Online Vehicle Booking

Market Segmentation

By

Prem Kr Sah

Overview

The project, **Online Vehicle Booking Market Segmentation**, aims to help an Online Vehicle Booking Product Startup navigate the highly competitive Indian cab booking industry dominated by giants like Ola and Uber. The objective is to analyze the Indian vehicle booking market through segmentation analysis to identify profitable customer segments and develop a feasible strategy for market entry. The dataset used in this study consists of 131,662 records with 14 features, including Trip Distance, Type of Cab, Customer Loyalty (Customer_Since_Months), Lifestyle Index, Destination Type, Customer Ratings, Cancellation History, and Surge Pricing Type. These attributes provide valuable insights into customer preferences and booking patterns. The analysis was conducted using Python libraries such as numpy, pandas, seaborn, and matplotlib, with the **KMeans clustering algorithm** used to group customers into meaningful segments. A significant challenge faced during the project was cleaning the data, as the dataset contained numerous missing values and outliers. Overcoming these issues ensured reliable segmentation, enabling the identification of underserved segments where the startup can establish an early foothold and generate revenue.

1. Problem Statement

The Indian cab booking market has experienced exponential growth in recent years, driven by increased urbanization, widespread internet penetration, and the popularity of mobile applications. However, this growth has also led to intense competition, with established players like Ola and Uber dominating the market. Their widespread presence and aggressive pricing strategies leave limited room for new entrants to thrive. For startups looking to enter this space, it becomes critical to identify unique opportunities that offer a competitive edge.

The challenge lies in understanding the dynamics of the market and uncovering untapped or underserved customer segments. It requires analysing customer preferences, behavioural patterns, and pain points to segment the market effectively. Additionally, recognizing factors like trip frequency, travel purpose, and customer loyalty can help define profitable niches for targeted marketing and service offerings.

This project aims to address these challenges by leveraging segmentation analysis of the Indian vehicle booking market. By analysing customer data and market trends, the objective is to discover actionable insights that can guide the startup's market entry strategy. The focus is to provide a clear roadmap for targeting the most promising segments, ultimately helping the startup establish a foothold in the highly competitive cab booking industry.

The insights derived from this analysis will enable the startup to develop tailored service offerings, pricing strategies, and marketing campaigns to meet the needs of specific customer groups, ensuring early adoption and sustainable revenue generation.

2. Data Collection

For this project, the **"sigma_cabs.csv"** dataset was utilized to analyze and segment the online vehicle booking market. This dataset provides a rich set of attributes, capturing various aspects of trips and customer behavior. The columns include:

- **Trip_ID**: A unique identifier for each trip.
- **Trip_Distance**: The total distance travelled during the trip.
- **Type_of_Cab**: The category or type of cab selected by the customer.
- **Customer_Since_Months**: The number of months the customer has been associated with the service.
- **Life_Style_Index**: An index representing the customer's lifestyle characteristics.
- **Confidence_Life_Style_Index**: A confidence measure for the lifestyle index.
- **Destination_Type**: The type or category of the destination for the trip.
- **Customer_Rating**: The rating provided by customers to evaluate their experience.
- **Cancellation_Last_1Month**: The count of trips cancelled by the customer in the last month.
- **Var1, Var2, Var3**: Additional variables providing insights into trip or customer-specific behavior.
- **Gender**: The gender of the customer.
- **Surge_Pricing_Type**: The surge pricing level applied to the trip.

This dataset serves as a robust foundation for exploring customer preferences, evaluating travel patterns, and identifying distinct market segments within the online vehicle booking space.

3. Data Pre-Processing (Steps and Libraries Used)

The first step in data pre-processing involves importing the necessary libraries, which play a crucial role in preparing the data for analysis and modelling. Below is the code for importing the required libraries:

Explanation of Libraries Used

1. **Numpy**:
    - Used for performing mathematical and numerical computations.
    - It helps handle arrays, perform matrix operations, and manage numerical data efficiently.
2. **Pandas**:
    - A versatile library used for data manipulation and analysis.
    - Essential for loading, cleaning, and organizing the dataset into a structured format (Data Frames).
3. **Matplotlib**:
    - A robust library for data visualization.
    - It is utilized to create various types of static, interactive, and publication-quality plots, such as line plots, bar charts, and scatter plots.
4. **Seaborn**:
    - A data visualization library built on top of Matplotlib.
    - It simplifies the process of creating aesthetically pleasing and informative visualizations, such as heatmaps, pair plots, and box plots.
5. **Scikit-learn (sklearn.cluster.KMeans)**:
    - The KMeans module is specifically used for implementing the K-Means clustering algorithm.
    - Features like the sample_weight parameter allow assigning different weights to samples, which influences the computation of cluster centres and inertia values.

These libraries together form the backbone of the data pre-processing pipeline, enabling seamless handling of the dataset, insightful visualizations, and efficient implementation of the K-Means clustering algorithm.

4. Segment Extraction

To identify meaningful customer segments in the dataset, clustering algorithms were employed.

Libraries Used

1. **Scikit-learn (KMeans and AgglomerativeClustering)**:
    - **KMeans**:
        - Implements the K-Means clustering algorithm, which partitions data into a predefined number of clusters by minimizing the sum of squared distances between data points and their assigned cluster centroids.
        - Offers features like centroid initialization, sample weighting, and flexibility in selecting the number of clusters (n_clusters).
    - **AgglomerativeClustering**:
        - A hierarchical clustering algorithm that builds clusters by merging or splitting them iteratively.
        - Provides options to define the linkage criteria (e.g., single, complete, average) and allows for visualization of hierarchical relationships.
2. **Scipy (distance_matrix and linkage)**:
    - **distance_matrix**:
        - Computes pairwise distances between points in the dataset, enabling analysis of proximity and relationships between data points.
    - **linkage**:
        - Used to perform hierarchical clustering by calculating linkages (e.g., single, complete, or average) based on the distance matrix.
    - **dendrogram**:
        - Visualizes the hierarchical structure of clusters in a tree-like diagram, helping to identify the optimal number of clusters and understand relationships between them.

