Team-Devendra

Electric vehicle segmentation Analysis

Team Members:

1. Ashish Kumar
2. Cherukuri Shraavan Sai Ram
3. Devendra Deshmane
4. Muhammad Asif Jaleel
6. Parag Jadhav
7. U Vijay Nikhil

Github link: devd1808/devd1808-Electric-vehicle-segmentation-Analysis (github.com)

Background

The electric vehicle industry in India is a growing industry. Electric Vehicles include a large range of vehicles from electric two - wheelers, three - wheelers (rickshaws), cars and electric buses. An electric vehicle can be classified on the basis of their attributes such as charging time, driving range, and the maximum load it can carry. Of these attributes, the two most important characteristics of an electric vehicle of concern to the consumer are: Driving range(i.e. Maximum distance an EV can run when fully charged) Charging time of batteries(i.e. the time required to fully charge the battery) and Charging time depends on the input power characteristics (i.e. input voltage and current), battery type and battery capacity. Therefore, such a vehicle is seen as a possible replacement for current-generation automobiles, in order to address the issue of rising pollution, global warming, depleting natural resources, etc

Problem Statement:

You are a team working under an Electric Vehicle Startup. The Startup is still deciding in which vehicle/customer space it will be develop its EVs. You have to analyse the Electric Vehicle market in India using Segmentation analysis and come up with a feasible strategy to enter the market, targeting the segments most likely to use Electric vehicles.

Fermi Estimation:

The main task is to find out the vehicle space in which the EV startup will develop its products. For this, we use a dataset which has the preferred car characteristics of the customers in the

market. As we are using the preferences and interests of the customers and also considering the frequency of purchase (as it is selling well in the market) and amount spent on purchasing, we can say that the segmentation criterion is psychographic and behavioural segmentation. And we use machine learning algorithms for segmentation of the dataset, after we have done with profiling it, we can then find the most optimal market segments to open in the markets.

Data Sources:

We have selected a dataset with instances of most popular cars in the market, we have selected it because considering the characteristics of best-selling cars in the market implies the preferences and interests of the people, that is, if the cars of such and such characteristics are selling widely it means that it is much preferred by the customers. And we can make the model of our electric car according to the preferences of the customers in the market segment. So, basically the segmentation criterion for the market segment analysis is psychographic (because of the preferences of the customers) and behavioural segmentation (because the best-selling electric cars (frequency) and amount spent on purchasing). And we made sure that there are enough instances in the dataset, so that there are enough instances in each segment after performing segmentation. We have chosen dataset with 14 attributes considering characteristics of the best-selling cars in the market, the attributes are:

- Brand
- Model
- Acceleration per sec
- Top Speed
- Range (int Km)
- Efficiency (Watt-hour per km)
- Fast Charge
- Rapid Charge
- Power Train
- Plug Type
- Body Style
- Segment
- Seats
- Price (in Euro)

Categorical attributes: Brand, Model, Rapid Charge, Power Train, Plug Type, Body Style, Segment

Numerical attributes: Acceleration per sec, Top Speed, Range (in Km), Efficiency, Fast Charge, Seats, Price (in Euro).

Classes of Categorical Variables:

Brand and Model will be removed, this will be later explained in data pre-processing
Rapid Charge: {Yes, No}
Power Train: {AWD, RWD, FWD}
Plug Type: {Type 2 CSS, Type 2, Type 2 CHAdeMO, Type 1 CHAdeMO}
Body Style: {Sedan, Hatchback, Liftback, SUV, MPV, Cabrio, Pickup, SPV}
Segment: {A, B, C, D, E, F, N, S}

Type of Numerical Variables:

Acceleration per sec: float
Top Speed: int
Range (in Km): int
Efficiency (Watt-hour per km): int
Fast Charge: float
Seats: int

Exploratory Data Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

So, we perform EDA on this dataset, to understand the dataset and draw rough estimations from the data which might be further useful in pre-processing step.

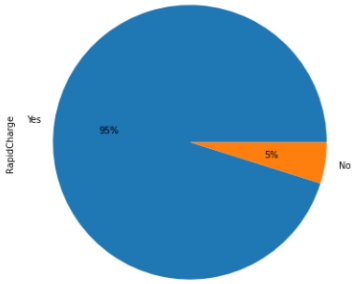
First we use unique() function for each categorical attribute , to find unique classes for each attribute. So that we will be able to understand what type of data is present in the dataset.

