Germany Cars For Sale Analysis

Presentation

Pourya Ebrahimi

Intro

Introduction:

- PBS Automobile Handler, a prominent car dealer in Germany with 20 years of experience, is venturing into the online market to tap into the growing online auto dealership sector. To strategize for this new chapter, PBS Automobile Handler aims to analyze the online marketplace landscape, focusing on AutoScout24, a leading platform for buying and selling cars and other vehicles.

Objectives :

- To analyze and segment the automotive market to understand the distribution and characteristics of different vehicle groups.

- To identify patterns and trends within the data that can inform marketing strategies, pricing decisions, and inventory management for automotive businesses.

Key Questions for Analysis:

- Which car brands dominate the market on AutoScout24?

- What is the distribution of the age of cars available on the platform?

- What types of fuel are prevalent among the listed vehicles?

- Is there a correlation between car prices and their colors?

- Does the presence of specific features influence the price of cars?

Data

Dataset:

Germany Used Cars Dataset 2023 from Kaggle.com

Data Cleaning:

The original dataset contains 251,079 rows and 15 columns.
Since the dataset comprises scraped data from the AutoScout24 website, it is evident that there are several areas requiring cleaning and preprocessing.

Missing values:

- *The missing values have been provided in the table .*
  For the fuel consumption which has most missing values ,I tried first to replace missing values with the values of exact same models to reduced the missing values.
- *Same approach has been used for Horse Power.*
- *For the missing values in the mileage I decided to use 0 instead and dropped remaining values.*

Mixed data types:

Moreover like any other scraped dataset we faced a lot of mixed values from other columns which needed be detected and dropped.

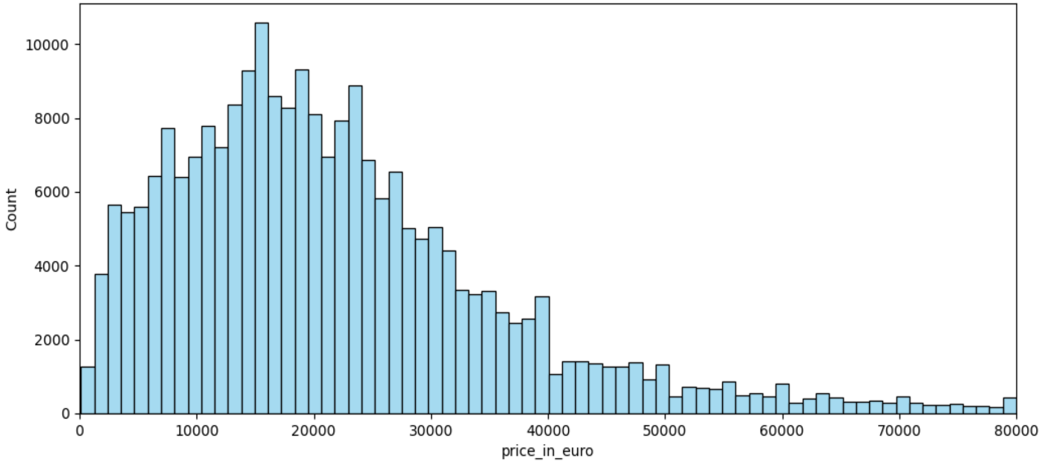
Feature	Missing Values
color	166
registration_date	4
power_kw	134
power_ps	129
fuel_consumption_l_100km	26873
mileage_in_km	152
offer_description	1

Data

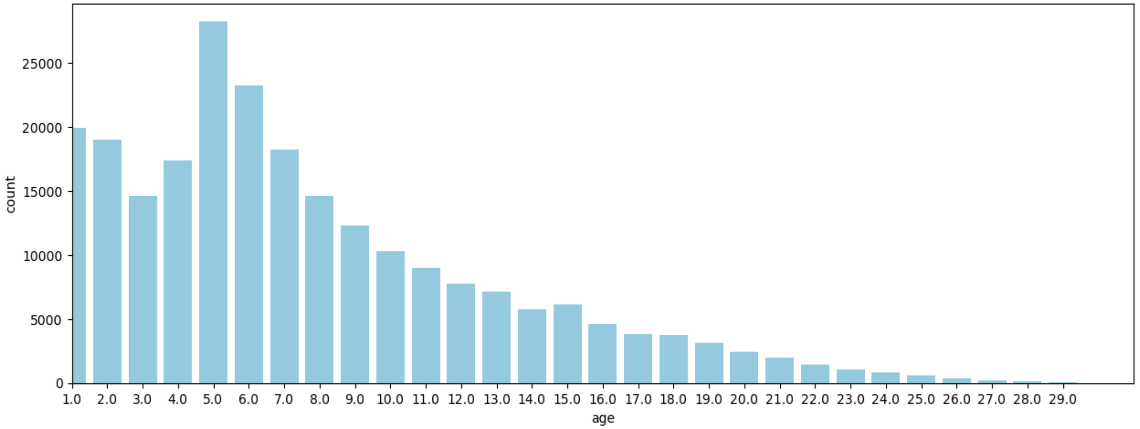
- The following table displays missing values identified after addressing mixed data types.
- Transmission values have been categorized into two types: Automatic and Manual. Semi-automatic values are also considered as Automatic.
- The rare models (less than 10) have been also omitted from the dataset.
- After cleaning process the new dataset has 238,633 rows and 11 columns.

