REPORT ON

ELECTRIC VEHICLE MARKET SEGMENTATION ANALYSIS

RAKHI SAU

02-02-2024

Abstract

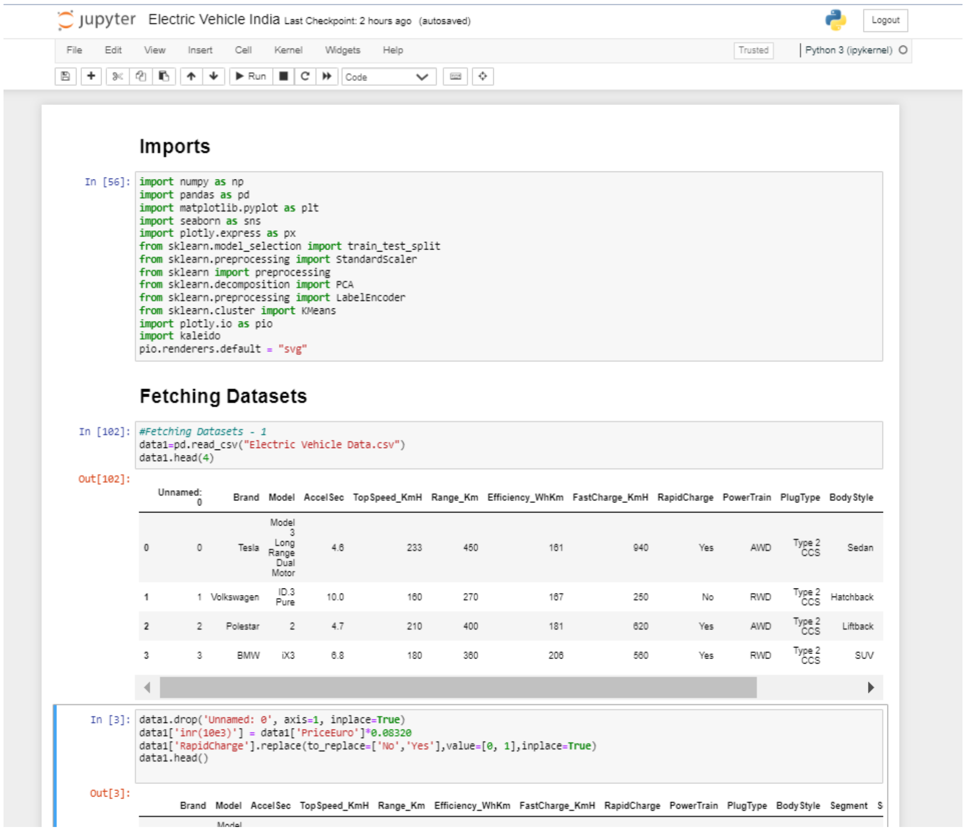
The electric vehicle (EV) market has experienced remarkable growth in recent years, driven by increasing environmental concerns, technological advancements, and government initiatives promoting sustainable transportation. This paper provides an overview of the current state and future prospects of the electric vehicle market. It examines key drivers such as government subsidies, regulations favoring EV adoption, and advancements in battery technology that have propelled the market forward. Additionally, the challenges facing the EV market, including concerns about range anxiety, charging infrastructure, and battery costs, are discussed. Furthermore, the paper explores emerging trends such as the integration of renewable energy sources into EV charging infrastructure, the development of autonomous electric vehicles, and the rise of shared mobility services. Finally, the potential impact of electric vehicles on reducing greenhouse gas emissions, improving air quality, and reshaping the automotive industry is highlighted, emphasizing the need for continued innovation and collaboration across various stakeholders to accelerate the transition to a sustainable transportation ecosystem.

1. Explain how and which ML model (algorithm) helped you in 2nd Project?

In electric vehicle (EV) market segmentation, machine learning (ML) models can be instrumental in identifying distinct groups of customers or market segments based on various characteristics or behaviors related to EV adoption. Here's how ML algorithms can be applied in this context:

1. **Data Collection and Preprocessing:**

- Initially, relevant data pertaining to potential customers or market segments need to be collected. This could include demographic information, geographic location, purchasing behavior, preferences, socio-economic factors, etc. Once collected, the data needs to be preprocessed to handle missing values, outliers, and to normalize or scale features as necessary.



