Assignment 2
Heart Disease Analysis

21st May 2018

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

Table of Contents

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. We selected random as the best model for heart disease

Introduction

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

The model evaluates cars according to the following concept structure:

1.Age

2.Sex

3.Chest Pain type

4.Resting Blood Pressure

5.Serum Cholesterol

6.Fasting Blood sugar

7.Resting ECG results

8.Max Heart rate

9.angina

10.Oldpeak

11.Slope

12. No. of Major vessels

13.Thal

14.heart-Disease

This report will narrate how the data is read, modelled using random forest, k nearest neighbours and decision tree algorithms. Since the dataset is free from errors, missing values and does not need any sanitary checks, it is directly taken into modelling and visualising stage. This report will also show which of the three machine learning models is best suited for the Heart dataset.

Methodology

Pandas feature in python is the main tool for analysis performed with the selected dataset. Pandas is a software library in python and is used in data manipulation and data analysis at a large scale.

Pandas offers various features which includes data structures and operations for manipulating numerical tables and even time-series data. The Pandas library features include Dataframe, which is an object used for manipulation, tools for reading, writing data between different file formats, Data alignments, Data reshaping, indexing, subsetting and lots more.

NumPy is another library for the python programming language, which is used to handle large multidimensional arrays and matrices. It also has an enormous collection of high-end mathematical functions to handle these arrays. The heart of NumPy's functionality lies in the concept of the "Nd-array", also known as n dimensional array.

Scikit learn is a library in python which features various types of classification, clustering and regressions techniques including random forests, gradient boosting and many more. Finally, the analysed data is pictured and visualised using a feature in python called as matplotlib. It is the plotting library in Python programming language.

The whole project starts with loading the precleaned data into a pandas data frame. It is then separated into two datasets (training set and test set) and fed into the three machine-learning models. The modelled data is then cross-validated and compared with the precision value obtained from the modelled data. If the variance between the cross-validated value and the precision value is negligible then the model is considered good. However, if the variance is high then the process is repeated to finetune and model the data again, till the variance is negligible. Finally, the precision value from the three modelled data is compared and the modelled data with the highest precision value is awarded the perfect model for the "Heart" dataset.

Results :

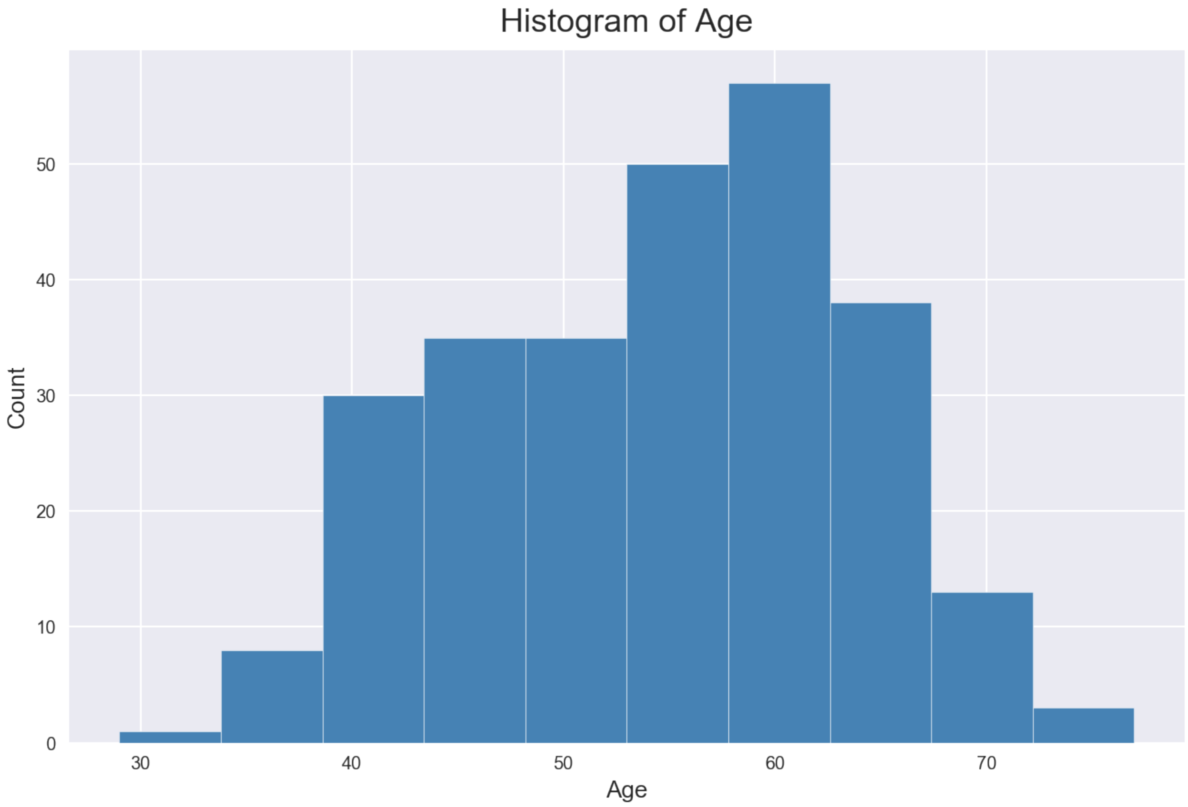
Data Exploration

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

Data has no missing values.

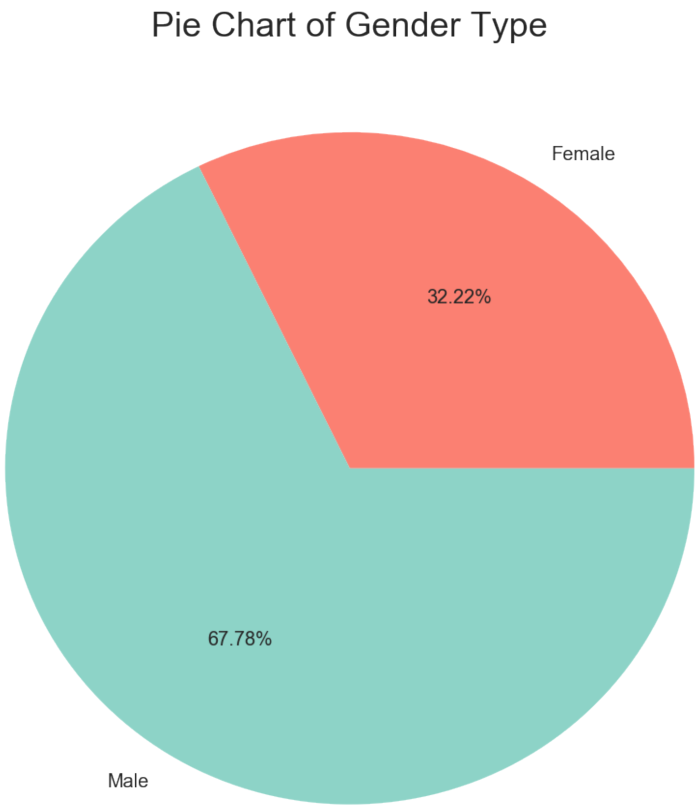
Feature Visualization – Age

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

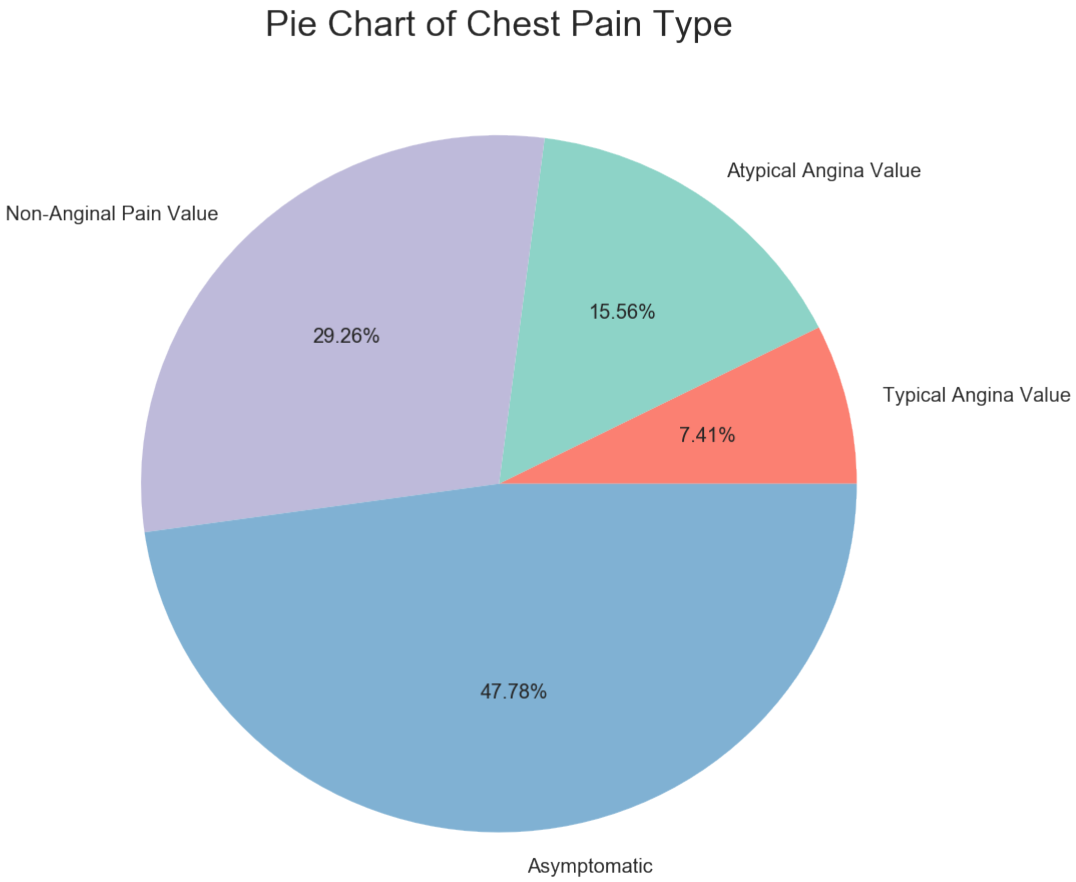
This is categorical variable. We have 4 different types of chest pain:

- 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.

