Report: Optimising NYC Taxi Operations

Include your visualisations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

1. Data Preparation

1.1. Loading the dataset

1.1.1. Sample the data and combine the files

Only 2023-year data is taken for this analysis. There were 12 parquet files for each month of 2023, and since the data was large, have taken 1% of data from each parquet file for analysis.

1.1.2. Ideally, keeping the total entries to around 300,000 to 400,000.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

- Have reset the index
- Dropped column "store_and_fwd_flag" as it mostly contains 'N' values, and it may not help with our analysis.

2.1.2. Combine the two airport fee columns

There were 2 airport fee columns, followed the below steps to combine them SAGNIK SAHA

• Filled the missing values in both the columns with median value of that column to avoid any data loss.

• Combine the two airport fee columns into single column ‘airport fee’.

• Dropped the duplicate Airport fee column.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

2.2.2. Handling missing values in passenger_count

- Imputed the NaN values with the median value in “passenger_count” column.
- Also found records with “passenger_count” a 0 and have imputed them with median value.

2.2.3. Handle missing values in RatecodeID

- Imputed NaN values in 'RatecodeID' with median value.

2.2.4. Impute NaN in congestion_surcharge

- Imputed NaN values in congestion_surcharge with median value

2.3. Handling Outliers and Standardising Values

2.3.1. Check outliers in payment type, trip distance and tip amount columns

- There are less trips with passenger count > 6, above once are mostly outliers.
- There are some records with RatecodeID 99, which is not standard value. Dropped them.
- There are some records with payment type 0, which is not standard value. Dropped them. SAGNIK SAHA
- There are some outliers in fare_amount, also there are trips where trip distance is < 1 and fare amount is > 300. Dropped them
- tip_amount looks, however, there seems to be 1/2 outliers, which upon validation of those records looks good.
- Very few records with trip_distance > 250, so dropped them.

3. Exploratory Data Analysis

3.1 General EDA: Finding Patterns and Trends

3.1.1 Classify variables into categorical and numerical

- VendorID: Categorical
- tpep_pickup_datetime: Numerical
- tpep_dropoff_datetime: Numerical
- passenger_count: Categorical
- trip_distance: Numerical

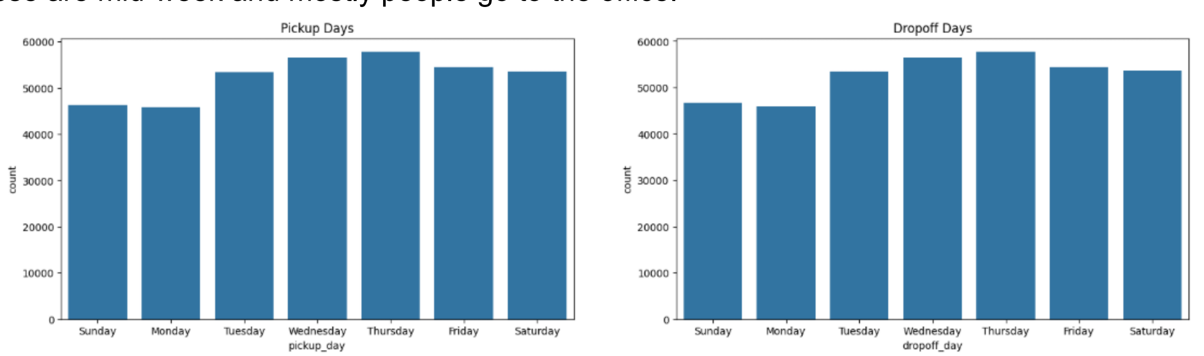
- RatecodeID: Categorical
- PULocationID: Numerical
- DOLocationID: Numerical
- payment_type: Categorical
- pickup_hour: Numerical
- trip_duration: Numerical

Below columns belong to numerical category

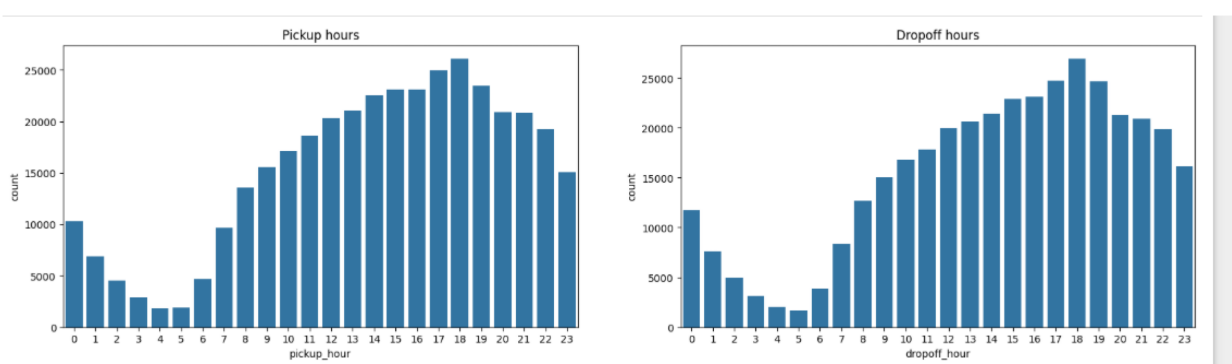
- fare_amount
- extra
- mta_tax
- tip_amount
- tolls_amount
- improvement_surcharge
- total_amount
- congestion_surcharge
- airport_fee

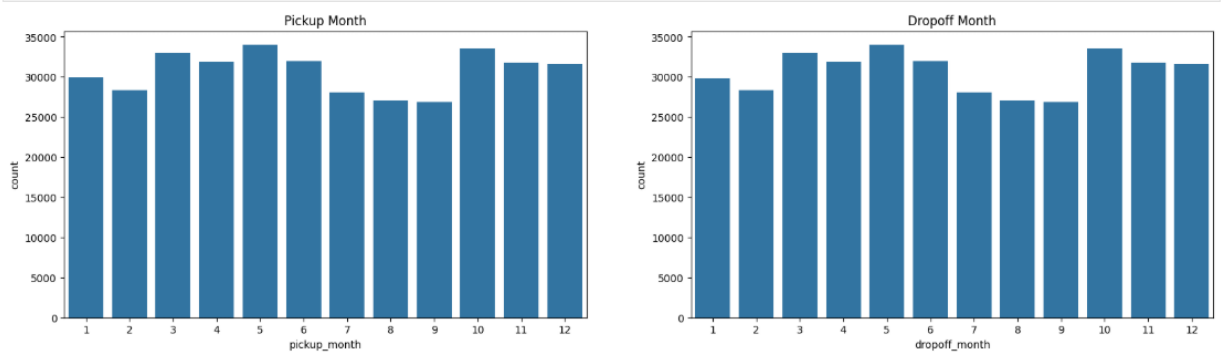
3.1.2 Analyse the distribution of taxi pickups by hours, days of the week, and months

- From the below plot, the busiest hours are 5:00 pm to 7:00 pm and that makes sense as this is the time when people return from their offices.
- From the plot, the busiest days are Wednesday and Thursdays, and that makes sense as these are mid-week and mostly people go to the office.



