Implementation

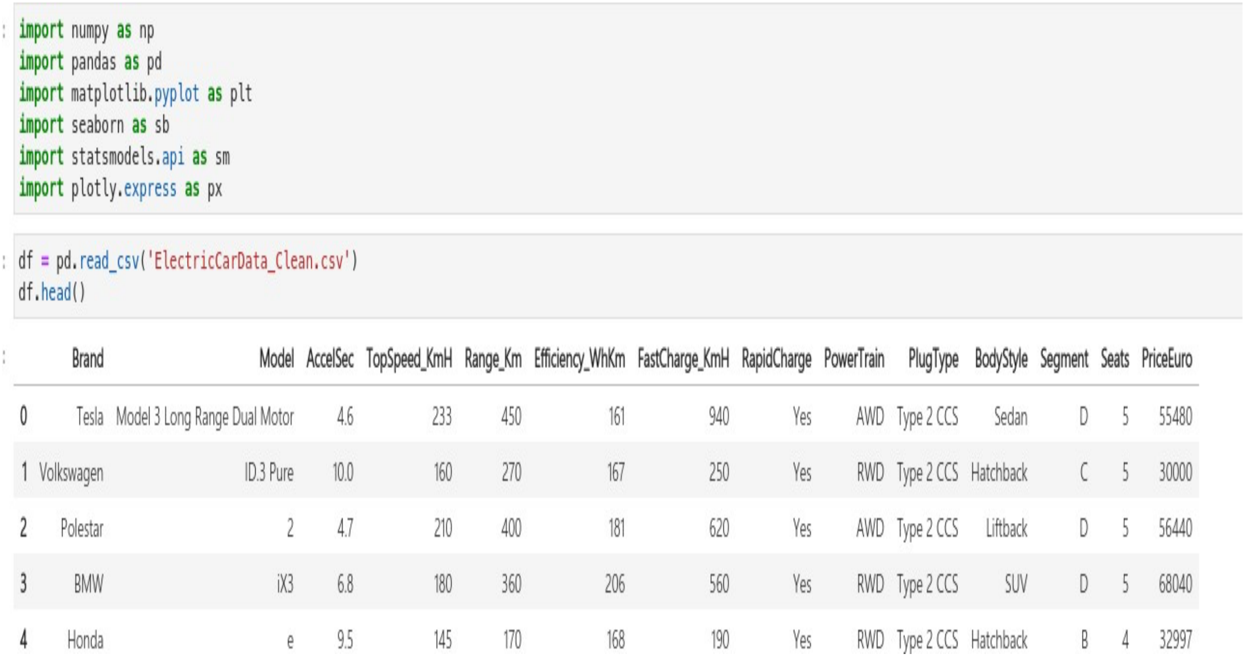
Packages/Tools used:

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.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb
import statsmodels.api as sm
import plotly.express as px
```

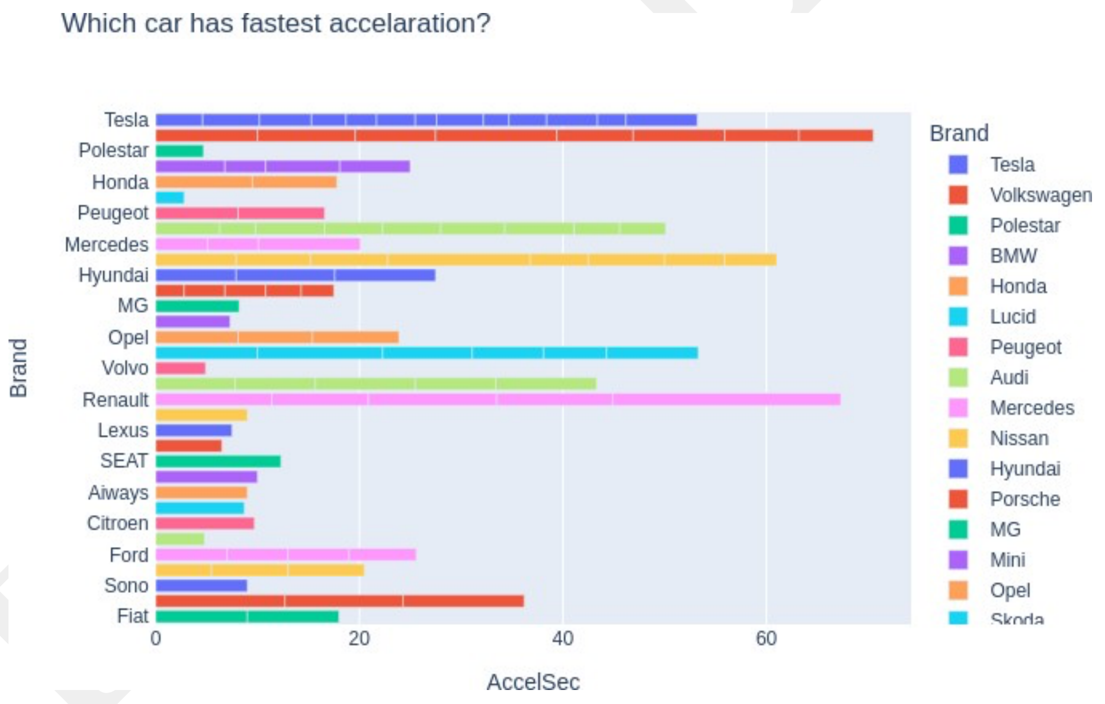
```
df = pd.read_csv('ElectricCarData_Clean.csv')
df.head()
```

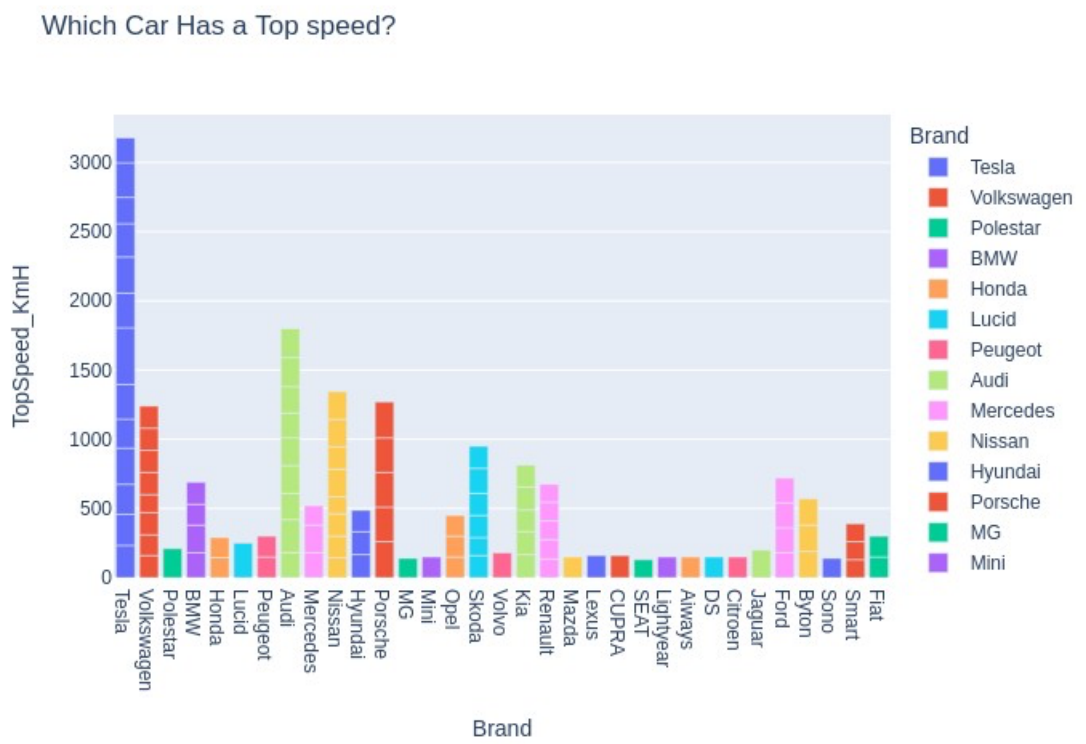
	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment	Seats	PriceEuro
0	Tesla	Model 3 Long Range Dual Motor	4.6	233	450	161	940	Yes	AWD	Type 2 CCS	Sedan	D	5	55480
1	Volkswagen	ID.3 Pure	10.0	160	270	167	250	Yes	RWD	Type 2 CCS	Hatchback	C	5	30000
