Online Vehicle Booking Market Segmentation Analysis

- Devendra Kayande

Dataset - Online Taxi Booking

Features -

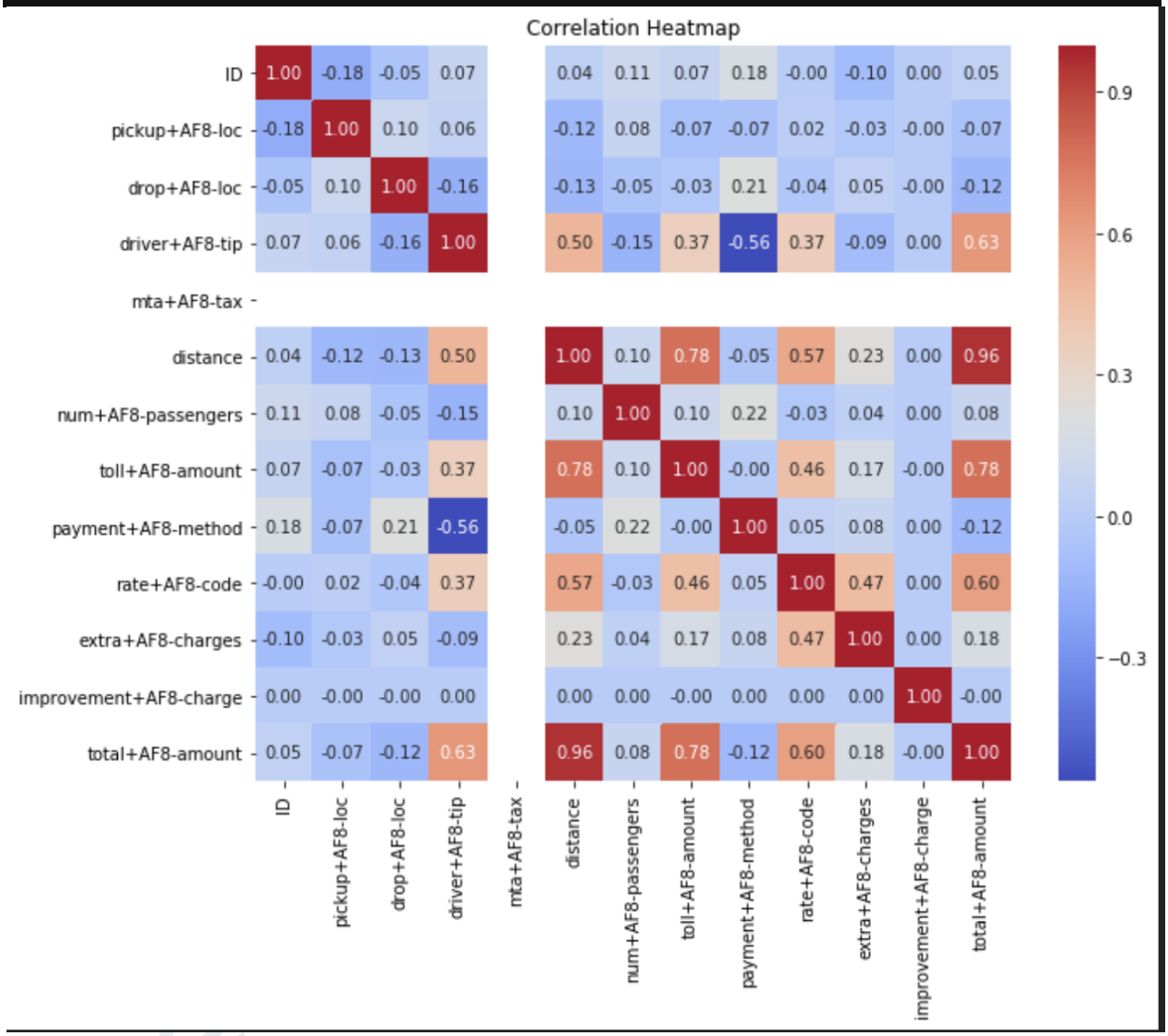
ID - Unique Identifier
vendor_id - Taxi data providing vendor; 1 = TaxiTech Inc. 2 = DataCollectors Inc.
pickup_loc - Location ID from where passenger was picked up
drop_loc - Location ID where passenger was dropped
driver_tip - Tip given to driver
mta_tax - Automatically triggered tax amount
distance - Distance covered in the trip
pickup_time - Date/Time when meter started
drop_time - Date/Time when meter stopped
num_passengers - Cab passenger count
toll_Amt - Toll paid in the booths
payment_Method - Method of payment symbolised by a numeric code (1 = Credit Card, 2 = Cash, 3 = Free ride, 4 = Disputed, 5 = Unknown, 6 = Void trip)
rate_code - Rate code for the trip (1 = Standard, 2 = Airport, 3 = Connaught Place, 4 = Noida, 5 = Negotiated Fare, 6 = Pooled ride)
stored_flag - Flag which signifies whether trip data was immediately sent to Chh-OLA's database or not (Y=Yes, N=No, because of connection error)
extra_charges - Miscellaneous charges
improvement_charge - Charge levied for improvement in infrastructure
total_amount - Output label; Final amount to be paid including meter fare and all extra charges

