**DSC 520 Final Project Template**

**Section 1 – Week 9 – Getting Started**

**The Heart Disease Dataset**

<https://www.kaggle.com/johnsmith88/heart-disease-dataset#heart.csv>

**Dataset**:

The Heart Disease Dataset is downloaded from the link given below. The data set is downloaded in csv file and analyzed using the R programming.

<https://www.kaggle.com/johnsmith88/heart-disease-dataset#heart.csv>

The data set have the following variables and their description

* age
* sex
* chest pain type (4 values) - cp
* resting blood pressure  - trestbps
* serum cholesterol in mg/dl - chol
* fasting blood sugar > 120 mg/dl -fbs
* resting electrocardiographic results (values 0,1,2) -restecg
* maximum heart rate achieved -thalach
* exercise induced angina - exang
* old peak = ST depression induced by exercise relative to rest  -oldpeak
* the slope of the peak exercise ST segment  - slope
* number of major vessels (0-3) coloured by fluoroscopy -ca
* thal: 0 = normal; 1 = fixed defect; 2 = reversable defect - thal

**Introduction**:

Heart diseases are the significant reason for human death rate. Right analysis and treatment at a beginning time will spare individuals from coronary illness and will diminish mortality rate because of heart issue. Since ten years different information mining  procedures have been utilized to encourage the forecast of heart diseases .when all is said in done forecast calculations for prepared with tremendous, known dataset to show up at a classifier which at that point predicts the diseases for obscure information with the assistance of ordering traits. These characteristics additionally called as highlights. In this work significant highlights are resolved for coronary illness expectation with known dataset utilizing relationship measures. The outcomes are introduced

**Research Question :**

The main idea of this project to model for predicting the person who is already diagnosed heart disease by using the pattern from the 14 variable descriptive data. The task comprises of two stages. Stage I centers around information preprocessing and investigation, as shrouded in this report. The model structure, approval and expectation are introduced in Phase II.

The remainder of this report is composed as follow. Segment 2 portrays the informational collections and their properties. segment 3 spreads information pre-handling. In segment 4, we investigate each characteristic and their between connections. In the last we summaries the result.

There may be many attributes related to a given prediction problem. But not all the attributes have strong association with the prediction. Hence finding the relevant attributes for a given prediction problem is important. In this work, relevant attributes for heart disease prediction are determined using correlation measure. As mentioned above, from link it is found that the 13 attributes (thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg, trestbps, chol, fbs) are beingused while predicting heart diseases. In order to find the weight or rank of these attributes an experiment has been conducted. In this experiment the correlation between each attribute and class label is found out.

**Packages** :

library(knitr)                      library(readr)

library(dplyr)                      library(ggplot2)

library(mlr)                        library(cowplot)

library(tidyverse)              library(corrplot)

library(qgraph)                 library(jtools)

library(caret)                     library(DataExplorer)

library(funModeling)      library(ggm)

**Section 2 – Week 10 – Cleaning Your Data and Exploratory Data Analysis**

* Data importing and cleaning steps are explained in the text and in the DataCamp exercises (tell me why you are doing the data cleaning activities that you perform) and follow a logical process.

**I am cleaning the data in order to remove all major errors and inconsistencies that are inevitable when multiple sources of data are getting pulled into one dataset. Cleaning up the data will make me more efficient because I will be able analyze the data to quickly get what is needed from the data. Data deduplication is key to efficient and accurate analysis on data. It entails getting rid of copies and silo-ed variants of the same data, so you only have one golden copy or as few copies as possible.**

**I will also be identifying and dropping duplicates in the data, since this will help me save time when analyzing data.**

* With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.