And then we use seaborn library for pair plot, this helps us the analyse the rough relation between 2 attributes. Some observations based on the pair plots (roughly):

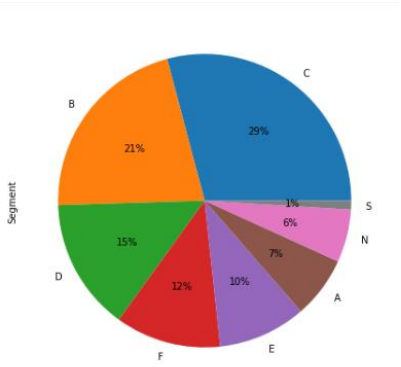
- TopSpeed and AccelSec has linear relation with negative correlation
- TopSpeed and PriceEuro has polynomial relation with probably degree 2
- FastCharge and Range has a linear relation with positive correlation
- Efficency vs other attributes plots has no specific pattern which indicates no relation between efficiency and other attributes

Plotting Pie-Charts for Categorical Variables:

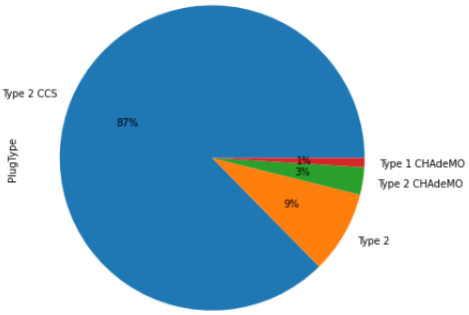
After plotting a pie chart for RapidCharge, we can see that disproportionate number of the instance values for these attributes is of one class which is Yes, so this does not provide us with much information. Hence, we can remove it from the dataset.



And then we plot pie chart for the attribute Segment, we can see that Segment C is the majority class. And all other classes are pretty much evenly distributed.



Then, we plot a pie chart for the attribute PlugType, and we can see more than 80 % of the instances is of class Type 2 CCS, Hence as disproportionate number of the instances is of one class, it provides us with a minimal information and we can safely remove it.

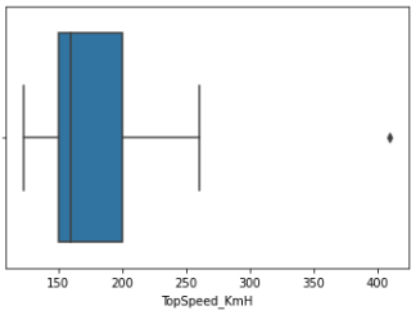


Plotting Bar Plot for numerical variable:

To have a visualization of numerical data, we basically convert a range of numerical values into a categorical classes and then we plot a bar plot, where x axis is the categorical classes that we have synthesized and y axis is the frequency of the particular class in the attribute.

In this dataset, we convert each numerical attributes into three categorical values (Low, Medium and High). And for deciding which range of values should be chosen for each categorical value, we take the help of box plot.

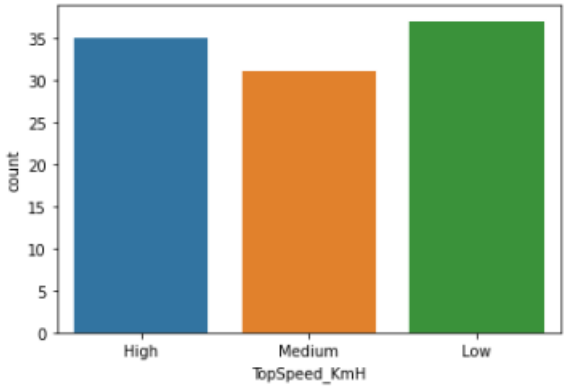
For instance, take the attribute TopSpeed_KmH, after plotting box plot for it.



