EDA Assignment Report

Introduction:

This report presents a comprehensive exploratory analysis of Yellow Taxi operations in NYC for the year 2023. The analysis is based on sampled and stratified trip data extracted from monthly parquet files and includes temporal, spatial, and financial metrics. The goal is to uncover demand patterns, revenue drivers, inefficiencies, and actionable insights to optimize fleet distribution and pricing strategies.

1. Data Preparation:

- 12 monthly .parquet files (Jan–Dec 2023) were sampled by hour and date to build a representative dataset.

- Combined dataset: ~1.89 million rows

- Geospatial zone data was merged using shapefiles to map pickup and dropoff locations to NYC Taxi Zones.

- Key columns cleaned:
    - Removed invalid values (e.g., negative fares, 0 distance)
    - Handled missing values (passenger_count, RatecodeID, congestion_surcharge, etc.)
    - Removed unrealistic outliers (e.g., distance > 250 miles)

2. Data Cleaning:

- Multiple cleaning steps were performed to ensure data consistency and accuracy before analysis:
    - ○ Removed rows with missing or zero `passenger_count`, `trip_distance`, and `fare_amount`
    - ○ Handled missing values in columns like `RatecodeID`, `store_and_fwd_flag`, and `congestion_surcharge` by imputing common values or using default charges
    - ○ Dropped rows with logically incorrect values such as:
    - ○ Trips with 0 fare but non-zero distance

- Trips where both pickup and dropoff locations were identical and trip_distance was zero
- Removed outliers:
- Trips with distance > 250 miles or fare > $10,000
- Passenger counts > 7 (rare cases likely due to data entry errors)
- Ensured proper datetime formatting and extracted hour, date, and day-of-week features
- Resolved duplicate columns (e.g., airport_fee vs Airport_fee) by checking consistency and retaining the accurate one

3. Exploratory Data Analysis (EDA):

3.1 Temporal Patterns

- **Peak hours**: 5 PM–9 PM (evening commute and nightlife)
- **Morning solo rides**: 7–10 AM (weekday commuters)
- **Highest trip volume**: Fridays and Saturdays
- **Quarterly revenue peak**: Q2 and Q4

3.2 Zonal Trends

- **Top pickup zones**: Midtown Center, JFK Airport, LaGuardia, Times Square
- **Top dropoff zones**: East Village, West Village, Midtown East
- **Zones with pickup/dropoff imbalance**: JFK and East Elmhurst (source-heavy), Freshkills Park (sink zones)

3.3 Passenger Insights

- Most common: 1–2 passengers per trip
- Higher passenger counts in airports and stations
- Group rides more frequent in evenings and weekends

3.4 Financial Insights

i. **Fare per mile**:

   1. Highest for trips <2 miles due to base fare

   2. Lower for longer trips but yields higher total fares

ii. **Tip behavior**:

   1. Tip % increases with trip distance

   2. Higher tips at night and weekends

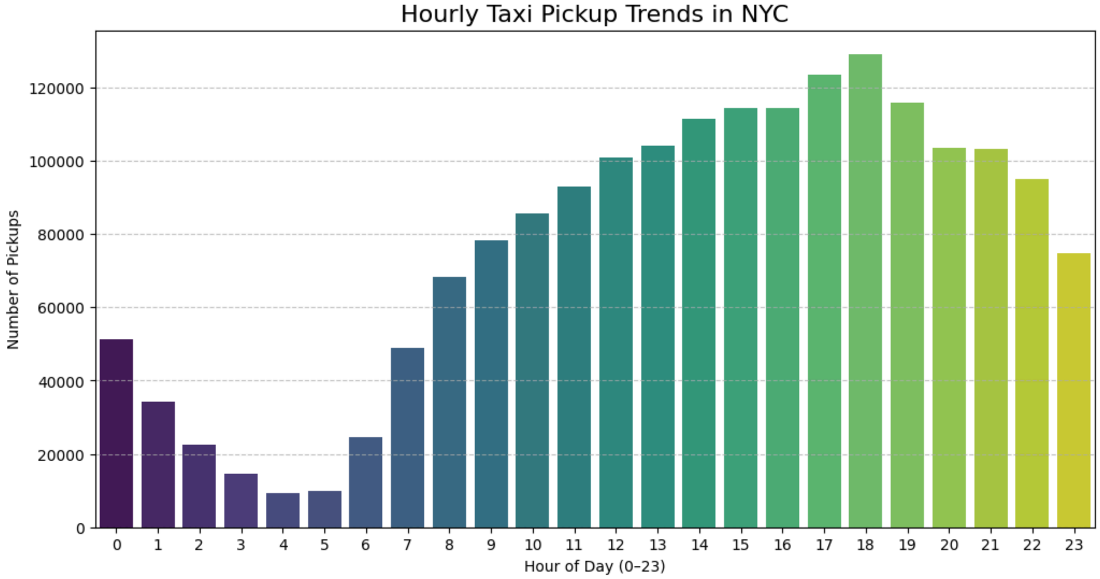
   3. Average tip %: ~18% for trips >25% tips

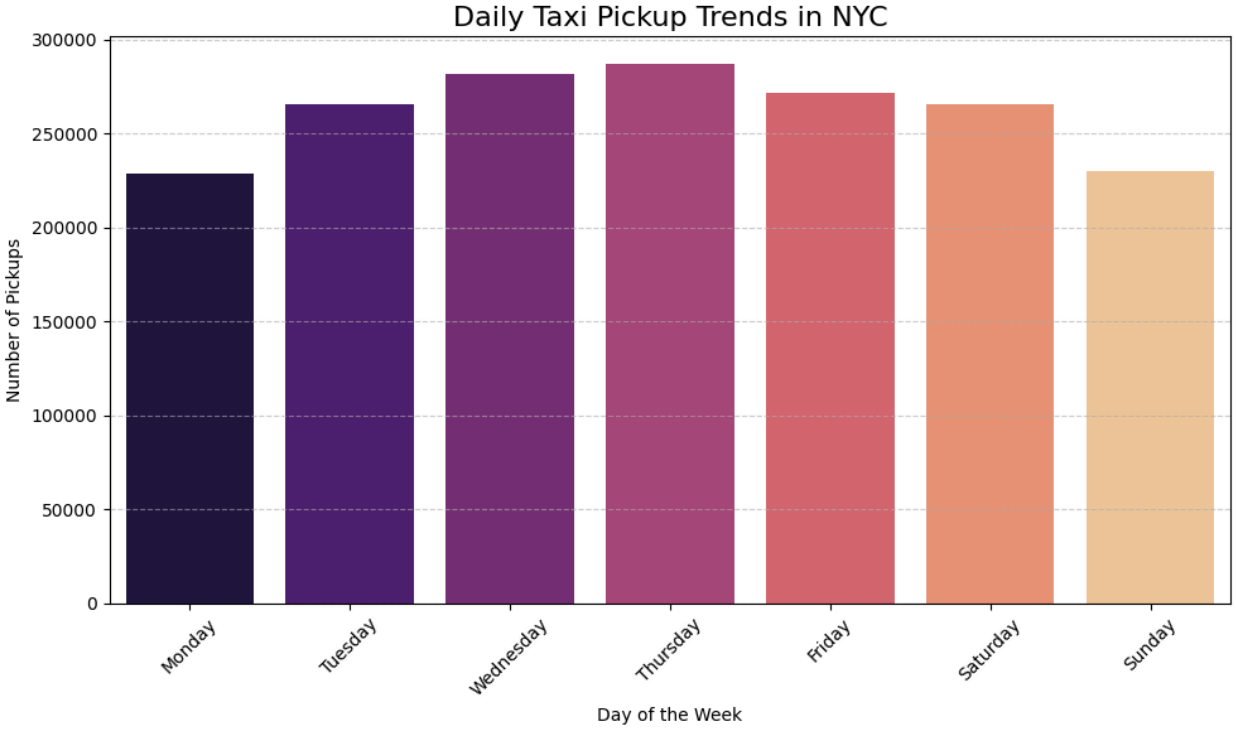
iii. **Surcharges**:

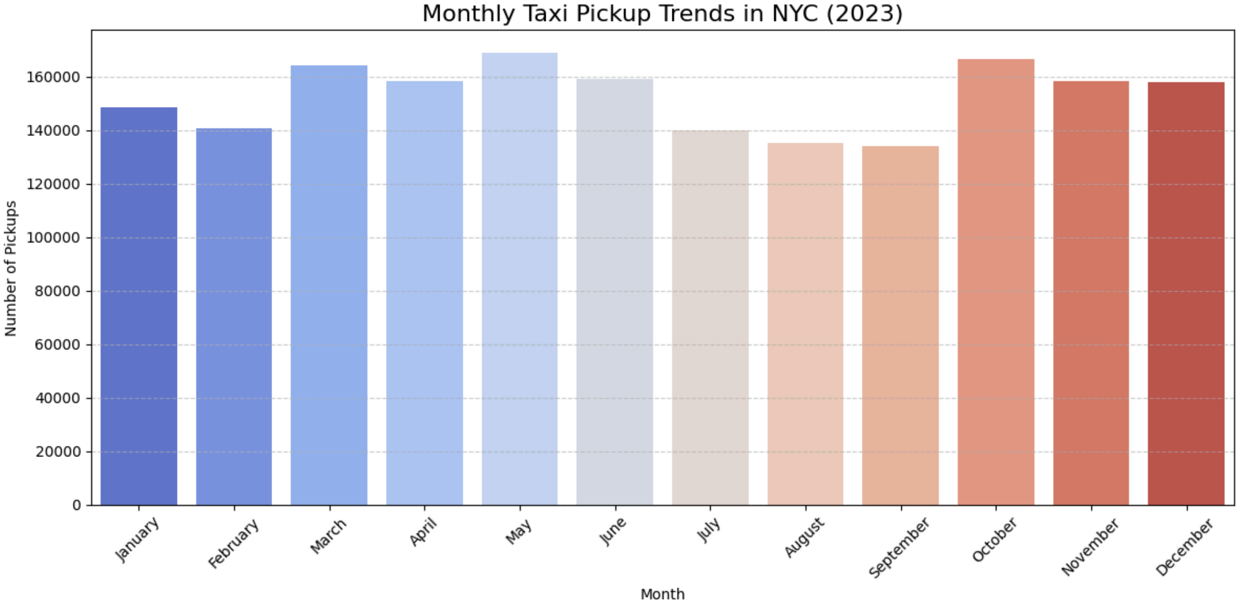
   1. Applied in predictable zones and times (congestion, airport fees, MTA tax)

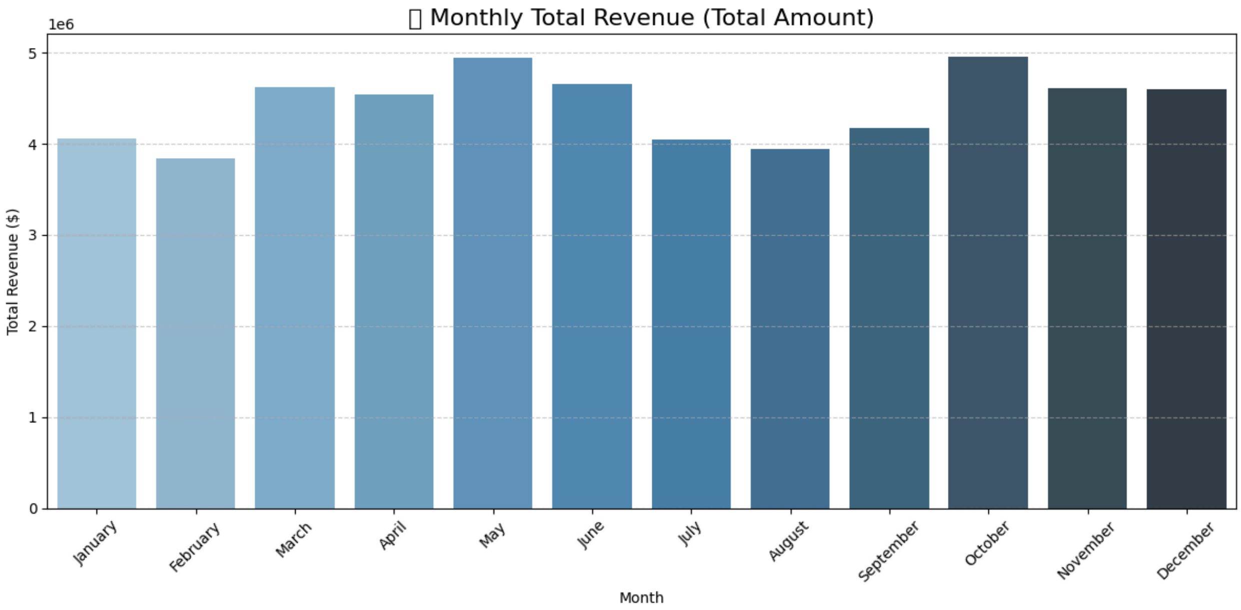
   2. Congestion surcharge mostly in Midtown during business hours

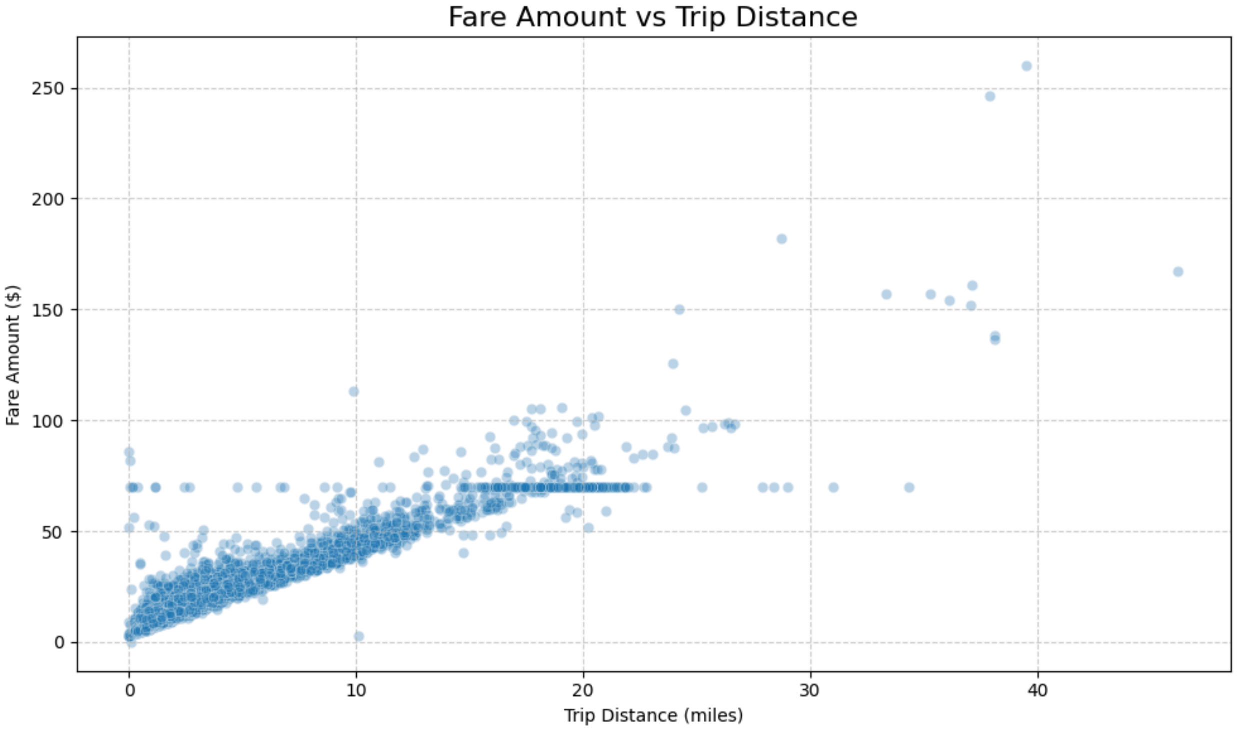
4. Visualisations:

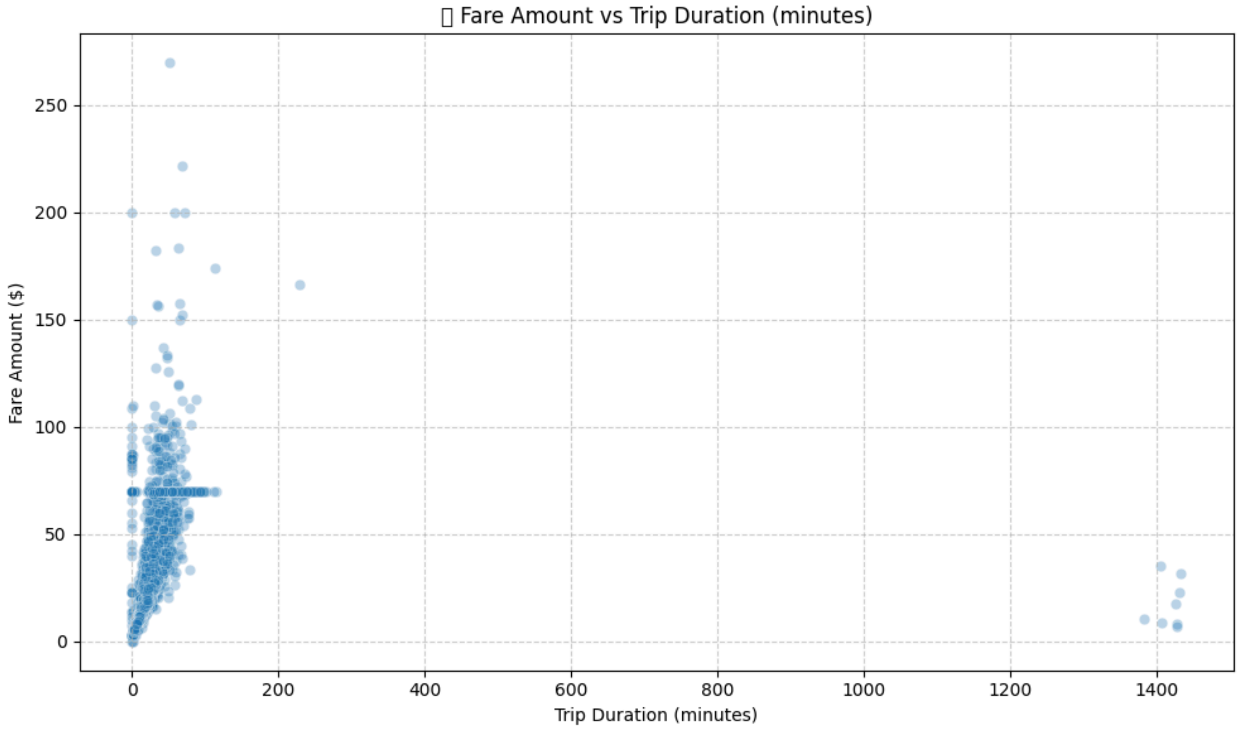


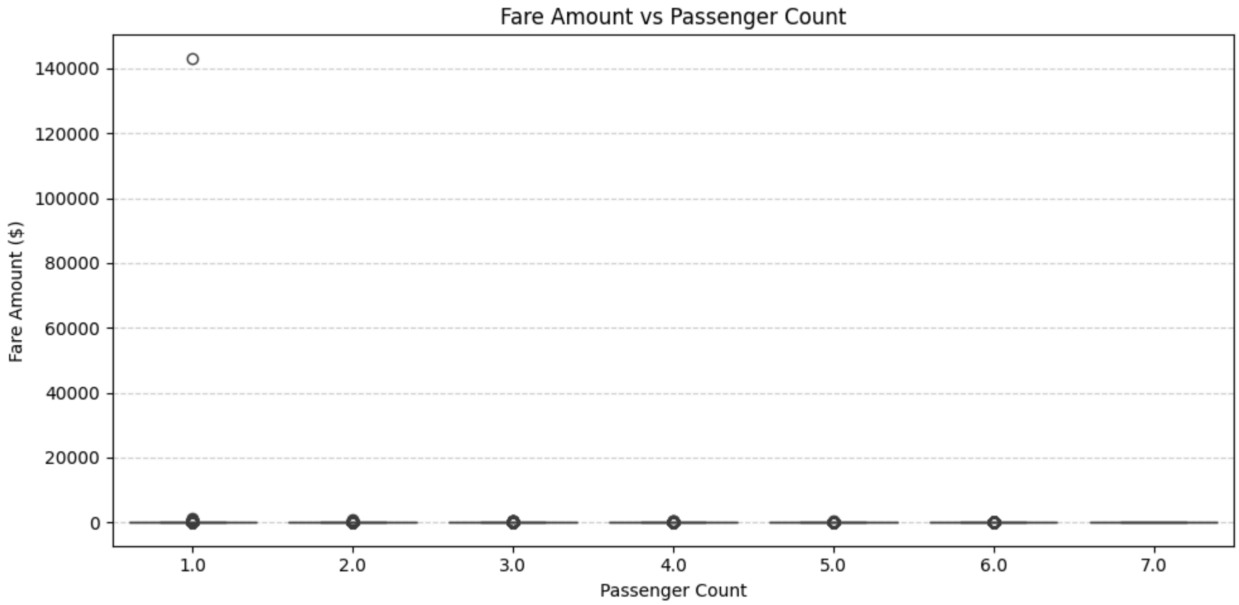

Daily Taxi Pickup Trends in NYC
300000
250000
200000
150000
100000
50000
0
Number of Pickups
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of the Week

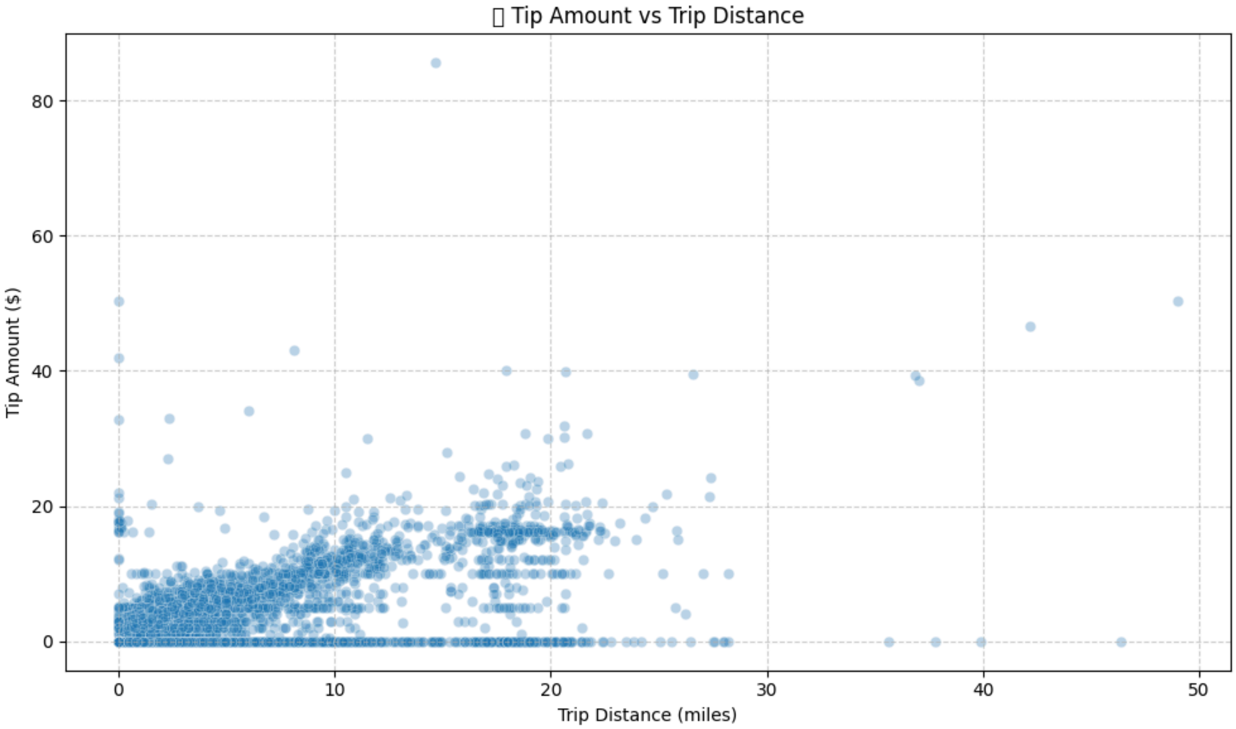

Monthly Taxi Pickup Trends in NYC (2023)
160000
140000
120000
100000
80000
60000
40000
20000
0
Number of Pickups
January
February
March
April
May
June
July
August
September
October
November
December
Month

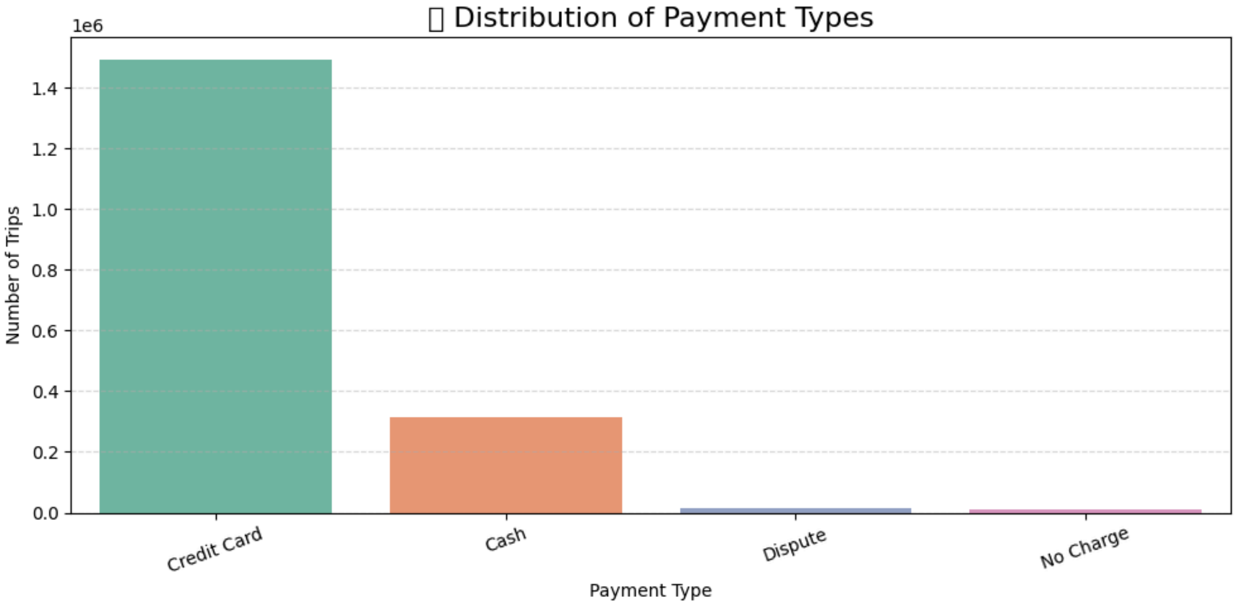

□ Monthly Total Revenue (Total Amount)
1e6
0
1
2
3
4
5
Total Revenue ($)
January
February
March
April
May
June
July
August
September
October
November
December
Month

