

# Report on Market Segmentation 2<sup>nd</sup> Project

## REQUIREMENTS:

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

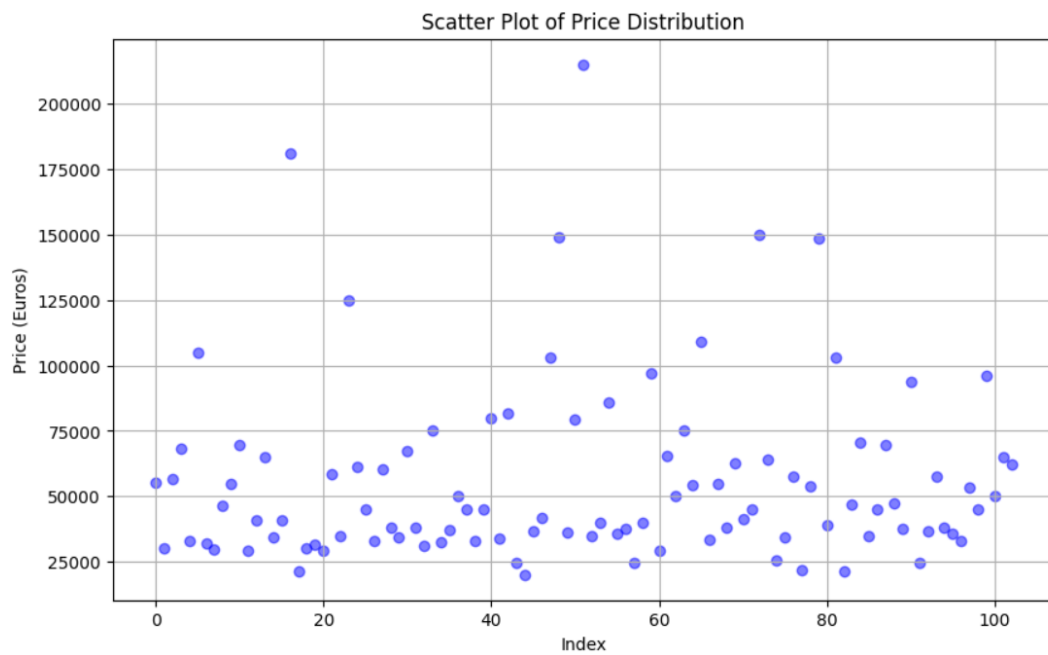
Exploratory Data Analysis (**EDA**) is a crucial step in data analysis that involves examining and understanding the structure, patterns, and characteristics of a dataset.

Exploratory Data Analysis (**EDA**) is a crucial step in data analysis that involves examining and understanding the structure, patterns, and characteristics of a dataset. A general outline of how we may approach EDA:

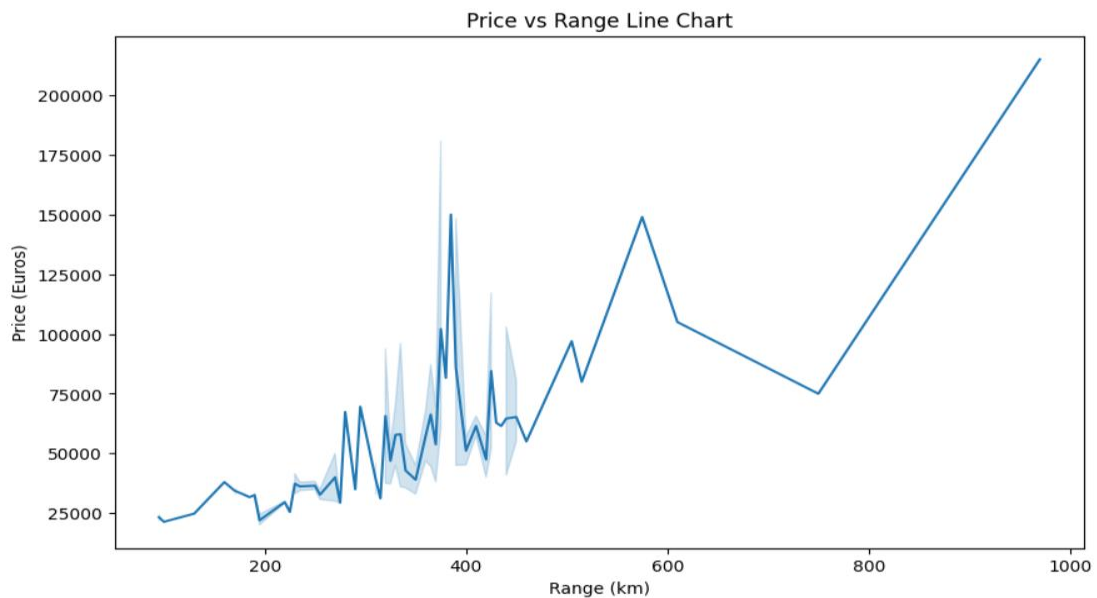
1. **Data Collection:** Gather the dataset you want to analyze. This could be from various sources such as databases, CSV files, APIs, etc.
2. **Initial Inspection:**
  - Check the first few rows of the dataset to understand its structure.
  - Look for missing values, outliers, and inconsistencies in 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:

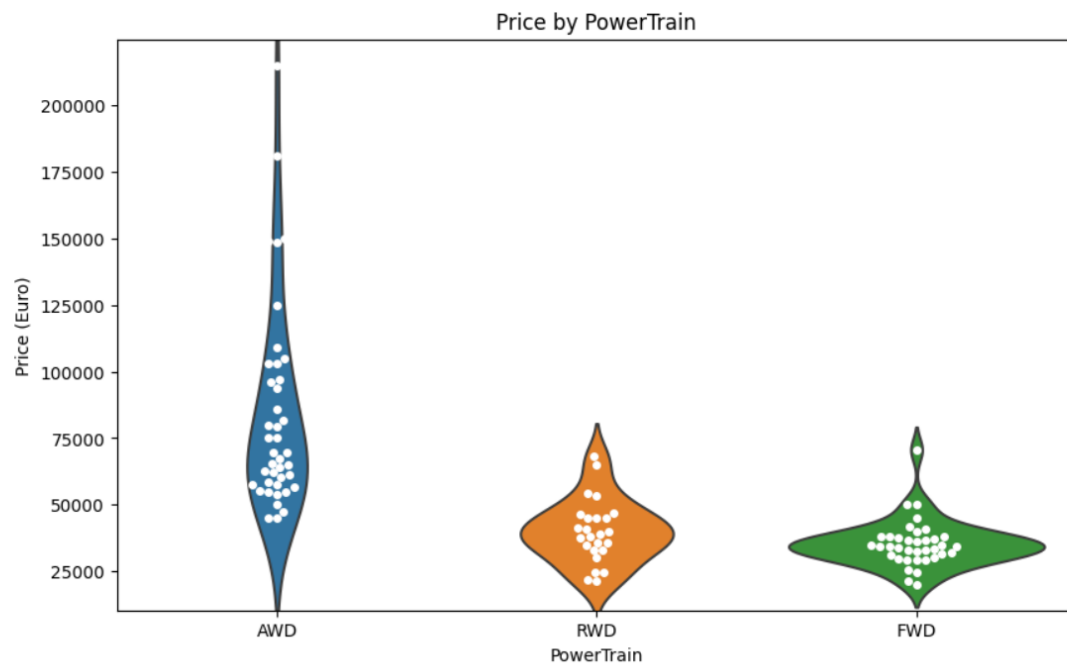


### Line Chart:

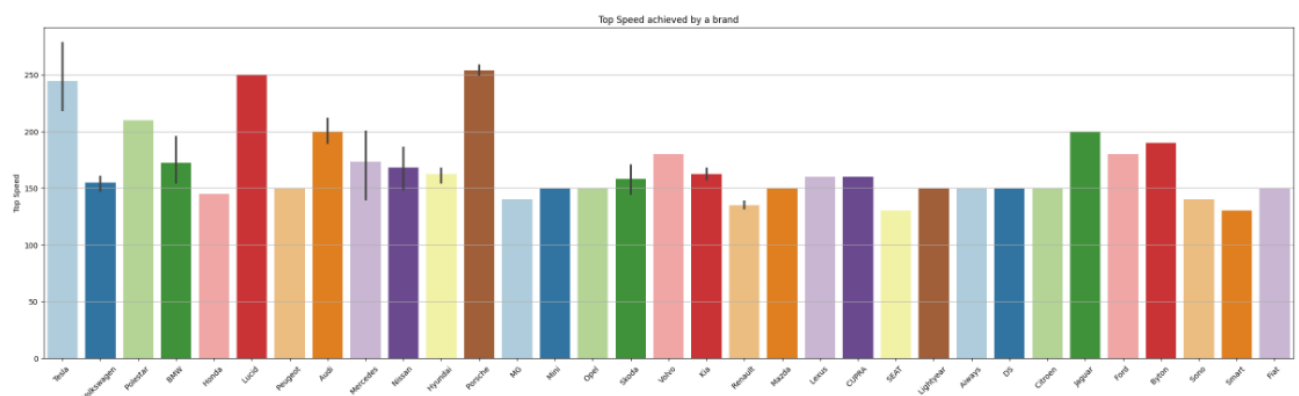


- In this visualize between range (kilometres) and price.

## Violin chart:



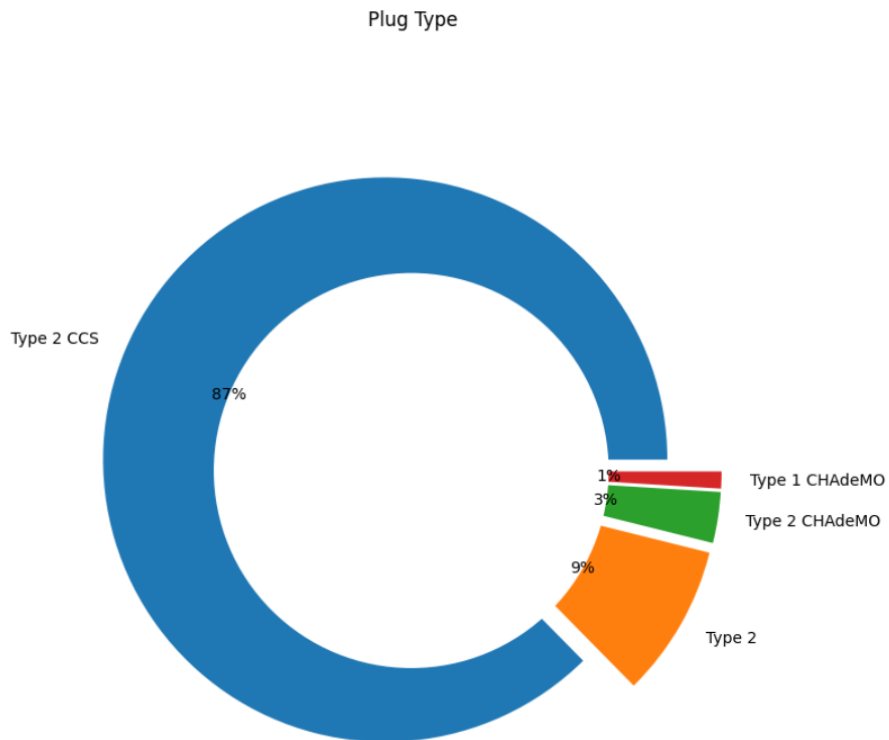
## Bar graph:



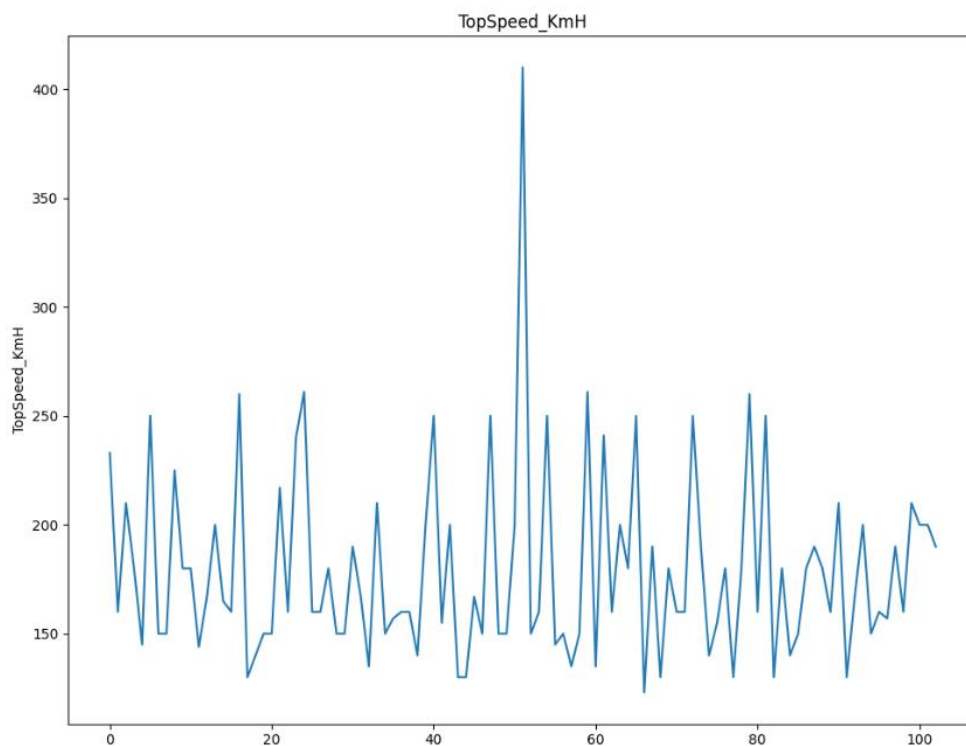
- In this we can visualize the top speed of each vehicle brand.

