Assignment 2

Heart Disease Analysis

**21st May 2018**

*Prepared by Rahul Gupta (s3635232) and Neha Voora (s3691382)*

Table of Contents

[Executive Summary 3](#_Toc515136323)

[Introduction 3](#_Toc515136324)

[Data Exploration 3](#_Toc515136325)

[Feature Visualization – Age 3](#_Toc515136326)

[Gender Type 3](#_Toc515136327)

[Chest Pain Type 4](#_Toc515136328)

[Resting Blood Pressure 5](#_Toc515136329)

[Serum cholesterol 6](#_Toc515136330)

[Fasting Blood Sugar 7](#_Toc515136331)

[Electrocardiograph 7](#_Toc515136332)

[Maximum Heart-Rate Achieved 8](#_Toc515136333)

[Exercise Induced Angina 9](#_Toc515136334)

[Old Peak 10](#_Toc515136335)

[Slope (ST Elevation) 10](#_Toc515136336)

[Major vessels by fluoroscopy 11](#_Toc515136337)

[Thalassemia 12](#_Toc515136338)

[Heart Disease – Target Variable 13](#_Toc515136339)

[Relationship Between variables 14](#_Toc515136340)

[Age and Resting Blood Pressure 14](#_Toc515136341)

[Age and Serum Cholesterol 15](#_Toc515136342)

[Heart Rate and Blood Pressure by Age 16](#_Toc515136343)

[Blood Sugar and Cholesterol 17](#_Toc515136344)

[Gender type and heart disease 18](#_Toc515136345)

[Chest pain with gender 19](#_Toc515136346)

[Electro cardio graph and oldpeak 20](#_Toc515136347)

[Angina with Major vessel 21](#_Toc515136348)

[ST Elevation with EC type 22](#_Toc515136349)

[Thalassemia and Heart Rate 23](#_Toc515136350)

[Discussion 24](#_Toc515136351)

[Conclusion 24](#_Toc515136352)

[Bibliography 24](#_Toc515136353)

# Executive Summary

Here we explored the statlog heart data to get insights about heart disease. Features are observed individually, in groups via visualization and machine learning algorithm.

Need to write few lines of analysis

# Introduction

This report explores the (Statlog (Heart) Data Set 2004) collected from UCI Machine learning repository.

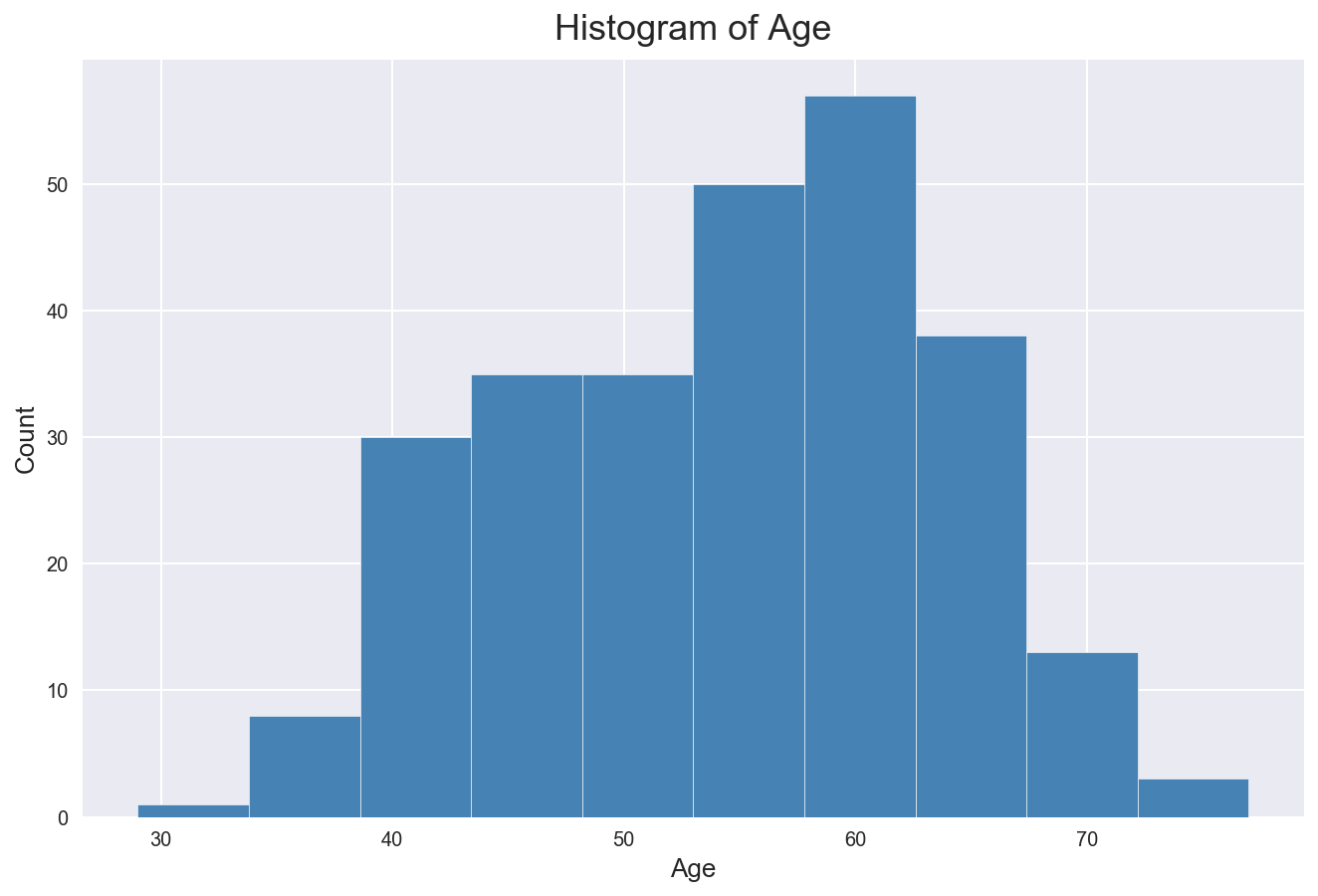
# Data Exploration

Here we’ll explore the features using python pandas and matplotlib library.

Data has no missing values.

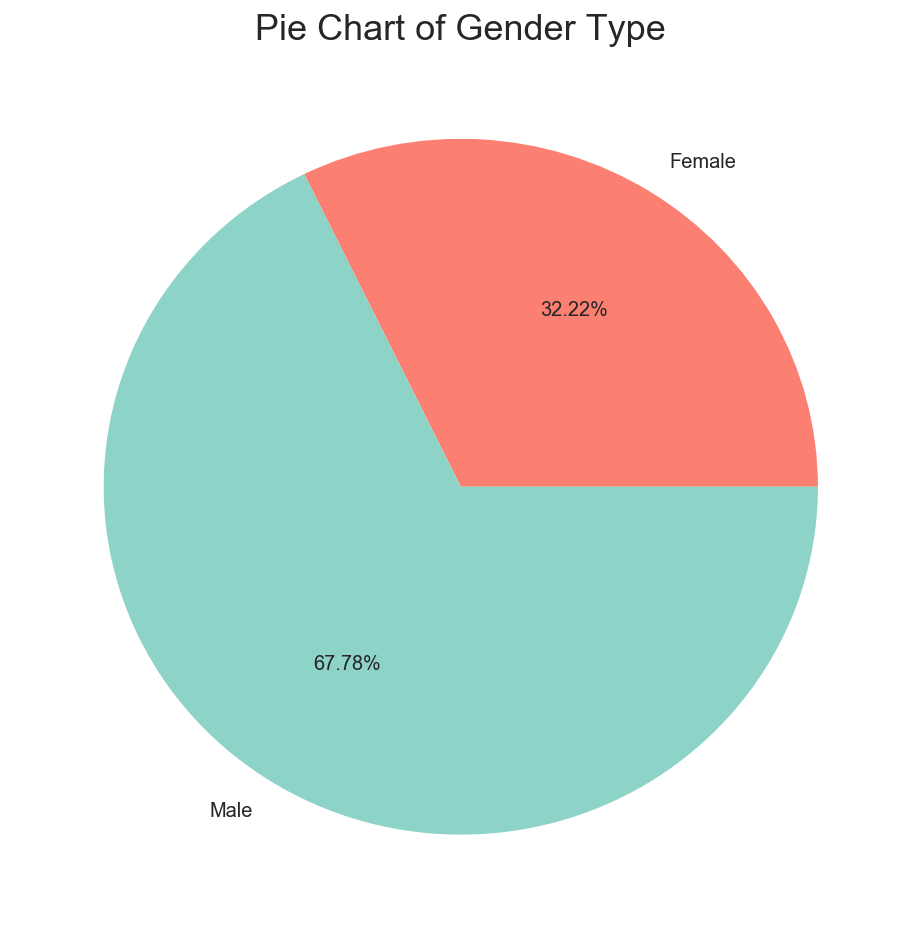
### Feature Visualization – Age

Age is a continuous variable and following visualization shows the distribution of age. Most of the data is between 50 and 70 years.



### Gender Type

Age is a nominal variable. 2/3 of sample collected are from males while remaining 1/3 are females



### Chest Pain Type

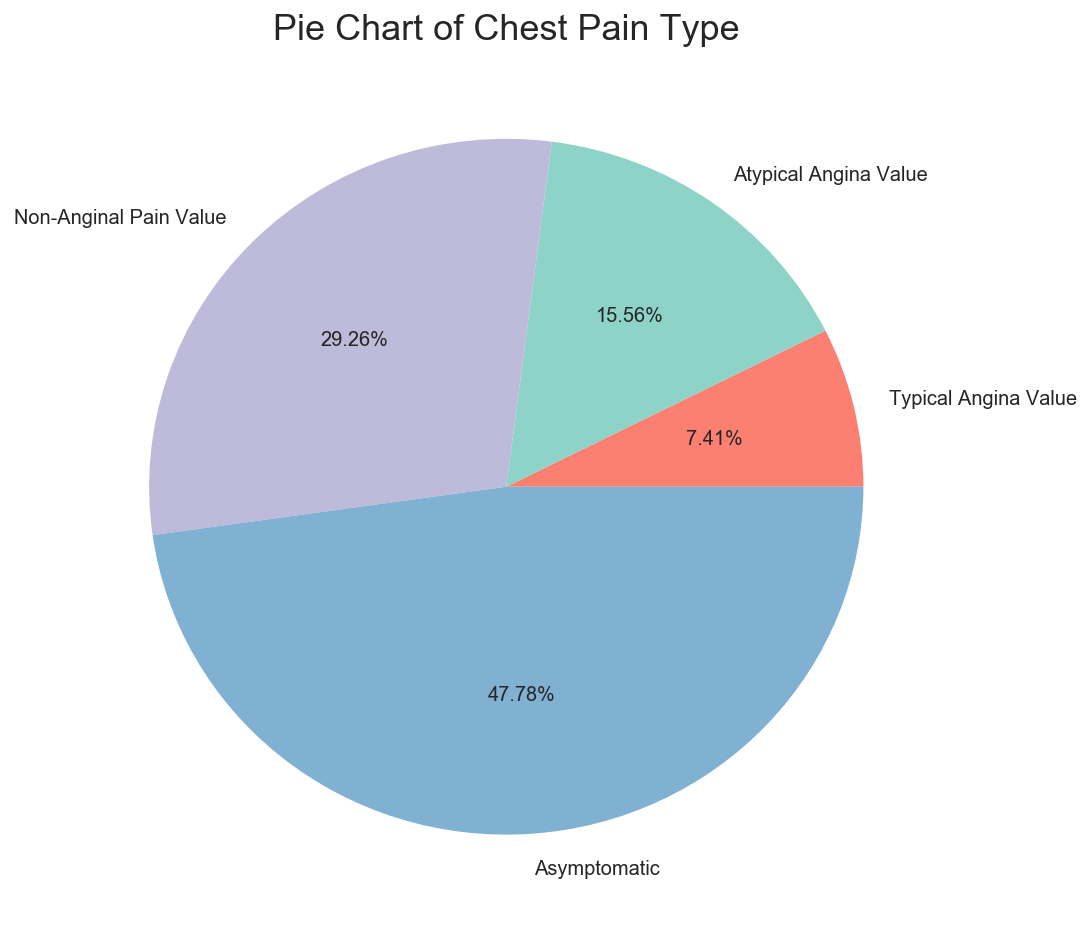
This is categorical variable. We have 4 different types of chest pain:

* Asymptomatic
* Non-Anginal Pain Value
* Atypical Angina Value
* Typical Angina Value

Chest pain is important factor for heart disease. Chest pain can also be due to problems in lungs, ribs, oesophagus, muscles, or nerves. *Asymptomatic* means neither causing nor exhibiting symptoms of any disease.

Non-angina chest pain is of short duration, typically less than 30 minutes or less than 5 seconds. it can be relieved immediately on lying down. Typical angina pain is presence of substernal chest pain or discomfort that was provoked by exertion or emotional stress. (Gore 2010). Atypical angina pain can last up to days while typical angina pain lasts from 3-15 minutes.

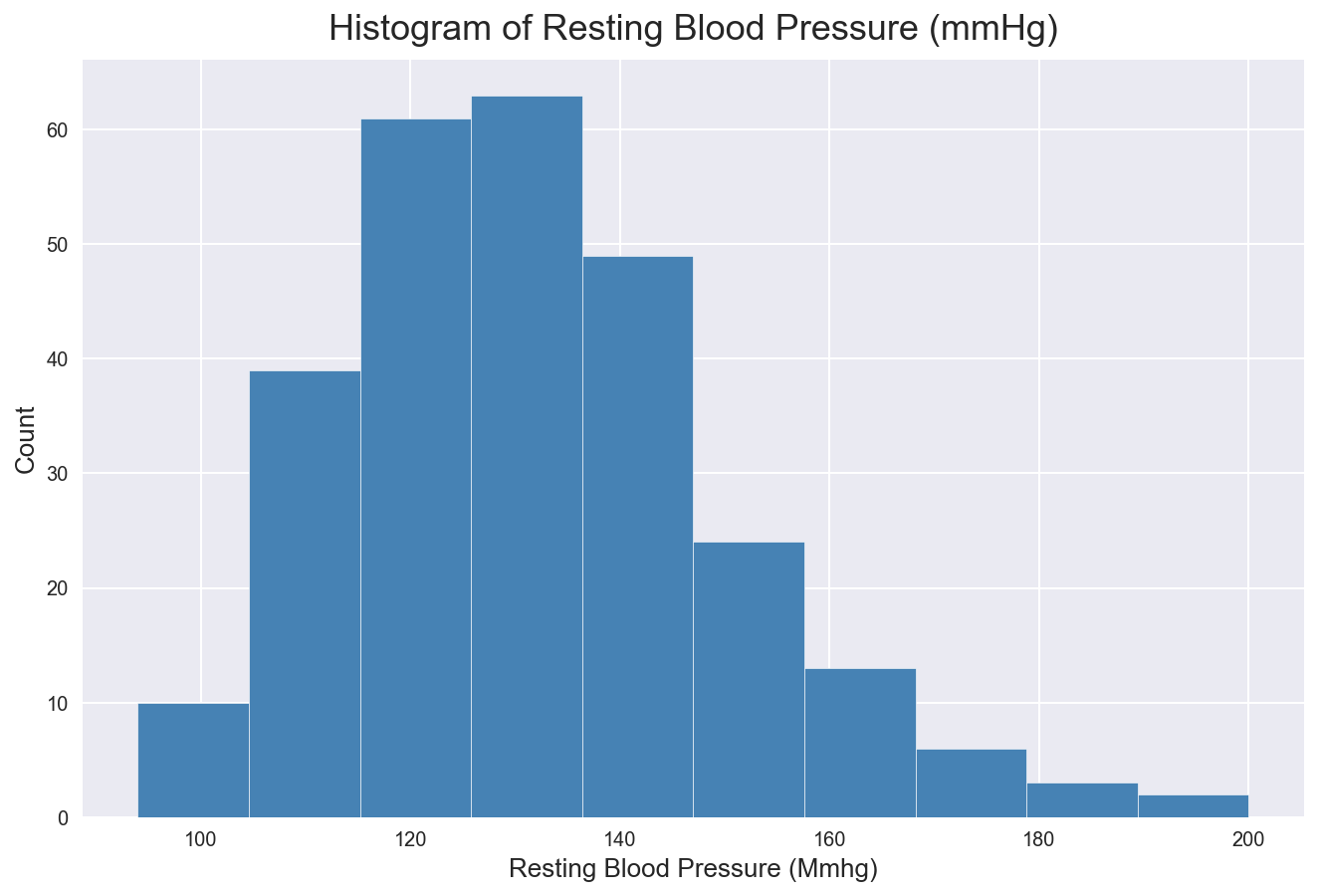
From the visualization, we can see that almost half samples have Asymptomatic chest pain. 30% have non-angina pain, while 15% have atypical angina pain.



### Resting Blood Pressure

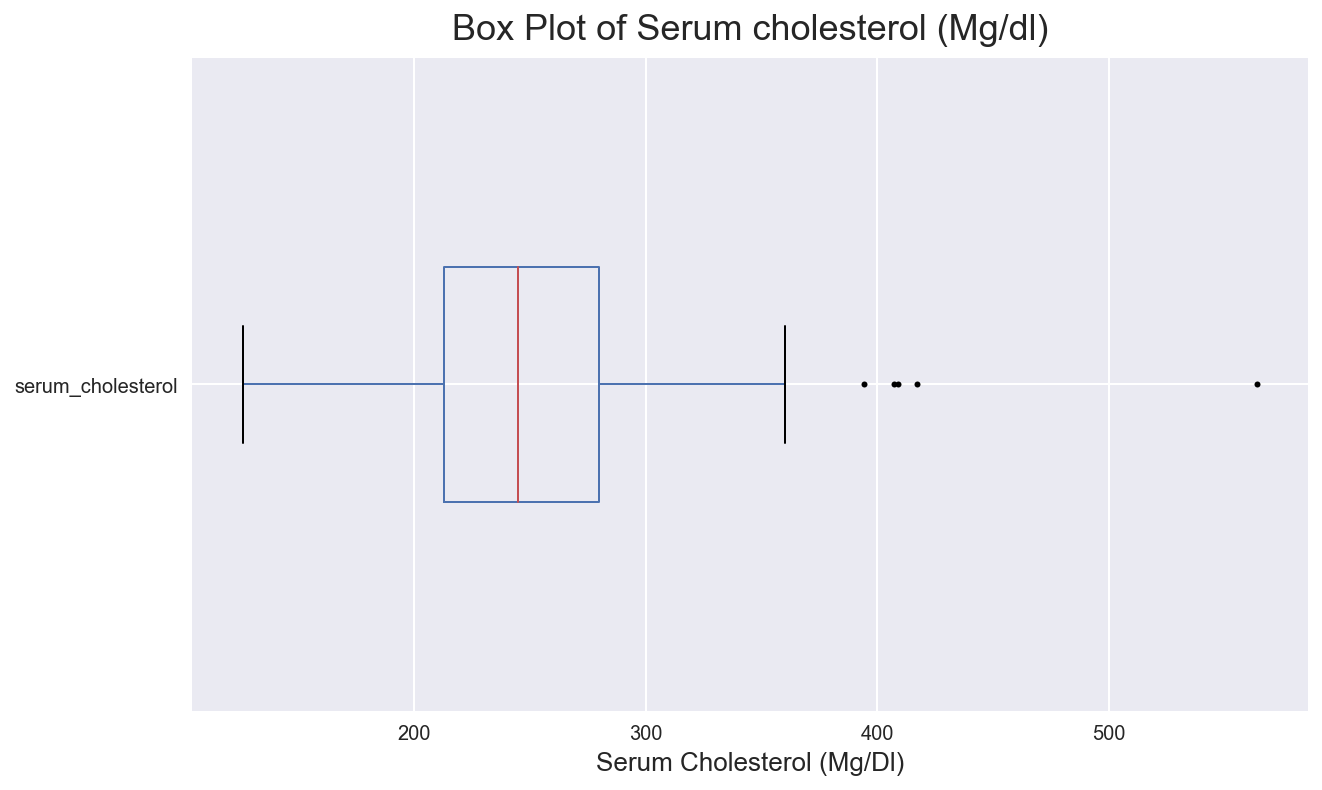
Blood pressure is the pressure of circulating blood in the walls of blood vessels. measured in millimetres of mercury (mmHg), above the surrounding atmospheric pressure. Normal resting blood pressure in an adult is around 120 systolic, and 80 diastolic.

Sample data has positive skewed distribution. Most of the pressure is between 120 and 160.



