Christine Zhou

Professor: Timothy Keaton

STAT 355 Project

Due Date: 04/12/2020

**Heart Disease Analysis Project**

**Introduction:**

For this project, I choose the dataset: Heart Disease UCI. I will first explore the data and come up with some interesting hypothesis from the correlations then test them using different methods. The goal for this project is to identify some common traits on the population with a higher risk of having heart disease.

Data from:

<https://www.kaggle.com/ronitf/heart-disease-uci/tasks?taskId=84>

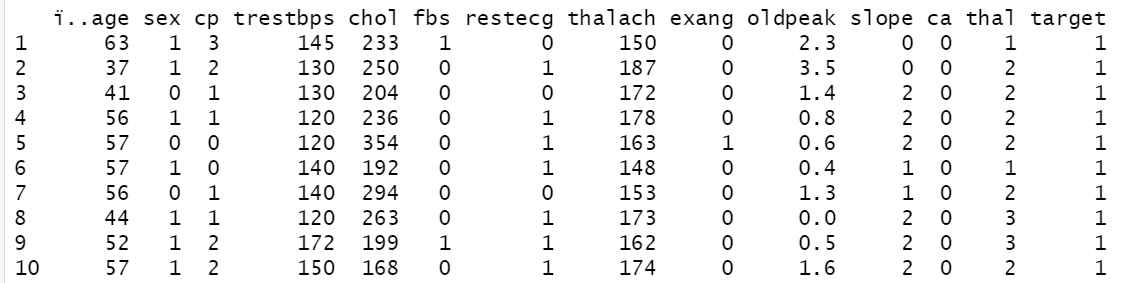
**Part 1: Data Exploration**

**Dataset Overview:**

This data set contains 303 patients (rows) each with 14 attributes (columns) such as their age and sex. Below is a more detailed description of the attributes:

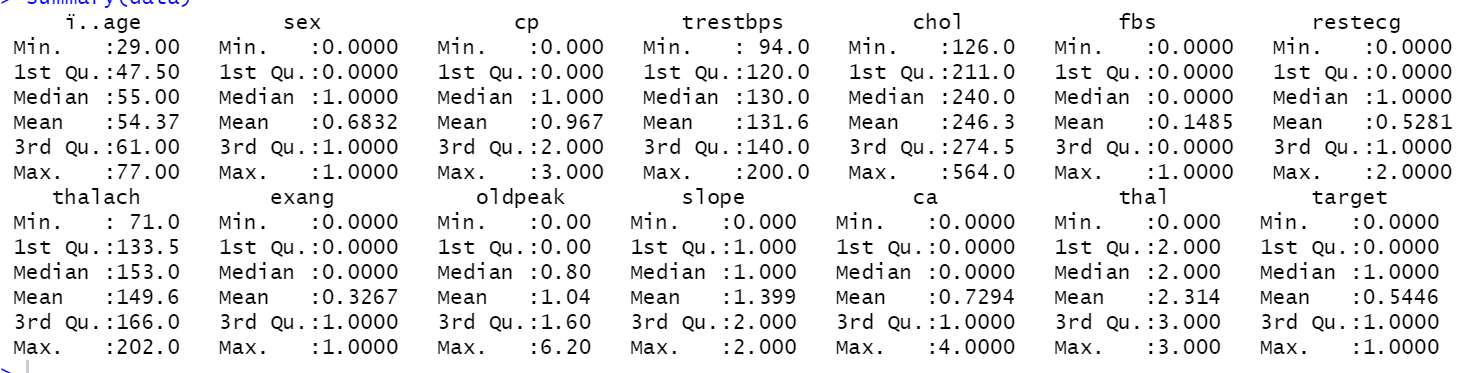
1. age
2. sex (1 = male; 0 = female)
3. chest pain type (values 0,1,2,3)
4. resting blood pressure (in mm Hg on admission to the hospital)
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina (1 = yes; 0 = no)
10. oldpeak = 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 flourosopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
14. target: 0 = heart disease, 1 = no heart disease

Here is an example of what the data looks like after importing into R:

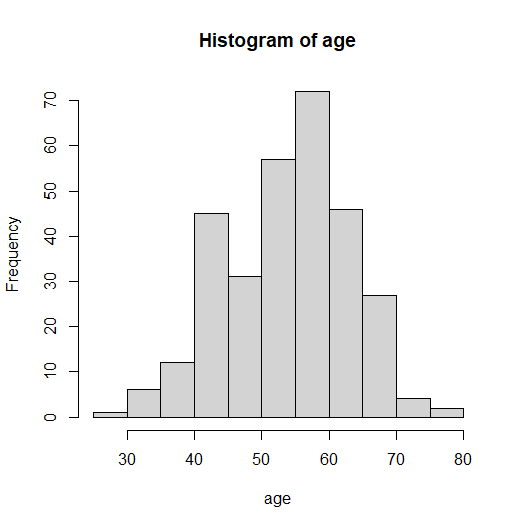
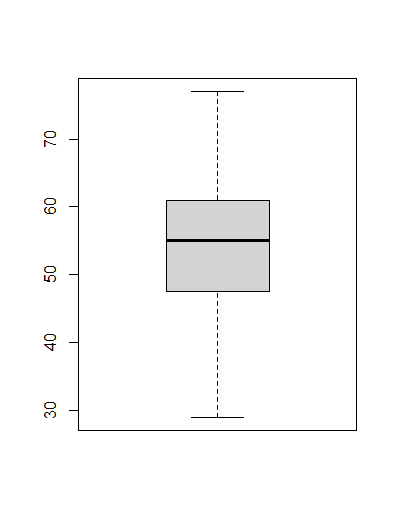


**Singe Variable Overview:**

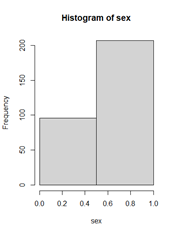
It’s time to look more thoroughly into each variable, here is an overview of the distribution of each variables in this dataset:



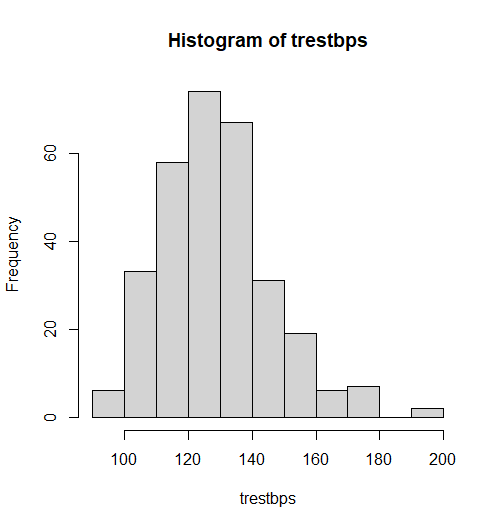
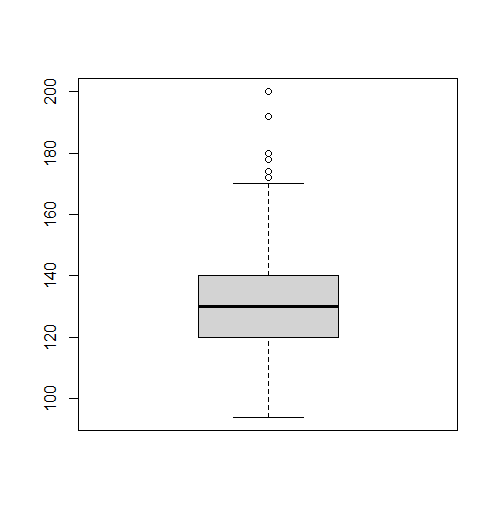
The best way to see the distribution of a single type of variable is to plot them on histogram or box plot:

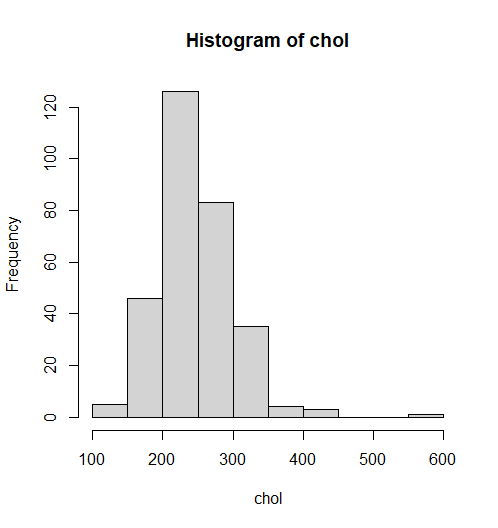
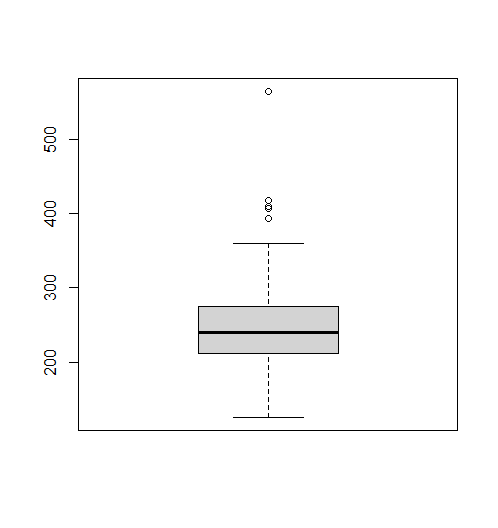
 

