**BUAN 6356.006 – Business Analytics with R (Group 16)** **Halfway Report**

**Team members:**

Draksharam, Vishnu Paschyanti (vxd200023)

Jawahar Vasagam, Premi (pxj220007)

Farzana M B, Ashika (axf210029)

Tallapally, Jahnavi (jxt200051)

Parthasarathi, Prriyamvradha (pxp220005)

**Dataset:**

<https://www.kaggle.com/code/namanmanchanda/heart-attack-eda-prediction-90-accuracy/data>

**Data Pre-processing:**

We loaded the selected dataset from Kaggle into R

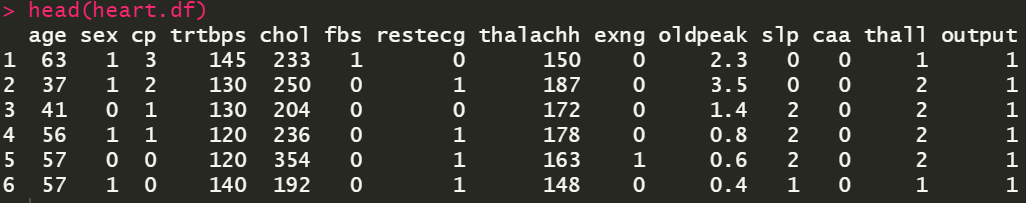
We found no null values in the dataset, so we did not have to remove any null values.



