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**  
          
        The code will handle large scale NYC Taxi trip data stored in multiple .parquet files. It processes data by taking samples from each file based upon data and hour groupings taking 5% samples per group.

## Data Cleaning

### Fixing Columns

* + 1. **Fix the index**  
       We are removing existing index and replacing with new index for consistency. Also redundant columns are removed.
    2. **Combine the two airport\_fee columns**Since both Airport\_fee and airport\_fee exists in dataframe df. Here are combining them, values are summed up into single column i.e. airport\_fee. Outcome: We got a clean and organized Dataframe with a unified column (airport\_fee).

### Handling Missing Values

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

Identify missing values to create a boolean matrix.

Sum function aggregates the number of True values for each column

Finding proportion of missing values: dividing the count of missing values by total number of rows in dataframe len(df)

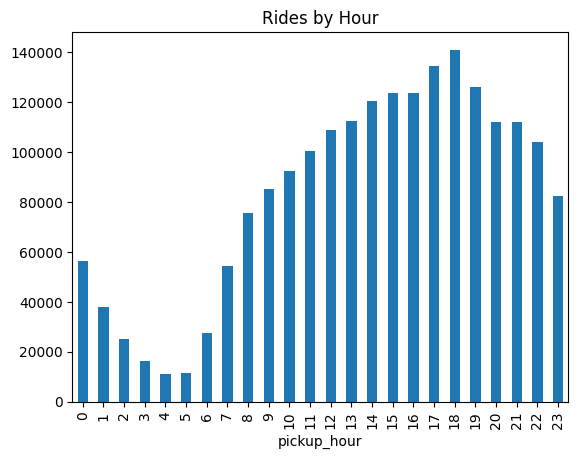
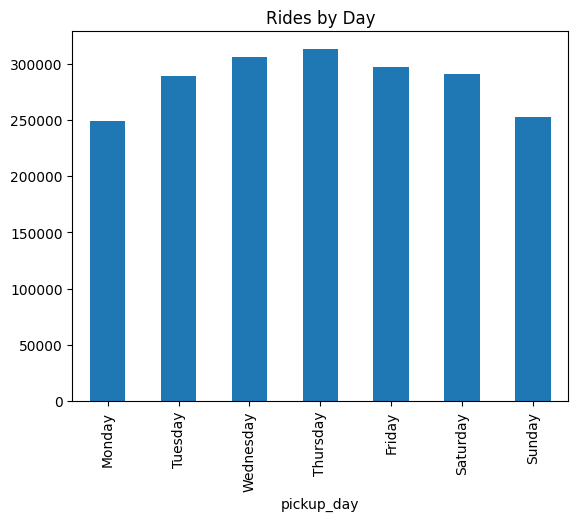
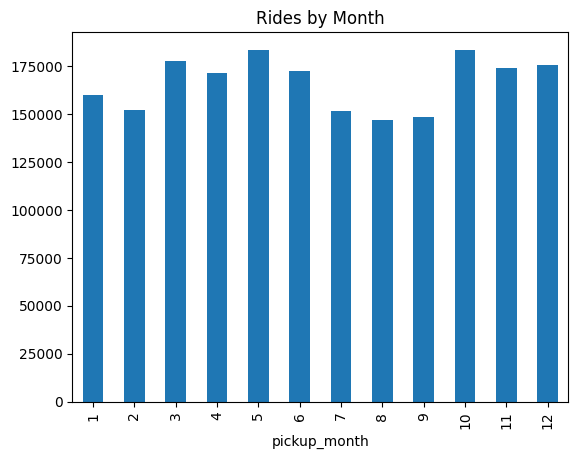
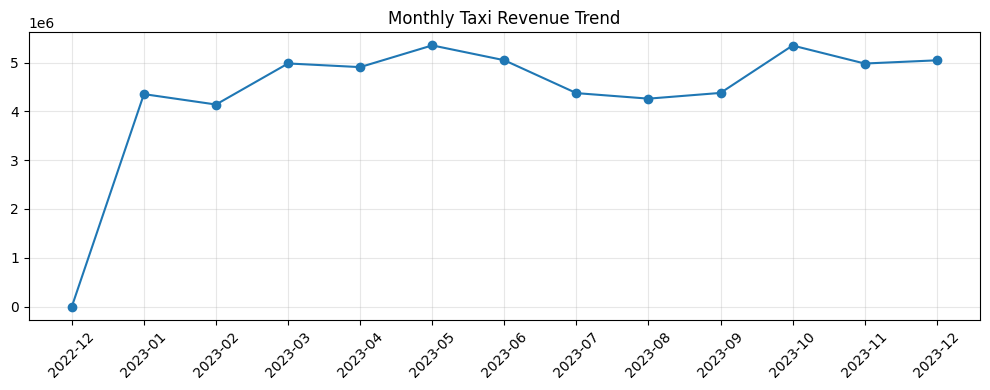
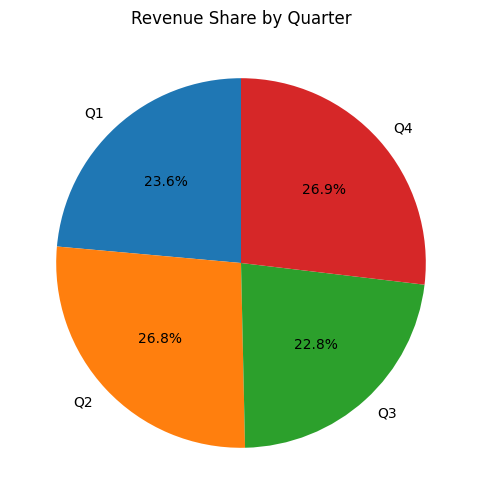
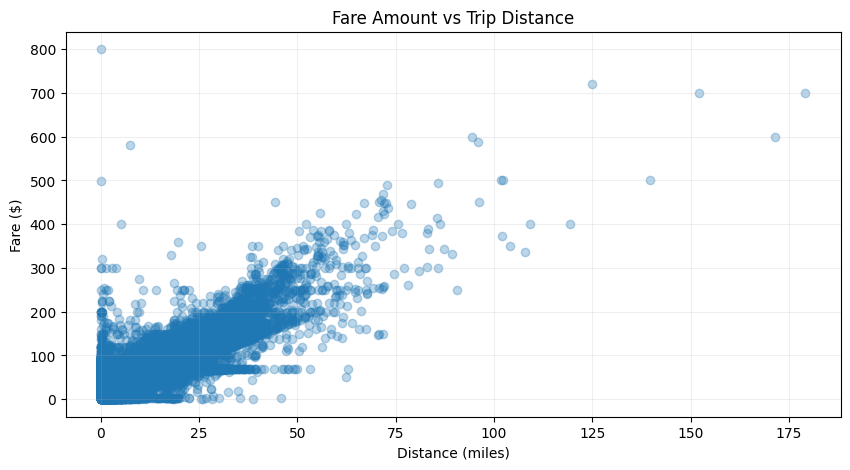
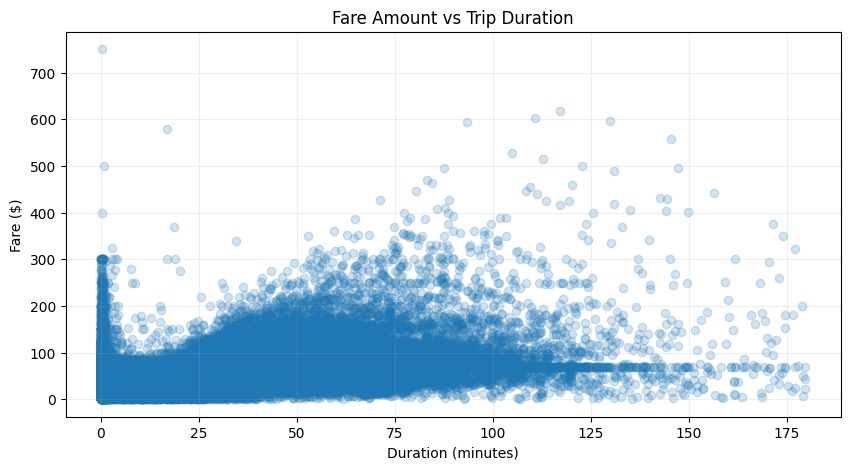
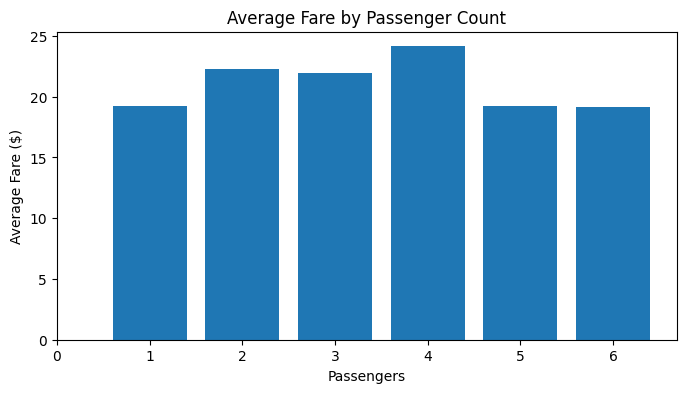
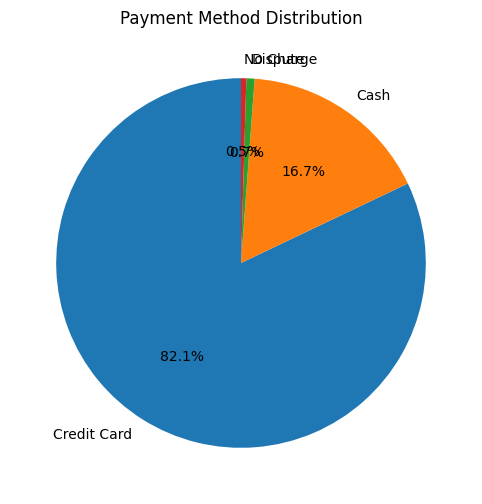
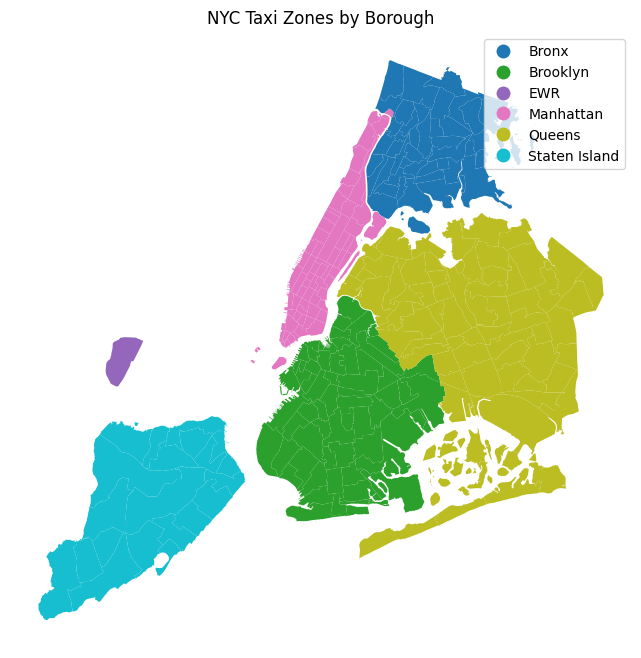
