# Heart_Disease Prediction Report

Dodgecarl Incila

3/8/2022

```
#for installation of these packages, please see the R file together of this report
library(caret)
library(tidyverse)
library(data.table)
library(dplyr)
```

```
setwd("C:/Users/ll/Desktop/DATA SCIENCE/Finals/Heart")
```

## INTRODUCTION

Heart disease refers to any condition affecting the heart. There are many types, some of which are preventable.

According to the Centers for Disease Control and Prevention (CDC), heart disease is the leading cause of death in the United States.

With the use of R and Rstudio, we will perform Exploratory Analysis (EDA) involving the heart disease data to further understand the nature and statistics of the disease and through Machine Learning modeling, we aim to predict a diagnosis of heart disease based on the available data in this report.

## THE DATA

Main: https://www.kaggle.com/cherngs/heart-disease-cleveland-uci

Download for report: https://raw.githubusercontent.com/inciladc/rfiles/main/heart101.csv

**Background of the Data**

I acknowledge the following for providing such data to work on:

Creators: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. University Hospital, Zurich, Switzerland: William Steinbr Creators: Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D. Donor: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

**Data Retrieval**

Retrieving Data from Github repo:

```
dat <- read.csv("https://raw.githubusercontent.com/inciladc/rfiles/main/heart101.csv")
```

**Inspect Raw Data**

```r
#inspect raw data
str(dat)
```

```
## 'data.frame':    297 obs. of  14 variables:
##  $ age      : int  69 69 66 65 64 64 63 61 60 59 ...
##  $ sex      : int  1 0 0 1 1 1 1 1 0 1 ...
##  $ cp       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ trestbps : int  160 140 150 138 110 170 145 134 150 178 ...
##  $ chol     : int  234 239 226 282 211 227 233 234 240 270 ...
##  $ fbs      : int  1 0 0 1 0 0 1 0 0 0 ...
##  $ restecg  : int  2 0 0 2 2 2 2 0 0 2 ...
##  $ thalach  : int  131 151 114 174 144 155 150 145 171 145 ...
##  $ exang    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak  : num  0.1 1.8 2.6 1.4 1.8 0.6 2.3 2.6 0.9 4.2 ...
##  $ slope    : int  1 0 2 1 1 1 2 1 0 2 ...
##  $ ca       : int  1 2 0 1 0 0 0 2 0 0 ...
##  $ thal     : int  0 0 0 0 0 2 1 0 0 2 ...
##  $ condition: int  0 0 0 1 0 0 0 1 0 0 ...
```

Our Heart Disease raw data consist of 297 rows and 14 variables.

**Attributes of the Variables**

1.) age: age in years

2.) sex: sex (1 = male; 0 = female)

3.) cp: chest pain type – Value 0: typical angina – Value 1: atypical angina – Value 2: non-anginal pain – Value 3: asymptomatic

4.) trestbps: resting blood pressure (in mm Hg on admission to the hospital)

5.) chol: serum cholestoral in mg/dl

6.) fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7.) restecg: resting electrocardiographic results – Value 0: normal – Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) – Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8.) thalach: maximum heart rate achieved

9.) exang: exercise induced angina (1 = yes; 0 = no)

10.) oldpeak = ST depression induced by exercise relative to rest

11.) slope: the slope of the peak exercise ST segment – Value 0: upsloping – Value 1: flat – Value 2: downsloping

