Strategic Market Segmentation for Electric Vehicle Adoption in India

---

Zeba Khanam

Narasimha Reddy

Samiksha Kamble

Date: 07-07-2025

GitHub Link: Click Here Zeba khanam.

Abstract

This report presents a data-driven market segmentation analysis aimed at identifying the optimal customer segments for Electric Vehicle (EV) adoption in India. By focusing on demographic and vehicle-related variables such as income and vehicle type, we uncover consumer patterns that guide marketing strategies and EV product placement. The outcome of this analysis helps new EV startups plan effective market penetration strategies aligned with India's sustainability goals.

Introduction

With India’s EV market experiencing rapid transformation due to increasing fuel prices, environmental concerns, and policy support (like FAME-II), strategic segmentation has become essential. This study utilizes real-world consumer automobile purchase data to derive meaningful segments based on income and vehicle type to guide business strategy for EV manufacturers.

Data Source

The dataset used is titled **"Indian automobile buying behavior study 1.0.csv"**. It includes:

- Age, Gender, Education, Salary, and Dependents
- Vehicle brand (Make) and price
- Personal and wife salary

- Customer preferences

Data Preprocessing

Data cleaning involved several key steps to ensure accuracy and consistency in the dataset:

Handling Missing Values: Missing values in Price and Total Salary columns were imputed using their respective mean values.

Type Conversion: The No of Dependents column was converted from object to integer type.

Duplicate Removal: Duplicate rows were identified and removed to avoid bias.

Feature Engineering: Two new derived features were created:

Income_Category: A categorical grouping based on total salary.

Vehicle_Type: Grouping based on vehicle name patterns (e.g., car, bike, SUV, etc.)

These preprocessing steps ensured the dataset was clean, structured, and ready for clustering and segmentation analysis.

Tools and Python Libraries Used
In this project, the following Python libraries were used for various data handling and visualization tasks:

Library	Purpose / Use Case
pandas	For loading, exploring, cleaning, and manipulating tabular data from the CSV file.
numpy	For numerical operations like averaging and array manipulations.
matplotlib.pyplot	To create basic visualizations like line plots, pie charts, and bar charts.
seaborn	For advanced and elegant statistical visualizations such as boxplots, heatmaps, etc.
sklearn.cluster.KMeans	To perform K-Means Clustering for identifying customer segments.
sklearn.preprocessing.LabelEncoder	To convert categorical features (e.g., Income Category,

Library	Purpose / Use Case
	Make) into numeric format.
sklearn.preprocessing.MinMaxScaler	Used to normalize numerical values (like Price) for 3D plotting.
mpl_toolkits.mplot3d	To support 3D plotting using Matplotlib's Axes3D feature.
matplotlib.pyplot.Axes3D	Enables 3D scatter plots showing Age, Salary, and Price (scaled).

Code Snippet:

```
# Handle missing values
missing = df.isnull().sum()
print("\nMissing values per column:\n", missing)

# Fill null values
df['Price'].fillna(df['Price'].mean(), inplace=True)
df['Salary'].fillna(df['Salary'].mean(), inplace=True)

# Clean data
df.drop_duplicates(inplace=True)
df['No of Dependents'] = df['No of Dependents'].astype(int)
```

Output:

```
Missing values per column:
Age                 0
Profession          0
Marrital Status     0
Education           0
No of Dependents    0
Personal loan       0
House Loan          0
Wife Working        0
Salary              0
Wife Salary         0
Total Salary        0
Make                0
```