### Serum cholesterol

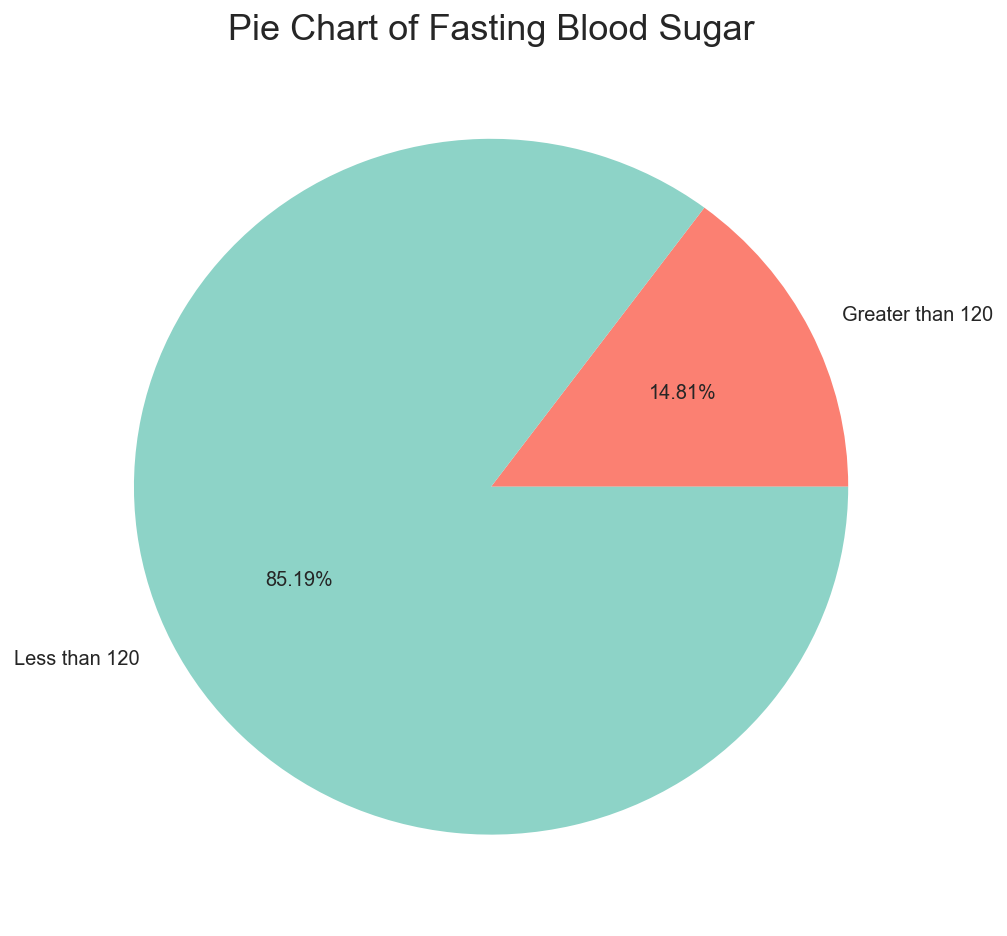
Cholesterol is a type of fat, also called as lipid. It travels through your bloodstream in molecules that can build up in arteries and restrict or block blood flow. *This is often associated with heart disease.* healthy serum cholesterol should be less than 200 mg/dl. Here in sample, most of the observation has high level of cholesterol, median is around 240 mg/dl. We have few outliers with more that 400 units.



### Fasting Blood Sugar

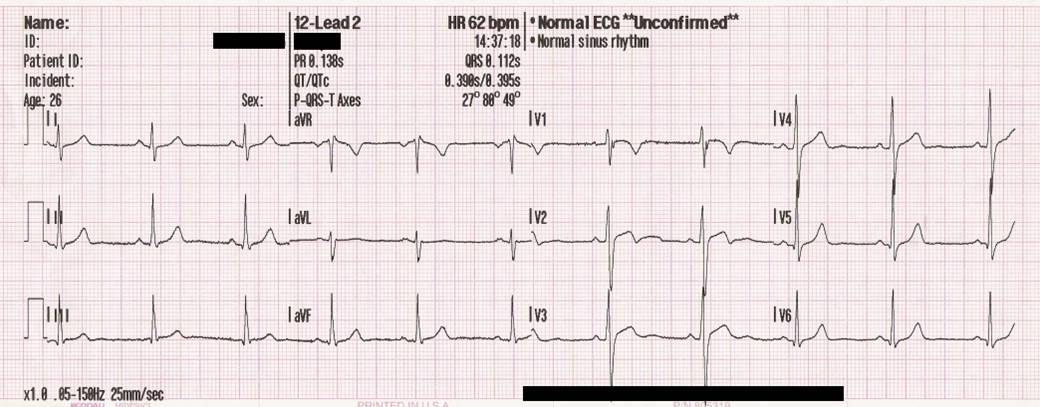
Blood sugar is the amount of glucose in blood. This is one of the major factor for Diabetes. Here the continuous variable is used as nominal variable. Fasting blood sugar level less than 100 mg/dL is normal. Up to 125 is considered as pre-diabetic.

Data is divided into greater than or less than 120. This will help up to identify if heart disease chances increase with blood sugar levels.



### Electrocardiograph

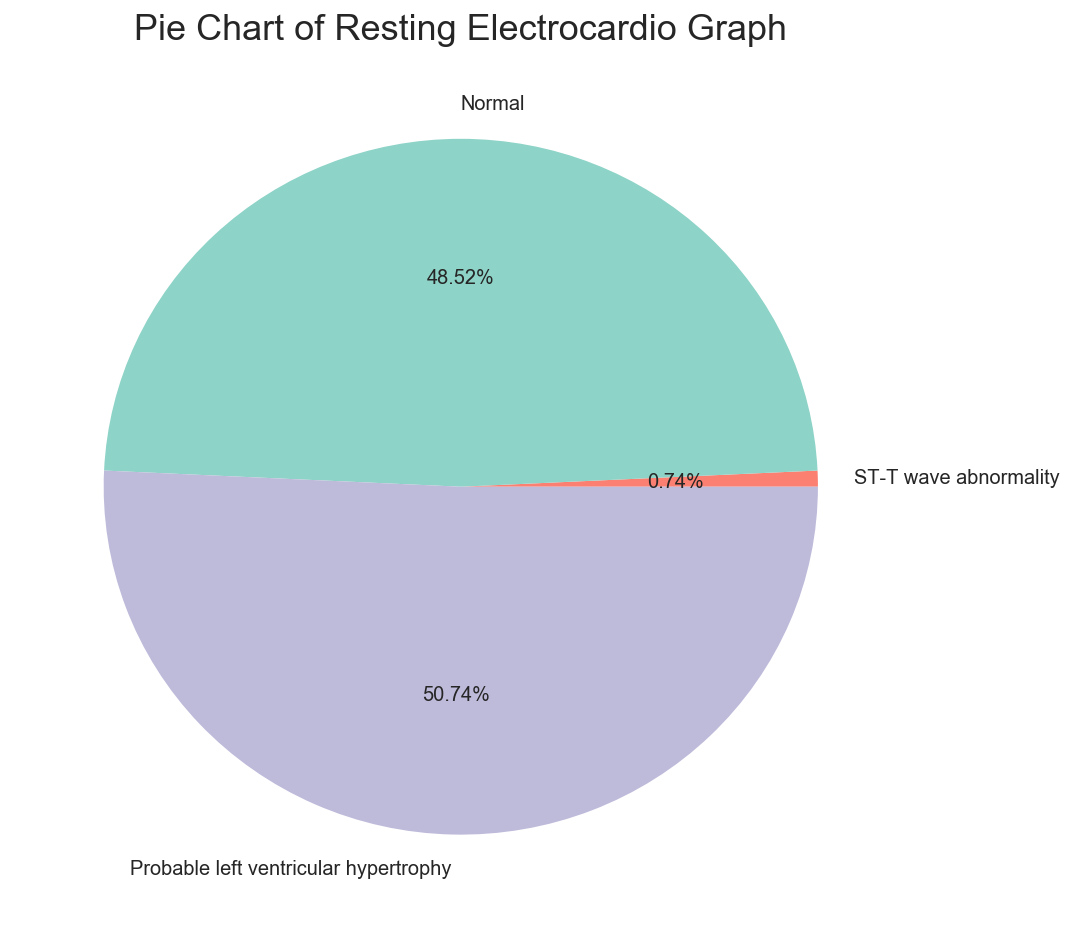
Electrocardiograph is recording of electrical activity of the heart over a period of time. It records tiny electrical changes on the skin that arise from the heart muscle's during each heartbeat.



*Figure 1 - Sample of ECG.*

*Source - Wikipedia*

ECG helps to detect heart disease. This can be important factor to identify presence of heart disease.



Our dataset has 3 types of ECG results

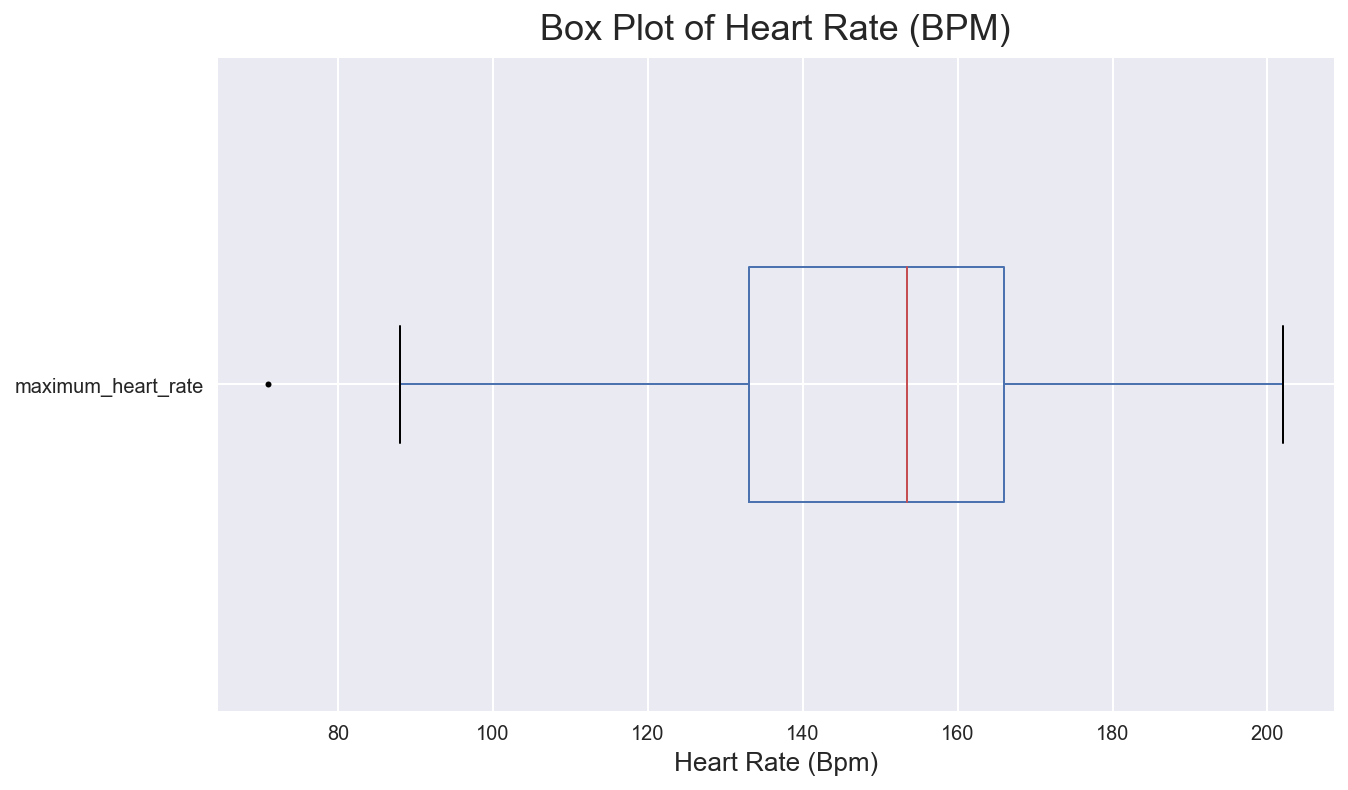
* Normal
* ST-T wave abnormality (Non-specific patterns)
* Probable left ventricular hypertrophy - it is enlargement and thickening of the walls of your heart's main pumping chamber

50% of sample has Probable left ventricular hypertrophy, 49% are normal while 1% have T wave abnormality in electro cardio graph results.

