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

To handle the large size of each file, we sampled 0.7% of all rows per file to keep the total data entries between 250,000 and 300,000, as suggested. We used the combination of date and hour as the sampling base to help ensure precision and relevance in time-sensitive data.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

We reset and cleaned the index to avoid misalignment from concatenation.

2.1.2. Combine the two airport_fee columns

We confirmed that the two columns have no conflicting values. The airport_fee column was kept as its naming convention is more in line with other columns, and we filled the missing values using Airport_fee.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

```
    # Find the proportion of missing value
    df.isna().mean() * 100
    # We have the exact same amount of mis
05]  ✓ 0.0s

VendorID                0.000000
tpep_pickup_datetime    0.000000
tpep_dropoff_datetime   0.000000
passenger_count         3.432780
trip_distance           0.000000
RatecodeID              3.432780
PULocationID            0.000000
DOLocationID            0.000000
payment_type            0.000000
fare_amount             0.000000
extra                   0.000000
tip_amount              0.000000
total_amount            0.000000
congestion_surcharge    3.432780
airport_fee             3.432780
dtype: float64
```

2.2.2. Handling missing values in passenger_count

As the passenger_count column has discrete values, we will use the median to fill in the missing and 0 values (impossible to have 0 passengers)

2.2.3. Handle missing values in RatecodeID

As the RatecodeID has discrete values, we will use the median to fill in the nulls and 99 values (99 equals to Null/unknown)

2.2.4. Impute NaN in congestion_surcharge

As the congestion_surcharge has discrete values, we will use the median to fill in the missing values

2.3. Handling Outliers and Standardising Values

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

There are a lot of extreme values for fare_amount, so we cut out the most extreme 1% of values for fare_amount in both tails.

The same goes for trip_distance, but we'll keep the trip_distance = 0 data, as there can be cases where the pick up and drop off occurred in the same zone

There are a lot of trips with abnormally long durations. We assume those are cases where the driver forgot to turn off the meter and will disregard them for the duration analysis

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 (datetime)

`tpep_dropoff_datetime`: Numerical (datetime)

`passenger_count`: Numerical

`trip_distance`: Numerical

`RatecodeID`: Categorical

`PULocationID`: Categorical

`DOLocationID`: Categorical

`payment_type`: Categorical

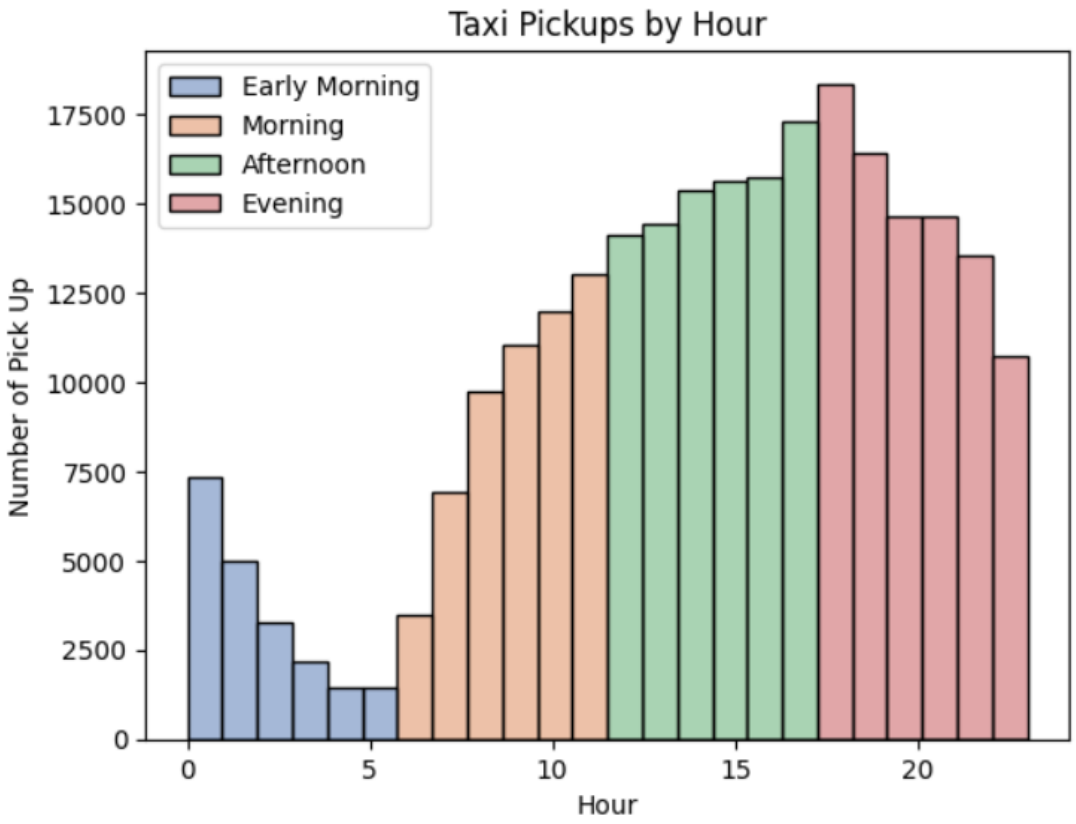
`pickup_hour`: Numerical

`trip_duration`: Numerical

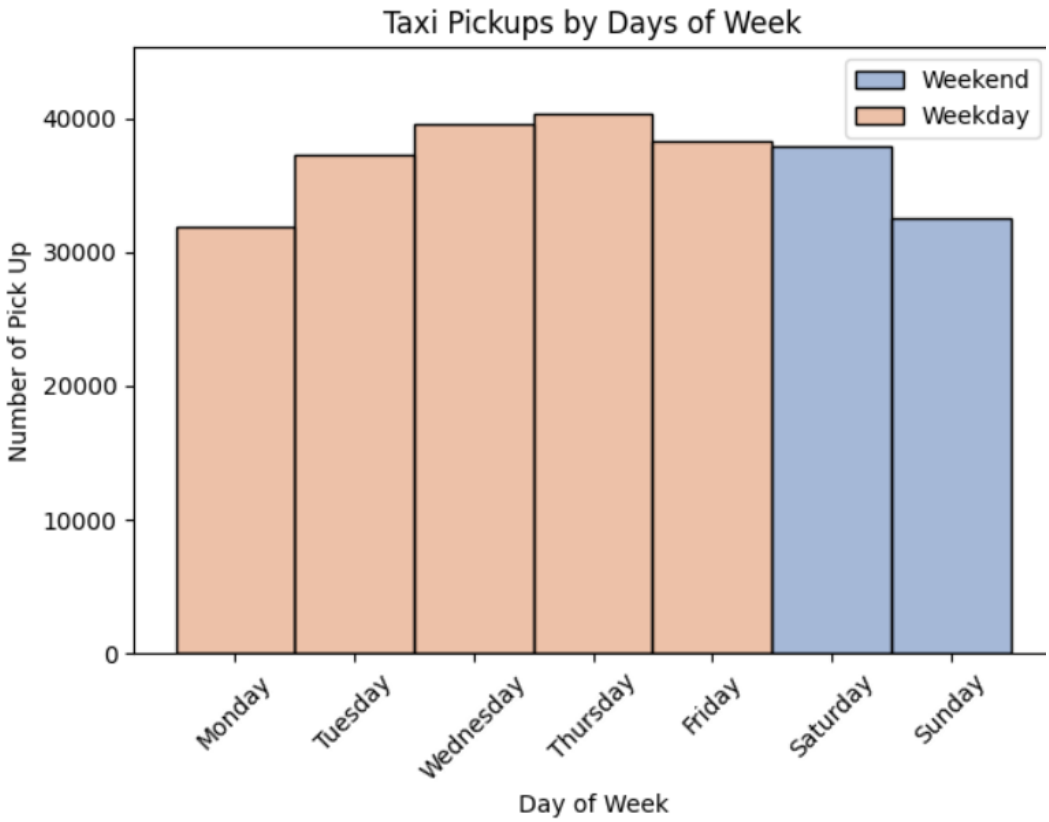
The following monetary parameters belong in the same category, is it categorical or numerical?

Answer: They are numerical

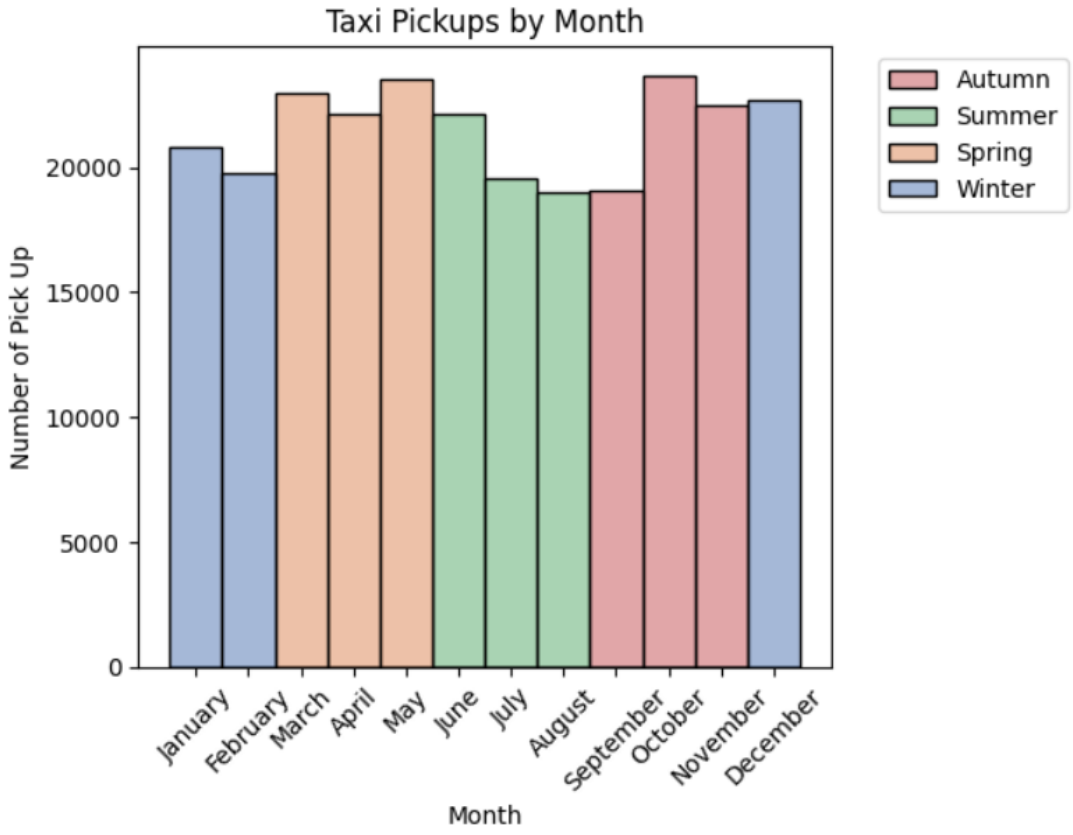
3.1.2. Analyse the distribution of taxi pickups by hours, days of the week, and months



Demand starts throughout the working hours and peaks at 5–7 PM.



Demands peak softly on Wednesday and Thursday, suggesting New Yorkers mostly use taxis to get to work.



Taxi usage is at its highest in spring, decreases gradually during summer, then peaks again in September. This can be explained due to spring being a favorable season for tourism, and September being the back-to-work and back-to-school month.

