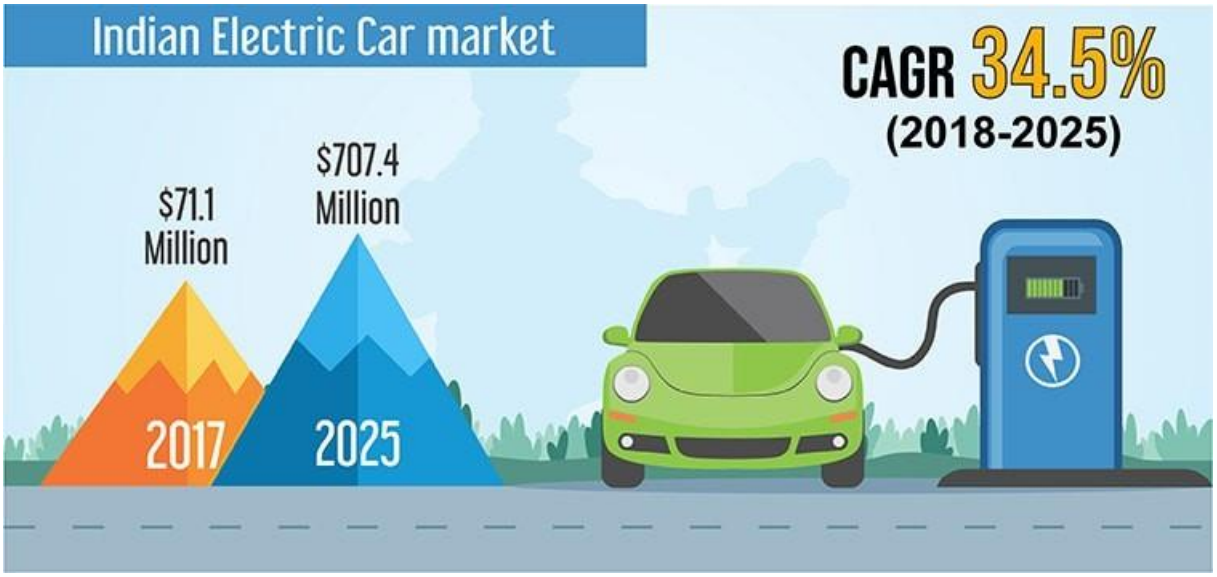
Electric Vehicle Market in India

Market Segmentation

(Sahil Devanand Khandait**)**



Electric Vehicle Market analysis in India

Electric cars are not going to take the market by storm, but it's going to be a gradual improvement.

(Carlos Ghosn)

Abstract

Market segmentation becomes a crucial tool for evolving transportation technology such as electric vehicles (EVs) in emerging markets to explore and implement for extensive adoption. EVs adoption is expected to grow phenomenally in near future as low emission and low operating cost vehicle, and thus, it drives a considerable amount of forthcoming academic research curiosity. The main aim of this study is to explore and identify distinct sets of potential buyer segments for EVs based on *psychographic, behavioral, and socio-economic* characterization by employing an integrated research framework of *'perceived benefits-attitude-intention'*. The study applied robust analytical procedures including cluster analysis, multiple discriminant analysis and Chi-square test to operationalize and validate segments from the data collected of 563 respondents using a cross-sectional online survey. The findings posit that the three distinct sets of young consumer groups have been identified and labelled as *'Conservatives', 'Indifferents', and 'Enthusiasts'* which are deemed to be buddying EV buyers The implications are recommended, which may offer some pertinent guidance for scholars and policy-makers to encourage EVs adoption in the backdrop of emerging sustainable transport market.

In this report we are going to analyse the data and solve the problem using **Fermi Estimation** by breaking down the problem.

KeyWords : *Electric vehicles, Market segmentation, Cluster analysis, Attitude towards electric vehicles, Subjective norms, Adoption intention, Sustainable transporta- tion.*

Data Collection

The data has been collected manually, and the sources used for this process are listed below :

- https://www.kaggle.com/datasets
- https://data.gov.in/
- https://www.data.gov/
- https://data.worldbank.org/
- https://datasetsearch.research.google.com/

Market Segmentation

Target Market:

The target market of Electric Vehicle Market Segmentation can be categorized into Geographic, SocioDemographic, Behavioral, and Psychographic Segmentation.

Behavioral Segmentation: searches directly for similarities in behavior or reported behavior.
Example: prior experience with the product, amount spent on the purchase, etc.

