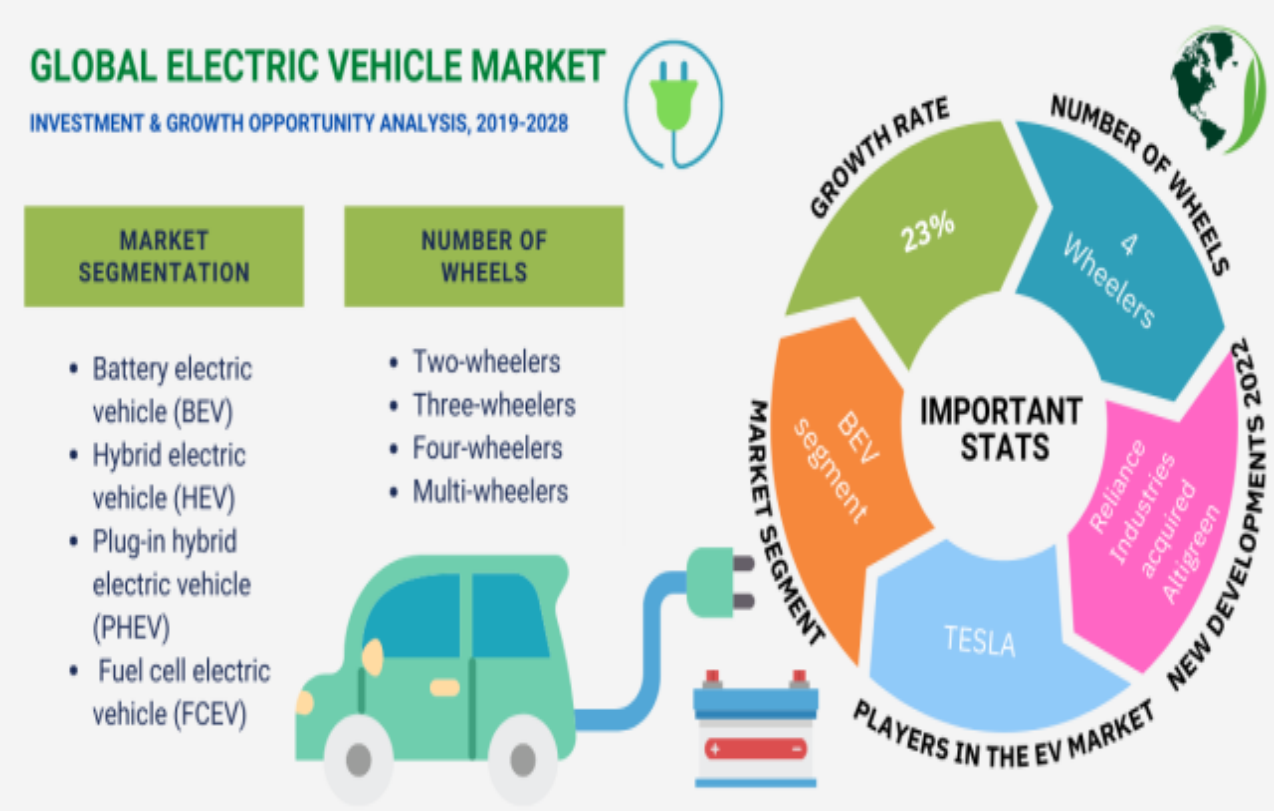
Electric Vehicles Market

Market Segmentation

Rutuja Thube



Abstract

In emerging regions, market segmentation becomes an essential tool for exploring and implementing changing transportation technologies, including electric vehicles (EVs), for widespread acceptance. Since EVs are low-emission and inexpensive to operate, their popularity is predicted to soar in the near future. As a result, they will likely drive a significant percentage of interest for upcoming scholarly research. This study's primary goal is to investigate and determine several groups of possible EV buyer categories based on psychographic, use an integrated research approach to characterize behavorial and socioeconomic aspects. "Perceived benefits-attitude-intention" framework. The research used strong analytical techniques such as Chi-square, multiple discriminant analysis, and cluster analysis test to operationalize and validate sections of the data.

Key Words : Electric vehicles, Market segmentation, Cluster analysis, Sustainable transportation, PCA, Feature Scaling.

Data Collection

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

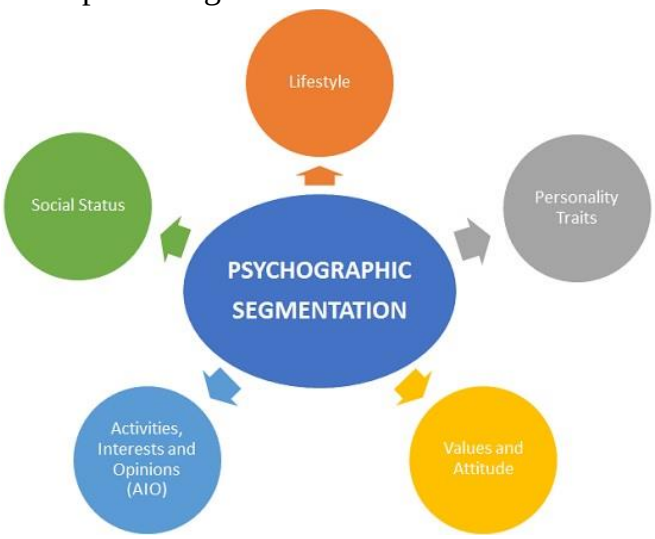
- https://www.kaggle.com/datasets
- https://www.data.gov/
- https://data.worldbank.org/
- https://datasetsearch.research.google.com/