### Maximum Heart-Rate Achieved

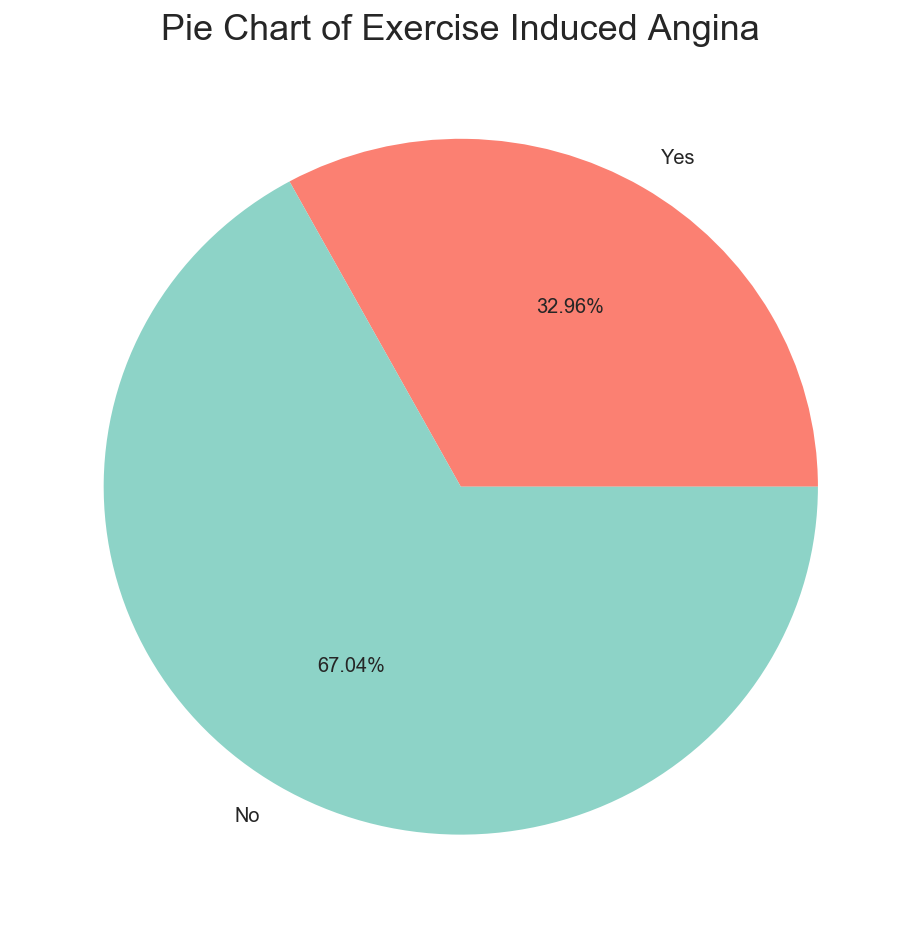
Heart rate is the number of times the heart beats per minute. Normal heart rate is between 60-100 bpm. Heart rate increases due physical activity. It is related to age too.

Most of the maximum heart rate in our sample lies between 130 and 170 bpm. There is one outlier with 70 bpm. Highly trained athletes can have a resting heart rate below 60 bpm.



### Exercise Induced Angina

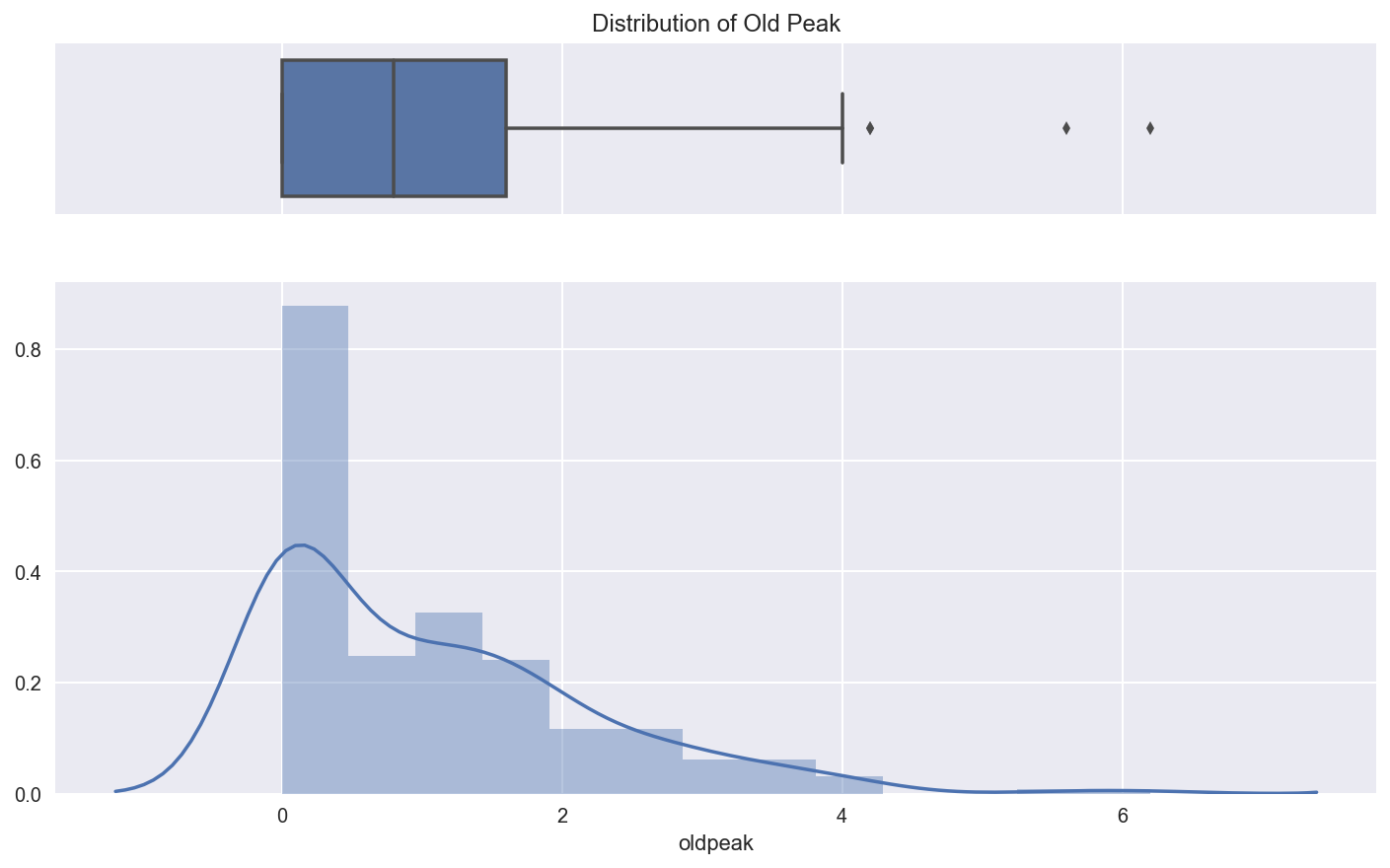
Angina is pain/discomfort that happens when your heart can't get enough blood and oxygen. (Heart Foundation 2018). Here 1/3 of patients has angina while 2/3 has not.



### Old Peak

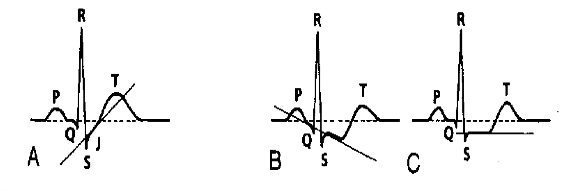
Here old peak is ST depression induced by exercise relative to rest. ST depression refers to a finding on an electrocardiogram, where the trace in the ST segment is abnormally low below the baseline. This is related with electro cardio graph. Old peak is significant if it is more than 1 mm (1 in our case)

In our dataset, most of the values are less than 2, values higher than that can be related to heart disease.



### Slope (ST Elevation)

Figure 2 explains the ST segment in electro cardio graph. ST slope refers to a finding on an electrocardiogram wherein the trace in the ST segment is abnormally high above the baseline. (Wikipedia 2018)

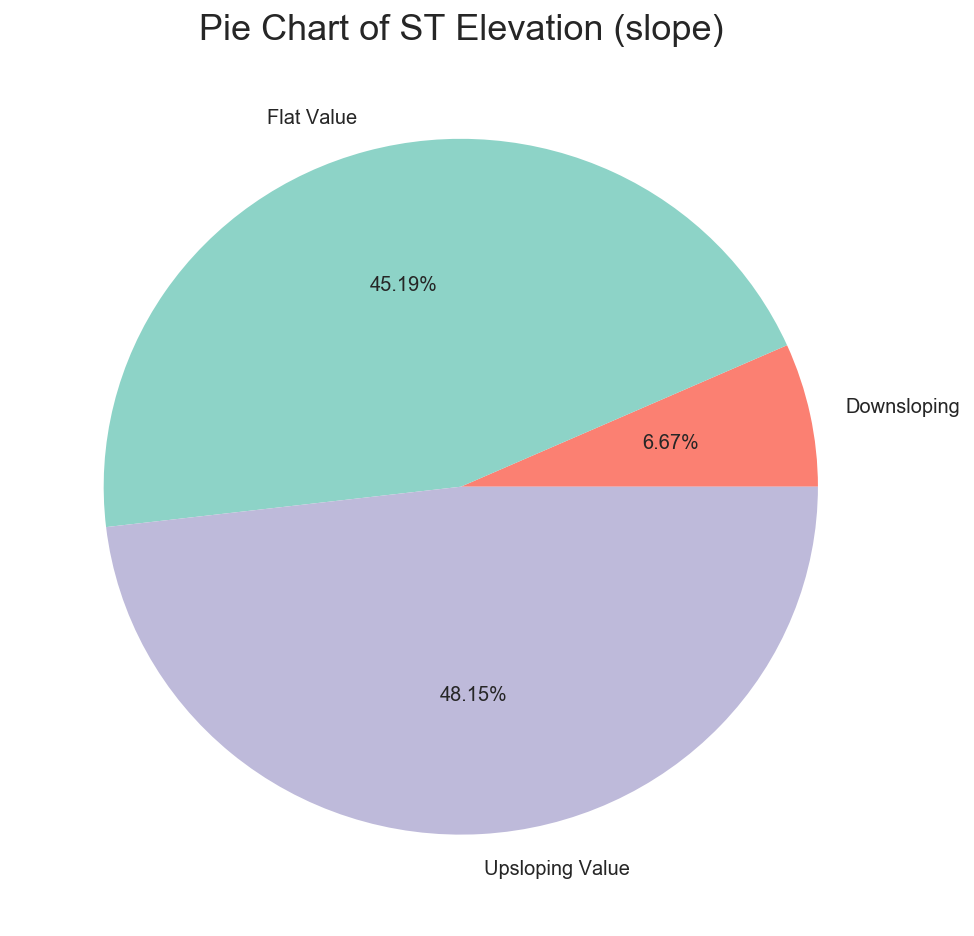


*Figure 2 – ST depression: upsloping (A), down sloping (B), horizontal (C).*

(Burns 2017)

* Down sloping ST depression indicates myocardial ischaemia.
* Upsloping ST depression is non-specific for myocardial ischaemia.

