**Informatics Institute of Technology**

**Department of Computing (BSc.) in Computer Science**

**Module: 5DATA001C.2**

**Machine Learning and Data Mining**

**Module Leader: Mr. Achala Aponso**

**Tutorial 1**

Date of Submission : 09/052022

Name : K.A.D.S.T.Kumarapeli

UOW No : W1790958

IIT No : 2019952

# Clustering Part

With the given problem description we will be able to perform all tasks whit in below steps.

Text

Description automatically generated

## EDA- Get High Level Idea about dataset

Text

Description automatically generated

From above figure we will be able to get an idea about the dataset and the data distribution with Min and Max Values and Some Statistical Measures of the Dataset.

There are 4710 total observations and 12 features in this dataset. There is a column called quality which is ordinal variable to represent quality of each red wine sample and it has four qualities such as 5,6,7,8 and in this project, I will map these values to 1,2,3,4 and it will be help in next stages.

Text

Description automatically generated with medium confidence

Missing Values!

Graphical user interface, text

Description automatically generated

There are not any missing values that we have to deal with in this dataset.

Chart, scatter chart

Description automatically generated

By using this figure, we can see alcohol level is the most correlated feature for the Quality.

## Outlier Removal and Feature Scaling - Preprocessing

**Outlier Removal**

Outlier is a silent killer, an observation that is numerically distant from the rest of the data or in a simple word it is the value which is out of the range.

Reasons for outliers,

\* Data Entry Errors: - Human errors such as errors caused during data collection, recording, or entry can cause outliers in data.

\* Data Entry Errors: - Human errors such as errors caused during data collection, recording, or entry can cause outliers in data.

\* Measurement Error: - It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty.

\* Natural Outlier: - When an outlier is not artificial (due to error), it is a natural outlier. Most of real-world data belong to this category.

For detect outliers we can use different techniques such as Hypothesis Testing, Z-score methos, Isolation Forest and Visualizing the Data. In this Project We use Z-score Method.

**IQR method**

Diagram

Description automatically generated

Figure : IQR method

Any Value which is -1.5\*IQR > x or 1.5\*IQR > x treated as outliers.

\* Q1 represents the 1st quartile/25th percentile of the data.

\* Q2 represents the 2nd quartile/median/50th percentile of the data.

\* Q3 represents the 3rd quartile/75th percentile of the data.

\* (Q1–1.5\*IQR) represent the smallest value in the data set and (Q3+1.5\*IQR) represent the largest value in the data set.

**Z-Score Method**

Using this method, we will be able to find how many Slandered Deviations value away from the mean.

Chart, histogram

Description automatically generated

Figure shows area under normal curve and how much area that standard deviation covers.

\* 68% of the data points lie between + or - 1 standard deviation.

\* 95% of the data points lie between + or - 2 standard deviation

\* 99.7% of the data points lie between + or - 3 standard deviation

By experimenting above mentioned IQR method I experienced most of the data points are eliminated with IQR method. Because of that reason we are using Z-score method to Scale and Remove Outliers in this Dataset.

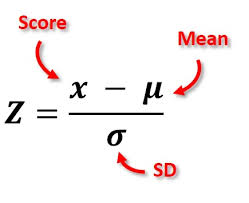


Figure : z-score formula

A screenshot of a computer screen

Description automatically generated with medium confidence

Figure : applying z-score normalization

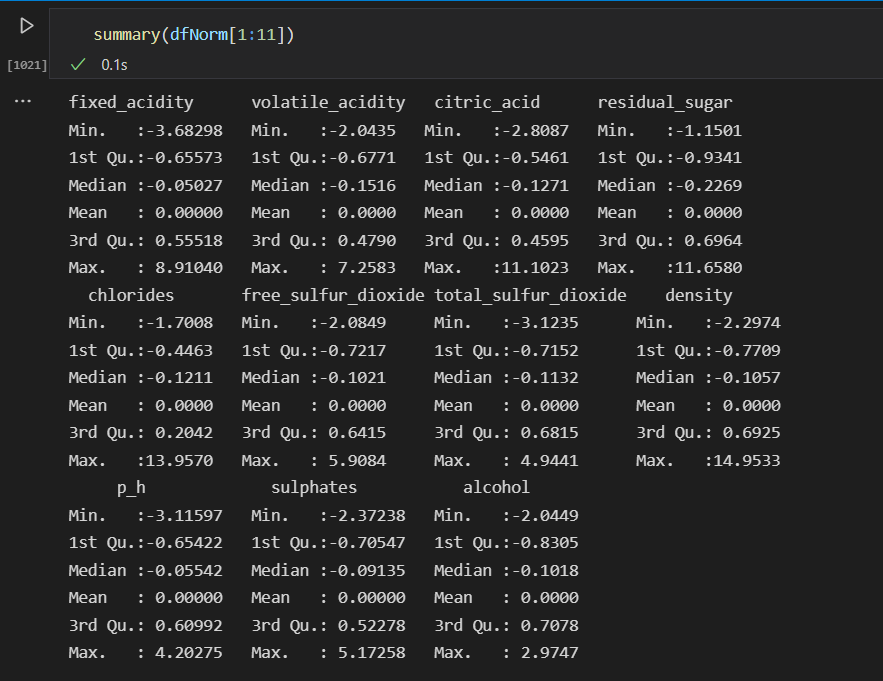


Figure : summary of normalized data frame

Eliminating all outliers using z-score,

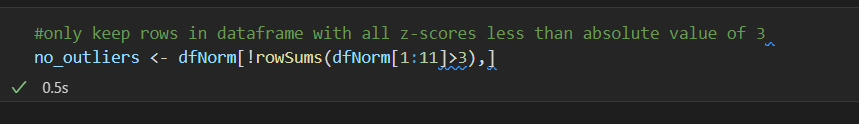


Figure : eliminating outliers

Below figure shows a summary of final dataset after scaling and outlier removal.

