EV_Market_Segmentation

Ramakrushna Das

Which EV vehicle to Produce

Process:

1. **Initial Data Preparation:**

- **Handling Categorical Features**:For features with a large number of categories, the top 10 most frequent categories were selected to simplify the dataset. These categories were transformed into binary columns, where each row was assigned a value of 1 if it belonged to the category or 0 otherwise.
- This approach reduced the dimensionality while focusing on the most relevant categories for the analysis.
- For features with fewer categories, one-hot encoding was applied to create binary columns for all categories. This ensured the preservation of all unique categories while maintaining compatibility with machine learning models.
- Tools and Libraries Used:Pandas was utilized for data manipulation tasks, such as counting unique categories and creating binary columns.NumPy was employed for conditional replacement of values during the transformation of categorical variables into binary columns.

2. **KMeans Clustering:**

- **Experimentation with Different Values of K**: A range of values for the number of clusters was tested, specifically from 2 to 9, to evaluate various segmentation scenarios and identify the optimal number of clusters.
- **Determination of the Optimal Number of Clusters:**Multiple evaluation metrics were employed to determine the best k value, including:
    - **Inertia**: To measure within-cluster sum-of-squares, with an elbow point indicating diminishing returns for increasing k.
    - **Silhouette Score**: To assess the separation between clusters, with higher scores suggesting better-defined clusters.
    - **Adjusted Rand Index (ARI)**: To measure the similarity between cluster assignments and ground truth or previous clustering results, providing a robust evaluation of cluster consistency.
    - **Segment-Level Stability Across Solutions**: Stability analysis was performed to ensure that the identified clusters were consistently robust across different runs of the model and varying initial conditions.

Based on these evaluations, k = 3 emerged as the optimal number of clusters, demonstrating strong segmentation stability, high silhouette scores, and meaningful patterns in the data.

- **Cluster Assignment**:The labels_ attribute from the trained KMeans model was used to assign each data point to one of the three clusters, enabling further analysis of the characteristics of each segment.

- **Gaussian Mixture Model and Information Criterion Analysis**:To complement the KMeans approach, a Gaussian Mixture Model was applied to identify clusters based on probabilistic distributions.The best number of clusters was determined using Information Criteria:
- **Akaike Information Criterion (AIC)**: Penalized overfitting while accounting for model complexity.

Bayesian Information Criterion (BIC): Added a stricter penalty for complexity, ensuring simpler and more interpretable models.

Integrated Completed Likelihood (ICL): Provided an alternative evaluation considering clustering membership probabilities.The analysis involved:

Fitting GMM models with kkk ranging from 2 to 9 and evaluating AIC, BIC, and ICL for each configuration.Stability and performance were visualized by plotting information criteria against the number of clusters, revealing k=3k = 3k=3 as the optimal solution based on lower AIC, BIC, and ICL values.

- **Cross-Comparison Between KMeans and GMM**:A contingency table compared cluster memberships from KMeans and GMM, revealing consistent groupings and validating the clustering process.Further sub-clustering within the GMM cluster segments was explored using KMeans for additional granularity.
- **Cluster Assignments**:Clustering labels from both KMeans and GMM were adjusted and aligned for interpretability.Specific clusters were analyzed in-depth, identifying key characteristics and segment contributions for business insights.
- **Validation Using Log-Likelihood**:The log-likelihood scores of different GMM models (MD.m4MD.m4MD.m4 and MD.m4aMD.m4aMD.m4a) were calculated to confirm model robustness and ensure reliable predictions.

3. Analysis of Cluster Mean:

- **WCSS Elbow Method:**
  A line plot of WCSS was used to determine the optimal number of clusters by identifying the "elbow" point where WCSS decreases less significantly.
- **Silhouette Scores:**
  A line plot of silhouette scores for different cluster numbers assessed the quality of clusters, with higher scores indicating better-defined clusters.
- **Adjusted Rand Index (ARI):**
  A boxplot of ARI values compared clustering solutions, helping identify the best number of clusters that closely matched the true structure.
- **Cluster Distribution (Histograms):**
  Histograms of cluster similarities showed the compactness of clusters, highlighting any outliers.
- **Segment Level Stability (SLSA) Plot:**
  A line plot visualized the stability of clusters across different segment numbers, confirming the robustness of the solution.
- **Information Criteria (AIC, BIC, ICL):**
  A line plot compared AIC, BIC, and ICL values across different cluster counts to select the optimal clustering model.

