Fadwa Elhor

Heart Disease

Data analysis

Table des matières

[I. Introduction 2](#_Toc14216838)

[II. Data Set Information 3](#_Toc14216839)

[III. Data Pre-processing 4](#_Toc14216840)

[IV. Data Cleaning and Transformation 6](#_Toc14216841)

[1. Missing Values 6](#_Toc14216842)

[2. Goal Values 7](#_Toc14216843)

[3. Gender Values 7](#_Toc14216844)

[4. Chol Values 8](#_Toc14216845)

[V. Data Exploration 8](#_Toc14216846)

[a) Age 8](#_Toc14216847)

[b) Gender 9](#_Toc14216848)

[c) Age VS Gender 10](#_Toc14216849)

[d) Cholesterol 11](#_Toc14216850)

[How Does High Cholesterol Cause Heart Disease? 11](#_Toc14216851)

[e) Chest Pain 13](#_Toc14216852)

[f) Relation between Variables 14](#_Toc14216853)

[g) Relation between Cholesterol , Gender and Heart Disease 15](#_Toc14216854)

[VI. Conclusion 16](#_Toc14216855)

# Introduction

The main objective of this project is to use R language to analyze a dataset related to heart disease. The dataset is available in UCI Center for machine learning and Intelligent systems: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>

**Why Heart Disease?**

* Heart disease or Cardiovascular diseases (CVDs) are the number 1 cause of death globally: more people die annually from CVDs than from any other cause\*.
* An estimated 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Of these deaths, 85% are due to heart attack and stroke\*.
* Over three quarters of CVD deaths take place in low- and middle-income countries\*.

\* Source: <https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>

# Data Set Information

This dataset contains 14 attributes from 4 different Hospital:

1. Hungarian Institute of Cardiology. Budapest
2. University Hospital, Zurich, Switzerland
3. University Hospital, Basel, Switzerland
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, Ohio, USA

**Attribute Information:**

1. **Age**
2. **Gender** : 0 for female, 1 for male
3. **CP** : chest pain: 4 values:

1: typical angina\*

2: atypical angina

3: non-anginal pain

4: asymptomatic

\* Angina is the medical term for chest pain or discomfort caused by a temporary disruption in the flow of blood and oxygen to the heart. People describe angina discomfort as a squeezing, suffocating or burning feeling – usually in the centre of the chest, behind the breastbone.

1. **Trestbps**: Resting blood pressure (in mm Hg on admission to the hospital)
2. **Chol:** serum cholestoral in mg/dl
3. **FBS:** fasting blood sugar:

FBS <120 mg/dL is normal 🡪 Value 0

FBS >120 mg/dL is diabete 🡪 Value 1

1. **RestECG**: resting electrocardiographic results, values:

0 - Hyperthropy

1- Normal

2 - st-t wave abnormality

1. **Thalach** : maximum heart rate achieved  in beats per minute (bpm)
2. **Exang :** exercise induced angina : 1 = yes; 0 = No
3. **Oldpeak** : ST depression induced by exercise relative to rest
4. **Slope:** the slope of the peak exercise ST segment
5. **CA:**  number of major vessels (0-3) colored by flourosopy
6. **thal**: The heart status as retrieved from Thallium test

3 = normal

6 = fixed defect;

7 = reversable defect

1. **Goal:**   0 indicates that the patient hasn’t heart disease, the values 1,2,3 and 4 patient diagnosed :

1- diagnosed with stage 1,

2- diagnosed with stage 2,

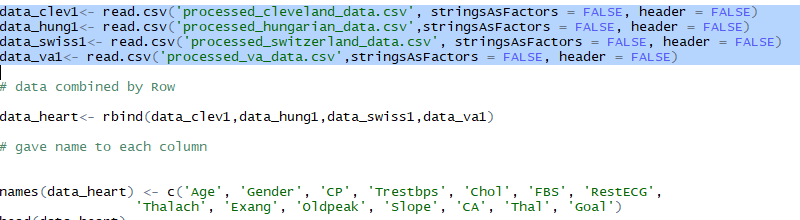
3- diagnosed with stage 3,

4- diagnosed with stage 4.

