EDA Final Project

Raj Thakkar, Abhilash Kuhikar, Dhruuv Agarwal, Darshan Shinde

April 27, 2019

Introduction to the problem:

Based on a report published by Centers for Disease Control and Prevention, about 610,000 people die of heart disease in the United States every year; that's 1 in every 4 deaths. Moreover, more than half of the deaths due to heart disease in 2009 were in men.

Men are usually more prone to heart diseases as compared to women. So in this project we wanted to focus on males and females separately in order to understand how does the impact of variable differ in men and women.

Description of the data

The data we are using in this project is Cleveland Heart Disease Data set available from UCI Machine Learning Repository. This database contains 14 attributes for 300 people. We tried to incorporate other Heart Disease Data sets available from UCI Machine Learning Repository but they had many missing values for the desired attributes

The variables and their corresponding description are given below:

Table 1: Data Dictionary

Name	Type	Description
Age	Continuous	Age (in years)
Gender	Discrete	Sex (1 = male; 0 = female)
CP	Discrete	Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
Trestbps	Continuous	Resting blood pressure (in mm Hg on admission to the hospital)
Chol	Continuous	Serum cholestoral in mg/dl
FBS	Discrete	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
RestECG	Discrete	resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
Thalach	Continuous	maximum heart rate achieved
Exang	Discrete	exercise induced angina (1 = yes; 0 = no)
Oldpeak	Continuous	ST depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = downsloping)
CA	Continuous	Number of major vessels (0-3) colored by flourosopy
Thal	Discrete	3 = normal; 6 = fixed defect; 7 = reversable defect
Goal	Discrete	The predicted attribute - diagnosis of heart disease (angiographic disease status) (Value 0 = < 50% diameter narrowing; Value 1 = > 50% diameter narrowing)

Research Question: How do the variables explain the likelihood of men and women having a heart disease? How does the effect differ for men and women?

First we need to make corrections to the Goal variable which tells us about the diagnosis of heart disease as it has values 0, 1, 2, 3 and 4. 1, 2, 3 and 4 represent that the person has a heart disease so they will also be encoded as 1 for the ease of analysis.

Correlation value quantifies the magnitude of relationship that exists between a pair of variables. So, we will start by looking at the correlation of Goal variable which is our dependent variable with the independent variables available in the dataset. The plot has been divided into 4 parts (2 for males and 2 for females) as number of variables is too large to be displayed in a single plot (see appendix for pair plots).

For the sake of Exploratory Data Analysis, we will only focus on variables whose absolute value of correlation is greater than 0.45 with the Goal variable. We are doing this to avoid over fitting of the model especially for females. Many other variables were considered in the beginning but later on we decided to go for a simple model for the ease of interpretation.

Thus the variables for males are:

1. Thalach (-0.467)

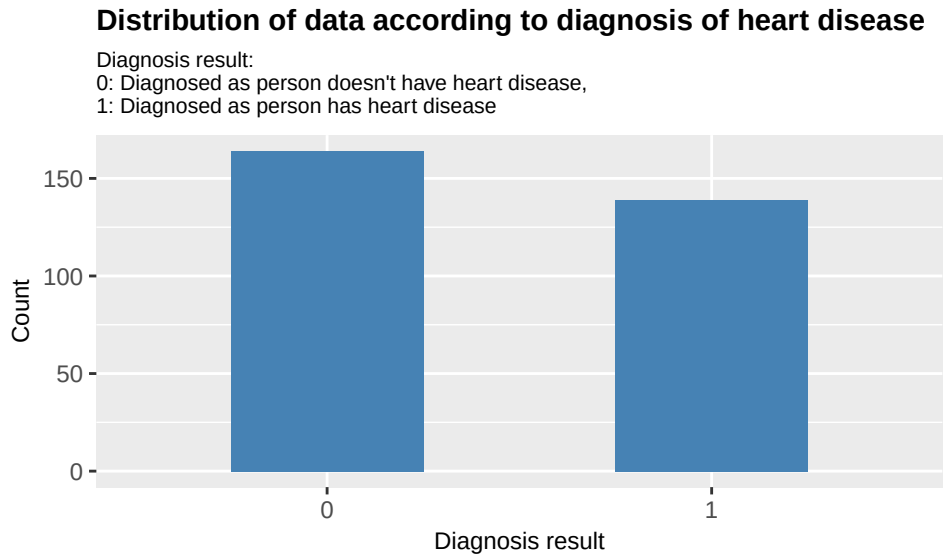
2. Oldpeak (0.456)
3. CA (0.467)

Thus the variables for females are:

1. Oldpeak (0.611)
2. CA (0.644)
3. Thal (0.707)

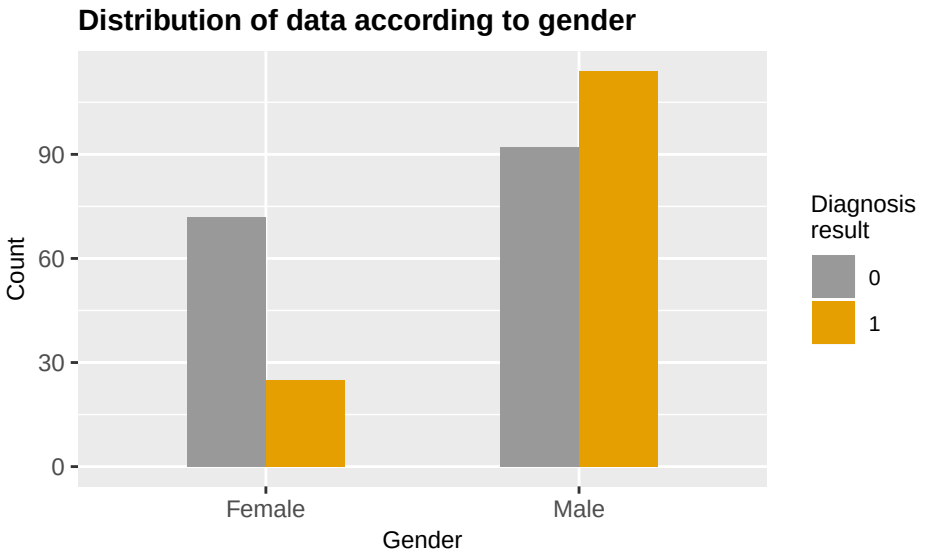
Out of these, only Thalach, Oldpeak and CA are continuous variables so we will do bi variate analysis using these continuous variables. We will also try to determine if there is an interaction between these variables. All the interactions have not been visualized due to the constraint on lengths of this report.

First we will check the distribution of the Goal variable which is 1 if the person has a heart disease and 0 otherwise



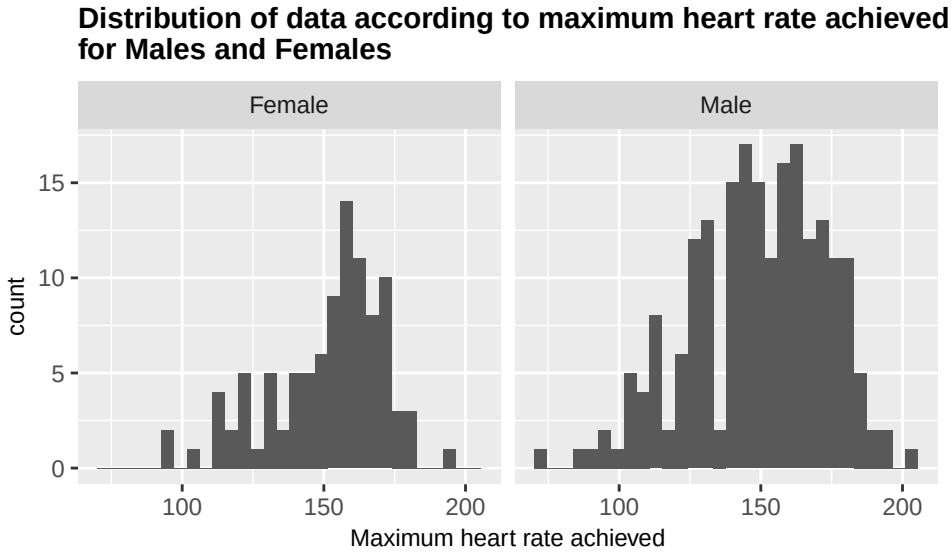
We can see that number of 0's and 1's in the Goal variable are not very different. Thus there is not a huge amount of class imbalance in this dataset.

