MARKET REPORT

# EV MARKET IN INDIA

This report aims to identify key segments to target for early market entry Indian electric vehicle (EV) market as a Startup.

# Market Segmentation Analysis of Electric Vehicle Market in India

**Team Aryan Report Contributions**

- Aryan: Sections 1 (1.1, 1.2 , 1.3 – 1.3.1/2/3), 2, 3.1, 6
- Bhaskar: Section 3.2
- Veena: Section 4
- Prateek: Section 5

**GitHub link:** https://github.com/Aryan092/EV-Market-Segmentation

## 1. Problem Statement

To be a successful start-up in a growing space such as the Electric Vehicle (EV) in a competitive market such as the one in India, it is important to properly define the foundation of the company. An effective target vehicle and customer space are imperative to thrive.

Hence this business feasibility report seeks to define these spaces through a Segmentation analysis of the EV market in India. The suggested strategy to enter the market will target the segments that are most likely to connect with our product from a standpoint of Demographic, Psychographic and Behavioural.

Following sub-sections detail the Pre-Segmentation study to define areas of interest.

## 1.1 Market Segmentation research Questions

The following questions are formulated in order to be able to identify the important segments whilst recognising the data requirement and availability restrictions present.

Questions formulated to direct the Market Segmentation Research:

A. What type of EV will the company produce?
B. Vehicular features that ought to be valued?
C. To whom will it sell?

## 1.2 Key Metrics identified

The key metrics are derived from the questions and emphasize the directly addressable factors.

A. Vehicular type
B. Consumer feature preference
C. User-base descriptors

## 1.3 Data Collection

To meet the metric identified the form of data necessary can be respectively categorised as follows:

A. EV type data, EV Sales data
B. Consumer Vehicular feature rating data
C. Datasets with buyer descriptor variables (i.e. age, income etc.)

| Reference | **Segmentation Questions** | **Key Metrics** | **Data Necessary** |
|---|---|---|---|
| A | What type of EV will the company produce? | Vehicular type | EV type data, EV Sales data |
| B | Vehicular features that ought to be valued? | Consumer feature preference | Consumer Vehicular feature rating data |
| C | To whom will it sell? | User-base descriptors | Datasets with buyer descriptor variables |

*Table 1.3.1 Pre-Segmentation study Summary*

### 1.3.1 Data Requirements

There is a threshold quantity and quality of data necessary to make market segmentation worth it. For effective market segmentation factors such as levels of missing data, data cleanliness, diversity of data sources and data source reliability are all relevant.

To ensure these were met a thorough investigation of the market data availability and relevance was conducted. Data exploration and pre-processing was done to investigate dataset behaviour and its compatibility with addressing the project questions.

### 1.3.2 Data Sources

Data was pooled from open source databases to ensure reproducibility and cross-examination of our code and results.

Sources:

- https://datasetsearch.research.google.com/
- https://trends.google.com/trends/explore
- https://kaggle.com/datasets

### 1.3.3 Data Pre-Processing

Primary libraries used:

- Pandas: To load datasets for processing
- Numpy: For Array-based calculations
- SKLearn: To perform clustering, pre-processing scaling methods and PCA analysis
- Matplotlib and Seaborn: To create visualizations to perform further analysis and showcase results

## 2. Vehicle type Segmentation

The *Fig. 2.1* shows the number of EV's in circulation per type from 2017 to 2023 in India. The plot shows noticeable growth across all vehicle types except electric buses. However, it is the two wheelers, i.e. predominately bikes and scooters, that not only are the EV's with most vehicles in circulation but also the ones that display the highest rate of growth over the recent half decade.

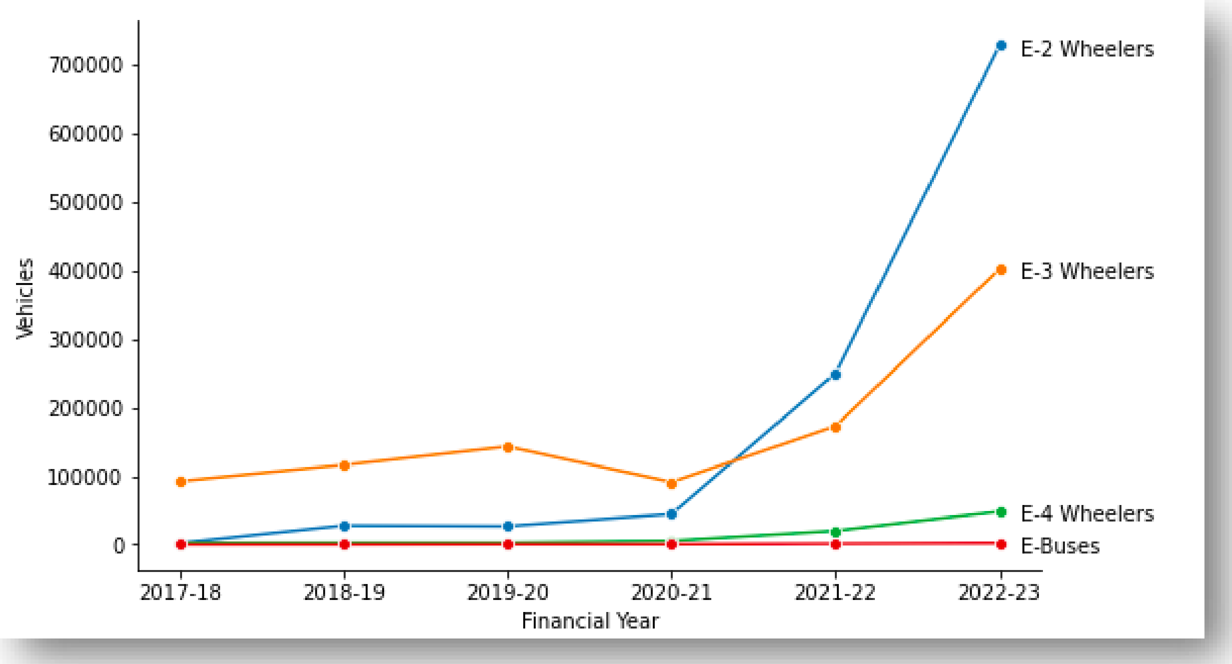


*Fig. 2.1: Electric Vehicle growth over the last 5 years for each type*

This underscores the growth trajectory of the electric two-wheeler and marks it as a dominant leader the EV market. Evaluating based solely on this, the start-up ought to invest in electric two-wheeler as its choice of EV. However, electric four wheeler has shown promise post pandemic (2020 onwards) in India and has proven to be a successful endeavour in overseas markets. This could, therefore, point towards a relatively untapped sector in the vast long-term outlook of the EV market in India.

Hence, the vehicular type segment analysis in this report will focus on the electric two and four wheeler. The analysis will place an emphasis on segmenting user ratings and analysing their sentiments on their features and performance.

## 3.1 Behavioural segmentation – Electric Two-Wheeler

### Data Collection

The data used here is from Kaggle, an open-source data science platform, and is originally scrapped from Bikewale, which is one of India's largest research two-wheeler research platforms. Bikewale offers research, pricing and marketplace information regarding all two-wheelers including electric two-wheelers.

The dataset comprises of electric two-wheeler customer reviews and ratings. The columns within the data set are explained as follows:

- ‘review’: Consumer sentiments expressed in sentences
    - No sentiment analysis was performed using this variable – assessed that the larger review sentiment was carried within the ratings and individual nit-pickings would not transfer in understanding the general market sentiment through segmentaion.
- ‘Used it for’: Purpose the bike was used for
- ‘Owned for’: Length of ownership
- ‘Ridden for’: Distance travelled on vehicle
- ‘rating’: Overall integer rating of the vehicle measured 1 to 5
- ‘Visual Appeal’, ‘Reliability’, ‘Performance’, ‘Service experience’, Extra features’, ‘Comfort’, ‘Maintenance cost’, ‘Value for Money’: Subjective integer rating on the particular feature or component of the vehicle measured 1 to 5
- ‘Model Name’: Model of two-wheeler used

### Exploratory Data Analysis (EDA)

First, we select numeric variables, i.e. the different ratings. These are our segmentation variables. We plot the correlation matrix to observe the inter-variable correlation.

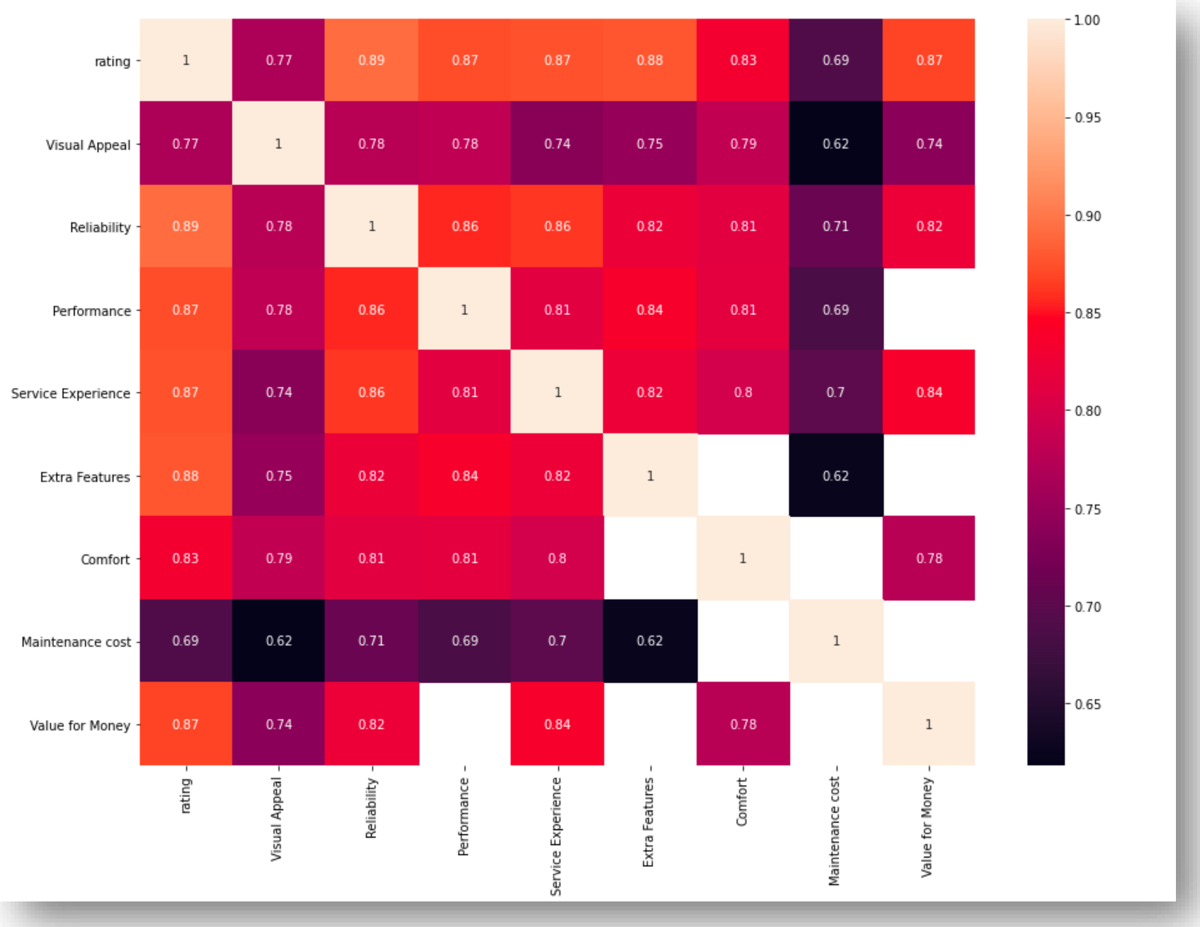


*Fig 3.1.1 Correlation Matrix for the Segmentation variables for the EV two-wheeler*

The major takeaway from the correlation matrix above (*Fig 3.1.1*) is the relatively high degree of corelation of almost all subjective ratings with the overall rating. Maintenance cost is the only rating that shows lower than 70% correlation. This could be either due to users viewing maintenance cost as a usage issue rather than vehicle issue thereby not accounting for it in the vehicle rating or it could be due to the abundance of missing data for this variable as detailed later on.

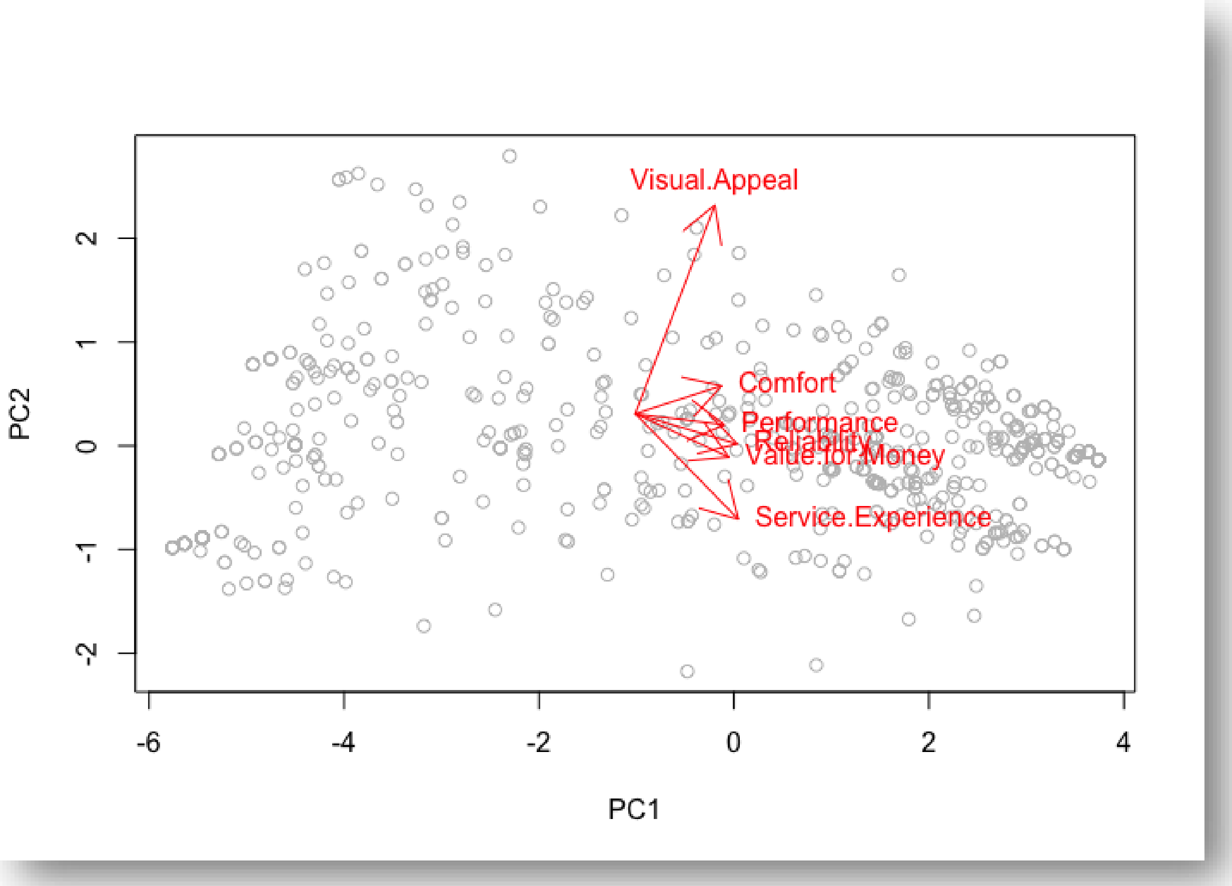


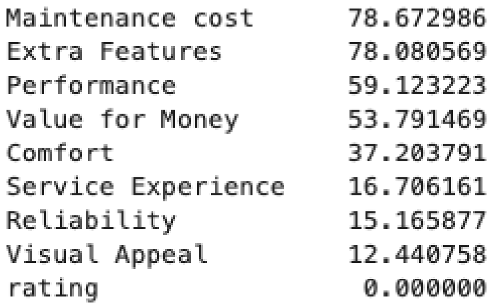
*Fig 3.1.2 Principal Component Analysis (PCA) map for the EV two-wheeler data set*

In EDA, a PCA map provides preliminary insights into how respondents rate attributes and which attributes tend to be rated similarly. In *Fig 3.1.2* , the Visual Appeal attribute plays a key role in the evaluation of two-wheelers in India. The remaining attributes align with what can be interpreted as positive perceptions – as all attributes point in the same direction in the perceptual chart.

