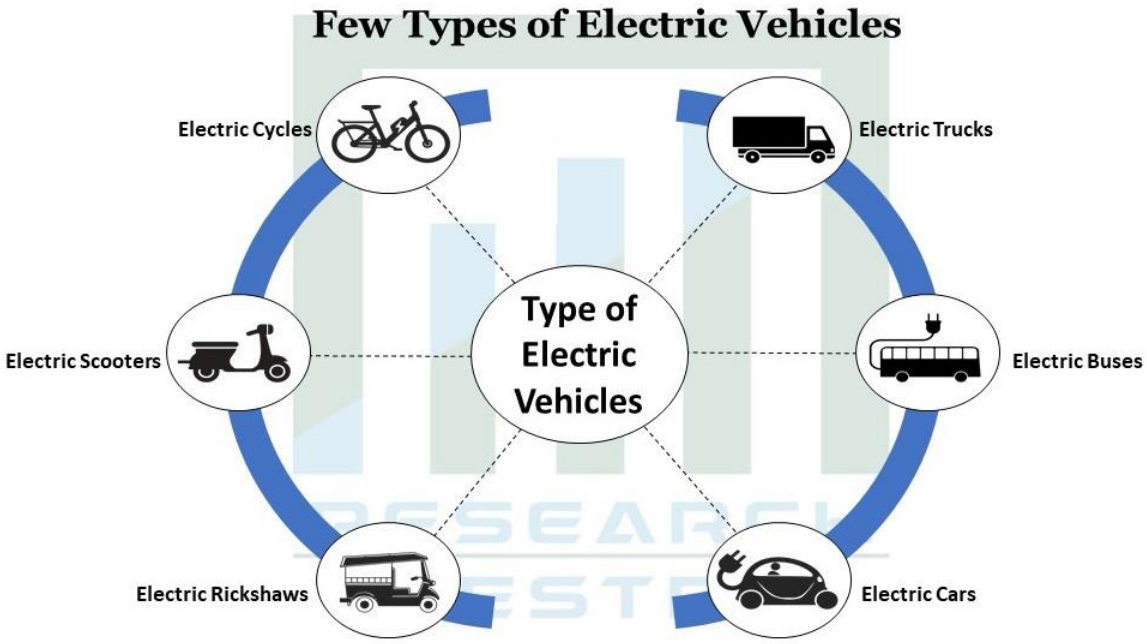
Electric Vehicle Market Segment Analysis: A Data-Driven Approach

Submitted by: Gyanpriya Misra
GitHub Link

Introduction:

The Electric Vehicle (EV) market has grown rapidly in recent years, driven by advancements in technology and increasing consumer demand for sustainable transportation options. As the EV ecosystem expands, it's essential to understand the various segments within the market to enable effective decision-making and strategy formulation. This report provides an analysis of the EV market based on multiple datasets related to charging infrastructure, vehicle types, and operational public charging stations (PCS), aiming to derive actionable insights.



Source: Research Nester

Market Segmentation

Market segmentation is a strategic approach that involves dividing a broad target market into smaller, more defined groups of consumers who share similar characteristics, needs, or behaviours. This process allows businesses to tailor their products, services, and marketing efforts to meet the specific demands of different segments effectively.

There are several bases for segmenting a market, including:

1. **Demographic Segmentation**: Dividing the market based on demographic factors such as age, gender, income, education, and family size.
2. **Geographic Segmentation**: Segmenting the market according to geographic location, such as region, city, or neighborhood, which can influence purchasing behaviours.
3. **Psychographic Segmentation**: Grouping consumers based on psychological factors, including lifestyle, values, interests, and personality traits.

4. **Behavioral Segmentation**: Analysing consumer behavior, including purchasing habits, brand loyalty, and usage rates, to create segments that reflect actual consumer engagement.

By employing market segmentation, businesses can identify niche markets, enhance customer satisfaction, and optimize marketing resources, ultimately leading to increased profitability and competitive advantage. It enables companies to focus their efforts on the most promising segments and to develop targeted strategies that resonate with specific groups of customers.

Data Collection and Sources:

Four primary datasets were collected for this analysis:

1. **EV Charging Station Dataset**:
   Contains information on various EV charging stations, including their region, type, power, and services offered.
   - **Source**: source_link
   - **Shape**: (202 rows, 9 columns)
2. **Vehicle Class Dataset**:
   Provides data on different vehicle classes and their total registration numbers.
   - **Source**: source_link
   - **Shape**: (16 rows, 2 columns)
3. **Operational Public Charging Stations (PCS) Dataset**:
   Details the number of operational public charging stations across different states.
   - **Source**: source_link
   - **Shape**: (34 rows, 2 columns)
4. **EV Final Dataset**:
   Comprehensive dataset featuring information about various EV stations, including capacity, vehicle type, payment modes, and station types.
   - **Source**: source_link
   - **Shape**: (2705 rows, 25 columns)

Data Cleaning:

To ensure data quality and consistency, the following cleaning steps were performed:

1. **EV Charging Station Dataset**:
    - **Dropped irrelevant columns**: Removed 'aux address', 'latitude', and 'longitude', as these were redundant or irrelevant for analysis.
    - **Checked for missing values**: No missing values found in the cleaned dataset.
2. **Vehicle Class Dataset**:
    - No irrelevant columns identified for removal.
    - **Checked for missing values**: No missing values found.
3. **Operational PCS Dataset**:
    - No irrelevant columns identified for removal.
    - **Checked for missing values**: No missing values found.
4. **EV Final Dataset**:
    - **Dropped irrelevant columns**: Removed 'latitude', 'longitude', 'logo_url', and an unnamed '0' column.
    - **Filled missing values**: Missing values in columns like 'zone', 'available', 'capacity', 'cost_per_unit', etc., were replaced with the string 'Unknown' to maintain consistency.
    - **Checked for duplicate rows**: None found.
    - **Ensured proper formats**:
        - Converted 'No. of Operational PCS' to numeric format.
        - Corrected 'open' and 'close' time formats to ensure consistency.

About Datasets:

1. EV Charging Station Dataset (cleaned dataset)

- **Count of Charging Stations**: The dataset consists of 202 charging stations.
- **Mean 'no'**: The average index value is 101.5, suggesting an evenly distributed sequence of station identifiers.
- **Standard Deviation**: A value of 58.46 shows some variation in the data, but not extreme, indicating that station identifiers are fairly evenly spaced.
- **Range**: The minimum identifier is 1, and the maximum is 202, confirming the station entries are sequentially numbered without missing values.

