Feynn Labs : Project 3

Link : https://github.com/kdb-source/EV-Market-Segmentation/blob/4664287eab2cfcb099900a3f3ae939cdc248024f/EVMarket_India_KAnalysis.ipynb

Market Segmentation : Electric Vehicle in India

Contributor: Kalyani Bhosale

Date : 2nd August 2024

1. Problem Statement :

The main aim of this study is to explore and identify distinct sets of potential buyer segments for EVs in India based on *Geographic* , *Demographic* , *psychographic, and behavioural* characterization by employing an integrated research framework. 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 .

In this report we analyse the Electric Vehicles Market in India using segments such as state wise sales, models, charging facility, type of vehicles (e.g., 2 wheelers, 3 wheelers, 4 wheelers etc.), Price, manufacturers, body type (e.g., Hatchback, Sedan, SUV, Autorickshaw etc.), range, plug types and much more. 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 transportation.*

2. 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. In the post-hoc approach to segmentation, the segments are identified based on the relationship among the multiple measured variables.

2.1 ML Model used :

- Supervised Machine Learning Algorithm used
  Linear regression : To identify relationships between the variable of interest and the inputs, and predict its values based on the values of the input variables.

- Unsupervised Machine Learning Algorithm used :
  1. K-Means Clustering
  2. Hierarchical Clustering

2.2 Market Segmentation Target Market:

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

- **Geographic Segmentation :it i**s a marketing strategy where target market derived based on geographical location. e.g. States, Region, City

- **Behavioural Segmentation:** searches directly for similarities in behaviour or reported behaviour. e.g. prior experience with the product, amount spent on the purchase, etc.
- **Psychographic Segmentation:** grouped based on beliefs, interests, preferences, aspirations, or benefits sought when purchasing a product. Suitable for lifestyle segmentation.
- **Demographic Segmentation:** includes age, gender, income and education. Useful in industries.

2 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/

4. Data Preprocessing

- **Data Cleaning**
  The data collected is compact and is partly used for visualization purposes and partly for clustering.

- **Implementation Packages/Tools used:**

  Python libraries such as NumPy, Pandas, Scikit-Learn, and SciPy are used for the workflow, and the results obtained are ensured to be reproducible.

  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.

- **Steps taken to preprocess the scraped raw data:**
  1. Ordinal encoded 'PowerTrain'
  2. Label encoded 'RapidCharge'
  3. Used Label Encoder and Standard Scaler package for preprocessing of the dataset.

- **Columns explanations:**

  1. 'Brand' and tells the manufacturers of electric vehicles.
  2. 'Model' tells the various of electric vehicles.
  3. 'AccelSec', 'Top Speed', 'PowerTrain' tells specification about the vehicles.
  4. 'Range_km', 'FastCharge_KmH', 'PlugType' and 'BodyStyle' tells us about range of vehicle per full charge, Fast charging is provided or not, type of charging plug and body style of vehicle respectively.

5. ‘Seats’ and ‘Price’ tells about the number of seats available on vehicle and their price.
6. State Name’ tells about the states of India.
7. ‘Two Wheeler’, ‘Three Wheeler’, ‘Four Wheeler’ and ‘Other’ tells about the type of vehicles in the market.

4.2 Exploratory Data Analysis

An Exploratory Data Analysis or EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for a company because it exposes trends, patterns, and relationships that are not readily apparent.

We analysed our dataset using univariate (analyse data over a single variable/column from a dataset), bivariate (analyse data by taking two variables/columns into consideration from a dataset) and multivariate (analyse data by taking more than two variables/columns into consideration from a dataset) analysis.

Data Analysis started with some 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 statistical 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.

