Electric Vehicle Market in India

Market Segmentation

Team Pradnya

Pradnya Kishor nabar

Jacob Charles Joy

Patel Raj Prashantkumar

Aditya kashyap

Shiva Chidambaram S

Abstract

Electric Vehicle(EVs) are the future of road transportation and is one of the leading business' in the motor world. In the past few years, a lot of electric vehicles have been introduced to the Indian market and are constantly being improved with a range of factors. This particular study aims to be looking at electric vehicles and their requirements and analyze the customers who are exploring EVs in order to come up with a strategy in which an EV startup can enter the market in India.

1. Introduction

There are a lot of EVs available in the market right now. In order to come up with a feasible strategy in which a company can introduce their vehicle into the Indian market, it would be best to perform market segmentation analysis using clustering. Clustering is a method in which we can group multiple data points in a dataset based on the similarity of values under the specified features. A number of segments have been taken into account for clusterings like behavioural segmentation, geographical segmentation and EV feature segmentation.

1.1 K-Means Clustering

K-Means clustering is an unsupervised machine learning technique using vector quantization that partitions a number of partitions into a number of clusters. The number of clusters for a particular dataset is usually defined by using the elbow method. It is an easy way to categorize and group unlabeled data. A certain number of centroids will be initialised in line with the number of clusters. The Euclidean distance between the data points and the centroids is calculated and accordingly, the data points are grouped together.

1.2 Elbow Method

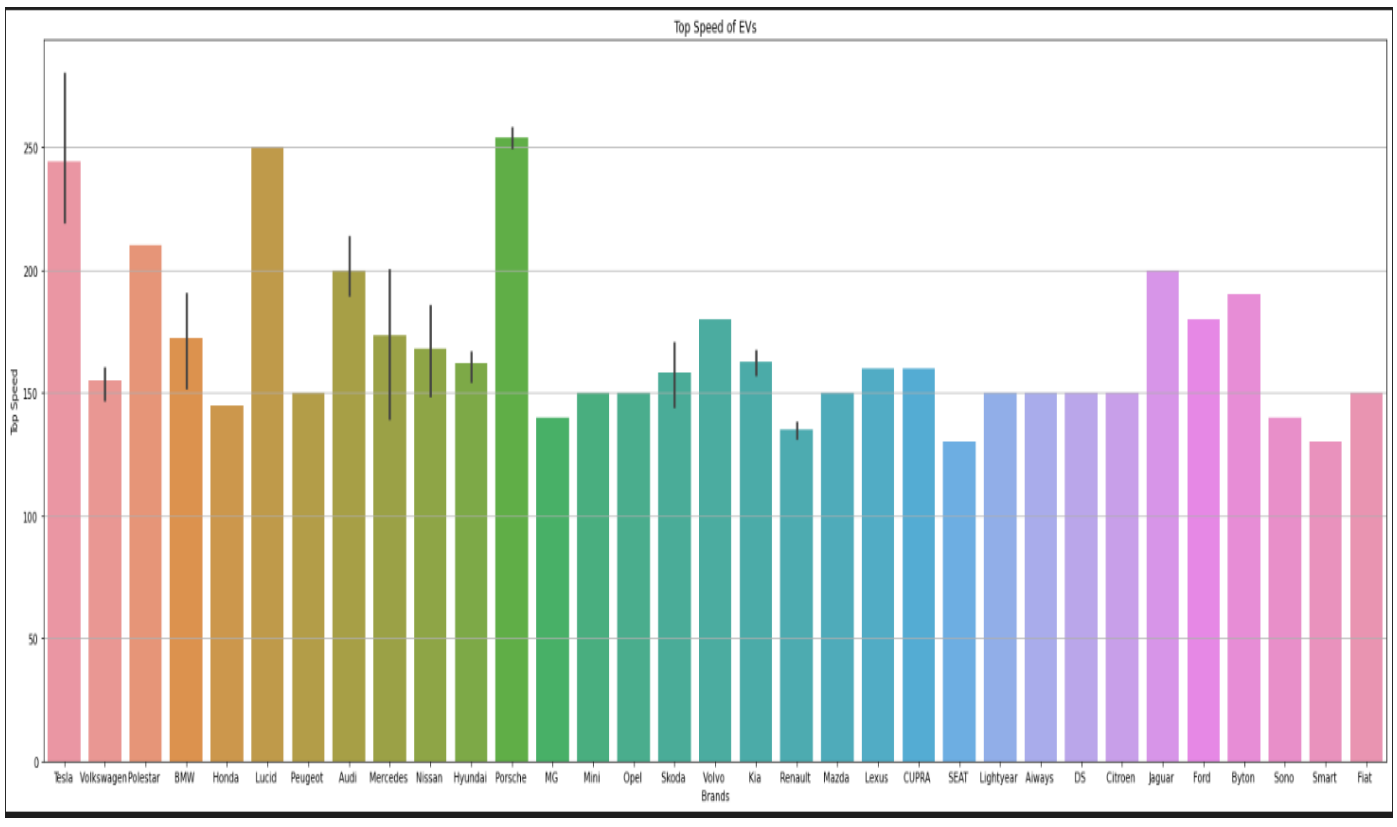
The elbow method is the method to find the appropriate number of clusters or centroids to be used when performing the K-Means clustering algorithm. It uses the concept of the Sum of Squared Errors. In each iteration, the algorithm calculates the sum of the squared distance between the centroid and data points. This value decreases each iteration until there is a distinct point("An Elbow") in which the difference between error s is minimal and it starts to flatline. This distinct point defines the number "K" or the number of clusters to be considered when creating the clustering algorithm.

EV Specification Segmentation

A company can look at the features of EVs being distributed and can choose a target segment based on what technical aspects of a car would be best for customers looking to buy a new EV themselves

Top Speed

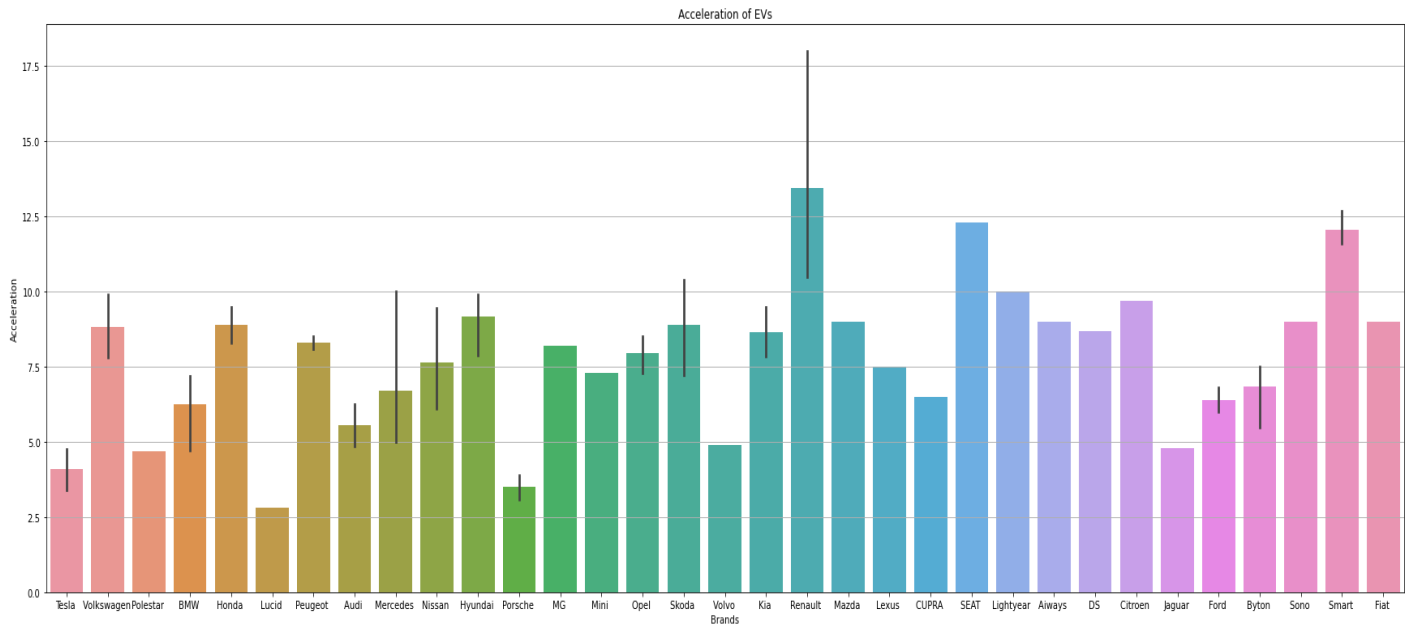
A lot of EVs are already being produced. Looking at current vehicles' top speed, we can understand what a new EV's top speed should be. We can visualise the data and get the average top speed of vehicles distributed right now.



The barplot helps us visualise and understand that most EV vehicles cap their top speed at 150km/h. The average top speed of most brands available in the market is about 180km/h. To make a vehicle that would likely attract customers and to get the vehicle to get more sales it would be best to create an EV in which the top speed caps a bit more than the average top speed of the vehicles currently available in the market right now(ex. 190km/h - 200km/h).

Acceleration

Aligned with top speed, it is also important to look at the acceleration of a vehicle. On legal roads and normal day-to-day driving, a person would almost never reach the top speed of a car. Therefore how fast a car can reach from one speed to another can be a very important feature to consider, especially in India where roads are crowded and a driver wants to maybe reach his/her point faster than he/she can.

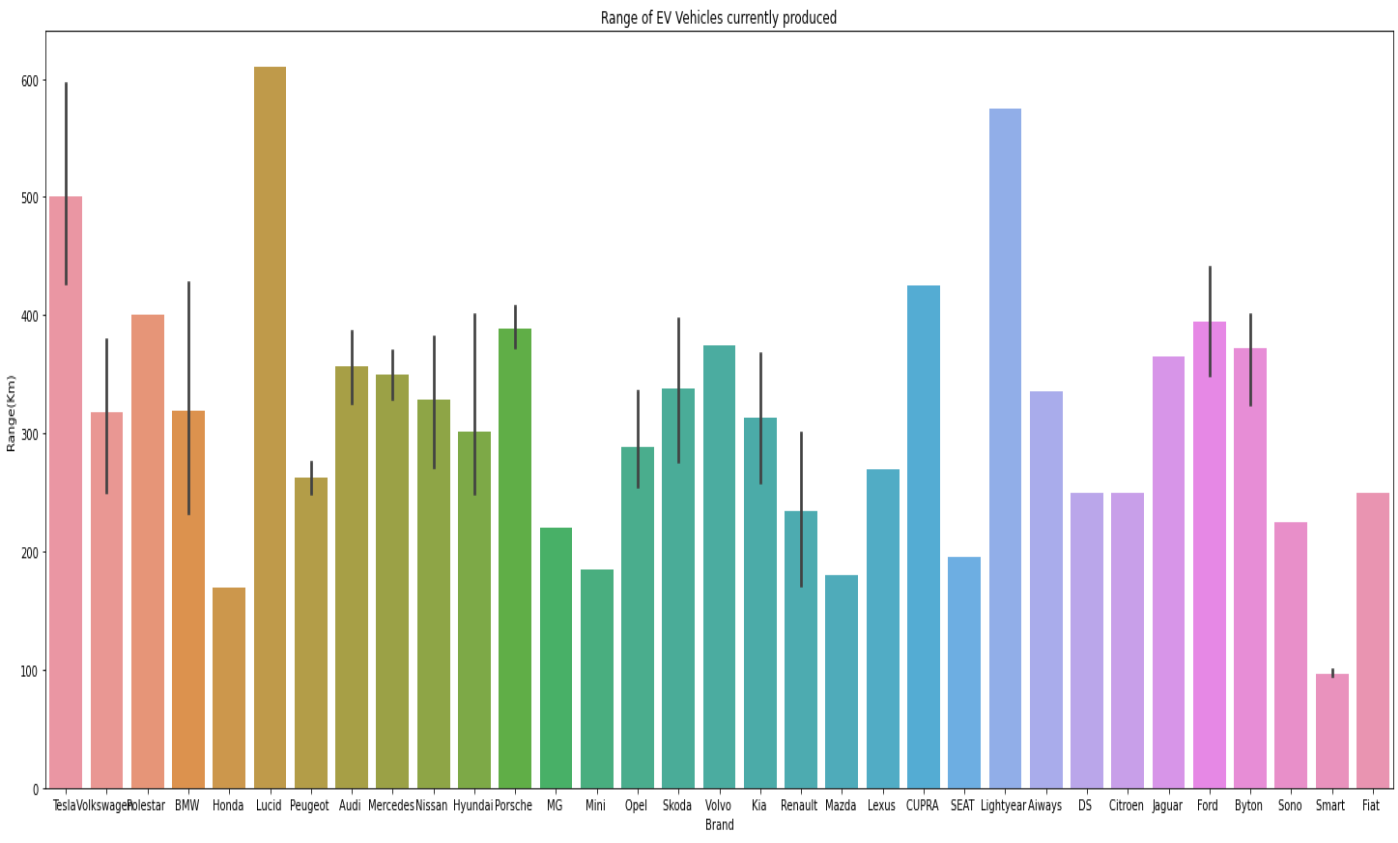


From the diagram, we can understand that most EVs present have an acceleration is about 8 - 9km/s. If we take the average of the values of acceleration we can see that the mean acceleration is about 7.4km/s. However, the average value is affected by the extreme values present in the dataset, for example, Renault has an acceleration of more than 12.5km/s. Therefore visualizing this data would be a good method to get a better understanding of what acceleration would be ideal for a new EV vehicle.

The barplot shows that having an acceleration of about 5-6 km/s.This can ensure that the EV being brought into the market would beat most of the competition that is in the market right now. It would be a vehicle that customers would more likely look into.

