|  |
| --- |
|  |

|  |
| --- |
| **Northeastern University**  Toronto, Canada |

**Final Project Report**

Intermediate Analysis

(ALY 6015)

**Submitted by**

Dhruvang Patel (NUID: 002195090)

Rushvil Patel (NUID:002964342)

Parva Patel(NUID:002195186)

Date: 20th May, 2022

**Guided By:**

Dr. Yvonne Leung

**INTRODUCTION:**

* In this study, a first analysis is done on the "heart" dataset, which comprises information about 300 patients of various ages and genders (UCI Machine Learning Repository: Heart Disease Data Set, n.d.) to understand the presence of heart disease in a patient. The dataset includes data on chest pain, maximum heart rate, age, gender, sugar level and cholesterol in order to determine whether or not a patient has heart disease.
* This report will solve the questions which will help in determining the heart disease and how they can be tackled.

**METHODS:**

1. **Hypothesis Testing**: This helps to study data for heart disease as well as determine the minimum criteria required.
2. **Linear and Logistic Regression**: To predict the target variable for the heart disease
3. **Generalized Linear Model:** Model the relationship between heart disease and other categorical variable.

**ANALYSIS:**

* **Description of data**

The dataset contains over 1000 observations of random patients with 13 attributes of information of patients which contains:

1. Age
2. Sex (1: Male, 0: Female)
3. chest pain (0: asymptomatic, 1 : atypical angina, 2 : no angina, 3: typical angina)
4. resting blood pressure
5. Cholesterol
6. Blood sugar (1: >120mg/dl, 0: <120mg/dl)
7. Resting ECG (0: normal, 1: having ST-T wave abnormality,

2: showing probable or definite left ventricular hypertrophy)

1. Maximum Heart Rate
2. Exercise induced angina (1: yes, 0:no)
3. Old peak = ST depression induced by exercise relative to rest
4. slope: the slope of the peak exercise ST segment (1: upsloping ,

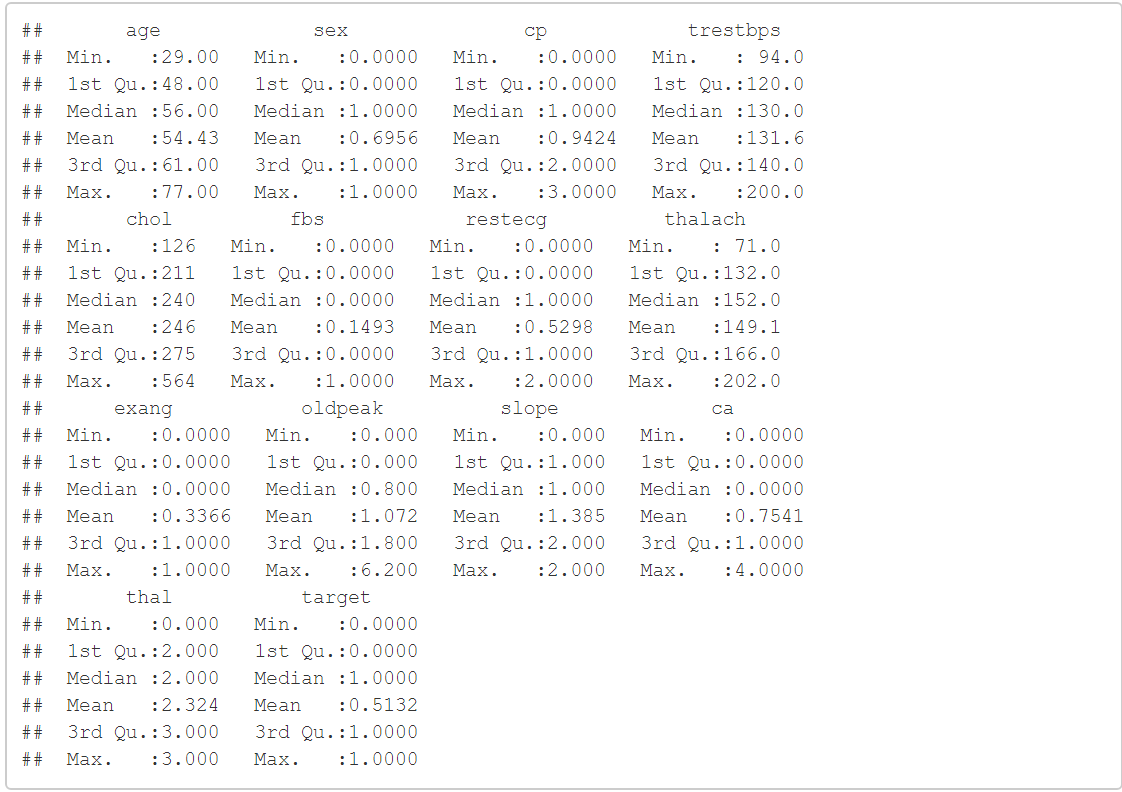
2: flat, 3: down sloping)