Imports

In [56]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn import preprocessing
from sklearn.decomposition import PCA
from sklearn.preprocessing import LabelEncoder
from sklearn.cluster import KMeans
import plotly.io as pio
import kaleido
pio.renderers.default = "svg"
```

Fetching Datasets

In [102]:
```
#Fetching Datasets - 1
data1=pd.read_csv("Electric Vehicle Data.csv")
data1.head(4)
```

Out[102]:

	Unnamed: 0	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle
0	0	Tesla	Model 3 Long Range Dual Motor	4.6	233	450	161	940	Yes	AWD	Type 2 CCS	Sedan
1	1	Volkswagen	ID.3 Pure	10.0	160	270	167	250	No	RWD	Type 2 CCS	Hatchback
2	2	Polestar	2	4.7	210	400	181	620	Yes	AWD	Type 2 CCS	Liftback
3	3	BMW	iX3	6.8	180	360	206	560	Yes	RWD	Type 2 CCS	SUV

In [3]:
```
data1.drop('Unnamed: 0', axis=1, inplace=True)
data1['inr(10e3)'] = data1['PriceEuro']*0.08320
data1['RapidCharge'].replace(to_replace=['No','Yes'],value=[0, 1],inplace=True)
data1.head()
```

Out[3]:

	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment	S
		Model											

2. **Feature Selection or Extraction:**

- ML models for segmentation require input features that effectively capture the differences between market segments. Feature selection techniques, such as correlation analysis or feature importance ranking from tree-based models, can be employed to identify the most relevant features. Additionally, feature extraction methods like Principal Component Analysis (PCA) can help reduce dimensionality while retaining important information.

3. **Model Selection:**

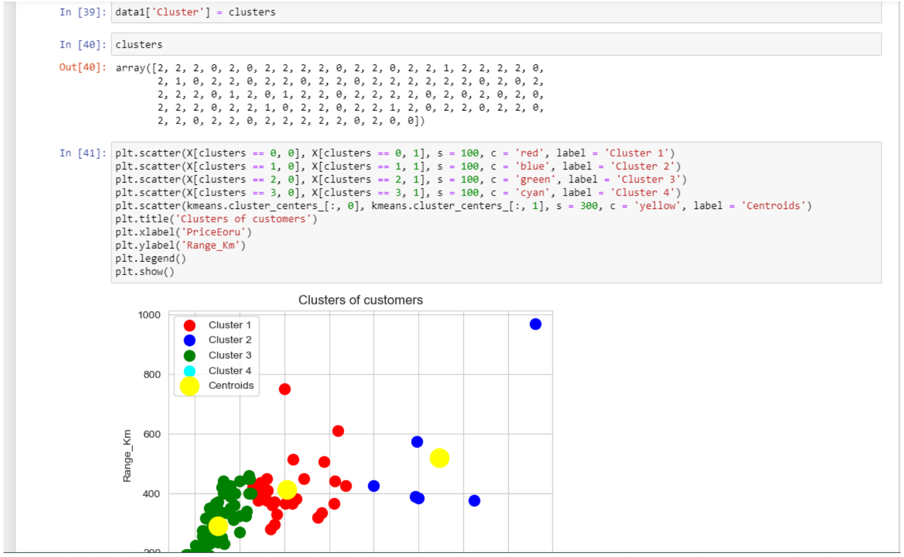
- Various ML algorithms can be utilized for market segmentation, depending on the nature of the data and the complexity of the segmentation task. Commonly used algorithms include:

  - K-means clustering: This algorithm partitions the data into K clusters based on similarities in feature space.

  - Hierarchical clustering: This method creates a tree of clusters, where clusters at lower levels are merged based on their similarity.

  - Gaussian Mixture Models (GMM): GMM assumes that the data is generated from a mixture of several Gaussian distributions and assigns probabilities to data points belonging to each cluster.

  - Self-Organizing Maps (SOM): SOM is an unsupervised learning technique that maps high-dimensional data onto a low-dimensional grid, preserving the topological relationships between data points.



```
In [39]: data1['Cluster'] = clusters

In [40]: clusters

Out[40]: array([2, 2, 2, 0, 2, 0, 2, 2, 2, 2, 0, 2, 2, 0, 2, 2, 1, 2, 2, 2, 2, 0,
                2, 1, 0, 2, 2, 0, 2, 2, 0, 2, 2, 0, 2, 2, 2, 2, 2, 2, 0, 2, 0, 2,
                2, 2, 2, 0, 1, 2, 0, 1, 2, 2, 0, 2, 2, 2, 2, 0, 2, 0, 2, 0, 2, 0,
                2, 2, 2, 0, 2, 2, 1, 0, 2, 2, 0, 2, 2, 1, 2, 0, 2, 2, 0, 2, 2, 0,
                2, 2, 0, 2, 2, 0, 2, 2, 2, 2, 2, 0, 2, 0, 0])

