ST1009 – EXPLORATORY DATA ANALYSIS

Identify Heart Disease at the Early Stage

Group 48

S 15401 - T. R. Gamhewage
s 15402 - H. D. N Gunawardana
s 15403 - S. S. Jagoda
s 15582 – P. G. M. D. Jayarathna

Introduction

Cardiovascular disease is a class of diseases that involve the heart or blood vessels. It is commonly known as CVD. CVDs are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders that includes,

- Coronary artery diseases (CAD) such as angina
- Myocardial infarction (commonly known as a heart attack)
- Stroke
- Heart failure
- Hypertensive heart disease
- Rheumatic heart disease
- Cardiomyopathy
- Abnormal heart rhythms
- Congenital heart disease
- Peripheral artery disease
- Thromboembolic disease
- And venous thrombosis

This may be caused by high blood pressure, smoking, diabetes mellitus, lack of exercise, obesity, high blood cholesterol, poor diet, excessive alcohol consumption, and poor sleep.

Identifying those at highest risk of CVDs and ensuring they receive appropriate treatment can prevent premature deaths. Access to noncommunicable disease medicines and basic health technologies in all primary health care facilities is essential to ensure that those in need receive treatment and counselling.

In this analysis, we will discuss about a dataset which includes data about a number of patients, who are having CVDs, from different different age groups.

Introduction to Dataset

This dataset contains thirteen different variables that the medical officers used to conclude the pain experienced by this people, is whether a heart disease or not. Both qualitative and quantitative variables can be found in this dataset. Most of them are qualitative and their respective categories are represented by a given value. For Example, the qualitative variable “CP” is indicated by 4 numerical values, with respect to their categories. (Described below). There are 6 quantitative variables and 7 qualitative variables in this dataset.

This data set contains data about a hospital record of CVD patients. It includes data from both men and women who are between 29 – 77 years. 303 individuals have contributed to creating this dataset.

Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Ageing is also associated with changes in the mechanical and structural properties of the vascular wall, which leads to the loss of arterial elasticity and reduced arterial compliance and may subsequently lead to coronary artery disease. As mentioned earlier, this dataset was taken to represent a wide area of age groups among the participants when creating the dataset. The lowest age recorded was 29 years and the maximum age recorded was 77 years. You can see it in the very first column on the dataset labelled as "Age".

Gender is the next important thing in cardiovascular diseases. Among men and women, there are differences in body weight, height, body fat distribution, heart rate, stroke volume, and arterial compliance. In the very elderly, age-related large artery plasticity and stiffness is more pronounced among women than men. Also, there is a saying, if a female has diabetes, she is more likely to develop heart disease than a male with diabetes. However, according to this dataset, 207 out of the 303 records were taken from men and the other 96 records were taken by women. In this dataset gender was recorded under the column "Sex" and Males are indicated by "1", while females are indicated by "0".

In the third column, there is the variable "CP", which stands for the level of chest pain that the patient has experienced. There are 4 levels of this pain, which are indicated by numbers 0 to 3. The levels are as follows.

0 – Typical Angina	1 – Atypical Angina
2 – Non- Anginal Pain	3 – Asymptomatic

In the fourth column of the dataset, we have the variable "trestbps", which stands for the Resting Blood Pressure (mmHg) of the person. And in the fifth column of the dataset, we have the variable "chol", which stands for the Person's cholesterol level (mg/dl). As the sixth variable, we have "fbs", the fasting blood sugar amount (mg/dl) of the person. Then we have the "restecg", the resting ECG measurement of the person. Then the variable "thalach", stands for the person's maximum heart rate achieved. Variable "exang", stands for the exercise-induced angina. "oldpeak" variable stands for the ST depression induced by exercise relative to rest.

As the last three variables, we have "slope", the slope of the peak exercise ST segment, "ca" The number of major vessels and at last, the variable "target" which gives the final result whether the person has heart disease or not.

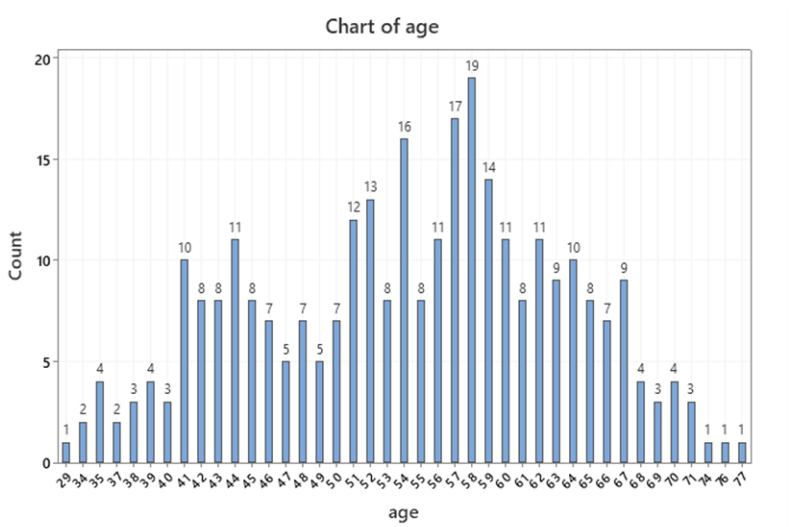
So, these are the variables that we have to analyze in our dataset.

