

Project Report On

2023

EV Market



Submitted By

Madhur Sharma

Naveen Kumar

Roshan

05/02/2023

Problem Statement: EV Market

Analysing the electric vehicle market in India using Segmentation analysis and deciding the strategy to enter the market, targeting the segments most likely used to use electric vehicles.

For this, we have decided to divide the market segmentation part into three segments.

1. Vehicle Feature Segmentation
2. Customer Behavioural Segmentation
3. Geographical Segmentation

Datasets used

Datasets used for above three market segmentation can be found at the following link:

https://drive.google.com/drive/folders/1Qcz-mhj4XoerI2T0uSBZrVEVtXGKDpVj?usp=share_link

Vehicle Feature Segmentation

Following Python libraries are imported for data analysis and visualization.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import seaborn as sb
import statsmodels.api as sm
from tqdm import tqdm
from google.colab import files
%pip install kaleido
import kaleido
from sklearn.preprocessing import StandardScaler,PowerTransformer
from sklearn.decomposition import PCA
from scipy.cluster.hierarchy import dendrogram , linkage
from sklearn.cluster import KMeans , MeanShift , estimate_bandwidth
from sklearn.datasets import make_blobs
from yellowbrick.cluster import KElbowVisualizer, SilhouetteVisualizer, InterclusterDistance
```

```

from collections import Counter

from sklearn.model_selection import cross_validate,train_test_split

from sklearn.linear_model import LinearRegression,LogisticRegression

from sklearn import metrics

from sklearn.metrics import r2_score,silhouette_score,confusion_matrix,accuracy_score

pd.set_option("display.precision",3)

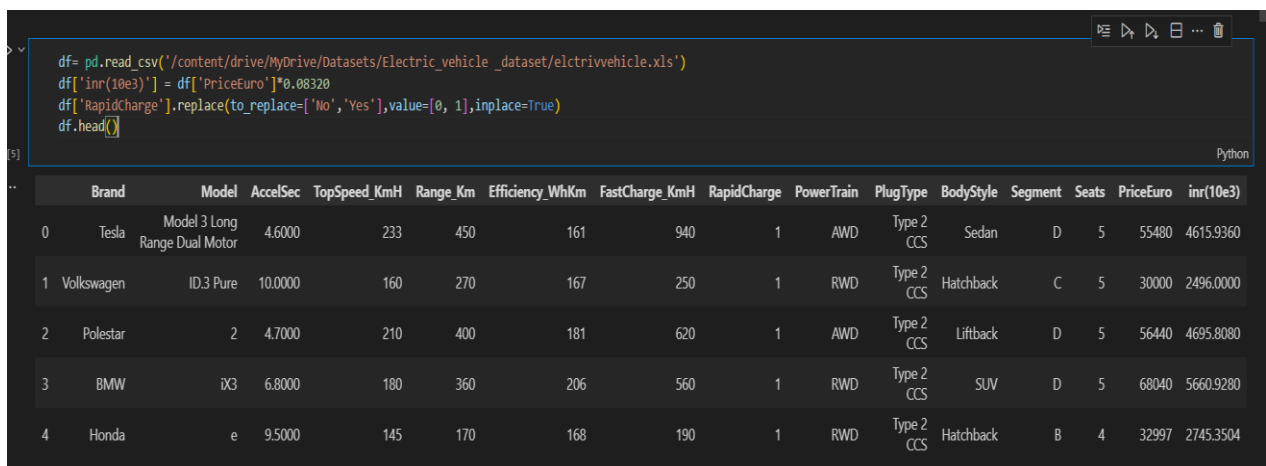
np.set_printoptions(precision=5, suppress=True)

pd.options.display.float_format = '{:.4f}'.format

import plotly.io as pio

We read the dataset using panda libraries and display its first five rows through df.head().

```



The screenshot shows a Jupyter Notebook interface. The code cell contains the following Python code:

```

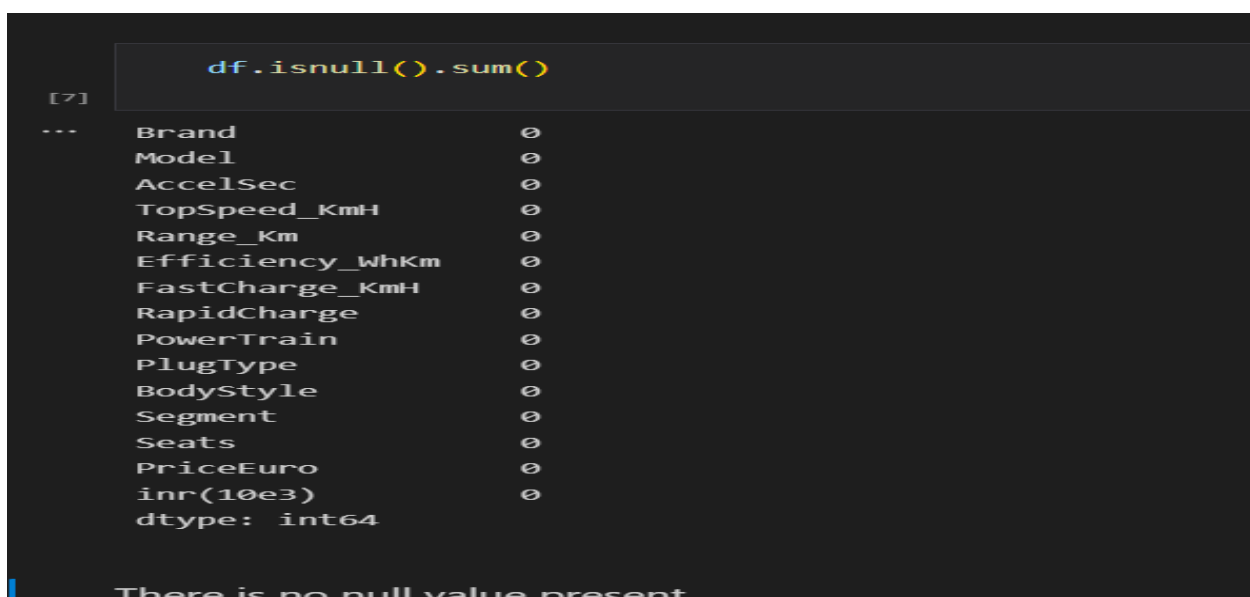
df= pd.read_csv('/content/drive/MyDrive/Datasets/Electric_vehicle _dataset/elctrivvehicle.xls')
df['inr(10e3)'] = df['PriceEuro']*0.08320
df['RapidCharge'].replace(to_replace=['No','Yes'],value=[0, 1],inplace=True)
df.head()

```

The output of the code is a table with 15 columns and 5 rows of data:

	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	1	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

Here we observe that there are 15 columns. It is very important to check null values if it is present. It is required to clean up null values before you pass your data to the machine learning model.



The screenshot shows a Jupyter Notebook interface. The code cell contains the following Python code:

```

df.isnull().sum()

```

The output of the code is a table showing the count of null values for each column:

Brand	0
Model	0
AccelSec	0
TopSpeed_KmH	0
Range_Km	0
Efficiency_WhKm	0
FastCharge_KmH	0
RapidCharge	0
PowerTrain	0
PlugType	0
BodyStyle	0
Segment	0
Seats	0
PriceEuro	0
inr(10e3)	0
dtype:	int64

There is no null value present

We can observe the statistics of the dataset using `df.describe()`.

