

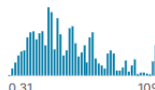

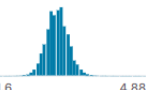
Report on Market Segmentation Analysis for Online Vehicle Booking

1. Objective

The primary goal of this analysis is to perform a comprehensive market segmentation on the online cab booking dataset to identify distinct customer groups. Market segmentation enables the company to better understand the diverse preferences, needs, and behaviors of its customers. This understanding is crucial for tailoring services, improving customer satisfaction, and enhancing the company's competitive edge. By delving into the characteristics of different customer segments, the company can allocate resources more efficiently and implement strategies that align closely with customer expectations.

2. Dataset Overview

The dataset utilized for this analysis, `sigma_cabs.csv`, contains rich information related to customer behavior and service dynamics in the cab booking market. Below are the key features and their implications for the analysis:

sigma_cabs.csv (7.68 MB)						
Detail Compact Column						
10 of 14 columns						
About this file						
History of customer service.						
▲ Trip_ID	# Trip_Distance	▲ Type_of_Cab	# Customer_Since_...	# Life_Style_Index	▲ Confidence_Life_...	▲ Destination_Type
ID for TRIP (Can not be used for purposes of modelling)	The distance for the trip requested by the customer	Category of the cab requested by the customer	Customer using cab services since n months; 0 month means current month	Proprietary index created by Sigma Cabs showing lifestyle of the customer based on their behaviour	Category showing confidence on the index mentioned above	Sigma Cabs divides an destination in one of 14 categories.
131662 unique values		B 24% C 21% Other (72404) 55%			B 31% C 27% Other (55340) 42%	A 5 B 2 Other (24510) 1
T0005689460	6.77	B	1	2.42769	A	A
T0005689461	29.47	B	10	2.78245	B	A

- **Trip_Distance:** Represents the distance traveled per trip, which helps in understanding whether customers predominantly use short-haul or long-haul services.

- **Type_of_Cab:** Indicates the category of cab preferred by customers, ranging from economy to luxury, shedding light on budget preferences.
- **Customer_Since_Months:** Shows the duration of customer association with the service, providing insights into customer loyalty and retention trends.
- **Life_Style_Index:** A calculated score representing customers' lifestyle attributes, correlating with their purchasing power and preferences.
- **Confidence_Life_Style_Index:** Reflects the reliability of the lifestyle index data, ensuring robust analysis.
- **Destination_Type:** Categorizes destinations into types such as airports, business districts, or residential areas, highlighting common travel purposes.
- **Customer_Rating:** Denotes customer satisfaction levels through ratings provided after each trip, directly tied to service quality.
- **Cancellation_Last_1Month:** Records cancellations made by customers within the past month, indicative of potential dissatisfaction or inconsistent needs.
- **Surge_Pricing_Type:** Identifies trips affected by dynamic pricing, offering insights into customer behavior during peak demand.

3. Data Preprocessing

Data preprocessing is a vital step to ensure the quality and reliability of the analysis. Here are the specific preprocessing techniques applied:

- **Handling Missing Data:**
 - Missing values in the **Type_of_Cab** column were imputed using the mode, which is the most frequently occurring value. This approach ensures consistency and retains the dataset's integrity.

```
[11] df.Type_of_Cab = df.Type_of_Cab.fillna('B')
```

- Missing values in numerical features were filled with appropriate statistical measures (e.g., mean or median) to minimize bias.

```
df.Life_Style_Index = df.Life_Style_Index.fillna(4)
```

```
df.Customer_Since_Months = df.Customer_Since_Months.fillna(df.Customer_Since_Months.median())
```

- **Encoding Categorical Variables:**

- Categorical variables, such as `Type_of_Cab`, `Destination_Type`, and `Gender`, were converted into numerical representations using Label Encoding. This step allows machine learning algorithms to process the data effectively.

