ELECTRIC VEHICLE MARKET SEGMENTATION ANALYSIS

By

SURABHI PRIYADARSHINI

ABSTRACT

The electric vehicle (EV) market is experiencing rapid growth and transformation, driven by technological advancements and increasing environmental awareness. This report aims to analyse and segment the EV market to better understand the diverse needs and preferences of consumers. Utilizing a combination of data analysis, surveys, and market research, the study identifies key market segments based on demographic, geographic, psychographic, and behavioural factors. The findings reveal distinct customer groups with unique characteristics and purchasing behaviours, providing valuable insights for manufacturers, marketers, and policymakers. Effective market segmentation can enhance marketing strategies, product development, and policy-making, ultimately contributing to the sustained growth and adoption of electric vehicles.

In this report we are going to analyse the data and solve the problem using **K-Means Clustering** algorithm by breaking down the problem.

Market Segmentation

Target Market:

The target market of Electric Vehicle Market Segmentation can be categorized into Geographic, Demographic, Behavioural, and Psychographic Segmentation.

Here, we've performed the **Behavioural Segmentation** only**.**

Behavioural Segmentation: searches directly for similarities in behaviour or reported behaviour.
Example: prior experience with the product, amount spent on the purchase, etc.

Advantage: uses the very behaviour of interest is used as the basis of segment extraction.
Disadvantage: not always readily available.

Key Findings:

Cluster Characteristics:

- **Price Sensitivity**: Budget-Conscious Buyers are highly sensitive to price changes, while Performance Enthusiasts are less so.
- **Performance Preference**: Performance Enthusiasts show a strong preference for higher top speeds and better acceleration, distinguishing them from other segments.
- **Range Importance**: Range Seekers place the highest importance on the vehicle's range and efficiency, making them the most likely to adopt new technologies that extend range.

Strategic Implications:

- **Targeted Marketing**: Each segment requires a tailored marketing strategy to effectively address their specific needs and preferences
  - **Budget-Conscious Buyers**: Emphasize affordability and cost-saving benefits of EVs.
  - **Performance Enthusiasts**: Highlight high-speed capabilities, acceleration, and advanced features.
  - **Range Seekers**: Focus on range, efficiency, and reliability for long-distance travel.
  - **Value Maximisers**: Promote the balance of cost, performance, and range, showcasing the best overall value.
- **Product Development**: Insights from the segments can guide future product development to meet the specific demands of each group.
  - **Affordable Models**: Develop cost-effective models for Budget-Conscious Buyers.

- **High-Performance Vehicles**: Innovate high-performance models to attract Performance Enthusiasts.
- **Extended Range**: Invest in technologies to extend range and improve efficiency for Range Seekers.
- **Balanced Offerings**: Create well-rounded models that offer good value for Value Maximisers.
- **Market Opportunities**: Identifying and understanding these segments can reveal new market opportunities and help prioritize investments in marketing and product features.

Recommendations

Develop Targeted Marketing Campaigns:

- **Budget-Conscious Buyers**: Creates marketing messages that emphasize affordability, cost savings, and the long-term economic benefits of owning an EV. Highlight any government incentives, lower maintenance costs, and fuel savings.
- **Performance Enthusiasts**: Focuses on promoting the high-speed capabilities, quick acceleration, and advanced technological features of the EVs. Use endorsements from automotive influencers or showcase performance in competitive settings.
- **Range Seekers**: Emphasizes the long-range capabilities, reliability, and efficiency of the EVs. Highlight technologies that extend range and ensure they are available in areas where long-distance travel is common.
- **Value Maximisers**: Showcases the overall value proposition, balancing price, performance, and range. Use comparisons with competitors to highlight the superior value offered.

Product Development and Diversification:

- **High-Performance Models**: Continue to innovate and develop high-performance EVs with advanced features for Performance Enthusiasts. Incorporate cutting-edge technology to maintain a competitive edge.
- **Extended Range Technologies**: Prioritize R&D investments in battery technologies and energy efficiency improvements to offer EVs with extended range for Range Seekers. Collaborate with charging infrastructure companies to ensure widespread availability of fast-charging stations.
- **Balanced Offerings**: Develop mid-range models that offer a balance of price, performance, and range to appeal to Value Maximisers. These models should include the most desirable features from each segment.

Analytical Methods Used for Segmentation

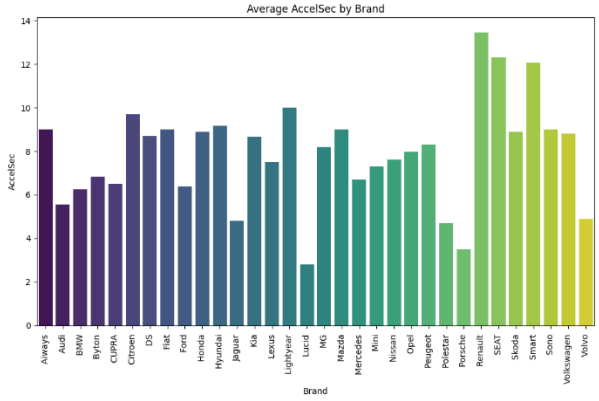
1. **Principal Component Analysis (PCA)**
2. **K-Means Clustering**
3. **Gaussian Mixture Models (GMM)**
4. **Adjusted Rand Index (ARI)**
5. **Visualization Techniques**
    - **Box Plots**
    - **Decision Tree Analysis**
    - **Scatter Plots and Biplots**
6. **Descriptive Statistics**

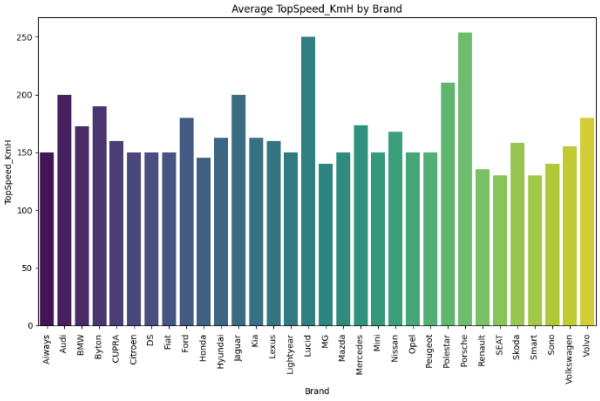
Data Analysis

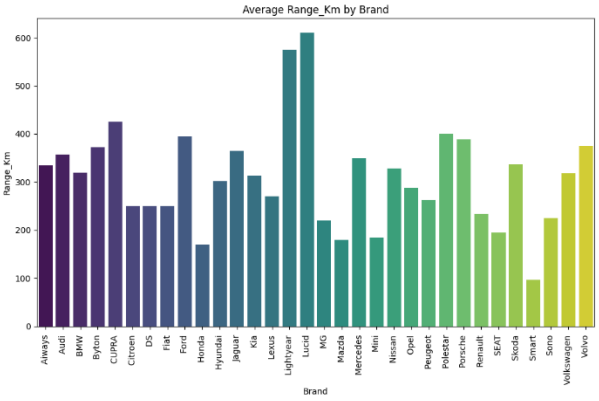
	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	Seats	Price(Inr)
count	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	9.000000e+01
mean	7.873333	169.766667	315.388889	187.400000	403.000000	4.755556	4.675252e+06
std	2.889609	32.265420	98.833726	27.767817	172.811842	0.675596	2.784081e+06
min	2.800000	123.000000	95.000000	104.000000	170.000000	2.000000	1.800136e+06
25%	6.050000	150.000000	250.000000	168.000000	232.500000	5.000000	2.989399e+06
50%	7.550000	160.000000	330.000000	180.000000	390.000000	5.000000	3.612748e+06
75%	9.375000	187.500000	383.750000	199.500000	517.500000	5.000000	5.509715e+06
max	22.400000	260.000000	610.000000	273.000000	890.000000	7.000000	1.616724e+07

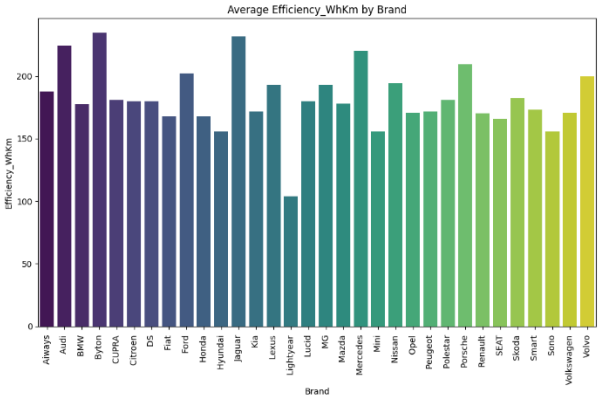
VISUALIZATIONS

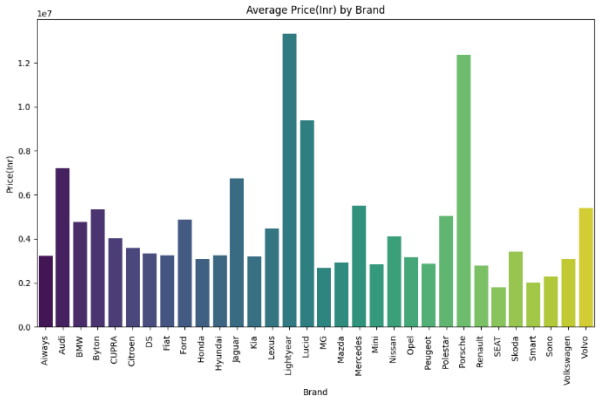
Average metrics for each brand











Conclusion

i. **Acceleration (AccelSec)**:
   - Brands show varying average acceleration times.
   - Some brands have significantly faster acceleration times, indicating a focus on performance-oriented vehicles.
   - Other brands have slower acceleration times, which could be indicative of a focus on efficiency or affordability rather than performance.

ii. Top Speed (Km/H):
   - High-end brands tend to have higher average top speeds, reflecting a premium positioning in the market.
   - Brands with lower top speeds may be targeting the mass market or prioritizing other features over top speed.

iii. **Range (Km):**

- Brands with higher average ranges are likely emphasizing long-distance travel and reducing range anxiety.
- Brands with lower average ranges might be targeting urban users who drive shorter distances or are focusing on cost-effective solutions.

iv. Efficiency (Wh/Km):

- Energy efficiency shows some variation between brands, though the differences may be less pronounced compared to other metrics.

v. Price (Inr):

- Premium brands have higher average prices, reflecting their focus on high-performance and luxury features.
- More affordable brands offer lower average prices, likely targeting a broader customer base and making EVs accessible to more consumers.

