    1       1
## 168                   0      70000            2.40          134    1       0
## 169                   0     270000            1.00          138    0       0
## 170                   1     305000            0.80          133    0       0
## 171                   0     263358            1.50          136    1       1
## 172                   0     325000            0.90          140    1       1
## 173                   1     176000            1.10          145    1       1
## 174                   0     189000            0.80          139    1       0
## 175                   1     281000            0.90          137    1       1
## 176                   0     337000            1.00          138    1       1
## 177                   0     105000            1.00          135    1       1
## 178                   0     132000            1.00          140    0       0
## 179                   0     267000            1.20          145    1       0
## 180                   0     279000            0.70          140    1       1
## 181                   0     303000            0.90          136    1       0
## 182                   0     221000            1.00          136    1       1
## 183                   0     265000            1.20          136    1       1
## 184                   1     224000            2.50          134    1       0
## 185                   0     219000            1.20          137    1       1
## 186                   0     389000            1.50          136    1       0
## 187                   0     153000            0.60          134    0       0
## 188                   0     365000            2.10          144    0       0
## 189                   1     201000            1.00          136    0       0
## 190                   1     275000            0.90          140    0       0
## 191                   0     350000            2.10          134    1       0
## 192                   0     309000            1.50          135    0       0
## 193                   0     260000            0.70          130    1       0
## 194                   0     160000            1.18          142    1       1
## 195                   1     126000            1.60          135    1       0
## 196                   0     223000            1.80          145    1       0
## 197                   1     263358            1.18          137    0       0
## 198                   0     259000            0.80          138    0       0
## 199                   1     279000            1.00          134    0       0
## 200                   0     263358            1.80          113    1       1
## 201                   0      73000            0.70          137    1       0
## 202                   1     377000            1.00          136    1       0
## 203                   1     220000            0.90          138    1       0
## 204                   1     212000            3.50          136    1       1
## 205                   0     277000            0.70          137    1       1
## 206                   0     362000            1.00          136    0       0
## 207                   0     226000            0.80          141    0       0
## 208                   0     186000            0.90          136    1       0
## 209                   0     283000            1.00          141    0       0
## 210                   1     268000            0.80          130    0       0
## 211                   1     389000            1.00          136    1       1
## 212                   1     147000            0.80          140    1       1
## 213                   0     481000            1.40          138    1       1
## 214                   1     244000            1.60          130    0       0
## 215                   1     290000            0.80          134    1       0
## 216                   1     203000            1.30          134    1       0
## 217                   1    358000             0.90          141   0       0
## 218                   1    151000             9.00          137   0       0
## 219                   0    271000             1.10          134   1       0
## 220                   1    371000             0.70          140   0       0
## 221                   0    263358             1.83          134   1       0
## 222                   0    194000             1.10          145   1       1
## 223                   0    365000             1.10          139   1       1
## 224                   0    130000             0.80          134   1       0
## 225                   0    504000             1.00          138   1       0
## 226                   0    265000             1.40          125   0       0
## 227                   0    189000             1.30          132   1       1
## 228                   1    141000             1.00          140   1       0
## 229                   0    237000             5.00          130   0       0
## 230                   0    274000             1.20          134   0       0
## 231                   0     62000             1.70          127   0       0
## 232                   0    185000             1.10          134   1       1
## 233                   0    255000             0.90          137   1       0
## 234                   0    330000             1.40          137   1       1
## 235                   0    305000             1.10          137   1       1
## 236                   1    406000             1.10          137   1       0
## 237                   1    248000             1.10          148   1       0
## 238                   0    173000             1.20          132   1       0
## 239                   0    257000             1.00          136   0       0
## 240                   0    263358             1.18          137   1       1
## 241                   1    533000             1.30          139   0       0
## 242                   0    249000             1.30          136   1       1
## 243                   0    255000             1.10          136   1       1
## 244                   1    220000             0.90          141   0       0
## 245                   0    264000             1.80          134   1       0
## 246                   0    282000             1.40          137   1       0
## 247                   0    314000             1.10          138   1       0
## 248                   0    246000             2.40          135   1       0
## 249                   0    301000             1.00          142   1       1
## 250                   0    223000             1.20          130   0       0
## 251                   1    404000             0.50          139   0       0
## 252                   0    231000             0.80          143   0       0
## 253                   1    274000             1.00          133   1       0
## 254                   1    236000             1.20          132   0       0
## 255                   1    263358             1.00          139   1       0
## 256                   1    334000             1.00          142   1       1
## 257                   0    294000             1.70          139   0       0
## 258                   1    253000             1.00          139   1       0
## 259                   0    233000             0.80          135   1       0
## 260                   0    308000             0.70          135   1       1
## 261                   0    203000             1.00          138   1       0
## 262                   0    283000             0.70          133   0       0
## 263                   0    198000             1.40          129   1       0
## 264                   0    208000             1.00          140   0       0
## 265                   0    147000             1.20          141   1       0
## 266                   0    362000             0.90          140   1       1
## 267                   0    263358             1.83          134   1       1
## 268                   0    133000             1.70          140   1       0
## 269                   0    302000             0.90          140   0       0
## 270                   0    222000             1.00          132   1       0
## 271                   1    263358             1.60          130   1       1
## 272                   0    221000             0.90          134   0       0
## 273                   0    215000             1.20          133   0       0
## 274                   0    189000             0.70          140   1       0
## 275                   0    150000             1.00          137   1       1
## 276                   1    422000             0.80          137   0       0
## 277                   0    327000             1.10          142   0       0
## 278                   0     25100             1.10          140   1       0
## 279                   0    232000             0.70          136   0       0
## 280                   0    451000             1.30          136   0       0
## 281                   0    241000             1.00          137   1       0
## 282                   0     51000             2.70          136   1       1
## 283                   0    215000             3.80          128   1       1
## 284                   0    263358             1.10          138   1       1
## 285                   0    279000             0.80          141   1       0
## 286                   0    336000             1.20          135   1       0
## 287                   0    279000             1.70          140   1       0
## 288                   0    543000             1.00          132   0       0
## 289                   0    263358             1.10          142   0       0
## 290                   0    390000             0.90          144   0       0
## 291                   0    222000             0.80          141   0       0
## 292                   0    133000             1.40          139   1       0
## 293                   0    382000             1.00          140   1       1
## 294                   0    179000             0.90          136   1       1
## 295                   1    155000             1.10          143   1       1
## 296                   0    270000             1.20          139   0       0
## 297                   0    742000             0.80          138   0       0
## 298                   0    140000             1.40          140   1       1
## 299                   0    395000             1.60          136   1       1
##     time DEATH_EVENT gender
## 1      4           1   male
## 2      6           1   male
## 3      7           1   male
## 4      7           1   male
## 5      8           1 female
## 6      8           1   male
## 7     10           1   male
## 8     10           1   male
## 9     10           1 female
## 10    10           1   male
## 11    10           1   male
## 12    10           1   male
## 13    11           1   male
## 14    11           1   male
## 15    12           0 female
## 16    13           1   male
## 17    14           1   male
## 18    14           1   male
## 19    15           1 female
## 20    15           1 female
## 21    16           0 female
## 22    20           1 female
## 23    20           1   male
## 24    22           0   male
## 25    23           1 female
## 26    23           1   male
## 27    24           1 female
## 28    26           1   male
## 29    26           1   male
## 30    26           1   male
## 31    27           1   male
## 32    28           1   male
## 33    28           1 female
## 34    29           0 female
## 35    29           1   male
## 36    30           1   male
## 37    30           1   male
## 38    30           1 female
## 39    30           0   male
## 40    30           1 female
## 41    31           1   male
## 42    32           1 female
## 43    33           1   male
## 44    33           0   male
## 45    33           1 female
## 46    35           1   male
## 47    38           1   male
## 48    40           1   male
## 49    41           1   male
## 50    42           1 female
## 51    43           1   male
## 52    43           1 female
## 53    43           1 female
## 54    44           1 female
## 55    45           1 female
## 56    50           1   male
## 57    54           0   male
## 58    54           0   male
## 59    55           1   male
## 60    59           1   male
## 61    60           1   male
## 62    60           1 female
## 63    60           0   male
## 64    61           1   male
## 65    63           0 female
## 66    64           1   male
## 67    65           1 female
## 68    65           1   male
## 69    66           1 female
## 70    67           1   male
## 71    68           0   male
## 72    71           0   male
## 73    72           1   male
## 74    72           0   male
## 75    73           1   male
## 76    73           1   male
## 77    74           0 female
## 78    74           0   male
## 79    74           0   male
## 80    74           0 female
## 81    75           0 female
## 82    76           0   male
## 83    77           1 female
## 84    78           0   male
## 85    78           1 female
## 86    79           0   male
## 87    79           0   male
## 88    79           0   male
## 89    79           0   male
## 90    79           0   male
## 91    80           0   male
## 92    80           0   male
## 93    82           0 female
## 94    82           1   male
## 95    83           0 female
## 96    83           0   male
## 97    83           0   male
## 98    85           0 female
## 99    85           0 female
## 100   86           0 female
## 101   87           0   male
## 102   87           0   male
## 103   87           0   male
## 104   87           0   male
## 105   87           0 female
## 106   88           1 female
## 107   88           0   male
## 108   88           0   male
## 109   88           0   male
## 110   88           0   male
## 111   90           1   male
## 112   90           0   male
## 113   90           0   male
## 114   90           1 female
## 115   91           0   male
## 116   91           0 female
## 117   94           0 female
## 118   94           0 female
## 119   94           0 female
## 120   95           1 female
## 121   95           0   male
## 122   95           0 female
## 123   95           0 female
## 124   95           0 female
## 125   96           1   male
## 126   97           0 female
## 127  100           1 female
## 128  104           0 female
## 129  104           0   male
## 130  105           0   male
## 131  106           0   male
## 132  107           0   male
## 133  107           0 female
## 134  107           0   male
## 135  107           0   male
## 136  107           0   male
## 137  107           0 female
## 138  108           0   male
## 139  108           0 female
## 140  108           0   male
## 141  109           1   male
## 142  109           0 female
## 143  109           0 female
## 144  110           0 female
## 145  111           1   male
## 146  112           0   male
## 147  112           0   male
## 148  113           0   male
## 149  113           1   male
## 150  115           0   male
## 151  115           1 female
## 152  117           0   male
## 153  118           0   male
## 154  119           0   male
## 155  120           0 female
## 156  120           0   male
## 157  120           0 female
## 158  120           0   male
## 159  121           0   male
## 160  121           0   male
## 161  121           0   male
## 162  121           0   male
## 163  123           0   male
## 164  126           1 female
## 165  129           1   male
## 166  130           1 female
## 167  134           0   male
## 168  135           1   male
## 169  140           0 female
## 170  145           0 female
## 171  145           0   male
## 172  146           0   male
## 173  146           0   male
## 174  146           0   male
## 175  146           0   male
## 176  146           0   male
## 177  147           0   male
## 178  147           0 female
## 179  147           0   male
## 180  147           0   male
## 181  148           0   male
## 182  150           1   male
## 183  154           1   male
## 184  162           1   male
## 185  170           1   male
## 186  171           1   male
## 187  172           1 female
## 188  172           1 female
## 189  172           0 female
## 190  174           0 female
## 191  174           0   male
## 192  174           0 female
## 193  175           0   male
## 194  180           0   male
## 195  180           1   male
## 196  180           1   male
## 197  185           0 female
## 198  186           0 female
## 199  186           0 female
## 200  186           0   male
## 201  186           0   male
## 202  186           0   male
## 203  186           0   male
## 204  187           0   male
## 205  187           0   male
## 206  187           0 female
## 207  187           0 female
## 208  187           0   male
## 209  187           0 female
## 210  187           0 female
## 211  188           0   male
## 212  192           0   male
## 213  192           0   male
## 214  193           1 female
## 215  194           0   male
## 216  195           0   male
## 217  196           0 female
## 218  196           1 female
## 219  197           0   male
## 220  197           0 female
## 221  198           1   male
## 222  200           0   male
## 223  201           0   male
## 224  201           0   male
## 225  205           0   male
## 226  205           0 female
## 227  205           0   male
## 228  206           0   male
## 229  207           0 female
## 230  207           0 female
## 231  207           1 female
## 232  208           0   male
## 233  209           0   male
## 234  209           0   male
## 235  209           0   male
## 236  209           0   male
## 237  209           0   male
## 238  210           0   male
## 239  210           0 female
## 240  211           0   male
## 241  212           0 female
## 242  212           0   male
## 243  212           0   male
## 244  213           0 female
## 245  213           0   male
## 246  213           0   male
## 247  214           1   male
## 248  214           0   male
## 249  214           0   male
## 250  214           0 female
## 251  214           0 female
## 252  215           0 female
## 253  215           0   male
## 254  215           0 female
## 255  215           0   male
## 256  216           0   male
## 257  220           0 female
## 258  230           0   male
## 259  230           0   male
## 260  231           0   male
## 261  233           0   male
## 262  233           0 female
## 263  235           1   male
## 264  237           0 female
## 265  237           0   male
## 266  240           0   male
## 267  241           1   male
## 268  244           0   male
## 269  244           0 female
## 270  244           0   male
## 271  244           0   male
## 272  244           0 female
## 273  245           0 female
## 274  245           0   male
## 275  245           0   male
## 276  245           0 female
## 277  245           0 female
## 278  246           0   male
## 279  246           0 female
## 280  246           0 female
## 281  247           0   male
## 282  250           0   male
## 283  250           0   male
## 284  250           0   male
## 285  250           0   male
## 286  250           0   male
## 287  250           0   male
## 288  250           0 female
## 289  256           0 female
## 290  256           0 female
## 291  257           0 female
## 292  258           0   male
## 293  258           0   male
## 294  270           0   male
## 295  270           0   male
## 296  271           0 female
## 297  278           0 female
## 298  280           0   male
## 299  285           0   male
```

Following this, we can visualise that str, dim with new variable and mean for both genders doesn’t affect by mutate function.

