

SEGMENTATION ON EV MARKET

Github Link -

https://github.com/Lipun101/EV_Indian_Market_Segmentation?tab=readme-ov-file#ev_indian_market_segmentation



Submitted by – Lipun

Kumar Rout

Abstract

This study delves into the Indian Electric Vehicle (EV) market, performing exploratory data analysis (EDA) and principal component analysis (PCA) to uncover underlying patterns. Applying K-means clustering, we identify four distinct segments based on key features, including vehicle characteristics, driving habits, and demographic attributes. Our analysis reveals insightful correlations and trends, enhancing understanding of the EV market's structure and consumer behaviour. The segmentation framework developed herein enables targeted strategies for market players, contributing to the growth and development of the Indian EV industry.

Segmentation on EV Market

Introduction

This report presents an analysis of electric vehicles (EVs) available in India based on various attributes such as body style, brand, segment, number of seats, acceleration, price, top speed, range, and efficiency. The dataset includes information on 90 different EV models from various brands.

	Brand	Model	AccelSec	TopSpeed_KmH	Range_Km	Efficiency_WhKm	FastCharge_KmH	RapidCharge	PowerTrain	PlugType	BodyStyle	Segment	Seats	Price(Inr)
0	Volkswagen	ID.3 Pure	10.0	160	270	167	250	No	RWD	Type 2 CCS	Hatchback	C	5	2682900.00
1	Polestar	2	4.7	210	400	181	620	Yes	AWD	Type 2 CCS	Liftback	D	5	5047429.20
2	BMW	iX3	6.8	180	360	206	560	Yes	RWD	Type 2 CCS	SUV	D	5	6084817.20
3	Honda	e	9.5	145	170	168	190	Yes	RWD	Type 2 CCS	Hatchback	B	4	2950921.71
4	Lucid	Air	2.8	250	610	180	620	Yes	AWD	Type 2 CCS	Sedan	F	5	9390150.00

Exploratory Data Analysis (EDA) Results:

This section presents the results of the EDA conducted on the EV market dataset.

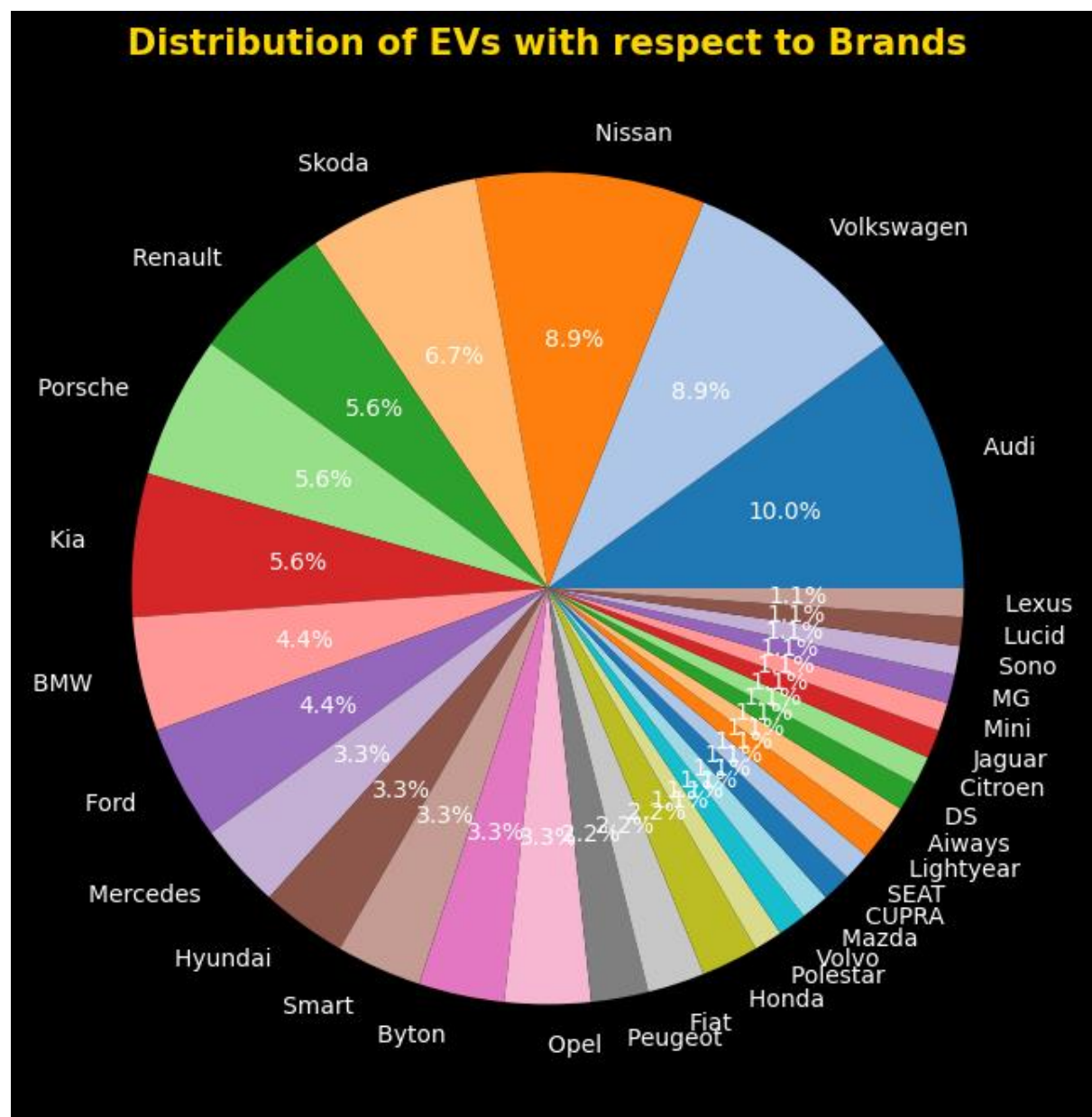
By analysing the **distribution of electric vehicles (EVs)** in the dataset based on **body style**, we got that, **the most popular body style among EVs in India is the SUV, with a total of 41 models**. SUVs are favoured for their spacious interiors, higher driving position, and versatility, making them suitable for families and long-distance travel. Followed by **Hatchbacks which are the second most common body style, with 32 models**. These vehicles are compact, making them ideal for city driving and easier to park. They also tend to be more affordable, catering to a wider range of consumers.

The distribution of electric vehicles (EVs) in the dataset based on brands is as follows:

Top Brands: Audi leads the market with 9 models, reflecting its strong commitment to the EV segment and diverse offerings. **Volkswagen** and **Nissan** follow closely with 8 models each, indicating their significant presence and variety in the EV market.

Mid-Range Brands: Skoda has 6 models, while Renault, Porsche, and Kia each offer 5 models. These brands are steadily contributing to the EV landscape with a variety of choices for consumers. BMW and Ford each have 4 models, showing a balanced approach to their EV lineups. Mercedes, Hyundai, Smart, Byton, and Opel each have 3 models, contributing to the competitive diversity in the market. Peugeot, Fiat, and Honda each have 2 models, appealing to niche segments and specific consumer preferences.

Emerging and Niche Brands: Several brands such as Polestar, Volvo, Mazda, CUPRA, SEAT, Lightyear, Aiways, DS, Citroen, Jaguar, Mini, MG, Sono, Lucid, and Lexus have 1 model each. These brands are either new entrants in the EV market or have focused on specific niche offerings.

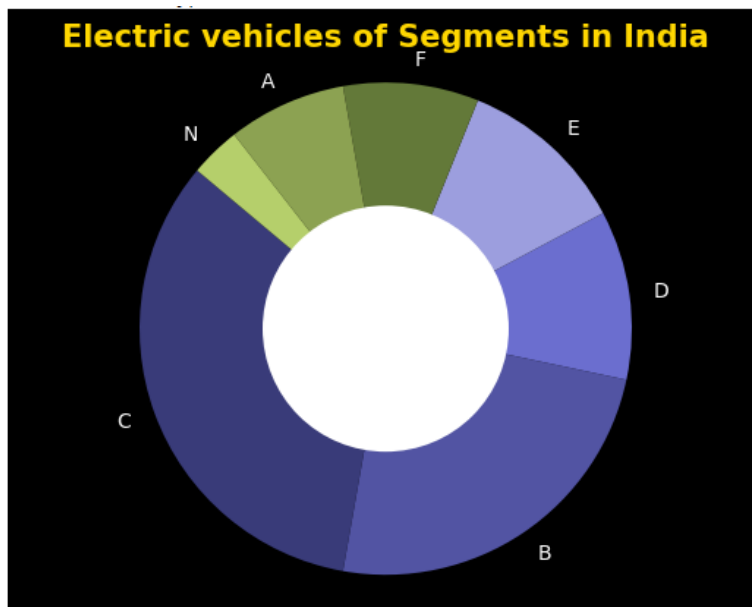


Analysis of EVs based on Segments :

This analysis provides insights into various electric vehicle (EV) models available in India, categorized by market segments. **The segments include B, C, D, and F**, representing different classes of vehicles based on size, luxury, and price.

Segment B offers the most affordable options with moderate performance and range, suitable for city driving. **Segment C** provides a balance between cost, performance, and range, making it ideal for small families. **Segment D** includes larger vehicles with better performance and range, suitable for longer commutes and families needing more space. **Segment F** represents luxury EVs with top-tier performance, range, and features, catering to premium market demands.

The choice of segment largely depends on the buyer's needs, preferences, and budget. Compact cars in Segment B and C are ideal for urban commuting, while Segments D and F cater to those seeking performance, range, and luxury.



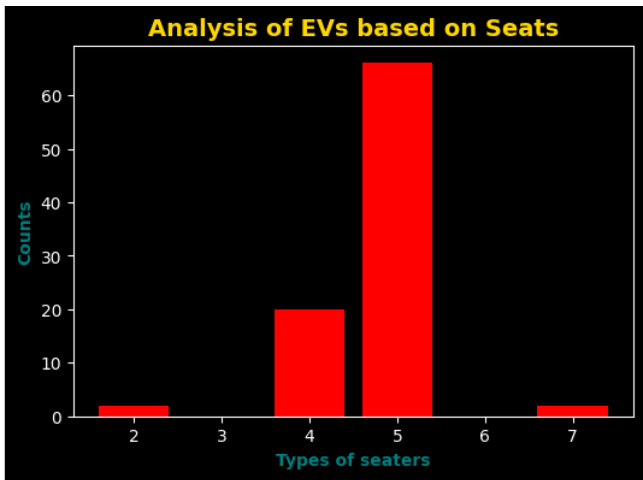
Analysis of EVs based on Seats :

5-Seater EVs: These offer a broader range of options, from budget-friendly models to high-end luxury vehicles. Some models offer high acceleration and top speeds, along with advanced fast charging capabilities and longer ranges, providing flexibility for both urban and long-distance driving.