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

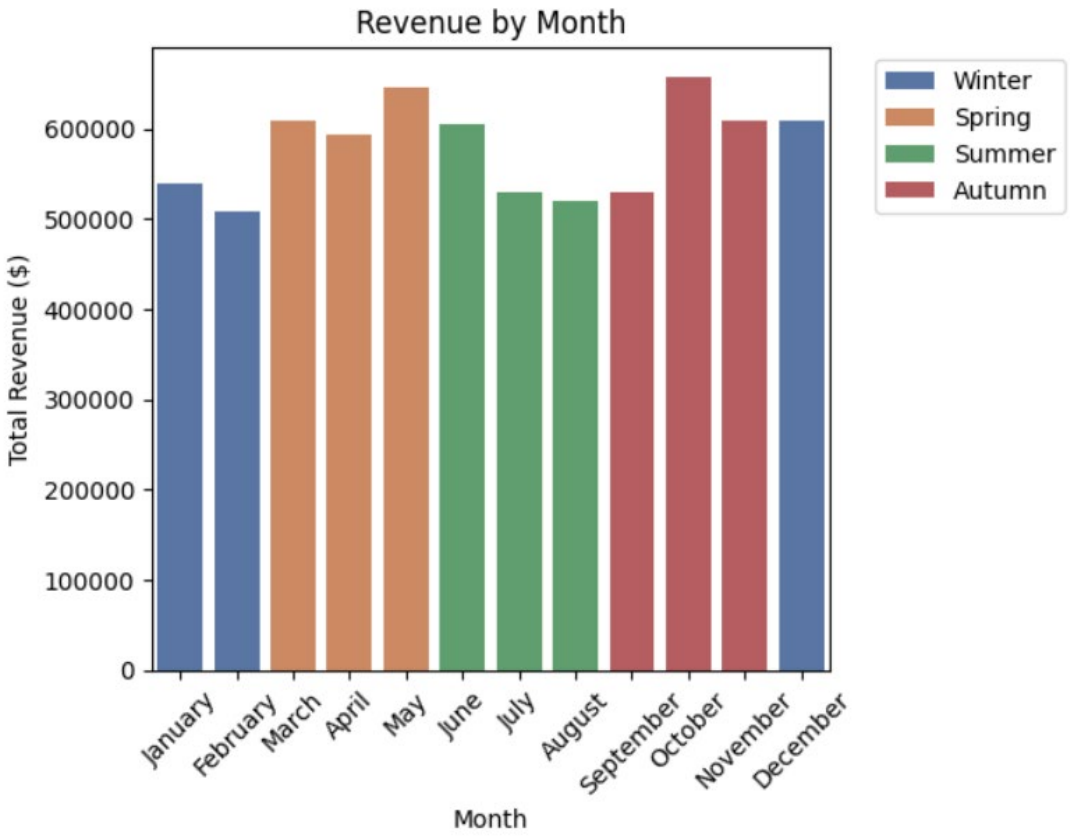
All the negative values have been dealt with by replacing them with the absolute values

Zero values for fare_amount have been considered outliers and were discarded.

tip_amount is almost always 0 for cash payment. We can discard those values when we do tip analysis.

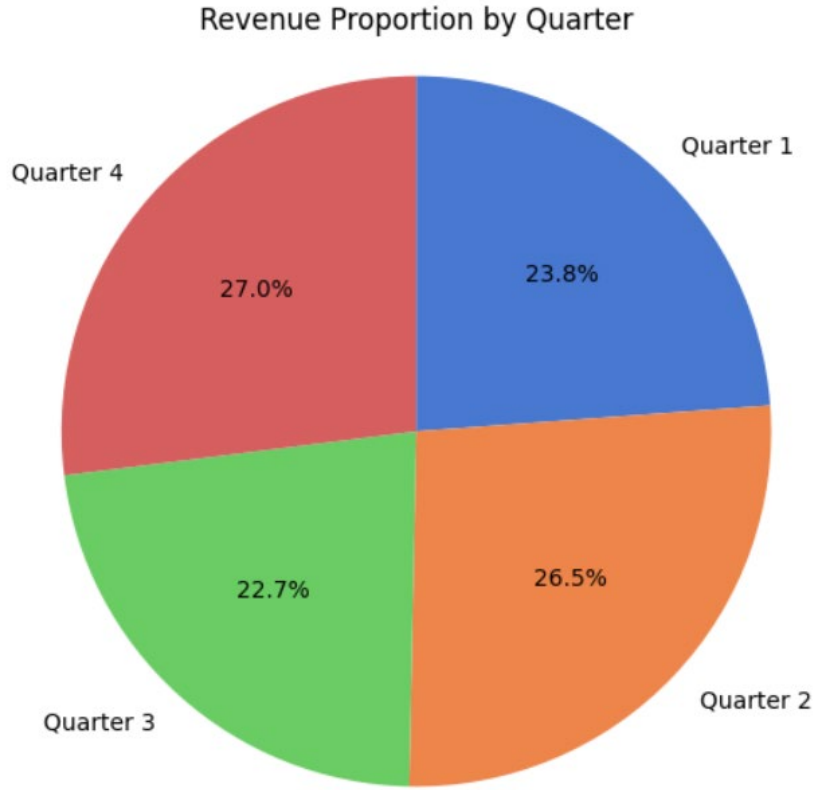
trip_distance equal to 0 are not impossible and can be kept as is. We can discard them for the trip distance-related analysis

3.1.4. Analyse the monthly revenue trends

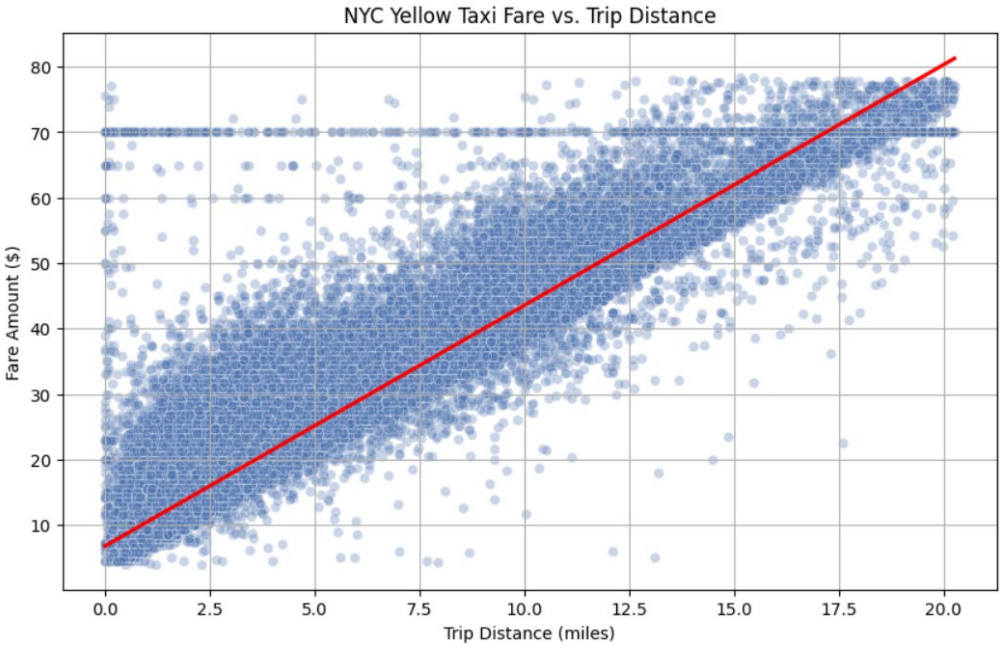


Monthly revenue peaks in spring due to pleasant weather, increased local activity, and early tourism. It then declines during the hot summer as locals travel and outdoor movement drops. Revenue rises again in September with the return of commuters, students, and business travelers.

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



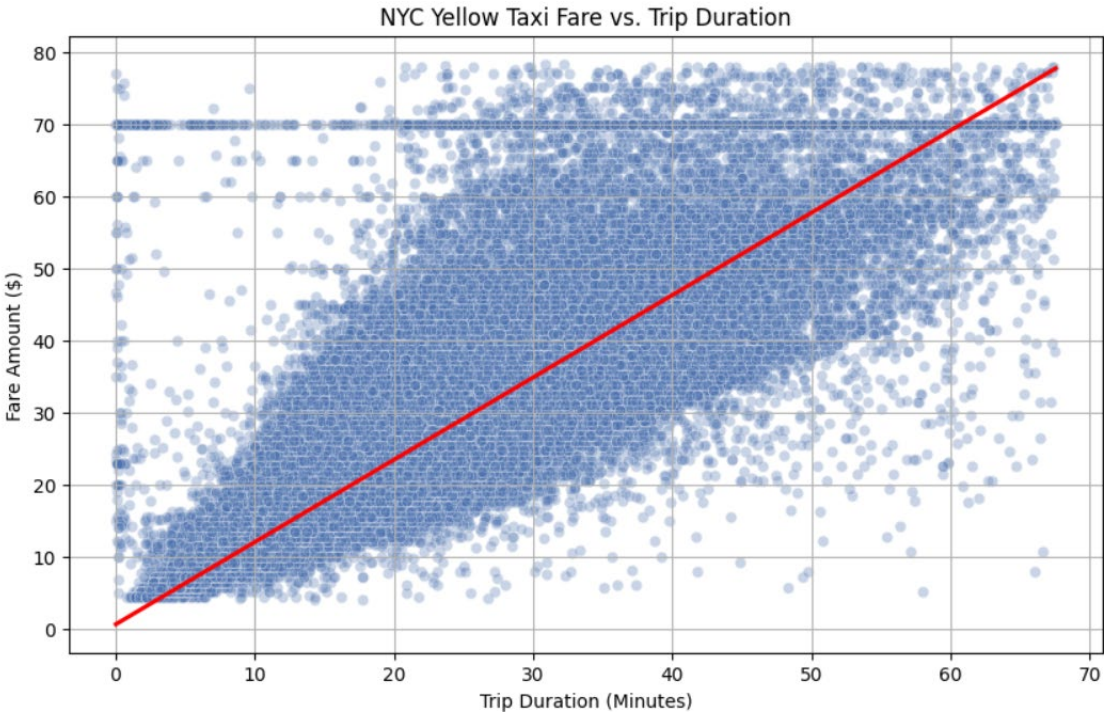
3.1.6. Analyse and visualise the relationship between distance and fare amount



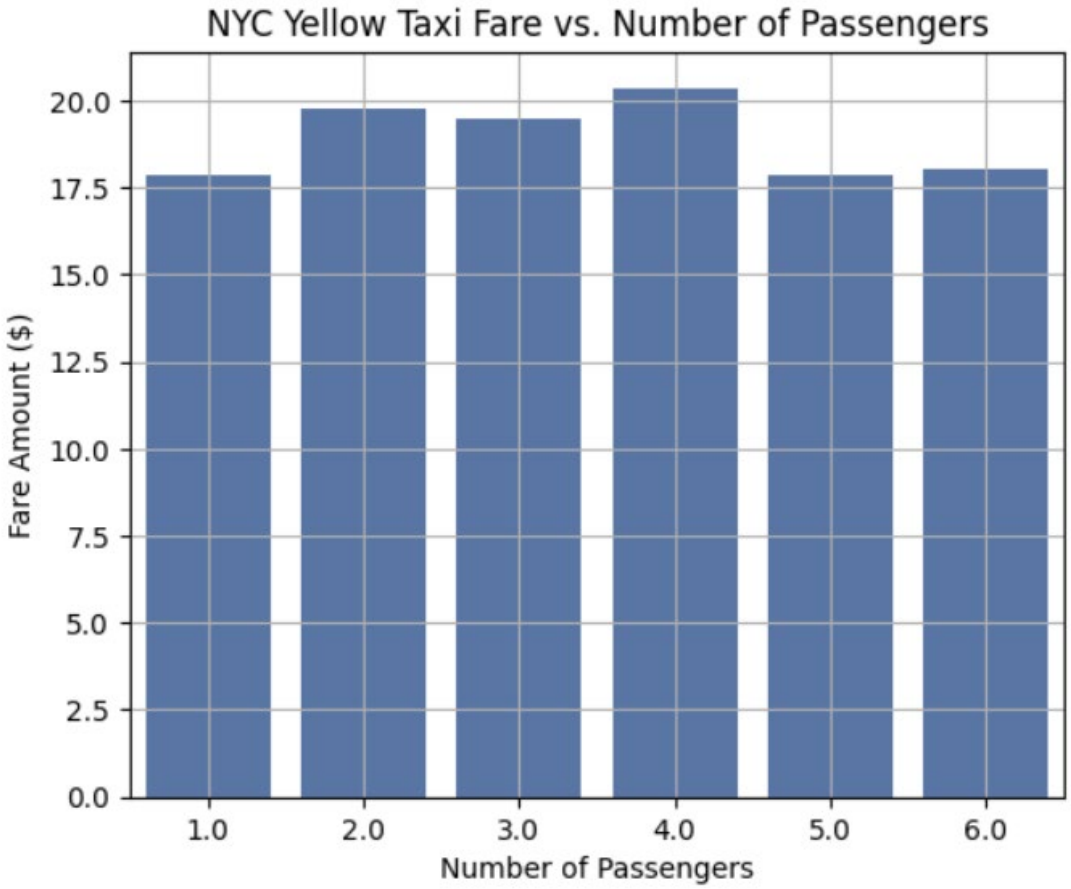
Fare amount generally increases with trip distance, showing a strong

positive linear relationship. Short trips have more variability due to minimum fare rules and flat fees, while longer trips show a more consistent fare-per-mile rate. Outliers often reflect airport flat fares or tolls.

