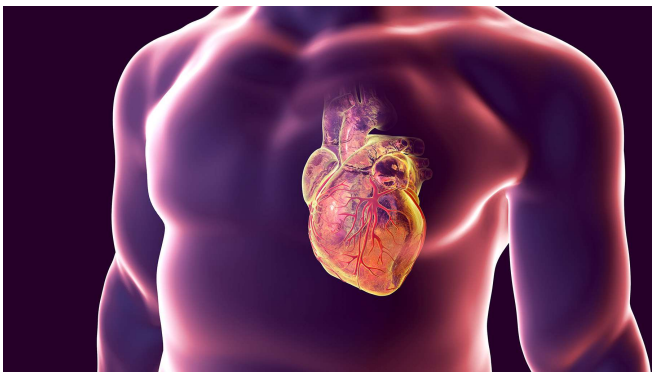


HEART DISEASE ANALYSIS

Submitted to

Prof. Guillaume Faddoul

Authors

Divya Raghunathan

Madhu Kaushik

Syed Asim

Aditya Tamhankar

ISYS-812: Programming and applications for Data Analytics - Python project report

Heart Diseases Report

Contents

1) Introduction

Heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease, heart rhythm problems (arrhythmias) and heart defects you're born with (congenital heart defects), among others.

In this project report, we will be applying data analysis techniques and statistics approaches and comparing the attributes for classifying whether a person is suffering from heart disease or not, using one of the most used dataset — Cleveland Heart Disease dataset from the UCI Repository.

Dataset can be found from Heart Disease UCI:
(https://www.kaggle.com/ronitf/heart-disease-uci)

1.1 Dataset description:

1.1.1 Number of Columns and Rows: 14 columns and 303 rows

1.1.2 Description of the data:

The data has 13 attributes which are independent factors. The 'goal' or dependent factor is the presence of heart disease in patients. We need to predict the presence of heart disease by analyzing the correlation between the attributes and the dependent factor.

1.2 Columns description

1.2.1 Age:

Displays the age of the individual.

1.2.2 Sex:

Displays the gender of the individual using the following format:
1 = male
0 = female

1.2.3 Chest-pain type:

Displays the type of chest-pain experienced by the individual using the following format:
1 = typical angina
2 = atypical angina
3 = non — anginal pain
4 = asymptotic

1.2.4 Resting Blood Pressure:

Displays the resting blood pressure value of an individual in mmHg (unit)

1.2.5 Serum Cholesterol:

Displays the serum cholesterol in mg/dl (unit)

1.2.6 Fasting Blood Sugar:

Compares the fasting blood sugar value of an individual with 120mg/dl.
If fasting blood sugar > 120mg/dl then: 1 (true)
else: 0 (false)

1.2.7 Resting ECG:

Displays resting electrocardiographic results
0 = normal
1 = having ST-T wave abnormality
2 = left ventricular hypertrophy

1.2.8 Max heart rate achieved:

Displays the max heart rate achieved by an individual.

1.2.9 Exercise induced angina:

#about eia
1 = yes
0 = no

1.2.10 ST depression induced by exercise relative to rest:

Displays the value which is integer or float.

1.2.11 Peak exercise ST segment

#peak
1 = upsloping
2 = flat
3 = down sloping

1.2.12 Number of major vessels (0–3) colored by fluoroscopy:

Displays the value as integer or float.

1.2.13 Thal:

Displays the thalassemia:
3 = normal
6 = fixed defect
7 = reversible defect

1.2.14 **Target** :

Displays whether the individual is suffering from heart disease or not:
0 = absence
1 = present

1.3 Missing values

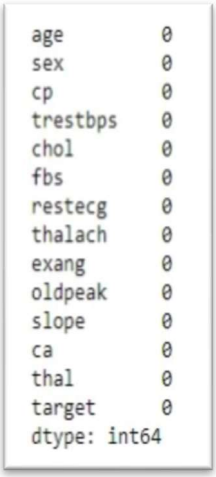


Figure 1: Missing values (screenshot from the code)

There are no missing values in this dataset.

1.4 Reason for selecting this dataset-

The dataset aligns with the project requirement of the data set. Since nowadays big companies like Google and Apple are more focused on bringing health as one of the parameters in upcoming projects.
Example: Google acquires Fitbit, Apple working towards more on health-related features in Apple Watch. So, as per the trend and growing concern of people towards health, health-related data is one of the most sorts after in the field of AI and ML. This project will help us to get a closer understanding of how things work around with the health data and how the analysis can contribute towards making a better prediction or reaching the precision of the prediction is accurate.

2) Facts derived from research on heart diseases

1. **Age:** Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.
2. **Sex**: Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman's risk is similar to a man's although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.
3. **Angina (Chest Pain)**: Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood
4. **Resting Blood Pressure**: Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.
5. **Serum Cholesterol**: A high level of low-density lipoprotein (LDL) cholesterol (the "bad" cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers your risk of heart attack.
6. **Fasting Blood Sugar:** Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body's blood sugar levels to rise, increasing your risk of heart attack.
7. **Resting ECG**: For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits.
8. **Max heart rate achieved:** The increase in the cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.
9. **Exercise induced angina:** The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. o Types of Angina a. Stable Angina / Angina Pectoris b. Unstable Angina c. Variant (Prinz metal) Angina d. Microvascular Angina.

10. **Peak exercise ST segment**: A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an 'equivocal' test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease

3) Driving question and subset questions

3.1 Driving question

"Finding how and which factors are key influencers in the heart diseases."

3.2 Subset questions

The below questions are answered through the course of this report.

Q1) From the dataset, what is the percentage and count of people suffering from heart diseases?

Q2) How different types of categorical variables impact the probability of a person having heart disease or not.

Q3) How different types of numerical variables impact the probability of a person having heart disease or not.

Q4) Which variables are "good" and "bad" predictors of Target (response variable)

Q5) What are the top 5 highly correlated variables with the response variable ?

Q6) How does the depression level of a person and type of their ECG slope determine if they have a heart disease?

Q7) How does the age of a person and their resting blood pressure and max heart rate determine if they have a heart disease?

Q8) How does the age of a person, their heart rate determine if they have a chest pain ?

Q9) How does the age of a person, their heart rate determine if they have a heart disease, based on their resting blood pressure ?

Q10) Are all the "good predictors" working well together in a model?

Q11) How well is the model performing with the test data?

4) Data Cleaning

The data cleaning process done in the project has been summarized in the steps below and the detailed steps can be found in the code file.

4.1 Initial dataset:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 2:First five rows of the inital dataset (screenshot from the code)

4.2 Missing values: There are no missing values in the dataset.

```
heart.isnull().values.any()

False
```

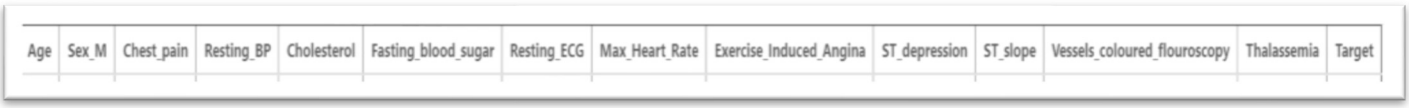
Figure 3: Missing values (screenshot from code)

4.3 Column names:

Since the column name in dataset is not very intuitive, we have renamed the columns so that it becomes easier to understand the data.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target

Figure 4: Column names before cleaning (screenshot from code)



Age	Sex_M	Chest_pain	Resting_BP	Cholesterol	Fasting_blood_sugar	Resting_ECG	Max_Heart_Rate	Exercise_Induced_Angina	ST_depression	ST_slope	Vessels_coloured_flouroscopy	Thalassemia	Target

Figure 5: Column names after cleaning(screenshot from code)

4.4 Thalassemia

Since Thalassemia column had two 0 values which didn't have any meaning, we removed the rows containing the 0 values from our dataset. Hence, now the number of rows has reduced to 301.

```
heart.dropna(subset=['Thalassemia'],inplace=True)
heart.shape
```

```
(301, 14)
```

Figure 6:The shape of the dataset after dropping thalassemia 0 rows (screenshot from code)

4.5 Changing data values to make it suitable for Logistic regression

We renamed the data values of the categorical values to be make it suitable for logistic regression. The first five rows are shown below:

	Age	Sex_M	Chest_pain	Resting_BP	Cholesterol	Fasting_blood_sugar	Resting_ECG	Max_Heart_Rate	Exercise_Induced_Angina	ST_depression	ST_slope	Vessels_coloured_flouroscopy	Thalassemia	Target
0	63	Male	Asymptomatic	145	233	Yes	Normal	150	No	2.3	Absent	0	1.0	1
1	37	Male	Non-anginal pain	130	250	No	ST Abnormality	187	No	3.5	Absent	0	2.0	1
2	41	Female	Atypical angina	130	204	No	Normal	172	No	1.4	Downsloping	0	2.0	1
3	56	Male	Atypical angina	120	236	No	ST Abnormality	178	No	0.8	Downsloping	0	2.0	1
4	57	Female	Typical angina	120	354	No	ST Abnormality	163	Yes	0.6	Downsloping	0	2.0	1

Figure 7: Changing the data values for categorical variables (screenshot from code)

4.6 Final dataset after cleaning

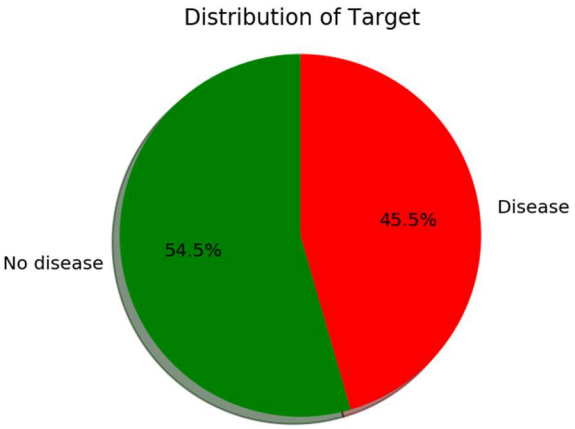
The first five rows of the final dataset after cleaning is shown in figure 7.

5) Data Analysis

In this section, we analyze all the variables and make comparisons through visual graphs and tables. The analysis has been divided into three categories- Univariate, bivariate and multivariate.

5.1 Univariate analysis

In univariate analysis, we analyze the number of people with heart disease (target= Response variable)



	Number of people	Percentage of disease
No Disease	137	54.49%
Have Disease	164	45.51%

Figure 8: Count and percentage of people with heart disease

From the figure 8- pie chart and table we can see that the count of people in the dataset having heart disease are more than those with no heart disease.

This answers question 1 of section 3.2 of the report

5.2 Bivariate analysis

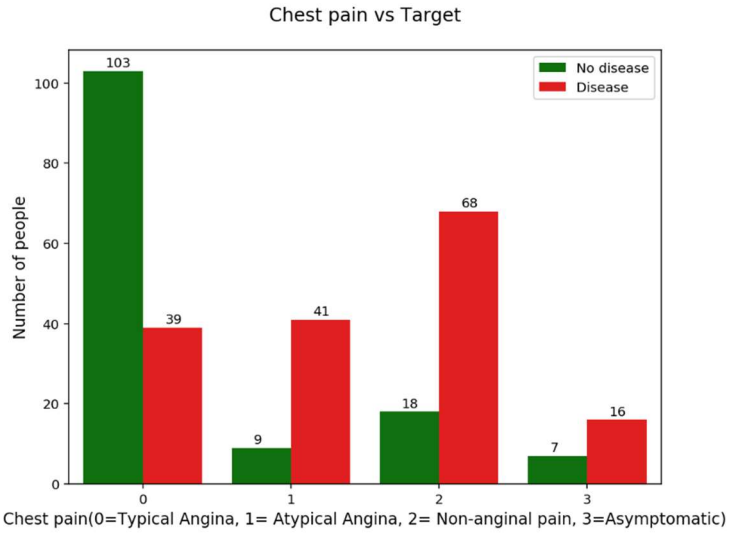
We have divided bivariate analysis in two sections depending on variable being categorical and numerical.

5.2.1 Categorical variable analysis

In this section, we have performed bivariate analysis for response variable -Target and each of the categorical variables- sex, chest pain, fasting blood sugar, resting ECG, exercise induced angina, slope, number of major vessels, thalassemia (8 variables)

This section answers question 2 of section 3.2 of the report

5.2.1.1 Chest pain and target



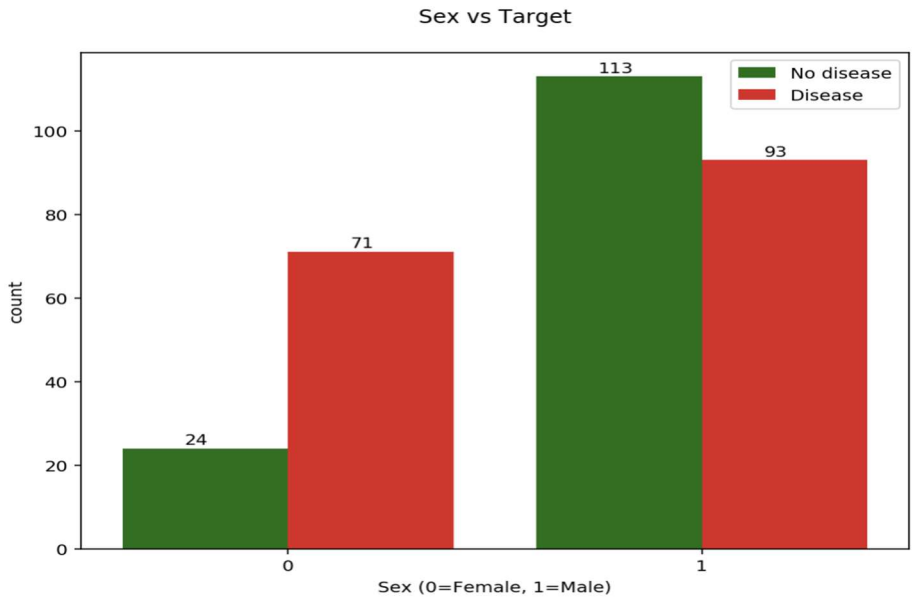
	Average_disease	Count of people
Typical Angina	27.46%	142
Atypical Angina	82.0%	50
Non-anginal pain	79.07%	86
Asymptomatic	69.57%	23

Figure 9: Comparing chest pain types to target (screenshot from code)

The following inferences can be made by looking at figure 9:

- People having Atypical angina, Non-anginal pain, Asymptomatic have more number of people with heart disease as compared to number of people do not have.
- For Typical angina (0), count for people having heart disease are much lower than with no disease.
- Comparatively, people suffering from Atypical angina have a higher chance of heart disease compared to the other three categories.
- Similarly, people suffering from Typical angina have the least chance of heart disease compared to the other three cheat pain categories.
- There is good disparity in avg disease rate between the different types of chest pain. The number of observations is distributed well between the four types. Chest pain is a good predictor variable and shall be included in the final model.

5.2.1.2 Sex and target



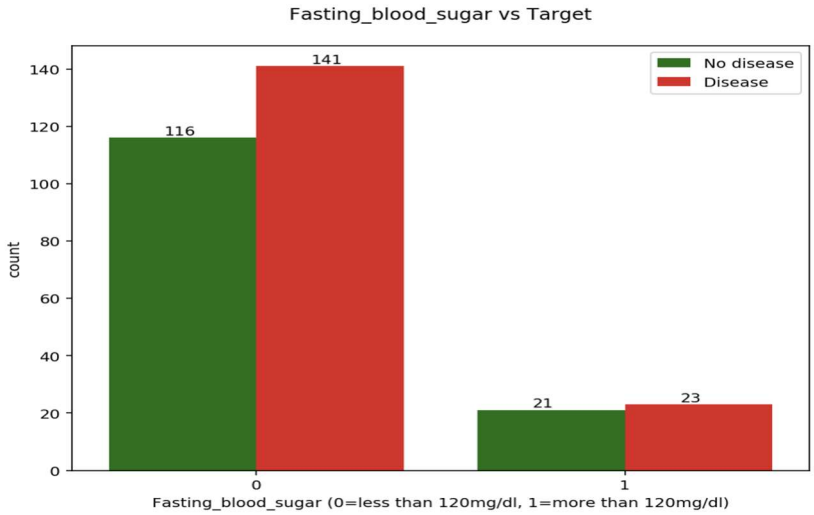
	Average_disease	Count of people
Females	74.74%	95
Males	45.15%	206

Figure 10: Comparing Sex to target(screenshot from code)

The following inferences can be made by looking at figure 10:

- Number of males are more than females
- Though the count of males is more in the dataset, females have a much higher rates of heart disease
- There is good disparity in Average disease rate between males and females. The number of observations is distributed well between males and females. Sex is a good predictor variable and shall be included in our model.