Range

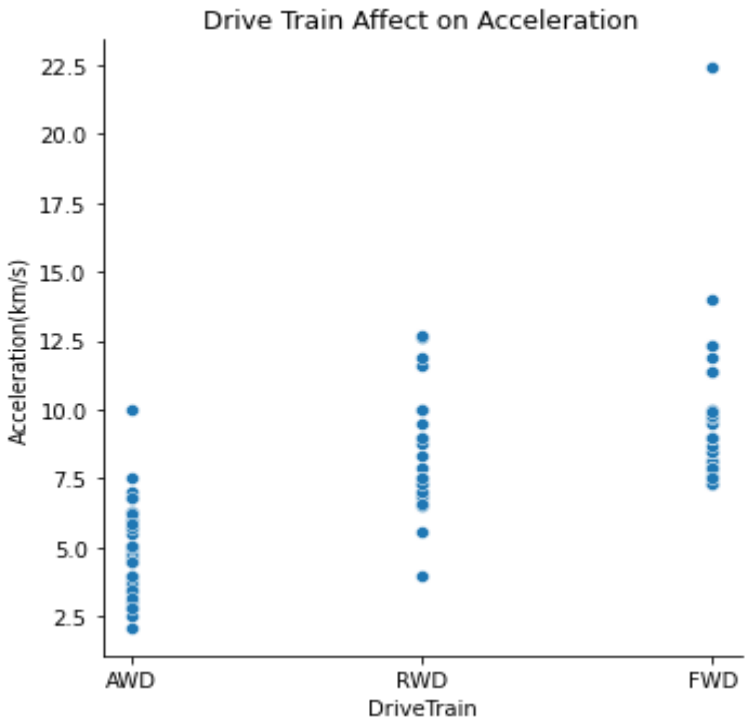
The speed of an EV is all well and good and is attractive to most customers. But one of the key features of an EV that almost all people take into consideration is the range that an EV can provide.

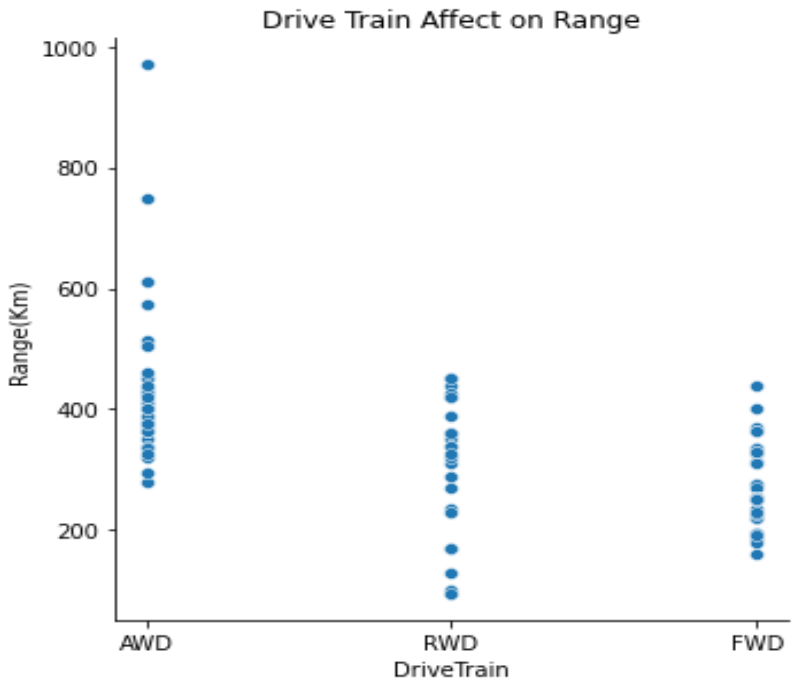


The barplot allows us to see that most EVs have a range of about 300km - 400km. Only a few brands like “Lucid” have a range of greater than 400km. In order for an EV to get into the early market, the vehicle should have a range of at least 500km and can range up to 600km if it is achievable. This would be a wise move for the company as only a few cars have a range of about 500km to 600km.

DriveTrain

The drive-Train of a vehicle specifies the number of motors an EV uses and how many wheels the motor controls itself.

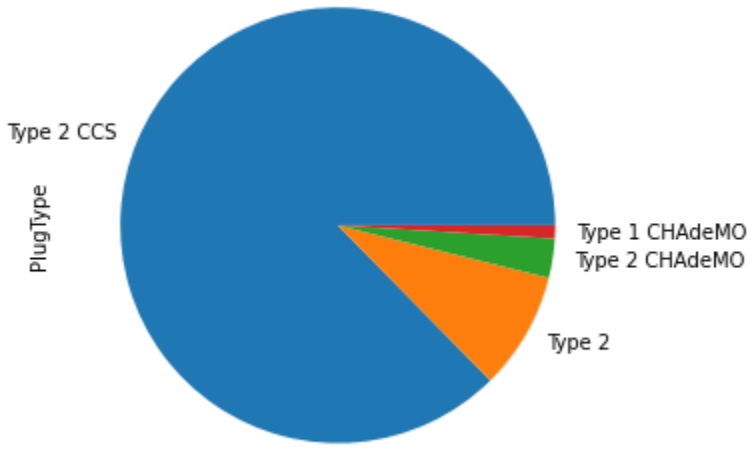




The drive train shows us that having an All Wheel Drive(AWD) vehicle would benefit both acceleration and range. This ensures that the company can build a faster car and a car that would give more miles as compared to an EV that is either Rear Wheel Drive(RWD) or Front Wheel Drive(FWD).

PlugType

Refers to the type of charging port that the EVs use.



The pie chart shows us that the “Type 2 CCS” plug type is the most common one used within the EV market and seems to be the most efficient one. Therefore using a “Type 2 CCS” itself for the startup would be a better idea as it would be the most researched and known one for consumers, therefore, giving them confidence in the charging capability of the vehicle

Advantages of Segmentation on EV Features/Specifications

- Knowing what makes a good EV vehicle in the current market right now and understanding what improvements can be made allows a company to understand what vehicle would more likely be purchased by a consumer
- A better vehicle than gives the best or the most of everything available right now in India would mean increased profits for the company.
- A comparison of the vehicle being developed and vehicles available not only in India but across the world is a good marketing strategy to advertise the way their vehicle is.

Implementation

Data Sources

In order to perform segmentation analysis to come up with a strategy for an EV startup we considered 3 datasets:

- ElectricCarData_Clean.csv
- Customer_Buying_Behavior.csv
- Dataset used

```
df = pd.read_csv("ElectricCarData_Clean.csv")

df
```

	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType
	Tesla	Model 3 Long Range Dual Motor	4.6	233	450	161	940	Yes	AWD	Type 2 CCS
	Volkswagen	ID.3 Pure	10.0	160	270	167	250	Yes	RWD	Type 2 CCS
	Polestar	2	4.7	210	400	181	620	Yes	AWD	Type 2 CCS
	BMW	iX3	6.8	180	360	206	560	Yes	RWD	Type 2 CCS
	Honda	e	9.5	145	170	168	190	Yes	RWD	Type 2 CCS

Data Preprocessing

Tools Used:

1. Sklearn(LabelEnoder)

Converting categorical variables to numbers

```
from sklearn.preprocessing import LabelEncoder

categorical = LabelEncoder()

df['RapidCharge'] = categorical.fit_transform(df['RapidCharge'])
```

	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	P
0	Tesla	Model 3 Long Range Dual Motor	4.6	233	450	161	940	1	
1	Volkswagen	ID.3 Pure	10.0	160	270	167	250	1	
2	Polestar	2	4.7	210	400	181	620	1	
3	BMW	iX3	6.8	180	360	206	560	1	
4	Honda	e	9.5	145	170	168	190	1	
...	...	...	...	...	...	...	...	...	
98	Nissan	Ariya 63kWh	7.5	160	330	191	440	1	
99	Audi	e-tron S Sportback 55 quattro	4.5	210	335	258	540	1	

Visualizing Data

- **Visualizing the number of brands that produce EVs**

```
brands = plt.figure(figsize=(30,10))

sns.barplot(x='Brand',y=range,data=df)

plt.grid(axis='y')

plt.title("Brands Producing EVs")

plt.xlabel("Brands")

plt.ylabel("Frequency")
```

- **Top Speed achieved by brands producing EV to get an estimate as to how much a new company should acheive**

```
top_speed = plt.figure(figsize=(30,10))

sns.barplot(x='Brand',y='TopSpeed_KmH',data=df)

plt.grid(axis='y')

plt.title("Top Speed of EVs")

plt.xlabel("Brands")

plt.ylabel("Top Speed")
```

- **Acceleration of EVs in the market**

```
acceleration = plt.figure(figsize=(30,10))

sns.barplot(x='Brand',y='AccelSec',data=df)

plt.grid(axis='y')

plt.title("Acceleration of EVs")

plt.xlabel("Brands")

plt.ylabel("Acceleration")
```

- **Range of EVs**

```
range = plt.figure(figsize=(25,10))

sns.barplot(x='Brand',y='Range_Km',data=df)

plt.title("Range of EV Vehicles currently produced")

plt.xlabel("Brand")

plt.ylabel("Range(Km)")
```

- **Plug Type**

```
plug_type = plt.figure(figsize=(10,5))

df['PlugType'].value_counts().plot.pie()
```

The pie chart shows that most EV producers use Type2 CCS cables

- **Drive Train Affect on Range**

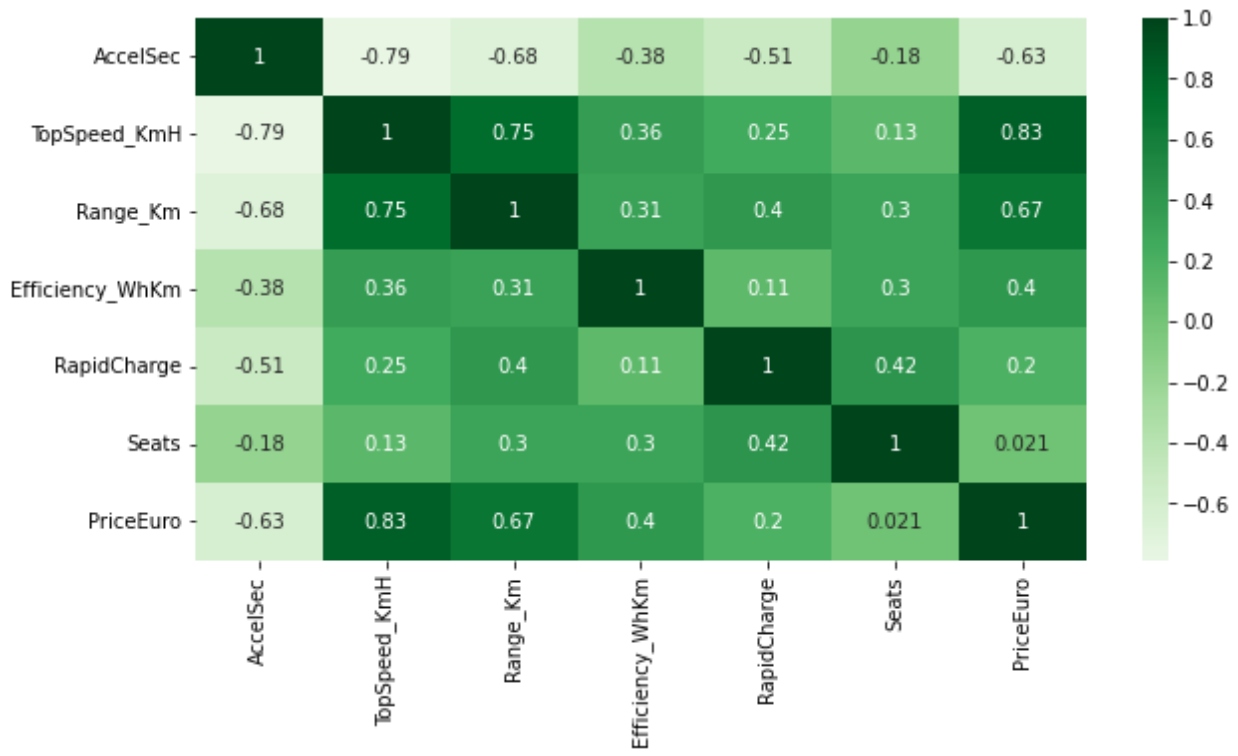
```
rng_dt = plt.figure(figsize=(15,10))

sns.relplot(data=df,x='PowerTrain',y='Range_Km')

plt.title("Drive Train Affect on Range")

plt.xlabel("DriveTrain")

plt.ylabel("Range(Km)")
```

- **Drive Train Affect on Acceleration**

```
spd_dt = plt.figure(figsize=(15,10))

sns.relplot(data=df,x='PowerTrain',y='AccelSec')

plt.title("Drive Train Affect on Acceleration")

plt.xlabel("DriveTrain")

plt.ylabel("Acceleration(km/s)")
```

- **Heatmap to understand which features have the most affect for an EV**

```
heatmap = plt.figure(figsize=(10,5))

sns.heatmap(df.corr(), center=0,cmap = 'Greens',annot=True)
```



From the heatmap, we can understand that from the features of an EV the range and top speed of a car have the highest positive correlation

