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.

## Data Preparation

* 1. Loading the dataset
     1. **Sample the data and combine the files**  
          
        Done in Jupyter Notebook.

## Data Cleaning

### Fixing Columns

* + 1. **Fix the index**  
         
       Done in Jupyter Notebook.
    2. **Combine the two airport\_fee columns**Done in Jupyter Notebook.

### Handling Missing Values

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

Done in Jupyter Notebook.

* + 1. **Handling missing values in passenger\_count**

Done in Jupyter Notebook.

* + 1. **Handle missing values in RatecodeID**  
       Done in Jupyter Notebook.
    2. **Impute NaN in congestion\_surcharge**  
        Done in Jupyter Notebook.

### Handling Outliers and Standardising Values

**Check outliers in payment type, trip distance, and tip amount columns**  
Potential outliers detected:

VendorID: 444 outliers

passenger\_count: 424621 outliers

trip\_distance: 249302 outliers

RatecodeID: 102267 outliers

payment\_type: 404026 outliers

fare\_amount: 197413 outliers

extra: 34047 outliers

mta\_tax: 17944 outliers

tip\_amount: 145673 outliers

tolls\_amount: 153802 outliers

improvement\_surcharge: 2259 outliers

total\_amount: 218083 outliers

congestion\_surcharge: 140898 outliers

airport\_fee: 160962 outliers

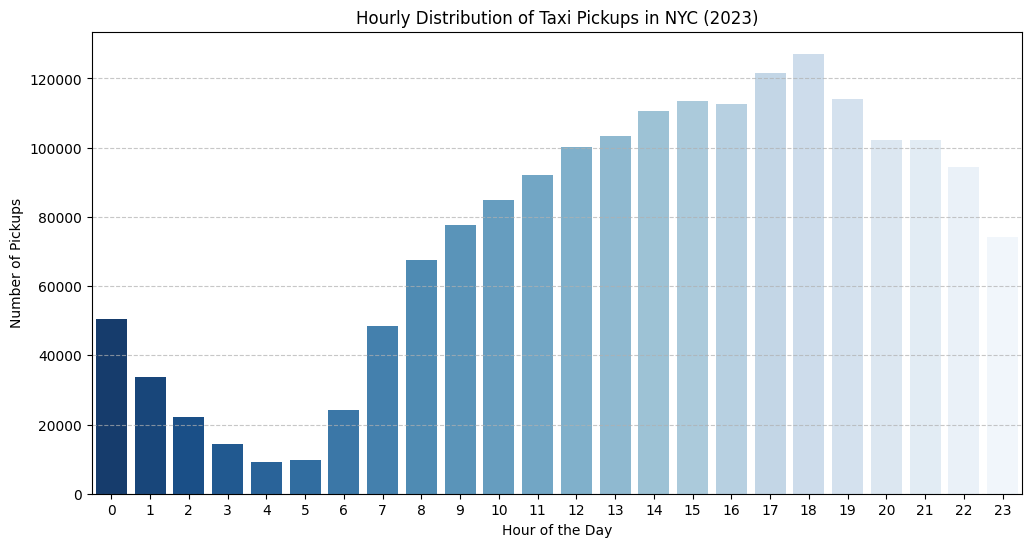
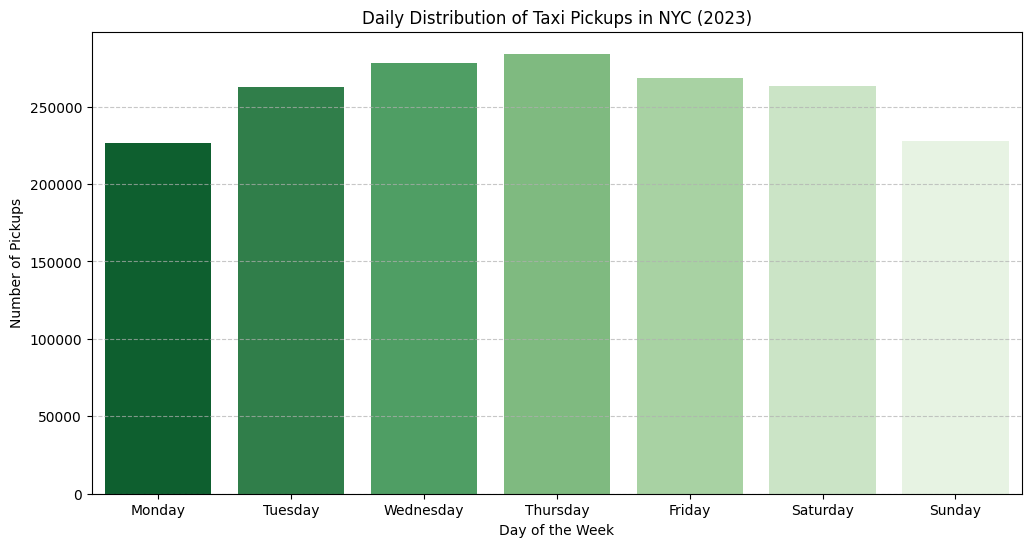
## Exploratory Data Analysis

