CLUSTERING ASSIGNMENT REPORT

- Imported the libraries and loaded the dataset
- Noted down the basic information of the data and found out what type of data we are dealing
- As everything is for 12 months some attributes were renamed
- No duplicates were found in the dataset
- Dropped the unnecessary columns i.e. ID#
- Obtained the statistical summary of the dataset's numerical values
- Went on with histogram visualization analysis
- Then inserted the kde function to clearly identify the distribution of the data per attribute
- Done with the box plot and as the data is raw and collected from the airlines the outliers are actual data and needed to be there for the clustering
- Plotted the Swarm Plot for the plotting of data to identify if cluster are possible to be formed
- Went on with the Heatmap visualization to identify the correlation between the attributes and what can be possibles clustered together
- Transformed the data using standard scaler method
- Plotted pair plot as to see the visualization between each attribute and its data
- Removed the null values using Z score method
- Went on with clustering the data using the three modes, K Means, Hierarchical, DB SCAN
- With the K means clustering and elbow method we found that they can be clustered into 3 clusters
- With hierarchical clustering only 2 clusters are formed, tried with different linkage methods and finalized ward's method to be accurate for the cluster
- And with DB scan most of the data went to be noise and the cluster was not clearly formed
- And plotted the data as per cluster separately and also grouped them by their labels and interpretated the characteristics of the clusters and also evaluated the clusters using Silhouette score