- From this plot, it's shows high taxi activity during may and oct months





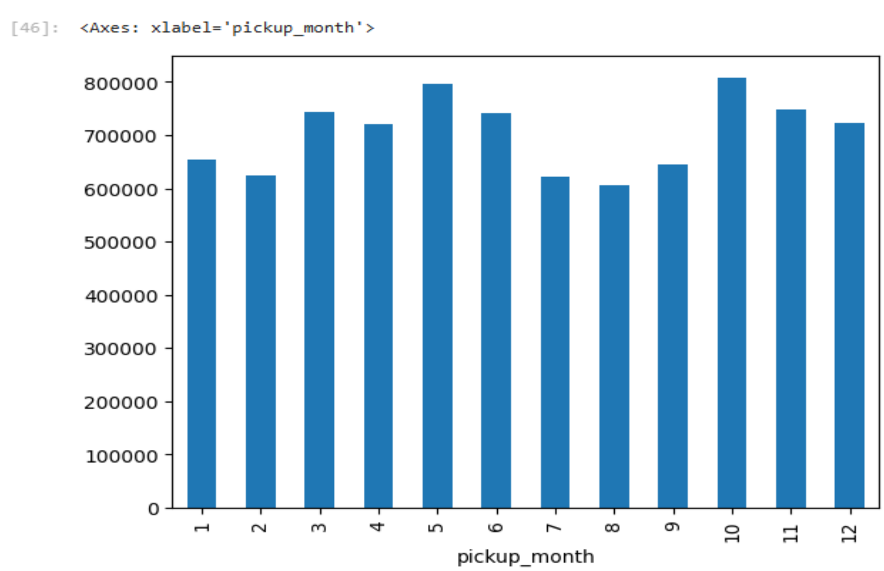
3.1.3 Filter out the zero/negative values in fares, distance and tips

- Removed the records with 0 value for fare amount, total amount, tip amount & trip distance, which may affect our visualization analysis.

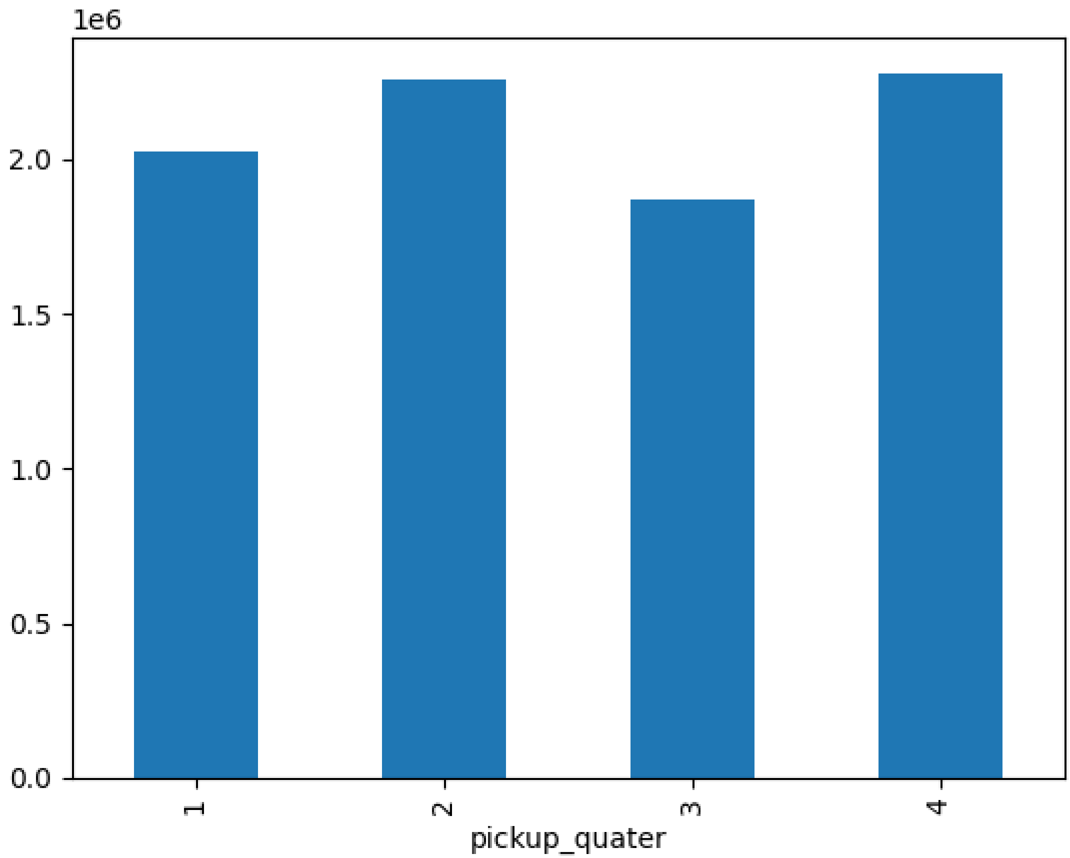
3.1.4 Analyse the monthly revenue trends

- The plot shows that the revenue is high mostly in may and oct months.

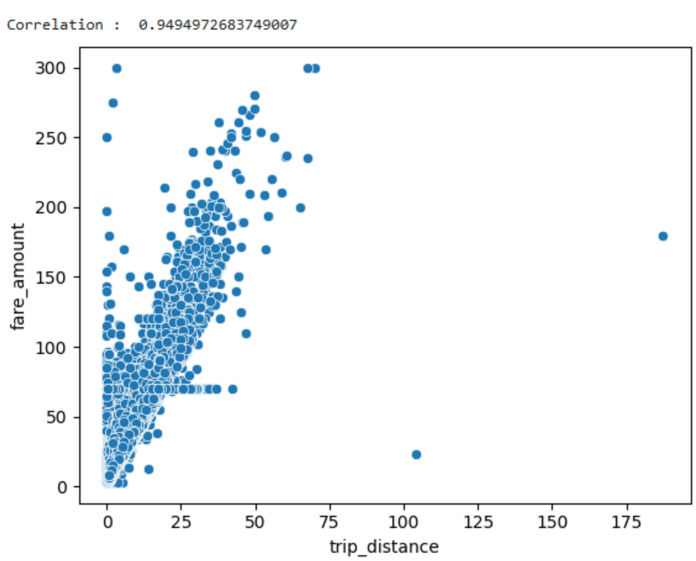
3.1.5 Find the proportion of each quarter's revenue in the yearly revenue



- This plot shows the revenue generated in the 2nd & 4th quarter is higher than 1st & 3rd quarter.

