HEART DISEASE

OPRE6359

Abstract

Heart Disease Statistics Using R

Amitesh Enugala

HEART DISEASE

Submitted by
Amitesh Enugala

Under the guidance of
Prof. Monica Brussolo

The University of Texas at Dallas
Richardson, TX 75080

May 2019

TABLE OF CONTENTS

INDEX

ACKNOWLEDGEMENT

My project on **Heart Disease** has been a great learning experience. I was exposed to a vast subject matter, concerns and arguments that helped me collectively assemble and shape the project.

I acknowledge Professor Monica Brussolo under whose guidance I were able to complete the project and effectively present its valuable benefits.

A greater share of inputs and knowledge through classes and assignments made this project report possible to its rightful accuracy.

To all our colleagues who have helped us either directly or indirectly, we are grateful for their valuable inputs.

LITERATURE

The goal of Heart Disease analytics is to provide best model for Healthcare Institutions with insights for analyzing and interpreting the risk of heart disease for individuals who are diagnosed, so that decisions for further test on heart disease can be reached quickly and efficiently. The challenge of Heart Disease analytics is to identify what data should be captured and how to use the data to model and predict capabilities, so the organization gets an optimal return on investment on its human capital.

Providing best Medical solution for an individual is a major stake for any Healthcare organization. But are there any reliable ways to figure out if and why the individuals are affected with Heart diseases? Most firms these days are already integrating the benefits of using analytics to introduce special efforts in predicting Heart condition and any diseases it possesses. Lot of factors play key role in identifying significant predictors in estimating the heart effects and meaning that can be interpreted using a statistical model language like R.

In our project, I have used Heart Disease Analytics dataset from UCI Machine Learning Repository which were reported by Hungarian Institute of Cardiology, University Hospital Switzerland, Long Beach and Cleveland Clinic foundation

BACKGROUND

Data set

Our data set represents 627 records and is composed of individuals who were diagnosed for Heart Diseases who have a heart disease, and some doesn't. It has 15 variables defining the best possible way to answer the below questions and insights.

Initially after loading the dataset, I have observed 323 records that had no significance for any of our analysis model and hence I decided to discard them. It is always recommended to run some basic checks and see if there are missing values or any unusual patterns amongst other things. Right from the very first correlation that I ran, I was clear about incorporating few changes to the dataset. I have consolidated the datasets of 4 regions namely Cleveland, Hungary, Switzerland, long Beach V.A and included a field called region.

OBJECTIVES

The main objectives that we had set out before working on the dataset were:

J. Is Maximum Heart rate achieved same in people with Heart Disease and no Heart Disease?

K. Is Maximum Heart rate achieved same in different Genders?

L. Are cholesterol Levels same in people with Heart Disease and no Heart Disease?

M. Are Cholesterol levels same in different Genders?

N. Is oldpeak (ST depression induced by exercise) same in people with Heart Disease and no Heart Disease?

O. Is Maximum Heart rate achieved same in all regions?

P. Maximum Heart rate achieved by an Individual

Q. When an individual is Diagnosed, does he have a Heart Disease?

R. Is Exercise Induced Angina (Exang) having any relation with Chest Pain(cp)?

DATA EXPLORATION

Read the Dataset

```
# reading dataset #
heart.df <- read.csv("C:/Users/amite/Downloads/stats project/processed.cleveland.csv",header=TRUE)
```

Dataset Details

```
> dim(heart.df)
[1] 303  15
```

Describe Dataset

```
> summary(heart.df)
      age             sex              cp           trestbps         chol            fbs              restecg          talach
 Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0   Min.   :126.0   Min.   :0.0000   Min.   :0.0000   Min.   : 71.0
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0   1st Qu.:202.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:133.5
 Median :56.00   Median :1.0000   Median :3.000   Median :130.0   Median :229.0   Median :0.0000   Median :1.0000   Median :153.0
 Mean   :54.44   Mean   :0.6799   Mean   :3.158   Mean   :131.7   Mean   :232.2   Mean   :0.1485   Mean   :0.9901   Mean   :149.6
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0   3rd Qu.:261.0   3rd Qu.:0.0000   3rd Qu.:2.0000   3rd Qu.:166.0
 Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0   Max.   :409.0   Max.   :1.0000   Max.   :2.0000   Max.   :202.0
     exang           oldpeak          slope             ca              thal            num              region
 Min.   :0.0000   Min.   :0.00   Min.   :1.000   Min.   :0.0000   Min.   :3.000   Min.   :0.0000   c:64
 1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:0.0000   h:90
 Median :0.0000   Median :0.80   Median :2.000   Median :0.0000   Median :3.000   Median :0.0000   s:73
 Mean   :0.3267   Mean   :1.04   Mean   :1.601   Mean   :0.6832   Mean   :4.736   Mean   :0.4653   v:76
 3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:7.000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :6.20   Max.   :3.000   Max.   :3.0000   Max.   :7.000   Max.   :1.0000
```

Meta Data

Attribute	Description
age	age in years
sex	Gender (1 = male; 0 = female)
cp	Chest Pain Type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
trestbps	resting blood pressure
chol	serum cholesterol in mg/dl
fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality (elevation > 0.05 mV) 2 = showing probable or definite left ventricular hypertrophy)
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by fluoroscopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect diagnosis of heart disease (angiographic disease status)

num (the predicted attribute	(0: < 50% less probability of heart disease 1: > 50% high probability of heart disease)
Region	(1 = Cleveland, 2 = Long Beach V.A, 3 = Switzerland, 4 = Hungary)

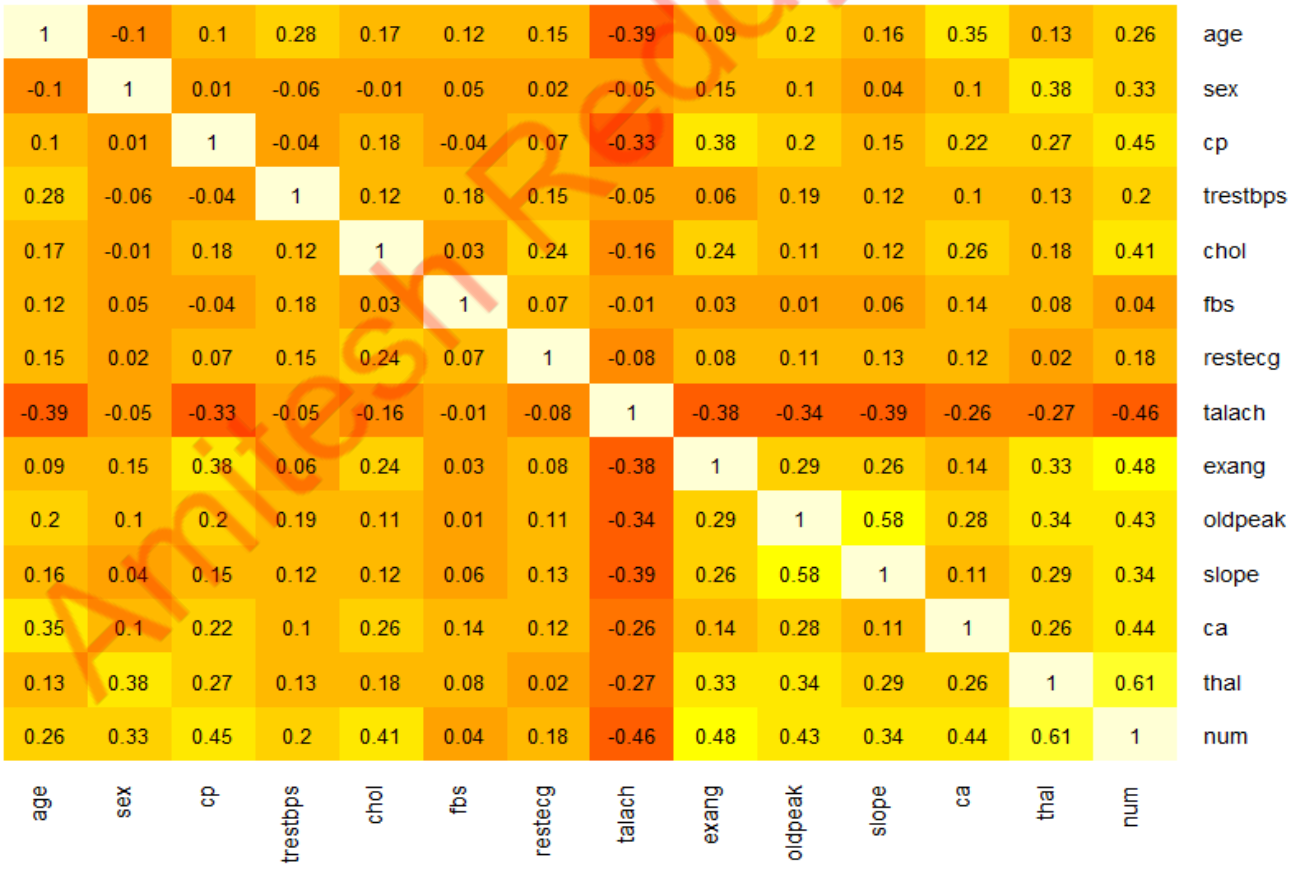
Correlation and changes to the dataset

To improve the correlation significance between various predictors, I made changes against few variable records. (*talach, cp, region*)

I have included a new variable "Region" by consolidating all the individual regions and naming each region by their Starting letter as indicated in Metadata. (1 = Cleveland, 2 = Long Beach V.A, 3 = Switzerland, 4 = Hungary). Removed Region for Correlation Matrix – Heatmaps

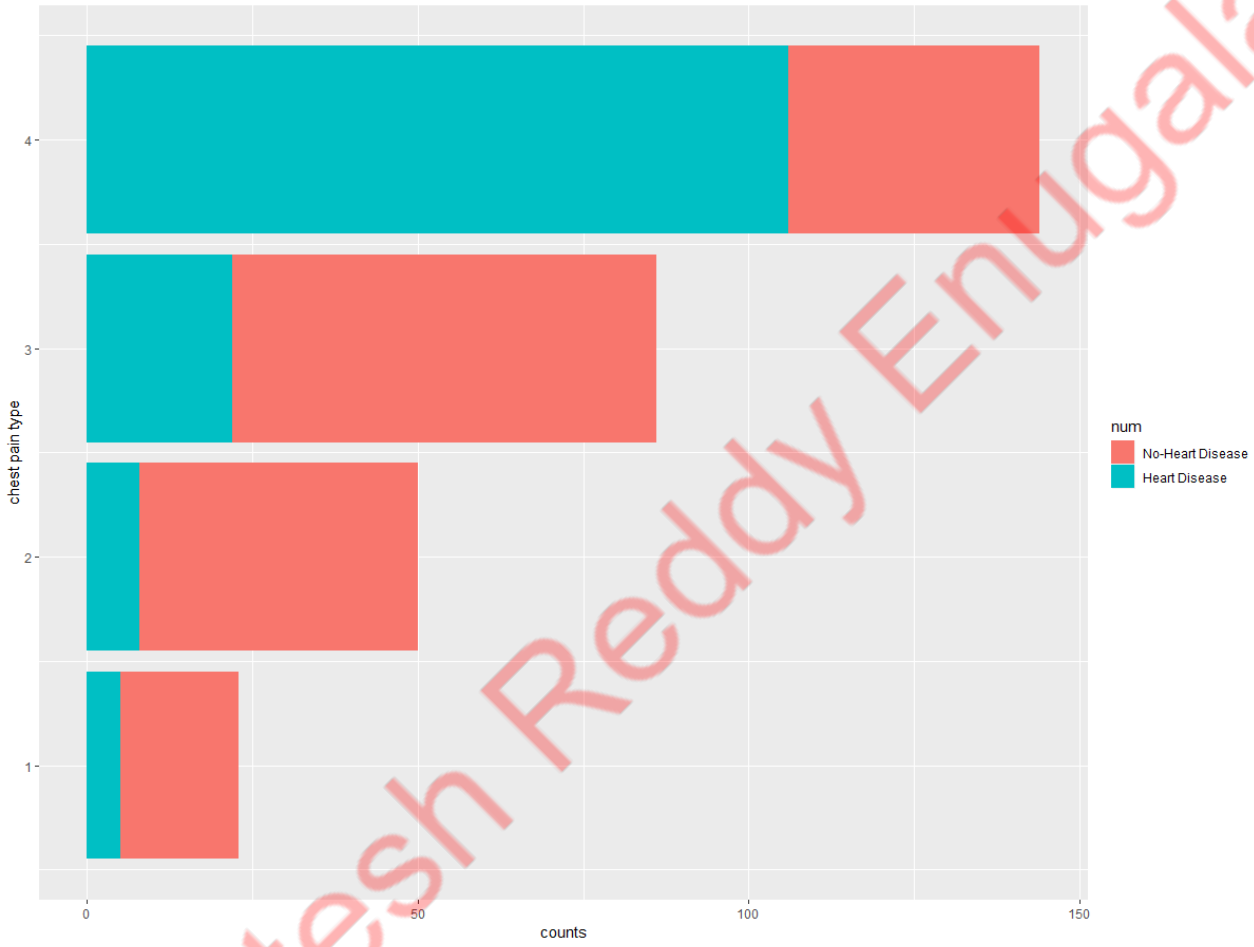
Significant correlation exists amongst majority of the variables

```
#Correlation matrix using Heat-Maps
heart1.df <- heart.df[,c(-15)]
heatmap.2(cor(heart1.df),Rowv = FALSE, Colv=FALSE,dendogram="none",cellnote=round(cor(heart1.df),2),
          notecol="black",key=FALSE,trace="none",margins=c(10,10))
```



Barplot to ascertain people who diagnosed with different chest pains using GGPLOT

```
heart1.df$num <- factor(heart1.df$num, levels = c(0,1), labels = c("No-Heart Disease","Heart Disease"))
#Barplot to ascertain people who diagnosed with different chest pains
ggplot(aes(x = cp), data = heart1.df) +
  geom_bar(aes(fill = num)) +
  xlab("chest pain type") +
  ylab("counts") +
  coord_flip()
```