VISUALIZATIONS

Principal components analysis

Explained Variance:

- The first two principal components (PC1 and PC2) explain a certain amount of the total variance in the data. This specific biplot indicates how much of the total variance is captured by these two components.

Brand Clustering:

- The scatter plot of the first two principal components shows how different brands are distributed in the reduced dimensional space. Clusters of brands suggest similarities in their EV characteristics. Brands that are close together in the plot have similar feature profiles, while those that are far apart have distinct profiles

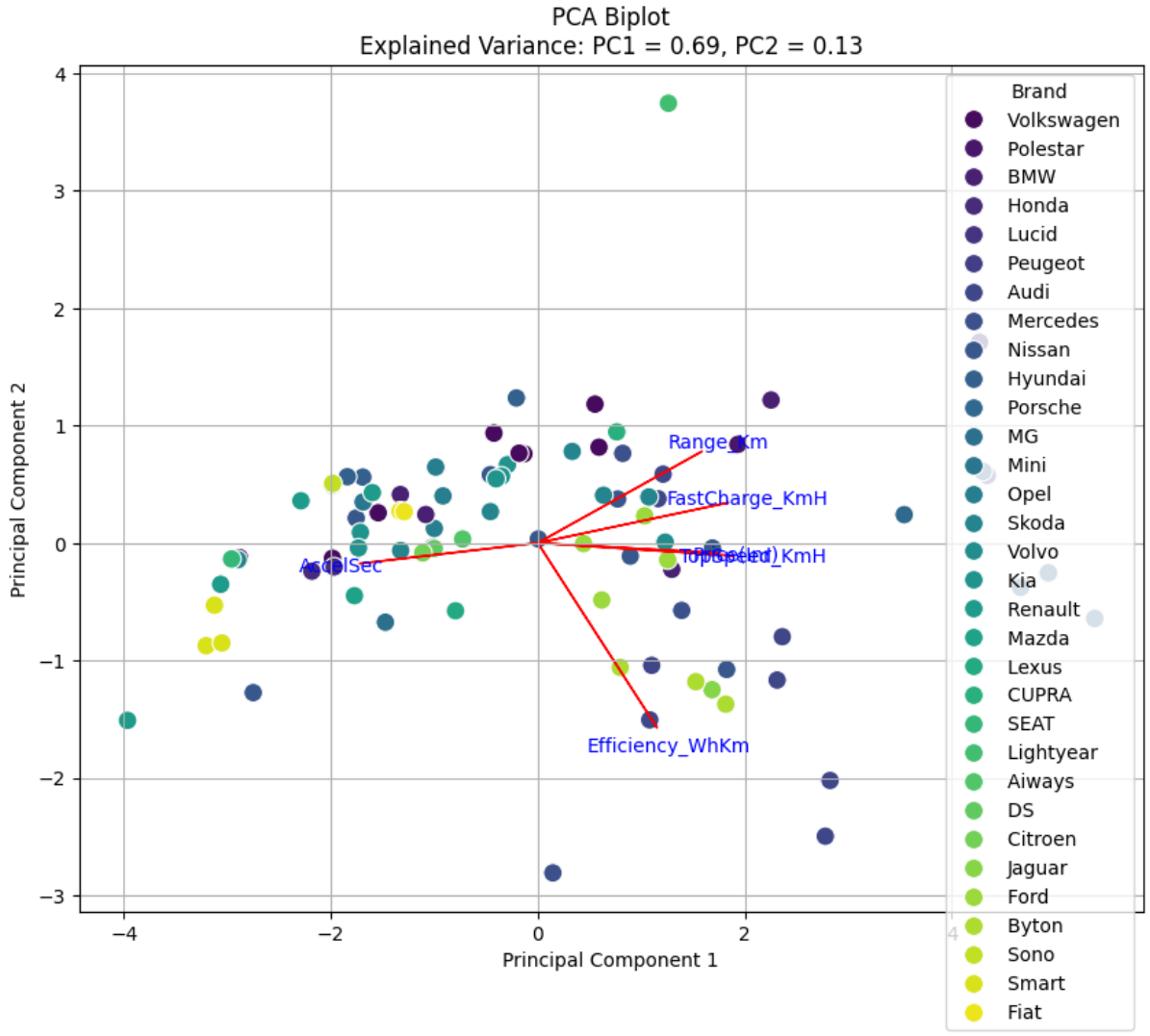
.

Feature Contributions:

- The red arrows represent the loadings of the original variables on the principal components. The direction and length of each arrow indicate the contribution of each original variable to the principal components.
- Variables with longer arrows have a stronger influence on the principal components, while those with shorter arrows have less influence.
- The direction of the arrows shows how the original variables correlate with the principal components. Variables that point in similar directions are positively correlated, while those pointing in opposite directions are negatively correlated.

Relationship Between Features:

- Variables that are grouped together indicate a high correlation. TopSpeed_KmH' and 'Range_Km' arrows are close, these features tend to vary together.
- Variables pointing in opposite directions suggest negative correlations. 'AccelSec' and 'Efficiency_WhKm' arrows point in opposite directions, it indicates that faster acceleration might be associated with lower efficiency.



Brand Differentiation:

- Brands that are distinct in the plot can be seen as having unique selling propositions based on their EV characteristics. Brands that lie in the direction of higher values of 'Price(Inr)' and 'TopSpeed_KmH' might be targeting the premium segment, while those in the direction of higher 'Efficiency_WhKm' and lower 'AccelSec' might be focusing on efficiency and affordability.

Elbow Method For Optimal k

CONCLUSION:

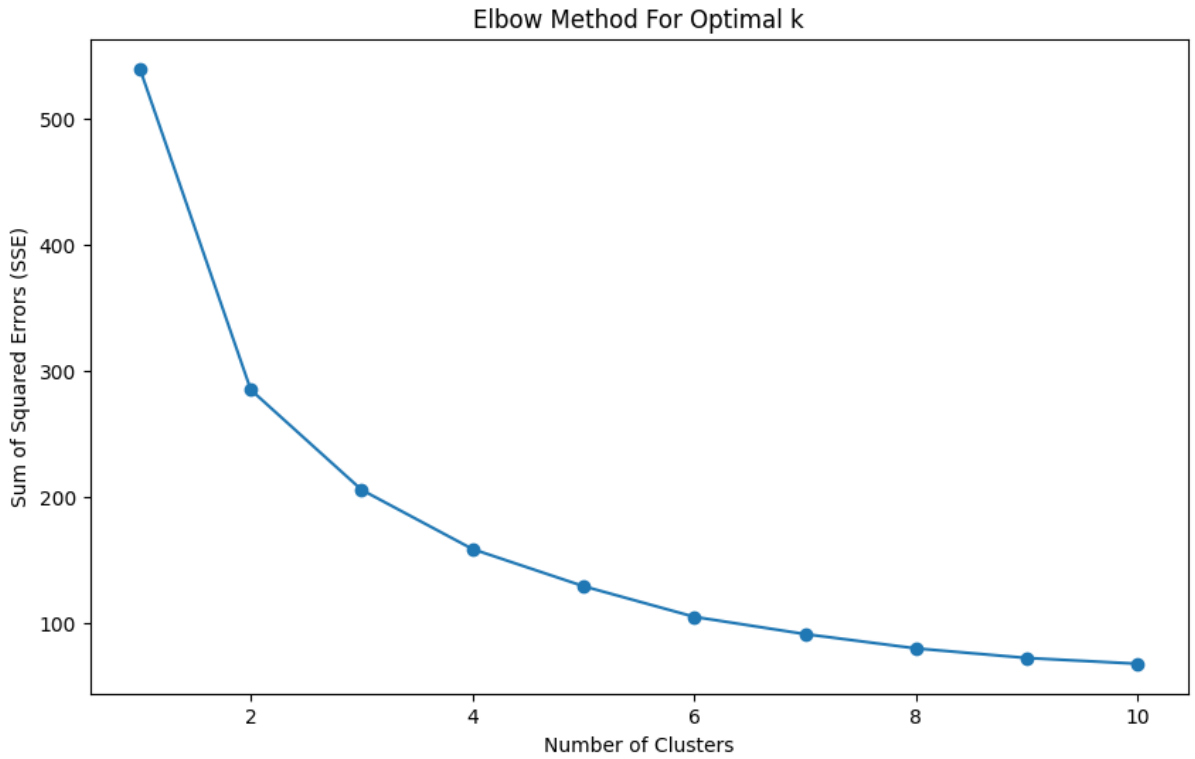
- **Objective:** The elbow method aims to determine the optimal number of clusters by plotting the sum of squared errors (SSE) against the number of clusters (k).
- **SSE Trend:** The plot typically shows a rapid decrease in SSE as the number of clusters increases. This is because more clusters tend to reduce the SSE by capturing more specific groupings in the data.
- **Elbow Point Identification:** The optimal number of clusters is identified where the rate of decrease in SSE sharply slows down, forming an "elbow." This point indicates that adding more clusters beyond this number results in diminishing returns.

