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.

## Data Preparation

* 1. Loading the dataset
     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.

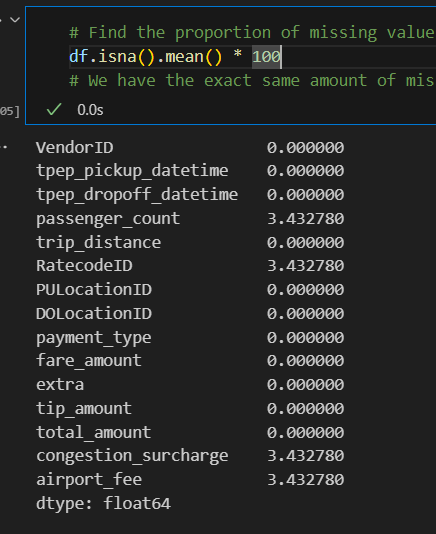
## Data Cleaning

### Fixing Columns

* + 1. **Fix the index**  
       We reset and cleaned the index to avoid misalignment from concatenation.
    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.

### Handling Missing Values

* + 1. **Find the proportion of missing values in each column**



* + 1. **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)

* + 1. **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. **Impute NaN in congestion\_surcharge**  
       As the congestion\_surcharge has discrete values, we will use the median to fill in the missing values

### Handling Outliers and Standardising Values

* + 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

## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 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

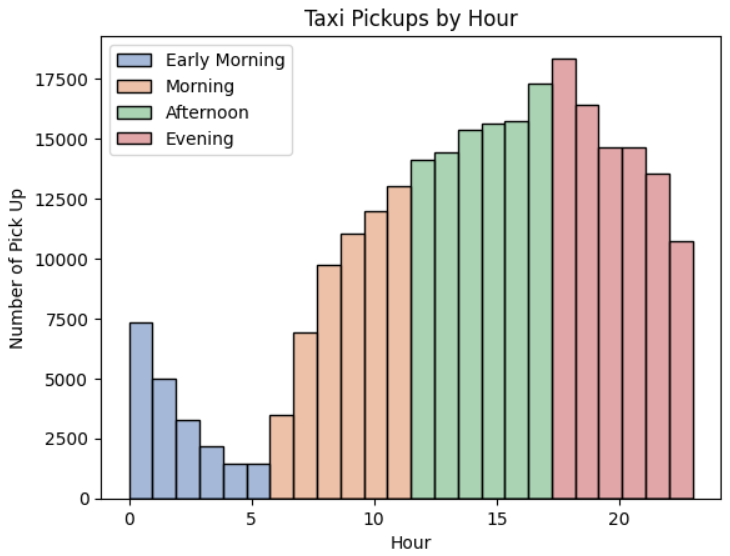
`pickup\_hour`: Numerical

`trip\_duration`: Numerical

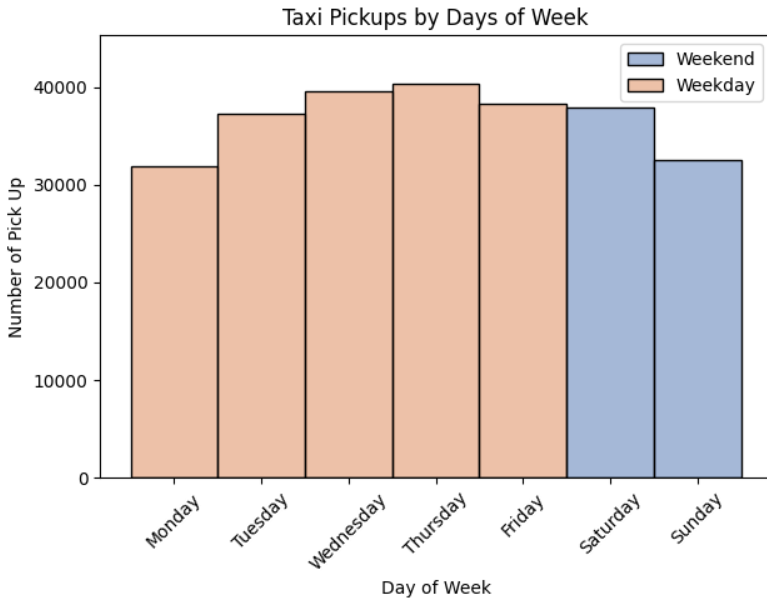
The following monetary parameters belong in the same category, is it categorical or numerical?

Answer: They are numerical

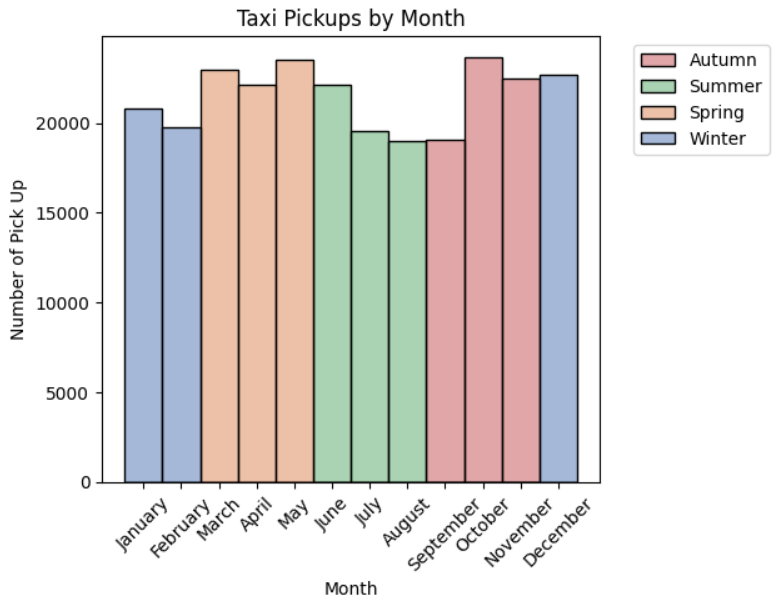
* + 1. **Analyse the distribution of taxi pickups by hours, days of the week, and months**

****

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.



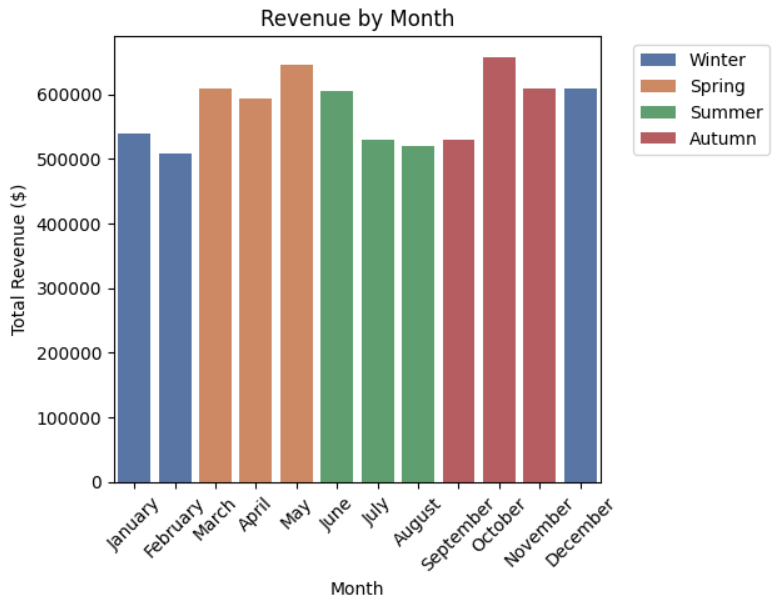
Taxi usage is at its highest in spring, decreases gradually during summer, then peaks again in September. This can be explained due to spring being a favorable season for tourism, and September being the back-to-work and back-to-school month.

