# Exploring Risk factors and Prediction of Risk of Heart Disease using machine learning

**HI 2021 PRACTICAL STATISTICS & PROGRAMMING USING R**

**University of Pittsburgh – School of Health and Rehabilitation Sciences**

**Guided by - Dr. Yanshan Wang, PhD, FAMIA**

**By : Sree Gayatri Anusha Mylavarapu**

## Introduction -

- Heart and Cardiovascular diseases are becoming significant global healthcare challenges causing high levels of mortality. These diseases are responsible for a lot of deaths annually in the world. These diseases encompass a range of conditions affecting the heart and blood vessels, including coronary artery disease, heart failure, stroke, and peripheral artery disease. Despite advances in medical science and healthcare, CVDs continue to impose a heavy burden on individuals, families, and healthcare systems.
- The importance of predicting cardiovascular diseases cannot be overstated. Early identification of individuals at risk of developing CVDs is crucial for implementing preventive measures, optimizing treatment strategies, and ultimately reducing the burden of these diseases on society. By accurately assessing an individual's risk of developing a cardiovascular condition, healthcare professionals can intervene proactively to mitigate risk factors, promote lifestyle modifications, and initiate appropriate medical interventions.
- To do this I will be using machine learning techniques to show the prediction of heart disease and help clinicians improve their care and diagnosis of heart diseases.
- The goal of this project is to explore the factors that can cause risk of heart disease and develop a predictive model with the help of machine learning and identify individuals with and without heart disease in my dataset.
- The dataset used in my analysis is "HEART DISEASE WANG.csv", and this is the link to my dataset https://www.kaggle.com/datasets/rishidamarla/heart-disease-prediction .
- The dataset I used was taken from Kaggle and this data comes from the University of California Irvine's machine learning repository.

- This dataset has 14 variables and 260 observations, and I have taken the summary of the dataset and done a descriptive statistics and this provides insights into the distribution, central tendency, and variability of these features among individuals in the dataset, aiding in understanding the demographic and clinical characteristics associated with heart disease risk.
- The outcome of my target variable is Heart Disease which is binary that means that there is presence or absence of heart disease.
- There have been studies that showed a relationship between causative factors like age, BP , cholesterol with the presence of heart disease but could not provide a detailed description of which factor may be the most evident or risk factor among these. So to check or test that out I would like to use machine learning techniques and build a robust model that can identify the factor that can show risk of heart disease and provide insights to health care professionals and clinicians to make targeted decisions in providing the best and preventive care.

**Aims -**

Hypothesis Testing -

Aim 1 - I would like to build a null hypothesis (H0) and an alternate hypothesis (H1) to check if there is any significant difference between the mean levels of age , cholesterol and BP with individuals with and without heart disease.

Null Hypothesis (H0) - There is no significant difference in the mean levels of age, cholesterol, and blood pressure between individuals with and without heart disease.

Alternate Hypothesis (H1) - There is significant difference in the mean values of BP, age and cholesterol between individuals with and without heart disease.

- This project also aims to perform statistical analysis and conduct T- tests between age cholesterol and BP in groups of two and additionally perform an ANOVA test as well to find if there is any relationship between cholesterol levels and type of chest pain.
- I have also created box plots, histograms and scatter plots to visually inspect any relationships between the variables that can show a presence of heart disease.

Aim 2 -

- Machine Learning model - The goal of my project is to implement machine learning techniques and test the variables to show or predict the presence or absence of heart disease. The variables I am testing are Age, BP and Cholesterol.
- I will be using the 'caret' package in R to run logistic regression as my machine learning model and also evaluate metrics like accuracy , sensitivity , ROC curve and confusion matrix to find out the results to test the prediction of presence or absence of heart disease.

## Quality Control -

There are some quality control methods that have to be used in my project to attain the results as needed and some of the measures are

1. Readability of the documentation -
    - Ensuring that the code we run is organized and performs well and makes it easy to understand the functionality.
    - Choosing the right variables while running the machine learning model is also necessary.

2. Handling missing values and data validation -
    - Checking the validity of the data and reordering according to the format wanted.
    - Checking and handling the missing values.
3. Testing the Code -
    - Code has to be tested to check for the robustness and any debugging errors have to be fixed.
4. Performance optimization -
    - The performance of the code has to be checked and by using different functions and models the code can show improvement.