5.2.1.3 Fasting blood sugar and target



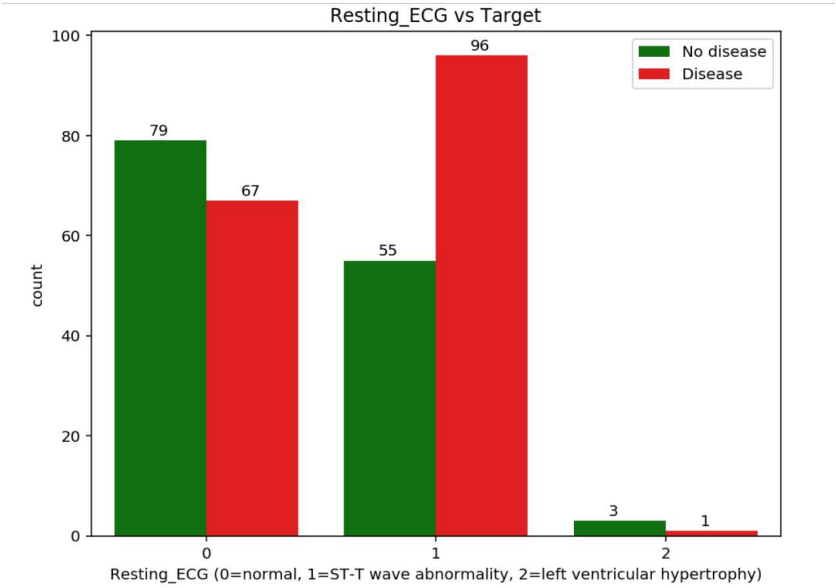
	Average_disease	Count of people
Less than 120mg/dl	54.86%	257
More than 120mg/dl	52.27%	44

Figure 11: Comparing fasting blood sugar to target(screenshot from code)

The following inferences can be made by looking at figure 11:

- People having a normal fasting sugar level are much more in number that those having higher fasting sugar level.
- The ratio of disease versus no disease is higher for people having a lower fasting sugar level.
- Fasting blood sugar doesn't seem to help us predict heart disease based on the blood sugar level. Most of the observations are in the first category. Hence, fasting_blood_sugar is not a good predictor and will not be included in the model.

5.2.1.4 Resting_ECG vs Target



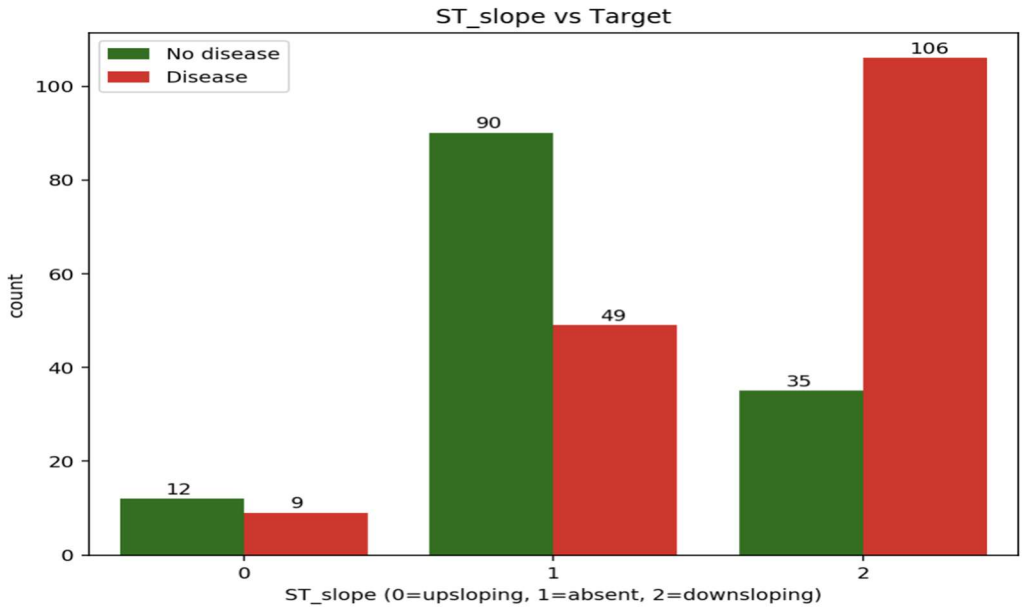
	Average_disease	Count of people
Normal	45.89%	146
ST-T wave abnormality	63.58%	151
Left ventricular hypertrophy	25.0%	4

Figure 12: Comparing Resting ECG to target (screenshot from code)

The following inferences can be made by looking at figure 12:

- People having ST-T wave abnormality have a higher count of diseased individuals as compared to those with no heart disease.
- Most number of observations are present in Normal and ST-T wave category
- There is good disparity in avg disease rate between the different types of Resting ECG. However, only 4 observations are present in the last category of Resting_ECG. We shall throw the variable in the standby pile and make a conclusion after analyzing all our variables.

5.2.1.5 ST slope and target



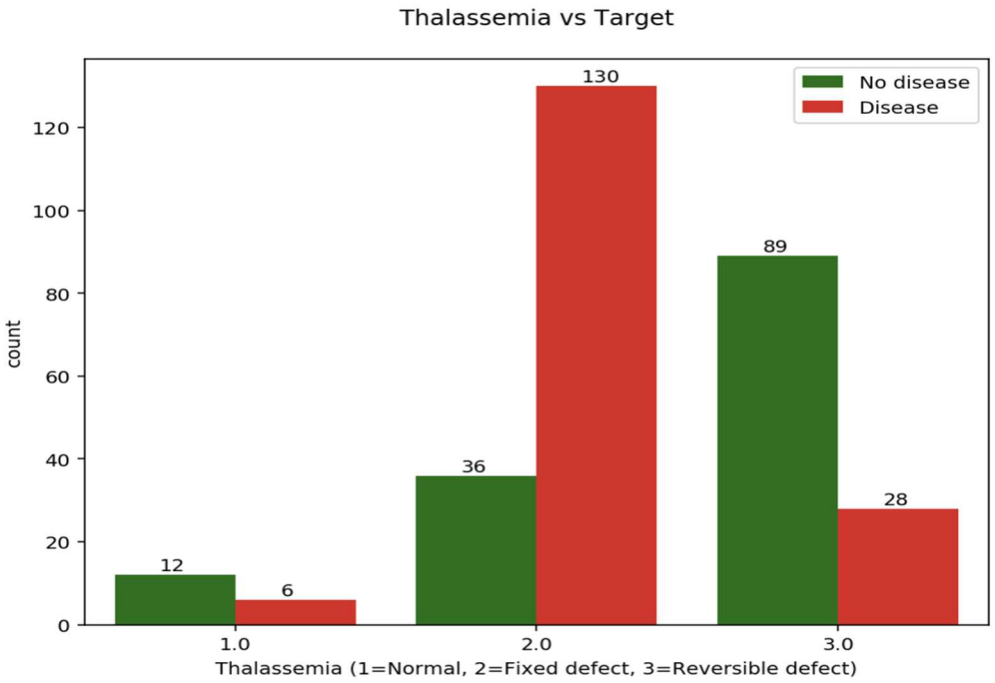
	Average_disease	Count of people
Upsloping	42.86%	21
Absent	35.25%	139
Downsloping	75.18%	141

Figure 13: Comparing ST slope to target(screenshot from code)

The following inferences can be made by looking at figure 13:

- In ECG patients who have a down sloping have a very high count of diseased individuals.
- Wherever the ST_slope shows a flat line, in those cases chances for heart disease are pretty low.
- When the slope of the ECG is downslope, the percentage of heart diseases is high.
- There is good disparity in avg disease rate between the different types of ST_Slope. The number of observations is distributed reasonably well between the three types. ST_slope is a good predictor variable and shall be included in our model.

