

Problem Statement: EV Market

Analysing the electric vehicle market in India using Segmentation analysis and deciding the strategy to enter the market, targeting the segments most likely used to use electric vehicles.

For this, we have decided to divide the market segmentation part into three segments.

1. Vehicle Feature Segmentation
2. Customer Behavioural Segmentation
3. Geographical Segmentation

Datasets used: Datasets used for above three market segmentation can be found at the following link:

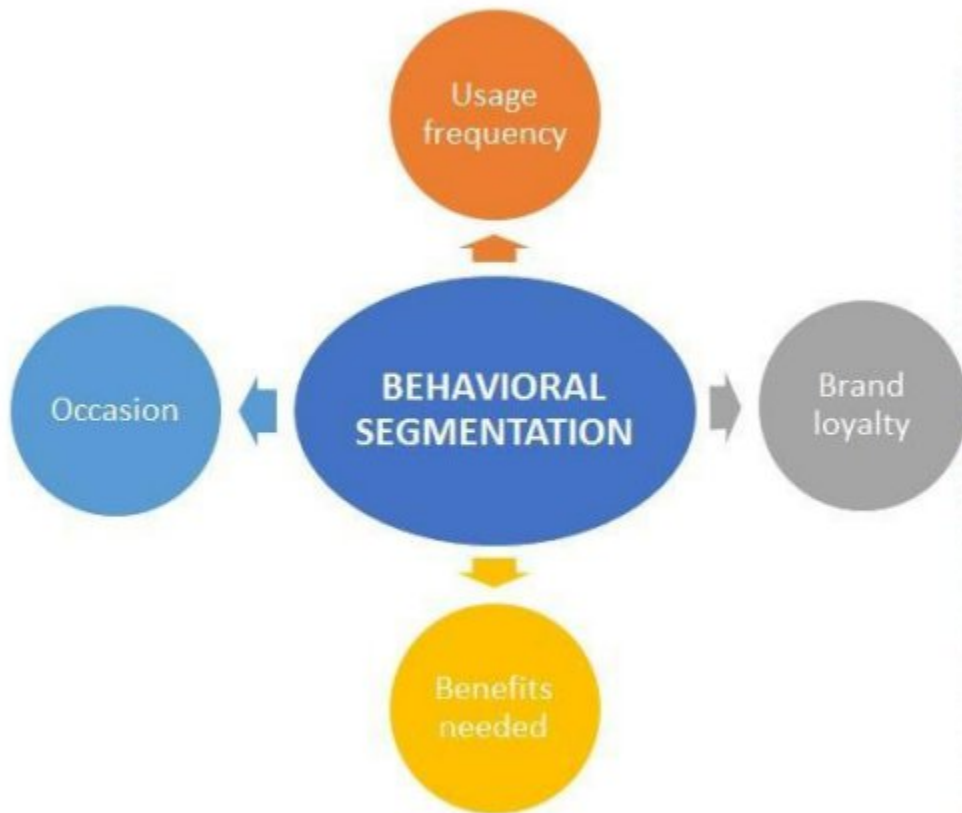
- <https://www.kaggle.com/datasets>
- <https://data.gov.in/>
- <https://data.worldbank.org/>
- <https://datasetsearch.research.google.com/>

Vehicle Feature Segmentation

Following Python libraries are imported for data analysis and visualization.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import seaborn as sb
import statsmodels.api as sm
from tqdm import tqdm
from google.colab import files
!pip install kaleido
import kaleido
from klearn.preprocessing import StandardScaler, PowerTransformer
from klearn.decomposition import PCA
from scipy.cluster.hierarchy import dendrogram, linkage
from klearn.cluster import KMeans, MeanShift, estimate_bandwidth
from klearn.datasets import make_blobs
from yellowbrick.cluster import KElbowVisualizer, SilhouetteVisualizer, InterclusterDistance
from collections import Counter
from klearn.model_selection import cross_validate, train_test_split
from klearn.linear_model import LinearRegression, LogisticRegression
from klearn.metrics import r2_score, silhouette_score, confusion_matrix, accuracy_score
```

Behavioural Segmentation of EV Market



Behavioural 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.

Behavioural segmentation refers to a process in marketing which divides customers into segments depending on their behaviour patterns when interacting with a particular business or website.

These segments could include grouping customers by-

1. Their attitude toward your product, brand or service.
2. Their use of your product or service.
3. Their overall knowledge of your brand and your brand's products.
4. Their purchasing tendencies, such as buying on special occasions like birthdays or holidays only, etc.

Going beyond the traditional demographic and geographic segmentation methods and utilizing behavioural data allows for the execution of more successful marketing campaigns.

At the very least, behavioural segmentation offers marketers and business owners a more complete understanding of their audience, thus enabling them to tailor products or services to specific customer needs. Below we take a look at four more benefits of behavioural segmentation.

We are using a dataset. By analysing dataset, we predict the customers behaviour towards buying EV vehicle. First of all, we read the dataset using panda library and display the first 5 rows so that we can know about the number of columns and their characteristics.