2	Polestar	2	4.7	210	400	181	620	Yes	AWD	Type 2 CCS	Liftback	D	5	56440
3	BMW	iX3	6.8	180	360	206	560	Yes	RWD	Type 2 CCS	SUV	D	5	68040
4	Honda	e	9.5	145	170	168	190	Yes	RWD	Type 2 CCS	Hatchback	B	4	32997

EDA

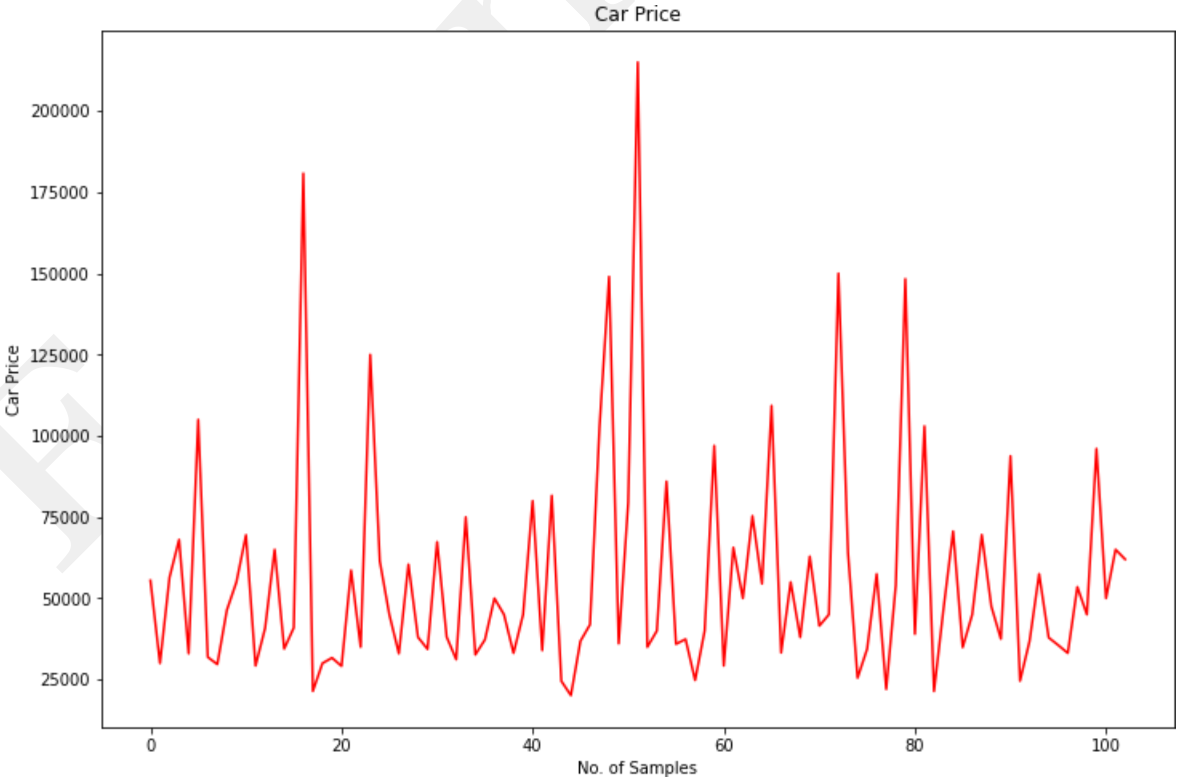
We start the Exploratory Data Analysis with some data Analysis drawn from the data without Principal Component Analysis and with some Principal Component Analysis in the dataset obtained from the combination of all the data we have. PCA is a statistical process that converts the observations of correlated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

Comparision of cars in our data



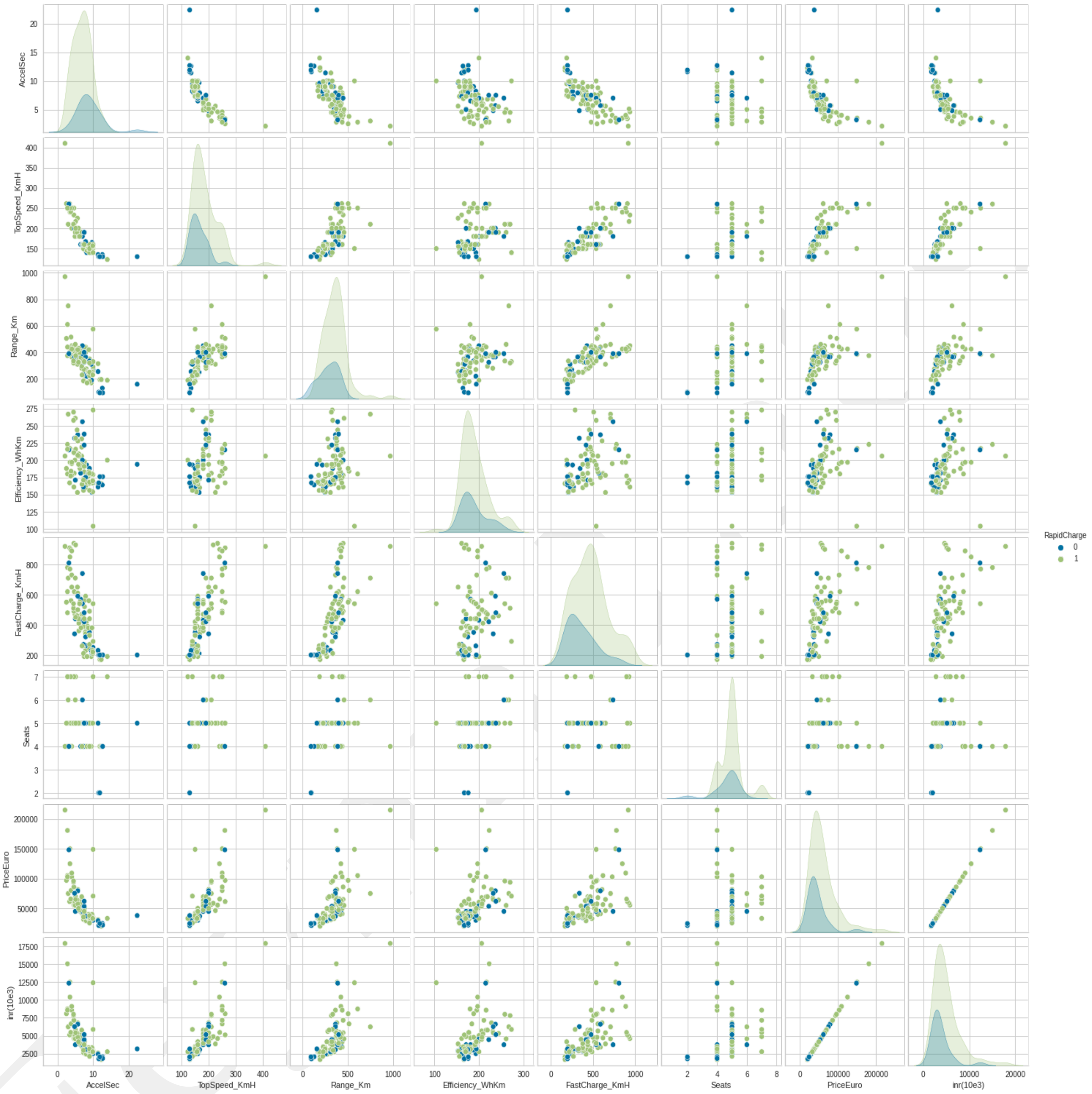


Which Car Has a Top speed?