- **KMeans vs. GMM Comparison:**
  A contingency table compared cluster assignments between KMeans and Gaussian Mixture Models (GMM) to assess which method better captured the data's structure.

4. **Refining the Insights:**

- **Thresholding**: To avoid visual clutter, you applied a threshold (0.01) to the data so that only meaningful contributions were considered in the visualizations.
- **Ensuring Consistency in Visuals**: Shifted all data to the positive side by adding a constant to the scaled data. This ensured all bars in the plots were on one side, making comparisons clearer.

5. Interpretation of Clusters Based on the visualization and analysis:

- Two_W_Personal emerged as a significant contributor, emphasizing the importance of two-wheelers for personal use in one cluster.
- Vehicle_Category_3_Wheelers dominated another cluster, highlighting the significance of three-wheelers in the market.
- Three_Wheeler_Goods and Three_Wheeler_Passenger were major contributors in another segment, confirming the critical role of three-wheelers for goods transport and passenger services.

6. Drawing Conclusions:

- You concluded that k = 3 was the optimal choice for clustering, providing meaningful segmentation.
- The dominance of three-wheelers and the importance of two-wheelers for personal use were clear takeaways.
- This analysis provides actionable insights for vehicle manufacturers or marketers, helping them focus on key segments such as three-wheelers for goods/passenger transport and two-wheelers for personal use.

The Graphs Visualization :

```
EV_Sales_Quantity                          127.27
MOTOR_CAR                                    0.04
GOODS_CARRIER                                0.04
M_CYCLE_SCOOTER                              0.04
MOTOR_CAB                                    0.04
BUS                                          0.04
THREE_WHEELER_PASSENGER                      0.04
AMBULANCE                                    0.04
MAXI_CAB                                     0.03
CONSTRUCTION_EQUIPMENT_VEHICLE               0.03
THREE_WHEELER_GOODS                          0.03
Karnataka                                    0.05
Maharashtra                                  0.05
Gujarat                                      0.05
Uttar_Pradesh                                0.04
Rajasthan                                    0.04
West_Bengal                                  0.04
Odisha                                       0.04
Kerala                                       0.04
Haryana                                      0.04
Tamil_Nadu                                   0.04
Others                                       0.55
Two_W_Personal                               0.12
Bus                                          0.07
Four_W_Shared                                0.05
Vehicle_Category_Two___Wheelers              0.14
Vehicle_Category_3_Wheelers                  0.13
Vehicle_Category_Four___Wheelers             0.09
Vehicle_Category_Bus                         0.09
Vehicle_Category_Others                      0.55
Year_Two_0Two_Two_                           0.47
Year_Two_0Two_3                              0.49
Year_Two_0Two_Four_                          0.04
dtype: float64
```

The average values of the transformed binary numeric segmentation variables indicate that:

a) EV_Sales_Quantity has a notable average of 127.27, representing significant activity.
b) "Others" is the most frequent category in both the state features (55%) and vehicle category features (55%).
c) Among vehicle types:

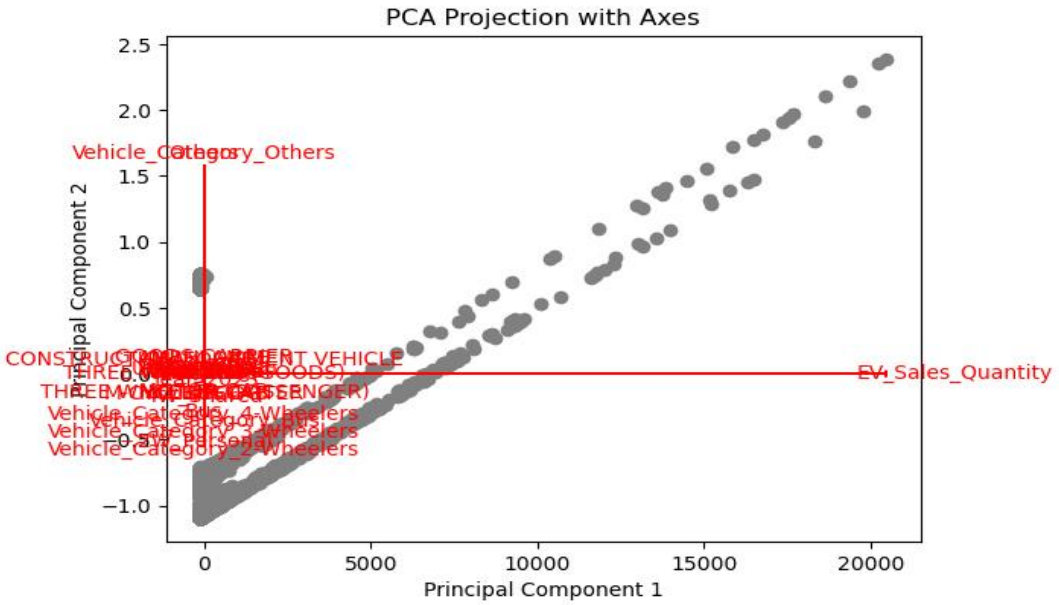
- Two_W_Personal accounts for 12%,
- Vehicle_Category_Two_Wheelers is 14%,
- Vehicle_Category_3_Wheelers is 13%,
- Vehicle_Category_Four_Wheelers and Vehicle_Category_Bus each account for 9%.

d) Among states:

- The most frequent are Karnataka, Maharashtra, and Gujarat (each 5%).
- Other states like Odisha, Kerala, and others account for 4% each.

e) Across years:

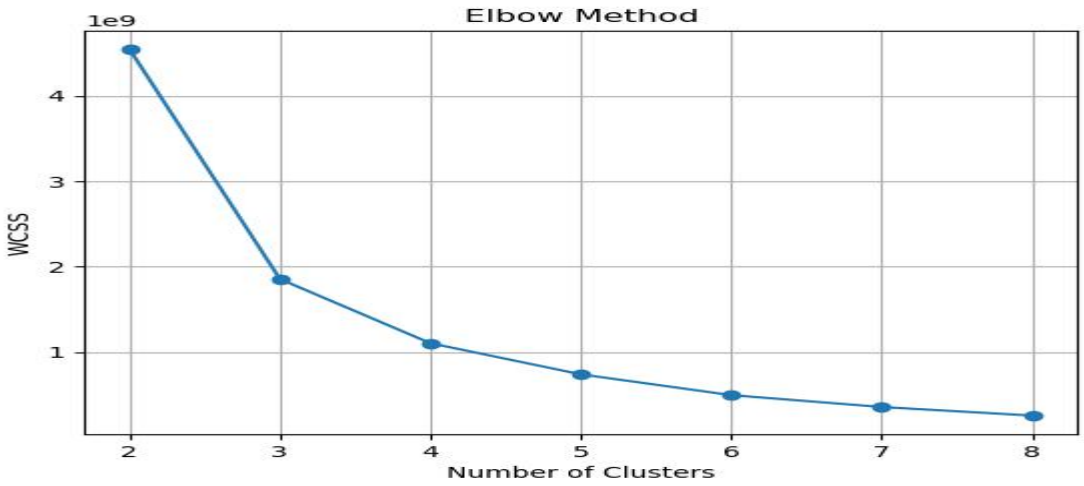
- The largest share is 2023 (49%), followed by 2022 (47%), with a small portion in 2024 (4%).



A perceptual map was created to show how people perceive EV Market based on various attributes.

