INTRODUCTION

Heart disease is the leading cause of death in both men and women. 647,000 people die of heart disease every year. In America about 805,000 people have a heart attack every year and 605,000 of these are a first heart attack. And about 1 in 5 heart attacks are silent so the damage has been done without the person being aware of it.

Currently cardiovascular disease in the costliest disease in America with the price tag of $555 million in year 2016. On Average an employee with cardiovascular disease costs their employer $1,100 more per year due to loss in productivity. If the study can help with prediction of heart disease, preventive measures can be taken to reduce cost.

In this study we will look into correlations between heart trends, and or other biomedical traits that can help us predict the presence of heart disease. If there is a correlation between some heart activities and, or other biomedical traits this study can be used to help with heart disease prevention. As a high risk patient including men, the elderly, individuals with diabetes and persons with family members who have suffered from heart disease, early detection and prevention is important as heart disease is a silent threat and shows no symptoms before occurrence.

Prevention is paramount because in the event of a heart attack if the any of the heart muscles are damaged they cannot regrow, also if any of the valves in the heart become stiff and calcified there is no way to restore the flexibility and it must be replaced or repaired.

DATA

Data Acquisition and Cleaning

The data set comes from UCI machine learning repository, the full dataset had 76 attributes but the most widely used for machine learning till date was a subset of 14 attributes from the experiments from Cleveland database. The variable that describes the presence of heart disease is called target where 0 is presence of disease and 1 is the absence of the disease. Column names used in the original data set were abbreviated making the variables unclear, hence each column was renamed to the full medical name.The data type of each columns was checked and corrected, so that all columns were numerical. The isnull and sum function was used to determine a count of the total number of missing values in each column all null and missing values were deleted. Outliers were also identified and deleted. Variables are listed below

1. age
2. sex
3. chest pain type (4 values)
   -- Value 0: asymptomatic
   -- Value 1: atypical angina
   -- Value 2: non-anginal pain
   -- Value 3: typical angina
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)
   -- Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria
   -- Value 1: normal
   -- Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
    0: down-sloping; 1: flat; 2: up-sloping
12. number of major vessels (0-3) colored by fluoroscopy
13. thalassemia: 3 = normal; 6 = fixed defect; 7 = reversible defect
14. Target 0 = heart disease 1 = no heart disease

DATA EXPLORATION

PLOT 1: y axis: resting blood pressure, x axis: cholesterol
In the first chart (Scatter plot) there is a comparison between resting blood pressure and cholesterol, there is no correlation between resting blood pressure and cholesterol also there is no correlation between these two variables as it relates to heart disease increases.

PLOT 2: y axis: cholesterol, x axis: age
The next chart we are see a slight positive correlation between the age and the cholesterol is also visible that as the older the individuals get the more likely they were to have heart disease. Hence there is a slight increase in the cholesterol levels as the individuals age.

PLOT 3: y axis: maximum heart rate achieved, x axis: age
There is a negative correlation between maximum heart rate and age, so as people age we can see that the heart rate slows down significantly, we can also see by that the number or yellow colored markers, heart disease is more prevalent in individuals as they age. And hence lower heart rate seems to be more prevalent with individuals with heart disease than not.

PLOT 4: y axis: maximum heart rate achieved, x axis: resting blood pressure
There is a no correlation between maximum heart rate and resting blood pressure, but as 0 signifies heart disease we see that at all levels of resting blood pressure there are cases of heart disease without any trend however as maximum heart rate increases we see a gentle decrease in heart disease cases. So we can again say that lower heart rate can be associated to heart disease.

PLOT 5: pie chart fasting blood sugar percentage levels
There was a significant number of people had their fasting blood sugar below 120 mg/dl and out of those there was about 56.2% had fasting blood sugar below 120 mg/dl had heart disease, and about 50% of the people with blood sugar above 120 mg/dl had heart disease so we can see about the same percentage of people had heart disease in these two cases so from this chart there doesn't seem to be any correlation between fasting blood sugar and heart disease.

