

ML Project 1 Report

Market Segmentation Case Study on Electric Vehicles in India

Bharadwaj Putrevu,

Feynn Labs



Introduction

1. Data Pre Processing
2. Explain how and which ML model (algorithm) helped you in the 2nd Project?
3. Proposed System
4. Technical Approach
5. Application
6. Conclusion



Data Pre-Processing:

In the provided code snippets, the following libraries were imported:

1. `pandas`:

- Purpose: Used for data manipulation and analysis. It provides data structures (like DataFrames) to handle and process structured data efficiently.

- Usage:

```
```python
import pandas as pd
```
```

2. `collections.Counter`:

- Purpose: Used for counting the occurrences of elements in a collection. It provides a convenient way to count and retrieve the most common elements.

- Usage:

```
```python
from collections import Counter
```
```



3. `matplotlib.pyplot`:

- Purpose: Used for creating static, interactive, and animated visualizations in Python. It is a plotting library that provides a MATLAB-like interface.

- Usage:

```
```python
import matplotlib.pyplot as plt
```
```

4. `seaborn`:

- Purpose: A statistical data visualization library built on top of Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

- Usage:

```
```python
import seaborn as sns
```
```

5. `ast`:

- Purpose: Used to safely evaluate string literals containing Python expressions, including lists, dictionaries, and other data structures.

