Electric Vehicle Market in India

Market Segmentation

Harshit Jain

Abstract

Market segmentation becomes a vital strategy for emerging markets to investigate and use for extensive adoption of emerging mobility technologies like electric vehicles (EVs). As a low emission and low operating cost vehicle, EV adoption is anticipated to increase drastically soon. As a result, it will stimulate a significant amount of future academic study interest. By utilising an integrated research framework of "perceived benefits-attitude-intention," the primary goal of this study is to examine and identify several sets of possible customer segments for EVs based on psychographic, behavioural, and socio-economic characterisation. To operationalize and validate segments from the data gathered from 563 respondents via a cross-sectional online survey, the study used rigorous analytical techniques like cluster analysis, multiple discriminant analysis, and the Chi-square test. The findings posit that the three distinct sets of young consumer groups have been identified and labelled as *'Conservatives', 'Indiffer- ents', and 'Enthusiasts'* which are deemed to be buddying EV buyers The implicationsare recommended, which may offer some pertinent guidance for scholars and policy- makers to encourage EVs adoption in the backdrop of emerging sustainable transport market.

In this report we are going to analyse 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, SocioDemographic, 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, aspirations, 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 this criteria is not the cause for customers product preferences then it does not provide sufficient market insight for optimal segmentation decisions

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.

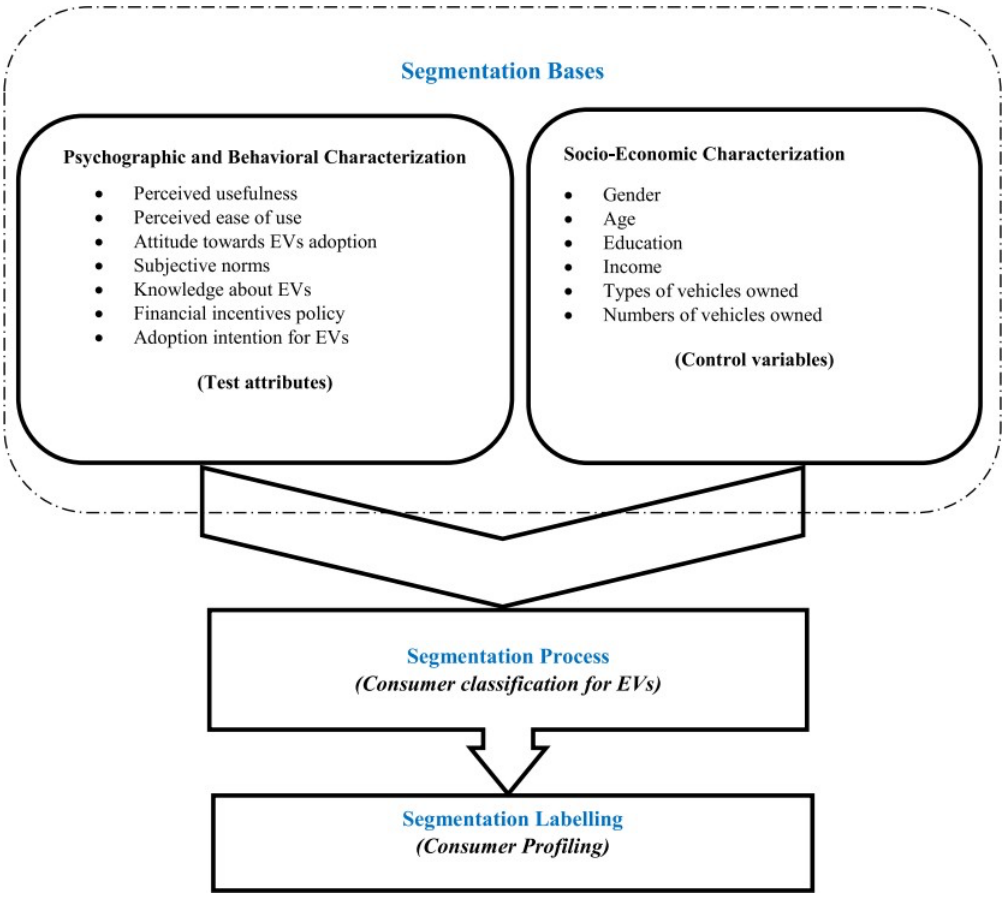


Figure 4: *Market Segmentation Electric Vehicles*

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:

1. **Numpy:** To calculate various calculations related to arrays.
2. **Pandas:** To read or load the datasets.
3. **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 [12]:
```
df = pd.read_csv('data.csv')
df.drop('Unnamed: 0', axis=1, inplace=True)
df['inr(10e3)'] = df['PriceEuro']*0.08320
df['RapidCharge'].replace(to_replace=['No','Yes'],value=[0, 1],inplace=True)
df.head()
```

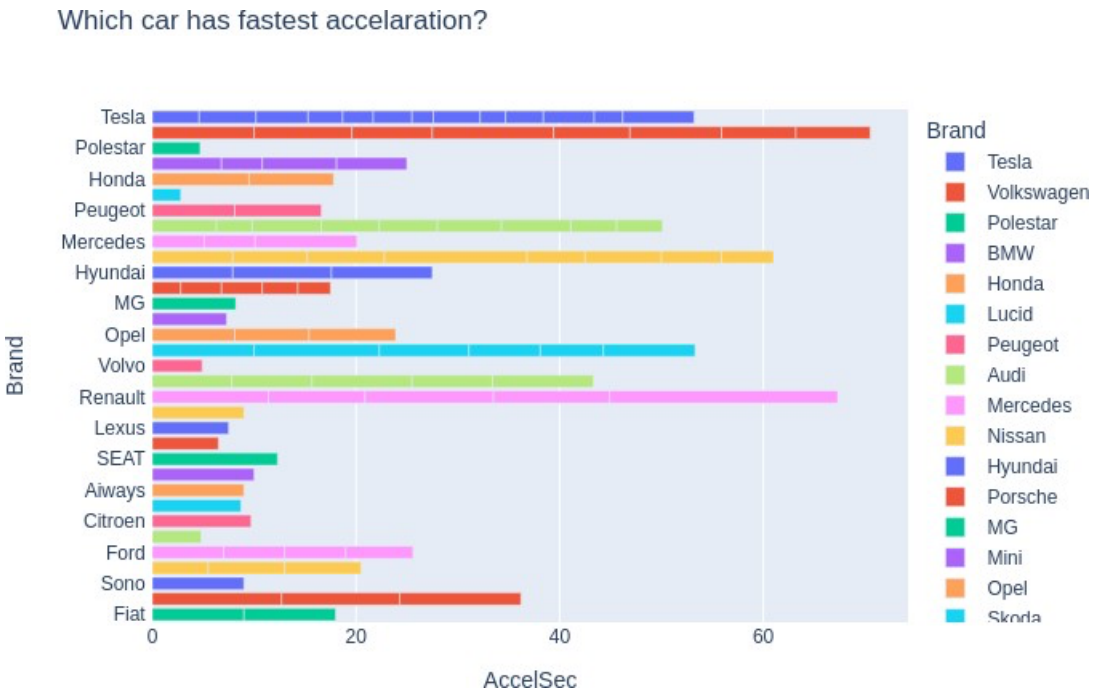
Out[12]:

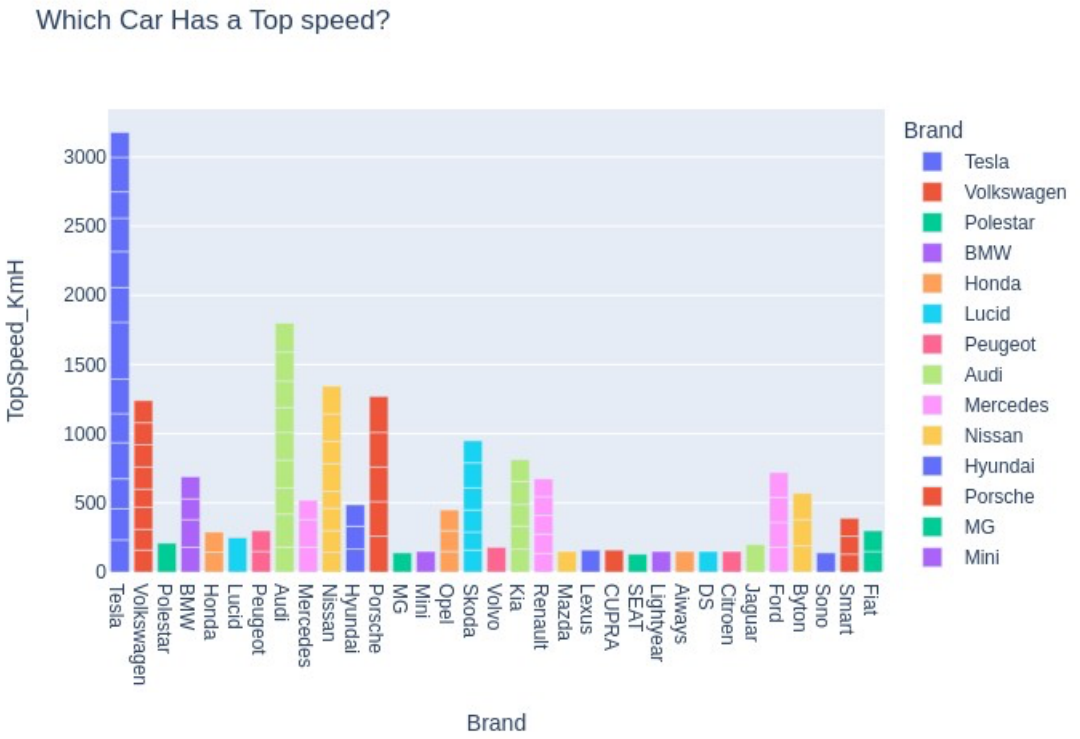
	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment	Se
0	Tesla	Model 3 Long Range Dual Motor	4.6000	233	450	161	940	1	AWD	Type 2 CCS	Sedan	D	
1	Volkswagen	ID.3 Pure	10.0000	160	270	167	250	0	RWD	Type 2 CCS	Hatchback	C	
2	Polestar	2	4.7000	210	400	181	620	1	AWD	Type 2 CCS	Liftback	D	
3	BMW	iX3	6.8000	180	360	206	560	1	RWD	Type 2 CCS	SUV	D	
4	Honda	e	9.5000	145	170	168	190	1	RWD	Type 2 CCS	Hatchback	B	

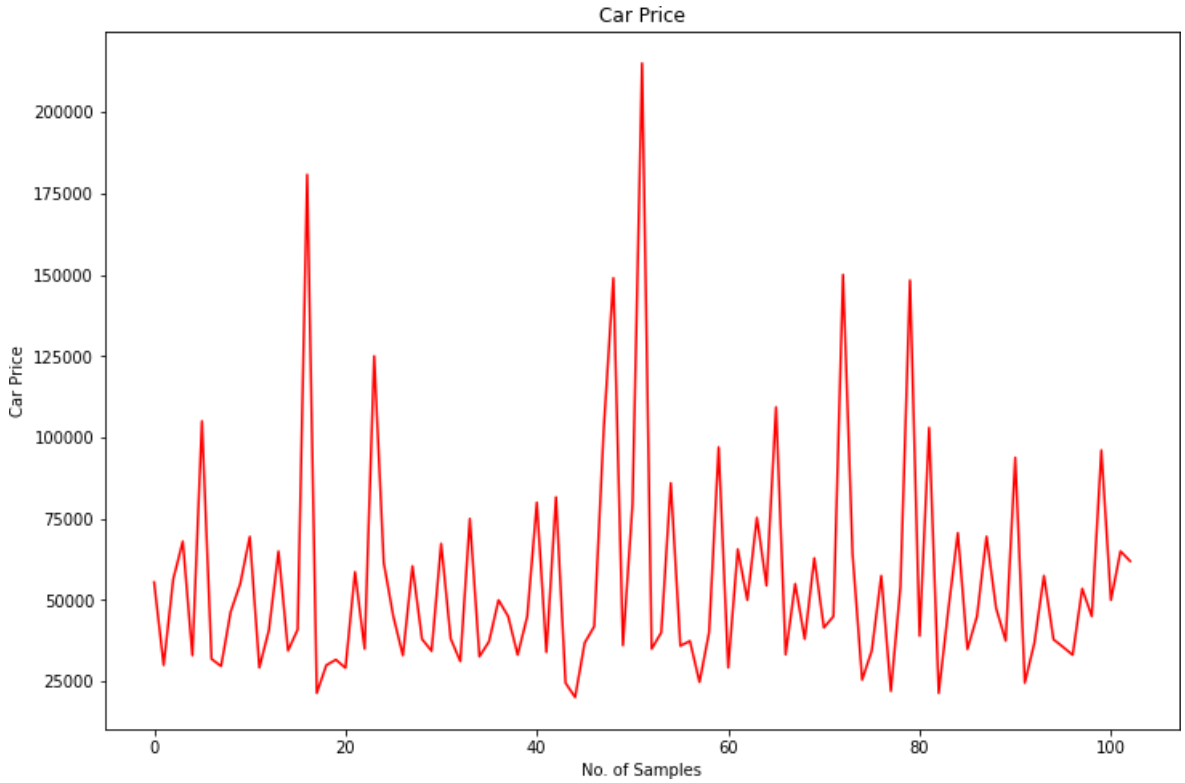
EDA

Beginning with some data analysis without principal component analysis and some principal component analysis in the dataset created by combining all of the data we have, we begin the exploratory data analysis. With the use of orthogonal transformation, PCA is a statistical technique that transforms the observations of correlated features into a set of linearly uncorrelated features. The Principal Components are the name given to these newly transformed features. The method aids in the cost-effectiveness of classification, regression, or any other type of machine learning by lowering the dimensions of the data.

