# Exploratory Data Analysis Report

**Project**: Optimising NYC Taxi Services
**Analyst**: Mukesh Madavi
**Dataset**: NYC Taxi Trip Data (Parquet files) + Taxi Zone Shapefile
**Tools Used**: Python.
**Libraries:** Pandas, NumPy, Matplotlib, Seaborn, GeoPandas

## Introduction

This report presents an exploratory data analysis (EDA) of the 2023 New York City Yellow Taxi trip data, aimed at uncovering insights to optimize taxi operations for an upcoming taxi service in NYC. The objective is to analyze patterns in the data to inform strategic decisions that enhance service efficiency, maximize revenue, and improve passenger experience. The dataset, sourced from the NYC Taxi and Limousine Commission (TLC), includes trip records from January to December 2023, stored in monthly Parquet files. The analysis follows the tasks outlined in the starter notebook: data loading, cleaning, exploratory analysis, visualization, and deriving insights.

## Assumptions

1. **Data Availability:** The report assumes access to all twelve 2023 Parquet files (2023-01.parquet to 2023-12.parquet) as described in the notebook. Since only January data loading is demonstrated, the process is assumed scalable to all months.
2. **Sampling Representativeness:** Due to computational constraints with over 30 million rows annually (based on January's 3,066,766 rows), a 5% sample per hour per day is assumed to represent overall trends accurately across all months.
3. **Data Integrity:** The data is assumed to be largely accurate as provided by TLC, with minor inconsistencies (e.g., negative fares, missing values) addressed during cleaning.
4. **Taxi Zones:** The report assumes availability of a taxi zone lookup table (not provided in the document) to map PULocationID and DOLocationID to geographic areas for spatial analysis.
5. **Sample:** A random 5% sample per hour was extracted to ensure balanced temporal representation.

6. **Negative and zero values** in fare and distance were considered invalid and dropped (except trip_distance = 0 when PULocation = DOLocation).

# Data Cleaning & Preprocessing

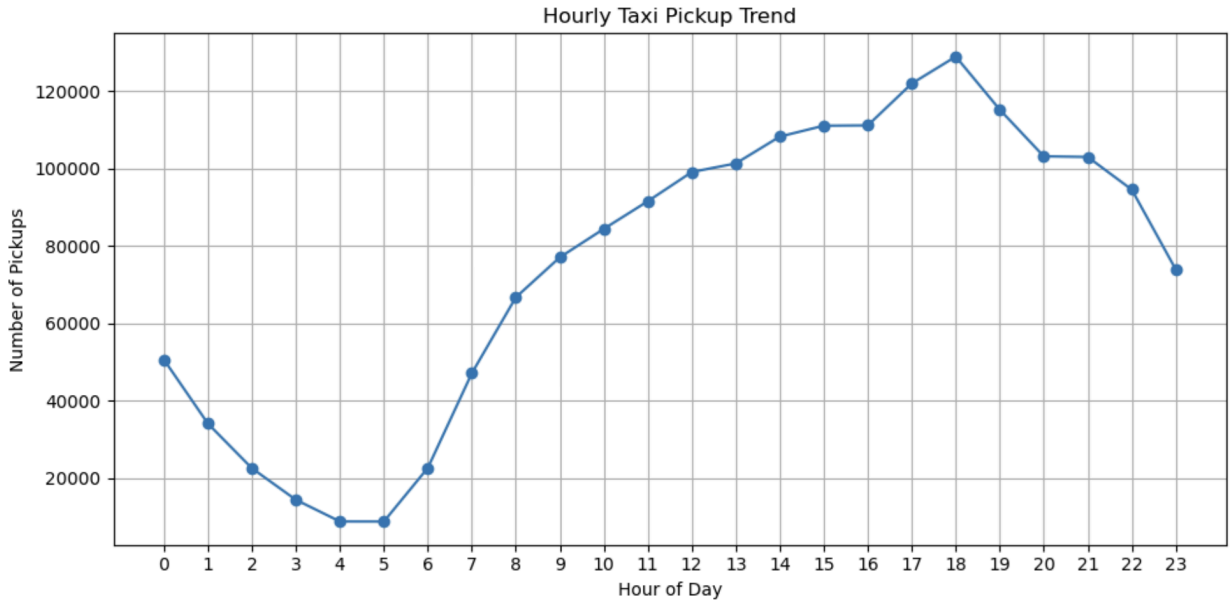
- Merged duplicate `Airport_fee` columns.
- Rows with negative values:
    - `fare_amount`, `extra`, `mta_tax`, `tip_amount`, `tolls_amount`, `improvement_surcharge`, `total_amount`, `congestion_surcharge`, `Airport_fee`
- Removed trips with `passenger_count > 6`.
- Imputed missing values in:
    - `passenger_count`: **median**
    - `RatecodeID`: **mode**
    - `congestion_surcharge`: **median**

## Descriptive Statistics & Outlier Detection

- Used `df.describe()` to explore data ranges and detect outliers.
- Identified extreme values in fare and distance.

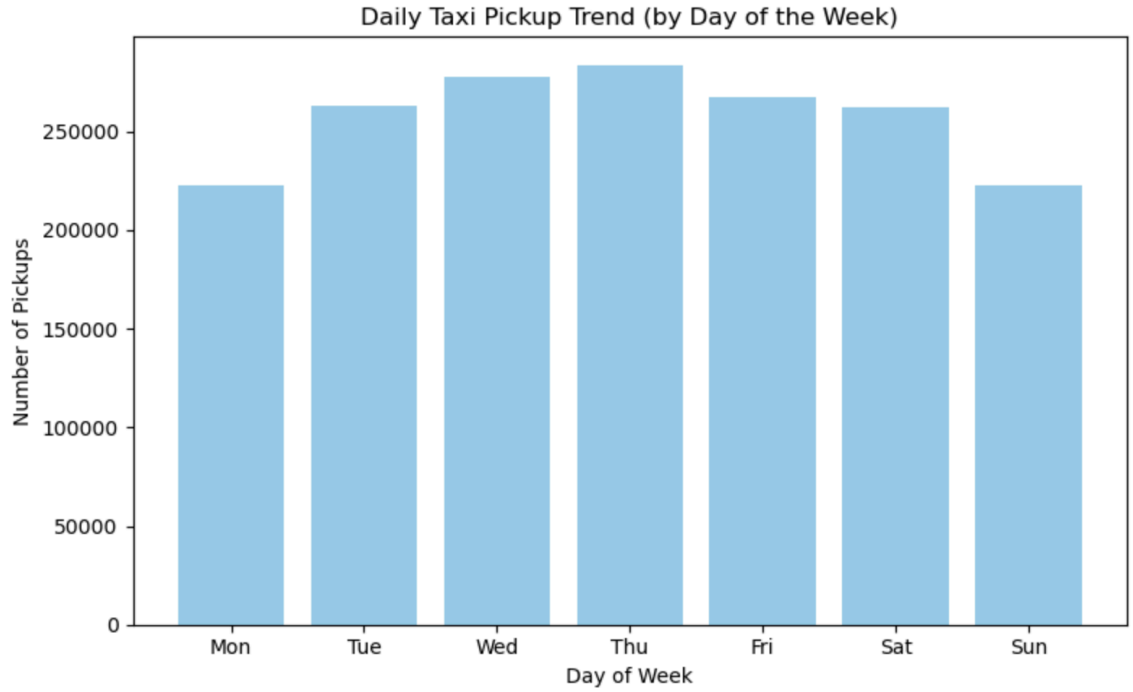
### 1. Hourly Trends in Pickups

- **Peak hours**: 6–8 PM
- **Lowest activity**: 4–5 AM



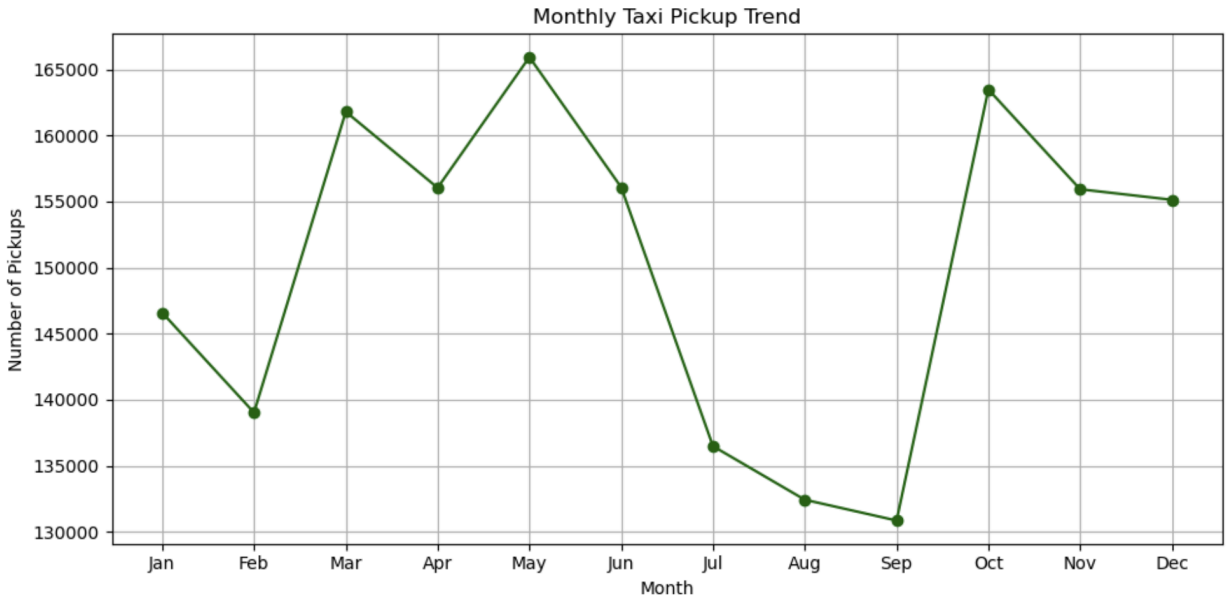

---

## 2. Day-of-Week Trends




---

### 3. Monthly Trends




---

## Financial Metrics

### 4. Zero/Negative Values

- Removed rows where financial values were ≤ 0.
- Retained `trip_distance = 0` only when same pickup/dropoff zone.