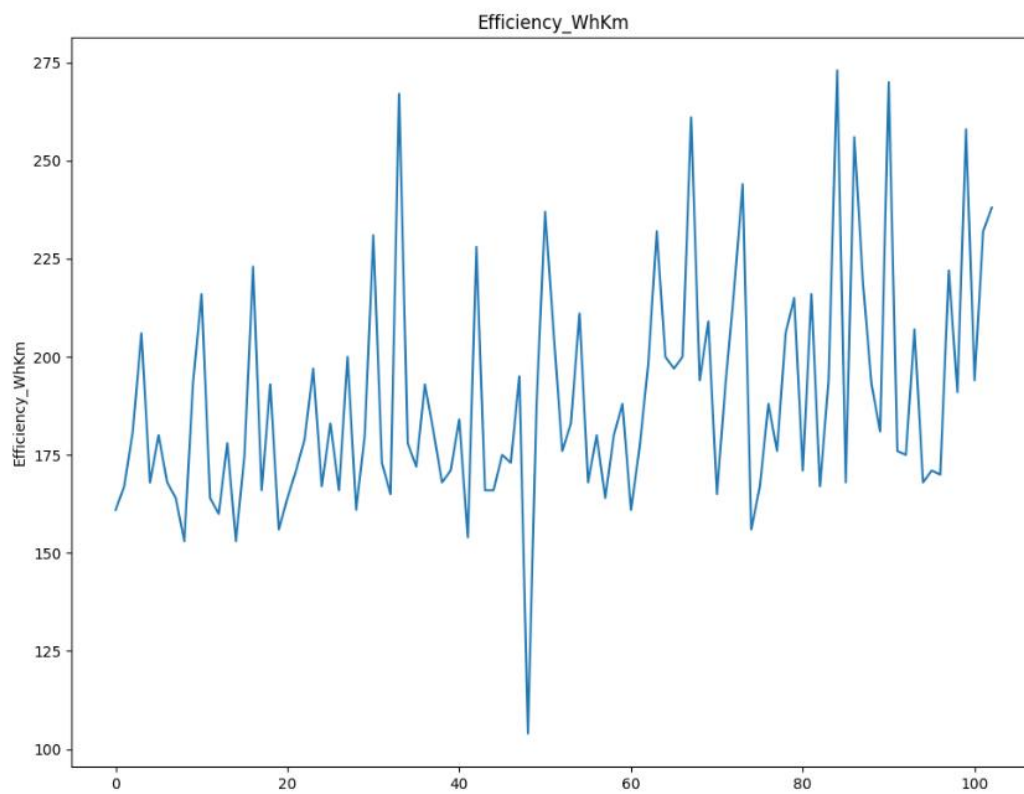
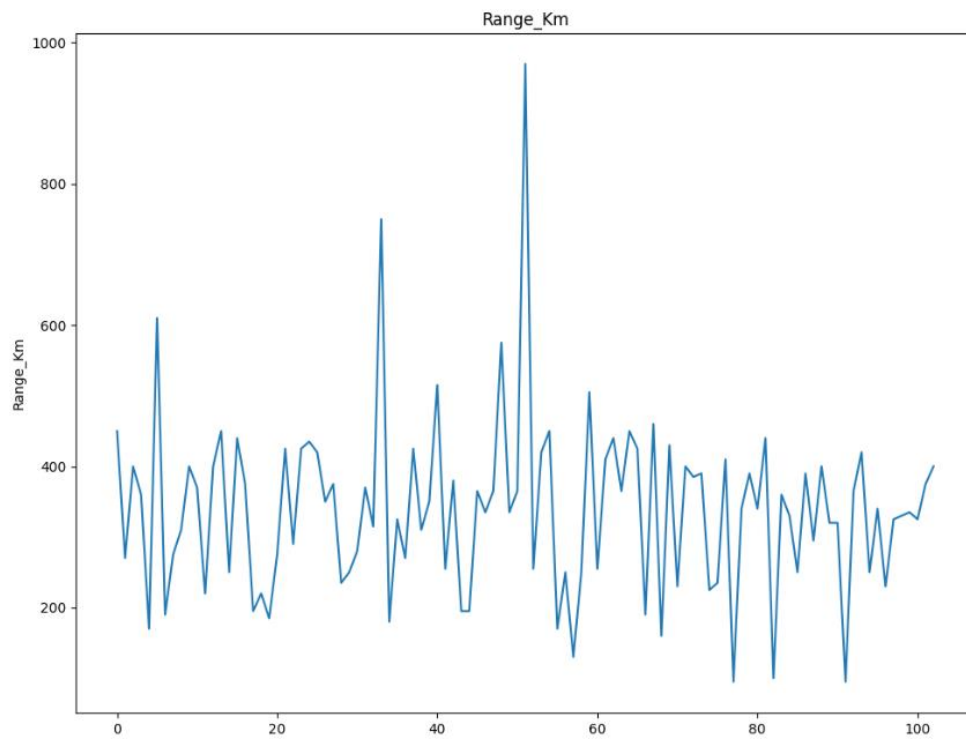
## Doughnut chart:

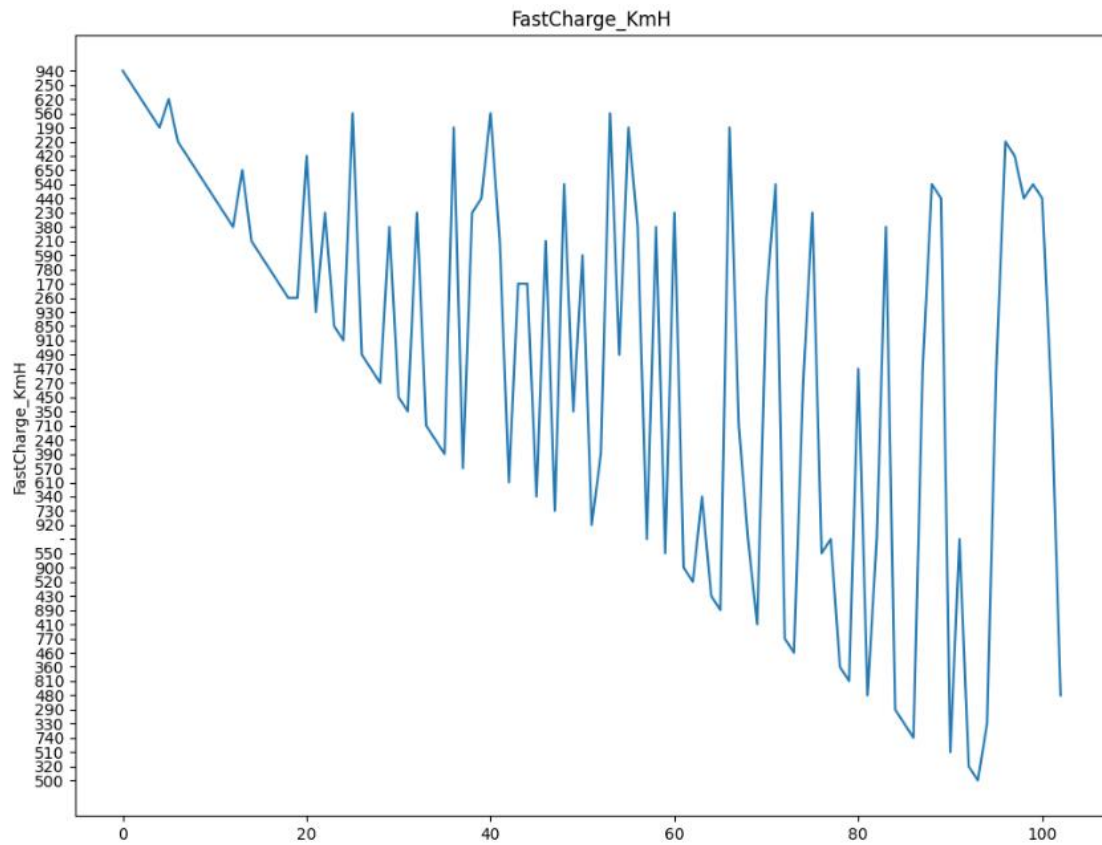
- In this we are looking at different types of plug time and their percentage of use.