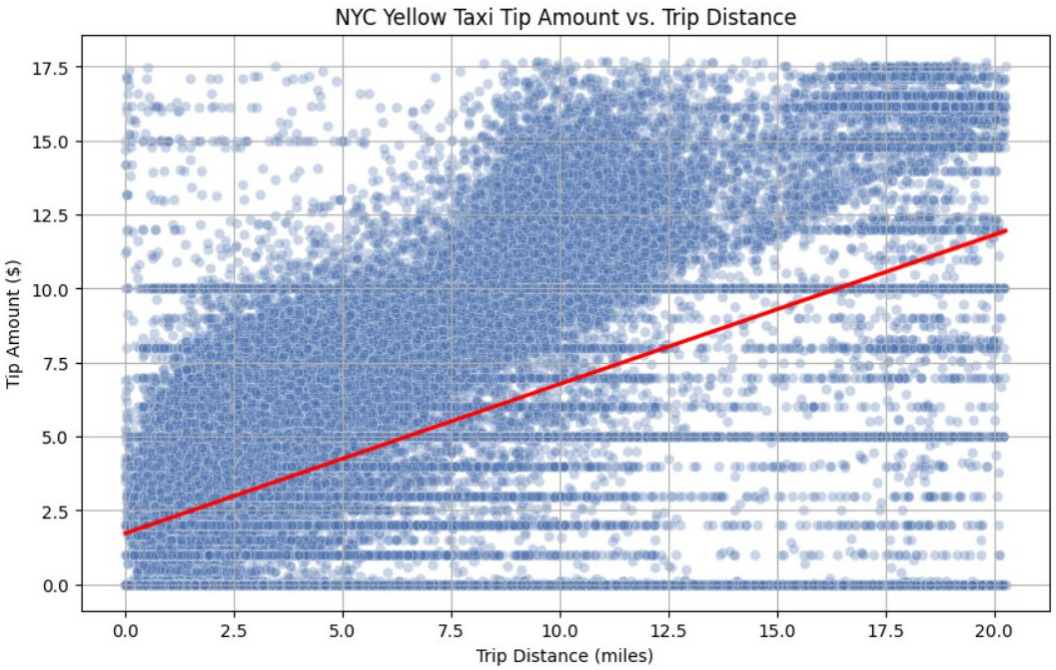
3.1.7. Analyse the relationship between fare/trip duration, fare/passengers, and tip/trip distance



Fare vs. Trip Duration: There's a positive correlation — longer durations lead to higher fares, with some variability due to traffic delays and waiting charges.



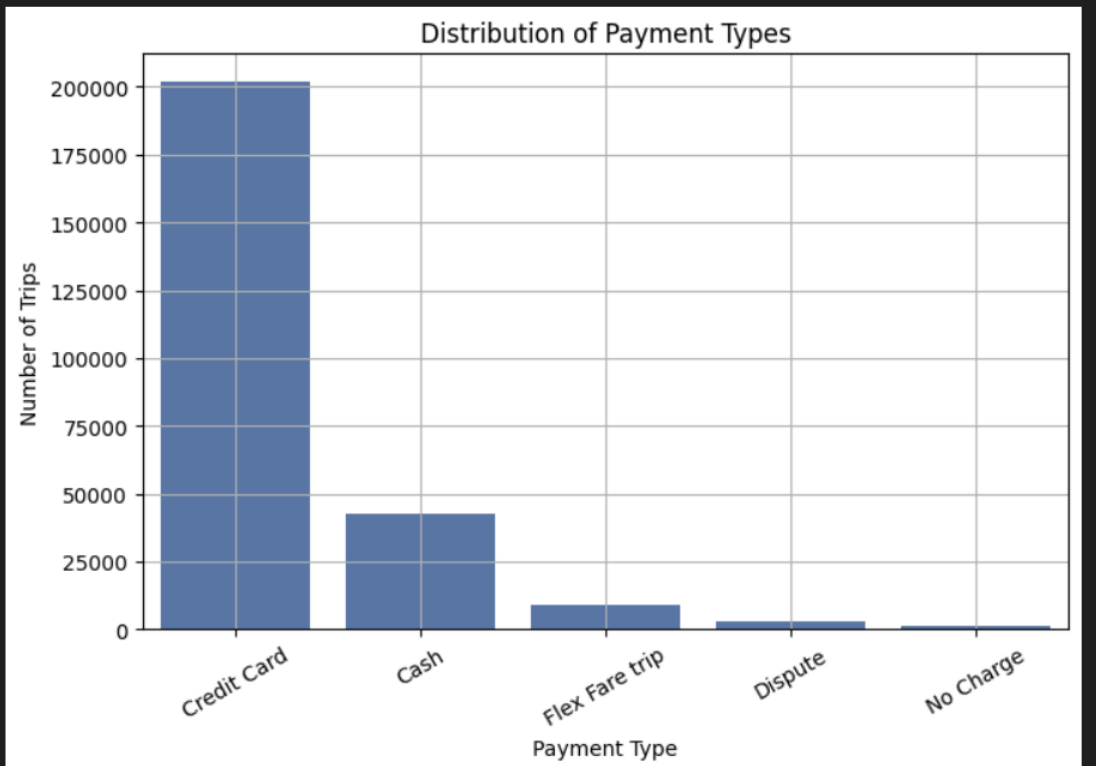
Fare vs. Passengers: No correlation. As passenger count increases, the average fare amount doesn't necessarily increase as well, meaning the fare per person decreases as the number of people sharing the ride increases.



Tip vs. Trip Distance: Tips generally increase with distance, as they are often a proportion of the fare amount. The longer the distance is, the higher the fare amount will be.

3.1.8. Analyse the distribution of different payment types

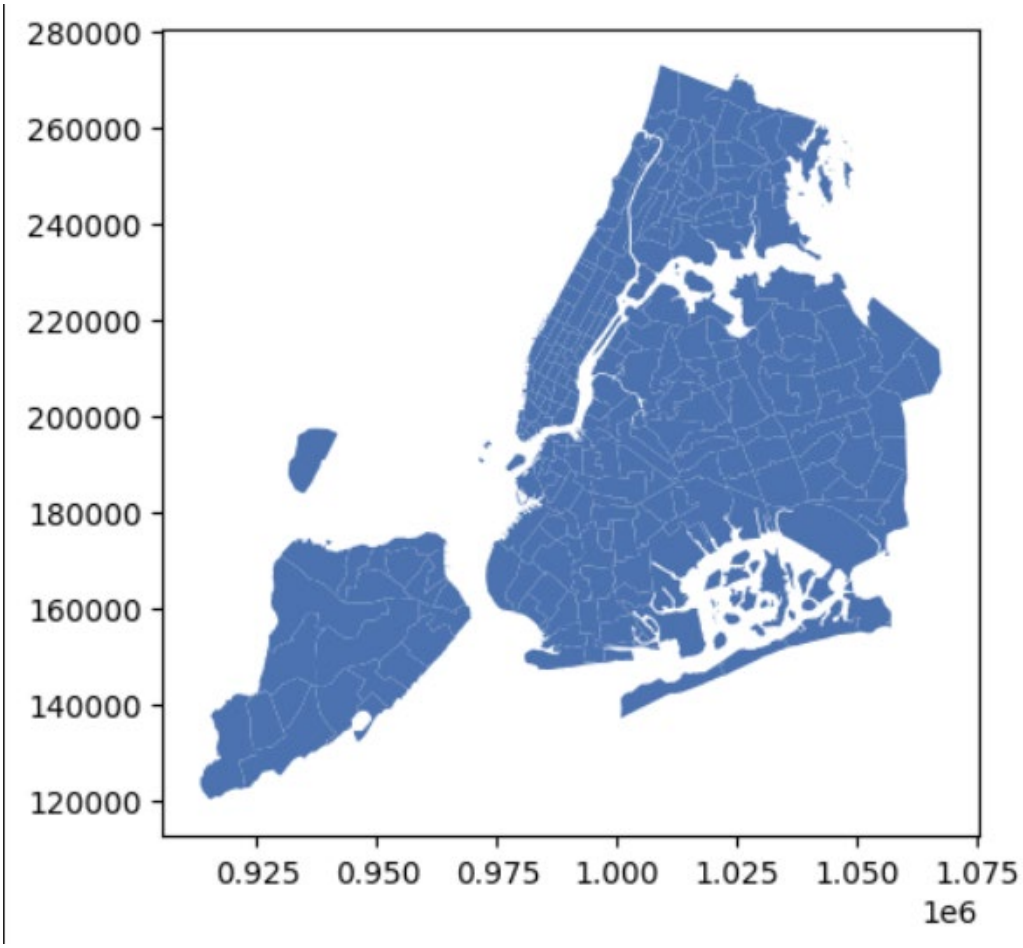
```
payment_label
Credit Card      78.400000
Cash             16.500000
Flex Fare trip    3.500000
Dispute           1.100000
No Charge         0.500000
Name: proportion, dtype: float64
```



The majority of taxi payments are made by credit card (78.4%), reflecting the convenience and popularity of cashless transactions. Flex Fare trips (3.5%) likely involve app-based pricing or promotions.

Disputes (1.1%) and No Charge trips (0.5%) are rare, indicating overall payment reliability and low incidence of billing issues.

3.1.9. Load the taxi zones shapefile and display it



3.1.10. Merge the zone data with trips data

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

print(zoned_df[zoned_df['LocationID'].isna()]['PULocationID'].unique()) #
We don't have zone information for LocationID 264 and 265, thus those
will be left out in the analysis

zoned_df.head()
```

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

```
location_count =
zoned_df.groupby('LocationID')['LocationID'].count().rename('trip_count').r
eset_index()

location_count
```

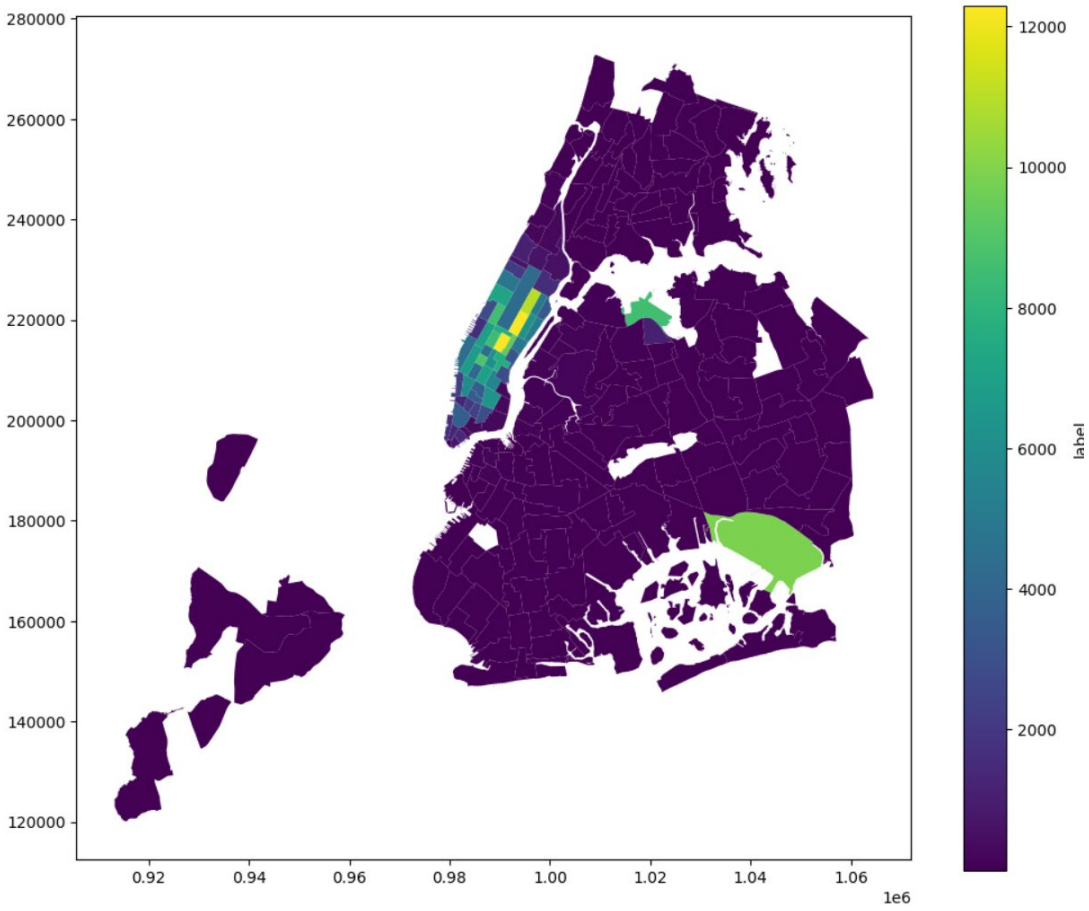
3.1.12. Add the number of trips for each zone to the zones dataframe