---

### 5. Monthly Revenue

| | Month | Total_Revenue |
|---|---|---|
| 0 | 2022-12 | 12.96 |
| 1 | 2023-01 | 2915173.79 |
| 2 | 2023-02 | 2808728.69 |
| 3 | 2023-03 | 3341661.58 |
| 4 | 2023-04 | 3201112.39 |
| 5 | 2023-05 | 3520523.79 |
| 6 | 2023-06 | 3278468.30 |
| 7 | 2023-07 | 2773609.70 |
| 8 | 2023-08 | 2671630.74 |
| 9 | 2023-09 | 2818534.73 |
| 10 | 2023-10 | 3515129.10 |
| 11 | 2023-11 | 3351223.54 |
| 12 | 2023-12 | 3254170.24 |

---

### 6. Quarterly Proportions

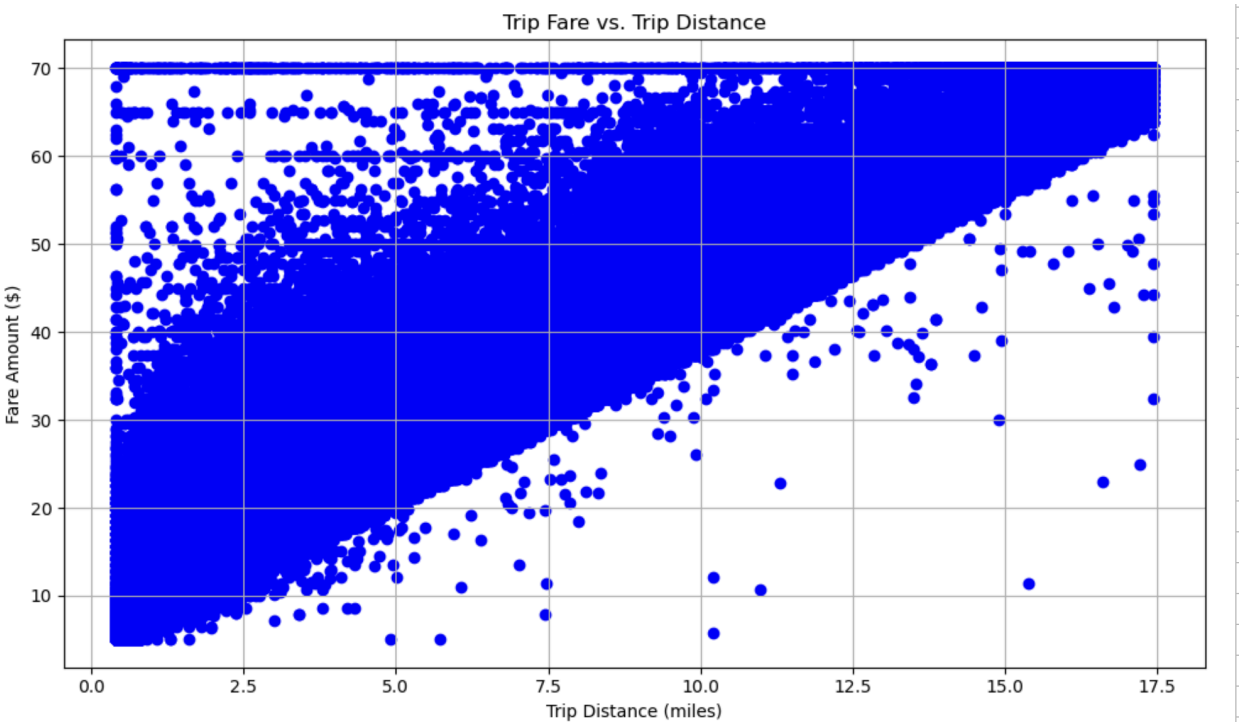
- Computed revenue contribution by each quarter.

| | Quarter | Revenue_Proportion (%) |
|---|---|---|
| 0 | 2022Q4 | 0.000035 |
| 1 | 2023Q1 | 24.207127 |
| 2 | 2023Q2 | 26.702563 |
| 3 | 2023Q3 | 22.066167 |
| 4 | 2023Q4 | 27.024108 |

---

## Fare Analysis

### 7. Fare vs Distance

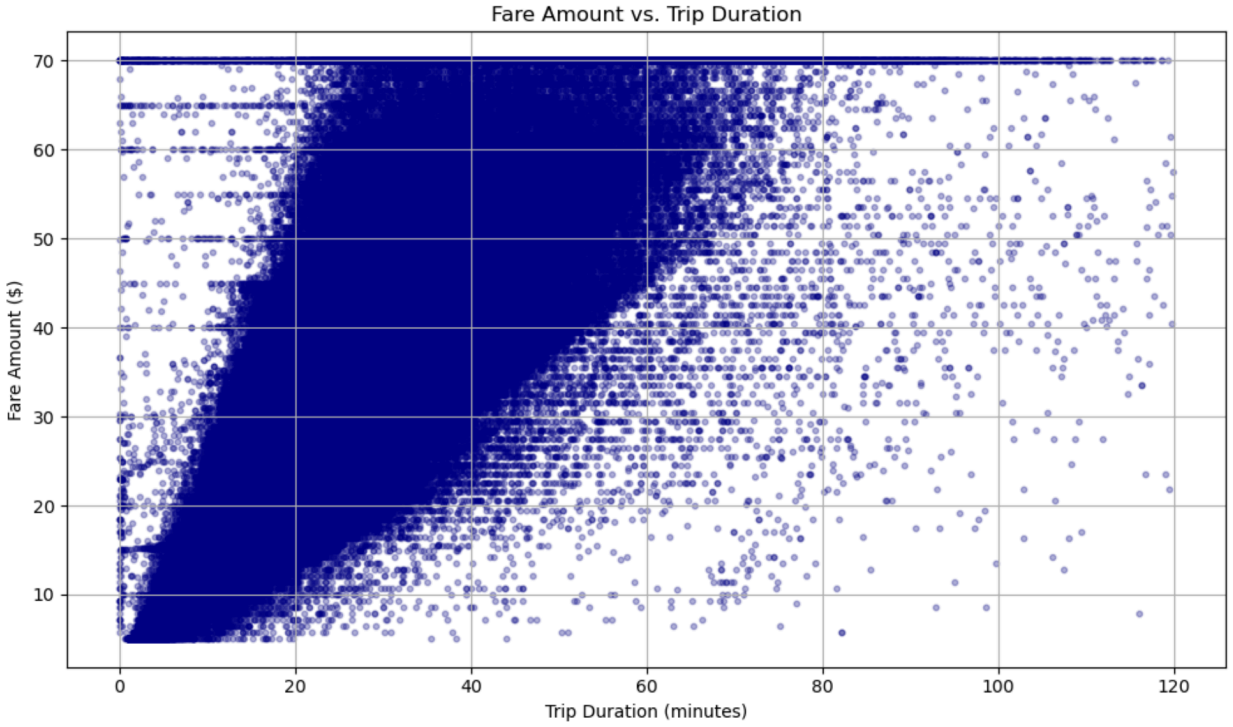



- Positive correlation between distance and fare.