```
str(myocard)
```

```
## ’data.frame’:    299 obs. of  14 variables:
##  $ age                     : num  75 55 65 50 65 90 75 60 65 80 ...
##  $ anaemia                 : int  0 0 0 1 1 1 1 1 0 1 ...
##  $ creatinine_phosphokinase: int  582 7861 146 111 160 47 246 315 157 123 ...
##  $ diabetes                : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ ejection_fraction       : int  20 38 20 20 20 40 15 60 65 35 ...
##  $ high_blood_pressure     : int  1 0 0 0 0 1 0 0 0 1 ...
##  $ platelets               : num  265000 263358 162000 210000 327000 ...
##  $ serum_creatinine        : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
##  $ serum_sodium            : int  130 136 129 137 116 132 137 131 138 133 ...
##  $ sex                     : int  1 1 1 1 0 1 1 1 0 1 ...
##  $ smoking                 : int  0 0 1 0 0 1 0 1 0 1 ...
##  $ time                    : int  4 6 7 7 8 8 10 10 10 10 ...
##  $ DEATH_EVENT             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ gender                  : chr  "male" "male" "male" "male" ...
```

```
dim(myocard)
```

```
## [1] 299  14
```

```
mean(myocard$gender == "female")
```

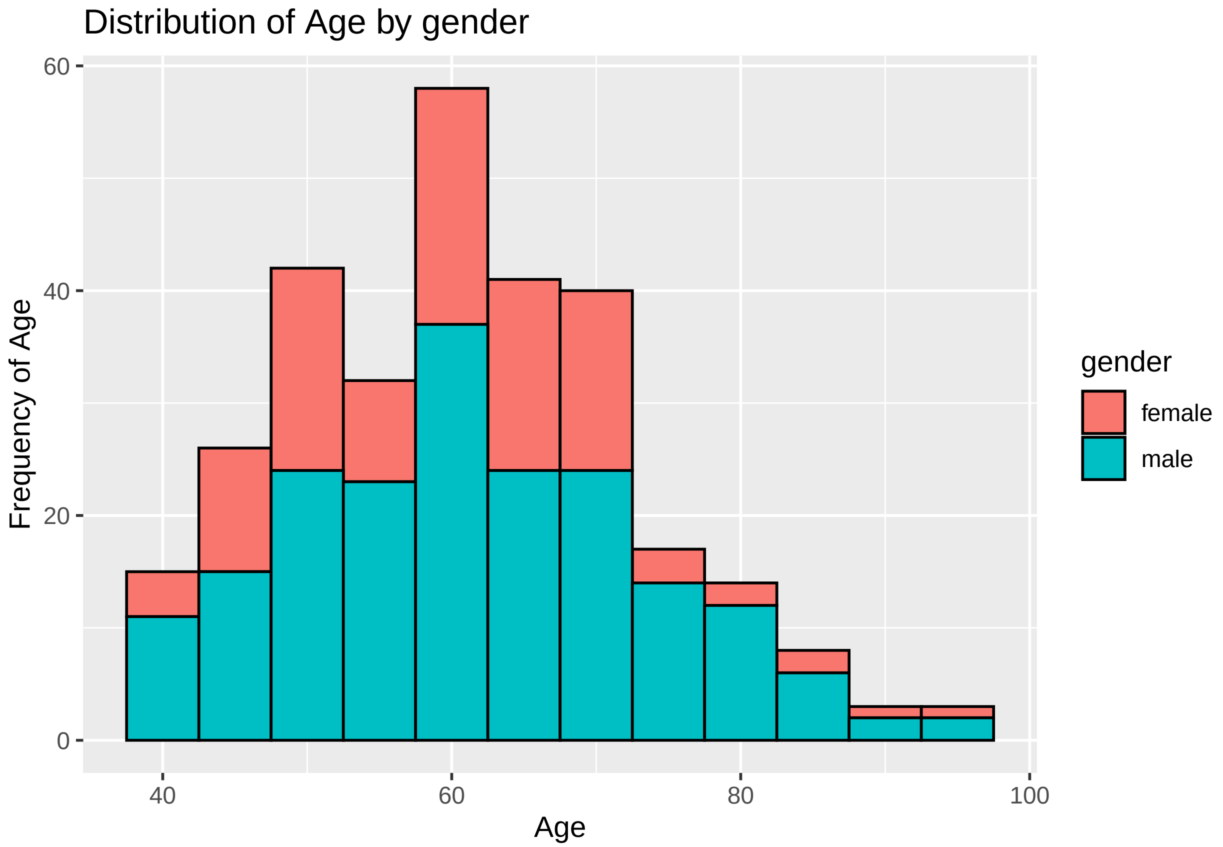
```
## [1] 0.3511706
```