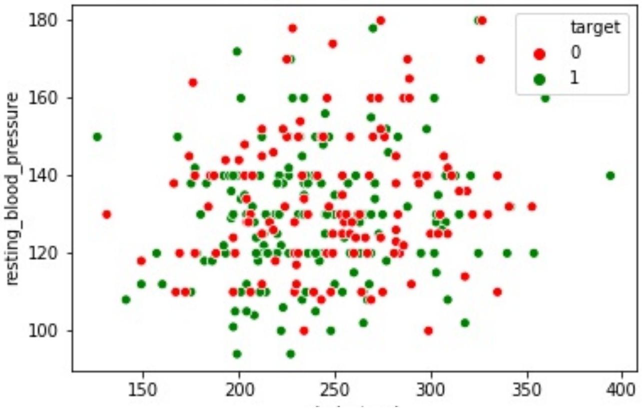
PLOT 6: pie chart exercise induced agina percentage levels
So we can see that out of tested patients the number of people with heart disease that have exercise induced agina is approximately 77% and the number of patients that have heart disease without exercise induced agina is approximately 30% there fore we can say that from these charts we can assume that there is a higher chance that someone with exercise induced agina to develop or have heart disease.

PLOT 7: bar chart comparing the types of thalassemia with presence of heart disease
There are 2 types of thalassemia which are either reversible which can be treated and corrected and the fixed type that comes from gene mutation and is not reversible. From the charts we can see that most people that do not have heart disease also do not thalassemia while approximately 64% of the people with heart disease had the reversible form of thalassemia which is a large percentage of people and from this chart we can suggest a correlation between reversible thalassemia and heart disease.

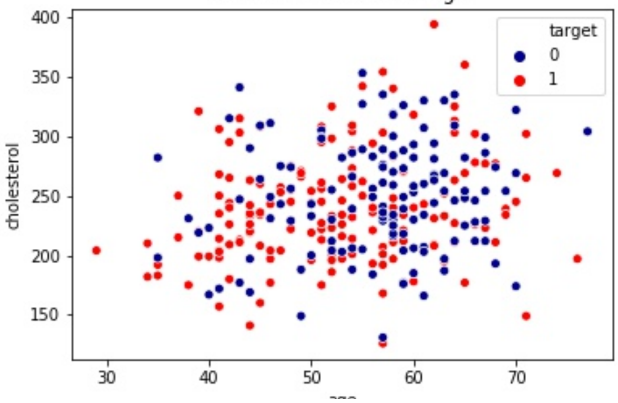
PLOT 8: bar chart comparing resting electrocardiographic results
There is no significant difference in results for the resting electrocardiograph so we can say there is no significant correlation directly between resting electrocardiograph results and heart disease.

PLOT 9: bar chart comparing chest pain and heart disease

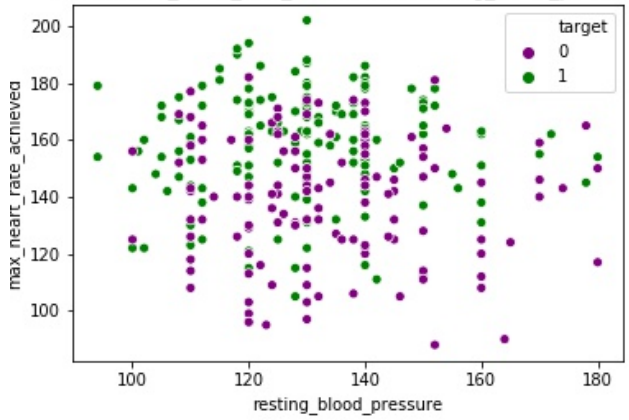
Plot 1 - Resting_blood_pressure vs Cholesterol



