ONLINE CAB BOOKING MARKET SEGMENTATION

- Abhishek A

Problem statement:

Analyse the Vehicle Booking market in India using Segmentation analysis and come up with a feasible strategy to enter the market, targeting the segments where there can be possible profit by offering Vehicle booking service.

Abstract:

In the era of digitalization and e-commerce, understanding the diverse and ever-evolving online consumer base is paramount for businesses seeking to thrive in the virtual marketplace. This project presents a comprehensive analysis of online market segmentation, focusing on identifying the primary target demographic for a specific product or service.

The study employs a data-driven approach, leveraging various data sources, including online behavior, demographics, and purchase history, to create a holistic dataset. This data is then subjected to advanced clustering techniques to reveal distinct consumer segments within the online market.

Through rigorous analysis and clustering methods, the project has successfully uncovered key insights into the target demographic for the chosen product or service. Notably, it is revealed that the primary target audience comprises individuals aged 18 to 30 years old. This age group demonstrates a strong affinity for the product or service under investigation, showcasing both a significant online presence and a propensity for making purchases within this category.

The findings of this research have far-reaching implications for marketing strategies, product development, and overall business planning. Armed with a clear understanding of the target demographic, businesses can tailor their marketing campaigns, messaging, and product offerings to effectively engage and cater to the identified consumer group.

In conclusion, this project underscores the importance of data-driven market segmentation in the online sphere and provides a valuable roadmap for businesses aiming to thrive in the digital landscape. By honing in on the 18 to 30-year-old demographic, organizations can enhance their online market presence and achieve more targeted and impactful results.

Data Collection:

Data was downloaded from kaggle.
1. Uber dataset
2. Complete customer and trip dataset.

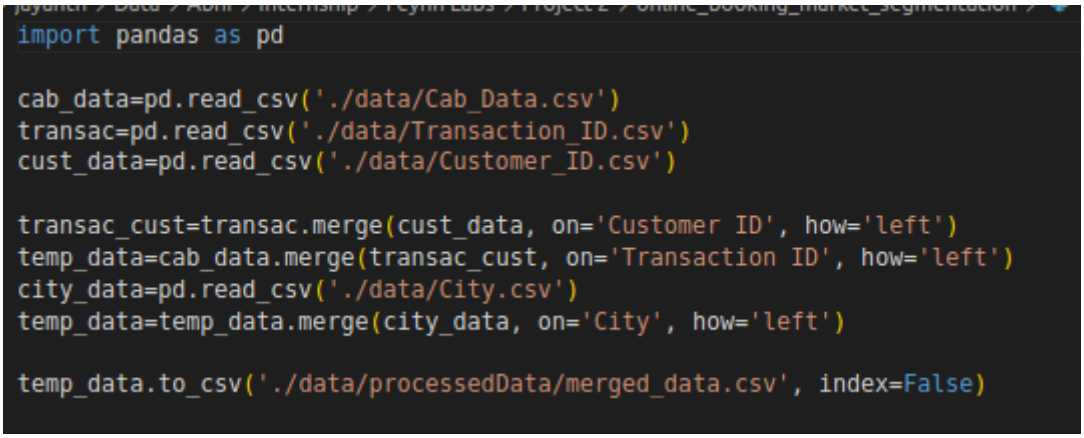
Data Preprocessing:

1. It imports the pandas library and reads three CSV (Comma Separated Values) files: Cab_Data.csv, Transaction_ID.csv, and Customer_ID.csv.

2. It merges the Transaction_ID.csv and Customer_ID.csv data based on a common column called 'Customer ID'. This step essentially combines information about transactions and customer details into one dataset.

3. Next, it merges this combined dataset with another CSV file called City.csv, based on a common column called 'City'. This step adds information about the city where each transaction occurred.

4. Finally, the code saves the merged and processed data into a new CSV file named merged_data.csv located in a folder called processedData.



```
import pandas as pd

cab_data=pd.read_csv('./data/Cab_Data.csv')
transac=pd.read_csv('./data/Transaction_ID.csv')
cust_data=pd.read_csv('./data/Customer_ID.csv')

transac_cust=transac.merge(cust_data, on='Customer ID', how='left')
temp_data=cab_data.merge(transac_cust, on='Transaction ID', how='left')
city_data=pd.read_csv('./data/City.csv')
temp_data=temp_data.merge(city_data, on='City', how='left')

temp_data.to_csv('./data/processedData/merged_data.csv', index=False)
```

Fig 1

It imports the pandas library and the LabelEncoder module from scikit-learn, a popular machine learning library.

It reads a CSV file named "merged_data.csv" located in the "./data/processedData/" directory into a DataFrame called data.

It uses the LabelEncoder to convert categorical variables ('Company', 'City', 'Payment_Mode', 'Gender') into numerical representations. This is necessary because many machine learning algorithms require numerical inputs.

It drops some columns ('Transaction ID', 'Customer ID', 'Population', 'Users') from the dataset as they may not be needed for the analysis or are not relevant.

It extracts year, month, and date information from the 'Date of Travel' column and creates three new columns ('dates', 'months', 'years') to store this information.