* + 1. **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

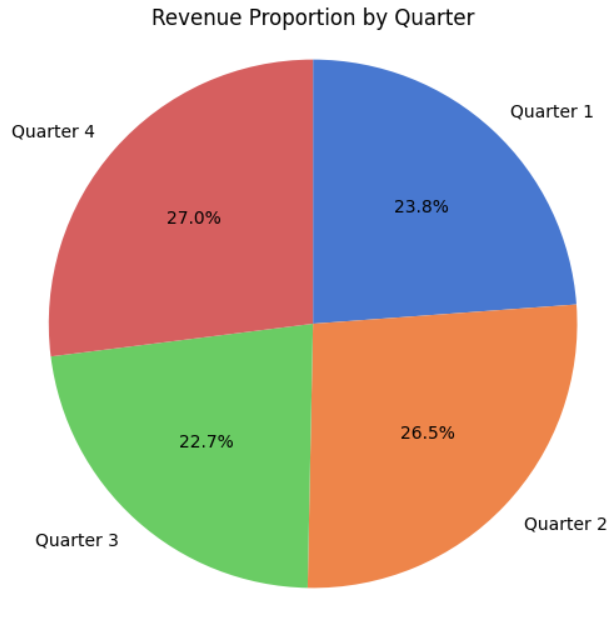
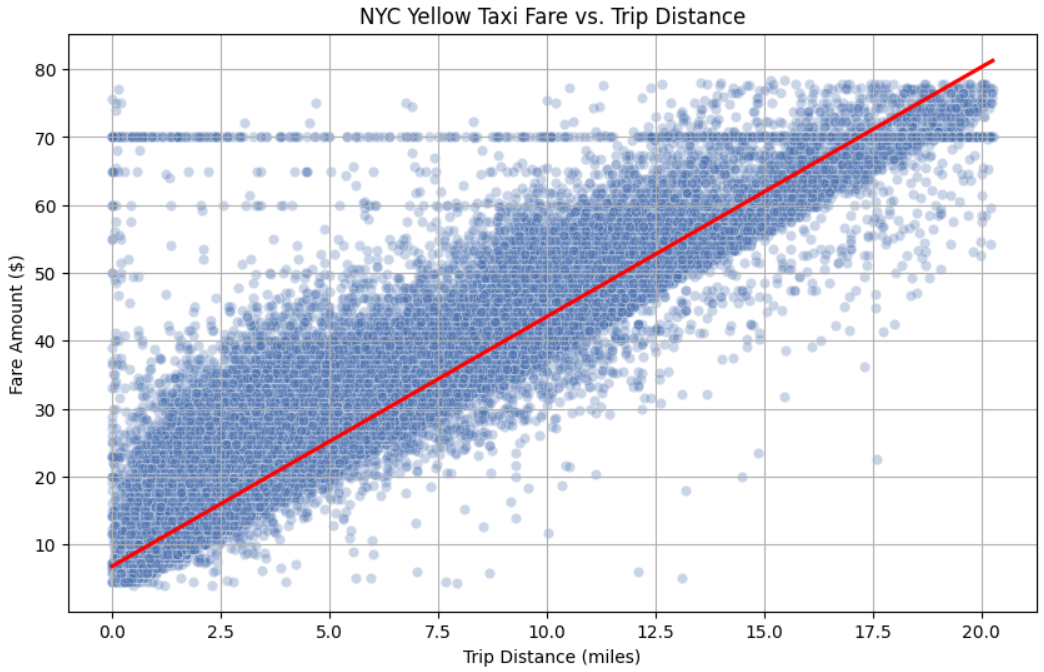
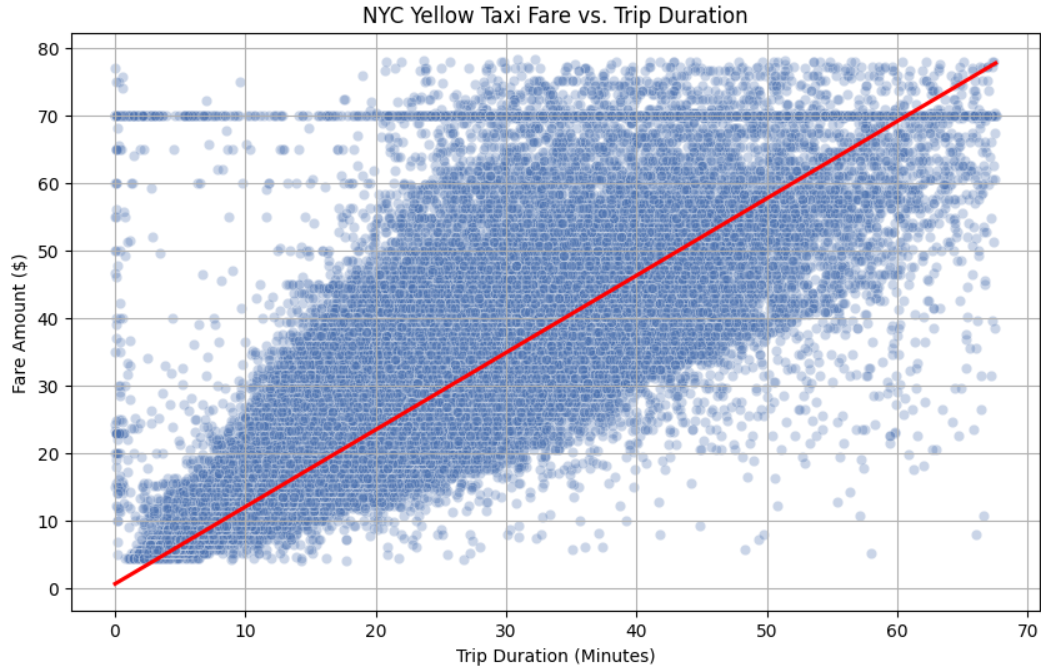
Zero values for fare\_amount have been considered outliers and were discarded.