Car Price

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



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

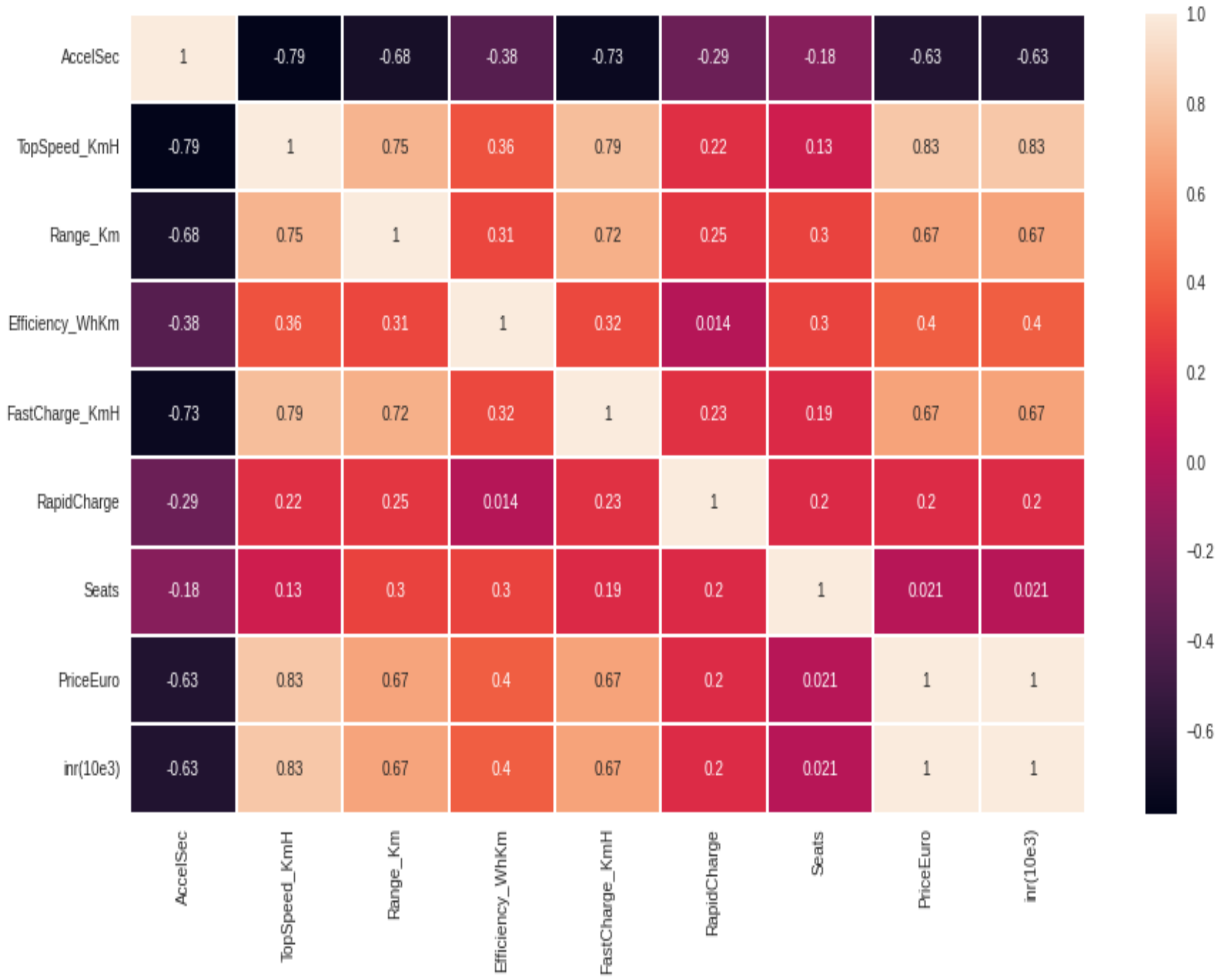
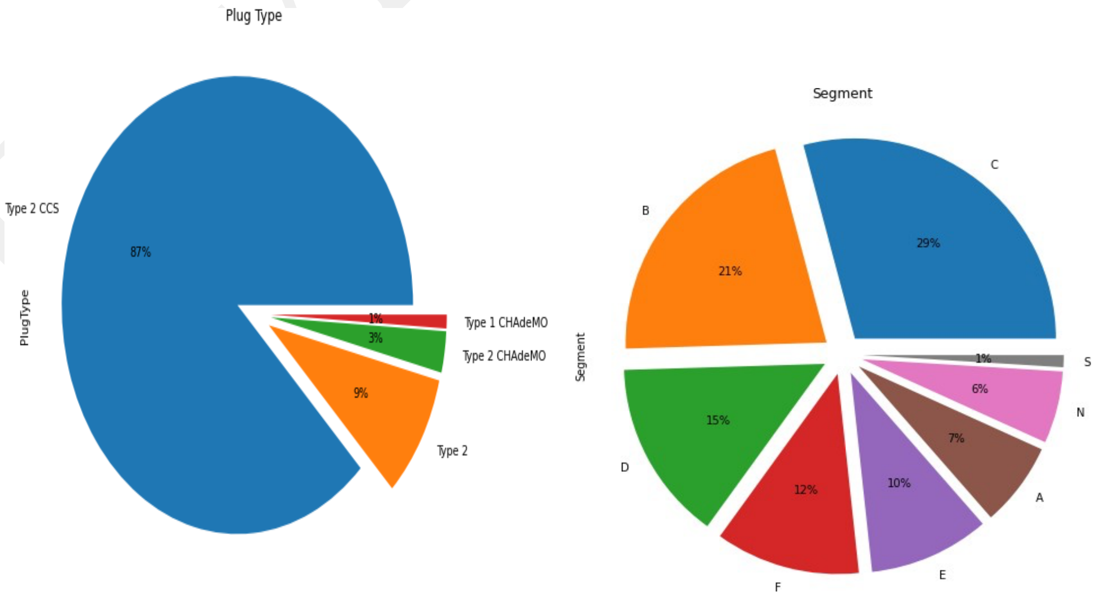
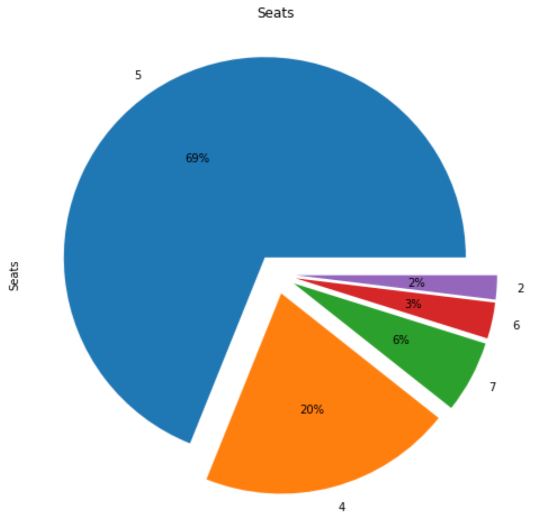
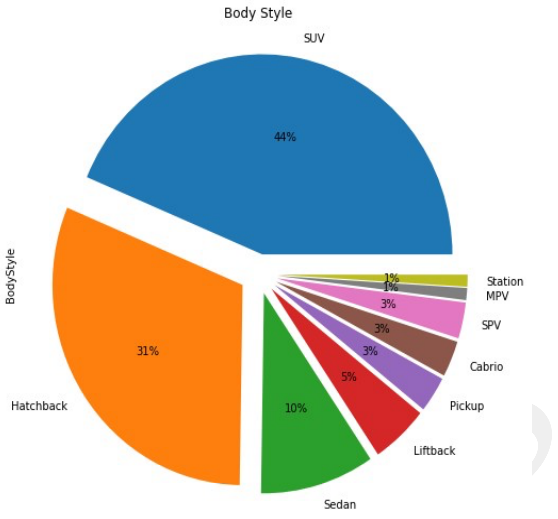


Figure 5: *Correlation Matrix for the dataset*







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

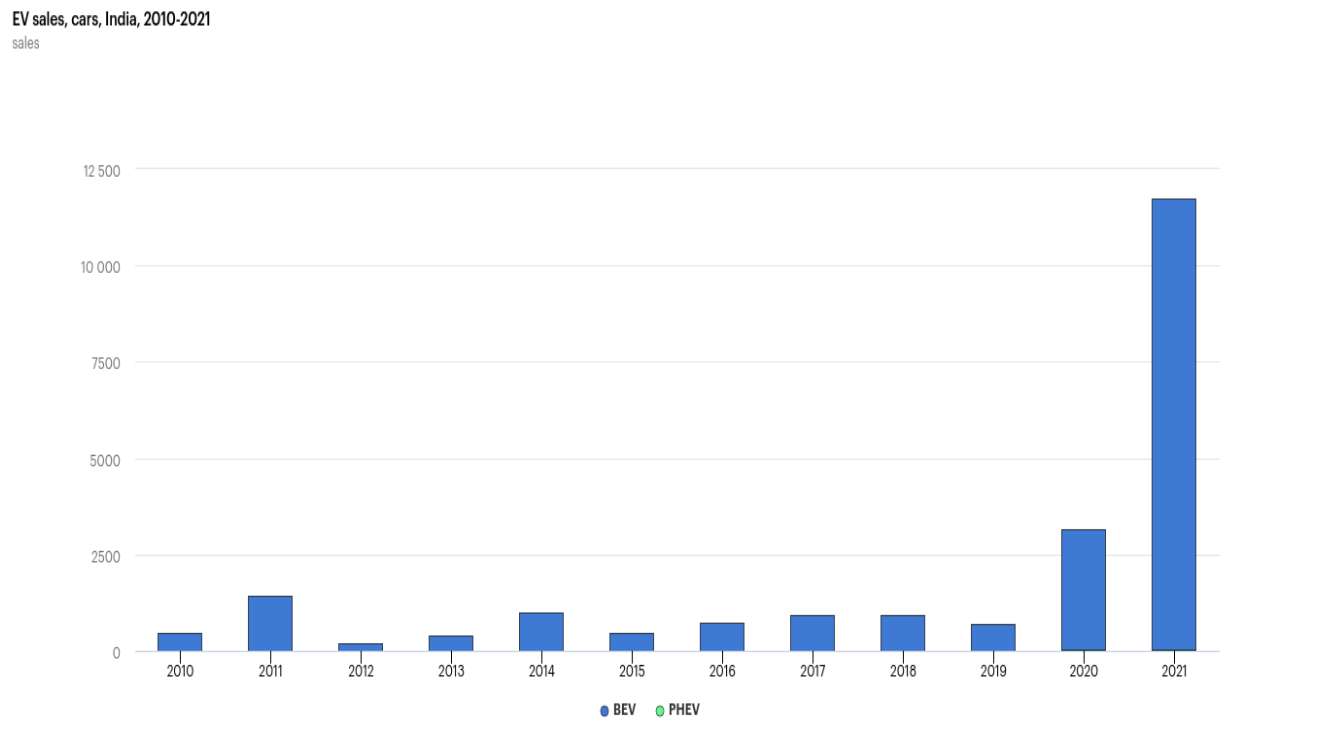
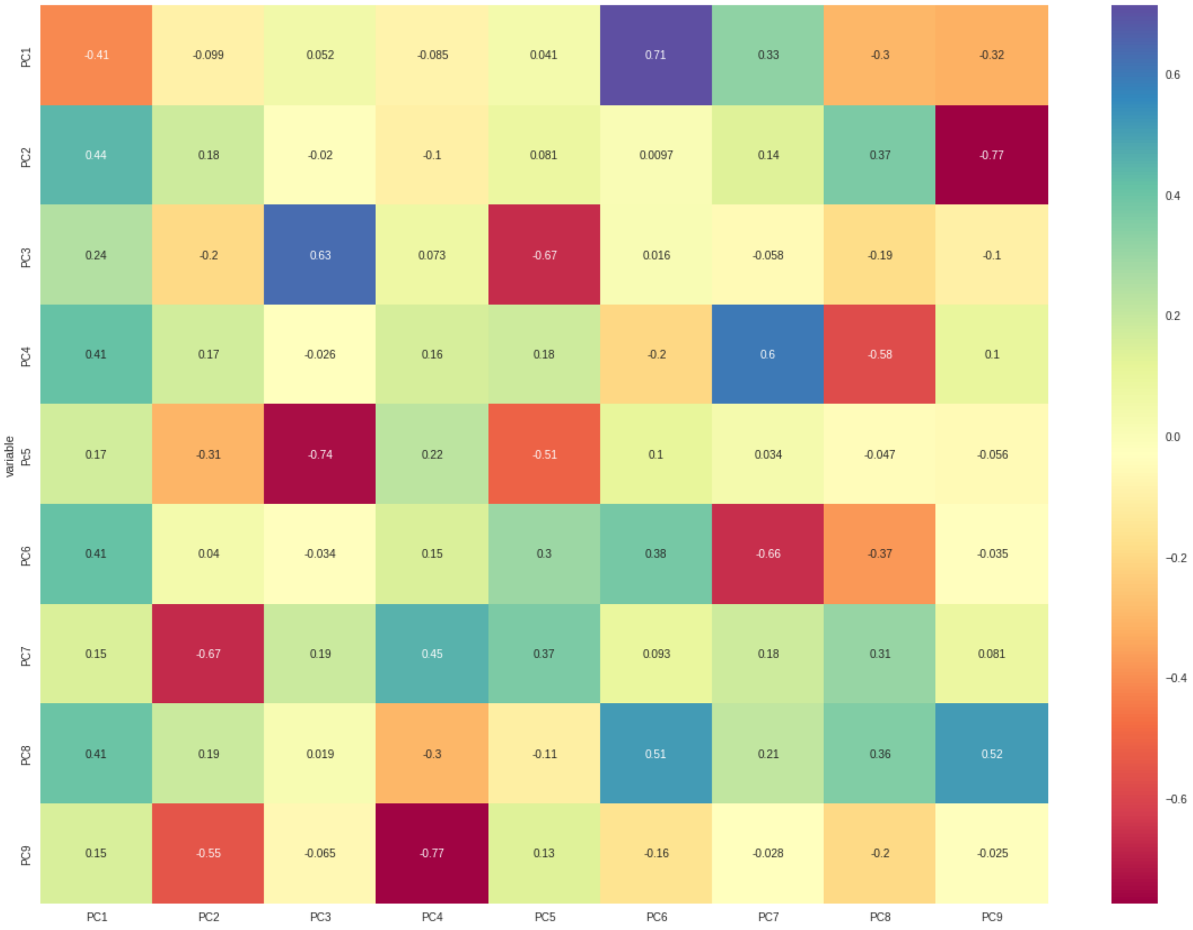


Figure 6: *Electric Cars sales in India*



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
PC1	-0.41	-0.099	0.052	-0.085	0.041	0.71	0.33	-0.3	-0.32
PC2	0.44	0.18	-0.02	-0.1	0.081	0.0097	0.14	0.37	-0.77