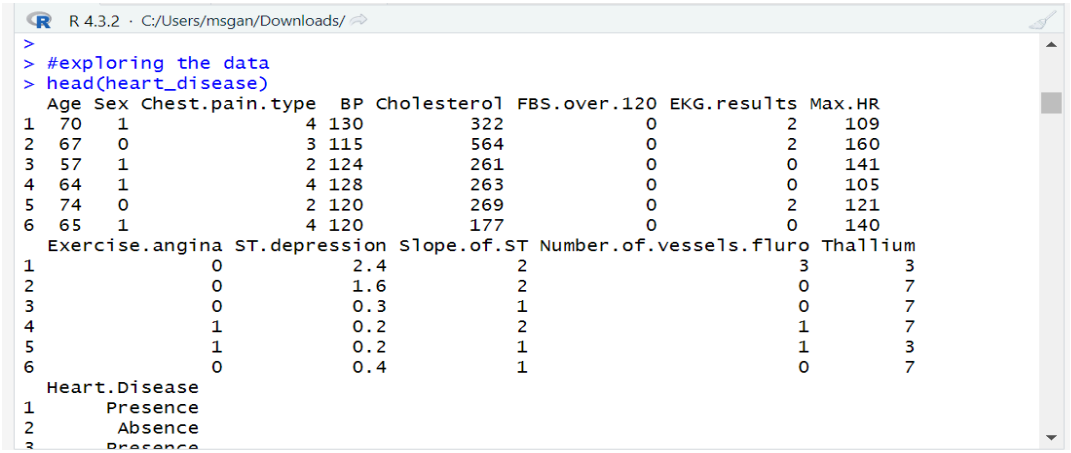
**Some metrics I have used in Quality control of my data set are**

- Handling missing values using the function colSums(is.na(heart_disease)) and this step is essential to make the model ready for testing and analysis.
- Identifying and removing outliers in my variables selected ie Age,BP and Cholesterol and created box plots using the function boxplot.stats() and outliers using (heart_disease_no_age_outliers, heart_disease_no_cholesterol_outliers, and heart_disease_no_bp_outliers) and removing outliers is essential to improve the reliability of the statistical testing and machine learning model.
- Data conversion - I have used the as.factor() function to change the type of my target variable as it is categorical.
- I have also used caret package to perform machine learning models and train and test the data as well to provide more accurate results and also to check the cross validation.
- Quality control is a basic measure and is generally done to obtain robust and meaningful results.

## Descriptive Analysis

In my code of heart disease prediction I have chosen 3 main variables: Age, Cholesterol and Blood Pressure.

- Loading of the dataset and exploration of the data



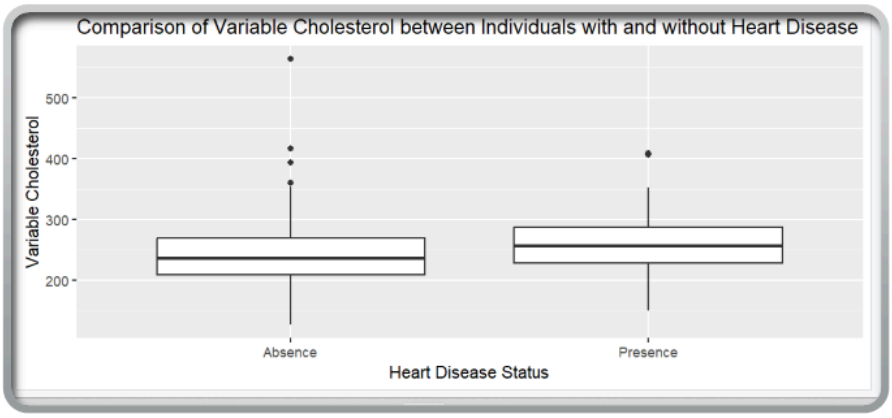
```
R 4.3.2 · C:/Users/msgan/Downloads/

> 
> #exploring the data
> head(heart_disease)
  Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR
1  70   1               4 130         322            0           2    109
2  67   0               3 115         564            0           2    160
3  57   1               2 124         261            0           0    141
4  64   1               4 128         263            0           0    105
5  74   0               2 120         269            0           2    121
6  65   1               4 120         177            0           0    140
  Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro Thallium
1               0           2.4           2                      3        3
2               0           1.6           2                      0        7
3               0           0.3           1                      0        7
4               1           0.2           2                      1        7
5               1           0.2           1                      1        3
6               0           0.4           1                      0        7
  Heart.Disease
1      Presence
2       Absence
3      Presence
```

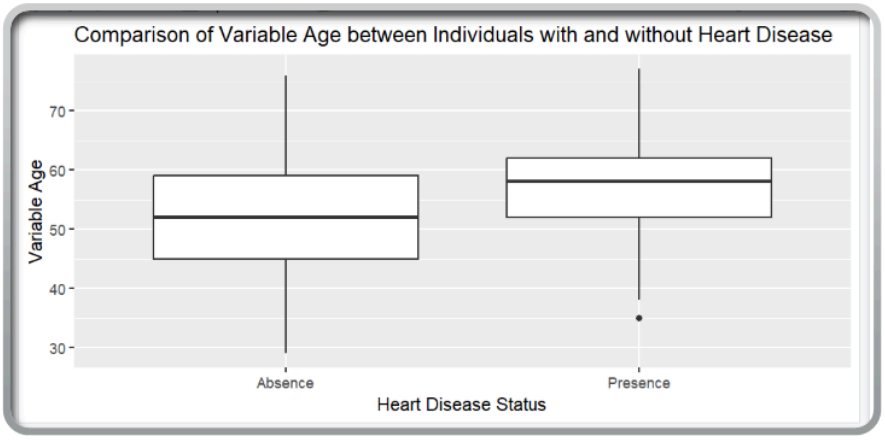
The dataset has been explored and it shows 270 observations and 14 variables