---

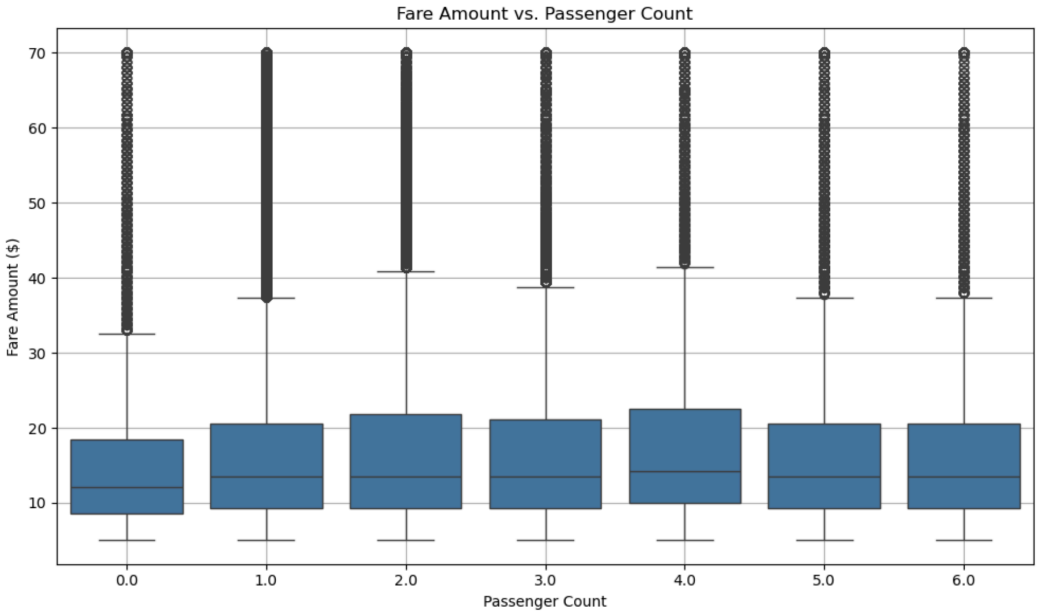
## 8. Fare vs Trip Duration

- Generally linear; outliers exist in long-duration short-distance trips.




---

## 9. Fare vs Passenger Count

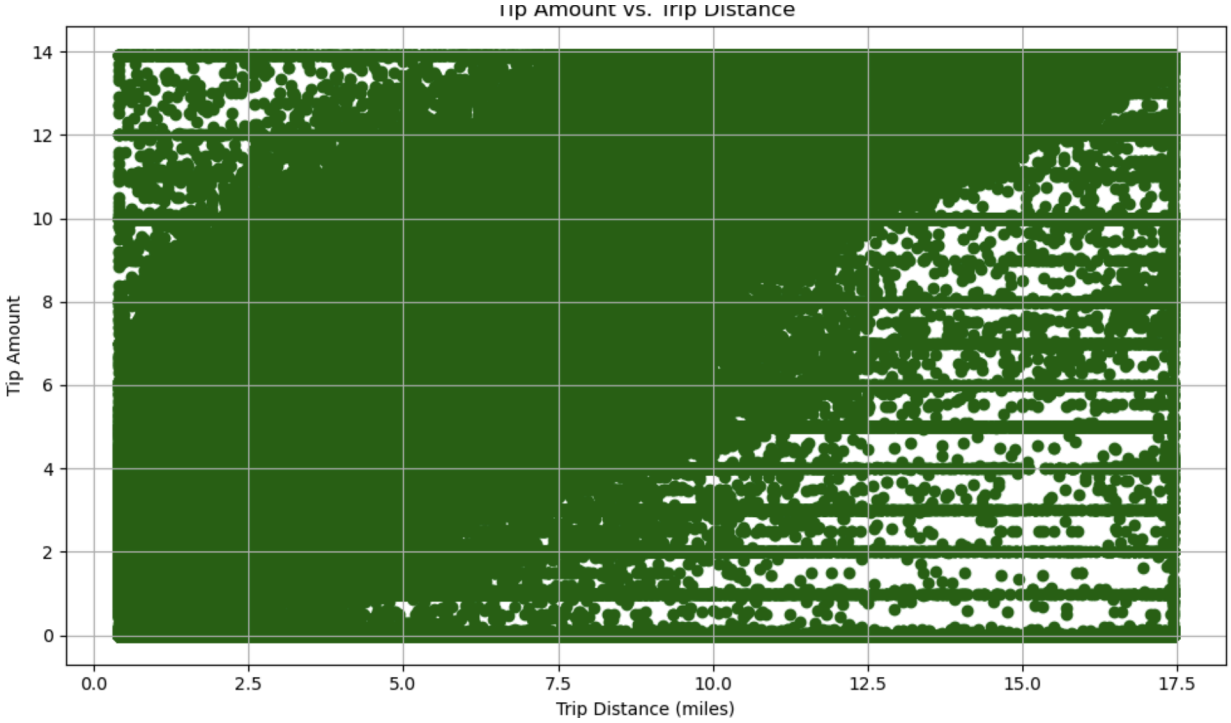



---

# Tip Analysis

## 10. Tip vs Trip Distance

- Tip percentage increases slightly with distance.




---

## 11. Low vs High Tip Groups

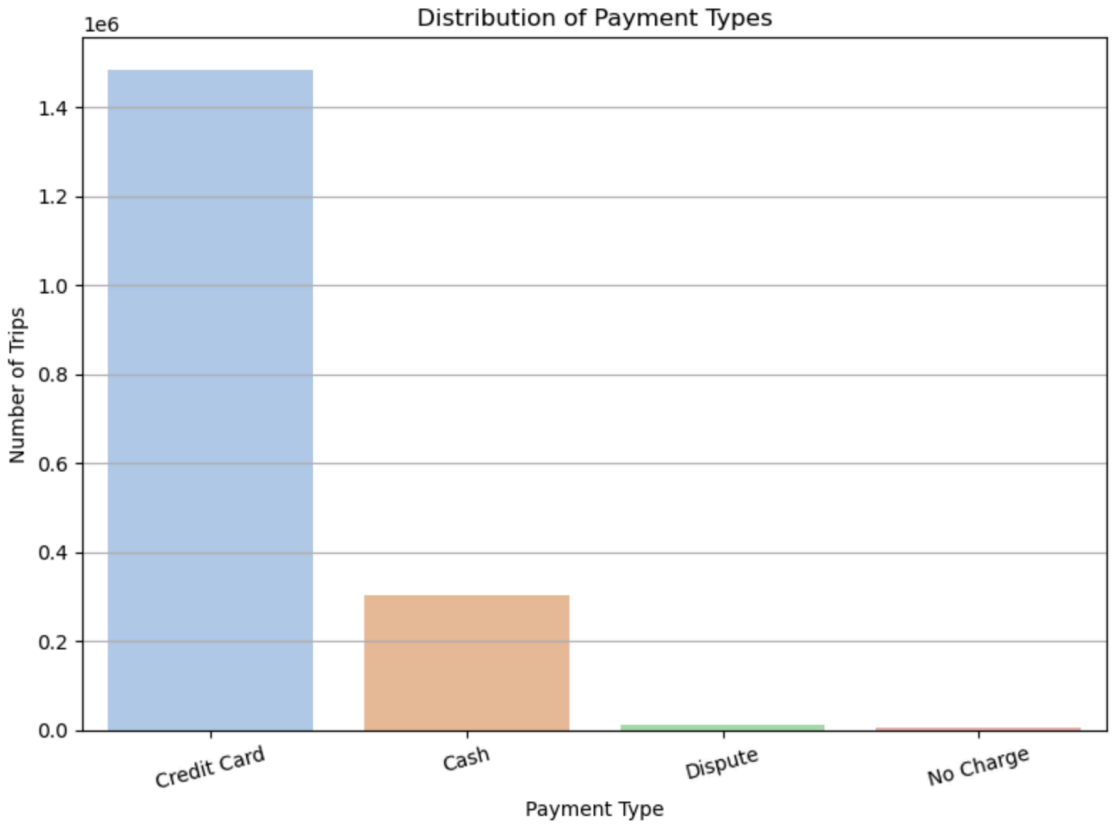
- Compared trips with `<10%` and `>25%` tip.
- Higher tip trips had longer distances and durations.

---

# Payment Method Breakdown

## 12. Payment Type Analysis

- **Credit Card** (Type 1) is the most common payment mode.