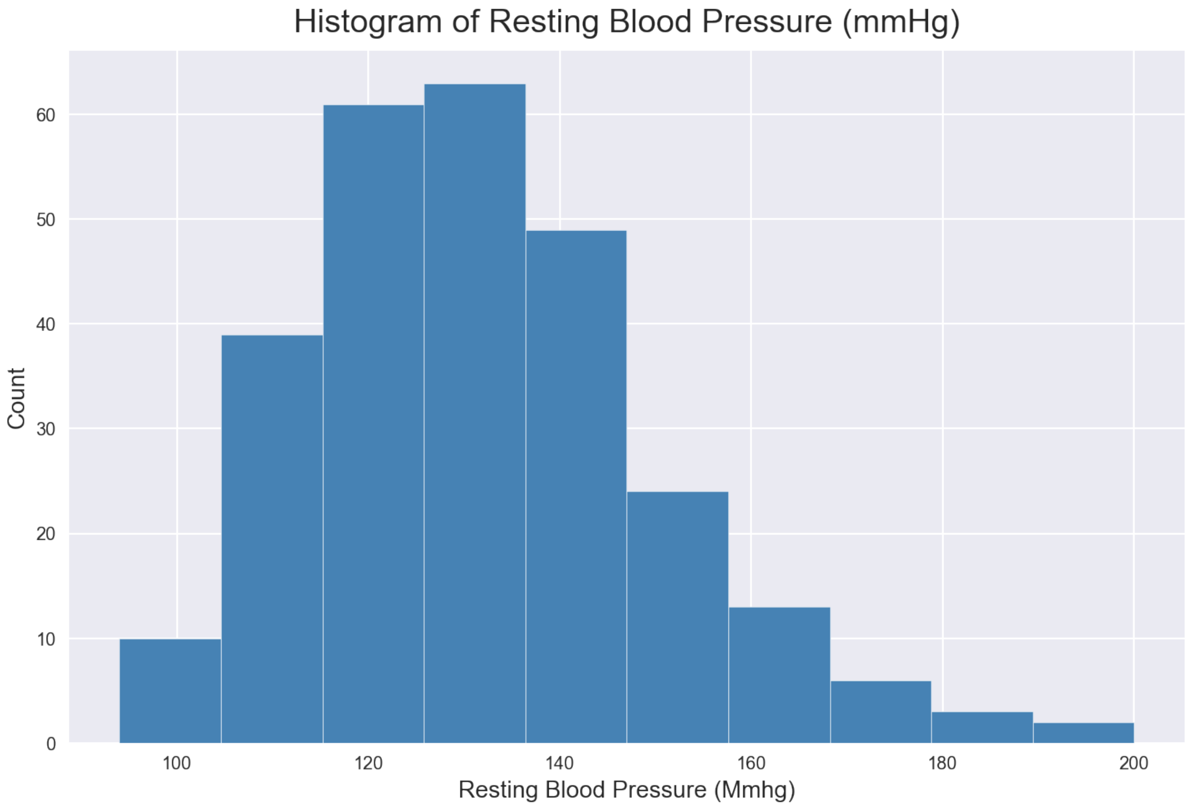
From the visualization, we can see that almost half samples have Asymptomatic chest pain. 30% have non-angina pain, while 15% have atypical angina pain.



Resting Blood Pressure

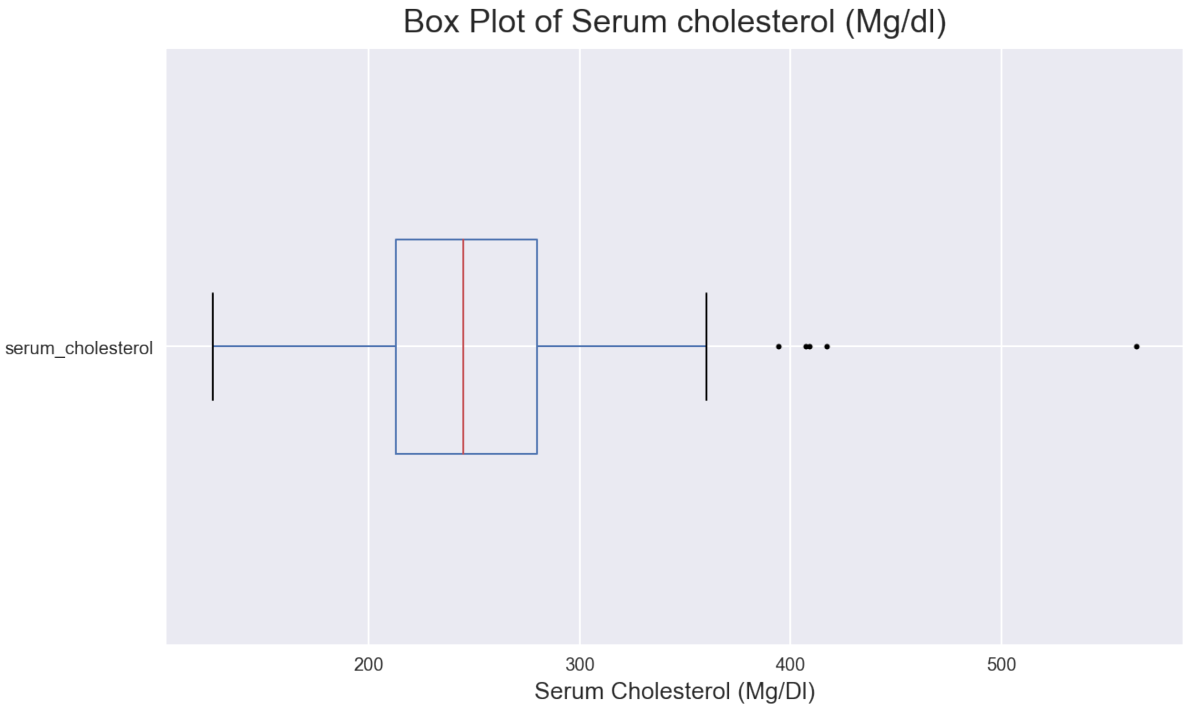
Blood pressure is the pressure of circulating blood in the walls of blood vessels. measured in millimetres of mercury (mmHg), above the surrounding atmospheric pressure. Normal resting blood pressure in an adult is around 120 systolic, and 80 diastolic.

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



Serum cholesterol

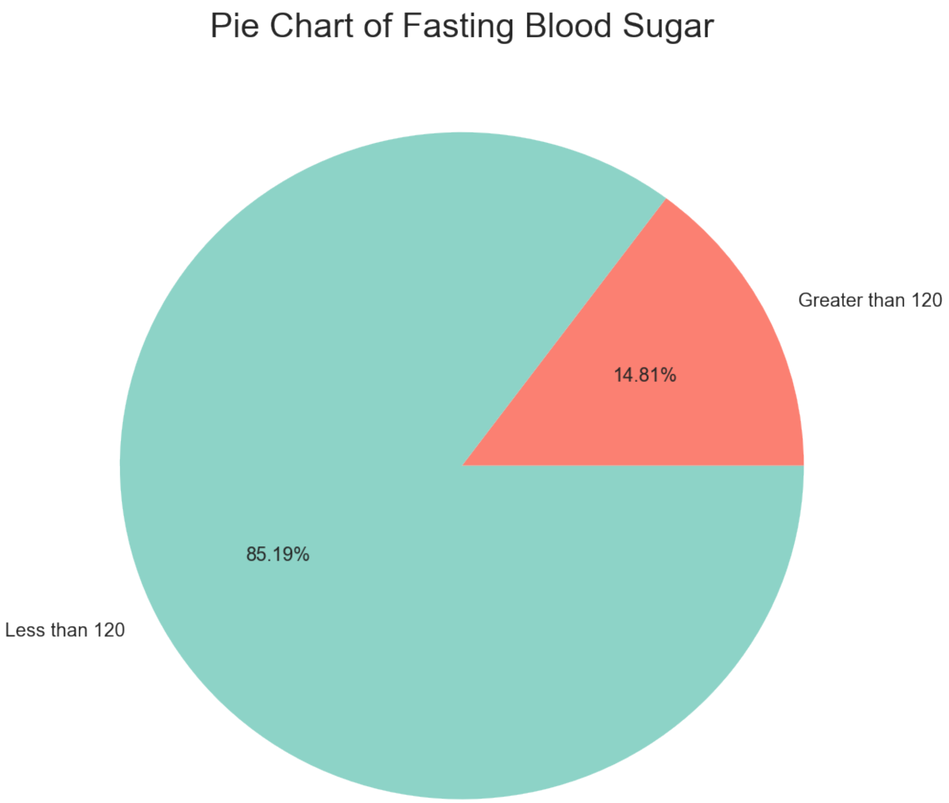
Cholesterol is a type of fat, also called as lipid. It travels through your bloodstream in molecules that can build up in arteries and restrict or block blood flow. *This is often associated with heart disease.* healthy serum cholesterol should be less than 200 mg/dl. Here in sample, most of the observation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

