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

We Iterated through each day and hour and sampled 1 percent of the data, this sampled data was appended to a pandas dataframe, and we saved this as a parquet file

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

index was reset

2.1.2. Combine the two airport_fee columns

Airport_fee had fewer null values, so we created a new column unified_airport_fee in which the missing values of Airport_fee was replaced by airport_fee values. Finally both the airport_fee column was dropped.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

df.isnull().mean()
This revealed that there were no null values

2.2.2. Handling missing values in passenger_count

No missing values but there were entries with passenger_ count zero. We replaced zero with the mode passenger count which is 1.

2.2.3. Handle missing values in RatecodeID

No null value in RatecodeID except that some values were 99 indicating this values were missing. We then replaced 99 with 7 so that we can analyse it later.

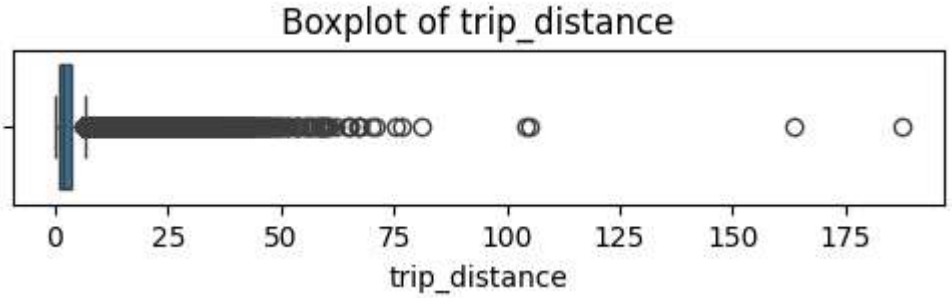
2.2.4. Impute NaN in congestion_surcharge

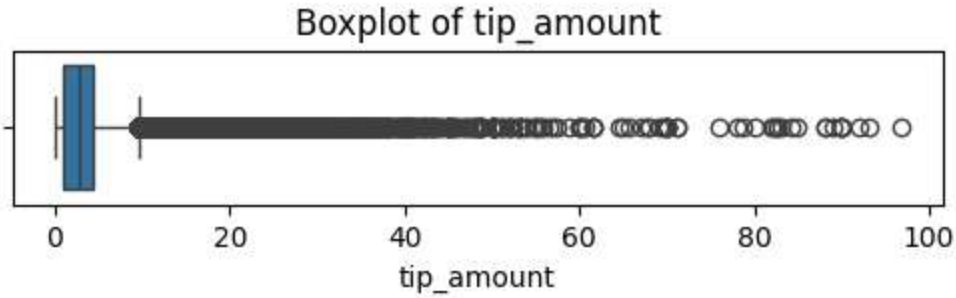
No NaN in congestion_surcharge

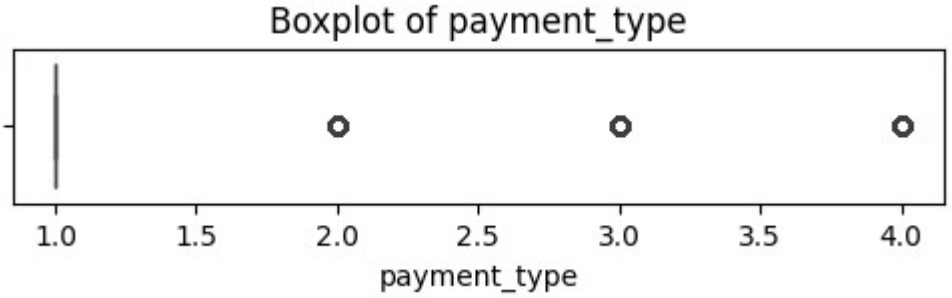
2.3. Handling Outliers and Standardising Values

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

trip distance had 48450 outliers while tip amount had 28220







The above plots were drawn after removing extreme outliers, Like trips having distance longer than 250 miles, And rides with near zero trip distance but having fare above 300.

3. Exploratory Data Analysis

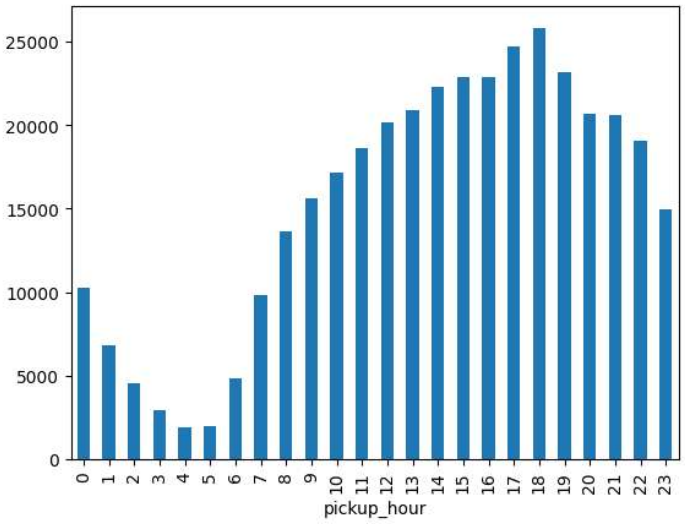
3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical

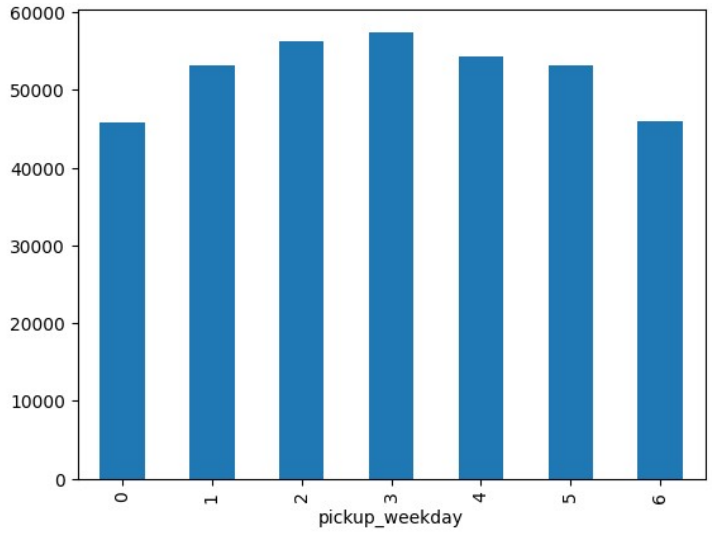
Categorical: VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, RatecodeID, PULocationID, DOLocationID, payment_type, pickup_hour
Numerical: passenger_count, trip_distance, trip_duration, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee

