**KMeans and DBscan Clustering**

**Instructions:**

Please share your answers wherever applicable in line with the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: MD SABIULLAH Batch ID:** PDS 07052024

**Topic: KMeans and DBscan Clustering**

**Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) is submitted along with the documentation, explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered a correct submission.**

**2. Ensure that you submit your assignments correctly and in full. Resubmission is not allowed.**

**3. Post the submission you can evaluate your work by referring to the keys provided. (Will be available only post the submission).**

**Hints:**

**1. Business Problem**

* 1. **What is the business objective?**
  2. **What are the constraints?**
  3. **Define success criteria**

**2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Table

Description automatically generated

[ SOLUTION ] :

| **Column Name** | **Column Description** | **Data Type** | **Range of Values** | **Min Value** | **Max Value** |
| --- | --- | --- | --- | --- | --- |
| **Activity Period** | **Numeric representation of the period of activity** | **int64** | **Represents different time periods** | **Varies** | **Varies** |
| **Operating Airline** | **Name of the airline operating the flight** | **object** | **Textual data (airline names)** | **N/A** | **N/A** |
| **Operating Airline IATA Code** | **IATA code representing the operating airline** | **object** | **3-letter codes (e.g., 'AA', 'DL')** | **N/A** | **N/A** |
| **GEO Region** | **Geographical region of the airline's operations** | **object** | **Various regions (e.g., North America)** | **N/A** | **N/A** |
| **Terminal** | **Airport terminal used by the airline** | **object** | **Different terminal names** | **N/A** | **N/A** |
| **Boarding Area** | **Specific boarding area within the terminal** | **object** | **Different boarding areas** | **N/A** | **N/A** |
| **Passenger Count** | **Number of passengers for a given period** | **int64** | **0 to varying high values** | **0** | **Varies** |
| **Year** | **Year of the recorded activity** | **int64** | **2010s - 2020s** | **Varies** | **Varies** |
| **Month** | **Month of the recorded activity** | **object** | **Month names (e.g., 'January', 'February')** | **N/A** | **N/A** |

**3. Exploratory Data Analysis (EDA):**

**3.1. Univariate analysis.**

**3.2. Bivariate analysis.**

**4. Data Pre-processing**

**4.1 Data Cleaning, Feature Engineering, etc.**

**5. Model Building**

**5.1 Build the model on the scaled data (try multiple options).**

**5.2 Perform the KMeans and DBscan clustering and find out the best model that minimizes Within the Sum of Squares. Compare the result with Hierarchical Clustering methods.**

**5.3 Validate the clusters (try with the different numbers of clusters), label the clusters, and derive insights (compare the results from multiple approaches).**

**6. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**7. Deploy the best model using Python Flask on the local machine.**

**Problem Statements:**

Global air travel has seen an upward trend in recent times. The maintenance of operational efficiency and maximizing profitability are crucial for airlines and airport authorities. Businesses need to optimize airline and terminal operations to enhance passenger satisfaction, improve turnover rates, and increase overall revenue.

The airline companies with the available data want to find an opportunity to analyze and understand travel patterns, customer demand, and terminal usage.

**CRISP-ML(Q) process model describes six phases:**

1. Business and Data Understanding

2. Data Preparation

3. Model Building

4. Model Evaluation

5. Deployment

6. Monitoring and Maintenance

**Objective**: Maximize the Sales

**Constraints**: Minimize the Customer Retention

**Success Criteria:**

Business Success Criteria: Increase the Sales by 10% to 12% by targeting cross-selling opportunities on current customers.

ML Success Criteria: Achieve a Silhouette coefficient of at least 0.6

Economic Success Criteria: The insurance company will see an increase in revenues by at least 8%

Data: Refer to the ‘AirTraffic\_Passenger\_Statistics.csv’ dataset.

**Questions to Trigger Your Thoughts:**

Q1. What is sklearn.preprocessing is mostly used for?

Q2. What are three scaling functions that can be imported from sklearn.preprocessing?

Q3. What library does the groupby function belong to?

Q4. How to save an output file?

Q5. How to perform scaling without using inbuilt functions?

Q6. What is the function used to reverse (inverse) the scaling function?

Q7. How are info() and describe() functions different?

Q8. How to write the X-axis label?

Q9. How to visually see more row labels when they are overlapped by one another?

Q10. Why plot elbow or scree plot to define the number of clusters?

Q11. What inbuilt Python function is used to label the clusters?

Q12. Is it important to scale data before clustering it?

Q13. Which libraries are used to implement DBscan clustering?

Q14. What is the DBscan algorithm?

Q15. What are the parameters used in DBscan?

Q16. How to perform cluster evaluation? Which are the techniques used for cluster evaluation?

Q17. Which attribute gives labels for the clusters?

Q18. How to visualize the clusters?

Q19. What are the files required for flask deployment?

Q20. What are the necessary libraries for flask?

Q21. Apart from Pickle which library we can use for saving and loading the model?

Q22. Why we are using @app.route?

Q23. Why are we using ‘/’ this inside the root?

Q24. What is GET & POST request I flask?

Q25. How do you collect the input data in the flask?

Q26. How would you call an HTML file in Flask?

Q27. Why do we use debug= true?