12.) ca: number of major vessels (0-3) colored by flourosopy

13.) thal: 0 = normal; 1 = fixed defect; 2 = reversable defect and the label

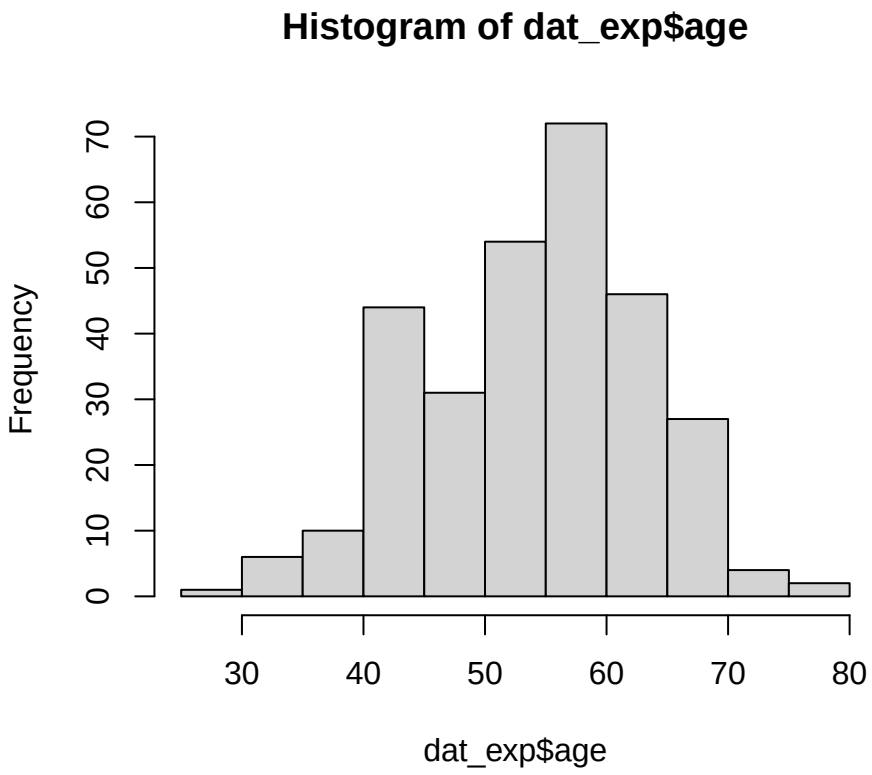
14.) condition: 0 = no disease, 1 = disease

## EXPLORING THE DATA

Let us now examine what the data tells us upon preliminary observation.

```r
#store main dat object to dat_exp for exploration purposes
dat_exp <- dat
```

**Age Distribution**




```
## [1] 297
```

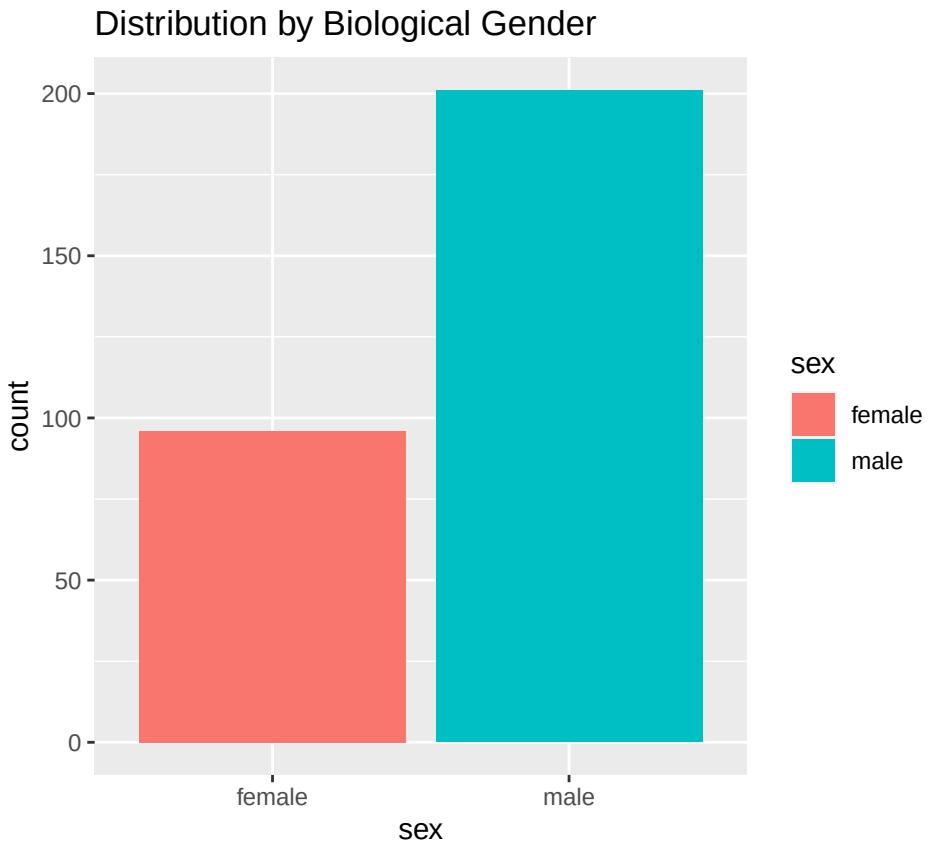
```
## [1] 54.54209
```

```
## [1] 29
```

```
## [1] 77
```