PC3	0.24	-0.2	0.63	0.073	-0.67	0.016	-0.058	-0.19	-0.1
PC4	0.41	0.17	-0.026	0.16	0.18	-0.2	0.6	-0.58	0.1
PC5	0.17	-0.31	-0.74	0.22	-0.51	0.1	0.034	-0.047	-0.056
PC6	0.41	0.04	-0.034	0.15	0.3	0.38	-0.66	-0.37	-0.035
PC7	0.15	-0.67	0.19	0.45	0.37	0.093	0.18	0.31	0.081
PC8	0.41	0.19	0.019	-0.3	-0.11	0.51	0.21	0.36	0.52
PC9	0.15	-0.55	-0.065	-0.77	0.13	-0.16	-0.028	-0.2	-0.025

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 eigenvaluesfor each individual PC. It shows the eigenvalues on the y-axis and the number of fac- tors 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.

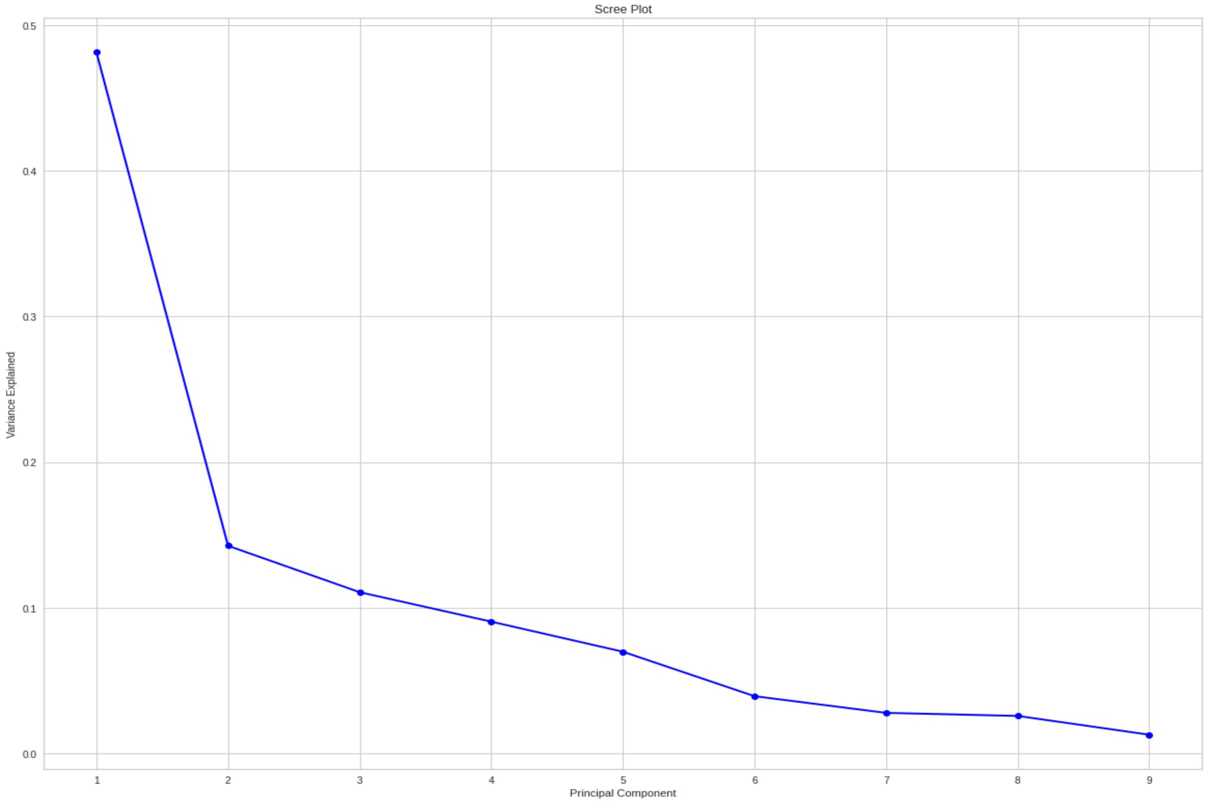


Figure 8: *Scree Plot for our Dataset*

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.

