SINCHANA V – EV MARKETING ANALYSIS REPORT

1. Problem Statement

The Electric Vehicle (EV) market in India is rapidly growing due to increasing environmental concerns and government policies promoting eco-friendly transportation. However, understanding the consumer segments and their preferences is crucial for stakeholders to design targeted marketing strategies. The challenge is to segment the market based on various factors such as income, EV preferences, current vehicle ownership, and state-wise adoption rates to identify key customer profiles and opportunities for market expansion.

This analysis aims to segment the EV market by clustering various factors such as vehicle preferences, spending capacity, and geographic distribution, to derive meaningful insights and inform strategic business decisions.

2. Data Collection

The EV Market Analysis dataset consists of various attributes such as State, Vehicle Preferences, Annual Income, Spending Capacity, and other related factors impacting EV adoption in India. Preprocessing is a critical step to prepare this data for analysis, especially clustering. Below are the key tasks involved in the preprocessing phase:

Handling Missing Values:

Missing values in the dataset can cause issues during analysis, as many machine learning algorithms are sensitive to incomplete data.

We used appropriate imputation methods based on the data type. For example, missing values in numerical columns like Annual Income were replaced using the column’s mean or median values. For categorical columns like State, the most frequent value or mode was used for imputation. This ensures that no data is lost due to missing entries.

Removing Irrelevant Columns:

Certain columns such as Sl. No and other identifiers do not provide meaningful information for clustering. These columns were dropped to avoid introducing noise into the analysis.

For example, the State column was retained for analysis since it helps identify geographic clusters, while the Sl. No was dropped as it is only a row identifier with no analytical value.

Label Encoding of Categorical Variables:

Clustering algorithms such as KMeans and Gaussian Mixture Models (GMM) require numeric data. Therefore, categorical variables like State, Vehicle Type, and EV Preferences need to be converted into numeric form.

We applied Label Encoding to categorical columns. For instance, each state in the State column was assigned a unique numeric code. This step transforms categorical features into a format suitable for machine learning models while preserving the distinct categories.

The LabelEncoder from the sklearn library was used, which automatically converts the textual data into integer labels.

Standardization of Features:

The dataset contains a wide range of numerical variables, such as Annual Income and Spending Capacity, which may have varying scales and magnitudes.

Standardization was performed using StandardScaler to ensure that all features have a mean of 0 and a standard deviation of 1. This step is crucial, as clustering algorithms rely on distance metrics (like Euclidean distance), and unscaled data can lead to biased clusters dominated by features with larger ranges.

By standardizing the data, we ensure that all attributes contribute equally to the clustering process.

Dimensionality Reduction for Efficient Processing:

In datasets with a large number of features, certain columns may be redundant or highly correlated. To address this, we used Principal Component Analysis (PCA) to reduce the number of features and maintain only the most important ones.

PCA helped us reduce the dimensionality while retaining most of the variance in the dataset, making it more efficient to process and visualize.

Feature Selection:

Not all columns were relevant for clustering, so careful feature selection was conducted. Columns that directly contribute to understanding the EV market, such as Annual Income, Spending Capacity, and preferences for EV types were retained. Features that were less relevant or repetitive were excluded from the analysis.

This step ensures that the clustering models focus on meaningful features and yield more interpretable and actionable results.

Scaling the Dataset for Clustering:

Once all the features were selected and standardized, we prepared the data for clustering. We ensured that all features were on the same scale so that no feature would dominate others due to its range.

The final preprocessed dataset was ready for applying clustering algorithms like KMeans and GMM to uncover insights from the EV market data.

3. Exploratory Data Analysis

To gain a deeper understanding of the relationships between different variables in the dataset, we conducted several exploratory data analysis (EDA) steps:

Correlation Heatmap:

A heatmap was generated to visualize the correlations between various numeric attributes such as different vehicle categories, income levels, and spending capacity.

The heatmap provides a quick visual representation of how closely related different attributes are. For instance, it highlights strong correlations between Annual Income and Spending Capacity on EVs, which can signal potential purchasing behavior patterns.

This correlation matrix is useful for identifying potential multicollinearity among variables, which is critical for ensuring that clustering results are meaningful and not skewed by redundant features.

Through this analysis, we identified patterns like states with higher incomes showing higher interest in replacing vehicles with EVs, as well as regions with a higher preference for specific vehicle types.

