Electric Vehicle Market in India

Market segmentation

Harsha Vardhan

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 profiling 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.

Implementation

Tools used:

- Numpy
- Pandas
- Seaborn
- Sk-Learn
- Matplotlib

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.

```
df = pd.read_csv('data.csv')
df.drop('Unnamed: 0', axis=1, inplace=True)
df['inr(10e3)'] = df['PriceEuro']*0.08320
df['RapidCharge'].replace(to_replace=['No','Yes'],value=[0, 1],inplace=True)
df.head()
```

	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment	Seats	PriceEuro	inr(10e3)
0	Tesla	Model 3 Long Range Dual Motor	4.6000	233	450	161	940	1	AWD	Type 2 CCS	Sedan	D	5	55480	4615.9360
1	Volkswagen	ID.3 Pure	10.0000	160	270	167	250	0	RWD	Type 2 CCS	Hatchback	C	5	30000	2496.0000
2	Polestar	2	4.7000	210	400	181	620	1	AWD	Type 2 CCS	Liftback	D	5	56440	4695.8080
3	BMW	iX3	6.8000	180	360	206	560	1	RWD	Type 2 CCS	SUV	D	5	68040	5660.9280
4	Honda	e	9.5000	145	170	168	190	1	RWD	Type 2 CCS	Hatchback	B	4	32997	2745.3504

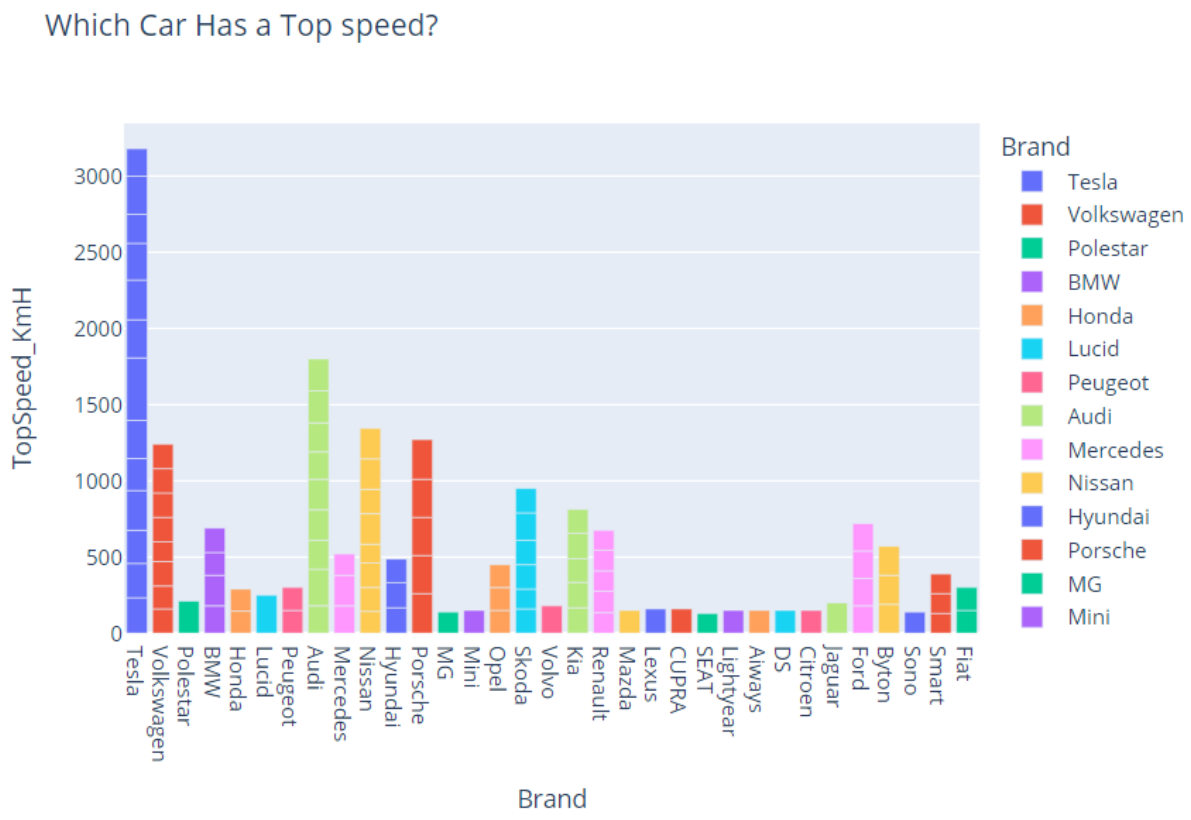
```
df[(df['Brand'] == 'Tesla ') | (df['Brand'] == 'BMW ')]
```

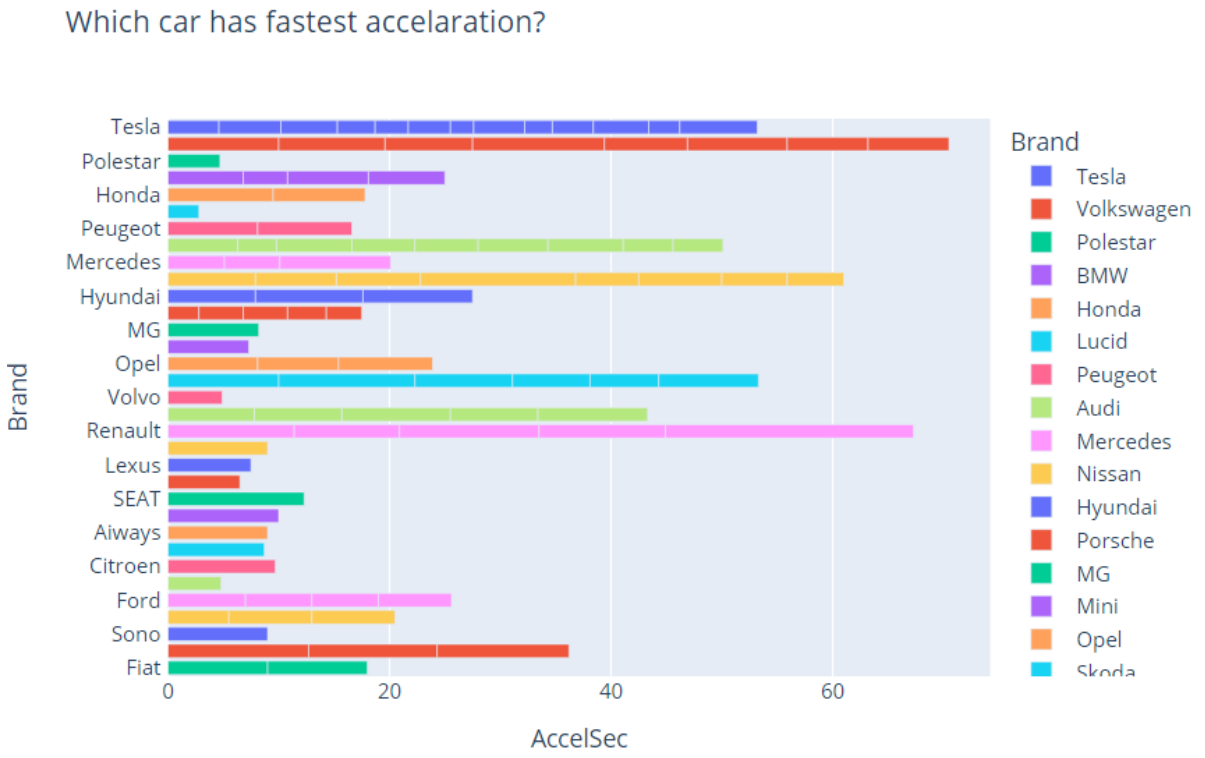
	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment	Seats	PriceEuro	inr(10e3)
0	Tesla	Model 3 Long Range Dual Motor	4.6000	233	450	161	940	1	AWD	Type 2 CCS	Sedan	D	5	55480	4615.9360
3	BMW	iX3	6.8000	180	360	206	560	1	RWD	Type 2 CCS	SUV	D	5	68040	5660.9280
8	Tesla	Model 3 Standard Range Plus	5.6000	225	310	153	650	1	RWD	Type 2 CCS	Sedan	D	5	46380	3858.8160
13	BMW	i4	4.0000	200	450	178	650	1	RWD	Type 2 CCS	Sedan	D	5	65000	5408.0000