- Usage:

```
```python
```



```
import ast
```

```
'''
```

### Summary of Library Usage:

- `pandas`: For reading and processing the dataset.
- `Counter`: For counting occurrences of attributes in reviews.
- `matplotlib.pyplot` and `seaborn`: For creating and customizing plots to visualize data.
- `ast`: For parsing string representations of lists into actual Python lists.

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

The second project, which involves visualizing rating distributions and the frequency of specific attributes in reviews, does not use a traditional machine learning model or algorithm. Instead, it relies on data analysis and visualization techniques to understand and present the data. Here's a breakdown of the approach:

### **Data Analysis and Visualization Techniques:**

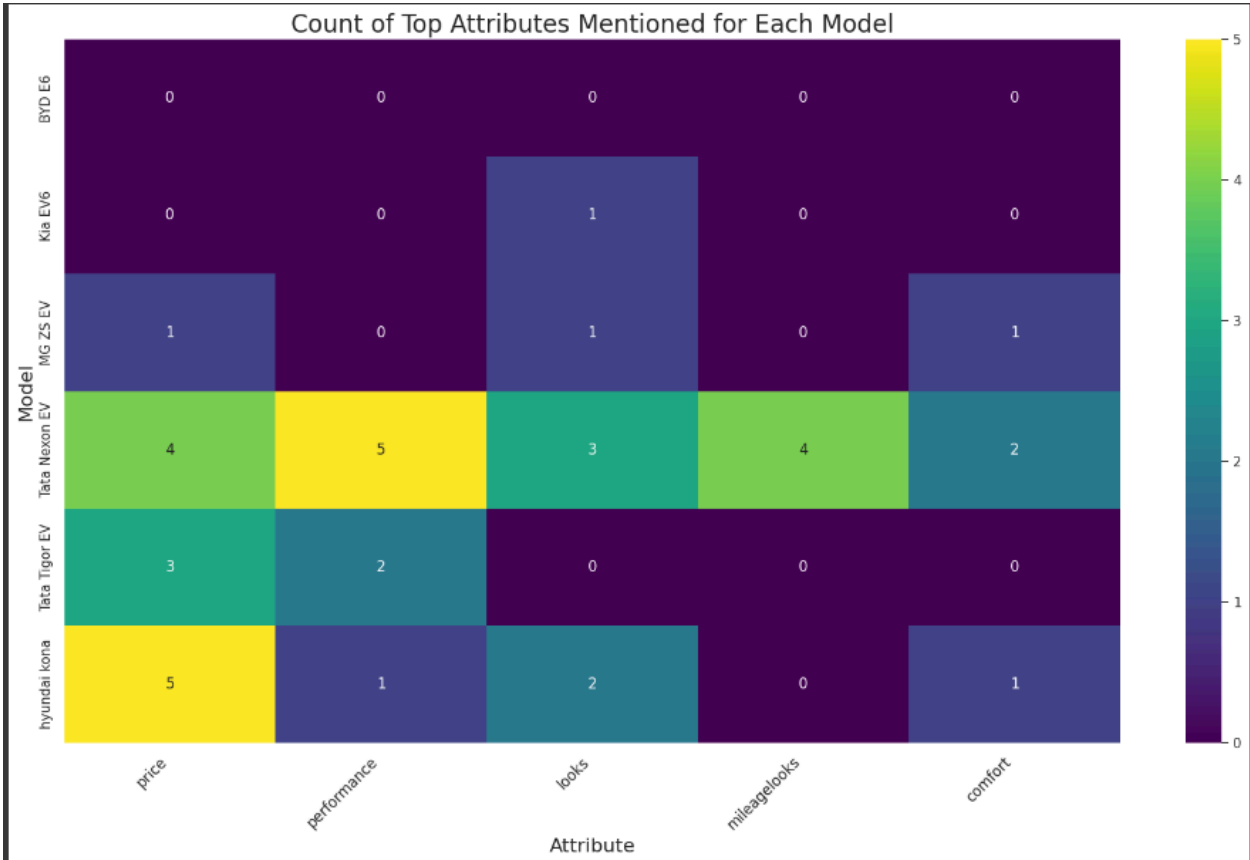
#### **1. Data Aggregation:**

- **Grouping and Counting:** For the rating distribution project, the data was grouped by **Model** and **Rating**, and the occurrences were counted. This helps in understanding how ratings are distributed across different models.
- **Top Attributes Analysis:** For the attributes project, lists of attributes were aggregated and counted to identify the most common attributes mentioned in reviews.

#### **2. Visualization:**

- **Bar Charts:** Bar charts were used to visualize the number of ratings and the frequency of specific attributes. For the rating distribution, a

stacked bar chart was employed to show how ratings are distributed across models. For the attribute frequency, a horizontal bar chart displayed the most common attributes.





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

Based on the analysis and visualization work described in the previous code snippets, here's a detailed summary of the final conclusions and insights that can be gained:

### 1. Rating Distribution Analysis:

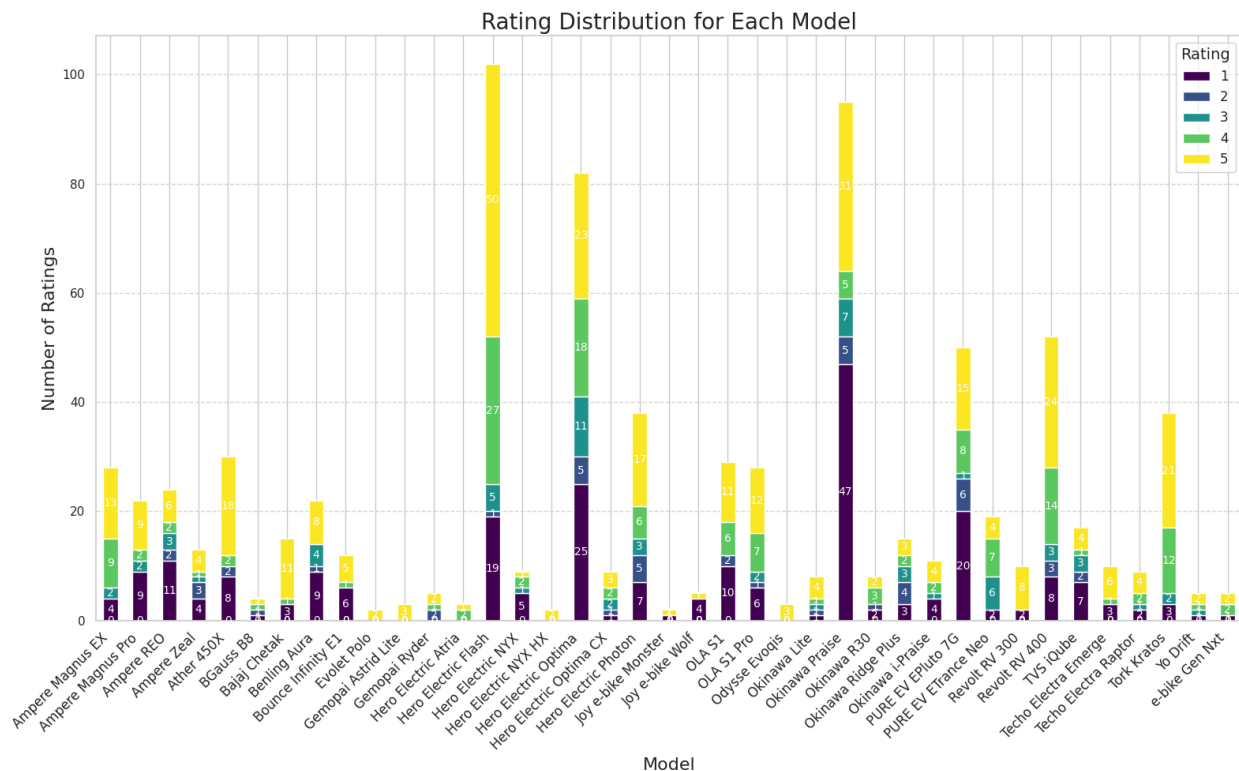