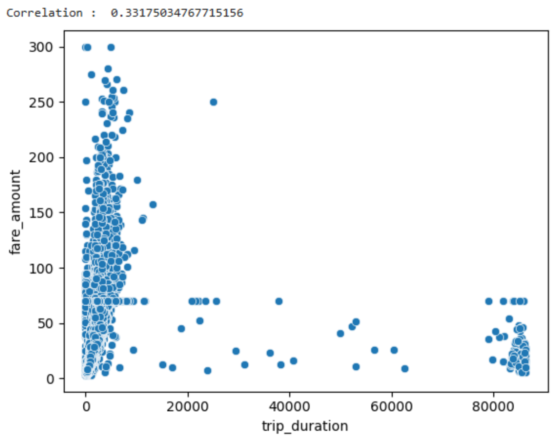


3.1.6 Analyse and visualise the relationship between distance and fare amount



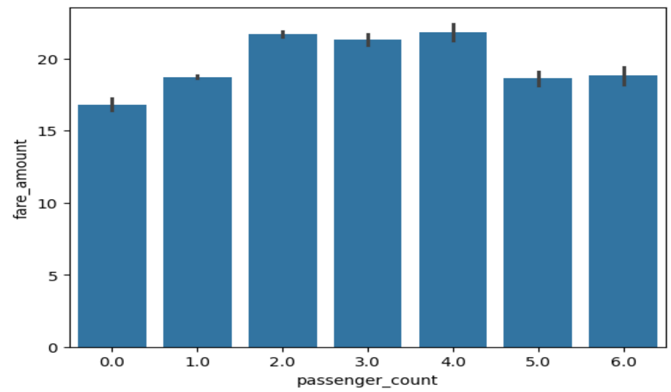
- There is a high positive correlation between trip distance & fare amount.

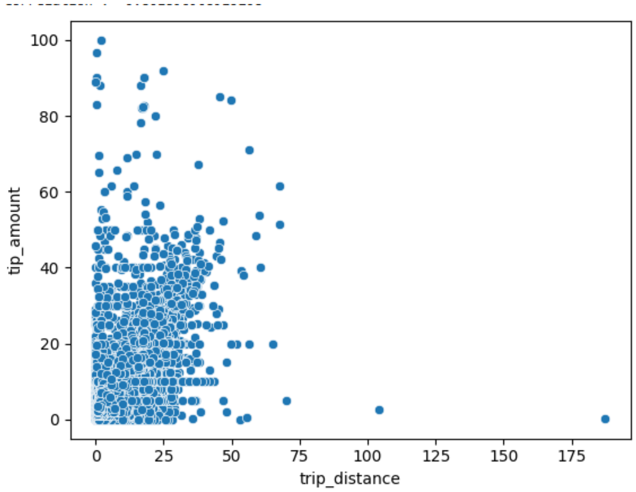
3.1.7 Analyse the relationship between fare/tips and trips/passengers



- There is a weak positive correlation between trip duration & fare amount.

- This shows that the Fare amount is higher for a trip with >1 passengers, passenger count - 4 being the top in the list.

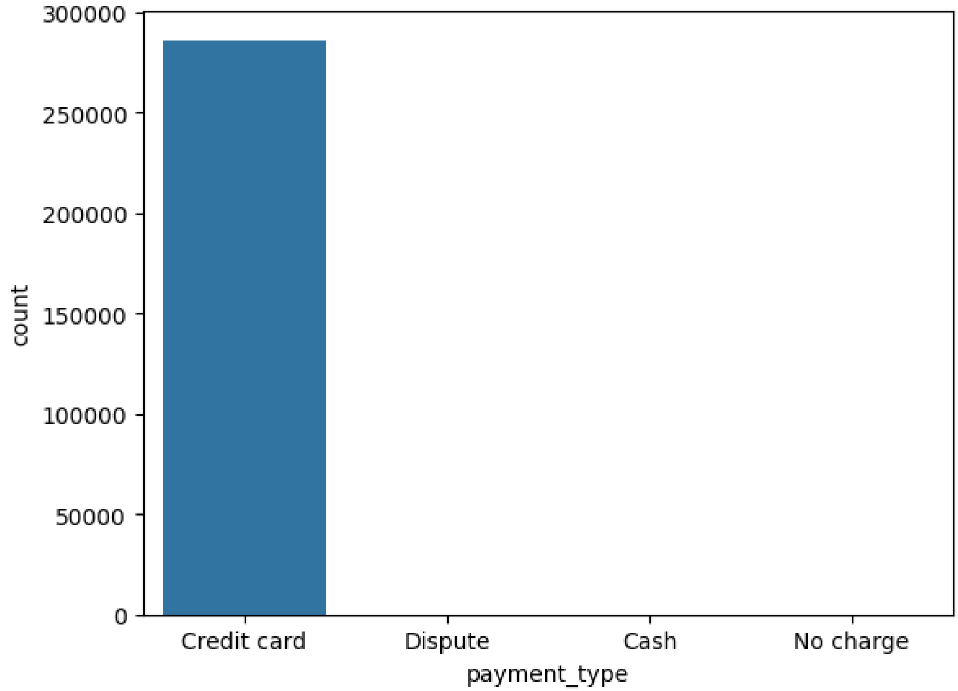




- There is a high positive correlation value between trip distance and tip amount.

3.1.8 Analyse the distribution of different payment types

- This plot shows that payment type 1 (Credit card) is the most common payment type.



3.1.9 Load the taxi zones shapefile and display it

```
import geopandas as gpd

# Read the shapefile using geopandas
zones = gpd.read_file('taxi_zones.shp')
zones.head()
```

✓ 0.0s

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...

3.1.10 Merge the zone data with trips data

taxi_sample_df_nonzero = taxi_sample_df_nonzero.merge(zones, left_on='PULocationID', right_on='LocationID', how='left')

3.1.11 Find the number of trips for each zone/location ID

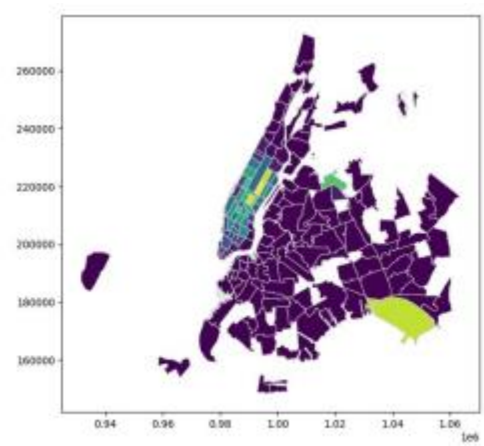
- Top 10 zones with highest number of trips

	zone	PULocationID	Number_of_trips
149	Upper East Side South	237	14076
97	Midtown Center	161	13421
148	Upper East Side North	236	12788
72	JFK Airport	132	12780
98	Midtown East	162	10660
80	LaGuardia Airport	138	10217
83	Lincoln Square East	142	9708
112	Penn Station/Madison Sq West	186	9677
143	Times Sq/Theatre District	230	8830
106	Murray Hill	170	8616