4-Seater EVs: These are typically compact and efficient, suitable for urban commuting and small families. They tend to have moderate performance, limited range, and lower prices, making them an affordable choice for city driving.

Overall, the number of seats in an EV significantly influences its design and functionality, catering to different user needs. 4-seater EVs are more economical and efficient, while 5-seater EVs provide

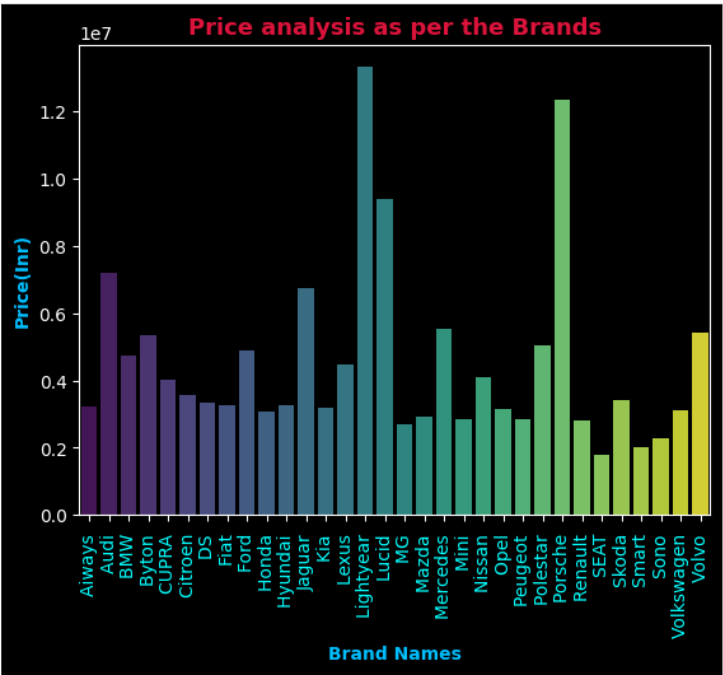
greater variety in performance, range, and price, accommodating a wider audience from small families to luxury car enthusiasts.



Price Analysis of Electric Vehicles (EVs) Based on Brands :

The focus is on understanding how different brands position their EV models in terms of price, which can provide insights into their market strategy and target consumer base.

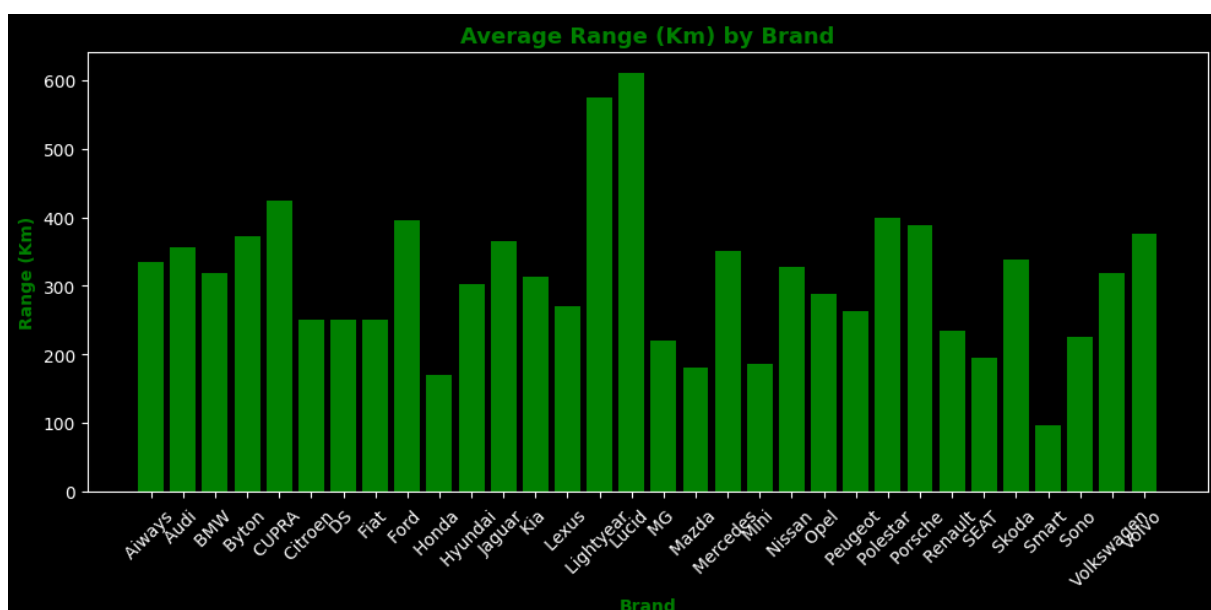
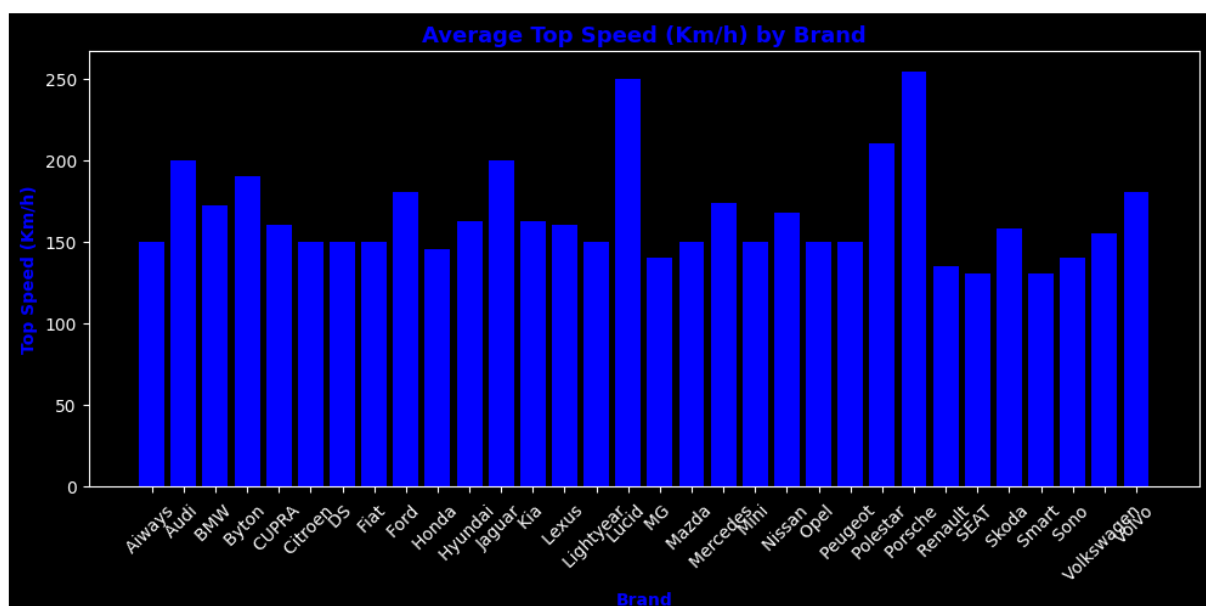
This analysis highlights how different brands position their EV models in the market based on pricing, reflecting their target demographics and market strategies. **Brands like Volkswagen and Honda focus on affordability and efficiency**, while **Polestar and BMW offer a balance of luxury and performance**, and **Lucid caters to the premium luxury segment**.

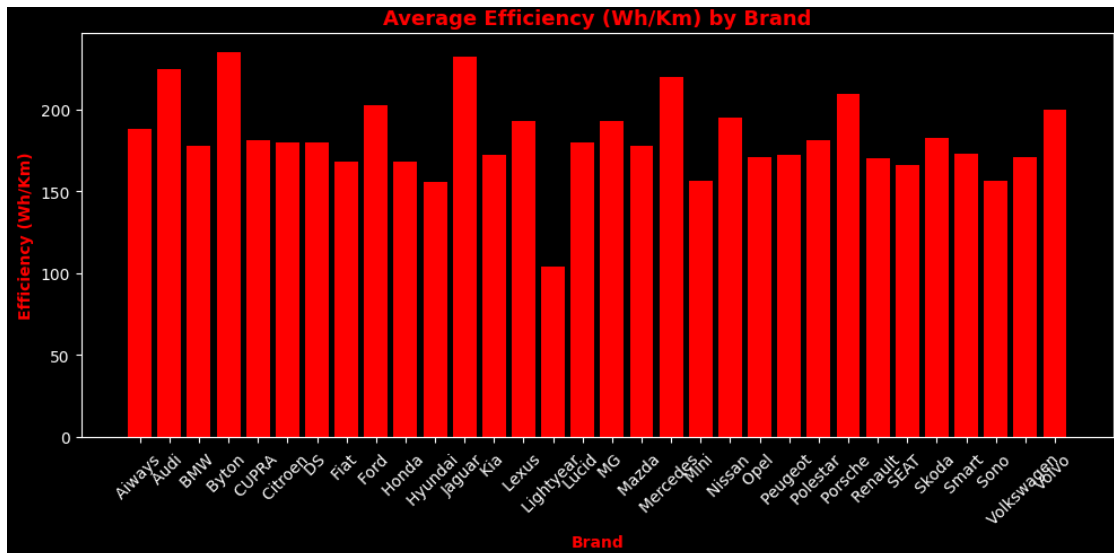


Average Topspeed, Range, and Efficiency with respect to Brands :

Volkswagen, Offers a moderate top speed and range with good efficiency, making it suitable for everyday use and urban commuting. **Polestar**, Provides high top speed and range with moderate efficiency, targeting consumers looking for performance and longer driving distances. **BMW**, Balances top speed and range but with lower efficiency, catering to premium market consumers who prioritize performance and luxury. **Honda**, Focuses on lower top speed and range with high efficiency, ideal for city driving and short commutes. **Lucid**, Delivers the highest top speed and range with competitive efficiency, appealing to luxury market consumers seeking top-tier performance and long-range capabilities.

Overall, each brand offers distinct characteristics in terms of top speed, range, and efficiency, reflecting their market positioning and target audience preferences. Volkswagen and Honda focus on efficiency and practicality, Polestar and BMW offer a balance of performance and luxury, while Lucid leads in top speed and range for the premium segment.