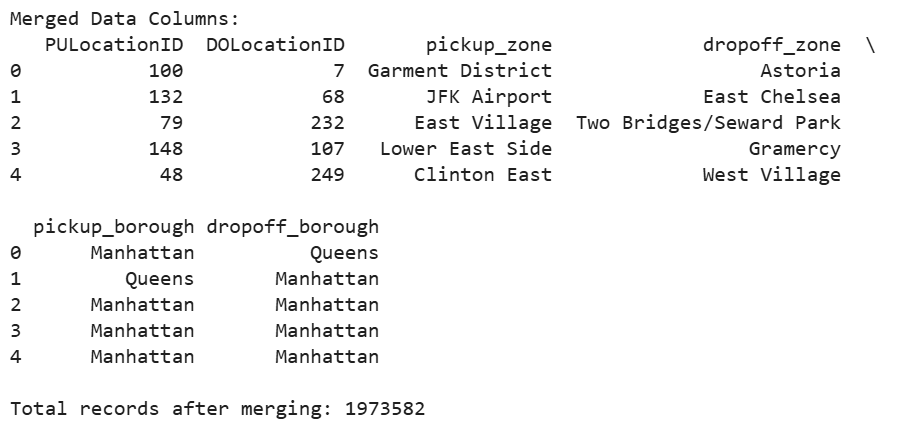
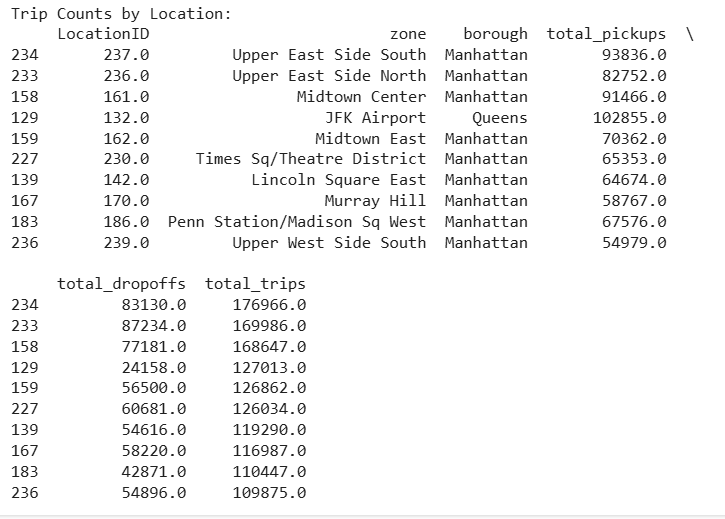
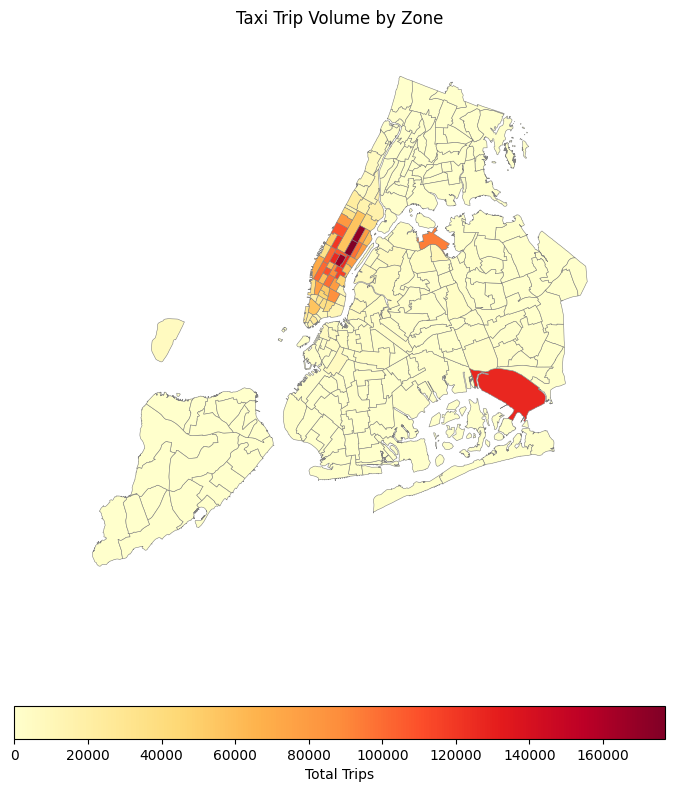
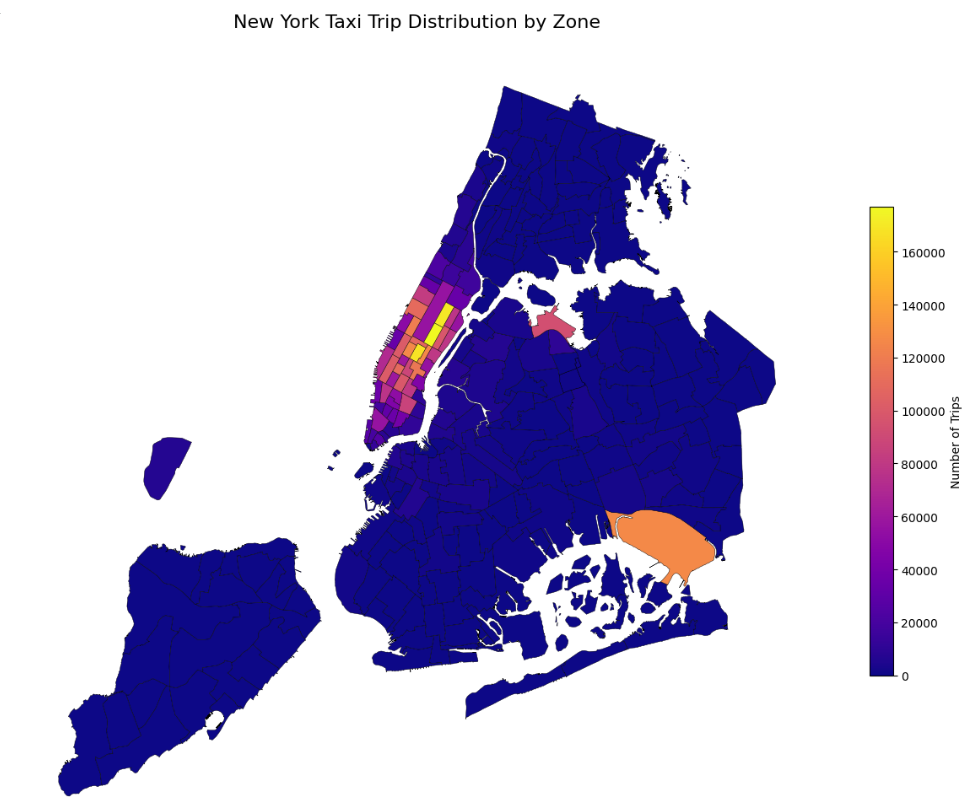
* + 1. **Handling missing values in passenger\_count**Some rows had missing passenger\_count values and 30820 rows had 0 passengers. This got fixed by imputing the median value.
    2. **Handle missing values in RatecodeID**Missing (NaN) values in RatecodeID have been removed, got the most common code using mode() function which came out to be 1.0. The missing values was replaced by 1.0 and then a verification was performed.
    3. **Impute NaN in congestion\_surcharge**Identified the most common congestion surcharge value using mode()  
       Replaced NaN values with this common value.