3.1.12 Add the number of trips for each zone to the zones dataframe

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	PULocationID	Number_of_trips
236	237	0.042213	0.000096	Upper East Side South	237	Manhattan	POLYGON ((993633.442 216961.016, 993507.232 21...	237.0	14076.0
160	161	0.035804	0.000072	Midtown Center	161	Manhattan	POLYGON ((991081.026 214453.698, 990952.644 21...	161.0	13421.0
235	236	0.044252	0.000103	Upper East Side North	236	Manhattan	POLYGON ((995940.048 221122.92, 995812.322 220...	236.0	12788.0
131	132	0.245479	0.002038	JFK Airport	132	Queens	MULTIPOLYGON (((1032791.001 181085.006, 103283...	132.0	12780.0
161	162	0.035270	0.000048	Midtown East	162	Manhattan	POLYGON ((992224.354 214415.293, 992096.999 21...	162.0	10660.0

3.1.13 Plot a map of the zones showing number of trips



OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	PULocationID	Number_of_trips
237	0.042213	0.000096	Upper East Side South	237	Manhattan	POLYGON ((993633.442 216961.016, 993507.232 21...	237.0	14076.0
161	0.035804	0.000072	Midtown Center	161	Manhattan	POLYGON ((991081.026 214453.698, 990952.644 21...	161.0	13421.0
236	0.044252	0.000103	Upper East Side North	236	Manhattan	POLYGON ((995940.048 221122.92, 995812.322 220...	236.0	12788.0
132	0.245479	0.002038	JFK Airport	132	Queens	MULTIPOLYGON (((1032791.001 181085.006, 103283...	132.0	12780.0
162	0.035270	0.000048	Midtown East	162	Manhattan	POLYGON ((992224.354 214415.293, 992096.999 21...	162.0	10660.0
138	0.107467	0.000537	LaGuardia Airport	138	Queens	MULTIPOLYGON (((1019904.219 225677.983, 102031...	138.0	10217.0
142	0.038176	0.000076	Lincoln Square East	142	Manhattan	POLYGON ((989380.305 218980.247, 989359.803 21...	142.0	9708.0
186	0.024696	0.000037	Penn Station/Madison Sq West	186	Manhattan	POLYGON ((986752.603 210853.699, 986627.863 21...	186.0	9677.0
230	0.031028	0.000056	Times Sq/Theatre District	230	Manhattan	POLYGON ((988786.877 214532.094, 988650.277 21...	230.0	8830.0
170	0.045769	0.000074	Murray Hill	170	Manhattan	POLYGON ((991999.299 210994.739, 991972.635 21...	170.0	8616.0

.

3.1.14.1 Conclude with results

Summary of Findings from Temporal Analysis:

Taxi Activity Trends:

- Peak pickup/drop-off hours are 5:00–7:00 PM, likely due to office commuters.
- Weekdays, especially Wednesday and Thursday, show higher activity than weekends.
- Taxi usage is highest during summer (May–June) and Q4 (Oct–Dec), likely due to holidays and festivals.

Revenue Trends:

- Revenue peaks in Q2 and Q4, aligning with periods of high taxi activity, with Q4 being the highest due to the festive season.

Financial Insights:

- Fare vs. Distance: Strong positive correlation—longer trips yield higher fares.
- Fare vs. Duration: Positive but weaker correlation than distance.
- Fare vs. Passenger Count: Higher fares for trips with more than one passenger, especially with 4 passengers.
- Tip vs. Distance: Positive correlation—longer trips tend to receive higher tips.

Busiest Pickup Zones:

- Top locations include Upper East Side South, Midtown Center, Upper East Side North, JFK Airport, and Midtown East.

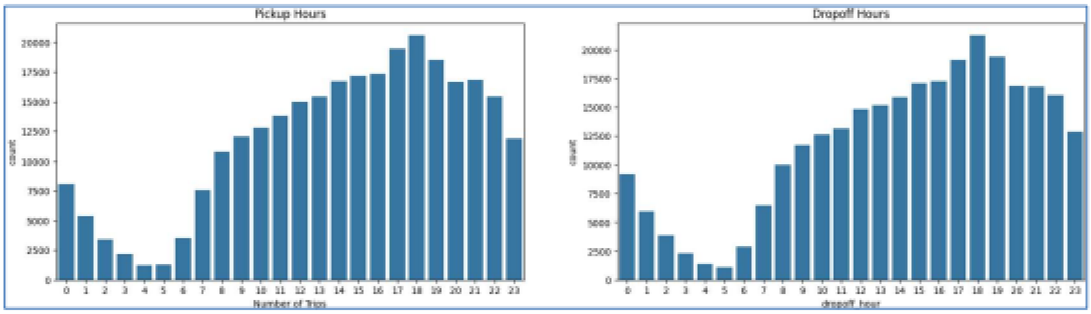
3.2 Detailed EDA: Insights and Strategies

3.2.14 **Identify slow routes by comparing average speeds on different routes**

	pickup_zone	PULocationID	dropoff_zone	DOLocationID	pickup_hour	speed
0	Seaport	209	Two Bridges/Seward Park	232	13	0.043579
1	East Elmhurst	70	LaGuardia Airport	138	6	0.085750
2	Seaport	209	Boerum Hill	25	22	0.106057
3	Midtown Center	161	Upper West Side North	238	7	0.117807
4	Midtown North	163	Financial District North	87	15	0.140078
5	Williamsburg (North Side)	255	Williamsburg (South Side)	256	2	0.141176
6	Queensbridge/Ravenswood	193	Queensbridge/Ravenswood	193	11	0.150000
7	Greenwich Village North	113	Park Slope	181	19	0.153191
8	Sutton Place/Turtle Bay North	229	Central Harlem	41	17	0.174780
9	Upper West Side North	238	West Village	249	1	0.196820

3.2.14.1 **Calculate the hourly number of trips and identify the busy hours**

- From the plot, the busiest hours are 5:00 pm to 7:00 pm and that makes sense as this is the time when people return from their offices.