This dataset does not have missing values and provides essential details on each station's type, power, and services offered. Removing unnecessary fields has simplified the dataset, making it ready for analysis of regional distribution and station types.

2. Vehicle Class Dataset

- **Unique Vehicle Classes**: The dataset includes 16 unique vehicle classes, representing a wide range of EV types in the market.
- **Top Vehicle Class**: "FOUR WHEELER (INVALID CARRIAGE)" has the highest registration count at 21,346, suggesting this class dominates registrations in the market.

The vehicle class dataset covers different EV categories with equal distribution, meaning each class is distinct and equally represented. This data is crucial for understanding the diversity of vehicles using public charging infrastructure.

3. Operational Public Charging Stations Dataset

- **Count of States with Public Charging Stations**: 34 states or regions have operational public charging stations, showing the geographical spread of the infrastructure.
- **Mean**: On average, each state has 357 operational PCS, though this varies greatly.
- **Standard Deviation**: A high standard deviation of 617.58 indicates significant variation in the number of charging stations per state.
- **Range**: The minimum number of stations in a state is 1, while the maximum is 3,079, highlighting the disparity in infrastructure between different regions.

This dataset reveals uneven distribution of charging infrastructure across states, with some states having far fewer stations than others. This information could be critical for identifying regions that require infrastructure development.

4. EV Final Dataset (cleaned dataset)

- **Open/Close Time**: Most charging stations in this dataset open at 00:00:00 and close at 23:59:59, suggesting that many stations operate 24 hours. However, some stations show open and close times outside this range, with the longest recorded open time being 10 hours.
- **Postal Code Distribution**: Postal codes vary widely, from 0 to 1.1 million, which could indicate data inconsistencies or a wide geographical range of stations.
- **Missing Data**: Several columns, such as "zone", "available", "capacity", and "cost_per_unit", have missing values that were replaced with "Unknown" to maintain the dataset's usability in further analysis.

The EV final dataset provides detailed information on charging station operations, but the high number of missing values for key attributes such as station capacity and cost per unit suggests gaps in reporting, which might limit the accuracy of some analyses. However, it still offers valuable insights into operational hours, station availability, and regional data through postal codes.

- ✓ ***Infrastructure Disparities****: There's a significant difference in the number of operational charging stations across regions, with some states having far more than others, which could influence access to EV infrastructure.*
- ✓ ***Vehicle Diversity****: The market includes a wide variety of vehicle classes, with a strong presence of four-wheeled electric vehicles.*
- ✓ ***Operational Consistency****: Most charging stations operate continuously (24 hours), suggesting that the infrastructure is designed to meet demand at all times.*

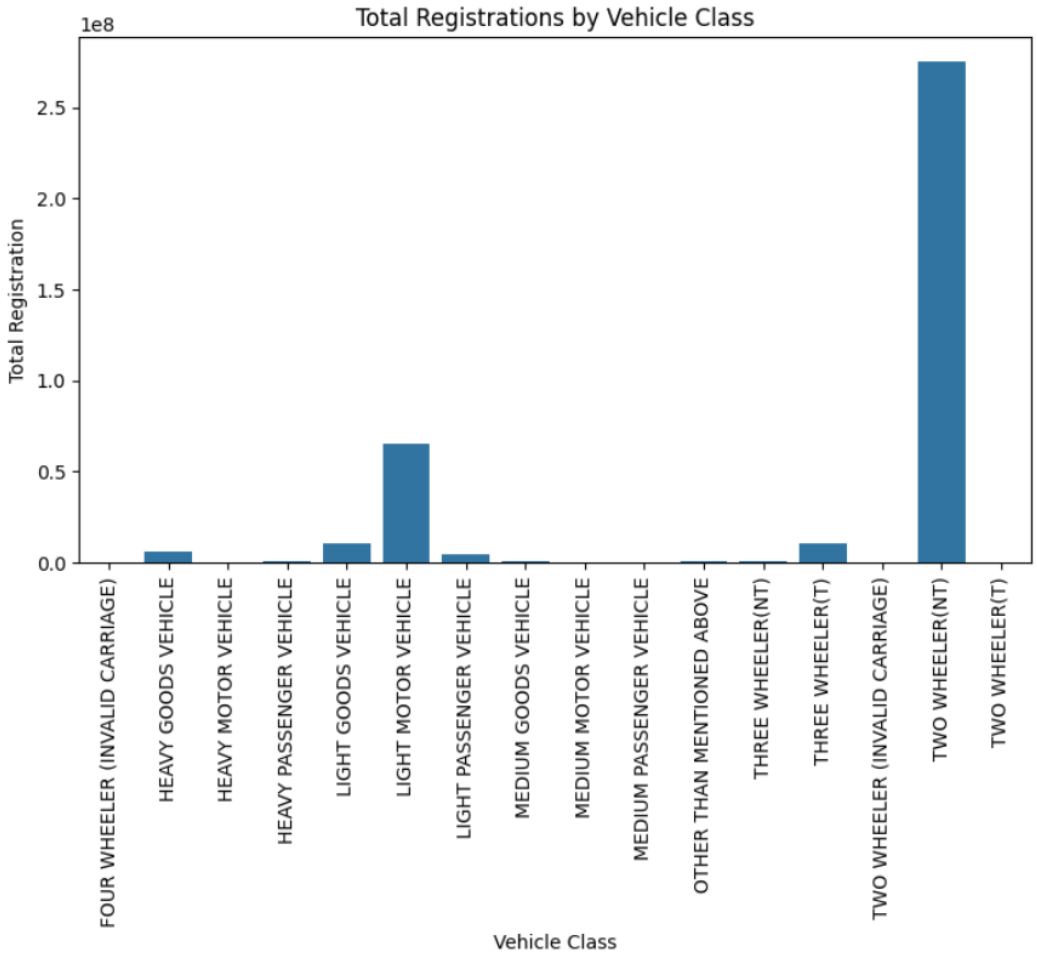
Exploratory Data Analysis (EDA)

Vehicle Class and Total Registrations

The analysis of vehicle classes and their total registrations provides a clear view of the dominant vehicle segments in the market. The data reveals that:

- **Two-Wheelers (NT)** dominate the market with a total registration of **274.97 million**, making it the largest registered vehicle category by a significant margin.
- **Light Motor Vehicles (LMV)** are the second most registered vehicle type, with **65.06 million** registrations.
- **Three-Wheelers (T)** follow with **10.70 million** registrations, demonstrating their prevalence in certain segments of the transport market.
- **Light Goods Vehicles (LGV)** and **Heavy Goods Vehicles (HGV)** have smaller but notable shares, with **10.25 million** and **5.87 million** registrations respectively.

