Electric Vehicle market in India using Segmentation analysis

Sayali Hatwar

Abstract

India's electric vehicle (EV) market is witnessing rapid growth, driven by government policies, rising fuel prices, and increasing environmental awareness. Key initiatives like the Faster Adoption and Manufacturing of Hybrid and Electric Vehicles (FAME) scheme aim to boost the adoption of EVs. Major automakers are entering the market, and the infrastructure for EVs, including charging stations, is gradually improving. However, challenges such as high initial costs, limited range, and inadequate charging infrastructure persist.

By effectively segmenting the market, EV companies in India can better meet the diverse needs of consumers, though they must balance the benefits with the inherent challenges to maximize market potential.

Analysing customer reviews can provide valuable insights into the behaviour of the Indian market concerning electric cars.

Attributes used in the project

To analyse the Indian market's behaviour regarding electric cars, we used several key attributes from customer reviews. These include overall review text, ratings for exterior design, comfort, performance, and fuel economy (related to battery efficiency and range). We also considered value for money, the vehicle's condition (new or used), distance is driven, overall rating, and the specific model's name. These attributes provide comprehensive insights into customer satisfaction and key areas for improvement and focus in the electric vehicle market in India.

Machine Learning Models used in the project

Sentiment Analysis, K-means clustering, and Principal Component Analysis (PCA) are powerful techniques for analysing and extracting insights from various attributes, particularly when dealing with behavioural data. Here is how each technique can be applied:

1. **Sentiment Analysis:**
   This technique is employed to determine the overall sentiment of customer reviews, categorizing them as positive, negative, or neutral. By analysing the text data, sentiment analysis helps in understanding customer emotions and attitudes towards specific features of electric vehicles.

2. **K-Means Clustering:**
   K-Means Clustering is used to segment customer reviews into distinct groups based on similarities in their feedback. This clustering technique enables the identification of different customer segments with common preferences and behaviours, facilitating targeted marketing strategies and personalized offerings.

3. **Principal Component Analysis (PCA):**
   Principal Component Analysis (PCA) transforms a multivariate dataset into a new set of uncorrelated variables called principal components, ordered by the amount of variance they capture. The first principal component holds the most variability, followed by the second, and so on. PCA retains the original data structure but views it from a different perspective. It uses the covariance or correlation matrix, depending on the scale and range of the variables. PCA is often used to reduce the dimensionality of data for visualization, typically focusing on the first few principal components that capture the most variation.

Elaboration on conclusion and insights gained from analysis

- We conducted our analysis using a dataset consisting of customer reviews for four-wheeler electric cars. These reviews contain detailed information about customer experiences and the facilities provided by electric cars.

- The electric vehicle market is growing rapidly but is still in its early stages. Attributes from customer reviews can help segment the market and understand different customer perspectives. By analysing these attributes, businesses can develop targeted marketing and sales strategies that address the specific needs of each segment, thus enhancing customer satisfaction and driving further growth in the industry.

- To ensure the best precision of our machine learning model, we meticulously pre-processed the data and conducted Exploratory Data Analysis (EDA). Preprocessing involved cleaning the data, handling missing values, and standardizing features as needed. During EDA, we explored the dataset to gain insights into its structure, distributions, and relationships between variables. These steps were crucial for understanding the data better and preparing it for modelling, ultimately aiming to achieve optimal performance and accuracy in our machine-learning endeavours.

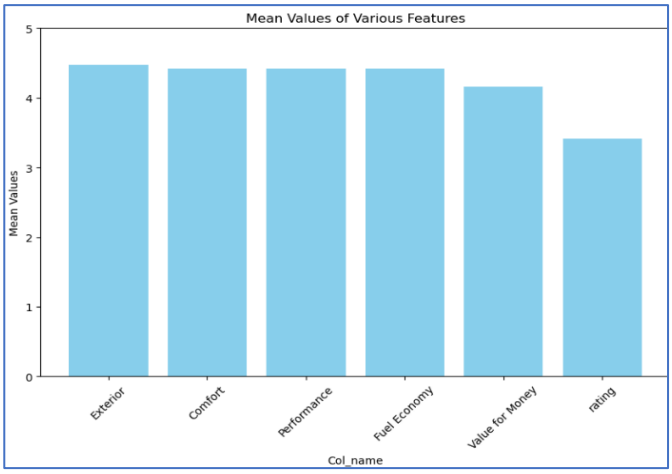
- In addition to obtaining a single sentiment score for each review, we have also computed the negative, positive, and neutral sentiment scores individually for every review using a sentiment intensity analyser.