Descriptive Analysis

Here, we are going to discuss about the spread of the data given in the respective dataset. For example, we're going to talk about the distribution of the variables, distribution of age, variance of blood sugar level, cholesterol level, blood pressure with the age and with the gender and etc. The following analysis will include graphical representations of the given data, such as tables, graphs, charts, histograms and etc.

(1) Distribution of Age

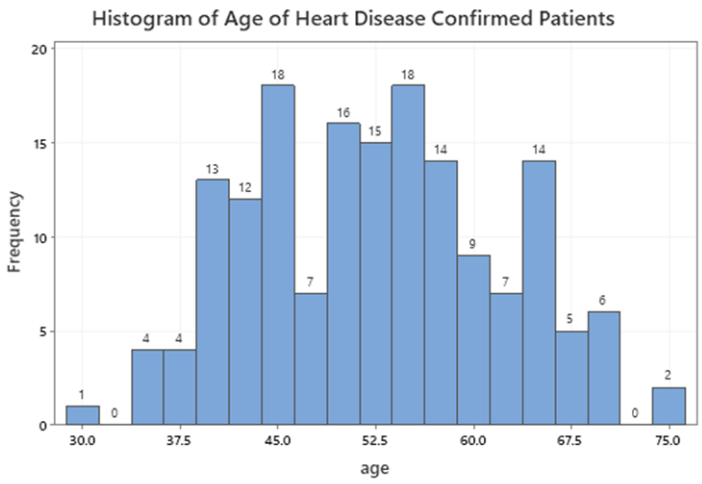
Age is the most important risk factor when considering cardiovascular diseases. So here, we are going to analyze the distribution of age of the given dataset.



This chart shows the distribution of age of the patients. There are ages from 29 years to 77 years. And the majority of the patients (19) are 58 years old. And the lowest age counts were 29,74,76 and 77 with only 1 patient from each age group.

When we take a closer look at this graph, we can see that there is a considerable number of patients are aged between 51 and 59 years. From all 303 patients, 118 individuals belong to the age group that mentioned above.

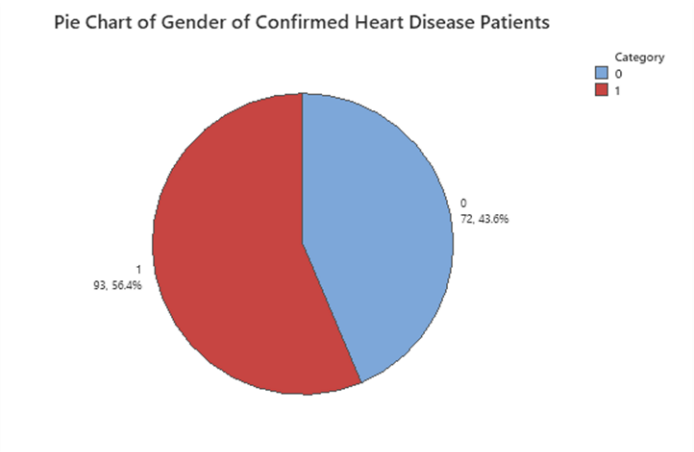
Now, let's consider the Age distribution of the patients who confirmed to have heart disease, (*click here to see the statistics of patients who confirmed to have a heart disease*)



Even here we can see that most of the Patients who confirmed to have a heart disease are aged 44 to 59 years old. The peak frequency [18] is attained at twice (bimodal) at around ages 45 & 55

(2) Gender

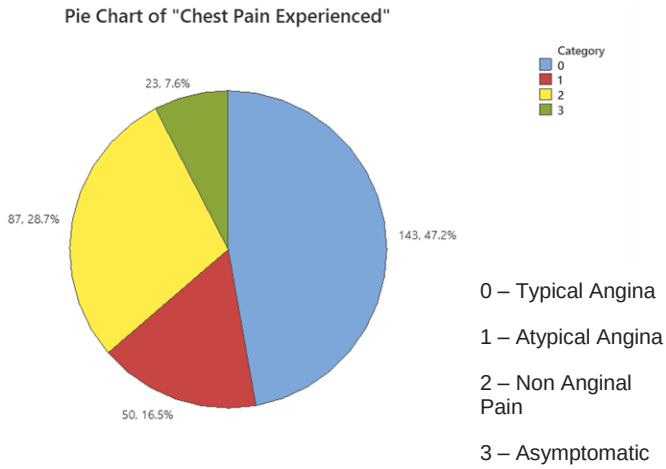
Let's consider the Gender variation among patients who confirmed to have a heart disease,