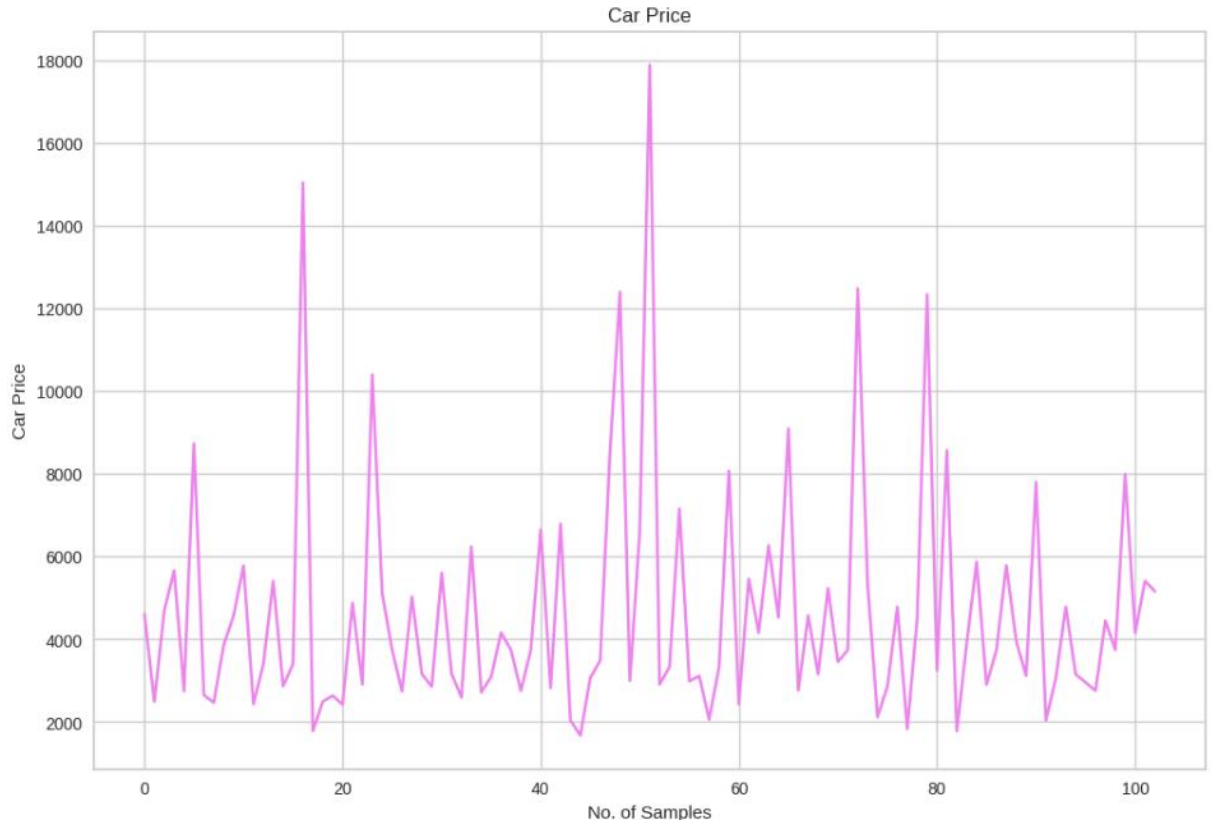
Exploratory Data Analysis

.

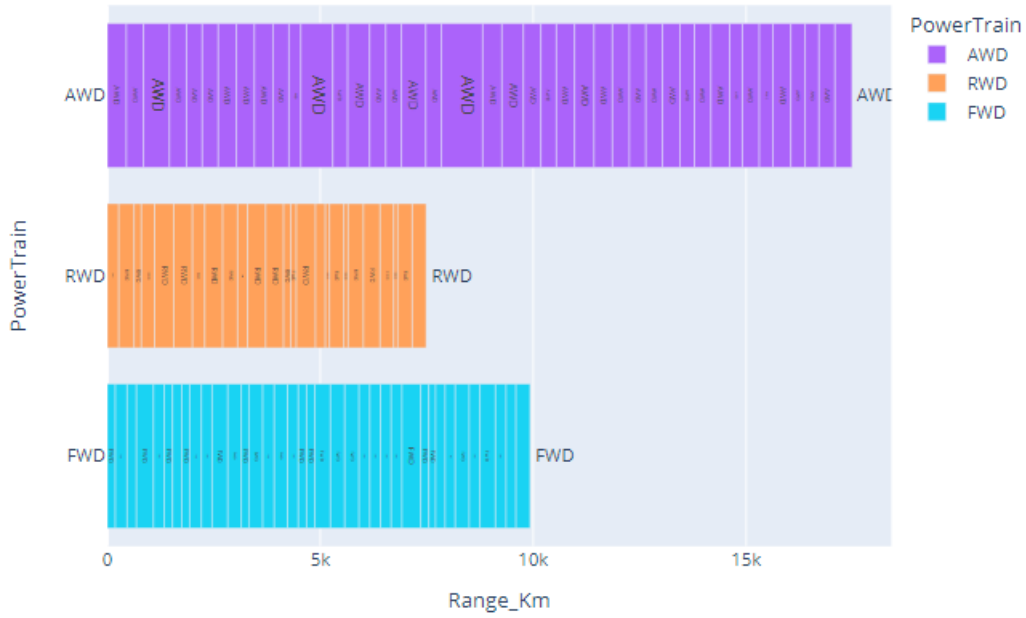
Comparison of cars :



