Market Segmentation Analysis of EV market

Prepared By: Sub-Group A

Dataset Link: https://drive.google.com/file/d/12d0WiYN-SjyLaRIwyG82Vw3ac-xrS0ri/view?usp=drive_link

Github Link: https://github.com/Muskanr8/EV_Market_Analysis_Team-A.git

Problem Statement:

To identify meaningful EV market segments in India using sales data so companies can understand regional adoption patterns and make data-driven decisions for targeted EV strategies.

Background

India's Electric Vehicle (EV) sector is expanding rapidly, supported by rising fuel prices, environmental concerns, government incentives, and increasing consumer interest. However, EV adoption is not uniform across India. Different regions demonstrate significant variations in EV purchasing behavior, total vehicle demand, infrastructure readiness, and reliance on petrol/diesel vehicles.

For companies aiming to enter or expand in the EV domain, understanding how the market is segmented becomes essential. Market segmentation enables businesses to develop targeted strategies, identify high-potential regions, and allocate resources where impact will be greatest. Using machine learning techniques such as KMeans clustering, businesses can uncover patterns in EV sales data that are not visible through traditional analysis.

Problem

- How does the Indian EV market naturally divide into meaningful segments based on sales behaviour?
- Which regions show high EV adoption and readiness, and which remain dominated by petrol/diesel vehicles?
- Can machine learning (KMeans clustering) reveal actionable market groups to support EV product planning and strategy?
- What insights do EV_Sales, Vehicle_Sales, and Petrol/Diesel_Sales provide about market maturity and consumer behavior?
- How can these data-driven segments help companies decide where to expand, what products to launch, and which regions need awareness or infrastructure support?

Objective

The objective of this project is to perform **data-driven segmentation of the Indian EV market** using machine learning techniques in order to identify distinct market groups based on variations in EV sales, petrol/diesel sales, and total vehicle sales.

This segmentation aims to help organizations:

- Understand differences in EV adoption and market maturity across regions
- Identify high-potential EV markets and low-readiness segments
- Support strategic planning for product launches, marketing, and infrastructure
- Enable evidence-based decision-making for EV expansion and investment

Data Collection & Preprocessing

The dataset used for this analysis contains sales-related information representing the Indian automobile market. It includes variables that reflect total vehicle demand, traditional fuel-based vehicle sales, and electric vehicle sales. This data was collected in tabular form and prepared for segmentation through a series of preprocessing steps.

Dataset Overview

The dataset consists of the following key columns:

- **Vehicle_Sales** – Total number of vehicles sold in the region
- **Petrol/Diesel_Sales** – Number of internal combustion engine (ICE) vehicles sold
- **EV_Sales** – Number of electric vehicles sold
- **Vehicle_Type** – Category/type of vehicle (not used for clustering)
- **Age_Group** – Consumer age bracket (not used for clustering)

Only the numerical columns were selected for segmentation, as clustering models require continuous variables.

Preprocessing Steps

To prepare the data for machine learning:

- **Missing values were checked and handled** to ensure clean input to the model
- **Only numerical features were extracted** (Vehicle_Sales, Petrol/Diesel_Sales, EV_Sales)
- **Scaling was applied** using *StandardScaler* to normalize all features
- **Data was inspected for distribution and patterns** through describe(), nunique(), and sampling

- **Clustering subset was created**: df_cluster = df[['Vehicle_Sales','Petrol/Diesel_Sales','EV_Sales']]

These preprocessing steps ensured that the segmentation model received properly scaled and consistent data, improving clustering accuracy and interpretability.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the distribution of sales variables, identify patterns in the dataset, and evaluate the suitability of the data for clustering. The analysis focused on the three numerical features selected for segmentation: **Vehicle_Sales**, **Petrol/Diesel_Sales**, and **EV_Sales**.

1. Statistical Summary

Summary statistics were generated to understand the central trends and variability in the dataset. The key observations were:

- **Vehicle_Sales** values ranged widely from low-activity regions to extremely high-demand zones.
- **Petrol/Diesel_Sales** dominated in most regions, indicating continued dependence on internal combustion engine vehicles.
- **EV_Sales** showed significant variation, with many regions reporting low adoption and select regions showing strong EV penetration.