Column	Missing Values	Action
Unnamed: 0	0	Column dropped
brand	0	No action required
model	0	No action required
color	166	dropped
registration_date	4	Column dropped
year	199	dropped
price_in_euro	199	dropped
power_kw	134	Column dropped
power_ps	447	dropped
transmission_type	144	dropped
fuel_type	483	dropped
fuel_consumption_l_100km	9861	dropped
fuel_consumption_g_km	90	Column dropped
mileage_in_km	152	dropped
offer_description	1	Column dropped

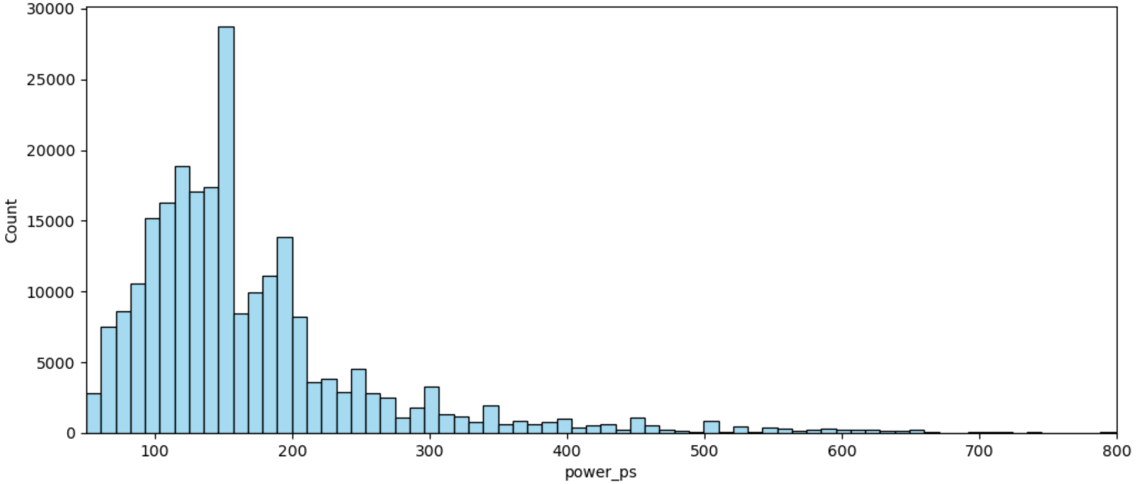
Exploring the Data



Price distribution

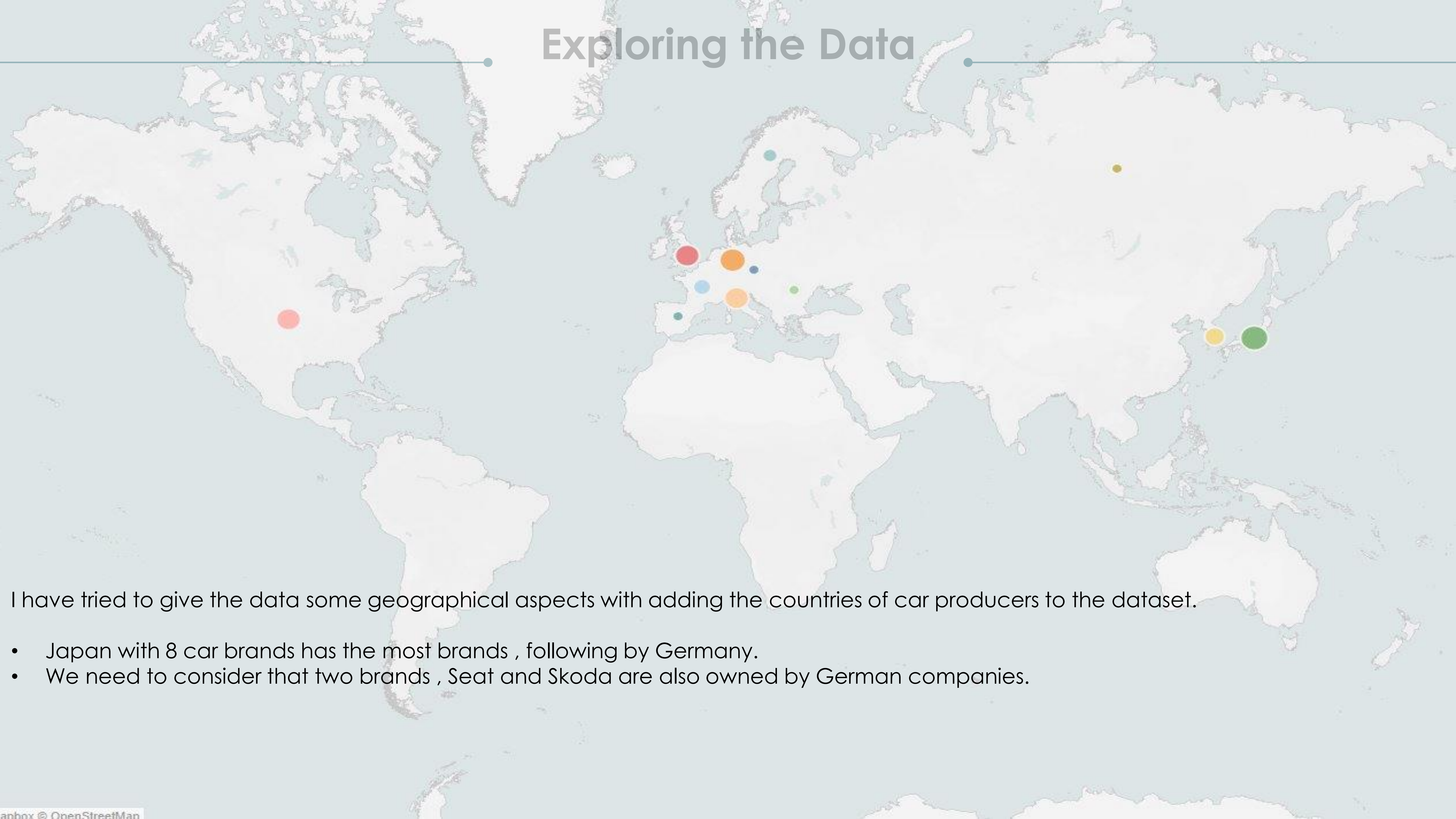


Cars age distribution



Horse Power distribution

Exploring the Data

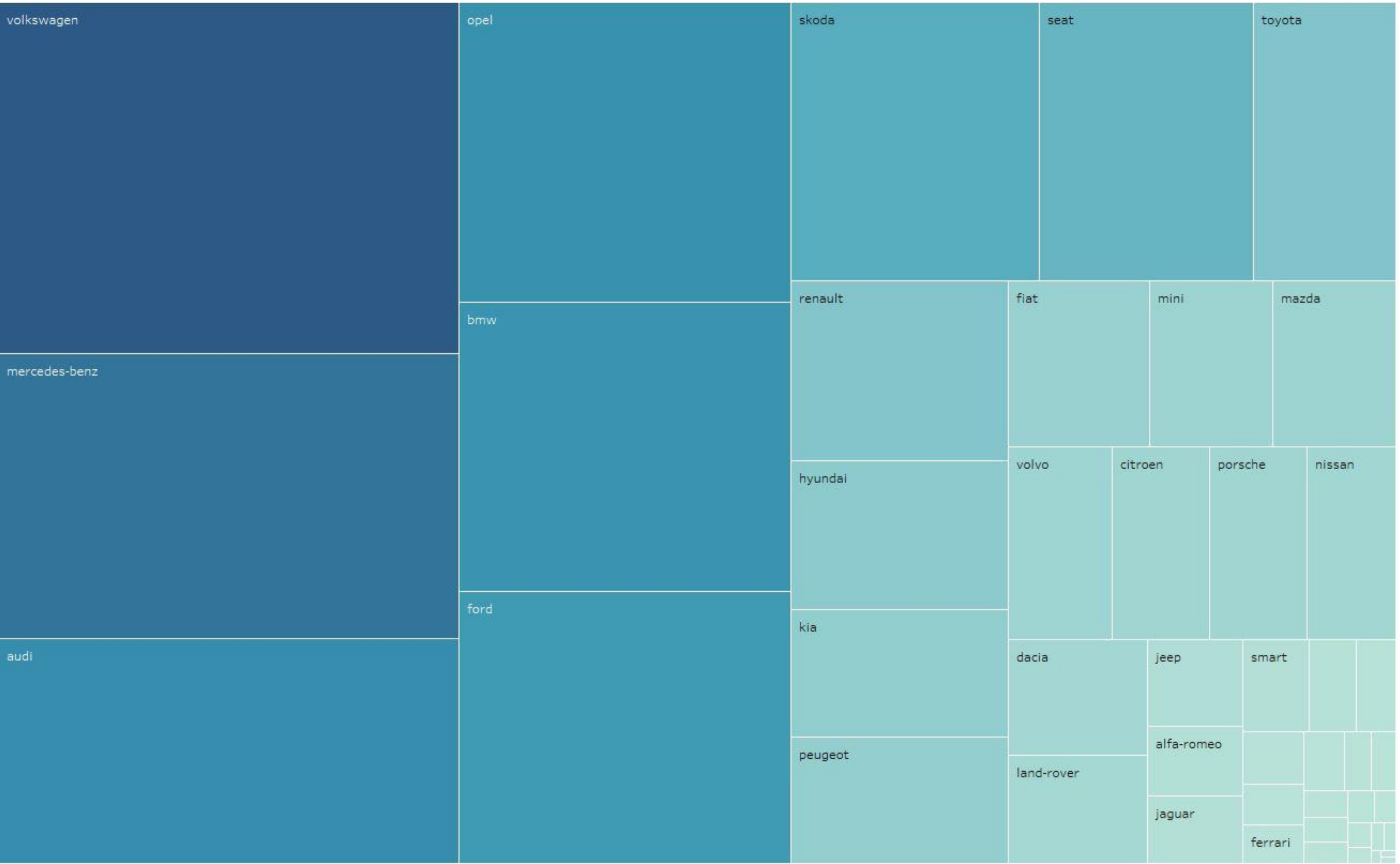


I have tried to give the data some geographical aspects with adding the countries of car producers to the dataset.

