# New York City Yellow Taxi Exploratory Data Analysis Report

*Submission by Snehal Nalawade*

1. **Objective:**

   In the case study I learned exploratory data analysis (EDA) with the help of dataset on yellow taxi rides in New York city. This assignment helped to understand why data analysis and EDA is an important step in the process of data science and machine learning world.

2. **Problem Statement:**

   As an analyst at an upcoming taxi operation in NYC, you are tasked to use the 2023 taxi trip data to uncover insights that could help optimize taxi operations. The goal is to analyze patterns in the data that can inform strategic decisions to improve service efficiency, maximize revenue, and enhance passenger experience.

3. **Data Understanding**

   The yellow taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.
   The data is stored in Parquet format (. parquet). The dataset is from 2009 to 2024. However, for this assignment, we will only be using the data from 2023.
   The data for each month is present in a different parquet file. We got twelve files for each of the months in 2023. The data was collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers like vendors and taxi hailing apps.

   3.1 Data Description
   You can find the data description here: Data Dictionary

   3.1.1 Trip Records Field Name description

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

2. tpep_pickup_datetime - The date and time when the meter was engaged.

3. tpep_dropoff_datetime - The date and time when the meter was disengaged.

4. Passenger_count - The number of passengers in the vehicle. This is a driver entered value.

5. Trip_distance - The elapsed trip distance in miles reported by the taximeter.

6. PULocationID - TLC Taxi Zone in which the taximeter was engaged

7. DOLocationID - TLC Taxi Zone in which the taximeter was disengaged

8. 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

9. Store_and_fwd_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

10. 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

11. 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 USD rush hour and overnight charges.

12. MTA_tax - 0.50 USD MTA tax that is automatically triggered based on the metered rate in use.

13. Improvement_surcharge - 0.30 USD improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.

14. Tip_amount – This field is automatically populated for credit card tips. Cash tips are not included.

15. Tolls_amount - Total amount of all tolls paid in trip.

16. total_amount - The total amount charged to passengers. Does not include cash tips.

17. Congestion_Surcharge - Total amount collected in trip for NYS congestion surcharge.

18. Airport_fee /airport_fee - 1.25 USD for pick up only at LaGuardia and John F. Kennedy Airports Although the amounts of extra charges and taxes applied are specified in the data dictionary, you will see that some cases have different values of these charges in the actual data.

3.1.2 Taxi Zones

Each of the trip records contains a field corresponding to the location of the pickup or drop-off of the trip, populated by numbers ranging from 1- 263. These numbers correspond to taxi zones, which may be downloaded as a table or map/shapefile and matched to the trip records using a join. This is covered in more detail in later sections.

## 4. Data Preparation

This is the first step of data analysis in which we performed sampling, cleaning, formatting and organizing data for further analysis.

4.1 Data Sampling - To ensure a representative subset of trip records for analysis while maintaining uniform coverage across different time periods, a stratified random sampling approach was employed. The sampling process was carried out as follows:

1. Data Source and Structure -
   The dataset consists of monthly parquet files named in the format YYYY-MM.parquet. Each file contains trip records with a pickup timestamp (tpep_pickup_datetime), which was used to extract the date and hour of each trip.
2. Sampling Strategy -
   The data was sampled in a structured manner to ensure uniform distribution across time. The process included:
   A. Extracting Unique Dates: Each monthly dataset was processed to extract all unique dates present in that month.
   B. Iterating Over Each Date: For every unique date, the dataset was filtered to retrieve all trip records corresponding to that specific day.
   C. Hourly Segmentation: Each day was further divided into 24 hourly segments (0 to 23 hours). The dataset was filtered to include only the records belonging to each specific hour.

D. Random Sampling Per Hour: Within each hour, 5% of the available records were randomly selected using a randomized selection method (sample(frac=0.05, random_state=42)) to ensure consistency and reproducibility of results. Later I used 300000 data for analysis for more accurate results.

3. Data Aggregation The hourly sampled data was combined to create a sampled dataset for the entire month. Once all monthly files were processed, the sampled data from each month was merged to form a final dataset covering the entire year.
4. Assumptions and Considerations It was assumed that all parquet files are stored in a single directory for processing. The random state (42) was used to ensure the sampling process is reproducible. Any new columns derived during this process were labeled with a _derived suffix to differentiate them from the original dataset attributes.

4.2 Data Cleaning

To ensure data integrity and improve the quality of analysis, a comprehensive data cleaning process was performed on the sampled dataset.

1. Dropping Columns Below unnecessary columns were dropped:
   a. Date and hour : these columns were created for sampling purpose and now not required, so dropped these columns.
   b. store_and_fwd_flag: Not useful for analysis, indicates if the trip record was held in vehicle memory before sending to the server.
   c. mta_tax: Fixed tax amount, does not provide variability.
2. Datatype Correction Below columns has incorrect data type and it was fixed as below:
   a. RatecodeID: It is parsed as float64 but as per the data dictionary the values could be 1, 2, 3, 4, 5, 6 and should be changed to integer.
   b. tpep_pickup_datetime and tpep_dropoff_datetime - It is parsed as object, but these dates should be changed to date.
3. Columns Merging - There are two airport fee columns airport_fee and Airport_fee. This is possibly an error in naming columns. These two columns were merged by adding the values of both columns and the new column was named as Airport_fee_combined

4. Fixing Negative Values During the data cleaning process, it was observed that certain financial columns contained negative values, which were not expected in

the dataset. Specifically, the below columns had negative entries that could have resulted from data entry errors or system inconsistencies.

Negative values removed from following columns-

a. improvement_surcharge
b. total_amount
c. congestion_surcharge
d. airport_fee_combined

To address this issue, a below function is applied to convert all negative values into 0 and then performed respective operations.

```
Final_trip_df[neg_col] = Final_trip_df[neg_col].applymap(lambda x: 0 if x < 0 else x)
print((Final_trip_df[neg_col]<0).sum())
```

5. Fixing Missing Values -

To ensure data completeness, missing values in critical columns were identified and addressed. The below columns contained missing entries, which could impact downstream analysis.

a. passenger_count
b. RatecodeID
c. congestion_surcharge

To handle these missing values effectively the mode function was applied to each column, The mode was chosen as it provides most occurred value. By filling in the missing values with the mode, the dataset was improved for reliability while minimizing distortions in the overall distribution.

6. Handling Outliers - To improve data quality and ensure meaningful analysis, outliers were identified and removed based on domain knowledge and logical constraints.

The following steps performed-

a. passenger_count: Entries where the passenger count exceeded 6 or was 0 were removed to maintain realistic trip records.
b. payment_type: Rows where payment_type was 0 were filtered out, as these entries were deemed invalid as per data dictionary.

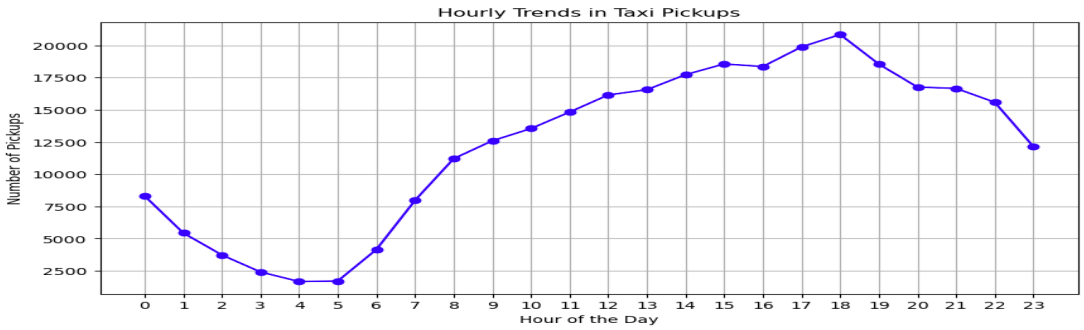
c. trip_distance: Trips shorter than 0.62 miles or longer than 120 miles were removed to exclude unrealistic or erroneous trip records. These outlier-handling measures helped refine the dataset, making it more representative of real-world scenarios.
d. RatecodeID – removed highest ratecodeid.

