**Description**

This dataset was developed initially to determine whether a glass type was float or non-float. Glass identification can help in crime investigation. With the identification of the type of glass which is left as evidence in the crime scene, the investigation can be boosted. In this dataset, different instances of such glasses are observed or considered as data samples. There are 214 observation samples in this data. For each sample, nine features of the sample are recorded. Those features are:

* Refractive Index (RI)
* Sodium (Na)
* Magnesium (Mg)
* Alumunium (Al)
* Silicon (Si)
* Potassium (K)
* Calcium (Ca)
* Barium (Ba)
* Iron (Fe)

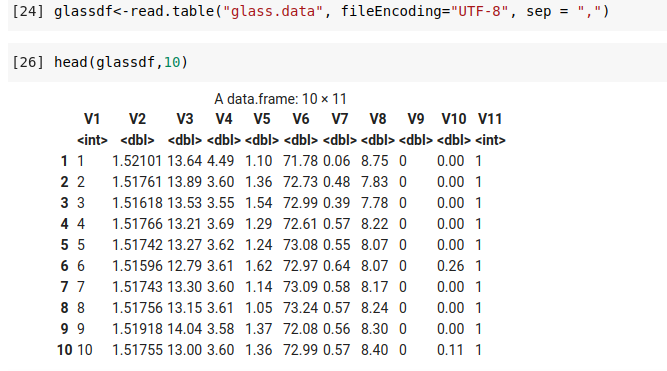
Along with such features, types of the glass referring to an individual sample are provided in a class column. The classes are maintained as follows:

1. Class 1 for building windows float processed
2. Class 2 for building windows non float processed
3. Class 3 for vehicle windows float processed
4. Class 4 for vehicle windows non float processed
5. Class 5 for containers
6. Class 6 for tableware
7. Class 7 for headlamps

This is a classification task in which new samples containing such features will be tested. The type of glass of such new samples should be identified by the algorithm developed.

**Data Preparation**

Initially the data was loaded using the read.table method in R. The data was downloaded from the UCI machine learning repo website. The columns of the data loaded are provided with their respective name. On observing the first 10 samples of the data, the following result was obtained.



After the loading of the dataset, the column names were changed according to the description of the dataset given in UCI ML repo website. After changing the column names following first 10 observations of the data was obtained.

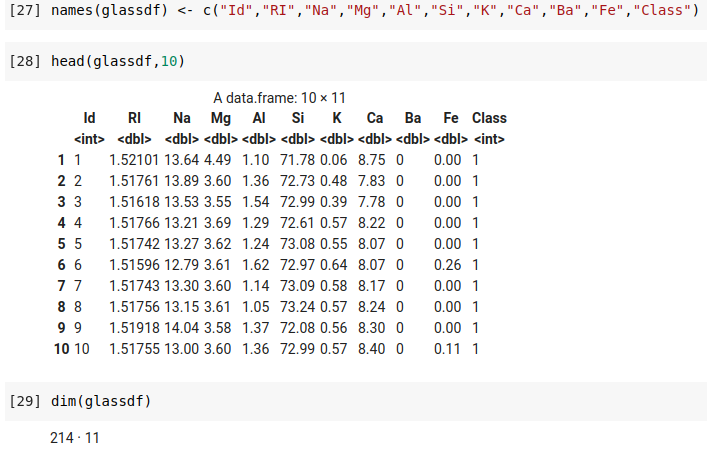


Fig: Screenshot showing the first 10 observations in glass identification dataset and dimension of the dataset

The dimension of the dataset suggests that there are 214 samples of such glass data and 10 features including the glass Id and a class representing the type of glass.

**Data Validation**

In this step, we check whether the number of columns and rows is equal to the number mentioned in the metadata or not and get the summary to compare with the one given in the glass.tag file.

Initially, we check whether there are null values in the dataset or not.