In [41]: plt.scatter(X[clusters == 0, 0], X[clusters == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
         plt.scatter(X[clusters == 1, 0], X[clusters == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
         plt.scatter(X[clusters == 2, 0], X[clusters == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
         plt.scatter(X[clusters == 3, 0], X[clusters == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
         plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'yellow', label = 'Centroids')
         plt.title('Clusters of customers')
         plt.xlabel('PriceEoru')
         plt.ylabel('Range_Km')
         plt.legend()
         plt.show()
```

4. **Model Training and Evaluation:**

- The selected ML model is trained on the preprocessed data to learn patterns and segment the market accordingly. Evaluation metrics such as silhouette score, Davies-Bouldin index, or intra-cluster coherence can be used to assess the quality of the segmentation and determine the optimal number of clusters (K).

5.**Interpretation and Actionable Insights:**

- Once the model has segmented the market, it's essential to interpret the results to understand the characteristics and preferences of each segment. This insight can then be used to tailor marketing strategies, product offerings, pricing, and distribution channels to better meet the needs of each segment.

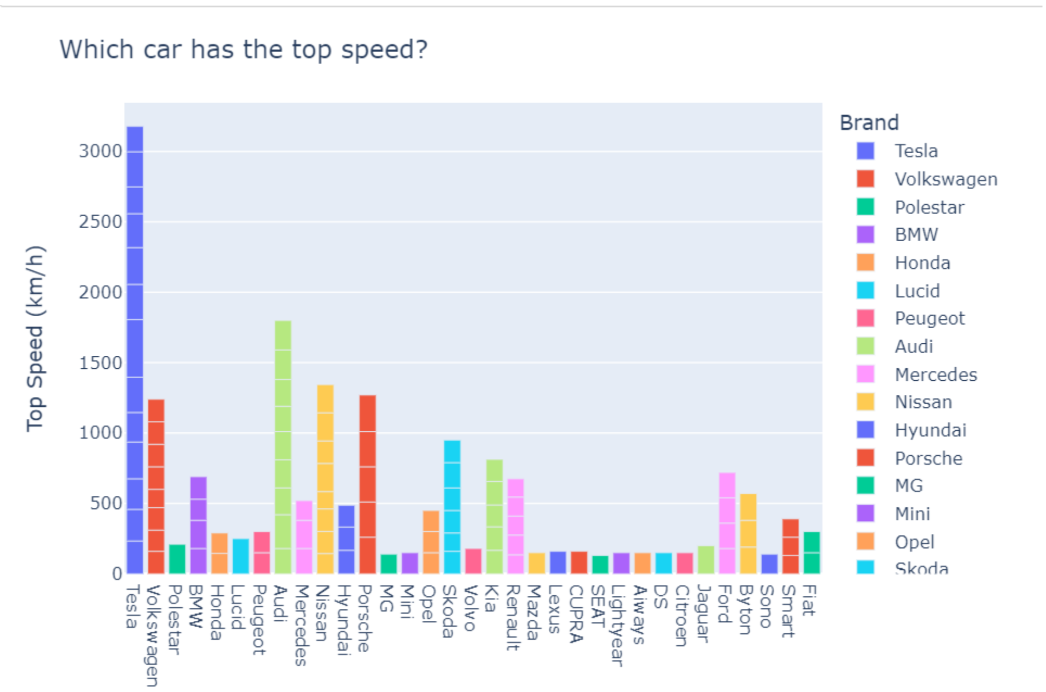
Overall, ML algorithms play a crucial role in automating the segmentation process, uncovering hidden patterns in data, and enabling targeted marketing efforts in the electric vehicle market. The choice of algorithm depends on factors such as the size and nature of the data, the desired level of interpretability, and the specific objectives of the segmentation analysis.

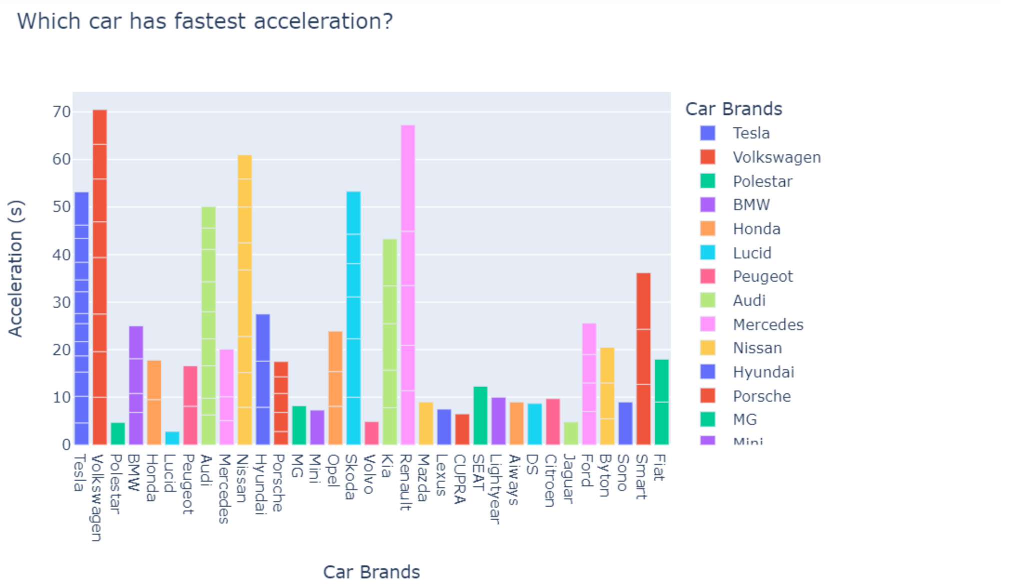
2. Elaborate on the final conclusion & insights gained from the research/analysis work.

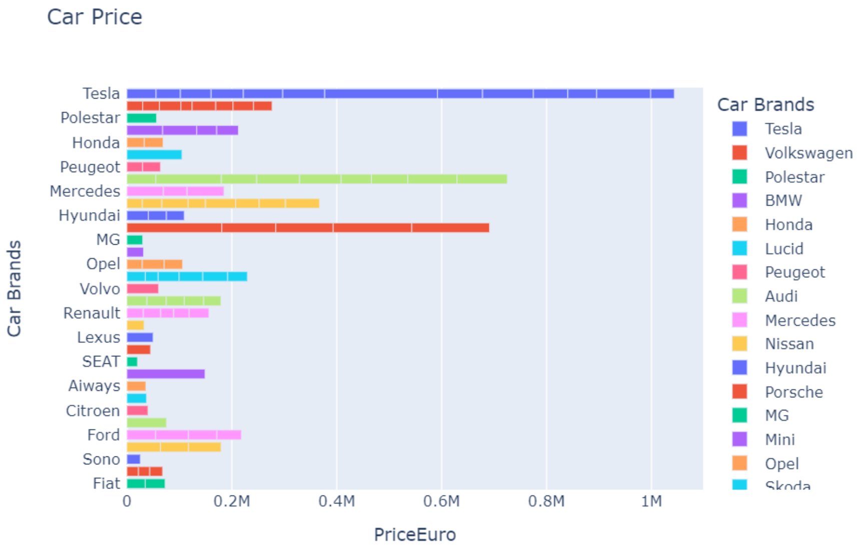
--EDA

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

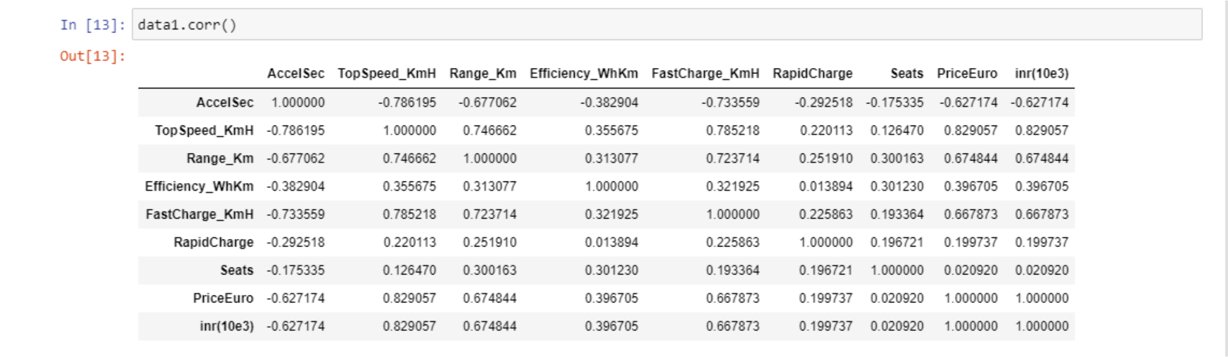
