Data pre-processing

Required libraries

In order to perform EDA and clustering on the collected data, the following Python libraries areused:

1. Pandas: for data handling/manipulation
2. Matplotlib and Seaborn: for data visualization
3. Scikit-learn: for the k-means clustering algorithm and some other algorithms

---

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

Pulling the datasets

[Dataset 1]

In [2]:
```
df1 = pd.read_csv(r'C:\Users\Milan\Downloads\ev_charger_dataset.csv')
```

In [3]:
```
df1
```

Out[3]:

	Region	2W	3W	4W	Bus	Chargers
0	Uttar Pradesh	9852	42881	458	197	207
1	Maharastra	38558	893	1895	186	317
2	Karnataka	32844	568	589	57	172
3	Tamil Nadu	25642	396	426	0	256
4	Gujarat	22359	254	423	22	228
5	Delhi	11756	5287	1578	186	72
6	Bihar	2388	10783	89	36	37
7	Assam	357	11547	42	0	20
8	Kerala	10345	308	578	0	131
9	Odisha	9540	253	89	0	18
10	Andhra Pradesh	14578	2587	524	0	266

Dataset 2

```
In [4]: df2 = pd.read_excel(r'C:\Users\Milan\Downloads\ev_charging_station_dataset.xlsx',sheet_name='Table 4',
```

```
In [5]: df2
```

Out[5]:

	State/UT	EV Charging Facility
0	Andhra Pradesh	65
1	Arunachal Pradesh	4
2	Assam	19
3	Bihar	26
4	Chandigarh	4
5	Chhattisgarh	51
6	Delhi	66
7	Goa	17
8	Gujarat	87
9	Haryana	114
10	Himachal Pradesh	13

Dataset 3

```
In [6]: df3 = pd.read_excel(r'C:\Users\Milan\Downloads\ev_market_india_dataset.xlsx')
```

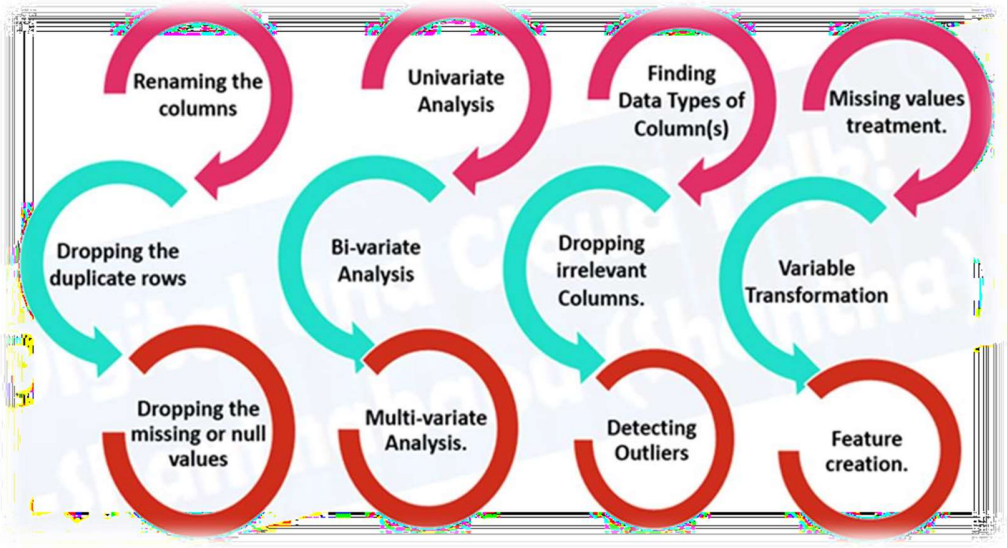
```
In [7]: df3
```

Out[7]:

	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	Power
0	Tesla	Model 3 Long Range Dual Motor	4.6	233	450	161	940	Yes	.
1	Volkswagen	ID.3 Pure	10.0	160	270	167	250	No	I
2	Polestar	2	4.7	210	400	181	620	Yes	.
3	BMW	iX3	6.8	180	360	206	560	Yes	I
4	Honda	e	9.5	145	170	168	190	Yes	I
...	...	...	...	...	...	...	...	...	
98	Nissan	Ariya 63kWh	7.5	160	330	191	440	Yes	
99	Audi	e-tron S Sportback	4.5	210	335	258	540	Yes	

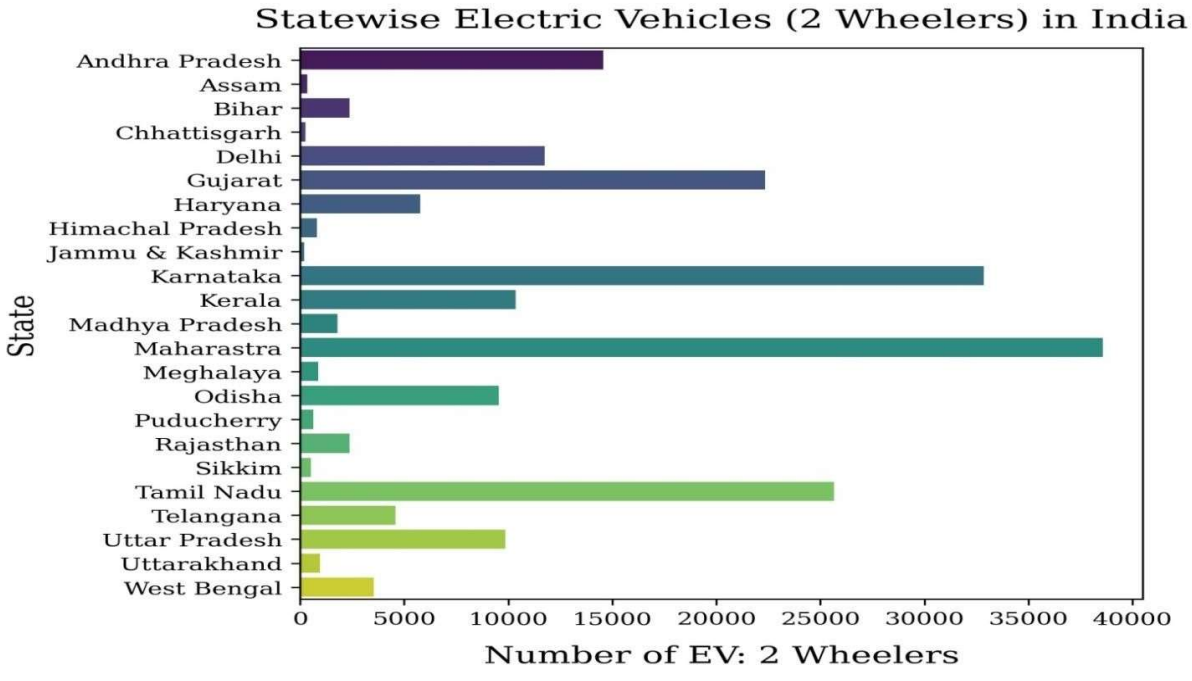
Exploratory Data Analysis

Exploratory Data Analysis, popularly abbreviated as EDA, is one of the most important steps inthe data science pipeline. It is the process of gaining the information present inside the data withthe help of summary statistics and visual representations. Keys features of this technique are presented in the below image.