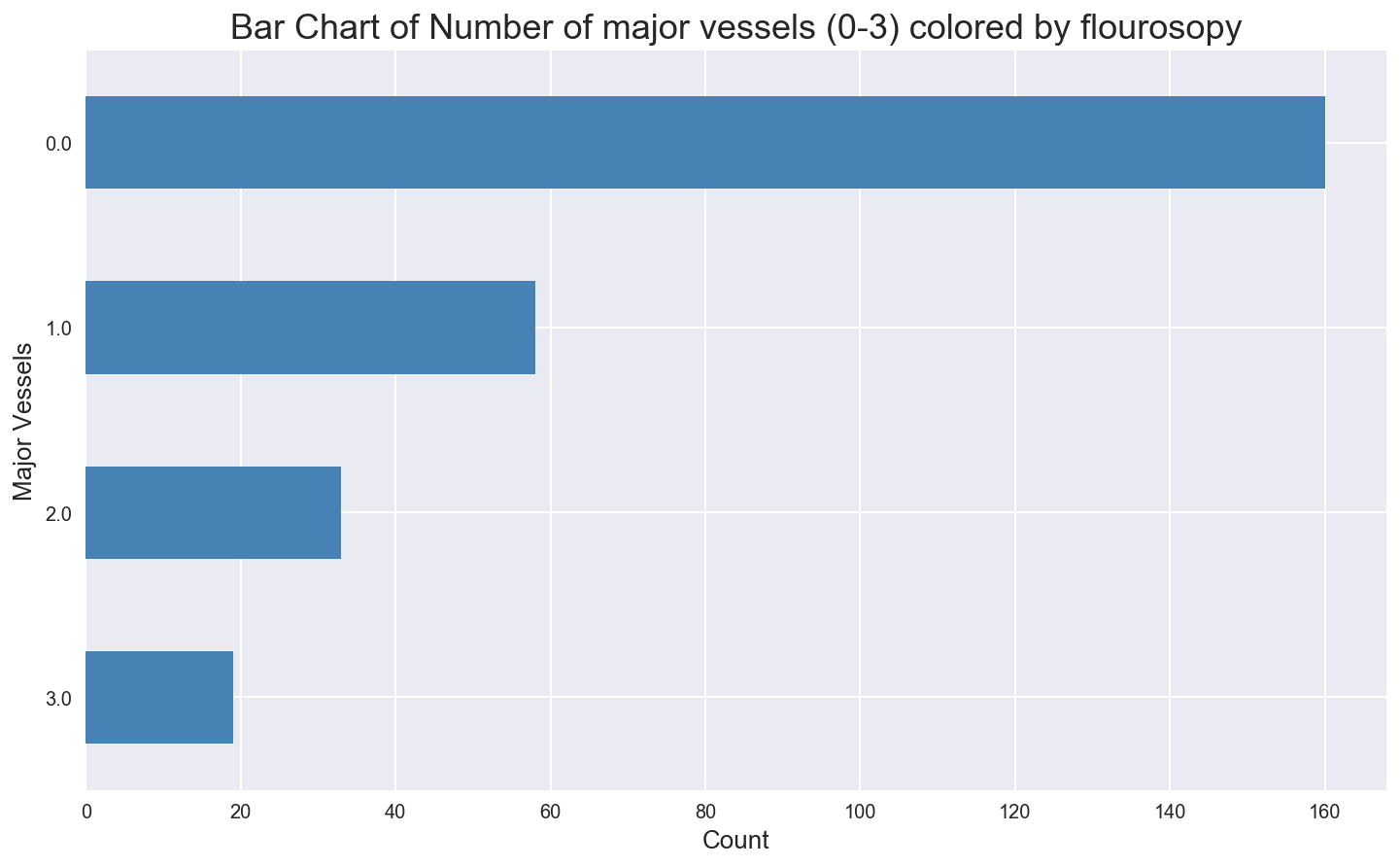
Pie chart of slope elevation has 45% flat, 48% upsloping while 7% as down sloping.



### Major vessels by fluoroscopy

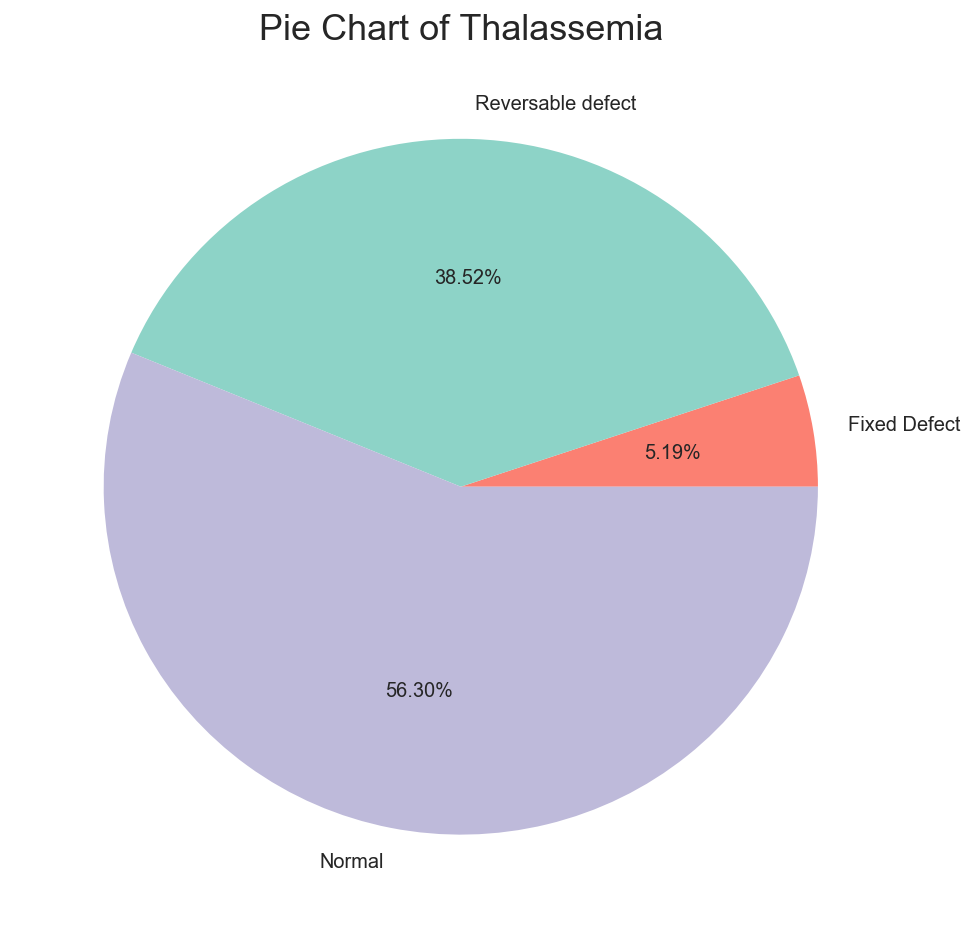
This is number of major vessels (0-3) coloured by fluoroscopy. Fluoroscopy, as an imaging tool, enables physicians to look at many body systems (John Hopkins Medicine 2018)

Our data sets have vessels as an ordinal variable. Mostly 0 major blood vessels are seen, followed by 1,2 and 3 using coloured fluoroscopy.



### Thalassemia

Thalassemia is an inherited blood disorder in which the body makes an abnormal form of haemoglobin (Healthline 2018)

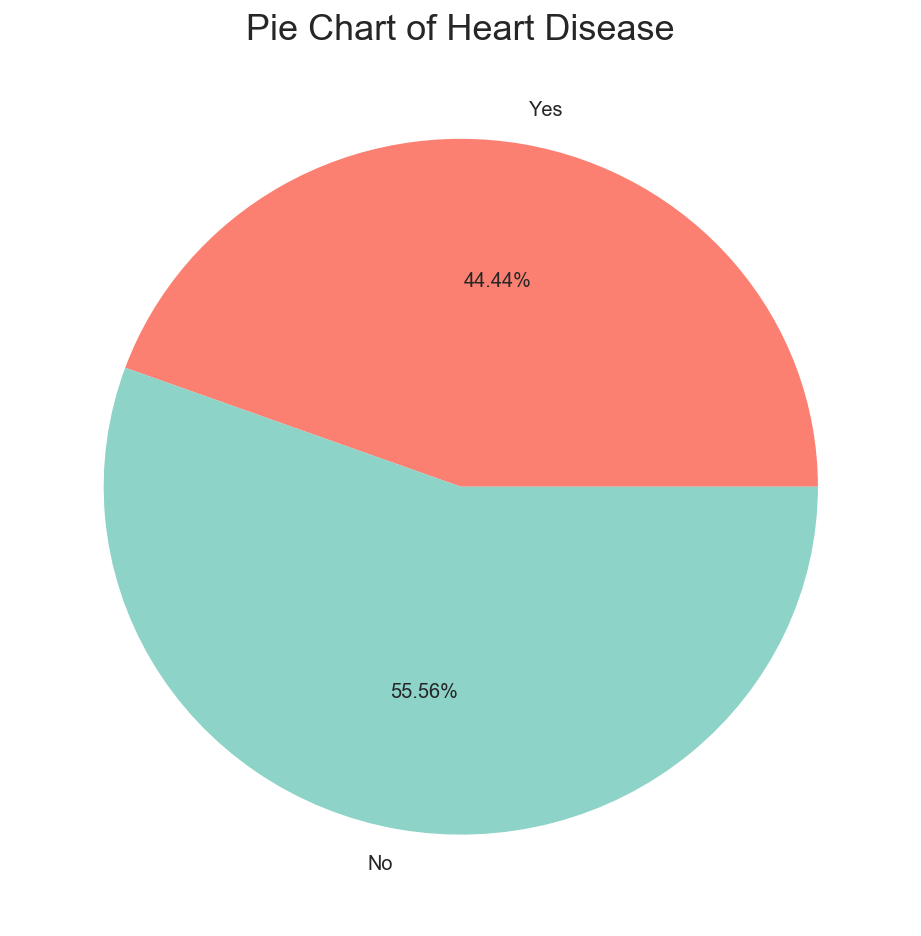


We have 3 types of variable for Thalassemia

1. 56% normal
2. 38% reversible defect
3. 5% fixed defect

### Heart Disease – Target Variable

This represents Absence (No) or presence (Yes) of heart disease. 56% sample has no heart disease while 44% have.



Next, we’ll analyse relationships between multiple variable.

Relationship Between variables  
From the univariate analysis, we understood the individual variables, variance and outliers. Most of the categorical variable are binary or factors up to 4.