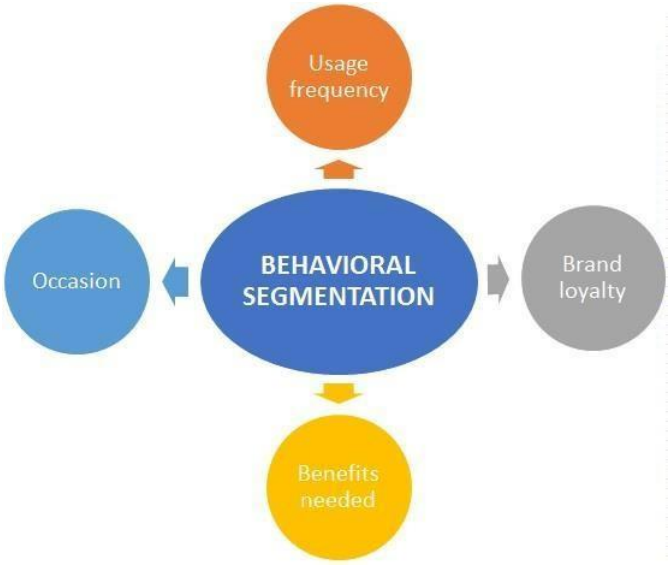


Figure 1: *Behavioral Segmentation*

Advantage: uses the very behavior of interest is used as the basis of segment extraction.

Disadvantage: not always readily available.

Psychographic Segmentation: grouped based on beliefs, interests, preferences, aspirations, or benefits sought when purchasing a product. Suitable for lifestyle segmentation. Involves many segmentation variables.

Advantage: generally more reflective of the underlying reasons for differences in consumer behavior.

Disadvantage: increased complexity of determining segment memberships for consumers.

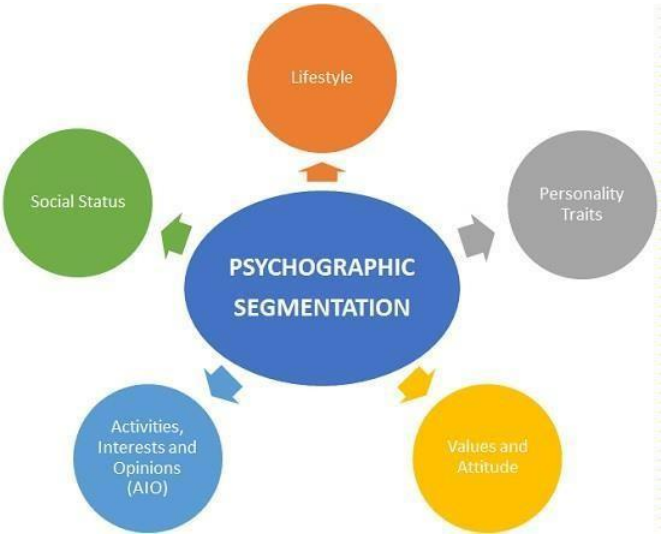


Figure 2: *Psychographic Segmentation*

Socio-Demographic Segmentation: includes age, gender, income and education. Useful in industries.

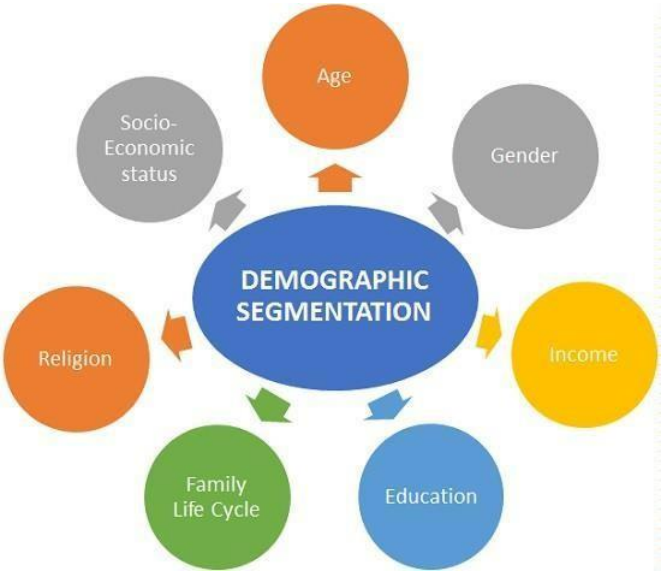


Figure 3: *Behavioral Segmentation*

Advantage: segment membership can easily be determined for every customer.

Disadvantage: if this criteria is not the cause for customers product preferences then it does not provide sufficient market insight for optimal segmentation decisions.

Segmenting for Electric Vehicle Market

The market segmentation approach aims at defining actionable, manageable, homogenous subgroups of individual customers to whom the marketers can target with a similar set of marketing strategies. In practice, there are two ways of segmenting the market-a-priori and post-hoc. An a-priori approach utilizes predefined characteristics such as age, gender, income, education, etc. to predefine the segments followed by pro- filing based on a host of measured variables *(behavioral, psychographic or benefit)*. In the post-hoc approach to segmentation on other hand, the segments are identified based on the relationship among the multiple measured variables. The commonality between both approaches lies in the fact that the measured variables determine the *'segmentation theme'*. The present study utilizes an a-priori approach to segmentation so as to divide the potential EV customers into sub-groups.