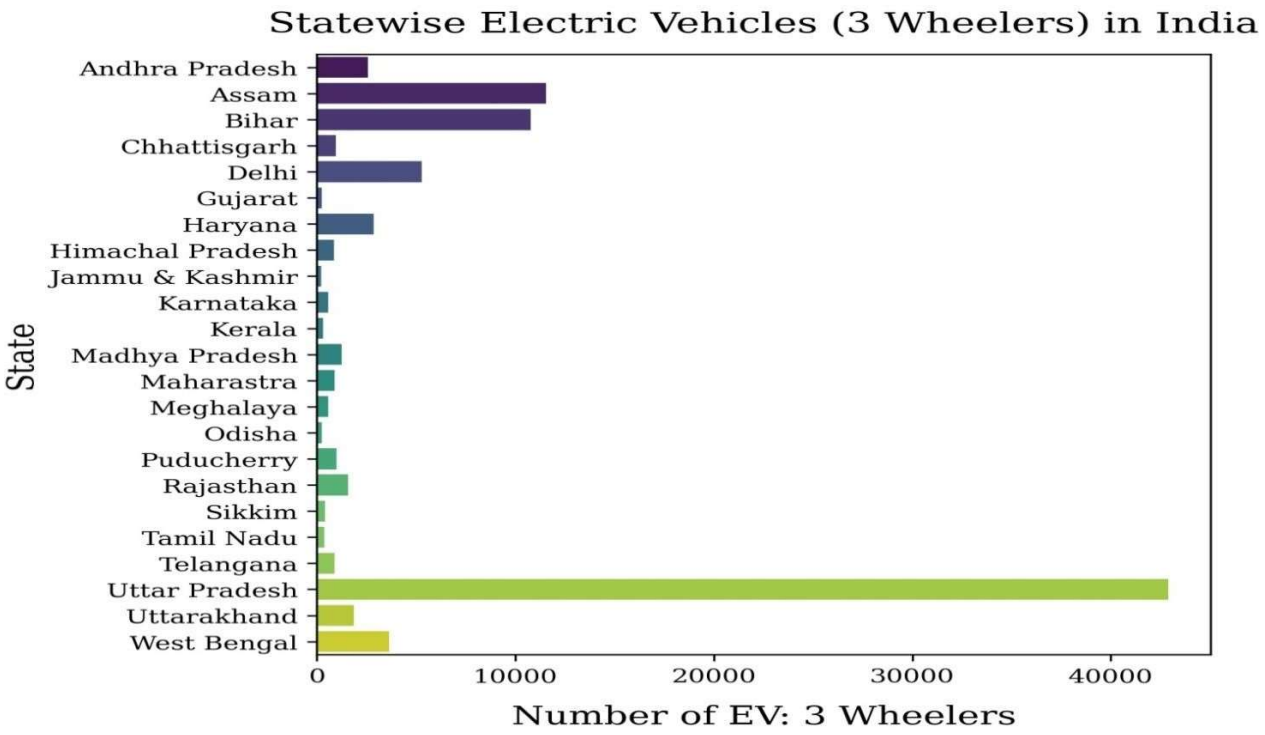


Implementing EDA on the datasets

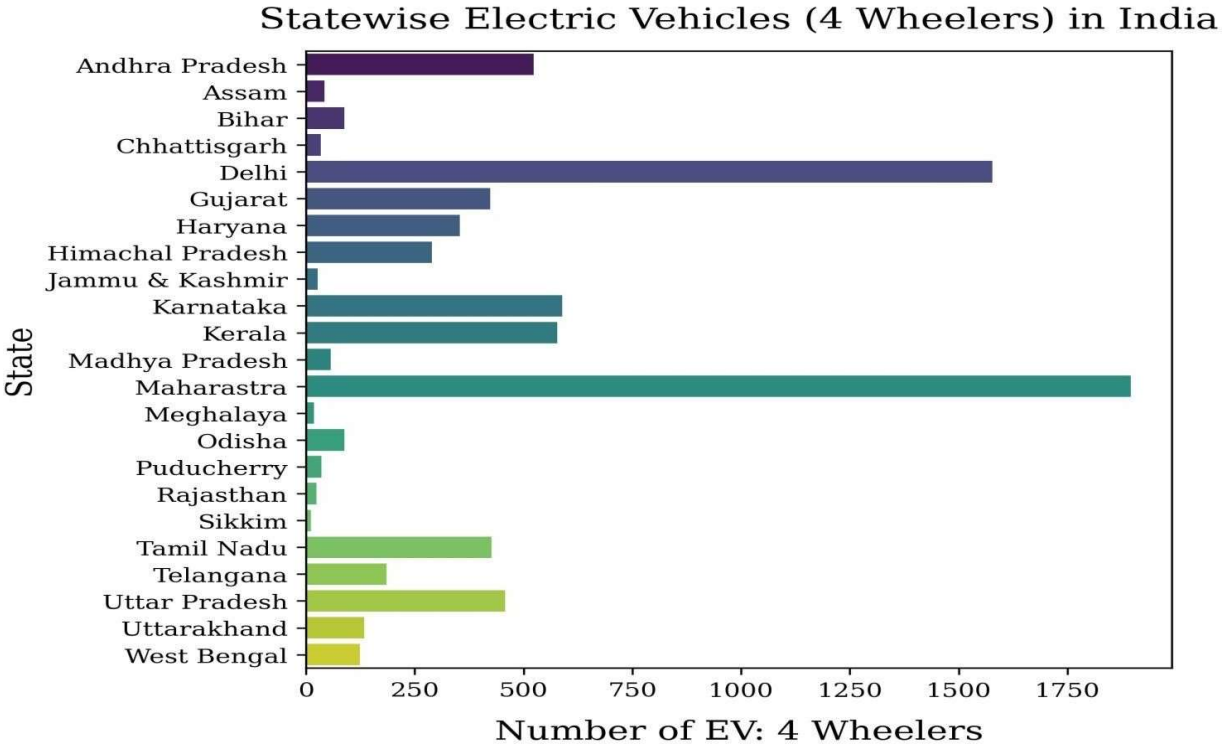
Number of 2-wheeler EVs in India