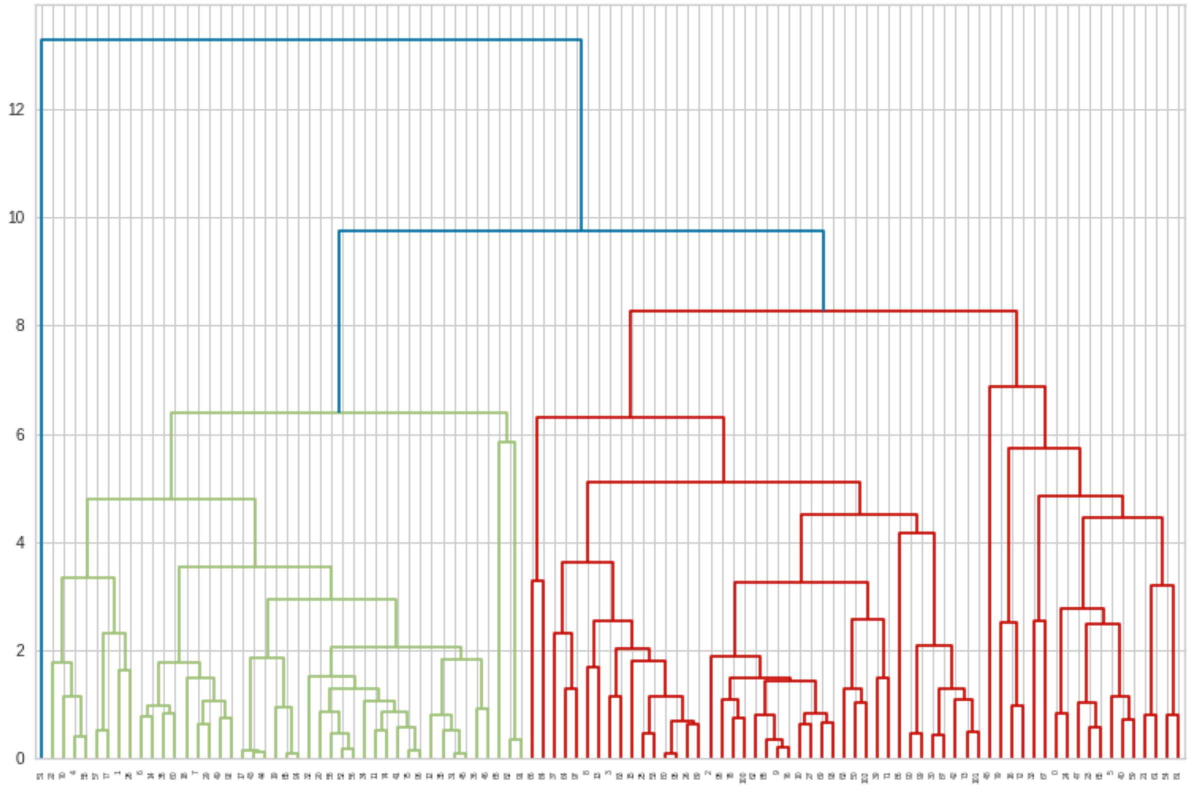


Figure 9: *Dendrogram Plot for our Dataset*

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.

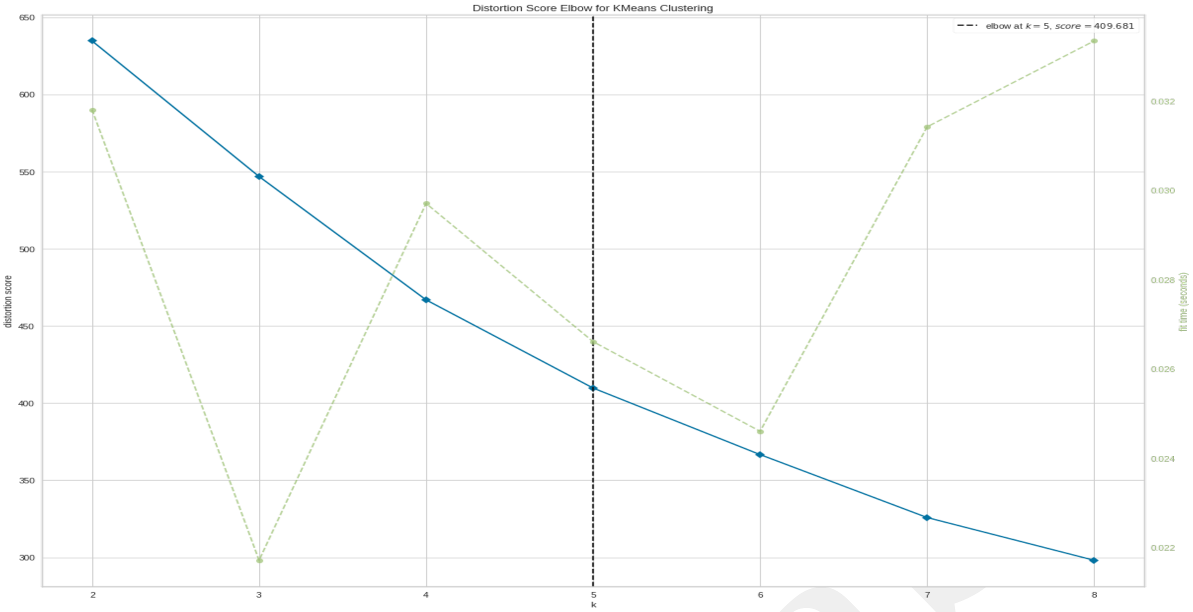


Figure 10: *Evaluating the cluters using Distortion*

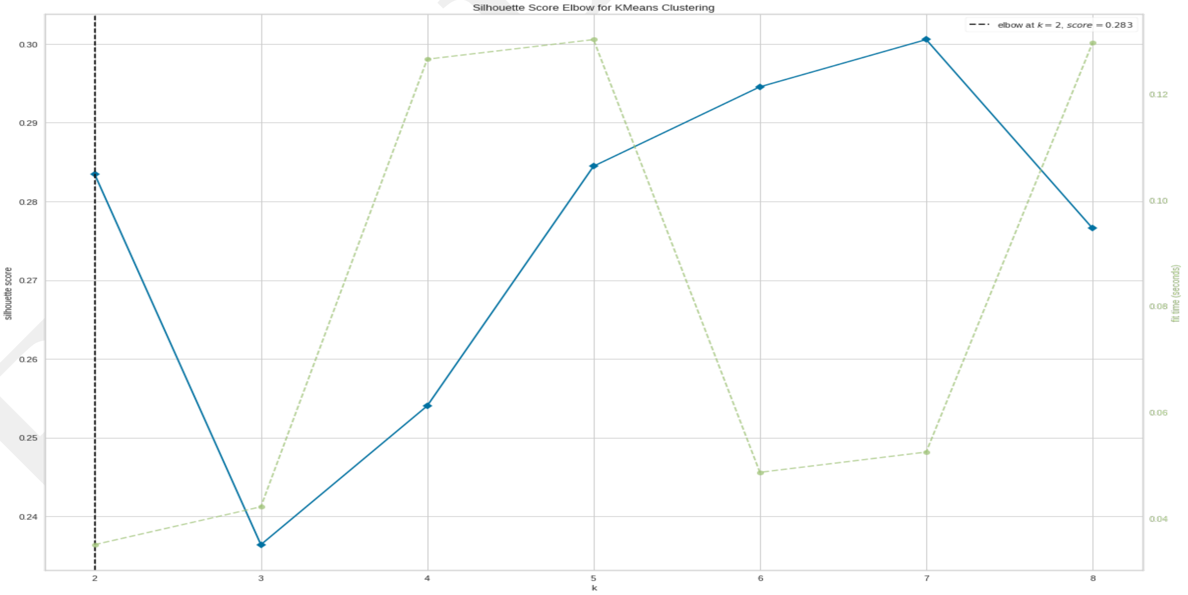


Figure 11: *Evaluating the cluters using silhouette*

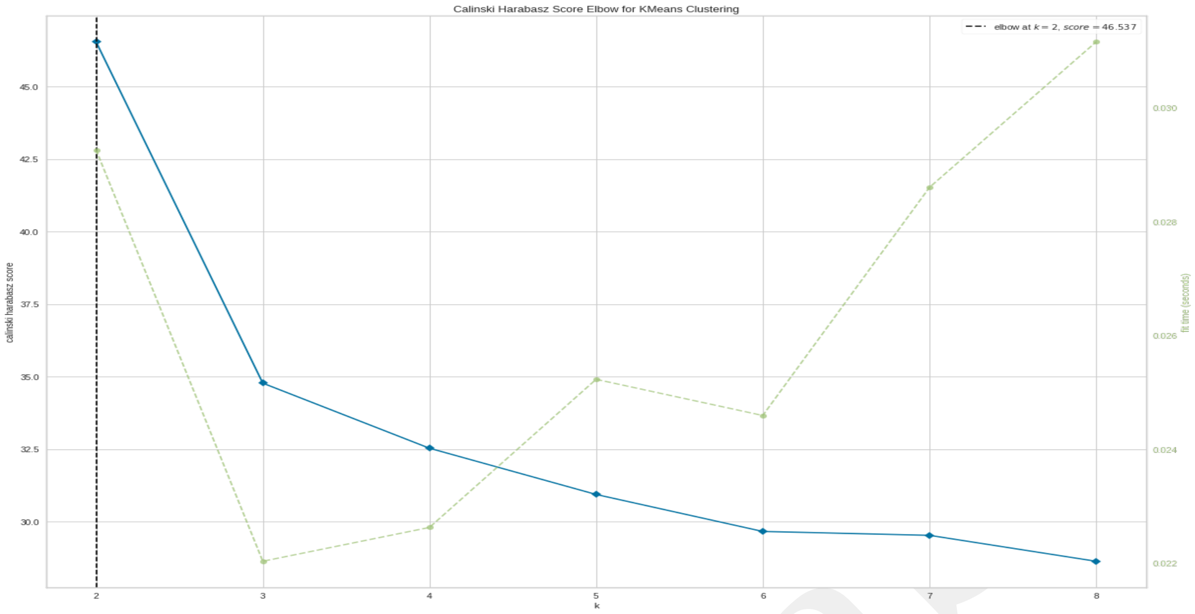


Figure 12: *Evaluating the cluters using calinski*$_{harabasz}$

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 computingthe centroid of each cluster. Below is a break down of how we can solve it mathematically,

The objective function is:

$$\overline{\underline{J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \lvert\lvert x^{i} - \mu_{k} \rvert\rvert}} \tag{1}$$

And M-step is :

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

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

Applications

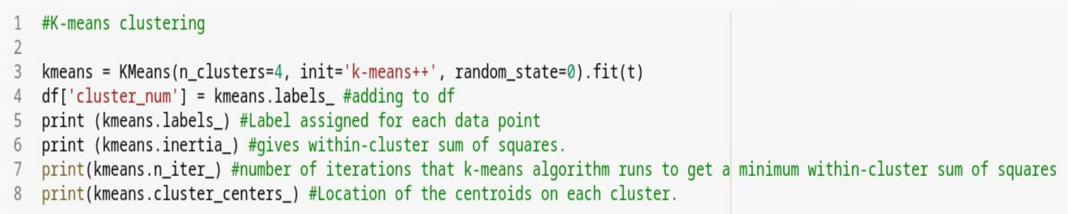
K means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc. The goal usually when we undergo a cluster analysis is either:

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.