Market Segmentation

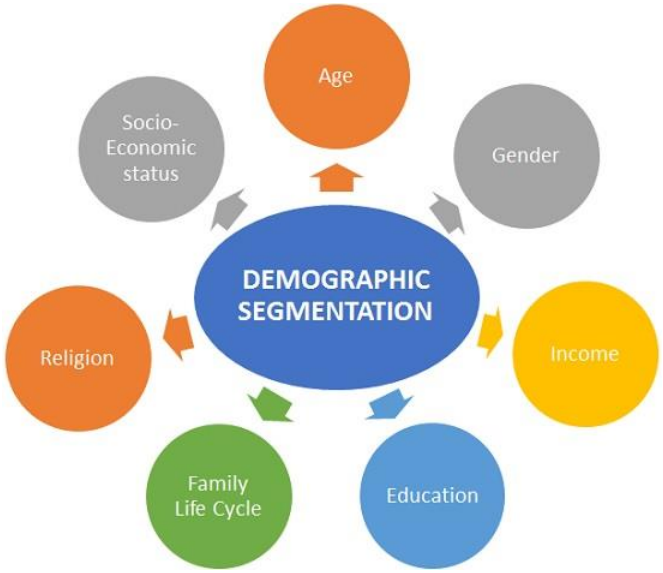
The target market of Electric Vehicle Market Segmentation can be categorized into Geographic, SocioDemographic, Behavioral, and Psychographic Segmentation.

1. **Psychographic Segmentation**:
   - Definition: Groups individuals based on lifestyle, values, attitudes, interests, and personality traits.

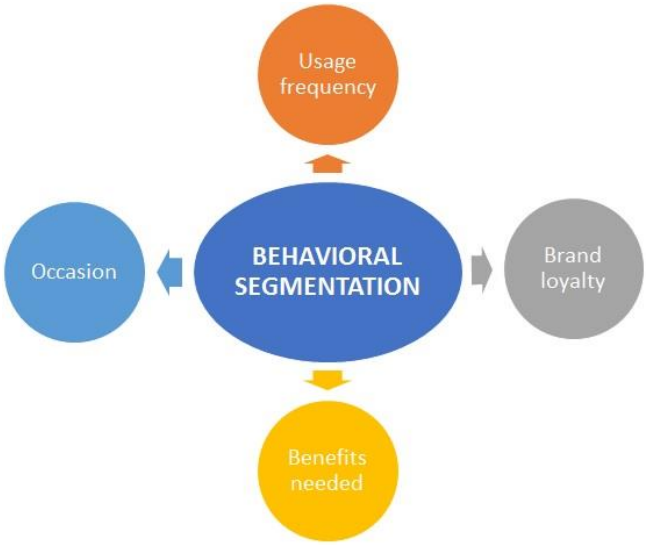
• Example: Segmenting car buyers based on their lifestyle choices, such as eco-conscious individuals preferring electric vehicles.



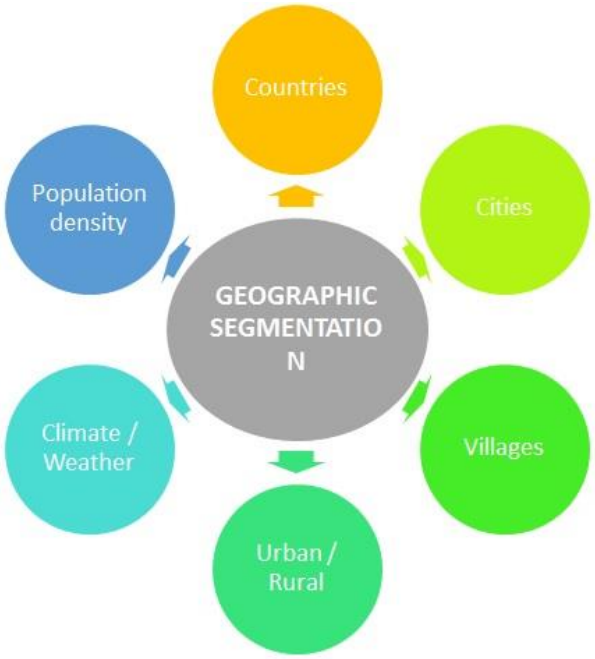
2. **Socio-Demographic Segmentation:**
   • Definition: Divides the market or dataset based on demographic variables such as age, gender, income, education, occupation, and family size.
   • Example: Segmenting car buyers by income level to target luxury vehicles to high-income groups.



3. **Behavorial Segmentation:**
   - Definition: Segments based on user behavior, such as purchasing patterns, usage frequency, brand loyalty, and product knowledge.
   - Example**:** Identifying frequent buyers of a particular car model and offering them loyalty rewards or upgrades.



4. **Geographic Segmentation:**
   - Definition: Segments the market based on geographic boundaries like region, country, city, climate, or urban vs. rural areas.
   - Example**:** A company may market different car models in urban and rural areas based on usage patterns and terrain.



Implementation

Packages and Tools Used

- Pandas (pandas): Used for data manipulation and analysis. It provides data structures like DataFrames to handle and analyze data efficiently.
- NumPy (numpy): Used for numerical operations, especially with arrays and matrices.
- Matplotlib (matplotlib): A plotting library used for creating static, animated, and interactive visualizations in Python.
- Seaborn (seaborn): A statistical data visualization library built on top of Matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.
- Scikit-Learn (sklearn): A comprehensive machine learning library in Python that provides simple and efficient tools for data mining, data analysis, and machine learning. Including algorithms such as KMeans, PCA and for encoding.