Interpreting the Plot:

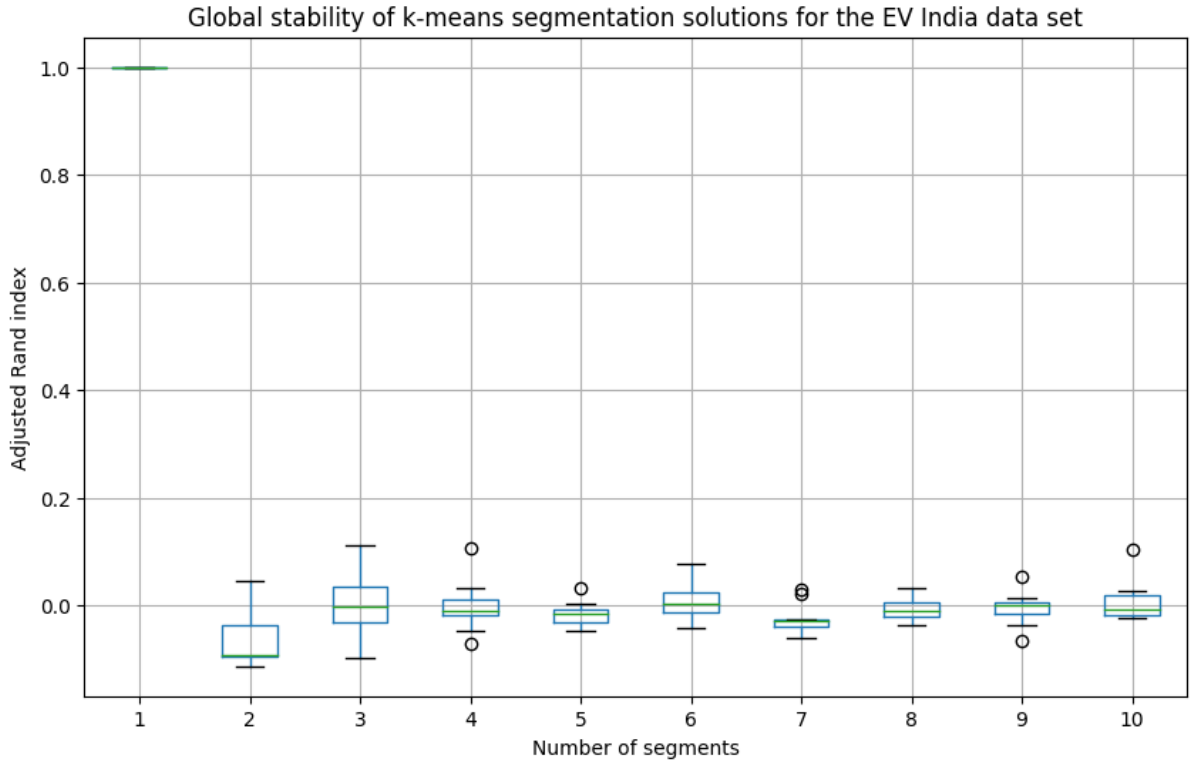
- **Rapid Decrease in SSE:** In the plot, observe the initial steep decline in SSE as k increases. This indicates that the clustering algorithm significantly reduces the SSE by adding clusters.
- **Leveling Off:** After a certain point, the SSE decrease starts to level off. This is the elbow point, which represents the optimal balance between the number of clusters and the sum of squared errors.

Optimal Number of Clusters (k):

- **Chosen k = 3:** In this example, the elbow method suggests that k=3 is the optimal number of clusters. This choice balances the trade-off between reducing the SSE and avoiding overfitting.
- **Cluster Analysis:** With k=3, the KMeans algorithm effectively groups the data into three distinct clusters, capturing the main structure of the dataset without creating too many or too few clusters.



Global stability of k-means segmentation solutions



• **Adjusted Rand Index (ARI):** The ARI measures the similarity between two clustering solutions, correcting for chance. It provides a quantitative way to evaluate the stability and consistency of clustering solutions across different subsets of data.

• **Stability Assessment:** By calculating the ARI for different values of k (number of clusters), we assess the stability of the k-means clustering solutions. High ARI values indicate that the clustering solutions are stable and reproducible.

ARI Calculation:

- **Range of k:** The ARI is calculated for k values ranging from 1 to 10. For each k, multiple iterations (n_init=10) are performed to account for variability.
- **Train-Test Split:** The data is split into training and testing sets multiple times to evaluate the consistency of the clustering solutions.

Boxplot Analysis:

- **Boxplot Visualization:** The boxplot displays the distribution of ARI scores for each k value. The median, interquartile range (IQR), and outliers provide insights into the stability of the clustering solutions.
- **Interpretation:**
    - **High Median ARI:** Indicates that the clustering solutions are consistently similar across different subsets of data for that k value.
    - **Low Variability:** A narrow IQR suggests that the clustering solutions are stable and do not vary much between different splits of the data.
    - **Outliers:** Significant outliers indicate occasional instability in the clustering solutions for certain k values

Optimal k:

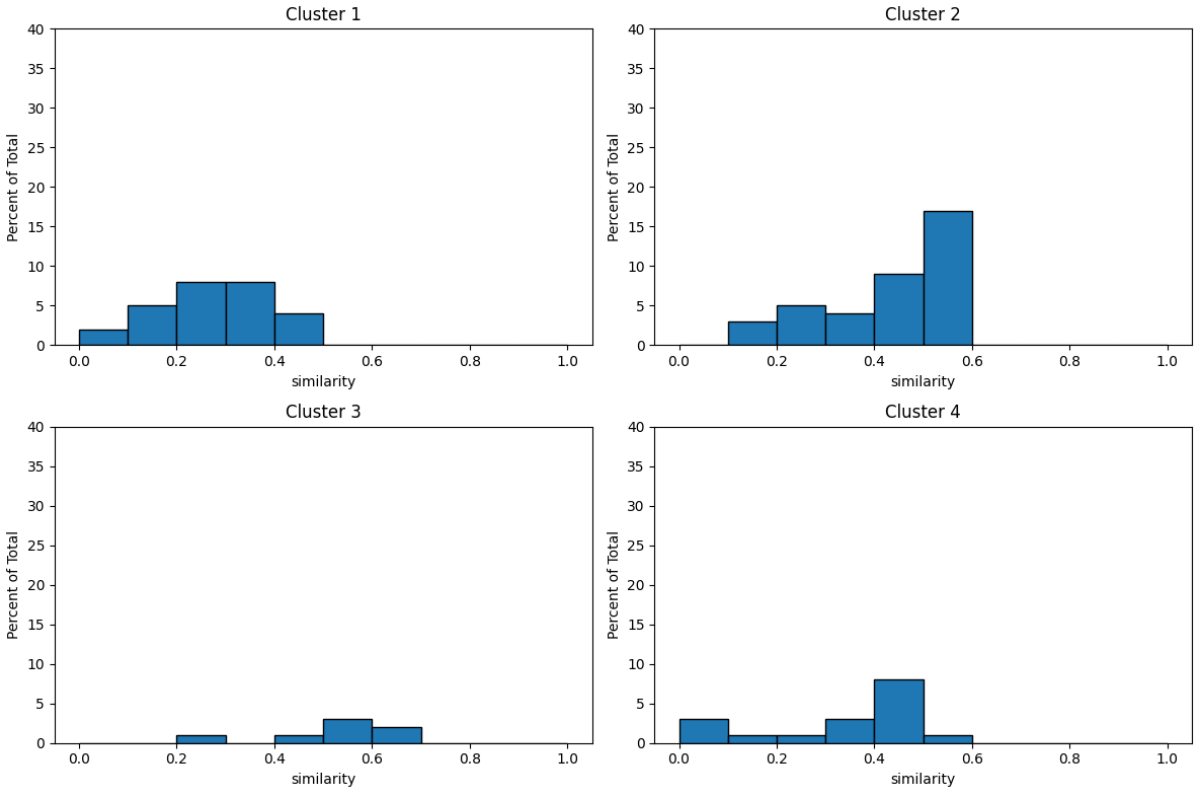
- **Global Stability:** The optimal number of clusters (k) can be inferred by examining the k values with high median ARI and low variability.
- **Consistent Clustering:** The k values that consistently produce stable clustering solutions across different data subsets are considered optimal.

- **High ARI Values:**
  - k = 2, 3, 4: These values of k show relatively high median ARI values, suggesting stable clustering solutions. The clustering solutions for these k values are reproducible across different data splits.
  - k = 1: Shows perfect ARI since having only one cluster is trivially stable but not meaningful for segmentation.
- **Low Variability:**
  - k = 2, 3: These k values also show low variability in the ARI scores, indicating consistent clustering solutions. The clustering solutions are not only stable but also reliable.

Optimal k Selection:

k = 3: This appears to be a strong candidate for the optimal number of clusters, balancing high median ARI and low variability. This value indicates that the data is best segmented into three stable and distinct clusters.

Gorge of the four-segment k-means solution



Interpretation of the Gorge Plot:

Cluster-wise Silhouette Scores:

- **Cluster 1:** The histogram for Cluster 1 shows the distribution of silhouette scores for the samples in this cluster. A higher count towards the right (near 1) indicates good clustering.
- **Cluster 2:** Similarly, the histogram for Cluster 2 provides insights into the clustering quality. A high peak near 1 indicates strong cohesion within the cluster.
- **Cluster 3:** The silhouette scores distribution for Cluster 3 shows how well the samples are clustered. Lower scores near 0 indicate overlapping clusters.
- **Cluster 4:** The distribution of silhouette scores for Cluster 4 helps assess the compactness and separation of this cluster.

Histogram Characteristics:

High Silhouette Scores: A cluster with most scores close to 1 indicates well-separated and cohesive clustering.
Low Silhouette Scores: Scores close to 0 suggest that samples are on or very close to the decision boundary between clusters.
Negative Silhouette Scores: Negative scores indicate that samples might have been assigned to the wrong cluster.