- The overall sentiment score for the dataset is approximately **0.36**, indicating a slightly **positive trend** in customer reviews. Upon analysis, we found that each review has been categorized into negative, positive, or neutral sentiment scores, providing more detailed insights into customer opinions and experiences with electric vehicles.

- After computing the mean ratings for columns containing integers on the same scale, we generated a bar chart to visualize their averages. The analysis unveiled that the mean ratings for Exterior, Comfort, performance, and Fuel Economy exhibit nearly identical values, while those for Value for Money and Overall Rating are marginally

lower. This indicates that customers express high satisfaction with factors such as exteriors, comfort, and fuel economy. However, they show a slight discontentment with aspects related to value for money and the overall rating.

The Bar chart



Insights from Principal Component Analysis

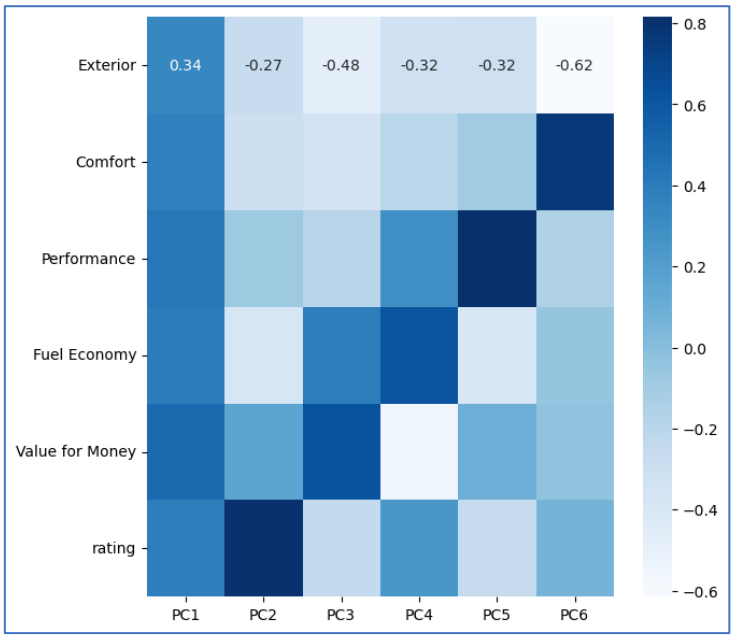
- The Principal Component Analysis (PCA) results indicate that PC1 captures the highest variability in customer reviews, potentially representing overall satisfaction or general quality of electric cars. PC2 and PC3 capture moderate variability and may represent specific aspects such as comfort, performance, or fuel economy.

- PC4 and PC5 capture additional variability, possibly related to factors like value for money or minor features. PC6 captures residual variance. Overall, attributes like exterior, comfort, performance, fuel economy, value for money, and ratings contribute differently to the overall variance in customer opinions, providing insights into various aspects of customer satisfaction and preferences.

- Later calculate the loadings matrix, which indicates how each original attribute contributes to each principal component and creates a data frame for easy interpretation of these loadings.

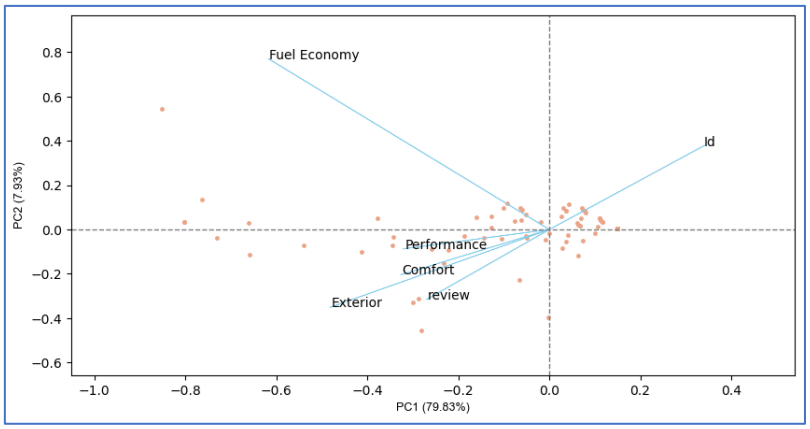
- Using “loadings_df” we created a heatmap to visualize the correlation between the attributes and principal components. The heatmap represents the strength and direction of correlation between each attribute and principal component, providing insights into how each attribute contributes to the principal components obtained from PCA.