Data-Preprocessing

The data collected is compact and is partly used for visualization purposes and partly for clustering. Python libraries such as NumPy, Pandas, Scikit-Learn are used for the workflow, and the results obtained are ensured to be reproducible.

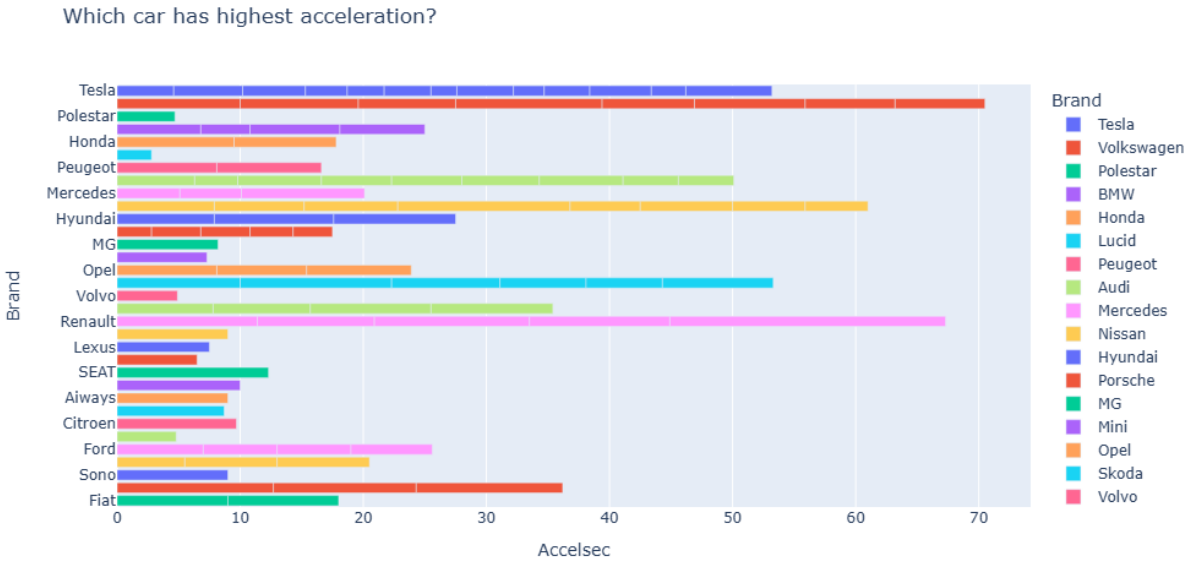
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
import plotly.express as px
import plotly.io as pio
```

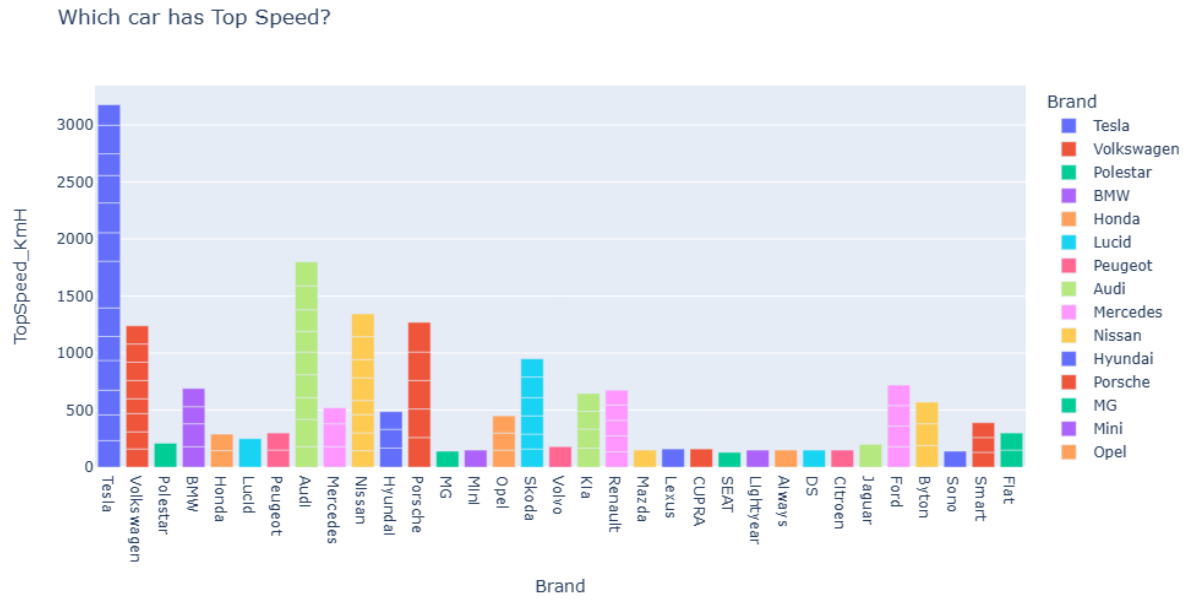
```
df=pd.read_csv("Electric Car Data.csv")
df.head()
```

	Brand	Model	Accelsec	TopSpeed_KmH	Range_Km	Battery_Pack Kwh	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	Plug Type	Body Style	Se
0	Tesla	Model 3 Long Range Dual Motor	4.6	233	460	70.0	161	940	Yes	AWD	Type 2 CCS	Sedan	
1	Volkswagen	ID.3 Pure	10.0	160	270	45.0	167	250	Yes	RWD	Type 2 CCS	Hatchback	
2	Polestar	2	4.7	210	400	75.0	181	620	Yes	AWD	Type 2 CCS	Liftback	
