**ACTIVE LEARNING PROJECT 1**

**DATASET:**

The given dataset involves prediction of heart disease based on some health factors as listed below.

1. Age   
 2. Sex   
 3. Chest pain type (4 values)   
 4. Resting blood pressure   
 5. Serum cholesterol in mg/dl   
 6. Fasting blood sugar > 120 mg/dl  
 7. Resting electrocardiographic results (values 0, 1, 2)  
 8. Maximum heart rate achieved   
 9. Exercise induced angina   
 10. Old peak = ST depression induced by exercise relative to rest   
 11. The slope of the peak exercise ST segment   
 12. Number of major vessels (0-3) colored by fluoroscopy   
 13. Thalassemia: (values 0, 1, 2, 3)

**Problem statement:**

Predict whether the person is having the heart disease (Binomial outcome) based upon the different health factors of the person and what are the various factors reasonable for the heart disease

**Tools used:**

R studio:

* Basic statics and outlier detection using box plots

Python:

* Reading the dataset from MySQL.
* Finding any missing values and identifying the data types of the variables.
* Analysing the spread of the data univariantely.
* Transformation of the data and building the logistic model with train and test data.

MySQL:

* Analysing the data spread of independent variable with the target variable.

Tableau:

* Different types of visualization

**Domain information**

ST Depression

<https://en.wikipedia.org/wiki/ST_depression>

Chest pain

<https://jamanetwork.com/journals/jama/fullarticle/201900>

Resting blood pressure

<https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/blood-pressure-vs-heart-rate-pulse>

Fasting blood sugar

<https://www.medicalnewstoday.com/articles/317466.php>

Serum Cholesterol

<https://www.medicalnewstoday.com/articles/321519.php>

Heart Rate

<https://www.cdc.gov/physicalactivity/basics/measuring/heartrate.htm>

Exercise induced angina

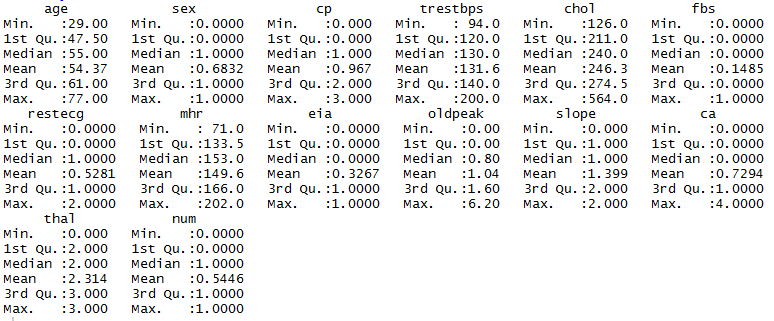
<https://www.medicalnewstoday.com/articles/323497.php>

Old peakedness

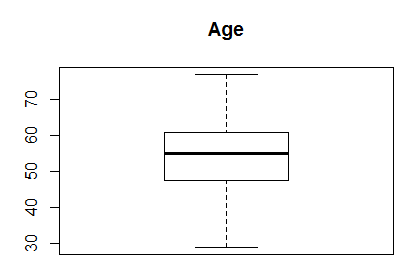
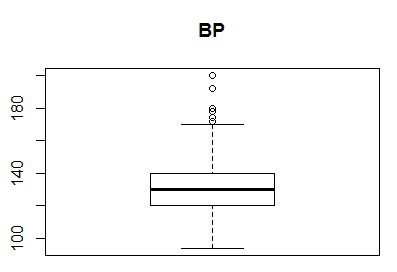
<https://www.sciencedirect.com/science/article/pii/073510979290044N>

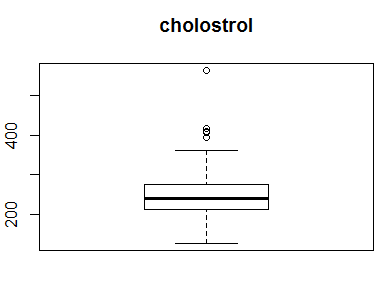
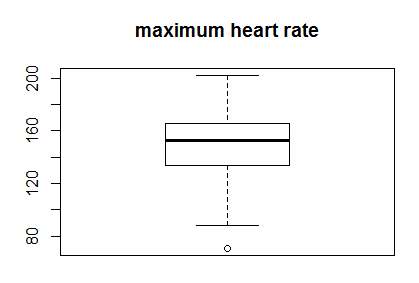
**R studio:**

* Basic statics and outlier detection using box plots.



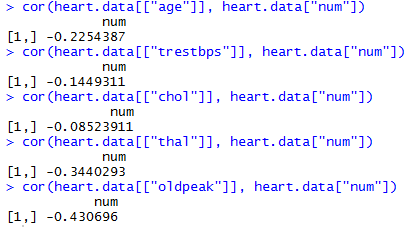
For the parameters BP, cholesterol and maximum heart rate there is a big difference between 3rd quartile and the maximum value which indicates there might be outliers present in the data.