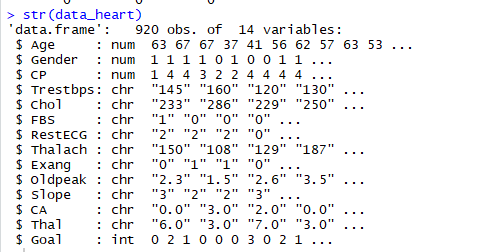
let’s try to analyze how each of those features is contributing to the disease

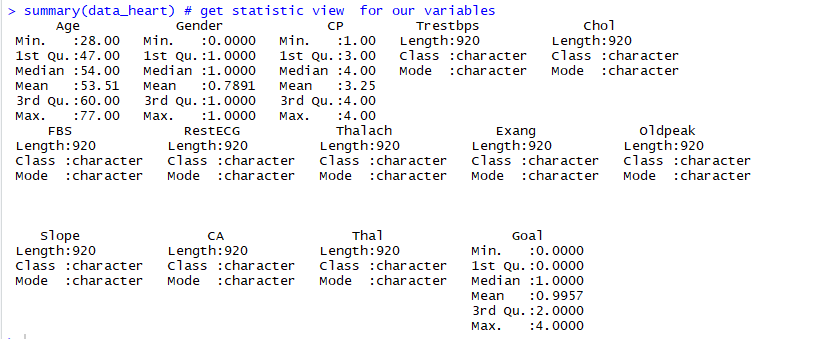
# Data Pre-processing

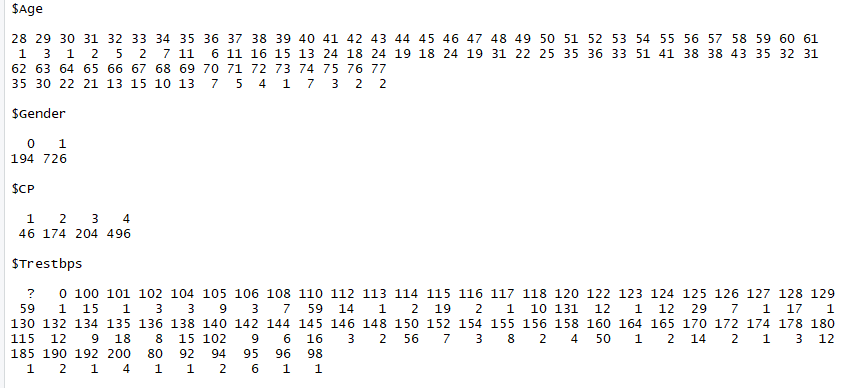
First step was to read the dataset from the CSV file and combine the 4 data in one dataframe:

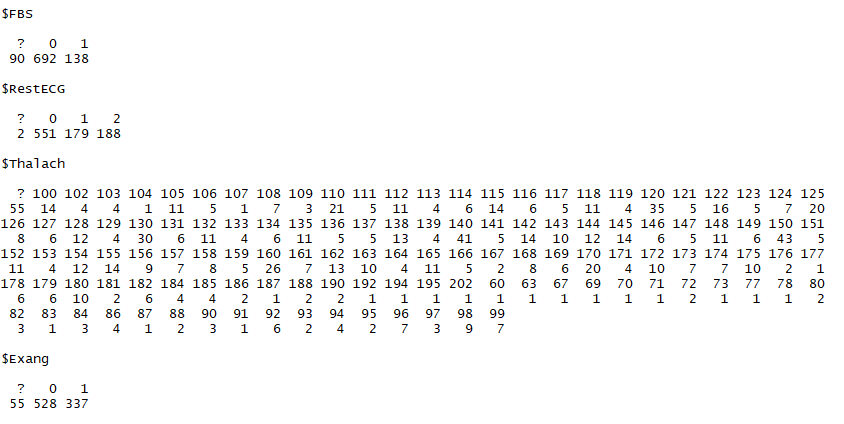


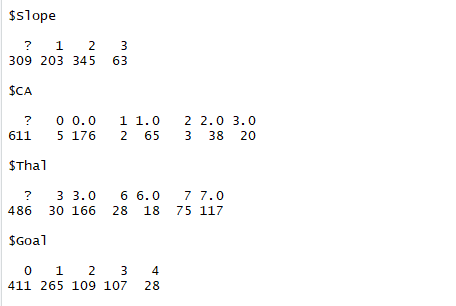
Second step was to get global information, statistic or summary view for all variables in the dataframe and table view:











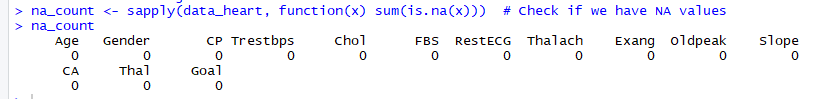
# Data Cleaning and Transformation

Using str and table function, we noticed the following anomalies:

* Some columns are not in correct data format : Gender, Chol
* Some columns reported unusual character “?”