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

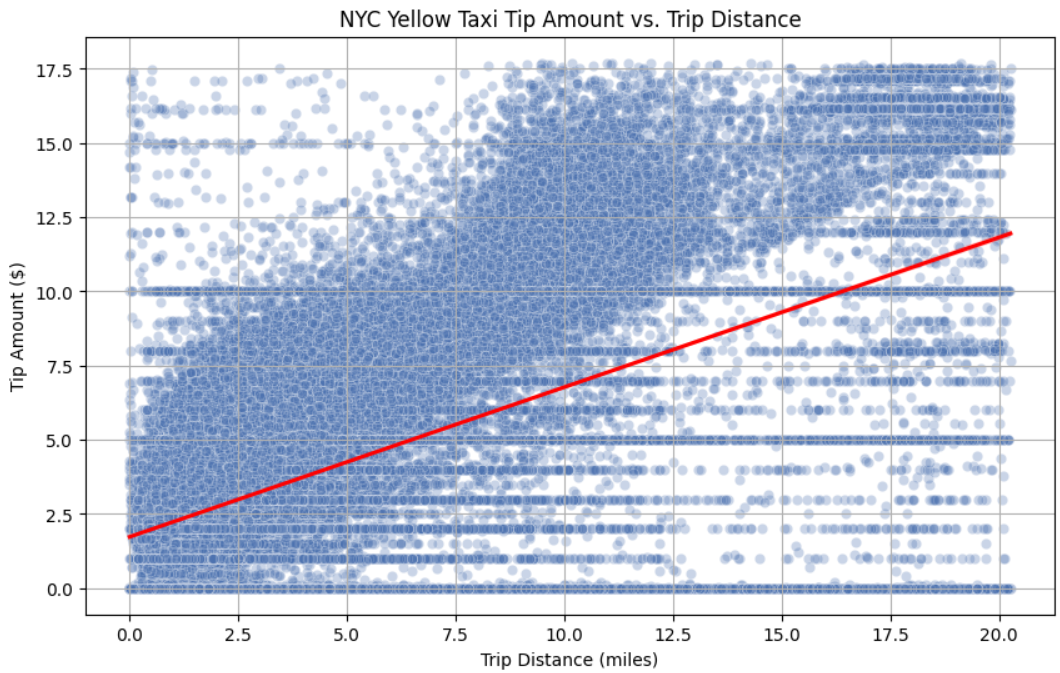
* + 1. **Analyse the monthly revenue trends  
         
       **

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.

* + 1. **Find the proportion of each quarter’s revenue in the yearly revenue  
       **
    2. **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. **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.

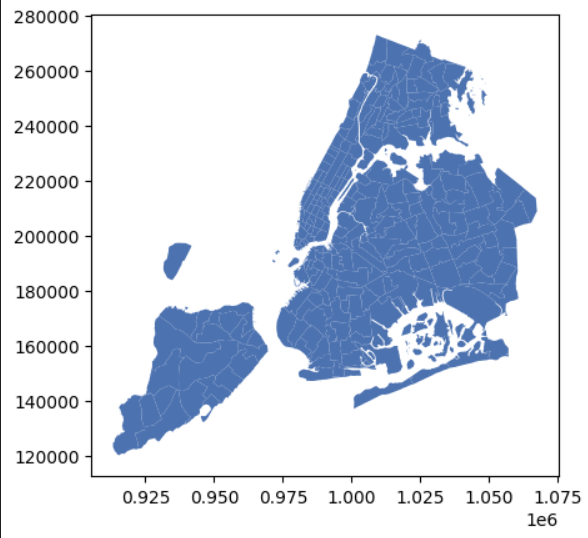


Tip vs. Trip Distance: Tips generally increase with distance, as they are often a proportion of the fare amount. The longer the distance is, the higher the fare amount will be.

* + 1. **Analyse the distribution of different payment types  
         
       **

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.

* + 1. **Load the taxi zones shapefile and display it  
         
       **
    2. **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()

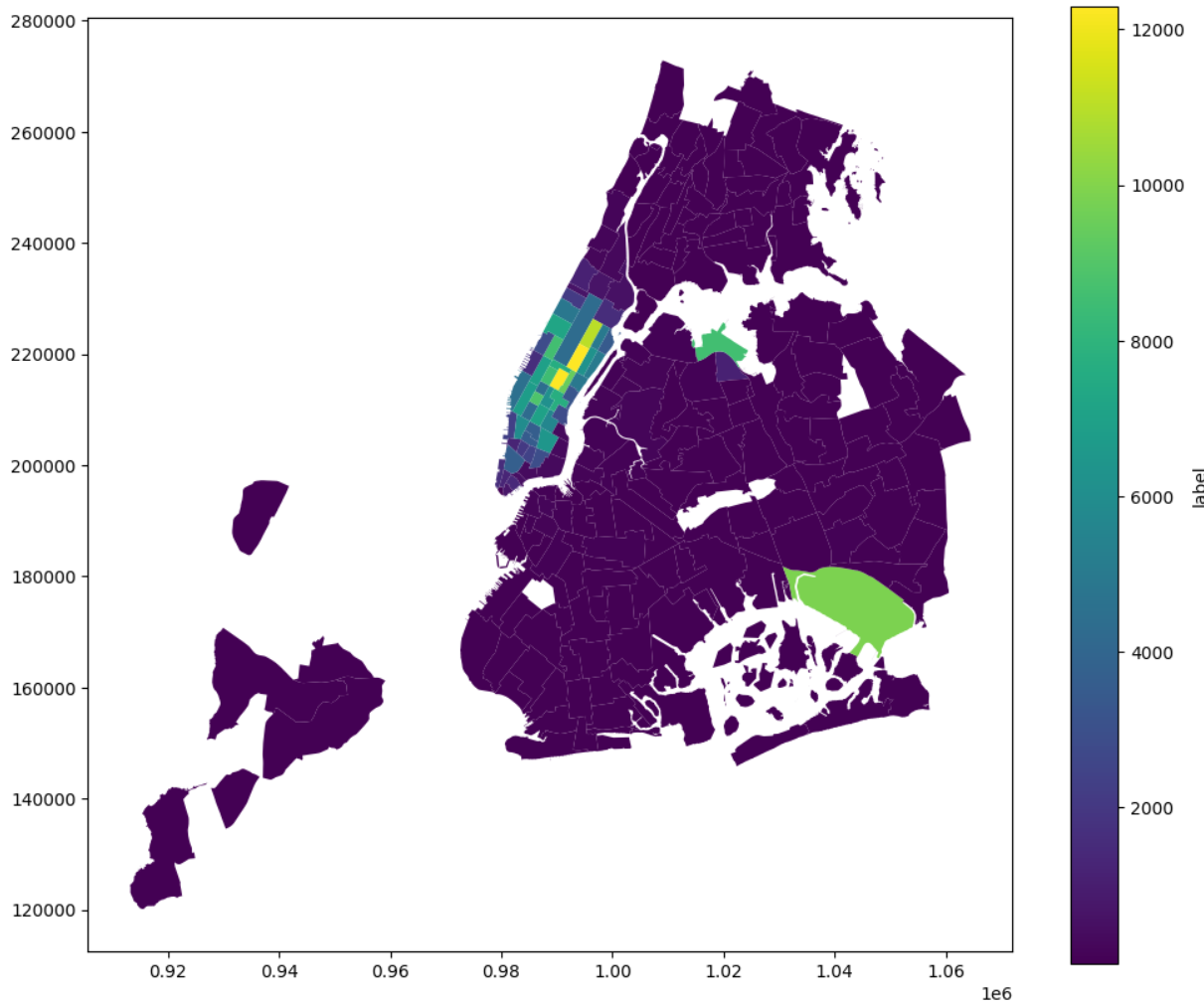
* + 1. **Find the number of trips for each zone/location ID**location\_count = zoned\_df.groupby('LocationID')['LocationID'].count().rename('trip\_count').reset\_index()

location\_count

* + 1. **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

* + 1. **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.