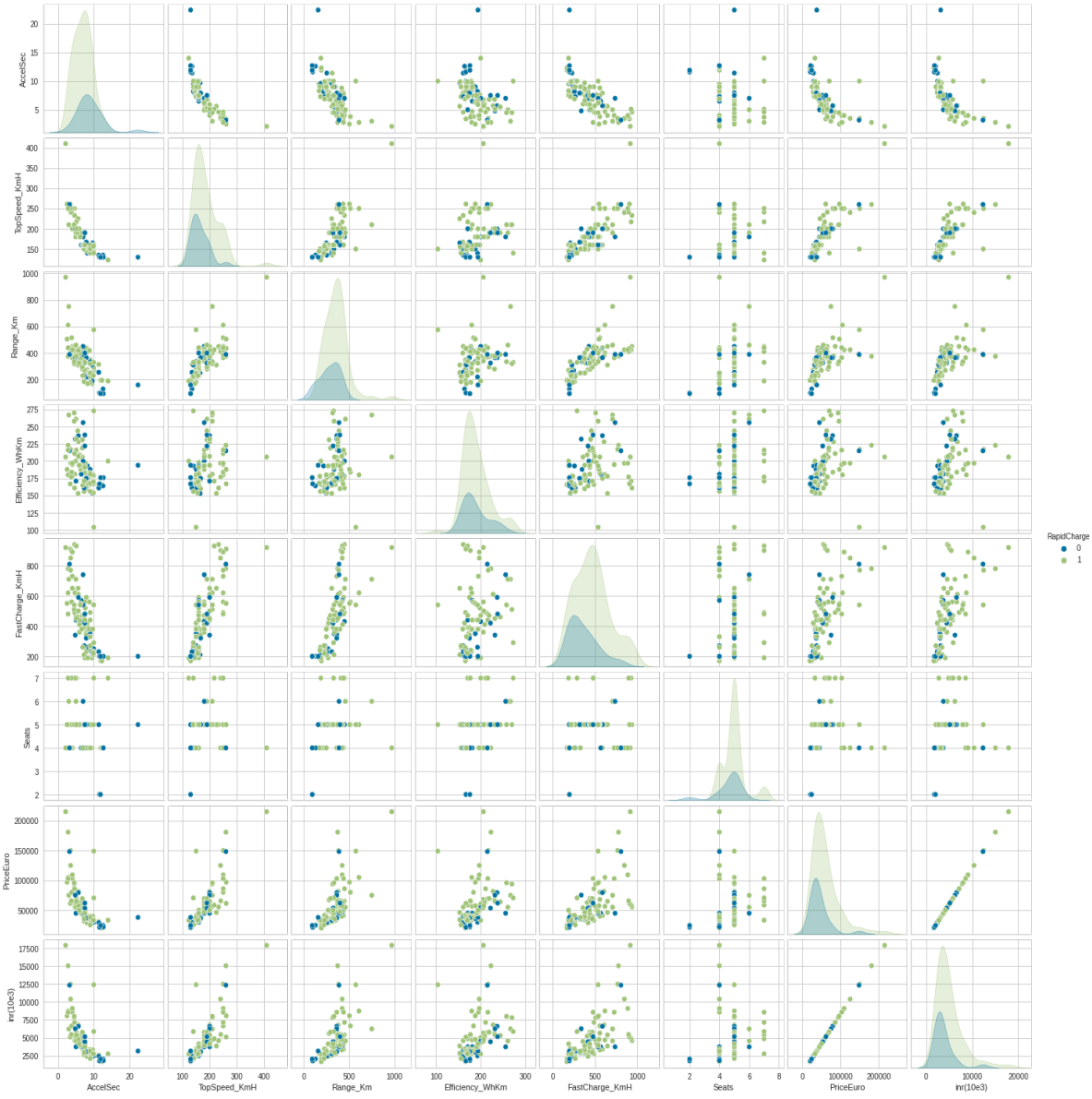
Comparision of cars in our data




Which Car Has a Top speed?
TopSpeed_KmH
3000
2500
2000
1500
1000
500
0
Tesla
Volkswagen
Polestar
BMW
Honda
Lucid
Peugeot
Audi
Mercedes
Nissan
Hyundai
Porsche
MG
Mini
Opel
Skoda
Volvo
Kia
Renault
Mazda
Lexus
CUPRA
SEAT
Lightyear
Aiways
DS
Citroen
Jaguar
Ford
Byton
Sono
Smart
Fiat
Brand
Brand
Tesla
Volkswagen
Polestar
BMW
Honda
Lucid
Peugeot
Audi
Mercedes
Nissan
Hyundai
Porsche
MG
Mini


Car Price
Car Price
200000
175000
150000
125000
100000
75000
50000
25000
0
20
40
60
80
100
No. of Samples

For Electric Vehicle Market one of the most important key is Charging:



Correlation Matrix: Simply said, a correlation matrix is a table that shows the correlation. It works well with variables that show a linear relationship to one another. different variables' coefficients. The heatmap in the next graphic shows how the matrix illustrates the correlation between all possible pairings of values. When the correlation coefficient between two variables is greater than 0.7, that relationship is typically regarded as strong.

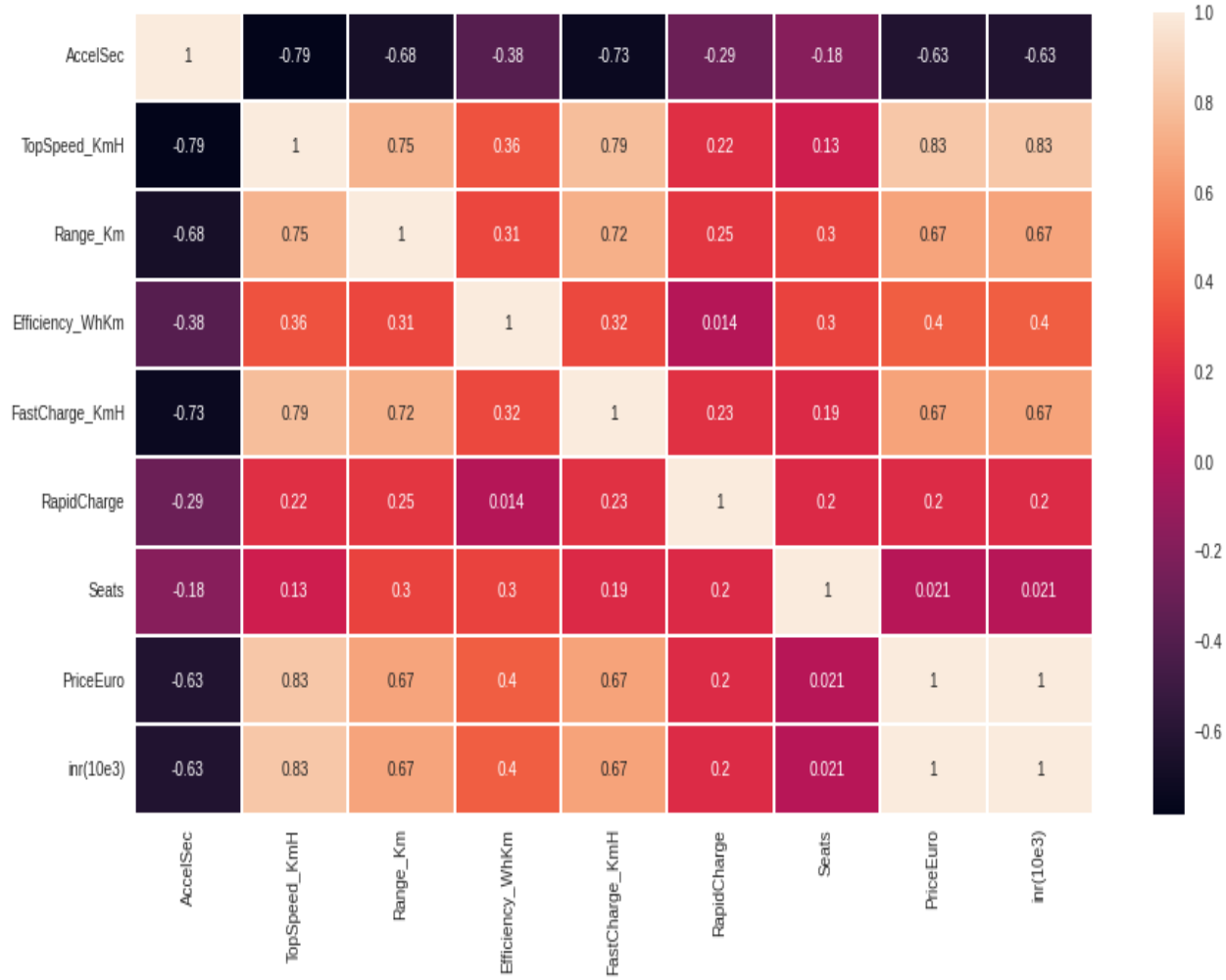
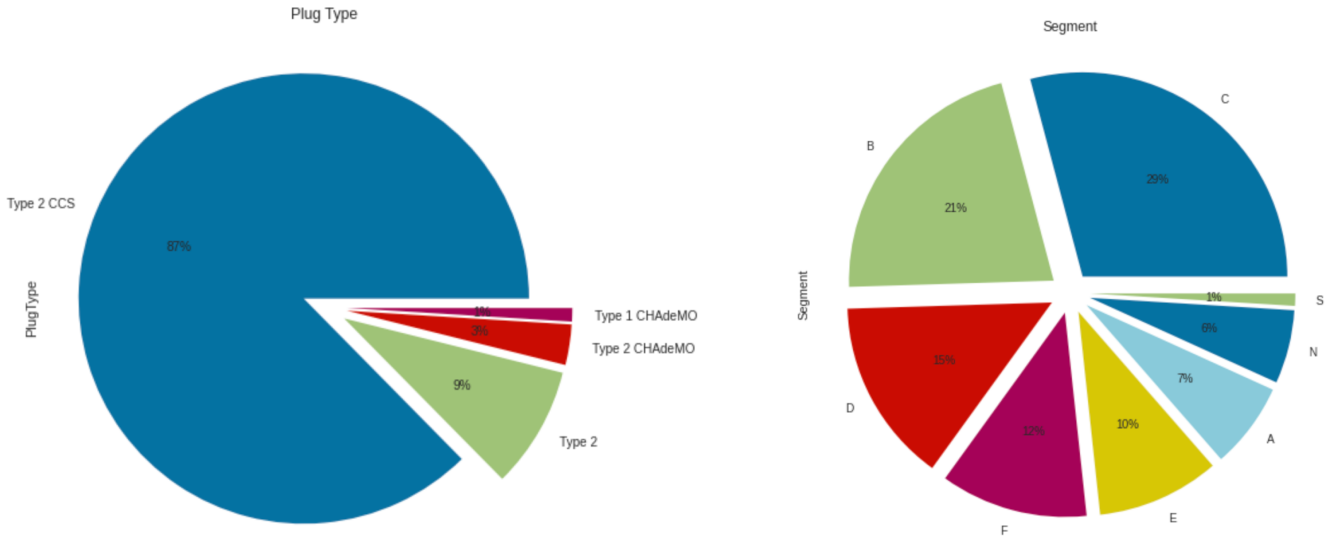
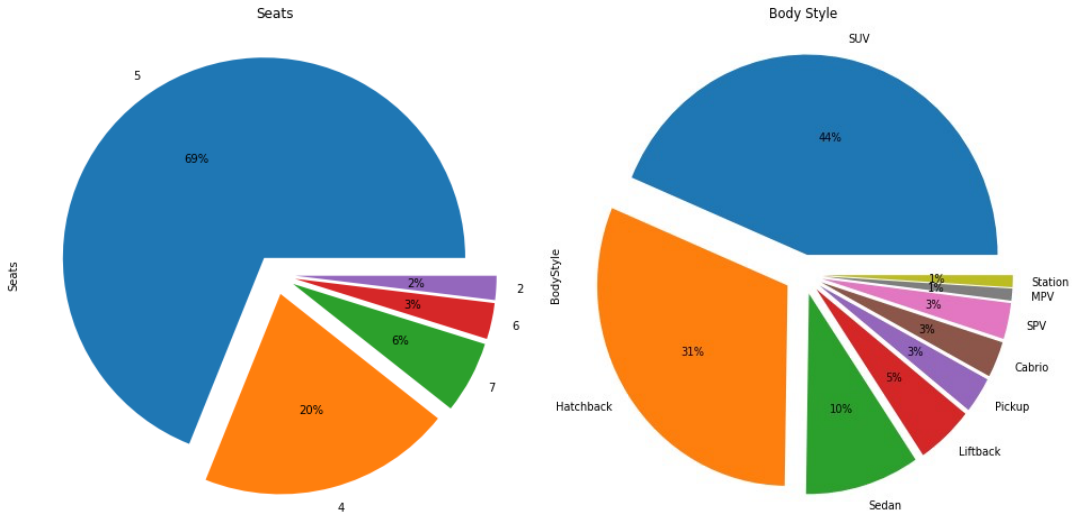