These initial exploratory insights represent valuable information for segment extraction. Results indicate that all attributes are strongly related to one another, and that the Visual appeal dimension may be critical in differentiating between groups of consumers later on in the segmentation analysis.

**Missing Data**

The following table highlights the percentage of missing data across the different features.



```
Maintenance cost      78.672986
Extra Features        78.080569
Performance           59.123223
Value for Money       53.791469
Comfort               37.203791
Service Experience    16.706161
Reliability           15.165877
Visual Appeal         12.440758
rating                 0.000000
```

*Table 3.1.1 Missing data percentage for the Segmentation variables in the data set*

According to *Table 3.1.1*, Several columns report a significant amount of missing data, with 4 columns missing over half of the data.

In cases where the amount of missing data is small relative to the size of your dataset and does not bias the analysis, one might choose to remove rows or columns with missing values altogether.

However, that does not seem to be the case here with  high degree of missing data for several variables. To ensure there is no bias in missing data the mean as per different groups were checked. No such bias was found for the data with high degree of missing data. “Occasional Commute” option for the “Owned for” variable showed several missing valued for the segmentation variables and was therefore dropped.

Therefore, due to the great amount of missing data where imputation would result in data bias and corruption and row removal result in great loss of data the two features with close to 80% missing data are dropped, “Maintenance cost” and “Extra Features”. Imputation is performed for each feature with the mean by group with the highest correlation feature instead of using the overall feature mean, which would lose some of the nuance within the data.

For example, the highest correlation feature for ‘Reliability’ is ‘Rating’. Now if there is a missing value in a row for ‘Reliability’ and ‘Rating’ = 1 (no missing values), that missing value in ‘Reliability’ is replaced by the mean of ‘Reliability’ when ‘Rating’ = 1.

### Extracting Segments

K-Means stands out as one of the most widely used Unsupervised Machine Learning Algorithms for tackling Classification Problems. This algorithm effectively partitions unlabelled data into distinct groups or clusters, based on shared features and patterns.

Consider our unlabelled multivariate dataset, encompassing diverse attributes such as Visual Appeal, Reliability, and Performance. Utilizing K-Means, the data is partitioned into multiple clusters, each representing entities with similar characteristics. Clustering is a pivotal technique in Unsupervised Learning Algorithms and enables the segregation of multivariate data into distinct groups solely based on inherent patterns, without requiring external supervision.

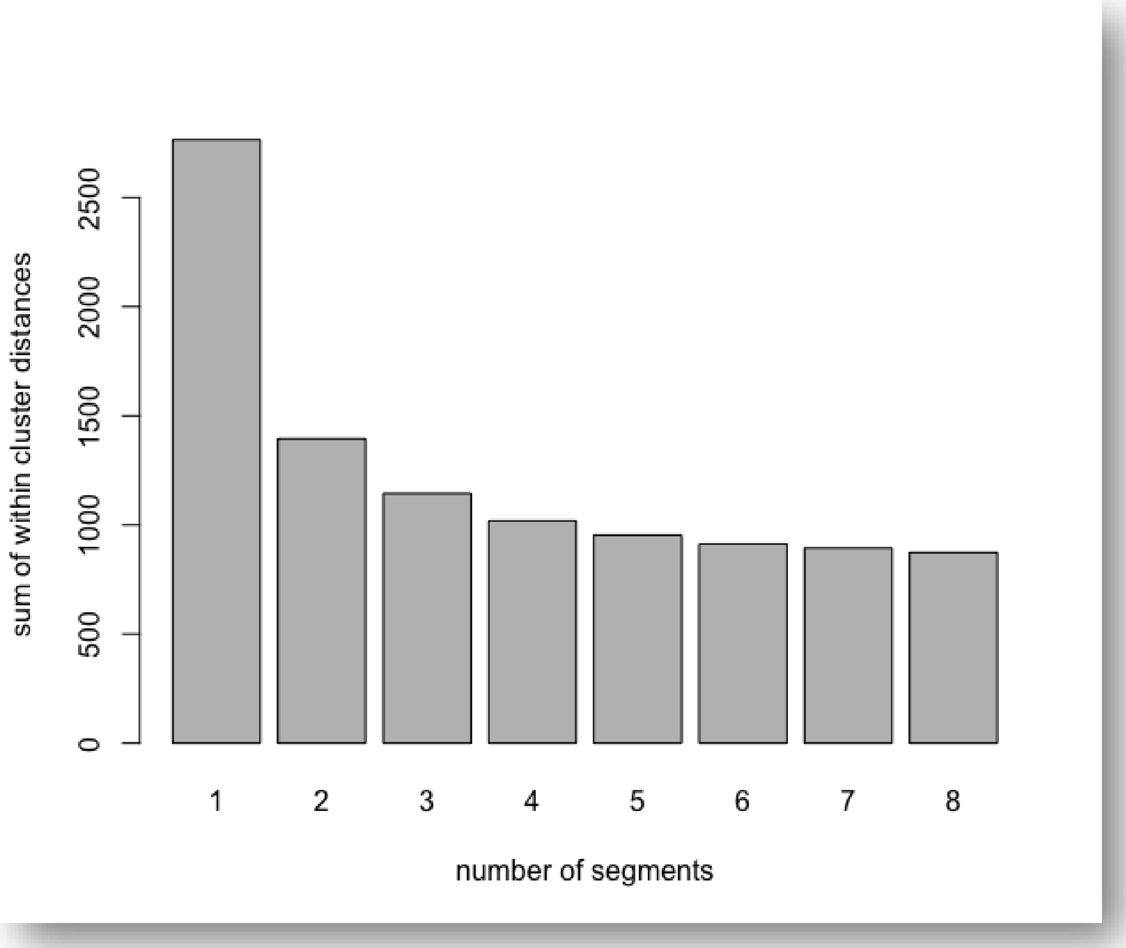


*Fig 3.1.3 Scree plot for the EV two-wheeler data set*

However we do not know in advance how many segments are optimal for our case hence we calculate and compare over a range for example 2 to 8. The goal here is to select clusters that capture segments comprising similar consumers while being notably distinct from members of other segments.

The scree plot in *Fig 3.1.3* is one way of comparing the number of segments. A significant elbow is desirable – a point where the sum of cluster distance decreases noticeably slowly while the number of segments increases. This point in our graph and hence our number of segments can be identified as 3.

We will try to consolidate this result by implementing a second approach of determining a good number of segments by using stability-based data structure analysis (SLSA). SLSA also attests to the nature of the segments forming naturally or artificially. The results of the segment analysis must be reproducible to be worth investing heavily. Therefore, analysis of the stability of the segment solution is necessary.

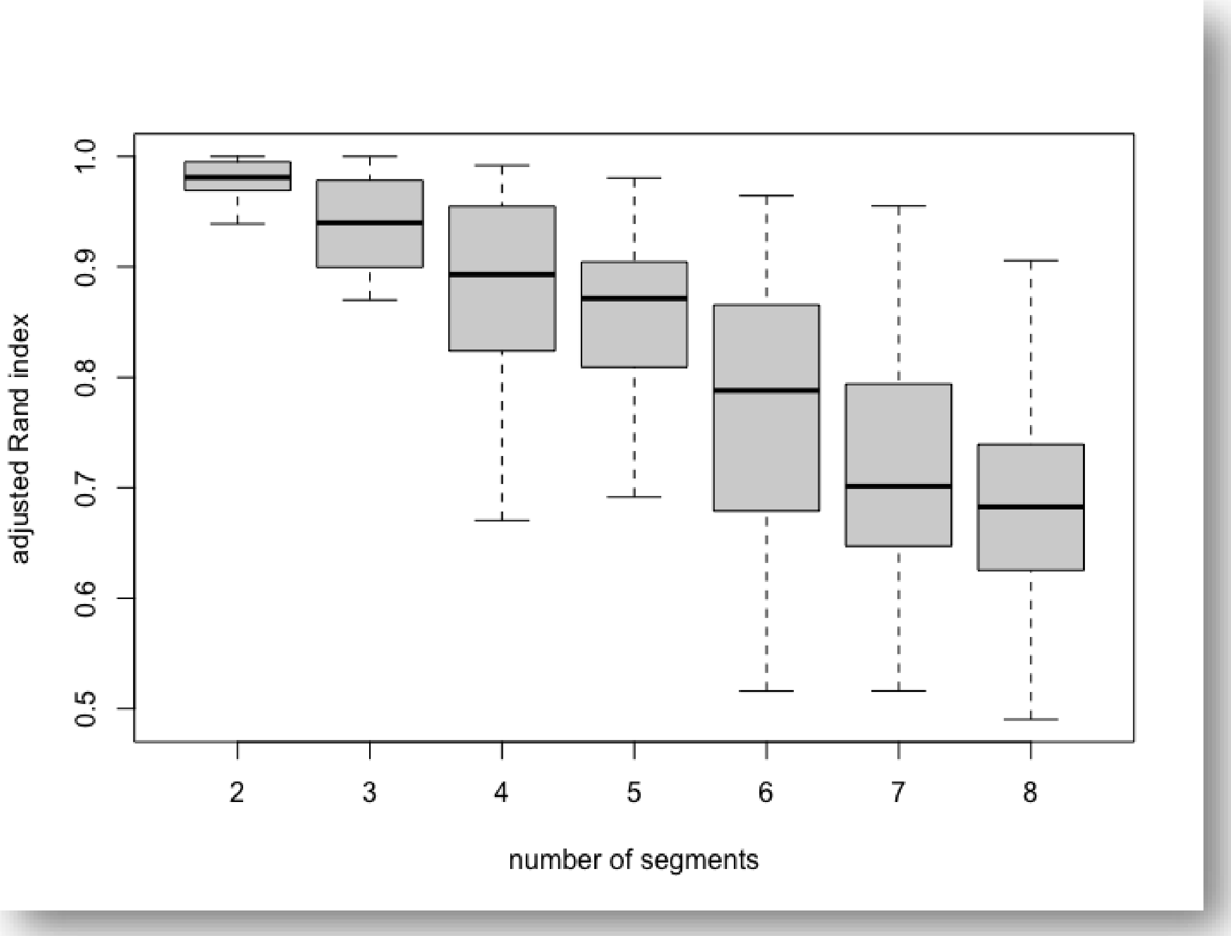


*Fig 3.1.4 Global Stability box blot of k-means segmentation solutions for EV two-wheeler*

Inspecting *Fig 3.1.4* points to the 2, 3, 4 and 5 segment solutions as being quite stable. Despite low number of segments not lacking interesting insights considering the same directional feature PCA plot (*Fig 3.1.2*) and the elbow method from the scree plot (*Fig 3.1.3*) the 3-segment solution still marks to be the best solution – the solution with the most mark segments with a high degree of replication. This is further backed by a noticeable drop in replicability in the 4-segment solution.

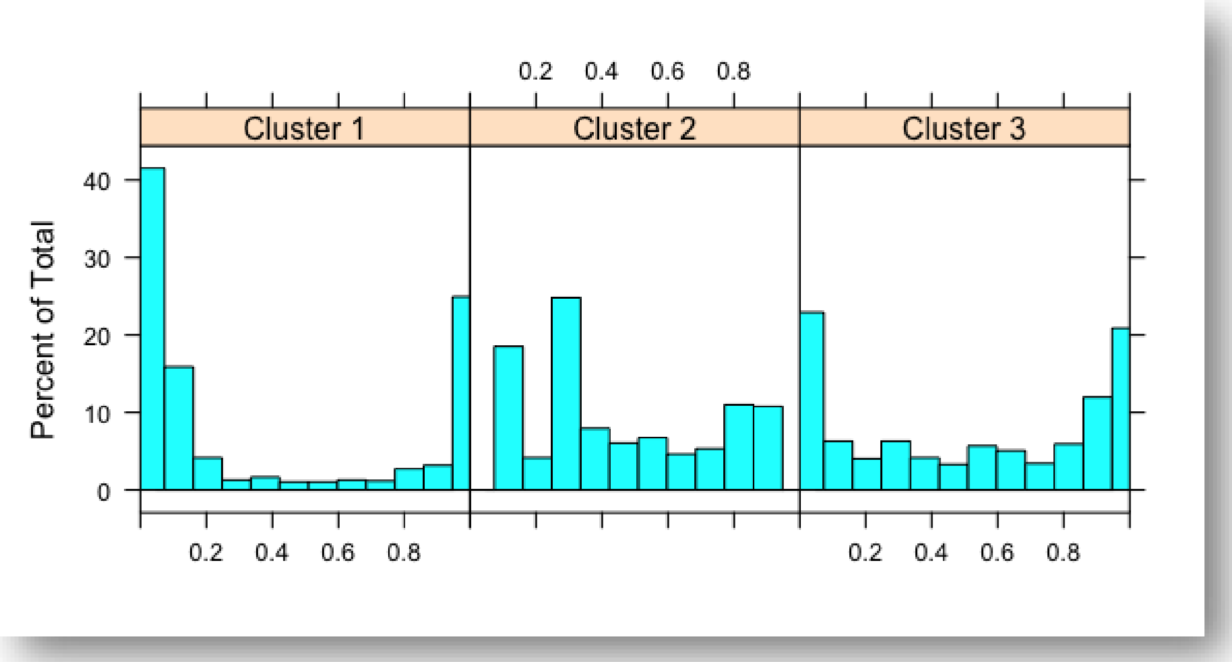


*Fig 3.1.5 Gorge plot of 3-segment k-means segmentation solutions for EV two-wheeler*

In *Fig 3.1.5,* we try to analyse the structure of the 3-segment solution. However, none of the segments shown in the gorge plot are well separated from the other segments, and proximity to at least one other segment is present as indicated by the similarity values all being distributed comparably between 0 and 1.

Finally, we analyse how segment level stability changes across solutions each time another segment is added.

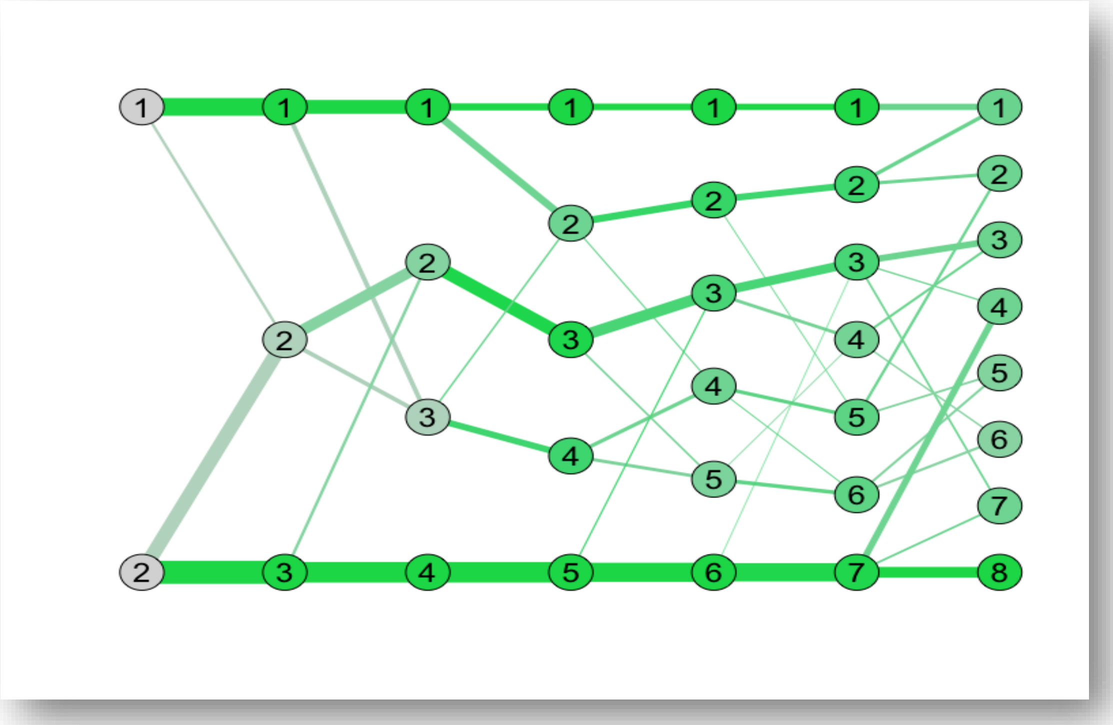


*Fig 3.1.6 SLSA for 2-8 segment solutions for the EV two-wheeler*

In *Fig 3.1.6*, the thick green lines indicate that many members of the segment to the left of the line move across to the segment on the right side of the line. Segment 2 in the two-segment solution (in the far left column of the plot) remains almost unchanged until the seven-segment solution, then it starts losing members. This seems in line with the global stability box plot. However keeping in mind earlier analysis, the SLSA plot in view of the earlier determination that the three-segment solution looks fine and but it seems segments 2 and 3 are nearly identical.

