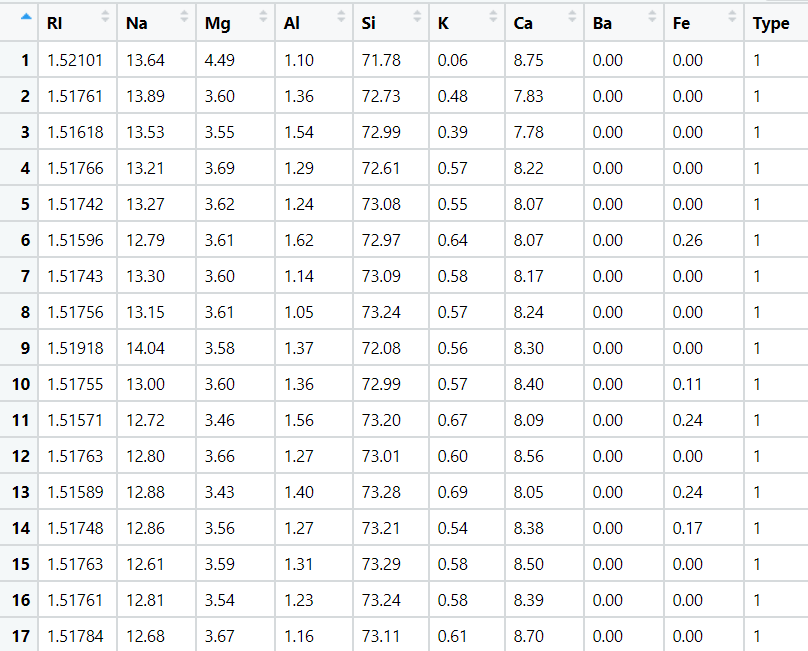
# Topic: K-Nearest Neighbor

1.) Prepare a model for glass classification using KNN. The dataset looks like this:



**Ans:**

**Data Analysis:**

Input Variables: Other Factors

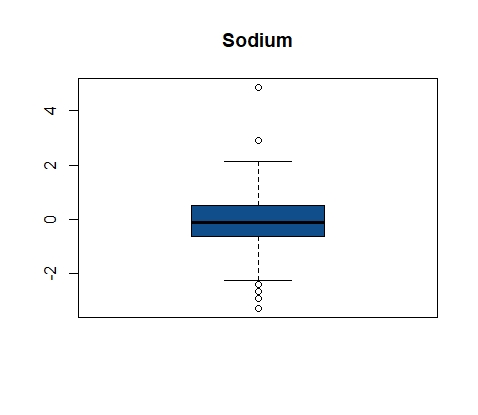
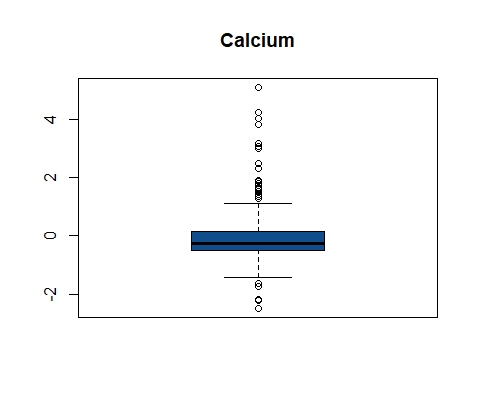
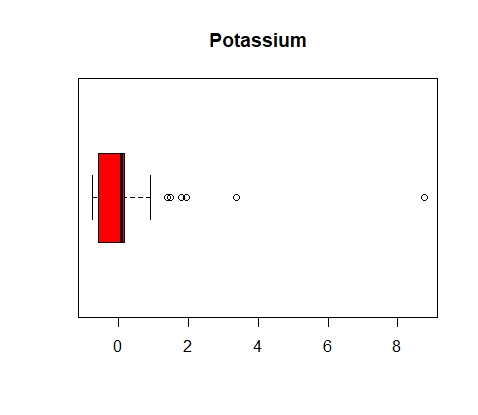
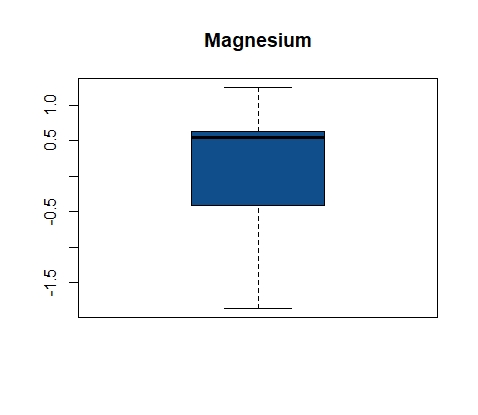
Output Variable: Type of glass

**Data Preprocessing:**

* Once the data analysis we need to process the data by removing unwanted columns and reordering the variable columns so the output will be at index 1.
* There are no NA values found, so no need to do imputation
* As the data of variables are not normal, we need to normalize or standardize the data so the values will be unit less.