5.2.1.6 Thalassemia and Target



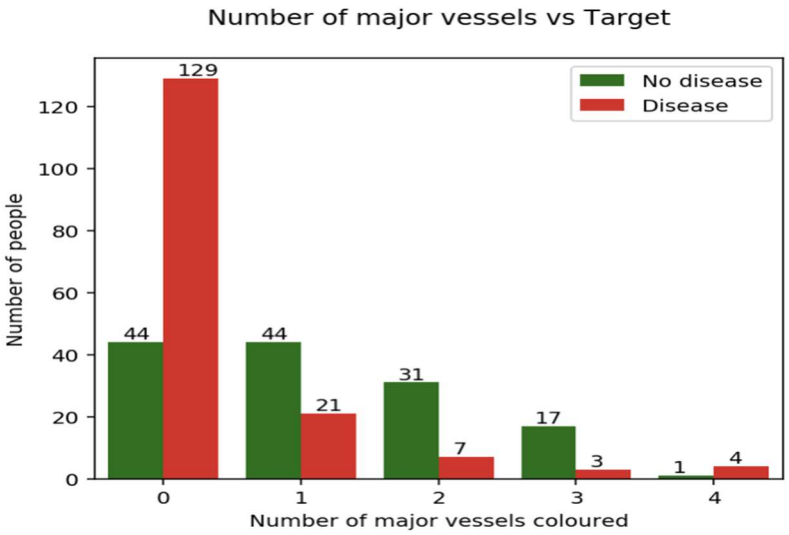
	Average_disease	Count of people
Normal	33.33%	18
Fixed defect	78.31%	166
Reversible defect	23.93%	117

Figure 14: Comparing Thalassemia to target(screenshot from code)

The following inferences can be made by looking at figure 14:

- Patients having a fixed defect are at a very high risk of having heart disease as compared to those with reversible defect and normal.
- There is good disparity in avg disease rate between the different types of Thalassemia. However, only 18 observations are present in the first category of Thalassemia. We shall throw the variable in the “maybe” pile and make a conclusion after analyzing all our variables.

5.2.1.7 Vessels colored fluoroscopy and target



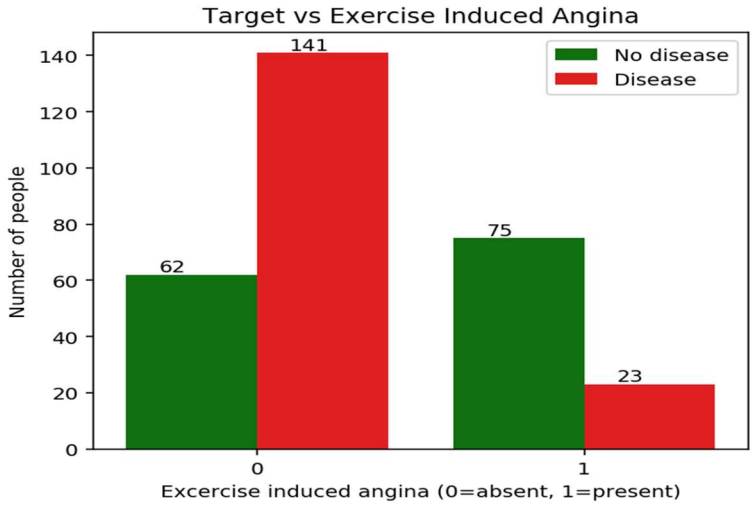
	Average_disease	Count of people
Vessels_coloured_flouroscopy		
0	74.57%	173
1	32.31%	65
2	18.42%	38
3	15.0%	20
4	80.0%	5

Figure 15: Comparing Vessels colored fluoroscopy to target (screenshot from code)

The following inferences can be made by looking at figure 15:

- Vessels colored fluoroscopy represents the number of major vessels. People with no major vessels are the most with a count of 173. They also have the highest rate of heart diseases
- There is good disparity in average disease rate between the different types of number of major vessels (vessels colored fluoroscopy) variable. Number of major vessels is a good predictor variable and shall be included in our model.

5.2.1.8 Exercise induced angina and Target



	Average_disease	Count of people
Angina	69.46%	203
No angina	23.47%	98

Figure 16: Comparing exercise induced angina to target (screenshot from code)

Exercise induced angina refers to the chest pain caused due to excessive exercise

The following inferences can be made by looking at figure 15:

- More number of people have exercise induced angina than people who do not in this dataset.
- Patients having angina due to exercise have a higher chance of having a heart disease as compared to those who do not have angina.
- There is good disparity in avg disease rate between the presence and absence of angina. However, most observations are for presence of angina and only a few for absence of angina. We shall throw the variable in the "maybe" pile and make a conclusion after analyzing all our variables.

5.2.2 Numerical variable analysis

In this section, we shall perform bivariate analysis for response variable -Target and each of the numerical variables-age, resting blood pressure, cholesterol, Maximum heart rate, ST depression (5 variables)

This section answers question 3 of section 3.2 of the report

5.2.2.1 Maximum heart rate vs Target

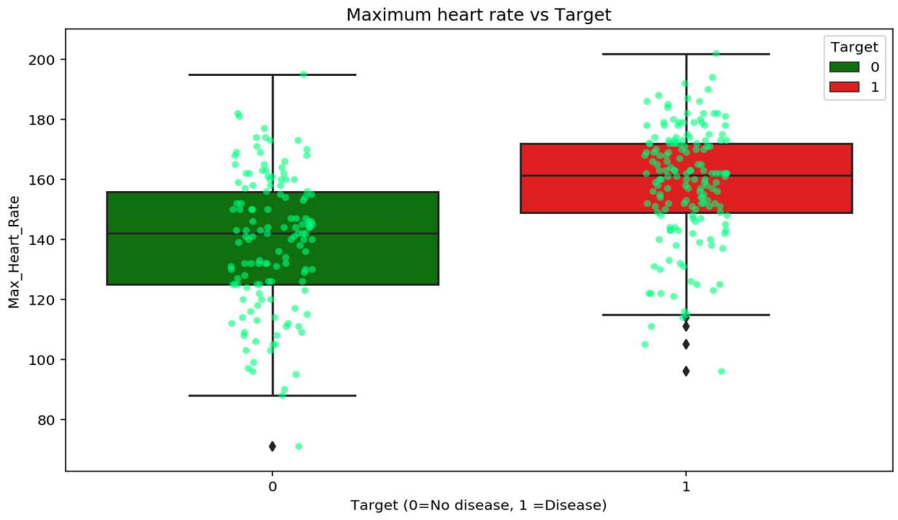


Figure 17: Comparing maximum heart rate and target (screenshot from code)

From figure 17, we can infer that the average heart rate for people with disease is more than the average heart rate of people without disease.

Since figure 17 is not insightful in providing information about the heart disease rate of people with different heart rates. We divided the heart rate into three categories. We divided the heart rate into three groups based on the minimum heart rate from the data set and maximum heart rate from the dataset.

	Percentage of disease	Number of people
Low heart rate	16.67	36
Medium heart rate	52.43	206
High heart rate	84.48	58

Figure 18: Comparing heart rate and heart disease (screenshot from code)

In figure 18, the percentage of disease column represents the number of people with disease in that category divided by total number of people multiplied by 100.The number of people column represents the total number of people in that category.

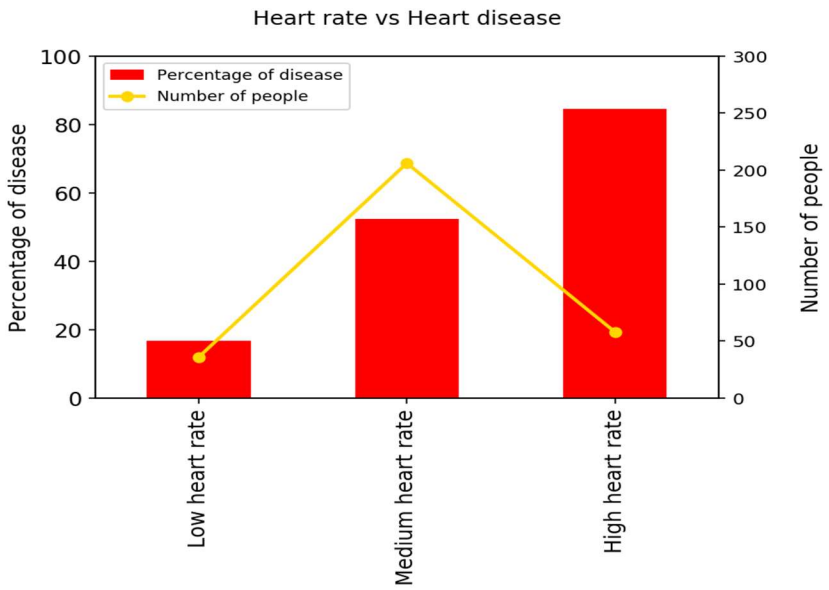


Figure 19: Comparing heart rate and target (screenshot from code)

In figure 19, the X axis represents the heart rate in groups based on figure 18, the primary Y axis represents the percentage of people with heart diseases and secondary Y axis represents number of people.

The following inferences can be made by looking at figure 18 and 19:

- The people with high heart rate have a higher chance of heart disease. It shows a clear pattern of- as the heart rate increases, the chances of heart diseases increase.
- Most of the observations are in the middle range.
- We shall throw the variable in the “maybe” pile for data modelling and make a conclusion after analyzing all our variables.