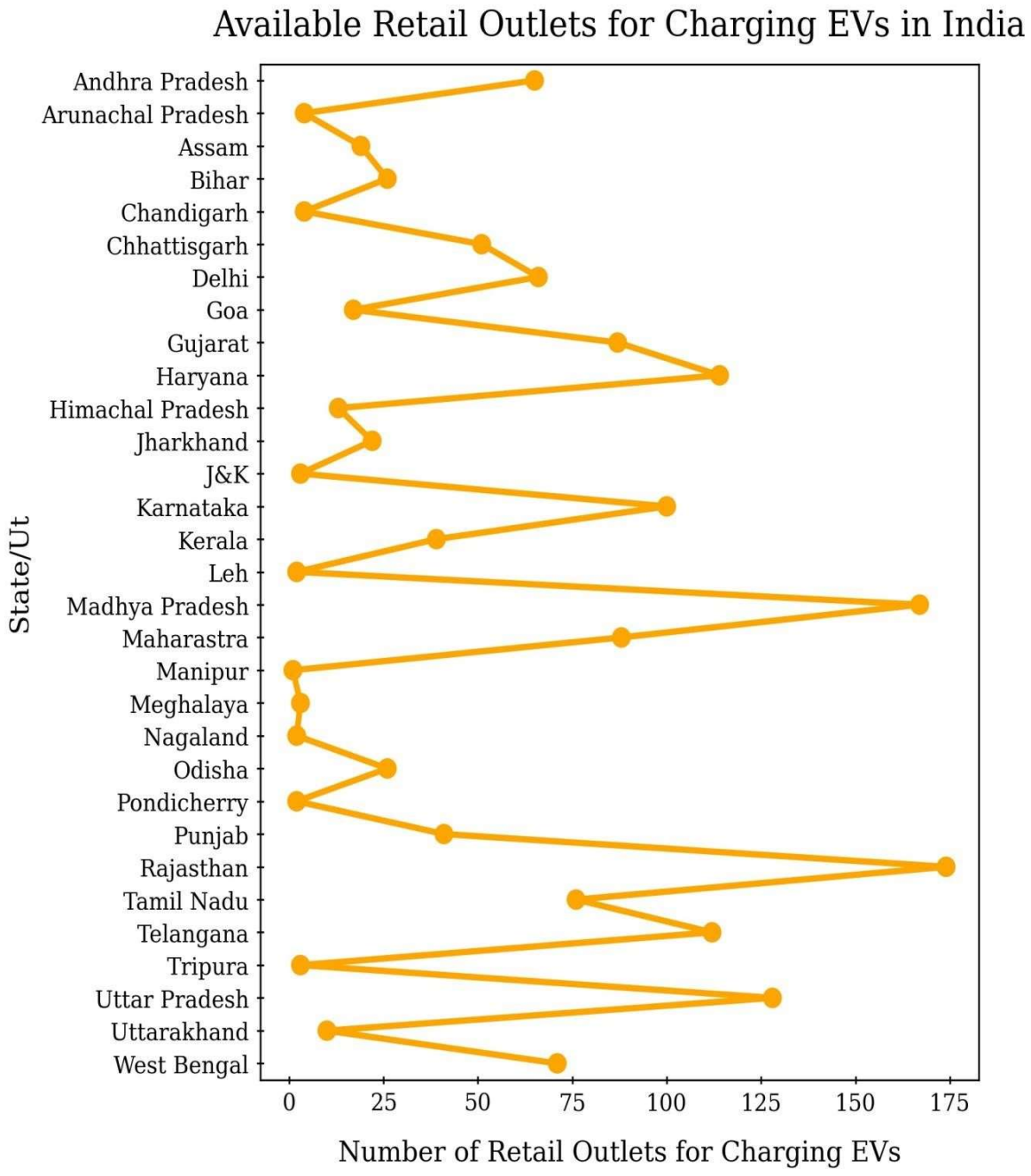
Number of 3-wheeler EVs in India



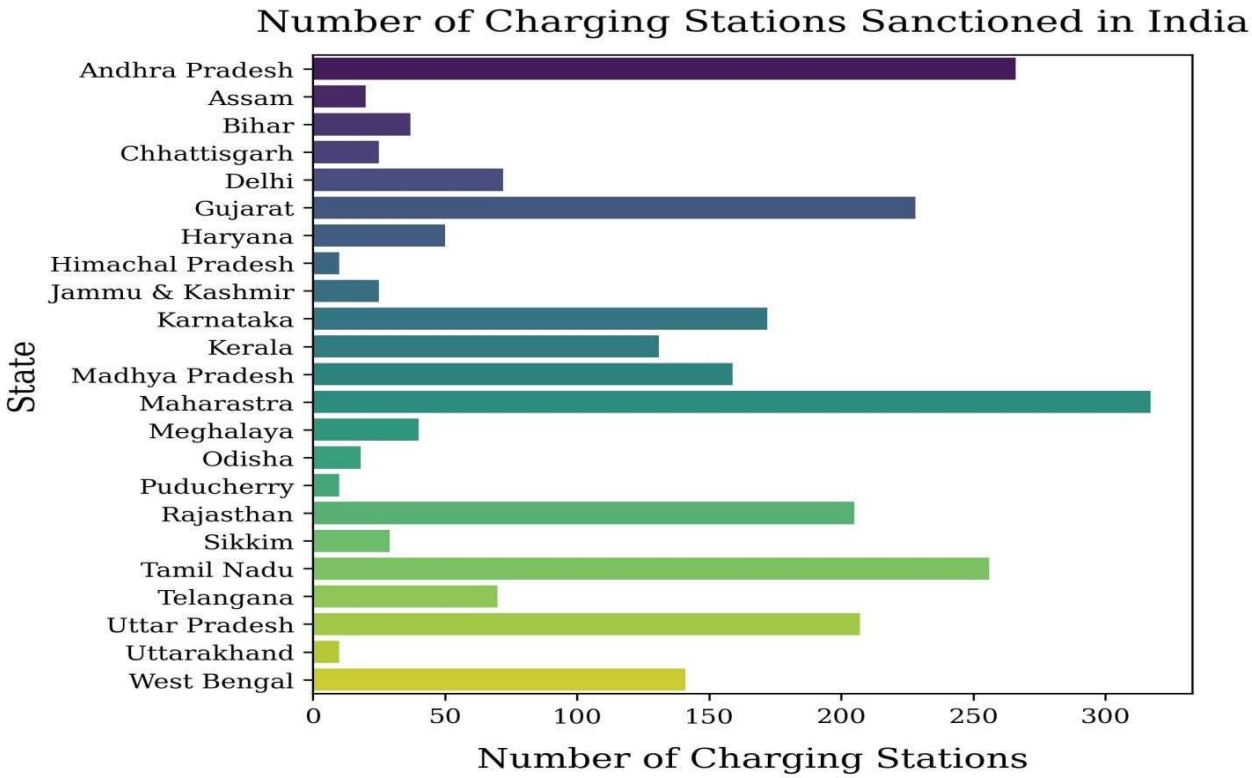
Number of 4-wheeler EVs in India



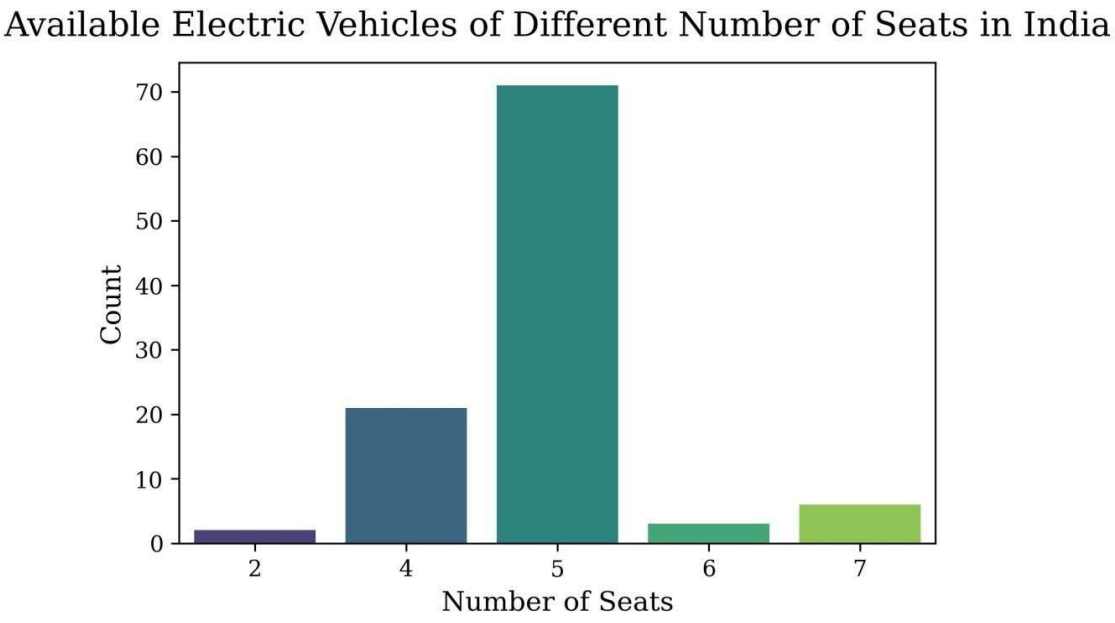