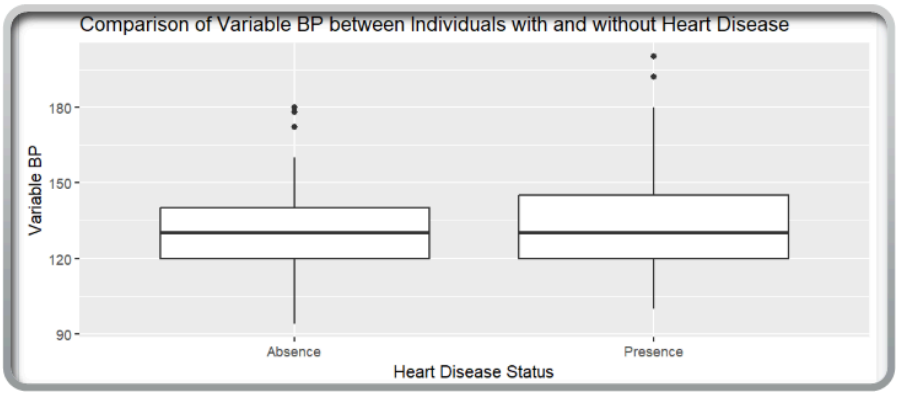
- Then I ran the code to check missing values and my dataset does not have any missing values implying that I don't have to have missing data imputation.

```
>
>
> #checking for missing values
> colSums(is.na(heart_disease))
                    Age                     Sex          Chest.pain.type
                      0                       0                        0
                     BP             Cholesterol             FBS.over.120
                      0                       0                        0
             EKG.results                 Max.HR          Exercise.angina
                      0                       0                        0
          ST.depression            Slope.of.ST Number.of.vessels.fluro
                      0                       0                        0
               Thallium           Heart.Disease
                      0                       0
>
```

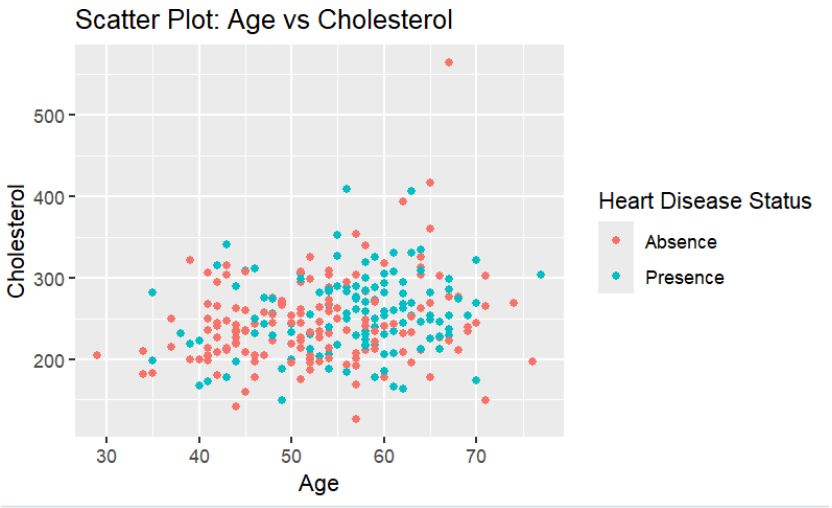
- I have then performed exploratory data analysis and created box plots, histograms and scatter plots using my variables age, bp and cholesterol.
- I have used ggplot and made 3 box plots using the variables Age, cholesterol and BP