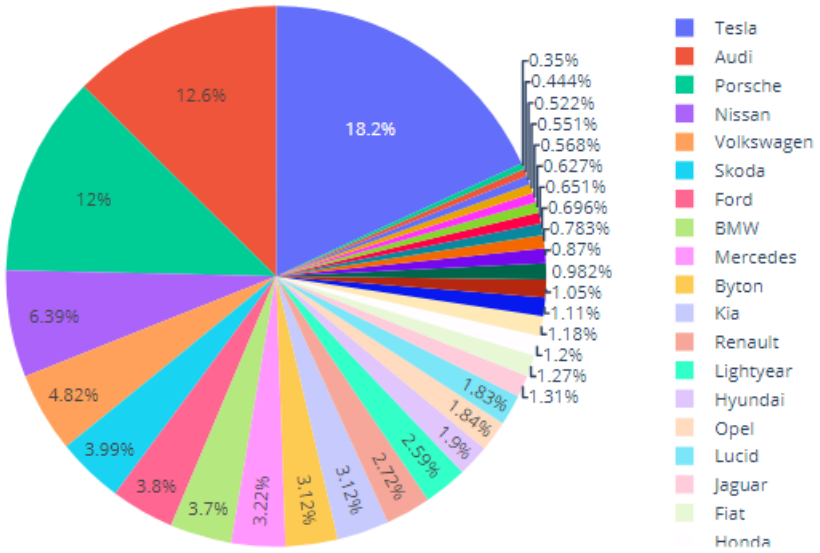

Which car has fastest accelaration?
Brand
Tesla
Polestar
Honda
Peugeot
Mercedes
Hyundai
MG
Opel
Volvo
Renault
Lexus
SEAT
Aiways
Citroen
Ford
Sono
Fiat
0
20
40
60
AccelSec
Brand
Tesla
Volkswagen
Polestar
BMW
Honda
Lucid
Peugeot
Audi
Mercedes
Nissan
Hyundai
Porsche
MG
Mini
Opel
Skoda


Car Price
Car Price
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
20
40
60
80
100
No. of Samples

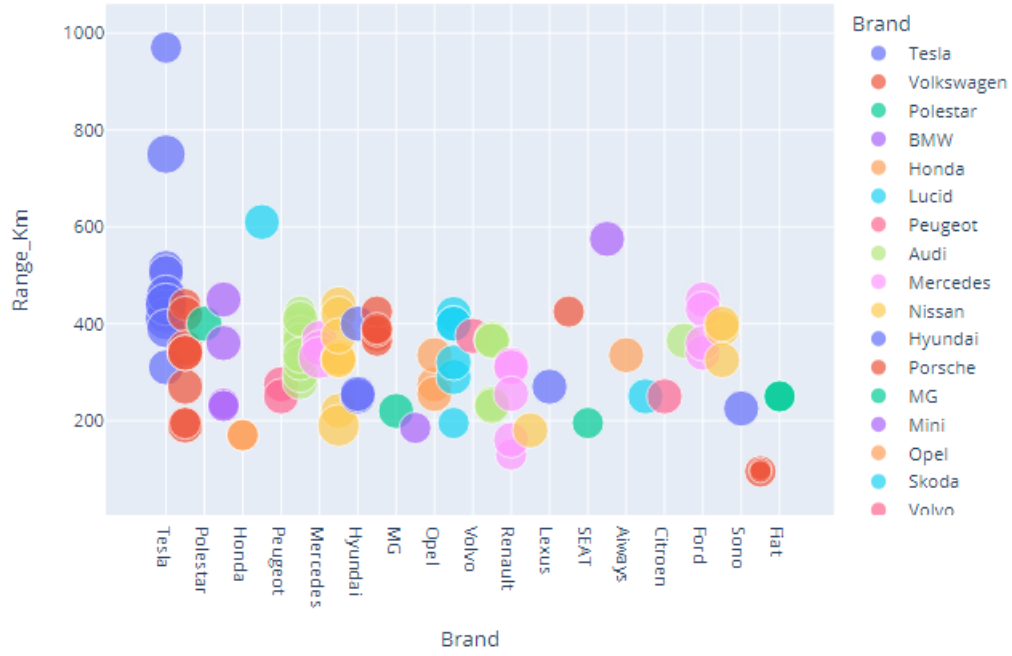
Comparision of power train :



Market Share or Value Proportion:

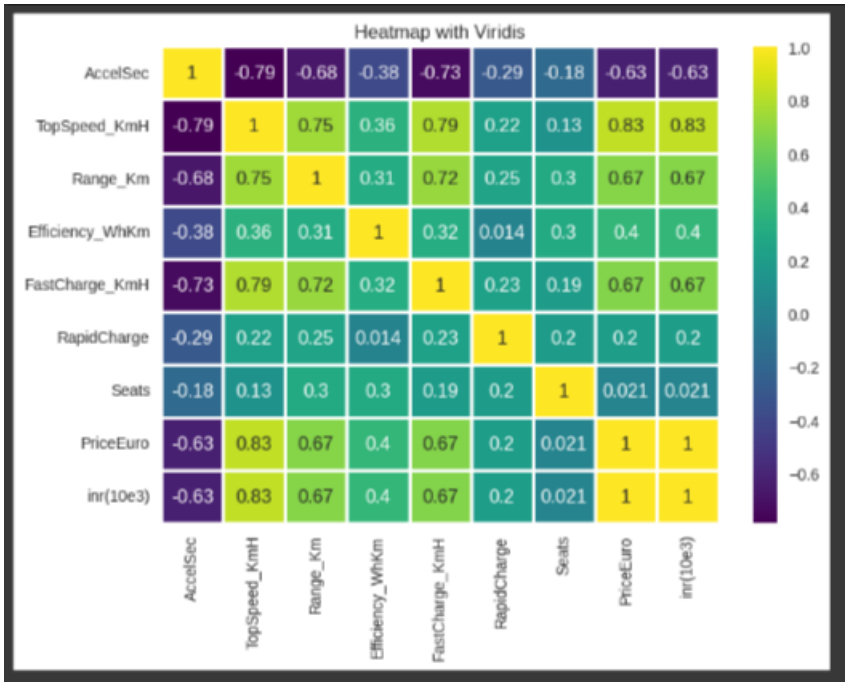


Comparison of the range and seating capacity of vehicles across different brands :

