**Introduction:** In this report, 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:**

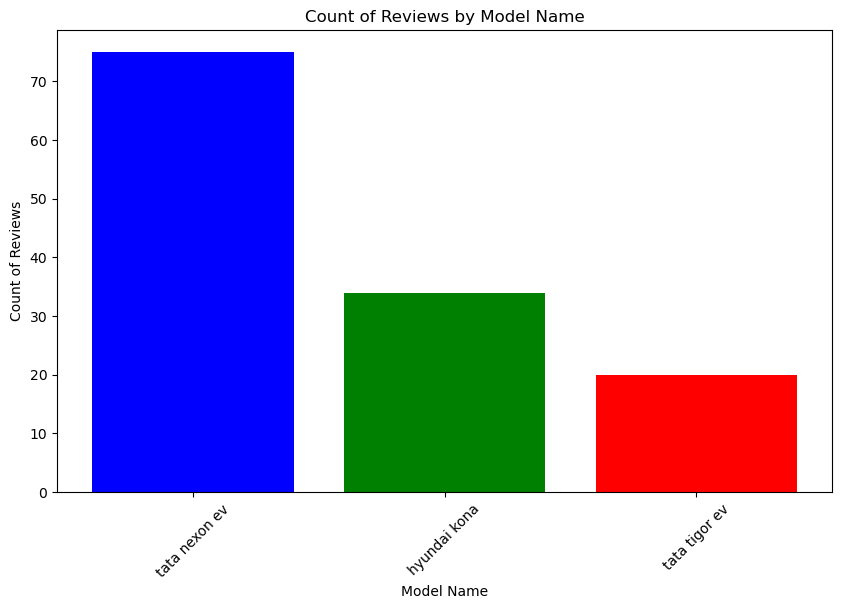
1. **Pandas** (**pandas**) - for data manipulation and analysis.
2. **NumPy** (**numpy**) - for numerical operations.
3. **Scikit-learn** (**sklearn**) - for machine learning tasks, including preprocessing (like **LabelEncoder**), decomposition (**PCA**), and clustering (**KMeans**).
4. **Matplotlib**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

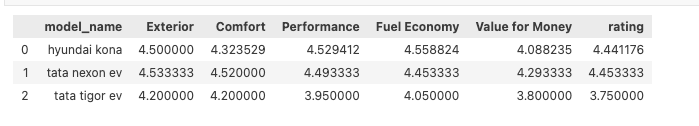
**Overview of Car Models:** The study encompasses three prominent EV models:

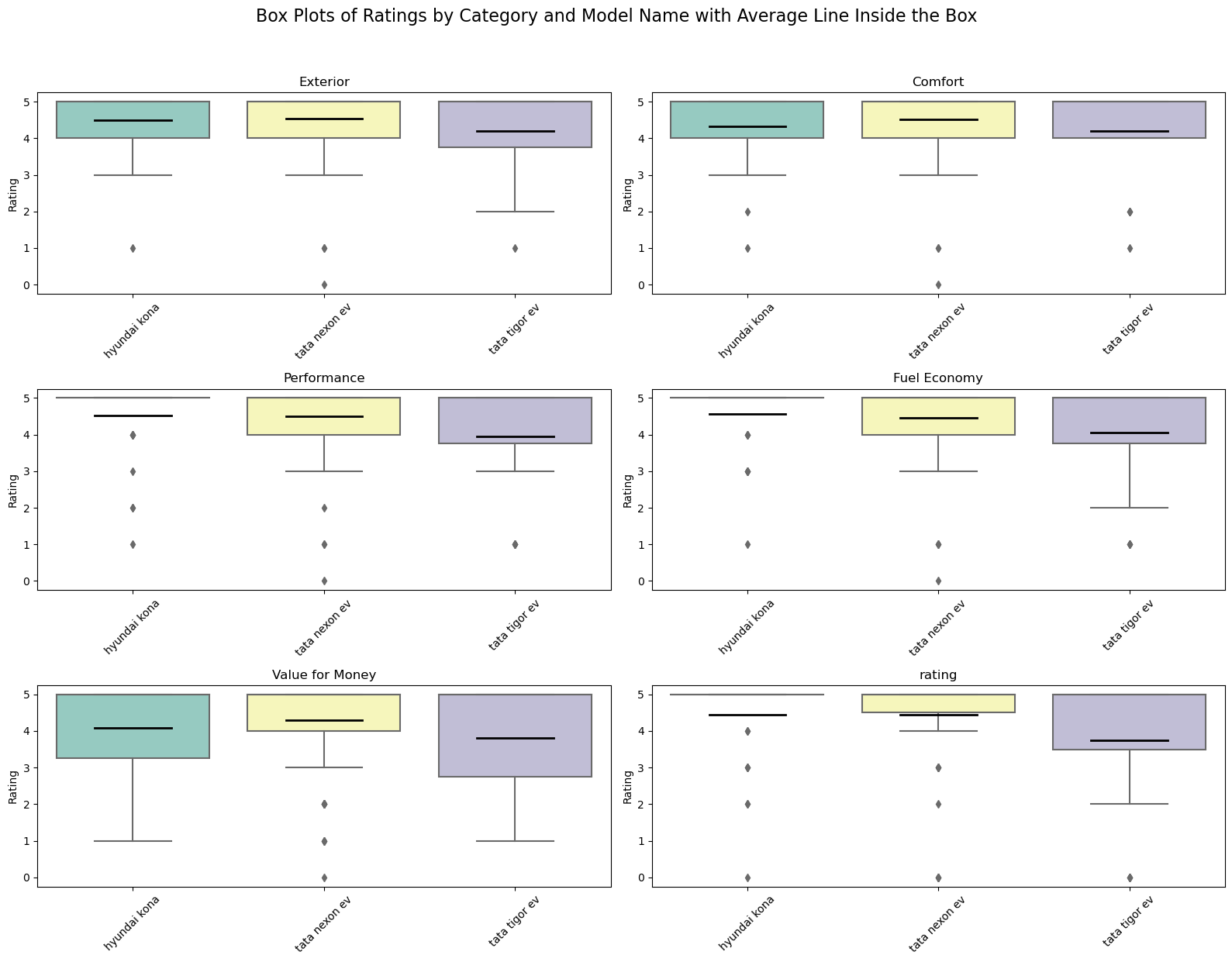
* **Tata Nexon EV:** The most reviewed model with 75 counts.
* **Hyundai Kona:** Mid-range in terms of review counts, totalling 34.
* **Tata Tigor EV:** The least reviewed with 20 counts.