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()

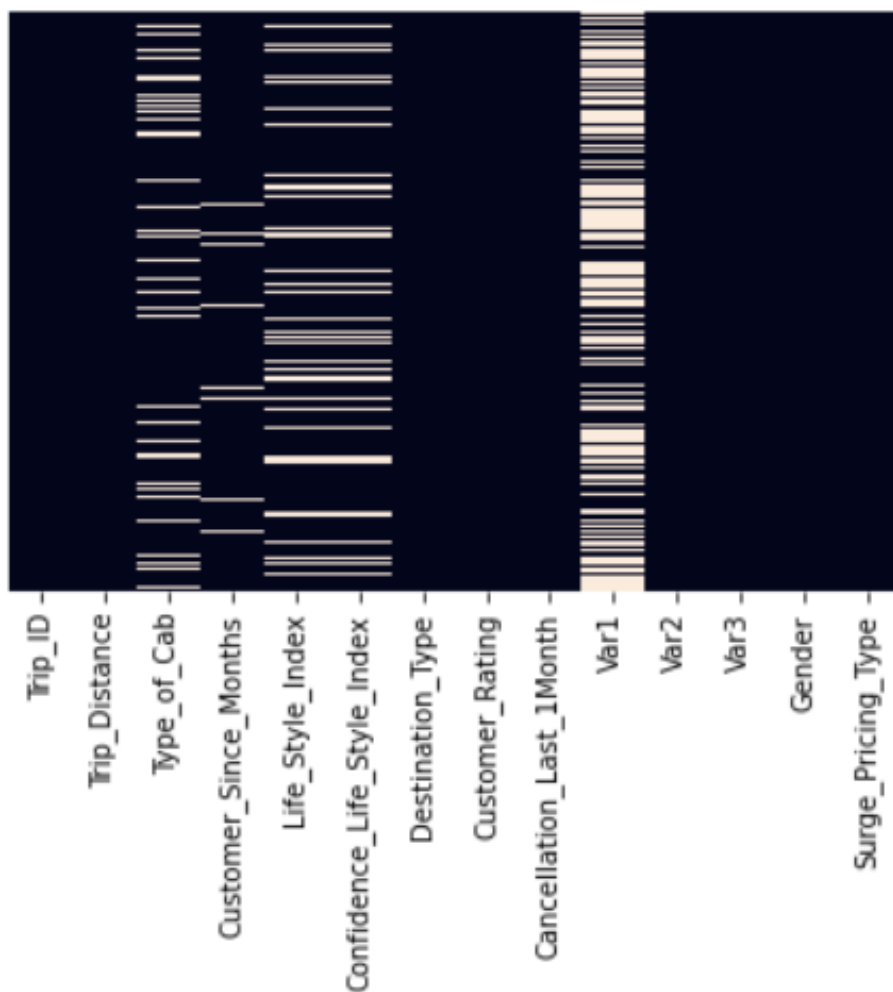
df["Type_of_Cab"]=le.fit_transform(df["Type_of_Cab"])
df["Confidence_Life_Style_Index"] =le.fit_transform(df["Confidence_Life_Style_Index"])
df["Destination_Type"]=le.fit_transform(df["Destination_Type"])
df["Gender"]=le.fit_transform(df["Gender"])
```

- **Scaling Numerical Features:**

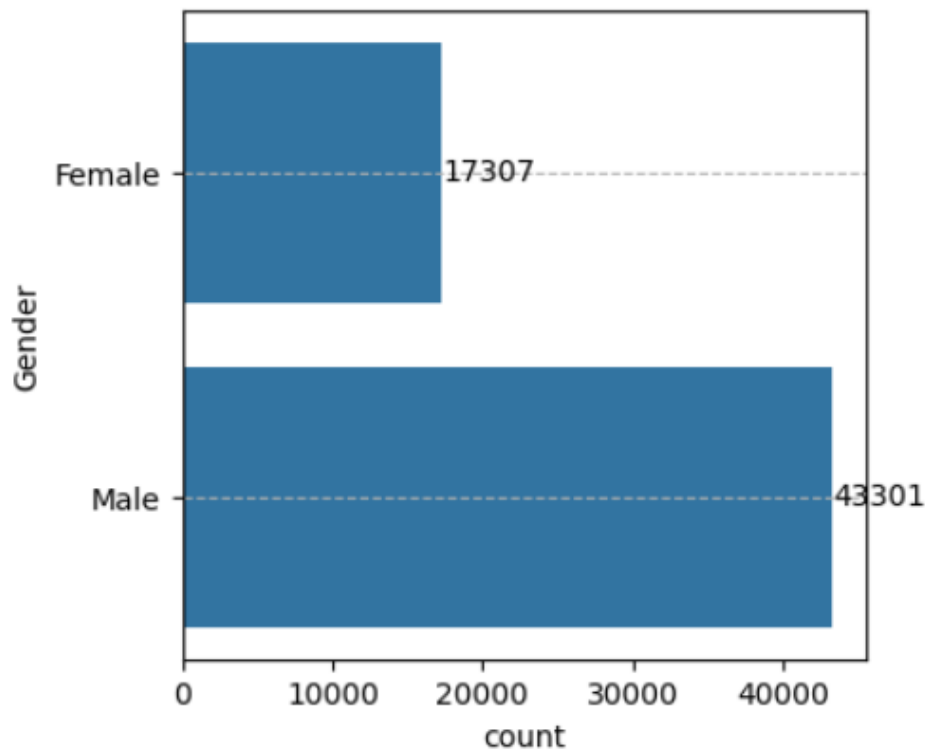
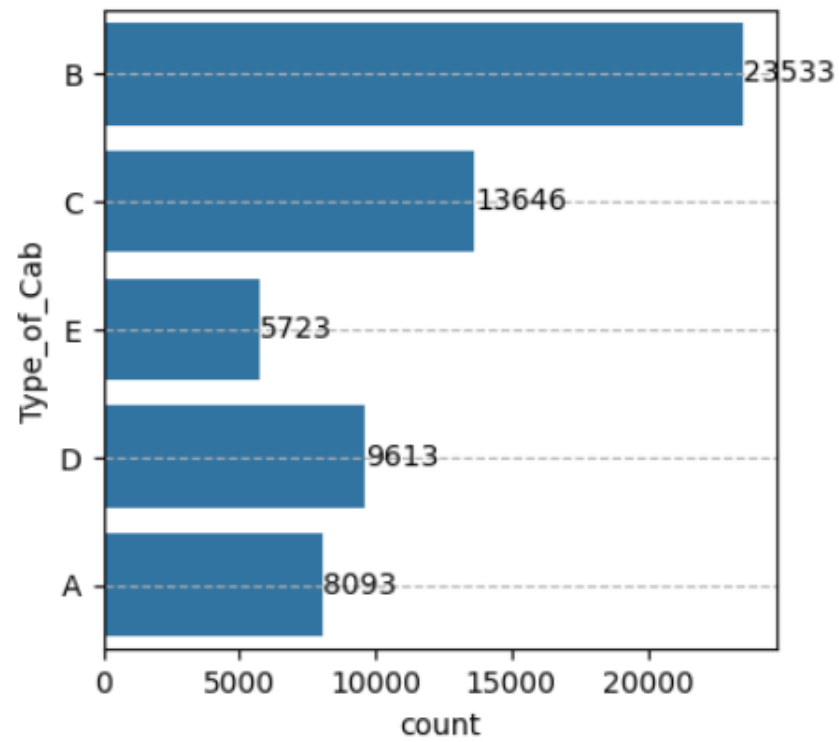