We have 297 observed data points with age distribution centering at 54.54yo. The youngest at 29yo and oldest at 77yo.

**Gender Distribution**

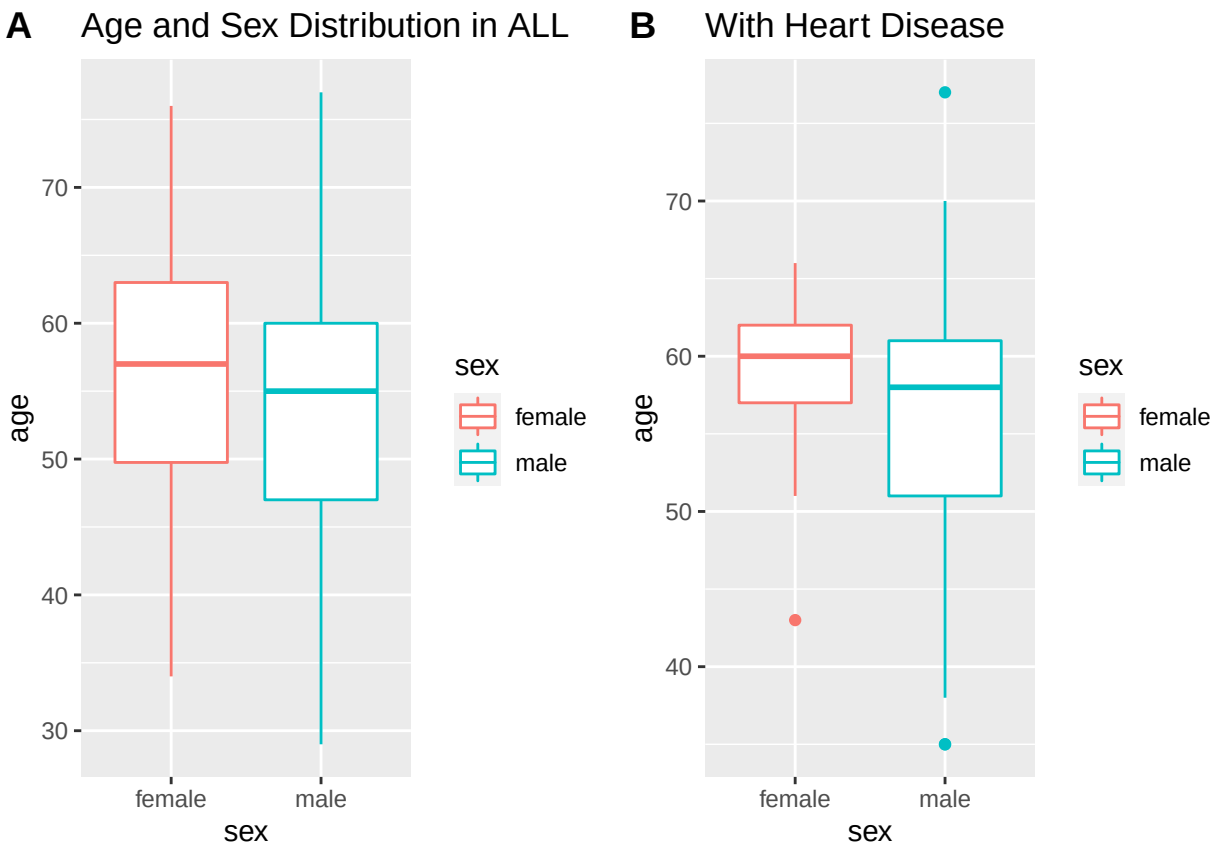



```
##   Total Male Female Difference
## 1   297  201     96        105
```

We have a total of 297 observations of which majority are males at 201 and females at 96. We have more males then females by 105 difference.

