Report: Optimising NYC Taxi Operations

Include your visualizations, 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

I sampled the data as asked (5% of every hour of the day) and combined the file and saved that to a different parquet file named - "Sampled_NYC_taxi_records"

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

I fixed the index to have a ordered index and also remove the first row, as its pickup date is 31 DEC 2022, which may disturb further analysis.

2.1.1.1. Combine the two airport_fee columns

In those 12 files for 12 months, the column “airport_fee” is named as “Airport_fee” in some of those files, as a result, we have two columns in the final file where one is NaN where other has a value. To combine them, I simply add them with fill value of 0 for NaN, as only one of them will have value, and finally drop the “Airport_fee” column.

- **Fix columns with negative (monetary) values**

extra	1	total_amount	81
mta_tax	74	congestion_surcharge	63
airport_fee	14	Improvement_surcharge	81

Fare amount has no negative rows. Only these columns have negative values. So, after I remove the rows where total amount is negative, surprisingly, all the other column that have negative values, now have none. That means these are the surcharges that adds up the negative total amount with either zero or close to zero fare amount.

I looked up if there are any rows where pickup time is actually less than or equal to the drop off time, and there are in fact some rows. Compared to the 19 lakhs rows we have, it's only approx. 800. The rows where pickup time and drop off time is exactly same, we can see the most of the trip distance column is zero, but not the fare amount. There are messy and unreliable, so I dropped them.

2.2. Handling Missing Value

2.2.1. Find the proportion of missing values in each column

Columns	Missing value proportion
VendorID	0.000000
tpep_pickup_datetime	0.000000
tpep_dropoff_datetime	0.000000
passenger_count	3.430697
trip_distance	0.000000
RatecodeID	3.430697
store_and_fwd_flag	3.430697
PULocationID	0.000000
DOLocationID	0.000000
payment_type	0.000000
fare_amount	0.000000
extra	0.000000
mta_tax	0.000000
tip_amount	0.000000
tolls_amount	0.000000
improvement_surcharge	0.000000
total_amount	0.000000
congestion_surcharge	3.430697
airport_fee	3.430697

All of the five columns have exact proportion of missing values, so there's a possibility that they're the same rows.

2.2.2. Handling missing values in passenger_count

passenger_count	68449
RatecodeID	68449
store_and_fwd_flag	68449
airport_fee	68449
congestion_surcharge	68449

I checked the number of missing values, where passenger count is missing, and we can see that these five columns have the same amount So, we can conclude that we are 100% sure that all of the 68410 rows where passenger count is NaN , all the other four columns were also negative in all rows. That’s why we got the exact same proportion of missing columns in the first place. As a result, after dropping these rows where passenger count is NaN, we would have no rows with missing values.

There are also more than 30000 rows where passenger count is 0, but the fare has normal values. It is surely not possible, and has some kind of reliability issues, that’s why I also dropped these rows.

2.2.3. Handle missing values in RatecodeID

Now, there is no need for handling missing values for rest of the columns, as now they have none. We can also see that from the output.

2.2.4. Impute NaN in congestion_surcharge

This is same for this column also. We can reverify from the output also that there is no missing rows for this column.

2.3. Handling Outliers and Standardising Values

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

There are already some obvious suggestions for outlier handling, so at first, I checked them myself.

Passenger count	Number of rows
1.0	1447902
2.0	292805
3.0	72501
4.0	40936
7.0	4
5.0	25386
6.0	16625
8.0	8
9.0	5

As we can see, there is only a handful of rows for passenger count above 6, so it will be best to just simply drop these rows.

Distance Tiers	Total rows
(0, 2]	1038552
(2, 5]	509369
(10, 50]	165838
(5, 10]	160203
(50, 100]	240
(100, 200]	17
(200, 250]	0
(250, 1000]	0

As we can see that, for trip_distance, there no rows above 200, and only 17 rows above 100. So, there’s no point dropping rows above 250.

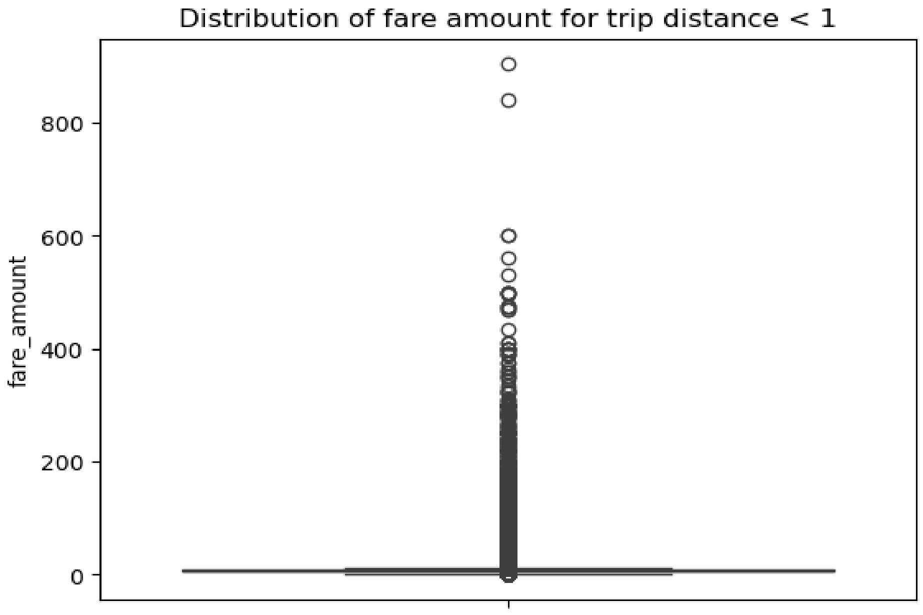


Figure 1

I considered 1 mile as a reasonably low trip distance, but for that we have a lot of values higher than usual. It can easily be declared as outlier, and as recommended I'm dropping the rows which has fare amount more than 300 dollars (47 rows).

There are also 15 rows where, trip distance and fare amount are 0, but the pickup location and drop off location are different, so I dropped them

I can see that there is no payment type "0", but with that I also noticed most of payment is done through credit card, but cash payment also has a considerable value. This has to be kept in mind for future reference.

For standardizing, I created three new columns, pickup_hour and dropoff_hour by the pickup and drop off datetime and trip_duration by taking the difference between these two columns. It will be used later.

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical

I first manually specify the categorical variables, as there are a few ordinal categorical variables. Then I specify rest of them as numerical columns.

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

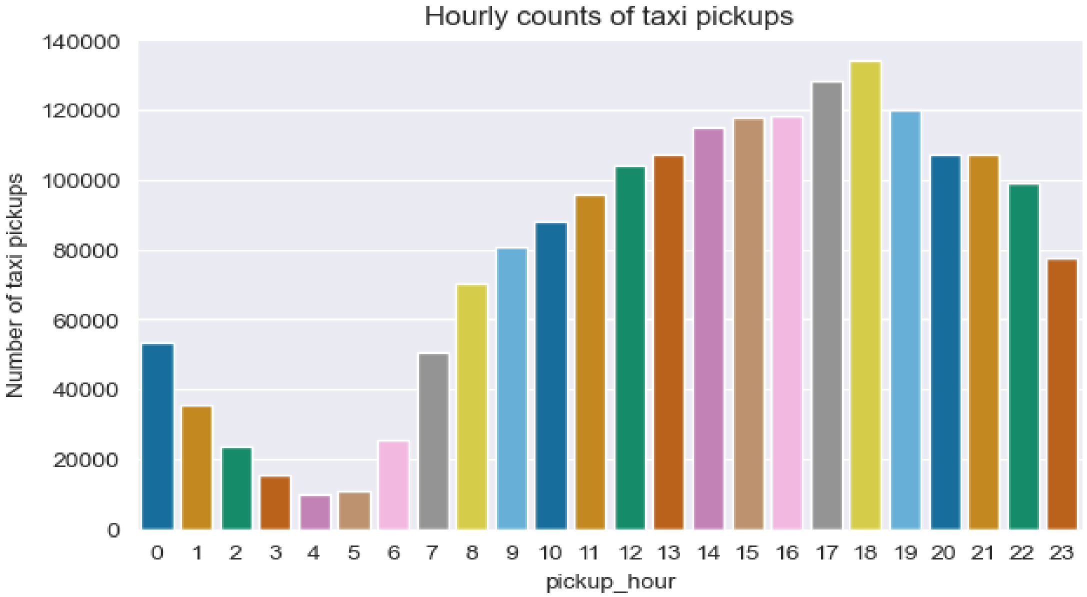


Figure 2

As we can clearly see, at late night and in the early morning, there are very few pickups, but as we get started with the day, pickup count is increasing and it peaks in the evening time.

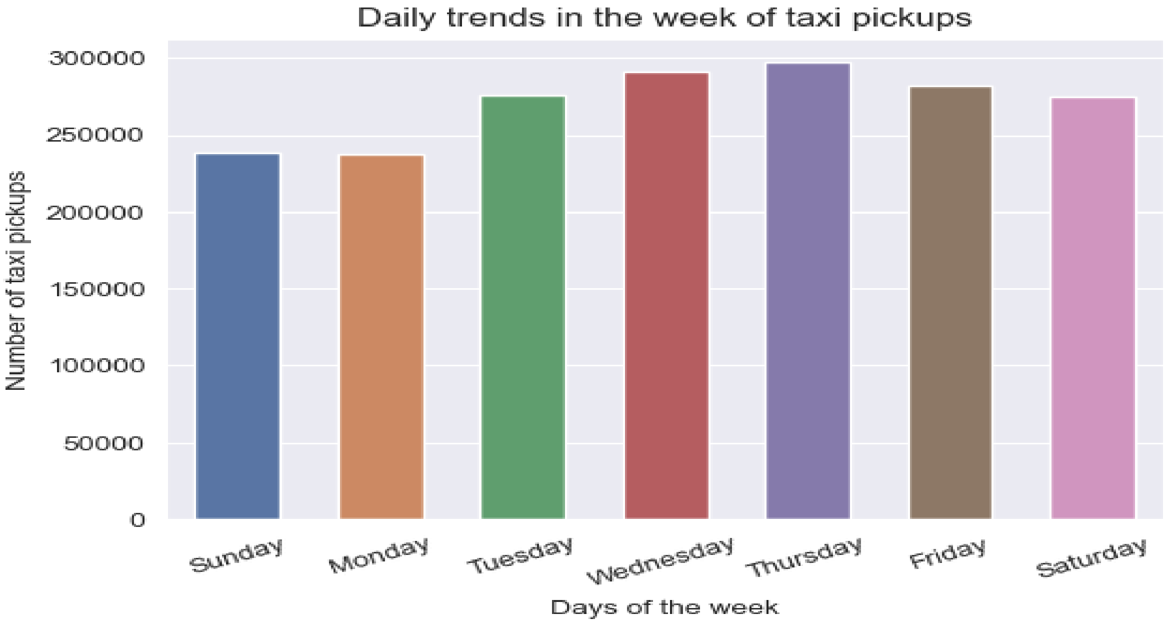


Figure 3

It’s not surprising, that weekdays have the most number of pickups on average.

From what I can observe, it's completely random. I don't find any clue of pattern from this plot.