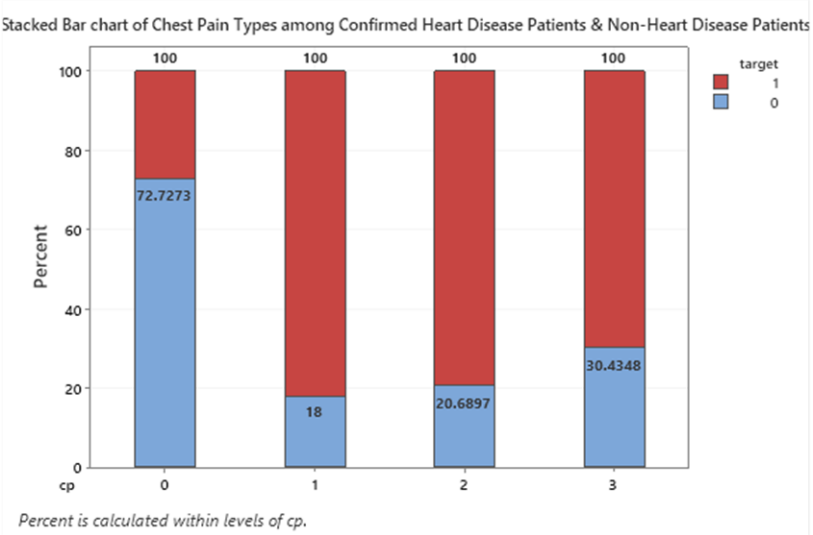
We can observe the majority of the patients who were identified as Heart patients were Male & it's 56.4%.

(3) The Chest Pain Experienced

This pie chart shows the level of the chest pain experienced by the patients. So, majority of them [47.2 %] (143/303) has experienced a typical angina meanwhile 28.7 % (87/303) of them experienced a non-anginal pain and 16.5 % (50/303) of them has experienced atypical angina. Also 7.6% (23/303) of the patients has experienced an asymptomatic pain in the chest.

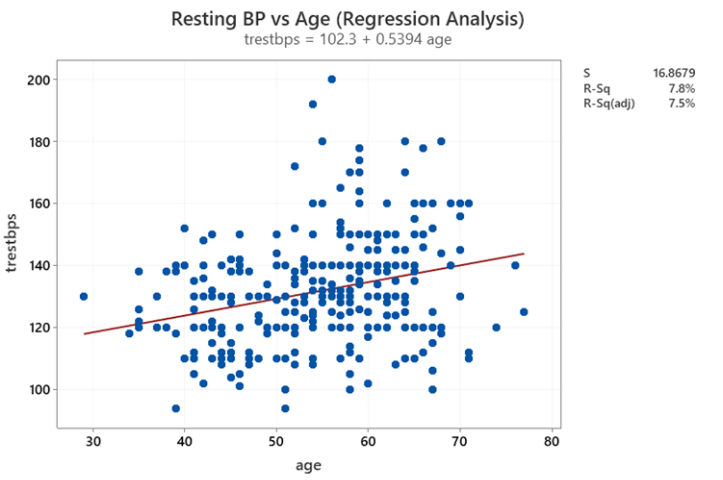


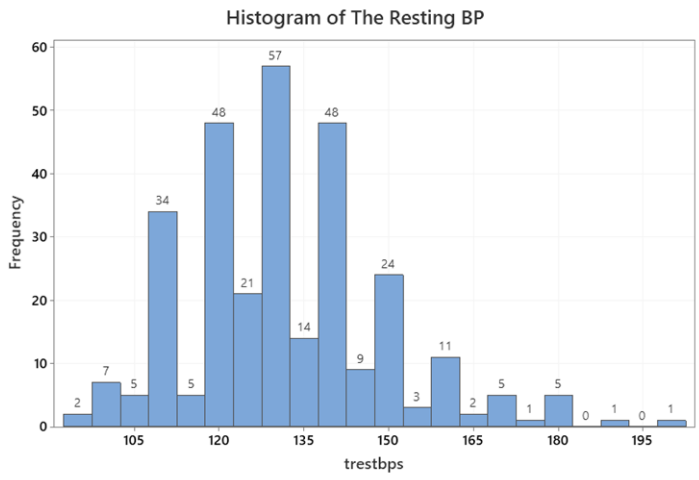
Let's consider the likelihood of each of these types to be a patient who confirmed to have a heart disease,



When considering each type individually, type 1 (atypical angina) has the most percentage [82%] of its patients being Patients who confirmed to have a heart disease, followed by type 2 (non-anginal pain) [79.3103%] & type 3 (asymptomatic) [69.5652%]. Type 0 (typical angina) has the lowest percentage of its patients being diagnosed as patients who confirmed to have a heart disease [27.2727%].