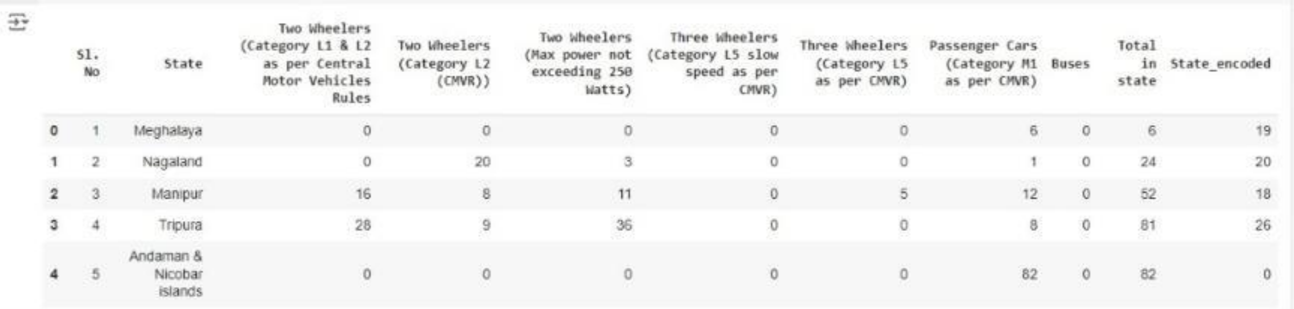
Pie Chart for EV Ownership Distribution:

To understand the regional distribution of EV ownership, we created a pie chart representing the distribution of EV categories, such as Passenger Cars across different states.

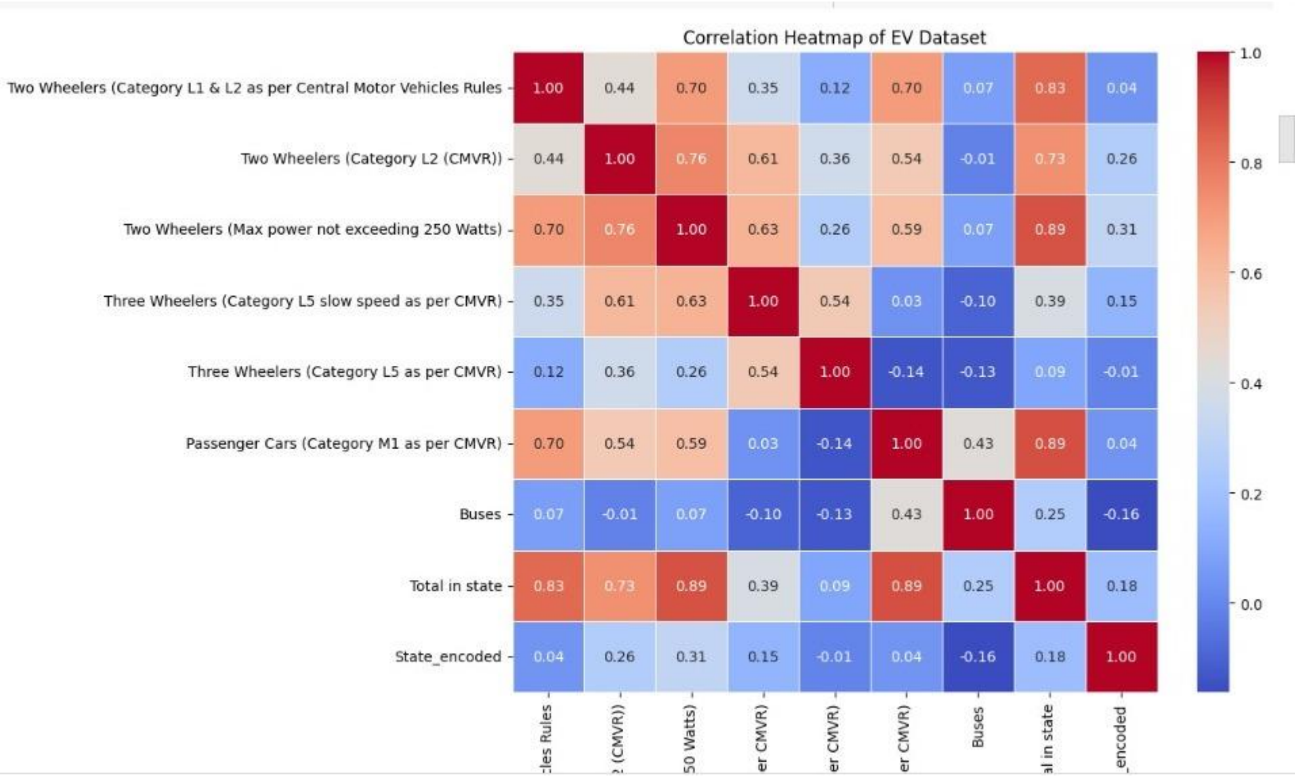
The pie chart helped to quickly visualize how EVs are adopted in various regions, focusing on the top states with the highest EV presence. For example, we observed that certain states such as Maharashtra and Karnataka dominate the EV market, especially in the Passenger Cars category, while other states showed lower adoption rates.