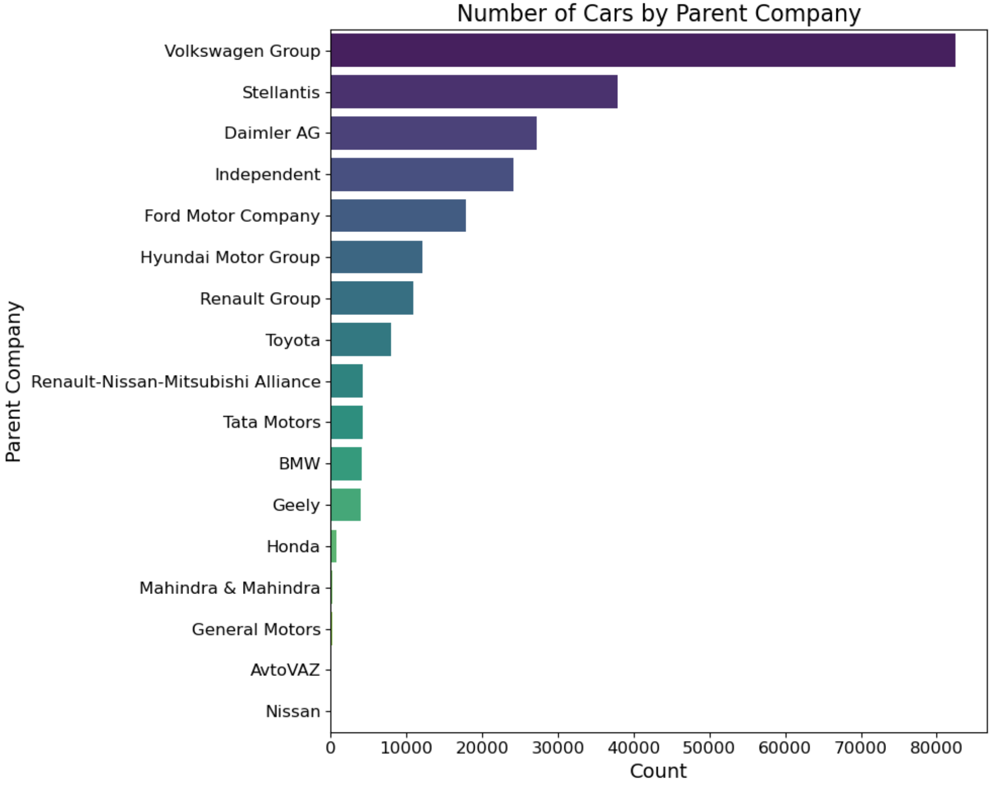
- Japan with 8 car brands has the most brands , following by Germany.
- We need to consider that two brands , Seat and Skoda are also owned by German companies.