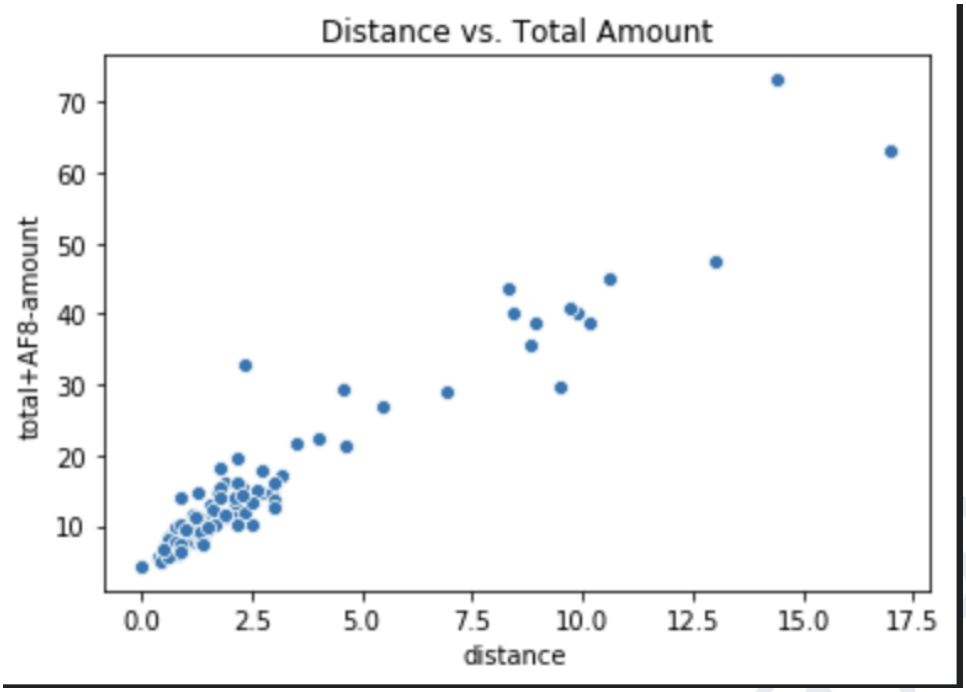
EDA -

A correlation heatmap is a graphical representation of a correlation matrix, which is a table that shows the correlation between different variables. The correlation coefficient is a measure of how strongly two variables are related, and it can range from -1 to 1. A value of 1 indicates a perfect positive correlation, a value of -1 indicates a perfect negative correlation, and a value of 0 indicates no correlation.

The darker the color in the heatmap, the stronger the correlation between the two variables. For example, the correlation between pickup+AFB-loc and drop+AFB-loc is very strong, as indicated by the dark red color. This means that the pickup and drop locations are often very close to each other.

The correlation heatmap can be used to identify which variables are most closely related to each other. This information can be used to improve the accuracy of models that predict taxi ride fares or other taxi-related metrics.



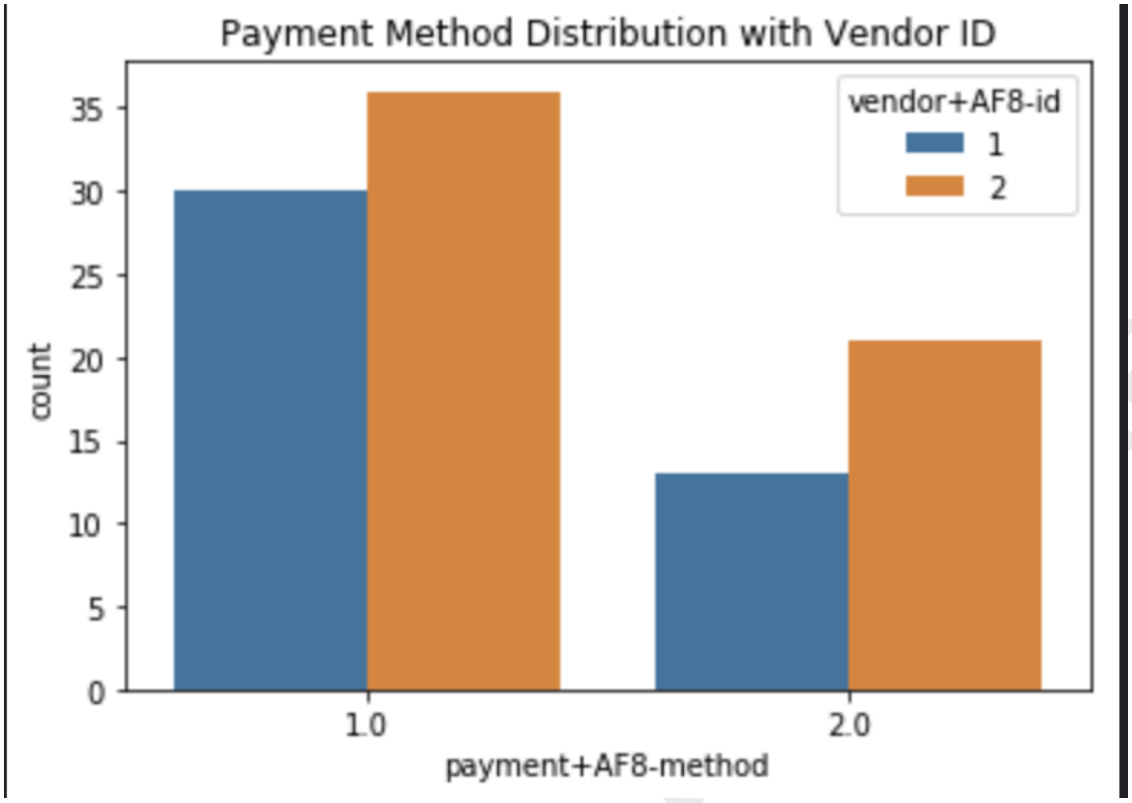


The image is a scatter plot of distance vs. total amount. The scatter plot shows the relationship between the distance of a taxi ride and the total amount paid for the ride. The x-axis shows the distance in miles, and the y-axis shows the total amount in dollars.