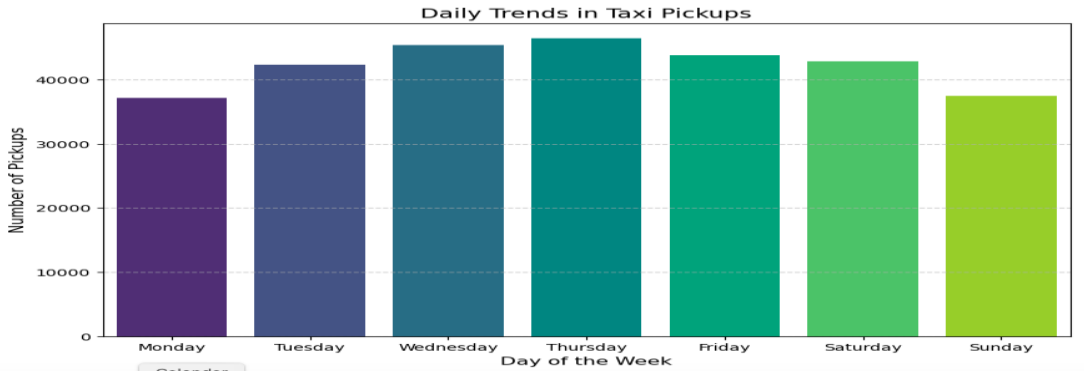
**5. Exploratory Data Analysis**

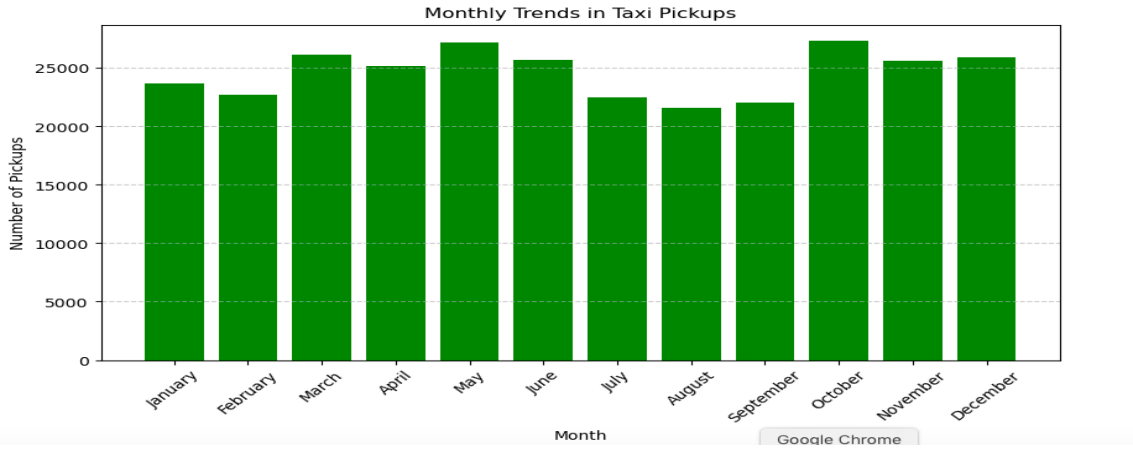
Analyzing and visualizing data to understand its main features, find patterns, and discover how different parts of the data are connected.

5.1 Temporal Analysis

5.1.1 Analyse the distribution of taxi pickups by hours, days of the week, and months.



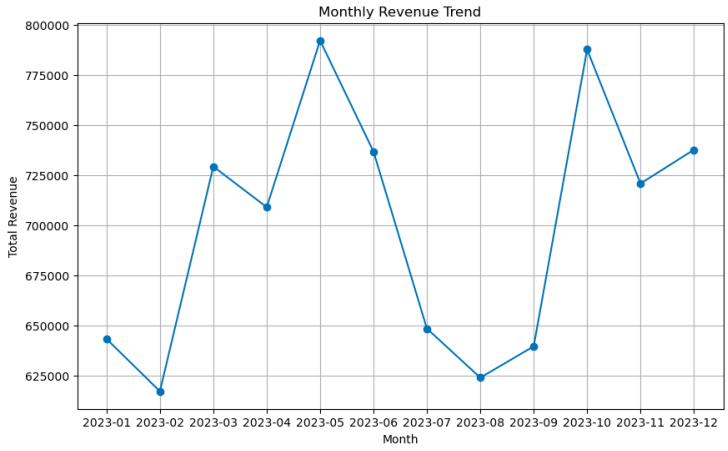



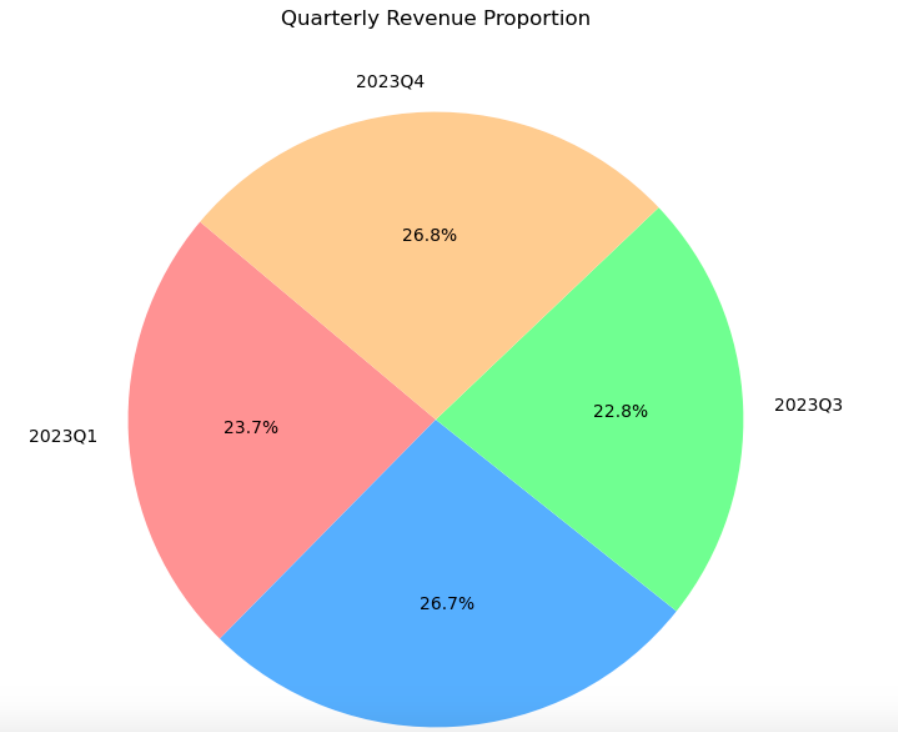



The charts reveal several key insights. The highest number of trips in May occurred on Thursdays. A clear pattern emerges, with taxi rides peaking in the morning, stabilizing around noon, and rising again in the evening. The busiest period is between 5:00 PM and 7:00 PM.