This indicates that electric vehicle initiatives and infrastructure planning should prioritize Two-Wheelers and Light Motor Vehicles due to their significant presence.

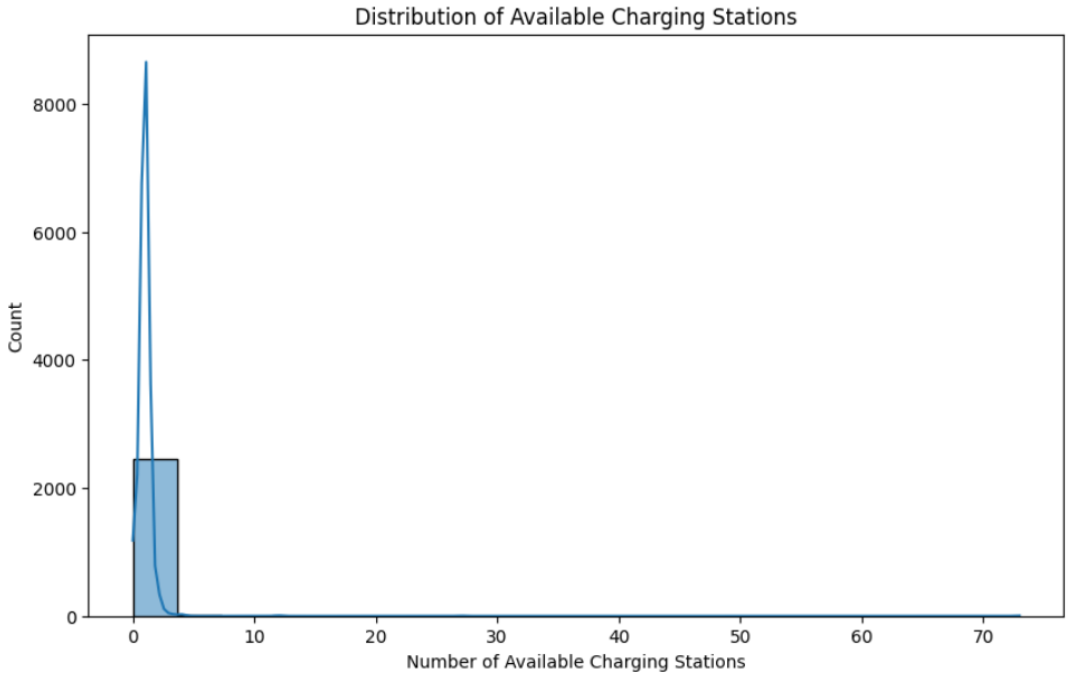


Charging Station Availability

The availability of charging stations is crucial for the adoption and growth of electric vehicles. The data on the number of available charging stations shows:

- A total of **2,467** charging stations were analysed.
- The average number of available charging stations is approximately **1 station per location**, with a standard deviation of **1.63**, indicating some variance across locations.
- The maximum availability at any given location is **73 charging stations**, but the majority of locations (75%) have **1 charging station**.

This suggests that most locations offer minimal charging capacity, highlighting the need for expansion in charging infrastructure to support a growing EV market.

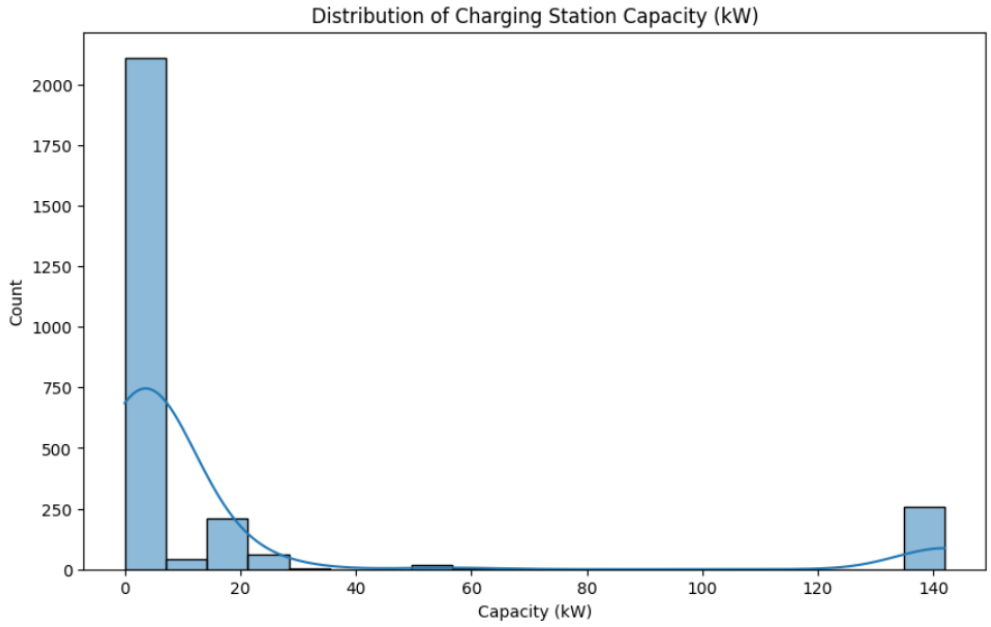


Charging Station Capacity

The capacity of charging stations, measured in kilowatts (kW), reflects their potential to serve vehicles efficiently. The analysis reveals:

- The average charging station capacity is **18.2 kW**, with a wide standard deviation of **40.7 kW**, indicating large variability in station capacity.
- The median and 75th percentile capacity is **3.3 kW**, suggesting that most stations offer relatively low charging speeds.
- The maximum observed capacity is **142 kW**, which reflects a few high-capacity stations designed for fast charging.

These figures suggest that while some high-capacity stations exist, most charging points are on the lower end of the spectrum, which may impact charging times and convenience for EV users.