5.2.2.2 Age and target

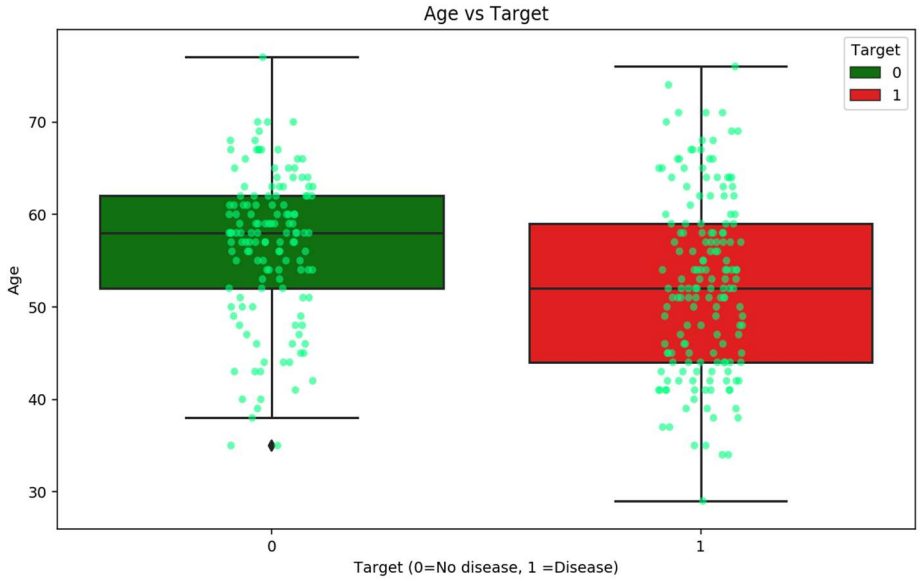


Figure 20: Comparing age and target (screenshot from code)

According to figure 20, most number of people with heart disease are in the middle aged group and older people. People having a heart disease have a lower age average as compared to those who do not have heart disease.

Since figure 20 is not insightful in providing information about the heart disease rate of people in different age groups. We divide the ages into three categories. We divided the ages into three groups based on the lowest age from the data set and highest age from the dataset.

	Percentage of disease	Number of people
Young people	71.43	7
Middle people	69.63	135
Old people	41.06	151

Figure 21: Comparing age and target (screenshot from code)

In figure 21, the percentage of disease column represents the number of people with disease in that category divided by total number of people multiplied by 100.The number of people column represents the total number of people in that category.

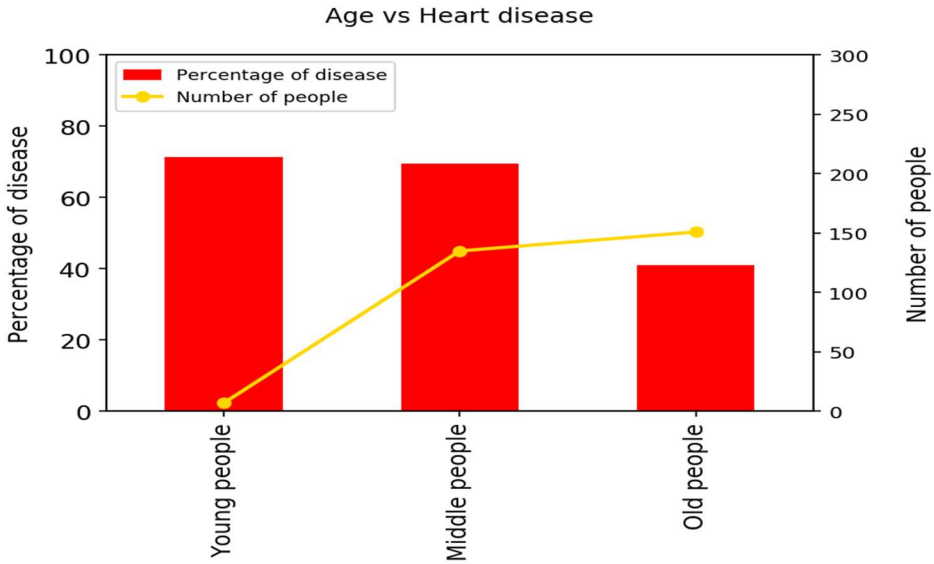


Figure 22:Comparing age and target (screenshot from code)

In figure 22, the X axis represents the age groups based on figure 21, the primary Y axis represents the percentage of people with heart diseases and secondary Y axis represents number of people.

The following inferences can be made by looking at figure 21 and 22:

We have distributed the individuals into three segments based on age i.e. young middle and old and the below details shows the count of individuals in each age group. This seems quite strange but as per the data, out of total 7 people in young age group below 35, 71.43 percent of them have heart disease. This is the maximum among all the three groups created above. The graph shows that younger people have a higher chance of heart diseases than older people, but there are just 7 young people. This variable is misleading and we shall not consider it for modelling.

5.2.2.3 Resting blood pressure and target

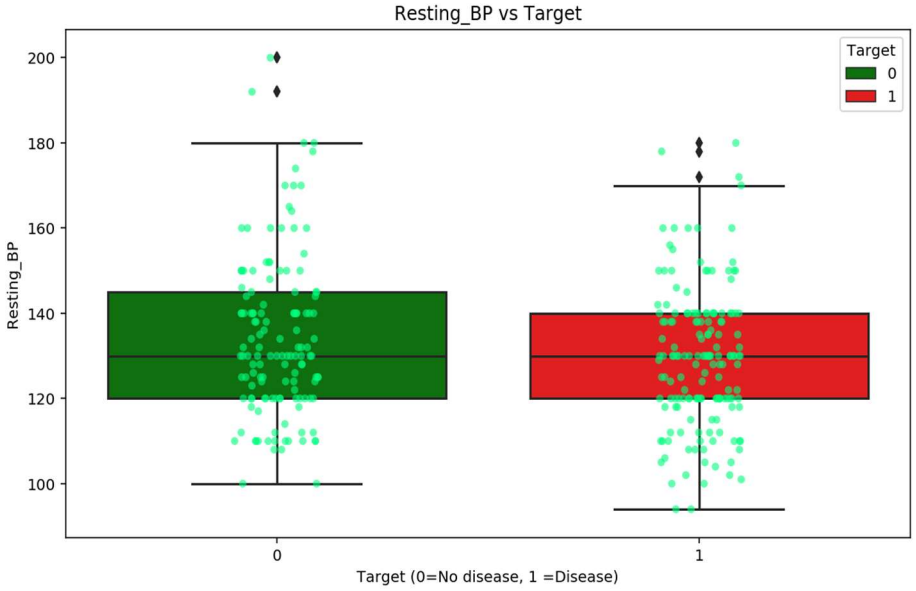


Figure 23: Comparing resting blood pressure and target (screenshot from code)

From figure 23, we cannot differentiate easily as to which level of resting blood pressure has more impact on the response variable since the average blood pressure rate of people with or without heart disease is almost the same.

Since figure 23 is not insightful in providing information about the heart disease rate of people with different blood pressure. We divide the blood pressure rates into three categories. We divided the blood pressures into three groups based on the lowest BP from the data set and highest BP from the dataset.

	Percentage of disease	Number of people
Low BP	66.67	6
Medium BP	58.59	198
High BP	100.00	65

Figure 24: Comparing resting blood pressure and target (screenshot from code)

In figure 24, the percentage of disease column represents the number of people with disease in that category divided by total number of people multiplied by 100.The number of people column represents the total number of people in that category.

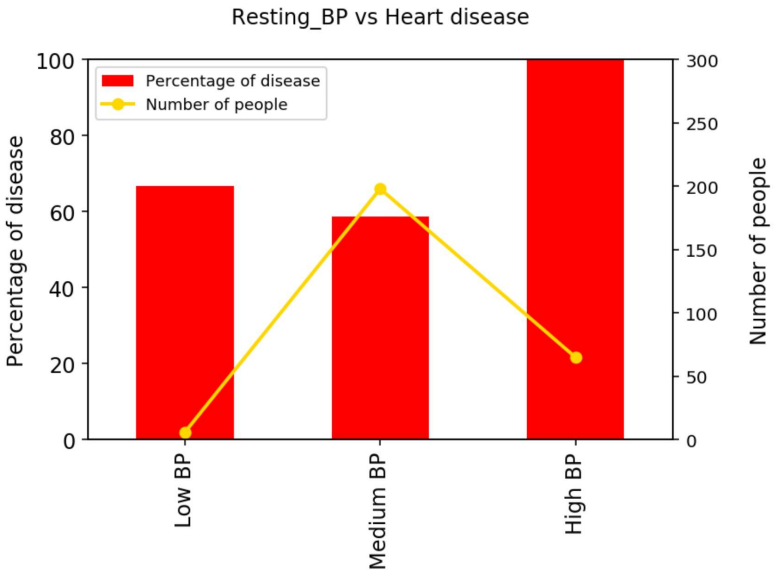


Figure 25: Comparing resting blood pressure and target (screenshot from code)

In figure 25, the X axis represents the age groups based on figure 24, the primary Y axis represents the percentage of people with heart diseases and secondary Y axis represents number of people.