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

Pickups by hour: 4pm to 8pm are the busiest hours

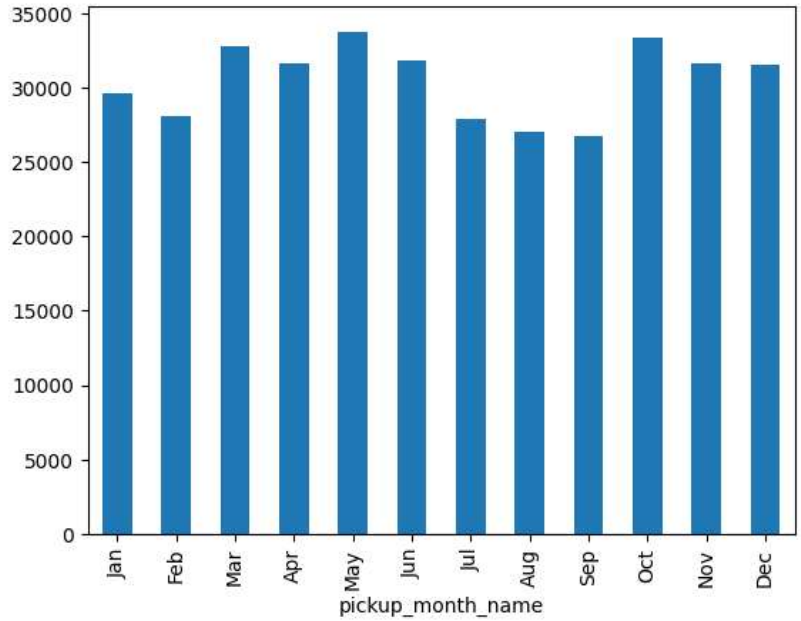


Pickups by days of the week:



We do not see much variations for day of week wise

Month wise pickups:

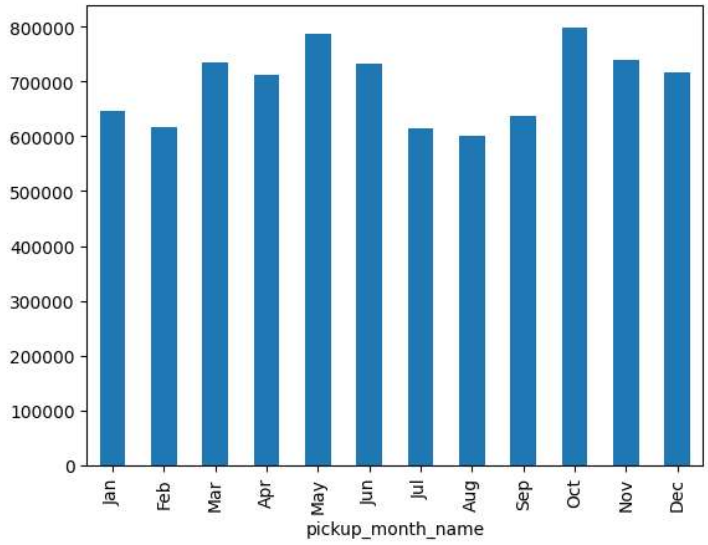


Pickups are relatively lower for months of Feb, July, Aug, Sept.

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

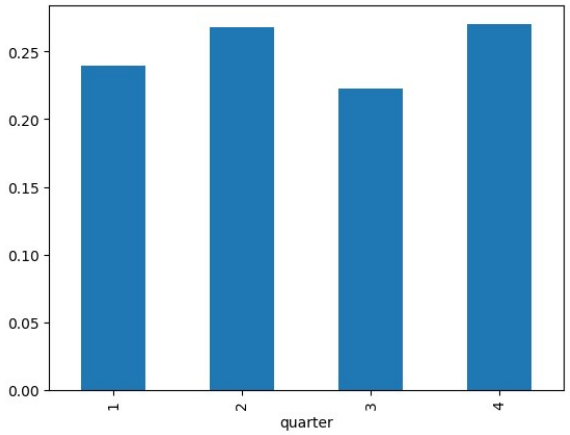
We created a df_copy which excluded these zero valued entries

3.1.4. Analyse the monthly revenue trends



Again months Feb, July, Aug, Sept have relatively lower contribution.

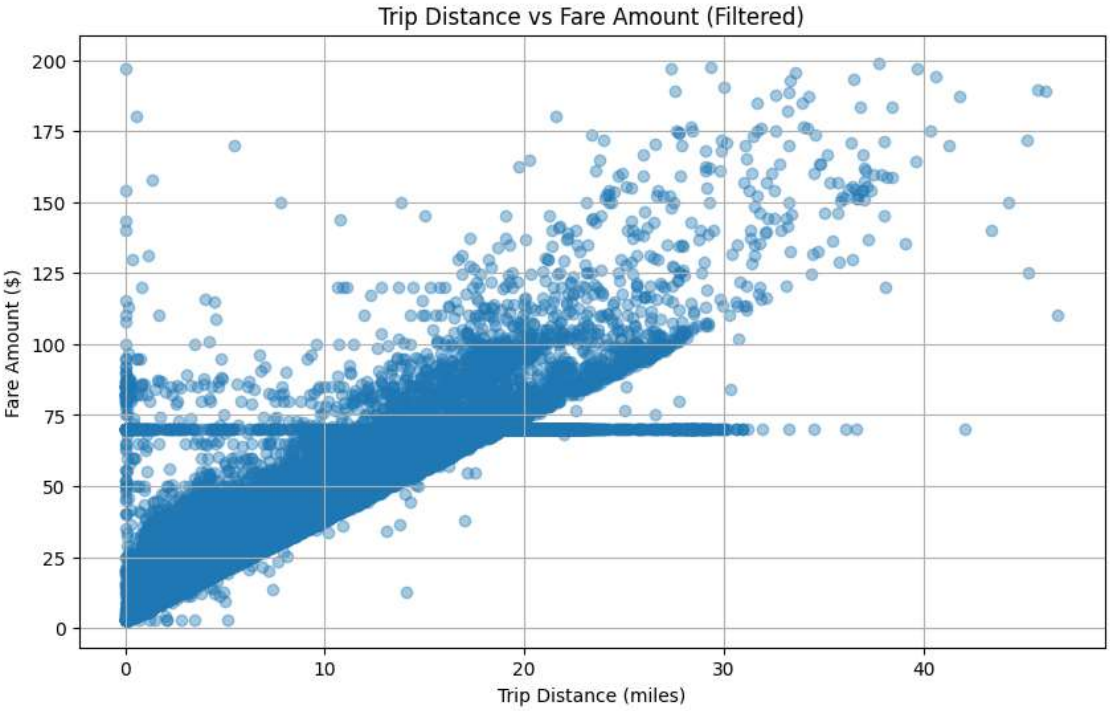
3.1.5. Find the proportion of each quarter's revenue in the yearly revenue



