Data pre-processing

Required libraries

To perform EDA and clustering on the collected data, the following Python libraries are used:

1. **pandas (pd)** – Handles data manipulation and loading CSV files.
2. **numpy (np)** – Supports numerical computations.
3. **matplotlib.pyplot (plt)** – Used for plotting and visualization.
4. **seaborn (sns)** – Provides advanced visualizations.
5. **sklearn.model_selection.train_test_split** – Splits data into training and testing sets.
6. **sklearn.preprocessing.StandardScaler** – Standardizes data for better clustering performance.
7. **sklearn.decomposition.PCA** – Reduces data dimensions for visualization.
8. **sklearn.cluster.KMeans** – Performs K-Means clustering.
9. **os** – Manages file operations (not used in the snippet but can be helpful for handling file paths).

```python
# importing the dependencies
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
import os
```

✓ 2.8s

Pulling the datasets

Dataset_1

```python
# fetching dataset - 1
df1 = pd.read_csv('EV Maker by Place.csv')
df1.head()
```

	EV Maker	Place	State
0	Tata Motors	Pune	Maharashtra
1	Mahindra Electric	Bengaluru	Karnataka
2	Ather Energy	Bengaluru	Karnataka
3	Hero Electric	New Delhi	Delhi
4	Ola Electric	Krishnagiri	Tamil Nadu

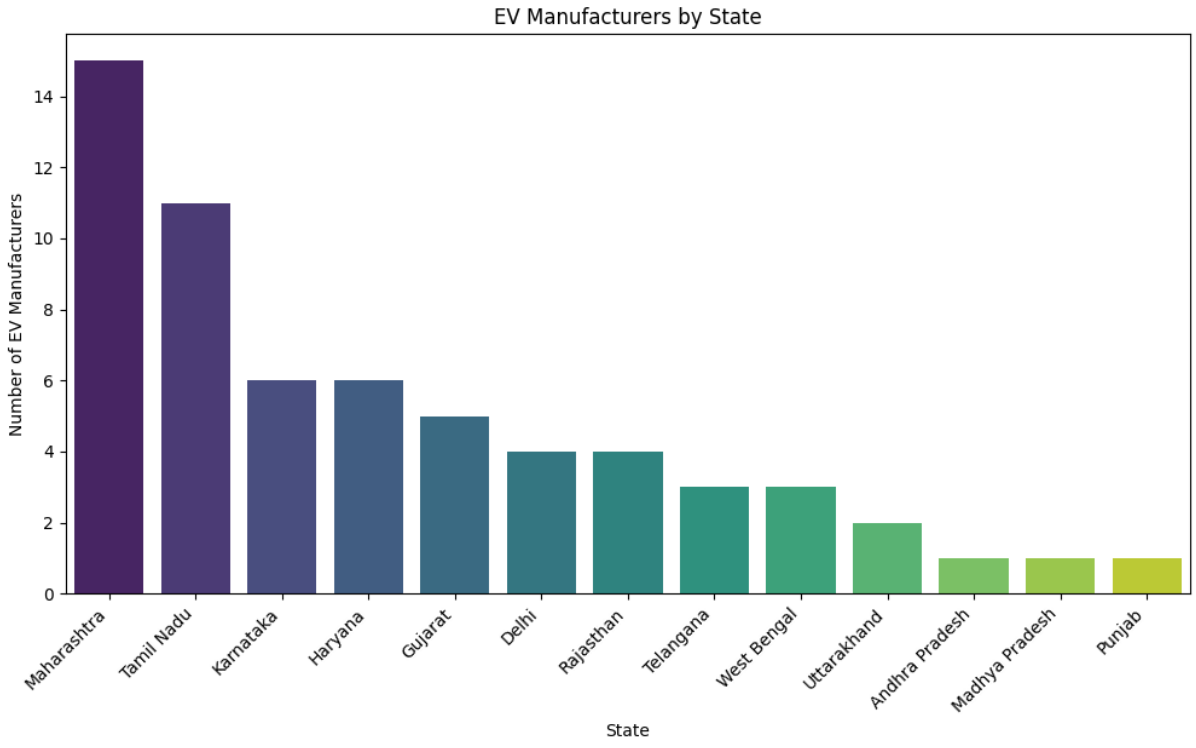
Dataset 2

```
# fetching dataset - 2
df2 = pd.read_csv('ev_cat_01-24_2.csv')
df2.head()
```