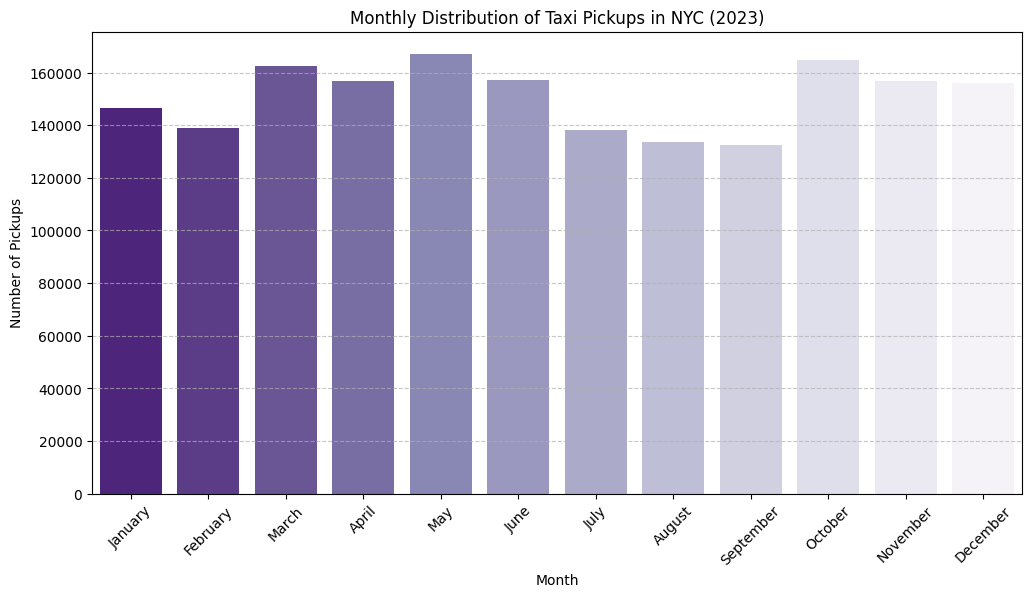
### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical  
         
       Numerical Variables (Continuous or Discrete)**

1. **passenger\_count**
2. **trip\_distance**
3. **fare\_amount**
4. **extra**
5. **mta\_tax**
6. **tip\_amount**
7. **tolls\_amount**
8. **improvement\_surcharge**
9. **total\_amount**
10. **congestion\_surcharge**
11. **airport\_fee**
12. **pickup\_hour**
13. **trip\_duration (if calculated from pickup and dropoff times)**

### **Categorical Variables**

1. **VendorID**
2. **RatecodeID**
3. **store\_and\_fwd\_flag**
4. **PULocationID**
5. **DOLocationID**
6. **payment\_type**
7. **pickup\_date (as a categorical time variable)**
8. **pickup\_hour (if treated as time category instead of numerical)**
   * 1. **Analyze the distribution of taxi pickups by hours, days of the week, and months  
          