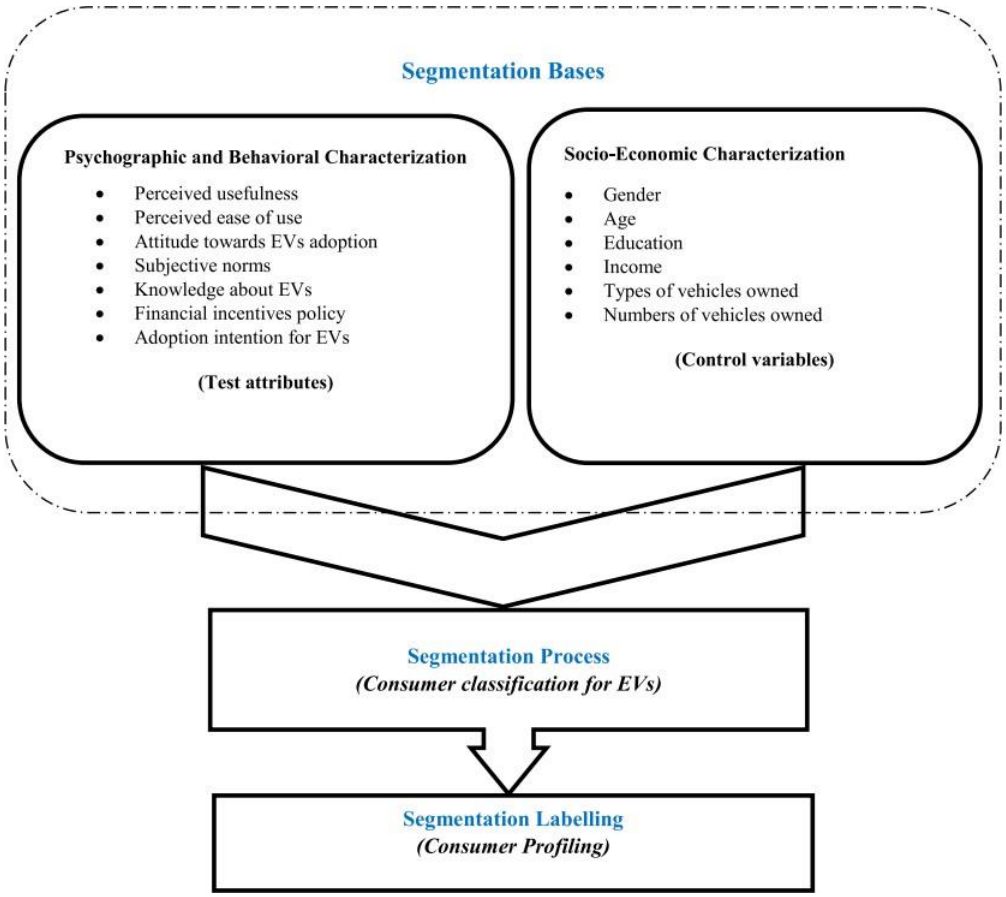


Figure 4: *Market Segmentation Electric Vehicles*

It is argued that the blended approach of *psychographic* and *socioeconomic attributes* for market segmentation enables the formulation of sub-market strategies which in turn satisfy the specific tastes and preferences of the consumer groups. Straughan and Roberts presented a comparison between the usefulness of *psychographic, demographic, and economic* characteristics based on consumer evaluation for eco-friendly products.

They pinpointed the perceived superiority of the psychographic characteristics over the socio-demographic and economic ones in explaining the environmentally-conscious consumer behavior and thus, the study recommended the use of psychographic characteristics in profiling the consumer segments in the market for eco-friendly products. The present study adds perceived-benefit characteristics guided by blended psychographic and socio-economic aspects for segmenting the consumer market.

Implementation on the Customer Review DataSet

Packages/Tools used:

1. **Numpy:** To calculate various calculations related to arrays.
2. **Pandas:** To read or load the datasets.
3. **SKLearn:** We have used LabelEncoder() to encode our values.
4. **Nltk toolkit** : We used it for the segemnatation of the Customer Review

Data-Preprocessing

Data Cleaning

The data collected is compact and is partly used for visualization purposes and partly for clustering. Python libraries such as NumPy, Pandas, Scikit-Learn, and SciPy are used for the workflow, and the results obtained are ensured to be reproducible.

```
import pandas as pd
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import nltk
```

```
[ ] df1 = pd.read_csv('2-wheeler-EV-bikewale.csv')
    df2 = pd.read_csv('4-wheeler-EV-cardekho.csv')
    df3 = pd.read_csv('4-wheeler-EV-carwale.csv')
```

```
df1.head(5)
```