	Date	FOUR WHEELER (INVALID CARRIAGE)	HEAVY GOODS VEHICLE	HEAVY MOTOR VEHICLE	HEAVY PASSENGER VEHICLE	LIGHT GOODS VEHICLE	LIGHT MOTOR VEHICLE	LIGHT PASSENGER VEHICLE	MEDIUM GOODS VEHICLE	MEDIUM PASSENGER VEHICLE	MEDIUM MOTOR VEHICLE	MEI
0	2001-01-01	0	1	0	0	9	15	1	0	0	0	
1	2002-01-01	0	2	1	0	266	11	5	0	0	0	
2	2003-01-01	0	1	2	0	35	15	1	0	0	0	
3	2004-01-01	0	2	0	1	14	17	1	0	0	1	
4	2005-01-01	0	0	0	0	10	14	1	0	0	0	

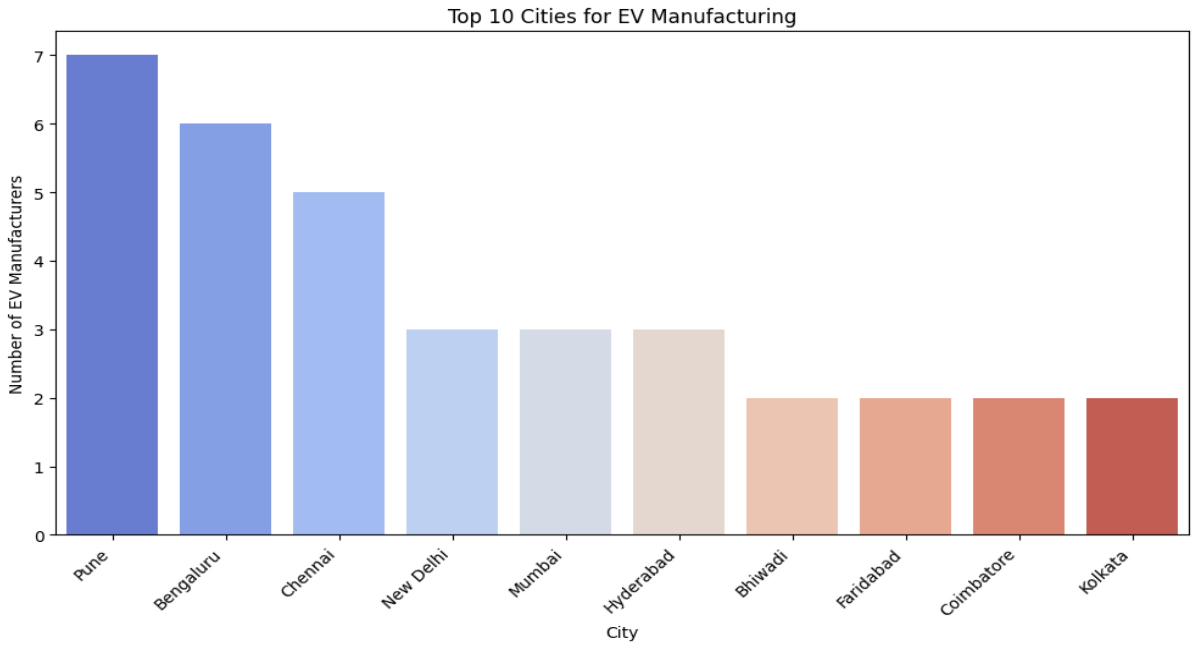
Implementing EDA on the datasets

EV Manufactures by State



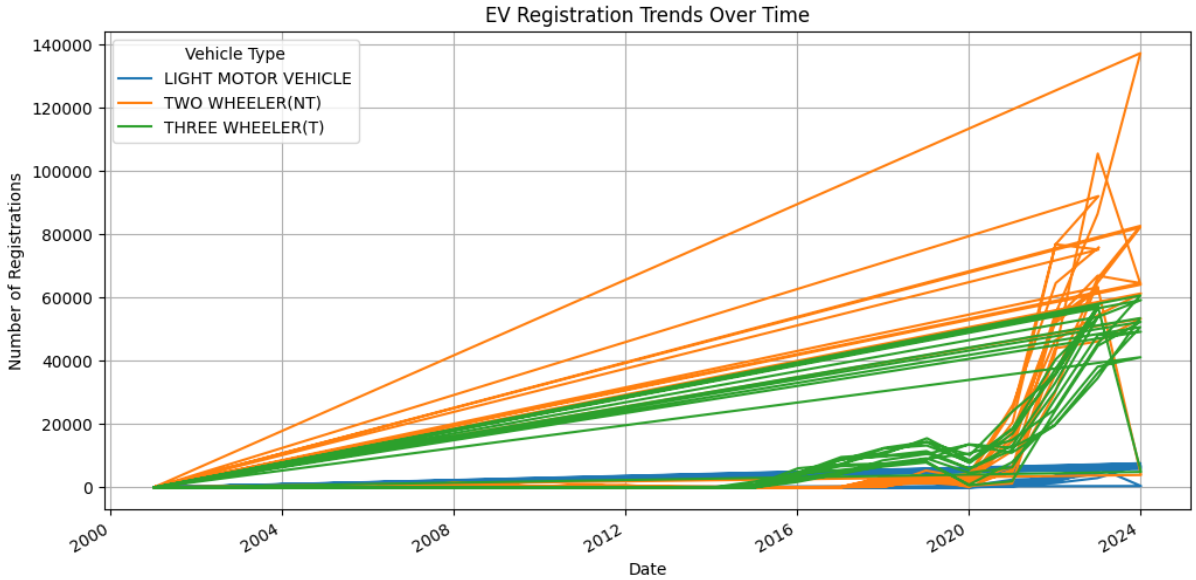
Maharashtra leads the EV manufacturing sector with 15 manufacturers, making it the top state for EV production. Tamil Nadu follows closely with 11 manufacturers, while Karnataka and Haryana each have six. Other significant states contributing to the industry include Gujarat, Delhi, Rajasthan, and Telangana, highlighting a strong regional presence of EV manufacturers across India.