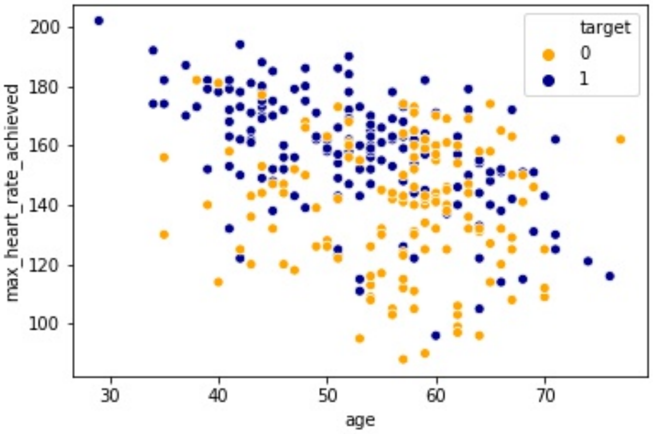
Plot 2 - Cholesterol vs Age



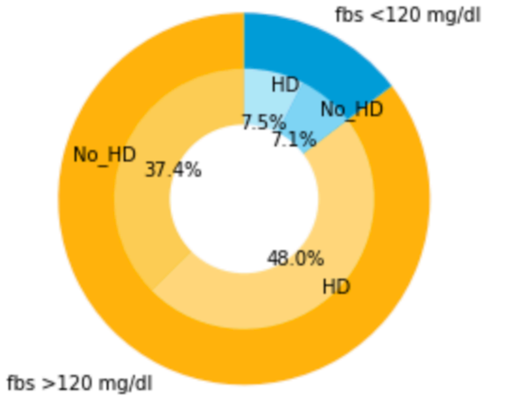
Plot 4 - max_heart_rate_achieved vs resting_blood_pressure



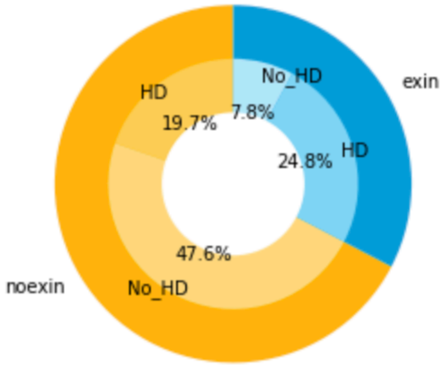
Plot 3 - max_heart_rate_achieved vs Age

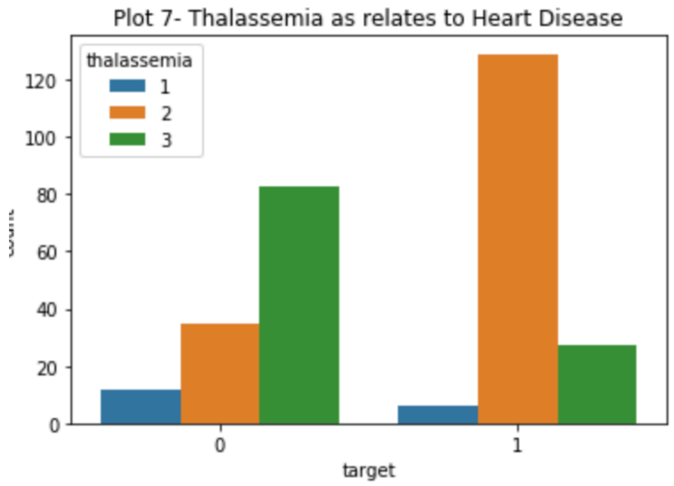


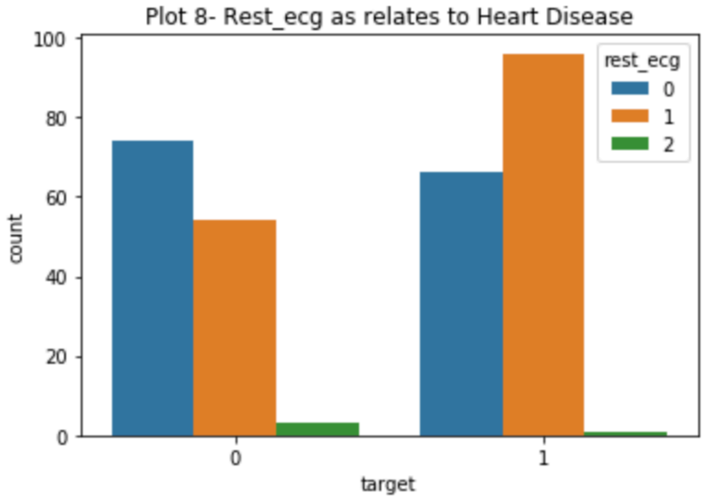
PLOT 5 - Fasing blood sugar to Heart Disease