Exploring the Data

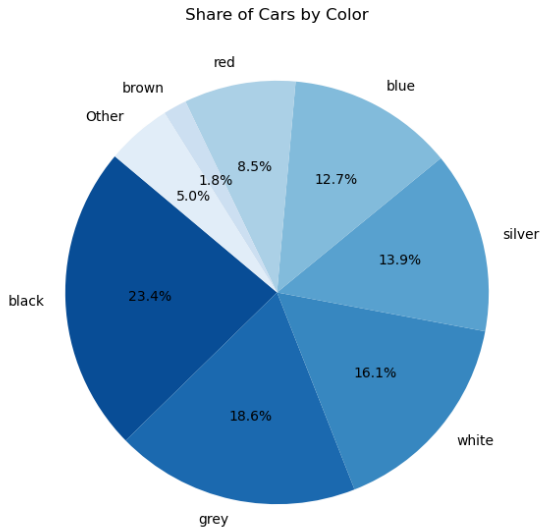
The brands presence

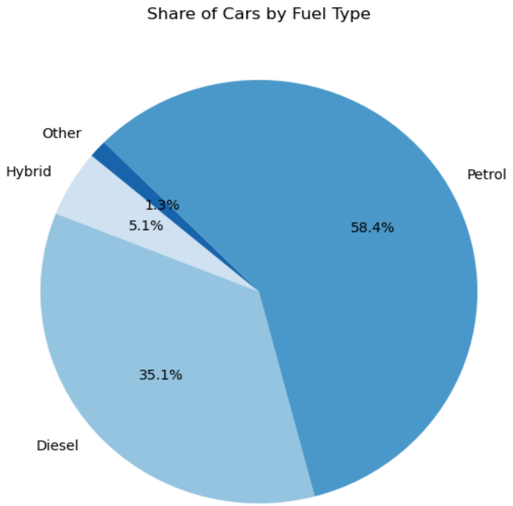


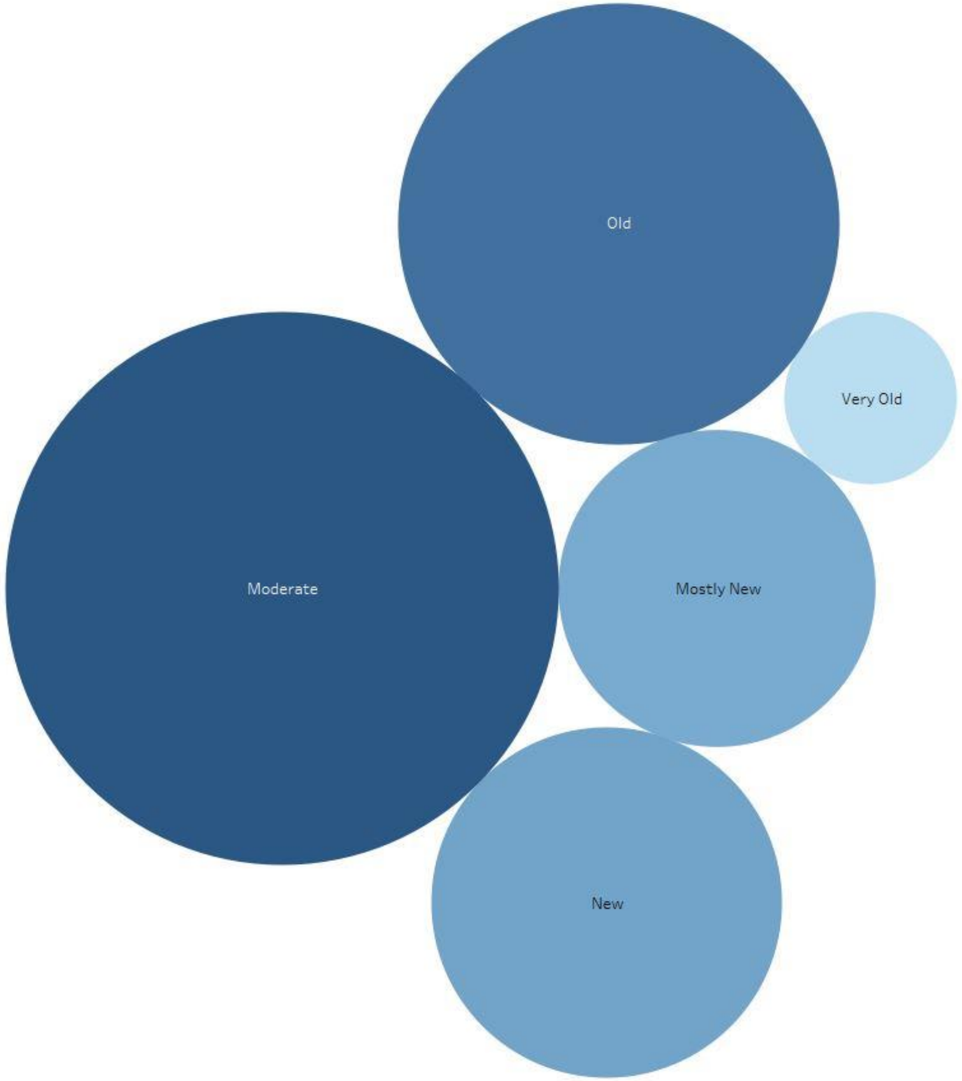
- I also included the parent company of the brands. It is evident that the Volkswagen Group holds a relatively dominant position.