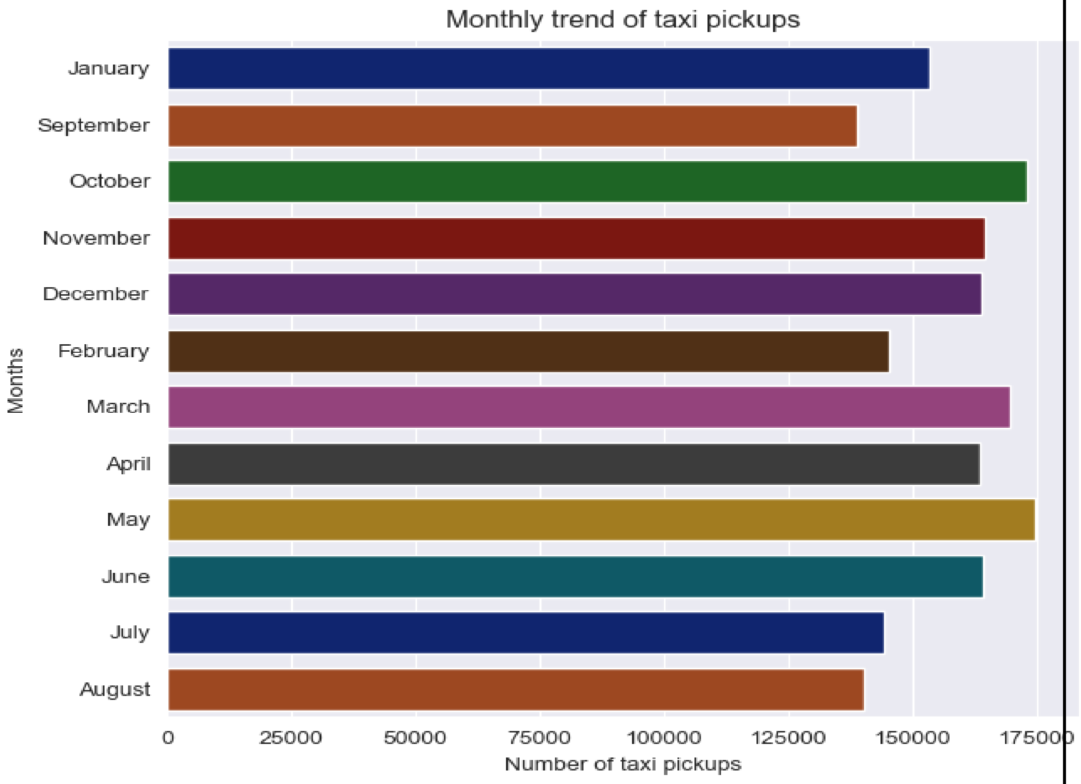


Figure 4

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

fare_amount	456
tip_amount	422776
total_amount	217
trip_distance	21866

These are the number of zeroes present in each of those columns. The possible reason behind a considerably large amount of zeroes in tip_amount column can be because of cash tips, as they are not part of our data so they are most probably marked as zero.

Now, I first take the rows where neither of fare amount of total amount is zero. Then I checked, the trip duration and fare amount columns for two conditions where trip distance is 0, one is pickup location and drop off location is the same, and another is different. The results are below.

Pickup location not equals to drop off location

Percentiles	Trip duration	Fare amount
0.5	0 days 00:16:57	18.45
0.75	0 days 00:33:40	36.60
0.90	0 days 00:53:16	57.20
0.99	0 days 01:32:00	92.80
1	6 days 06:14:54	275.00

Pickup location == drop off location

Percentiles	Trip duration	Fare amount
0.50	0 days 00:00:15	10.00
0.75	0 days 00:00:48	70.00
0.90	0 days 00:05:28	83.00
0.99	0 days 00:39:28.	178.00
1.00	1 days 16:33:03	300.00

Analysing both we can see that, the rows where pickup and drop off location are the same, they have reasonably good data except some outliers as the median is only 34 seconds, buy the rows where the pickup and drop off location are different despite having trip distance zero, they have normal range of data except some outliers. So, in both cases we just need to get rid of the outlies.

Percentiles	Trip duration	Fare amount
0.75	0 days 00:11:34	47.20
0.99	0 days 01:16:50	158.41
1.00	6 days 06:14:54	300.00

That’s why, we analyse the top percentiles for trip distance = 0, and found that we can take values up to 99 percentiles. After that, those extreme values appear.

3.1.4. Analyse the monthly revenue trends

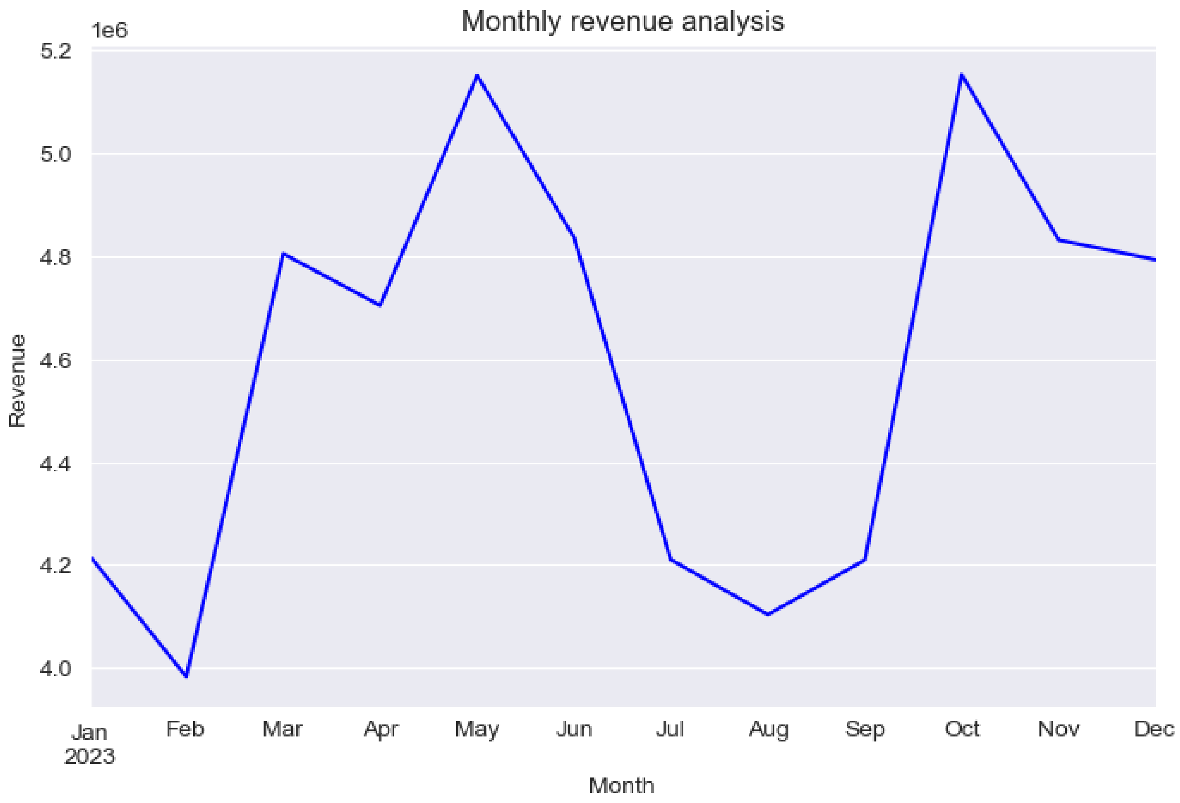


Figure 5

We can see some peaks and drops in the plot. I have some reasonable guesses behind these, the drop in January and February can be because of extreme cold and snowfall and the people in New York take their vacation in summer period which affects daily commuting, that explains the drop in July, August and September.

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

Quarter	Revenue Share
2023Q1	23.648
2023Q2	26.712
2023Q3	22.771

2023Q4	**26.869**

Here, we can see that the second quarter and fourth quarter have the most amount of revenue share. Clearly, because the first quarter consists of January, February and the second quarter consists of July, August and September.

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

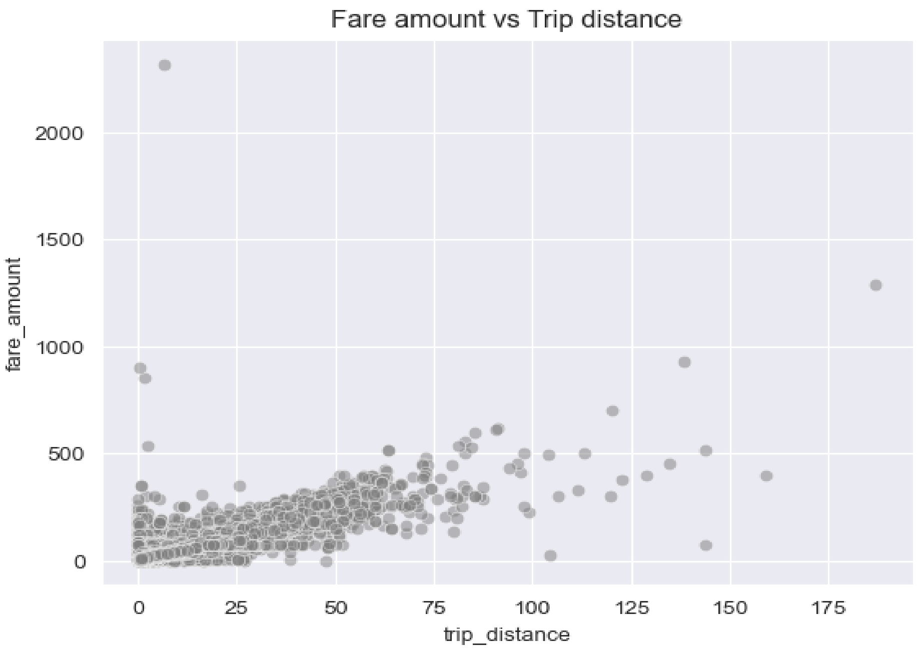


Figure 6

As we can see and also confirm from the correlation score which is 0.94, trip distance and fare amount are highly correlated which is expected also.

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

From the plot we can see that, Fare amount and trip duration is positively correlated. We can be surer of this fact by the correlation score of these two which is 0.26, which tells us moderated positive correlated between these

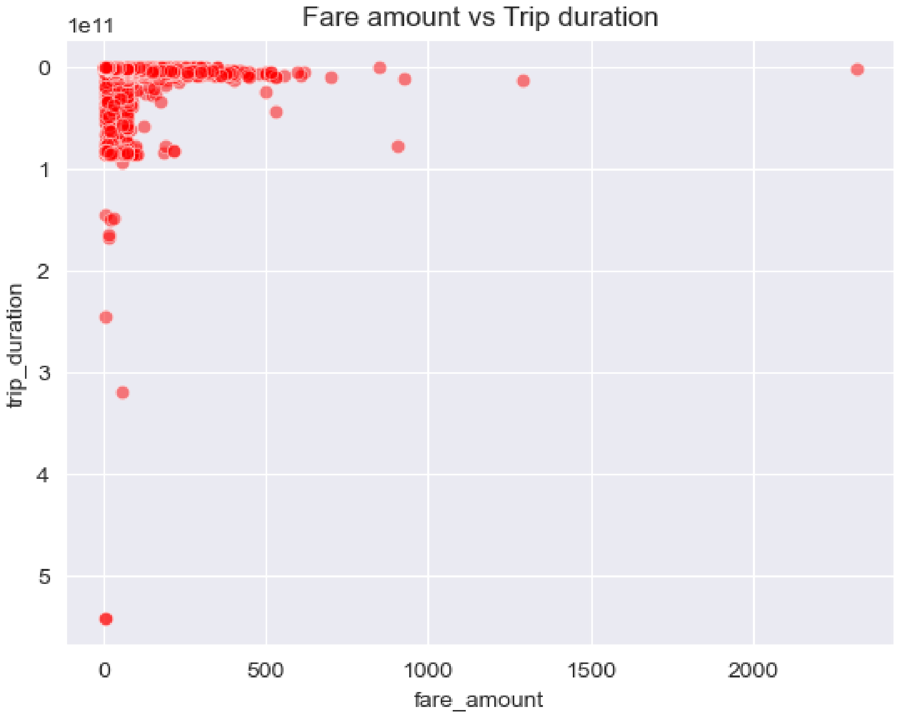


Figure 7

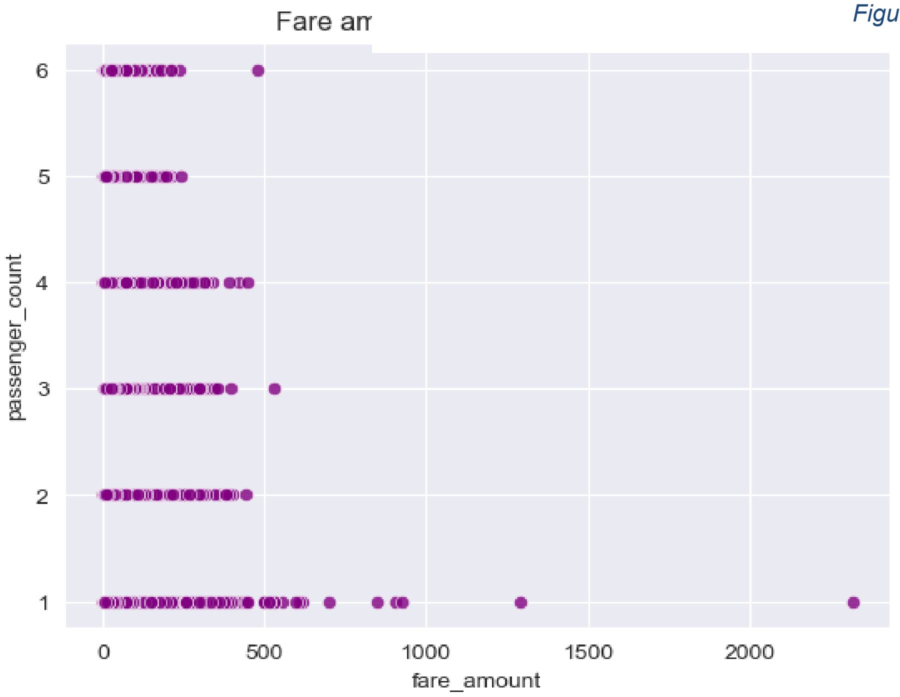


Figure 8

It very clear that these two has no correlation between them and completely independent of each other. Their correlation score is 0.04.