(4) Distribution of the resting blood pressure





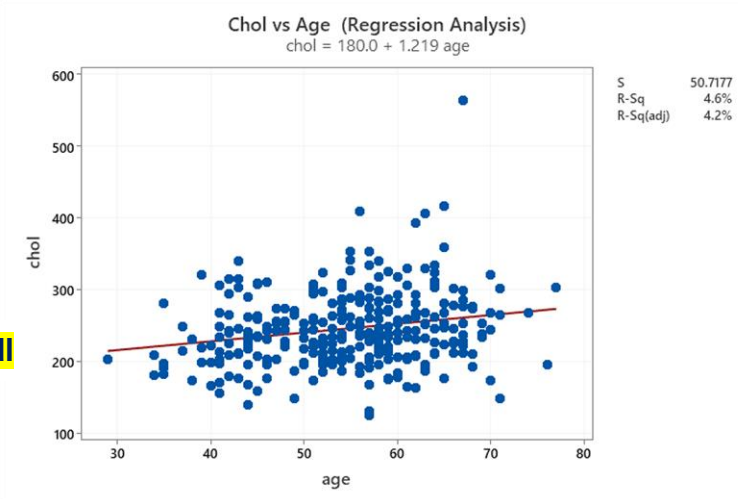
The scatterplot above, shows the distribution of the resting blood pressure, with the increment of age. According to that, the resting blood pressure gradually increases with the increment of age. The two variables have a positive relationship. Also, we have done a regression analysis about the two variables, and it shows that there is a gradient of 0.5394 mmHg in the resting blood pressure for a unit difference of age. So according to the analysis, we can calculate the resting BP for a given age by the following equation.

The Resting BP = 102.3 + (0.5394 * Age) mmHg

When we study the histogram of the resting blood pressure, we can see that the histogram is slightly right skewed. Most of the people have a normal resting blood pressure between 107.5 mmHg and 142.5 mmHg. 227 people out of 303 belongs to this region. The maximum resting blood pressure recorded was 200 mmHg and the minimum was 94 mmHg.

(5) Distribution of Blood Cholesterol Level

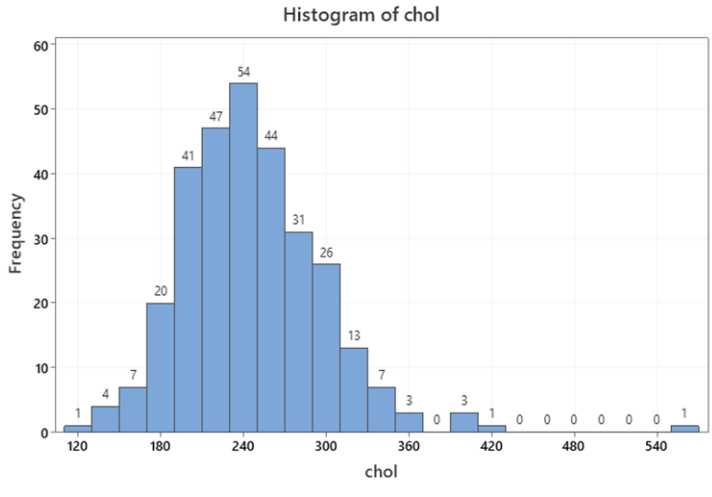
Here also we are going to analyze the distribution of the blood cholesterol level of this dataset by using a regression analysis. According to the scatterplot below, which was created from the regression analysis, show that the distribution of blood cholesterol level also has a positive relationship with the age. It shows that when the age increases in one unit, the blood pressure also increases with a gradient of 1.219 mg/dl. So according to the analysis, we can calculate the blood cholesterol level for a given age by the following equation.



The Blood Chol Level = 180.0 + (1.219 * Age) mg/dl

Also, we have created a histogram of blood cholesterol level to analyze the distribution of the variable "chol".

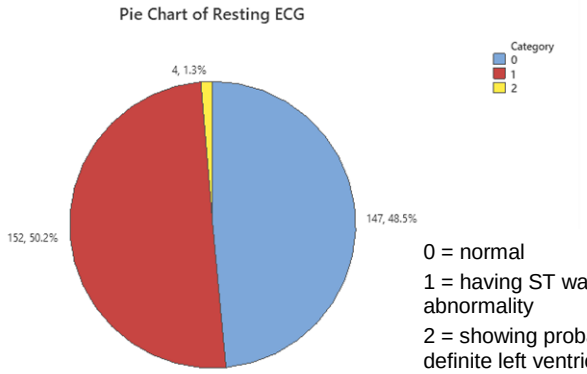
When we study the histogram of the blood cholesterol level, we can see that the histogram is right skewed. Most of the people have a normal blood cholesterol level between 190 mg/dl and 270 mg/dl. 186 people out of 303 belongs to this group. Also, two records were found that exceeding the blood cholesterol level of 400 mg/dl. Minimum blood cholesterol level recorded was 126 mg/dl and the maximum recorded value was 564 mg/dl.