**> tbl\_df(heart)**

**# A tibble: 1,025 x 14**

**age sex cp trestbps chol**

***<int>* *<int>* *<int>* *<int>* *<int>***

**1 52 1 0 125 212**

**2 53 1 0 140 203**

**3 70 1 0 145 174**

**4 61 1 0 148 203**

**5 62 0 0 138 294**

**6 58 0 0 100 248**

**7 58 1 0 114 318**

**8 55 1 0 160 289**

**9 46 1 0 120 249**

**10 54 1 0 122 286**

**# … with 1,015 more rows, and 9 more**

**# variables: fbs *<int>*,**

**# restecg *<int>*, thalach *<int>*,**

**# exang *<int>*, oldpeak *<dbl>*,**

**# slope *<int>*, ca *<int>*, thal *<int>*,**

**# target *<int>***

**> str(heart)**

**'data.frame': 1025 obs. of 14 variables:**

**$ age : int 52 53 70 61 62 58 58 55 46 54 ...**

**$ sex : int 1 1 1 1 0 0 1 1 1 1 ...**

**$ cp : int 0 0 0 0 0 0 0 0 0 0 ...**

**$ trestbps: int 125 140 145 148 138 100 114 160 120 122 ...**

**$ chol : int 212 203 174 203 294 248 318 289 249 286 ...**

**$ fbs : int 0 1 0 0 1 0 0 0 0 0 ...**

**$ restecg : int 1 0 1 1 1 0 2 0 0 0 ...**

**$ thalach : int 168 155 125 161 106 122 140 145 144 116 ...**

**$ exang : int 0 1 1 0 0 0 0 1 0 1 ...**

**$ oldpeak : num 1 3.1 2.6 0 1.9 1 4.4 0.8 0.8 3.2 ...**

**$ slope : int 2 0 0 2 1 1 0 1 2 1 ...**

**$ ca : int 2 0 0 1 3 0 3 1 0 2 ...**

**$ thal : int 3 3 3 3 2 2 1 3 3 2 ...**

**$ target : int 0 0 0 0 0 1 0 0 0 0 ...**

**There are 14 variables and 1025 observations and 13 variables are integers and 1 variable is numeric**

* What do you not know how to do right now that you need to learn to import and cleanup your dataset?

**There may be many attributes related to a given prediction problem. But not all the attributes have strong association with the prediction. As of now I don’t have know all the relevant attributes for a given prediction. I will have to find all those attributes.**

**Section 3 – Week 11 – Starting Your Writeups**

* Discuss how you plan to uncover new information in the data that is not self-evident.

**I plan to uncover new information in the data by using histograms and other graphs**

* What are different ways you could look at this data to answer the questions you want to answer?

**I will see if the dependent variables are associated with each other and how all the dependent variables are associated with independent variables. Also we see the distribution behavior of the data if it is left skewed, right skewed, data having outliers.**

corr <- cor(heart)

corrplot(corr,method = "number",type="lower",order = "hclust",title = "Correlations between Variables ")