It drops the original 'Date of Travel' column as it has been replaced by the new 'dates', 'months', and 'years' columns.

```python
import pandas as pd
from sklearn.preprocessing import LabelEncoder

data = pd.read_csv("./data/processedData/merged_data.csv")

le = LabelEncoder()
data['Company'] = le.fit_transform(data['Company'])
data['City'] = le.fit_transform(data['City'])
data["Payment_Mode"] = le.fit_transform(data['Payment_Mode'])
data['Gender'] = le.fit_transform(data['Gender'])

data.drop(['Transaction ID', "Customer ID", 'Population', 'Users'], axis = 1, inplace = True)

years = []
months = []
dates = []
for i in range(len(data)):
    date, month, year = data['Date of Travel'][i].split('-')
    years.append(int(year))
    months.append(int(month))
    dates.append(int(date))

data['dates'] = dates
data['months'] = months

data.drop(['Date of Travel'], axis = 1, inplace = True)
```

Fig 2

Data Analysis:

Based on the graph, it's evident that the Yellow Cab company has a larger number of customers compared to the Pink Cab company. In simpler terms, more people choose to use Yellow Cabs than Pink Cabs based on the data shown in the graph.

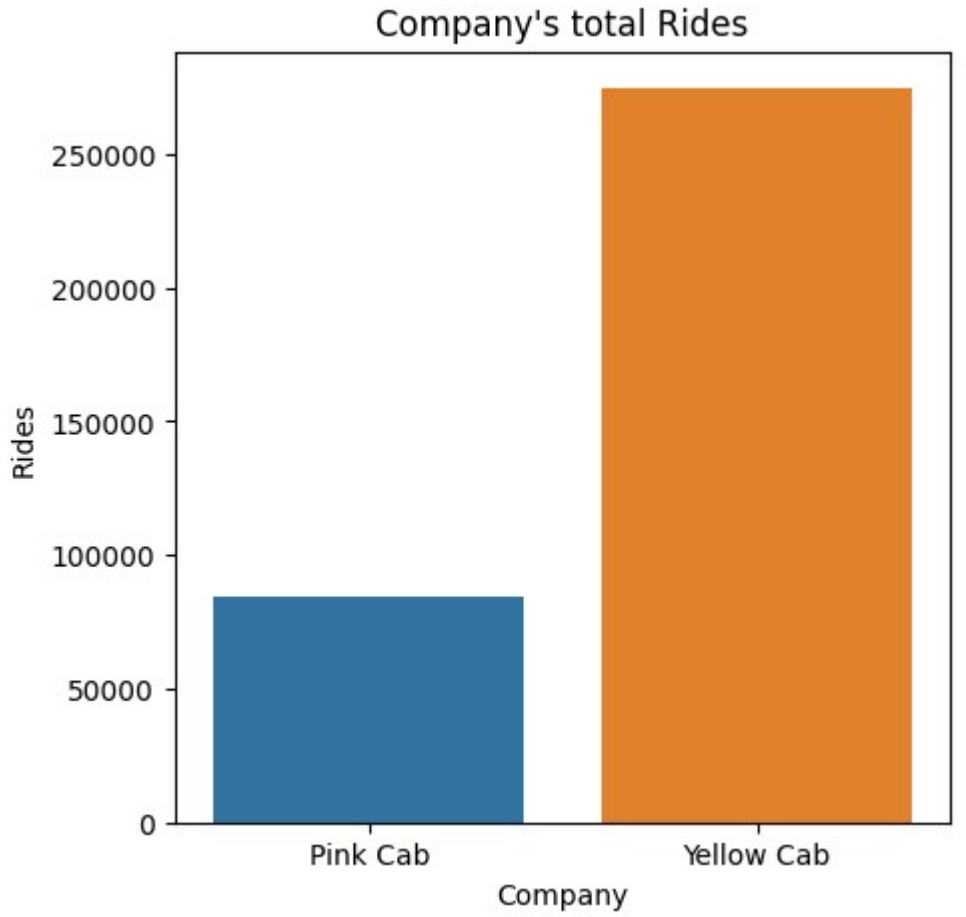


Fig 3

It appears that major cities such as New York, Chicago, Los Angeles, and Washington, D.C. have a higher number of customers according to the data. In other words, these big cities seem to attract more taxi customers compared to other places shown in the graph.