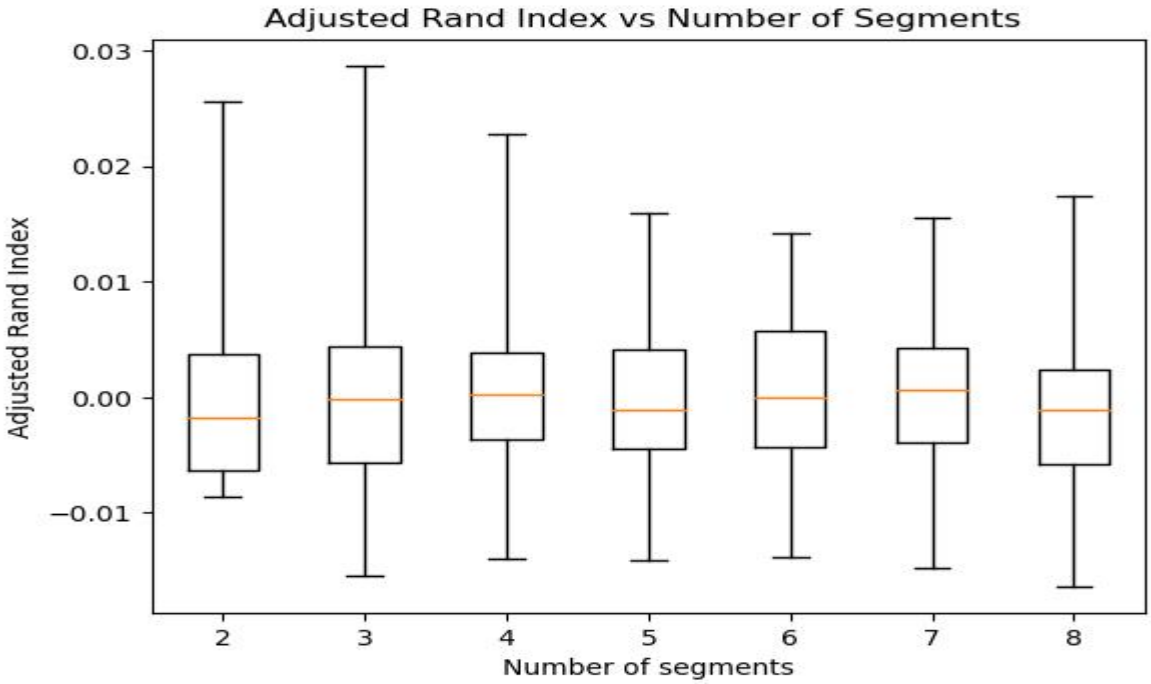
- EV_Sales_Quantity is strongly aligned with Principal Component 1, indicating that it plays a significant role in the data along this axis.
- The feature Vehicle_Category_Others shows strong alignment with Principal Component 2, highlighting its distinct impact.
- Other vehicle categories, such as Two-Wheelers and Four-Wheelers, cluster closely near the origin, suggesting less variability or influence compared to the dominant features.

- Features like Construction Equipment Vehicle and Three-Wheelers show weaker contributions as they are located closer to the center.



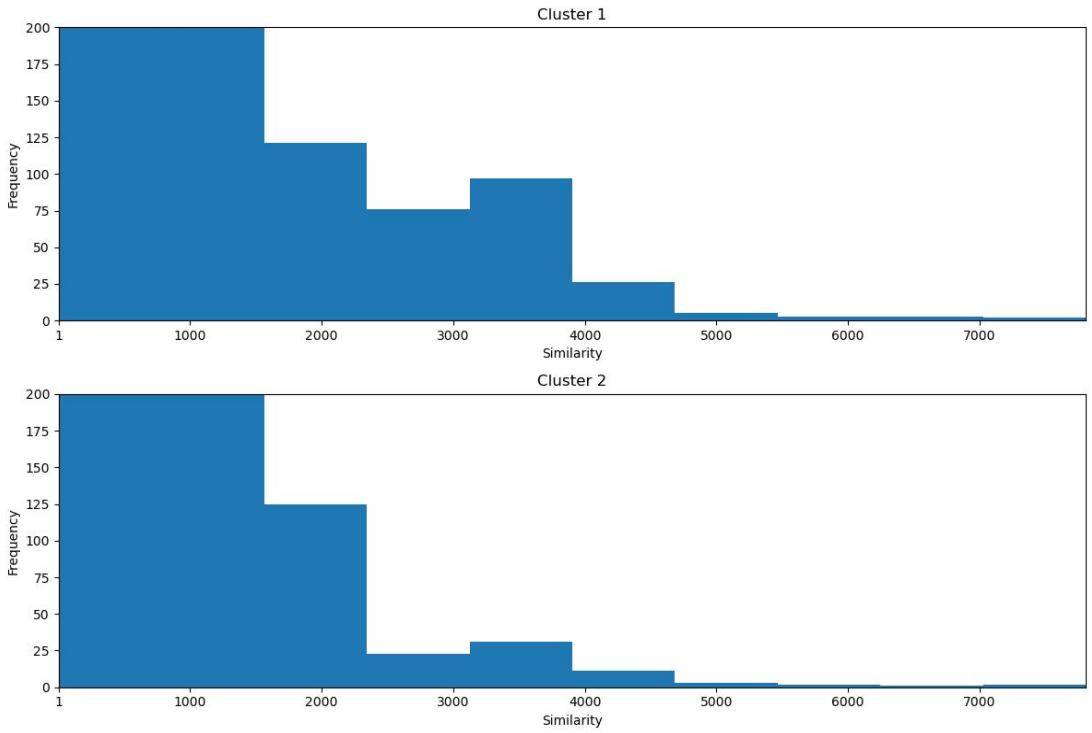
The line plot shows the Elbow Method for selecting the optimal number of clusters. The x-axis represents the number of clusters, and the y-axis represents the Within-Cluster Sum of Squares (WCSS), which measures the compactness of clusters. Key observations include:

- There is a noticeable sharp drop in WCSS from 2 to 3 clusters, indicating that adding more clusters initially reduces variability significantly.
- After 3 clusters, the reduction in WCSS slows down, and the curve starts to flatten.
- This suggests that 3 clusters might be a reasonable choice, as adding more clusters after this point provides diminishing returns in reducing variability.



The boxplots in the figure illustrate the stability (measured by Adjusted Rand Index) for different numbers of segments. The median stability for each number of segments is represented by the orange line within each box.

- Stability is highest for solutions with 2, 3, and 4 segments, as these have smaller variability and a higher central value compared to solutions with more segments.
- The stability decreases significantly as the number of segments increases beyond 3, indicating that the groupings become less consistent when the segmentation process is repeated.
- The 3-segment solution is the most reliable and stable choice, balancing meaningful market differentiation with consistent reproducibility.
- While the 4-segment solution offers some stability, the 3-segment solution is even more stable and consistently replicable across different iterations.



The histograms in the image show the distribution of similarity values for two clusters.

Cluster 1:

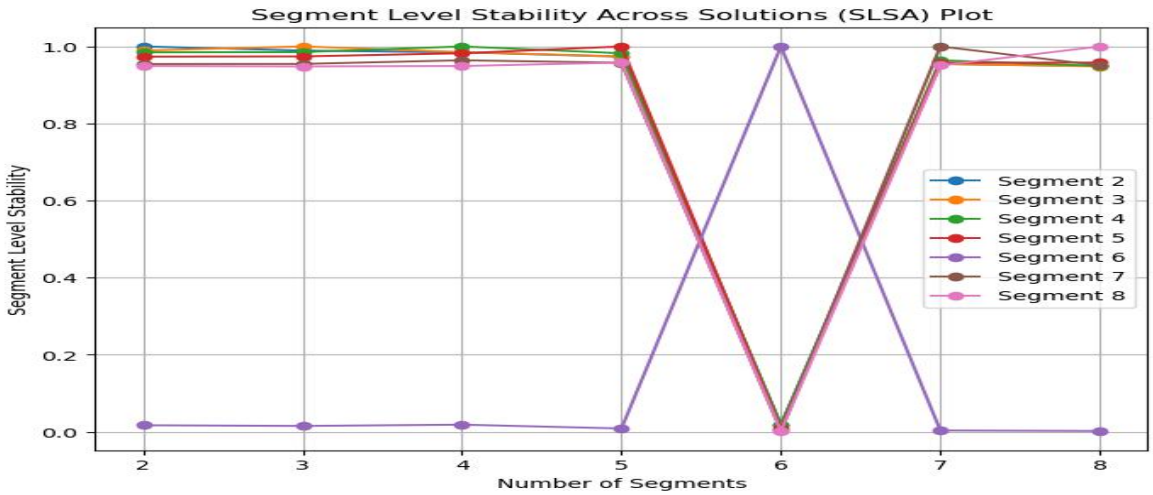
- The majority of similarity values are concentrated between 0 and 2000.
- There's a smaller peak around 3000.
- The distribution is right-skewed, meaning there are more values on the lower end.

Cluster 2:

- The distribution is similar to Cluster 1, with the majority of values between 0 and 2000.
- There's also a smaller peak around 3000.
- The distribution is also right-skewed.

