REPORT ON 2nd PROJECT

REQUIREMENTS

To perform this we need a data set and a python libraries such as Pandas, Matplotlib, Seaborn ,Scikit-learn and kaleido.

Importing libraries:

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

```
import warnings
warnings.filterwarnings("ignore")
```

```
from scipy.spatial import ConvexHull
```

```
#Import the libraries
import plotly.express as px
import kaleido
```

Loading the dataset:

Step- 2 We need to Load the Dataset

```
df=pd.read_csv("Electric-car-data.csv")
```

```
data = pd.read_csv("behavioural_dataset.csv")
```

Dataset 1-

```
df
```

	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment	S
0	Tesla	Model 3 Long Range Dual Motor	4.6	233	450	161	940	Yes	AWD	Type 2 CCS	Sedan	D	
1	Volkswagen	ID.3 Pure	10.0	160	270	167	250	Yes	RWD	Type 2 CCS	Hatchback	C	
2	Polestar	2	4.7	210	400	181	620	Yes	AWD	Type 2 CCS	Liftback	D	
3	BMW	iX3	6.8	180	360	206	560	Yes	RWD	Type 2 CCS	SUV	D	
4	Honda	e	9.5	145	170	168	190	Yes	RWD	Type 2 CCS	Hatchback	B	
...	...	...	...	...	...	...	...	...	...	...	...	...	
98	Nissan	Ariya 63kWh	7.5	160	330	191	440	Yes	FWD	Type 2 CCS	Hatchback	C	
99	Audi	e-tron S Sportback 55 quattro	4.5	210	335	258	540	Yes	AWD	Type 2 CCS	SUV	E	
100	Nissan	Ariya e-4ORCE 63kWh	5.9	200	325	194	440	Yes	AWD	Type 2 CCS	Hatchback	C	
101	Nissan	Ariya e-4ORCE 87kWh Performance	5.1	200	375	232	450	Yes	AWD	Type 2 CCS	Hatchback	C	
102	Byton	M-Byte 95 kWh 2WD	7.5	190	400	238	480	Yes	AWD	Type 2 CCS	SUV	E	

103 rows × 14 columns

Dataset 2-

	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
...	...	...	...	...	...	...	...	...
94	27	Business	Single	Graduate	0	No	2400000	1600000
95	50	Salaried	Married	Post Graduate	3	No	5100000	1600000
96	51	Business	Married	Graduate	2	Yes	2200000	1100000
97	51	Salaried	Married	Post Graduate	2	No	4000000	1500000
98	51	Salaried	Married	Post Graduate	2	Yes	2200000	1100000

99 rows × 8 columns

Performing EDA:

Exploratory Data Analysis (EDA) is a crucial step in data analysis that involves examining and understanding the structure, patterns, and characteristics of a dataset.

Exploratory Data Analysis (EDA) is a crucial step in data analysis that involves examining and understanding the structure, patterns, and characteristics of a dataset. Here's a general outline of how you might approach EDA:

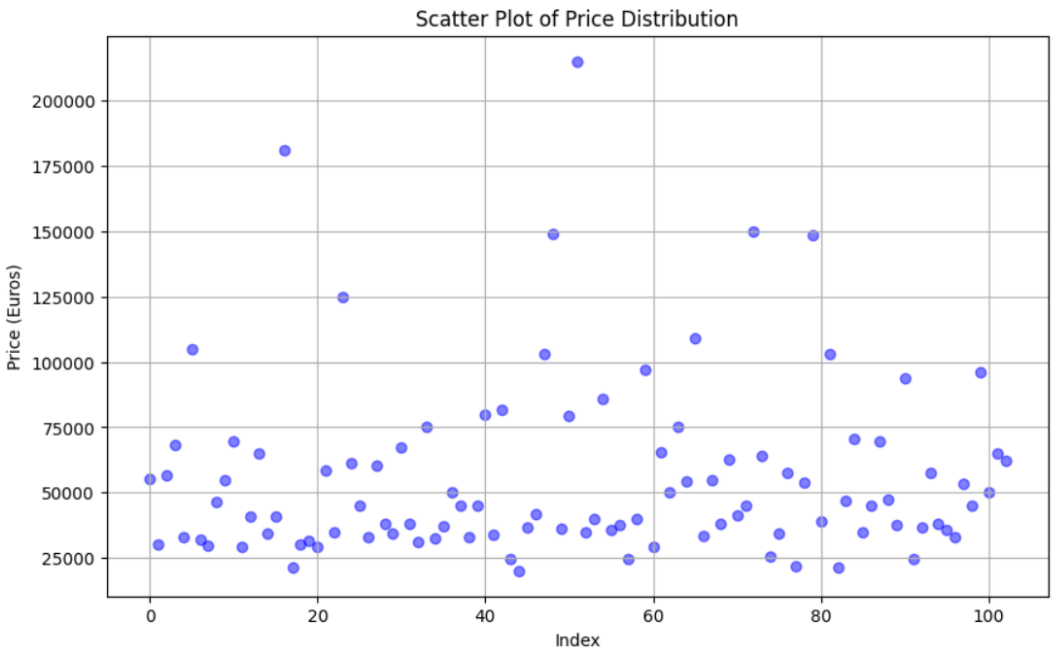
1. **Data Collection**: Gather the dataset you want to analyze. This could be from various sources such as databases, CSV files, APIs, etc.
2. **Initial Inspection**:
    - Check the first few rows of the dataset to understand its structure.
    - Look for missing values, outliers, and inconsistencies in the data.
3. **Summary Statistics**:
    - Calculate basic statistics such as mean, median, mode, standard deviation, minimum, maximum, etc., for numerical columns.
    - For categorical variables, count the frequency of each category.
4. **Data Visualization**:
    - Use plots such as histograms, box plots, scatter plots, and bar charts to visualize the distribution of numerical data, identify outliers, and understand relationships between variables.
    - For categorical data, use bar charts, pie charts, and stacked bar charts to visualize the distribution of categories.
5. **Correlation Analysis**:
    - Calculate correlation coefficients (e.g., Pearson correlation for numerical variables) to understand the linear relationship between variables.
    - Visualize correlations using heatmaps for better interpretation.
6. **Feature Engineering**:
    - Create new features if necessary based on domain knowledge and insights gained during EDA.
    - Transform variables (e.g., log transformation for skewed data) to make them more suitable for modelling.
7. **Handling Missing Values and Outliers**:
    - Decide on strategies to handle missing values (e.g., imputation, deletion) based on the extent of missingness and domain knowledge.
    - Identify and potentially remove outliers that can significantly affect analysis and modelling.
8. **Data Quality Check**:
    - Validate data quality by cross-checking against domain knowledge and business rules.
    - Ensure data consistency and correctness.
9. **Data Segmentation**:

- If applicable, segment the data based on certain criteria to analyse subsets separately (e.g., segmenting customers based on demographics).

