**HEART DISEASE PREDICTION**

Group Members

AAKASH SARAP

ANISH NITIN SOMAIAH

SHIVANI GROVER

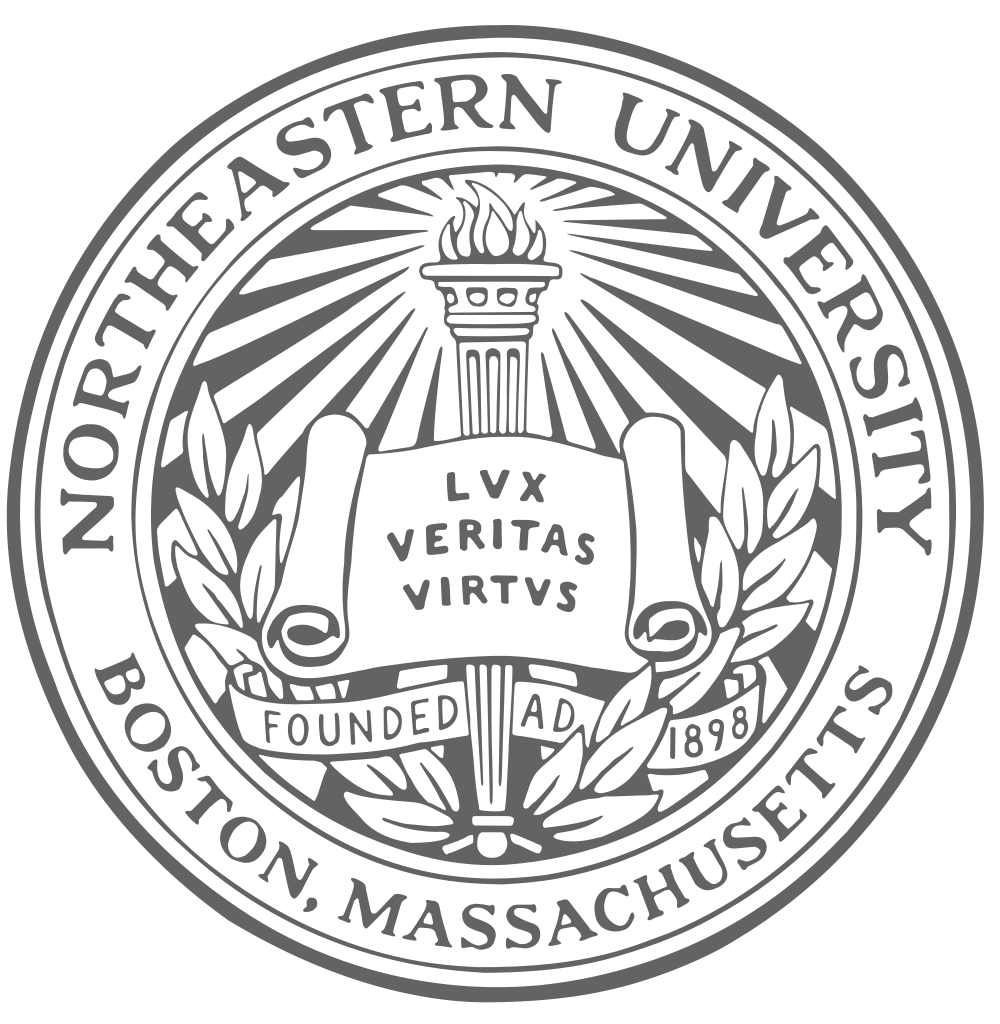
SUSHMITA JADHAV

WEEK 6: FINAL PROJECT

ALY 6015: 21282 Intermediate Analytics

Instructor: VALERIY SHEVCHENKO

Date: February 17, 2019



**INTRODUCTION**

Heart Disease refers to a condition which affects the normal functioning of the heart. Some of the heart diseases include blood vessels, such as coronary artery disease; rhythmic problems, birth defects related to heart and so on. Heart disease also includes narrowing or blocking of blood vessels leading to heart attack, chest pain or stroke. Some other conditions which involve weakening of heart muscles are also termed as heart disease.

Business Questions to be answered:

* **What are the factors that affect heart disease?**
* **What is the relationship between factors and the possibility of heart disease?**
* **How strong is the relationship between factors and the possibility of heart disease?**

**About the dataset:**

The dataset is retrieved from the UCI Machine Learning Repository. We have chosen this topic in order to predict whether an individual would face Heart Disease in the future or not. The dataset contains 14 variables namely, age, sex, chest pain, resting blood pressure, fasting blood sugar, cholesterol, resting ECG, maximum heart rate, exercise induces angina, old peak, the slope of peak exercise, number of major vessels, thallium heart scan and result.

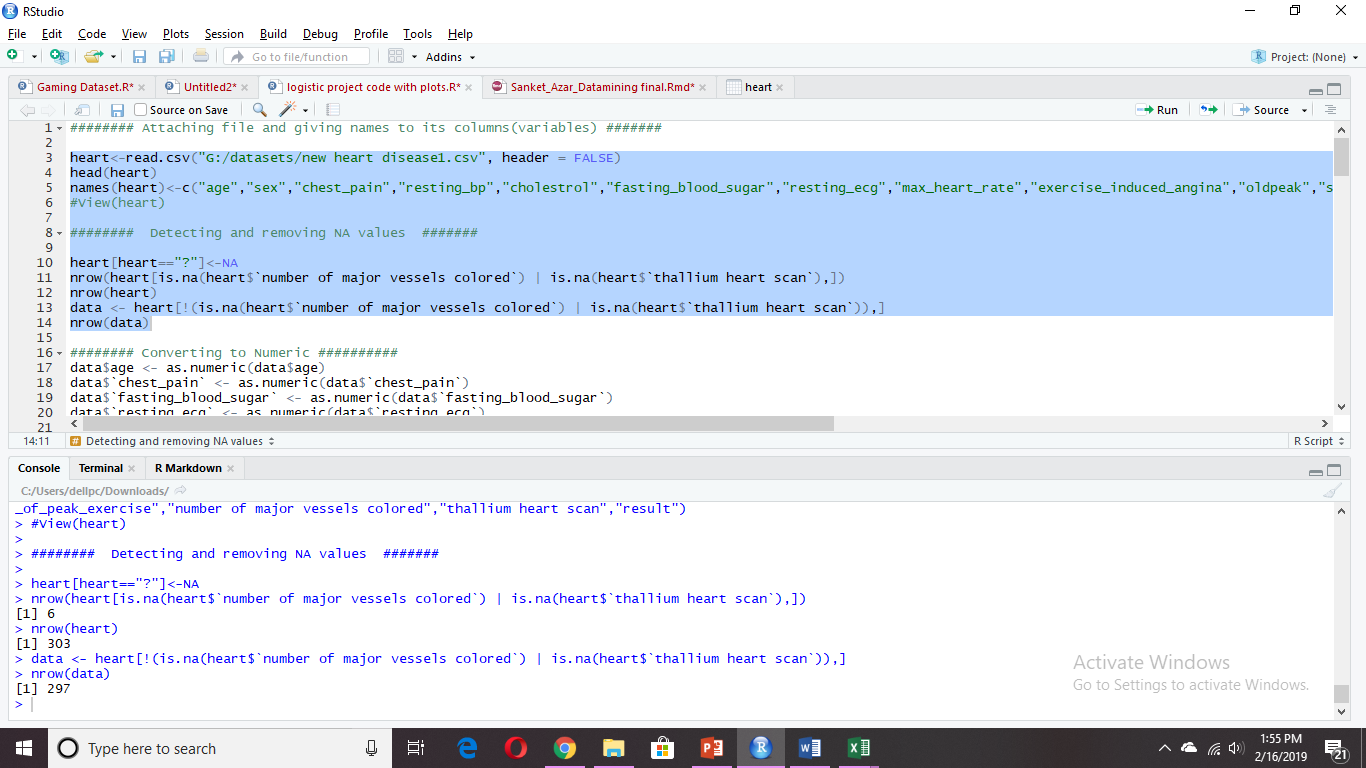
The variable definition is as follows:

|  |  |
| --- | --- |
| **Age** | the age group analyzed is 29- 77 with maximum people of 58-59 years old. |
| **Sex** | sex=0; individual is female |
|  | sex=1; individual is male |
|  | There are 97 females and 206 males analyzed is this dataset |
| **Chest\_pain** | It defines the type of chest pain experienced |
| **Resting\_bp** | Blood pressure in mm Hg on admission to the hospital |
| **Fasting\_blood\_sugar** | fasting blood sugar > 120 mg/dl - 1 = true; 0 = false |
| **Cholesterol** | serum cholesterol in mg/dl |
| **Resting\_ecg** | resting electrocardiographic results |
| **Max\_heart\_rate** | maximum heart rate achieved |
| **Exercise\_induced\_angina** | (1 = yes; 0 = no) |
| **Old peak** | ST depression induced by exercise relative to rest |
| **Slope\_of\_peak\_exercise** | the slope of the peak exercise ST segment |
| **Number of major vessels** **colored** | number of major vessels (0-3) colored by fluoroscopy |
| **Thallium heart scan** | 3 = normal; |
|  | 6 = fixed defect; |
|  | 7 = reversable defect |
| **Result** | The result variable is the prediction which identifies as positive or negative for an individual suffering with any heart disease. |
|  | If result = 0; no heart disease |
|  | result = 1; has heart disease |