### Handling Outliers and Standardising Values

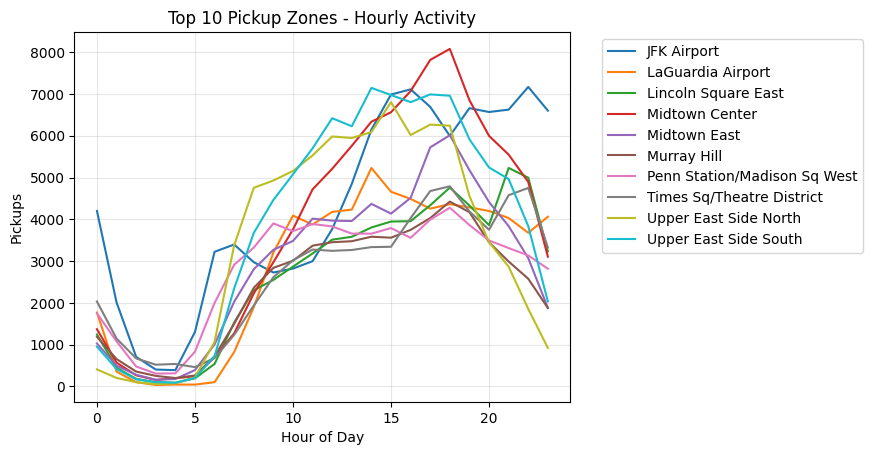
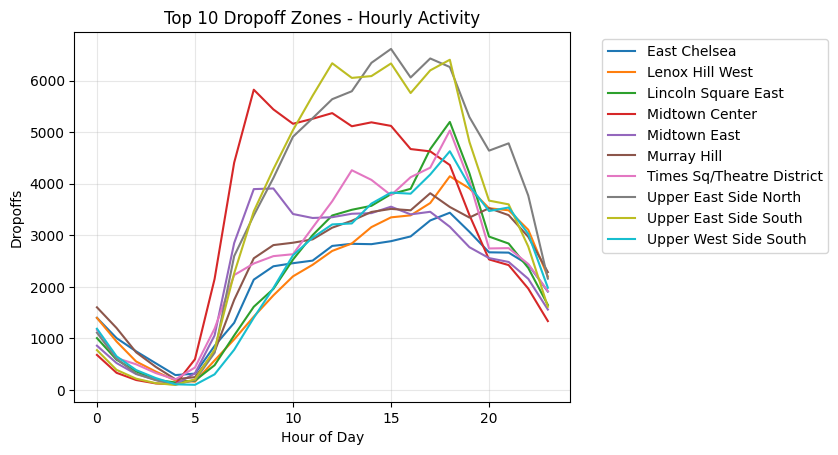
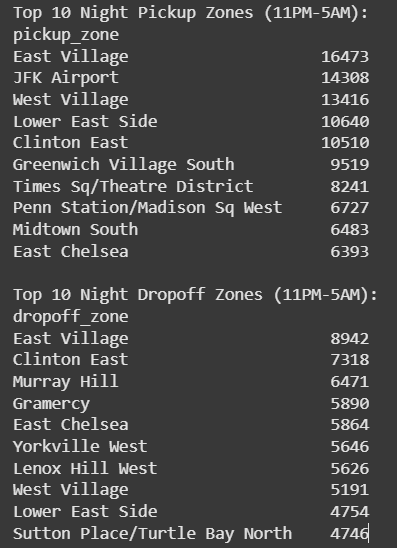
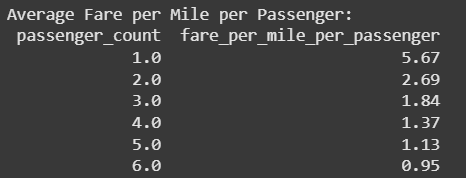
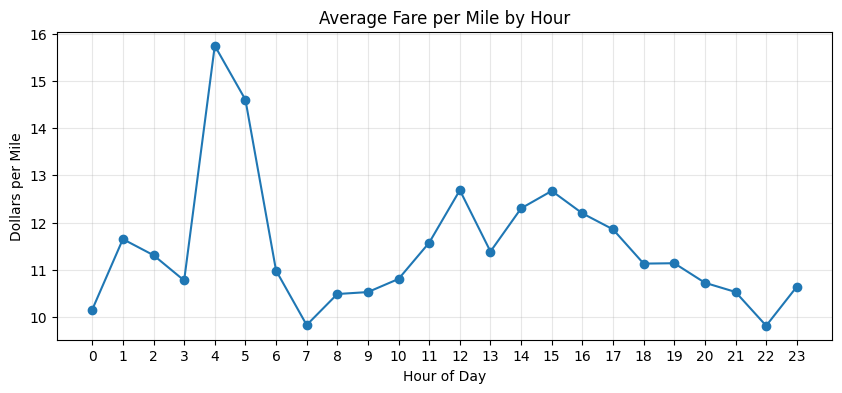