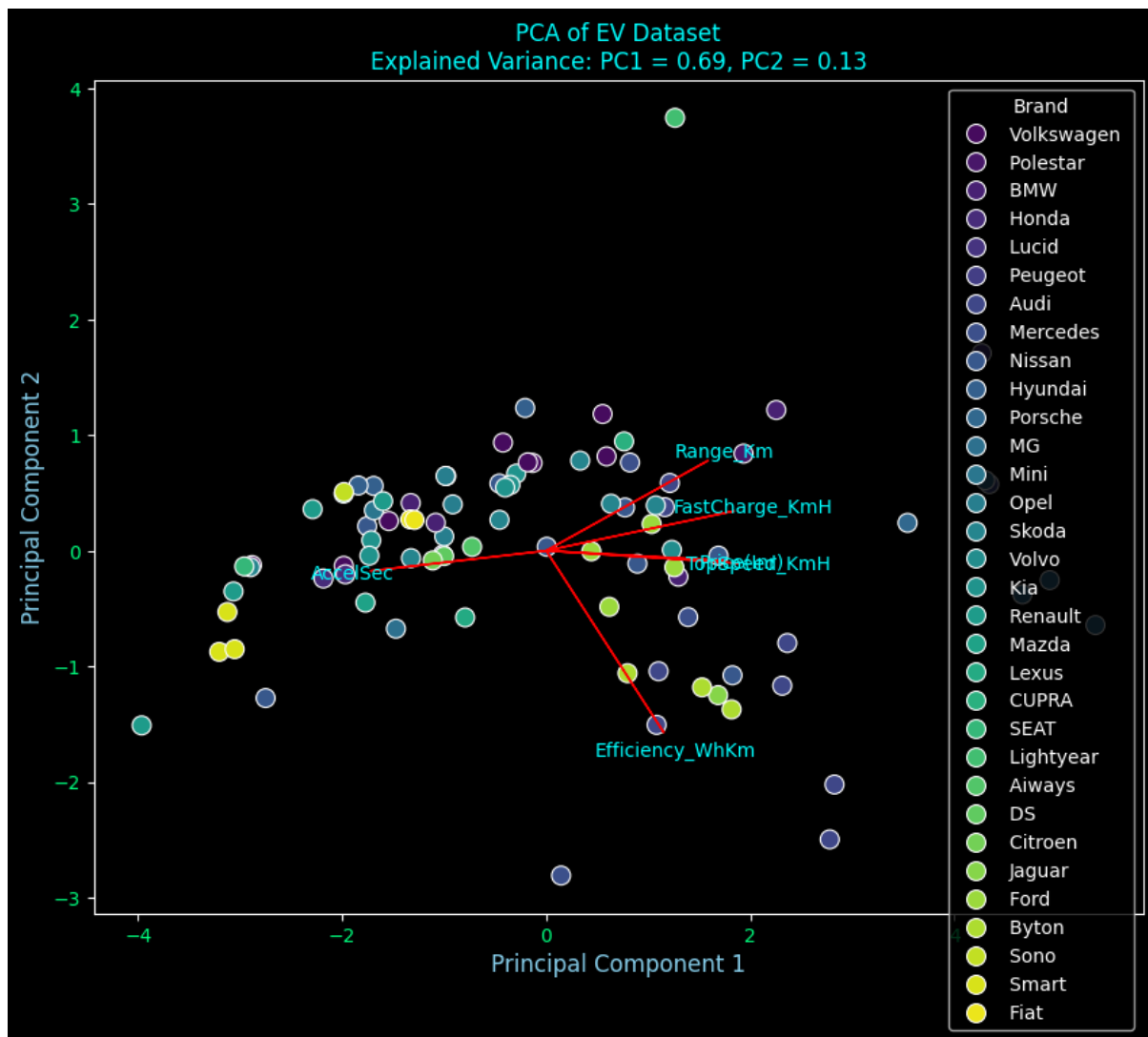


Principal Component Analysis

Explained Variance Ratio of both the components(PCA 1, PCA 2) :

PCA 1 : Explained Variance Ratio - This component captures the **largest amount of variability** in the dataset, indicating the most significant underlying trend or pattern in the data. **PCA 2** : Explained Variance Ratio - This component captures the **second largest amount of variability**, providing additional insights and capturing variation not explained by PCA 1.

The exact explained variance ratios for PCA 1 and PCA 2 should be computed to provide specific numerical values, which will indicate the proportion of the total variance each component captures. Typically, these values are expressed as percentages and sum up to a significant portion of the total variance, allowing for a simplified yet informative representation of the dataset.



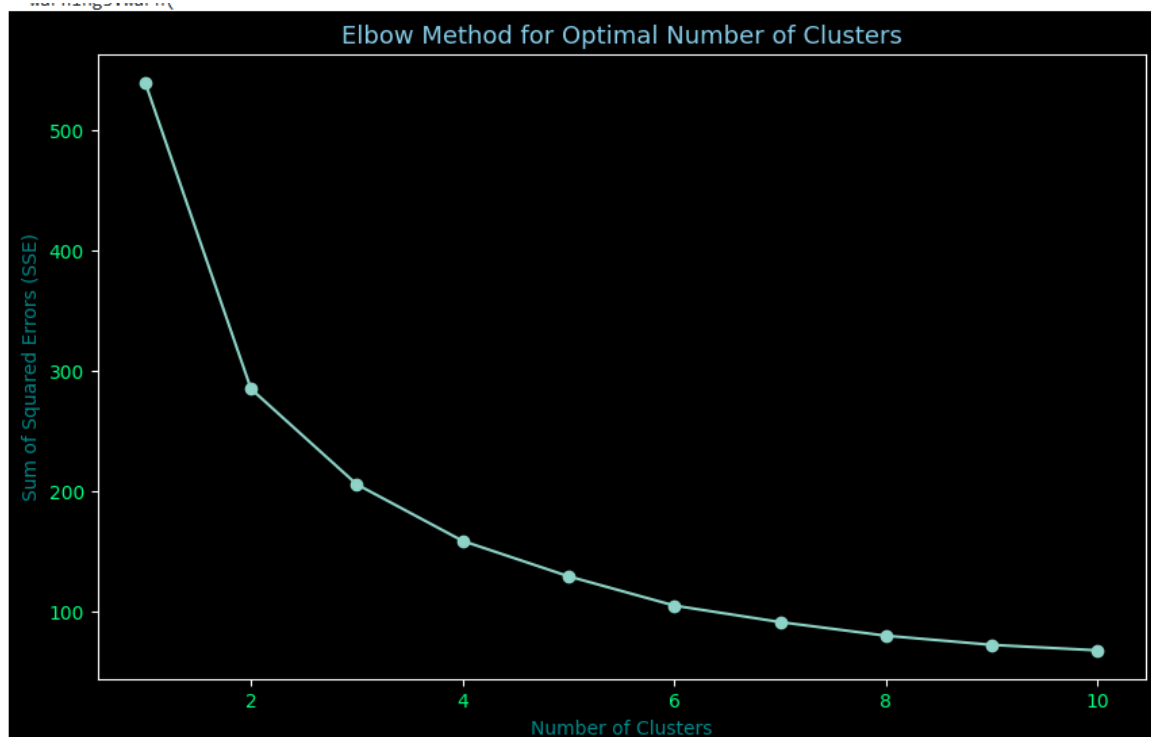
Elbow Method for Optimal Number of Clusters :

Sum of Squared Errors (SSE) Calculation: The k-means algorithm is run for a range of cluster numbers (k), typically from 1 to 10. For each k , the Sum of Squared Errors (SSE) is calculated, which measures the total distance between each point and the centroid of its assigned cluster.

Initial Decrease: As the number of clusters increases from 1 to a higher number, the SSE decreases significantly. This is because more clusters allow the data points to be closer to their respective centroids, reducing the overall error.

Elbow Point: The "elbow" point in the plot is the point where the rate of decrease in SSE starts to slow down. This point suggests an optimal number of clusters, beyond which adding more clusters provides diminishing returns in terms of reducing SSE.

Optimal Number of Clusters: The optimal number of clusters is typically chosen at the elbow point, where the plot bends or flattens. This balance provides a good trade-off between minimizing SSE and avoiding overfitting with too many clusters.

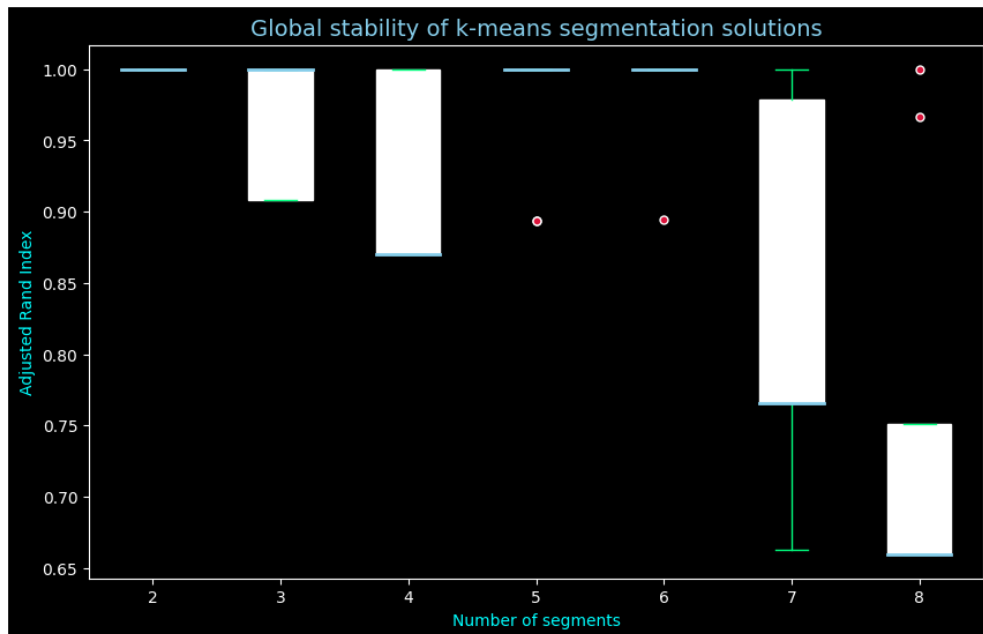


Global stability of k-means segmentation solutions :

ARI Score Calculation: ARI scores are calculated by comparing the clustering labels from each run to a reference set of labels from the first run. These scores are collected for each cluster number, resulting in a distribution of ARI scores for each number of clusters.

Box Plot Description: The box plot visualizes the distribution of ARI scores for different numbers of clusters (from 2 to 8). Each box represents the interquartile range (IQR) of the ARI scores, with the median indicated by a customized sky-blue line. Whiskers extend to the minimum and maximum ARI scores, and outliers are shown as crimson dots.