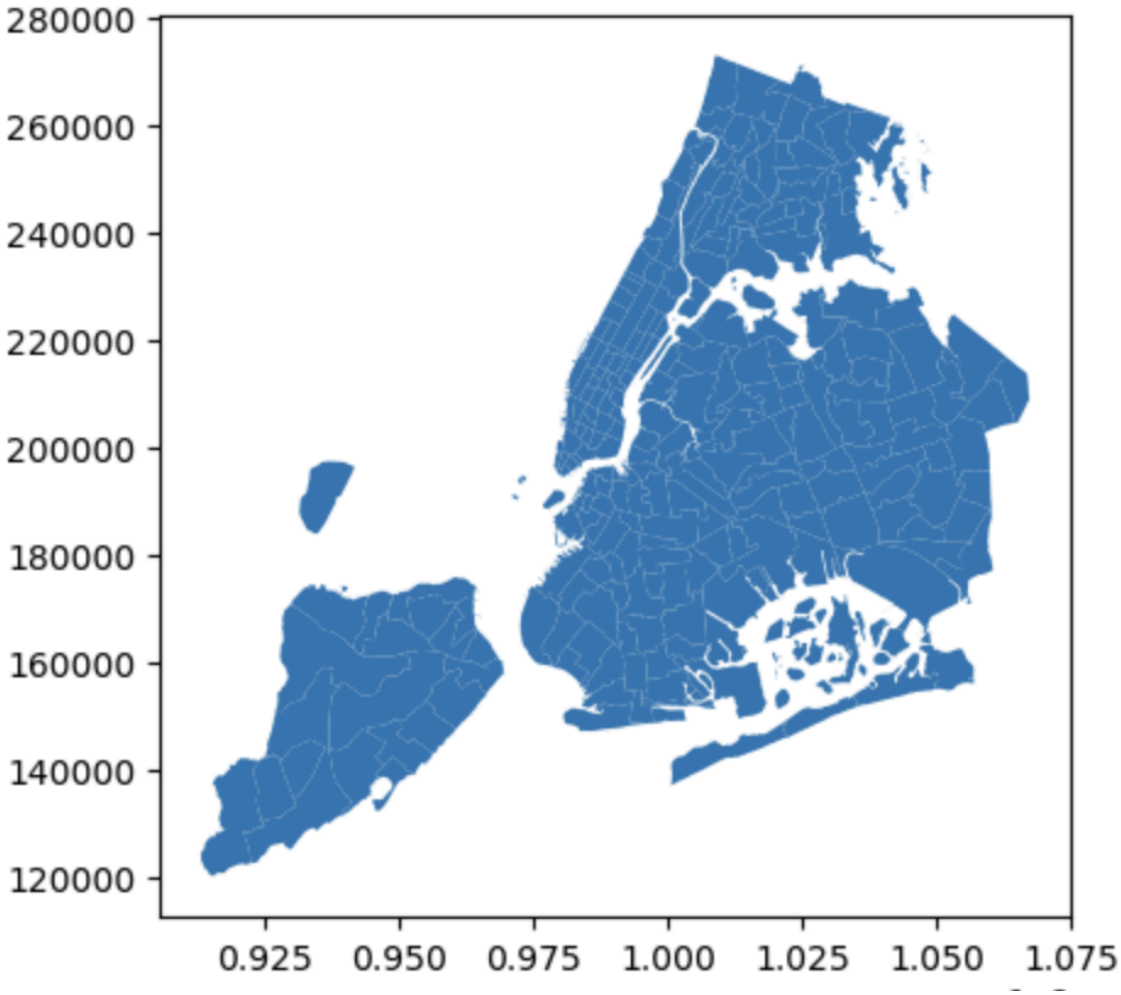



---

# Location

## 13. Trips by Pickup Zone

- Top zones include airports, midtown, and downtown.
- Zone plot:

- 


---

### 14. Zone Information

- Number of trips per zone:

| | zone | trip_count |
|---|---|---|
| 221 | Upper East Side South | 89606 |
| 148 | Midtown Center | 86583 |
| 220 | Upper East Side North | 78528 |
| 149 | Midtown East | 67050 |
| 172 | Penn Station/Madison Sq West | 65338 |
| 130 | Lincoln Square East | 62222 |
| 126 | LaGuardia Airport | 62216 |
| 214 | Times Sq/Theatre District | 61151 |
| 157 | Murray Hill | 55728 |
| 150 | Midtown North | 54667 |

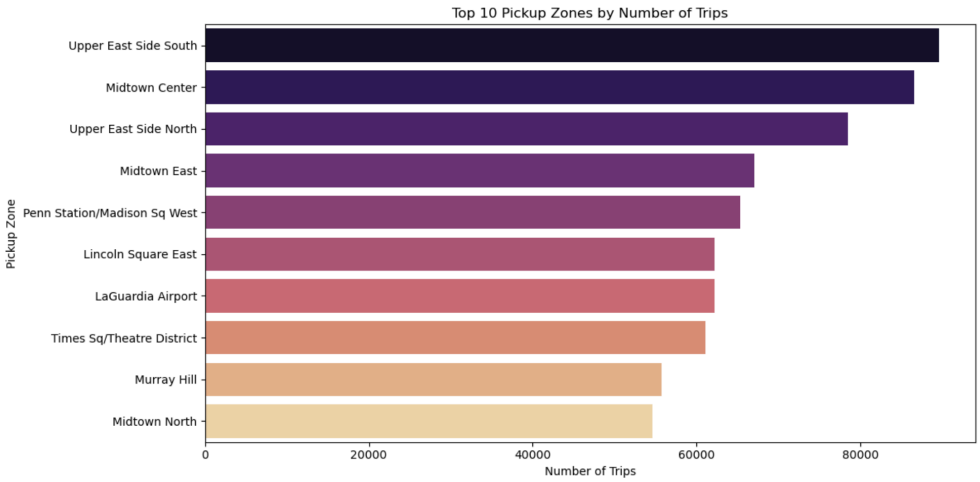
- Top Zones with highest number of trips:

| | zone | borough | trip_count |
|---|---|---|---|
| **236** | Upper East Side South | Manhattan | 89606 |
| **160** | Midtown Center | Manhattan | 86583 |
| **235** | Upper East Side North | Manhattan | 78528 |
| **161** | Midtown East | Manhattan | 67050 |
| **185** | Penn Station/Madison Sq West | Manhattan | 65338 |
| **141** | Lincoln Square East | Manhattan | 62222 |
| **137** | LaGuardia Airport | Queens | 62216 |
| **229** | Times Sq/Theatre District | Manhattan | 61151 |
| **169** | Murray Hill | Manhattan | 55728 |
| **162** | Midtown North | Manhattan | 54667 |

- Top pickup zones

Top 10 pickup zones

| | LocationID | num_pickups | zone |
|---|---|---|---|
| **0** | 237 | 89606 | Upper East Side South |
| **1** | 161 | 86583 | Midtown Center |
| **2** | 236 | 78528 | Upper East Side North |
| **3** | 162 | 67050 | Midtown East |
| **4** | 186 | 65338 | Penn Station/Madison Sq West |
| **5** | 142 | 62222 | Lincoln Square East |
| **6** | 138 | 62216 | LaGuardia Airport |
| **7** | 230 | 61151 | Times Sq/Theatre District |
| **8** | 170 | 55728 | Murray Hill |
| **9** | 163 | 54667 | Midtown North |

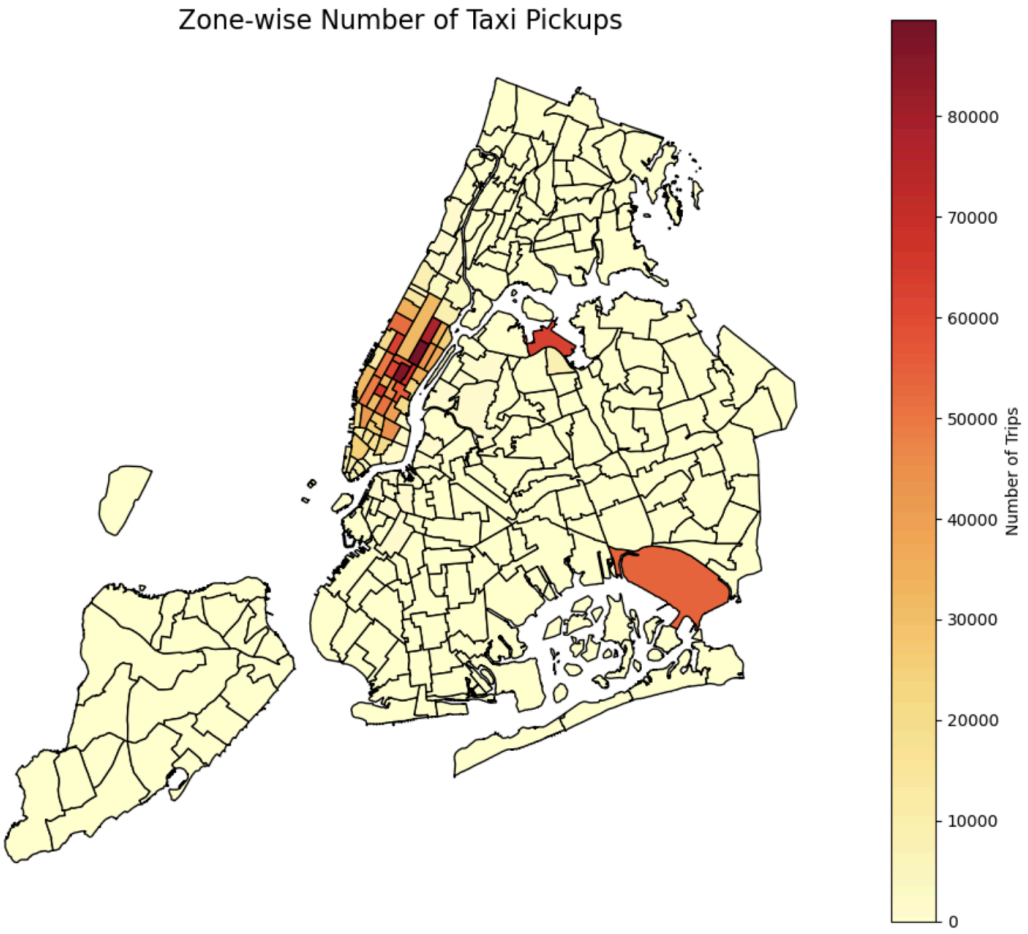



- Top drop zones:

Top 10 dropoff zones

| | LocationID | num_dropoffs | zone |
|---|---|---|---|
| **0** | 236 | 82690 | Upper East Side North |
| **1** | 237 | 79552 | Upper East Side South |
| **2** | 161 | 72239 | Midtown Center |
| **3** | 170 | 54881 | Murray Hill |
| **4** | 230 | 54711 | Times Sq/Theatre District |
| **5** | 162 | 52584 | Midtown East |
| **6** | 142 | 52232 | Lincoln Square East |
| **7** | 239 | 51414 | Upper West Side South |
| **8** | 141 | 48820 | Lenox Hill West |
| **9** | 68 | 46812 | East Chelsea |

