The Working Procedure of Challenge-1

Operations:

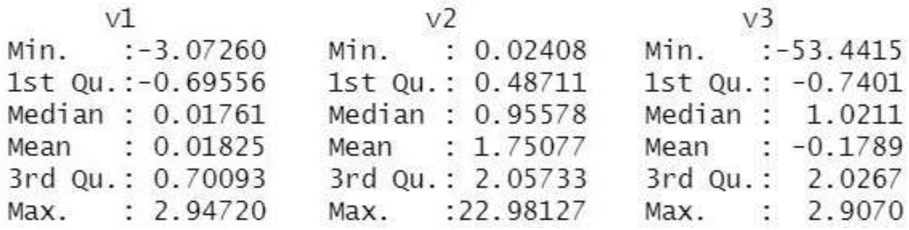
- Building a “Linear Regression” model on a “Training Dataset” and testing it on another “Testing Dataset”.
- Building a “KNN Regression” model on a “Training Dataset” and testing it on another “Testing Dataset”.

Dataset:

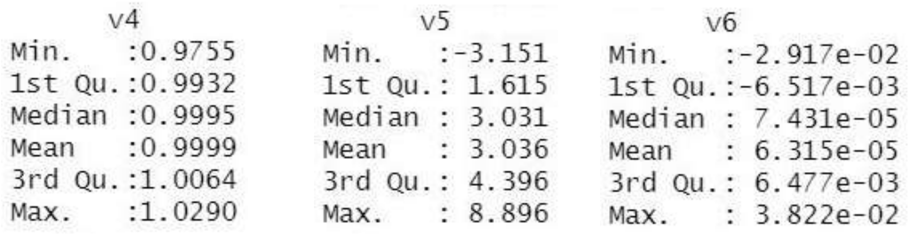
- **Training Dataset**: A dataset of 1000 observations of one “ID/Index variable”, 9 “Predictor Variable”(V1, V2, V3, V4, V5, V6, V7, V8, V9), and one “Target Variable”(Y)
- **Testing Dataset**: A dataset of 100 observations of one “ID/Index variable” and 9 “Predictor Variable”(V1, V2, V3, V4, V5, V6, V7, V8, V9).

Data Analysis: Here, in the Training Dataset 9 predictor variable is found. So, to analyze the predictor variables and their relation with the Target data, the Summarizing, Histogram, Boxplot, Bi-variate, and Correlation Analyses have been conducted.

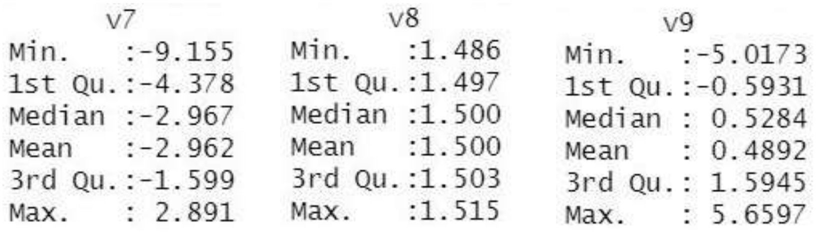
- **Summary Results:** The summarizing operation was conducted on the data set to get an overview of the Minimum value, Maximum value, Mean value, Median value, 1st, 2nd, and also 3rd quartile value and also the numerical relation, contrast in the variable’s distribution.



```
       v1                  v2                 v3
 Min.   :-3.07260   Min.   : 0.02408   Min.   :-53.4415
 1st Qu.:-0.69556   1st Qu.: 0.48711   1st Qu.: -0.7401
 Median : 0.01761   Median : 0.95578   Median :  1.0211
 Mean   : 0.01825   Mean   : 1.75077   Mean   : -0.1789
 3rd Qu.: 0.70093   3rd Qu.: 2.05733   3rd Qu.:  2.0267
 Max.   : 2.94720   Max.   :22.98127   Max.   :  2.9070
```



```
       v4                v5                 v6
 Min.   :0.9755   Min.   :-3.151   Min.   :-2.917e-02
 1st Qu.:0.9932   1st Qu.: 1.615   1st Qu.:-6.517e-03
 Median :0.9995   Median : 3.031   Median : 7.431e-05
 Mean   :0.9999   Mean   : 3.036   Mean   : 6.315e-05
 3rd Qu.:1.0064   3rd Qu.: 4.396   3rd Qu.: 6.477e-03
 Max.   :1.0290   Max.   : 8.896   Max.   : 3.822e-02
```



```
       v7                  v8                  v9
Min.   :-9.155    Min.   :1.486    Min.   :-5.0173
1st Qu.:-4.378    1st Qu.:1.497    1st Qu.:-0.5931
Median :-2.967    Median :1.500    Median : 0.5284
Mean   :-2.962    Mean   :1.500    Mean   : 0.4892
3rd Qu.:-1.599    3rd Qu.:1.503    3rd Qu.: 1.5945
Max.   : 2.891    Max.   :1.515    Max.   : 5.6597
```