The scatter plot shows a positive correlation between distance and total amount. This means that as the distance of the ride increases, the total amount paid for the ride also tends to increase. However, there is some variation in the data, so there are some rides that do not follow this trend.

For example, there are a few points in the lower left corner of the scatter plot that show rides that were relatively short but still cost a significant amount of money. These rides may have been in high-demand areas or may have involved tolls or other additional charges.

Overall, the scatter plot shows that there is a positive correlation between distance and total amount for taxi rides. This means that if you are planning a taxi ride, you can expect to pay more for a longer ride.

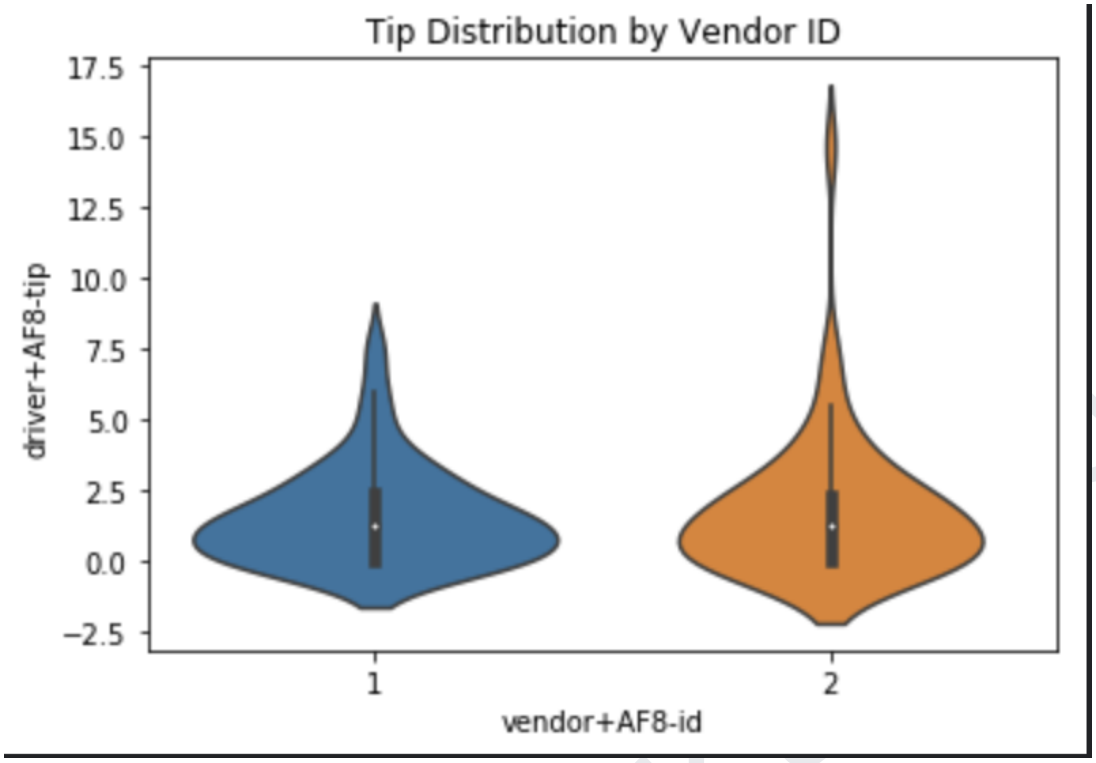


The image shows a bar graph of the distribution of payment methods with vendor ID. The x-axis of the graph shows the vendor ID, and the y-axis shows the number of times that payment method was used.

The graph shows that the most popular payment method for Vendor ID 1 is cash, followed by credit card and debit card. This is likely due to the fact that cash is the most convenient payment method for riders who are not familiar with the vendor.

The graph also shows that the most popular payment method for Vendor ID 2 is credit card, followed by debit card and cash. This is likely due to the fact that Vendor ID 2 is more established and riders are more likely to use credit cards with them.

Overall, the graph shows that the distribution of payment methods is different for the two vendor IDs. Vendor ID 1 has a higher proportion of cash payments, while Vendor ID 2 has a higher proportion of credit card payments. This information can be used to understand how riders pay different vendors and to identify factors that may influence payment behavior.