### Missing Values

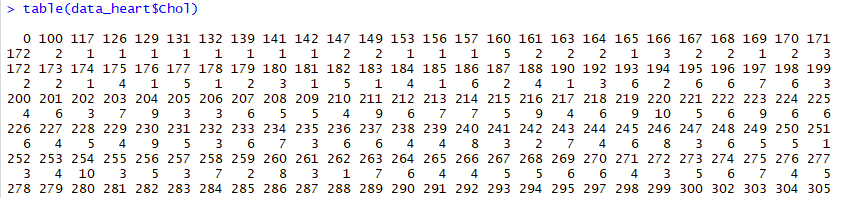
When we look for a missing value we get a wrong results:



The missing value are represented by “?”

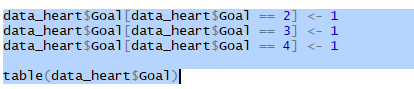
the column Chol is one of most important data to explore, that's why we will deleate the missing value in this column





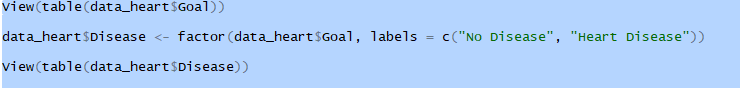
### Goal Values

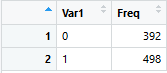
In the “Goal” column we will group the data into 2 categories : ‘no heart disease’ (value of 0) and ‘displaying heart disease’ (value of 1) so it became binary.





For a comprehensible data displaying we will add a new column “Disease” with a “Goal” variable as “heart disease” and “no heart disease” value

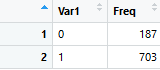
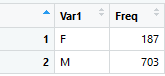


 🡪 

### Gender Values

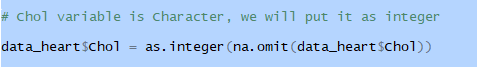
Also, for Gender variable, we will modify the value from 0 and 1 to “F” and “M”



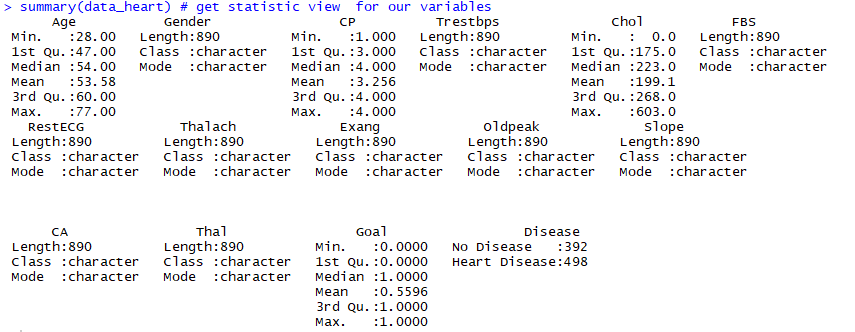
 🡪 

### Chol Values

Cholesterol is measured mg/dl, but in our data Chol variable is “character”, we have to transform to integer



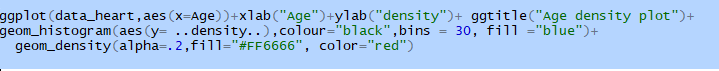
After cleaning, our data looks like:

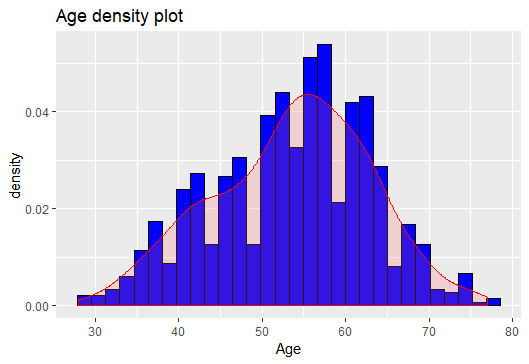


# Data Exploration

### Age

Patients from age 28 years to 77 years were included in this data set





The Histogram is showing that most of patients’ age was in between 45-65 years.

