REPORT SUMMARY ON

ONLINE VEHICLE BOOKING SYSTEM

Presented By:

Kunal Verma

Rohit Valsetwar

Archana Rajeshwar

Gunda Sri Harini

ABSTRACT

In the highly competitive Indian market for cab booking services, our Online Vehicle Booking Product Startup is seeking an alternative segment that can provide early market entry and generate revenue. To achieve this, we conducted a segmentation analysis of the vehicle market in India. The aim was to identify segments where our vehicle booking service can potentially generate profits.

Methods

We utilized segmentation analysis techniques and examined various factors to understand the Indian vehicle market. By analyzing data using machine learning algorithms, we sought to identify viable segments for our services.

Results

The study employed three different machine learning algorithms – Clustering & Regression techniques - implemented in Python. Among these algorithms, the model based on the Regression algorithm demonstrated the best performance. It achieved a higher R squared error, indicating its effectiveness in predicting suitable districts for bone supplement production.

Conclusion

By applying machine learning algorithms and segmentation analysis, we were able to identify the most suitable districts for bone supplement production. These findings will aid our Online Vehicle Booking Product Startup in formulating a feasible strategy to enter the vehicle booking market and target profitable segments in India.

1. PROBLEM STATEMENT

Our team is part of an Online Vehicle Booking Product Startup that aims to establish a strong presence in the Indian market, considering the intense competition from industry giants like Ola and Uber in the cab booking sector. To gain an early foothold and generate revenue, we are seeking an alternative segment within the vehicle market in India that can be targeted effectively.

The task at hand is to conduct a comprehensive analysis of the Indian vehicle market using segmentation techniques. By leveraging segmentation analysis, we aim to identify specific segments within the market that hold the potential for profitability, taking into account factors such as geographic location, demographic characteristics, psychographic attributes, and behavioral patterns.

In this project, our focus is to develop a feasible strategy that allows us to penetrate the market by offering our vehicle booking services to the identified segments. To achieve this, we will analyze various segmentation dimensions such as geographic preferences, customer demographics, psychographic traits (such as lifestyle and preferences), and behavioral patterns. Additionally, we will consider other relevant factors such as price sensitivity, service expectations, availability of vehicles, and technological preferences within the target segments.

By thoroughly understanding and analyzing the vehicle market in India through segmentation analysis, we aim to devise an effective strategy that positions our Online Vehicle Booking Product Startup to capture early market share and generate sustainable revenue.

2. DATA COLLECTION

Data collection for performed by 2 teams, each managing distinct datasets for project analysis. Link to datasets are provided below:

Dataset 1: https://github.com/sriharinigunda/Online-Vehical-Bookings-Market-Segmentation/blob/main/OLA_trips_dataset.csv

Dataset 2: https://github.com/MYSTIC-HUNTER/Online-Vehicle-Booking/blob/main/Cab_Aggregator_Problem_Dataset.csv

Each column of Dataset 1 explained below:

- Booking id: A unique identifier assigned to each booking made through the online vehicle booking service.
- Booking_date_time: The date and time when the booking was made.
- Distance_travelled: The distance covered during the trip.
- Time_taken: The duration of the trip, typically measured in minutes or hours.
- Commission_base_cost: The base cost of the trip that includes the commission charged by the online vehicle booking service.
- Driver_base_cost: The base cost paid to the driver for the trip.
- Total_tax: The total tax amount associated with the trip.
- Total_trip_cost: The overall cost of the trip, including the base cost, tax, and any additional charges.
- Ratings: The ratings provided by the customers to rate their experience with the trip or the driver.

Each column of Dataset 2 explained below:

- City: The name of the city where the vehicle booking data was collected.
- Zone: The specific zone or area within the city where the vehicle booking data was recorded.
- Week Number: The number corresponding to the week during which the data was collected.
- Day: The day of the week when the vehicle bookings took place.
- 4 hour windows: Divisions of time into four-hour intervals to track booking patterns throughout the day.

- Date: The specific date when the vehicle bookings occurred.
- Time stamp date: The date and time when the vehicle booking data was recorded.
- Demand: The number of vehicle booking requests or demand during a specific time period.
- Fulfillment %: The percentage of booking requests that were successfully fulfilled by providing a vehicle.
- Avg. Price: The average price or fare for the vehicle bookings.
- Avg. Trip Size: The average number of passengers or trip size per booking.
- Cancellation % (Total): The overall percentage of booking cancellations, including both driver and customer cancellations.
- Driver Cancellation %: The percentage of bookings canceled by drivers.
- Customer Cancellation %: The percentage of bookings canceled by customers.
- Avg. ETA: The average estimated time of arrival for the vehicles to reach the customers.

3. DATA PRE-PROCESSING: (STEPS AND LIBRARIES USED)

Importing Libraries: firstly, we will import the libraries for our model, which is part of data pre-processing. The code is given below:

/*import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans */

• Numpy we have imported for the performing mathematics calculation.
• Matplotlib is for plotting the graph, and pandas are for managing the dataset.
• Seaborn is for data visualization library, it is based on matplotlib.
• Scikit-learn have sklearn.cluster.KMeans module to perform K-Means clustering. While computing cluster centers and value of inertia, the parameter named sample_weight allows sklearn.cluster.KMeans module to assign more weight to some samples

4. SEGMENT EXTRACTION:

Advantage for K-Means:

• Relatively simple to implement.

• Scales to large data sets.

• Guarantees convergence.

• Can warm-start the positions of centroids.

• Easily adapts to new examples.

• Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

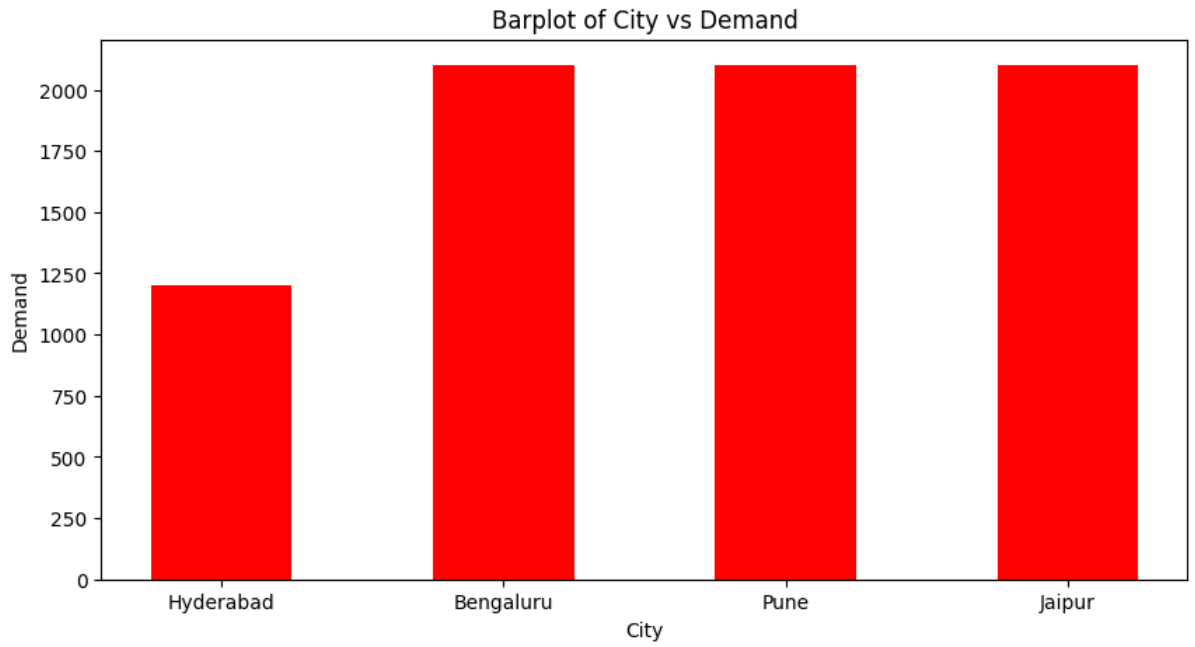
5. EXPLORATORY DATA ANALYSIS

An Exploratory Data Analysis, or EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for a company because it exposes trends, patterns, and relationships that are not readily apparent.

We analysed our dataset using univariate (analyze data over a single variable/column from a dataset), bivariate (analyze data by taking two variables/columns into consideration from a dataset) and multivariate (analyze data by taking more than two variables/columns into consideration from a dataset) analysis.