Figure 5: *Correlation Matrix for the dataset*





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

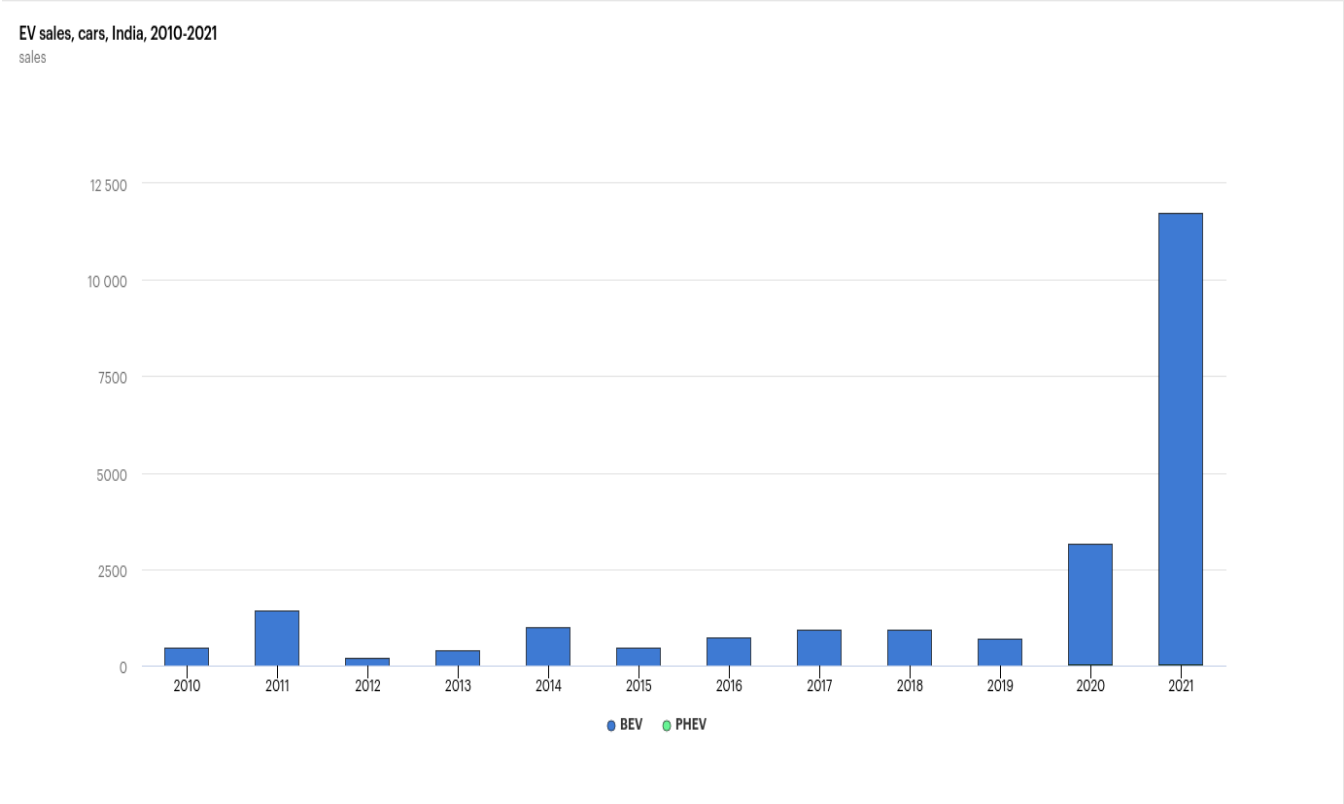


Figure 6: *Electric Cars sales in India*

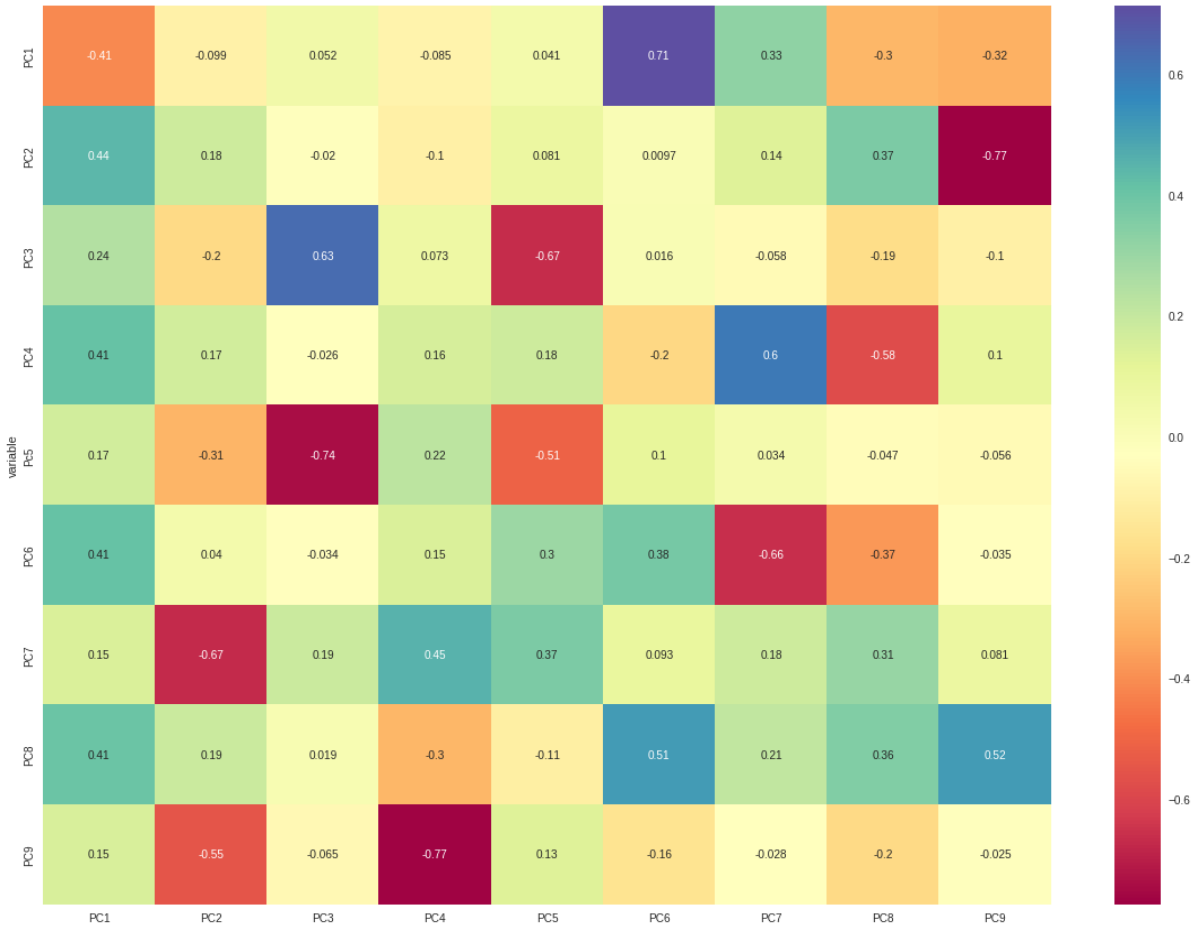


Figure 7: *Correlation matrix plot for loadings*

Scree Plot: 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 scree 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 scree 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.

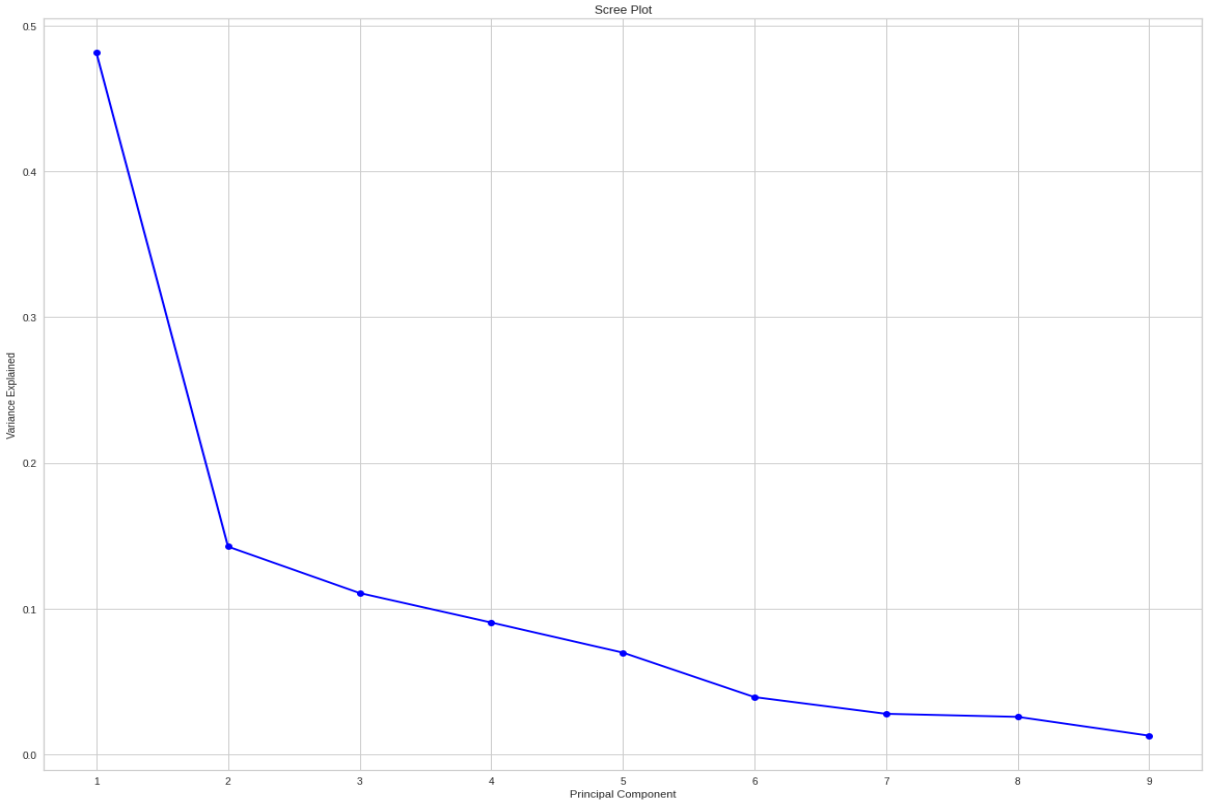


Figure 8: *Scree Plot for our Dataset*

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 chose 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.