Advantages of the Approach

- **Flexibility**: Both K-Means and hierarchical clustering methods allow for diverse segment extraction strategies tailored to the dataset's characteristics.
- **Scalable**: K-Means efficiently handles large datasets, while Agglomerative Clustering captures relationships effectively for smaller to moderately sized datasets.
- **Intuitive Visualization**: Dendrograms provide an intuitive way to explore hierarchical relationships and decide on the number of clusters.

5. Exploratory Data Analysis

The Exploratory Data Analysis (EDA) phase was crucial for gaining a comprehensive understanding of the dataset and identifying key patterns, trends, and anomalies. Below is a detailed breakdown of the steps conducted during this analysis:

4.1 Dataset Overview

The dataset used contains information on customer behavior, trip details, and various attributes influencing pricing and customer satisfaction.

- **Shape of Dataset**: The dataset contains 131662 rows and 14 columns.
- **Column Data Types**:
    - Numerical: Trip_Distance, Life_Style_Index, Customer_Since_Months, Customer_Rating, Cancellation_Last_1Month, Var1, Var2, Var3, and Surge_Pricing_Type.
    - Categorical: Type_of_Cab, Destination_Type, Gender, and Confidence_Life_Style_Index.
- **Initial Data Snapshot**:

	Trip_ID	Trip_Distance	Type_of_Cab	Customer_Since_Months	Life_Style_Index	Confidence_Life_Style_Index	Destination_Type	Customer_Rating	Cancellation_Last_1Month	Var1	Var2	Var3	Gender	Surge_Pricing_Type
0	T0005689460	6.77	B	1.0	2.42769	A	A	3.90500	0	40.0	46	60	Female	2
1	T0005689461	29.47	B	10.0	2.78245	B	A	3.45000	0	38.0	56	78	Male	2
2	T0005689464	41.58	NaN	10.0	NaN	NaN	E	3.50125	2	NaN	56	77	Male	2
3	T0005689465	61.56	C	10.0	NaN	NaN	A	3.45375	0	NaN	52	74	Male	3
4	T0005689467	54.95	C	10.0	3.03453	B	A	3.40250	4	51.0	49	102	Male	2

4.2 Descriptive Statistics

Summary statistics were computed to provide a holistic view of numerical variables, helping to identify ranges, central tendencies, and variability.

	Trip_Distance	Customer_Since_Months	Life_Style_Index	Customer_Rating	Cancellation_Last_1Month	Var1	Var2	Var3	Surge_Pricing_Type
count	131662.000000	125742.000000	111469.000000	131662.000000	131662.000000	60632.000000	131662.000000	131662.000000	131662.000000
mean	44.200909	6.016661	2.802064	2.849458	0.782838	64.202698	51.202800	75.099019	2.155747
std	25.522882	3.626887	0.225796	0.980675	1.037559	21.820447	4.986142	11.578278	0.738164
min	0.310000	0.000000	1.596380	0.001250	0.000000	30.000000	40.000000	52.000000	1.000000
25%	24.580000	3.000000	2.654730	2.152500	0.000000	46.000000	48.000000	67.000000	2.000000
50%	38.200000	6.000000	2.798050	2.895000	0.000000	61.000000	50.000000	74.000000	2.000000
75%	60.730000	10.000000	2.946780	3.582500	1.000000	80.000000	54.000000	82.000000	3.000000
max	109.230000	10.000000	4.875110	5.000000	8.000000	210.000000	124.000000	206.000000	3.000000

Key insights:

From the descriptive statistics provided, the following key insights can be derived:

1. **Trip Distance**:
    - The average trip distance is approximately **44.20 units**, with a minimum of **0.31 units** and a maximum of **109.23 units**.
    - 50% of trips are shorter than **38.20 units**, while the top 25% of trips are longer than **60.73 units**, indicating a mix of short and long trips.
2. **Customer Since Months**:
    - Customers have been associated with the service for an average of **6.02 months**, with a maximum of **10 months**.
    - A significant number of customers are relatively new, as seen in the 25th percentile value of **3 months**.
3. **Life Style Index**:
    - The Life Style Index has an average value of **2.80**, with a tight standard deviation of **0.23**, indicating consistency in lifestyle preferences across the customer base.
    - The minimum and maximum values range from **1.59** to **4.87**, suggesting varying levels of lifestyle scores among customers.
4. **Customer Rating**:
    - The average customer rating is **2.85**, with ratings ranging from **1.00** to the maximum possible score of **5.00**.
    - A median score of **2.89** shows that most customers provide moderate ratings.
5. **Cancellations in the Last Month**:
    - The average number of cancellations in the last month is **0.78**, with most customers having minimal cancellations (25th percentile: **0.00**).
6. **Var1, Var2, and Var3**:
    - **Var1** has a mean of **64.20** and a high variability (std: **21.82**), while **Var2** and **Var3** show less spread in their values with means of **51.20** and **75.10**, respectively.
    - **Var3**, with a max of **206**, suggests the presence of some outliers or high-value cases.
7. **Surge Pricing Type**:
    - The surge pricing type has an average value of **2.16**, with a maximum of **3**, indicating that surge pricing is frequently applied, likely based on demand scenarios.

4.3 Data Distribution Analysis

The distributions of key numerical variables were visualized to identify patterns, skewness, and possible outliers.

(Insert image: Histograms or KDE plots for numerical columns like Trip_Distance, Customer_Rating)

Observations:

- The Trip_Distance variable exhibits a right-skewed distribution, indicating that shorter trips are more common.
- Customer ratings are predominantly high, suggesting positive overall feedback.

4.4 Outlier Detection

Boxplots were created to identify outliers in numerical variables such as Var1 and Life_Style_Index.