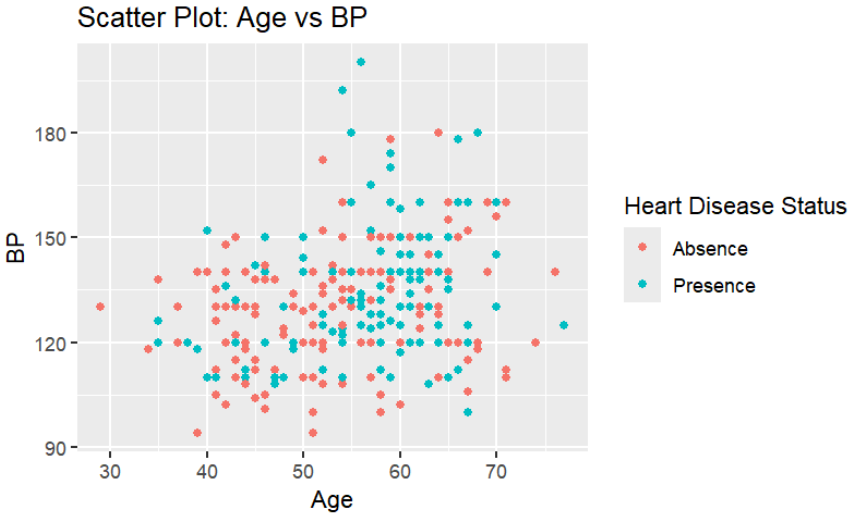


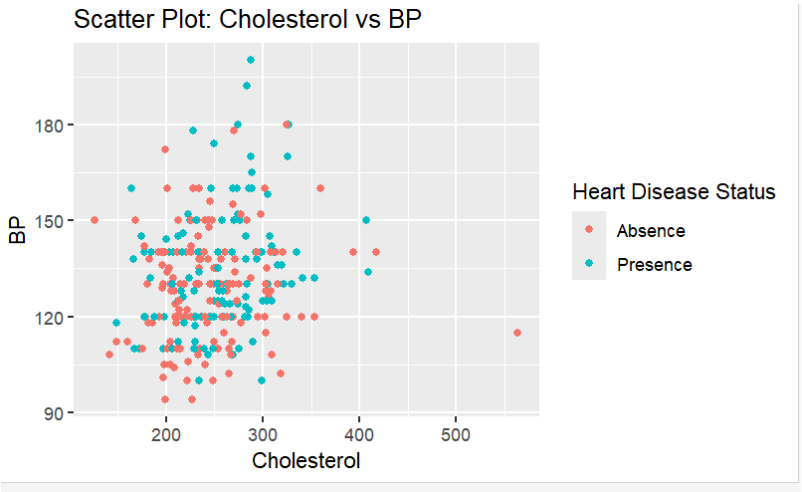


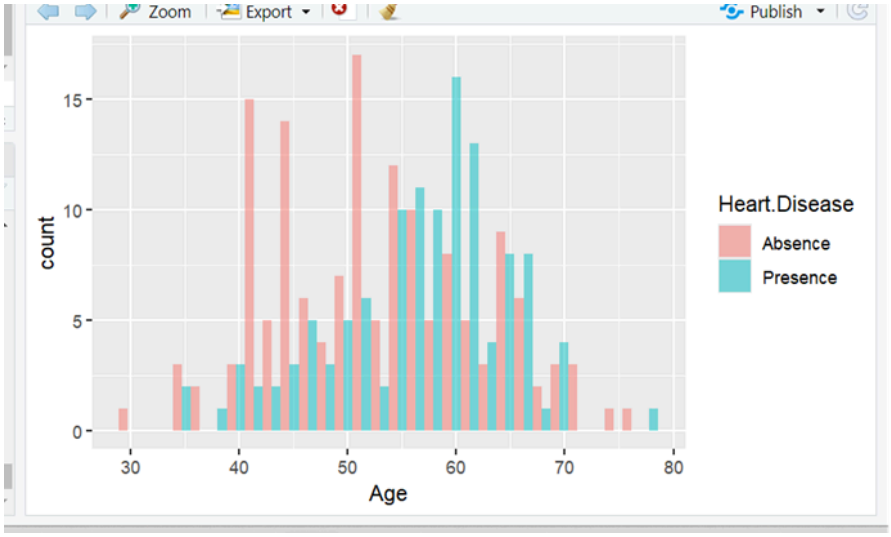

- Through the box plots we can observe that there is significant difference in presence and absence of heart disease associated with cholesterol and age but not a significant difference in BP.
- I have also created 3 scatter plots to show the visualization