	review	Used it for	Owned for	Ridden for	rating	Visual Appeal	Reliability	Performance	Service Experience	Extra Features	Comfort	Maintenance cost	Value for Money	Model Name
0	We all checked the bike's capacity to be 150 k...	Daily Commute	Never owned	NaN	1	3.0	4.0	NaN	NaN	NaN	4.0	NaN	1.0	TVS iQube
1	Performance is very poor on this bike. The cha...	Everything	> 1 yr	< 5000 kms	1	3.0	1.0	NaN	1.0	NaN	3.0	NaN	3.0	TVS iQube
2	I purchased this in April 2022 and the sales s...	Daily Commute	< 3 months	< 5000 kms	3	4.0	4.0	NaN	2.0	NaN	5.0	NaN	2.0	TVS iQube
3	If any issues come in scooty parts not availab...	Daily Commute	6 months-1 yr	5000-10000 kms	1	1.0	1.0	NaN	1.0	NaN	1.0	NaN	1.0	TVS iQube
4	Don't buy this vehicle unless you have a near ...	Daily Commute	6 months-1 yr	< 5000 kms	1	3.0	4.0	NaN	1.0	NaN	3.0	NaN	2.0	TVS iQube

EDA

We start the Exploratory Data Analysis with some data Analysis drawn from the data without Principal Component Analysis and with some Principal Component Analysis in the dataset obtained from the combination of all the data we have. PCA is a statis- tical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the Principal Components. The process helps in reducing dimensions of the data to make the process of classification/regression or any form of machine learning, cost-effective.

- **Remove Duplicates**: Ensure no duplicate reviews are present.
- **Handle Missing Values**: Identify and address any missing values in the dataset.
- **Text Preprocessing**: Clean the text data by:

1. Converting text to lowercase
2. Removing punctuation, numbers, and special characters
3. Removing stop words (common words that do not carry significant meaning)
4. Tokenization (splitting text into individual words)
5. Lemmatization/Stemming (reducing words to their base or root form)

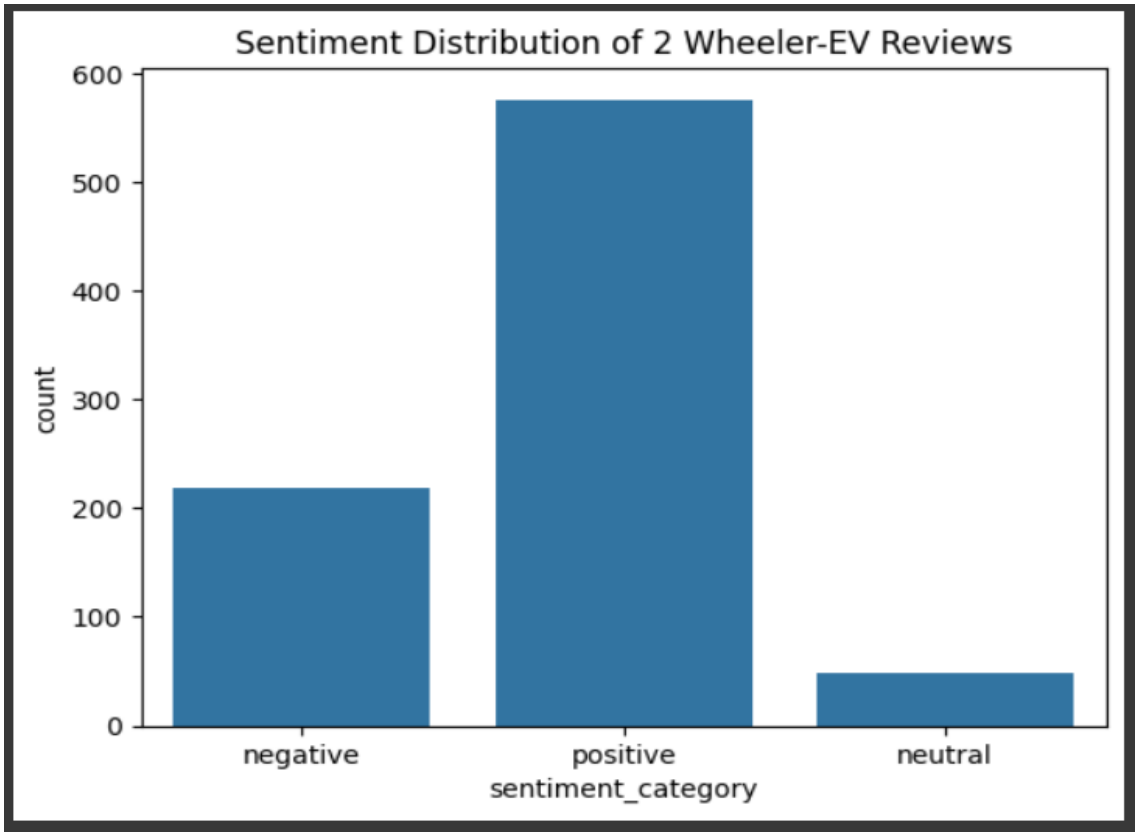
```python
def clean_text(text):
    # Handle potential non-string values
    if not isinstance(text, str):
        return ""  # Or any other suitable handling for non-string values

    # Remove HTML tags
    text = re.sub(r'<[^>]+>', '', text)

    # Remove special characters and numbers
    text = re.sub(r'[^a-zA-Z]', ' ', text)

    # Convert to lowercase and split into words
    words = text.lower().split()

    # Remove stop words and lemmatize
    words = [lemmatizer.lemmatize(word) for word in words if word not in stop_words]

    # Join the words back into a string
    return ' '.join(words)

# Apply the clean_text function to the 'Review' column of each DataFrame
df1['review'] = df1['review'].apply(clean_text)
df2['review'] = df2['review'].apply(clean_text)
df3['review'] = df3['review'].apply(clean_text)
```