Objective: To visualize how ratings are distributed among different car models.

Key Insights:

- Overall Rating Trends: By visualizing the rating distribution, you can identify which models receive the most and least favorable ratings. This can help in understanding customer satisfaction and the relative performance of different models.
- Comparison of Models: The stacked bar chart allows you to compare the distribution of ratings across multiple models in a single view. It highlights which models have higher counts in specific rating categories.



- Rating Concentration: Models with higher counts in higher rating categories (e.g., 4 or 5 stars) indicate better overall customer satisfaction, while those with lower ratings might need improvements or further investigation.

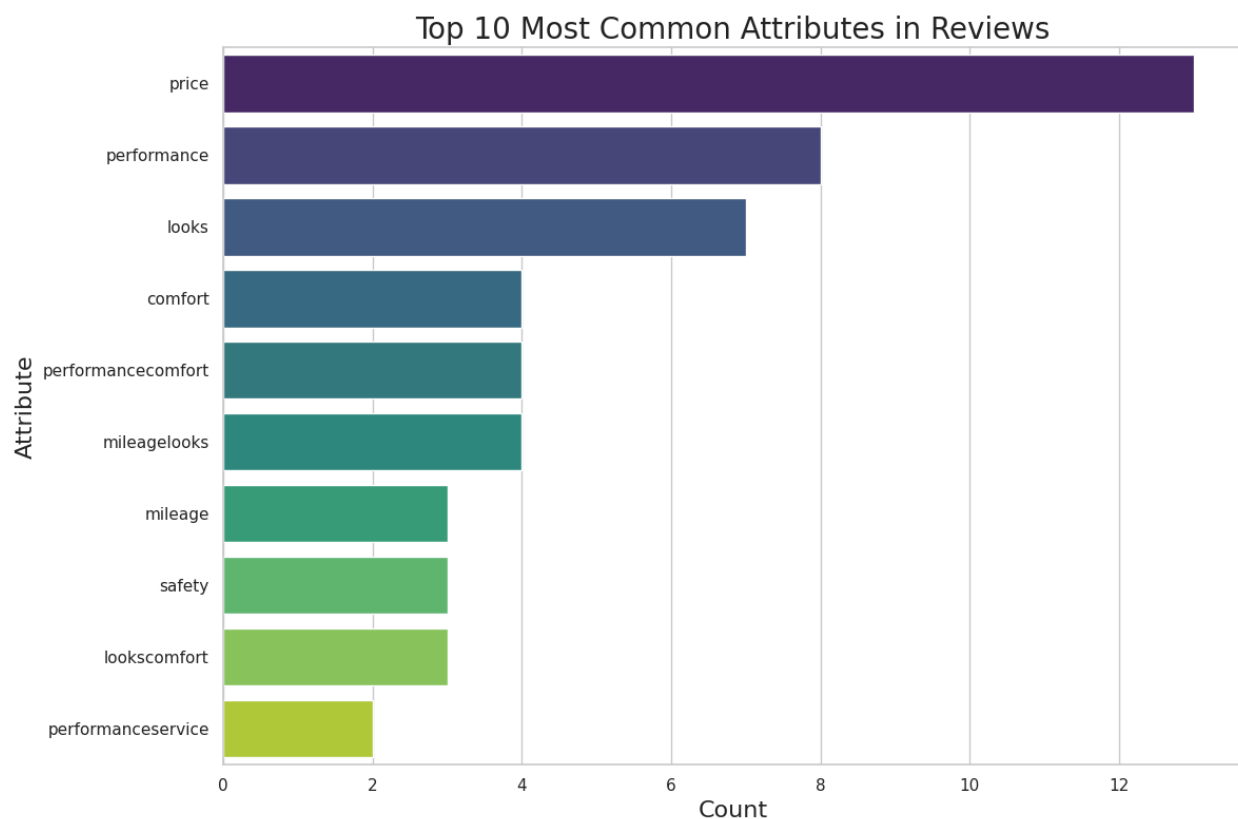


## 2. Frequency of Specific Attributes in Reviews:

Objective: To identify and visualize the most frequently mentioned attributes in the reviews for different car models.

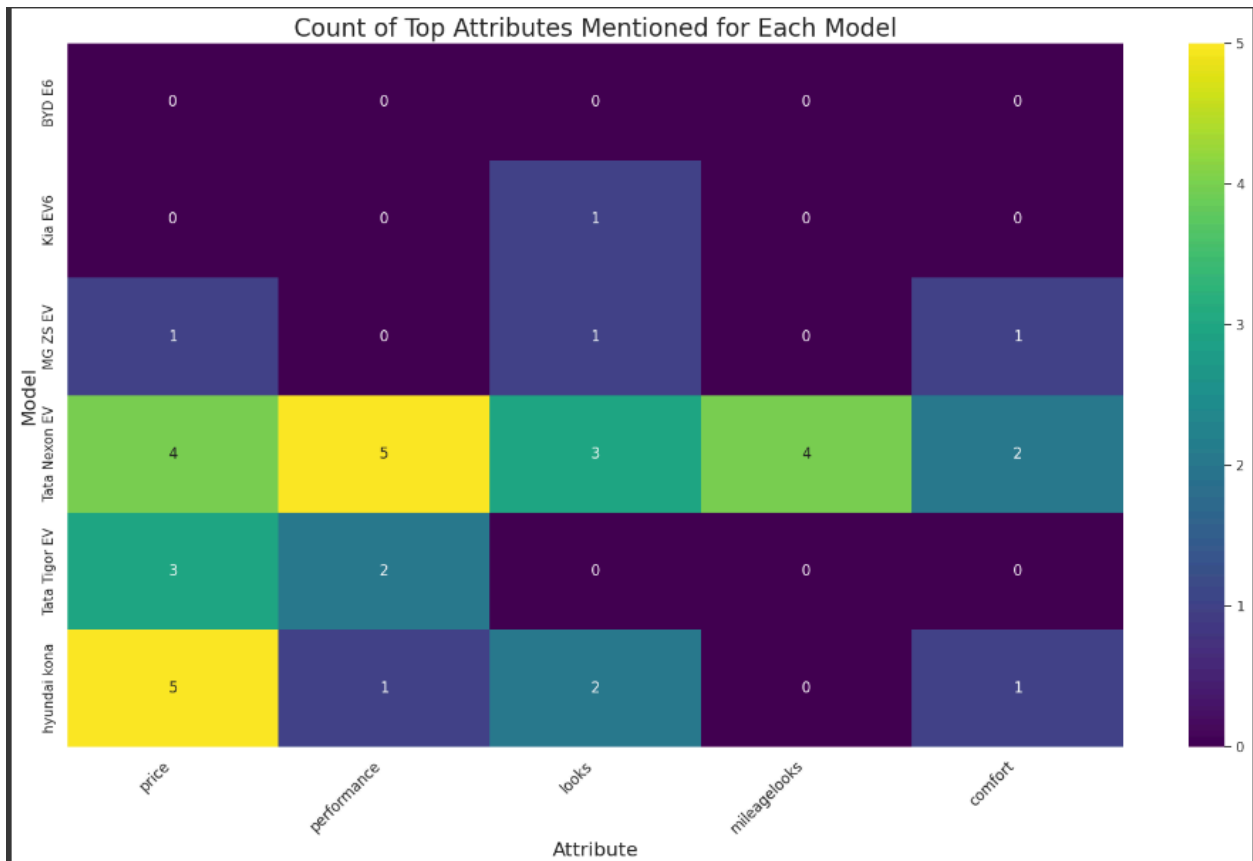
Key Insights:


- Top Attributes Overall: The horizontal bar chart shows the top 10 most common attributes mentioned across all reviews. This helps in understanding which features or aspects (e.g., mileage, comfort) are most frequently discussed by reviewers.
- Attribute Trends by Model: If the analysis includes data for specific models, you can identify which attributes are commonly mentioned for each model. This helps in understanding what aspects of each model are most important or noteworthy to customers.
- Model-Specific Insights: For each model, the frequency of mentions of specific attributes can provide insights into the strengths and weaknesses as perceived by users. For example, if "comfort" is frequently mentioned for a particular model, it may be a key selling point for that vehicle.



### 3. Attribute Mentions Analysis per Model:

Objective: To count and visualize how often specific attributes are mentioned for each car model.