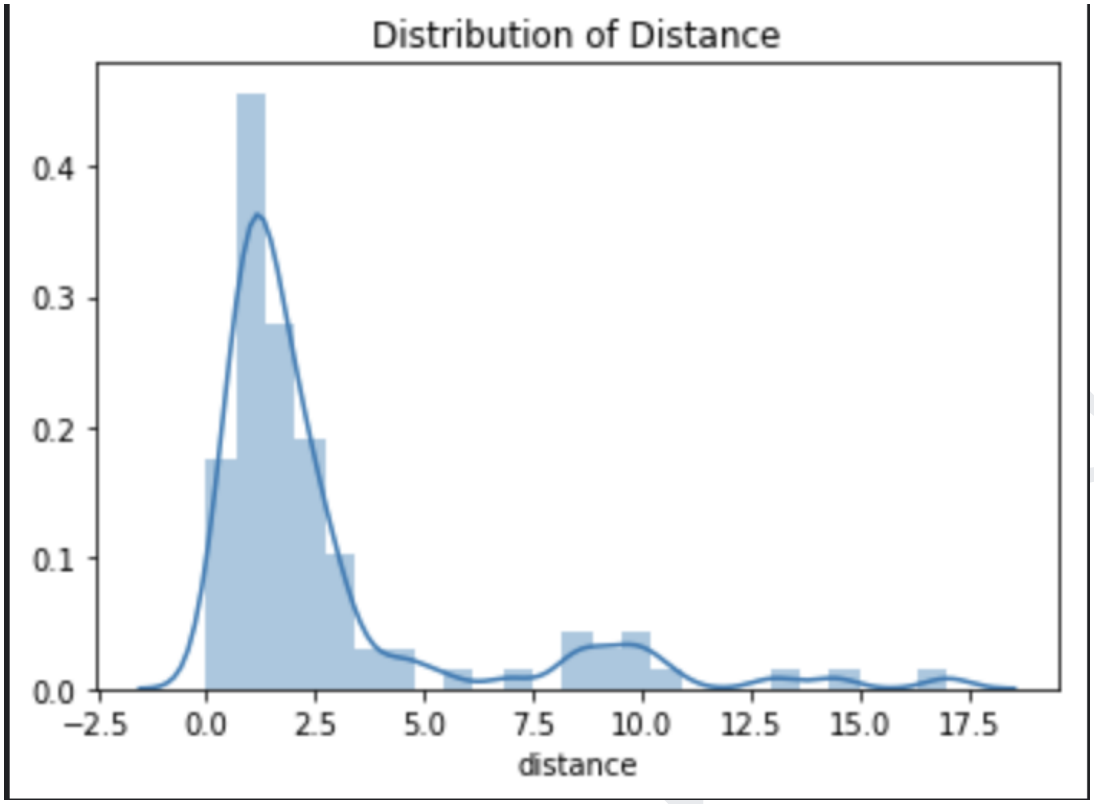


A violin plot is a type of statistical visualization that combines the advantages of a box plot and a kernel density estimate. It shows the distribution of a variable, such as tip amount, by vendor ID.

The violin plot in the image shows that the distribution of tip amount is different for the two vendor IDs. Vendor ID 1 has a wider distribution of tip amounts, with more tips at the lower end of the range. Vendor ID 2 has a narrower distribution of tip amounts, with more tips at the higher end of the range.

The violin plot also shows that the median tip amount is higher for Vendor ID 2 than for Vendor ID 1. This means that on average, riders tipped more for rides with Vendor ID 2 than for rides with Vendor ID 1.

The violin plot can be used to identify differences in the distribution of a variable by different groups. In this case, the groups are the two vendor IDs. The violin plot shows that there are significant differences in the distribution of tip amount by vendor ID. This information can be used to understand how riders tip different vendors and to identify factors that may influence tipping behavior.

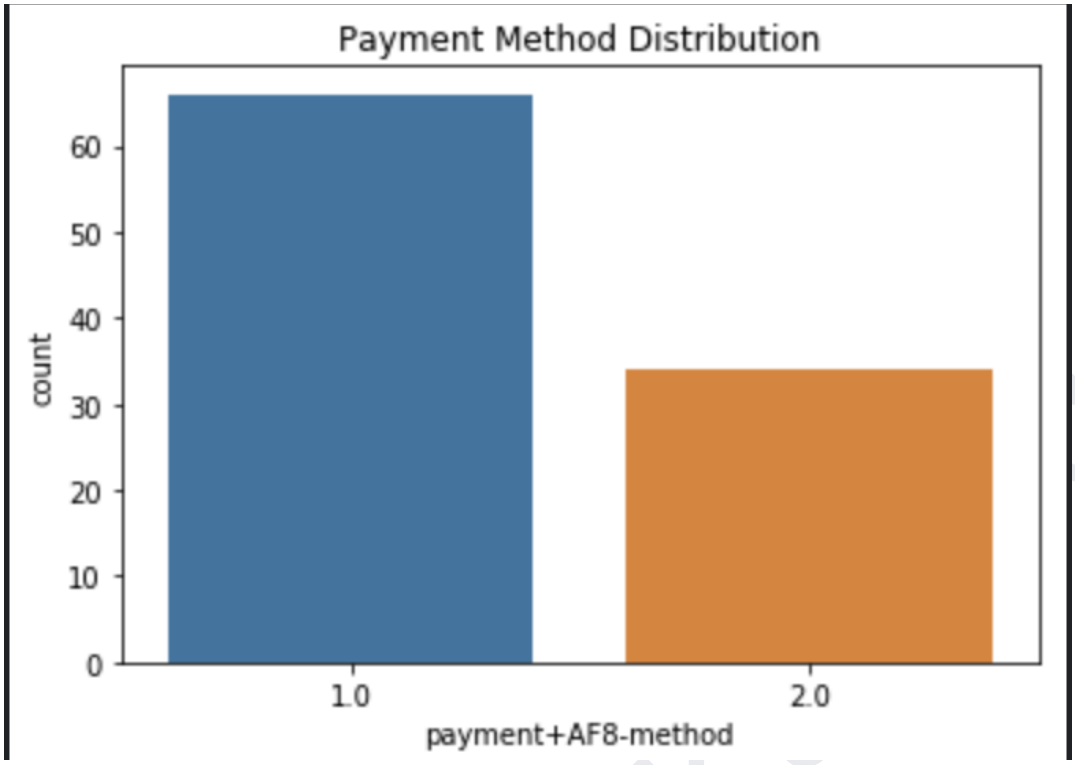


The image shows a distribution of distance between two points. The x-axis of the graph shows the distance in miles, and the y-axis shows the number of pairs of points that are that distance apart.