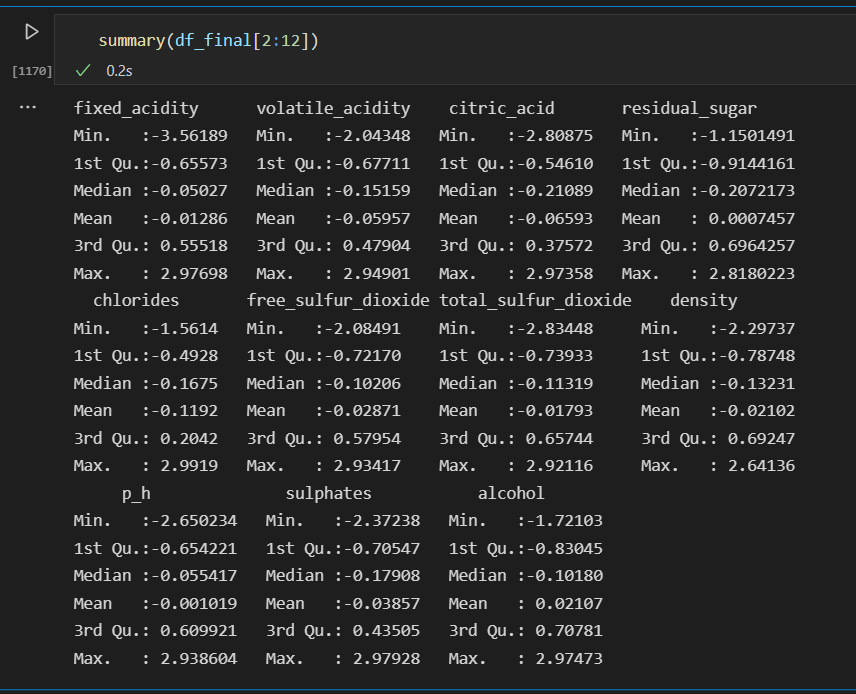
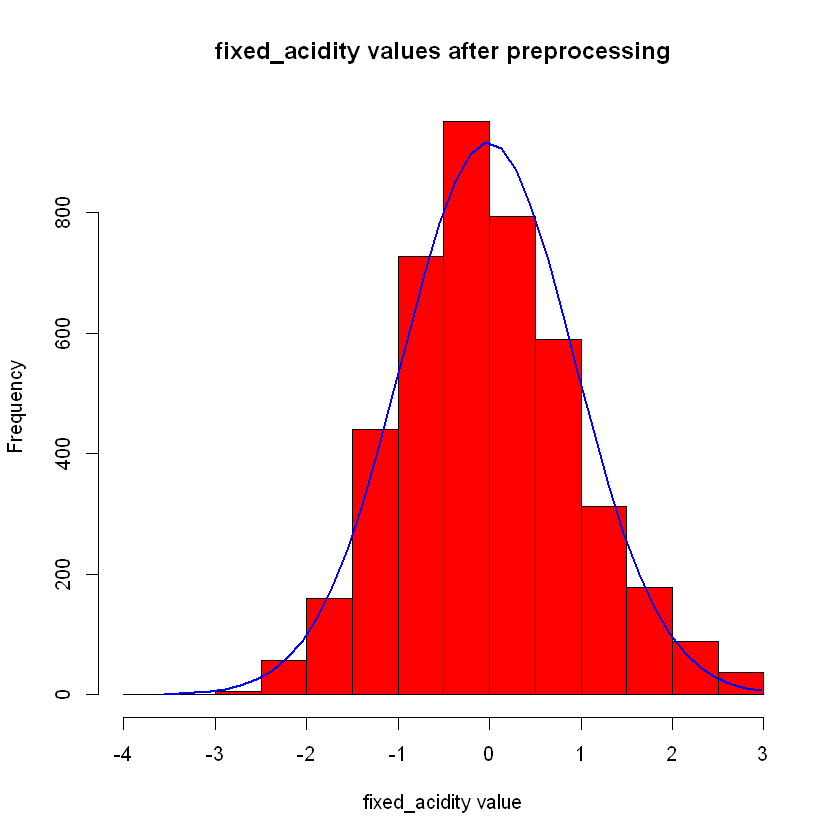
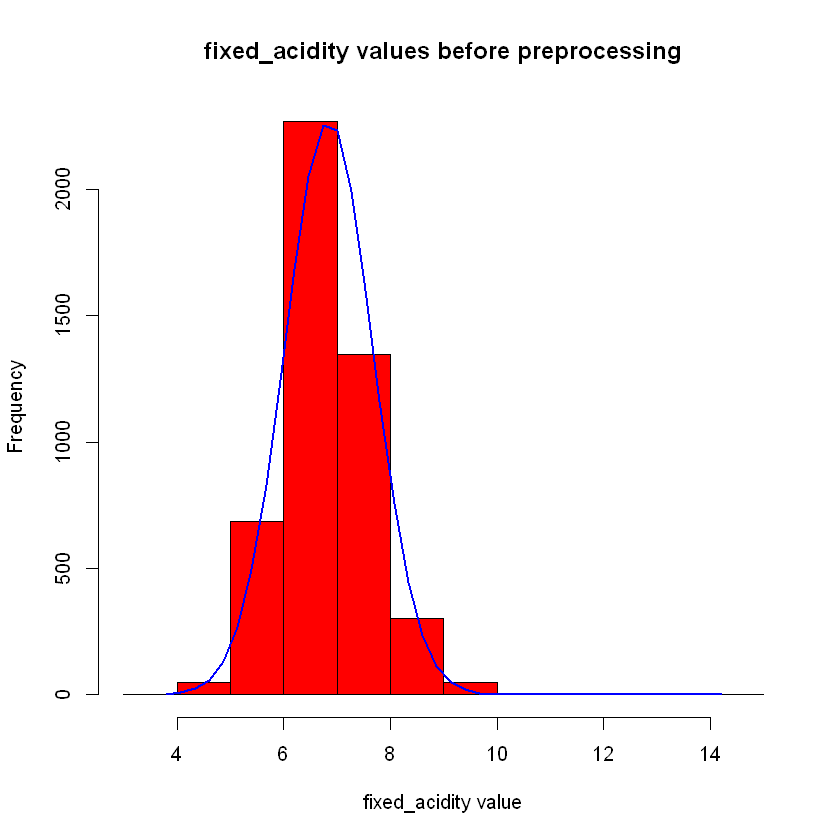