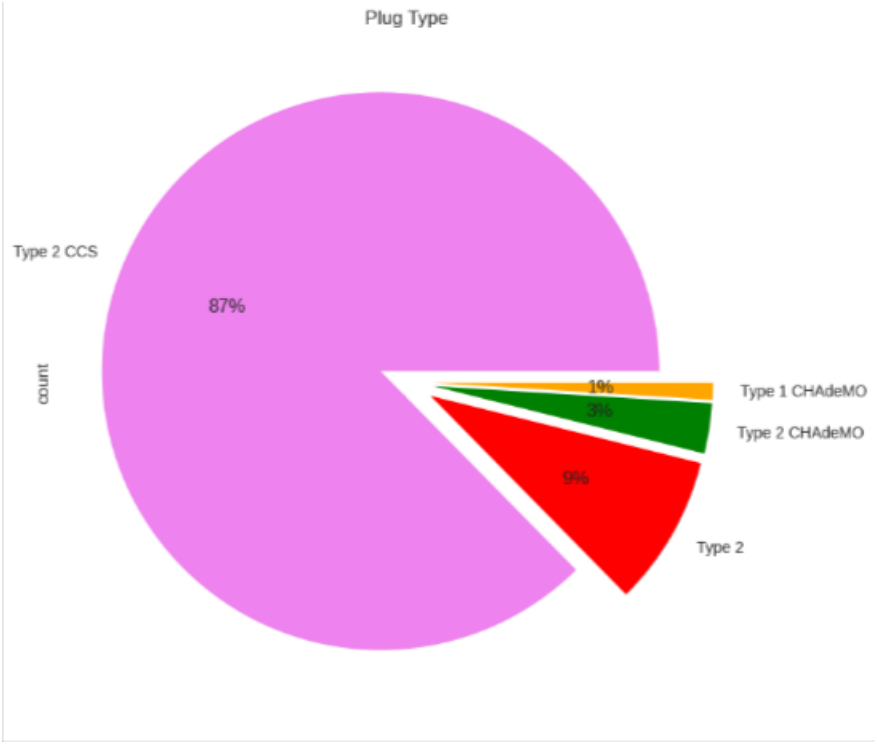


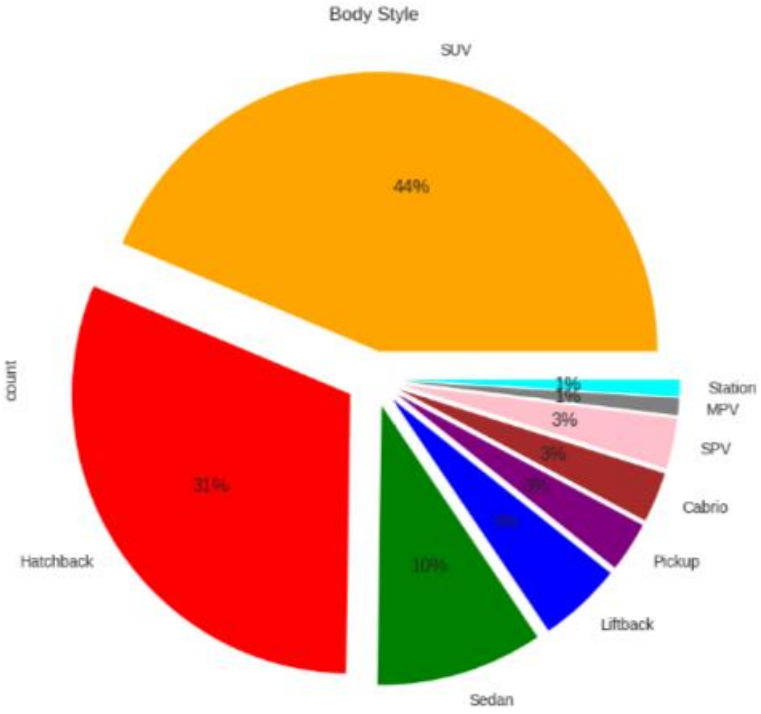
Correlation Matrix :

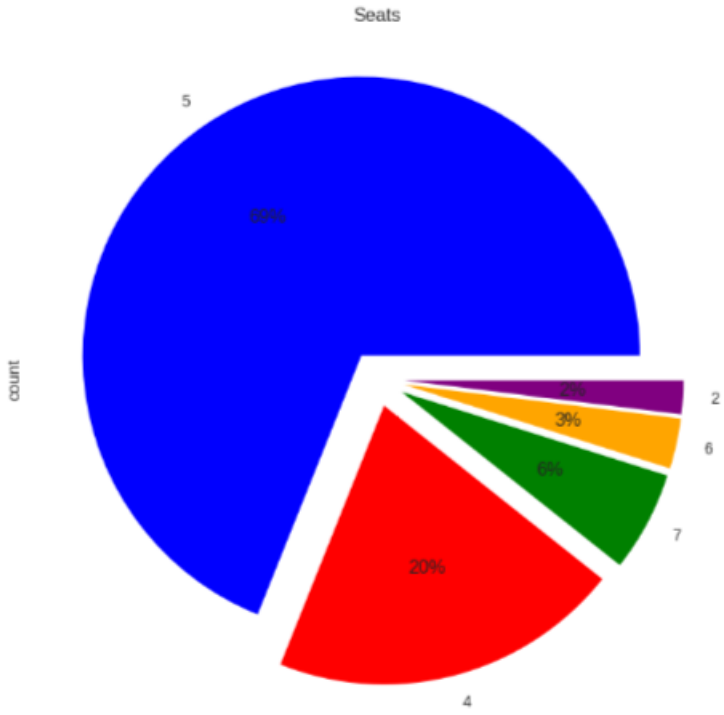
A correlation matrix is a table that displays the correlation coefficients between multiple variables. It is particularly useful for identifying linear relationships between variables. The matrix shows the correlation between all possible pairs of values, typically visualized through a heatmap as illustrated in the figure below. A relationship between two variables is generally considered strong when their correlation coefficient is greater than 0.7.