## 5.2 Financial Analysis

### 5.2.1 Analyse the revenue distribution trend

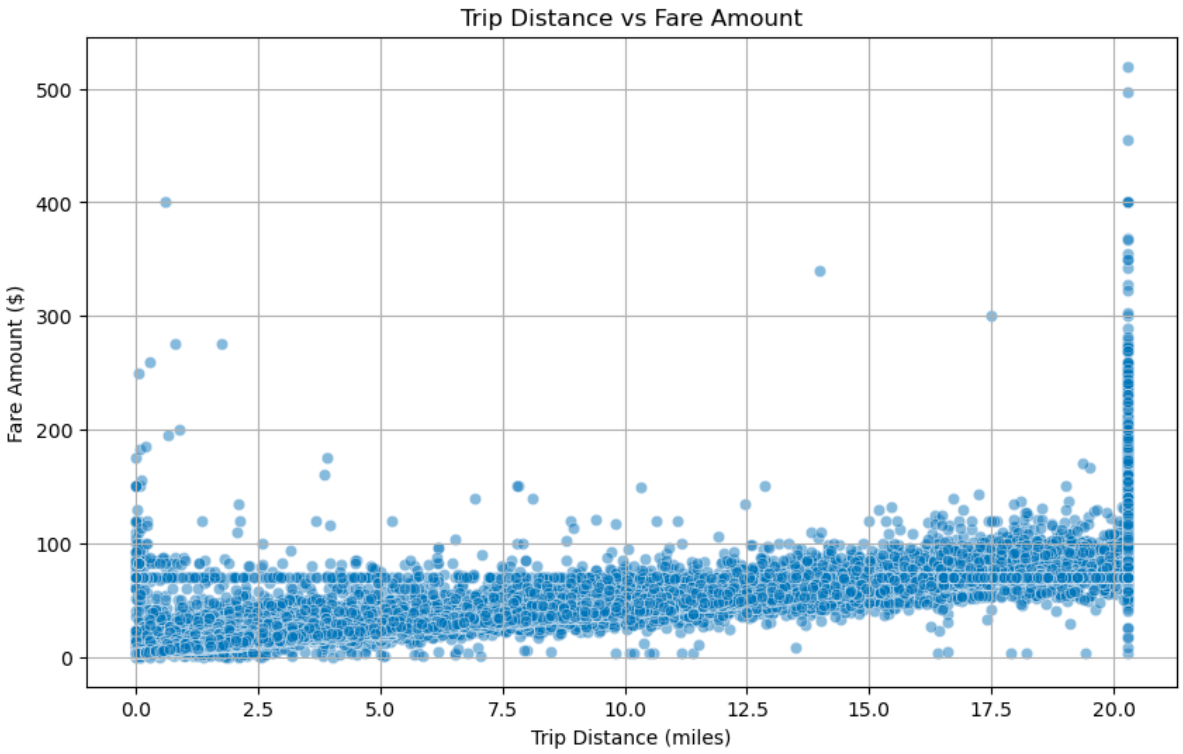




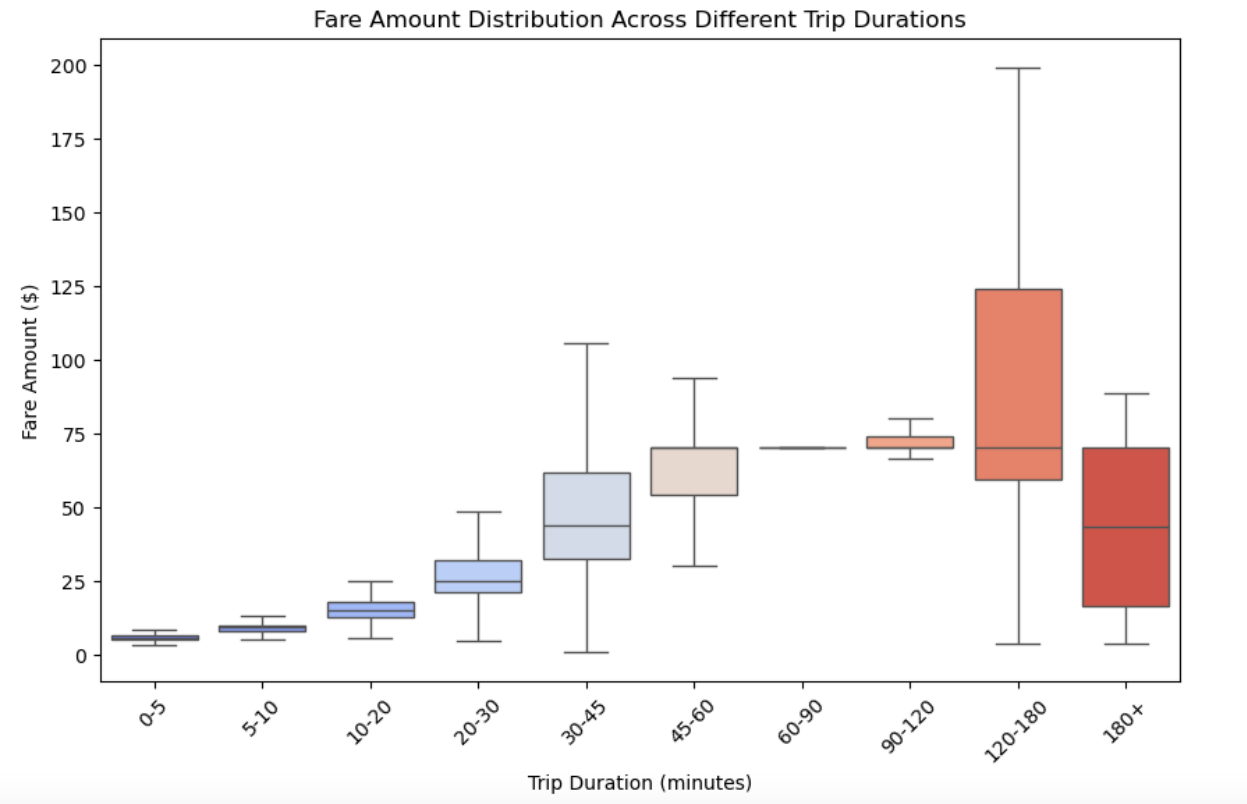

The key insights from the above charts are that the highest revenue was generated in Quarter 4, with October being the peak month.

Conversely, the lowest revenue occurred in Quarter 3, with August being the bottom month. Most of the revenue was generated by the trips completed in daytime.

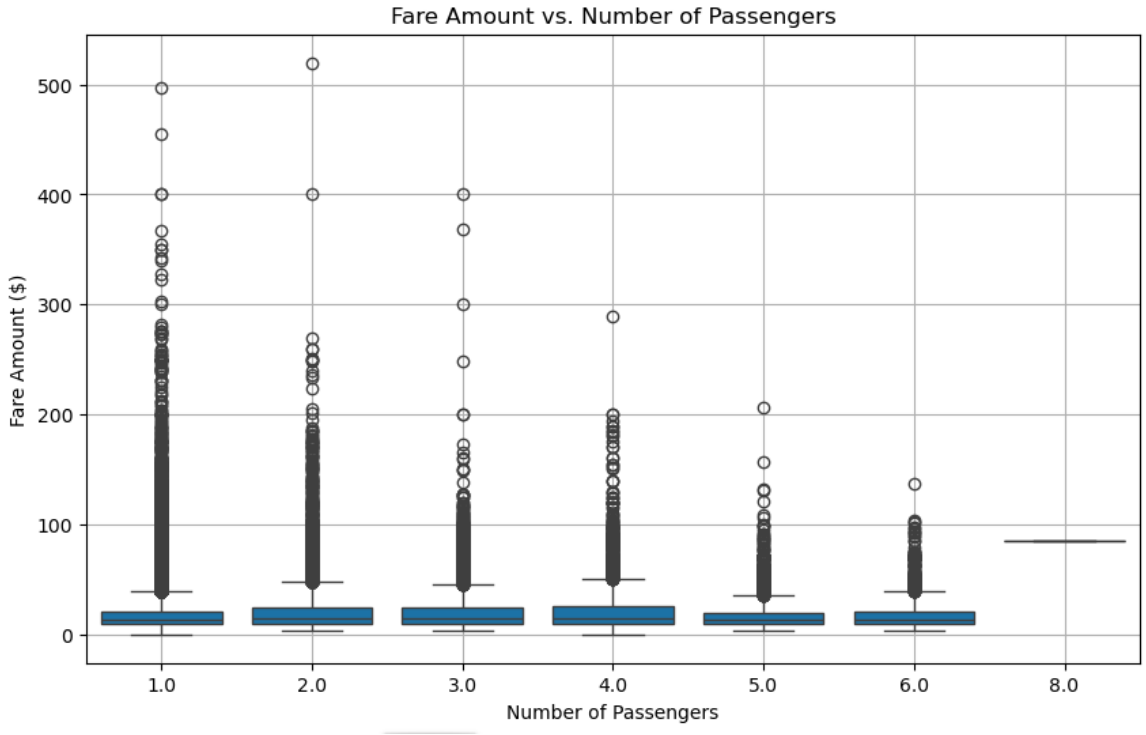
5.2.2 Relationship of Fare Amount with Trip Distance, Trip Duration and Number of Passengers



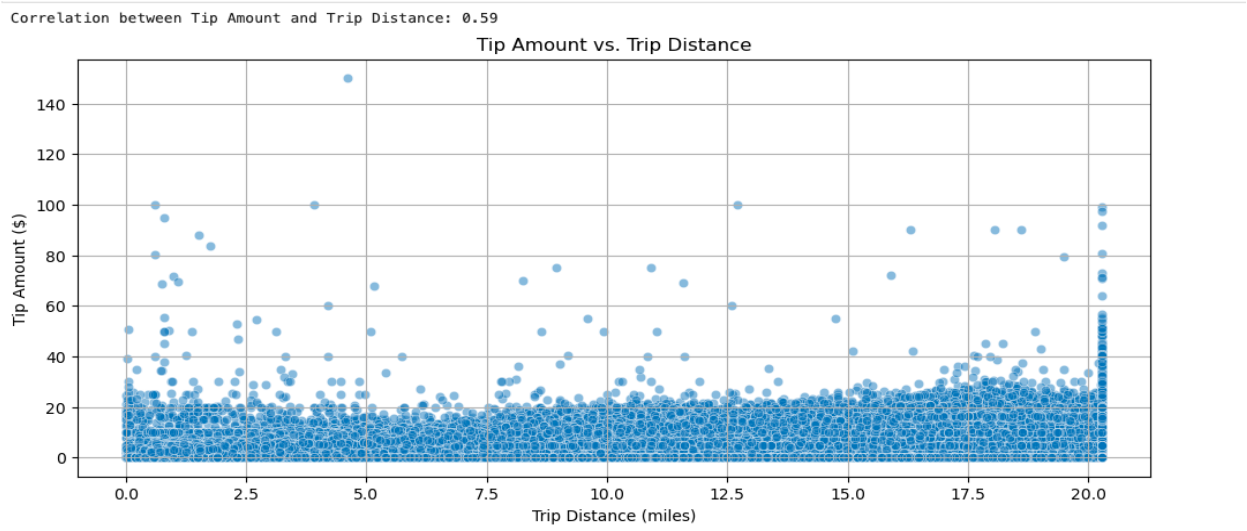

1. Relationship between T**rip Distance and Fare Amount** There seems to be a positive correlation between trip distance and fare amount. As the trip distance increases, the fare amount generally increases as well. The data points are spread out, indicating variability in fare amounts for similar trip distances. This could be due to factors like traffic, time of day, or additional charges. The majority of the data points are clustered in the lower range of trip distances (0-20 miles) and fare amounts (0-200 dollars), suggesting that most trips are relatively short and inexpensive.




2. Relationship between **Trip Duration and Fare Amount T**here seems to be a positive correlation between trip duration and fare amount. As the trip duration increases, the fare amount also tends to increase. This is expected since longer trips typically cost more. The data points are spread out, indicating variability in fare amounts for similar trip durations. This could be due to different rates, traffic conditions, or other variables affecting the fare. Most of the data points are clustered in the lower range of trip durations (0-400 minutes) and fare amounts (0-200). This suggests that the majority of trips are relatively short and inexpensive.



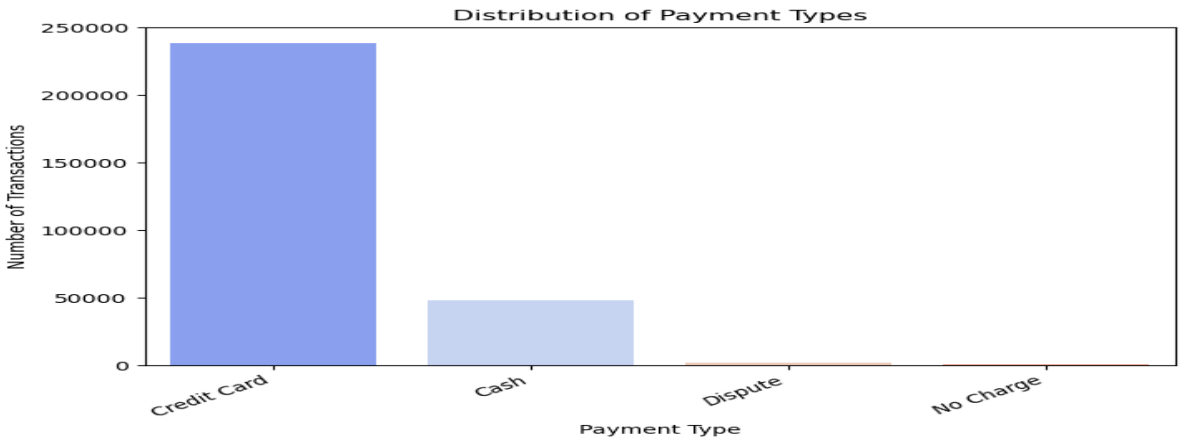

3. Relationship between **Number of Passengers and Fare Amount** The correlation between the number of passengers and the fare amount is weak. We can observe that single passengers tend to travel longer distances compared to groups, as the fare amount is higher for single passengers. This indicates a positive correlation between trip distance and fare amount.