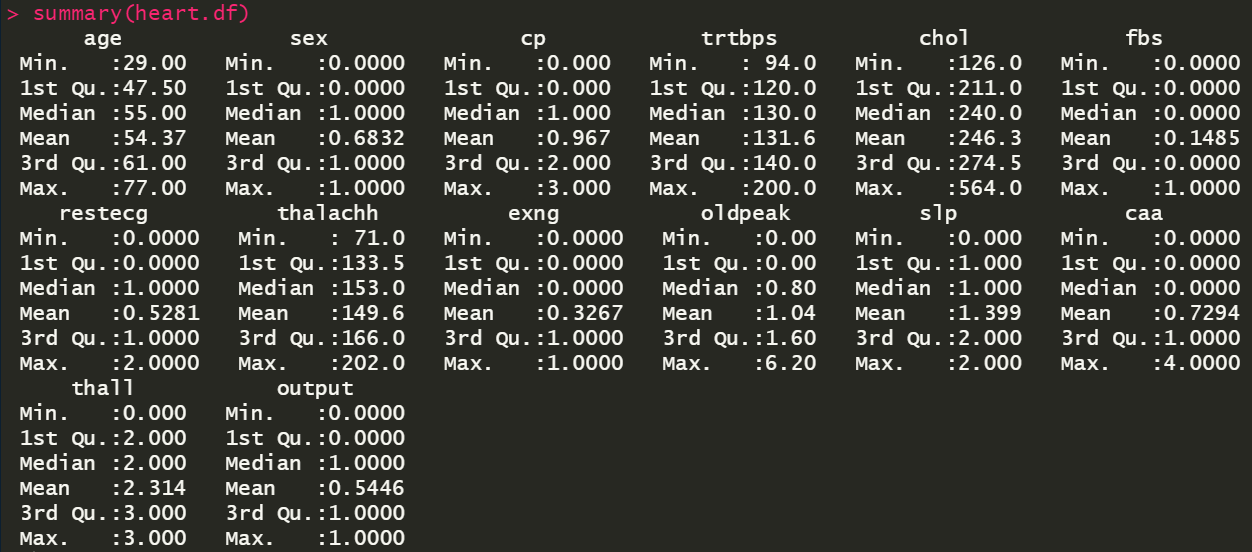
**Exploratory Data Analysis:**

**Summary:**

The following image shows the sample data of the Heat Attack prediction dataset



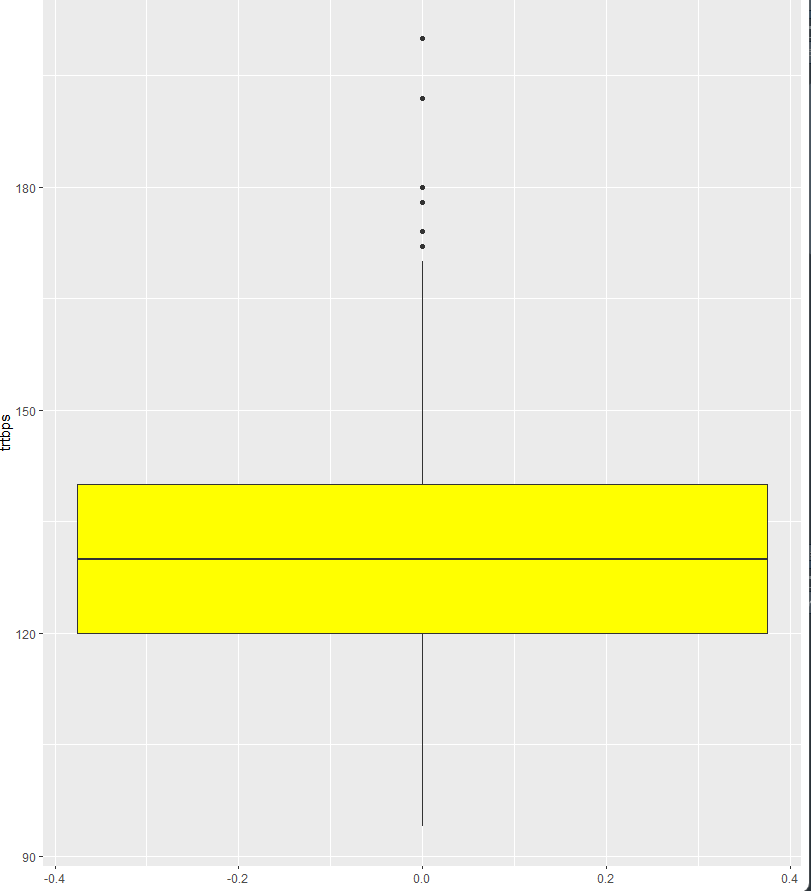
The following image shows the summary statistics for the different variables that we have in our dataset.



**Finding Outliers using Box Plot:**

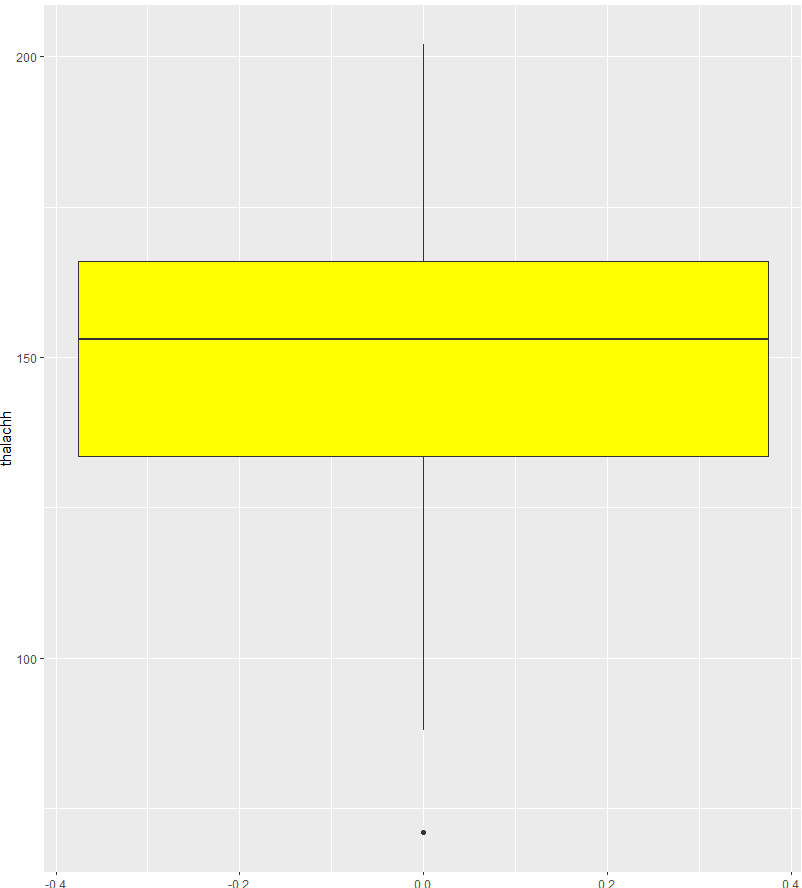
1. **Trtbps**

Resting Blood Pressure (in mm Hg)



We can see that we were able to observe 6 outliers for resting blood pressure out of which 3 are 180 or above. Since blood pressure readings more than 180/110 are considered hypertensive emergencies, we do not remove these outliers.