> corr

age sex cp trestbps chol

age 1.00000000 -0.10324030 -0.07196627 0.27112141 0.21982253

sex -0.10324030 1.00000000 -0.04111909 -0.07897377 -0.19825787

cp -0.07196627 -0.04111909 1.00000000 0.03817742 -0.08164102

trestbps 0.27112141 -0.07897377 0.03817742 1.00000000 0.12797743

chol 0.21982253 -0.19825787 -0.08164102 0.12797743 1.00000000

fbs 0.12124348 0.02720046 0.07929359 0.18176662 0.02691716

restecg -0.13269617 -0.05511721 0.04358061 -0.12379409 -0.14741024

thalach -0.39022708 -0.04936524 0.30683928 -0.03926407 -0.02177209

exang 0.08816338 0.13915681 -0.40151271 0.06119697 0.06738223

oldpeak 0.20813668 0.08468656 -0.17473348 0.18743411 0.06488031

slope -0.16910511 -0.02666629 0.13163278 -0.12044531 -0.01424787

ca 0.27155053 0.11172891 -0.17620647 0.10455372 0.07425934

thal 0.07229745 0.19842425 -0.16334148 0.05927618 0.10024418

target -0.22932355 -0.27950076 0.43485425 -0.13877173 -0.09996559

fbs restecg thalach exang oldpeak

age 0.121243479 -0.13269617 -0.390227075 0.08816338 0.20813668

sex 0.027200461 -0.05511721 -0.049365243 0.13915681 0.08468656

cp 0.079293586 0.04358061 0.306839282 -0.40151271 -0.17473348

trestbps 0.181766624 -0.12379409 -0.039264069 0.06119697 0.18743411

chol 0.026917164 -0.14741024 -0.021772091 0.06738223 0.06488031

fbs 1.000000000 -0.10405124 -0.008865857 0.04926057 0.01085948

restecg -0.104051244 1.00000000 0.048410637 -0.06560553 -0.05011425

thalach -0.008865857 0.04841064 1.000000000 -0.38028087 -0.34979616

exang 0.049260570 -0.06560553 -0.380280872 1.00000000 0.31084376

oldpeak 0.010859481 -0.05011425 -0.349796163 0.31084376 1.00000000

slope -0.061902374 0.08608609 0.395307843 -0.26733547 -0.57518854

ca 0.137156259 -0.07807235 -0.207888416 0.10784854 0.22181603

thal -0.042177320 -0.02050406 -0.098068165 0.19720104 0.20267203

target -0.041163547 0.13446821 0.422895496 -0.43802855 -0.43844127

slope ca thal target

age -0.16910511 0.27155053 0.07229745 -0.22932355

sex -0.02666629 0.11172891 0.19842425 -0.27950076

cp 0.13163278 -0.17620647 -0.16334148 0.43485425

trestbps -0.12044531 0.10455372 0.05927618 -0.13877173

chol -0.01424787 0.07425934 0.10024418 -0.09996559

fbs -0.06190237 0.13715626 -0.04217732 -0.04116355

restecg 0.08608609 -0.07807235 -0.02050406 0.13446821

thalach 0.39530784 -0.20788842 -0.09806817 0.42289550

exang -0.26733547 0.10784854 0.19720104 -0.43802855

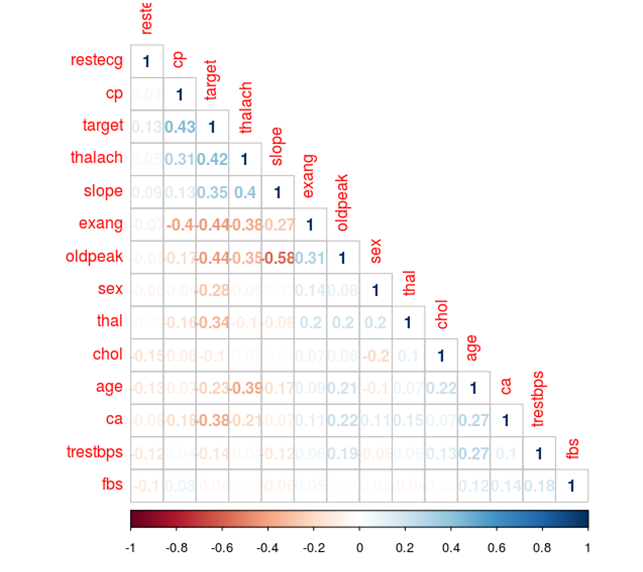
oldpeak -0.57518854 0.22181603 0.20267203 -0.43844127