4. Relationship between **Trip Distance and Tip Amount** There seems to be a positive correlation between trip distance and tip amount. As the trip distance increases, the tip amount also tends to increase. While there is a general upward trend, there

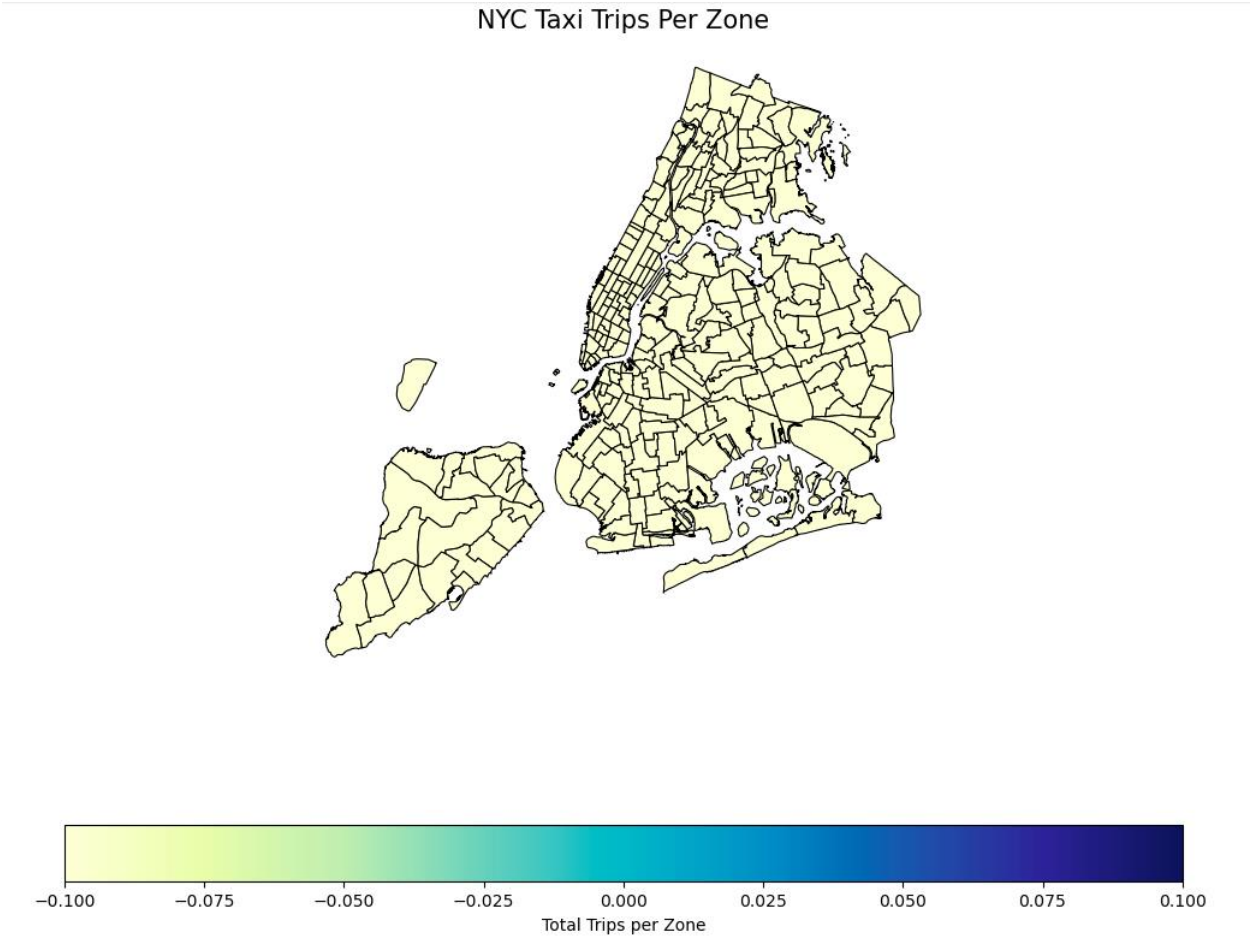
is significant variability in tip amounts for similar trip distances. This suggests that factors other than distance (such as service quality, passenger generosity, or fare amount) may influence tipping behavior. The majority of the data points are clustered in the lower range of trip distances (0-40 miles) and tip amounts (0-100). This suggests that most trips are relatively short, and tips are modest. The highest tip amounts are observed for trips around 60-80 miles, but these are less frequent.

5.2.3 Analyse the distribution of different payment types (payment_type).




5.3 Geographical Analysis

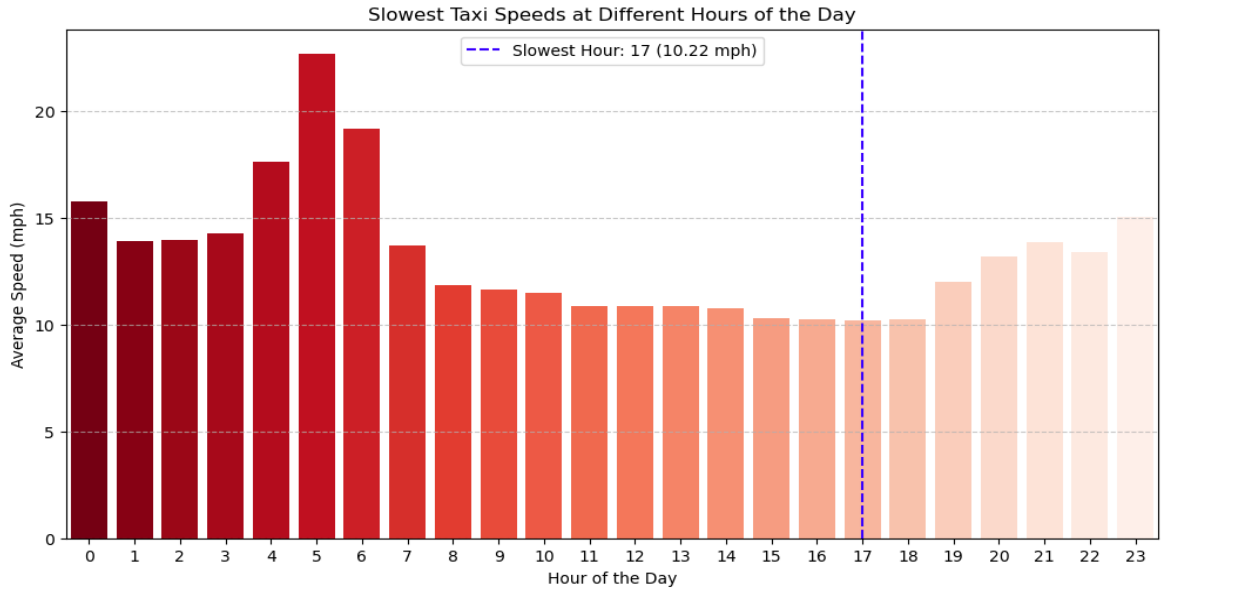
5.3.1 Total Trips/Zone vs Zone-wise Number of Trips