Exploring the Data







The Age of the cars

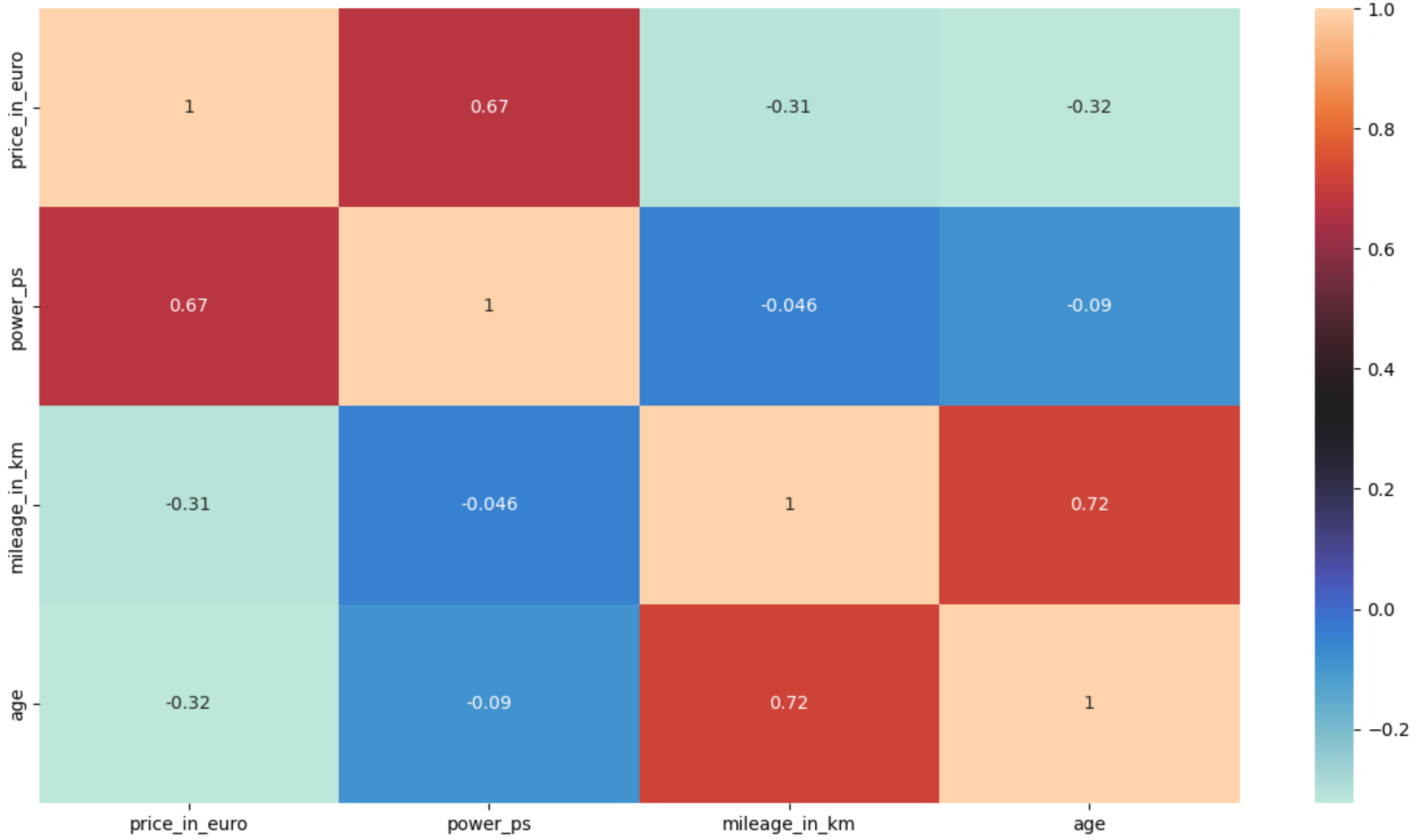
Exploring the Data

Correlation:

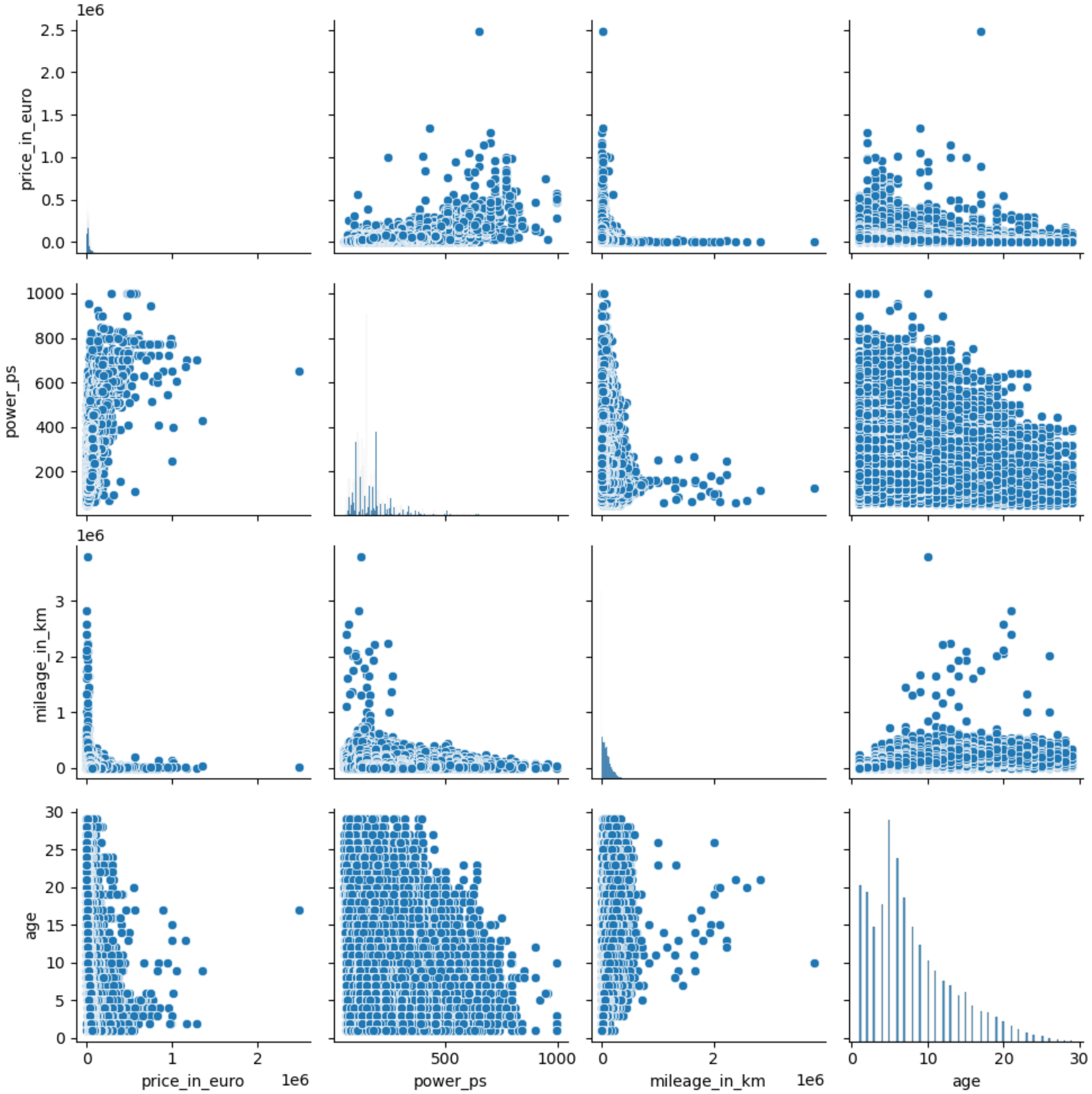
- The strongest correlation is between mileage and age of the car .
- The second strongest is relation is between price and Horse Power.