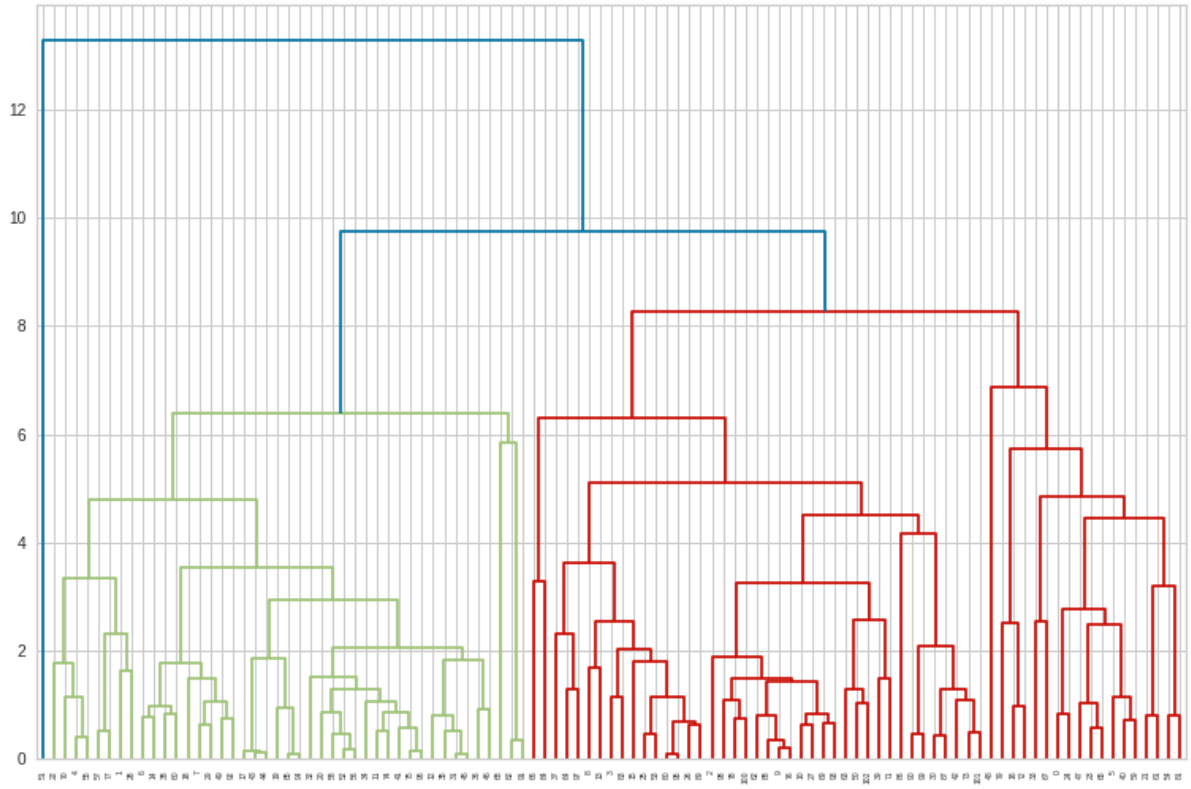


Figure 9: *Dendrogram Plot for our Dataset*

Elbow Method

The Elbow method is a popular method for determining the optimal number of clusters. The method is based on calculating the Within-Cluster-Sum of Squared Errors (WSS) for a different number of clusters (k) and selecting the k for which change in WSS first starts to diminish. The idea behind the elbow method is that the explained variation changes rapidly for a small number of clusters and then it slows down leading to an elbow formation in the curve. The elbow point is the number of clusters we can use for our clustering algorithm.