Now we will look at the distribution of Males and Females in the given dataset. Usually heart disease is more associated with males so chances are that the data was collected primarily for males.

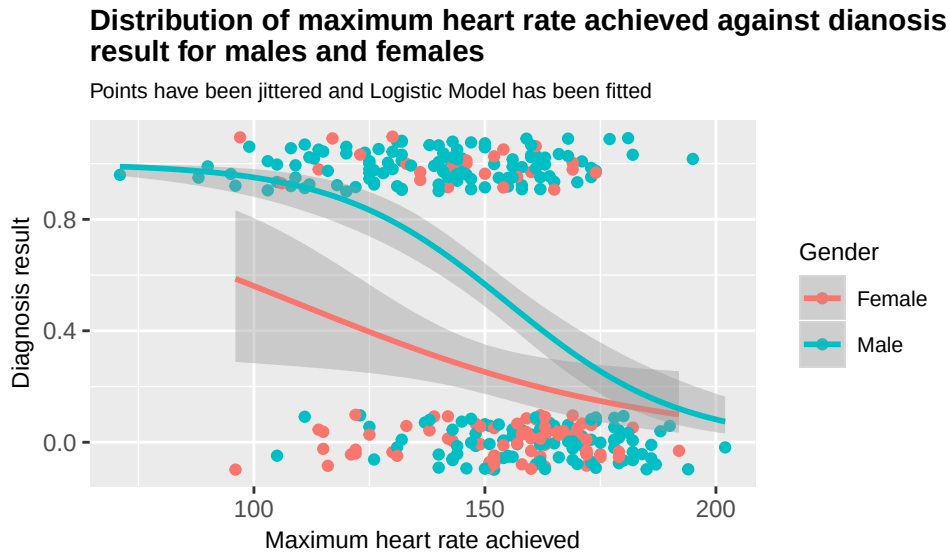


We can see that male participants in the study are almost twice as compared to the female participants. Moreover, percentage of females with heart disease (~25%) is very less as compared to percentage of males with heart disease (~55%) in the given data.

Now we will focus on the variable thalach which tells us about the maximum heart rate achieved by a person. We will first look at the distribution of values of thalach variable



We can see that the distribution is approximately normal and slightly right skewed for both males and females. Since the skew is not huge, we will not be using any transformations for the thalach variable.

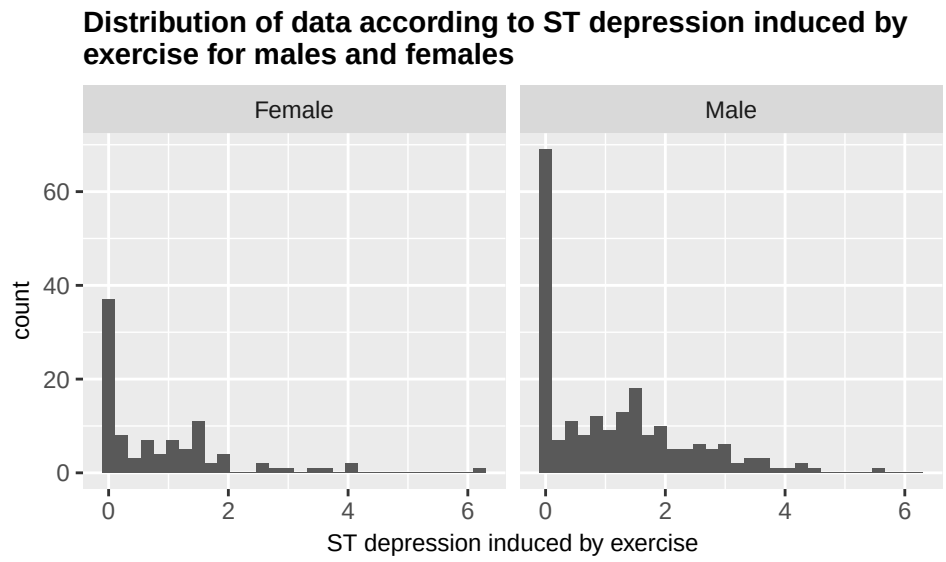


A oerson's maximum heart rate plays a major role in setting your aerobic capacity-the amount of oxygen you are able to consume. Several large observational studies have indicated that a high aerobic capacity is associated with a lower risk of heart attack and death.

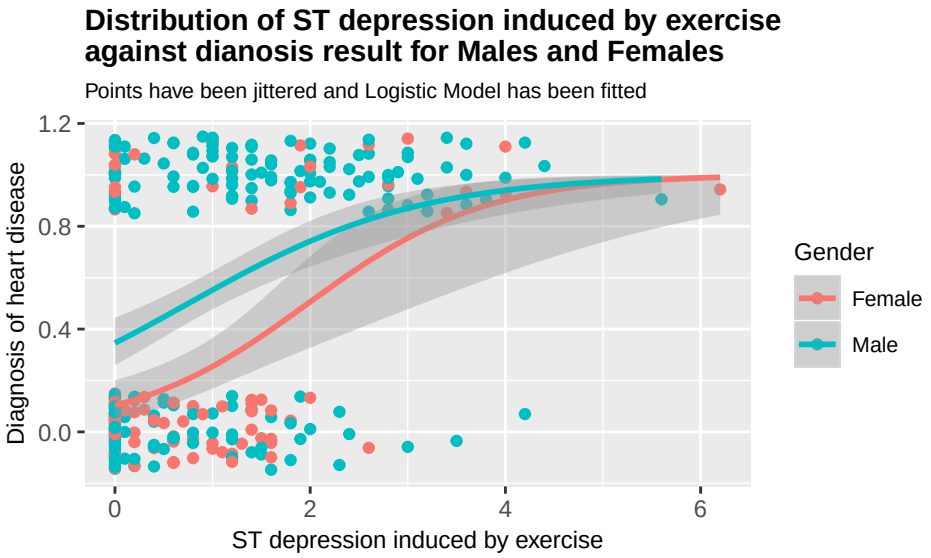
We can see that at the same values of max heart rate(Thalach), men are at higher risk of having a heart disease.

One more important factor to determine if a person has a heart disease is old peak. To understand its meaning we referred to description by Anthony L. Komaroff, MD, an internal medicine specialist (see appendix).

