# INTRODUCTION TO DATA ANALYTICS PROJECT REPORT

## INSTRUCTOR: DR. SREEJA SR

## TOPIC - 2

DONE BY:

GROUP ID: -G14

| NAME | ROLL NUMBER |
|---|---|
| MOHAN MOPADA | S20200010134 |
| YASHASWI | S20200010073 |
| SRI NITYA | S20200010143 |
| SIDDU PUTCHALA | S20200010173 |
| SAKETH CHAMALLA | S20200010046 |

# INDEX

Definitions

- **OUTLIERS:**

Outlier Analysis is a process that involves identifying the anomalous observation in the dataset. Outliers are nothing but an extreme value that deviates from the other observations in the dataset.

*IQR = Q3 – Q1*

*Outliers range = <=Q1 – 1.5*IQR, >= Q3 + 1.5*IQR*

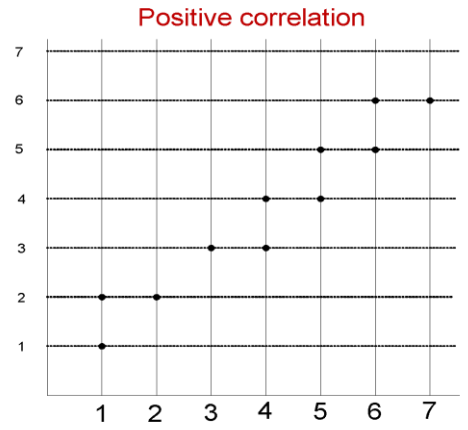
- CORRELATION:

- In statistics, the word correlation is used to denote some form of association between two variables.

  ` Ex-Weight is correlated with height.

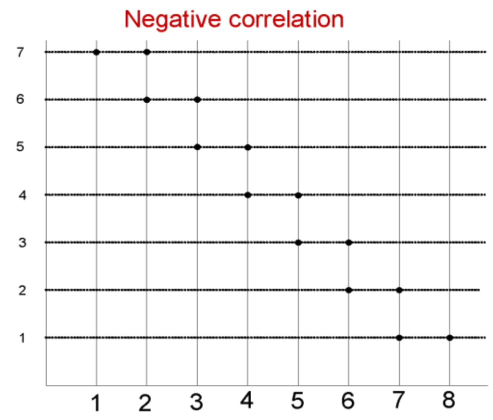
- The correlation may be positive, negative, or zero.
- **Positive correlation:** If the value of attribute A increases with the increase in the value of attribute B and vice-versa.

  Ex- Relation between rate of change of velocity and acceleration.



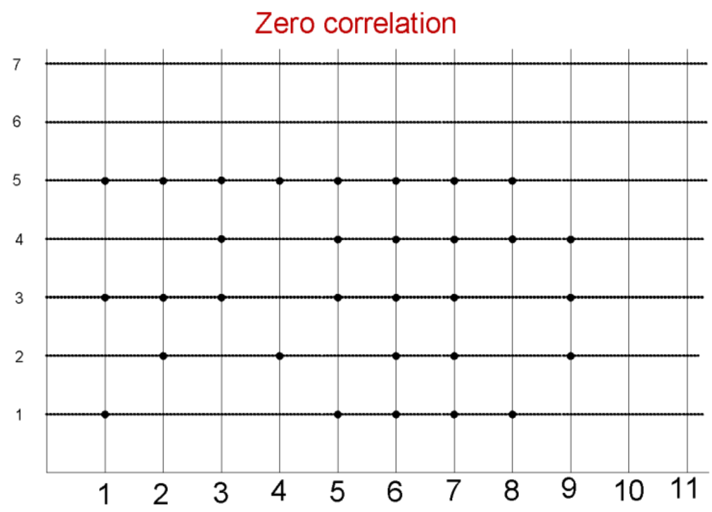

- **Negative correlation:** If the value of attribute A decreases with the increase in the value of attribute B and vice-versa.

  Ex-Relation between the speed of the train and time taken to reach the destination.



Definitions

- **Zero correlation:** When the values of attribute A varies at random with B and vice-versa.
  Ex- Relationship between the amount of tea drunk and level of intelligence.




**Correlation Coefficient**

- Correlation coefficient is used to measure the degree of association.
- It is usually denoted by r.
- The value of r lies between +1 and -1.
- Positive values of r indicate positive correlation between two variables, whereas, negative values of r indicate negative correlation.
- r = +1 implies perfect positive correlation, and otherwise.
- The value of r nearer to +1 or -1 indicates a high degree of correlation between the two variables.
- r = 0 implies, there is no correlation.
- There are three methods known to measure the correlation coefficients
  - Karl Pearson's coefficient of correlation-
    This method is applicable to find correlation coefficient between two numerical attributes
    r* > 0 Positively correlated | r* < 0 Negatively correlated
    r* >= 0.7 & r* <= -0.7 are said to b highly correlated.
  - Spearman correlation: Spearman correlation evaluates the monotonic relationship. The Spearman correlation coefficient is based on the ranked values for each variable rather than the raw data.
    A ρ value of +1 means a perfect association of rank.
    A ρ value of 0 means no association of ranks
    A ρ value of -1 means a perfect negative association between ranks.