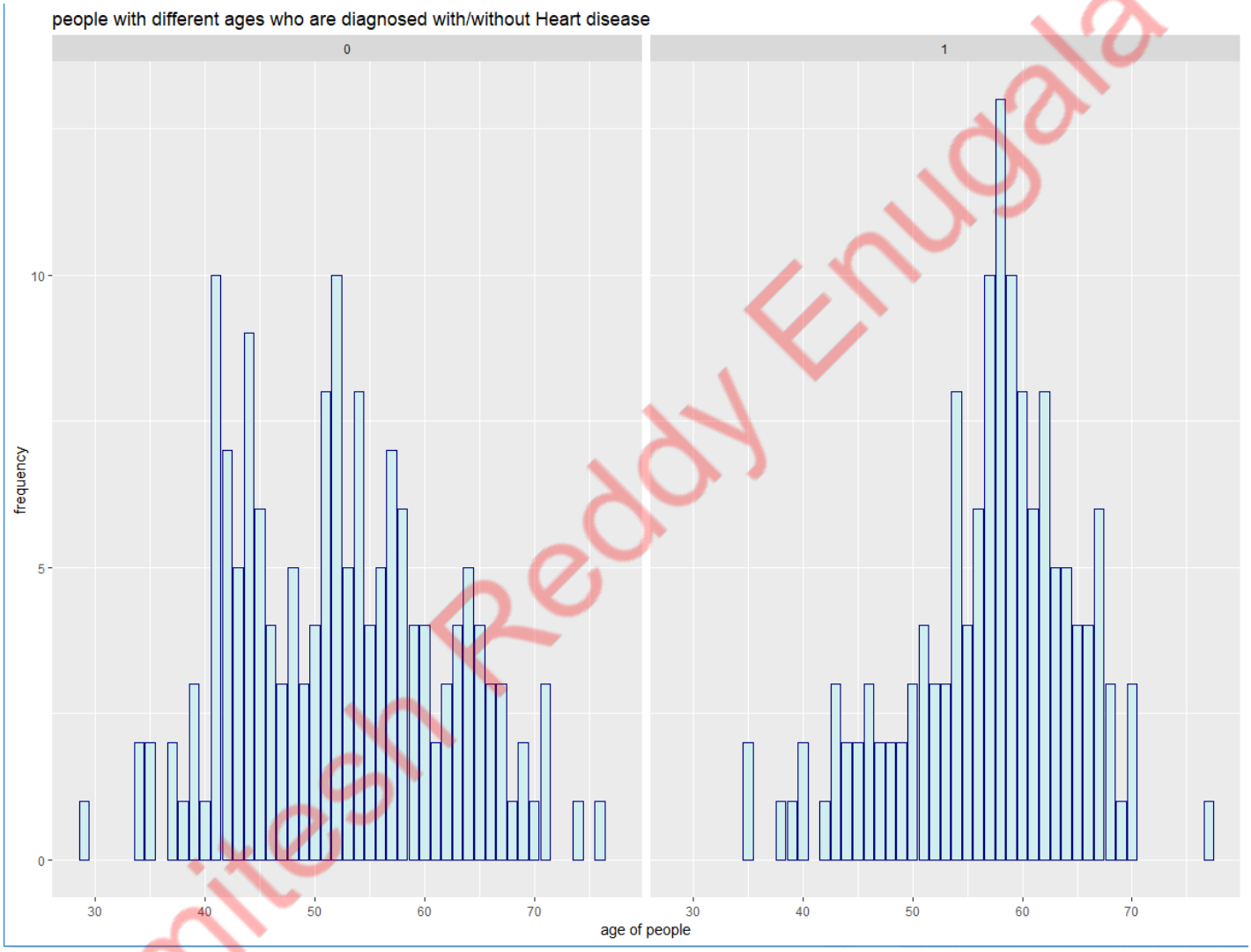


Interpretation

- Highest cases were diagnosed for Chest pain type = 4 (asymptomatic) and majority of cases resulted in heart disease
- For Chest pain type = 3(non-anginal pain) Majority of cases were not resulted in Heart disease
- For Chest pain type = 2(atypical angina) it has highest probability of not having a heart disease
- For Chest Pain type = 1(typical angina) only fewer cases are recorded

Barplot of people with different ages who are diagnosed with/without Heart disease

```
#Barplot of people with different ages who are diagnosed with/without Heart disease #
ggplot(aes(x=heart.df$age),data=heart.df) +
  geom_bar(fill = 'lightcyan2', color='navy') +
  xlab("age of people") +
  ylab("frequency") +
labs(title = "people with different ages who are diagnosed with/without Heart disease ") +
  facet_wrap(~num)
```



Interpretation

- There is a high frequency of having Heart Disease around the age 60
- Frequency of having heart disease in people with low age are pretty low

MODEL ANALYSIS

After running descriptive diagnostics on the dataset, lets move on to predictive analytics. In this section we aim to answer the questions that will help the Healthcare institutions to mitigate the heart disease effect on individuals. This analysis is important in the sense that it assists Doctors or other medical leaders to analyze the factors that shows a cause and effect on heart and to take proactive actions.

T-tests:

The central idea of using T-Tests in our project is to use to determine whether there is a significant difference between the means of two groups. With all inferential statistics, and the dependent variable fits a normal distribution.

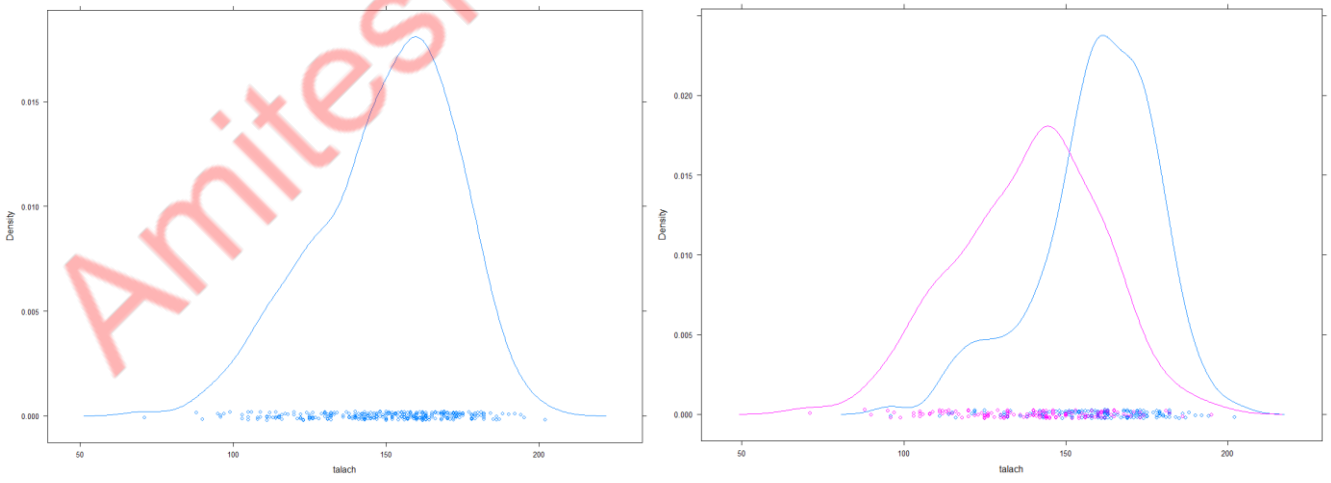
Problem Statement 1:

Is Maximum Heart rate achieved same in people with Heart Disease and no Heart Disease?

Code:

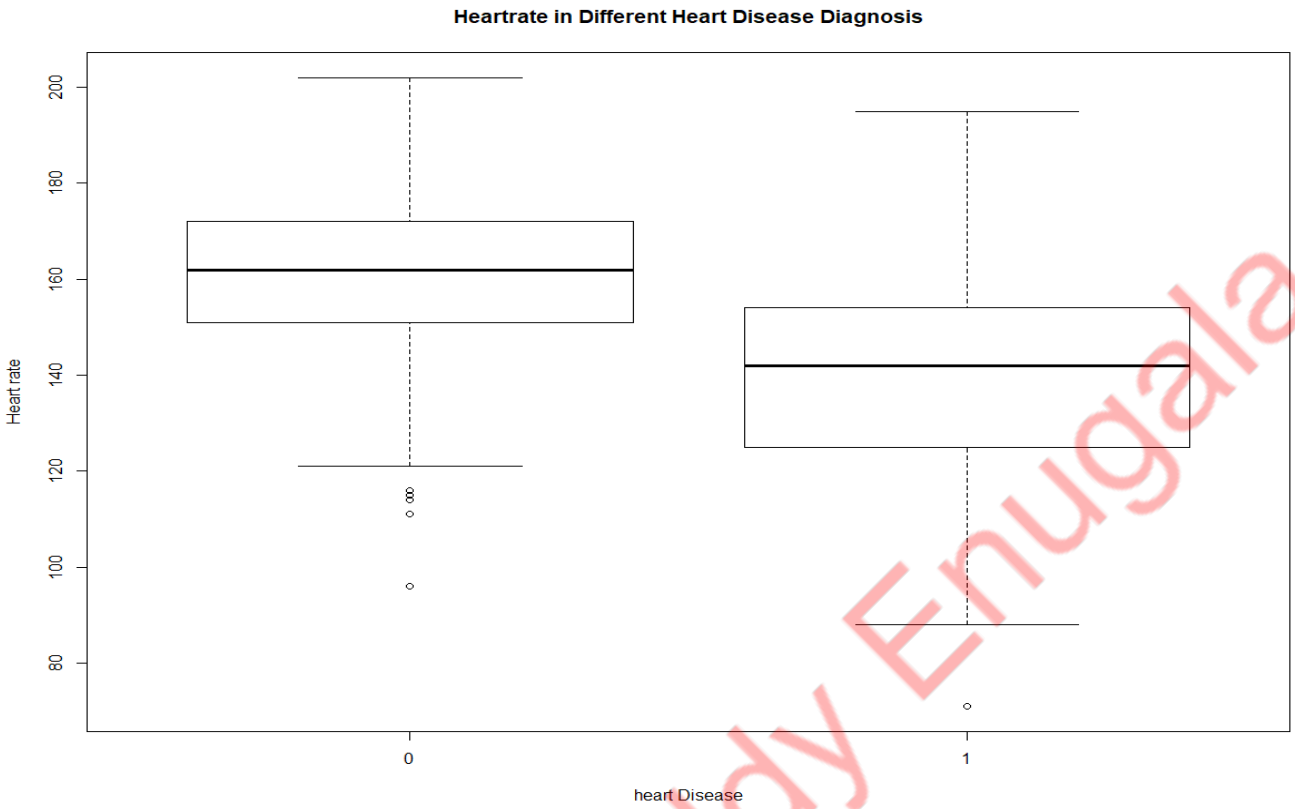
```
# density plot to show heart rate is normally distributed #
densityplot(~talach, auto.key = TRUE, data = heart.df)
densityplot(~talach, groups = num, auto.key = TRUE, data = heart.df)
# box plot to show heart rate distributed between people with Heart Disease and No- Heart Disease#
boxplot(talach~num,data=heart.df, main="Heartrate in Different Heart Disease Diagnosis",
        xlab="heart Disease", ylab="Heart rate")
# t-test to show significence of Maximum Heart rate achieved for people with heart Disease and no heart Disease #
t.test(talach~num, var.equal=FALSE, data=heart.df)
t.test(log(talach)~num, var.equal=TRUE, data=heart.df)
```

Check for normality: Distribution of Maximum heart rate achieved(talach)



These Graphs shows us that Maximum Heart rate achieved is most likely to be normal fig1 and

Normal when taken between people with Heart Disease and No-Heart Disease fig 2



This Clearly indicates that there is a difference hence we can proceed with T-Test for further analysis

Test Statistics:

```
> t.test(talach~num, var.equal=FALSE, data=heart.df)

        Welch Two Sample t-test

data:  talach by num
t = 8.7852, df = 273.88, p-value < 0.00000000000000022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 16.20680 25.56809
sample estimates:
mean in group 0 mean in group 1
       159.3272        138.4397

> t.test(log(talach)~num, var.equal=TRUE, data=heart.df)

        Two Sample t-test

data:  log(talach) by num
t = 8.6457, df = 301, p-value = 0.0000000000000003232
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1134785 0.1803596
sample estimates:
mean in group 0 mean in group 1
       5.063587        4.916667
```

- From the above t-tests we can say that we can reject null hypothesis and conclude that there is a difference in Maximum Heart rate achieved in people with Heart Disease and No-Heart Disease.
- p-value < 0.00000000000000022 shows a strong indication to reject Null Hypothesis.
- People with no Heart Disease have 15% more Maximum Heart rate achieved than people with heart Disease with a CI of 12% to 20% (taken by making log transformations)

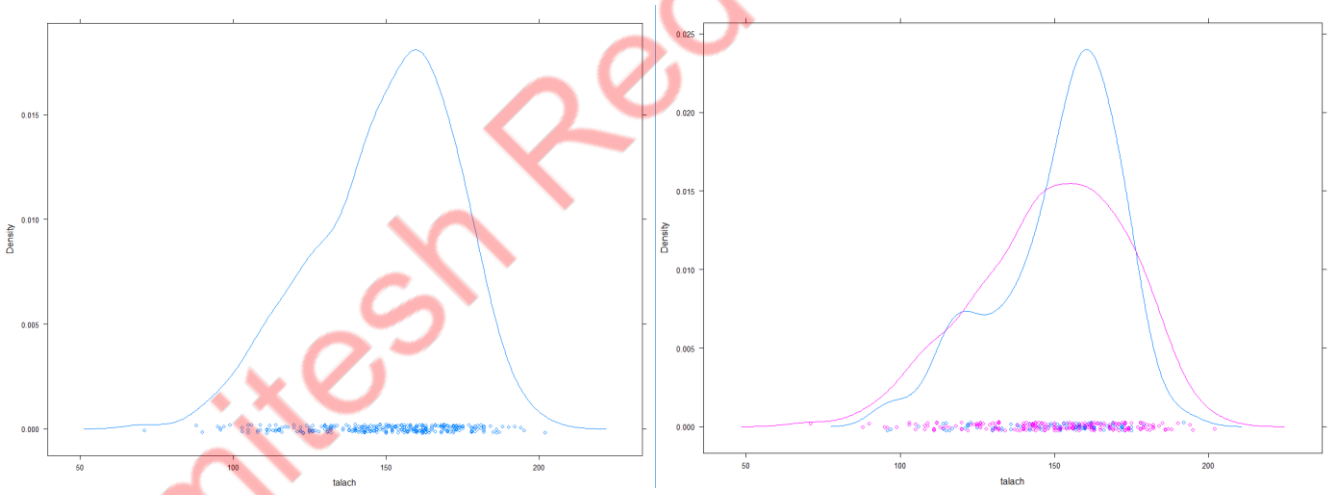
Problem Statement 2:

Is Maximum Heart rate achieved same in different Genders?

Code:

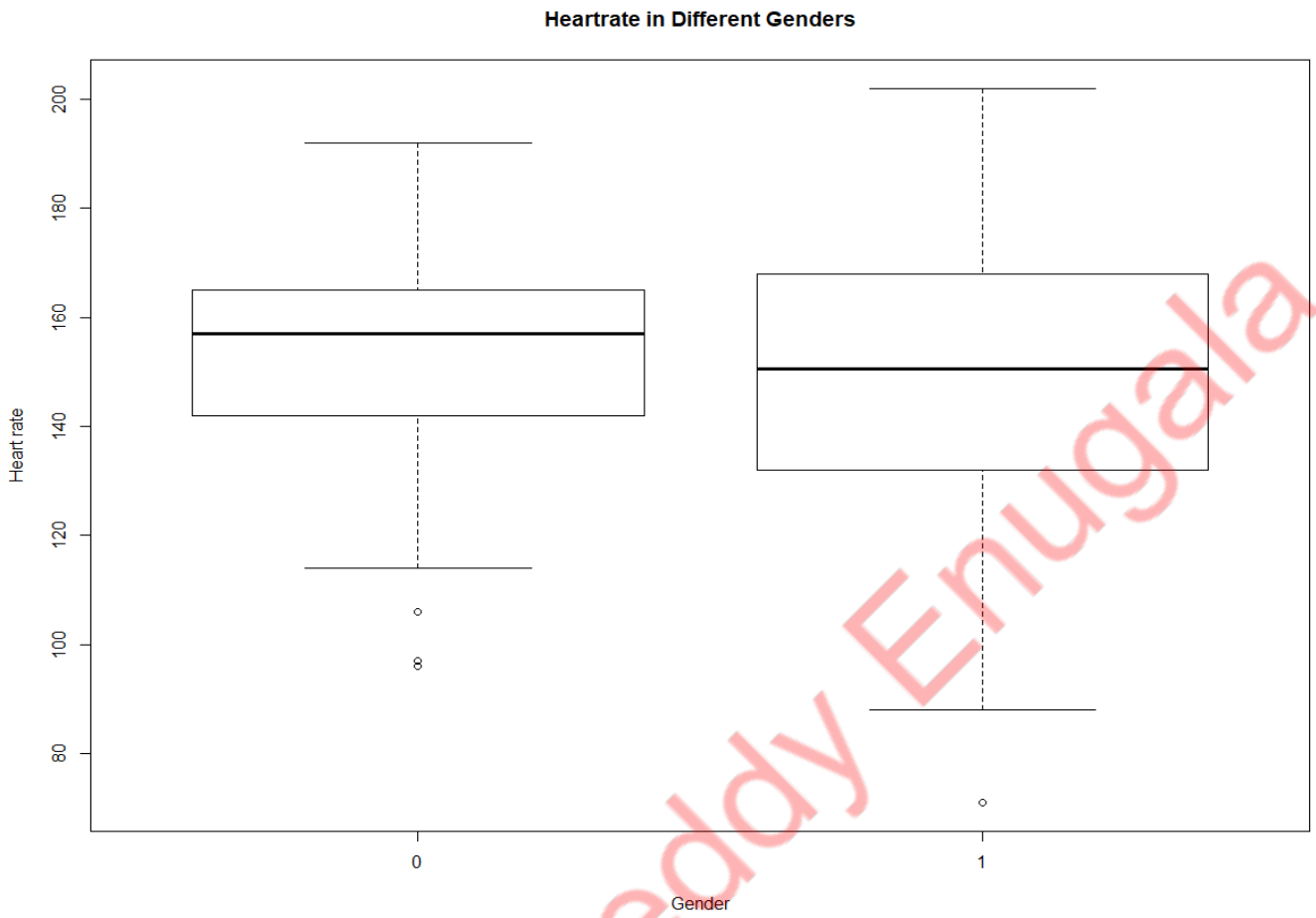
```
# density plot to show heart rate is normally distributed #
densityplot(~talach, auto.key = TRUE, data = heart.df)
densityplot(~talach, groups = sex, auto.key = TRUE, data = heart.df)
# box plot to show heart rate distribution in different Genders#
boxplot(talach~sex,data=heart.df, main="Heartrate in Different Genders",
        xlab="Gender", ylab="Heart rate")
# t-test to show significence of Maximum Heart rate achieved for different genders#
t.test(talach~sex, var.equal=FALSE, data=heart.df)
t.test(log(talach)~sex, var.equal=FALSE, data=heart.df)
```