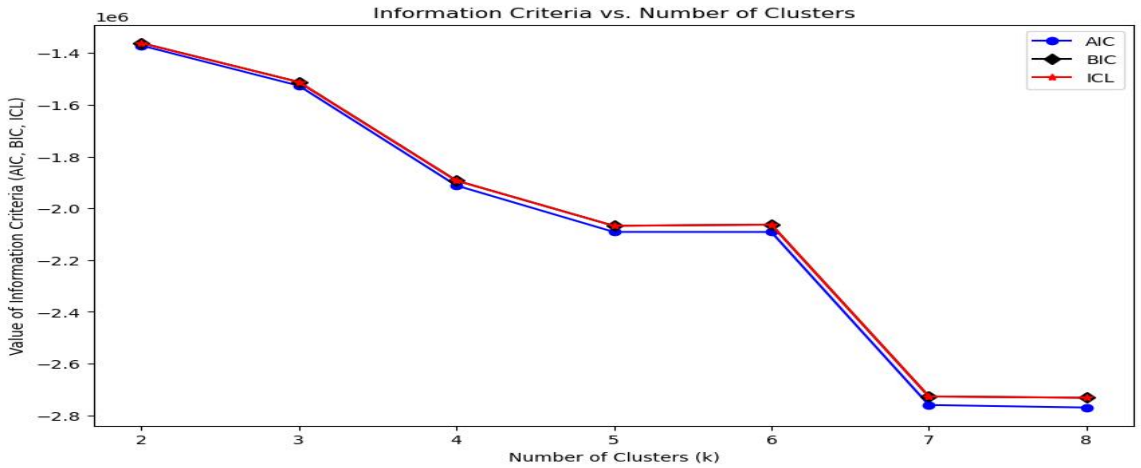
Both Clusert 1 and 2:

- Both clusters have a similar distribution of similarity values.
- The majority of values are on the lower end, indicating that the segments within each cluster are not very similar to each other.
- The smaller peaks around 3000 suggest that there might be some segments within each cluster that are more similar to each other.
- The majority of similarity values in both clusters are concentrated between 0 and 2000.



The Segment Level Stability Across Solutions (SLSA) plot shows the stability of different segments (2 to 8) as the number of segments in the solution increases.

- **Segment 2 and Segment 3:** These segments have high stability throughout the range of solutions, indicating that they remain consistent across different numbers of segments.
- **Segment 4 and Segment 5:** These segments show a decrease in stability as the number of segments increases. This suggests that they might be less stable and could be further divided into smaller segments.
- **Segment 6, Segment 7, and Segment 8:** These segments exhibit low stability overall, indicating that they are not consistent across different solutions. This could be due to the complexity of the data or the limitations of the clustering algorithm.