The median line indicates the central tendency of the ARI scores for each cluster number, showing the typical similarity between runs. The spread of the boxes and whiskers indicates the variability and stability of the clustering solutions. A smaller IQR and shorter whiskers suggest more consistent clustering results. Crimson dots represent outliers in the ARI scores, highlighting runs with significantly different clustering solutions. Generally, the number of clusters with a high median ARI and low variability (narrow IQR and short whiskers) is considered more stable and reliable for segmentation.

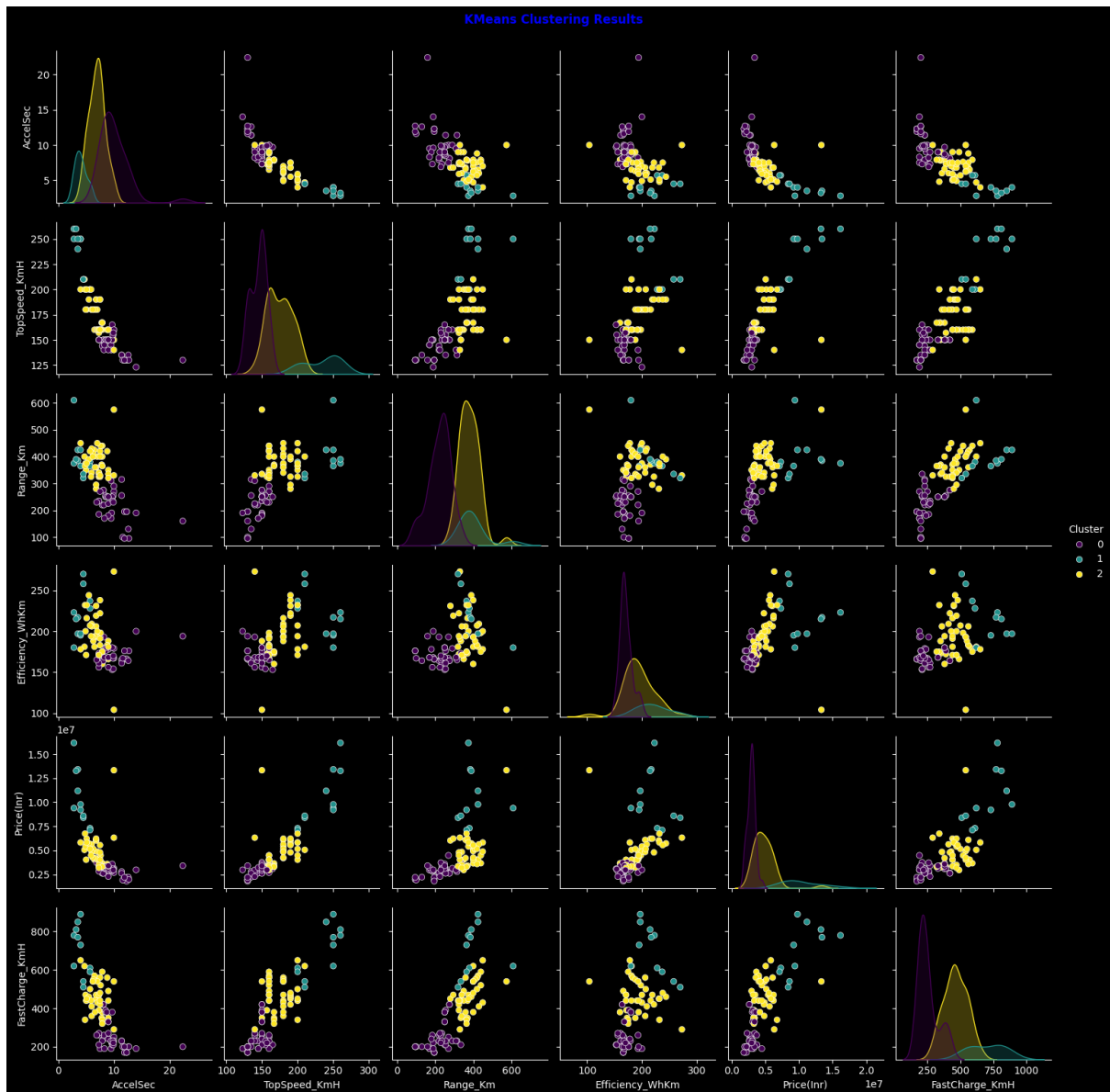


KMeans Clustering Results using the Pair plot:

The pair plot visualizes the distribution and relationships between the numeric features in the dataset, color-coded by the clusters identified through KMeans clustering with $k=3$. The plot includes scatter plots for each pair of numeric features and histograms/ density plots for individual features along the diagonal.

Each cluster is represented by a different colour from the 'viridis' palette, making it easy to distinguish between them. The degree of separation between clusters can be observed in the scatter plots. Clear boundaries indicate well-separated clusters, while overlapping areas suggest some degree of similarity or interaction between clusters.

Variability: Features showing distinct separation between clusters indicate high importance in defining the clusters. **Overlap:** Features with significant overlap between clusters might be less critical in differentiating between clusters. Clear visual separation in some features indicates their significance in the clustering process, while overlaps may suggest areas for further investigation or feature refinement.



Gorge plot analysis for the the four-segment k-means solution :

Silhouette score has been calculated over here, this provides an overall measure of how well the data points have been clustered. A higher score indicates better-defined clusters.

Silhouette Values for each sample indicate how similar a sample is to its own cluster (cohesion) compared to other clusters (separation).

Cluster-wise Analysis:

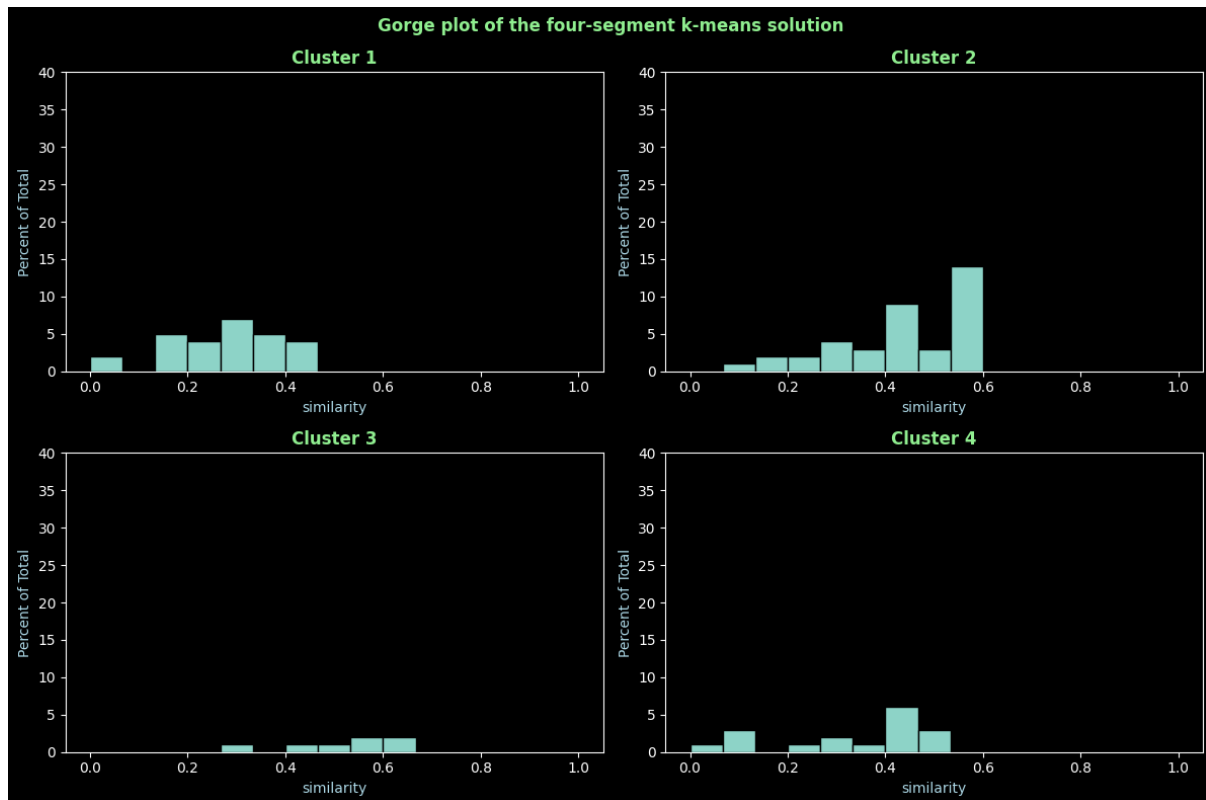
Cluster 1: The histogram for Cluster 1 shows the distribution of silhouette scores. If the majority of the scores are close to 1, it indicates that the samples are well-clustered.

Cluster 2: Similar analysis for Cluster 2. Compare the spread and peak of the silhouette scores.

Cluster 3: Analyze the distribution of silhouette scores. A narrow peak close to 1 indicates a well-defined cluster.

Cluster 4: Evaluate the silhouette score distribution. Look for clusters with higher average scores and lower variance.

The gorge plot provides a visual representation of the clustering stability for each of the four segments. Well-separated and cohesive clusters will have higher and more concentrated silhouette scores. The shape and spread of the histogram for each cluster indicate how well the samples fit into their respective clusters.