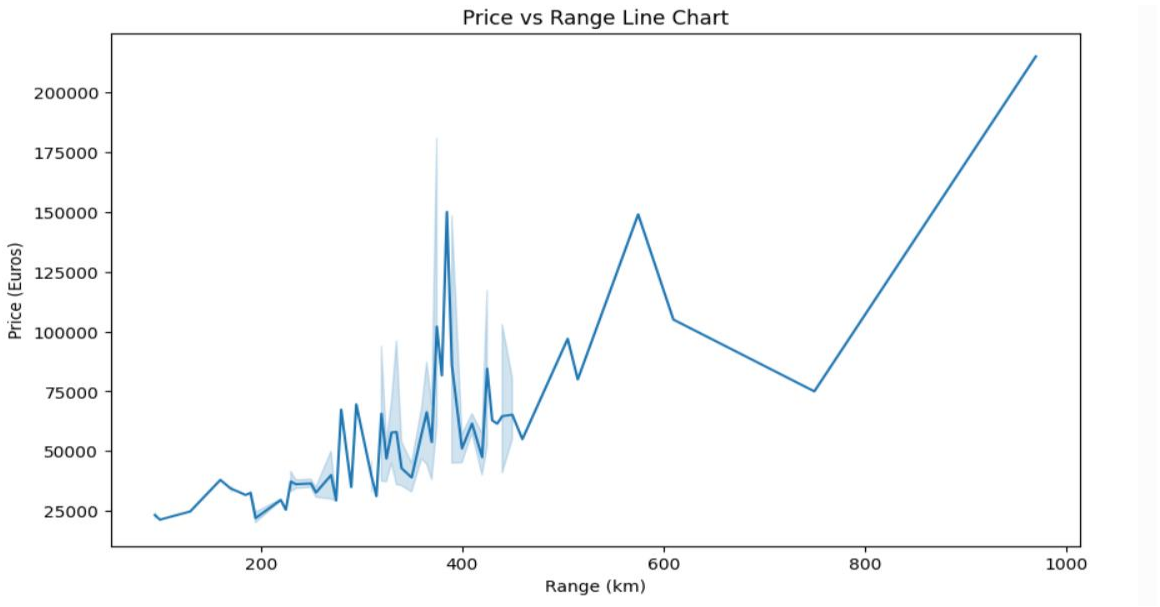
10. **Documentation**:
    - Document your findings, insights, and decisions made during EDA.
    - Summarize key takeaways and prepare

Implementing EDA:

Scatter plot:

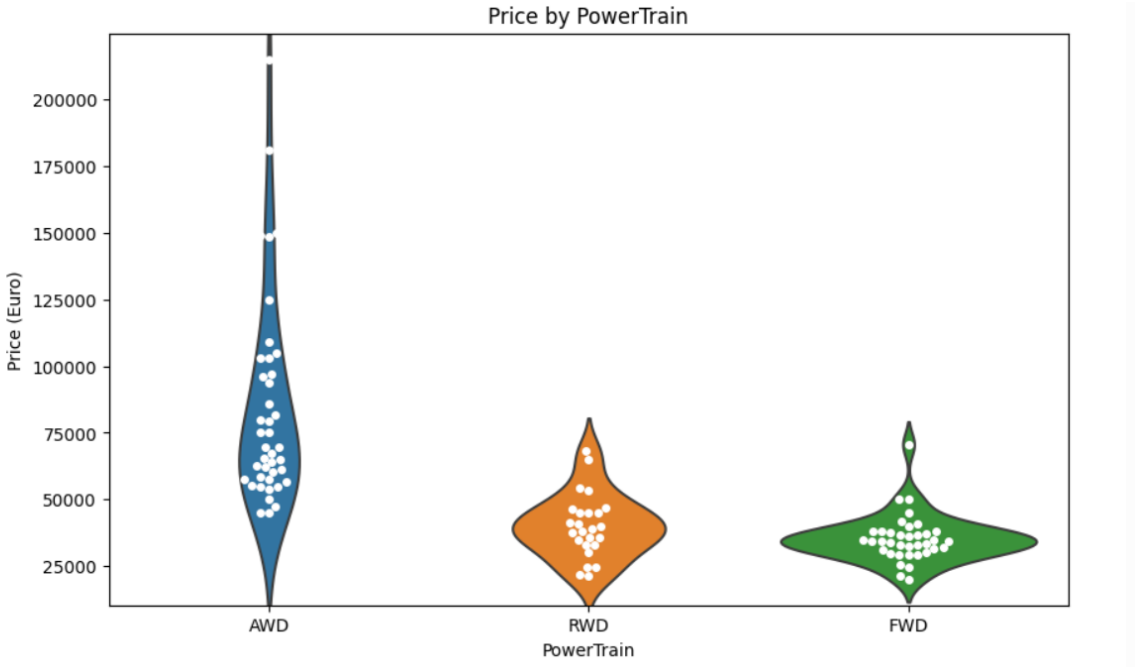


Line chart :



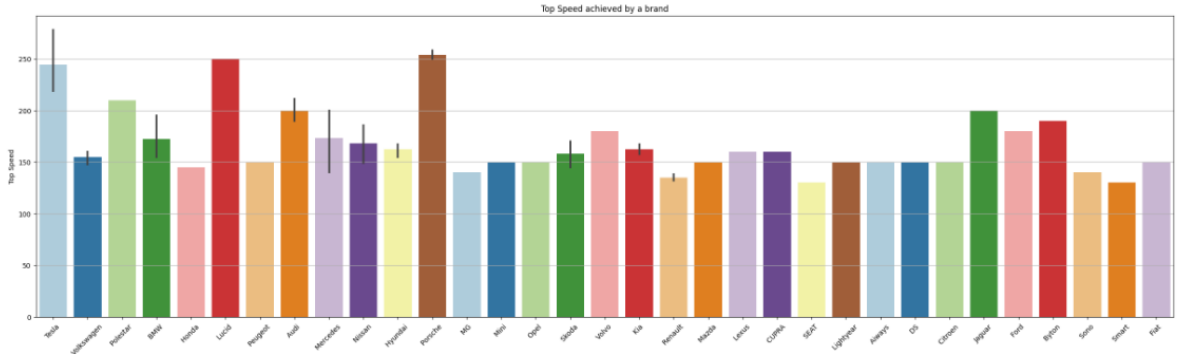
In this visualize between range (kilometres) and price.