```
df.describe()
```

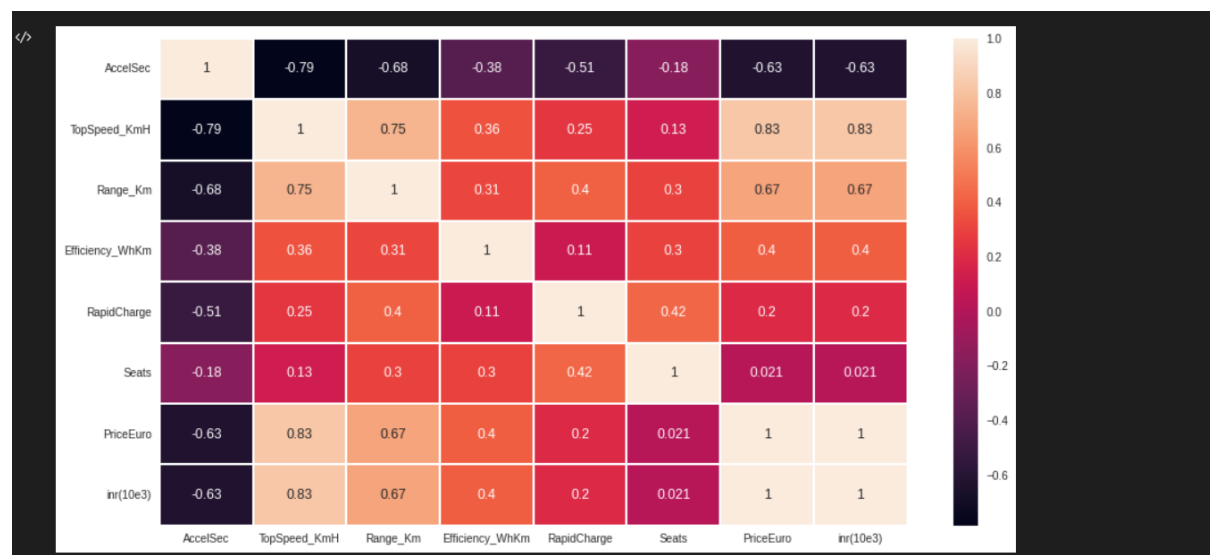
	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	RapidCharge	Seats	PriceEuro	inr(10e3)
count	103.0000	103.0000	103.0000	103.0000	103.0000	103.0000	103.0000	103.0000
mean	7.3961	179.1942	338.7864	189.1650	0.9515	4.8835	55811.5631	4643.5221
std	3.0174	43.5730	126.0144	29.5668	0.2160	0.7958	34134.6653	2840.0042
min	2.1000	123.0000	95.0000	104.0000	0.0000	2.0000	20129.0000	1674.7328
25%	5.1000	150.0000	250.0000	168.0000	1.0000	5.0000	34429.5000	2864.5344
50%	7.3000	160.0000	340.0000	180.0000	1.0000	5.0000	45000.0000	3744.0000
75%	9.0000	200.0000	400.0000	203.0000	1.0000	5.0000	65000.0000	5408.0000
max	22.4000	410.0000	970.0000	273.0000	1.0000	7.0000	215000.0000	17888.0000

It is observed that maximum top speed of the vehicle is 410km/h and the average price of the car is 55811.5631(in Euro). We can also observe that minimum number of seats in the car is 2 and maximum seat is 7. While most of the car has 5 seats.

Correlation of data

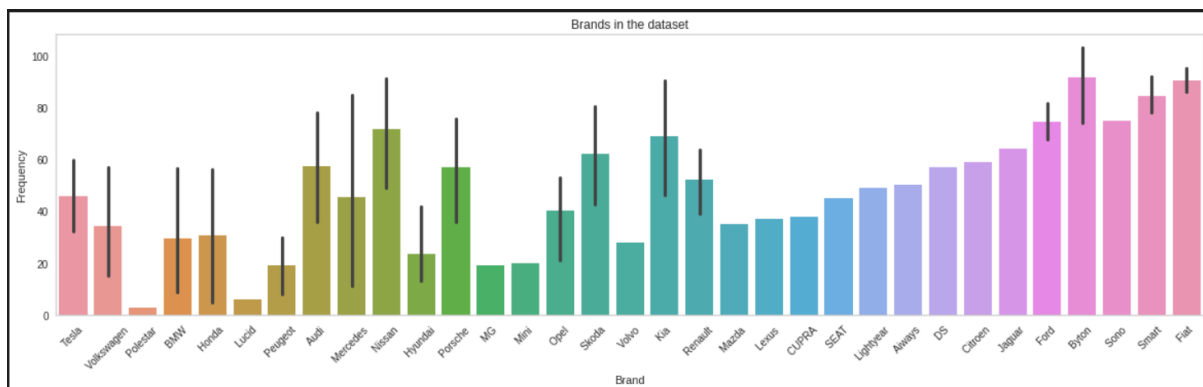
The correlation coefficient helps in measuring the extent of the relationship between two variables in one figure. Correlation facilitates the decision-making in the business world. It reduces the range of uncertainty as predictions based on correlation are likely to be more reliable and near to reality.

We have used Heatmap to show correlation between Data.

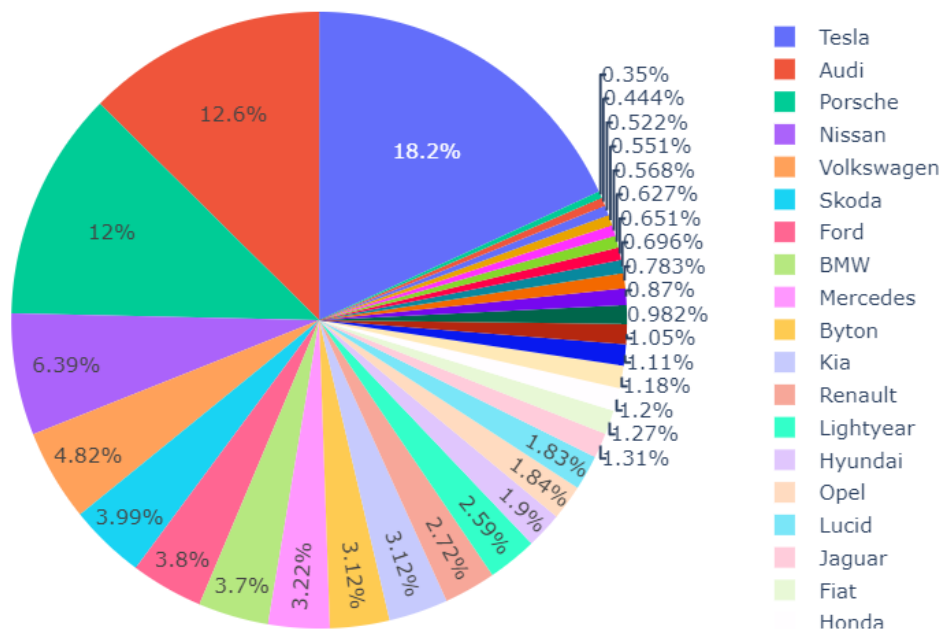


There are total two column that are highly correlated first one is price and Topspeed 0.83 positively correlated and this is obvious thing if price is increase so we get more features and second one correlated feature is topspeed and range is 0.75 positively correlated

Frequency of Brands in the dataset

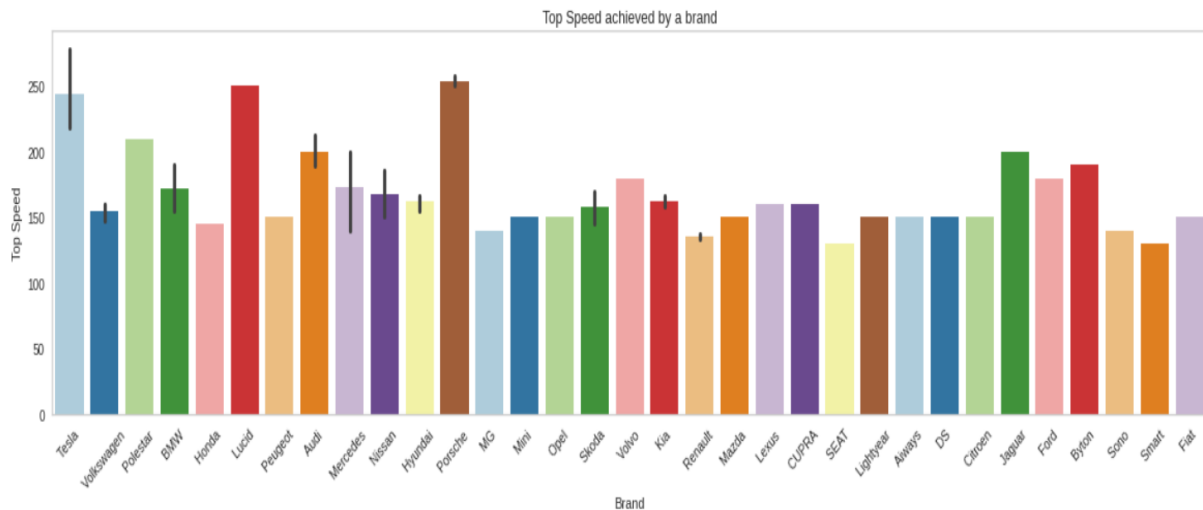


```
fig = px.pie(df, names = 'Brand', values = 'inr(10e3)')
pio.show(fig)
```



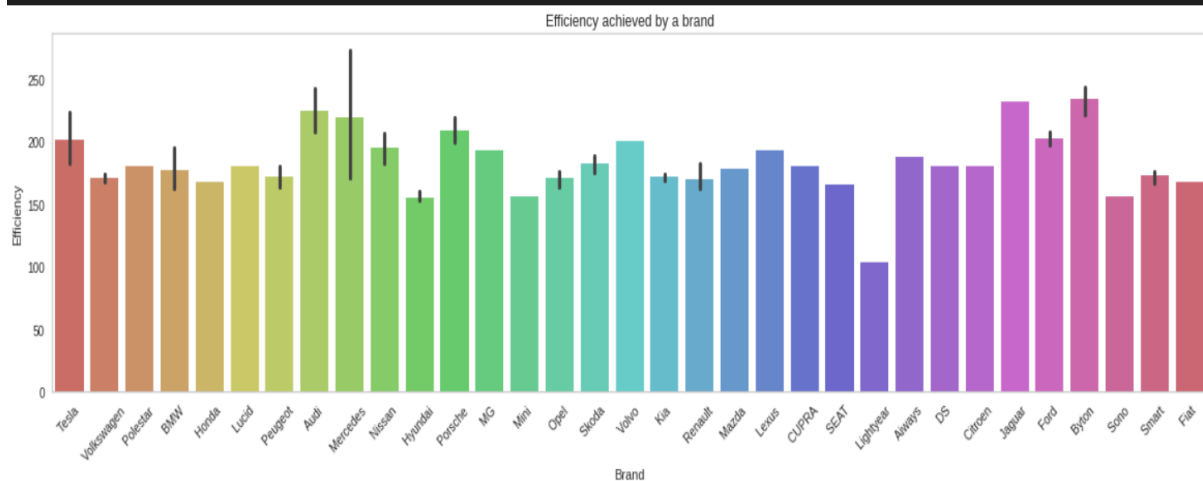
Byton , Fiat and smart are the prominent brands and Polestar being the least

Top Speed of the car



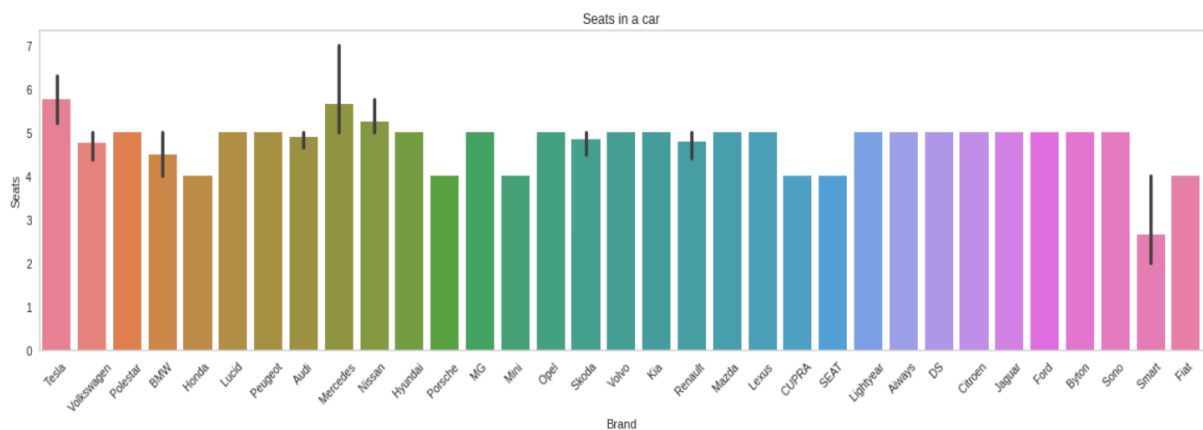
Tesla, Lucid, Porsche produces the fastest car and the Smart slowest.

Car efficiency



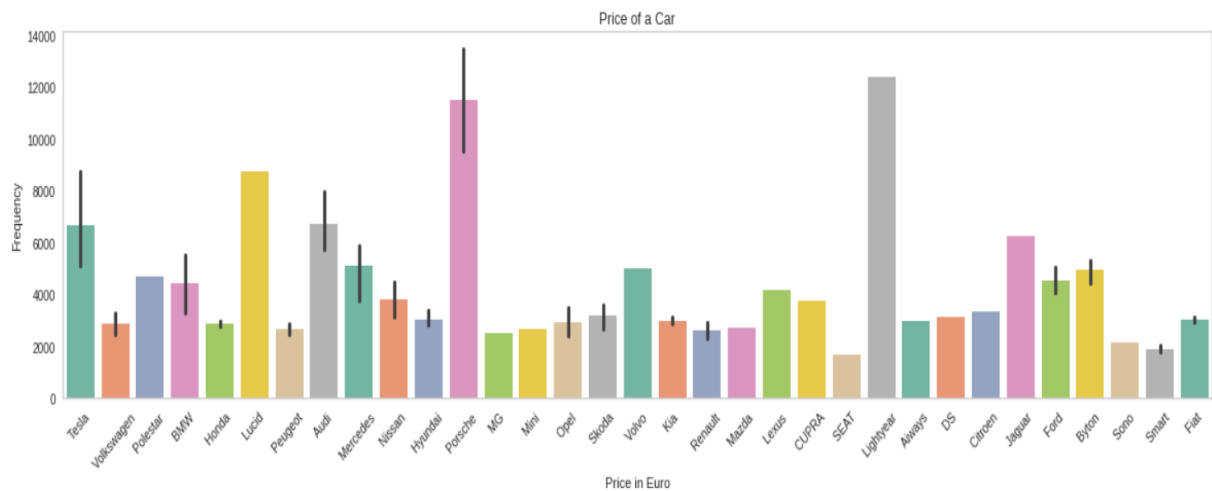
Byton , Jaguar and Audi are the most efficient and Lightyear the least.

Number of seats in car



Mercedes, Tesla, Nissan has the highest number of seats while smart has lowest.

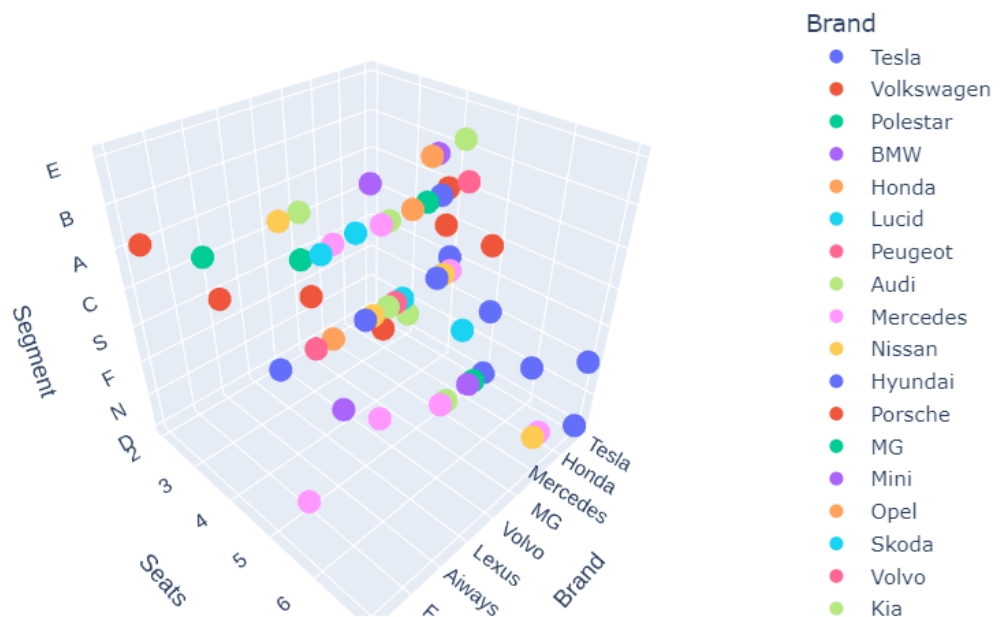
Price of car



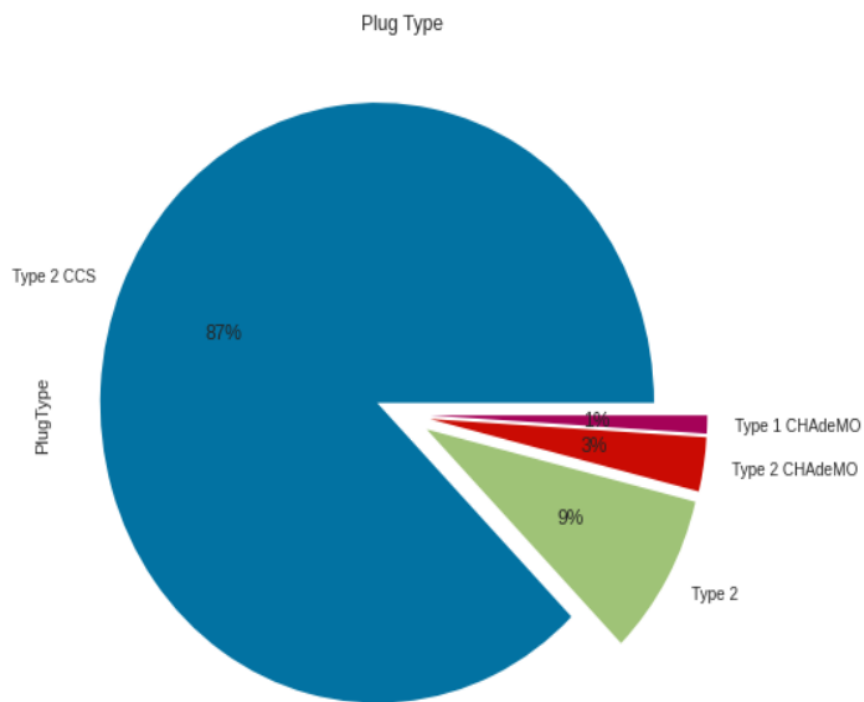
Lightyear, Porsche and Lucid are the most expensive and SEAT and Smart the least

3D visualization of brands,seats, segment