We can see that the age of the patients averages around 55 years old with a normal distribution. There are no outliers in age. The sample in this dataset is aimed towards older populations since I assume that older population tends to have more health problems.

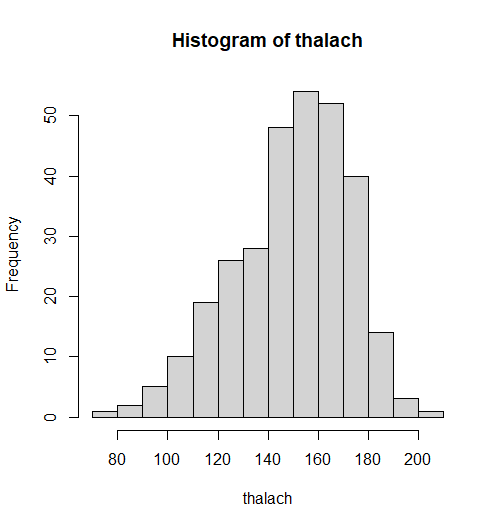
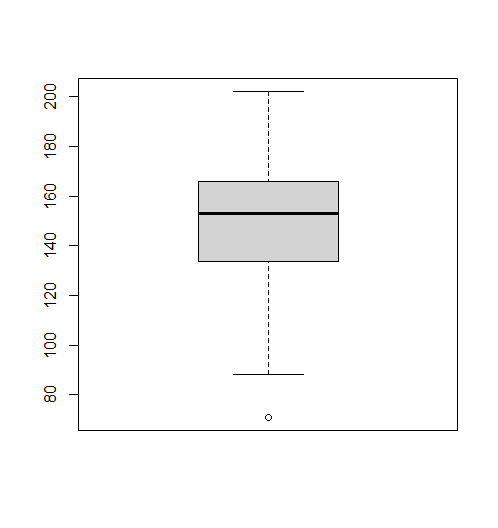


There are much more male patients than female patients. That means that this data could be biased towards male population.

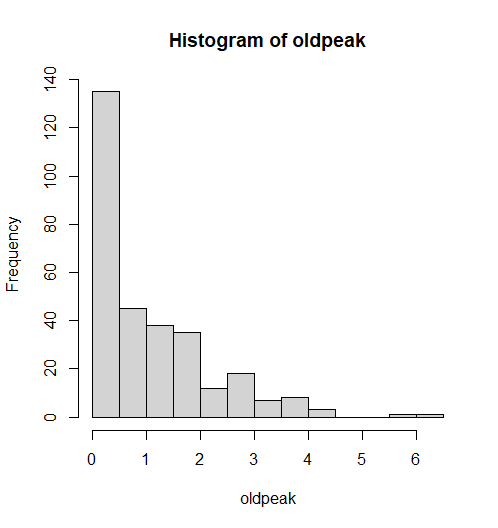
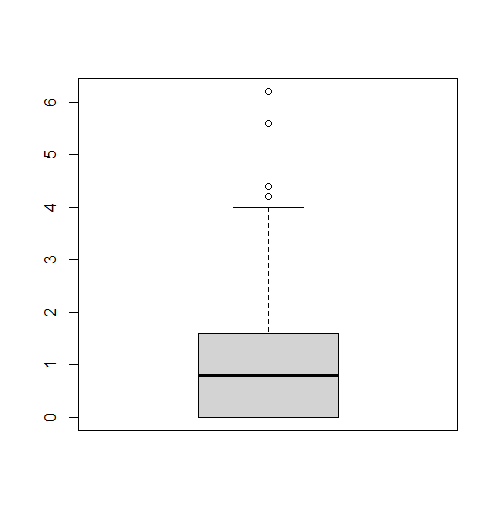
 