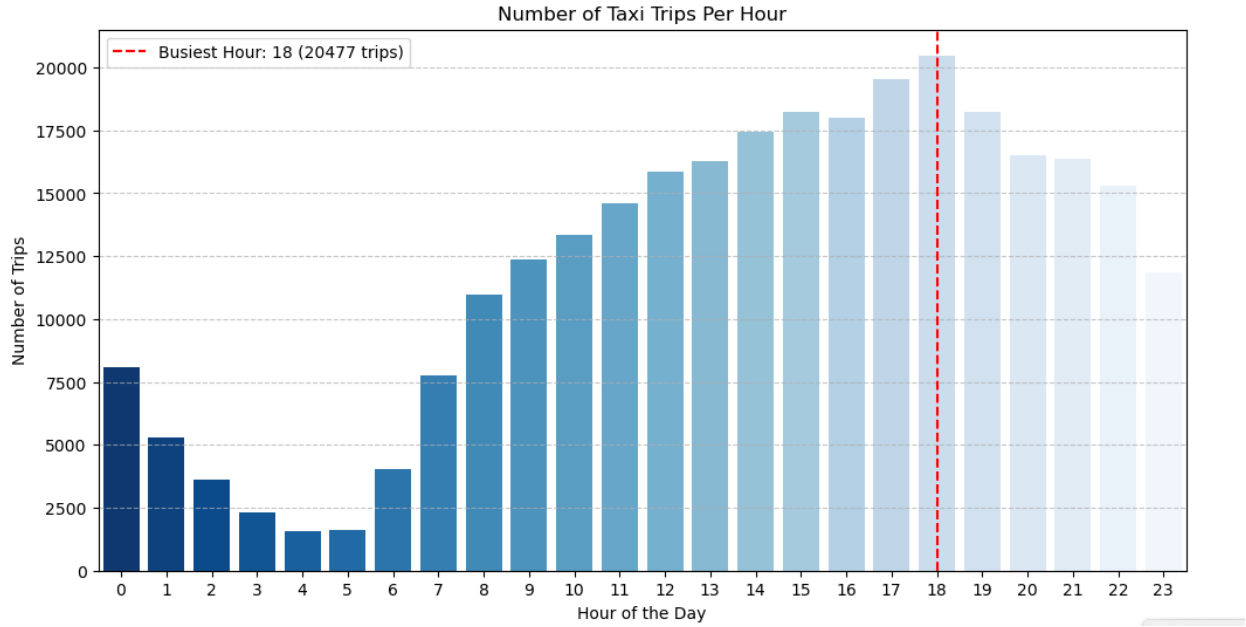


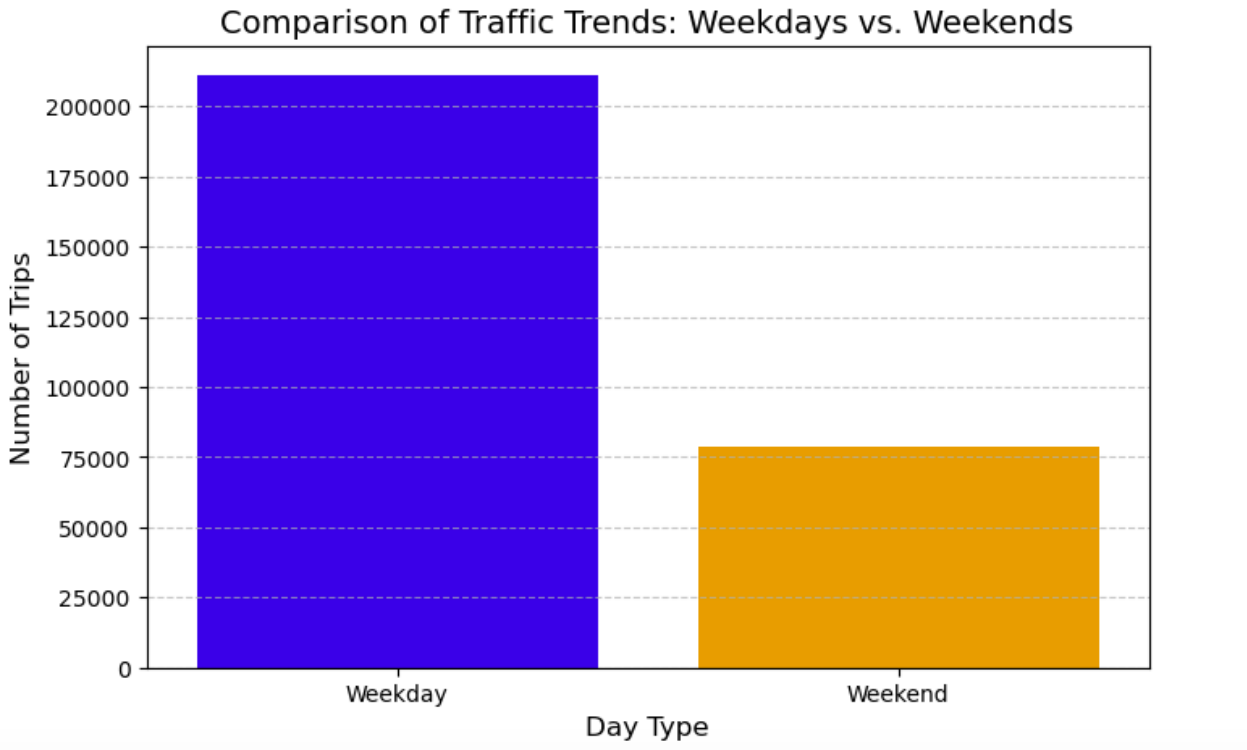

| | zone | borough | total_trips |
|---|---|---|---|
| **0** | Newark Airport | EWR | 0.0 |
| **165** | Morningside Heights | Manhattan | 0.0 |
| **167** | Mott Haven/Port Morris | Bronx | 0.0 |
| **168** | Mount Hope | Bronx | 0.0 |
| **169** | Murray Hill | Manhattan | 0.0 |
| **170** | Murray Hill-Queens | Queens | 0.0 |
| **171** | New Dorp/Midland Beach | Staten Island | 0.0 |
| **172** | North Corona | Queens | 0.0 |
| **173** | Norwood | Bronx | 0.0 |
| **174** | Oakland Gardens | Queens | 0.0 |

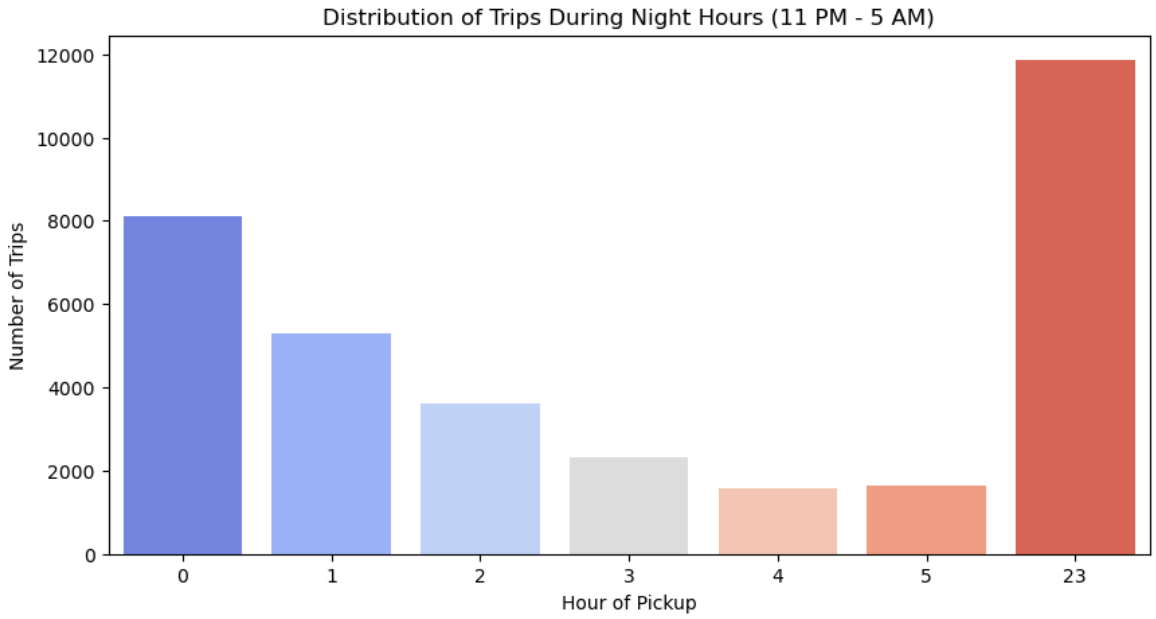
## 5.4 Operational Efficiency Analysis

### 5.4.1 Traffic Trend Analysis

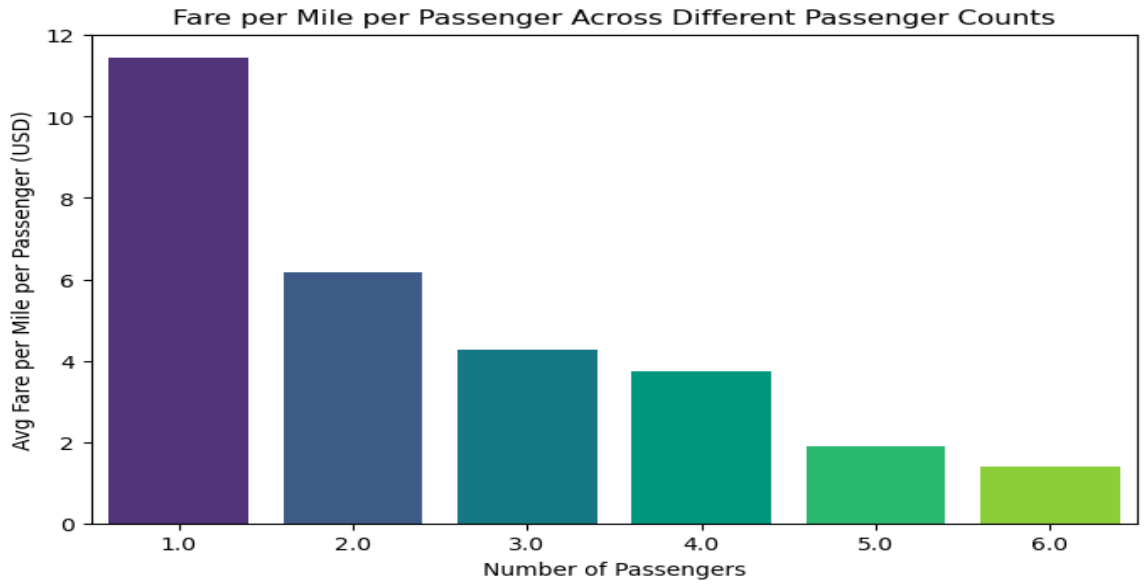

Slowest Taxi Speeds at Different Hours of the Day
Slowest Hour: 17 (10.22 mph)
Average Speed (mph)
Hour of the Day