```
fig = px.scatter_3d(df,x = 'Brand',y = 'Seats',z = 'Segment',color='Brand')
pio.show(fig)
```

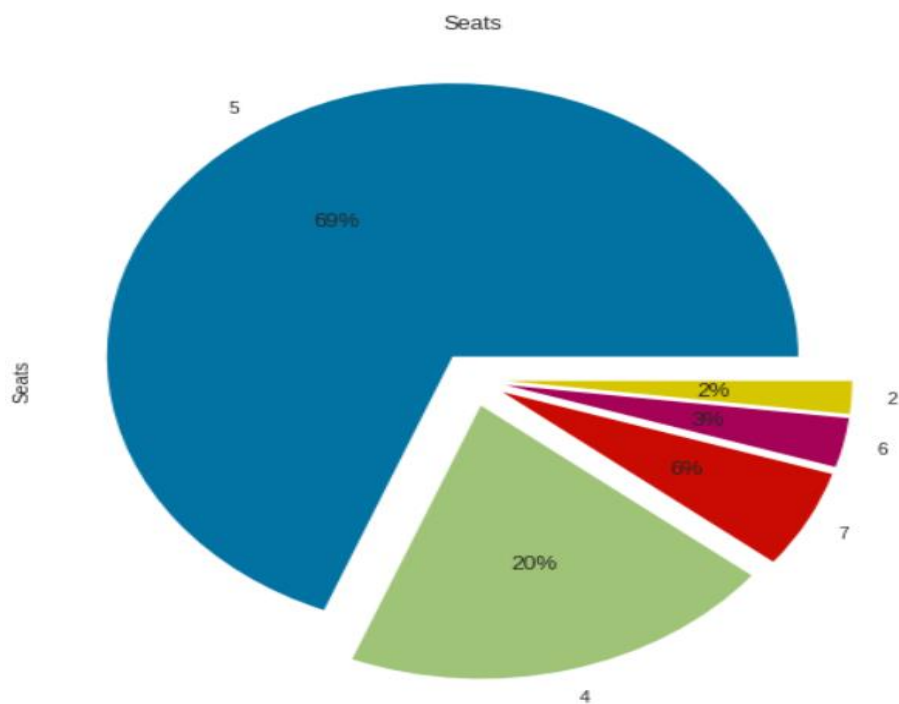


Types of plug used



Type 2 CCS is used widely.

Number of seats in the car

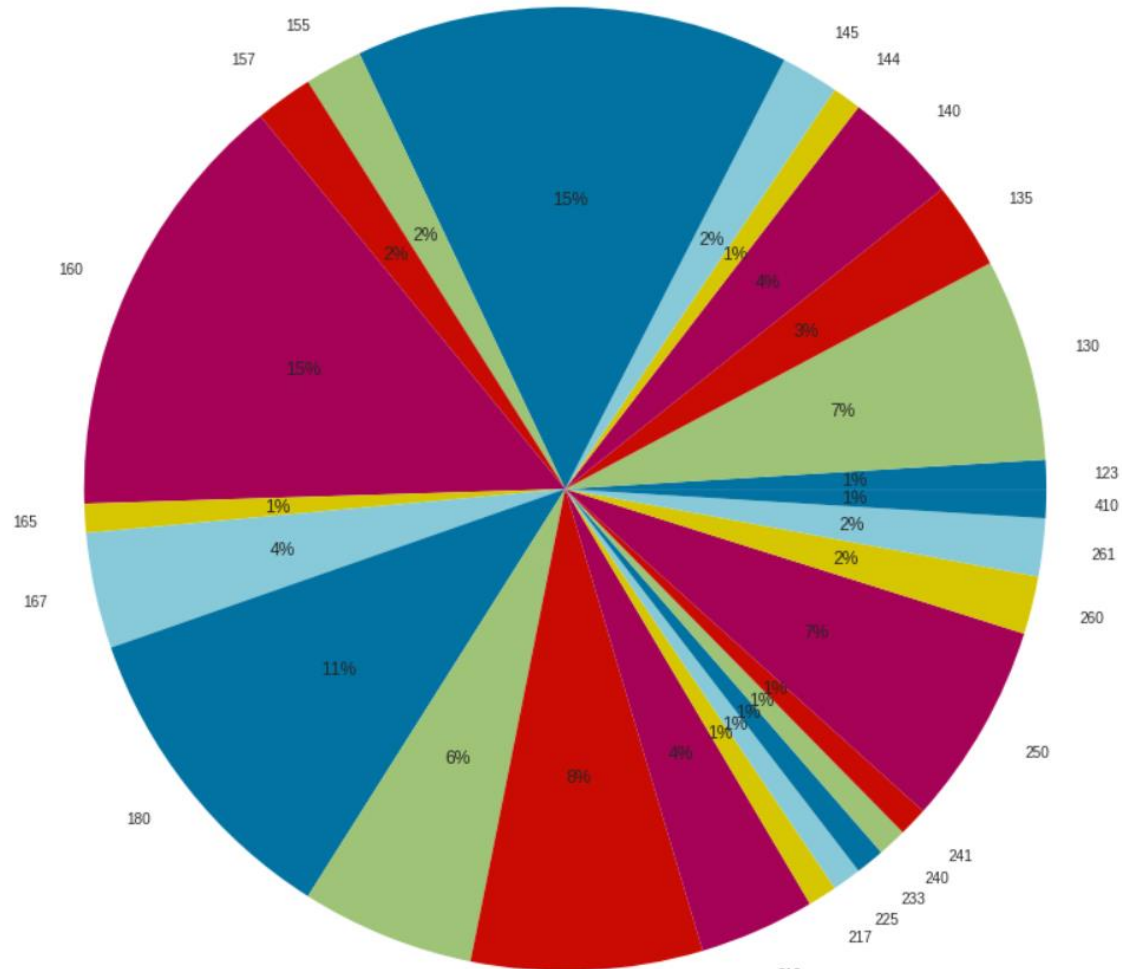


is 5 seated.

69% of total car

Cost based on top speed