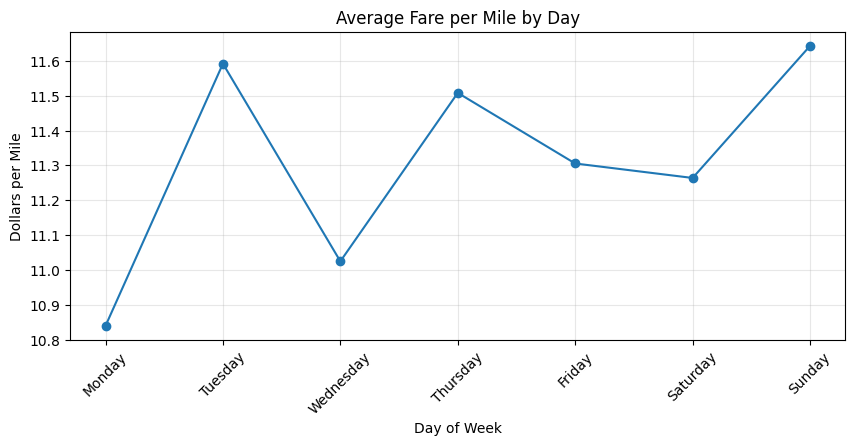
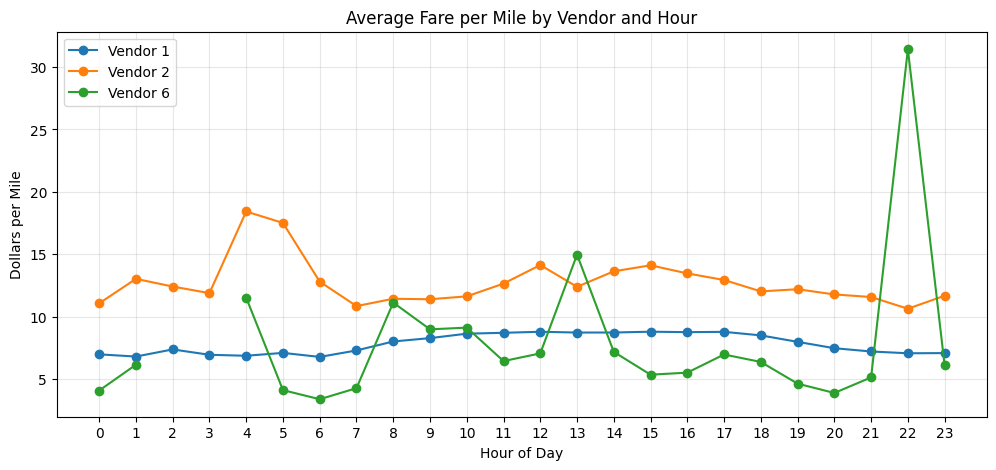
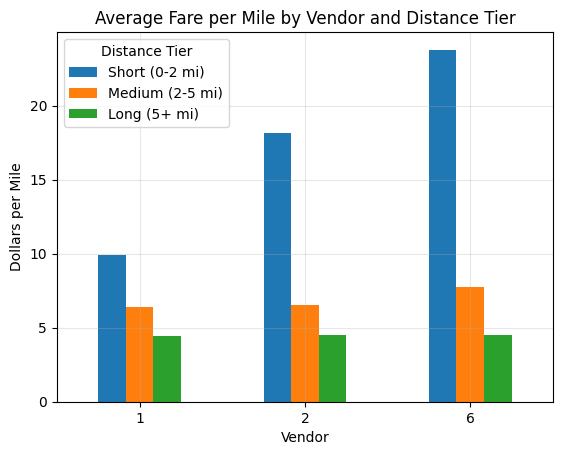
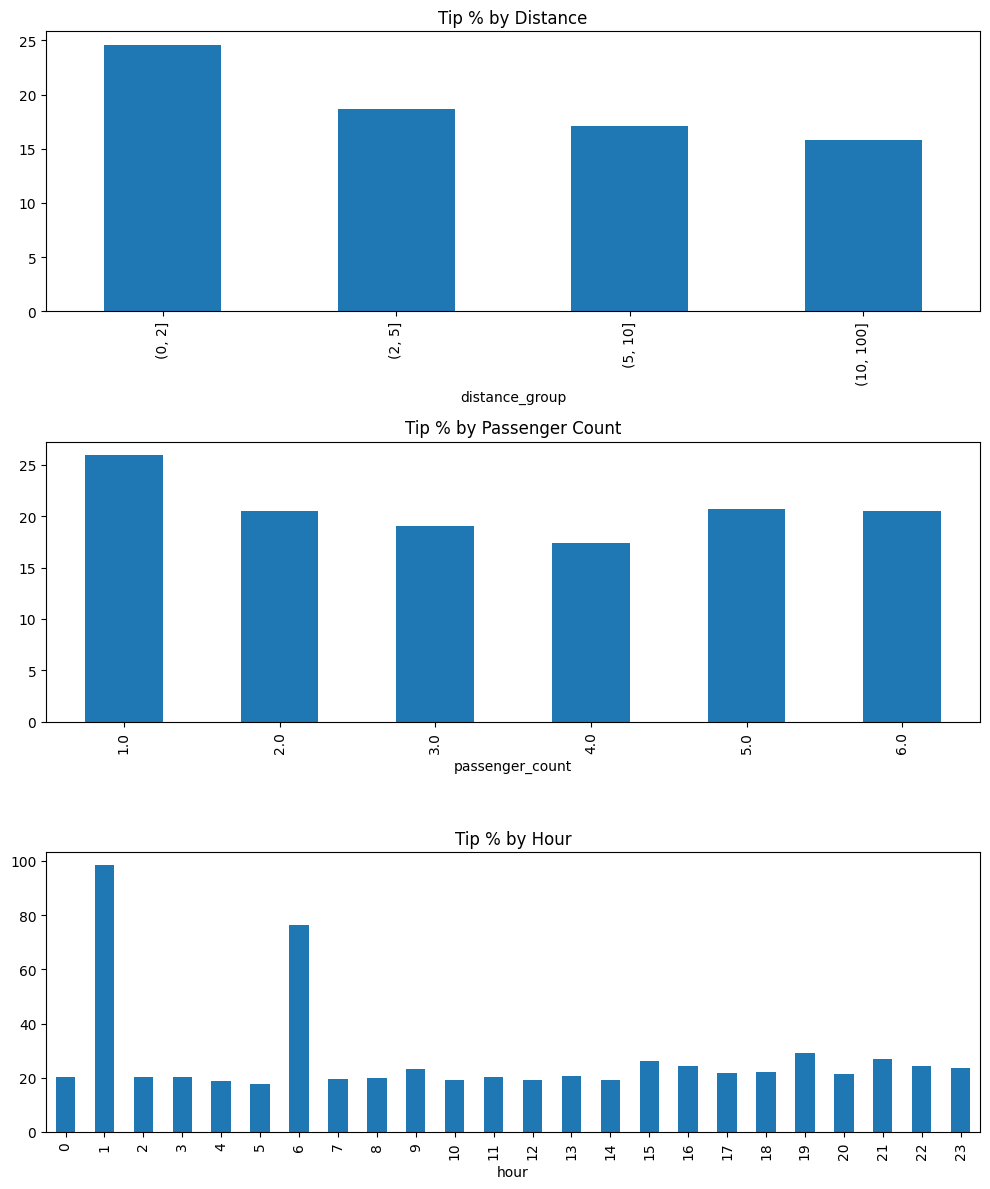
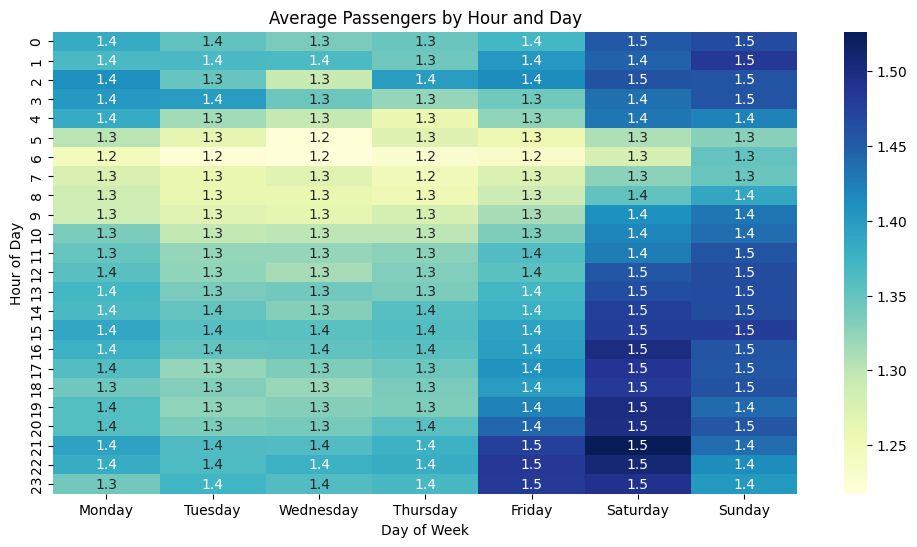