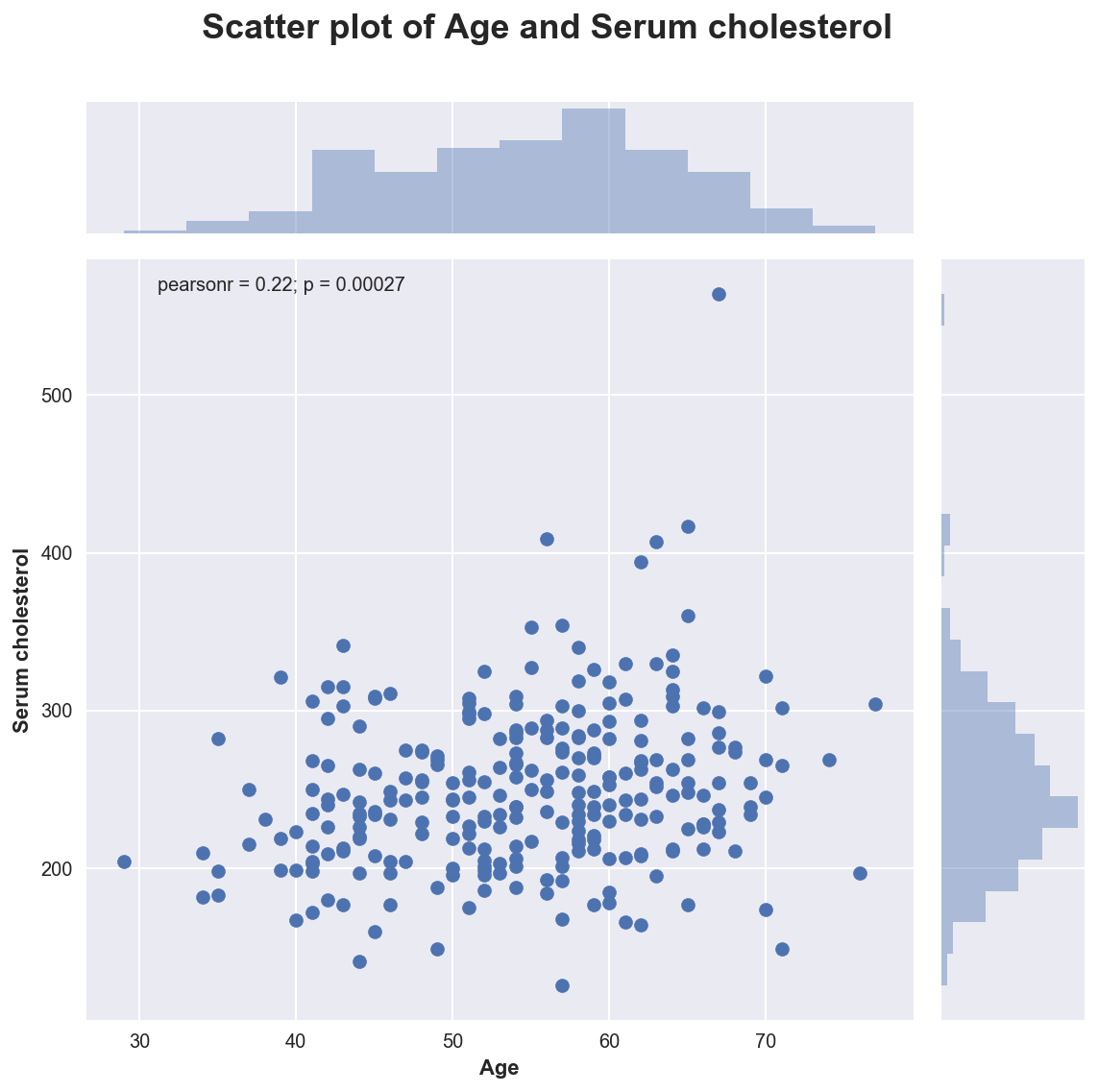
### Age and Resting Blood Pressure

There is small positive significant correlation between age and blood pressure. Pearson correlation is 0.27. We notice that people with higher age has slightly more resting BP than those with young age.



### Age and Serum Cholesterol

There is small positive significant correlation between age and blood pressure. This is similar to blood pressure, cholesterol is higher among old people.



Next, we can see relationship between blood pressure and heart rate by gender type.

### Heart Rate and Blood Pressure by Age

Here we notice:

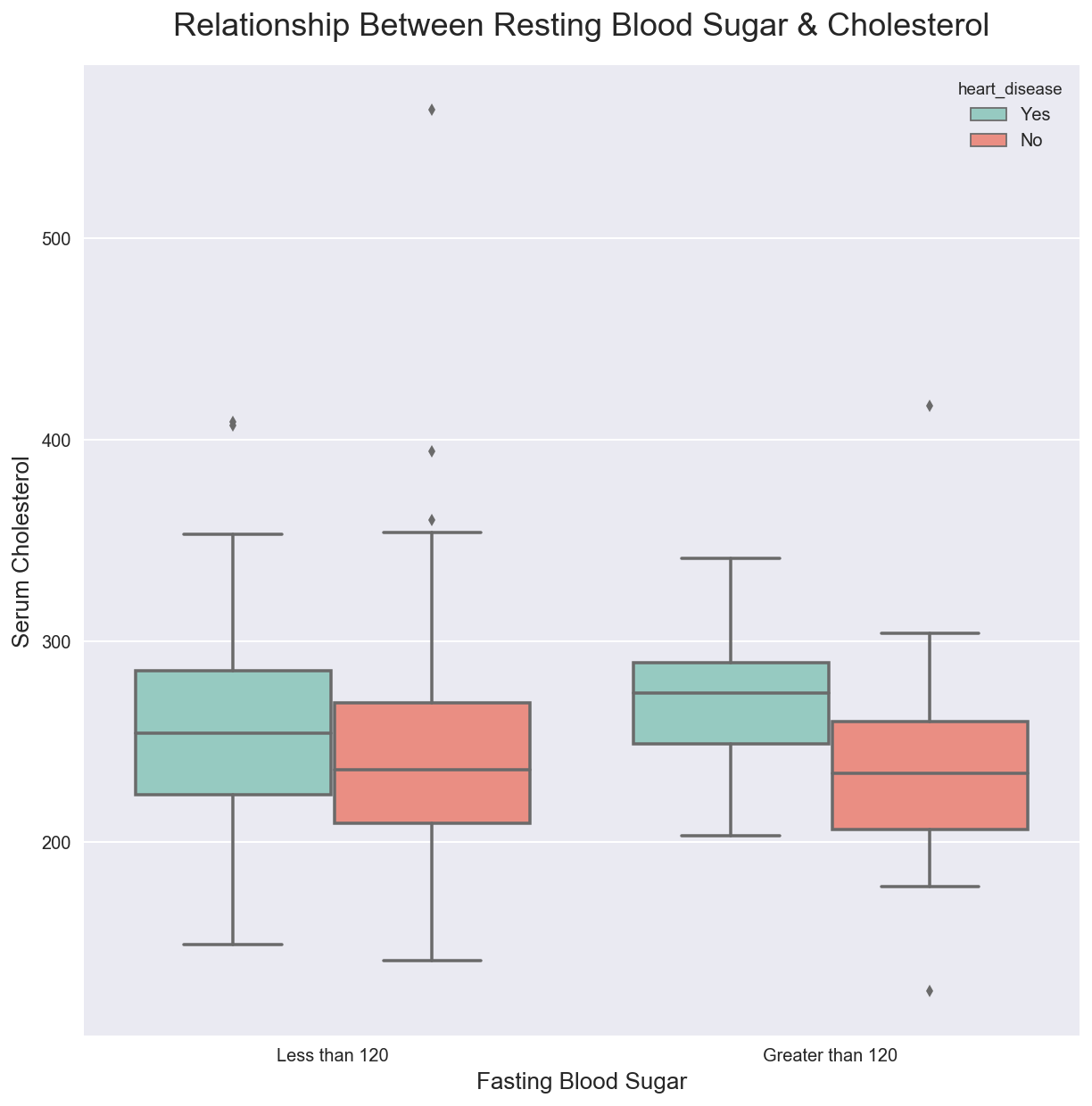
* Females generally have higher blood pressure than males
* Females with higher BP and heart rate have more tendency of heart disease
* Males even with normal BP and heart rate gets heart disease. There must be another factor for disease.



### Blood Sugar and Cholesterol

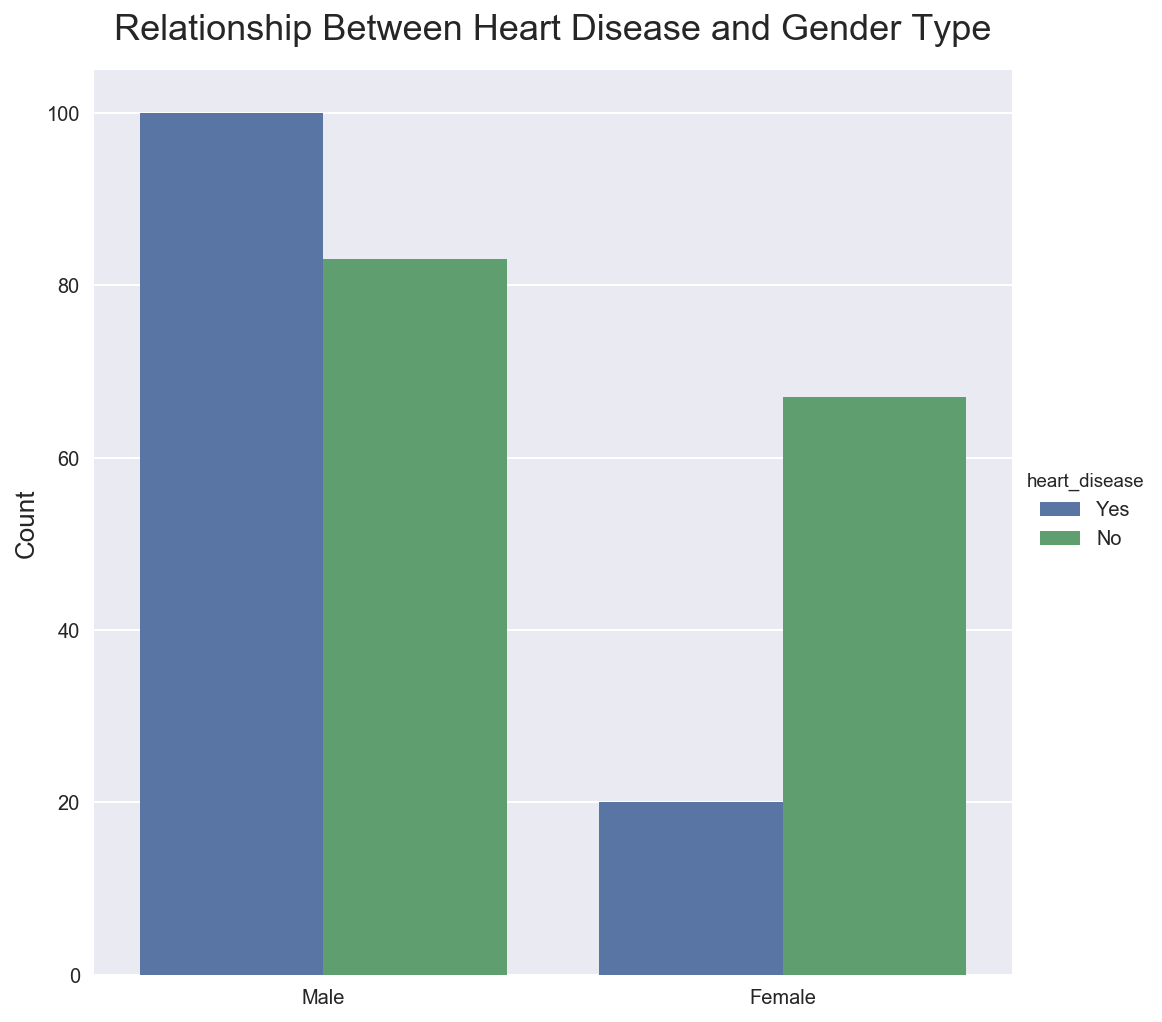
Here we notice:

* People with higher cholesterol have higher chances of heart disease
* Having blood sugar increase the chances even more.
* People having no heart disease have almost similar median value for cholesterol
* These factors are useful in identifying heart disease



### Gender type and heart disease

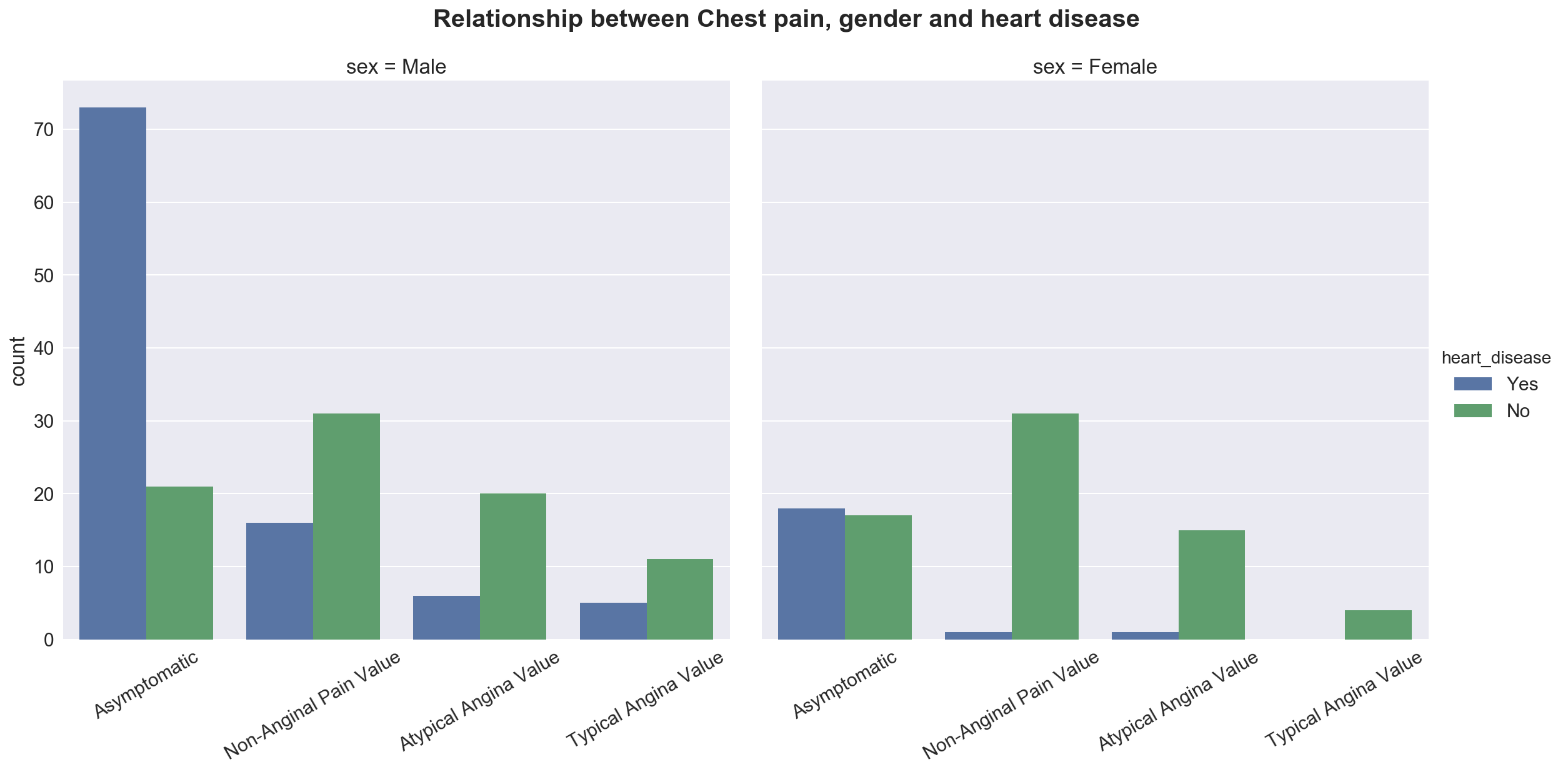
Male count with heart disease is 4 times than female. This shows an interesting result than chances of getting heart disease for males are higher.



### Chest pain with gender

*Asymptomatic chest pain* means neither causing nor exhibiting symptoms of any disease.

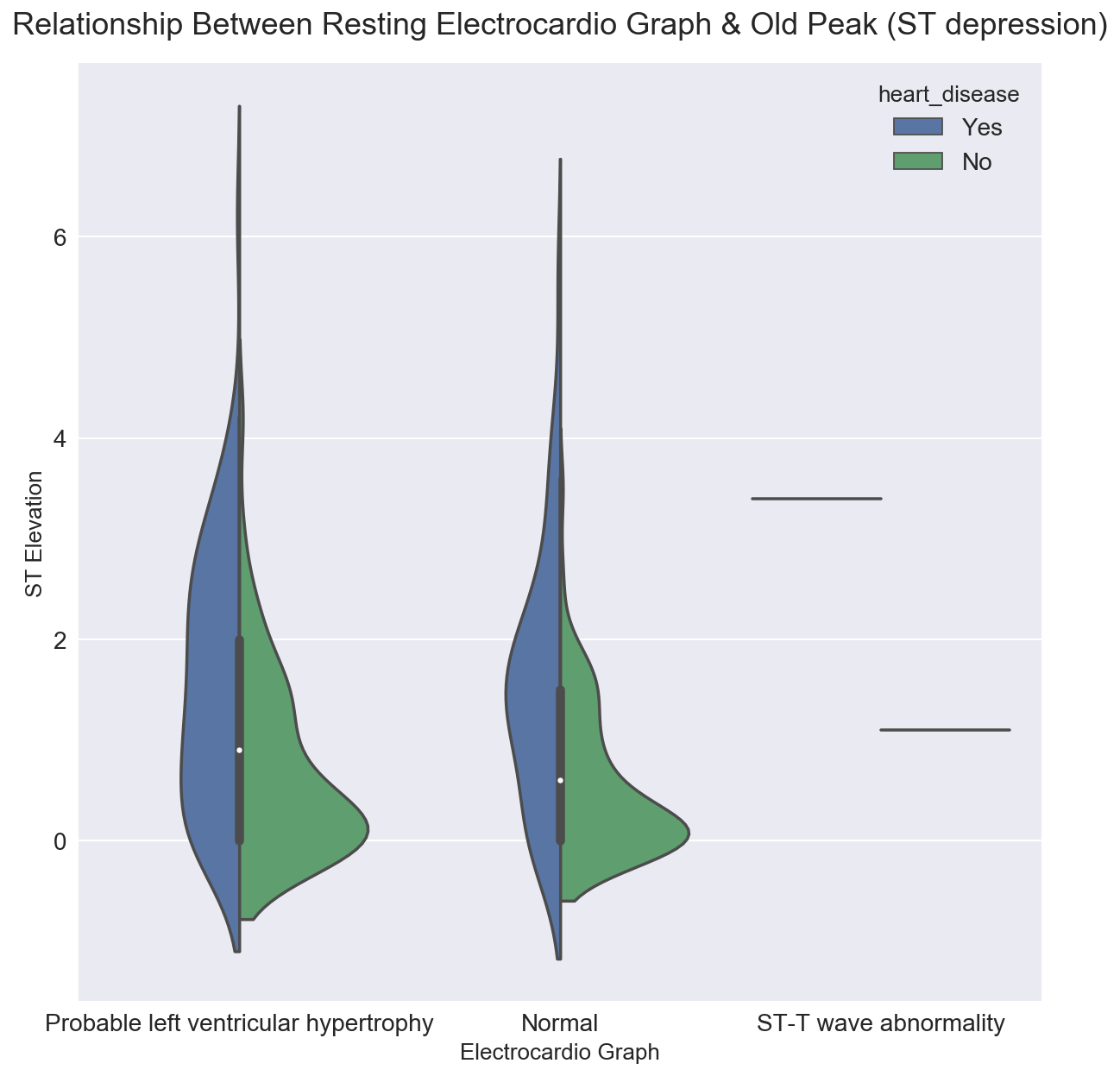
* Here chest pain *Asymptomatic* indicates that it’s not highly related with heart disease
* Other types of chest pain are not significant for both males and females



### Electro cardio graph and oldpeak

*From the univariate analysis we have seen that old peak (ST depression) is part of electro cardio graph*