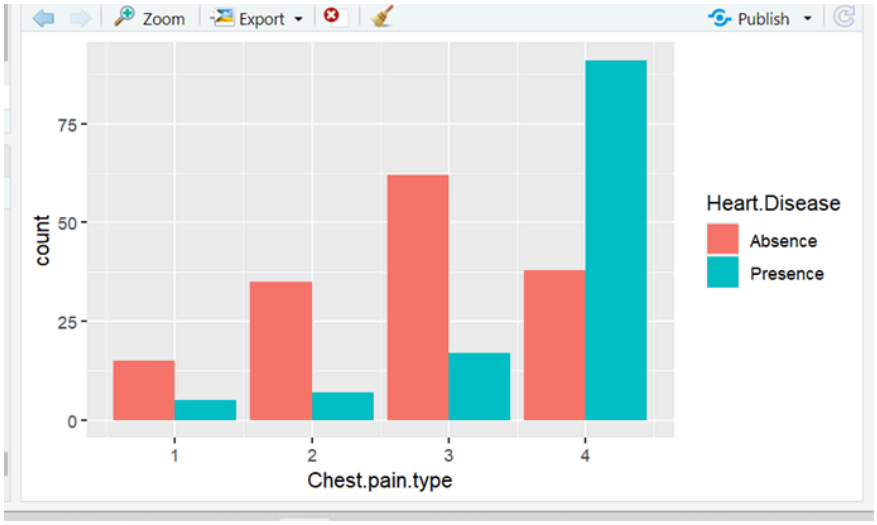

Scatter Plot: Age vs Cholesterol
Cholesterol
500
400
300
200
30
40
50
60
70
Age
Heart Disease Status
Absence
Presence


Scatter Plot: Age vs BP
BP
180
150
120
90
30
40
50
60
70
Age
Heart Disease Status
Absence
Presence


Scatter Plot: Cholesterol vs BP
BP
180
150
120
90
200
300
400
500
Cholesterol
Heart Disease Status
Absence
Presence

- I have additionally visualized histograms and done exploratory data analysis to visualize the distribution of variables



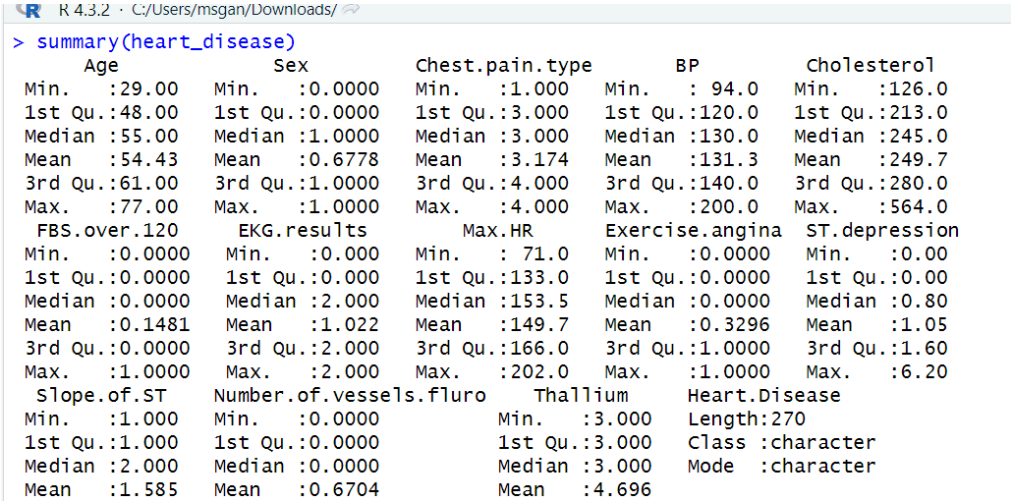

- It is clear from the plot that certain types of chest pain are more common in one group versus the other.

Outlier detection and removal -

- The code has outliers for age, cholesterol and bp and those have been identified and removed to make the model perform better.
- After removal of outliers I have combined the dataset without outliers.

Overall the descriptive statistics helps the model to quantify the variables and helps in subsequent steps to perform machine learning and help identify the key factors associated with heart disease.

- In measures of central tendency I found out the mean, median, min, maximum values of all variables.



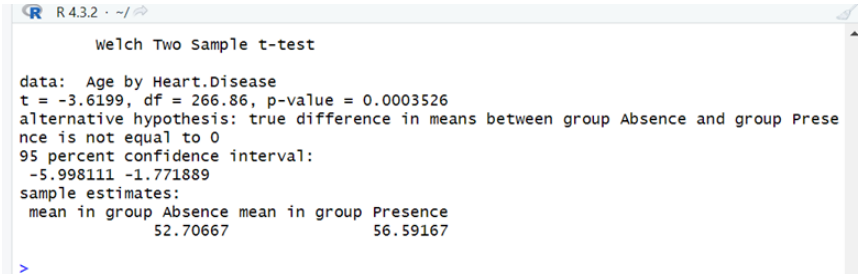
```
R 4.3.2 · C:/Users/msgan/Downloads/
> summary(heart_disease)
      Age             Sex         Chest.pain.type       BP          Cholesterol
 Min.   :29.00   Min.   :0.0000   Min.   :1.000   Min.   : 94.0   Min.   :126.0
 1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:120.0   1st Qu.:213.0
 Median :55.00   Median :1.0000   Median :3.000   Median :130.0   Median :245.0
 Mean   :54.43   Mean   :0.6778   Mean   :3.174   Mean   :131.3   Mean   :249.7
 3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:140.0   3rd Qu.:280.0
 Max.   :77.00   Max.   :1.0000   Max.   :4.000   Max.   :200.0   Max.   :564.0
  FBS.over.120     EKG.results        Max.HR      Exercise.angina  ST.depression
 Min.   :0.0000   Min.   :0.000   Min.   : 71.0   Min.   :0.0000   Min.   :0.00
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:133.0   1st Qu.:0.0000   1st Qu.:0.00
 Median :0.0000   Median :2.000   Median :153.5   Median :0.0000   Median :0.80
 Mean   :0.1481   Mean   :1.022   Mean   :149.7   Mean   :0.3296   Mean   :1.05
 3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60
 Max.   :1.0000   Max.   :2.000   Max.   :202.0   Max.   :1.0000   Max.   :6.20
  Slope.of.ST    Number.of.vessels.fluro    Thallium     Heart.Disease
 Min.   :1.000   Min.   :0.0000          Min.   :3.000   Length:270
 1st Qu.:1.000   1st Qu.:0.0000          1st Qu.:3.000   Class :character
 Median :2.000   Median :0.0000          Median :3.000   Mode  :character
 Mean   :1.585   Mean   :0.6704          Mean   :4.696
```