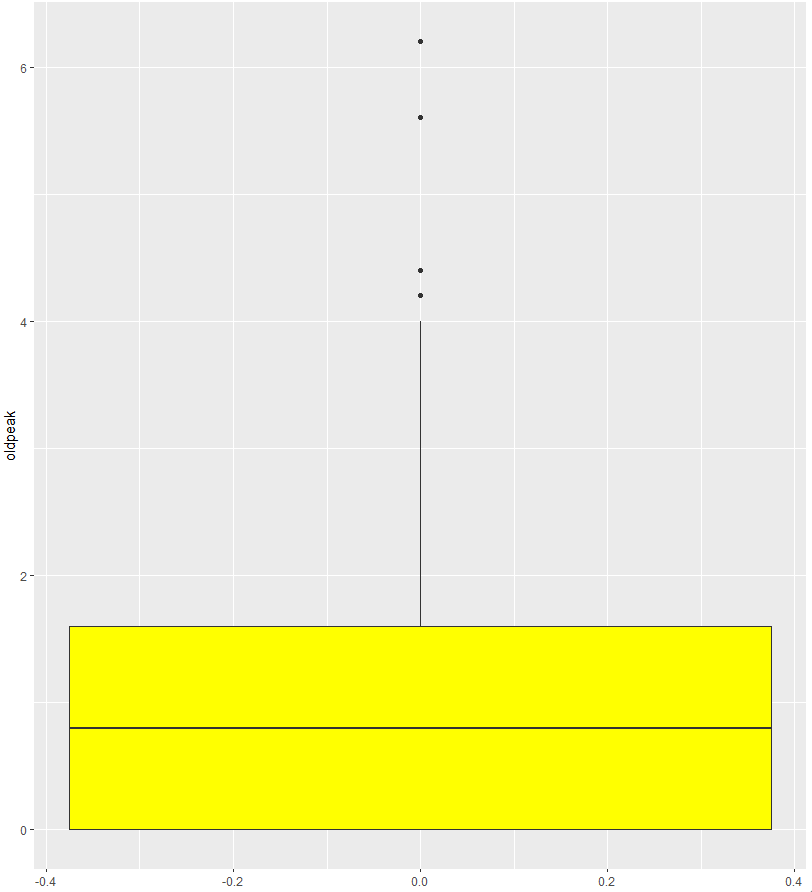
1. **Thalachh**

Maximum Heart Rate Achieved (Scale of 71 – 202) 

We observe one outlier that is below 75 but since it lies in the normal resting heart rate range, we do not remove it since it indicates a healthy subject

1. **Oldpeak**

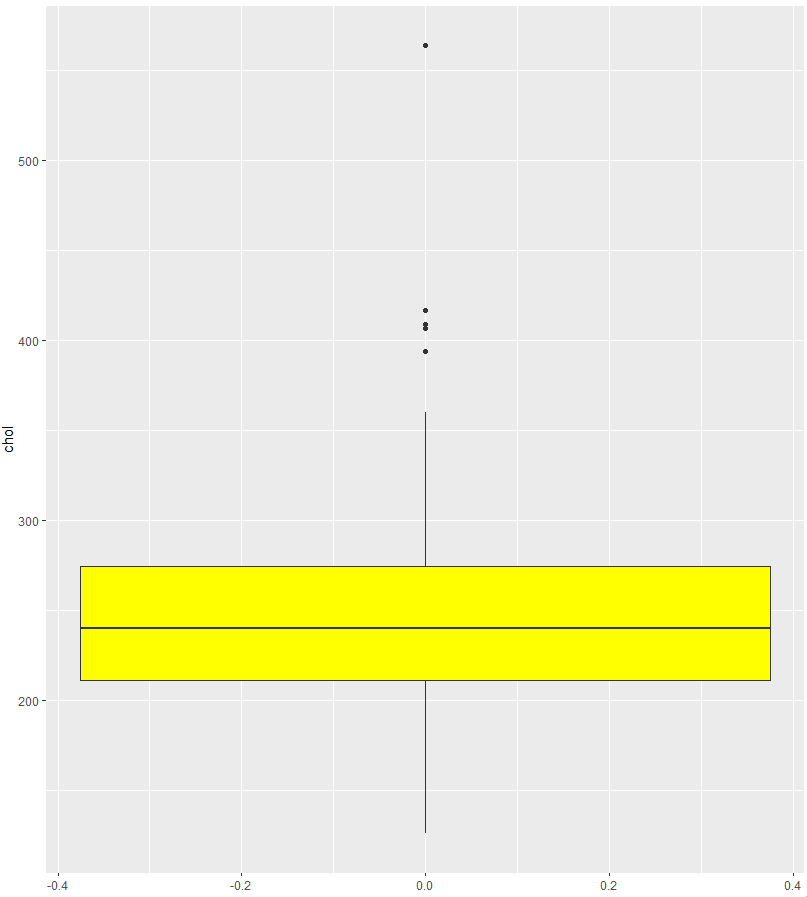
ST Depression Induced by Exercise Relative to Rest