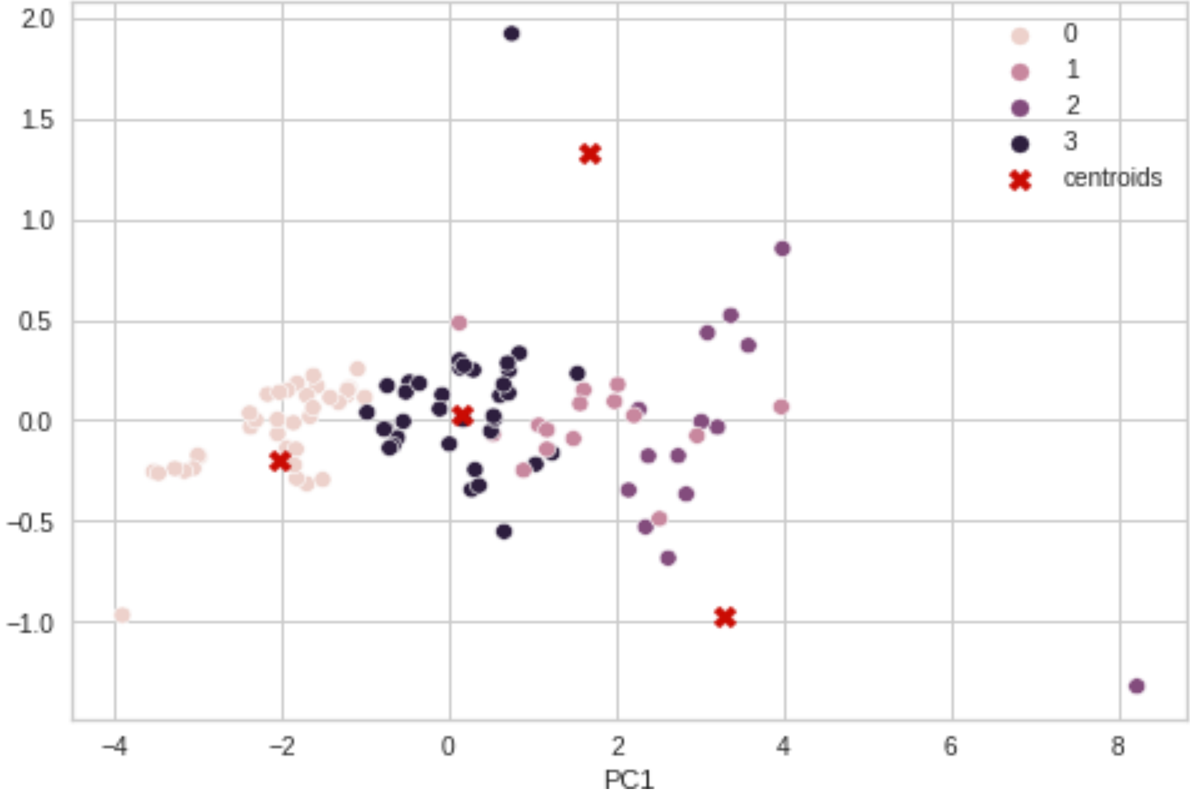
The **k-means clustering algorithm** performs the following tasks:

- 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



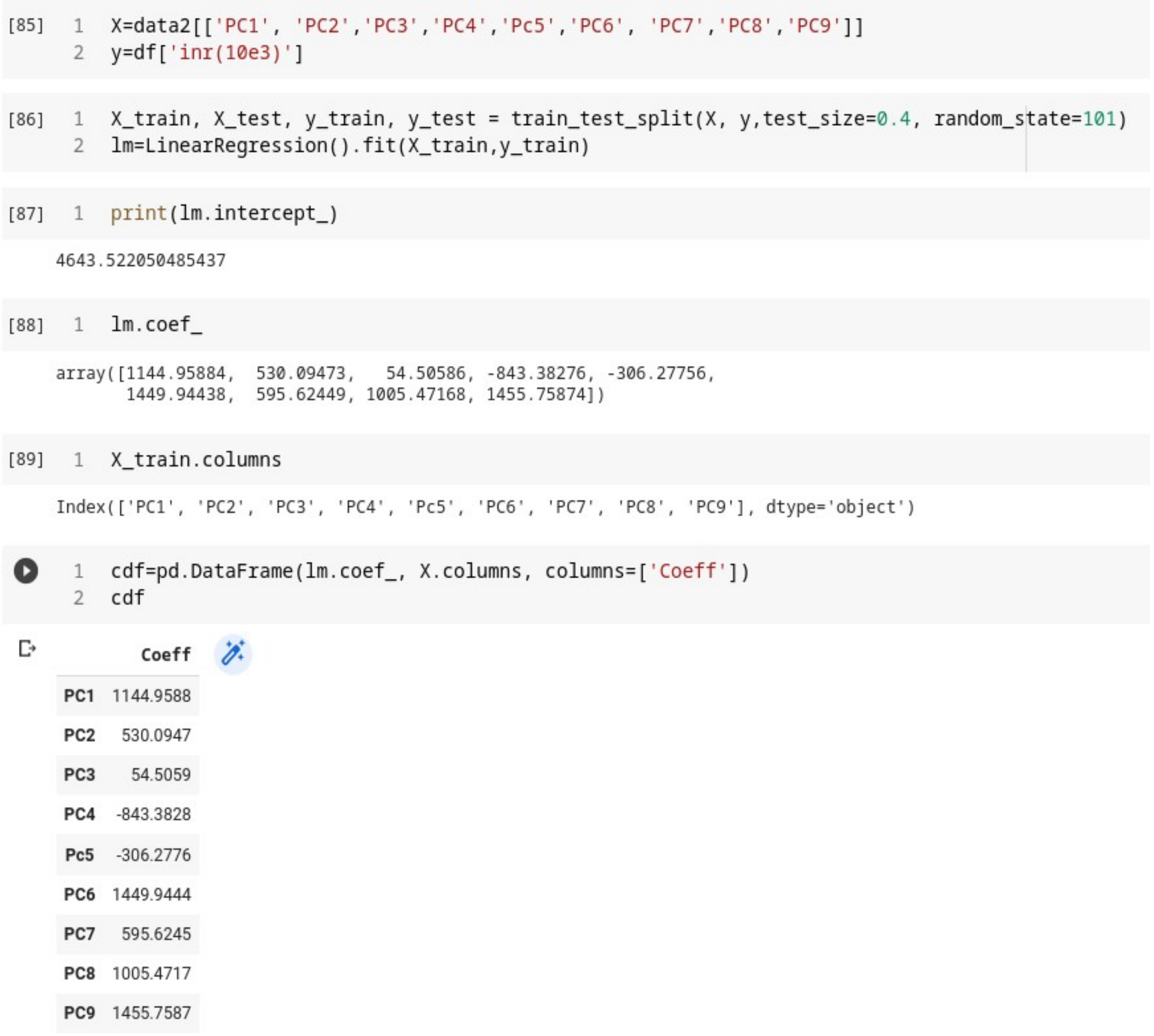
```
#K-means clustering

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

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

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

```
4643.522050485437
```

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

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

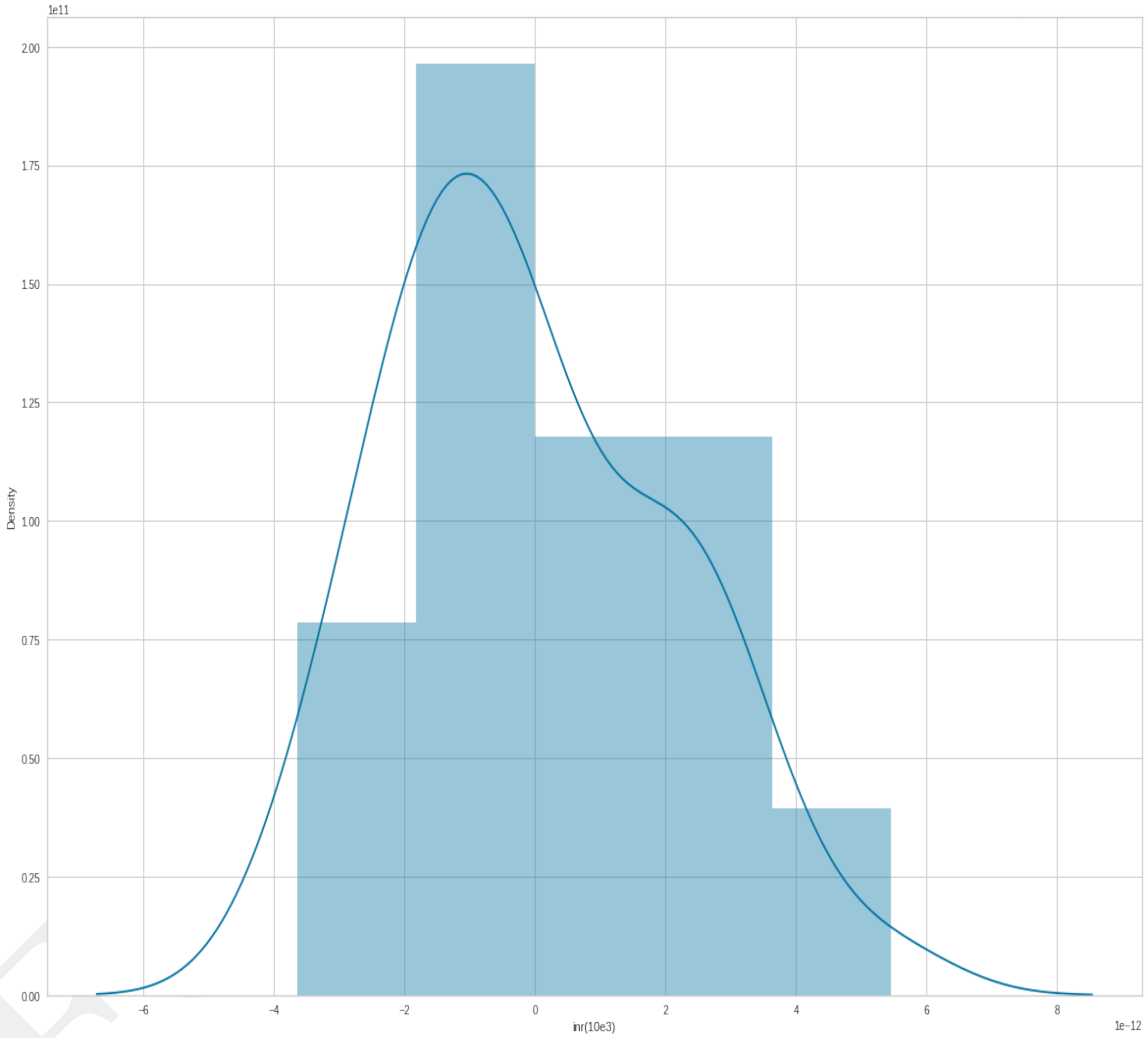
```
[89]
X_train.columns
```

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

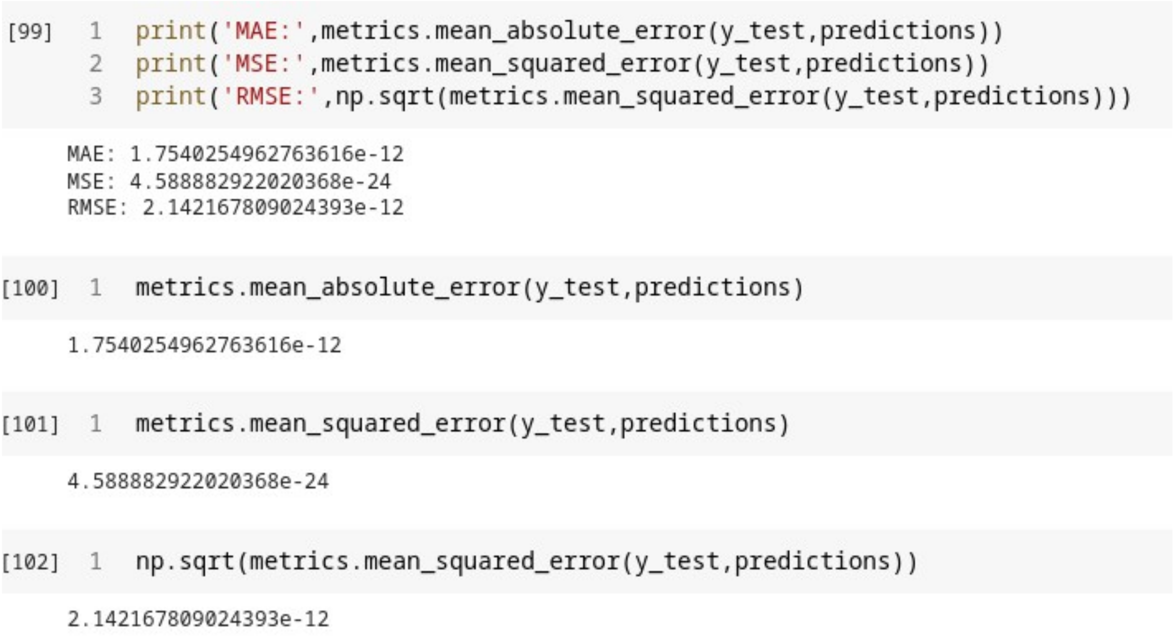
```
cdf=pd.DataFrame(lm.coef_, X.columns, columns=['Coeff'])
cdf
```

	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]  print('MAE:',metrics.mean_absolute_error(y_test,predictions))
      print('MSE:',metrics.mean_squared_error(y_test,predictions))
      print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test,predictions)))
```

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