**Age, Gender and Heart Disease**

Let us examine how Age and Gender affects Heart Disease




In our data set, there are more males who have heart disease then females. Males tend to develop the disease at younger age then females. This could be due to the fact that there are more males than females in our data set.

```
##   Total_Male with_disease percent_male
## 1        201          112     55.72139
```

```
##   Total_Female with_disease percent_female
## 1           96           25       26.04167
```

```
##   Average_age_in_all
## 1           54.54209
```

```
##   Age_avg_withheartdisease
## 1                 56.75912
```

However the numbers tell us, while there are more male in our data set percentage shows that 55% of males have the disease and 26% on females. Hence males still have more heart disease than females.

On age in all, data suggest that heart disease show on average of 56.76 years old.

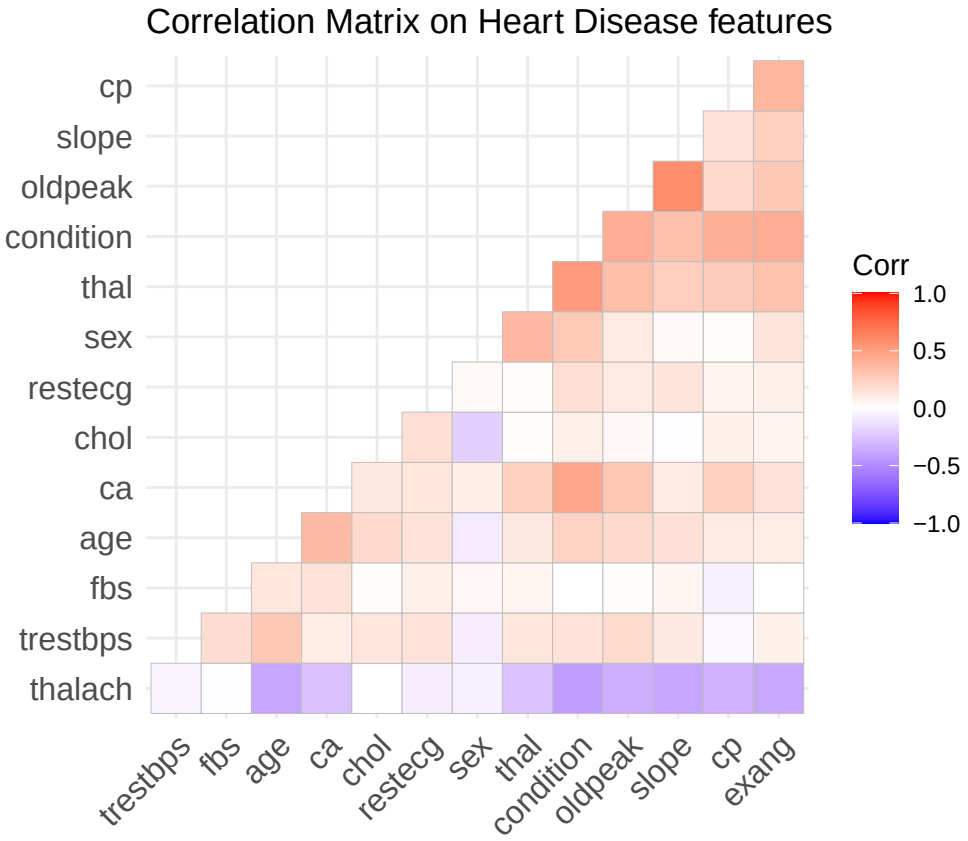
## FURTHER INVESTIGATION

Gender and Age is innate in all of us. But why others have the heart disease and others don't have? Let us examine underlying reasons, based on the data we have, what relationship there is into developing a Heart Disease and see if we can use them as a predictors in building our prediction model for the disease.