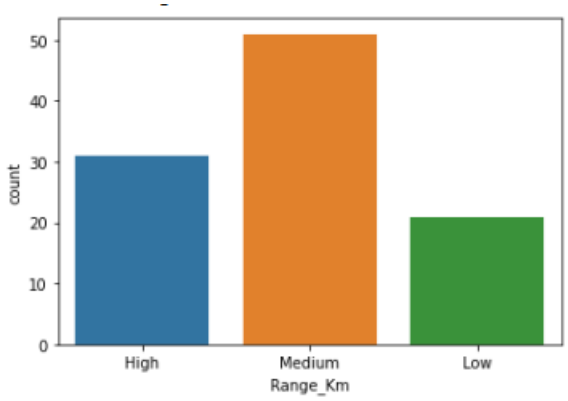
From the above box plot we can analyse assign the following:
less than 160: Low

between 160 and 190: Medium
greater than 190: High

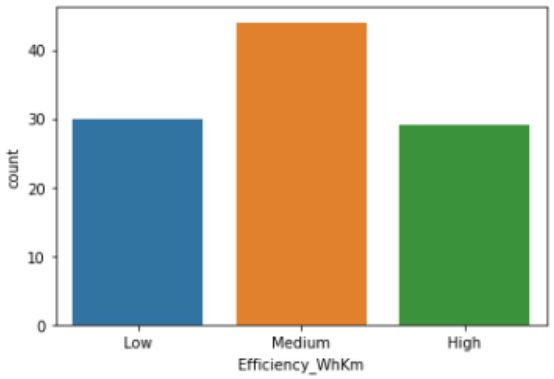
Based on this we can convert the instances of numerical attribute to a list of values with categorical values. From Bar plot we can see that the values in each ranges are pretty evenly distributed and number of values in the class Low (<160) are slightly higher. The bar plot for this processed attribute can be seen below:



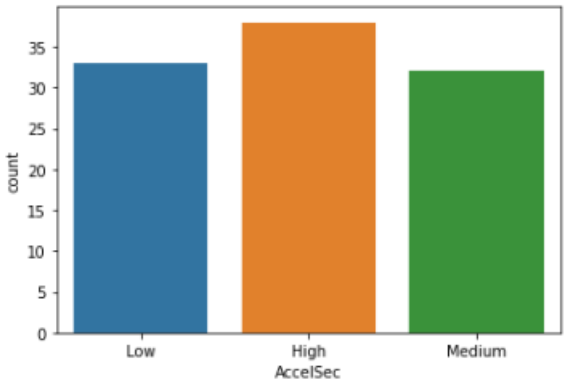
The similar thing can be done for numerical attributes such as Range_Km, Efficiency_WhKm and AccelSec, and rough conclusions can be drawn based on each bar plot.
The bar plot for Range_Km is (less than 250: Low; between 250 and 400: Medium; greater than 400: High):



The bar plot for Efficiency_Whkm is (less than 170: Low; between 170 and 200: Medium; greater than 200: High):



The bar plot for AccelSec is (less than 6: Low; between 6 and 8: Medium; greater than 8: High):



Data Pre-processing:

The packages used for Data Pre-processing are:

pandas - where we use drop(), to remove irrelevant or redundant attributes in the dataset
- where we use apply(), to apply normalization formula on the numerical attributes
- where we use get_dummies(), to convert categorical variables into numerical variables

Removing irrelevant attributes:

We remove the irrelevant attributes in the dataset such as Brand and Model. As this is an Electric Vehicle start-up, already existing brand and model of the good selling cars should not be considered in as a segmentation variable. On the other hand, the characteristics of these cars should be considered as those are the reason why the particular car is doing well in the market. So , we remove the attributes Brand and Model ,because these cannot be included or used in anyway, for developing the product.

Handling the missing values:

We have missing values present in FastCharge_KmH, this makes the attribute FastCharge_KmH to be wrongly considered as a categorical variable whereas it should be considered as float. So we take the mean of the attribute of the remaining instances where the values are not missing and then we replace the missing values with the mean calculated. And we make sure after that the FastCharge_KmH is considered as float only as intended.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103 entries, 0 to 102
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Brand            103 non-null    object
 1   Model            103 non-null    object
 2   AccelSec         103 non-null    float64
 3   TopSpeed_KmH     103 non-null    int64
 4   Range_Km         103 non-null    int64
 5   Efficiency_WhKm  103 non-null    int64
 6   FastCharge_KmH   103 non-null    object
 7   RapidCharge      103 non-null    object
 8   PowerTrain       103 non-null    object
 9   PlugType         103 non-null    object
 10  BodyStyle        103 non-null    object
 11  Segment          103 non-null    object
 12  Seats            103 non-null    int64
 13  PriceEuro        103 non-null    int64
dtypes: float64(1), int64(5), object(8)
memory usage: 11.4+ KB
```

Here we can see that in Dtype for FastCharge_KmH is object, that means it is considered as the categorical variable.

In below diagram we can see that '-' is considered as a value for the variable ,consequently considering as a categorical variable