```
mean(myocard$gender == "male")
```

```
## [1] 0.6488294
```

3.2.1 Distribution of Age

```
myocard %>% ggplot(aes(age, fill = gender)) + geom_histogram(bins = 30, binwidth = 5, color = "black")
```

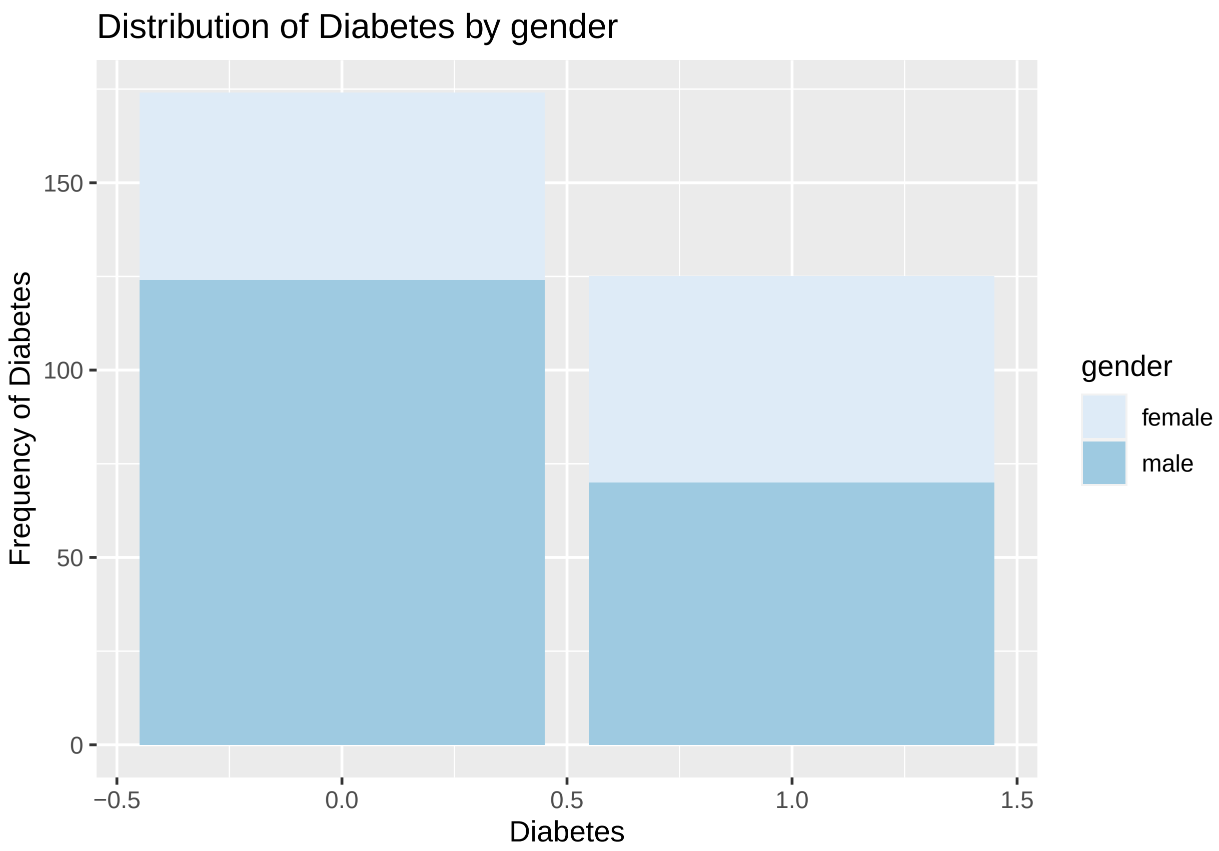


Age is one of most risk for developing predictions of diseases (Nicolli T, 2012). In our case, we can notice the distribution of age by gender. We see that the peak of age is around 60 and previous see before, and there is more man than a woman in our case.

3.2.2 Distribution of Diabetes

Diabetes is a common disease where your blood glucose is too high. That means your body doesn't produce enough insulin or can't produce any at all. The diagnostic of Diabetes can lead to several predictions of disease, such as MI. In our case, we are assuming that all patients who have Diabetes are 1 and without Diabetes is 0.

```
myocard %>% ggplot(aes(diabetes)) + geom_bar(aes(fill = gender)) + scale_fill_brewer(palette="Blues") +
        xlab("Diabetes") + ylab("Frequency of Diabetes") + ggtitle("Distribution of Diabetes by gender")
```



```
myocard %>% group_by(diabetes) %>% summarise(n = n()) %>% head()
```

```
## # A tibble: 2 x 2
##   diabetes     n
##      <int> <int>
## 1        0   174
## 2        1   125
```

We can visualise the number of patients who have Diabetes as around 125 and without Diabetes as 174.

3.2.3 Distribuition of Diabetes by gender

What's the proportion of Diabetes by gender?

```
myocard %>% group_by(gender) %>% summarise(diabetes = mean(diabetes == 0)) %>%
        filter(gender == "female")
```

```
## # A tibble: 1 x 2
##   gender diabetes
##   <chr>     <dbl>
## 1 female    0.476
```

```
myocard %>% group_by(gender) %>% summarise(diabetes = mean(diabetes == 1)) %>%
        filter(gender == "female")
```

```
## # A tibble: 1 x 2
##   gender diabetes
##   <chr>     <dbl>
## 1 female    0.524
```

```
myocard %>% group_by(gender) %>% summarise(diabetes = mean(diabetes == 0)) %>%
        filter(gender == "male")
```

```
## # A tibble: 1 x 2
##   gender diabetes
##   <chr>     <dbl>
## 1 male      0.639
```

```
myocard %>% group_by(gender) %>% summarise(diabetes = mean(diabetes == 1)) %>%
        filter(gender == "male")
```

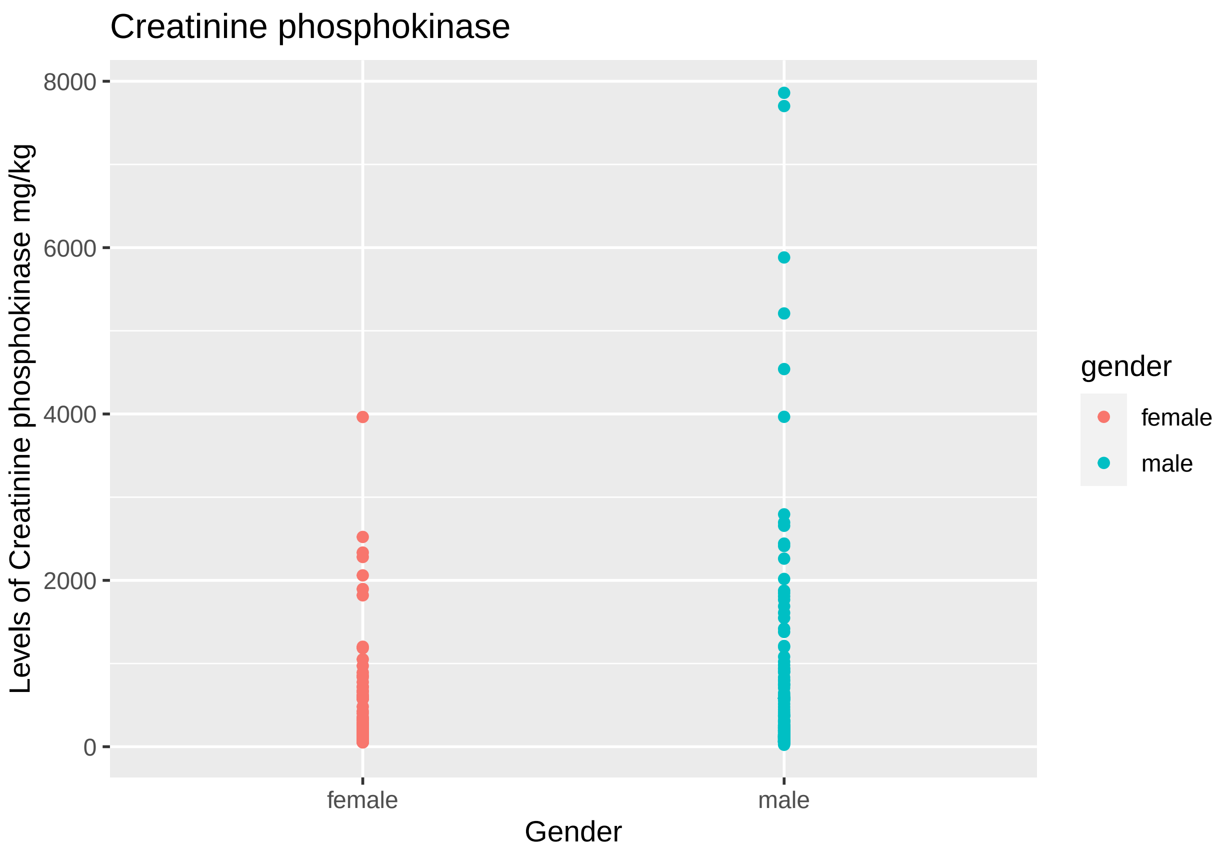
```
## # A tibble: 1 x 2
##   gender diabetes
##   <chr>     <dbl>
## 1 male      0.361
```

Filtering by gender, we can find more women with Diabetes (0.524) than men (0.361).

3.2.4 Levels of Creatinine phosphokinase

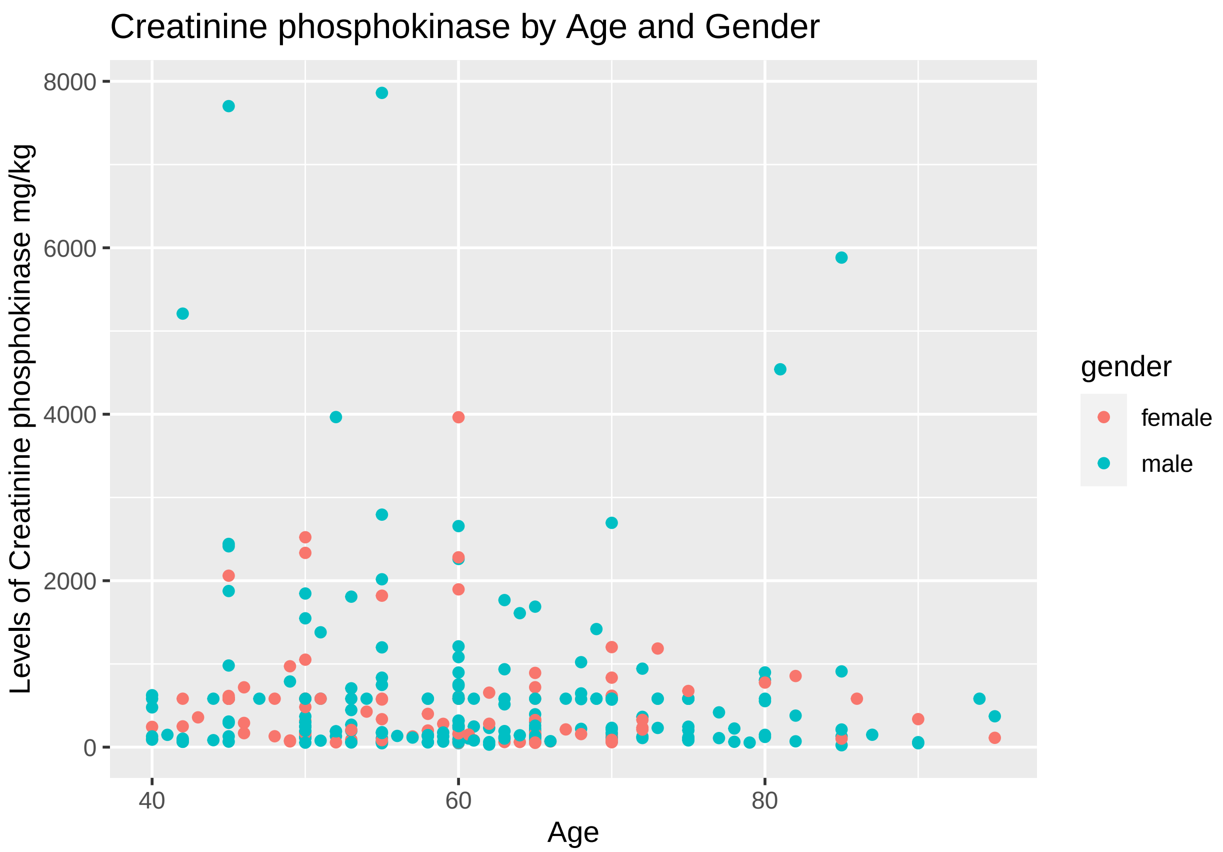
Creatinine phosphokinase, also know as Creatine Kinase (CK), is an enzyme found in our body. It is located mainly in the heart, brain, and skeletal muscle. When your body has damage, CK increase into your body, meaning something is not right. Also, CK levels are important MI markers. Elevated levels of CK have used to diagnose a case of MI (Patel R, 2021).