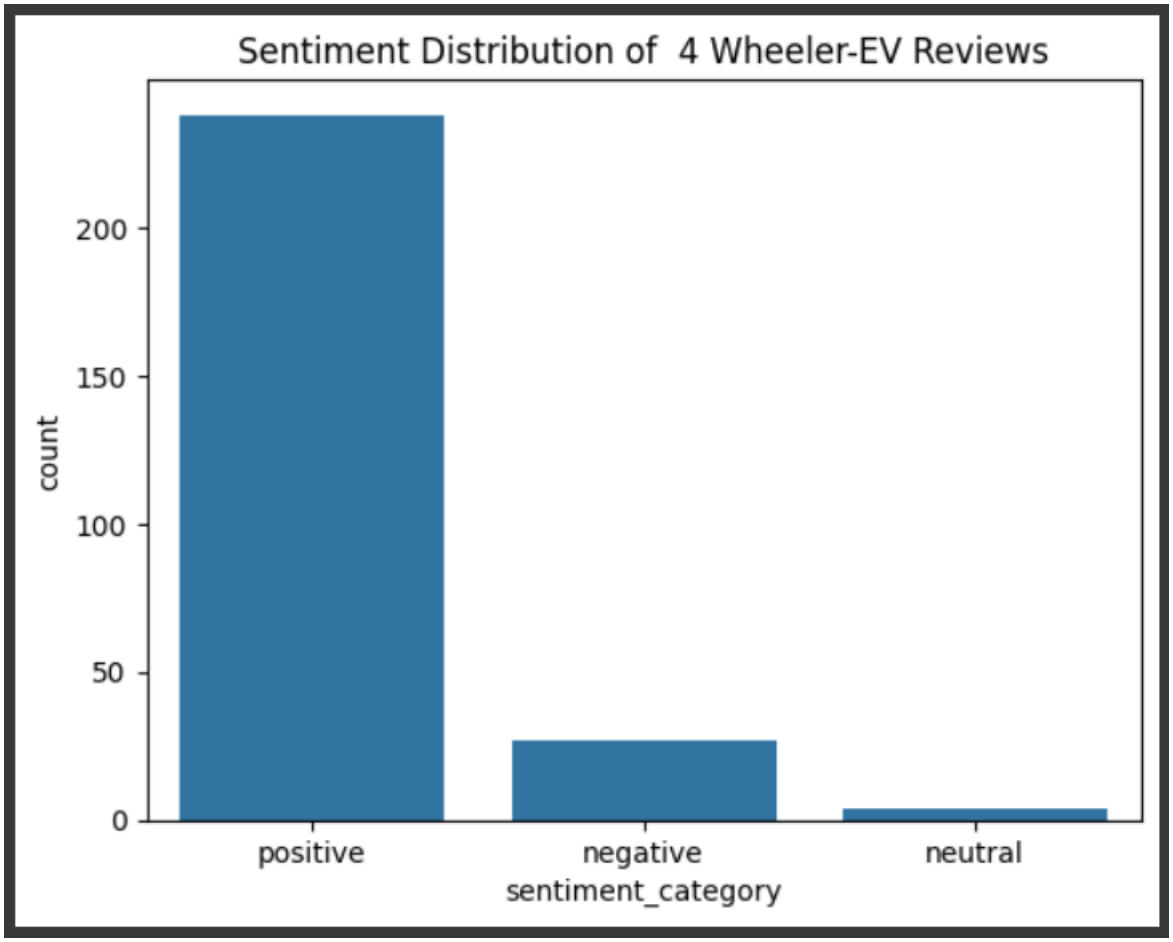
Code we used for the Eda of the dataset

```python
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import nltk

nltk.download('vader_lexicon')
sid = SentimentIntensityAnalyzer()

# Replace 'df1' with the correct DataFrame name if you intended to use one of those
df1['sentiment'] = df1['review'].apply(lambda x: sid.polarity_scores(x)['compound'])  # Use 'review' column if it contains the cleaned reviews
df1['sentiment_category'] = df1['sentiment'].apply(lambda x: 'positive' if x > 0 else ('negative' if x < 0 else 'neutral'))

# Repeat for df2 and df3 if needed
df2['sentiment'] = df2['review'].apply(lambda x: sid.polarity_scores(x)['compound'])  # Use 'review' column if it contains the cleaned reviews
df2['sentiment_category'] = df2['sentiment'].apply(lambda x: 'positive' if x > 0 else ('negative' if x < 0 else 'neutral'))

df3['sentiment'] = df3['review'].apply(lambda x: sid.polarity_scores(x)['compound'])  # Use 'review' column if it contains the cleaned reviews
df3['sentiment_category'] = df3['sentiment'].apply(lambda x: 'positive' if x > 0 else ('negative' if x < 0 else 'neutral'))
```

To see the distribution of the customer review we used the sentiment intensity analysis and we divide the customer review into the 'POSITIVE', 'NEGATIVE', 'NEUTRAL'.