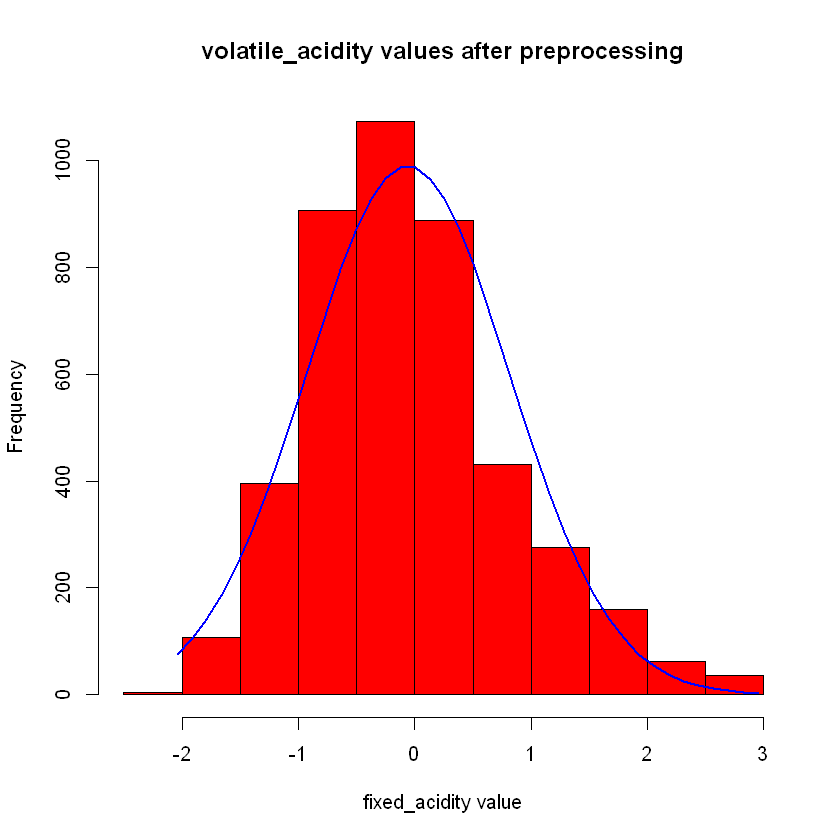
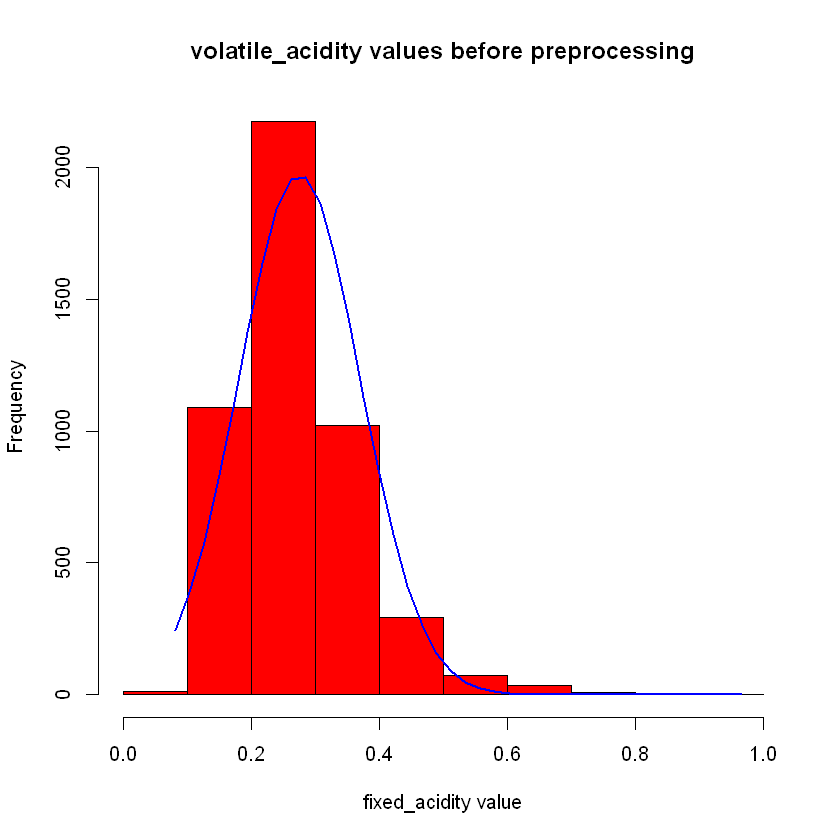
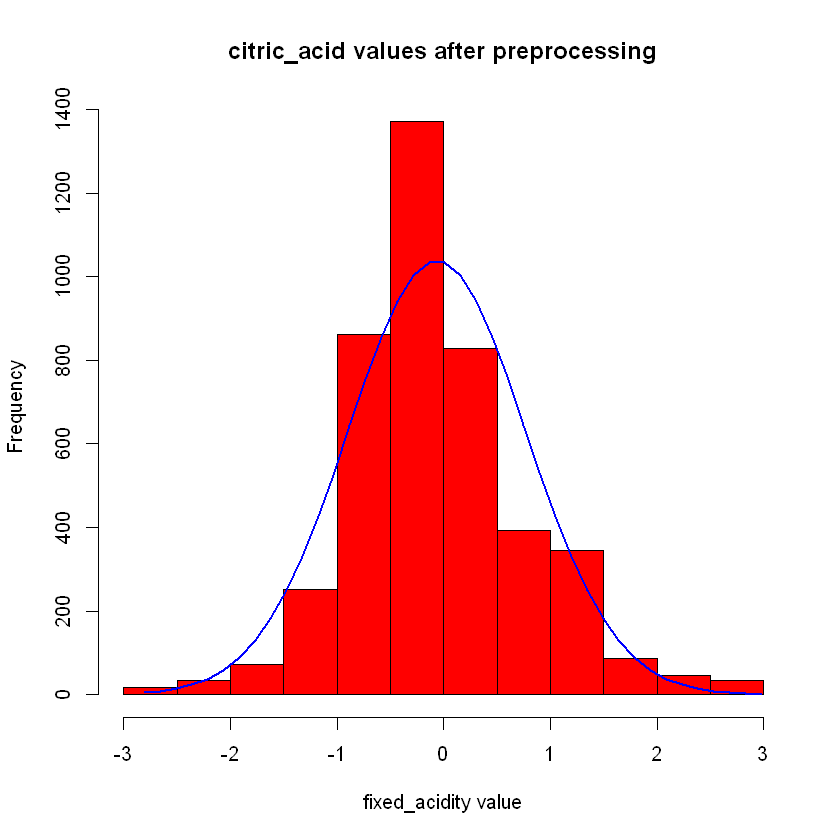