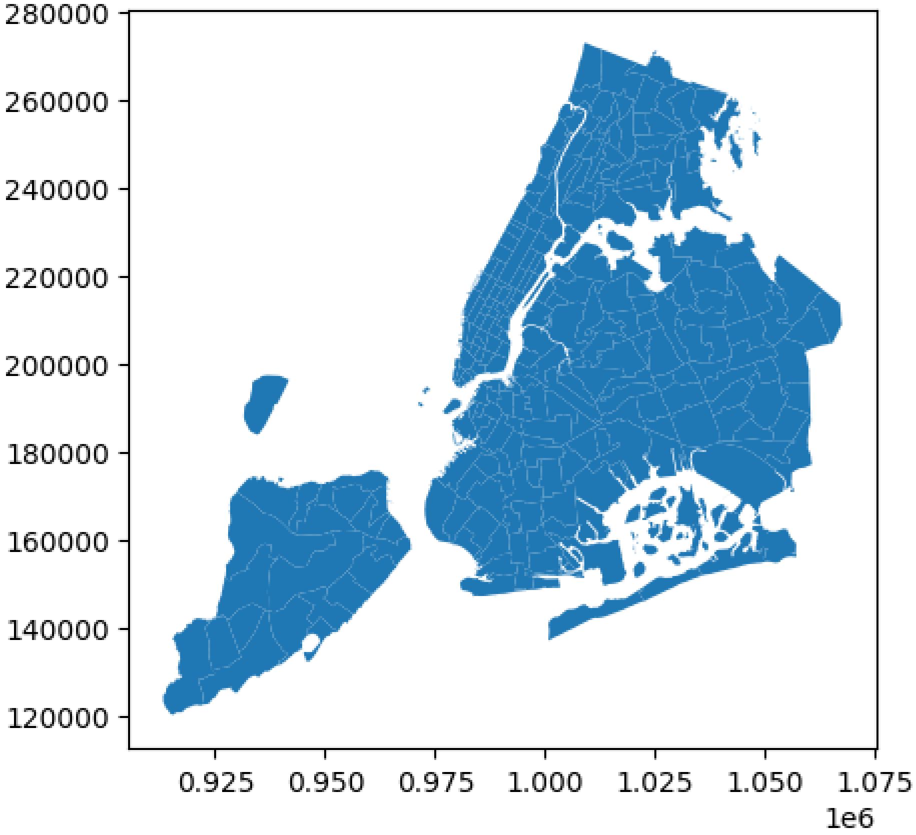

Fare Amount vs Trip Distance
0
50
100
150
200
250
Fare Amount ($)
0
10
20
30
40
Trip Distance (miles)

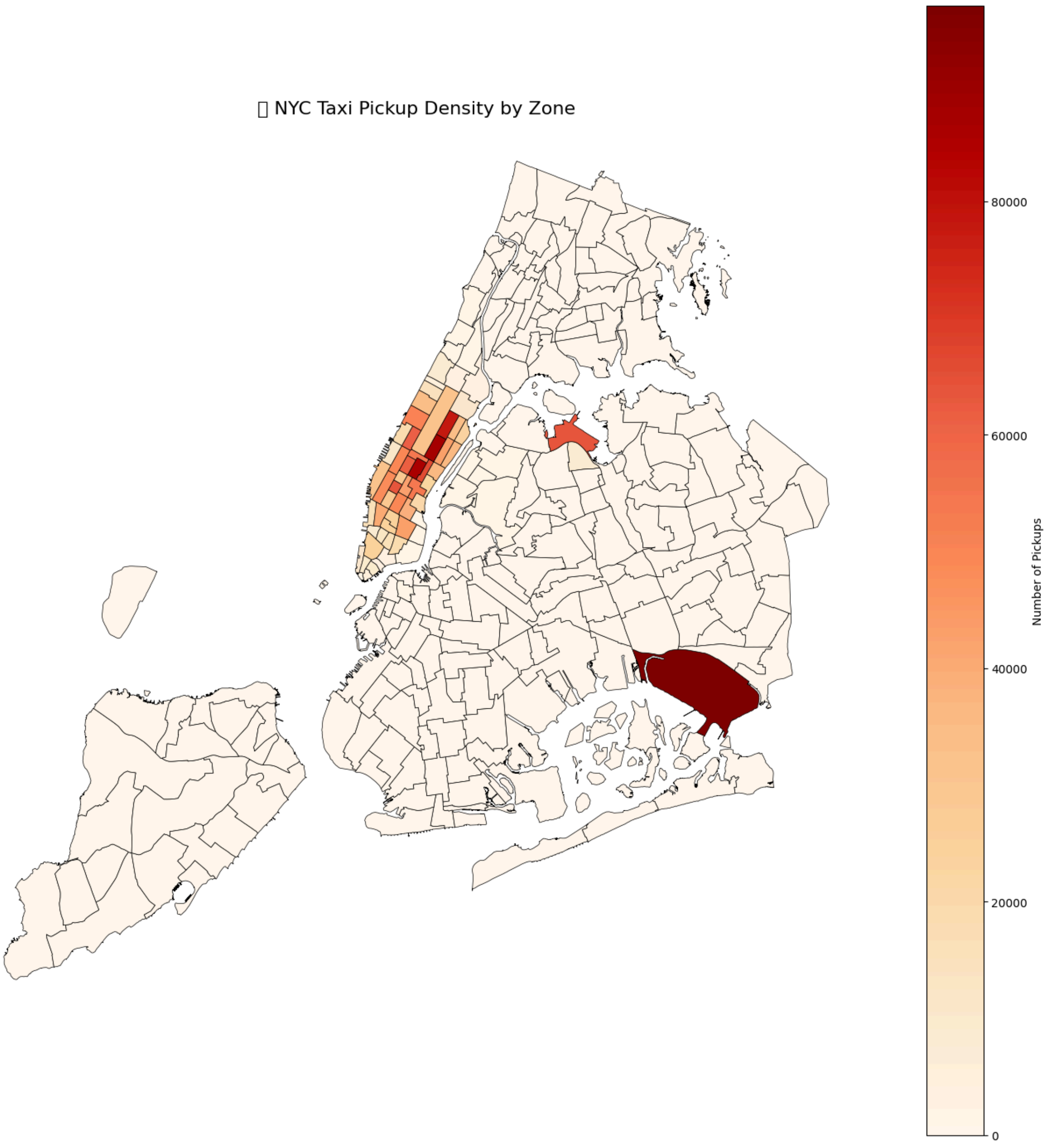

□ Fare Amount vs Trip Duration (minutes)
Fare Amount ($)
250
200
150
100
50
0
0
200
400
600
800
1000
1200
1400
Trip Duration (minutes)


Fare Amount vs Passenger Count
Fare Amount ($)
140000
120000
100000
80000
60000
40000
20000
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
Passenger Count

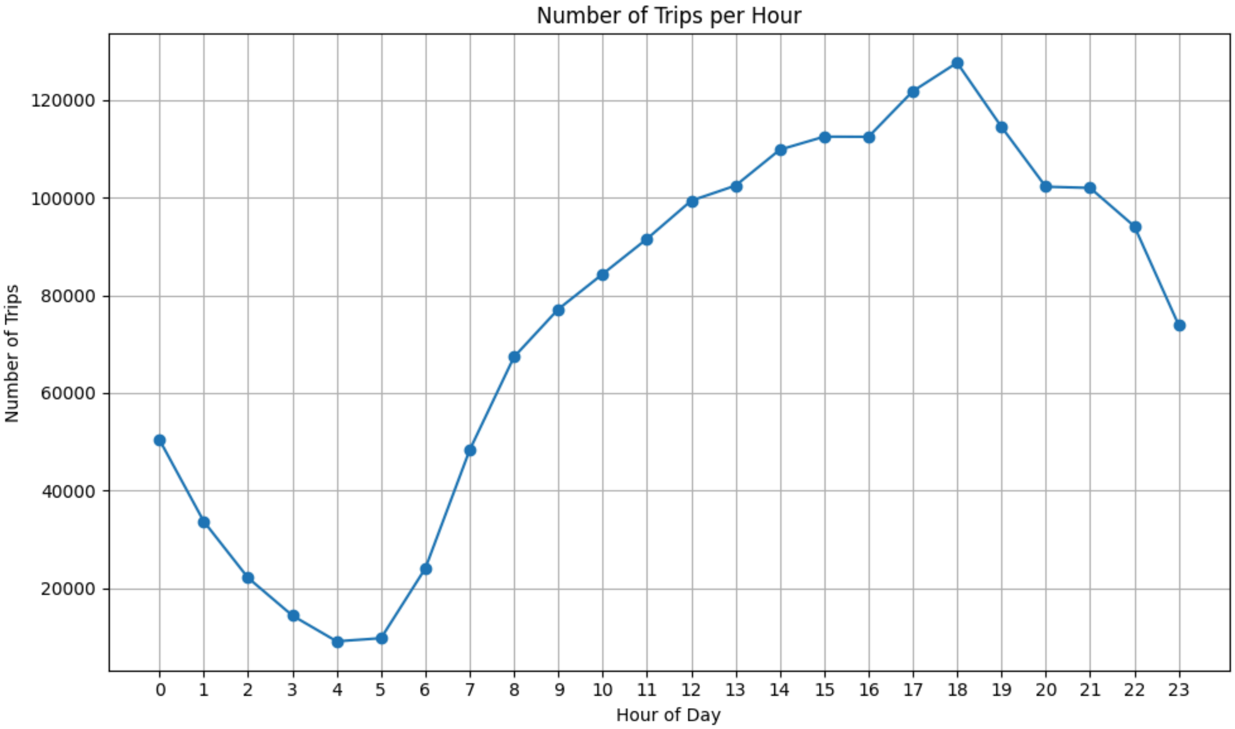

□ Tip Amount vs Trip Distance
Tip Amount ($)
80
60
40
20
0
Trip Distance (miles)
0
10
20
30
40
50