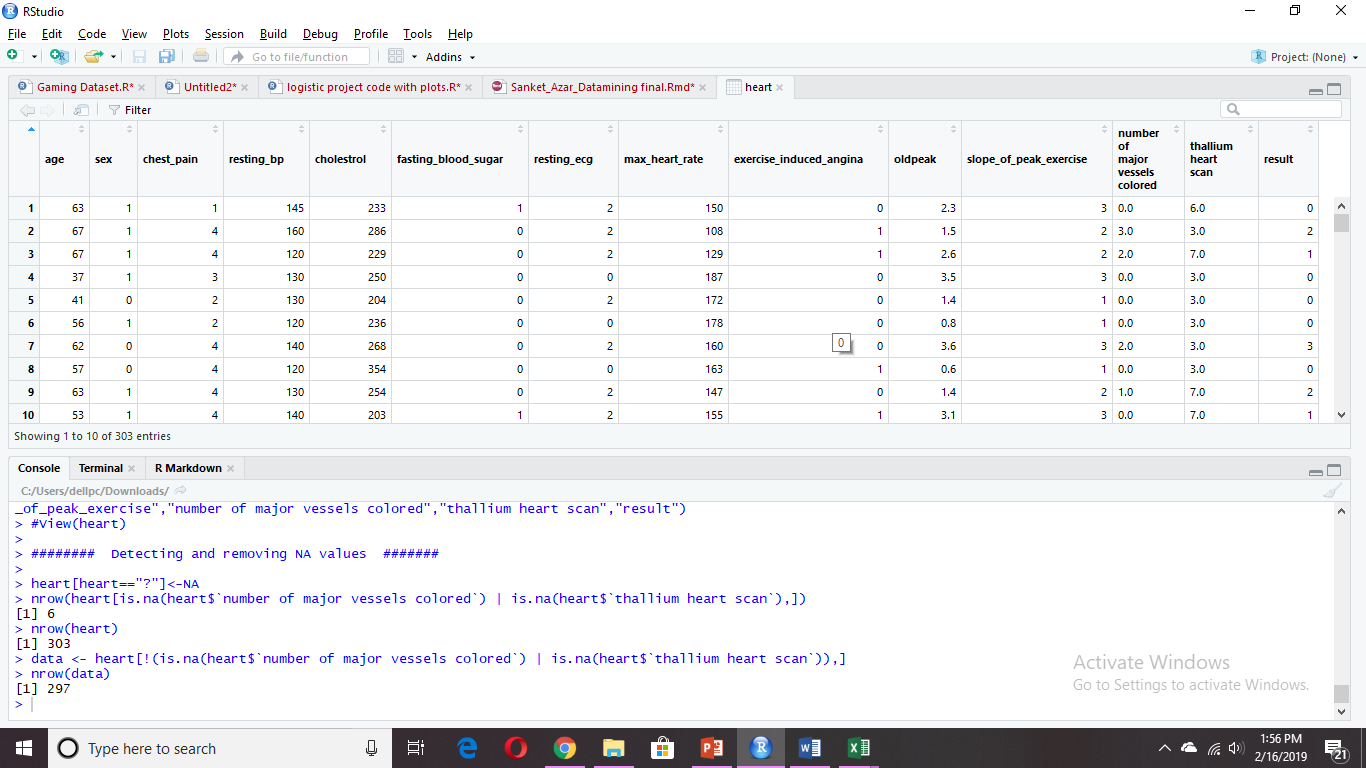
**Data Cleaning**

It refers to removing the undesired variables and NA values from the dataset. Our dataset had ‘?’ for no value. We converted it to NA And then removed them using is.a() function. We only had 6 NA values which were just 2% of the entire data, so, we removed those 6 rows. Initially, we had 303 observations, after removing those NA values 297 observations are used for analysis.

Also, to facilitate smooth access, the variables were named using names() function.



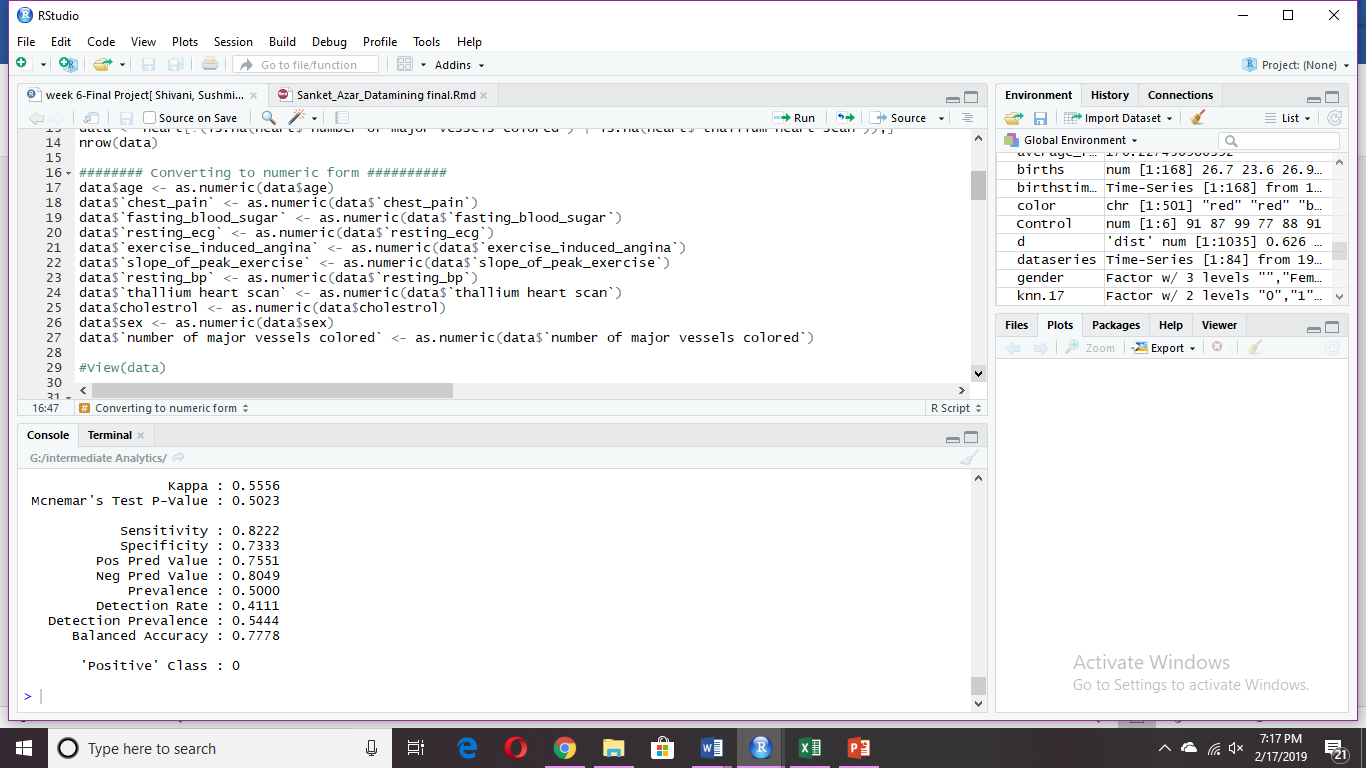
This is the dataset we have selected.



**ANALYSIS**

**KNN classification**

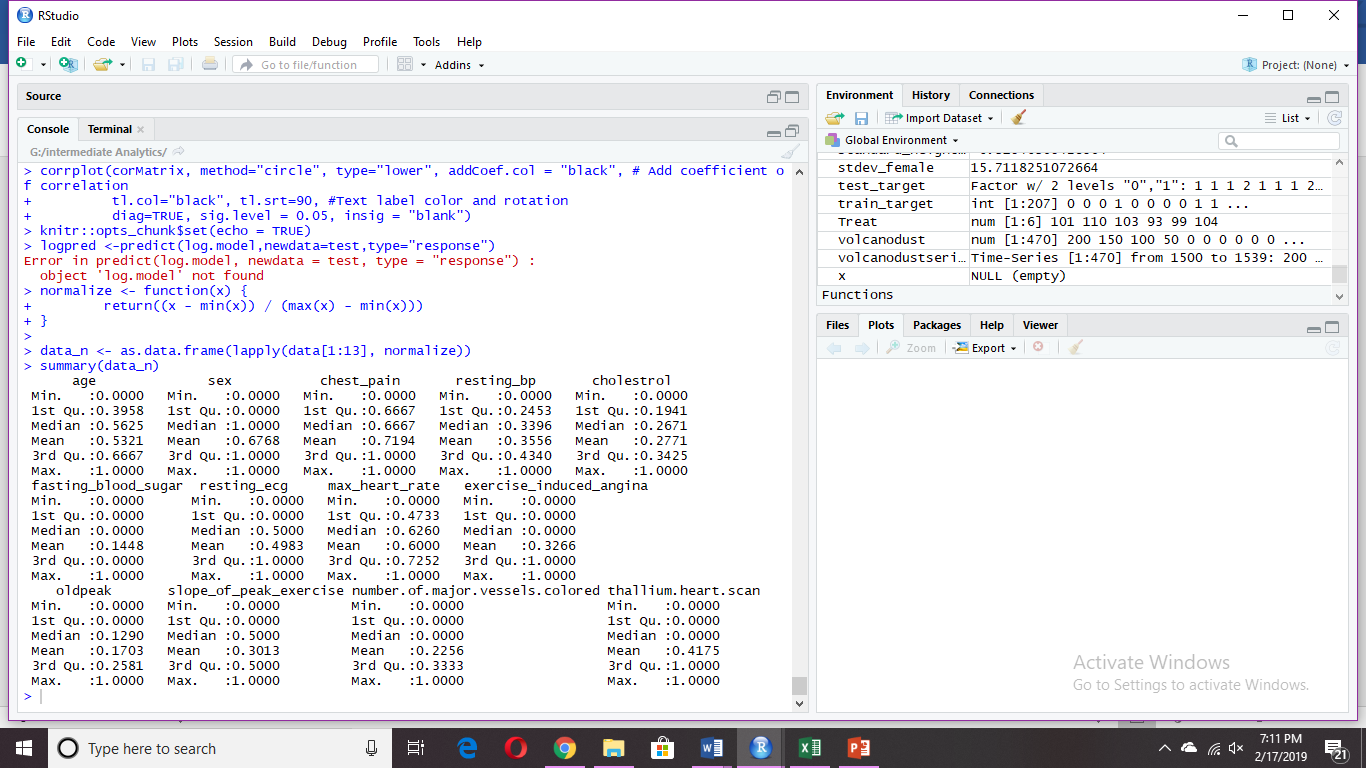
To implement KNN classification, all the desired variables should be converted to numeric form.



