FeyNN Labs Assignment

NAME : Uppara Madhurya Sri
Date : 26/08/24

1.Market Analysis of Electric Vehicle market in India

Project link : https://github.com/madhu1403/data_science_projects/tree/main

BACKGROUND:

The project is designed to enhance the understanding of electric vehicle (EV) data by exploring and segmenting the dataset based on various attributes. The objective is twofold:

1. **Clustering Analysis:** Using Principal Component Analysis (PCA) and K-Means clustering, the project aims to identify distinct clusters of EVs with similar performance and characteristics. This clustering helps in recognizing patterns and relationships between different vehicle features, such as top speed, range, and efficiency, allowing for targeted analysis and comparison of vehicle types.
2. **Predictive Modeling:** By applying regression analysis, the project seeks to model and predict the financial value of EVs based on their attributes. This involves understanding how features like acceleration, range, and powertrain influence the price, providing valuable insights for stakeholders and potential buyers.

Overall, the project combines exploratory data analysis, dimensionality reduction, and predictive modeling to offer a comprehensive view of the EV market, facilitating better decision-making and strategic planning.

DATA:

The data used in the report are obtained from the following sources:

https://github.com/madhu1403/data_science_projects/blob/main/data.csv

EDA(Exploratory Data Analysis):

1. Importing Libraries

- The code begins by importing necessary libraries such as `numpy`, `pandas`, `matplotlib`, `seaborn`, `statsmodels`, and others. These libraries are used for data manipulation, visualization, statistical analysis, and machine learning.

2. Reading the Data

- The data is read from a CSV file named `data.csv`, which is located in the `EVMarket-India` directory. The dataset contains various attributes of electric vehicles, including brand, price, top speed, acceleration, range, efficiency, and more.

3. Data Cleaning

- The code drops an unnecessary column (`Unnamed: 0`) and converts the price from Euros to INR. It also encodes the `RapidCharge` column, converting `Yes` and `No` into binary values (1 and 0).

4. Data Overview

- **`df.head()`**: Displays the first few rows of the dataset.
- **`df.info()`**: Provides information about the data types and non-null values in each column.

- `df.isnull().sum()`: Checks for missing values in the dataset.
- `df.describe()`: Provides descriptive statistics, such as mean, standard deviation, min, and max for numerical columns.

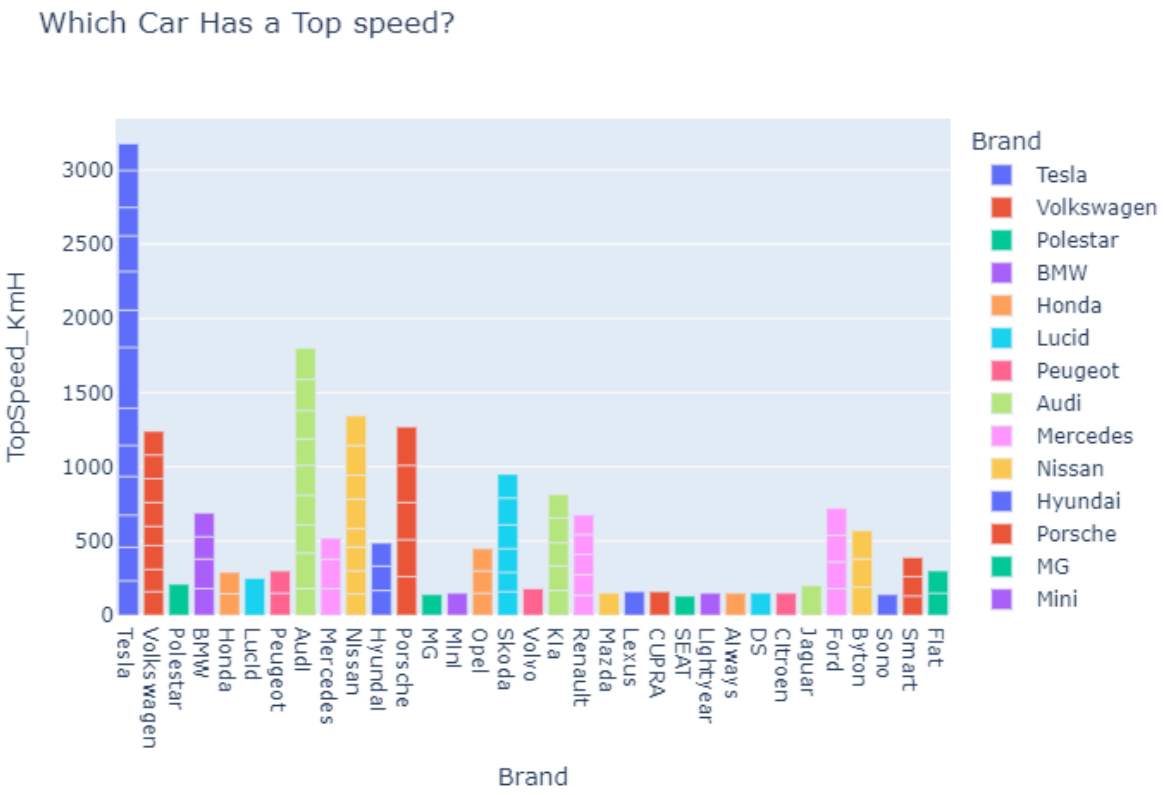
5. Descriptive Statistics

- `df.describe()`: Provides summary statistics of the numerical columns in the dataset, including count, mean, standard deviation, min, max, and quartiles.

6. Data Visualization

- Various visualizations are created using `plotly` and `seaborn` to explore the relationships between different variables in the dataset.

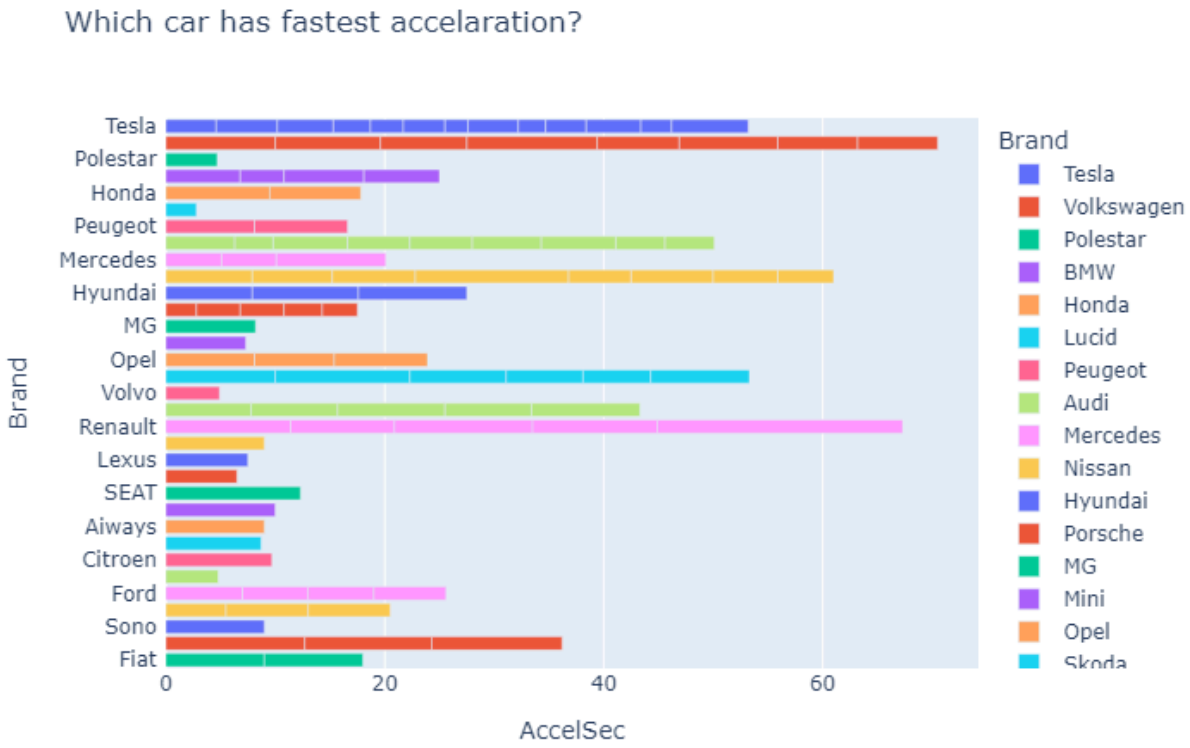
- **Top Speed by Brand**
  - A bar chart (`px.bar`) shows the top speed of cars for each brand. This helps identify which brands produce the fastest cars.



  - Brands like **Tesla,Audi** and **Porsche** likely offer the highest top speeds, positioning themselves as leaders in the performance segment. These brands cater to consumers seeking high-speed electric vehicles, emphasizing their technological prowess and engineering capabilities. In contrast, brands like

Nissan or **Hyundai** might focus more on affordability and efficiency, with slightly lower top speeds.

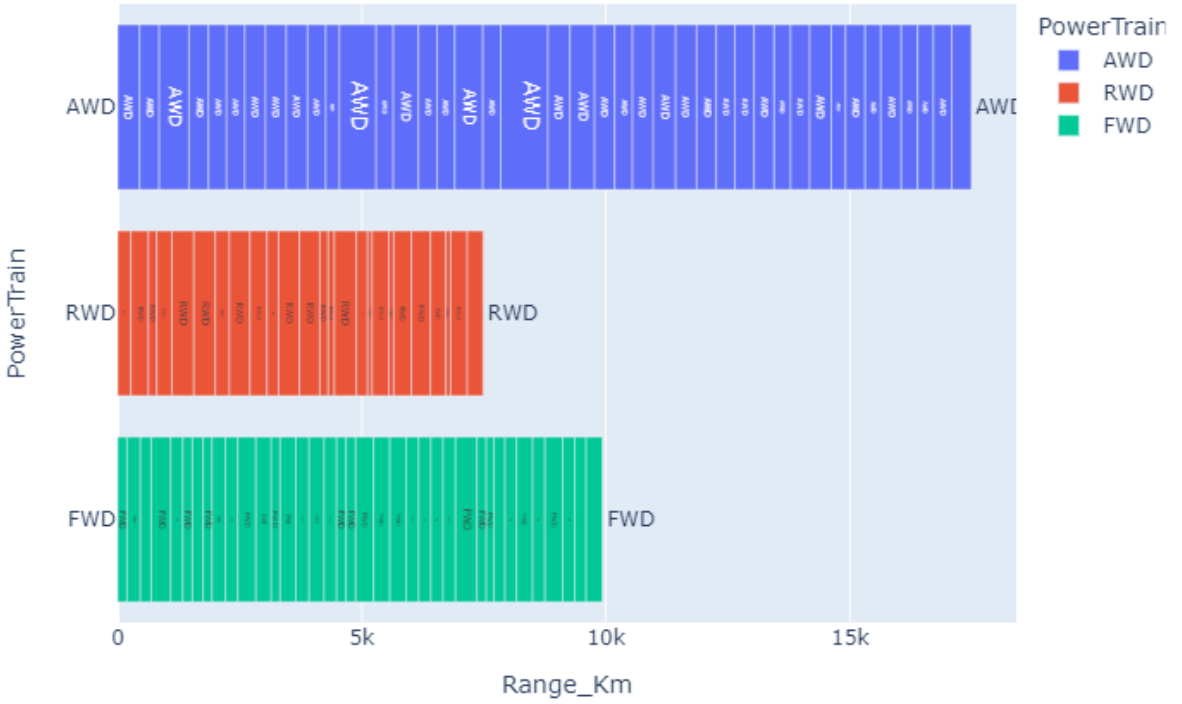
- **Acceleration by Brand**
  - A bar chart displays the acceleration times (in seconds) for cars by brand, highlighting which brands have the fastest acceleration.**Tesla,Volksvagen,Honda** and **Porsche** are expected to show the fastest acceleration times, reinforcing their image as high-performance, luxury EV brands. Consumers interested in sportier, high-acceleration vehicles are likely drawn to these brands. Other brands like **BMW** or **Audi** may also perform well, while brands such as **Chevrolet** or **Kia** might focus on balanced performance with more emphasis on practicality.