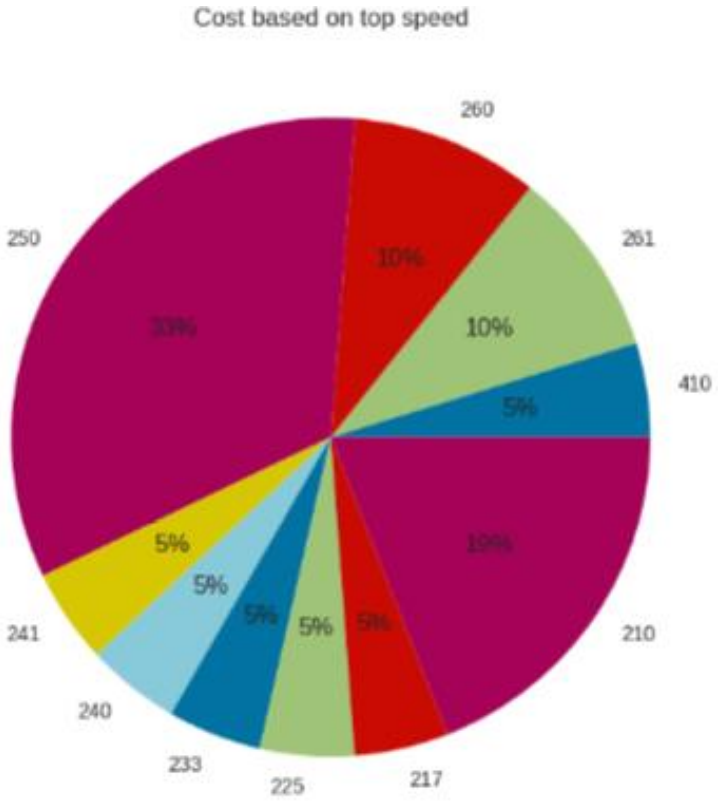


Some other Key observations :