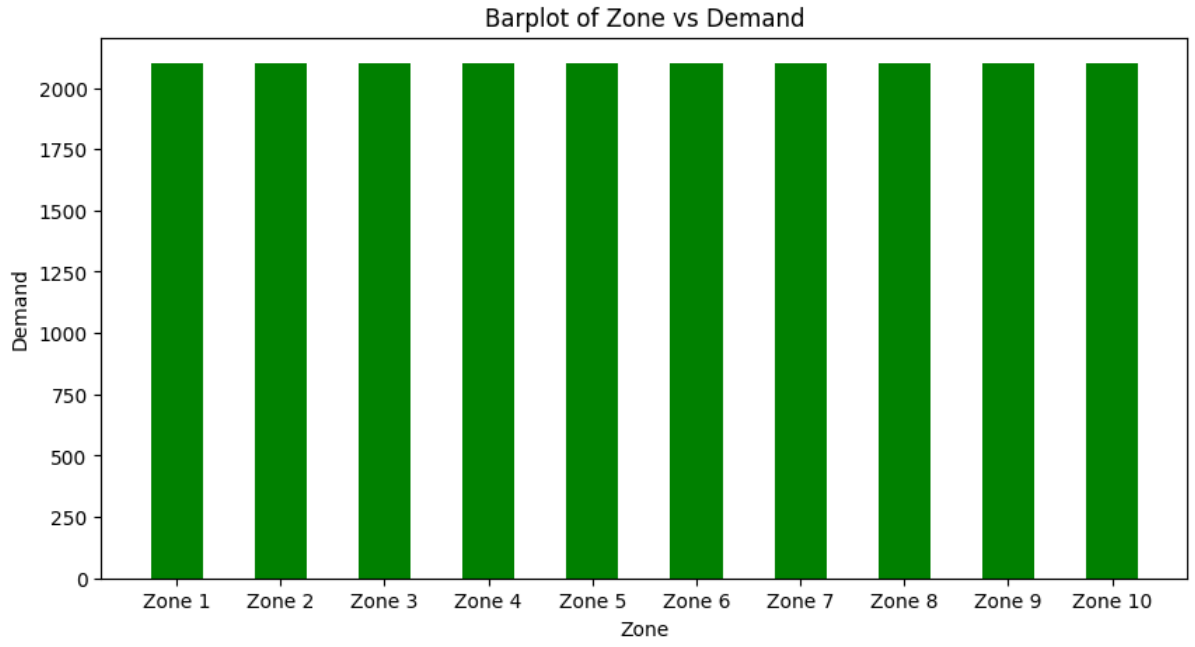
```
plt.figure(figsize=(10, 5))
plt.bar(df['City'], df['Demand'],color='red', width=0.5)
plt.title('Barplot of City vs Demand')
plt.xlabel('City')
plt.ylabel('Demand')
plt.show()
```

This Bar graph shows the diversity of the data geographically. We can see that we have Demands in different Cities.



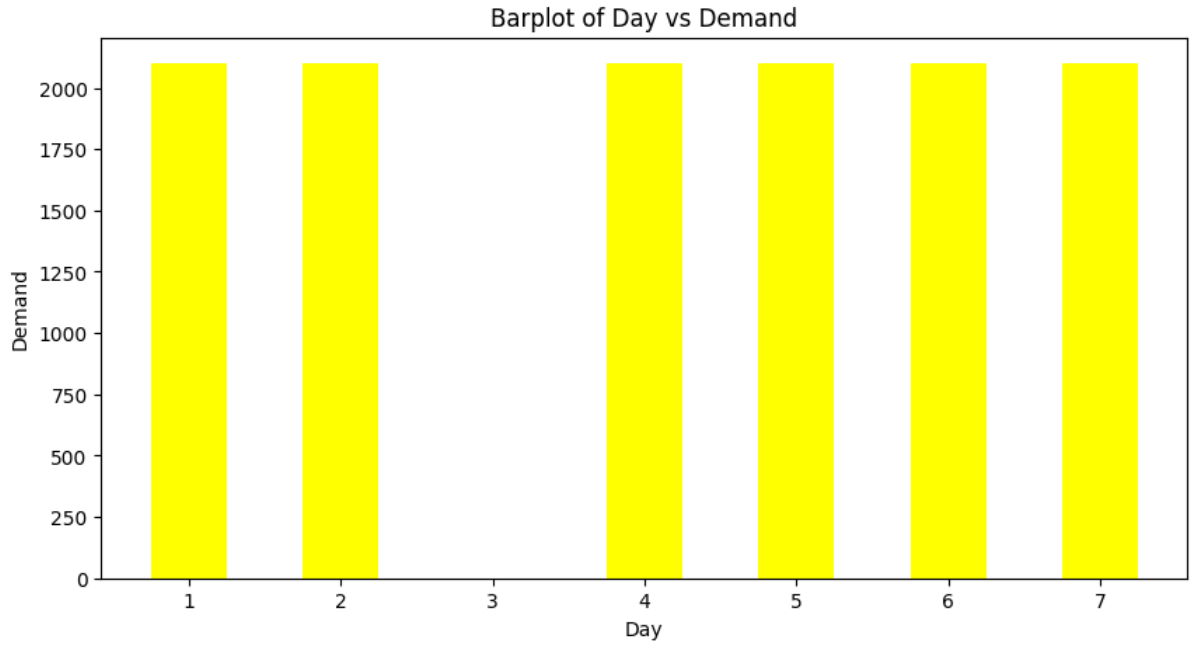
```
plt.figure(figsize=(10, 5))
plt.bar(df['Zone'], df['Demand'],color='green', width=0.5)
plt.title('Barplot of Zone vs Demand')
plt.xlabel('Zone')
plt.ylabel('Demand')
plt.show()
```

This Bar graph shows the diversity of the data geographically. We can see that we have Demands in different City Zones.



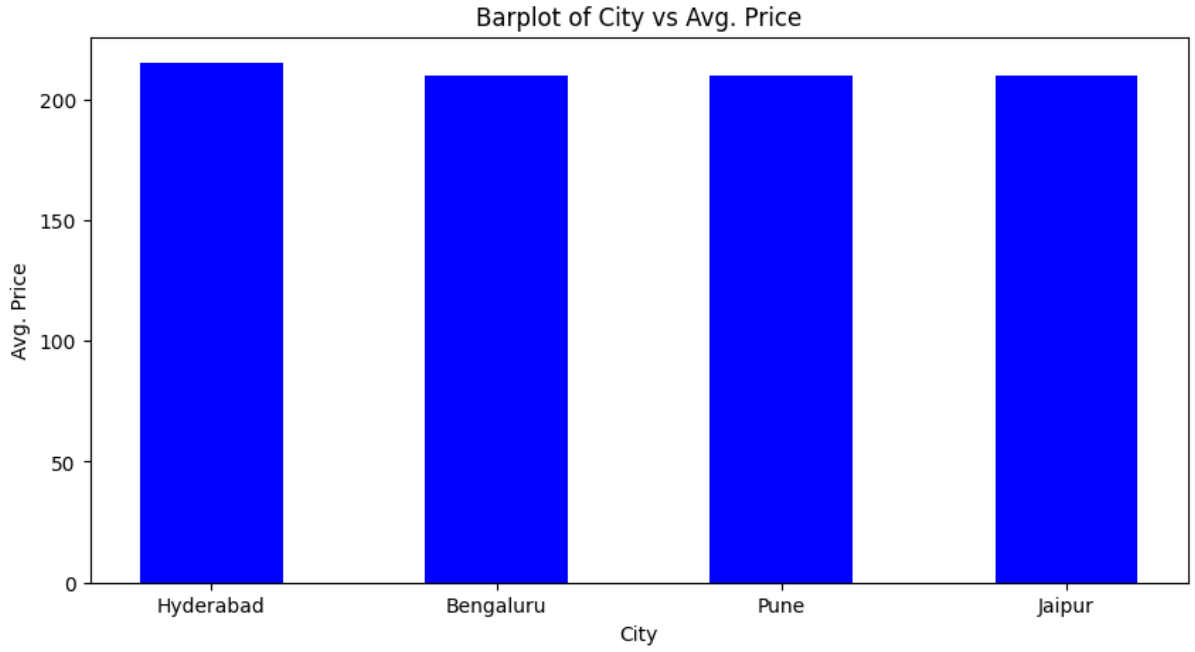
```
plt.figure(figsize=(10, 5))
plt.bar(df['Day'], df['Demand'],color='yellow', width=0.5)
plt.title('Barplot of Day vs Demand')
plt.xlabel('Day')
plt.ylabel('Demand')
plt.show()
```

This Bar graph shows the diversity of the data chronologically. We can see that we have Demands on different days.



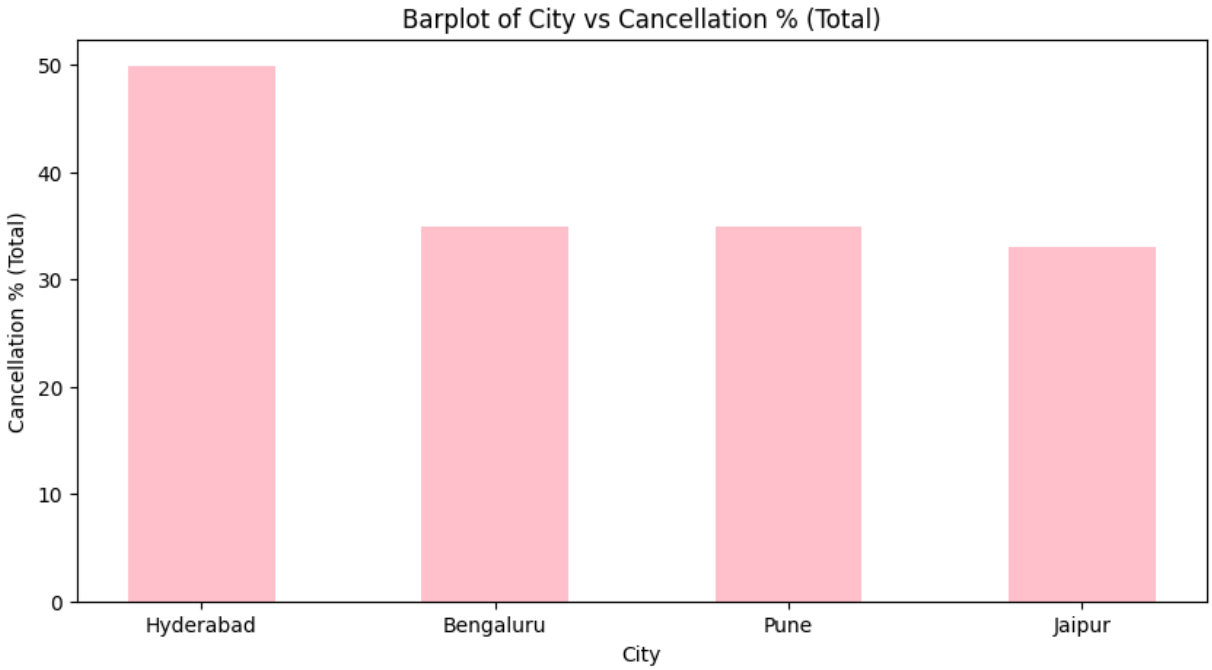
```
plt.figure(figsize=(10, 5))
plt.bar(df['City'], df['Avg. Price'],color='blue', width=0.5)
plt.title('Barplot of City vs Avg. Price')
plt.xlabel('City')
plt.ylabel('Avg. Price')
plt.show()
```