Sentiment Distribution in our data

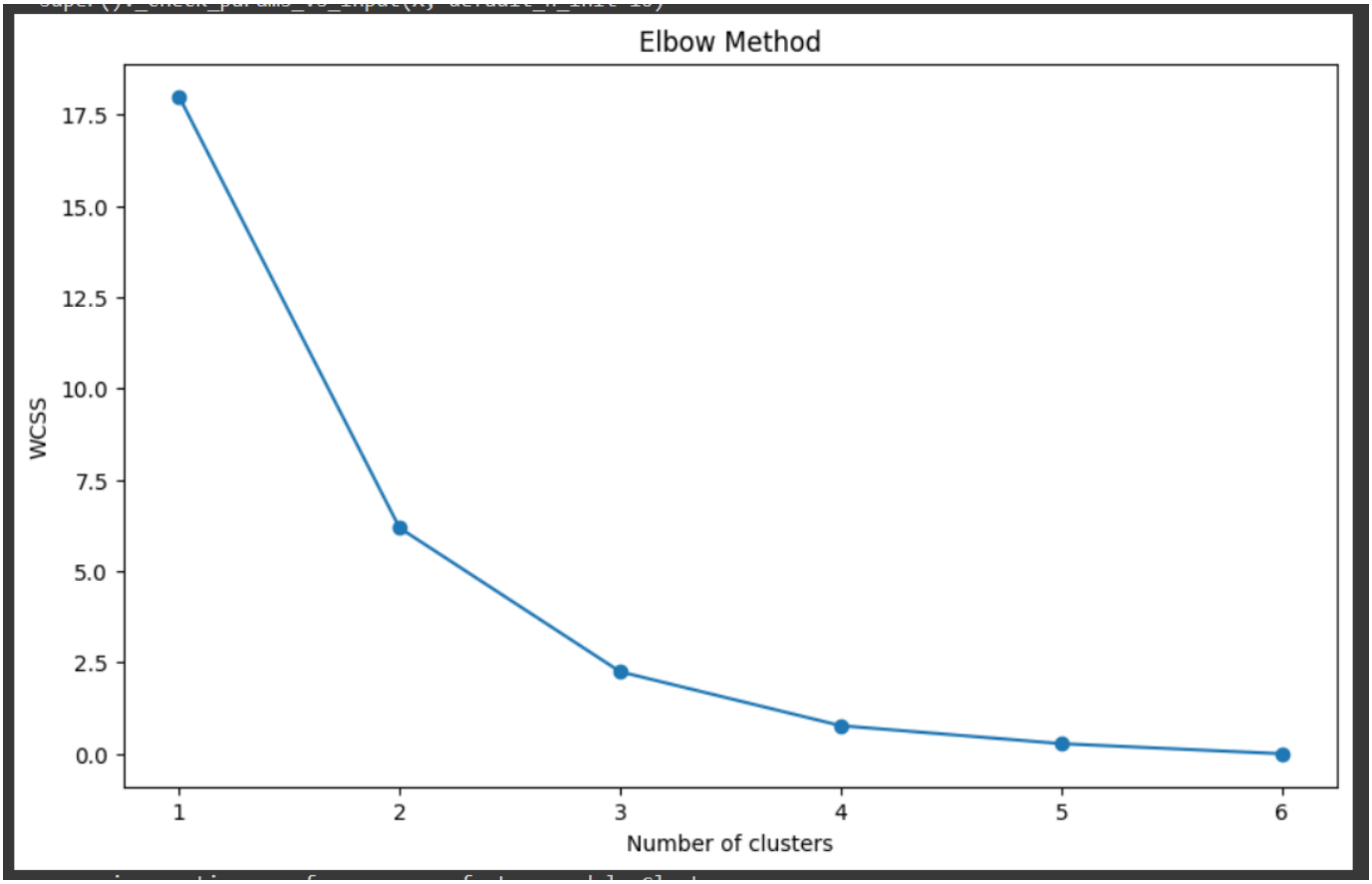




Elbow Method

The Elbow method is a popular method for determining the optimal number of clusters. The method is based on calculating the Within-Cluster-Sum of Squared Errors (WSS) for a different number of clusters (k) and selecting the k for which change in WSS first starts to diminish. The idea behind the elbow method is that the explained variation changes rapidly for a small number of clusters and then it slows down leading to an elbow formation in the curve. The elbow point is the number of clusters we can use for our clustering algorithm.