The following inferences can be made by looking at figure 24 and 25:

A total of 6 people are present in the lowest BP category. The graph shows that people with highest blood pressure have a 100% chance of having heart disease. Hence, this seems like a good variable and we shall add it to modelling.

5.2.2.4 Cholesterol vs Target

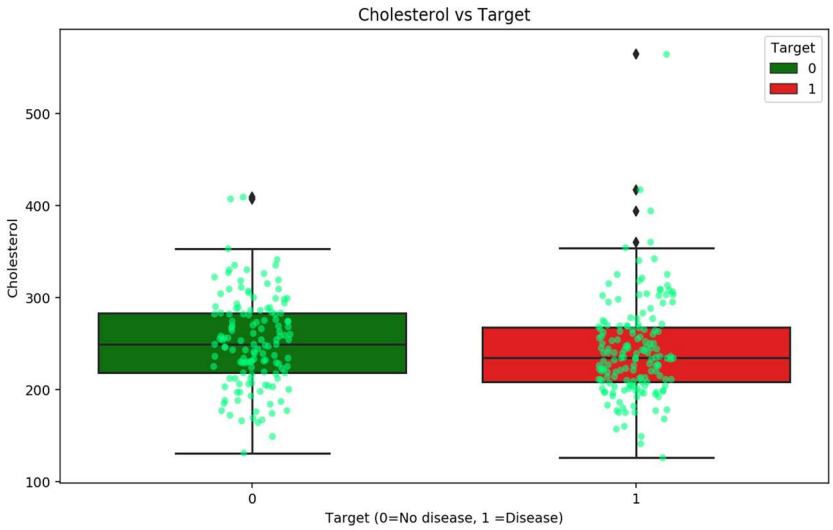


Figure 26: Comparing cholesterol and target (screenshot from code)

From figure 26, we cannot differentiate easily as to which level of cholesterol has more impact on the response variable since the average cholesterol rate of people with or without heart disease is almost the same.

Since figure 26 is not insightful in providing information about the heart disease rate of people with different cholesterol. We divide the cholesterol rates into three categories. We divided the blood pressures into three groups based on the lowest cholesterol from the data set and highest cholesterol from the dataset.

	Percentage of disease	Number of people
Normal Cholesterol	71.43	7
Mildly High Cholesterol	58.04	224
High Cholesterol	41.43	70

Figure 27: Comparing cholesterol and target (screenshot from code)

In figure 27, the percentage of disease column represents the number of people with disease in that category divided by total number of people multiplied by 100.The number of people column represents the total number of people in that category.

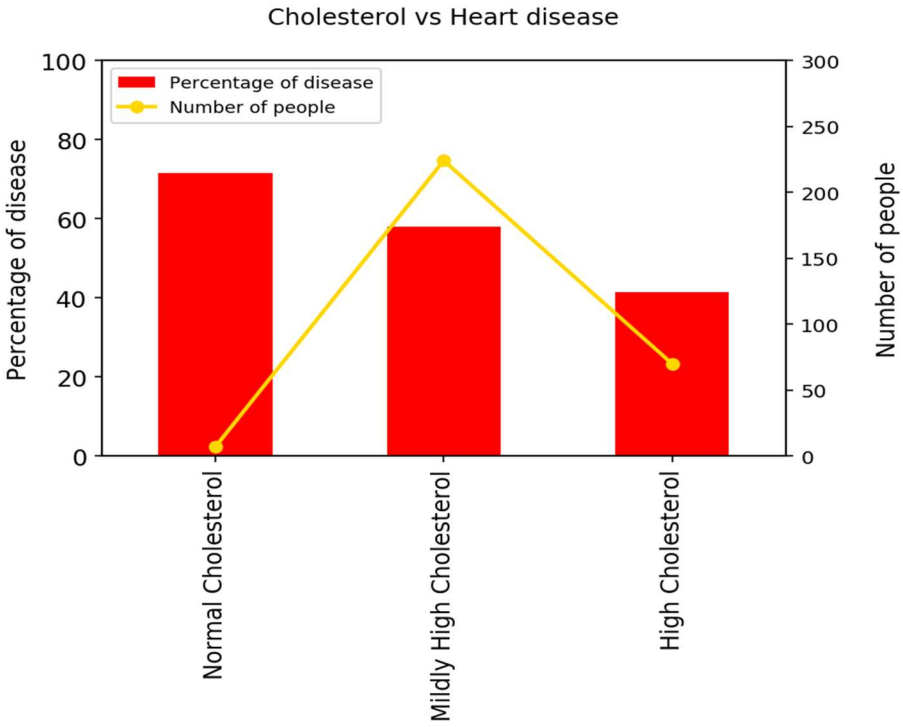


Figure 28: Comparing cholesterol and target (screenshot from code)

In figure 28, the X axis represents the cholesterol levels based on figure 28, the primary Y axis represents the percentage of people with heart diseases and secondary Y axis represents number of people.

The following inferences can be made by looking at figure 27 and 28:

The graph shows that normal cholesterol have a higher chance of heart diseases than mildly high and high cholesterol, but most of the observations are distributed to the "Mildly high" range cholesterol. Cholesterol is not a good predictor and we shall not consider it for modelling.

5.2.2.5 ST depression and target

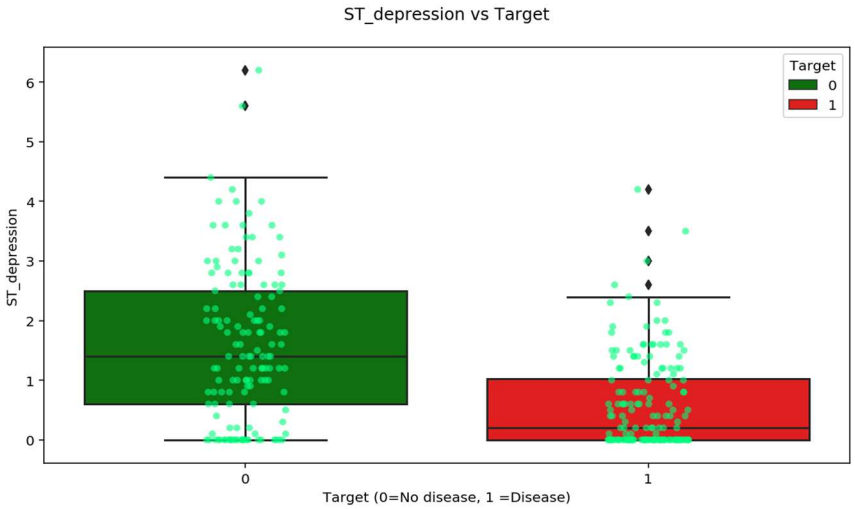


Figure 29: Comparing ST depression and target (screenshot from code)

From figure 29, we can see that people with higher ST depression have a higher risk of heart disease.

Since figure 29 is not insightful in providing information about the heart disease rate of people with different ST depression. We divide the ST depression rates into three categories. We divided the ST depression rates into three groups based on the lowest ST depression from the data set and highest ST depression from the dataset.

	Percentage of disease	Number of people
Normal ST	10.00	40
Mild ST	53.21	156
Fatal ST	73.33	105

Figure 30: Comparing ST_depression and target (screenshot from code)

In figure 30, the percentage of disease column represents the number of people with disease in that category divided by total number of people multiplied by 100.The number of people column represents the total number of people in that category.