Tip amount has a strong positive correlation with trip distance as a lot of points very close to the 45-degree angle. Their correlation score which is 0.57 supports this argument.

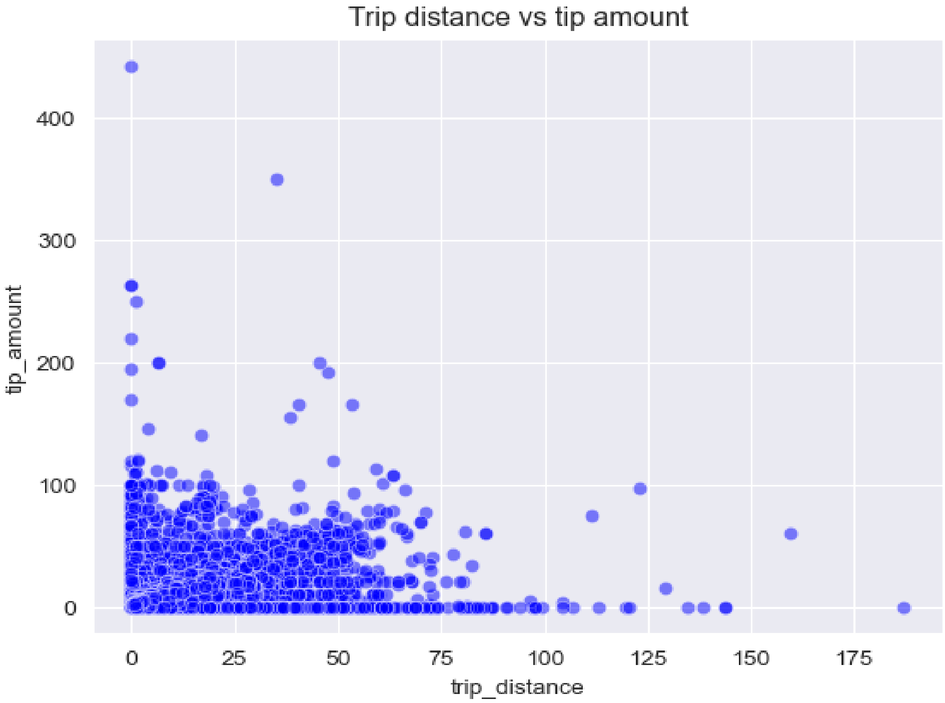


Figure 9

3.1.8. Analyse the distribution of different payment types

We can clearly see that most of the payment are done via credit card and cash though cash payments are considerably lesser than credit card payments. 3 represents no charge which is very less due to maybe coupon or some offers and 4 is code for dispute which can happen some times.

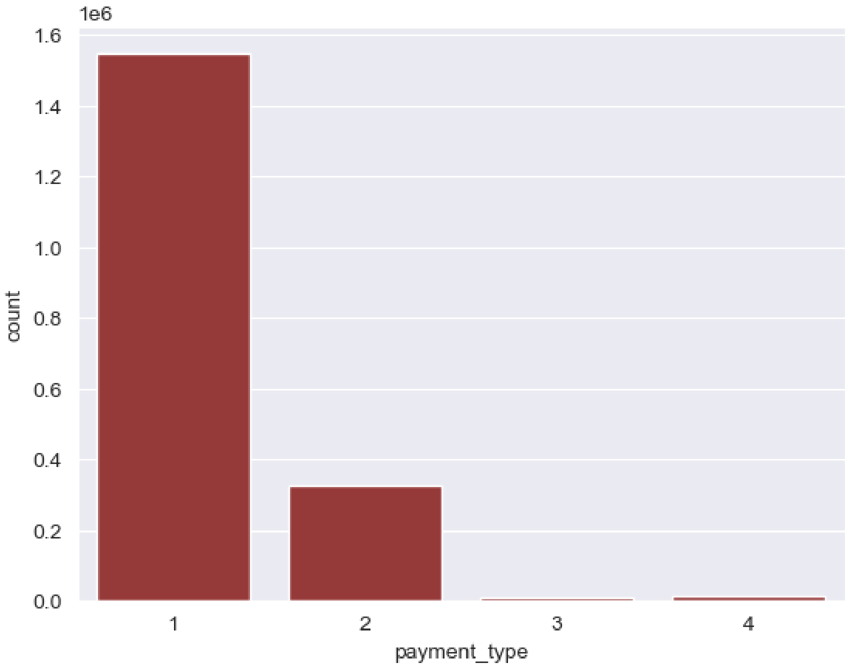


Figure 10

3.1.9. Load the taxi zones shapefile and display it

This is the zones shapefile, which may further give us many visual interpretations which are otherwise very hard to derive.

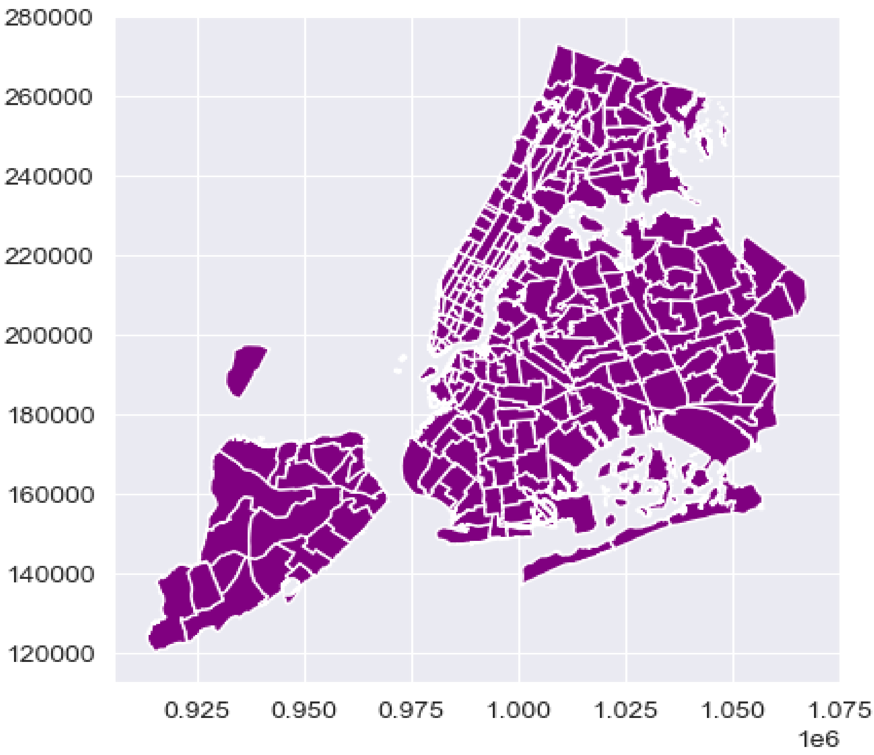


Figure 11

3.1.10. Merge the zone data with trips data

So, I’ve merged the zone data with trips data using location id and pickup location id as a common column.

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

For each one, I group by it, and determine its size, it gives us the number of rows in each zone or location id.

3.1.12. Add the number of trips for each zone to the zones data frame

We create another data frame joining zones and trips count data and named it as zones_trips.

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

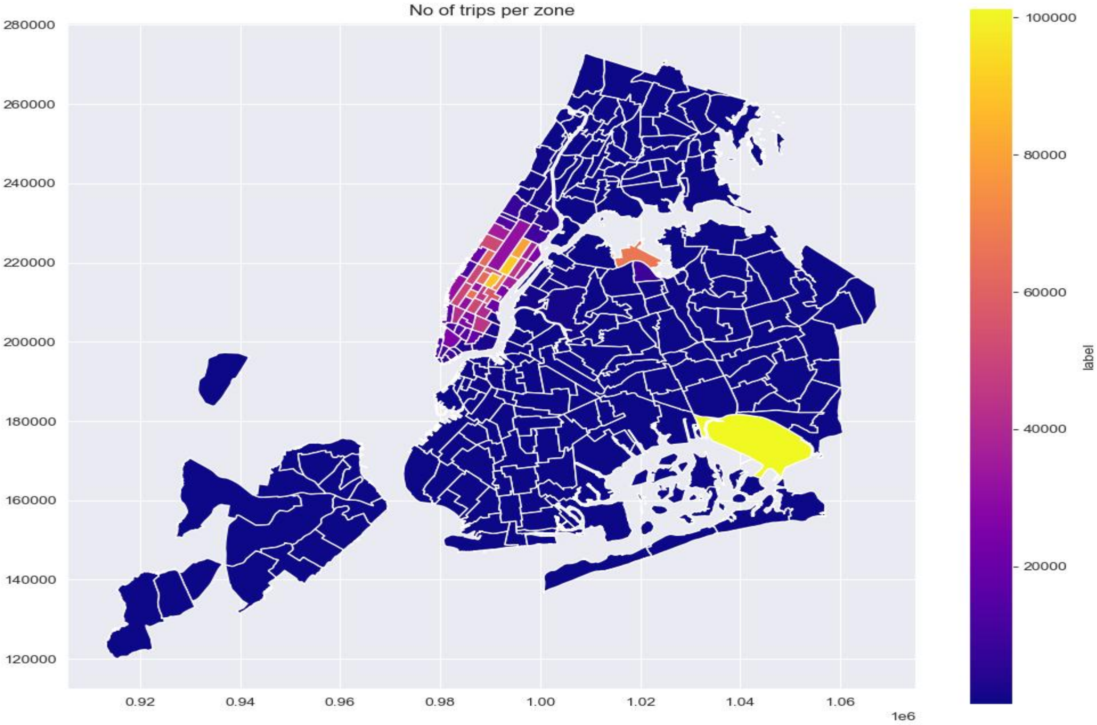


Figure 12

It can be seen that the most number of trips are at Airport and Manhattan area.

I created a custom zone group using the borough, to visualize better.