Comparision of cars in our data.







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.



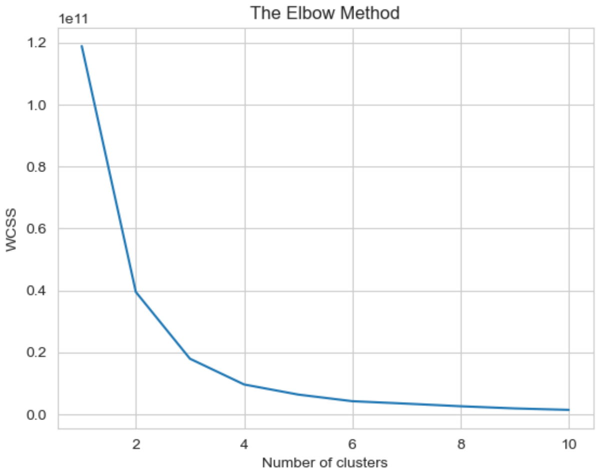
```
In [13]: data1.corr()
```

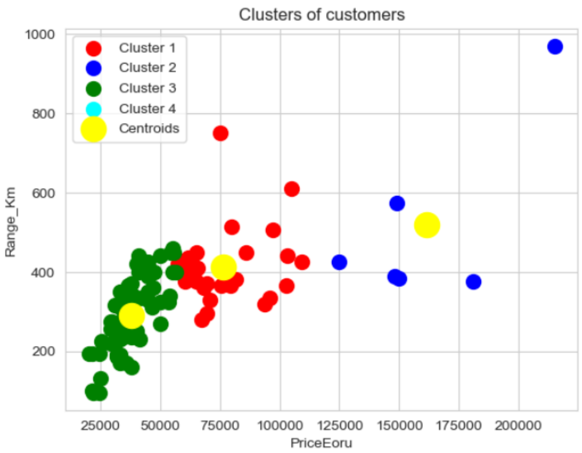
```
Out[13]:
```

	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	Seats	PriceEuro	inr(10e3)
AccelSec	1.000000	-0.786195	-0.677062	-0.382904	-0.733559	-0.292518	-0.175335	-0.627174	-0.627174
TopSpeed_KmH	-0.786195	1.000000	0.746662	0.355675	0.785218	0.220113	0.126470	0.829057	0.829057
Range_Km	-0.677062	0.746662	1.000000	0.313077	0.723714	0.251910	0.300163	0.674844	0.674844
Efficiency_WhKm	-0.382904	0.355675	0.313077	1.000000	0.321925	0.013894	0.301230	0.396705	0.396705
FastCharge_KmH	-0.733559	0.785218	0.723714	0.321925	1.000000	0.225863	0.193364	0.667873	0.667873
RapidCharge	-0.292518	0.220113	0.251910	0.013894	0.225863	1.000000	0.196721	0.199737	0.199737
Seats	-0.175335	0.126470	0.300163	0.301230	0.193364	0.196721	1.000000	0.020920	0.020920
PriceEuro	-0.627174	0.829057	0.674844	0.396705	0.667873	0.199737	0.020920	1.000000	1.000000
inr(10e3)	-0.627174	0.829057	0.674844	0.396705	0.667873	0.199737	0.020920	1.000000	1.000000

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 10. 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.



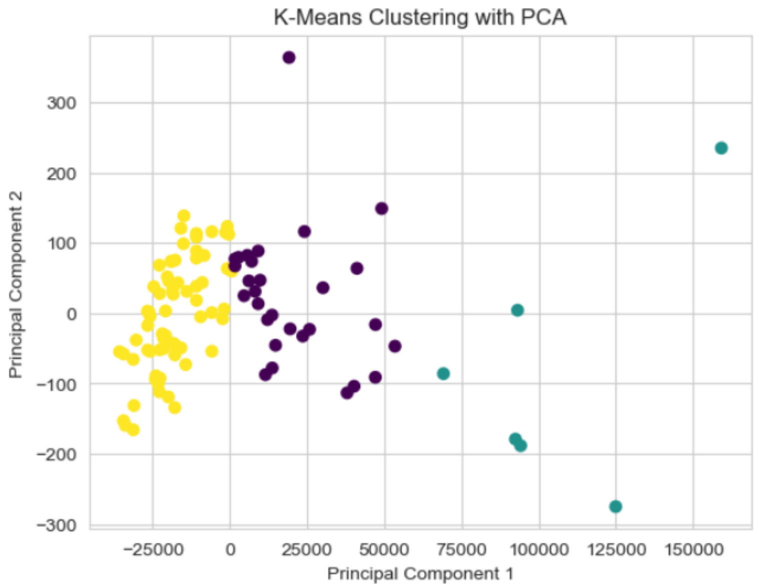


Cluster 1, Cluster 2, Cluster 3, and Cluster 4. Each data point is represented by a colored circle, and the axes are labeled "PriceEur" (presumably price in Euros) and "Range_Km" (range in kilometers). There are also four red "x" markers, which I assume represent the centroids of each cluster.

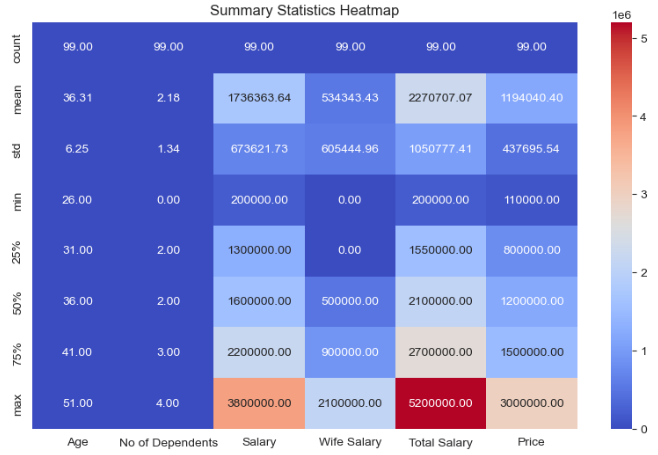
Cluster distribution: The clusters are spread out somewhat evenly across the chart, with Cluster 1 in the upper left, Cluster 2 in the lower left, Cluster 3 in the upper right, and Cluster 4 in the lower right.