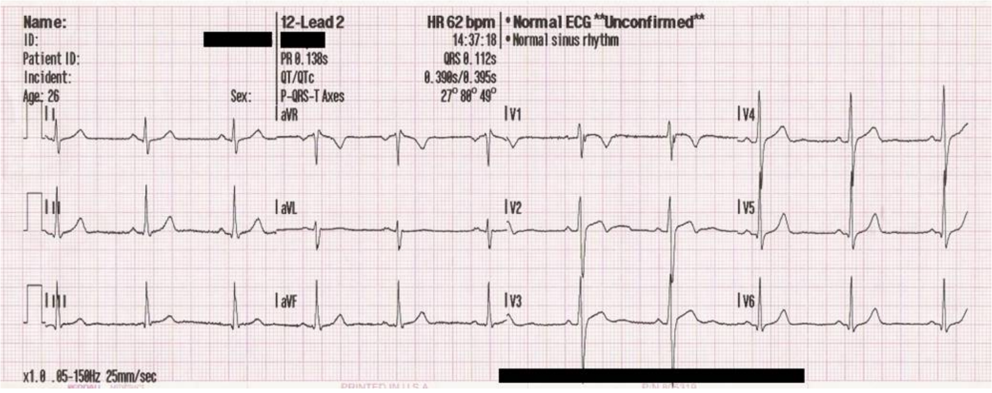
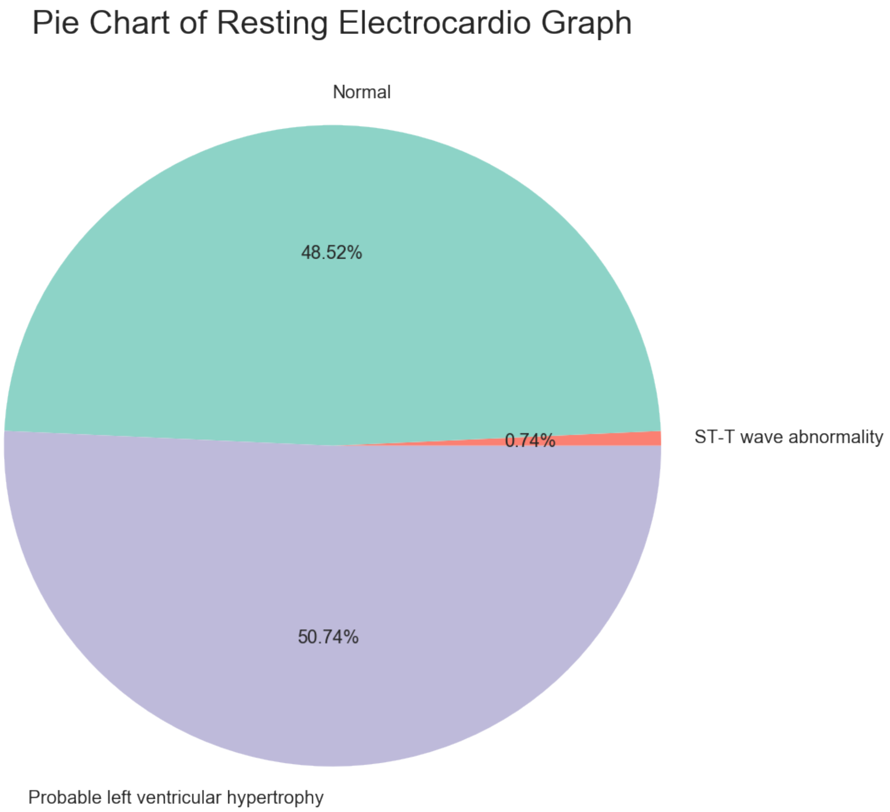


Figure 1 - Sample of ECG.
Source - Wikipedia

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

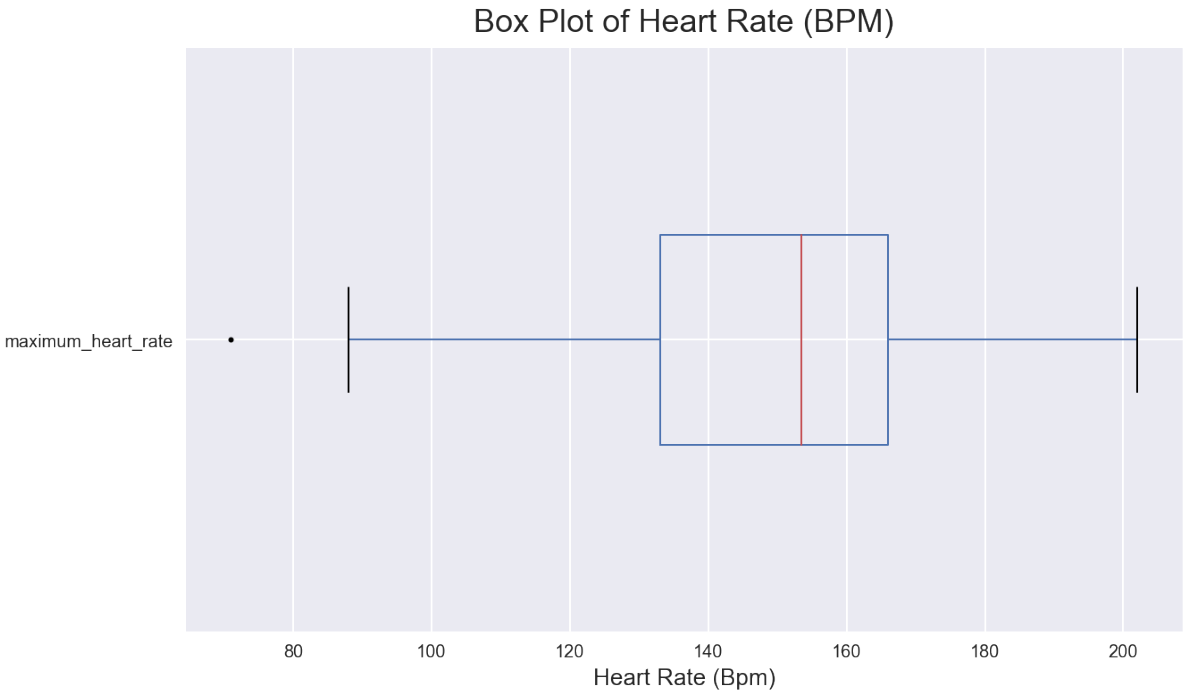
- Normal
- ST-T wave abnormality (Non-specific patterns)
- Probable left ventricular hypertrophy - it is enlargement and thickening of the walls of your heart's main pumping chamber

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

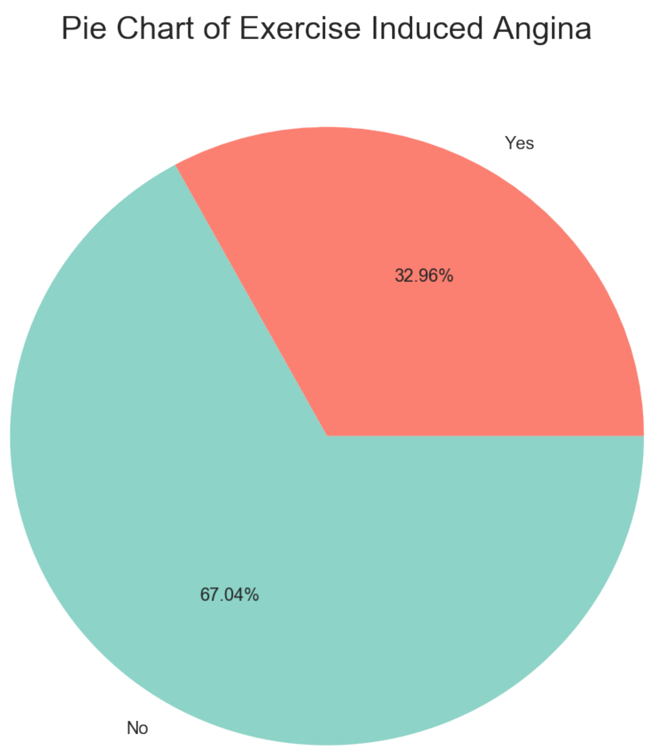
Heart rate is the number of times the heart beats per minute. Normal heart rate is between 60-100 bpm. Heart rate increases due physical activity. It is related to age too.

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

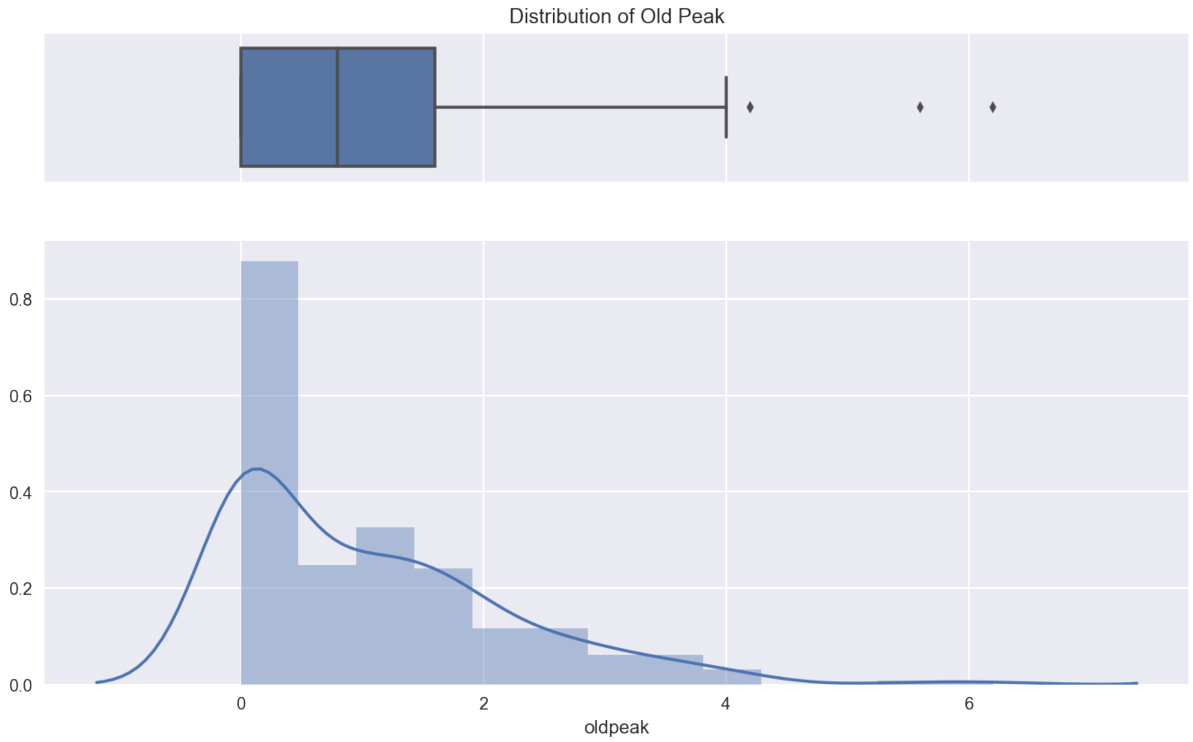
Angina is pain/discomfort that happens when your heart can't get enough blood and oxygen. (Heart Foundation 2018). Here 1/3 of patients has angina while 2/3 has not.