**Correlation of Features**

Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.

```
##              age   sex    cp trestbps  chol   fbs restecg thalach exang oldpeak
## age         1.00 -0.09  0.11     0.29  0.20  0.13    0.15   -0.39  0.10    0.20
## sex        -0.09  1.00  0.01    -0.07 -0.20  0.04    0.03   -0.06  0.14    0.11
## cp          0.11  0.01  1.00    -0.04  0.07 -0.06    0.06   -0.34  0.38    0.20
## trestbps    0.29 -0.07 -0.04     1.00  0.13  0.18    0.15   -0.05  0.07    0.19
## chol        0.20 -0.20  0.07     0.13  1.00  0.01    0.17    0.00  0.06    0.04
## fbs         0.13  0.04 -0.06     0.18  0.01  1.00    0.07   -0.01  0.00    0.01
## restecg     0.15  0.03  0.06     0.15  0.17  0.07    1.00   -0.07  0.08    0.11
## thalach    -0.39 -0.06 -0.34    -0.05  0.00 -0.01   -0.07    1.00 -0.38   -0.35
## exang       0.10  0.14  0.38     0.07  0.06  0.00    0.08   -0.38  1.00    0.29
## oldpeak     0.20  0.11  0.20     0.19  0.04  0.01    0.11   -0.35  0.29    1.00
## slope       0.16  0.03  0.15     0.12 -0.01  0.05    0.14   -0.39  0.25    0.58
## ca          0.36  0.09  0.24     0.10  0.12  0.15    0.13   -0.27  0.15    0.29
## thal        0.12  0.37  0.27     0.13  0.02  0.05    0.01   -0.26  0.32    0.34
## condition   0.23  0.28  0.41     0.15  0.08  0.00    0.17   -0.42  0.42    0.42
##             slope    ca  thal condition
## age         0.16  0.36  0.12      0.23
## sex         0.03  0.09  0.37      0.28
## cp          0.15  0.24  0.27      0.41
## trestbps    0.12  0.10  0.13      0.15
## chol       -0.01  0.12  0.02      0.08
## fbs         0.05  0.15  0.05      0.00
## restecg     0.14  0.13  0.01      0.17
## thalach    -0.39 -0.27 -0.26     -0.42
## exang       0.25  0.15  0.32      0.42
## oldpeak     0.58  0.29  0.34      0.42
## slope       1.00  0.11  0.26      0.33
## ca          0.11  1.00  0.25      0.46
## thal        0.26  0.25  1.00      0.52
## condition   0.33  0.46  0.52      1.00
```




Right off the bat we see which among the variables we can use that lean towards positive correlation for "condition" which is our goal in building the predictive model. oldpeak, slope, cp, thal and exang are leaning more towards the positive,Corr 1, for Condition.

**Predictors for the Model**

Now that we are able to see the relationship between variable to our target, condition, let us create a new object that stored data for modeling

```
#store predictors on object mod_dat
mod_dat <- dat[,c(3,9,10,11,13,14)]

#examine structure of mod_dat
str(mod_dat)
```

```
## 'data.frame':    297 obs. of  6 variables:
##  $ cp       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ exang    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak  : num  0.1 1.8 2.6 1.4 1.8 0.6 2.3 2.6 0.9 4.2 ...
##  $ slope    : int  1 0 2 1 1 1 2 1 0 2 ...
##  $ thal     : int  0 0 0 0 0 2 1 0 0 2 ...
##  $ condition: int  0 0 0 1 0 0 0 1 0 0 ...
```

From 13 variables (minus condition as this is our expected output) from the original data set, we now only have 5 variables based on the correlation leaning towards the positive for "condition"