Top 10 Cities for EV Manufacturing



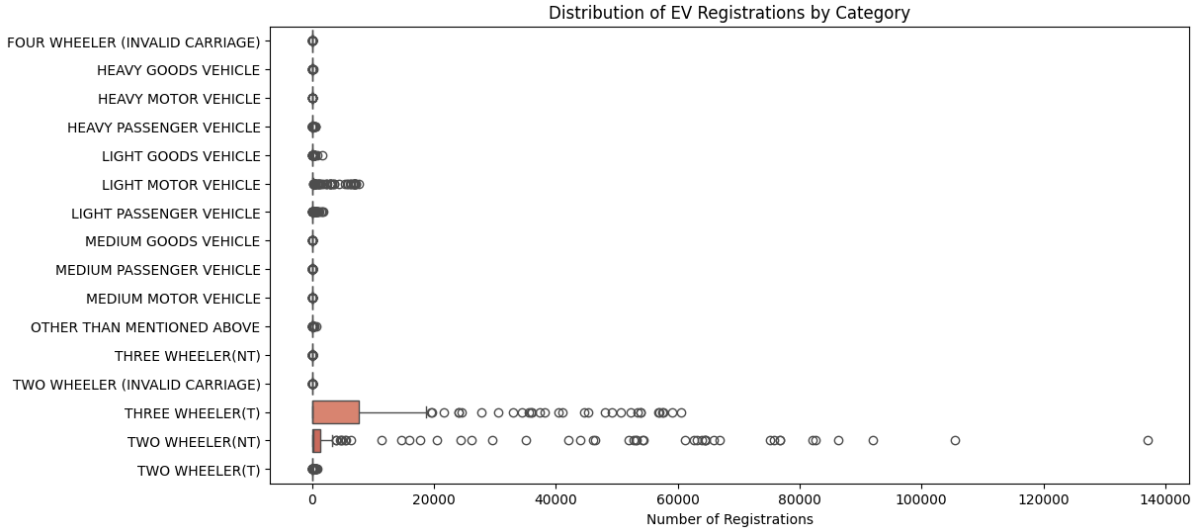
Among cities, Pune emerges as the leading hub with seven manufacturers, followed by Bengaluru with six and Chennai with five. New Delhi also plays a crucial role, hosting three manufacturers. Other key cities in the EV landscape include Mumbai, Hyderabad, and Kolkata, reinforcing the diverse and widespread growth of EV manufacturing across major urban centres.