* + 1. **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.

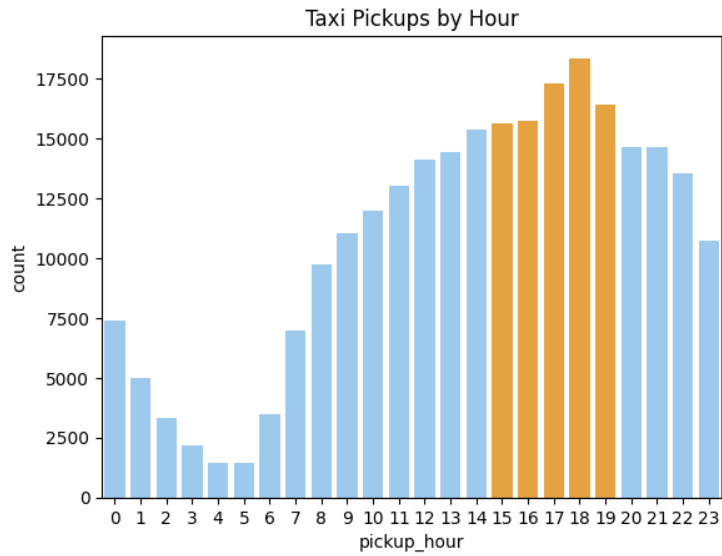
Fare Drivers: Fare increases with distance and duration; more passengers reduce fare per person.

Tips: Higher on longer trips.

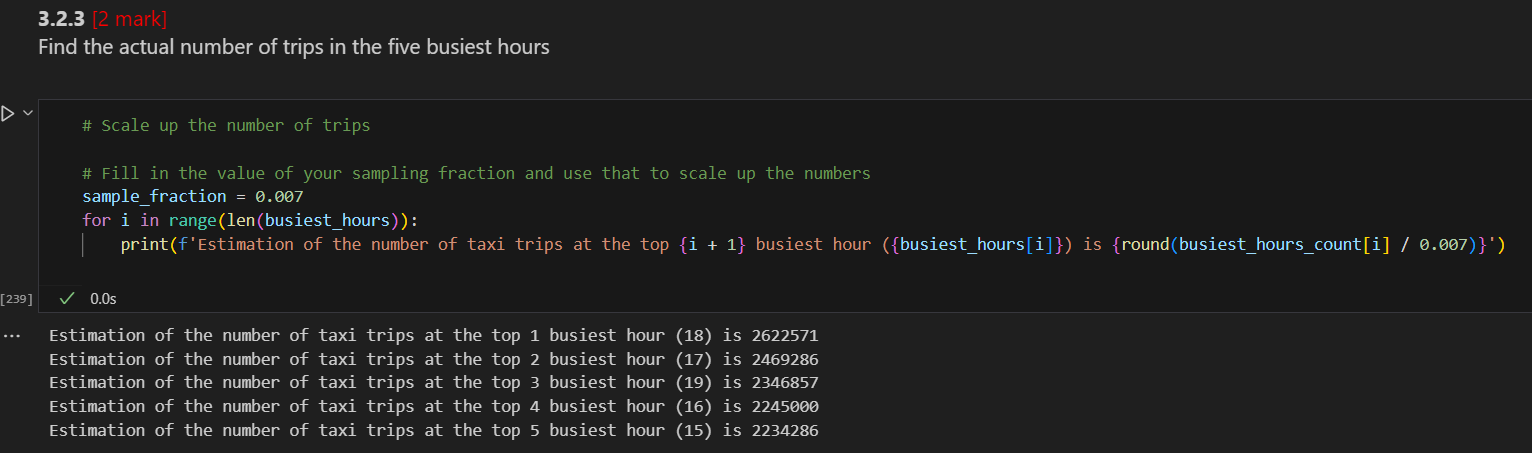
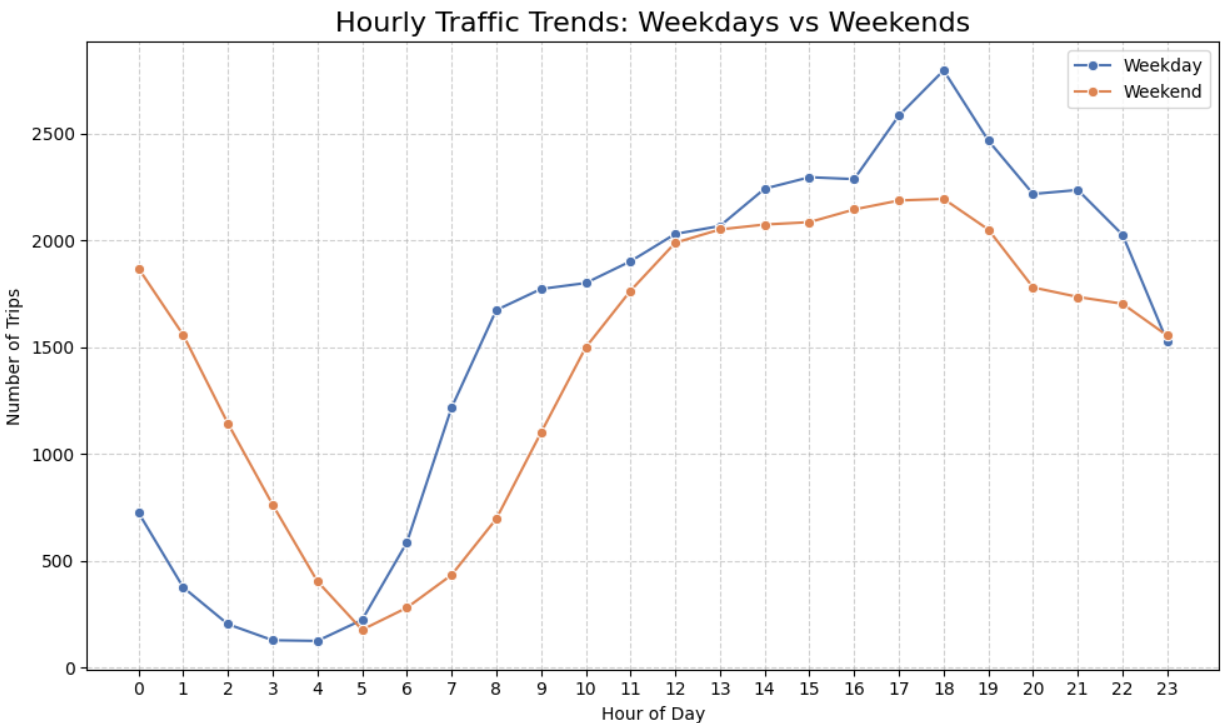
Hot Zones: Midtown and Financial District have the most trips; airports also see high activity.