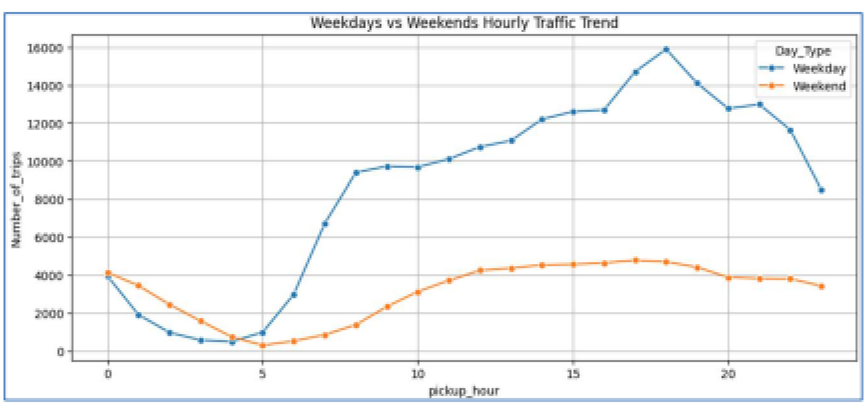
3.2.15 **Scale up the number of trips from above to find the actual number of trips**

- I have taken the sample frac as 0.01
- Below are the actual number of trips for the top 5 busiest hours

pickup_hour	Number of trips
18	2059300.0
17	1948000.0
19	1850700.0
16	1732600.0
15	1716400.0

3.2.16 **Compare hourly traffic on weekdays and weekends**

- From the plot, the busiest hours are 5:00 pm to 7:00 pm on weekdays and that makes sense as this is the time when people return from their offices.
- And on weekends, the trips are more during the late night hours due to the holiday.



3.2.17 **Identify the top 10 zones with high hourly pickups and drops**

3.2.18 **Find the ratio of pickups and dropoffs in each zone**

Top 10 Pickup Zones:

	pickup_zone	pickup_hour	Number_of_trips

Top 10 Highest Pickup/Dropoff Ratios:

	zone	pickup_trip_counts	dropoff_trip_counts	pickup_dropoff_ratio
63	East Elmhurst	1284.0	92	13.956522
116	JFK Airport	12793.0	2668	4.794978
125	LaGuardia Airport	10224.0	3519	2.905371
201	South Jamaica	27.0	15	1.800000
174	Penn Station/Madison Sq West	9678.0	6001	1.612731
39	Central Park	4889.0	3460	1.413006
235	West Village	6894.0	5062	1.361912
101	Greenwich Village South	3855.0	2881	1.338077
149	Midtown East	10660.0	8250	1.292121
91	Garment District	4261.0	3503	1.216386

Top 10 Lowest Pickup/Dropoff Ratios:

	zone	pickup_trip_counts	dropoff_trip_counts	pickup_dropoff_ratio
11	Bay Ridge	1.0	152	0.006579
160	Newark Airport	6.0	742	0.008086
211	Stuyvesant Heights	2.0	206	0.009709
206	Spuyten Duyvil/Kingsbridge	1.0	91	0.010989
54	Crown Heights North	5.0	387	0.012920
229	Washington Heights North	6.0	461	0.013015
185	Ridgewood	2.0	119	0.016807
84	Flushing	2.0	105	0.019048
14	Bedford	5.0	255	0.019608
183	Rego Park	1.0	51	0.019608

3.2.19 **Identify the top zones with high traffic during night hours**

'Top 10 pickup zones during night hours'

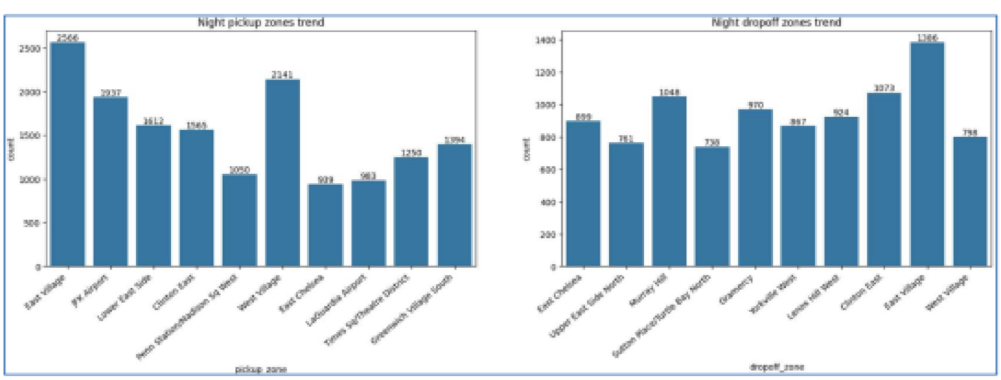
	pickup_zone	Trip Count
0	East Village	2566
1	West Village	2141
2	JFK Airport	1937
3	Lower East Side	1612
4	Clinton East	1565
5	Greenwich Village South	1394
6	Times Sq/Theatre District	1250
7	Penn Station/Madison Sq West	1050
8	LaGuardia Airport	983
9	East Chelsea	939

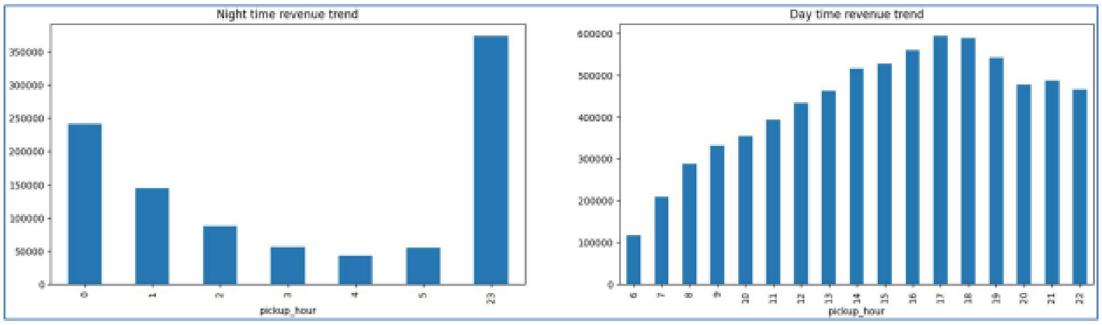
'Top 10 dropoff zones during night hours'

	dropoff_zone	Trip Count
0	East Village	1386
1	Clinton East	1073
2	Murray Hill	1048
3	Gramercy	970
4	Lenox Hill West	924
5	East Chelsea	899
6	Yorkville West	867
7	West Village	798
8	Upper East Side North	761
9	Sutton Place/Turtle Bay North	738

3.2.20 **Find the revenue share for nighttime and daytime hours**

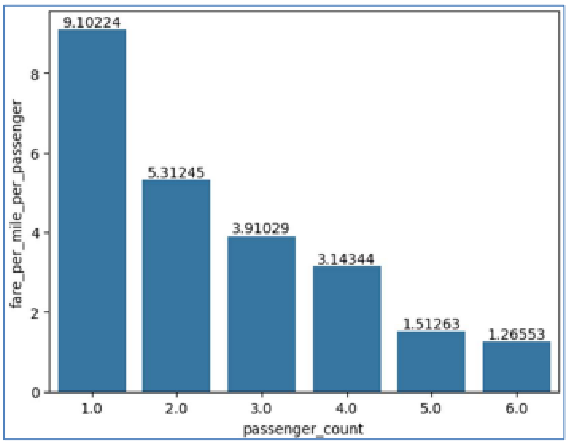
- Nighttime Revenue Share: 11.98%
- Daytime Revenue Share: 88.02%
- Day time revenue share is more than nighttime, because taxi activity is more in daytime.