```
plt.figure(figsize=(20,15))
plt.title('Cost based on top speed')
plt.pie(x=df3["inr(10e3)"],labels=df3.index,autopct='%1.0f%%')
plt.show()
```



Regression

A regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables. A regression model is able to show whether changes observed in the dependent variable are associated with changes in one or more of the explanatory variables.

It does this by essentially fitting a best-fit line and seeing how the data is dispersed around this line. Regression helps economists and financial analysts in things ranging from asset valuation to making predictions.

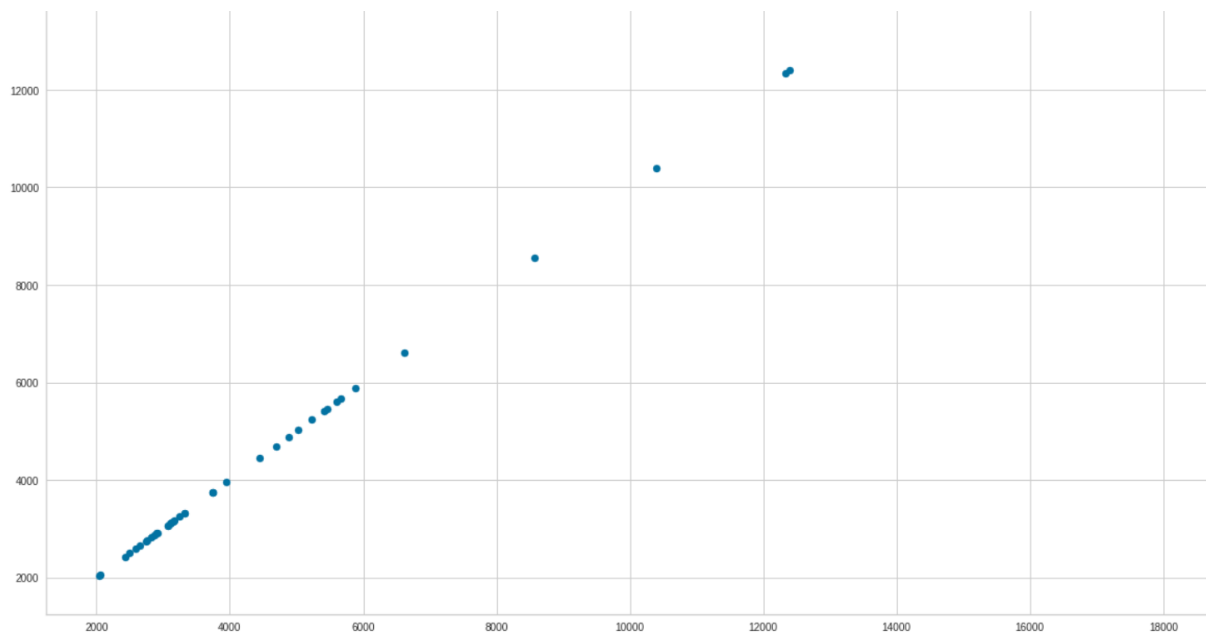
Putting independent variable as x and dependent variables as y

```
df['PowerTrain'].replace(to_replace=['RWD','AWD','FWD'],value=[0, 2,1],inplace=True)
x=df[['AccelSec','Range_Km','TopSpeed_KmH','Efficiency_WhKm','RapidCharge','PowerTrain']]
y=df['PriceEuro']
```

x

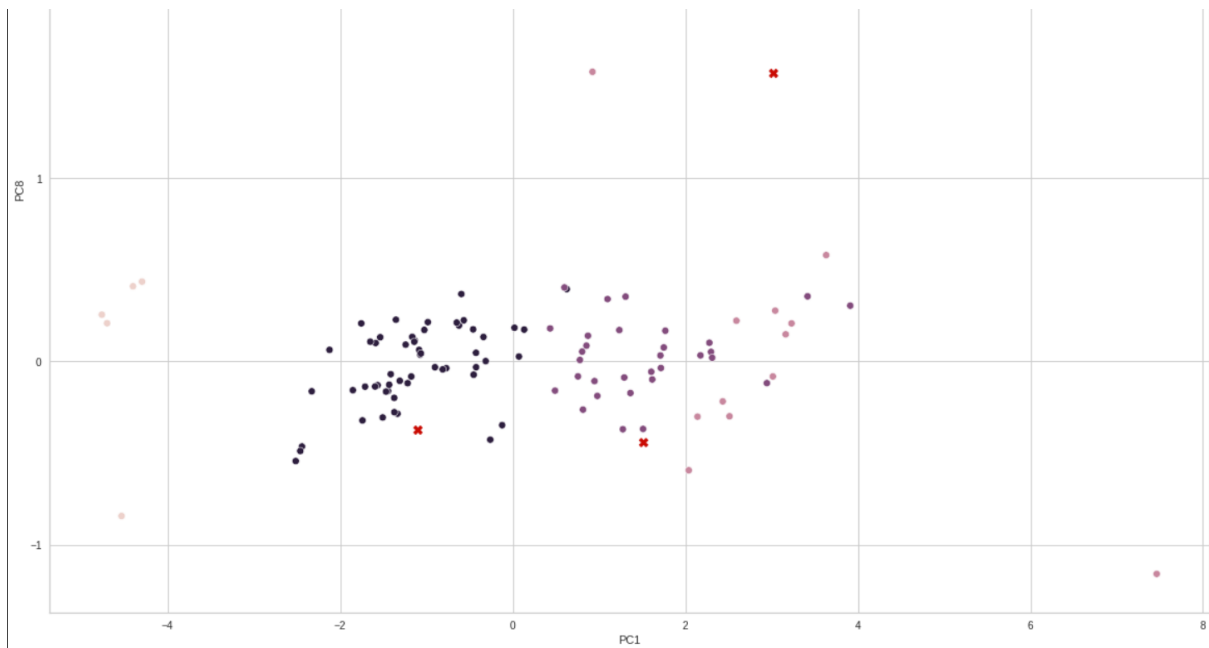
	AccelSec	Range_Km	TopSpeed_KmH	Efficiency_WhKm	RapidCharge	PowerTrain
0	4.6000	450	233	161	1	2
1	10.0000	270	160	167	1	0
2	4.7000	400	210	181	1	2
3	6.8000	360	180	206	1	0
4	9.5000	170	145	168	1	0
...
98	7.5000	330	160	191	1	1
99	4.5000	335	210	258	1	2
100	5.9000	325	200	194	1	2
101	5.1000	375	200	232	1	2
102	7.5000	400	190	238	1	2

103 rows × 6 columns



Clustering:

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.



Principal Component Analysis

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It 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**. It is one of the popular tools that is used for exploratory data analysis and predictive modelling. It is a technique to draw strong patterns from the given dataset by reducing the

the

```
pca = PCA(n_components=8)
t = pca.fit_transform(x)
data2 = pd.DataFrame(t, columns=['PC1', 'PC2', 'PC3', 'PC4', 'Pc5', 'PC6', 'PC7', 'PC8'])
data2
```

	PC1	PC2	PC3	PC4	Pc5	PC6	PC7	PC8
0	1.5113	0.2120	-1.0143	-0.9438	0.4121	-0.8232	0.3279	-0.3689
1	-1.7406	-0.5828	-0.6710	0.6083	0.3941	0.1555	-0.3973	-0.3224
2	1.2930	0.0209	-0.3737	-0.7301	-0.0892	-0.6444	0.2832	-0.0881
3	0.0213	-0.1154	-0.0850	1.5579	0.1865	0.2337	-0.4378	0.1833
4	-2.3280	0.2449	-0.7962	0.5904	-0.7448	0.2406	-0.5019	-0.1635
...
98	-0.3383	-0.4627	-0.0185	0.1724	-0.1474	0.0289	0.1420	0.1329
99	2.2795	0.2302	1.6805	0.3698	-1.0874	0.1321	-0.3490	0.1014
100	0.8151	-0.1643	0.1789	-0.7220	-0.4616	-0.4929	0.0801	-0.2633
101	1.6176	-0.0897	1.0125	0.0003	-0.7315	-0.2048	0.2896	-0.0988
102	1.2771	-0.2213	1.3252	-0.0193	-0.5180	0.3873	0.5720	-0.3703

103 rows × 8 columns

variances.

Conclusion 1:

- There are two columns namely price and top speed that is highly correlated as 0.83. It is obvious that when we buy a car of high price, we get many features and top speed is one of them.
- Byton, Smart, Fiat have highest number of vehicles in the market while Polestar and Lucid have lowest.
- Porsche, Lucid and Tesla produce fastest cars while smart slowest.
- Byton, Jaguar, and Audi are the efficient brands while Lightyear has lowest in efficiency.
- Most of the cars have 5 seats.
- Porsche and Lightyear produces expensive cars while SEAT and Smart produces cheap car.
- Most of the cars have Type 2 CCS plug charger.
- Most cars are either SUV or Hatchbacks.

2. Behavioral Segmentation of EV Market

Behavioural segmentation refers to a process in marketing which divides customers into segments depending on their behaviour patterns when interacting with a particular business or website.

These segments could include grouping customers by:

- Their attitude toward your product, brand or service;
- Their use of your product or service,
- Their overall knowledge of your brand and your brand's products,
- Their purchasing tendencies, such as buying on special occasions like birthdays or holidays only, etc.

Going beyond the traditional demographic and geographic segmentation methods and utilizing behavioural data allows for the execution of more successful marketing campaigns.

At the very least, behavioural segmentation offers marketers and business owners a more complete understanding of their audience, thus enabling them to tailor products or services to specific customer needs. Below we take a look at four more benefits of behavioural segmentation.

We are using a dataset. By analysing dataset, we predict the customers behaviour towards buying EV vehicle.

First of all, we read the dataset using panda library and display the first 5 rows so that we can know about the number of columns and their characteristics.

```
# This data contains the details about consumers who purchased an EV
df = pd.read_csv("/content/drive/MyDrive/Datasets/Electric_vehicle _dataset/behavioural_dataset.csv")
```

```
df.head()
```

	Age	Profession	Marrital Status	Education	No of Dependents	Personal loan	Total Salary	Price
0	27	Salaried	Single	Post Graduate	0	Yes	800000	800000
1	35	Salaried	Married	Post Graduate	2	Yes	2000000	1000000
2	45	Business	Married	Graduate	4	Yes	1800000	1200000
3	41	Business	Married	Post Graduate	3	No	2200000	1200000
4	31	Salaried	Married	Post Graduate	2	Yes	2600000	1600000

In this dataset, we observe that there are 8 columns namely Age, Profession, Marital Status, Education, No of Dependents, Personal loan, Total Salary and Price.

Description of the dataset