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)

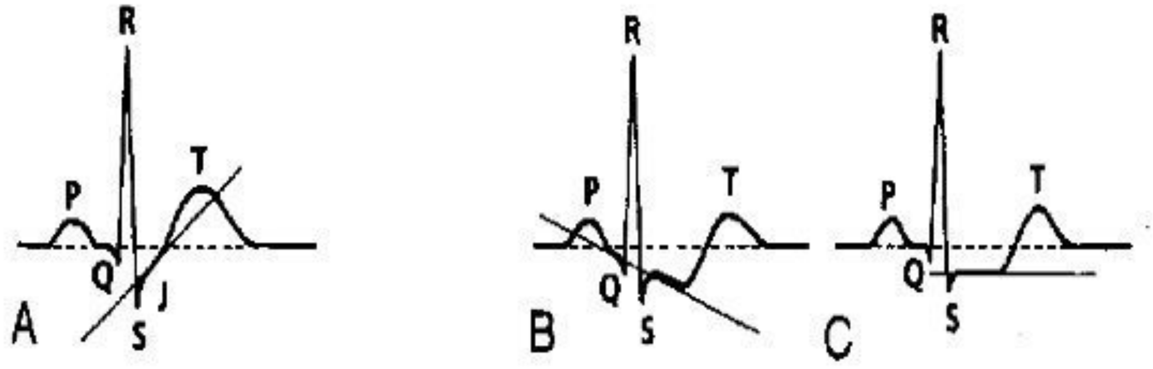
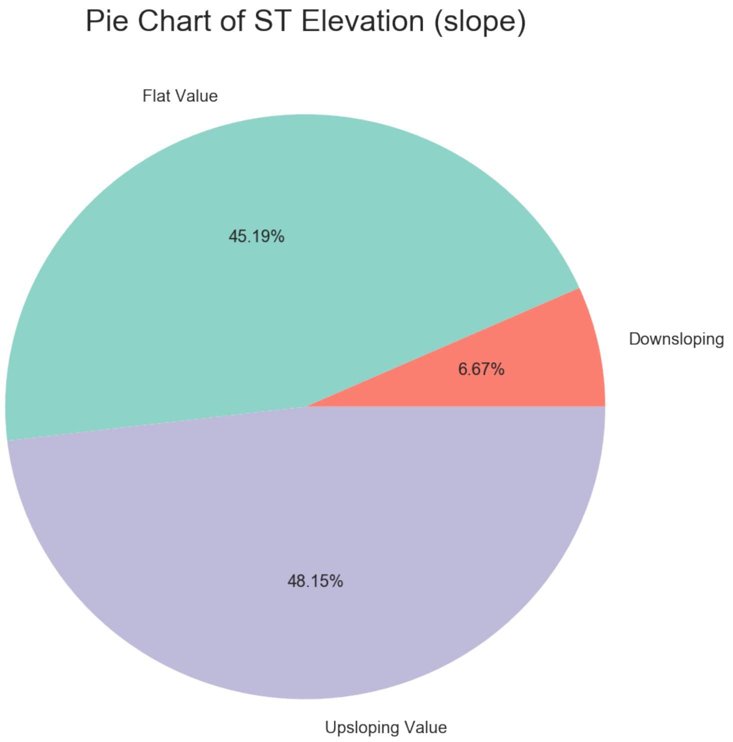


Figure 2 – ST depression: upsloping (A), down sloping (B), horizontal (C).
(Burns 2017)

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

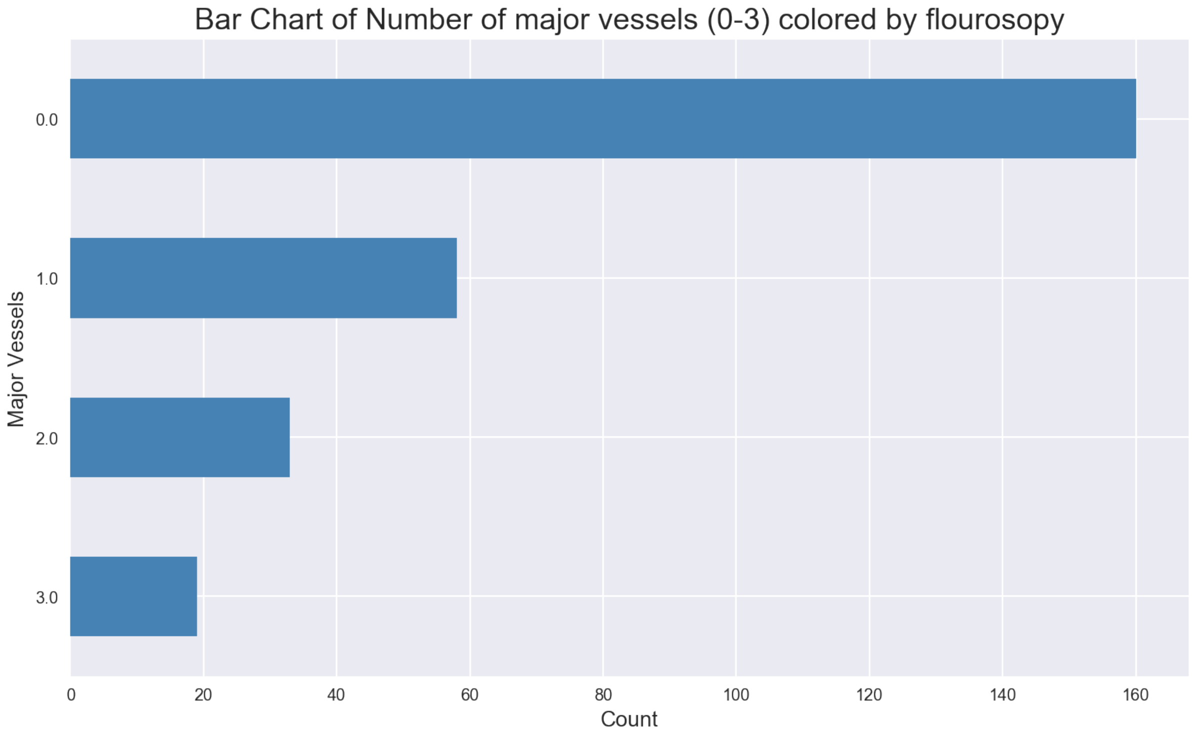
Pie chart of slope elevation has 45% flat, 48% upsloping while 7% as down sloping.



Major vessels by fluoroscopy

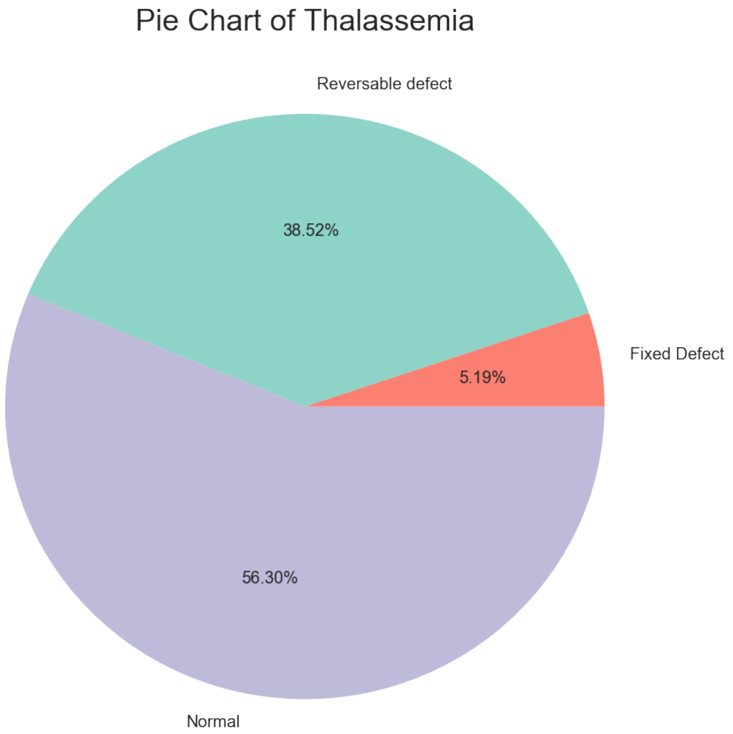
This is number of major vessels (0-3) coloured by fluoroscopy. Fluoroscopy, as an imaging tool, enables physicians to look at many body systems (John Hopkins Medicine 2018)

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)

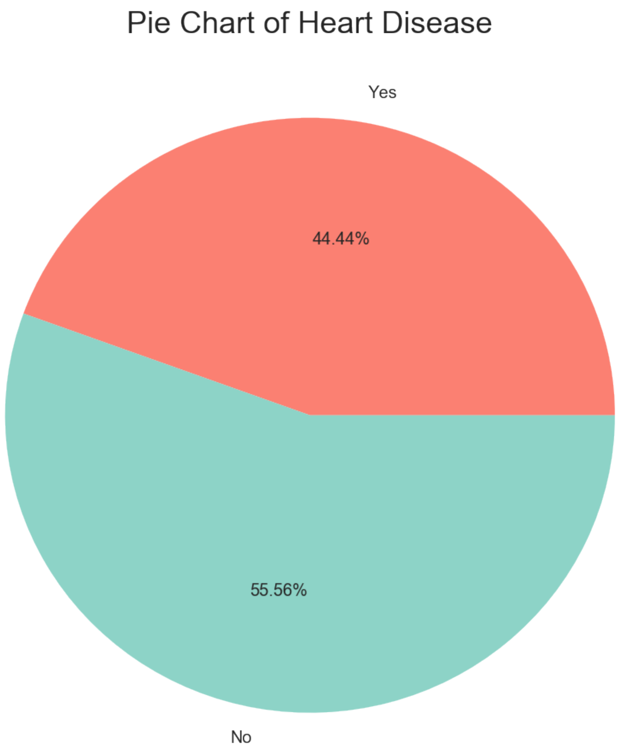


We have 3 types of variable for Thalassemia

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.