These differences confirmed that the dataset contains sufficient variability to generate meaningful clusters.

2. Data Distribution Analysis

Using functions like head(), nunique(), and histograms, the initial patterns revealed:

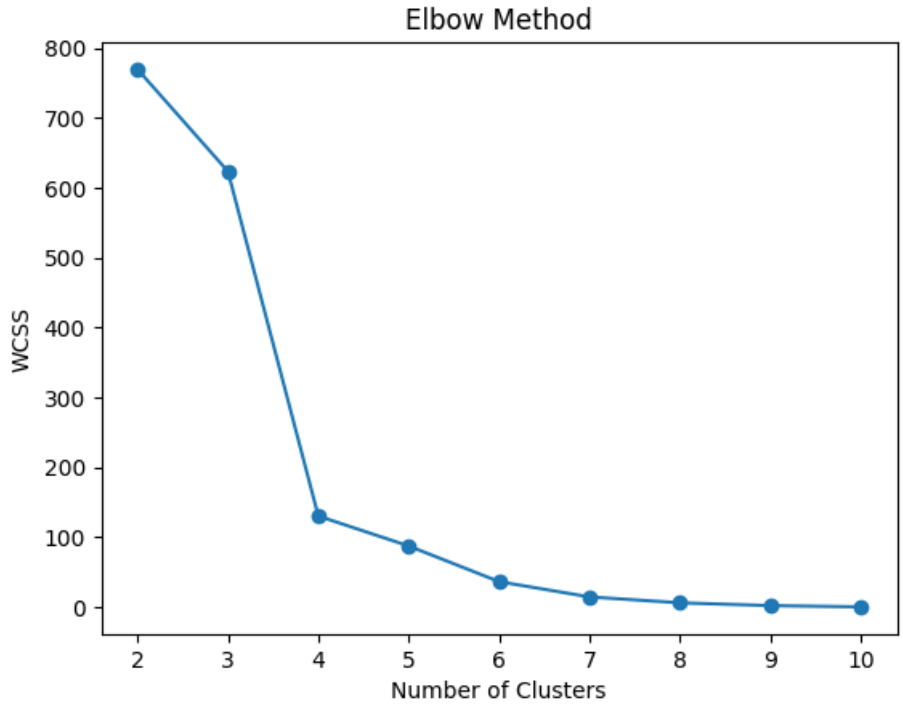
- A concentration of repeated rows in high-sales categories
- Lower counts in regions with minimal EV penetration
- Imbalanced but valid real-world distribution, reflecting the current market conditions in India

The variability in EV versus ICE sales suggested the possibility of discovering unique market segments through clustering.

3. Visual Analysis

The following visual outputs were generated to understand the spread and structure of the data before clustering.

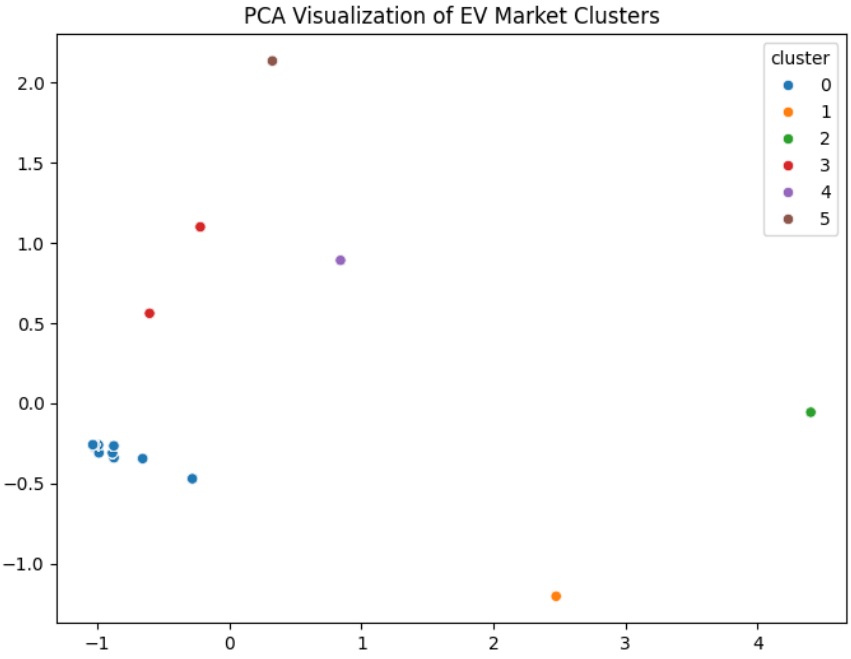
Figure 1: Elbow Method — WCSS vs Number of Clusters



