ONLINE VEHICLE BOOKING MARKET

Shubhang Sharma

Vishnu Sai Teja Nagabandi

Shubham Mojidar

INTRODUCTION

The online vehicle booking market refers to the industry where consumers can book various types of vehicles, such as cars, motorcycles, buses, and taxis, online through websites or mobile applications. This market has grown rapidly in recent years due to the increased availability of internet access and the convenience it offers.

The online vehicle booking market is highly competitive, with numerous companies offering their services. Some of the largest players in the market include Uber, Lyft, and Ola. These companies have disrupted the traditional taxi and car rental industries by offering a more convenient and affordable alternative to traditional transportation.

Consumers can book vehicles for various purposes, such as airport transfers, city tours, business trips, and personal travel. Online vehicle booking platforms typically offer a variety of vehicle types to choose from, including economy, standard, luxury, and premium cars, as well as motorcycles, buses, and other types of vehicles.

To book a vehicle online, customers typically need to provide their pickup and drop-off locations, travel dates, and other relevant information. The booking platform then searches for available vehicles that meet the customer's requirements and provides them with a quote. Customers can compare prices and select the vehicle that best suits their needs and budget.

The online vehicle booking market has revolutionized the way people travel, providing a convenient and affordable alternative to traditional transportation. It is expected to continue to grow in the coming years as more people turn to online booking platforms for their transportation needs.

MARKET SEGMENTATION

Market segmentation is the process of dividing a larger market into smaller subgroups of consumers with similar needs, characteristics, or behaviors. The goal of market segmentation is to identify specific groups of customers that a business can target with tailored marketing messages and product offerings.

In the context of the online vehicle booking market, market segmentation involves identifying different segments of customers based on various factors such as geographic, demographic, and behavioral. Each of these segments has its unique characteristics and needs and targeting them requires a different approach.

Geographic segmentation divides a market into smaller subgroups based on location, allowing businesses to tailor their marketing efforts to specific customer needs and preferences. It can be useful for businesses that operate in different regions or countries and want to target customers with location-specific needs.

Demographic segmentation is a powerful marketing strategy that divides a market based on demographic characteristics such as age, gender, income, education, occupation, marital status, and family size. This helps businesses create targeted marketing campaigns that appeal to specific customer groups and understand overall trends and patterns in the market.

Behavioral segmentation is a powerful marketing strategy that divides a market based on customers' behaviors, attitudes, and preferences towards a product or service. It helps businesses understand their customers on a deeper level and create targeted marketing campaigns that increase customer engagement and loyalty. Behavioral segmentation can be based on factors such as occasion, benefits sought, loyalty status, user status, and readiness to buy.

In summary, market segmentation is a crucial tool for any online vehicle booking startup looking to enter the Indian market. By identifying the right customer segments based on geographic, demographic, and behavioral factors, businesses can tailor their marketing efforts to meet the needs of specific customer groups and ultimately succeed in this rapidly evolving and highly competitive market.

MARKET ANALYSIS

Market analysis for an online vehicle booking startup involves examining the market trends, identifying the target customers, analyzing the competition, and evaluating the overall industry landscape. Here are some key factors to consider:

1. **Market size and growth potential:** According to a report by Allied Market Research, the global car rental market size was valued at $92.92 billion in 2019 and is expected to reach $214.04 billion by 2027, growing at a CAGR of 10.7% from 2020 to 2027. This indicates significant growth potential for the industry.
2. **Target customers:** The target customers for an online vehicle booking startup are individuals and businesses that require transportation services. The startup could focus on specific customer segments such as budget-conscious travelers, luxury vehicle enthusiasts, or corporate clients. The graph could represent the percentage of each customer segment in the overall market.
3. **Competition:** There are many established players in the online vehicle booking market such as Uber, Lyft, and Turo. The graph could represent the market share of each competitor and the startup in the overall market.
4. **Customer satisfaction:** Customer satisfaction is critical for the success of an online vehicle booking startup. The graph could represent the satisfaction levels of customers using different online vehicle booking platforms, highlighting the areas where the startup can differentiate itself and improve its services.

STEPS FOR ANALYSIS

STEP 1: IMPORTING LIBRARIES

The first step in any data analysis project is to import the necessary libraries that will be used to process and analyze the data. In this project, several libraries were imported including pandas, NumPy, seaborn, matplotlib, sklearn and warnings.

1. **Pandas:** Pandas is a library for data manipulation and analysis. It provides a data structure called a DataFrame, which is similar to a table in a relational database. Pandas is commonly used for data cleaning, data wrangling, and data exploration.
2. **NumPy:** NumPy is a library for numerical computing in Python. It provides a powerful array data structure and functions for array manipulation, linear algebra, and statistical analysis. NumPy is commonly used in scientific computing and data analysis.
3. **Seaborn:** Seaborn is a visualization library built on top of matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is commonly used for data visualization and exploration.
4. **Matplotlib:** Matplotlib is a library for creating static, animated, and interactive visualizations in Python. It provides a low-level interface for creating basic plots, such as line plots, scatter plots, and histograms. Matplotlib is widely used in scientific computing and data analysis.
5. **Scikit-learn (sklearn):** Scikit-learn is a library for machine learning in Python. It provides a wide range of machine learning algorithms for classification, regression, clustering, and dimensionality reduction. Scikit-learn is commonly used in data science and machine learning projects.