Hypothesis:

- If the car has more power it would be more expensive.



Analysis



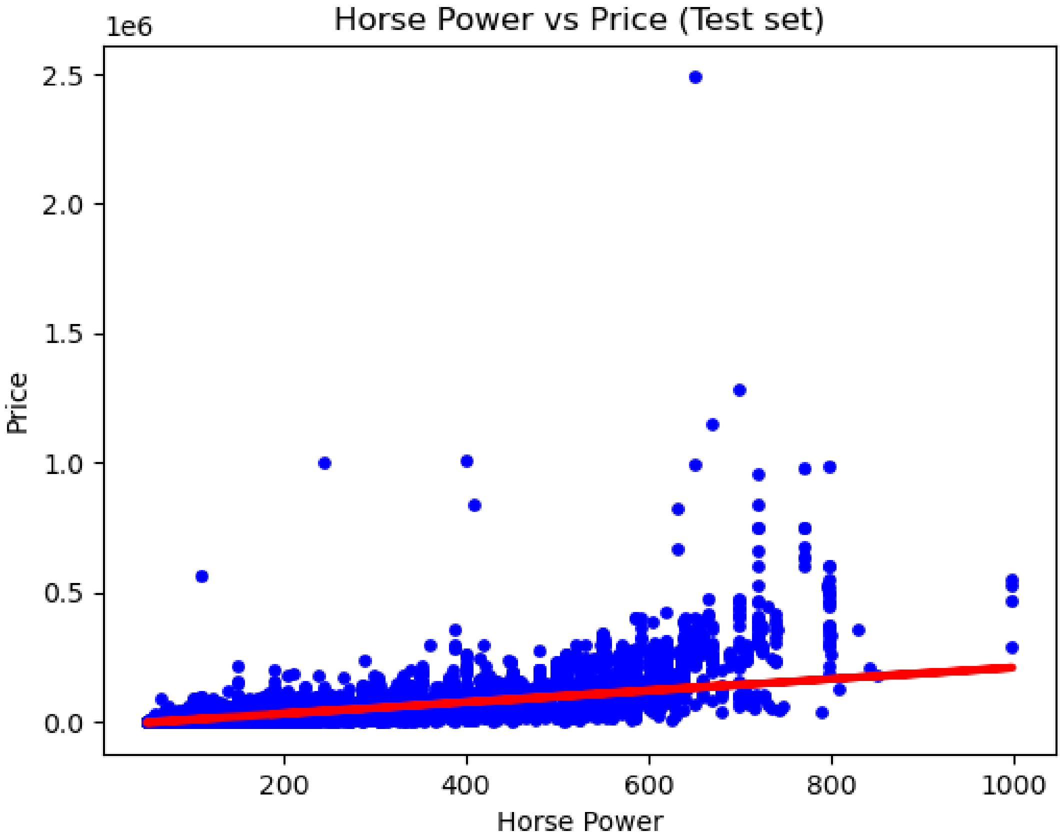
- In this pair plot we see the relations between different values.

- The scatterplot also supports our hypothesis.

Analysis

Regression Analysis:

- The outcome of the regression analysis does not support our hypothesis, indicating a significant discrepancy between the predicted and actual values of the price. This suggests that additional factors may be influencing the results, necessitating a more detailed analysis to identify these components.