3	BMW	iX3	6.8	180	360	74.0	206	560	Yes	RWD	Type 2 CCS	SUV	
4	Honda	e	9.5	145	170	28.5	168	190	Yes	RWD	Type 2 CCS	Hatchback	

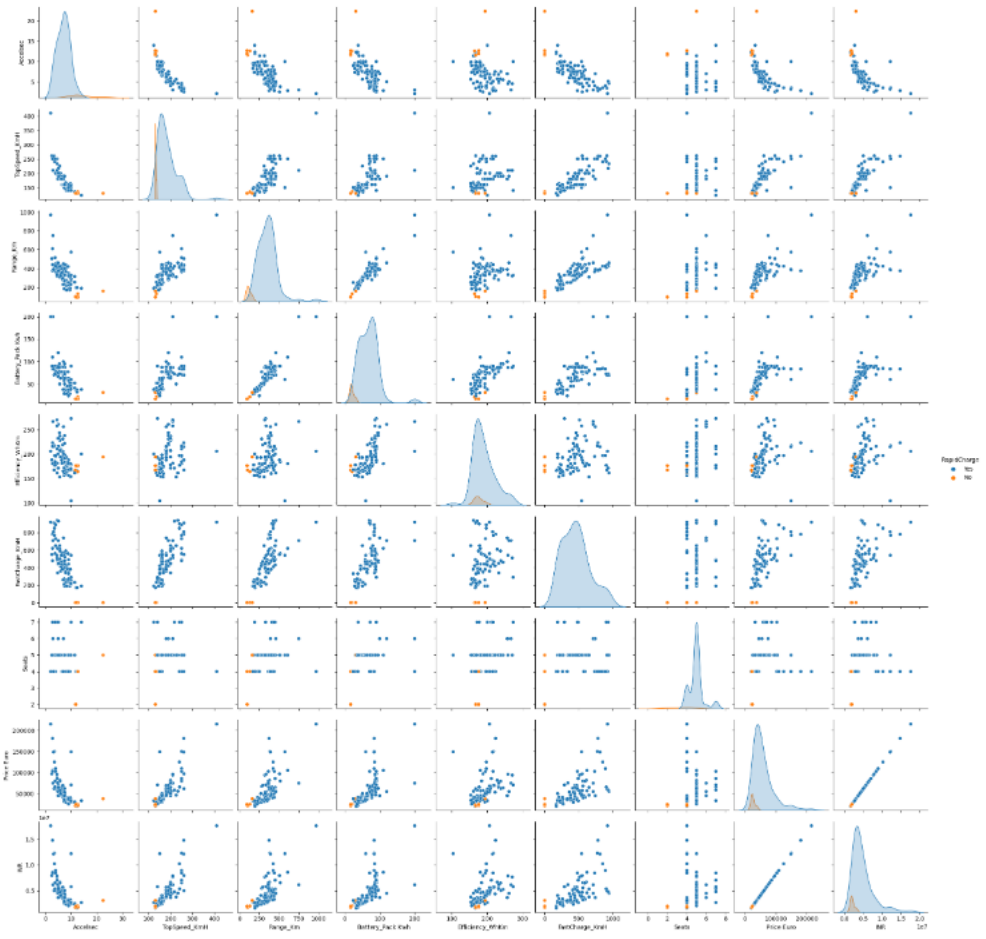
EDA

We begin the exploratory data analysis by analysing a portion of the data that was collected without principal component analysis and a portion of the dataset that was created by combining all of the available data. PCA is a statistical technique that uses orthogonal transformation to turn a set of correlated feature observations into a set of linearly uncorrelated features. The Principal Components are the newly altered features. The method aids in the reduction of data dimensionality, which lowers the cost of machine learning processes such as regression and classification.



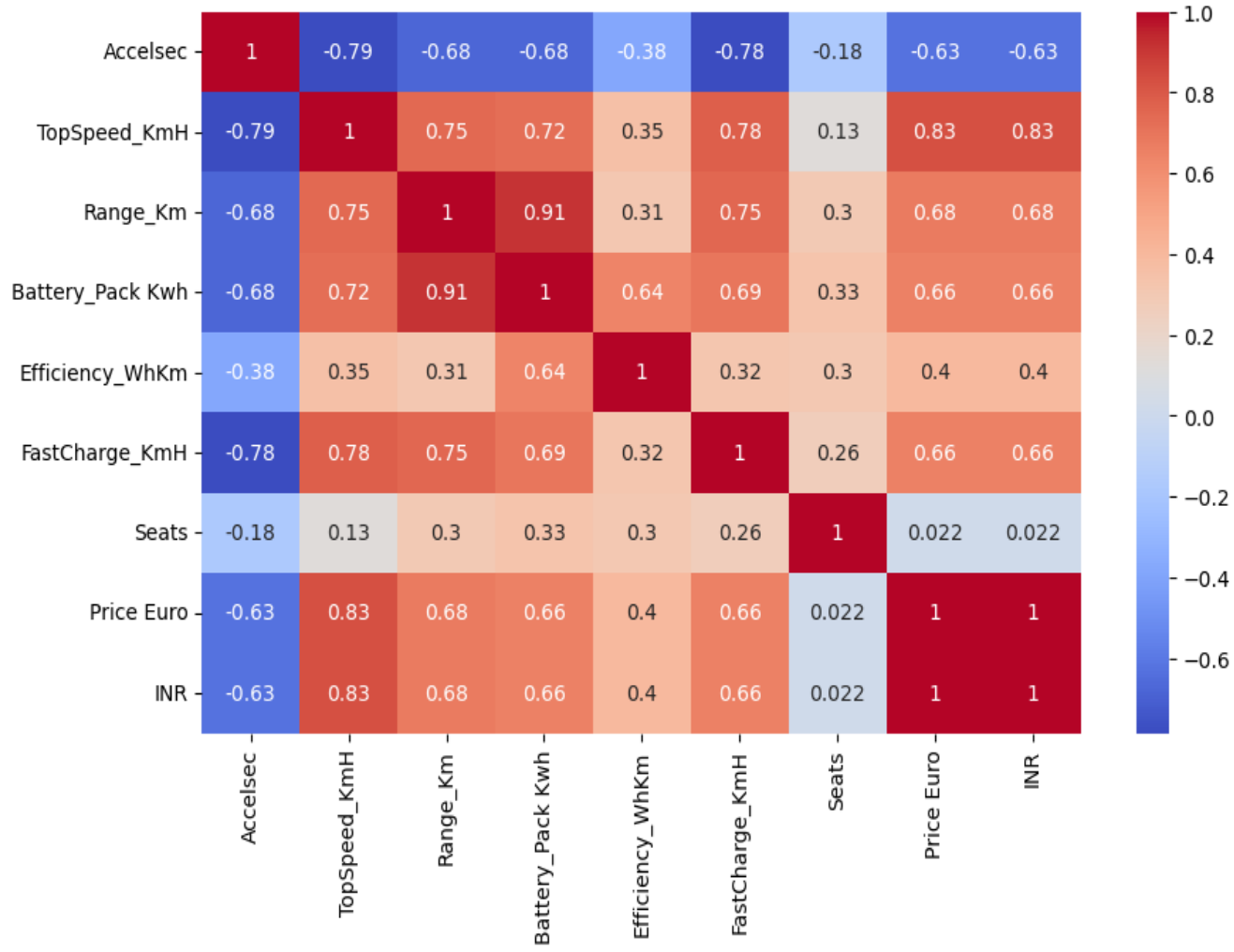


For Electric Vehicle Market one of the most important key is Charging:

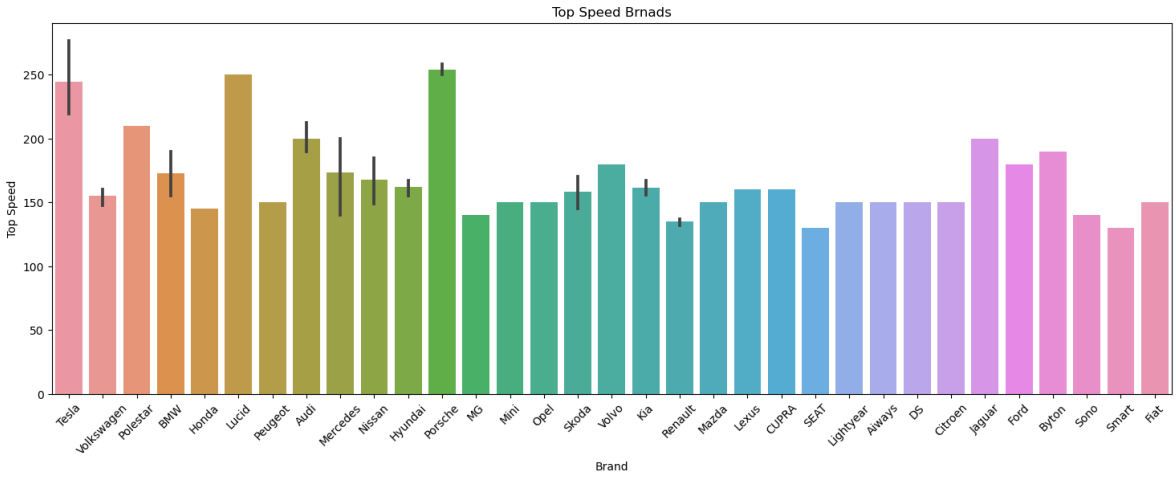