        **

****

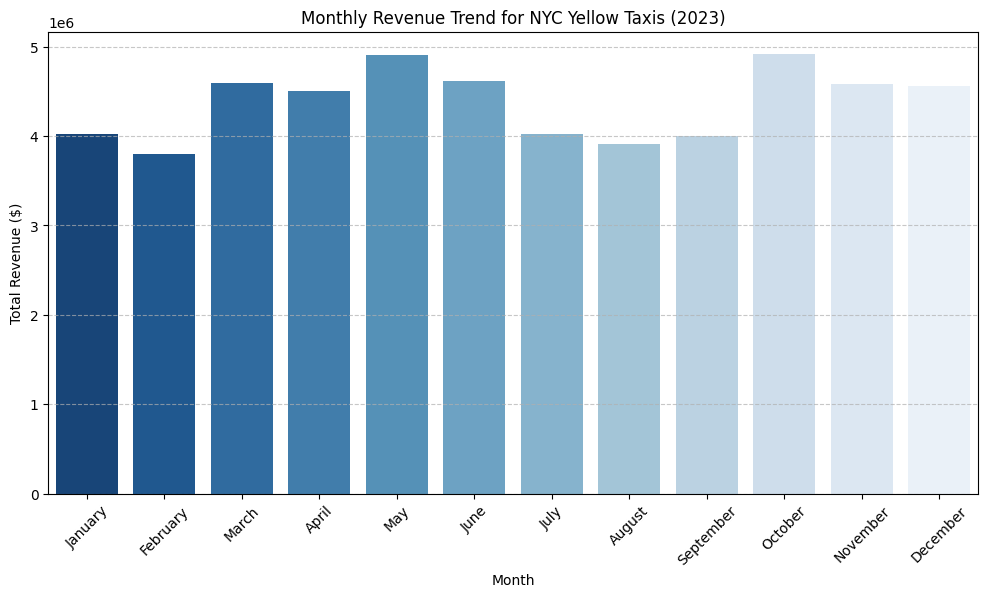
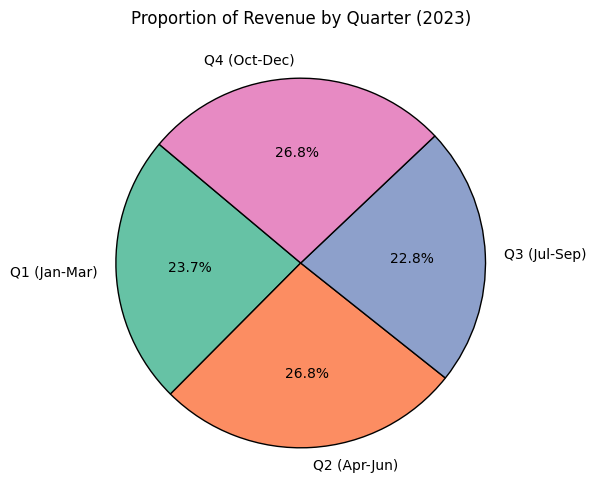
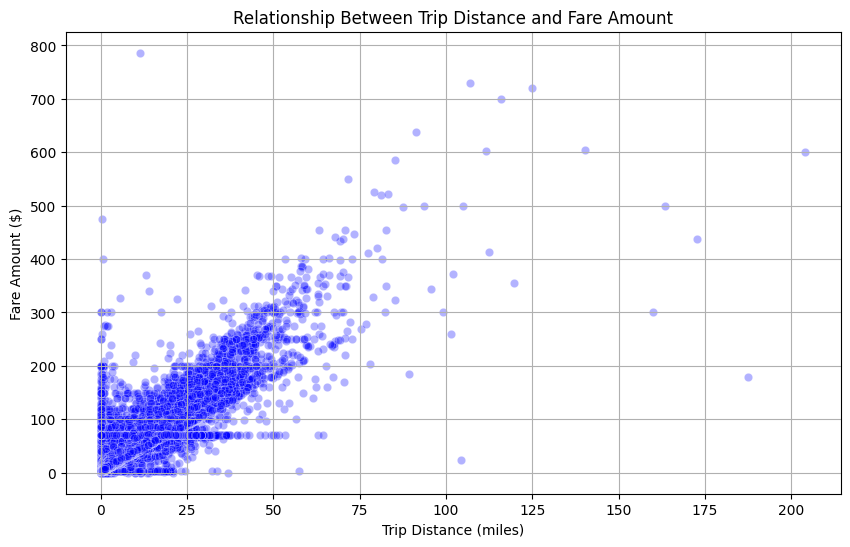
**Filter out the zero/negative values in fares, distance, and tips  
  