slope 1.00000000 -0.07344041 -0.09409006 0.34551175

ca -0.07344041 1.00000000 0.14901387 -0.38208529

thal -0.09409006 0.14901387 1.00000000 -0.33783815

target 0.34551175 -0.38208529 -0.33783815 1.00000000



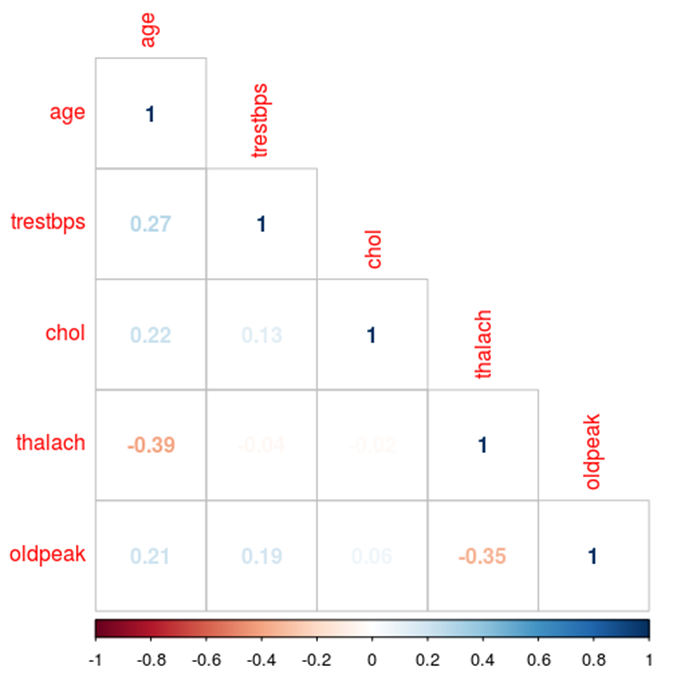
Oldpeak has the highest correlation coefficient of 0.58 which indicating that these variables have good association with each other. So we can say that depression induced by exercise relative to rest is associated with the slope of the peak exercise ST segment - slope

* Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

**In this data I am not slicing the data and creating the new variable like by using the log or square root transformation.**

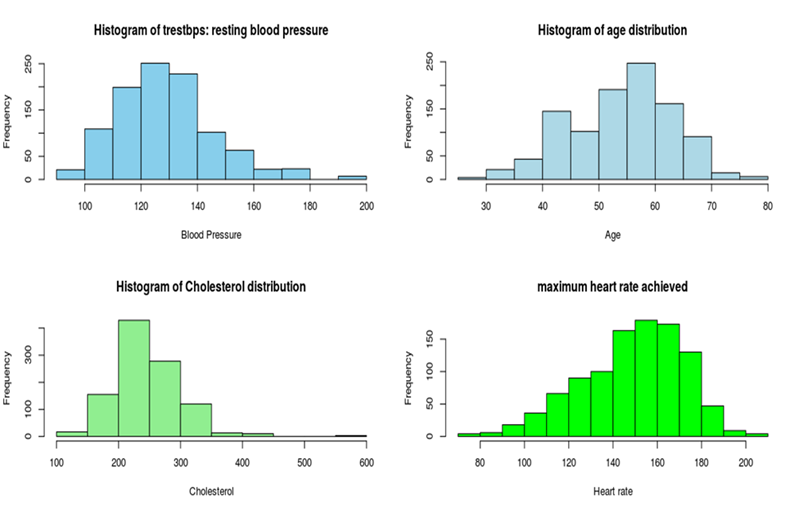
* **How could you summarize your data to answer key questions?**

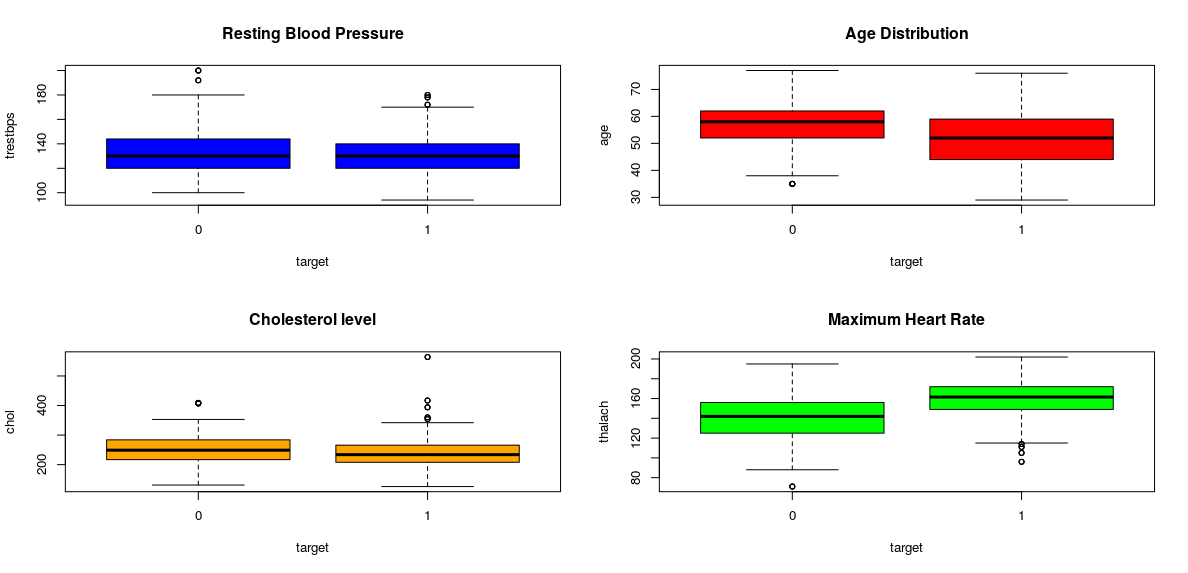
I have summarize the data as below :