Fig-1: The summary result of the variables.

- **Histogram:** Then the histogram plot was drawn for each variable to find the distribution of data for each variable. From histogram the skewness of the data distribution was found.

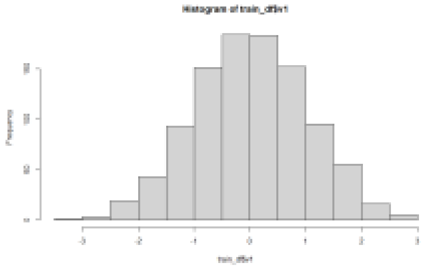


Fig-2(a): Data Distribution of V1

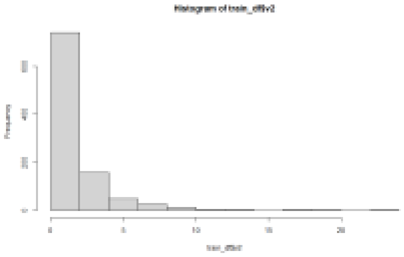


Fig-2(b): Data Distribution of V2

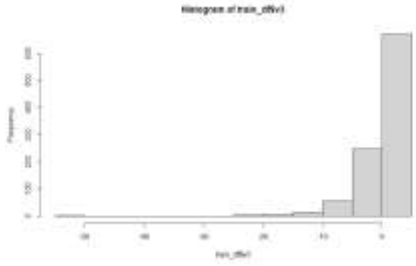


Fig-2(c): Data Distribution of V3

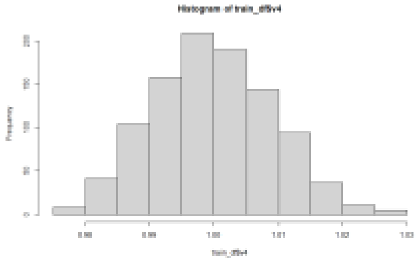


Fig-2(d): Data Distribution of V4

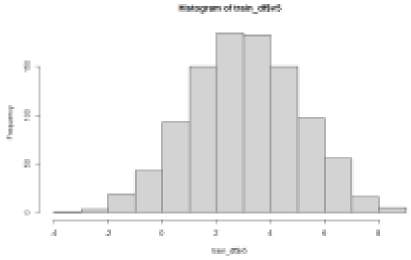


Fig-2(e): Data Distribution of V5

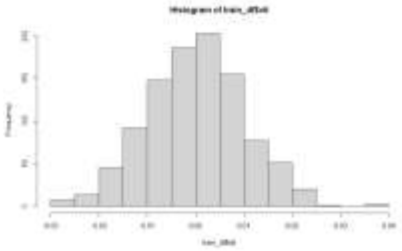


*Fig-2(f): Data Distribution of V*6

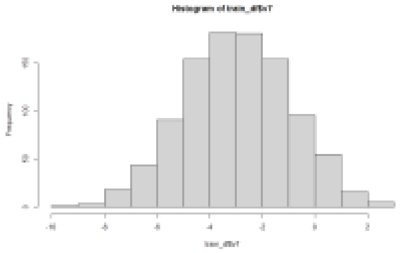


Fig-2(g): Data Distribution of V7

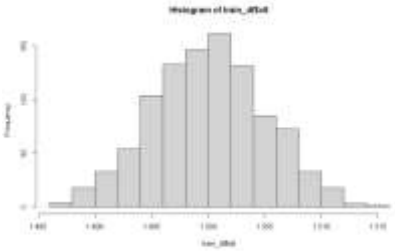


Fig-2(h): Data Distribution of V8

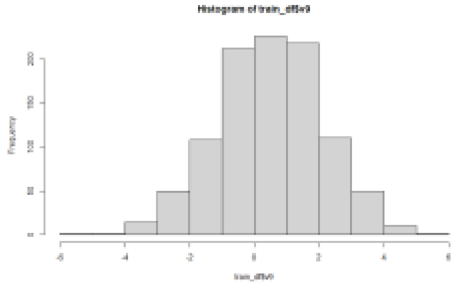