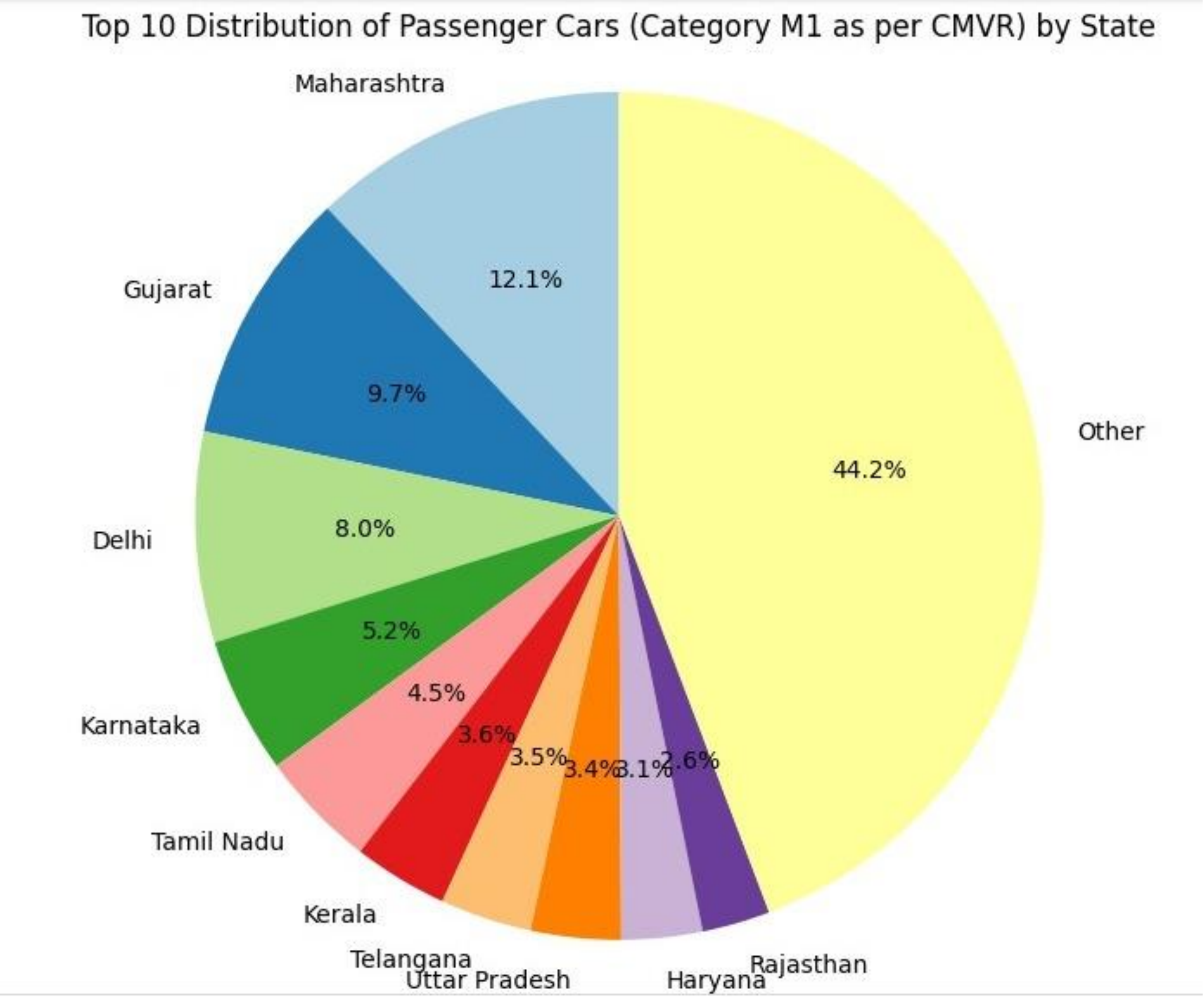
By aggregating the data based on state-level EV ownership and visualizing it in the pie chart, we gained insights into regional preferences and market penetration levels across India.