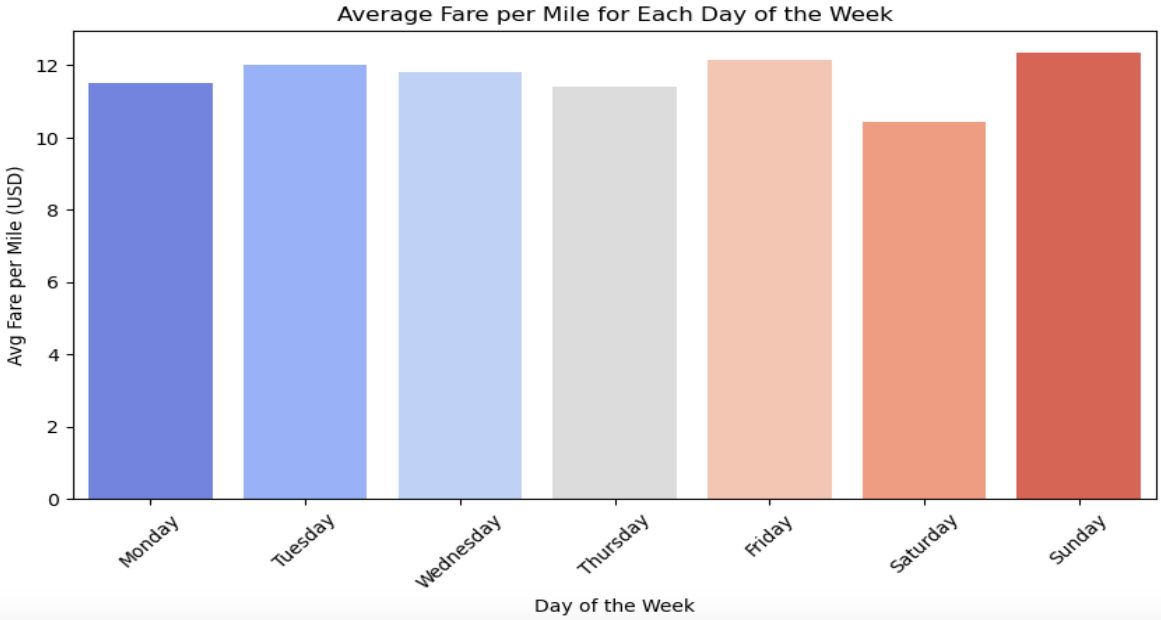

Number of Taxi Trips Per Hour
Busiest Hour: 18 (20477 trips)
Number of Trips
Hour of the Day

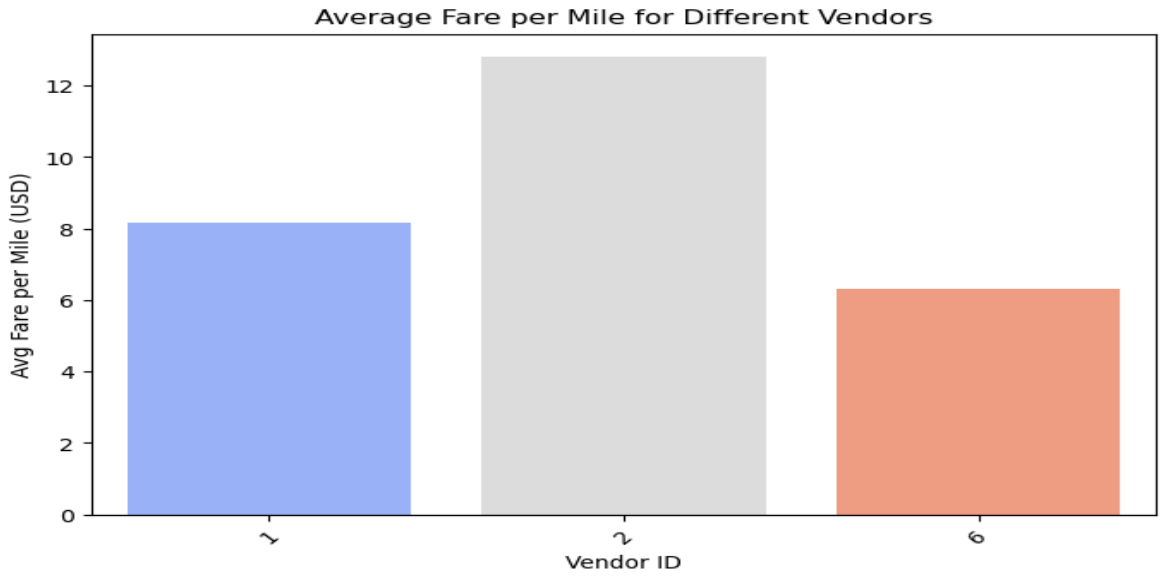


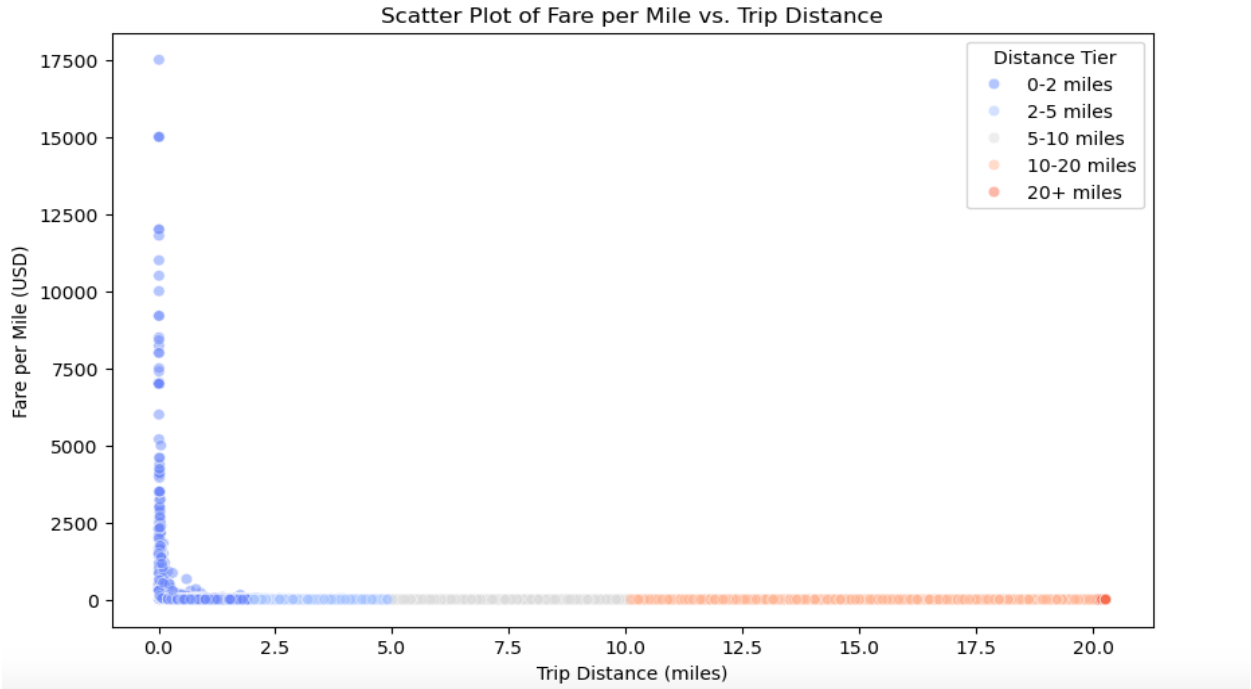




5.4.2 Pricing Strategy Analysis

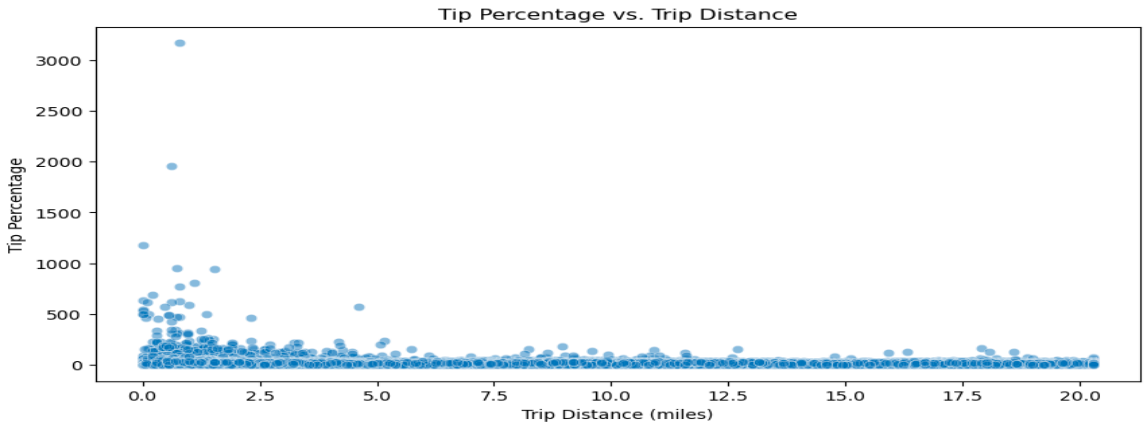

Fare per Mile per Passenger Across Different Passenger Counts
Avg Fare per Mile per Passenger (USD)
0
2
4
6
8
10
12
1.0
2.0
3.0
4.0
5.0
6.0
Number of Passengers