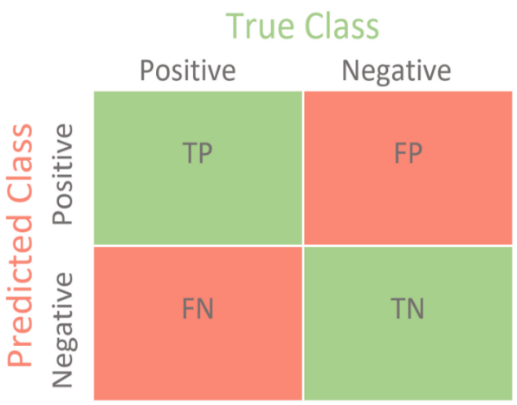
Definitions

- CHI-SQUARED Test of correlation: This method is applicable to categorical data.
  Null hypothesis: Two values are independent
  Significance level: 0.05
  If the obtained ‘p’ value is less than significance level then we reject our null hypothesis and conclude that there is a relationship between two attributes else there is no relationship between two attributes.

- **Apriori**:
  P(A), P(B)

- **Posterior probability**:
  P(A|B)

- **Conditional probability**:
  If events are dependent then their probability is expressed by conditional probability. The probability that A occurs given that B is denoted by P(A|B) = P(B|A)*P(A)/P(B) i.e., P(A ∩ B)/P(B)
- **Confusion matrix:**




- Accuracy = (TP+TN)/(TP+TN+FP+FN)
- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN)
- F1-score = 2*precision*Recall/(Precision + Recall)

# 2.Classification

Various approaches to Solve the classification problems are:

1. Decision trees
2. SVM
3. Naïve Bayes Classifier
4. Logistic Regression
   and so on...

In the problem statement it is mentioned to use GNB Classifier.

**GNB Classifier**:

Naive Bayes Classifiers are based on the Bayes Theorem. One assumption taken is the strong independence assumptions between the features. These classifiers assume that the value of a particular feature is independent of the value of any other feature. In a supervised learning situation, Naive Bayes Classifiers are trained very efficiently. Naive Bayed classifiers need a small training data to estimate the parameters needed for classification. Naive Bayes Classifiers have simple design and implementation and they can apply to many real-life situations.

When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be-

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

All the features in the dataset are not completely categorical. The feature set contains both numerical as well as categorical. So, it is not the accurate approach to implement the gaussian naïve bayes classifier for the entire data.

Alternative methods which we can use when our data set consists of both categorical and numerical features are:

Mixed NB(Gaussian + Categorical)

***Approach 1:***

We will convert the continuous variables into categorical ones through binning. Then we will train a categorical model on all of those features.

***Approach 2:***

We need to train two separate models using continuous and categorical independent variables. Then we will take prediction probabilities from these two models and use them for training the final model.

But in problem it's explicitly mentioned that we need to GNB Classifier. So, we are using the GNB Classifier for the Classification.

# 3.PROBLEM STATEMENT

# Heart Attack Prediction using the GNB Classifier in R

# 4.ABOUT THE DATA SET

Heart Attack Analysis & Prediction Dataset

The attributes/features in the dataset are:

**Age** – *Age of the Patient* – Numerical

**Sex** – *Sex of the Patient* - categorical

**Cp** – *Chest pain type* – categorical

*Value 1: Typical Angina*

*Value 2: Atypical angina*

*Value 3: Non-Anginal Pain*

*Value 4: Asymptomatic*

**Trtbps** – *Resting Blood Pressure* - Numerical

**Chol** – *Cholesterol in mg/dl* - Numerical

**Fbs** – *Fasting Blood Sugar* - Categorical

**Restecg** – *Resting Electrocardiographic results* - Categorical

**Thalachh** – *Maximum heart rate achieved* - Numerical

**Exng** – *Exercise Induces Angina* - Categorical

**Oldpeak** – *Previous Peak* - Numerical



**Slp**- *The slope of the peak exercise ST segment* – Categorical

*Value 1: upsloping*

*Value 2: flat*

*Value 3: downsloping*

**Caa** – Number of Major Vessels (0-3) coloured by fluoroscopy Categorical

**Thall** – 3=>Normal , 6=>fixed defect , 7=>Reversable Defect - Categorical

**Output** – *diagnosis of heart disease* – Label (Categorical)

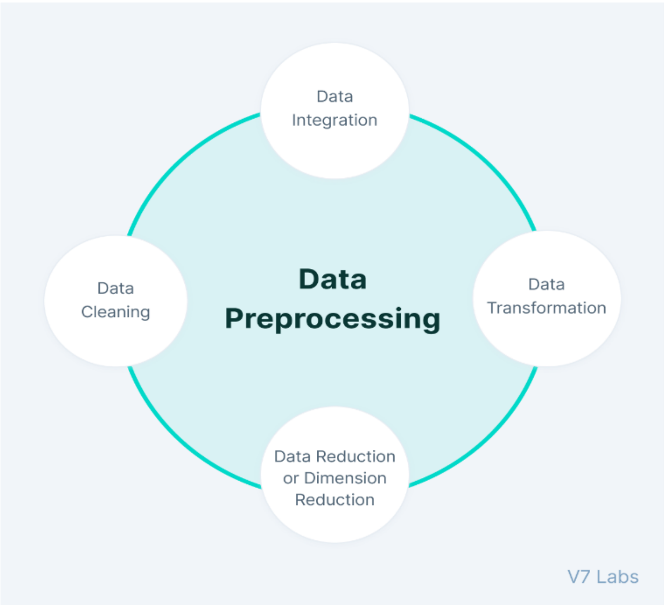
*Value 0: < 50% diameter narrowing*

*Value 1: >50% diameter narrowing*

**Total Number of rows in the data set – 303**

# 5.DATA PREPROCESSING

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.