6. **Warnings:** Warnings is a module in Python used to handle warning messages. It provides a way to control how warning messages are displayed and allows for the handling of warning messages as exceptions. Warnings are commonly used in data analysis and machine learning to handle deprecated functions or unexpected behavior.

```
# Import the useful libraries.

import pandas as pd, numpy as np
import datetime as dt
import matplotlib.pyplot as plt, seaborn as sns
%matplotlib inline

from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
```

STEP 2: LOADING DATASETS

The next step after importing libraries is to load the dataset for further processing and model building. We have use different datasets including the following, every dataset used are in the CSV format:

1. **Uber Request Data** – Uber is facing driver cancellation and non-availability of cars leading to loss of potential revenue. It consists of variables like Request id, Pickup point, Driver id, Status, Request timestamp, Drop timestamp.

```
# Read the data set of "Uber Request Data" in uber_df.

uber_df = pd.read_csv("Uber Request Data.csv")
uber_df
```

	Request id	Pickup point	Driver id	Status	Request timestamp	Drop timestamp
0	619	Airport	1.0	Trip Completed	11/7/2016 11:51	11/7/2016 13:00
1	867	Airport	1.0	Trip Completed	11/7/2016 17:57	11/7/2016 18:47
2	1807	City	1.0	Trip Completed	12/7/2016 9:17	12/7/2016 9:58
3	2532	Airport	1.0	Trip Completed	12/7/2016 21:08	12/7/2016 22:03
4	3112	City	1.0	Trip Completed	13-07-2016 08:33:16	13-07-2016 09:25:47
...	...	...	...	...	...	...
6740	6745	City	NaN	No Cars Available	15-07-2016 23:49:03	NaN
6741	6752	Airport	NaN	No Cars Available	15-07-2016 23:50:05	NaN
6742	6751	City	NaN	No Cars Available	15-07-2016 23:52:06	NaN
6743	6754	City	NaN	No Cars Available	15-07-2016 23:54:39	NaN
6744	6753	Airport	NaN	No Cars Available	15-07-2016 23:55:03	NaN

6745 rows × 6 columns

```
# Checking the number of rows and columns in the dataframe

uber_df.shape
```

```
(6745, 6)
```

2. **New York City Taxi Trip Duration** - The competition dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this playground competition. Based on individual trip attributes, participants should predict the duration of each trip in the test set.

3. **Uber drives -** The dataset contains Start Date, End Date, Start Location, End Location, Miles Driven and Purpose of drive (Business, Personal, Meals, Errands, Meetings, Customer Support etc.)

```
dframe = pd.read_csv('data.csv')
```

```
dframe.shape
```

```
4]: (1156, 7)
```

```
dframe.head()
```

5]:

	START_DATE*	END_DATE*	CATEGORY*	START*	STOP*	MILES*	PURPOSE*
0	1/1/2016 21:11	1/1/2016 21:17	Business	Fort Pierce	Fort Pierce	5.1	Meal/Entertain
1	1/2/2016 1:25	1/2/2016 1:37	Business	Fort Pierce	Fort Pierce	5.0	NaN
2	1/2/2016 20:25	1/2/2016 20:38	Business	Fort Pierce	Fort Pierce	4.8	Errand/Supplies
3	1/5/2016 17:31	1/5/2016 17:45	Business	Fort Pierce	Fort Pierce	4.7	Meeting
4	1/6/2016 14:42	1/6/2016 15:49	Business	Fort Pierce	West Palm Beach	63.7	Customer Visit

After loading the dataset various pre-analysis and cleaning of data was done including treating the missing values, comping up with relations between the variables of the dataset etc.

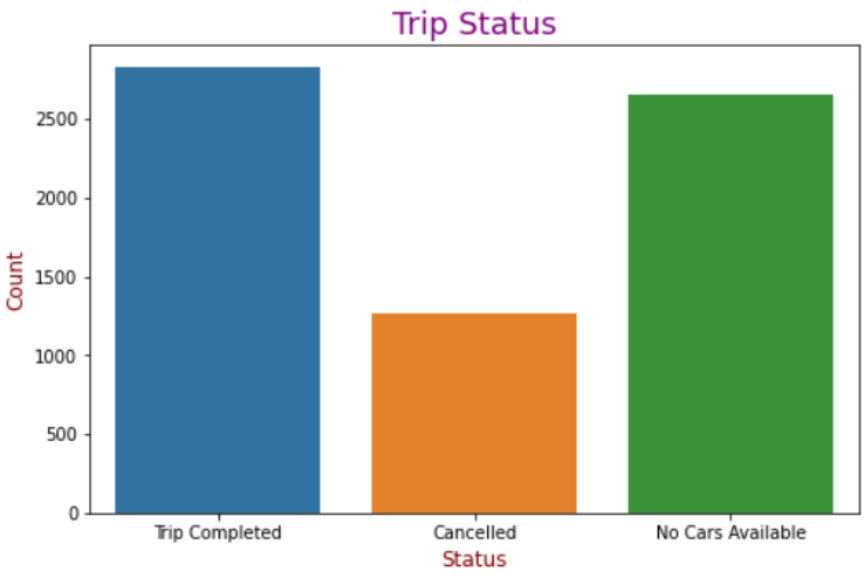
STEP 3: VISUALIZING THE DATA AND EDA

The next step after pre-processing the data is to visualize it and come up with the better approach to start with. Visualizing of the data helps to determine the important and relevant variables to work with so as to ensure better performance from the model and generating the most accurate and promising solutions for the problem statement.