Retail outlets in India for charging EVs



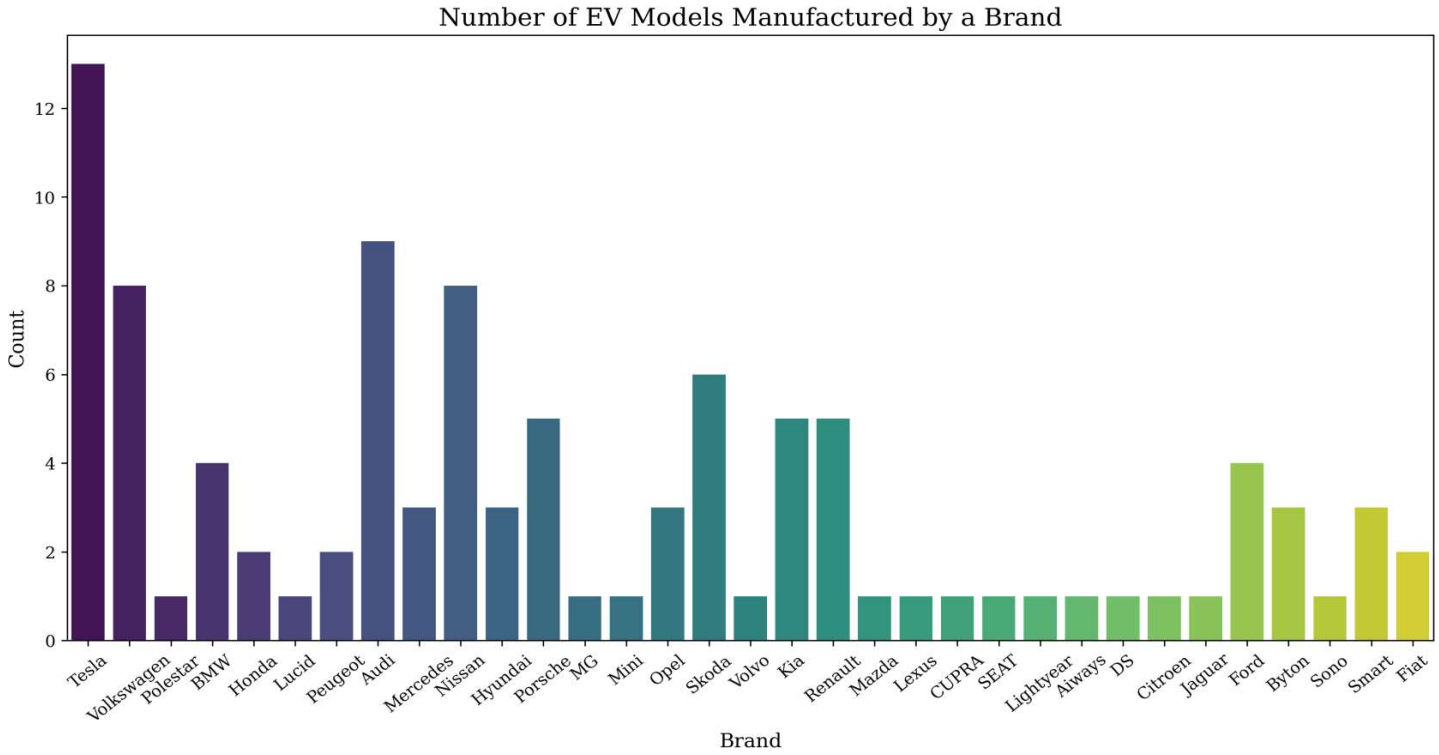
Number of charging stations sanctioned by Government of India