### Gender

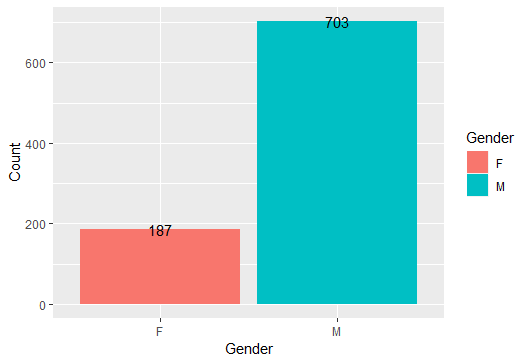
Out of all patients, around 21% were female



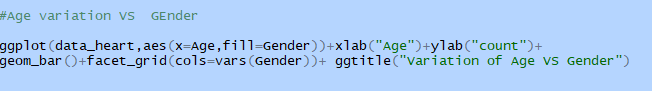


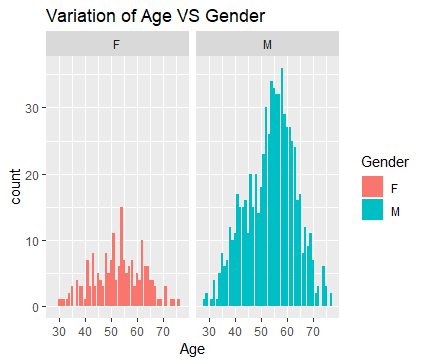
Let visualize the distribution of “Gender” Variable:





### Age VS Gender



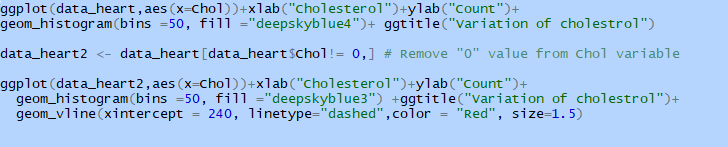


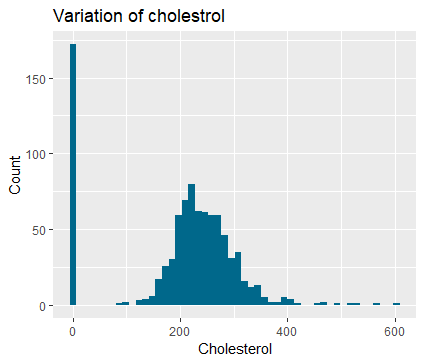
### Cholesterol

## How Does High Cholesterol Cause Heart Disease?

When there is too much cholesterol in your [blood](https://www.webmd.com/heart/anatomy-picture-of-blood), it builds up in the walls of your [arteries](https://www.webmd.com/heart/picture-of-the-arteries), causing a process called [atherosclerosis](https://www.webmd.com/heart-disease/what-is-atherosclerosis), a form of [heart disease](https://www.webmd.com/heart-disease/ss/slideshow-visual-guide-to-heart-disease). The arteries become narrowed and [blood](https://www.webmd.com/a-to-z-guides/rm-quiz-blood-basics) flow to the [heart](https://www.webmd.com/heart/picture-of-the-heart) muscle is slowed down or blocked. The blood carries oxygen to the [heart](https://www.webmd.com/heart-disease/rm-quiz-know-heart), and if not enough blood and oxygen reach your heart, you may suffer [chest pain](https://www.webmd.com/pain-management/guide/whats-causing-my-chest-pain). If the blood supply to a portion of the heart is completely cut off by a blockage, the result is a [heart attack](https://www.webmd.com/heart-disease/guide/heart-disease-heart-attacks)\*.

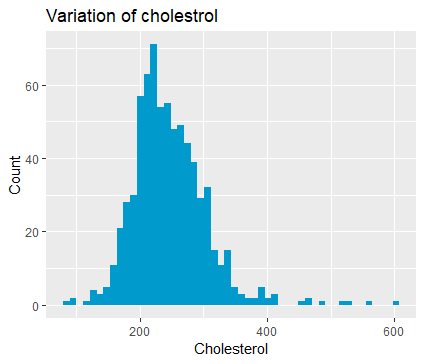
\*source: <https://www.webmd.com/heart-disease/guide/heart-disease-lower-cholesterol-risk#1>





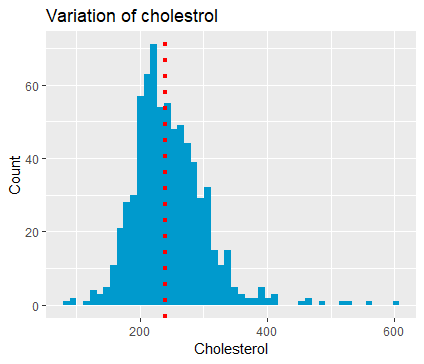
We can see that we have a huge number of patient with Choesterol variable equal 0, maybe it’s due to a missing value.