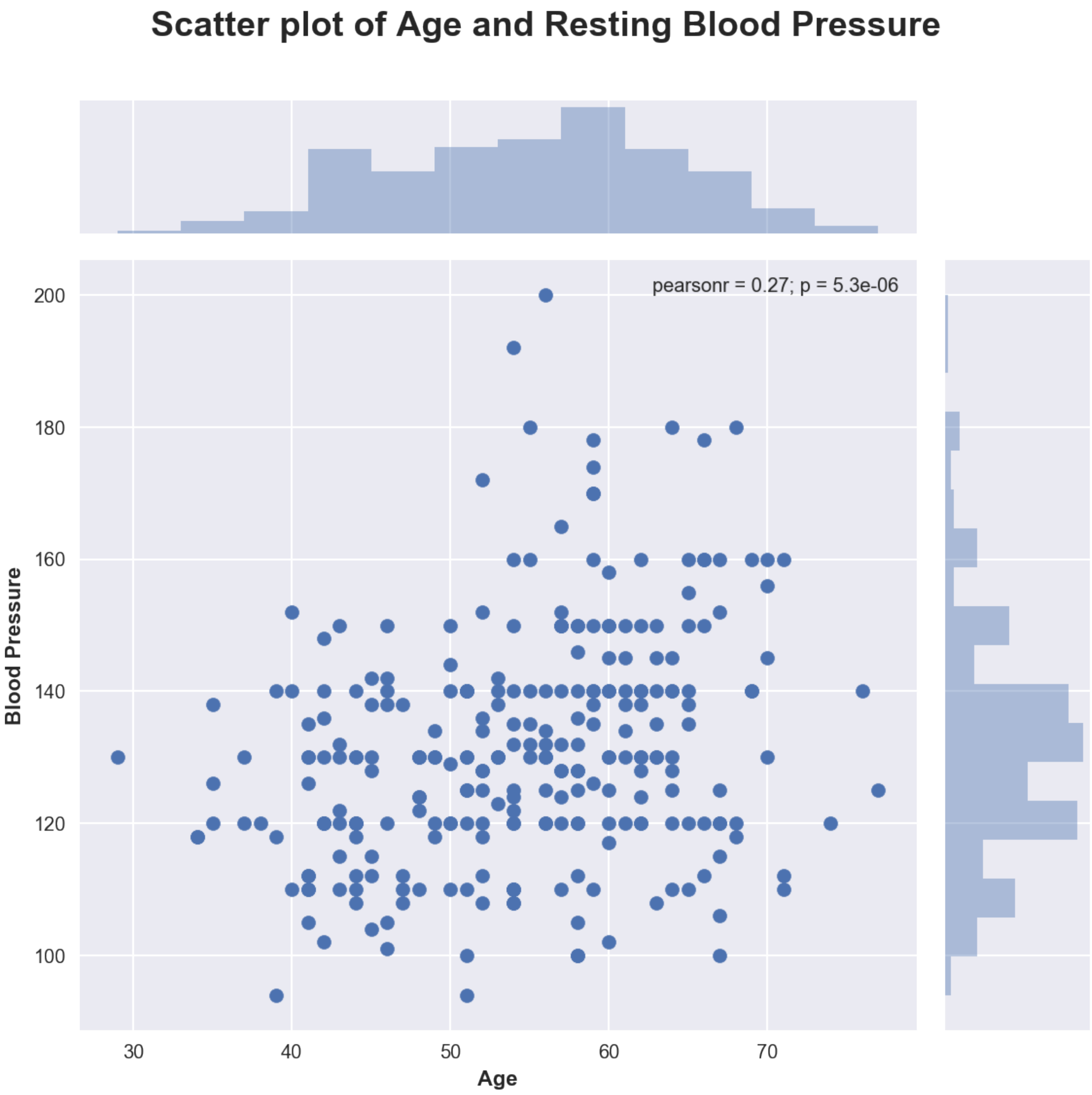
Next, we’ll analyse relationships between multiple variable.

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.

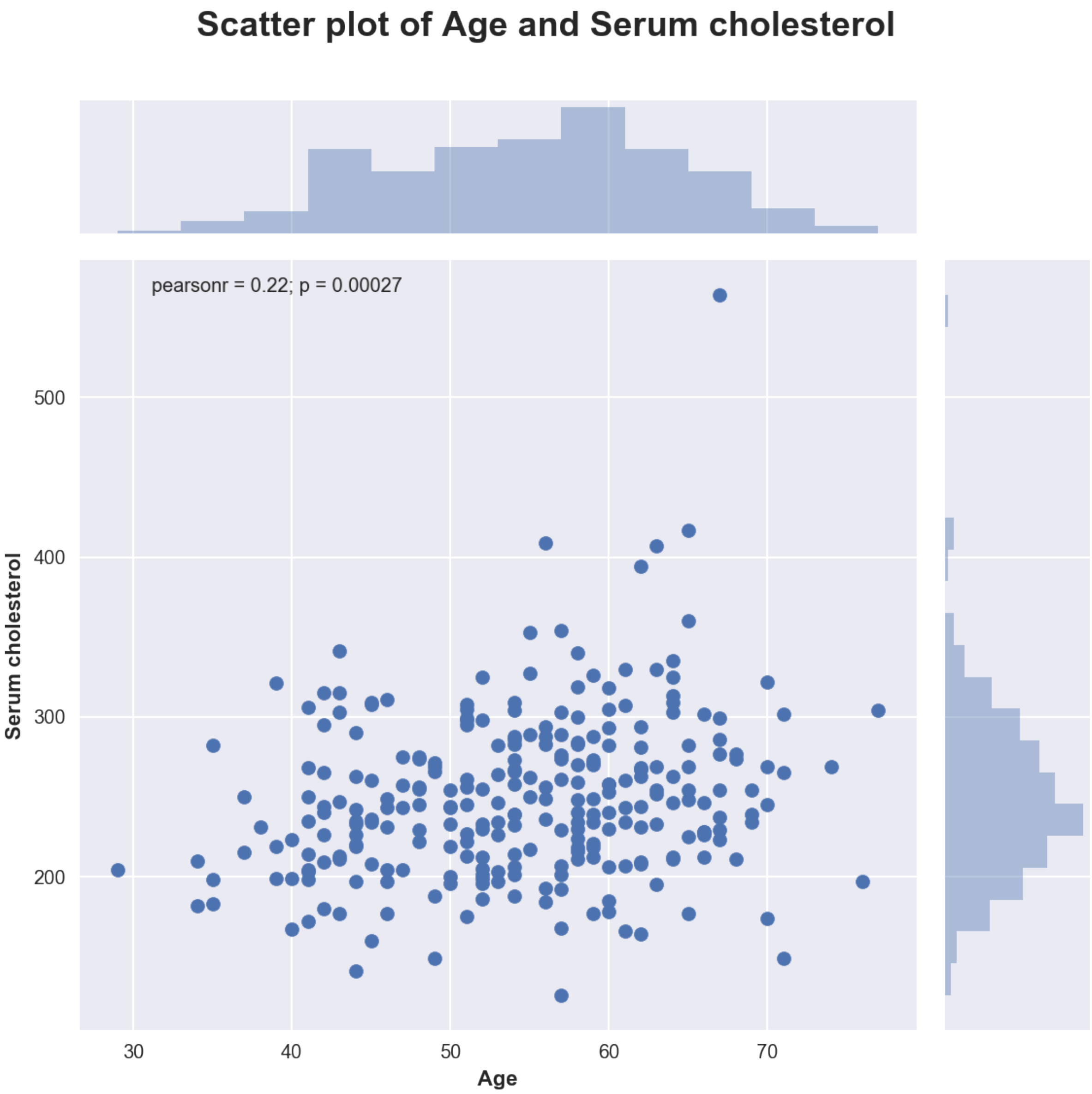
Age and Resting Blood Pressure

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.

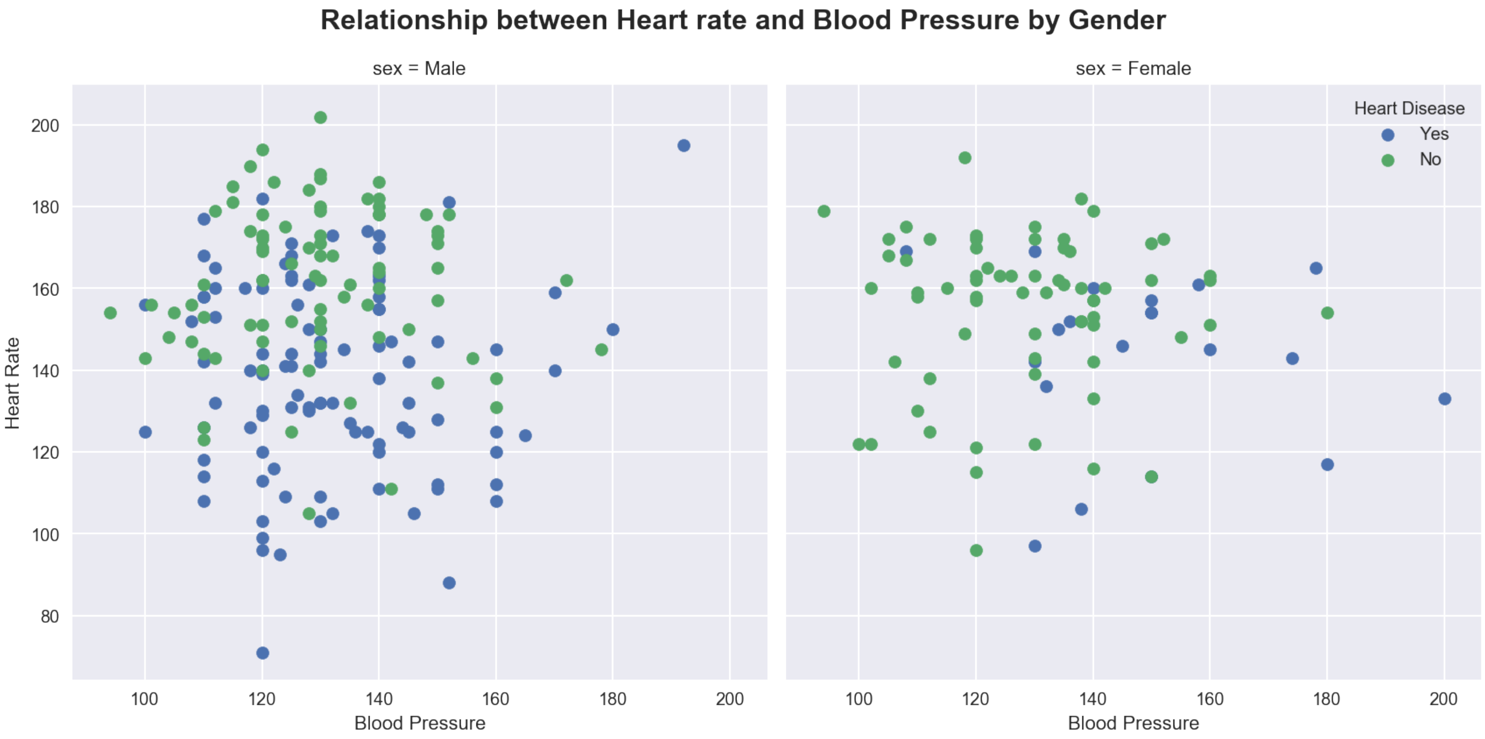


Next, we can see relationship between blood pressure and heart rate by gender type.