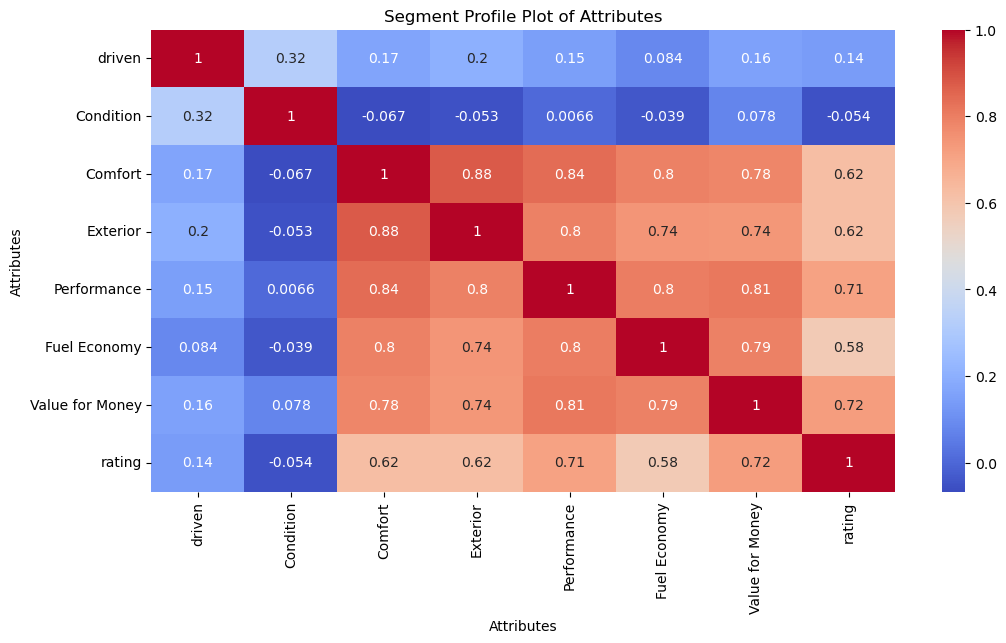
**Model Performance:**

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.