```
# This data contains the details about consumers who purchased an EV
df = pd.read_csv("/content/drive/MyDrive/Datasets/Electric_vehicle _dataset/behavioural_dataset.csv")

df.head()
```

	Age	Profession	Marrital Status	Education	No of Dependents	Personal loan	Total Salary	Price
0	27	Salaried	Single	Post Graduate	0	Yes	800000	800000
1	35	Salaried	Married	Post Graduate	2	Yes	2000000	1000000
2	45	Business	Married	Graduate	4	Yes	1800000	1200000
3	41	Business	Married	Post Graduate	3	No	2200000	1200000
4	31	Salaried	Married	Post Graduate	2	Yes	2600000	1600000

In this dataset, we observe that there are 8 columns namely Age, Profession, Marital Status, Education, No of Dependents, Personal loan, Total Salary and Price.

Description of the dataset

```
df.describe()
```

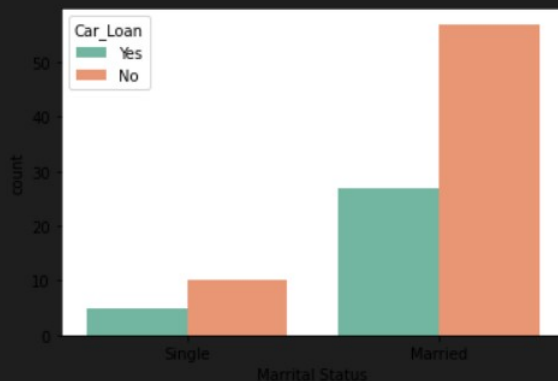
	Age	No of Dependents	Total Salary	Price
count	99.000000	99.000000	9.900000e+01	9.900000e+01
mean	36.313131	2.181818	2.270707e+06	1.194040e+06
std	6.246054	1.335265	1.050777e+06	4.376955e+05
min	26.000000	0.000000	2.000000e+05	1.100000e+05
25%	31.000000	2.000000	1.550000e+06	8.000000e+05
50%	36.000000	2.000000	2.100000e+06	1.200000e+06
75%	41.000000	3.000000	2.700000e+06	1.500000e+06
max	51.000000	4.000000	5.200000e+06	3.000000e+06

Here we observe that maximum age of the person who bought EV car is 51 and mean of such ages is 36.31 years. Maximum salary of the person who bought the car is 5200000 while minimum salary is 200000 and the average of the salary is 2270707.

Checking the marital status of the person with car loan

