**Clustering Report: EV Dataset Analysis**

**INTRODUCTION**

The objective of this project is to perform clustering on an Electric Vehicle (EV) dataset to identify distinct patterns in sales and distribution across different states and vehicle categories. Clustering helps in segmenting the data into meaningful groups based on similarities.

**DATASET OVERVIEW**

The dataset contains the following features:

* **Year**: The year of sales data.
* **Month\_Name**: The month in which the sales occurred.
* **Date**: Specific date of entry.
* **State**: The region where sales occurred.
* **Vehicle\_Class**: The classification of the vehicle (e.g., bus, ambulance, tractor).
* **Vehicle\_Category**: Higher-level categorization of vehicles.
* **Vehicle\_Type**: The type of electric vehicle.
* **EV\_Sales\_Quantity**: The count of electric vehicle sales.

**TECHNIQUES USED**

**------DATA PREPROCESSING**

To prepare the dataset for clustering, the following steps were performed:

* **Handling Categorical Data**: Since clustering algorithms require numerical input, categorical variables (State, Vehicle\_Class, Vehicle\_Category, Vehicle\_Type) were encoded using **Label Encoding**.
* **Feature Selection**: Selected relevant numerical and categorical features for clustering.
* **Feature Scaling**: Standardized the numerical data using **StandardScaler** to ensure all features contribute equally to the clustering process.

**------CLUSTERING METHOD: K-MEANS**

The **K-Means Clustering** algorithm was used for segmentation. The key steps involved:

* **Finding the Optimal Number of Clusters**: The **Elbow Method** was used to determine the best value for K by plotting inertia (sum of squared distances from points to their assigned cluster center) against different K values.
* **Applying K-Means**: Once the optimal K was identified, K-Means was applied to partition the dataset into clusters.
* **Cluster Assignment**: The dataset was updated with a new column, Cluster, indicating the assigned group.

**KEY FINDINGS**

* The dataset was successfully clustered into K groups (e.g., 4 clusters based on the elbow method).
* Clusters revealed meaningful patterns related to EV sales across different states and vehicle types.
* The pairplot visualization provided insights into how features such as Year and EV\_Sales\_Quantity contributed to cluster formation.

**VISUALIZATION**

* **Elbow Curve**: Helped identify the optimal number of clusters.
* **Pairplot Analysis**: Showed how clusters differ based on selected numerical features.

**CONCLUSION & INSIGHTS**

The clustering approach successfully grouped EV sales data into meaningful segments. This segmentation can help policymakers, manufacturers, and marketers better understand the distribution and trends in electric vehicle sales across regions and categories.

**Insights:**

* EV sales exhibit regional patterns, with certain states having higher adoption rates.
* Specific vehicle types, such as buses and commercial EVs, form distinct clusters.
* Temporal trends indicate an increasing trend in EV adoption over the years.

**FUTURE IMPROVEMENTS & MARKET SEGMENTATION ENHANCEMENTS**

**------ENHANCEMENTS WITH ADDITIONAL TIME & BUDGET:**

If given additional time and budget, improvements to the market segmentation project could include:

* **Datasets Collection**: Adding columns such as Charging Infrastructure, Government Incentives, EV Battery Capacity, and Customer Demographics.
* **Additional ML Models**:
  + **DBSCAN** for detecting anomalies in EV sales patterns.
  + **Hierarchical Clustering** to create a detailed segmentation tree.
  + **Gaussian Mixture Models (GMM)** for soft clustering (probabilistic segmentation).

**-------ESTIMATED MARKET SIZE:**

The estimated market size for the **EV domain** (non-segmented) is expected to reach **$1.1 trillion by 2030**, with an annual growth rate of **23% CAGR** globally.

**-------TOP 4 FEATURES FOR OPTIMAL MARKET SEGMENTATION:**

1. **State** - Geographic location plays a major role in EV adoption.
2. **Vehicle\_Type** - Different types of EVs appeal to different customer segments.
3. **EV\_Sales\_Quantity** - Sales volume determines demand intensity.
4. **Year** - Trends in adoption over time help in forecasting future growth.