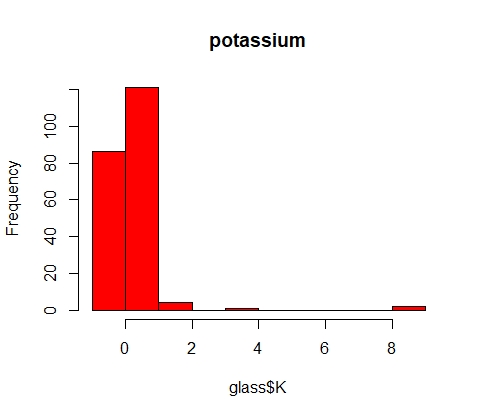
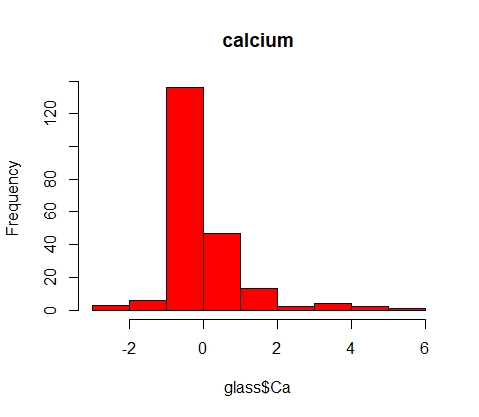
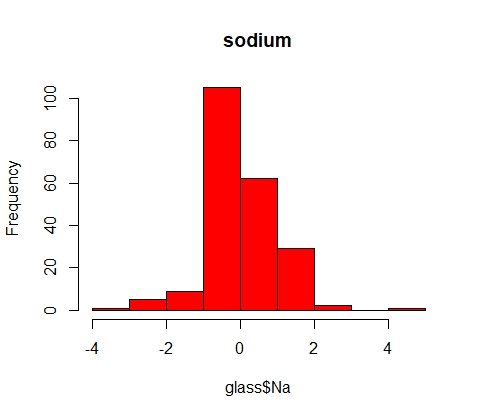
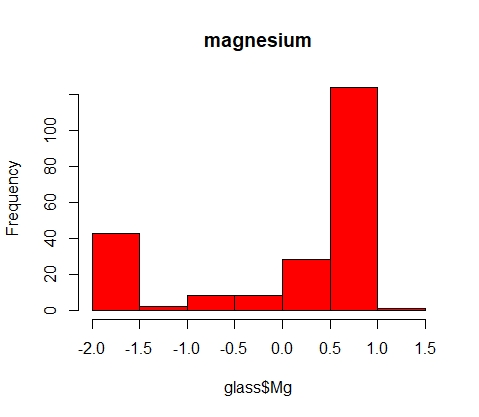
**Exploratory Data Analysis:**

**Box Plot Representation:**

****

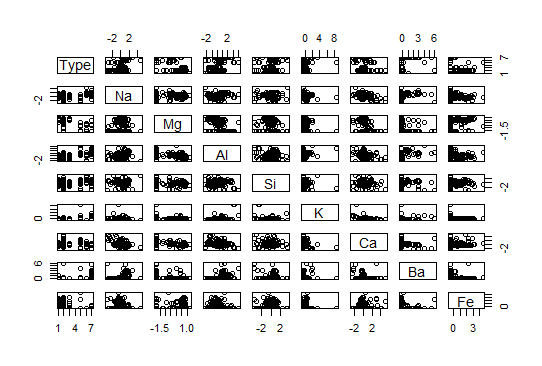
* From the box plot we can say that the individual variables has been effected with outliers