4.2.1 Comparison of cars in our data

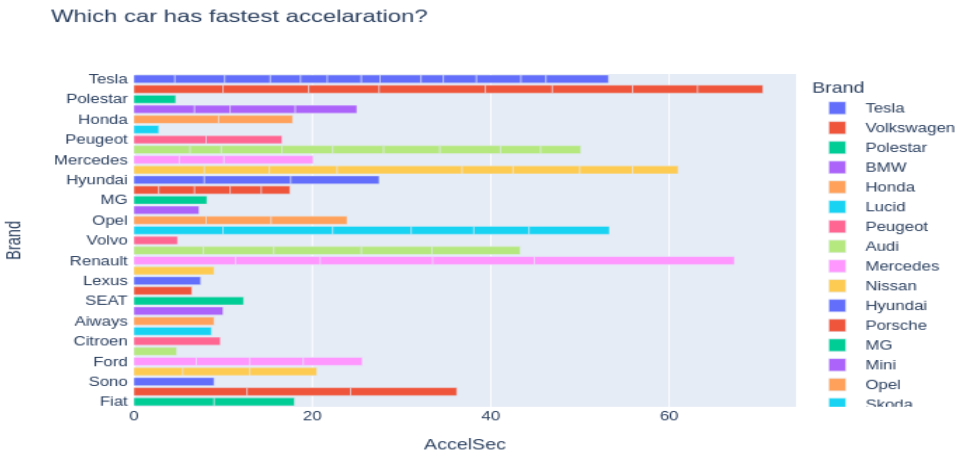


Figure 1 : Fastest Acceleration

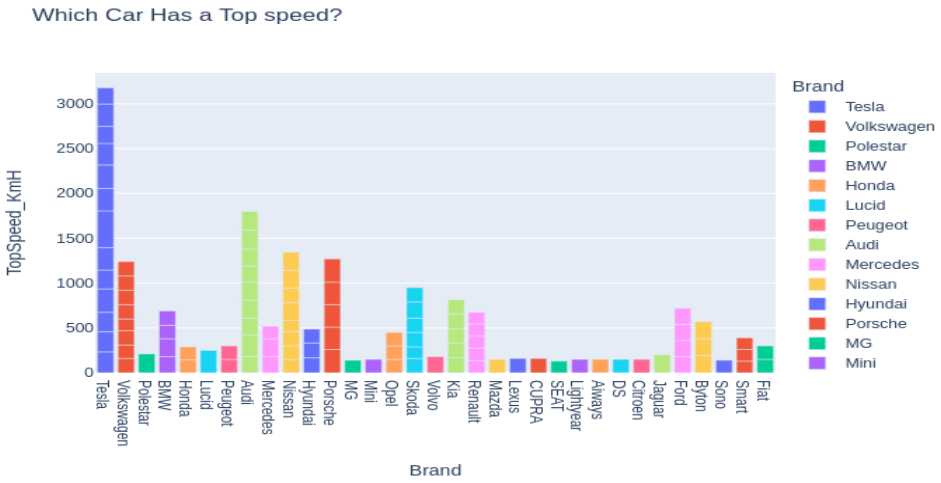


Figure 2 : Top speed

- Line plot representing Car Price range and variations

Figure 3 : Car Price

- **For Electric Vehicle Market one of the most important key is Charging:**

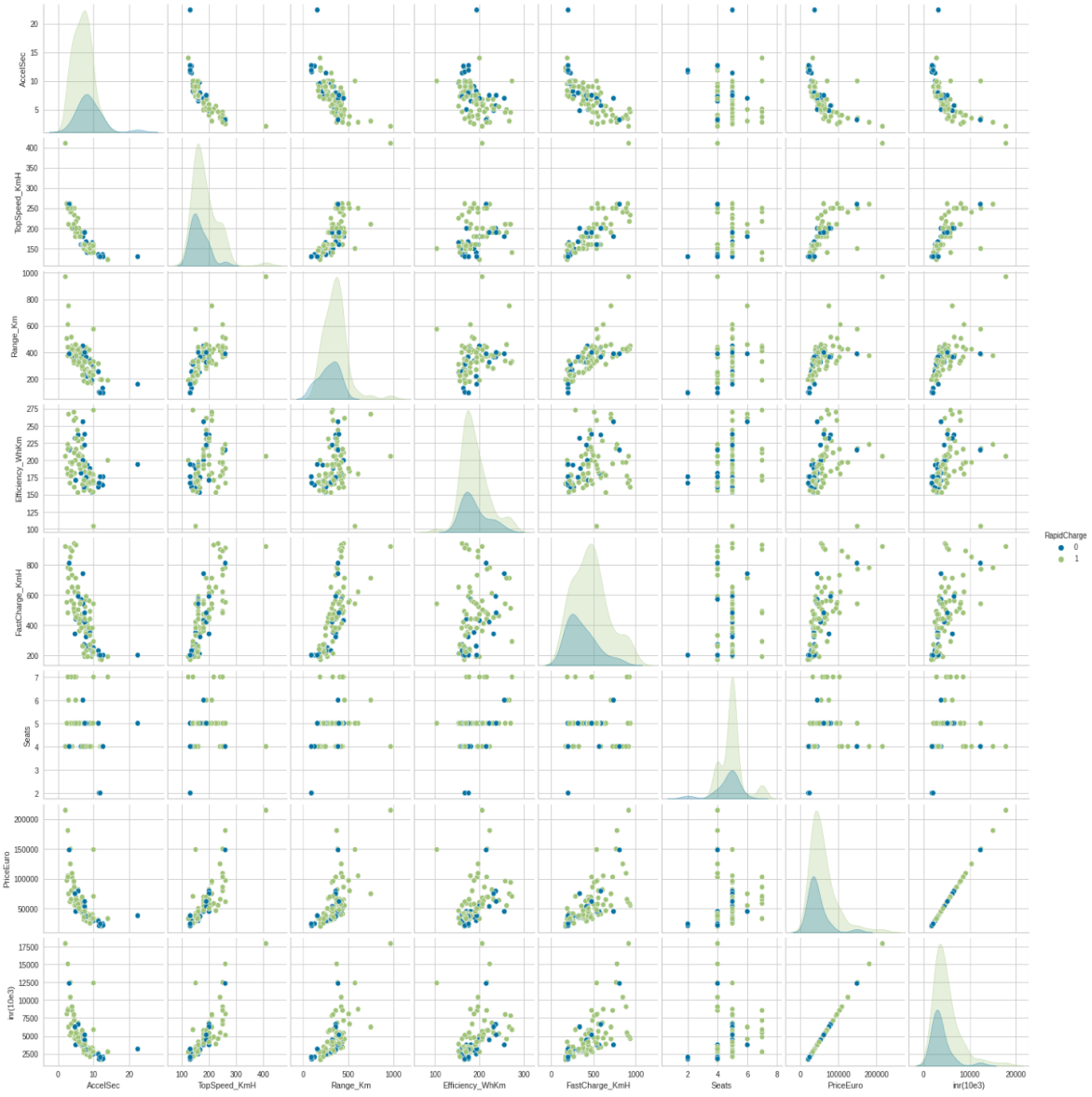


Figure 4 : Charging

4.2.2 Correlation Matrix:

A correlation matrix is simply a table that displays the correlation. It is best used in variables that demonstrate a linear relationship between each other. Coefficients for different variables. The matrix depicts the correlation be- tween all the possible pairs of values through the heatmap in the below figure. The relationship between two variables is usually considered strong when their correlation coefficient value is larger than 0.7.

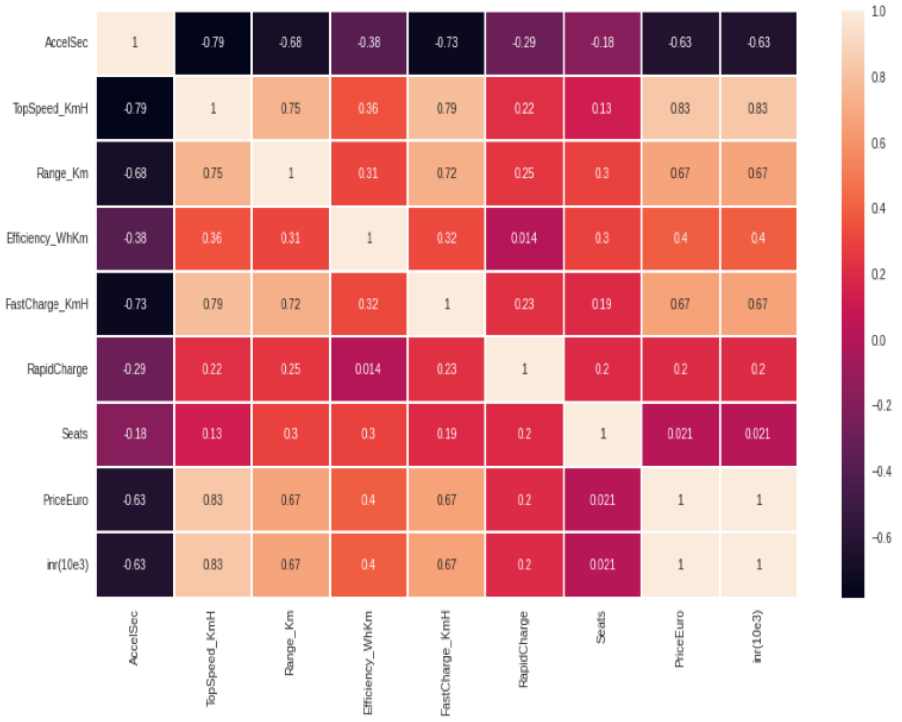
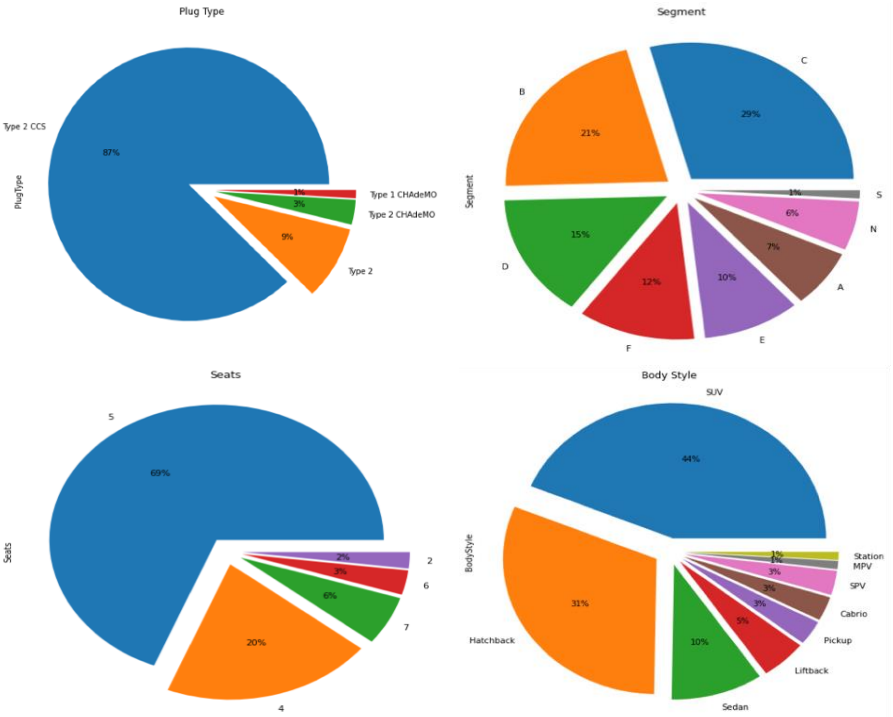