Check for normality: Distribution of Maximum heart rate achieved(talach)



These Graphs shows us that Maximum Heart rate achieved is most likely to be normal fig1 and

Normal when taken between Male and Female (purple – male, blue-Female) fig 2



This Clearly indicates that there is not much difference hence we can proceed with T-Test for further analysis

Test Statistics:

```
> t.test(talach~sex, var.equal=FALSE, data=heart.df)

        Welch Two Sample t-test

data:  talach by sex
t = 0.90442, df = 223.85, p-value = 0.3667
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.808276  7.572564
sample estimates:
mean in group 0 mean in group 1
       151.2268        148.8447

> t.test(log(talach)~sex, var.equal=FALSE, data=heart.df)

        Welch Two Sample t-test

data:  log(talach) by sex
t = 1.0988, df = 226.79, p-value = 0.273
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01642649  0.05783761
sample estimates:
mean in group 0 mean in group 1
       5.009295        4.988590
```

- From the above t-tests we can say that we cannot reject null hypothesis and conclude that there is no much difference in Maximum Heart rate achieved in different genders.
- p-value = 0.3667 shows a strong indication not to reject Null Hypothesis.
- We found out that both Genders have same Maximum Heart Rate achieved with CI of -2.8 to 7.57

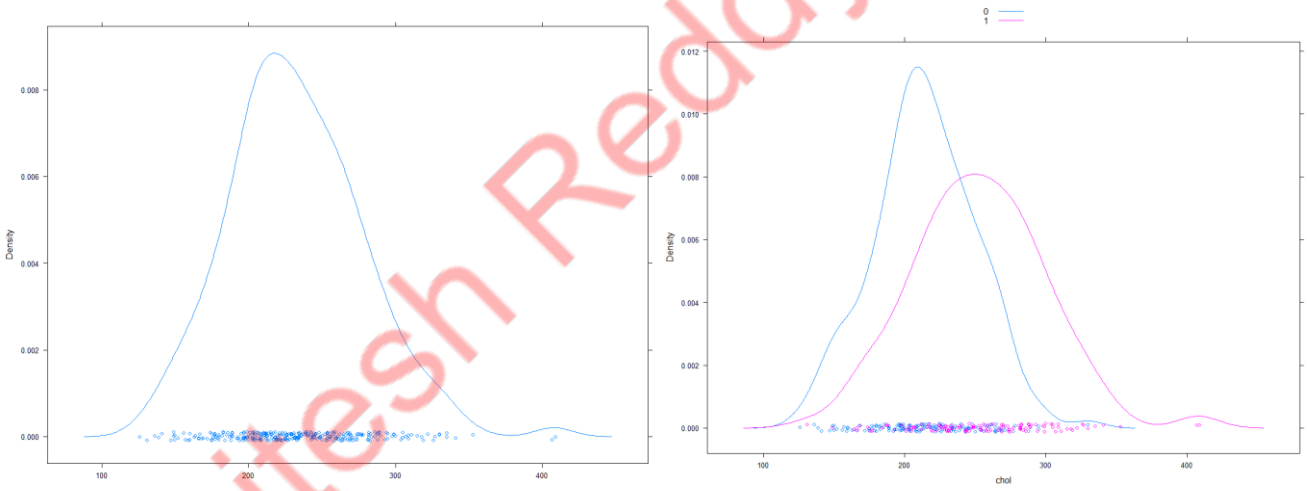
Problem Statement 3:

Are cholesterol Levels same in people with Heart Disease and no Heart Disease?

Code:

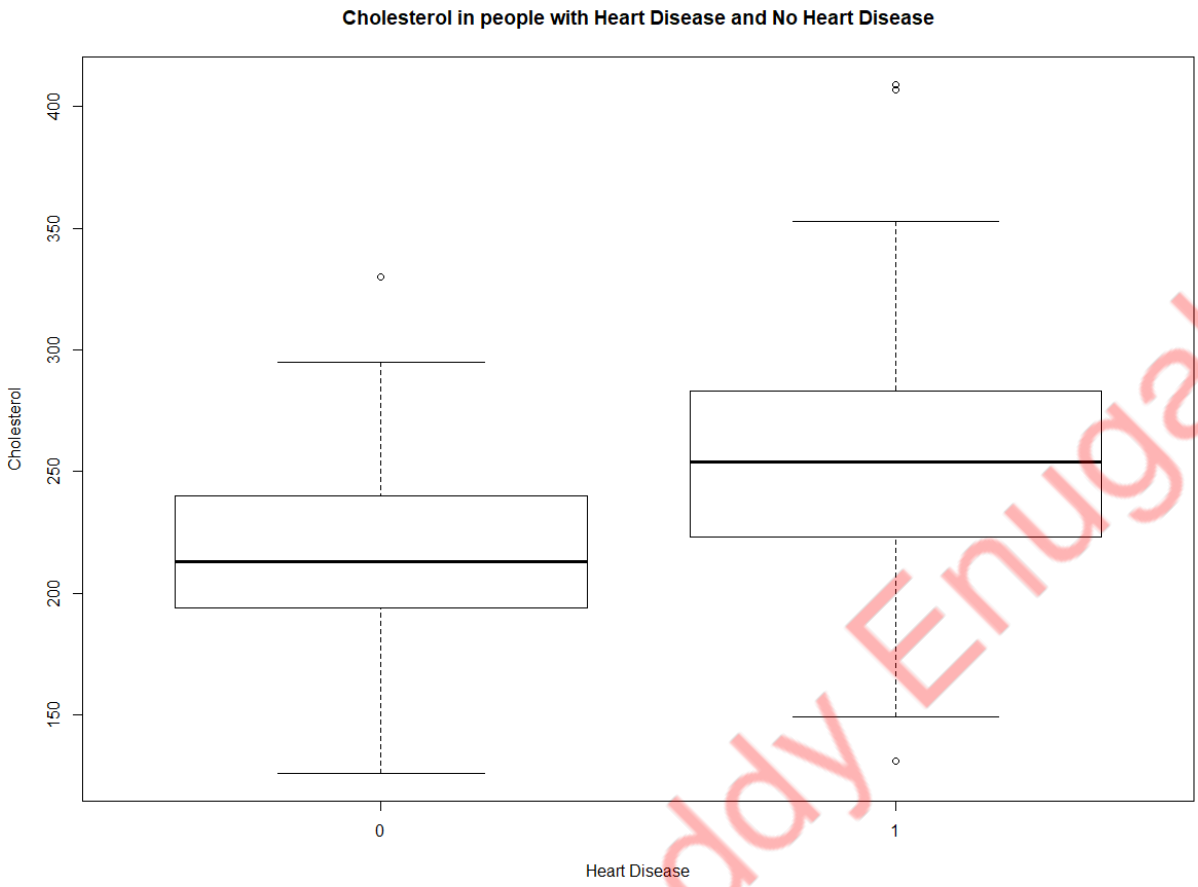
```
# density plot to show cholesterol is normally distributed #
densityplot(~chol, auto.key = TRUE, data = heart.df)
densityplot(~chol, groups = num, auto.key = TRUE, data = heart.df)
# box plot to show Cholesterol distribution in people with Heart Disease and NO Heart Disease#
boxplot(chol~num,data=heart.df, main="Cholesterol in people with Heart Disease and No Heart Disease",
        xlab="Heart Disease", ylab="Cholesterol")
# t-test to show cholestrol is different in different for people with heart attack and no heart attack #
t.test(chol~num, var.equal=FALSE, data=heart.df)
t.test(log(chol)~num, var.equal=TRUE, data=heart.df)
```

Check for normality: Distribution of Cholesterol(chol)



These Graphs shows us that Cholesterol is most likely to be normal fig1 and

Normal when taken between people with Heart Disease(purple) and No-Heart Disease(blue) fig 2



This Clearly indicates that there is a significant difference hence we can proceed with T-Test for further analysis

Test Statistics:

```
> t.test(chol~num, var.equal=FALSE, data=heart.df)

        Welch Two Sample t-test

data:  chol by num
t = -7.7571, df = 258.11, p-value = 0.000000000000202
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -47.81614 -28.45441
sample estimates:
mean in group 0 mean in group 1
       214.4321        252.5674

> t.test(log(chol)~num, var.equal=TRUE, data=heart.df)

        Two Sample t-test

data:  log(chol) by num
t = -7.6027, df = 301, p-value = 0.000000000000372
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2014931 -0.1186317
sample estimates:
mean in group 0 mean in group 1
       5.353469        5.513531
```

- From the above t-tests we can say that we can reject null hypothesis and conclude that there is difference in Cholesterol values for people with Heart Disease and No Heart Disease
- p-value = 0.000000000000202 shows a strong indication to reject Null Hypothesis.
- Cholesterol levels is 17% less in people without Heart Disease than people with Heart Disease with of CI 12% to 19%

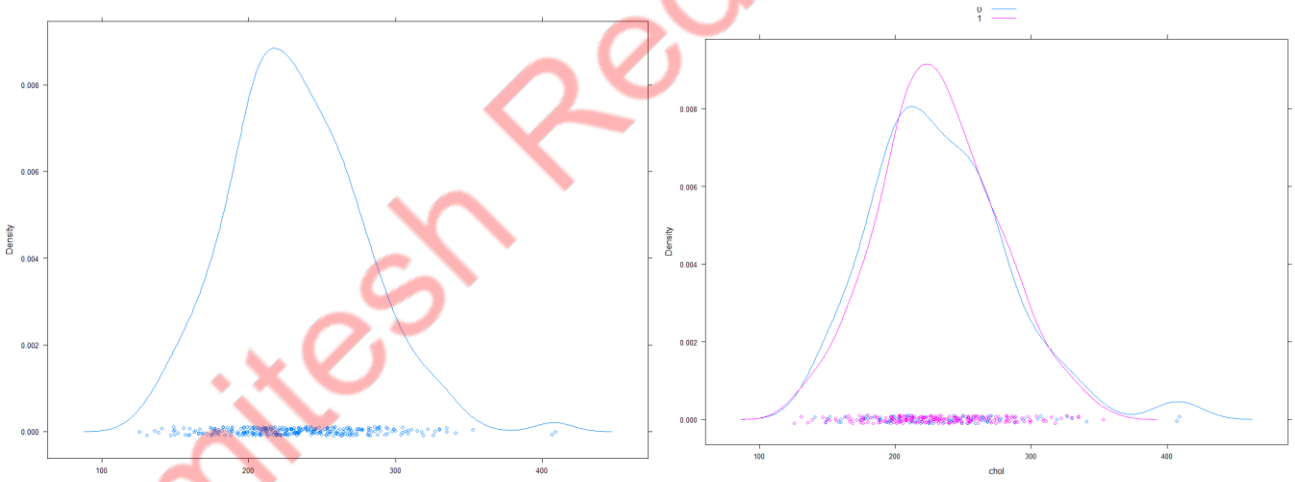
Problem Statement 4:

Are Cholesterol levels same in different Genders?

Code:

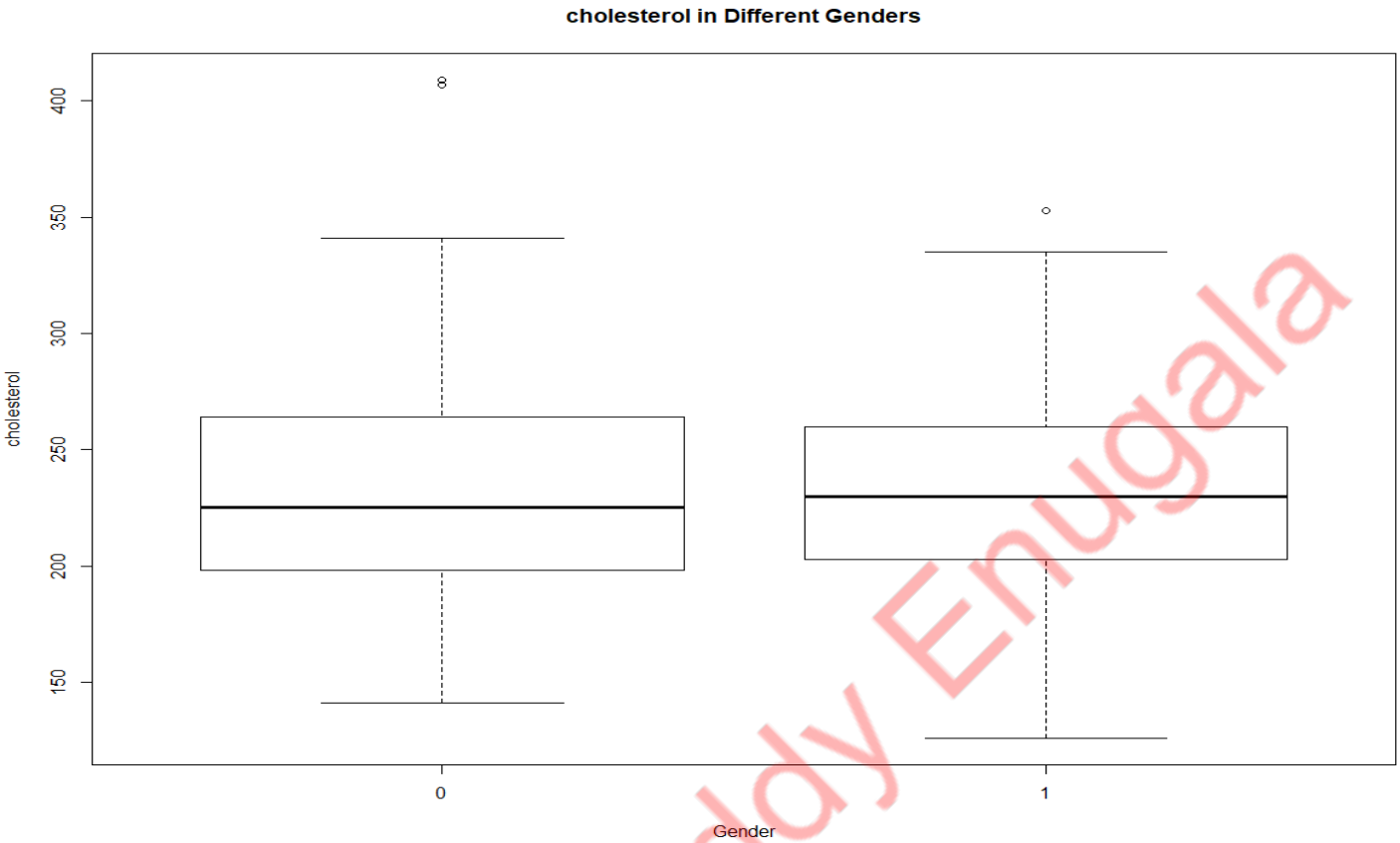
```
# density plot to show cholesterol is normally distributed  #
densityplot(~chol, auto.key = TRUE, data = heart.df)
densityplot(~chol, groups = sex, auto.key = TRUE, data = heart.df)
# box plot to show cholesterol distribution in different Genders#
boxplot(chol~sex,data=heart.df, main="cholesterol in Different Genders",
        xlab="Gender", ylab="cholesterol")
# t-test to show cholestrol is different in different for gender #
t.test(chol~sex, var.equal=FALSE, data=heart.df)
t.test(log(chol)~sex, var.equal=TRUE, data=heart.df)
```