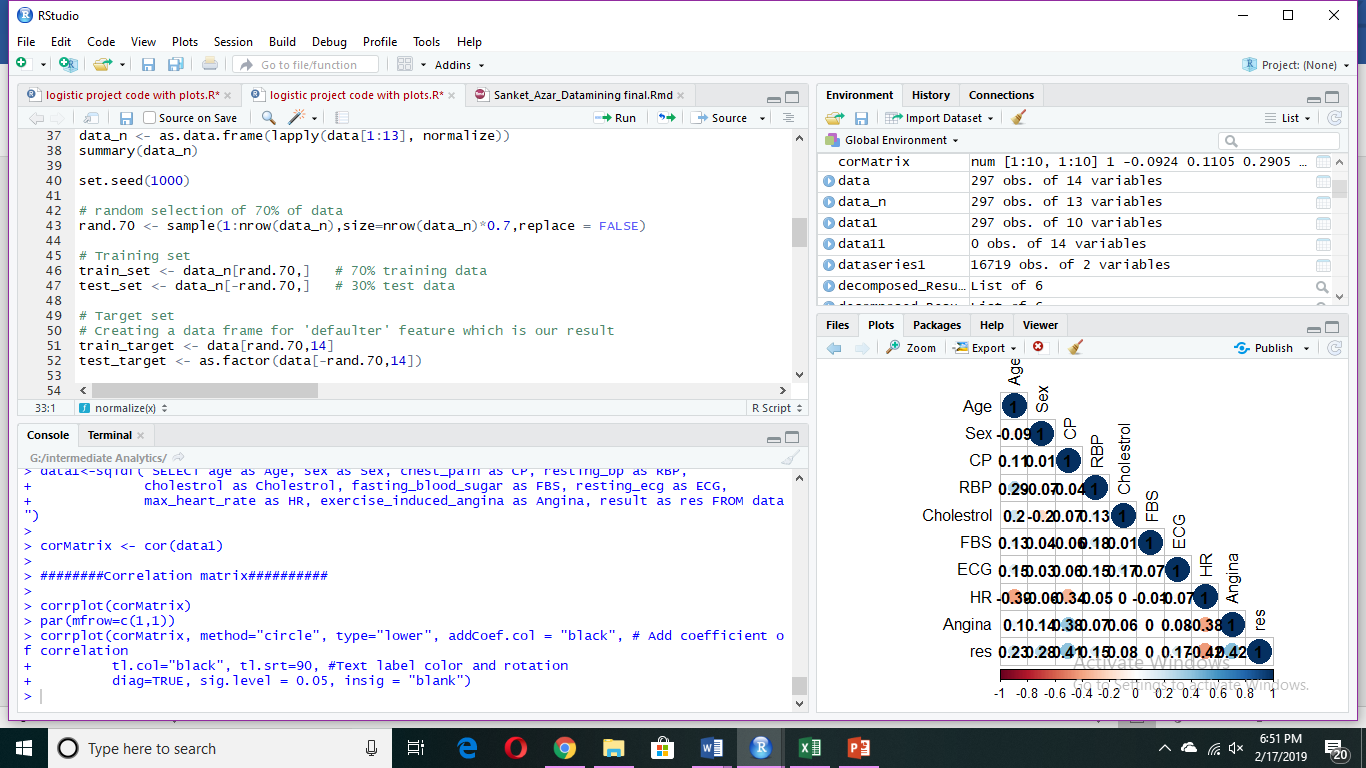
The parameters in the dataset have a different kind of scales, thus we normalized the data.

As we can see the below summary for ‘data\_n’, all the values have been normalized and lie between 0 & 1.

**Normalized data variables**

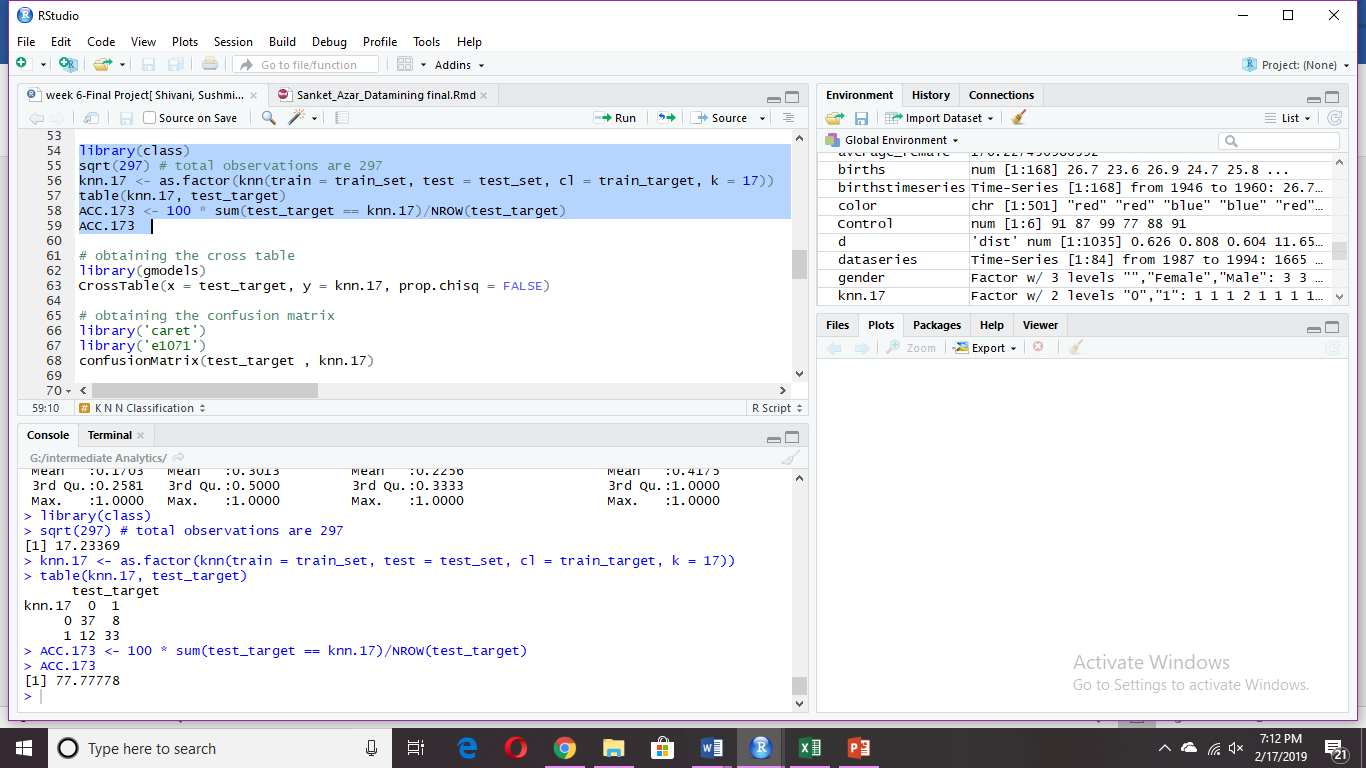


The dataset is divided into test and train dataset such that 70% of randomly selected data points are in train set and 30% of them in the test set so that the model created with train dataset could be cross-verified with test dataset. We have used the sample() to select the random sample data and the set.seed(1000) so as to fetch the same random sample every time we run the code.