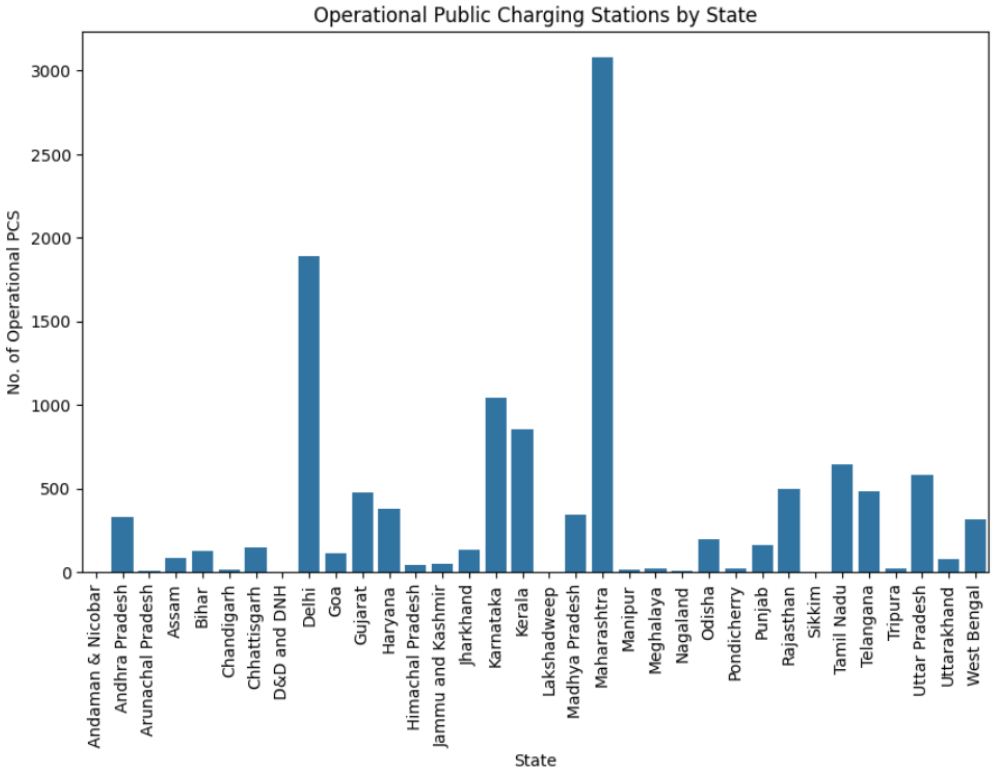


Operational Public Charging Stations by State

The distribution of operational public charging stations across states highlights the regions leading in EV infrastructure. The top five states are:

- **Maharashtra** leads significantly with **3,079** operational public charging stations.
- **Delhi** follows with **1,886** stations, showing strong EV infrastructure in the capital.
- **Karnataka** and **Kerala** have **1,041** and **852** operational stations respectively.
- **Tamil Nadu** rounds out the top five with **643** stations.

This indicates a strong regional focus on electric vehicle infrastructure, particularly in the western and southern parts of India, with Maharashtra being the frontrunner.



Station Types and Power Types

The types of charging stations and power sources provide insights into the available infrastructure:

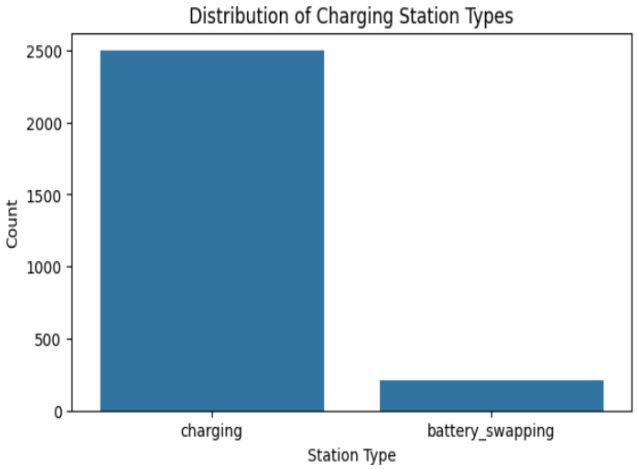
5.1 Station Types

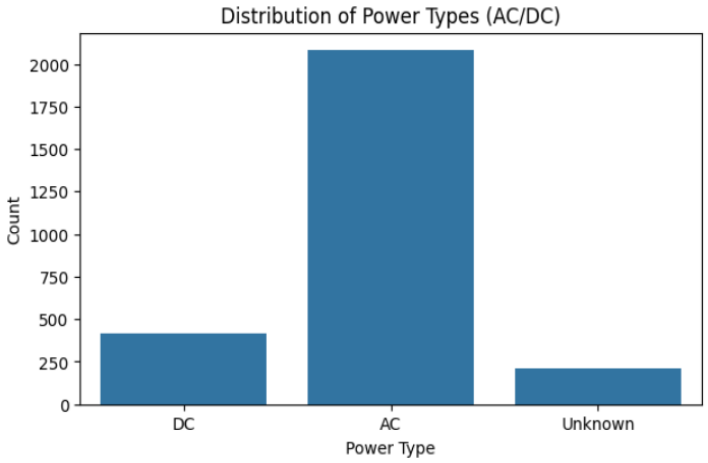
Out of **2,705** stations analysed, **2,498** are regular **charging stations**, while **207** support **battery swapping**.
This suggests that the majority of infrastructure is focused on standard charging, but there is a growing segment of battery-swapping stations catering to a different business model.

5.2 Power Types (AC/DC)

- **2,079** stations use **AC power**, while **418** utilize **DC power**, indicating that AC-powered stations are the predominant type.
- **208** stations have unspecified or **unknown** power types.

The large number of AC stations reflects the slower but more widely available charging method, while DC stations, typically used for fast charging, are less common but crucial for reducing charging times.





Extracting Segments

1. Segmenting by Vehicle Type: We'll use the registration data to identify distinct market segments (e.g., two-wheelers, three-wheelers, etc.).
2. Segmenting by Charging Capacity: Segment the market based on the type and capacity of charging stations (e.g., low capacity vs. high capacity stations).
3. Geographical Segmentation: Segment based on regional factors, such as the distribution of operational public charging stations by state.

Profiling Segments

Profile each segment by analysing the following characteristics:

1. Vehicle class registrations.
2. Charging station availability and capacity.
3. Geographical location of charging stations.

Segment by vehicle type registration

1. Two-Wheelers (NT) and Three-Wheelers (T) represent the largest market segments, indicating a high demand for charging stations for these vehicle types.
2. Light Motor Vehicles also constitute a significant portion, meaning charging infrastructure catering to these vehicles is essential.

Segmenting charging stations based on capacity

1. Low capacity stations (<= 15 kW) dominate the market, but there's a growing need for high capacity stations to support faster charging for heavier and more demanding vehicles (e.g., light goods vehicles, commercial vehicles).
2. High capacity stations (e.g., 142 kW) are a small segment, but they are essential for long-distance travellers and commercial EV fleets.