(6) Distribution of Resting ECG

The following pie chart shows the distribution of the resting ECG measurement of the given dataset.

According to this, 50.2% (152/303) of the patients has a normal ECG count and 48.5% (147/303) of them has a ST wave abnormality.

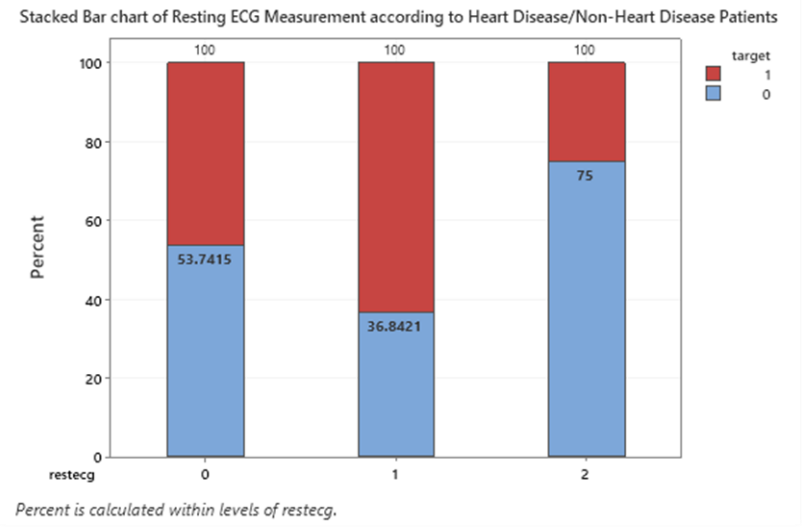


0 = normal

1 = having ST wave abnormality

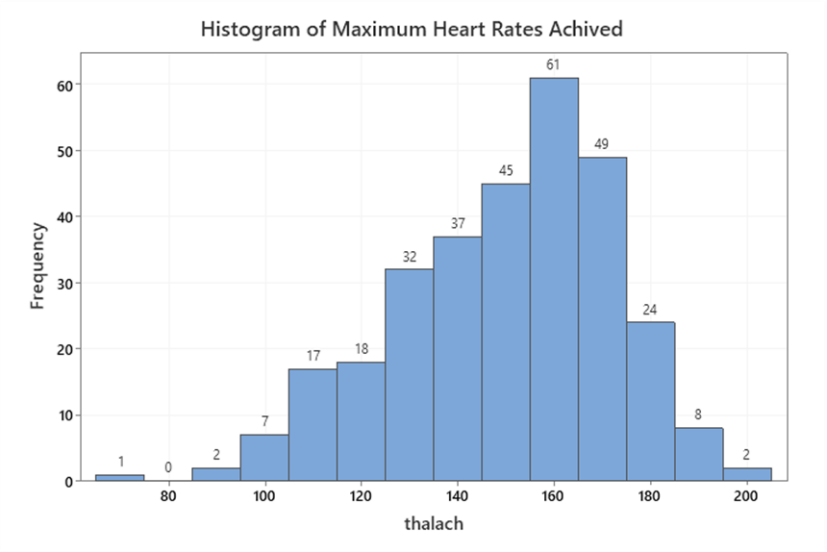
2 = showing probable or definite left ventricular hypertrophy by Estes' criteria.

Let's consider the likelihood of each of these types to be a Patient who is confirmed to have a heart disease,



We can see that When considering each type individually, type 1 (having ST-T wave abnormality) has the most percentage [63.1579%] of its patients being Patients who are confirmed to have a heart disease, followed by type 0 (normal)) [46.2585.%]. Type 2 (showing probable or definite left ventricular hypertrophy by Estes' criteria) has the lowest percentage of its patients being diagnosed as Patients who are confirmed to have a heart disease [25%].

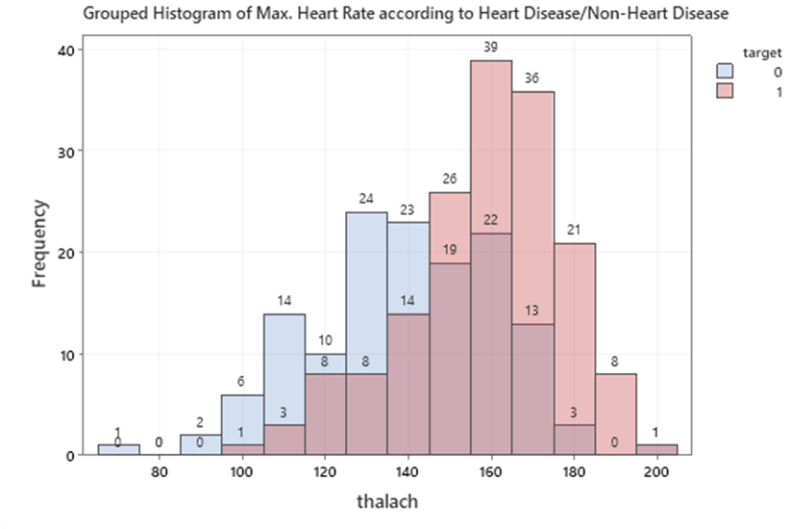
(7) Distribution of Maximum Heart Rate Achieved