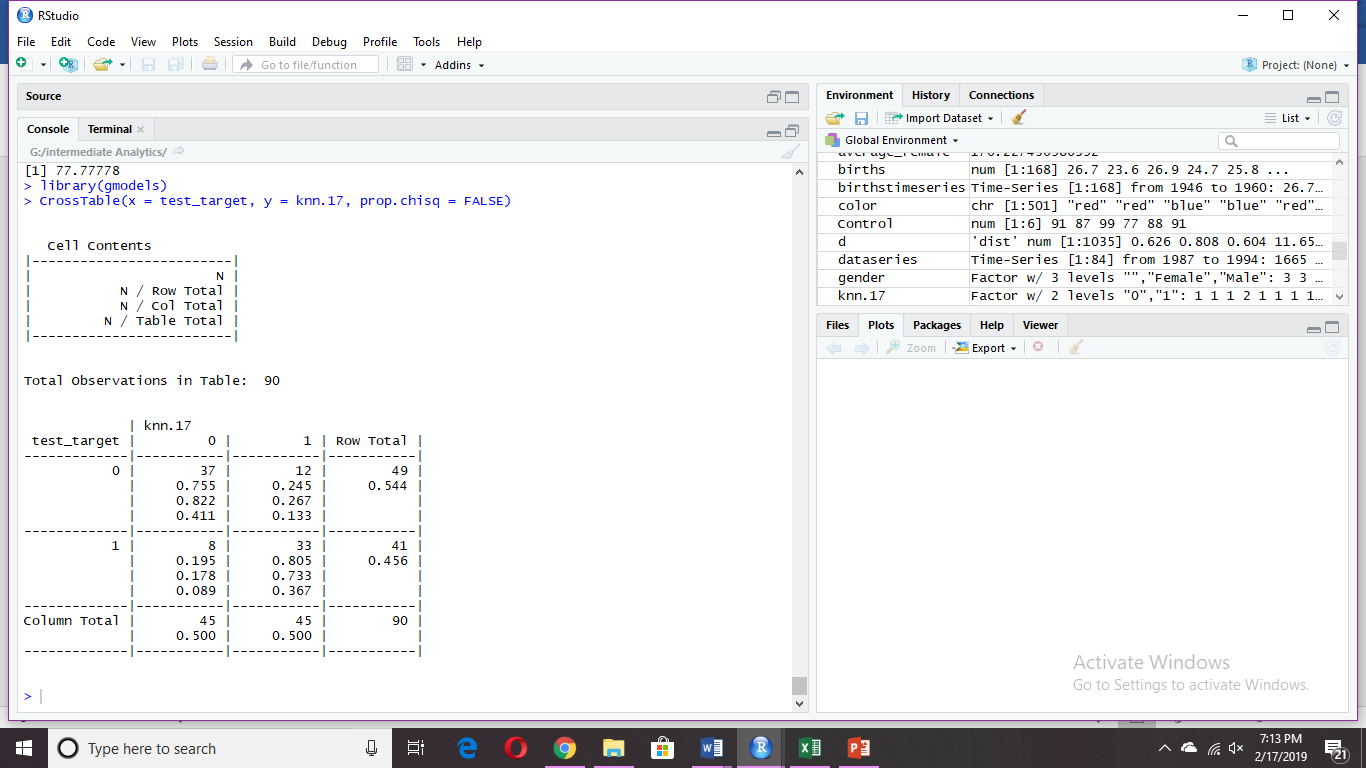


**Implementing KNN- classification**

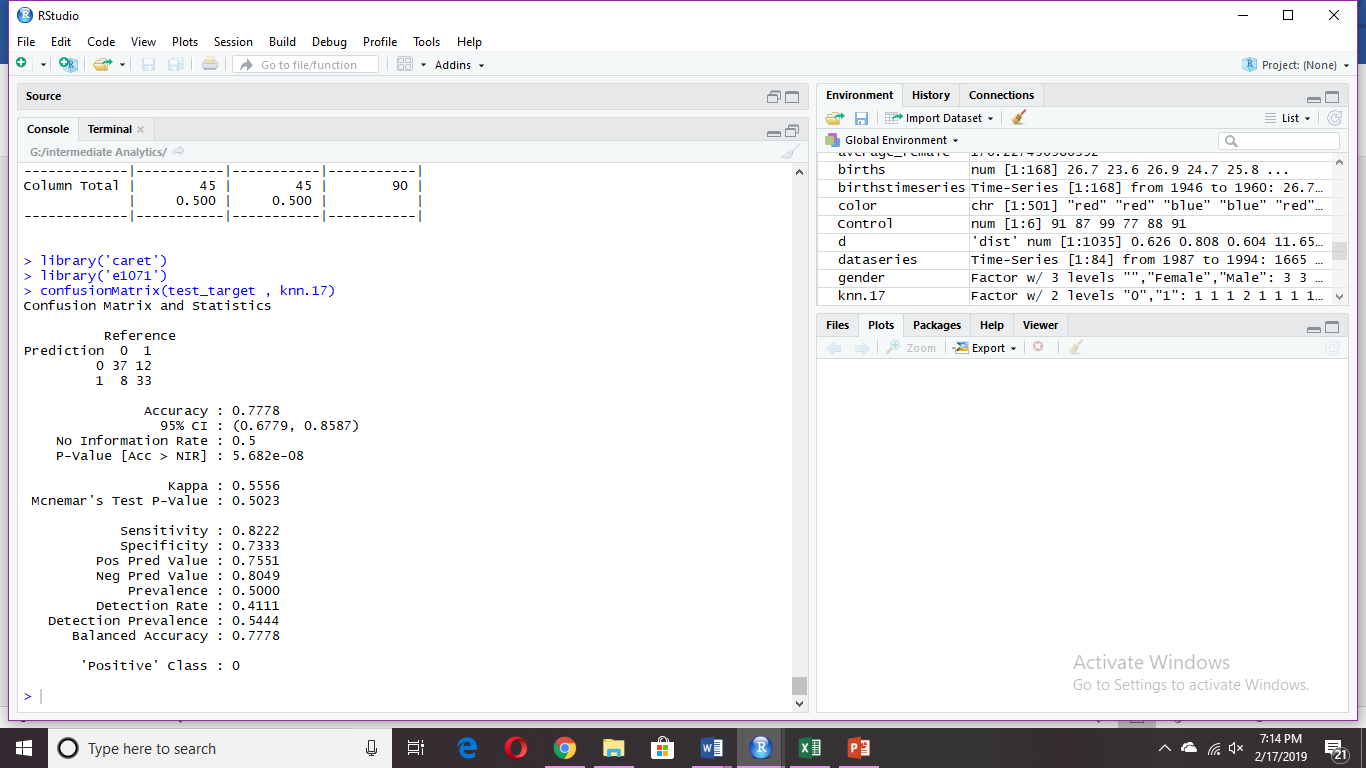
We need to identify the optimum value of k to minimize the error. Generally, we take k as an odd number nearest to the square root of the total number of observations. So, we take k = 17.