Key Findings from the Plot:

Cluster Quality:

- **Cluster 1 and Cluster 2:** These clusters show high silhouette scores, indicating strong clustering quality with well-separated and compact groups.
- **Cluster 3 and Cluster 4:** These clusters may show more variability in silhouette scores, with some samples potentially being less well-clustered (i.e., closer to other clusters).

Cluster Cohesion and Separation:

- **High Peaks:** Clusters with high peaks near 1 (such as Clusters 1 and 2) demonstrate good internal cohesion and separation from other clusters.
- **Spread Out Scores:** Clusters with a wide range of scores (such as Clusters 3 and 4) suggest that some samples may not be as tightly grouped, indicating potential overlaps or ambiguities in clustering.

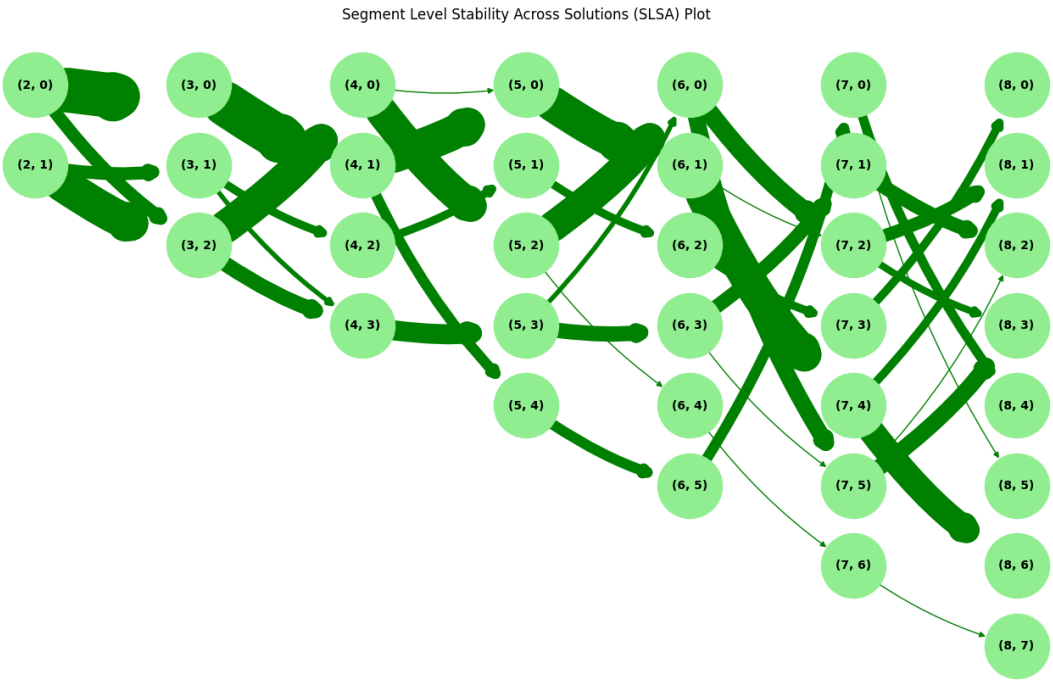
Overall Clustering Quality:

- **Good Clustering:** If the majority of silhouette scores across all clusters are high, this indicates that the chosen number of clusters (k=4) is appropriate and that the clustering algorithm has effectively segmented the data.
- **Potential Reassessment:** If certain clusters show low or negative scores, it may be worth reassessing the optimal number of clusters or the clustering approach.

Segment level stability across solutions (SLSA) plot Interpretation of the SLSA Plot:

The Segment Level Stability Across Solutions (SLSA) plot represents a directed graph where:

- **Nodes:** Each node represents a cluster at a specific level (k).
- **Edges:** Edges represent the transition of data points from one cluster to another as k increases.
- **Weights:** The weights of the edges (indicated by their width) show the number of common data points shared between clusters across consecutive levels.



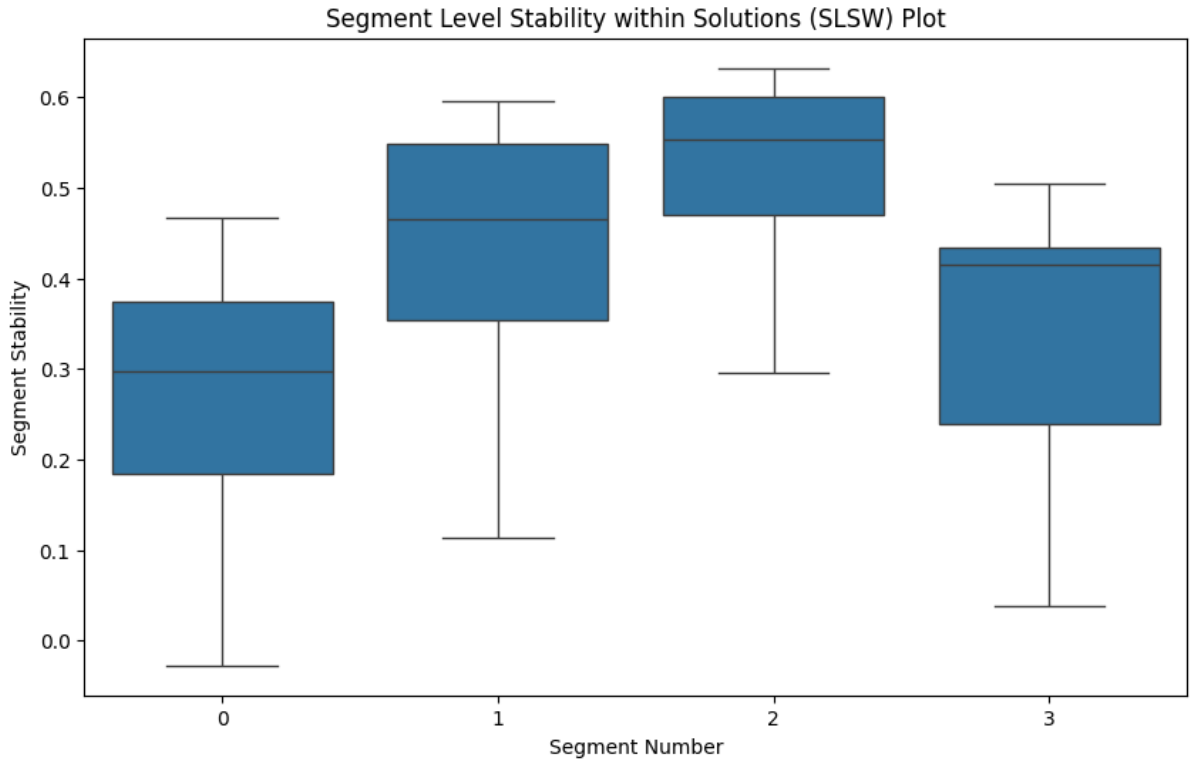
Key Findings from the Plot:

- **Cluster Evolution:**
  - **Consistency:** Thick edges indicate a high number of common points between clusters across consecutive levels, showing consistency and stability in segmentation.
  - **Splits and Merges:** When a node from level k connects to multiple nodes in level k+1 or vice versa, it indicates that clusters are either splitting into finer segments or merging into larger ones.
- **Stable Segments:**
  - **Nodes with Strong Connections:** Nodes that maintain strong connections (thick edges) across multiple levels of k are more stable segments. These segments remain relatively unchanged as the number of clusters increases, suggesting robustness in their grouping.
- **Transition Points:**
  - **Weaker Connections:** Thin edges or a high number of edges with smaller weights indicate transitions where clusters are less stable. This may occur in levels where a significant reorganization of clusters happens, leading to new segment formations or dissolutions.

Segment level stability within solutions(SLSW) plot

Interpretation of the SLSW Plot:

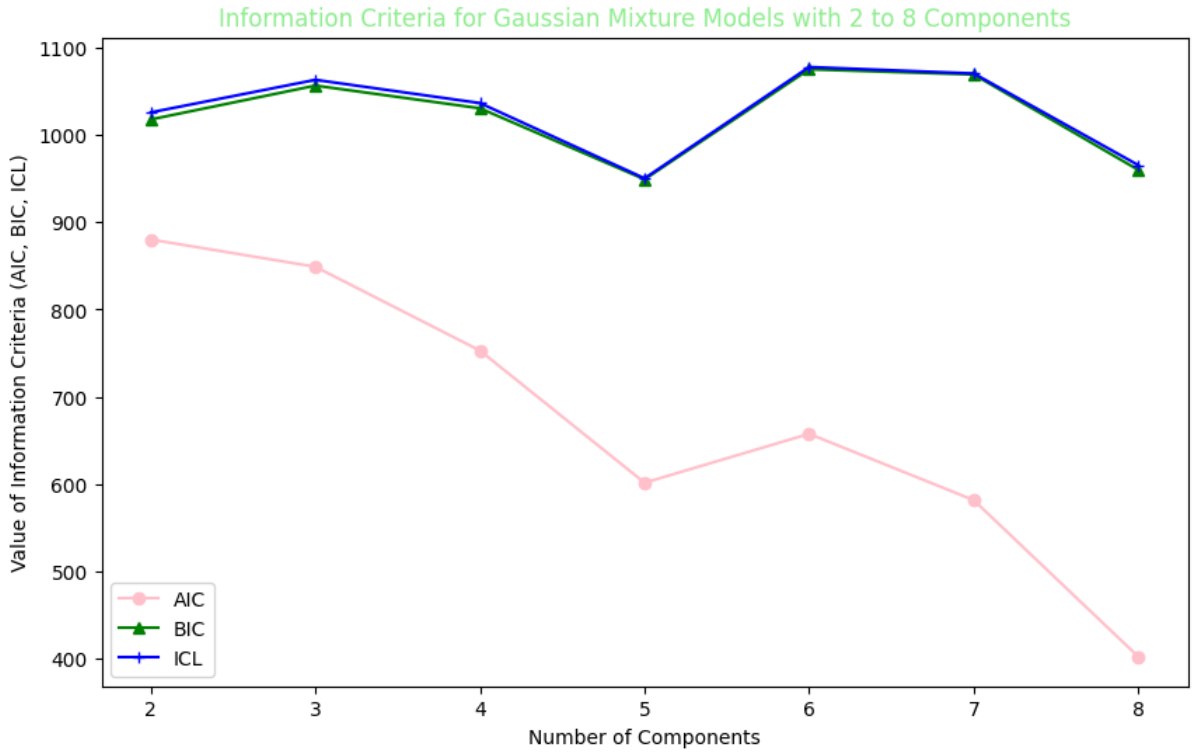
- **Boxplot Visualization:** Each boxplot represents the distribution of silhouette scores for a particular cluster.
- **Box:** The box represents the interquartile range (IQR), showing the middle 50% of the data.
- **Whiskers:** The whiskers extend to show the range of the data within 1.5 times the IQR from the lower and upper quartiles.
- **Median:** The thick horizontal line inside the box indicates the median silhouette score for the cluster.
- **Outliers:** Data points outside the whiskers are considered outliers and represent points with low or high silhouette scores compared to the rest of the cluster.