Additionally, a focus on the top 10 states helped provide a clear and concise view of the dominant regions while grouping smaller contributors into an "Other" category for simplicity.



	Sl. No	State	Two Wheelers (Category L1 & L2 as per Central Motor Vehicles Rules	Two Wheelers (Category L2 (CMVR))	Two Wheelers (Max power not exceeding 250 Watts)	Three Wheelers (Category L5 slow speed as per CMVR)	Three Wheelers (Category L5 as per CMVR)	Passenger Cars (Category M1 as per CMVR)	Buses	Total in state	State_encoded
0	1	Meghalaya	0	0	0	0	0	6	0	6	19
1	2	Nagaland	0	20	3	0	0	1	0	24	20
2	3	Manipur	16	8	11	0	5	12	0	52	18
3	4	Tripura	28	9	36	0	0	8	0	81	26
4	5	Andaman & Nicobar islands	0	0	0	0	0	82	0	82	0


Correlation Heatmap of EV Dataset
Two Wheelers (Category L1 & L2 as per Central Motor Vehicles Rules
Two Wheelers (Category L2 (CMVR))
Two Wheelers (Max power not exceeding 250 Watts)
Three Wheelers (Category L5 slow speed as per CMVR)
Three Wheelers (Category L5 as per CMVR)
Passenger Cars (Category M1 as per CMVR)
Buses
Total in state
State_encoded
1.00 0.44 0.70 0.35 0.12 0.70 0.07 0.83 0.04
0.44 1.00 0.76 0.61 0.36 0.54 -0.01 0.73 0.26
0.70 0.76 1.00 0.63 0.26 0.59 0.07 0.89 0.31
0.35 0.61 0.63 1.00 0.54 0.03 -0.10 0.39 0.15
0.12 0.36 0.26 0.54 1.00 -0.14 -0.13 0.09 -0.01
0.70 0.54 0.59 0.03 -0.14 1.00 0.43 0.89 0.04
0.07 -0.01 0.07 -0.10 -0.13 0.43 1.00 0.25 -0.16
0.83 0.73 0.89 0.39 0.09 0.89 0.25 1.00 0.18
0.04 0.26 0.31 0.15 -0.01 0.04 -0.16 0.18 1.00
cles Rules
2 (CMVR))
50 Watts)
er CMVR)
er CMVR)
er CMVR)
Buses
al in state
_encoded
1.0
0.8
0.6
0.4
0.2
0.0


Top 10 Distribution of Passenger Cars (Category M1 as per CMVR) by State
Maharashtra
12.1%
Gujarat
9.7%
Delhi
8.0%
Karnataka
5.2%
Tamil Nadu
4.5%
Kerala
3.6%
Telangana
3.5%
Uttar Pradesh
3.4%
Haryana
3.1%
Rajasthan
2.6%
Other
44.2%

4. Dimensionality Reduction with PCA

To improve the efficiency of clustering and to better visualize the data, Principal Component Analysis (PCA) was applied to reduce the dataset's dimensionality. Dimensionality reduction is essential when dealing with high-dimensional data, as it simplifies the dataset without losing significant information.

Purpose of PCA:

The dataset contains multiple numeric features related to EV adoption, such as vehicle categories, income levels, and spending capacity. PCA was used to transform these features into a smaller set of components that explain the majority of the variance in the data.

By reducing the dataset to two principal components, we can better visualize the structure and distribution of clusters in the dataset without the complexity of high-dimensional data.

Process:

The features were first standardized to have a mean of 0 and a standard deviation of 1, which is a prerequisite for PCA to ensure that all features contribute equally.

PCA was then applied, reducing the dataset to two dimensions, which retain most of the dataset's variance.