We observe 4 outliers that are greater than 4 in this variable but since higher oldpeak values can indicate serious heart conditions, we do not remove them.

1. **Chol**

Cholesterol (in mg/dl)

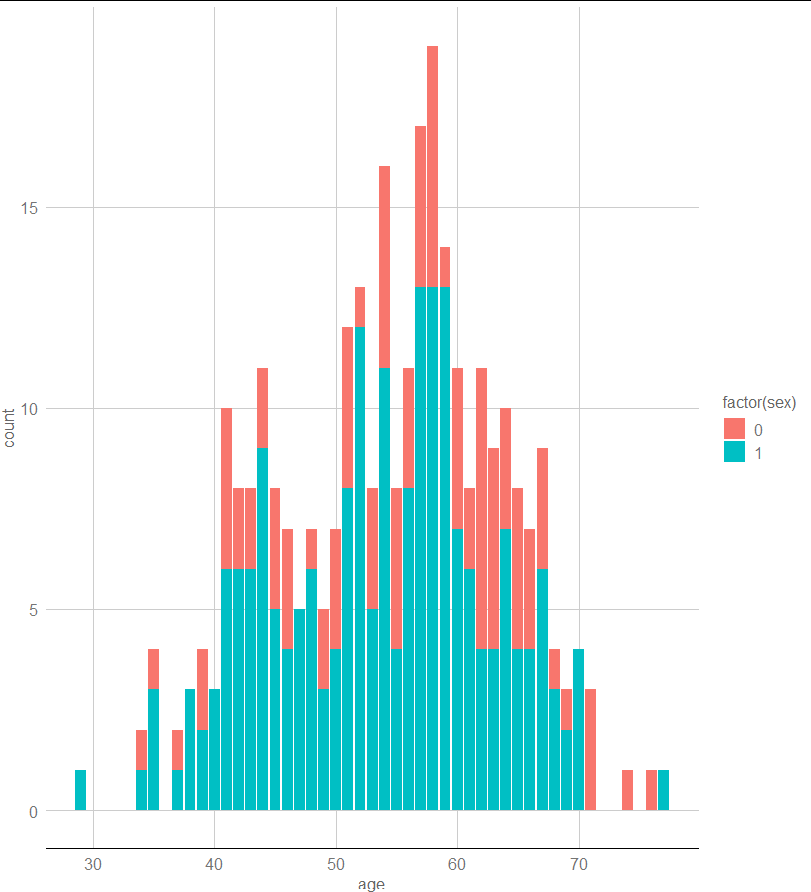


We observed 5 outliers that are around 400 or more. The normal cholesterol levels are around 200-240 so we do not remove these outliers since they represent critical levels which would require immediate attention.