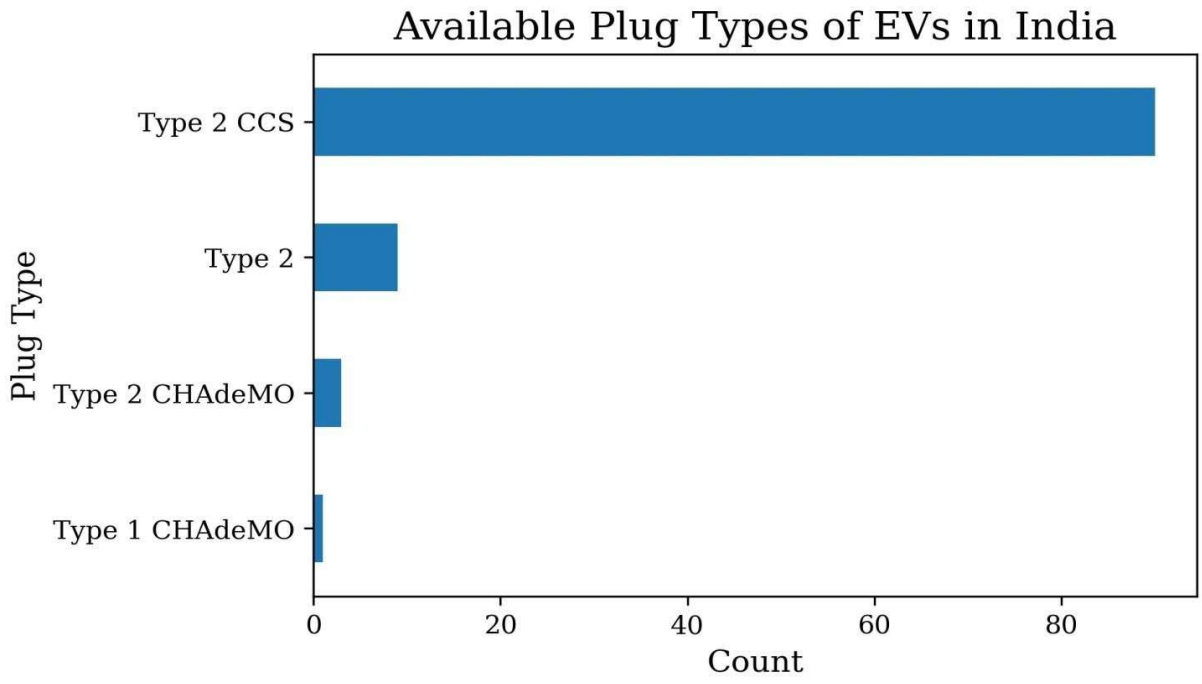
Choices for the number of seats for EVs in India