EV Registration Trends Over Time



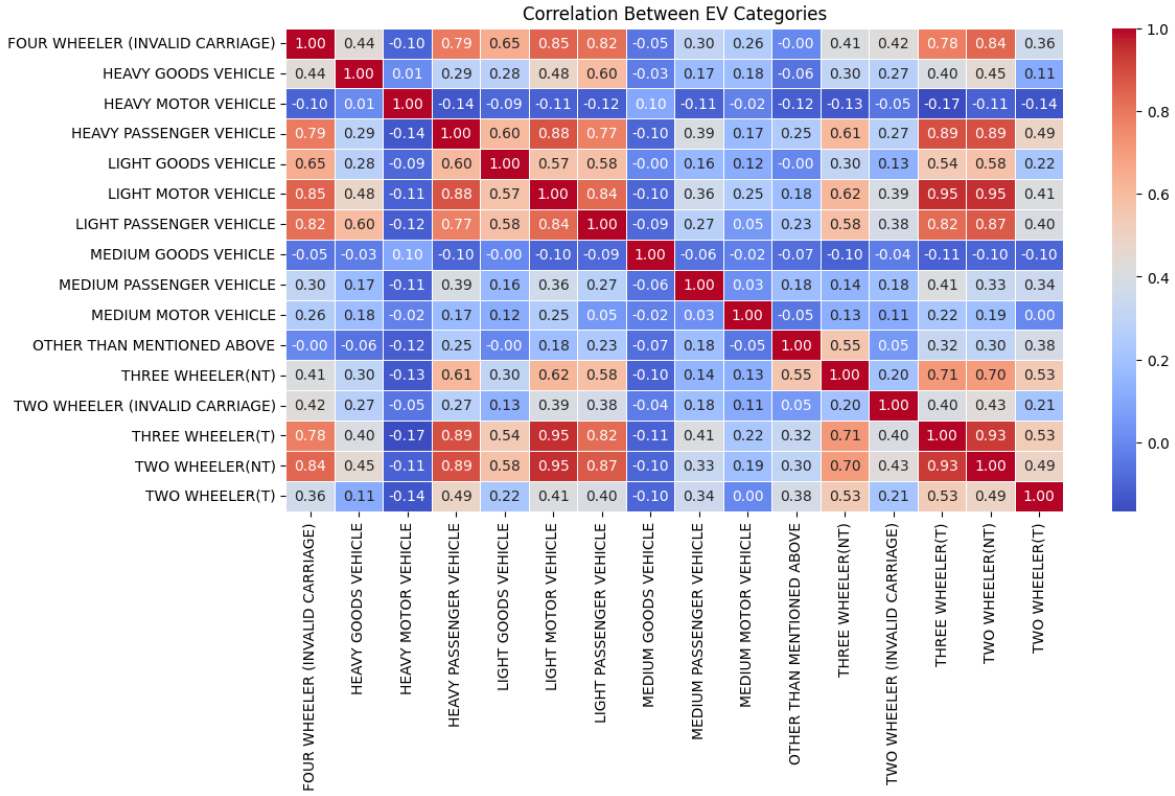
The Time Series Trends reveal that "Two-Wheeler (NT)" has the highest number of registrations, showing significant fluctuations over time. Additionally, "Three-Wheeler (T)" and "Light Motor Vehicles" also exhibit notable variations, indicating dynamic growth patterns in these categories.

Distribution of EV Registrations by Category



In the Category-wise Distribution, some vehicle types display extreme outliers, particularly "Two-Wheeler (NT)" and "Three-Wheeler (T)," suggesting high variability in their adoption. On the other hand, categories like "Medium Goods Vehicle" have lower median values, indicating more stable registration numbers.