**Ggplot – Histogram**

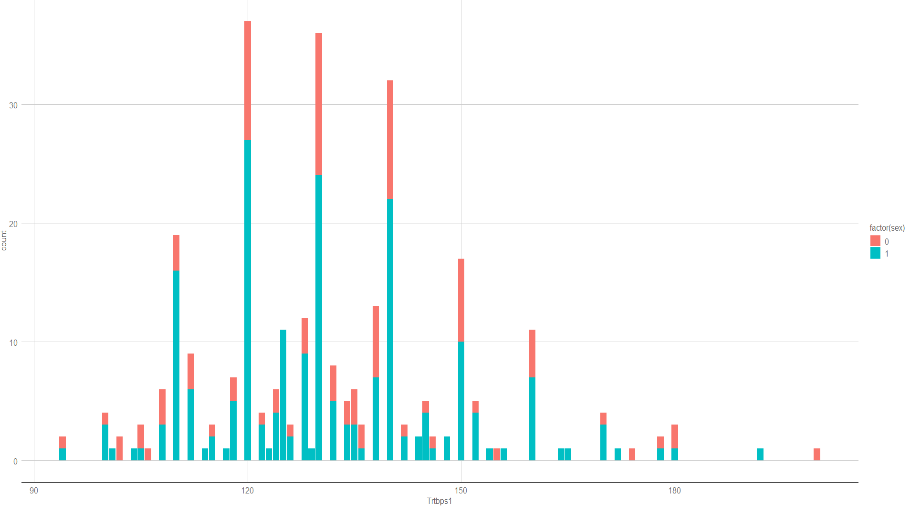
1. **Age**

The histogram below shows us the graphical representation of the data gathered on age. The minimum value is 29 years, mean is 54.37 years, and the maximum value is 77 years.



1. **Trtbps**

This graph shows us the resting blood pressure graphical data. We can see that the minimum value is 94 units, mean is 131.6 units, and the maximum value is 200 units.



**Scatterplot**



As shown above, elderly women between the ages of 60-70 have higher chances of having cholesterol. Whereas coming to men, we cannot really infer which ages of men have higher chances of cholesterol.

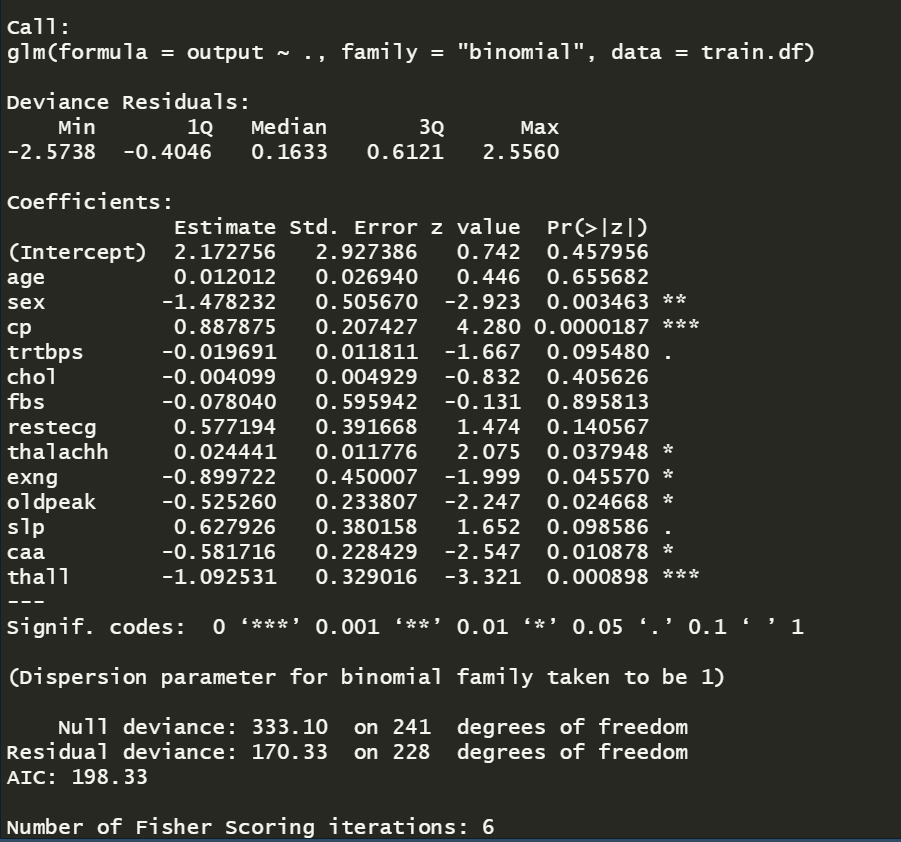
**Data Modelling**

**Splitting the dataset:**

The data set was first split into two parts: training data set and validation data set. 80% of the data set was used for training, while the remaining 20% was used for validation data set.