We will first look at the distribution of the variable 'Oldpeak'

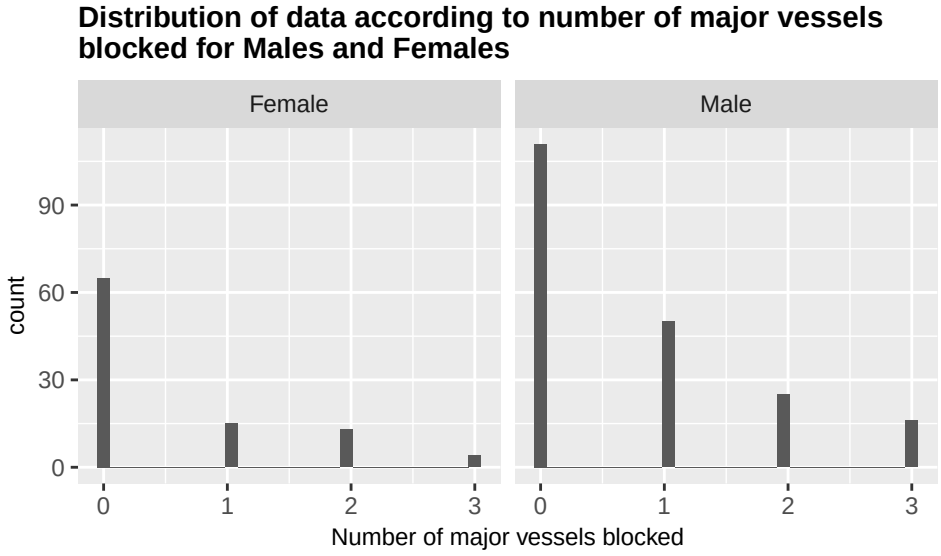


From the graph above, we can see that value of ST depression induced by exercise is 0 for most of the males and females in the dataset. Since this basically means that they don't have any ST depression and taking a log of this variable will lead to a loss of data for most of the participants as log(0) = -Infinity, we will not use any transformation for Oldpeak variable.

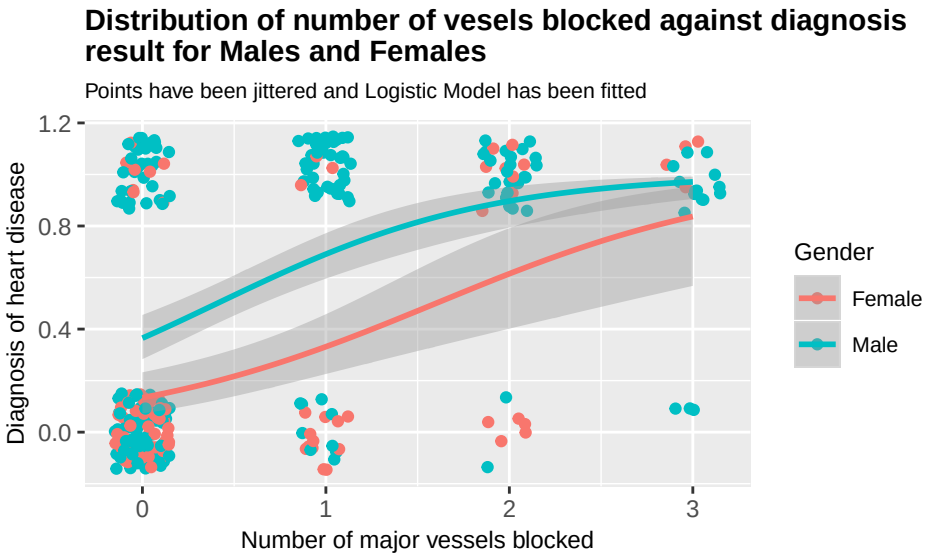


We can see that at low oldpeak values there's a significant difference of probabilities of heart diseases between males and females. The probability increases steadily with the value of oldpeak.

Number of major vessels blocked in a person's heart will definitely help us identify if a person has a heart disease.



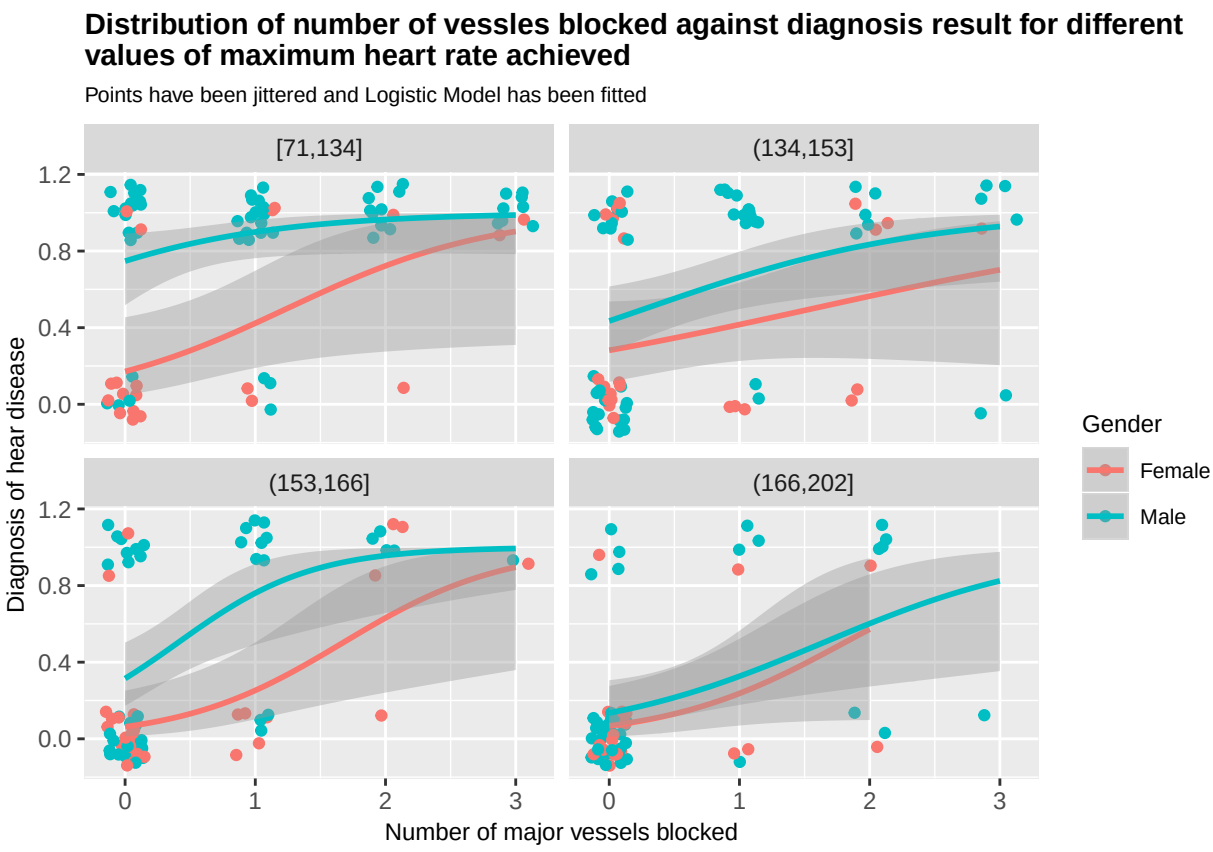
We can see that even though this is a continuous variables, it only takes 4 values i.e. 0, 1, 2 and 3 thus we cannot comment on the distribution of this variable as such.



The probability of Females getting a heart disease is lower than the probability of a Male getting a hear disease when they have equal number of heart vessels blocked. The probability increases steadily with increase in the number of vessels blocked.

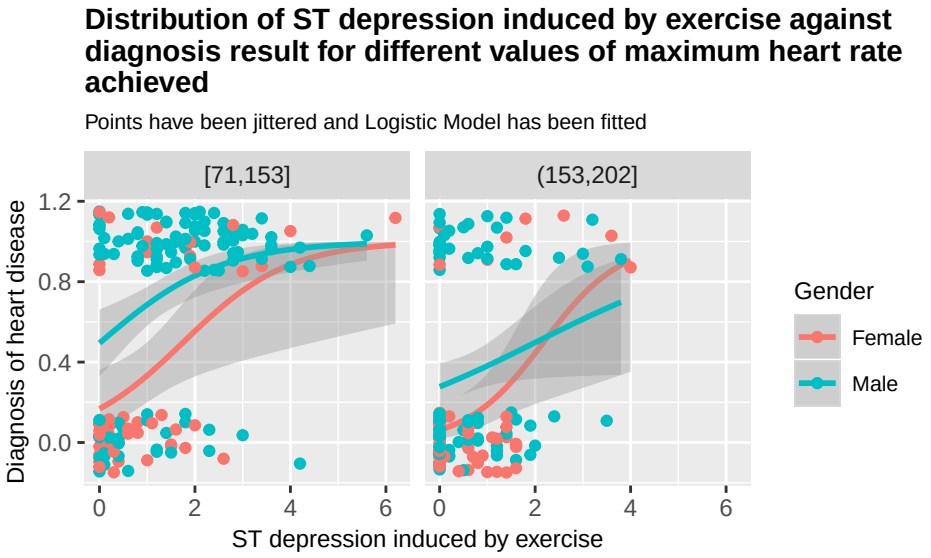
Now we will try to determine if we need an interaction between the 2 set of variables in our model.