### 5.1 DATA CLEANING

The data can have many irrelevant and missing parts which might effect the performance of our model in a negative manner. In order to avoid this we follow a some techniques.

Data pre-processing / Data cleaning

**5.1.1**Missing Data: Remove or replace the missing values with an appropriate measure of central tendency.

**5.1.2**In our data set there were no missing values present.

**5.1.3** There is a redundant row(duplicate) in the 165$^{th}$ column and it is being removed.

**5.1.4**Outliers:

Outliers in the categorical data can also be said to the problem of class imbalance. This means that the data is not in similar proportion.
But we should not remove the outlier without knowing the importance of class.
Outliers present in the numerical attributes:

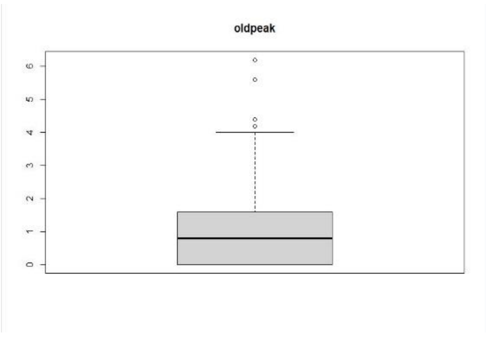
Oldpeak => 4 Outliers

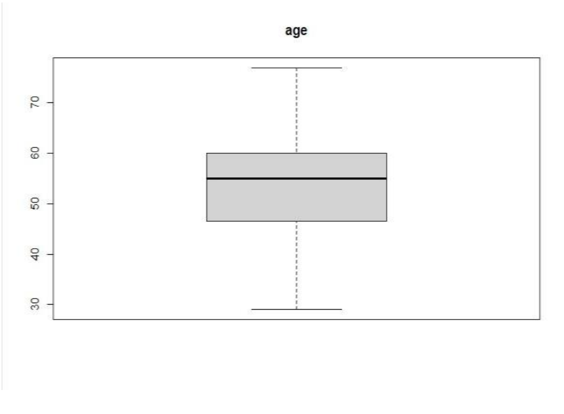
Thalachh => 1 Outlier