Price and range relationship: There seems to be a general trend of higher prices associated with longer ranges. For example, Cluster 3 and 4, which are located on the right side of the graph with higher ranges, also tend to have higher prices. However, there are also exceptions, such as some data points in Cluster 1 that have longer ranges than some data points in Cluster 2, even though Cluster 2 is generally priced higher.

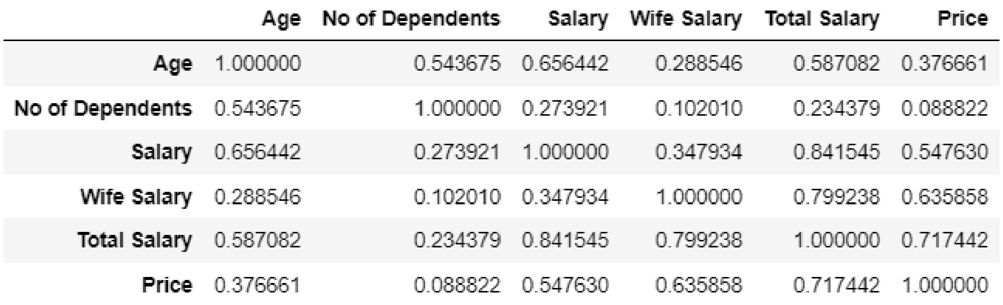
Cluster centroids: The centroids (red "x" markers) are located in the center of their respective clusters, which suggests that they are representative of the data points within each cluster.



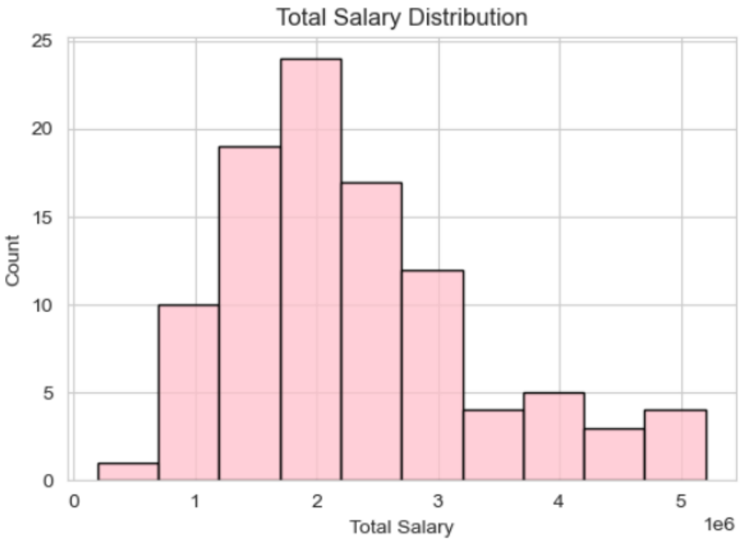
2^nd Dataset:

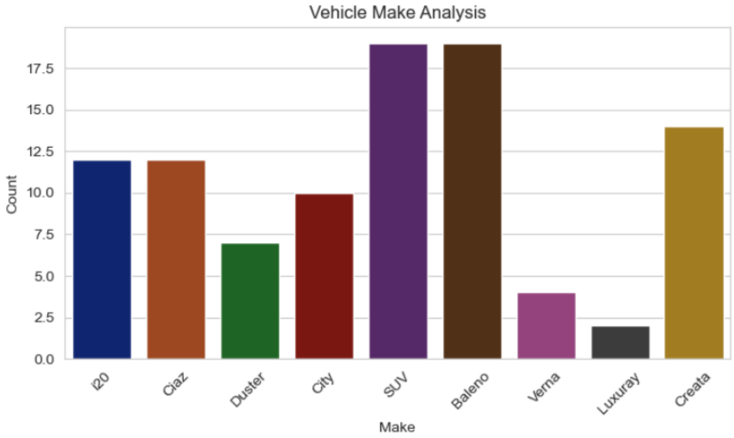


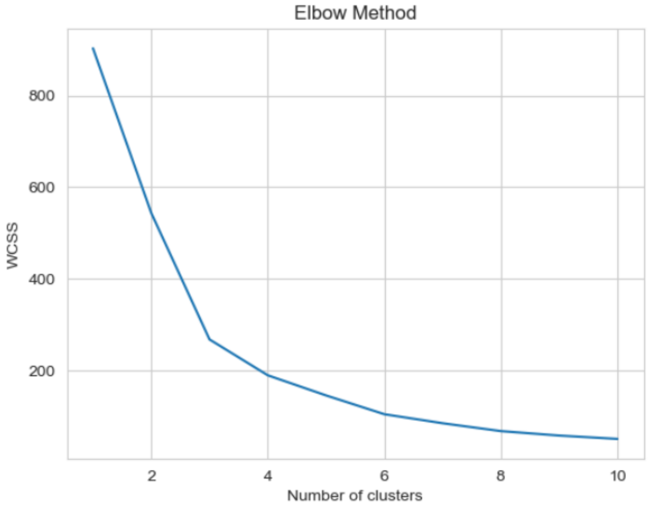
	Age	No of Dependents	Salary	Wife Salary	Total Salary	Price
count	99.00	99.00	99.00	99.00	99.00	99.00
mean	36.31	2.18	1736363.64	534343.43	2270707.07	1194040.40
std	6.25	1.34	673621.73	605444.96	1050777.41	437695.54
min	26.00	0.00	200000.00	0.00	200000.00	110000.00
25%	31.00	2.00	1300000.00	0.00	1550000.00	800000.00
50%	36.00	2.00	1600000.00	500000.00	2100000.00	1200000.00
75%	41.00	3.00	2200000.00	900000.00	2700000.00	1500000.00
max	51.00	4.00	3800000.00	2100000.00	5200000.00	3000000.00

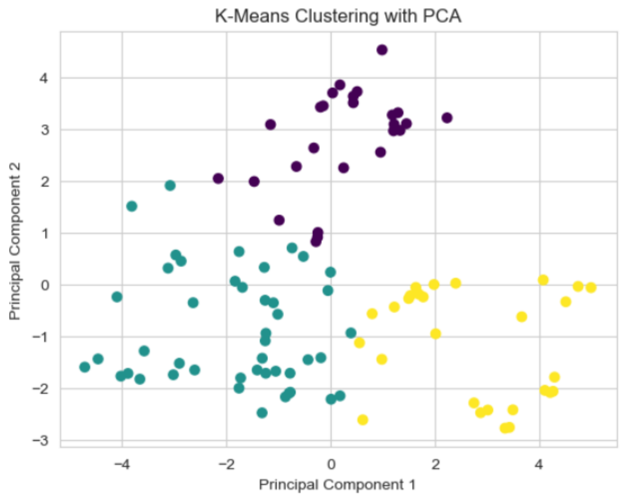


	Age	No of Dependents	Salary	Wife Salary	Total Salary	Price
Age	1.000000	0.543675	0.656442	0.288546	0.587082	0.376661
No of Dependents	0.543675	1.000000	0.273921	0.102010	0.234379	0.088822
Salary	0.656442	0.273921	1.000000	0.347934	0.841545	0.547630
Wife Salary	0.288546	0.102010	0.347934	1.000000	0.799238	0.635858
Total Salary	0.587082	0.234379	0.841545	0.799238	1.000000	0.717442
Price	0.376661	0.088822	0.547630	0.635858	0.717442	1.000000


Total Salary Distribution
25
20
15
10
5
0
Count
0
1
2
3
4
5
Total Salary
1e6


Vehicle Make Analysis
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Count
i20
Ciaz
Duster
City
SUV
Baleno
Verna
Luxuray
Creata
Make


Elbow Method
800
600
400
200
WCSS
2
4
6
8
10
Number of clusters



3. **How will you improve upon the Market Segmentation Project given additional time & some budget to purchase data? (in terms of Datasets collection - name what columns points you will search for & what additional ML models you would like to try)**