Quarter wise proportion gives a better picture, we can confidently say that Q1 and Q3 have lower revenue proportions

3.1.6. Analyse and visualise the relationship between distance and fare amount

Fare amount increases linearly with trip distance

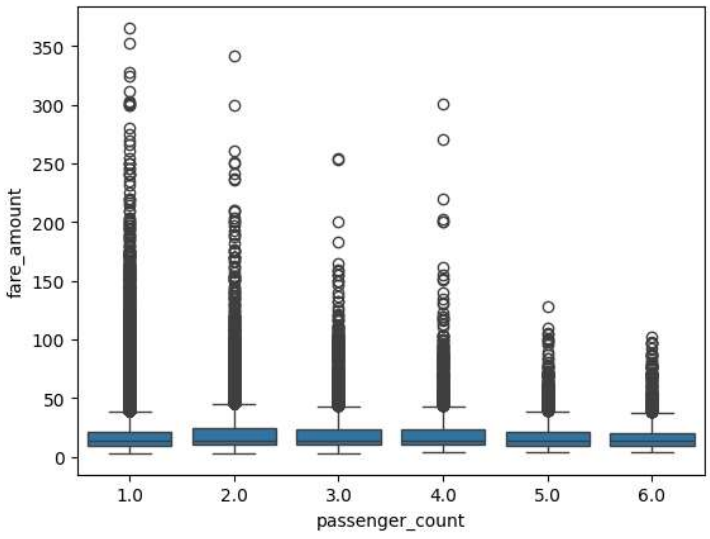


We can also notice a flat line at 70$, The points on this line had RatecodeID as 2 which indicates it was JFK flat fare, hence it did not vary with distance.

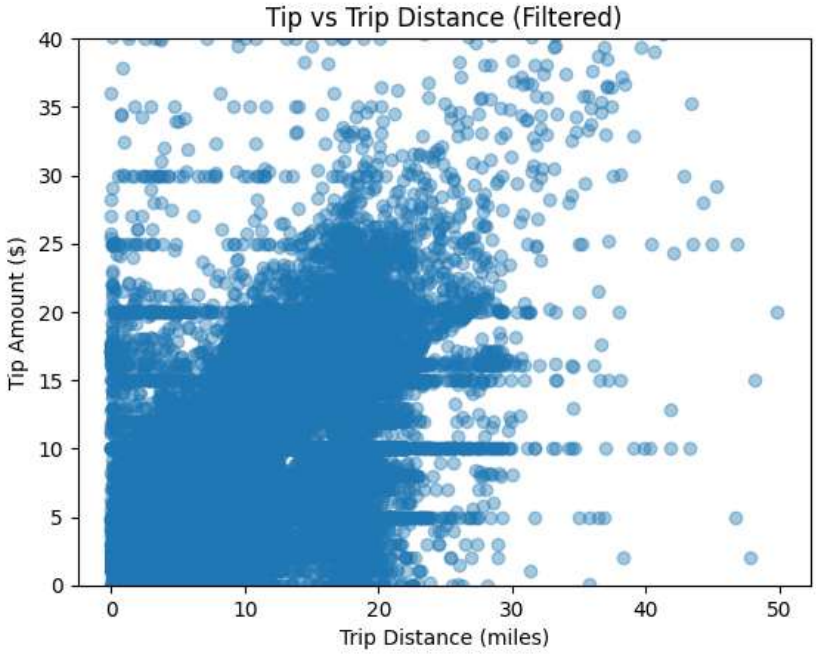
The correlation value between fare amt and trip distance was 0.954 which means they are highly positively related

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

Fare amount v/s passenger count: No strong relation



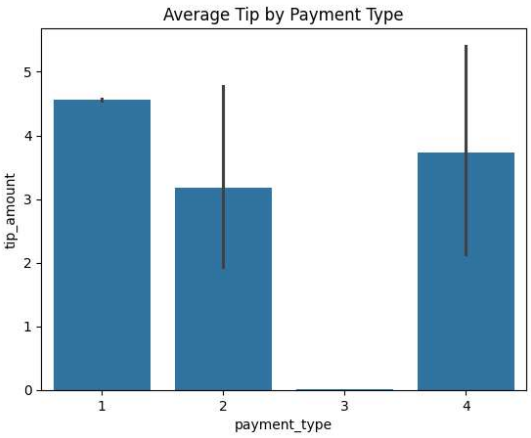
Tip vs distance



For longer rides the tip increases linearly with distance.

We can also notice horizontal lines at 5$, 10$, 15$ and son on upto 30$, This is obvious because its human tendency to give tips on multiples of 5.