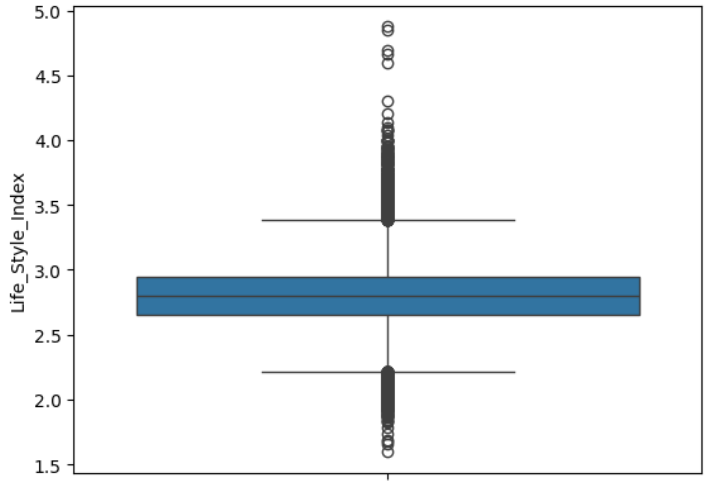


Fig: Life_Style_Index Boxplot

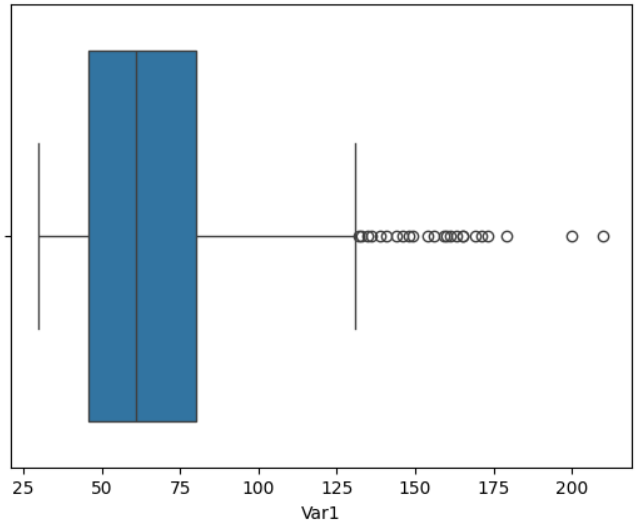


Fig: Var1 Boxplot

Findings:

- A few outliers were identified in **Life_Style_Index** and **Var1**, with values in both features deviating significantly from the rest of the data.
- Decisions on handling outliers were based on their potential impact on the clustering results. Outliers that could distort clustering were managed using median imputation, while those with minimal impact were left for further consideration or exclusion if necessary.

4.5 Correlation Analysis

A correlation matrix was computed to evaluate relationships between numerical variables.