The KElbowVisualizer function fits the KMeans model for a range of clusters values between 2 to 8. As shown in Figure, the elbow point is achieved which is highlighted by the function itself. The function also informs us about how much time was needed to plot models for various numbers of clusters through the green line.

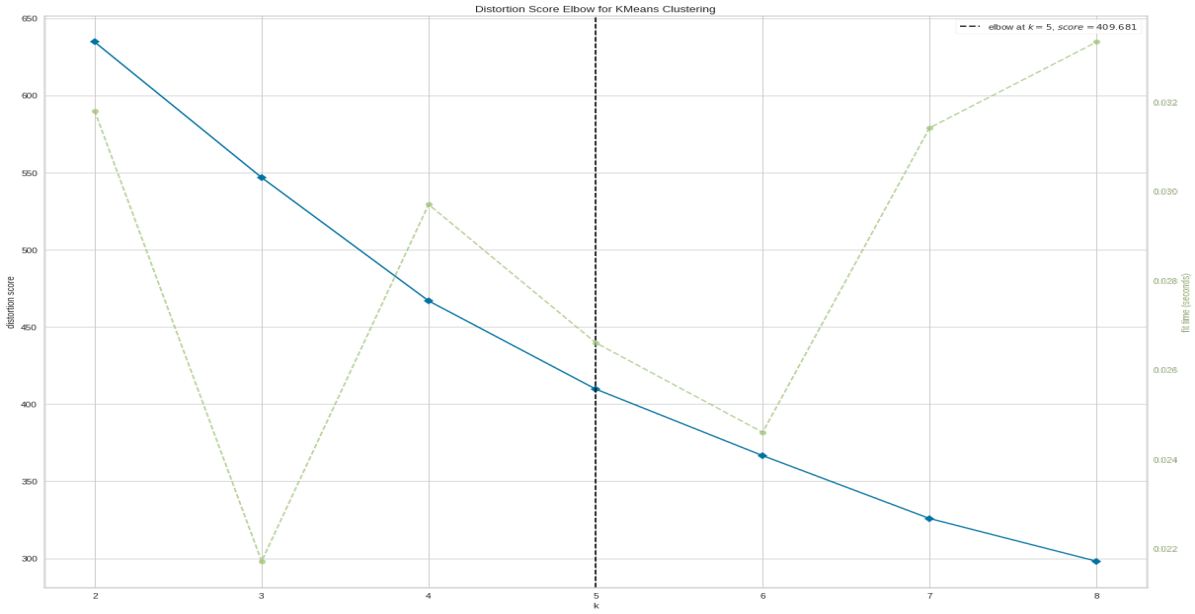


Figure 10: *Evaluating the cluters using Distortion*

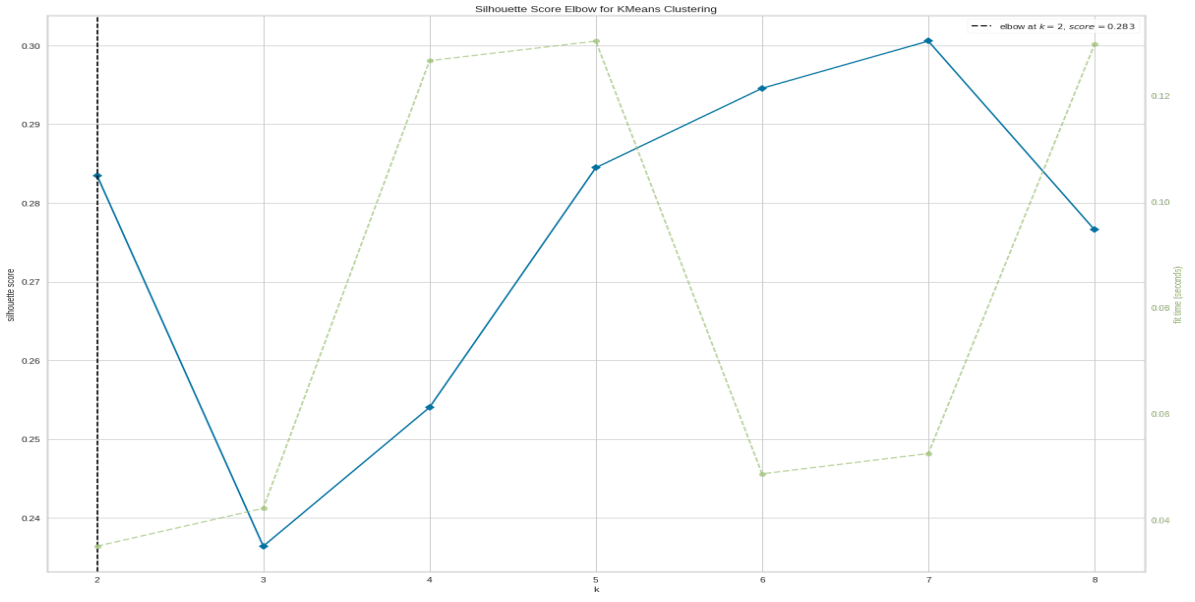


Figure 11: *Evaluating the cluters using silhouette*

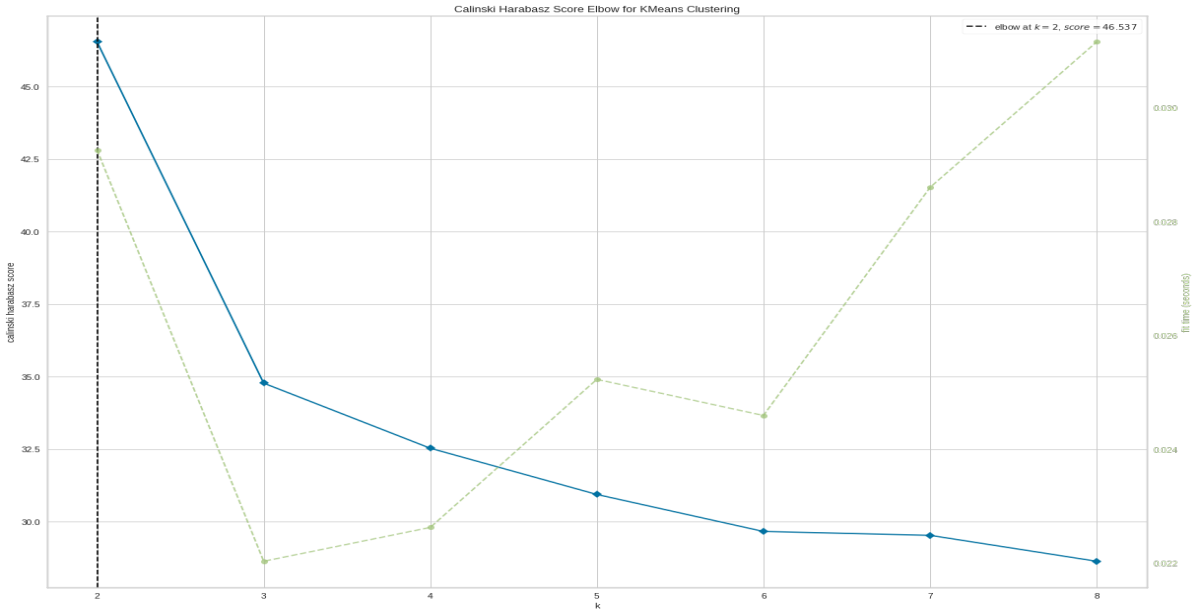


Figure 12: *Evaluating the cluters using calinski$_{harabasz}$*

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 homogeneous (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:

$$\underline{J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} ||x^i - \mu_k||} \tag{1}$$

And M-step is :

$$\frac{\partial J}{\partial \mu_k} \stackrel{\Sigma}{=} 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}}$$

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.