Fig-2(i): Data Distribution of V9

- **Boxplot:** Then from the boxplot was plotted to visualize the data distribution and analyze its quartile values.

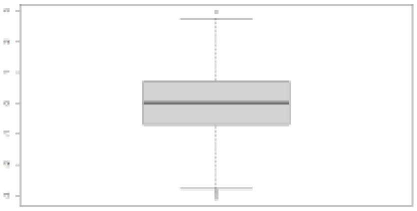


Fig-3(a): Box Plot for V1

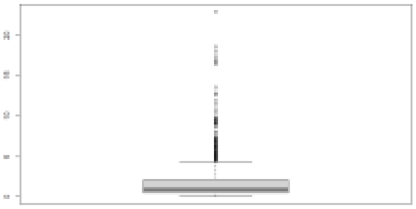


Fig-3(b): Box Plot for V2

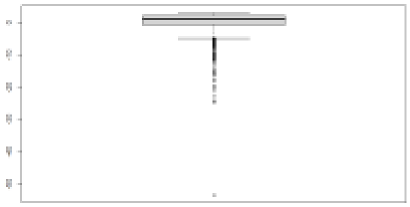


Fig-3(c): Box Plot for V3

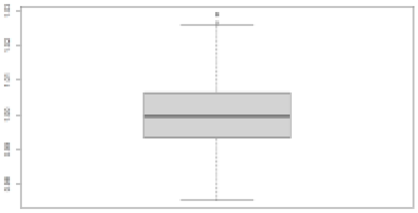


Fig-3(d): Box Plot for V4

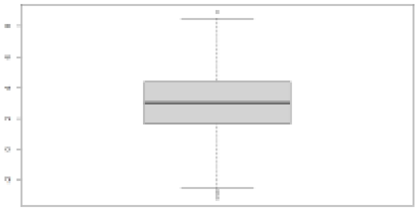


Fig-3(e): Box Plot for V5

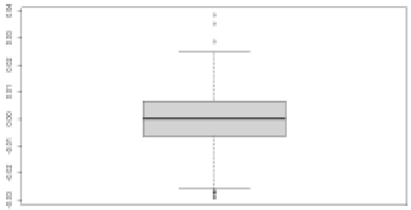


Fig-3(f): Box Plot for V6

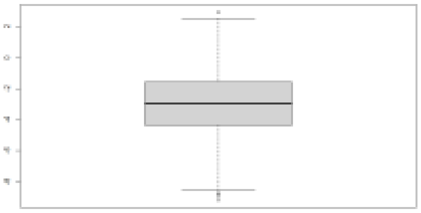


Fig-3(g): Box Plot for V7

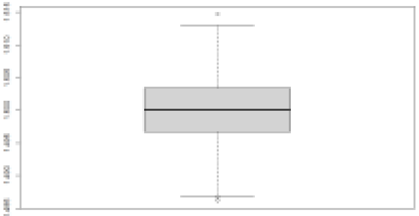


Fig-3(h): Box Plot for V8

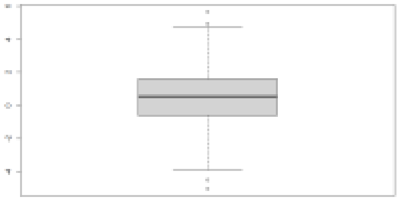


Fig-3(i): Box Plot for V9

- **Bi-variate Analysis:** Finally the pair plot was drawn to find out any outliers.