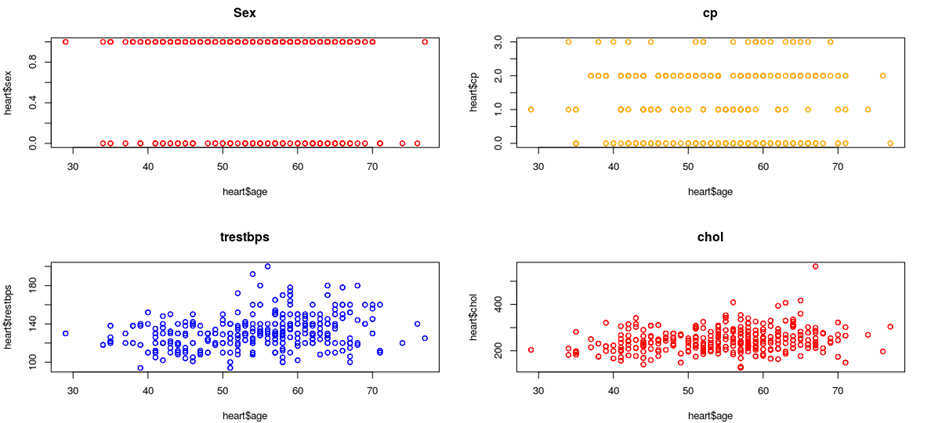


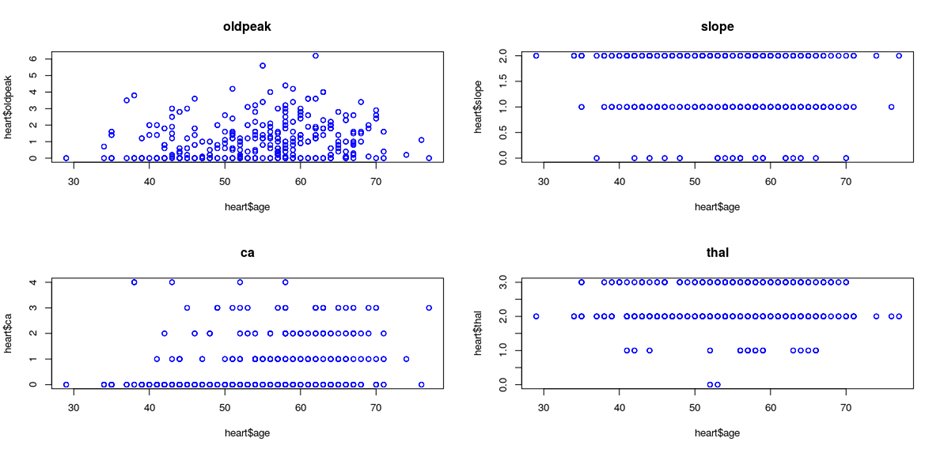
* What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