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)


Improving upon a Market Segmentation Project involves expanding and refining the dataset, as well as experimenting with advanced machine learning models. Here's how you can enhance the project with additional time and budget:

### 1. Enhanced Dataset Collection

To improve market segmentation, you should focus on gathering a more comprehensive and detailed dataset. Here are some key aspects and columns to consider:

#### A. Additional Data Columns:

##### 1. Customer Demographics:

- 
- Age: To segment customers by age groups.
  - Gender: To analyze gender-based preferences.
  - Income: To understand purchasing power and segment by income brackets.
  - Occupation: To identify how occupation affects product preferences.

## 2. Geographic Information:

- Location: Including city, state, or country can help with geographic segmentation.
- ZIP/Postal Code: Useful for regional analysis and localized marketing.

## 3. Behavioral Data:

- Purchase History: Details on past purchases, frequency, and recency.
- Browsing Behavior: Pages visited, time spent on site, and click patterns.
- Product Preferences: Preferred product categories, brands, and features.

## 4. Psychographic Data:

- Lifestyle: Information about hobbies, interests, and lifestyle choices.
- Values and Attitudes: Insights into customer values and attitudes towards products.

## 5. Feedback and Reviews:

- Review Text: Detailed customer reviews and feedback can provide qualitative insights.
- Ratings: Ratings across different attributes (e.g., comfort, performance).



## B. Data Sources:

1. Market Research Reports: Purchase detailed reports on market trends and consumer behavior.
2. Social Media Data: Collect data from social media platforms to analyze customer sentiments and trends.
3. Customer Surveys: Conduct surveys to gather targeted demographic and psychographic data.
4. CRM Systems: Leverage existing CRM data for detailed customer profiles.