Interaction between Thalach and CA:



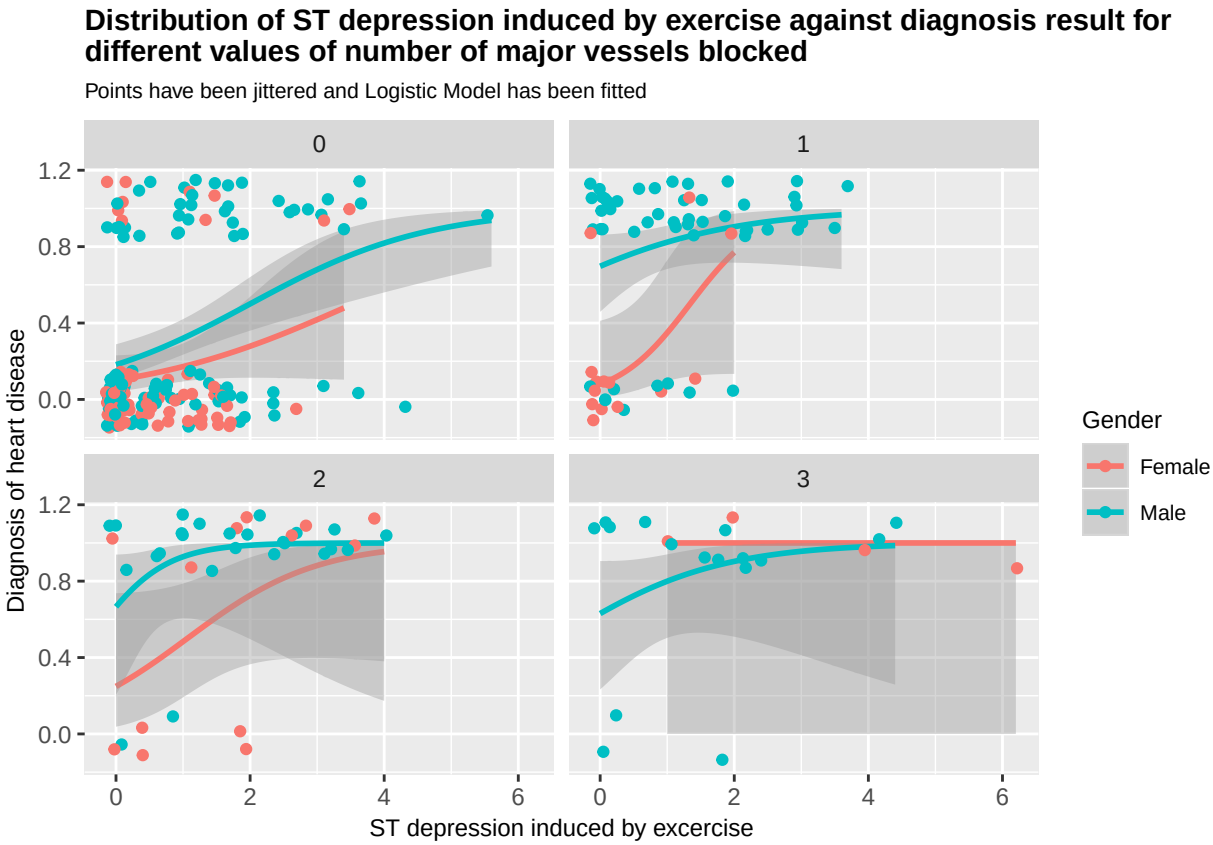
The slope of CA doesn't change a lot for different values of Thalach variable which tells us about the maximum heart rate achieved by an individual in the dataset. Thus we cannot say that there is an interaction between Thalach and CA variable for either males or females.

Interaction between Thalach and Oldpeak:



From the graph above, we can see that slope of oldpeak for males and females don't change a lot with different values of Thalach variable which tells us about the maximum heart rate achieved by an individual in the dataset. Thus we cannot say that there is an interaction between Thalach and CA variable for either males or females.

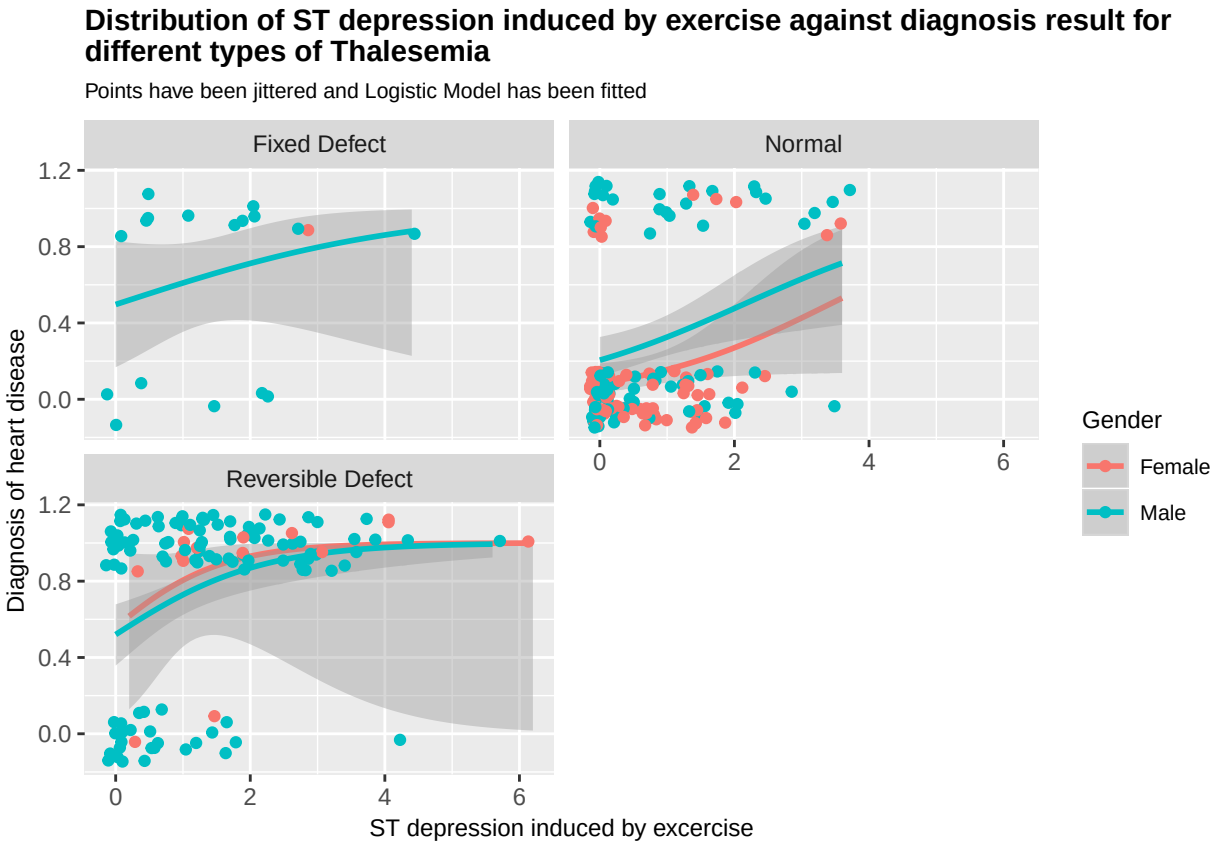
Interaction between Oldpeak and CA:



The slope of Thalch doesn't change a lot for males with different value of CA i.e. the number of major vessels blocked. While the slopes are very different for females with different values of CA or number of major vessels blocked, we see that the slope changes for female population in the graphs, especially when CA is 0 (vessels colored) vs other values of CA.

