MTH209: DATA SCIENCE LAB 2

Optimal Taxi Business Management Plan

Team Members

Indraj Prajapat, MSc Statistics, (231080044)

Chandan Kumar Singh, MSc Statistics, (231080030)

Gaurav Tomar, MSc Statistics, (231080039)

Ketan Saini, BS SDS, (220523)

Vandan Neema, BS SDS, (221165)

VISION BEHIND OUR PROJECT

In today's ever-changing transportation industry, taxi companies must prioritize efficiency and customer satisfaction. By combining data science methods and statistical concepts, we hope to transform current taxi management practices and improve service quality while streamlining operations.

Drawing inspiration from a comprehensive case study, our project is centred around utilizing advanced statistical operations and data analysis methodologies to harness insights from vast amounts of taxi trip data. The dataset, sourced from reputable transportation authorities, provides a rich repository of information including trip distance, fare, and duration, among other variables.

Our objective is to detect patterns and trends in the dataset through statistical operations like mean, median, and standard deviation. This will facilitate a more profound comprehension of the factors that impact trip dynamics and customer preferences. In addition, I will use data visualization techniques to create graphs that illustrate the ideal values of variables, emphasizing distributions and trends that are important for making well-informed decisions.

Our project aims to provide taxi businesses with useful insights for effective resource allocation, route optimization, and pricing strategies by utilizing an inventive approach. Our goal is to turn traditional taxi operations into data-driven, agile businesses that can thrive over time in a constantly changing market by utilizing the power of data and statistical analysis.

Contents

Aim:

To analyse historical taxi trip data and develop a comprehensive business plan for starting and expanding a taxi company, including loan repayment strategies, profit projection, and expansion plans, leveraging insights from data analysis and predictive modelling techniques.

Libraries used:

- `requests`
- `pyarrow.parquet`
- `pandas`
- `numpy`
- `io`
- `matplotlib.pyplot`
- `seaborn`
- `statsmodels.api`
- `datetime`
- `scipy`
- `pulp`

Introduction to our dataset

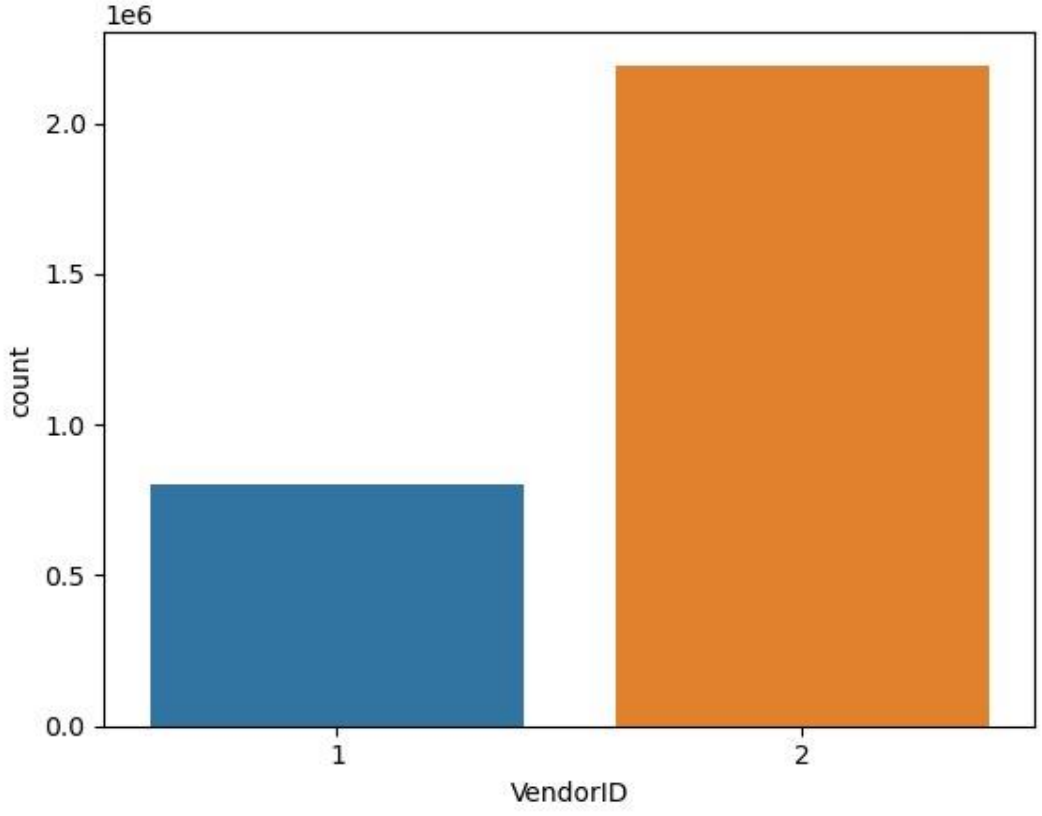
we obtained our dataset from https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page .(refer above image)

- *VendorID : A code indicating the TPEP provider that provided the record.*
- 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc

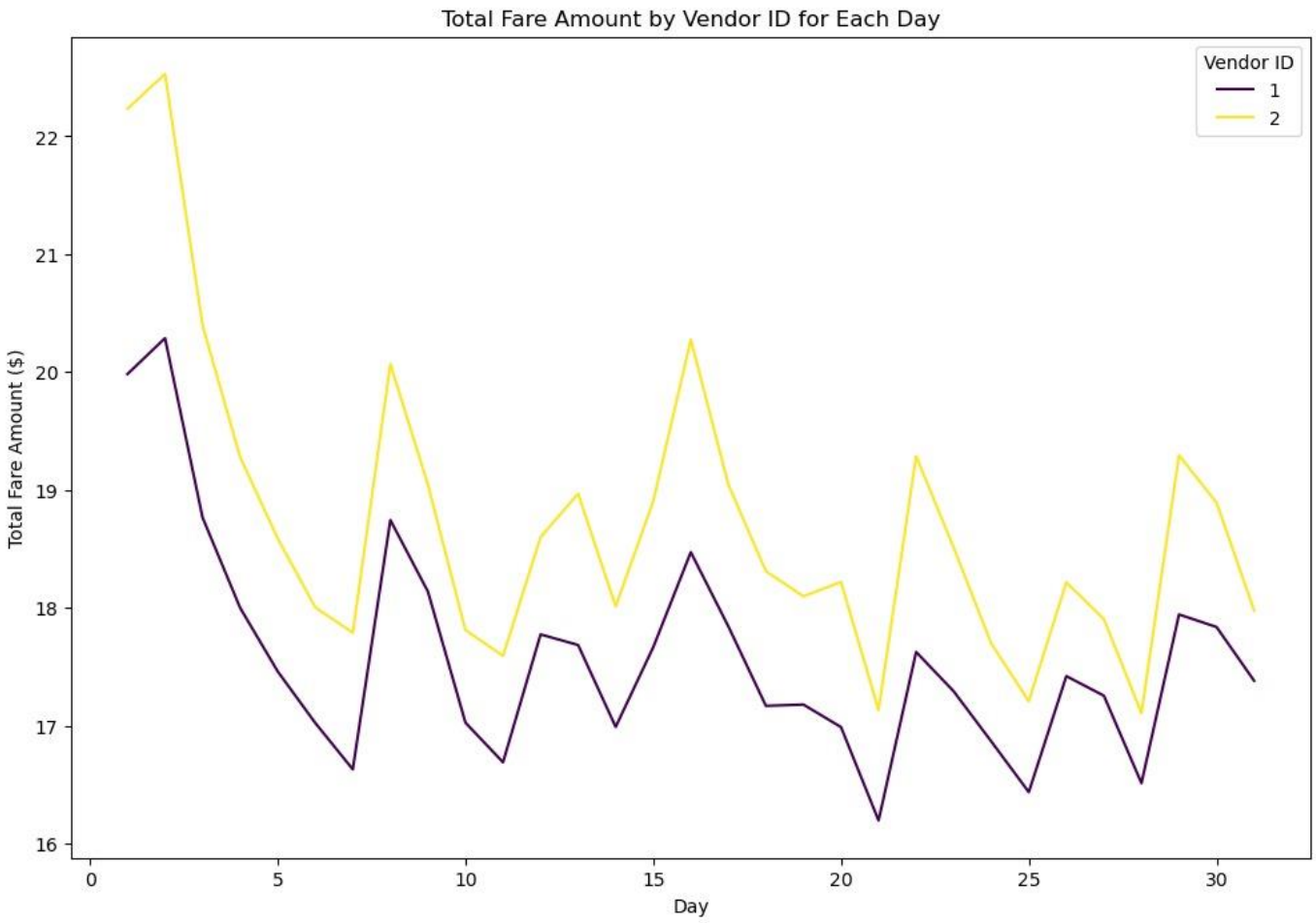
- *tpep_pickup_datetime : The date and time when the meter was engaged.*
- *tpep_dropoff_datetime : The date and time when the meter was disengaged.*
- *Passenger_count : The number of passengers in the vehicle.*
- This is a driver-entered value
- *Trip_distance : The elapsed trip distance in miles reported by the taximeter.*
- PULocationID : TLC Taxi Zone in which the taximeter was engaged.
- DOLocationID : TLC Taxi Zone in which the taximeter was disengaged.
- RateCodeID : The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride Store_and_forward_flag : This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server.
- Y= store and forward trip N= not a store and forward trip Payment type : A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip Fare amount : The time-and-distance fare calculated by the meter.
- Extra : Miscellaneous extras and surcharges. Currently, this only includes the 0.50 and 1 rush hour and overnight charges. Tip amount : Tip amount – This field is automatically populated for credit card tips. Cash tips are not included. Tolls amount : Total amount of all tolls paid in trip. Total amount : The total amount charged to passengers. Does not include cash tips. Congestion Surcharge : Total amount collected in trip for NYS congestion surcharge.

Graphs

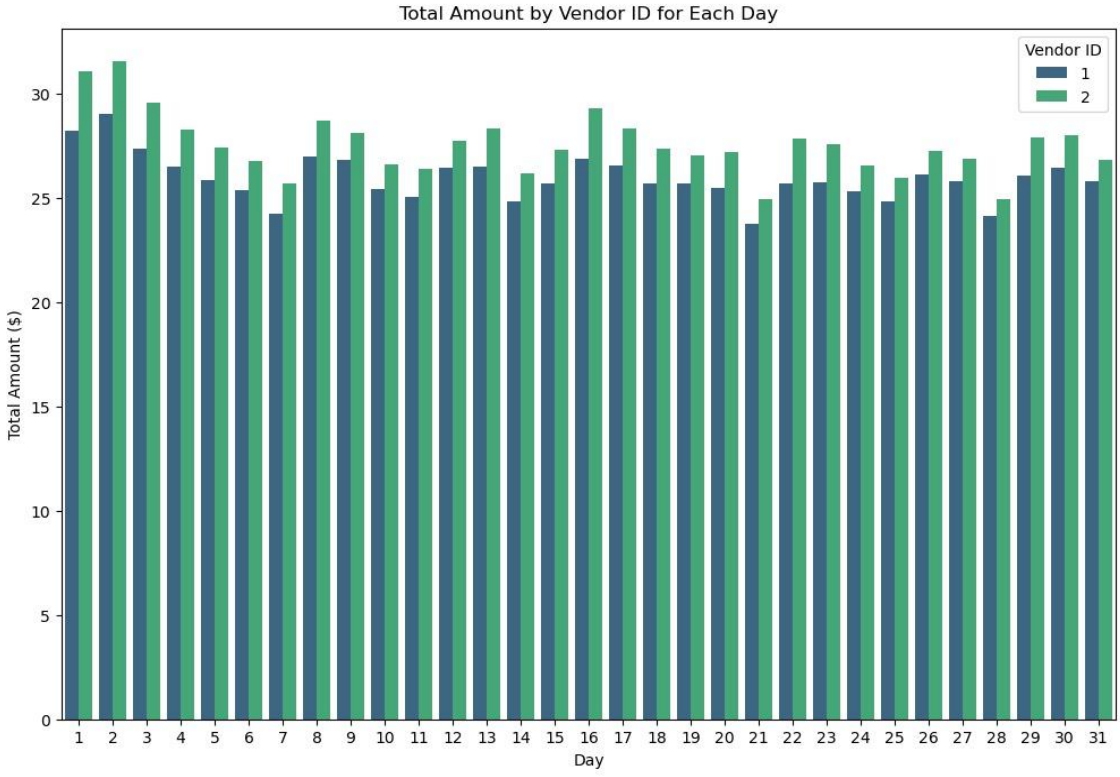
This graph shows the record count for vendorID.



1 = Creative Mobile Technologies, LLC; 2 = VeriFone Inc

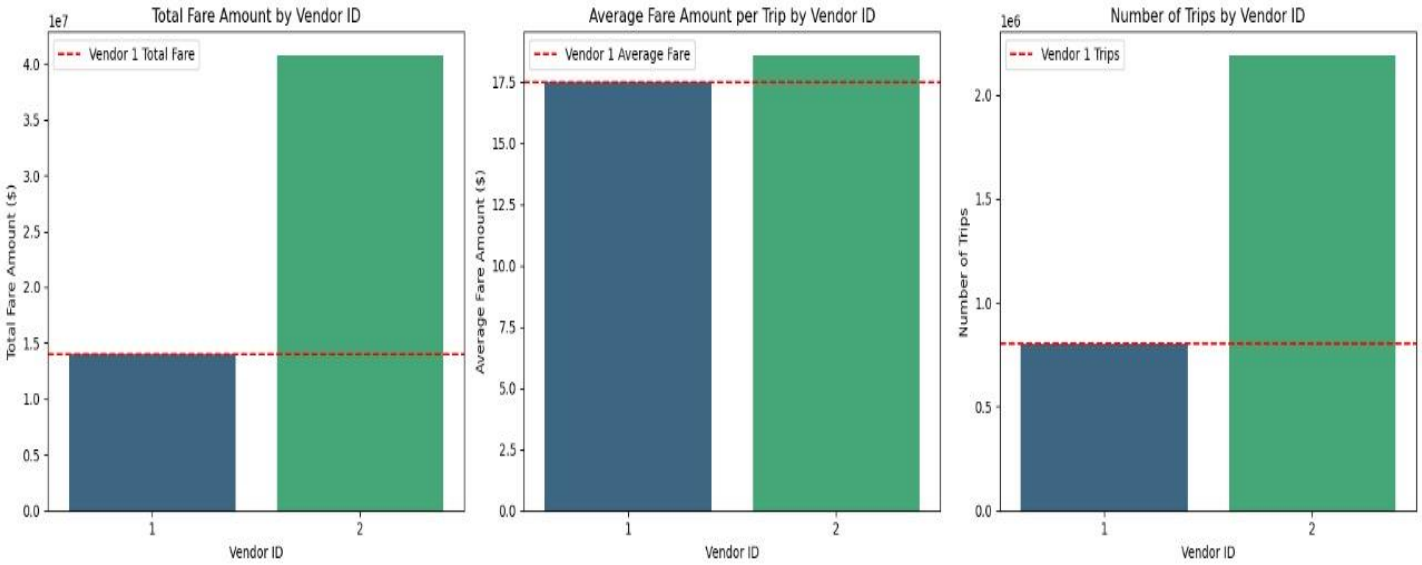


For this we Convert *datetime column to day.* This graph depicts the relationship between the Total Fare Amount by VendorID for Each Day for a month.