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

### Detailed EDA: Insights and Strategies

* + 1. **Identify slow routes by comparing average speeds on different routes  
         
       **
    2. **Calculate the hourly number of trips and identify the busy hours  
         
       **

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

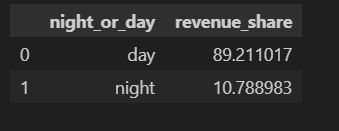
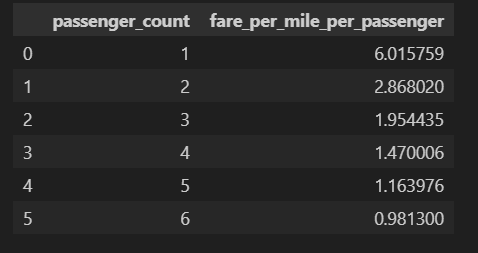
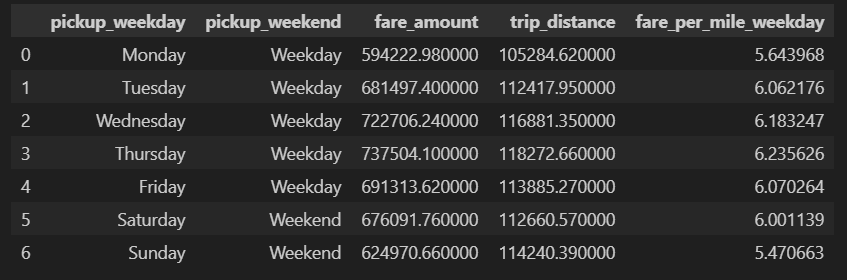
* + 1. **Scale up the number of trips from above to find the actual number of trips  
         
       **
    2. **Compare hourly traffic on weekdays and weekends  
         
       **

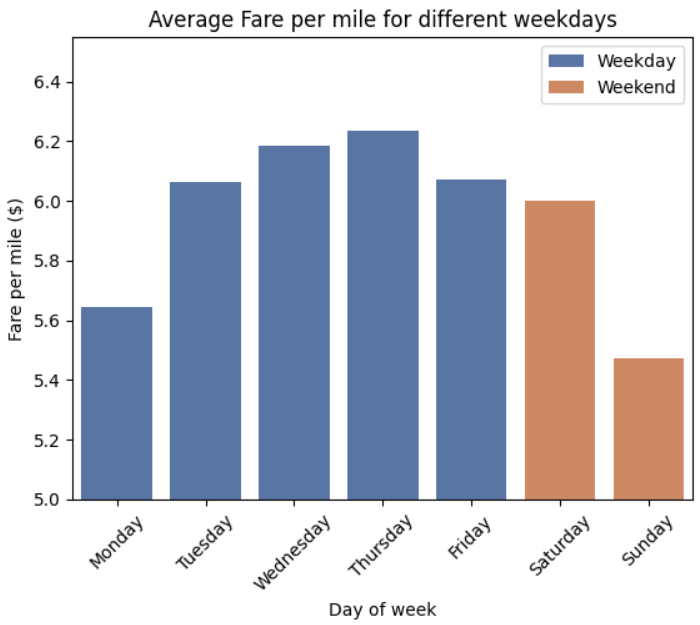
Taxi demand on weekdays is consistently higher during work hours, while weekend demand is much higher after midnight until 5 AM, driven by nightlife.

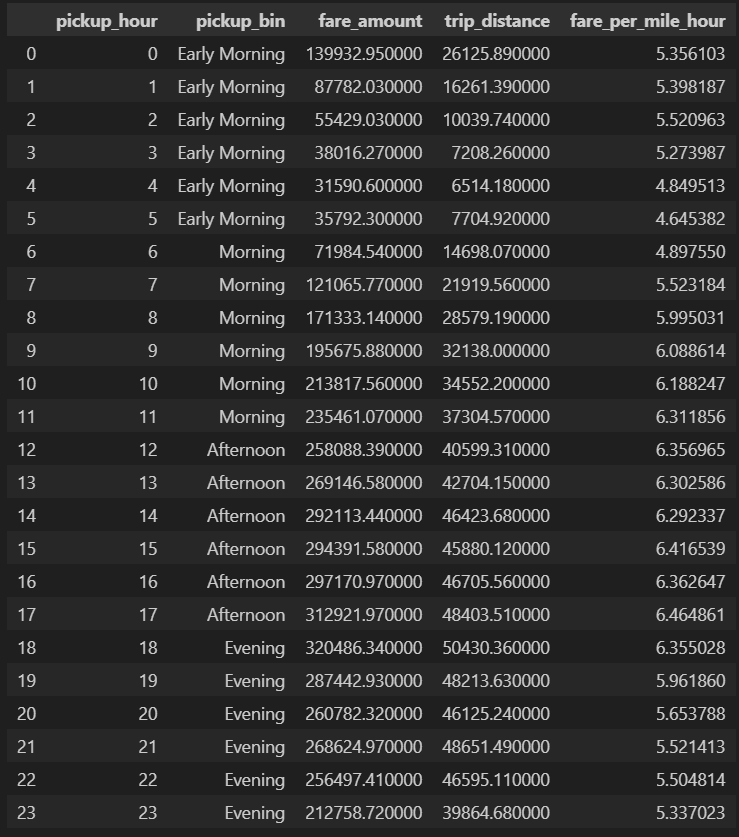
Knowing busy and quiet hours helps optimize driver shifts, reduce wait times, and boost revenue.

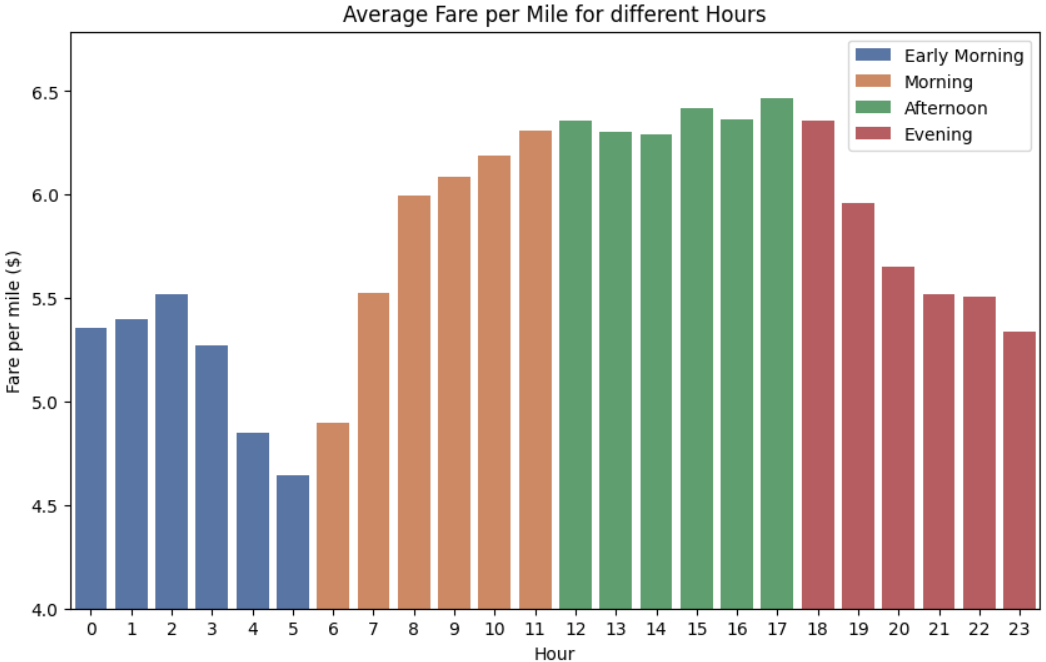
* + 1. **Identify the top 10 zones with high hourly pickups and drops  
       **