```
df['FastCharge_KmH'].unique()
```

```
array(['940', '250', '620', '560', '190', '220', '420', '650', '540',
       '440', '230', '380', '210', '590', '780', '170', '260', '930',
       '850', '910', '490', '470', '270', '450', '350', '710', '240',
       '390', '570', '610', '340', '730', '920', '-', '550', '900', '520',
       '430', '890', '410', '770', '460', '360', '810', '480', '290',
       '330', '740', '510', '320', '500'], dtype=object)
```

After we have replaced the missing values with the mean of the attribute of remaining instances, we can see FastCharge_KmH is considered as float

```
df['FastCharge_KmH'].unique()
```

```
array([940.  , 250.  , 620.  , 560.  , 190.  , 220.  , 420.  , 650.  ,
       540.  , 440.  , 230.  , 380.  , 210.  , 590.  , 780.  , 170.  ,
       260.  , 930.  , 850.  , 910.  , 490.  , 470.  , 270.  , 450.  ,
       350.  , 710.  , 240.  , 390.  , 570.  , 610.  , 340.  , 730.  ,
       920.  , 434.56, 550.  , 900.  , 520.  , 430.  , 890.  , 410.  ,
       770.  , 460.  , 360.  , 810.  , 480.  , 290.  , 330.  , 740.  ,
       510.  , 320.  , 500.  ])
```

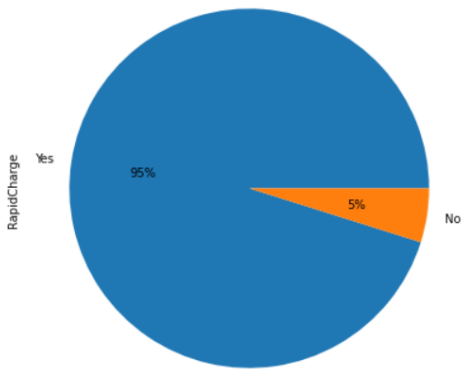
Removing redundant attributes:

PlugType and RapidCharge are considered as redundant attributes because they have number of one of the classes disproportionately higher than the remaining classes ,that is, the attributes are kind of imbalanced. In these cases, when one of the classes is present in majority of instances, there is no valuable information that this data can provide us as almost all the instances are having the same value.

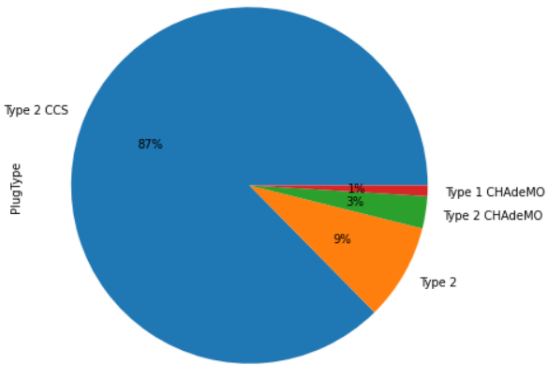
PlugType and RapidCharge are considered to be redundant variables from the results based on the EDA where the pie chart for each categorical variable is obtained.

From the below two pie charts for RapidCharge and PlugType we can see that the one sector (blue) is covering more than 80 % of the area of the circle which means that majority of the instances in that particular attribute is of one class, so we can eliminate these two attributes as they provide minimal information.