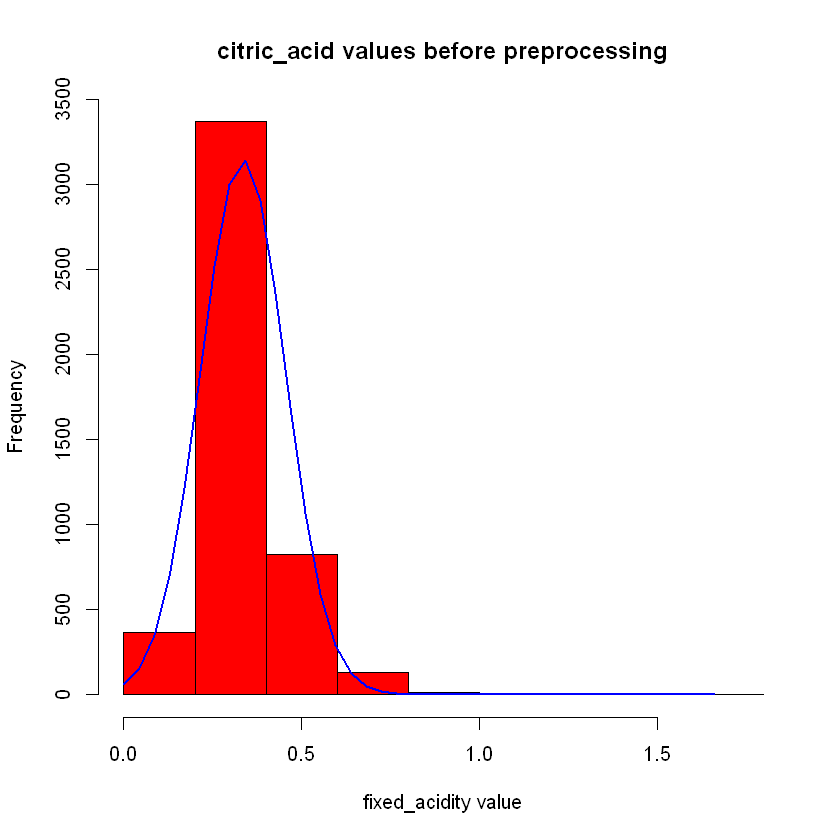
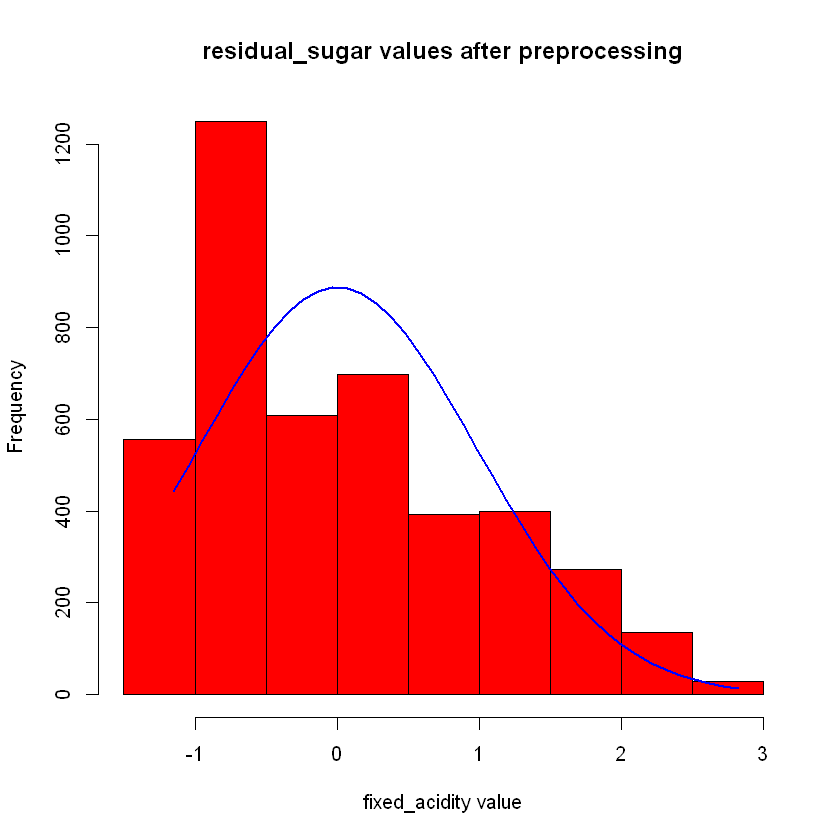
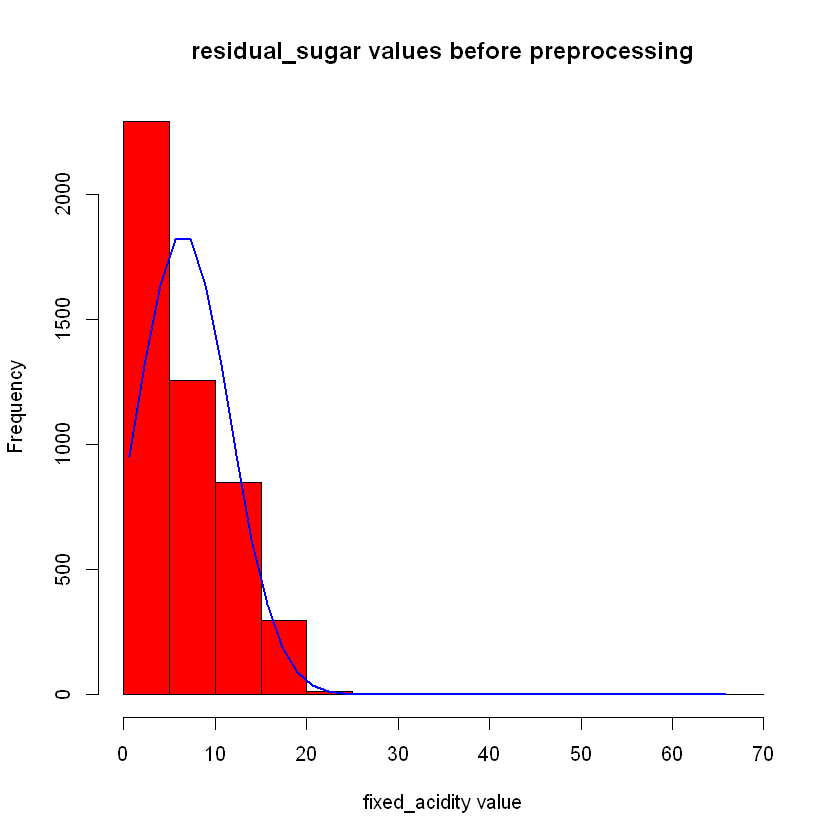