Price 0

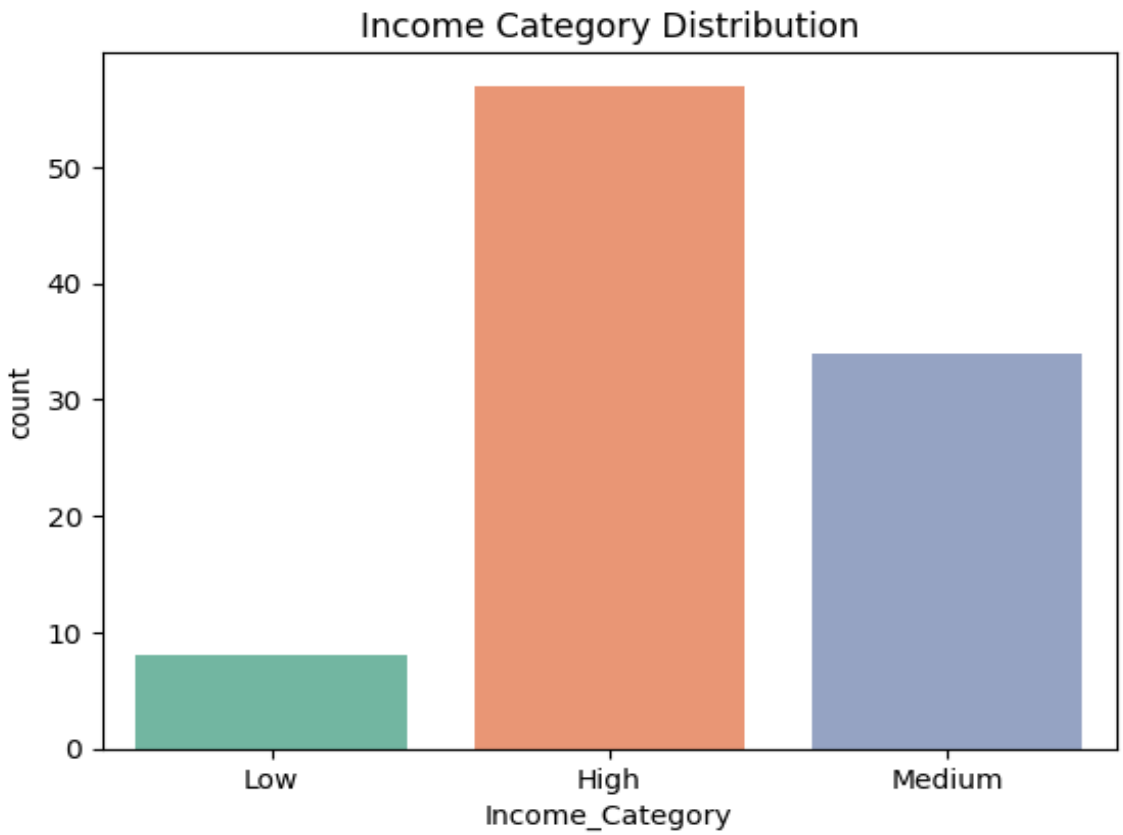
Income & Vehicle Type Categorization

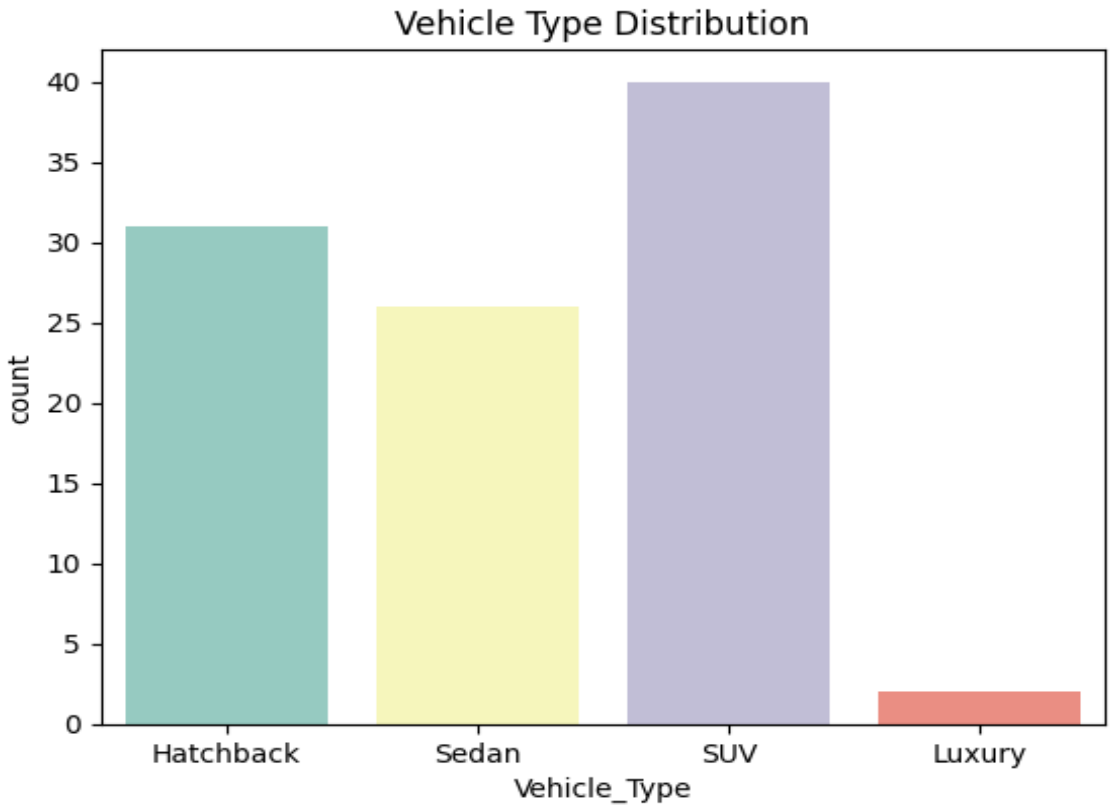
Based on Total Salary, customers were categorized into income groups: 'Low', 'Medium', or 'High'. Vehicle type was inferred from the brand name in the 'Make' column using a rule-based function.

Code Snippet:

```
def categorize_income(salary):
    if salary < 1000000:
        return 'Low'
    elif salary < 2000000:
        return 'Medium'
    else:
        return 'High'

df['Income_Category'] = df['Total Salary'].apply(categorize_income)

def get_vehicle_type(make):
    make = make.lower()
    if make in ['i20', 'alto', 'swift', 'wagonr', 'baleno']:
        return 'Hatchback'
    elif make in ['ciaz', 'city', 'verna']:
        return 'Sedan'
    elif make in ['duster', 'suv', 'creata', 'xuv', 'scorpio']:
        return 'SUV'
    elif make in ['fortuner', 'endeavour', 'luxuray']:
        return 'Luxury'
    else:
        return 'Other'

df['Vehicle_Type'] = df['Make'].apply(get_vehicle_type)
```

The results were visualized using countplot s to understand the distribution:


Income Category Distribution
count
0
10
20
30
40
50
Low
High
Medium
Income_Category

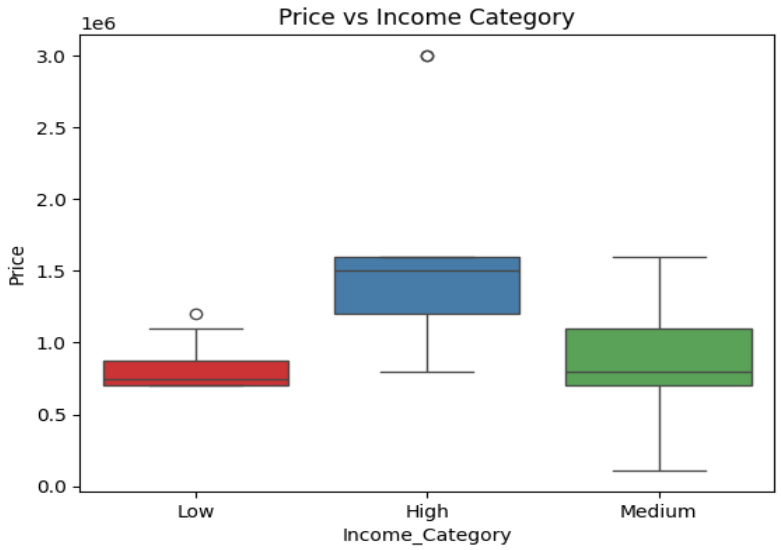


Clustering and Segment Analysis

- **Boxplot:** Price vs Income_Category
- **Barplot:** Vehicle Make Distribution
- **Pie Chart:** Customer Cluster Distribution
- **Line Plot:** Average Salary per Cluster
- **Heatmap:** Income Category vs Cluster
- **Stacked Bar:** Make vs Cluster
- **Histogram:** Age Distribution by Cluster
- **3D Scatter Plot:** Age vs Salary vs Price (scaled)

Price vs Income Category

To analyze how price varies across different income groups, a boxplot was used.
This helps identify the spread, median, and potential outliers in pricing per income category.