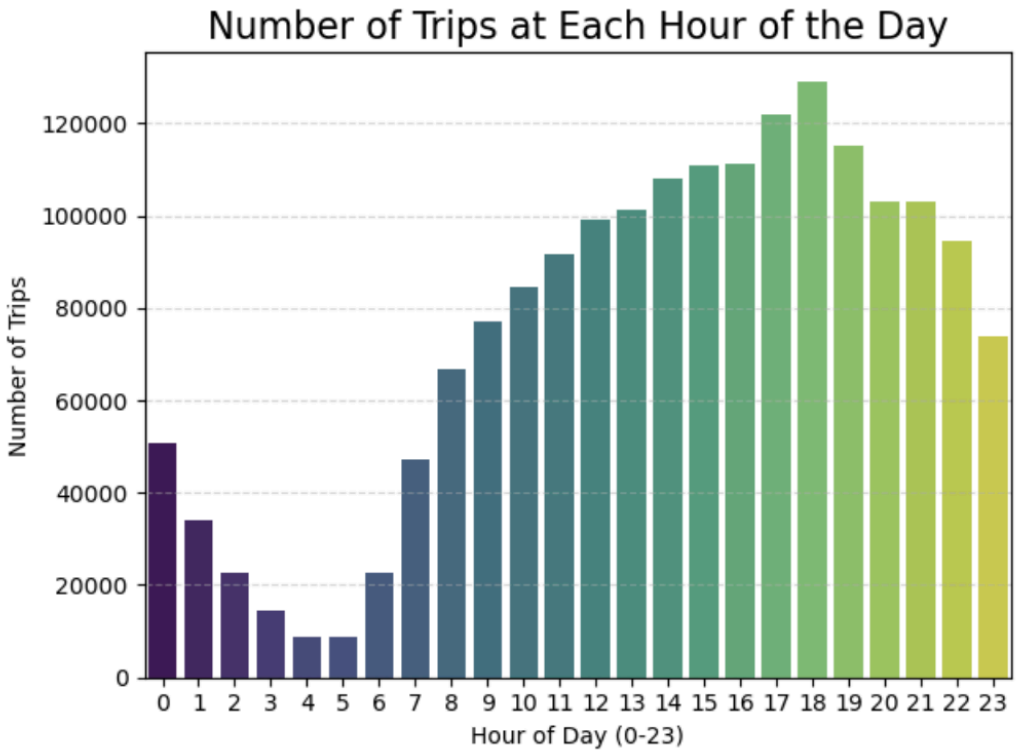
- zone-wise trips:



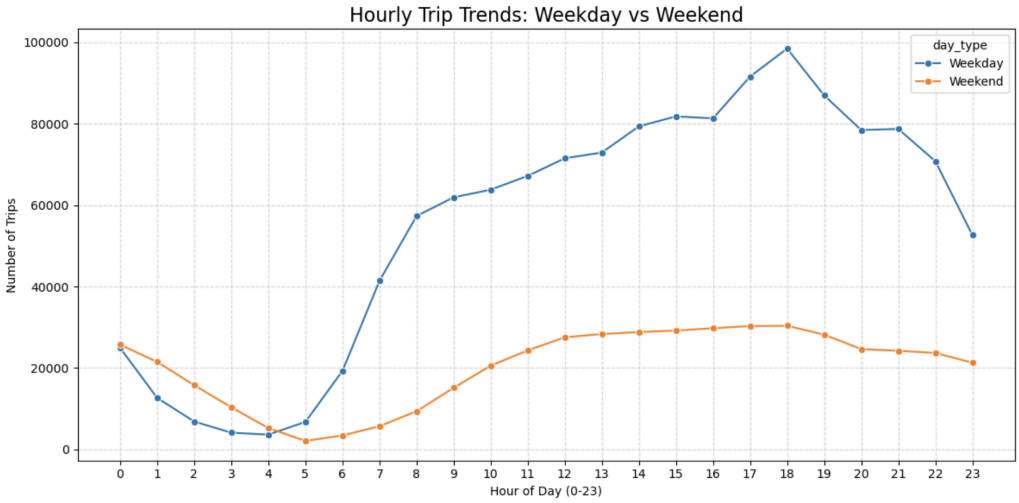

- Routes with the slowest speed:

| | PULocationID | DOLocationID | hour | trip_duration_hrs | trip_distance | avg_speed_mph |
|---|---|---|---|---|---|---|
| **88835** | 213 | 32 | 21 | 59.277778 | 1.160 | 0.019569 |
| **116006** | 261 | 87 | 4 | 22.908889 | 0.510 | 0.022262 |
| **13958** | 50 | 237 | 1 | 23.897222 | 2.350 | 0.098338 |
| **32006** | 97 | 195 | 15 | 23.608333 | 2.550 | 0.108013 |
| **20967** | 74 | 43 | 17 | 11.965417 | 1.345 | 0.112407 |
| **9936** | 45 | 233 | 22 | 23.903889 | 2.780 | 0.116299 |
| **62619** | 145 | 82 | 0 | 23.603889 | 2.800 | 0.118625 |
| **23248** | 76 | 72 | 6 | 13.936389 | 1.700 | 0.121983 |
| **48078** | 137 | 90 | 3 | 11.901111 | 1.525 | 0.128139 |
| **70161** | 161 | 162 | 5 | 4.712833 | 0.624 | 0.132404 |

- Number of trips per each hour:

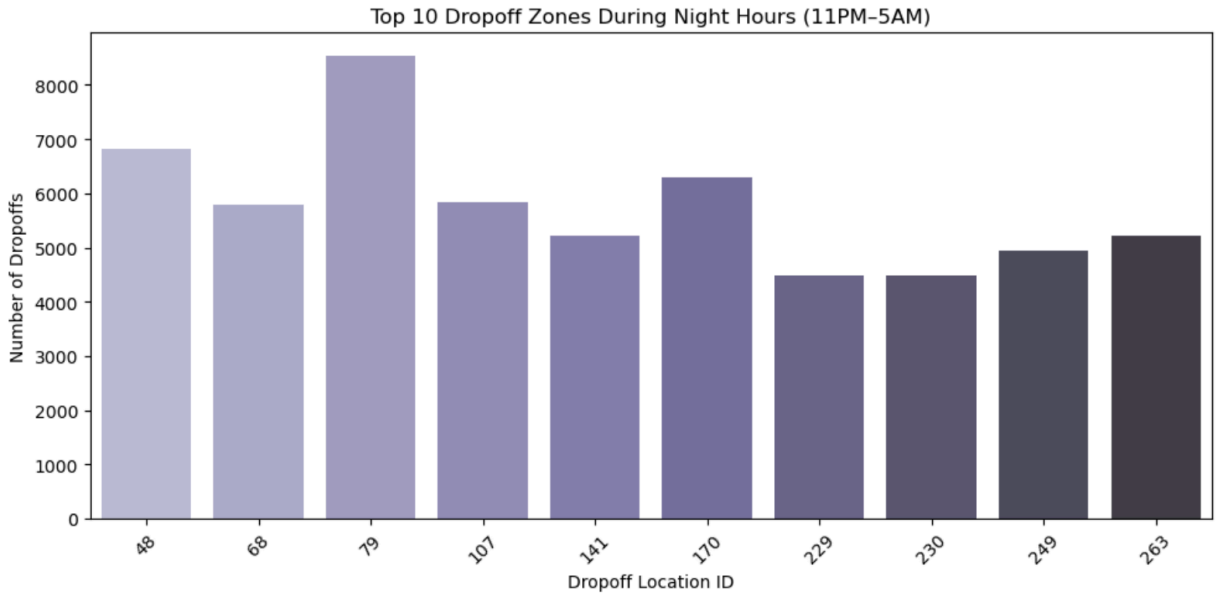



- Trips: weekday vs weekend




---

## 15. Night-Time Zones (11PM to 5AM)




- Nighttime top pickup and drop zones:

```
Top 10 Nighttime Pickup Zones:
                          zone  night_pickups
0                 East Village          15938
1                 West Village          12942
2                 Clinton East          10483
3              Lower East Side          10065
4                  JFK Airport           9266
5       Greenwich Village South           8995
6     Times Sq/Theatre District           8341
7  Penn Station/Madison Sq West           7187
8                Midtown South           6261
9                 East Chelsea           6176

Top 10 Nighttime Dropoff Zones:
                          zone  night_dropoffs
0                 East Village            8537
1                 Clinton East            6809
2                  Murray Hill            6287
3                     Gramercy            5826
4                 East Chelsea            5796
5               Lenox Hill West            5223
6                Yorkville West            5211
7                  West Village            4948
8  Sutton Place/Turtle Bay North            4495
9      Times Sq/Theatre District            4482
```