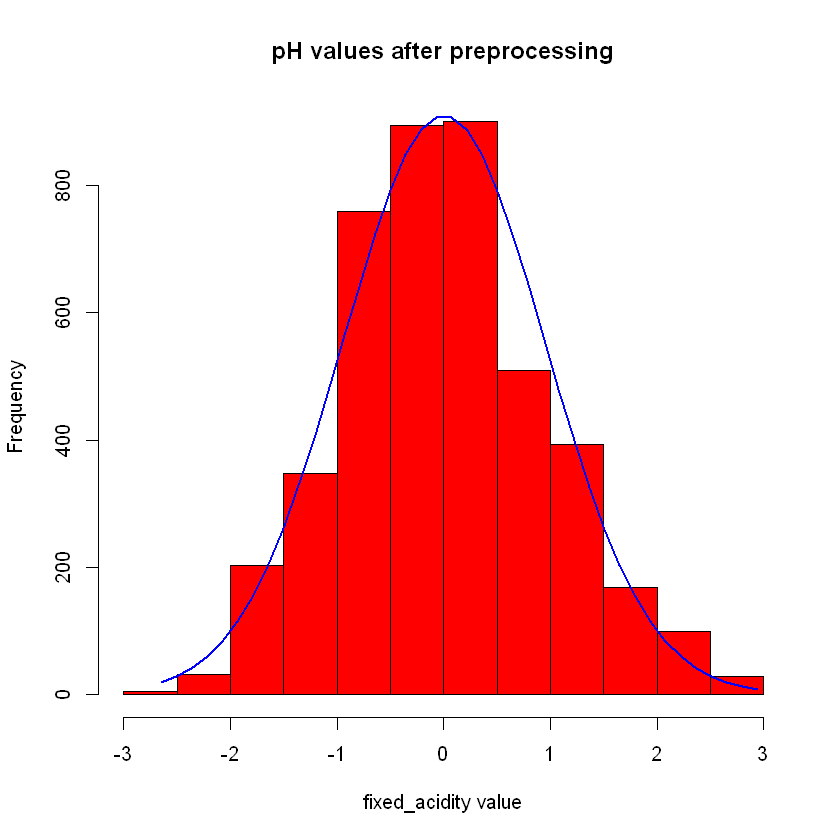
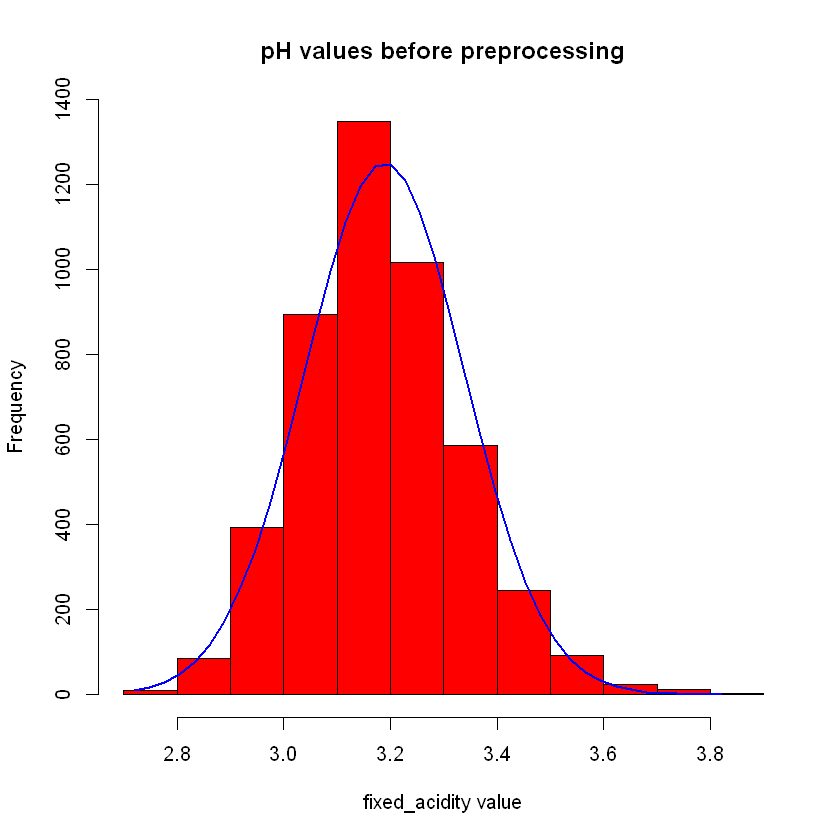
Figure : summary of scaled and outlier removed data frame