To get a best vie of our distribution, we will remove the records with this value in our graphe.

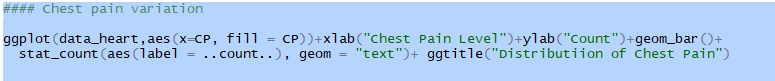


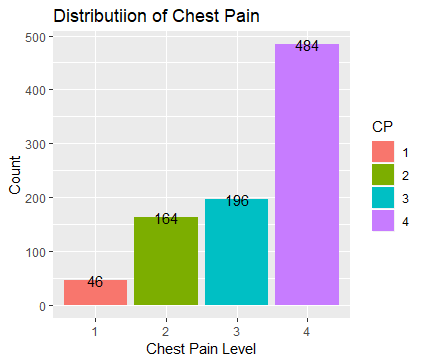
Total cholesterol levels less than 200 milligrams per deciliter (mg/dL) are considered desirable for adults. A reading between 200 and 239 mg/dL is considered borderline high and a reading of 240 mg/dL and above is considered high\*

\* Source: <https://www.medicalnewstoday.com/articles/315900.php>



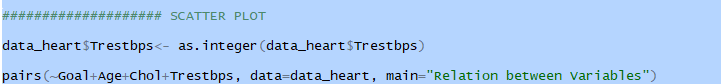
### Chest Pain

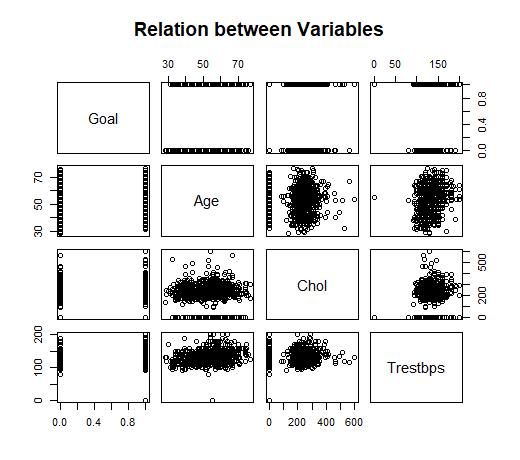




### Relation between Variables

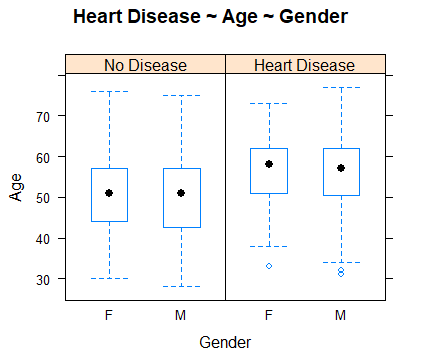
Let see the relation between heart disease variable and other variables:



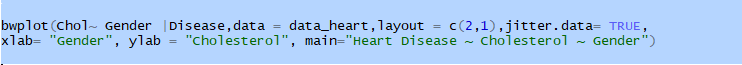


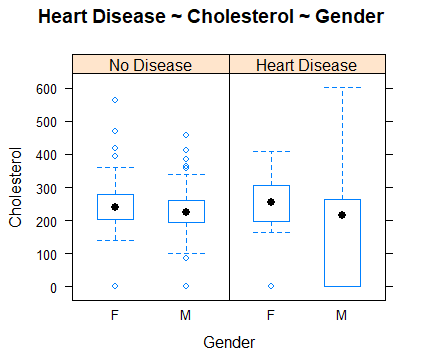
We notice that there is no direct relation between Heart disease and each other variables, but it’s an equation between Age, Gender, Cholesterol level, Blood pressure and several other parameters as showen in the graph below





### Relation between Cholesterol , Gender and Heart Disease





# Conclusion

In this report, we didn’t analyse all variables variation, but we were able to get a good idea about the parameters can cause the heart disease.