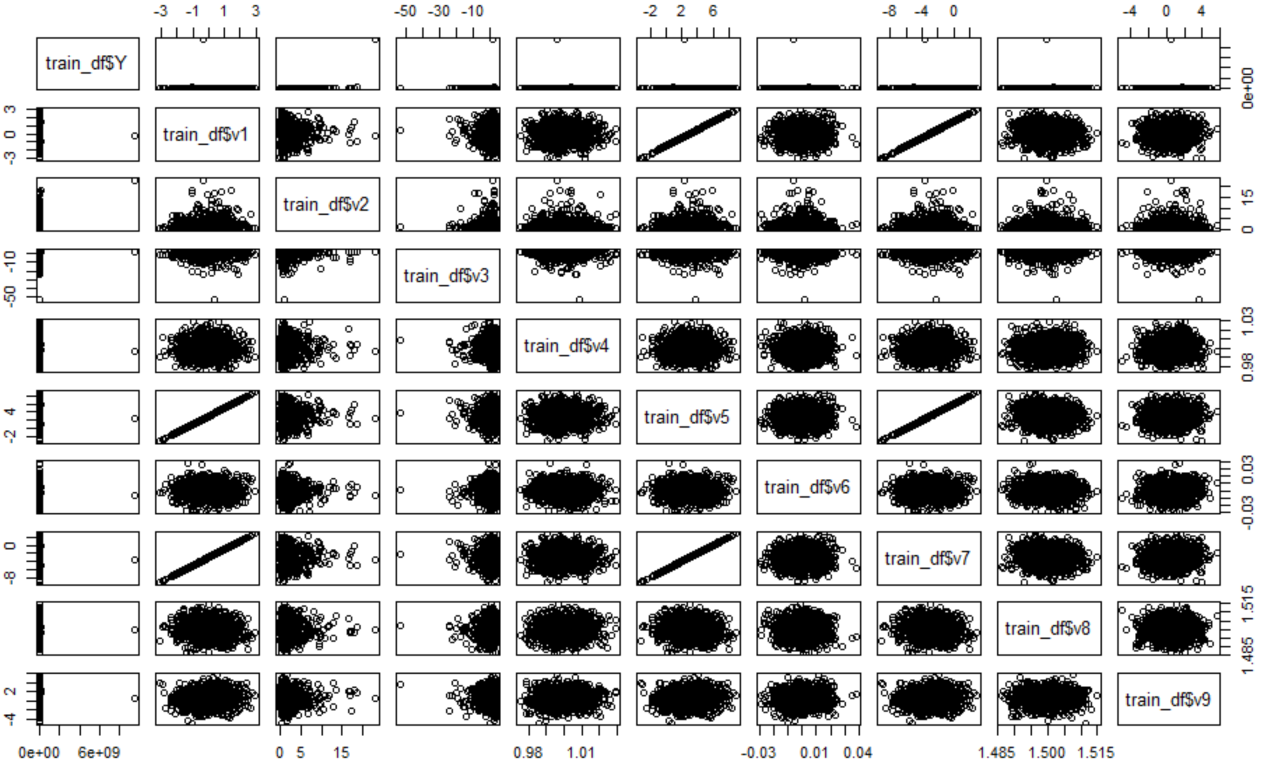
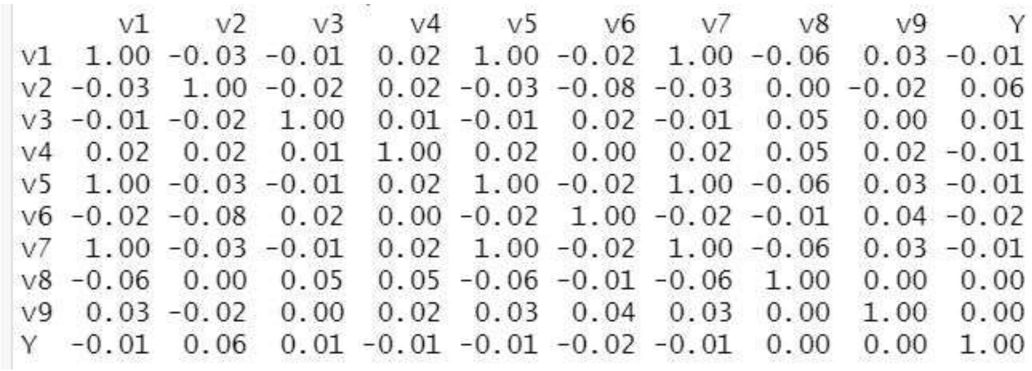


Fig-4: Bi-variate result between all the variables in the dataset for finding the outliers.