## BUILDING THE MODEL USING LINEAR REGRESSION

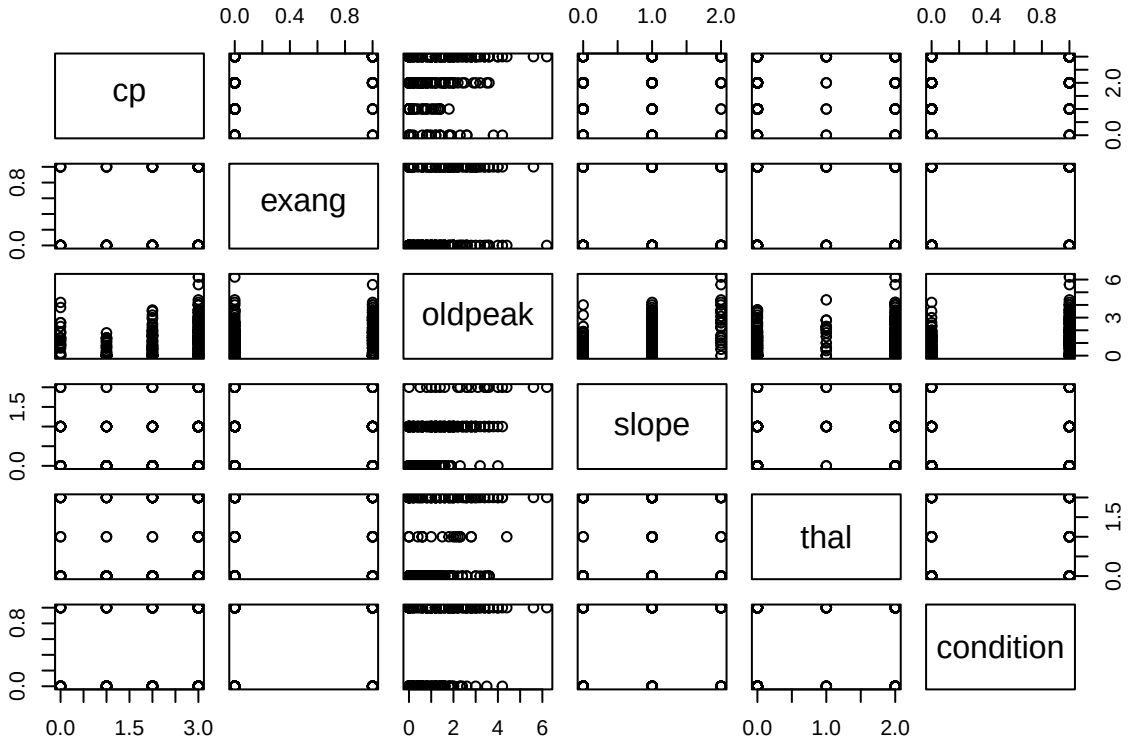
Linear regression is one of the most commonly used predictive modelling techniques.It is represented by an equation Y = a + bX + e, where a is the intercept, b is the slope of the line and e is the error term. This equation can be used to predict the value of a target variable based on given predictor variable(s)

As required in capstone we will use Linear Regression,and see if it is suitable or not in modeling for our heart disease prediction.

### Linearity of Features

Let us check the linearity of our features before we start.

```
pairs(mod_dat[1:6])
```




We could really not see any linearity amongst the features of the data.

Linear regression may not be a suitable for classification output. However let us see what we can learn in running regression.

### Running regression

```
#run regression
result <- lm(condition~cp+exang+oldpeak+slope+thal, data = mod_dat)

#view summary
summary(result)
```

```
##
```

```
## Call:
## lm(formula = condition ~ cp + exang + oldpeak + slope + thal,
##     data = mod_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82791 -0.23589 -0.04655  0.24262  1.09184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.09184    0.05671  -1.619  0.10647
## cp           0.10924    0.02516   4.341 1.96e-05 ***
## exang        0.17490    0.05353   3.267  0.00122 **
## oldpeak      0.07809    0.02419   3.228  0.00139 **
## slope        0.05491    0.04424   1.241  0.21551
## thal         0.17329    0.02568   6.747 8.13e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3807 on 291 degrees of freedom
## Multiple R-squared:  0.4284, Adjusted R-squared:  0.4186
## F-statistic: 43.63 on 5 and 291 DF,  p-value: < 2.2e-16
```

cp and thal prove to be most significant among the features based on the P-values.