## Data Analysis - Machine Learning

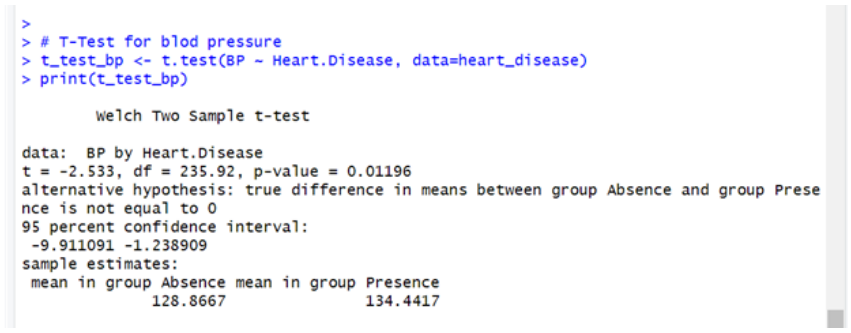
After following the previous steps of Data Exploration, visualization and outlier removal I have run the code for some machine learning algorithms.

T tests

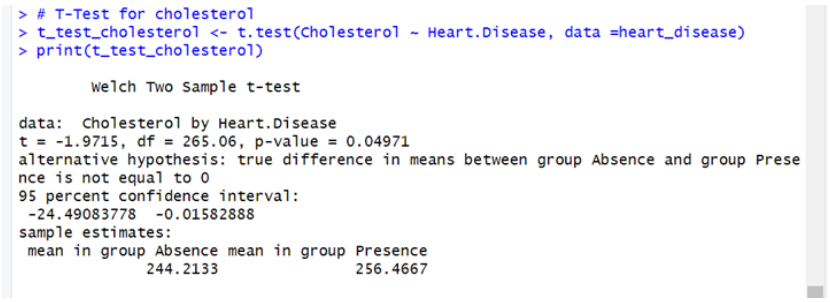
- I have performed T- tests with 3 variables for age , cholesterol and BP.



```
R 4.3.2 · ~/

        Welch Two Sample t-test

data:  Age by Heart.Disease
t = -3.6199, df = 266.86, p-value = 0.0003526
alternative hypothesis: true difference in means between group Absence and group Prese
nce is not equal to 0
95 percent confidence interval:
 -5.998111 -1.771889
sample estimates:
 mean in group Absence mean in group Presence
              52.70667               56.59167

>
```



```
>
> # T-Test for blod pressure
> t_test_bp <- t.test(BP ~ Heart.Disease, data=heart_disease)
> print(t_test_bp)

        Welch Two Sample t-test

data:  BP by Heart.Disease
t = -2.533, df = 235.92, p-value = 0.01196
alternative hypothesis: true difference in means between group Absence and group Prese
nce is not equal to 0
95 percent confidence interval:
 -9.911091 -1.238909
sample estimates:
 mean in group Absence mean in group Presence
              128.8667               134.4417
```



```
> # T-Test for cholesterol
> t_test_cholesterol <- t.test(Cholesterol ~ Heart.Disease, data =heart_disease)
> print(t_test_cholesterol)

        Welch Two Sample t-test

data:  Cholesterol by Heart.Disease
t = -1.9715, df = 265.06, p-value = 0.04971
alternative hypothesis: true difference in means between group Absence and group Prese
nce is not equal to 0
95 percent confidence interval:
 -24.49083778  -0.01582888
sample estimates:
 mean in group Absence mean in group Presence
              244.2133               256.4667
```

Age -

The test shows statistical significance in the mean ages between individuals with and without heart disease and on an average the individuals with detected heart disease were over 56 years than those without heart disease of 52 years (mean age).