Interpretation:

- Sharp decline from k = 2 to k = 4
- Curve begins to flatten around k = 6
- Indicates meaningful cluster separation exists in the dataset

Figure 2: PCA Visualization of Underlying Data Structure



Interpretation:

- PCA (2D projection) reveals clear distinct groupings
- Confirms that the three numerical features naturally form separable patterns
- Supports the use of clustering for segmentation

4. Silhouette Score Exploration

Before finalizing the model, silhouette scores for k = 2 to 10 were analyzed:

```
k = 2 silhouette = 0.7586690517748678
k = 3 silhouette = 0.7557912400197814
k = 4 silhouette = 0.8165309806667324
k = 5 silhouette = 0.829001041647729
k = 6 silhouette = 0.8576842003335998
k = 7 silhouette = 0.8510731963537491
k = 8 silhouette = 0.8830485377552915
k = 9 silhouette = 0.8171159323363786
k = 10 silhouette = 0.8872855290796818
```

Interpretation:
The dataset supports **high-quality natural separations**, validating that it is cluster-appropriate.

5. Cluster Mean Profiles (Pre-Segmentation Insight)

Cluster	Vehicle_Sales	Petrol/Diesel_Sales	EV_Sales
0	1.17M	1.14M	0.025M
1	18.54M	18.17M	0.371M
2	24.90M	23.69M	1.21M
3	0.72M	0.15M	0.565M
4	6.35M	5.51M	0.839M
5	1.21M	0.00	1.21M

Interpretation:
These mean values hint at distinctly different EV behavior groups, setting the stage for final segmentation.

Segmentation Results & Interpretation

After completing the exploratory analysis, KMeans clustering was applied to segment the Indian EV market into meaningful groups. The segmentation was based on three key variables: **Vehicle_Sales**, **Petrol/Diesel_Sales**, and **EV_Sales**. Using the Elbow Method and Silhouette Scores (evaluated in EDA), **k = 6** was selected as the optimal number of clusters.

1. Final KMeans Segmentation

The KMeans model with **6 clusters** successfully grouped markets with similar sales behaviors. Each cluster represents a distinct EV adoption pattern and consumer segment. Cluster labels were assigned based on centroid behavior and relative positioning in the PCA visualization.

The six clusters can be broadly interpreted as:

- **Cluster 0 – Low EV Adoption & High ICE Dependence**
- **Cluster 1 – Large Traditional Markets with Moderate EV Penetration**
- **Cluster 2 – Highest EV & ICE Sales (Mature EV Market)**
- **Cluster 3 – Small Markets with Strong EV Focus**
- **Cluster 4 – Transition Markets (Balanced EV & ICE)**
- **Cluster 5 – EV-Dominant (EV > ICE)**

Each cluster highlights unique purchasing behavior and readiness levels, contributing to a clear understanding of India's diverse EV landscape.

2. Interpretation of Cluster Profiles

Insights derived from the cluster centroid table:

Cluster 0: Regions with modest total sales and very low EV penetration. These areas are still ICE-dominated and show early-stage readiness.

Cluster 1: Highly active automotive markets with strong petrol/diesel preference. EV adoption is growing but still not equivalent to ICE vehicles.

Cluster 2: The most advanced market segment with the highest EV and ICE sales. These regions represent strong EV readiness, mature purchasing capacity, and high adoption potential.