3.1.8. Analyse the distribution of different payment types



- 1= Credit card

- 2= Cash

- 3= No charge

- 4= Dispute

3.1.9. Load the taxi zones shapefile and display it

3.1.10. Merge the zone data with trips data

zones data was merged into the df_copy using a left inner join on PULocationID

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

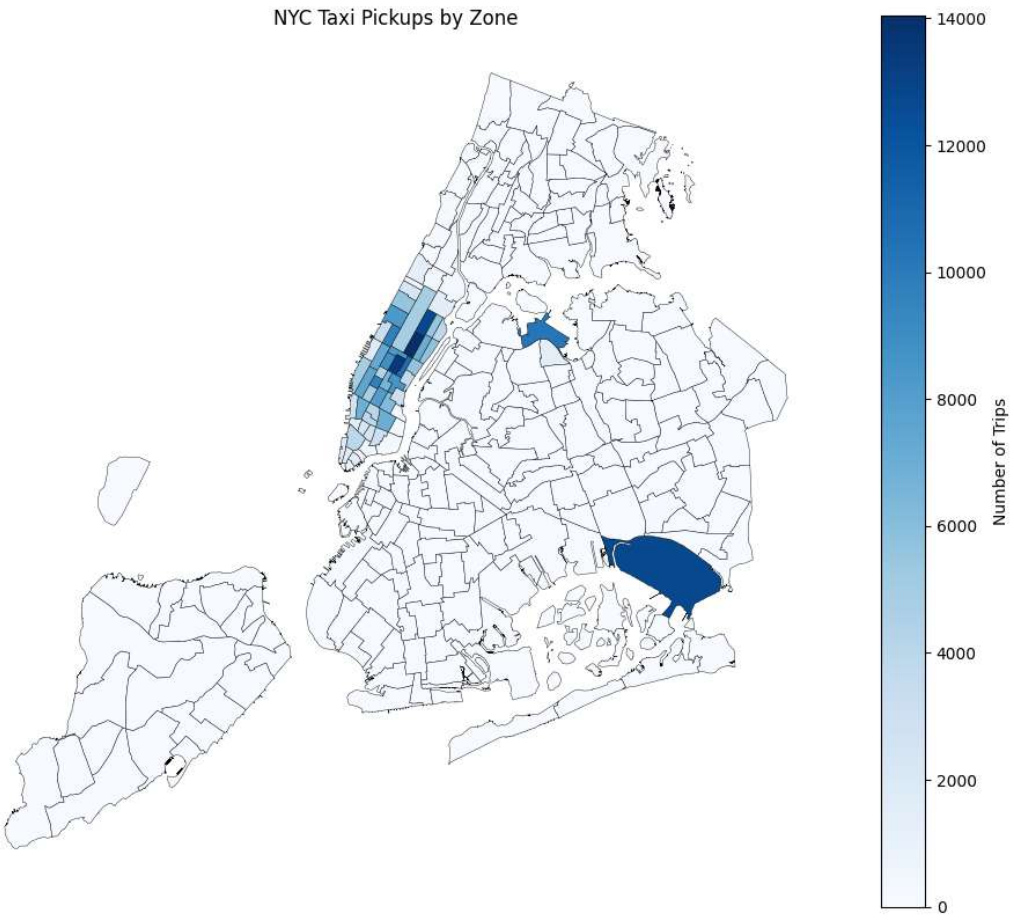
3.2.1. Find the number of trips for each zone/location ID

value_counts() method used to find the zone wise trip count and stored in a df trip_counts

3.2.2. Add the number of trips for each zone to the zones dataframe

The trip_counts record is merged back to zones geodata frame using a left inner join on LocationID

3.2.3. Plot a map of the zones showing number of trips



3.2.4. Conclude with results

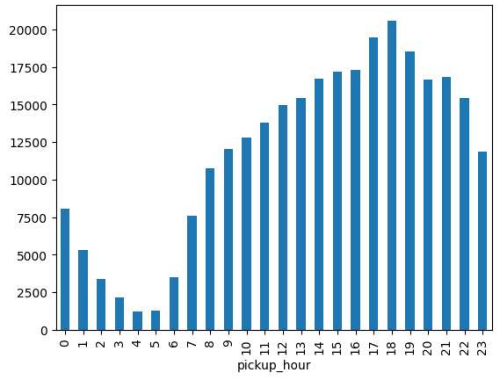
3.3. Detailed EDA: Insights and Strategies

3.3.1. Identify slow routes by comparing average speeds on different routes

We group the data by pickup, dropoff locations and the pickup hour. Now on this grouped aggregated data we calculate the average speed.

Sorting the data gives us slowest routes, some routes are having speed as slow as 0.04 miles per hour. Also many routes are having avg speed below 1 and 2 miles per hour.

3.3.2. Calculate the hourly number of trips and identify the busy hours



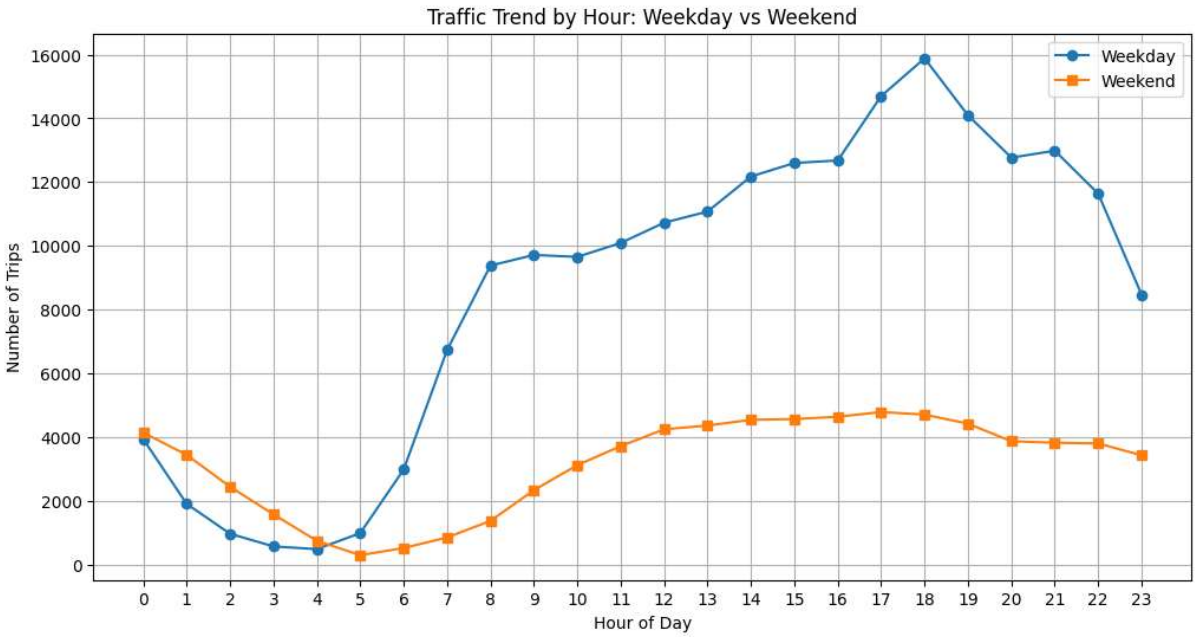
But this data is 1percent of our records