Cholesterol level -

There is a statistically significant difference in the mean cholesterol levels but not as significant as age as the p - value is close to 0.05 and the individuals with heart disease have higher mean cholesterol levels of Approx 256.47 compared to those without heart disease which is 244.21.

Blood Pressure -

The difference in the mean blood pressure between both the variables is also statistically significant although it was not clearly indicated in the box plot and the individuals with heart disease have shown to have higher BP ove avg 133.44mmHg than those without heart disease of 128.87mmHg.

ANOVA test

In addition to the T - test I have also performed ANOVA test as my data set has more than two types of chest pain.

- I have done the ANOVA test to check if there are any significant differences in the mean cholesterol level across different types of chest pain.
- The p- value is 0.138 which exceeds the normal required value of 0.05 used to determine the statistical significance.
- As the P - value is greater than 0.05 we do not have enough evidence to reject null hypothesis.
- We cannot confidently say that different types of chest pain have different average cholesterol levels and we can say that the type of chest pain has no evidence or association to its statistical significance.

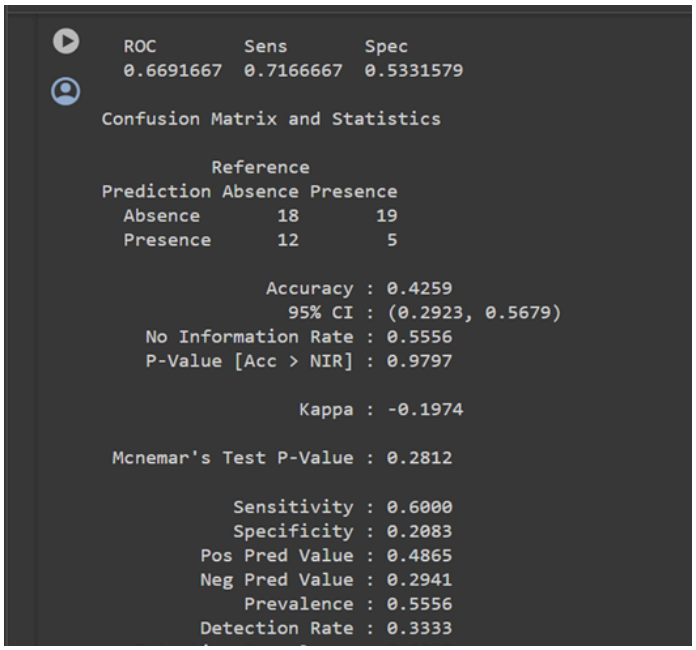
```
> # Anova for more than two-groups, as we have more than two types of chest pain
> anova_result <- aov(Cholesterol ~ Chest.pain.type, data=heart_disease)
> summary(anova_result)
                Df Sum Sq Mean Sq F value Pr(>F)
Chest.pain.type   1   5881    5881   2.211  0.138
Residuals       268 712743    2659
>
```

- I have used the caret package to run the machine learning model and I have chosen logistic regression.
- For the machine learning model development I have split the code into training and testing sets using the data partition function and set up a training control function for cross validation method and the number of folds.
- The code trains the logistic regression model using the training and testing sets with age , BP and cholesterol and heart disease as the target variable.
- I have also done evaluation metrics like accuracy, sensitivity , specificity , ROC , kapaa and Mcnemar's test p value and confusion matrix.
- Building a machine learning model is the second main step in this project and logistic regression and the other evaluation metrics will provide insights in showing the accuracy and reliability of the model in predicting heart disease.
- By the use of exploratory data analysis, hypothesis testing and the use of machine learning my code can provide a comprehensive approach to factors associated with risk of heart disease.

## **Results**

The overall results i observed after training and testing the datasets are

- The model I chose here is Logistic regression
- I have used the Caret package to run my machine learning model where I split my data into training and testing sets.
- I have also done cross validation with 5 - folds
- I have also done ROC and Confusion matrix



```
    ROC        Sens       Spec
  0.6691667  0.7166667  0.5331579

Confusion Matrix and Statistics

          Reference
Prediction Absence Presence
  Absence       18       19
  Presence      12        5

               Accuracy : 0.4259
                 95% CI : (0.2923, 0.5679)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 0.9797

                  Kappa : -0.1974

 Mcnemar's Test P-Value : 0.2812

            Sensitivity : 0.6000
            Specificity : 0.2083
         Pos Pred Value : 0.4865
         Neg Pred Value : 0.2941
             Prevalence : 0.5556
         Detection Rate : 0.3333
```

- The ROC AUC score is 0.669,suggesting the model has moderate ability to differentiate between the presence and absence of heart disease.
- The model has a sensitivity(TPR) of approximately 0.717,meaning it correctly identifies 71.7% of actual heart disease cases.
- The specificity(TNR) is relatively low at 0.533, indicating that it correctly identifies 53.3% of the actual non disease cases.
- The accuracy of the model is 0.4259, which is below the No Information Rate(NIR) of 0.5556.The NIR is the largest class percentage in the data, and an accuracy less than this rate suggests the model is not performing well.
- The p-value for accuracy is greater than the NIR is 0.9797,which is not significant,indicating that the model is not performing better than a no-information classifier that would always predict the most frequent class.