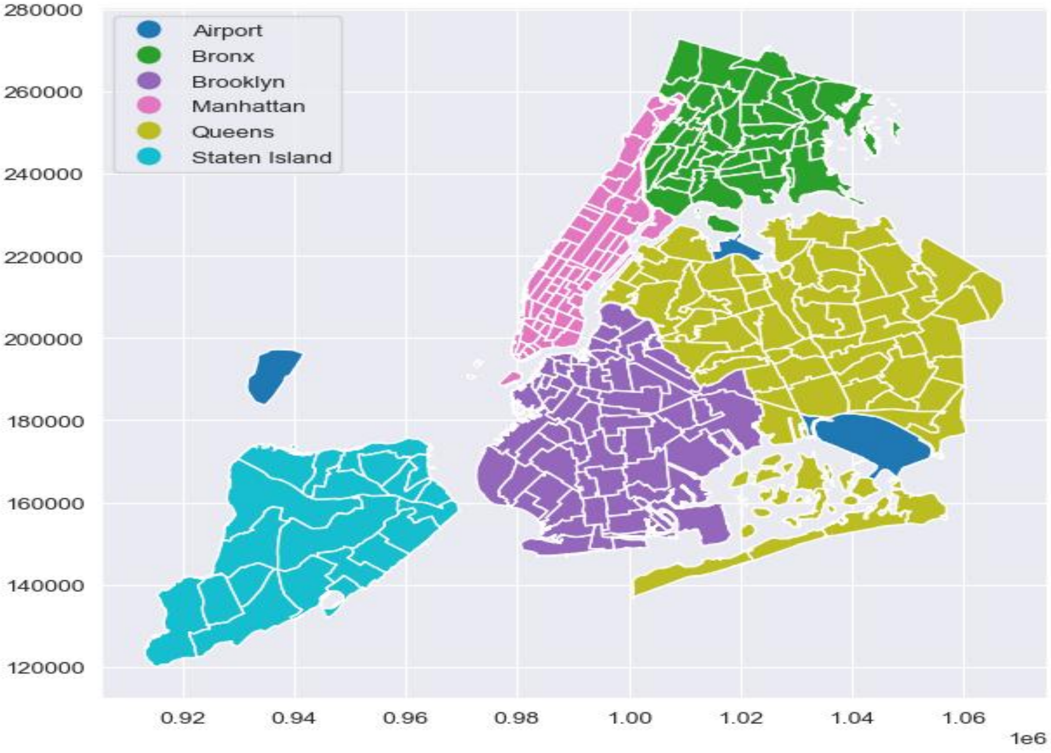


Figure 13

3.1.14. Conclude with results

Till this point we can conclude the followings:

1. The majority of trips happens at the daytime and evening hours.
2. Fare amount has zero correlation with passenger count so neither vendor nor drivers should focus on how many passengers are in the trip.
3. Airport and Manhattan have the maximum number of trips.
4. Tip amount has a moderate positive correlation with trip distance
5. The revenue share of quarter 1 and 3 is lower than that of quarter 2 and 4.

3.2. Detailed EDA: Insights and Strategies

3.2.1. Identify slow routes by comparing average speeds on different hours of day

Pick Up location ID	Drop off location ID	Pickup hour	Average trip duration
88	198	23	23.963333
161	121	20	23.963333
228	13	0	23.962222
236	217	12	23.935833
233	223	18	23.928056
40	65	21	23.907222
80	233	16	23.907222
48	106	17	23.901389
164	217	14	23.878611
48	247	23	23.873056

These are top 10 routes that has the highest trip duration. Now, either these routes are long trips or they’re congested and have more traffic.

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

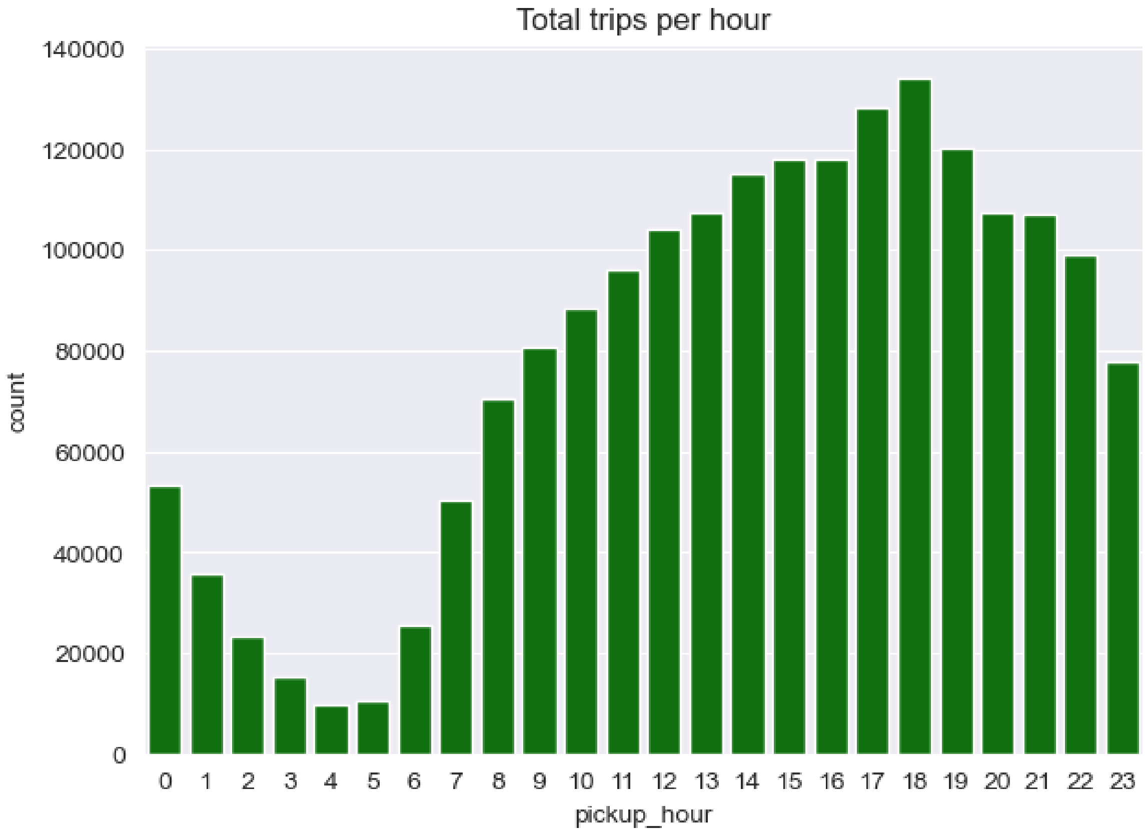


Figure 14

No surprise, the total trips per hour starts increasing after 5 AM and peaks at the afternoon and evening time.6 pm in the evening is the busiest hour according to our analysis.

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

18	2681380.0
17	2561920.0
19	2403460.0
16	2361560.0
15	2360160.0

These are the actual number of trips, in the top 5 busiest hours. Because I took a 5% of data from each hour, that’s why I scaled up the data by the same ratio.

3.2.4. Compare hourly traffic on weekdays and weekends

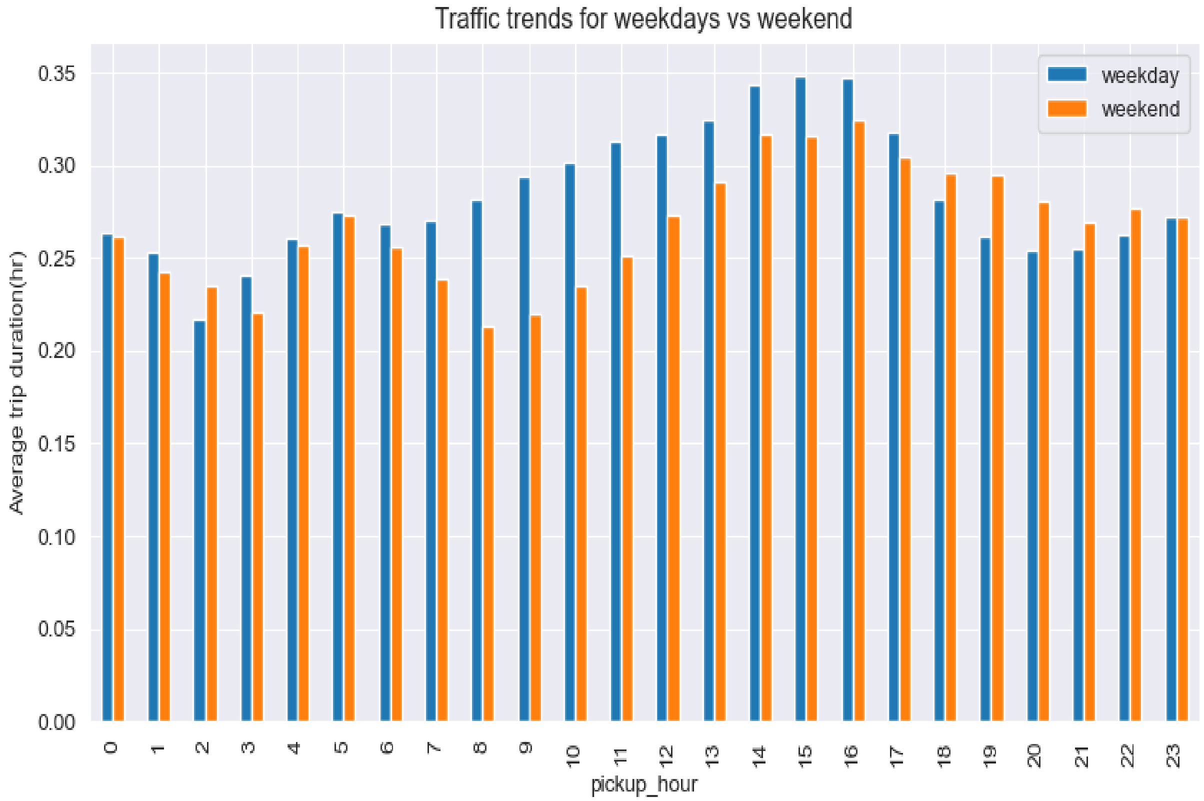


Figure 15

It is very clear that in weekdays, during the late morning and day time (from 8 am to 2 pm), average trip duration is significantly higher than that of weekend. But the evening trips are slightly higher in weekend time than that of weekday.

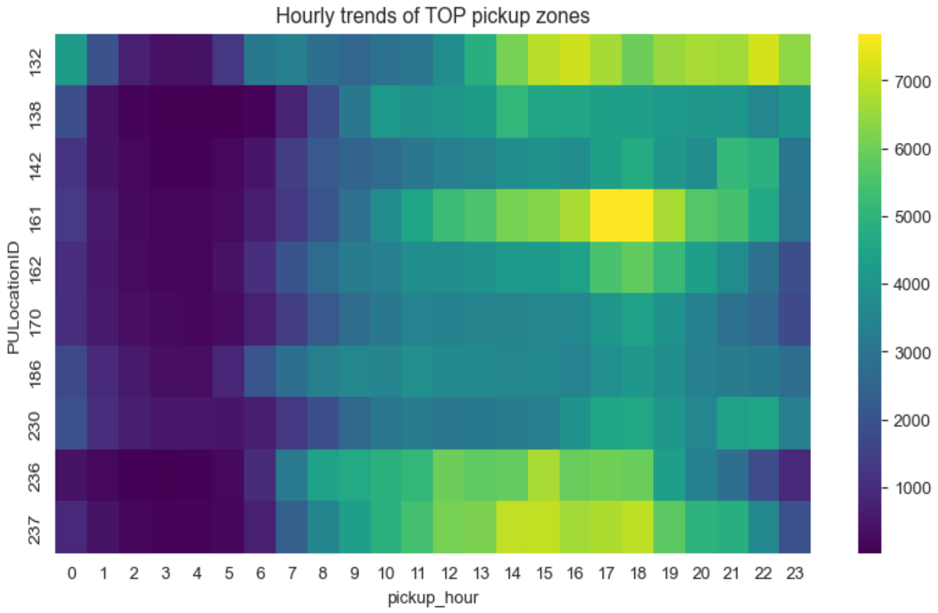


Figure 16

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

Here we can see that the top 10 pickup zones and drop off zones with high hourly pickups and in which hour it peaks the most.