3.3.3. Scale up the number of trips from above to find the actual number of trips

Pickup hour: No of trips

18	2058700
17	1948400
19	1851600
16	1731300
15	1716300

3.3.4. Compare hourly traffic on weekdays and weekends



3.3.5. Identify the top 10 zones with high hourly pickups and drops

Top 10 Pickup Zones with Names and Boroughs:

	LocationID	zone	borough	Pickup_Count
0	237	Upper East Side South	Manhattan	14053
1	161	Midtown Center	Manhattan	13428
2	236	Upper East Side North	Manhattan	12786
3	132	JFK Airport	Queens	12752
4	162	Midtown East	Manhattan	10667
5	138	LaGuardia Airport	Queens	10249
6	142	Lincoln Square East	Manhattan	9709
7	186	Penn Station/Madison Sq West	Manhattan	9688
8	230	Times Sq/Theatre District	Manhattan	8829
9	170	Murray Hill	Manhattan	8591

Top 10 Dropoff Zones with Names and Boroughs:

Key observation is Manhattan followed by Queens is the major Pickup borough.

Top 10 Dropoff Zones with Names and Boroughs:

	LocationID	zone	borough	Dropoff_Count
0	236	Upper East Side North	Manhattan	13524
1	237	Upper East Side South	Manhattan	12507
2	161	Midtown Center	Manhattan	11120
3	170	Murray Hill	Manhattan	8703
4	239	Upper West Side South	Manhattan	8517
5	142	Lincoln Square East	Manhattan	8307
6	162	Midtown East	Manhattan	8242
7	141	Lenox Hill West	Manhattan	7992
8	230	Times Sq/Theatre District	Manhattan	7734
9	68	East Chelsea	Manhattan	7254

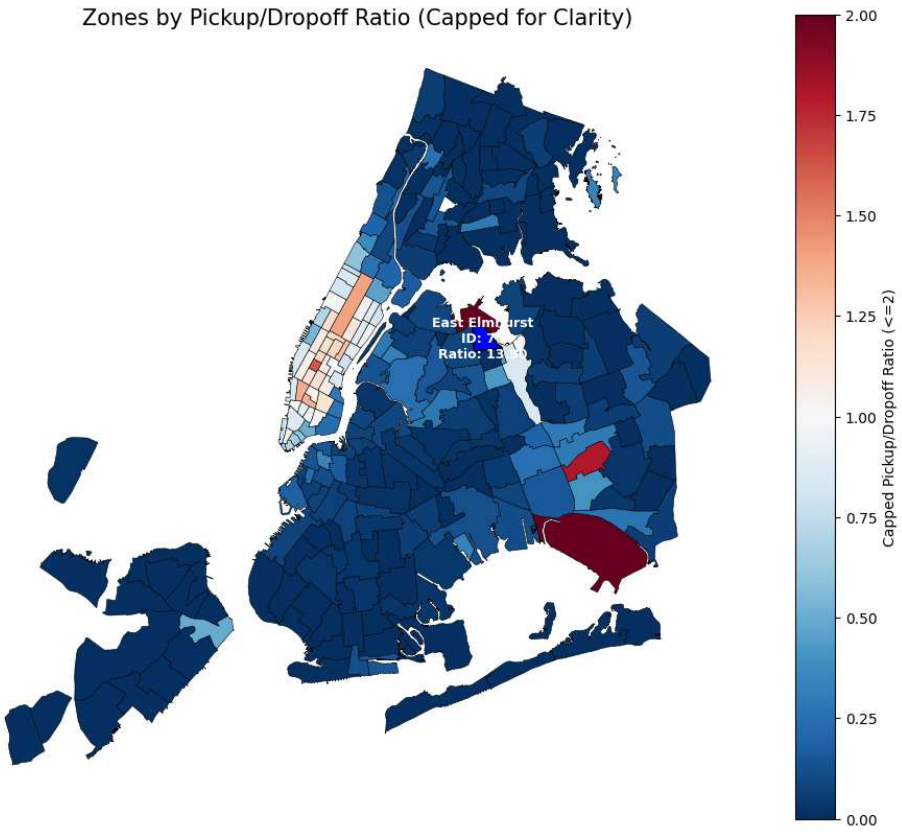
3.3.6. Find the ratio of pickups and dropoffs in each zone

Top 10 Pickup/Dropoff Ratios:

	pickups	dropoffs	dropoff_safe	pickup_drop_ratio
70	1293.0	93	93	13.903226
132	12752.0	2673	2673	4.770670
138	10249.0	3522	3522	2.909994
215	27.0	15	15	1.800000
186	9688.0	5997	5997	1.615474
43	4881.0	3472	3472	1.405818
249	6892.0	5065	5065	1.360711
114	3856.0	2862	2862	1.347310
162	10667.0	8242	8242	1.294225
100	4269.0	3518	3518	1.213474

We can also plot the zones colored by their pickup/dropoff ratios

The below plot was made capping out East Elmhurst due to it having highest pickup to drop ratio.



3.3.7. Identify the top zones with high traffic during night hours

Top 10 Night Pickup Zones with Names and Boroughs:

	LocationID	zone	borough	Night_Pickup_Count
0	79	East Village	Manhattan	2575
1	249	West Village	Manhattan	2137
2	132	JFK Airport	Queens	1949
3	148	Lower East Side	Manhattan	1607
4	48	Clinton East	Manhattan	1564
5	114	Greenwich Village South	Manhattan	1391
6	230	Times Sq/Theatre District	Manhattan	1241
7	186	Penn Station/Madison Sq West	Manhattan	1041
8	138	LaGuardia Airport	Queens	993
9	68	East Chelsea	Manhattan	945