```
zone_count = zones.merge(location_count, how='left', on='LocationID')

print(zone_count[zone_count['trip_count'].isna()]['LocationID'].unique())

zone_count
```

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



The zone map reveals that Manhattan dominates taxi activity, with the highest number of trips concentrated in Midtown, Times Square, and the Financial District. Outer boroughs like Queens, Brooklyn, and the Bronx show significantly lower activity, except near LaGuardia and John F. Kennedy airports.

This suggests that taxi services are highly centralized, catering mostly to business, tourism, and dense residential zones, highlighting potential opportunities for fleet redistribution or targeted service expansion in underserved areas.

3.1.14. Conclude with results

Busiest Times: Peak hours are 5–7 PM; busiest days are Wednesday and Thursday.

Seasonal Trends: Demand and revenue peak in spring and September, dip during summer.

Revenue: Q2 and Q4 contribute the most to yearly earnings.

Fare Drivers: Fare increases with distance and duration; more passengers reduce fare per person.

Tips: Higher on longer trips.

Hot Zones: Midtown and Financial District have the most trips; airports also see high activity.

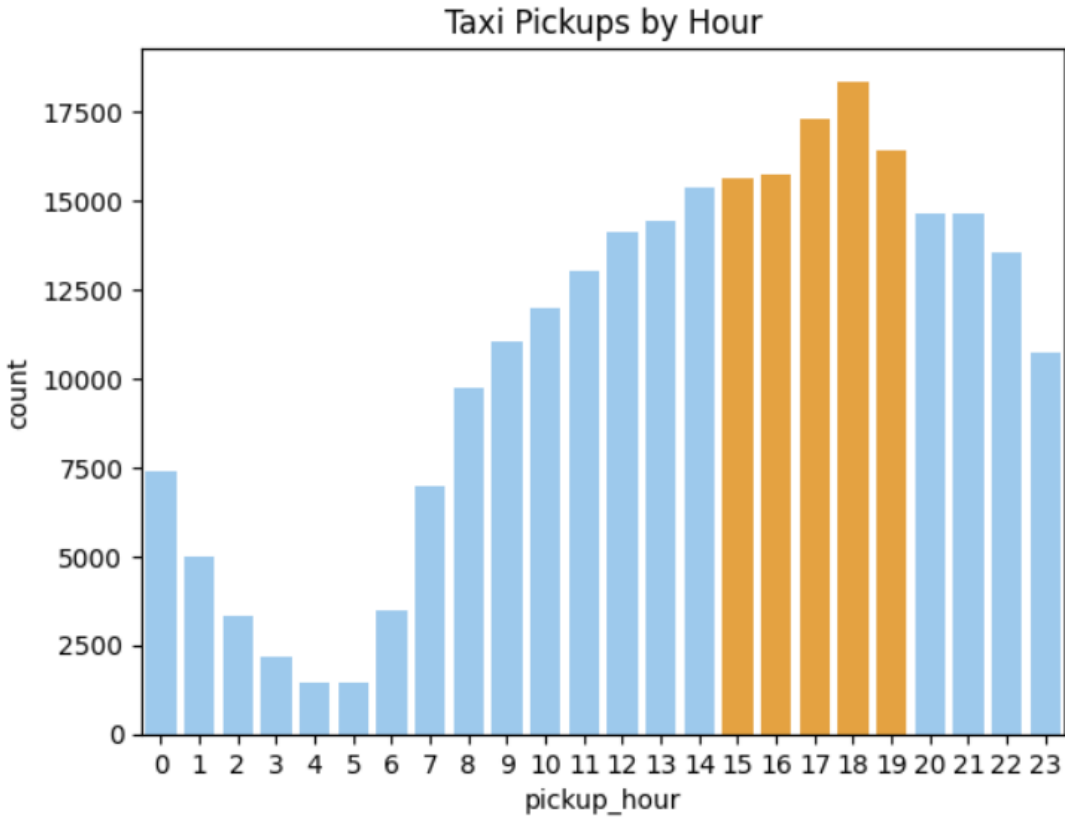
Insight: Taxi services are centralized in Manhattan, with growth potential in outer boroughs.

3.2. Detailed EDA: Insights and Strategies

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

	route	pickup_hour	miles_per_hour	trip_count	average_speed	speed_difference	slower_from_average
0	Lower East Side - Lower East Side	0	5.495298	11	14.128728	8.633429	0.611055
1	Greenwich Village South - Lower East Side	1	6.979724	13	14.003382	7.023659	0.501569
2	Lower East Side - Lower East Side	2	5.709006	8	14.448585	8.739579	0.604874
3	Lower East Side - Lower East Side	3	7.392682	10	16.259046	8.866363	0.545319
4	East Village - East Village	4	7.452322	8	17.560982	10.108660	0.575632
5	Garment District - Midtown Center	5	10.611321	22	19.111822	8.500501	0.444777
6	Clinton East - Times Sq/Theatre District	6	7.989393	14	16.095760	8.106367	0.503634
7	Garment District - Midtown Center	7	6.690575	29	12.982792	6.292217	0.484658
8	West Chelsea/Hudson Yards - Midtown South	8	5.339175	8	10.989847	5.650672	0.514172
9	Penn Station/Madison Sq West - Clinton East	9	4.367105	10	10.650390	6.283285	0.589958
10	Penn Station/Madison Sq West - Times Sq/Theatr...	10	4.001085	21	10.381412	6.380327	0.614591
11	Penn Station/Madison Sq West - Times Sq/Theatr...	11	4.084870	49	9.995541	5.910672	0.591331
12	Penn Station/Madison Sq West - Garment District	12	3.365201	18	9.644675	6.279474	0.651082
13	Garment District - Garment District	13	1.996031	8	9.918556	7.922525	0.798758
14	Times Sq/Theatre District - Times Sq/Theatre D...	14	3.635147	8	9.485315	5.850168	0.616761
15	Midtown South - Midtown South	15	2.883374	10	9.185108	6.301734	0.686082
16	Times Sq/Theatre District - Times Sq/Theatre D...	16	3.214935	29	9.388705	6.173770	0.657574
17	Times Sq/Theatre District - Times Sq/Theatre D...	17	2.861301	21	9.790050	6.928750	0.707734
18	Times Sq/Theatre District - Times Sq/Theatre D...	18	3.745168	32	10.834261	7.089094	0.654322
19	Times Sq/Theatre District - Times Sq/Theatre D...	19	4.416773	20	10.848389	6.431616	0.592864
20	Times Sq/Theatre District - Times Sq/Theatre D...	20	4.017856	12	11.841833	7.823976	0.660707
21	Times Sq/Theatre District - Times Sq/Theatre D...	21	4.665405	21	12.438052	7.772648	0.624909
22	Clinton East - Times Sq/Theatre District	22	5.714998	9	12.719416	7.004418	0.550687
23	Times Sq/Theatre District - Times Sq/Theatre D...	23	5.420769	10	13.727763	8.306994	0.605124

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



The 5 busiest hours are from 3 PM to 7 PM.

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

3.2.3 [2 mark]
Find the actual number of trips in the five busiest hours

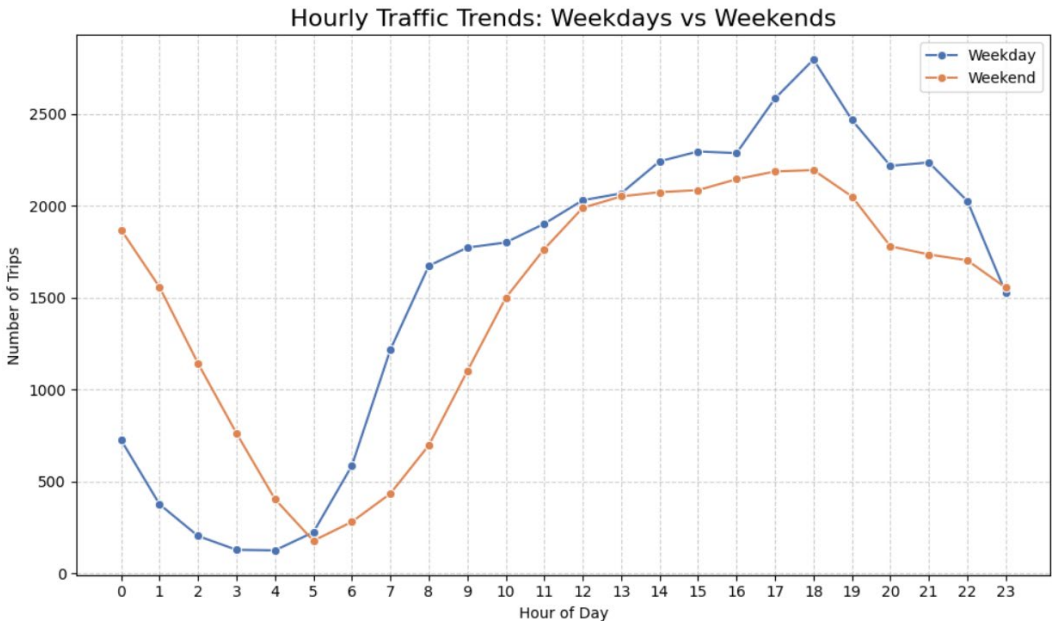
```
# Scale up the number of trips

# Fill in the value of your sampling fraction and use that to scale up the numbers
sample_fraction = 0.007
for i in range(len(busiest_hours)):
    print(f'Estimation of the number of taxi trips at the top {i + 1} busiest hour ({busiest_hours[i]}) is {round(busiest_hours_count[i] / 0.007)}')
```

[239] ✓ 0.0s