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)

* + 1. **Find the ratio of pickups and dropoffs in each zone  
         
       **
    2. **Identify the top zones with high traffic during night hours  
         
       **
    3. **Find the revenue share for nighttime and daytime hours  
       **
    4. **For the different passenger counts, find the average fare per mile per passenger  
         
       **
    5. **Find the average fare per mile by hours of the day and by days of the week  
         
       **

****

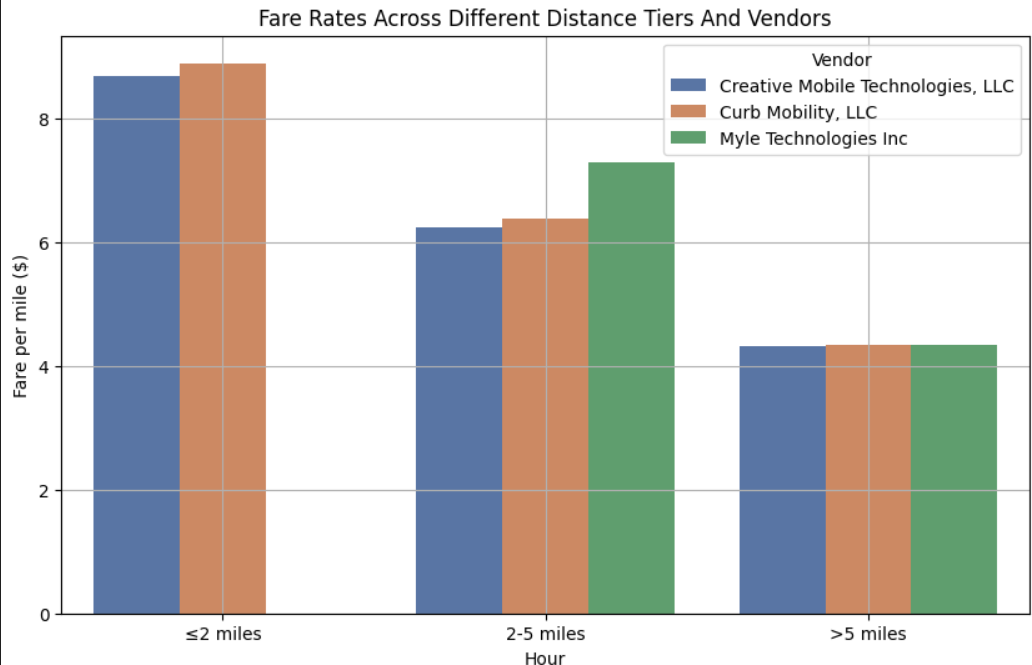
****

****

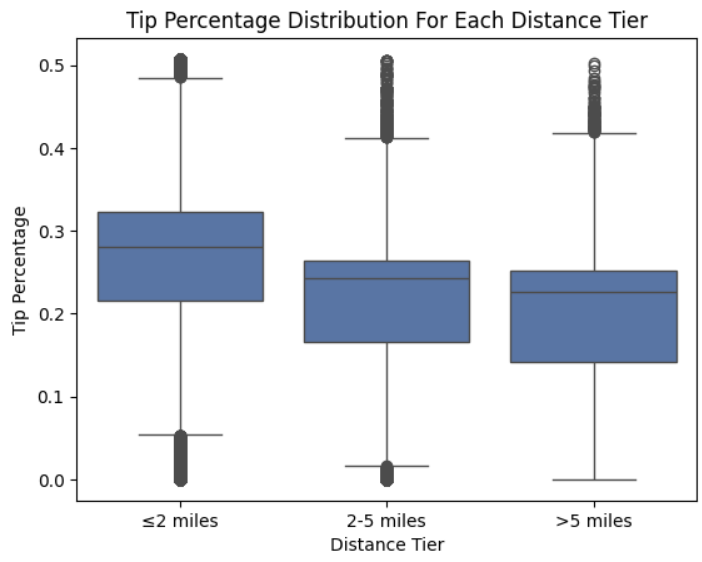
* + 1. **Analyse the average fare per mile for the different vendors  
         
       **

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

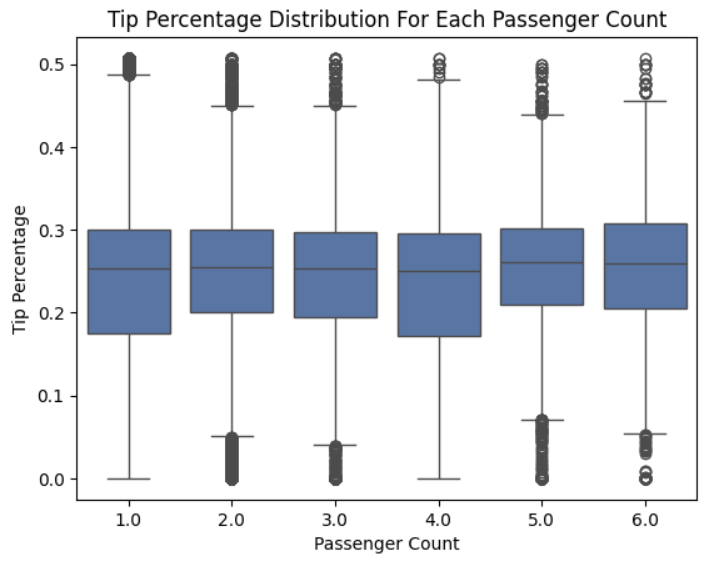
Myle Technologies has a much lower fare per mile overall, but shows sharp peaks at 10 AM, 2 PM, and 8 PM, indicating possible dynamic pricing or service model differences. Other possibilities include low sample size or errors in data entry.

* + 1. **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.