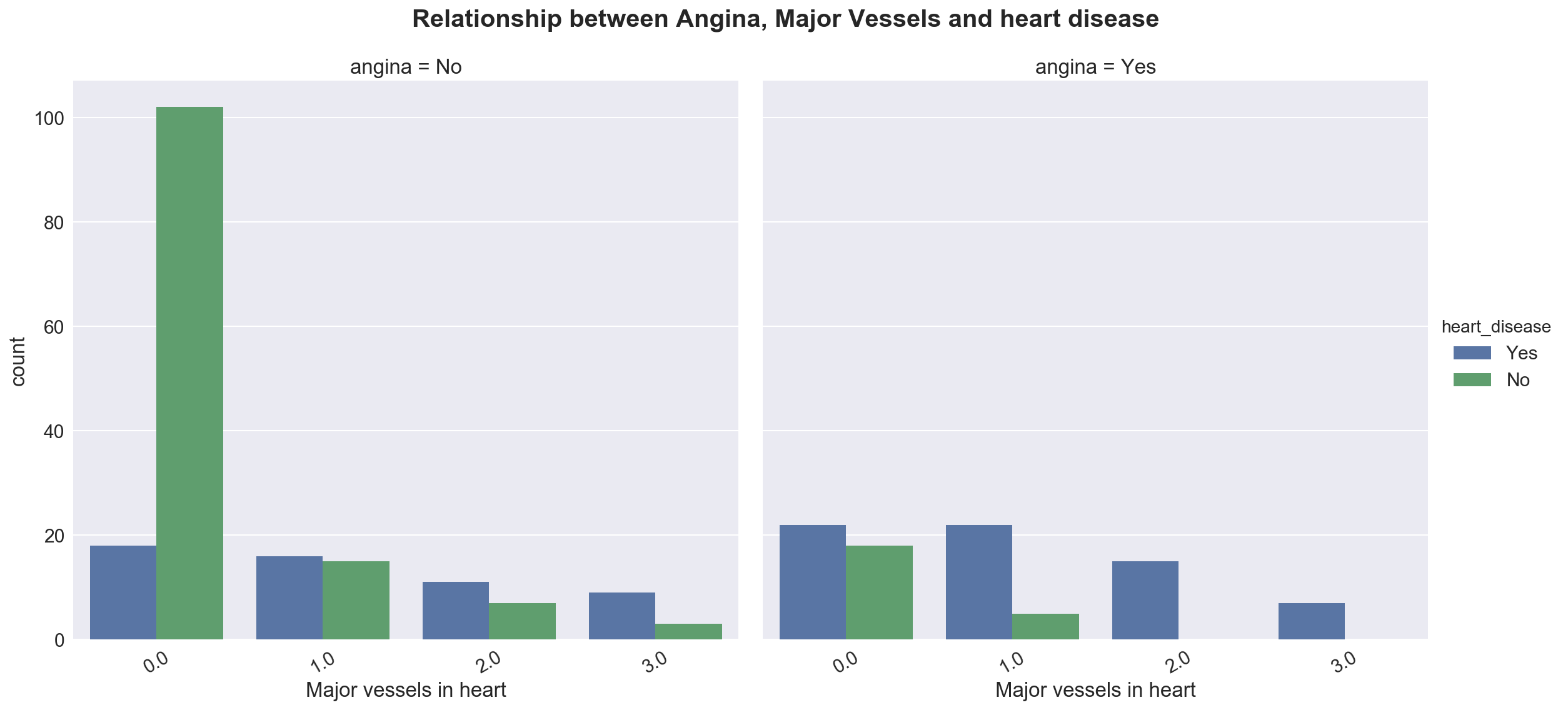
* People with high ST elevation has higher old peak value
* Distribution is around zero for absence of heart disease



### Angina with Major vessel

*Angina is the heart pain and major vessel by fluoroscopy is related with heart*.

* Distribution of heart disease is similar for all major vessels seen by fluoroscopy
* Samples with angina heart pain has high chances of heart disease

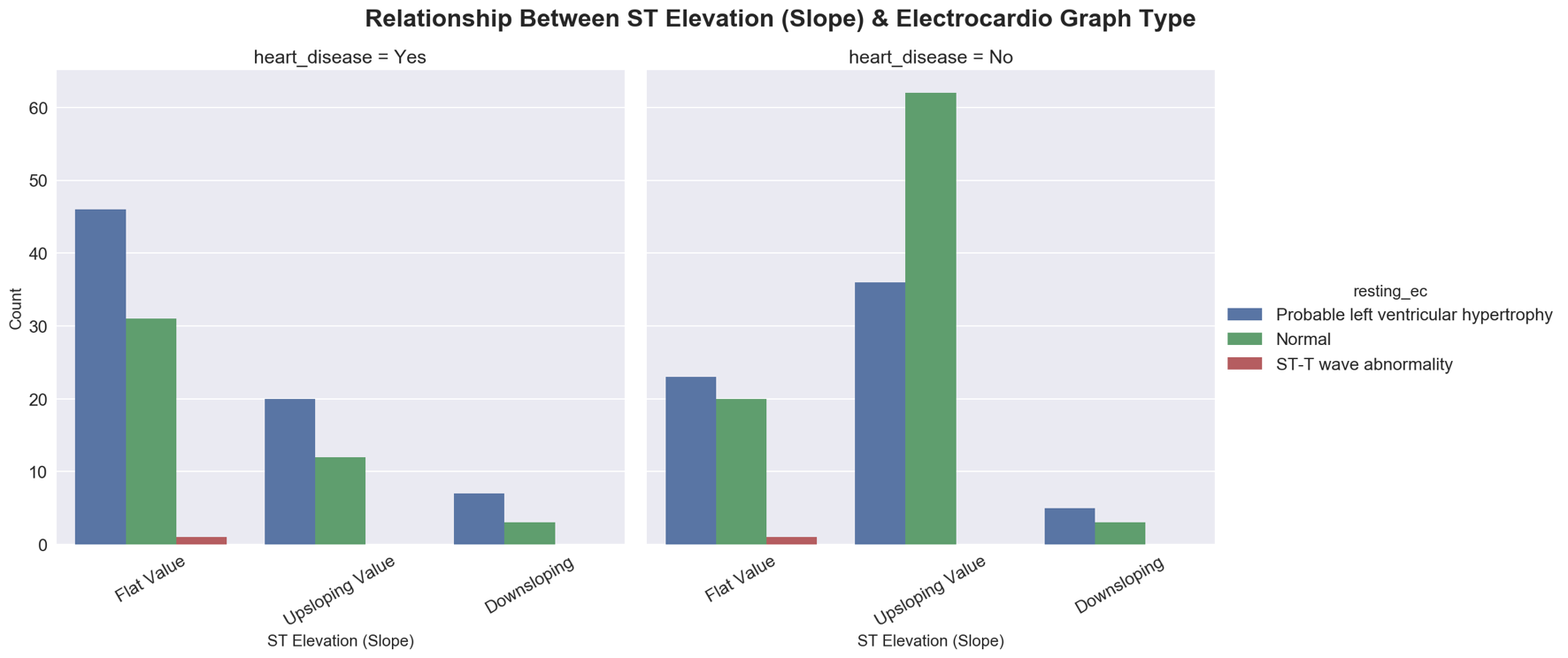


### ST Elevation with EC type

*Both variables are related with electro cardio graph results*

* Here we are observing an interesting pattern. Electro cardio graph results is different for presence and absence of heart disease
* Samples with flat value has high chances of having heart disease
* Normal EC results will not have heart disease in most of the cases
* Upsloping value will have less chances of heart disease

Electro cardio graph reveals a good prediction for heart disease.

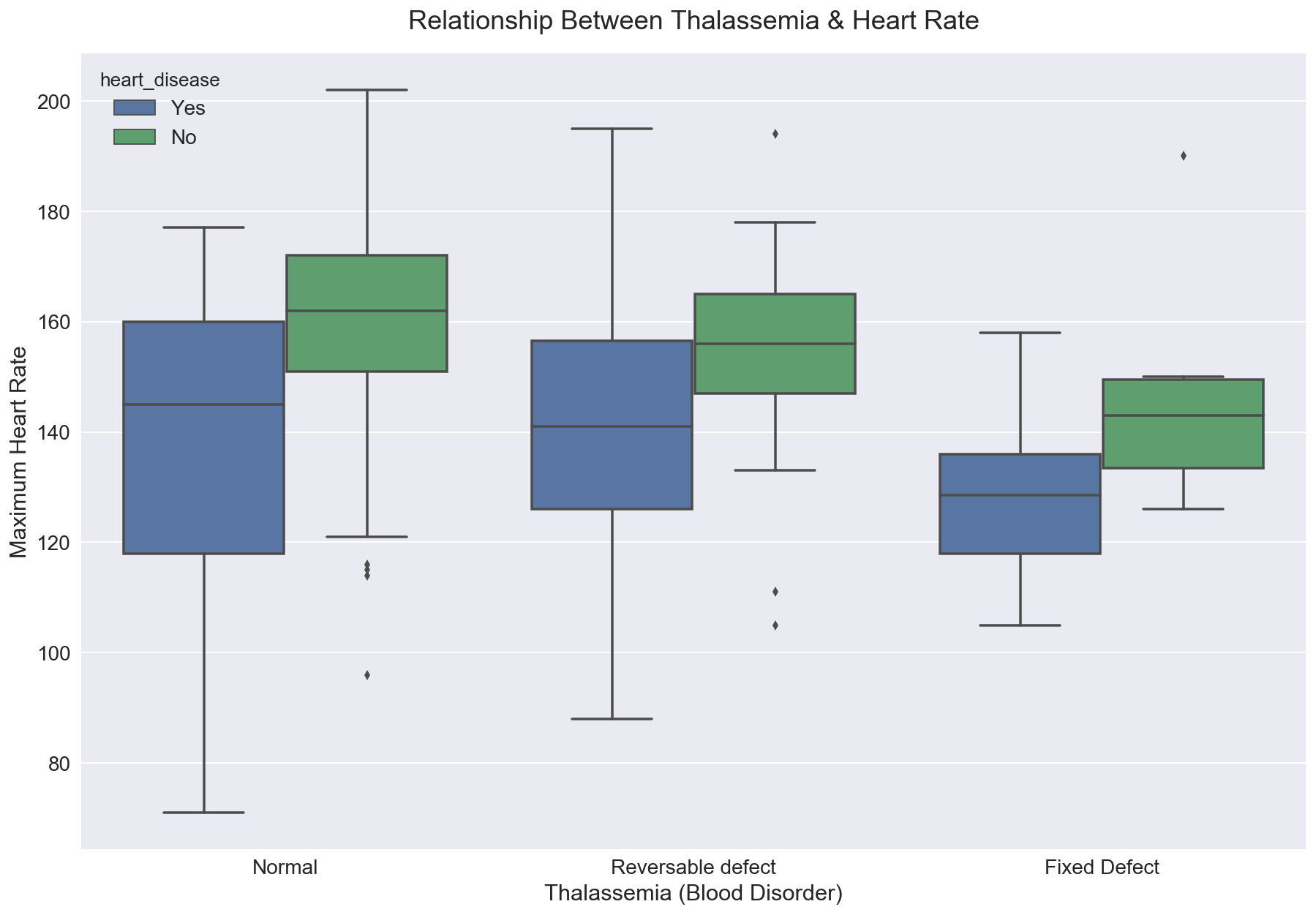


### Thalassemia and Heart Rate

*Thalassemia is a heart disorder*

* Thal types are not proving enough evidence for heart disease
* Mostly sample with lower heart rate has heart disease
* Chances are, people with high heart rate are athlete and not a good indication for heart disease

Electro cardio graph reveals a good prediction for heart disease.



# Discussion

# Conclusion

# Bibliography

Burns, Edward. 2017. *Myocardial Ischaemia.* 3 4. https://lifeinthefastlane.com/ecg-library/myocardial-ischaemia/.

Gore, Joel M. 2010. *Typical Angina vs. Atypical Chest Pain.* 1 June. https://www.jwatch.org/jc201007070000002/2010/07/07/typical-angina-vs-atypical-chest-pain.

Healthline. 2018. *thalassemia.* https://www.healthline.com/health/thalassemia.

Heart Foundation. 2018. *Angina Facts.* https://www.heartfoundation.org.au/your-heart/heart-conditions/angina.

John Hopkins Medicine. 2018. *Fluoroscopy Procedure.* https://www.hopkinsmedicine.org/healthlibrary/test\_procedures/orthopaedic/fluoroscopy\_procedure\_92,P07662.

2004. *Statlog (Heart) Data Set.* http://archive.ics.uci.edu/ml/datasets/statlog+(heart).

Wikipedia. 2018. *ST Elevation.* https://en.wikipedia.org/wiki/ST\_elevation.