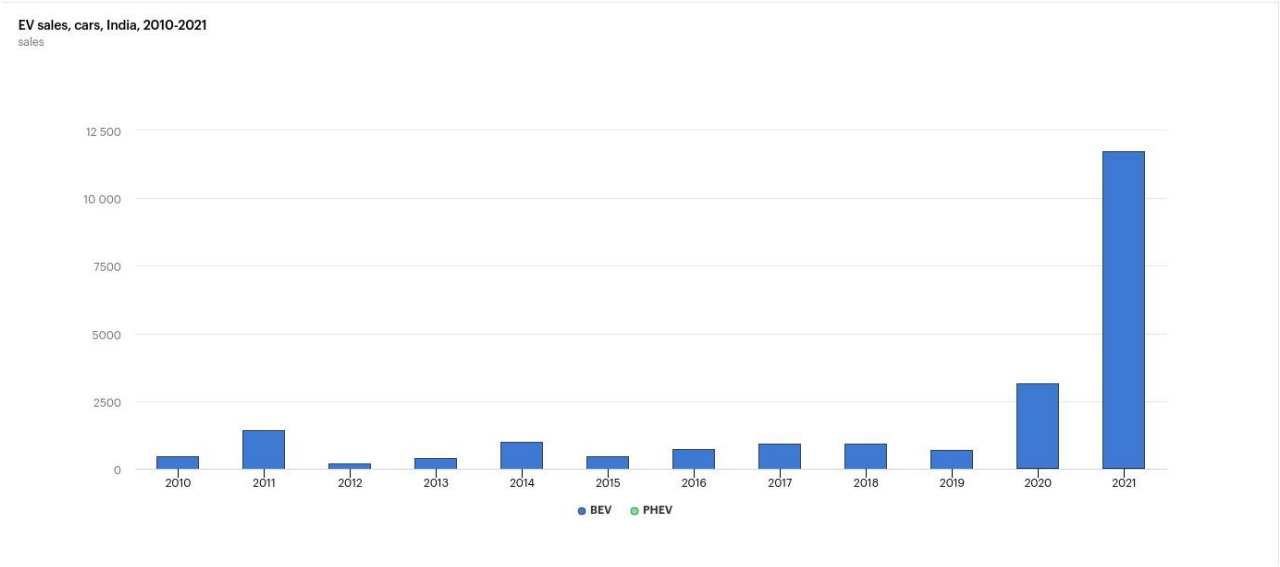




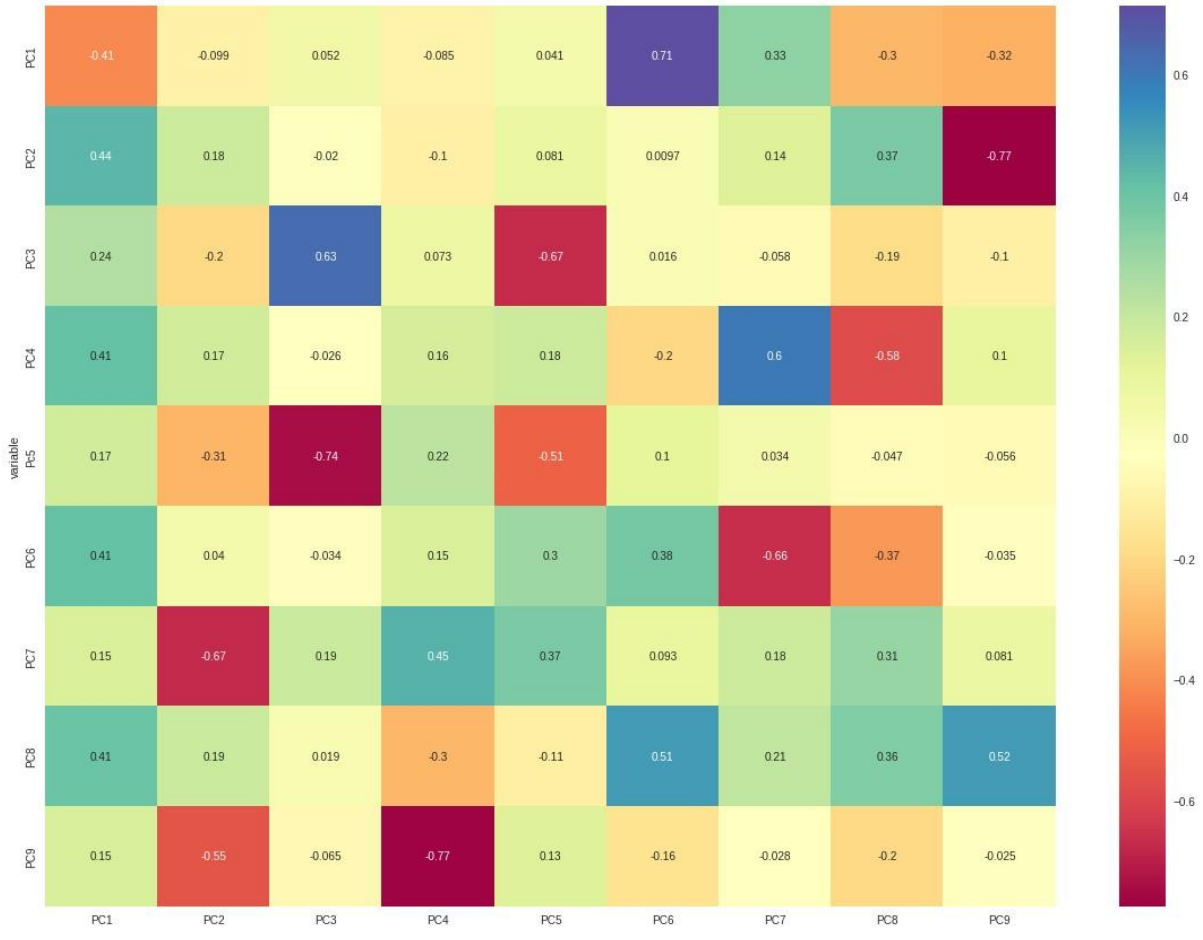

Seats
count
5
55%
2
2%
6
3%
7
6%
4
20%


Cost based on top speed
260
10%
261
10%
410
5%
210
10%
217
5%
225
5%
233
5%
240
5%
241
5%
250
20%

Now we can see that the requirements of what type of cars are most needed for customers and from the past 10 years there is a rapid growth of Electric vehicles usage in India

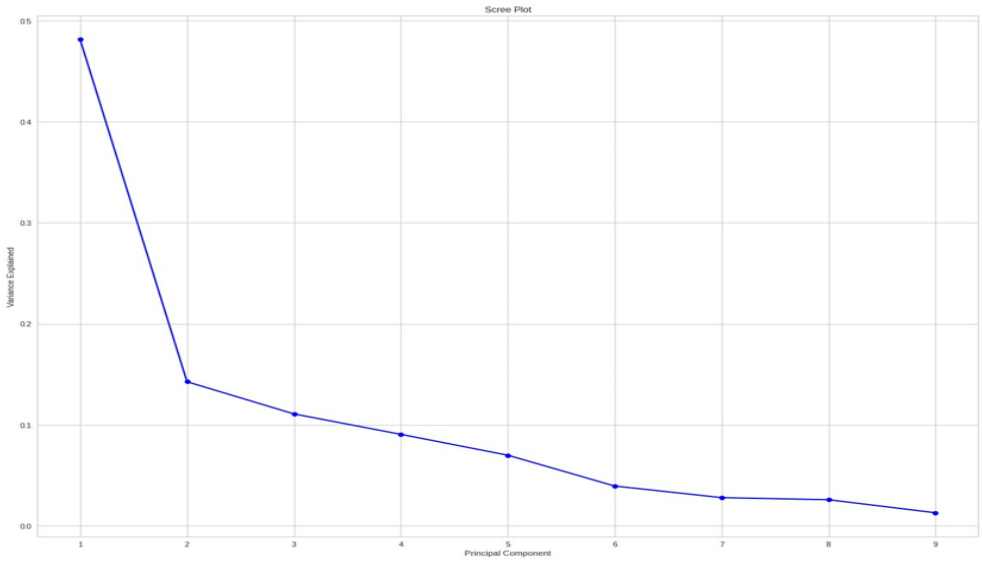


Correlation matrix plot for loadings :



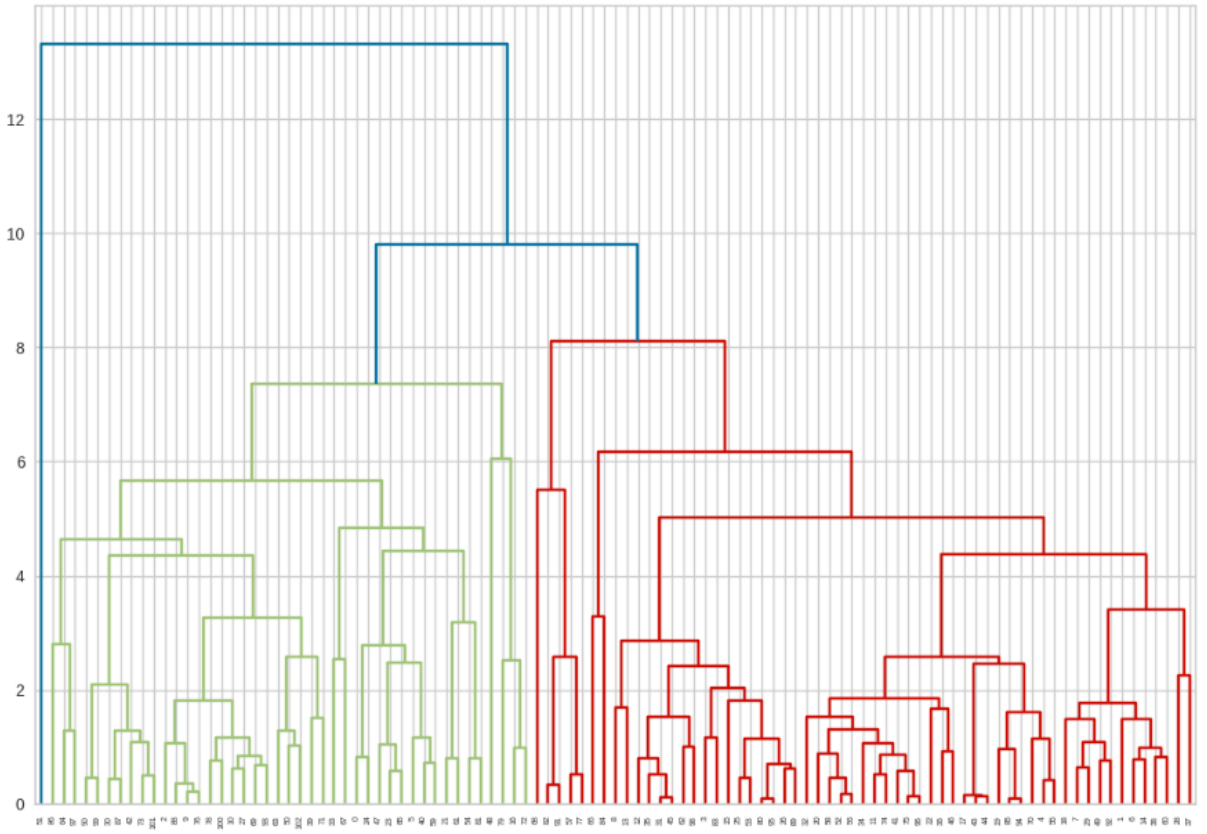
Scree Plot

A scree plot is a common graphical method used to determine the number of principal components (PCs) to retain in an analysis. It is a simple line plot that shows the eigenvalues for each individual PC. The y-axis represents the eigenvalues, while the x-axis represents the number of factors. The plot typically displays a downward curve, starting high on the left, dropping quickly, and then flattening out. This shape occurs because the first component usually explains a significant portion of the variability, the next few components explain a moderate amount, and the latter components explain only a small fraction of the overall variability. The scree plot criterion involves identifying the “elbow” point in the curve and selecting all components just before the line flattens out. Additionally, the selected PCs should be able to describe at least 80% of the variance in the data.