## Line chart:

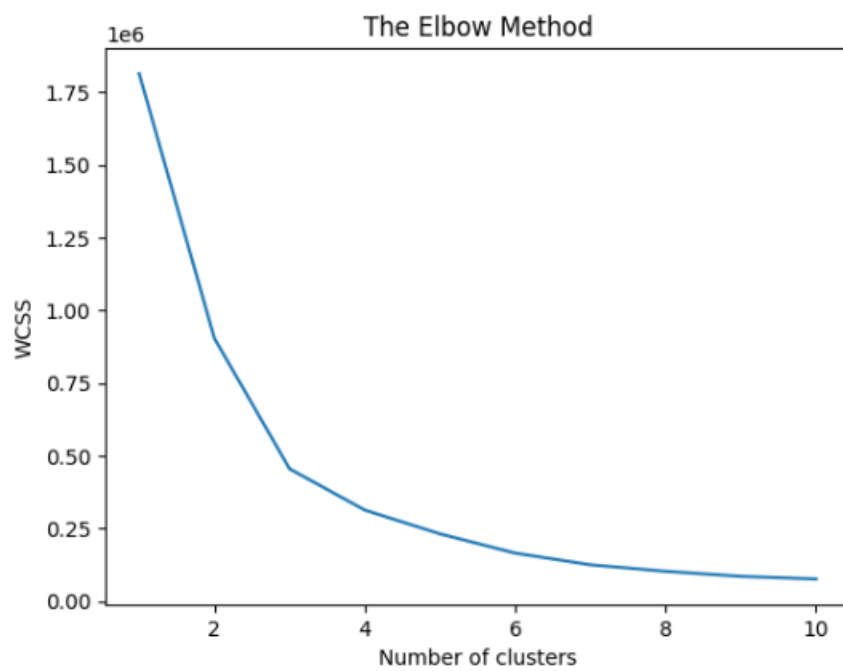






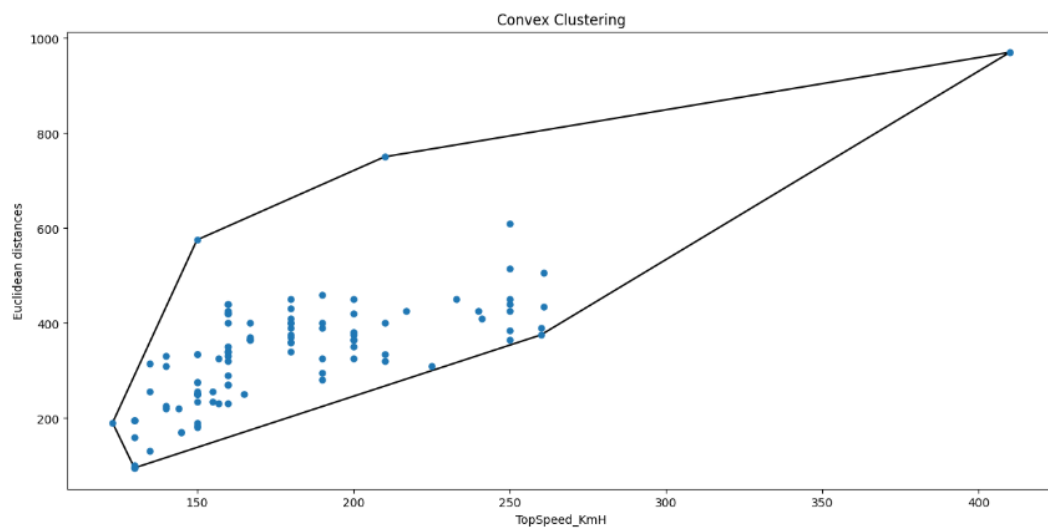
## K-MEANS CLUSTERING

### Elbow Method:



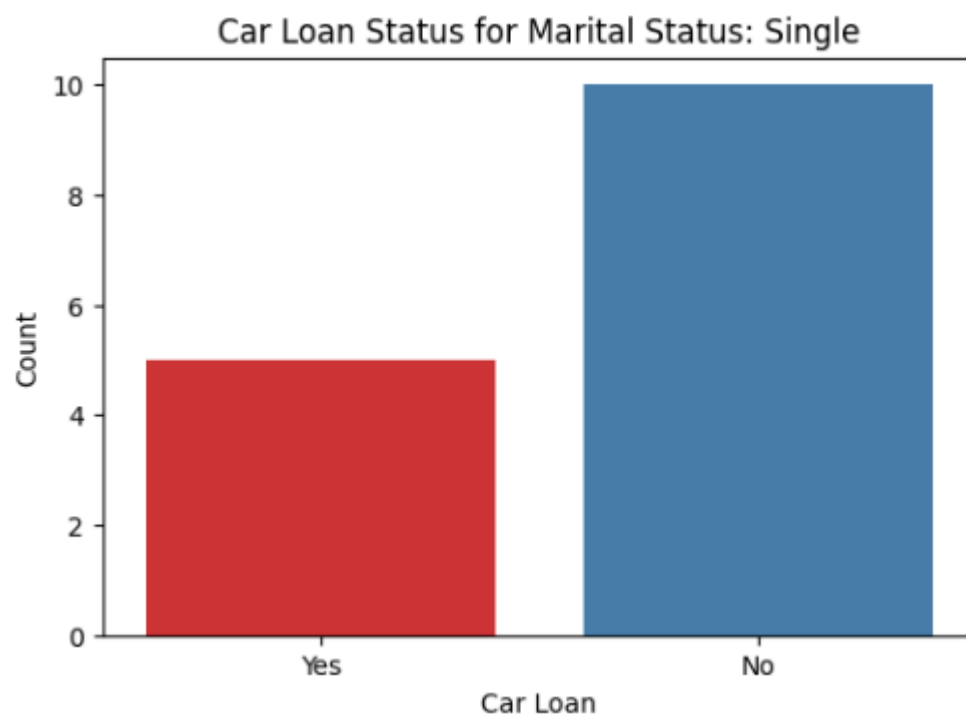


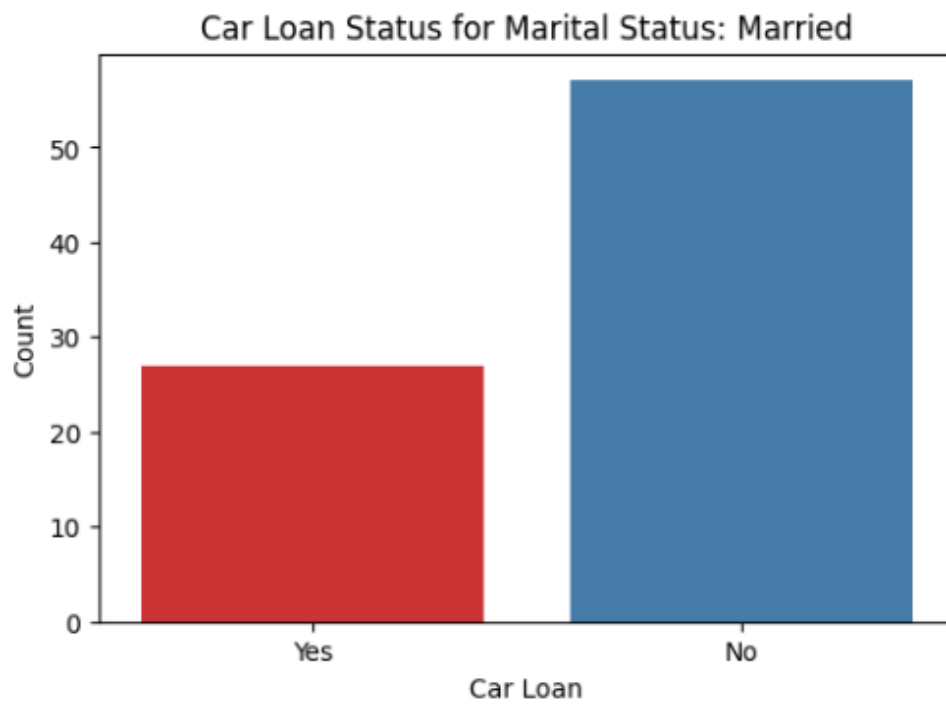
### Convex Clustering:



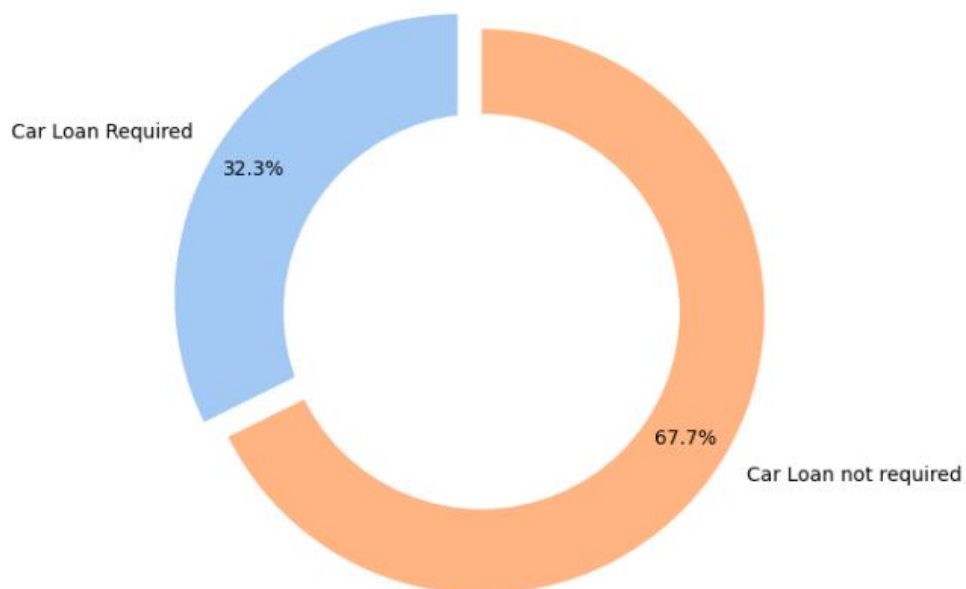
### Data Set-2:

- Plotting the count plot on different marital status.