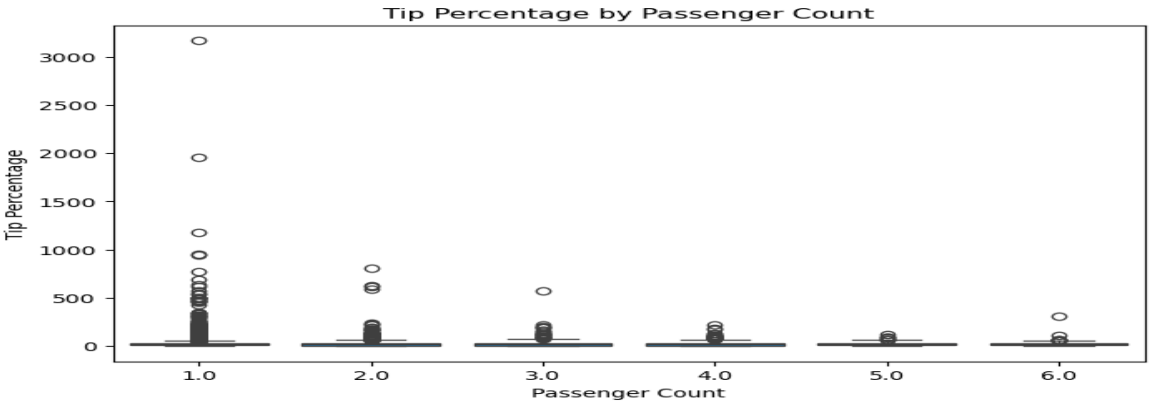

Average Fare per Mile for Each Day of the Week
Avg Fare per Mile (USD)
0
2
4
6
8
10
12
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week

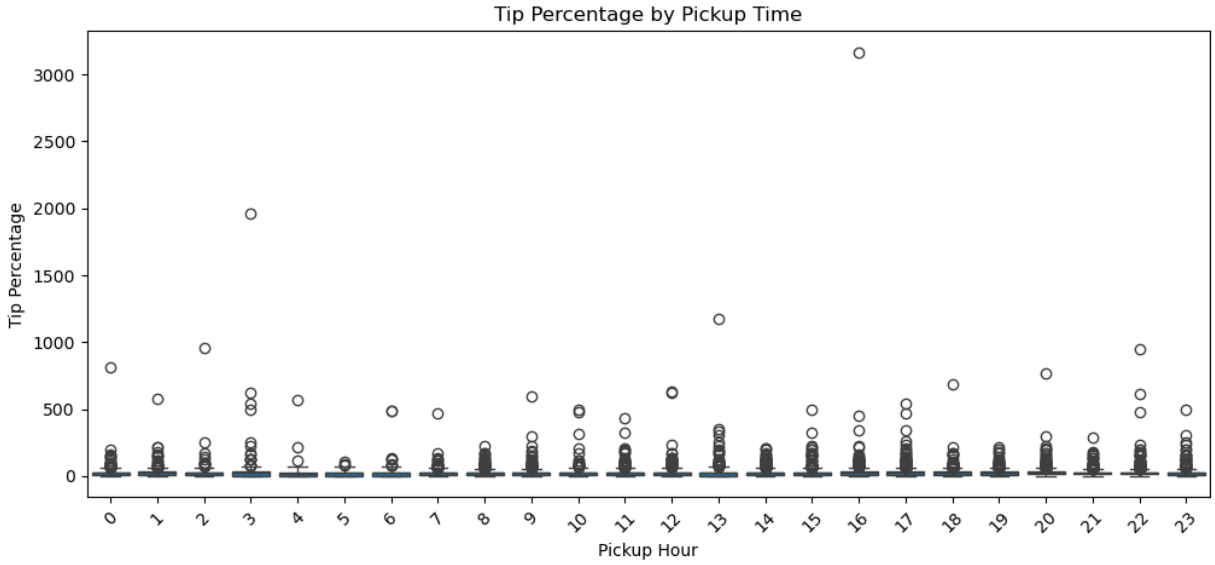


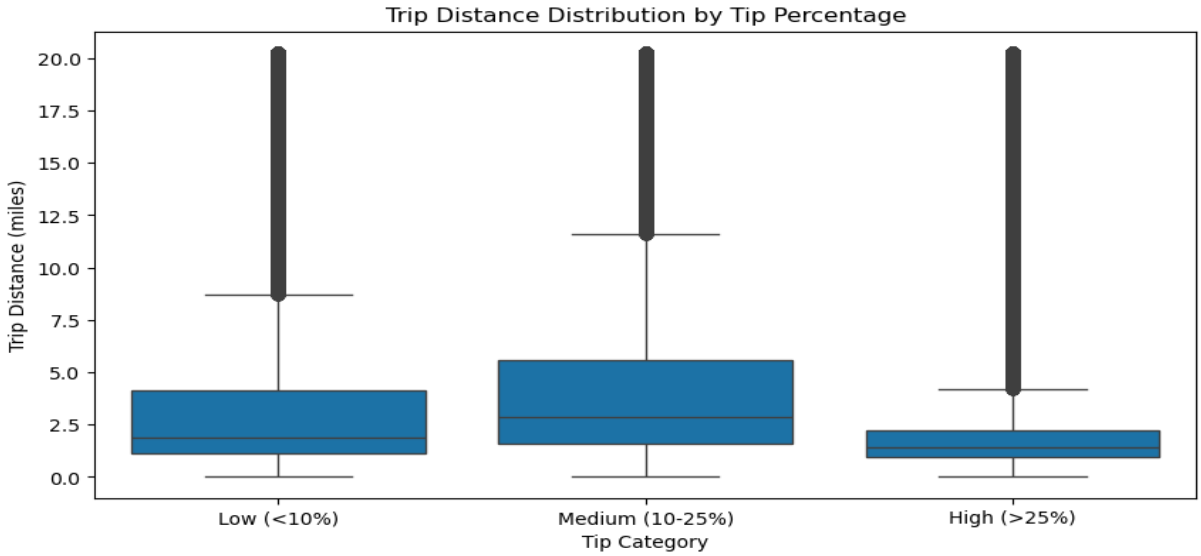




### 5.4.3 Customer Experience Analysis

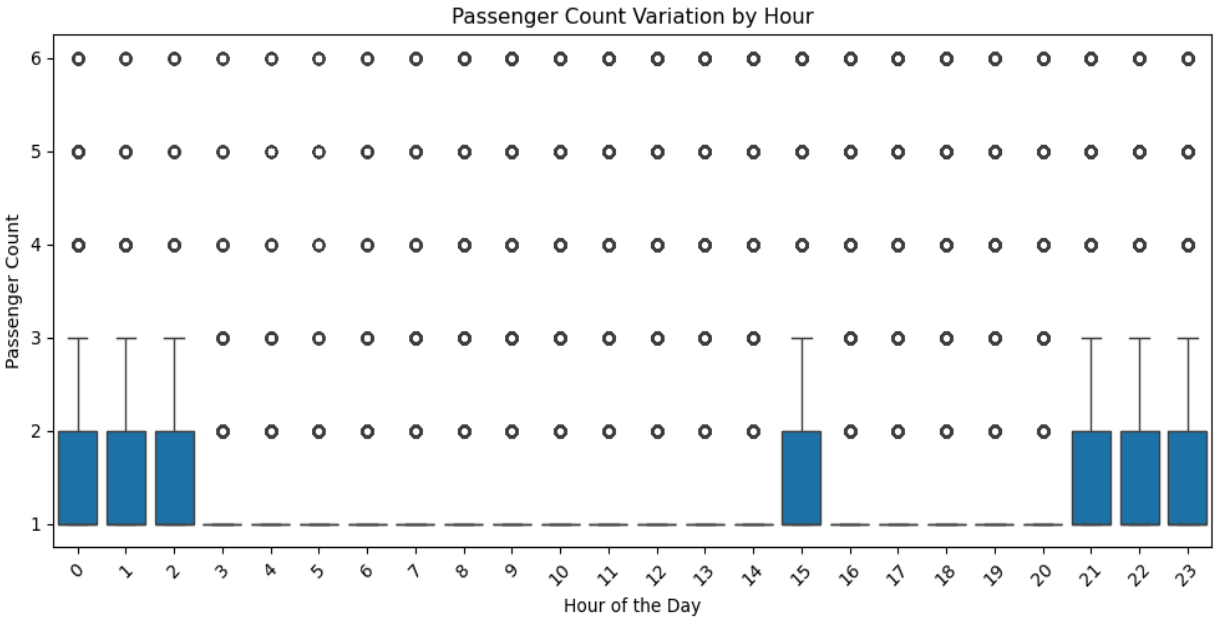

Tip Percentage vs. Trip Distance
Tip Percentage
Trip Distance (miles)
0
500
1000
1500
2000
2500
3000
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