## 16. Pickup to Dropoff Ratio

```
Top 10 pickup/dropoff ratio zones:
                           zone  pickup_dropoff_ratio  num_pickups  \
70                 East Elmhurst              9.311703       8036.0
128                  JFK Airport              5.775360      54144.0
134            LaGuardia Airport              2.951703      62216.0
181  Penn Station/Madison Sq West              1.606501      65338.0
110       Greenwich Village South              1.401011      24938.0
42                  Central Park              1.381344      31706.0
243                 West Village              1.359356      41773.0
157                 Midtown East              1.275103      67050.0
100              Garment District              1.213778      30728.0
156                Midtown Center              1.198563      86583.0

     num_dropoffs
70          863.0
128        9375.0
134       21078.0
181       40671.0
110       17800.0
42        22953.0
243       30730.0
157       52584.0
100       25316.0
156       72239.0

Bottom 10 pickup/dropoff ratio zones:
                zone  pickup_dropoff_ratio  num_pickups  num_dropoffs
0     Newark Airport              0.009709          1.0         103.0
73     East Flushing              0.027972          4.0         143.0
251  Windsor Terrace              0.032787         20.0         610.0
246       Whitestone              0.036415         13.0         357.0
192        Ridgewood              0.038576         39.0        1011.0
155   Middle Village              0.039249         23.0         586.0
259              NaN              0.043680        113.0        2587.0
30        Bronx Park              0.044643          5.0         112.0
108       Greenpoint              0.046310        187.0        4038.0
36    Bushwick South              0.047248         91.0        1926.0
```

---

## 17. Zone-wise Passenger Count

- fare per mile per passenger for different passenger counts:

| | passenger_count | fare_per_mile_per_passenger |
|---|---|---|
| 0 | 1.0 | 7.839757 |
| 1 | 2.0 | 3.888855 |
| 2 | 3.0 | 2.615773 |
| 3 | 4.0 | 1.972602 |
| 4 | 5.0 | 1.526023 |
| 5 | 6.0 | 1.293405 |

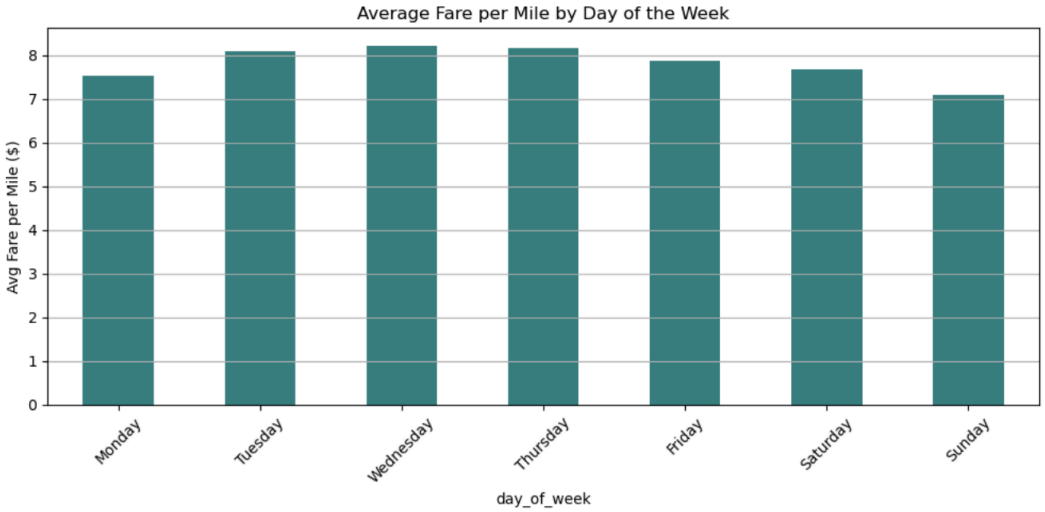
---

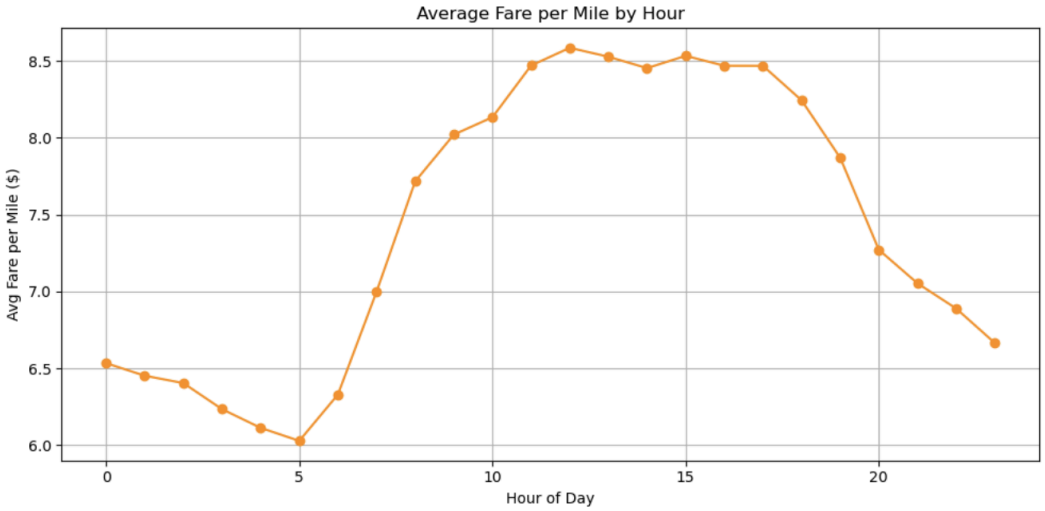
# Fare Efficiency

## 18. Fare per Mile

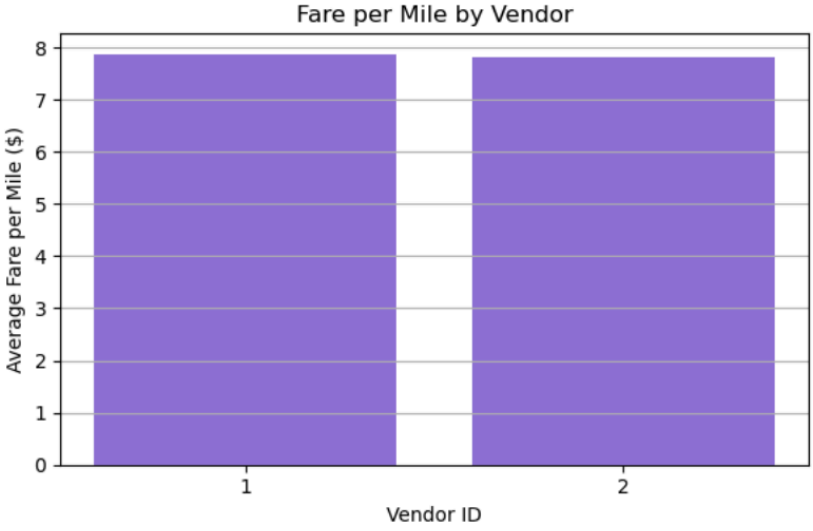
- Average fare per mile for different days and for different times of the day:

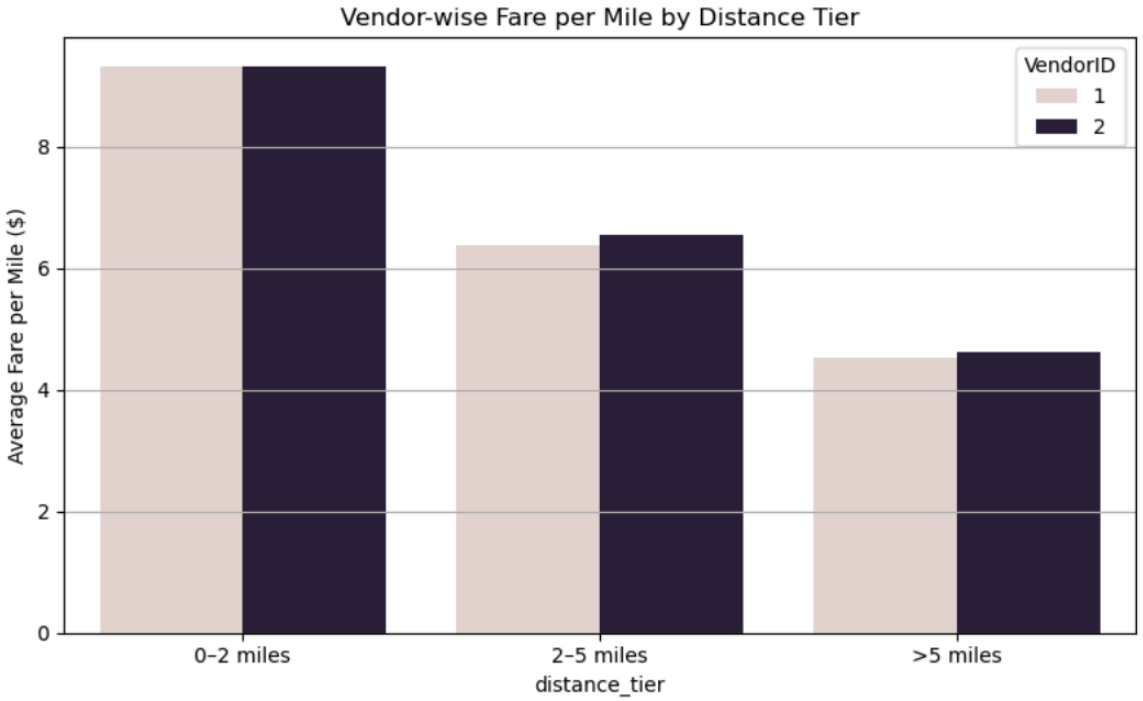
```
day_of_week
Monday       7.528458
Tuesday      8.093732
Wednesday    8.204448
Thursday     8.152178
Friday       7.873955
Saturday     7.679405
Sunday       7.089355
Name: fare_per_mile, dtype: float64
hour
0     6.533907
1     6.451990
2     6.404206
3     6.234563
4     6.112821
5     6.028163
6     6.329691
7     6.998509
8     7.716850
9     8.021019
10    8.134187
11    8.470535
12    8.585718
13    8.527780
14    8.453685
15    8.533609
16    8.469107
17    8.468624
18    8.244191
19    7.870212
20    7.271458
21    7.053337
22    6.889485
23    6.665359
```




| | VendorID | fare_per_mile |
|---|---|---|
| 0 | 1 | 7.869566 |
| 1 | 2 | 7.819692 |