Figure 5 : *Correlation Matrix for the dataset*

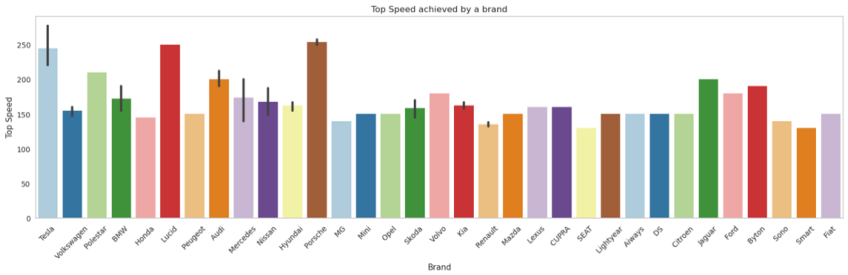
Electric Vehicle Range have strong positive correlation with Top Speed per Km driving. It might be sufficient to predict Electric Vehicle Range and then calculate range in Top Speed per Km.

Electric Vehicle Range have a strong negative correlation with Acceleration per second.

4.3Extracting Segments : [Plug type, Seats, Body style]

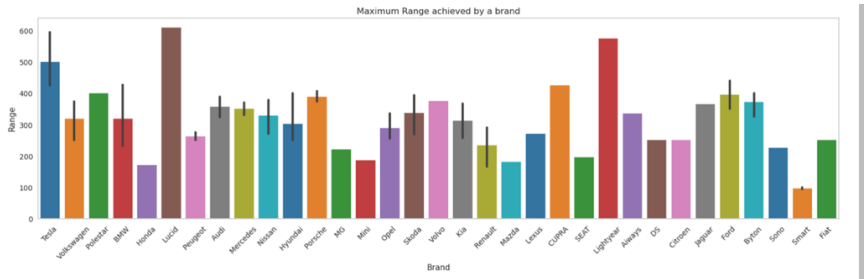


- **Top speed achieved by Car**



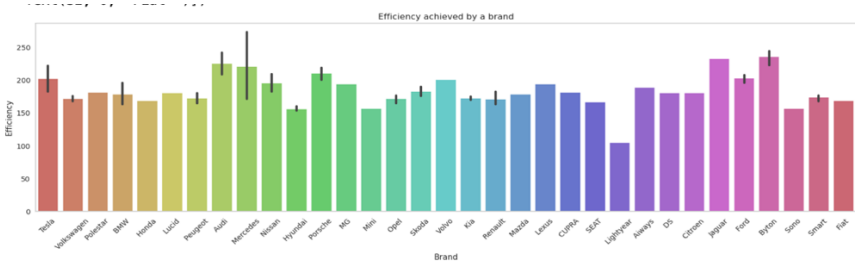
Porsche, Lucid and Tesla produce the fastest cars and Smart the lowest

- **Range achieved by Car**



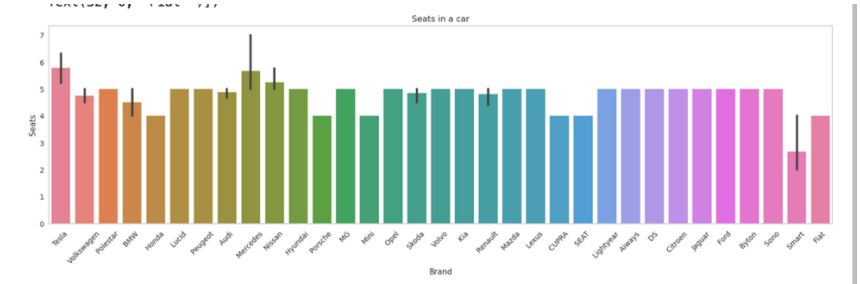
Lucid, Lightyear and Tesla have the highest range and Smart the lowest