Comparison between VendorID 1 & 2

After that we average fare amount per trip for each vendorID, calculate the number of trips served by each vendorID, followed by calculating total fare amount and number of trips and plotting the comparison between them.



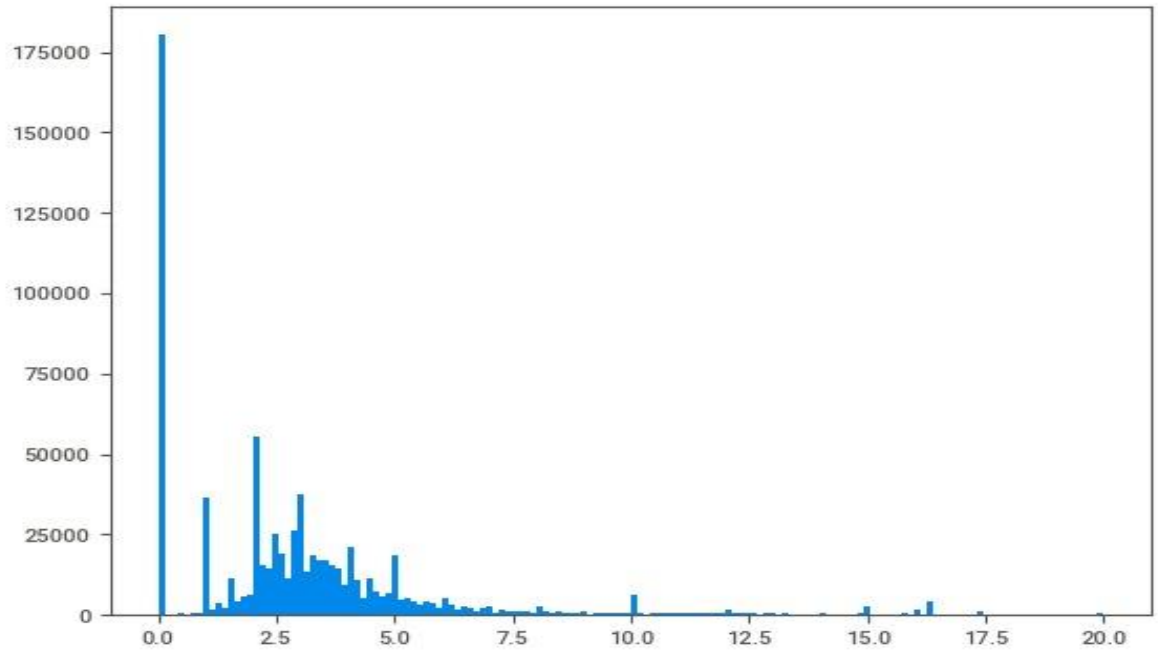
Total Fare Amount: Show that while VendorID 2 handles more trips, VendorID 1 generates a comparable total fare amount.

Average Fare Amount per Trip: Highlight that VendorID 1 has a higher average fare amount per trip compared to VendorID 2.

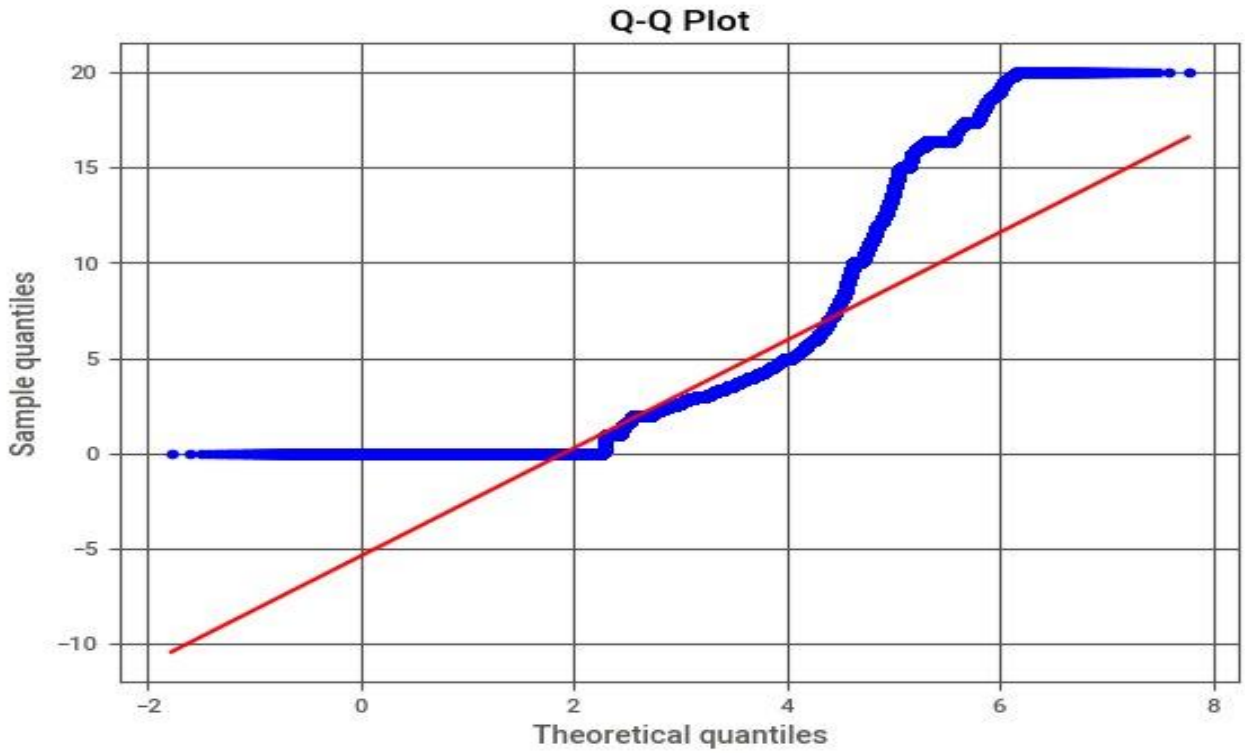
Number of Trips: Emphasize that although VendorID 1 serves fewer trips, it still generates a significant total fare amount due to the higher average fare per trip.

Tip Amount

After that we make a dataset of fare amount, extra, mta_tax, tip amount, tolls_amount, improvement_surcharge, total_amount' congestion_surcharge, airport_fee. Then we plot the density plot for tip amount with tip amount on x-axis and frequency on y-axis.

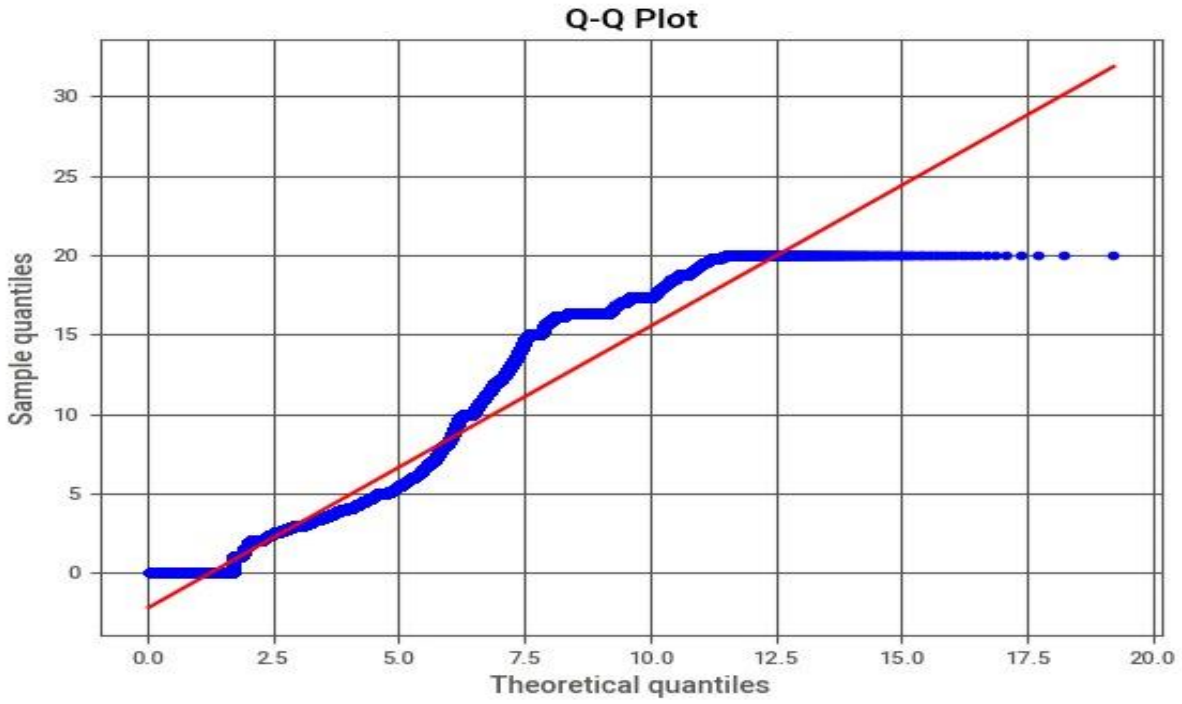


Then we generated the summary statistics for the tip amount and found out the confidence interval for the same.
After that we generated the Q-Q plot between the theoretical sample quantiles.



Blue line deviate from red so it doesn't follow normal distribution.

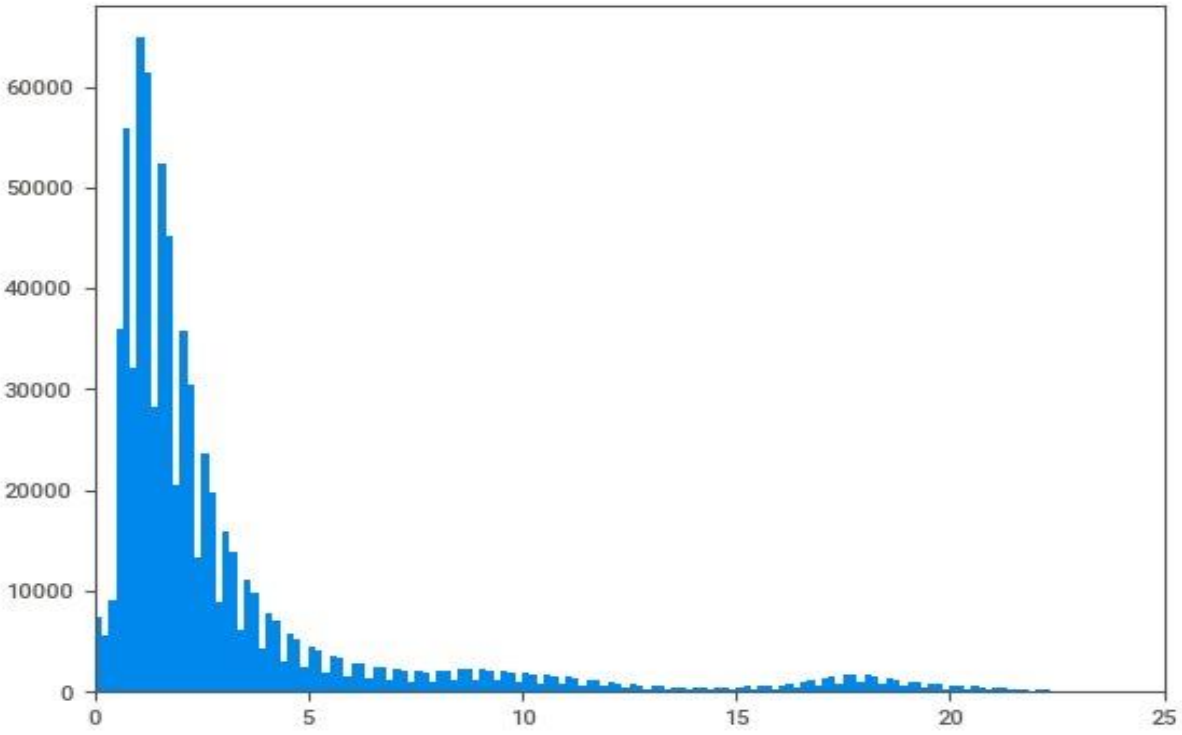
We assume that the tip amount follows gamma distribution with rate parameter 3.
Now we will check for that and plot another Q-Q Plot for the same.



By the above plot we can clearly see that, it nearly follows Gamma Distribution.

Trip Distance

Then we generated the summary statistics for the trip distance and found out the confidence interval for the same.
After that we generated the histogram with trip distance on x-axis and frequency on y-axis.