Violin chart:



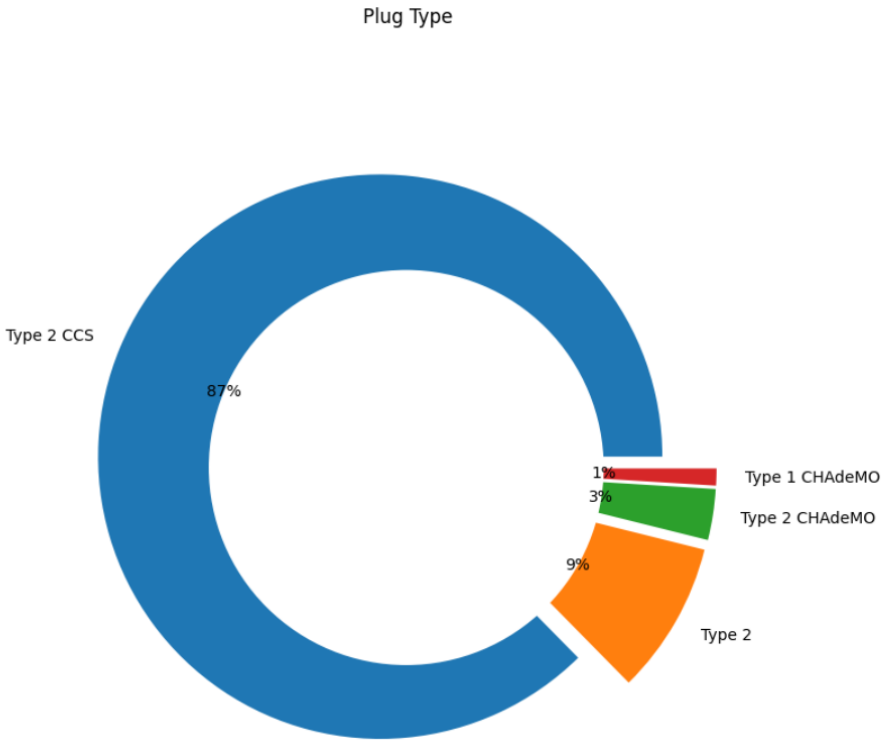
Bar graph:

In this we are going to explore the top speed of each vehicle brand.

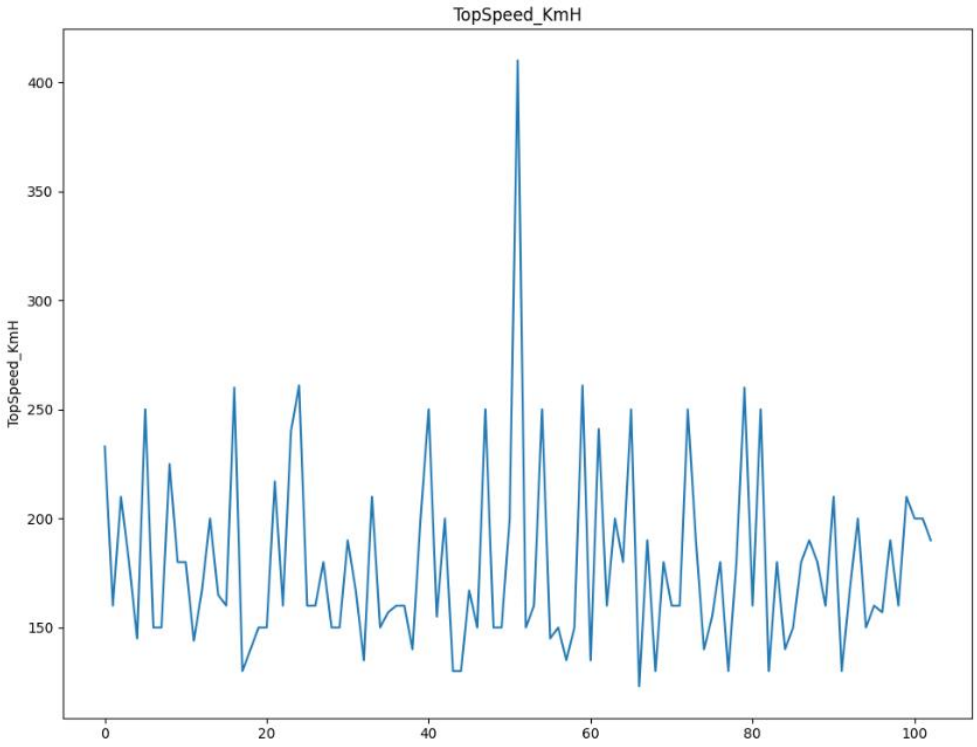


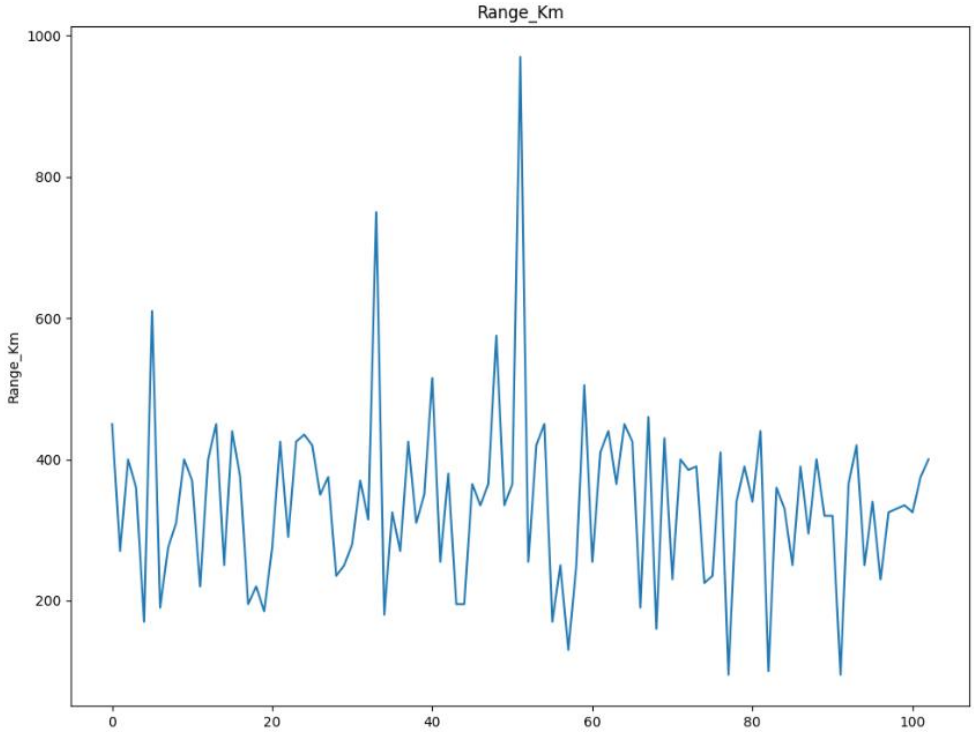
Doughnut chart:

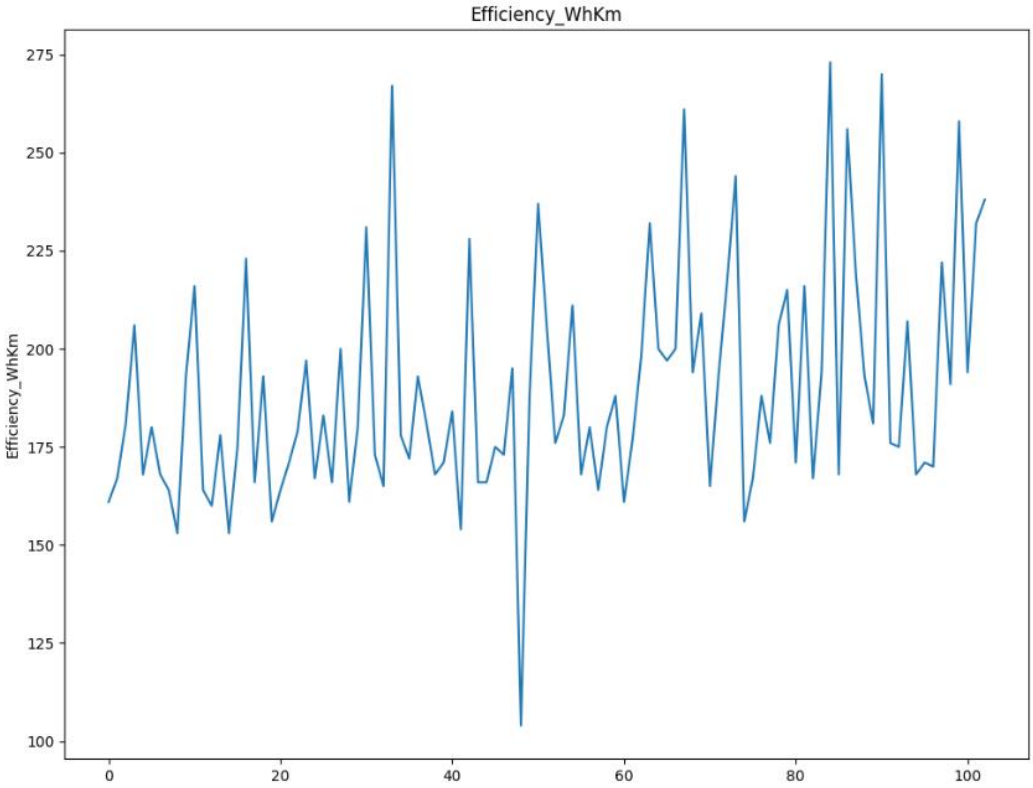
In this we are looking at different types of plug time and their percentage of use.

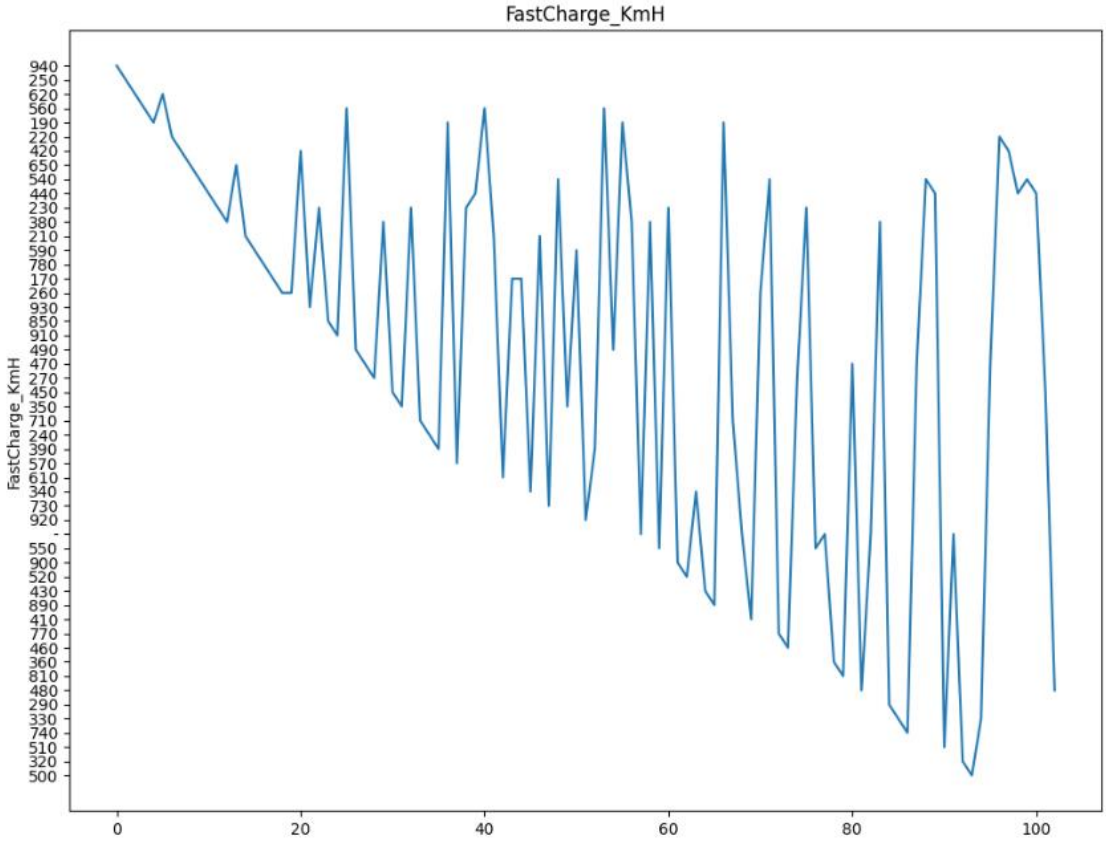


Line chart:


TopSpeed_KmH
TopSpeed_KmH
400
350
300
250
200
150
0
20
40
60
80
100

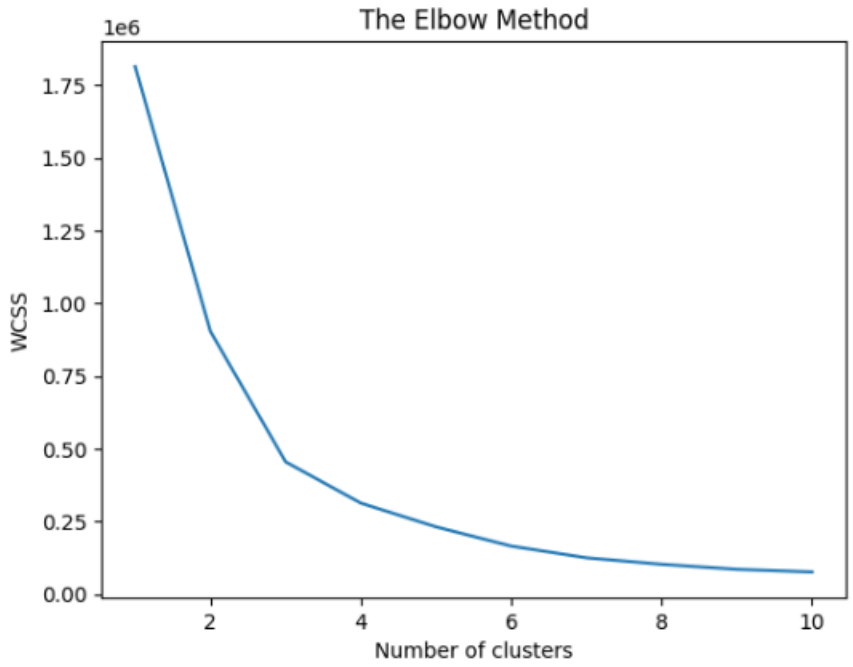

Range_Km
Range_Km
1000
800
600
400
200
0
20
40
60
80
100


Efficiency_WhKm
Efficiency_WhKm
275
250
225
200
175
150
125
100
0
20
40
60
80
100