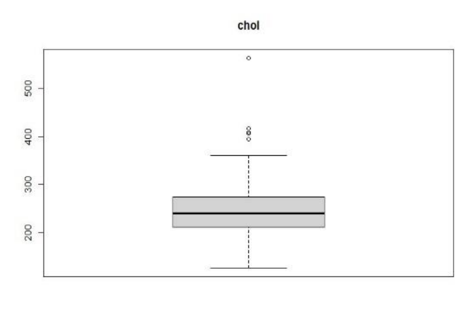
Chol => 3 Outliers

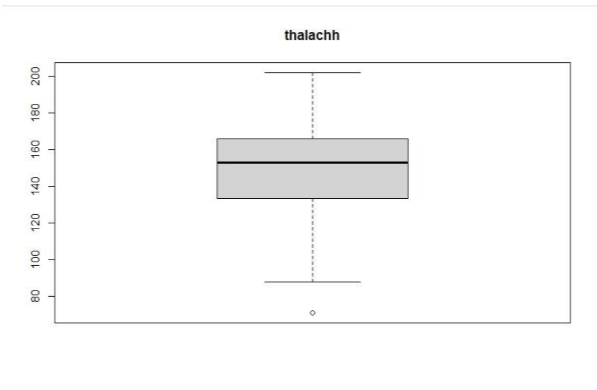
Trtbps => 6 Outliers

Age => No Outliers



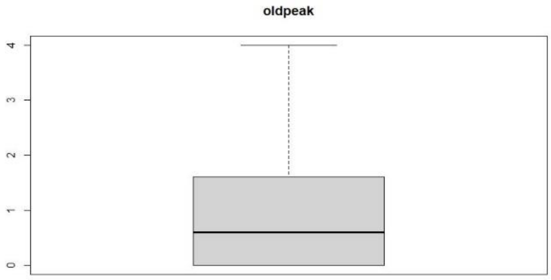



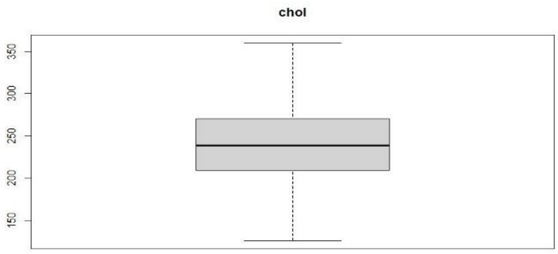



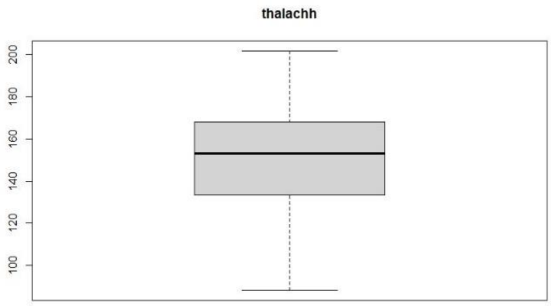



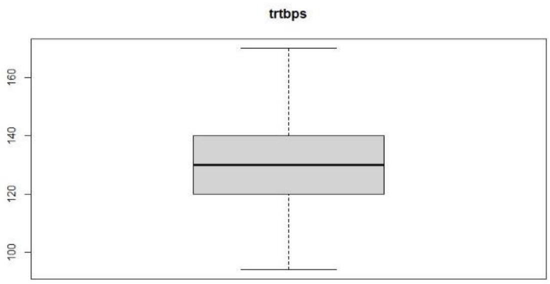
Data pre-processing / Data cleaning

**5.1.5** Removing outliers




## 5.2 DATA INTEGRATION/REDUCTION

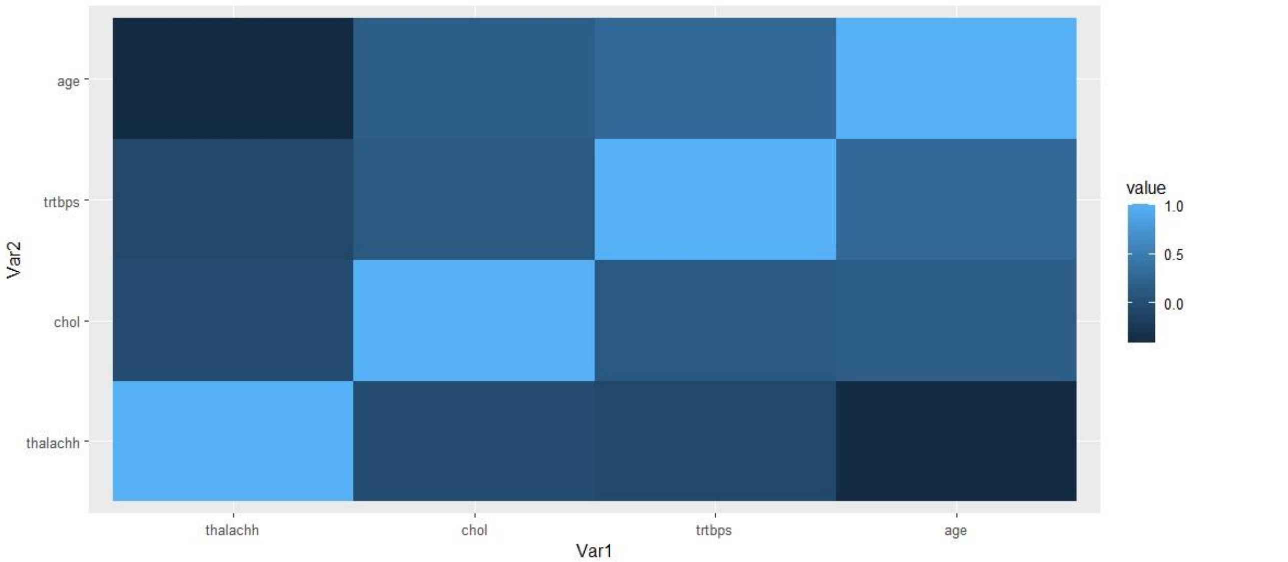
*Data reduction* is used to reduce the amount of data and thereby reduce the costs associated with data analysis.

**5.2.1** Correlation analysis of numerical data

- Continuous features are: oldpeak, thalachh, chol, trtbps, age
- Among them oldpeak doesn't follow the normal distribution.
- So, we find the Pearson's correlation between the features { thalachh, chol, trtbps, age }
- Now use Spearman Correlation to find the correlation with oldpeak to remaining continuous features.