Group by state to see distribution of public charging stations

1. There are stark differences in the number of charging stations across states. Some states are well-equipped, while others may present opportunities for expansion.
2. High-density EV areas should be the focus of high-capacity charging station investments.

Data Preprocessing

1. Handle Missing Values: Handled the missing values earlier by replacing them with suitable values. For clustering, we need to ensure there are no missing values remaining in the dataset.
2. Encoding Categorical Variables: Since clustering algorithms work with numerical data, we'll encode the categorical variables. We'll use LabelEncoder for simple labels and OneHotEncoder for more complex cases.
3. Scaling Numerical Features: It's important to scale the numerical features (e.g., capacities, power types, etc.) since clustering algorithms are sensitive to feature scaling. We'll use StandardScaler for this.

Modeling

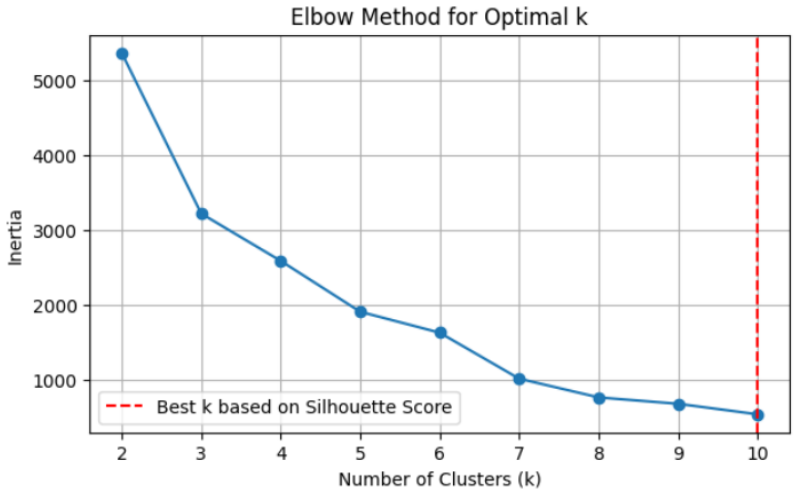
Modeling is the process of creating predictive or descriptive algorithms to extract patterns, relationships, and insights from data. In the context of the EV market segment analysis, modeling helps to analyse large datasets and identify trends, patterns, or anomalies that may not be easily visible through exploratory data analysis (EDA) alone. By building predictive models, we can forecast future trends, segment the market, and make data-driven recommendations.
It enables grouping of customers or regions into specific segments based on behaviours, needs, or attributes, improving targeted strategies.

K-Means Algorithm

K-Means is a popular clustering algorithm used to partition a dataset into K distinct, non-overlapping subgroups (clusters). The algorithm works by assigning data points to clusters based on their features, with the aim of minimizing the variance within each cluster while maximizing the variance between clusters. This is achieved through an iterative process that involves the following steps:

1. **Initialization**: Randomly select K initial centroids from the dataset.
2. **Assignment**: Assign each data point to the nearest centroid, forming K clusters.
3. **Update**: Calculate the new centroids by taking the mean of all data points assigned to each cluster.
4. **Convergence**: Repeat the assignment and update steps until the centroids no longer change significantly or a maximum number of iterations is reached.

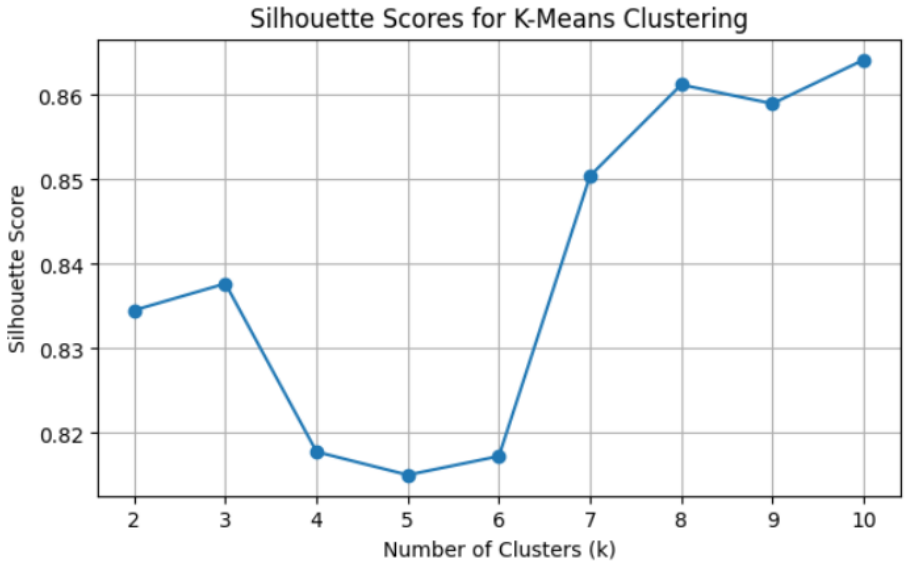
The effectiveness of the K-Means algorithm can be assessed using the silhouette score, which measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters.



In this analysis, the K-Means algorithm was evaluated for different values of K (the number of clusters), and the silhouette scores were calculated as:

```
K-Means Silhouette Score for k=2: 0.8345414164222698
K-Means Silhouette Score for k=3: 0.8376889402978663
K-Means Silhouette Score for k=4: 0.8177798143408805
K-Means Silhouette Score for k=5: 0.8150416983443757
K-Means Silhouette Score for k=6: 0.8172635816298474
K-Means Silhouette Score for k=7: 0.8504442843794148
K-Means Silhouette Score for k=8: 0.861228058586249
K-Means Silhouette Score for k=9: 0.8589878683456159
K-Means Silhouette Score for k=10: 0.8641655272428672
```