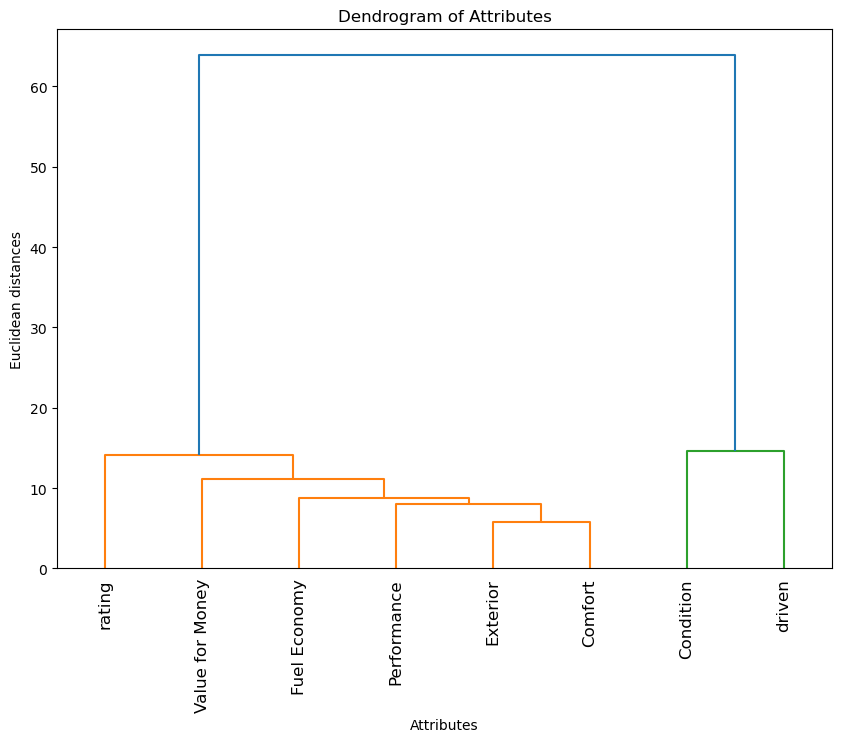
****

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.



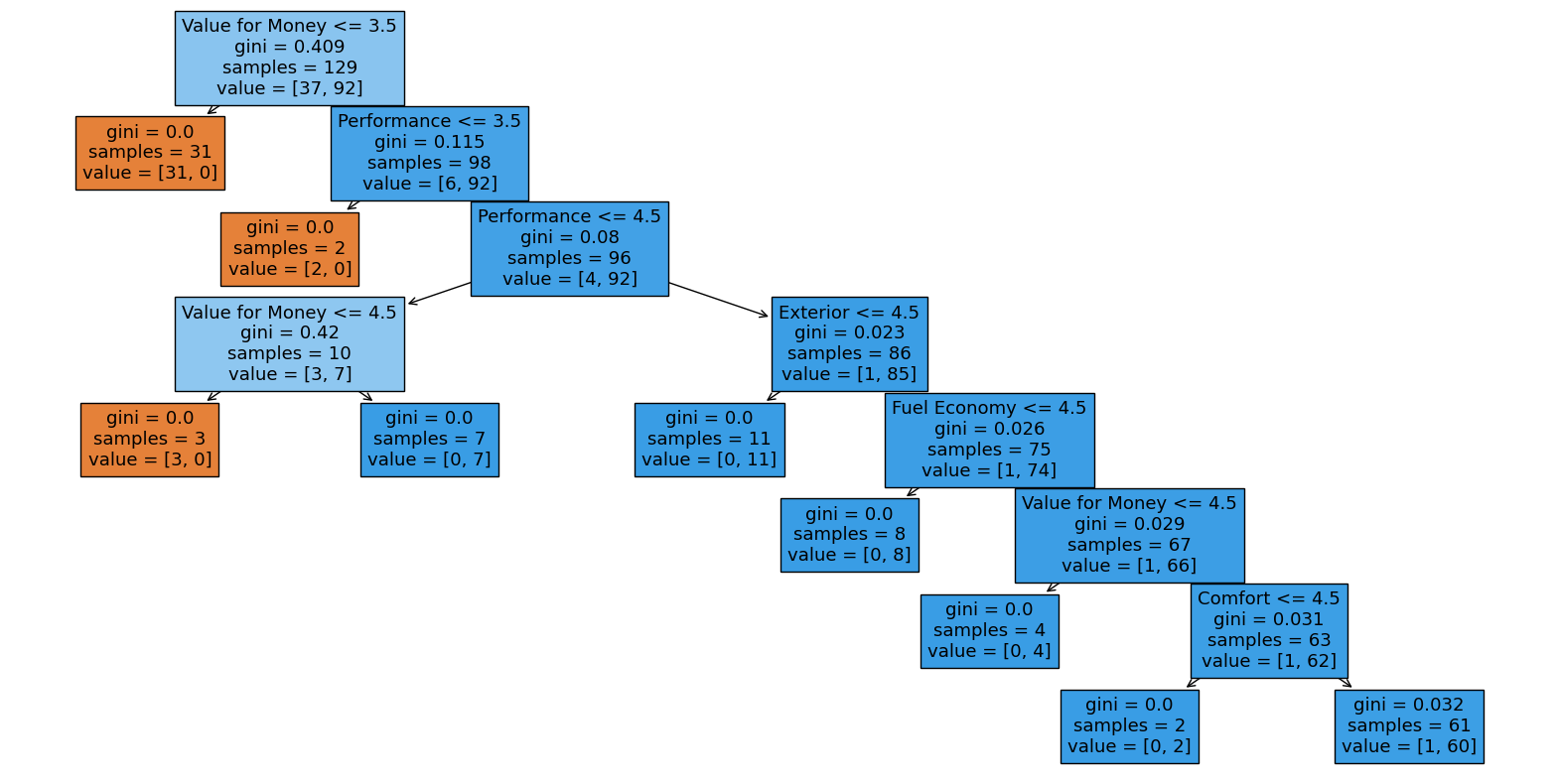
**Dendrogram Conclusions:**

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:

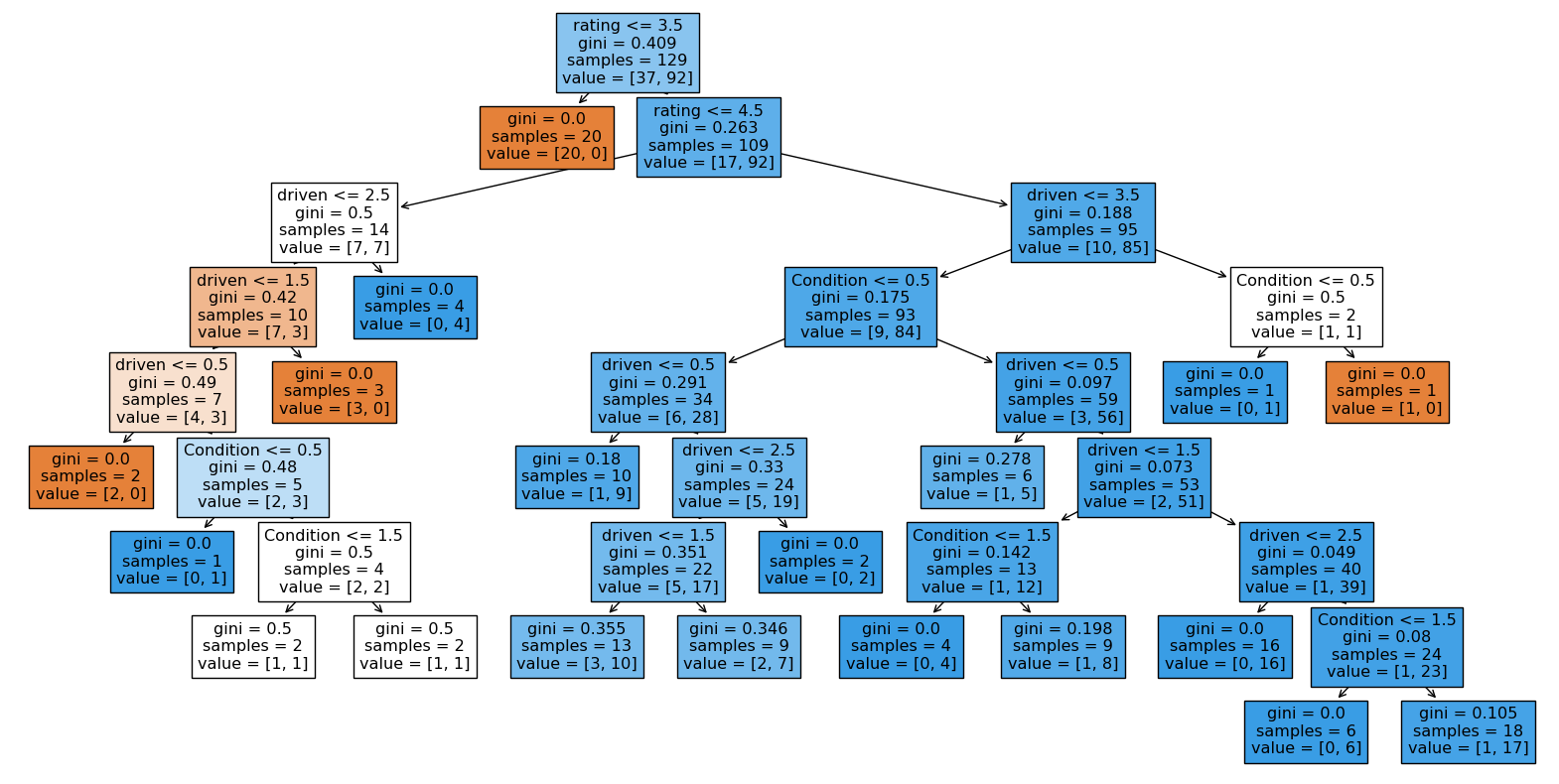
* Value for Money,
* Performance,
* Exterior,
* Fuel Economy,
* and Comfort.

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.



**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.