- **Car efficiency**



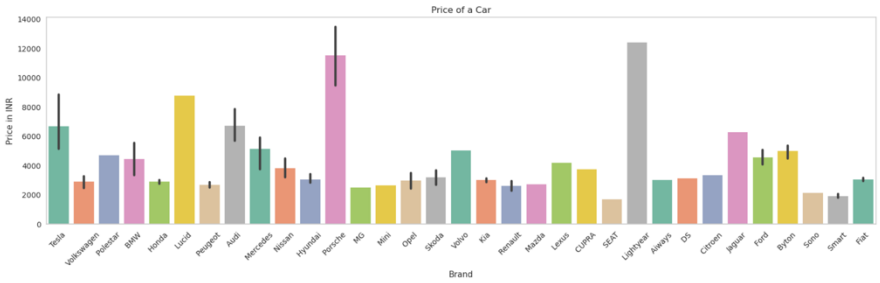
Byton , Jaguar and Audi are the most efficient and Lightyear the least

- **No of seats in each Car**



Mercedes, Tesla and Nissan have the highest number of seats and Smart the lowest

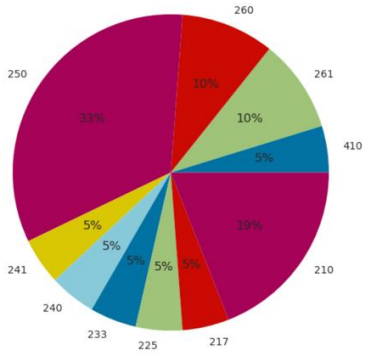
- **Price of cars (in INR)**

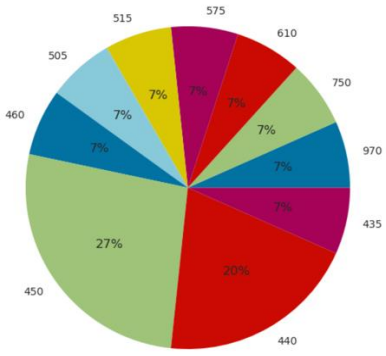


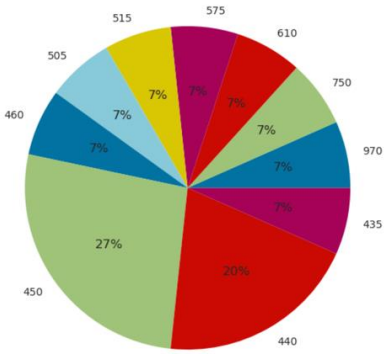
4.4 Profiling and Describing the Segments

Sorting the Top Speeds and Maximum Range in accordance to the Price with head () we can view the Pie Chart.

Pie chart : 1 Pie chart : 2. Pie chart : 3







Cost based on top speed max range **Cost based on max range** **Top speed based on**

4.5 Analysis and Approaches used for Segmentation

4.4.1 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.

- **Dendrogram**

This technique is specific to the agglomerative hierarchical method of clustering. The agglomerative hierarchical method of clustering starts by considering each point as a separate cluster and starts joining points to clusters in a hierarchical fashion based on their distances. To get the optimal number of clusters for hierarchical clustering, we make use of a dendrogram which is a tree-like chart that shows the sequences of merges or splits of clusters. If two clusters are merged, the dendrogram will join them in a graph and the height of the join will be the distance between those clusters. As shown in Figure, we can chose the optimal number of clusters based on hierarchical structure of the dendrogram. As highlighted by other cluster validation metrics, four to five clusters can be considered for the agglomerative hierarchical as well.