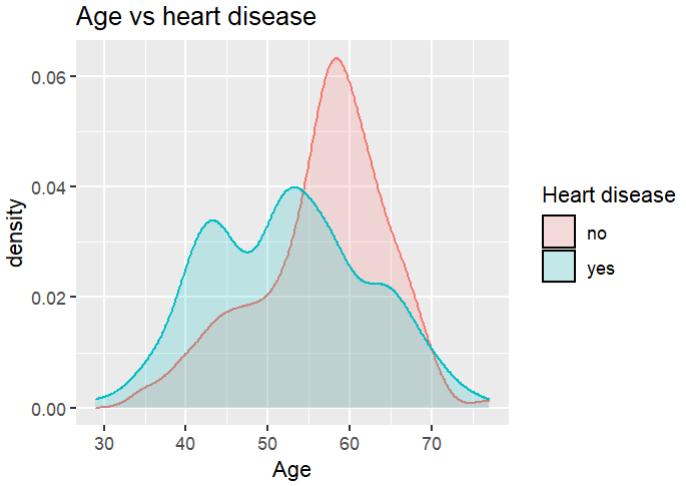
1. ca: number of major vessels (0-3) coloured by fluoroscopy
2. thal ( 1 = normal; 2 = fixed defect; 3 = reversable defect)
3. Heart disease (0: No, 1: yes)

Here heart disease is a predictor variable and all other attributes are categorical variable which can be used to predict the heart disease.

* **Exploratory Data Analysis:**

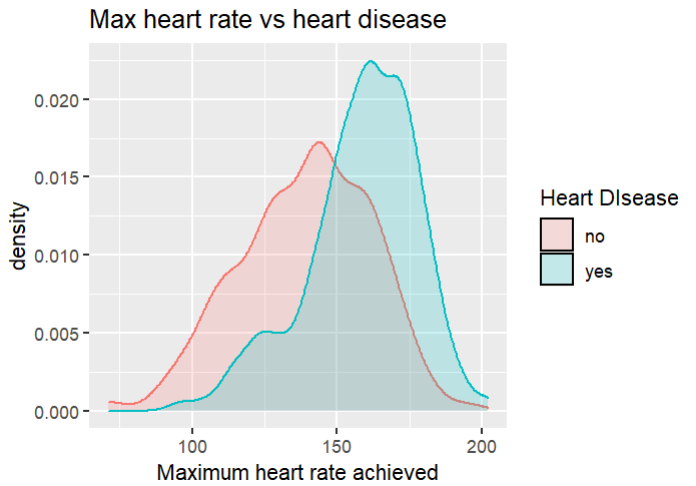
Summary of dataset is figured by summary() to know the dataset.

  
This will give us information about the statistical information such as range, mean, median of the data.

****

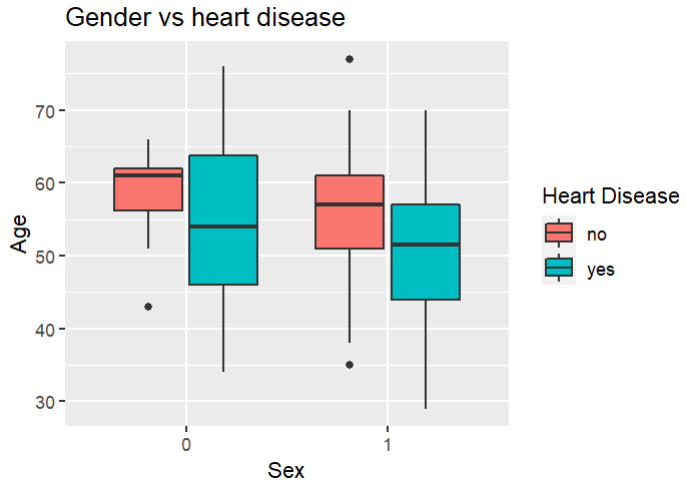
*Figure 1: Age vs heart disease*

Figure 1 shows some intriguing findings, since the dataset shows that middle aged individuals have more heart disease and older persons have less heart disease.