Check for normality: Distribution of Cholesterol(chol)



These Graphs shows us that Cholesterol is most likely to be normal fig1 and

Normal when taken between Male and Female (purple – male, blue-Female) fig 2



This Clearly indicates that there is not much difference hence we can proceed with T-Test for further analysis

Test Statistics:

```
> t.test(chol~sex, var.equal=FALSE, data=heart.df)

        Welch Two Sample t-test

data:  chol by sex
t = 0.24528, df = 165.3, p-value = 0.8065
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.33833  13.27137
sample estimates:
mean in group 0 mean in group 1
       233.1753        231.7087

> t.test(log(chol)~sex, var.equal=TRUE, data=heart.df)

        Two Sample t-test

data:  log(chol) by sex
t = 0.09254, df = 301, p-value = 0.9263
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.04608801  0.05063651
sample estimates:
mean in group 0 mean in group 1
       5.429499        5.427225
```

- From the above t-tests we can say that we cannot reject null hypothesis and conclude that there is not much difference in Cholesterol values for different Genders
- p-value = 0.8065 shows a strong indication not to reject Null Hypothesis.
- Cholesterol levels is 17% less in people without Heart Disease than people with Heart Disease with of CI Same levels with CI of -10.33833 to 13.27137

Wilcoxson Rank Test: Non-Parametric test of Significance

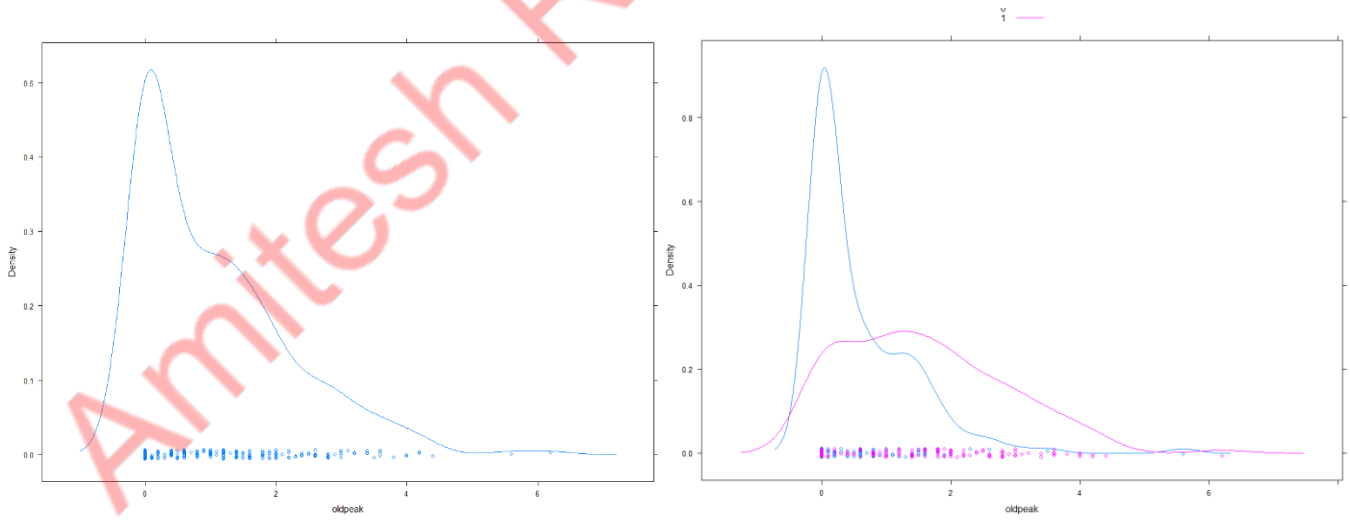
Problem Statement:

Is oldpeak (ST depression induced by exercise) same in people with Heart Disease and no Heart Disease?

Code:

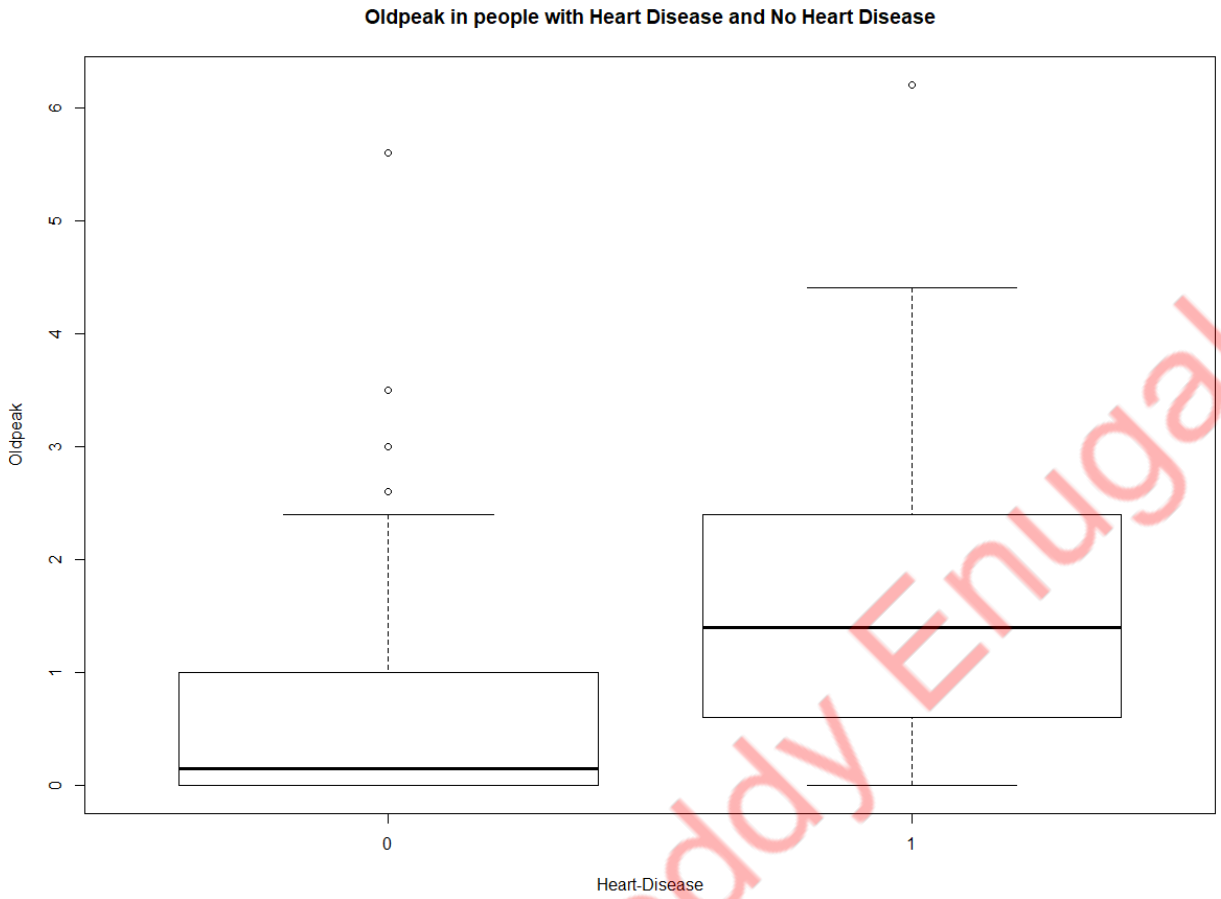
```
# density plot to show cholesterol is normally distributed  #
densityplot(~chol, auto.key = TRUE, data = heart.df)
densityplot(~chol, groups = sex, auto.key = TRUE, data = heart.df)
# box plot to show cholesterol distribution in different Genders#
boxplot(chol~sex,data=heart.df, main="cholesterol in Different Genders",
        xlab="Gender", ylab="cholesterol")
# t-test to show cholestrol is different in different for gender #
t.test(chol~sex, var.equal=FALSE, data=heart.df)
t.test(log(chol)~sex, var.equal=TRUE, data=heart.df)
```

Check for normality: Distribution of ST depression induced by exercise (oldpeak)



These Graphs shows us that oldpeak is not normal fig1 and

Not Normal when taken between people with Heart Disease(purple) and No-Heart Disease(blue) fig 2



This Clearly indicates that there is much difference hence we can proceed with Wilcoxson Rank Sum test for further analysis

Test Statistics:

```
> wilcox.test(rank ~ num, conf.int = TRUE, exact = TRUE, data = heart.df)

        Wilcoxon rank sum test with continuity correction

data:  rank by num
W = 5578, p-value = 0.000000000000005235
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -107.00005  -64.49997
sample estimates:
difference in location
             -85.00001
```

- From the above U-tests we can say that we can reject null hypothesis and conclude that there is difference in oldpeak values for people with Heart Disease and No Heart Disease
- p-value = 0.000000000000005235 shows a strong indication to reject Null Hypothesis.
- People with no Heart Disease has 15% less oldpeak than people with heart Disease with CI of 7% to 35%

ANOVA

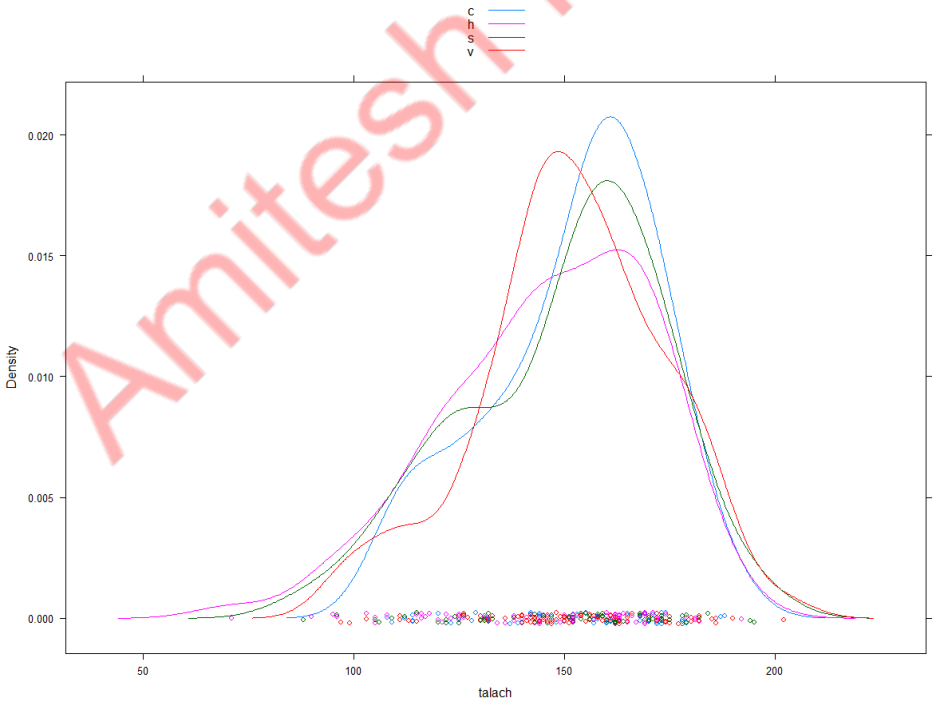
Problem Statement:

Is Maximum Heart rate achieved same in all regions?

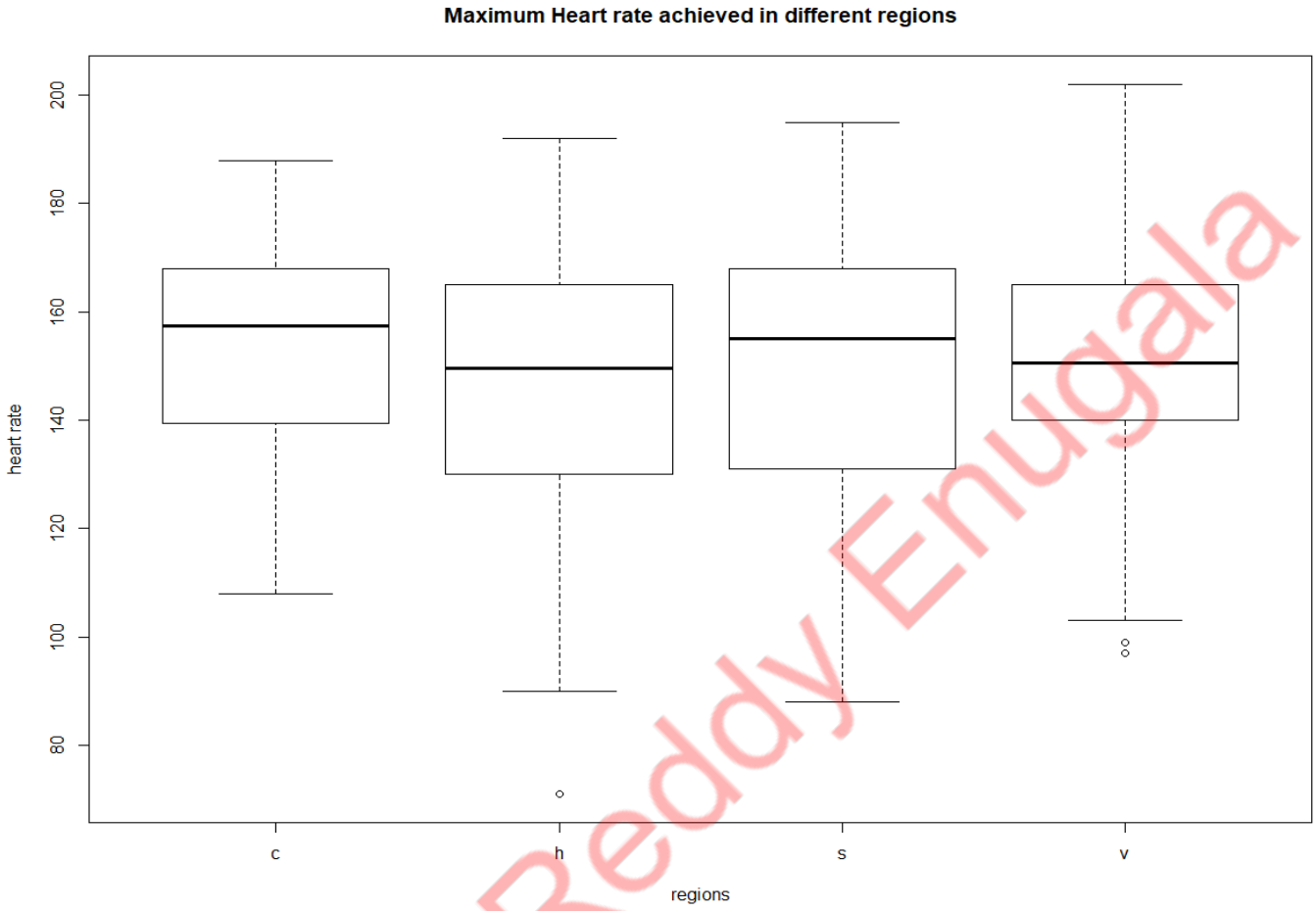
Code:

```
#Anova
# density plot of heart rate for different regions #
densityplot(~talach, groups = region, auto.key = TRUE, data = heart.df)
# box plot to show Maximum Heart rate Achieved in different regions#
boxplot(talach~region,data=heart.df, main="Maximum Heart rate achieved in different regions",
        xlab="regions", ylab="heart rate")
#anova for heart rate for different regions #
anova(lm(talach~region, data=heart.df))
# pairwise comparison of each region #
fit.contrast(lm(talach~region, data=heart.df), "region",c(-1,1,0, + 0),conf.int= .95)

fit.contrast(lm(talach~region, data=heart.df), "region",c(-1,0,1, + 0),conf.int= .95)

fit.contrast(lm(talach~region, data=heart.df), "region",c(-1,0,0, + 1),conf.int= .95)

fit.contrast(lm(talach~region, data=heart.df), "region",c(-0,-1,1, + 0),conf.int= .95)

fit.contrast(lm(talach~region, data=heart.df), "region",c(-0,-1,0, + 1),conf.int= .95)

fit.contrast(lm(talach~region, data=heart.df), "region",c(-0,0,-1, + 1),conf.int= .95)

#residuals vs fitted plot #
aov1 = aov(lm(chol~region, data=heart.df))
plot(aov1,which=1)
plot(aov1,which=2)
plot(aov1,which=3)
```