Heatmap



- The heatmap shows that PC1 is mainly influenced by the exterior, performance, fuel economy, and value for money. PC2 is strongly affected by ratings. PC3 inversely relates to the exterior and strongly relates to value for money. PC4 is influenced by value for money and fuel economy. PC5 is driven by performance, while PC6 is most affected by comfort and inversely by the exterior. These insights highlight the key attributes contributing to each principal component.

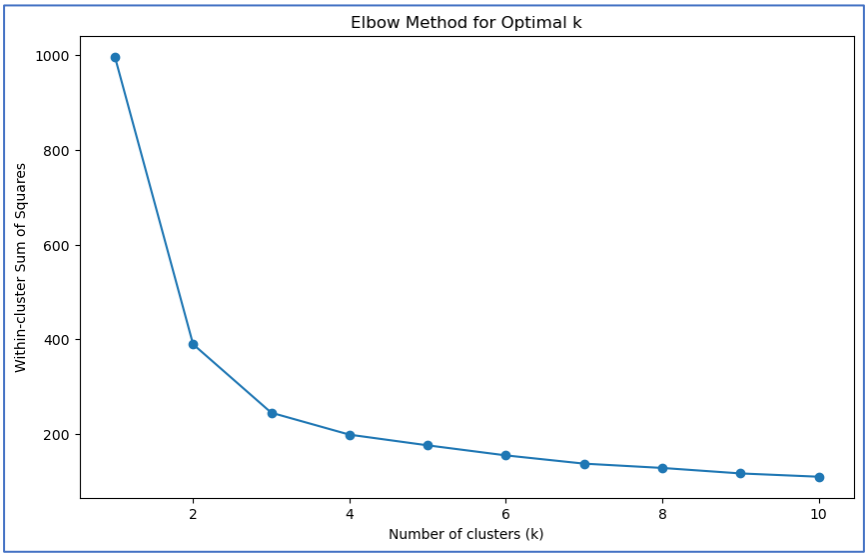
The Biplot



- The biplot shows that PC1 (79.83% variance) is primarily influenced by 'Id', while PC2 (7.93% variance) is strongly influenced by 'Fuel Economy'. 'Performance', 'Comfort', and 'Review' contribute to both PC1 and PC2, with a balanced influence. 'Exterior' has a moderate impact on both components. Most data points are clustered near the origin, indicating low variability, while a few are spread out, showing some variability in the dataset. This helps us understand which attributes drive the variance in customer reviews and how they relate to the principal components.

Insights from K-Means Clustering

Elbow Method

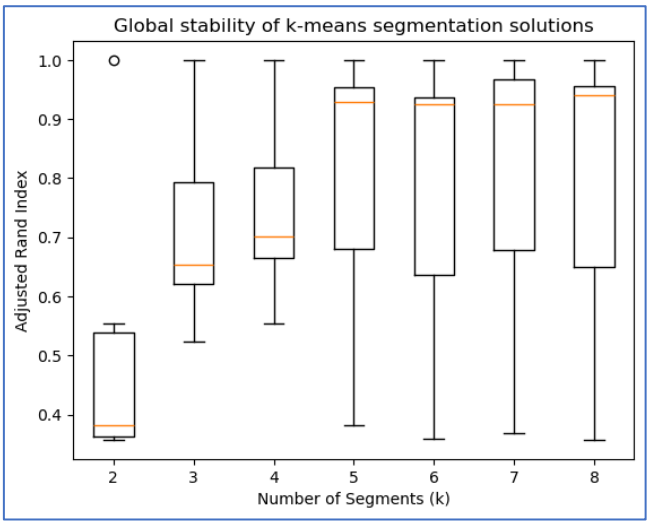


- The Elbow Method plot illustrates the within-cluster sum of squares (inertia) against the number of clusters (k). The goal is to identify the optimal number of clusters for K-Means clustering. Here are the insights from the plot:

    1. Inertia Decrease: The within-cluster sum of squares sharply decreases as the number of clusters increases from 1 to 3, indicating a significant improvement in clustering quality.
    2. Elbow Point: The "elbow" point, where the rate of decrease slows down, appears to be around k=3. This suggests that 3 clusters might be the optimal number, balancing between clustering accuracy and simplicity.
    3. Diminishing Returns: After k=3, the decrease in inertia becomes more gradual, indicating diminishing returns in clustering quality with the addition of more clusters.

- The plot suggests that the optimal number of clusters is likely around 3, as adding more clusters beyond this point results in only marginal improvements in the clustering quality.

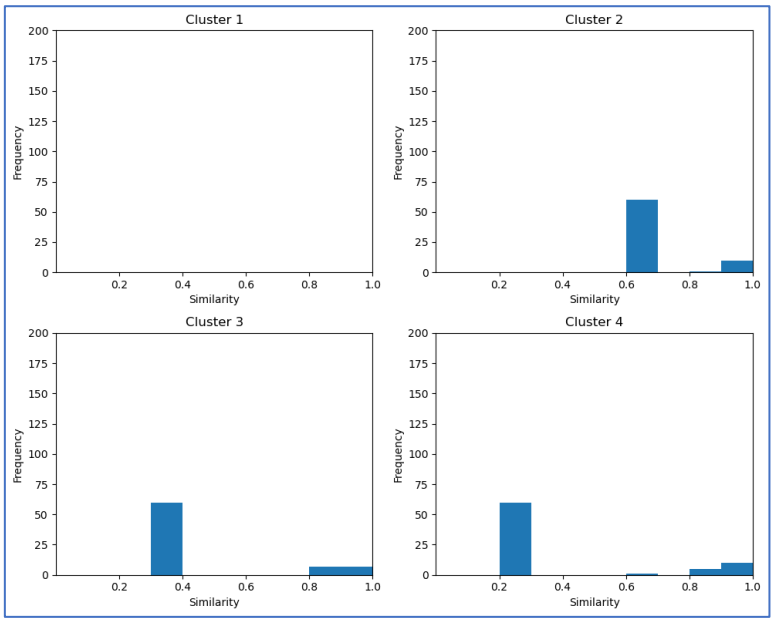
The Vertical Box Plot



- The vertical boxplots display the stability distribution for each number of segments. Higher stability indicates better results. Analysis of the figure suggests that the two-, three-, and four-segment solutions are quite stable. However, the two and three-segment solutions lack differentiation in market insights. Increasing segments to five leads to a significant drop in stability.

- Hence, the four-segment solution is considered optimal, offering a reasonable balance between segmentation depth and stability.

The Gorge plot of the four-segment k-means solution.



- The segments depicted lack clear separation, with similarity values ranging between 0.3 and 0.7, indicating proximity to other segments. Segments 2, 3, and 4 show high stability across varying segment numbers, suggesting they are consistent.

- However, Segment 1 displays instability, drawing members from multiple segments in different solutions, indicating it may not be a reliable target segment despite the overall suitability of the four-segment solution.

Insight Using Mixture of Distribution

- The confusion matrix highlights that K-Means and GMM result in different cluster assignments for the same dataset. This indicates variability in the clustering algorithms' interpretations of the data, suggesting that the clusters are not very distinct or are interpreted differently by each model.
  Below are some insights we got from the confusion matrix:

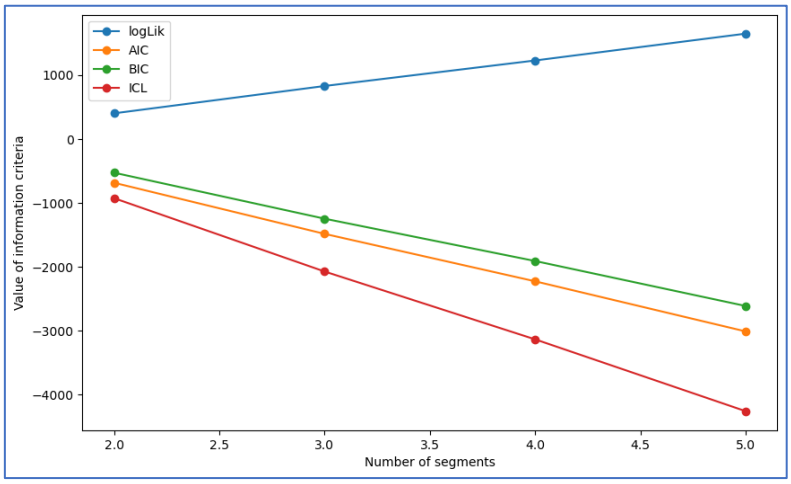
- High Disagreement: There is a significant disagreement between the K-Means and GMM clustering assignments. For example, K-means Cluster 0 is split between GMM's Clusters 1 (69 pt) and 2 (23 pt.).

- Clear Assignments in Some Clusters: K-Means' Cluster 1 is entirely(8 pt.) mapped to GMM's Cluster 3, indicating some agreement in this region of the data.

- Misclassifications: The points in K-Means' Cluster 2 and Cluster 3 show some degree of misclassification when compared to GMM clusters.

- Cluster Stability: The variation in cluster assignments suggests that the data might not have very well-defined clusters or the two models interpret the underlying structure differently.

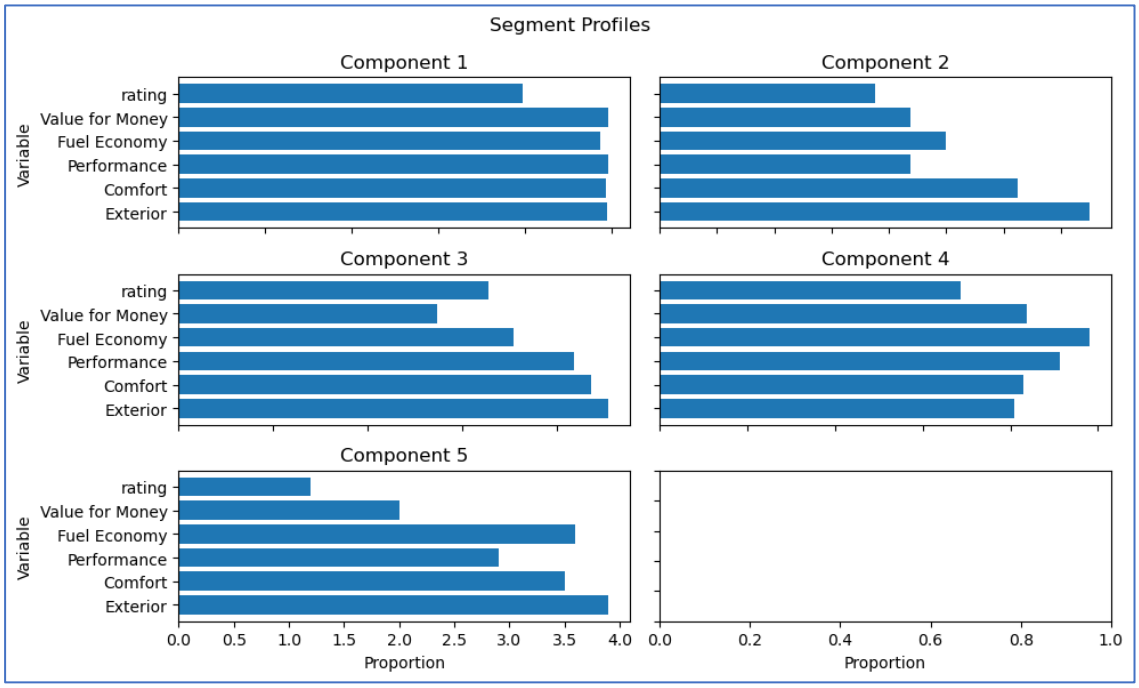
Information Criteria Plot



- The information criteria plot illustrates the behaviour of the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Integrated Completed Likelihood (ICL) as the number of components in the Gaussian Mixture Model (GMM) varies.
- The above information criteria plot shows how AIC, BIC, and ICL change with varying numbers of components. Although these criteria decrease notably until four components, strict inference theory suggests extracting seven segments. However, a more pragmatic view Favors four segments, as beyond this, the decline in criteria becomes less pronounced.