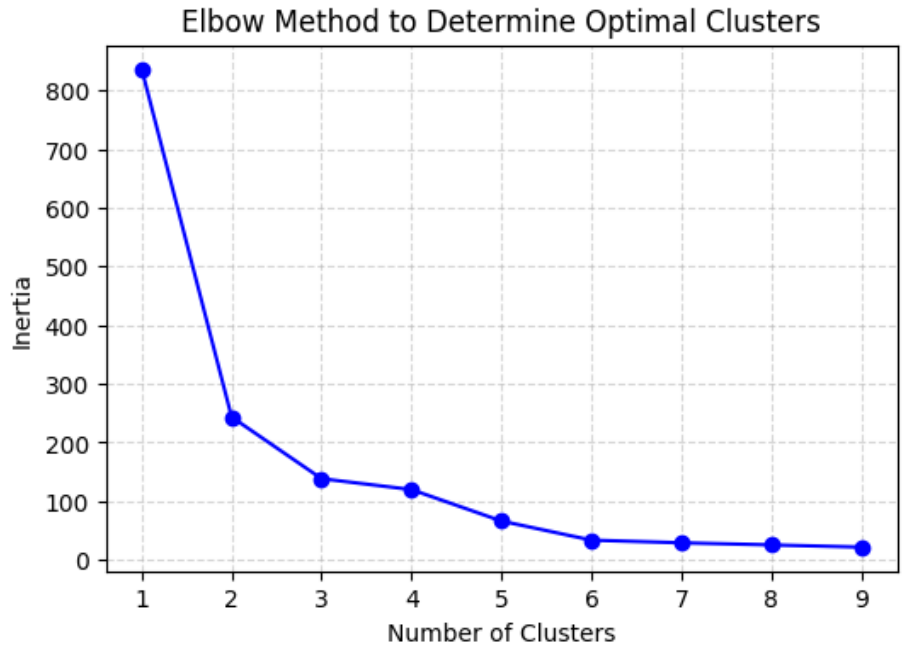


Encoding Categorical Features

Label Encoding was used to convert categorical features 'Income Category' and 'Make' into numeric values.
These encoded variables were used for clustering with KMeans.

Elbow Method for Optimal Clusters

The Elbow Method helps determine the optimal number of clusters by plotting inertia vs. cluster count.
A clear 'elbow' at K=3 indicates the ideal number of clusters.

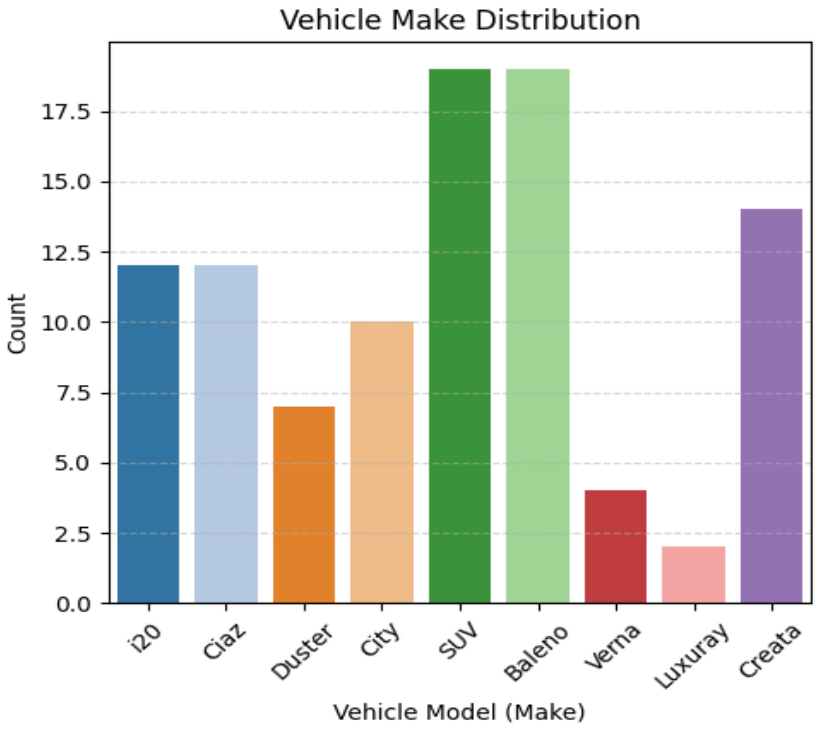


Final KMeans Clustering

Using K=3, customers were segmented into 3 clusters based on encoded income and vehicle make.