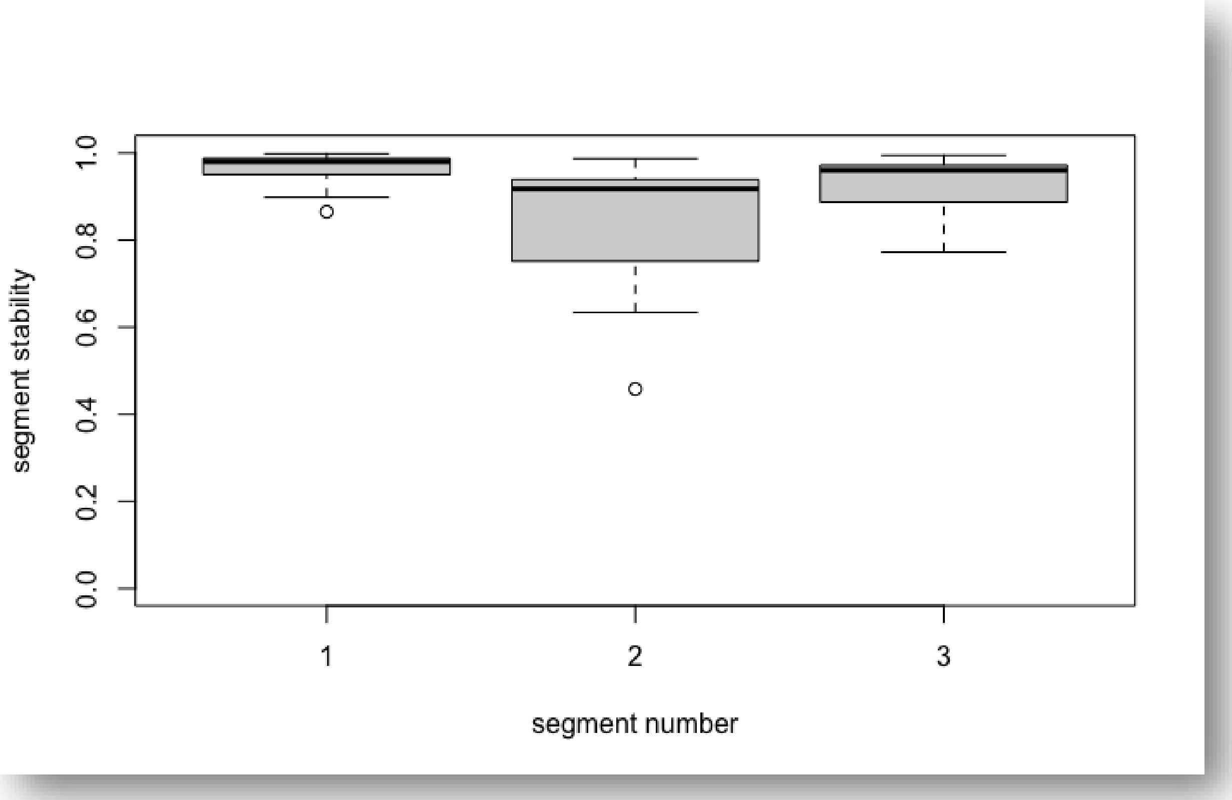


*Fig 3.1.7 SLSW for 2-8 segment solutions for the EV two-wheeler*

*Fig 3.1.7* shows the segment level stability within solutions (SLSW) for the 3-segment solution. Segment 2 is marginally the least stable across replications, followed by segments 3 and 1. Segment 3 is the most stable. All segments unaspiringly display a high level of stability which follows on from the SLSA plot.

**Profiling Segments**

The segmentation analysis is complete and the market segments have been extracted. To understand our solution several easy-to-understand visualisations are produced.

The segment profile plot shows the key characteristics of each market segment. While highlighting the major differences between segments. To aid in interpretation, similar attributes have been positioned close to one another through hierarchical cluster analysis.

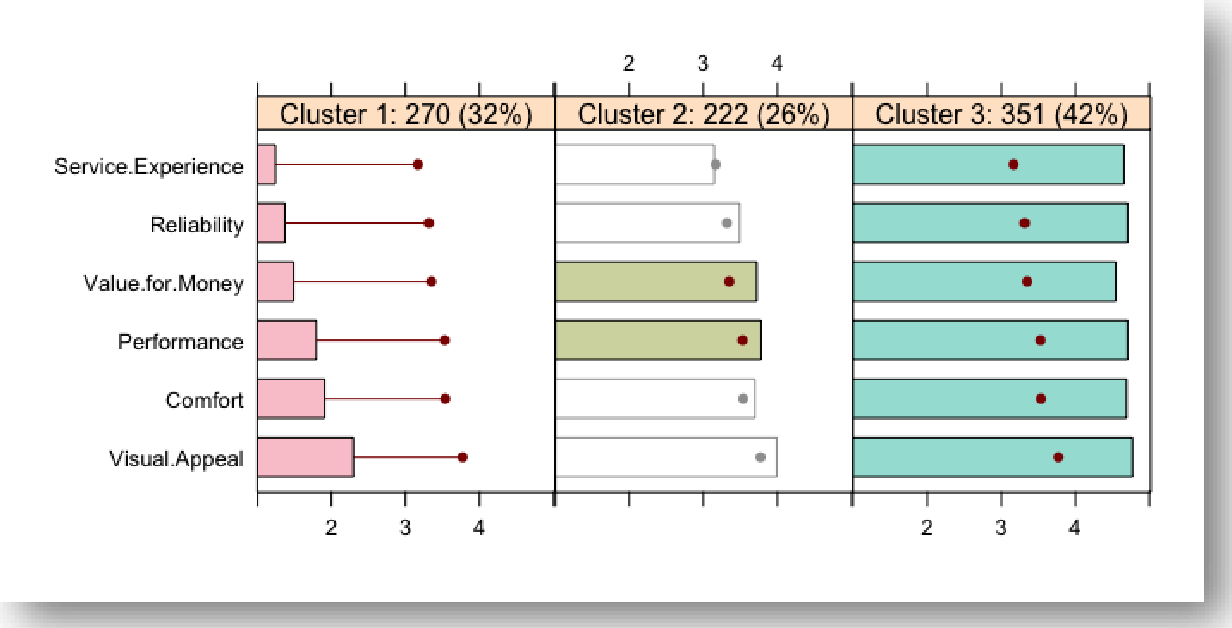


*Fig 3.1.8 Segment profile plot for the EV two-wheeler*

Looking at *Fig 3.1.8*, segment 1 thinks EV two-wheelers are generally low, significantly below average, across the board whereas segment 3 thinks highly, above average, in every metric. These are distinctly different perceptions. Segment 2 puts emphasis on EV two-wheelers as value for money and high performing.

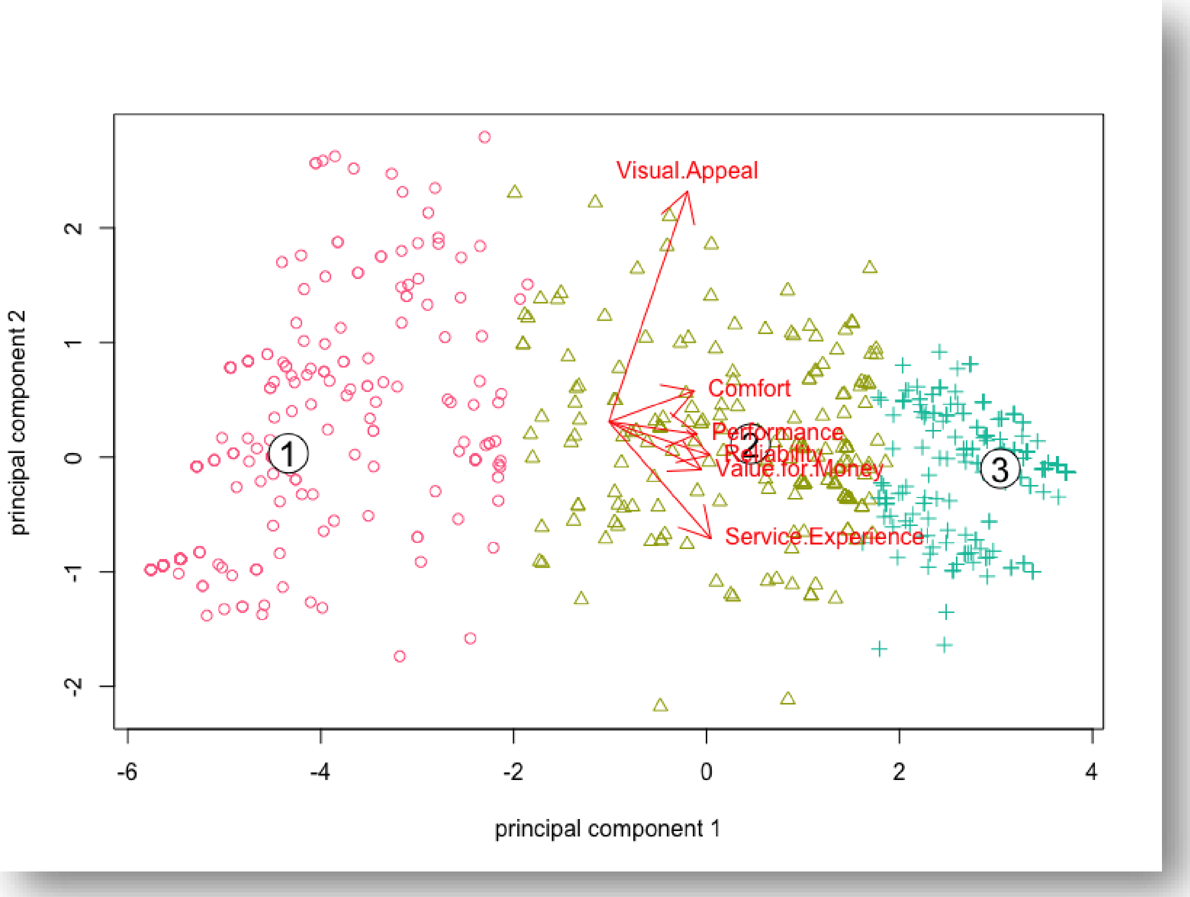


*Fig 3.1.9 Segment separation plot for the EV two-wheeler*

*Fig 3.1.9* is the PCA plot from earlier with the newly found segments differentiated by colour. The segment separation plot reflects in many ways the segment profile plot from earlier with largely three distinct perceptions ranging from dislike-like (moderate)-love.

### Describing Segments

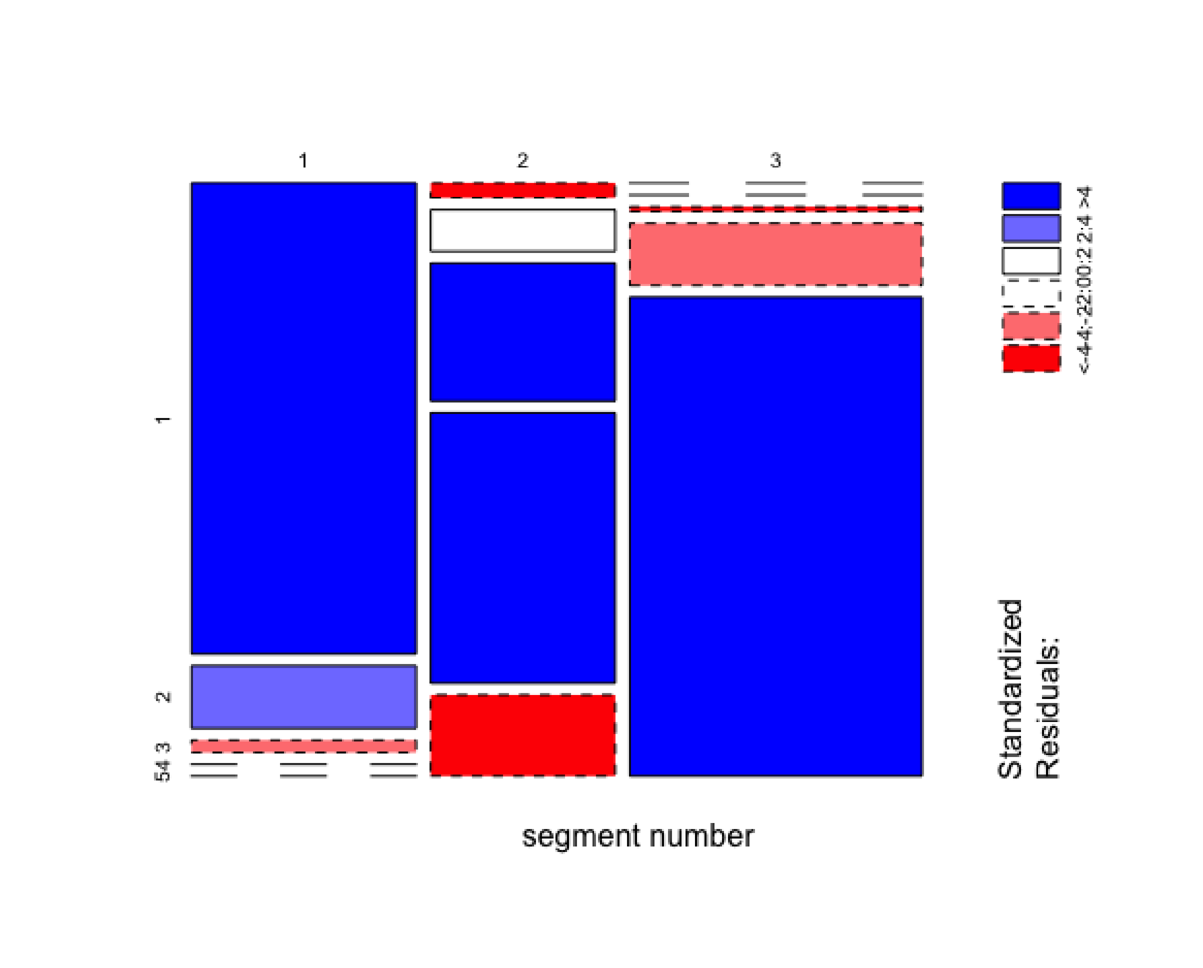


*Fig 3.1.10 Shaded mosaic plot for cross-tabulation of segment membership and overall rating for the EV two-wheeler data set*

The mosaic plot in *Fig 3.1.10* plots segment number along the x-axis, and the overall user rating their EV two-wheeler along the y-axis. The mosaic plot reveals a strong and significant association between these two variables. Members of segment 1 (depicted in the first column) do not highly rate their EV two-wheelers whatsoever shown by the completely empty boxes bottom left and seldom find them average (rating = 3), as indicated by the bottom left box being coloured in red. In stark contrast, members of segment 3 are significantly more likely to rate an EV two-wheeler highly, as indicated by the dark blue box in the bottom right of the plot. At the same time, these consumers are less likely dislike or find EV bikes average, as indicated by the completely empty boxes and very small red box at the top right of the plot. Members of segment 2 appear to have the moderate feelings towards EV bikes (rating largely 3-4); their generally have an above average liking with a small likelihood to fall into extreme liking or disliking for the product.