Correlation Analysis: Correlation Analysis was also conducted to see the relation between the predictor variables and the target variable.



```
      v1    v2    v3    v4    v5    v6    v7    v8    v9     Y
v1  1.00 -0.03 -0.01  0.02  1.00 -0.02  1.00 -0.06  0.03 -0.01
v2 -0.03  1.00 -0.02  0.02 -0.03 -0.08 -0.03  0.00 -0.02  0.06
v3 -0.01 -0.02  1.00  0.01 -0.01  0.02 -0.01  0.05  0.00  0.01
v4  0.02  0.02  0.01  1.00  0.02  0.00  0.02  0.05  0.02 -0.01
v5  1.00 -0.03 -0.01  0.02  1.00 -0.02  1.00 -0.06  0.03 -0.01
v6 -0.02 -0.08  0.02  0.00 -0.02  1.00 -0.02 -0.01  0.04 -0.02
v7  1.00 -0.03 -0.01  0.02  1.00 -0.02  1.00 -0.06  0.03 -0.01
v8 -0.06  0.00  0.05  0.05 -0.06 -0.01 -0.06  1.00  0.00  0.00
v9  0.03 -0.02  0.00  0.02  0.03  0.04  0.03  0.00  1.00  0.00
Y  -0.01  0.06  0.01 -0.01 -0.01 -0.02 -0.01  0.00  0.00  1.00
```

Fig-5: Correlation Analysis for determining the linear relationship between the predictor and the target variable.

Decision Inferred from Data Analysis:

Skewness: From the summary's Max-Min, Quartile values, and the Histogram analysis, it is clear that variables V1, V4, V5, V6, V7, V8, V9 are almost normally distributed. The distribution of variable V2 is skewed to the left side (Minimum value side) and the distribution of variable V3 is skewed to the right side (Maximum value side).

Outliers: From the boxplot and the Bi-variate analysis, it is clear that variable V2 has an outlier at the maximum value side and variable V3 has an outlier at the Minimum value side.

Linear Relationship: From the Correlation Analysis, it is clear that none of the predictor variables is linearly related to the target variable.

Outlier Treatment:

V2: To remove the outliers at the maximum value side, the value of the 1.5 times of the 80th quantile value has been taken as the distribution's upper limit. Thus, the difference between the mean and the median was possible to minimize without significantly altering the data in variable, V2.

V3: To remove the outliers at the minimum value side, the value of the 0.1 times of the 5th quantile value has been taken as the distribution's lower limit. Thus, the difference between the mean and the median was possible to minimize without significantly altering the data in variable, V3.

Linear Regression:

Though none of the predictor variables are linearly related to the target variable, the Linear Regression has technique has been conducted on the training dataset. And, as expected, the model's output was not good or unacceptable.

KNN Regression:

In the training dataset, we have the data for 1000 subjects. So, the seed value was selected as,

The seed value, $K = \sqrt{n} = \sqrt{1000} \approx 31$.

To perform the KNN regression, the IBk() function from the RWeka package has been used. It is a package of a collection of a bunch of Machine Learning and Data Mining operations such as Data Pre-processing, Classification, Regression, Clustering, Association Rules, and Visualization. This package has been developed/ written in Java. So, to use this package rJava package will also be needed.

The .Rdata File Generation:

In the .Rdata file there are four variables available,

fit = The output of the Linear Regression model.
KNN_model = The output of the KNN Regression model.
lm_pred = The predicted values of the target variable Y by the Linear Regression model.
KNN_pred = The predicted values of the target variable Y by the KNN Regression model.