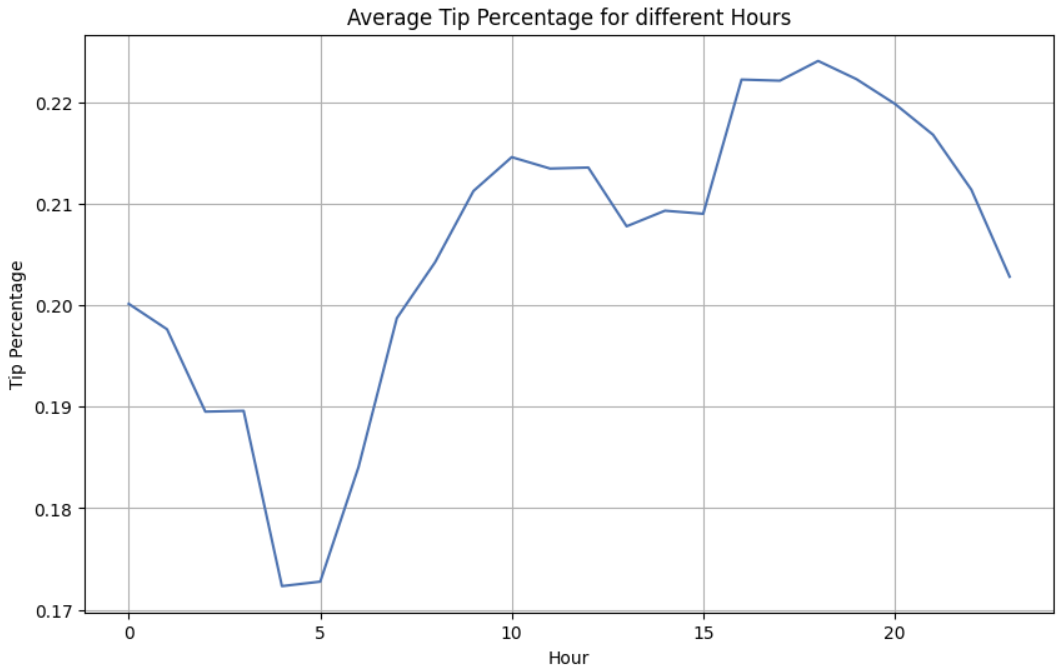
* + 1. **Analyse the tip percentages  
       **Tip percentage is higher on shorter trips, likely because passengers round up small fares or follow flat tipping habits (e.g. $2–$5), which results in a larger percentage of the fare.

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

****

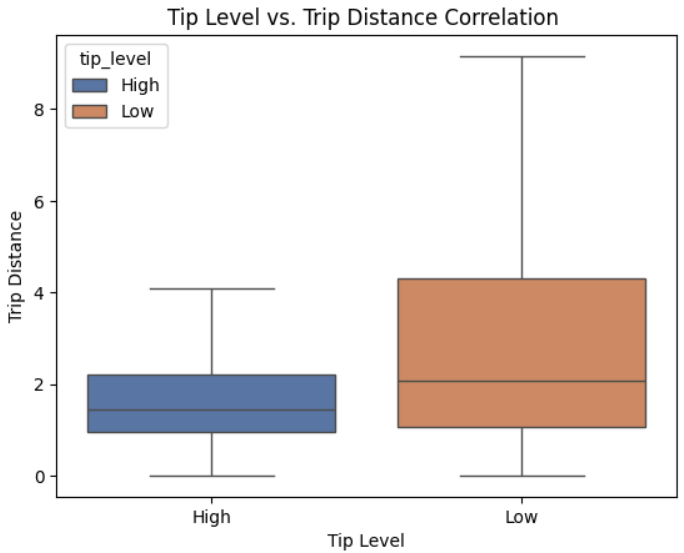
Tip percentage remains consistent across different passenger counts, suggesting that group size does not significantly influence tipping behavior.

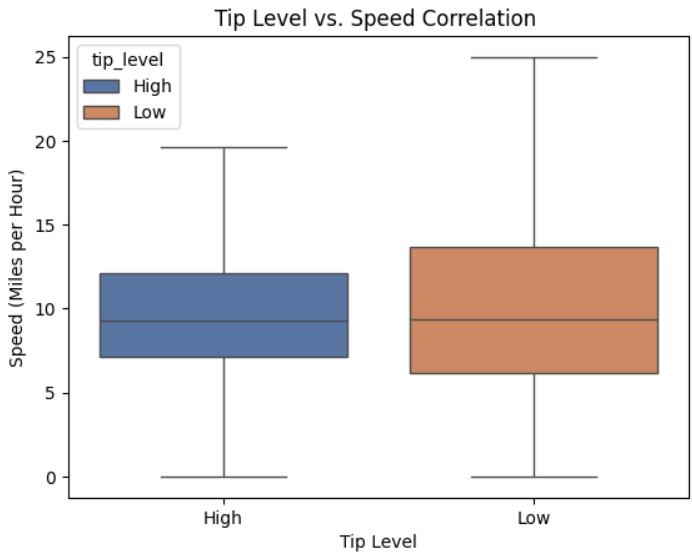
This implies tipping is more closely tied to fare size, payment method, or trip quality rather than how many people are sharing the ride.



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.

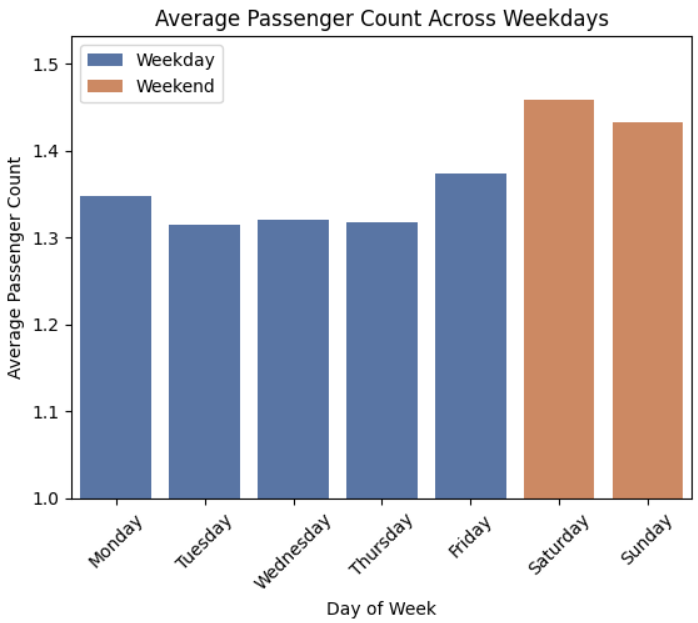
It peaks again from 4–6 PM, possibly due to evening commutes or generous tipping after work hours.





Tip 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.