****

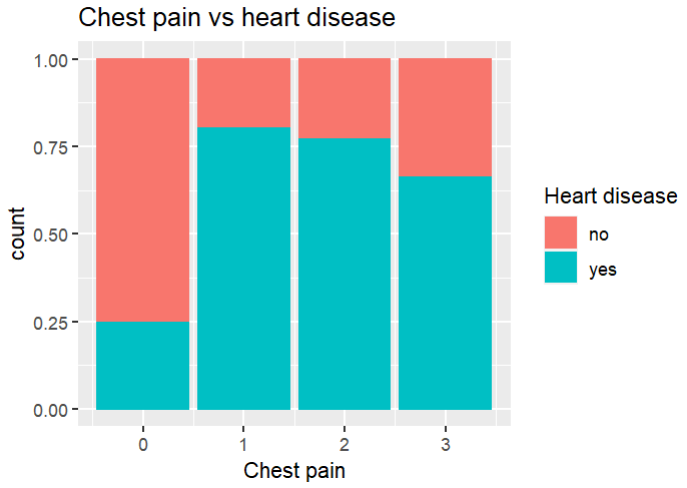
*Figure 3: Density plot of max heart rate for heart disease*

Figure 3 depicts people's highest heart rates and their impact on the heart. It is clear that if the heart rate exceeds 150, the likelihood of developing heart disease increases.

****

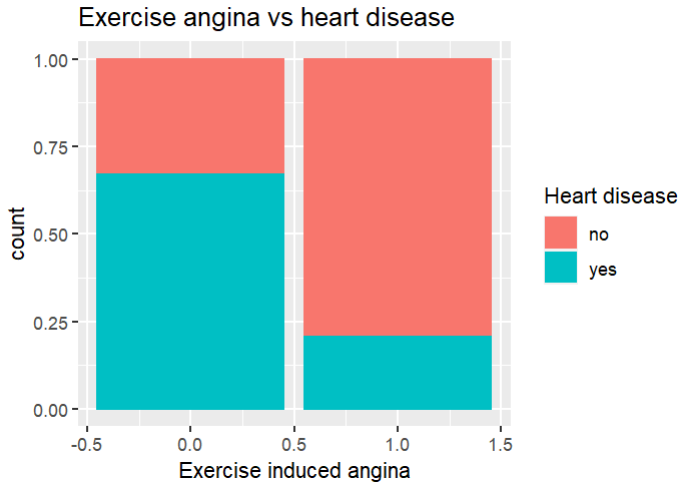
*Figure 4: Barplot of gender and heart disease*

Figure 4 shows the difference between the male and female having heart disease or not where 0 is female and 1 is male.



*Figure 2: Barplot of types of chest pain*

Figure 2 shows that atypical angina has the highest risk of heart disease which is 1, followed by no angina(2), normal angina(3), and asymptomatic angina(0), which have the lowest risk of heart disease. Asymptomatic angina means absent of chest pain and not showing any symptoms.



*Figure 5: bar plot of exercise induced angina*

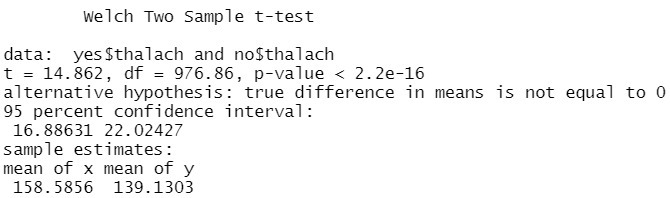
Angina is a symptom, not a medical condition. This sign means you're more likely to develop heart disease, which can be deadly. As indicated in the graph above, persons with exercise-induced angina are more likely than those who do not to acquire heart disease.

QUESTION 1: What happens if an individual reaches maximum heart rate?

t-test can be performed to find if there is any difference between the means of maximum heart rate achieved for people having heart disease or not at 95% confidence interval.

Null Hypothesis (Ho): Mean of max. heart rate is same.

Alternate Hypothesis(H1): Mean of max. heart rate is not same.