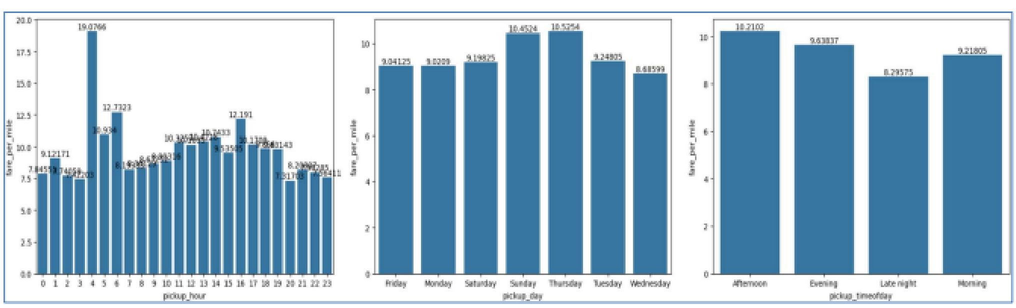


3.2.21 **For the different passenger counts, find the average fare per mile per passenger**

- This plot shows that the fare per mile per passenger is higher for trips with passenger count – 1
- There is a downward trend in avg fare per mile >1 passenger count.



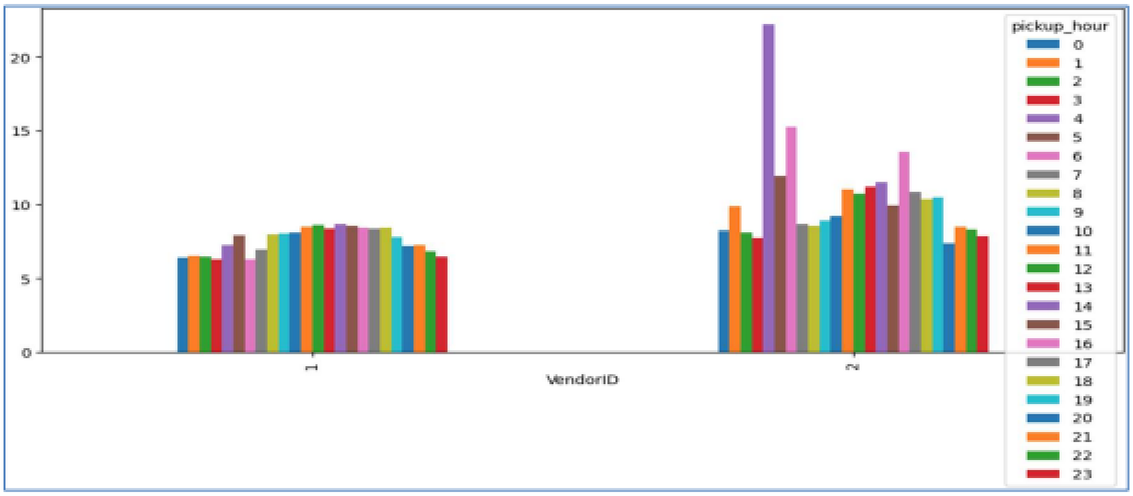
3.2.22 **Find the average fare per mile by hours of the day and by days of the week**



- Avg fare per mile is mostly high during the early morning (4am-6am and evening (4pm – 6pm).
- On weekdays, avg fare per mile is high on Thursday, aligning with the higher taxi activity on weekdays.
- On weekends, avg fare per mile is high on Sunday, due to high taxi activity during late night on weekends.
- Avg fare per mile higher for Afternoon time, followed by Evening, Morning and late night.

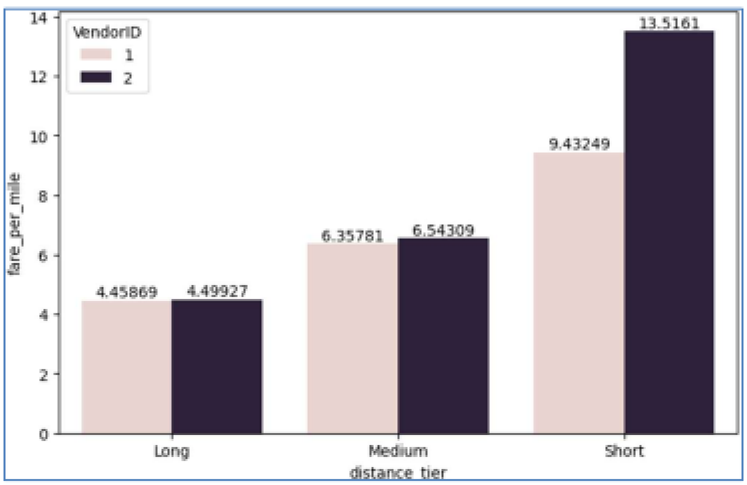
3.2.23 **Analyse the average fare per mile for the different vendors**

- The far per mile is higher for vender 2 than vendor 1.



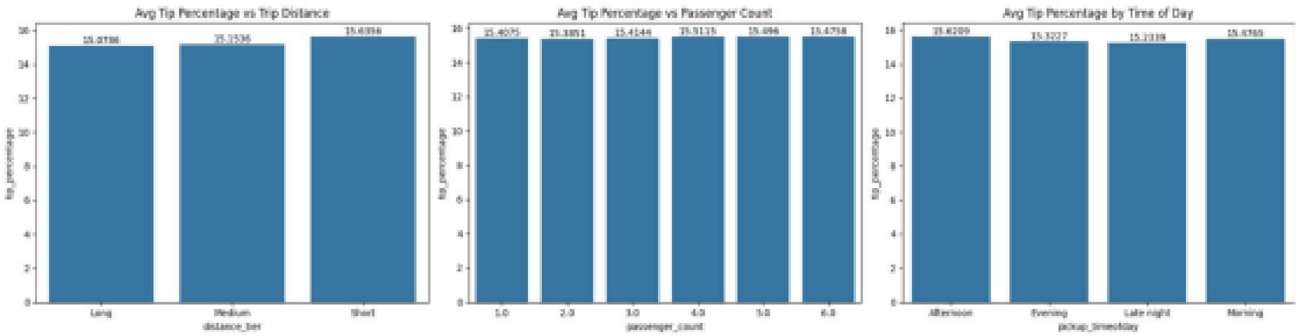
3.2.24 **Compare the fare rates of different vendors in a distance-tiered fashion**

- Similar observation as above, vendor 2 is higher fare per mile for all types of distance tier.