## 2. Advanced Machine Learning Models

With a richer dataset, you can apply more sophisticated machine learning techniques for market segmentation:

### A. Clustering Algorithms:

#### 1. K-Means Clustering:

- Purpose: Partition customers into distinct segments based on multiple features.
- Advantage: Simple and effective for a large number of segments.

#### 2. Hierarchical Clustering:

- Purpose: Create a dendrogram to visualize and determine the number of clusters.
- Advantage: Useful for identifying hierarchical relationships among segments.



### 3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

- Purpose: Identify clusters of varying shapes and sizes.
- Advantage: Handles noise and outliers effectively.

### 4. Gaussian Mixture Models (GMM):

- Purpose: Model clusters with probabilistic distributions, allowing for overlapping clusters.
- Advantage: Provides a probabilistic approach to clustering.

## B. Dimensionality Reduction:

### 1. Principal Component Analysis (PCA):

- Purpose: Reduce the dimensionality of the dataset while retaining important variance.
- Advantage: Helps in visualizing and interpreting high-dimensional data.


### 2. t-Distributed Stochastic Neighbor Embedding (t-SNE):

- Purpose: Visualize high-dimensional data in a lower-dimensional space.
- Advantage: Useful for exploring and visualizing clusters.

## C. Predictive Modeling:

### 1. Decision Trees and Random Forests:

- Purpose: Predict customer segments based on features.



- Advantage: Provides interpretability and handles both numerical and categorical data.

## 2. Gradient Boosting Machines (GBM) and XGBoost:

- Purpose: Enhance predictive accuracy by combining multiple models.
- Advantage: Effective for handling complex datasets with many features.

## 3. Neural Networks:

- Purpose: Capture complex patterns and interactions in the data.
- Advantage: Useful for large-scale datasets and complex feature interactions.