This is to be expected considering the high levels of correlation to overall rating observed in the EDA correlation matrix and is further reinforced following the missing data imputation strategy.

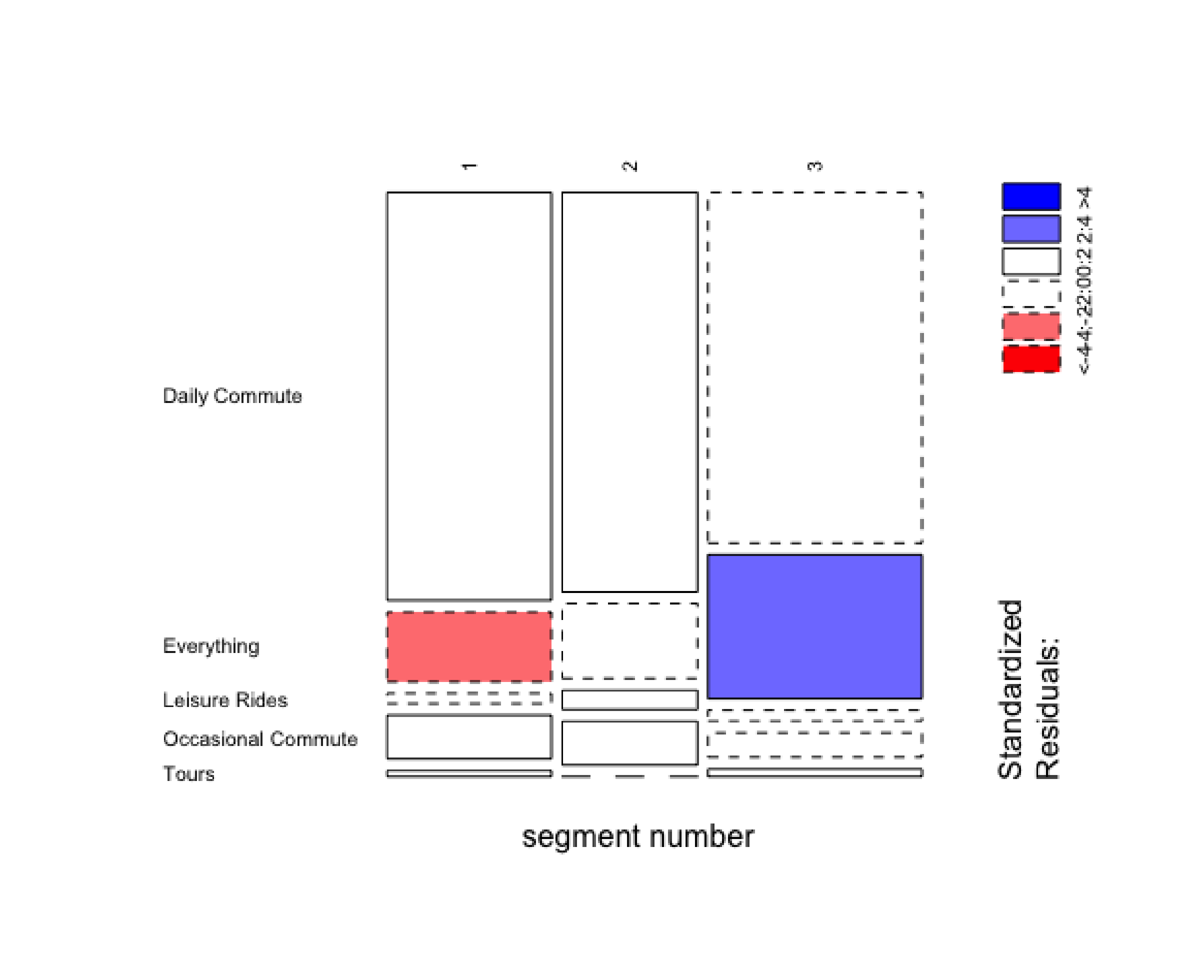


*Fig 3.1.11 Shaded mosaic plot for cross-tabulation of segment membership and usage style for the EV two-wheeler data set*

In *Fig 3.1.11*, the descriptor variable (Used it for) is plotted along the y-axis. This mosaic plot offers the following additional insights about our market segments: segment 1 and segment 2 have a similar use case distribution with an emphasis on daily commute. Segment 1 has a very low likelihood to use EV bikes for everything. Segment 3 contains more individuals who use EV bikes for everything and have a higher tendency to use their EV bikes for everything rather than a specific use case.

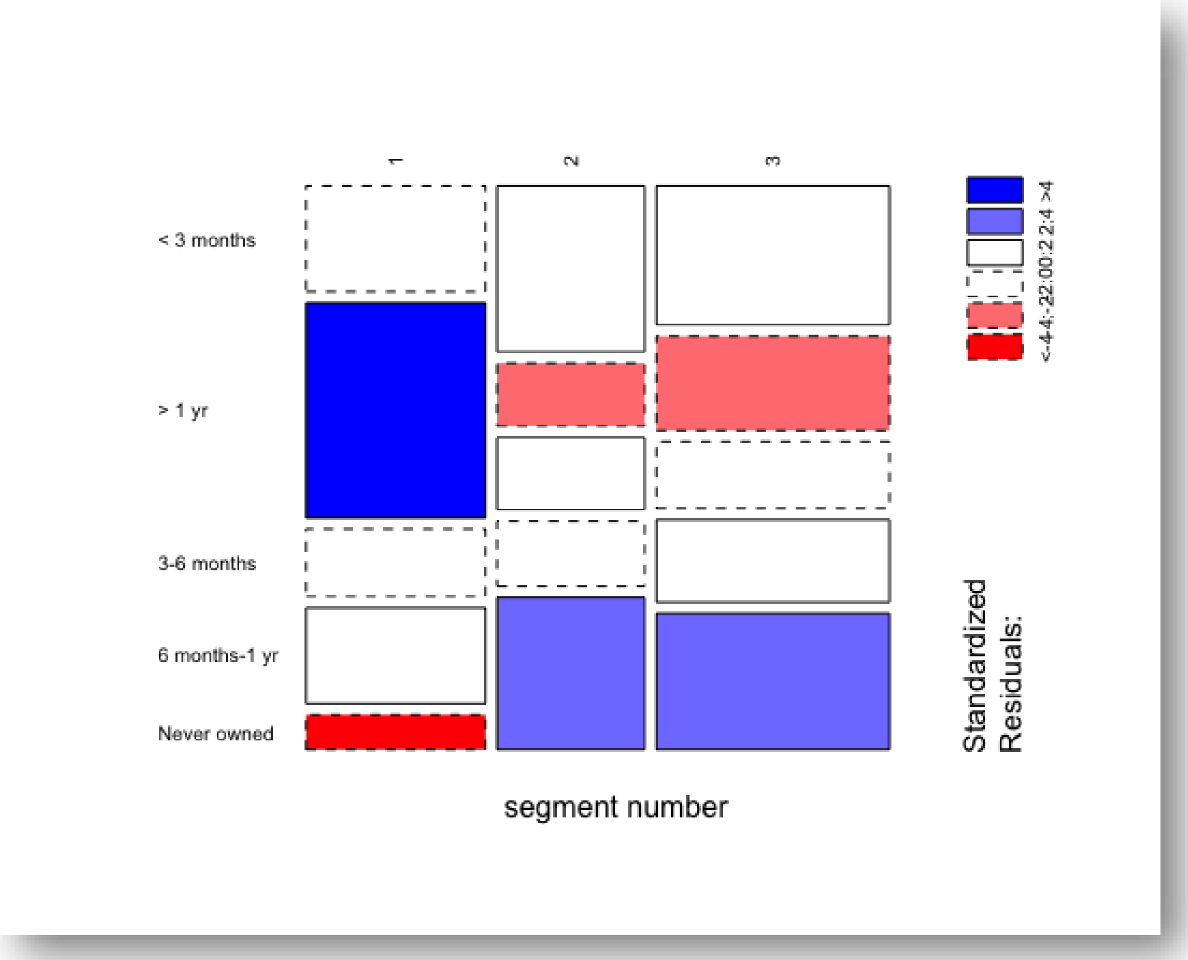


*Fig 3.1.12 Shaded mosaic plot for cross-tabulation of segment membership and length of ownership for the EV two-wheeler data set*

In *Fig 3.1.12*, the descriptor variable (Owned for) is plotted along the y-axis. Segment 2 and segment 3 have a similar distribution with regards to their length of ownership. In both these segments, users tend to have a greater likelihood to never own the vehicle. Segment 1, the haters, contains significantly more individuals owning the product for a length greater than 1 year. (as depicted by the larger blue box for the >1 yr category).

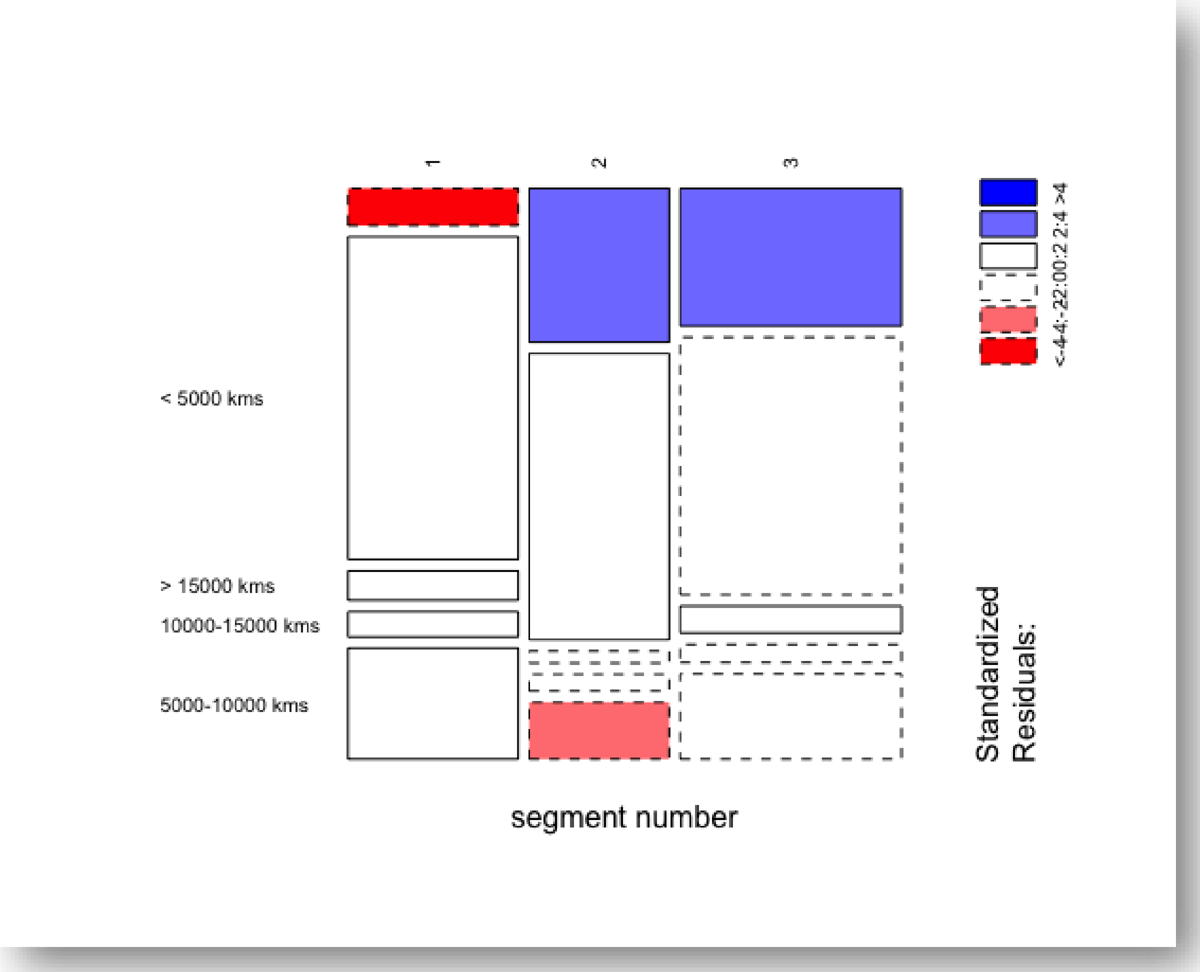


*Fig 3.1.13 Shaded mosaic plot for cross-tabulation of segment membership and distance ridden for the EV two-wheeler data set*

In *Fig 3.1.13*, the descriptor variable (Ridden for) is plotted along the y-axis. Segment 2 and 3 once again largely contain users that tend to have a greater likelihood to never own the vehicle. Segment 1, the haters, have very low likely to never have owned the vehicle. This segment is also largely made up of users with less than 5000 kilometres travelled despite having the vehicle for more than 1 year makes sense when you consider that daily commute is what many users segment 1 use their vehicles for.

**Most Optimal Market Segment**

The market segmentation highlights Segment 3 as the largest group (42%) and a group of users who love the product, therefore, naturally should be the core target segment. Targeting this group would entail developing an all-purpose electric two-wheeler with a use case of everything. The low ownership becomes problematic from a sales point of view hence converting these individuals into buyers will be critical.

Segment 1, on the other hand, comprise of owners that highlight the shortcomings in the current available electric two-wheelers. This userbase would demand a higher quality but would perhaps be willing to pay higher for those product improvements given their underlying interest – higher purchase rate despite low subjective returns. A space of unsatisfied buyers ought to be targeted. The use case of daily commute would need to be targeted here.

### Conclusion

EV two-wheeler owners of higher durations tend to rate the vehicle lower this is most likely due to metrics such as performance and reliability going down with usage and hence decreasing their value for money. The deterioration of the EV two-wheeler points to key area to exploit which would be the reliability of the vehicle for longer durations. Obtaining consistency should high on the agenda when engineering our vehicle.

Users who enjoyed their vehicle tended to not own the vehicle. Hence perhaps a rental business model for EV two-wheelers should be explored further. Visual appeal was a standout feature and something to pay attention to when engineering the outer design of the vehicle. Lastly, the EV two-wheeler should be promoted as an all-purpose vehicle, targeting the users who enjoyed the product the most.

### Suitable Early Market Strategy

To analyse the suitability of locations in India to enter the EV early market we refer to the Innovation Adoption Life Cycle. The Technology Adoption Life Cycle is a sociological model that describes the adoption or acceptance of a new product or innovation, according to the demographic and psychological characteristics of defined adopter groups. *(Wikipedia)*

The process of adoption over time can be illustrated as a classical normal distribution. As a major technological disruptor that is still building and defining its market, innovators and early adaptors become natural early EV market target demographics.

We defined demographic and psychological characteristics of our dataset they were:

a. Behavioural: ‘Value of money’, ‘Visual Appeal’, ‘Reliability’, ‘Performance’

b. Psychological: Quality over quantity,

Big cities with high levels of pollution with a young, educated and informed crowd should be targeted

## 3.2 Behavioural segmentation – Electric Four-Wheeler

**Introduction:** In this section, we analyse the performance and segmentation of three electric car models: Tata Nexon EV, Hyundai Kona, and Tata Tigor EV. The data includes user ratings on various attributes such as exterior, comfort, performance, fuel economy, and value for money. The electric vehicle (EV) market is rapidly evolving, with consumer preferences becoming increasingly complex. Understanding these preferences is crucial for manufacturers to enhance product offerings and for marketers to create targeted campaigns. This report evaluates the key attributes of three EV models and segments the consumer market based on the data collected.

**Key attributes –**

1. **Exterior**: The design and build quality of the car's outer appearance, including style and materials.
2. **Comfort**: The level of convenience and ease inside the cabin, including seating and ride quality.
3. **Performance**: How well the car drives, including acceleration, handling, and stability.
4. **Fuel Economy**: The efficiency of fuel use, crucial for cost-saving and environmental impact.
5. **Value for Money:** The overall worth based on features, performance, and cost relative to similar cars.