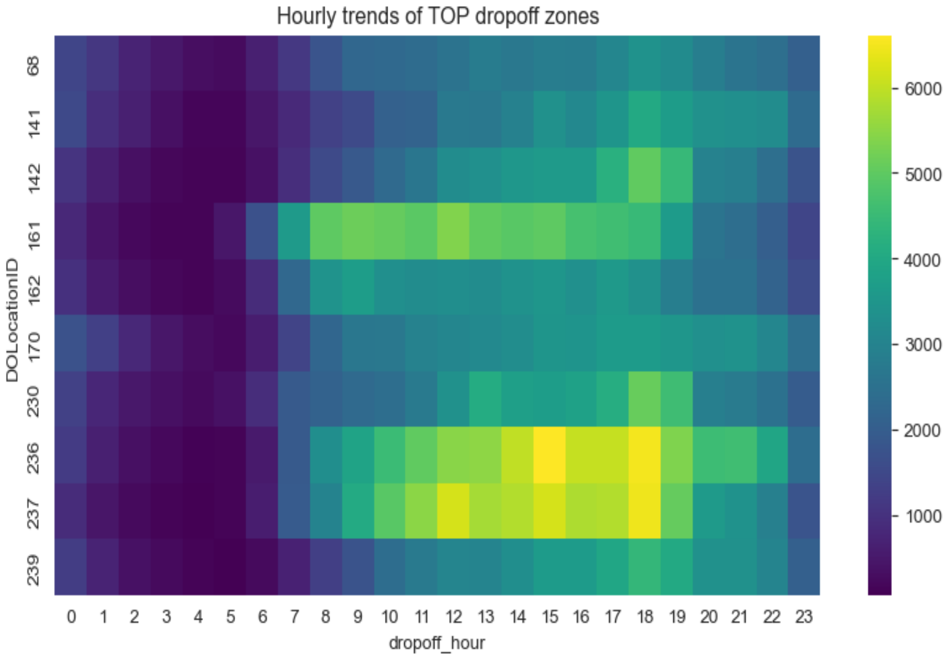


Figure 17

3.2.6. Find the ratio of pickups and drop offs in each zone

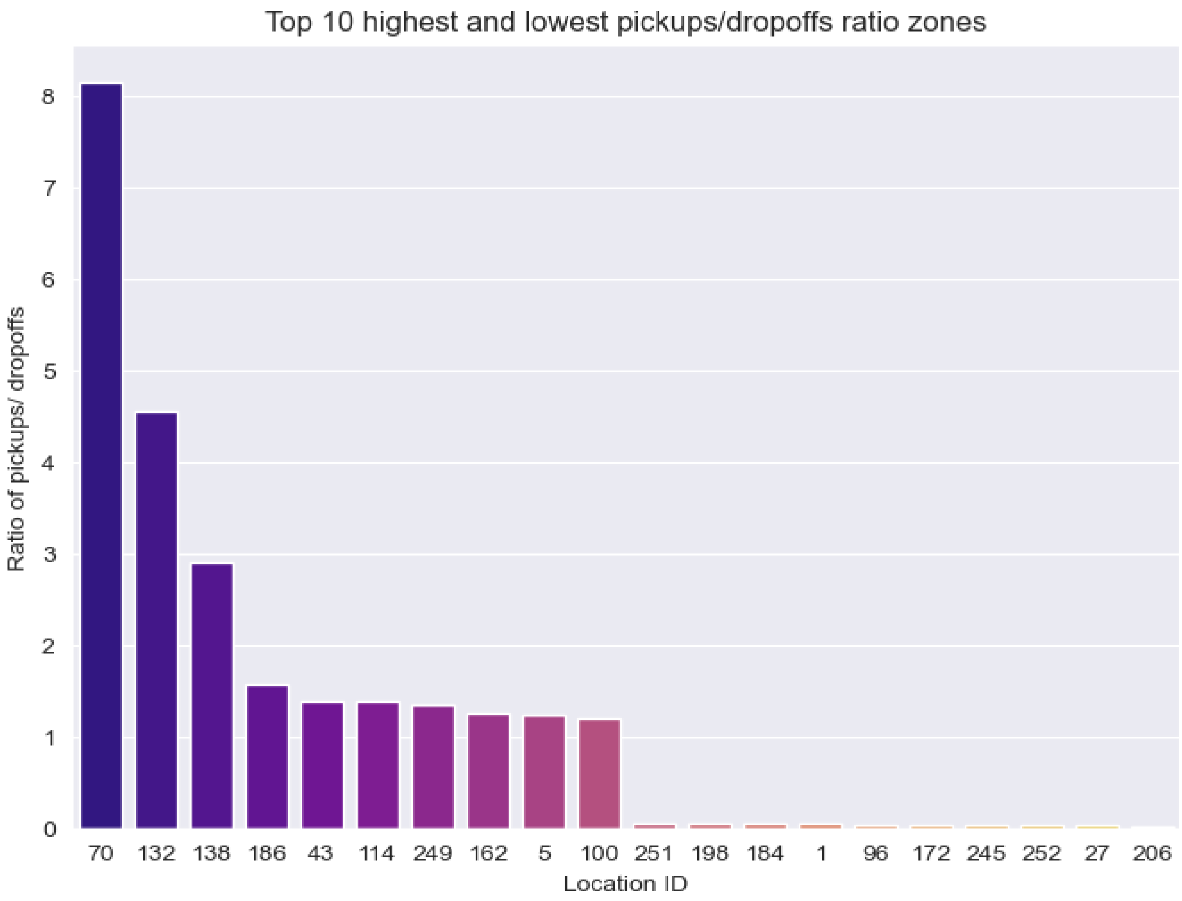
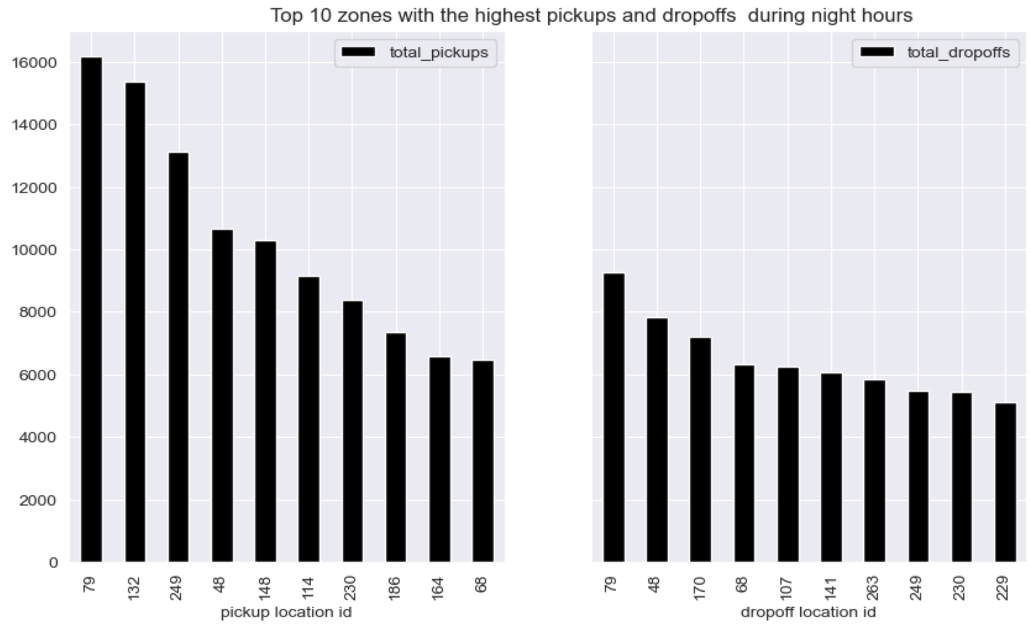


Figure 18

The first 10 are the highest ratio of pickups and drop offs and the last 10 are the lowest in the same category.

The location with low ratio, means there's only drop offs, driver can't find new pickups ride there. We have to keep this in mind.

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



As we can see, the top zones in pickup and drop offs during night hours.

Figure 19

3.2.8. Find the revenue share for nighttime and daytime hours

From here, we can clearly see that the major amount of revenue comes from Daytime hours. Nighttime don't contribute that much except at 11 PM at night.

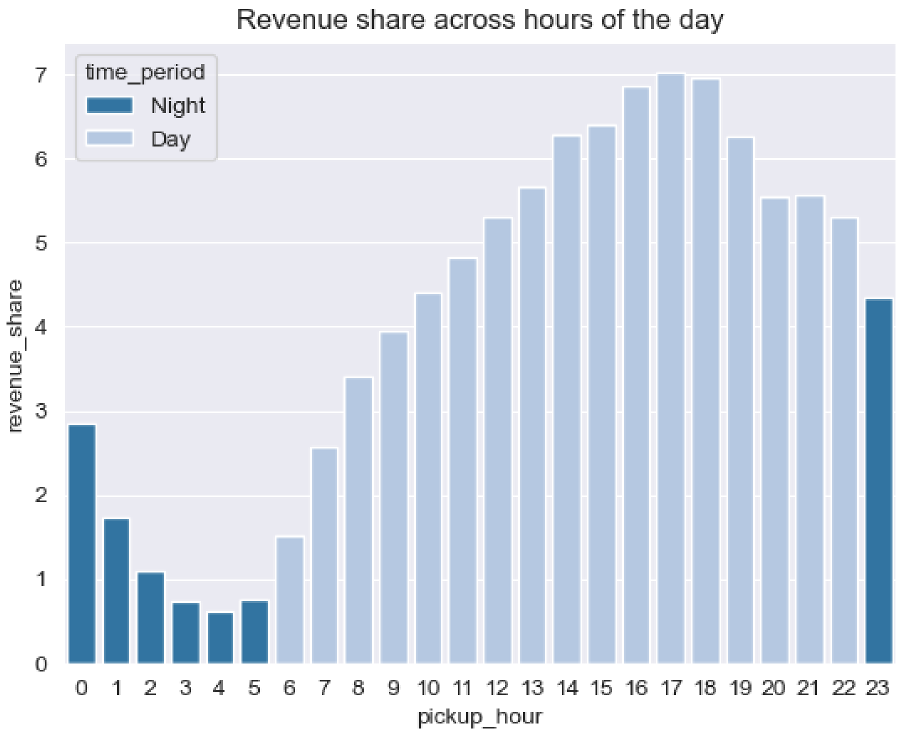


Figure 20

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

This is expected behavior. Usually fare amount doesn’t depend on passenger counts. That’ why more passenger count means avg fare per person becomes lower.

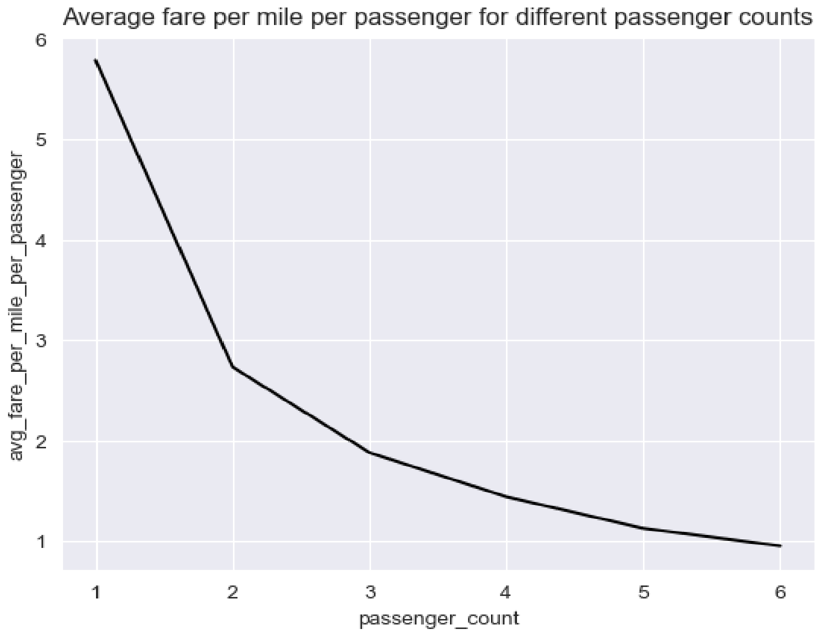


Figure 21

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

Due to less demand, the average fare per mile during late night and early morning is low and contrarily the peak office hours have comparatively higher average fare per mile.

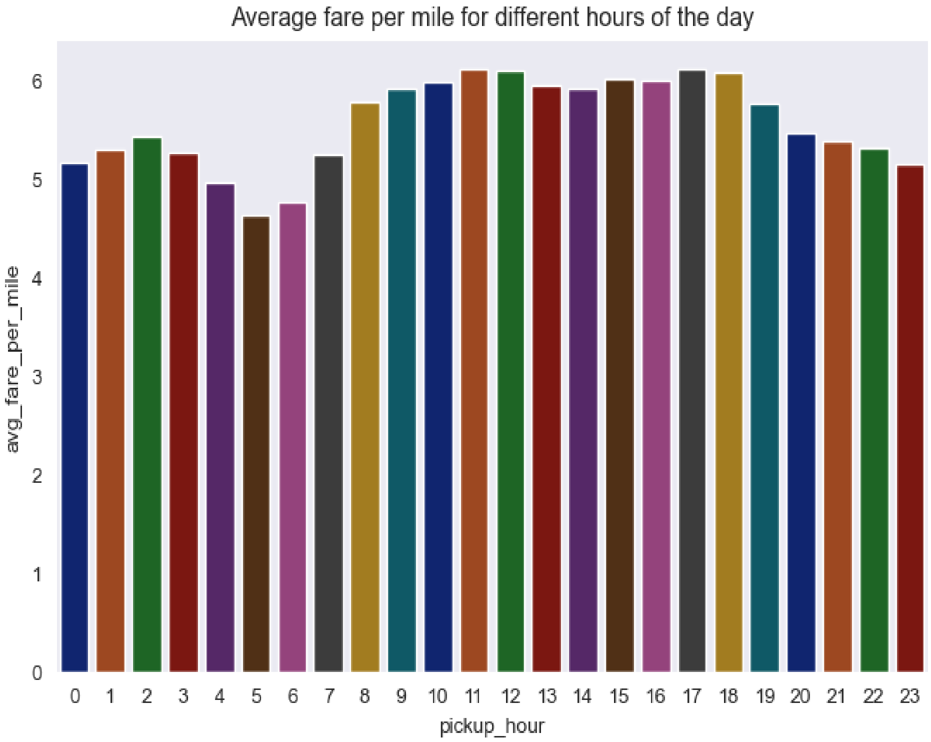


Figure 22

Weekdays have a higher average fare than weekends. It is completely normal as expected.