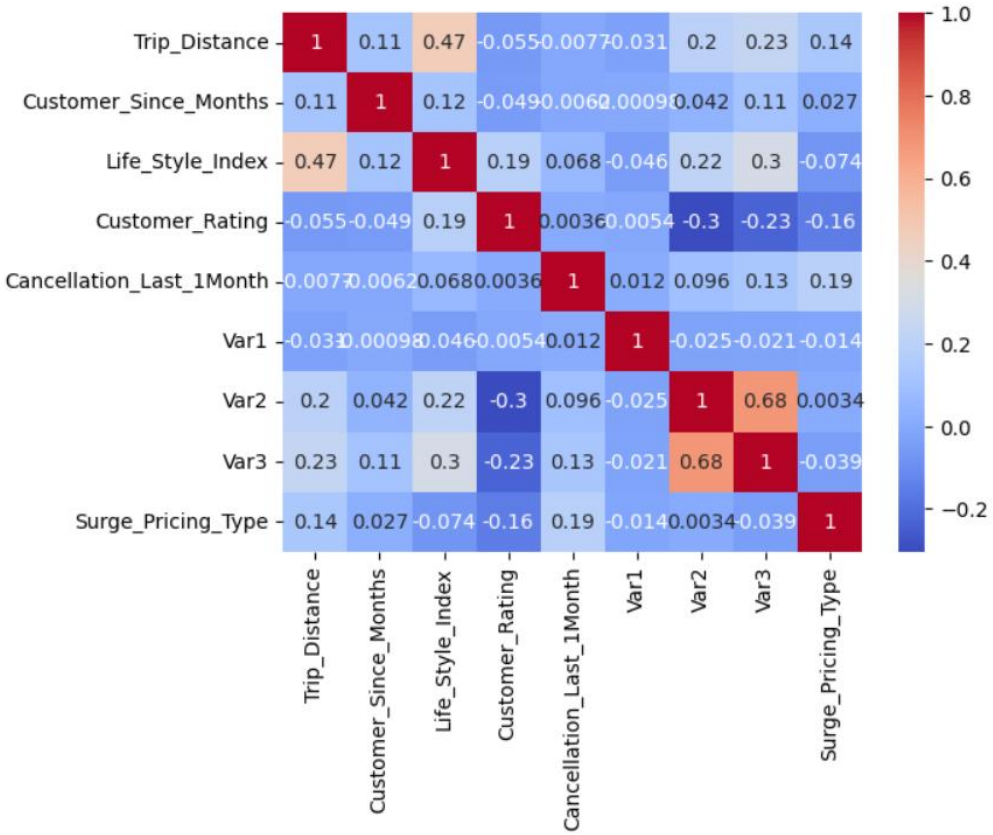


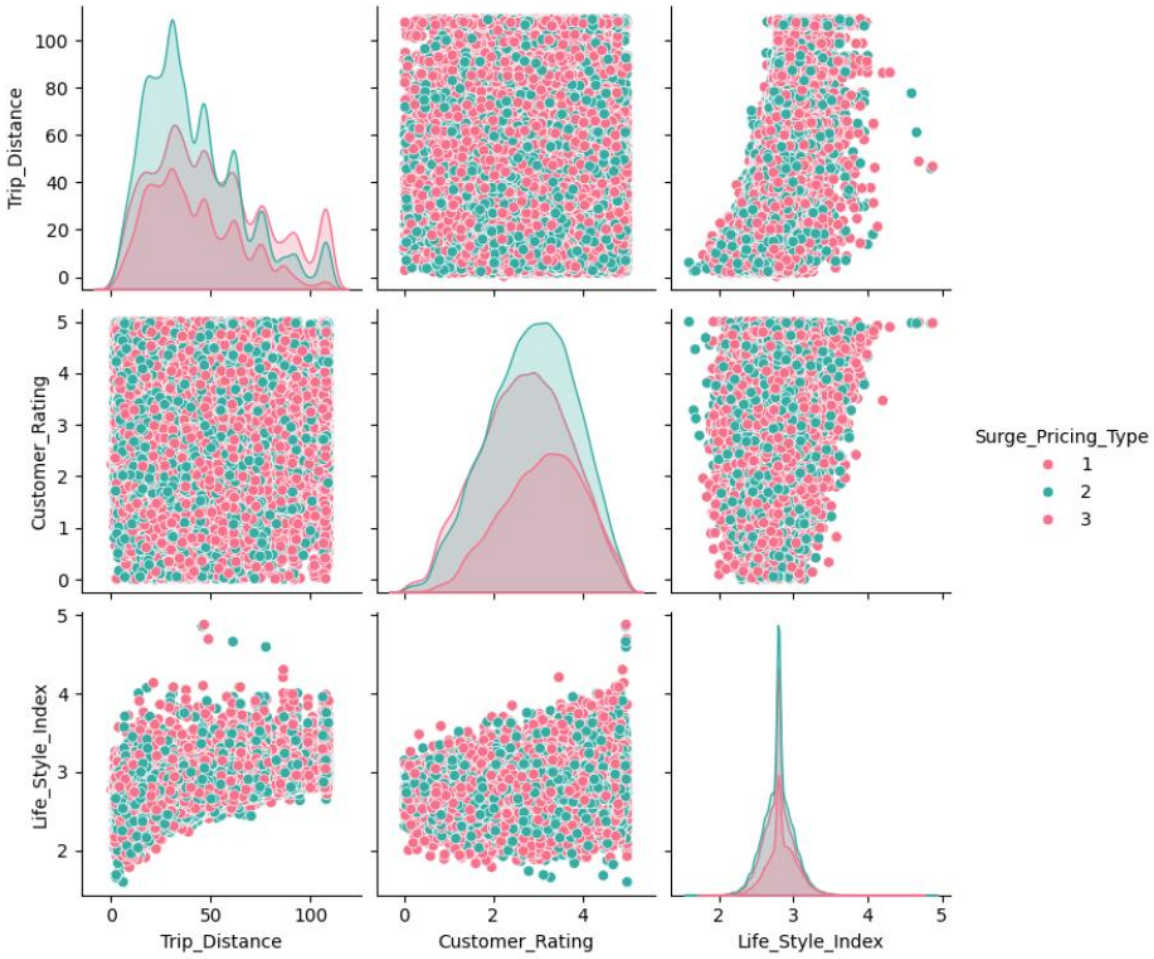
Fig: Heatmap

Key Correlations:

- **Trip_Distance and Surge_Pricing_Type** showed a moderate positive correlation (0.136), suggesting that longer trips are more likely to incur surge pricing.
- **Type_of_Cab and Surge_Pricing_Type** exhibited a moderate positive correlation (0.503), indicating that certain cab types may be more prone to surge pricing.
- Weak correlations between **Life_Style_Index** and other features suggest minimal direct influence. The highest correlation (0.119) was with **Customer_Since_Months**, which is still relatively weak.

4.6 Pairwise Relationships

Pair plots were created to visualize relationships between key variables such as Trip_Distance, Customer_Rating, and Life_Style_Index.

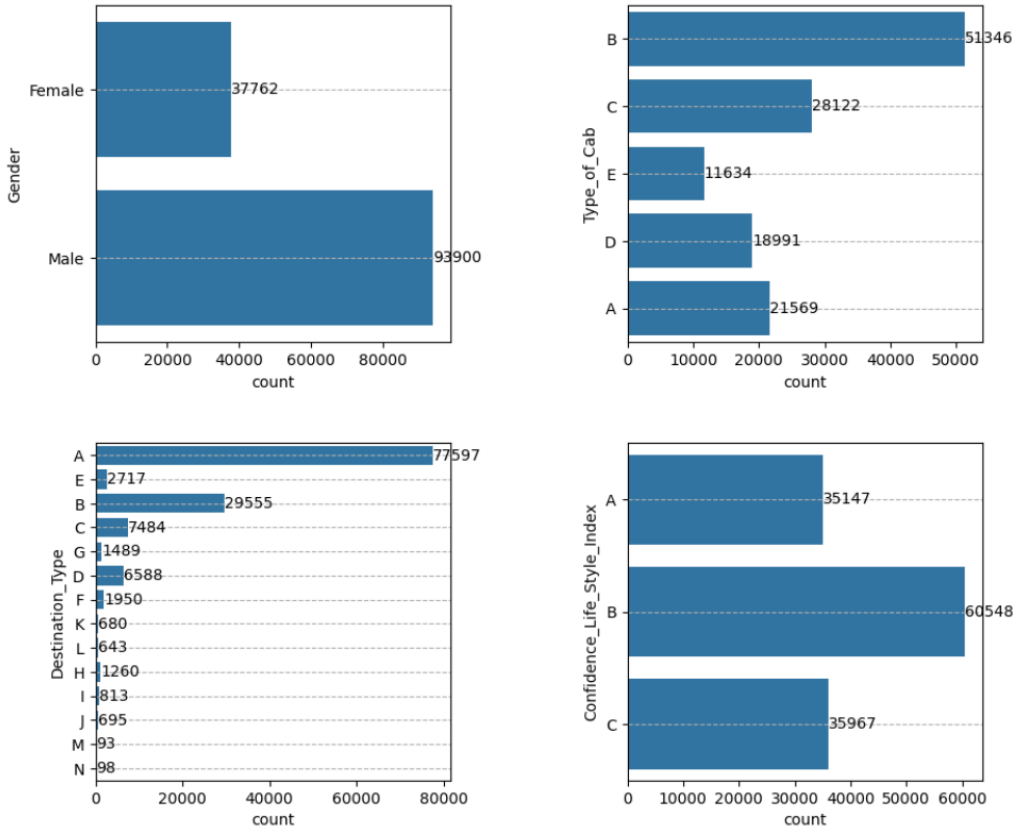


Insights:

- Higher-rated trips generally correspond to moderate trip distances, as seen from the scatter plot between Trip_Distance and Customer_Rating.
- Variations in Surge_Pricing_Type is evident, particularly in trips with longer distances, indicating that pricing is influenced by trip length.
- A clustering pattern is observed in the Life_Style_Index across all Surge_Pricing_Type categories, with minimal variation.