```
#Plotting a pie chart
plt.figure(figsize=[9,7])
df['RapidCharge'].value_counts().plot.pie(autopct='%.0f%%')
plt.show()
```




```
#Plotting a pie chart
plt.figure(figsize=[9,7])
df['PlugType'].value_counts().plot.pie(autopct='%.0f%%')
plt.show()
```



Converting categorical to numerical variable:

For the application of clustering algorithm, it is required for the dataset to have all numerical variables for better results. So we convert the categorical variables to numerical variables using one hot encoding. One hot encoding is the right choice because there exists no ordinal relation between the classes of each categorical attributes. In this, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to

those columns. In a converted attribute (from the categorical value), an instance is assigned 1, if it had that categorical value in the previous dataset and we give 0 if it is not the categorical value for the instance. To implement this we use get_dummies function in pandas library.

Normalization:

Normalization avoids raw data and various problems of datasets by creating new values and maintaining general distribution as well as a ratio in data.

Normalization Formula

$$X_{normalized} = \frac{(x - x_{minimum})}{(x_{minimum} - x_{minimum})}$$

We use the above normalization formula to normalize the numerical attributes of the dataset to bring them on the same scale. It also improves the performance and accuracy of machine learning models using various techniques and algorithms.

The numerical attributes on which we apply normalization are:

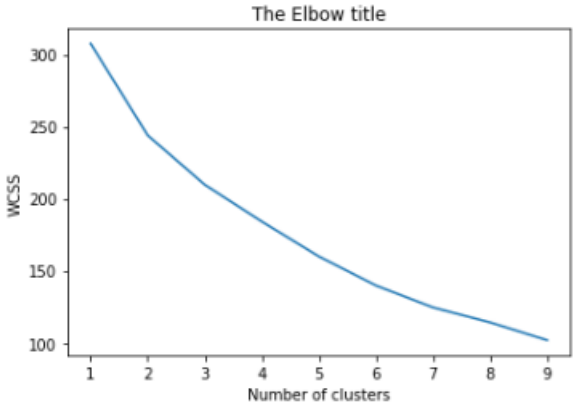
- AccelSec
- TopSpeed_KmH
- Range_Km
- Efficiency_WhKm
- FastCharge_KmH
- PriceEuro

Note: Before we normalize the dataset, we copy the dataset to a new variable called df_prev, because after we find which segment each instance belongs to we should combine the results with the dataset before normalization, to perform analysis based on clusters obtained from clustering machine learning algorithm.

Segmentation Extraction:

After pre-processing of the dataset we use machine learning clustering algorithm to divide the dataset into segments. In this case, we use **distance based** method for the extraction of segments, more specifically we use K means algorithm (partitioning clustering algorithm) to perform segmentation. Partitioning clustering methods such as K means divide these data into subsets (market segments) such that consumers assigned to the same market segment are as similar to one another as possible, while consumers belonging to different market segments are as dissimilar as possible. The representative of a market segment is referred to in many partitioning clustering algorithms as the centroid. For the *k*-means algorithm based on the squared Euclidean distance, the centroid consists of the column-wise mean values across all members of the market segment.
Now that we have decided the partitioning clustering that we use for the segment extraction, the further step is to determine the number of segments as the algorithm requires the specification of the number of segments. For this we systematically repeat the

extraction process for different numbers of clusters (or market segments), and then select the number of segments that leads to either the most stable overall segmentation solution, or to the most stable individual segment. We implement this by plotting a graph using Elbow method. In the Elbow method, we are actually varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease. WCSS value is largest when K = 1. When we analyze the graph we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph starts to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters. In this graph the K value could be 5 or 6, keeping the number of instances in mind, we can choose the minimum of the two that is we take K value as 5.



Now that we have found the optimal number of segments, We can perform segmentation using partitioning clustering algorithm (K means) with K (=5) segments.

After we cluster the given data into 5 segments, we can now create a separate attribute cluster containing the label of each cluster to the corresponding instance. As K value is 5, the labels of the attribute are {0,1,2,3,4}. And then we concatenate this attribute with df_prev which is the dataset pre-processed but just before normalization. After performing the clustering algorithm, we can see how many instances are present in each cluster.