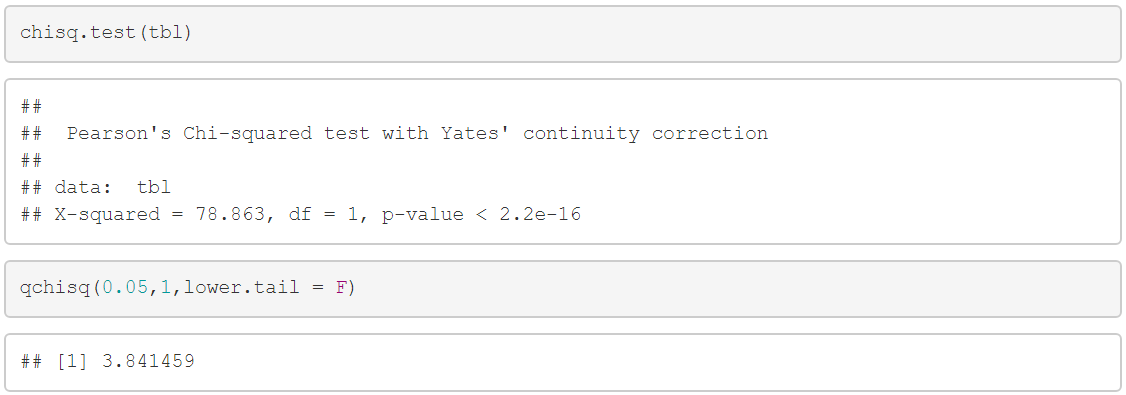
Here the p-value is less than 0.05 which means null hypothesis is rejected which says that the mean of maximum heart rate for people having heart disease and those who do not have is different.

QUESTION 2: Is there difference between gender for heart disease and is it dependent with heart disease?

A chi-squared test with a 95% confidence interval is used to determine whether gender is related to heart disease.

Null hypothesis(H0): heart disease and gender are independent.

Alternative Hypothesis(H1): heart disease and gender are dependent.



The critical value is 3.84 which is less than the x-squared value of the test which means null hypothesis is rejected. Gender and heart disease are dependent on each other.



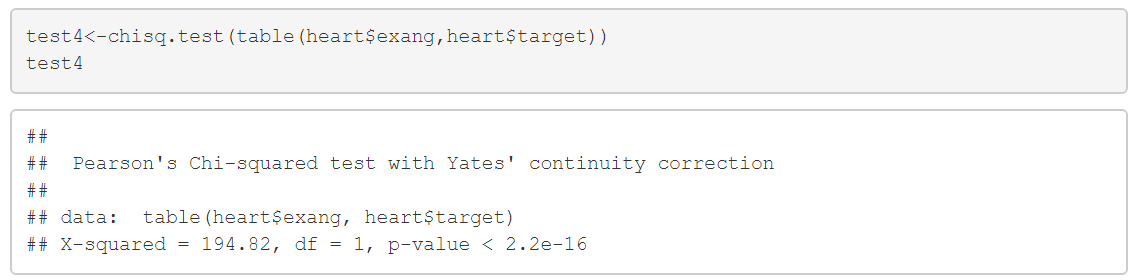
From the above table we can say that 72% females have heart disease whereas 42% male have heart disease. This means female have more heart disease than male. This analysis is a bit interesting as females have more heart disease than males.

QUESTION 3: Do exercise-induced angina symptoms suggest a higher risk of heart disease?

Here chi-squared test is used to see whether exercise- induced angina is dependent on heart disease or not.

Null hypothesis(H0): heart disease and exercise angina are independent.