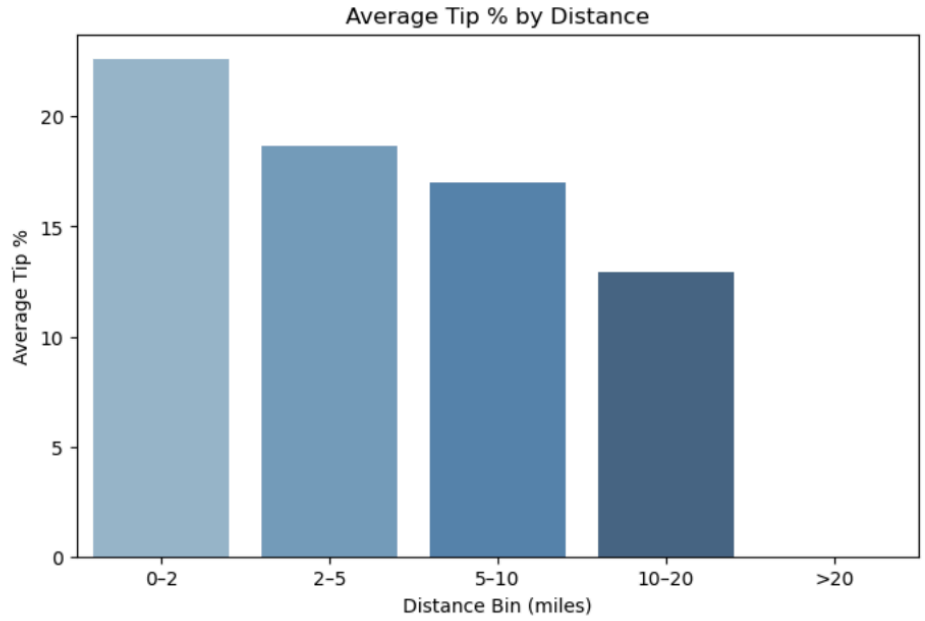



---

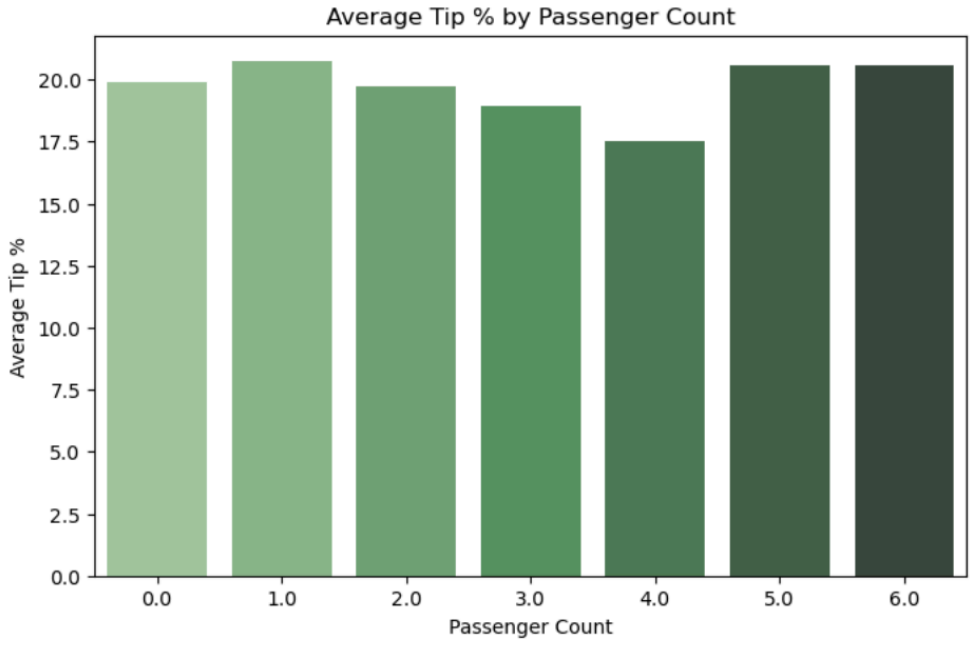
### 19. Tip percentages based on distances, passenger counts and pickup times

- Short trips (0–2 mi): Higher fare/mile
- Long trips: Lower fare/mile

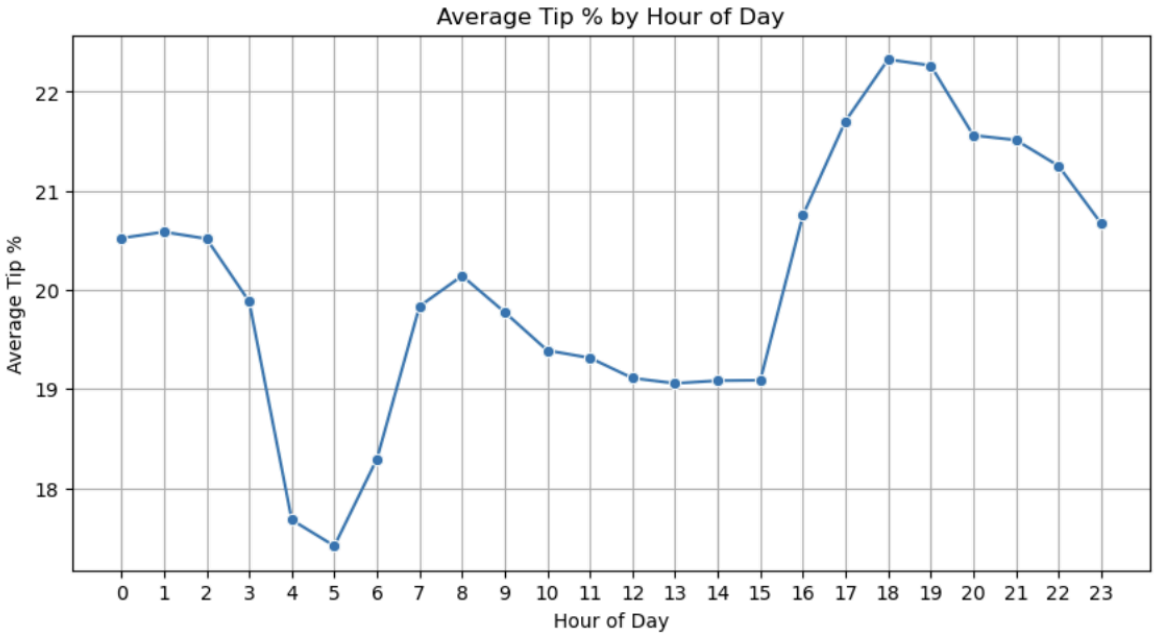
- Distance:




- Passenger Count:



- Hour of the Day:

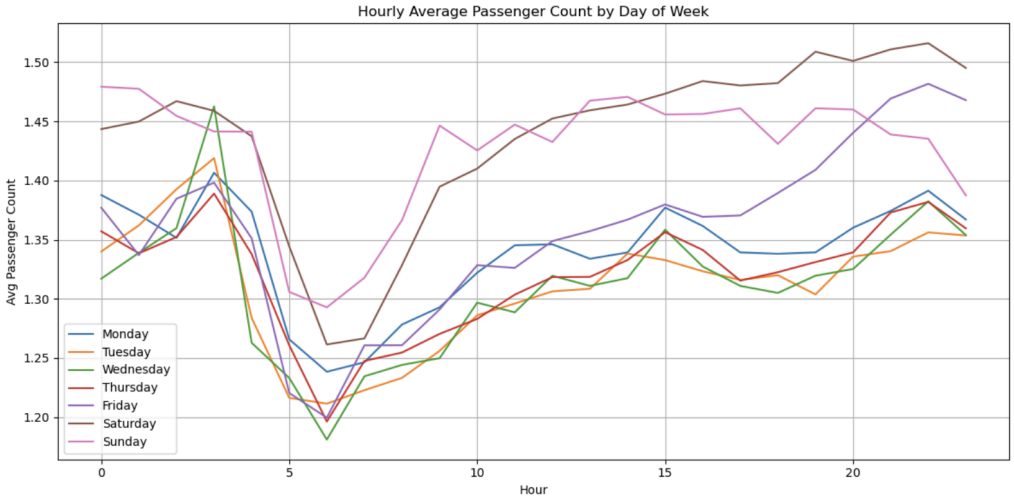



- Trips with tip percentage < 10% to trips with tip percentage > 25%:

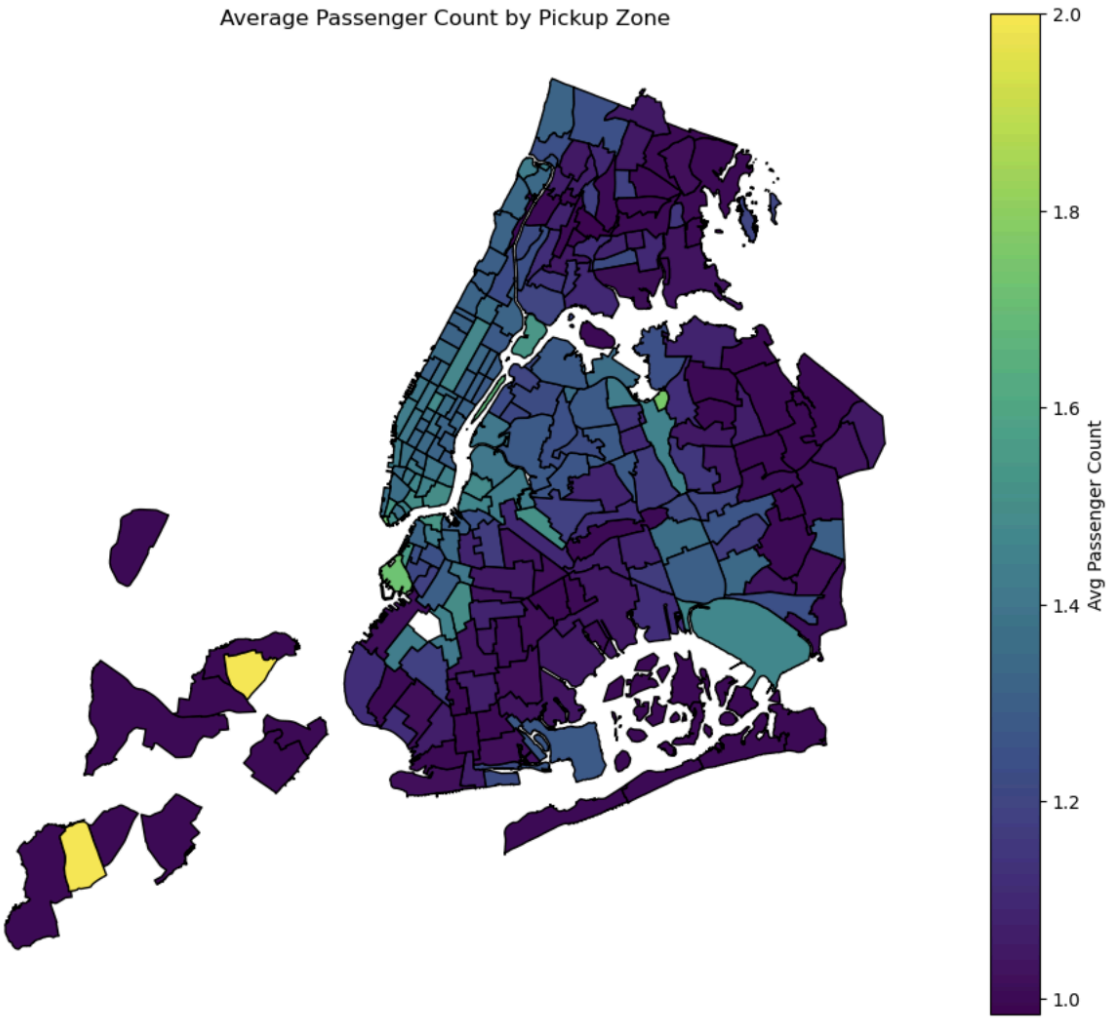
Comparison of key features for low and high tip percentage trips:

| | Low Tip (<10%) | High Tip (>25%) |
|---|---|---|
| **fare_amount** | 20.084637 | 13.371235 |
| **trip_distance** | 3.528294 | 2.034643 |
| **passenger_count** | 1.414962 | 1.346668 |
| **hour** | 13.978006 | 14.624679 |

20. How passenger count varies across hours and days:

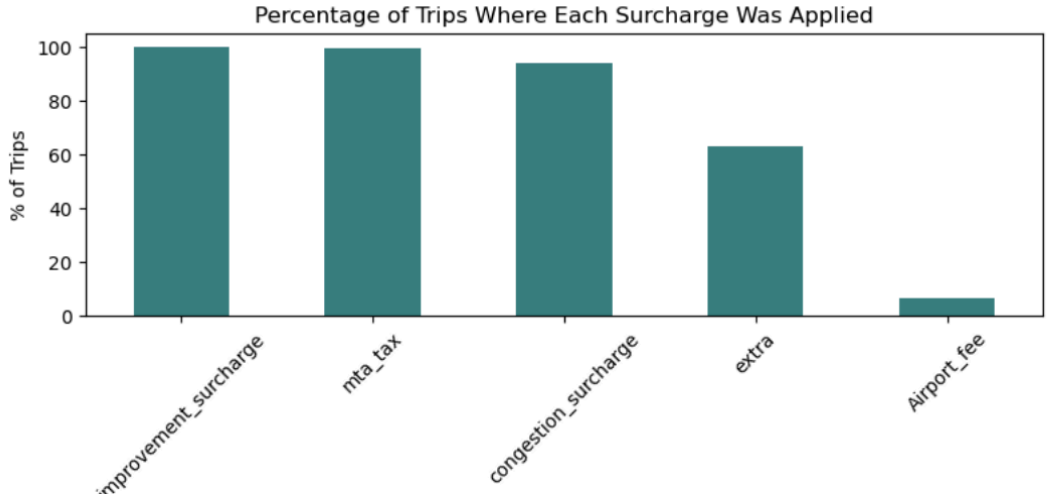


21. How does passenger count vary across zones:



## 22. How often is each surcharge applied?

| | applied_count | applied_pct |
|---|---|---|
| improvement_surcharge | 1799807.0 | 100.00 |
| mta_tax | 1797305.0 | 99.86 |
| congestion_surcharge | 1697159.0 | 94.30 |
| extra | 1135762.0 | 63.10 |
| Airport_fee | 118983.0 | 6.61 |




---

# Conclusion:

1. Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies
    - Shift Cabs with the Clock: Demand peaks at predictable hours — align driver shifts accordingly to cover rush hours efficiently.
    - Zone-Based Rebalancing: Regularly shift idle cabs toward hotspots like airports and entertainment districts using real-time trip data.
    - Avoid Known Slow Routes: Use historical speed data to reroute around bottlenecks, especially during peak hours.
    - Match Vehicle Size to Party Size: Most trips are 1–2 passengers. Save larger vehicles for when they're really needed.
    - Late-Night Focus: Certain areas stay active past midnight — dedicate a lean night fleet to serve those zones.
    - Use Tip Trends: Trips with higher tips often mean better routes or service. Study those to train and guide drivers.
    - Monitor Fare Efficiency: Keep an eye on fare-per-mile, especially on short trips, to flag any inconsistencies or overcharging.
    - Forecast Demand Spikes: Use past trends to predict where and when surges will hit — prep the fleet in advance.

2. Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.
    - Follow the Flow: Position more cabs in zones with consistently high pickups during morning and evening peaks.
    - Weekend Rebalancing: Entertainment and nightlife zones heat up on weekends — shift availability accordingly.
    - Zone-Level Micro-Targeting: Use high-resolution trip trends to identify under-served but high-demand micro-zones.
    - Event-Based Deployments: Keep flexible units ready to cover spikes near stadiums, theaters, or transit hubs.
    - Monthly Seasonality: Boost coverage in tourist-heavy areas during peak travel months (e.g., December, summer).
    - Off-Hour Optimization: Late-night activity clusters around fewer zones — concentrate limited night fleets there.
    - Tip & Fare Data Clues: Zones with better fare-per-mile and higher tip percentages are likely your most valuable — prioritize them.

3. Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.
    - Tiered Pricing by Distance: Short trips (0–2 mi) show higher fare-per-mile — slightly premium pricing here could lift revenue with minimal impact on demand.
    - Time-Based Adjustments: Demand spikes during commute hours and late nights suggest room for dynamic pricing during these windows.
    - Passenger Count Factor: Larger groups are often associated with higher fares — explore per-passenger surcharges for 4+ riders.
    - Low-Tip Routes: Routes or zones with consistently low tip percentages might benefit from fare adjustments or bundled pricing to compensate.
    - Weekend Premiums: Traffic and demand trends suggest weekend surcharges could be viable in high-activity zones.
    - Competitive Benchmarking: Use fare-per-mile comparisons with other vendors across common zones to fine-tune base rates without underpricing or overshooting.