**Hierarchical 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: \_\_\_\_\_\_\_\_\_\_\_\_\_ Batch ID:** \_\_\_\_\_\_\_\_\_\_\_

**Topic: Hierarchical 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:**



**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 hierarchical clustering and visualize the clusters using a dendrogram.**

**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?**

**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 operational efficiency

**Constraints**: Maximize the financial health

**Success Criteria:**

Business Success Criteria: Increase the operational efficiency by 10% to 12% by segmenting the Airlines.

ML Success Criteria: Achieve a Silhouette coefficient of at least 0.7

Economic Success Criteria: The airline companies will see an increase in revenues by at least 8% (hypothetical numbers)

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

**Questions to Trigger Your Thoughts:**

Q1. Why it is important to define the objectives for any Business problem?

Q2. How to maintain the quality of the Machine Learning model developed for the Business problem?

Q3. What is the first document created/drafted for any ML project?

Q4. How to load data with multiple sheets?

Q5. What are the Auto EDA techniques?

Q6. What are four business moments, and what insights we can draw from them?

Q7. Write the techniques in data Pre-Processing.

Q8. When we use label encoding and one-hot encoding?

Q9. What is the technique to remove outliers?

Q10. What are the techniques to check whether the data is normally distributed or not?

Q12. How to make data scale-free?

Q13. **What types of graphs are used to depict the bivariate analysis?**

Q14. **What do you mean by bivariate frequency distribution?**

Q15. Which libraries are used in Hierarchical clustering?

Q16. What is the difference between Agglomerative clustering and Divisive Clustering?

Q17. Which metric is used to find distance/similarities between two data points and between a record and a cluster?

Q18. What are the parameters needed to plot the Dendrogram?

Q19. How to perform cluster evaluation? Which are the techniques used for cluster evaluation?

Q20. **What do the Silhouette coefficient, Calinski Harbaz, and Davies-Bouldin Index indicate in hierarchical clustering?**