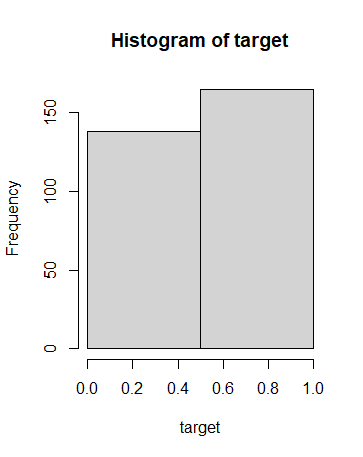
The resting blood pressure and the cholesterol level of the patients are both right skewed with few outliers in the upper distribution.

The maximum heart rate achieved is left skewed with 1 outlier on the lower range. The mean is aroun 150 beats per min.

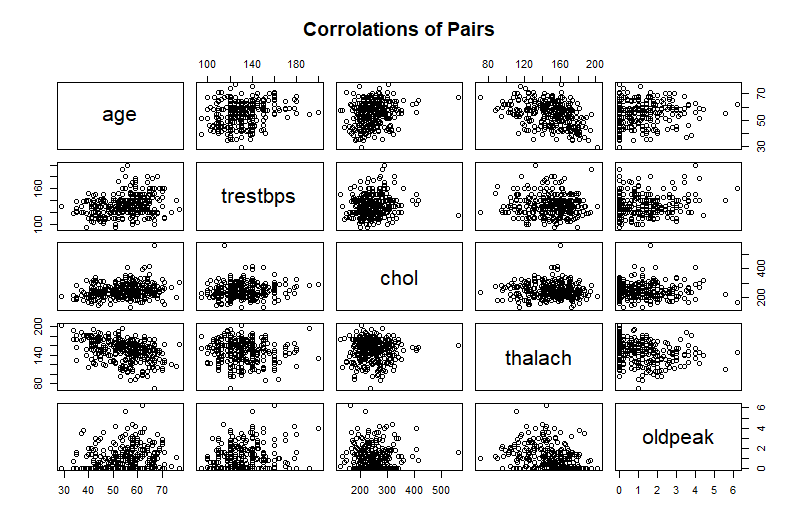
The ST depression level is very right skewed with few upper outliers. Mostly because it has a lower limit of 0. It is a test that indicates the probability of coronary artery disease. Deeper and more widespread (higher number) ST depression generally indicates more severe or extensive disease.



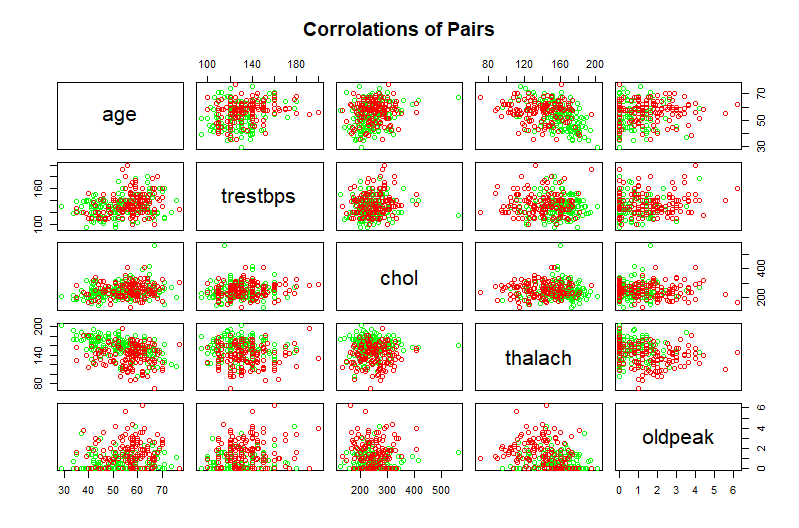
There are more patients without heart disease (target = 1) than with heart disease (target = 0).