**Improved Model**

This time we will run regression using cp and thal only.

```
#we will use cp and thal only for this regression
imp <- lm(condition~cp+thal,data = mod_dat)
```

**ANOVA test**

```
anova(result,imp)
```

```
## Analysis of Variance Table
##
## Model 1: condition ~ cp + exang + oldpeak + slope + thal
## Model 2: condition ~ cp + thal
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1    291 42.184
## 2    294 48.003 -3    -5.819 13.38 3.324e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Prediction**

We will use our improved model to run prediction for heart disease where,

cp: chest pain type – Value 0: typical angina – Value 1: atypical angina – Value 2: non-anginal pain – Value 3: asymptomatic

thal: 0 = normal; 1 = fixed defect; 2 = reversible defect

```
#cp values = 0,1,2,3
#thal values = 0,1,2
#Scenario 1: typical angina, normal thal
predict(imp,data.frame(cp=0,thal=0))
```

```
##           1
## -0.05681502
```

```
#Scenario 2: non-anginal pain, normal thal
predict(imp,data.frame(cp=1,thal=0))
```

```
##          1
## 0.09377201
```

```
#Scenario 3: asymptomatic, fixed defect
predict(imp,data.frame(cp=3,thal=1))
```

```
##         1
## 0.6261871
```

## BUILDING THE MODEL USING RANDOM FOREST

For our prediction model, we will be using RANDOM FOREST. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning.

**Preparation** Make sure that randomForest is installed and the library is run in your rstudio

```
#Install randomforest package if not yet installed
#If giving you an error, update your randomForest package and run rfNews() and re run library
#install.packages("randomForest")
library(randomForest)
library(caret)
```

### Data to use for RF model

We will be using the mod_dat object used in the previous sections. We already filtered down to the most correlated to "condition" which is our desired ouput.

```
#check data structure
str(mod_dat)
```

```
## 'data.frame':    297 obs. of  6 variables:
##  $ cp       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ exang    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak  : num  0.1 1.8 2.6 1.4 1.8 0.6 2.3 2.6 0.9 4.2 ...
##  $ slope    : int  1 0 2 1 1 1 2 1 0 2 ...
##  $ thal     : int  0 0 0 0 0 2 1 0 0 2 ...
##  $ condition: int  0 0 0 1 0 0 0 1 0 0 ...
```

```
#set seed for reproducibility
set.seed(222)
```

### Data Partition

For training and testing the model, will create train and test sets from the mod_dat data set:

```
#store mod_dat to dp for partition
dp <- mod_dat

#set condition as a factor
dp$condition <- as.factor(dp$condition)

#set particion at 70/30 for train and test respectively
ind <- sample(2, nrow(dp), replac=T, prob = c(0.7,0.3))
rf_train <- dp[ind==1,]
rf_test <- dp[ind==2,]

#See structures of train and test set
str(rf_train)
```

```
## 'data.frame':    205 obs. of  6 variables:
##  $ cp       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ exang    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ oldpeak  : num  1.8 2.6 1.4 2.3 2.6 0.9 4.2 0.2 0 0.8 ...
##  $ slope    : int  0 2 1 2 1 0 2 1 0 0 ...
##  $ thal     : int  0 0 0 1 0 0 2 2 0 0 ...
##  $ condition: Factor w/ 2 levels "0","1": 1 1 2 1 2 1 1 2 2 2 ...
```

```
str(rf_test)
```

```
## 'data.frame':    92 obs. of  6 variables:
##  $ cp       : int  0 0 0 0 0 0 1 1 1 1 ...
##  $ exang    : int  0 1 0 0 0 1 0 1 0 0 ...
##  $ oldpeak  : num  0.1 1.8 0.6 1 1.9 1.4 0 0 0 0 ...
##  $ slope    : int  1 1 1 0 1 0 0 1 0 0 ...
##  $ thal     : int  0 0 2 0 2 2 0 1 0 0 ...
##  $ condition: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 2 ...
```

We have 214 obs on train set and 83 obs on test set.

### Run RF on Train set

```
#
rf <- randomForest(condition~cp+thal+exang+oldpeak+slope, data = rf_train)
print(rf)
```