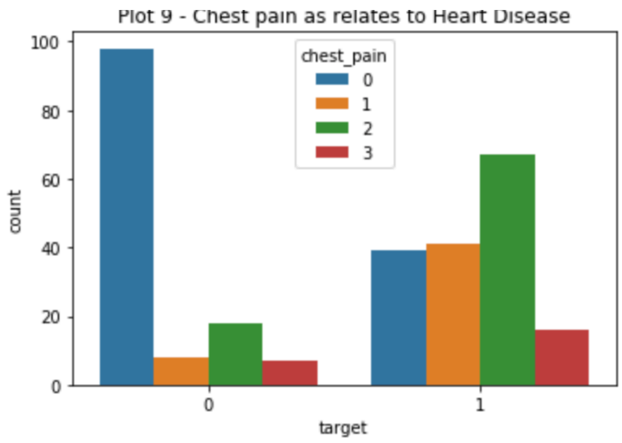


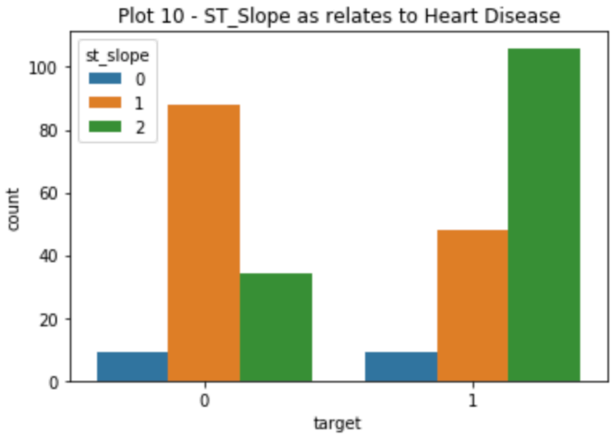
PLOT 6 - Excersice induced agina to Heart Disease











From the results of this graph the most prevalent type of chest pain in the individuals with heart disease was asymptomatic chest pain, this was found in high number, though there was also a sizable number of people without heart disease that also have that type of heart pain so having asymptomatic chest pain may be an indication of heart disease.

PLOT 10: bar chart comparing st_slope and heart disease
The predictions from the chart suggest that people without heart disease have up-sloping ST segment while those with heart disease have a high number of flat ST segment, suggesting a correlation between flat ST segment and heart disease.

Findings on Data exploration

Out of the 14 attributes selected we are trying to determine which of these attribute exhibit any behaviors that can help us predict the presence of heart disease. By using simple charts we will try to find any surface level correlation to heart disease and from our first four analysis. Out of these four characteristics: resting blood pressure, cholesterol, maximum heart rate and age, In plot 4 Comparing maximum heart rate and resting blood pressure, heart disease patients are found spread out through all ranges of resting blood pressure but on maximum heart rate axis they concentrated on the lower region where the maximum heart rates are lower, similarly looking at plot 3 where maximum heart rate is compared to age, as a person gets older we see that their maximum heart rate decreases also we see that heart disease patients increase, showing again correlation between heart disease and maximum heart rate. In plot 1 resting blood pressure to cholesterol there is no correlation in this scatter plot because the patients with the heart disease is evenly mixed in with those without heart disease on both axes. Hence for the individual data from our surface level graph observation a low maximum heart rate is a strong variable in heart disease detection.

As we look into the categorical data from plot 6 to plot 9, there are six attributes that we make observations of. From the graphs we see that there seems to be high correlation between heart disease and the following variables: exercise induced agina (plot 6), reversible form of thalassemia in (plot 7), asymptomatic chest pain (plot 9) and a flat st_slope (plot 10). While fasting blood sugar (plot 5) and resting electrocardiograph (plot 8) does not show strong correlation to heart disease. So in conclusion from the results of this surface level overview of the data there are five attributes that if are found in a patient that predicts a high probability of heart disease which are low maximum heart rate, exercise induced agina, reversible form of thalassemia, asymptomatic chest pain and a flat st_slope.