--

1. Data Collection:

- Acquire demographic data: Collect data on age, gender, income level, education level, occupation, marital status, and household size. This data will provide deeper insights into consumer behavior and preferences.
- Purchase geospatial data: Obtain data on geographic location, such as ZIP code, city, and state, to analyze regional trends and preferences.
- Obtain psychographic data: Collect information on lifestyle, interests, values, attitudes, and personality traits to understand consumer psychographics.
- Gather purchase history data: Acquire data on past purchases, including product categories, frequency of purchase, and average transaction value.
- Obtain social media data: Collect data from social media platforms to analyze user interactions, sentiment, and preferences.
- Purchase market research reports: Invest in market research reports to gain industry-specific insights and trends.

2. Feature Engineering:

- Conduct feature engineering to create new features from the collected data, such as calculating average income by ZIP code, creating age groups, or deriving consumer segments based on lifestyle indicators.
- Use techniques like one-hot encoding, binning, and feature scaling to preprocess categorical and numerical variables.

3. **Machine Learning Models:**

- Experiment with advanced clustering algorithms: Apply algorithms like hierarchical clustering, DBSCAN, or Gaussian mixture models to identify more complex patterns in the data.
- Try ensemble clustering techniques: Utilize ensemble methods like hierarchical clustering combined with k-means or DBSCAN to improve clustering accuracy and robustness.
- Explore dimensionality reduction techniques: Implement techniques like t-distributed stochastic neighbor embedding (t-SNE) or uniform manifold approximation and projection (UMAP) to visualize high-dimensional data and identify meaningful clusters.