```
df.describe()
```

	Age	No of Dependents	Total Salary	Price
count	99.000000	99.000000	9.900000e+01	9.900000e+01
mean	36.313131	2.181818	2.270707e+06	1.194040e+06
std	6.246054	1.335265	1.050777e+06	4.376955e+05
min	26.000000	0.000000	2.000000e+05	1.100000e+05
25%	31.000000	2.000000	1.550000e+06	8.000000e+05
50%	36.000000	2.000000	2.100000e+06	1.200000e+06
75%	41.000000	3.000000	2.700000e+06	1.500000e+06
max	51.000000	4.000000	5.200000e+06	3.000000e+06

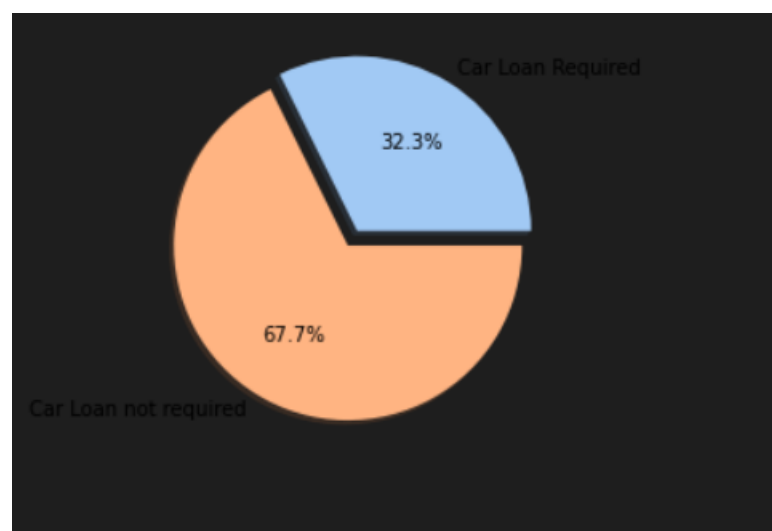
Here we observe that maximum age of the person who bought EV car is 51 and mean of such ages is 36.31 years. Maximum salary of the person who bought the car is 5200000 while minimum salary is 2000000 and the average of the salary is 2270707.

Checking the marital status of the person with car loan



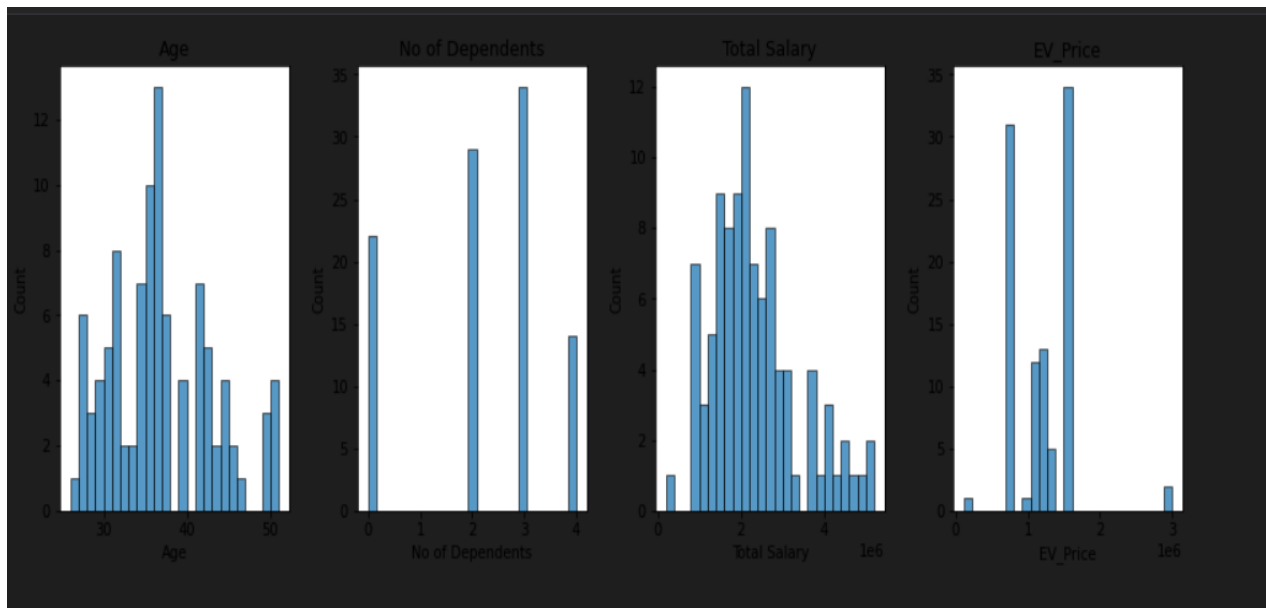
By seeing the plot , It is clear that there are more persons(single or married) do not need car loan to buy the car.

Car Loan



67.7% of the total person who bought EV car do not need car loan.

Plotting the frequency of customers against Age, No of Dependents, Total Salary and EV_price.



Persons having the age between 30 and 40 has bought most of cars.

Person having 4 family members are less attractive towards buying cars.

Persons who have medium level salary bought Ev cars more.

EV cars having medium price range are sold a lot.

KModes clustering

KModes clustering is one of the unsupervised Machine Learning algorithms that is used to cluster **categorical variables**.

KMeans uses mathematical measures (distance) to cluster continuous data. The lesser the distance, the more similar our data points are. Centroids are updated by Means. But for categorical data points, we cannot calculate the distance. So we go for KModes algorithm. It uses the dissimilarities (total mismatches) between the data points. The lesser the dissimilarities the more similar our data points are. It uses Modes instead of means.

In K-means clustering when we used categorical data after converting it into a numerical form. it doesn't give a good result for high-dimensional data.

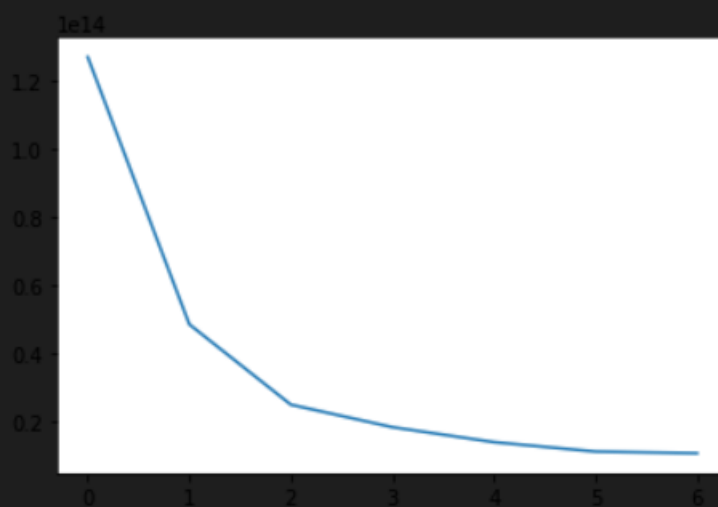
So, Some changes are made for categorical data.

- Replace Euclidean distance with Dissimilarity metric
- Replace Mean by Mode for cluster centres.
- Apply a frequency-based method in each iteration to update the mode.

Finding optimal number of clusters for KPrototypes

```
cost = []
for num_clusters in list(range(1,8)):
    kproto = KPrototypes(n_clusters=num_clusters, init='Cao')
    kproto.fit_predict(cluster_data, categorical=[1,2,3,5])
    cost.append(kproto.cost_)

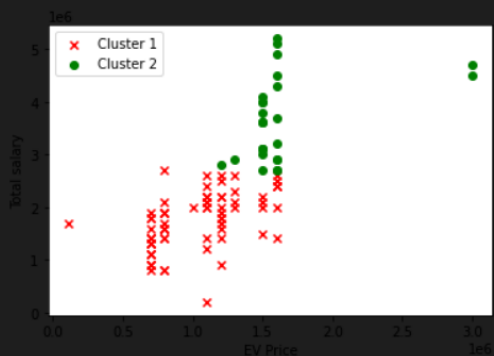
plt.plot(cost)
plt.show()
```



There is bending in the line at point 2. Hence number of clusters formed is 2 for better visualizations.

```
plt.scatter(Cluster_0.EV_Price, Cluster_0['Total Salary'],color='red', marker = 'x', label = 'Cluster 1')
plt.scatter(Cluster_1.EV_Price, Cluster_1['Total Salary'],color='green', label = 'Cluster 2')
plt.legend(loc="upper left")

plt.xlabel('EV Price')
plt.ylabel('Total salary')
plt.show()
```



There is clear difference in the segments in comparing salary and price of Ev cars.

Conclusions 2

- The maximum age of the person who bought EV car is 51 year and mean of such ages is 36.31 years. Maximum salary of the person who bought the car is 5200000 while minimum salary is 2000000 and the average of the salary is 2270707.
- there are more persons (single or married) do not need car loan to buy the car.
- 67.7% of the total person who bought EV car do not need car loan.
- Persons having the age between 30 and 40 has bought most of cars.
- Person having 4 family members are less attractive towards buying cars.
- Persons who have medium level salary bought Ev cars more.
- EV cars having medium price range are sold a lot.

3.Geographical Segmentation of EV market

Geographic segmentation is a marketing strategy used to target products or services at people who live in, or shop at, a particular location. It works on the principle that people in that location have similar needs, wants, and cultural considerations. By understanding what people in that area require, brands can target more relevant marketing messages and suitable products to customers who are then aware and more likely to buy.

We have dataset and we read this through pandas library and know about number of charging stations present in the different parts of India.