STATISTICAL TESTING

In relation with the aim of this project we are trying to find out some of the biological traits of individuals who have heart disease. This data set consists of 14 attributes, however the attributes that were tested are in this research are ten as stated:

1. chest pain type (4 values)
2. resting blood pressure
3. serum cholesterol in mg/dl
4. fasting blood sugar > 120 mg/dl
5. resting electrocardiographic results (values 0,1,2)
6. maximum heart rate achieved
7. exercise induced angina
8. oldpeak = ST depression induced by exercise relative to rest
9. the slope of the peak exercise ST segment
10. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

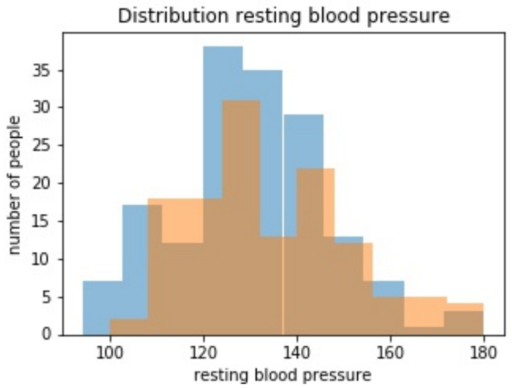
These attributes consist of continuous variable and categorical variables. The continuous variable were analyzed using the two sample t test for equal means these test have some assumptions that need to be verified to ensure that the test yields correct results. The t test assumptions are:

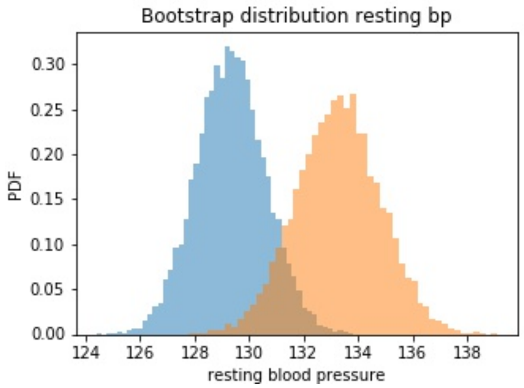
1. Independent observations
2. Normal distribution
3. Equal variances

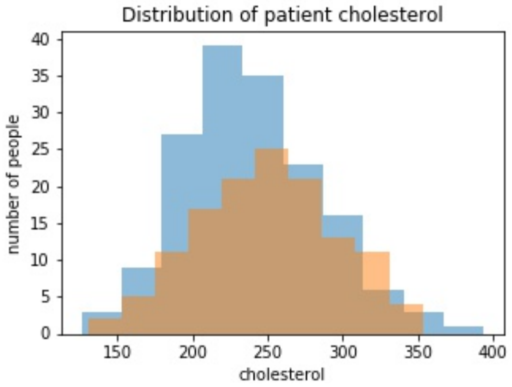
Hypothesis Test Procedures and Assumption Checks for Continuous Variables

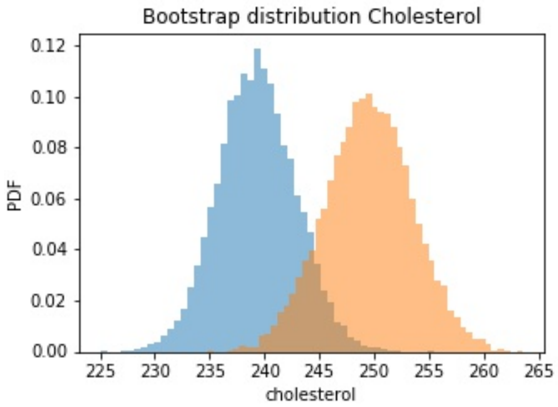
The procedures used to test the continuous variables are as follows:

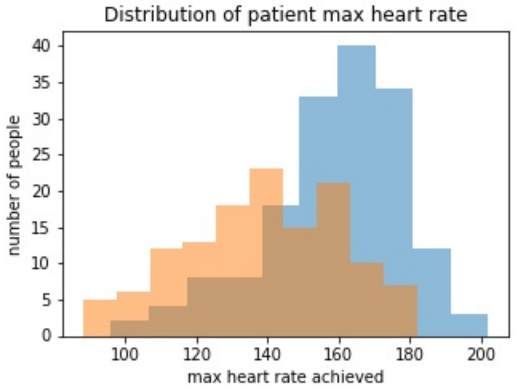
4. The Shapiro-wilk test is run to check for normality of the distributions. However the t test normal distribution assumption is satisfied if the distribution of means from random samples of data forms a normal distribution.