The number of females with 3 major vessels blocked to number of females with 0 vessels blocked. Thus we cannot confidently say if an interaction between ST depression induced by exercise (Oldpeak) and Number of major vessels blocked (CA) exists for females as it can lead to over fitting.

Interaction between Oldpeak and Thal:



There are not enough data points for females with Fixed defect Thalesemia so we cannot comment on the interaction of Thal and Oldpeak variable for Females. For Males, the slopes are almost the same and thus we cannot say that there is an interaction between Oldepeak and Thal variables.

Thus, we will not be adding any interactions in the model since, we are likely to over fit the model by adding interactions especially for females.

Model Building

We are trying to compare the models fitted for male and female. For that it is required to have same variables in both model. Hence, we are trying to fit a simple additive logit model by considering the union of variables found which are highly correlated for male and female data. So, we are using Thalach, Oldpeak, CA and Thal variables. We still have the issue of over fitting because of less data available for females and hence we dropped the plan of adding any interaction between the variables in model.

Let's try to fit the simple additive model on male data.

```
## [1] "The confusion matrix for males is given below:"
```

```
##
## all.male.pred  0  1
##         pred 0 75 20
##         pred 1 17 94
```

From the confusion matrix above, the accuracy of model fitted for males is:

$$(75 + 94)/(75 + 94 + 20 + 17) = 82.04\%$$

The equation of the model fitted for males is:

$$Pr(HeartDisease) = logit^{-1}(3.78 - 0.04 * Thal + 0.52 * Oldpeak + 1.17 * CA$$
$$-0.33 * ThalNormal + 1.25 * ThalReversibleDefect)$$

Now let's try to fit a normal additive model on female data.

```
## [1] "The confusion matrix for females is given below:"
```

```
##
## all.female.pred  0  1
##          pred 0 69  9
##          pred 1  3 16
```

From the confusion matrix above, the accuracy of model fitted for males is:

$$(69 + 16)/(69 + 16 + 9 + 3) = 87.63\%$$

The equation of the model fitted for males is:

$$Pr(HeartDisease) = logit^{-1}(13.41 - 0.01 * Thal + 0.52 * Oldpeak + 0.84 * CA - 15.11 * ThalNormal$$
$$-12.03 * ThalReversibleDefect)$$

The coefficients of Logistic Regression are difficult to interpret. One way to understand their effects is to look at the Odds Ratio which tells us about the change in odds of a person having a hear disease when independent variable increases by 1 unit. The odds ratio for model fitted for males is given below:

Table 2: Odds Ratio for Males

	OR
Thalach	-3.63
Oldpeak	69.04
CA	221.44
ThalNormal	-28.12
ThalReversible Defect	247.59

From the above table, we get some interesting information. From the data in above table we can say that for every unit of increase in maximum heart rate achieved chances of having heart disease decreases by 3.63 % for males. Unit increment in ST depression induced by exercise leads to 69.03% increase in chances of having heart disease in males. There is huge impact of number of major vessels in which blockage has been observed during fluoroscopy on chances of having heart disease. For every single unit of increment of CA variable, it increase the chances of having heart disease by 221.44%. The odds of male with Normal Thalassemia is 28.11% lower than the male with Fixed defect Thalassemia. Similarly, the odds of a male with Reversible defect Thalassemia are 247.58% greater than the male with Fixed defect Thalassemia.

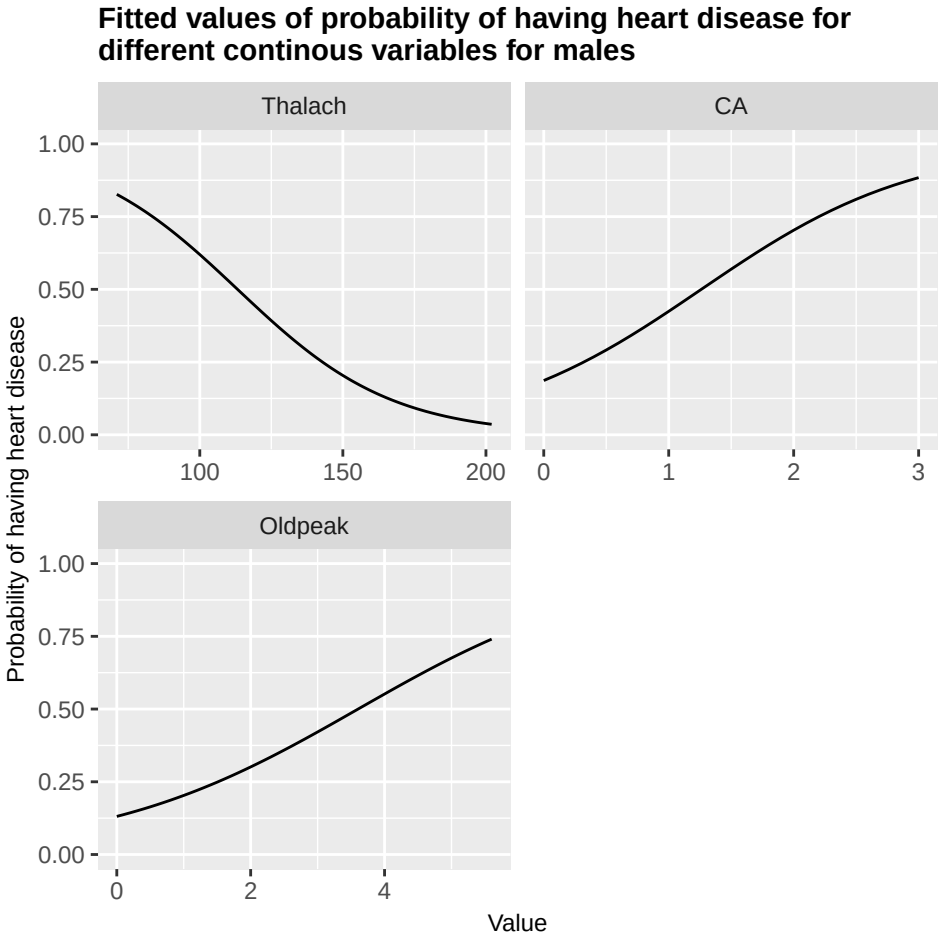
Now let's observe the results we got for female data. We can say that for every unit of increase in maximum heart rate achieved chances of having heart disease decreases by 0.68% for females. Unit increment in ST depression induced by exercise lead to 68.79% Increase in chances of having heart disease in females. There is somewhat high impact of number of major vessels in which blockage has been observed during fluoroscopy on chances of having heart disease. For every single unit of increment in it, it increase the chances of having heart disease by 131.92% for females. The odds of female with Normal Thalassemia is 100% lower than the female with Fixed defect Thalassemia. Similarly, the odds of a female with Reversible defect Thalassemia are 100% lower than the female with Fixed defect Thalassemia.