```
myocard %>% ggplot(aes(gender, creatinine_phosphokinase)) + geom_point(aes(color = gender)) +
        ylab("Levels of Creatinine phosphokinase mg/kg") + xlab("Gender") +
        ggtitle("Creatinine phosphokinase")
```



We can notice that's men have a little higher levels of CK rather than women.

```
myocard %>% ggplot(aes(age, creatinine_phosphokinase)) + geom_point(aes(color = gender)) +
        ylab("Levels of Creatinine phosphokinase mg/kg") + xlab("Age") +
        ggtitle("Creatinine phosphokinase by Age and Gender")
```



As sawn before, age can be risk factors for some disease. We can analyse the CK by age and gender. Following this, what're the mean levels between women and men?

```
myocard %>% group_by(gender) %>% summarise(creatinine_phosphokinase = mean(creatinine_phosphokinase))
```

```
## # A tibble: 2 x 2
##   gender creatinine_phosphokinase
##   <chr>                     <dbl>
## 1 female                     477.
## 2 male                       639.
```

```
myocard %>% group_by(gender) %>% summarise(creatinine_phosphokinase = sd(creatinine_phosphokinase))
```

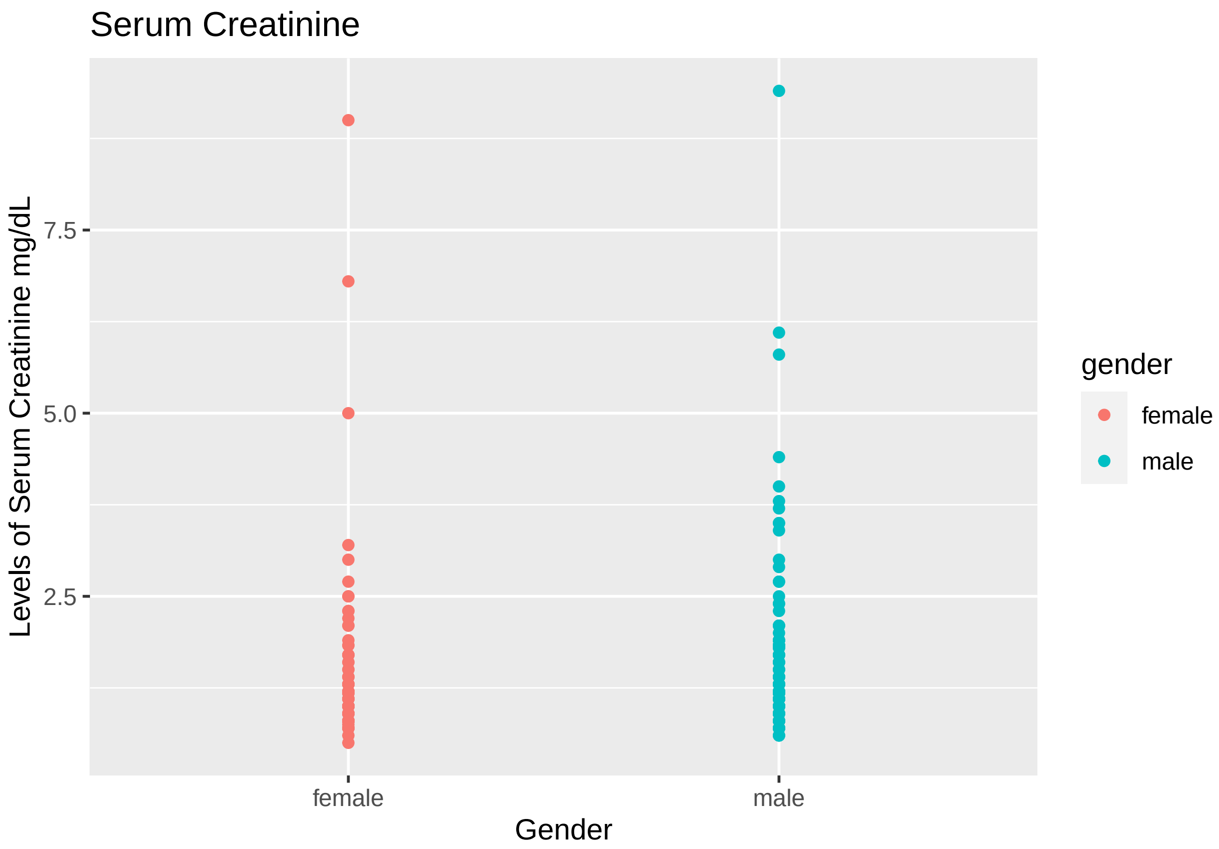
```
## # A tibble: 2 x 2
##   gender creatinine_phosphokinase
##   <chr>                     <dbl>
## 1 female                     611.
## 2 male                      1115.
```

According to the literature, the normals levels of CK is around 200 ul/L. We can see the man have mean and standard deviation higher than women. Showing to us the man has higher risk than women.

3.2.5 Serum Creatinine

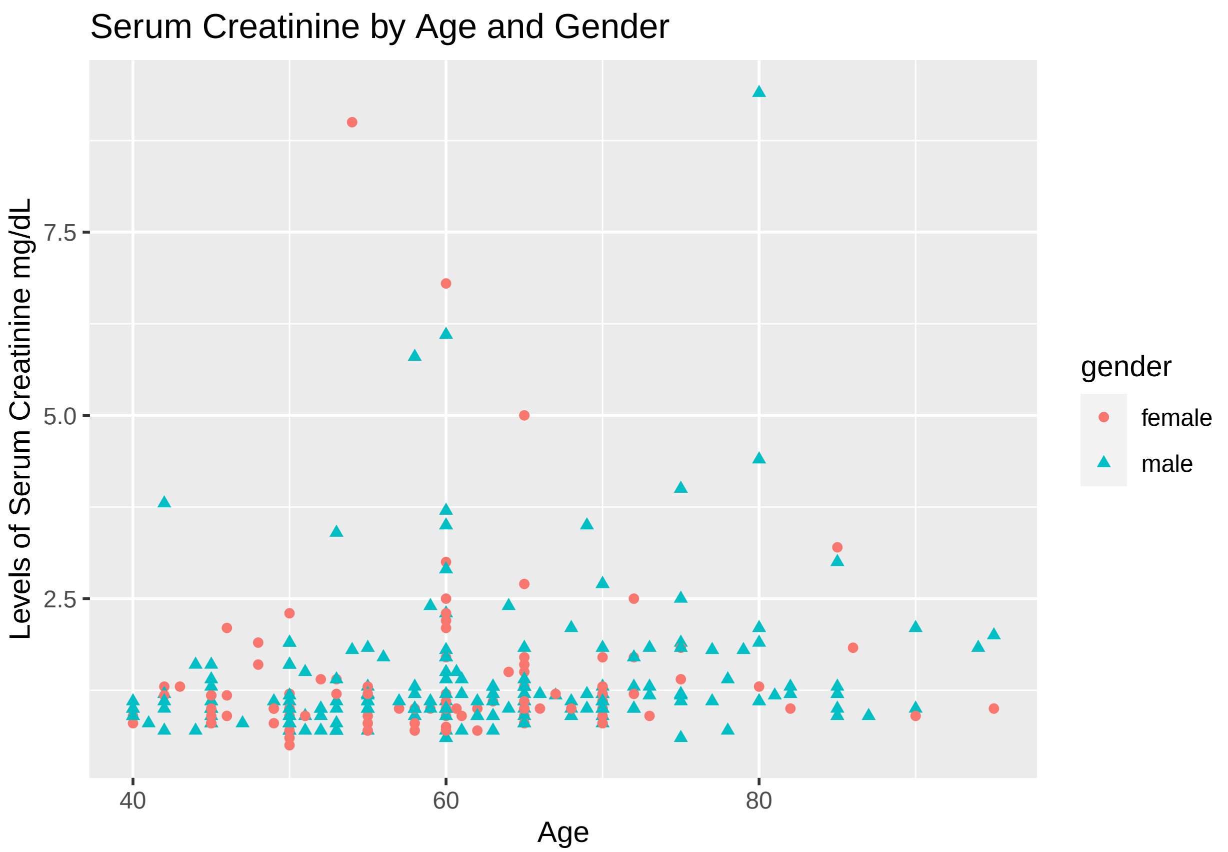
The creatinine levels measure how well your kidneys perform their work of filtering waste from your blood. That's mean if your kidney has an injury or damage, the levels of creatinine can arise.