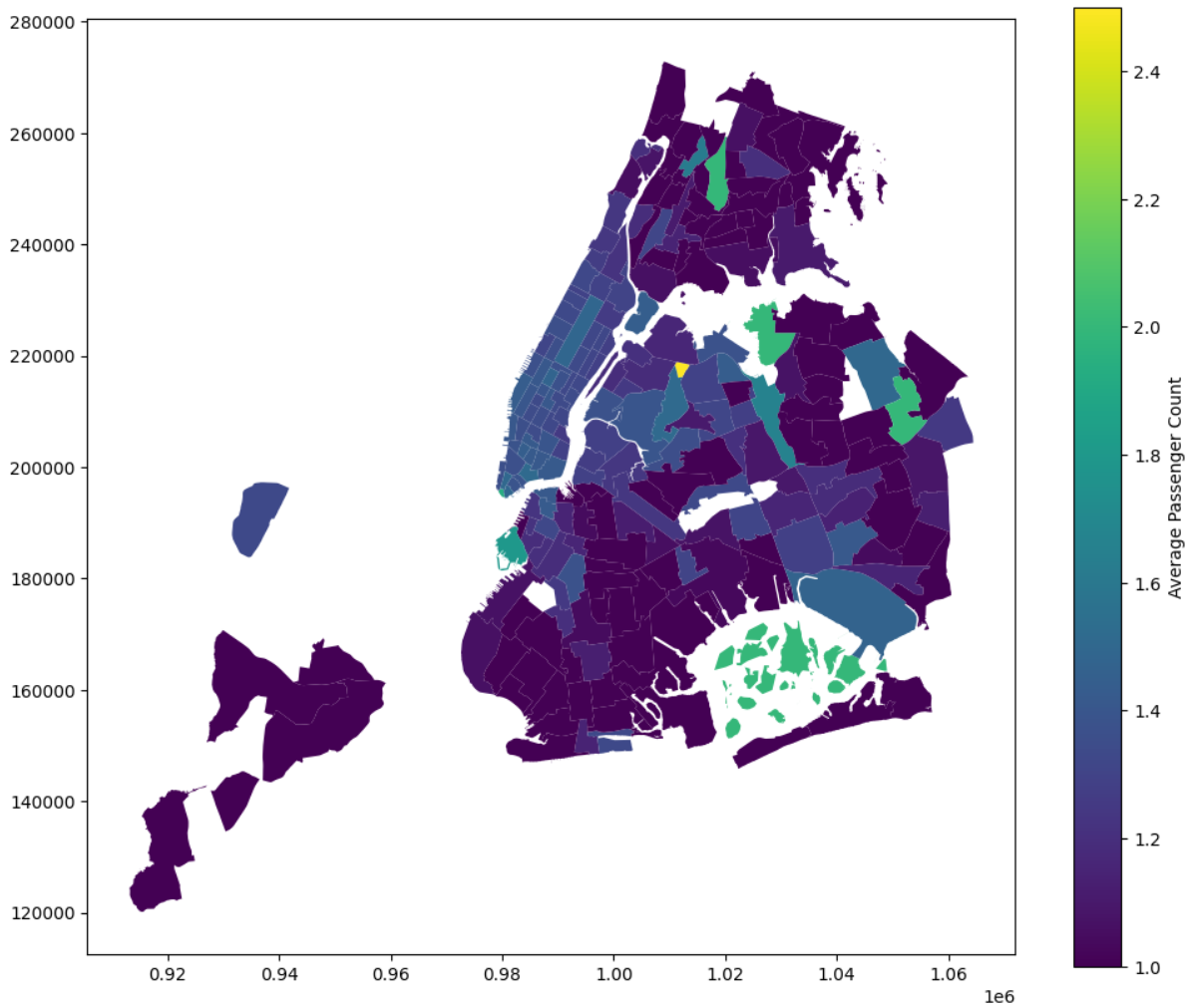
* + 1. **Analyse the trends in passenger count**



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.

* + 1. **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.

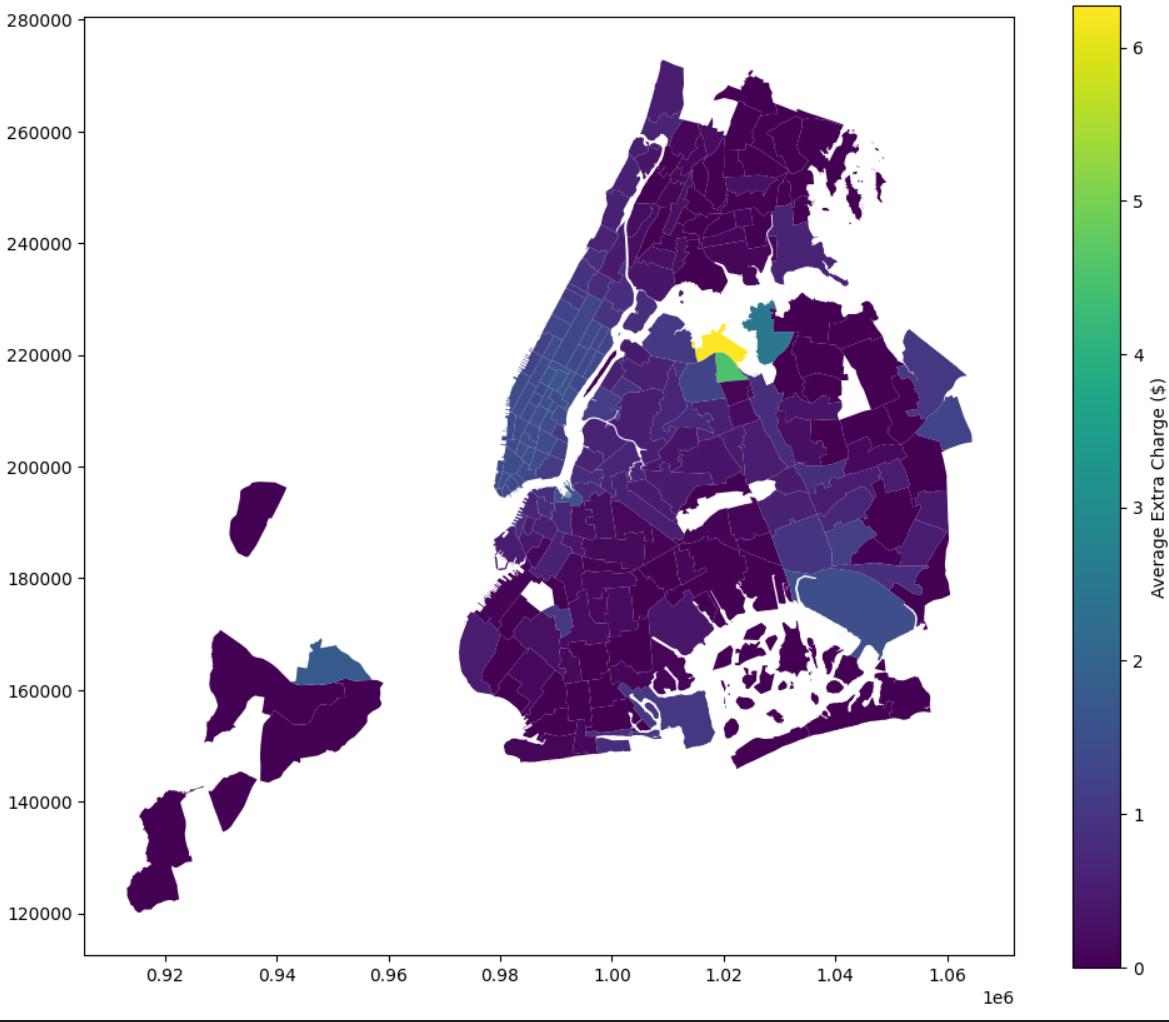
* + 1. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.  
         
       **The variation in average extra charges can be explained as follows:

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.

## Conclusions

### Final Insights and Recommendations

* + 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.

* + 1. **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.

* + 1. **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.