Correlation Matrix

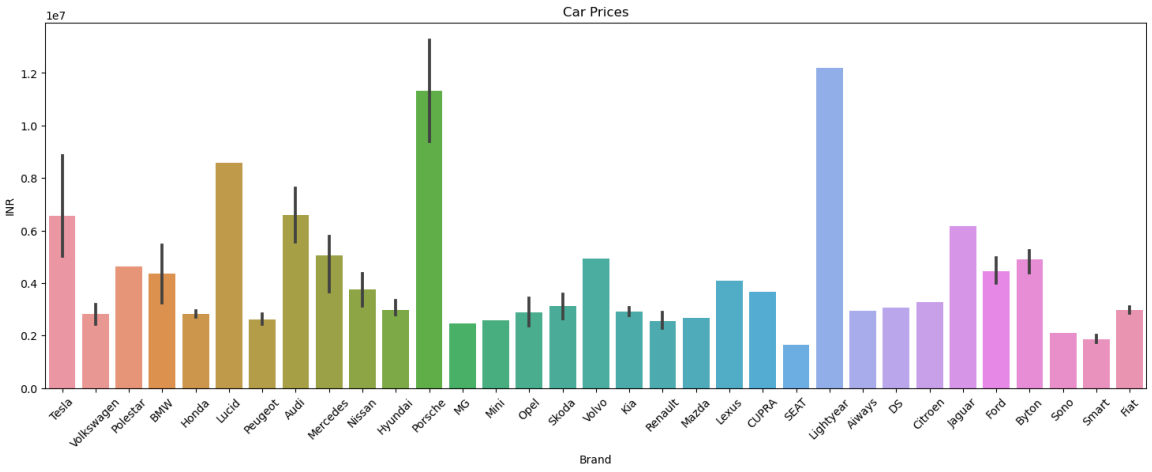
A correlation matrix is simply a table that displays the correlation. It is best used in variables that demonstrate a linear relationship between each other. Coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values through the heatmap in the below figure. The relationship between two variables is usually considered strong when their correlation coefficient value is larger than 0.7.



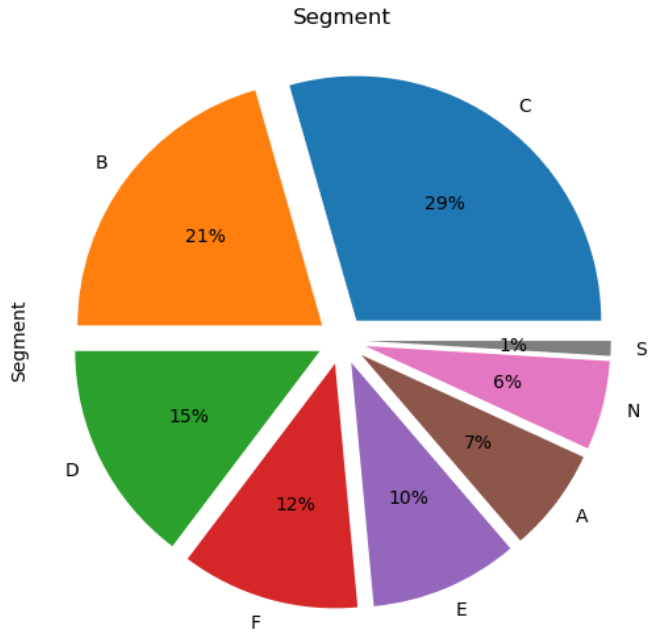
Top speeds achieved by the cars of a brand



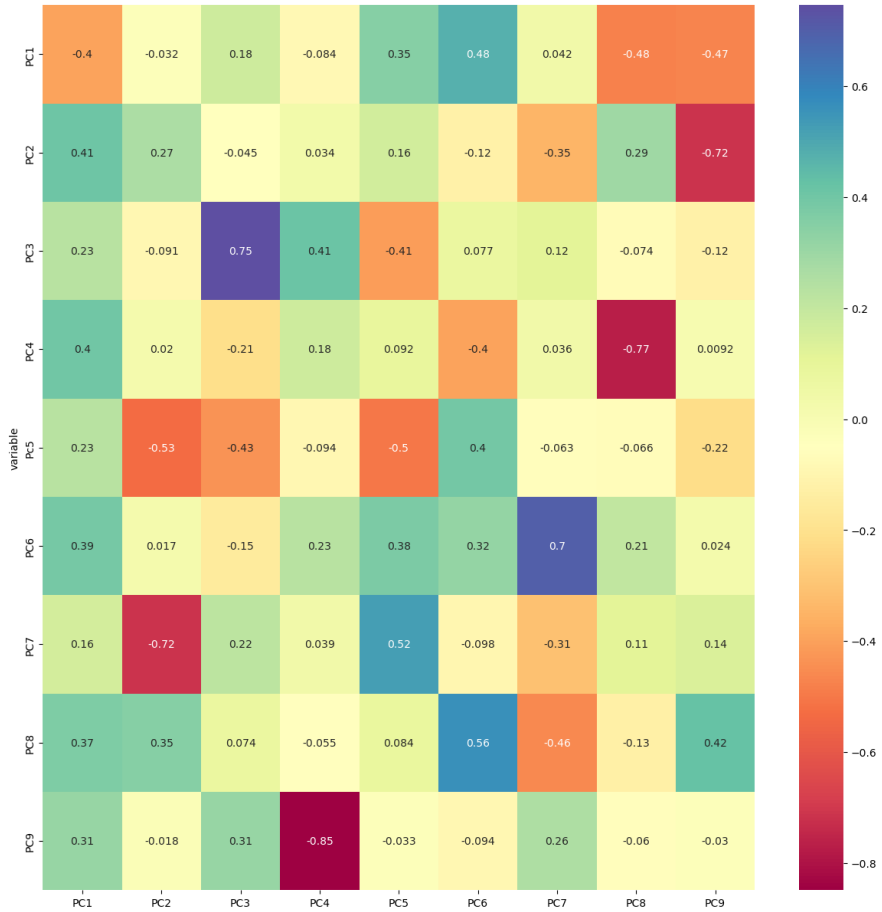
Price of cars (in INR)



Segment in which the cars fall under

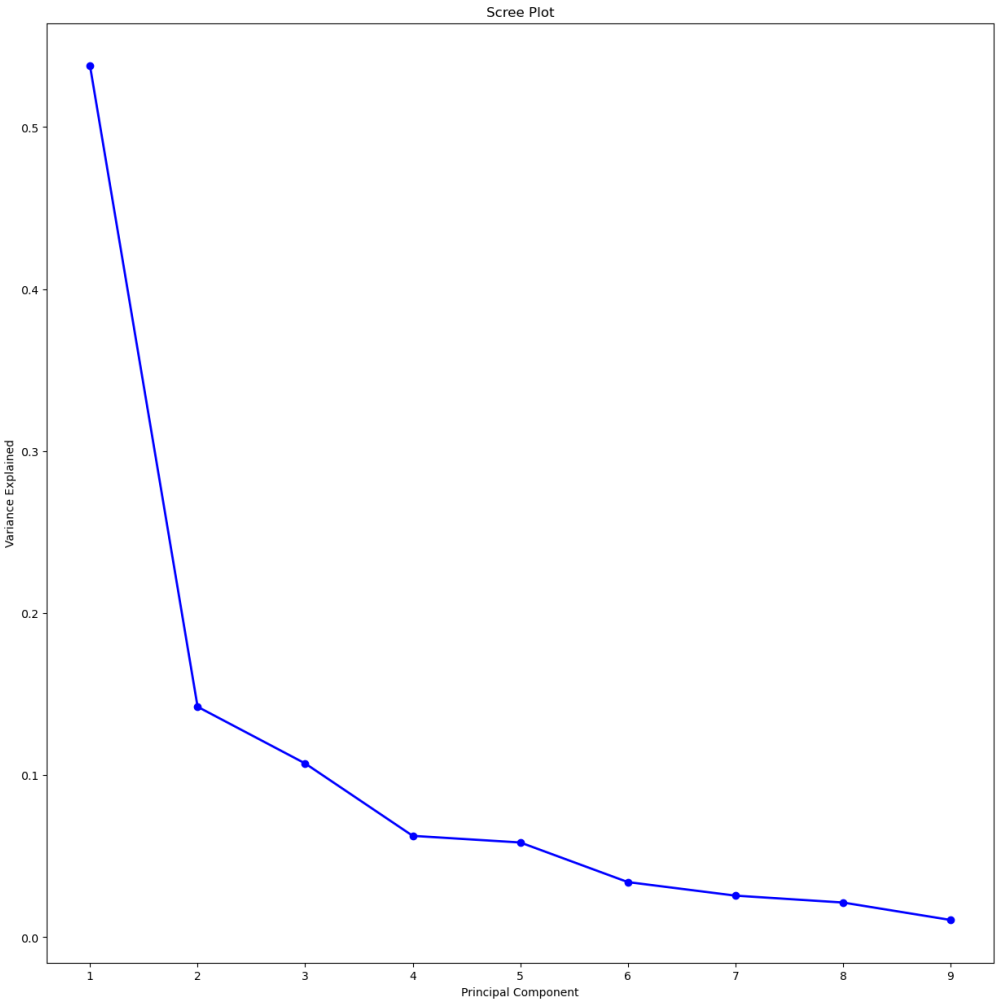


Correlation Matrix for loadings :



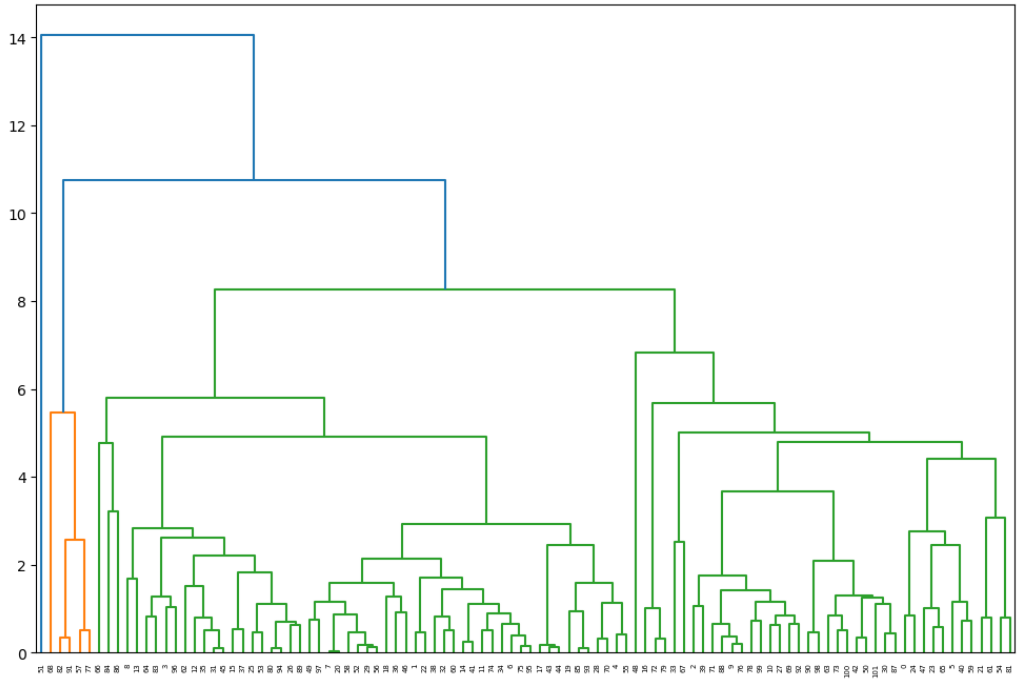
Scree Plot