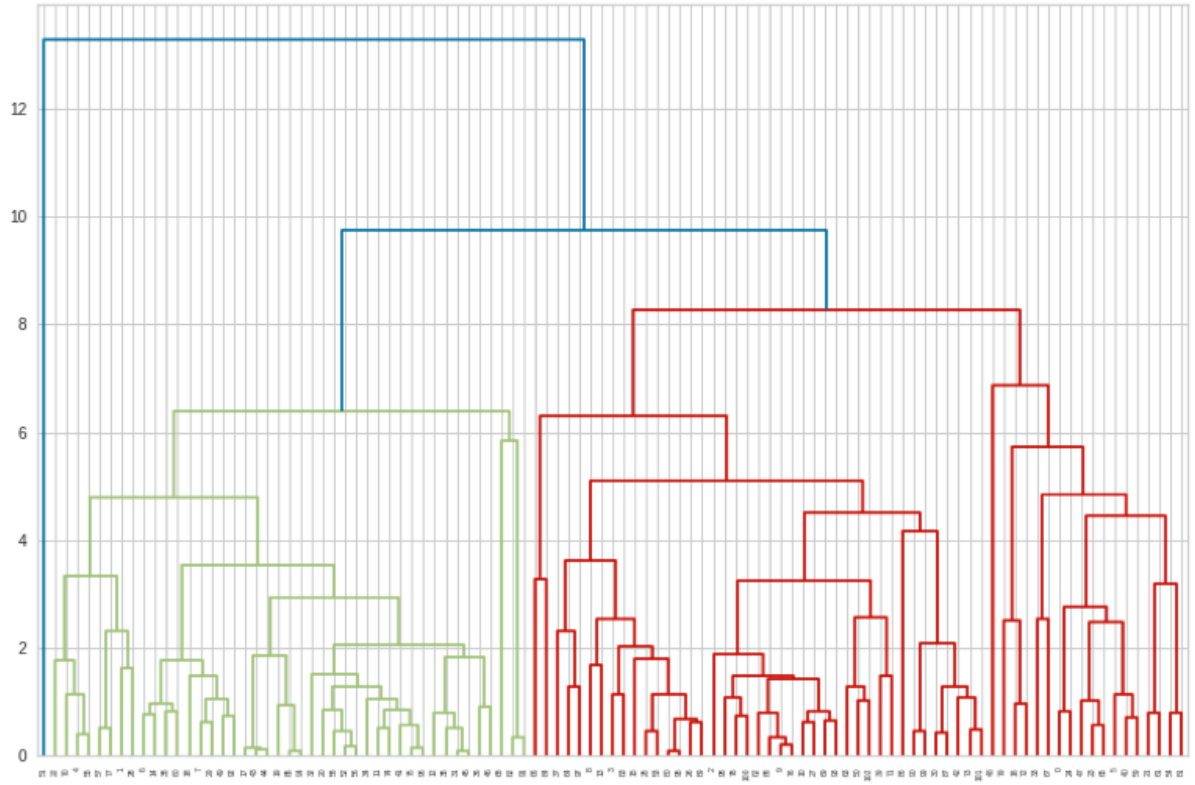


Figure 6: *Dendrogram Plot for our Dataset*

- **Elbow Method**

The Elbow method is a popular method for determining the optimal number of clusters. I have actually varying the number of clusters (K) from 1 – 10. For each value of K, I am calculating WCSS (Within-Cluster-Sum of Squared Errors) and selecting the k for which change in WCSS first starts to diminish. WCSS is the sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow.

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. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1 and thus creating an elbow shape. The elbow point is the number of clusters we can use for our clustering algorithm. From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters.

```
wcss = []

for i in range(1, 11):
  kmean = KMeans(n_clusters=i, init='k-means++', random_state=90)
  kmean.fit(X_pca)
  wcss.append(kmean.inertia_)
```

```
# plotting the results of Elbow

plt.figure(figsize=(4,3))
plt.title('Plot of the Elbow Method', size=15, family='serif')
plt.plot(range(1, 11), wcss)
plt.xticks(range(1, 11), family='serif')
```

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.

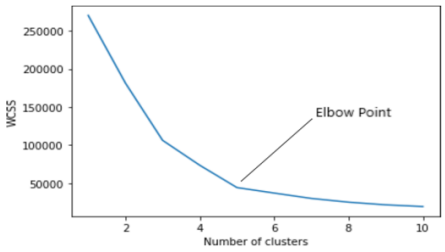


Fig.1: *Elbow Curve*

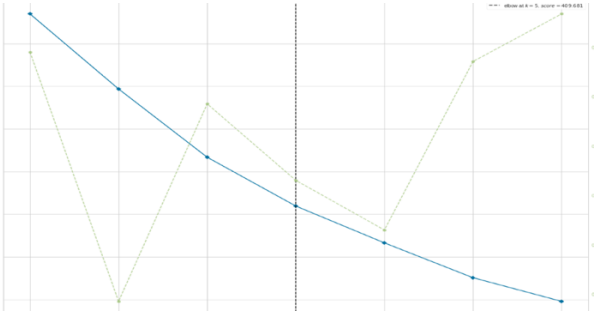


Fig.2: *Evaluating the clusters using Distortion*

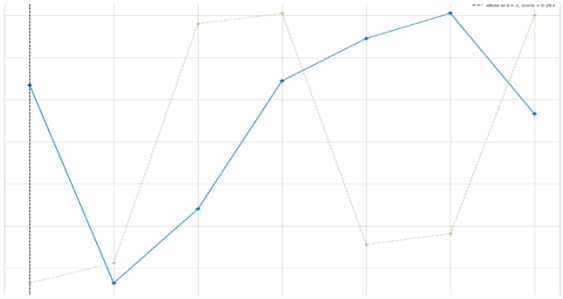


Fig.3: *Evaluating the clusters using silhouette*.

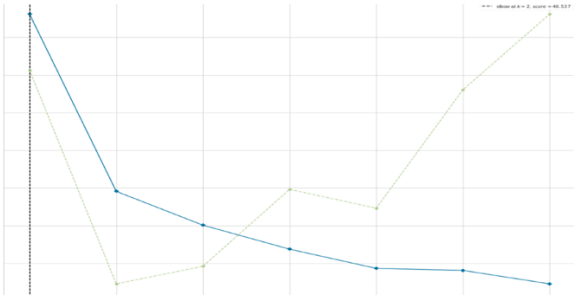


Fig.4: *Evaluating the clusters using calinski_harabasz*

K-Means Algorithm

K Means algorithm is an iterative algorithm that tries to partition the dataset into pre-defined distinct non-overlapping 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 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 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 breakdown of how we can solve it mathematically,