- **Range by PowerTrain**

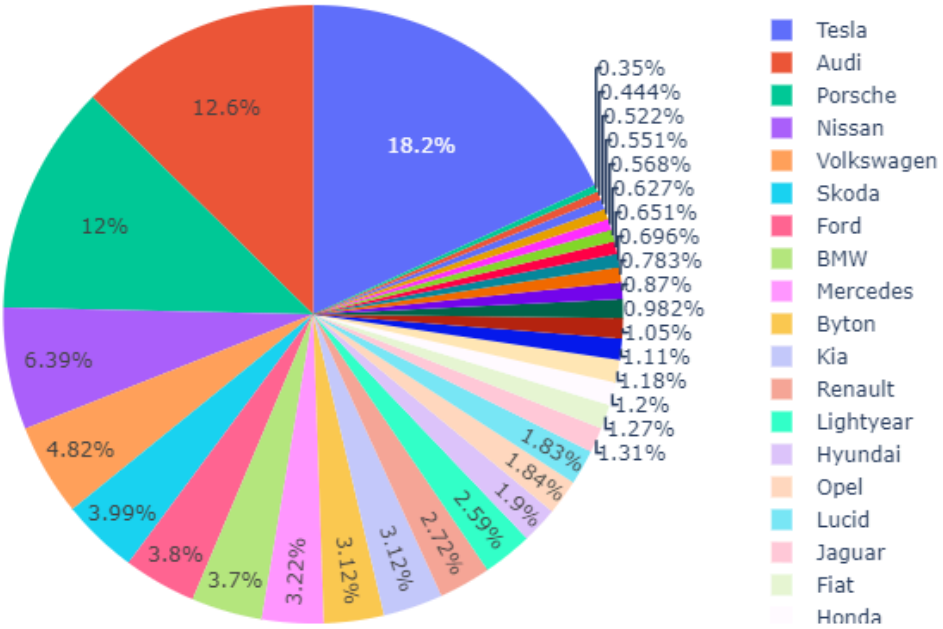
  Vehicles with Tesla's electric powertrain likely show the longest range, making them attractive to consumers who are concerned about range anxiety. This gives Tesla an edge in markets where charging infrastructure is still developing. Hybrid powertrains from brands like Toyota may offer a good balance between range and fuel efficiency, appealing to consumers who prioritize versatility.

- A bar chart shows the range of cars by their powertrain type (RWD, AWD, FWD), giving insights into how the drivetrain affects range.



- **Price by Brand:**

- A pie chart (`px.pie`) shows the price distribution of cars by brand, indicating which brands have more expensive cars.

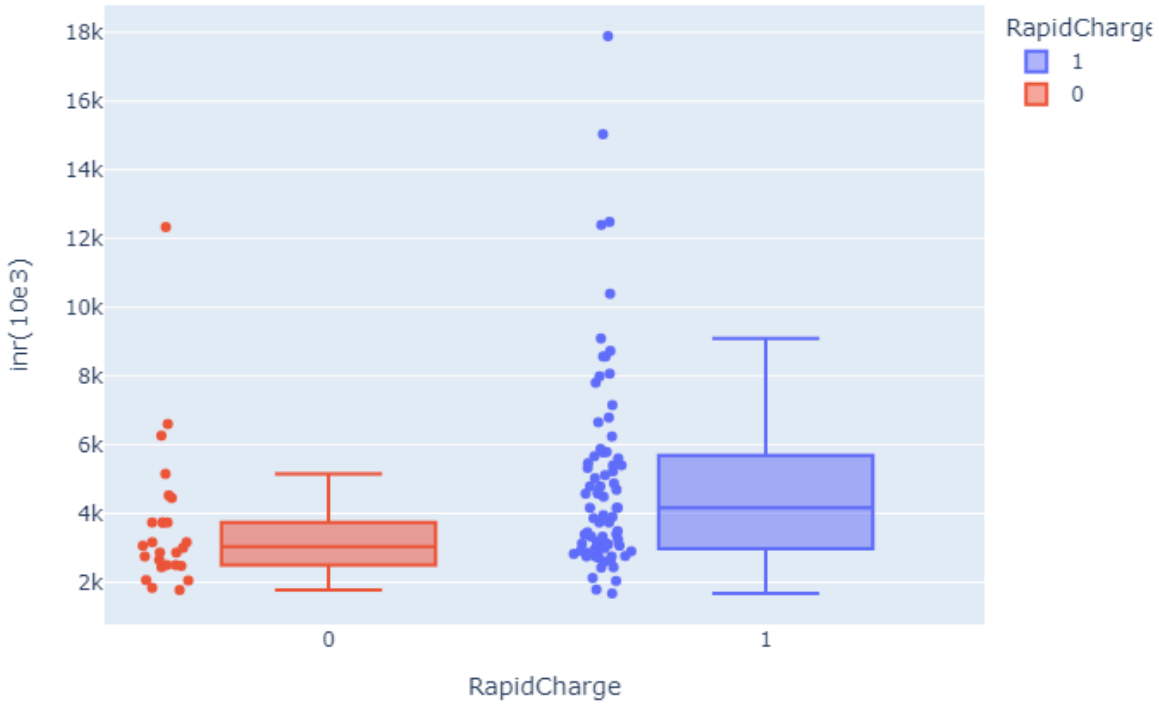


Tesla likely dominates the chart followed by Audi ,Porsche, reflecting their strong presence and brand loyalty in the EV market. **Nissan**, with its popular Leaf model, might also hold a significant share, especially in regions where affordable EVs are in demand. Brands like **BMW** and **Chevrolet** may have smaller but notable shares, targeting different niches within the EV market

- **Price vs. Rapid Charging Capability**

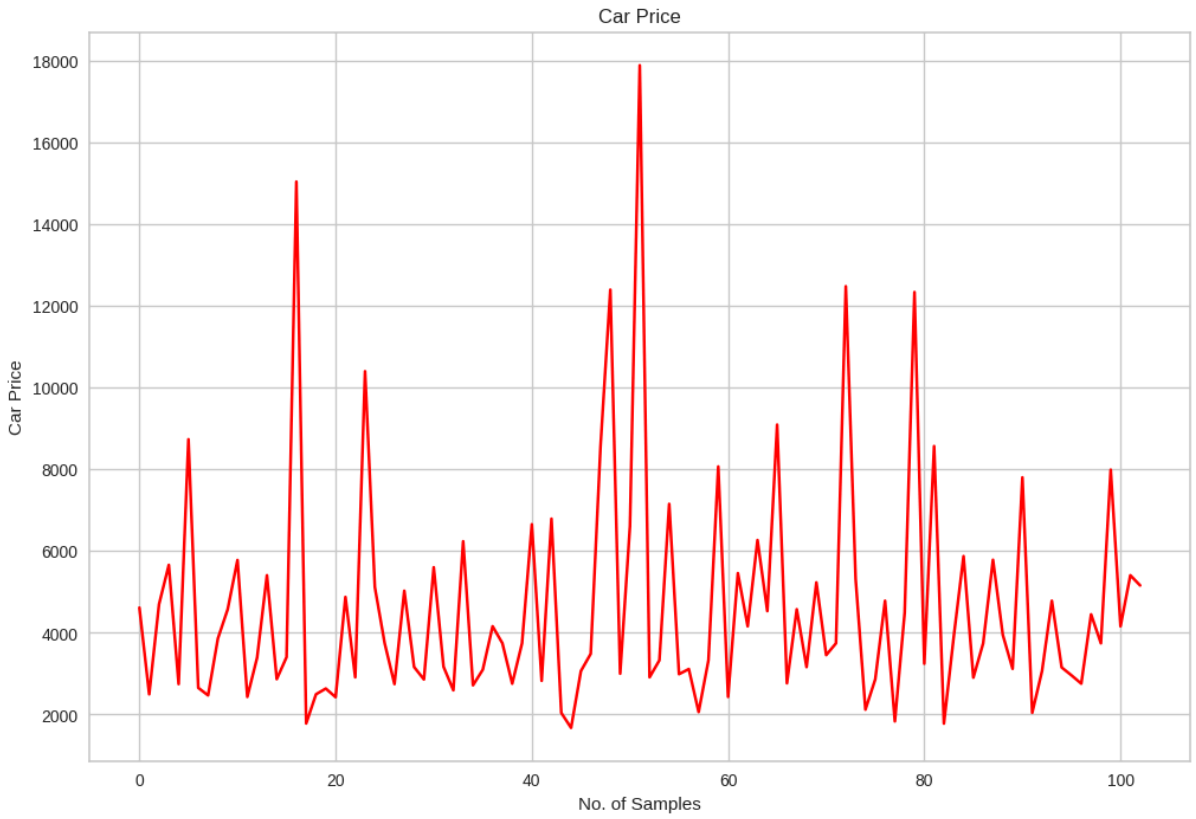
    - A box plot compares car prices based on whether they have rapid charging capability.Vehicles from **Tesla** and **Porsche** with rapid charging capabilities are likely positioned in the higher price range, emphasizing the premium nature of these features. Rapid charging is a significant selling point for these brands, appealing to consumers willing to pay more for the convenience of faster charging times. Lower-cost brands like **Hyundai** or **Chevrolet** might

offer rapid charging as well, but possibly at a lower premium.



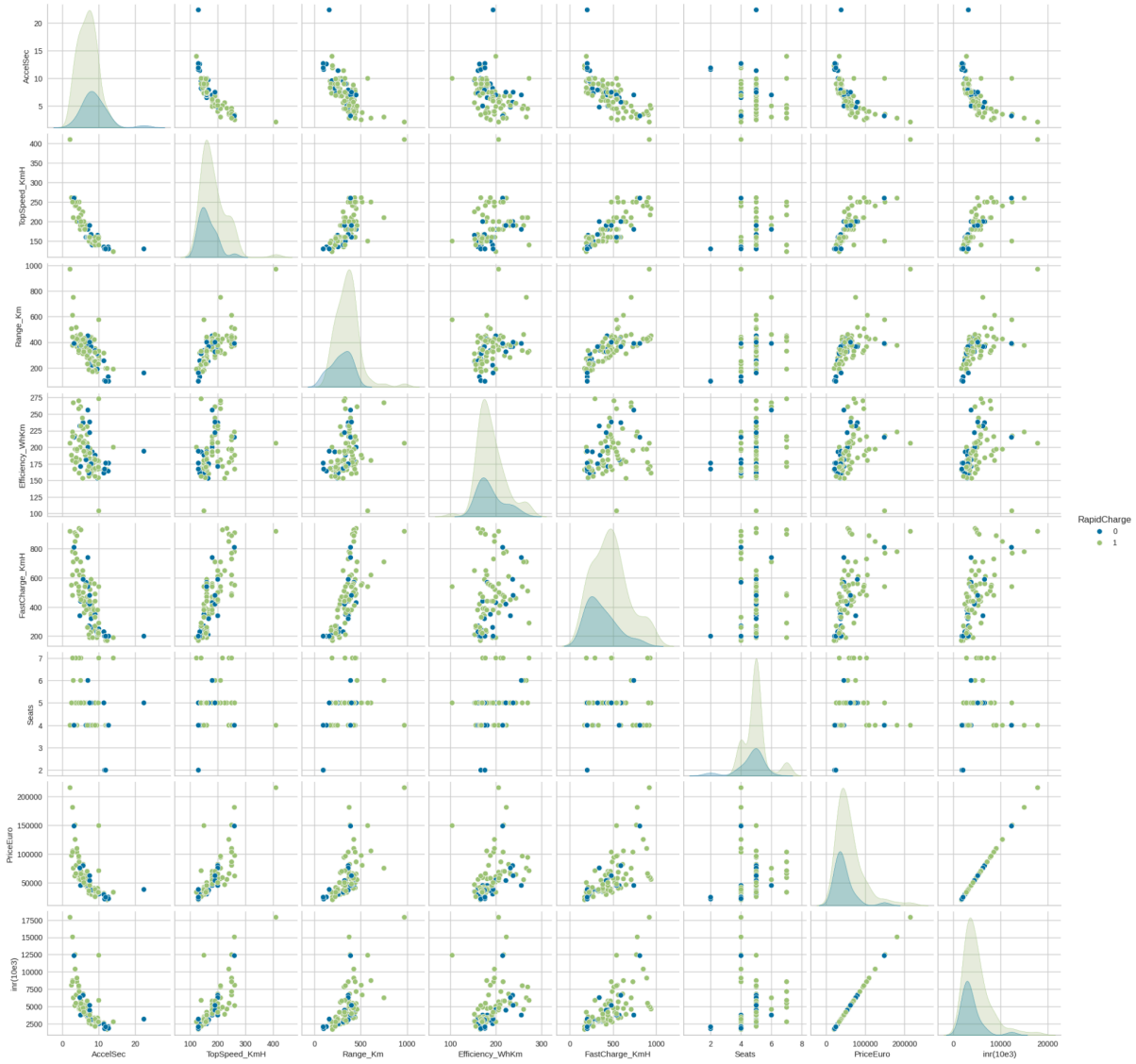
Price Distribution

- A line plot of car prices in INR (`df['inr(10e3)'].plot`) provides a visual distribution of car prices.



Pairplot by Rapid Charger Presence

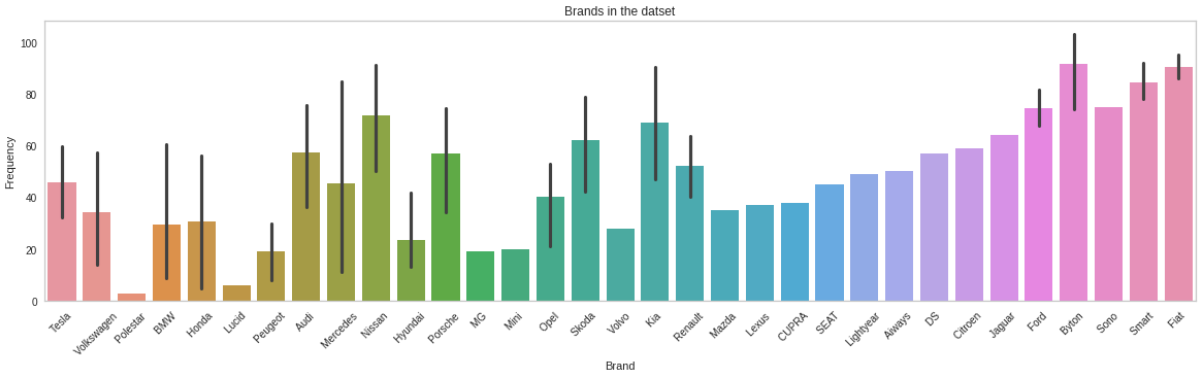
A pairplot created using `seaborn` visualizes the relationships between all numerical variables, color-coded by the presence of a rapid charger.