Heart Rate and Blood Pressure by Age

Here we notice:

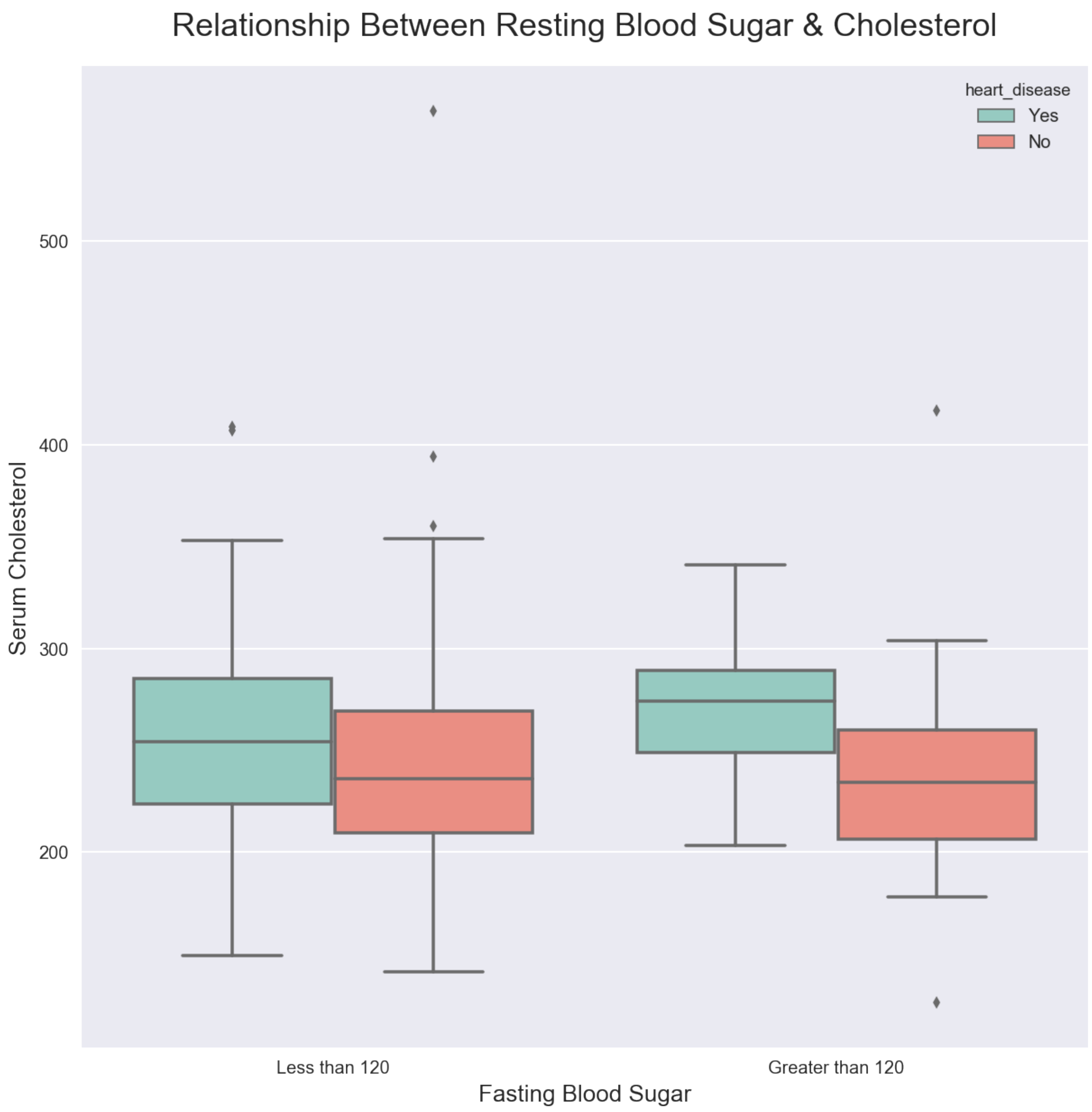
- Females generally have higher blood pressure than males
- Females with higher BP and heart rate have more tendency of heart disease
- Males even with normal BP and heart rate gets heart disease. There must be another factor for disease.



Blood Sugar and Cholesterol

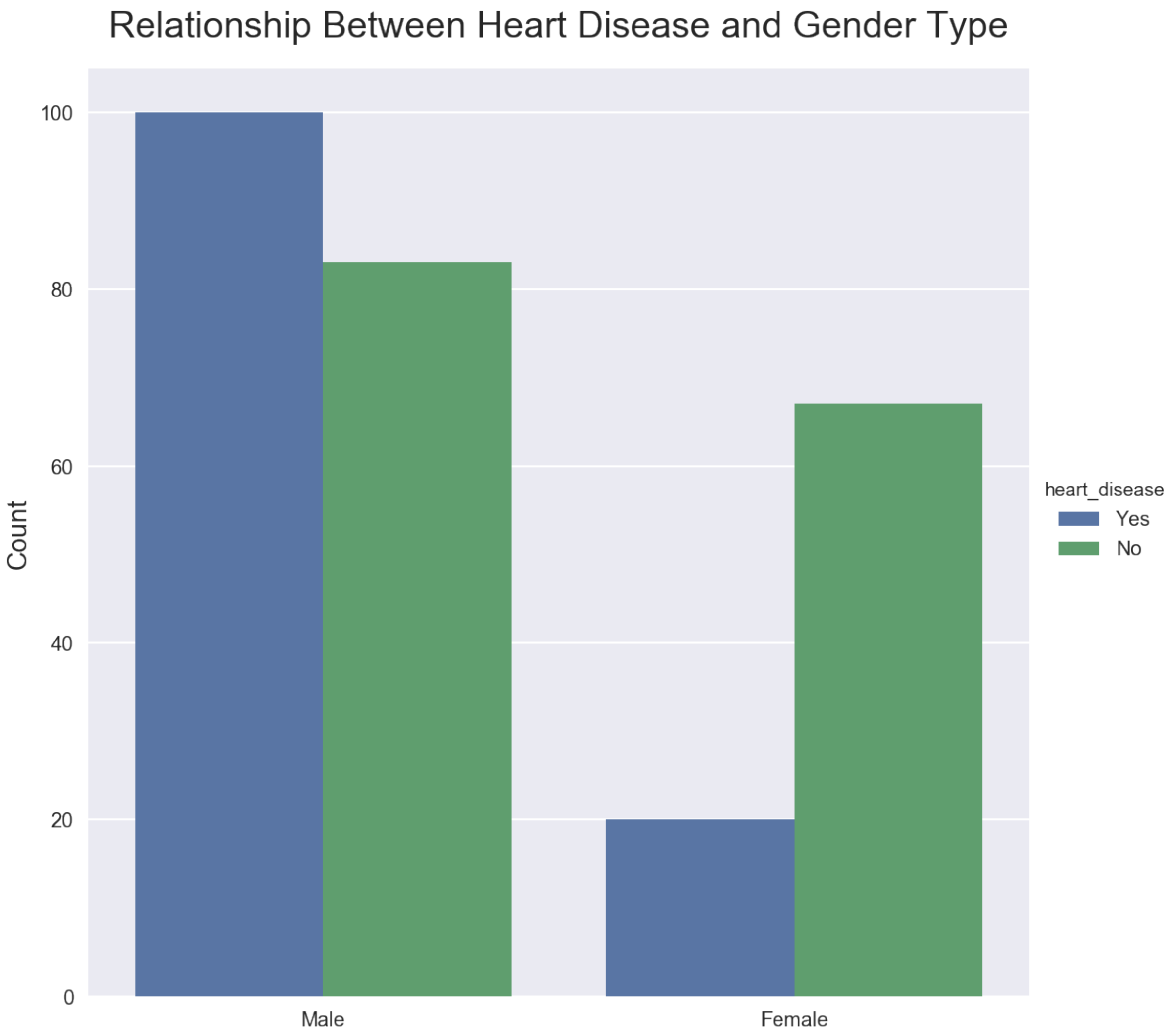
Here we notice:

- 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

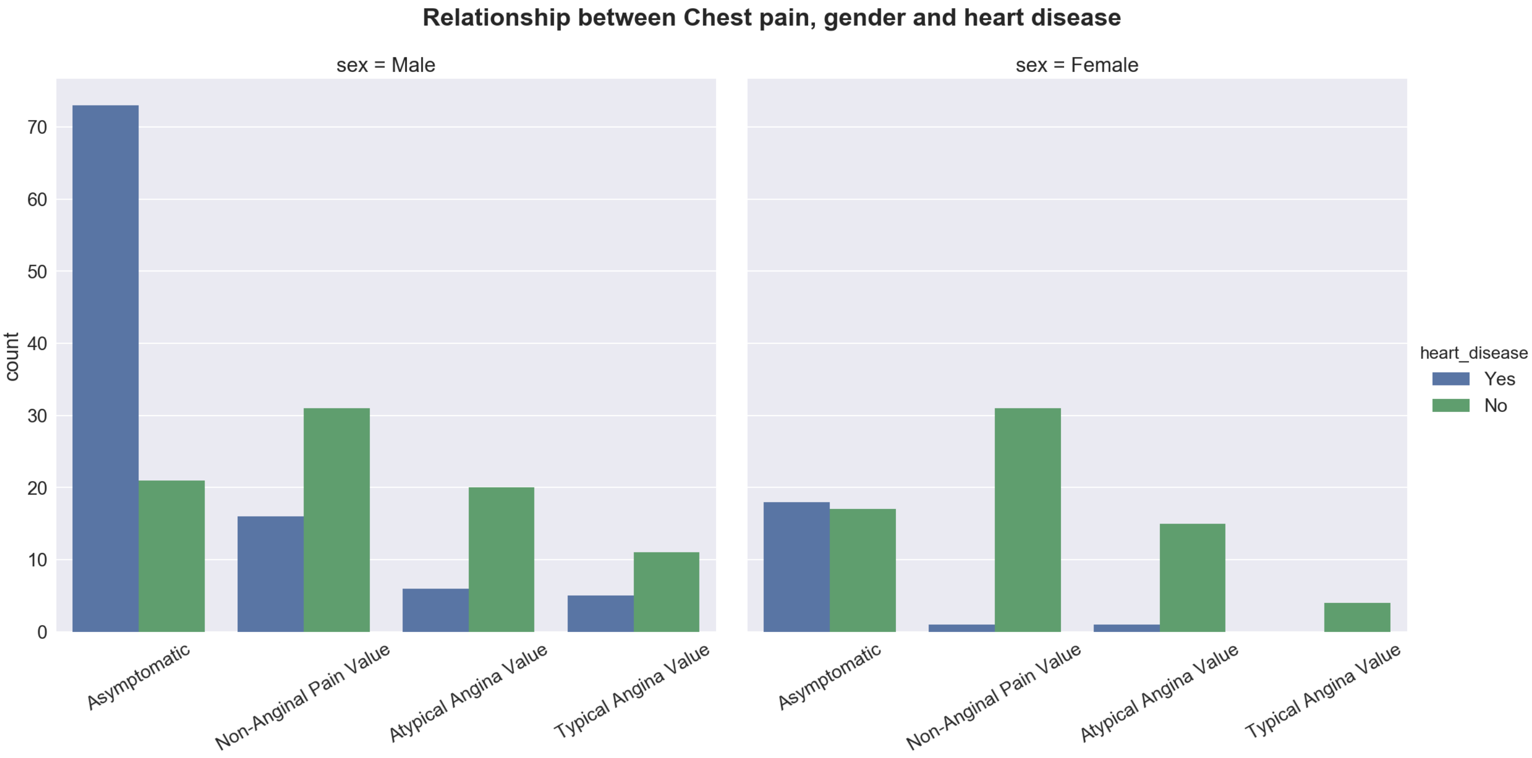
Male count with heart disease is 4 times than female. This shows an interesting result than chances of getting heart disease for males are higher.