```
data_with_clusters['Clusters'].value_counts()
```

```
3    29
4    24
1    19
0    16
2    15
Name: Clusters, dtype: int64
```

Now we that we are done with segment extraction, we move onto next step which is profiling, that is we analyze attributes of instances in each segment and try to figure out their common characteristics.

Profiling and Describing Potential Segments:

In profiling the segments, we try to analyse the attributes of instances in each of the segments and find the common characteristics of the data points within the clusters.
In each segment, for numerical attributes we find the mean of that particular attribute and maximum and minimum value in that attribute, to estimate the range of the values in the particular segment. And for categorical attributes, we find the mode of classes in the attribute of each segment and generalize the characteristic of the segment as the mode. And if we find that two or more classes of an attribute are present in a segment are at almost same frequency, then we consider all the classes which have almost similar frequencies to that of the mode class.

Segment 0:

Number of instances in segment 0 is 16

Numerical Value Summary:

Segement 0	Range of Values	Average value
AccelSec	2.1 - 10	4.03
Top Speed (KmH)	150 - 410	247.5
Range (Km)	310 - 970	474
Efficiency (WhKm)	104 - 223	184.1
Fast Charge (KmH)	540 - 940	736.9
Price Euro	46380 - 215000	109193
Seats	4 and 5 seater	4.56

Categorical Value Summary:

Segment 0	Mode Class(s)
Power Train	AWD
Body Style	Sedan
Segment	F

From the above summary we can see the range of values of each numerical attribute. And we can see that seats in segment 0 is 4 and 5 seater and as the average is 4.56 we can infer that the number of 5 seaters is slightly greater than 4 seaters in segment 0. And EV's in segment 0 has Power Train of AWD, Body Style of Sedan and belongs to segment F.
Note:

F-class vehicles are known to be vehicles with higher performance, better handling, comfort, more top quality equipment and an stupendous design.

Segment 1:

Number of instances in segment 1 is 19

Numerical Value Summary:

Segement 1	Range of Values	Average value
AccelSec	7.3 – 22.4	9.92
Top Speed (KmH)	123 - 167	147.68
Range (Km)	160 - 400	278.15
Efficiency (WhKm)	154 - 273	178.6
Fast Charge (KmH)	190 - 435	308.7
Price Euro	29146 - 70631	36443

Seats	4 – 7 seater	5.15

Categorical Value Summary:

Segment 1	Mode Class(s)
Power Train	FWD
Body Style	SUV
Segment	B

From the above summary we can see the range of values of each numerical attribute. And we can see that seats in segment 1 is 4,5,6 and 7 seater and as the average is 5.15 we can infer that the number of 4 and 5 seaters is greater than remaining seaters in segment 1. And EV's in segment 1 has Power Train of FWD, Body Style of SUV and belongs to segment B.
Note:
Segment B SUV : These are cars that have more street presence than a similarly priced sedan or hatchback. However, they aren't too rugged or even too powerful.

Segment 2:
Number of instances in segment 2 is 15
Numerical Value Summary:

Segement 2	Range of Values	Average value
AccelSec	6.5 – 12.7	9.86
Top Speed (KmH)	130 - 160	142.33
Range (Km)	95 - 440	207.33
Efficiency (WhKm)	156 - 181	168.2
Fast Charge (KmH)	170 - 590	327.2
Price Euro	20129 - 45000	30655.6
Seats	2 - 4 seater	3.73

Categorical Value Summary:

Segment 2	Mode Class(s)
Power Train	RWD
Body Style	Hatch Back
Segment	A, B

From the above summary we can see the range of values of each numerical attribute. And we can see that seats in segment 2 is 2,3 and 4 seater and as the average is 3.73 we can infer that the number of 4 seaters is greater than remaining seaters in segment 1. And EV's in segment 2 has Power Train of RWD, Body Style of Hatch Back and belongs to segment A, B.

Note:
Segment A: A-segment cars are budget-oriented and hence they are small in size and less on features. These have small engines which are efficient.
Segment B: B-segment cars or small hatchbacks are best suited for you if your driving requirements are daily commuting and occasional long drives. They are slightly larger than A-segment vehicles and offer more stability on the road.

Segment 3:
Number of instances in segment 3 is 29

Numerical Value Summary:

Segement 3	Range of Values	Average value
AccelSec	2.8 – 7.5	5.64
Top Speed (KmH)	160 - 250	195.44
Range (Km)	280 - 750	393.96
Efficiency (WhKm)	171 - 270	217.48
Fast Charge (KmH)	340 - 930	534.14
Price Euro	45000- 102990	65418.96
Seats	5 - 7 seater	5.37

Categorical Value Summary:

Segment 3	Mode Class(s)
Power Train	AWD
Body Style	SUV
Segment	D, E