```
cs_highways = pd.read_csv('/content/drive/MyDrive/Datasets/Electric_vehicle _dataset/Geographical Segmentation/Datasets/CS_Highway.csv')
```

cs_highways

Sl. No.	Highways/Expressways	Charging Stations
0	1 Mumbai - Pune Expressway	10
1	2 Surat-Mumbai Expressway	30
2	3 Mumbai - Delhi Highway	124
3	4 Mumbai - Panaji Highway	60
4	5 Mumbai - Nagpur Highway	70
5	6 Mumbai - Bengaluru Highway	100
6	7 Agra-Nagpur	80
7	8 Kolkata- Nagpur	120
8	9 Chennai- Nagpur	114
9	Total	Total 708

Mumbai-Delhi Highway has highest number of charging stations followed by Kolkata-Nagpur Highway.

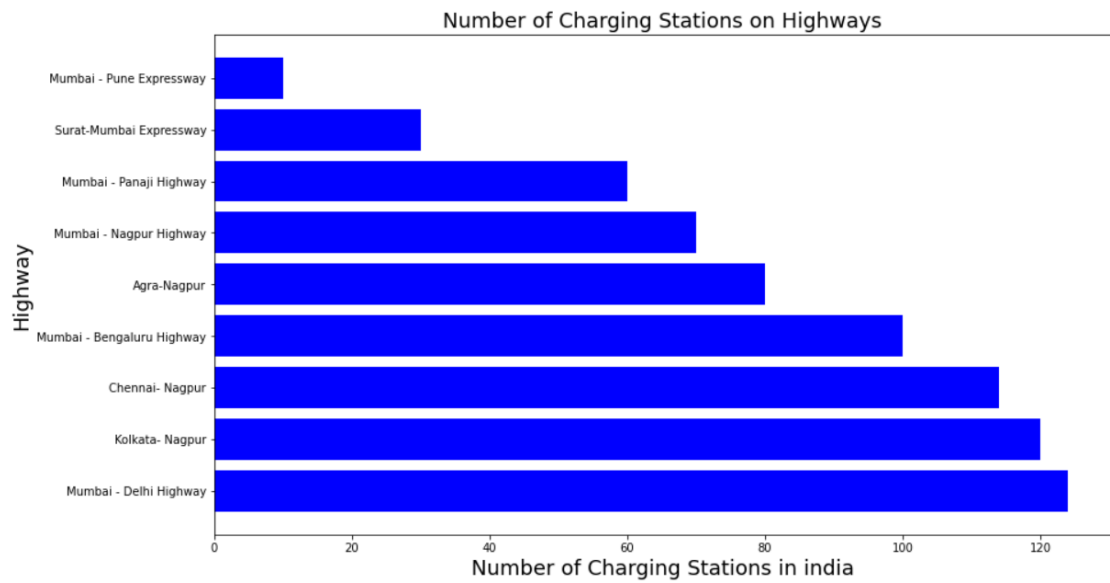
Mumbai-Pune Expressway has least number of charging stations.

Number of electric vehicle charging stations sanctioned in different states and union territories

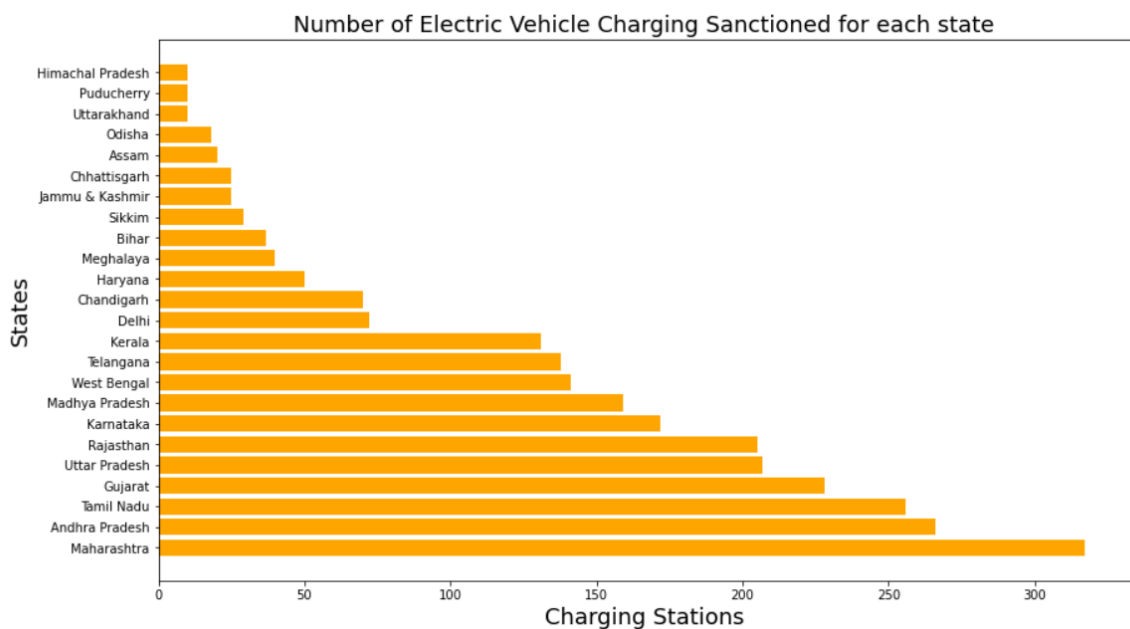
	State/UT-wise	Number of Electric Vehicle Charging Sanctioned
0	Maharashtra	317
1	Andhra Pradesh	266
2	Tamil Nadu	256
3	Gujarat	228
4	Uttar Pradesh	207
5	Rajasthan	205
6	Karnataka	172
7	Madhya Pradesh	159
8	West Bengal	141
9	Telangana	138
10	Kerala	131
11	Delhi	72
12	Chandigarh	70
13	Haryana	50
14	Meghalaya	40
15	Bihar	37
16	Sikkim	29
17	Jammu & Kashmir	25
18	Chhattisgarh	25
19	Assam	20
20	Odisha	18
21	Uttarakhand	10
22	Puducherry	10
23	Himachal Pradesh	10
24	Total	2636

Maharashtra has highest number of charging stations sanctioned followed by Andhra Pradesh and total sanctioned all over India is 2636.

Number of charging stations on highway