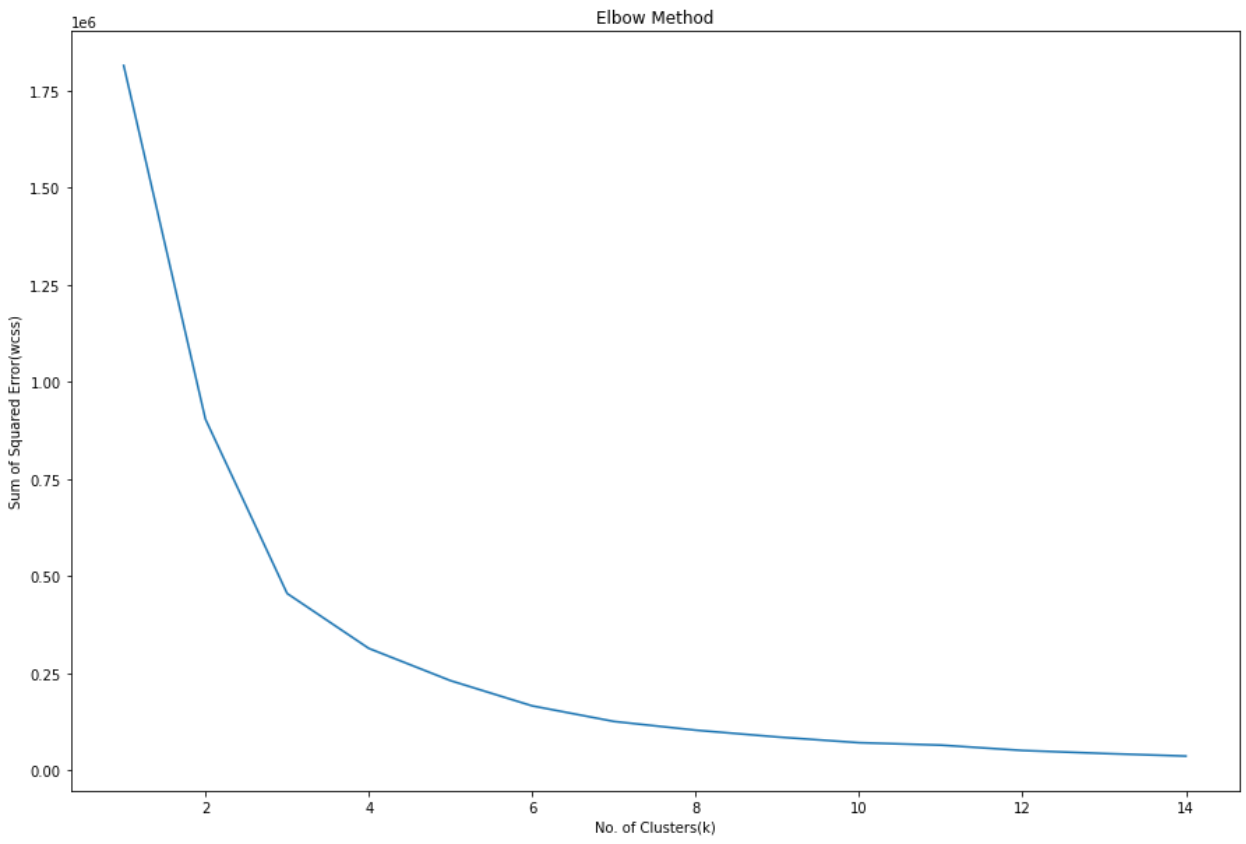
- **K-Means Clustering Algorithm**

Updating wcss array

```
- x = df.iloc[:,[3,4]]
- X
- from sklearn.cluster import KMeans
- wcss = []
- k= np.arange(1,15)
- for i in k:
-     kmeans = KMeans(n_clusters=i, init='k-means++', random_state=45)
-     kmeans.fit(x)
-     wcss.append(kmeans.inertia_)
-
```

Elbow Method

```
plt.figure(figsize=(15,10))

plt.plot(k,wcss)

plt.title("Elbow Method")

plt.xlabel("No. of Clusters(k)")

plt.ylabel("Sum of Squared Error(wcss)")

plt.show()
```



The elbow method shows us that the number of clusters we must use for clustering is about 6 as there is a distinct point where the slope starts to flatline

- **Adding the Clusters to the dataset**

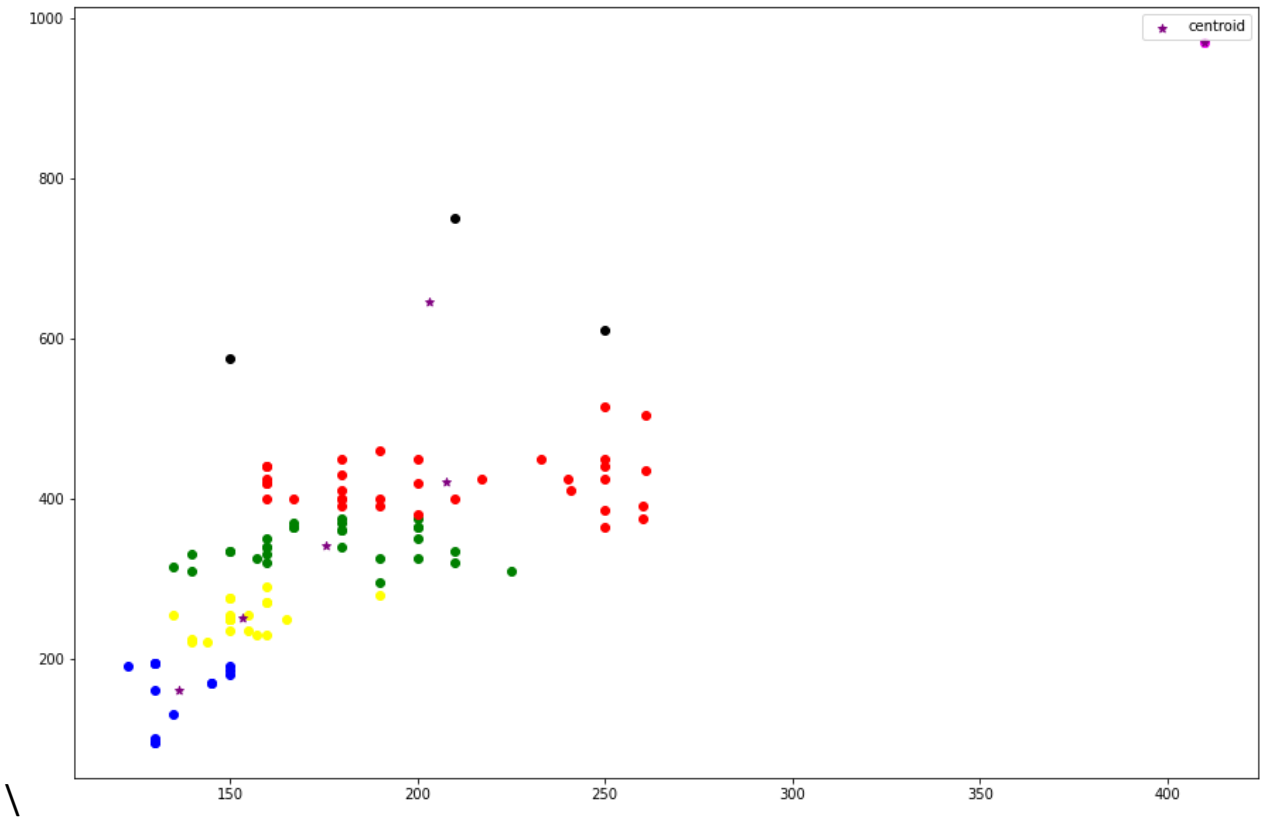
```
kmeans = KMeans(n_clusters=6)

x_predict = kmeans.fit_predict(x)
```

```
x_predict

df['cluster'] = x_predict

df
```

Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment	Seats	PriceEuro	cluster
450	161	940	1	AWD	Type 2 CCS	Sedan	D	5	55480	2
270	167	250	1	RWD	Type 2 CCS	Hatchback	C	5	30000	1
400	181	620	1	AWD	Type 2 CCS	Liftback	D	5	56440	2
360	206	560	1	RWD	Type 2 CCS	SUV	D	5	68040	5
170	168	190	1	RWD	Type 2 CCS	Hatchback	B	4	32997	4

- **Plotting the Clusters using a scatter plot**

```
kmeans.cluster_centers_

df1 = df[df.cluster==0]

df2 = df[df.cluster==1]

df3 = df[df.cluster==2]

df4 = df[df.cluster==3]

df5 = df[df.cluster==4]

df6 = df[df.cluster==5]



plt.figure(figsize=(15,10))

plt.scatter(df1['TopSpeed_KmH'],df1['Range_Km'],color='green')

plt.scatter(df2['TopSpeed_KmH'],df2['Range_Km'],color='red')

plt.scatter(df3['TopSpeed_KmH'],df3['Range_Km'],color='black')

plt.scatter(df4['TopSpeed_KmH'],df4['Range_Km'],color='blue')

plt.scatter(df5['TopSpeed_KmH'],df5['Range_Km'],color='magenta')

plt.scatter(df6['TopSpeed_KmH'],df6['Range_Km'],color='yellow')

plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color =
'purple',marker='*',label='centroid')

plt.legend()

plt.show()
```



\

Plotting the clusters shows that there are 4 main clusters in the dataset therefore we could have 4 segments

Problem Statement

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 discriminate 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 labeled as 'Conservatives', 'Indifferent', and 'Enthusiasts' which are deemed to be budding EV buyers The implications are recommended, which may offer some pertinent guidance for scholars and policymakers to encourage EVs adoption in the backdrop of emerging sustainable transport market.

In this report we are going to analyze 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.

Data Collection

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

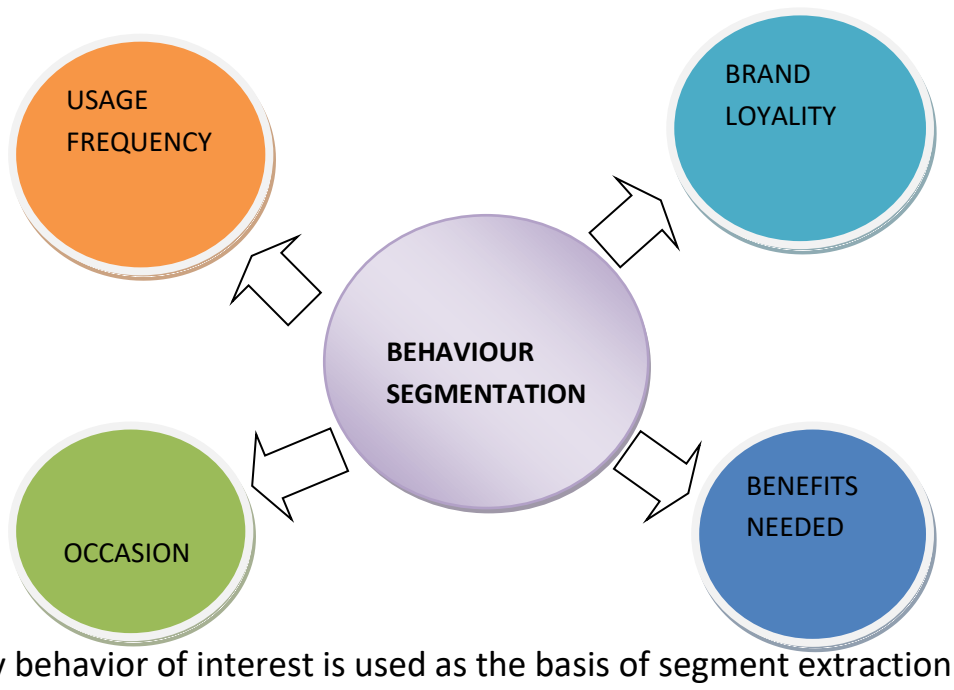
- https://www.kaggle.com/datasets
- https://data.gov.in/
- https://www.data.gov/
- https://data.worldbank.org/
- https://datasetsearch.research.google.com/

Market Segmentation

Target Market

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

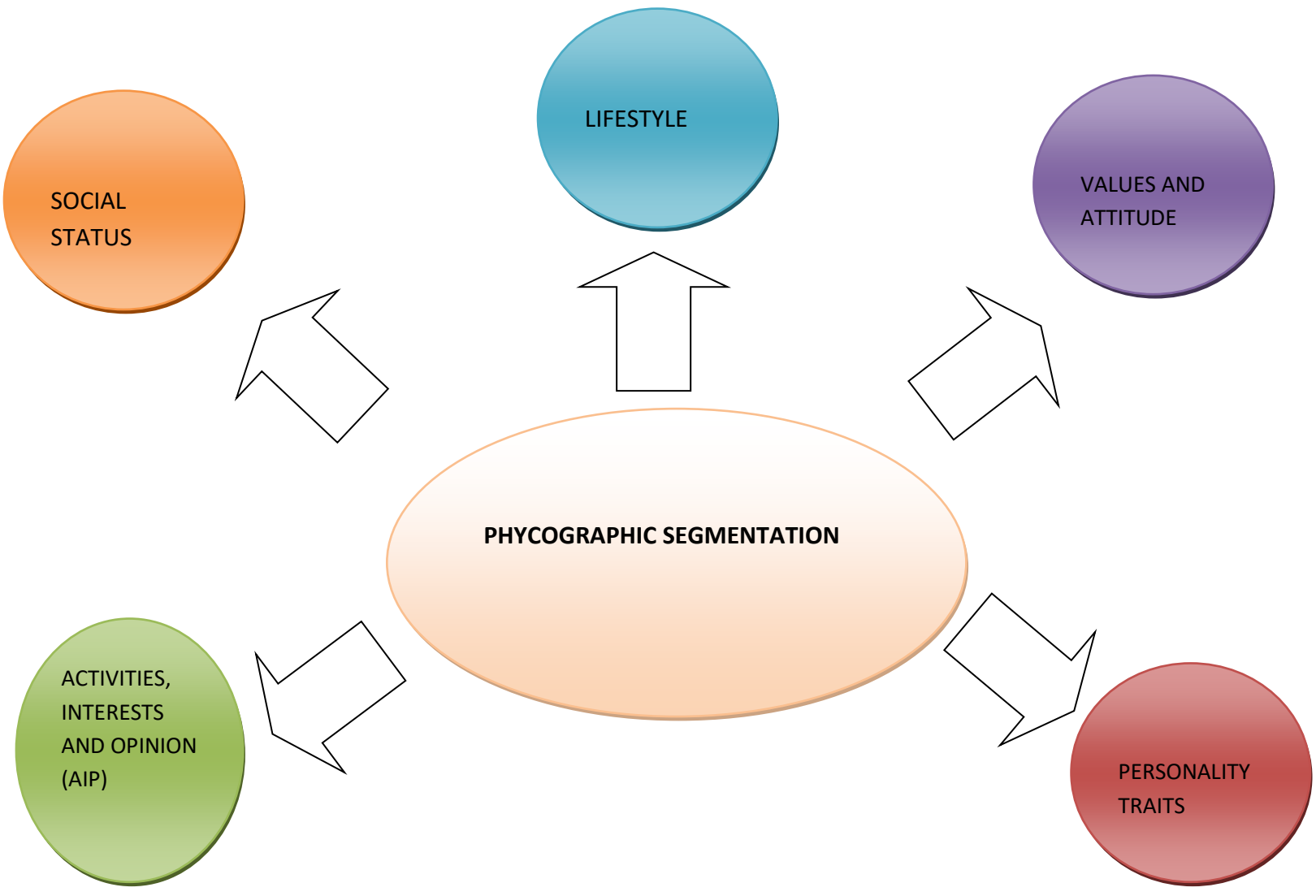
Behavioral Segmentation: Searches directly for similarities in behavior or reported behavior.
Example: Prior experience with the product, amount spent on the purchase, etc.