- Evaluate deep learning models: Experiment with deep learning architectures like autoencoders or self-organizing maps (SOMs) to learn hierarchical representations of the data and extract latent features for segmentation.

4. **Model Evaluation and Validation:**

- Use cross-validation techniques to assess model performance and generalization.
- Evaluate clustering algorithms based on metrics such as silhouette score, Davies–Bouldin index, or Calinski–Harabasz index.
- Validate segmentation results by comparing them with external benchmarks or conducting customer surveys to assess segment relevance and effectiveness.

5. **Integration and Deployment:**

- Integrate the selected models into a scalable and automated pipeline for real-time or batch segmentation.
- Develop visualization tools and dashboards to communicate segmentation insights effectively to stakeholders.
- Deploy the segmentation solution in production environments for ongoing monitoring and updating as new data becomes available.

By incorporating these strategies, the Market Segmentation Project can achieve more comprehensive and accurate segmentation results, leading to better targeted marketing strategies, product customization, and overall business performance.

4. What is the estimated Market Size for your Market Domain (non-segmented) in Numbers?

--

Global AI Market: The global AI market is estimated to reach $1.56 trillion by 2028, according to Grand View Research. LLMs are considered a key driver of this growth, with applications in various sectors like healthcare, finance, retail, and manufacturing.

Natural Language Processing (NLP) Market: The NLP market is expected to reach $43.2 billion by 2026, according to MarketsandMarkets. LLMs represent a significant advancement in NLP capabilities, paving the way for new applications like chatbots, virtual assistants, and content creation tools.

Enterprise Adoption: LLMs are increasingly being adopted by enterprises across various departments, including customer service, marketing, and research & development. The potential cost savings and productivity gains from LLM-powered solutions contribute to their growing adoption.