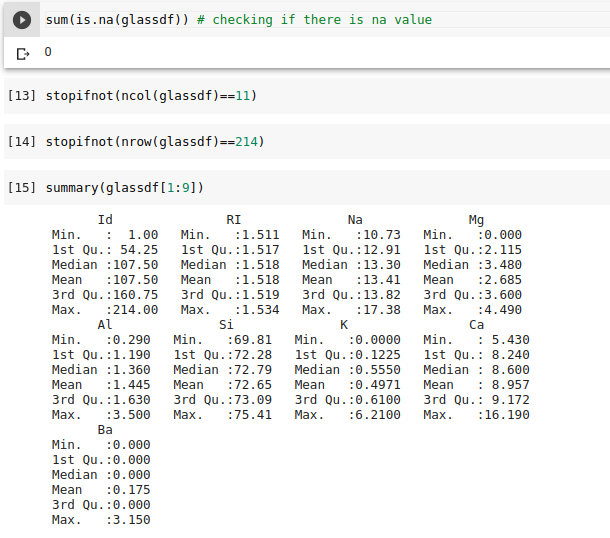


Fig: Screenshot representing data validation of the glass identification dataset.

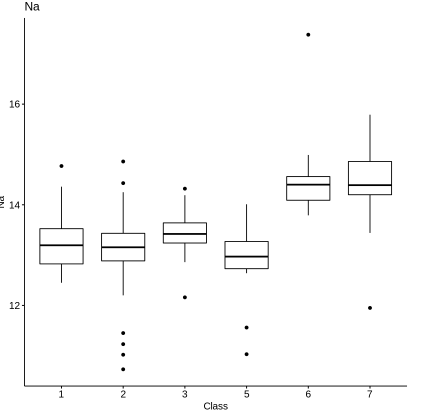
**Data Analysis**

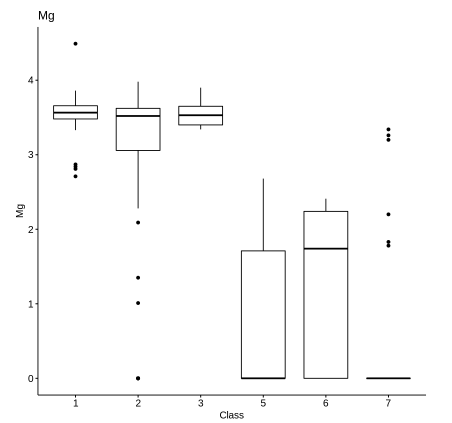
**Box plot**

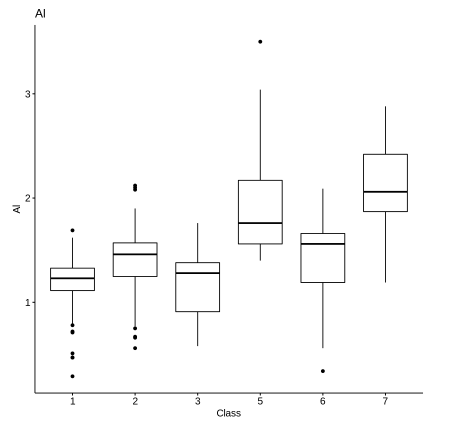
Box Plot shows some outliers present in the data that needs further cleaning. They are useful because they show the average score of the dataset. They are specifically used to get a visual representation of the dispersion of a data set, signs of skewness, mean values and so on.

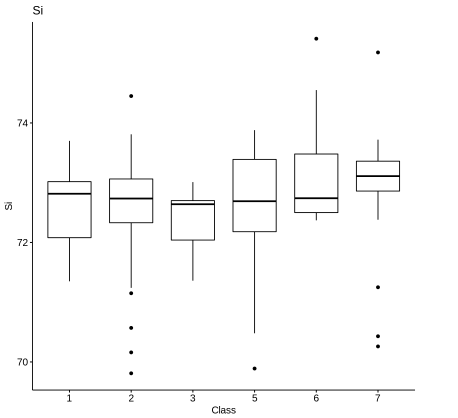
“ggpubr” and “ggplot2” package of R was utilized to generate box plots for the dataset.