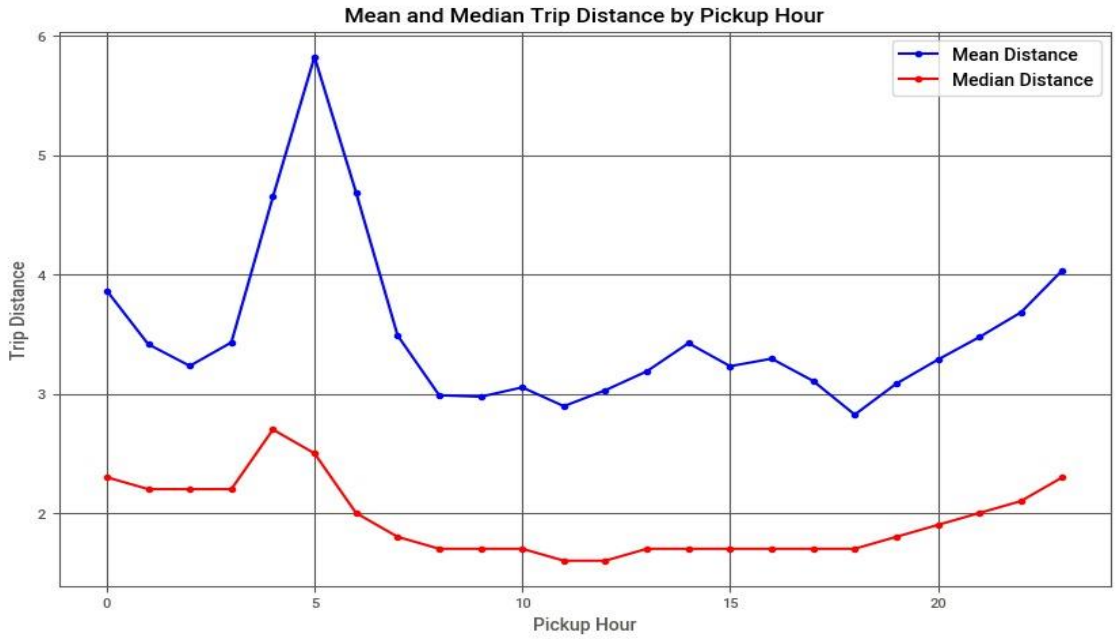


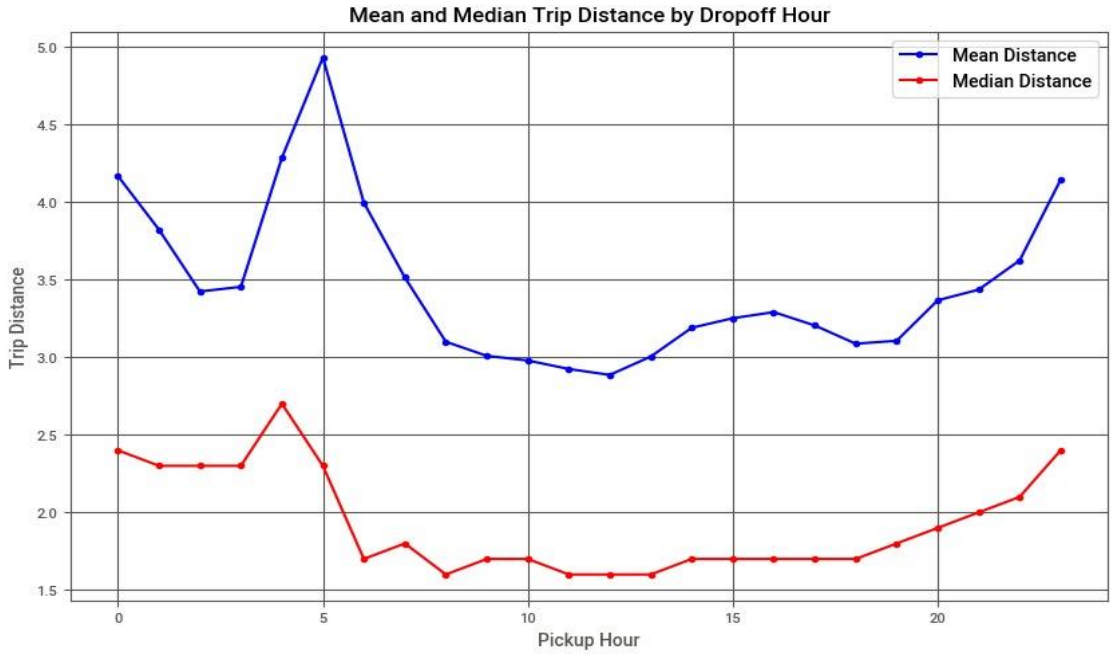
Then we generate the Q-Q Plot for it.

By the above graph we can see that it follows Gamma distribution.

Trip Pickup Time and Dropoff time

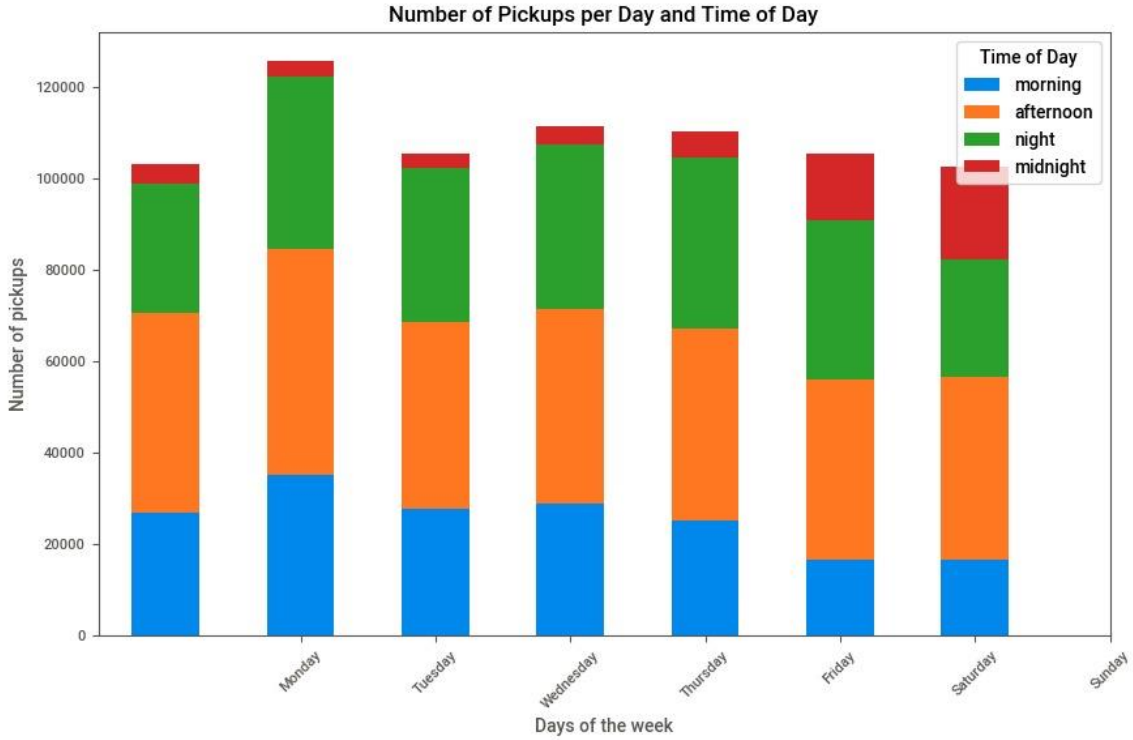
Assuming 'tpep_pickup_datetime' is a string representing datetime values we Extract the day of the week from the pickup datetime, followed by Separating data for weekdays and weekends, and Perform independent samples t-test, interpreting the results in the end.
After that convert the series to datetime dtype and separate date and time components.
We have a DataFrame named df with columns ‘pickup_datetime’ and ‘trip_distance’. Extract hour from ‘pickup_datetime’. Group by pickup hour and calculate mean and median trip distances then Plotting the line plot for mean trip distances and Plotting the line plot for median trip distances.





Trips started during 4:00 PM to 6:00 PM tended to be the shortest, and trips started between 4:00 AM and 6:00 AM were the longest. The surge in long-distance trips during the morning is likely driven by trips to the airports or other long-distance rides.

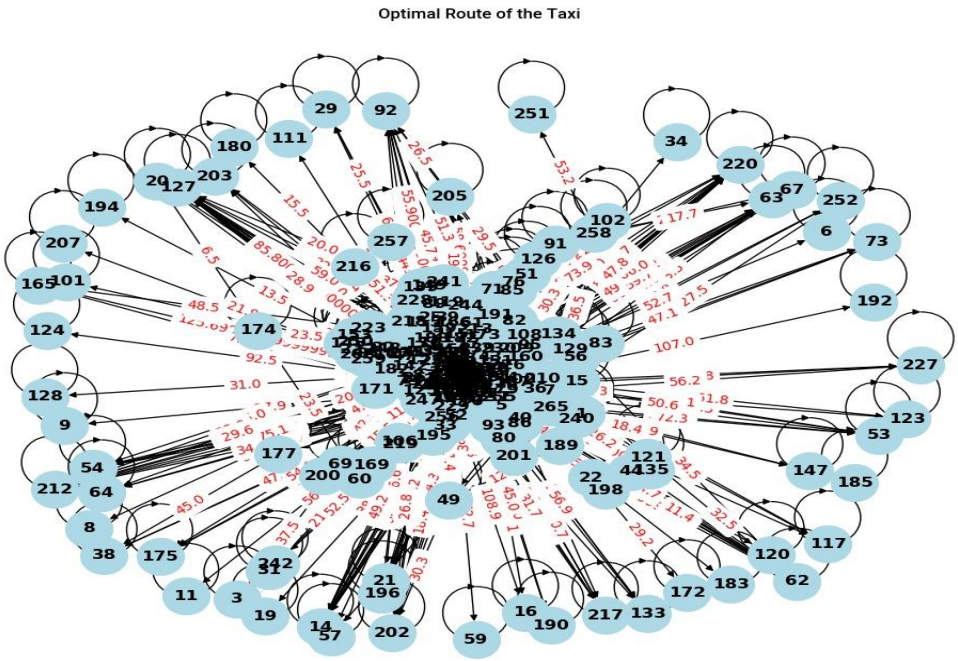
Pickup Amount



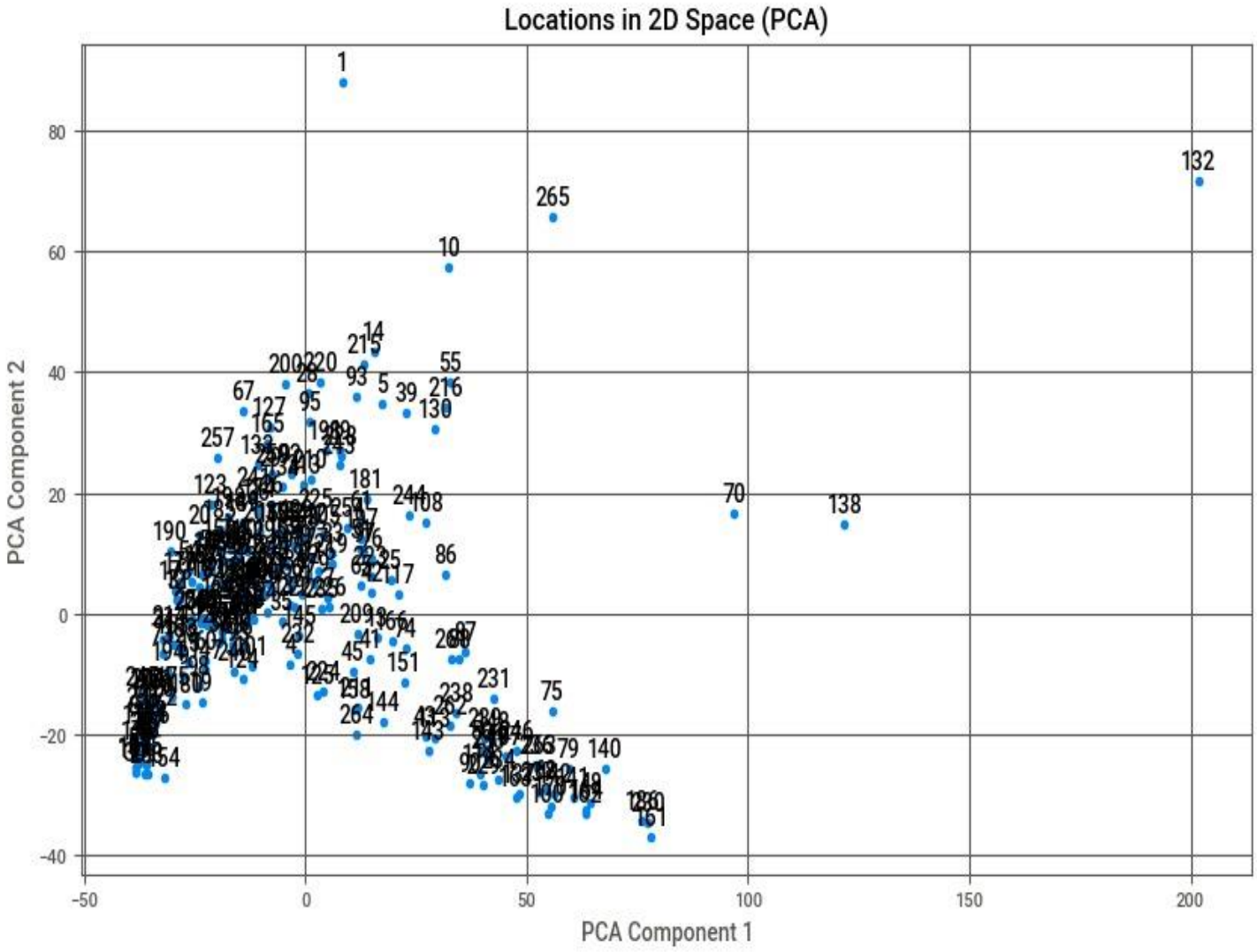
From the graph we see that, We get most no. of pickups during daytime and lowest during midnight for all the days of the week.