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

Disadvantage: Not always readily available.

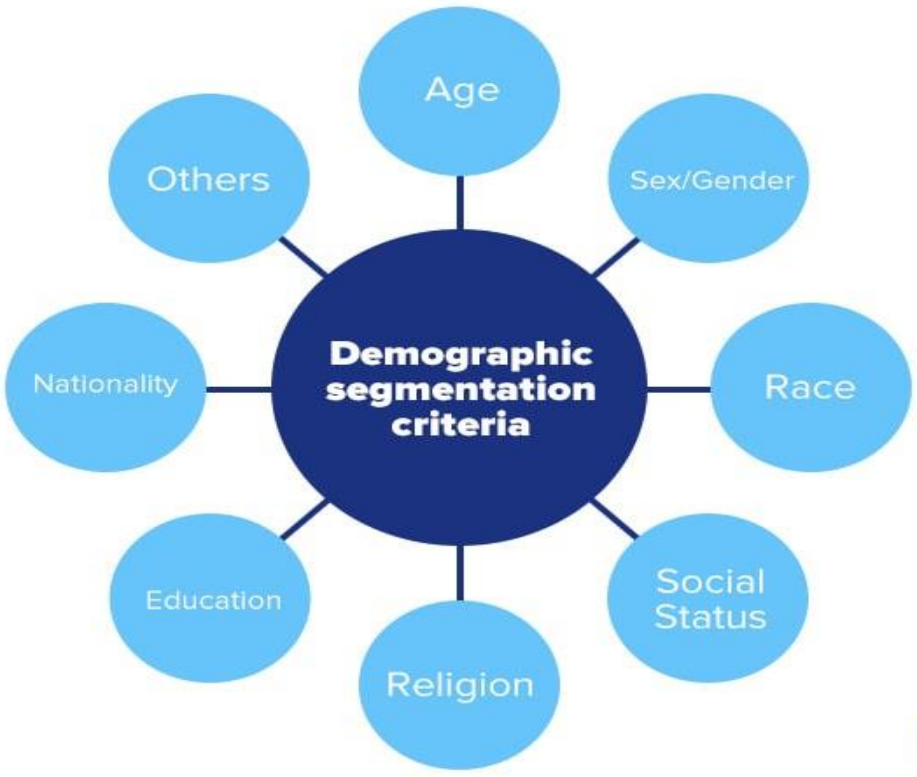
Psychographic Segmentation: Grouped based on beliefs, interests, preferences, aspi-rations, or benefits sought when purchasing a product. Suitable for lifestyle segmentation. Involves many segmentation variables.



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

Disadvantage: increased complexity of determining segment memberships for consumers.

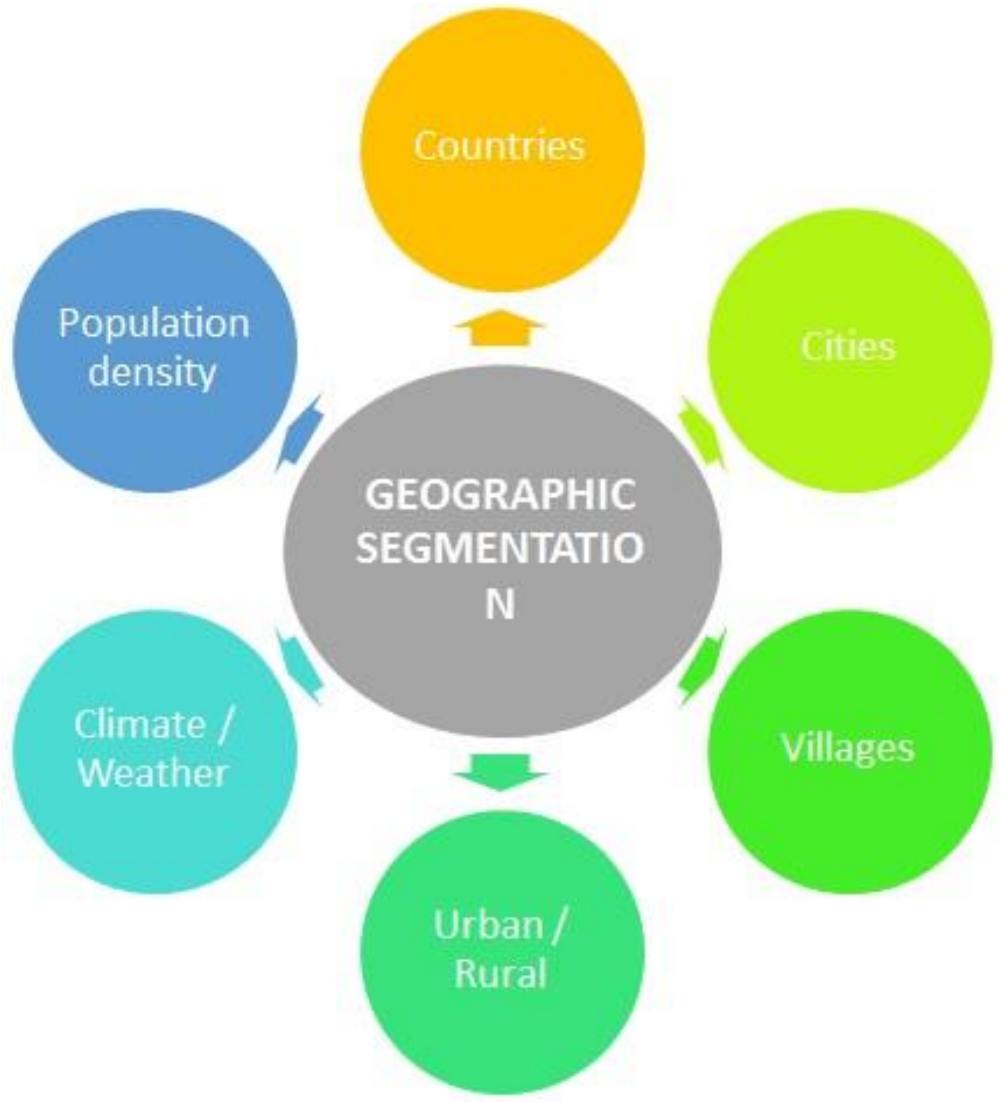
Socio-Demographic Segmentation: includes age, gender, income and education. Useful in industries.



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

Disadvantage: If these criteria are not the cause for customer's product preferences then it does not provide sufficient market insight for optimal segmentation decisions.

Geographic Segmentation: segmenting your audience based on the region they live or work in. This can be done in any number of ways: grouping customers by the country they live in, or smaller geographical divisions, from region to city, and right down to postal code.



Advantage: Geographic segmentation allows small businesses with limited budgets to be more cost effective. The findings that result from geographic segmentation allow small businesses to focus their marketing efforts specifically on their defined area of interest, therefore avoiding inefficient spending.

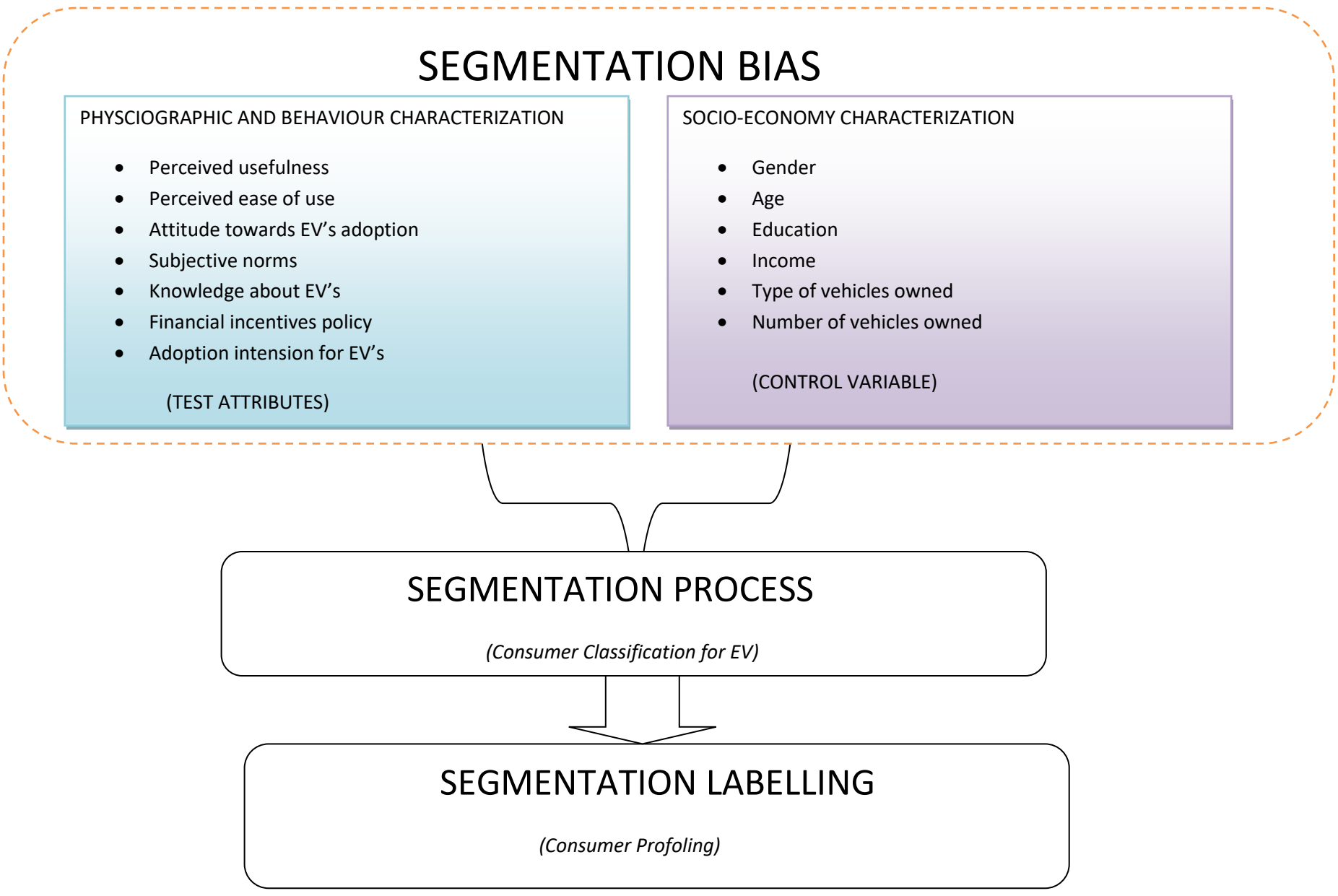
Disadvantage: Companies often do not rely solely on geographic segments to determine their target market.

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.

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

Packages/Tools used:

Numpy: To calculate various calculations related to arrays.

Pandas: To read or load the datasets.

SKLearn: We have used LabelEncoder() to encode our values.

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.