```
Estimation of the number of taxi trips at the top 1 busiest hour (18) is 2622571
Estimation of the number of taxi trips at the top 2 busiest hour (17) is 2469286
Estimation of the number of taxi trips at the top 3 busiest hour (19) is 2346857
Estimation of the number of taxi trips at the top 4 busiest hour (16) is 2245000
Estimation of the number of taxi trips at the top 5 busiest hour (15) is 2234286
```

3.2.4. Compare hourly traffic on weekdays and weekends



Taxi demand on weekdays is consistently higher during work hours, while weekend demand is much higher after midnight until 5 AM, driven by nightlife.

Knowing busy and quiet hours helps optimize driver shifts, reduce wait times, and boost revenue.

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

	zone	pickup_count	pickup_rank
0	Midtown Center	12298	1
1	Upper East Side South	12261	2
2	Upper East Side North	11031	3
3	JFK Airport	9850	4
4	Midtown East	9457	5
5	Penn Station/Madison Sq West	8847	6
6	LaGuardia Airport	8574	7
7	Lincoln Square East	8560	8
8	Times Sq/Theatre District	8474	9
9	Murray Hill	7688	10

	zone	dropoff_count	dropoff_rank
0	Upper East Side North	11477	1
1	Upper East Side South	10915	2
2	Midtown Center	10114	3
3	Times Sq/Theatre District	7845	4
4	Murray Hill	7684	5
5	Midtown East	7511	6
6	Lincoln Square East	7426	7
7	Upper West Side South	7109	8
8	Lenox Hill West	6899	9
9	East Chelsea	6594	10

Assume that we are taking the top 10 zones with the most pick up/drop off trips, and not the top 10 combination of zones and hours (which means one zone can appear multiple times)

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

Top 10 zones with the highest pickup/drop ratio

	zone	pickup_count	dropoff_count	pick/drop_ratio
0	East Elmhurst	1109	107	10.364486
1	JFK Airport	9850	2033	4.845057
2	LaGuardia Airport	8574	3289	2.606871
3	Penn Station/Madison Sq West	8847	5798	1.525871
4	Greenwich Village South	3531	2534	1.393449
5	Central Park	4289	3250	1.319692
6	West Village	5945	4535	1.310915
7	Midtown East	9457	7511	1.259087
8	Garment District	4370	3563	1.226495
9	Midtown Center	12298	10114	1.215938

Top 10 zones with the lowest pickup/drop ratio

	zone	pickup_count	dropoff_count	pick/drop_ratio
102	South Ozone Park	28	233	0.120172
103	South Williamsburg	10	104	0.096154
104	Greenpoint	56	591	0.094755
105	Inwood	13	145	0.089655
106	Bushwick South	25	280	0.089286
107	Bedford	25	293	0.085324
108	Washington Heights North	38	451	0.084257
109	Flushing	14	183	0.076503
110	Roosevelt Island	11	144	0.076389
111	Ridgewood	11	147	0.074830

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

	zone	night_pickup_count	night_pickup_rank
0	East Village	1905	1
1	West Village	1369	2
2	Lower East Side	1269	3
3	Clinton East	1232	4
4	Greenwich Village South	1038	5
5	JFK Airport	1001	6
6	Times Sq/Theatre District	989	7
7	Penn Station/Madison Sq West	938	8
8	Upper East Side South	825	9
9	East Chelsea	814	10

	zone	night_dropoff_count	night_dropoff_rank
0	East Village	986	1
1	Upper East Side North	917	2
2	Murray Hill	879	3
3	Clinton East	863	4
4	East Chelsea	788	5
5	Upper East Side South	780	6
6	Midtown Center	767	7
7	Gramercy	729	8
8	Times Sq/Theatre District	671	9