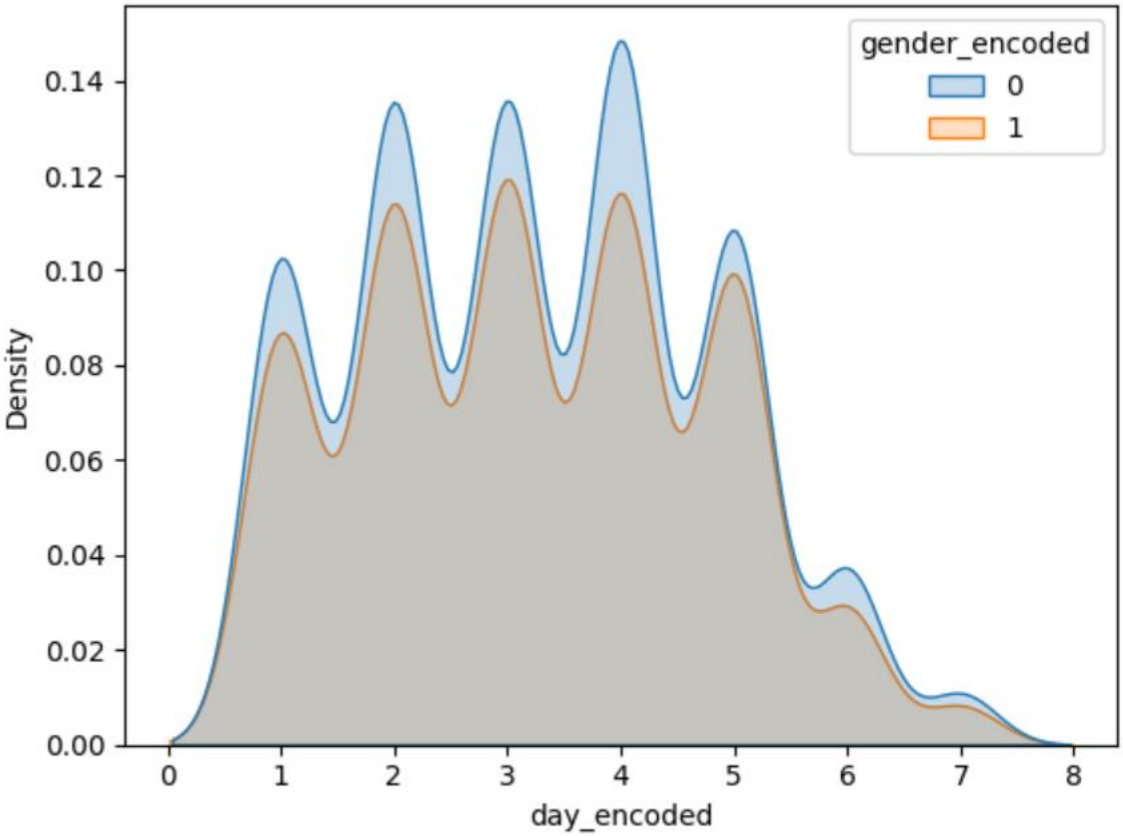
This Bar graph shows the diversity of the data geographically. We can see that we have Demands in different Cities.

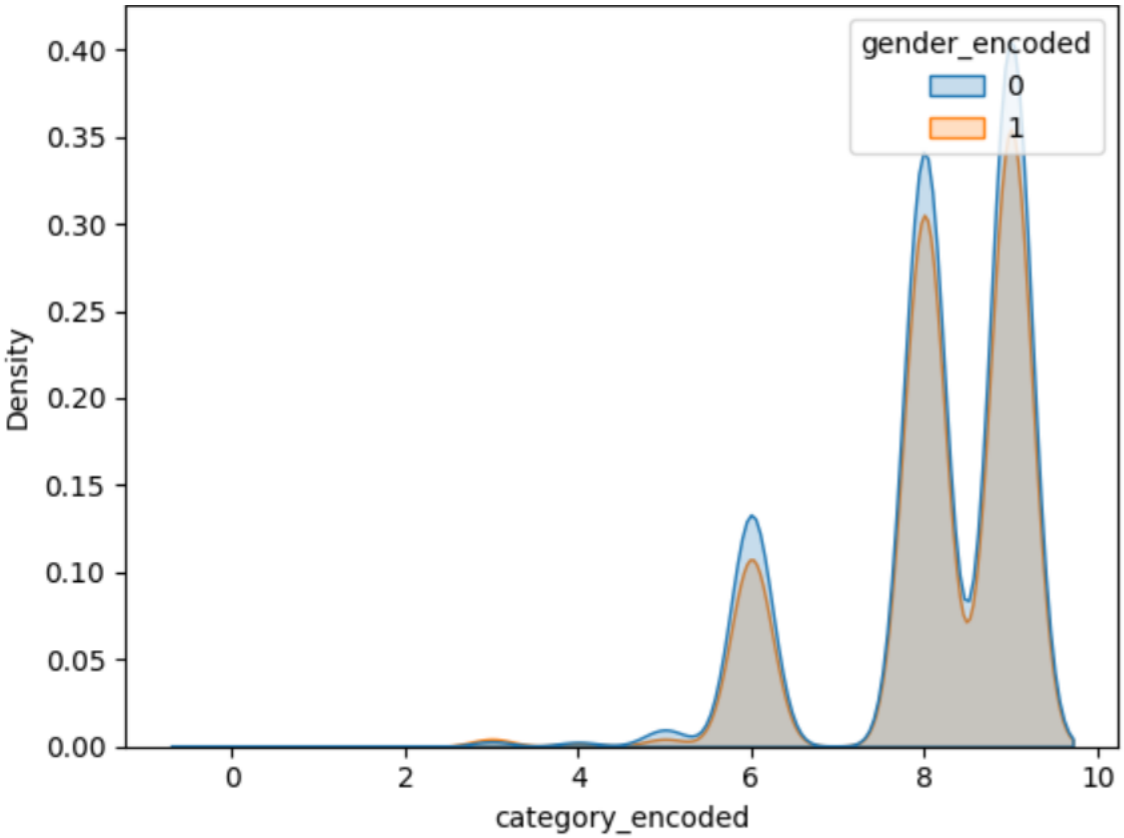


```
plt.figure(figsize=(10, 5))
plt.bar(df['City'], df['Cancellation % (Total)'],color='pink',
width=0.5)
plt.title('Barplot of City vs Cancellation % (Total)')
plt.xlabel('City')
plt.ylabel('Cancellation % (Total)')
plt.show()
```

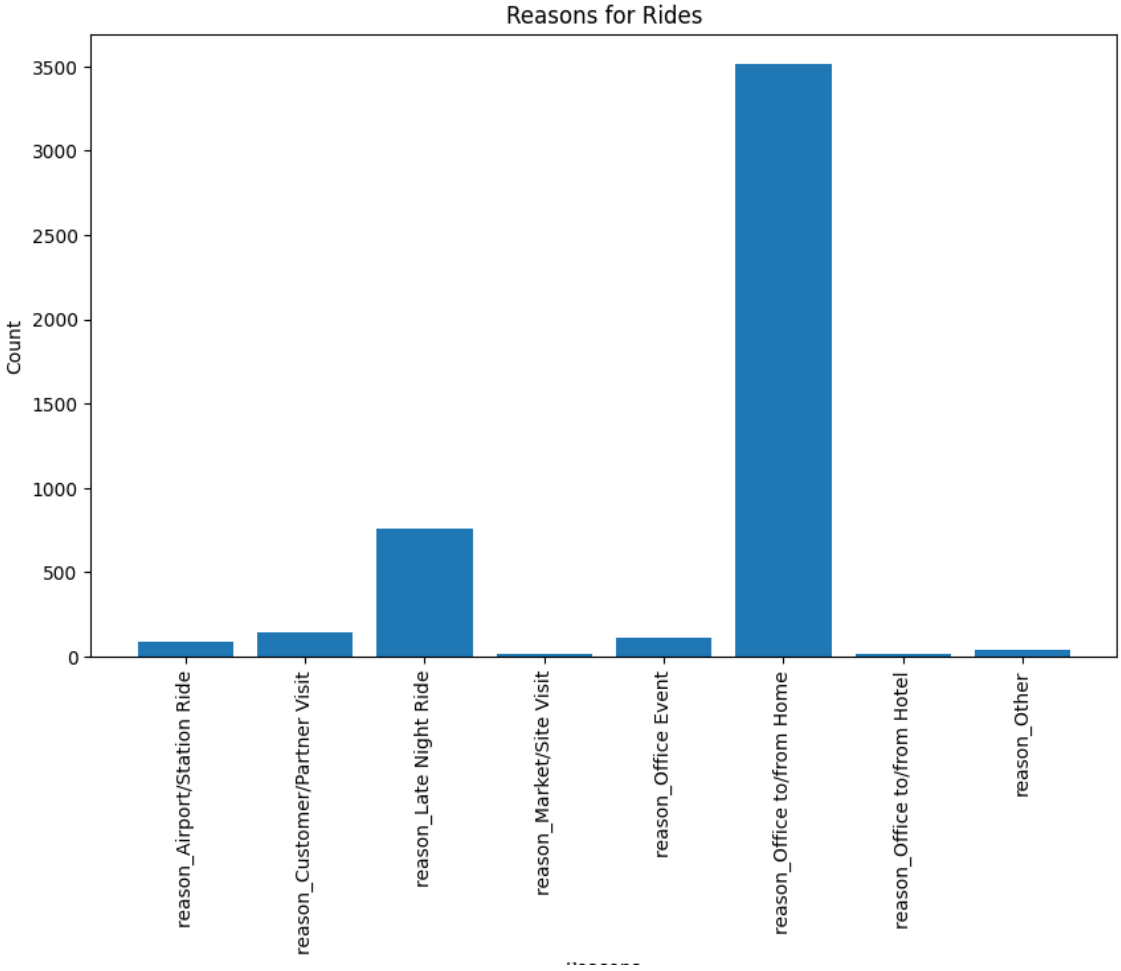
This Bar graph shows the diversity of the data geographically. We can see that we have chances of overall cancellations in various cities.