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.io as pio
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, f1_score, recall_score, precision_score
from sklearn.linear_model import LinearRegression
```

In [3]:
```
df = pd.read_csv("ElectricCarData_Clean.csv")
df
```

Out[3]:

	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyl
0	Tesla	Model 3 Long Range Dual Motor	4.6	233	450	161	940	Yes	AWD	Type 2 CCS	Seda
1	Volkswagen	ID.3 Pure	10.0	160	270	167	250	Yes	RWD	Type 2 CCS	Hatchba
2	Polestar	2	4.7	210	400	181	620	Yes	AWD	Type 2 CCS	Liftba
3	BMW	iX3	6.8	180	360	206	560	Yes	RWD	Type 2 CCS	SU
4	Honda	e	9.5	145	170	168	190	Yes	RWD	Type 2 CCS	Hatchba
...	...	...	...	...	...	...	...	...	...	...	.
98	Nissan	Ariya 63kWh	7.5	160	330	191	440	Yes	FWD	Type 2 CCS	Hatchba
99	Audi	e-tron S Sportback 55 quattro	4.5	210	335	258	540	Yes	AWD	Type 2 CCS	SU
100	Nissan	Ariya e-4ORCE 63kWh	5.9	200	325	194	440	Yes	AWD	Type 2 CCS	Hatchba
101	Nissan	Ariya e-4ORCE 87kWh Performance	5.1	200	375	232	450	Yes	AWD	Type 2 CCS	Hatchba
102	Byton	M-Byte 95 kWh 2WD	7.5	190	400	238	480	Yes	AWD	Type 2 CCS	SU

103 rows × 14 columns

EDA

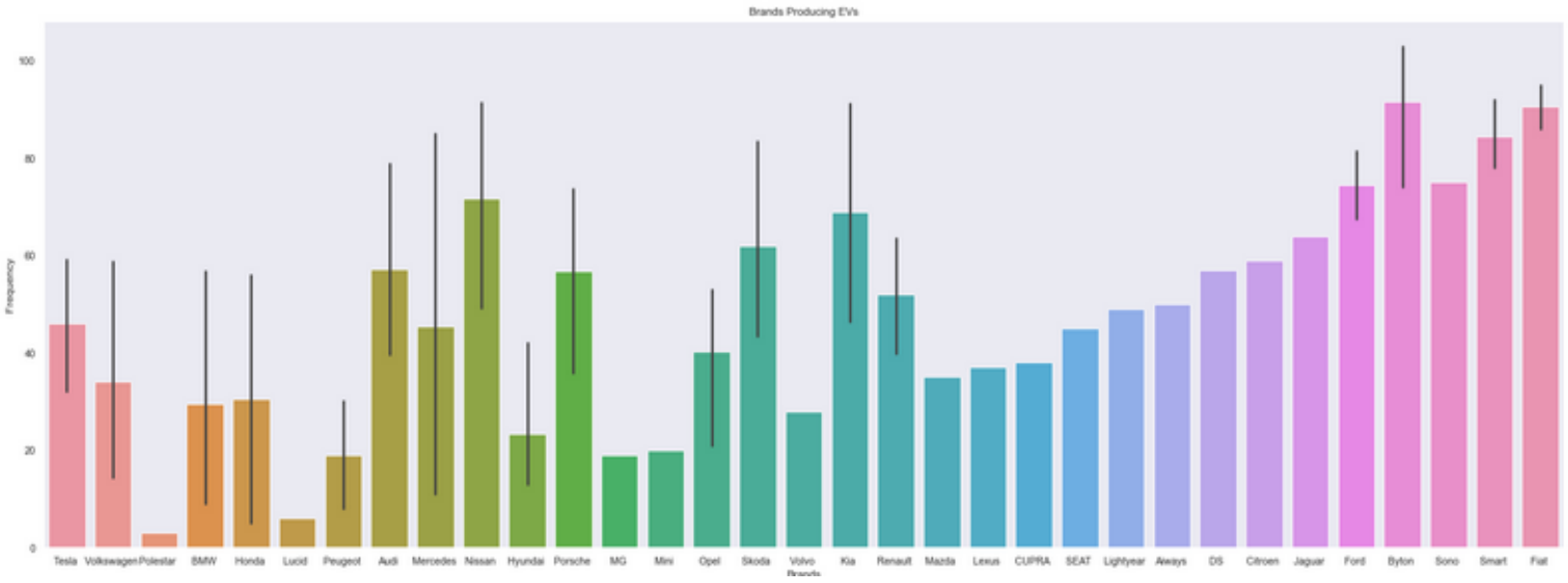
We start the Exploratory Data Analysis with some data Analysis drawn from the data without Principal Component Analysis and with some Principal Component Analysis in the dataset obtained from the combination of all the data we have. PCA is a statistical process that converts the observations of correlated 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.

Comparison of cars in our data

Visualizing the number of brands that produce EVs

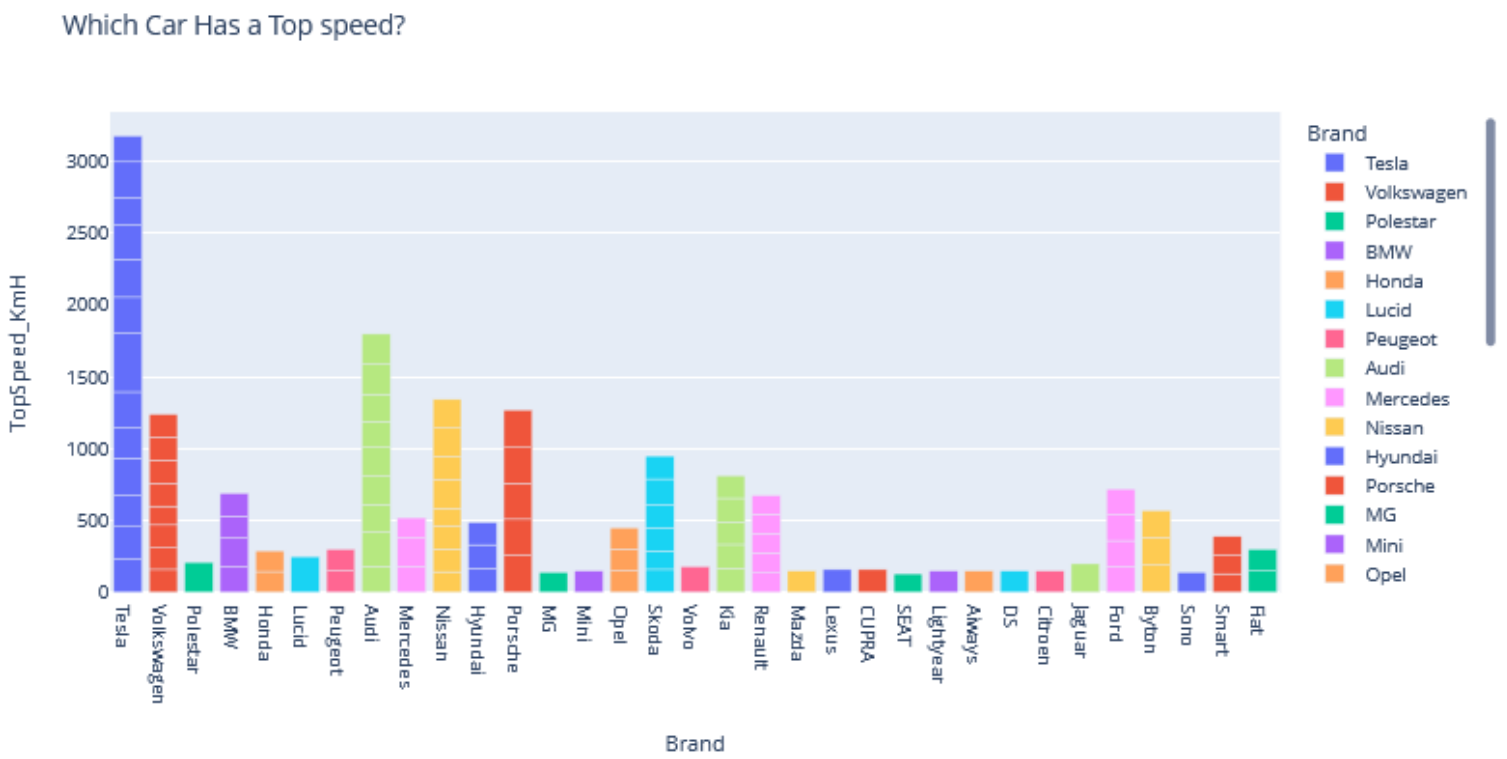
In [214]:
```
brands = plt.figure(figsize=(30,10))
sns.barplot(x='Brand',y=range,data=df)
plt.grid(axis='y')
plt.title("Brands Producing EVs")
plt.xlabel("Brands")
plt.ylabel("Frequency")
```

Out[214]: Text(0, 0.5, 'Frequency')



Which car has top speed

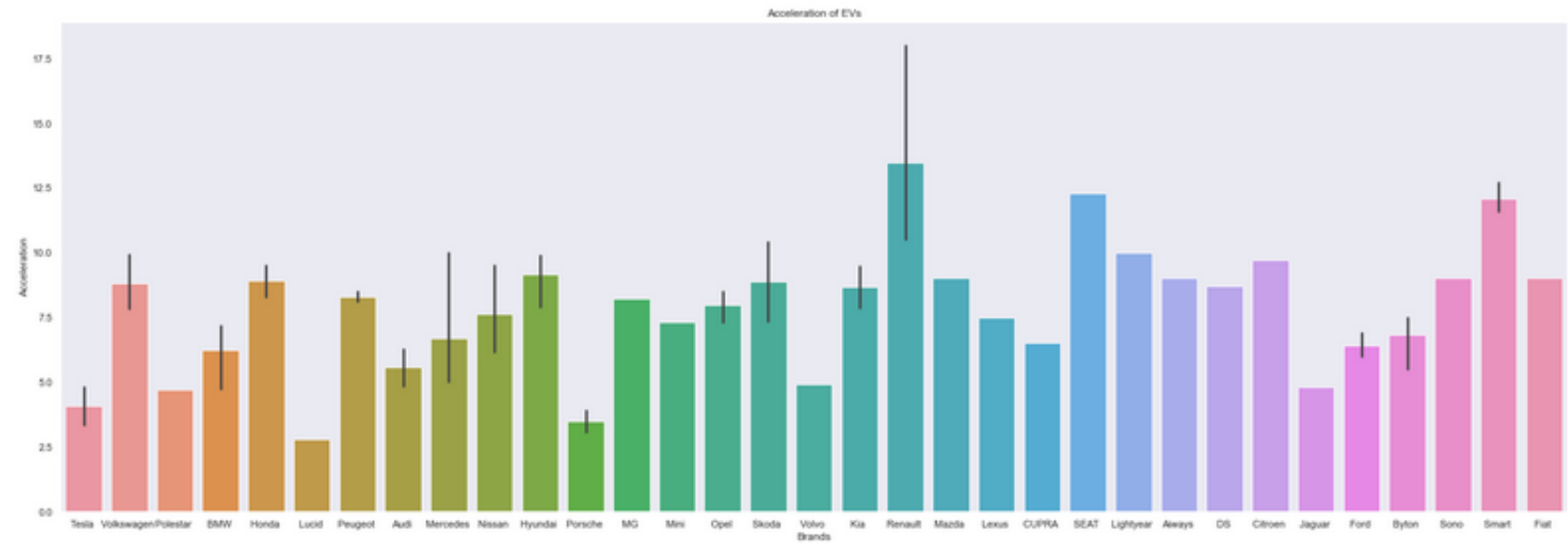
In [13]:
```
fig = px.bar(df,x='Brand',y = 'TopSpeed_KmH',color = 'Brand',title = 'Which Car Has a Top speed?',labels = {'x':'Car Brands','y
pio.show(fig)
```



Which car has fastest accelaration?

```
top_speed = plt.figure(figsize=(30,10))
sns.barplot(x='Brand',y='AccelSec',data=df)
plt.grid(axis='y')
plt.title("Acceleration of EVs")
plt.xlabel("Brands")
plt.ylabel("Acceleration")
```

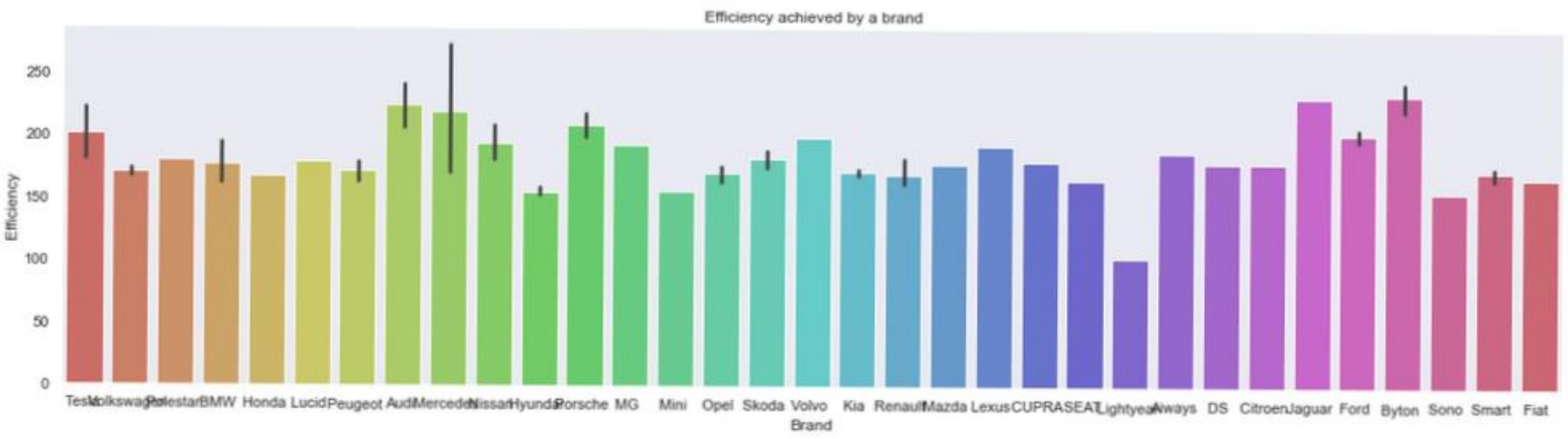
```
Text(0, 0.5, 'Acceleration')
```



Car Efficiency

```
ax= plt.figure(figsize=(20,5))
sns.barplot(x='Brand',y='Efficiency_WhKm',data=df,palette='hls')
plt.grid(axis='y')
plt.title('Efficiency achieved by a brand')
plt.xlabel('Brand')
plt.ylabel('Efficiency')
```

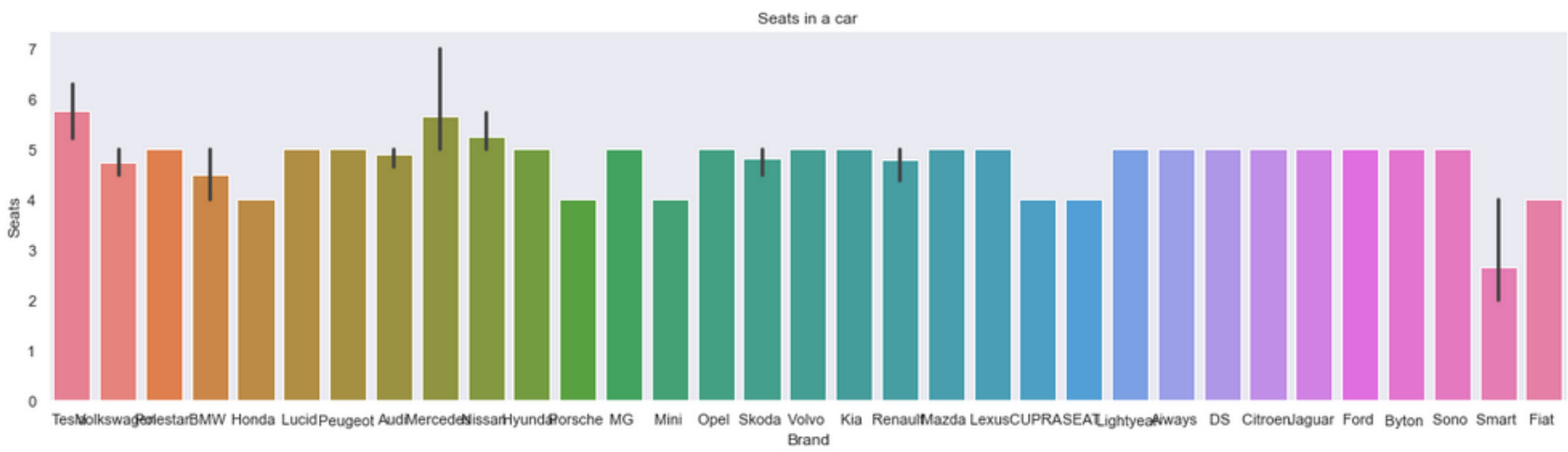
```
Text(0, 0.5, 'Efficiency')
```



Number of seats in car