Extracting Segments using dendrogram

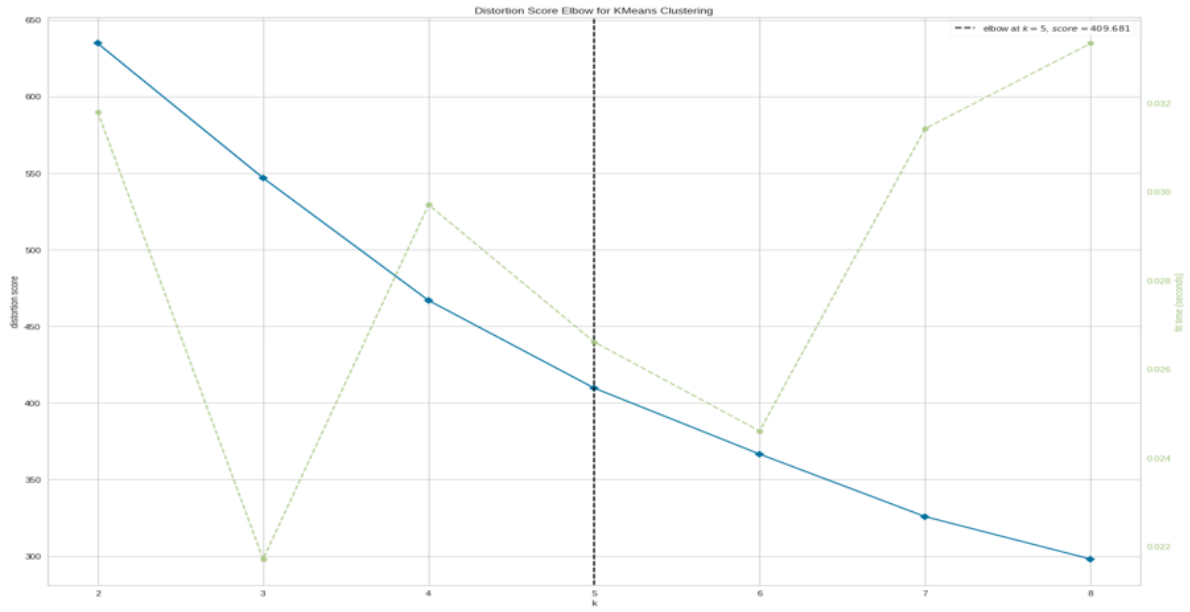
This technique is specific to the agglomerative hierarchical method of clustering. This method starts by treating each data point as its own cluster and then sequentially merges points and clusters based on their distances in a hierarchical manner. To determine the optimal number of clusters, we use a dendrogram—a tree-like chart that illustrates the sequence of merges or splits among clusters. In a dendrogram, when two clusters merge, they are connected by a line, and the height of the line represents the distance between the clusters. As shown in the figure, the optimal number of clusters can be determined based on the hierarchical structure of the dendrogram. Other cluster validation metrics suggest that four to five clusters may also be appropriate for agglomerative hierarchical clustering.



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.

Evaluating the clusters using Distortion



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 this use case, our focus is to derive the subgroups of the Indian population to target for Electric vehicles based on their socio-demographic and psychographic features. We have used the K-Means algorithm which is one of the simplest and fastest unsupervised algorithms that solves the well-known clustering problem.

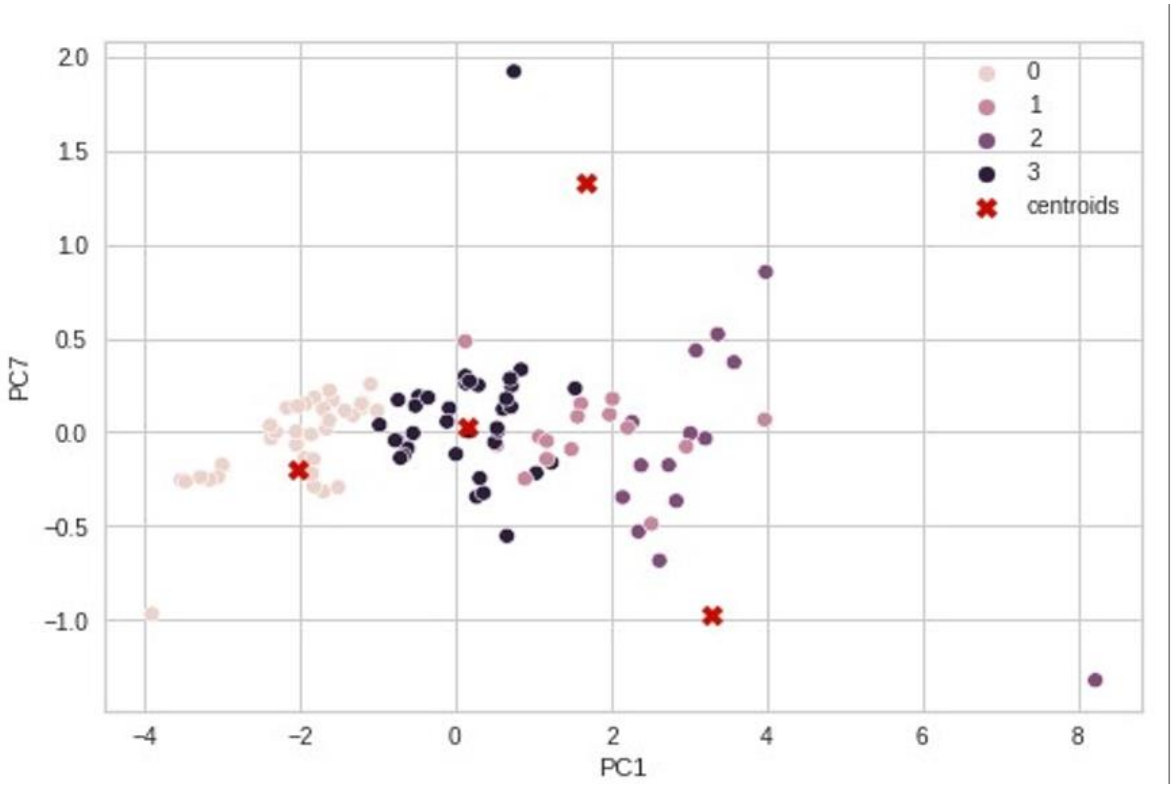
K-Means:

K-means is one of the simplest unsupervised learning algorithms that solves the clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). Data points inside a cluster are homogeneous and heterogeneous to peer groups.