gender_encoded
0
1
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
1
2
3
4
5
6
7
8
day_encoded

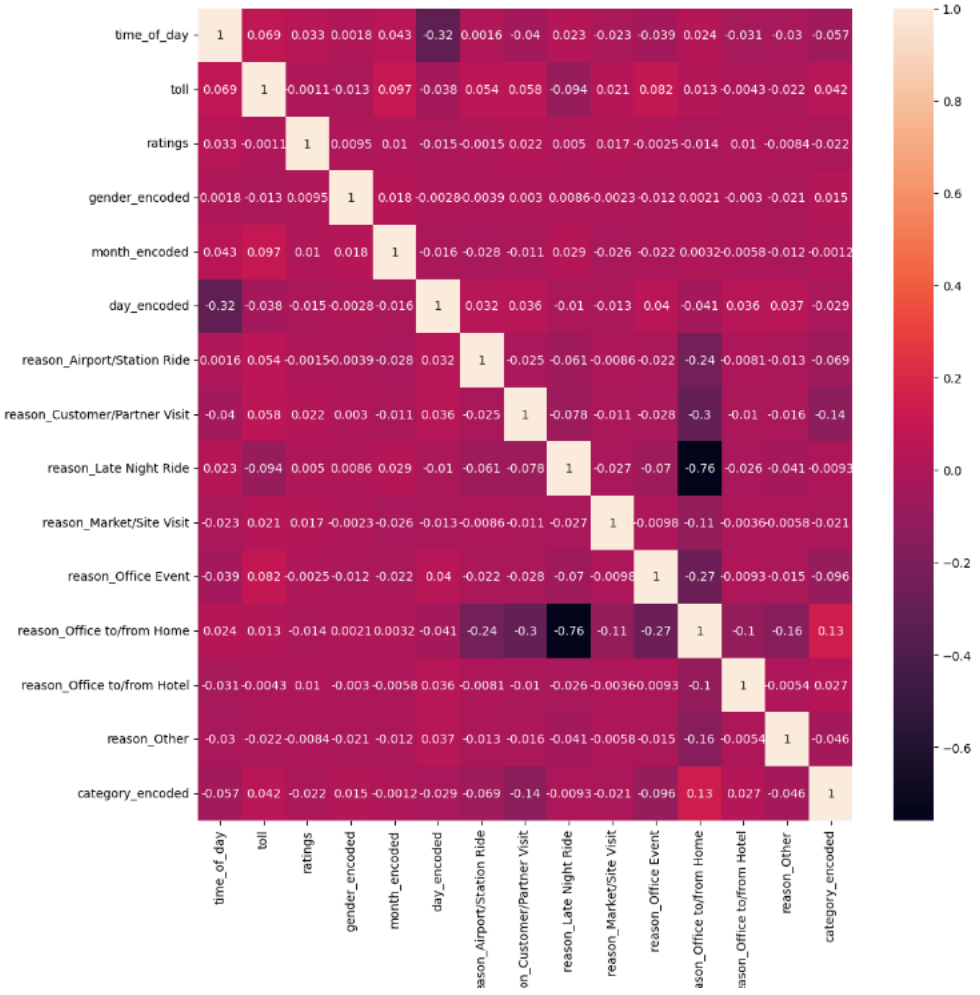

gender_encoded
0
1
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
category_encoded

From dataset 2 KDE plots helped us understanding the density. From the given plot it shows that most of the people who book cabs are females as compared to males. This gives an better understanding of the market where gender can be one way to describe the segment.



The above bar plot explains that most reasons for rides are reason from office to / from home or late night rides. This gives us an potential market for dividing the segments based on the psychographic segmentation and we also found that working people mostly book cabs for going to office.

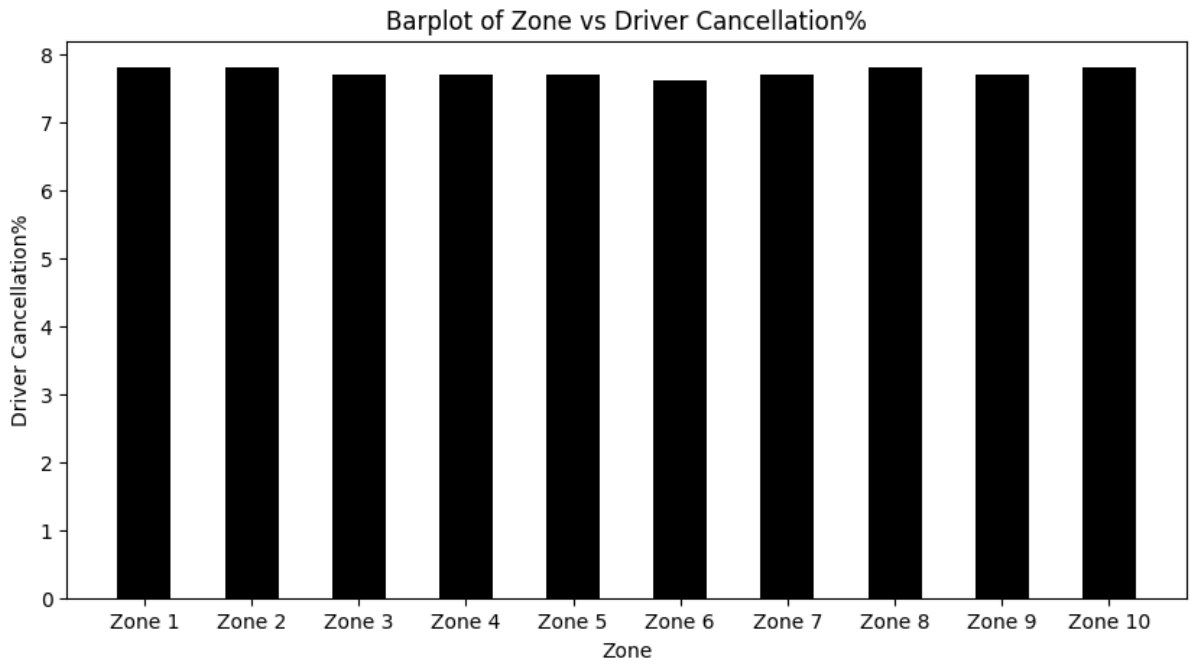
```
plt.figure(figsize=(12,12))
sns.heatmap(df.corr(), annot=True)
plt.show()
```



Heatmap provides us with the most relevant features in the 2nd dataset.

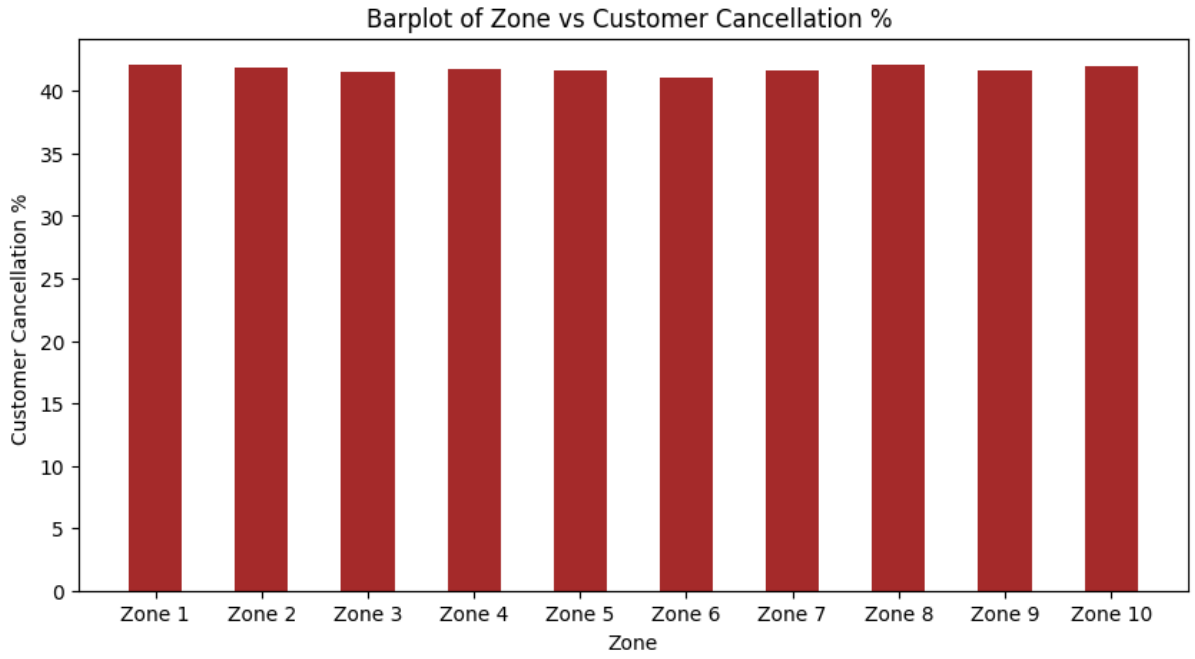
```
plt.figure(figsize=(10, 5))
plt.bar(df['Zone'], df['Driver Cancellation%'],color='black',
width=0.5)
plt.title('Barplot of Zone vs Driver Cancellation%')
plt.xlabel('Zone')
plt.ylabel('Driver Cancellation%')
plt.show()
```