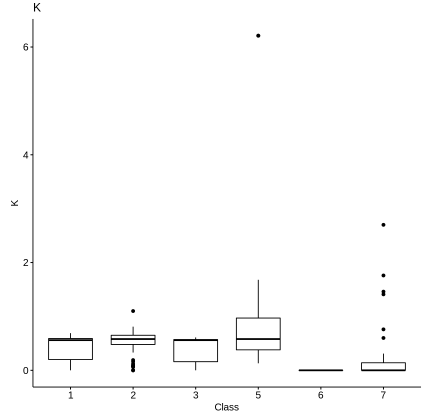
Following box plot was generated for the feature columns like “Na”, “Mg”, “Al”, “Si”, “K”, “Ca” and “Ba”.

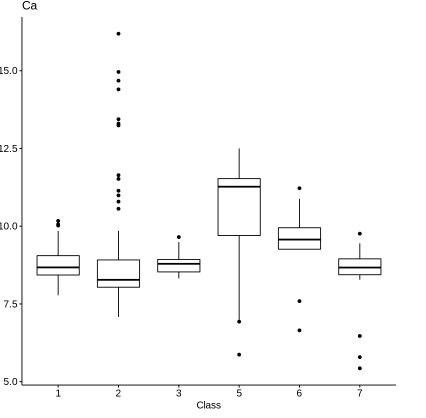












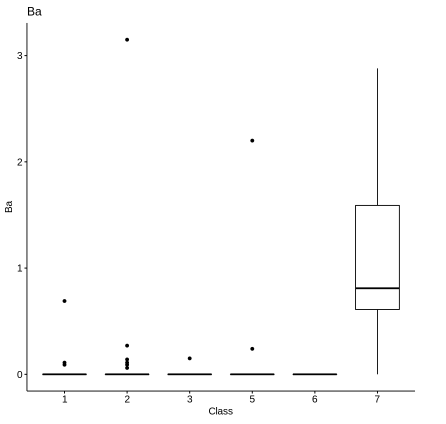


Fig: Box plot of the dataset attributes with respect to class

**Correlation Plot**

Correlation plot was performed on the dataset to evaluate the linear relationship measure of the features and class of the dataset. Following result was obtained:

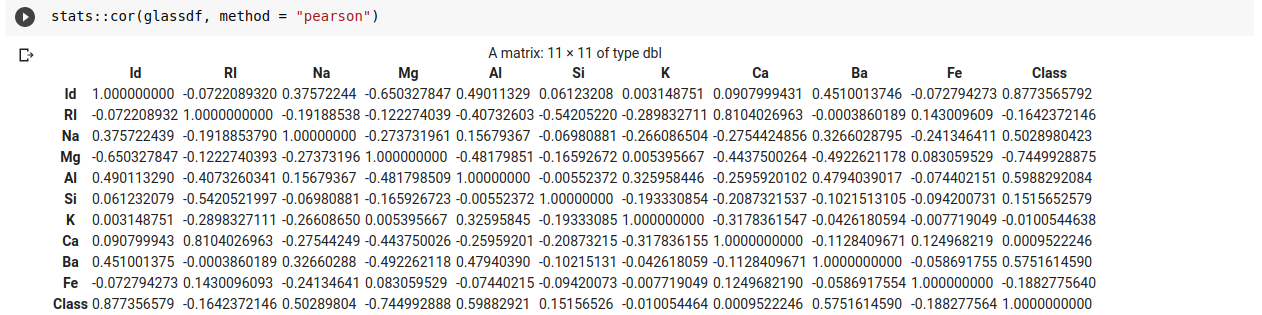


Fig: Correlation plot of the features and class of the dataset

The correlation plot suggests that two best oxides that best predicts the refractive index of the glass are:

1. Calcium (Ca)
2. Iron (Fe)

Similarly, we can draw conclusion from the correlation plot that the two best oxides that best predicts the class of the glass are:

1. Aluminium (Al)
2. Barium (Ba)

**Scatter plot**

The function pair.panels is the “psych” R package is used to generate a scatter plot of the matrices in which we have bivariate scatter plot below the diagonal, histograms of the data of the column in the diagonal and Pearson correlation above the diagonal. Following scatter plot was generated for the feature columns of the dataset.