In this step various graphs were plotted including count plot, factor plot, histogram and scatter plot.

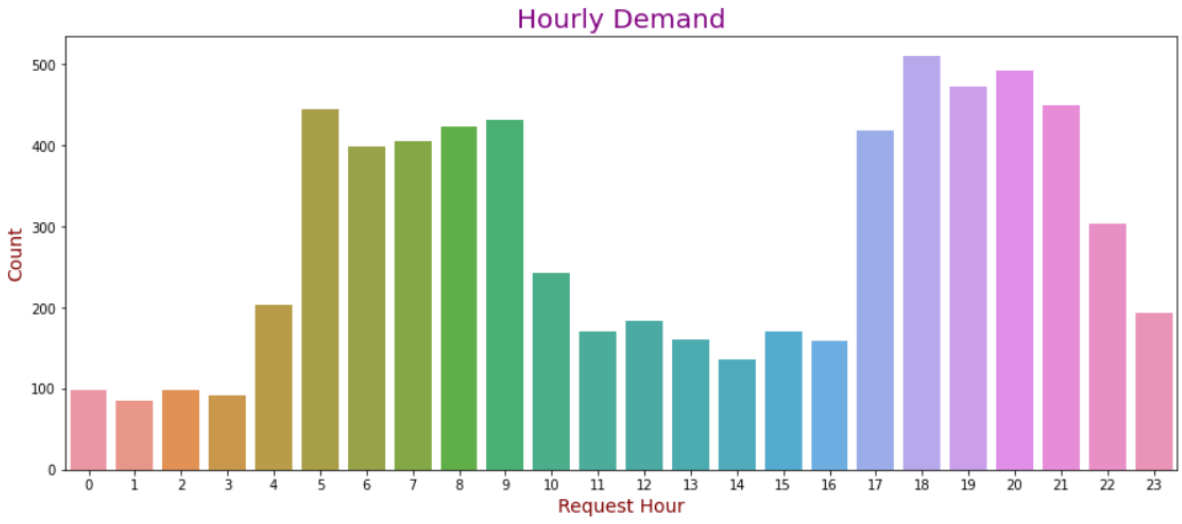
1. **Count plot:** Count plot is a type of plot used to visualize the frequency distribution of categorical data. It is a bar plot where the height of each bar represents the number of occurrences of a particular category. In seaborn, count plot is created using the countplot() function.
2. **Factor plot:** Factor plot is a type of plot used to visualize the distribution of a numerical variable across different levels of one or more categorical variables. It is a versatile plot that can be used to create different types of plots such as bar plot, box plot, violin plot, etc. In seaborn, factor plot is created using the factorplot() function.
3. **Scatter plot:** A scatter plot is a type of plot used to visualize the relationship between two numerical variables. Each point on the plot represents a pair of values for the two variables. Scatter plots are useful for identifying patterns or trends in the data, as well as for detecting outliers or anomalies. In Python, scatter plots can be created using the scatter() function in matplotlib or using the scatterplot() function in seaborn.
4. **Histogram:** A histogram is a type of plot used to visualize the distribution of a numerical variable. The variable is divided into a set of intervals called bins, and the height of each bar represents the frequency or count of data points that fall into that bin. Histograms are useful for identifying the shape of the distribution, such as whether it is symmetric or

skewed, as well as for detecting outliers or anomalies. In Python, histograms can be created using the hist() function in matplotlib or using the histplot() function in seaborn.



INFERENCES:

- Total Demand = 6745
- Total Supply = 2831
- Supply-Demand Gap = 3914
- This shows only 42% of total demand was met there is gap of 58% of supply due to trip cancellation or cabs availability.