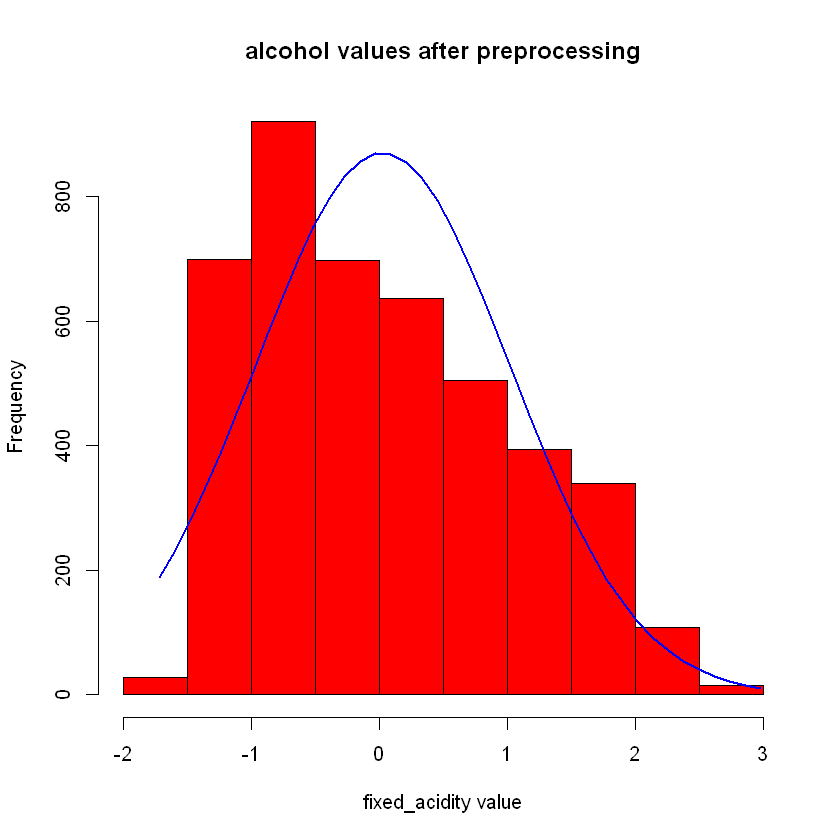
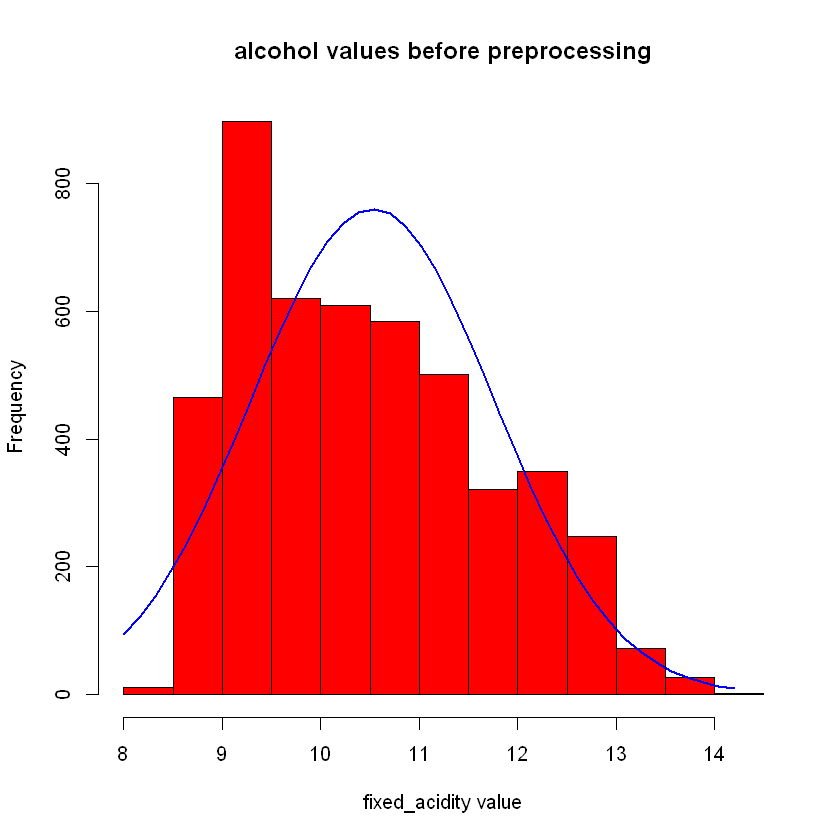