According to the histogram, the distribution of the maximum heart rate achieved is left skewed. The minimum value recorded was 71 bpm and the maximum was 202 bpm. Most of the people (155) had their maximum heart rate recorded between 145 bpm and 175 bpm.

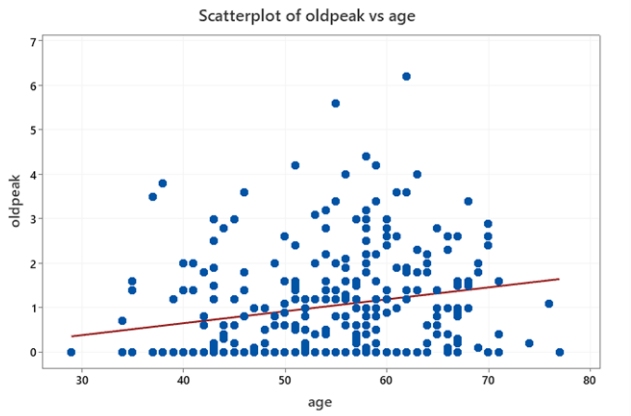
Let's consider the maximum heartrate of the Patients who are confirmed to have a heart disease vs Who are not.

Here, we can see that generally both histograms have negative (left) skewness. But it is clear that generally Confirmed Heart Patients has a much higher max. Heart rate when compared to non-Heart disease. Majority of the Confirmed Heart Disease Patients have their max. Heart Rate in the range of 150bmp to 190bpm.



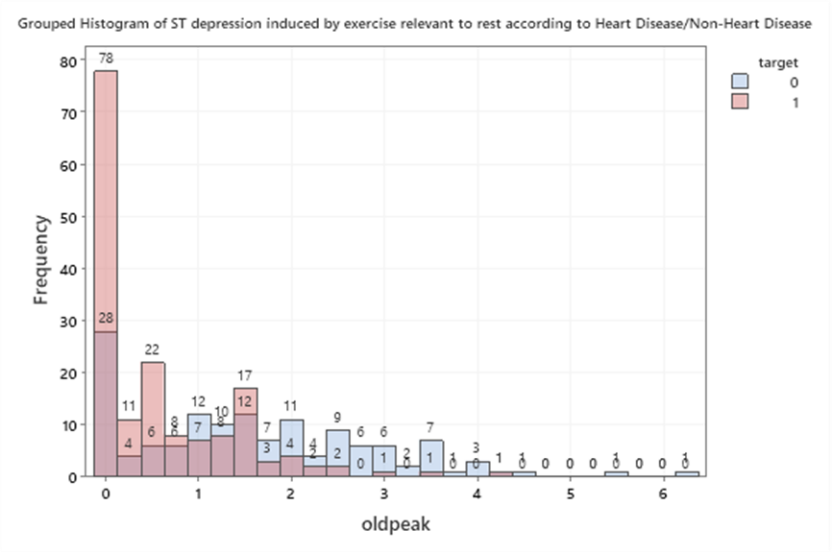
(8) Oldpeak vs Age

The variable "oldpeak" stand for the ST depression included by exercise relative to the rest. The following scatterplot shows the distribution of the variable "Oldpeak" against the increment of the age of the patients.



According to this scatterplot, "oldpeak" variable and the age of the patients has a positive relationship. That means the ST depression included by exercise relative to the rest, increases with the increment of the age.

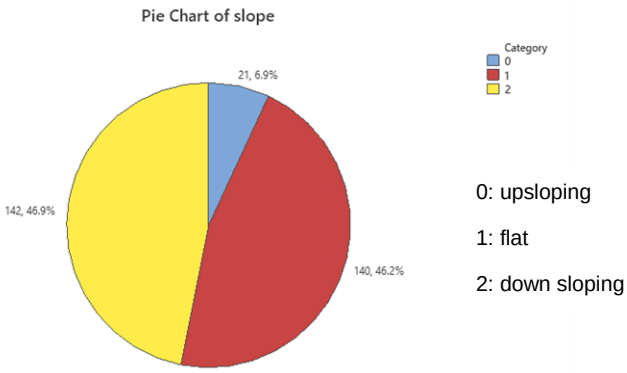
Let's consider the distribution of Confirmed heart disease patients vs Non-Heart disease patients.