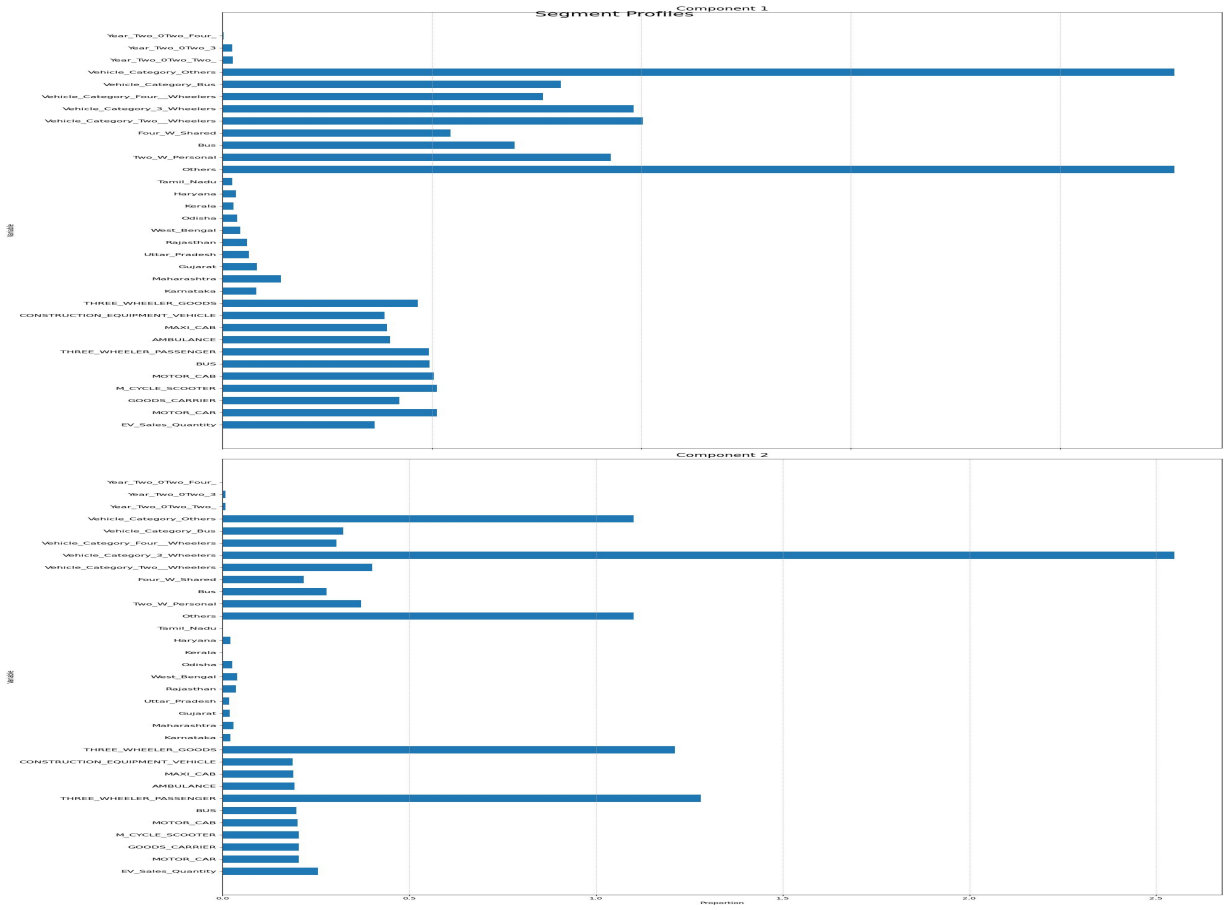
- **AIC:** The AIC decreases rapidly as the number of clusters increases, reaching a minimum around k=6. After that, it starts to increase again.
- **BIC:** The BIC also decreases initially, but it reaches a minimum at k=4 and then starts to increase more gradually than the AIC.
- **ICL:** The ICL shows a similar trend to the BIC, with a minimum around k=4 and a gradual increase afterward.

Interpretation From AIC,BIC,ICL

- **AIC:** Suggests that a model with 6 clusters might be a good fit, as it minimizes the information loss. However, it is important to note that AIC tends to favor more complex models.
- **BIC:** Suggests that a model with 4 clusters might be a good compromise between model complexity and fit. BIC penalizes model complexity more heavily than AIC.
- **ICL:** Also supports a model with 4 clusters, indicating a good balance between model fit and complexity.

Gaussian Mixture KMeans	1	2	3
1	18397	41	2396
2	0	58	0
3	1	91	176

- **Component 1 (Gaussian Mixture)**:Mostly made up of 18,397 members from Segment 1 (KMeans).A small portion (2,396 members) overlaps with Segment 3 (KMeans).
- **Component 2 (Gaussian Mixture)**:Almost entirely overlaps with Segment 2 (KMeans) (58 members).
- **Component 3 (Gaussian Mixture)**:Largely made up of 176 members from Segment 3 (KMeans).A smaller portion (91 members) overlaps with Segment 2 (KMeans).
- **Stable Segments:**Segment 2 and Segment 3 from K-Means are highly similar to Component 1 and Component 3 of the GMM. Specifically, the majority of members from K-Means Segment 1 belong to GMM Component 1, while the majority of members from K-Means Segment 3 belong to GMM Component 3.
- K-Means Segment 2 shows some overlap with GMM Component 2, but the overlap is smaller compared to other segments. Some members from K-Means Segment 2 also belong to GMM Component 3.



Component 1:

1. **Key Factors with High Influence:**

- ✓ Vehicle_Category_Others has the largest positive contribution to this cluster. This suggests that vehicles in the "Others" category dominate this segment.
- ✓ Two_W_Personal also has a significant positive contribution, indicating the prominence of two-wheelers used for personal purposes.