It is a popular technique for calculating, using a graphical representation, how many PCs should be kept. The plot, which is a straightforward line segment plot, displays the eigenvalues for every PC. The y-axis displays the eigenvalues, and the x-axis displays the number of factors. It shows a descending gradient at all times. Most scree plots have a wide view. comparable in shape, with a steep beginning on the left, a rapid descent, and a flattening out eventually. This is due to the fact that the first element typically clarifies most of the variability, the subsequent components account for a modest portion, and the final components account for a negligible portion of the total variability. The criteria for scree plots searches for the curve's "elbow" and chooses all the parts right before the line flattens out. The proportion of variance plot: The selected PCs should be able to describe at least 80% of the variance.



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.



Elbow Method

In the k-means clustering algorithm we randomly initialize k clusters and we iteratively adjust these k clusters till these k-centroids riches in an equilibrium state. However, the main thing we do before initializing these clusters is that determine how many clusters we have to use. For determining K(numbers of clusters) we use Elbow method. Elbow Method is a technique that we use to determine the number of centroids(k) to use in a k-means clustering algorithm. In this method to determine the k-value we continuously iterate for k=1 to k=n (Here n is the hyperparameter that we choose as per our requirement). For every value of k, we calculate the within-cluster sum of squares (WCSS) value.

Analysis and Approaches used for Segmentation

Clustering

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group." It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns. It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabelled dataset. After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.