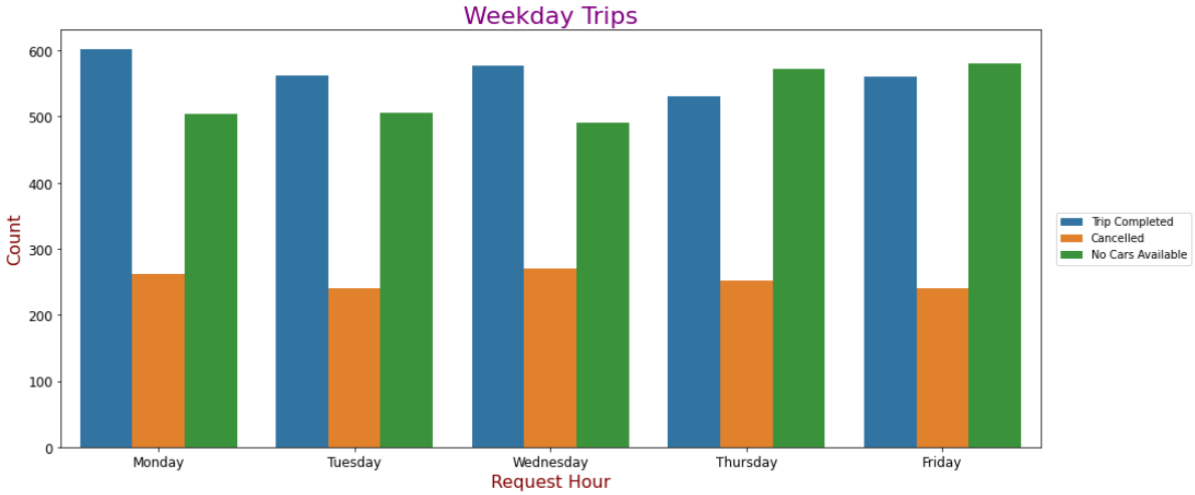


INFERENCES:

- Overall peek-hours of demand for Uber Cab ride are higher in morning is between 4:00 to 10:00 and at night is between 17:00 to 22:00.

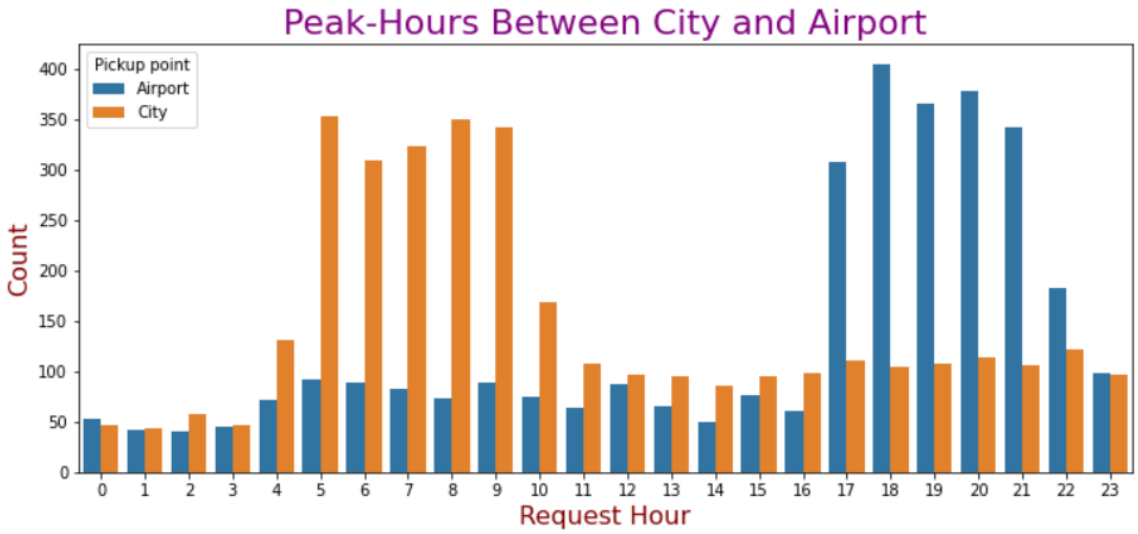
INFERENCES:

- In City, demand is greater than supply during early morning and in mid-morning (4am to 10am).
- In Airport, demand is higher than supply during evening and at night (5pm to 12am).



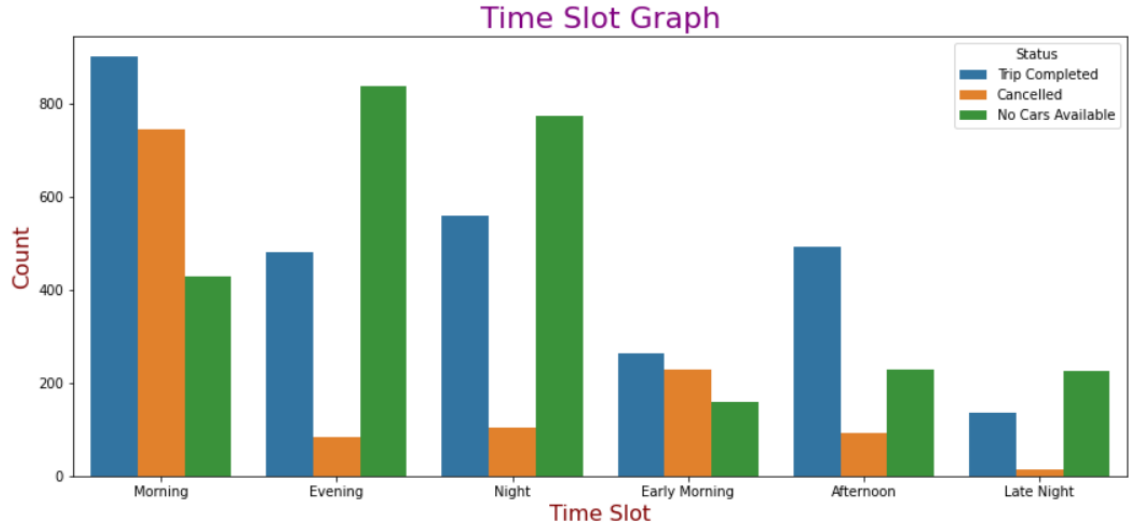
INFERENCES:

- From the above graph we can see that:
  1. Number of trips completed is higher on Monday and least on Thursday.
  2. Number of trips cancelled is higher on Wednesday.
  3. Number of no-cars availability is higher on Thursday and Friday and slightly lower on other days



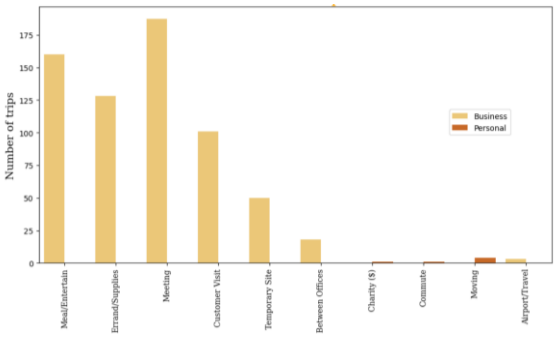
INFERENCES:

- For the trips from City to Airport between 05:00-10:10 in the morning demand is higher while demand for the trips from Airport to City between 17:00-22:00 at night is higher.



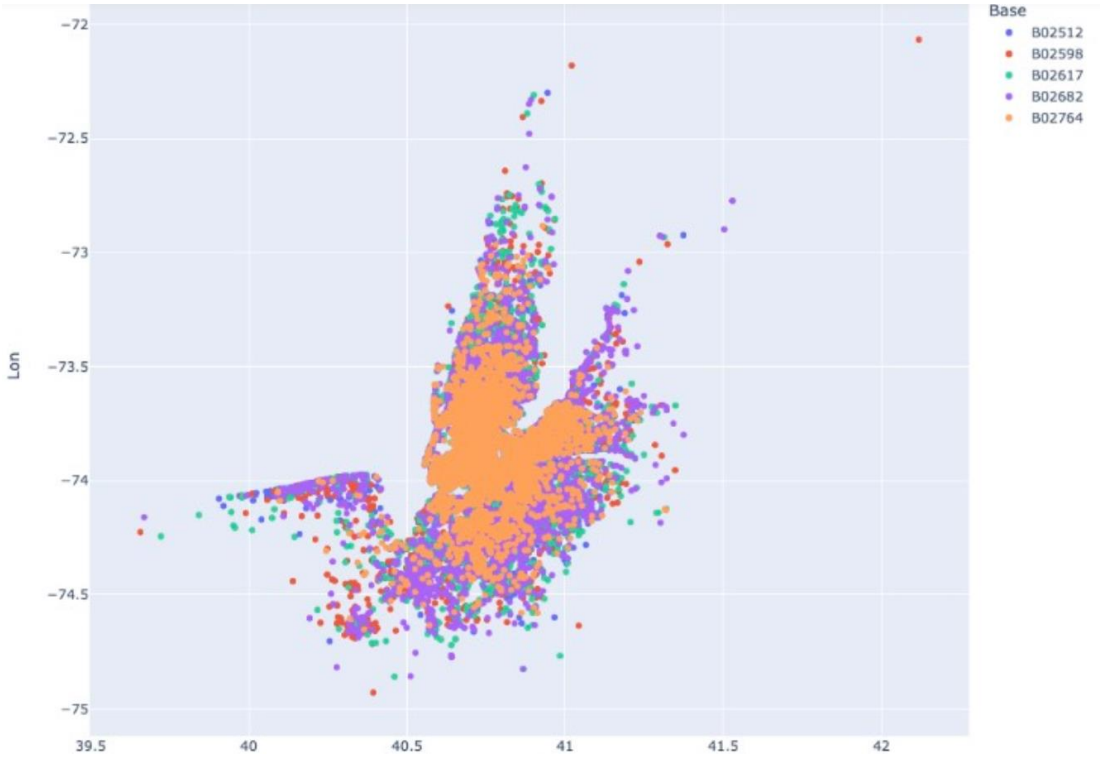
INFERENCES:

- As we can see there is more rush in morning and evening slot as comparison to other time slots.
- And we can also observe that there is high demand in evening and at night and Uber is not able to meet up the demand as the no cars availability bar is higher in both slots which shows there is lack of supply.