This Bar graph shows the diversity of the data geographically. We can see that we have chances of cancellation



```
plt.figure(figsize=(10, 5))
plt.bar(df['Zone'], df['Customer Cancellation %'],color='brown',
width=0.5)
plt.title('Barplot of Zone vs Customer Cancellation % ')
plt.xlabel('Zone')
plt.ylabel('Customer Cancellation % ')
plt.show()
```

This Bar graph shows the diversity of the data geographically. We can see that we have chances of cancellations by customer in various cities.

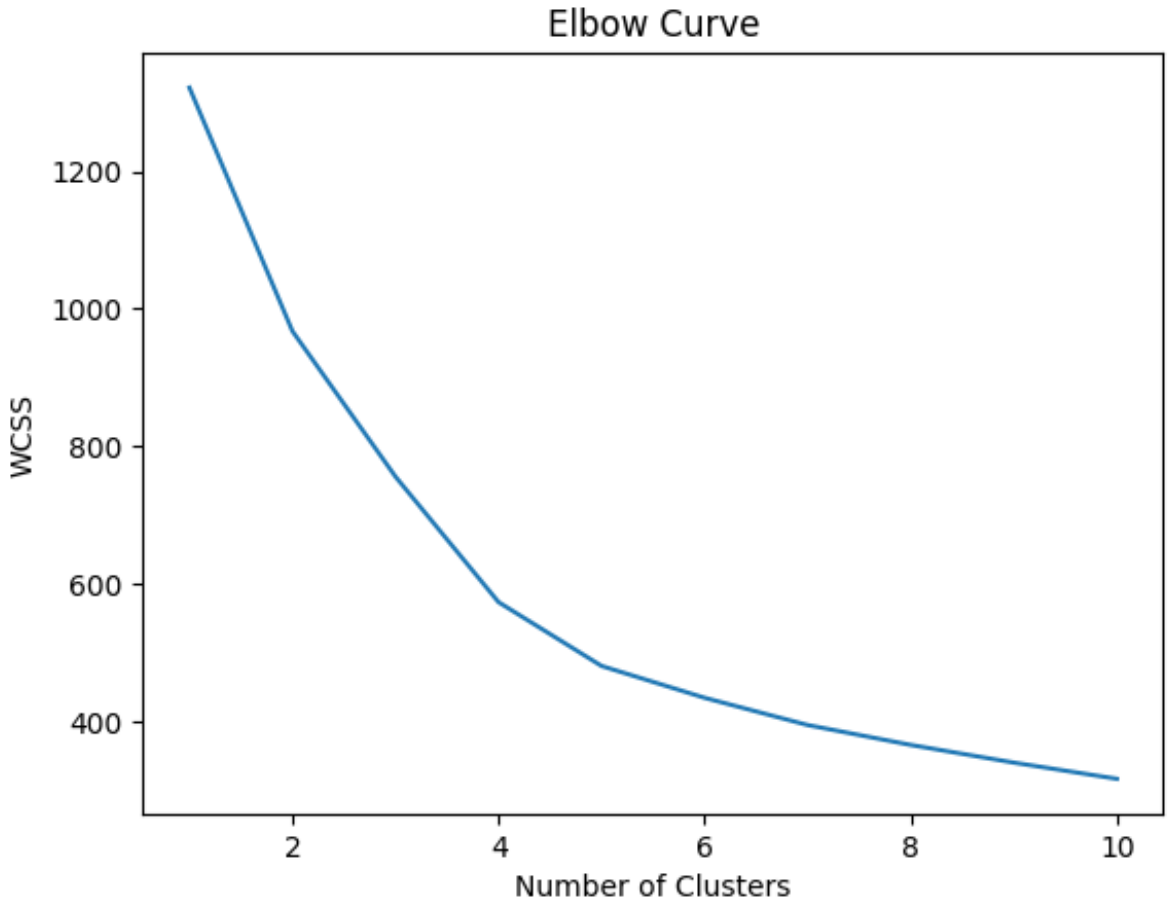


K-Means Clustering

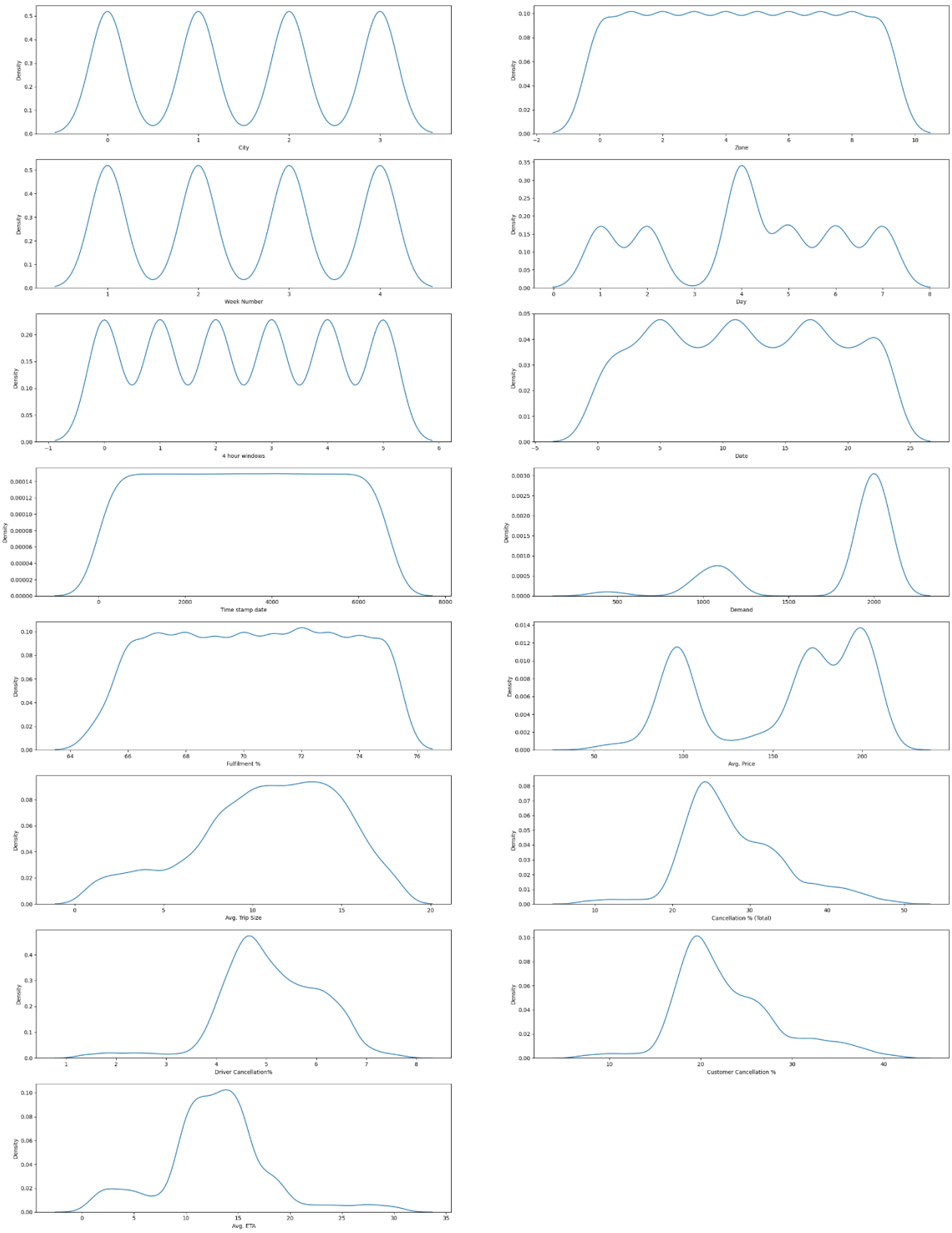
K-Means Clustering is an unsupervised learning algorithm used in machine learning and data science for clustering problems. It automatically groups unlabelled data into distinct clusters based on similarities. The algorithm minimizes the distances between data points and their respective cluster centroids. 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 number of clusters by identifying the point where additional clusters no longer significantly improve the model's performance.

```
from sklearn.cluster import KMeans
wcss = []
for k in range(1,15):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(df)
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(15,5))
plt.plot(range(1,15),wcss)
plt.xlabel("number of k (cluster) value")
plt.ylabel("wcss")
plt.show()
```