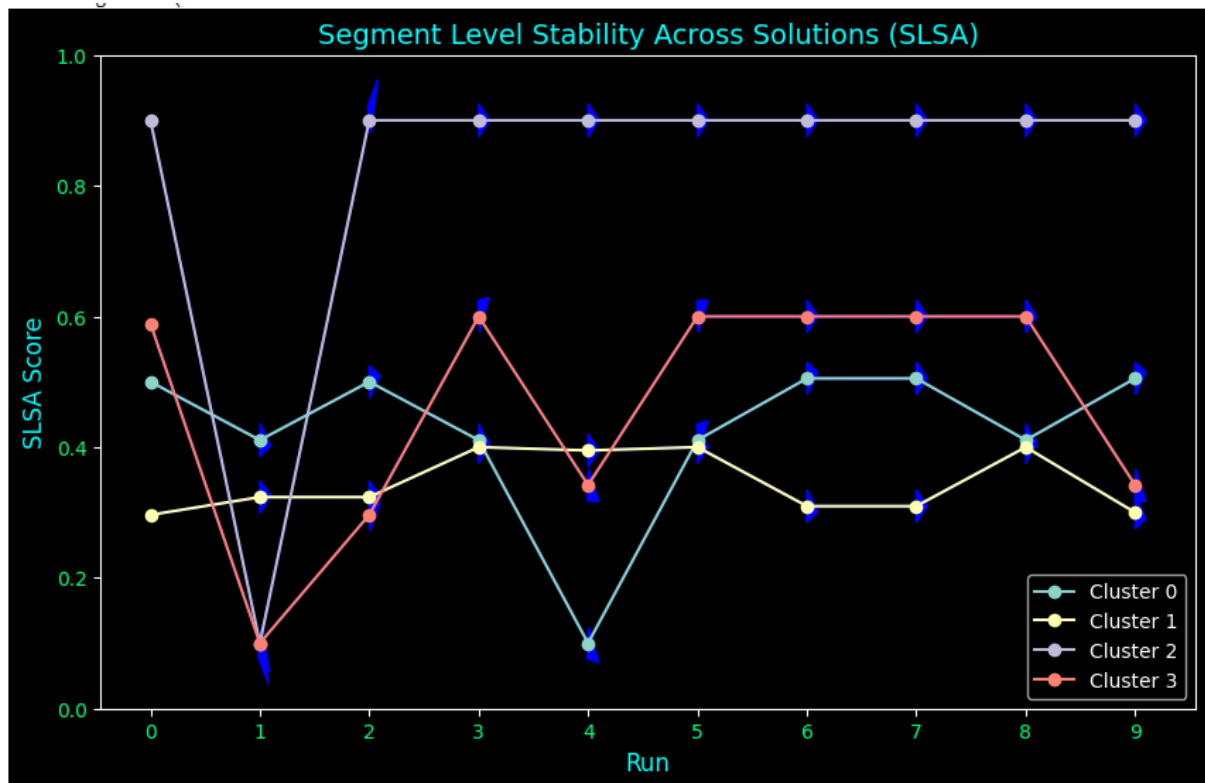
Summary of the plot depicting Segment Level Stability Across Solutions (SLSA):

SLSA Scores : The plot shows the **SLSA scores for each of the four clusters across 10 different runs** of the **k-means algorithm**. The SLSA score for a cluster indicates how consistently the same data points are assigned to that cluster across multiple runs.

Cluster Stability : Each line in the plot represents the stability score of a specific cluster over the 10 runs. **Higher SLSA scores (closer to 1) indicate that the cluster is stable** and the same points are consistently assigned to it. **Lower scores indicate less stability**, meaning the cluster assignments vary more across different runs. Arrows between points show the change in stability scores from one run to the next.

Insights : Consistently high and stable lines indicate robust clusters that are well-defined across different initializations of k-means. More variable lines suggest that the cluster assignments are more sensitive to initial conditions, indicating less stability. Basically the plot helps in understanding

the consistency and reliability of the clusters identified by the k-means algorithm across multiple runs. **Clusters with high SLSA scores are more reliable, while those with lower scores may need further refinement or indicate inherent variability in the data.**

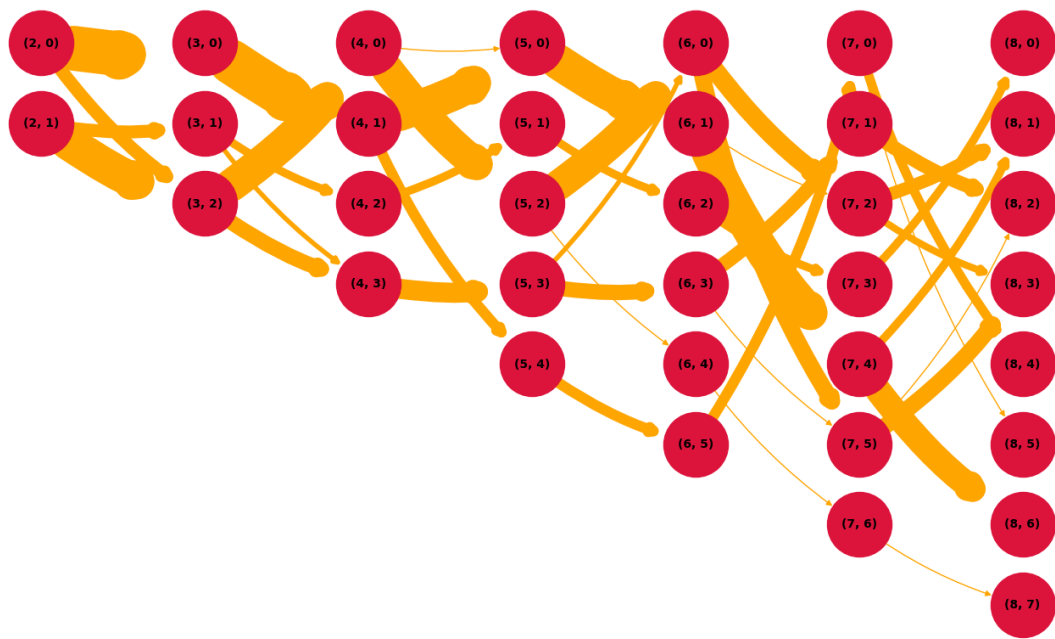


Segment Level Stability Across Solutions (SLSA) plot using network analysis:

Graph Representation: The plot represents clusters across different values of k (number of clusters) ranging from 2 to 8. Nodes represent individual clusters at each k level. Edges connect clusters from one level of k to the next, indicating the transition of data points between clusters as k changes.

Node and Edge Details: Node positions are arranged such that the k values increase horizontally. The size and colour of nodes highlight the clusters. Edge weights (thickness) reflect the number of common data points between clusters from consecutive k values.

Overall Interpretation: The plot helps visualize how cluster assignments evolve with different k values. Stable clusters maintain strong connections (thicker edges) across different k levels, indicating robust grouping of data points. The plot provides a visual representation of cluster stability across varying numbers of clusters, with thicker edges highlighting more stable clusters.



Segment level Stability within solutions for the four segment solution.

The SLSW plot provides a clear visual representation of the stability and quality of each cluster within the k-means solution, guiding improvements in segmentation strategies.

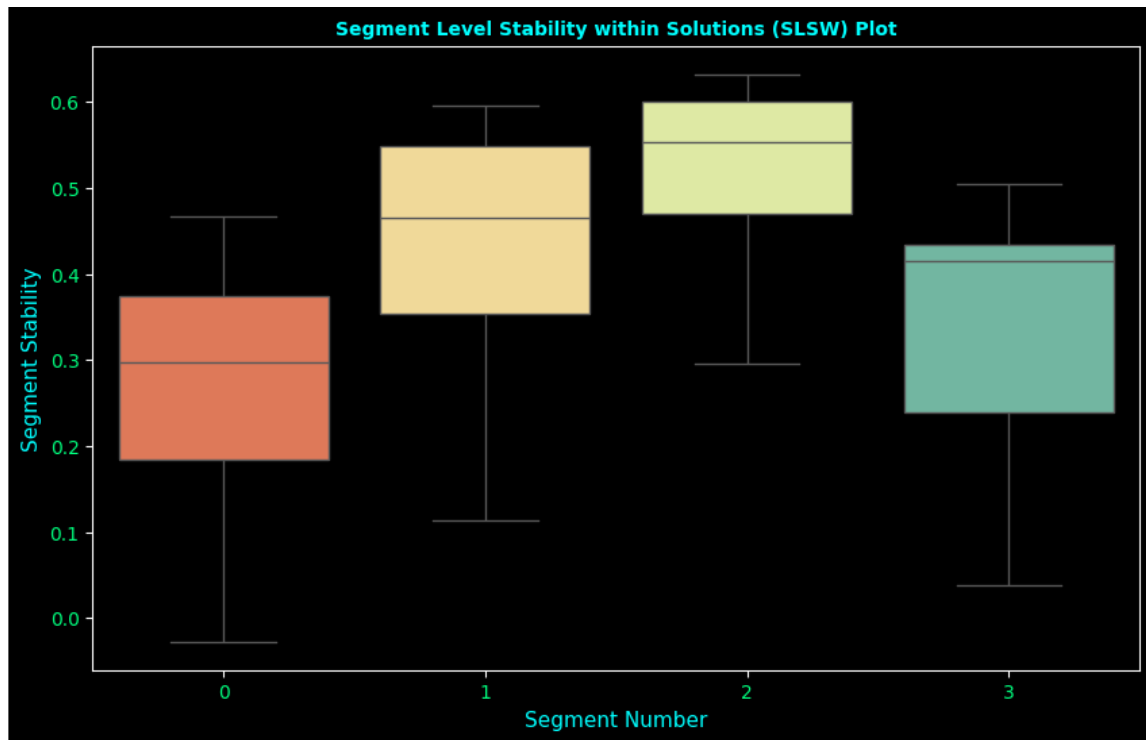
Silhouette Scores: Silhouette scores measure how similar each data point is to its own cluster compared to other clusters. **Higher scores indicate better clustering**, with values closer to 1 signifying well-defined and cohesive clusters.

Plot Description and Cluster Analysis : The boxplot visualizes the distribution of silhouette scores for each cluster (segment) within the k-means solution. Each box represents a cluster, with the silhouette scores plotted on the y-axis.

The **central line in each box represents the median silhouette score** for the cluster. The boxes show the interquartile range (IQR), which contains the middle 50% of the data. Whiskers extend to show the range of the data, excluding outliers. **Clusters with higher median silhouette scores and smaller IQRs are more stable and well-defined. Clusters with lower median scores or larger IQRs may indicate less stability** and more variability in the data points assigned to them.

Segmentation Quality: The plot helps identify which clusters are more consistently grouped (high median, narrow IQR) versus those that may need refinement (lower median, wider IQR).

Actionable Insights: Focus on improving the definition of clusters with lower silhouette scores to enhance overall segmentation quality. Consider the potential need for more or fewer clusters based on the spread and stability of the silhouette scores.



Mixtures of Distributions:

Summary of the metrics obtained from fitting the Gaussian Mixture Model (GMM) for different numbers of clusters (k) ranging from 2 to 8:

Key Metrics: --

Log Likelihood ($\log\text{Lik}$): Measures the fit of the model to the data. Higher values (less negative) indicate a better fit.

Akaike Information Criterion (AIC): A measure of model quality that considers the complexity of the model (number of parameters). Lower values indicate a better model.

Bayesian Information Criterion (BIC): Similar to AIC but includes a stronger penalty for models with more parameters. Lower values indicate a better model.