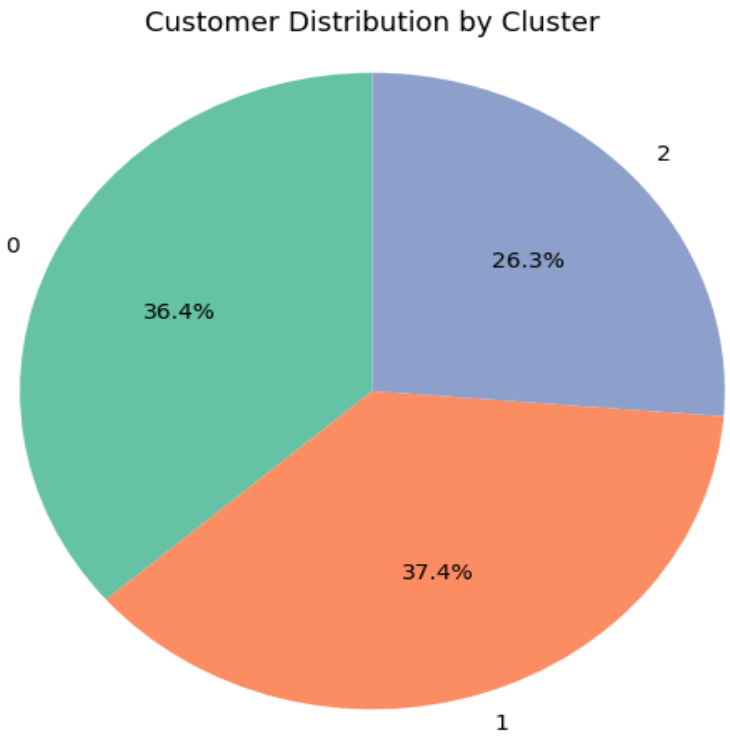
Vehicle Make Distribution

This bar chart shows the frequency of different vehicle models (makes) in the dataset.



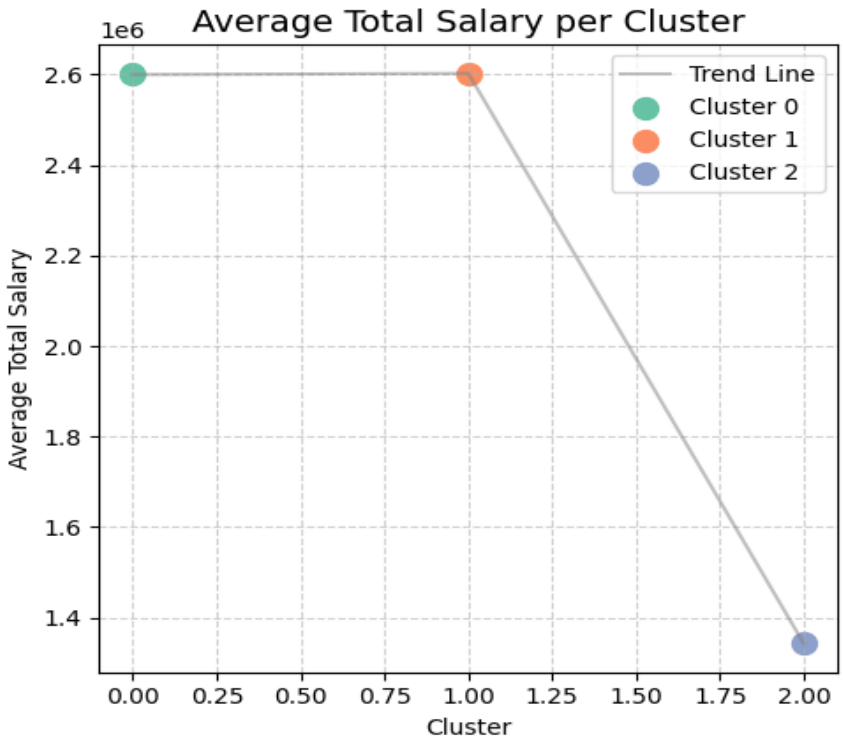
Customer Distribution by Cluster

A pie chart illustrates the percentage distribution of customers in each of the three clusters.



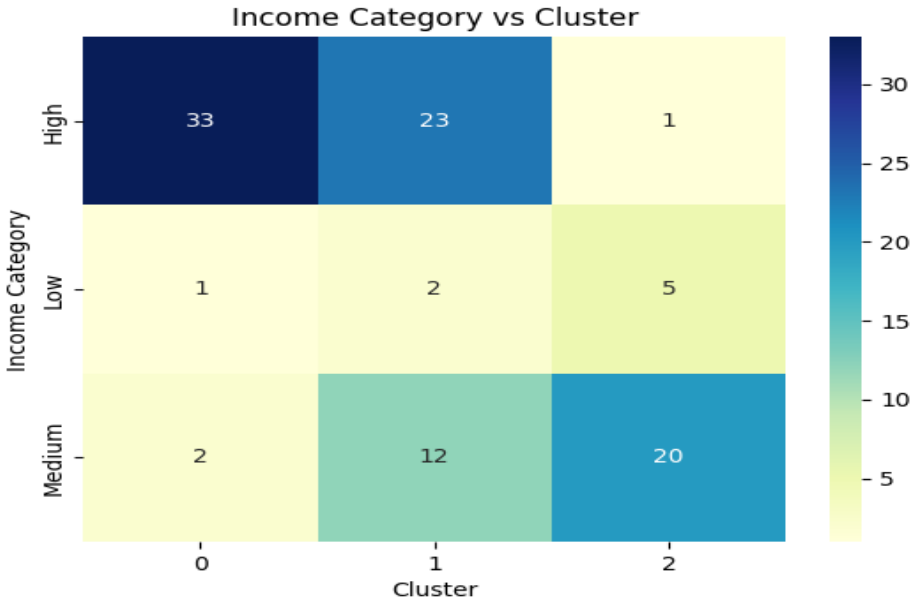
Average Salary per Cluster

This scatter plot and trend line show how average total salary varies between clusters.