```
myocard %>% ggplot(aes(gender, serum_creatinine)) + geom_point(aes(color = gender)) +
        ylab("Levels of Serum Creatinine mg/dL") + xlab("Gender") +
        ggtitle("Serum Creatinine")
```



The levels of Serum creatinine is slightly the same between gender.

```
myocard %>% ggplot(aes(age, serum_creatinine)) + geom_point(aes(color = gender, shape = gender)) +
        ylab("Levels of Serum Creatinine mg/dL") + xlab("Age") +
        ggtitle("Serum Creatinine by Age and Gender")
```



We can visualise the Serum creatine Age and gender. We can see there are no differences between gender. The levels of Serum creatinine between woman and men are quite the same. We can see this by the mean and standard deviation.

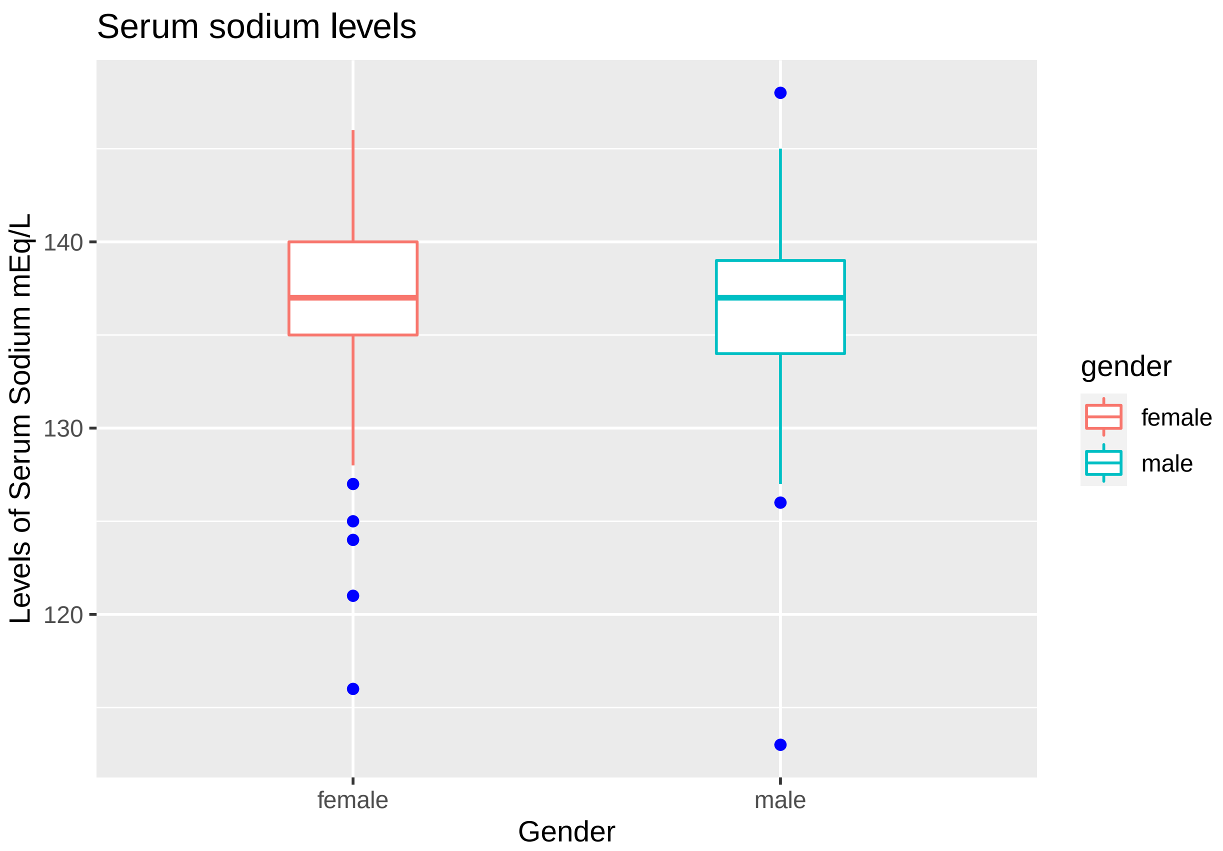
```
myocard %>% group_by(gender) %>% summarise(mean(serum_creatinine), sd(serum_creatinine))
```

```
## # A tibble: 2 x 3
##   gender ‘mean(serum_creatinine)‘ ‘sd(serum_creatinine)‘
##   <chr>                     <dbl>                   <dbl>
## 1 female                     1.38                   1.12
## 2 male                       1.40                   0.989
```

3.2.6 Serum Sodium Levels

Alterations in sodium levels is a risk for any disease. High levels of serum sodium can lead to high pressure, but lower levels can predict MI. This graph shows us the Sodium serum levels, and we can notice for both gender is around 137. Normals levels are about 135 to 145 mEq/L.

```
myocard %>% ggplot(aes(gender, serum_sodium)) +
        geom_boxplot(aes(color = gender), outlier.colour = "blue", width = 0.3) +
        ylab("Levels of Serum Sodium mEq/L") + xlab("Gender") +
        ggtitle("Serum sodium levels")
```



We can see, the levels of serum sodium in both gender in our case doesn't have difference. Also, that can observe by the mean and standard deviation.

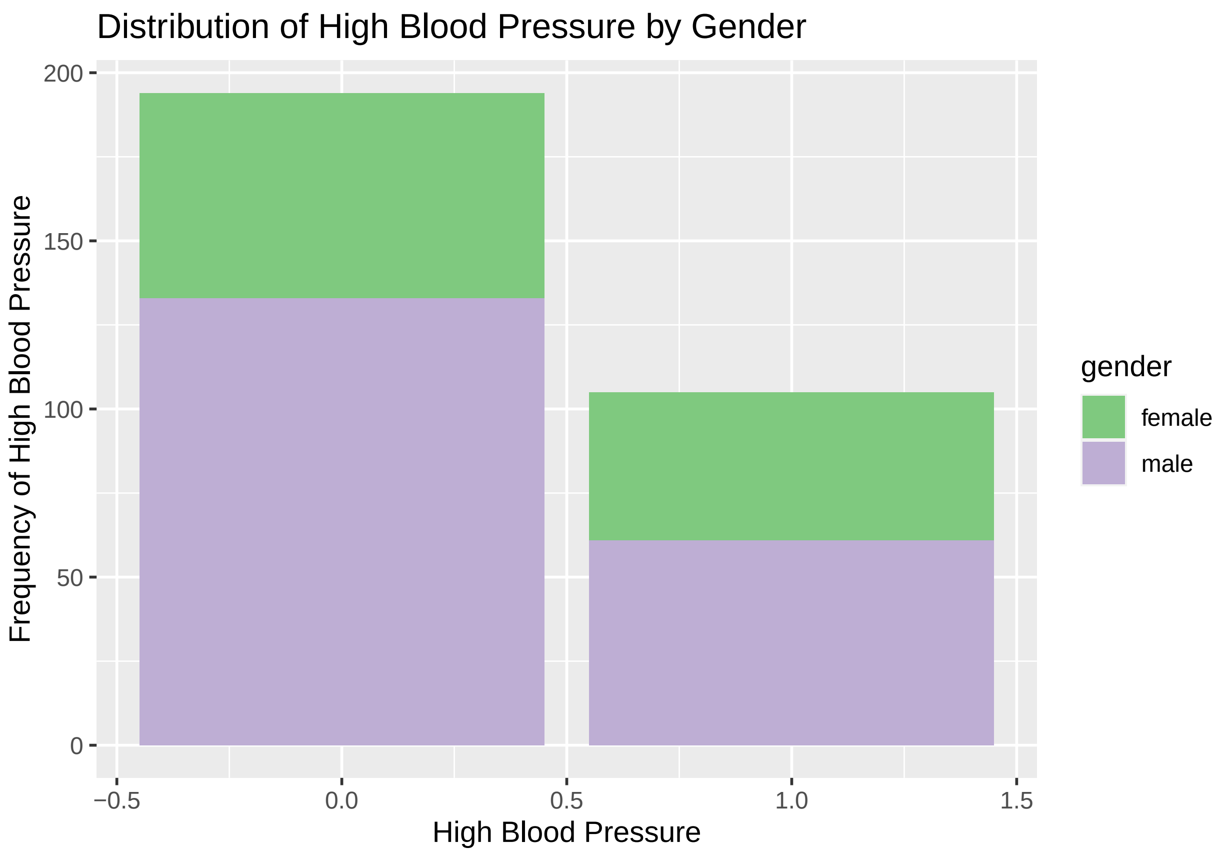
```
myocard %>% group_by(gender) %>% summarise(mean(serum_sodium), sd(serum_sodium))
```

```
## # A tibble: 2 x 3
##   gender `mean(serum_sodium)` `sd(serum_sodium)`
##   <chr>                 <dbl>              <dbl>
## 1 female                 137.               4.90
## 2 male                   137.               4.13
```

3.2.7 High Blood Pressure

High blood pressure is medically known as hypertension. When somebody has high blood pressure, the heart is working too hard to pump the blood for all the body. In our case, the variable high blood pressure is measurable by meaning the number 1 is the patients who have high pressure and 0 who don't have pressure.