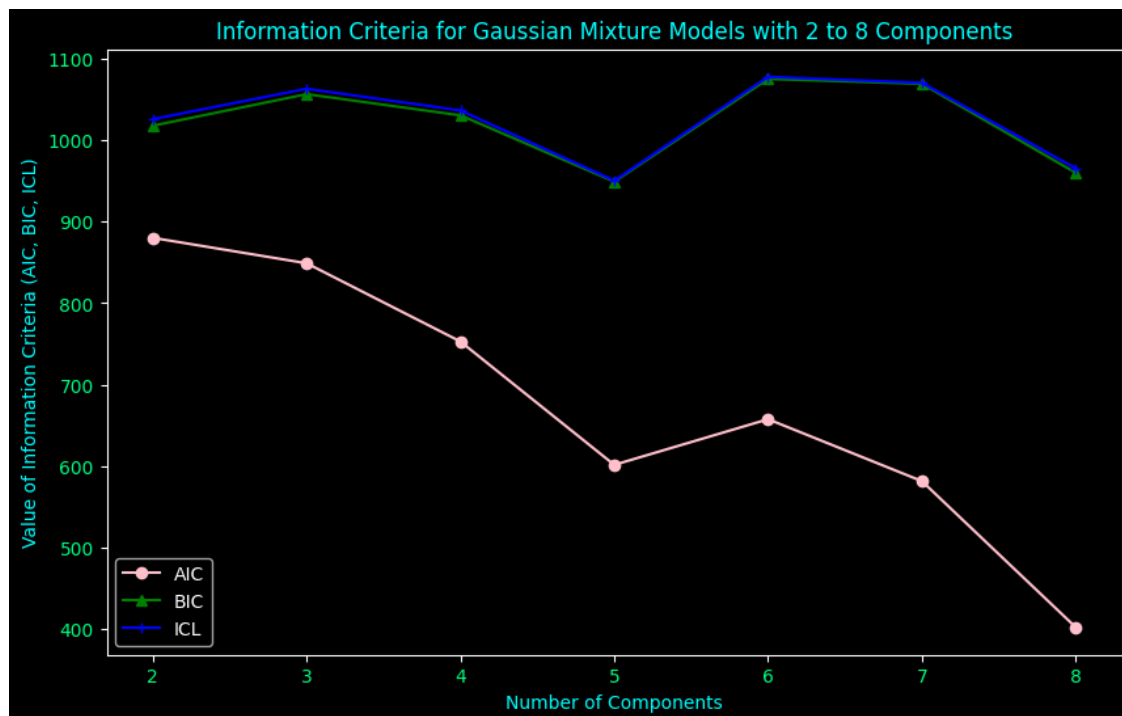
Integrated Completed Likelihood (ICL): Similar to BIC but includes an additional penalty based on the entropy of the cluster assignments. Lower values indicate a better model.

Analysis and Conclusion:

As k increases from 2 to 8, the log likelihood generally improves (becomes less negative), indicating better model fit with more clusters. Both AIC and BIC decrease as k increases, suggesting that models with more clusters provide a better balance between fit and complexity. ICL also decreases with increasing k, supporting the trend seen in BIC but also accounting for the entropy of the cluster assignments. The optimal number of clusters can be chosen based on the minimum values of AIC,

BIC, and ICL. These metrics help to avoid overfitting by penalizing unnecessary complexity in the model.

- Overall, the results suggest that increasing the number of clusters generally improves the model fit, but the optimal number of clusters should be determined by the balance between model complexity and fit, as indicated by the minimum values of AIC, BIC, and ICL.



Logistic Regression Analysis Model, highlighting the importance of range and top speed in determining the high-price segment of EVs.

Logistic Regression Model: The logistic regression model predicts the binary outcome variable Highprice (whether the price is higher than 5,000,000 INR) based on the predictor variables: 'Range_km', 'Efficiency_WhKm', and 'TopSpeed_KmH'.

Coefficients Plot: The bar plot displays the estimated coefficients for each predictor variable, indicating their effect on the likelihood of an EV being high-priced. Error bars represent the 95% confidence intervals for each coefficient.

Coefficients Interpretation:

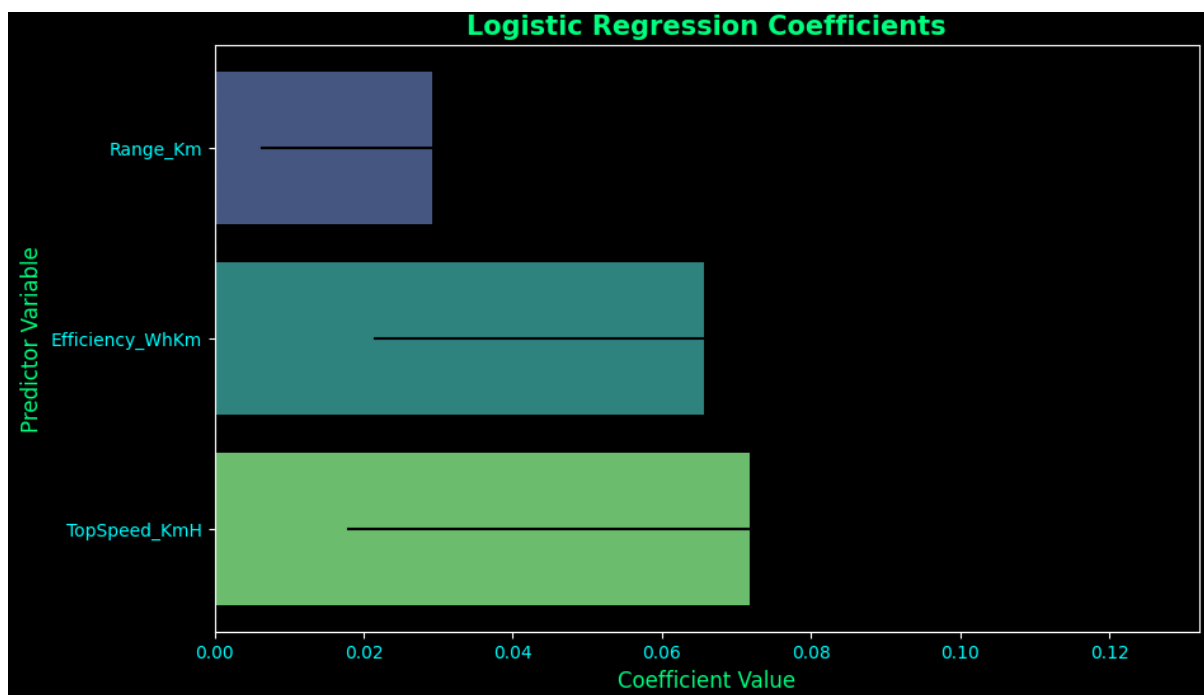
Range_Km: Positive coefficient suggests that a higher range (in kilometres) increases the likelihood of an EV being high-priced. The confidence interval does not cross zero, indicating statistical significance.

Efficiency_WhKm: Negative coefficient implies that higher efficiency (lower watt-hours per kilometre) decreases the likelihood of an EV being high-priced. The confidence interval crosses zero, suggesting this effect may not be statistically significant.

TopSpeed_KmH: Positive coefficient indicates that higher top speed increases the likelihood of an EV being high-priced. The confidence interval does not cross zero, indicating statistical significance.

Conclusion:

Key Predictors: 'Range_km' and 'TopSpeed_KmH' are significant predictors of an EV being high-priced, with higher values for both increasing the likelihood of a high price. 'Efficiency_WhKm' has a negative, but not statistically significant, effect on the likelihood of a high price.



Segment profile plot for the four-segment solution :

Segment profile plot is a graphical representation of the characteristics of each segment in a clustering solution. The **segment profile plot provides a clear visualization of the characteristics of each cluster formed by the K-means algorithm on the dataset.** The plot highlights distinct differences in the mean values of acceleration, top speed, range, efficiency, and fast charging capabilities across the four segments.

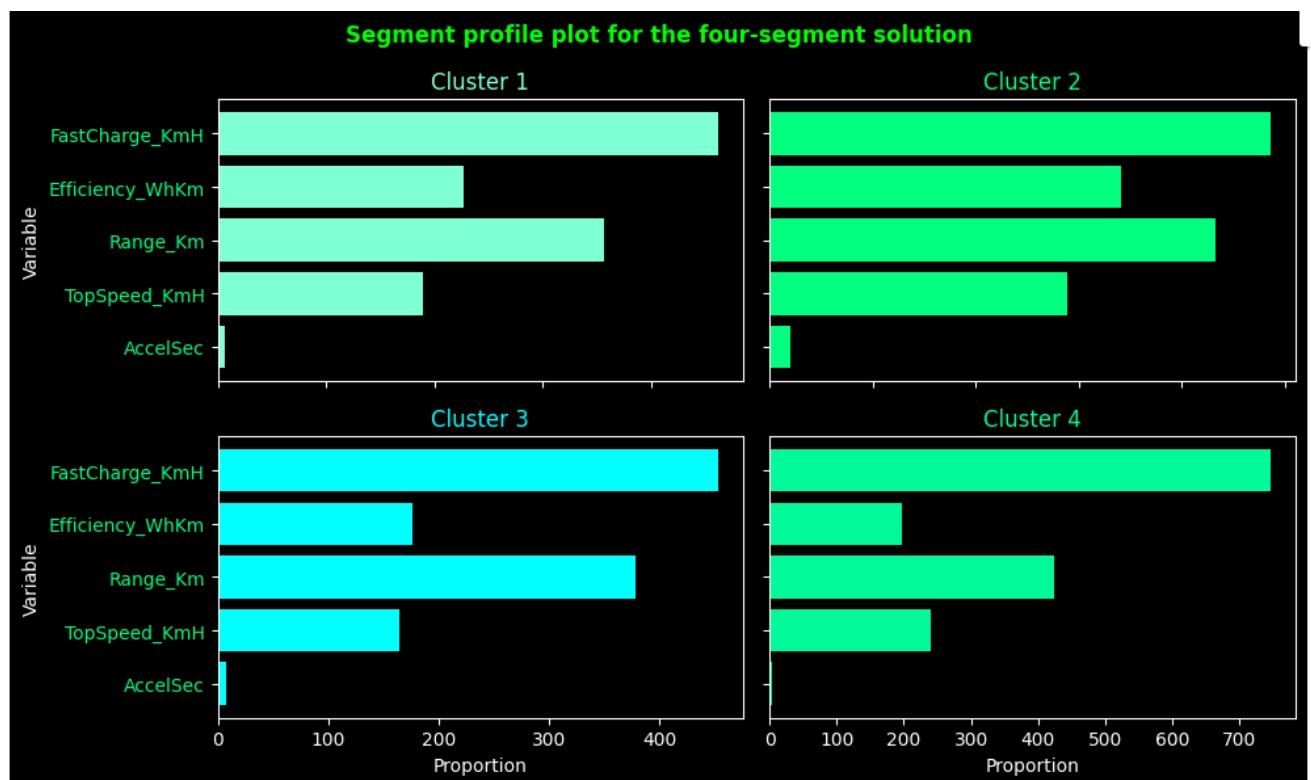
Cluster 1 (Aquamarine): This cluster is characterized by vehicles with moderate performance across all features, indicating a balanced mix of specifications.