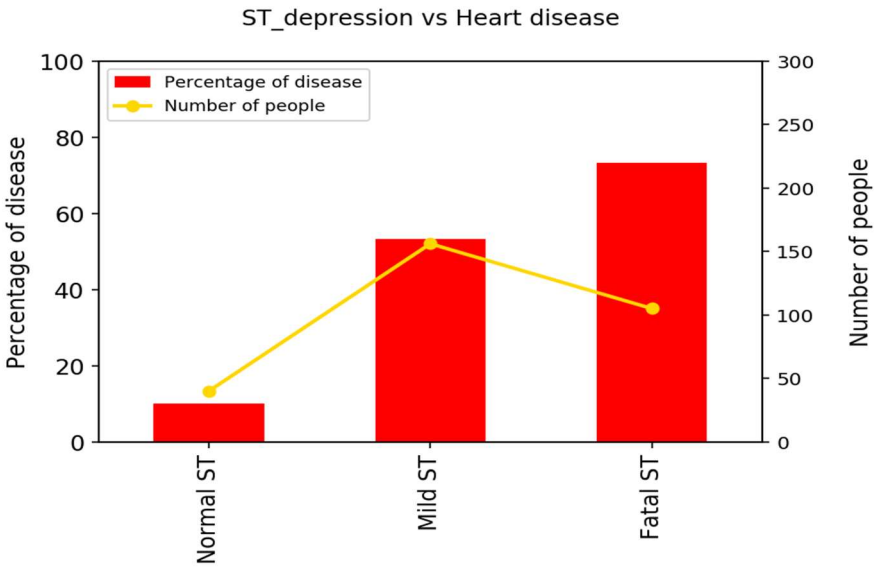


Figure 31: Comparing ST_depression and target (screenshot from code)

In figure 31, the X axis represents the ST depression levels based on figure 20, the primary Y axis represents the percentage of people with heart diseases and secondary Y axis represents number of people.

The following inferences can be made by looking at figure 30 and 31:

The above graph shows that people with high ST_depression have a higher chance of heart disease. It shows a clear pattern of- as the ST depression rate increases, the chances of heart diseases increase. The observations are reasonably distributed between the three ranges. Hence, ST depression is a good predictor and we shall use it for modelling.

5.2.3 Result of Bivariate analysis:

This section address Q4 of section 3.2

We have classified the variables as “good”, “bad” and “maybe good” is based on its ability to predict if a person has heart disease or not.

Considering observations and results produced by performing bivariate analysis, we have grouped predictors as “Good Predictor”, “Bad Predictor” and “Maybe a good predictor” as seen in figure 32.

	Good	Bad	Maybe
0	Chest_pain	Fasting_Blood_Sugar	Maximum heart rate
1	Sex	Age	Cholesterol
2	ST_Slope	-	Thalassemia
3	ST_depression	-	Resting_ECG
4	Number of major vessels	-	Excercise_Induced_Angina

Figure 32: Table summarizing good, bad and maybe good predictors.

5.3 Correlation and heat map

Correlation is the method of analyzing the dependability of two variables with each other.

1. Positive correlation would mean that both the variables are moving in the same direction
2. Neutral correlation : No relationship in the change of the variable
3. Negative correlation would mean that the variables are moving in the opposite direction

If any two variables have high positive or high negative correlation, the model performance will be reduced. If one variable has high positive or negative correlation with the response variable, the model performance will increase. The correlation heatmap will be one of the factors for model selection.

Now, to find the top five predictors which are correlated to the response variable "Target", we construct heatmap:

Figure 33: Heat map (screenshot from code)

From the above figure 33 matrix we observe that the top 5 predictors are :

(The below answers Q5 of section 3.2 of this report)

1. Max_Heart_rate: having a correlation of 0.42
2. Chest pain: having a correlation of -0.43
3. Exercise induced angina: having a correlation of -0.44
4. ST depression: having a correlation of -0.43
5. Vessels colored fluoroscopy: Having correlation of -0.39

As, Maximum heart rate, Thalassemia, Exercise Induced Angina has good correlation with target, so we shall move the variable from "Maybe" to "Good"

	Good	Bad	Maybe
0	Chest_pain	Fasting_Blood_Sugar	Cholesterol
1	Sex	Age	Resting_ECG
2	ST_Slope	-	-
3	ST_depression	-	-
4	Number of major vessels	-	-
5	Excercise_Induced_Angina	-	-
6	Thalassemia	-	-
7	Maximum heart rate	-	-

Figure 34: Table summarizing good, bad and maybe good predictors.

5.4 Multivariate analysis

In this case, we shall perform analysis for three or more variables.

5.4.1 ST depression , ST slope and Target

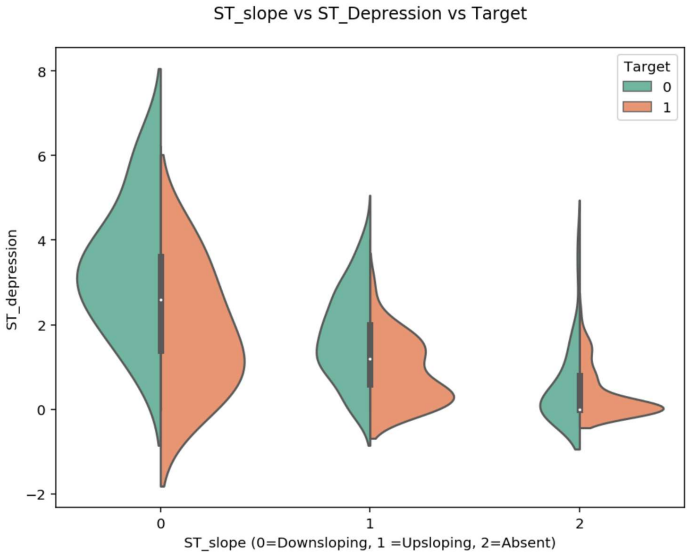


Figure 35 : Comparing ST slope, ST depression and target

The following inference can be made from figure 35,

This section address Q6 of section 3.2

- We can see that people having lower ST depression from 0 to 4 and with down sloping have higher chance of having a heart disease.
- The average level of ST depression corresponding to the target type keeps decreasing for every ST slope level.

5.4.2 Sex, fasting blood sugar and target

Figure 36: Sex, fasting blood sugar and target

The following inference can be made from figure 36,

This section address Q7 of section 3.2

- We can see that females having lower fasting blood sugar have higher chances of heart diseases.
- More number of males and females with and without heart disease have low fasting blood sugar

5.4.3 Maximum heart rate, Age and Angina

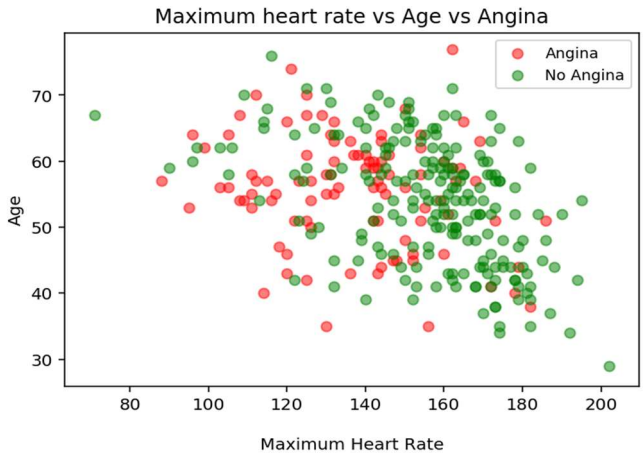


Figure 37: Maximum heart rate, age and angina

The following inference can be made from figure 37,

This section address Q8 of section 3.2

- We observe that people in the age group 45-60 and within max heart rate range 145-185 have no angina.
- People with higher heart heart rate have more changes of heart pain compared to people with lower heart rate.

5.4.4 Maximum heart rate, Age, Resting BP and Target

Figure 38: Maximum heart rate, age , resting BP and target

The following inference can be made from figure 38,

This section address Q9 of section 3.2

- People with heart disease have a higher heart rate.
- Most of the patients having heart disease start from heart rate of 140 and above.
- Circles of large size i.e. having more resting BP are mostly green which means people with mid-range(120-130) resting BP do not have heart disease.

6) Data modelling - Logistic Regression

In this section we would run the Logistic Regression on all the numerical as well as categorical variables by first converting them into dummy variables and then create a dataframe to add the coefficient values and also convert the coefficient values into exponential form.