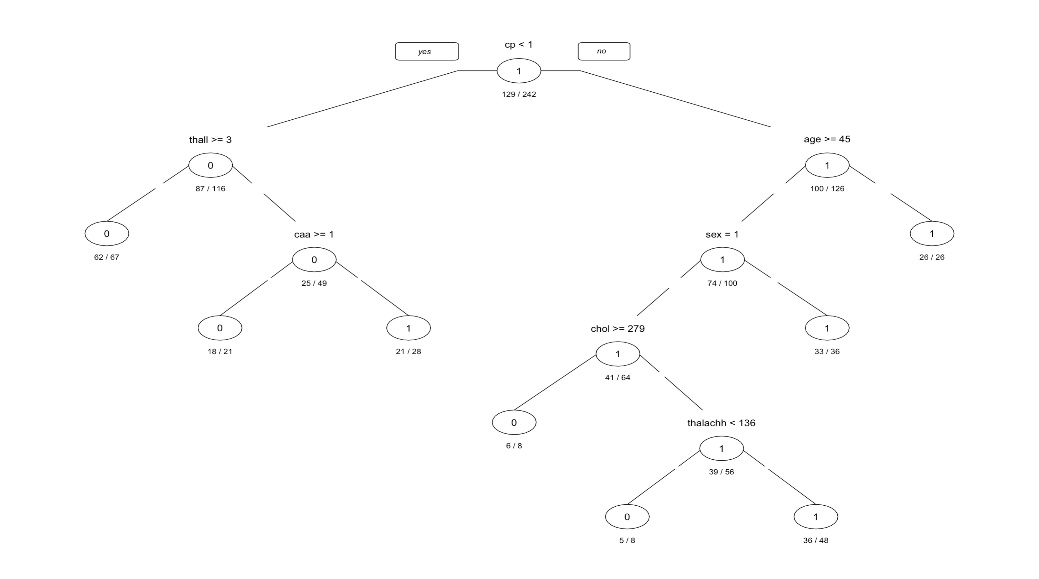
**Logistic Regression:**

It extends the idea of linear regression to the situation where the outcome variable is categorical and the appropriate regression analysis to conduct when the dependent variable is binary. Sex, cp, trestbps, ca, thalach, exang, slope are significant variables. The null deviance shows how well the response is predicted by the model with nothing but an intercept. The residual deviance shows how well the response is predicted by the model when the predictors are included. Residual deviance is a measure of error. The smaller the residual deviance, the better the predictive power of the model. In the output we get the residual deviance smaller than the null deviance, or a logistic regression model has some predictive power, and the variables will have some explanatory power. Below is a summary of the logistic regression model which has an accuracy of 82%.



**Decision Tree**

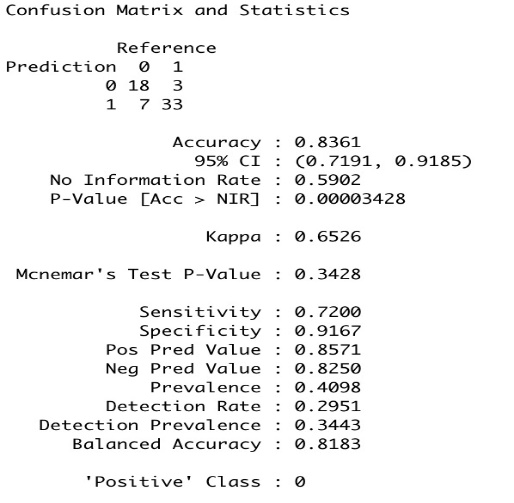
The most effective and well-known technique for categorization and prediction is the decision tree. A decision tree is a flow chart like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. From the decision tree we get the following rule with the most percentage cover of cases.



**Confusion Matrix:**

The confusion matrix is a measurement that tracks the number of errors, including false positives and false negatives, to show how well a classification model performs. We display the accuracy of our training data using the confusion matrix.

**Confusion matrix for logistic regression:**



**Confusion Matrix for Decision Matrix:**