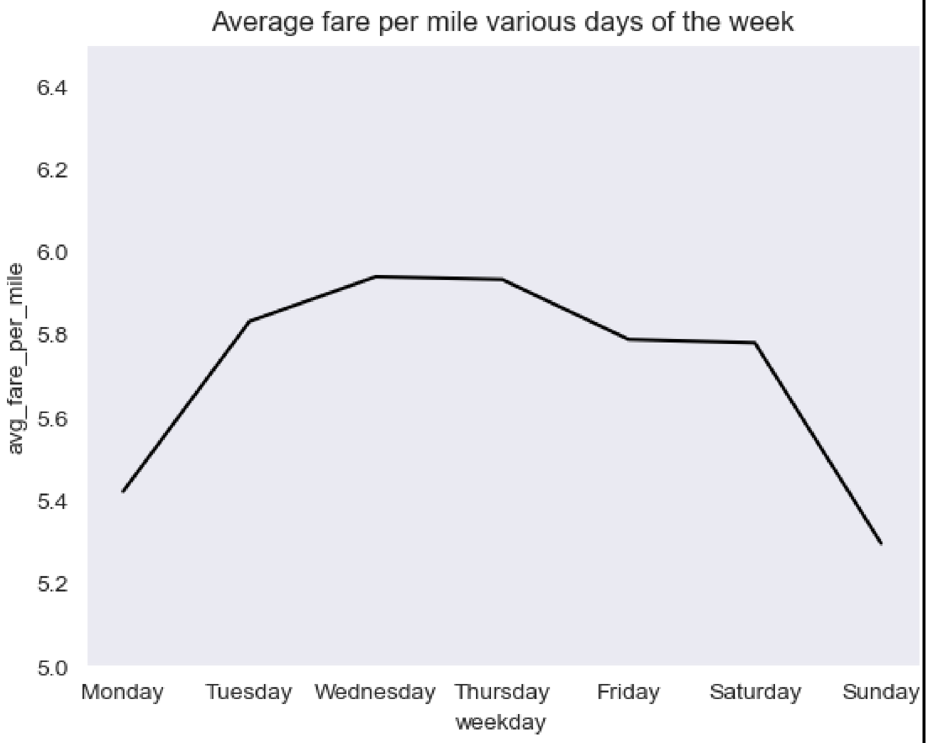


Figure 23

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

On average the average fare per mile is similar, the ups and downs are similar to the hourly variation of average fare per mile.

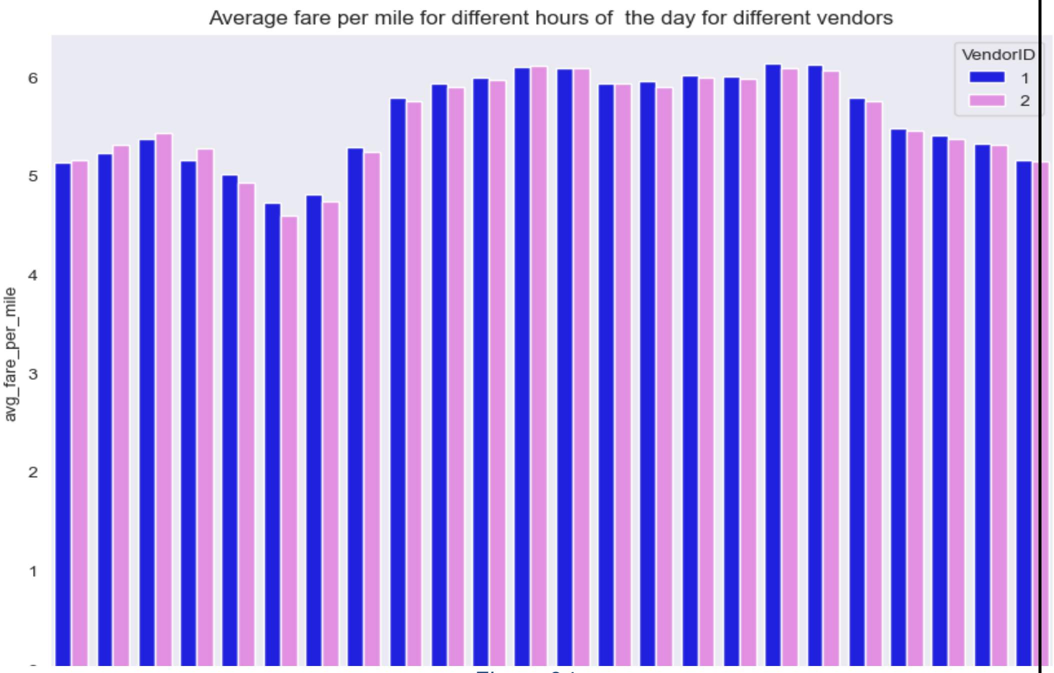


Figure 24

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

This is expected behavior. Short trips generate more revenue as their average fare per mile is more than that of long trips.

But the fare rate of vendor 2 is consistently high than that of vendor 1.

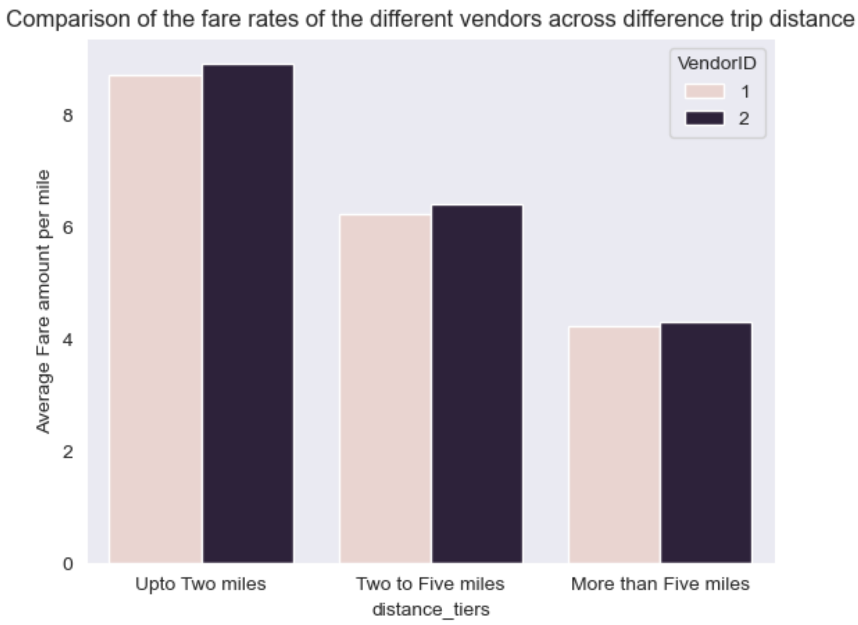


Figure 25

3.2.13. Analyse the tip percentages

This is a surprisingly unexpected. The tip percentage is decreasing as the trip distance gets longer. Maybe passengers get irritated for long trip duration or congested topic etc.

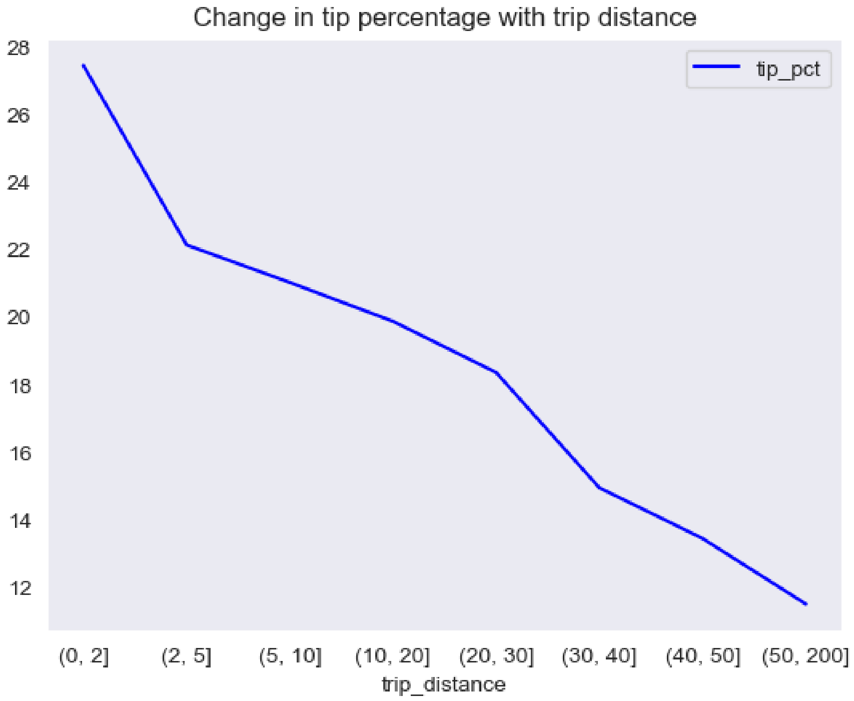


Figure 26

This looks totally random, and the difference is very small. I can’t conclude any trends from this plot.