**Histogram Representation:**

****

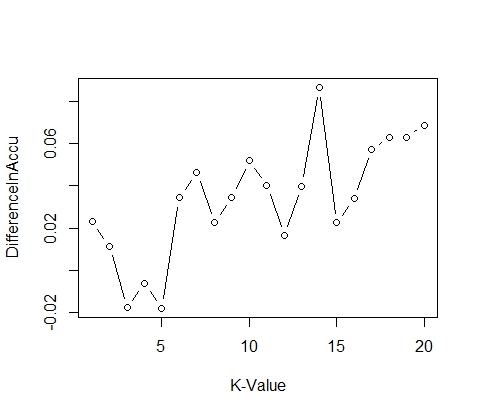
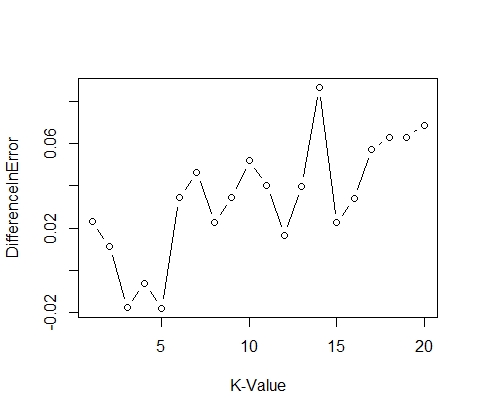
* For Potassium and Calcium, the data is right skewed and the data for sodium is quite normally distributed and when checking with Magnesium the data is left skewed.

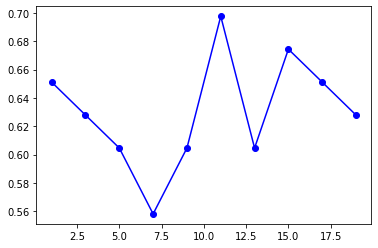
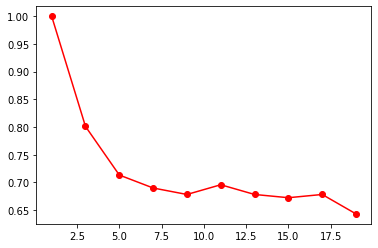
**Scatter Plot:**

****

**Model Building with KNN:**

* The data is divided into training and test with the partition of 80% and 20%.
* Once data is divided, we built a model with a random K value and find the accuracy of the test and train data
* At k = 9 , the test accuracy = 0.8139 and train accuracy is 0.9181,which shows the model is over fit we need standardize the data to get the right fit.
* To standardize the data here, we are using Z score standardization.
* Even after standardization the accuracy of the test data is still same, so we need to tune the K value and check the accuracy level.
* Instead of taking single k value at each, we can create a loop function for the range of K value where we can determine the best value for k and the right fit
* Here are the below graphs for K value for accuracy and error for both R and Python