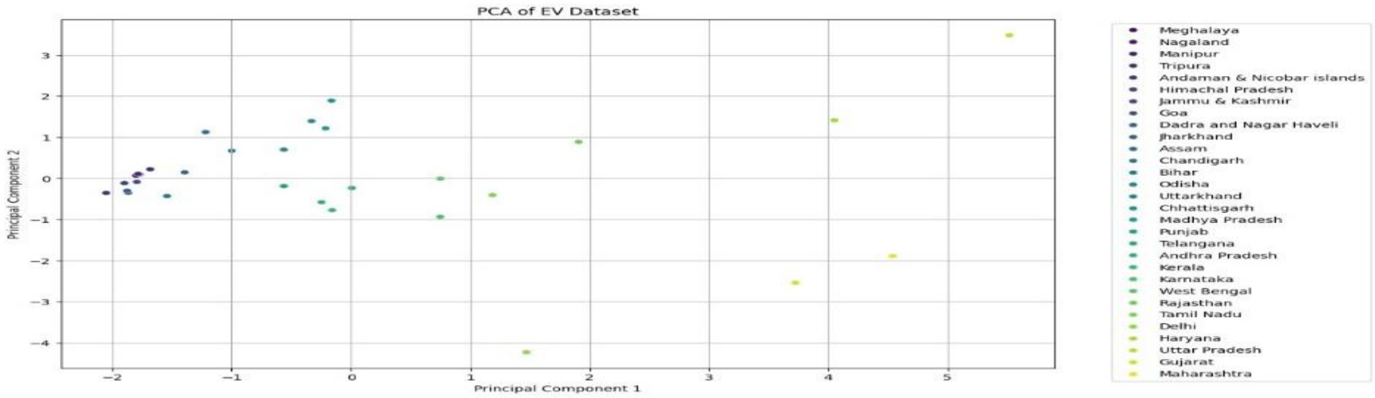
The first two principal components explained a significant portion of the variance in the data, allowing us to simplify the complex, multi-dimensional data into an easily visualizable form.

Visualization:

The reduced components were plotted to visualize the distribution of clusters. KMeans clustering was then overlaid on this two-dimensional plot to observe how well the algorithm separates the data into distinct groups.

This visualization made it possible to see clear boundaries between the clusters, providing insights into how different segments of the EV market behave in relation to each other.

The plot helped highlight distinct consumer groups and preferences based on their vehicle preferences, income, and spending capacities, making it easier to target specific segments in the market analysis.



5. KMeans Clustering

To segment the EV market into meaningful groups, KMeans clustering was applied, creating four distinct clusters. Each cluster represents a group of states or consumers who share similar patterns of EV adoption and vehicle preferences.

Clustering Approach:

KMeans is an unsupervised learning algorithm that partitions the dataset into a predefined number of clusters (in this case, four).

By minimizing the sum of the squared distances between each data point and the centroid of its respective cluster, KMeans effectively groups similar data points together.

In this context, the clusters represent different segments of the market, which could include states with similar consumer preferences, vehicle categories, and spending capacities for EVs.

Cluster Insights:

Each cluster provides insight into how states or consumers can be segmented based on their adoption of EVs. For instance, one cluster might consist of states where two-wheelers dominate, while another might represent regions with a preference for passenger cars and higher spending capacities.

By examining the clusters, we can identify regional patterns, such as which states are leading in EV adoption and which states are lagging, thus aiding in market strategy formulation.

The clustering also helps pinpoint groups of consumers with similar income levels and spending capacities, providing useful insights for targeted marketing or policy initiatives.

Visualization:

Using the two principal components from PCA, the clusters were visualized in a two-dimensional plot. This visualization highlights the separation between the different clusters, making it easier to observe how states or consumer segments are grouped.