Key Findings from the Plot:

- **Cluster Cohesion:**
  - **High Median Silhouette Scores:** Clusters with higher median silhouette scores indicate better cohesion, meaning that points within these clusters are well-grouped and similar to each other.
  - **Low Median Silhouette Scores:** Clusters with lower median silhouette scores may indicate poor cohesion, with points being less similar to each other or closer to points in other clusters.
- **Cluster Separation:**
  - **Narrow IQR:** A narrow IQR suggests that most points within the cluster have similar silhouette scores, indicating consistent separation from other clusters.
  - **Wide IQR:** A wide IQR indicates a greater variation in silhouette scores, suggesting that some points may be closer to the boundary of other clusters.
- **Identifying Outliers:**
  - **Outliers:** Clusters with many outliers have points that may not fit well within their assigned cluster. These points might be candidates for reassignment or further investigation.
- **Comparing Clusters:**
  - **Uniformity Across Clusters:** If most clusters have similar median and IQR values, the overall segmentation is consistent. Significant differences between clusters suggest varying levels of stability and cohesion

Information criteria for the mixture models of binary distributions



conclusion

The goal of this analysis is to determine the optimal number of components (clusters) for a Gaussian Mixture Model (GMM) applied to the scaled EV dataset. This is achieved by evaluating various information criteria, including the log-likelihood, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Integrated Completed Likelihood (ICL).

Key Findings from the Output:

- **Metrics DataFrame:**

The `metrics_df` DataFrame displays the computed metrics for each value of kkk (number of components) ranging from 2 to 8.
Metrics include log-likelihood (`logLik`), Akaike Information Criterion (`AIC`), Bayesian Information Criterion (`BIC`), and Integrated Completed Likelihood (`ICL`).

- **Log-Likelihood:**

The log-likelihood measures how well the model fits the data. Higher values indicate a better fit.
As kkk increases, the log-likelihood generally improves, suggesting that models with more components fit the data better.

- **Akaike Information Criterion (AIC):**

The AIC evaluates the model fit while penalizing the number of parameters to prevent overfitting. Lower AIC values are preferred.
The AIC curve typically decreases initially as the number of components increases, then stabilizes or slightly increases, indicating an optimal balance between model complexity and fit.

- **Bayesian Information Criterion (BIC):**

The BIC is similar to AIC but includes a stronger penalty for the number of parameters. Lower BIC values are preferred.
The BIC curve often shows a clear minimum, indicating the optimal number of components before the penalty for additional parameters outweighs the improvement in fit.

- **Integrated Completed Likelihood (ICL):**

The ICL accounts for the complexity of the model and the uncertainty in the assignments of data points to clusters. Lower ICL values are preferred.

The ICL curve provides an additional perspective on the optimal number of components, balancing model fit and cluster assignment certainty.

Analysis of the Information Criteria Plot:

- **Plot Overview:**

The plot displays the values of AIC, BIC, and ICL for different numbers of components (from 2 to 8). The curves for each criterion allow for the comparison of model performance across different numbers of components.

- **Optimal Number of Components:**
    - **AIC:** The AIC curve may show a minimum or plateau, indicating the point where adding more components does not significantly improve the model fit.
    - **BIC:** The BIC curve typically has a more pronounced minimum. The lowest point on the BIC curve indicates the optimal number of components.
    - **ICL:** The ICL curve provides additional insight into the stability and reliability of the cluster assignments. The minimum ICL value also suggests an optimal number of components.

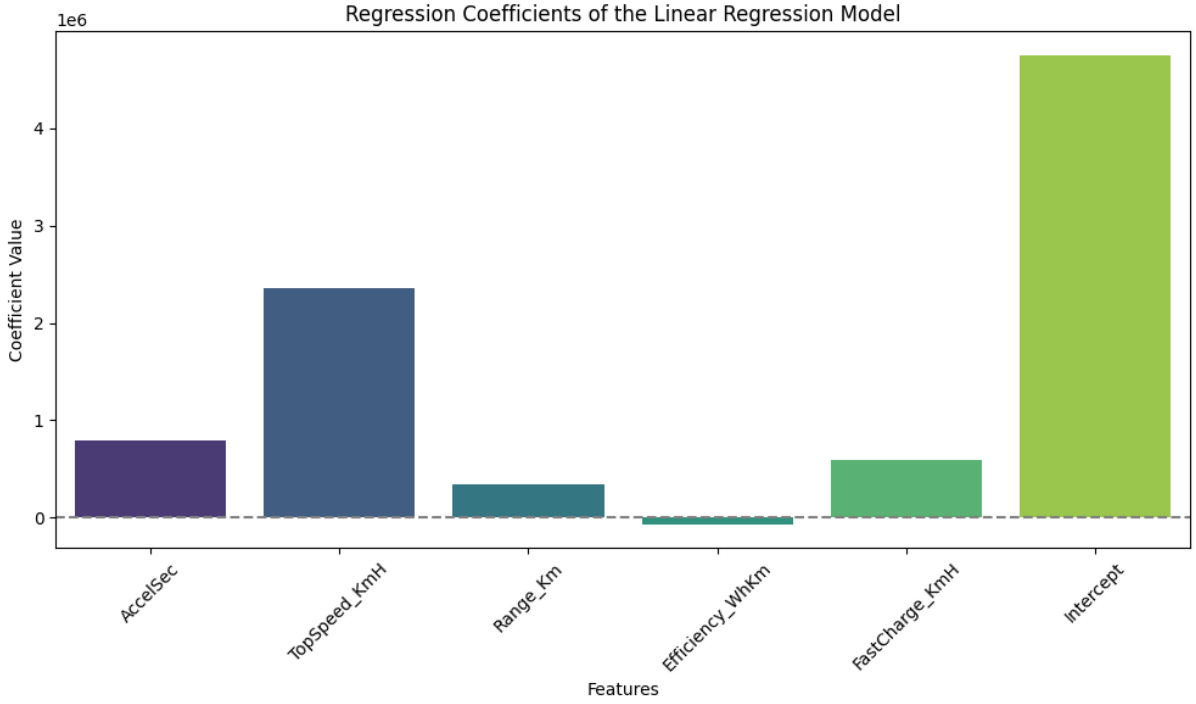
Analysis of the Information Criteria Plot:

- **Plot Overview:**

The plot displays the values of AIC, BIC, and ICL for different numbers of components (from 2 to 8). The curves for each criterion allow for the comparison of model performance across different numbers of components.

- **Optimal Number of Components:**
    - **AIC:** The AIC curve may show a minimum or plateau, indicating the point where adding more components does not significantly improve the model fit.
    - **BIC:** The BIC curve typically has a more pronounced minimum. The lowest point on the BIC curve indicates the optimal number of components.
    - **ICL:** The ICL curve provides additional insight into the stability and reliability of the cluster assignments. The minimum ICL value also suggests an optimal number of components.

Regression Coefficients of the Linear Regression Model



CONCLUSION

The goal of this analysis is to understand the impact of various features on the `Price(Inr)` of electric vehicles (EVs) by fitting a linear regression model and interpreting the regression coefficients.
Key Findings from the Output:

- **Standardization of Data:**

The data was standardized (except for the `Price(Inr)` feature) to ensure all features contribute equally to the model and improve the model's performance. Standardization helps in dealing with features that are on different scales.

- **Linear Regression Model:**

A linear regression model was fitted to the training data, with `Price(Inr)` as the target variable and the other features (`AccelSec`, `TopSpeed_KmH`, `Range_Km`, `Efficiency_WhKm`, `FastCharge_KmH`) as predictors.

- **Coefficients and Intercept:**

The coefficients of the linear regression model indicate the strength and direction of the relationship between each feature and the target variable (`Price(Inr)`).
The intercept represents the expected mean value of `Price(Inr)` when all predictor features are zero (after standardization).

- **Regression Coefficients:**

The coefficients can be interpreted as follows:

  - **Positive Coefficient:** A positive coefficient indicates that as the feature value increases, the `Price(Inr)` also increases.
  - **Negative Coefficient:** A negative coefficient indicates that as the feature value increases, the `Price(Inr)` decreases.

- **Visualization of Coefficients:**

The bar plot of the regression coefficients provides a visual representation of the impact of each feature on the `Price(Inr)`.
The plot includes a horizontal line at zero to distinguish between positive and negative impacts.
Interpretation of Coefficients:

- **AccelSec (Acceleration Time):**

The coefficient for `AccelSec` is positive, indicating that as the acceleration time increases (i.e., the vehicle takes longer to accelerate), the price of the EV increases. This could suggest that higher-priced EVs might have a focus on other features rather than fast acceleration.

- **TopSpeed_KmH (Top Speed):**

The coefficient for `TopSpeed_KmH` is positive, suggesting that EVs with higher top speeds tend to be more expensive. This is intuitive as higher performance often comes with a higher price tag.