**Obtaining the Cross Table**



**Obtaining the Confusion matrix**

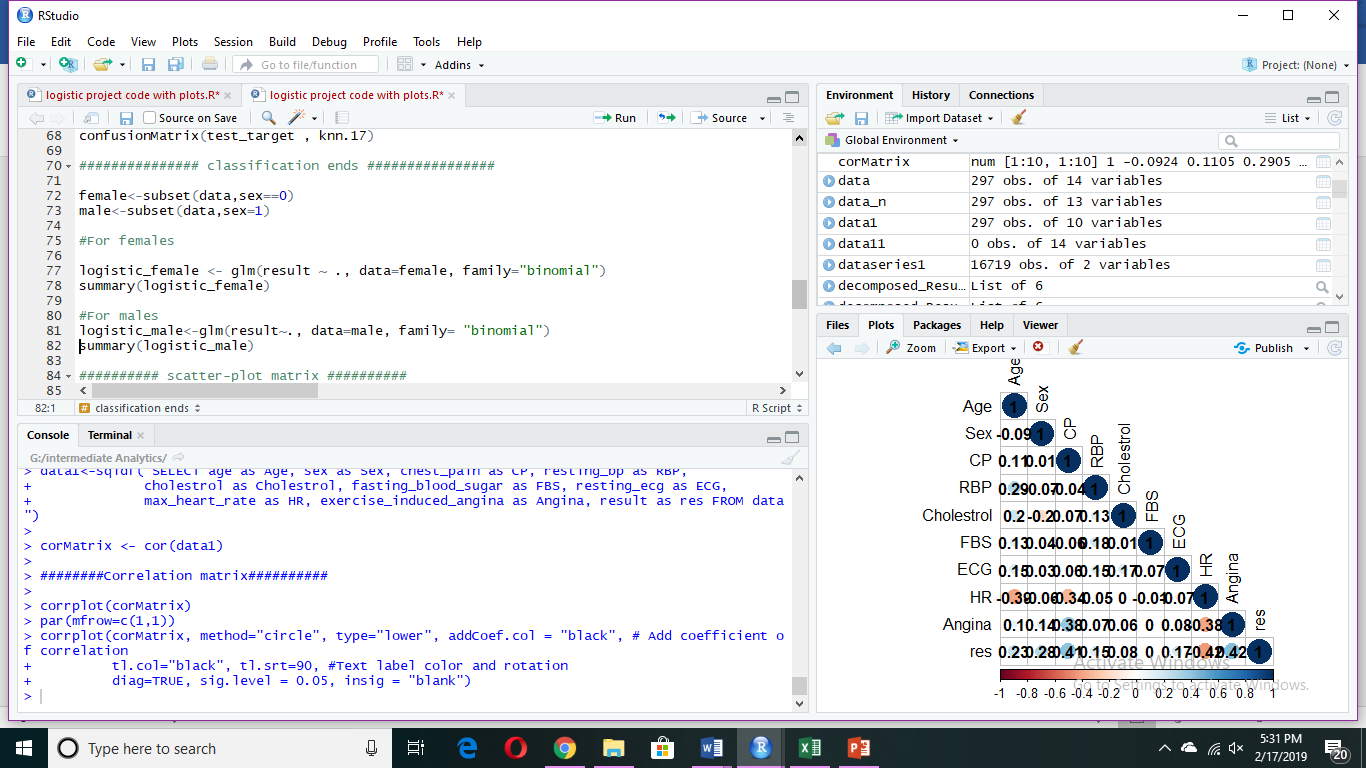


**Interpretation**

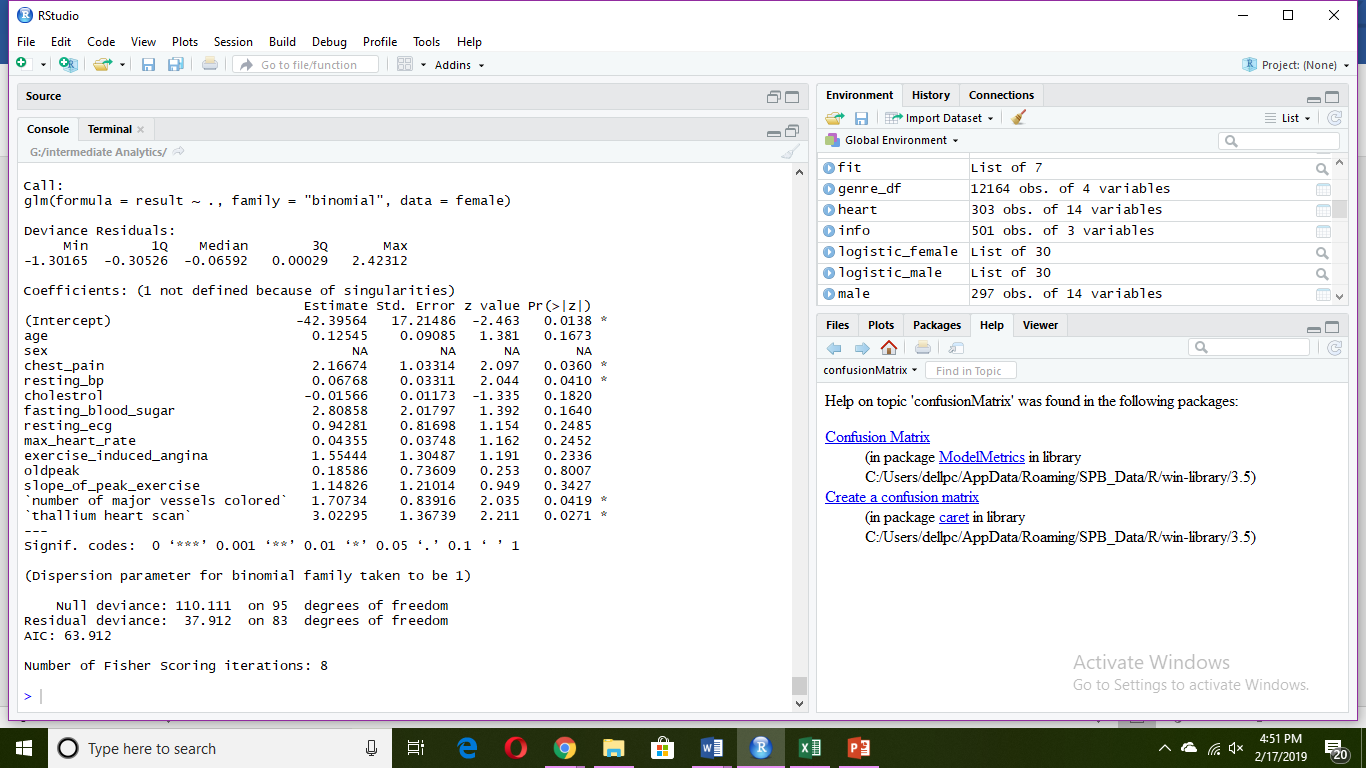
* From the cross table, we can infer that our Test data consisted of 90 observations.
* Out of which37 cases have been accurately predicted (True Negatives) as patients without heart disease. Also, 33 cases out of 90 were accurately predicted (True Positives) as the patients with Heart Disease. While 20 cases were incorrectly predicted, that is, 12 of them were predicted to have heart disease when they did not have and 8 were not predicted of having heart disease while they had the disease.
* Our KNN prediction classification model has an accuracy of 77.78% as shown in the above Confusion Matrix at a confidence level of 95%
* Moreover, sensitivity (proportion of people with the disease and positive result) of the test is 82.2% and the specificity (proportion of people without disease and negative result) of the test is 73.3%.
* Balanced accuracy and actual accuracy are the same indicating that the accuracy cannot be improved than the acquired 78% value.

**Implementing Logistics Regression**

Logistic Regression is considered for males and females separately (considering sex to be a dominant factor)



Output for Females dataset



**Regression Model (Females)**

Y=2.166\*(chest\_pain) +0.067\*(resting\_bp)+1.71\*(number of major vessels colored) + 3.02\*(thallium heart scan) – 42.39

P(Y)=1/1+e^(-Y) (model equation for result)

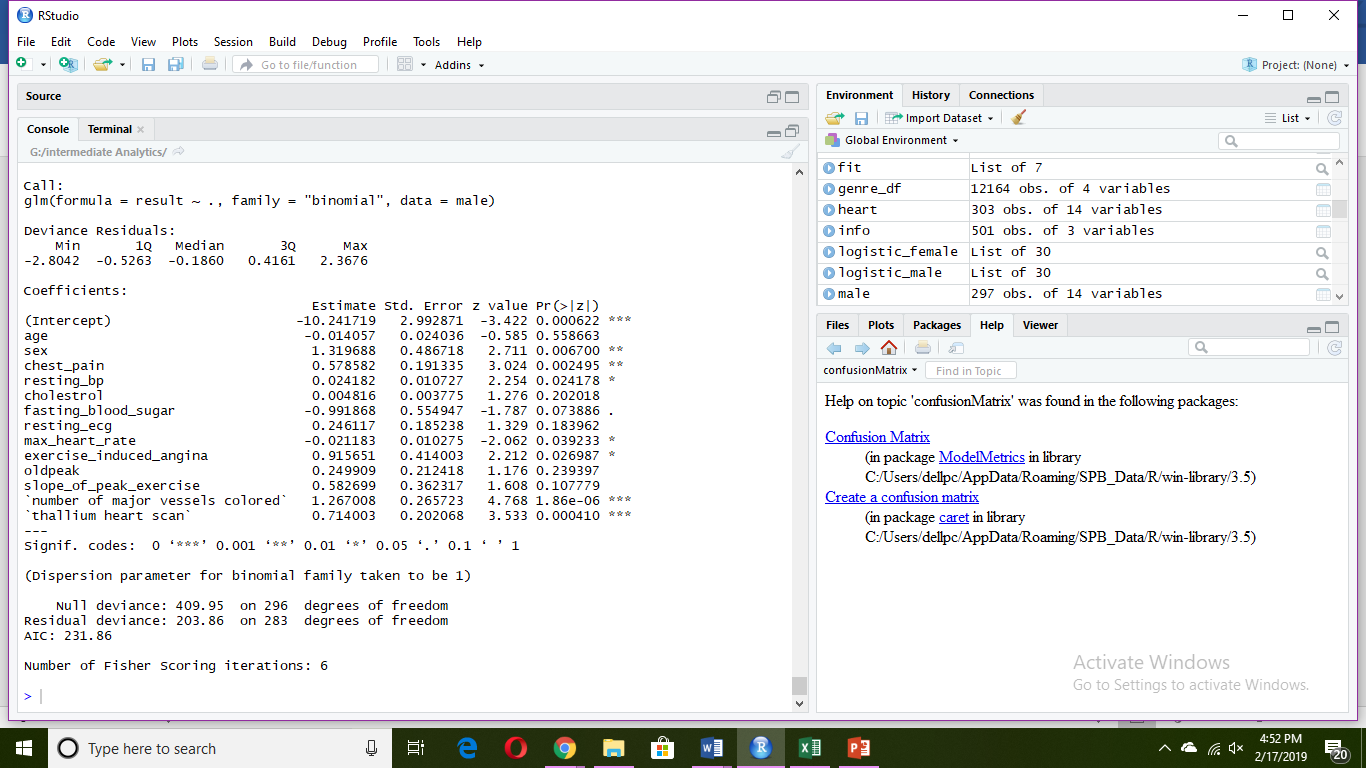
* Logistic regression is done for the female subset to find out the best predictor variables for our given dataset.
* It has been observed that 4 out of 13 variables form the optimum predictor variables for the female subset, namely chest pain, resting blood pressure, number of major vessels colored and thallium heart rate. The regression model has been shown above using the coefficients obtained.
* The variables affecting can be determined from their p-values obtained after performing z-test. For logistic regression, the Null hypothesis is: the result variable is independent upon the variable considered.

Alternate hypothesis: the result variable is dependent upon the variable considered.

For the given four variables the p-values are less than 0.05 for the 95% significance level. Therefore, we reject the null hypothesis for all these variables and accept the alternate hypothesis and create a model considering these four variables.

* From the above output we also obtain the min= (-1.30); max= (2.42); median= (-0.0659); quantile1= (-0.305); quantile3= (0.00029) and the degrees of freedom =83 for the residual deviance.

Output for Males dataset



**Regression Model (Males)**

Y=0.578\*(chest\_pain) + 1.26\*(number of major vessels colored) + 0.71\*(thallium heart scan) +0.02418\*(resting\_bp) – 0.02118\*(max\_heart\_rate) + 0.9156\*(exercise\_induced\_angina) – 0.9918\*(fasting\_blood\_sugar) - 10.2

(Y)=1/1+e^(-Y) (model equation)

Logistic regression is done for the male subset to find out the best predictor variables for our given dataset.