The pair plot may reveal that **Tesla** vehicles, while more expensive, generally offer better ranges and faster charging times, reinforcing their premium positioning. **Chevrolet** and **Nissan** could show a focus on affordability with good, but not top-tier, performance in range and charging. This helps highlight the different trade-offs consumers make when choosing between brands.

7. Brand Frequency

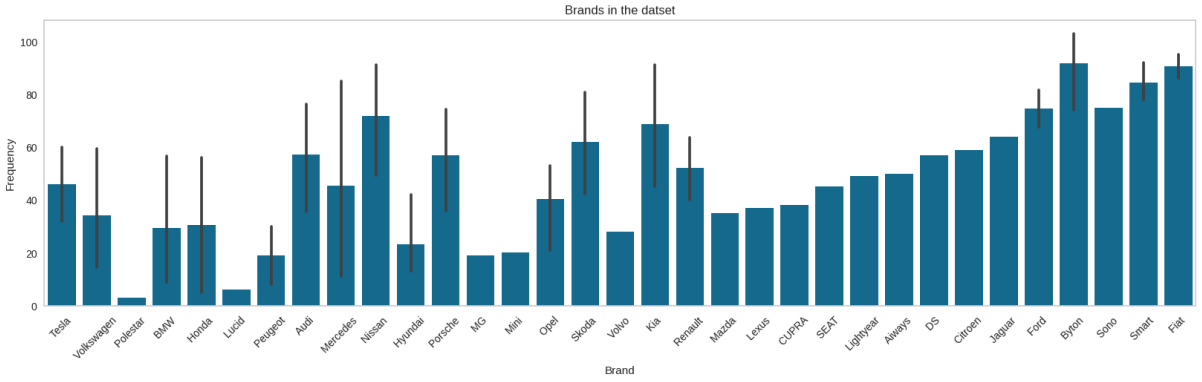
- A bar plot (`sb.barplot`) shows the frequency of each brand in the dataset, highlighting which brands are most common in the dataset.



Brands like Byton,Ford,smat,fiat,ford,nissan might have the highest counts, indicating a broad range of models or significant representation in the market. This suggests these brands have diverse offerings, appealing to different consumer needs. Brands with lower counts could be more niche or new entrants to the EV market, focusing on specific segments or innovative features.

8. Top Speed by Brand

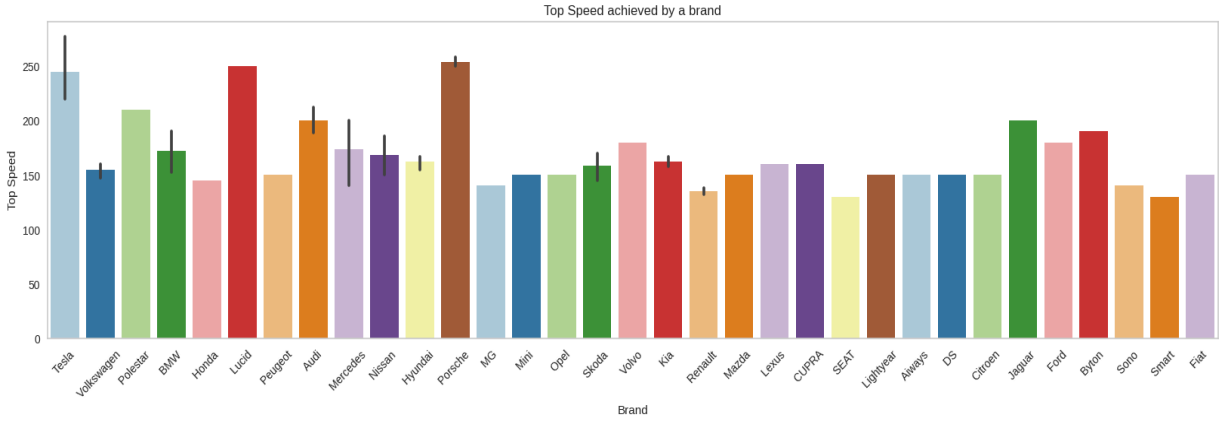
- A bar plot shows the top speed of cars for each brand, indicating which brands produce the fastest vehicles.



Similar to the earlier top speed plot, this one likely reinforces the positioning of brands like **Tesla** and **Porsche** as leaders in high-speed electric vehicles. Their focus on top speeds highlights their commitment to performance and technological excellence, attracting enthusiasts who prioritize speed. Other brands with lower speeds might emphasize other strengths like efficiency, safety, or affordabilit

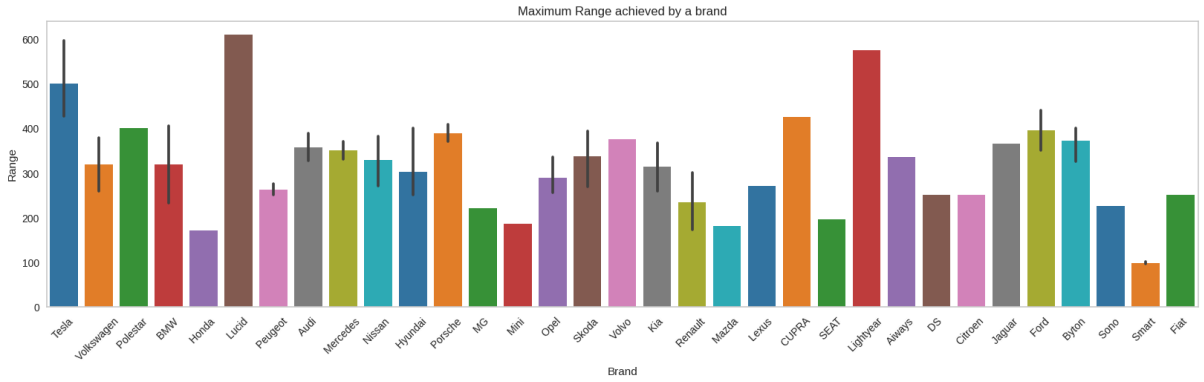
9. Range by Brand

- A bar plot shows the maximum range that cars from each brand can achieve, helping to identify which brands offer the longest range vehicles.



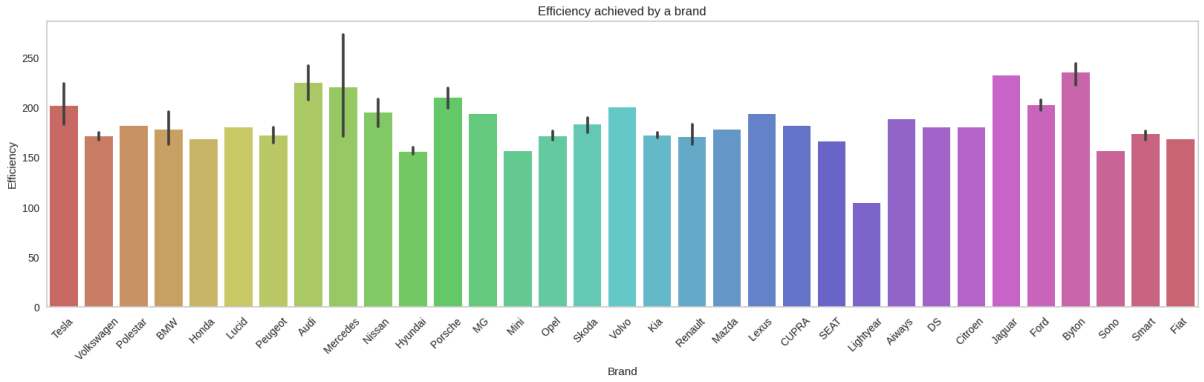
10. Car Efficiency by Brand

- A bar plot visualizes the efficiency (in Wh/km) of cars by brand, showing which brands produce the most energy-efficient vehicles.



11. Seats by Brand

- A bar plot displays the number of seats in cars for each brand, indicating which brands offer more seating capacity.

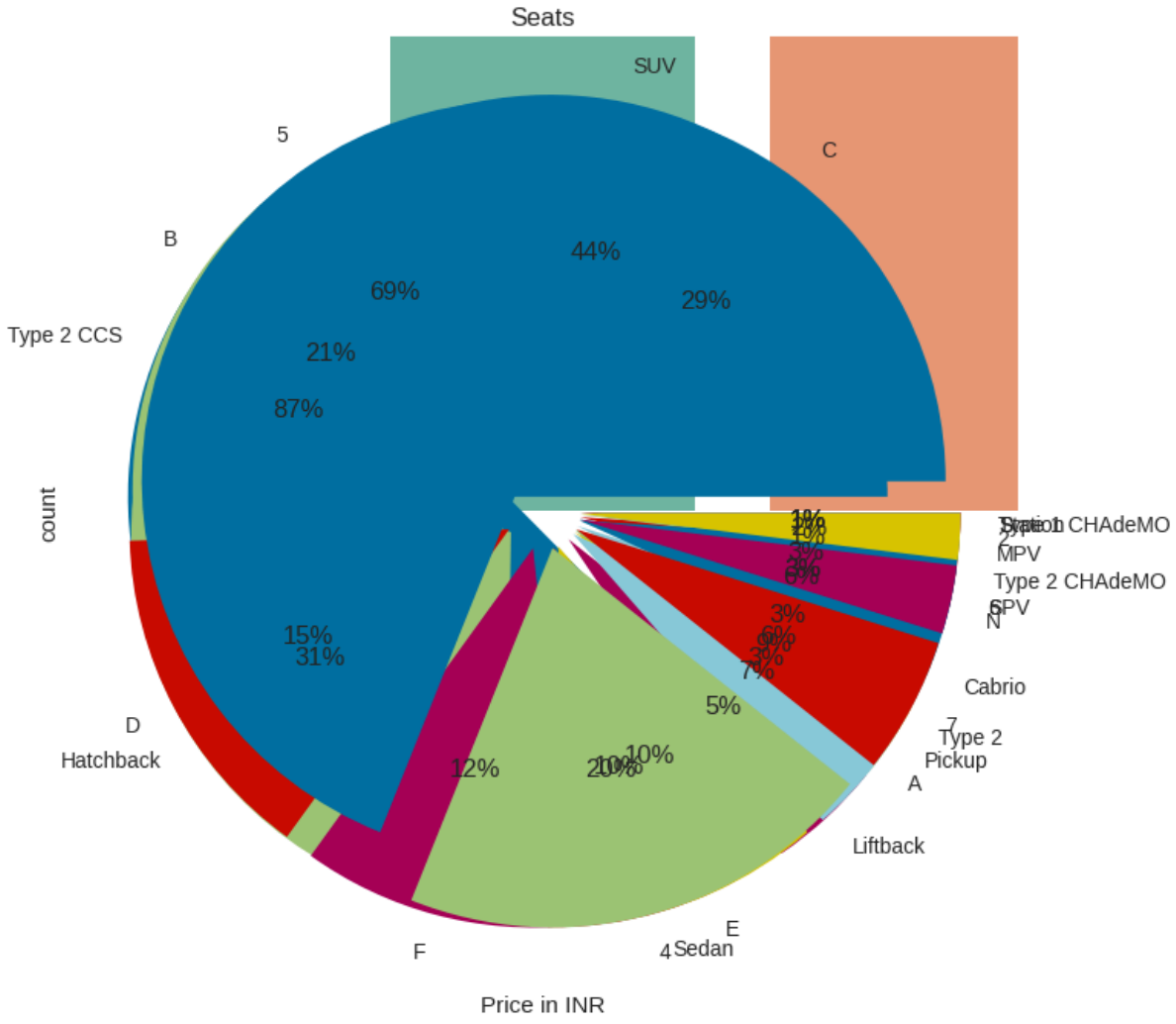


12. Price of Cars by Brand

- A bar plot visualizes the price of cars (in INR) by brand, showing which brands have the most expensive and the least expensive cars.