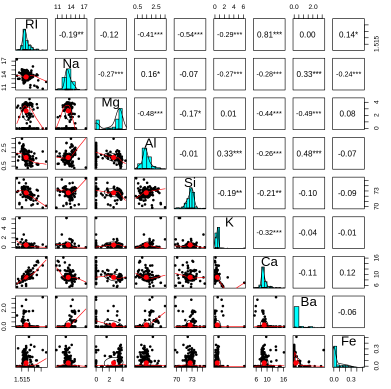


Fig: Scatter plot of the features of glass dataset

**MANOVA test**

Multivariate variance analysis (MANOVA) of many dependent variables is essentially an ANOVA. In other words, ANOVA measures the mean difference between two or more groups, while MANOVA tests the mean difference between two or more vectors. It can be evaluated concurrently using a multivariate variance analysis where there are several response variables (MANOVA).

MANOVA assumptions

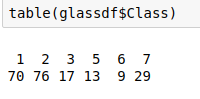
- The answer vector should usually be assigned to

- Homogeneity of variances across the predictor spectrum.

- Linearity between all pairs of variables that are dependent

- Adequate sample size

- Lack of an outlier, univariate or multivariate



Above table shows a minimum of 9 data which is greater than the number of classes. Thus, the first assumption is verified. We make the use of libraries such as “tidyverse”, “rstatix”, “car”, “broom” for this task. Initially, we should remove any outliers present in the data. To do this we perform,

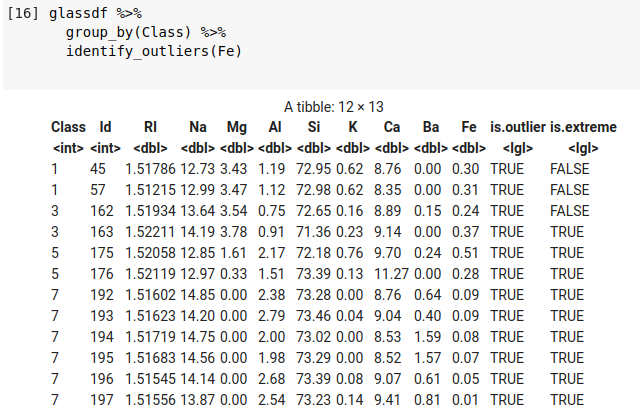
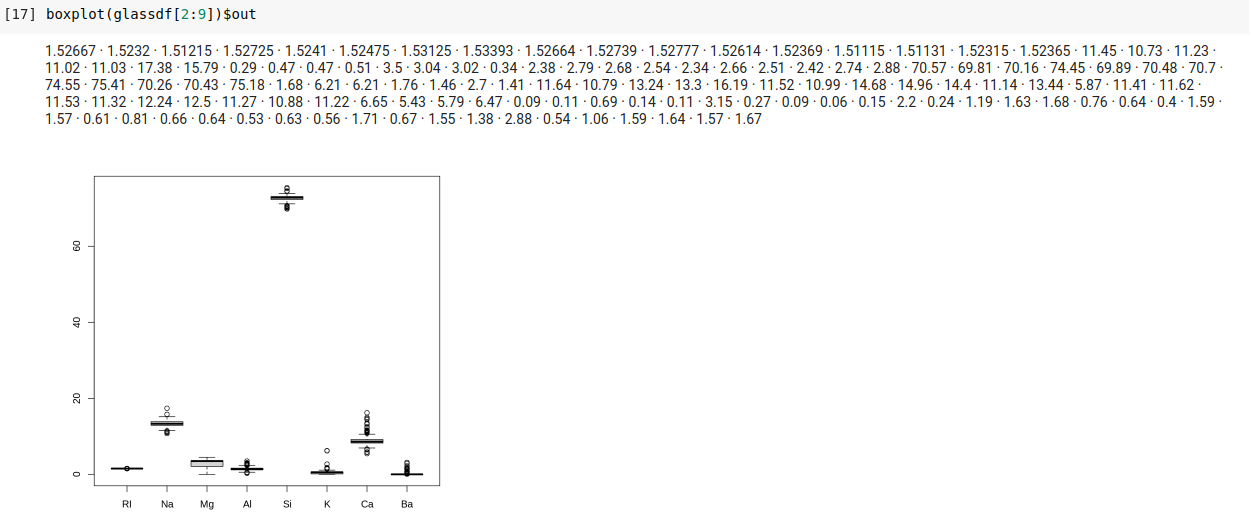
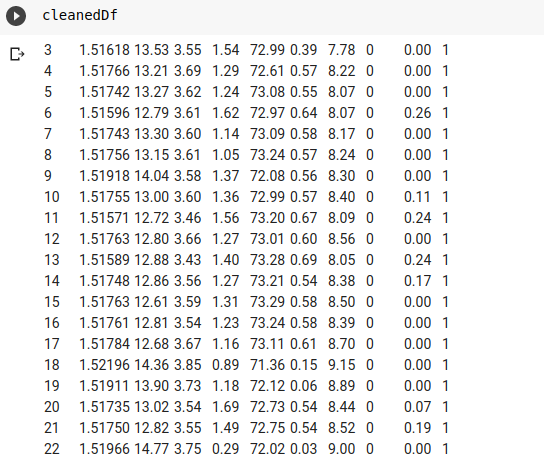