The KElbowVisualizer function fits the KMeans model for a range of clusters values between 2 to 8. As shown in Figure, the elbow point is achieved which is highlighted by the function itself. The function also informs us about how much time was needed to plot models for various numbers of clusters through the green line.


Elbow Method
WCSS
Number of clusters
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
1
2
3
4
5
6

Analysis and Approaches used for Segmentation

Clustering

Clustering is one of the most common exploratory data analysis techniques used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean- based distance or correlation-based distance.
The decision of which similarity measure to use is application-specific. Clustering analysis can be done on the basis of features where we try to find subgroups of samples based on features or on the basis of samples where we try to find subgroups of features based on samples.

K-Means Algorithm

K Means algorithm is an iterative algorithm that tries to partition the dataset into pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k means algorithm works is as follows:

- Specify number of clusters K.
- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

The approach k-means follows to solve the problem is **expectation maximization** The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. Below is a break down of how we can solve it mathematically,

The objective function is:

(1) $$\boxed{J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} ||x^i - \mu_k||}$$

And M-step is :

$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}}$$

Applications

K means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc. The goal usually when we undergo a cluster analysis is either:

1. Get a meaningful intuition of the structure of the data we're dealing with.
2. Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviors of different subgroups.

The **k-means clustering algorithm** performs the following tasks:

- Specify number of clusters K
- Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
- Compute the sum of the squared distance between data points and all centroids.
- Assign each data point to the closest cluster (centroid).
- Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.
- Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

According to the Elbow method, here we take K=3 clusters to train KMeans model.

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

# Determine the optimal number of clusters using the elbow method
wcss = []
for i in range(1, min(11, X.shape[0] + 1)):
    kmeans = KMeans(n_clusters=i, random_state=42)
    kmeans.fit(X)
    wcss.append(kmeans.inertia_)

# Plot the elbow method
plt.figure(figsize=(10, 6))
plt.plot(range(1, min(11, X.shape[0] + 1)), wcss, marker='o')  # Adjust the x-axis range accordingly
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()

# Apply K-Means with the chosen number of clusters (e.g., 3)
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(X)

# Add cluster labels to the merged dataframe
merged_df['Cluster'] = clusters

# Verify the dataframe with cluster labels
print(merged_df)
```

Implementation on the Customer Review DataSet

Packages/Tools used:

1. **Numpy:** To calculate various calculations related to arrays.
2. **Pandas:** To read or load the datasets.
3. **SKLearn:** We have used LabelEncoder() to encode our values.
4. **MatplotLib :** We used it for the plotting of the map.
5. **Seaborn :** We also used it for the visvalization and the plotting.

Data-Preprocessing

Data Cleaning

The data collected is compact and is partly used for visualization purposes and partly for clustering. Python libraries such as NumPy, Pandas, Scikit-Learn, and SciPy are used for the workflow, and the results obtained are ensured to be reproducible.

```
import pandas as pd
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import sklearn
from scipy import stats
import os
import seaborn as sns

# Example dataset
data1 = pd.read_csv('EV_India.csv')
data2 = pd.read_csv('ev station in india.csv')
data3 = pd.read_csv('ev-charging-stations-india.csv')