Optimal Route for Taxi

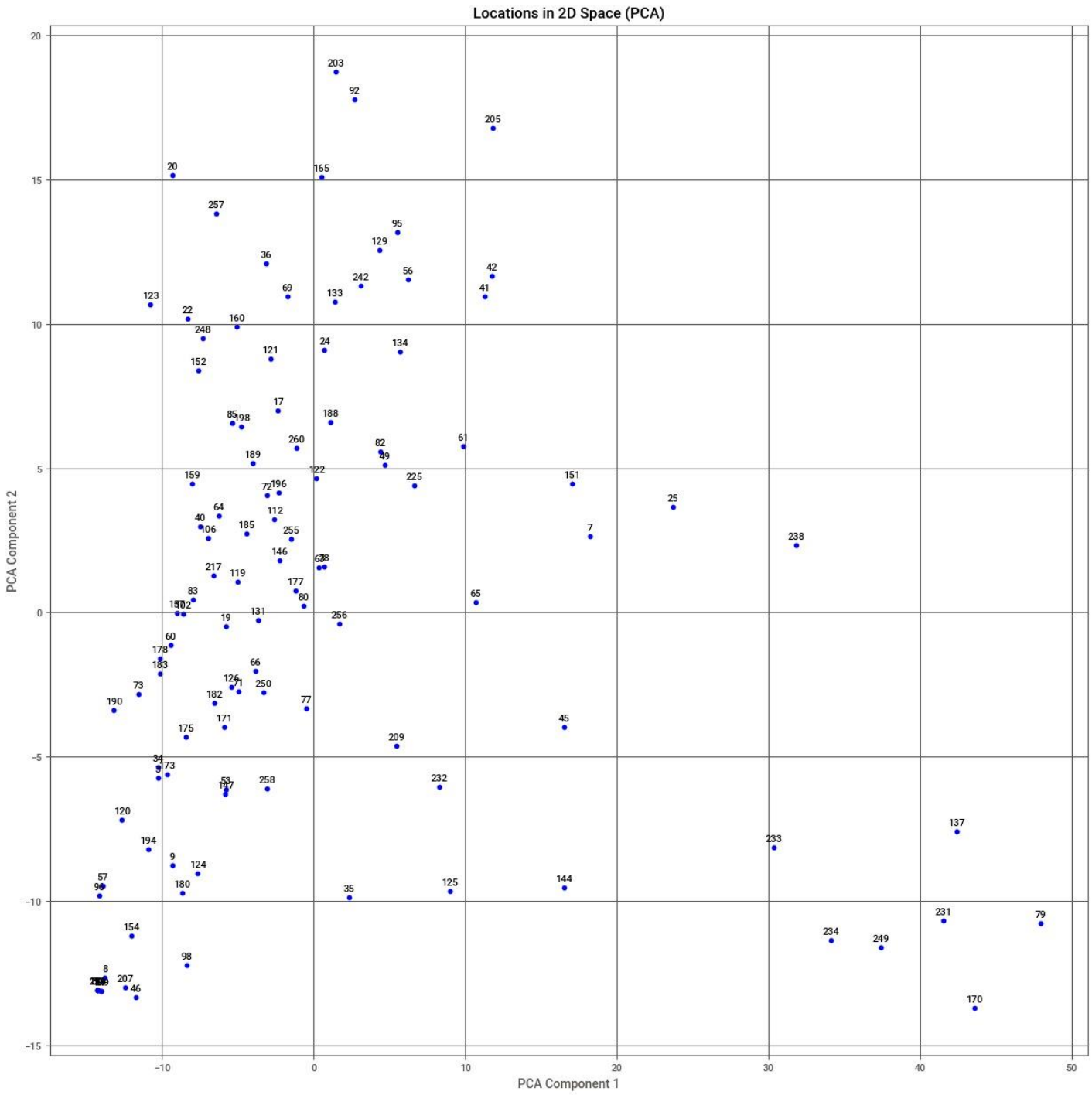
Next, we calculate the profit associated with each trip based on the fare amount. Finally, we initialize a linear programming problem to maximize the total profit. This process filters the Data Frame df to only include data for trips that occurred on the first day of the month. By adding the



profits from each journey, we were able to define the objective function that would maximize the overall profit. After doing some calculations we found the optimum Route for taxi.
Next, we visualization of taxi pickup and drop-off locations in a 2D space using Principal Component Analysis (PCA). We first create a distance matrix representing the distances between each pair of pickup and drop-off locations. Then, it applies PCA to reduce the dimensionality of the distance matrix to two components. These two components capture the most significant variations in the distances between locations. Finally, we plot the locations in the 2D space, with each location represented by a point, and annotates each point with its corresponding location ID. This visualization provides insight into the spatial distribution of taxi pickup and drop off locations, allowing for a better understanding of the geographical patterns within the dataset.



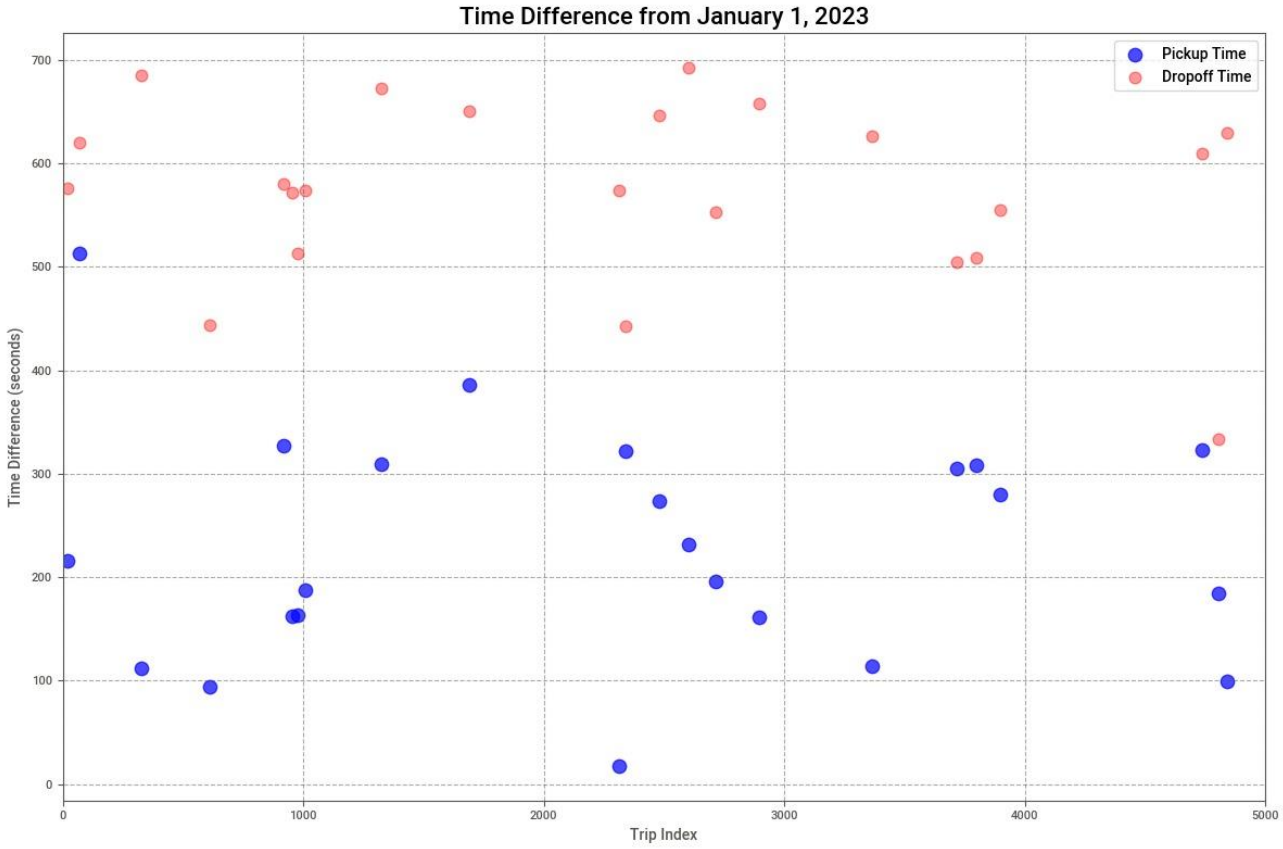
Now we present a refined visualization of taxi pickup and drop off locations in a 2D space using Principal Component Analysis (PCA), with outliers removed. This visualization provides a clearer and more accurate representation of the spatial distribution of taxi pickup and drop off locations, aiding in understanding geographical patterns within the dataset.

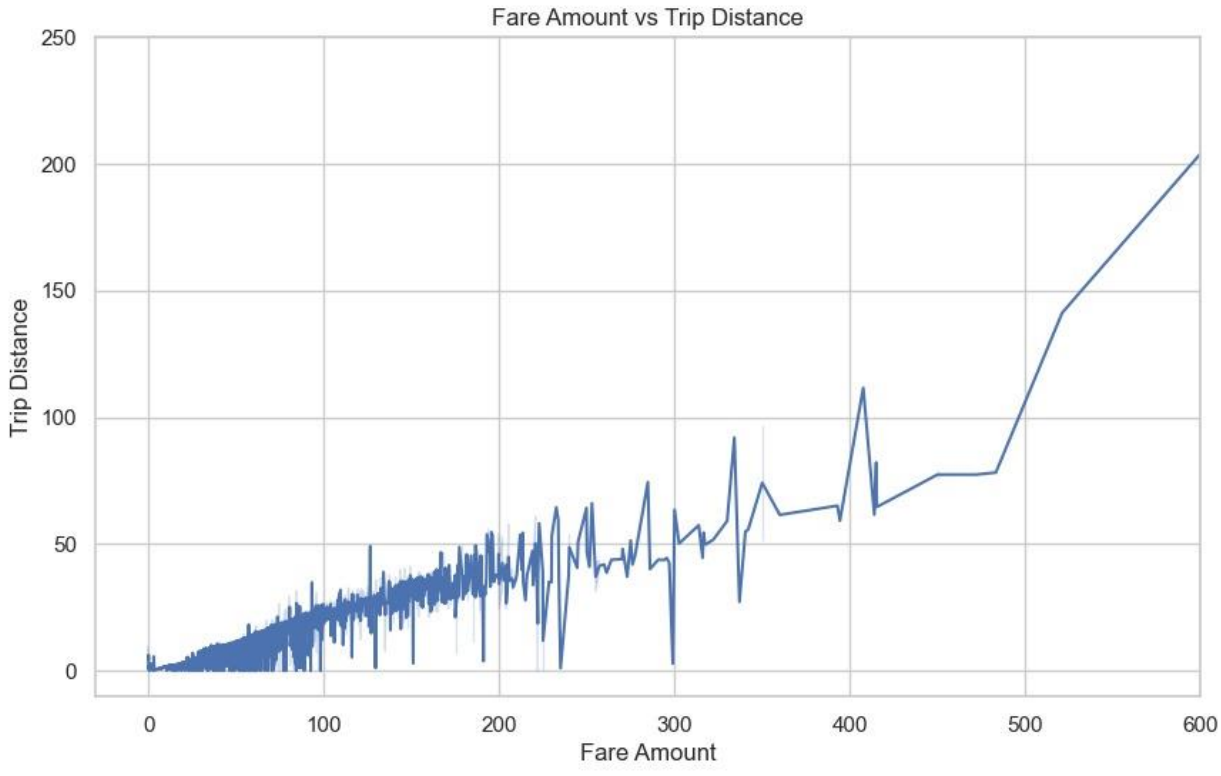


Now we try to find a model for this. After that we begin training of a linear regression model to predict the drop off location ('DOLocationID') based on the pickup location ('PULocationID') from the provided dataset. Next, we initializes and fits a linear regression model using the training data. After fitting the model, we makes predictions on the testing set and calculates the Mean Squared Error (MSE) to evaluate the model's performance. The printed MSE value provides insight into how well the model generalizes to unseen data, with lower MSE indicating better predictive performance.

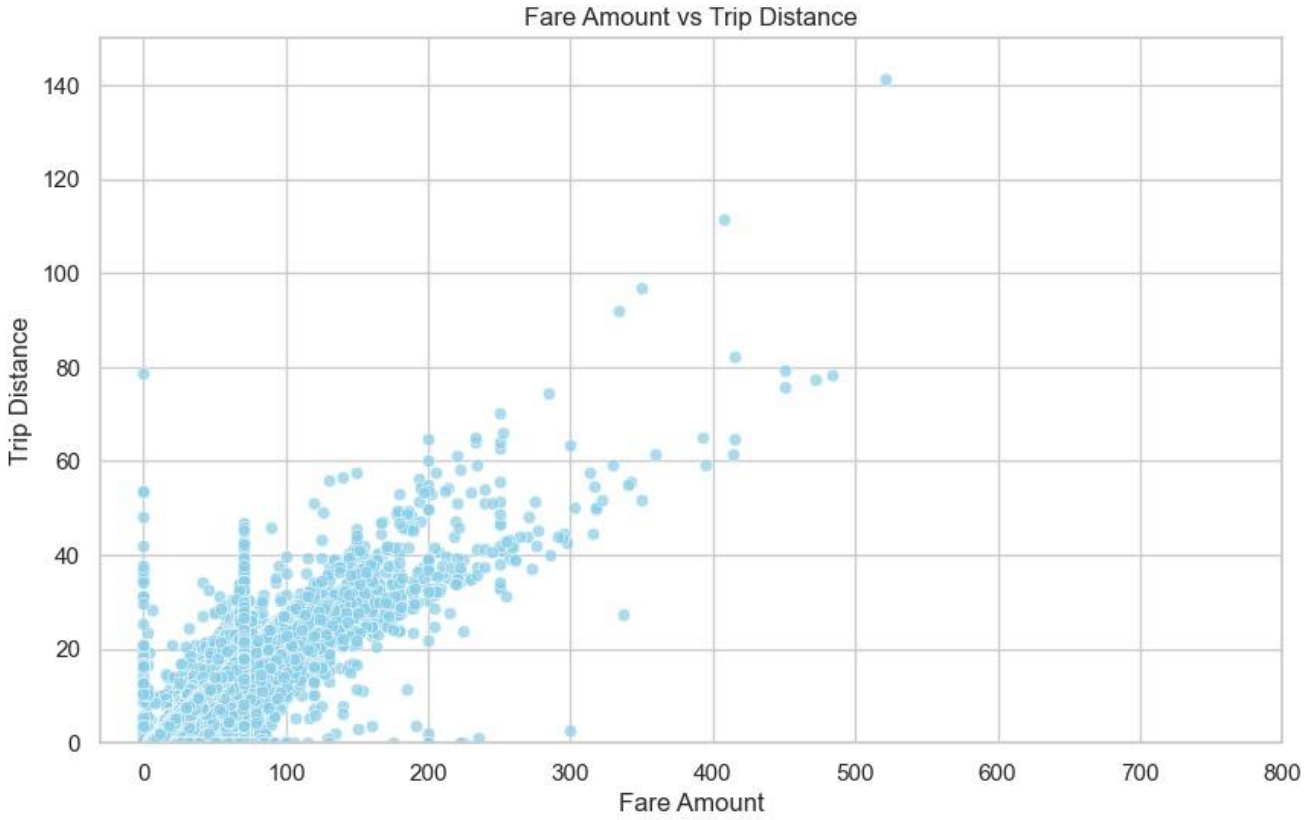
Time , Distance , Fare Amount

The below scatter plot depicts the temporal distribution of taxi trips, with blue dots representing pickup times and red dots representing drop off times. This visualization allows for the observation of patterns and trends in the timing of taxi pickups and drop-offs on the specified day, aiding in understanding temporal dynamics within the dataset.