- Continuous features like `Trip_Distance` and `Life_Style_Index` were standardized using `StandardScaler`. Standardization ensures that features are on a comparable scale, preventing bias in clustering algorithms.

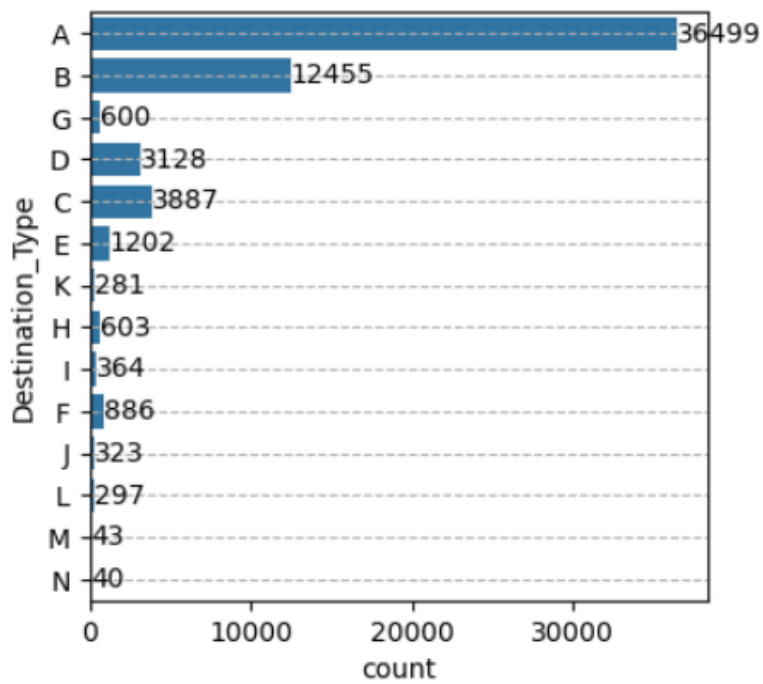
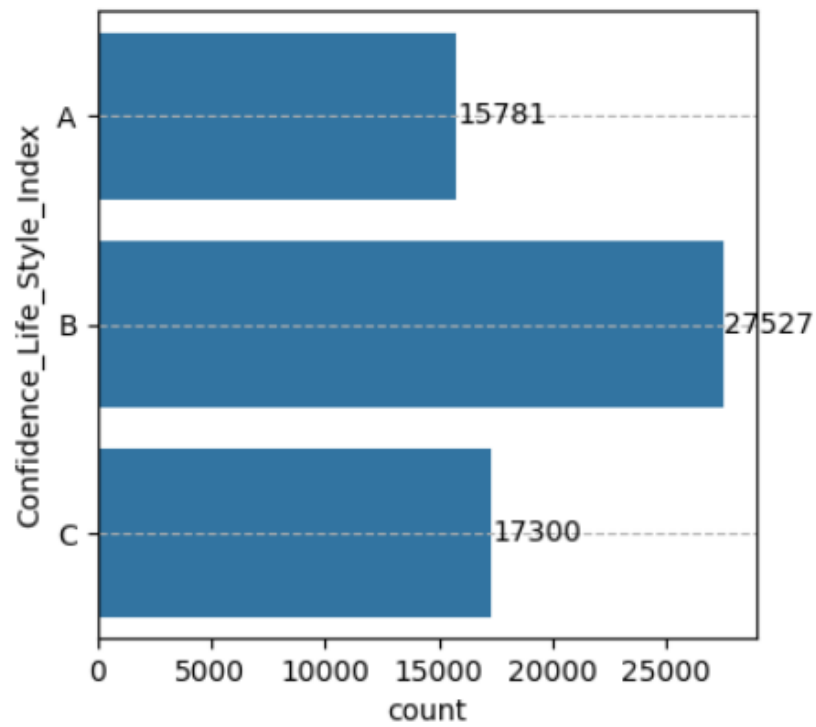
```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit_transform(df[columns])
```

4. Exploratory Data Analysis (EDA)