```
##
## Call:
##  randomForest(formula = condition ~ cp + thal + exang + oldpeak +      slope, data = rf_train)
##                Type of random forest: classification
##                      Number of trees: 500
```

```
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 22.93%
## Confusion matrix:
##    0  1 class.error
## 0 85 20   0.1904762
## 1 27 73   0.2700000
```

### Predict using RF model

```
#predict
p1 <- predict(rf,rf_train)

#see confusion matrix
confusionMatrix(p1, rf_train$condition, positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 99 11
##          1  6 89
##
##                Accuracy : 0.9171
##                  95% CI : (0.8705, 0.9509)
##     No Information Rate : 0.5122
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8338
##
##  Mcnemar's Test P-Value : 0.332
##
##             Sensitivity : 0.8900
##             Specificity : 0.9429
##          Pos Pred Value : 0.9368
##          Neg Pred Value : 0.9000
##              Prevalence : 0.4878
##          Detection Rate : 0.4341
##    Detection Prevalence : 0.4634
##       Balanced Accuracy : 0.9164
##
##        'Positive' Class : 1
##
```

### Apply on test set

```
p2 <- predict(rf,rf_test)
confusionMatrix(p2,rf_test$condition,positive = "1")
```

```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction  0  1
##          0 47 13
##          1  8 24
##
##                Accuracy : 0.7717
##                  95% CI : (0.6725, 0.8528)
##     No Information Rate : 0.5978
##     P-Value [Acc > NIR] : 0.0003311
##
##                   Kappa : 0.5146
##
##  Mcnemar's Test P-Value : 0.3827331
##
##             Sensitivity : 0.6486
##             Specificity : 0.8545
##          Pos Pred Value : 0.7500
##          Neg Pred Value : 0.7833
##              Prevalence : 0.4022
##          Detection Rate : 0.2609
##    Detection Prevalence : 0.3478
##       Balanced Accuracy : 0.7516
##
##        'Positive' Class : 1
##
```

## RESULTS

### Linear Model

Using cp and thal as predictors on our final model

| Scenario | Condition (1) |
|---|---|
| (1) cp = 0 thal = 0 | -0.05681502 |
| (2) cp = 1 thal = 0 | 0.09377201 |
| (3) cp = 3 thal = 1 | 0.6261871 |

Looking at the results, our 3rd scenario, based on the given parameters show to be more closer to be diagnosed of heart disease at 0.62 vs scenario 1 which shows a negative value with parameters cp=0,thal=0

Is this related to Heart Disease? I am not sure as I am not in the medical field and I cannot picture out what cp or thal is. However base on the prediction outputs, Scenario 3 may have a worse heart condition than the other two data wise, Hence, if I were to advise this person on scenario 3 to be careful with his heart health and have a follow up with his doctor.

### Random Forest Model

| RF Model testing | Accuracy | Sensitivity | Specificity |
|---|---|---|---|
| P1 on Training Set | 0.9336 | 0.8925 | 0.9661 |
| P2 on Testing Set | 0.686 | 0.6136 | 0.7619 |

Our RF model using all predictors on training set shows high accuracy and acceptable Sensitivity and specificity. Our goal for the heart disease prediction is to determine a "positive" prediction of the disease hence sensitivity of detecting (True Positive) is important.

Applying our model on the two data sets, shows different values on Sensitivity. Even Accuracy is impacted greatly. While we have good numbers in our Training set, but Our goal to have a model that is Sensitive enough to detect Condition = 1, needs to be improved at 61% in the training set.

**Limitation** One of the limitation I would highlight in this report is the number of observations available. With only 297 data points to work on, we have not maximized the ability of Random Forest. Also, Resources such as acceptable computing machine used and learning curve of the vast knowledge there is in data science.

## CONCLUSION

Machine Learning and Data Science allows us to look at problem-solving in a different perspective. Although one limitation I see going through the entire course is data. If the data is erroneous or lack therof, it will impact every analysis every modeling we build.

Machine Learning as a tool and technique will take us to great heights in data-driven decision making.

This report is a culmination of the possibilities Data Science can do for us. From building businesses to solving medical impossibilities, we are in the age where data is power. And Data Science is where it all begin.