9	Lenox Hill West	665	10

3.2.8. Find the revenue share for nighttime and daytime hours

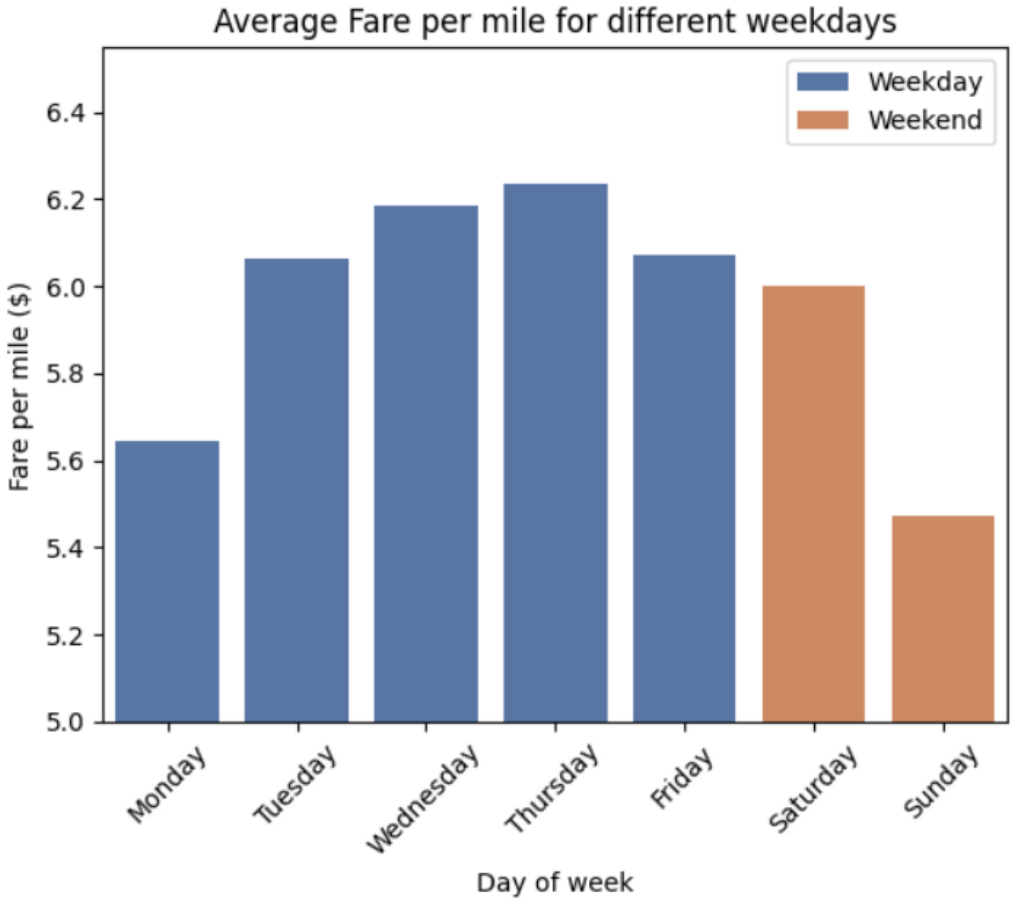
	night_or_day	revenue_share
0	day	89.211017
1	night	10.788983

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

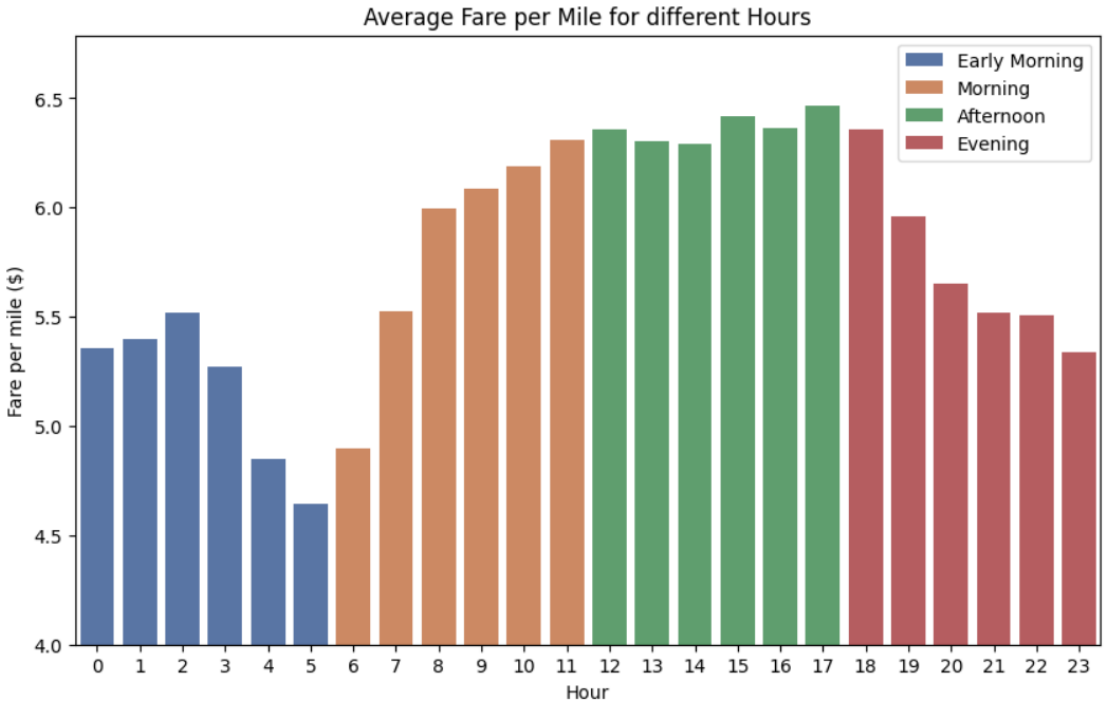
	passenger_count	fare_per_mile_per_passenger
0	1	6.015759
1	2	2.868020
2	3	1.954435
3	4	1.470006
4	5	1.163976
5	6	0.981300

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

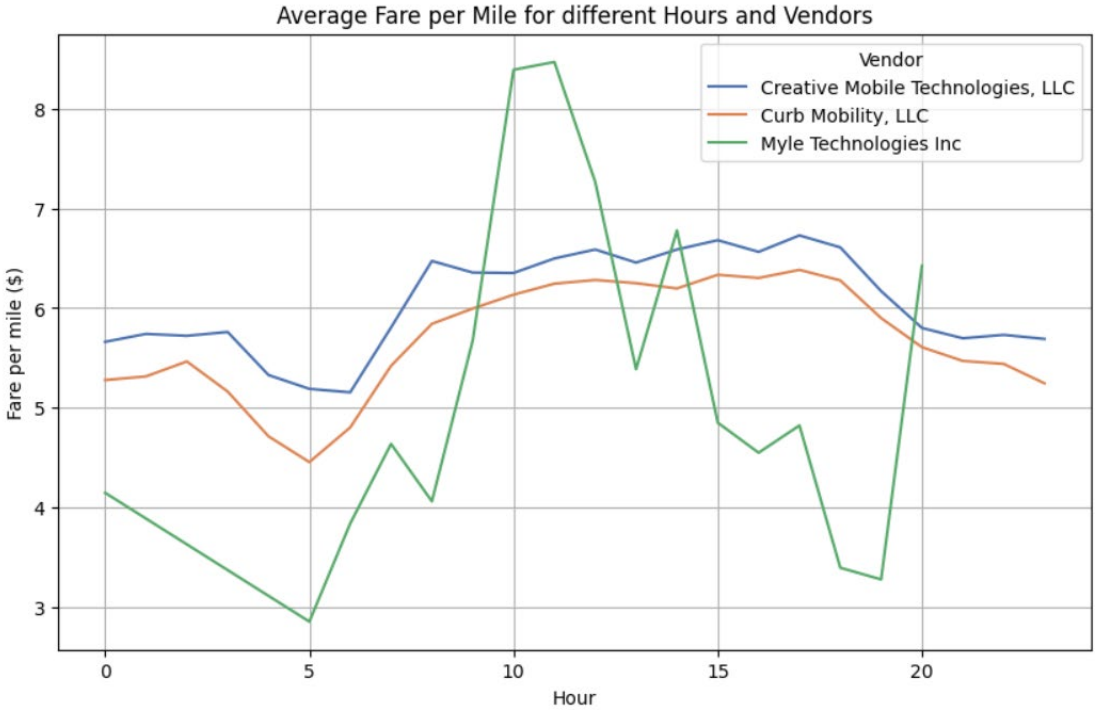
	pickup_weekday	pickup_weekend	fare_amount	trip_distance	fare_per_mile_weekday
0	Monday	Weekday	594222.980000	105284.620000	5.643968
1	Tuesday	Weekday	681497.400000	112417.950000	6.062176
2	Wednesday	Weekday	722706.240000	116881.350000	6.183247
3	Thursday	Weekday	737504.100000	118272.660000	6.235626
4	Friday	Weekday	691313.620000	113885.270000	6.070264
5	Saturday	Weekend	676091.760000	112660.570000	6.001139
6	Sunday	Weekend	624970.660000	114240.390000	5.470663


Average Fare per mile for different weekdays
Weekday
Weekend
Fare per mile ($)
6.4
6.2
6.0
5.8
5.6
5.4
5.2
5.0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of week

	pickup_hour	pickup_bin	fare_amount	trip_distance	fare_per_mile_hour
0	0	Early Morning	139932.950000	26125.890000	5.356103
1	1	Early Morning	87782.030000	16261.390000	5.398187
2	2	Early Morning	55429.030000	10039.740000	5.520963
3	3	Early Morning	38016.270000	7208.260000	5.273987
4	4	Early Morning	31590.600000	6514.180000	4.849513
5	5	Early Morning	35792.300000	7704.920000	4.645382
6	6	Morning	71984.540000	14698.070000	4.897550
7	7	Morning	121065.770000	21919.560000	5.523184
8	8	Morning	171333.140000	28579.190000	5.995031
9	9	Morning	195675.880000	32138.000000	6.088614
10	10	Morning	213817.560000	34552.200000	6.188247
11	11	Morning	235461.070000	37304.570000	6.311856
12	12	Afternoon	258088.390000	40599.310000	6.356965
13	13	Afternoon	269146.580000	42704.150000	6.302586
14	14	Afternoon	292113.440000	46423.680000	6.292337
15	15	Afternoon	294391.580000	45880.120000	6.416539
16	16	Afternoon	297170.970000	46705.560000	6.362647
17	17	Afternoon	312921.970000	48403.510000	6.464861
18	18	Evening	320486.340000	50430.360000	6.355028
19	19	Evening	287442.930000	48213.630000	5.961860
20	20	Evening	260782.320000	46125.240000	5.653788
21	21	Evening	268624.970000	48651.490000	5.521413
22	22	Evening	256497.410000	46595.110000	5.504814
23	23	Evening	212758.720000	39864.680000	5.337023



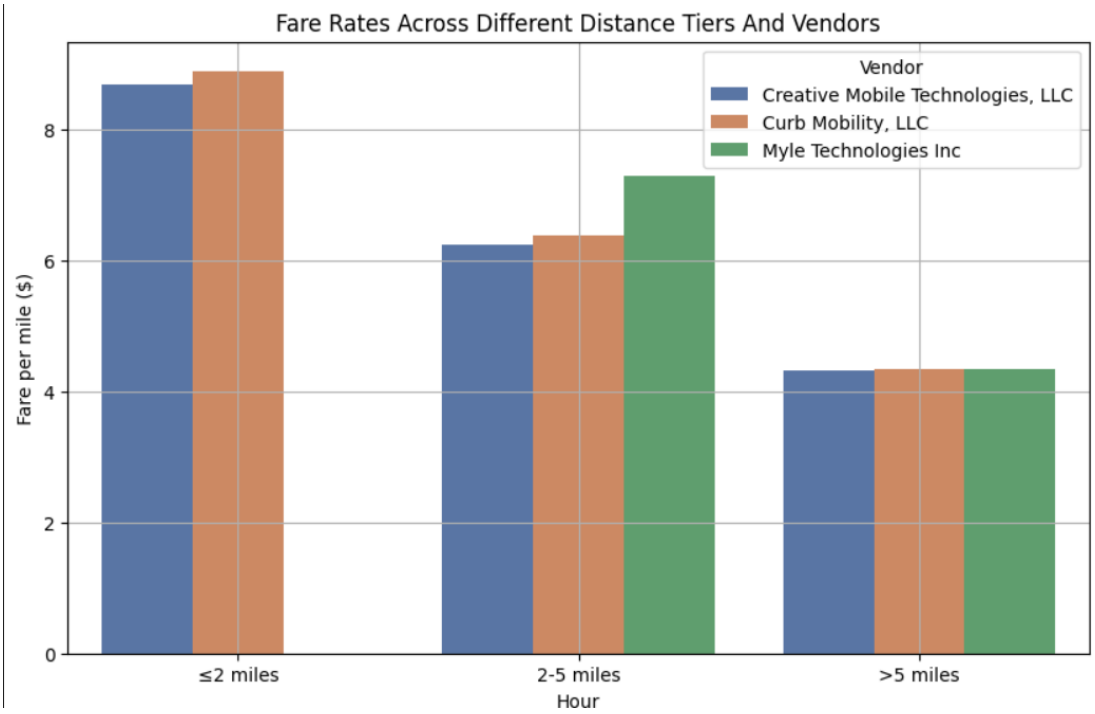
3.2.11. Analyse the average fare per mile for the different vendors



Creative Mobile and Curb show similar average fare per mile, with Creative consistently slightly higher.

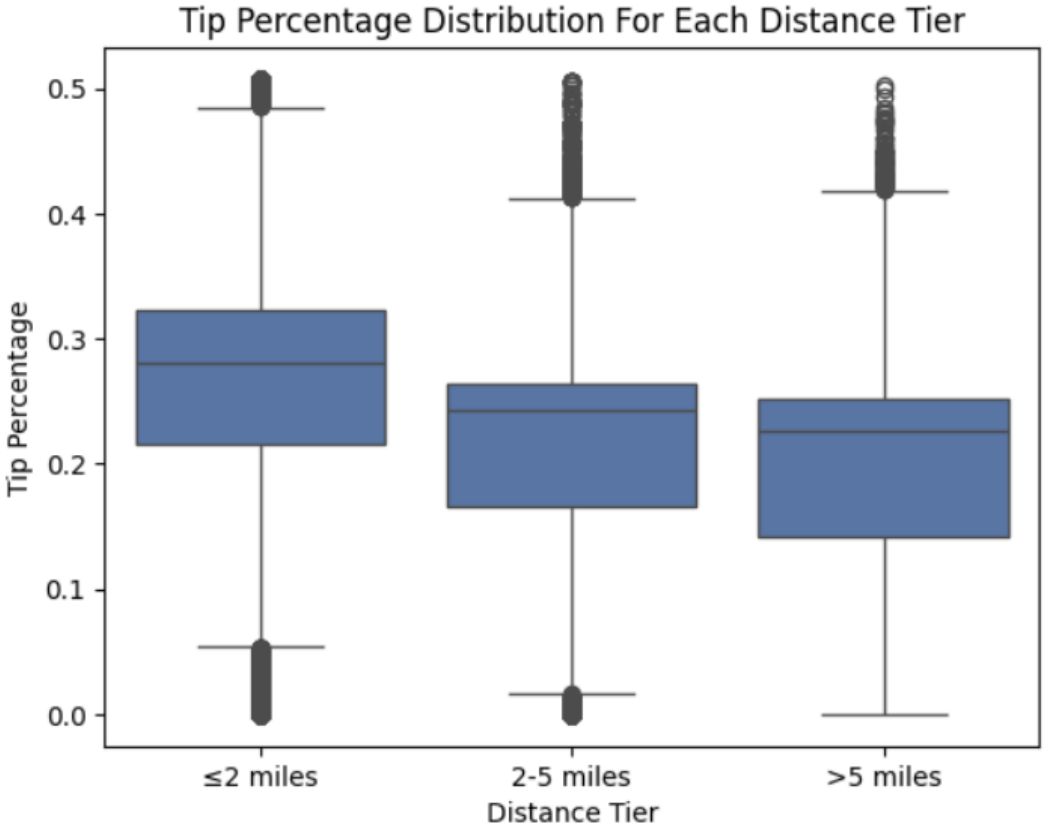
Myle Technologies has a much lower fare per mile overall, but shows sharp peaks at 10 AM, 2 PM, and 8 PM, indicating possible dynamic pricing or service model differences. Other possibilities include low sample size or errors in data entry.

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



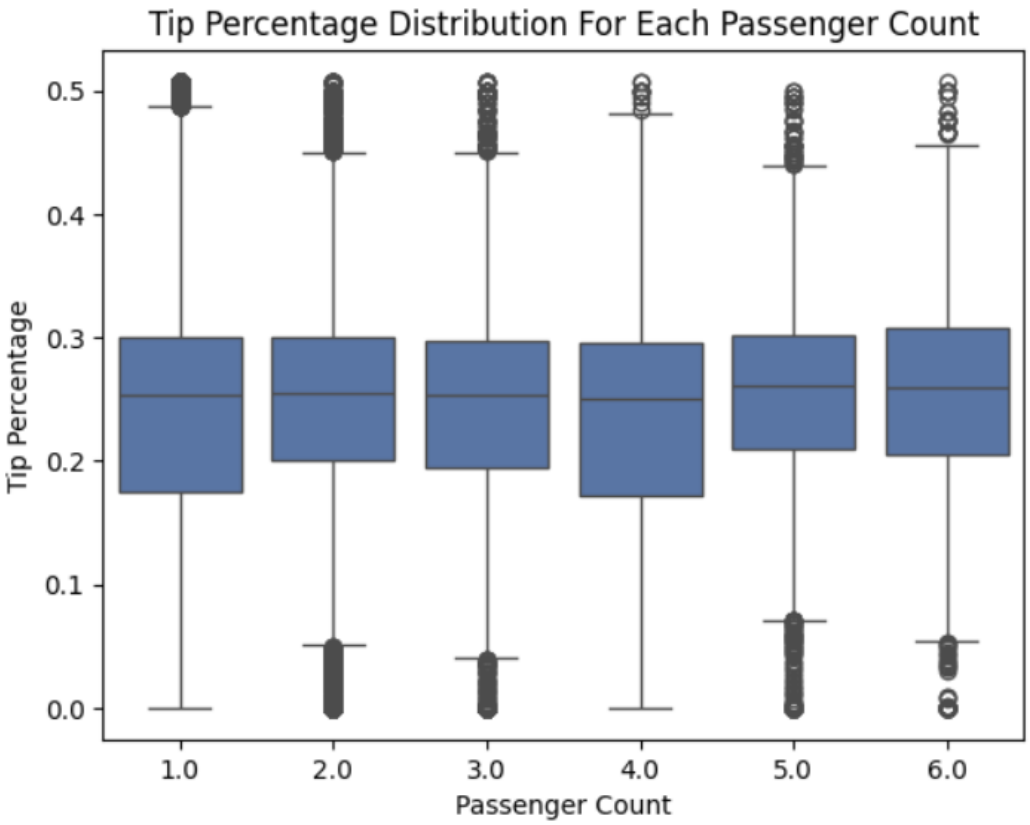