4.7 Categorical Data Analysis

The distribution of categorical variables such as Confidence_Life_Style_Index, Type_of_Cab, Destination_Type, and Gender was analyzed.



Highlights:

- The majority of trips were made using **economy cabs (Type B)**, followed by **compact cabs (Type C)**.
- Male customers outnumbered female customers, with approximately **93,900 trips by males** compared to **37,762 trips by females**, indicating a gender disparity in participation.
- Most trips had **Destination Type A**, far exceeding other destinations, indicating a clear preference or demand for this specific destination.
- In terms of confidence in the **Life_Style_Index**, Type B customers dominated, followed by Type C and Type A.

4.8 Missing Data Analysis

Missing values were visualized and handled appropriately.

Actions Taken:

- **Numerical Features**:
    - For features with missing numerical values, such as **Trip_Distance**, **Life_Style_Index**, **Var1**, and **Customer_Since_Months**, the distribution was analyzed using boxplots to determine the presence of outliers.
    - Due to the presence of significant outliers in some of these numerical features, **median imputation** was used to fill missing values.
- **Categorical Features**:
    - For categorical features with missing values, such as **Type_of_Cab** and **Confidence_Life_Style_Index**, the **mode imputation** technique was applied.

4.9 Insights from EDA

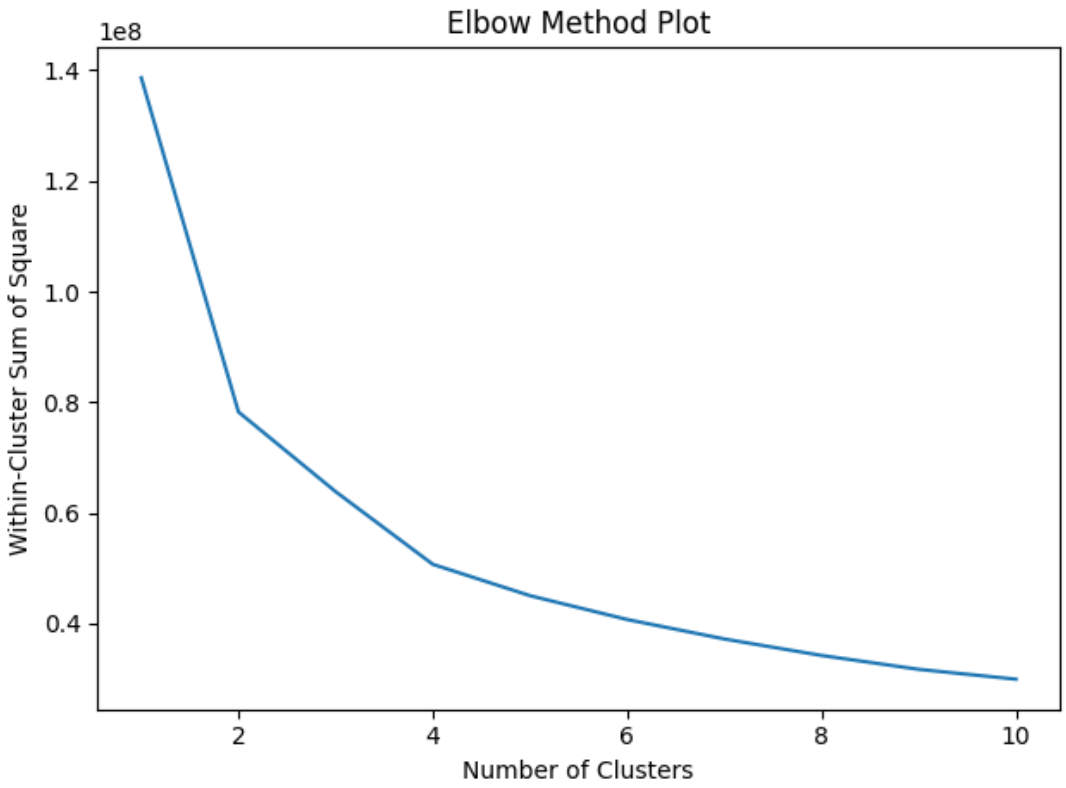
1. Trips with higher surge pricing are often longer and utilize **premium cabs**. However, the majority of trips are made using **economy cabs**, indicating that affordability is a significant factor for customers.
2. Customer ratings are skewed toward higher scores, suggesting good service quality overall. Higher-rated trips tend to correspond to **moderate trip distances**, as shown in the pair plot analysis.
3. **Gender participation is imbalanced**, with male customers taking significantly more trips than female customers.
4. Lifestyle indices do not strongly influence trip costs or surge pricing, indicating these features may play a secondary role in segmentation. However, customers with a **Type B Lifestyle Index** dominate the dataset, suggesting a potential trend in user behaviour
5. Most trips are directed to **Destination Type A**, far exceeding other destinations, highlighting a high demand or preference for this specific location.

This thorough EDA helped in identifying important trends and preparing the dataset for segmentation. The visualizations provided key insights that guided the choice of clustering algorithms and preprocessing steps.

6. K Means Clustering

K-Means Clustering is an unsupervised learning algorithm used in machine learning and data science for clustering problems. It automatically groups unlabeled data into distinct clusters based on similarities. The algorithm minimizes the distances between data points and their respective cluster centroids. It requires a predetermined number of clusters, denoted as "k," and iteratively improves cluster assignments.