Below mentioned are the results of Logistic Regression:

	Coefficient	Odd_ratios
Intercept	0.330410	1.391539
Resting_BP	-0.005365	0.994649
Cholesterol	-0.002251	0.997752
Max_Heart_Rate	0.032993	1.033543
ST_depression	-0.378064	0.685187
Vessels_coloured_flouroscopy	-0.758033	0.468587
Thalassemia	-0.720582	0.486469
Sex_M_Male	-1.372085	0.253578
Chest_pain_Atypical angina	-0.106021	0.899405
Chest_pain_Non-anginal pain	0.625917	1.869961
Chest_pain_Typical angina	-1.038400	0.354021
Resting_ECG_Normal	-0.097494	0.907108
Resting_ECG_ST Abnormality	0.409967	1.506768
Exercise_Induced_Angina_Yes	-0.660569	0.516557
ST_slope_Downsloping	0.402236	1.495164
ST_slope_Upsloping	-0.252875	0.776565
Fasting_blood_sugar_Yes	0.065380	1.067564

Figure 39: Logistic Regression

The following inference can be made from figure 39,

- The odds of person having heart disease increases by a factor of 1.854476 when the chest pain is non angina rather than typical angina
- Chance of a heart disease increase by a factor of 1.489851 if the resting ECG has a ST abnormality rather than normal
- For one unit increase in Max Heart rate, chances of having a heart disease increase by 0.032993
- Chance of a heart disease increase by a factor of 1.489851 if the resting ECG has a ST abnormality rather than normal
- The odds of person having heart disease increases by a factor of 1.067564 when the fasting blood sugar is more than 120mg/dl.
- Chance of a heart disease increase by a factor of 1.495164 if the ST_slope has a downsloping rather than normal

7) Data Modelling- Model comparison

In this section, based on the analysis of all the variables in the data analysis section, we would create different models and select the best among them which can predict in a patient chances of having a heart disease. We divided our dataset into 80 percent training and 20 percent validation.

With reference to Fig. 34 on page 31 we have created three models:

7.1 Model I

This model contains all the "Good" variables only i.e Chest_pain, Sex_M, ST_slope, ST_depression, Vessels_coloured_flouroscopy, Exercise_Induced_Angina, Max_Heart_Rate.

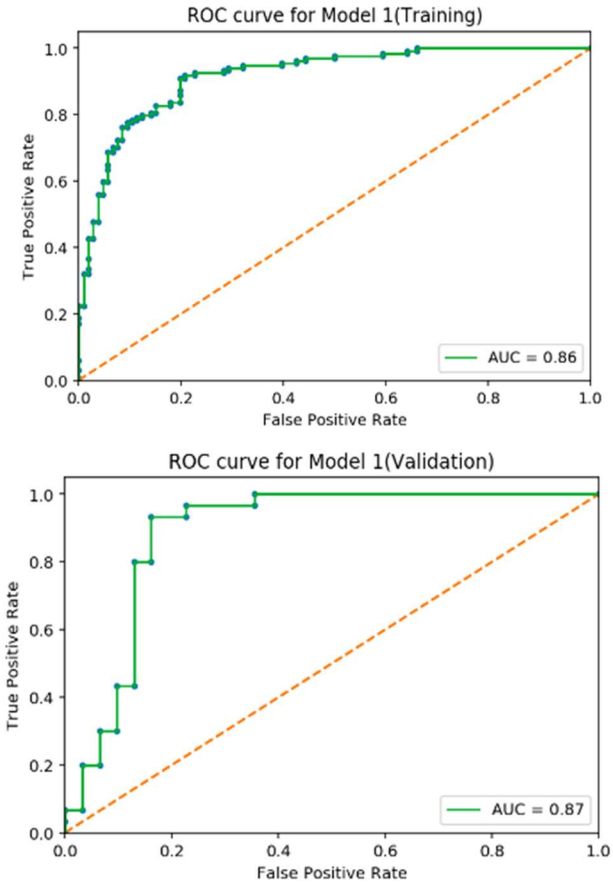


Figure 40: Model I

The following inference can be made from figure 39,

- Accuracy rate for Training dataset is 86 percent
- Accuracy rate for Validation dataset is 87 percent

7.2 Model II

This model contains all the “Good” plus “May be Good” variables i.e Chest_pain, Sex_M, ST_slope, ST_depression, Vessels_coloured_flouroscopy, Exercise_Induced_Angina, Max_Heart_Rate, Resting_ECG, Cholesterol.

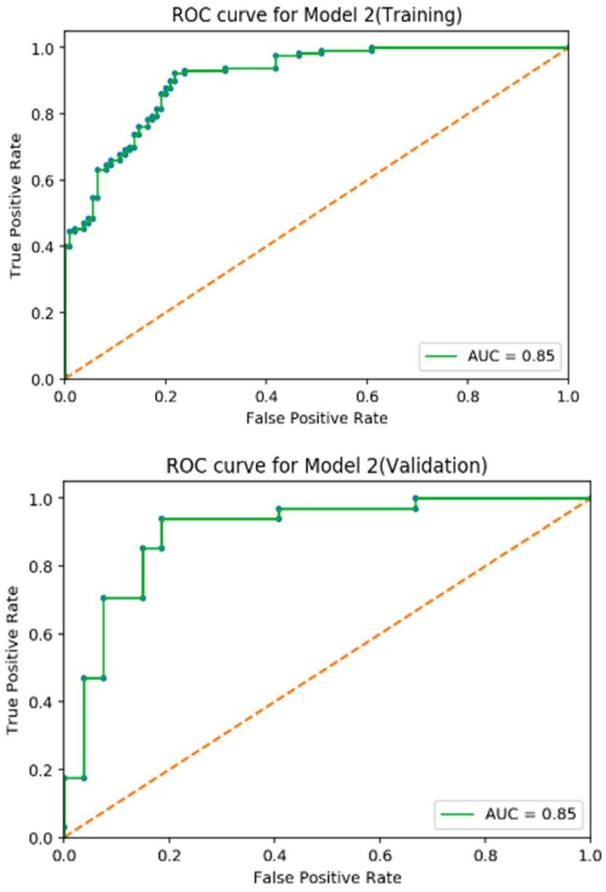


Figure 41: Model II

The following inference can be made from figure 40,

- Accuracy rate for Training dataset is 85 percent
- Accuracy rate for Validation dataset is also close to 85 percent

7.3 Model III

This model contains all the “Good”, “May be Good” and ”Bad” variables i.e Chest_pain, Sex_M, ST_slope, ST_depression, Vessels_coloured_flouroscopy, Exercise_Induced_Angina, Max_Heart_Rate, Resting_ECG, Cholesterol, Fasting_Blood_Sugar and Age.

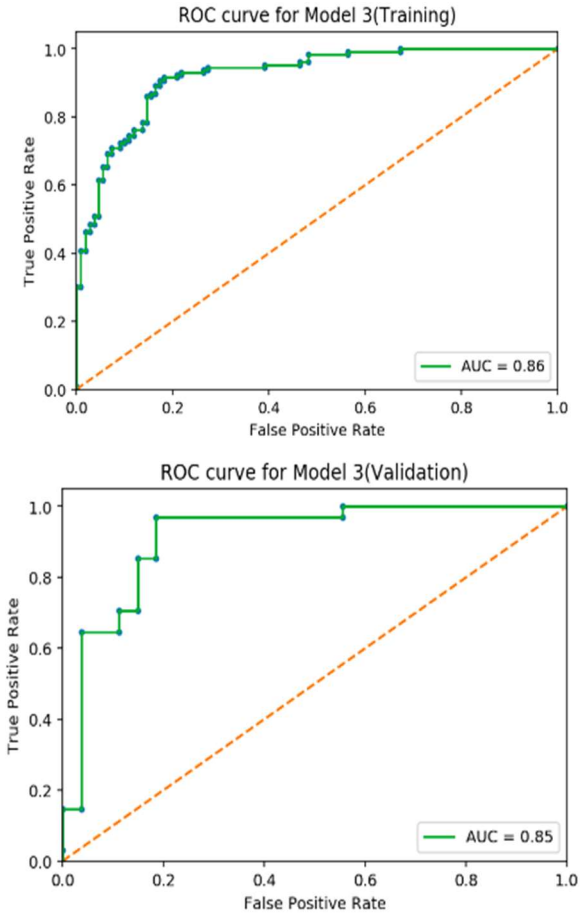


Figure 42: Model III

The following inference can be made from figure 41,

- Accuracy rate for Training dataset is close to 86 percent
- Accuracy rate for Validation dataset is 85 percent

7.4 Conclusion of modelling:

- Out of all the three models, “Model I“ gives us the best results for both training as well as the validation dataset.
- Which means that the variables under the ‘Good’ category are only required to predict the chances of a patient having a heart disease or not.
- When we try to add more variables to the model, then it gets over fitted and it was observed that the accuracy of the model starts decreasing

8) Final conclusion

From all the above analysis, correlation and data modelling we conclude that according to this dataset:

- Non angina pain has a most impact out of the four types of chest pains as the odds of person having heart disease increases by a factor of 1.854476 when the chest pain is non angina rather than typical angina
- Chance of a heart disease increase by a factor of 1.489851 if the resting ECG has a ST abnormality rather than normal
- Females under middle age are more prone to heart disease
- Downsloping ST slope has the most impact on having a heart disease
- More the ST Depression more are the chances of heart disease

The dataset was limited to 301 observations for a limited location – Cleveland hospital, Hence, any prediction and analysis was restricted to limited range of data for each variable.