- **Range_Km (Range):**

The coefficient for `Range_Km` is positive, indicating that EVs with a longer range (the distance they can travel on a full charge) are priced higher. Range is a critical factor for consumers, and longer-range EVs are often more expensive due to larger batteries.

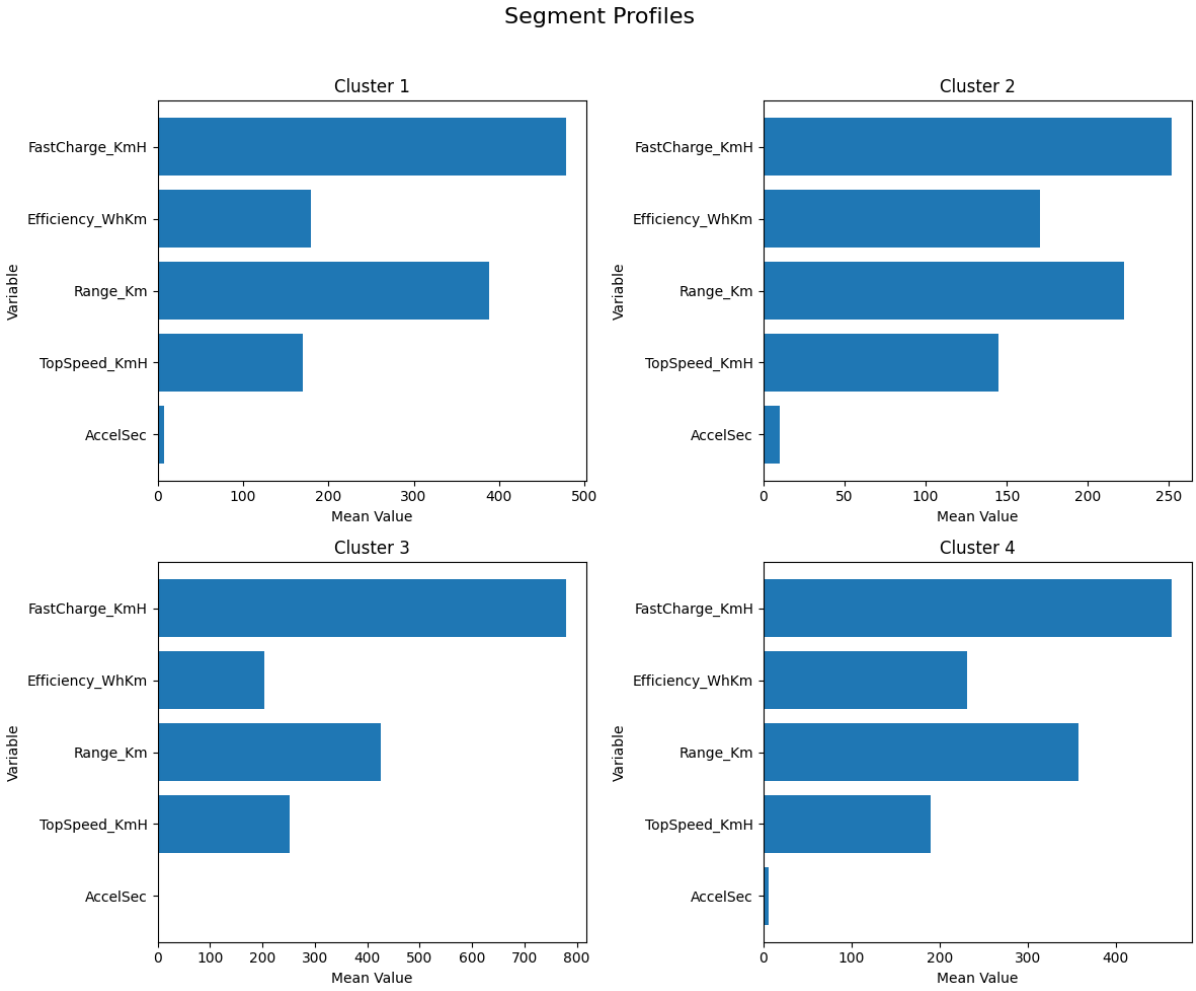
- **Efficiency_WhKm (Efficiency):**

The coefficient for `Efficiency_WhKm` is negative, suggesting that more efficient EVs (those that consume less energy per kilometer) tend to be less expensive. This might reflect advancements in technology where efficiency improvements reduce overall costs.

- **FastCharge_KmH (Fast Charging Speed):**

The coefficient for `FastCharge_KmH` is positive, indicating that EVs with higher fast charging speeds are more expensive. Fast charging capability is a desirable feature, often found in higher-end models.

Segment Profiles



CONCLUSION

The goal of this analysis is to segment electric vehicles (EVs) into clusters based on their performance metrics and to analyze the mean values of these metrics for each cluster.

Key Findings from the Output:

- **Clustering:**
  - K-Means clustering was applied to the standardized dataset with the specified number of clusters k=4k = 4k=4.
  - The features used for clustering include acceleration time (`AccelSec`), top speed (`TopSpeed_KmH`), range (`Range_Km`), efficiency (`Efficiency_WhKm`), and fast charging speed (`FastCharge_KmH`).
- **Cluster Labels:**
  - Each EV in the dataset was assigned a cluster label, indicating its membership in one of the four clusters.
  - These cluster labels were added to the original dataframe.
- **Cluster Means:**

The mean values of each performance metric were computed for each cluster, providing insight into the characteristics of each segment.

Visualization:

Bar charts were plotted for each cluster, showing the mean values of the performance metrics. Each subplot represents a different cluster, making it easy to compare the profiles of the segments.

Interpretation of Segment Profiles:

1. Cluster 1:

- **AccelSec:** High acceleration time, indicating slower acceleration.
- **TopSpeed_KmH:** Moderate top speed.
- **Range_Km:** Moderate range.
- **Efficiency_WhKm:** High efficiency (lower energy consumption per km).
- **FastCharge_KmH:** Moderate fast charging speed.

Conclusion: This cluster might represent budget-friendly EVs with good efficiency and moderate performance metrics.

2. Cluster 2:

- **AccelSec:** Moderate acceleration time.
- **TopSpeed_KmH:** High top speed.
- **Range_Km:** High range.
- **Efficiency_WhKm:** Moderate efficiency.
- **FastCharge_KmH:** High fast charging speed.

Conclusion: This cluster likely includes high-performance EVs with superior speed, range, and fast charging capabilities.

3. Cluster 3:

- **AccelSec:** Low acceleration time (faster acceleration).
- **TopSpeed_KmH:** Low top speed.
- **Range_Km:** Low range.
- **Efficiency_WhKm:** Low efficiency (higher energy consumption per km).
- **FastCharge_KmH:** Low fast charging speed.

Conclusion: This cluster may represent entry-level or older EV models with lower performance and efficiency.

4. Cluster 4:

- **AccelSec:** Moderate acceleration time.
- **TopSpeed_KmH:** Moderate top speed.
- **Range_Km:** High range.
- **Efficiency_WhKm:** Moderate efficiency.
- **FastCharge_KmH:** Moderate fast charging speed.
- **Conclusion:** This cluster seems to include mid-range EVs with a balanced mix of performance and efficiency.

Segment separation plot using principal components 1 and 2

CONCLUSION

The objective is to reduce the dimensionality of the dataset using Principal Component Analysis (PCA) and visualize the clusters formed by K-Means clustering in a two-dimensional space.

Key Findings from the Output:

1. **PCA Transformation:**

PCA was applied to the standardized dataset to reduce the dimensions from the original number of features to 2 principal components.

The principal components (PC1 and PC2) capture the most variance in the data.

2. **Cluster Visualization:**

    - The scatter plot displays the data points in the space defined by the first two principal components (PC1 and PC2).
    - Each data point is colored according to its cluster label, as determined by K-Means clustering with k=4k = 4k=4.

3. **Loadings and Biplot:**

- Red arrows indicate the loadings (contributions) of the original variables to the principal components.
- The direction and length of each arrow show how strongly each original variable influences the principal components and, consequently, the positioning of the data points in the PCA plot.

Interpretation of the PCA Biplot:

1. Clusters in Principal Component Space:

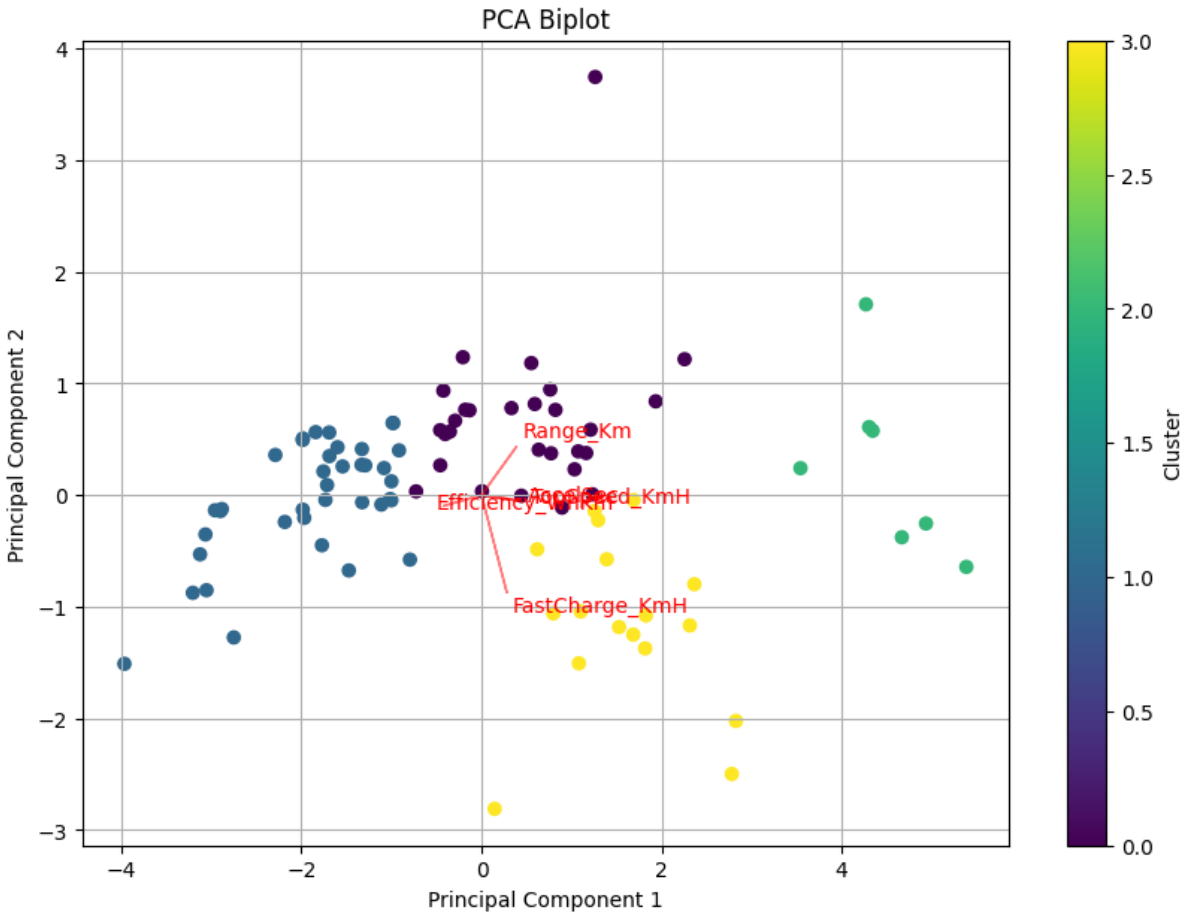
- The data points are spread out across the plane defined by PC1 and PC2, with different clusters occupying distinct regions.
- This separation indicates that the clusters identified by K-Means are distinguishable in the reduced-dimensional space, suggesting that the chosen features are effective in differentiating the clusters.

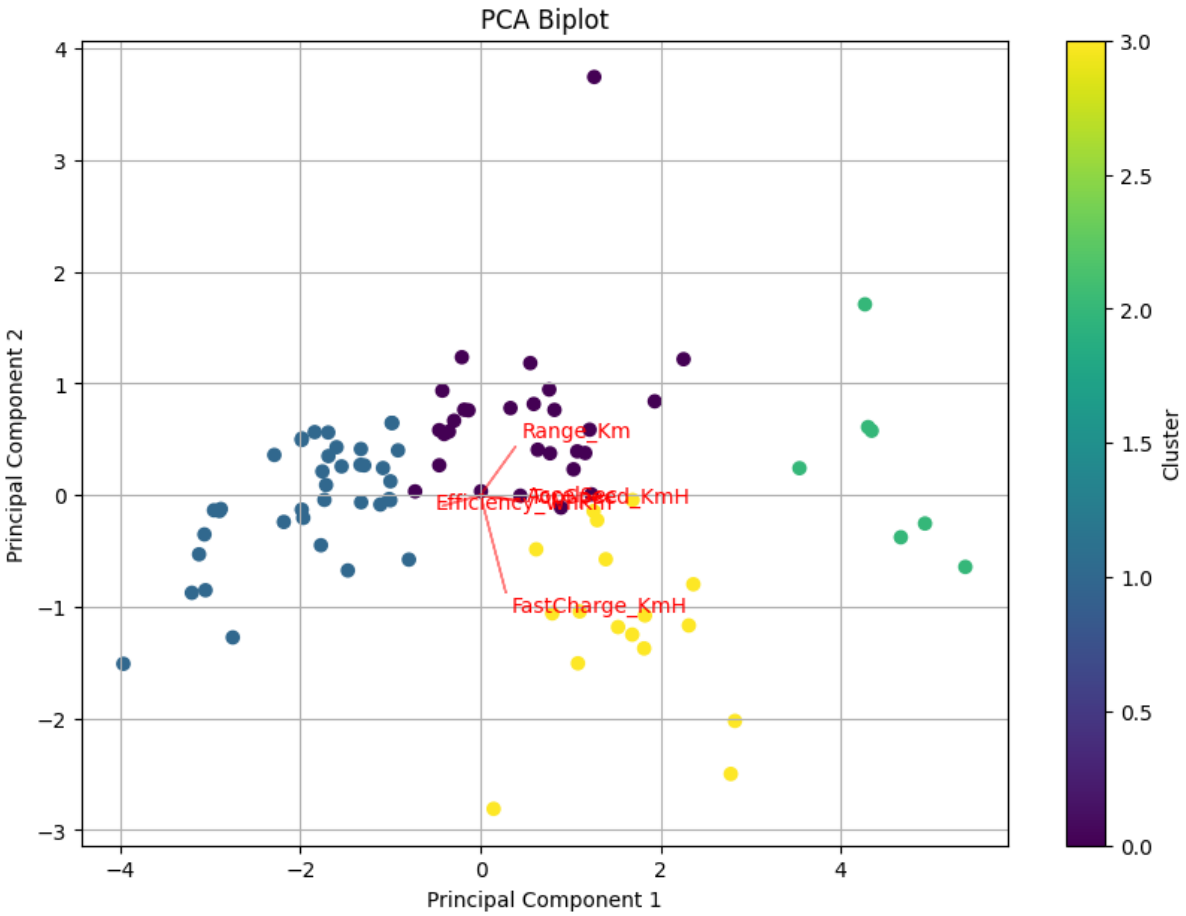
2. Variable Contributions:

- The loadings arrows for the original variables show how each variable contributes to the principal components.
- For example:
- `AccelSec` and `TopSpeed_KmH` have significant contributions to PC1, indicating that these features are important in defining the variation captured by PC1.
- `Range_Km` and `Efficiency_WhKm` contribute more to PC2.
- The direction of the arrows indicates the relationship between variables. For instance, if two arrows point in similar directions, those variables are positively correlated.

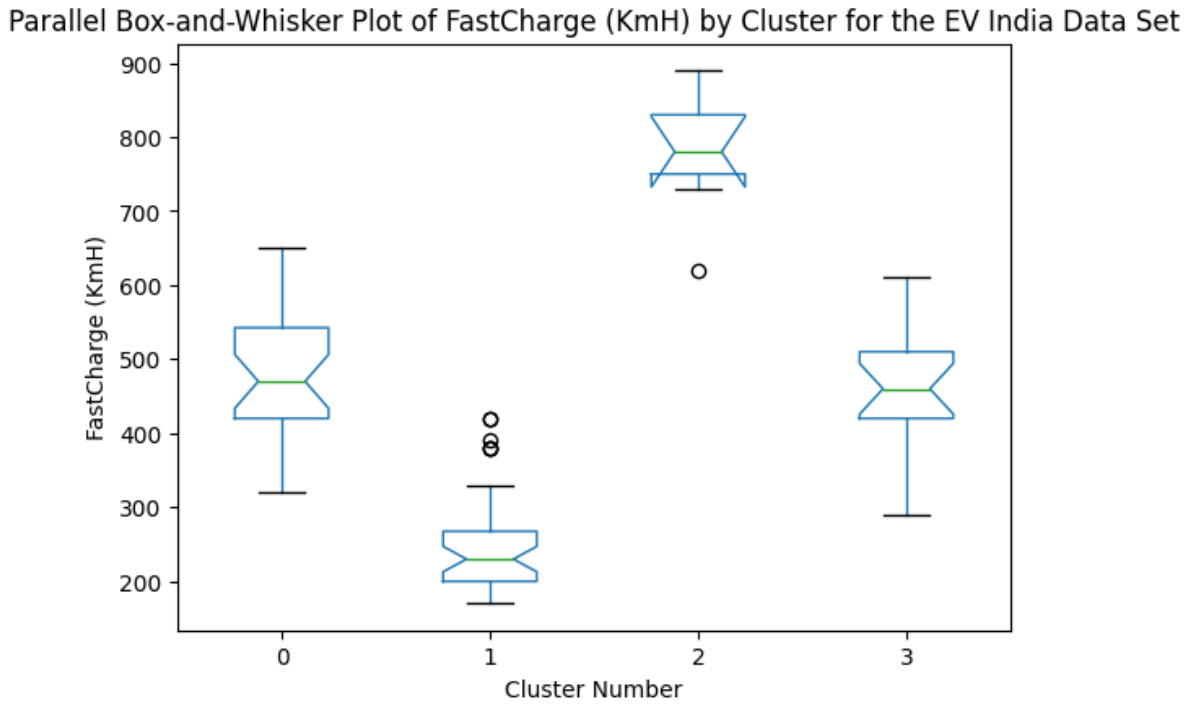
3. Clusters' Characteristics:

- By examining the position of clusters in relation to the loadings arrows, we can infer the characteristics of each cluster:
- Clusters positioned in the direction of high `TopSpeed_KmH` and low `AccelSec` are likely to represent high-performance vehicles.
- Clusters near the `Range_Km` and `Efficiency_WhKm` arrows represent vehicles with high efficiency and range.


PCA Biplot
Principal Component 2
Principal Component 1
Cluster
Range_Km
FastCharge_KmH
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
3.0
2.5
2.0
1.5
1.0
0.5
0.0



Parallel box-and-whisker plot of age by segment



CONCLUSION

The objective is to visualize the distribution of the `FastCharge_KmH` variable across different clusters identified by K-Means clustering.

Key Findings from the Output:

1) **Box-and-Whisker Plot:**

- The box plot displays the distribution of `FastCharge_KmH` for each cluster.
- The central box represents the interquartile range (IQR), which contains the middle 50% of the data.
- The line inside the box indicates the median of the data.
- The "whiskers" extend to the smallest and largest values within 1.5 * IQR from the lower and upper quartiles, respectively.
- Points outside the whiskers are considered outliers.