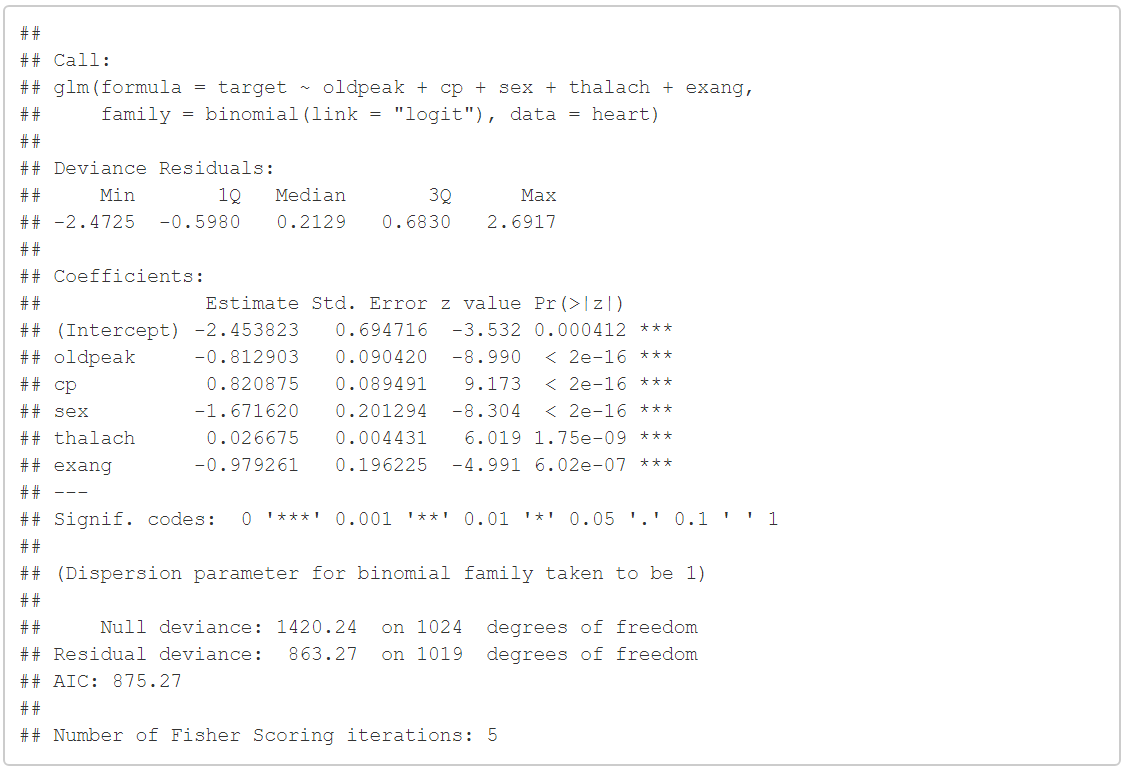
Alternative Hypothesis(H1): heart disease and exercise angina are dependent.



Here p-value is less than 0.05 which means null hypothesis is rejected and exercise induced angina is dependent on heart disease.

QUESTION 4: Can we predict heart disease using significant variables?

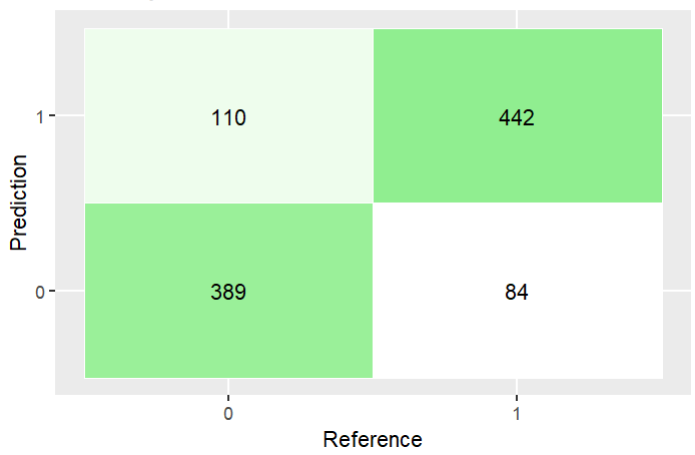
To find the response variable to generalized linear regression is used to see the significance of the response variable which were chosen i.e. old peak, chest pain, sex, maximum heart rate and exercise induced angina.



The information for the five predictor variables is shown in Figure. All the variables have p-value less than 0.05 which means all of them are statistically significant with the heart disease. An average change of 0.02 in the log probabilities of the heart disease taking on a value of 1 is connected with a one unit rise in the maximum heart rate.

Confusion matrix is created to see if we can predict that a person has heart disease or not.

For creating confusion matrix, confusionMatrix() function is used.



As seen in figure 4,

True positive: 442 patients predicted correctly by heart disease.