- **Visualizing Missing Data:** A heatmap was created to visualize missing values, helping identify patterns and determine the extent of missing data across columns.
- **Feature Distribution Analysis:**
 - Histograms and boxplots were generated to understand the distribution of features like `Customer_Rating`, `Trip_Distance`, and `Surge_Pricing_Type`. For example, `Type_of_Cab` distribution revealed a significant preference for mid-tier cab options (category B).
 - The analysis of `Customer_Since_Months` highlighted a mix of long-term loyal customers and newer users.





5. Market Segmentation Approach

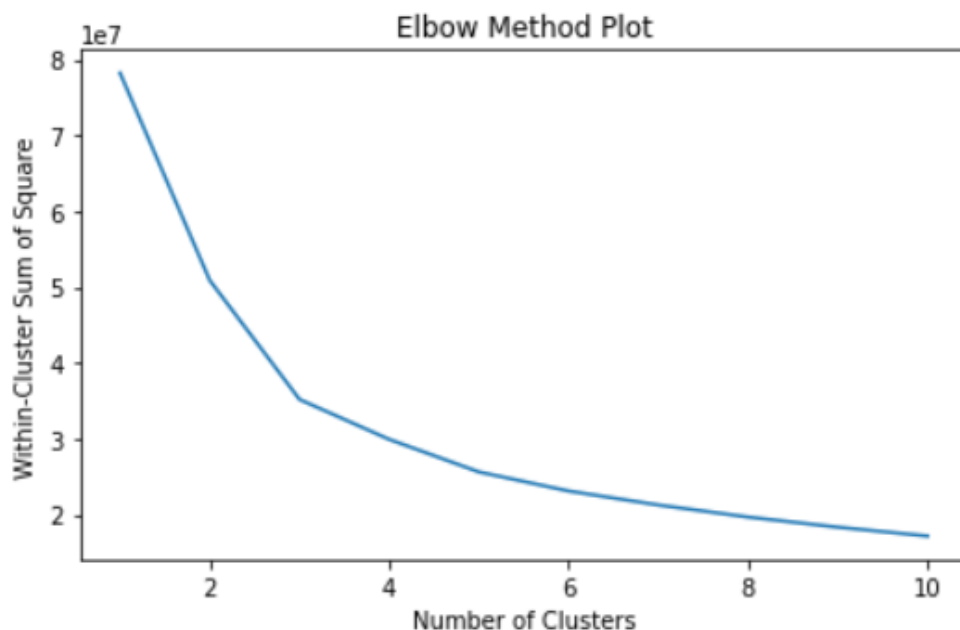
Market segmentation was achieved using clustering techniques to group customers based on their behaviors and preferences.

- **Choosing Clustering Method:** The K-Means clustering algorithm was selected

due to its efficiency and effectiveness in segmenting data into distinct clusters. This algorithm minimizes intra-cluster variance while maximizing inter-cluster differences.