From the above summary we can see the range of values of each numerical attribute. And we can see that seats in segment 3 is 5,6 and 7 seater and as the average is 5.37 we can infer that the number of 5 seaters is greater than remaining seaters in segment 3. And EV's in segment 3 has Power Train of AWD, Body Style of SUV and belongs to segment D, E.
Segment D: Mid-sized family car where the interior and plethora of luxury items take precedence over the powertrain and drivetrain, making them more difficult to maneuver.
Segment E: The E stands for Executive luxury cars. They are much longer than mid-size sedans. E-segments cars in India include some massive and large vehicles and some opulent rides with a long wheelbase. These are known among the business class since they start with the letter E and exude excellence and luxury.

Segment 4:
Number of instances in segment 4 is 24
Numerical Value Summary:

Segement 4	Range of Values	Average value
AccelSec	5.1 - 10	8.2
Top Speed (KmH)	140 - 200	162
Range (Km)	180 - 440	312
Efficiency (WhKm)	153 - 232	179.8
Fast Charge (KmH)	190 - 560	370
Price Euro	25500 - 65000	39670.8
Seats	5 seater	5

Categorical Value Summary:

Segment 4	Mode Class(s)
Power Train	FWD, RWD
Body Style	Hatch Back , SUV
Segment	C

From the above summary we can see the range of values of each numerical attribute. And we can see that seats in segment 4 is only 5 seater. And EV's in segment 4 has Power Train of FWD,RWD, Body Style of SUV and Hatch Back and belongs to segment C.
Segment C:These are described as 'medium cars'. C-segment category cars are good balance between the interior space and compact exterior dimensions
We can see that of the above 5 segments, segment 3 and 4 are the potential target segments. In segment 3, we can see that most are segment D and E, these segment include mid-sized family cars to luxury cars. The price in euros for this segment is in the range of 45000- 102990. And mostly include cars with 5 seaters.
In segment 4, we can see that most are of segment C, these are described as medium cars or family cars. The price in Euros for this segment is in the range of 25500 -65000, these cars are recommended for people with reasonable budget, not much expensive compared to segment 3 cars. And most cars include 5 seaters.

Target Segment:

The target segment which we choose from the 5 segments is segment 3. This segment consists of 29 instances. The summary of the segment 3 is:
Numerical Value Summary:

Segement 3	Range of Values	Average value
AccelSec	2.8 – 7.5	5.64
Top Speed (KmH)	160 - 250	195.44
Range (Km)	280 - 750	393.96
Efficiency (WhKm)	171 - 270	217.48
Fast Charge (KmH)	340 - 930	534.14
Price Euro	45000- 102990	65418.96
Seats	5 - 7 seater	5.37

Categorical Value Summary:

Segment 3	Mode Class(s)
Power Train	AWD
Body Style	SUV
Segment	D, E

The Acceleration range of these cars is 2.8 to 7.5 m/s^2. And top speed of these cars in segment 3 is 160 to 250 KmH. And efficiency of the cars in this segment is in the range 171 to 270 Whkm. And the price range in this segment is from 45,000 euros to 102,990 euros. And most preferred seats in this segment is 5 seater. And Power train of these segment car is AWD, an all-wheel drive vehicle (AWD vehicle) is one with a powertrain capable of providing power to all its wheels, whether full-time or on-demand. For the cars in this segment the body style is SUV (A sport utility vehicle or SUV is a car classification that combines elements of road-going passenger cars with features from off-road vehicles, such as raised ground clearance and four-wheel drive). And cars here belongs to segment D and E, which means these include mid-sized family cars to luxury cars.

The most optimal market segment to open in the market is segment 3, that is, if we manufacture cars with acceleration in the range of 2.5 to 7.8 m/s^2, and top speed in the range of 160 to 250 KmH we can be more profitable in the market. And cars with powertrain AWD, and SUV body style also seems to increase the profits. Throughout the analysis, it seems that 5 seaters are the most preferred seats in the market. So with this vehicle characteristics, it seems we can be more profitable in the market.

Customizing the Market 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 costs of production, segment targeted, ability of the market to pay, supply - demand and a host of other direct and indirect factors. There can be several types of pricing strategies, each tied in with an overall business plan.

Product: Refers to the item actually being sold. The product must delivera 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 rightlocation. 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, contests and prizes.

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.