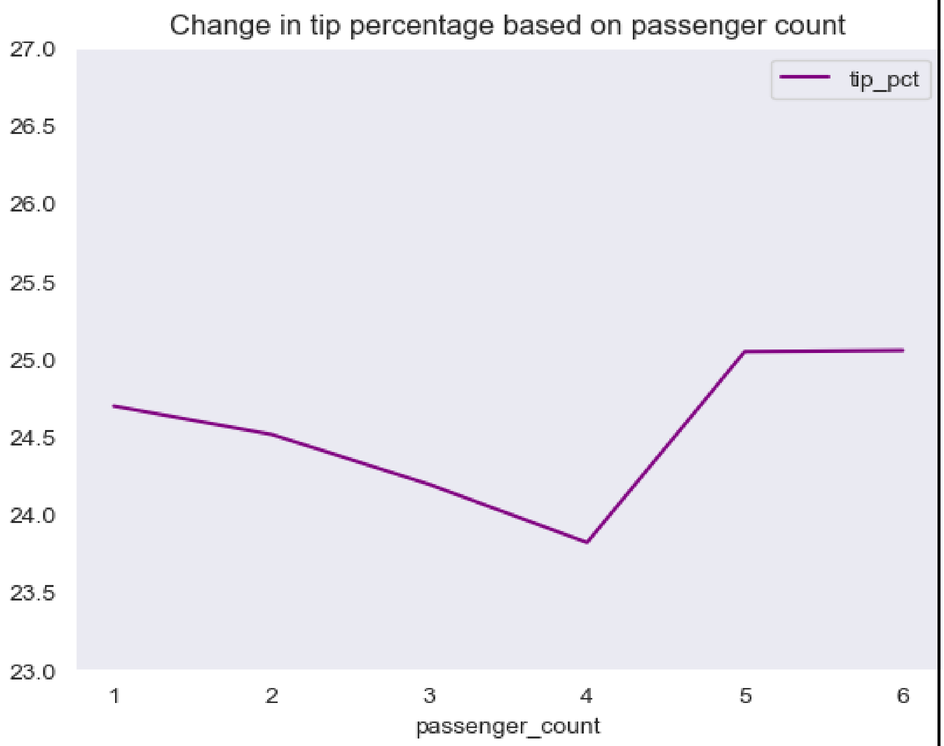


Figure 27

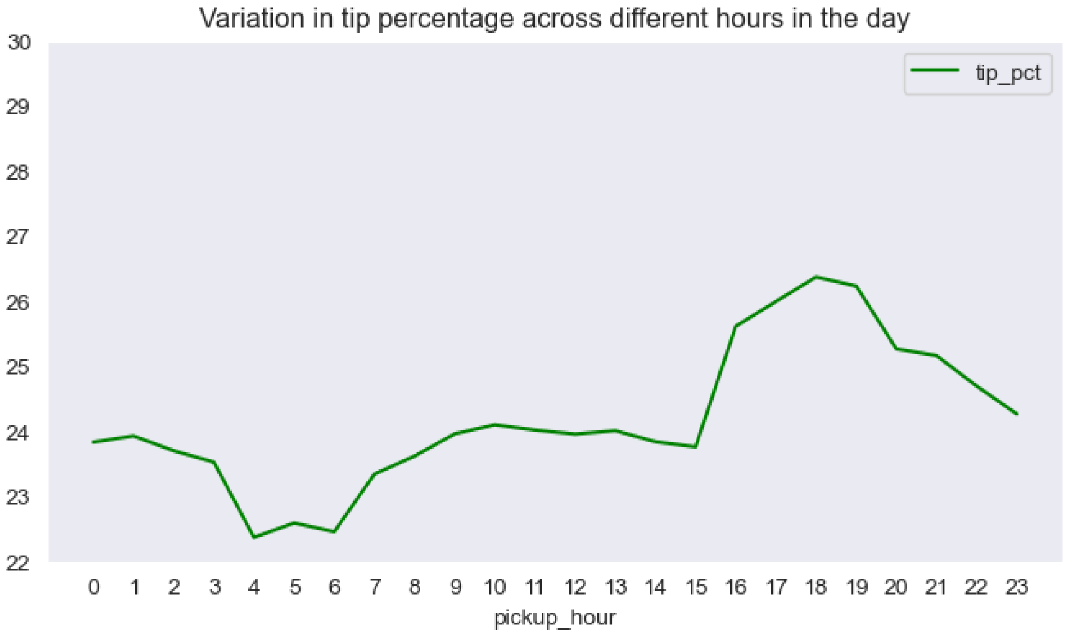


Figure 28

In late night and early morning, the tip percentage is considerably lower than during peak hours.

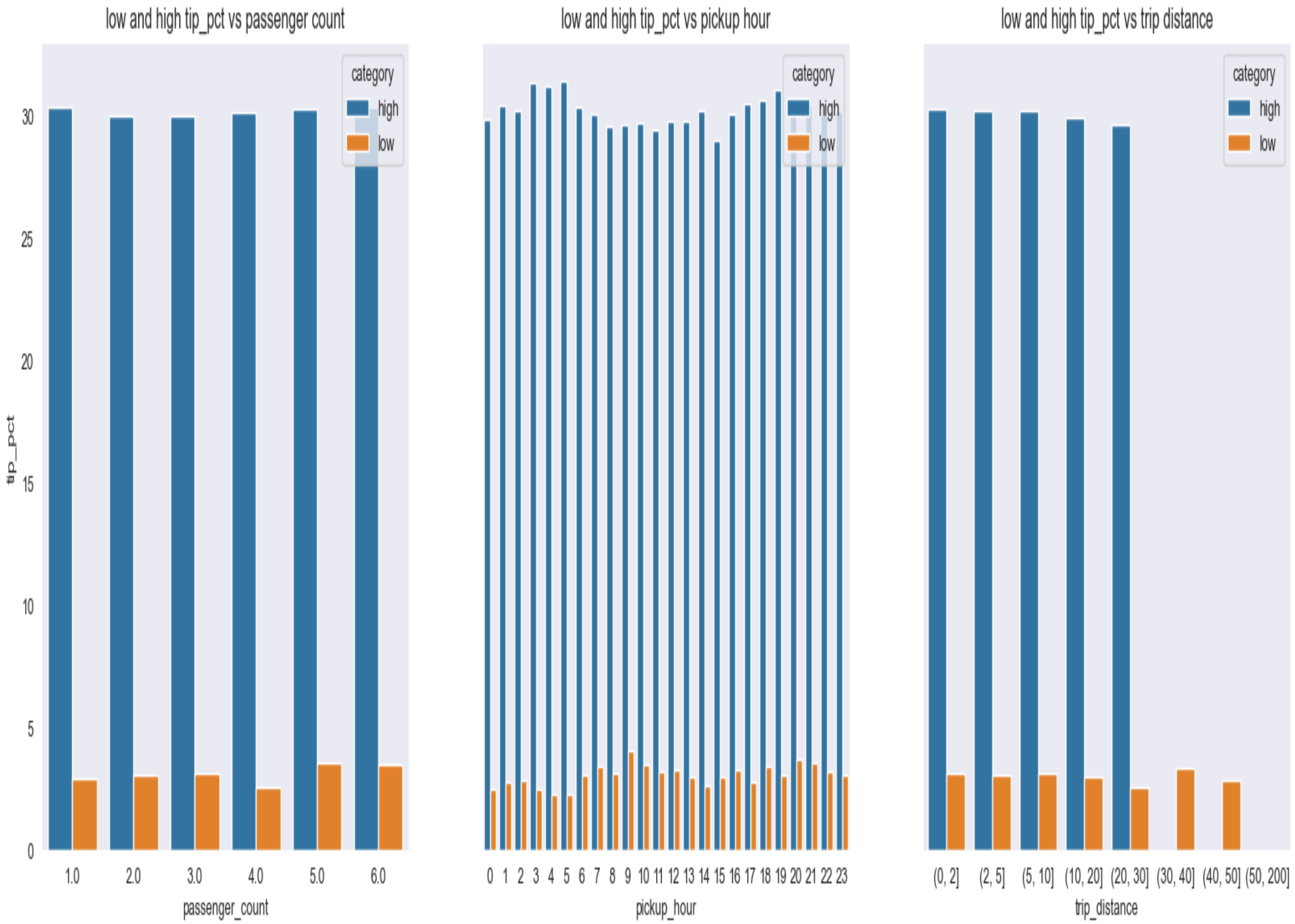


Figure 29

This is variation in lower (less than 10) and higher (more than 25) tip percentages across different categories. We can see here also, there is no higher tip percentage bar in long distance tier. That explains the dip in tip percentage with the increase of distance.

3.2.14. Analyse the trends in passenger count

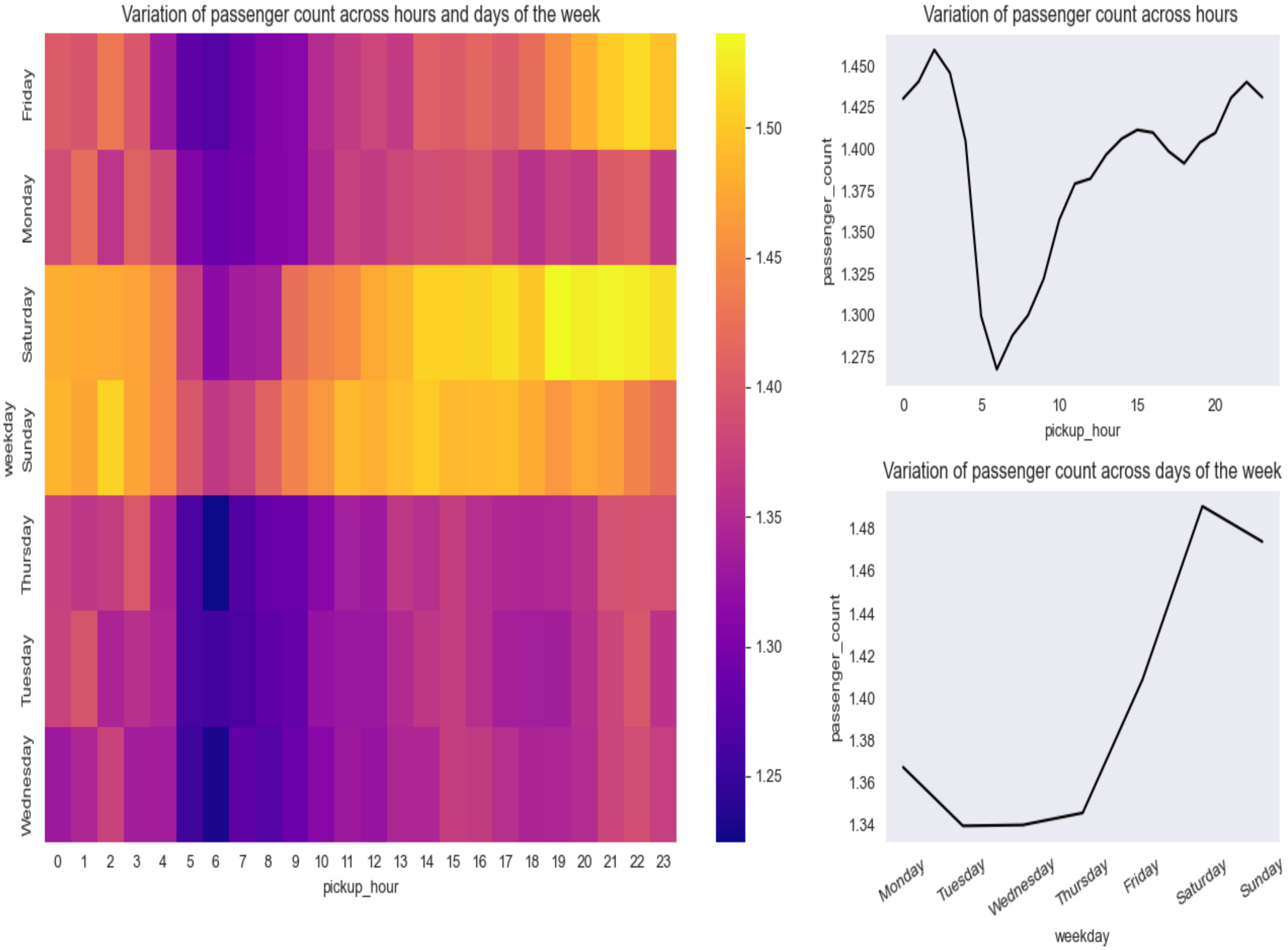


Figure 30

We can conclude the fact that people mostly travel alone in early mornings they travel in groups during evenings and weekends

3.3. Analyse the variation of passenger counts across zones

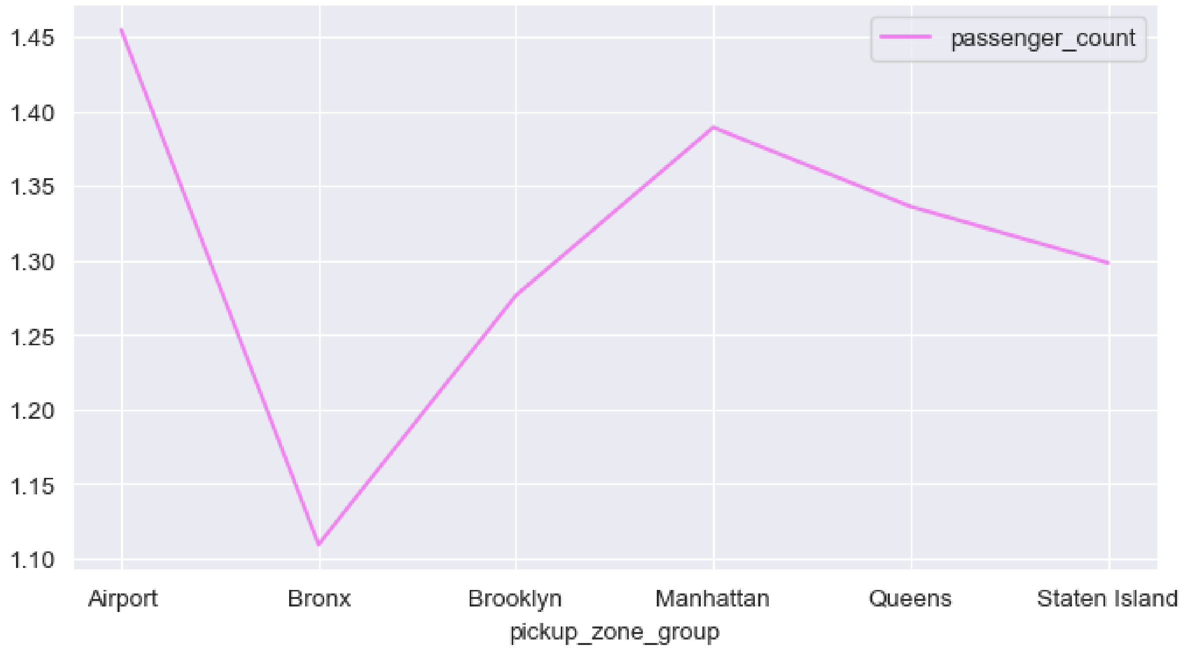


Figure 31

As we've seen before, the maximum trips happen at airport and Manhattan, and here we can confirm that fact. The average passenger count is also the highest in these areas.

