Topics in Data Analytics
Project -1

Topic : Topic 4
Name : G Naga Tarun Kumar
Roll no : B170594CS
Mobile no : 9182946856

The implementation is done using R programming.
Tool used : R-Graphical User Interface(R-GUI)
Methodology:

- ➢ “mlbench” library has been installed
- ➢ The ”Glass” database has been imported in the form of a dataframe from the above library.
- ➢ Initially, the duplicates (if any) of the dataset were removed.

1.) fivenum() function is used to get the five number summary (Min_value, lower quartile, median, upper quartile, Max_value) of each attribute of the dataset namely RI, Na, Mg, Al, Si, K, Ca, Ba, Fe after removing the Null values for each attribute using na.rm=TRUE. The Grid is divided into nine equal plots using par(mfrow=c(1,1)). Box plots are drawn for every attribute boxplot() function.

2.) The quantile() function is used to find the 25th and 75th percentile for each attribute of the dataset. The IQR value of each attribute has been calculated using the IQR() function. The InterQuartile Range of each attribute has been determined using the below formulas.

Upper limit = Q2 + y*1.5
Lower limit = Q1 - y*1.5

where Q1 is 25th percentile, Q2 is 75th percentile, y is IQR value.
Then the dataset is filtered using the subset() function and the new dataset after removing outliers is obtained i.e, the rows containing the attribute values in the [Lower limit, Upper limit] are considered. Thus, the outliers are removed.

3.) The new data set after removing outliers is taken and a five number summary of each attribute is calculated using fivenum() function after removing the null values. The grid is divided into nine equal plots and box plots are drawn for each attribute using boxplot() function.

4.) The grid is divided into nine columns using par(mfrow=c(1,9)) function. The original and new datasets were considered and a five number summary of each attribute is calculated in both the datasets. Both the box plots for each attribute were plotted in the same column i.e, same plot and can be differentiated using the tags old and new on y-axis. So, they can be compared easily.

Assumptions:

- ➢ The axis of the box plot is taken HORIZONTAL.
- ➢ The box plots of all the attributes are drawn side by side in the same grid.