**Libraries Used:**

1. **Pandas** (**pandas**) - for data manipulation and analysis.
2. **NumPy** (**numpy**) - for numerical operations.
3. **Scikit-learn** (**sklearn**) - for machine learning tasks, including preprocessing (like **LabelEncoder**), decomposition (**PCA**), and clustering (**KMeans**).
4. **Matplotlib** and **Seaborn** - for data visualization.
5. **Plotly** (if used) - for interactive plots.

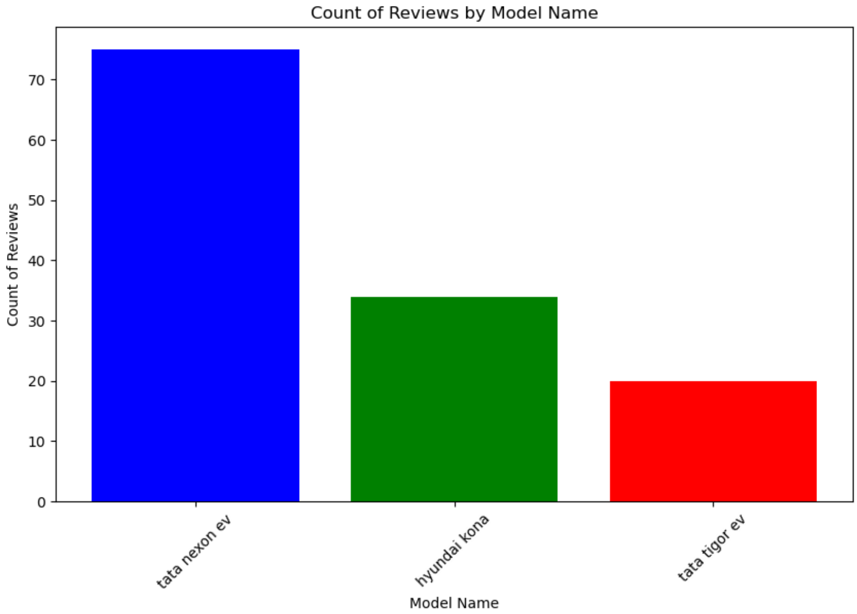
**Data Preprocessing for Dendrogram (Hierarchical Clustering) & Tree Formation (Decision Trees)**

1. **Data Cleaning**: Remove or impute missing values since hierarchical clustering algorithms can be sensitive to such gaps in data.
2. **Feature Selection**: Choose relevant features that contribute to the grouping of data points.
3. **Data Encoding**: Convert categorical variables into a numerical format if present, as hierarchical clustering algorithms require numerical input.

4. **Data Normalization/Standardization**: Normalize or standardize the data to ensure that all features contribute equally to the distance calculations. This is crucial because hierarchical clustering uses distance metrics (like Euclidean distance) to determine the similarity between data points.

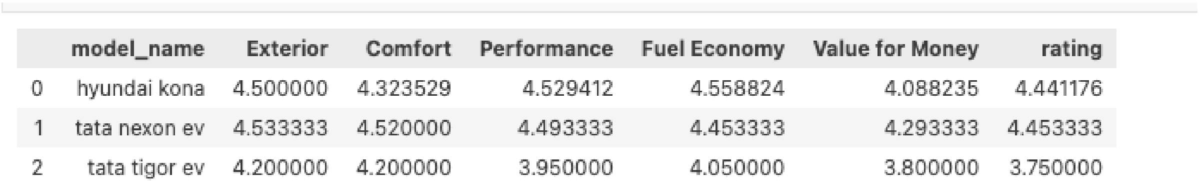
**Overview of Car Models:** The study encompasses three prominent EV models:

- **Tata Nexon EV:** The most reviewed model with 75 counts.
- **Hyundai Kona:** Mid-range in terms of review counts, totalling 34.
- **Tata Tigor EV:** The least reviewed with 20 counts.

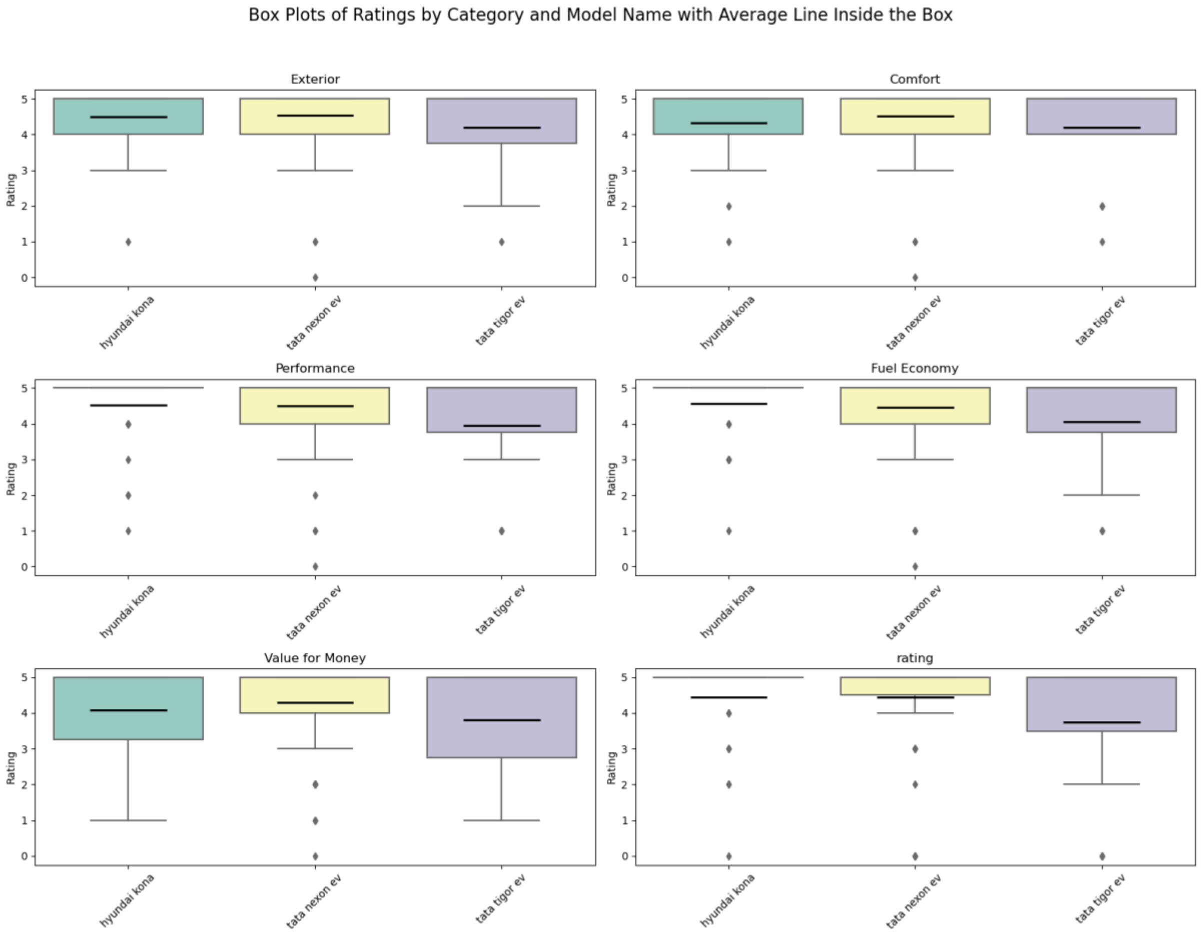


**Model Performance:**

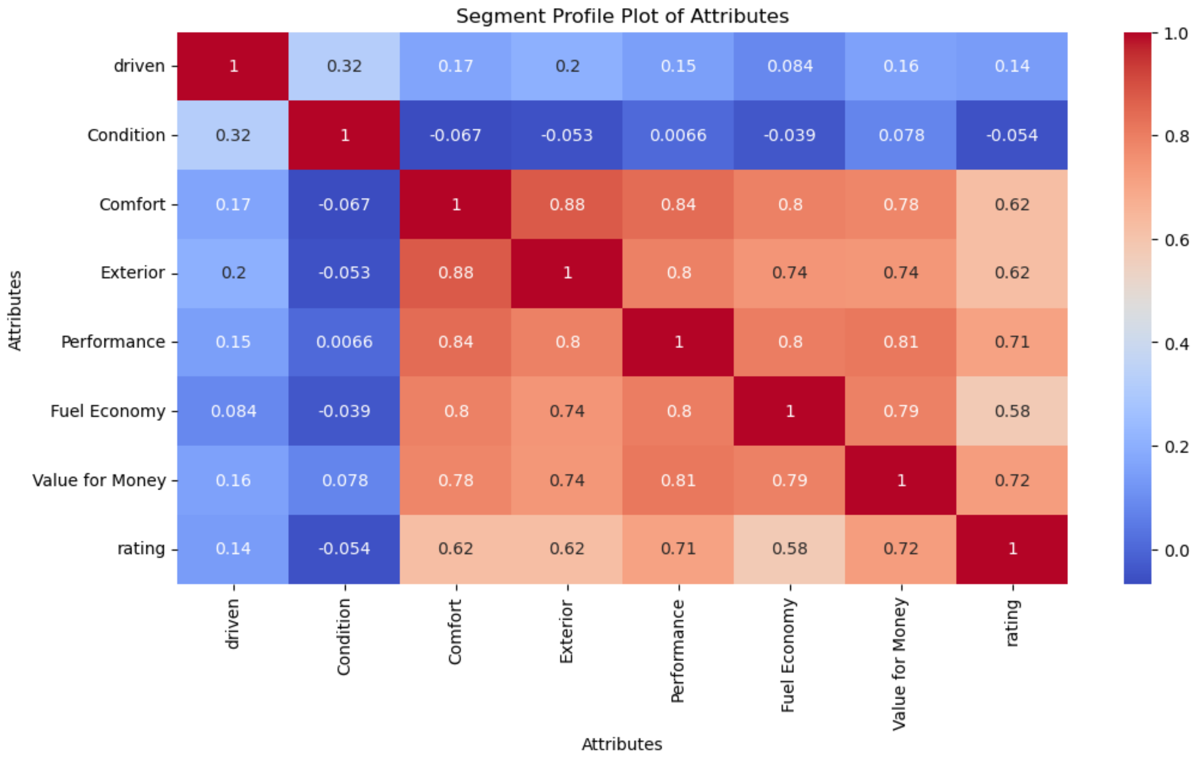
The table provides the mean ratings for each car model across different attributes. Hyundai Kona received high ratings in all aspects, especially in performance and fuel economy. Tata Nexon EV also performed well across most attributes, while Tata Tigor EV received comparatively lower ratings.



| | model_name | Exterior | Comfort | Performance | Fuel Economy | Value for Money | rating |
|---|---|---|---|---|---|---|---|
| 0 | hyundai kona | 4.500000 | 4.323529 | 4.529412 | 4.558824 | 4.088235 | 4.441176 |
| 1 | tata nexon ev | 4.533333 | 4.520000 | 4.493333 | 4.453333 | 4.293333 | 4.453333 |
| 2 | tata tigor ev | 4.200000 | 4.200000 | 3.950000 | 4.050000 | 3.800000 | 3.750000 |

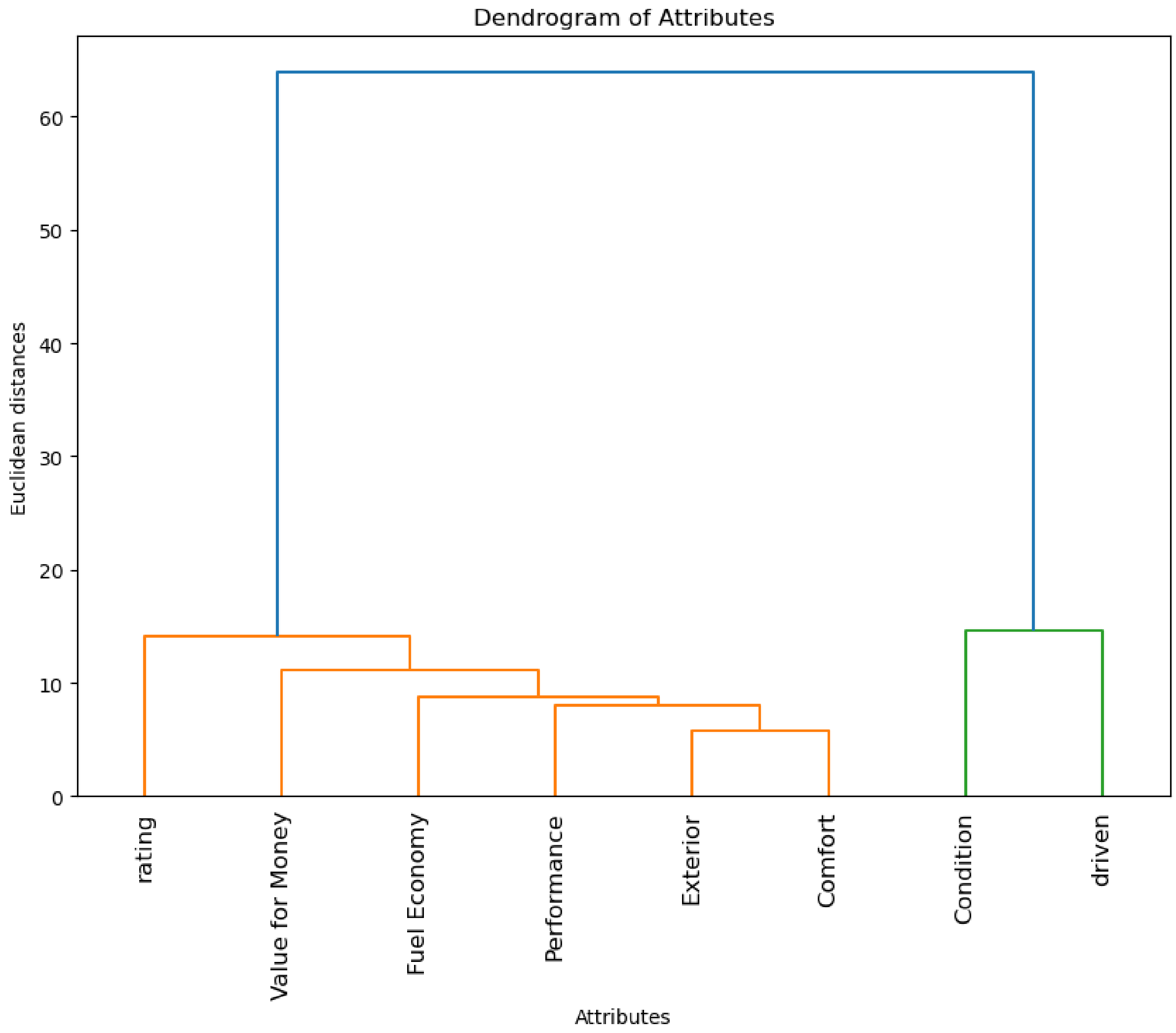


Mean ratings for each model are calculated to assess performance across six attributes: Exterior, Comfort, Performance, Fuel Economy, Value for Money, and Overall Rating. Tata Nexon EV, leads in most attributes, closely followed by Hyundai Kona suggesting strong market performance. Tata Tigor EV lags behind, indicating potential areas for improvement.



## Correlation Matrix Heatmap Conclusions:

1. **High Correlation Pairs**: There are several pairs of attributes with a high degree of correlation (close to 1). Notably, Comfort and Exterior, Performance and Exterior, and Performance and Comfort show a very strong positive relationship, with coefficients above 0.8. This indicates that these attributes tend to receive similar evaluations, suggesting that improvements or deficits in one are likely to affect the perception of the others.
2. **Moderate Correlation**: The overall rating of the product or service has a moderate correlation with most attributes, except 'driven' and 'condition'. This suggests that while the rating is somewhat influenced by these attributes, there are other factors at play that determine the overall satisfaction or rating.
3. **Low Correlation**: The 'driven' and 'condition' attributes have low correlations with the rest, implying they are considered somewhat independently when evaluating the product or service.
4. **Independence of 'driven' Attribute**: The attribute 'driven' is the least correlated with the overall rating, which may indicate that the frequency of use or the extent to which a product has been used does not significantly impact the overall satisfaction or perceived quality.