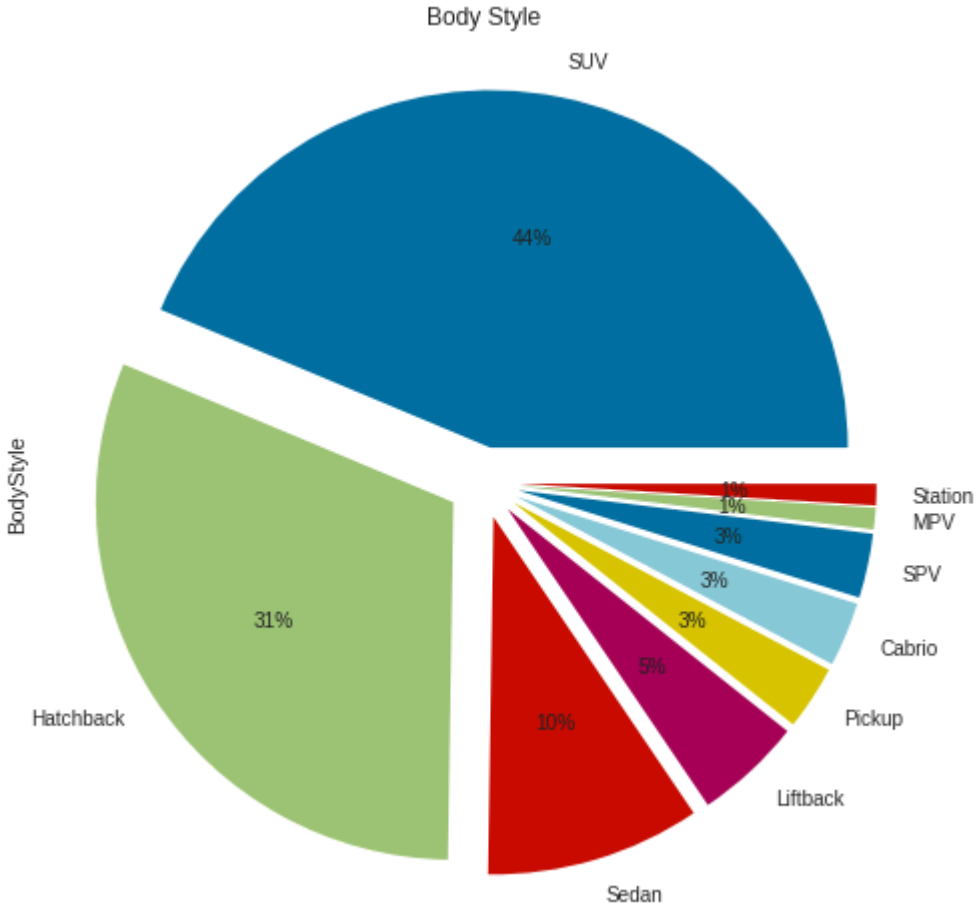
13. Plug Type Distribution

- A pie chart shows the distribution of plug types used for charging, indicating the most common and least common plug types among the cars.



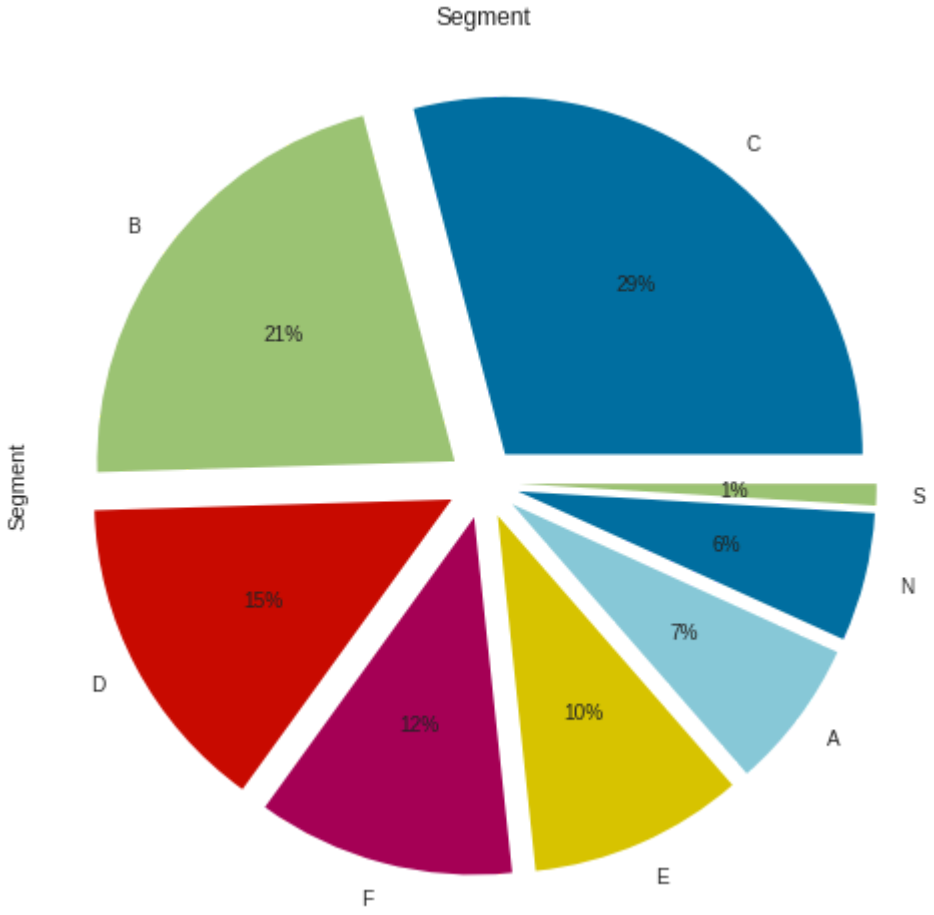
14. Body Style Distribution

- A pie chart displays the distribution of body styles (e.g., SUV, hatchback) among the cars in the dataset.



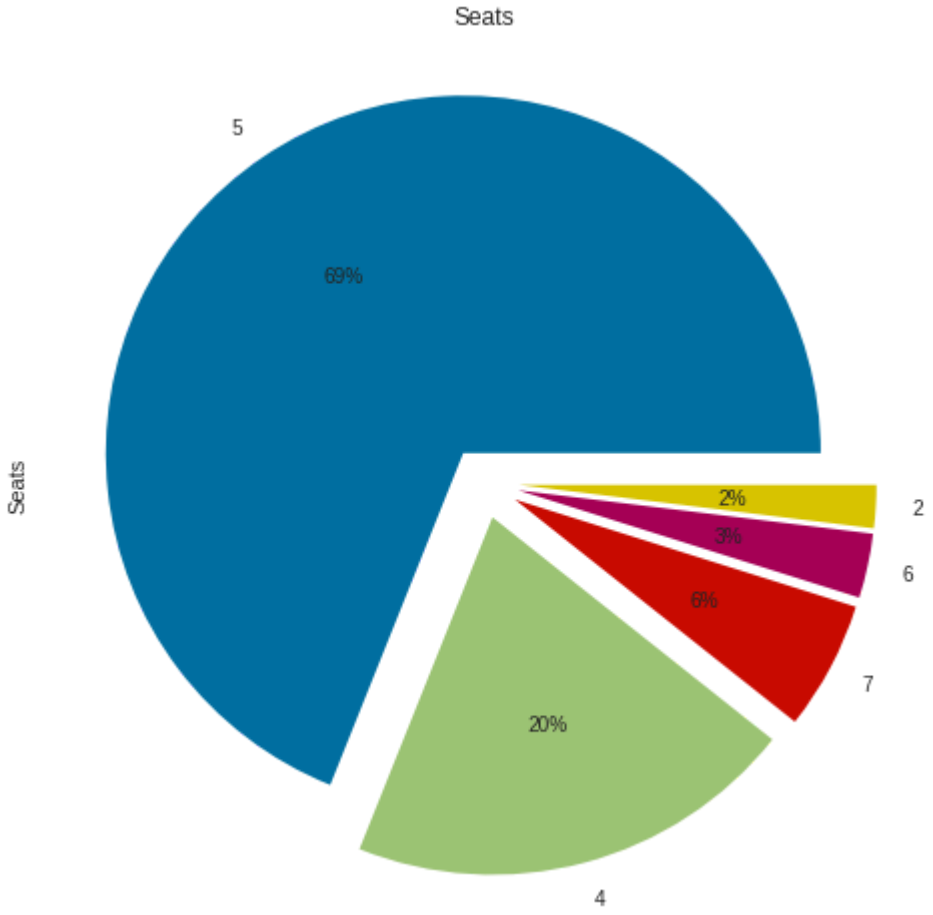
15. Segment Distribution

- A pie chart shows the distribution of cars by segment, highlighting which segments (e.g., C, B) are most common.



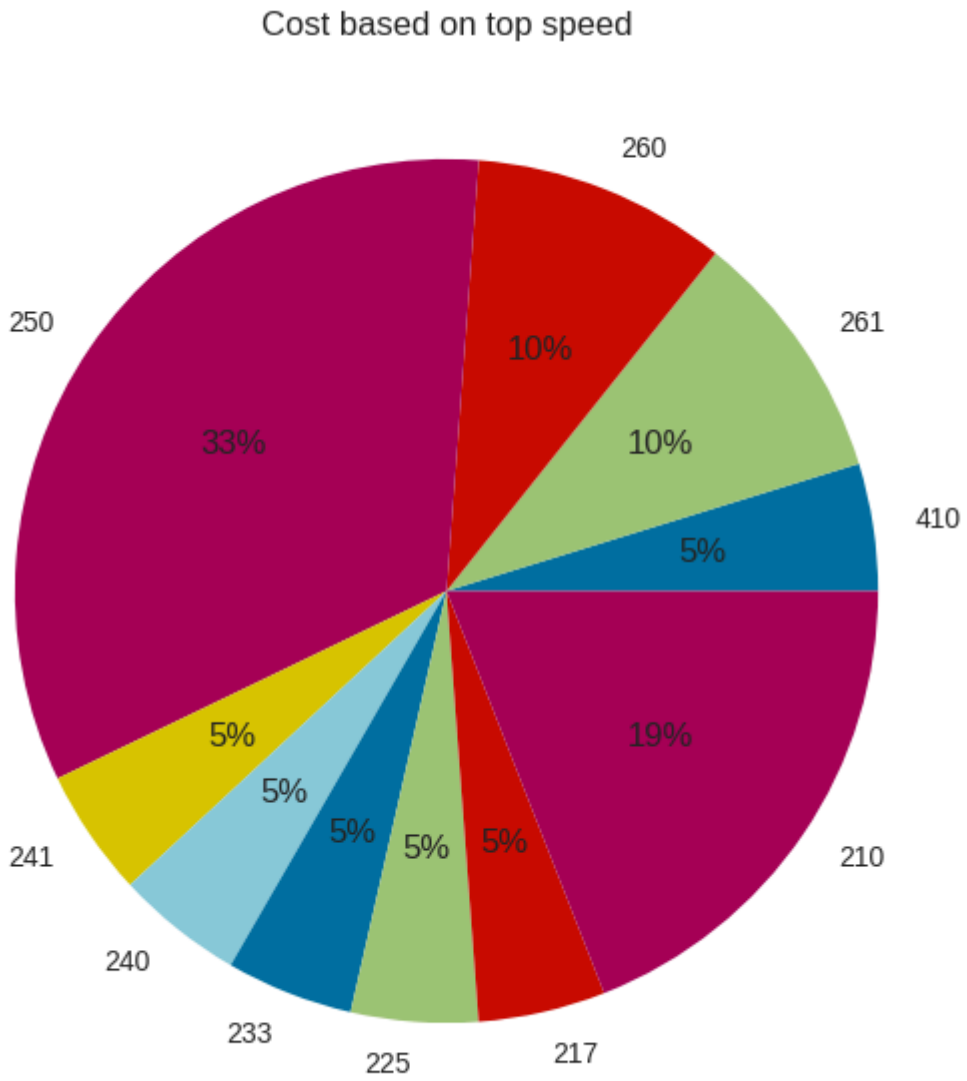
16. Seats Distribution

- A pie chart displays the distribution of cars based on the number of seats, indicating the most common seating capacities.