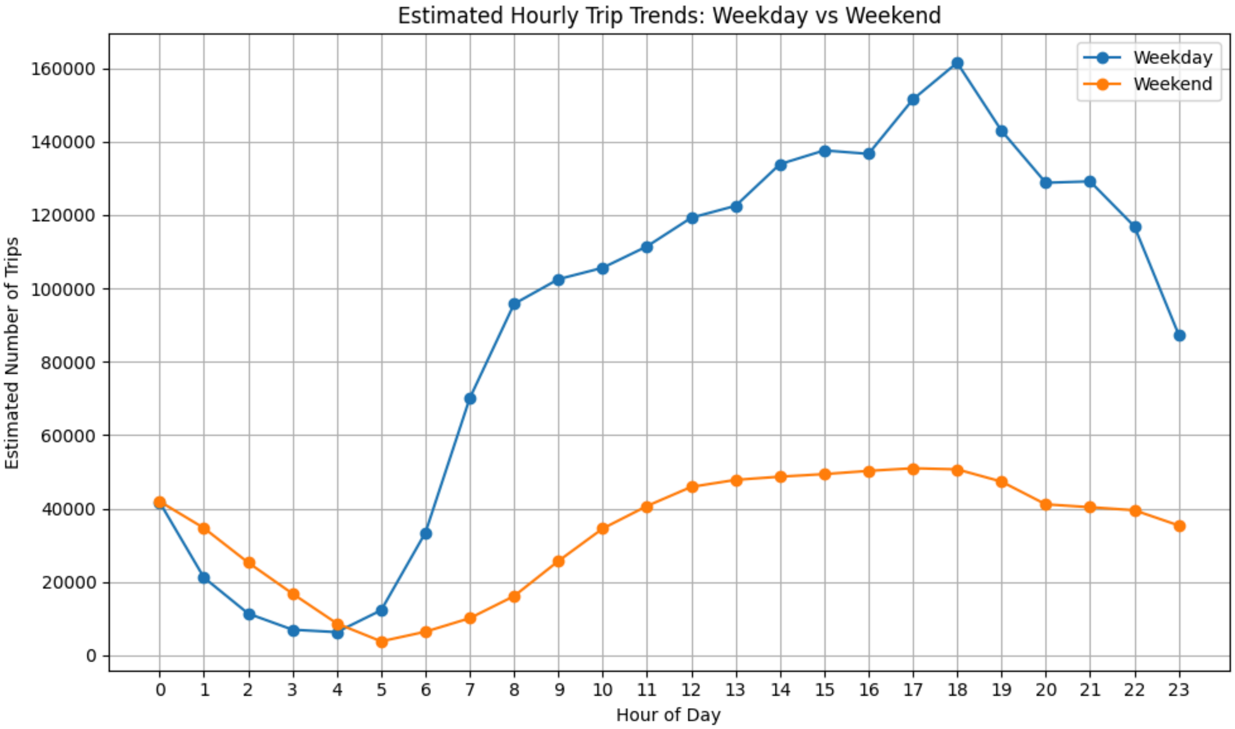

□ Distribution of Payment Types
1e6
Number of Trips
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Credit Card
Cash
Dispute
No Charge
Payment Type

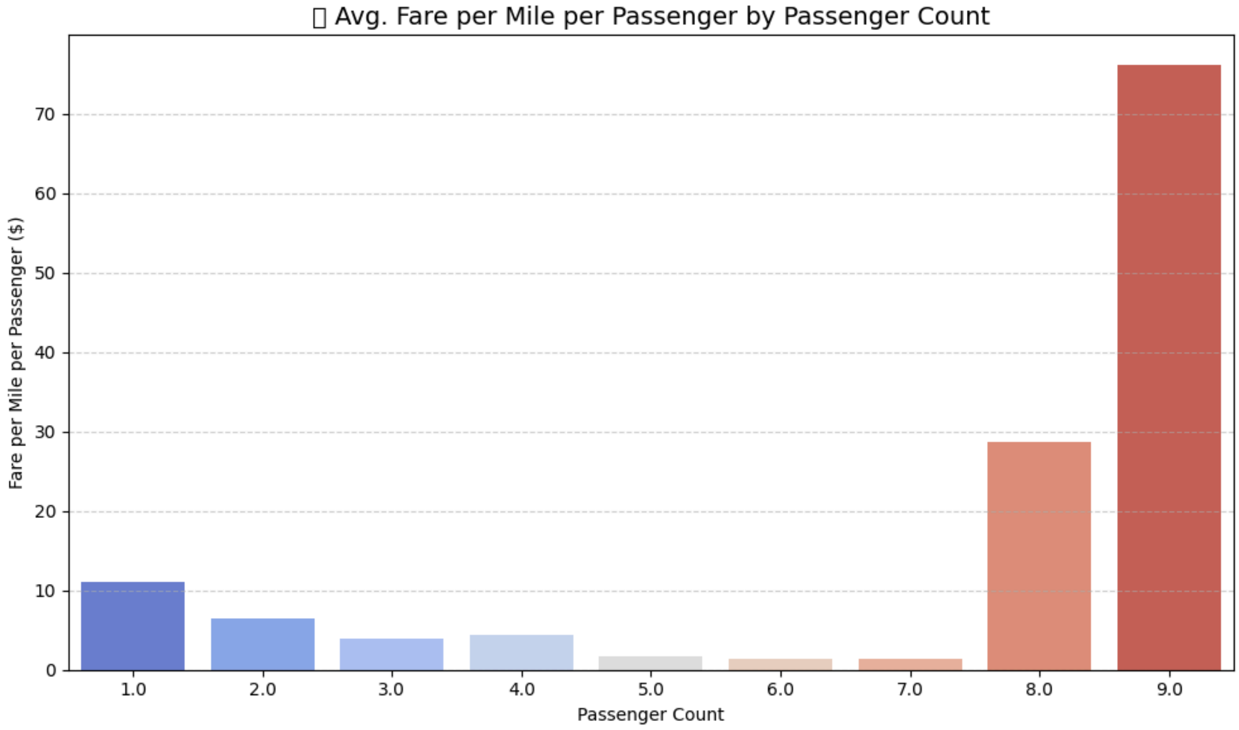

280000
260000
240000
220000
200000
180000
160000
140000
120000
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1e6



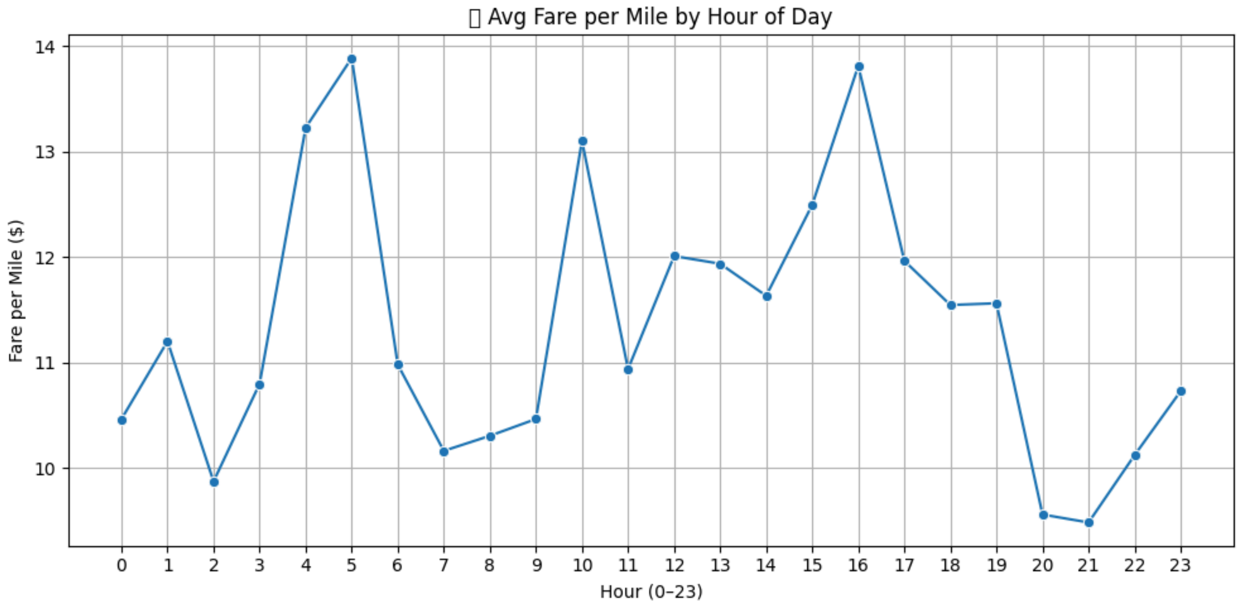
□ NYC Taxi Pickup Density by Zone


Number of Trips per Hour
Number of Trips
120000
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day