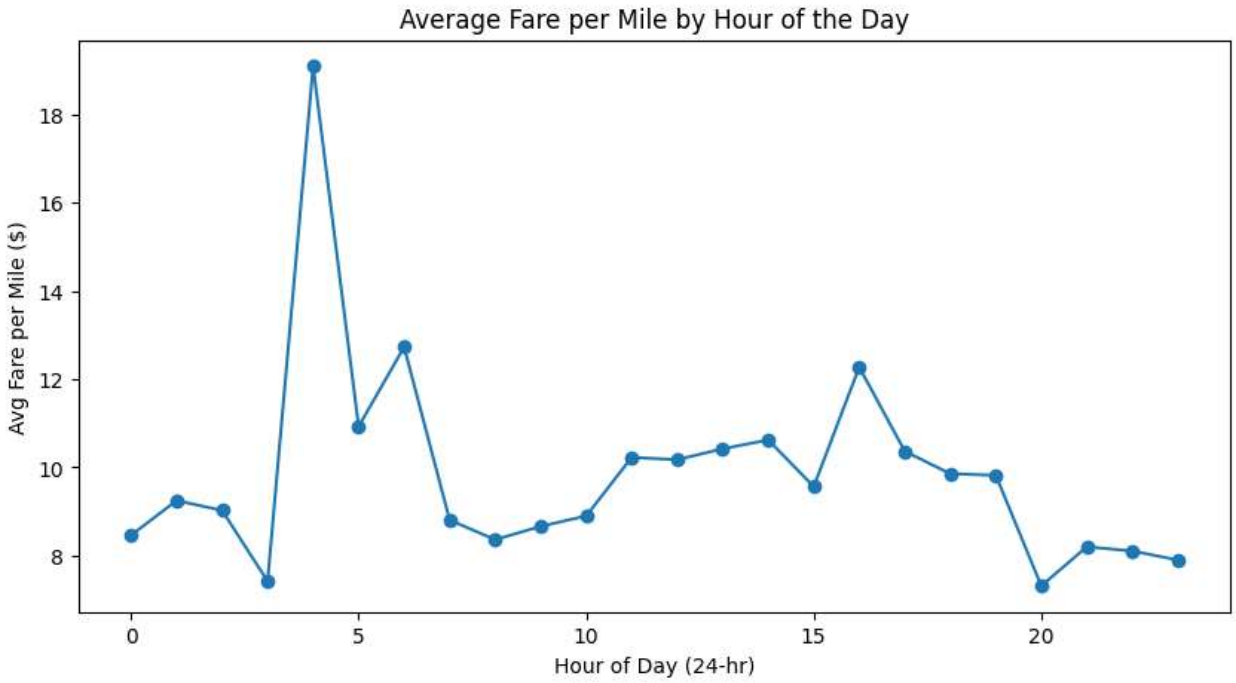
3.3.8. Find the revenue share for nighttime and daytime hours

Time Category	Revenue proportions (%)
DAY	**88.01**
NIGHT	**11.98**

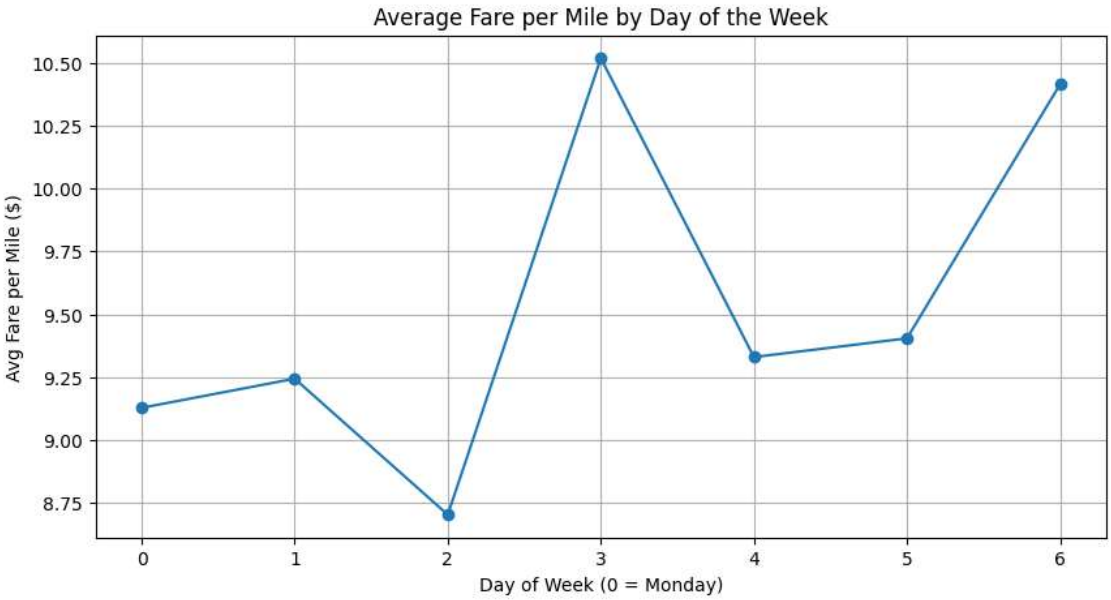
3.3.9. For the different passenger counts, find the average fare per mile per passenger

	passenger_count	fare_per_mile	fare_per_mile_per_passenger
0	1.0	9.150703	9.150703
1	2.0	10.653140	5.326570
2	3.0	11.810095	3.936698
3	4.0	14.703518	3.675879
4	5.0	7.579811	1.515962
5	6.0	7.590740	1.265123

3.3.10. Find the average fare per mile by hours of the day and by days of the week



The Fare per mile peaks around 4 am, This is also high around 7 am and 4 pm.



The Fare per mile does not vary much numerically

3.3.11. Analyse the average fare per mile for the different vendors

Vendor ID	Average Fare
1	**7.88$**
2	**10.08$**

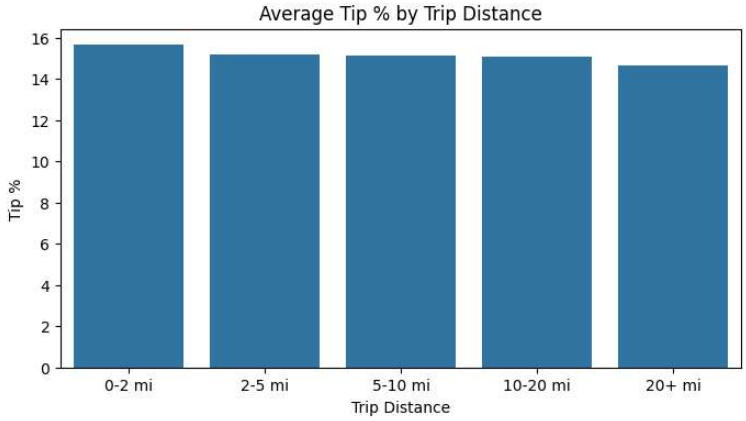
3.3.12. Compare the fare rates of different vendors in a distance-tiered fashion