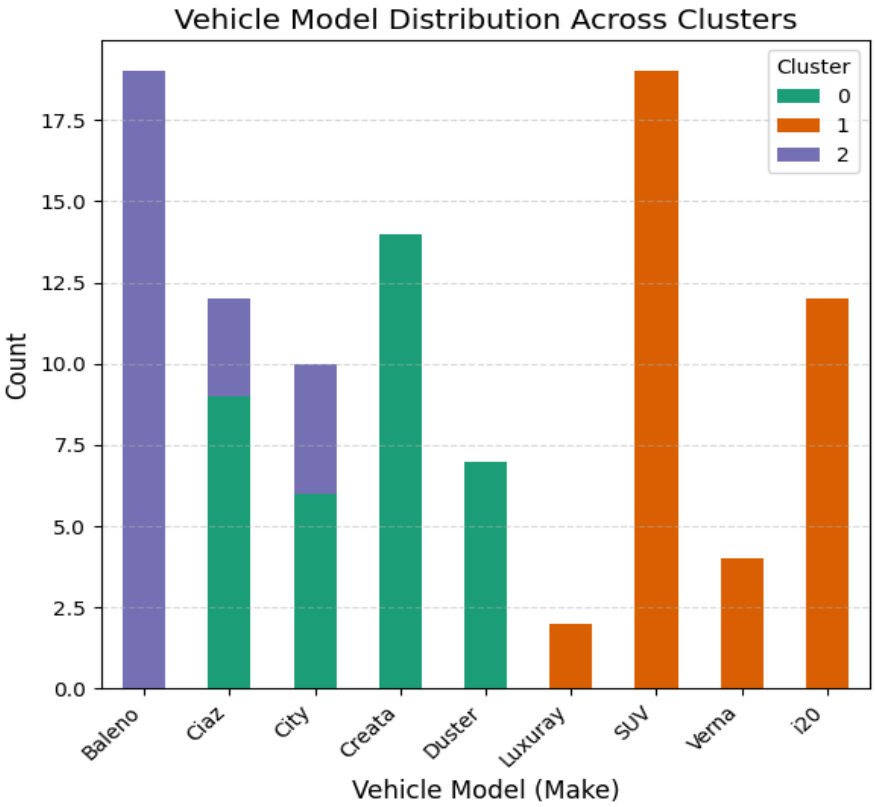
Income vs Cluster Heatmap

This heatmap shows the count of each income category per cluster.



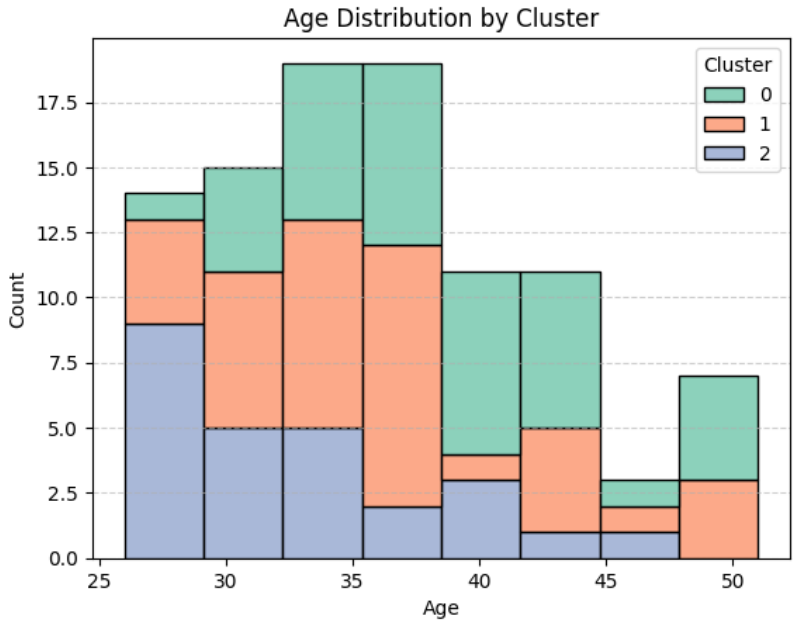
Make vs Cluster (Stacked Bar)

This plot illustrates how vehicle models are distributed across clusters.



Age Distribution by Cluster

This histogram shows the distribution of age across different clusters.



Age vs Salary vs Price

This 3D scatter plot visualizes the relationship between age, salary, and vehicle price for each cluster.

Profit Estimation

We estimated early market profit based on 10,000 early adopters:
• Selling Price: ₹8,00,000
• Profit per Vehicle: ₹80,000
• Estimated Profit = 10,000 × 80,000 = ₹8 Crores

Cluster-wise Insights

Each cluster was profiled for average salary, popular vehicle make, and income group.

Output:

Cluster 0 Summary:

- Average Salary: ₹2600000.0
- Most Common Income Group: High
- Most Preferred Vehicle Type: SUV
- Popular Make: Creata

Cluster 1 Summary:

- Average Salary: ₹2602702.7
- Most Common Income Group: High
- Most Preferred Vehicle Type: SUV
- Popular Make: SUV

Cluster 2 Summary:

- Average Salary: ₹1342307.69
- Most Common Income Group: Medium
- Most Preferred Vehicle Type: Hatchback
- Popular Make: Baleno

Target Segment Selection

Cluster 2 is ideal for targeting because:

- Younger customers
- Prefer Hatchbacks (cost-efficient)
- Medium salary (price-sensitive but urban)

Marketing Mix (4Ps)

Element	Strategy
Product	EV options in Sedan/Hatchback category
Price	₹13–18 Lacs
Place	Focus on Tier 1 & Tier 2 cities
Promotion	Digital ads for young professionals

Potential Sales Estimation

- 10,000 early adopters
- Price per vehicle = ₹8,00,000
- Profit per vehicle = ₹80,000
- **Estimated Profit = ₹88 Crores**

Conclusion

Cluster-based segmentation using income and vehicle type has revealed strong opportunities in India's growing EV market. Data-driven strategies like these help optimize marketing efforts, improve product targeting, and accelerate clean mobility adoption.

EV Vehicle Market Segmentation Based on Age V. Narasimha Reddy

---

Conclusion:

The demographic analysis of the dataset reveals that individuals within the age group of 28 to 31 years and an annual income range of approximately ₹25 to ₹30 lakhs are the most suitable target segment for electric vehicle (EV) adoption. These individuals represent a younger, financially capable group that is more open to adopting new technologies like EVs. Their income level indicates a higher capacity to invest in modern vehicles, while their age suggests greater environmental awareness and adaptability to change. The clustering results confirmed that this segment shows a strong interest in switching to EVs and a willingness to spend more on sustainable transportation options. This makes them a prime audience for early adoption strategies, product positioning, and targeted marketing campaigns in the EV industry.