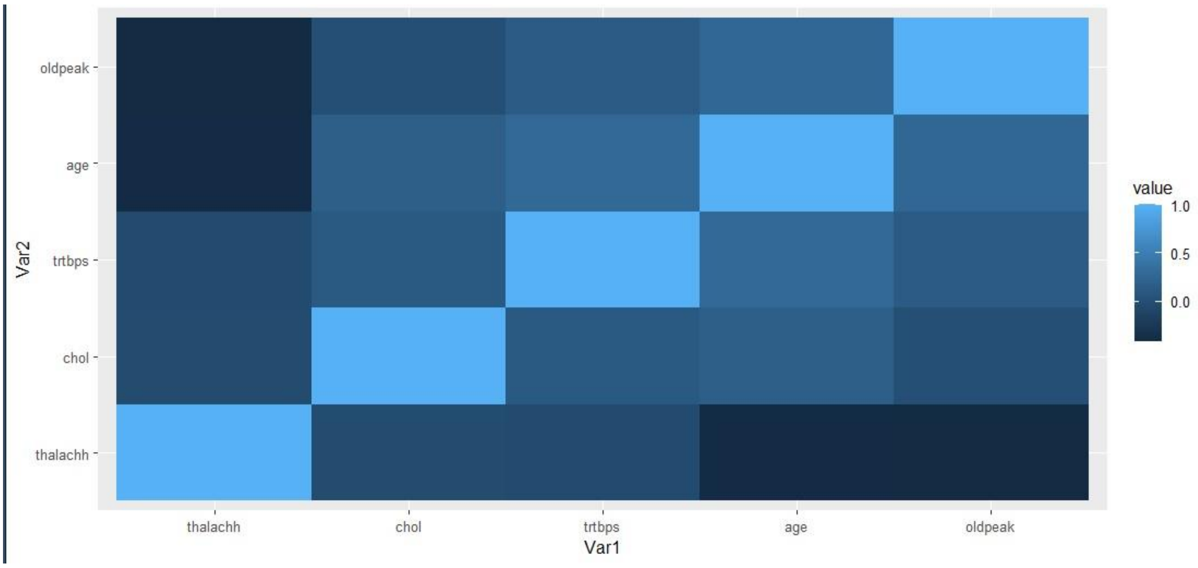
Data pre-processing / Data reduction | Data transformation

**Pearson**




We can clearly observe None of the features are highly correlated

**Spearmen's**




We can clearly observe None of the features are highly correlated

**5.2.2** Correlation analysis of categorical data

**Chi-squared**

- sex vs cp -> p = 0.02597 *(<0.05 => Correlated)*
- sex vs fbs -> p = 0.3572 *(>0.05 => Not Correlated)*
- sex vs restecg -> p = 0.06513 *(>0.05 => Not Correlated)*
- sex vs exng -> p = 0.003654 *(<0.05 => Correlated)*
- sex vs slp -> p = 0.5392 *(>0.05 => Not Correlated)*
- sex vs caa -> p = 0.06259 *(>0.05 => Not Correlated)*
- sex vs thall -> p = 1.901E -10 *(<0.05 => Correlated)*

- restecg vs fbs -> p = 0.4019 *(>0.05 => Not Correlated)*
- fbs vs slp -> p = 0.1269 *(>0.05 => Not Correlated)*
- fbs vs caa -> p = 0.1015 *(>0.05 => Not Correlated)*
- restecg vs slp -> p = 0.08721 *(>0.05 => Not Correlated)*
- restecg vs caa -> p = 0.4868 *(>0.05 => Not Correlated)*
- slp vs caa -> 0.1237 *(>0.05 => Not Correlated)*

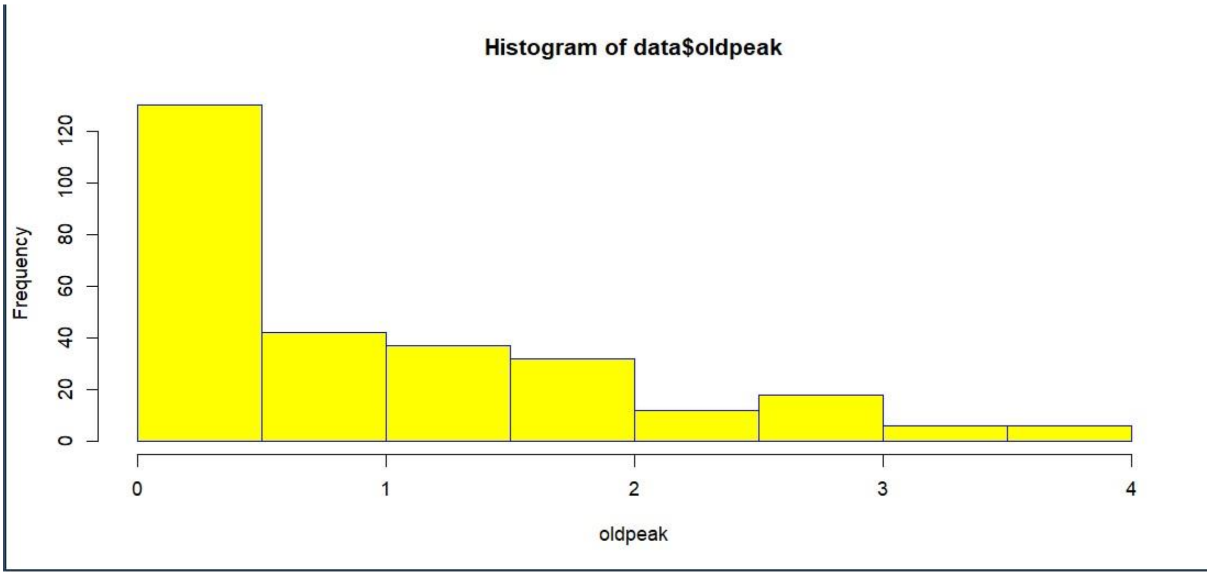
  The features {***sex, cp***}, {***sex, exng***}, {***sex,thall***} are correlated.
  The features {***sex, exng***}, {***sex,thall***} are highly correlated so they are redundant attributes.
  We can **remove** the features *exng* and *thall* from the dataset.