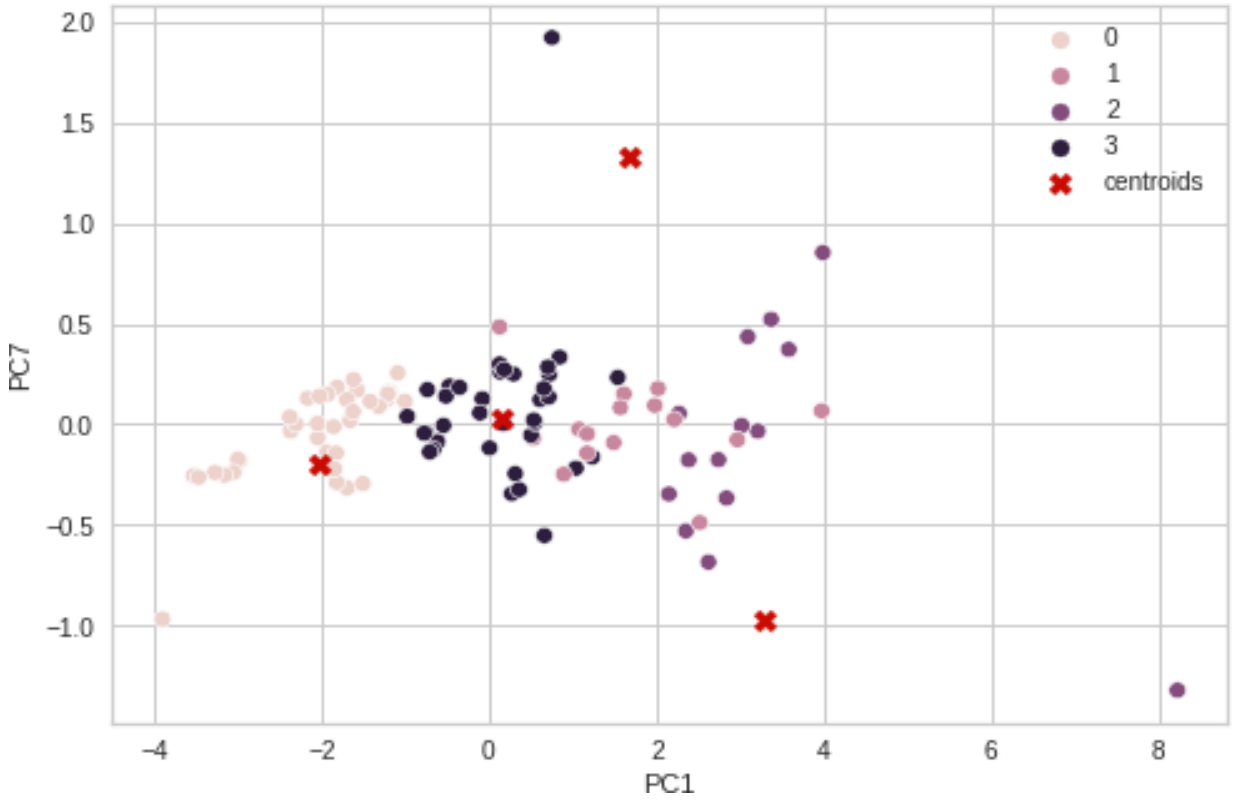
The **k-means clustering algorithm** performs the following tasks:

- 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.

According to the Elbow method, here we take K=4 clusters to train KMeans model. The derived clusters are shown in the following figure

```
#K-means clustering

kmeans = KMeans(n_clusters=4, init='k-means++', random_state=0).fit(t)
df['cluster_num'] = kmeans.labels_ #adding to df
print (kmeans.labels_) #Label assigned for each data point
print (kmeans.inertia_) #gives within-cluster sum of squares.
print(kmeans.n_iter_) #number of iterations that k-means algorithm runs to get a minimum within-cluster sum of squares
print(kmeans.cluster_centers_) #Location of the centroids on each cluster.
```



Prediction of Prices most used cars

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models targets prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Here we use a linear regression model to predict the prices of different Electric cars in different companies. X contains the independent variables and y is the dependent Prices that is to be predicted. We train our model with a splitting of data into a 4:6 ratio, i.e. 40% of the data is used to train the model.

LinearRegression().fit($X_{train,ytrain}$) command is used to fit the data set into model. The values of intercept, coefficient, and cumulative distribution function (CDF) are described in the figure.

```
[85]  1  X=data2[['PC1', 'PC2','PC3','PC4','Pc5','PC6', 'PC7','PC8','PC9']]
      2  y=df['inr(10e3)']
```

```
[86]  1  X_train, X_test, y_train, y_test = train_test_split(X, y,test_size=0.4, random_state=101)
      2  lm=LinearRegression().fit(X_train,y_train)
```

```
[87]  1  print(lm.intercept_)
```

```
4643.522050485437
```

```
[88]  1  lm.coef_
```

```
array([1144.95884,  530.09473,   54.50586, -843.38276, -306.27756,
       1449.94438,  595.62449, 1005.47168, 1455.75874])
```

```
[89]  1  X_train.columns
```

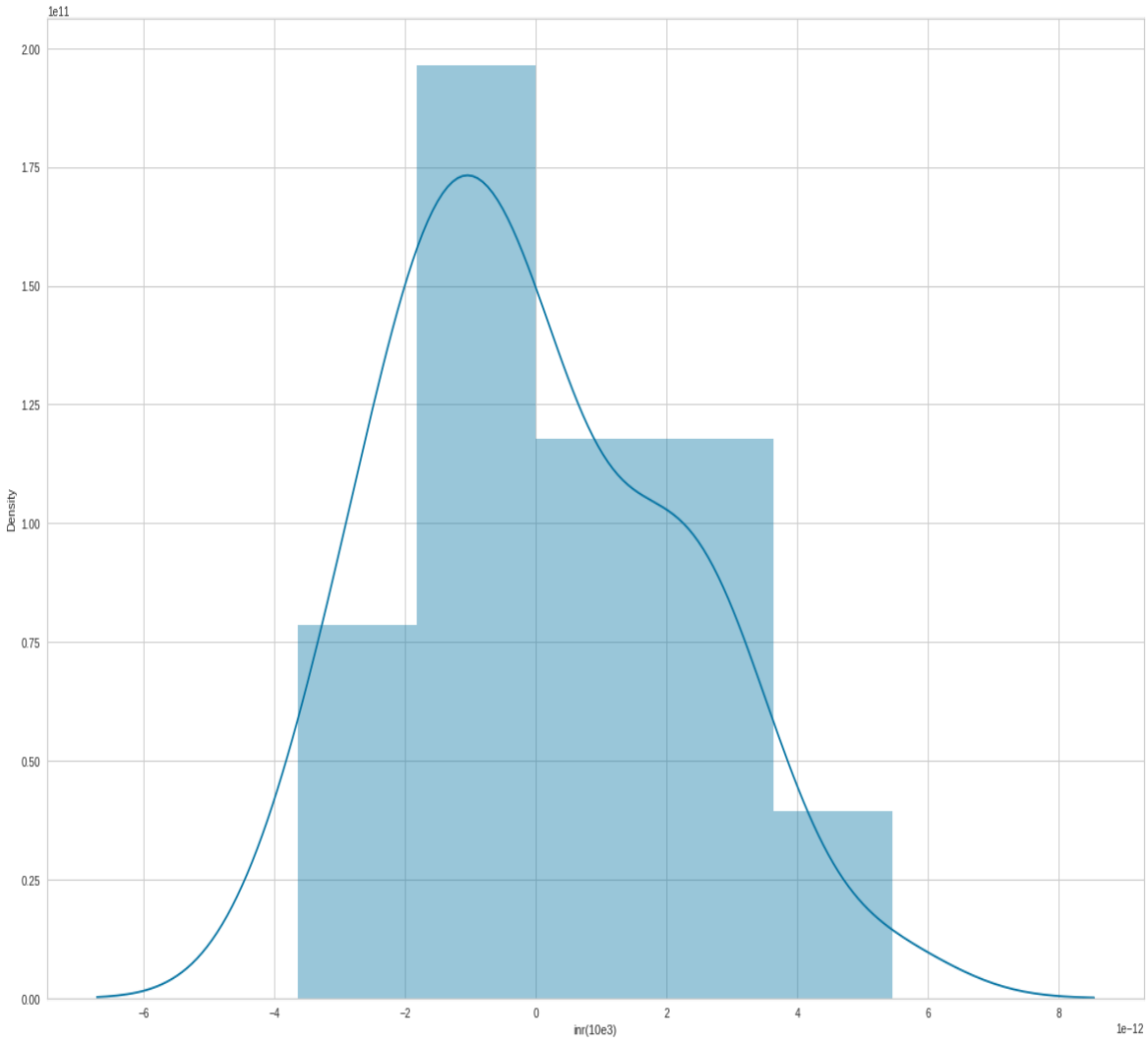
```
Index(['PC1', 'PC2', 'PC3', 'PC4', 'Pc5', 'PC6', 'PC7', 'PC8', 'PC9'], dtype='object')
```

```
      1  cdf=pd.DataFrame(lm.coef_, X.columns, columns=['Coeff'])
      2  cdf
```