# Display the dataset
data1.head(5)
```

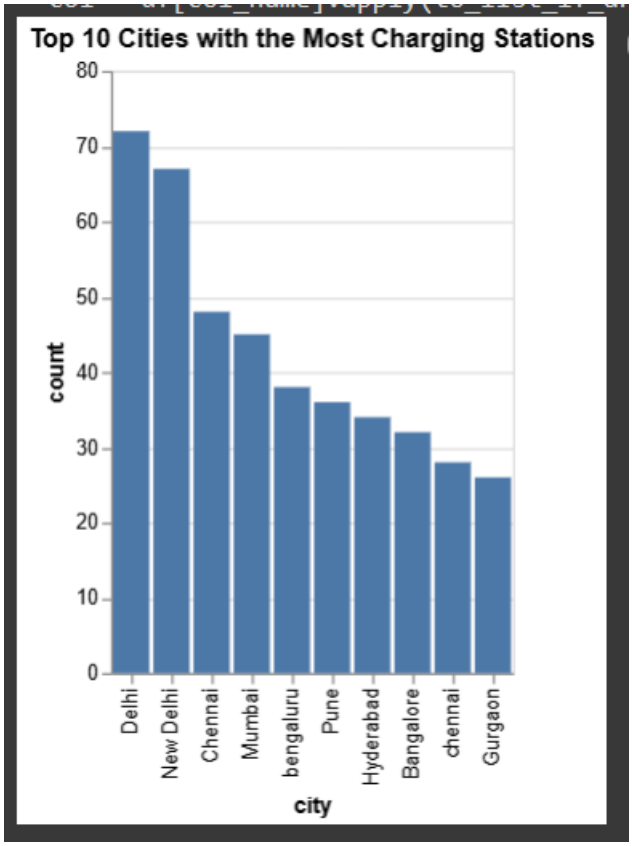
	Sr. No.	State Name	Total Electric Vehicle	Total Non-Electric Vehicle	Total
0	1	Andaman & Nicobar Island	162	1,46,945	1,47,107
1	2	Andra Pradesh	NaN	NaN	NaN
2	3	Arunachal Pradesh	20	2,52,965	2,52,985
3	4	Assam	64766	46,77,053	47,41,819
4	5	Bihar	83335	1,04,07,078	1,04,90,413

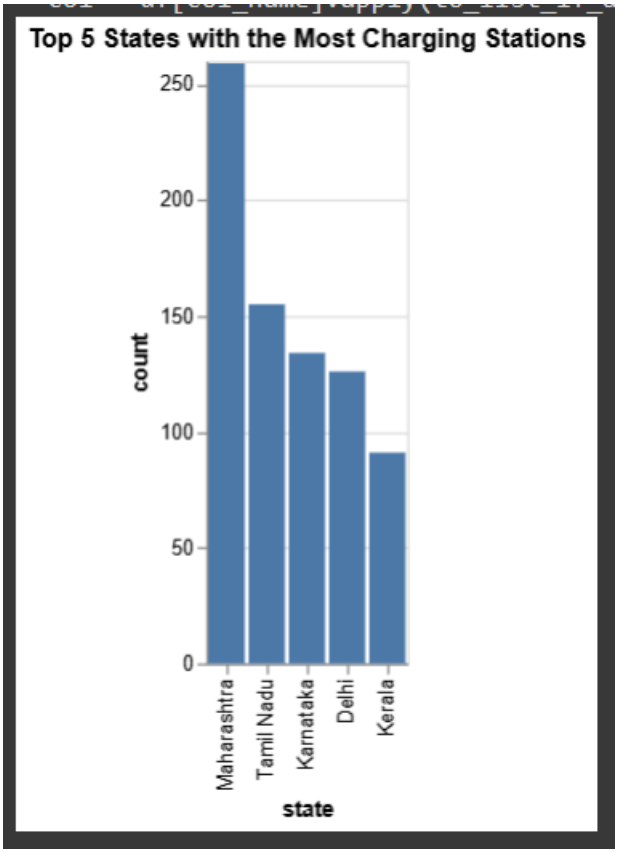
EDA

We start the Exploratory Data Analysis with some data Analysis drawn from the data without Principal Component Analysis and with some Principal Component Analysis in the dataset obtained from the combination of all the data we have. PCA is a statis- tical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the Principal Components. The process helps in reducing dimensions of the data to make the process of classification/regression or any form of machine learning, cost-effective.

- **Remove Unnecessary data**: Ensure no duplicate data are present.
- **Handle Missing Values**: Identify and address any missing values in the dataset.

And here are charts that we got from dataset :


Top 10 Cities with the Most Charging Stations
count
0
10
20
30
40
50
60
70
80
Delhi
New Delhi
Chennai
Mumbai
bengaluru
Pune
Hyderabad
Bangalore
chennai
Gurgaon
city


Top 5 States with the Most Charging Stations
count
0
50
100
150
200
250
Maharashtra
Tamil Nadu
Karnataka
Delhi
Kerala
state

Target Segments:

So from the analysis we can see that the optimum targeted segment should be belonging to the following categories:

1. Geographic Segmentation

- State
- city
- lattitude
- longitude
- address
- State/UT

2. Demographic Segmentation

- Owned for: This can give an indication of how long different age groups or demographics keep their vehicles, though it's more indicative of behavior.
- Model/model_name: Different demographics might prefer different models.

3. Behavioral Segmentation

- Used it for: This describes how the vehicle is used (e.g., commuting, leisure), which is directly tied to user behavior.
- Owned for/Ridden for: These columns describe the duration of ownership and usage, which are key behavioral indicators.
- Performance: User perceptions of performance can indicate behavior-related preferences and priorities.
- Comfort: Similar to performance, comfort is a behavioral preference.

4. Psychographic Segmentation

- Review: Reviews often contain psychographic data, reflecting users' attitudes, values, and lifestyle choices.
- Comfort: While also behavioral, comfort preferences can reflect underlying lifestyle and values.
- Performance: User perceptions of performance can also reflect lifestyle and values.

5. Socio-Demographic Segmentation

- Rating: Ratings might correlate with socio-demographic factors, reflecting satisfaction levels across different socio-demographic groups.
- Model/model_name: Different socio-demographic groups may prefer different models, possibly due to income level, status, etc.

References

[1] Deepak Jaiswal,Arun Kumar Deshmukh(2022) *Who will adopt electric vehicles? Segmenting and exemplifying potential buyer heterogeneity and forthcoming research*, Journal of Retailing and Consumer Services .

[2] Dolnicar, S., Grun Bettina, amp; Leisch, F. (2019). *Market segmentation analysis understanding it, doing it and making it useful*. Springer Nature.

[3] McDonald, M., amp; Dunbar, I. (2003). /emphMarket segmentation. Butterworth-Heinemann.