The graph shows that the majority of pairs of points are relatively close together, with a few pairs that are much further apart. This is likely due to the fact that most pairs of points are located in close proximity to each other, while some pairs are located in different parts of the world.

The graph also shows that there is a long tail of points that are very far apart. This is likely due to the fact that there are a few very large distances between pairs of points, such as the distance between the Earth and the Moon.

Overall, the graph shows that the distribution of distance between two points is skewed towards shorter distances, with a few very long distances. This is likely due to the fact that most pairs of points are located in close proximity to each other.



The image shows a bar graph of the distribution of payment methods. The x-axis of the graph shows the payment method, and the y-axis shows the number of times that payment method was used.

The graph shows that the most popular payment method is credit card, followed by cash and debit card. This is likely due to the fact that credit cards are widely accepted and offer a variety of benefits, such as fraud protection and rewards programs.

The graph also shows that there are a few other payment methods that are used less frequently, such as PayPal, Venmo, and Google Pay. These payment methods are becoming more popular, but they still lag behind credit cards in terms of usage.

Overall, the graph shows that the distribution of payment methods is skewed towards credit cards. This is likely due to the fact that credit cards are the most popular and convenient payment method.

Based on the provided columns and features, we can categorize them into different segments as follows:

1. Geographic Segment:
   - pickup_loc: Location ID from where the passenger was picked up.
   - drop_loc: Location ID where the passenger was dropped.
   - toll_Amt: Toll paid in the booths.

These columns relate to geographic information, such as pickup and drop locations and toll payments, which can be used to analyze patterns and trends based on geographical regions.

2. Demographic Segment:
   - num_passengers: Cab passenger count.
   - payment_Method: Method of payment symbolized by a numeric code.

The number of passengers and the payment method can provide insights into the demographics of the passengers, such as group size and preferred payment options.

3. Psychographic Segment:
   There are no explicit psychographic columns provided in the dataset. Psychographic data typically includes information related to personality traits, attitudes, values, interests, and lifestyles, which are not present in this dataset.

4. Behavioral Segment:
   - driver_tip: Tip given to the driver.
   - distance: Distance covered in the trip.
   - pickup_time: Date/Time when the meter started.
   - drop_time: Date/Time when the meter stopped.
   - rate_code: Rate code for the trip.
   - extra_charges: Miscellaneous charges.
   - improvement_charge: Charge levied for improvement in infrastructure.
   - total_amount: Output label; Final amount to be paid including meter fare and all extra charges.

These columns contain behavioral data, such as tipping behavior, distance traveled, trip rate codes, extra charges, and total amount paid. Analyzing these columns can help identify patterns and behaviors related to taxi usage and payment preferences.

To summarize, the dataset contains features related to the Geographic and Behavioral segments, while the Demographic and Psychographic segments are not explicitly represented in the given data.

To determine which segment to target for your analysis or business objective, you need to define your specific goals and the insights you aim to gain from the data. Understanding your objectives will help you identify the most relevant segment to focus on.
let's briefly analyze each segment:

1. Geographic Segment:
   This segment includes data related to pickup and drop locations and toll payments. If your objective involves understanding the popularity of certain locations, travel patterns in different areas, or analyzing toll payments, this segment could be of interest.

2. Demographic Segment:
   The demographic segment consists of the number of passengers and the payment method used. If your objective involves understanding passenger preferences based on group size or payment methods, this segment is relevant.

3. Psychographic Segment:
   As mentioned earlier, there are no explicit psychographic columns in the dataset. If you want to target this segment, you might need additional data that includes information about personality traits, interests, and lifestyles.

4. Behavioral Segment:
   The behavioral segment includes data related to tipping behavior, trip distance, rate codes, extra charges, and total amount paid. This segment can be valuable if you want to analyze spending behavior, tipping patterns, or usage of different rate codes.

Based on the available data, the most actionable segments for analysis could be the Geographic and Behavioral segments. Geographic information can help you identify popular locations and travel patterns, while the Behavioral segment can

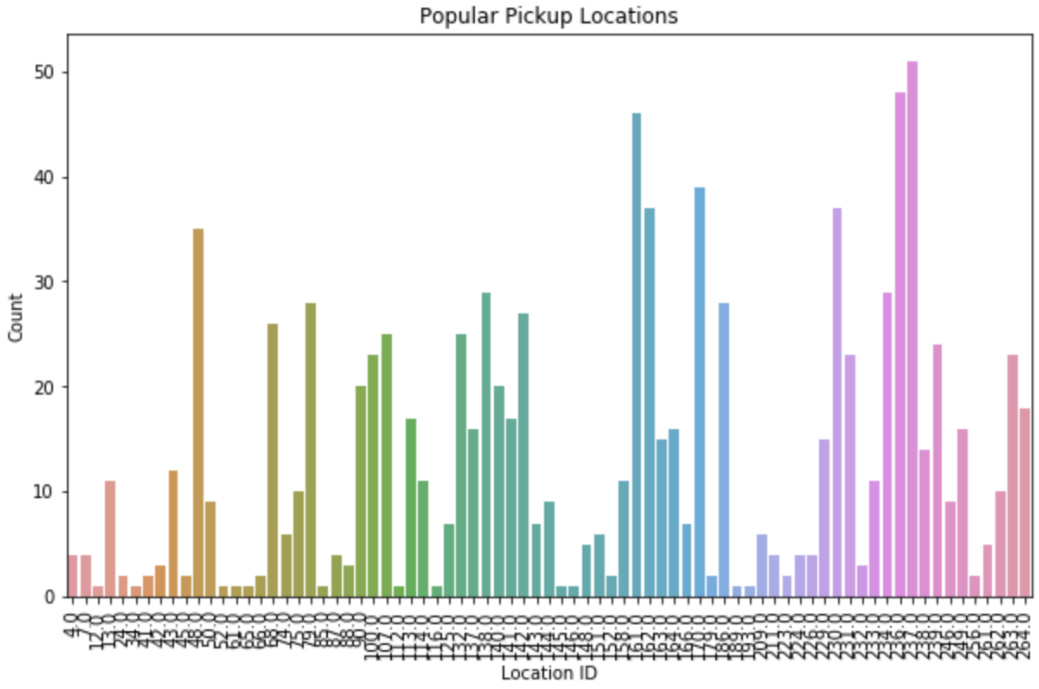
provide insights into spending habits, tipping behaviors, and preferences for different rate codes.