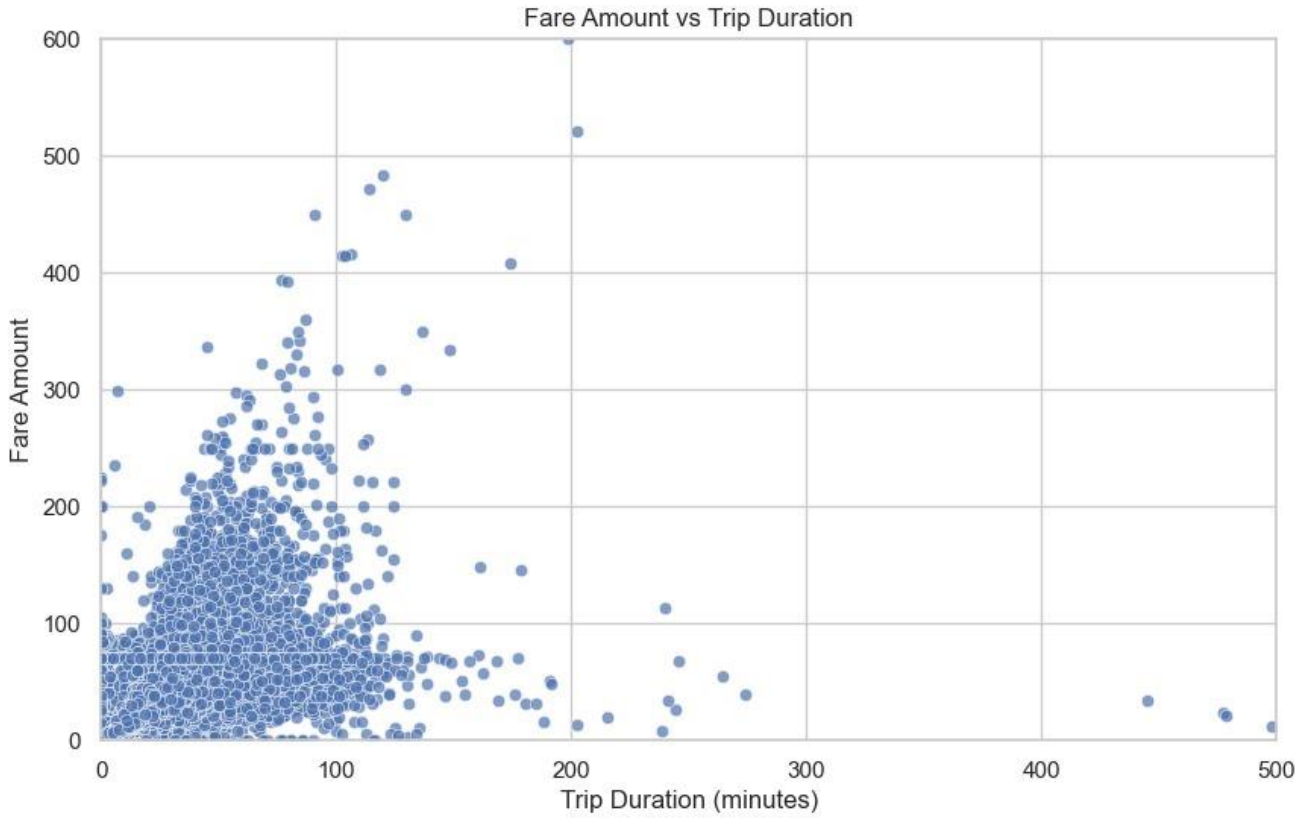




Now we made the scatter plot, individual data points are plotted, allowing for better visibility of individual data points and any potential outliers.

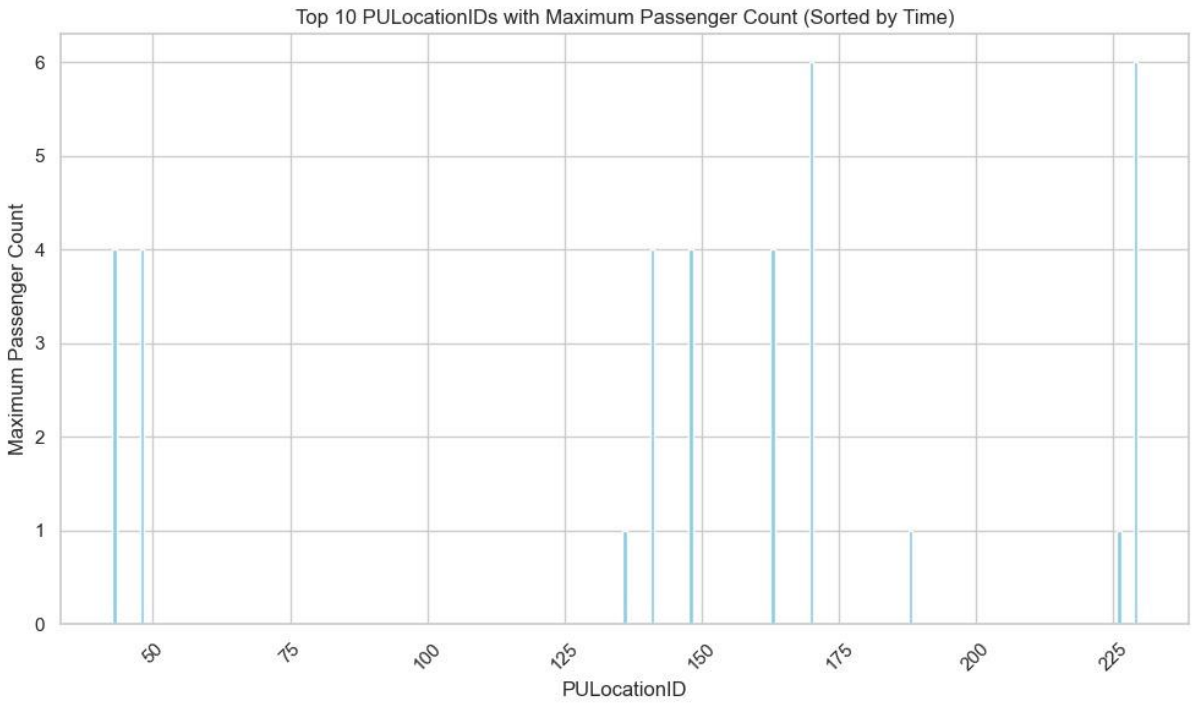


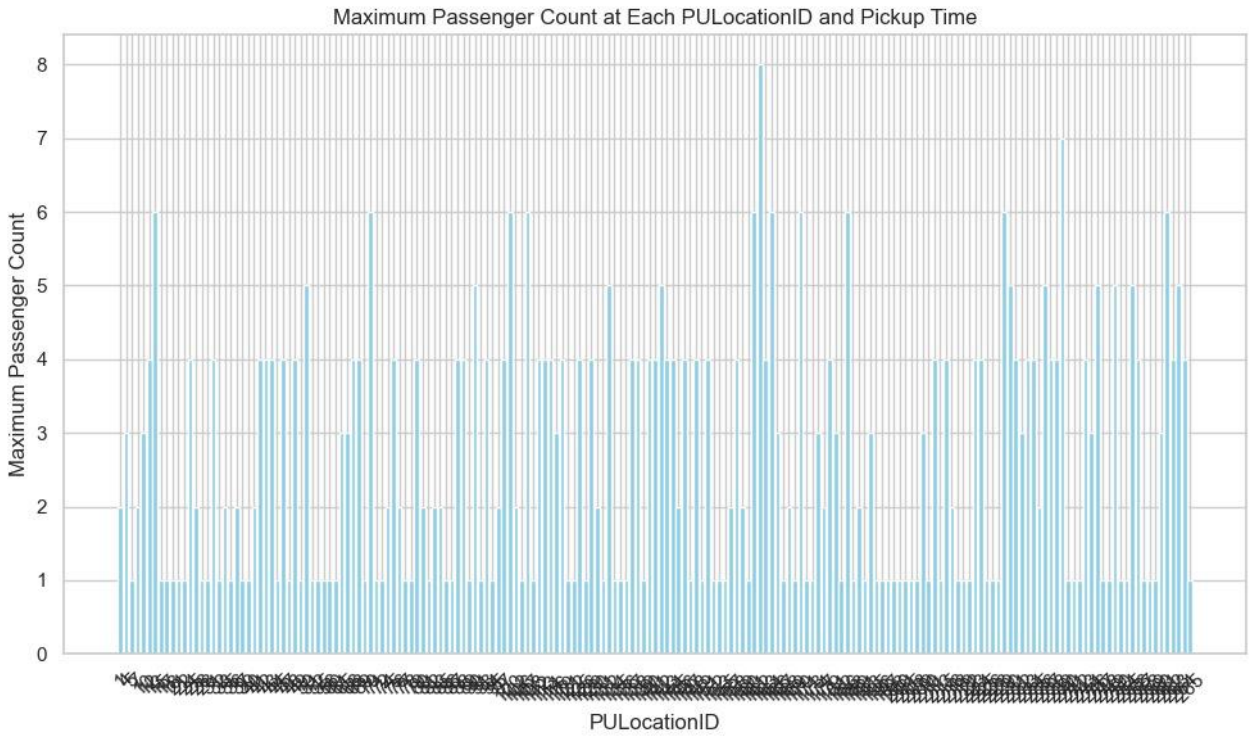
The scatter plot offers a more granular view of individual data points and potential outliers, while the line plot provides a smoother representation of the trend between the variables.



Finding Busiest Pickup Location And Time

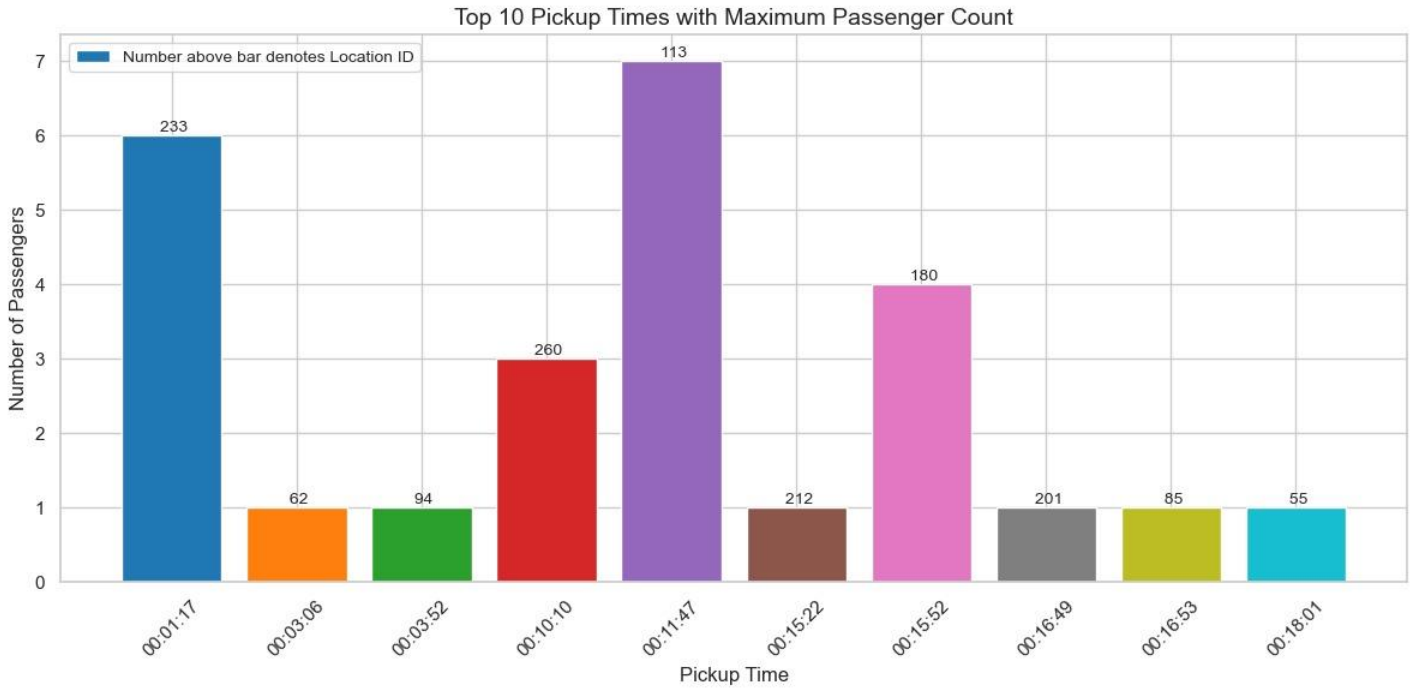
We analyse taxi pickup data for January 1st. It identifies the pickup time with the maximum total passengers for each pickup location and visualizes this information through two plots. The first plot displays the maximum passenger count at each pickup location, while the second plot shows the top N pickup locations with the highest maximum passenger counts, sorted by pickup time.