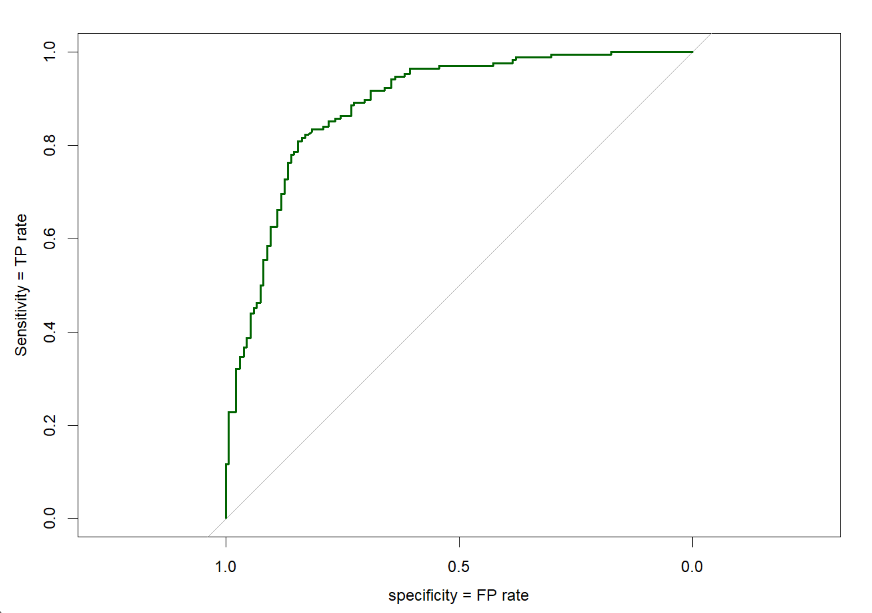
False-positive: 110 patients of not having heart disease were wrongly predicted as a having heart disease.

False-negative: 84 patients of heart disease were wrongly predicted as not having heart disease.

True Negative: 389 patients of not having heart disease were correctly predicted by the model.

Here false-negative values are more harmful as patients with heart disease were told they do not have heart disease and later on it can worsen the situation. So main concern would be reducing the false negative.

The ROC curve is a graph that shows the relationship between specificity and sensitivity. The model becomes increasingly accurate as the curve moves closer to the top-left corner. ROC curve for the decided model is:

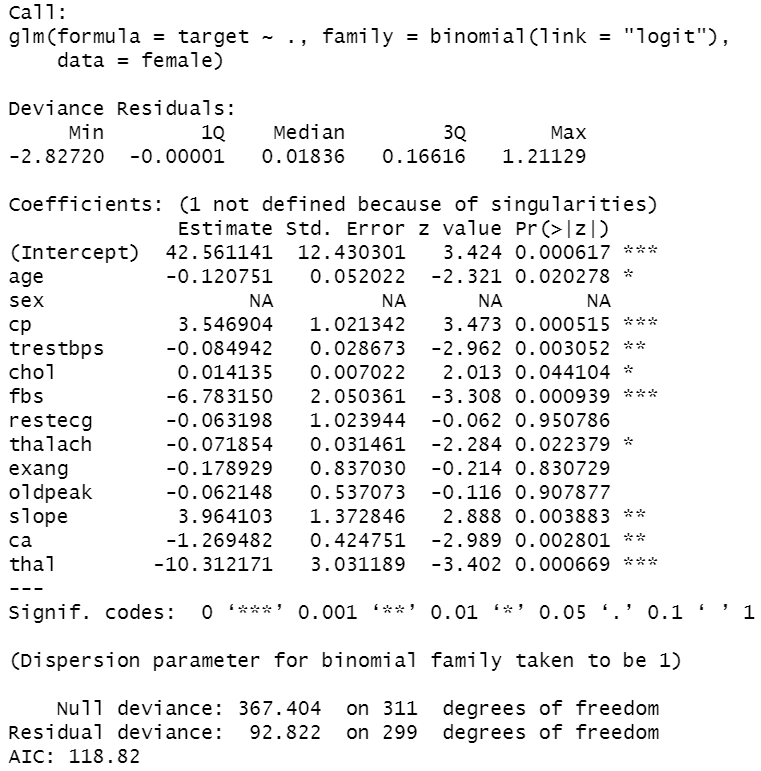




The AUC of the ROC curve is 88.6% which is pretty accurate which means 88.6% of the time we can predict the heart disease.

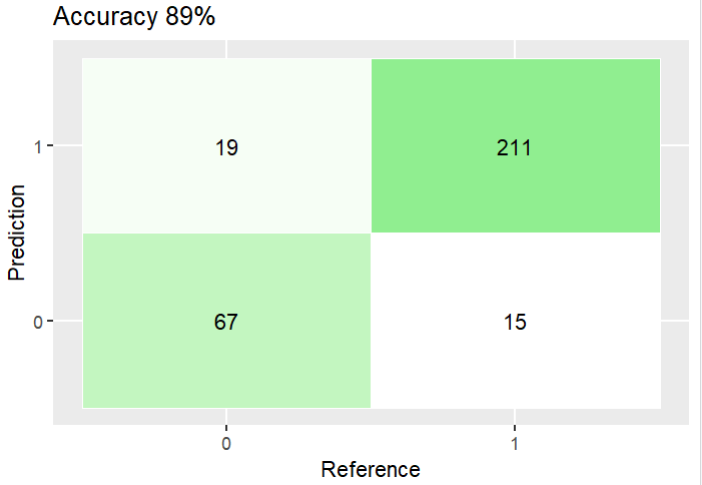
Question 5: What factors affect heart disease in female?