Check for normality: Distribution of Heart rate by regions



These Graphs shows us that Maximum Heart Rate achieved is normal



This Clearly indicates that there is not much difference hence we can proceed with Anova test for further analysis

Test Statistics:

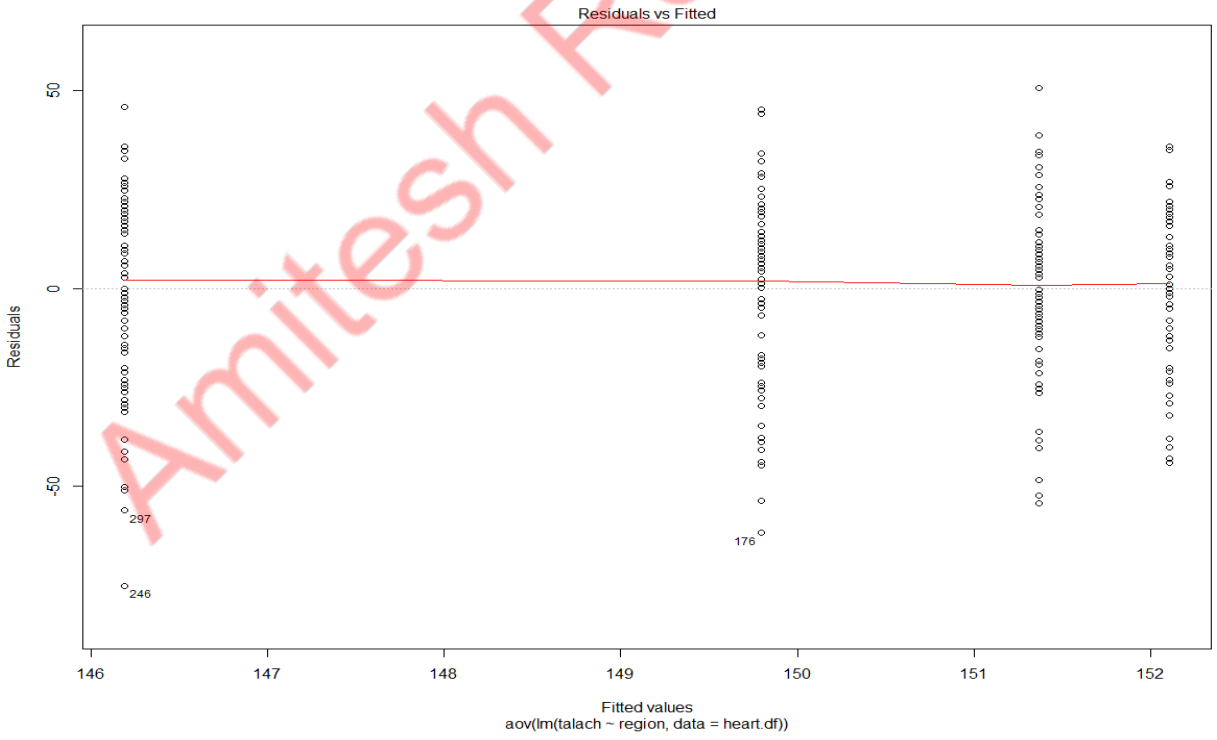
```
> #anova for heart rate for different regions #
> anova(lm(talach~region, data=heart.df))
Analysis of Variance Table

Response: talach
           Df Sum Sq Mean Sq F value Pr(>F)
region      3   1691  563.55  1.0778 0.3587
Residuals 299 156336  522.86
```

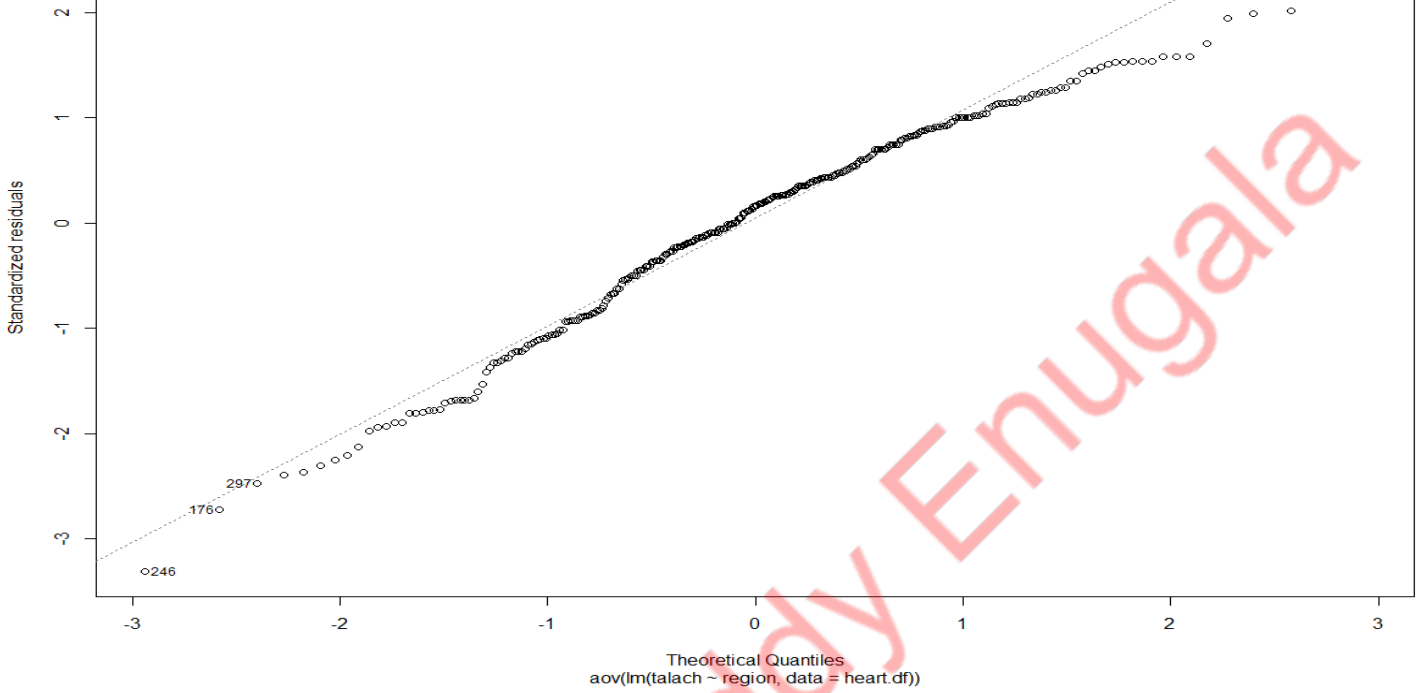
- From the above Anova tests we can say that we cannot reject null hypothesis and conclude that there is not much difference in oldpeak values for people with Heart Disease and No Heart Disease
- p-value = 0.3587 shows a strong indication not to reject Null Hypothesis.

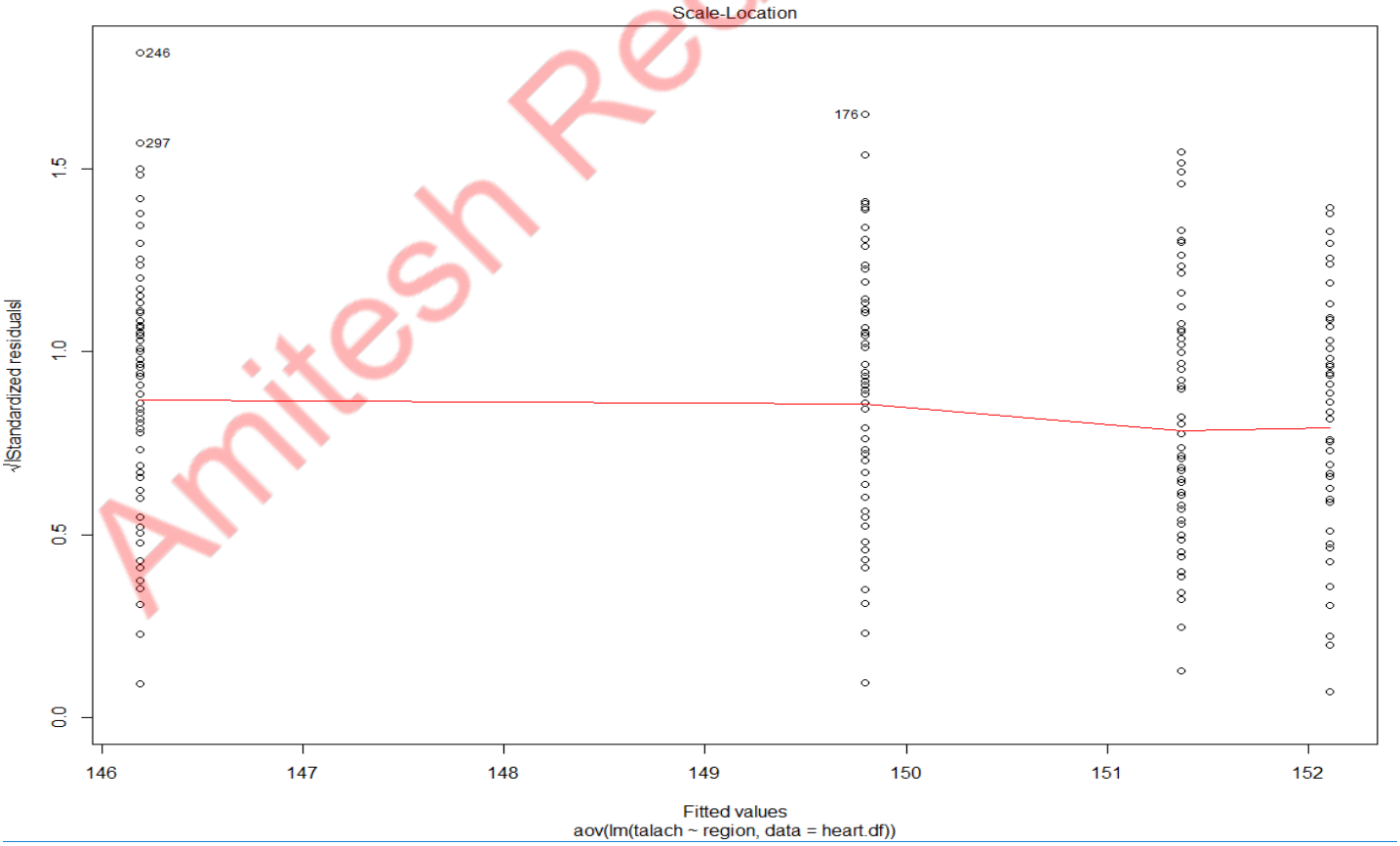
Continue the Test Statistics with Pairwise Comparison for detailed Explanation

```
> # pairwise comparison of each region #
> fit.contrast(lm(talach~region, data=heart.df), "region",c(-1,1,0, + 0),conf.int= .95)
                      Estimate Std. Error   t value  Pr(>|t|)  lower CI upper CI
region c=( -1 1 0 0 ) -5.920486   3.738888 -1.583488 0.1143673 -13.27836 1.437383
attr(,"class")
[1] "fit_contrast"
> fit.contrast(lm(talach~region, data=heart.df), "region",c(-1,0,1, + 0),conf.int= .95)
                      Estimate Std. Error    t value  Pr(>|t|)  lower CI upper CI
region c=( -1 0 1 0 ) -2.314854   3.915635 -0.5911823 0.5548449 -10.02055 5.390841
attr(,"class")
[1] "fit_contrast"
> fit.contrast(lm(talach~region, data=heart.df), "region",c(-1,0,0, + 1),conf.int= .95)
                       Estimate Std. Error    t value  Pr(>|t|)  lower CI upper CI
region c=( -1 0 0 1 ) -0.7409539   3.879365 -0.1909988 0.8486562 -8.375271 6.893363
attr(,"class")
[1] "fit_contrast"
> fit.contrast(lm(talach~region, data=heart.df), "region",c(-0,-1,1, + 0),conf.int= .95)
                     Estimate Std. Error  t value  Pr(>|t|)  lower CI upper CI
region c=( 0 -1 1 0 ) 3.605632   3.601675 1.001099 0.3175886 -3.482211 10.69347
attr(,"class")
[1] "fit_contrast"
> fit.contrast(lm(talach~region, data=heart.df), "region",c(-0,-1,0, + 1),conf.int= .95)
                     Estimate Std. Error  t value  Pr(>|t|)  lower CI upper CI
region c=( 0 -1 0 1 ) 5.179532   3.562209 1.454023 0.1469889 -1.830644 12.18971
attr(,"class")
[1] "fit_contrast"
> fit.contrast(lm(talach~region, data=heart.df), "region",c(-0,0,-1, + 1),conf.int= .95)
                     Estimate Std. Error   t value  Pr(>|t|)  lower CI upper CI
region c=( 0 0 -1 1 ) 1.573901   3.747299 0.4200094 0.6747803 -5.800519  8.94832
attr(,"class")
[1] "fit_contrast"
```

We can say that P value in all pairwise Comparisons has P-value > 0.05 Hence we can say that Maximum Heart Rate achieved is almost same







From the above graphs we can say that there is not much Variance in Heart rate in various Regions

LINEAR REGRESSION

Problem Statement:

Maximum Heart rate achieved by an Individual.

Running Linear Regression

Linear regression is the most basic and commonly used predictive analysis. Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables. At the center of the regression analysis is the task of fitting a single line through a scatter plot. It consists of 3 stages:

1) analyzing the correlation and directionality of the data,
2) estimating the model, i.e., fitting the line, and
3) evaluating the validity and usefulness of the model.