2. **Moderate Contributions:**

- ✓ Four_W_Shared and Bus contribute moderately, likely reflecting their secondary presence in this segment.

3. **State-Level Influence:**

- ✓ The regional factors (like specific states) have smaller or negligible contributions, indicating a lower correlation with the clustering for Component 1.
- ✓ The first component appears to be dominated by diverse vehicle categories, especially personal-use two-wheelers and other miscellaneous vehicles.

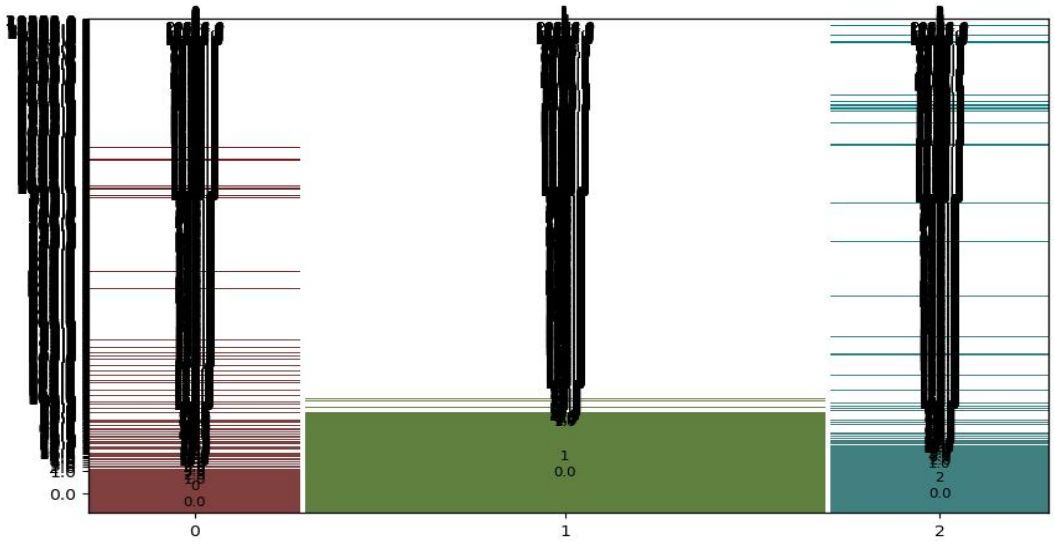
Component 2:

1. **Key Factors with High Influence:**

- ✓ Vehicle_Category_3_Wheelers has the most significant positive contribution, emphasizing the dominance of three-wheelers in this cluster.
- ✓ Three_Wheeler_Goods and Three_Wheeler_Passenger also strongly contribute to this segment, confirming the importance of this vehicle category.

2. **State-Level Influence:**

- ✓ Others (possibly non-state-specific factors) has moderate positive contributions. States and regions seem to have minimal influence on this component.
- ✓ The second component strongly focuses on three-wheelers, both for goods and passenger transport.



The mosaic plot illustrates a clear relationship between segment membership and the level of EV sales.

Observations:

- **Segment 1:** Individuals in this segment have significantly lower EV sales compared to other segments.
- **Segment 2:** This segment exhibits a more balanced distribution of EV sales.
- **Segment 3:** This segment shows the highest level of EV sales.

Overall, the plot reveals that different segments have distinct levels of EV sales, with Segment 3 demonstrating the highest level of adoption.

Solution For the company :

- You concluded that k = 3 was the optimal choice for clustering, providing meaningful segmentation.
- The dominance of three-wheelers and the importance of two-wheelers for personal use were clear takeaways

Conclusion :

After experimenting with values of *k* ranging from 2 to 9 using 100 samples, the optimal number of clusters was determined to be **k = 3**. This segmentation provided meaningful insights into the dataset, highlighting the key contributors within each cluster.Two_W_Personal exhibits a significant positive contribution, indicating the prominence of two-wheelers used for personal purposes in one of the clusters.Vehicle_Category_3_Wheelers stands out with the most significant positive contribution, emphasizing the dominance of three-wheelers in another cluster.Three_Wheeler_Goods and Three_Wheeler_Passenger also strongly contribute, reinforcing the importance of three-wheelers specifically for goods transport and passenger services in this segment.

Click here for the code link