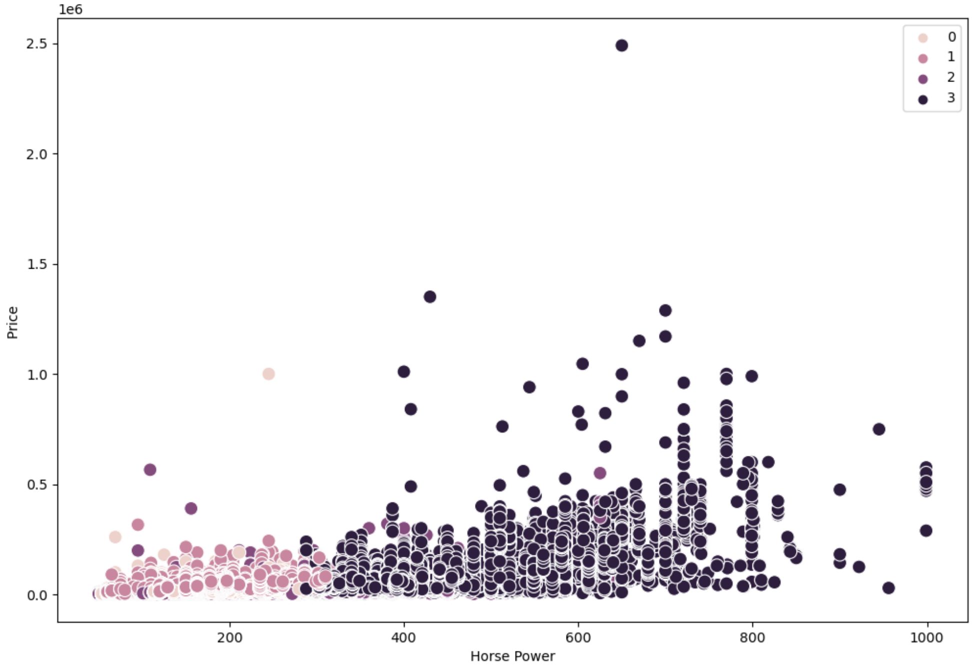
Analysis

Cluster 3
Count: 73,200 cars
Price: Moderate (mean: €16,037.79, median: €14,486.50)
Age: Older (mean: 9.43 years, median: 9 years)
Power: Average (mean: 148.71 PS, median: 140 PS)
Mileage: High (mean: 117,949.43 km, median: 106,500 km)
This cluster likely represents older, moderately priced cars with high mileage, indicating they are well-used but maintained an average power level. The moderate price points to a market for used, reliable vehicles.

Cluster 2
Count: 110,108 cars
Price: Higher (mean: €27,930.56, median: €24,990)
Age: Newer (mean: 3.66 years, median: 4 years)
Power: Average to slightly below average (mean: 149.70 PS, median: 150 PS)
Mileage: Low (mean: 38,050.73 km, median: 28,500 km)
This cluster appears to consist of relatively new, higher-priced cars with lower mileage, suggesting they are likely newer models that have retained much of their value and have not been heavily used.

Cluster 1 : Count: 36,472 cars
Price: Lower (mean: €8,069.62, median: €5,500)
Age: Very Old (mean: 17.70 years, median: 17 years)
Power: Average (mean: 146.25 PS, median: 136 PS)
Mileage: Very High (mean: 180,408.40 km, median: 173,000 km)
Vehicles in this cluster are characterized by their very old age, low price, and very high mileage. These are likely economy or older vehicles that have been significantly depreciated over time.

Cluster 0 :Count: 18,412 cars
Price: Very High (mean: €83,153.54, median: €56,870)
Age: Newer (mean: 5.74 years, median: 6 years)
Power: High (mean: 420.02 PS, median: 392 PS)
Mileage: Moderate (mean: 60,064.72 km, median: 50,000 km)
This cluster is distinct for its very high price and high power vehicles, which are relatively new and have moderate mileage. This group likely includes luxury, high-performance, or premium cars that are sought after for their brand, performance, and features rather than practicality.



Recommendations

Recommendations:

- Targeted Marketing Strategies: Based on the clustering analysis, marketing efforts can be more effectively directed. For example, vehicles in Cluster 2 - newer, higher-priced cars with low mileage - can be marketed to consumers looking for almost new vehicles without the brand-new price tag. Conversely, vehicles in Cluster 1 offer opportunities for targeting budget-conscious consumers interested in older models.

- Inventory Management: Dealerships could adjust their inventory to match the profiles of the most common clusters in their region. For instance, if Cluster 3 vehicles are prevalent in the market, stocking up on older, moderately priced cars with reasonable power and high mileage might meet consumer demand more effectively.

- Pricing Strategy: Insights from the clusters can help in setting competitive prices. Vehicles in Cluster 0, being very high-priced and high-powered, might allow for a premium pricing strategy, while those in Cluster 1 might require more competitive pricing to attract buyers looking for economical options.

Future Directions

- Data Enrichment: Incorporating additional features, such as vehicle condition, or more detailed model information, could provide a richer basis for segmentation.

- Temporal Analysis: Conducting this analysis periodically could help in understanding market dynamics and trends over time, offering insights into how consumer preferences evolve.

- Validation with External Data: Cross-referencing the findings with external market data, customer surveys, or sales data could help validate and refine the analysis.
- In conclusion, while this analysis provides valuable insights into the automotive market, it should be viewed as a starting point for deeper exploration. Future efforts should aim to address the limitations mentioned, leveraging a broader data set and incorporating more dynamic market factors to refine and validate the findings.

Limitations

DATA QUALITY

Data Quality and Completeness: As the analysis is based on scraped data, it's subject to the limitations inherent in such data, including potential inaccuracies, missing values, and biases in the data collection process. The insights derived should be seen as indicative rather than definitive.

.

Generalizability

The findings are specific to the dataset and the time period from which it was collected. While the clustering provides useful insights, they may not be universally applicable across different geographic regions or market segments.

Thank You

Connections