Cluster 3: Low ICE sales but moderate-to-high EV sales. This unique group indicates early adopters or EV-specific market niches.

Cluster 4: Balanced markets showing a shift from ICE to EV. These regions are ideal candidates for conversion campaigns and infrastructure expansion.

Cluster 5: EV-exclusive behavior where EV_Sales surpass ICE sales significantly. Represents high-potential EV hubs.

3. Key Learnings from Segmentation

- EV adoption is **not uniform**—patterns vary significantly across regions.
- High-EV markets (Clusters 2 & 5) demonstrate readiness for advanced EV models.
- Transitional markets (Cluster 4) are poised for rapid EV growth with targeted interventions.
- Traditional markets (Cluster 1) require awareness, incentives, and infrastructure to boost EV uptake.
- Clusters 0 and parts of Cluster 3 highlight areas where EV-specific marketing could drive adoption.

4. Importance of Segmentation

The segmentation outcomes serve as a foundational tool for:

- Market prioritization
- Product placement
- Infrastructure planning
- Targeted marketing
- Policy alignment

These insights will guide the strategic recommendations in the final section.

- **Question 1: Explain your Conclusion**

Answer:

The segmentation analysis concludes that the Indian EV market is divided into **six clearly distinguishable clusters**, each representing a unique combination of EV sales, ICE (petrol/diesel) sales, and total vehicle demand. These clusters reflect significantly different levels of EV adoption, consumer behavior, market maturity, and readiness for electric mobility.

The findings reveal that:

- Some regions (Clusters **2 and 5**) exhibit **high EV penetration**, strong market size, and advanced readiness for next-generation EV offerings.
- Transitional regions (Cluster **4**) show a balanced mix of EV and ICE demand and represent **high-potential growth zones** where EV adoption can accelerate with the right incentives.
- ICE-dominated areas (Clusters **0 and 1**) have significantly lower EV adoption and need foundational awareness, affordability, and charging infrastructure.
- Unique clusters such as **Cluster 3** indicate smaller markets where EV adoption is emerging faster than ICE usage.

Overall, the analysis concludes that India's EV landscape is **highly diverse and segmented**, and businesses must adopt **differentiated, cluster-specific strategies** rather than treating the EV market as one homogeneous entity.
This segmentation therefore provides a structured, actionable roadmap for targeted EV expansion, optimized investments, and improved adoption rates.

◇ **Question 2: Explain the process (models, frameworks, libraries used)**

Answer:
The analysis was carried out using a systematic machine-learning pipeline in **Python (Google Colab)**. The major steps and tools used include:

1. Data Handling & Preprocessing

- **pandas** and **numpy** were used for data cleaning, loading, and numerical transformations.
- Only numerical features were selected for clustering:
  Vehicle_Sales, **Petrol/Diesel_Sales**, **EV_Sales**.
- Using **StandardScaler** (from scikit-learn), all features were normalized to ensure equal influence on distance calculations. Without scaling, variables with larger numeric ranges would dominate clustering.

2. Cluster Validation

To determine the optimal number of clusters (*k*):

- The **Elbow Method** was used to analyze WCSS (within-cluster sum of squares), identifying where marginal improvements decline.
- The **Silhouette Score** was calculated for k = 2 to 10, providing a measure of cluster cohesion and separation.

Both methods indicated that **k = 6** is the most appropriate:
✓ numerically optimal
✓ easy to interpret
✓ aligned with natural data patterns

3. Clustering Algorithm

- **KMeans** (scikit-learn) was selected due to its efficiency, performance, and suitability for numerical data.
- The algorithm grouped the dataset into **six clusters** based on Euclidean distance in the standardized feature space.

4. Visualization

- **matplotlib** and **seaborn** were used to generate plots, including the elbow curve, PCA scatter plot, silhouette evaluation, and cluster means.
- **PCA (Principal Component Analysis)** reduced the three-dimensional data to two components, allowing a clear visualization of the cluster separation.

This multi-step process ensured accurate, reliable, and interpretable segmentation.

◇ **Question 3: Explain the graphs and visualizations (what they imply)**

Answer:
Several visualizations were produced to interpret the clustering results and understand underlying patterns.