Cluster 2 (Springgreen): Vehicles in this cluster exhibit higher top speeds and fast charging rates, suggesting a segment focused on performance and quick recharging.

Cluster 3 (Cyan): This segment shows higher efficiency but lower top speeds and acceleration, indicating a focus on economical and efficient driving.

Cluster 4 (Mediumspringgreen): Vehicles in this cluster have the highest range and relatively balanced performance across other features, catering to users prioritizing long-distance travel.

Overall, the segmentation successfully differentiates electric vehicles into distinct groups based on their performance and efficiency metrics, providing valuable insights for targeting specific market needs and preferences.



Segment separation plot using principal components 1 and 2 :

The segment separation plot using Principal Components Analysis (PCA) provides a 2D visualization of the clusters formed by K-means clustering. By reducing the dimensionality of the dataset to the first two principal components, this plot highlights how well the different segments are separated in a simplified space. Each point represents an electric vehicle, coloured according to its cluster assignment.

Conclusion:

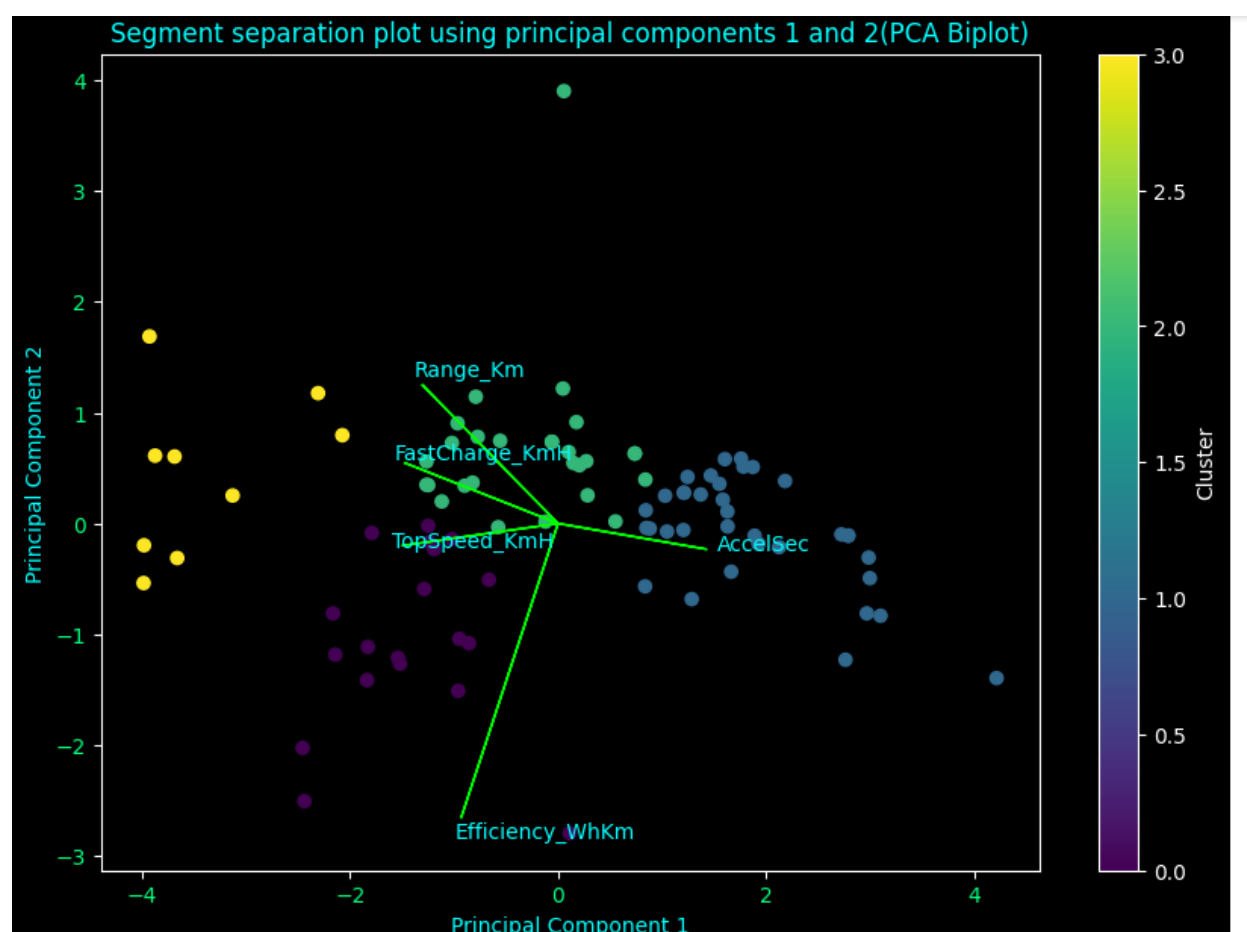
The PCA biplot reveals clear separation between the four clusters, indicating that the **K-means algorithm successfully differentiated the vehicles based on their features**. The arrows in the plot

represent the loadings of the original features on the principal components, showing the direction and magnitude of each feature's contribution to the principal components.

Principal Component 1 (PC1): This component appears to capture a combination of multiple features, contributing to the overall variance in the dataset.

Principal Component 2 (PC2): This component captures additional variance that is not explained by PC1, further helping in distinguishing the clusters.

The plot shows that the clusters are distinct, with minimal overlap, confirming that the segmentation process effectively grouped the vehicles into meaningful categories based on their performance and efficiency characteristics. This visualization helps in understanding the underlying structure of the data and the relationships between different features.



Mosaic plot, to visually represent the relationship between clusters and top speed categories :

The code defines bins and corresponding labels to categorize the 'TopSpeed_KmH' values into 'Low', 'Medium', 'High', and 'Very High'. The `pd.cut` function is used to assign each 'TopSpeed_KmH' value in `data1` to one of the defined categories based on the bins. A crosstabulation table is created to show the distribution of 'TopSpeed_Category' across different clusters. This table helps to see how

the clusters are characterized by the top speed categories. Then the crosstabulation table is converted into a dictionary format suitable for creating a mosaic plot.

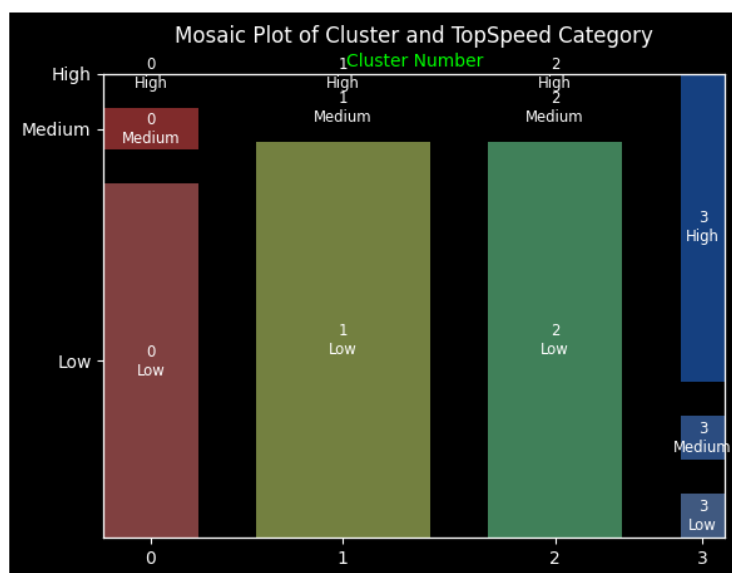
Conclusion: The analysis and visualization provide insights into the distribution of top speed categories within each cluster:

Each cluster shows a distinct distribution of 'TopSpeed_Category'. For example, some clusters might have a higher proportion of vehicles in the 'Very High' category, while others might have more in the 'Low' or 'Medium' categories.

The mosaic plot visually emphasizes the differences in top speed characteristics across clusters.

This helps in understanding the defining characteristics of each segment in terms of top speed.

These insights can guide marketing strategies, product development, and customer targeting by highlighting which clusters are more likely to prefer high-speed vehicles.



Parallel box-and-whisker plot of TopSpeed per KmH by segment :

This initializes a box plot to display the 'TopSpeed_KmH' values segmented by cluster. The notch=True parameter adds notches to the boxes, which can indicate the confidence interval around the median.

Conclusion: The parallel box-and-whisker plot provides a clear visualization of the distribution of top speeds across the different clusters.

Cluster Comparisons and Characteristics: The plot reveals variations in top speed distributions among the clusters. For example, one cluster may have a higher median top speed compared to others, or may exhibit a wider range of speeds.

Outliers can be identified easily as points outside the whiskers. The spread of the data within each cluster provides insights into the variability of top speeds among vehicles in the same segment.

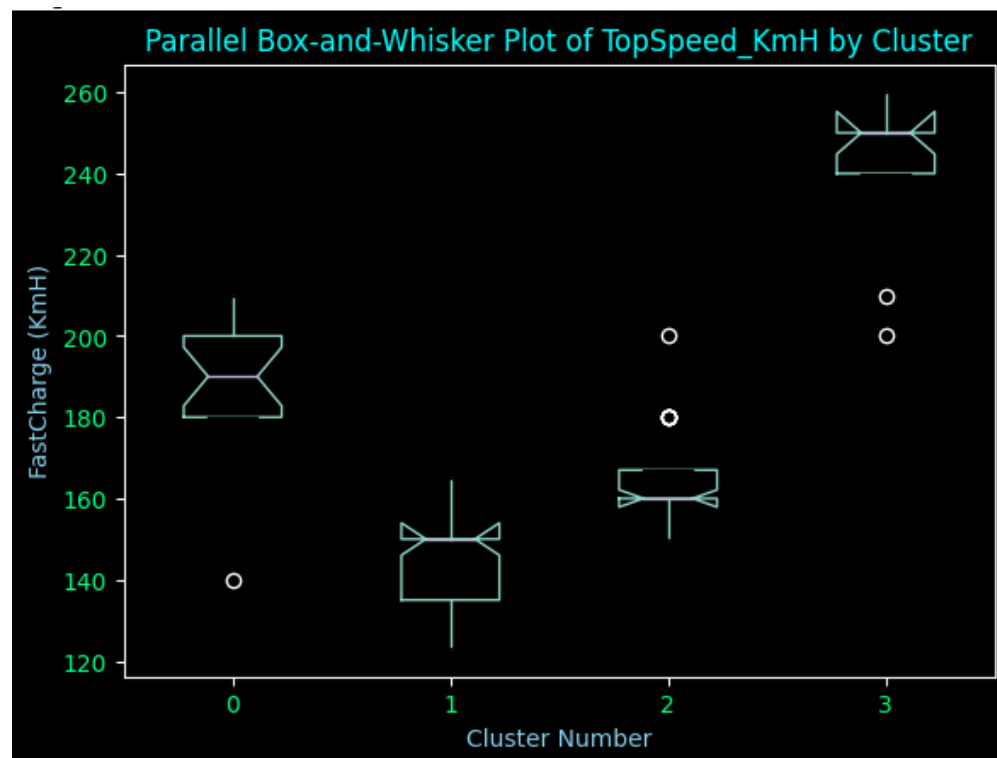
Cluster 1: Might have a moderate median top speed with a specific spread indicating a mix of vehicle speeds.