Research & Development: LLMs are still under active development, with ongoing research exploring their capabilities and potential applications. This continuous innovation suggests further expansion of the market in the future.

5. **Name top 4 Variables/features which can be used to create most optimal Market Segments for your Market Domain.**

-- Identifying the top variables or features for creating optimal market segments depends on the specific domain and objectives of the segmentation project. However, in the context of AI and NLP markets, some key variables/features that can be useful for segmentation include:

1. **Industry Vertical:** Segmenting based on the industry verticals where AI and NLP technologies are applied can be highly relevant. Different industries have unique needs, challenges, and use cases for AI and NLP solutions. For example, healthcare, finance, retail, and customer service are some of the industries where AI and NLP adoption is widespread, each with its own set of requirements and preferences.

2. **Technology Adoption Stage:** Segmenting based on the stage of technology adoption can provide insights into the maturity of AI and NLP implementation within organizations. This could include segments such as early adopters, mainstream adopters, and laggards, each with distinct characteristics and priorities.

3. **Geographical Region:** Segmenting based on geographical regions can help account for differences in market dynamics, regulatory environments, cultural factors, and

language preferences. Global, regional, or country-level segmentation can provide valuable insights into localized trends and opportunities.

4. **Organization Size:** Segmenting based on the size of organizations, such as small and medium-sized enterprises (SMEs) versus large enterprises, can be relevant for understanding variations in AI and NLP adoption, budget allocations, decision-making processes, and technology requirements.

GitHub Link -- https://github.com/rakhisau/Feynn-Labs-Internship/blob/main/Electric%20Vehicle%20India.ipynb