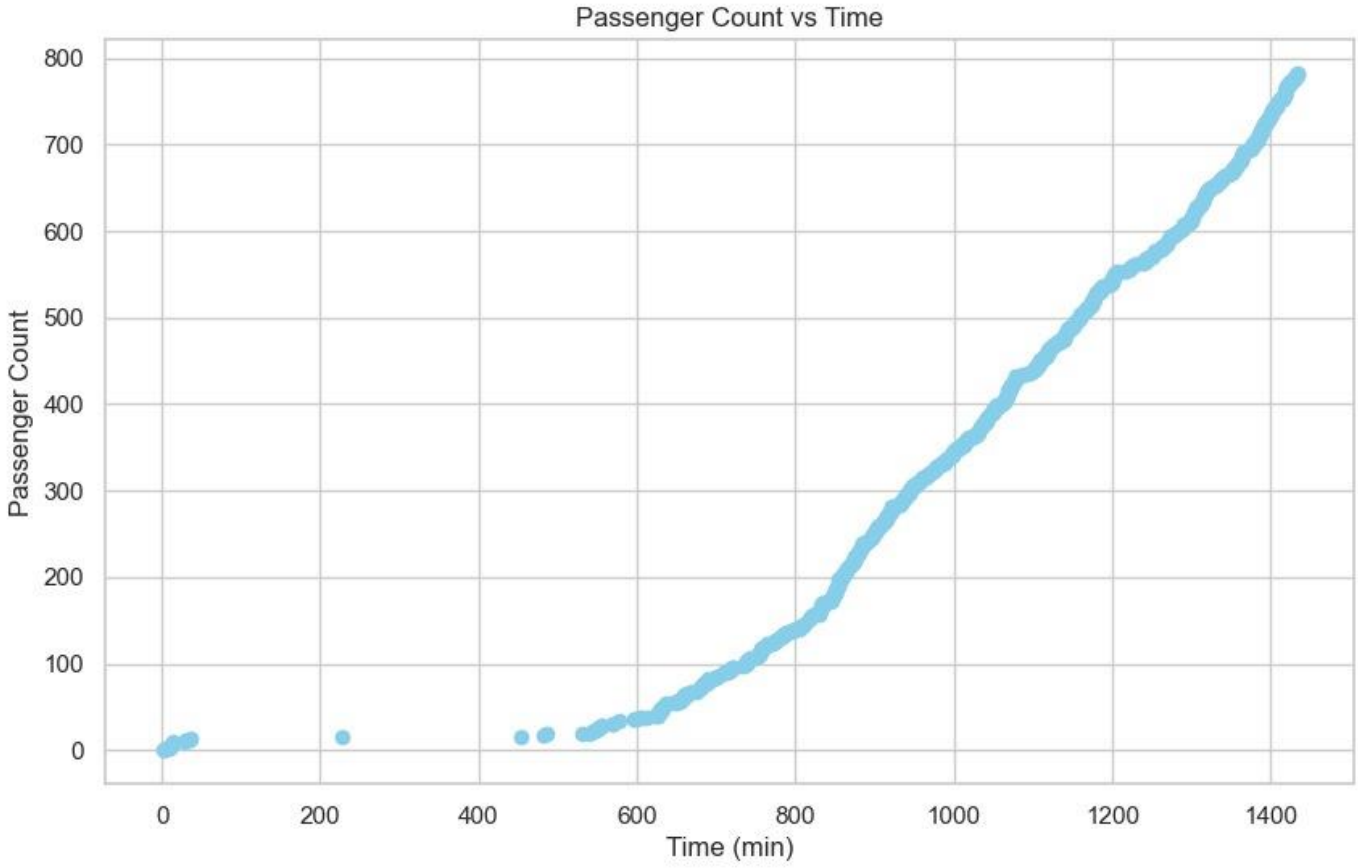


These visualizations help identify the busiest pickup locations and times, providing insights into areas and periods of high demand for taxi services.

Next, we aim to find the best busiest pickup times. This visualization helps to understand the fluctuating patterns of passenger demand across multiple pickup locations.



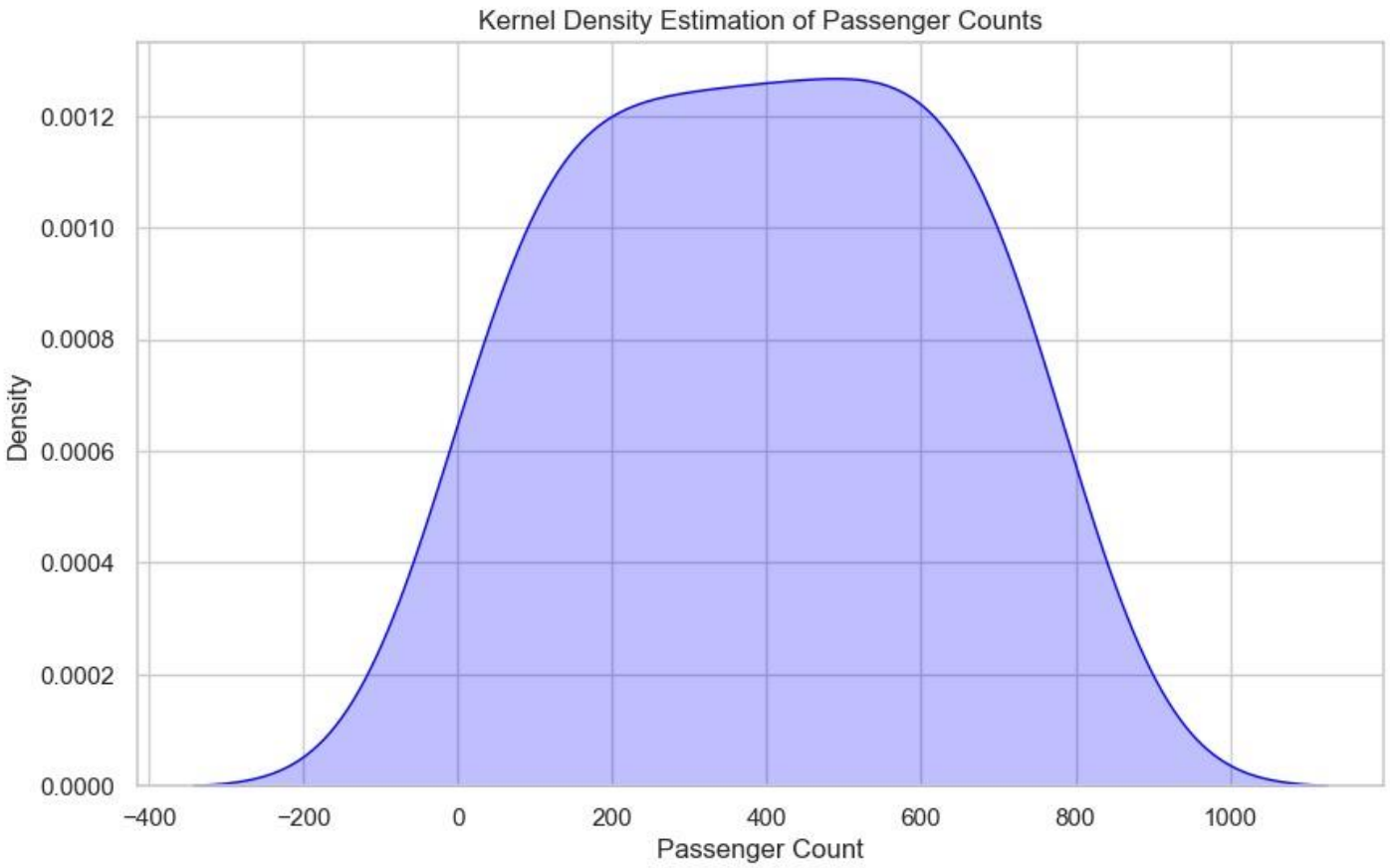
We analyse the passenger count dynamics over time at a specific pickup location on 15th Jan. calculates the time difference in minutes from a reference time (January 15, 2023) for each pickup, and then plots the cumulative sum of passenger counts against time.



For Poisson Process following condition must be satisfied:

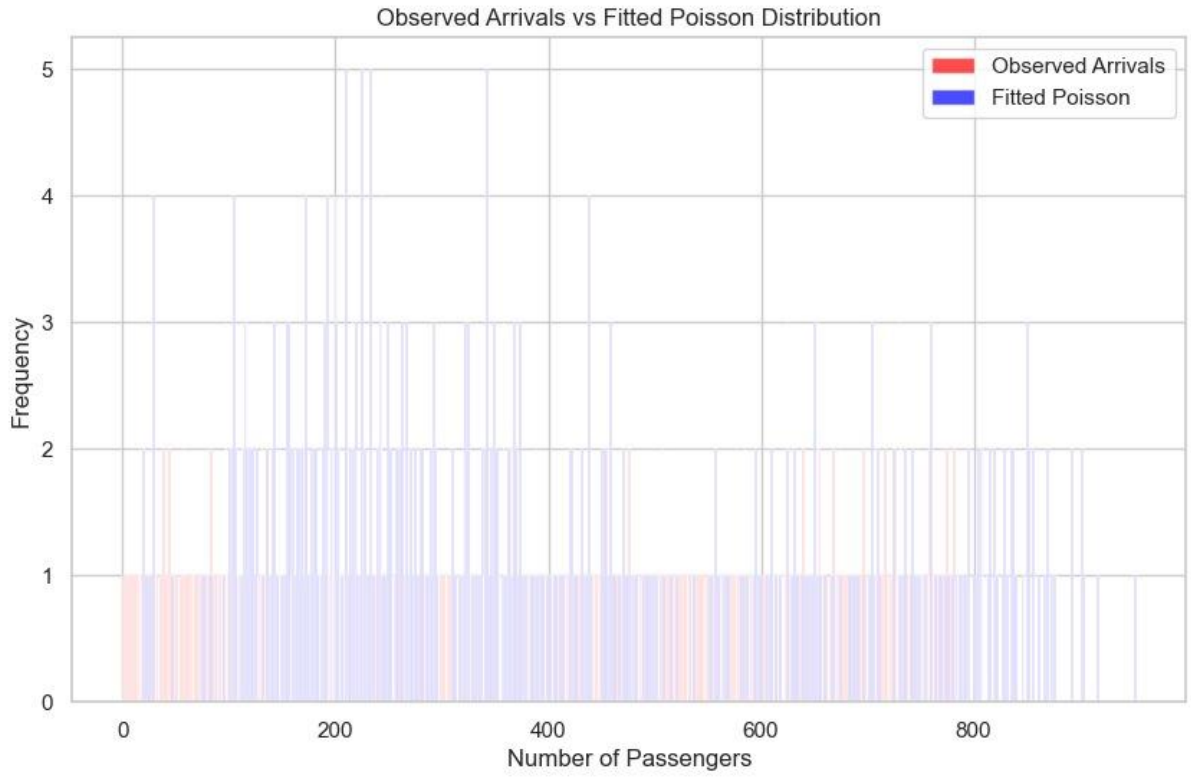
1. This includes events occurring independently within fixed intervals.

2. Constant rate of occurrence across different intervals.

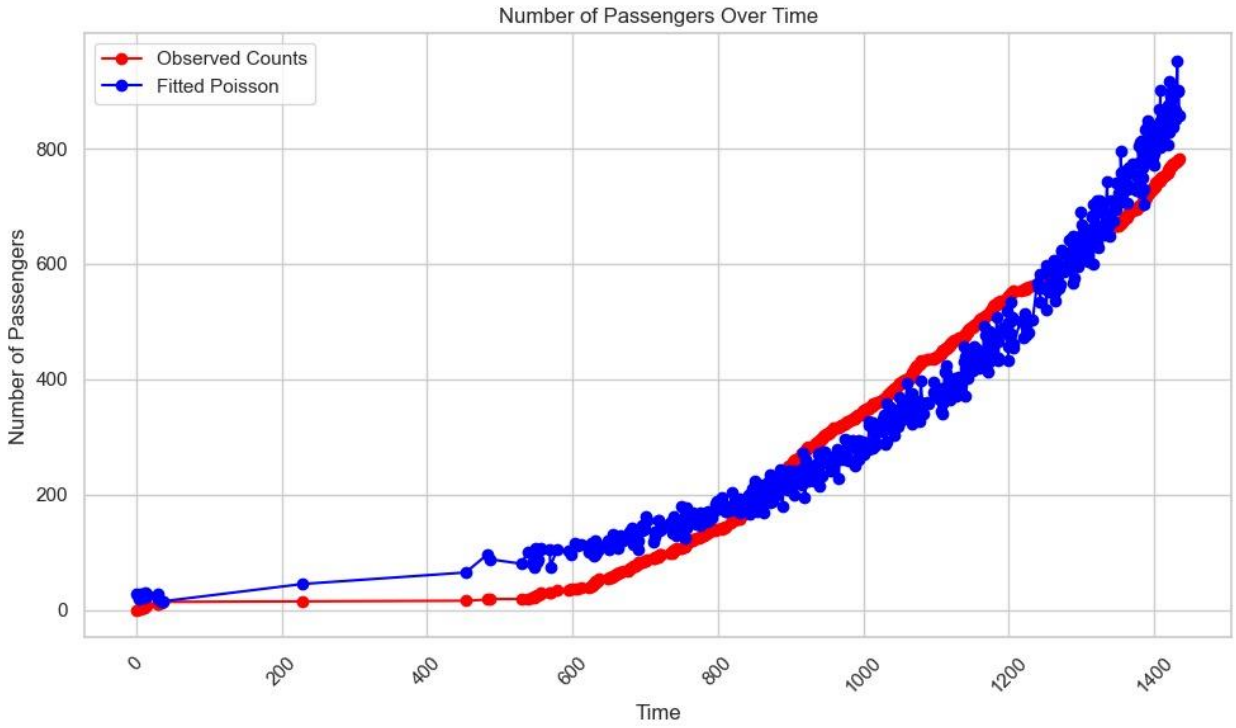


Based on this KDE plot, the distribution of passenger counts at pickup location 138 on January 15th seems to be relatively symmetric around a central value, indicating a consistent pattern in the number of passengers per pickup at this location on that particular day.

We evaluate the Poisson regression model's a good fit by comparing observed passenger counts to those predicted by the fitted Poisson distribution. This comparison helps to determine how well the Poisson model describes the distribution of passenger counts at the specified pickup location and time.