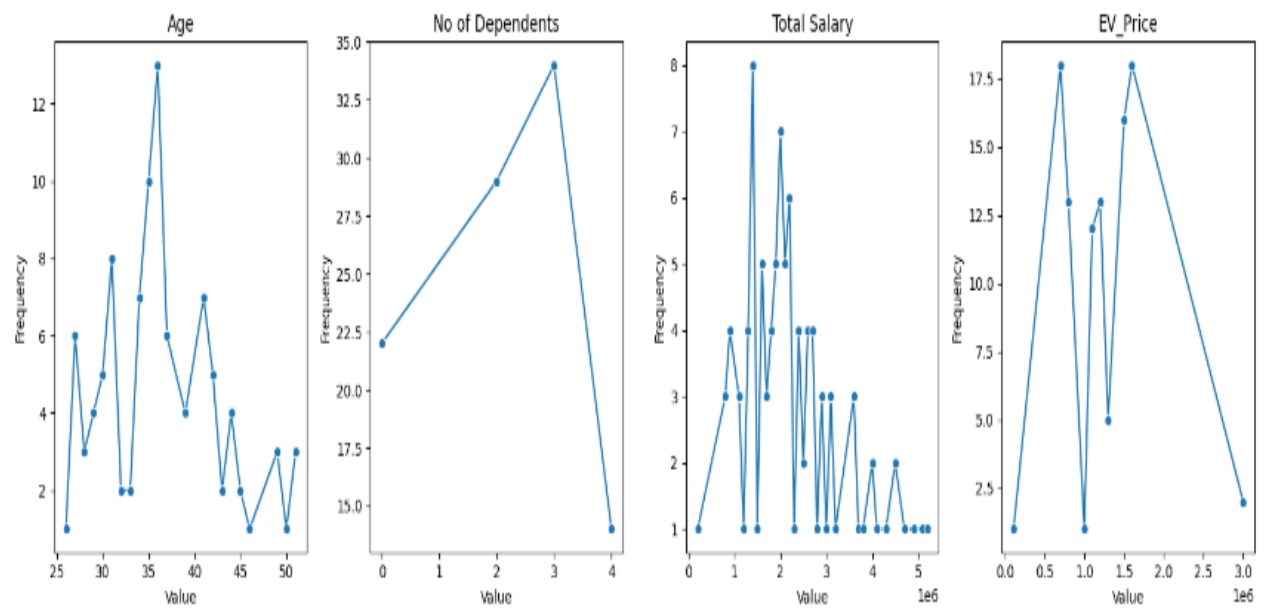




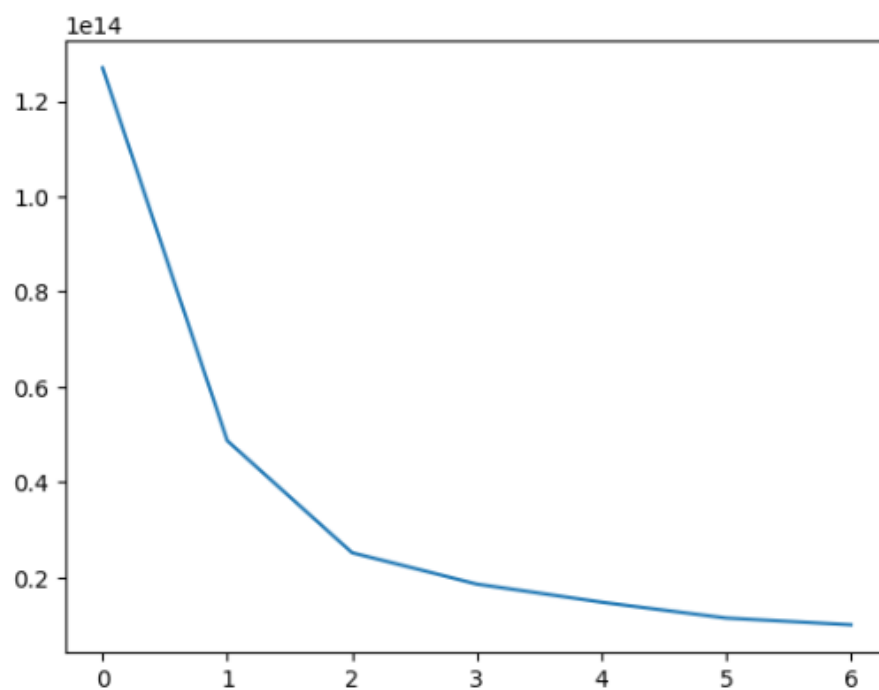
Plotting the Doughnut chart car loans:

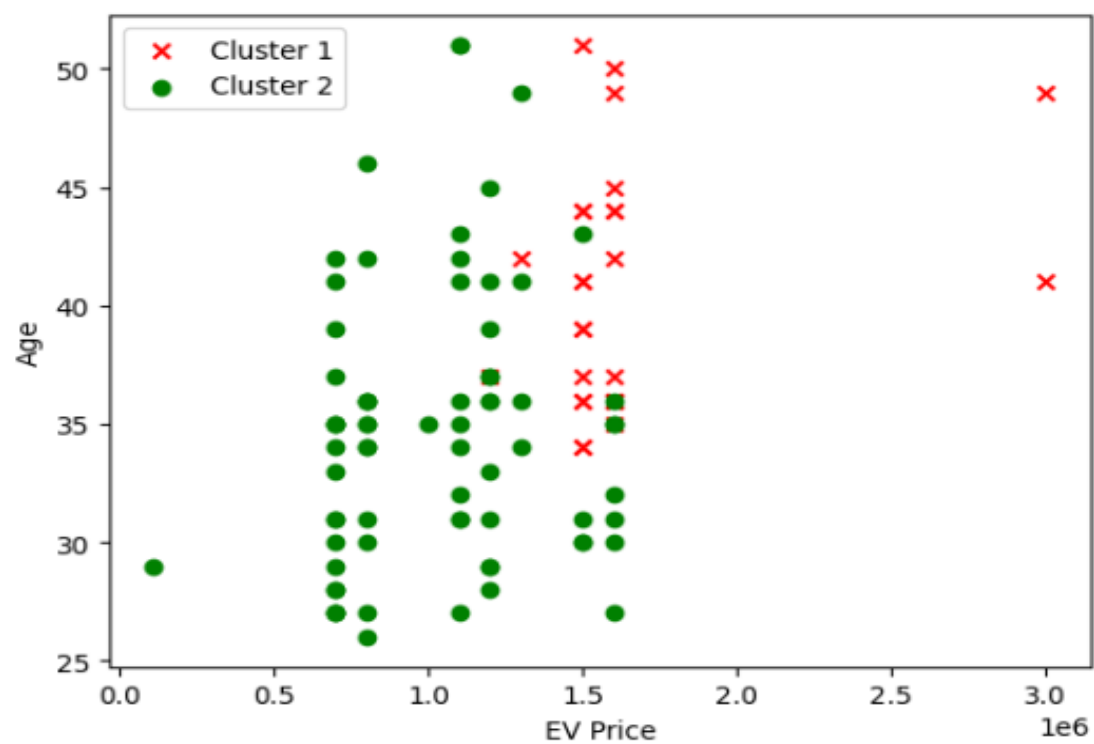
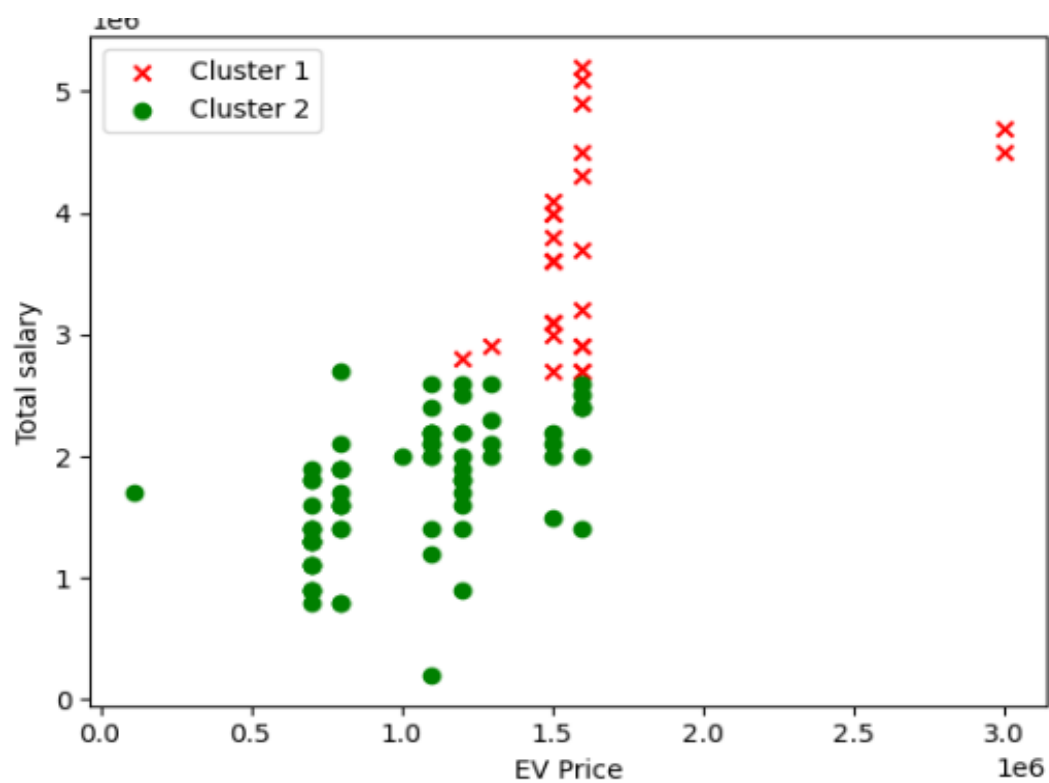


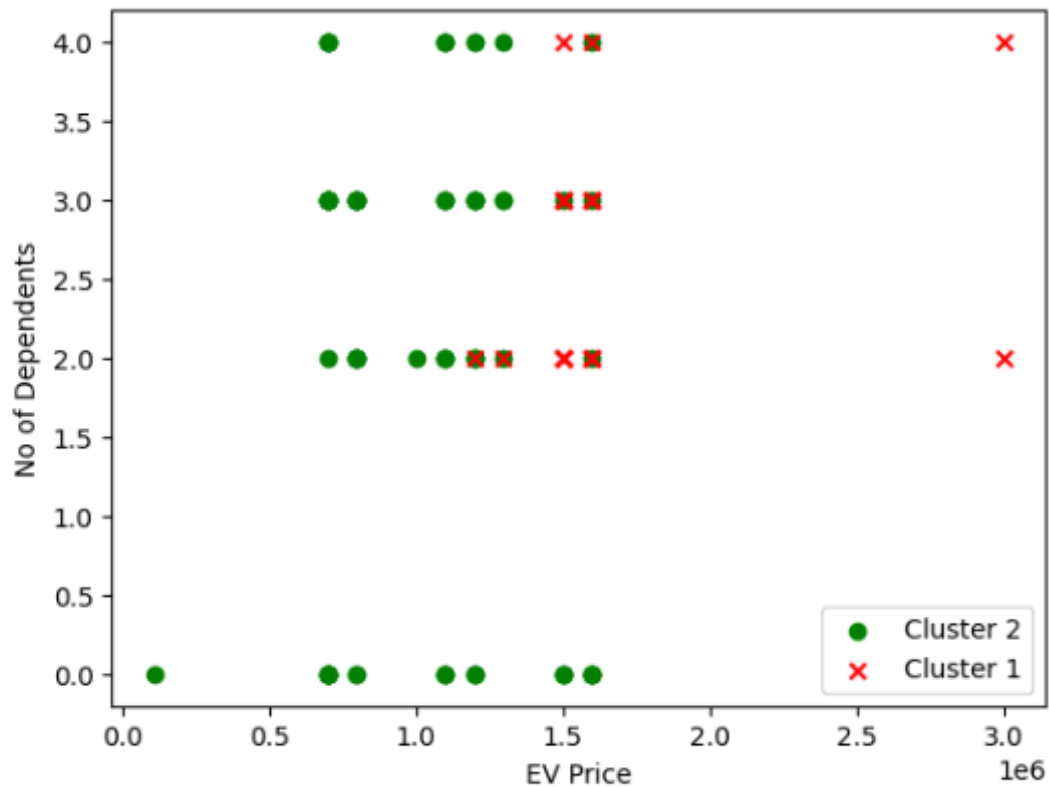
## Plotting the frequency of the different customers:



## Optimal number of Clusters Using Elbow method:







### Conclusion & Insights gained from the Analysis:

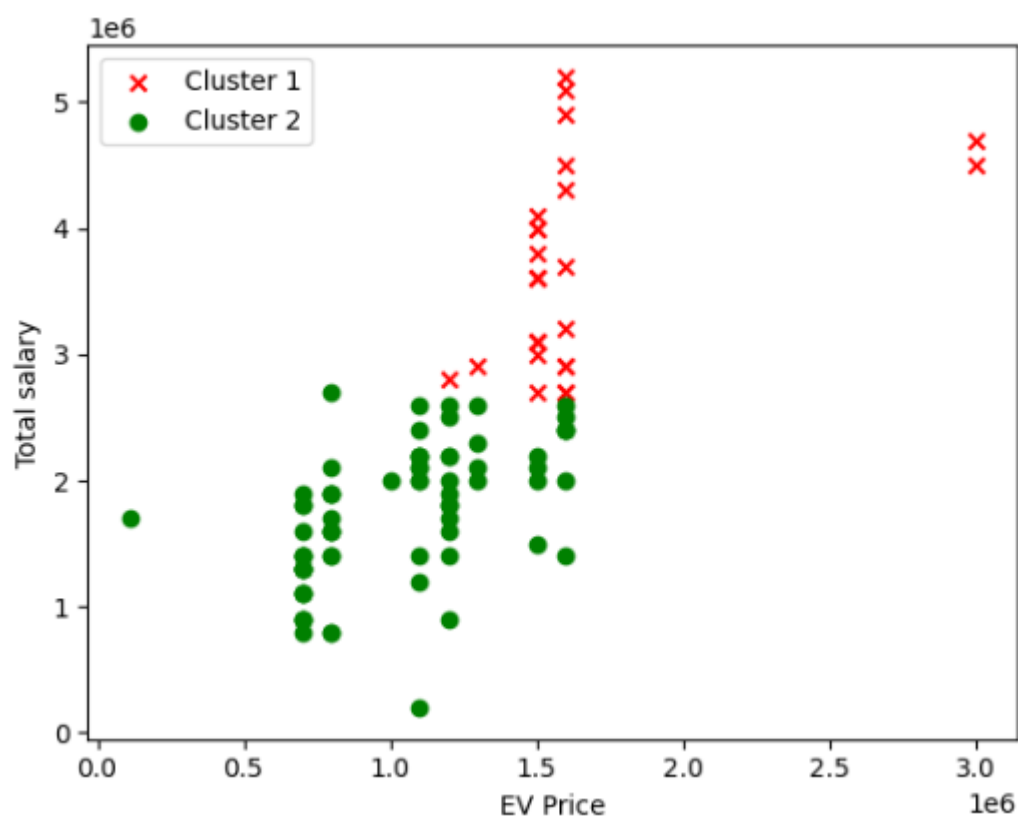
1. **Data Exploration:** EDA uncovered key patterns in features like range, price, and speed, guiding the selection of meaningful variables for segmentation and improving overall data comprehension through visualization.
2. **Clustering and Segmentation:** Applying K-means clustering effectively segmented the data, with the Elbow Method identifying optimal clusters, resulting in clear, distinct groupings that can inform targeted strategies.
3. **Future Enhancements:** Additional analysis on columns such as range, charging time, and price, along with tools like Kaleido and the Average Silhouette method, could further refine segmentation quality and visualization.

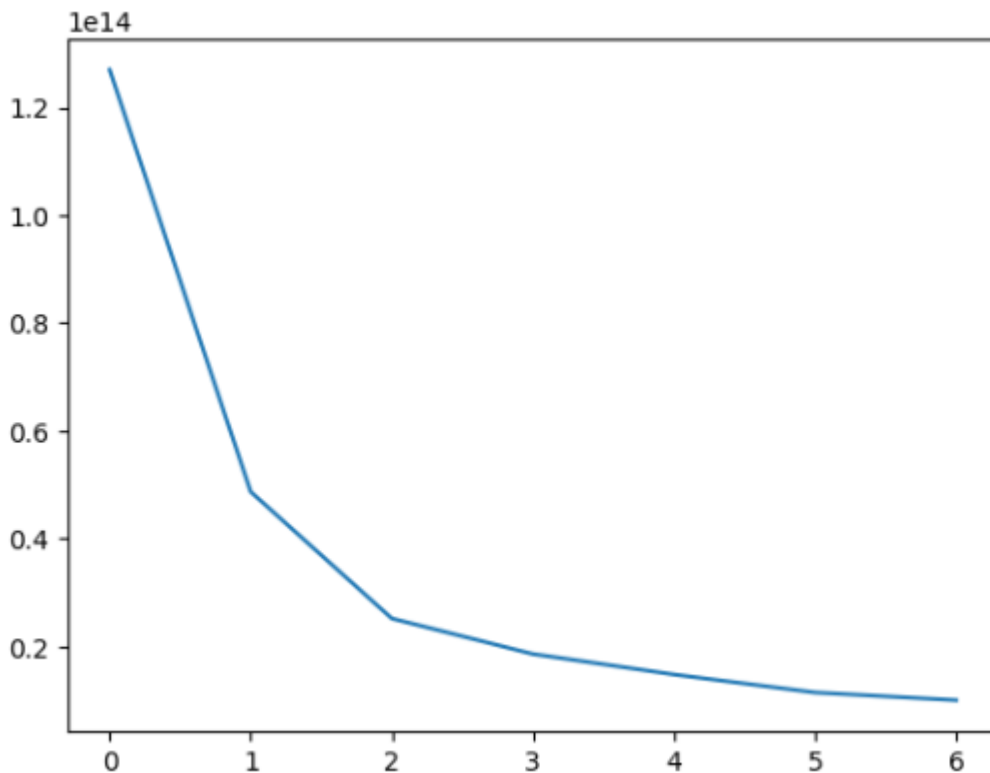
## SEGMENTATION APPROACH

**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:





**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.

### **Estimated Market Size of the EV market Market Domain in Numbers**

The global electric vehicle (EV) market, valued at approximately \$250 billion in 2023, is experiencing rapid expansion. Driven by factors such as government incentives, technological advancements, and the increasing availability of charging infrastructure, the EV market is expected to reach around \$1.3 trillion by 2030, growing at a compound annual growth rate (CAGR) of 22-24%.

## **Top 4 Variables/features which can be used to create most optimal Market Segments for your Market Domain:**

1. **Vehicle Range (Battery Capacity):** Range influences consumer choice significantly, as it determines the vehicle's usability for different driving needs (e.g., urban vs. long-distance travel).
2. **Charging Time:** This feature addresses convenience, as shorter charging times are often preferred, especially for consumers in regions with limited access to fast-charging infrastructure.
3. **Price Range:** Price is crucial for segmenting by affordability, with categories like budget, mid-range, and premium, catering to different income groups and market demands.
4. **Vehicle Type** (e.g., Sedan, SUV, Commercial): Vehicle type aligns with customer lifestyle and usage patterns, allowing segmentation based on use cases (e.g., family vehicles, luxury cars, commercial fleets).

These features together support creating precise segments based on performance, affordability, convenience, and utility preferences.

Kesarla ManiKanta Venkatesh Github link: [Click Here](#)