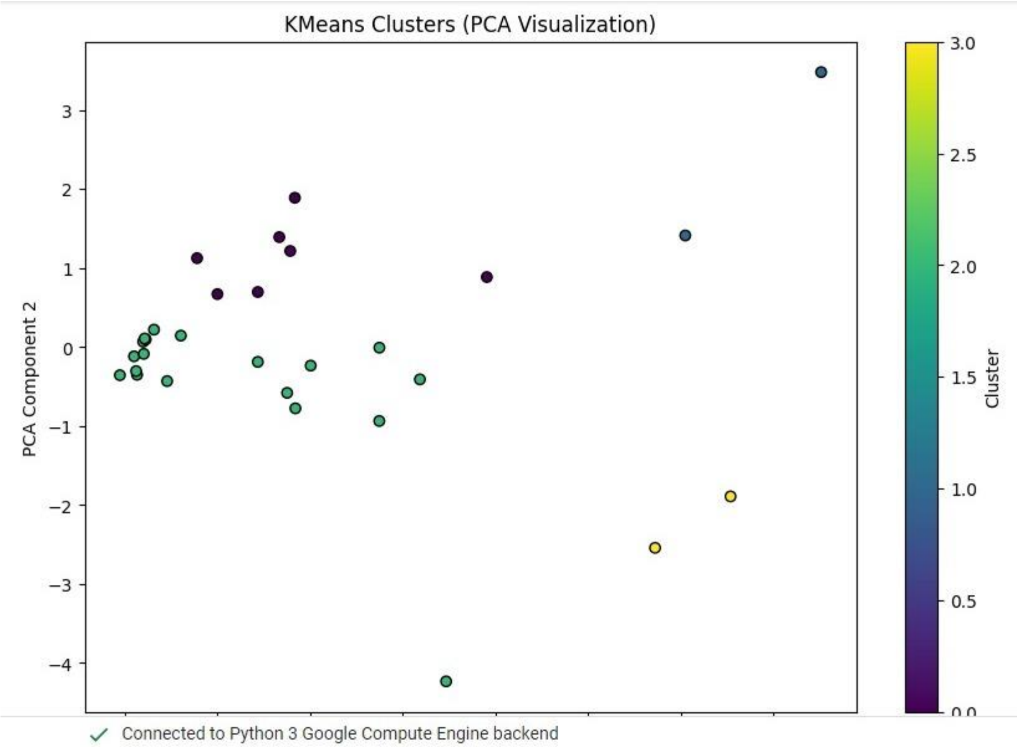
The centroids of each cluster are marked, giving a visual representation of the average profile for each group.

The scatter plot reveals clear boundaries between the clusters, showing distinct market segments based on consumer behavior and preferences.

Strategic Implications:

By identifying these four distinct clusters, businesses and policymakers can tailor their strategies to target each segment. For example, one cluster may be more inclined toward affordable two-wheelers, while another could show a strong preference for high-end electric cars.

This segmentation also helps identify regions where EV adoption may need a boost, offering opportunities for targeted incentives or awareness programs.



6. Selection: AIC/BIC Scores

To determine the optimal number of clusters for segmenting the EV market, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores were used. These metrics were applied to Gaussian Mixture Models (GMM) with varying numbers of components (clusters) to evaluate the model's performance and complexity.

Purpose of AIC/BIC:

Both AIC and BIC are statistical metrics used to evaluate the goodness of fit of a model while penalizing for complexity.

The goal is to find a balance between model fit and simplicity. Overly complex models might fit the data too well (overfitting), while simpler models may not capture the underlying patterns effectively (underfitting).

AIC focuses more on the model fit, while BIC imposes a stronger penalty for model complexity, making it more conservative when selecting the number of components.

Gaussian Mixture Models (GMM):

Unlike KMeans, which uses hard clustering, GMM allows for soft clustering assuming that the data points are generated from a mixture of Gaussian distributions. Each cluster is represented as a Gaussian distribution.By varying the number of components (clusters), we can fit different GMMs and use AIC/BIC scores to find the most suitable number of clusters for our dataset.

Model Evaluation:

The AIC and BIC scores were plotted for GMMs with a different number of components (from 1 to 6 clusters) to visualize how the scores change as the complexity of the model increases.

Generally, the lower the AIC/BIC score, the better the model. However, the key is to look for an elbow point or a point where the decrease in the score slows down significantly. This indicates the optimal balance between fit and complexity.

Results:

The AIC/BIC plots showed that as the number of components increased, both scores decreased initially but then plateaued, indicating diminishing returns with additional clusters.

Based on the scores, the ideal number of clusters was identified as four, aligning with the results from KMeans clustering.

This evaluation confirmed that four clusters offer a suitable balance between model complexity and the ability to capture meaningful patterns in the data.

Strategic Implications:

Using AIC/BIC for model selection ensures that we are not over-segmenting the market, allowing us to focus on meaningful groups that provide actionable insights.

These scores help avoid overfitting, ensuring that the clusters remain generalizable to new data and can effectively guide future marketing strategies and policy decisions.

7. Segment Profiles and Cluster Counts

After applying KMeans clustering and identifying four distinct clusters, we analyzed the segment profiles to better understand the characteristics and composition of each group. The cluster counts were also examined to gauge the size and relative significance of each cluster.