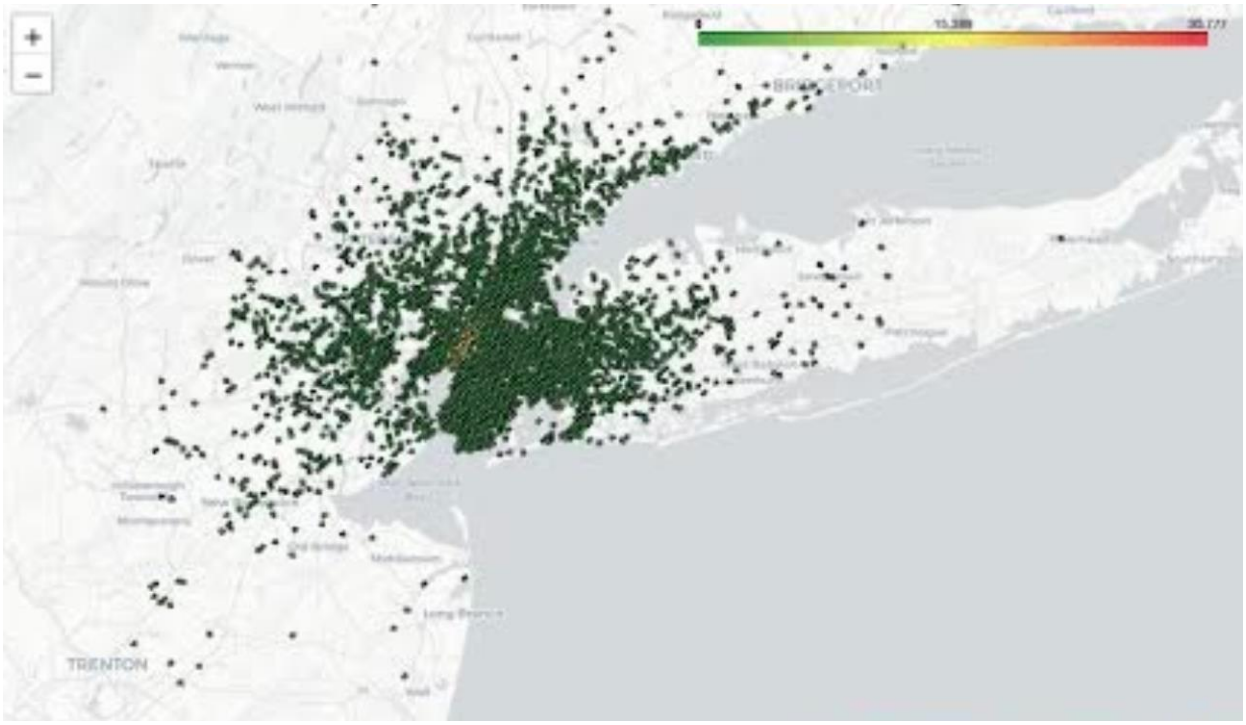
INFERENCES:

- Total number of Small Distance passengers are: 288
- Total number of Medium Distance passengers are: 288
- Total number of Large Distance passengers are: 274
- Total number of Super Large Distance passengers are: 305
- The cities with most of the pickups are:
  1. Cary
  2. Mohrsville
  3. Whitebridge
  4. Islamabad
  5. Durham
- The same is also true for the greatest number of drops so it makes sense on calling them the busiest cities judging by the data available



INFERENCES:

- The preferred method for visualizing data when the size of the geographical region is unimportant is a hex map.
- To find patterns and trends in the data, analyses the hex map. For instance, you might see that prices are higher in some regions or those with significant traffic, or that prices tend to be lower at particular times of the day.
- Also, you can examine taxi rates by examining various geographical regions. You may, for instance, compare prices between various neighborhoods or between downtown and the suburbs. You can learn from this how taxi businesses base their pricing on the locations of their clients.
- Examining how taxi rates alter depending on different geographic areas is part of the geographic study of taxi fares. As it offers information on how taxi firms decide their prices based on the location of their customers, this can be helpful for both taxi companies and customers.



INFERENCES:

- Geography Factors Affecting:

    1. **Distance:** A significant factor influencing taxi prices is travel distance. On longer travels or trips that require the cab to leave the city borders, for instance, prices could be higher.
    2. **Location:** The pick-up and drop-off locations' locations can have an impact on taxi prices. For instance, prices may be greater for travel to or from inconvenient locations, such rural or suburban districts.
    3. **Competition:** Taxi prices may vary depending on the degree of competition in a certain area. The increased competition in locations with several cab providers may result in cheaper prices.
    4. **Service type:** The kind of service that taxi firms provide might have an impact on prices. For services that offer more facilities or a higher caliber of service, such as luxury or premium services, the cost may be greater.

- Taxi firms can analyse these variables to change their pricing methods to remain competitive and draw consumers, and customers can select the most affordable alternative for their needs. A consumer might decide to choose a taxi service, for instance, that charges less for rides to or from particular locations or that experiences more competition in their neighborhood to benefit from reduced rates. Taxi companies, on the other hand, may adjust their fares based on demand and competition in different geographic areas in order to maximize their revenue.

STEP 4: DATA PREPROCESSING

Data preprocessing is the process of cleaning, transforming, and organizing raw data into a more usable format for analysis. It is a crucial step in data analysis, as the accuracy and effectiveness of any model or analysis is highly dependent on the quality of the data.