Count of Zero/Negative Values in Financial Parameters & Trip Distance:**

**fare\_amount 0**

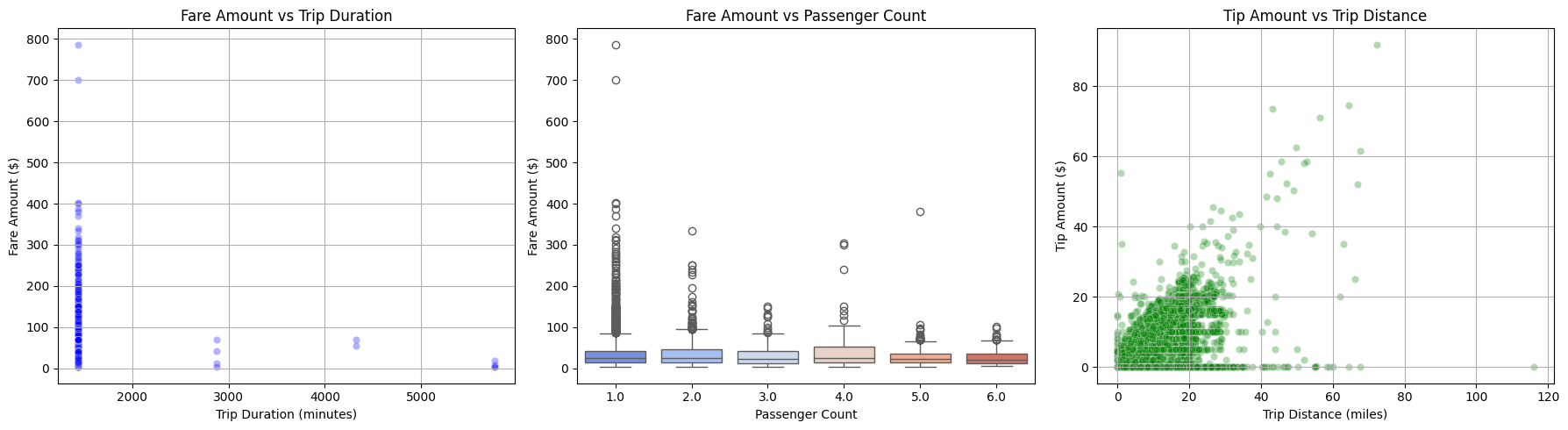
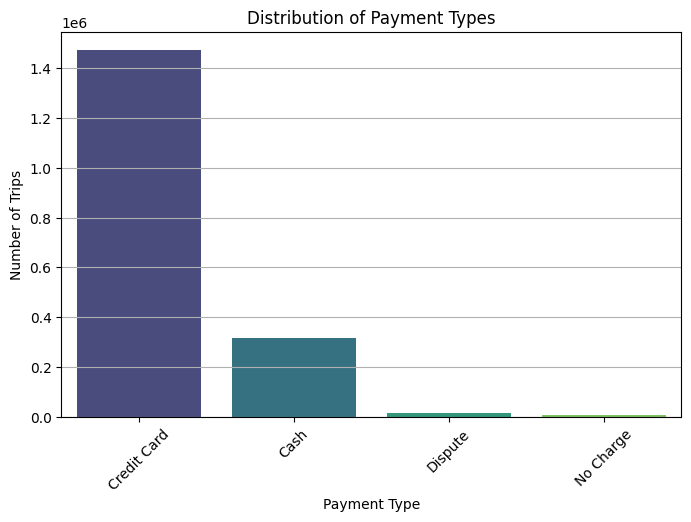
**tip\_amount 409675**

**total\_amount 0**

**trip\_distance 21879**

* + 1. **Analyze the monthly revenue trends  
         
       **
    2. **Find the proportion of each quarter’s revenue in the yearly revenue  
       **
    3. **Analyse and visualize the relationship between distance and fare amount**

**Correlation between Trip Distance and Fare Amount: 0.9442**

* + 1. **Analyze the relationship between fare/tips and trips/passengers  
       **
    2. **Analyze the distribution of different payment types  
         