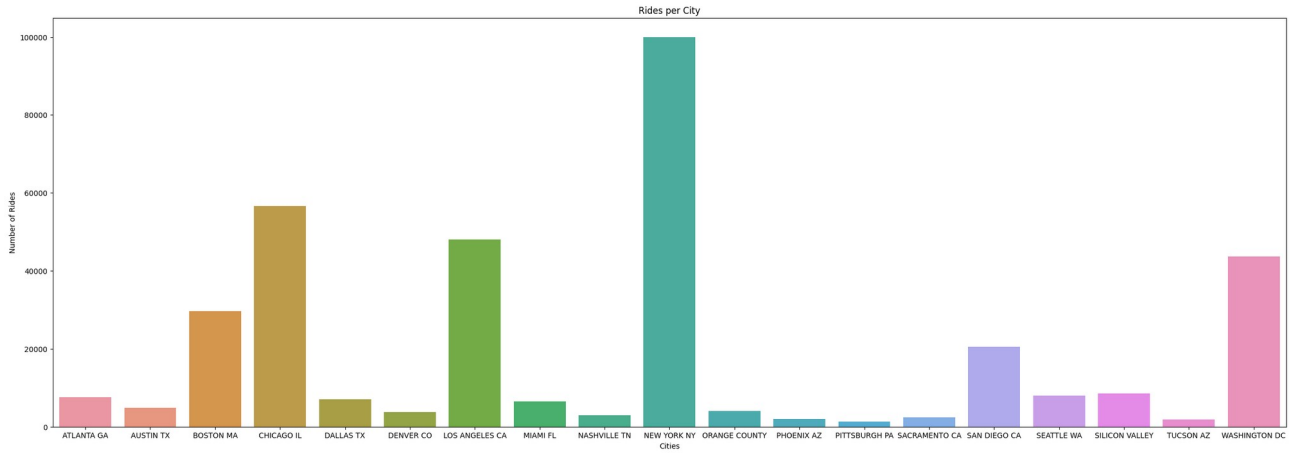


Fig 4

The graph provides data on how far, on average, taxis drive in a single day. This information can be important for various purposes, such as understanding taxi usage patterns, estimating fuel consumption, or assessing the overall activity of taxi services. It gives you an idea of the typical distance covered by taxis on any given day.

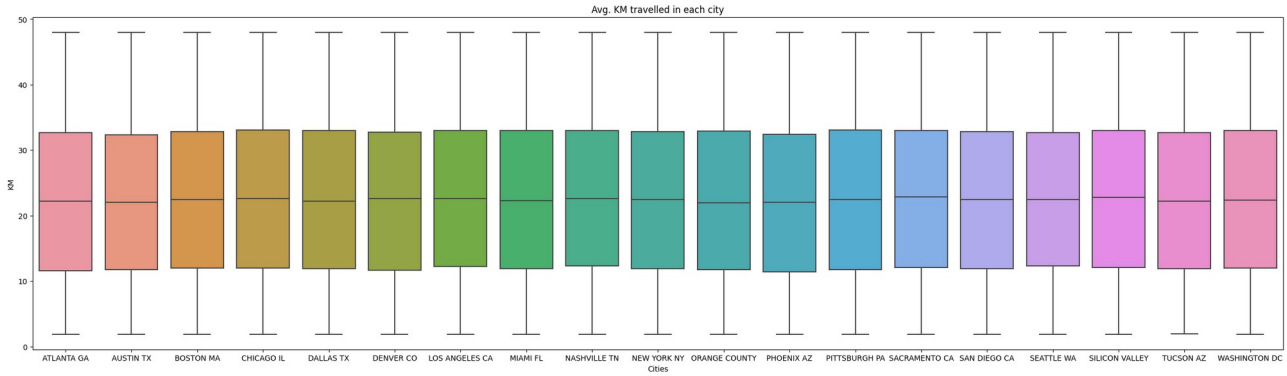


Fig 5

The graph is giving us an idea of how expensive or affordable it is to take a taxi in various cities. Each bar on the graph represents a different city, and the height of the bar shows the average cost of a taxi ride in that city.

For example, if a city has a high bar on the graph, it means that, on average, taking a taxi there tends to be relatively costly. On the other hand, if a city has a low bar, it suggests that taxi rides in that city are, on average, more affordable.

This information can be valuable for travelers and residents to understand the cost of transportation in different places, and it might influence their choice of transportation or travel plans.

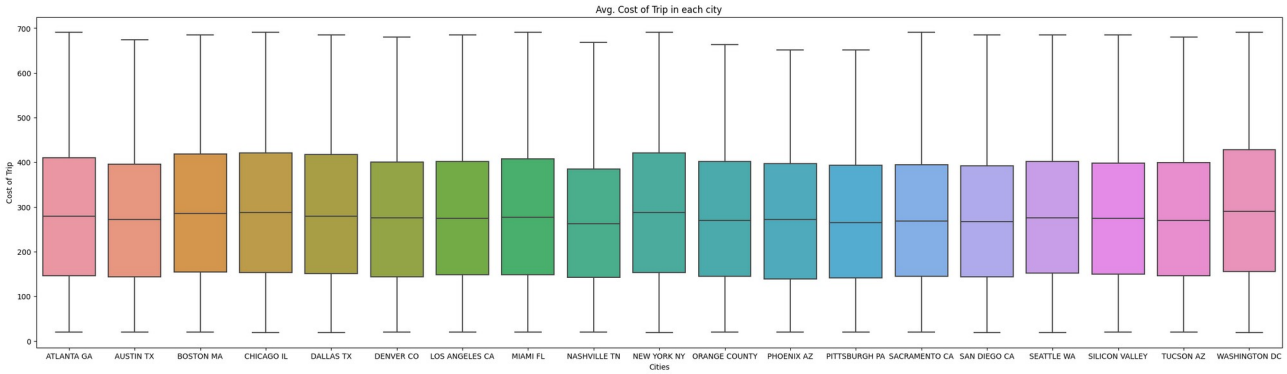


Fig 6

By looking at the graph, you can get an idea of the distribution of customers across these age groups and see which age group has the highest or lowest number of customers. This information can be useful for businesses to understand their customer demographics and tailor their services accordingly.