	VendorID	Distance Tier	Average Fare per Mile
0	1	0-2 miles	9.540645
3	2	0-2 miles	13.743683
1	1	2-5 miles	6.436262
4	2	2-5 miles	6.551485
2	1	5+ miles	4.474127
5	2	5+ miles	4.501630

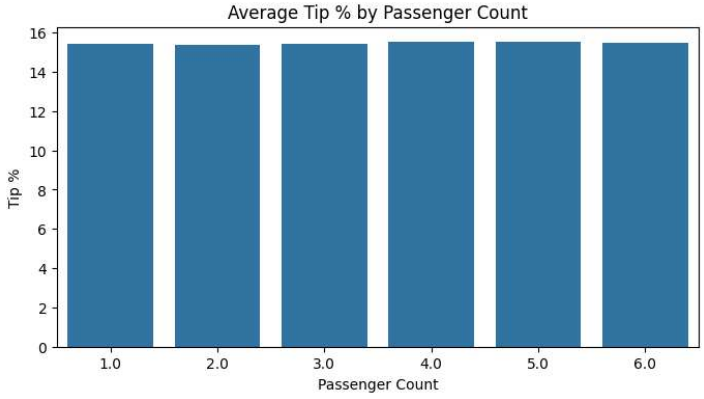
We observe for shorter trips (below 2 miles) Vendor 1 is more cheaper

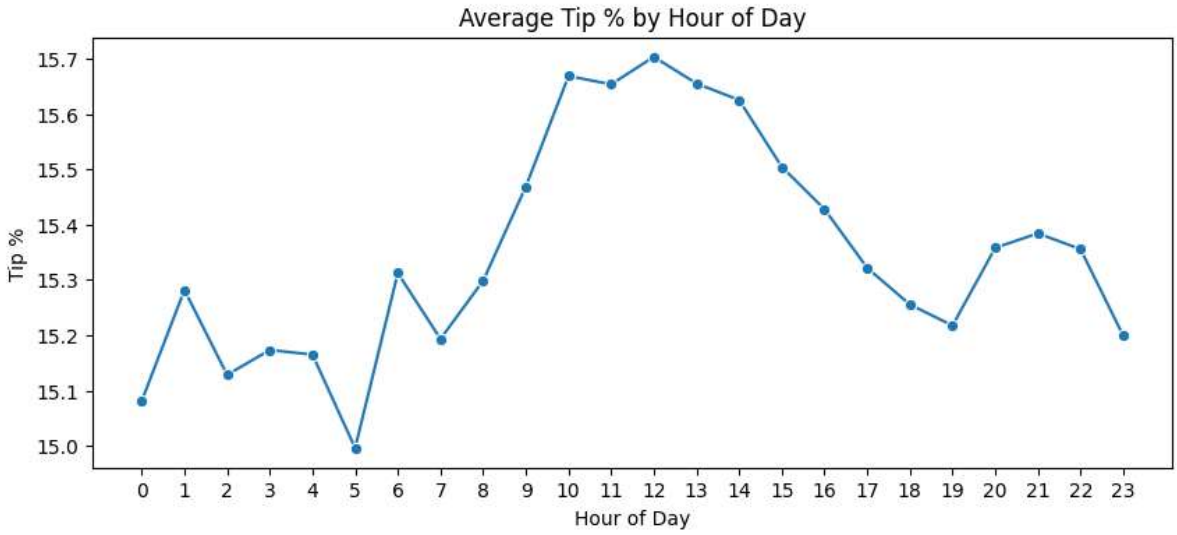
3.3.13. Analyse the tip percentages

Tips increase with distance but the tip/fare percentage remains same



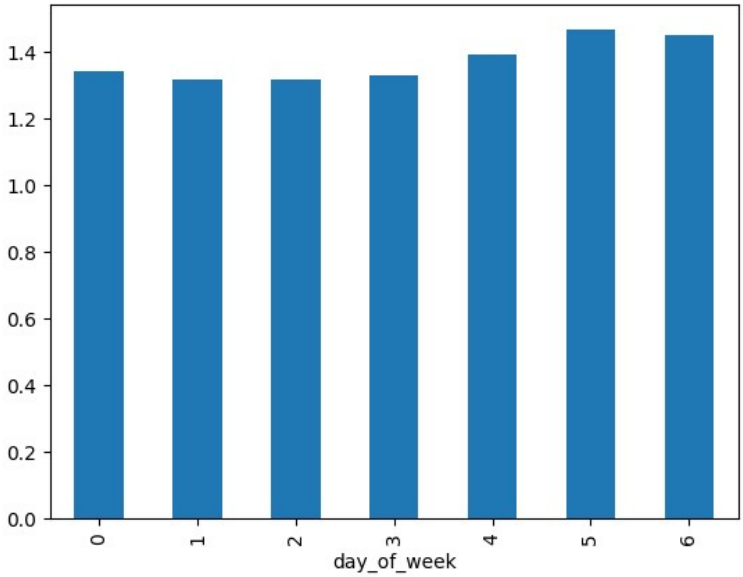
Passenger count also doesn't affect tip percentage