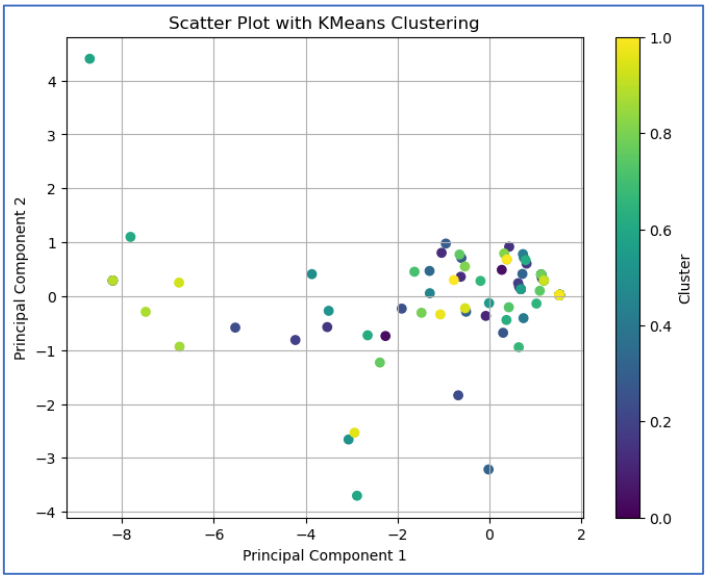
Insight from Profiling Segments

Segment Profiles



- The graph displays segment profiles generated by k-means clustering. Each subplot represents a cluster, showing the mean proportions of variables. The insight we have got:
  1. Visualizes differences between segments based on mean variable proportions and helps to identify which variables characterize each cluster.
  2. Variations in bar lengths indicate segment differences. Higher bars signify variables more influential in defining segments.
  3. Guides targeted marketing and product customization strategies and highlights areas for focused attention on each segment's preferences.

The Scatter Plot

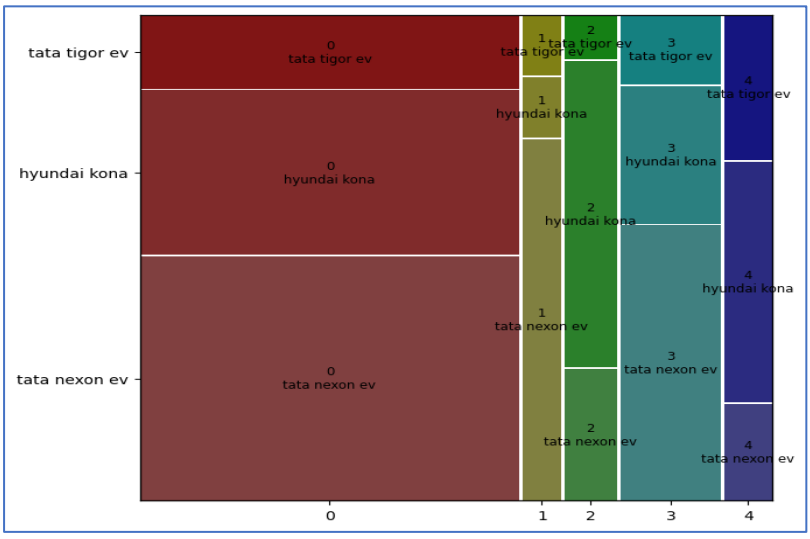


- The scatter plot visualizes the data points in the principal component space with colours representing different clusters. Here's a summary:

1. Each point represents a data sample projected onto Principal Component 1 and Principal Component 2.
2. Colours indicate different clusters assigned by the KMeans clustering algorithm. Points closer together tend to belong to the same cluster. Colours help distinguish clusters visually.
3. Reveals how well the data separates into distinct clusters. Provides an overview of cluster distribution and potential overlaps.