```
plot_frequency(cs_sanctioned,'State/UT-wise','Number of Electric Vehicle  
Charging Sanctioned','Charging Stations','States',  
               'Number of Electric Vehicle Charging Sanctioned for each  
state','orange')
```



Maharashtra has the highest number of charging stations sanctioned.

States to target based on charging stations sanctioned

Maharashtra

Andhra Pradesh

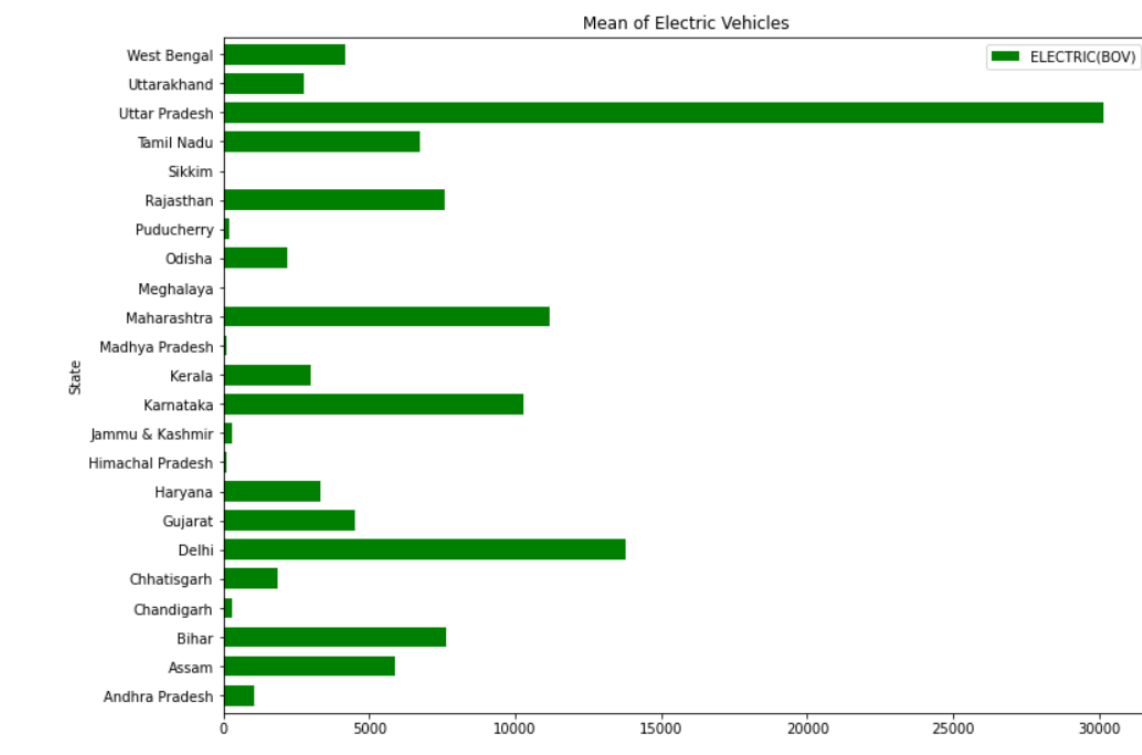
Tamilnadu

Gujarat

Uttar Pradesh

Rajasthan

Average number of electric vehicles in different states

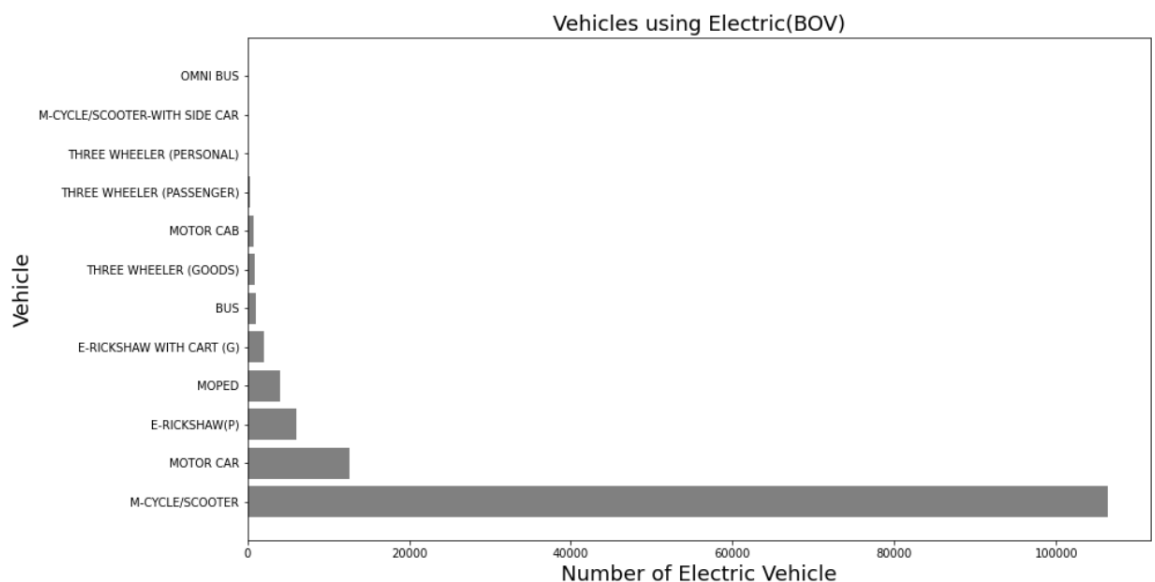
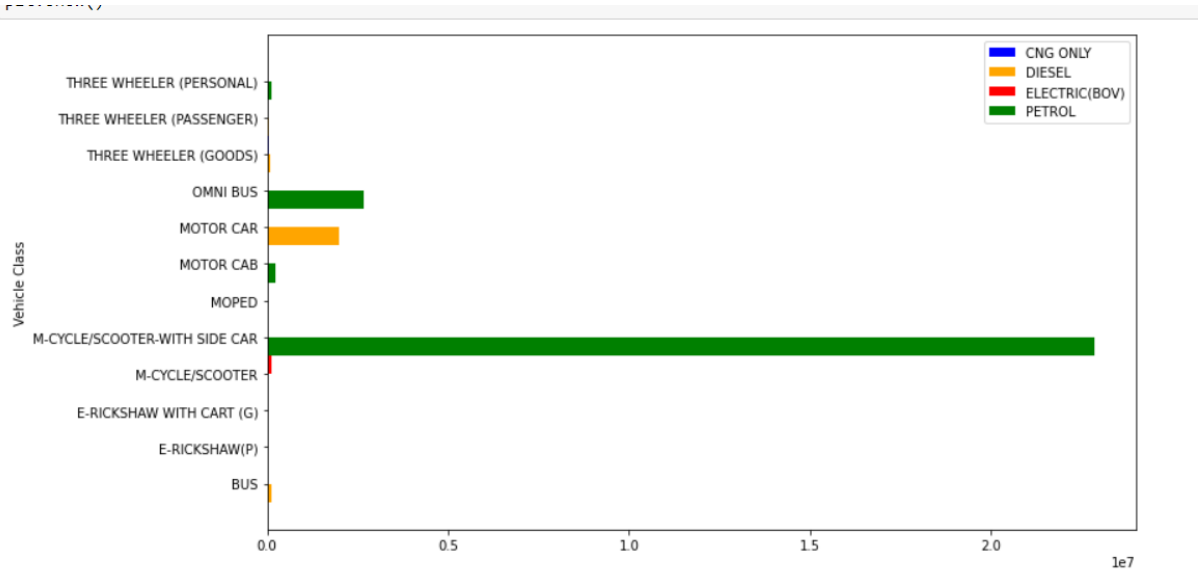


Uttar Pradesh has the highest number of electric vehicles in India followed by Delhi, Maharashtra, Karnataka, Rajasthan, Bihar.

Sikkim and Himachal Pradesh has lowest number of electric vehicles in India.

Visualizing the type of fuel used for every vehicle class & which vehicle space uses Battery operated fuel (EV)

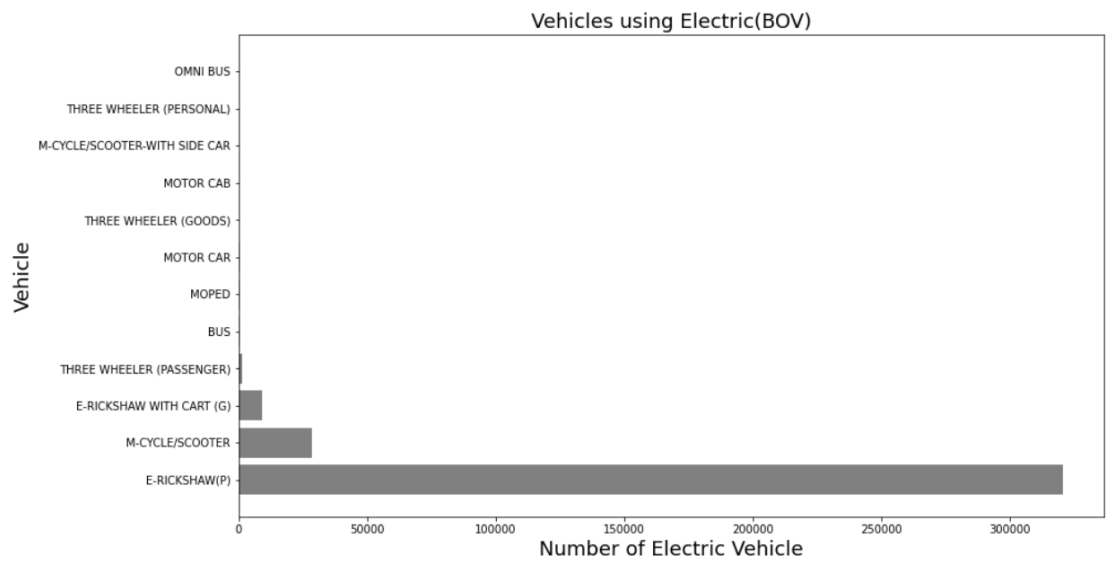
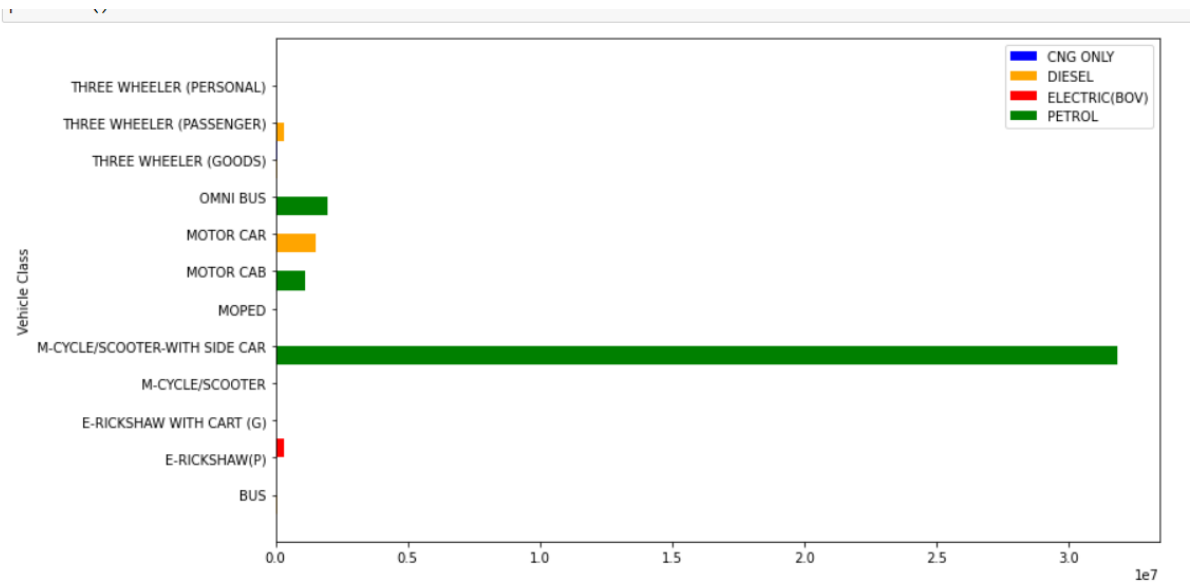
1. Maharashtra



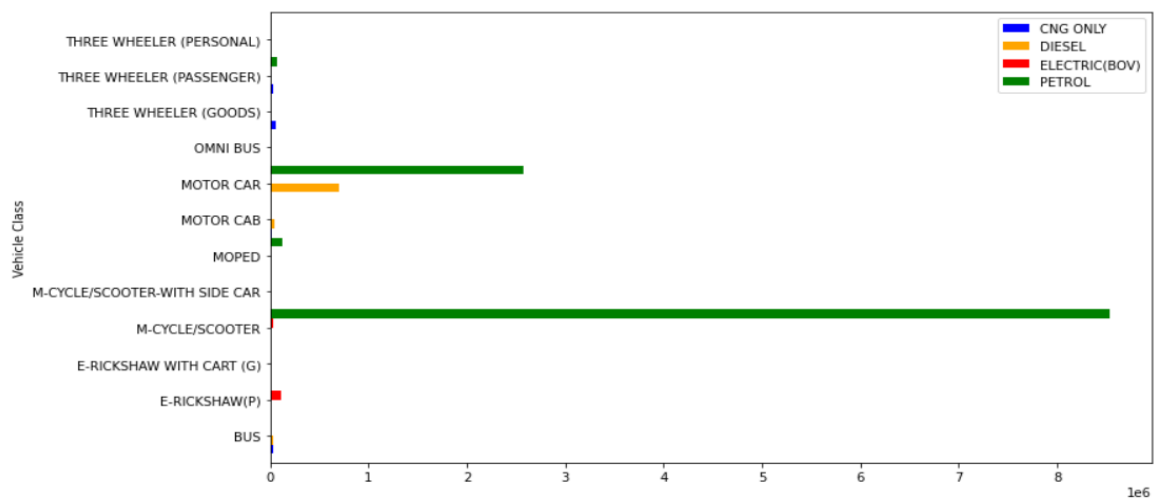
In Maharashtra, very less number of vehicles use Battery operated vehicle out of which M-Cycle/Scooter has largest share on the road. So we should focus on M-cycle/Scooter in Maharashtra.

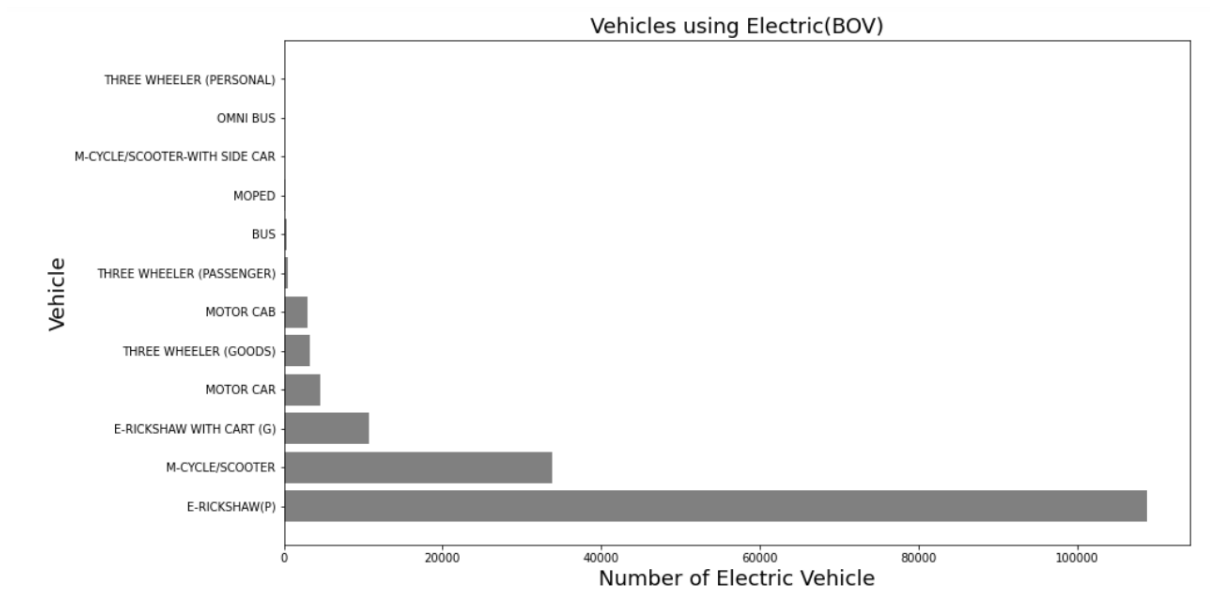
2. Uttar Pradesh

We should focus on E-rickshaw market in UP as it is highest number and also to look for other opportunities such as E-Motorcycle/Scooter.

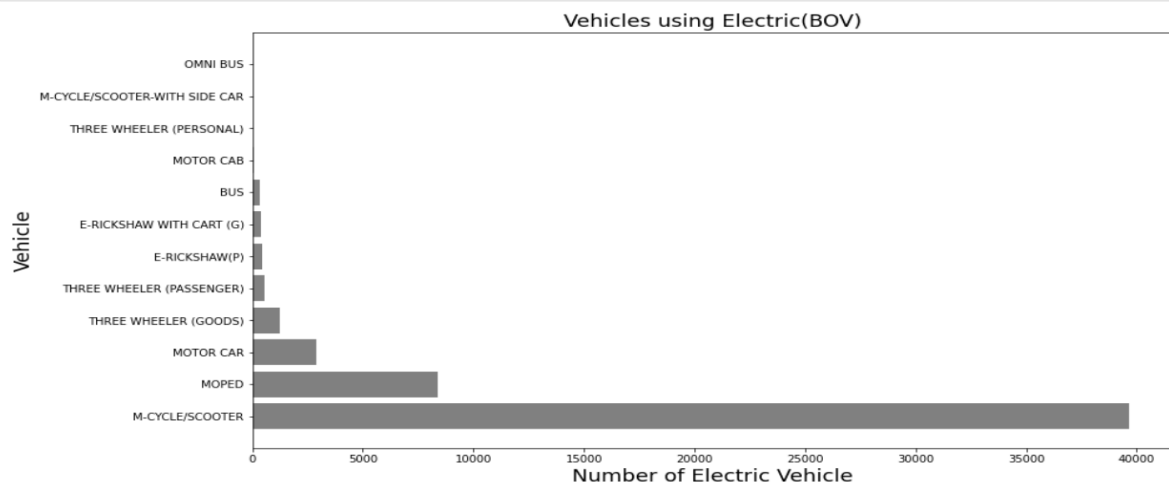
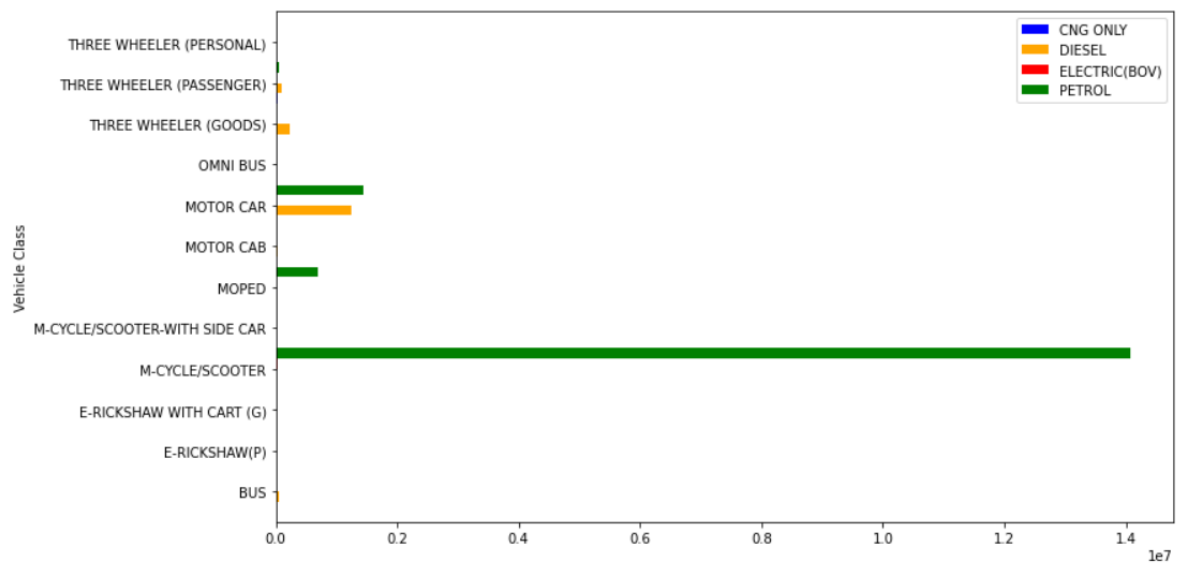


3 Delhi



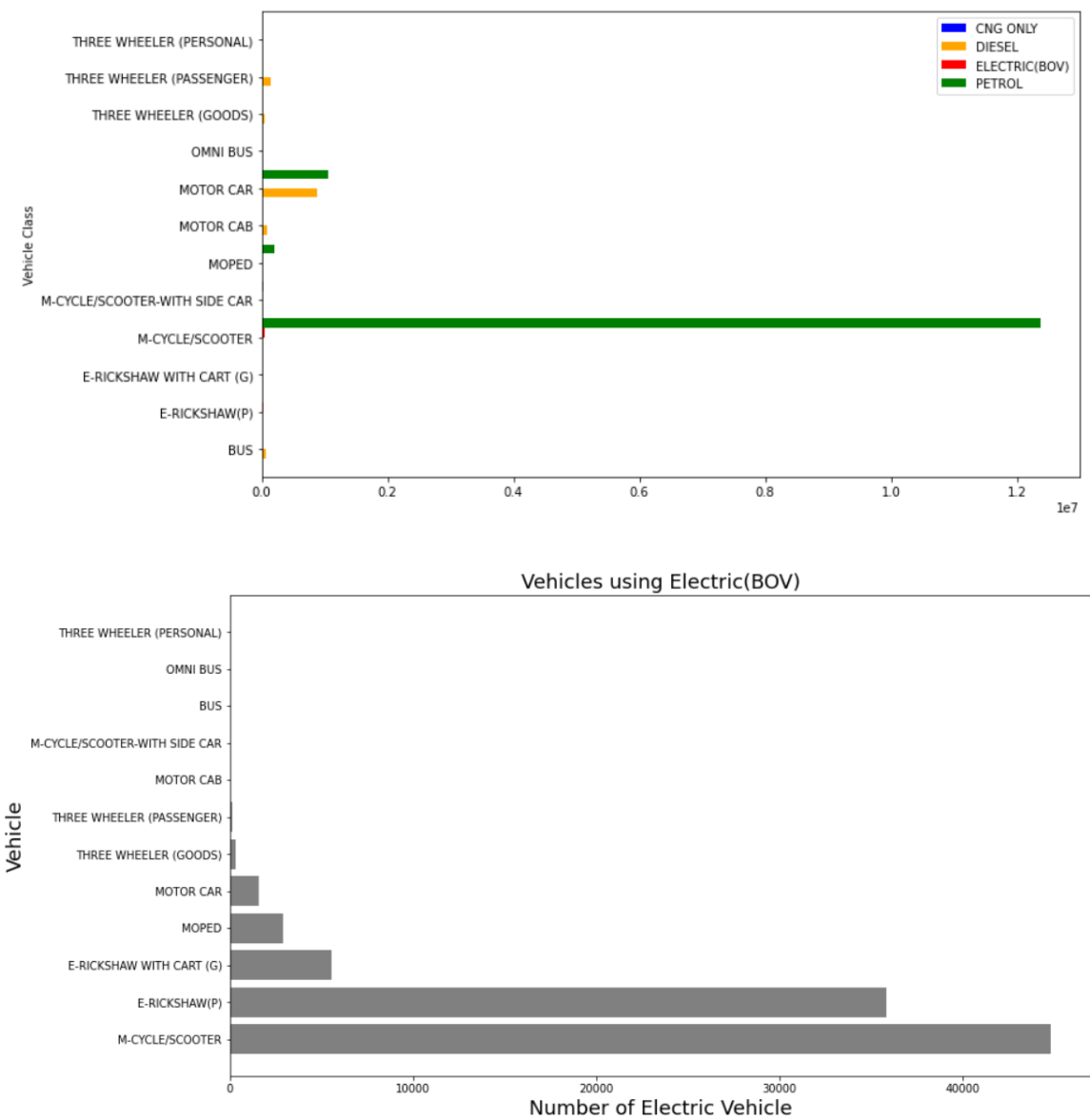


4 Gujarat



In Gujarat, M-Cycle/Scooter is in highest number among all electric vehicles followed by MOPED.

5 Rajasthan

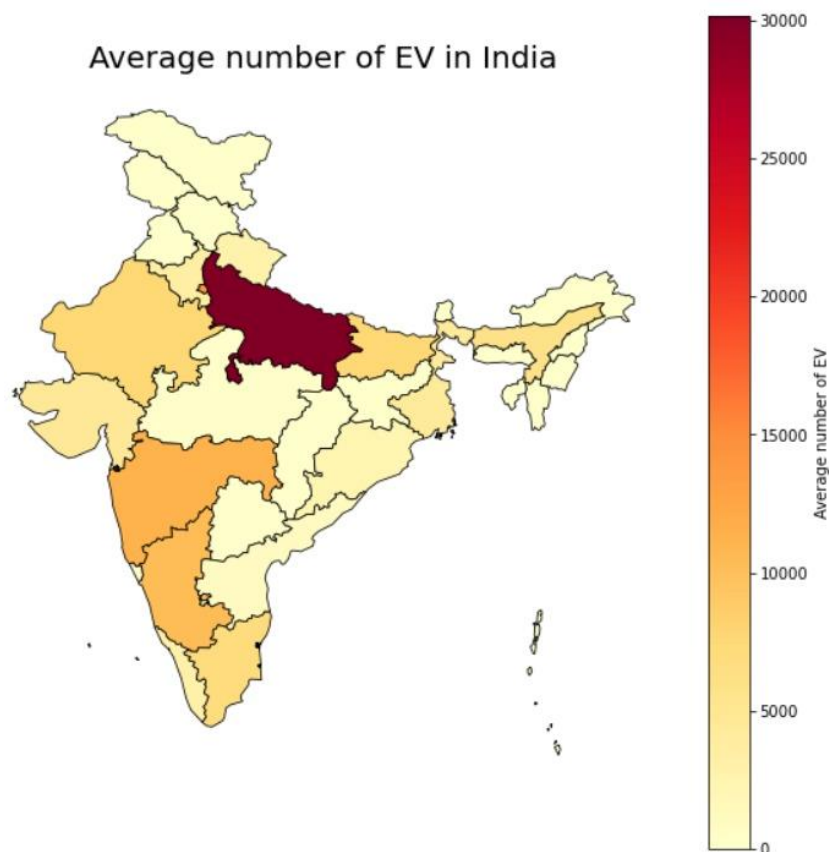


In Rajasthan, M-Cycle/Scooter is in highest number among all electric vehicles followed by E-Rickshaw(P).

Conclusion 3:

- Mumbai-Delhi Highway has highest number of charging stations followed by Kolkata-Nagpur Highway.
- Mumbai-Pune Expressway has least number of charging stations.