To determine which factors influence heart disease in women, logistic regression is used to identify the variables that influence heart disease in women.



Variables of significance may be detected after running the glm model, and four response variables were chosen: chest discomfort, exercise-induced angina, thalassemia, and fasting blood sugar.

The response variable was used to create a confusion matrix to test if these characteristics might predict heart disease in females.

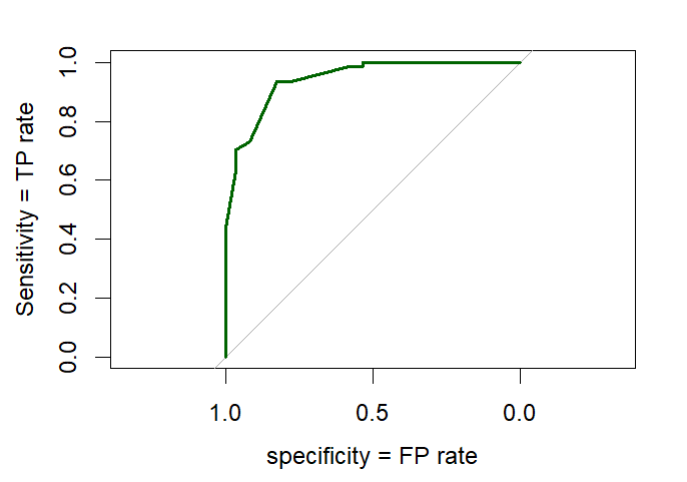


True positive: 211 female patients predicted correctly by heart disease.

False-positive: 19 female patients of not having heart disease were wrongly predicted as a having heart disease.

False-negative: 15 female patients of heart disease were wrongly predicted as not having heart disease.

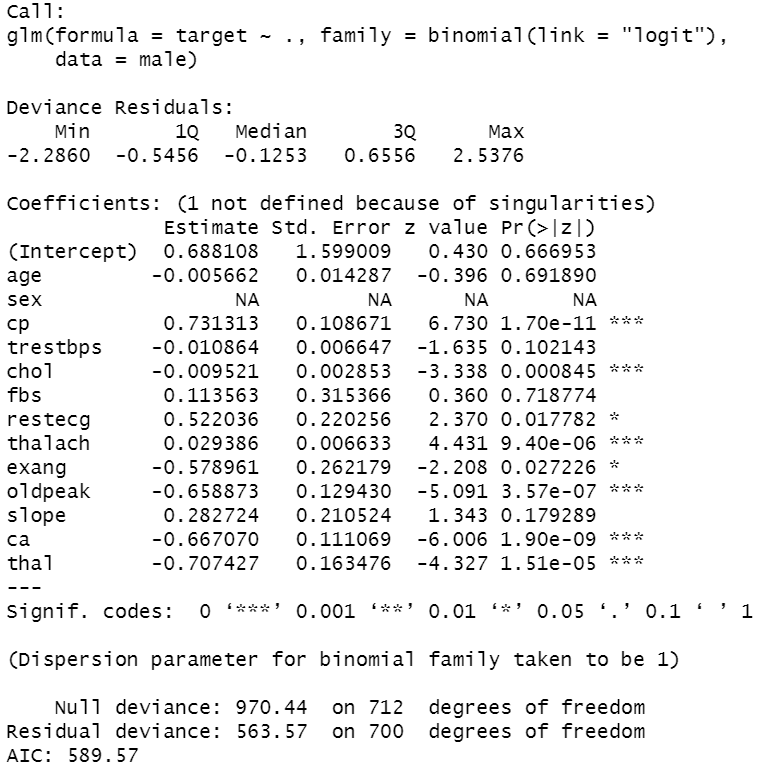
True Negative: 67 female patients of not having heart disease were correctly predicted by the model.



The AUC of the ROC curve is 0.94, indicating that this model can predict female heart disease 94% of the time.