  The objective function is:

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

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:

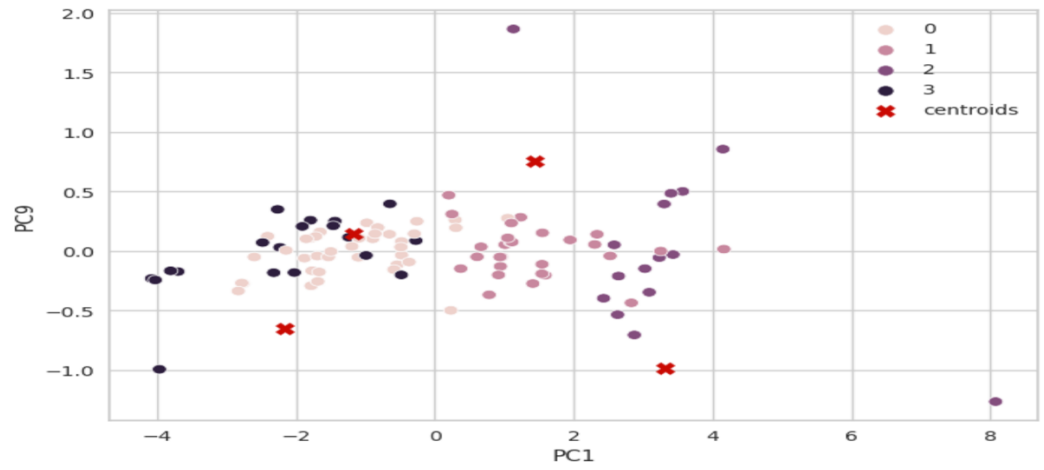
1. Get a meaningful intuition of the structure of the data we're dealing with.
2. Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviours of different subgroups.

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 datapoints that belong to each cluster.
- Keep iterating until there is no change to the centroids. i.e assignment of datapoints to clusters isn't changing.

- According to the Elbow method, here we take K=4 clusters to train KMeans model. The derived clusters are shown in the following figure

```
#Visulazing clusters
sb.scatterplot(data=df_pca, x="PC1", y="PC9", hue=kmeans.labels_)
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1],
            marker="X", c="r", s=80, label="centroids")
plt.legend()
plt.show()
```



- **Prediction of Prices most used cars**

Linear regression is a machine learning algorithm based on supervised learning to perform a regression task. Regression models targets prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Here we use a linear regression model to predict the prices of different Electric cars in different companies. X contains the independent variables and y is the dependent Prices that is to be predicted. We train our model with a splitting of data into a 4:6 ratio, i.e. 40% of the data is used to train the model.

LinearRegression().fit(X_{train,y_train}) command is used to fit the data set into model. The values of intercept, coefficient, and cumulative distribution function (CDF) are described in the figure.

```
X=df_pca[['PC1', 'PC2','PC3','PC4','PC5','PC6', 'PC7','PC8','PC9']]
y=df['inr(10e3)']
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.4, random_state=101)
lm=LinearRegression().fit(X_train,y_train)
```

```
print(lm.intercept_)
```

```
4643.522050485438
```

```
lm.coef_
```

```
array([ 1101.58721,  -741.20904,   208.53617,   508.32246,   122.3533 ,
        1579.00686,   333.61147, -1079.99512,  1461.72269])
```

```
X_train.columns
```

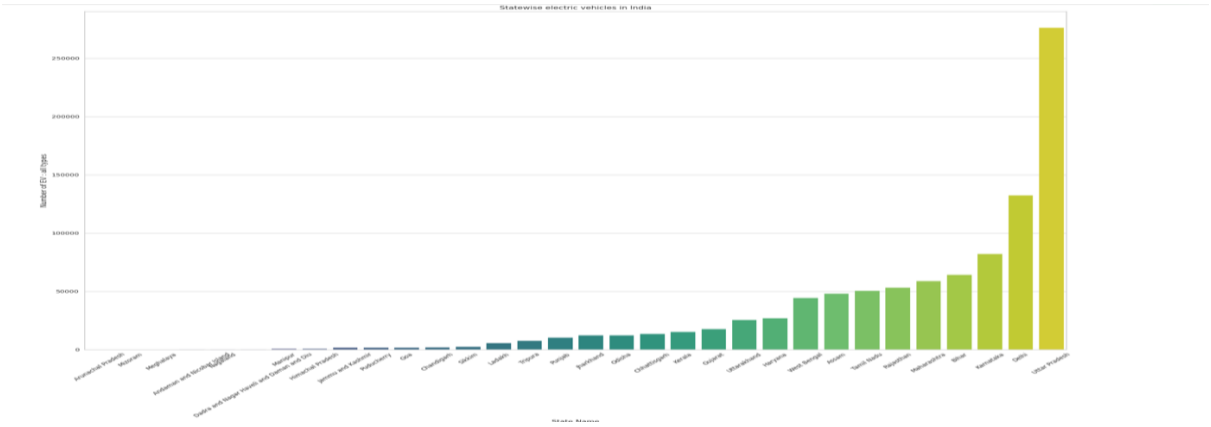
```
Index(['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7', 'PC8', 'PC9'], dtype='object')
```