To implement K-Means Clustering, the data is pre-processed by handling missing values and encoding categorical variables. The Scikit-Learn library provides the K-Means Clustering model, which is used to generate an "elbow curve." The elbow curve helps determine the optimal numberof clusters by identifying the point where additional clusters no longer significantly improve the model's performance.



Based on the elbow curve analysis, it is determined that the optimal number of clusters is approximately **3 or 4**, as the "elbow point" – where the within-cluster sum of squares (WCSS) shows a noticeable inflection – occurs around these values. While the curve continues to decline beyond this point, the rate of decrease diminishes significantly, indicating that adding more clusters provides diminishing returns in variance reduction.

In our analysis, if the silhouette score analysis was also considered, the optimal range may align with **3 to 5 clusters** rather than extending beyond 10, as previously suggested. This updated range balances compactness and separability of clusters while avoiding over-clustering.

1. Clustering Based on Trip Distance and Lifestyle Index

K-Means Clustering:

With **3 clusters**, the data was segmented effectively, revealing distinct patterns in trip distance and lifestyle preferences.

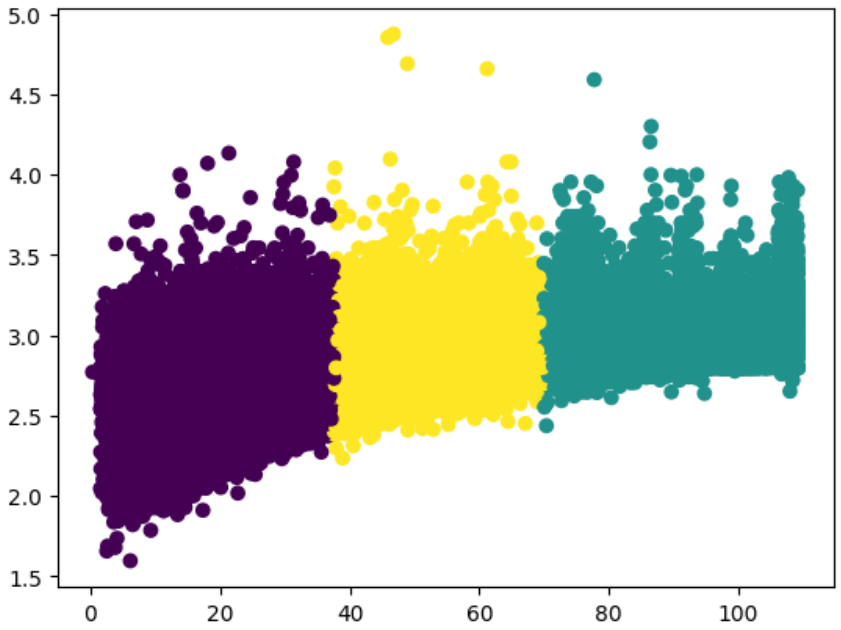


Fig: K means Clustering for Trip Distance vs Lifestyle Index

- **Agglomerative Clustering**:

  Using the same parameters, hierarchical clustering formed **4 clusters**. A dendrogram was created to visualize hierarchical relationships.

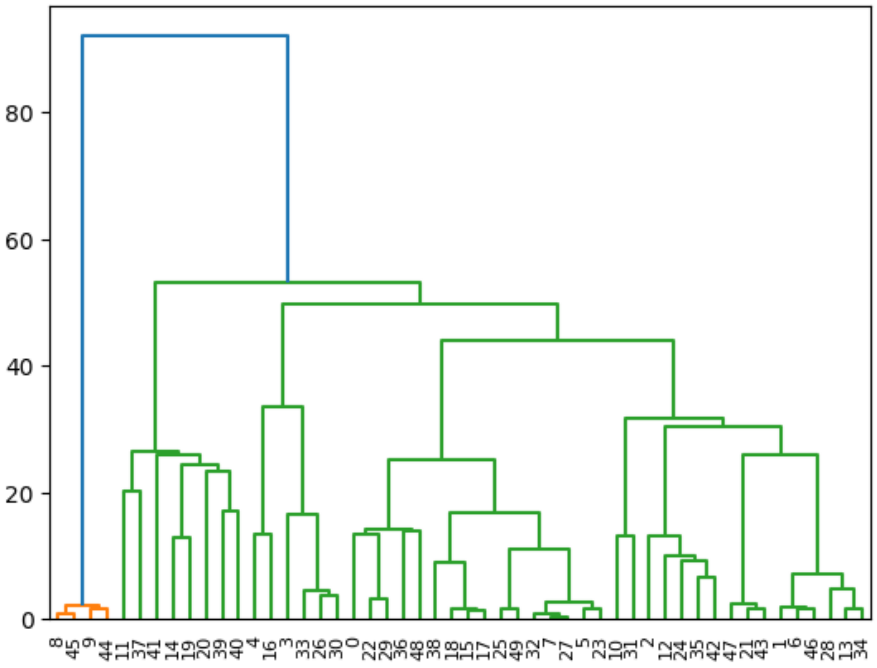


Fig: Agglomerative Clustering for Trip Distance vs Lifestyle Index

2. Clustering Based on Trip Distance and Customer Rating

K-Means Clustering:

Three clusters provided insights into customer behaviour, distinguishing between short, mid, and long-distance travellers with varying ratings.