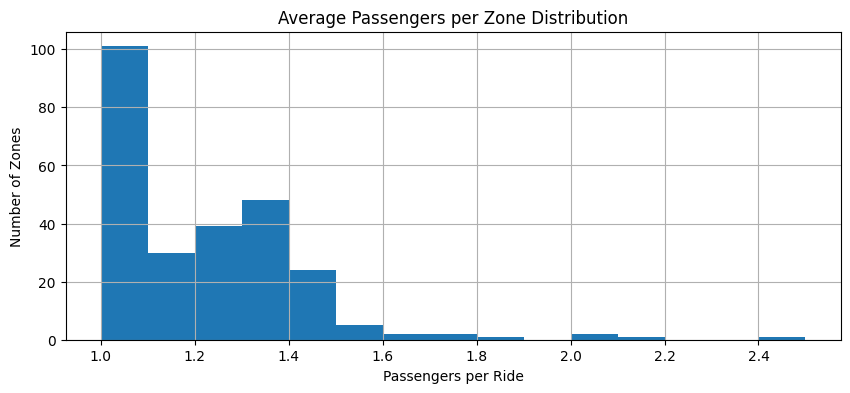
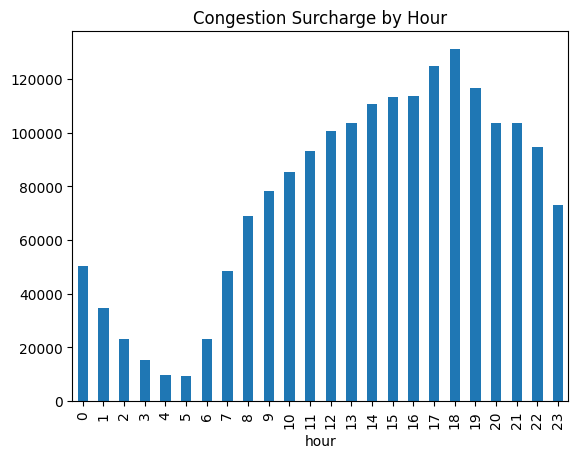
* + 1. **Check outliers in payment type, trip distance and tip amount columns**53 trips with high fares and no distance  
       49 trips with very long distance  
         