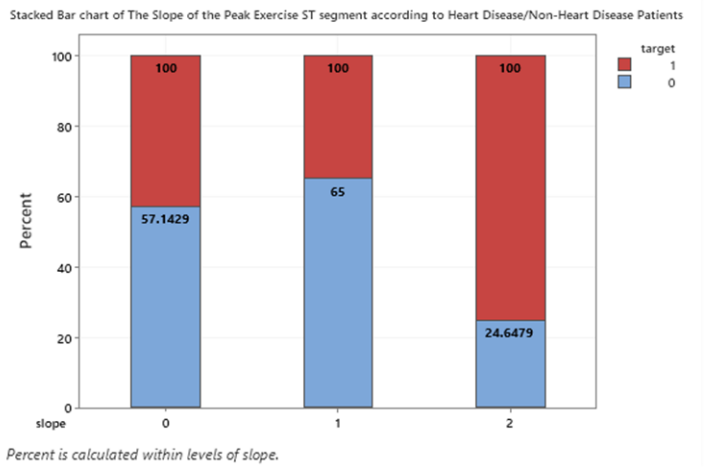
Here, we can see that generally both histograms extremely positive (right) skewed. But it is clear that generally Confirmed Heart Patients has a much lower ST depression induced by exercise relative to rest compared to non-Heart disease. Majority of the Confirmed Heart Disease Patients have a ST depression induced by exercise relative to rest value less than 1.

(9) slope of the peak exercise ST segment

This pie chart shows the slope of the peak exercise ST segment. Most of the results (46.2%) were down slopping and 46.9% were flat (normal).



Let's consider the likelihood of each of the types of slopes to be a Patient who is confirmed to have a heart disease,

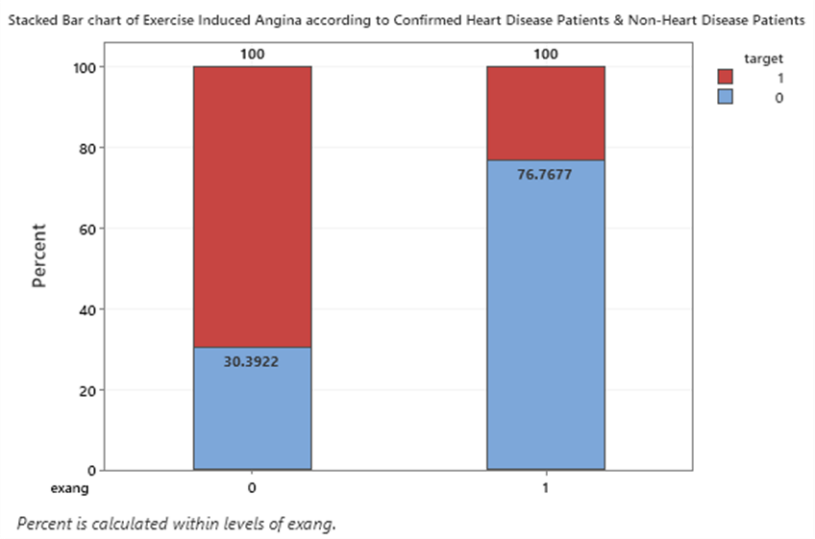


When considering the slope of the peak exercise ST segment we can see that,

Type 2 (downsloaping) has the most percentage [75.3521%] of its patients having a Heart Disease followed by type 0 (upsloping) [42.8571%]. Type 1 (flat) has the lowest percentage [35%] of its patients being diagnosed as Heart patients.

(10) Exercise induced angina

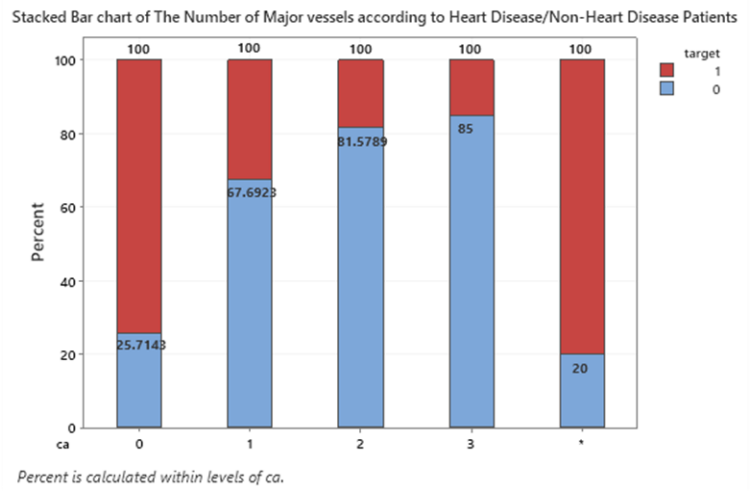
Let's consider the relationship of Exercise induced angina with Heart Disease,



We can see that when Exercise Induced Angina is absent, 69.6078% of the patients are Heart patients, but when Exercise Induced Angina is present, only 23.2323% of patients are confirmed to have a heart disease.