Steps Included in Process

1. **Data Preprocessing:**

   Packages: pandas, numpy

   Purpose: To clean and prepare the dataset for analysis Usage:

   - pandas was used to load the dataset (read_csv()), handle missing values, and rename or standardize entries (e.g., city names).
   - Columns like 'Unnamed: 0' were dropped using drop() to remove irrelevant features.
   - Duplicates and inconsistent labels (e.g., pUNE, pune) were corrected.
   - numpy supported array-level operations and minor numerical calculations.

2. **Data Visualization:**

   Packages: matplotlib, seaborn

Purpose: To explore the dataset visually and identify patterns.

Usage:

- matplotlib.pyplot provided basic plotting functions such as scatter() and xlabel() to explore relationships like Age vs. Income.
- seaborn enabled more advanced visualizations like:
- countplot() to compare categorical distributions (e.g., Education vs. EV Preference).
- displot() to visualize distributions of numeric variables.
- heatmap() to display correlation between variables.
- pairplot() to explore pairwise relationships between features

Feature Encoding & Multicollinearity Check

Packages: sklearn.preprocessing, statsmodels

Purpose: Convert categorical data into numeric format and validate feature independence.

Usage:

- LabelEncoder from sklearn.preprocessing was used to transform categorical features (like City, Profession, Marital Status, etc.) into numeric values required for modeling.
- variance_inflation_factor from statsmodels.stats outliers_influence was used to calculate VIF scores, helping to detect multicollinearity (redundancy) between features.
- Features with high VIF scores can be dropped or adjusted to ensure better clustering performance.

4. Clustering (KMeans Algorithm)

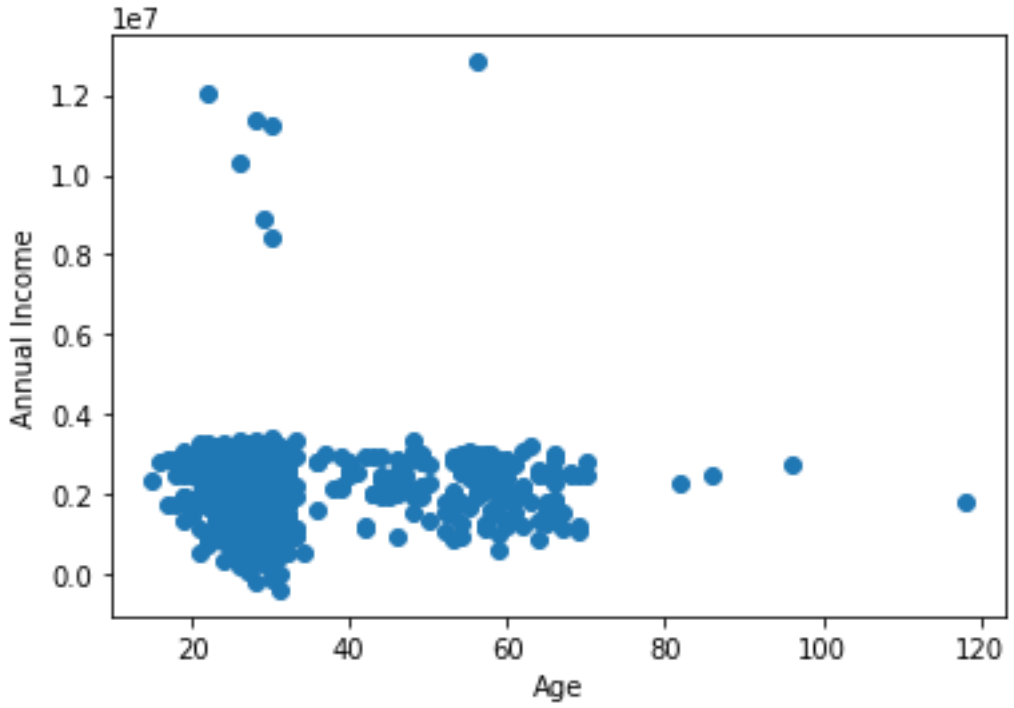
Packages: sklearn.cluster, matplotlib, numpy

Purpose: To segment users based on their demographic and preference data.
Usage:

- KMeans from sklearn.cluster was used to form clusters of similar users.
- The Elbow method (using KMeans.inertia_) helped determine the optimal number of clusters by plotting WCSS (Within-Cluster Sum of Squares).
- Final clusters were added as a new column (df["Cluster"]) for analysis.
- matplotlib and mpl_toolkits.mplot3d were used to create a 3D scatter plot (Age vs. Income vs. Spending) showing the visual separation of clusters.