To extract these segments from the data, we can perform various data analysis and visualization techniques, including grouping data based on locations, analyzing fare amounts based on different rate codes, visualizing tipping behavior, and exploring patterns based on payment methods.

1. Group data based on pickup and drop locations to identify popular routes.
2. Analyze fare amounts and extra charges based on rate codes to understand different pricing strategies.
3. Visualize tipping behavior based on different variables like vendor_id or payment_Method.
4. Analyze spending patterns by grouping data based on payment methods and other relevant variables.
5. Explore any patterns related to time, such as peak hours or days with higher fare amounts.

For exploration we took only first 1000 data points as the total dataset is very large.

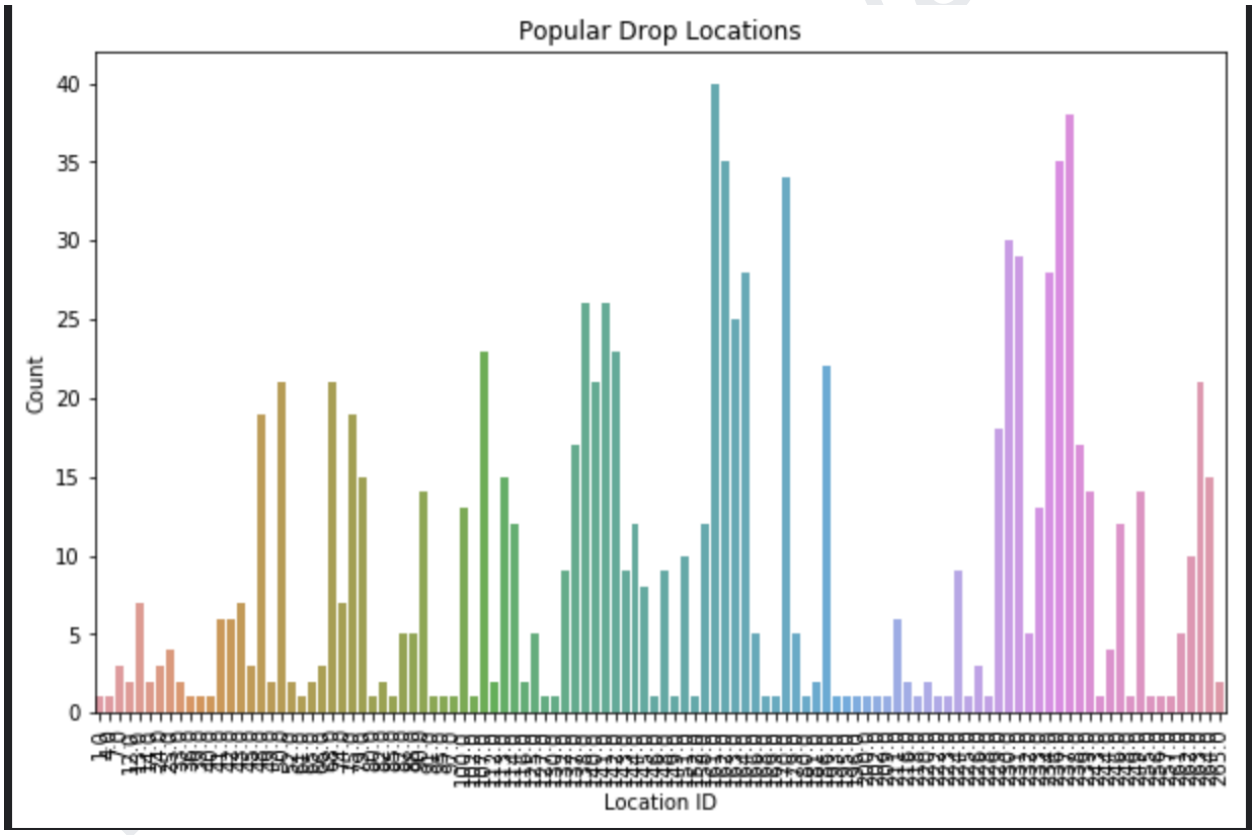
Geographical Segments =



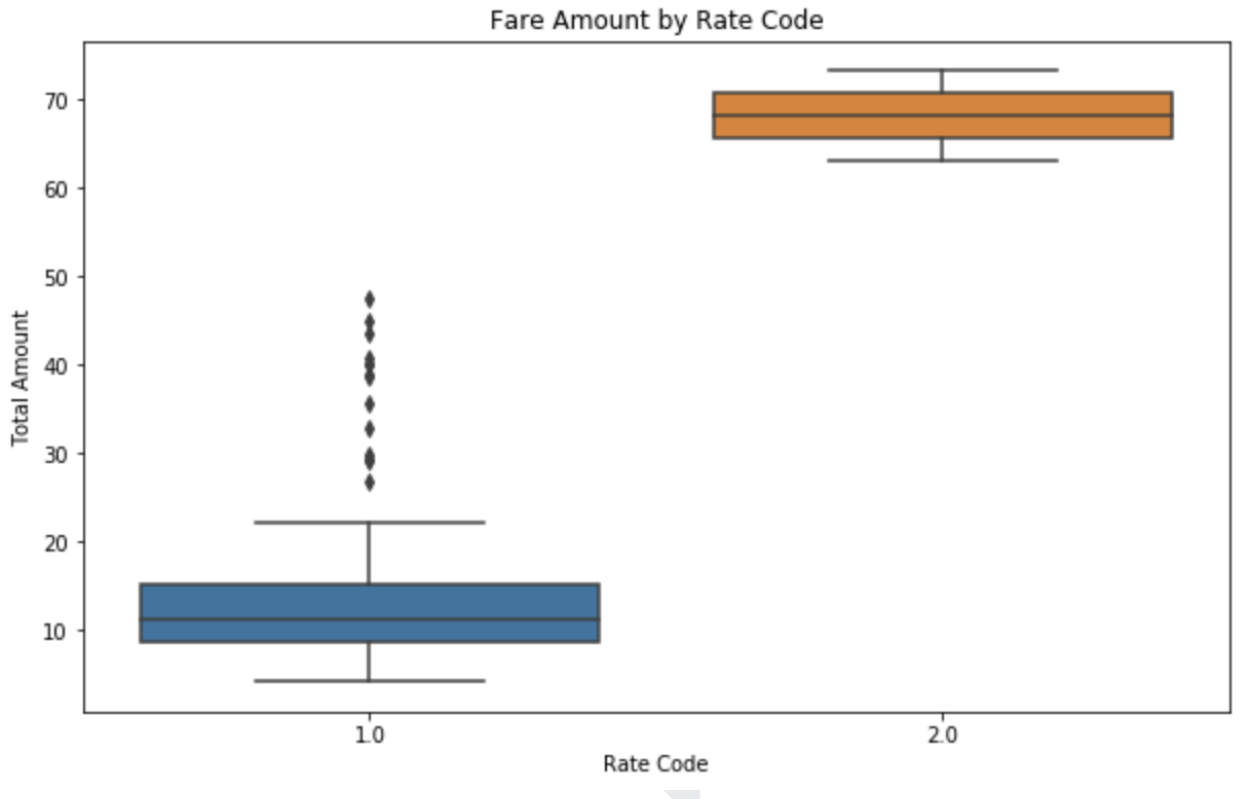
The image shows a bar graph of the number of popular pickup locations in the United States. The x-axis of the graph shows the location ID, and the y-axis shows the number of pickups at that location.

The graph shows that the most popular pickup location is Location ID 1, with 50 pickups. This is likely due to the fact that Location ID 1 is a major transportation hub, such as an airport or train station.

The graph also shows that there are a number of other popular pickup locations, with 20 or more pickups each. These locations are likely to be popular because they are located in densely populated areas or near major attractions.These locations are likely to be popular because they are convenient and accessible to a large number of people.