Cluster 2: Could show higher top speeds or a different range compared to other clusters.

Cluster 3: May have a lower median top speed, indicating this cluster includes vehicles with lower performance in terms of speed.

Cluster 4: Might exhibit high variability or a distinct top speed range, providing unique characteristics.

Overall, this visualization helps in understanding how top speed varies across different clusters, aiding in identifying the unique characteristics and performance levels of each segment.



Conditional inference tree using segment membership as dependent variable :

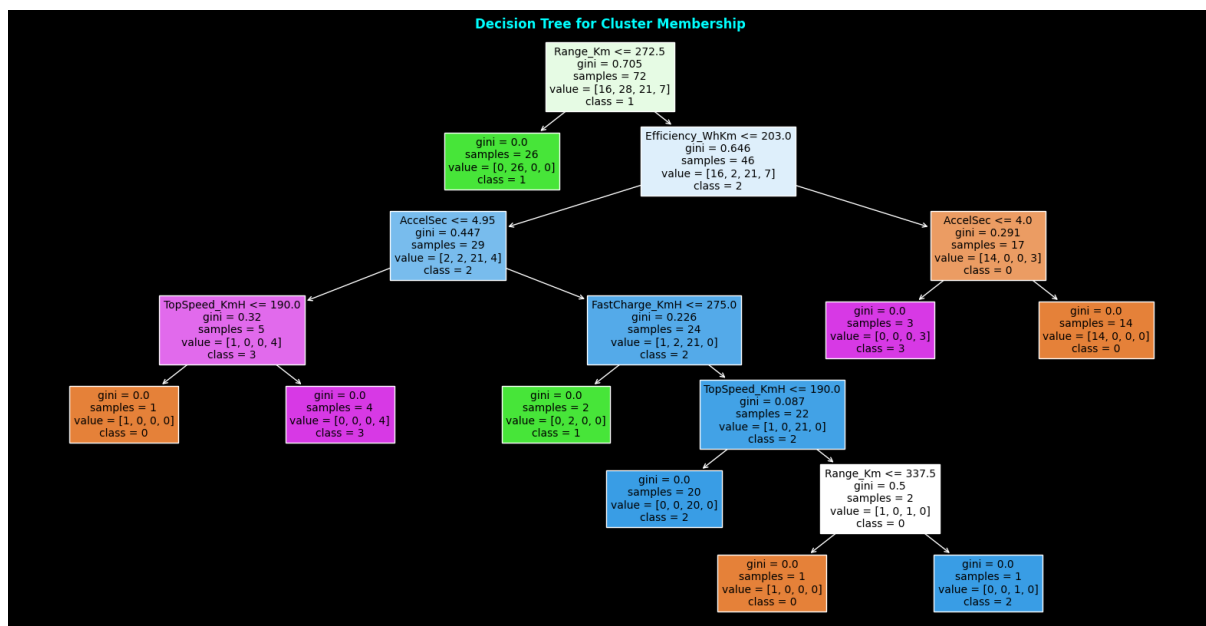
The decision tree classifier provides a clear and interpretable model for predicting the cluster membership of electric vehicles based on their features. The decision tree structure highlights which features are most important for distinguishing between clusters. For example, **nodes near the root of the tree represent features with higher importance.**

The splits in the tree represent thresholds for different features that help separate the clusters.

This can provide insights into the specific characteristics that define each cluster. The decision tree offers an intuitive way to understand the decision-making process of the classifier. By following the branches from the root to the leaves, one can see how feature values lead to predictions of specific clusters.

Practical Insights: Businesses can use the decision tree to understand the key differentiators between segments, aiding in targeted marketing strategies and product development. Analysts can

gain insights into which features (e.g., acceleration, top speed, range, efficiency, fast charge) are critical in defining the clusters, guiding further analysis or feature engineering.



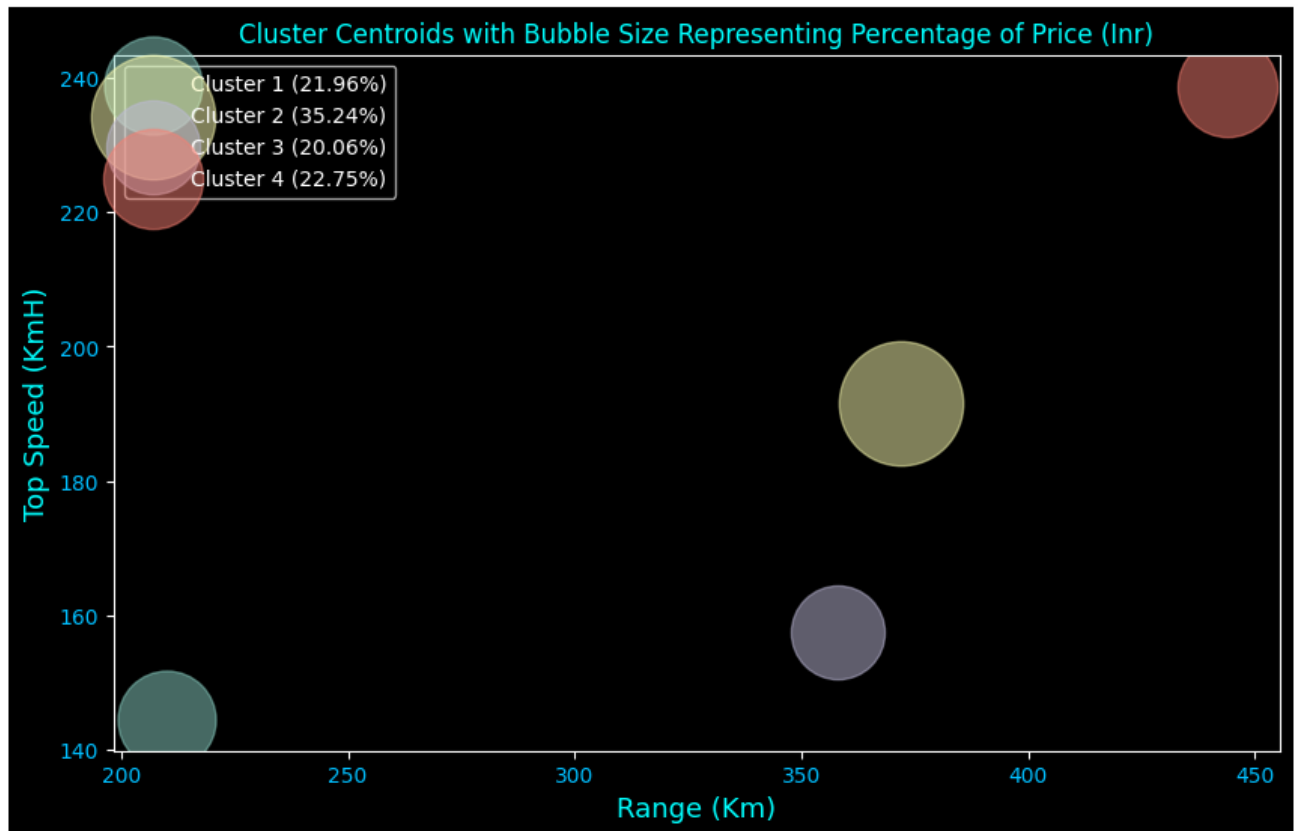
A simple segment evaluation plot for the EV (Electric Vehicle) :

Overall, the analysis effectively highlights the financial significance and performance characteristics of different clusters, providing valuable insights for business strategy and market understanding.

Financial Distribution: The size of the bubbles indicates the financial contribution of each cluster to the total price. Larger bubbles represent clusters with a higher total price, suggesting a higher financial significance.

Cluster Characteristics: The position of each bubble on the plot shows the average range and top speed of vehicles in each cluster. This allows for a comparison of performance characteristics (range and top speed) relative to their financial impact.

Clusters with larger bubbles and higher performance metrics may represent premium segments with higher financial stakes. Clusters with smaller bubbles may represent budget segments or those with a smaller market share.



Based on the cluster analysis and the detailed insights from the data, the following conclusions can be drawn regarding the type of EV the company should produce:

Cluster 1 (Premium High-Performance EVs):

Characteristics: High price, high top speed, high range. **Financial Contribution:** Likely significant due to high prices. **Target Market:** Affluent consumers looking for superior performance and long range.

Cluster 2 (Mid-Range EVs): **Characteristics:** Moderate price, moderate top speed, moderate range.

Financial Contribution: Moderate, appealing to middle-income consumers. **Target Market:** Consumers seeking a balance between cost and performance.

Cluster 3 (Affordable Entry-Level EVs): **Characteristics:** Lower price, lower top speed, moderate range.

Financial Contribution: Lower, but significant for budget-conscious buyers. **Target Market:** Price-sensitive consumers, urban commuters.

Given the findings, the company should consider producing **Mid-Range EVs** with balanced features for the following reasons:

Market Demand: Mid-range EVs cater to a broad segment of the market, balancing affordability with desirable performance characteristics. This can appeal to a significant portion of consumers who are neither looking for the cheapest nor the most expensive options.

Financial Viability: The moderate pricing of mid-range EVs ensures a good balance between cost and profit, making it a financially viable option for the company.

Scalability: Focusing on mid-range EVs allows the company to scale production more effectively, catering to diverse customer needs without the high production costs associated with premium models.

Market Penetration: Mid-range EVs can penetrate various market segments, including both urban and suburban areas, appealing to both individual consumers and potential fleet buyers.

Conclusion: The company should **primarily focus on producing Mid-Range EVs** that offer a balance between price, performance, and range. By focusing on mid-range EVs, the company can establish a strong market presence while also exploring niche segments for premium and entry-level models as supplementary products.

* ----- X ----- *