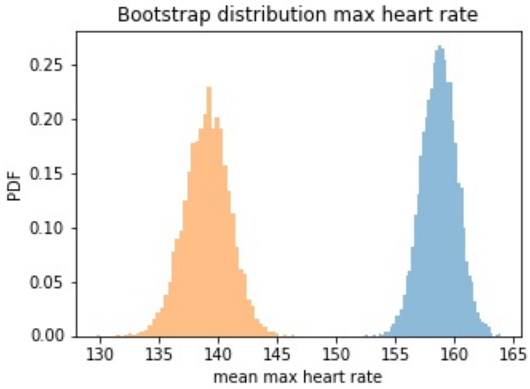


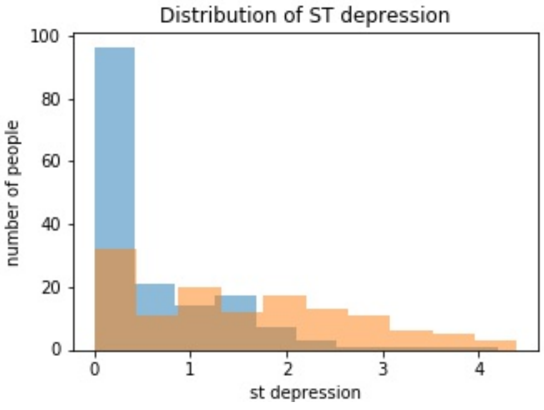


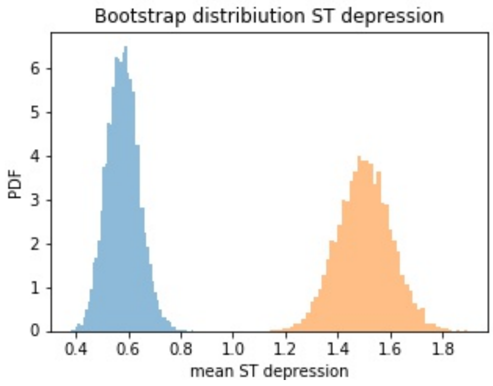


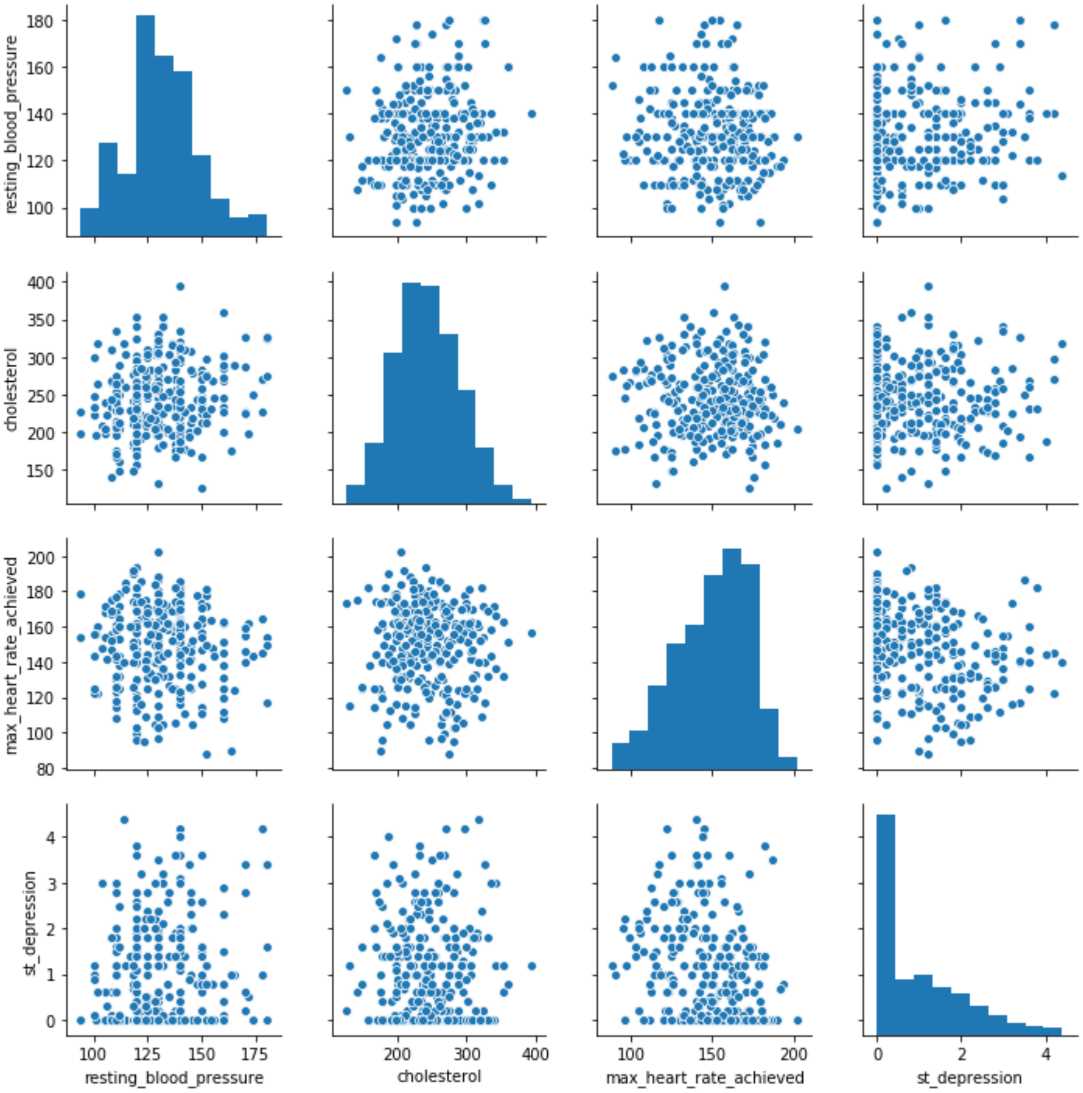












5. Bootstrap method is used to plot the distribution of means from random samples to ensure the second criterion is met.

6. Levene's test for equal variances is used before the test is run to ensure the scipy code is set to the right variant to ensure accuracy of results.

7. The two sample t test for independence is run. With the null hypothesis stating there is no difference between the means of the two samples. The p value is compared to the t statistic, if less then null hypothesis is rejected.

Also in order to see how these variables relate to one another the Pearsons test for correlation was run on the continuous variables to see how the continuous variable interact with one another

Test for Categorical Variables

The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. With the null hypothesis stating there is no independence between the variable in question and heart disease.

Results

There were 4 continuous variables : cholesterol, resting blood pressure, maximum heart rate, ST depression and 6 categorical attributes tested : fasting blood sugar, exercise induced aging, thalassemia, resting electrocardiograph, chest pain, and the slope of the peak exercise ST segment.