Again, Creative Mobile and Curb show very similar pricing ranges, while Myle Technologies has a noticeably higher rate for the 2-5 miles range. It's also worth noting that Myle Technologies doesn't have any data for the under 2-mile range, indicating that they either don't operate in that range (seems unlikely) or another sign of data entry errors.

3.2.13. Analyse the tip percentages



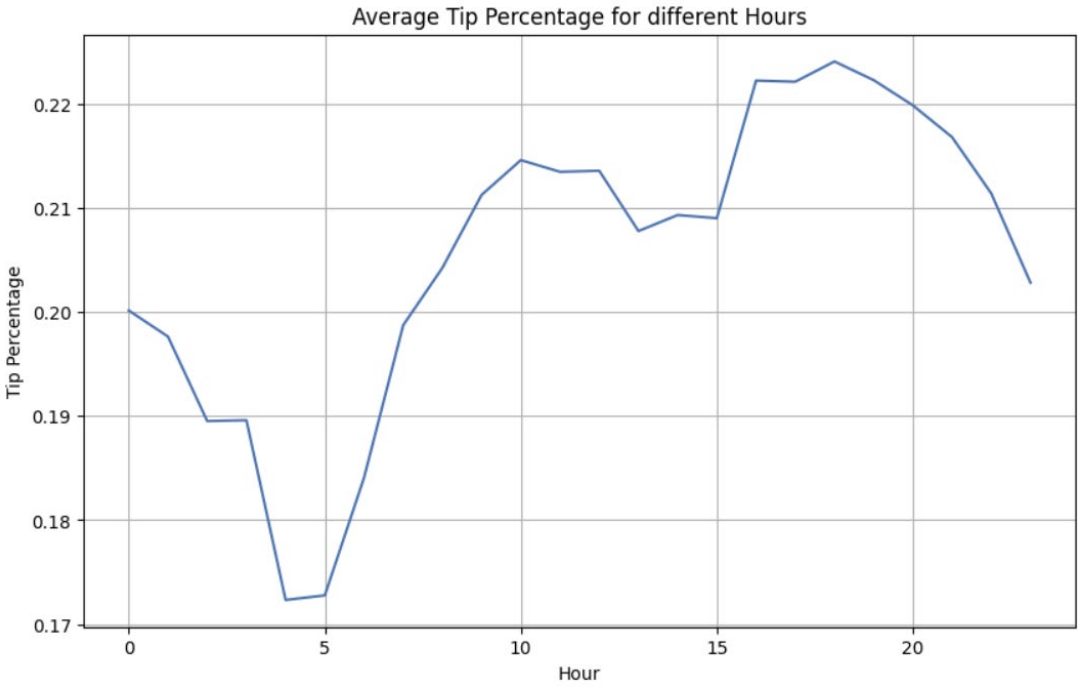
Tip percentage is higher on shorter trips, likely because passengers round up small fares or follow flat tipping habits (e.g. $2–$5), which results in a larger percentage of the fare.

In longer trips, even if the tip amount increases, it makes up a smaller proportion of the total fare.



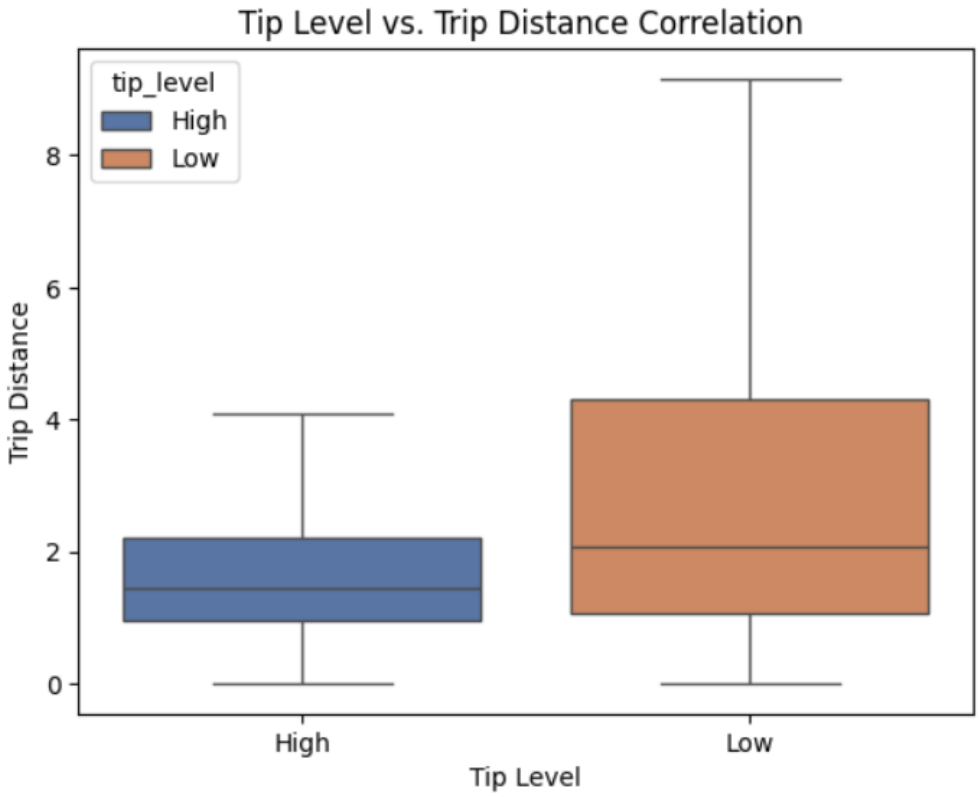
Tip percentage remains consistent across different passenger counts, suggesting that group size does not significantly influence tipping behavior.

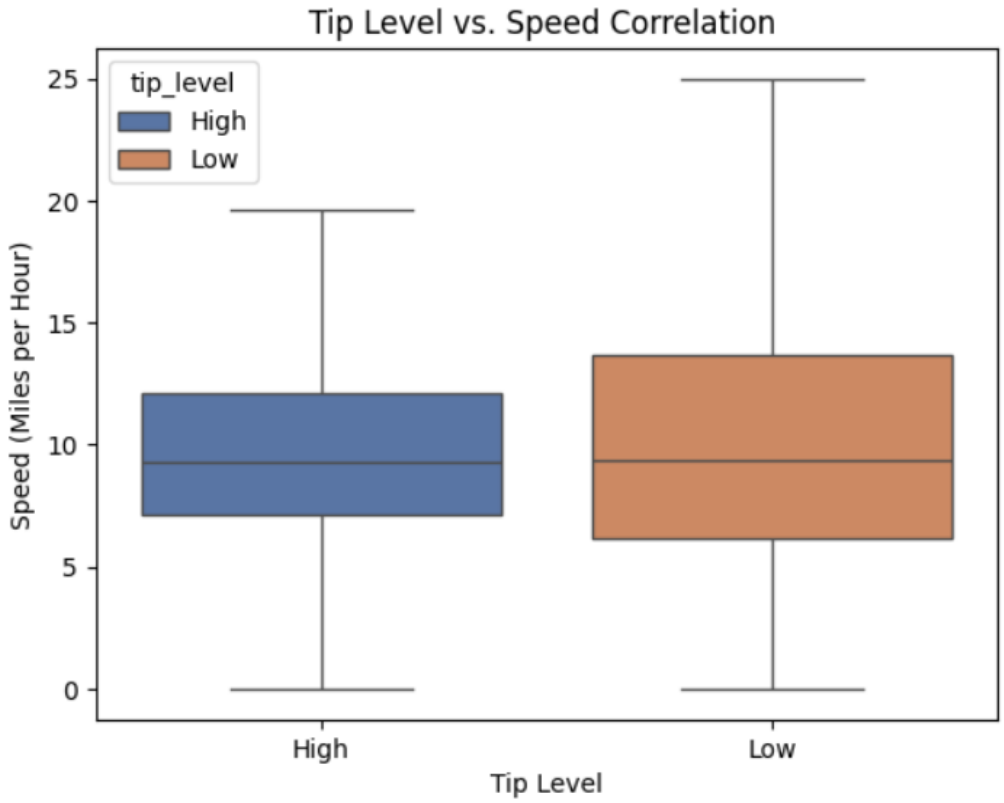
This implies tipping is more closely tied to fare size, payment method, or trip quality rather than how many people are sharing the ride.



Tip percentage is lowest at 5 AM, likely due to quick, non-leisure trips. It rises until 10 AM as more commuters and business travelers ride, then dips midday.

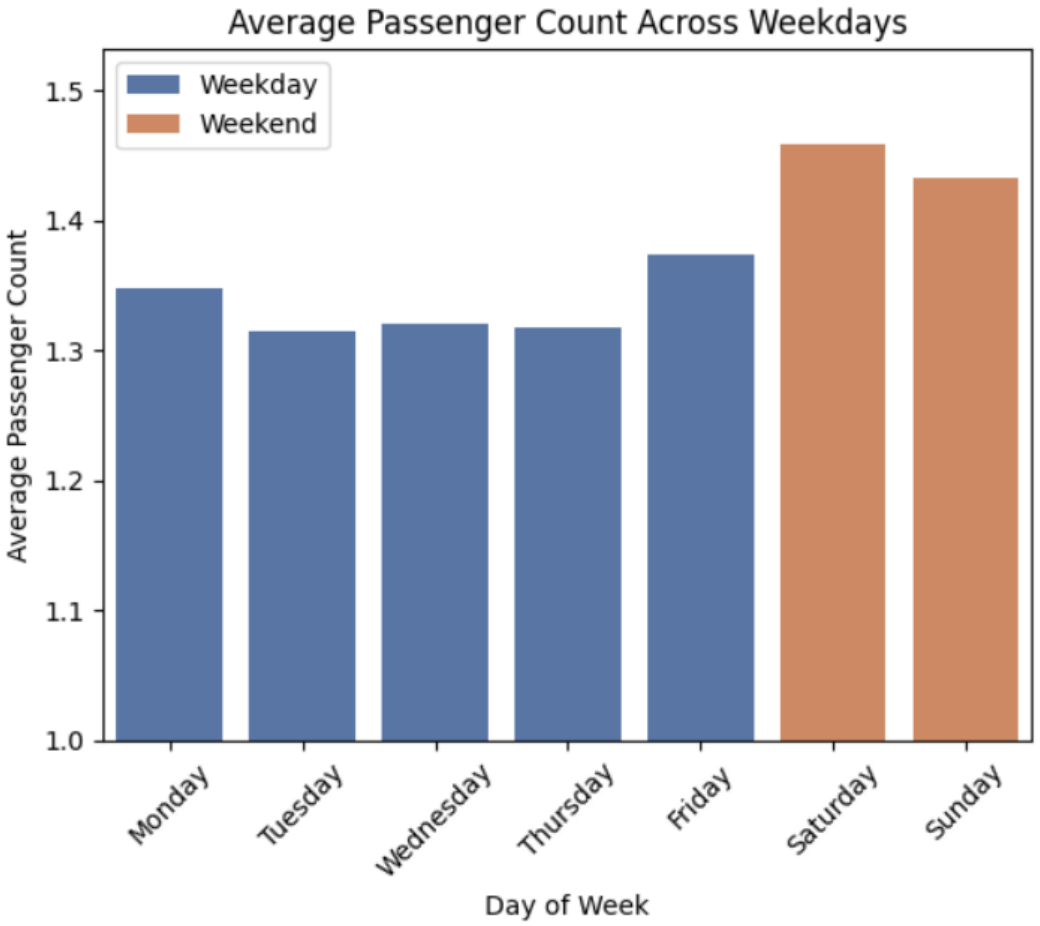
It peaks again from 4–6 PM, possibly due to evening commutes or generous tipping after work hours.


Tip Level vs. Trip Distance Correlation
tip_level
High
Low
Trip Distance
0
2
4
6
8
High
Low
Tip Level

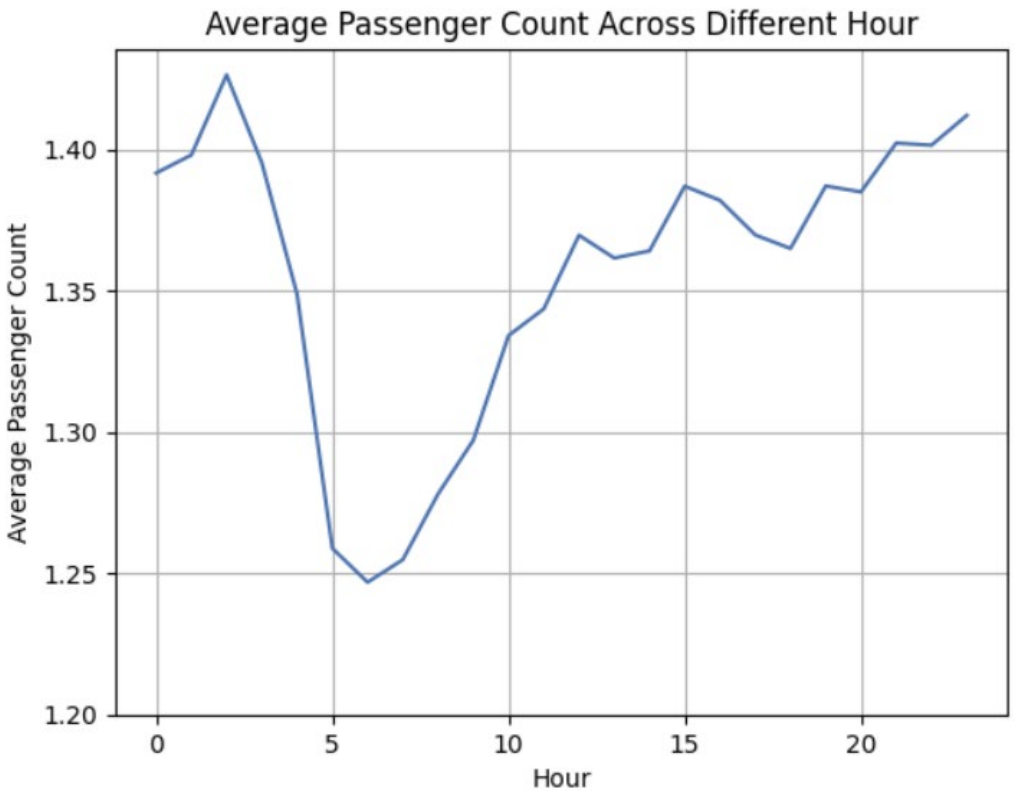

Tip Level vs. Speed Correlation
tip_level
High
Low
Speed (Miles per Hour)
0
5
10
15
20
25
High
Low
Tip Level

Tip percentage is generally higher for shorter trips, likely due to flat tipping habits. However, it remains consistent regardless of ride speed, suggesting passengers tip based on fare amount or distance, not how fast the ride was.

3.2.14. Analyse the trends in passenger count

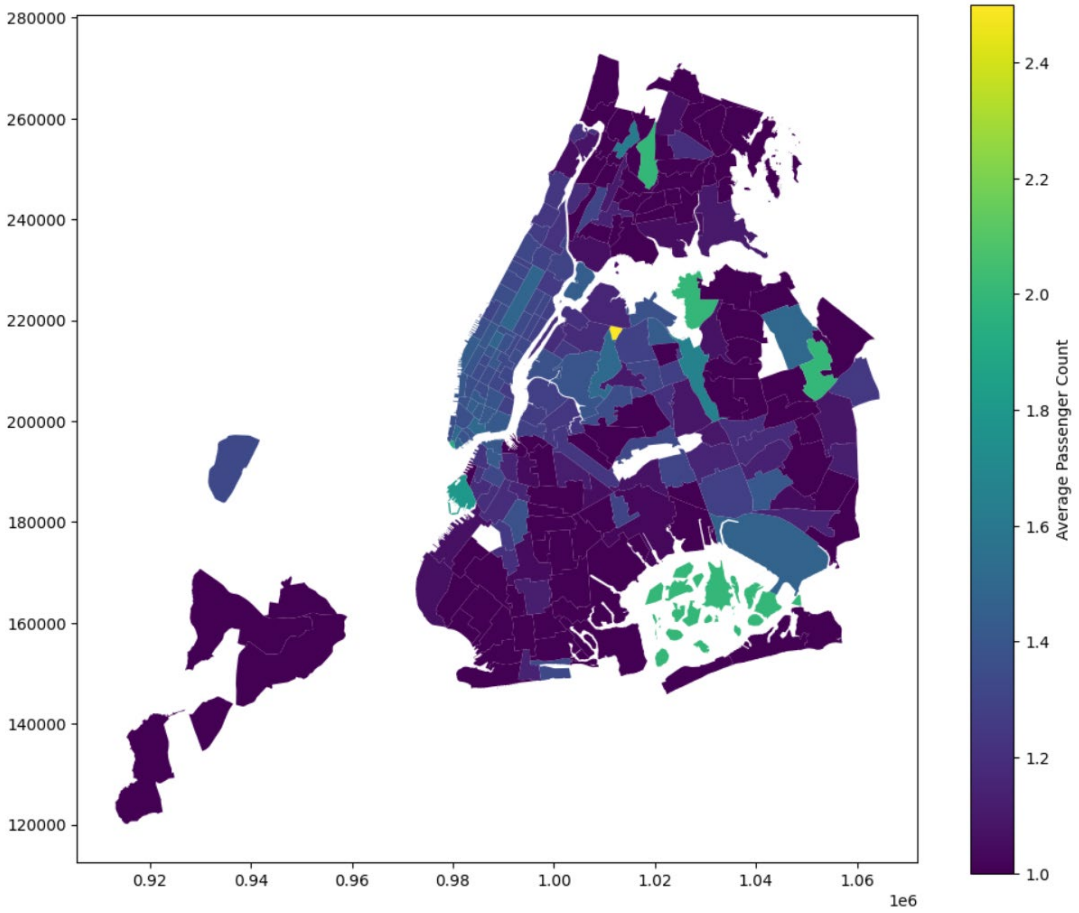


Passenger count per trip is higher on weekends, likely due to group outings, social events, and nightlife, where people tend to travel together more often than on weekdays.



Passenger count per trip gradually increases from 5 AM, reaching its peak around 1 AM, reflecting a shift from solo morning commutes to late-night group outings and social travel.

3.2.15. Analyse the variation of passenger counts across zones



Passenger counts vary notably across zones:

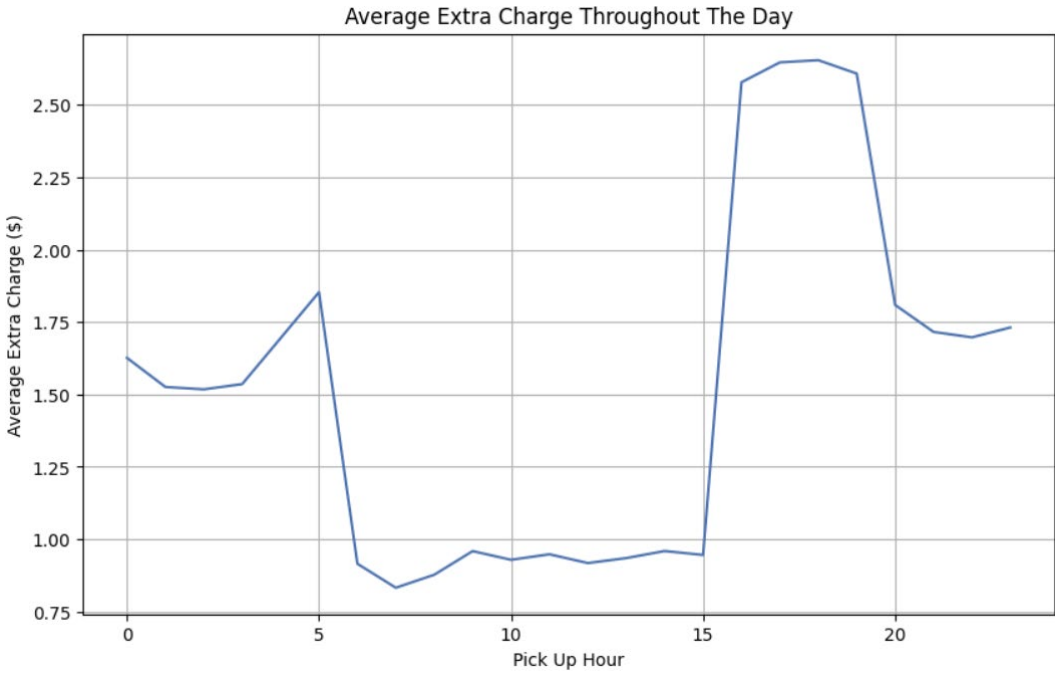
Midtown, Times Square, and Financial District have higher average passenger counts, likely due to tourists, business groups, and shared rides

Residential zones and outer boroughs tend to have lower counts, reflecting solo or routine local trips.

Airport zones show moderate to high passenger counts, often from group airport transfers.

This pattern highlights how zone type (commercial, residential, tourist) influences group travel behavior.

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



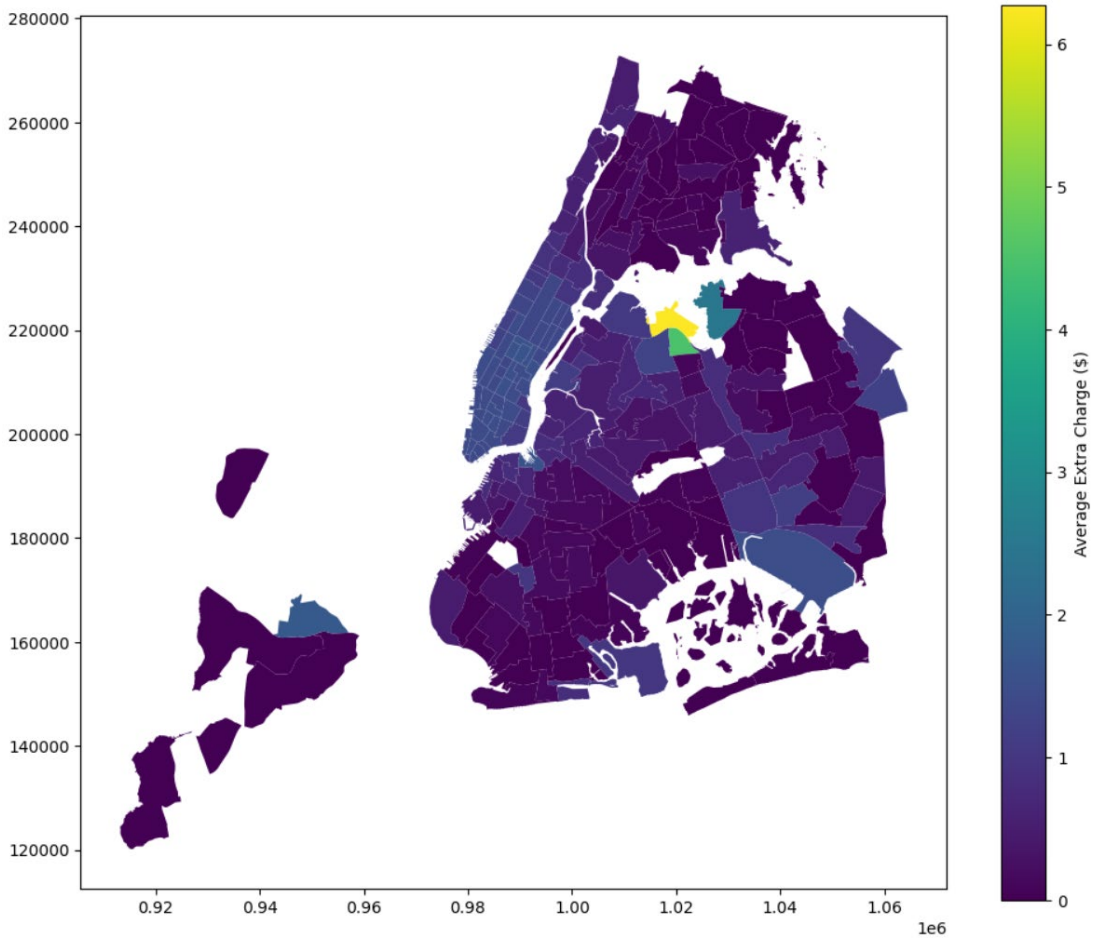
The variation in average extra charges can be explained as follows:

5 AM Peak: Likely due to early morning surcharges for late-night or early-morning rides, airport pickups, or shift change premiums when demand is low but operational costs are higher.