**Correlations:**

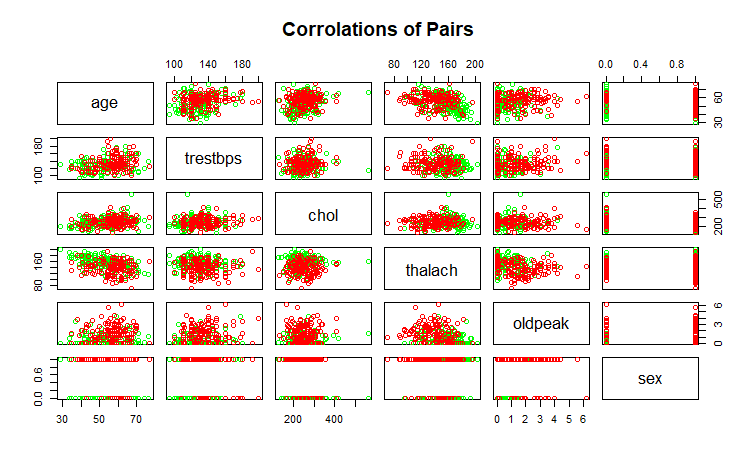
Now let’s look at the correlation between different variables:



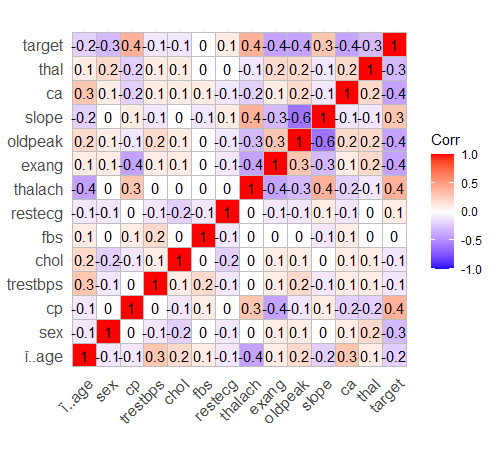
There is not much correlations between each pairs of variables other than a downward trend between age and maximum heart rate. This is a good thing because that way each variables can contribute to the prediction of the population with a higher chance of having heart disease.



This time I colored the patients diagnosed with red representing population diagnosed heart disease and green without. The difference between some are big enough to be noticed by eyes. For example, the pair with cholesterol level (chol) and max heart rate (thalach) shows that the patients with lower maximum heart rate has a higher chance of having heart disease and that cholesterol level has almost no impact on the chance of getting heart disease.



For this plot I added the variable sex and it seems that older male patients have higher change of having heart disease than female patients. The data could be biased because there are more male patients in this dataset.



I also made a correlation matrix using ggplot library. The variables that have the most correlation with the target are cp, thalach, exang, oldpeak, and ca. It also looks like there is a relatively strong correlation between the variables thalach, exang, oldpeak, and slope.

**Part 2 Hypotheses Testing**

In the part 2 of the project I will be using different tests to test my hypothesis of different variables and their association with heart disease and with each other.

**Question 1: Max Heart Rate vs. Target**

In the correlation graph above it seemed that the maximum heart rate has a relatively strong correlation with heart disease. In this question I will be using 2-sided t-test and Wilcoxon test to see if there is actually a difference between the average max heart rate (thalach) of patients with heart disease and without heart disease (target). I will be using independent t-test since the variables are from different patients.

H0: There is no difference between the average max heart rate of patients with heart disease and without heart disease.

H1: There is a difference between the average max heart rate of patients with heart disease and without heart disease.

𝛼 = 0.05

T-test summary:

data: hd and nhd

t = -7.953, df = 269.9, p-value = 5.019e-14

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-24.15912 -14.57132

sample estimates:

mean of x mean of y

139.1014 158.4667

Since p-value < 5.019e-14 < 𝛼 = .05, reject H­­0. There is sufficient evidence that the average max heart rate of patients with heart disease is different from patients without heart disease.

It makes sense to assume that the heart of a heart disease patient would be less active than a heart without disease. So, I will now do a 1-sided t-test predicting that the average max heart rate of heart disease patient would be lower than the average from heart disease free patient.

H0: The average max heart rate of patients with heart disease is not lower than the average for patients without heart disease.

H1: The average max heart rate of patients with heart disease is lower than the average for patients without heart disease.

𝛼 = 0.05

T-test summary:

data: hd and nhd

t = -7.953, df = 269.9, p-value = 2.509e-14

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:

-Inf -15.34629

sample estimates:

mean of x mean of y

139.1014 158.4667

Since p-value < 5.019e-14 < 𝛼 = .05, reject H­­0. There is sufficient evidence that the average max heart rate of patients with heart disease is lower than the average for patients without heart disease.