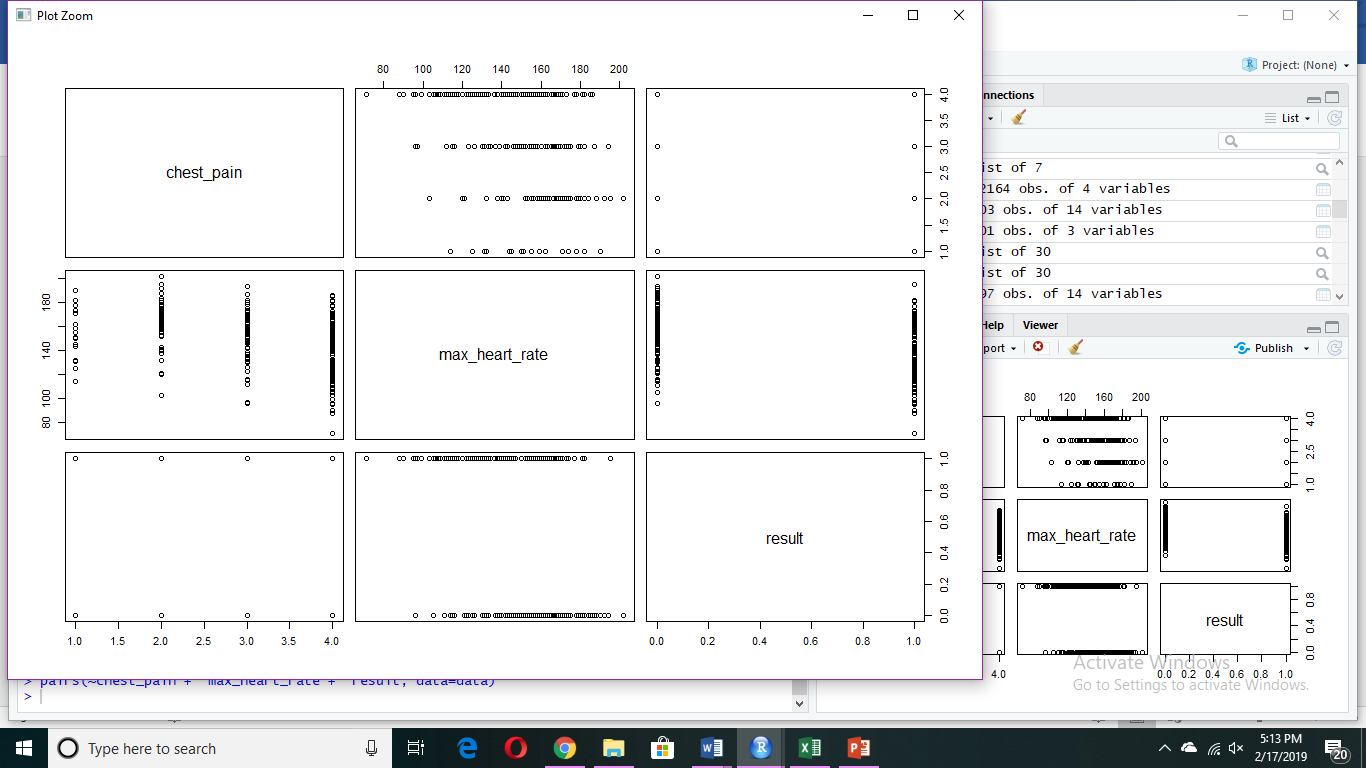
* It has been observed that 6 out of 13 variables form the optimum predictor variables for the male subset, namely chest pain, resting blood pressure, a number of major vessels colored and thallium heart rate, maximum heart rate, and angina induced due to heavy exercise. The regression model has been shown above using the coefficients obtained.
* The variables affecting can be determined from their p-values obtained after performing z-test. For logistic regression, the Null hypothesis is: the result variable is independent upon the variable considered.

Alternate hypothesis: the result variable is dependent upon the variable considered.

For the given four variables the p-values are less than 0.05 for the 95% significance level. Therefore, we reject the null hypothesis for all these variables and accept the alternate hypothesis and create a model considering these four variables.

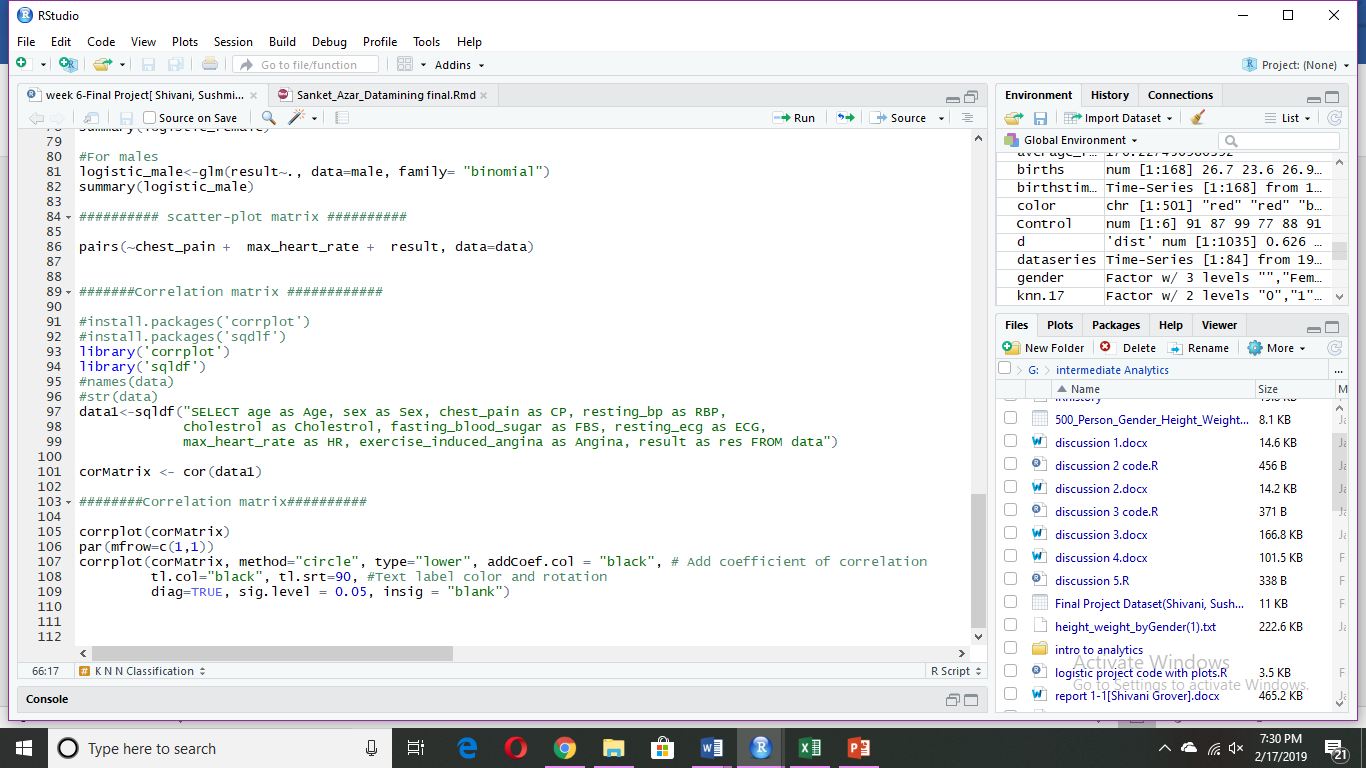
* From the above output we also obtain the min= (-2.80); max= (2.36); median= (-0.186); quantile1= (-0.526); quantile3= (0.416) and the degrees of freedom =283 for the residual deviance.
* The effect of fasting\_blood\_sugar on the result is not very significant. But, in order to obtain an accurate model, all the affecting factors are considered.