Chest pain with gender

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

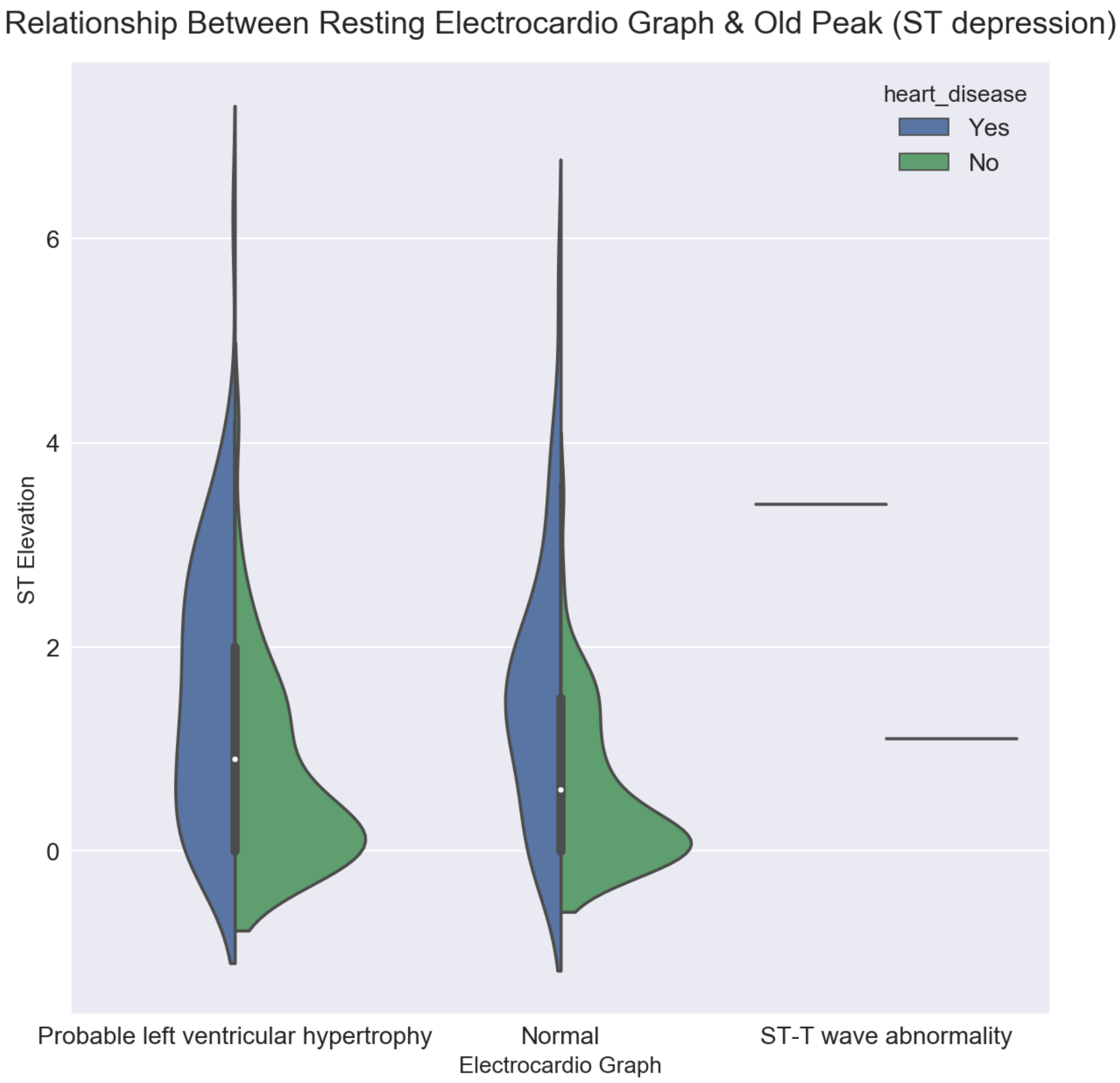
- Here chest pain *Asymptomatic* indicates that it's not highly related with heart disease
- Other types of chest pain are not significant for both males and females



Electro cardio graph and oldpeak

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

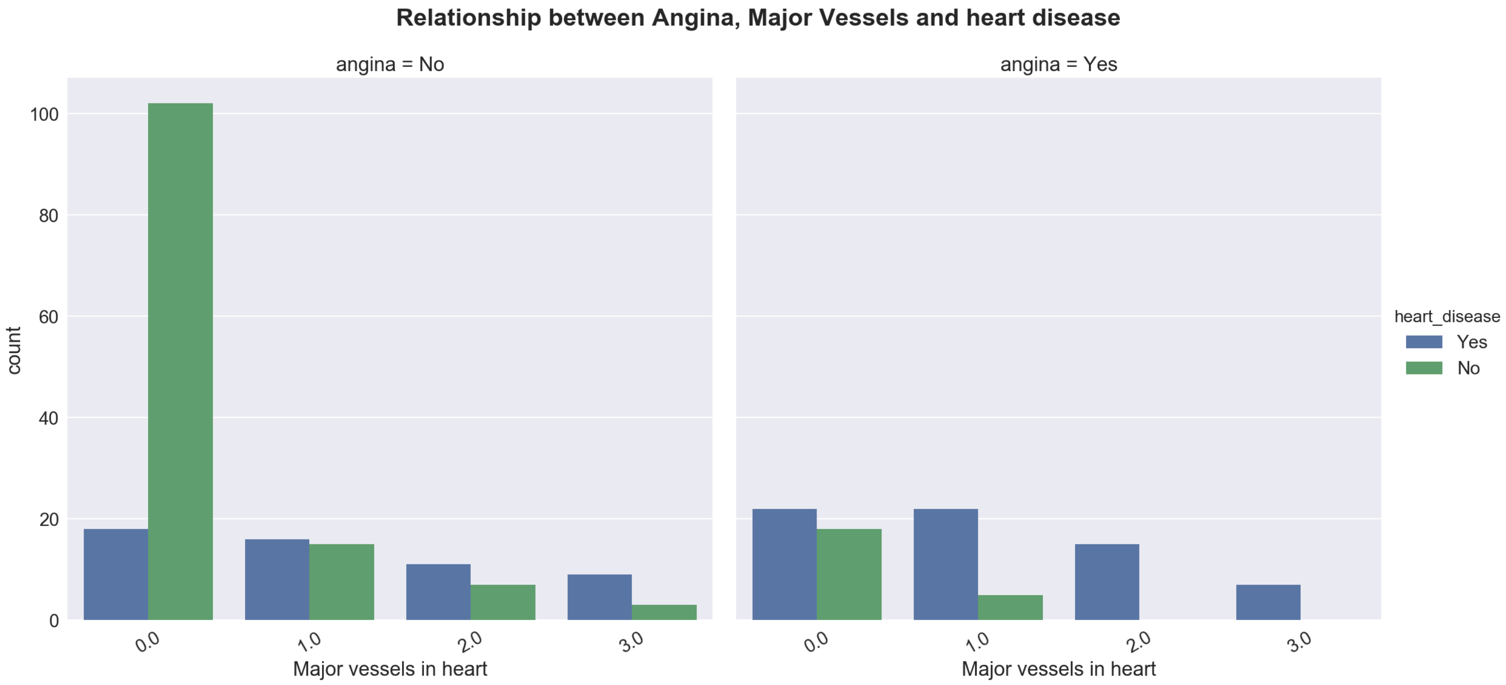
- People with high ST elevation has higher old peak value
- Distribution is around zero for absence of heart disease



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

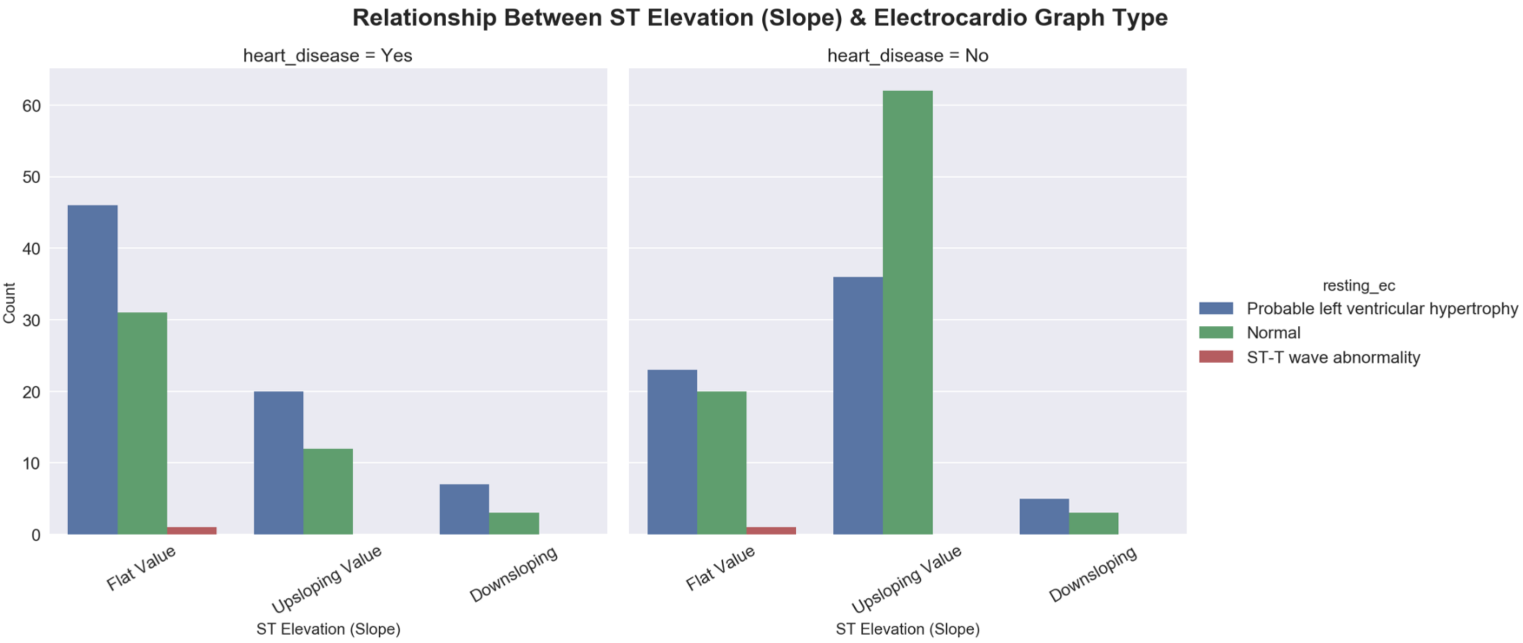


ST Elevation with EC type

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.