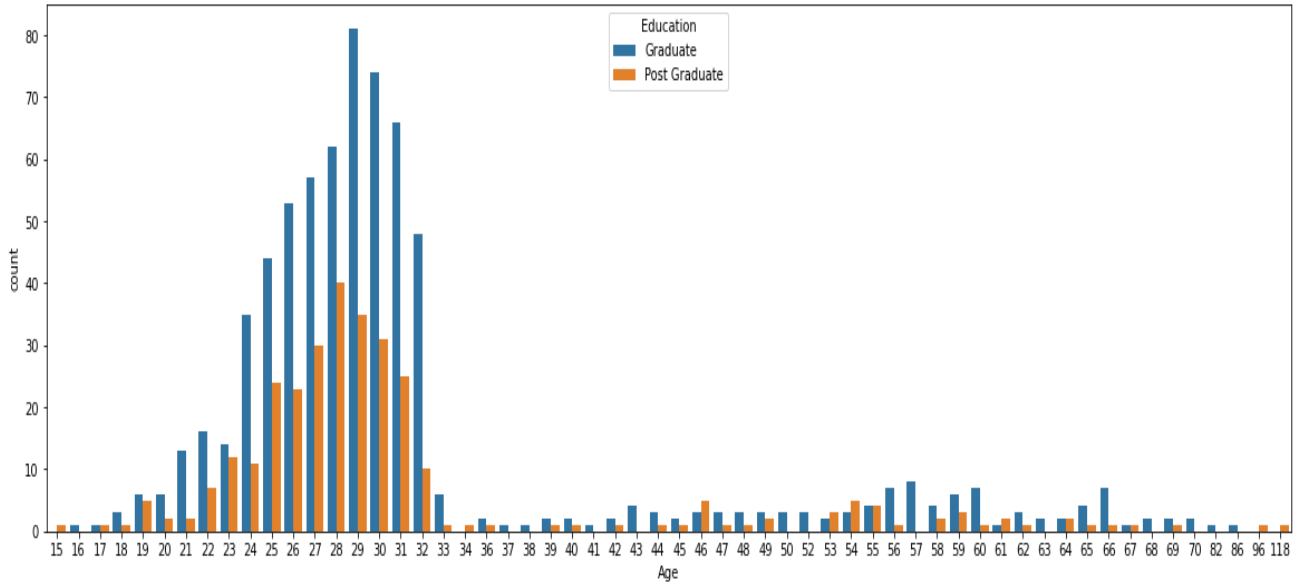
Visualization:

Age vs Annual Income (Scatter Plot)



This graph helps identify which age groups earn more annually. From your clustering, it's evident that individuals aged 28–31 have higher incomes, aligning with the EV target market.

Age vs Education (Countplot)

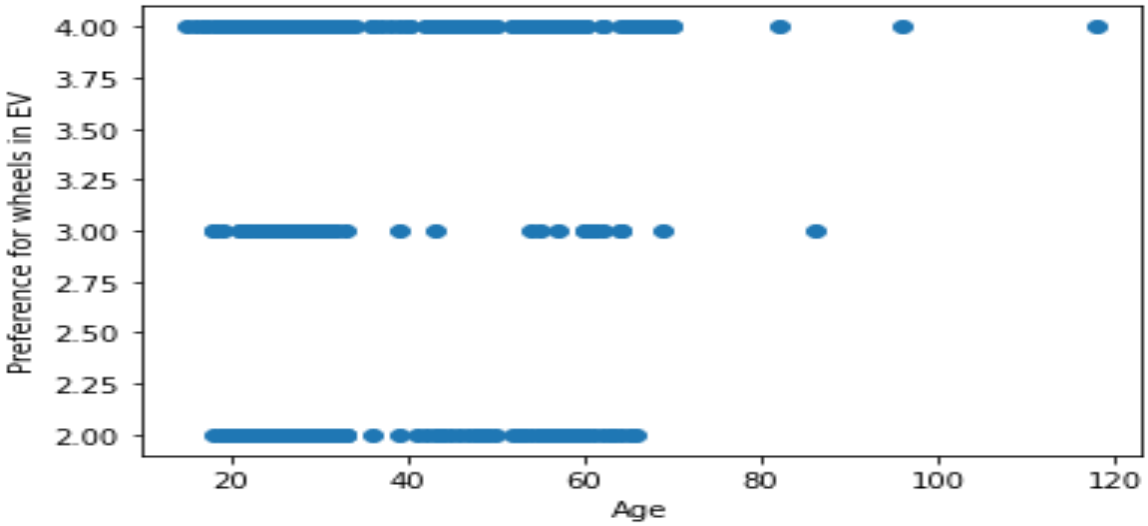


This reveals that educated individuals in their late 20s and early 30s are more common, supporting demographic targeting based on both age and education.

Age vs Preference for Wheels in EV (Scatter Plot)

Understand if age influences the preference between 2-wheeler vs. 4-wheeler EVs.

Typically, younger users (under 30) tend to prefer 2-wheelers, while older users lean toward 4-wheelers. This helps in producttype targeting within age groups.



**** To Whom EV Vehicles will sell ? ****

Electric vehicles (EVs) should be primarily targeted at individuals in the **age group of 28 to 31 years with an annual income ranging from ₹25 to ₹30 lakhs**. This demographic segment has shown the highest interest in adopting EVs, coupled with the financial capacity to invest in such technology. These individuals are typically well-educated, tech-aware, and environmentally conscious, making them more open to innovation and sustainability. Focusing on this age-income bracket allows for strategic marketing and product positioning that aligns with their preferences and spending ability, thereby increasing the likelihood of successful adoption and long-term engagement with EV brands.

GITHUB LINK: https://github.com/Narsi14/Vehicle-Segmentation

Market Segmentation Analysis Notes

Samiksha Kamble
GitHub Link Here

Step 1: Deciding (not) to Segment

- Market segmentation requires long-term strategic commitment and is not a short-term initiative.
- Segmentation incurs costs — surveys, focus groups, packaging, advertising, etc.
- It should only be pursued if the expected profit increase outweighs the costs.
- Strategic Changes Required - New or modified products, pricing and distribution changes, tailored communication strategies.
- Must be decided at the executive level and communicated across the organisation.

Step 2: Specifying the Ideal Target Segment

- Step 2 involves defining segment evaluation criteria to guide data collection and target segment selection.
- This step requires user (organizational) input, not just technical analysis.
- Two types of Segment Evaluation Criteria:
    - Knock-out and attractiveness criteria are distinct and serve different purposes.
        - Knock-out Criteria – Non-negotiable, must-have.
          Segments must:
            - Be homogeneous – members within a segment should be similar.
            - Be distinct – clearly different from other segments.
            - Be large enough – to justify a tailored marketing mix.
            - Match organizational strengths – company must be able to serve the segment.
            - Be identifiable – must be recognizable in the marketplace.

        - Be reachable – possible to communicate and deliver value to them.