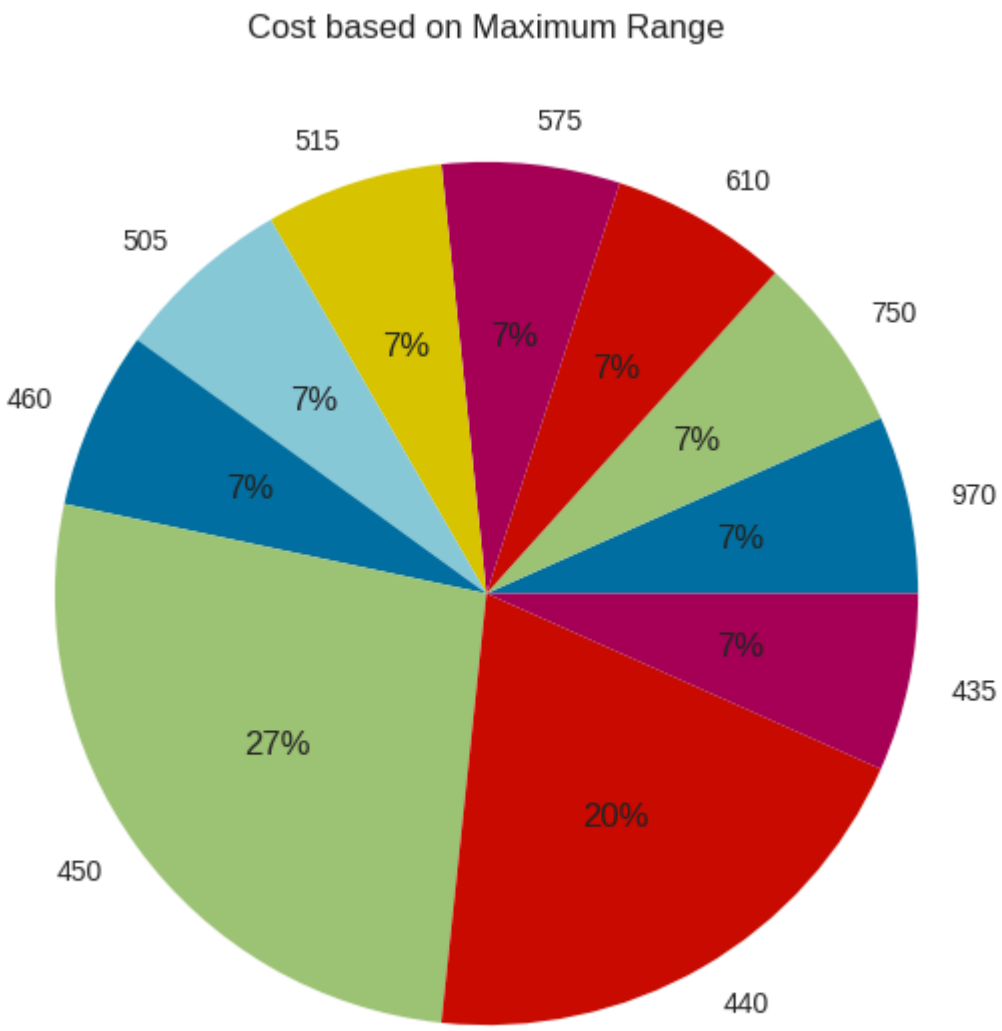


17. Cost Analysis Based on Top Speed and Range

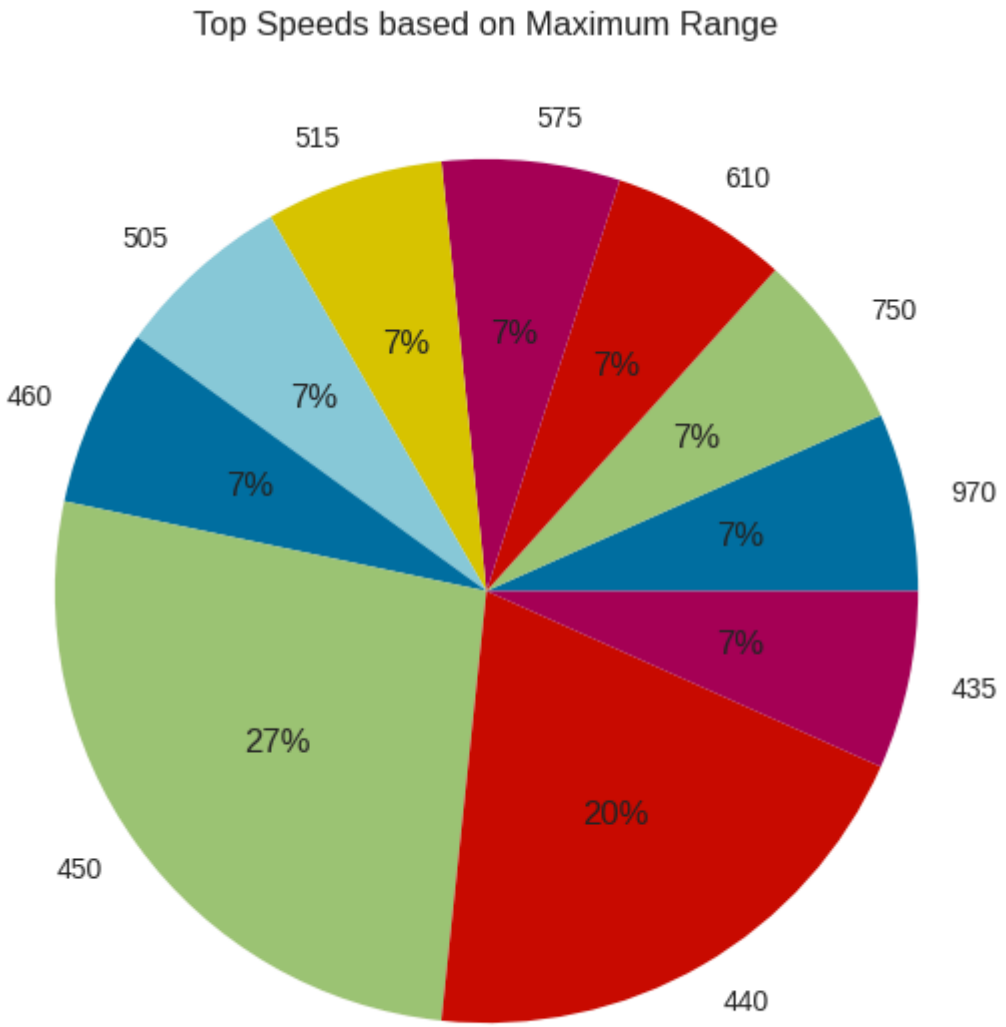
- **Cost by Top Speed**: A pie chart visualizes the cost distribution based on the top speed of cars.



- **Cost by Range**: A pie chart shows the cost distribution based on the maximum range of cars.


Cost based on Maximum Range
575
515
610
505
750
7%
7%
7%
460
7%
7%
970
7%
7%
435
27%
20%
450
440

- **Top Speed by Range**: A pie chart visualizes the top speed distribution based on the maximum range of cars.



18. Regression Analysis

- **Independent and Dependent Variables**: The independent variables (x) include acceleration, range, top speed, efficiency, rapid charging capability, and powertrain type. The dependent variable (y) is the price in Euros.
- **OLS Regression**: An Ordinary Least Squares (OLS) regression model is fitted to the data to analyze the relationship between the independent variables and the price. The model summary provides insights into the significance of each variable.

19. Train-Test Split and Linear Regression

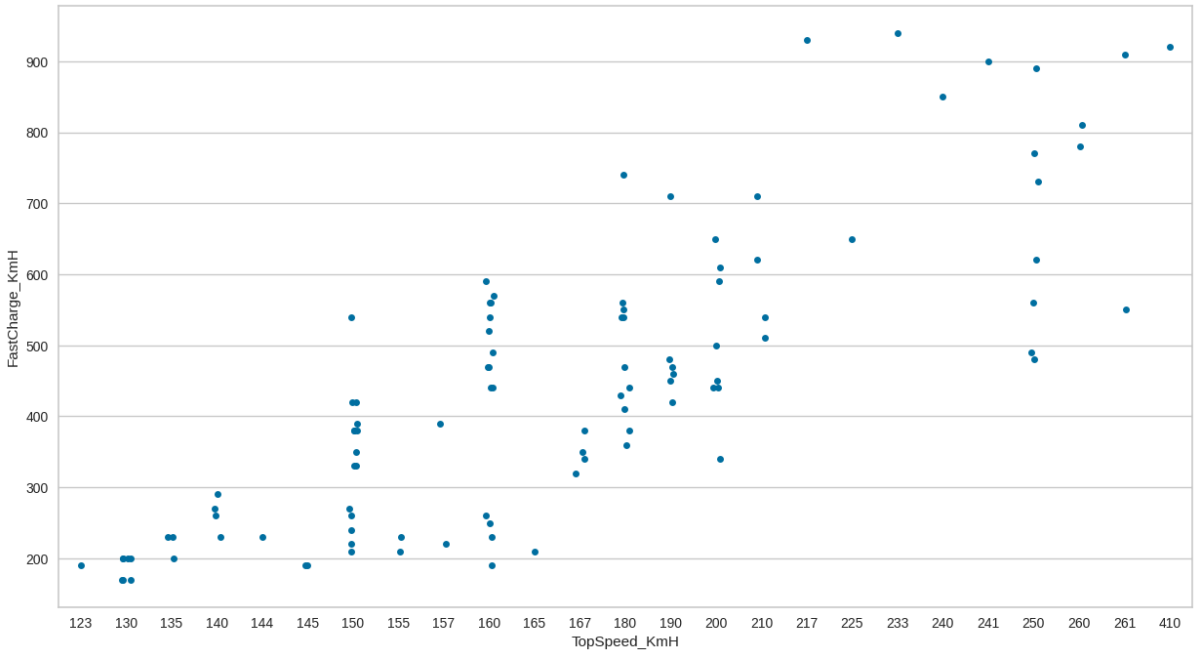
- The data is split into training and testing sets, and a linear regression model is fitted to the training data. The model is used to predict prices on the test set, and the R-squared value is calculated to evaluate the model's performance.

20. Logistic Regression for Rapid Charging

- Logistic regression is performed to predict whether a car has rapid charging capability based on its price. The confusion matrix is generated to evaluate the performance of the logistic regression model.

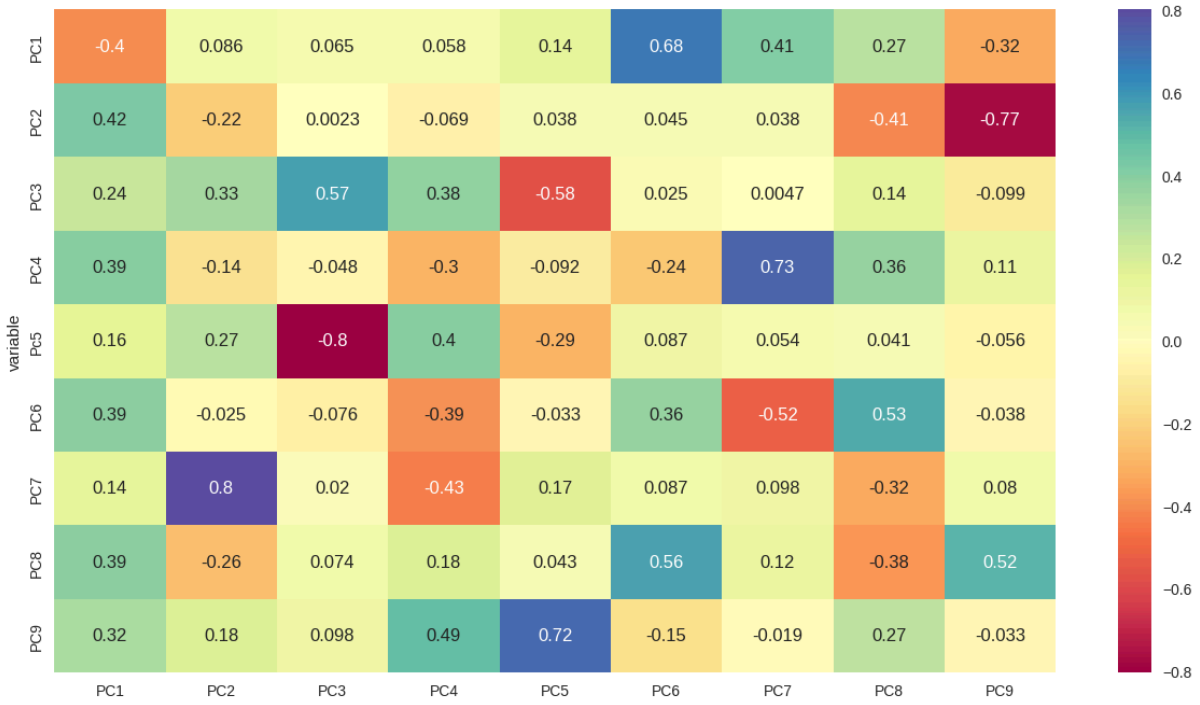
21. Clustering Analysis

- **Strip Plots**: Strip plots visualize the relationship between top speed and fast charging capability, as well as top speed and efficiency.



5. Heatmap of Loadings

A heatmap is plotted to visualize the correlation coefficients between the original features and the principal components.