```
myocard %>% ggplot(aes(high_blood_pressure)) + geom_bar(aes(fill = gender)) +
        scale_fill_brewer(palette = "Accent") + ylab("Frequency of High Blood Pressure") +
        xlab("High Blood Pressure") + ggtitle("Distribution of High Blood Pressure by Gender")
```



We can see the more patients doesn't have high blood pressure. We can easily see by the code below the proportions of males and females in high blood pressure.

```
myocard %>% group_by(gender) %>% summarise(high_blood_pressure = mean(high_blood_pressure == 0)) %>% fi
```

```
## # A tibble: 1 x 2
##   gender high_blood_pressure
##   <chr>                <dbl>
## 1 male                 0.686
```

```
myocard %>% group_by(gender) %>% summarise(high_blood_pressure = mean(high_blood_pressure == 1)) %>% fil
```

```
## # A tibble: 1 x 2
##   gender high_blood_pressure
##   <chr>                <dbl>
## 1 male                 0.314
```

```
myocard %>% group_by(gender) %>% summarise(high_blood_pressure = mean(high_blood_pressure == 0)) %>% fil
```

```
## # A tibble: 1 x 2
##   gender high_blood_pressure
##   <chr>                <dbl>
## 1 female               0.581
```

```
myocard %>% group_by(gender) %>% summarise(high_blood_pressure = mean(high_blood_pressure == 1)) %>% fil
```

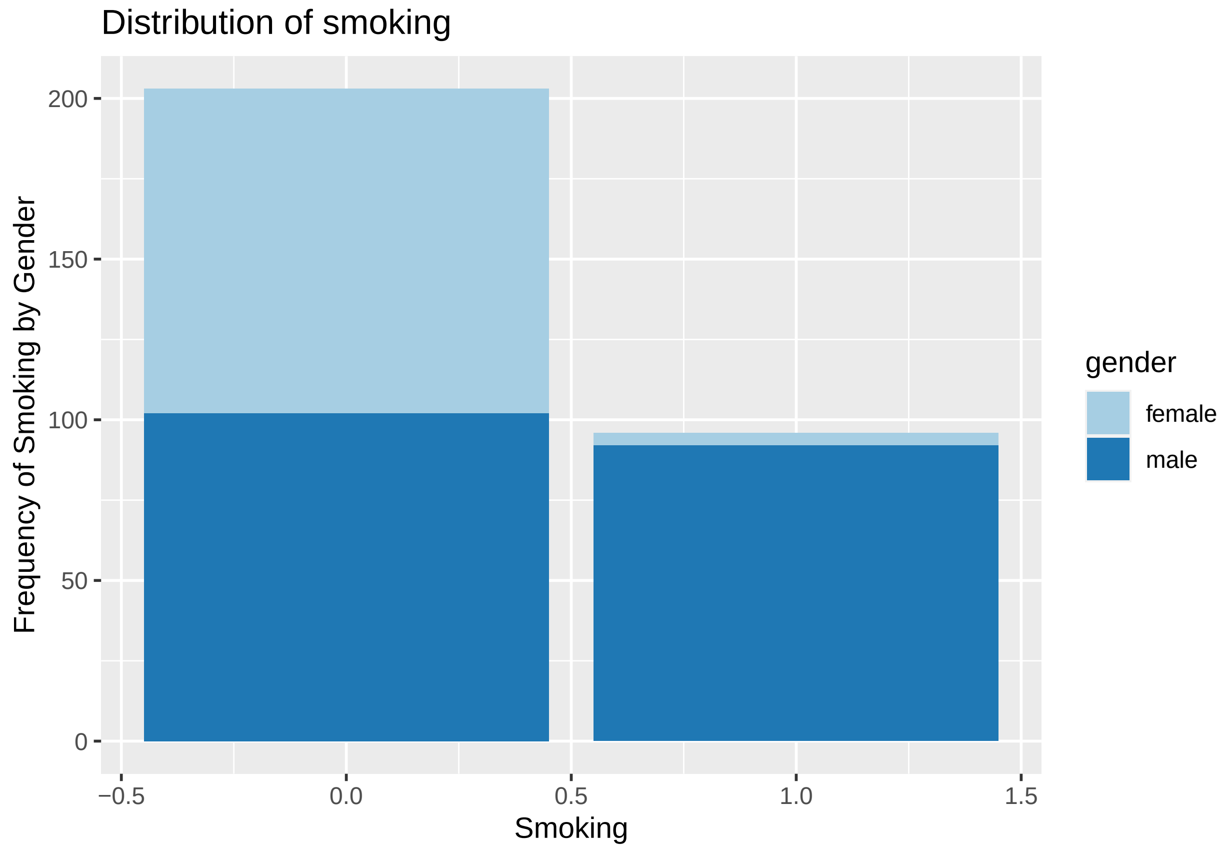
```
## # A tibble: 1 x 2
##   gender high_blood_pressure
##   <chr>                <dbl>
## 1 female               0.419
```

We can quickly notice that there are more patients male with high blood pressure than males don't have.

3.2.8 Distribution of Smoking

Smoking is the most significant health issue worldwide. Besides that, smoking causes problems breathing and lung cancer. People who smoke are two to four times more likely to get MI. In our case, the variable smoking is measurable by meaning the number 1 is the patients who somking and 0 who don't smoking.

```
myocard %>% ggplot(aes(smoking)) + geom_bar(aes(fill = gender)) +
        scale_fill_brewer(palette = "Paired") + ylab("Frequency of Smoking by Gender") +
        xlab("Smoking") + ggtitle("Distribution of smoking")
```



And now, what the proportion of woman and men smoking?

```
myocard %>% group_by(gender) %>% summarise(smoking = mean(smoking == 0)) %>% filter(gender == "male")
```

```
## # A tibble: 1 x 2
##   gender smoking
##   <chr>    <dbl>
## 1 male     0.526
```

```
myocard %>% group_by(gender) %>% summarise(smoking = mean(smoking == 1)) %>% filter(gender == "male")
```

```
## # A tibble: 1 x 2
##   gender smoking
##   <chr>    <dbl>
## 1 male     0.474
```

```
myocard %>% group_by(gender) %>% summarise(smoking = mean(smoking == 0)) %>% filter(gender == "female")
```

```
## # A tibble: 1 x 2
##   gender smoking
##   <chr>    <dbl>
## 1 female   0.962
```

```
myocard %>% group_by(gender) %>% summarise(smoking = mean(smoking == 1)) %>% filter(gender == "female")
```