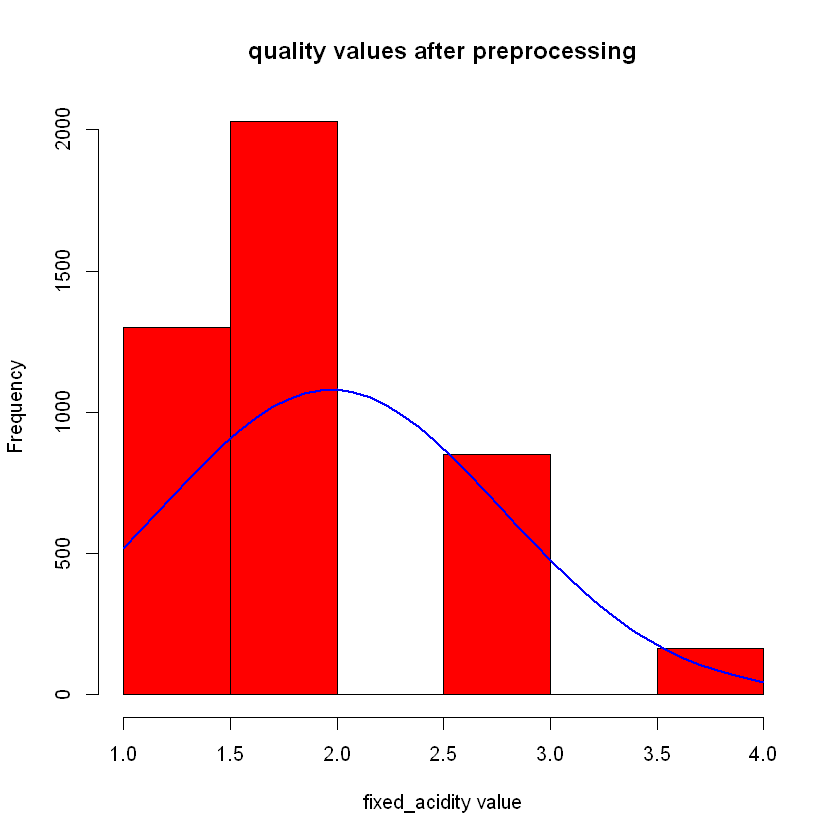
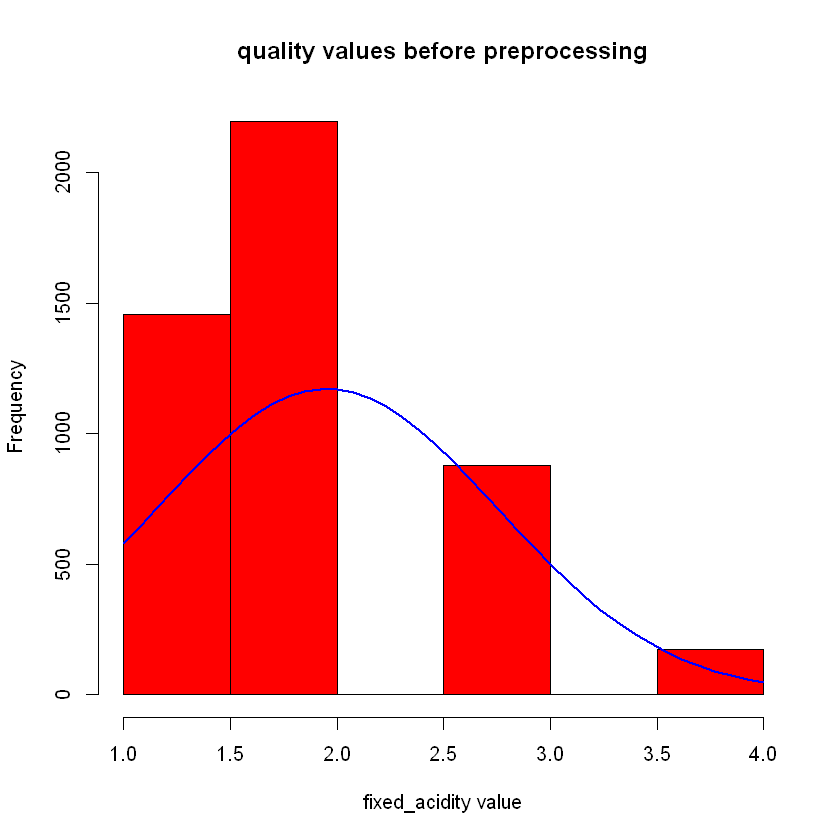