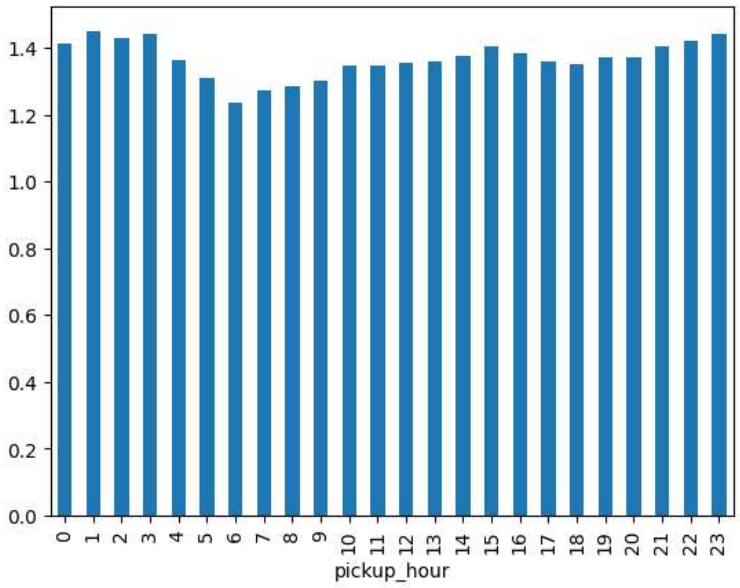


Tip percentage is highest between 10 AM and 2 PM

3.3.14. Analyse the trends in passenger count

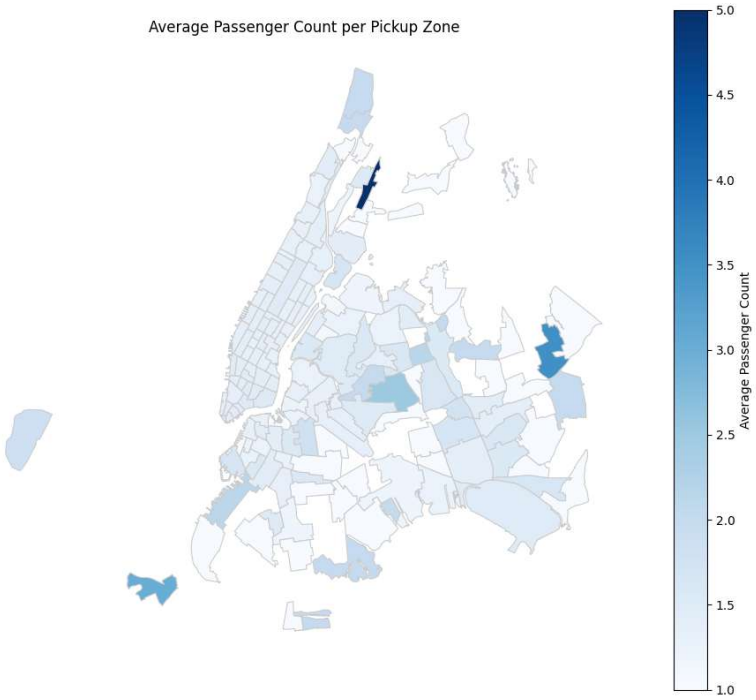


Average passenger count remains between 1.34 and 1.45

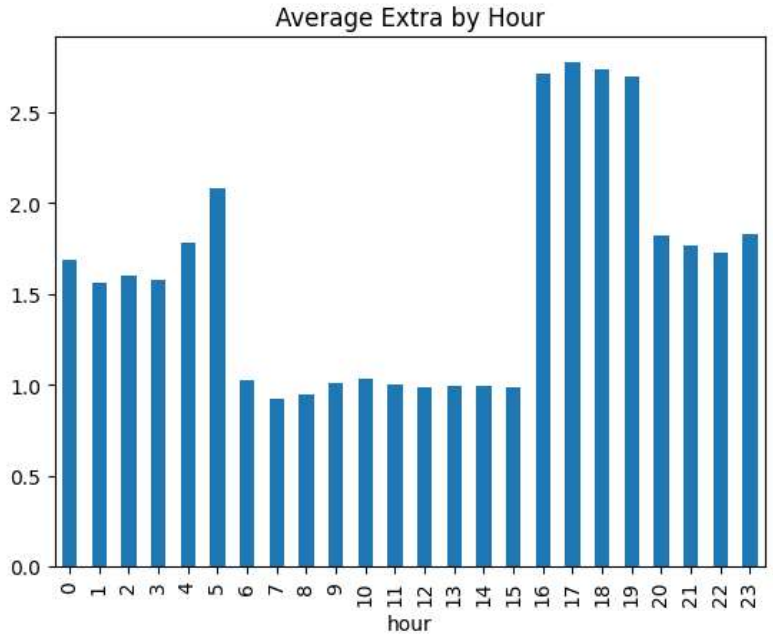


Average passenger count doesn't vary much

3.3.15. Analyse the variation of passenger counts across zones



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



Extra charges have a structured pattern:

8pm to 5am = 1.5$

5am to 3pm = 1$

4pm to 8pm = 2.5$

4. Conclusions

4.2. Final Insights and Recommendations

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

Optimizing routing and despatching:

Time of Day Effects: The busiest hours were observed between 5 PM to 8 PM, especially on weekdays, with increased demand for pickups in Manhattan. Dispatching more cabs preemptively in these hours can reduce passenger wait times.

Inefficient Routes: Some routes consistently show slow speeds, especially in congested areas during peak hours. For example, average speeds dropped to as low as 0.04 mph in certain zones,

often due to traffic congestion or signals. These bottlenecks should be avoided or rerouted using real-time traffic data.

Pickup/Dropoff Ratios: Some regions are popular for pickups and other for dropoffs, using the PU/DO ratios we can devise strategies to optimize dispatching, like matching two nearest zones one having High PU/DO ratio to the low PU/DO ratio so cabs from a zone which is popular for dropoff can be sent to its nearest zone high PU/DO zone for pickups again.
For eg: East Elmhurst (Location ID - 70) has a exceptionally high PU/DO ratio maybe because of its airport. So to optimize routing, we can dispatch cabs from neighbouring zones low PU/DO to minimize waiting times

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

High-Demand Zones: Most zones in Manhattan are hotspots for both pickups and drop-offs. Placing idle cabs strategically in or near these areas can increase trip frequency and revenue.

Temporal Positioning:

Temporal analysis shows that trip volume peaks in the evening on weekdays, while weekends have a more uniform trip distribution, with notable midnight travel.

Weekday Strategy: More cabs should be dispatched and available between 5–8 PM. Reduced fleet may suffice after midnight.

Weekend Strategy: Fleet availability should remain consistently high throughout the day, especially ensuring night coverage due to steady demand.

Zone-specific strategy: Zones with higher tip percentages often correlate with longer trips and better service, especially in affluent neighborhoods or airport routes. Positioning cabs in such zones can increase driver incentives and revenue.

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

Fare per Mile by Distance Tiers:

For short trips (< 2 miles), Vendor A generally offers slightly higher per-mile rates than Vendor B.

For longer trips (> 5 miles), fares level out, making competitive pricing essential.

Night vs. Day Revenue Share: Night trips contribute less to overall trip volume but can generate proportionally higher average fares (due to surcharges and longer distances). Maintain night surcharges to encourage availability during these hours.

Tips & Service Quality: Higher tip percentages correlate with:

*Longer distances,

*Lower passenger counts (1–2),

*Midday and evening pickups.

There is a potential that the company can explore personalized service or premium options during these periods to encourage tipping and improve driver earnings.

Flat rates: The JFK flat rates can be relaxed as it is not feasible for nearby zones, and it will lead to drivers rejecting trips from far away zones