**Dendrogram Conclusions:**

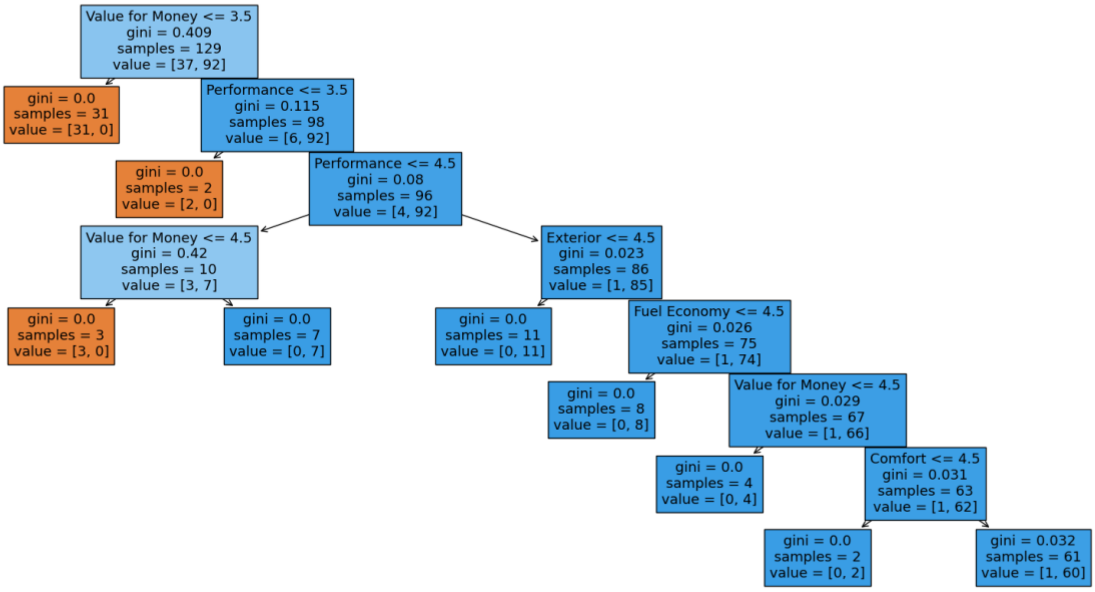
1. **Cluster Formation**: The dendrogram shows how attributes are grouped based on their similarity. The attributes that are most similar are joined together at the lower distances, and as we move up the y-axis, the less similar attributes are joined.
2. **Major Clusters**: There are two main clusters. The first includes Rating, Value for Money, Fuel Economy, Performance, and Exterior. This cluster suggests that these attributes are perceived similarly by consumers or have a strong relationship in how they influence each other.
3. **Distinct Attributes**: The 'driven' attribute stands alone with a significant distance from the 'condition' cluster. This reinforces the conclusion from the heatmap that 'driven' is perceived quite differently from the other attributes.
4. **Implications for Product Strategy**: The clustering of attributes like Fuel Economy, Performance, and Exterior suggests that these might be key areas to focus on for improving product ratings and customer satisfaction. 'Driven' may

require separate consideration or targeted strategies as it does not cluster tightly with other attributes.

**Decision tree** this particular tree is classifying samples based on several features:

- Value for Money,
- Performance,
- Exterior,
- Fuel Economy,
- and Comfort.

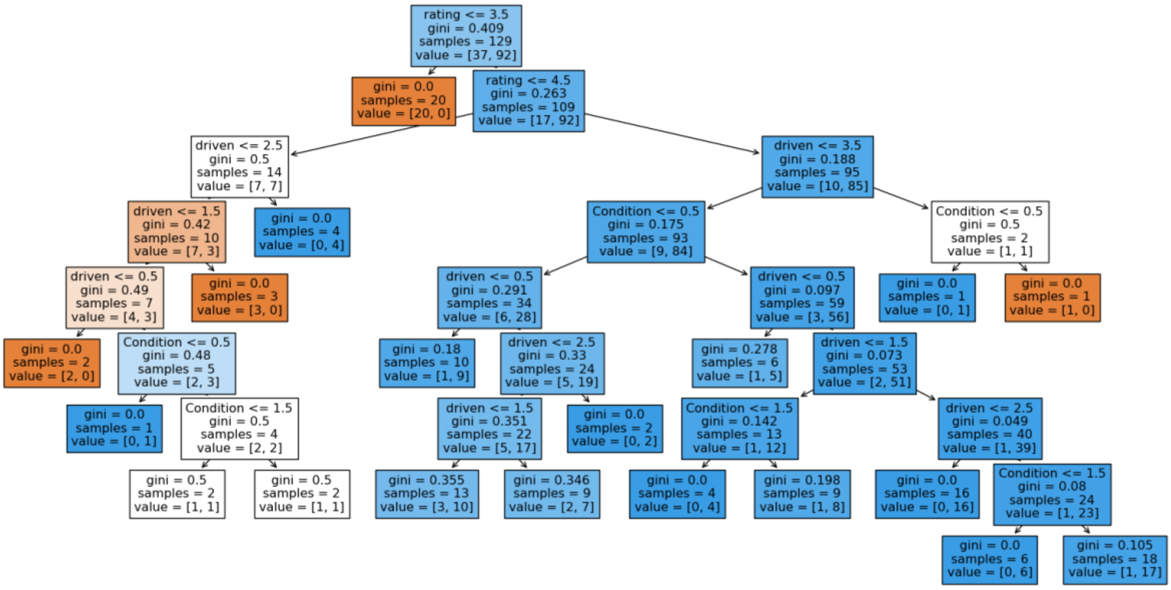
The tree uses the Gini impurity index as a measure of the quality of the splits; a Gini index of 0 represents a perfect separation of classes.



**Conclusion:**

- The feature **'Value for Money'** is the most critical factor in determining the classification of samples in this decision tree model, which suggests that it might be the most significant predictor of the target variable.
- When 'Value for Money' is 3.5 or less, the model predicts a single class with high confidence (31 samples are classified in one class with no impurity).

- **'Performance'** is the next most critical feature, further splitting the dataset with a low Gini index, suggesting that performance ratings are also a strong predictor after considering the value for money.
- Features **'Exterior'**, 'Fuel **Economy'**, and **'Comfort'** also contribute to classification but are considered after 'Value for Money' and 'Performance', indicating their secondary importance in the predictive model.


rating <= 3.5 gini = 0.409 samples = 129 value = [37, 92]
gini = 0.0 samples = 20 value = [20, 0]
rating <= 4.5 gini = 0.263 samples = 109 value = [17, 92]
driven <= 2.5 gini = 0.5 samples = 14 value = [7, 7]
driven <= 3.5 gini = 0.188 samples = 95 value = [10, 85]
driven <= 1.5 gini = 0.42 samples = 10 value = [7, 3]
gini = 0.0 samples = 4 value = [0, 4]
Condition <= 0.5 gini = 0.175 samples = 93 value = [9, 84]
Condition <= 0.5 gini = 0.5 samples = 2 value = [1, 1]
driven <= 0.5 gini = 0.49 samples = 7 value = [4, 3]
gini = 0.0 samples = 3 value = [3, 0]
driven <= 0.5 gini = 0.291 samples = 34 value = [6, 28]
driven <= 0.5 gini = 0.097 samples = 59 value = [3, 56]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 1 value = [1, 0]
gini = 0.0 samples = 2 value = [2, 0]
Condition <= 0.5 gini = 0.48 samples = 5 value = [2, 3]
gini = 0.18 samples = 10 value = [1, 9]
driven <= 2.5 gini = 0.33 samples = 24 value = [5, 19]
gini = 0.278 samples = 6 value = [1, 5]
driven <= 1.5 gini = 0.073 samples = 53 value = [2, 51]
gini = 0.0 samples = 1 value = [0, 1]
Condition <= 1.5 gini = 0.5 samples = 4 value = [2, 2]
driven <= 1.5 gini = 0.351 samples = 22 value = [5, 17]
gini = 0.0 samples = 2 value = [0, 2]
Condition <= 1.5 gini = 0.142 samples = 13 value = [1, 12]
driven <= 2.5 gini = 0.049 samples = 40 value = [1, 39]
gini = 0.5 samples = 2 value = [1, 1]
gini = 0.5 samples = 2 value = [1, 1]
gini = 0.355 samples = 13 value = [3, 10]
gini = 0.346 samples = 9 value = [2, 7]
gini = 0.0 samples = 4 value = [0, 4]
gini = 0.198 samples = 9 value = [1, 8]
gini = 0.0 samples = 16 value = [0, 16]
Condition <= 1.5 gini = 0.08 samples = 24 value = [1, 23]
gini = 0.0 samples = 6 value = [0, 6]
gini = 0.105 samples = 18 value = [1, 17]

**Conclusions:**

- **'Rating'** is the most significant predictor for splitting the data, as it is the first division in the decision tree.
- **'Driven'** is the second most used feature for making decisions within the tree, suggesting that it has a considerable impact on the outcome after 'rating'.
- **'Condition'** appears as a splitting attribute in subsequent levels, especially after the dataset has already been split by 'rating' and 'driven', indicating its relevance in finer categorizations.

## 4. Vehicular Feature Segmentation

Market segmentation is an important strategy for emerging markets to study and deploy growing transportation technologies such as electric vehicles (EVs) to achieve widespread acceptance. EV usage is predicted to skyrocket shortly as a low-emission and low-cost vehicle, sparking a significant amount of future academic study interest. The primary goal of this study is to investigate and identify unique sets of possible EV buyer groups based on psychographic, behavioral, and socioeconomic characteristics, using an integrated research framework of 'perceived benefits-attitude-intention'.

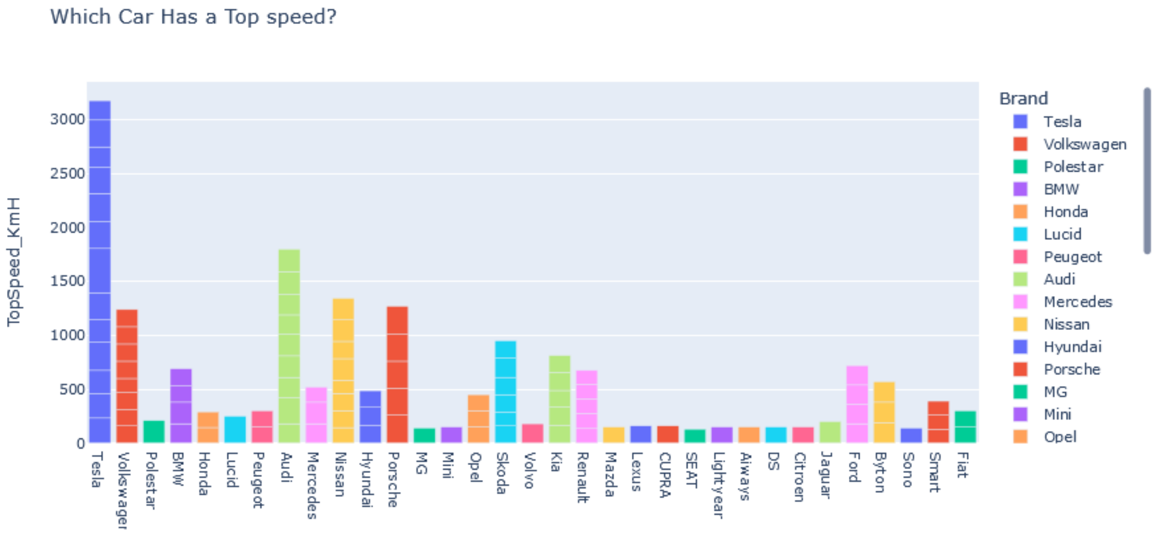
Libraries:

- **SKLearn:** Simple and efficient tools for predictive data analysis
- **Seaborn:** Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **Plotly:** The Plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering various statistical, financial, geographic, scientific, and 3-dimensional use cases.
- **Matplotlib :** Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
- **Numpy:** Caffe-based Single Shot-Multibox Detector (SSD) model used to detect faces
- **Pandas:** pandas is a fast, powerful, flexible, and easy-to-use open-source data analysis and manipulation tool, built on top of the Python programming language.
- Statsmodels: Python library that lets you analyze data, estimate statistical models, and run statistical tests. There is a comprehensive set of descriptive statistics, statistical tests, charting routines, and outcome statistics for each type of data and estimator.

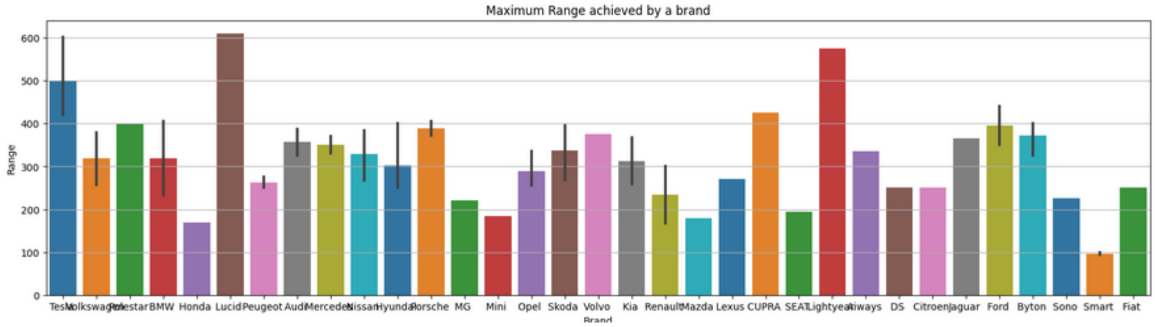
Data Preprocessing: Data pre-processing is a critical stage in the development of a machine learning model. Initially, data may not be clean or in the appropriate format for the model, resulting in misleading results. Data pre-processing involves transforming data into the format that we require. It is used to deal with noise, duplication, and missing values in a dataset. Data pre-processing operations include importing, dividing, scaling, and scaling attributes. Preprocessing of data is essential to improve model accuracy.