3.2.13 Analyse the tip percentages

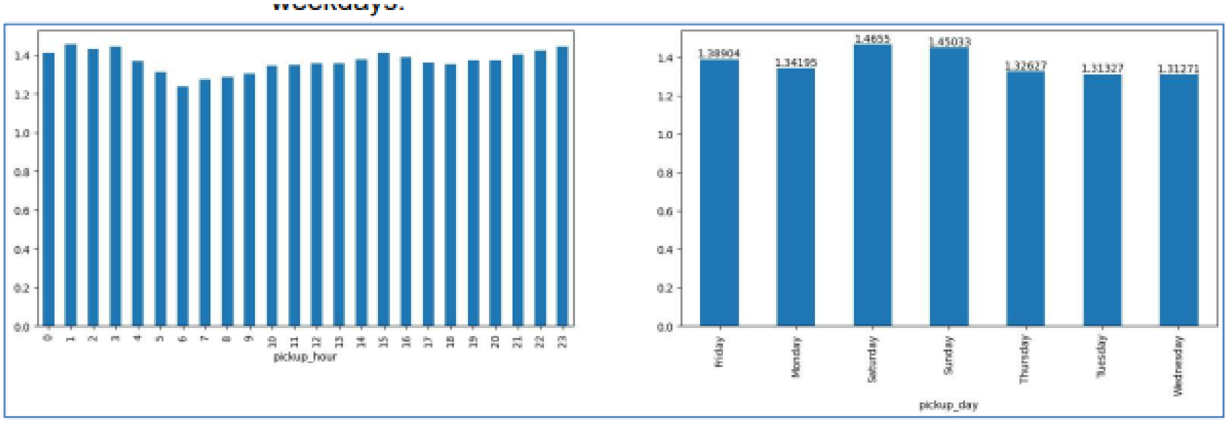
- Tip % is almost same across all passenger count and trip distance tier.
- Trips during certain times of the day, such as late night tend to have lower tip percentages.



•

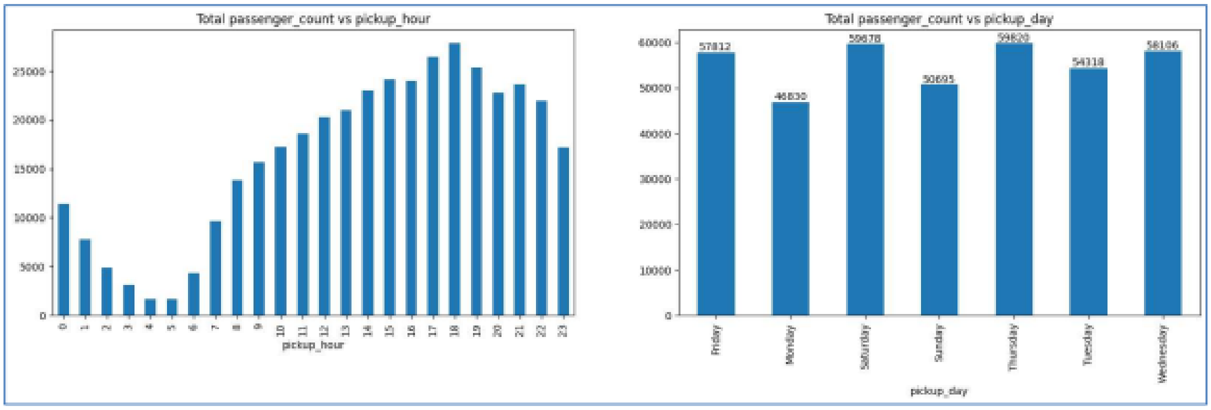
3.2.14. Analyse the trends in passenger count

- Avg passenger count is mostly high during the evening & late night.
- Avg passenger count is mostly high during the weekends compared to weekdays.



- Total passenger travelled on evening hours are more compared to late night & early morning, aligning with out previous finding where busiest hour is during the evening time (5pm – 7pm).
- The total passenger count on weekdays (Wednesdays & Thursdays) is higher compared to weekends. This also aligns with our previous findings, where taxi activity is more on weekdays

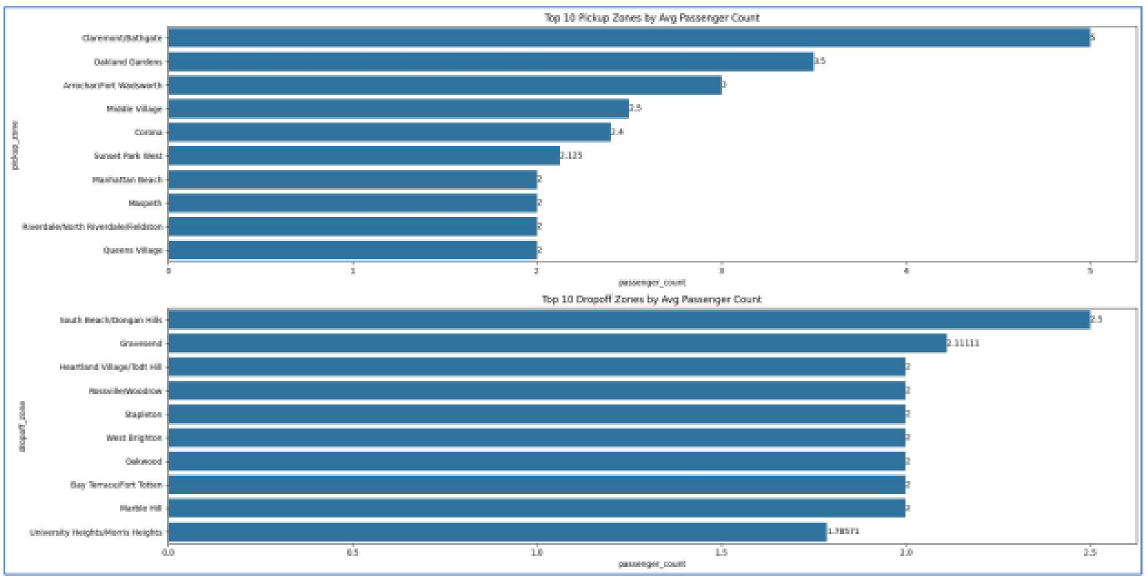
where taxi activity is more on weekdays.



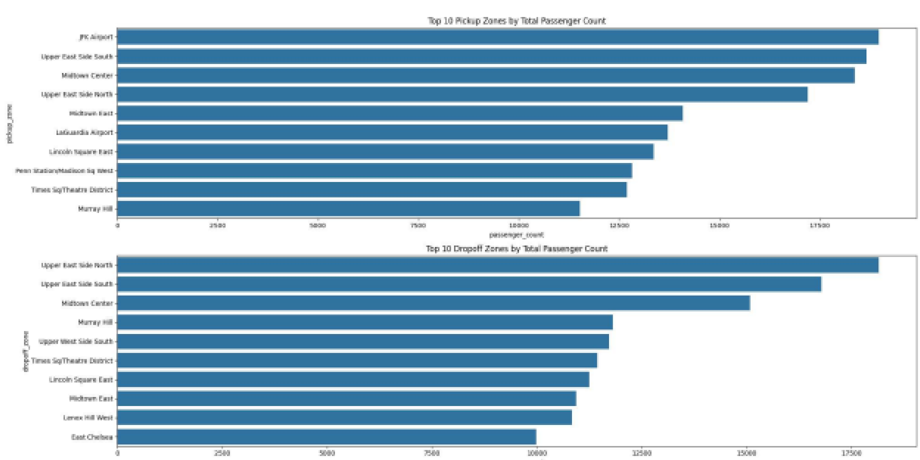
•

3.2.14 Analyse the variation of passenger counts across zones

- Top 10 zones where avg passenger count travelled.