(11) The number of major vessels

Let's consider the relationship of the number of major vessels with Heart Disease,

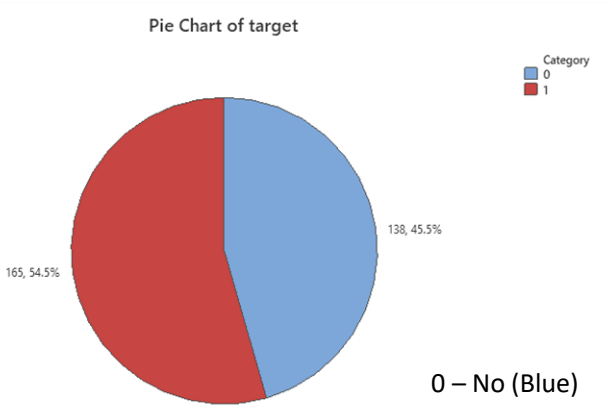


There are 5 empty entries in the dataset.

When considering each type individually, type 0 has the most percentage [74.2851%] of its patients being Patients who are confirmed to have a heart disease, followed by type 1 [32.3077%] & type 2 [18.4211%]. Type 3 has the lowest percentage of its patients being diagnosed as Heart patients [15%].

(12) Target

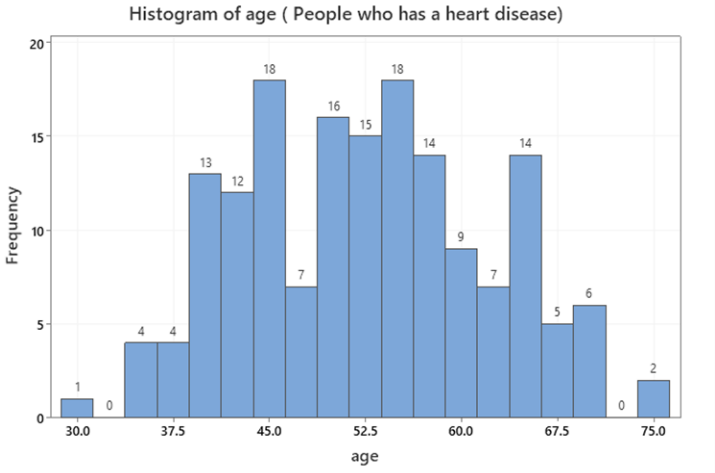
This variable shows the final decision of the medical officers, regarding each individual person is having a heart disease or not.

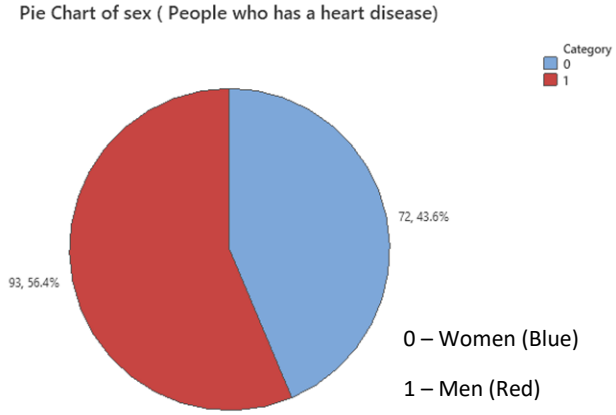


0 – No (Blue)

1 – Yes (Red)

This pie chart shows the distribution of the “target” variable. 54.5% (165) from the whole 303 patients were confirmed to have a heart disease and the other 45.5 % were found tested negative for having a heart disease. Now lest take a look at the distribution of ages of the patients who happen to have a heart disease. The mean age was 52 years.





0 – Women (Blue)

1 – Men (Red)

According to the pie chart, 56.4% (93/165) of the patients who has the heart diseases are men, and 43.6% of them (72/165) are women. According to the histogram, the peak of the heart diseased patients were recorded from the ages of 43-47 years and from 53-57 years by recording 18 patients from each age group.

Conclusion

There is no one method or indicator to determine whether one has a Heart Disease or not. But by looking at some basic Statistics we could determine whether a person is at the risk of becoming a Heart patient or not.

One's risk increases if their,

- Age is in the range of 44 to 59 years
- Gender is Male
- Experiences Atypical Anginal or non-anginal chest pains

He/She can confirm these suspicions by visiting a medical facility, where they will test to make sure.

Some of the Statistics in the reports that will increase the risk of a person being a Heart patient,

- If your maximum heart rate exceeds 150bpm
- If things like your Cholesterol level, Fasting Blood Sugar, Resting Blood Pressure are high.
- If your - Resting electrocardiographic measurement has having ST-T wave abnormality
- If your ST depression induced by exercise relative to rest is less than 1
- If your slope of the peak exercise ST segment is downsloping
- If your number of major vessels is zero