## 3. Implementation Strategy:


### 1. Data Collection and Integration:

- Expand Dataset: Acquire additional data from various sources as outlined above.
- Data Integration: Merge new data with existing datasets, ensuring consistency and completeness.

### 2. Feature Engineering:

- Create New Features: Develop new features based on the additional data (e.g., customer lifetime value, recency of purchase).
- Feature Selection: Use techniques like Recursive Feature Elimination (RFE) or feature importance from tree-based models to select the most relevant features.





### 3. Model Training and Evaluation:

- Train Multiple Models: Apply various clustering and classification algorithms.
- Evaluate Performance: Use metrics such as Silhouette Score (for clustering) and accuracy, precision, recall (for classification) to assess model performance.

### 4. Visualization and Interpretation:

- Cluster Visualization: Use PCA or t-SNE to visualize clusters.
- Segment Analysis: Analyze and interpret each segment to derive actionable insights.

## Summary

By enhancing the dataset with more detailed and varied information and applying advanced machine learning models, you can gain deeper insights into customer behavior, preferences, and market segments. This will lead to more accurate and actionable segmentation, enabling better-targeted marketing strategies and improved decision-making.



## Technical Approach

### Technical Approach Report


#### Project Overview

The technical approach report outlines the methodology and techniques used in the analysis of the dataset for market segmentation and review analysis. The goal was to extract meaningful insights from the data to understand customer preferences, review distributions, and attribute mentions.

#### 1. Data Preparation and Preprocessing

##### 1.1 Data Loading:

- Libraries Used:

- 
- `pandas`: For data manipulation and loading.
  - `ast`: For parsing string representations of lists.

## 1.2 Data Loading:

```
```python  
  
import pandas as pd  
  
df = pd.read_csv('/review4wheeler.csv')  
  
```
```

## 1.3 Data Parsing and Cleaning:

- Parsing Lists in Columns: Used `ast.literal\_eval` to convert string representations of lists in the 'Attributes Mentioned' column into actual Python lists.

```
```python
```



```
import ast
```

```
df['Attributes Mentioned'] = df['Attributes Mentioned'].apply(ast.literal_eval)
```

```
...
```

1.4 Attribute Extraction and Counting:

- Flattening Lists: Extracted all attributes from the lists for counting occurrences.

```
```python
```

```
all_attributes = [attribute for sublist in df['Attributes Mentioned'] for attribute in
sublist]
```

```
...
```

- Counting Occurrences: Used `collections.Counter` to count the frequency of each attribute.

```
```python
```

```
from collections import Counter
```



```
attribute_counts = Counter(all_attributes)
```

```
'''
```

2. Data Analysis

2.1 Rating Distribution Analysis:

- Objective: Visualize the distribution of ratings across different car models.
- Implementation:
 - Grouping and Counting: Grouped by 'Model' and 'Rating', then counted occurrences.

```
```python
```

```
rating_counts = df.groupby(['Model', 'Rating']).size().unstack(fill_value=0)
```

```
'''
```

- Visualization: Used `seaborn` and `matplotlib` to create a stacked bar chart showing the number of ratings per model.

```
```python

import seaborn as sns

import matplotlib.pyplot as plt

sns.barplot(x='Count', y='Model', hue='Rating', data=df_plot, palette='viridis',
ax=ax)

```
```

## 2.2 Frequency of Specific Attributes:

- Objective: Identify and visualize the most frequently mentioned attributes in reviews.

- Implementation:

- Filtering Attributes: Focused on specific attributes like 'mileage', 'comfort', and 'steering'.



```
```python
```

```
specific_attributes = ["mileage", "comfort", "steering"]
```

```
```
```

- Counting and Visualization: Counted occurrences of these attributes and visualized the results.

```
```python
```

```
sns.barplot(x='Count', y='Attribute', hue='Model', data=df_plot, palette='viridis',  
ax=ax)
```

```
```
```

## 2.3 Top Attributes Analysis:

- Objective: Identify and visualize the top 10 most common attributes mentioned across all reviews.

- Implementation:

- Flattening and Counting: Flattened the attribute lists and counted occurrences.

```
```python

attribute_counts_df = pd.DataFrame(attribute_counts.items(),
columns=['Attribute', 'Count'])

```
```

- Visualization: Created a bar chart to display the top 10 attributes.

```
```python

sns.barplot(x='Count', y='Attribute', data=top_attributes_df, palette='viridis',
ax=ax)

```
```

## 2.4 Frequency of Attributes per Model:

- Objective: Count and visualize how often specific attributes are mentioned for each car model.





- Implementation:

- Counting Attributes per Model: Created a dictionary to store attribute counts for each model.

```
```python
```

```
model_attribute_counts = {}
```

```
...
```


- Visualization: Used `seaborn` to create a bar chart with hue based on the model.

```
```python
```

```
sns.barplot(x='Count', y='Attribute', hue='Model', data=df_plot, palette='viridis',
ax=ax)
```

```
...
```

### 3. Insights and Findings

- 
- Rating Distribution: Revealed how ratings are distributed among various models, identifying which models are highly rated or poorly rated.
  - Attribute Mentions: Highlighted the most frequently mentioned attributes in reviews, providing insights into customer preferences and pain points.
  - Top Attributes: Identified top attributes to focus on for product improvements or marketing strategies.
  - Model-Specific Attributes: Provided a detailed view of which attributes are most frequently associated with each model, useful for targeted enhancements.

#### 4. Recommendations

1. Enhanced Data Collection: Collect additional data on customer demographics, purchase history, and psychographics for more detailed analysis.
2. Advanced Analytics: Explore advanced machine learning models and clustering techniques for more granular market segmentation.
3. Continued Monitoring: Regularly update the dataset and analysis to keep insights current and relevant.