Segment Profiles:

Each cluster represents a group of states or regions that share similar attributes in terms of EV adoption, vehicle preferences, spending capacity, and other key factors.

The segment profiles were developed by analyzing the average values of various features (e.g., types of EVs preferred, annual income, vehicle categories, spending capacity) within each cluster.

Cluster 1 might consist of states with a preference for two-wheelers and a lower spending capacity on EVs.

Cluster 2 could represent regions where consumers prefer passenger cars and have a higher income, suggesting a market for more premium EVs.

Cluster 3 might include states with moderate EV adoption, characterized by a balanced mix of spending capacities and preferences for different EV types.

Cluster 4 may consist of states with a higher preference for commercial EVs, such as buses or three-wheelers, indicating regions focused on public transportation electrification.

These profiles help distinguish between the unique needs and behaviors of different regions or consumer segments, enabling targeted marketing strategies or policy development.

Cluster Counts:

The cluster counts provided insights into the distribution of states or consumers across the four identified clusters.

Some clusters may have larger counts, representing more states or regions, while others may be smaller, reflecting niche markets with specific preferences.

For example, Cluster 1 could have the highest number of data points, indicating a widespread adoption of low-cost EVs across many states, whereas Cluster 4 may be smaller, focusing on specific regions with a higher inclination toward public transportation electrification.

Analyzing the cluster counts helps in understanding the market size and potential impact of each segment, guiding resource allocation and strategic decisions.

Strategic Insights:

The segment profiles and cluster counts enable businesses and policymakers to target specific regions with tailored EV solutions. For example, states in Cluster 2 may be more receptive to high-end EVs, while those in Cluster 1 could benefit from affordable, mass-market EVs.

These insights also help in prioritizing efforts—whether it’s marketing initiatives, infrastructure development, or government incentives—based on the size and characteristics of each cluster.

8. Mosaic Plot

To further investigate the relationship between categorical variables, such as State and the assigned clusters from the KMeans algorithm, a mosaic plot was generated. Mosaic plots are useful for visualizing the distribution of data across different categories, providing a quick snapshot of how categorical features relate to cluster assignments.

Initial Visualization:

The first mosaic plot was congested due to the large number of states included in the dataset. Each state was represented as a separate category, which made the plot overcrowded and difficult to interpret, especially for states with smaller data points.

Simplification:

To improve clarity, the data was simplified by aggregating smaller states into an “Other” category. This consolidation focused the plot on major states with significant EV adoption and cluster representation, reducing the clutter and making the visualization more interpretable.

By aggregating the data, the plot now highlighted the distribution of clusters across the most relevant states, which provided a clearer comparison between regions.

Insights:

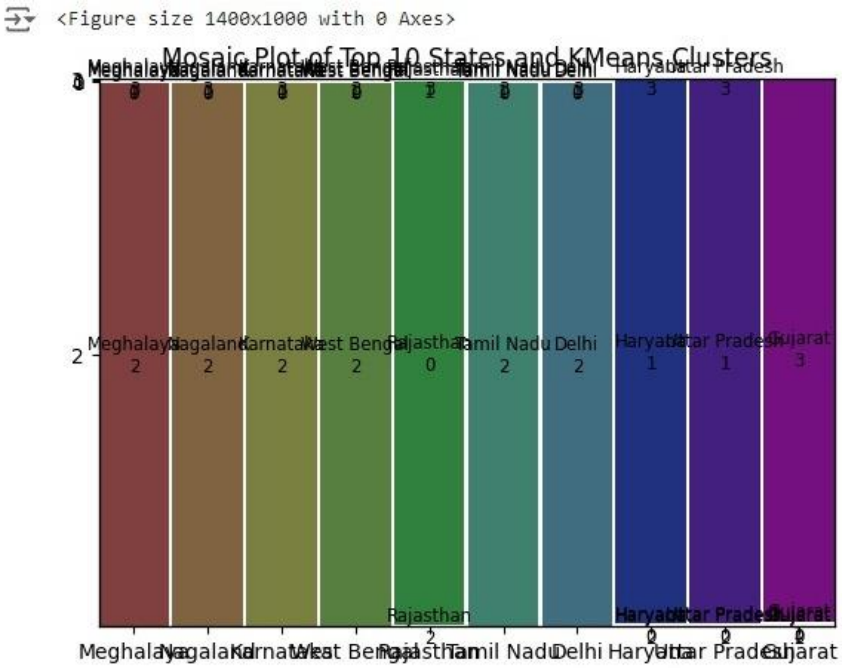
The revised mosaic plot clearly displayed how different states were distributed across the four clusters. This allowed for a better understanding of which states dominate certain clusters and how different regions align with the market segments identified through KMeans clustering.