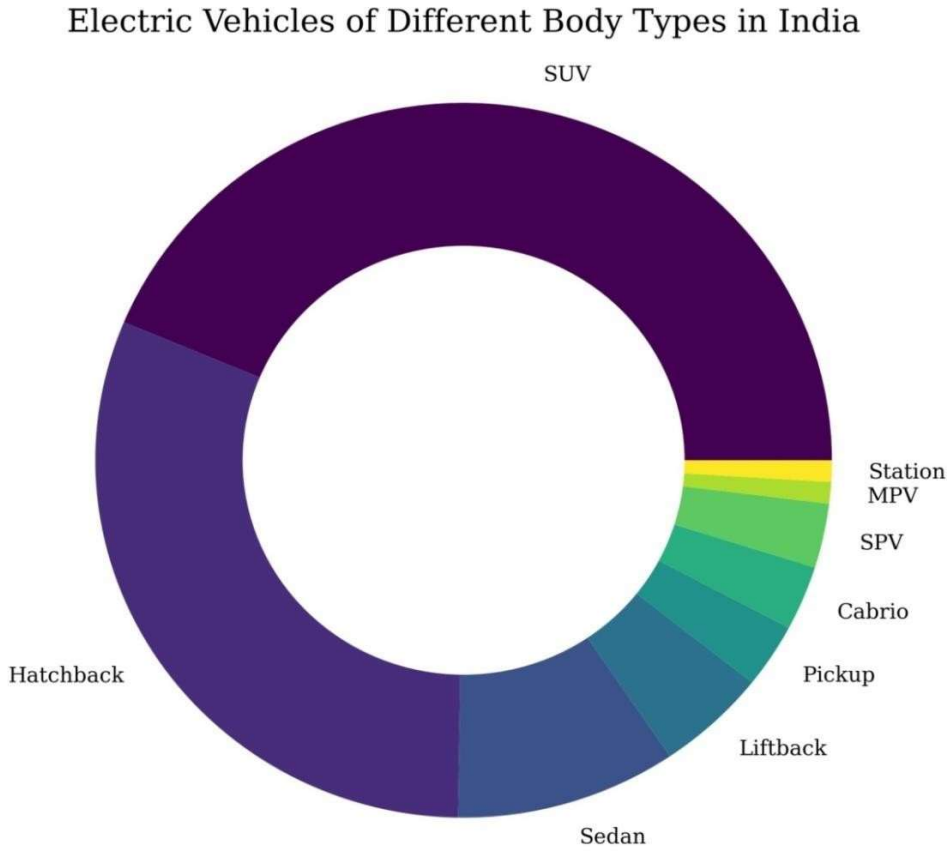
Top EV manufacturing brands in India



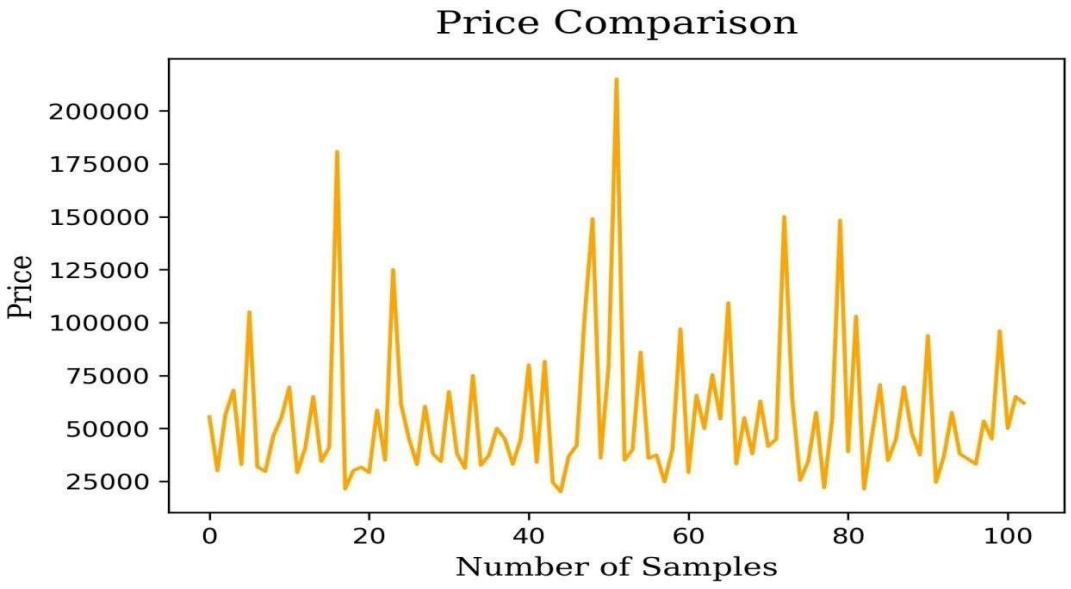
Types of EV plugs available in India



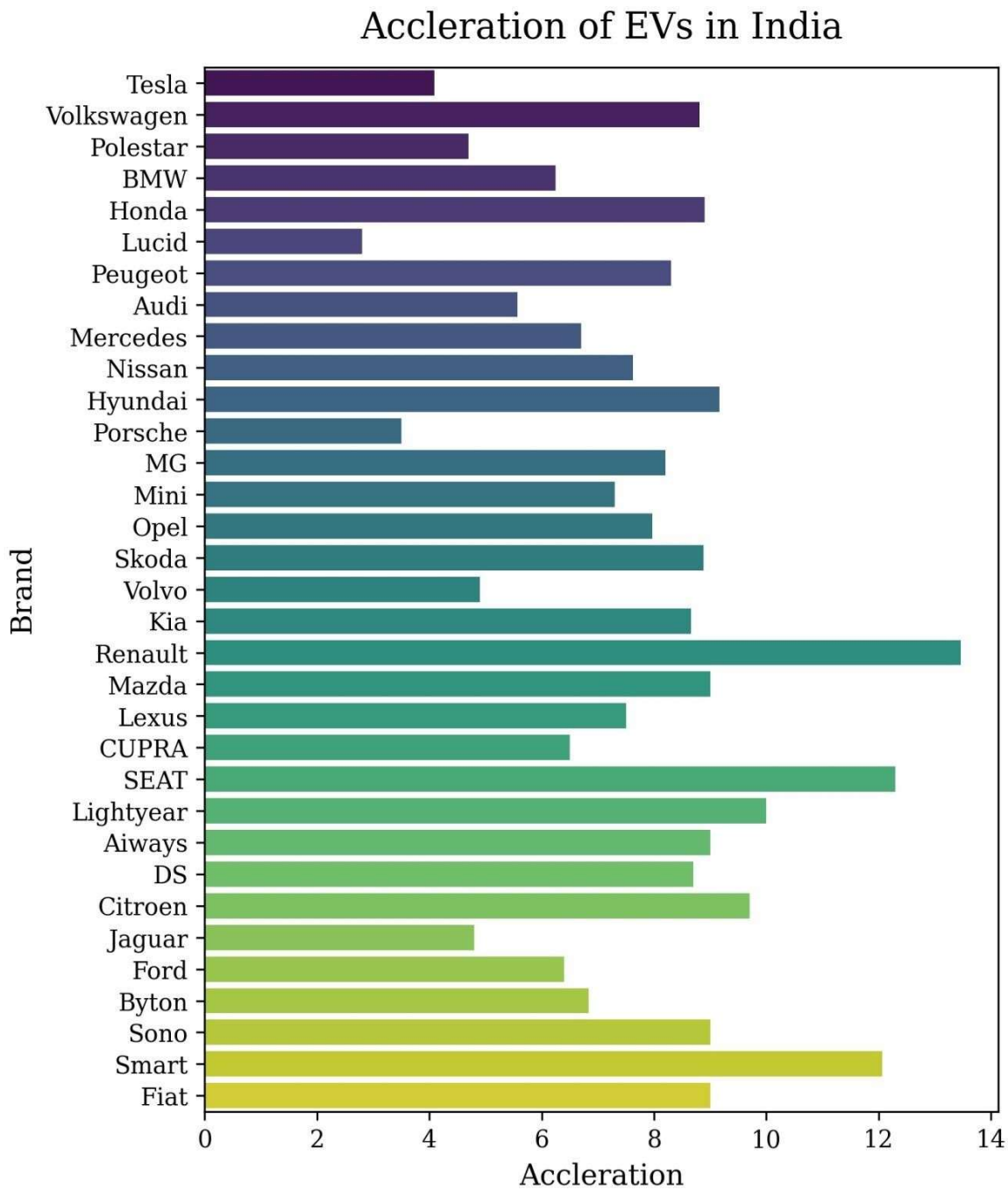