```
ax= plt.figure(figsize=(20,5))
sns.barplot(x='Brand',y='Seats',data=df,palette='husl')
plt.grid(axis='y')
plt.title('Seats in a car')
plt.xlabel('Brand')
plt.ylabel('Seats')
```

```
Text(0, 0.5, 'Seats')
```



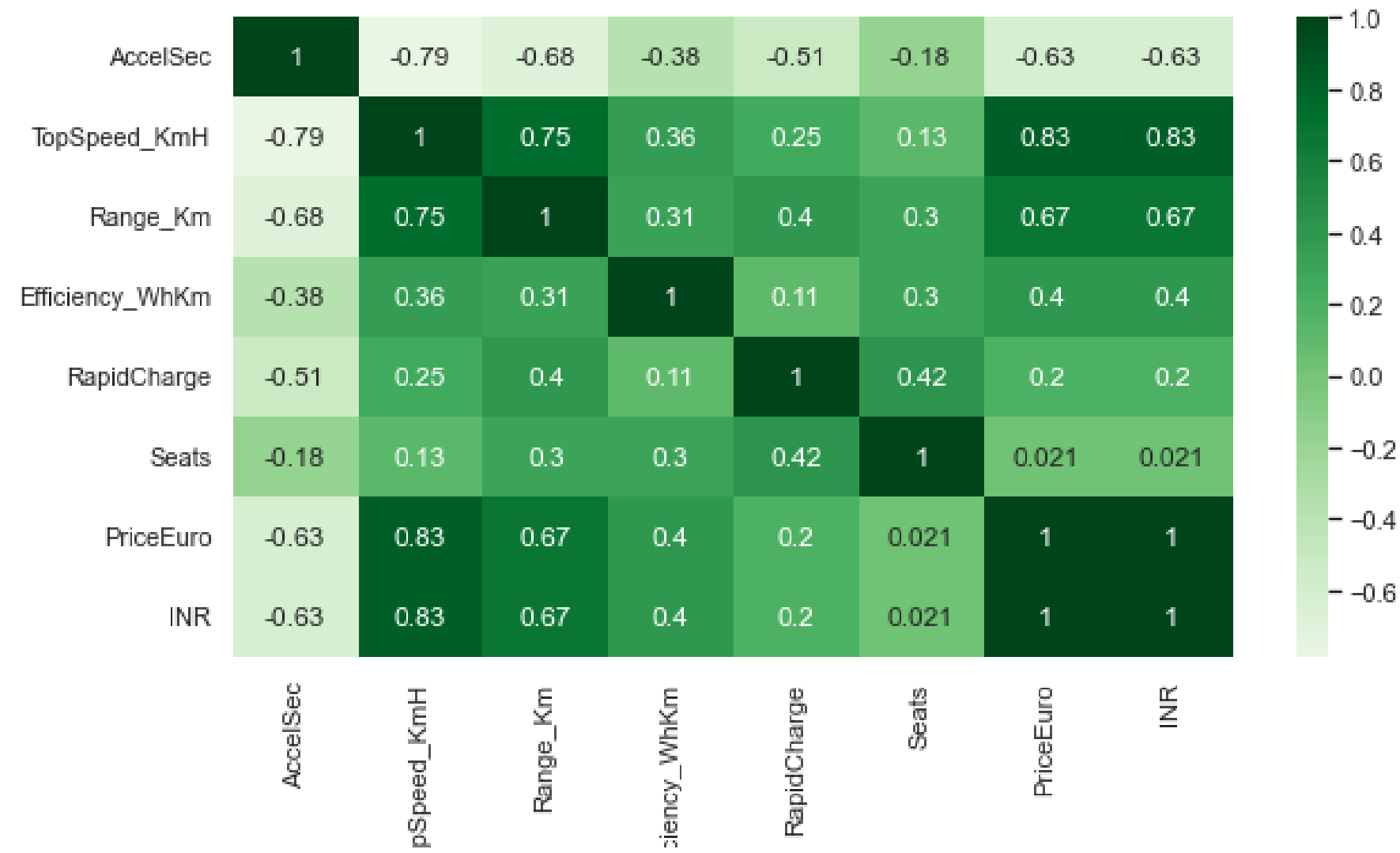
Correlation Matrix:

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 between all the possible pairs of values through the heat map in the below figure. The relationship between two variables is usually considered strong when their correlation coefficient value is larger than 0.7

Heatmap to show the correlation of the data

```
heatmap = plt.figure(figsize=(10,5))
sns.heatmap(df.corr(), center=0,cmap = 'Greens',annot=True)
```

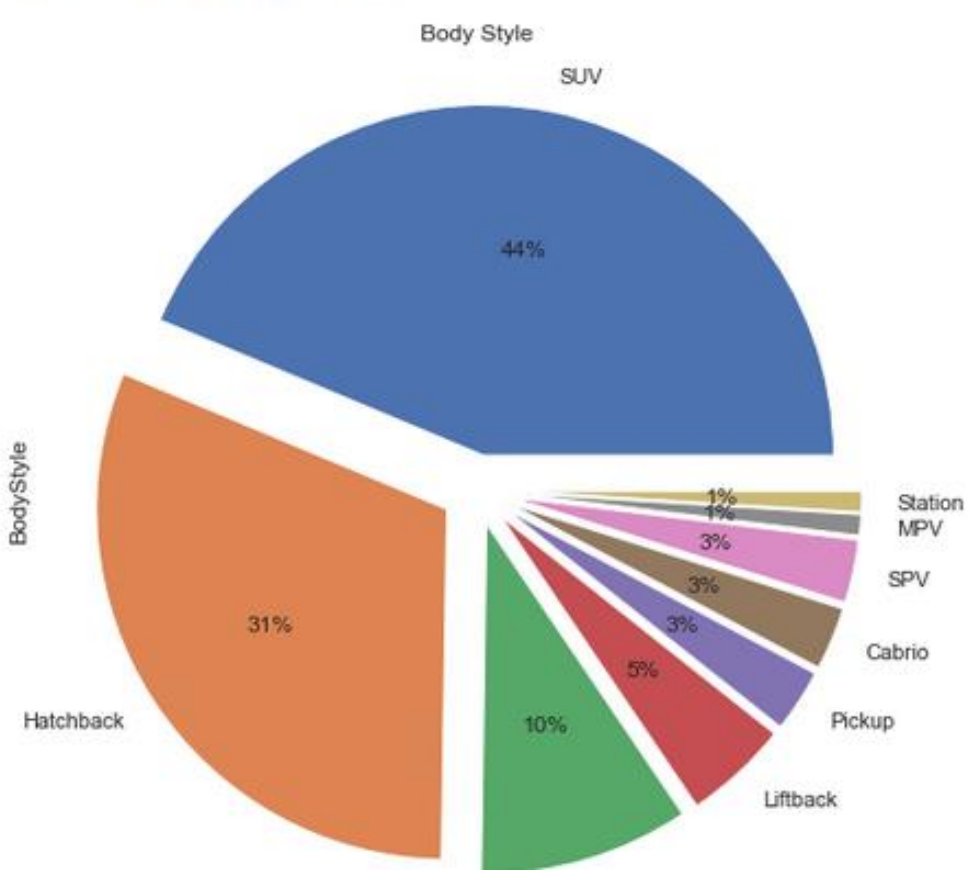
```
<AxesSubplot:>
```



Cars and their body style