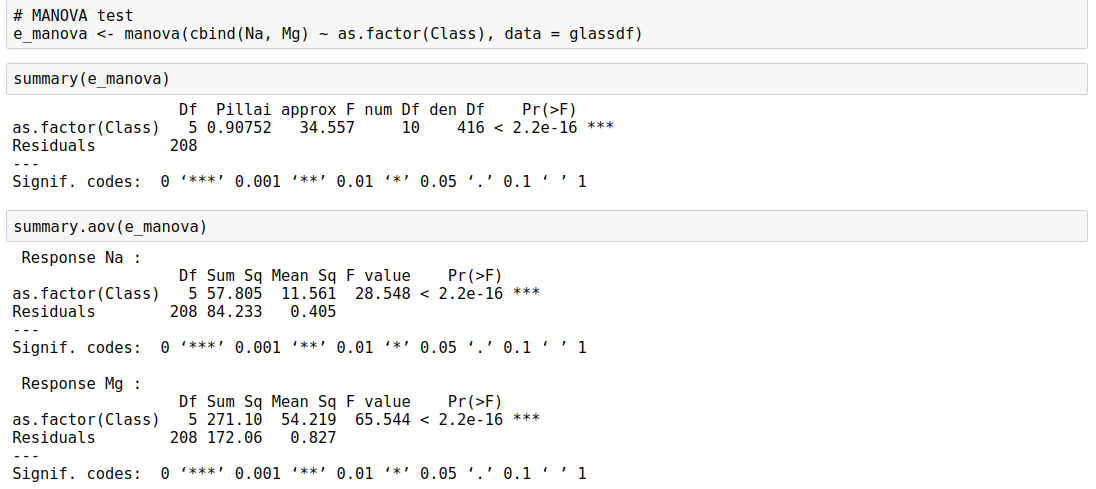
Fig: Removing outliers in the dataset



Finally, we obtain the clean dataset as shown below:



Finally, we perform MANOVA test result by executing following code:



Fig” Results of MANOVA test