Download: dataset here

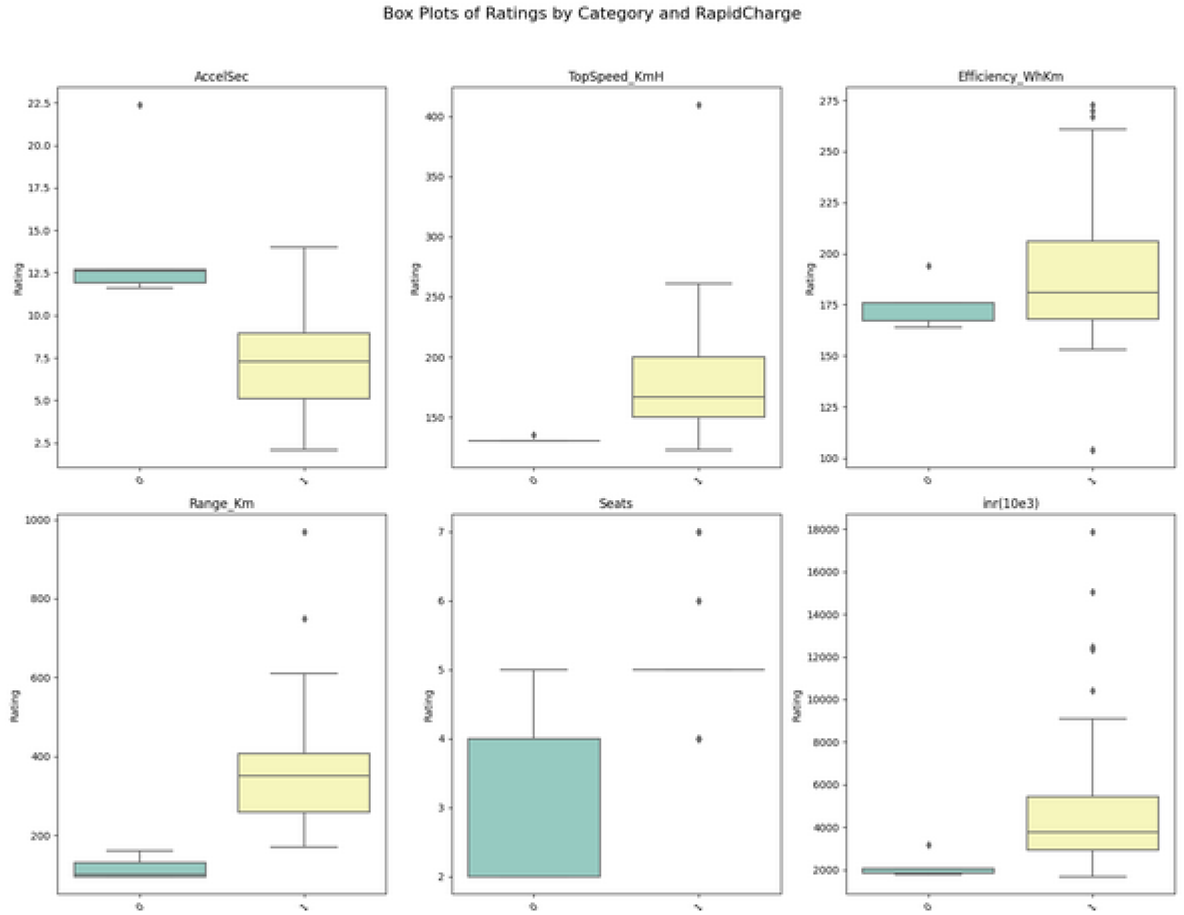
Results:



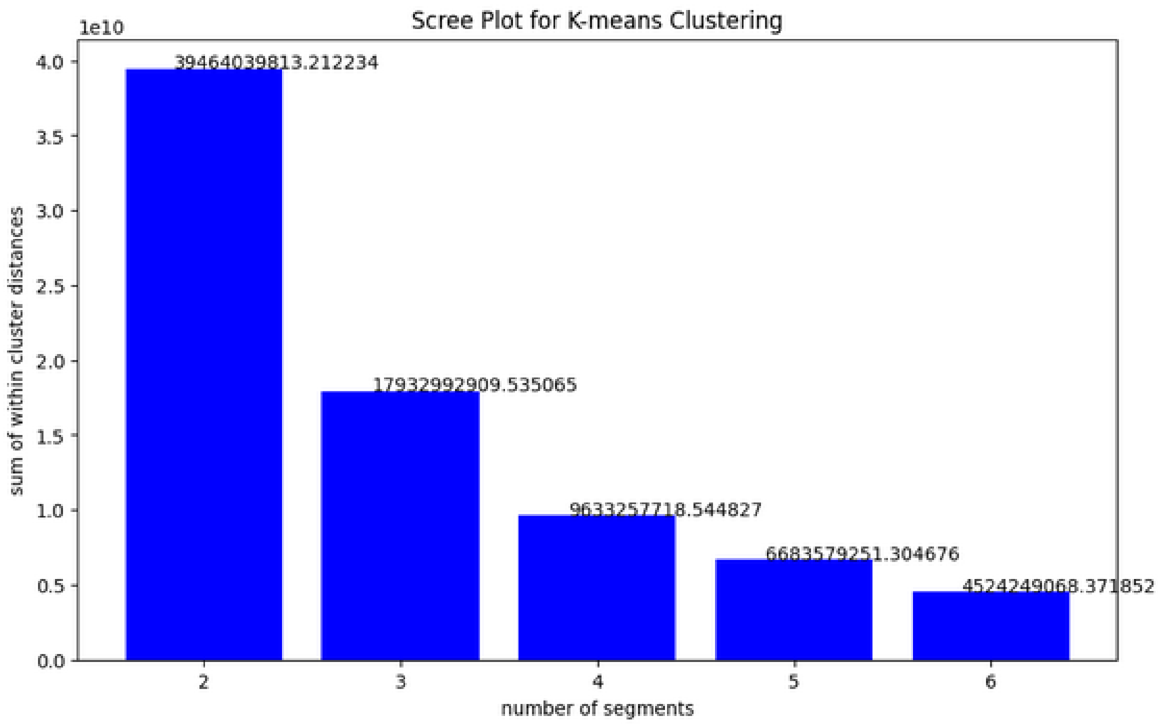
There are different types of brands and models in the dataset, the above visual helps us in finding the top speed with respect to the brand to give appropriate output. In the above chart, Tesla is on the top with high speed following with Audi and Volkswagen. This brand in market has huge demand for its speed.



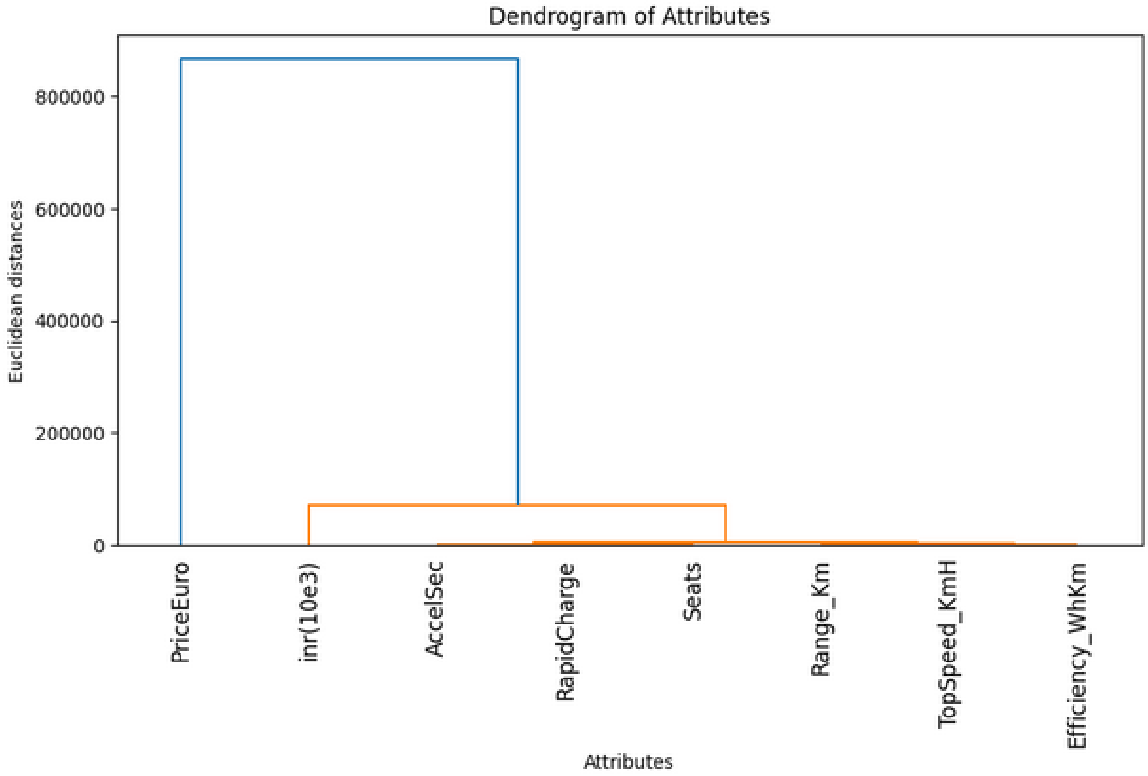
Each brand vehicle has some range, based on the mileage of the brand an electric vehicle can be sold. So, here the Lucid has around 600+ range followed by Light year which is 590, and then Tesla. Finding the maximum range of the brands of electric vehicles will know the high mileage for future data to calculate which brand vehicle is better.



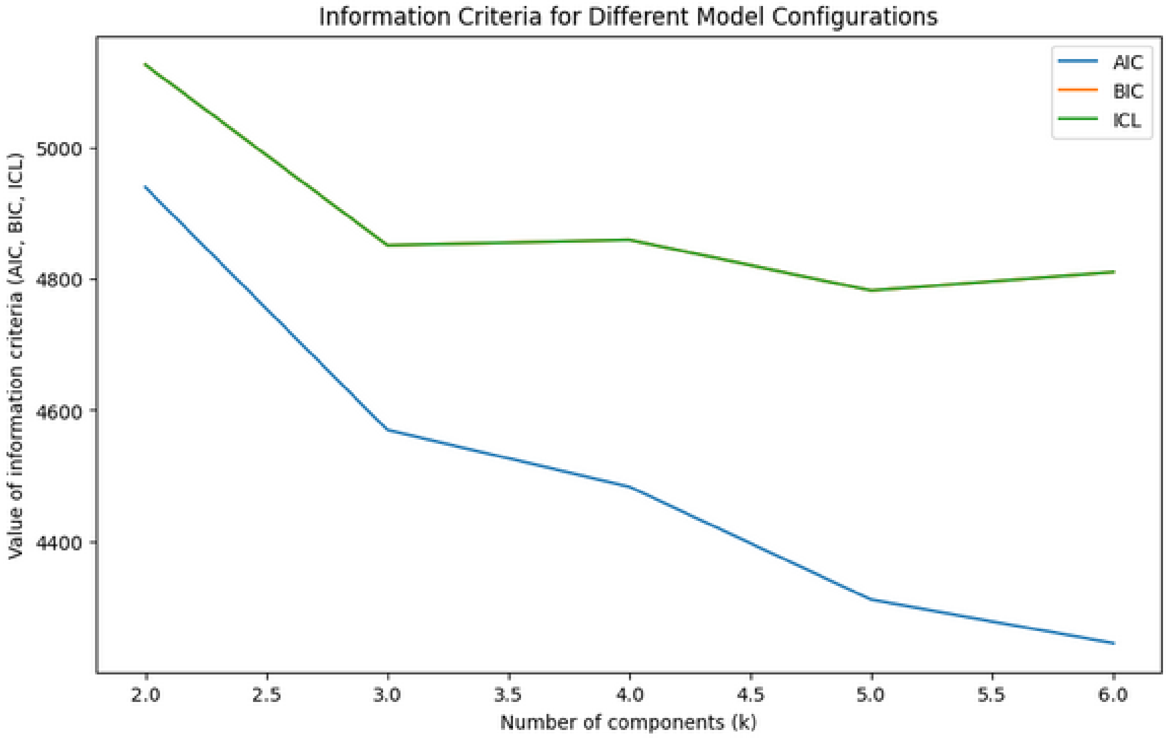
Based on Rapid Charge the following columns show the performance, the columns are AccelSec, TopSpeed, Efficiency, Range, Seats, and Price. As observed the Rapid Charge is high for most of the following data but least at seats.



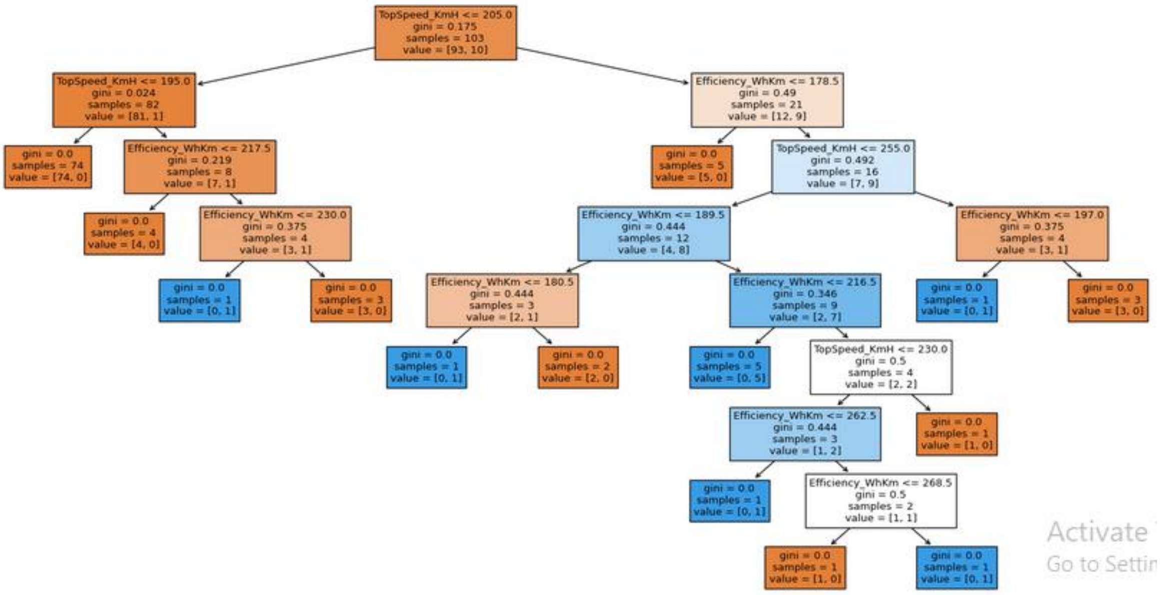
KMeans clustering on a dataset, assesses the stability of the clustering, and analyzes the resulting clusters. The results are stored in various data structures for further analysis and comparison. The scree plot helps identify the "elbow" or the point where the decrease in inertia becomes less significant, indicating that adding more clusters may not significantly improve the clustering quality.

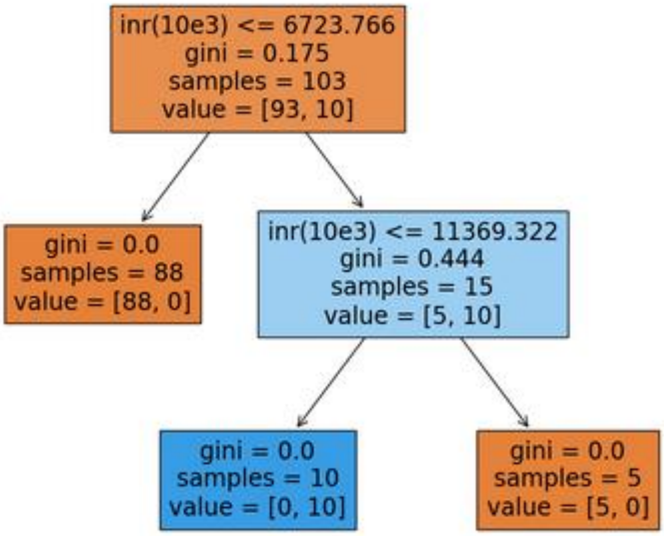


In the above visual we see there are two major clusters they are PriceEuro and other involves AccelSec, RapidCharge, Seats, Range_km, Top Speed_kmH,Efficieny_whkm.The attributes in x-axis and Euclidean distance in y-axis, where the major part is divided by inr and seats, further seats are divided into 5 subdivisions they are accel sec, rapid-charge, range_km,topspeed_kmh, and efficiency. The distinct attribute is the price of Euro as it's just a conversion of price into euro. This shows how the Dendrogram has created a relationship between clusters and attributes.



Information criteria for the mixture models of binary distributions with 2 to 6 components(segments) for the Electric Vehicle data set. The information criteria values AIC, BIC and ICL on the *y*-axis for the different number of components (segments) on the *x*-axis. As can be seen, the values of all information criteria decrease quite dramatically until three components (market segments) are reached. If the information criteria are strictly applied based on statistical inference theory, the ICL recommends – by a small margin – the extraction of six market segments. The BIC also points to six market segments. The AIC values decrease beyond six market segments, indicating that at least seven components are required to suitably fit the data. Three market segments might be a good solution if a more pragmatic point of view is taken; this is the point at which the decrease in the information criteria flattens visibly.