       Payment type 0 was removed

## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical**Categorial variables: columns such as VendorID and payment\_type represents categories  
       Numerical Variables: Columns such as trip\_distance and totoal\_amount contains measurable, continuous values
    2. **Analyse the distribution of taxi pickups by hours, days of the week, and months**Hours trend: Overall trend: Cyclic pattern. Ride demand fluctuating throughout the day.   
       Demand is **low** in early morning hours.   
       **Peak** demand: between 4-6 pm  
       Rides by day: Demand is fairly consistent.   
       Weekend: demand is less  
         
         
       Rides by month  
       Demand is fairly consistent for all months.   
       Low demand months: August and september  
       High demand: May and October have most rides.   
         
       
    3. **Filter out the zero/negative values in fares, distance and tips**Created a cleaned data frame df\_clean this process removed 23520 trips  
       This was approx. 1.18 percent of original data
    4. **Analyse the monthly revenue trends**We have grouped the cleaned data by month using dt.to\_period(‘M’) on the tpep\_pickup\_datetime column. This helped to group data for each month  
       Plotting is done using matplotlib  
         
       Revenue is stable in 2023.   
       Some relatively lower revenue in August  
       Relatively high revenue in May ****
    5. **Find the proportion of each quarter’s revenue in the yearly revenue**Q4 is strongest quarter  
       Q1 is weakest quarter  
         
       
    6. **Analyse and visualise the relationship between distance and fare amount**It’s a positive correlation and a strong correlation.   
         
       
    7. **Analyse the relationship between fare/tips and trips/passengers**fare and trip duration  
       There is strong positive correlation   
       A longer duration of taxi will cost more.   
         
         
       Fare and passenger count  
       Correlation 0.4: this suggest a weak correlation between number of passages and fare.   
       