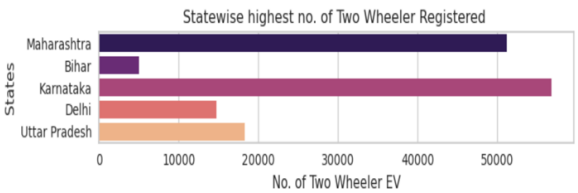
```
cdf=pd.DataFrame(lm.coef_, X.columns, columns=['Coeff'])
cdf
```

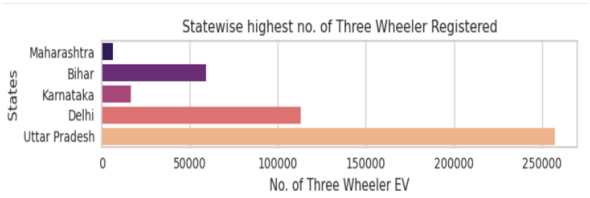
	Coeff
PC1	1101.5872
PC2	-741.2090
PC3	208.5362
PC4	508.3225
PC5	122.3533
PC6	1579.0069
PC7	333.6115
PC8	-1079.9951

- Geographical Segmentation :

The bar graph below shows the diversity of the data geographically.







5. Most optimal -Target Market Segments:

So from the analysis I can see that the Factors that affects EV Range are Top Speed, Efficiency, Acceleration, Segment, Seats and Price. Optimum targeted segment should be belonging to the following categories:

- **Behavioural :** Mostly from our analysis there are cars with 5 seats, most cars are either SUV or Hatchback and most used plug type is 2 CCS and plug type 1 CHAdeMO is the least used.

- **Demographic :**

1. *Top Speed & Range* : With a large area of market the cost is dependent on Top speeds and Maximum range.
2. *Efficiency* : Mostly the segments are with most efficiency.

- **Psychographic :**
  Price : From the above analysis, the price range is between 16,00,000 to 1,80,00,000.

- **Geographic** : We can see that, the maximum amount of sale in states Karnataka and Maharashtra and Uttar Pradesh; and minimum amount of sale in Arunachal Pradesh, Meghalaya, Mizoram, Nagaland, Manipur, Dadra and Nagar Haveli and Daman and Diu.

6. **Customizing the Marketing Mix**

• **Price:** *refers to the value that is put for a product. It depends on segment targeted, ability of the companies to pay, ability of customers to pay supply - demand and a host of other direct and indirect factors.*

• **Product:** *refers to the product actually being sold, the product must deliver a minimum level of performance; otherwise even the best work on the other elements of the marketing mix won't do any good.*

• **Place:** *refers to the point of sale. In every industry, catching the eye of the consumer and making it easy for her to buy it is the main aim of a good distribution or 'place' strategy*.

• **Promotion:** *To make the product and service known to the user and trade, direct & Indirect marketing required through advertising, word of mouth, press reports, commissions, etc. It can also include consumer schemes, contests and prizes, awards to the trade.*

7. Final conclusion :

- Our target segment should contain cars with most **Efficiency**, contains **Top Speed** and **price** between **16 to 180 lakhs** with mostly with **5 seats**.
- EV market would expect growth in metropolitan areas of India. States like Maharashtra and Karnataka present significant market potential.
- Established players like Tata Motors, Mahindra & Mahindra, and new entrants like Tesla and Ola Electric shape the competitive landscape.

8. Additional datasets requirement and ML models can be helpful improve upon the Market Segmentation EV

8.1 Fields which can be considered:

- Social media usage
- Online shopping frequency
- Concern about climate change
- Interest in sustainable products
- Vehicle maintenance habits
- Word-of-mouth referrals
- First-time buyers

8.2 ML models can be helpful improve Segmentation:

- Segmentation based on behaviour: ML models can analyse customer behaviour, such as driving patterns, charging habits, and vehicle usage, to create segments with similar characteristics.
- Propensity scoring: ML algorithms can assign scores to customers based on their likelihood of adopting EVs, allowing for focused marketing efforts.

9. Estimated Market Size for EV Market Domain (non-segmented) :

- Units : 2030 : 5,000,000 units (Source: India Energy Storage Alliance)
- Revenue (INR) : 2030: ₹2,00,000 crores (Source: NITI Aayog)
- Growth Rate : CAGR (2025-2030): 30% (Source: NITI Aayog)

10. Top 4 Variables/features which can be used to create most optimal Market Segments :

- Vehicle preferences
- Technology Adoption & Vehicle Usage Patterns
- Socio-Economic Factors
- Purchase influencers