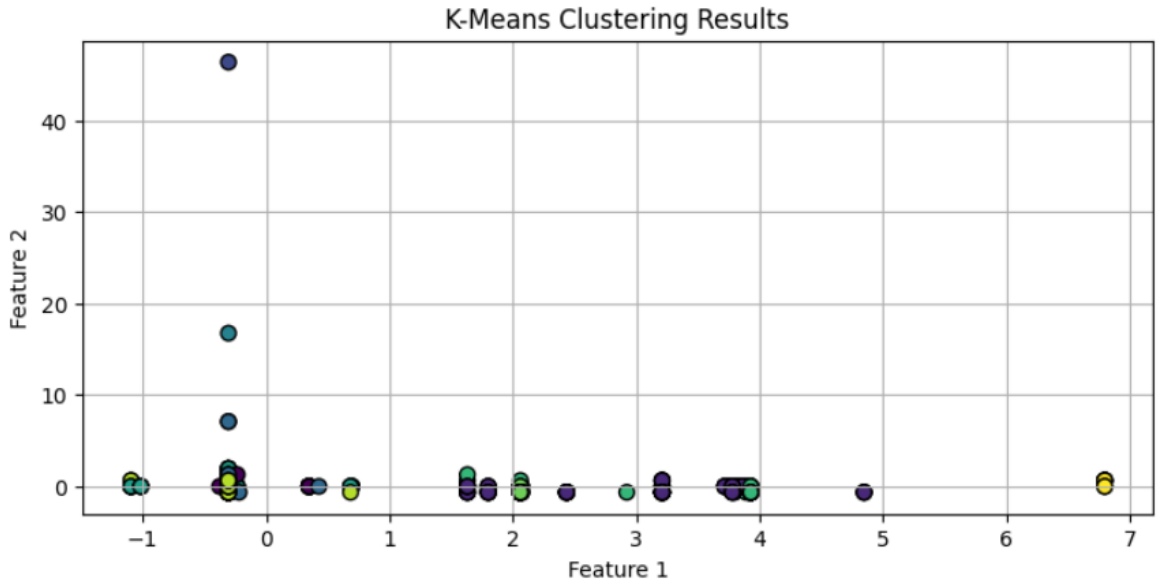
The graph for above values is as shown below:




```
Best K for K-Means: 10
```

The highest silhouette score was achieved with K=10, indicating that this is the optimal number of clusters for this dataset. The scores consistently increased as K increased, suggesting that additional clusters improve the separation of data points into distinct groups.
This result implies that segmenting the market into ten distinct clusters may provide valuable insights into customer behavior or preferences, enabling more targeted strategies in the EV market.

Plotting K-Means Clustering results:



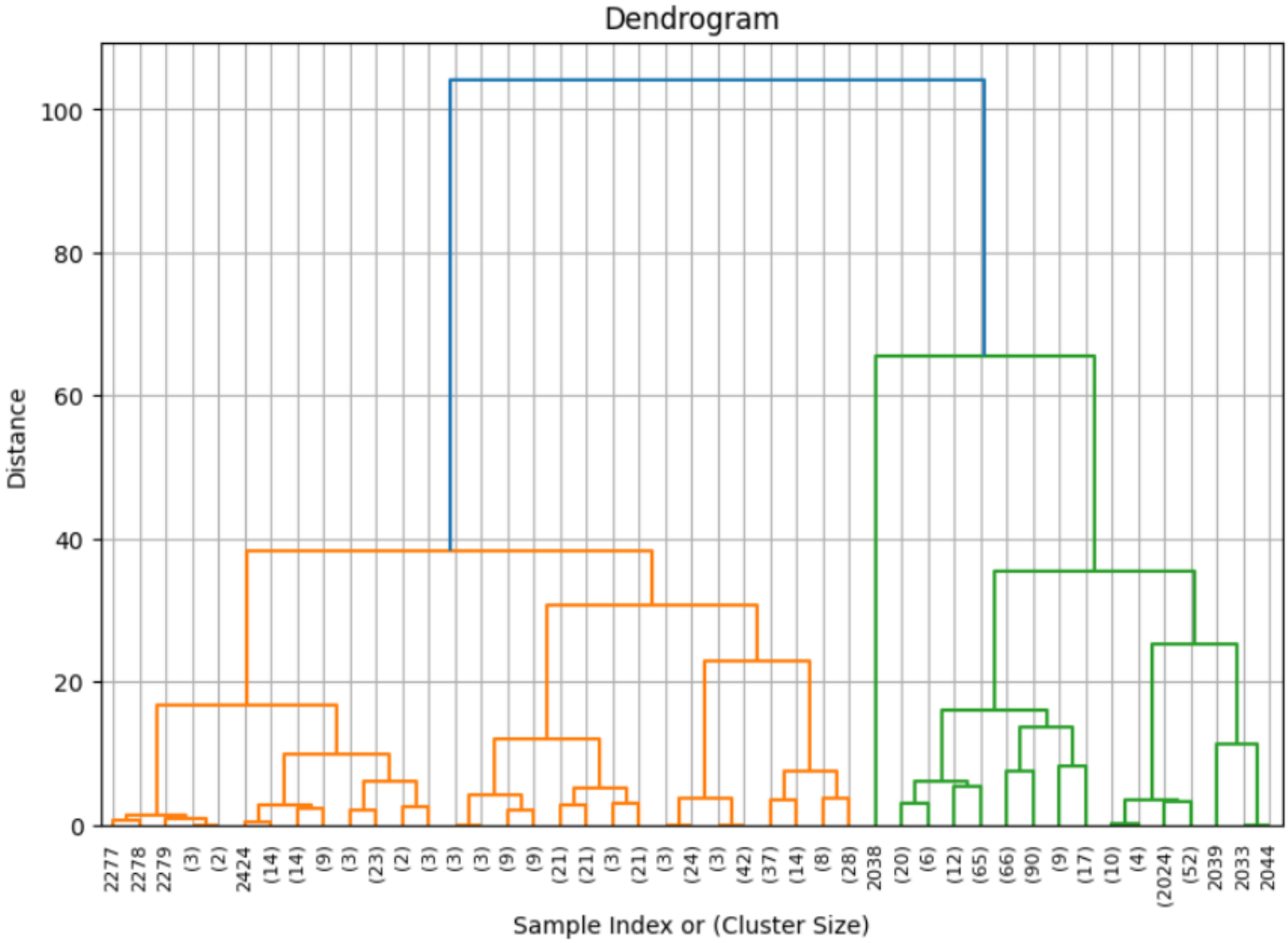
Hierarchical Clustering

Hierarchical clustering is an unsupervised learning technique used to group similar data points based on their features. Unlike K-Means clustering, which partitions data into a specified number of clusters, hierarchical clustering builds a tree-like structure (dendrogram) that shows the relationships between clusters at various levels of similarity. This method allows for more flexible cluster definitions and can help identify nested groups within the data.

```
Hierarchical Clustering Silhouette Score: 0.8624821215169947
```

The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where a higher score indicates better-defined clusters. The silhouette score of 0.862 obtained from hierarchical clustering indicates that the clusters are well-separated and distinctly defined, demonstrating the effectiveness of this clustering approach for the dataset at hand.