Presence of outliers is confirmed using the boxplot,hence while building the model the data has to be scaled down before applying to the algorithm

**Correlation**



Old peakdness and thalasemmia are highly correlated with the heart disease

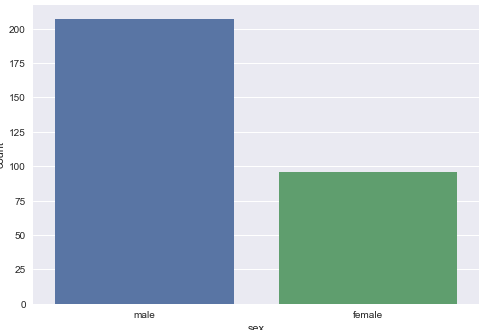
**Python:**

The dataset has 303 rows with 14 columns.

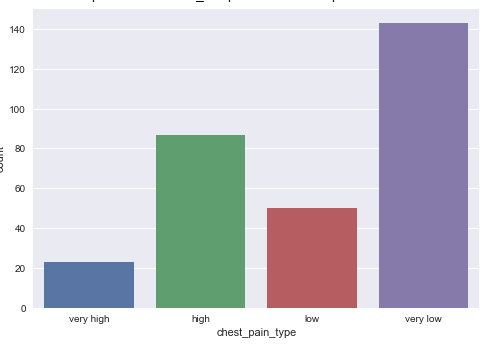
There are no null values in the dataset.

**Univariant analysis**

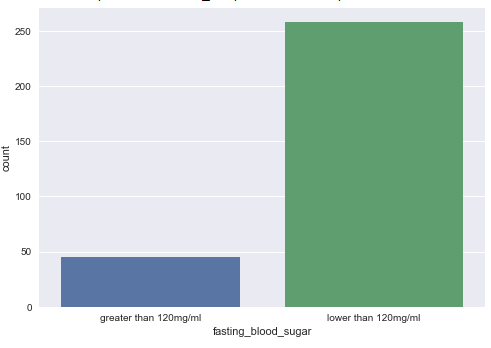
1. Number of males are more in the given dataset



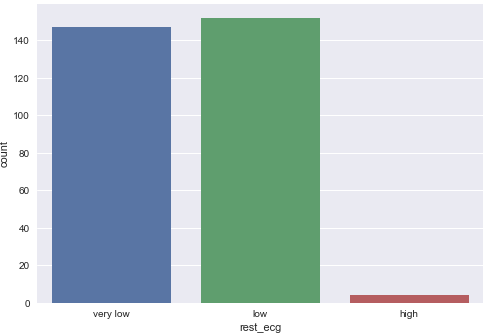
2. Most of the people are present with the chest pain of very low to high.



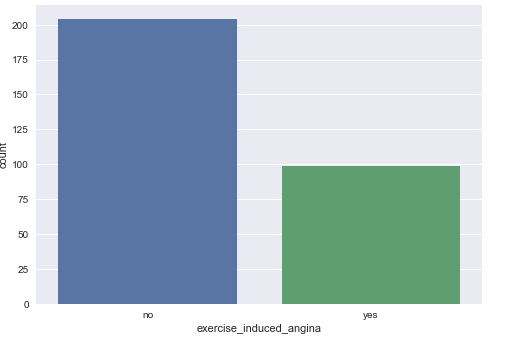
3. Blood sugar level less than 120 are more when compared to FBS>120



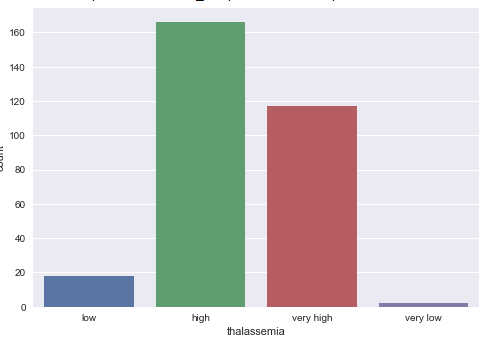
4. Most of the people are not having high rest ecg



5. People with exercise induced angina are more

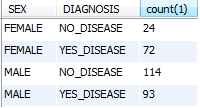


6. Thalassemia is more in the given dataset ranging from high to very high

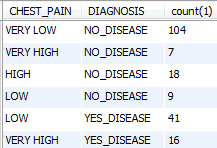




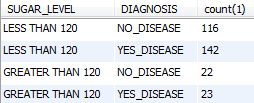
**Bivariate analysis with the target variable using SQL**



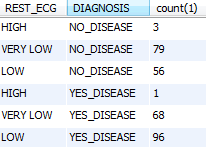
* Male tends to have heart disease more than female



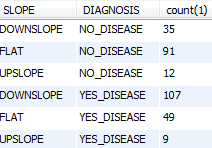
* If the chest pain is very low there is less chance of heart disease



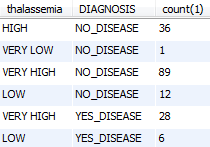
* Sugar level less than 120 along with other health factor seems to create heart disease.



* Rest ECG alone doesn’t influence the heart disease since there is equal spread of data with yes and no disease



* People with downslope tends to have greater risk of having heart disease.



* Surprisingly people with very high thalassemia tends to have less risk of heart disease



**Visualizations**

<https://public.tableau.com/profile/chandiraprakash#!/vizhome/ChandraPrakash_SSN_HeartDisease_Data_AnalysisandVisualization/Story1?publish=yes>

**Model building:**

**Used logistic regression:**





**Used Naïve Bayes:**