	Coeff
PC1	1144.9588
PC2	530.0947
PC3	54.5059
PC4	-843.3828
Pc5	-306.2776
PC6	1449.9444
PC7	595.6245
PC8	1005.4717
PC9	1455.7587

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.

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



```
[99]  1  print('MAE:',metrics.mean_absolute_error(y_test,predictions))
      2  print('MSE:',metrics.mean_squared_error(y_test,predictions))
      3  print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test,predictions)))
```

```
MAE: 1.7540254962763616e-12
MSE: 4.588882922020368e-24
RMSE: 2.142167809024393e-12
```

```
[100]  1  metrics.mean_absolute_error(y_test,predictions)
```

```
1.7540254962763616e-12
```

```
[101]  1  metrics.mean_squared_error(y_test,predictions)
```

```
4.588882922020368e-24
```

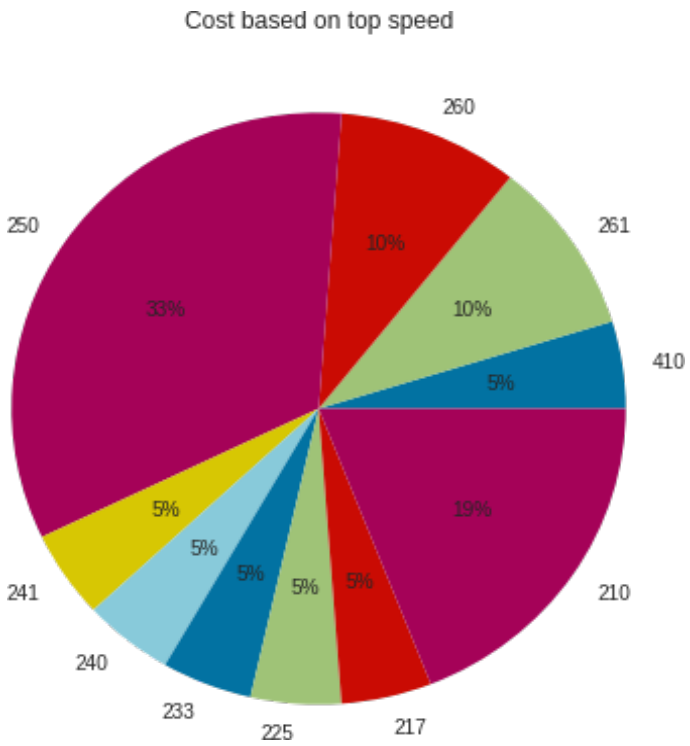
```
[102]  1  np.sqrt(metrics.mean_squared_error(y_test,predictions))
```

```
2.142167809024393e-12
```

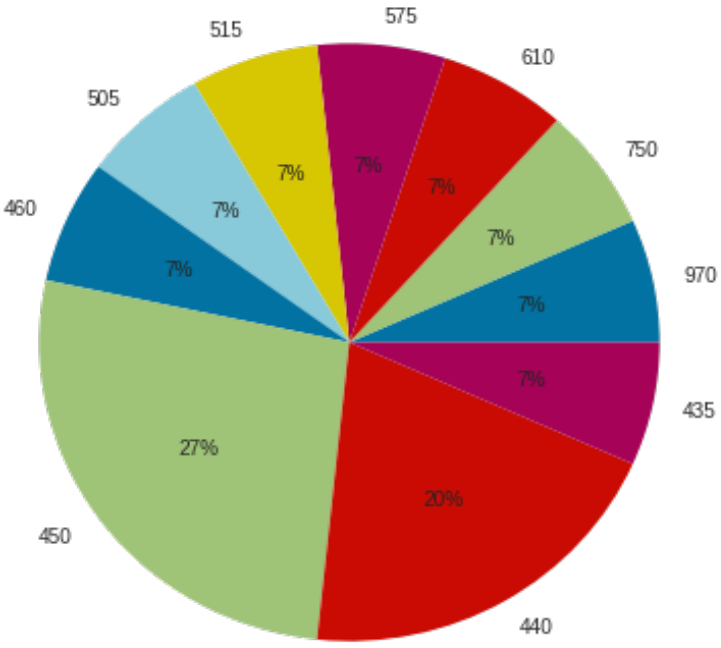
Profiling and Describing the Segments

Sorting the Top Speeds and Maximum Range in accordance to the Price with head () we can view the Pie Chart.

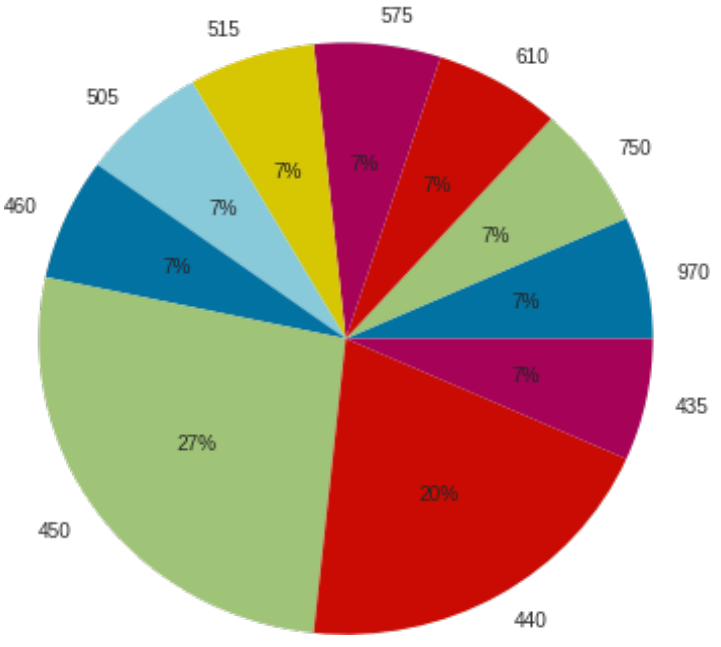
Pie Chart:



Cost based on Maximum Range



Top Speeds based on Maximum Range



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.

Finally, our target segment should contain cars with most **Efficiency**, contains **Top Speed** and price between **16 to 180 lakhs** with mostly with **5 seats**.

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.

[3] McDonald, M., amp; Dunbar, I. (2003). /emphMarket segmentation. Butterworth-Heinemann