Analysing through Dendrogram:
The dendrogram plot visually represents the clustering process, illustrating how data points are merged into clusters based on their similarity. By analysing the dendrogram, we can determine the appropriate number of clusters and understand the hierarchical relationships among the data points.



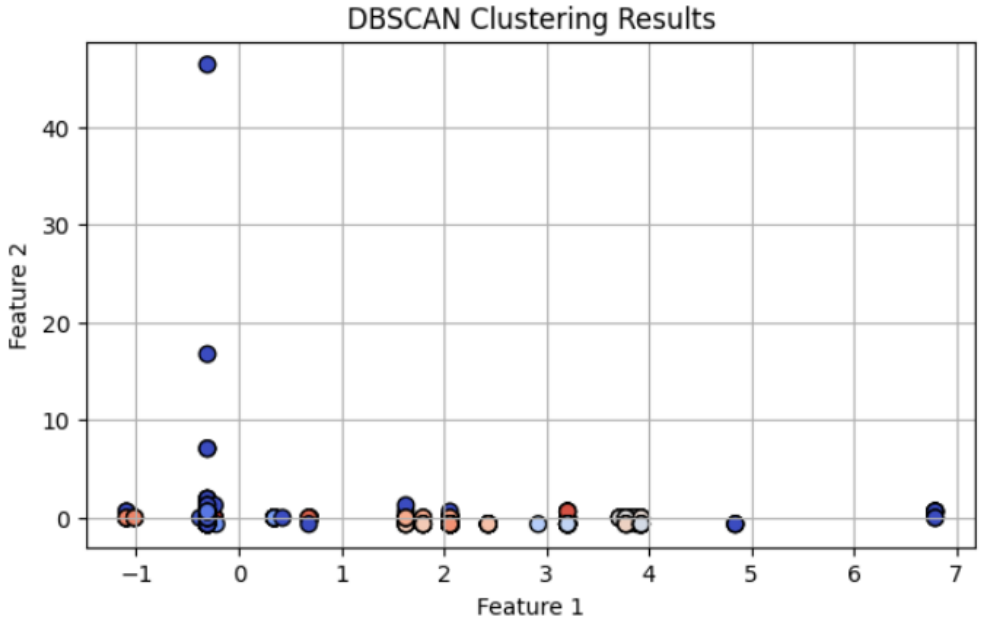
By Visualizing dendrogram, for approximate results, even four clusters suffice business needs. While for better results, 10 clusters can be formed but with increase cost.

DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an unsupervised clustering algorithm that identifies clusters based on the density of data points in a given space. Unlike K-Means, which requires specifying the number of clusters beforehand, DBSCAN groups data points that are closely packed together while marking points that lie alone in low-density regions as outliers. This makes it particularly effective for identifying clusters of varying shapes and sizes, as well as for handling noise in the data.

```
DBSCAN Silhouette Score: 0.7461664680185108
```

The silhouette score for DBSCAN is 0.746, indicating a relatively good level of separation between clusters. While this score is lower than those achieved with K-Means and hierarchical clustering, it still suggests that the clusters are reasonably well-defined. A silhouette score above 0.5 generally indicates that the clustering is effective, but it may also imply the presence of noise or overlapping clusters.



This result highlights DBSCAN's ability to capture the underlying structure of the dataset, particularly in cases where clusters are not spherical in shape. However, the lower silhouette score compared to the other algorithms may indicate some challenges in separating clusters distinctly. Further tuning of DBSCAN's parameters could help improve the clustering performance and enhance the silhouette score.

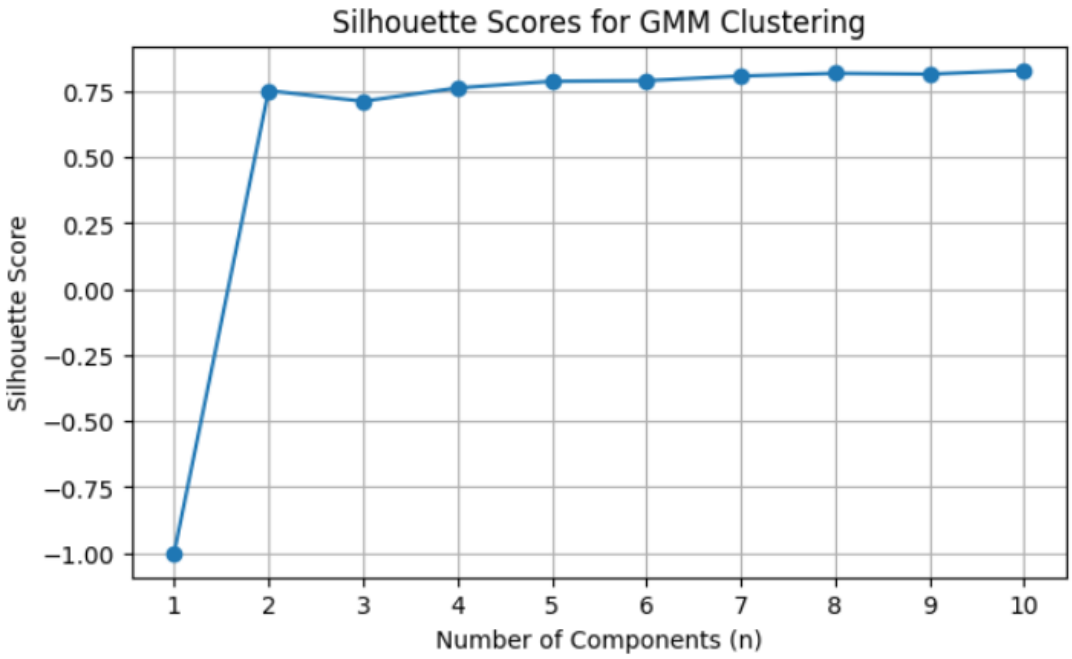
Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM) are a probabilistic model used for clustering that assumes the data is generated from a mixture of several Gaussian distributions with unknown parameters. Unlike K-Means, which assigns each data point to a single cluster, GMM provides a soft clustering approach, where each data point can belong to multiple clusters with certain probabilities. This makes GMM particularly useful for identifying clusters in datasets where the clusters may overlap or have different shapes and sizes.