```
# Plotting the Car loan status with respect to Marrital Status
sns.countplot(x='Marrital Status', hue='Car_Loan', data=df, palette='Set2')
plt.show()
```

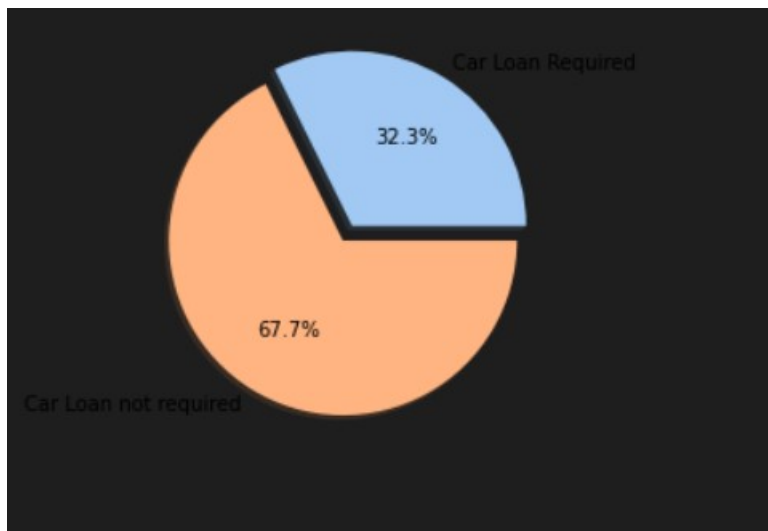
<matplotlib.axes._subplots.AxesSubplot at 0x7f38408e3f40>



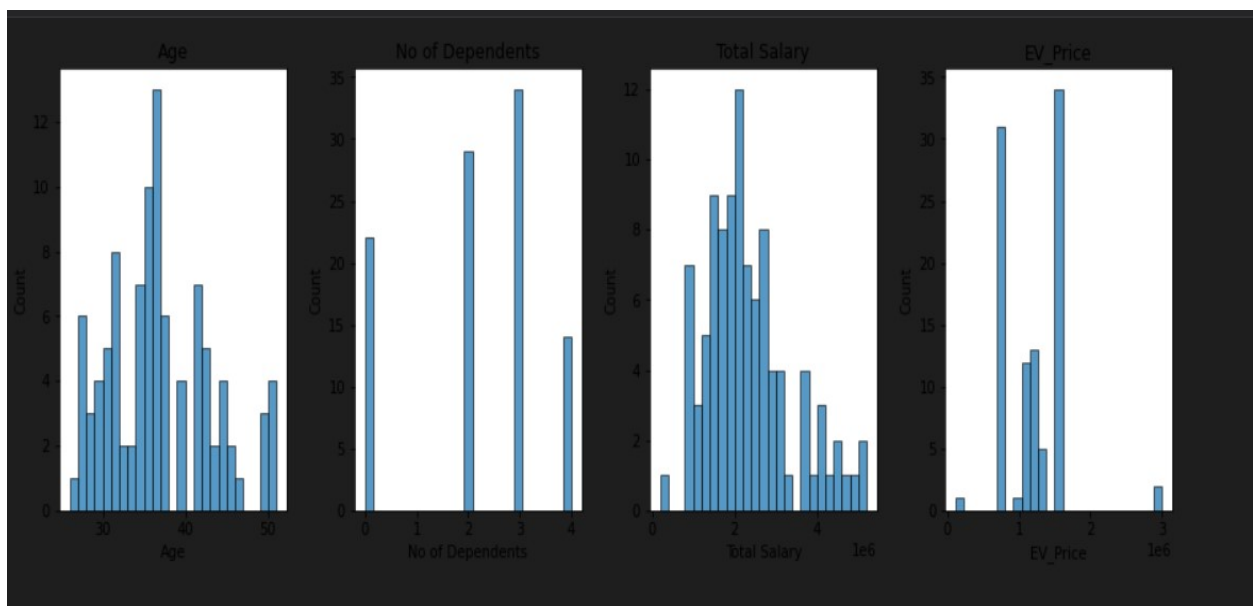
By seeing the plot , It is clear that there are more persons(single or married) do not need car loan to buy the car.

Car Loan

67.7% of the total person who bought EV car do not need car loan.



Plotting the frequency of customers against Age, No of Dependents, Total Salary and EV_price.



1. Persons having the age between 30 and 40 has bought most of cars.

2. Person having 4 family members are less attractive towards buying cars.
3. Persons who have medium level salary bought Ev cars more.
4. EV cars having medium price range are sold a lot

KModes clustering

KModes clustering is one of the unsupervised Machine Learning algorithms that is used to cluster categorical variables.

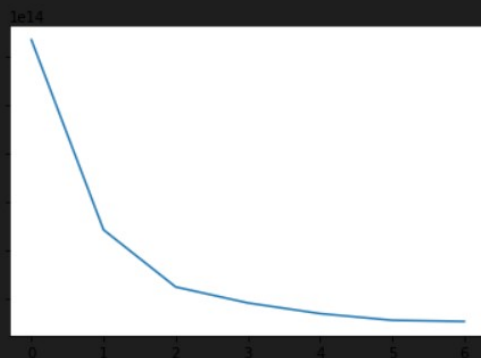
KMeans uses mathematical measures (distance) to cluster continuous data. The lesser the distance, the more similar our data points are. Centroids are updated by Means. But for categorical data points, we cannot calculate the distance. So we go for KModes algorithm. It uses the dissimilarities (total mismatches) between the data points. The lesser the dissimilarities the more similar our data points are. It uses Modes instead of means In K-means clustering when we used categorical data after converting it into a numerical form. it doesn't give a good result for high-dimensional data. So, Some changes are made for categorical data.