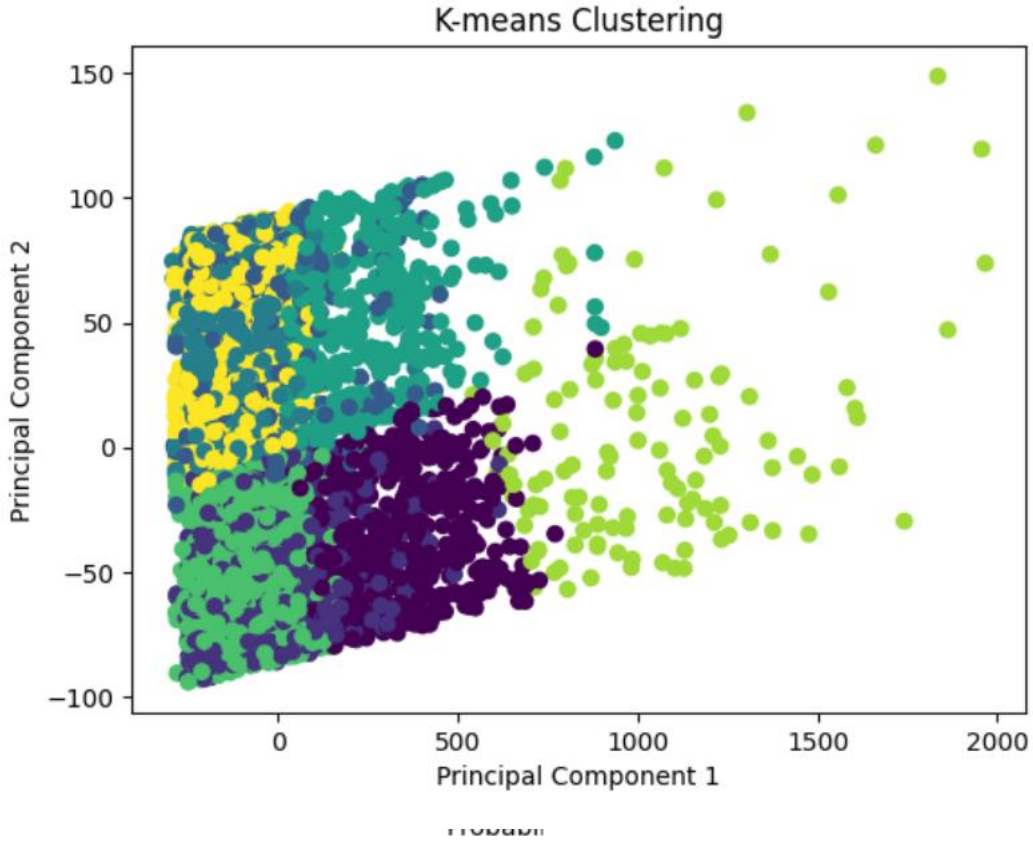


Based on the elbow curve analysis, it is determined that the optimal number of clusters is approximately 10. In our analysis, the silhouette score exhibits a significant increase beyond the value of 10, and the optimal range for the score falls between 8 and 10.



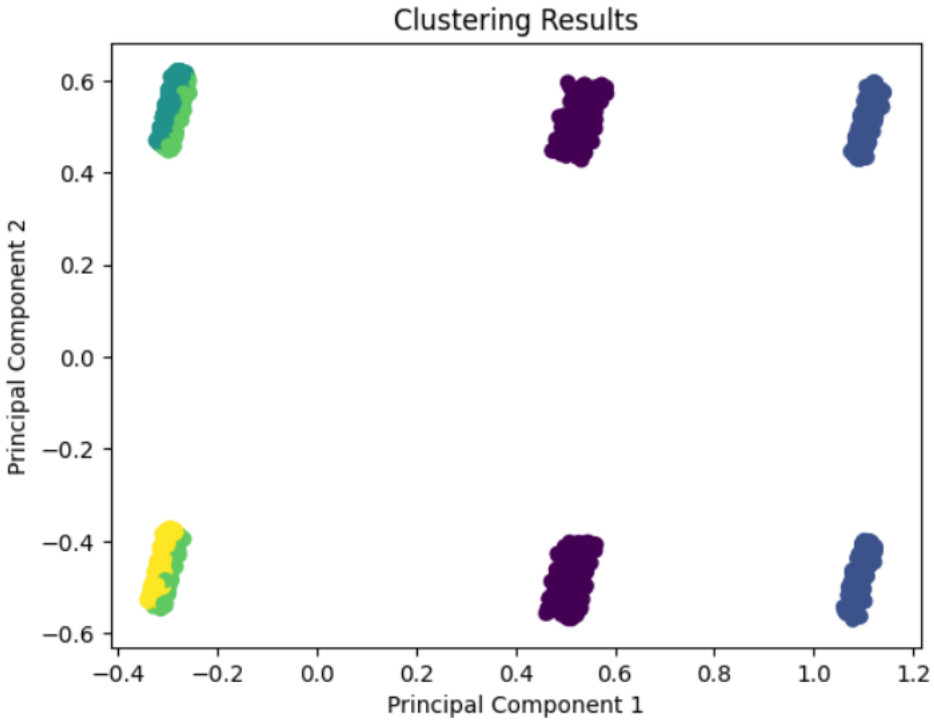
For all these graphs most of the data points lie within 4 and none go lesser than 3, this means the overall rating for vehicles.

This is for 2nd dataset , Based on the elbow curve analysis, it is determined that the optimal number of clusters is approximately 10. In our analysis, the silhouette score exhibits a significant increase beyond the value of 10, and the optimal range for the score falls between 8 and 10.

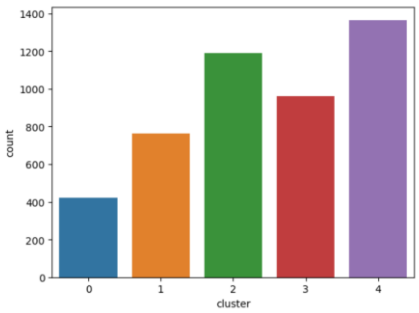



```
5    1120
7    1031
1     691
3     628
0     491
2     440
4     423
6     126
Name: cluster_kmeans, dtype: int64
```

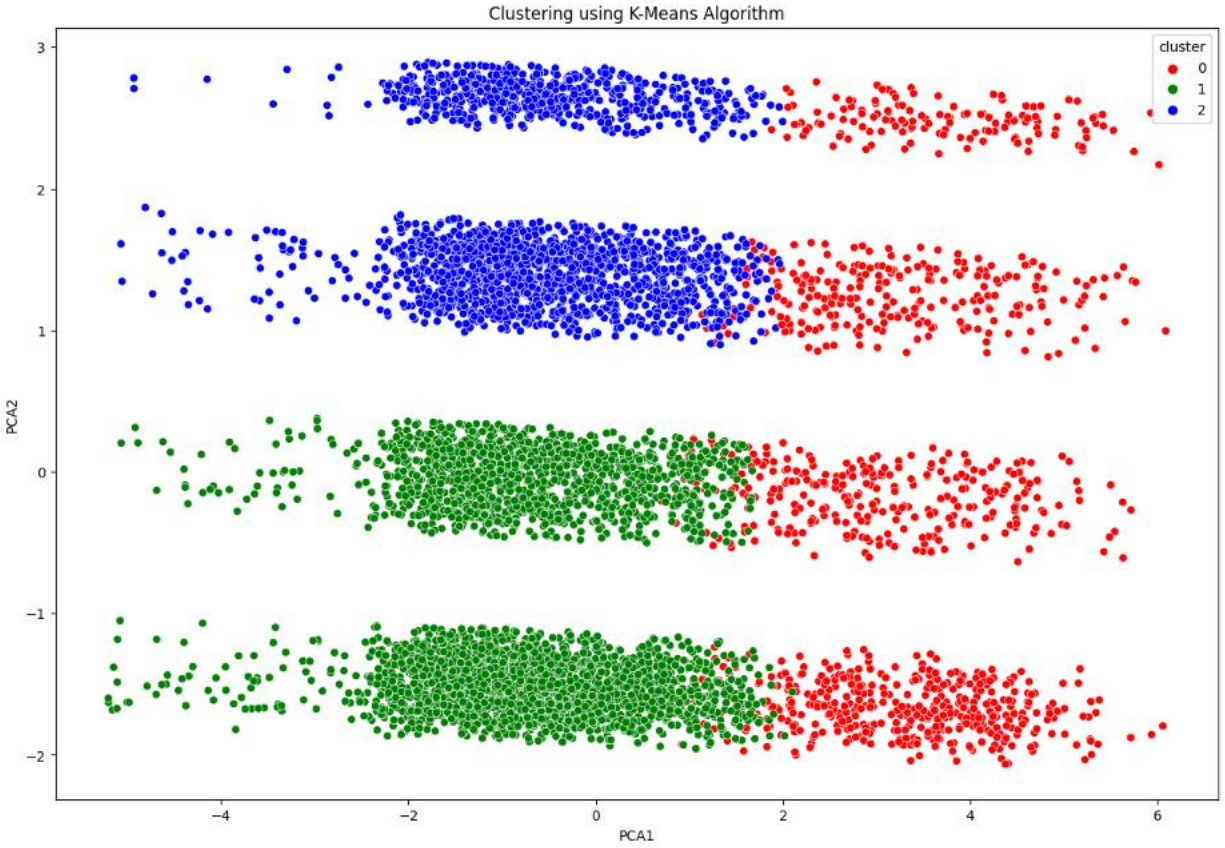
Results for kmeans clustering on 2nd dataset as it can be visualized maximum cluster counts are in the cluster 5 which can be predominant cluster for our market. Total number of segments found are 8.




```
4    1365
2    1189
3     960
1     761
0     420
Name: cluster, dtype: int64
```



Results for kmeans clustering on 2nd dataset as it can be visualized maximum cluster counts are in the cluster 4 which can be predominant cluster for our market. Total number of segments found are 5.