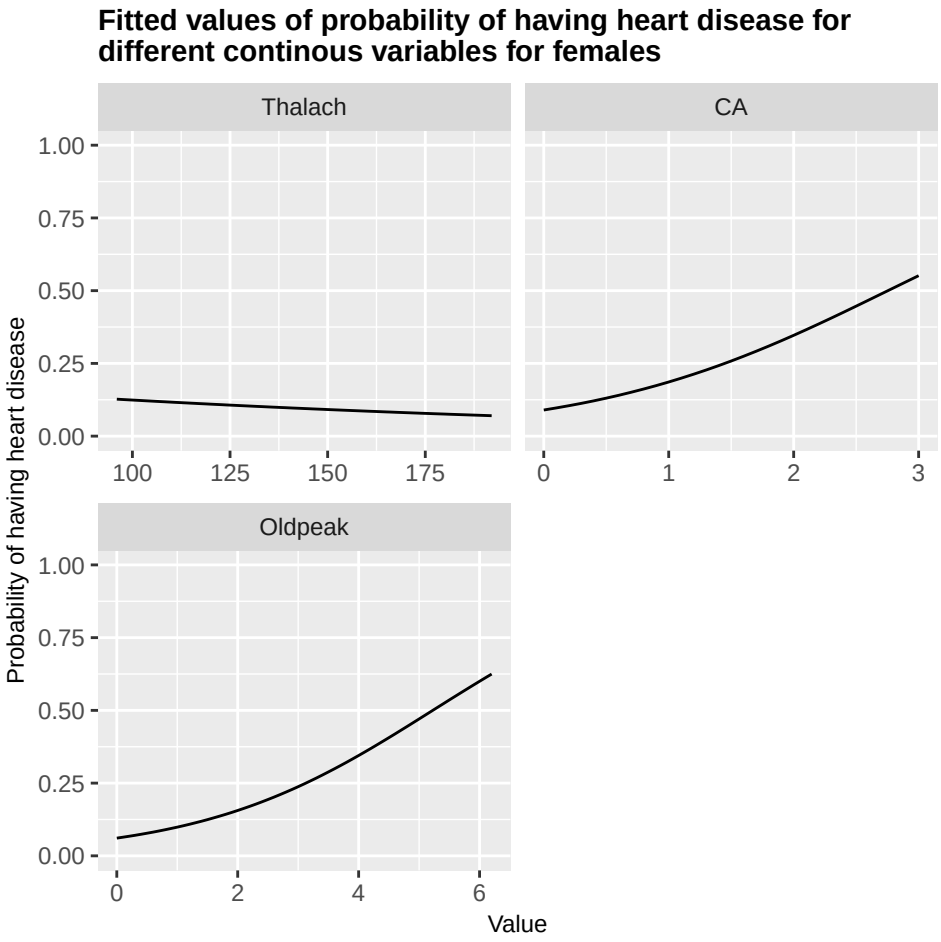
Table 3: Odds Ratio for Males

	OR
Thalach	-0.68
Oldpeak	68.79
CA	131.92
ThalNormal	-100.00
ThalReversible Defect	-100.00

Now let's try to visualize the fit of the model for every variable in case of both males and females. Here Thal is categorical variable has only 3 different possible values to it is technically not feasible to plot fitted graph for it. So, we tried to find a separate box plot for the Thal variable showing the range of predicted values for Goal variables with different values of Thal (See appendix).

For other three variables we have fitted model on new grid data frame and predicted the chances of having heart disease for them. For each variable a grid has been implemented and the other independent variables are set to their median value for the sake of visualization





From the graphs above, we can clearly see that impact of Maximum heart rate achieved (Thalach) is very different on likelihood of males and females having a heart disease. Similarly, there is some difference in impact of No. of major vessels blocked (CA) on likelihood of males and females having a heart disease. While the impact of ST depression induced by exercise (Old peak) on likelihood of males and females having a heart disease is almost same.