- The kappa stat is-0.1974,which is negative,signifying that the model is performing worse than chance.
- Mcnemar's test has a p-value of 0.2812 -there is no significant bias in the model's error rates between the two classes.
- PPV or precision is 0.4865, indicating that when the model predicts heart disease presence,it is correct about 48.65% of the time.
- NPV is 0.2941, indicating that when the model predicts heart disease absence,it is correct about 29.41% of the time.
- The model's prevalence,the actual rate of heart disease presence in the sample,is 55.56%.
- Detection rate -the rate of true positives, is 33.33%
- The detection prevalence-total number of positive predictions made by the model - 68.52%.
- The Balanced Accuracy-the average of sensitivity and specificity- is 40.42%, indicating poor performance across both classes.
- The 'Positive' class is labeled as 'Absence,' which can be confusing. usually,'Positive' refers to the presence of a condition.

**Limitation -**

- Limited feature set – The performance and accuracy of the data would perform much better with different variables than the one I chose here. So exploring other variables may give different results.
- Data quality – This dataset has inconsistencies so that can be improved
- Imbalanced data

- Other complex machine learning models like random forest , decision tree , SVM can be used between other variables to achieve more effective results.
- Investigation of  class imbalance in the target variable and apply techniques like oversampling, undersampling, or class weighting to mitigate the imbalance and improve model performance.
- Model tuning.

**Conclusion -**

- The analysis presented in this project aims to identify risk factors associated with heart disease and develop predictive models for risk assessment.
- The exploratory data analysis and statistical hypothesis testing conducted in this project can help identify significant differences in variables such as age, cholesterol levels, blood pressure, and chest pain types between individuals with and without heart disease.
- Statistical hypothesis testing using t-tests confirmed significant differences in the means of these variables across the two groups. Additionally, one-way ANOVA indicated significant variations in cholesterol levels among different chest pain types.
- To predict heart disease presence, a logistic regression model was trained on age, cholesterol, and blood pressure. The model's performance was further evaluated using metrics such as sensitivity, specificity derived from the confusion matrix.
- The Area Under the Curve (AUC) demonstrated the model's ability to discriminate between individuals with and without heart disease.
- While the analysis provided valuable insights and a predictive model, there are potential limitations and areas for improvement. Incorporating additional relevant features, handling data quality issues, addressing class imbalance, exploring more complex

models, and conducting external validation could enhance the model's predictive power and generalizability.

- Overall, this study contributes to the understanding of heart disease risk factors and the development of predictive models, which can aid in early identification and preventive strategies for cardiovascular health.

**REFERENCES** :


- Greenland, P., Knoll, M. D., Stamler, J., Neaton, J. D., Dyer, A., Garside, D. B., & Wilson, P. W. (2003). Major risk factors as antecedents of fatal and nonfatal coronary heart disease events. *JAMA*, *290*(7), 891. https://doi.org/10.1001/jama.290.7.891
- Khot, U. N., Khot, M. B., Bajzer, C., Sapp, S., Ohman, E. M., Brener, S. J., Ellis, S. G., Lincoff, A. M., & Topol, E. J. (2003). Prevalence of conventional risk factors in patients with coronary heart disease. *JAMA*, *290*(7), 898. https://doi.org/10.1001/jama.290.7.898
- Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q. V., Litsch, K., . . . Dean, J. M. (2018). Scalable and accurate deep learning with electronic health records. *Npj Digital Medicine*, *1*(1). https://doi.org/10.1038/s41746-018-0029-1
- Rana, J. S., Mukamal, K. J., Morgan, J. P., Muller, J. E., & Mittleman, M. A. (2004). Obesity and the risk of death after acute myocardial infarction. *American Heart Journal˜the ‰American Heart Journal*, *147*(5), 841–846. https://doi.org/10.1016/j.ahj.2003.12.015
- Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, *25*(10), 1419–1428. https://doi.org/10.1093/jamia/ocy068
- Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, Á., Lanas, F., McQueen, M., Budaj, A., Пайс, П., Varigos, J., & Liu, L. (2004). Effect of potentially modifiable risk

factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet*, *364*(9438), 937–952. https://doi.org/10.1016/s0140-6736(04)17018-9

- Zheng, Y., Liu, Q., Chen, E., Ge, Y., & Zhao, J. (2014). Time series classification using Multi-Channels Deep Convolutional neural networks. In *Lecture notes in computer science* (pp. 298–310). https://doi.org/10.1007/978-3-319-08010-9_33
- Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2017). Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association, 24(2), 361-370. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5391725/