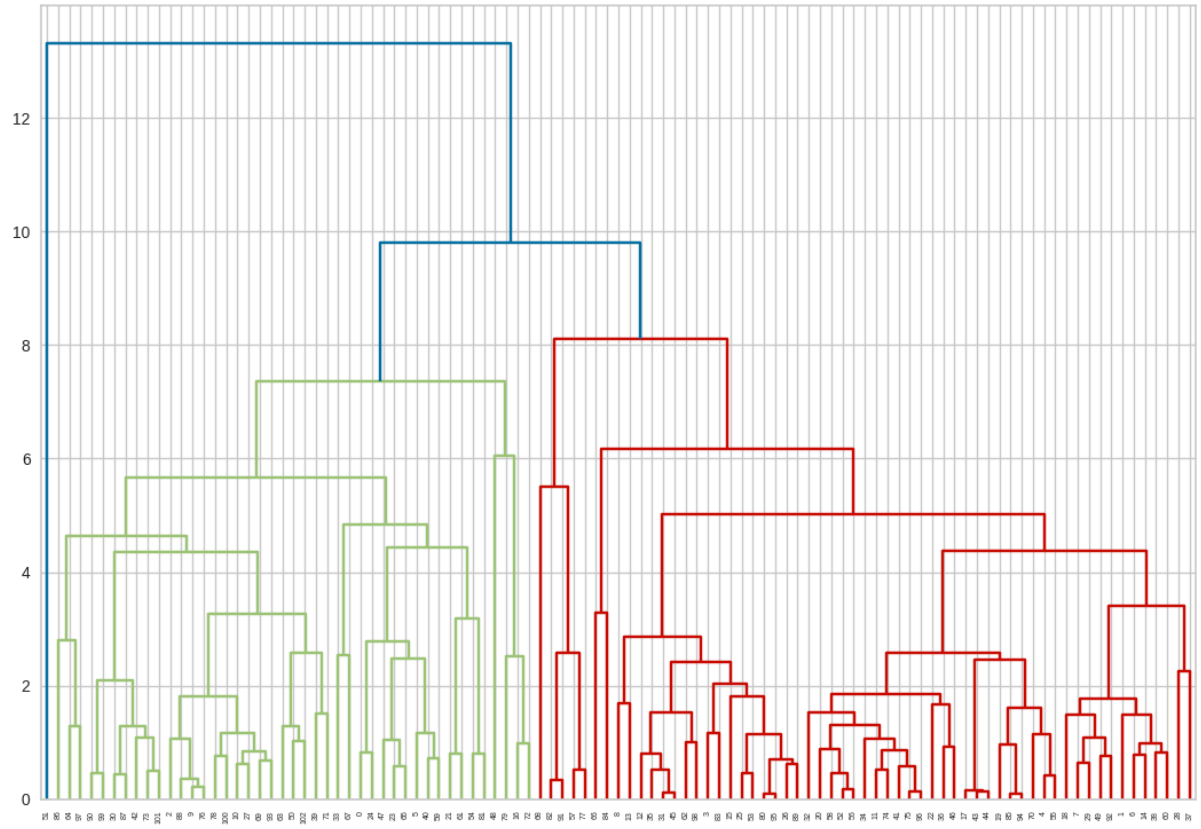
Heatmap: The heatmap visually shows which features are most correlated with each principal component. High correlation values are indicated by more intense colors.

Strong correlations might be observed between **price** and **range** or **top speed**, particularly for luxury brands like **Tesla**, **Porsche**, and **BMW**, indicating that higher-priced models tend to offer superior performance. A negative correlation between **price** and **charging time** could suggest that more expensive models generally have faster charging capabilities, which is a key differentiator in the premium EV market.

6. Hierarchical Clustering (Dendrogram)

Hierarchical clustering is applied to the PCA-transformed data, and a dendrogram is plotted.

Dendrogram: The dendrogram represents the hierarchical relationships between clusters. The 'complete' linkage method is used, which considers the maximum distance between elements when forming clusters.

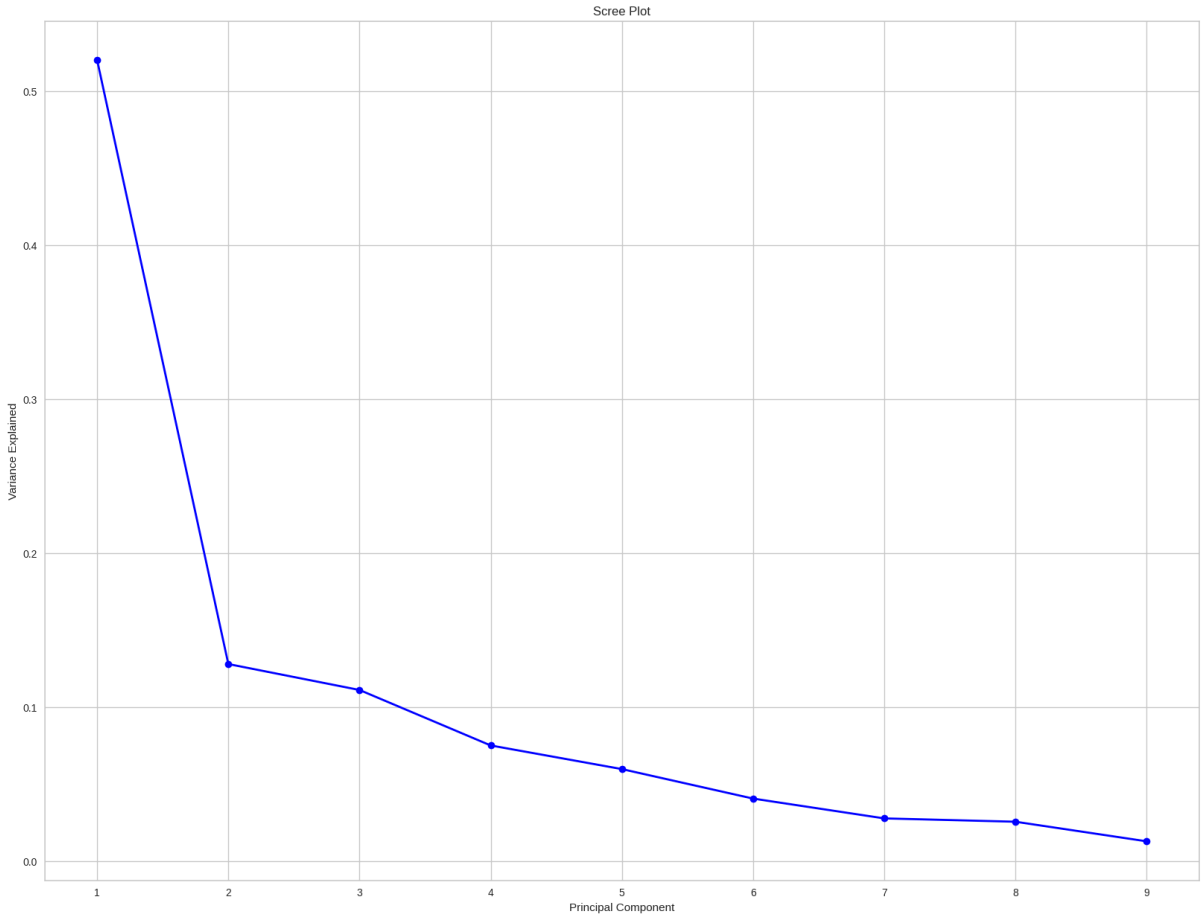


- **PCA and K-Means Clustering**:
  - Principal Component Analysis (PCA) reduces the dimensionality of the data.
  - K-Means clustering is applied to group the cars into clusters based on the principal components.
  - Various methods like the elbow method, silhouette method, and Calinski-Harabasz index are used to determine the optimal number of clusters.
  - The clusters are visualized using scatter plots, and the cluster centers are highlighted.

7. Scree Plot

A scree plot is created to show the explained variance ratio of each principal component.

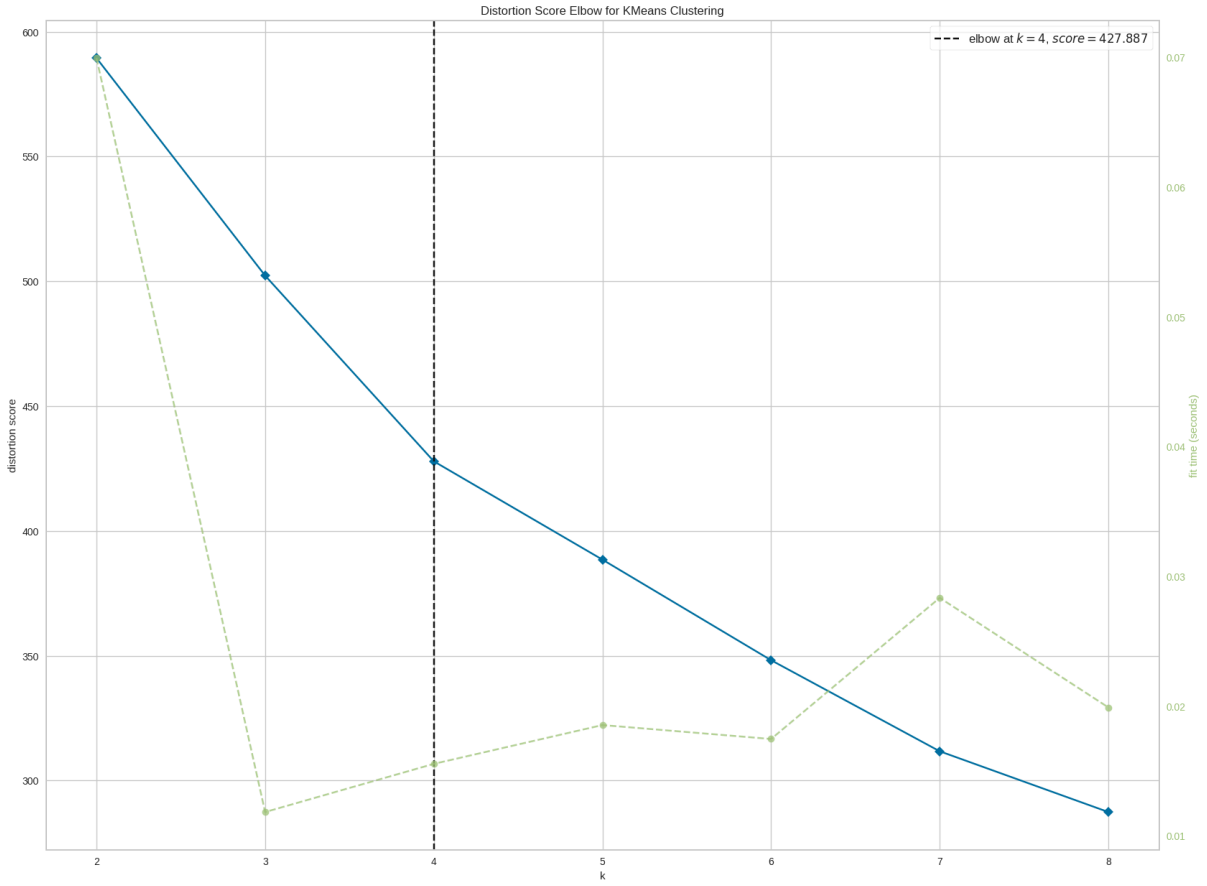
Scree Plot: This plot shows how much variance each principal component explains. The elbow point can help determine the optimal number of components to retain.