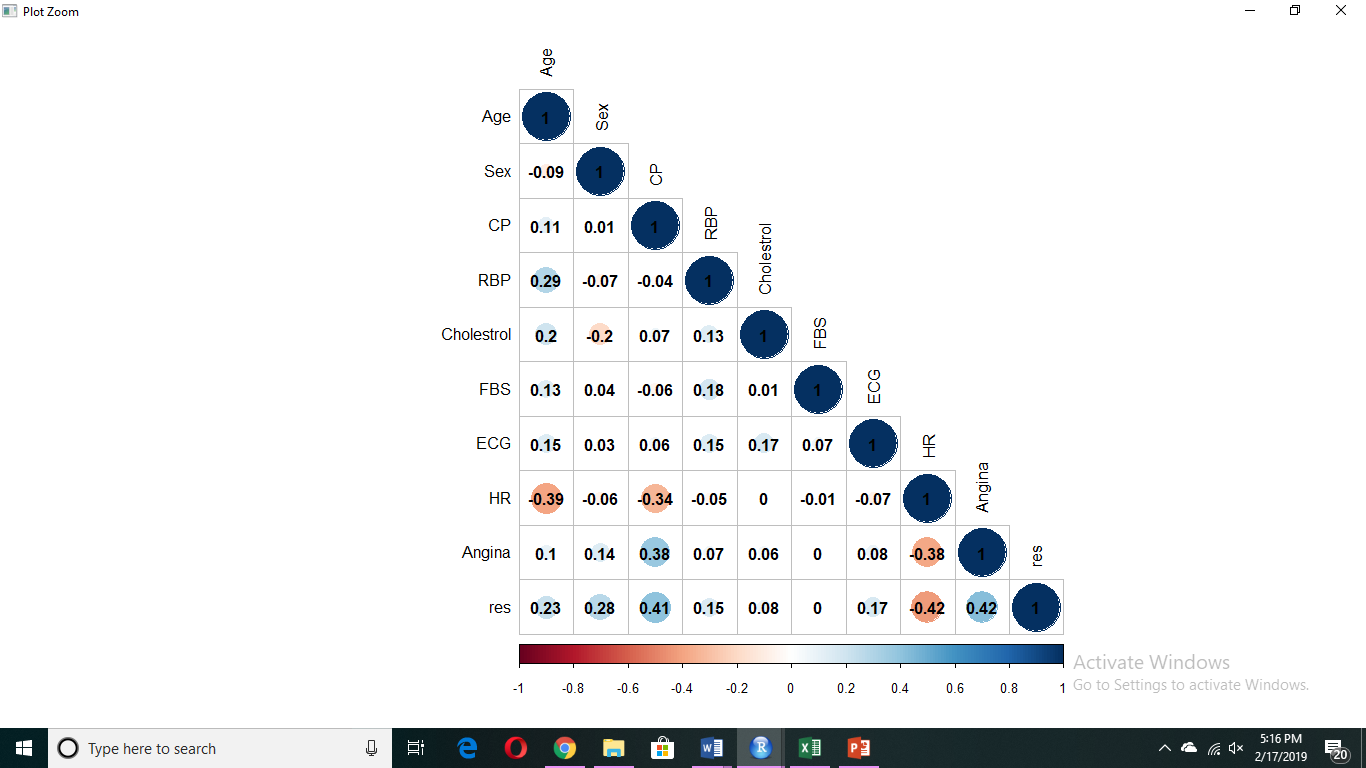
**Plotting the scatter plot matrix and correlation matrix**



The above is a matrix of scatter plots which shows the relationship between result with chest\_pain and max\_heart\_rate (on separate plots).

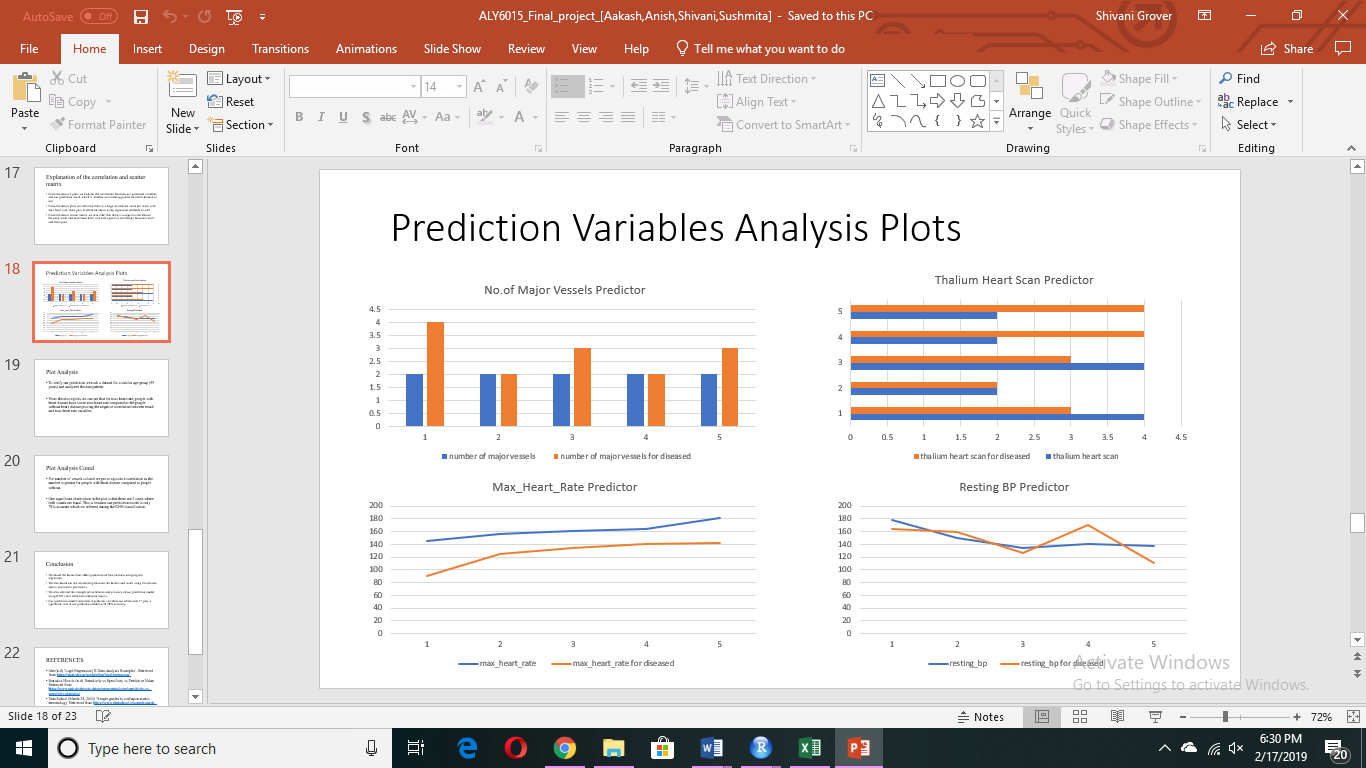
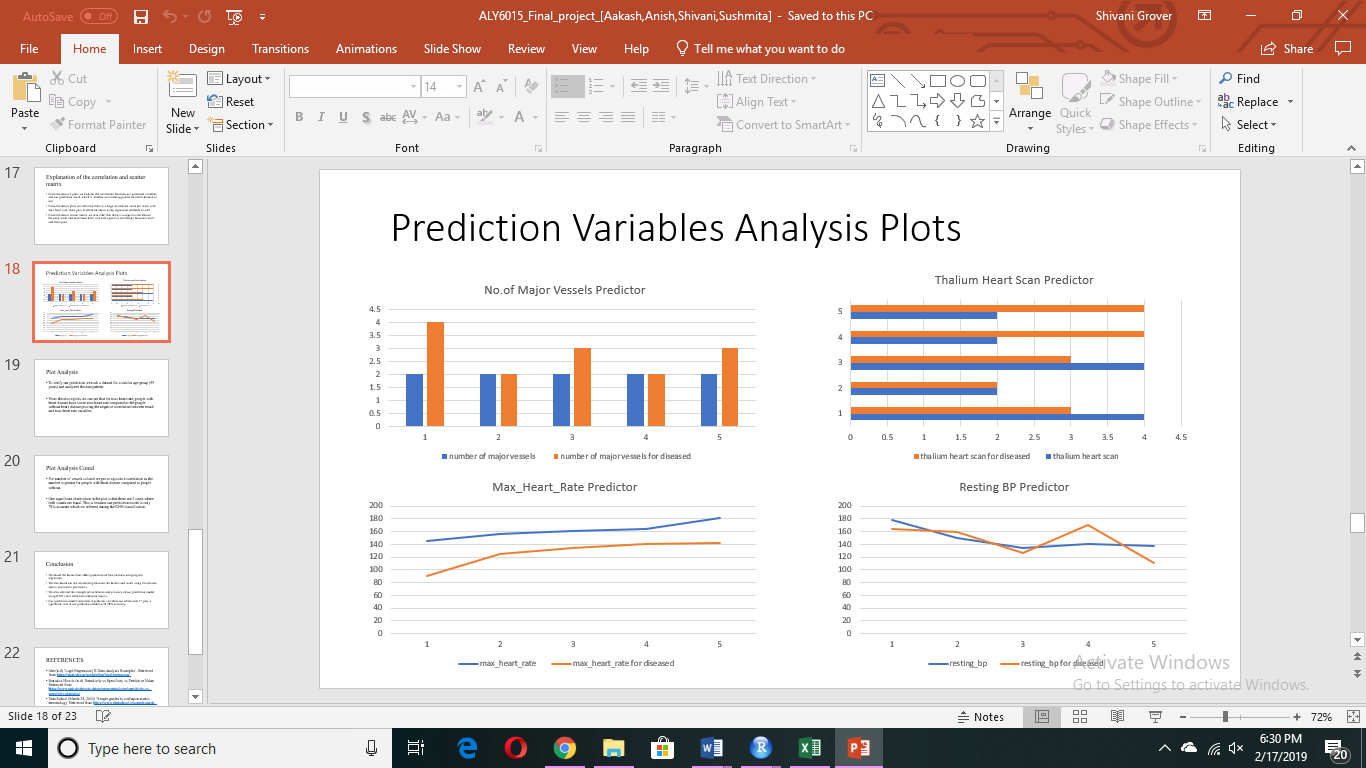
From the above scatter matrix we also infer that there is a negative correlation between the result and maximum heart rate and a positive correlation between result and chest pain. This indicates that when chest pain increases the chances of the person having heart disease also increase.