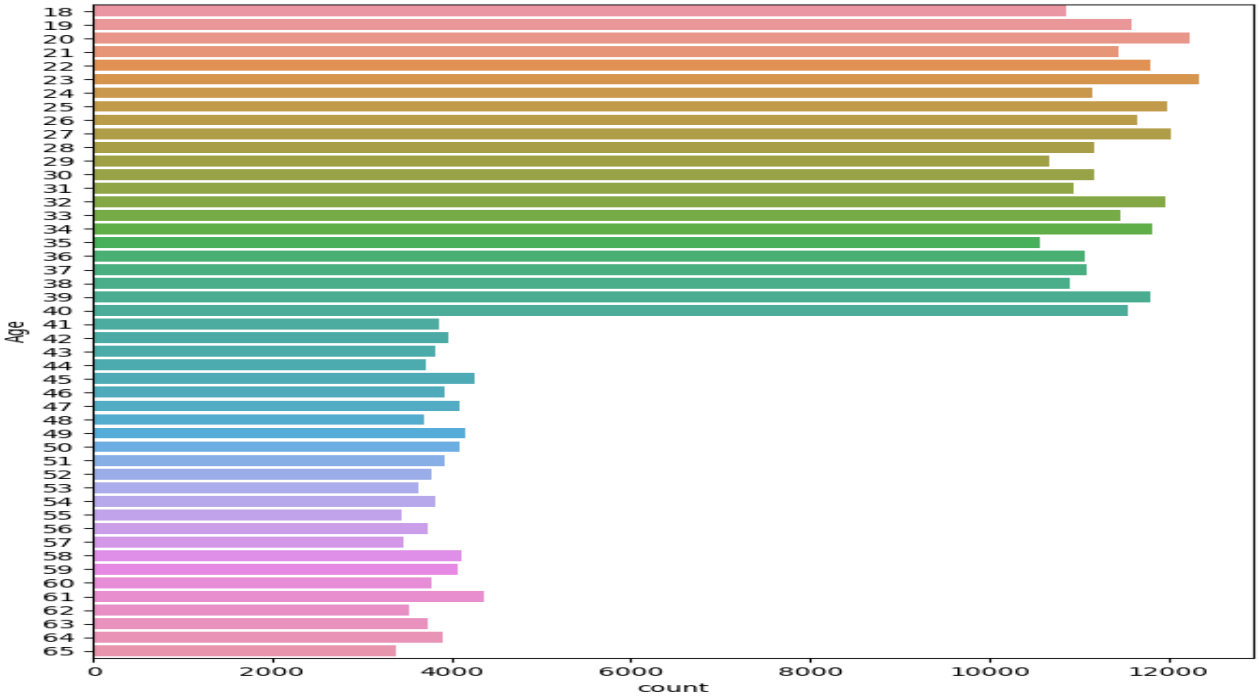


Fig 7

The graph likely displays different payment methods that customers use when taking a taxi. These payment modes could include options like cash, credit cards, debit cards, mobile apps, or any other means by which customers settle their taxi fares. By looking at the graph, you can understand which payment methods are more commonly used by taxi riders, and this information can be valuable for taxi companies and businesses to tailor their services and payment options to meet customer preferences.

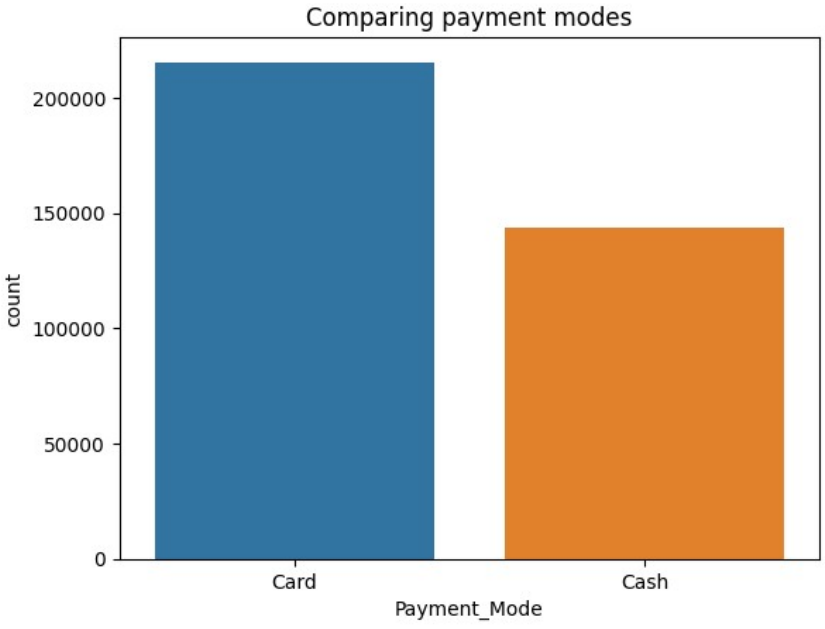


Fig 8

Distance Traveled (KM): This is the number of kilometers you travel during your taxi ride. It could be a short trip covering just a few kilometers or a long journey covering many kilometers.