All three continuous variables cholesterol, resting blood pressure, maximum heart rate and ST depression are found to have non normal distributions hence the bootstrap method was used to find the distribution of the means, which were normally distributed allowing us to use the two sample t test. Using Levene’s test we determined cholesterol and resting blood pressure have equal variances while maximum heart rate and ST depression do not, using this to apply the appropriate conditions in the t test. From the t test we determine that for all four continuous variables the absolute value of the t statistic is greater than the than the pvalue hence we reject the null hypothesis meaning that the mean values of the patients with and without heart disease are significantly different. From these studies all the continuous variables are significant in helping us to determine the presence of heart disease.
There was also correlation matrix between these four variable and age there was no real strong correlation between these variables as the strongest correlation was between age and maximum heart rate which was negative correlation of 0.4.

The chi square test for independence is carried on the categorical variables of the six categorical variables fasting blood sugar, exercise induced agina, thalassemia, resting electrocardiograph, chest pain, and the slope of the peak exercise ST segment, two attributes fasting blood sugar and resting electrocardiographic results have a pvalues greater than the alpha which is set to 0.01 hence we fail to reject the null hypothesis which means we cannot prove there any association or relationship between those two variable and heart disease. While all the other variables have pvalues lower than the 0.01 so there is a high probability that there is association between these variables and heart disease.