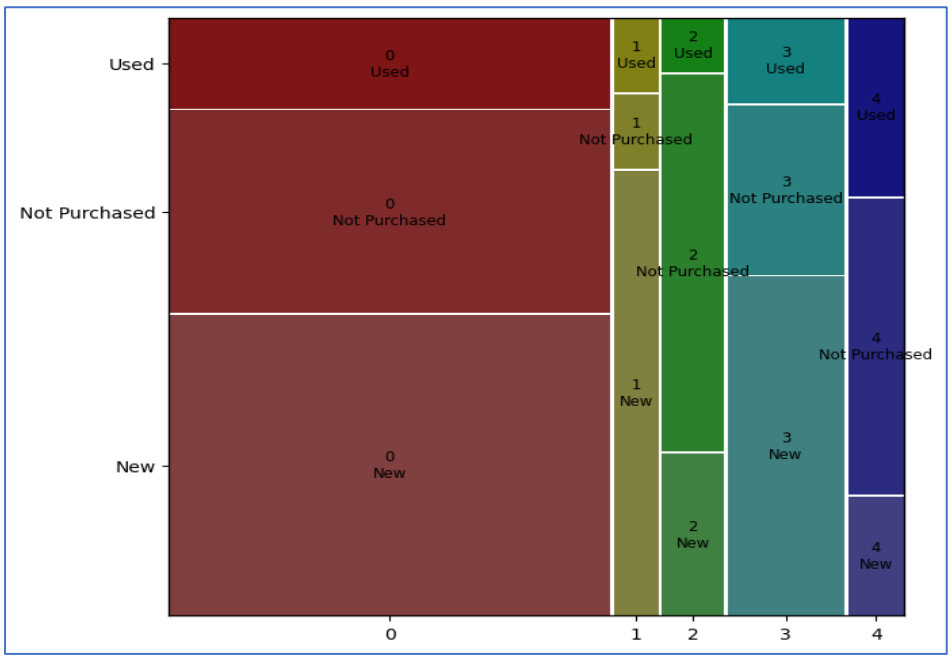
Insight from Describing Segment

The Mosaic Plot for Car Models



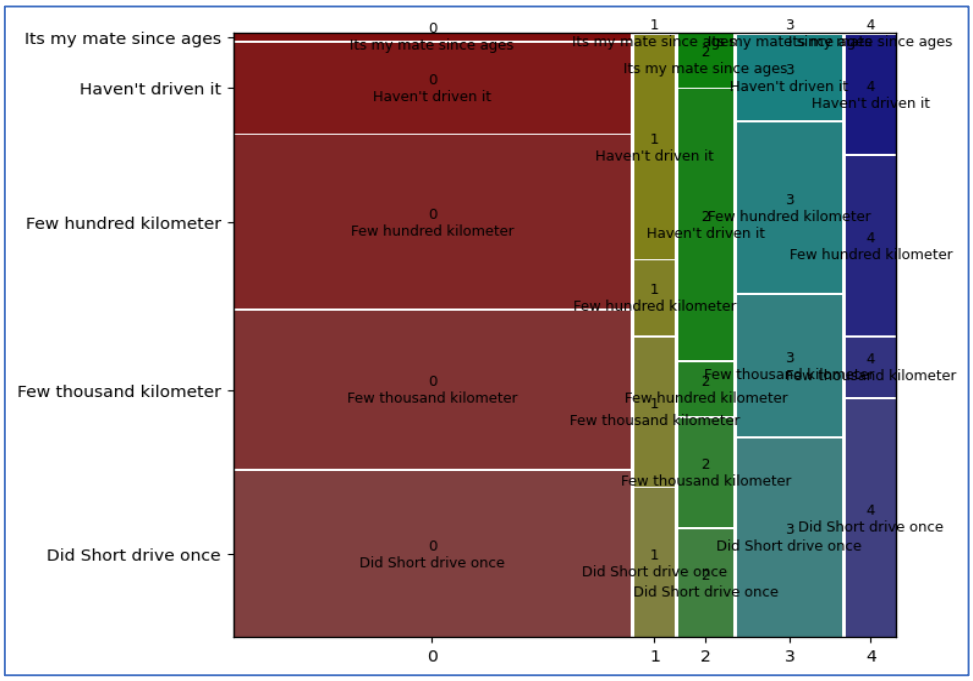
- This visualization helps in understanding how different segments(clusters) of consumers have different preferences for car models, which can be useful for market segmentation analysis and targeted marketing strategies. The insight we got:
    1. Cluster 0 has a significant number of consumers who prefer “Tata Nexon ev”.
    2. Cluster 1 has a mix of preferences, with some consumers choosing the “Tata tigor ev” and others choosing “Hyundai Kona”.
    3. Cluster 2 has a notable preference for “tata nexon ev”.
    4. Clusters 3 and 4 show a more diverse distribution of car model choices.