TopSpeed_KmH <= 205.0
gini = 0.175
samples = 103
value = [93, 10]
TopSpeed_KmH <= 195.0
gini = 0.024
samples = 82
value = [81, 1]
Efficiency_WhKm <= 178.5
gini = 0.49
samples = 21
value = [12, 9]
gini = 0.0
samples = 74
value = [74, 0]
Efficiency_WhKm <= 217.5
gini = 0.219
samples = 8
value = [7, 1]
gini = 0.0
samples = 5
value = [5, 0]
TopSpeed_KmH <= 255.0
gini = 0.492
samples = 16
value = [7, 9]
gini = 0.0
samples = 4
value = [4, 0]
Efficiency_WhKm <= 230.0
gini = 0.375
samples = 4
value = [3, 1]
Efficiency_WhKm <= 189.5
gini = 0.444
samples = 12
value = [4, 8]
Efficiency_WhKm <= 197.0
gini = 0.375
samples = 4
value = [3, 1]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 3
value = [3, 0]
Efficiency_WhKm <= 180.5
gini = 0.444
samples = 3
value = [2, 1]
Efficiency_WhKm <= 216.5
gini = 0.346
samples = 9
value = [2, 7]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 3
value = [3, 0]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 2
value = [2, 0]
gini = 0.0
samples = 5
value = [0, 5]
TopSpeed_KmH <= 230.0
gini = 0.5
samples = 4
value = [2, 2]
Efficiency_WhKm <= 262.5
gini = 0.444
samples = 3
value = [1, 2]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Efficiency_WhKm <= 268.5
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
Activate
Go to Settir


inr(10e3) <= 6723.766
gini = 0.175
samples = 103
value = [93, 10]
gini = 0.0
samples = 88
value = [88, 0]
inr(10e3) <= 11369.322
gini = 0.444
samples = 15
value = [5, 10]
gini = 0.0
samples = 10
value = [0, 10]
gini = 0.0
samples = 5
value = [5, 0]

**Conclusion**:

- The first division in the decision tree, "rapid charge," is the most significant predictor for data splitting.
- After 'seats' is the second most used element in decision-making within the tree, indicating its significant impact on outcomes.
- 'Range_Km' is a splitting property that appears after ‘rapid charge ‘ and 'driven', showing its relevance in finer categorizations.

The highest speed is Tesla (410), high-frequency Byton. Maximum Range Lucid and light-year, Plug type 2 CCS, body shape is hatchback and SUV, preferred seats are 5, the rapid charge is excellent in efficiency. The clusters are separated into six segments. The optimal number of clusters in terms of stability: 2. Stability score.

Decision tree This tree classifies samples based on a few features:

- Top speed (km/h) • Range (km) • Efficiency (Wh/km) • Price (Euro) • INR (10e3)

The tree measures the quality of the splits using the Gini impurity index; a Gini index of 0 denotes a complete separation of classes.

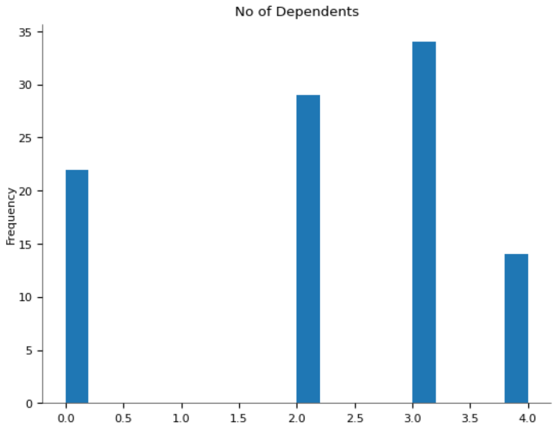
- The feature 'TopSpeed_KmH' is the most important factor in deciding sample categorization in this decision tree model, indicating that it may be the most relevant predictor.

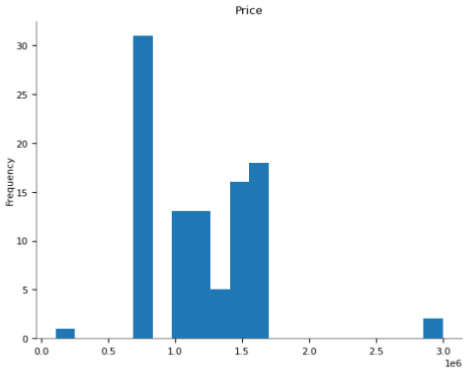
## 5. Demographic Segmentation

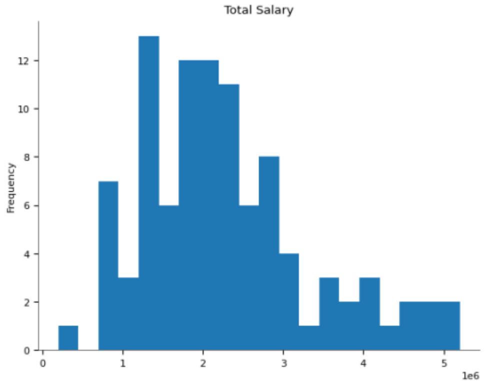
**ABOUT THE DATASET :-**

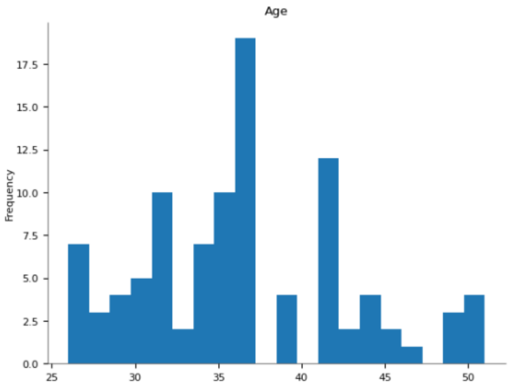
- CONSISTS OF EIGHT ATTRIBUTES.
- NAMELY -> Age, Profession, Marital Status, Education, No of Dependents, Personal loan, Total salary, Price.
- Datatypes -> (Age, No of Dependents, Price, Total Salary) : integer & (Profession, Marital status, Education, Personal loan): object data type
- No empty values in between.
- Modules used: NumPy, pandas, matplotlib, seaborn
- Having 99 rows and 8 columns.

**NUMERICAL DISTRIBUTIONS :-**

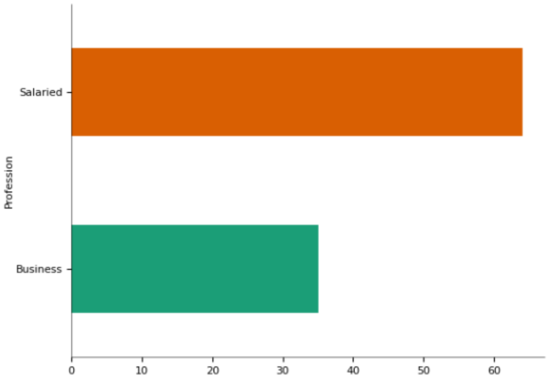


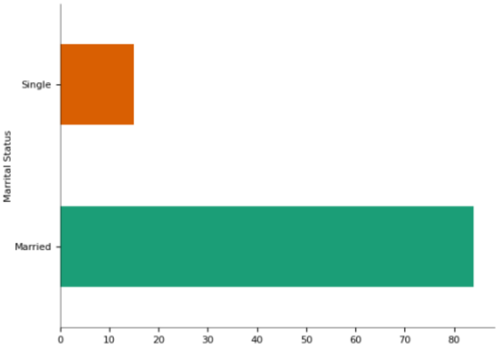


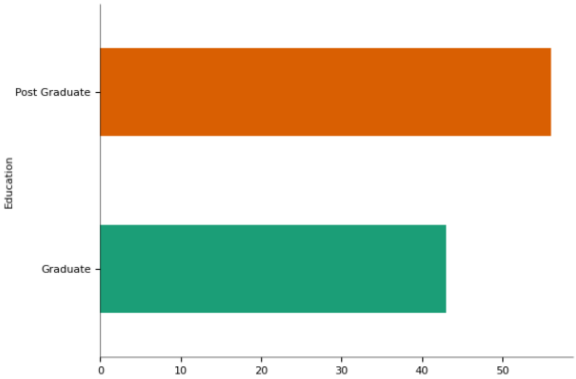


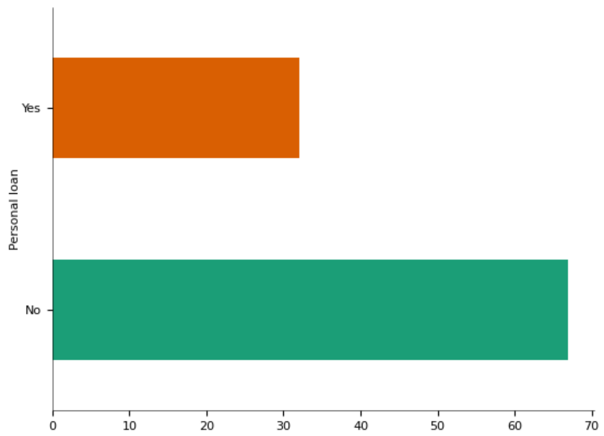


## CATEGORICAL DISTRIBUTIONS:-

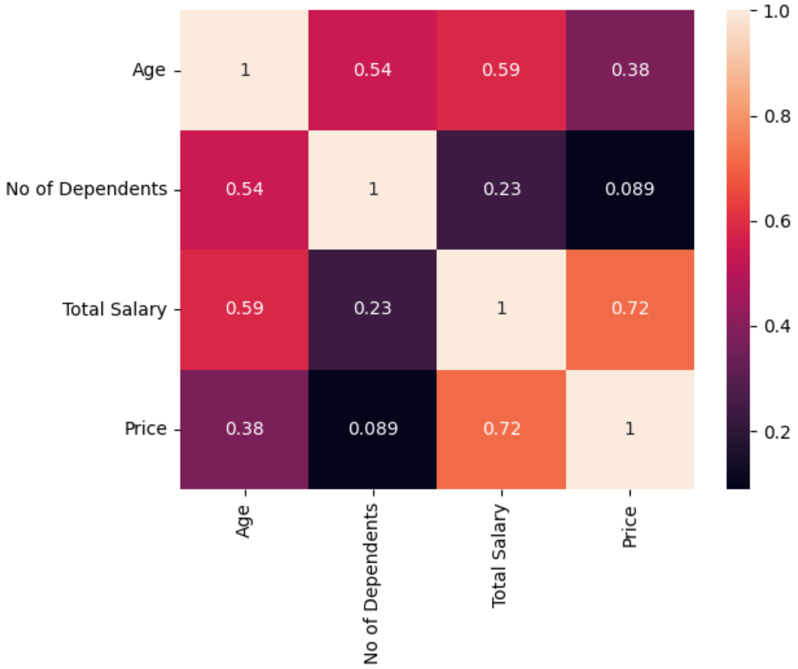








## CORRELATION BETWEEN NUMERICAL ATTRIBUTE:-



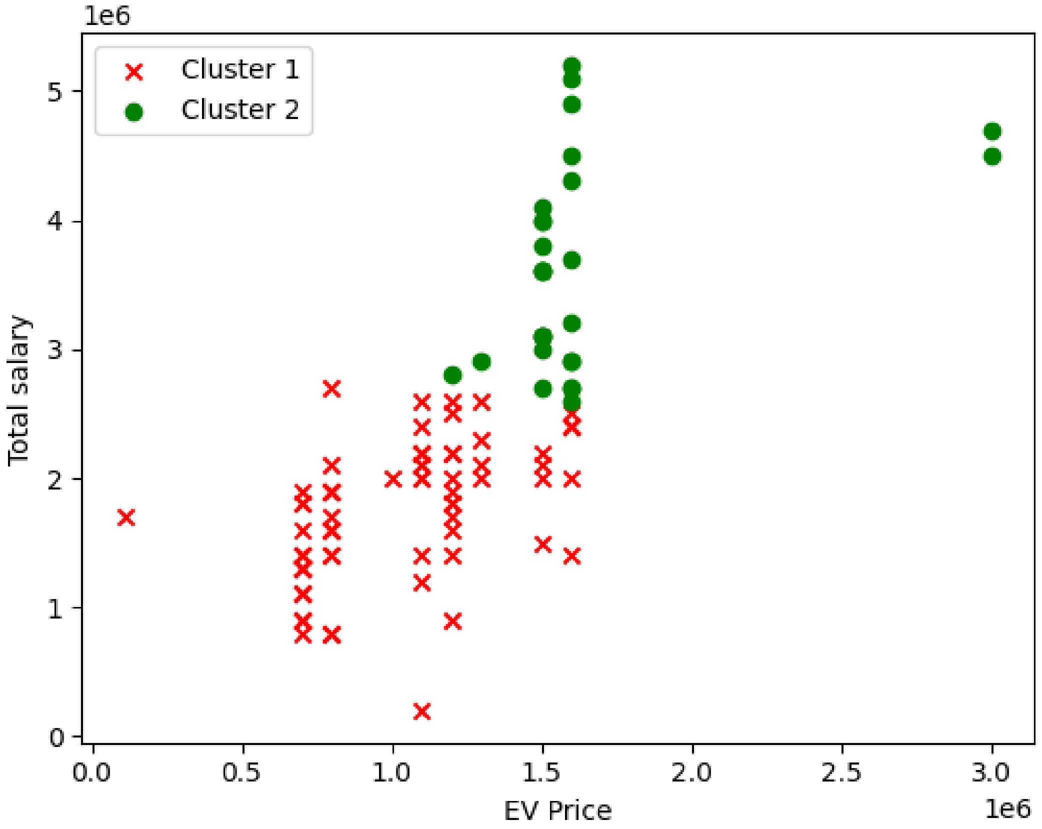
- Here,high correlation between Salary & Price.

[{"index":"count","Age":"99.0","No of Dependents":"99.0","Total Salary":"99.0","Price":"99.0"},{"index":"mean","Age":"36.313131313131315"," No of Dependents":"2.1818181818181817","Total Salary":"2270707.0707070706","Price":"1194040.4040404041"},{"index":"std","Age":"6.246054206941395","No of Dependents":"1.3352645259872389","Total Salary":"1050777.4114525313","Price":"437695.54042274354"},{"index":"min","Age":"26.0","No of Dependents":"0.0","Total Salary":"200000.0","Price":"110000.0"},{"index":"25%","Age":"31.0","No of Dependents":"2.0","Total Salary":"1550000.0","Price":"800000.0"},{"index":"50%","Age":"36.0","No of Dependents":"2.0","Total Salary":"2100000.0","Price":"1200000.0"},{"index":"75%","Age":"41.0","No of Dependents":"3.0","Total Salary":"2700000.0","Price":"1500000.0"},{"index":"max","Age":"51.0","No of Dependents":"4.0","Total Salary":"5200000.0","Price":"3000000.0"}]

- There, are earning professionals like businessmen, and salaried., having other dependent members in their families (0-4).
- All of them are Educated(Graduate or Post-Grad), having age from 26 yrs to 51 yrs.
- Salaries ranging from 8 lpa to 51 lpa.

**SEGMENTATION:-**

- SCATTER ANALYSIS USING KPROTOTYPE.
- Changed the numerical features to float types.
- Performed analysis for Price and Salary earned.



**RESULT:-**

EV Price around 1.5 lack are bought much.

## 6. Report Conclusion

1. Develop an EV for the Indian market: The unique needs and requirements highlighted over the course of this report briefly outline the nature of that vehicle.

2. Understand your audience: The demographic is one that is educated and informed that a values for quality and reliability.

3. Limit cost: The large consumer base is one that is price sensitive. Limiting costs generally a good strategy.

4. Partner with the government: Addressing a nature-centric and renewability subject makes the EV market desirable from a governing perspective of sustainability. Obtaining grants, tax breaks and subsidies are ideal ways to support and gain momentum as a Strat-Up. On the other side, keeping up to date with safety, battery and sustainability regulations are a must.

5. Target fleet operators: Fleet operators, such as taxi companies and delivery services, are potential customers for EVs. Targeting fleet operators could help you to reach a large number of vehicles.

6. Rental Service: Something to consider as a large consumer base is very interested but perhaps not capable of affording.

7. Know the competition and build a competitive advantage

8. Funding: Private funding is necessary for a research intensive market such as the electric vehicles.

## Individual Github Repositories

Aryan: https://github.com/Aryan092/EV-Market-Segmentation

Bhaskar: https://github.com/bhaskar0402/feynnlab-EV-market

Prateek: https://github.com/PrateekKumar135/FEYN-LAB2

## References

[1] Technology adoption life cycle (2023a) Wikipedia. Available at: https://en.wikipedia.org/wiki/Technology_adoption_life_cycle

[2] Electric two-wheelers in India. Available at: https://aeee.in/wp-content/uploads/2022/07/ICA-AEEE-Whitepaper-2022.pdf

[3] Electric car vector art, icons, and graphics for free download Vecteezy. Available at: https://www.vecteezy.com/free-vector/electric-car

[4] Dolnicar, S. (2018) Market segmentation analysis. Springer Nature.