Cost of Trip: This is the amount of money you have to pay for the taxi ride. It includes the base fare and any additional charges that might apply, such as per-kilometer rates or surcharges.

The graph likely plots these two variables on its axes. As you move along the horizontal axis (left to right), you're looking at different distances traveled, and as you move along the vertical axis (bottom to top), you're seeing how the cost of the trip changes.

In essence, the graph helps you understand how the cost of a taxi ride increases or decreases as you travel different distances. For example, it might show that shorter trips have lower costs, while longer trips have higher costs, which is a common pricing structure for taxi services.

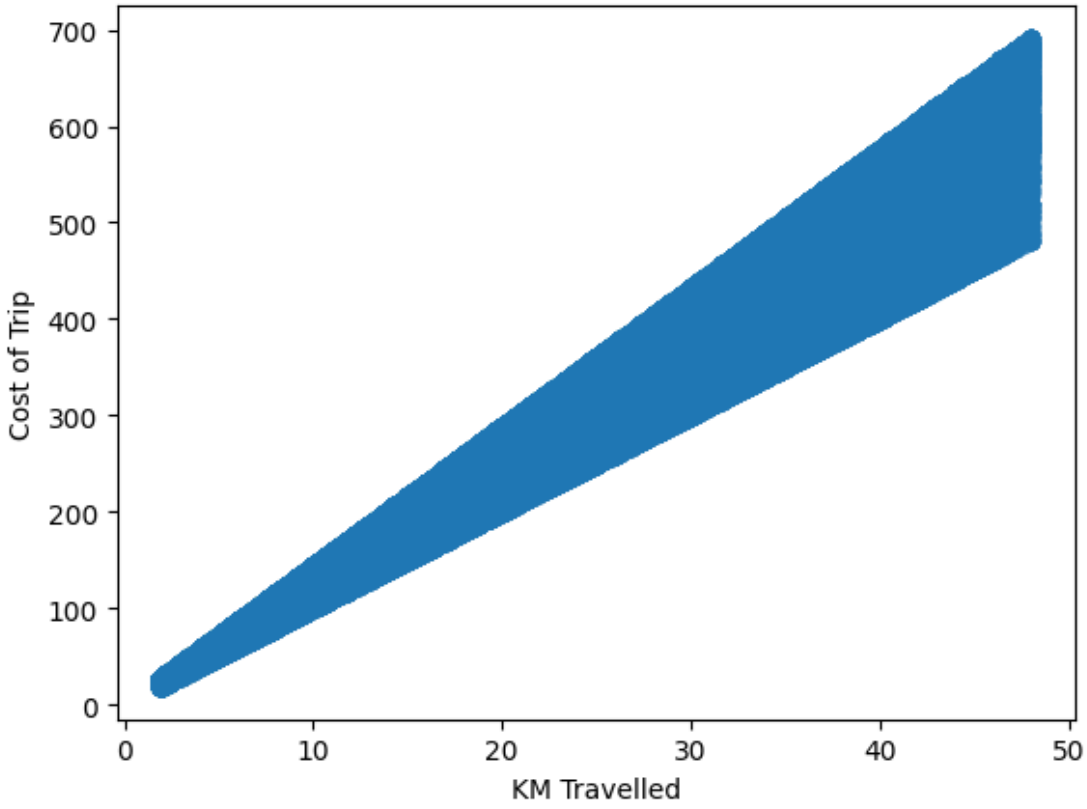
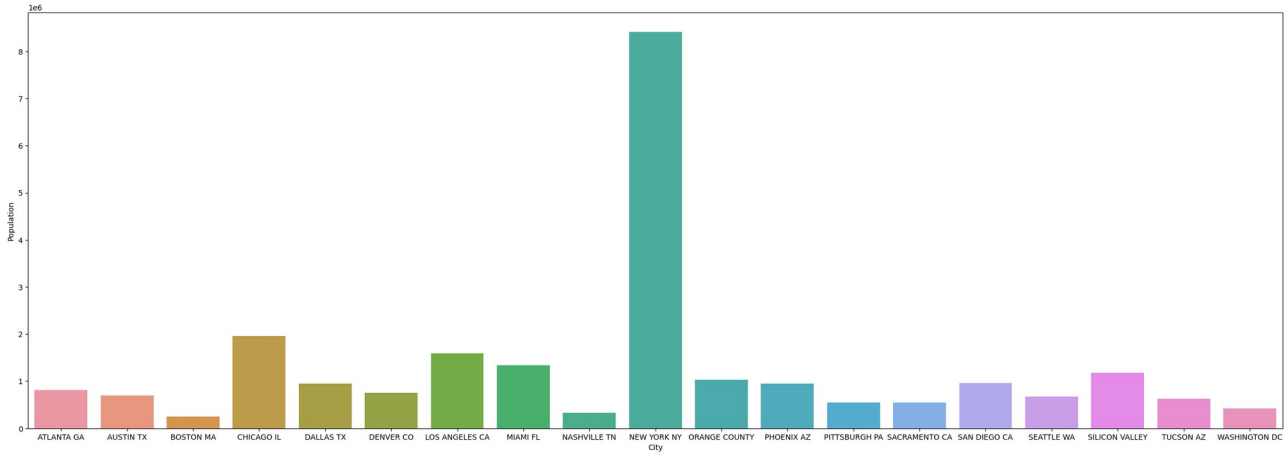