Mosaic Plot for Car Conditions



- The mosaic plot visualizes the relationship between cluster no. and consumers' choice for car condition. Here are some insights:
    1. Cluster 0 has a dominant segment with many consumers who have not purchased a car.
    2. Clusters 1 & 4 have balanced distribution among conditions and potential for diversified marketing.
    3. Clusters 2 & 3 prefer new cars and new models.
    4. The shading indicates deviations from expected frequencies, with darker shades showing stronger associations. This analysis helps tailor marketing strategies to different segments based on their preferences and market potential.

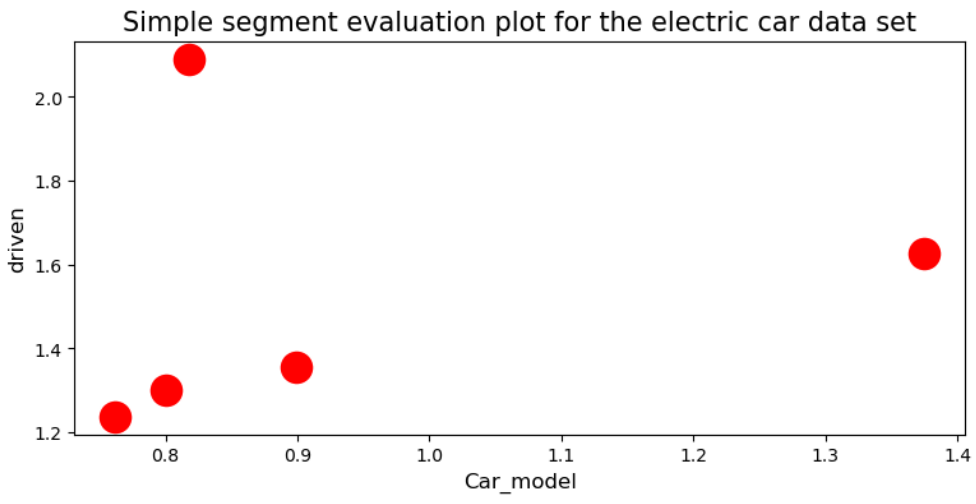
Mosaic Plot for consumers driven feature



- This mosaic plot shows the relation between cluster no. and consumers' driving experience. Insights are:
  1. Cluster 0 has a dominant segment with diverse driving experiences.
  2. Clusters 1 & 4 have balanced distribution, requiring diversified marketing
  3. Clusters 2 & 3 have a preference for “Few thousand kilometres” and “Few hundred kilometres”, suggesting a focus on moderately experienced drivers.
  4. Potential Market: A significant number of consumers in the “Haven’t driven it” category, indicating opportunities for test drives and introductory offers.

Insight for Target Selection

Scatter Plot



- The scatter plot provides a clear visual representation of the relationship between car models and consumers' driving experiences. This information can be used to develop targeted marketing strategies, such as promoting test drives for models with low driving experience or leveraging brand loyalty for models with high driving experience.
- Spread-out dots suggest models appealing to a varied range of driving experiences.
- Market opportunities lie when we focus on models with high “Haven’t driven it” for test drives and loyalty-building for models with “It's my mate since age”.

Features Affecting Market Segmentation

- Market segmentation based on the most driven car by customers and the car model name helps in understanding customer preferences effectively. By analysing which car models are most commonly driven and preferred by customers, businesses can tailor their marketing strategies to target specific segments of the market more accurately.
- This approach allows for better product customization, more efficient marketing campaigns, and improved customer satisfaction. Gathering data from sources like car sales websites, surveys, and market research helps in identifying popular car models and segmenting the market accordingly, leading to more informed business decisions.