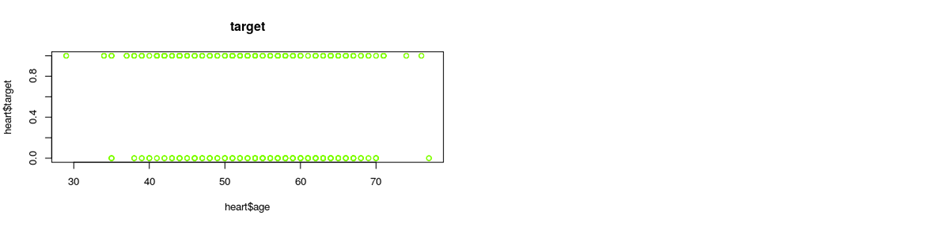
**Histogram, Boxplot, Scatter plot, Correlation Matrix plot as shown below :**

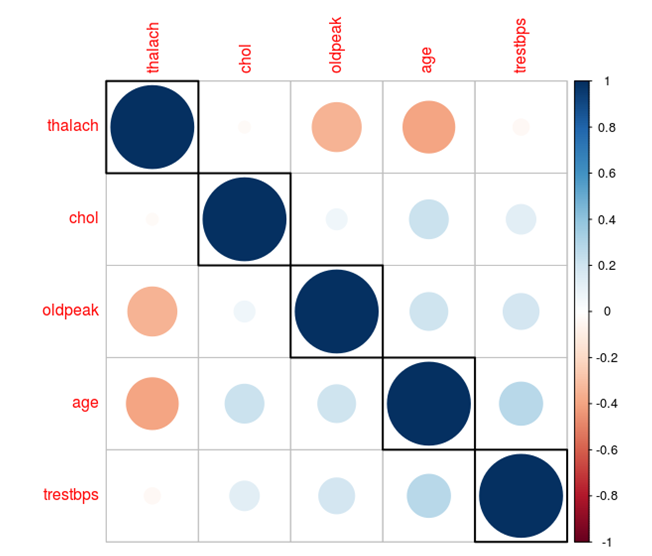
****

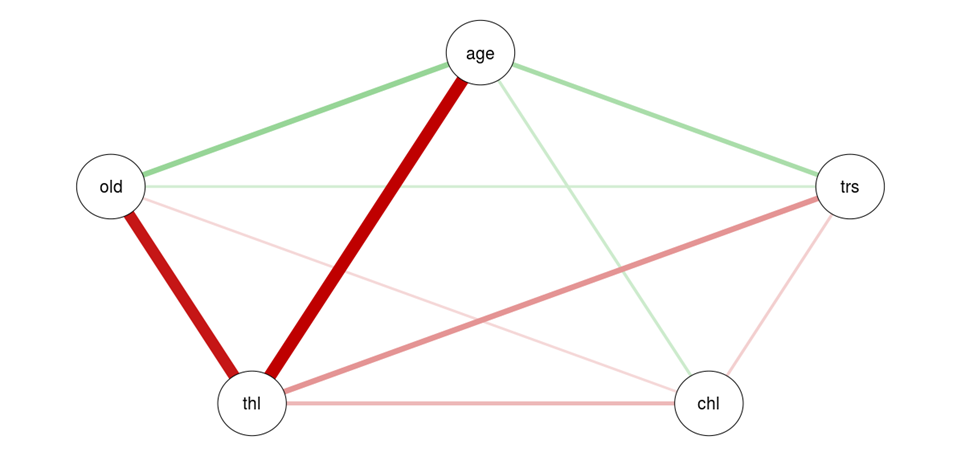


****

****

****

****

****

* What do you not know how to do right now that you need to learn to answer your questions?

**Right now I know most about the data. Their association with each other. The magnitude od association.**

* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

**Correlation and regression analysis technique on incorporating the machine learning technique.**

**Section 4 – Week 12**

* Overall, write a coherent narrative that tells a story with the data as you complete this section.