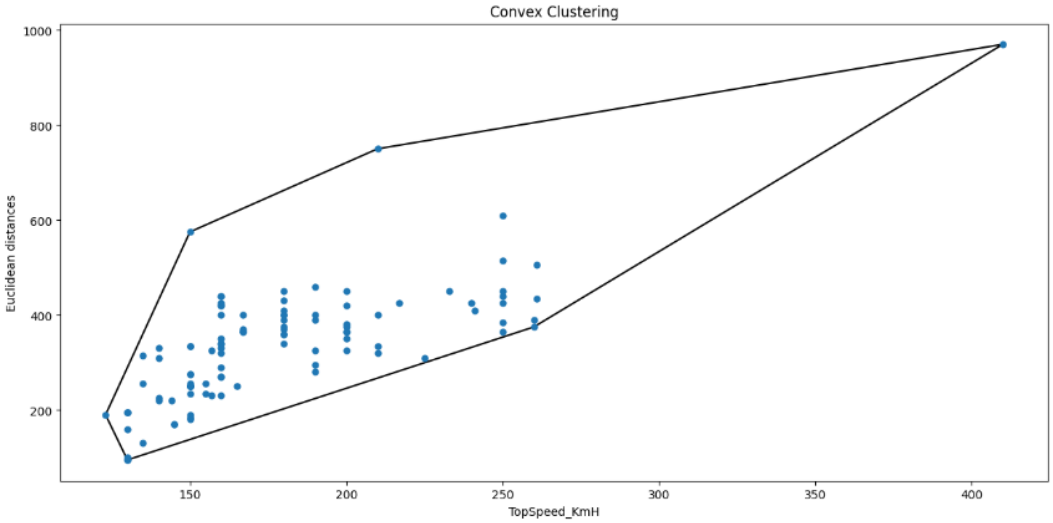

FastCharge_KmH
FastCharge_KmH
0
20
40
60
80
100

K-MEANS CLUSTERING

Elbow Method:

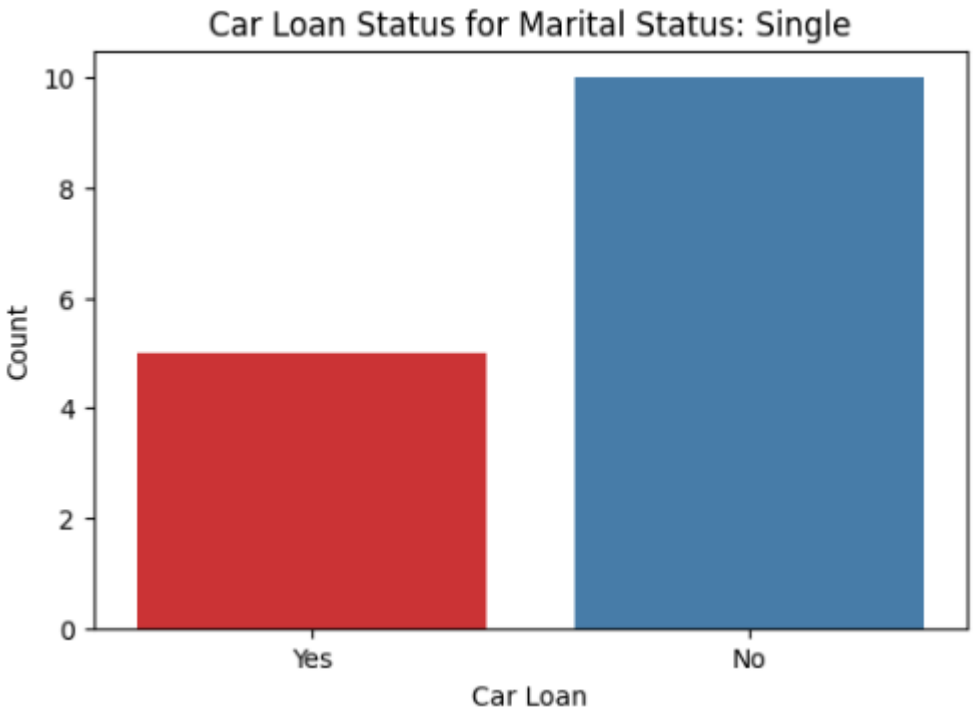


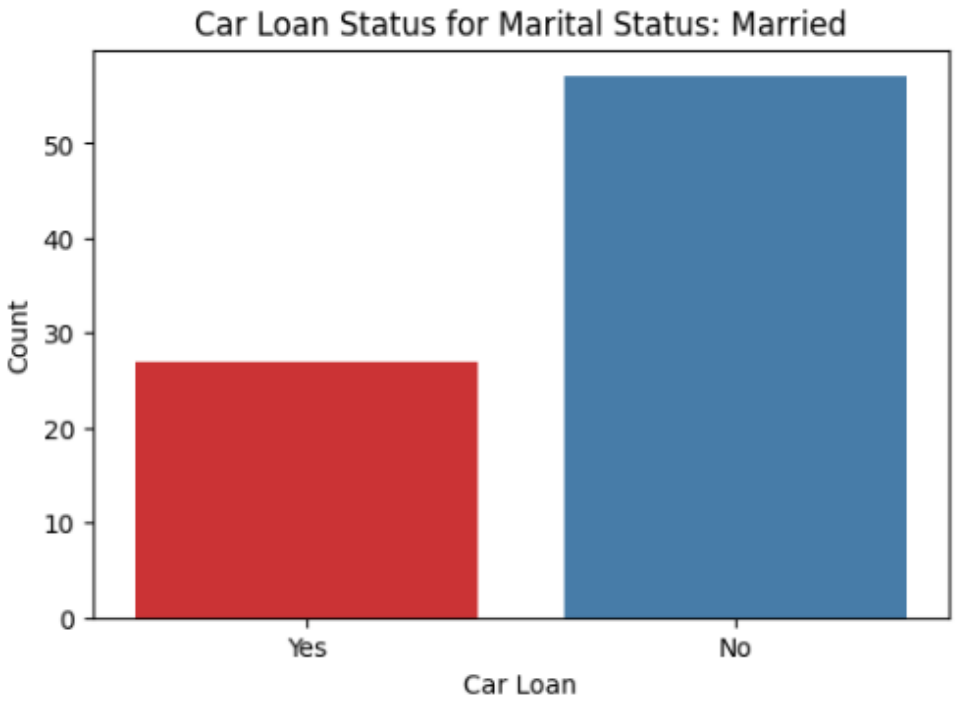
Convex Clustering:



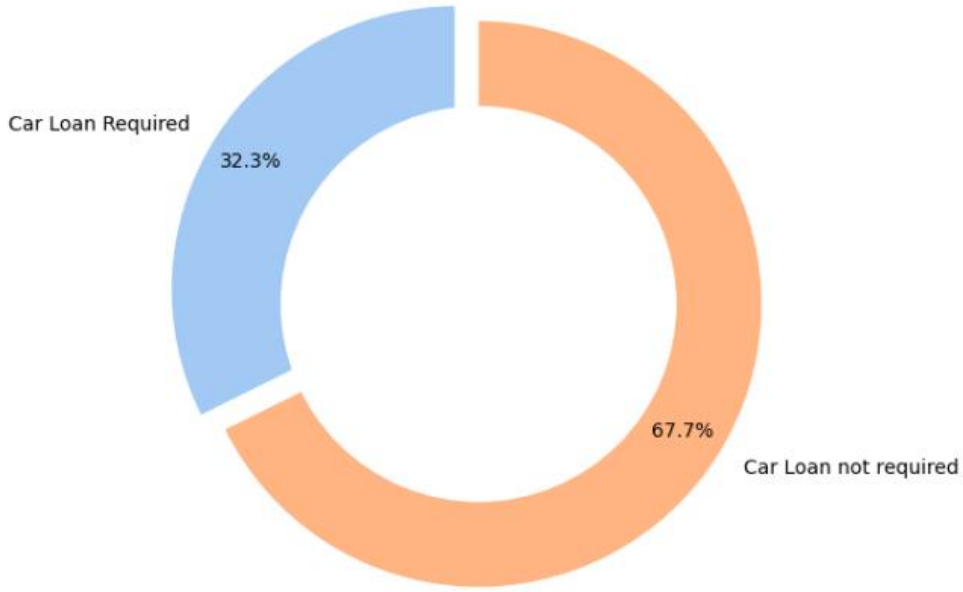
2ND DATA SET

Plotting the count plot on different martial status.

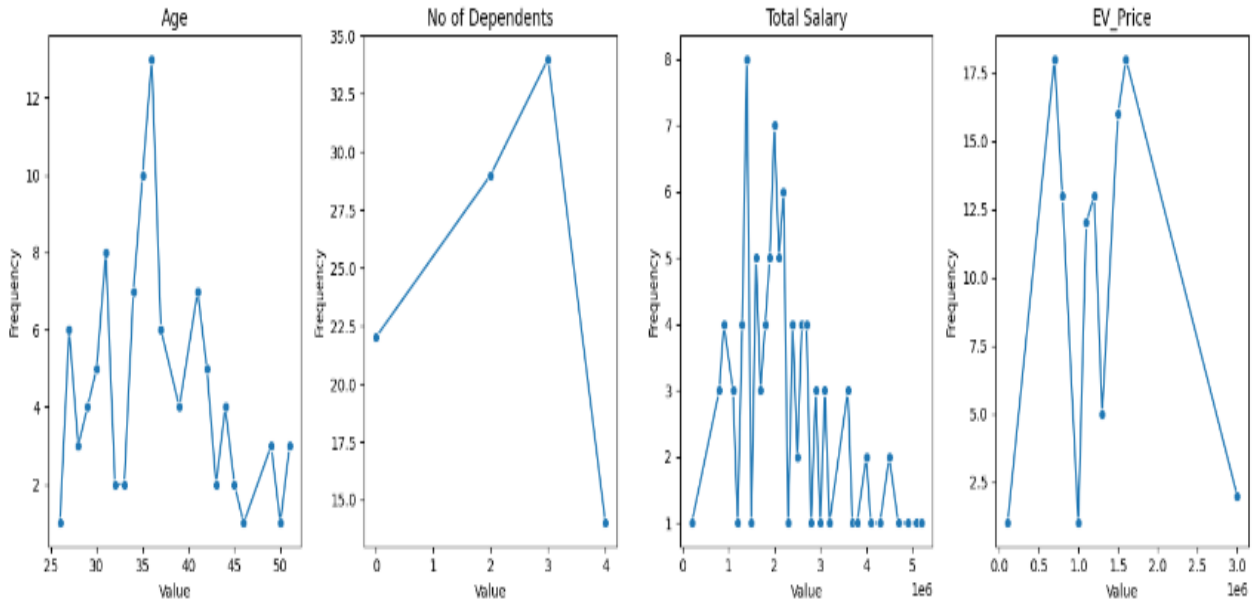




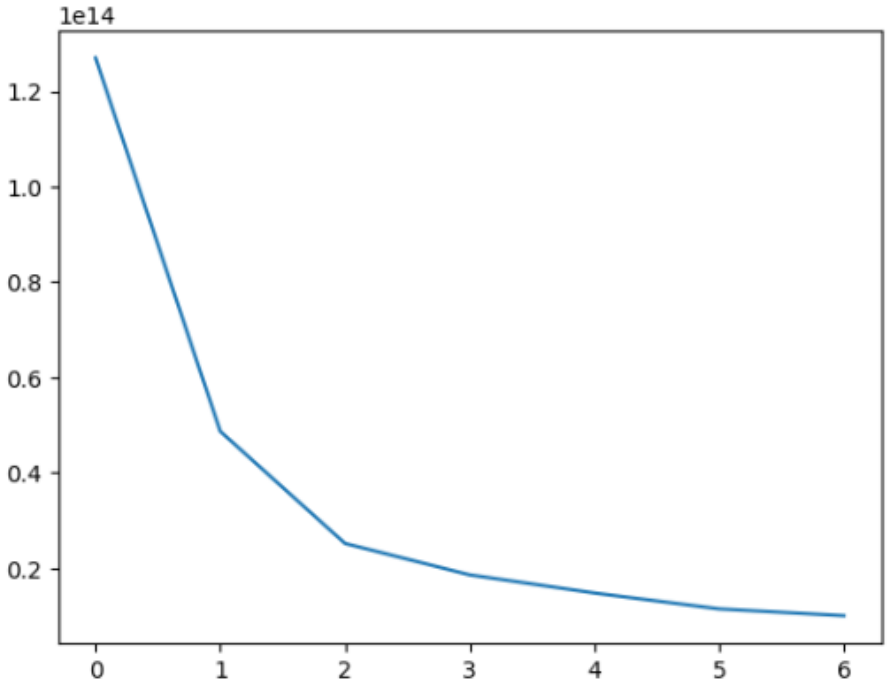
Plotting the Doughnut chart car loans:

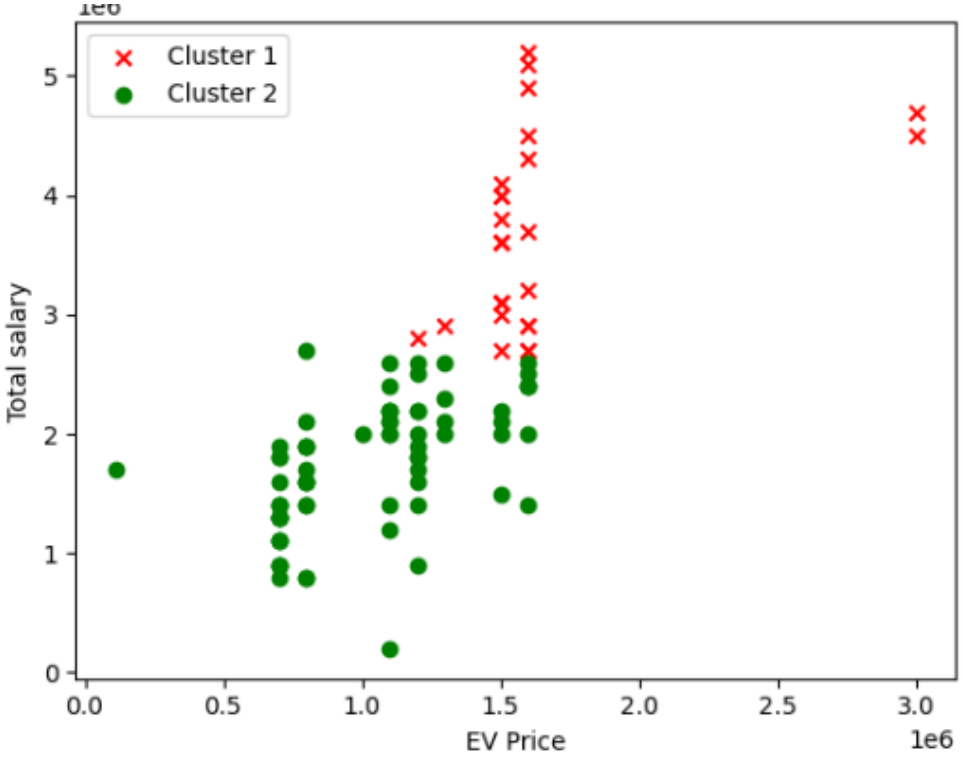


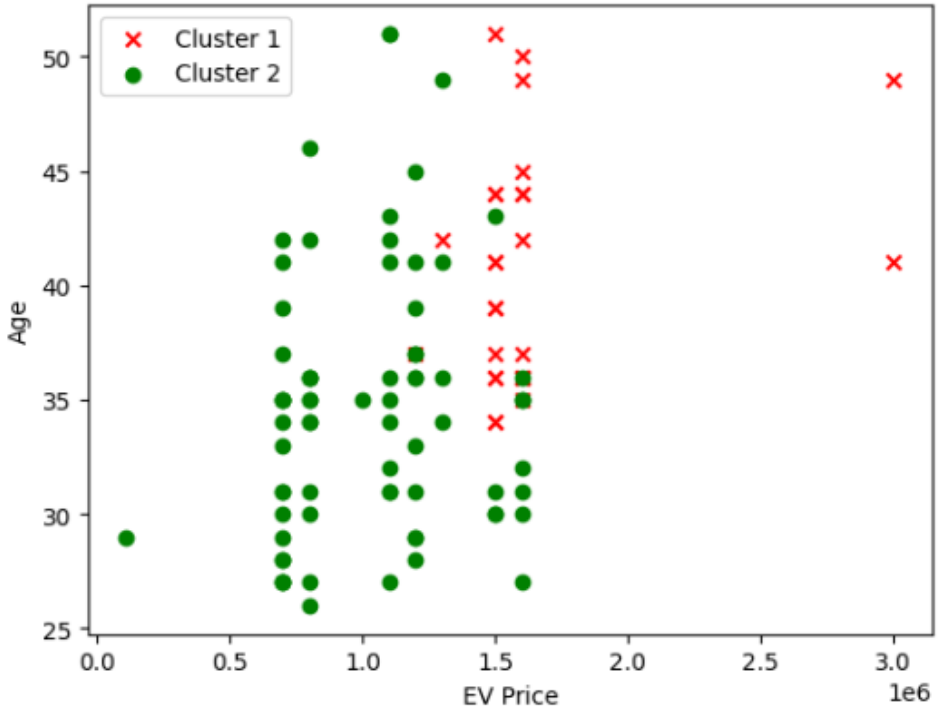
Plotting the frequency of the different customers:

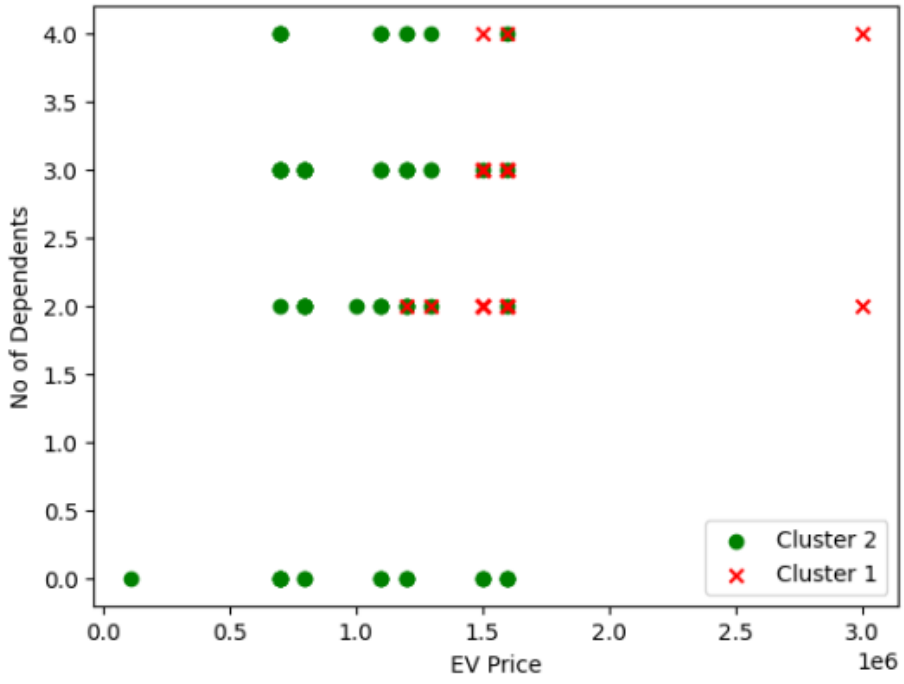


Optimal number of Clusters Using Elbow method:






Cluster 1
Cluster 2
Age
EV Price
1e6

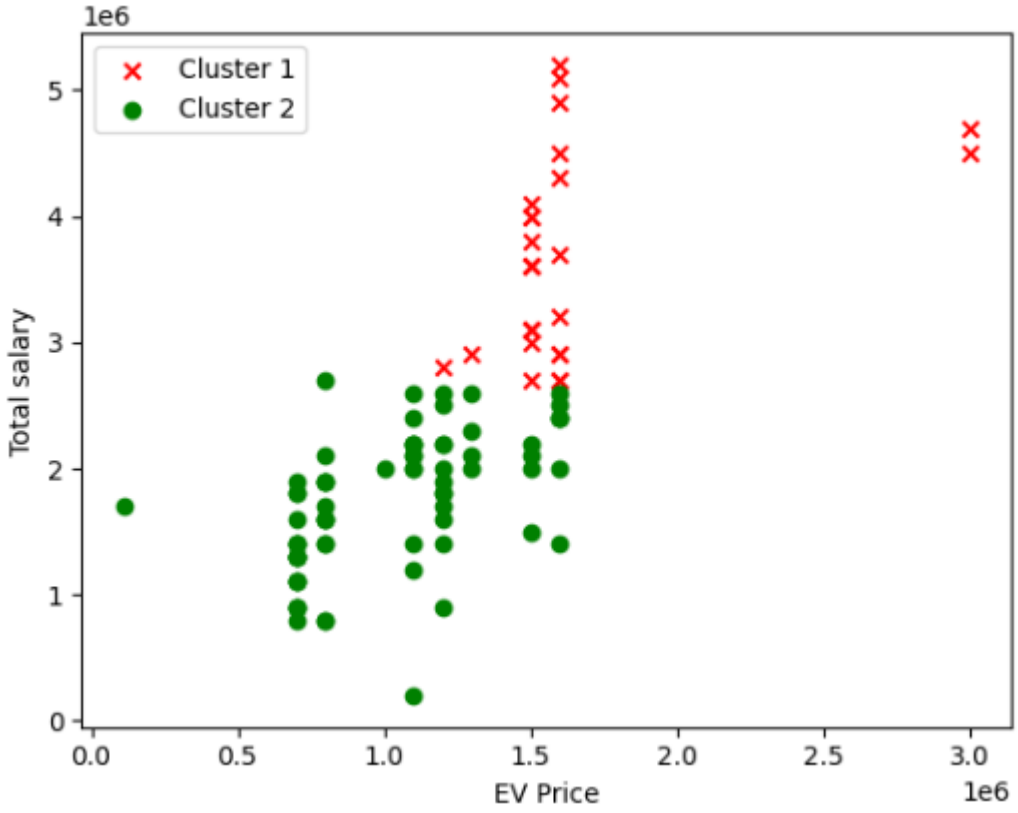

No of Dependents
EV Price
1e6
Cluster 2
Cluster 1

SEGMENTATION APPROACH

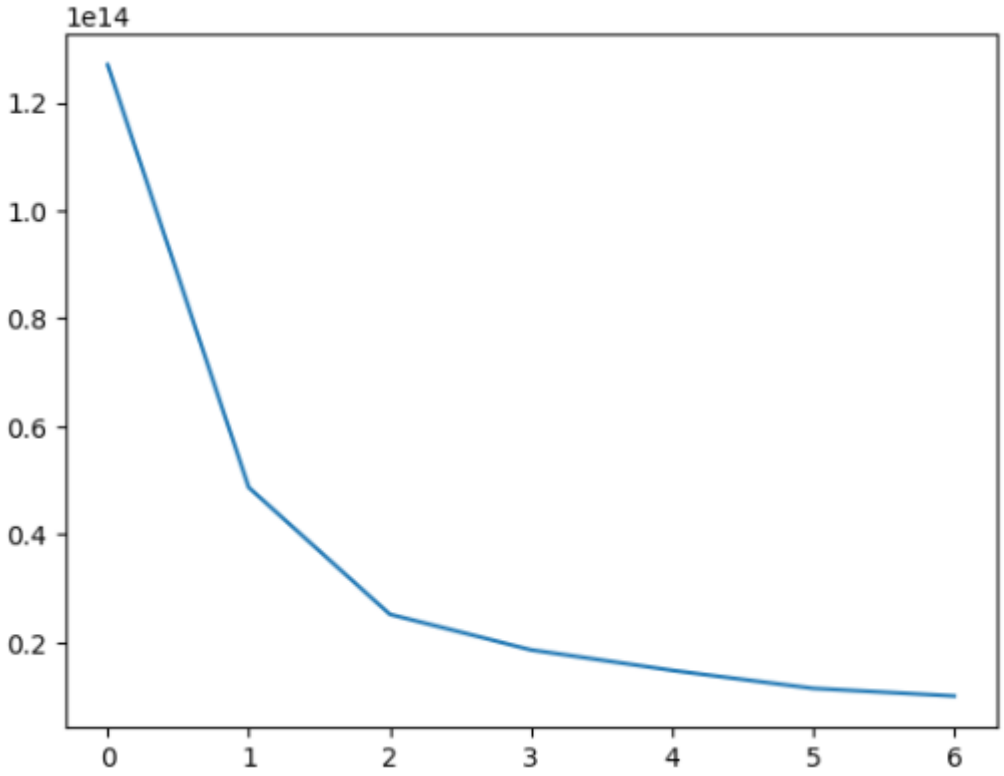
Clustering : Clustering is a technique used in unsupervised machine learning to group similar data points together. In Python, you can perform clustering using libraries such as Scikit-learn or Keras (with TensorFlow backend).

K-means clustering : K-means clustering is a popular unsupervised machine learning algorithm used for clustering data into distinct groups.

Here is the result:



Elbow Method : The Elbow Method is a technique used to determine the optimal number of clusters (K) in K-means clustering. It involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow" point where the rate of decrease in WCSS slows down significantly. This point indicates the optimal number of clusters.



Kaleido: Is a Python library that enables the generation of static image files (such as PNG, JPEG, SVG) directly from Plotly figures without relying on an external rendering engine. This library is particularly useful for creating high-quality static images of interactive Plotly visualizations.

If I had given additional time, I would like to search columns such as range and charging time. I feel this will keep evolving in terms of providing the comfort to the customers. The other column will be the price.

I will try kaleido. Which I just imported the library in this project but couldn't work on. I was getting the errors and I was unable to make on time. The other model that I would like to work on is Average Silhouette method. This measures how similar a cluster is to its cluster set compared to other cluster.