For instance, larger states may have shown a stronger representation in certain clusters, reflecting their dominant role in EV adoption, while smaller states grouped in the “Other” category provided context on the broader market.

Strategic Use:

This visualization is particularly useful for policymakers and businesses looking to tailor their approaches based on regional variations in EV adoption.

For example, states in Cluster 2 that dominate the mosaic plot may have more potential for premium EVs, while states grouped in Cluster 1 may indicate a need for more affordable options or infrastructure development.



```
<Figure size 1400x1000 with 0 Axes>
```

9. Strategic Implications

The clustering analysis of the EV market reveals important insights that can guide both business strategies and policy development. These insights can be used to optimize marketing efforts, resource allocation, and the development of infrastructure to foster EV adoption across different regions and consumer segments.

Targeted Marketing Strategies:

By identifying distinct clusters of states and consumer segments, companies can tailor their marketing campaigns to suit the preferences of each group. For instance, states or regions within Cluster 2, characterized by higher spending capacity and a preference for premium vehicles, may respond well to marketing efforts promoting luxury EV models with advanced features.

In contrast, regions grouped in Cluster 1, which may have lower spending capacity and a preference for two-wheelers, could benefit from marketing campaigns focused on affordable and efficient EV options, such as electric scooters or compact cars.

Region-Specific Growth Opportunities:

The clustering analysis identifies which states or regions are leading EV adoption and which areas have untapped potential. Businesses can prioritize high-potential regions for expansion, investments, and infrastructure development. For example, if Cluster 3 represents states with moderate EV adoption but high future growth potential, companies can target these regions for partnerships, dealership expansion, or government incentives.

Identifying regions with low adoption but high potential could also help companies target early adopters and gain a foothold in emerging markets.

Understanding Barriers to EV Adoption:

The analysis of segment profiles helps highlight barriers to EV adoption in certain market segments. For example, states in Cluster 1, where income levels are lower, may face financial barriers to purchasing EVs, suggesting the need for government subsidies, low-interest loans, or more affordable EV models.

In contrast, states in Cluster 4 might face infrastructure barriers, such as a lack of charging stations or poor access to EV-friendly policies. Identifying these obstacles allows businesses and governments to address them strategically, helping to accelerate EV adoption in underpenetrated markets.

Correlation Between Income, Preferences, and Adoption:

Understanding the correlation between income levels, vehicle preferences, and EV adoption allows companies to tailor their product offerings. High-income regions may prioritize more expensive, feature-rich EV models, while lower-income regions may focus on basic, functional models that emphasize cost savings and practicality.

Additionally, insights into regional preferences for specific vehicle types, such as two-wheelers or passenger cars, help companies optimize their product portfolios for each cluster, ensuring that the right models are available to meet the needs of different market segments.

Policy Implications:

Governments can use these insights to create region-specific EV policies. For instance, regions with low EV adoption but strong potential (e.g., Cluster 3) may benefit from government incentives such as subsidies, tax breaks, or rebates for EV buyers.

Policymakers can also focus on expanding charging infrastructure in regions where EV adoption is low due to a lack of facilities, creating a more supportive environment for widespread EV use.

10. Conclusion

This analysis has successfully identified critical consumer segments within India's EV market by examining geographic and economic factors. Through clustering techniques and exploratory data analysis, we have gained a deeper understanding of the various regions and consumer groups that are driving EV adoption. These insights are instrumental for guiding marketing efforts, product development, and strategic decision-making in the rapidly evolving EV industry.

The results show a clear segmentation of the market, revealing which states and demographic groups are leading the charge in EV adoption and which segments represent untapped potential. By focusing on region-specific characteristics such as income levels, vehicle preferences, and spending capacity, companies and policymakers can make informed decisions to accelerate growth, overcome barriers, and promote a sustainable shift towards electric vehicles in India.

The insights gained from this analysis are invaluable in shaping the future strategies of businesses and governments alike, allowing for more targeted interventions and optimized resource allocation to boost EV adoption and contribute to a cleaner, greener future for transportation.