Body types of EVs in India



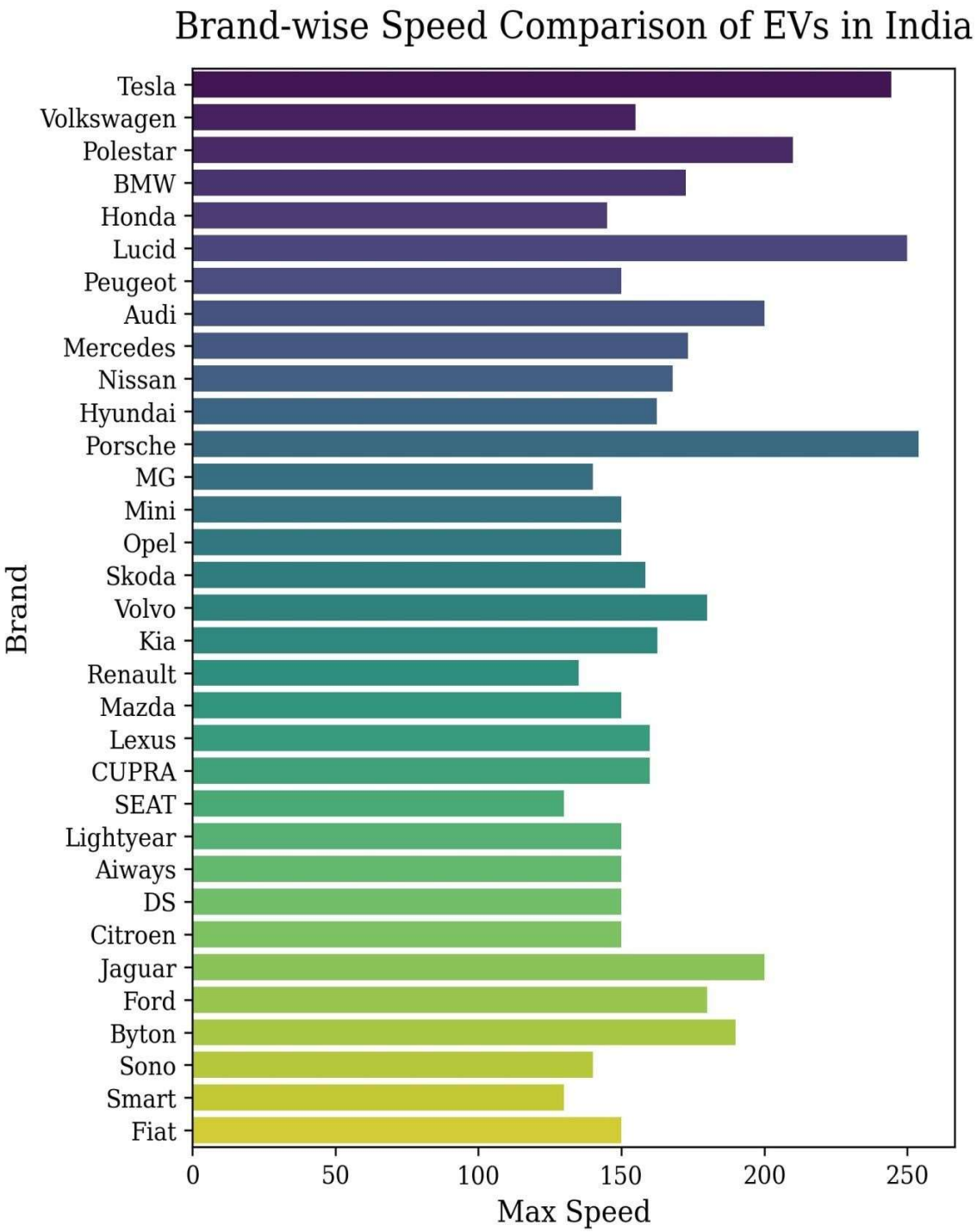
Price comparison of different brands of EVs in India



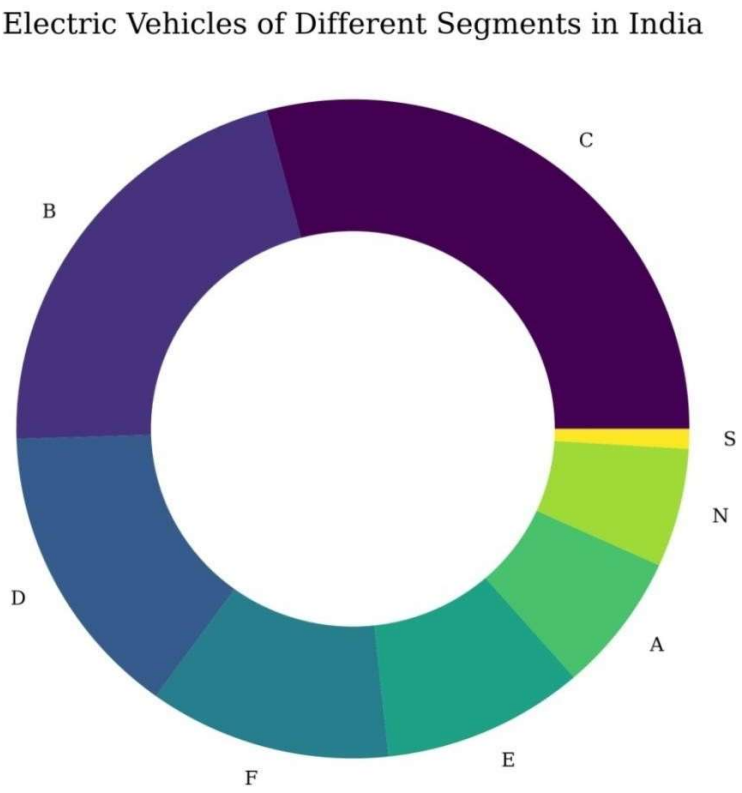
Comparison of different brands of EVs based on acceleration