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

```
1.7540254962763616e-12
```

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

```
4.588882922020368e-24
```

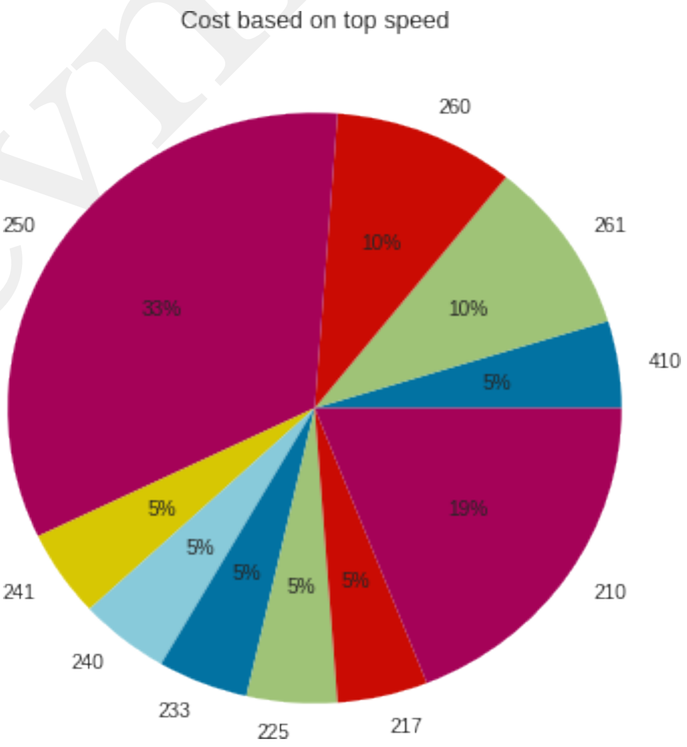
```
[102]  np.sqrt(metrics.mean_squared_error(y_test,predictions))
```

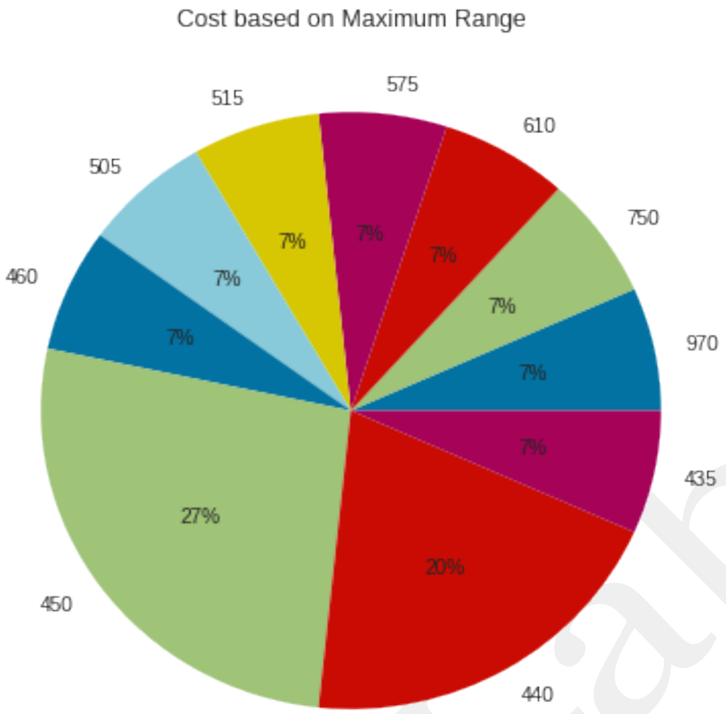
```
2.142167809024393e-12
```

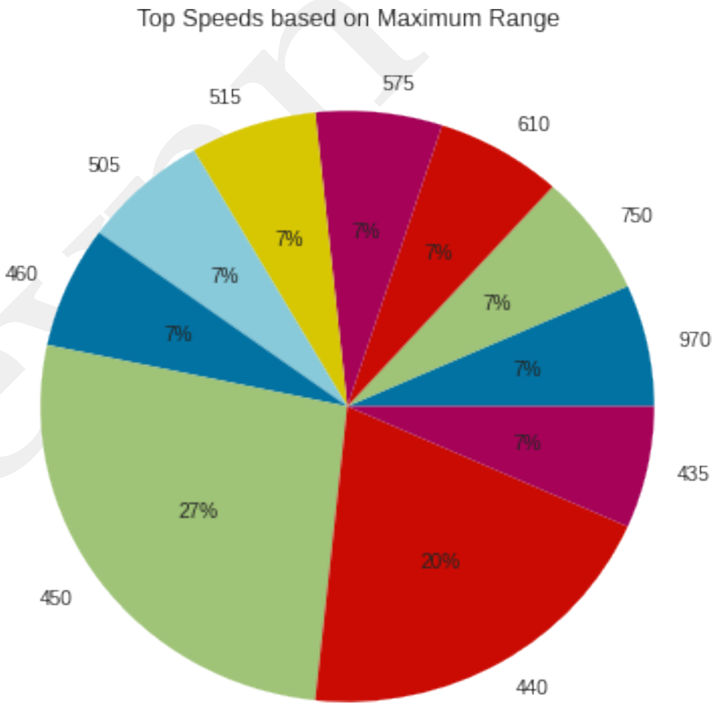
Profiling and Describing the Segments

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
515
575
610
505
750
460
970
435
450
440
7%
7%
7%
7%
7%
7%
7%
7%
27%
20%


Top Speeds based on Maximum Range
515
575
610
505
750
460
970
435
450
440
7%
7%
7%
7%
7%
7%
7%
7%
27%
20%

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:

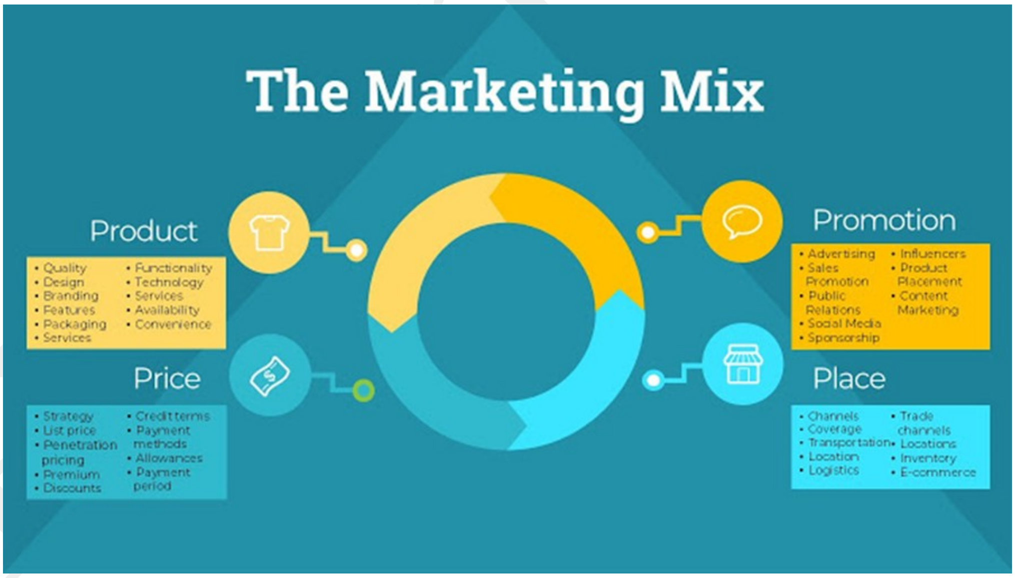
- *Top Speed & Range* : With a large area of market the cost is dependent on Topspeeds and Maximum range of cars.
- *Efficiency* : Mostly the segments are with most efficiency.

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**.

Customizing the Marketing Mix



The Marketing Mix

Product

- Quality
- Design
- Branding
- Features
- Packaging
- Services
- Functionality
- Technology
- Services
- Availability
- Convenience

Price

- Strategy
- List price
- Penetration pricing
- Premium
- Discounts
- Credit terms
- Payment methods
- Allowances
- Payment period

Promotion

- Advertising
- Sales Promotion
- Public Relations
- Social Media
- Sponsorship
- Influencers
- Product Placement
- Content Marketing

Place

- Channels
- Coverage
- Transportation
- Location
- Logistics
- Trade channels
- Locations
- Inventory
- E-commerce

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 tar-geted, ability of the companies to pay, ability of customers to pay supply - de- mand 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 thebest 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 con-sumer and making it easy for her to buy it is the main aim of a good distributionor '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 canalso include *consumer schemes, direct marketing, contests and prizes*.

All the elements of the marketing mix influence each other. They make up the businessplan 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.