    - Attractiveness Criteria – Used to compare and rank qualifying segments.
        - Used after knock-out filtering to evaluate and compare remaining segments.
        - Not binary – segments are rated based on how well they satisfy each criterion.
        - Selected and weighted based on organizational needs.
        - Examples : Market size, growth rate, profitability, competitive advantage, price sensitivity, buyer/supplier power, fit with company strengths and image, technological volatility, socio-political factors

- Implementing a Structured Process:
    - Use a Segment Evaluation Plot with Segment Attractiveness (x-axis) and Organizational Competitiveness (y-axis).
        - Choose no more than 6 criteria to keep the process practical.
        - Determine criteria before data collection to ensure relevant information is captured.
        - Assign weights to each attractiveness criterion:
            - Common method: team distributes 100 points among the criteria based on importance.
            - Final weights are negotiated and ideally approved by an advisory committee.
        - Include representatives from multiple organizational units to:
            - Gain diverse perspectives.
            - Ensure buy-in for implementation.

Step 3: Collecting Data

- Segmentation Variables

- Segmentation variable: Used to divide the market into segments (e.g., gender).
- Commonsense segmentation: Uses a single characteristic.
- Data-driven segmentation: Uses multiple variables (e.g., benefits sought).
- Descriptor variables: Describe segments in detail.
- High-quality empirical data is essential for accurate segmentation.
- Data sources: Surveys (most common), observational studies, experiments.

- Segmentation Criteria
  - Broader category like geographic, psychographic, behavioral. It refers to the type of information used for segmentation.
  - Choosing the right criterion requires market knowledge and cannot be outsourced entirely.
  - Geographic Segmentation
    - Based on location.
    - Pros: Easy to assign/target.
    - Cons: May not reflect preferences or values.
  - Socio-Demographic Segmentation
    - Includes age, gender, income, education.
    - Pros: Easy to identify.
    - Cons: Explains only ~5% of behavioral variance.
  - Psychographic Segmentation
    - Based on beliefs, values, motivations, lifestyles.
    - Pros: Better captures underlying behavior.
    - Cons: Hard to measure; requires valid data.

    - Behavioural Segmentation
        - Based on consumer behavior (e.g., purchase history).
        - Pros: Actionable and relevant.
        - Cons: Data may be unavailable for non-customers.

- Choice of Variables
    - Select only relevant variables.
    - If there are too many variables → fatigue, noise, poor segmentation.
    - Response Options - Binary and metric data preferred.
    - Response Styles - Biases (e.g., always agreeing) distort results. Minimize through design; remove biased responses if needed.
    - Sample Size
        - Formann: ≥ 2^p (binary); better: 5·2^p.
        - Qiu & Joe: ≥ 10·p·k.
        - Dolnicar et al.: ≥ 60·p (simple) or ≥ 70·p (complex).
    - Larger samples improve accuracy but have diminishing returns.

Step 3: Profiling Segments

- Profiling is essential in data-driven segmentation, not commonsense segmentation.
- Goal is to understand and describe each segment and compare them for strategic insights.
- Profiling process:
    - Describe segments individually.
    - Compare segments against each other and the overall sample.
- Challenges:

    - Users struggle to interpret segmentation results.
    - Oversimplified summaries → too trivial.
    - Detailed tables → too complex.

- Accurate profiling leads to better marketing decisions.
- Segment Profiling with Visualisations
    - Visualisations:
        - Simplify complex data and enhance interpretation.
        - Help detect marker variables (defining characteristics).
        - Improve communication to non-technical stakeholders.
- Segment Profile Plots:
    - Show how each segment differs from the average.
    - Variables can be reordered using hierarchical clustering for clarity.

Step 4: Describing Segments

- Segment profiling uses segmentation variables; segment describing uses extra variables (descriptor variables like age, gender, income) to better understand segments and tailor marketing strategies.
- Descriptor variables add depth to segmentation.
- Visual tools (like charts) make segment comparisons clearer and easier than tables. They help avoid reading too much into small differences.
- Visualizations (e.g., mosaic plots) highlight meaningful differences and patterns.
- Use cross-tabs and visual tools like stacked bar charts and mosaic plots to compare categorical descriptors (e.g., gender or income) across segments.
- Metric Descriptor Variables - Numerical descriptors (like age or moral values) help understand how segments differ in measurable ways.
- Conditional Histograms - Show distributions of numerical variables (like age) for each segment separately; useful for exploration but not always easy to compare.

- Parallel Box-and-Whisker Plots - Boxplots show how segments differ in metrics like age or moral values, and they highlight significant differences using confidence intervals and box widths.

- Modified SLSA Plot - Tracks how segments and their traits (like moral obligation) stay stable or change across different segmentation models.

- Boxplots and SLSA are better for comparison than histograms.

- Regression Models

    - Models like linear or logistic regression help predict segment based on descriptors.

    - Logistic regression is used when the outcome is categorical (e.g., which segment).

    - Binary Logistic Regression

        - Predicts whether someone is in a specific segment (yes/no) using variables like age or values.

        - Uses age and moral obligation to predict if someone is in a particular segment. Results show which traits matter most.

        - Coefficients tell how each variable affects the probability of being in the segment.

        - Plots show how age or moral values affect the likelihood of segment membership.

    - Multinomial Logistic Regression

        - Used when there are more than two segments. Predicts which one a person is likely to belong to.

        - Coefficients tell how descriptors (like age or values) change the odds of belonging to each segment.

        - Tests whether the descriptors as a group are useful predictors.

        - Can predict actual segment or probability of belonging to each one.

    - Tree-Based Methods

        - Decision trees predict outcomes by splitting data into groups step-by-step.

- Easy to intepret and handle complex data, but results can change a lot with small data changes.
- The method of repeatedly splitting data to make groups more similar internally.
- Trees differ in how they split, choose variables, stop growing, and make predictions.
- Each split shows how outcomes are distributed; deeper levels give more specific predictions.
- Multiclass Prediction: Trees can predict multiple segments (not just yes/no) at once.