1. Replace Euclidean distance with Dissimilarity metric
2. Replace Mean by Mode for cluster centres.
3. Apply a frequency-based method in each iteration to update the mode.

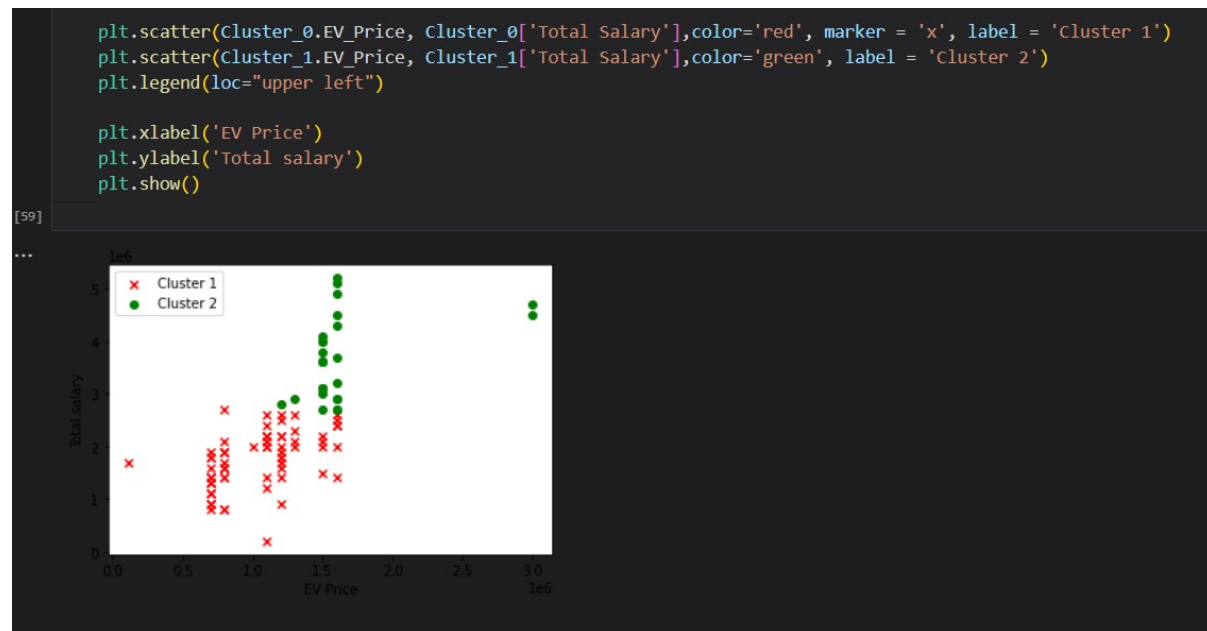
Finding optimal number of clusters for KPrototypes

```
cost = []
for num_clusters in list(range(1,8)):
    kproto = KPrototypes(n_clusters=num_clusters, init='Cao')
    kproto.fit_predict(cluster_data, categorical=[1,2,3,5])
    cost.append(kproto.cost_)

plt.plot(cost)
plt.show()
```



There is bending in the line at point 2. Hence number of clusters formed is 2 for better visualizations.



There is clear difference in the segments in comparing salary and price of Ev cars.

Conclusions 2

1. The maximum age of the person who bought EV car is 51 year and mean of such ages is 36.31 years. Maximum salary of the person who bought the car is 5200000 while minimum salary is 2000000 and the average of the salary is 2270707.
2. There are more persons (single or married) do not need car loan to buy the car.
3. 67.7% of the total person who bought EV car do not need car loan.
4. Persons having the age between 30 and 40 has bought most of cars.
5. Person having 4 family members are less attractive towards buying cars.
6. Persons who have medium level salary bought Ev cars more.
7. EV cars having medium price range are sold a lot.