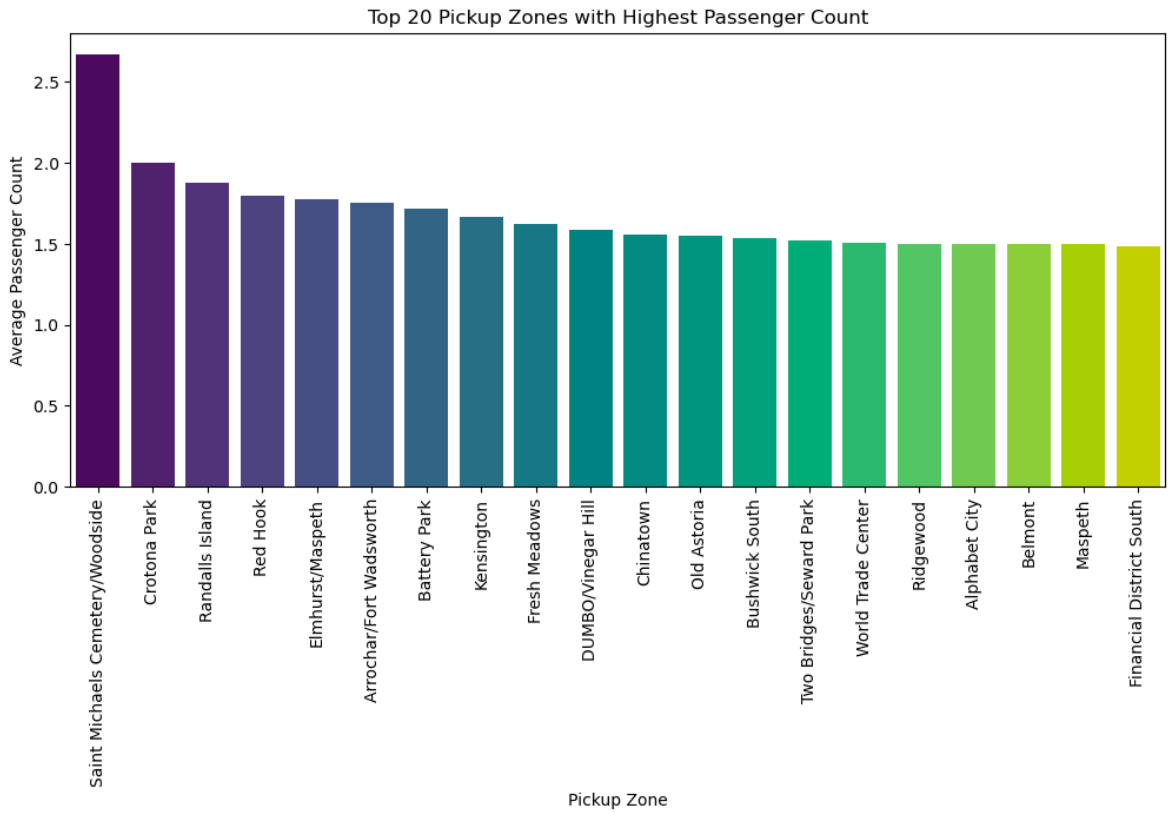

Tip Percentage by Passenger Count
Tip Percentage
Passenger Count
0
500
1000
1500
2000
2500
3000
1.0
2.0
3.0
4.0
5.0
6.0


Tip Percentage by Pickup Time
Tip Percentage
Pickup Hour
0
500
1000
1500
2000
2500
3000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

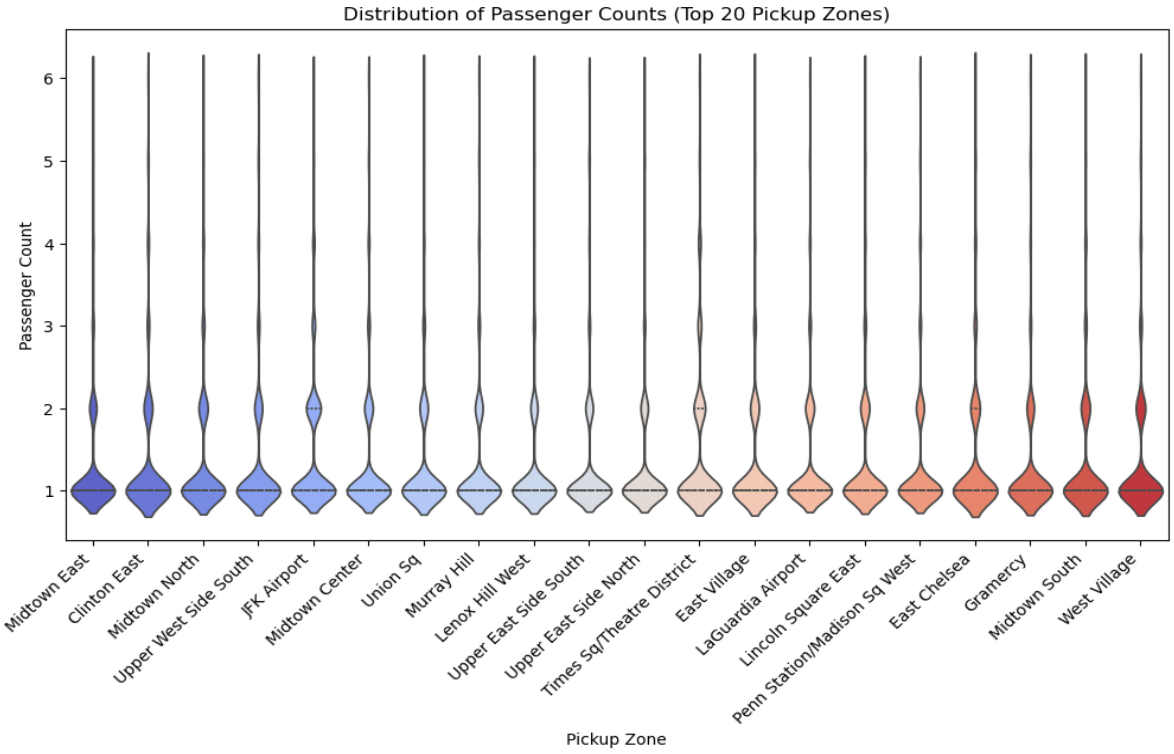


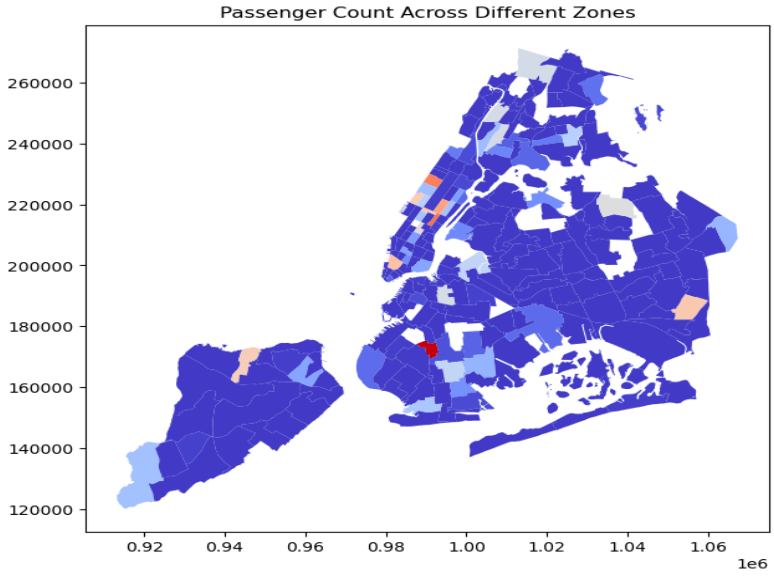

### 5.4.4 Variation of Passenger Count Analysis

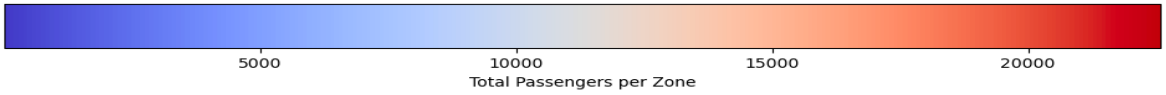





### 5.4.5 Surcharges Analysis




## 6 Conclusion

Final Insights and Recommendations

6.1.1 Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.

Based on the analysis of demand patterns and operational inefficiencies, below are some recommendations:

1. Allocate more cabs during peak day hours (6 AM to 10 PM) based on the analysis at section 3.2.2.
2. Implement surge pricing in high-demand zones during daytime peak periods.
3. Adjust pricing according to the time of day and day of the week, based on the analysis of average fare per mile for different hours and days. (Derived from section 3.2.4 and 3.2.10)
4. As an outcome for the analysis conducted at section 3.2.7, we should increase number of cabs in high-demand pickup and drop-off zones during night hours (11 PM to 5 AM).
5. Repositioning algorithms can be introduced for cabs positioning to fulfill the demand surges in these areas.

6.1.2 Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analyzing trip trends across time, days and months.

Suggestions on strategically positioning cabs based on trip trends:

1. From the above heatmap we can identify zones with high demand during peak hours (e.g., rush hour, evenings). Position more cabs in these high-demand zones during peak times to reduce wait times and improve customer satisfaction.

2. Observe how demand fluctuates throughout the day.

3. Adjust cab deployment accordingly, increasing presence during peak periods and reducing it during lulls.

4. Deploy more cabs on weekdays during peak hours and adjust deployment for weekend demand patterns, which might be concentrated in specific areas or times. (Derived from 3.2.4)

5. Position cabs strategically to efficiently serve short and long-distance trips, optimizing overall efficiency.

6.1.3 Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.

1.  Below are recommendations for data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors

2. Monthly revenue is very low in July, August, September company can offer competitive price as compared to other vendor during these months which can increase pickup during that time and revenue will increase.

3. The correlation between Trip Duration and Fare Amount is 0.32 which is very low.

4. The company can impose a waiting charge for the ride which will increase the correlation between these two variables.

5. A fair amount depending on the number of passengers can also increase the revenue for the company.

6. Consider using machine learning models to predict demand elasticity for various distances. This would allow more precise price adjustments.