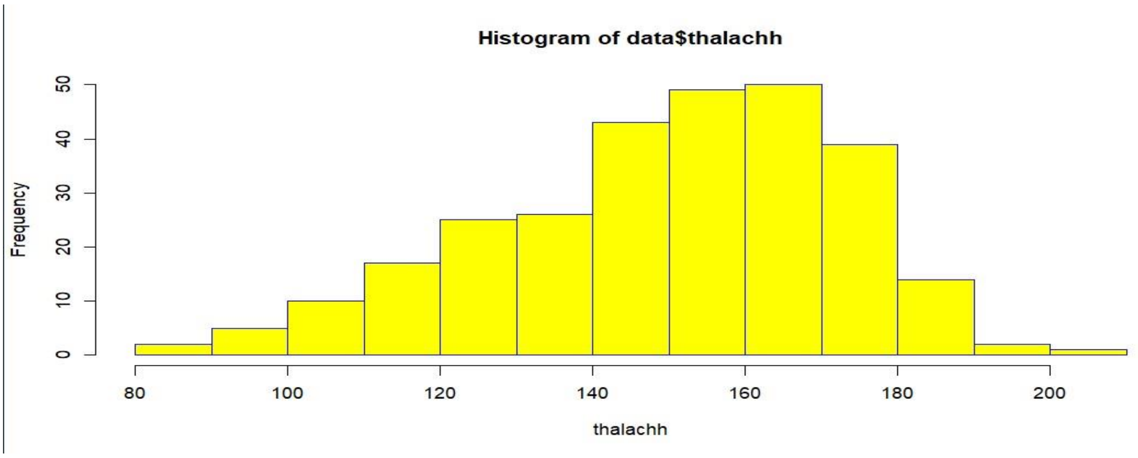
**5.3DATA TRANSFORMATION**:

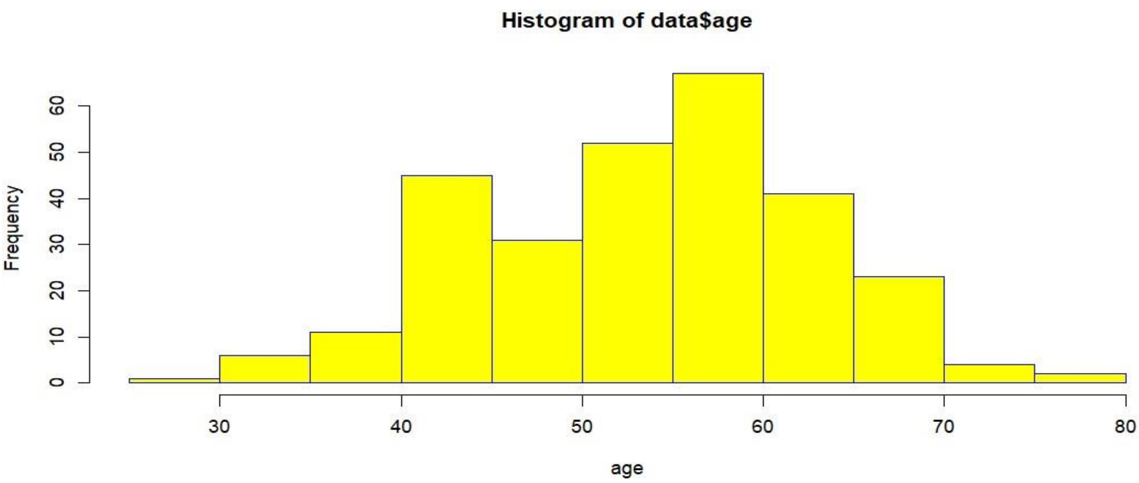
As Naïve Bayes Algorithm is based on probability not on distance. So, it doesn't require feature scaling

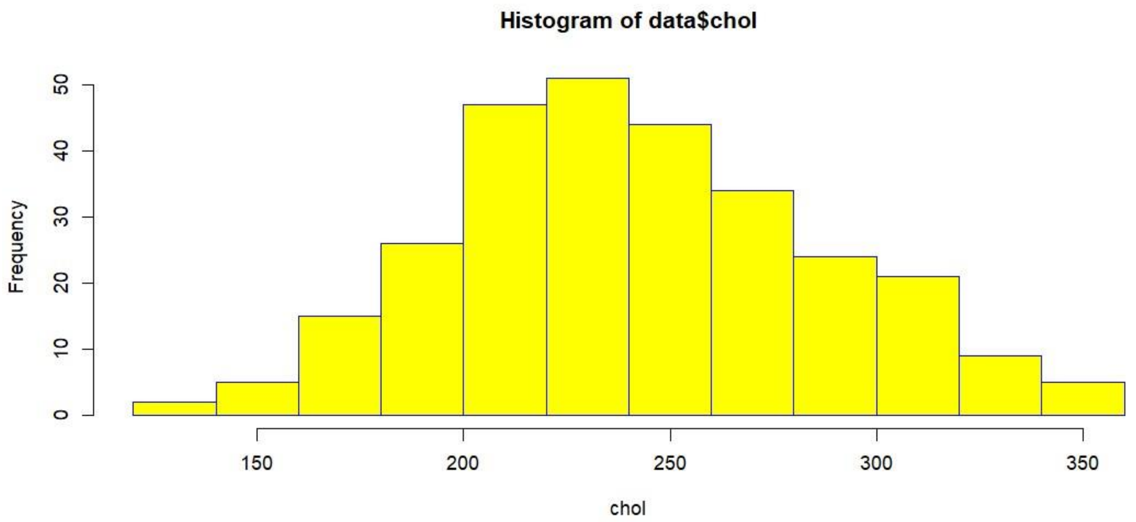
# 6.DATA VISUALIZATION(Histograms)