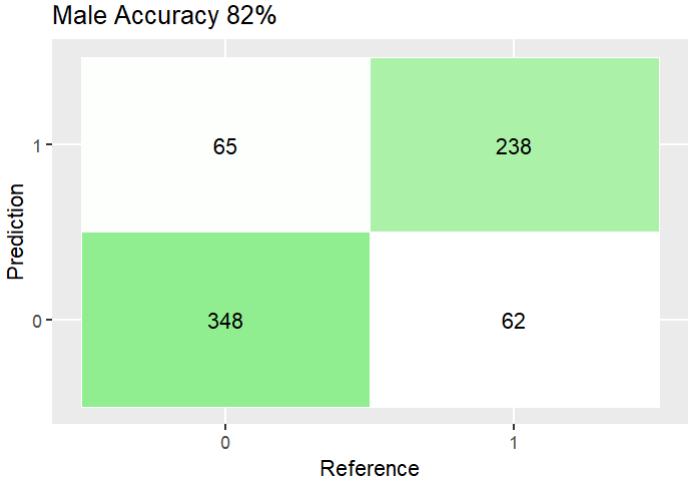
Question 6: What factors affect heart disease in male?

Logistic regression is used to find the variables that impact heart disease in men in order to establish which factors influence heart disease in men.



After running the glm model, significant factors were identified, and four response variables were selected: maximum heart rate, chest pain, old peak and thalassemia

The confusion matrix was created using the response variable to see if these factors may predict heart disease in men.

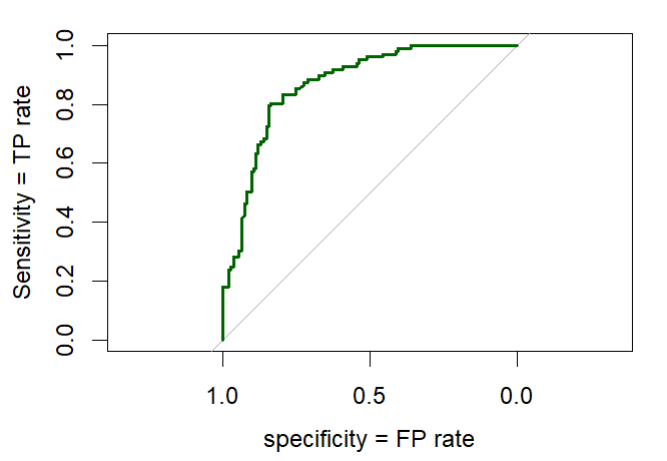


True positive: 238 male patients predicted correctly by heart disease.

False-positive: 65 male patients of not having heart disease were wrongly predicted as a having heart disease.

False-negative: 62 male patients of heart disease were wrongly predicted as not having heart disease.

True Negative: 348 male patients of not having heart disease were correctly predicted by the model.



The ROC curve's AUC is 0.87, suggesting that this model can accurately identify male heart disease 87% of the time.

**CONCLUSION:**

* In the initial analysis, several simple graphs and EDA were used to assist understand the dataset, and knowledge regarding heart disease and its symptoms was acquired from the dataset.
* We dug further into the research to see which factors influence heart disease.
* If you have the old peak, chest pain, sex, maximum heart rate, and exang, you may use the model built in this report to forecast heart disease.
* Furthermore, the data revealed that those in their forties and fifties had a higher risk of heart disease. It may be claimed that the danger is higher in this age because of the increased stress and job load.
* Surprisingly, females have more heart diseases than males in our database.
* The factors affecting the heart diseases in male are maximum heart rate, chest pain, old peak and thalassemia whereas in females the factors are chest discomfort, exercise-induced angina, thalassemia, and fasting blood sugar.

**REFERENCES:**

1. *Chi-squared Test of Independence | R Tutorial*. (n.d.). Chi-Squared Test of Independence | R Tutorial; www.r-tutor.com. Retrieved May 15, 2022, from <http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence>
2. Zach, & posts by Zach, V. all. (2021, April 1). *How to Create a Confusion Matrix in R (Step-by-Step)*. Statology; www.statology.org. <https://www.statology.org/confusion-matrix-in-r/>
3. *What is a ROC Curve and How to Interpret It - Displayr*. (2018, July 5). Displayr; www.displayr.com. <https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it>
4. - robk@statmethods.net, R. K. (n.d.). *Quick-R: Generalized Linear Models*. Quick-R: Generalized Linear Models; www.statmethods.net. Retrieved May 15, 2022, from <https://www.statmethods.net/advstats/glm.html>