Wilcoxon test summary:

data: hd and nhd

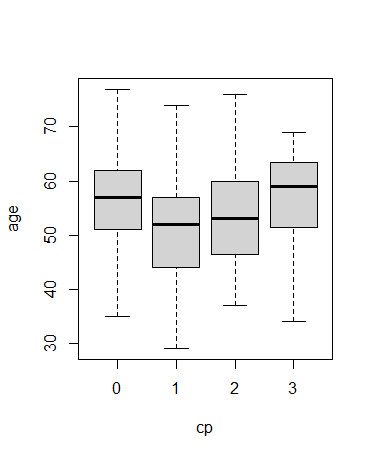
W = 5732, p-value = 4.898e-14

alternative hypothesis: true location shift is less than 0

Since p-value < 4.898e-14 < 𝛼 = .05, reject H­­0. There is sufficient evidence that the average max heart rate of patients with heart disease is lower than the average for patients without heart disease even without needing to assume that the CLT has kicked in and that the sample mean comes from a Normal distribution.

**Question 2: How does different types of chest pain affect different age group**

It seemed that there is a link between patients’ age and the types of chest pain they experience. So, I decided to compare the mean of patients’ age across the different types of chest pain.



Here is a boxplot for the age distribution with different types of chest pain. It seems like the patients who experience type 1 chest pain are on average younger than the patients who are experiencing type 0 chest pain.

Let’s validate the observation by doing an ANOVA test to compare patients’ age across the different types of chest pain they experience.

H0: There is no difference between the patients’ age across the different types of chest pain they experience.

H1: There is at least one chest pain that has a different age group from the others

𝜶 = .05

Summary of ANOVA test:

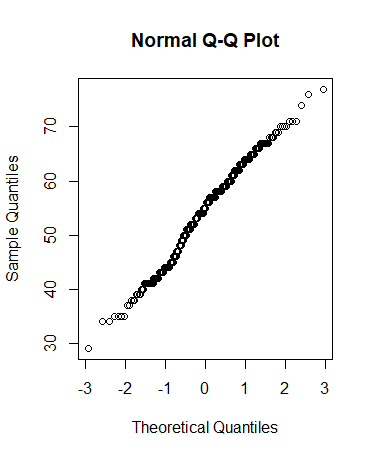
Df Sum Sq Mean Sq F value Pr(>F)

cp 1 117 117.41 1.425 0.233

Residuals 301 24793 82.37

Since 𝒑-value = 0.233 > 𝜶 = .05, fail to reject 𝑯𝟎. There is no sufficient evidence that at least one of the chest pains has different age group from the others. The difference between the ages for the types of chest pains are not big enough for me to conclude that there is a variance. This could be caused by too little data for one or more types of chest pain.

Checking assumptions for ANOVA:



The variances do look similar across groups. Those points seem to be generally following a straight line. So, the Normal assumption is probably valid

**Question 3: Does interaction between variables affect their correlation with the target**

In this question I will create a multiple regression model with cp, thalach, exang, oldpeak, and ca to predict heart disease (target), both with and without interactions.

Firstly, I want to see if there is a difference between with 2-way interaction and without interactions.

𝛼 = .05

H0 = There is no difference with and without interactions for predicting heart disease.

H1 = At least one interaction is useful for predicting the heart disease.

Analysis of Variance Table:

Model 1: target ~ cp + thalach + exang + oldpeak + ca

Model 2: target ~ cp \* thalach \* exang \* oldpeak \* ca

Res.Df RSS Df Sum of Sq F Pr(>F)

1 297 41.736

2 271 35.590 26 6.1458 1.7999 0.01166 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Since 𝒑 = 0.01166 < 0.05, reject 𝑯𝟎. This means that we do have sufficient evidence to say that at least one interaction is useful for predicting heart disease.

Now it’s time to see which interactions has the most correlation:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.5610721 0.6052154 2.579 0.01043 \*

cp -0.6131047 0.3535273 -1.734 0.08401 .

thalach -0.0045445 0.0037441 -1.214 0.22589

exang -2.1988298 0.8255786 -2.663 0.00820 \*\*

oldpeak -1.5682160 0.5360062 -2.926 0.00373 \*\*

ca 0.0140594 0.4863251 0.029 0.97696

cp:thalach 0.0040163 0.0022039 1.822 0.06951 .