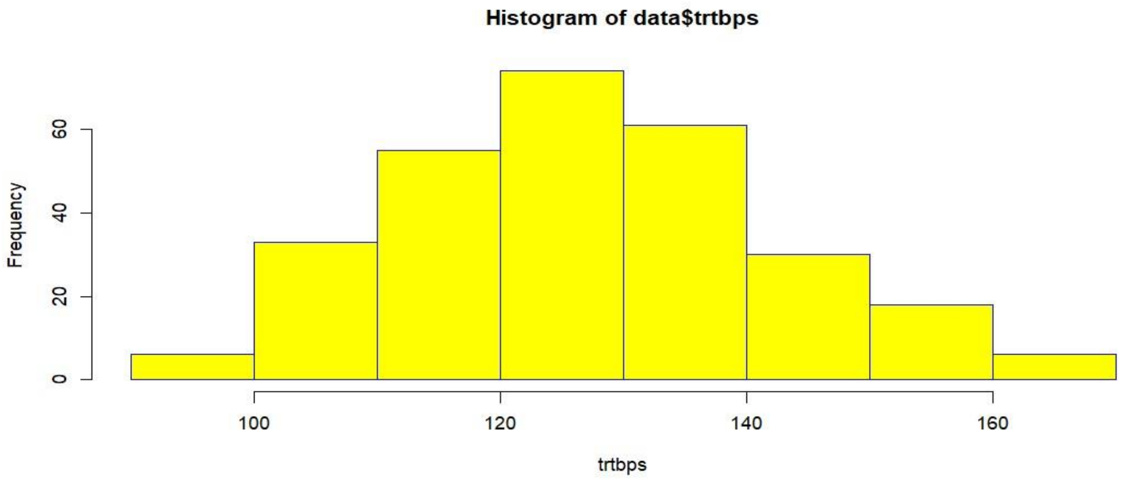
**Numerical Features (**Checking the distribution of the data**)**



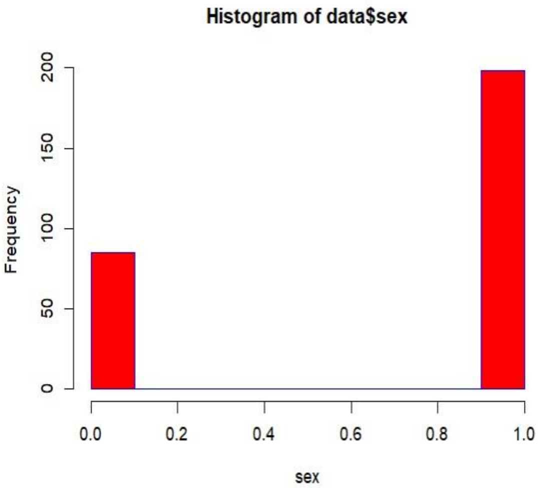

Histogram of data$thalachh
Frequency
0
10
20
30
40
50
80
100
120
140
160
180
200
thalachh