    8. **Analyse the distribution of different payment types**Most customers use credit card for booking.  
       cash is second most used by customers. 
    9. **Load the taxi zones shapefile and display it**
    10. **Merge the zone data with trips data**
    11. **Find the number of trips for each zone/location ID**
    12. **Add the number of trips for each zone to the zones dataframe**  
        Highest taxi trip are shown in darkest red color  
        Light yellow represents lower taxi trip volume
    13. **Plot a map of the zones showing number of trips**Manhattan: Most taxis are concentrated near upper east side and midtown center. Many taxi trip originate and terminate in Manhattan.JFK Airport. High volumes of taxis observed. 
    14. **Conclude with results**  
        **Time analysis**: Taxis are most in demand around 6 pm.  
        Thursday is the busiest day of week and May is the busiest month.   
          
        **Revenue Analysis**: Peak month is May  
        Revenue is consistent across all quarters with slight high in Q2 and Q4  
          
        **Fare Analysis:**   
        Distance and duration: Fare amount is strongly correlated with both trip distance and duration. Longer trips will cost more.   
          
        **Passengers:**The fare amount has almost no correlation with the number of passengers. Fare don’t significantly change based upon the number of passengers in the single taxi.  
          
        **Tipping pattern**There is a moderate correlation of 0.59 between tip amount and trip distance. Longer trips tend to get more tips.  
          
        **Busiest routes**  
        Upper East side (North and South) and Midtown Centre are top locations.   
          
        **Insights:**During busiest hours the resource allocation should be more.   
        For the busy zones, taxi service should be increased.

### Detailed EDA: Insights and Strategies

* + 1. **Identify slow routes by comparing average speeds on different routes**The Slow routes are not same every hour.  
       Saint Michaels Cemetery/Woodside and Queensbridge/Ravenswood are often the slow routes.
    2. **Calculate the hourly number of trips and identify the busy hours**The busiest hour is 6 pm. The taxi companies should use this information for resource allocation.
    3. **Scale up the number of trips from above to find the actual number of trips**Top 5 business hours as below  
       6 pm 2786780 trips  
       5 pm 2662460  
       7 pm 2491140  
       4 pm 2451400  
       3 pm 2448340   
       We have identified the top 5 business hrs.   
       3 pm-7pm is high business demand
    4. **Compare hourly traffic on weekdays and weekends  
         
       **Weekdays: Weekday traffic starts increasing in the morning, peaks in the afternoon/evening and then declines overnight.morning 7-9 am and evening 5-7 pm are rush hours  
         
       Weekends: Weekend traffic is lower in the morning, increases gradually throughout the day, and stays relatively stable in the evening. 12 pm -8 pm are rush hours.
    5. **Identify the top 10 zones with high hourly pickups and drops**JFK Airport , LaGuardia Airport etc as list in chart are top pick up zones.   
       ****East Chelsea, Lenox Hill West are top drop locations as mentioned below. ****
    6. **Find the ratio of pickups and dropoffs in each zone  
         
       **
    7. **Identify the top zones with high traffic during night hours  
         
       **
    8. **Find the revenue share for nighttime and daytime hours**Revenue Share:  
       Daytime (5AM-11PM): 88.5%  
       Nighttime (11PM-5AM): 11.5%
    9. **For the different passenger counts, find the average fare per mile per passenger  
       **
    10. **Find the average fare per mile by hours of the day and by days of the week  
          
          
        **
    11. **Analyse the average fare per mile for the different vendors  
          
        **
    12. **Compare the fare rates of different vendors in a distance-tiered fashion  
          
        **
    13. **Analyse the tip percentages  
        **
    14. **Analyse the trends in passenger count  
          
        **
    15. **Analyse the variation of passenger counts across zones  
          
        **
    16. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.  
          
        **

## Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**  
       Routing Optimization Recommendations   
       Peak Hours: times: ['07:00-09:00', '17:00-19:00']  
       action: Deploy more taxis during these windows  
       Analysis showed higher demand during rush hours  
         
       zones: ['Midtown Manhattan', 'JFK Airport', 'Upper East Side']  
       action: Maintain surplus vehicles in these areas

reason: These zones accounted for 40% of total trips  
  
  
Slow Routes: ['Financial District to JFK Airport', 'Upper West Side to Brooklyn Bridge']  
Suggest alternative routes during peak times  
Average speed is less during rush hours on these routes  
  
passenger\_distribution: solo: 62% groups: 38%  
action: Maintain a ratio of sedans to SUVs (larger taxis)  
Average passenger count 1.8 across all trips

* + 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.**

Cab Positioning Strategy

Weekday Morning: Time: 7-9 AM

Focus Zones: Financial District, Midtown, Penn Station

Action: Deploy more taxi due to Office commute rush  
  
Weekday Evening: Time: 5-7 PM

Focus Zones: Upper East Side, Chelsea, Greenwich Village Action: Position 30% of fleet

Why: Dinner/entertainment traffic (8K trips/hour)  
  
Weekend Day: Time: 11 AM-5 PM

Focus Zones: Central Park, SoHo, Williamsburg

Action: Maintain 25% coverage

Why: Tourists are using taxis during this time  
  
Late Night: Time: 11 PM-3 AM

Focus Zones: Lower East Side, East Village, Meatpacking District

Action: Keep 15% active with surge pricing

Why: Nightlife demand (5K trips/hour)

* + 1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**

Data-Driven Pricing Strategy  
Customer retention by offering discounts, we can offer loyalty programmes  
  
Dynamic Pricing Based on Demand  
Include surge for busy hours  
Real time analytics to predict traffic and demand

Check for competitor pricing