I am running the linear regression algorithm keeping "Maximum Heart Rate Achieved" as the dependent variable. The model is trained and validated on total data

```
#Linear Regression
heart.df <- heart.df[,c(-15,-16)]

#Step 1: Linear Regression using all variables#
lm1 <- lm(talach ~ ., data = heart.df)
summary(lm1)

#Step 2: variable selection using backward selection#
lm2 <- step(lm1, direction = "backward")
summary(lm2)

#Step 3: Linear regression with all single and interaction effects between 2 variables#
lm3 <- lm(talach ~ (age+sex+cp+trestbps+chol+fbs+restecg+exang+oldpeak+slope+ca+thal+num)^2, data = heart.df)
summary(lm3)

#Step 4: variable selection using backward selection#
lm4 <- step(lm3, direction = "backward")
summary(lm4)
#to find if there is an interaction effect#
anova(lm2,lm4)

# Comparison of best model
AIC(lm2)
AIC(lm4)

#predicting Goodness of model
predict.heart <- predict(lm1,heart.df)
e <- heart.df$talach-predict.heart

hist(e)
plot(heart.df$talach,e) + abline(0,0)

#gains
gain <- gains(heart.df$talach,predict.heart,groups=10)

#lift chart
plot(c(0,gain$cume.pct.of.total*sum(predict.heart))~c(0,gain$cume.obs),xlab = "# cases", ylab = 'cumilative',main = "", type="l")
lines(c(0,sum(predict.heart))~c(0,dim(heart.df)[1]),lty=5)

#decile chart
height <- gain$mean.resp/mean(heart.df$talach)
midpoint <- barplot(height,names.arg=gain$depth, ylim= c(0,9),
                    col ="blue", xlab = "percentile", ylab = "Decile lift", main = "Decile-chart")
text(midpoint,height+0.5,labels=round(height,1), cex=.8)
```

Step 1: Running Linear Regression on all variables

Now we run Linear Regression on all Variables

```
> #Step 1: Linear Regression using all variables#
> lm1 <- lm(talach ~ ., data = heart.df)
> summary(lm1)

Call:
lm(formula = talach ~ ., data = heart.df)

Residuals:
    Min      1Q  Median      3Q     Max
-58.733 -10.238   1.853  11.960  48.541

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 191.56295   12.97228  14.767 < 0.0000000000000002 ***
age          -0.78401    0.13106  -5.982          0.0000000065 ***
sex           1.58685    2.60451   0.609             0.542823
cp           -2.45800    1.29011  -1.905             0.057737 .
trestbps      0.14617    0.06534   2.237             0.026034 *
chol          0.02996    0.02615   1.145             0.252966
fbs           1.66681    3.04369   0.548             0.584370
restecg       0.62631    1.10723   0.566             0.572065
exang        -8.29898    2.64222  -3.141             0.001859 **
oldpeak      -0.42271    1.19690  -0.353             0.724219
slope        -8.03298    2.14040  -3.753             0.000211 ***
ca           -0.31932    1.32353  -0.241             0.809520
thal          0.35892    0.71556   0.502             0.616339
num         -11.03779    3.56846  -3.093             0.002174 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.14 on 289 degrees of freedom
Multiple R-squared:  0.3983,    Adjusted R-squared:  0.3713
F-statistic: 14.72 on 13 and 289 DF,  p-value: < 0.00000000000000022
```

We can see few records are not significant Hence we a selection process is required

Step 2: Selecting Best Model by Backward Regression

```
> #Step 2: variable selection using backward selection#
> lm2 <- step(lm1, direction = "backward")
Start:  AIC=1769.87
talach ~ age + sex + cp + trestbps + chol + fbs + restecg + exang +
    oldpeak + slope + ca + thal + num

           Df Sum of Sq    RSS    AIC
- ca        1      19.2  95099 1767.9
- oldpeak   1      41.0  95121 1768.0
- thal      1      82.8  95163 1768.1
- fbs       1      98.7  95179 1768.2
- restecg   1     105.3  95185 1768.2
- sex       1     122.1  95202 1768.3
- chol      1     431.7  95512 1769.2
<none>                   95080 1769.9
- cp        1    1194.3  96274 1771.7
- trestbps  1    1646.7  96727 1773.1
- num       1    3147.7  98228 1777.7
- exang     1    3245.7  98326 1778.0
- slope     1    4634.0  99714 1782.3
- age       1   11772.8 106853 1803.2

Step:  AIC=1767.93
talach ~ age + sex + cp + trestbps + chol + fbs + restecg + exang +
    oldpeak + slope + thal + num

           Df Sum of Sq    RSS    AIC
- oldpeak   1      53.1  95152 1766.1
- thal      1      82.5  95182 1766.2
- fbs       1      88.5  95188 1766.2
- restecg   1     105.0  95204 1766.3
- sex       1     122.1  95221 1766.3
- chol      1     417.8  95517 1767.3
<none>                   95099 1767.9
- cp        1    1211.6  96311 1769.8
- trestbps  1    1696.7  96796 1771.3
- exang     1    3229.7  98329 1776.0
- num       1    3463.6  98563 1776.8
- slope     1    4639.1  99738 1780.4
- age       1   12841.2 107941 1804.3
```

```
Step:  AIC=1766.1
talach ~ age + sex + cp + trestbps + chol + fbs + restecg + exang +
    slope + thal + num

           Df Sum of Sq    RSS    AIC
- thal      1      75.3  95228 1764.3
- fbs       1      98.9  95251 1764.4
- restecg   1     103.2  95256 1764.4
- sex       1     123.2  95276 1764.5
- chol      1     450.6  95603 1765.5
<none>                   95152 1766.1
- cp        1    1214.7  96367 1767.9
- trestbps  1    1654.5  96807 1769.3
- exang     1    3289.1  98442 1774.4
- num       1    3725.8  98878 1775.7
- slope     1    6742.0 101894 1784.8
- age       1   12997.6 108150 1802.9

Step:  AIC=1764.34
talach ~ age + sex + cp + trestbps + chol + fbs + restecg + exang +
    slope + num

           Df Sum of Sq    RSS    AIC
- restecg   1      83.7  95311 1762.6
- fbs       1     109.7  95337 1762.7
- sex       1     187.6  95415 1762.9
- chol      1     441.1  95669 1763.7
<none>                   95228 1764.3
- cp        1    1196.3  96424 1766.1
- trestbps  1    1697.7  96925 1767.7
- exang     1    3248.9  98477 1772.5
- num       1    3918.9  99147 1774.6
- slope     1    6674.5 101902 1782.9
- age       1   13006.7 108234 1801.1

Step:  AIC=1762.61
talach ~ age + sex + cp + trestbps + chol + fbs + exang + slope +
    num

           Df Sum of Sq    RSS    AIC
- fbs       1     116.7  95428 1761.0
- sex       1     193.0  95504 1761.2
- chol      1     529.5  95841 1762.3
<none>                   95311 1762.6
- cp        1    1191.4  96503 1764.4
- trestbps  1    1770.3  97082 1766.2
- exang     1    3266.7  98578 1770.8
- num       1    3892.8  99204 1772.7
- slope     1    6599.6 101911 1780.9
- age       1   12925.7 108237 1799.1


Step:  AIC=1760.98
talach ~ age + sex + cp + trestbps + chol + exang + slope + num

           Df Sum of Sq    RSS    AIC
- sex       1     217.3  95645 1759.7
- chol      1     537.1  95965 1760.7
<none>                   95428 1761.0
- cp        1    1215.3  96643 1762.8
- trestbps  1    1951.3  97379 1765.1
- exang     1    3243.4  98671 1769.1
- num       1    3955.3  99383 1771.3
- slope     1    6540.6 101969 1779.1
- age       1   12810.5 108239 1797.1

Step:  AIC=1759.66
talach ~ age + cp + trestbps + chol + exang + slope + num

           Df Sum of Sq    RSS    AIC
- chol      1     438.3  96084 1759.0
<none>                   95645 1759.7
- cp        1    1482.5  97128 1762.3
- trestbps  1    1822.8  97468 1763.4
- exang     1    3188.9  98834 1767.6
- num       1    3909.3  99555 1769.8
- slope     1    6757.7 102403 1778.3
- age       1   13699.8 109345 1798.2

Step:  AIC=1759.05
talach ~ age + cp + trestbps + exang + slope + num

           Df Sum of Sq    RSS    AIC
<none>                   96084 1759.0
- cp        1    1517.5  97601 1761.8
- trestbps  1    1868.0  97952 1762.9
- exang     1    3058.1  99142 1766.5
- num       1    3473.1  99557 1767.8
- slope     1    6903.5 102987 1778.1
- age       1   13443.6 109527 1796.7
```

Now we get a best optimum Model with Single interaction Effect

```
> summary(lm2)

Call:
lm(formula = talach ~ age + cp + trestbps + exang + slope + num,
    data = heart.df)

Residuals:
    Min      1Q  Median      3Q     Max
-59.565 -10.275   2.127  11.778  45.409

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 201.48835   10.22606  19.703 < 0.0000000000000002 ***
age          -0.79220    0.12310  -6.435       0.000000000495 ***
cp           -2.71028    1.25350  -2.162              0.03141 *
trestbps      0.15062    0.06279   2.399              0.01706 *
exang        -7.96333    2.59446  -3.069              0.00234 **
slope        -8.33508    1.80739  -4.612       0.000005951970 ***
num          -8.78121    2.68455  -3.271              0.00120 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.02 on 296 degrees of freedom
Multiple R-squared:  0.392,     Adjusted R-squared:  0.3797
F-statistic:  31.8 on 6 and 296 DF,  p-value: < 0.00000000000000022
```

Now this model has better Adj – R Squared .3797 than previous model which is .3713

Now Lets check for Interaction Effects

Step 3: Including all Single and Interaction effects Between two Variables

```
> summary(lm3)

Call:
lm(formula = talach ~ (age + sex + cp + trestbps + chol + fbs +
    restecg + exang + oldpeak + slope + ca + thal + num)^2, data = heart.df)

Residuals:
    Min      1Q  Median      3Q     Max
-49.682  -7.314   1.202   9.654  38.693

Coefficients:
                Estimate  Std. Error t value Pr(>|t|)
(Intercept)  281.4779846 117.0018856   2.406  0.01700 *
age           -1.1793105   1.6986447  -0.694  0.48828
sex           -4.8409327  37.6163562  -0.129  0.89772
cp           -32.8357622  16.9361819  -1.939  0.05386 .
trestbps       0.8568625   0.8548853   1.002  0.31734
chol          -0.2990145   0.3467594  -0.862  0.38950
fbs           21.0648862  54.5302116   0.386  0.69967
restecg      -37.8790622  16.8572550  -2.247  0.02567 *
exang         65.8665029  44.3291099   1.486  0.13881
oldpeak       31.5742747  17.8403732   1.770  0.07820 .
slope         12.8502238  33.3816716   0.385  0.70066
ca             9.8007564  21.3295897   0.459  0.64635
thal         -27.3723435  12.1089307  -2.261  0.02481 *
num           -5.9018248  69.4439670  -0.085  0.93235
age:sex       -0.0514061   0.3793631  -0.136  0.89234
age:cp         0.2216002   0.1908988   1.161  0.24702
age:trestbps  -0.0075594   0.0093341  -0.810  0.41893
age:chol       0.0004047   0.0041239   0.098  0.92193
age:fbs       -0.3103207   0.5750057  -0.540  0.58998
age:restecg    0.1199847   0.1635091   0.734  0.46388
age:exang     -0.1751980   0.3893414  -0.450  0.65318
age:oldpeak   -0.0729945   0.1965735  -0.371  0.71076
age:slope     -0.1308949   0.3292758  -0.398  0.69138
age:ca         0.2295732   0.1859176   1.235  0.21827
age:thal       0.1573715   0.1127908   1.395  0.16441
age:num        0.1044329   0.5902174   0.177  0.85973
sex:cp         4.6171888   3.7207762   1.241  0.21601
sex:trestbps  -0.1813000   0.2094071  -0.866  0.38760
sex:chol       0.0214219   0.0731918   0.293  0.77005
sex:fbs       11.2450190   9.3662804   1.201  0.23126
sex:restecg    3.9716645   3.3226909   1.195  0.23331
sex:exang    -18.8330481   8.9280816  -2.109  0.03609 *
sex:oldpeak   -5.2858371   3.7145212  -1.423  0.15621
sex:slope      2.3545226   6.4253595   0.366  0.71440
sex:ca        -3.6478621   4.3448865  -0.840  0.40210
sex:thal       4.8998611   2.6286391   1.864  0.06371 .
sex:num      -11.1257145  13.2761575  -0.838  0.40297
cp:trestbps    0.0188140   0.0944618   0.199  0.84232
cp:chol        0.0650269   0.0461985   1.408  0.16073
cp:fbs        -6.2345187   4.9203452  -1.267  0.20652
cp:restecg     1.1288985   1.6614816   0.679  0.49760
cp:exang      -0.1665732   4.2023585  -0.040  0.96842
cp:oldpeak    -1.4238695   1.7154399  -0.830  0.40746
cp:slope      -2.2451431   3.3477036  -0.671  0.50318
```

```
cp:ca             -2.1401343   2.1310337  -1.004  0.31640
cp:thal            1.0507592   1.0826550   0.971  0.33289
cp:num            -6.0220713   5.0996996  -1.181  0.23898
trestbps:chol     -0.0004522   0.0023324  -0.194  0.84645
trestbps:fbs       0.1465983   0.2084955   0.703  0.48275
trestbps:restecg   0.0429210   0.0916901   0.468  0.64019
trestbps:exang    -0.0224995   0.1933642  -0.116  0.90748
trestbps:oldpeak  -0.0835482   0.0875087  -0.955  0.34080
trestbps:slope    -0.2398064   0.1886017  -1.271  0.20495
trestbps:ca       -0.0800876   0.1078802  -0.742  0.45869
trestbps:thal      0.0464181   0.0514586   0.902  0.36806
trestbps:num       0.1307344   0.2435520   0.537  0.59198
chol:fbs          -0.0285754   0.1039430  -0.275  0.78365
chol:restecg       0.0676309   0.0344529   1.963  0.05096 .
chol:exang        -0.1210693   0.0843326  -1.436  0.15259
chol:oldpeak      -0.0863006   0.0313390  -2.754  0.00641 **
chol:slope         0.0545607   0.0651489   0.837  0.40327
chol:ca            0.0078432   0.0382112   0.205  0.83757
chol:thal          0.0180693   0.0224540   0.805  0.42188
chol:num           0.0642364   0.1195977   0.537  0.59176
fbs:restecg       -7.3734808   4.0156900  -1.836  0.06774 .
fbs:exang         18.7547594  11.4734716   1.635  0.10362
fbs:oldpeak        1.8198573   4.2258661   0.431  0.66716
fbs:slope          2.2453490   7.0253974   0.320  0.74958
fbs:ca            -1.2970519   4.8828046  -0.266  0.79078
fbs:thal          -2.0796157   2.4025987  -0.866  0.38771
fbs:num            8.3543104  13.1242858   0.637  0.52511
restecg:exang     -6.1690865   3.2483262  -1.899  0.05891 .
restecg:oldpeak   -0.3375703   1.4785205  -0.228  0.81962
restecg:slope      5.8944201   2.5950982   2.271  0.02413 *
restecg:ca         0.6510537   1.6647015   0.391  0.69612
restecg:thal       0.0282190   0.9412324   0.030  0.97611
restecg:num       -5.9227121   4.6022998  -1.287  0.19954
exang:oldpeak      1.6735184   3.2493549   0.515  0.60707
exang:slope       -5.6403034   6.2717631  -0.899  0.36951
exang:ca           0.6057570   3.9440330   0.154  0.87808
exang:thal        -2.2725220   1.9133641  -1.188  0.23628
exang:num         13.9375711  11.0952790   1.256  0.21044
oldpeak:slope      4.6057907   1.9714343   2.336  0.02042 *
oldpeak:ca         0.2954029   1.4703579   0.201  0.84097
oldpeak:thal      -0.4974085   0.9195156  -0.541  0.58912
oldpeak:num        9.7726027   4.4012828   2.220  0.02746 *
slope:ca          -4.1676480   3.7106443  -1.123  0.26265
slope:thal         1.8303395   1.7657161   1.037  0.30111
slope:num        -14.2504243   9.6266454  -1.480  0.14028
ca:thal            0.4242357   1.1068214   0.383  0.70189
ca:num             0.1246135   5.6511216   0.022  0.98243
thal:num          -0.2324640   2.4754475  -0.094  0.92527
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.59 on 211 degrees of freedom
Multiple R-squared:  0.5868,    Adjusted R-squared:  0.4085
F-statistic: 3.292 on 91 and 211 DF,  p-value: 0.000000000000654
```

This model has better Adj R squared than previous Model, Lets find the Best model by this from Variable Selection