```
df['BodyStyle'].value_counts().plot.pie(figsize=(8,15),autopct='%.0f%%',explode=(0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1))
plt.title('Body Style')
```

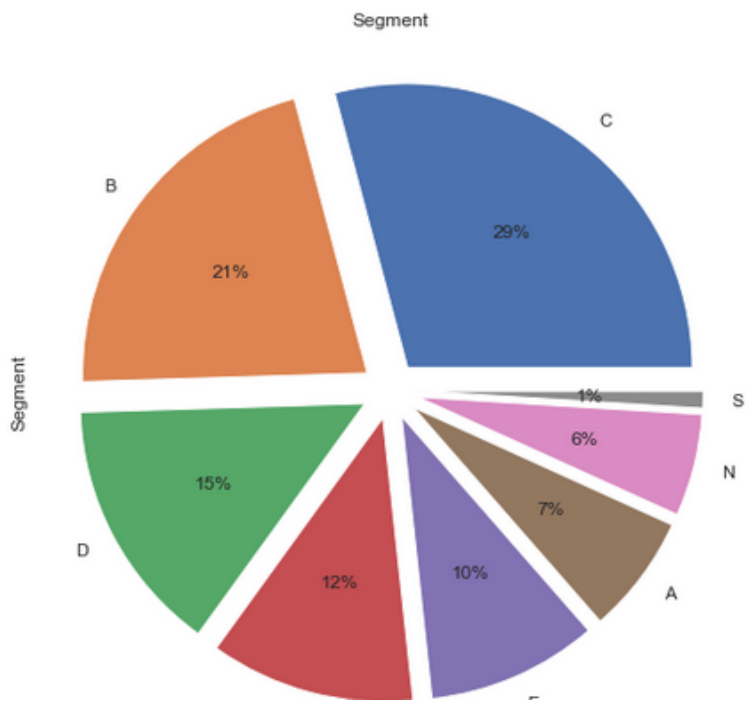
```
Text(0.5, 1.0, 'Body Style')
```



Segment in which the cars fall under

In [298]:
```
df['Segment'].value_counts().plot.pie(figsize=(8,15),autopct='%.0f%%',explode=(0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1))
plt.title('Segment')
```

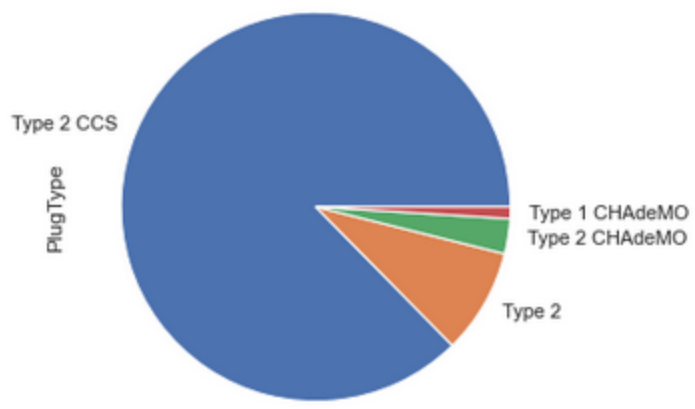
Out[298]: Text(0.5, 1.0, 'Segment')



plug type

In [224]:
```
plug_type = plt.figure(figsize=(10,5))
df['PlugType'].value_counts().plot.pie()

```

Out[224]: <AxesSubplot:ylabel='PlugType'>

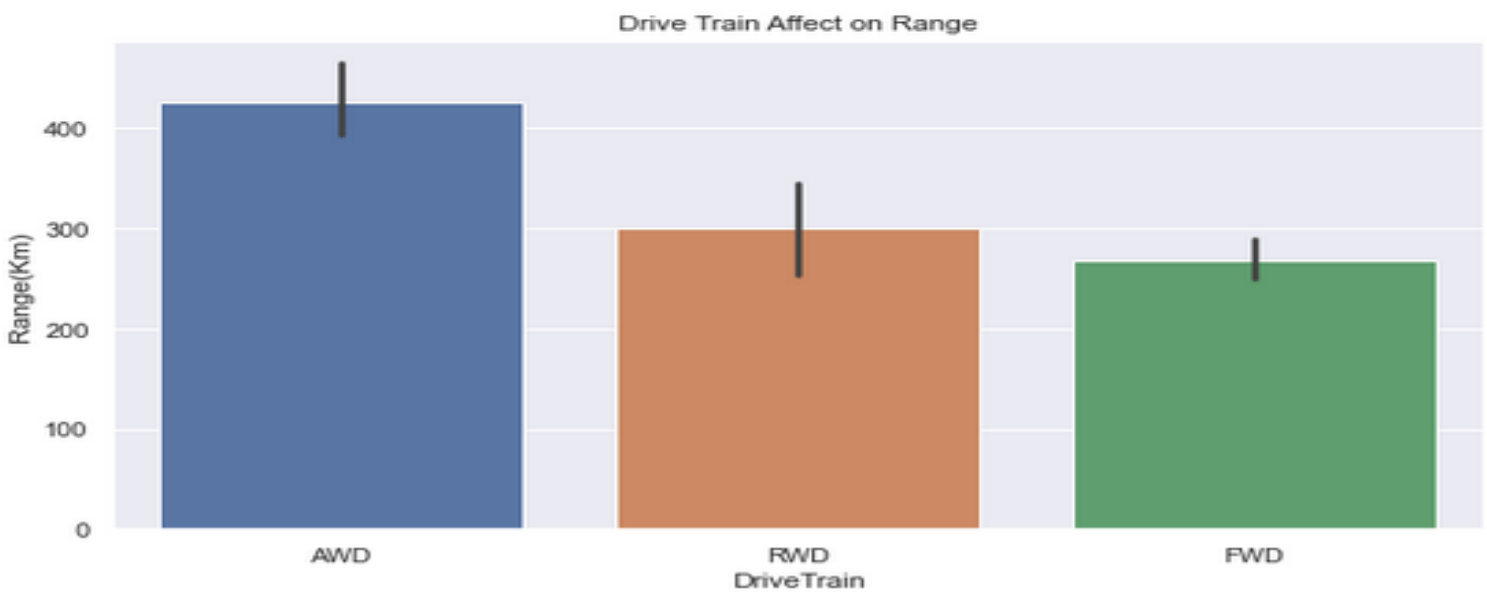


- The pie chart shows that most EV producers use Type2 CCS cables

Drive Train Affect on Range

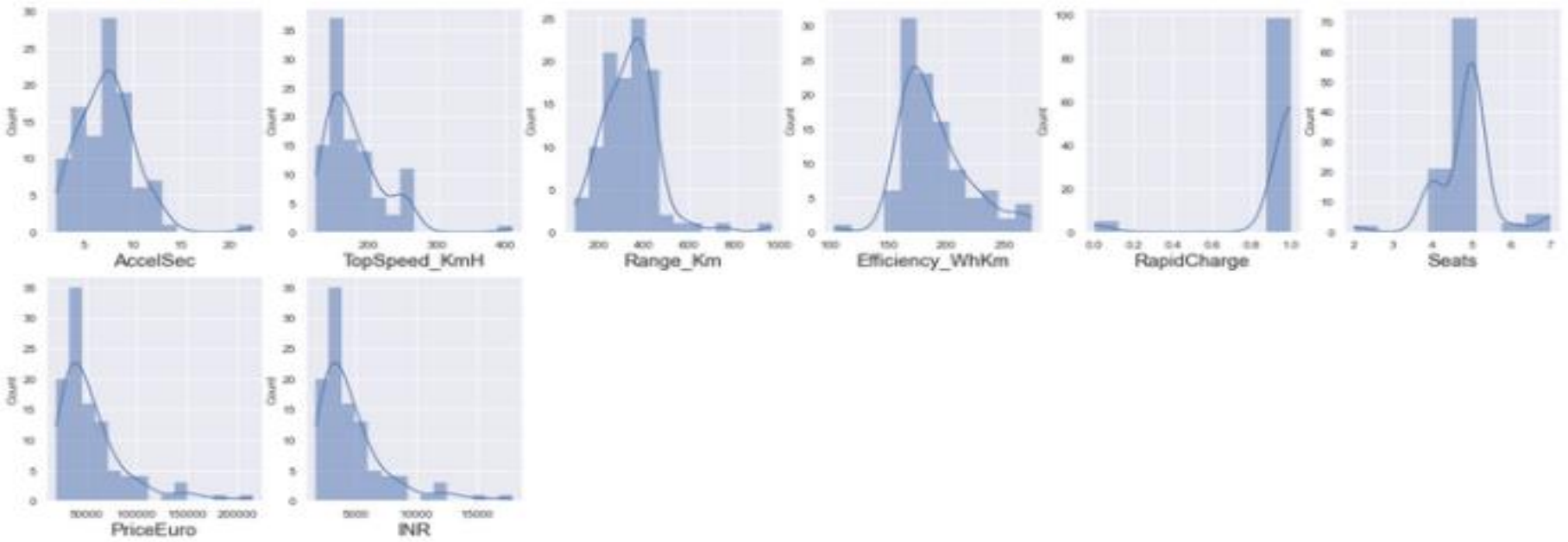
```
rng_dt = plt.figure(figsize=(10,5))
sns.barplot(data=df,x='PowerTrain',y='Range_Km')
plt.title("Drive Train Affect on Range")
plt.xlabel("DriveTrain")
plt.ylabel("Range(Km)")

```

Text(0, 0.5, 'Range(Km)')

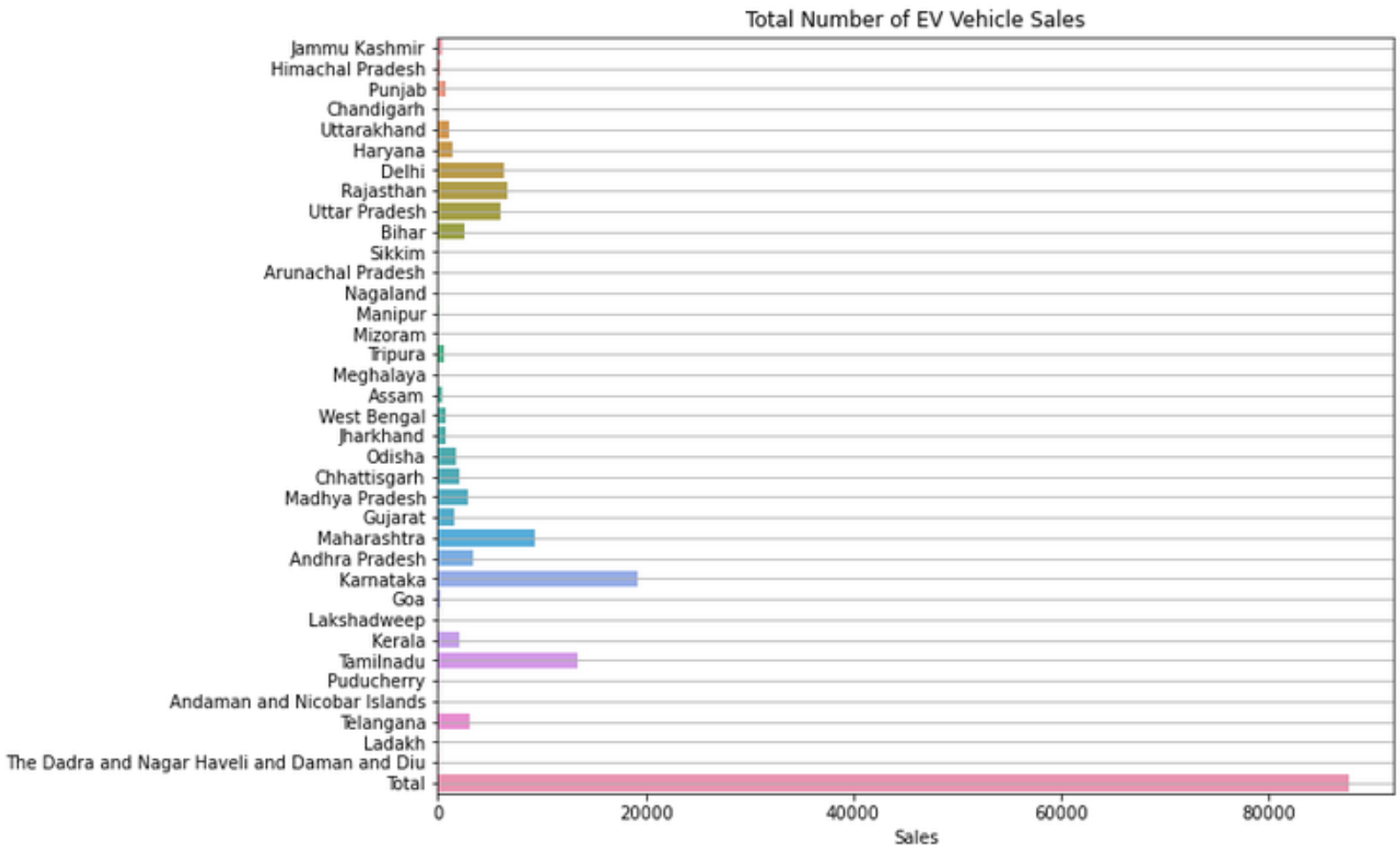


Checking for normal distribution of data

```
plt.figure(figsize=(25, 10))
col = 1

for i in df.columns:
    if df[i].dtype == "object":
        continue
    elif col<14:
        plt.subplot(2, 6, col)
        sns.histplot(df[i], kde=True, linewidth=0)
        plt.xlabel(i, y=-0.1, fontsize=18)
    col=col+1
```



Now we can see that the requirements of the EV Car Sales in every state in India in recent 5 years.



Screen Plot

It is a common method for determining the number of PCs to be retained via graphical representation. It is a simple line segment plot that shows the eigenvalues for each individual PC. It shows the eigenvalues on the y-axis and the number of factors on the x-axis. It always displays a downward curve. Most screen plots look broadly similar in shape, starting high on the left, falling rather quickly, and then flattening out at some point. This is because the first component usually explains much of the variability, the next few components explain a moderate amount, and the latter components only explain a small fraction of the overall variability. The screen plot criterion looks for the “elbow” in the curve and selects all components just before the line flattens out. The proportion of variance plot: The selected PCs should be able to describe at least 80% of the variance.

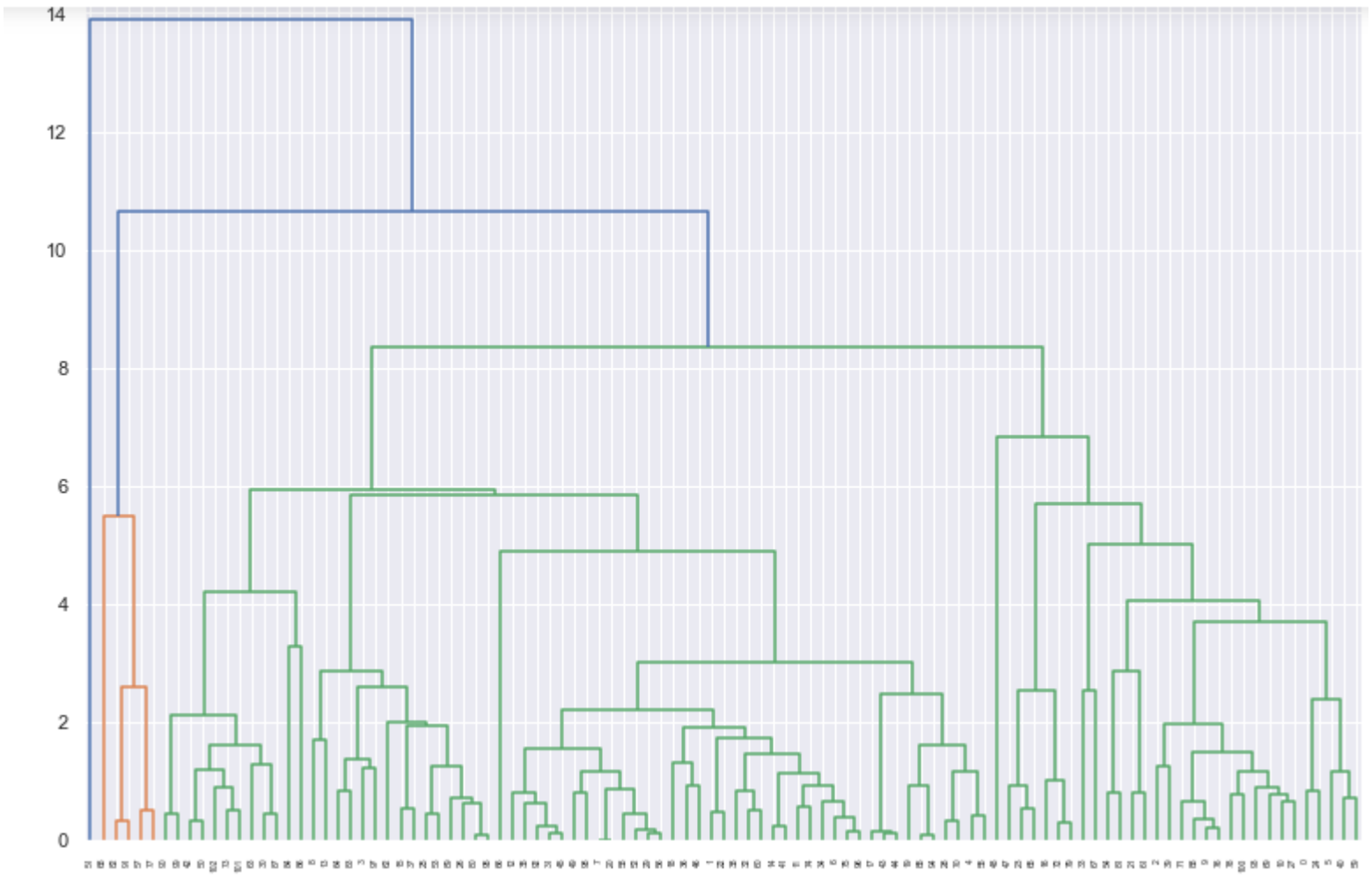
Screenplot