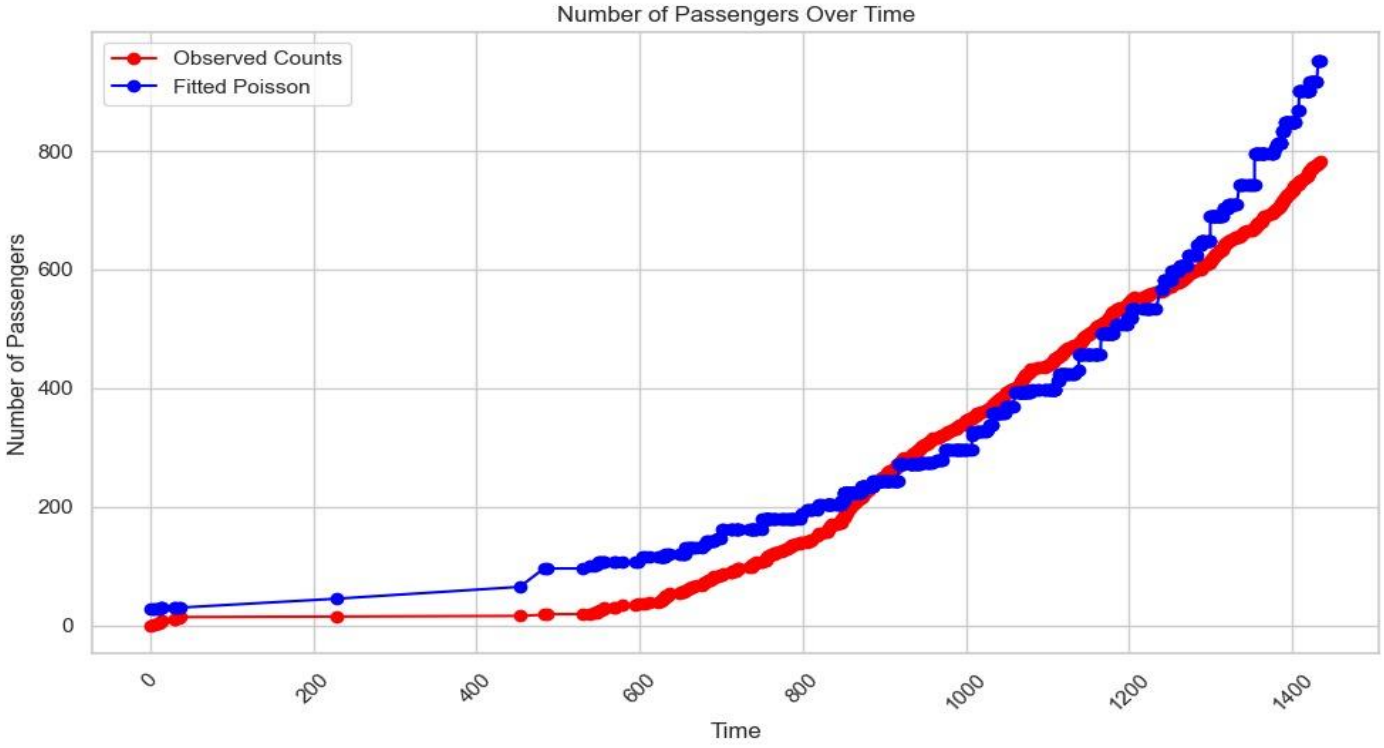


Then we plot a line graph showing the observed counts of passengers over time alongside the fitted Poisson distribution.

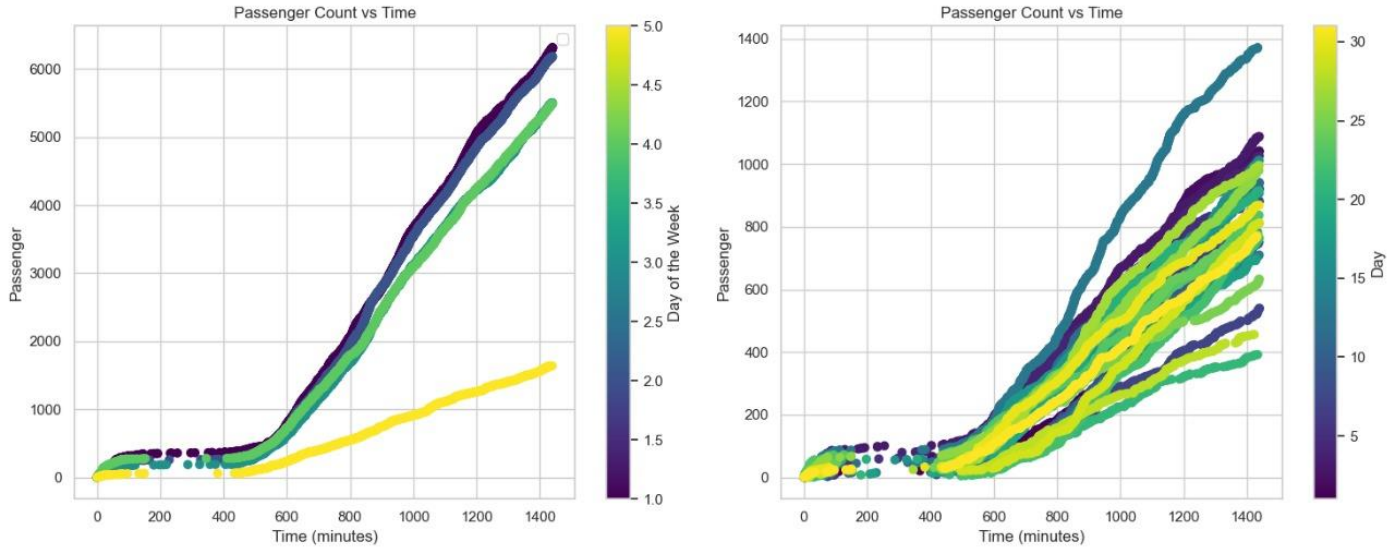


this plot visually compares the observed passenger counts over time with the passenger counts predicted by the fitted Poisson distribution, allowing for an assessment of how well the Poisson model describes the passenger count dynamics at the given pickup location and time.

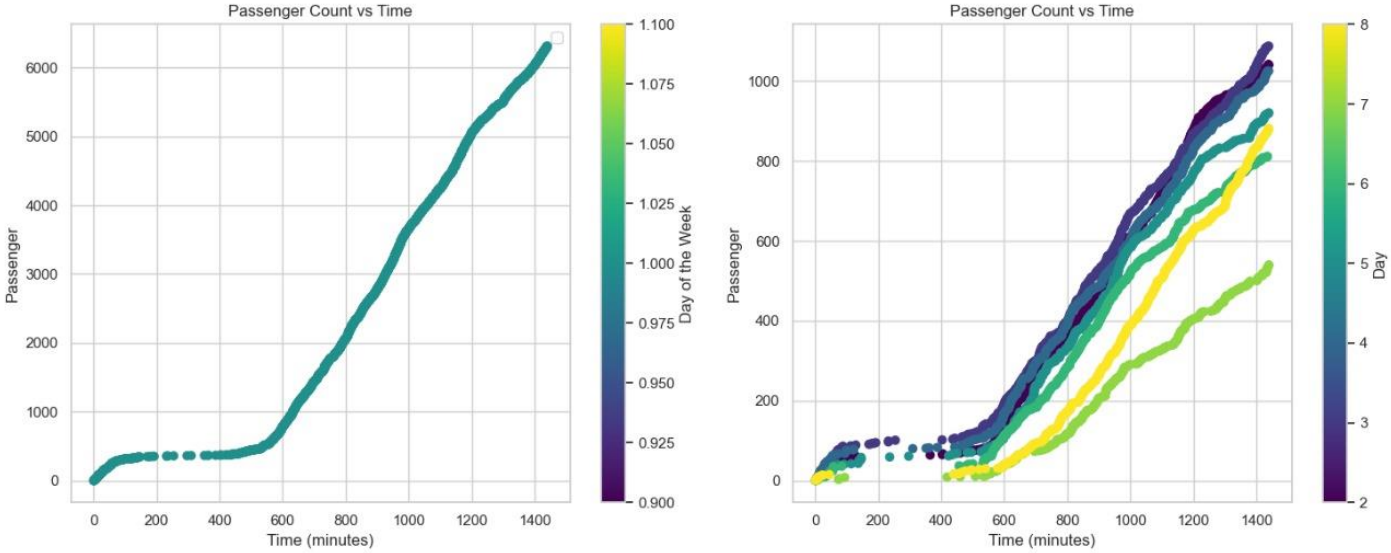
The previous graph iterates through the fitted Poisson distribution, updating each value to the maximum of either the current or previous value. This effectively ensures that the fitted Poisson distribution does not decrease with time, resulting in a non-decreasing trend. We then plot the observed passenger counts over time against this adjusted fitted Poisson distribution. In contrast to the previous graph , which simply depicted the fitted Poisson distribution without any adjustments, this updated code ensures that the fitted Poisson distribution maintains a non-decreasing trend overtime.



This graph depicts the dynamics of passenger count over time at the pickup location, taking data in terms of weeks. It generates side-by-side scatter plots in which the colour of each data point represents the day of the week. This allows for the comparison of passenger counts on different days of the week, as well as their variation over time at the specified pickup location.

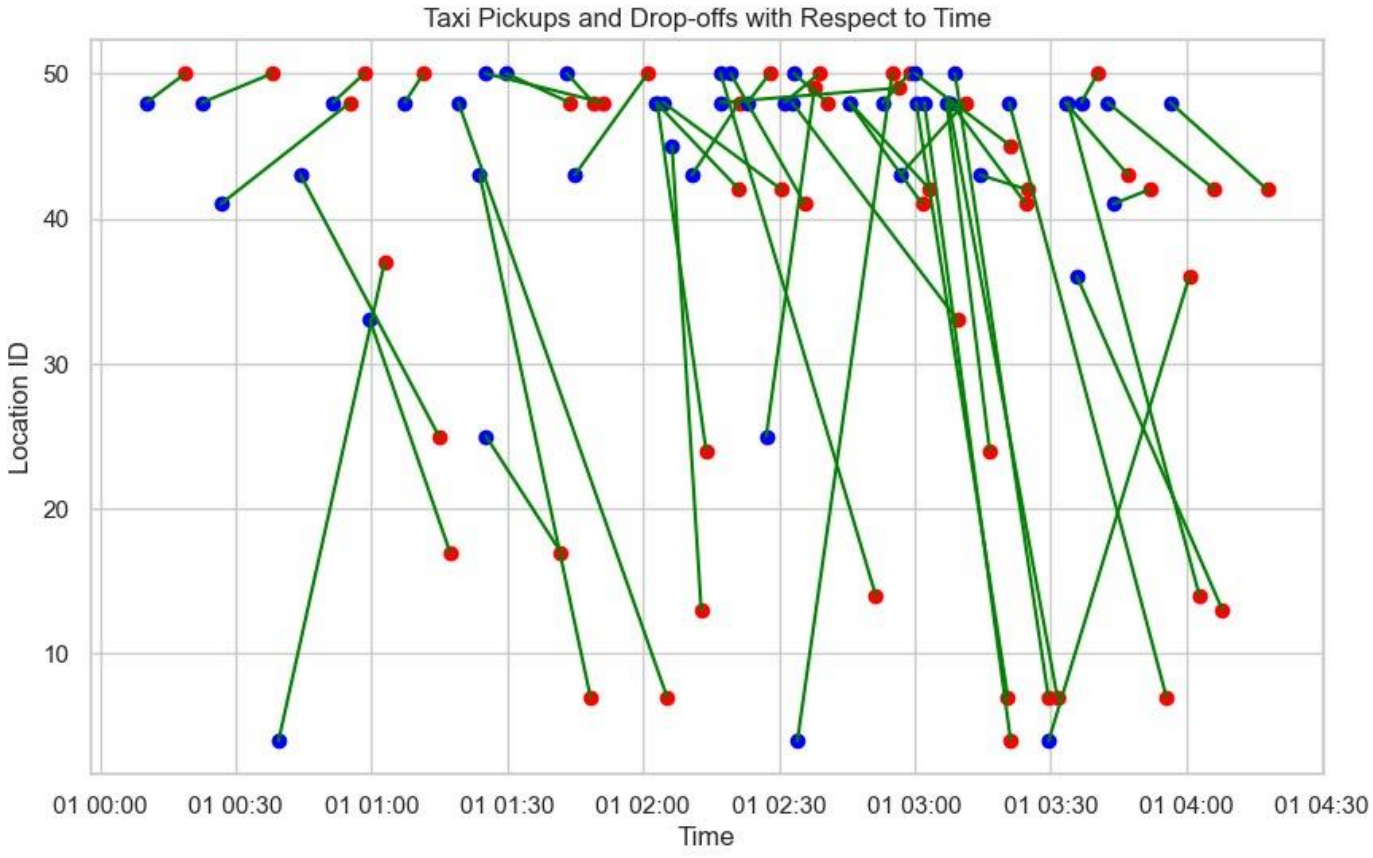


Compared to the previous graph, this one provides a more focused view by zooming in on the passenger count dynamics for a specific week, allowing for a more in-depth examination of the variation in passenger counts over time during that week. It also maintains the comparison aspect by displaying the selected week alongside the data from the previous graph.