- The objective of K means is to minimize the sum of squared distances between all points and the center of the cluster.
- The K-Means algorithm can be summarized as follows:
    1. Define the number of clusters K
    2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
    3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
    - 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.

K-Means Clustering

The task of the K-means clustering algorithm is to divide the data into K clusters. Each cluster has a centroid. The sum of the squared distances between the data points and the centroids of the clusters is minimized. We have used the K-means clustering algorithm from the scikit-learn library to extract the cluster. We have used K=4 to get our cluster to describe the market segmentation.



Prediction of Prices most used cars

Linear regression is a machine learning algorithm based on supervised learning. It performs 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.

Regression for data2

Add blockquote

```
X=data2[['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])
```

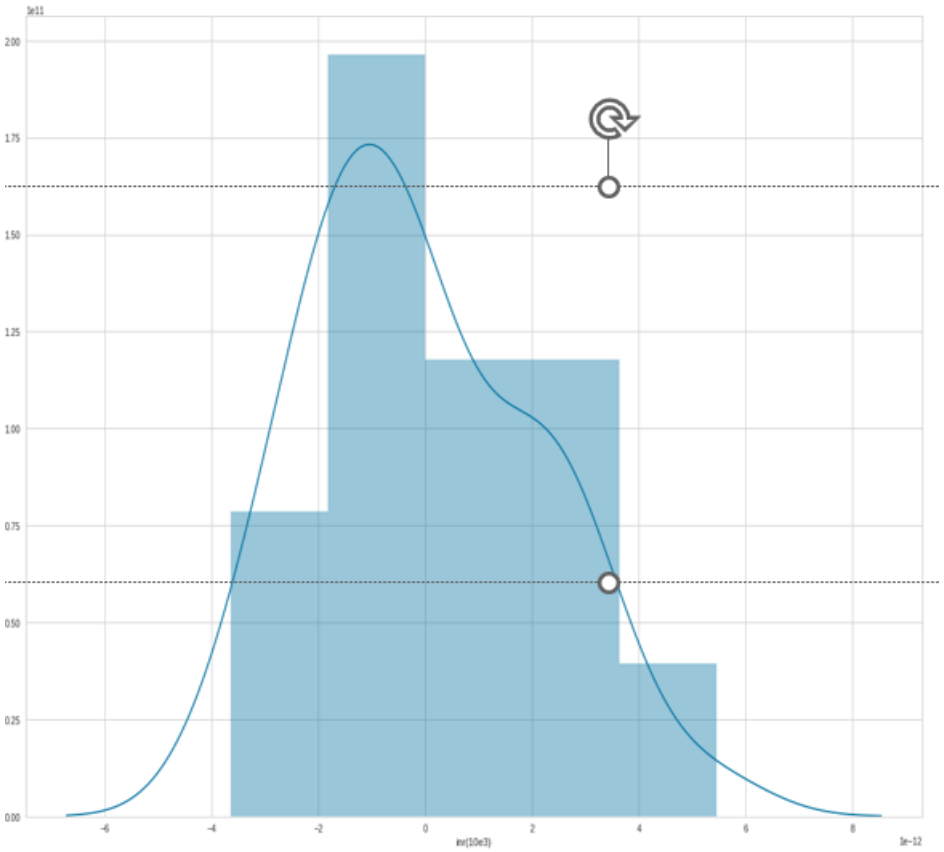
```
X_train.columns
```

```
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
PC9	1461.7227

After completion of training the model process, we test the remaining 60% of data on the model. The obtained results are checked using a scatter plot between predicted values and the original test data set for the dependent variable and acquired similar to a straight line as shown in the figure and the density function is also normally distributed.



Metrics :

```
print('MAE:',metrics.mean_absolute_error(y_test,predictions))
print('MSE:',metrics.mean_squared_error(y_test,predictions))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test,predictions)))
```

```
MAE: 2.2629094365540715e-12
MSE: 1.0768119045556378e-23
RMSE: 3.281481227366138e-12
```

```
metrics.mean_absolute_error(y_test,predictions)
```

```
2.2629094365540715e-12
```

```
metrics.mean_squared_error(y_test,predictions)
```

```
1.0768119045556378e-23
```

```
np.sqrt(metrics.mean_squared_error(y_test,predictions))
```

```
3.281481227366138e-12
```

Conclusion :

Behavior : Segmenting potential customers based on their behaviour. This can include factors such as their purchasing habits, brand loyalty, and product usage.

Demographics : Segmenting potential customers based on demographic factors such as age, gender, income, education, and occupation.

Psychographics :Segmenting potential customers based on their lifestyle, personality, values, and interests.

Efficiency :Segmenting potential customers based on efficiency. This can include factors such as energy consumption, battery life, and charging time.

Cars : Segmenting potential customers based on the type of car. This can include factors such as the size, features, and performance of the car.

Segmentation Strategy

Segmenting the potential customer base for Electric Vehicles based on the above mentioned parameters we get the following four segments:

1. **Conservative Buyers**
2. **Early Adopters**
3. **Mass Market Buyers**
4. **Innovators**

Marketing Mix

Price

Setting a price point that reflects the value of the electric vehicle and is attractive to potential customers. This may include pricing strategies such as discounts, financing options, and rebates.

Product

Ensuring that the electric vehicle meets the needs and preferences of potential customers. This may include offering a range of models with different features, battery capacities, and performance levels.

Place

Making the electric vehicle available to potential customers through a variety of channels. This may include online sales, dealerships, and partnerships with other retailers.

Promotion

Promoting the electric vehicle to potential customers through a variety of channels. This may include advertising, social media, and events.

Identified Segments:

1. **Cars with 5 seats:** The maximum people tend to buy cars having 5 seats.
2. **Top speed and Maximum range:** People are willing to pay a higher price for cars with a top speed and maximum range.
3. **Price Range:** People tend to buy cars within the range of 16 to 180 lakhs.

Git Hub LIink : https://github.com/HarshaManam49/Ev-market-segmentation