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. 1

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.

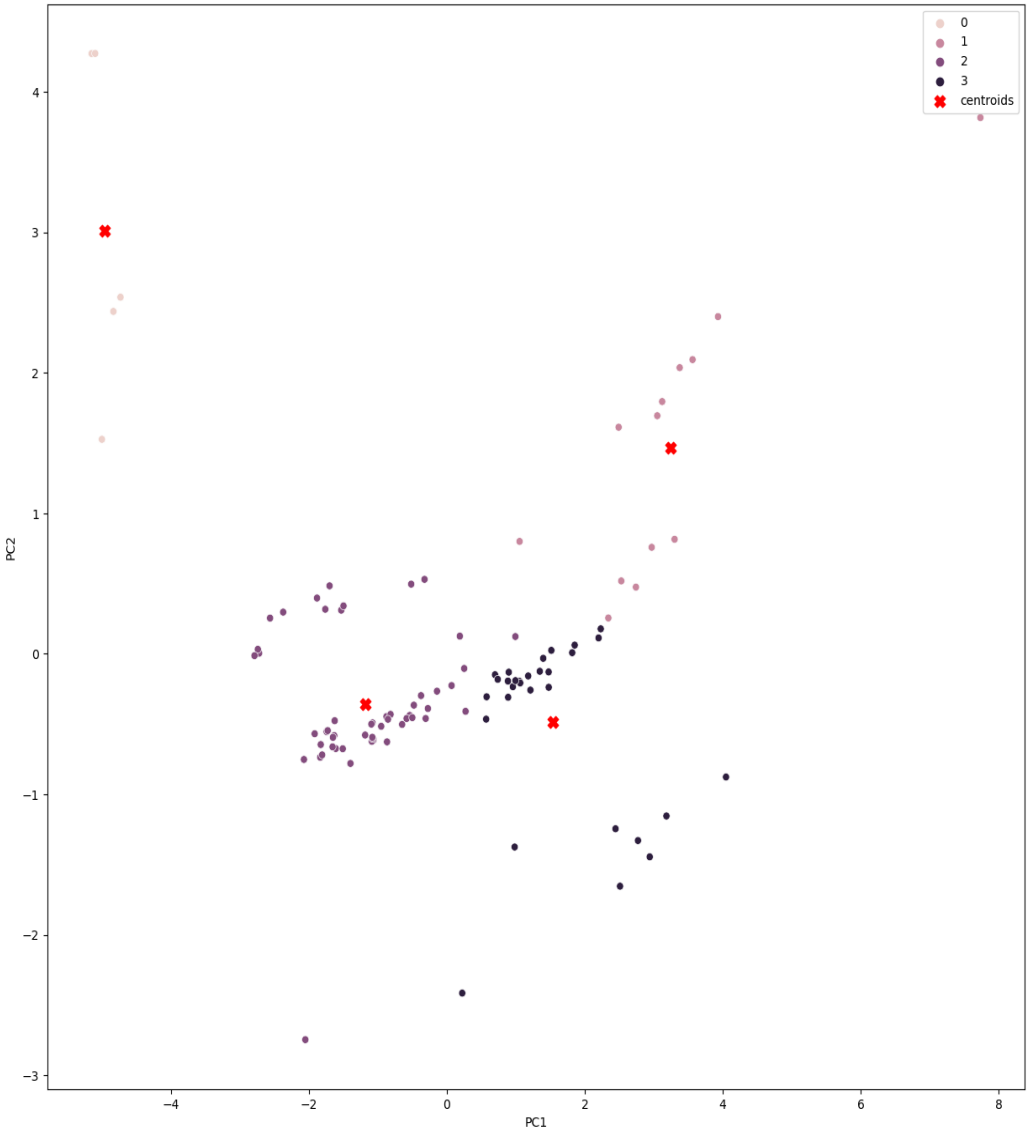
• Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters is not changing.

According to the Elbow method, here we take K=4 clusters to train K-Means model.

The derived clusters are shown in the following figure:

```
kmeans = KMeans(n_clusters=4, init='k-means++', random_state=0).fit(c)
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.
```

```
#Visulazing clusters
sns.scatterplot(data=data2, x="PC1", y="PC2", hue=kmeans.labels_)
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1],
            marker="X", c="r", s=80, label="centroids")
plt.legend()
plt.show()
```



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.