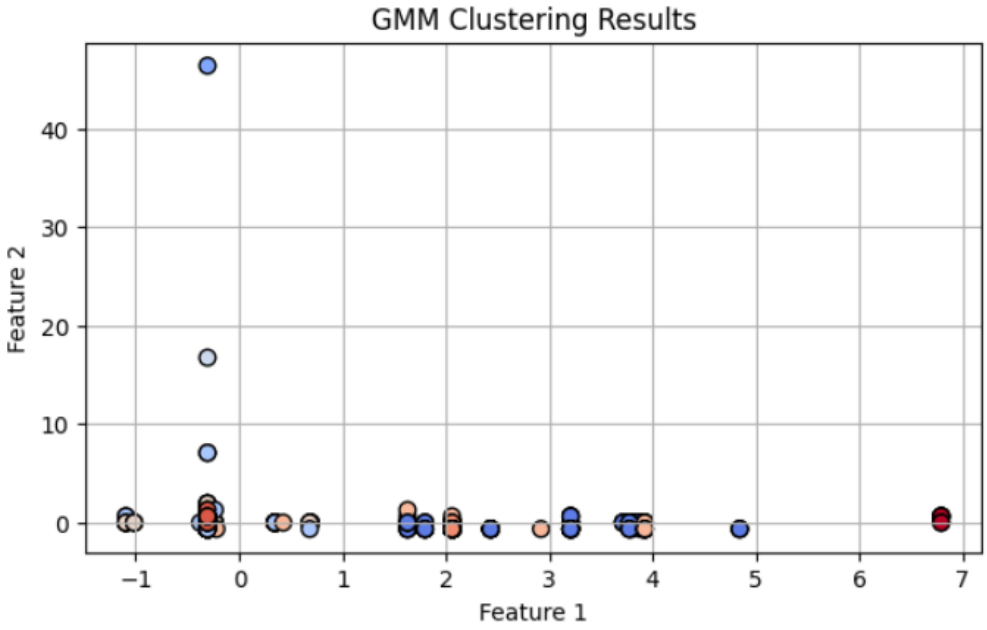
```
Best number of components for GMM: 10
Best GMM Silhouette Score: 0.8296719276108886
```

For the GMM applied in this analysis, the best number of components identified is 10, suggesting that the data can be effectively modelled with ten distinct Gaussian distributions. The silhouette score achieved is 0.830, indicating a strong separation between clusters, although it is slightly lower than the scores obtained with K-Means and hierarchical clustering.

This score reflects the model's ability to capture the underlying structure of the dataset, effectively grouping similar data points while maintaining a clear distinction between different clusters. The combination of the optimal number of components and the high silhouette score suggests that GMM is a suitable method for this analysis, providing flexibility in cluster assignments and robustness against noise.



Observe the GMM Clustering results.



Comparison And extract Best Model

Observe the scores of all models below:

```
--- Clustering Results Comparison ---
------------------------------------------
K-Means:
  Best K: 10
  Silhouette Score: 0.8641655272428672
------------------------------------------
Hierarchical:
  Silhouette Score: 0.8624821215169947
------------------------------------------
DBSCAN:
  Silhouette Score: 0.7461664680185108
------------------------------------------
GMM:
  Best Components: 10
  Silhouette Score: 0.8296719276108886
------------------------------------------
Best Clustering Model: K-Means
```

The evaluation of clustering models reveals varying degrees of effectiveness in capturing the underlying structures of the dataset, as indicated by their respective silhouette scores. Here's a comparison of the clustering models based on their silhouette scores:

1. **K-Means**:
    - **Best K**: 10
    - **Silhouette Score**: 0.8642
    - **Insights:** K-Means exhibits the highest silhouette score, indicating that it effectively partitions the dataset into well-separated clusters. The high score suggests that the clusters formed are dense and distinct, making K-Means the most suitable model for this data.

2. **Hierarchical Clustering**:
    - **Silhouette Score**: 0.8625
    - **Insights:** Hierarchical clustering closely follows K-Means in performance with a very similar silhouette score. This indicates that the hierarchical method also creates well-separated clusters.

However, the slight difference suggests that while hierarchical clustering is effective, it may be less efficient than K-Means for this specific dataset.

3. **DBSCAN**:
    - **Silhouette Score**: 0.7462
    - **Insights:** DBSCAN shows a significantly lower silhouette score compared to K-Means and hierarchical clustering. This indicates that the clusters formed are less distinct and may overlap more, suggesting that DBSCAN struggled with the dataset's density variations or noise. This model might be more appropriate for datasets with clear density differences and less overlapping clusters.

4. **Gaussian Mixture Model (GMM)**:
    - **Best Components**: 10
    - **Silhouette Score**: 0.8297
    - **Insights:** GMM offers a soft clustering approach, and while it achieves a decent silhouette score, it still falls short compared to K-Means and hierarchical clustering. The score suggests that the clusters are reasonably well-separated, but the overlaps indicate that some data points may not be assigned to the most appropriate clusters, highlighting the limitations of GMM in this context.

Conclusion:

Best Clustering Model: **K-Means** is determined to be the best clustering model for this dataset based on its superior silhouette score, which reflects the model's effectiveness in creating distinct clusters. The results suggest that K-Means is well-suited for this analysis, providing clear and well-defined groupings of data points.

Hierarchical clustering is also a strong candidate, while DBSCAN and GMM may require further tuning or may be better suited for different types of datasets. Overall, the comparison highlights the importance of selecting the appropriate clustering method based on the specific characteristics of the dataset being analysed.

Insights and Recommendations:

1. Adopt K-Means for Market Segmentation:

Based on the clustering analysis, K-Means demonstrated the highest silhouette score (0.8642), indicating effective segmentation of the EV market. It is recommended to utilize K-Means for creating targeted marketing strategies and optimizing resource allocation.

2. Investigate Hierarchical Clustering Results:

Hierarchical clustering showed a similar silhouette score (0.8625) to K-Means. This suggests that further analysis with hierarchical clustering could provide additional insights. It is recommended to visualize these clusters and analyse their characteristics for a deeper understanding of customer segments.

3. Explore DBSCAN for Density-Based Segmentation:

Although DBSCAN produced a lower silhouette score (0.7462), its strength lies in identifying dense regions in data. If there are specific geographical areas with high EV adoption rates or charging station concentrations, DBSCAN could be beneficial in identifying these dense clusters, especially in urban planning and infrastructure development.

4. Leverage Gaussian Mixture Models for Flexibility:

GMM achieved a silhouette score of 0.8297, indicating reasonable clustering with the flexibility of soft assignments. This model can be beneficial for understanding customer behavior where overlaps in segment characteristics may exist. Further refinement of GMM parameters may yield better results.

5. Utilize Clustering Insights for Business Strategies:

The clusters identified through K-Means and hierarchical clustering can be utilized to tailor marketing campaigns