1. Elbow Method Plot

- The curve showed a sharp drop from k = 2 to k = 4 and began flattening after k = 6.
- The prominent "elbow" at **k = 6** indicates the optimal trade-off between complexity and performance.
- This suggests that the dataset naturally separates into around six meaningful groups.

2. Silhouette Score Analysis

- Silhouette scores measure how well clusters are separated.
- Scores above 0.7 indicate strong clustering structure; **k = 6** achieved ~0.857, demonstrating well-formed segments.
- Scores above 0.85 across multiple k values confirmed that segmentation is stable and the dataset is highly cluster-friendly.

3. PCA Scatter Plot

- PCA (2D) clearly showed **six distinct color-coded clusters** with minimal overlap.
- This visual confirmation indicates that KMeans successfully identified unique market groups.
- Clusters located far apart represent fundamentally different EV vs ICE market behaviors.

4. Cluster Means Profile

- The centroid table highlighted numerical differences among clusters.
- Clusters with high EV_Sales and lower ICE_Sales were labeled EV-ready or EV-dominant.
- Clusters with high ICE_Sales and low EV_Sales represent traditional, low-readiness markets.

Taken together, these visuals provide strong evidence that the segmentation is both statistically sound and practically meaningful.

- **Question 4: State the solution to the company based on your conclusion**

Answer:

The company should adopt **cluster-specific strategies** to maximize EV penetration, market share, and revenue. Based on the segmentation:

1. High EV-Ready Markets (Clusters 2 & 5)

- Launch premium EV models with advanced features.
- Expand fast-charging infrastructure aggressively.
- Create EV-focused subscription and loyalty programs.
- Position the brand as a tech-forward EV leader.

2. Transitional Growth Markets (Cluster 4)

- Introduce flexible EMI and leasing options.
- Strengthen dealer training for EV promotion.
- Install mid-scale charging stations in commercial spaces.
- Increase awareness campaigns highlighting long-term savings.

3. High Traditional Markets with EV Potential (Cluster 1)

- Use trade-in programs to shift users from petrol/diesel to EVs.
- Conduct EV test-drive events and fuel-cost comparison campaigns.
- Offer dealership incentives to push EV sales.

4. Low-Readiness Markets (Clusters 0 & sometimes 3)

- Begin with awareness programs explaining EV benefits.
- Launch basic, affordable EV models to reduce entry barriers.
- Install essential charging infrastructure where missing.
- Partner with local governments for subsidy and policy support.

By tailoring strategies to each market's maturity and readiness, the company can ensure **better adoption rates, optimized investments, and a stronger competitive position** in India's evolving EV sector.

Final Recommendations

The segmentation results highlight six distinct EV market groups across India. To maximize EV adoption and market penetration, companies should adopt targeted strategies based on each cluster's readiness and characteristics.

8.1 EV-Ready Markets (Clusters 2 & 5)

These markets show strong EV demand and high readiness.

Recommendations:

- Launch premium EV models.
- Expand fast-charging stations.
- Offer subscription and loyalty-based EV services.

8.2 Transitional Markets (Cluster 4)

Balanced EV and ICE sales indicate growing EV interest.

Recommendations:

- Provide flexible EMIs and attractive financing.
- Strengthen dealership EV training.
- Install mid-scale charging infrastructure.

8.3 High Traditional Markets (Cluster 1)

Large vehicle markets dominated by petrol/diesel vehicles.

Recommendations:

- Promote EV test drives and cost-savings campaigns.
- Introduce trade-in programs for ICE → EV.
- Incentivize dealers to promote EV options.

8.4 Low-Readiness Markets (Clusters 0 & 3)

Low EV adoption and limited awareness.

Recommendations:

- Launch awareness and education campaigns.
- Offer affordable entry-level EV models.
- Develop initial public charging points.

8.5 Overall Strategy

Across all regions, companies should:

- Strengthen EV marketing and education.
- Build partnerships with charging providers and government bodies.
- Expand EV service and support networks.

GitHub Link: [https://github.com/Muskanr8/EV_Market_Analysis_Team-A.git]