```
for i in range(1, 11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++',
                    max_iter = 300, n_init = 10, random_state = 0)
    kmeans.fit(df)
    wcss.append(kmeans.inertia_)
```

- **Determining Optimal Clusters:** The Elbow Method was applied to determine the ideal number of clusters. By plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters, the “elbow point” was identified as the point of diminishing returns.



- **Selecting Features for Clustering:** Key features, including Trip_Distance, Life_Style_Index, Customer_Rating, and Surge_Pricing_Type, were selected for clustering based on their ability to capture customer preferences and service usage patterns.

6. Cluster Insights and Interpretation

- **Cluster Characteristics:**
 - **Cluster 1:** Comprised of customers with a high lifestyle index, high customer ratings, and frequent cab usage. These customers exhibit loyalty and are inclined towards premium services.
 - **Cluster 2:** Includes newer customers with moderate ratings and less frequent usage. This segment represents an opportunity to build loyalty and increase engagement.

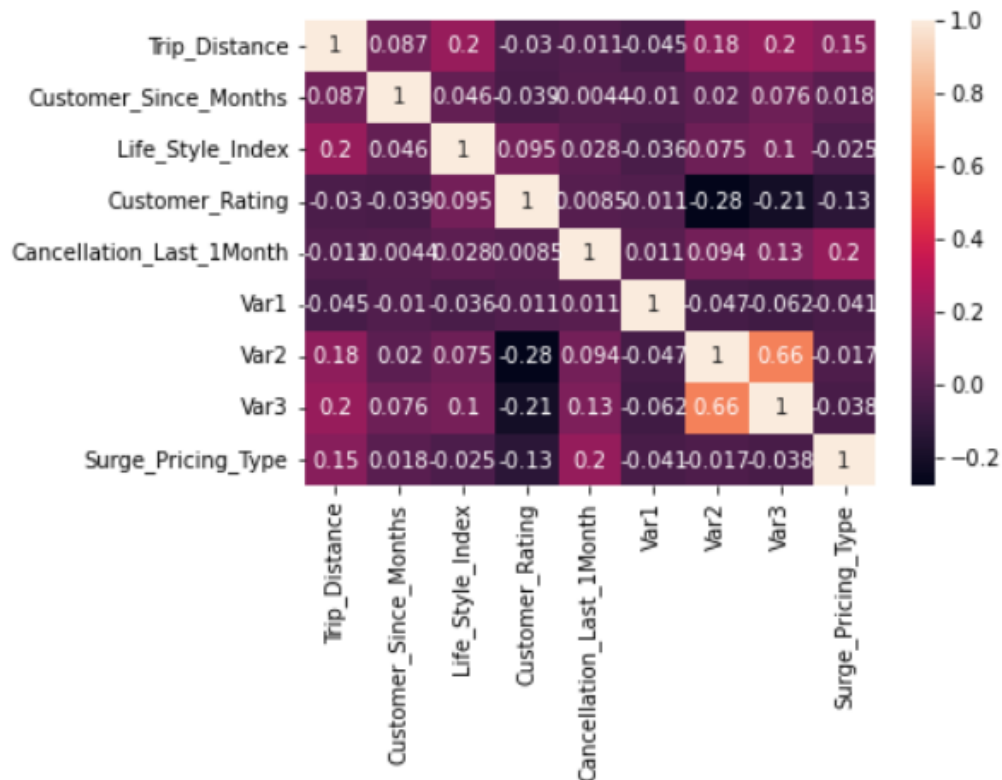
- **Cluster 3:** Consists of budget-conscious customers who tend to cancel bookings frequently. Retention efforts can focus on addressing their cost concerns.

- **Demographic and Behavioral Trends:**

- Analysis of gender distribution within clusters revealed differing preferences for cab types and trip purposes. For instance, females may prefer shorter trips, while males may dominate longer commutes.
- Surge pricing analysis indicated that certain customer segments are less sensitive to price hikes, providing an opportunity for dynamic pricing strategies.

7. Visualization and Analysis

- **Heatmap of Correlations:** A correlation heatmap was used to identify relationships among variables. For example, a strong correlation was observed between `Life_Style_Index` and `Type_of_Cab`, indicating that lifestyle heavily influences cab preferences.



- **Cluster Visualization:** PCA (Principal Component Analysis) was applied to reduce dimensionality and visualize clusters in a 2D space. This visualization provided clear distinctions among the clusters, aiding interpretation.

8. Strategic Recommendations

Based on the insights from clustering, the following strategies are recommended:

- **Segment-Specific Marketing Campaigns:**
 - **Cluster 1:** Focus on premium offerings, loyalty programs, and personalized services to maintain satisfaction and engagement.
 - **Cluster 2:** Launch introductory promotions, referral incentives, and awareness campaigns to convert new users into loyal customers.
 - **Cluster 3:** Design cost-effective packages and offer flexible cancellation policies to address concerns and improve retention.
- **Dynamic Pricing Models:** Develop pricing strategies that align with the willingness to pay of different customer segments, especially during peak demand periods.
- **Enhanced Customer Retention Strategies:**
 - Collect regular feedback to identify pain points and implement service improvements.
 - Provide targeted offers and rewards to encourage repeat usage among less frequent users.

9. Conclusion

This market segmentation analysis highlights the diverse characteristics and preferences of the customer base. By implementing the suggested strategies, the company can better cater to its customers, drive growth, and build a sustainable competitive advantage. The insights from this analysis underscore the importance of tailored services, strategic pricing, and targeted marketing in the highly competitive online cab booking industry.