Fig 9

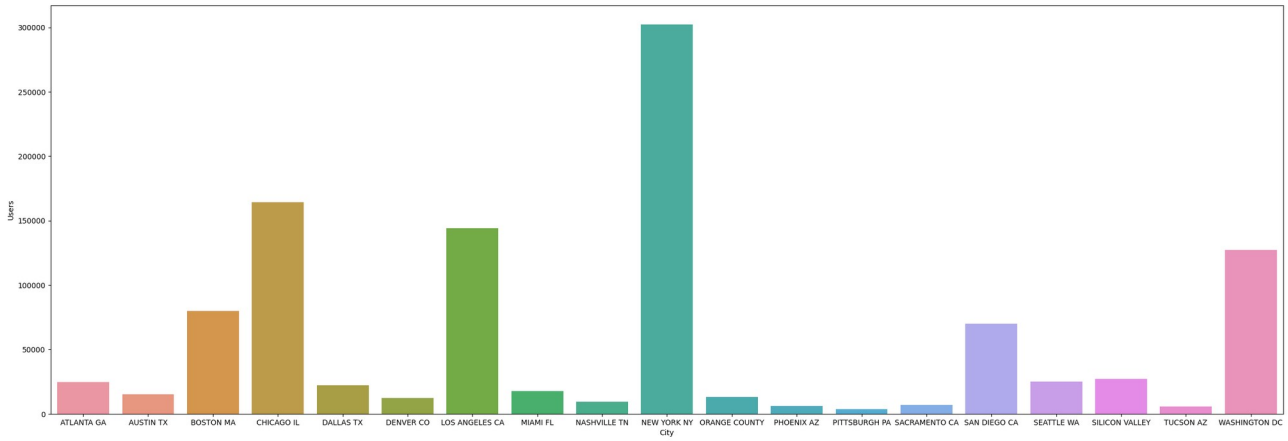
The graph provides information about the population of different cities. It indicates that major cities like New York, Chicago, Los Angeles, and Washington, D.C. have larger populations compared to other cities on the graph. These cities have more residents living within their boundaries.

The higher population in these major cities can explain why they might have more taxi customers. With more people residing in these areas, there is a larger pool of potential customers who may need taxi services for daily transportation, work commutes, or various other reasons. The demand for taxis often correlates with the population size of a city because more people generally mean more transportation needs.

So, in simpler terms, the graph shows that these big cities have more residents, and this larger population can contribute to the higher number of taxi customers in these areas.



The graph displays the number of cab users in each city, then the statement means that cities like New York, Chicago, Los Angeles, and Washington, D.C. have more people who use taxi services compared to other cities on the graph. In simpler terms, these major cities have a higher demand for taxis, suggesting that more residents and visitors in those areas rely on cabs for transportation.



Principal Component Analysis (PCA) is a mathematical technique used in data analysis to simplify complex datasets. It's especially useful when dealing with a large amount of data with many variables. PCA aims to reduce the dimensionality of the data while preserving its important information.

Here's what's happening with PCA in the context of the graph:

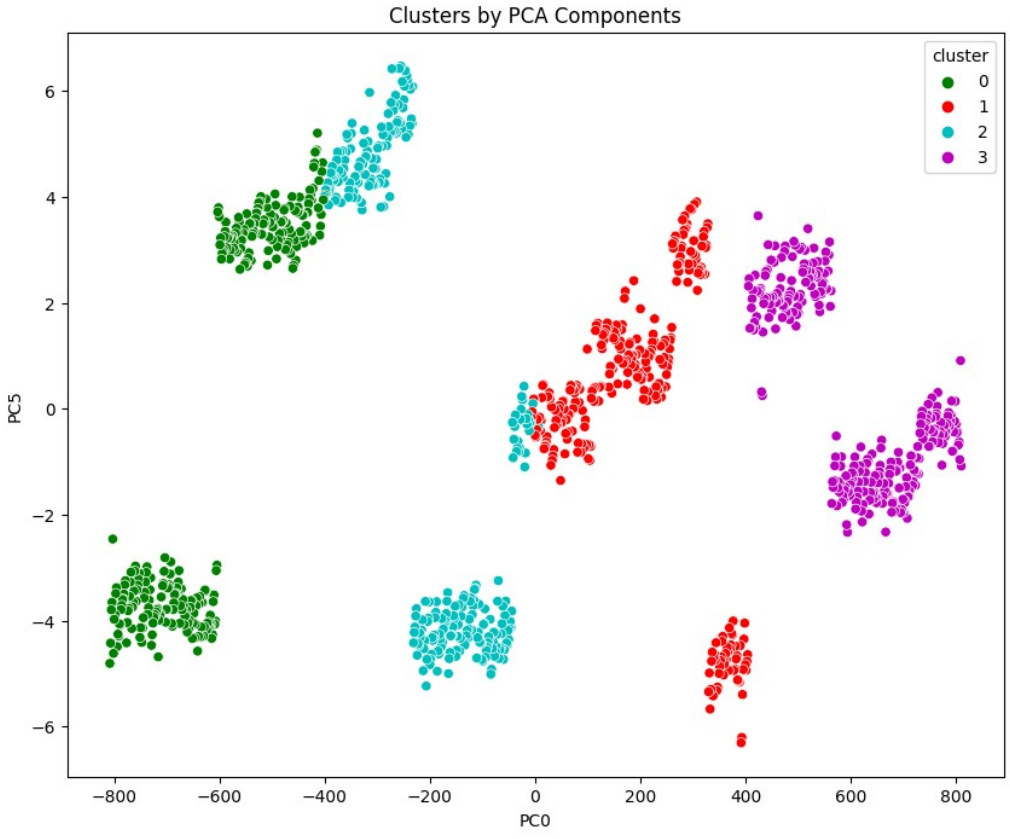
Input Data: There is a dataset called "data1." This dataset likely contains a lot of data points, each with multiple features or variables. These features could be things like age, income, location, or any other relevant information, depending on the context of the data.

Dimension Reduction: PCA takes this high-dimensional data and transforms it into a lower-dimensional representation. It does this by finding new variables, called principal components, that capture the most significant variation in the data. These principal components are combinations of the original features.

Visualization: The graph you're looking at is a visual representation of the results of PCA applied to "data1." It likely shows how much of the total variation in the data is explained by each principal component. Each point or line on the graph might represent a principal component, and the graph can help you understand which components are the most important in describing the data.

Insight: By analyzing this graph, we can gain insights into the structure and patterns within "data1." We can see which variables or components contribute the most to the overall variation in the dataset. This can be valuable for tasks like data compression, visualization, or feature selection.

In summary, the graph labeled "Principal Component Analysis (PCA) for data1" is a visual representation of how PCA was used to simplify and understand a complex dataset called "data1." It helps you see which aspects of the data are most significant and how they contribute to the overall variation in the dataset.



K-means Clustering Analysis:

K-means clustering is a technique used in data analysis to group similar data points together. Imagine you have a set of data points on a graph, and you want to categorize them into distinct groups or clusters based on their similarities. K-means clustering helps you do just that.

Here's how it works:

1. Choose a Number of Clusters (K): You start by deciding how many clusters (K) you want to divide your data into. For example, if you think there are three distinct groups in your data, you would choose K=3.

2. Initial Cluster Centers: The algorithm randomly selects K points in your data as the initial cluster centers. These points serve as the centers around which the clusters will form.

3. Assign Data Points to Nearest Cluster: Each data point is assigned to the cluster whose center is closest to it.

4. Recalculate Cluster Centers: The center of each cluster is recalculated based on the data points assigned to it.

5. Repeat Steps 3 and 4: Steps 3 and 4 are repeated iteratively until the cluster centers no longer change significantly or a predefined number of iterations is reached.

6. Result: After the algorithm converges, you have your data grouped into K clusters, and each data point belongs to one of these clusters.

Elbow Method:

Now, let's talk about the elbow method, which is used to determine the optimal number of clusters (K) in a K-means analysis.