Results for kmeans clustering on 1st dataset as it can be visualized.

	City	Zone	Week Number	Day	4 hour windows	Date	Time stamp date	Demand	Fulfilment %	Avg. Price	Avg. Trip Size	Cancellation % (Total)	Driver Cancellation%	Customer Cancellation %	Avg. ETA
0	0.994648	4.792150	2.460303	4.124888	2.486173	11.855486	3351.446922	1003.583408	69.933095	152.855486	10.063336	38.922658	6.101963	32.824799	20.078947
1	1.602732	4.451073	2.955394	3.850293	2.505436	16.186507	4555.575969	1905.060496	70.486200	155.454976	11.003903	25.728910	4.875216	20.860106	11.497519
2	1.598410	4.424453	1.710239	4.674453	2.498012	4.139662	1208.619781	1894.740060	70.433400	155.739066	11.059145	25.965557	4.907704	21.061680	11.520328

It gives us an clear understanding of how features belong to each cluster.

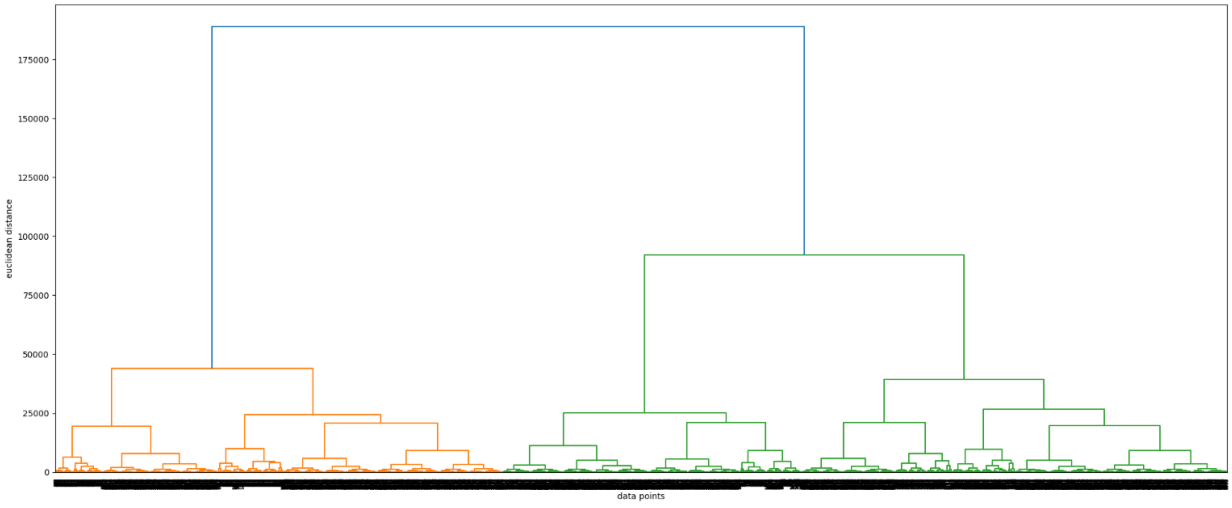
Hierarchical Clustering Algorithm

In hierarchical clustering, the importance or significance of features is not explicitly determined by the algorithm itself. Hierarchical clustering is a distance-based method that focuses on finding the similarity or dissimilarity between samples based on their feature values. However, you can still gain insights into the importance of features indirectly through the clustering results. Here's how you can analyze the importance of features in hierarchical clustering: 1. Calculate feature importance scores: After clustering, you can calculate the feature importance scores using various methods such as: - Feature importance in each cluster: Calculate the mean or median feature values within each cluster and compare them. Features that exhibit larger differences between clusters are likely to contribute more to the clustering result. - Feature variance: Calculate the variance of each feature within each cluster. Features with higher variances are likely to have more discriminative

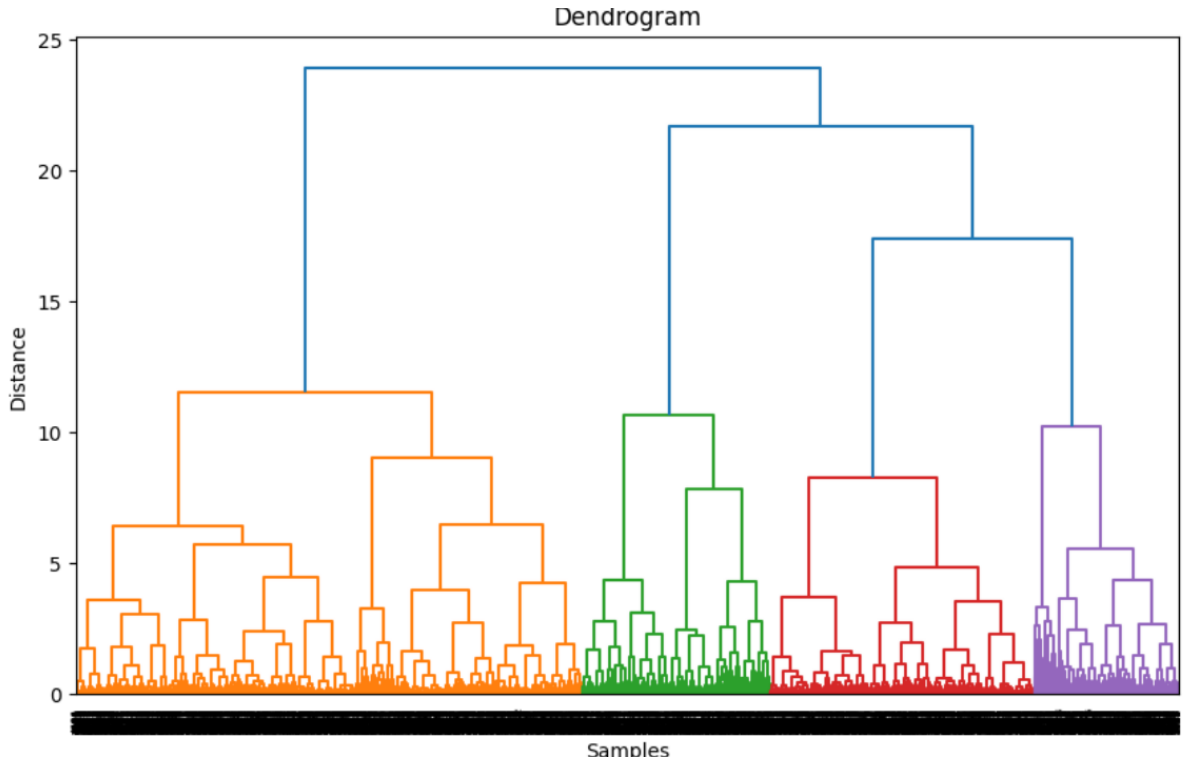
power for clustering. 2. Visualize feature importance: You can create visualizations to understand the importance of features. For example: - Box plots: Create box plots of each feature for each cluster. Features with wider or more separated box plots between clusters indicate higher importance. - Heatmaps: Create a heatmap that displays the average feature values for each cluster. Features that show distinct patterns or differences across clusters are considered important. By performing these analyses, you can gain insights into which features are more influential in driving the clustering results and thus understand the importance of features in the market segmentation obtained from hierarchical clustering.

In a dendrogram, each data point is represented as a leaf node, and the similarity between data points is represented by the height of the branches that join them. The higher the branch, the more dissimilar the data points.

The dendrogram provides insights into the hierarchical structure of the data points, allowing you to identify clusters and their relationships. The vertical height of the branches indicates the dissimilarity between data points or clusters, while the horizontal lines show how the clusters are merged during the hierarchical clustering process.



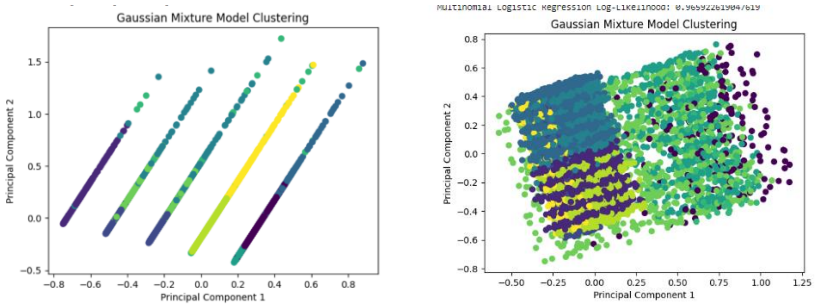
Hierarchical Clustering on dataset 1



Hierarchical Clustering on dataset 2

Gaussian Mixture Model (GMM)

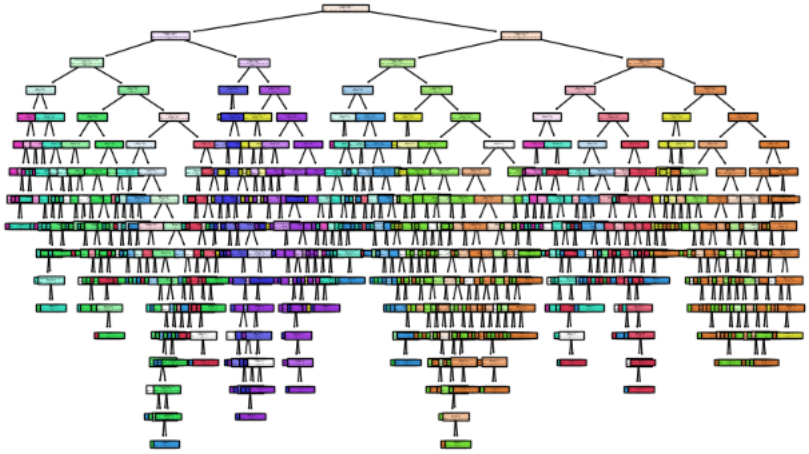
Gaussian Mixture Model (GMM) is a statistical model that assumes the data is generated from a mixture of Gaussian distributions. In the context of market segmentation, GMM can be used to identify underlying groups or clusters wihin a market based on the characteristics of the data.