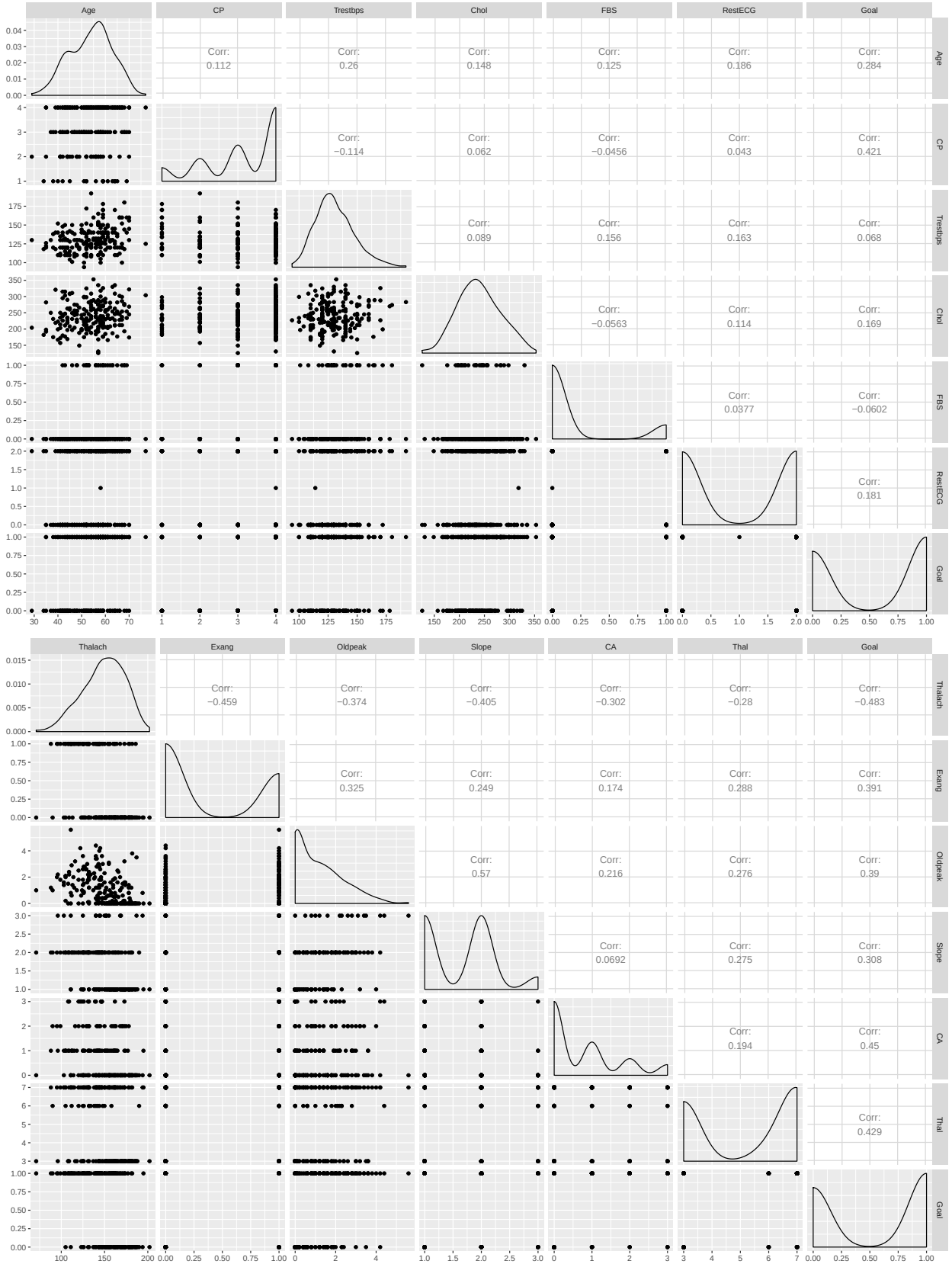
Limitations

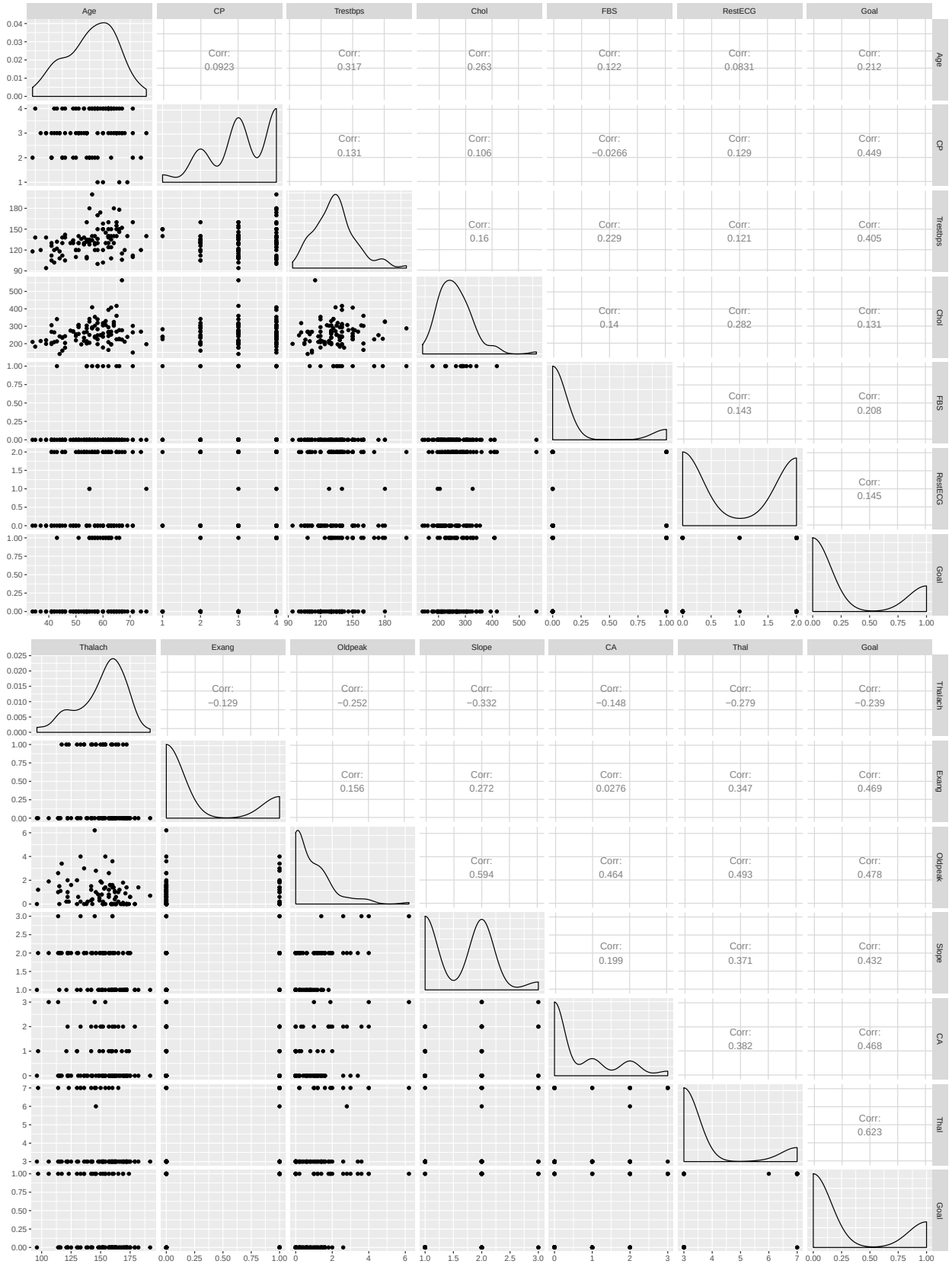
One of the biggest limitations of the dataset available was the less number of females available in the dataset. This prevented us from using more variables in the model and identifying more interactions as doing so would lead to over fitting for females. To understand the effect better, we need to do a study for equal number of males and females and analyse the difference in impact of different variables (possibly more variables then the ones used in this project) on the likelihood of males and females having a heart disease.

Conclusion

From the results above, we can see that the impact of variables on males and females is drastically different except for the Old peak variable. The Odds ratio have helped us quantify the effect of different variables on the odds of a male or females having a heart disease. Thus, we can say that the variables do affect the males and females differently based on the limited data we have. The results may be slightly biases for females because of the less data available.

Appendix


Age
CP
Trestbps
Chol
FBS
RestECG
Goal
Thalach
Exang
Oldpeak
Slope
CA
Thal
Corr: 0.112
Corr: 0.26
Corr: 0.148
Corr: 0.125
Corr: 0.186
Corr: 0.284
Corr: −0.114
Corr: 0.062
Corr: −0.0456
Corr: 0.043
Corr: 0.421
Corr: 0.089
Corr: 0.156
Corr: 0.163
Corr: 0.068
Corr: −0.0563
Corr: 0.114
Corr: 0.169
Corr: 0.0377
Corr: −0.0602
Corr: 0.181
Corr: −0.459
Corr: −0.374
Corr: −0.405
Corr: −0.302
Corr: −0.28
Corr: −0.483
Corr: 0.325
Corr: 0.249
Corr: 0.174
Corr: 0.288
Corr: 0.391
Corr: 0.57
Corr: 0.216
Corr: 0.276
Corr: 0.39
Corr: 0.0692
Corr: 0.275
Corr: 0.308
Corr: 0.194
Corr: 0.45
Corr: 0.429



Explanation of oldpeak variable:

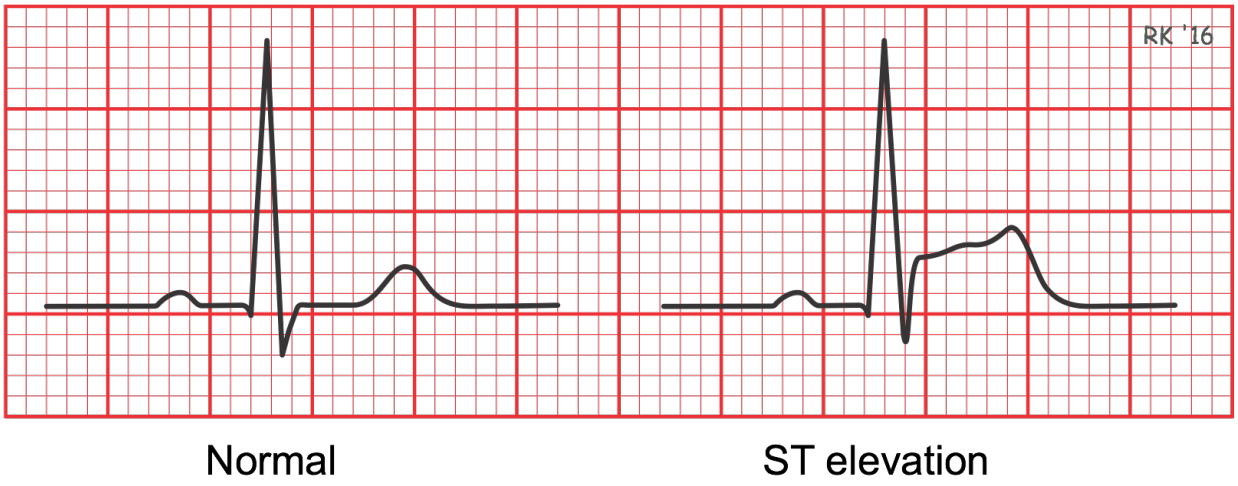
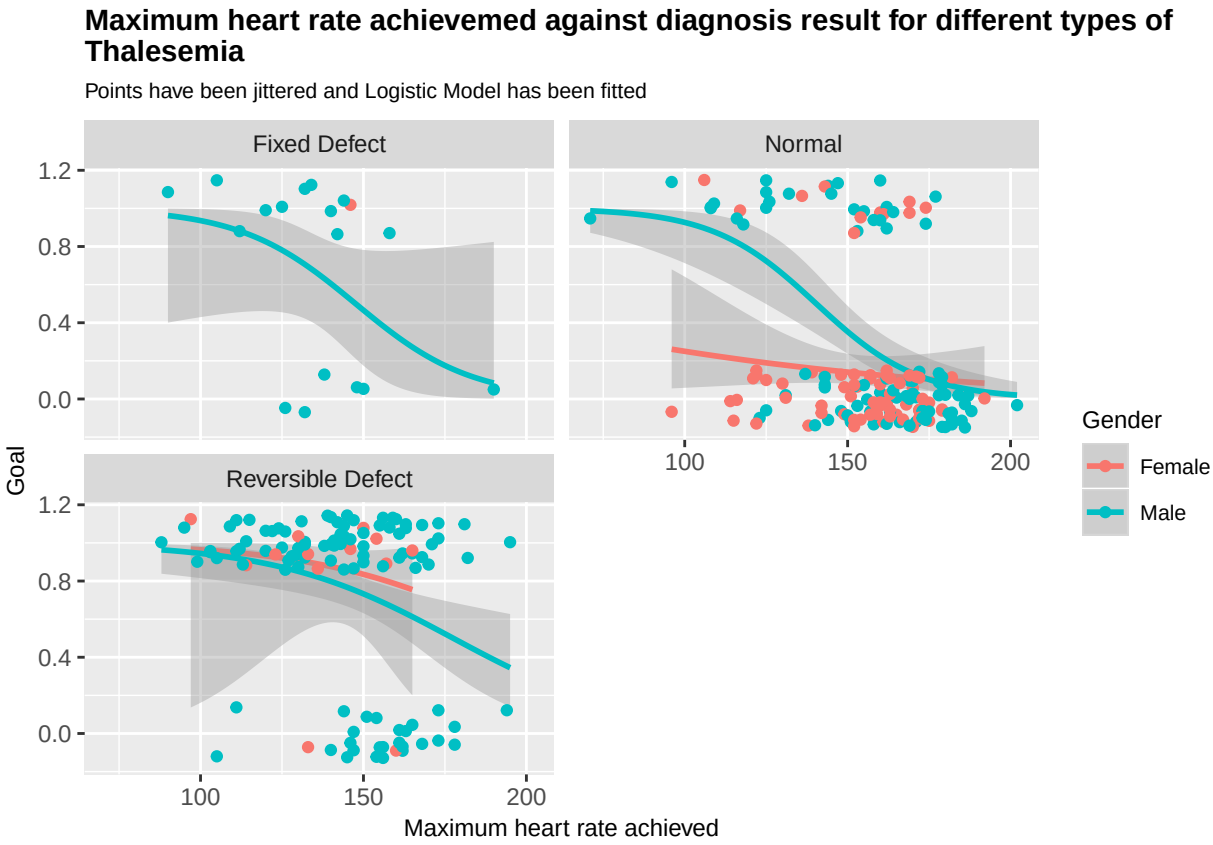


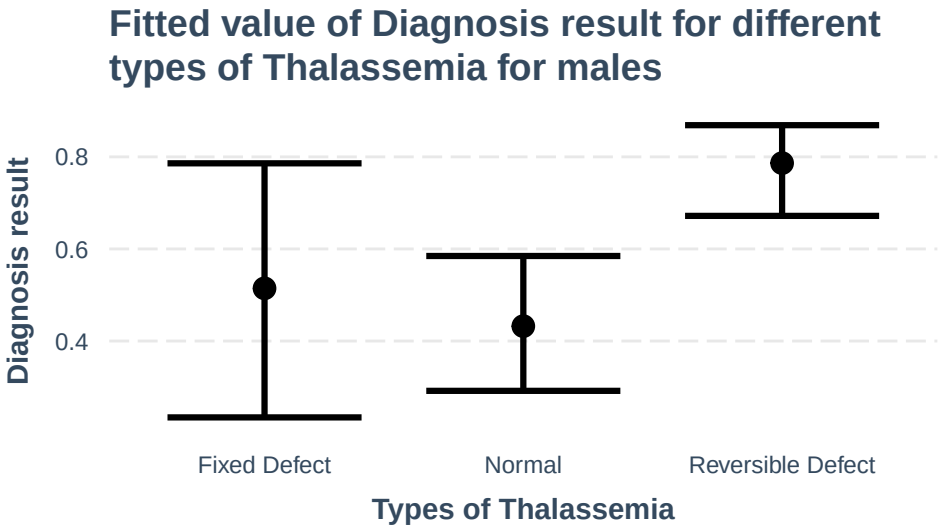
Figure 1: ST segment of ECG

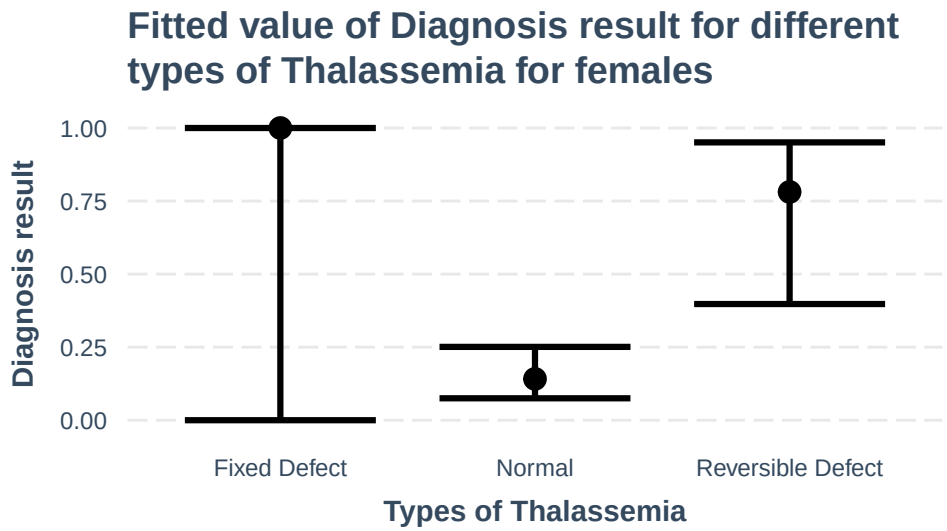
"An electrocardiogram (ECG) measures the heart's electrical activity. The waves that appear on it are labeled P, QRS, and T. Each corresponds to a different part of the heartbeat. The ST segment represents the heart's electrical activity immediately after the right and left ventricles have contracted, pumping blood to the lungs and the rest of the body. Following this big effort, ventricular muscle cells relax and get ready for the next contraction. During this period, little or no electricity is flowing, so the ST segment is even with the baseline or sometimes slightly above it. The faster the heart is beating during an ECG, the shorter all of the waves become. The shape and direction of the ST segment are far more important than its length. Upward or downward shifts can represent decreased blood flow to the heart from a variety of causes, including heart attack, spasms in one or more coronary arteries (Prinzmetal's angina), infection of the lining of the heart (pericarditis) or the heart muscle itself (myocarditis), an excess of potassium in the bloodstream, a heart rhythm problem, or a blood clot in the lungs (pulmonary embolism)."

Interaction between Thalach and Thal



For Males, the slope of Thalach for Males doesn't change a lot with the type of Thalassemia detected. For Females, we don't have enough data to know if the slope changes or not. Thus we cannot say that there is an interaction between Thalach and Thal.





From the graphs above, we can clearly see that effect of Normal Thalassemia is very different for males and females. The range predicted values of goal is more for males as compared to females. Similarly there is a difference in the predicted value of goal for males and females even for Fixed Defect Thalassemia and Reversible Defect Thalassemia.