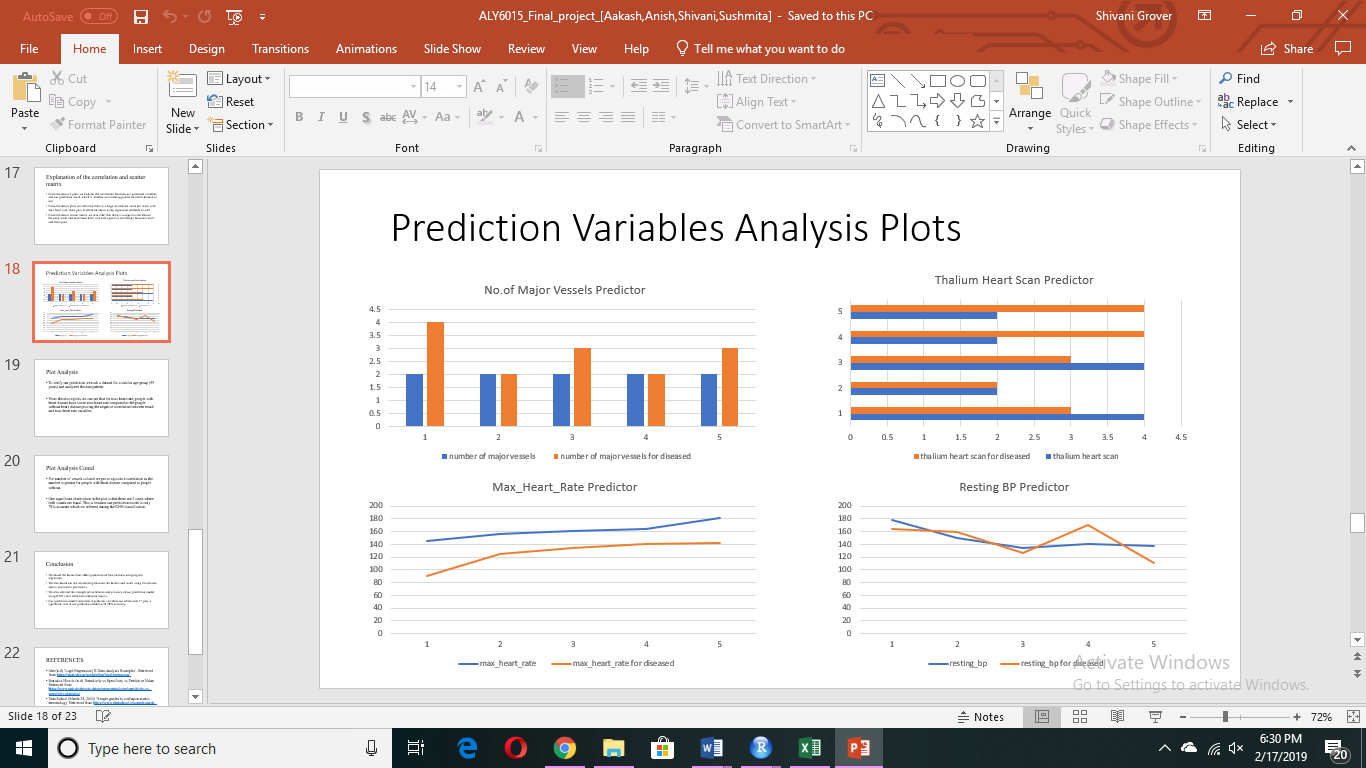


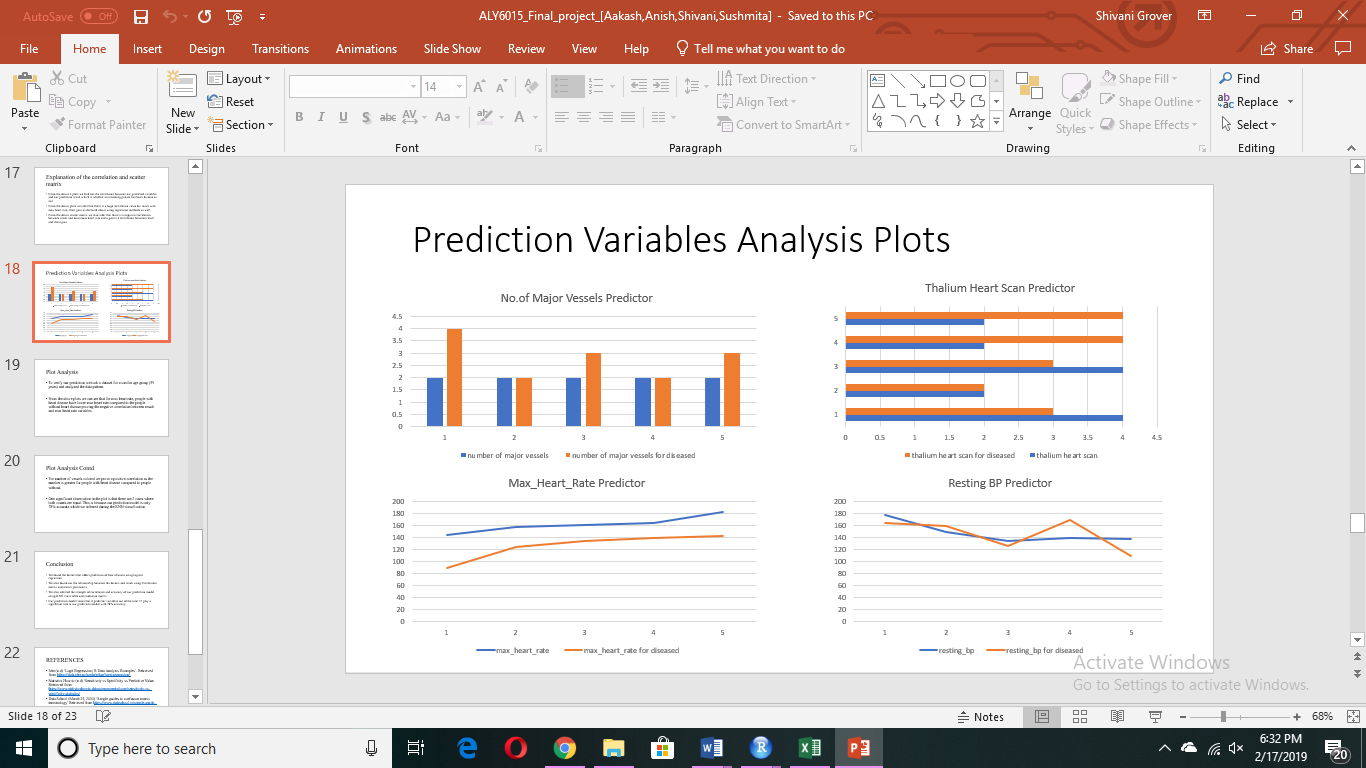


* From the above 2 plots, we find out the correlation between our predicted variables and our prediction result, which is whether an incoming patient has heart disease or not.
* From the above plots, we infer that there is a huge correlation value for the result with max heart rate, chest pain as declared above using regression methods as well.
* The above two matrices also show that there are some variables which are completely independent of each other, such as HR and cholesterol; FBS and result and angina with FBS.
* Using the above matrix, it can be concluded that the variables are not highly correlated. Thus, making the results of regression acceptable.

**Prediction Variables Analysis Plots**







**Interpretation of the above plots**

* To verify our prediction, we took a dataset for a similar age group (59 years) and analyzed the data pattern.
* From the above plots we can see that for max heart rate, people with heart disease have lower max heart rate compared to the people without heart disease proving the negative correlation between result and max heart rate variables.
* For a number of vessels colored we prove a positive correlation as the number is greater for people with heart disease compared to people without.
* One significant observation in the plot is that there are 2 cases where both counts are equal. This is because our prediction model is only 78% accurate which we inferred during the KNN classification.

**CONCLUSION**

1. Using Logistic Regression, we determine that chest pain, resting blood pressure, a number of major vessels, thallium heart scan are the factors that are significant for prediction of heart disease in females. While for males, the significant factors for prediction of heart disease are chest pain, the number of major vessels, resting blood pressure, maximum heart rate, exercise-induced pain, and thallium heart scan.
2. From the correlation plots we infer that there is a huge correlation value for the result with max heart rate and chest pain which could be verified using logistic regression method as well.
3. For max heart rate, people with heart disease have lower max heart rate compared to the people without heart disease proving the negative correlation between result and max heart rate variables.
4. For a number of vessels colored, we prove a positive correlation exists with the result, as the number is greater for people with heart disease compared to people who do not have heart disease.
5. One significant observation in the plot is that there are 2 cases where both counts are equal. This is because our prediction model is only 78% accurate which we inferred during the KNN classification.
6. Our prediction model states that 4 predictor variables out of the total 13 play a significant role in our prediction model with 78% accuracy.

Therefore, for any new patients if we have their chest pain type, thallium scan output, a number of vessels colored and resting blood pressure values then we could predict whether the person would have heart ailments or not.

**REFERENCES**

**Dataset reference link:** <http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/heart.dat>

1. Mayo Clinic (n.d) ‘Heart Disease’. Retrieved from <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>
2. Idre (n.d) ‘Logit Regression | R Data Analysis Examples’. Retrieved from <https://stats.idre.ucla.edu/r/dae/logit-regression/>
3. Statistics How to (n.d) ‘Sensitivity vs Specificity vs Predictive Value. Retrieved from <https://www.statisticshowto.datasciencecentral.com/sensitivity-vs-specificity-statistics/>
4. Data School (March 25, 2014) ‘Simple guides to confusion matrix terminology’ Retrieved from <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>