Gaussian Mixture Model performed on both the datasets

Prediction

Methods Used : Binary Logistic Regression, Multinomial Logistic Regression, Tree Based Method-(Decision Method)

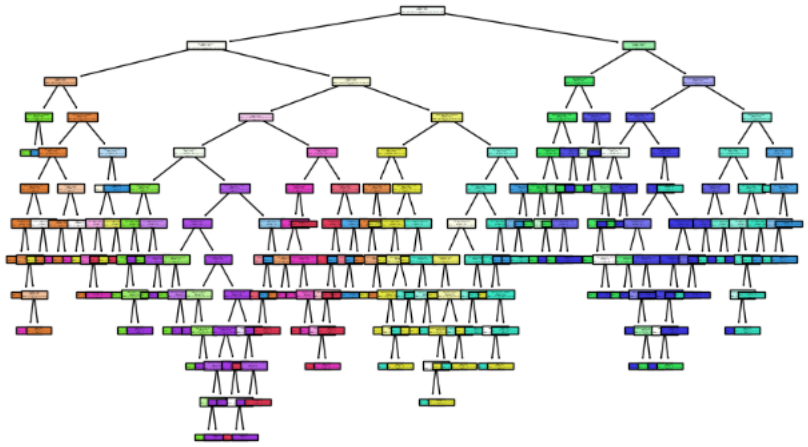


```
Binary Logistic Regression Coefficients:
Demand: 13.164)
Fulfilment %: 0.165)
Avg. Price: 0.141)
Avg. Trip Size: 1.798)
Cancellation % (Total): 0.057)
Driver Cancellation%: 0.022)
Customer Cancellation %: 0.062)
Avg. ETA: 0.975)

Multinomial Logistic Regression Coefficients:
Demand: 17.868, 0.422, 0.355, 2.211, 0.112, 0.120, 0.110, 1.022
Fulfilment %: 8.319, 7.935, 7.879, 4.566, 0.486, 2.417, 1.027, 1.256
Avg. Price: 2.544, 0.325, 6.427, 9.208, 2.008, 1.938, 2.031, 0.139
Avg. Trip Size: 7.894, 8.384, 7.568, 4.413, 0.474, 2.263, 0.971, 1.199
Cancellation % (Total): 6.139, 7.274, 8.721, 4.081, 0.385, 1.850, 0.801, 0.864
Driver Cancellation%: 4.979, 0.223, 11.319, 0.988, 2.786, 0.486, 3.411, 1.253
Customer Cancellation %: 4.129, 0.540, 8.637, 0.931, 2.075, 0.221, 2.389, 3.369
Avg. ETA: 7.785, 0.211, 1.321, 0.658, 0.899, 4.371, 1.881, 0.996

Binary Logistic Regression Log-Likelihood: 1.0
Multinomial Logistic Regression Log-Likelihood: 0.965922619047619
```

Dataset1

The Algorithm used for the prediction is Logistic regression, it is used because of its categorical data handling capability.

Log-Likelihood: 23.576419559532688



Binary Logistic Regression Coefficients:
distance_travelled: 1.767
time_taken: -8.606
commission_base_cost: -0.051
driver_base_cost: 3.368
total_tax: 1.805
total_trip_cost: 2.577
ratings: -14.856

Multinomial Logistic Regression Coefficients:
distance_travelled: -1.160, 4.647, 0.357, -2.477, -1.162, -1.817, 19.198
time_taken: -0.248, -0.599, -0.390, -0.439, -0.443, -0.438, -21.219
commission_base_cost: -5.237, -1.520, -2.855, -3.689, -3.459, -3.579, -10.262
driver_base_cost: 2.215, 3.999, 0.482, 3.383, 2.048, 2.846, 13.811
total_tax: 4.113, 4.007, 0.020, 3.194, 1.703, 2.526, -8.971
total_trip_cost: -4.213, -10.831, -2.568, -2.937, -2.906, -2.918, 15.954
ratings: 0.674, -2.876, 3.013, 3.187, 3.407, 3.024, -2.067

Binary Logistic Regression Log-Likelihood: 0.955959595959596
Multinomial Logistic Regression Log-Likelihood: 0.8909090909090909

Dataset2

The Algorithm used for the prediction is Logistic regression, it is used because of its categorical data handling capability.

Profiling and describing potential segments

Market segmentation based on gender, reason for cab bookings, zones, and cost has revealed interesting insights into potential segments. These factors have emerged as predominant features that significantly influence consumer behavior and preferences in the transportation industry. By examining these variables, we can profile and describe the potential segments to better understand their characteristics and tailor our marketing strategies to meet their specific needs. One segment that stands out is the "Female Commuters" segment. This segment primarily consists of women who frequently use cab services for their daily commute. They value safety, convenience, and reliability in their transportation choices. With safety being a top concern, this segment prefers cab bookings that offer features like female drivers or enhanced security measures. They are likely to prioritize zones with well-lit streets, populated areas, and access to public transportation hubs.

In the transportation industry, market segmentation reveals key segments such as "Female Commuters," who value safety, convenience, and affordability, and "Business Travelers," who prioritize efficiency and comfort. "Leisure Explorers" seek flexibility and access to tourist spots, while "Cost-Conscious Commuters" prioritize affordability without compromising on essential features. Tailoring marketing strategies to these segments allows businesses to effectively engage their target audience.

Understanding and catering to segments like "Female Commuters," "Business Travelers," "Leisure Explorers," and "Cost-Conscious Commuters" allows businesses to tailor marketing strategies. This includes highlighting safety, convenience, and affordability for female commuters, offering specialized services for business travelers, emphasizing popular destinations for leisure explorers, and providing cost-saving options for cost-conscious commuters. Regular monitoring and updates are necessary to adapt to evolving customer preferences and market trends, driving customer satisfaction and business growth.

Selection of Target Segment:

Selecting the appropriate target segment requires careful consideration of business goals, market size, growth potential, and competitive landscape. Based on the analysis, we recommend focusing on the "Female Commuters" segment. With a strong demand for cab services from female customers, targeting this segment presents significant opportunities for growth. By tailoring services to address their specific needs, businesses can establish a reputation for providing safe, reliable, and affordable transportation options.
The selected target segments for the cab booking industry are the "Urban Business Districts" and "Residential Suburbs" zones. For the Urban Business Districts, the focus should be on providing efficient and time-saving transportation solutions, catering to professionals who frequently travel within the district for work. For the Residential Suburbs, affordability, convenience, and reliability are key factors, and offering competitive pricing, personalized promotions, and subscription-based services can attract residents who rely on cab services for daily commuting or occasional trips.

Customizing the Marketing Mix

when targeting the "Female Commuters" segment in the cab booking industry, companies should focus on providing specialized product features such as female drivers, enhanced safety measures, and comfortable interiors. They should offer competitive pricing, transparent pricing structures, occasional discounts, and loyalty rewards to attract and retain customers. Promotional activities should emphasize the commitment to safety, reliability, and customer-centricity, utilizing various channels and collaborations with influencers or organizations promoting women's empowerment. Convenient access to services through strategic placement of pick-up points, partnerships with key locations, and user-friendly digital platforms is essential. Exceptional service, including driver training, customer feedback mechanisms, and responsive customer support, should be prioritized. Continuous monitoring and adaptation of marketing strategies based on customer feedback and market trends are crucial for success.

Conclusion:

In conclusion, our market segmentation analysis has identified distinct customer segments within the cab booking industry. Each segment possesses unique characteristics and preferences that can guide strategic marketing efforts. By targeting the "Female Commuters" segment, businesses can capitalize on the demand for safe and affordable transportation options. However, the final decision on the target segment should consider the company's capabilities, resources, and market dynamics. By aligning the business strategy with the selected segment's characteristics and preferences, companies can maximize their chances of success in the competitive cab booking market.

Link to github profile with codes and datasets well documented.

https://github.com/sriharinigunda/Online-Vehical-Bookings-Market-Segmentation/blob/main/Online%20Vehical%20Booking%20.ipynb

https://colab.research.google.com/drive/1KxEQBjYZBXNSjmGLg7ewtag4Wzs7bx5g?usp=sharing#scrollTo=yY16RD82EQAM

https://github.com/MYSTIC-HUNTER/Online-Vehicle-Booking

https://github.com/archanarajeshwar/market_segmentation_online_vehicle_booking