```
PC_values = np.arange(pca.n_components_) + 1
plt.plot(PC_values, pca.explained_variance_ratio_, 'o-', linewidth=2, color='blue')
plt.title('Scree Plot')
plt.xlabel('Principal Component')
plt.ylabel('Variance Explained')
plt.show()
```

Extracting Segments

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 choose 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.



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

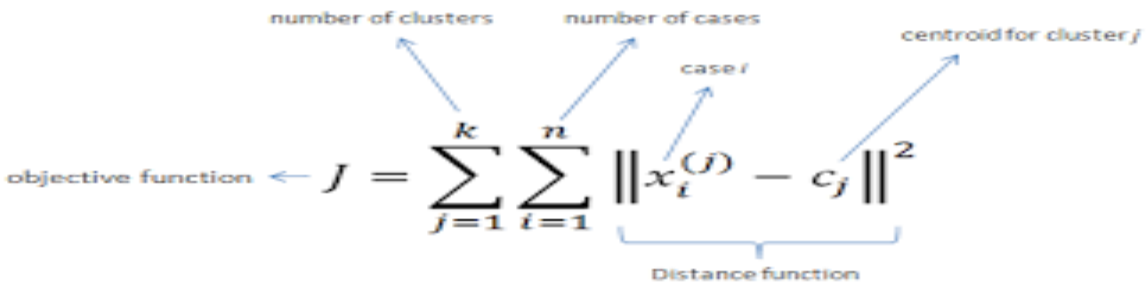
K Means algorithm is an iterative algorithm that tries to partition the dataset into pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (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 homogenerous (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:



And M-step is :

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j \|x^{(i)} - \mu_j\|^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

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.

K-Means Clustering Algorithm

In [21]:
```
#From the heatmap we can understand that from the features of an EV the range and top speed of a car has the highest positive coorela
x = df.iloc[:,[3,4]]
x
```

Out[21]:

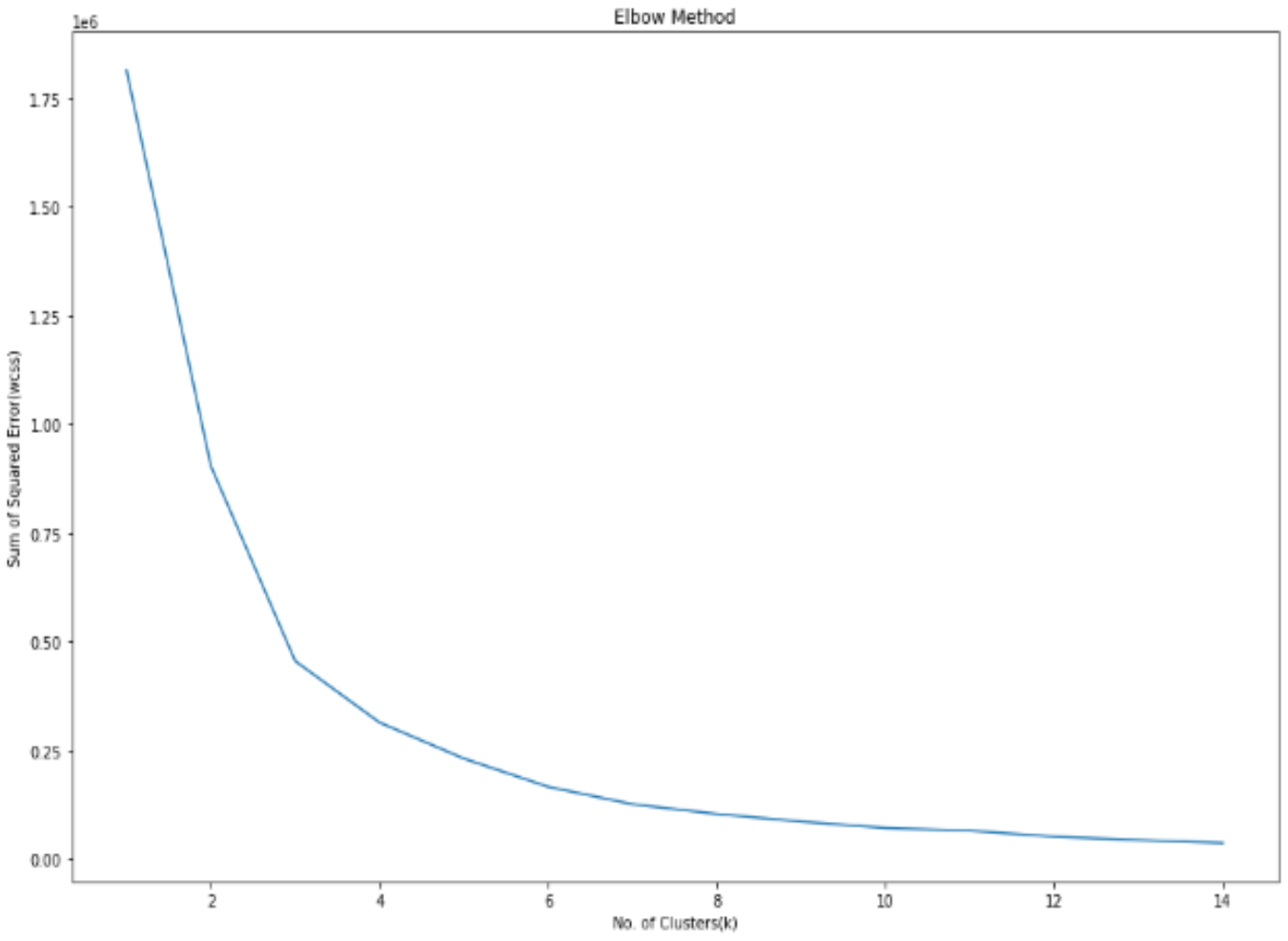
	Top$peed_KmH	Range_Km
0	233	450
1	160	270
2	210	400
3	180	360
4	145	170
...	...	...
98	160	330
99	210	335
100	200	325
101	200	375
102	190	400

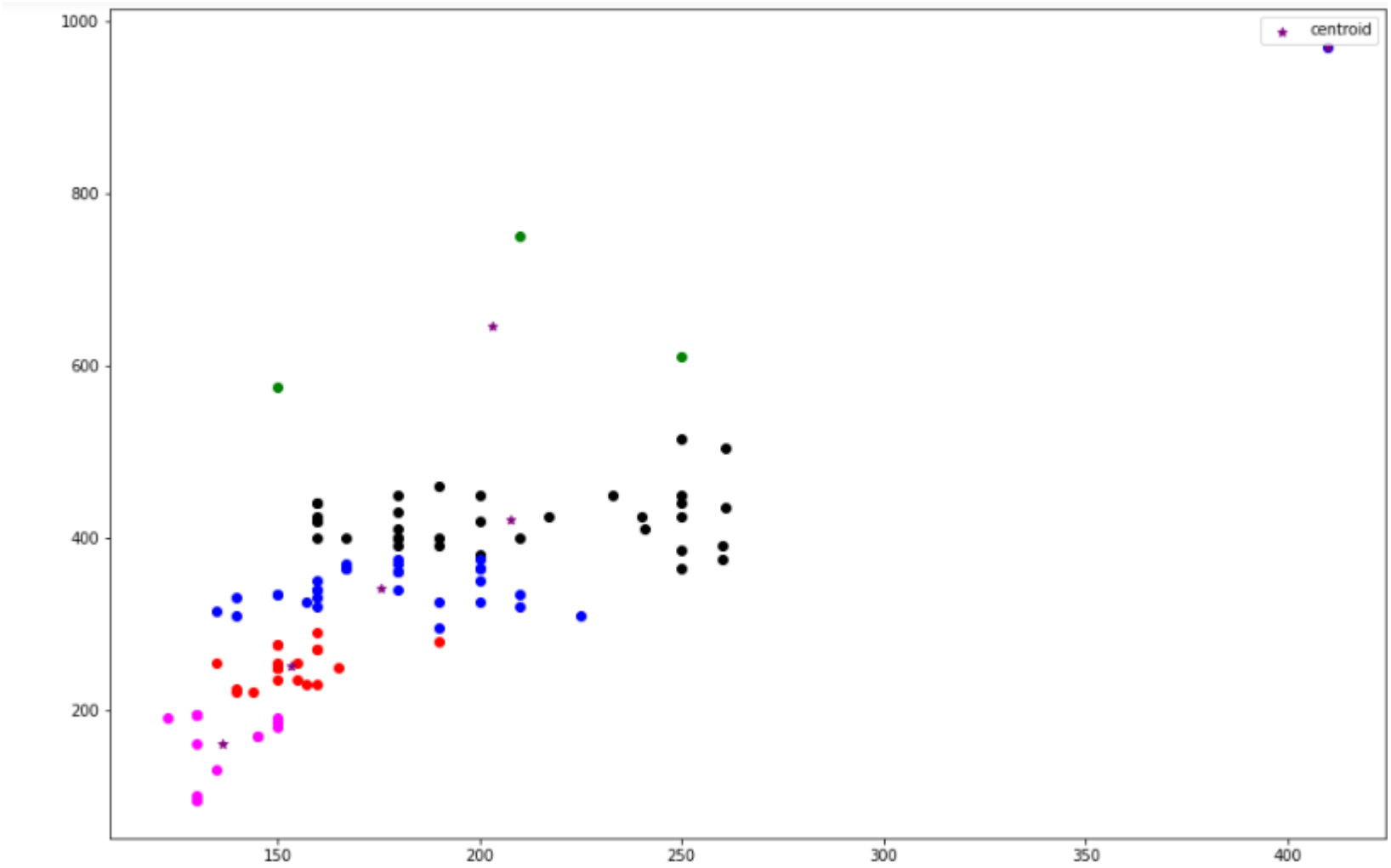
103 rows × 2 columns

In [22]:
```
from sklearn.cluster import KMeans
```

In [23]:
```
wcss = []
k= np.arange(1,15)
for i in k:
    kmeans = KMeans(n_clusters=i, init='k-means++', random_state=45)
    kmeans.fit(x)
    wcss.append(kmeans.inertia_)


```

In [24]:
```
plt.figure(figsize=(15,10))
plt.plot(k,wcss)
plt.title("Elbow Method")
plt.xlabel("No. of Clusters(k)")
plt.ylabel("Sum of Squared Error(wcss)")
plt.show()

```





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(Xtrain,ytrain)

command is used to fit the data set into model. The values of intercept, coefficient, and cumulative distribution function (CDF) are describ

Regression for PCA(Data2)

In [273]:
```
X=data2[['PC1', 'PC2','PC3','PC4','Pc5','PC6', 'PC7','PC8','PC9']]
y=df['INR']
```

In [277]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.3, random_state=21)
ls=LinearRegression().fit(X_train,y_train)
```

In [278]:
```
ls.intercept_
```

Out[278]:
```
4643.522050485438
```

In [279]:
```
ls.coef_
```

Out[279]:
```
array([1061.42554674,  920.77714129,  190.08894001, -264.41258254,
        185.54152362, 1652.17431742, -784.78832366, 1073.55879753,
       1172.29994606])
```

In [282]:
```
cof=pd.DataFrame(ls.coef_, X.columns, columns=['Coeff'])
cof
```

Out[282]:

	Coeff
PC1	1061.425547
PC2	920.777141
PC3	190.088940
PC4	-264.412583
Pc5	185.541524
PC6	1652.174317
PC7	-784.788324
PC8	1073.558798
PC9	1172.299946

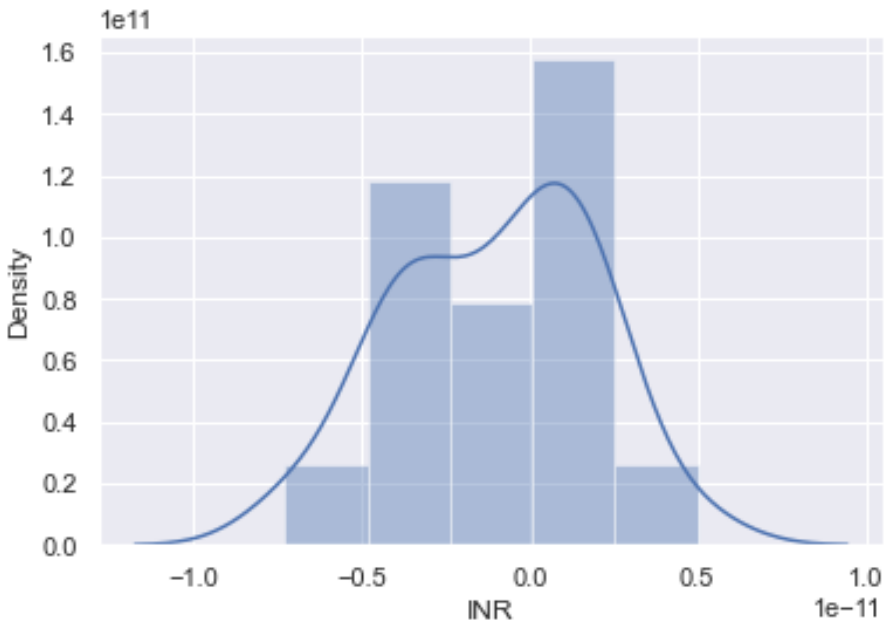
In [283]:
```
predict=ls.predict(X_test)
predict
```

Out[283]:
```
array([ 3486.5792, 10400.    ,  2858.8352,  4877.184 ,  5408.    ,
        2999.9424,  2862.08  ,  5233.28  , 17888.    ,  2043.808 ,
        3744.    ,  4160.    ,  2432.2688,  6655.168 ,  3243.7184,
```

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.

```
In [285]:  1  sb.distplot((y_test-predict))
```

```
Out[285]: <AxesSubplot:xlabel='INR', ylabel='Density'>
```



The metrics of the algorithm, Mean absolute error, Mean squared error and mean square root error are described in the below figure:

```
In [289]:  1  from sklearn import metrics
```

```
In [291]:  1  print('MAE:',metrics.mean_absolute_error(y_test,predict))
           2  print('MSE:',metrics.mean_squared_error(y_test,predict))
           3  print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test,predict)))
```

```
MAE: 2.5011104298755527e-12
MSE: 9.46754906746681e-24
RMSE: 3.0769382618874253e-12
```

Target Segments

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

Behavioral:

Mostly from our analysis there are cars with 5 seats.

Demographic:

- Top Speed & Range: With a large area of market the cost is dependent on Top speeds and Maximum range of cars.
- 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.

Geographical:

It depends on the number of Electric vehicle sale in recent 4-5 years in every state of India.

Customizing the Marketing Mix

The marketing mix refers to the set of actions, or tactics, that a company uses to promote its brand or product in the market. The 4Ps make up a typical marketing mix -Price, Product, Promotion and Place.

- 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 – In this case, the service. 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. Retailers pay a premium for the right location. In fact, the mantra of a successful retail business is 'location, location, location'.
- Promotion: this refers to all the activities undertaken to make the product or service known to the user and trade. This can include advertising, word of mouth, press reports, incentives, commissions and awards to the trade. It can also include consumer schemes, direct marketing.

All the elements of the marketing mix influence each other. They make up the business plan for a company and handle it right, and can give it great success. The marketing mix needs a lot of understanding, market research and consultation with several people, from users to trade to manufacturing and several others.

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.

GitHub Link

1. https://github.com/pradnya2613/Electric-Vehicle-Market-in-India
2. https://github.com/Patelraj8694/EV_Segmentation
3. https://github.com/Adityakashyap9569/EV-vehicle
4. https://github.com/OsamaMrBean/EV-market-segmentation
5. https://github.com/itsjacobjoy/EV-Market-Segmentation