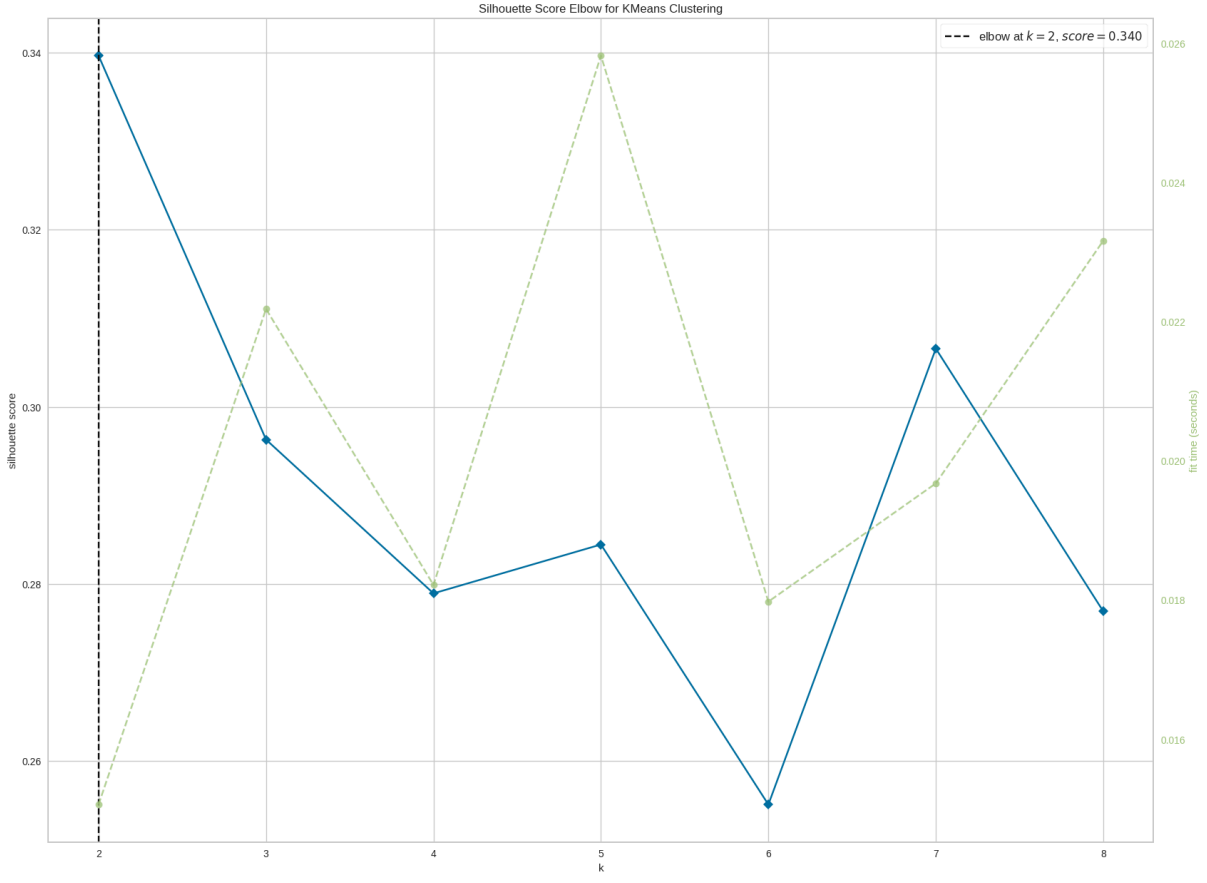


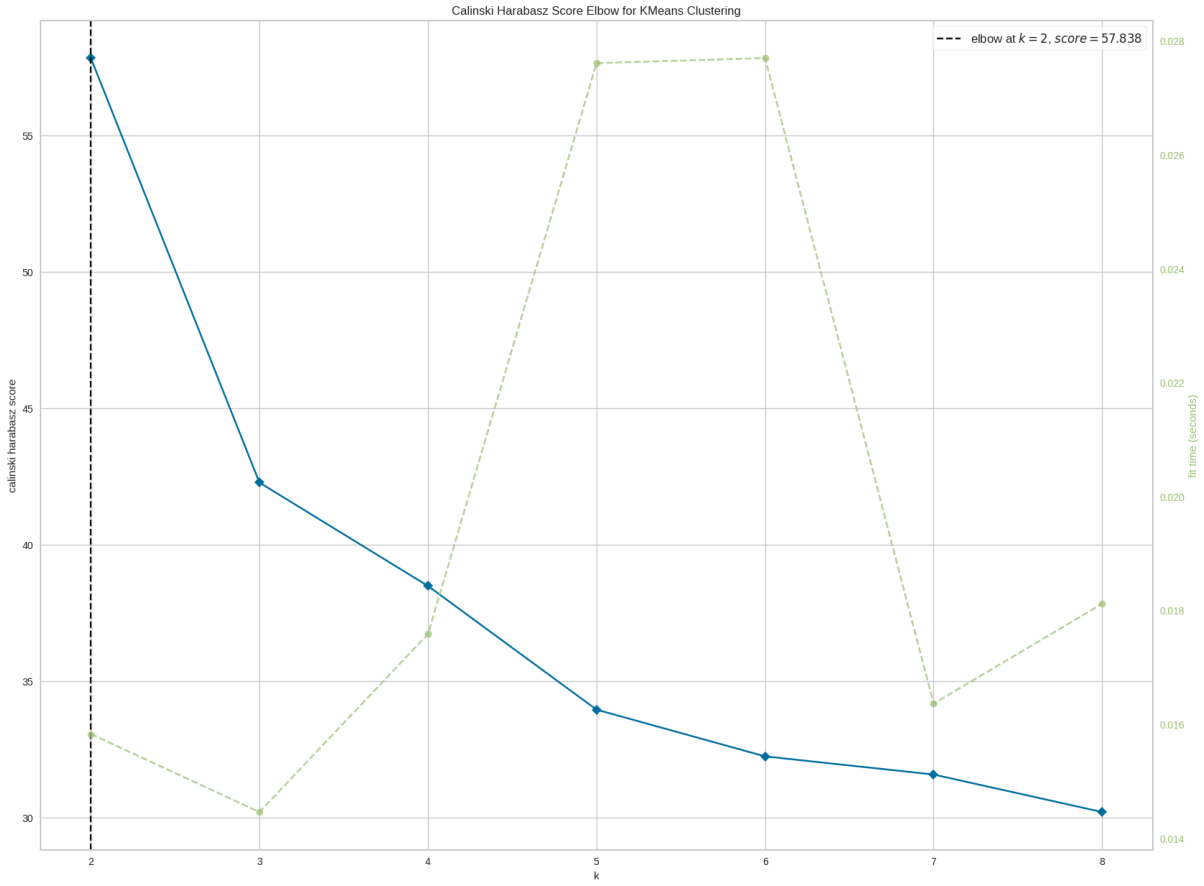
8. K-Means Clustering and Elbow Method

The elbow method is used to determine the optimal number of clusters for K-Means clustering.

Elbow Method: The visualizer shows different metrics (distortion, silhouette score, and Calinski-Harabasz index) to help determine the optimal number of clusters.


Distortion Score Elbow for KMeans Clustering
elbow at $k = 4$, $score = 427.887$
distortion score
fit time (seconds)
k
600
550
500
450
400
350
300
0.07
0.06
0.05
0.04
0.03
0.02
0.01
2
3
4
5
6
7
8

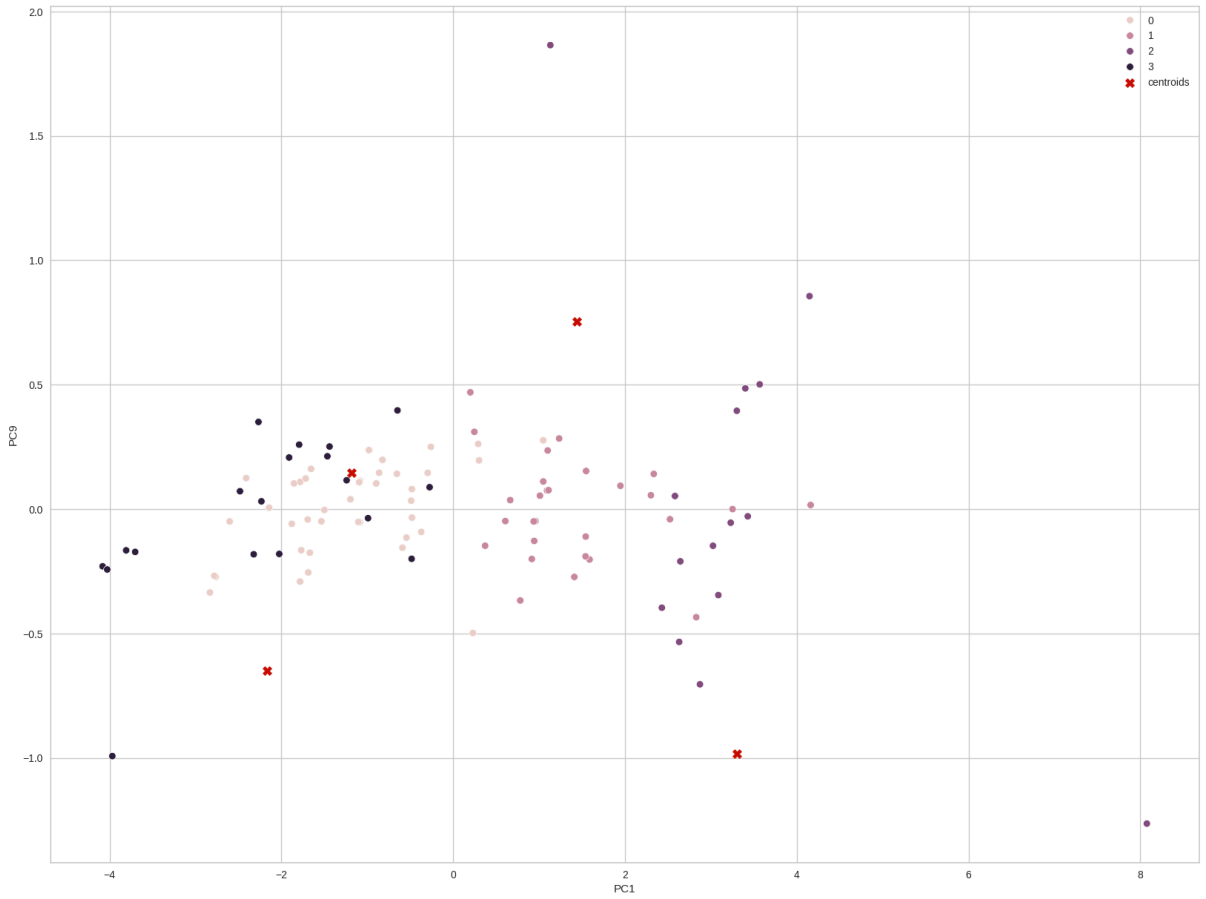

Silhouette Score Elbow for KMeans Clustering
elbow at $k = 2$, $score = 0.340$
silhouette score
fit time (seconds)
k
0.34
0.32
0.30
0.28
0.26
0.026
0.024
0.022
0.020
0.018
0.016
2
3
4
5
6
7
8



Cluster Visualization

The clusters are visualized using a scatter plot of the first two principal components.

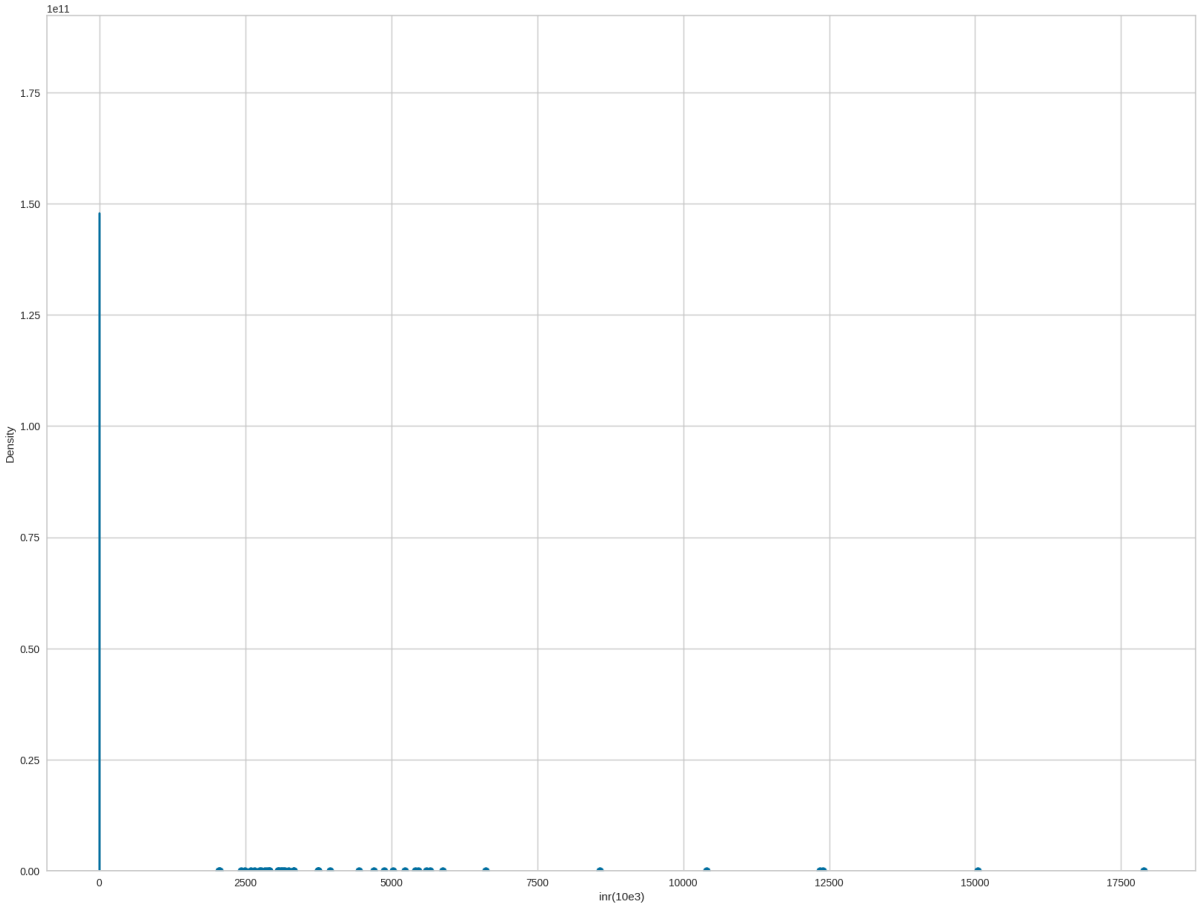
Scatter Plot: This plot shows the data points colored by their cluster labels, with cluster centroids marked with red X's.



22. Regression on Principal Components

- A linear regression model is fitted to the principal components obtained from PCA to predict the price in INR. The model's coefficients, predictions, and residuals are analyzed, and metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are calculated to evaluate the model's performance.
- Predictions and Evaluation:

  Scatter Plot of Predictions: This plot shows the relationship between actual and predicted values. Ideally, the points should lie close to the diagonal, indicating good predictions.