- Maharashtra has the highest number of charging stations sanctioned.
- Uttar Pradesh has the highest number of electric vehicles in India followed by Delhi, Maharashtra, Karnataka, Rajasthan, Bihar.
- Sikkim and Himachal Pradesh have lowest number of electric vehicles in India.
- In Maharashtra, very less number of vehicles use Battery operated vehicle out of which M-Cycle/Scooter has largest share on the road. So, we should focus on M-cycle/Scooter in Maharashtra.
- We should focus on E-rickshaw market in UP as it is highest number and also to look for other opportunities such as E-Motorcycle/Scooter.
- We should focus on E-rickshaw(P) market in Delhi as it is highest number and also to look for other opportunities such as E-Motorcycle/Scooter.
- In Gujarat, M-Cycle/Scooter is in highest number among all electric vehicles followed by MOPED.
- In Rajasthan, M-Cycle/Scooter is in highest number among all electric vehicles followed by E-Rickshaw(P).



Distribution of Electric Vehicles Across India

From above map, we see that Uttar Pradesh has highest number of EVs followed by Maharashtra. So, we should focus on these areas for our EVs

Now here is the state wise list of type of electric vehicles used in the different states of India.

Vehicle Space to target based on State

Sr.	State	Vehicle Type (Based on Rank)
1	Maharashtra	I. Cycle / Scotter II. Car III. E - Rickshaw IV. Moped
2	Uttar Pradesh	I. E - Rickshaw II. Cycle / Scotter III. 3 Wheeler (Passenger)
3	Delhi	I. E - Rickshaw II. Cycle / Scotter III. Car IV. 3 Wheeler (Goods)
4	Karnataka	I. Cycle / Scotter II. 3 Wheeler (Passenger) III. Car IV. Moped
5	Gujarat	I. Cycle / Scotter II. Moped III. Car IV. 3 Wheeler (Goods)
6	Rajasthan	I. Cycle / Scotter II. E - Rickshaw III. Moped IV. Car

Name	Github Link
Naveen Kumar	https://github.com/Naveenkumar900/EV-Market-Segmentation
Madhur Sharma	https://github.com/Madhursharma98/EVs-Market-Segmentation
Roshan	https://github.com/roshananduri/EV_Market_Segmentation

End