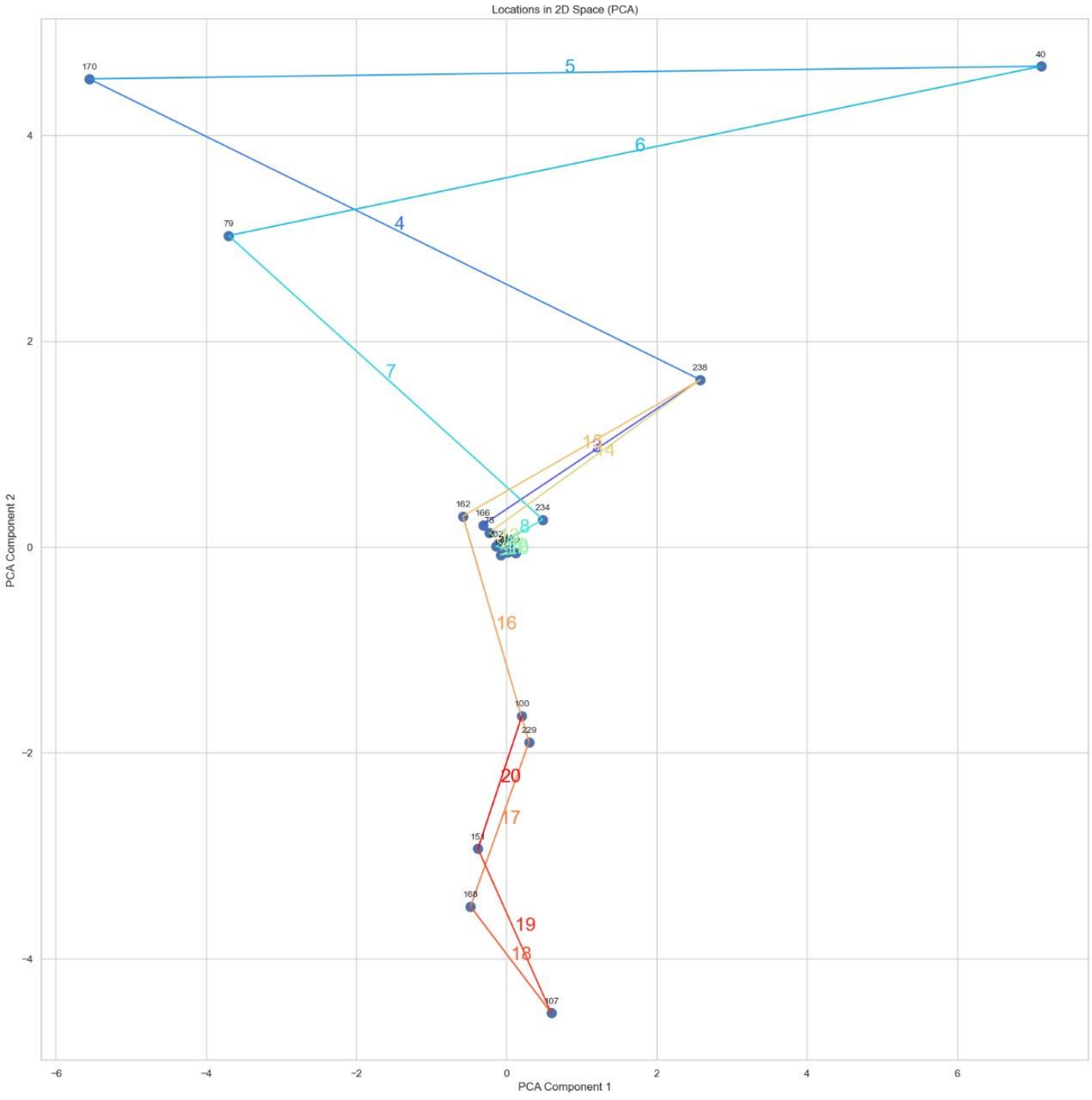
Visualizing taxi pickups and drop-offs with respect to time and location

This visualization provides insights into the temporal and spatial distribution of taxi pickups and drop-offs, allowing for the analysis of taxi patterns and routes over time.



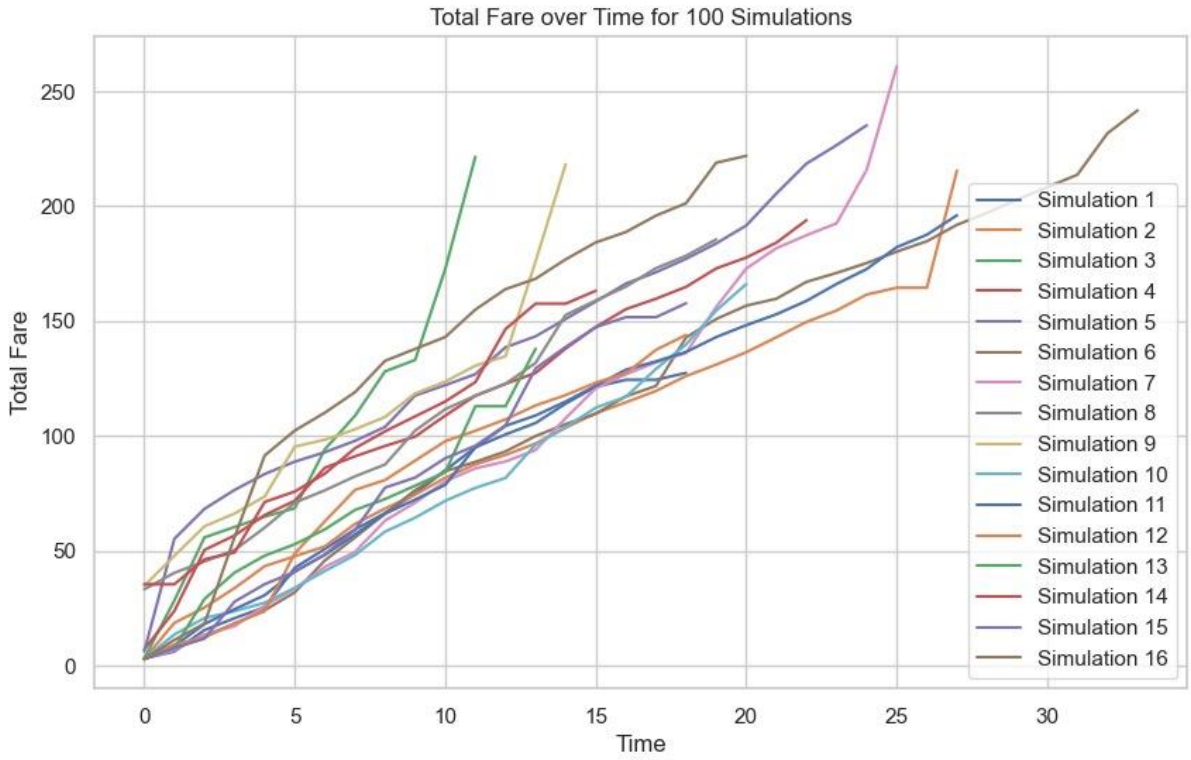
Simulation

we are simulating taxi routes and passengers' pickup and drop-off locations based on a Poisson process model overall simulates taxi operations, estimates passenger counts, generates taxi routes, and visualizes the simulated routes and passenger activities in a 2D space.

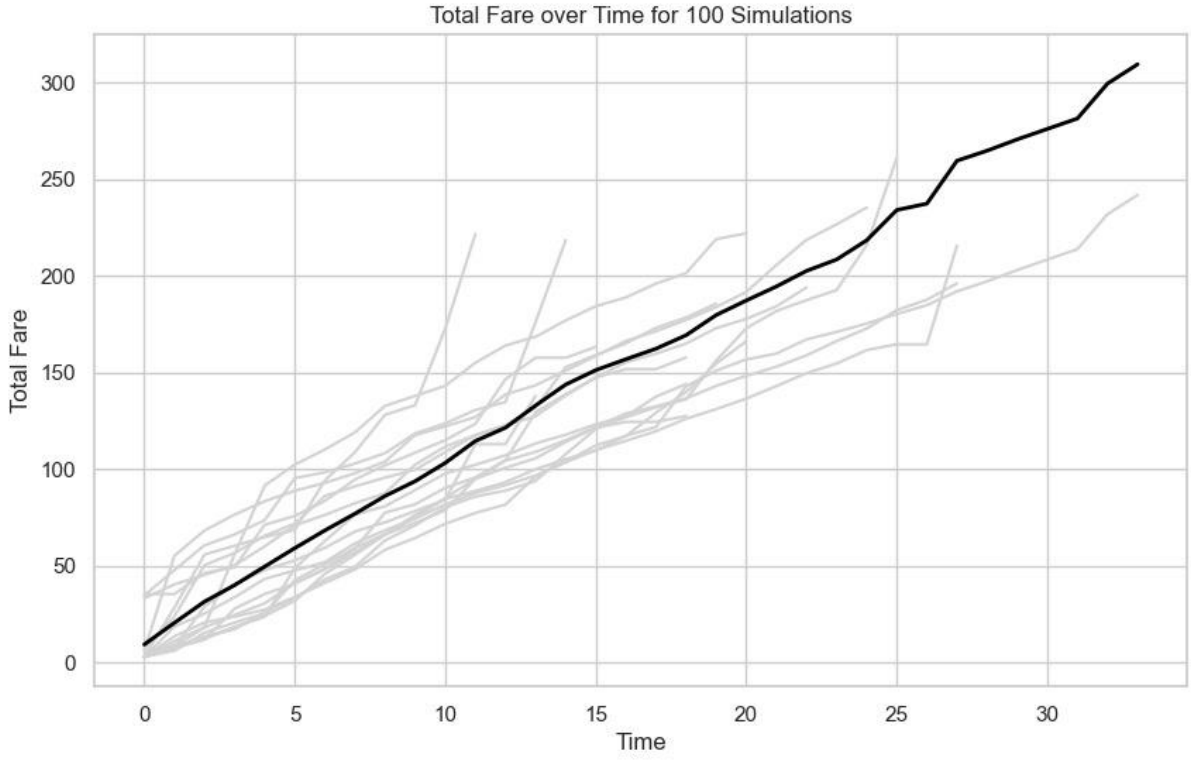


Each route is plotted with a corresponding route number, and the routes are color-coded based on their sequence.

Now we simulate taxi operations with dynamic pricing, considers passenger demand, and attempts to optimize profit using linear programming, providing insights into the profitability of taxi services under varying conditions.



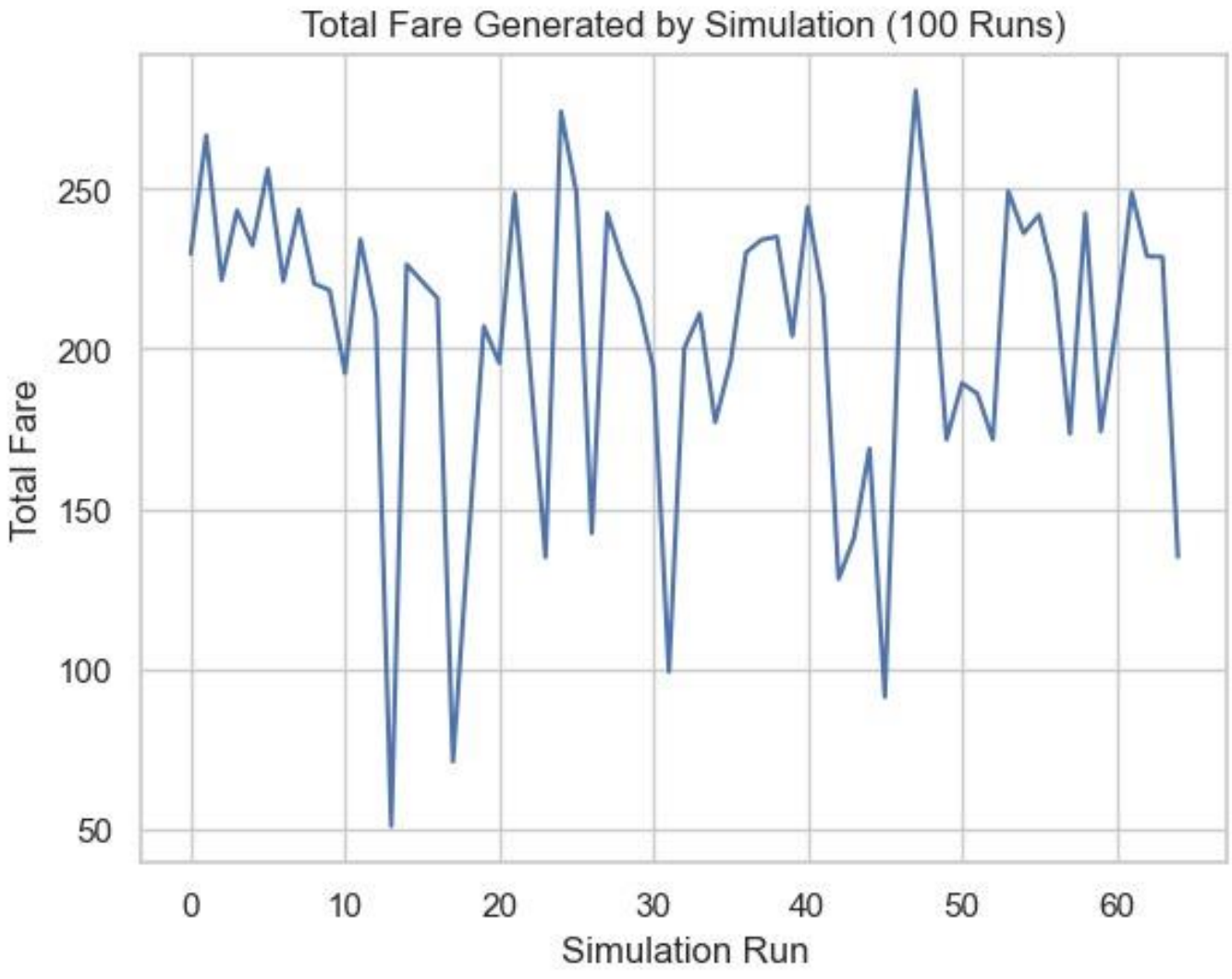
This plot shows the total fare over time for 100 simulations to visualize the variability in earnings across different simulations.



The above plot attempt to display the average simulation scenario.

Mean Revenue

We conduct further analysis on the results of a taxi simulation. Initially, we combine the outcomes of multiple simulation runs into a single Data Frame, enabling aggregated insights. Through grouping we calculated the total fare generated at each time step and computes the mean fare across all simulation runs. Subsequently, we proceed to individually run the simulation 100 times, calculating the total fare generated in each run and storing these values in a list. To visualize the variability in total fare across different simulation runs, we plot the total fare for each run. These analyses offer valuable insights into the average fare trends over time and the variance in total fares among different simulation instances, contributing to a deeper understanding of the simulation outcomes and their reliability.



From the above graph we can say that our mean revenue is 193$.

Conclusion

Using the output of our simulations we have drawn following conclusions: -

1. Mean Revenue is 193$/Day (Single Taxi)
2. Average Distance travelled by a taxi is 70 miles.

Some of the data that we are using based on the current trends is as follows: -

1. According to the data acquired from nyc.gov the car mostly preferred by taxi company cost around 29000$. Mileage of car is around 18$
2. Average salary of a cab driver is around 130$ in NYC.
3. Cost of Gas per Gallon averages around 4$

Daily Expenses of running a single taxi

70/18 (Fuel Consumption) x 4$ = 15.56$

Cab Driver Salary = 130$

Total = 145.56$

Remaining profit after removing daily expenses.

193$ - 145.56$ = 47.44$

Monthly Profit

47.44$ x 30 = 1423.2$

Annual Profit

1423.2$ x 12 = 17078.4$

Annual Maintenance Cost

2000$

So estimated annual profit averages around 15k.

Load Repayments

Buying a taxi cost about 30k.

Annual revenue generation is 15k

Bank usually provides 16% EMI in NYC

Based on these values:-

It will take around 6 and a Quarter Year to repay the loan.