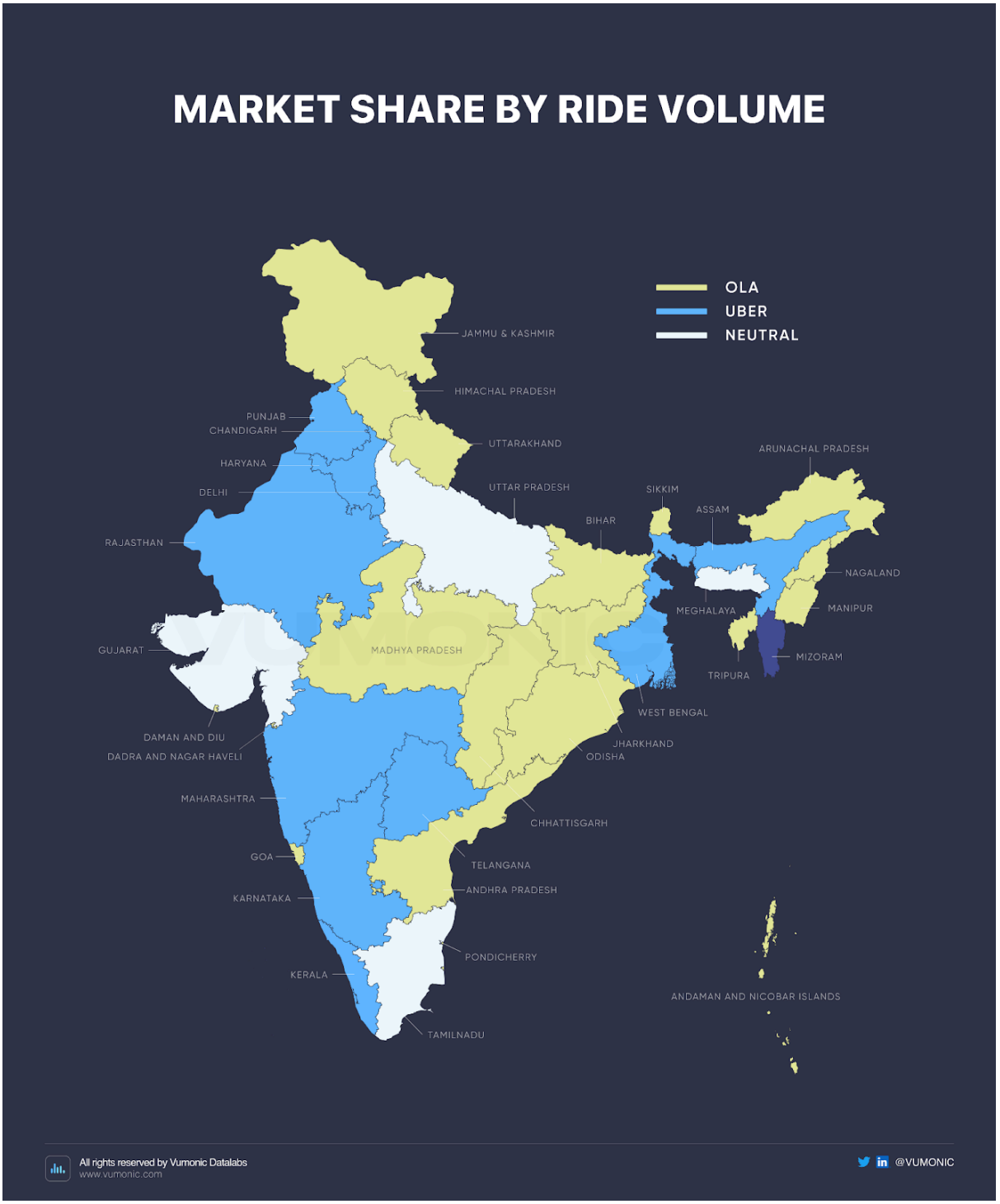


This analysis provides a detailed exploration of the EV market in India, offering insights into various aspects of electric vehicles, including their performance, efficiency, pricing, and market segmentation. The combination of visualizations, regression, and clustering analysis helps to uncover patterns and relationships in the data, making it a comprehensive study of the Indian EV market.

2. Online Vehicle Booking Market : ola and uber

Project link:
https://github.com/madhu1403/data_science_projects/blob/main/ola-vs-uber.ipynb

Background



–

Overview of the Ride-Hailing Market

The ride-hailing industry has experienced significant growth in recent years, driven by the increasing demand for convenient and cost-effective transportation options. Major players in this market, such as Ola and Uber, have established themselves as dominant forces, offering a range of services to cater to diverse customer needs.

- **Ola**: Founded in 2010, Ola is an Indian ride-hailing company that has expanded its operations both domestically and internationally. It offers various services, including ride-sharing, auto-rickshaw rentals, and electric vehicle options.

- **Uber**: Established in 2009, Uber is a global leader in the ride-hailing industry. It operates in numerous countries, providing a range of services from standard rides to luxury options and food delivery through Uber Eats.

Importance of Customer Reviews

Customer reviews play a crucial role in shaping the reputation and success of ride-hailing services. They provide valuable insights into user experiences, highlighting areas of strength and opportunities for improvement. Analyzing these reviews helps companies understand customer sentiment, address issues, and enhance service offerings.

Objectives of the Analysis

The primary objective of this analysis is to evaluate customer sentiment from reviews of Ola and Uber. By performing sentiment analysis on a substantial dataset of customer reviews, we aim to:
- **Understand Customer Perceptions**: Identify how customers feel about the services provided by Ola and Uber.
- **Identify Common Complaints and Praise**: Highlight frequent issues and positive aspects mentioned in the reviews.
- **Enhance Competitive Strategies**: Provide actionable insights to improve service quality and competitiveness in the market.

This analysis will help both Ola and Uber leverage customer feedback to refine their services and maintain a competitive edge in the ride-hailing industry.

Based on the data analysis and sentiment analysis conducted for Ola and Uber customer reviews, here is a project report summarizing the findings and conclusions:

DATA:

The data used in the report are obtained from the following sources:

https://www.kaggle.com/code/khushipitroda/ola-vs-uber/input?select=Ola+Customer+Reviews.csv

---

Project Report: Sentiment Analysis of Ola and Uber Customer Reviews

1. Introduction

This report presents the findings from an analysis of customer reviews for Ola and Uber, focusing on sentiment analysis. The goal was to understand customer perceptions and identify key insights that could help in enhancing service quality and gaining a competitive edge.

2. Data Overview

Datasets
- **Ola Customer Reviews**: 357,698 reviews.
- **Uber Customer Reviews**: 1,069,616 reviews.

Columns Analyzed
- Review description
- Rating

3. Data Preprocessing

Text Cleaning
The text data from the reviews was preprocessed to improve the quality of the text used for sentiment analysis. This included:
- **Removing Stopwords**: Unnecessary words that do not contribute to the meaning were removed.
- **Text Normalization**: Converted all text to lowercase and removed non-alphabetic characters.
- **Stemming**: Words were reduced to their root forms to ensure uniformity in analysis.

4. Sentiment Analysis

Methodology
- **Feature Extraction**: Text reviews were transformed into numerical features using the Bag-of-Words approach.
- **Model Training**: A Multinomial Naive Bayes classifier was used for sentiment classification, where reviews were classified into positive or negative sentiments based on ratings.

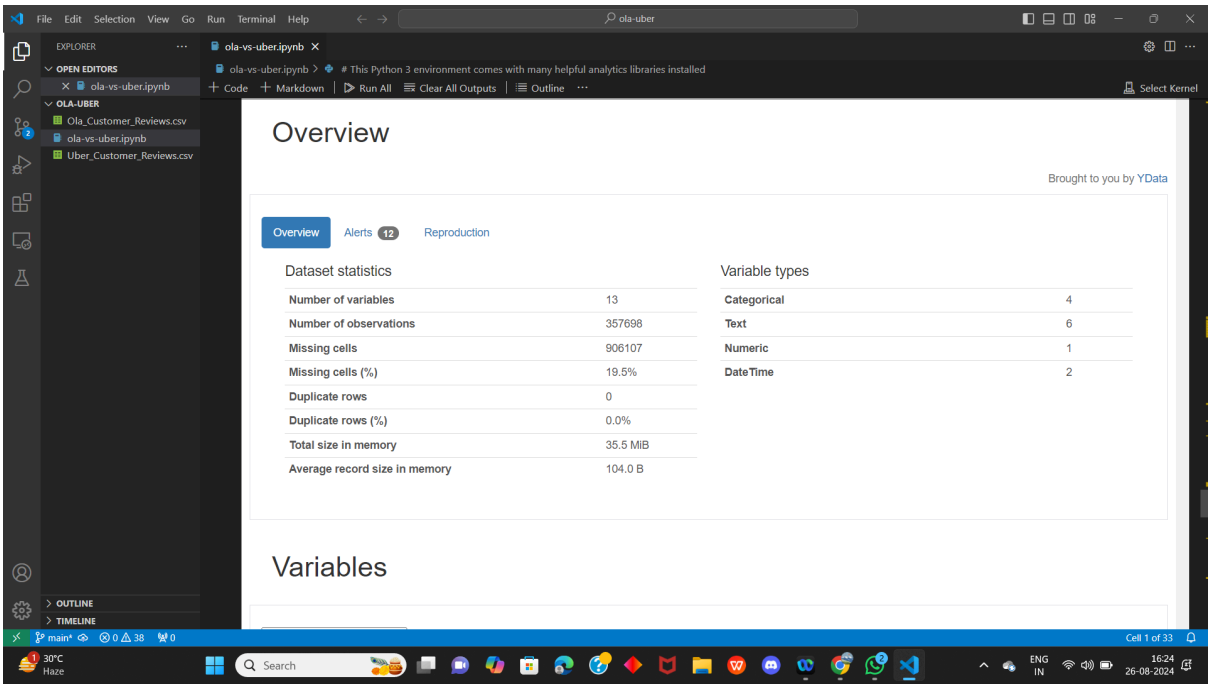
Results

Model Performance
- **Accuracy**: The model achieved an accuracy of approximately 89.75%, indicating a reliable classification of sentiments.
- **Classification Report**:

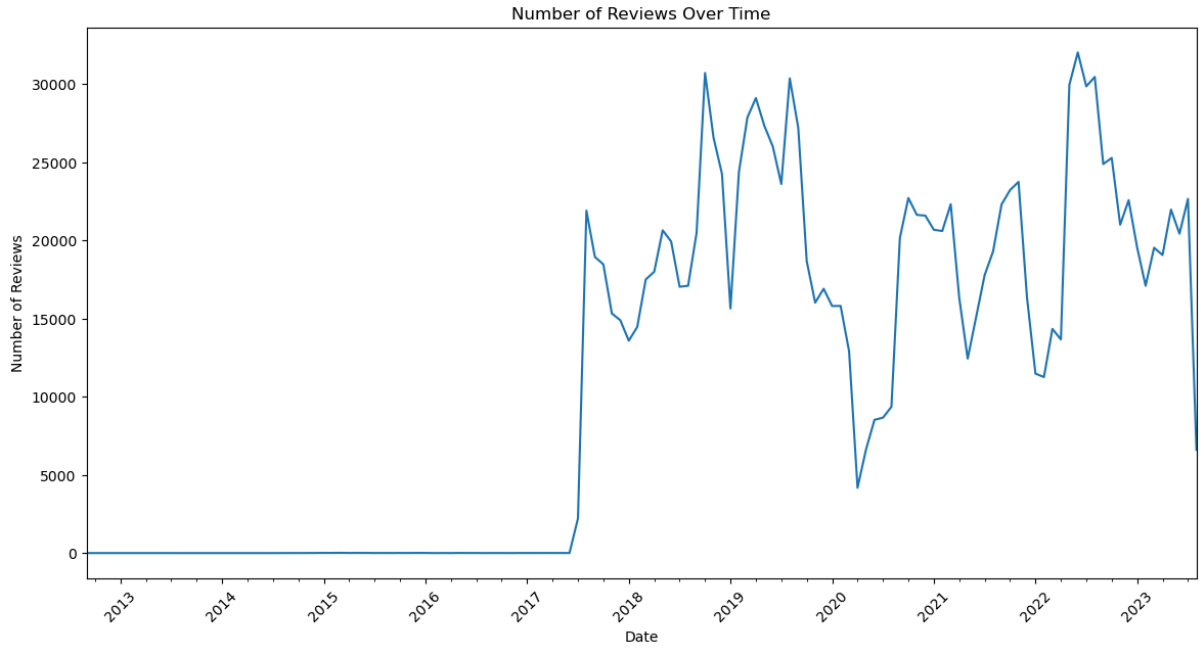
- **Precision**: The model demonstrated a precision of 0.91 for negative reviews and 0.89 for positive reviews.
- **Recall**: The recall rates were 0.92 for negative and 0.87 for positive reviews.
- **F1-Score**: The F1-score, which balances precision and recall, was 0.90 for both classes.

Example Prediction
For a sample review, "This drive was amazing! Bad driver tho", the model predicted a negative sentiment. This demonstrates the model's ability to distinguish between positive and negative feedback based on the review content.

5. Exploratory Data Analysis (EDA)



Overview
A Pandas Profiler report was generated for the Ola dataset, providing a comprehensive analysis of the data. Key findings included:
- **Distribution of Ratings**: The ratings distribution helps identify the general sentiment of the customer base.
- **Frequent Issues**: Common words and phrases from negative reviews highlight frequent complaints and areas for improvement.



6. Conclusions

Customer Sentiments
- **Positive Feedback**: Many reviews highlight satisfaction with services, though specific positive aspects were not analyzed in depth.
- **Negative Feedback**: Common complaints include issues with the app’s functionality and service quality. Addressing these concerns could enhance overall customer satisfaction.

Recommendations
- **Service Improvement**: Focus on addressing frequent issues identified in negative reviews to improve service quality.
- **Feature Enhancement**: Invest in refining app functionalities that customers have reported as problematic.
- **Competitive Strategy**: Use insights from both Ola and Uber reviews to identify gaps in the market and develop strategies to address unmet needs.

Future Work
Further analysis could involve a comparative study between Ola and Uber to identify which aspects of service are more critical to users and how each company performs in those areas.

This report provides a high-level overview of the sentiment analysis and key findings from the customer reviews of Ola and Uber. The insights gained can inform strategic decisions and operational improvements.