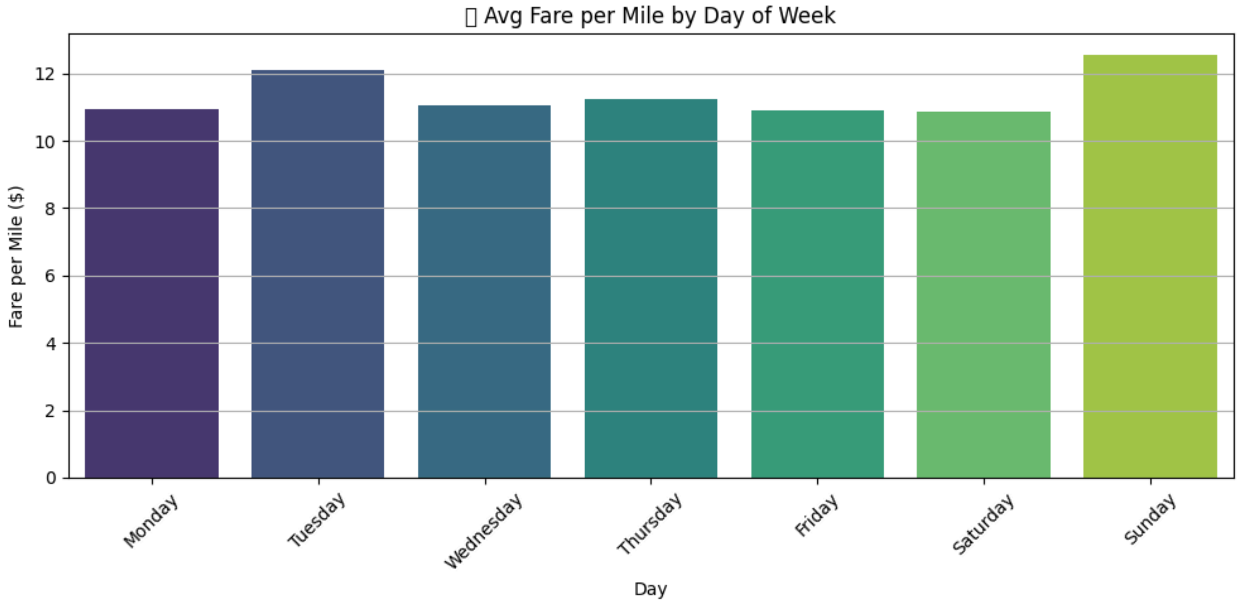

Estimated Hourly Trip Trends: Weekday vs Weekend
Weekday
Weekend
Estimated Number of Trips
160000
140000
120000
100000
80000
60000
40000
20000
0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day

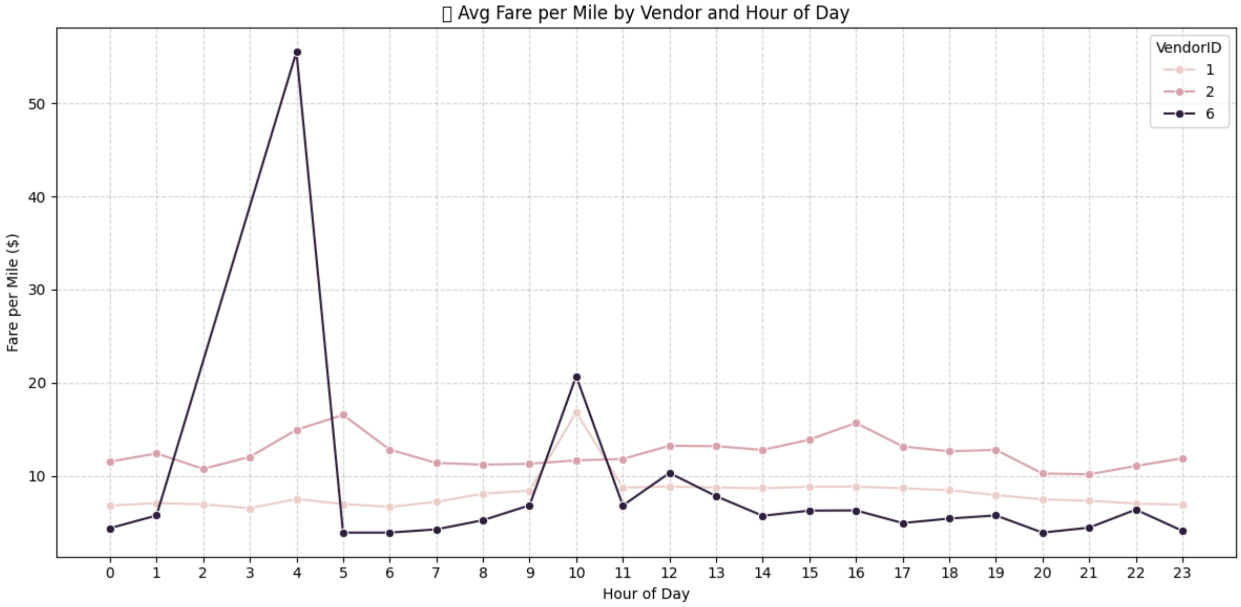


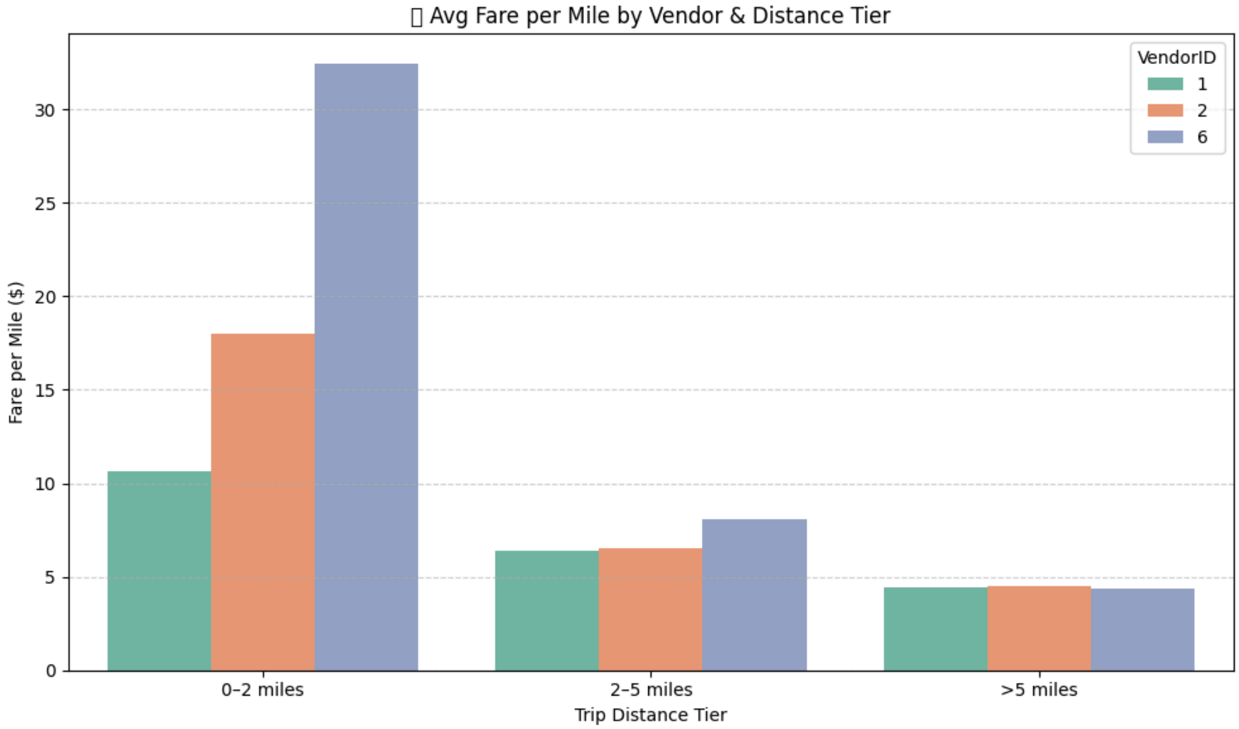
□ Avg. Fare per Mile per Passenger by Passenger Count