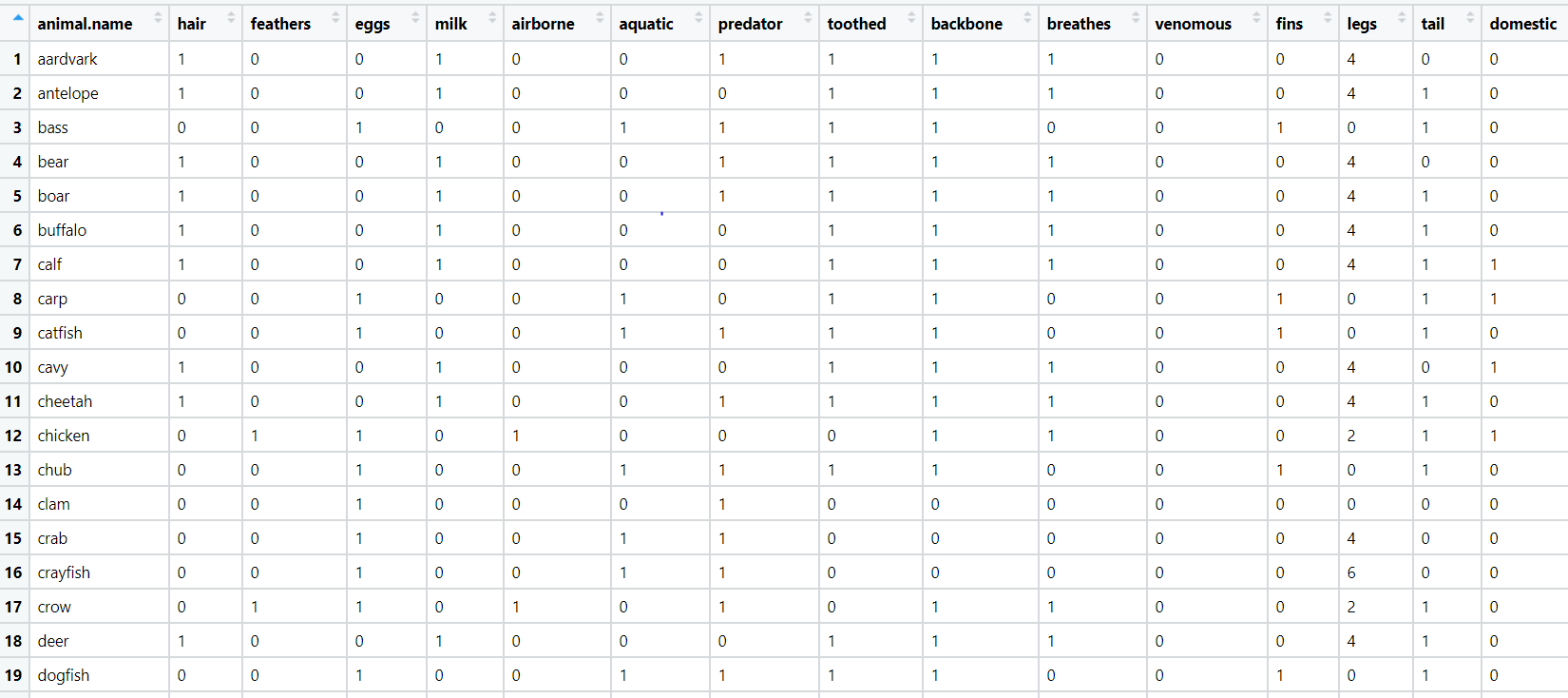




**Conclusion:**

* As per the above analysis the higher we take the 'k' value the lesser the accuracy.So the best model is considered when the 'k' value is 8

2.) Implement a KNN model to classify the animals in categories.



**Ans:**

**Data Analysis:**

Input Variables: Other Factors

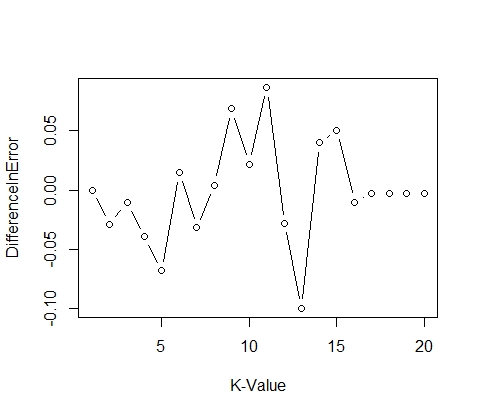
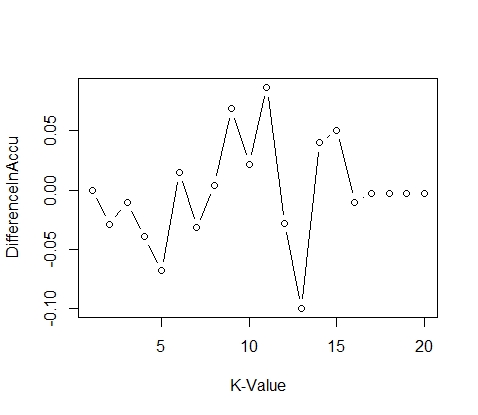
Output Variable: Type of glass

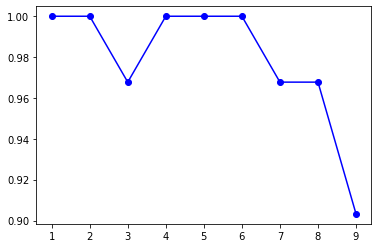
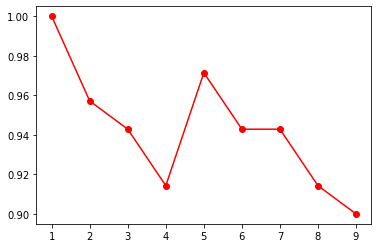
**Data Preprocessing:**

* Once the data analysis we need to process the data by removing unwanted columns and reordering the variable columns so the output will be at index 1.
* There are no NA values found, so no need to do imputation.
* Here no normalization is required as the values given the data are normal

**Model Building with KNN:**

* The data is divided into training and test with the partition of 70% and 30%.
* Once data is divided, we built a model with a random K value and find the accuracy of the test and train data (Sq rt(train data/2)
* At k = 7, the test accuracy = 0.97 and train accuracy is 0.94, which shows the model is over fir we need standardize the data to get the right fit.
* Instead of taking single k value at each, we can create a loop function for the range of K value where we can determine the best value for k and the right fit
* Here are the below graphs for K value for accuracy and error for both R and Python





**Conclusion:**

* The k value at 7 gives the test and train accuracy as 0.97 and 0.94 which is also a Right fit model

**Note:**

* **When the data divided in a sequential order the accuracy rate has been decreased drastically, so another model was built with the random data selection which gives the best output accuracy and right fit model at k=7.**

**Hints:**

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy. Ex: Feature Engineering, Hyper Parameter tuning etc.
3. All the codes (executable programs) are running without errors
4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here.