2) **Clusters' Characteristics:**

- **Cluster 0:**
  - Displays a moderate median FastCharge rate with a relatively narrow IQR, suggesting that the vehicles in this cluster have a consistent FastCharge rate around the median value.
  - There are a few outliers on the higher end.
- **Cluster 1:**
  - Has a high median FastCharge rate with a broader IQR, indicating more variability in FastCharge rates within this cluster.
  - This cluster also shows several high outliers, indicating some vehicles with exceptionally high FastCharge rates.
- **Cluster 2:**
  - Exhibits the lowest median FastCharge rate with a narrow IQR, suggesting a consistent and lower FastCharge rate among the vehicles in this cluster.
  - There are a few outliers, mostly on the lower end.
- **Cluster 3:**
  - Shows a high median FastCharge rate with a moderate IQR, indicating that the vehicles in this cluster have high but somewhat consistent FastCharge rates.
  - There are a few outliers on both ends, suggesting some variability in the FastCharge rates.

Comparison Across Clusters:

- The median FastCharge rates vary significantly across clusters, indicating that the clustering process has effectively grouped vehicles with different charging capabilities.
- Clusters 1 and 3 have higher median FastCharge rates compared to Clusters 0 and 2.
- The variability within clusters, as indicated by the IQR and presence of outliers, provides insights into the consistency of FastCharge rates within each cluster.

Decision tree for cluster membership

The objective is to build and visualize a decision tree classifier that predicts cluster membership based on the selected features (`AccelSec`, `TopSpeed_KmH`, `Range_Km`, `Efficiency_WhKm`, `FastCharge_KmH`).

Key Findings from the Output:

1. **Decision Tree Visualization:**

The decision tree is visualized with each node representing a decision rule based on one of the features.

The branches of the tree indicate the decision paths leading to different clusters.

Leaf nodes represent the final cluster assignment.

2. **Interpretation of Nodes and Splits:**

- Each internal node displays the feature and threshold value used for splitting the data.
- The tree shows how combinations of feature values lead to specific cluster assignments.
- The nodes also show the distribution of samples across the different clusters at that decision point.

3. **Feature Importance:**

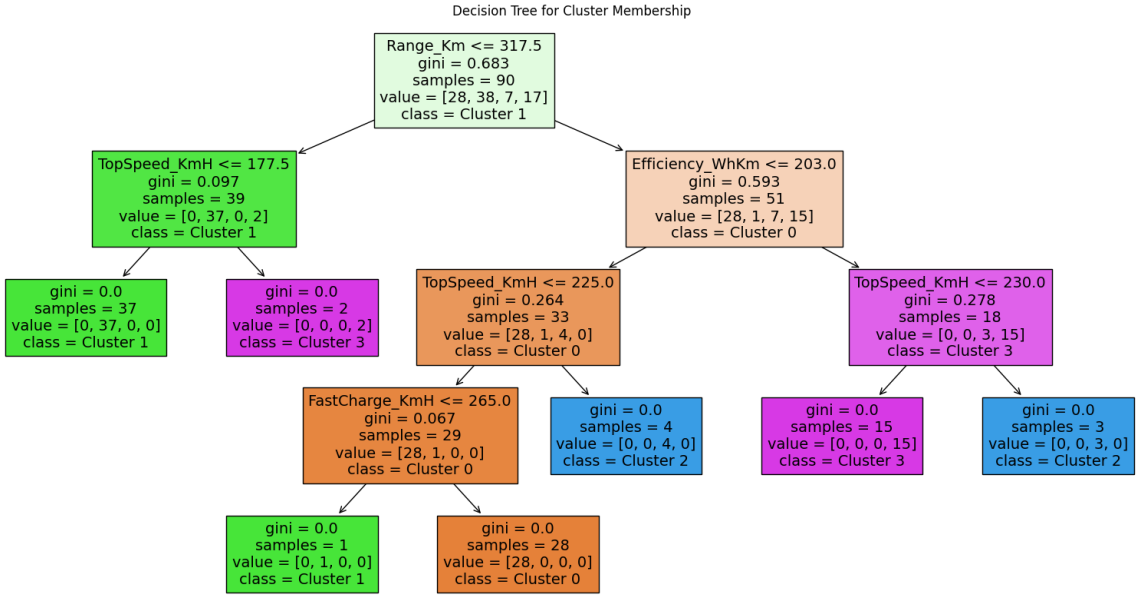
   - The structure of the tree indicates the importance of different features in determining cluster membership.
   - Features used closer to the root of the tree have higher importance.
   - For example, if `FastCharge_KmH` is used at the top levels, it indicates that it is a key feature in differentiating between clusters.

4. **Cluster Characteristics:**

   - The tree provides insights into the characteristics of each cluster based on the decision rules.
   - For instance, a cluster might be characterized by high values of `TopSpeed_KmH` and low values of `AccelSec`.
   - This helps in understanding the distinguishing features of each cluster.

5. **Model Interpretation:**

   - The decision tree model is interpretable and allows for understanding how decisions are made to classify vehicles into different clusters.
   - It provides a clear and visual representation of the decision-making process.



simple segment evaluation plot

The objective is to visualize the cluster centroids for electric vehicle data based on `Range_Km` and `TopSpeed_KmH` features, with bubble size representing the percentage of total `Price(Inr)`.

Key Findings from the Output:

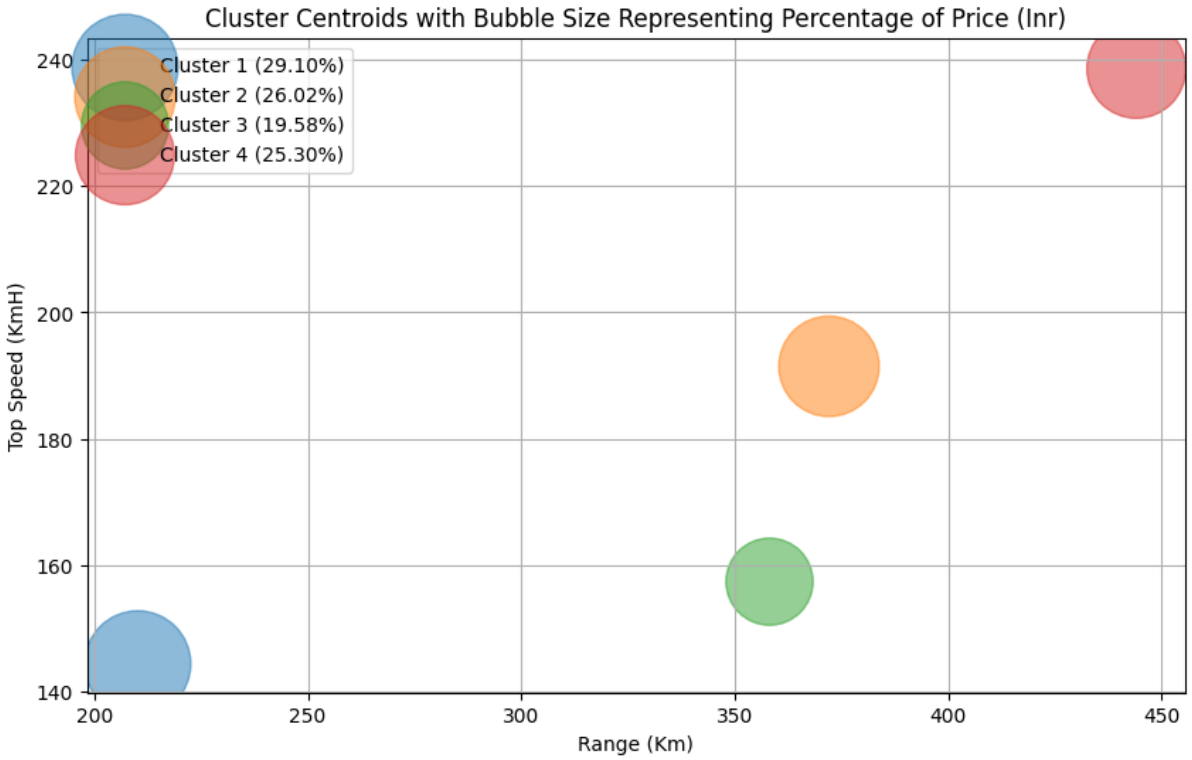
I. **Cluster Centroids:**
- The plot displays the centroids of four clusters, showing their mean values for `Range_Km` and `TopSpeed_KmH`.
- Each cluster centroid is represented by a bubble on the scatter plot.

II. **Bubble Size and Price Contribution:**
- The size of each bubble is proportional to the percentage of the total `Price(Inr)` contributed by the vehicles in that cluster.
- Larger bubbles indicate clusters that contribute a higher percentage of the total price.

III. **Cluster Characteristics:**
- The position of each bubble on the `Range_Km` and `TopSpeed_KmH` axes indicates the average performance of vehicles in that cluster.
- This helps in understanding the trade-offs between range and top speed for different clusters.

IV. **Cluster Insights:**
- **Cluster 1:** Vehicles in this cluster might have moderate range and top speed with a significant contribution to the total price, indicating balanced performance.
- **Cluster 2:** Vehicles with higher range but possibly lower top speed, contributing moderately to the total price.
- **Cluster 3:** Vehicles with high top speed but lower range, showing another segment of the market focusing on performance.
- **Cluster 4:** Vehicles with the highest contribution to total price and possibly a unique combination of range and top speed.

Visualization Benefits:
- The bubble plot provides a clear visual representation of how different clusters compare in terms of average range, top speed, and price contribution.
- It allows for quick identification of clusters that dominate the market in terms of total price.


Cluster Centroids with Bubble Size Representing Percentage of Price (Inr)
Cluster 1 (29.10%)
Cluster 2 (26.02%)
Cluster 3 (19.58%)
Cluster 4 (25.30%)
Top Speed (KmH)
Range (Km)
240
220
200
180
160
140
200
250
300
350
400
450