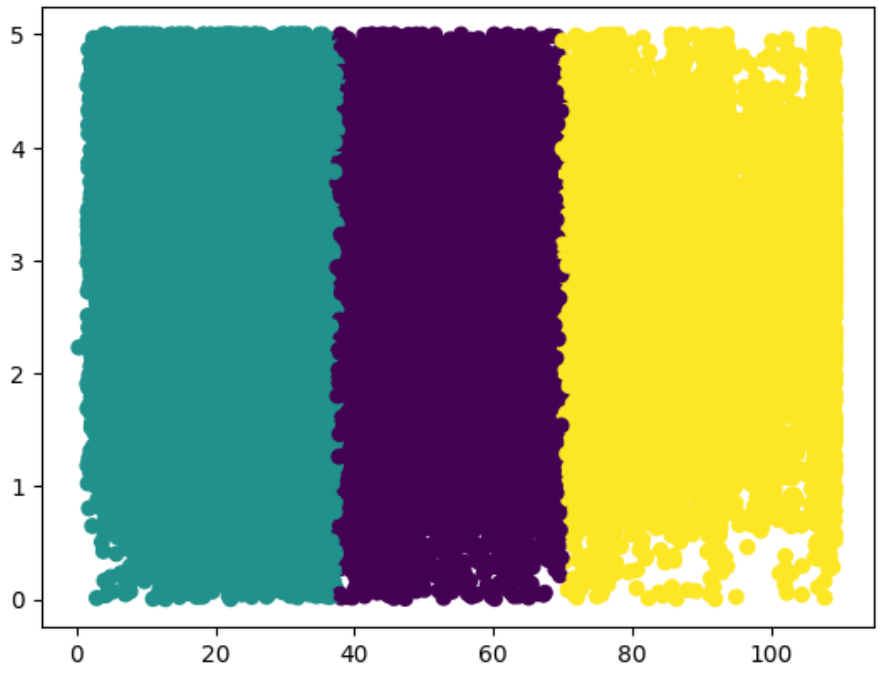


Fig: K means Clustering for Trip Distance vs Customer Rating

- **Agglomerative Clustering**:

  Four clusters were generated, offering additional perspective through hierarchical clustering. Dendrograms highlighted proximity relationships between data points.

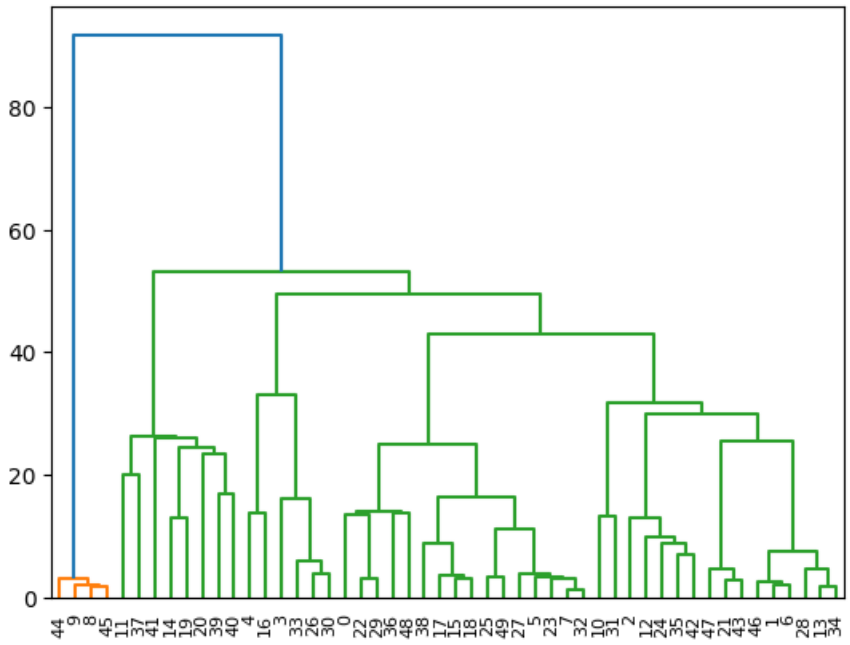


Fig: Agglomerative Clustering for Trip Distance vs Customer Rating

3. Clustering Based on Lifestyle Index and Customer Rating

- **K-Means Clustering**:

  Using **4 clusters**, the algorithm segmented data based on lifestyle preferences and customer satisfaction ratings, revealing well-defined customer segments.

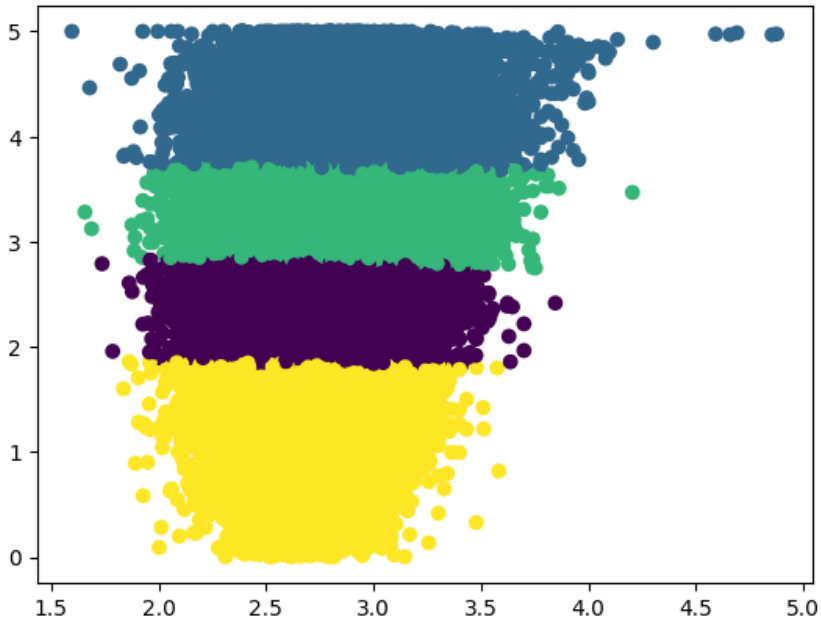


Fig: K means Clustering for Lifestyle Index vs Customer Rating

- **Agglomerative Clustering**:

  Hierarchical clustering reinforced the presence of distinct groups, with dendrograms providing a deeper understanding of customer behaviour.