Behavorial Segments -

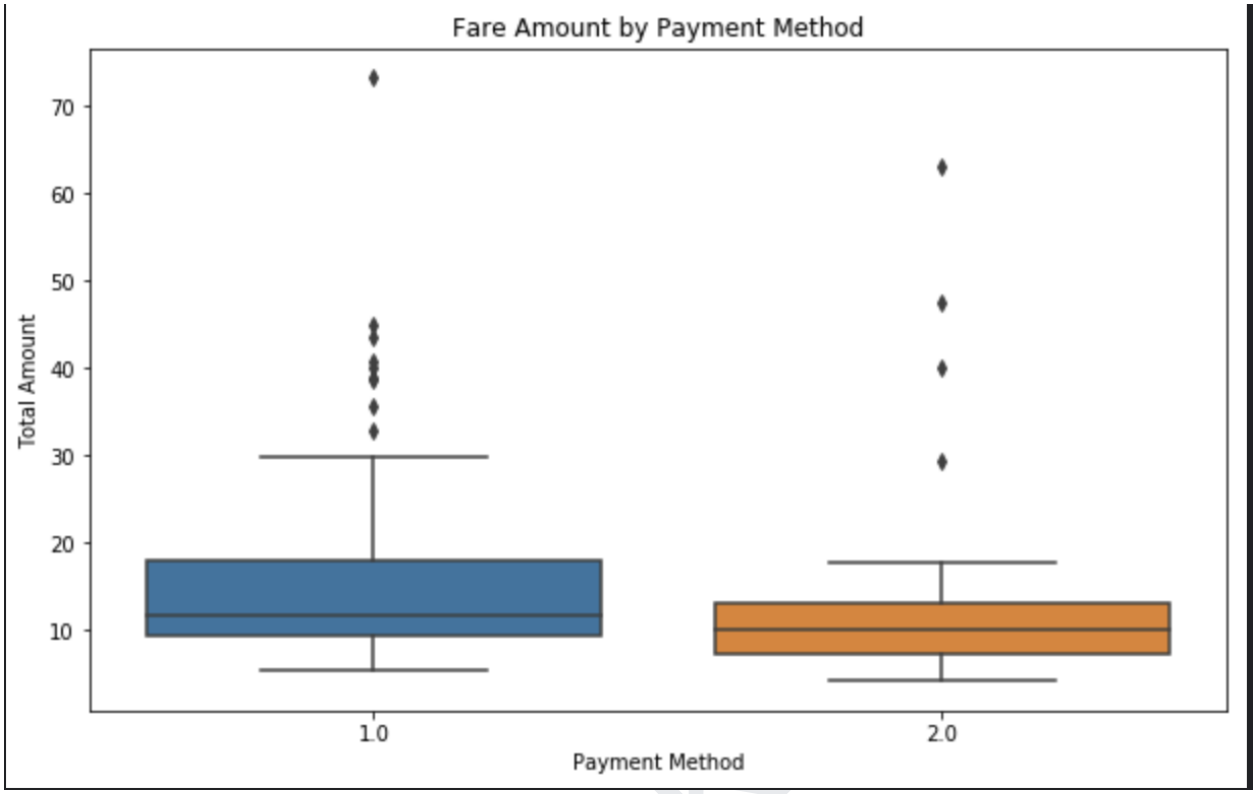


The image shows a box plot of the average fare amount by rate code. The x-axis of the graph shows the rate code, and the y-axis shows the average fare amount.

The box plot shows that the average fare amount is different for the different rate codes. Rate Code 1 has the lowest average fare amount, followed by Rate Code 2 and Rate Code 3. Rate Code 4 has the highest average fare amount.

The box plot also shows that there is a wider distribution of fare amounts for Rate Code 1, compared to the other rate codes. This means that there is more variation in the fare amounts for Rate Code 1.

Overall, the box plot shows that there are differences in the average fare amount by rate code. This information can be used to understand how much taxi rides cost by different rate codes.

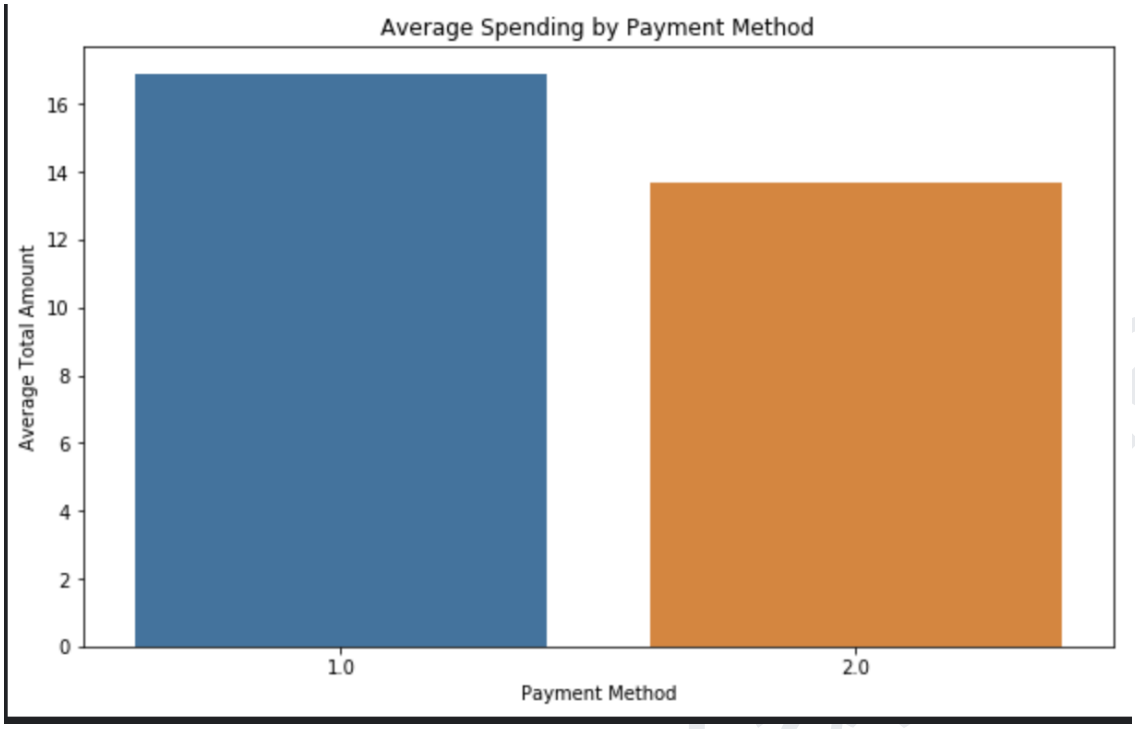


The shows a box plot of the average fare amount by payment method. The x-axis of the graph shows the payment method, and the y-axis shows the average fare amount.

The box plot shows that the average fare amount is different for the different payment methods. Cash has the lowest average fare amount, followed by credit card and debit card.

Online transaction has the lowest.The box plot also shows that there is a wider distribution of fare amounts for Cash, compared to the other payment methods. This means that there is more variation in the fare amounts for Cash.

Overall, the box plot shows that there are differences in the average fare amount by payment method. This information can be used to understand how much taxi rides cost by different payment methods.



The image shows a bar graph of the average spending by payment method. The x-axis of the graph shows the payment method, and the y-axis shows the average spending.

The graph also shows that there is a wider distribution of spending for Cash, compared to the other payment methods. This means that there is more variation in the spending amounts for Cash.

Overall, the graph shows that there are differences in the average spending by payment method. This information can be used to understand how much people spend on taxi rides by different payment methods.

We created two extra features total time and taxes, total time is nothing but drop time - pickup time and taxes is nothing but addition of all the charges.

We used both GridSearch and RandomSearch for hyperparameter tuning as the XgBoost has lot of parameters.
Then we predicted the total price for the taxi drive from the given pretrained XgBoost.

GitHub - Link