cp:exang 3.2854230 1.2458288 2.637 0.00884 \*\*

thalach:exang 0.0119087 0.0054088 2.202 0.02853 \*

cp:oldpeak 0.6722232 0.2811906 2.391 0.01750 \*

thalach:oldpeak 0.0095415 0.0035371 2.698 0.00742 \*\*

exang:oldpeak 1.7637473 0.6178794 2.855 0.00464 \*\*

cp:ca -0.0405189 0.3705018 -0.109 0.91300

thalach:ca -0.0020168 0.0032131 -0.628 0.53073

exang:ca 0.4564966 0.7391852 0.618 0.53738

oldpeak:ca 0.3064885 0.3333721 0.919 0.35873

cp:thalach:exang -0.0177868 0.0074143 -2.399 0.01712 \*

cp:thalach:oldpeak -0.0042347 0.0018731 -2.261 0.02457 \*

cp:exang:oldpeak -2.1328413 0.7563091 -2.820 0.00516 \*\*

thalach:exang:oldpeak -0.0117828 0.0042253 -2.789 0.00567 \*\*

cp:thalach:ca 0.0012100 0.0023122 0.523 0.60118

cp:exang:ca -0.7233849 0.9848484 -0.735 0.46327

thalach:exang:ca -0.0023976 0.0050074 -0.479 0.63245

cp:oldpeak:ca -0.1165509 0.2717142 -0.429 0.66830

thalach:oldpeak:ca -0.0017541 0.0022506 -0.779 0.43644

exang:oldpeak:ca -0.4669013 0.4460496 -1.047 0.29615

cp:thalach:exang:oldpeak 0.0121201 0.0044312 2.735 0.00665 \*\*

cp:thalach:exang:ca 0.0024096 0.0069308 0.348 0.72837

cp:thalach:oldpeak:ca 0.0002865 0.0017605 0.163 0.87084

cp:exang:oldpeak:ca 2.1669578 1.3025725 1.664 0.09735 .

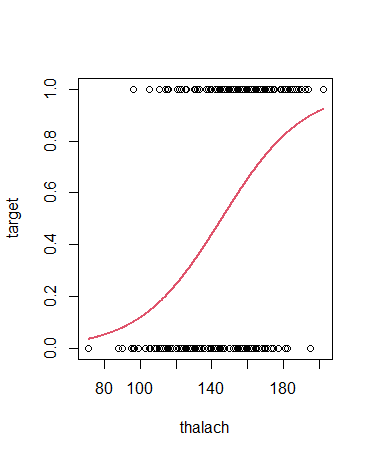
thalach:exang:oldpeak:ca 0.0031861 0.0031323 1.017 0.30997

cp:thalach:exang:oldpeak:ca -0.0155697 0.0103240 -1.508 0.13269

All the highlighted interactions have a p-value less than 0.01, which means that they are the most useful in predicting heart disease.

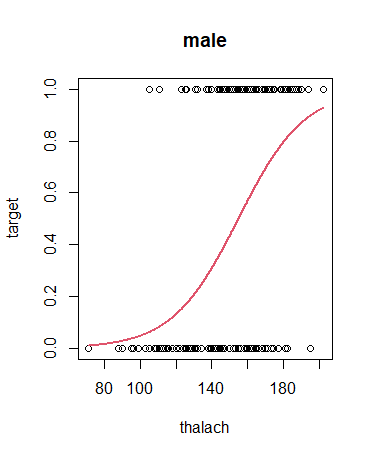
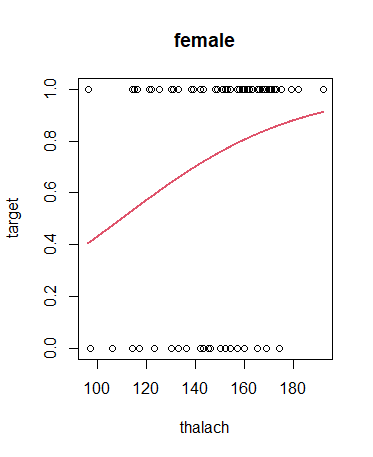
**Question 4: Estimating probability of heart disease based on variables**

For question 4 , I will build a logistic regression model to estimate the probability the patient has heart disease based on maximum heart rate and then separate the data by sex.



The slope of the logistic regression = 1.044931 , that means that as maximum heart rate increases by 1 bpm, the odd of not having heart disease are predicted to be 1.044931 times higher, which is not a lot.

Now I will build a logistic regression model with separated sex:

It seems that from the plot, the logistic regression is not a good fit for predicting the probability of female patient having heart disease based on maximum heart rate.