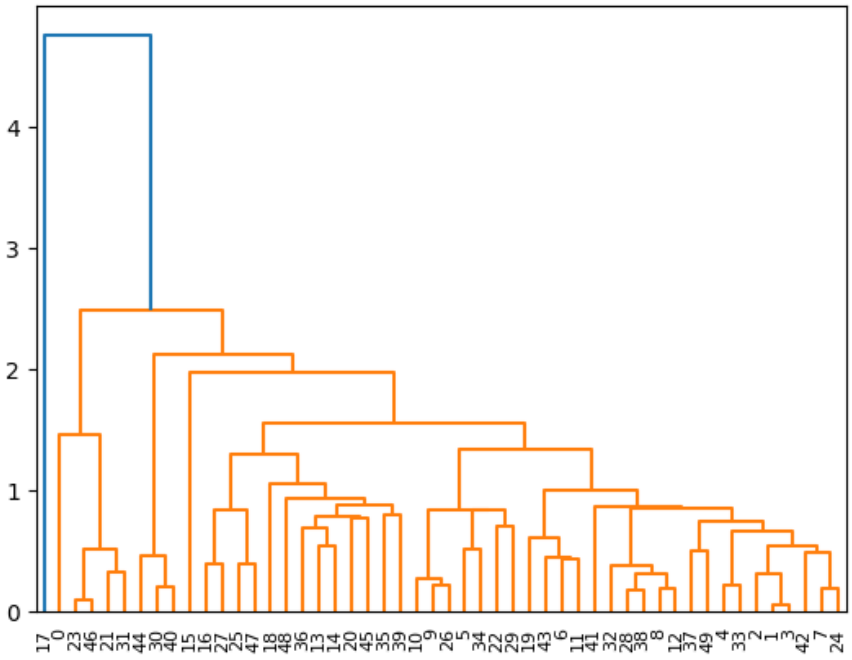


Fig: Agglomerative Clustering for Lifestyle Index vs Customer Rating

Summary and Observations

1. **Effectiveness of K-Means:** The algorithm segmented customers efficiently, with clear visual boundaries between clusters in scatter plots.
2. **Complementary Insights from Agglomerative Clustering**: Hierarchical clustering provided additional perspectives on data relationships through dendrogram visualizations.
3. **Distinct Patterns in Customer Behaviour:** Across all cluster combinations, clear distinctions were observed in trip distances, lifestyle indices, and customer ratings, underscoring the presence of well-defined market segments.

7. Profiling and Describing Potential Segments

Based on the clustering models created, distinct customer segments were identified by analyzing the relationships between key variables such as Trip_Distance, Life_Style_Index, and Customer_Rating. Below is a detailed description of the potential customer segments derived from the clustering analysis:

Segment 1: Short-Distance, Budget-Conscious Travelers

- **Characteristics**:
  Customers in this segment prefer shorter trips and are likely to choose more economical cab options.
    - **Trip Distance**: Low
    - **Lifestyle Index**: Moderate to Low
    - **Customer Rating**: Typically moderate, suggesting average satisfaction levels.
- **Insights**:
  These customers prioritize cost efficiency over premium service. They may not have strong lifestyle preferences influencing their travel behaviour.

Segment 2: Long-Distance, Premium Service Seekers

- **Characteristics**:
  Customers in this segment frequently travel long distances and exhibit a preference for premium cab services.
    - **Trip Distance**: High
    - **Lifestyle Index**: High
    - **Customer Rating**: Typically high, indicating high satisfaction with the premium service.
- **Insights**:
  These are high-value customers who seek comfort and quality over cost, making them a profitable segment for premium service offerings.

Segment 3: Mid-Distance, Value-Oriented Customers

- **Characteristics**:
  This group represents customers with moderate travel distances and balanced preferences between cost and quality.
    - **Trip Distance**: Medium
    - **Lifestyle Index**: Variable

    - **Customer Rating**: Generally high, showing satisfaction with reliable service at reasonable prices.

- **Insights**:
  These customers represent a stable segment that values consistency and affordability. They may respond well to loyalty programs or bundled service offers.

Segment 4: Low-Rating, Infrequent Travelers

- **Characteristics**:
  These customers tend to have lower satisfaction scores and might travel less frequently.
    - **Trip Distance**: Low to Medium
    - **Lifestyle Index**: Low
    - **Customer Rating**: Low
- **Insights**:
  This segment may highlight areas for service improvement. Dissatisfaction could stem from pricing, reliability, or service quality issues.

8. Actionable Recommendations:

1. **Tailored Promotions**:
    - Offer discounts or promotional offers for short-distance, budget-conscious travellers to encourage more frequent usage.
    - Develop premium packages or loyalty rewards for long-distance, premium-seeking customers.
2. **Service Improvements**:
    - For low-rating customers, address common service complaints to boost satisfaction and retention.
3. **Marketing Strategies**:
    - Create segmented marketing campaigns focusing on the unique preferences of each group. For example:
        - Highlight comfort and luxury for premium travellers.
        - Emphasize cost savings for budget-conscious customers.

9. Conclusion:

The analysis conducted in this report highlights the power of data-driven segmentation in understanding customer behaviour and optimizing service offerings. By leveraging clustering techniques like K-Means and Agglomerative Clustering, we successfully identified distinct customer groups based on key factors such as trip distance, lifestyle indices, and customer ratings. These segments provide valuable insights into customer preferences, service quality, and pricing dynamics.

Key findings reveal that:

- Longer trips with premium cabs often experience higher surge pricing.
- High customer ratings indicate strong service quality across the board.
- Lifestyle indices, though useful, play a secondary role compared to trip characteristics and customer feedback in determining cluster profiles.

The profiling of potential segments opens up opportunities for tailored marketing strategies, personalized customer experiences, and targeted service improvements. By addressing the unique needs and behaviours of each segment, businesses can enhance customer satisfaction and drive revenue growth.

Looking forward, integrating advanced machine learning models and additional features like time of travel, traffic patterns, and weather conditions could further refine the segmentation process. This analysis serves as a foundation for continuous improvement and innovation in customer-centric decision-making.

In conclusion, this study demonstrates the critical role of data science in uncovering actionable insights and empowering businesses to make informed, strategic decisions.

10. Codes

All the codes used in this project can be found on Feynn-Labs-Internship-2025/Project 3 - Online Vehicle Booking Market Segmentation at main · PremKr1122/Feynn-Labs-Internship-2025