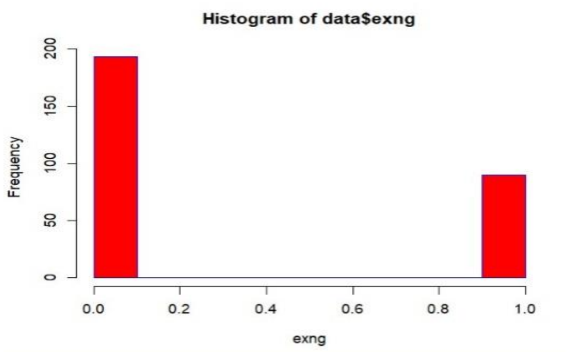

Histogram of data$age
Frequency
0
10
20
30
40
50
60
30
40
50
60
70
80
age

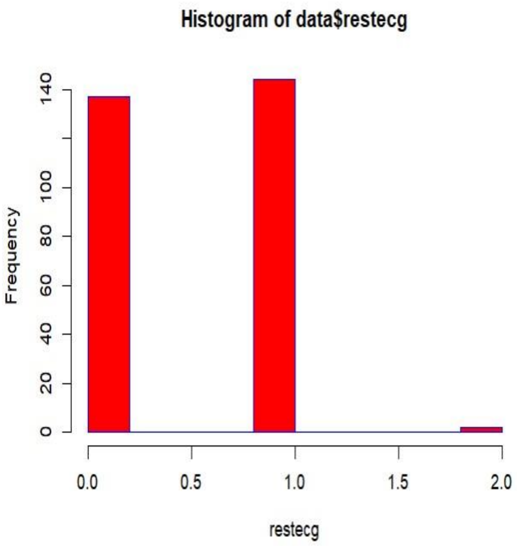

Histogram of data$chol
Frequency
0
10
20
30
40
50
150
200
250
300
350
chol

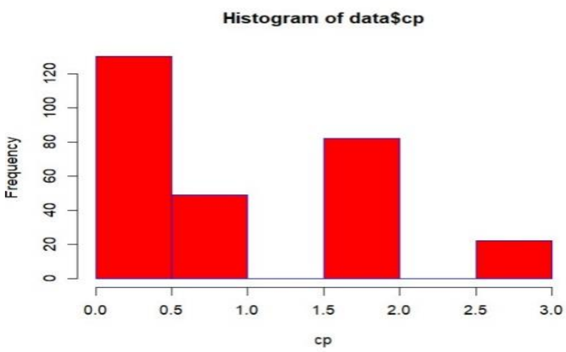


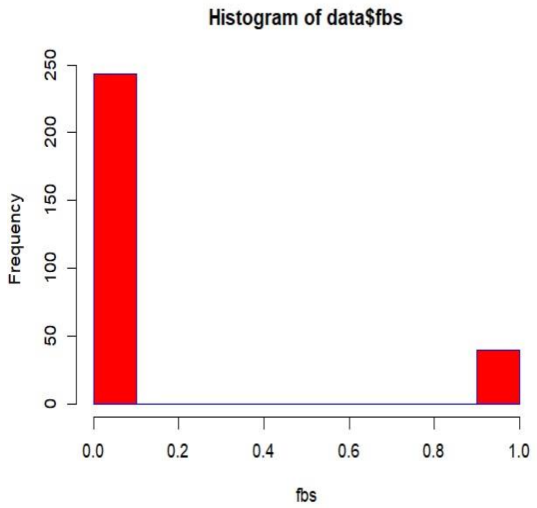

## Categorical Features

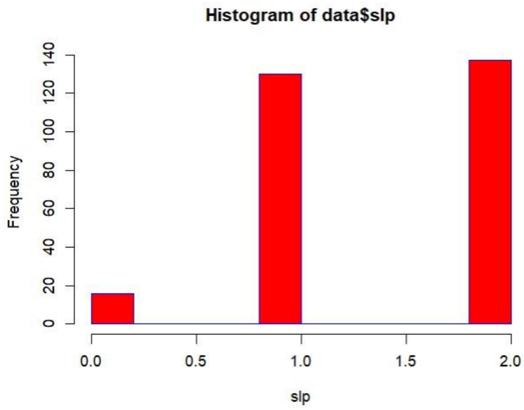


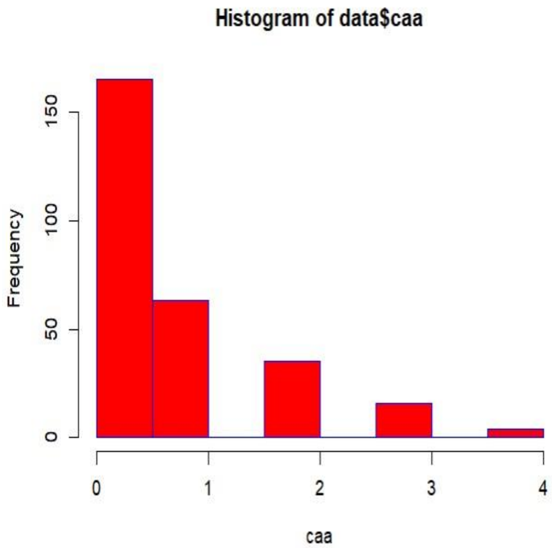






# 7.Data Sampling

The dataset is split into train and test in the ratio 7:3 respectively using random subsampling

Dimensions of train: 199 12
Dimensions of test: 84 12

# 8. GNB Classification

## GNB Model Summary

```
================================= Gaussian Naive Bayes =================================

- Call: gaussian_naive_bayes(x = X_train, y = Y_train)
- Samples: 199
- Features: 11
- Prior probabilities:
    - 0: 0.4171
    - 1: 0.5829

----------------------------------------------------------------------------------------
```

## GNB Model prediction – Confusion Matrix

```
   y_pred
     0  1
0   31 11
1    2 40
```

# 9. PERFORMANCE METRICS-RESULTS

Here the Positive class is 0 - i.e., person having less chance of heart attack

```
               Accuracy : 0.8452
                 95% CI : (0.7499, 0.9149)
    No Information Rate : 0.6071
    P-Value [Acc > NIR] : 1.871e-06

                  Kappa : 0.6905

 Mcnemar's Test P-Value : 0.0265

            Sensitivity : 0.9394
            Specificity : 0.7843
         Pos Pred Value : 0.7381
         Neg Pred Value : 0.9524
              Precision : 0.7381
                 Recall : 0.9394
                     F1 : 0.8267
             Prevalence : 0.3929
         Detection Rate : 0.3690
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.8619

       'Positive' Class : 0
```