**Age, Resting Blood Pressure (TRESTBPS),Serum Cholesterol (CHOL),Maximum Heart Rate Achieved (THALACH) and ST Depression Induced by Exercise Relative to Rest (OLDPEAK) There’s no significant linear correlation between any two of features. Blood pressure, cholesterol level and ST depression shows a weak positive linear correlation with age while maximum heart rate achieved shows a weak negative correlation with age. Maximum heart rate achieved also shows a weak negative linear relationship with ST depression induced by exercise relative to rest.**

**The Histogram of resting blood pressure is right skewed, showing that few patients have had an extremely higher blood pressure. When comparing Histograms separately for levels of target feature, we can see patients having a heat disease showed a higher blood pressure compared to patients not having a heart disease.**

**Patients from age 29 years to 77 years were included in this data set. The Histogram of age is little skewed to the left, showing the average age is little lower than the median age. Around 50% of patients’ age was in between 45-65 years. There is no visualized difference in ages for patients with or without heart disease. That is age may not be a major factor to diagnose a hear disease.**

**The distribution of patients’serum cholesterol level is highly right skewed, showing that few patients have had extremely high cholesterol levels. When we compare this distribution separately for patients with a heart disease and patients without a heart disease, the healthy patients’ distribution is leptokurtic. That means, there were many healthy peaple who had there cholesterol level around 200-220 mg/fl than patients with a hear disease.**

**The histogram for maximum heart rate achieved by patients is left skewed as few patients showed a comparatively low heart rate. The separate Histograms for two levels of target feature show healthy people have had a quite higher maximum heart rate (around 160) compared to the maximum heart rate (150) of patients with a heart disease. Further, the Histogram for health people is leptokurtic (There are many people around the peak point).**

* Summarize the problem statement you addressed.

**Heart diseases are the significant reason for human death rate. Right analysis and treatment at a beginning time will spare individuals from coronary illness and will diminish mortality rate because of heart issue. Since ten years different information mining procedures have been utilized to encourage the forecast of heart diseases .when all is said in done forecast calculations for prepared with tremendous, known dataset to show up at a classifier which at that point predicts the diseases for obscure information with the assistance of ordering traits. These characteristics additionally called as highlights. In this work significant highlights are resolved for coronary illness expectation with known dataset utilizing relationship measures. The outcomes are introduced**

* Summarize how you addressed this problem statement (the data used and the methodology employed).

**The main idea of this project to model for predicting the person who is already diagnosed heart disease by using the pattern from the 14 variable descriptive data. The task comprises of two stages. Stage I centers around information preprocessing and investigation, as shrouded in this report. The model structure, approval and expectation are introduced in Phase II. The remainder of this report is composed as follow. Segment 2 portrays the informational collections and their properties. segment 3 spreads information pre-handling. In segment 4, we investigate each characteristic and their between connections. In the last we summaries the result.**

* Summarize the interesting insights that your analysis provided.

**There may be many attributes related to a given prediction problem. But not all the attributes have strong association with the prediction. Hence finding the relevant attributes for a given prediction problem is important. In this work, relevant attributes for heart disease prediction are determined using correlation measure. As mentioned above, from link it is found that the 13 attributes (thal, ca, exang, oldpeak, thalach, cp, slope, sex, age, restecg, trestbps, chol, fbs) are beingused while predicting heart diseases. In order to find the weight or rank of these attributes an experiment has been conducted. In this experiment the correlation between each attribute and class label is found out.**

* Summarize the implications to the consumer (target audience) of your analysis.

**Since the method is probabilistic, so there will always be problem of some risk of not getting the same result. To the problem of heart disease there could be so many variables that can influence heart related proble. All those data may not be part of the analysis. So the heart problems with those factors could be the implication in detecting.**

* Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

**Choose a relevant dataset that is representative of the whole population. Choose the right parameters for your analysis. Even after all this care and attention, don’t be surprised if your data still needs preprocessing before you can analyze it accurately. Preprocessing often takes a long time and significant effort because it has to address several issues related to the original data — these issues include:**

* **Any values missing from the data.**
* **Any inconsistencies and/or errors existing in the data.**
* **Any duplicates or outliers in the data.**
* **Any normalization or other transformation of the data.**
* **Any derived data needed for the analysis.**