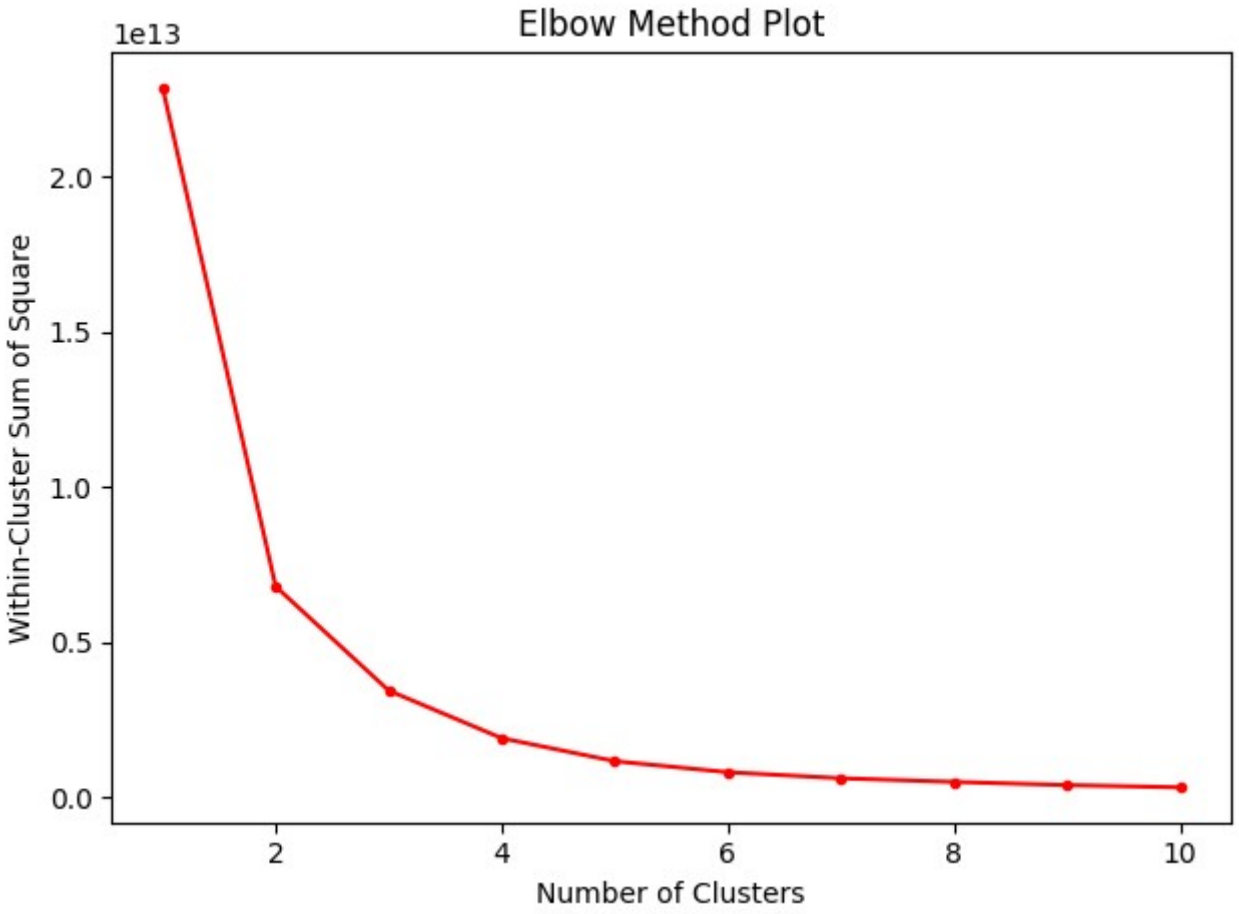
- The idea behind the elbow method is to run the K-means clustering algorithm for different values of K (e.g., from 1 to 10) and calculate the sum of squared distances from each data point to its assigned cluster center for each K value.

- These sums of squared distances are plotted on a graph, with K on the x-axis and the sum of squared distances on the y-axis.

- The "elbow" of the curve on the graph represents a point where adding more clusters doesn't significantly reduce the sum of squared distances. In other words, it's where the rate of improvement starts to level off.

- The number of clusters (K) corresponding to the elbow point is considered the optimal number of clusters for your data. It represents a balance between having enough clusters to capture the data's structure and avoiding too many clusters, which might overcomplicate the interpretation.

So, when you mention "K-means clustering analysis after elbow method," you're likely describing the results of applying K-means clustering to your data and using the elbow method to determine the most appropriate number of clusters for your specific dataset.



Segment Extraction

To summarize, segment extraction based on the data involved using a clustering technique, specifically K-means clustering with 4 clusters as determined by the elbow method. This process allowed you to group similar data points together into four distinct segments or clusters. These clusters represent subsets of your data that share common characteristics or patterns, and they can be used for various analytical and decision-making purposes.

Each segment likely captures a unique aspect or behavior within the dataset, and exploring these segments can provide valuable insights into your data. It's important to delve deeper into each cluster to understand what makes them distinct and how they can inform your business or research objectives. Segment extraction through clustering helps in data-driven decision-making, target audience identification, and tailoring strategies to meet the specific needs of each segment.

The choice is of using 4 clusters in the segment extraction via K-means clustering.

Segment 1: Age Group 18-30
- This segment likely represents a cluster of customers who fall within the age group of 18 to 30 years old.
- Understanding this demographic can help tailor marketing strategies and promotions that resonate with younger customers.
- Consider offering promotions or services that are particularly appealing to this age group, such as student discounts or rideshare partnerships.

Segment 2: Fridays and Afternoons
- This segment may include data points related to specific times and days when taxi rides are in high demand, particularly on Fridays and during the afternoons.
- This insight can inform driver scheduling and dispatch strategies. You might allocate more resources during these peak times to meet the increased demand.
- It's also an opportunity to implement dynamic pricing models if applicable, allowing you to maximize profits during high-demand periods.

Segment 3: Year-End Month
- This segment could represent data related to rides and profits during the year-end month.
- It's common for people to travel more during the holiday season for shopping, visiting family, or attending events.
- Consider running holiday-themed promotions, extending service hours, or offering packages for year-end events to capitalize on the increased demand and potentially boost profits.

By recognizing these patterns within your data and aligning your strategies accordingly, you can optimize your taxi service operations. Keep in mind that data-driven decision-making is an ongoing process. Regularly monitor and analyze data to adapt to changing customer preferences and market dynamics, ensuring your services remain competitive and profitable.