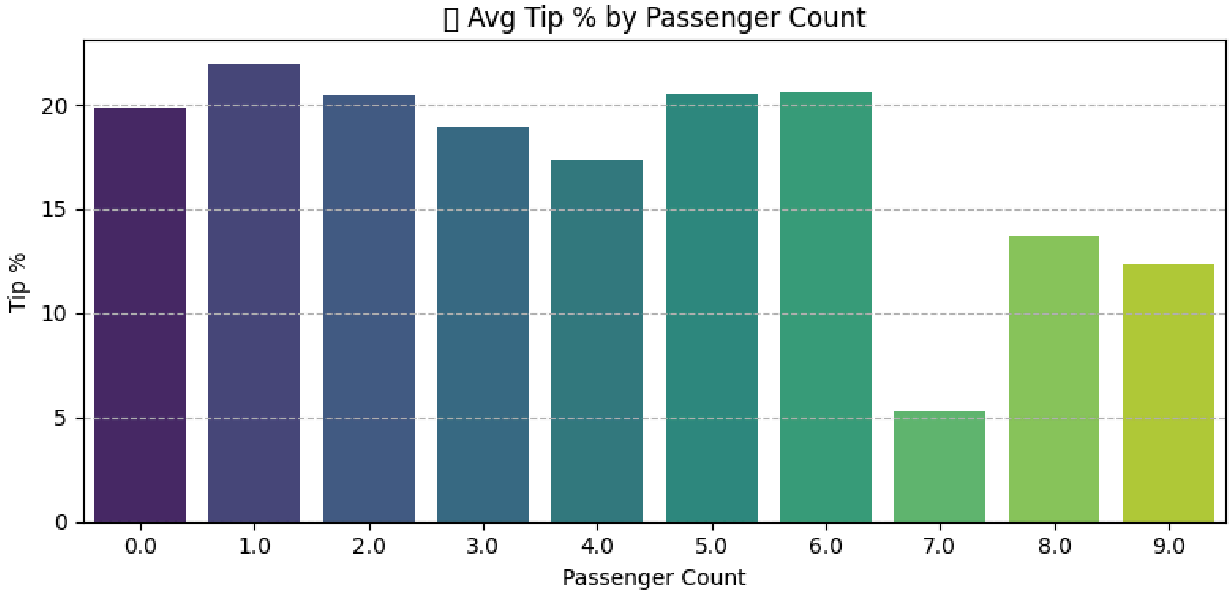


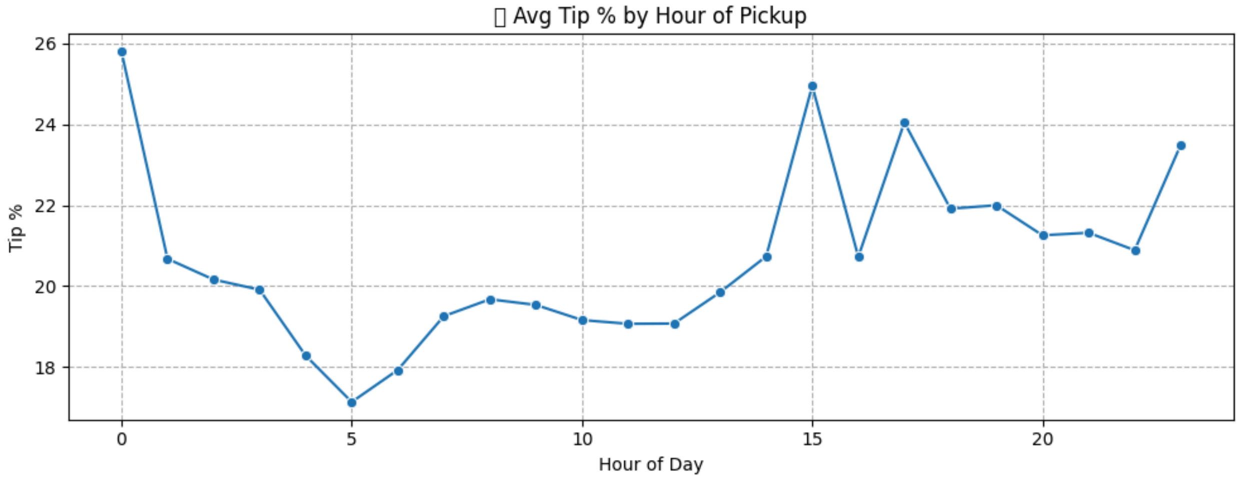
□ Avg Fare per Mile by Hour of Day


□ Avg Fare per Mile by Day of Week
Fare per Mile ($)
0
2
4
6
8
10
12
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day

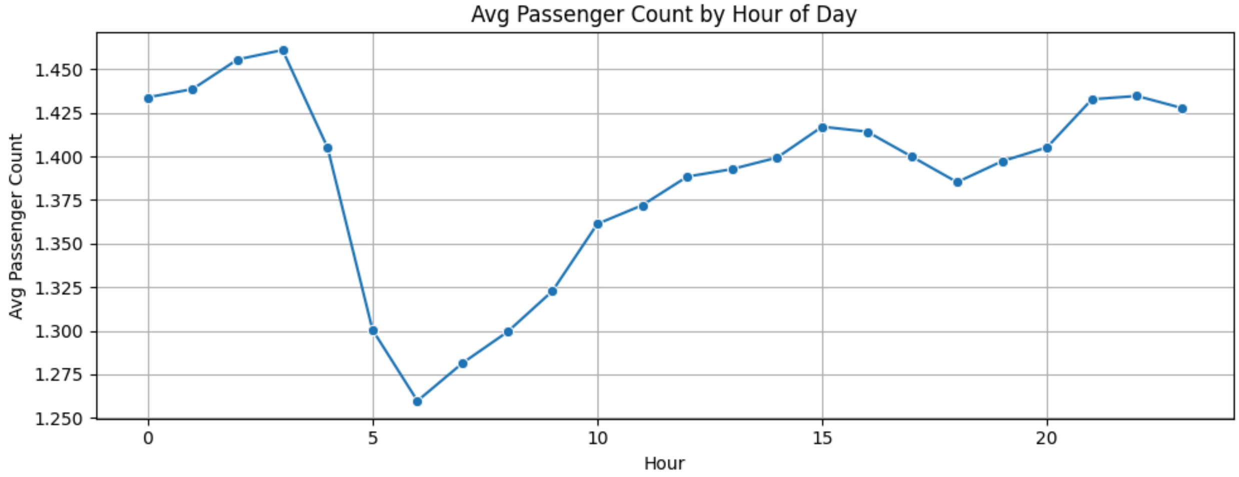

□ Avg Fare per Mile by Vendor and Hour of Day
Fare per Mile ($)
10
20
30
40
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day
VendorID
1
2
6


□ Avg Fare per Mile by Vendor & Distance Tier
VendorID
1
2
6
Fare per Mile ($)
0
5
10
15
20
25
30
0–2 miles
2–5 miles
>5 miles
Trip Distance Tier

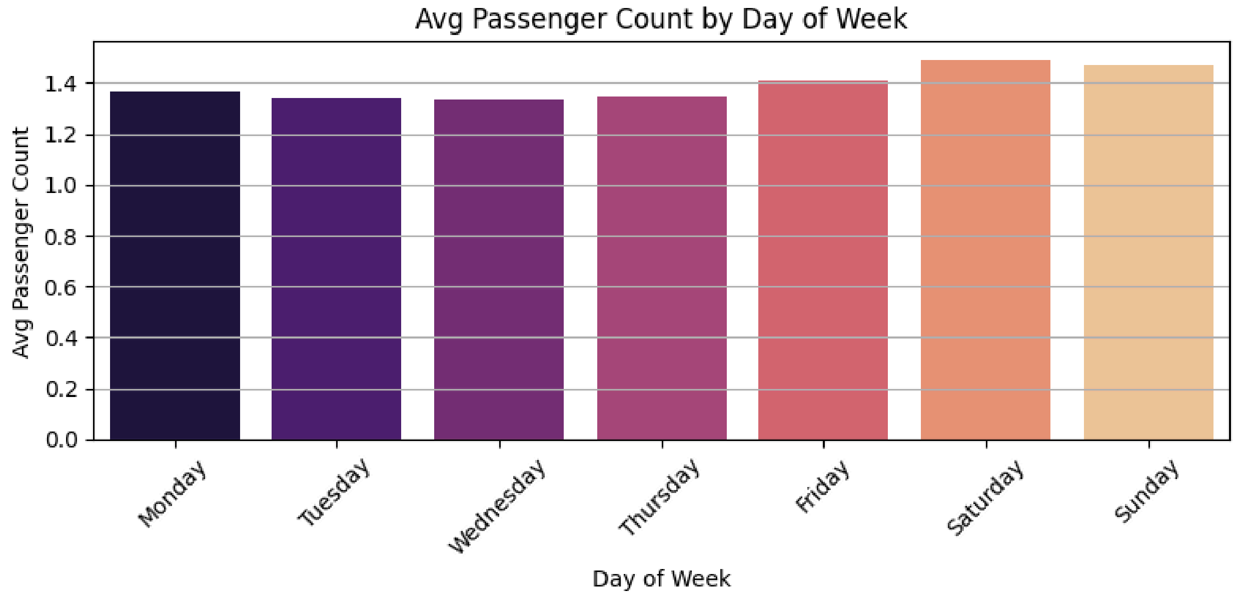

□ Avg Tip % by Passenger Count
Tip %
0
5
10
15
20
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
Passenger Count