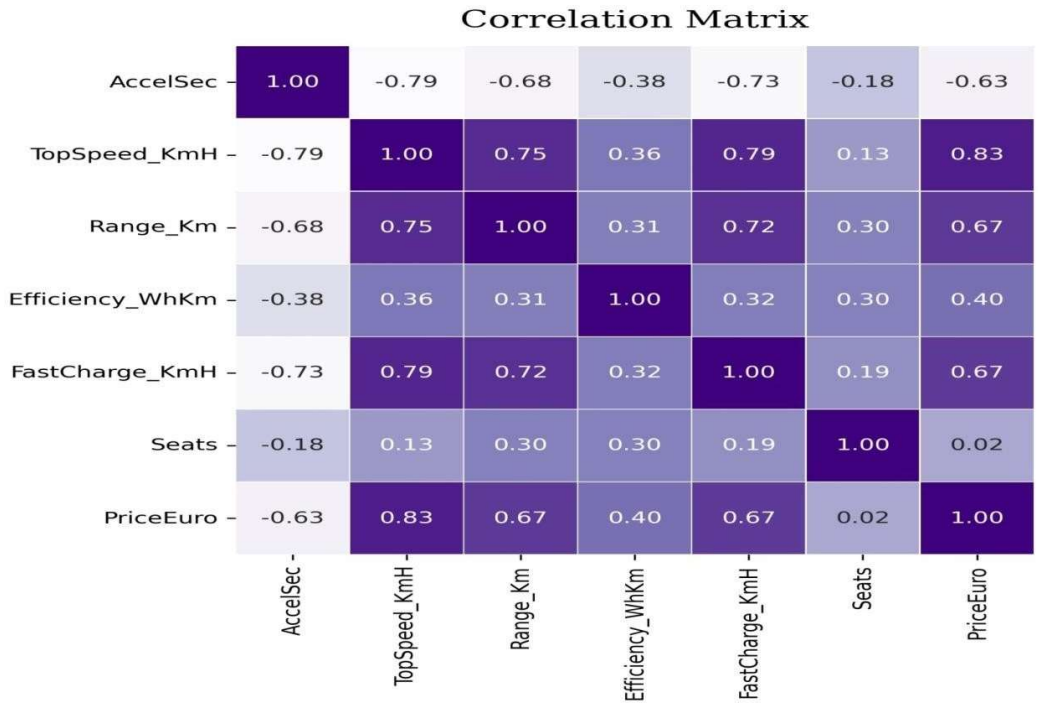
Comparison of different brands of EVs based on speed



EV Segments in India



Correlation Matrix



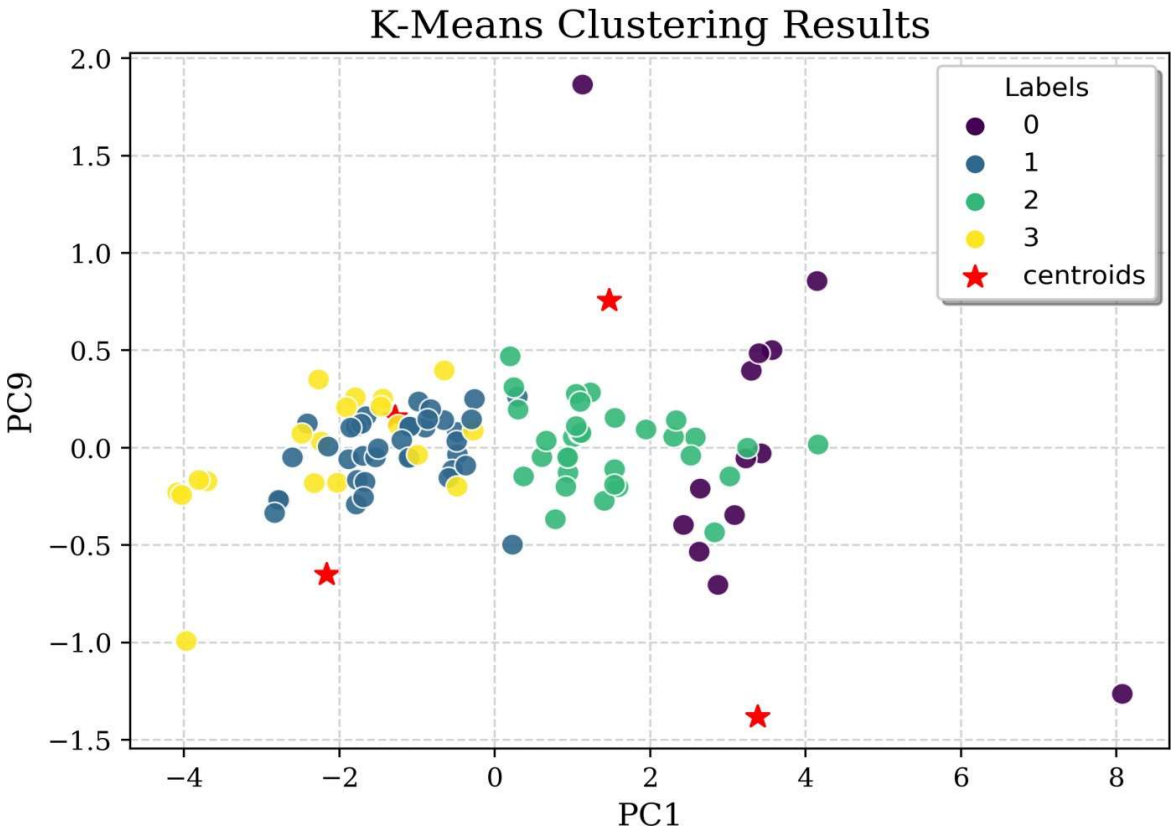
Segmentation Approaches

Clustering

Clustering is an unsupervised machine learning technique of grouping similar data points into clusters. The sole objective of this technique is to segregate datapoints with similar traits and place them into different clusters. There are several algorithms to perform clustering on data such as k-means clustering, hierarchical clustering, density-based clustering etc.

K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm whose job is to group the unlabelled dataset into different clusters where each datapoint belongs to only one cluster. Here, K is the number of clusters that need to be created in the process. The algorithm finds its applicability into a variety of use cases including market segmentation, image segmentation, image compression, document clustering etc. The below image is the results of clustering on one of ourdatasets.



The K-Means Algorithm works the following way:

1. Specify the number of clusters, i.e. K
2. Select K random points in the dataset. These points will be the centroids (centres) of eachof the K clusters.
3. Assign each data point in the dataset to one of the K centroids, based on its distance fromeach of the centroids.
4. Consider this clustering to be correct and reassign the Centroids to the mean of theseclusters.
5. Repeat Step 3. If any of the points change clusters, Go to step 4. Else Go to step 6.
6. Calculate the variance of each of the clusters.
7. Repeat this clustering 'n' number of times until the sum of variance of each cluster isminimum.

Principle Component Analysis

Principal component analysis (PCA) is a linear dimensionality-reduction technique that is used toreduce the dimensionality of large data sets by transforming a large set of variables into a smallerone while preserving most of the information present in the large set.

Elbow Method

The Elbow method is a way of determining the optimal number of clusters (k) in K-Means Clustering. It is based on calculating the Within Cluster Sum of Squared Errors (WCSS) for a different number of clusters (k) and selecting the k for which change in WCSS first starts to diminish. When you plot its graph, at one point the line starts to run parallel to the X-axis and that point, known as the Elbow Point, is considered as the best value for the k (as 4 in the belowfigure).