Step 4: Selecting best Model after including Interaction effect

```
Coefficients:
                Estimate Std. Error t value      Pr(>|t|)
(Intercept)    246.35616   39.42153   6.249 0.00000000165 ***
age             -1.77978    0.48295  -3.685      0.000277 ***
sex            -19.55625   10.40338  -1.880      0.061241 .
cp             -25.34527   10.49382  -2.415      0.016407 *
trestbps         0.63984    0.17543   3.647      0.000319 ***
chol            -0.26614    0.11011  -2.417      0.016329 *
fbs             11.00255   12.01417   0.916      0.360612
restecg        -15.65534    6.25596  -2.502      0.012941 *
exang           27.45110   13.94719   1.968      0.050093 .
oldpeak          7.17460    5.97415   1.201      0.230855
slope           62.91332   16.76257   3.753      0.000215 ***
ca              -6.84523    7.92932  -0.863      0.388770
thal           -17.19300    5.29309  -3.248      0.001312 **
num             11.58585   13.39992   0.865      0.388035
age:cp           0.25141    0.12599   1.995      0.047022 *
age:slope       -0.33115    0.21391  -1.548      0.122809
age:ca           0.25223    0.13077   1.929      0.054831 .
age:thal         0.14345    0.06655   2.155      0.032032 *
sex:cp           5.39512    3.00347   1.796      0.073595 .
sex:exang      -15.40859    5.65860  -2.723      0.006901 **
sex:oldpeak     -4.19151    2.33587  -1.794      0.073895 .
sex:thal         4.03947    1.88297   2.145      0.032847 *
sex:num        -14.39871    7.89732  -1.823      0.069402 .
cp:chol          0.06503    0.03416   1.904      0.058018 .
cp:fbs          -7.27104    3.36906  -2.158      0.031819 *
cp:slope        -5.08435    1.83427  -2.772      0.005972 **
cp:num          -4.59096    3.28649  -1.397      0.163616
trestbps:slope  -0.33872    0.10207  -3.318      0.001033 **
chol:restecg     0.04815    0.02449   1.966      0.050339 .
chol:exang      -0.07142    0.05379  -1.328      0.185426
chol:oldpeak    -0.06112    0.01988  -3.074      0.002336 **
chol:thal        0.02631    0.01423   1.850      0.065487 .
fbs:restecg     -7.09626    2.99394  -2.370      0.018500 *
fbs:exang       20.70144    7.06963   2.928      0.003708 **
fbs:slope        8.60009    4.39827   1.955      0.051602 .
restecg:exang   -6.11256    2.39092  -2.557      0.011133 *
restecg:slope    4.75783    1.79810   2.646      0.008635 **
oldpeak:slope    3.61013    1.47323   2.450      0.014918 *
oldpeak:num      5.14342    2.28608   2.250      0.025283 *
slope:ca        -4.10757    2.07853  -1.976      0.049178 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.41 on 263 degrees of freedom
Multiple R-squared:  0.5516,    Adjusted R-squared:  0.4851
F-statistic: 8.296 on 39 and 263 DF,  p-value: < 0.00000000000000022
```

We can say that this model has the Best Adj R Squared and Can proceed by Checking AIC to decide which model to choose

Step 5: Comparison of Best model among all and test for Interaction Effect by ANOVA Two-way Test

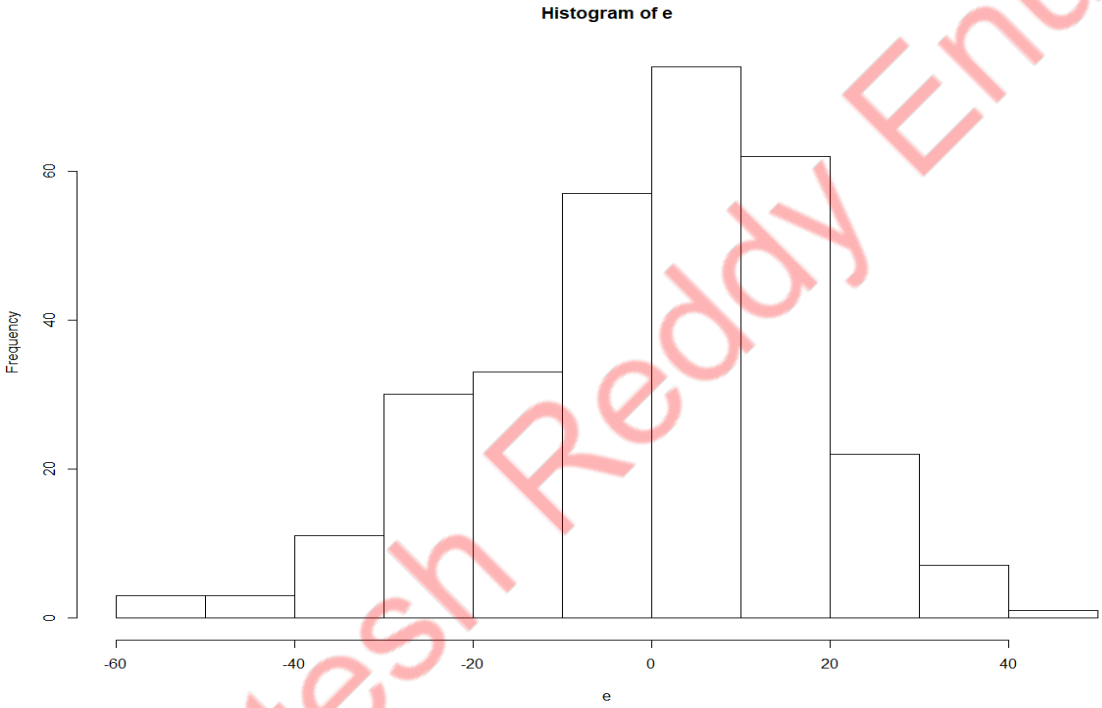
```
> # Comparison of best model
> AIC(lm2)
[1] 2620.927
> AIC(lm4)
[1] 2594.649
```

Hence, we can say that the 4th model with all single and interaction effects is the best

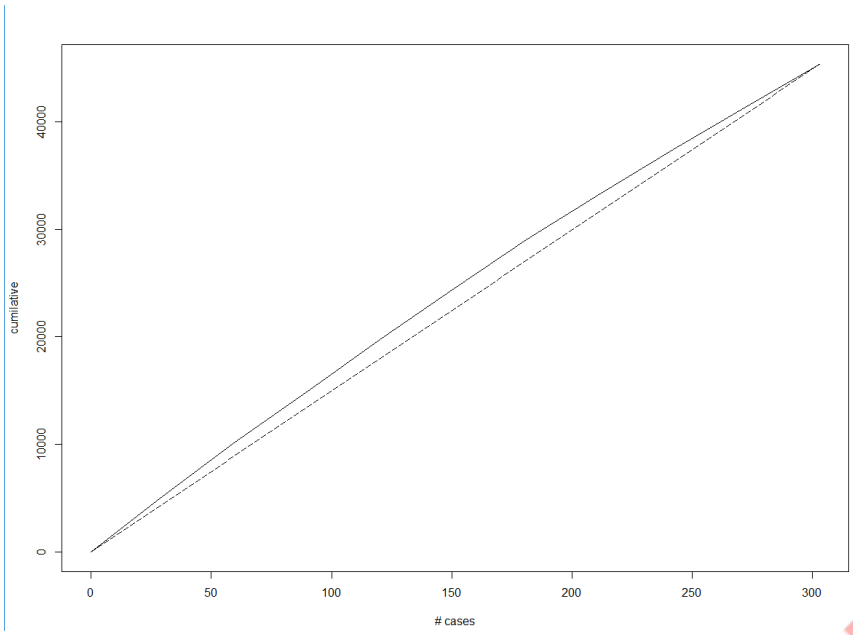
```
> anova(lm2,lm4)
Analysis of Variance Table

Model 1: talach ~ age + cp + trestbps + exang + slope + num
Model 2: talach ~ age + sex + cp + trestbps + chol + fbs + restecg + exang +
    oldpeak + slope + ca + thal + num + age:cp + age:slope +
    age:ca + age:thal + sex:cp + sex:exang + sex:oldpeak + sex:thal +
    sex:num + cp:chol + cp:fbs + cp:slope + cp:num + trestbps:slope +
    chol:restecg + chol:exang + chol:oldpeak + chol:thal + fbs:restecg +
    fbs:exang + fbs:slope + restecg:exang + restecg:slope + oldpeak:slope +
    oldpeak:num + slope:ca
  Res.Df   RSS Df Sum of Sq      F       Pr(>F)
1    296 96084
2    263 70858 33     25226 2.8373 0.000002021 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we can see that p value is significant, Hence we can say that from ANOVA Two way Test interaction effect plays a significant role in the Model

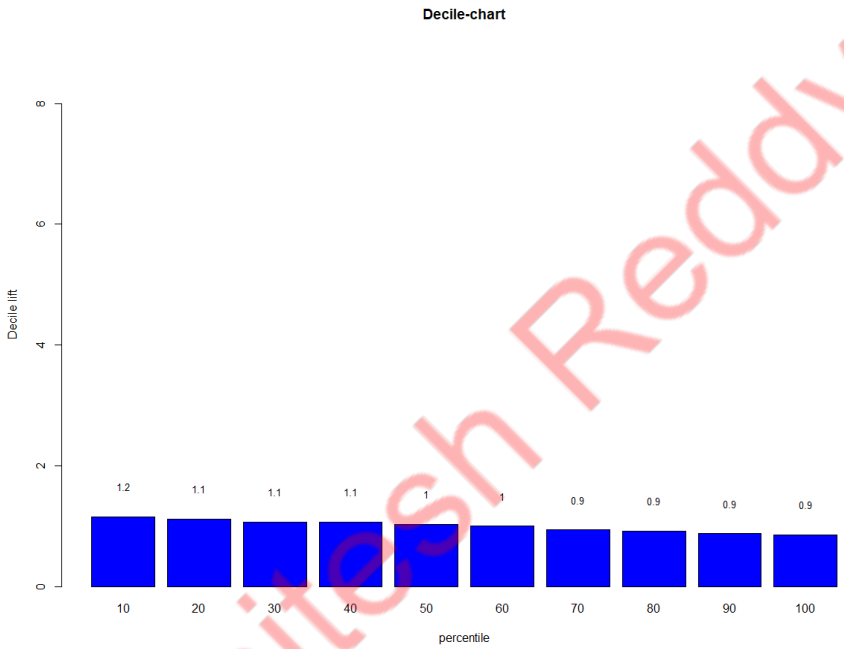
Step 6: Discussion on how good our model is!



We can say that Most of the Residual is around zero and this model is good for estimating values



- As seen from the above lift chart, it is evident that the model curve has comparatively more area (covers more variation) under it compared to the naïve rule represented by the straight line.



- Decile chart follows an i deal structure representing maximum variation covered in initi al deciles.
- This can be considered a s good model where the deciles are de creasing in order from st art to end.
- Looking at the first decil e, we can say that this model perform s 1.4 time better than th e one with Naïve rule.

Interpretations:

- We can see that age, trestbps, slope is the most significant Variables in Model
- Slope has the highest estimates in +ve direction among all Variables (Good Slope makes a person heart beat b etter)
- Sex, cp, restecg, thal, sex*exang, sex*num has highest estimates among all variables in -ve direction (Males h ave lesser Heart Beat than Women and heart beat is high during exercise)
- By residual plot we can say that most Residuals are around 0 hence this Model is good to predict
- By lift Chart we can see the Curve is above Naïve's Rule hence it is a good model to predict values
- By Decile Chart we can say that the maximum Variation covered by initial deciles is not much hence there cou ld be high residual values
- Overall This Model is a Good Model, but the Predicted Values may be deviated from original values

Strengths: Linear regression is straightforward to understand, explain and can be regularized to avoid over fitting. In addition, linear models can be updated easily with new data.

Weaknesses: Linear regression performs poorly when there are non-linear relationships. They are not naturally flexible enough to capture more complex patterns and adding the right interaction terms or polynomials can be tricky and time -consuming.

LOGIT and PROBIT

Problem Statement:

When an individual is Diagnosed, does he have a Heart Disease?

Logistic vs Probit Regression

Logistic regression extends the idea of linear regression to situation where outcome variable is categorical. It is widely used, especially where a structured model is used to explain or predict.

We make a model using logistic regression to predict if a person has Heart Disease or not.

```
# logit and Probit
#Step 1 Run a logit with all variables
lg1 <- glm(num~.,data=heart.df,family='binomial')
options(scipen=999)
summary(lg1)

#Step 2 variable selection using backward selection to choose best Logit model#
lg2 <- step(lg1, direction = "backward")
summary(lg2)


#Step 3 Run the same model with Probit
Pg2 <- glm(num ~ chol + cp + thal + sex + oldpeak + trestbps + exang + talach + fbs,
           family = binomial(link="probit"), data = heart.df)
summary(Pg2)

#step 4 CHoosing between Logit and Probit
AIC(lg2)
AIC(Pg2)

#Accuracy of the Model
pre = predict.attack > 0.95
#confusion matrix
table(heart.df$num, pre)
#Accuracy
mean(pre == heart.df$num)*100

#Estimating the goodness of Model
predict.attack <- predict(lg3,heart.df,type='response')
e <- heart.df$num-predict.attack

plot(heart.df$num,e) + abline(0,0)

hist(e)

 #ROC curve
library(ROCR)
ROCRpred <- prediction(predict.attack, heart.df$num)
ROCRperf <- performance(ROCRpred, 'tpr','fpr')
plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
```

Step 1: Run Logistic Model with all Variables

```
> summary(lg1)

Call:
glm(formula = num ~ ., family = "binomial", data = heart.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5123  -0.1434  -0.0045   0.0746   2.7268

Coefficients:
             Estimate Std. Error z value     Pr(>|z|)
(Intercept) -27.30230    5.67381  -4.812 0.000001494 ***
age          -0.01324    0.03684  -0.359    0.719255
sex           4.15331    0.96486   4.305 0.000016731 ***
cp            1.45030    0.32117   4.516 0.000006309 ***
trestbps      0.07027    0.01823   3.856    0.000115 ***
chol          0.05927    0.01198   4.948 0.000000751 ***
fbs          -1.38379    0.81272  -1.703    0.088632 .
restecg       0.25553    0.29495   0.866    0.386300
talach       -0.05530    0.01753  -3.154    0.001610 **
exang         1.87131    0.63929   2.927    0.003421 **
oldpeak       0.81501    0.35669   2.285    0.022318 *
slope         0.39729    0.62942   0.631    0.527914
ca            1.48177    0.39325   3.768    0.000165 ***
thal          0.58113    0.16156   3.597    0.000322 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 418.591  on 302  degrees of freedom
Residual deviance:  99.142  on 289  degrees of freedom
AIC: 127.14

Number of Fisher Scoring iterations: 8
```

We can see few Variables which aren't Significant, Hence we run a model selection to choose the best one

Step 2: Use backward Selection Process to Choose the Best Model

```
> #Step 2 variable selection using backward selection to choose best Logit model#
> lg2 <- step(lg1, direction = "backward")
Start:  AIC=127.14
num ~ age + sex + cp + trestbps + chol + fbs + restecg + talach +
    exang + oldpeak + slope + ca + thal

           Df Deviance    AIC
- age       1   99.272 125.27
- slope     1   99.534 125.53
- restecg   1   99.906 125.91
<none>          99.142 127.14
- fbs       1  102.236 128.24
- oldpeak   1  105.251 131.25
- exang     1  108.548 134.55
- talach    1  112.194 138.19
- thal      1  114.165 140.16
- ca        1  117.669 143.67
- trestbps  1  118.336 144.34
- cp        1  127.586 153.59
- sex       1  129.277 155.28
- chol      1  149.624 175.62

Step:  AIC=125.27
num ~ sex + cp + trestbps + chol + fbs + restecg + talach + exang +
    oldpeak + slope + ca + thal