```
## # A tibble: 1 x 2
##   gender smoking
##   <chr>    <dbl>
## 1 female  0.0381
```

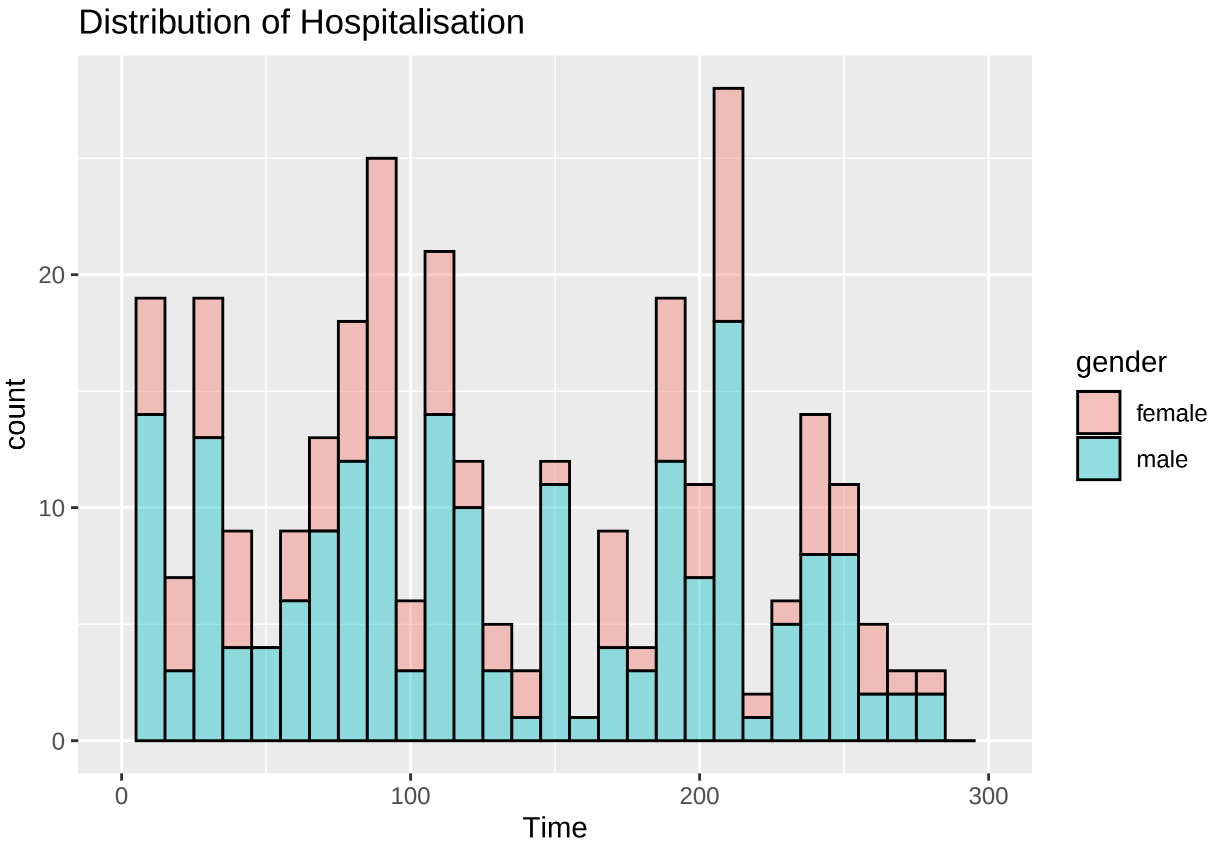
We can notice the number of women doesn't smoke higher than woman smoking, either comparing with man smoking or not.

3.2.9 Distribution of Time (Hospitalisation)

Time of hospitalisation is another risk for MI. Studies have shown the times of hospitalisation associated with comorbidity increase of risk for MI. As sad before, the Time variable is counting how many time the patients has of hospitalisation.

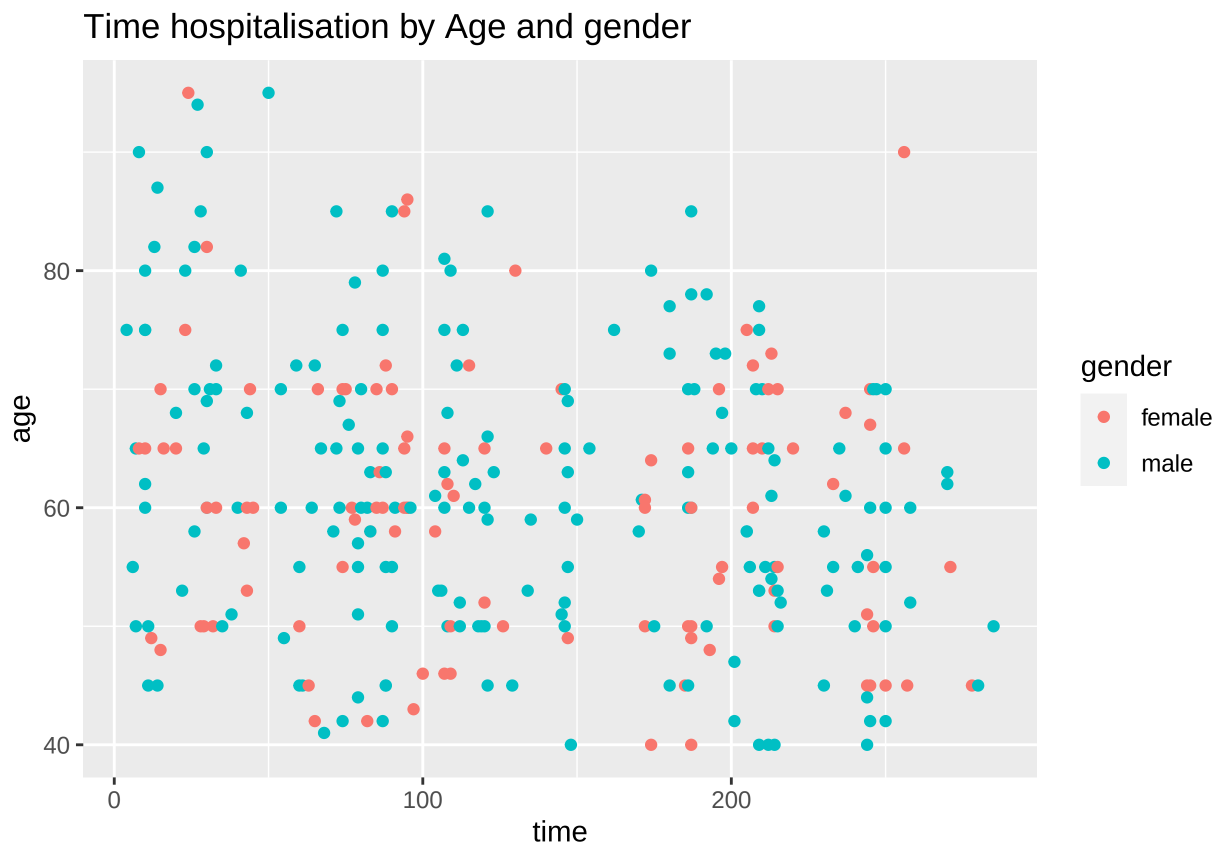
```
myocard %>% ggplot(aes(time, fill = gender)) +
        geom_histogram(binwidth = 10, bins = 5, color = "black", alpha = 0.4) +
        xlim(c(0,300)) +
        xlab("Time") + ggtitle("Distribution of Hospitalisation")
```

```
## Warning: Removed 4 rows containing missing values (geom_bar).
```



Observing this histogram graph, we can see the distribution of data by the time is not a normal distribution. We can see a high peak in around 200 times.

```
myocard %>% ggplot(aes(time, age)) + geom_point(aes(color = gender)) + ggtitle("Time hospitalisation by
```



We can notice in this graph that it's hard to know the distribution, but we can see a man has shown more in the scatterplot, but the number of man in our case is higher than the woman.

4.0 Linear Regression

Until now, we have observed the distribution of comorbidity in our case MI. The second part of this project is to build a machine learning model, and a linear regression model is considered a machine learning model.

Multiple Linear Regression is a technique statistical that use several variables exploratory to predict the outcome of a response variable. We have several variable exploratories (independent variable) and response (in our case, time-variable) outcomes.

```
fit <- lm(time ~ age + sex + high_blood_pressure + smoking + creatinine_phosphokinase + serum_creatinine
fit
```

```
##
## Call:
## lm(formula = time ~ age + sex + high_blood_pressure + smoking + 
##     creatinine_phosphokinase + serum_creatinine + serum_sodium, 
##     data = myocard)
##
## Coefficients:
##              (Intercept)                       age                       sex  
##                67.925116                 -1.232983                 -0.408229  
##      high_blood_pressure                   smoking  creatinine_phosphokinase  
##               -30.351623                 -5.279724                 -0.003481  
##         serum_creatinine              serum_sodium  
##                -8.168100                  1.195768  
```

```
summary(fit)
```

```
##
## Call:
## lm(formula = time ~ age + sex + high_blood_pressure + smoking +
##     creatinine_phosphokinase + serum_creatinine + serum_sodium,
##     data = myocard)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -152.453  -59.491   -9.518   68.181  152.762
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               67.925116 139.584667   0.487  0.62689
## age                       -1.232983   0.371034  -3.323  0.00100 **
## sex                       -0.408229  10.183948  -0.040  0.96805
## high_blood_pressure      -30.351623   9.134630  -3.323  0.00101 **
## smoking                   -5.279724  10.312968  -0.512  0.60907
## creatinine_phosphokinase  -0.003481   0.004492  -0.775  0.43904
## serum_creatinine          -8.168100   4.296581  -1.901  0.05828 .
## serum_sodium               1.195768   0.997905   1.198  0.23178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.41 on 291 degrees of freedom
## Multiple R-squared:  0.1025, Adjusted R-squared:  0.08087
## F-statistic: 4.745 on 7 and 291 DF,  p-value: 4.572e-05
```

```
tidy(fit)
```

```
## # A tibble: 8 x 5
##   term                     estimate std.error statistic p.value
##   <chr>                       <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)              67.9     140.         0.487  0.627
## 2 age                      -1.23      0.371     -3.32   0.00100
## 3 sex                      -0.408    10.2       -0.0401 0.968
## 4 high_blood_pressure     -30.4       9.13      -3.32   0.00101
## 5 smoking                  -5.28     10.3       -0.512  0.609
## 6 creatinine_phosphokinase  -0.00348   0.00449   -0.775  0.439
## 7 serum_creatinine         -8.17      4.30      -1.90   0.0583
## 8 serum_sodium              1.20      0.998      1.20   0.232
```

4.1 Linear Regression and Results

The overall quality of the multiple linear regression can be assessing following three quantify display in summary function.

Residual Standard Error (RSE): this represents the average of the outcomes and the predicted values by the model. Which lowest RSE, is better for fit the model in our data. We can notice we have 74.41 RSE in our project results, and meaning is a high value for RSE. Multiple R-square and Adjust R-square: the multiple correlations between three or more variables. It tells us how useful the predictor variables are at predicting the value of the response variable. In our case, we can see the Multiple R-square is 0.1025. It is a good result however, it doesn't mean it is a good fit. Adjust R-square: This is a correction for the number of x variables included in the predictive model. Adjust R-square near to 1 indicates that the regression model has explained a large proportion of the variability in the outcome. In our case is around 0.080. F-statistical: It's the overall significance of the model. It assesses whether at least one predictor variable has a non-zero coefficient. In statistical p-value significant is <= 0.05, our case show 4.572e-05 for p-value, showing us very strong p-value.

What the intercept of the model?