From below figure we will be able to figure it out it doesn’t affect to Quality feature distribution to Scale and the Outlier removal task.



Scaling and outlier removal is very important task when doing ML use case, it will be helpful to increase performance and the accuracy of the Machine Learning Model.

## Define the number of cluster centers

**Manual Method**

This red wine dataset originally consisted with 1-10 range qualities but in this dataset, we will be able to detect 4 qualities which are 5,6,7 and 8. So there will be four clusters in this data frame and further clarification we must process with different methods and techniques to evaluate number of clusters in this red wine dataset.

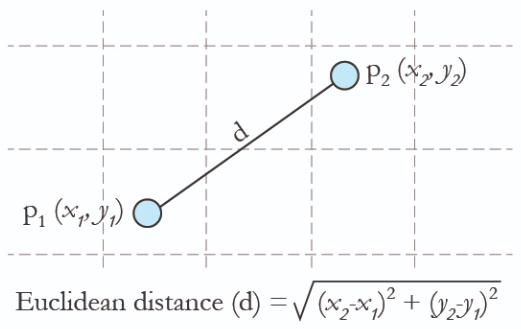
**Automated Methods**

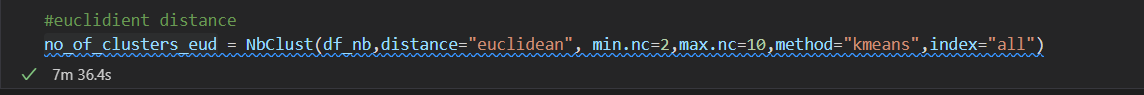
**1.NBclust**

**Text

Description automatically generated**

**i. Euclidian Distance**





Text

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Text

Description automatically generated

**ii. Manhattan Distance**

Text

Description automatically generated

Chart, diagram

Description automatically generated with medium confidence

Text

Description automatically generated

**2.Elbow Method**

**A screenshot of a computer

Description automatically generated with medium confidence**

In below figure k value is initializing 1 to 10 and it will calculate the “WSS” (With in cluster Sum of Square) and return. We need to plot this output to get a better idea about that. This elbow point is the optimal cluster value of k. in below figure it is k=4

Chart, line chart, histogram

Description automatically generated

**3.Silhouette Method**

This will calculate average Silhouette with each k value and display the optimal solution.

**Graphical user interface, text

Description automatically generated**

Chart, line chart

Description automatically generated

Above figure silhouette method returned optimal number of clusters as 2.

As above mentioned, automated tools most of the tools suggested k=2 for optimal clustering and elbow method suggest k=4, from this idea and as form the instructions given in the problem we will perform k-means clustering from k=2, k=3 and k=4 in next chapter and evaluate each clusters using BSS/TSS and confusion matrix.

## K-means Clustering

From above chapter we discussed some automated and manual methods to select optimal cluster amount for k-means algorithm, in this chapter we will experimenting k=2, k=3 and k=4 as number of clusters and evaluate which is the “winning” clustering approach.

To evaluate this process, we use WSS (Between cluster Sums of Square) over TSS (Total cluster Sums of Square) and Confusion Metrix calculations such as accuracy/precision and recall.



Figure : confusion matrix

**K=2**

A picture containing graphical user interface

Description automatically generated

For further analysis we will divide data frame into x and y, x: all numerical features / y: quality.

Graphical user interface, application

Description automatically generated

Above figure shows executing k-means inbuilt function in r with data and number of clusters. Then we will see the results.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Then we will be able to create confusion matrix, below figure shows the confusion matrix,

Text

Description automatically generated

Text

Description automatically generated

Then we visualize this result with the help of visualization libraries.

Chart, scatter chart

Description automatically generated

**K=3**

Graphical user interface, application

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

**K=4**

Graphical user interface, application

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Chart

Description automatically generated with medium confidence

Visualizing all together,

Scatter chart

Description automatically generated

**Compression Table**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster amount | BSS/TSS | Accuracy | Recall/Sensitivity | Specificity |
| k=2 |  |  |  |  |
| k=3 |  |  |  |  |
| K=4 |  |  |  |  |