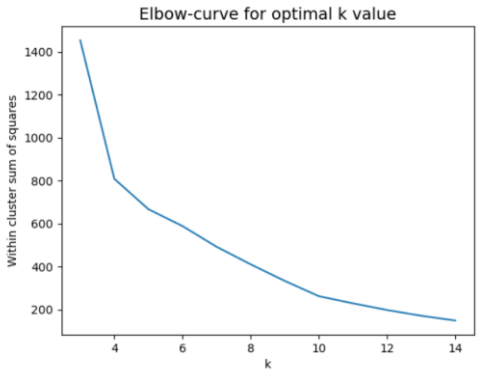
The following are the steps involved in data preprocessing:

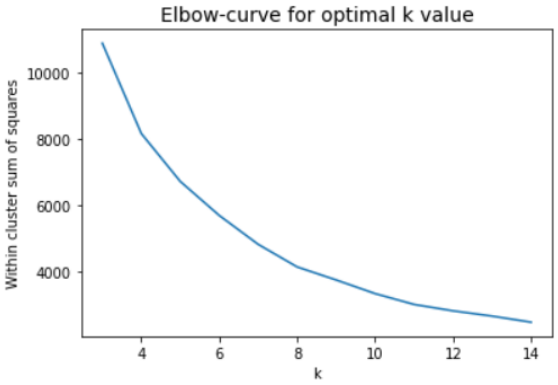
1. **Data Cleaning:** This step involves removing any irrelevant, incomplete, or inconsistent data from the dataset. This may involve removing duplicates, handling missing values, and correcting inconsistent data. But there were some columns which doesn't required to be dropped such as the 'Driver id'.
2. **Data Formatting:** This step involves formatting the data into a specific format for analysis. This may include converting categorical data to numerical data, or converting the data into a specific data structure for analysis. For instances, k-means clustering doesn't work with categorical variables so all categorical variables are transformed/formatted to their respective integer dummy variables.
3. **Data Transformation:** This step involves converting the data into a more useful format. This may include data normalization, feature scaling, or data discretization. Data normalization involves scaling the data to a specific range, while feature scaling involves scaling individual features to improve model performance. After formatting the variables, they are required to be scaled down to the same scale for better visualization and performance of the model.
4. **Data Integration:** This step involves combining data from different sources into a single dataset. This may involve merging data from different files or databases, or combining data from different data types.

STEP 5: OPTIMAL NUMBERS OF CLUSTER

Determining the optimal number of clusters in k-means is an important step in the clustering process as it can significantly impact the results of the analysis. There are several methods that can be used to determine the optimal number of clusters:

1. **Elbow method:** This method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters. The WCSS measures the sum of squared distances between each data point and its assigned centroid. The optimal number of clusters is the point at which the rate of decrease in WCSS slows down, forming an elbow-like shape in the plot.





INFERENCES:

1. **Graph I**
    - As the Elbow-curve is coming out to be smooth and making it difficult to come up with an optimal number of clusters. It can either 4 or 5 clusters
    - But when analyzed closely, we can consider 5 clusters to be optimal for our K-Means model.
    - For better clarification let's consider Silhouette Analysis too.

2. **Graph II**
    - The steep points of the Elbow Curve represent the optimal number of clusters be considered.
    - We can consider 5 Points in this case as the optimal number of clusters for the dataframe.

2. **Silhouette method:** This method involves computing the average silhouette width for each number of clusters. The silhouette width measures the similarity of data points within a cluster compared to points in other clusters. The optimal number of clusters is the point at which the average silhouette width is maximized.

```
For n_clusters=3, the silhouette score is 0.32659150403378706
For n_clusters=4, the silhouette score is 0.3379859613441783
For n_clusters=5, the silhouette score is 0.34653699707280783
For n_clusters=6, the silhouette score is 0.37493495216082573
For n_clusters=7, the silhouette score is 0.3808224669568404
For n_clusters=8, the silhouette score is 0.38919678832137117
For n_clusters=9, the silhouette score is 0.41421878348963087
For n_clusters=10, the silhouette score is 0.4088027113455968
For n_clusters=11, the silhouette score is 0.4255280435867989
For n_clusters=12, the silhouette score is 0.42127621467398674
For n_clusters=13, the silhouette score is 0.42859713842263036
For n_clusters=14, the silhouette score is 0.4354170727358296
```

INFERENCES:

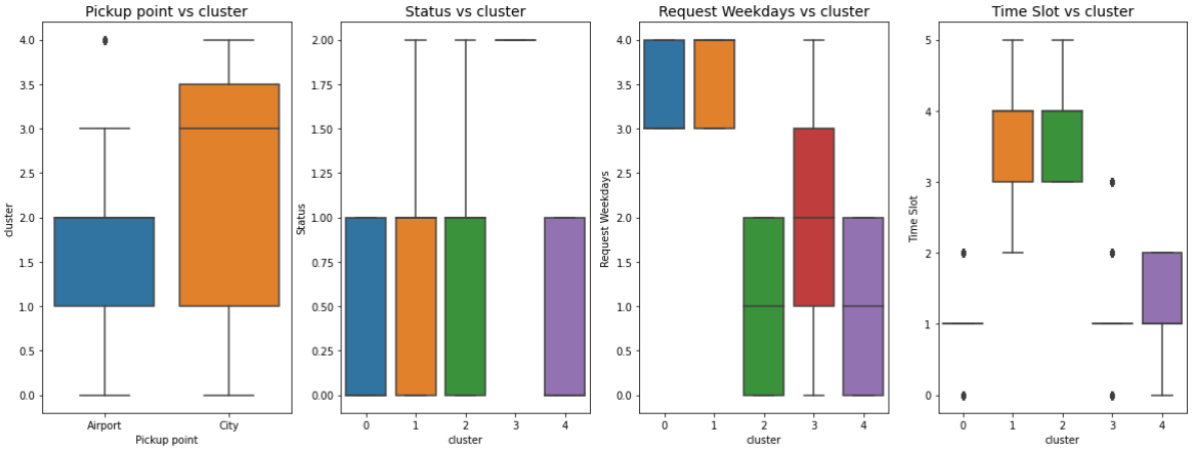
- As we can see, the silhouette value of 5 clusters is about 0.346. But Silhouette analysis says higher the value of the cluster, better is the model performance and is the optimal number of clusters.

- From Silhouette analysis, 14 clusters give us the highest score but in real world it is very difficult to analysis 14 clusters over a dataset having 6000+ data points.
- So, from Elbow-curve and Silhouette analysis we have 5 clusters for model building.

STEP 6: K-MEANS CLUSTERING

K-means is a popular clustering algorithm that partitions a dataset into k clusters based on their similarity. It is an iterative algorithm that works by first randomly selecting k points as the initial cluster centroids, and then assigning each data point to the nearest centroid. The algorithm then computes the mean of the points assigned to each centroid and updates the centroid to this mean. This process is repeated until the cluster centroids no longer change or a maximum number of iterations is reached.

To use k-means, one typically needs to specify the number of clusters, k, that the algorithm should create. One common approach to determining k is to use the elbow method, which involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and selecting the k at which the rate of decrease in WCSS slows down (i.e., forming an elbow shape in the plot).



CONCLUSIONS

Conclusions from Behavioral-Segmentation:

1. From the clusters we can interpret that most of the people book cabs in the city rather near the airport.
2. As shown from the boxplot there are very few people who are frequently booking cabs from the airport. 14.6% of the people are the ones who are frequently booking cabs from the airports.
3. Time Slot vs cluster graph depicts the following: -
    - From cluster 0, 19.3% and 20.3% cabs are more frequently booked in the Early morning and Afternoon respectively.
    - From cluster 3, 20.4%, 8.2% and 4.4% cabs are more frequently booked in the Early morning, Afternoon and Evening respectively.
4. The people who are frequently booking cabs near airport are the one either who are arriving or departing from the airport. The time during which the bookings are happening is either during evening or at night (5pm to 12am).
5. The people who are booking Early Morning (4am to 6 am) are the ones who are traveling to their work place. Also due to heavy bookings, there is shortage of cabs leading to high cancelling of trips by the drivers throughout all weekdays. During evening and night (5pm to 12am), same scenario repeats and in return no cabs are available.

Conclusions from Geographic-Segmentation:

1. **Consumer demographics:** Taxi businesses can learn about the characteristics of their clients by looking at where they pick up and drop off their passengers. For instance, they can determine whether neighborhoods or places are more well-liked by particular age groups or socioeconomic groups.
2. **Peak hours:** Geography segmentation can also be used to pinpoint the hours when demand for taxis is at its highest. During times of high demand, this information can be used to enhance taxi availability and pricing.
3. **Competition analysis:** Taxi businesses can also use geographic segmentation to study local competition. For instance, they can determine which areas are better supplied by rival businesses and modify their price or marketing plans appropriately.
4. **Traffic patterns:** Taxi firms can learn about traffic patterns in various places by examining taxi routes and pick-up/drop-off points. To reduce trip time and raise customer happiness, they can use this to improve their routing and dispatching algorithms.
5. **Popular destination:** Taxi companies can pinpoint popular sites in various areas by looking at pickup and drop-off locations. This data can be used to focus marketing efforts or change prices to encourage travel to such places.

Conclusions from Demographic-Segmentation:

1. The factors that are being considered for the demographic analysis are the purpose of the trip, category to which it belongs to (Business / Personal) and finally the class cation of the trip based on the distance travelled by the passenger.
2. Most of the business trips that are being booked are majorly for outstation travel, Although the trips are intercity the traffic between the start and stop locations are equally distributed with almost negligible differences.
3. The frequency of booking is highest in Cary, Mohrsville, Whitebridge, Islamabad, Durham
4. The Distances travelled are classified into 4 categories at it makes sense too to have almost equal number of trips in all of them as the categories 'Medium Distance', 'Super Large Distance', 'Large Distance', 'Small Distance' are made on the basis of inter-Quartile ranges of distance travelled.
5. Based on the three features that are being considered for the demographic segmentation analysis the optimum number of clusters was fund to be 5 ,with the K-means clustering.

Link to the GitHub of Online Vehicle Booking Market Segmentation:
https://github.com/Shubhang7100/Online-Vehicle-Booking-Market