```
fit$coef[1]
```

```
## (Intercept)
##    67.92512
```

Looking back at the display of the summary function, we can see the age and high_blood_pressure has to asterisk. We can see the strong p-value between time and these two variables. However, we can see the plot doesn't show strong relation when we plot the linear regression model.

4.1.2 Predict age by time

```
time_vs_age_hbp <- lm(time ~ age + high_blood_pressure, data = myocard)
time_vs_age_hbp
```

```
##
## Call:
## lm(formula = time ~ age + high_blood_pressure, data = myocard)
##
## Coefficients:
##         (Intercept)                  age  high_blood_pressure
##             222.740               -1.354              -28.744
```

```
tidy(time_vs_age_hbp)
```

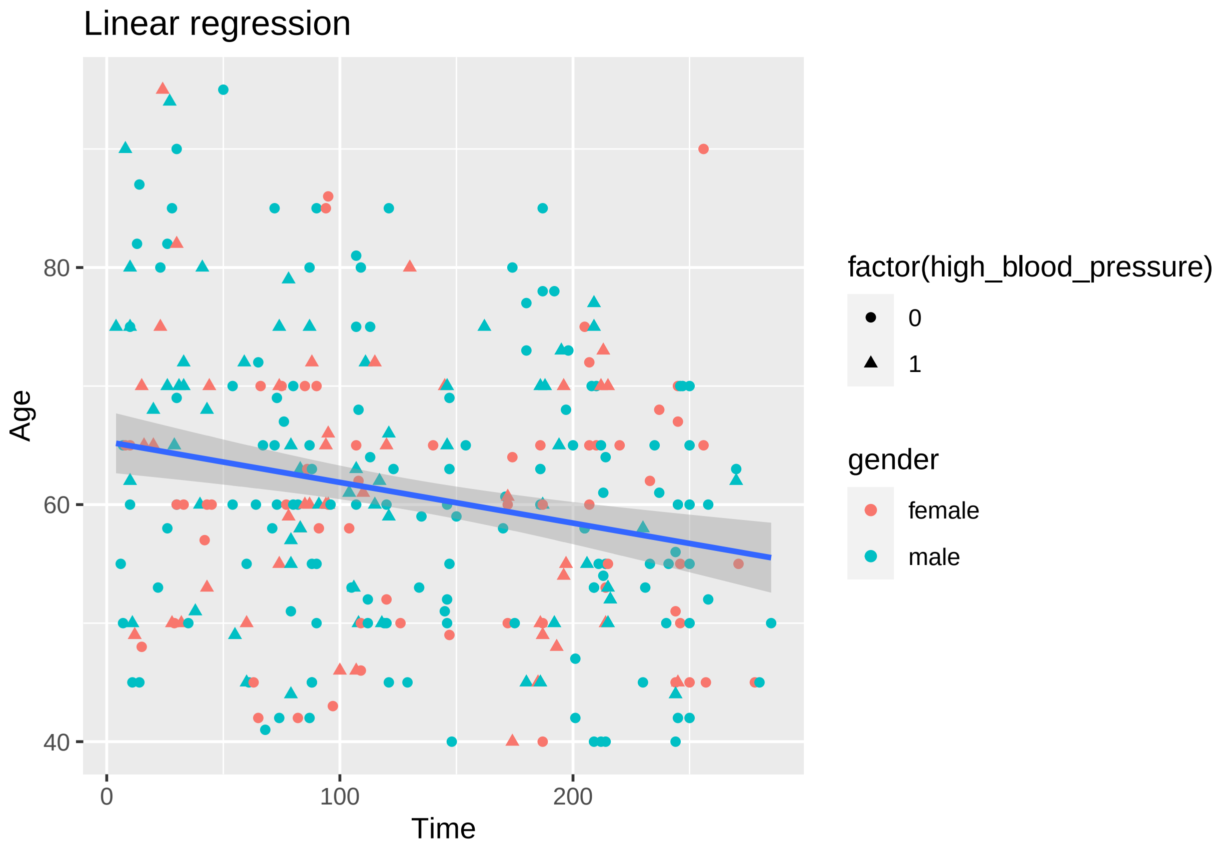
```
## # A tibble: 3 x 5
##   term                estimate std.error statistic  p.value
##   <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)           223.      22.6        9.87 4.64e-20
## 2 age                    -1.35     0.365     -3.71 2.48e- 4
## 3 high_blood_pressure   -28.7      9.08      -3.16 1.71e- 3
```

```
summary(time_vs_age_hbp)
```

```
##
## Call:
## lm(formula = time ~ age + high_blood_pressure, data = myocard)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -150.798  -60.870   -9.381   67.473  159.969
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         222.7404    22.5594   9.874  < 2e-16 ***
## age                  -1.3543     0.3651  -3.709 0.000248 ***
## high_blood_pressure -28.7444     9.0830  -3.165 0.001714 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.64 on 296 degrees of freedom
## Multiple R-squared:  0.08129,    Adjusted R-squared:  0.07508
## F-statistic:  13.1 on 2 and 296 DF,  p-value: 3.551e-06
```

```
myocard %>% ggplot(aes(time, age)) + geom_point(aes(color = gender, shape = factor(high_blood_pressure)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We can see for each increase at the variable age, the time decrease -1.35 and high_blood_pressure -28.7.

4.1.3 Conclusion

Until now, we made a multiple linear regression, trying to find out if the comorbidity or health issue predicts the time in hospitals in patients MI. In our results, we found a significant relationship between age and high_blood_pressure with time. Specially we found for each new data add in age the time decrease -1.354 and each new data in high_blood_pressure decrease time in -28.744.

5.0 Machine Learning

The large of information in data healthcare have increasingly risen in the last decade. Understanding and quantify extensive healthcare data can lead to the expected risk and predict disease in patients. Machine learning is the study of tools and methods for identifying data patterns (Wienns J, 2017). Understanding these methods can be used to predict the risk for disease and predict the future. This project uses two parameters in machine learning, LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis).

5.1 Linear Discriminant Analysis - LDA

Linear discriminant analysis is used as a tool for classification, dimension reduction, and data visualisation. Also is a linear machine learning algorithm used for multi-class classification. LDA seeks to best separate (or discriminate) the samples in the training dataset by their class value. For this, we are classifying the data by the median in the time. We are using the mean value on time and considering this "high" or "low" for hospitalisation time. Value is separated into "high" for risk to MI and "low" risk to MI. At this part of the project, we are going to use all the variables in it. For this step, we are split the myocard data set into train and test. The train data set has 10%, and the test has 90% of the original data. Also, it is an important method that evaluates the accuracy of the dataset.

Until now, we create the gender variable for better visualisation of data. However, for doing a machine learning algorithm, we are going to exclude this variable.

```
myocard <- subset(myocard, select = -14)
```

```
mean(myocard$time)
```

```
## [1] 130.2609
```

```
myocard <- myocard %>% mutate(time = ifelse(time > 130, "high", "low"))

myocard$time <- factor(myocard$time, levels = c("high", "low"))
```

```
set.seed(1)
test_index <- createDataPartition(myocard$time, times = 1, p = 0.1, list = FALSE)
train_set <- myocard[-test_index, ]
test_set <- myocard[test_index, ]

lda_fit <- train(time ~., method = "lda", data = train_set)
lda_predict <- predict(lda_fit, test_set)
confusionMatrix(lda_predict, test_set$time)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction high low
##       high    9   7
##       low     5  10
##
##                Accuracy : 0.6129
##                  95% CI : (0.4219, 0.7815)
##     No Information Rate : 0.5484
##     P-Value [Acc > NIR] : 0.2960
##
##                   Kappa : 0.2282
##
##  Mcnemar's Test P-Value : 0.7728
```

```
##
##             Sensitivity : 0.6429
##             Specificity : 0.5882
##          Pos Pred Value : 0.5625
##          Neg Pred Value : 0.6667
##              Prevalence : 0.4516
##          Detection Rate : 0.2903
##    Detection Prevalence : 0.5161
##       Balanced Accuracy : 0.6155
##
##        'Positive' Class : high
##
```

When we look at our results on sensitivity and specificity parameters, we also conclude that on our model LDA: the sensitivity 0.6429 means 64% of patients have the risk for MI and fails in 36%. While specificity 0.5882 means 58% doesn't have relations with time and fails 42% and an Accuracy of 0.6129.

5.2 Quadratic Discriminant Analysis - QDA

QDA is a variant of LDA in which an individual covariance matrix is estimated for every class of observations. Discriminant analysis is used to determine which variables discriminate between two or more naturally occurring groups.

```
qda_fit <- train(time ~., method = "qda", data = train_set)
qda_predict <- predict(qda_fit, test_set)
confusionMatrix(qda_predict, test_set$time)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction high low
##       high    8  11
##       low     6   6
##
##                Accuracy : 0.4516
##                  95% CI : (0.2732, 0.6397)
##     No Information Rate : 0.5484
##     P-Value [Acc > NIR] : 0.8965
##
##                   Kappa : -0.0733
##
##  Mcnemar's Test P-Value : 0.3320
##
##             Sensitivity : 0.5714
##             Specificity : 0.3529
##          Pos Pred Value : 0.4211
##          Neg Pred Value : 0.5000
##              Prevalence : 0.4516
##          Detection Rate : 0.2581
##    Detection Prevalence : 0.6129
##       Balanced Accuracy : 0.4622
##
##        'Positive' Class : high
##
```

In our case, when we look at our results on sensitivity and specificity parameters, we also conclude that on our model QDA: the sensitivity 0.5714 means 57% of patients have the risk for MI and fails in 43%. While specificity 0.3529 means 35% doesn't have relations with time and fails 65% and an Accuracy of 0.4516.

Accuracy is a parameter that generally describes how the model performs across the data. It's a relation between the number of correct prediction to the total prediction on test data. For example, in our data, we can find the better Accuracy on the LDA model is 0.6129. This algorithm has 61% classifying patients low or high time in hospitalisation.

6.0 Conclusion

Our project tried whether the patient's time in the hospital is at risk for myocardial infarction. For this, we performed a linear regression analysis in the first part, verifying the relationship between the variables. In this part, we saw that age and time are related. In the second part, we set up a machine learning LDA and QDA model. We were checking the specificity and sensitivity of the model. We can conclude that our model can predict with an accuracy of around 60% that the time of the hospitalized patient added to the other comorbidities increases the risk of MI.

7.0 References

Zumel Nina, Mount Jhon: Practical Data Science with R. ed; Manning, 2019. book https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4804079/ https://www.sciencedirect.com/science/article/abs/pii/S0009912018311974?via%3Dihub https://ascpt.onlinelibrary.wiley.com/doi/10.1002/cpt.1803 https://www.sciencedirect.com/science/article/pii/S0960982212008159 https://www.ncbi.nlm.nih.gov/books/NBK546624/