Low from 6 AM to 3 PM: During this period, extra charges are low due to steady, regular demand and no significant surcharges for rush hour or traffic.

4 PM to 7 PM Peak: Extra charges rise sharply due to rush hour traffic congestion, surge pricing, and evening commute rates.

After 7 PM: Extra charges drop but still stay higher than in the morning, reflecting nighttime premiums for events and lower driver availability.



Extra charges are notably higher in airport zones and busy areas due to several factors:

Airports: Airport pickups incur additional costs such as surcharges, tolls, and the time spent waiting for passengers. These locations also tend to have higher operational costs for drivers, particularly during busy travel periods, leading to higher extra charges for passengers.

Busy Areas: In highly congested zones like Midtown Manhattan or Times Square, extra charges rise due to surge pricing driven by high demand. The increased traffic congestion in these areas leads to longer trip durations, which in turn increases costs. Demand spikes in these locations, particularly during peak hours, further drive up the cost of rides.

4. Conclusions

4.1. Final Insights and Recommendations

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

Demand-based routing: Utilize real-time data to predict demand spikes in busy areas and airport zones, optimizing routing to avoid congestion. Implement dynamic dispatching systems that place cabs in high-demand zones during peak times, reducing idle time and increasing efficiency.

Time-of-day scheduling: Adjust driver shifts to align with peak demand hours, focusing on commute times (8–10 AM, 5–7 PM) and nightlife hours (10 PM–1 AM). This reduces the need for drivers to wait idle and increases the likelihood of quick, profitable fares.

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.

Manhattan & Busy Zones: Focus on placing cabs in Midtown, Times Square, and other high-traffic areas, especially during rush hour and peak seasons (spring and September). These zones see consistent demand and can benefit from surge pricing.

Airports & Transportation Hubs: Position cabs near JFK, LaGuardia, and big stations to capture airport and transit-based rides, where demand is high throughout the day.

Outer Boroughs: Adjust fleet positioning in Queens, Brooklyn, and the Bronx based on time-of-day demand. These areas tend to have lower demand but can be optimized for afternoon and evening travel. Fleet movements should consider low-traffic hours in these zones.

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

Dynamic Pricing: Implement surge pricing during periods of high demand (e.g., rush hour, weekends, evenings). Focus on airports and high-traffic zones where there is consistent demand and longer trips.

Time-based Discounts: Consider offering discounts or incentives during low-demand periods (e.g., midday or late-night rides) to attract

customers.

Fare Transparency: Revise pricing models to stay competitive with other vendors. Offering flat fares or package pricing for regular commuters can help build customer loyalty while ensuring profitability.