           Df Deviance    AIC
- slope     1   99.639 123.64
- restecg   1   99.974 123.97
<none>          99.272 125.27
- fbs       1  102.379 126.38
- oldpeak   1  105.583 129.58
- exang     1  108.626 132.63
- talach    1  113.435 137.44
- thal      1  114.185 138.19
- ca        1  118.236 142.24
- trestbps  1  119.610 143.61
- cp        1  127.672 151.67
- sex       1  130.813 154.81
- chol      1  150.165 174.16
```

```
Step:  AIC=123.64
num ~ sex + cp + trestbps + chol + fbs + restecg + talach + exang +
    oldpeak + ca + thal

           Df Deviance    AIC
- restecg   1  100.453 122.45
<none>          99.639 123.64
- fbs       1  102.494 124.49
- exang     1  109.176 131.18
- oldpeak   1  111.875 133.88
- talach    1  115.949 137.95
- thal      1  116.516 138.52
- ca        1  118.242 140.24
- trestbps  1  119.922 141.92
- cp        1  128.332 150.33
- sex       1  130.826 152.83
- chol      1  150.370 172.37

Step:  AIC=122.45
num ~ sex + cp + trestbps + chol + fbs + talach + exang + oldpeak +
    ca + thal

           Df Deviance    AIC
<none>          100.45 122.45
- fbs       1   103.36 123.36
- exang     1   109.73 129.73
- oldpeak   1   112.89 132.90
- thal      1   116.52 136.52
- talach    1   117.37 137.37
- ca        1   118.61 138.62
- trestbps  1   121.32 141.32
- cp        1   128.84 148.84
- sex       1   132.44 152.44
- chol      1   157.42 177.42
```

```
> summary(lg2)

Call:
glm(formula = num ~ sex + cp + trestbps + chol + fbs + talach +
    exang + oldpeak + ca + thal, family = "binomial", data = heart.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5730  -0.1405  -0.0055   0.0905   2.5821

Coefficients:
             Estimate Std. Error z value    Pr(>|z|)
(Intercept) -26.81295    5.22905  -5.128 0.000000293 ***
sex           4.17893    0.94730   4.411 0.000010270 ***
cp            1.43689    0.31727   4.529 0.000005928 ***
trestbps      0.06857    0.01718   3.992 0.000065601 ***
chol          0.06032    0.01164   5.181 0.000000220 ***
fbs          -1.33470    0.80426  -1.660    0.097005 .
talach       -0.05733    0.01639  -3.497    0.000470 ***
exang         1.83302    0.63007   2.909    0.003623 **
oldpeak       0.94161    0.29923   3.147    0.001651 **
ca            1.39730    0.37253   3.751    0.000176 ***
thal          0.55862    0.14858   3.760    0.000170 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 418.59  on 302  degrees of freedom
Residual deviance: 100.45  on 292  degrees of freedom
AIC: 122.45

Number of Fisher Scoring iterations: 8
```

Now we have generated a Best Logit Model, Lets move further by running it with Probit

Step 3: Now Run Probit Model with the best Model from Logistic Regression

```
> summary(Pg2)

Call:
glm(formula = num ~ chol + cp + thal + sex + oldpeak + trestbps +
    exang + talach + fbs, family = binomial(link = "probit"),
    data = heart.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8495  -0.2890  -0.0031   0.2139   2.9137

Coefficients:
             Estimate Std. Error z value      Pr(>|z|)
(Intercept) -9.962332   1.803494  -5.524 0.00000003315 ***
chol         0.024288   0.004100   5.924 0.00000000313 ***
cp           0.637010   0.136663   4.661 0.00000314419 ***
thal         0.300567   0.065071   4.619 0.00000385428 ***
sex          1.587987   0.339898   4.672 0.00000298354 ***
oldpeak      0.307975   0.113108   2.723       0.006472 **
trestbps     0.024367   0.006962   3.500       0.000466 ***
exang        0.601337   0.272862   2.204       0.027538 *
talach      -0.026721   0.006741  -3.964 0.00007365299 ***
fbs         -0.116520   0.327515  -0.356       0.722012
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 418.59  on 302  degrees of freedom
Residual deviance: 132.74  on 293  degrees of freedom
AIC: 152.74

Number of Fisher Scoring iterations: 8
```

Now we have Generated a Probit Model, let’s move further in discussing Best Model among Logit and Probit

Step 4: Choosing Between Logit and Probit

```
> #step 4 CHoosing between Logit and Probit
> AIC(lg2)
[1] 122.453
> AIC(Pg2)
[1] 152.7372
```

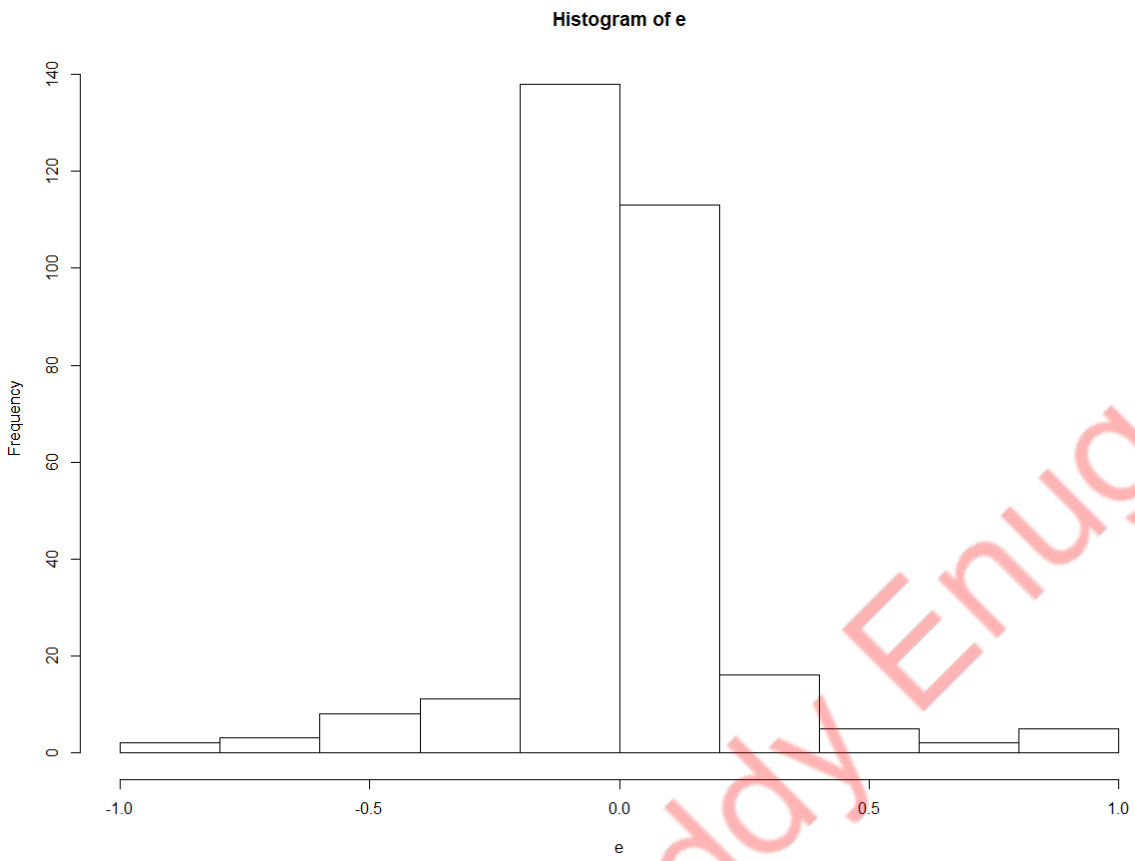
We can Observe that AIC values for Logit is pretty Much low than the values for Probit.

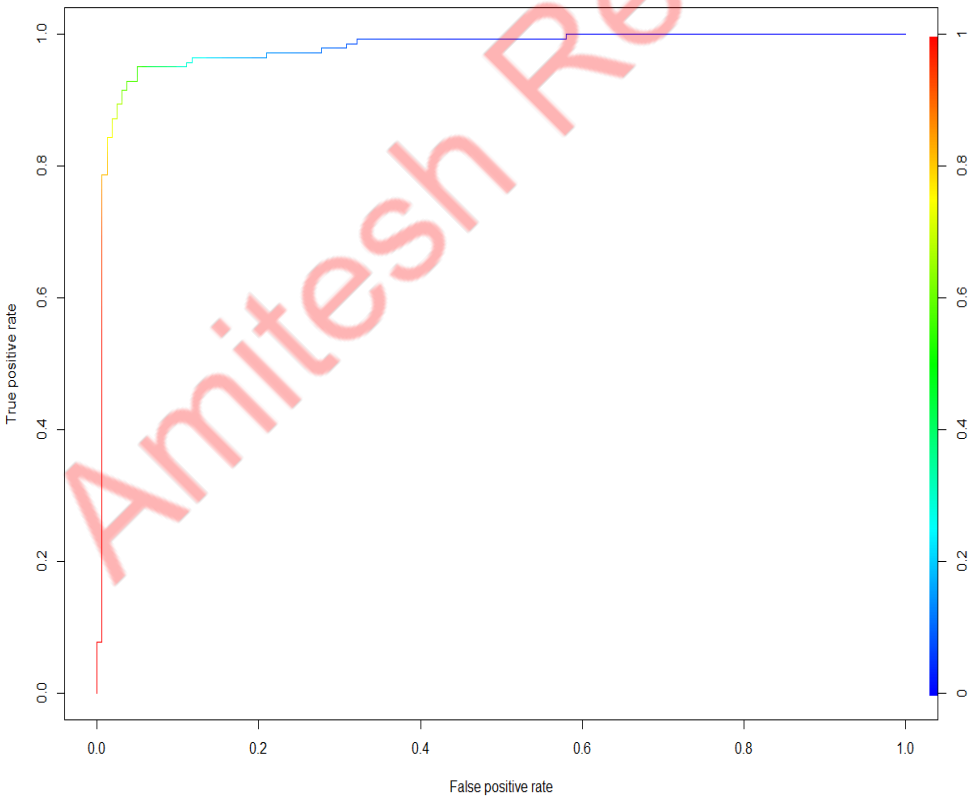
We can Choose Logit Model for further Evaluation of Goodness of Fit.

Step 5: Measuring the Goodness of Fit

```
> #confusion matrix
> table(heart.df$num, pre)
   pre
    FALSE TRUE
  0   161    1
  1    55   86
> #Accuracy
> mean(pre == heart.df$num)*100
[1] 81.51815
```

Here we can see that at 95% probability Threshold we are exhibiting 81% accuracy, This indicates that this is a best model to proceed and estimate





- ROC curve explains about the goodness of curve
- More the Area covered more is the Accuracy
- We can see the area Covered by the Curve is ~ 95%
- Hence this Model is Better for Estimating the Values

Inferences:

- Sex, cp, trestbps, chol, talah, ca, thal the most Significant variables
- Sex, cp, exang, ca has the highest odds among all the Variables that determines more probability of having a disease
- Fbs has the highest odds among all variables that determines highest probability of not having a disease
- The odds ratio of men having heart Disease is 22 times more than Female
- The Residual Plot shows us that most of the residuals are between -0.2 to 0.2 which is of high precision
- ROC Curve Indicates that the model is best in indicating the Values

CHI SQUARED

Problem Statement:

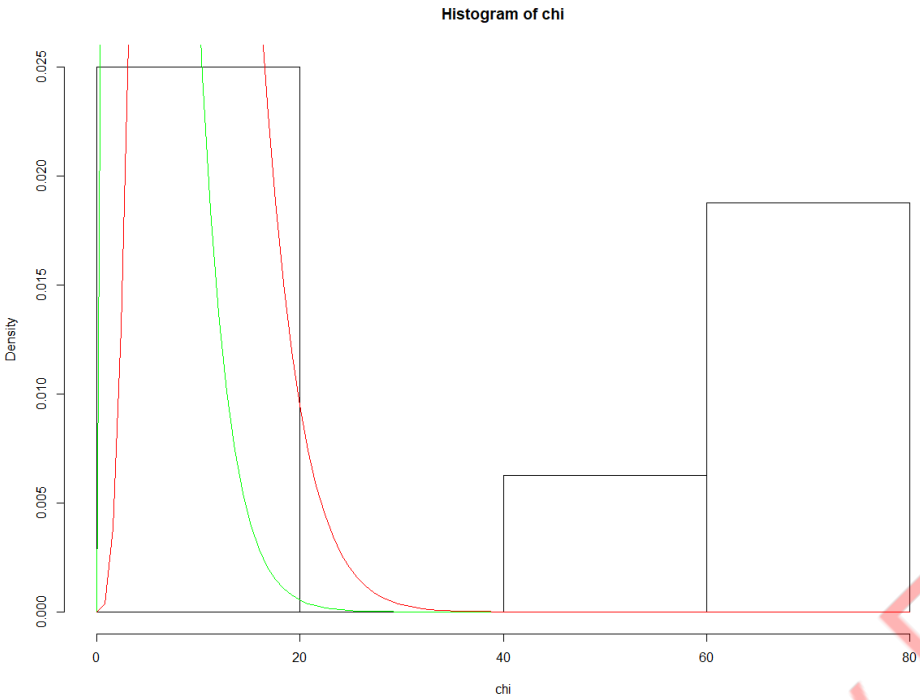
Is Exercise Induced Angina (Exang) having any relation with Chest Pain(cp)?

Chi Squared is used to explain the significance of dependency of two individual Variables

Now we want to calculate does Pain in Chest is due to Exercise Induced by Angina or not. For this we use this method to find it out

```
#chisquared to say sex and num are dependent

chi <- table(heart.df$exang,heart.df$cp)

chisq.test(chi)

hist(chi, prob=TRUE)
curve( dchisq(x, df=5), col='green', add=TRUE)
curve( dchisq(x, df=10), col='red', add=TRUE )

#It may be easier to compare the therotical curve to a density estimate rather than the histogram#
lines( density(x), col='orange')
```

```
> chi <- table(heart.df$exang,heart.df$cp)
> chi

     1  2  3  4
  0 19 46 75 64
  1  4  4 11 80
> chisq.test(chi)

        Pearson's Chi-squared test

data:  chi
X-squared = 66.009, df = 3, p-value = 0.00000000000003052
```



Orange Line Indicates the distribution of Chi with respect to the Histogram. By looking into the lines we can say that there is some dependency

RESULTS

What type of People Get Heart Disaease?

People with high cp, exang, cholesterol tend to get Heart Diseases – t-test, Linear Regression, Heat Map

Will the individual with Chest Pain tend to get Heart Disease? Which individual?

Not all Individuals tend to get heart Disease who has chest pain. But people with chest pain and high levels of cholesterol, fasting blood pressure and slope have highest chances of Heart Disease – Chi-Squared

What is the likelihood of the Individual Getting a Heart Disease?

Sex, cp, trestbps, chol, talah, ca, thal the most Significant variables which determine the likelihood of getting a Heart Disease. Males have high odds of getting Heart Disease, People with low trestbps & talah, high cp and thal are most likely to have a heart Disease – Logit Model

How are our findings and regression model can build up a story?

The results for Logit Model tend to have high Accuracy of 92% for .70 threshold, 87% accuracy for .8 threshold, 86% for .9 threshold. Such high accuracies narrate that our model is the best fit – Logit Model

Does people with chest pain is due to ST depression induced by angina?

No people who Even have ST depression induced by Angina are not diagnosed with Heart Disease, but there is a high probability of having a Heart Disease

REFERENCES & CITATIONS

Book
The Statistical Sleuth A Course in Methods of Data Analysis -THIRD EDITION
Fred L. Ramsey (Oregon State University)
Daniel W. Schafer (Oregon State University)
Website
R- Graph Gallery. 201. https://www.r-graph-gallery.com/portfolio/ggplot2-package/

UCI Machine Learning Repository - https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Kaggle - https://www.kaggle.com/ronitf/heart-disease-uci

R- Statistics - http://r-statistics.co/