- The top 10 zones with total passenger count travelled.

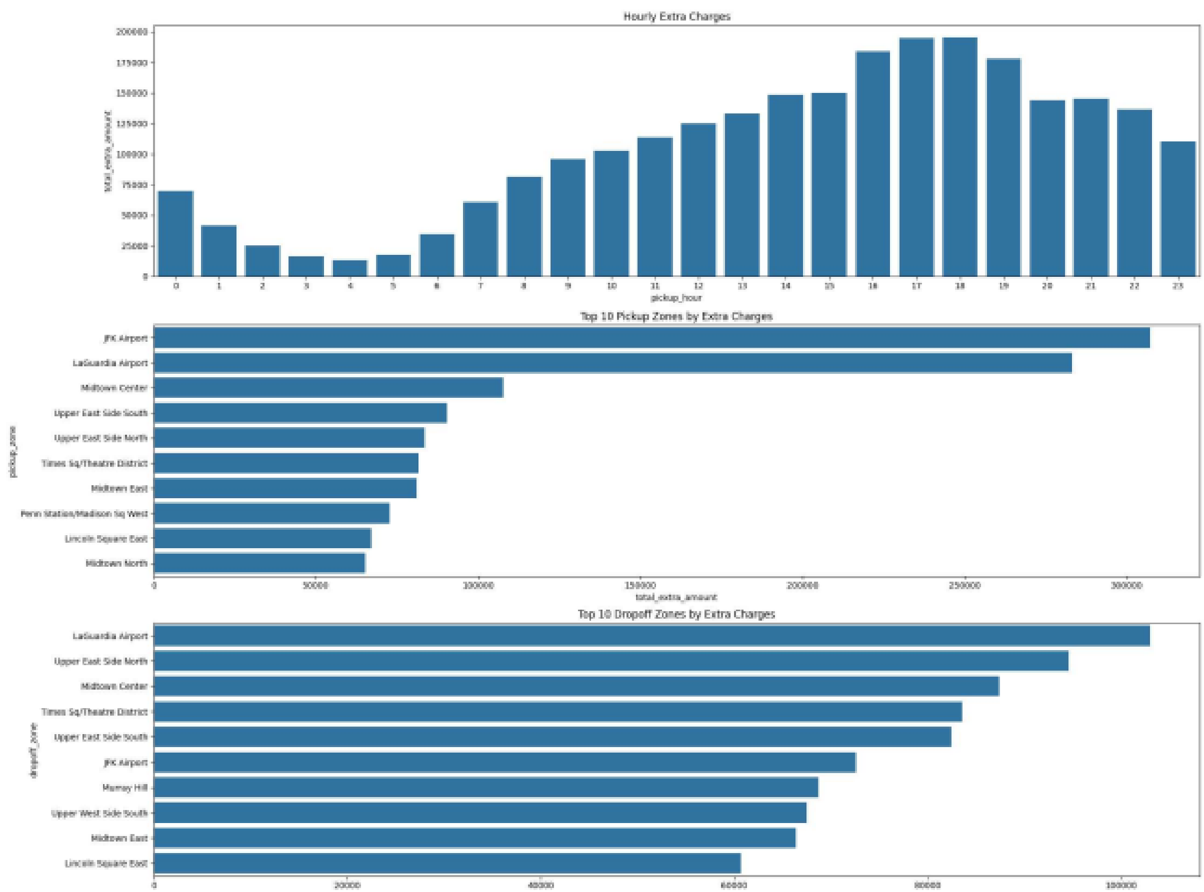


3.2.15 Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.

- Frequency of surcharge type applied across all trips

	Surcharge_Type	Frequency
2	tip_amount	282922
4	improvement_surcharge	282919
1	mta_tax	281393
5	congestion_surcharge	268650
0	extra	179521
6	airport_fee	23192
3	tolls_amount	22645

- Extra charges are applied mostly during the busiest evening hours (4pm – 7pm), due to high demand and office closing time.
- Top zones where extra charges are applied are JFK Airport, LaGuardia Airport, Midtown



center, Upper East Side Nort, Times Sqr.

Conclusions

3.3 Final Insights and Recommendations

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

Peak Hour Management

Weekday rush: 5–7 PM; weekend late-night: 11 PM–5 AM → allocate extra taxis.

Route Optimization

- Reroute around slow segments using average-speed and live-traffic data.
- Prioritize high-demand zones (e.g., Upper East Side South, Midtown Center, JFK).
- Scale back in low-demand areas off-peak.

Customer Experience

- Position cabs in busy zones during peaks to cut wait times.
- Offer off-peak/low-demand discounts to boost utilization.

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

High-Demand Zones

- Focus on Upper East Side South, Upper East Side North, Midtown Center, Midtown East, and JFK Airport—especially during peak hours—to cut wait times

Weekday Evenings

- Deploy extra cabs in business districts and residential areas from 5–7 PM.

Weekend Evenings & Late Nights

- Position cabs in shopping, entertainment, and tourist spots from 7 PM–5 AM.

Seasonal Surges

- Add dedicated “seasonal” cabs in summer (May–June) and holidays (Oct–Dec).

Partnerships

- Collaborate with hotels, businesses, and event organizers for on-demand fleets.

Ongoing Optimization

- Continuously monitor and forecast trip data to adjust positioning proactively.

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

- Implement dynamic pricing based on demand patterns. Increase fares during peak hours (5:00 PM - 7:00 PM) and late-night hours (11:00 PM - 5:00 AM) to maximize revenue.
- Offer discounts or promotions during off-peak hours to encourage more rides and improve utilization.
- Adjust fare rates for different distance tiers. For short trips (<= 2 miles), maintain competitive rates to attract more customers. For medium (2-5 miles) and long trips (> 5 miles), increase fare rates slightly to maximize revenue.
- Monitor competitor pricing and adjust rates to remain competitive while ensuring profitability.

- Use real-time data to dynamically adjust pricing based on current demand and supply conditions.
- Implement surge pricing during high-demand periods and in high-demand zones to manage demand and increase revenue.
- Offer loyalty programs or incentives for frequent riders to encourage repeat business and improve customer retention.