REPORT

Insights:

- Given dataframe has 'Class' column which has only two outputs "0"(Tuple is NOT Fraudulent) and "1"(Tuple is Fraudulent).
- Column 'Time' denoting the number of seconds elapsed between this transaction and the first transaction is always increasing.
- Colums V1 to V28 contains values which are neither NULL nor FLOAT.

Inferences:

- The Standard Deviation of the Column 'Time' is almost 50000 times the Standard Deviation of other Columns, so we have to remove this column from our dataframe because if we have included this Column than it would have dominated the other columns features, and reducing the accuracy.
- In data pre-processing, Normalisation is not done. In Normalisation every column gets into one range, and in our dataframe, if we do Normalisation, the column 'Amount' having vast range before Normalisation will get less dominance over other columns features. And we know that the 'Amount' column should be of the heighest dominant feature in detecting the Fraudulent.

Result :

1. Yes, we need data pre-proccessing because there are many tuples in our dataframe which have values neither NULL nor FLOAT and many tuples have duplicate values, so in data cleaning I have removed all these datasets.
2. No, we don't need Normalisation in this dataframe, because it would have decreases the dominant feature of 'Amount' column in detecting the Fraud, since all columns would have same range.
3. Comparing clustering result with actual result using:
    a. Root Mean Squared Error: In this I got 0.1454.
    b. Correlation between class and predicted column : [[1. , 0.0150] , [0.0150 , 1.]].
4. Kmeans clustering algorithm gives us around 98 percent accuracy, and this is the closest I could get to the actual result. There are two more clustering algorithms DBSCAN and Agglomerative clustering which gives less accuracy. Theoritically also, we can infer that DBSCAN is not useful for this dataframe. we know that there will be only two clusters, since the 'Class' column has only two values, and we use DBSCAN when we don't have any information about the number of clusters of the dataframe. So DBSCAN shouldn't be used. Similarly I can infer for Agglomerative clustering algorithm.

Conclusion:

I have run a clustering algorithm on the data provided by the top banks of the country. With the help of k-means algorithm, I have clustered our data in two datasets. I have also tried the

Normalization using min-max and z-score algorithm on the dataset but the accuracy got reduced which was due to scaling of ‘Amount’ column. Finally I got result with 97.88 percent accuracy.