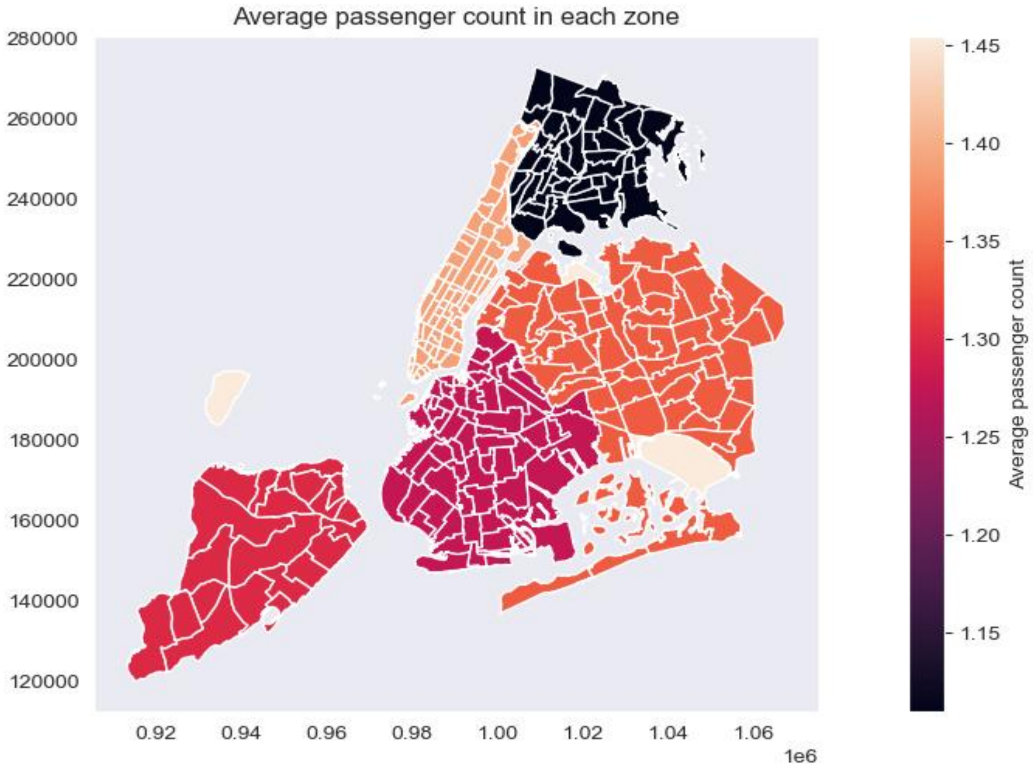


Figure 32

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

First, I take all the surcharges and extra charges columns to count as extra charges. But when I noticed the value counts of them, mta_tax and improvement_surcharge are mostly constant across the rows. So there is not much of a variation. That's why I excluded them from my analysis.

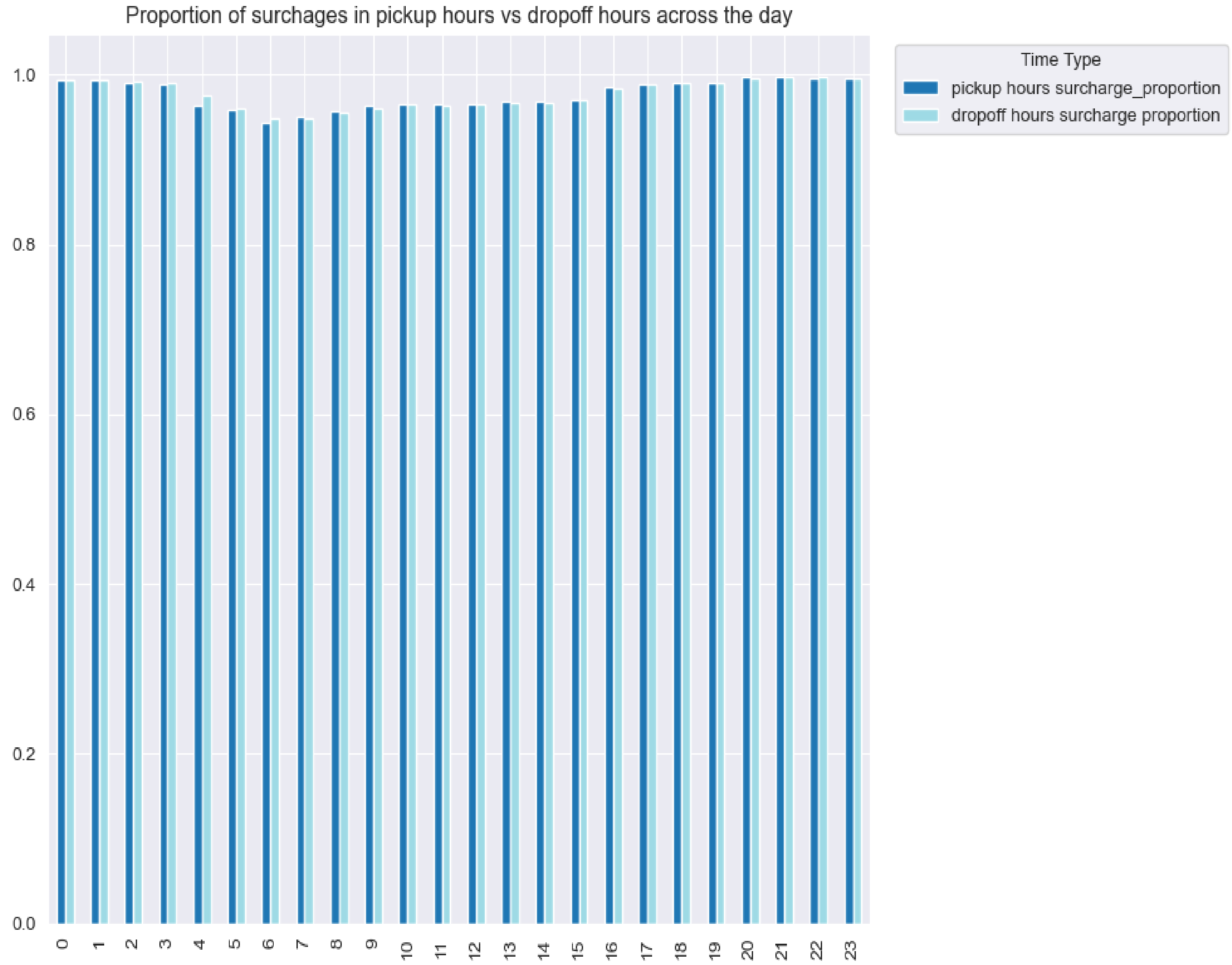


Figure 33

Here we can see that, late night has significantly more surcharge proportion than that of other times of the day and pickup and drop off surcharge proportion is almost similar in every hour except few ones.

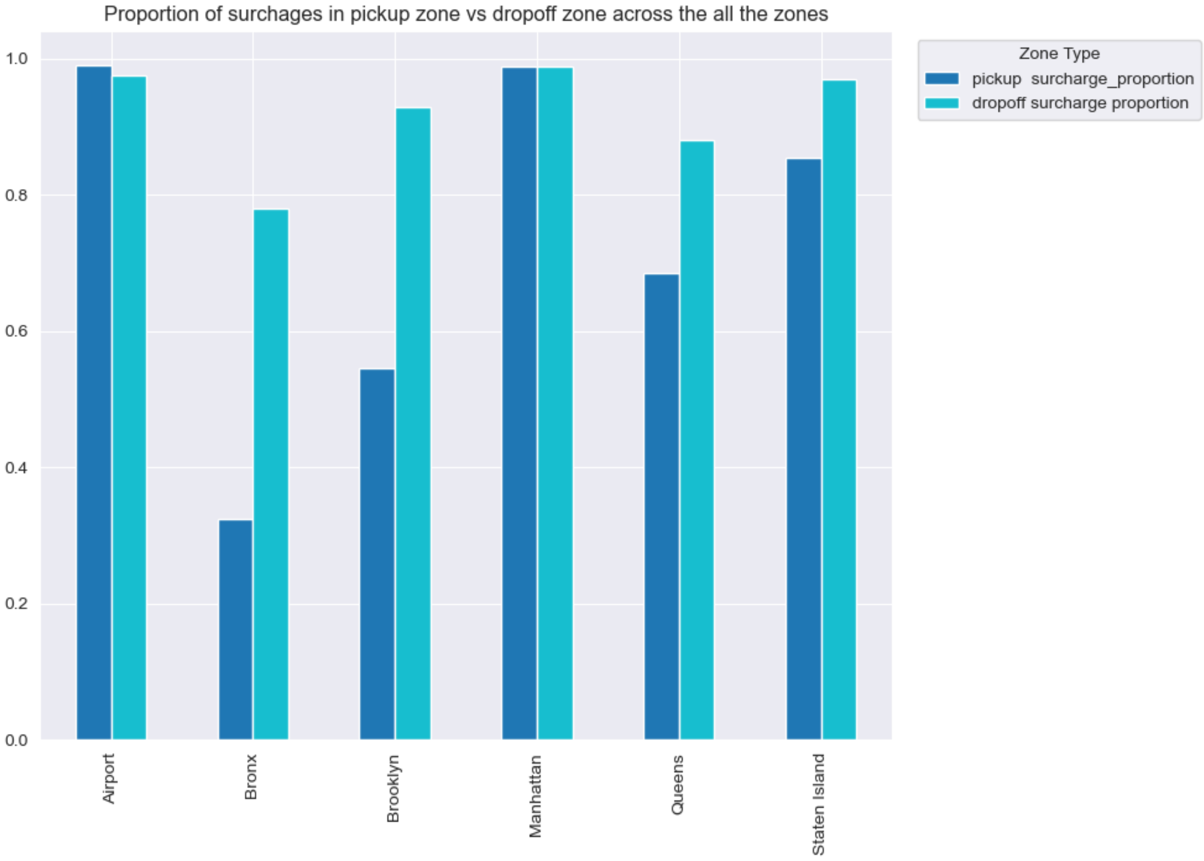


Figure 34

This is a very surprising finding. Except Airport and Manhattan, rest of the places have massive difference between pickup and drop off surcharge proportion. That means in these places drivers can't find that many passengers after drop off.

4. Conclusions

4.1. Final Insights and Recommendations

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

- Routing should consider destination also, not only pickup especially trips going towards Airport and Manhattan, as these trips should expect congestion and surcharges.
- During peak daytime hours, avoid congested routes as possible.
- During early morning hours, fewer cabs are needed due to low demand.

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

- More cabs should be placed in outer boroughs during morning and daytime as many trips start here.
- Maintain good cab availability in Airport and Manhattan zones, as many trips end there and these two places have the majority number of trips.
- Reduce supply during very early morning hours.
- Increase supply during evening and weekends, when passenger count is high.

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

- Pricing should be zone aware: Higher and unavoidable surcharges in Manhattan and airports should be reflected
- Avoid aggressive pricing for long-distance trips, as tip percentage is lower.
- Daytime pricing is the most important, as most of the revenue comes from daytime.
- Keep pricing competitive in outer borough areas to attract demand.
- Short trips have very high tip percentage, so the fare for short trips can be increased slightly.
- There could be a cashback feature for credit card payments to lessen the cash payments, then the total tips amount will increase which will lead more revenue.