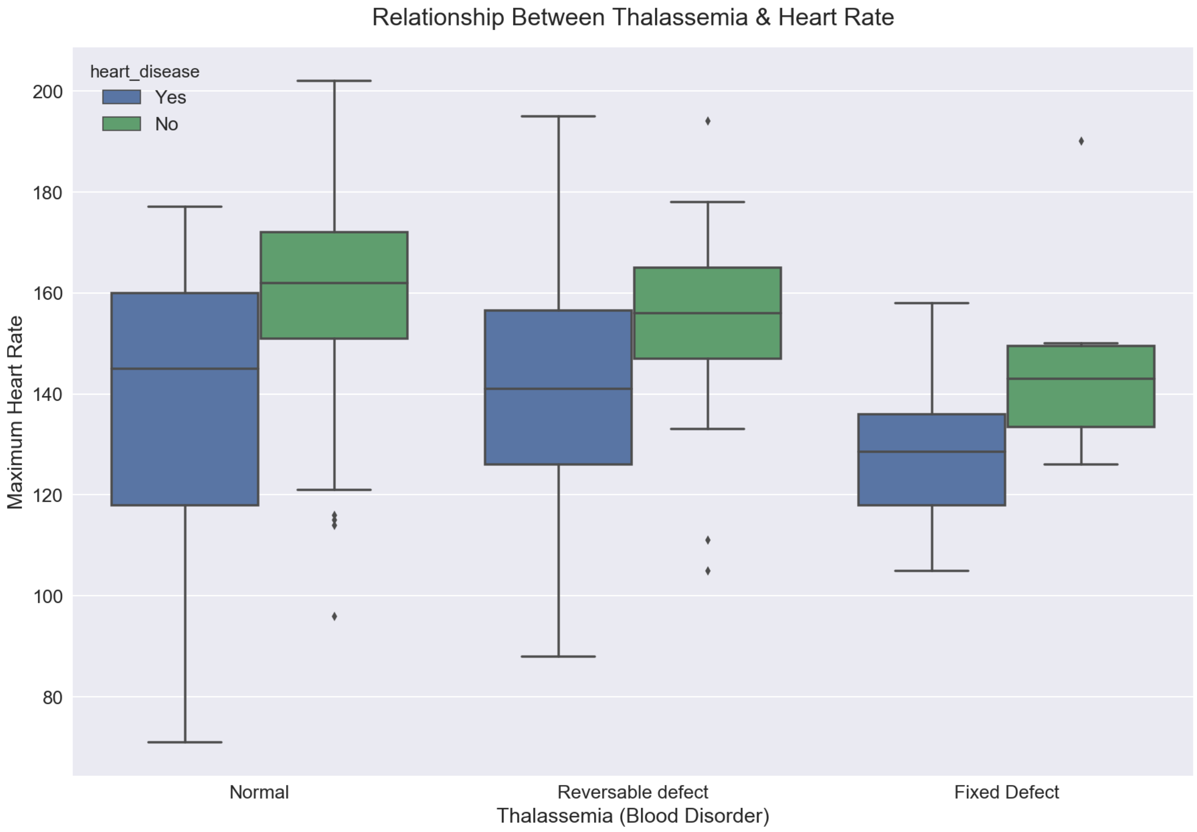


Thalassemia and Heart Rate

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.



Data Modelling

Depending upon the dataset previously selected the classification technique is used for modelling the dataset. Three classification models, namely Decision tree, Random forest and k-nearest Neighbors are chosen for data modelling. The following steps are done in order to model the given dataset:

a) The dataset is split into “Test Dataset” and “Training Dataset”.

b) The data is trained using appropriate values for each parameter.

c) You need to show how do you choose this value, and justify why you choose it

d) The accuracy of the model on the test dataset is tested and the performance of the model is reported on the following factors:

i. Confusion Matrix

ii. Classification error rate

iii. Precision

iv. Recall

v. F1-Score

1. Decision tree

To start with, Decision tree is one of the most popular and best predictive modelling approaches used in machine learning (supervised learning), data mining and in advances statistics. Decision tree algorithms come handy in solving any kind of real life data science problems. Even problems of classification and regression can easily be solved using this concept. Decision tree are of two types, namely Classification tree (when the predicted final result belongs to the class of the original data) and Regression tree (when the predicted result can be a real number). In this technique, the sample data is split into more than two sets/branches based on a most significant differentiator present in the input data. They are simple to compute and interpret and also can handle both categorical and numerical data. Also, very large amounts of data can be analysed and interpreted in a very short period of time which aids human decision making more precise than other techniques. Even though Decision tree approaches have a lot of advantages, they also have a considerable number of drawbacks such as they can be very non-robust i.e., a small change in the training data can make a very huge change in the tree and in the final predictions. Another major drawback is the major problem of overfitting. A lot of data mining software packages provide implementations of these decision tree algorithms, such as Salford Systems, Matlab, SAS Enterprise Miner and R .

Terminologies in Decision tree:

- **Root Node:** This depicts the whole population or data sample and it gets further classified into more than two homogenous data sets.
- **Splitting:** This is the process of classifying a node into two or more subdivided nodes

- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
- **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.

- **Pruning:** The process of omitting/removing the subdivided nodes of a decision tree. This is the inverse process of splitting.
- **Branch / Sub-Tree:** The subdivided part of an entire decision tree.
- **Parent and Child Node:** When a node is subdivided into two more nodes it is the parent node, on the other hand, the subdivided nodes are the child.

```
[[37  7]

[13 24]]
```

The above shows the confusion matrix attained from modelling the data using Decision tree. With KFold(n_splits=10,random_state=4)

```
          precision   recall   f1-score   support


       0     0.74     0.84     0.79      44

       1     0.77     0.65     0.71      37


avg / total     0.76     0.75     0.75      81
```

This is the classification report of the decision tree. This shows that the precision is 76% and the recall, f1- score are 75% and 75% respectively. The f1- score is calculated as the average of precision and recall.

Cross-validation:

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

```
[fold 0] score: 0.74074

[fold 1] score: 0.74074
```

```
[fold 2] score: 0.77778

[fold 3] score: 0.77778

[fold 4] score: 0.70370

[fold 5] score: 0.77778

[fold 6] score: 0.77778

[fold 7] score: 0.77778

[fold 8] score: 0.62963

[fold 9] score: 0.81481
```

The results in this K-fold cross validation doesn’t not show the same percentage as Precision ,recall or f1-score. And differs in almost all the folds.

2. Random forest

Random forest is capable of performing both regression and classification. It is a highly versatile machine learning technique and it can handle large number of features. It is also extremely accurate is estimating which of the selected variables is important in the data which is modelled. It can be used to solve any kind of data science

problem with a combination of several models for a prediction.

```
[[40,  4],

[11, 26]]
```

The above shows the confusion matrix attained from modelling the data using Random Forest. With KFold(n_splits=10,random_state=4)

```
             precision    recall  f1-score   support


          0       0.78      0.91      0.84        44

          1       0.87      0.70      0.78        37


avg / total       0.82      0.81      0.81        81
```

Cross-validation:

```
[fold 0] score: 1.00000

[fold 1] score: 0.88889

[fold 2] score: 0.88889

[fold 3] score: 0.85185

[fold 4] score: 1.00000

[fold 5] score: 0.96296

[fold 6] score: 0.92593

[fold 7] score: 0.92593

[fold 8] score: 0.96296

[fold 9] score: 0.96296
```

3. K-Nearest Neighbors algorithm:

K-Nearest Neighbors is a non-parametric method in machine learning used for classification and regression. In classification, an entity/object is grouped by the majority vote among its neighbouring objects. The object is then alloted to the group, most similar among its neighbouring objects. Whereas, in the K-nearest regression technique, the result corresponds to the property value for the object/entity. The mean value of the neighbours in the k cluster is the corresponding value of the object. The K-nearest neighbours are selected from the group of objects from which the class or the object property value is known.

For continuous variables, the metric used for evaluation is the Euclidean distance.

For discrete variables, overlap metric (also known as Hamming distance) is used for text classification.

Dimensionality reduction is performed before applying k-Nearest neighbour algorithm. Even though there are advantages, there are also disadvantages when the distribution is skewed.

```
[43,  1],

[ 1, 36]]
```

The above shows the confusion matrix attained from modelling the data using K Nearest Neighbours. With KFold(n_splits=10,random_state=None, Shuffle=false)

```
             precision    recall  f1-score   support

          0       0.98      0.98      0.98        44
          1       0.97      0.97      0.97        37

avg / total       0.98      0.98      0.98        81
```

Cross-validation:

```
[fold 0] score: 1.00000
[fold 1] score: 1.00000
[fold 2] score: 1.00000
[fold 3] score: 1.00000
[fold 4] score: 1.00000
[fold 5] score: 1.00000
[fold 6] score: 1.00000
[fold 7] score: 1.00000
[fold 8] score: 1.00000
[fold 9] score: 0.59259
```

Discussion

- Hot encoding is not used for categorical variable
- More advance models can be used like neural network
- Regression models can be applied too which is not part of this report

Conclusion

We successfully using python libraries to explore, visualize and fit machine learning models in heart dataset. We found many feature variables are useful for heart disease prediction. After applying supervised learning models i.e. decision tree, random forest and K nearest neighbours.

We found random forest as the best prediction model for heart dataset.

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.