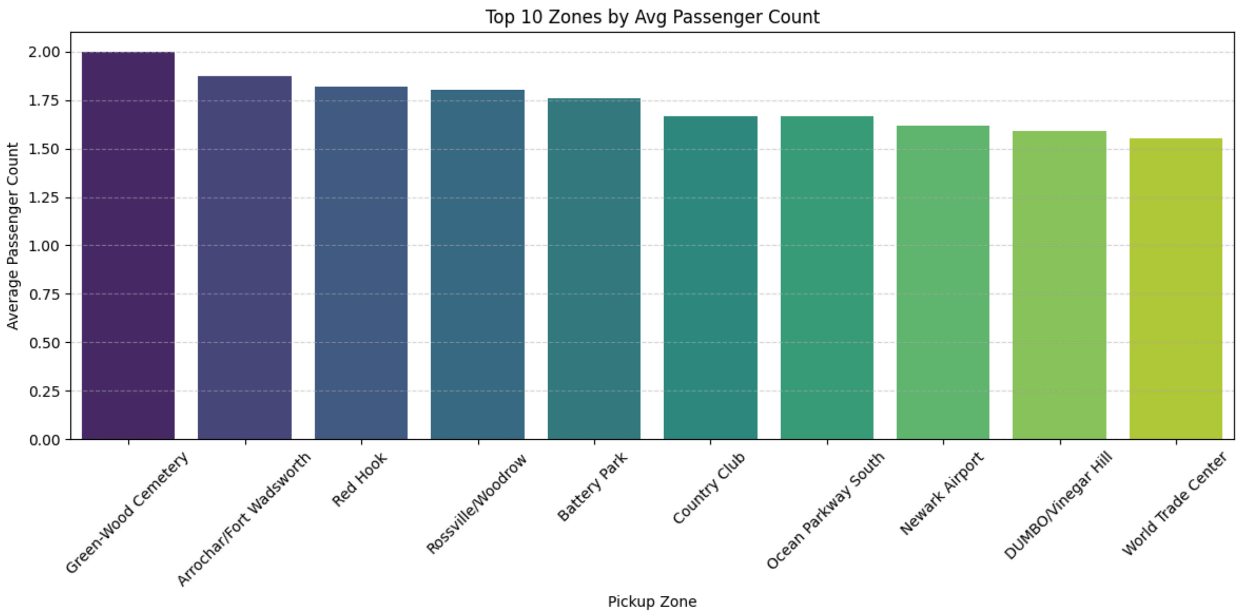
□ Avg Tip % by Hour of Pickup

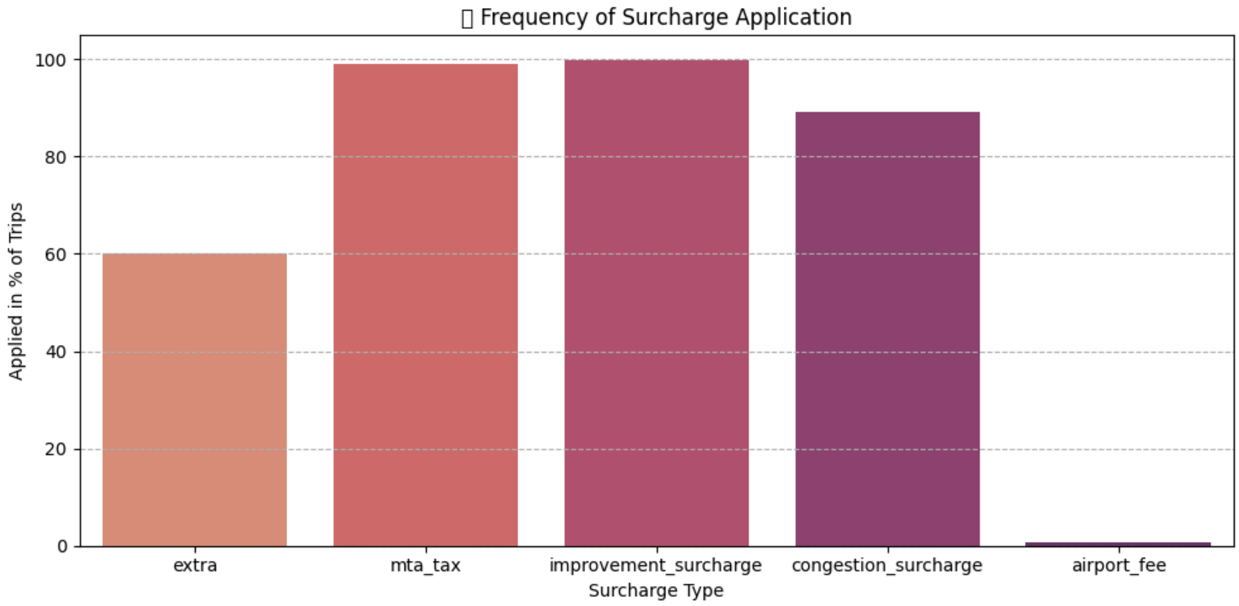


Avg Passenger Count by Hour of Day



Avg Passenger Count by Day of Week





5. Conclusion:

- This EDA reveals clear trends in passenger behavior, fare distribution, and operational inefficiencies. Data-backed strategies for pricing, dispatching, and routing can significantly enhance both rider experience and revenue for vendors.
- Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies:
    - Prioritize fleet availability at airport and transit zones - Pickup/dropoff ratios show high outbound demand from JFK, LGA,

and Penn Station, indicating a need to reposition idle cabs into these zones proactively.

    - Avoid fleet idle time in low-return zones - Residential areas like Freshkills Park and Whitestone see mostly drop-offs with almost no pickups, leading to inefficient returns.
    - Match driver count with commuter traffic during weekday mornings - Solo passenger rides peak between 8–10 AM, indicating commuter patterns where quick solo pickups dominate.
    - Strengthen operations over weekends, especially Friday and Saturday nights - These days show maximum trip volume and revenue, particularly in entertainment-focused neighborhoods.
    - Encourage longer trips to drive up tip percentage - Tip percentage rises notably for trips longer than 5 miles, likely due to higher engagement or perceived value.
    - Avoid congestion zones during low demand to reduce customer cost - Congestion surcharges are applied heavily in Manhattan, especially during work hours.
    - Plan route dispatches to avoid repeated surcharge areas when possible - Airport and congestion surcharges occur predictably; optimizing paths around them can reduce ride abandonment.
    - Use pickup/dropoff ratio to reposition idle vehicles - Zones like East Elmhurst have 8x more pickups than drop-offs, suggesting a need to route idle taxis back here.
    - Assign driver shifts based on peak demand patterns - Time-based trends clearly show when and where customer demand surges, enabling smarter shift design.

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

  1. Airports (JFK, LGA) — Always Keep Cabs Nearby

  These zones have many more pickups than drop-offs.

  Keep some cabs always available near JFK and LaGuardia, even during mid-day or late-night hours.

  2. Nightlife Zones (East Village, West Village, Midtown) — High Demand in Evenings

  Peak demand and higher tips between 6 PM to 2 AM, especially on Friday to Sunday.

  Position cabs near bars, clubs, and popular food areas during evenings.

3. Work Hubs (Penn Station, Midtown East) — Morning Commute Focus

Weekday mornings (7 AM to 10 AM) have many solo rides.

Use more solo cabs here during these hours to meet fast-moving commuter demand.

Weekend Nights (Times Square, LES, Meatpacking) — Boost Coverage

Friday and Saturday nights see many short trips with surcharges.

Keep extra cabs nearby (not directly in hotspots to avoid congestion) and rotate them in when others finish trips.

5. High Fare/Tip Zones — Priority Areas

Places like East Chelsea and Murray Hill give better fare per mile and higher tip percentages.

Send idle cabs to these areas during off-peak hours.

6. Busy Seasons (April–June, Oct–Dec) — Increase Fleet

Q2 and Q4 show higher trip revenue.

Add more cabs especially near hotels, parks, and tourist areas during these months.

7. Use Pickup/Dropoff Ratios

Some zones like East Elmhurst have 8x more pickups than drop-offs.

Monitor and return cabs back into such zones using trip ratio tracking.

8. Avoid Congestion Zones Midday (if demand is low)

Midtown has high congestion surcharges but not always higher fares.

Between 11 AM to 2 PM, shift cabs to SoHo, UES, or Brooklyn zones.