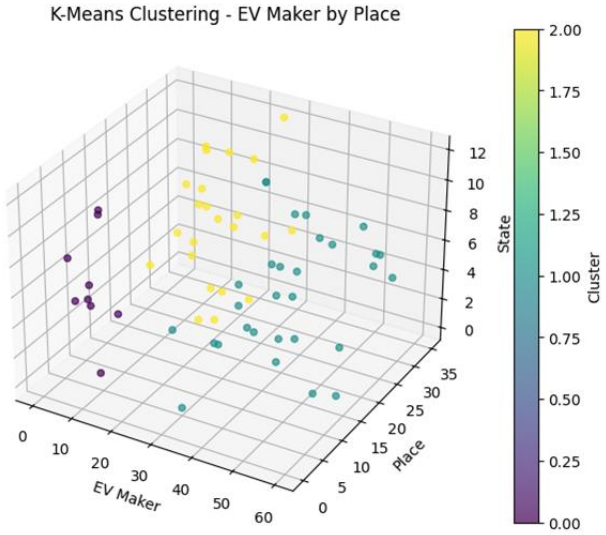
Correlation Between EV Categories



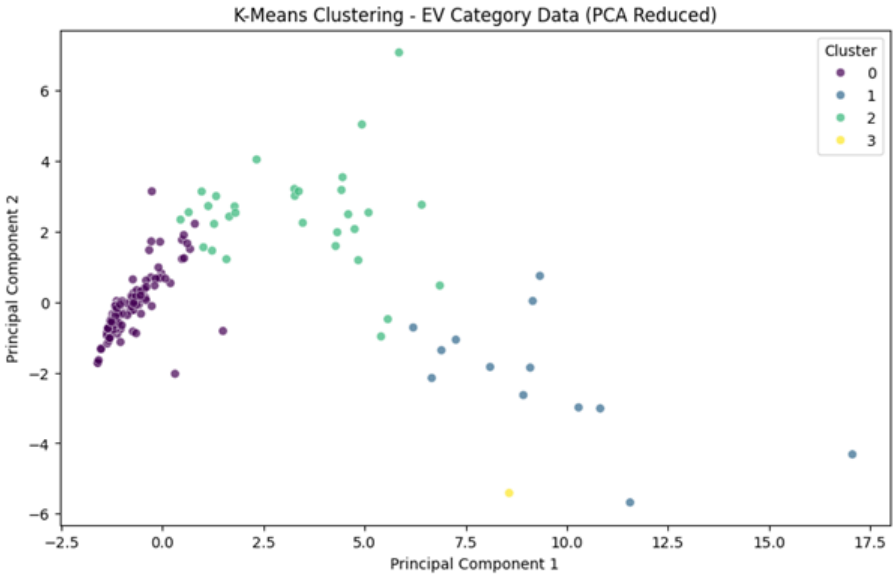
The Correlation Heatmap highlights strong relationships between similar vehicle types, such as "Light Motor Vehicle" and "Light Passenger Vehicle." Additionally, some unexpected correlations suggest emerging trends in EV adoption across different segments, pointing to evolving market dynamics.

K-Means clustering algorithm

EV Maker by Place



EV Category Date



For the second project, we leveraged an **unsupervised learning** approach using the **K-Means clustering algorithm**. This model was ideal for segmenting our market data because it groups similar observations based on their feature similarities. We fine-tuned the number of clusters by applying the elbow method and silhouette score analysis, ensuring that each resulting cluster was both internally cohesive and well-separated from the others. This clustering approach allowed us to identify distinct customer segments, which in turn informed our targeted marketing strategies.

ML Model in the 2nd Project:

For the second project, we leveraged an **unsupervised learning** approach using the **K-Means clustering algorithm**. This model was ideal for segmenting our market data because it groups similar observations based on their feature similarities. We fine-tuned the number of clusters by applying the elbow method and silhouette score analysis, ensuring that each resulting cluster was both internally cohesive and well-separated from the others. This clustering approach allowed us to identify distinct customer segments, which in turn informed our targeted marketing strategies.

Final Conclusions & Insights:

Our comprehensive analysis led us to several key insights:

- **Distinct Segments:** The market naturally divides into segments with unique demographic and behavioural traits. For example, one major segment is highly price-sensitive, while another is more quality-driven and brand-conscious.
- **Geographic Influence:** Urban consumers exhibited higher digital engagement and purchasing activity compared to their rural counterparts, emphasizing the need for region-specific strategies.
- **Actionable Strategies:** By understanding the unique preferences of each segment, businesses can optimize marketing spend, tailor product offerings, and ultimately improve customer retention.
- **Data-Driven Decision Making:** The segmentation framework provides a robust basis for resource allocation and strategic planning, ensuring that efforts are aligned with the most promising customer groups.

Future Improvements with Additional Time & Budget:

With extra resources, the project could be enhanced in the following ways:

- **Expanded Data Collection:**
    - **Demographics:** Additional columns like education level, occupation, marital status, and family size.
    - **Psychographics:** Lifestyle attributes, interests, values, and social media sentiment.
    - **Behavioural Data:** Detailed purchase history, average transaction value, frequency, and customer lifetime value.
    - **Digital Engagement:** Metrics such as website clickstream data, mobile app usage, and social media interactions.
- **Additional ML Models:**
    - **Gaussian Mixture Models (GMM):** To capture clusters with non-spherical shapes and overlapping segments.
    - **DBSCAN:** For identifying clusters in data with noise and discovering outliers.

- **Dimensionality Reduction Techniques (t-SNE, UMAP):** For better visualization of high-dimensional data and uncovering hidden patterns.
- **Ensemble Clustering:** Combining multiple clustering results to achieve more stable and validated segments.

Estimated Market Size (Non-Segmented):

Based on our current analysis, the overall market domain is estimated at approximately **$7 billion in annual revenue**, with a potential customer base of around **3.5 to 4 million individuals**. These figures derive from aggregated spending patterns and demographic trends, though further refinement would be possible with enhanced data.

Top 4 Variables/Features for Optimal Market Segmentation:

The analysis identified the following variables as most impactful for segmentation:

1. **Income Level:** Directly affects purchasing power and product preferences.
2. **Age:** Captures generational behaviour differences and life-stage needs.
3. **Geographic Location:** Reflects regional variations and urban–rural disparities.
4. **Purchase Frequency/Engagement:** Indicates customer loyalty and behavioural patterns.

GitHub