Unsupervised Learning

Project - 60 Marks

General Instructions:

1. Submission of all the parts is expected in 1 notebook only
2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only
3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.
4. If output for any code cell is missing, 50% marks will be deducted.

Domain: Automobile

Context:

The data concerns city-cycle fuel consumption in miles per gallon to be predicted in terms of 3 multivalued discrete and 5 continuous attributes.

Data Description:

The data concerns city-cycle fuel consumption in miles per gallon.

1. mpg: continuous
2. cylinders: multi-valued discrete
3. displacement: continuous
4. horsepower: continuous
5. weight: continuous
6. acceleration: continuous
7. model year: multi-valued discrete
8. origin: multi-valued discrete
9. car name: string (unique for each instance)

Project Objective:

To understand K-means Clustering by applying on the Car Dataset to segment the cars into various categories.

- **Steps and Tasks:**

1. **Data Understanding: 10marks**
   a. Read 'Car name.csv' as a DataFrame and assign it to a variable. [1 Mark]
   b. Read 'Car-Attributes.json as a DataFrame and assign it to a variable. [1 Mark]
   c. Merge both the DataFrames together to form a single DataFrame [2 Mark]

d. Print 5 point summary of the numerical features and share insights. [1 Marks]

2. **Data Preparation and Analysis: 20marks**
   a. Check and print feature-wise percentage of missing values present in the data and impute with the best suitable approach. [2 Mark]
   b. Check for duplicate values in the data and impute with the best suitable approach. [1 Mark]
   c. Plot a pairplot for all features. [1 Marks]
   d. Visualize a scatterplot for 'wt' and 'disp'. Datapoints should be distinguishable by 'cyl'. [1 Marks]
   e. Share insights for Q2.d. [1 Marks]
   f. Visualize a scatterplot for 'wt' and 'mpg'. Datapoints should be distinguishable by 'cyl'. [1 Marks]
   g. Share insights for Q2.f. [1 Marks]
   h. Check for unexpected values in all the features and datapoints with such values. [2 Marks]
   [Hint: '?' is present in 'hp']

3. **Clustering: 30marks**
   a. Apply K-Means clustering for 2 to 10 clusters. [3 Marks]
   b. Plot a visual and find elbow point. [2 Marks]
   c. On the above visual, highlight which are the possible Elbow points. [1 Marks]
   d. Train a K-means clustering model once again on the optimal number of clusters. [3 Marks]
   e. Add a new feature in the DataFrame which will have labels based upon cluster value. [2 Marks]
   f. Plot a visual and color the datapoints based upon clusters. [2 Marks]
   g. Pass a new DataPoint and predict which cluster it belongs to. [2 Marks]

Submission Format:

1. .ipynb *(Jupyter Notebook)* **and**
2. .html *(Jupyter Notebook > File > Download as > HTML)*

5 Marks will be deducted if submission in any of the formats is missing.