       **

**Payment Type Distribution:**

**payment\_type**

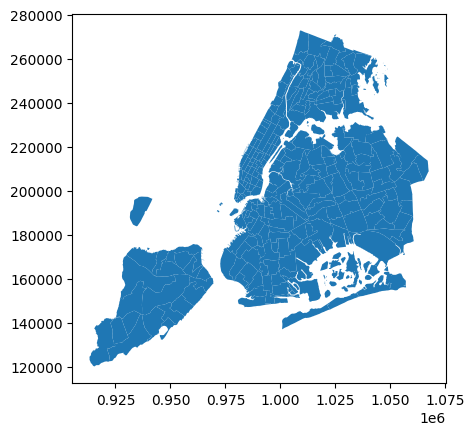
**Credit Card 1472672**

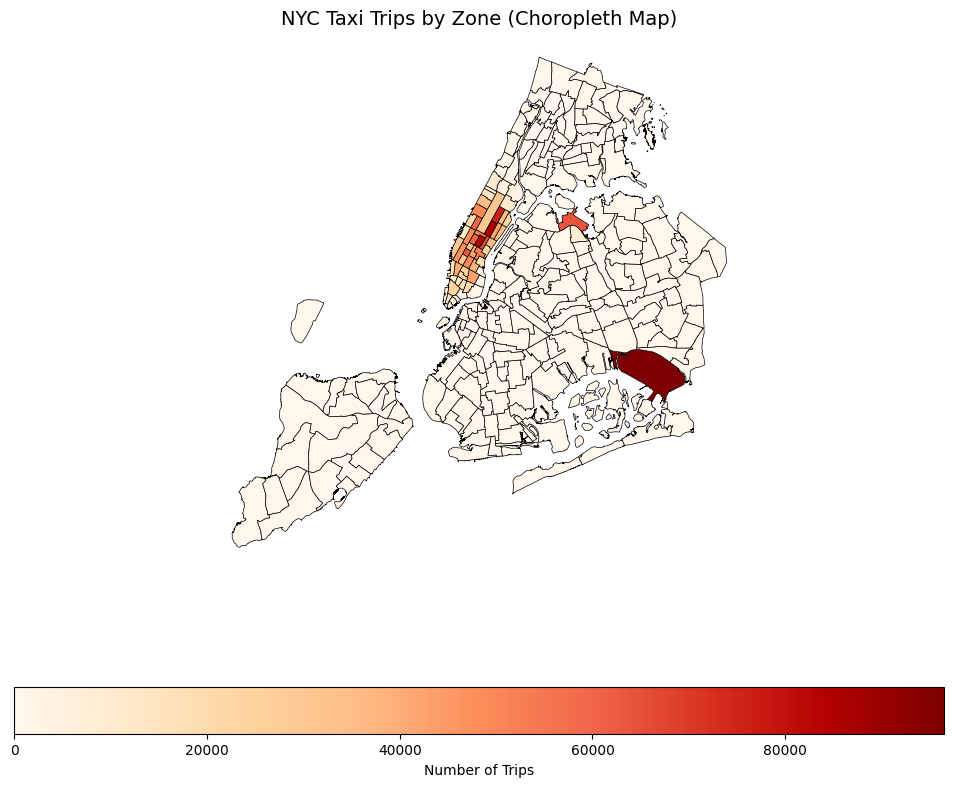
**Cash 316180**

**Dispute 13563**

**No Charge 8883**

* + 1. **Load the taxi zones shapefile and display it  
         
       Done in Jupyter Notebook.**

****

* + 1. **Merge the zone data with trips data  
         
       Done in Jypyter Notebook.**
    2. **Find the number of trips for each zone/location ID  
         
       Done in Jypyter Notebook. 257 rows × 2 columns**
    3. **Add the number of trips for each zone to the zones dataframe  
         
       Done in Jupyter Notebook.**
    4. **Plot a map of the zones showing the number of trips  
       **
    5. **Conclude with results**

**Temporal Analysis: Taxi Demand Over Time Busiest Hours: Peak demand is observed between 6 PM – 8 PM, likely due to evening rush hours. Busiest Days: Fridays and Saturdays show the highest number of trips, suggesting increased activity during weekends. Busiest Months: May and December had the most taxi rides, possibly due to tourism and holiday seasons.**

**💰 Financial Analysis: Revenue & Fare Trends Revenue Trends: Total earnings peaked during Q2 (April-June) and Q4 (Oct-Dec), aligning with increased seasonal demand.**

**Quarterly Revenue Breakdown:**

**Q1: Lowest revenue, likely due to colder months and reduced travel. Q2 & Q4: The highest revenue is due to increased tourism and holiday shopping. Trip Distance vs. Fare:**

**Strong positive correlation—longer trips tend to have higher fares. Some outliers exist where short distances have very high fares, possibly due to tolls or fixed pricing. Trip Duration vs. Fare:**

**Moderate correlation—longer trip durations lead to higher fares, but exceptions exist due to traffic congestion. Passenger Count vs. Fare:**

**No strong correlation—fares depend more on distance and duration than passenger count. Tip Amount vs. Trip Distance:**

**Positive relationship—longer trips generally yield higher tips, possibly due to better service experience.**

**🗺️ Geographical Analysis: Pickup Hotspots & Busiest Zones Busiest Pickup Locations:**

**Manhattan dominates taxi activity, especially near Midtown, Times Square, and financial districts. Airports (JFK & LaGuardia) show consistently high trip counts. Least Busy Zones:**

**Peripheral areas like Staten Island and some parts of The Bronx have significantly lower demand.**

**🔍 Key Insights & Recommendations Optimize Fleet Availability**

**Deploy more taxis during evening rush hours and weekend nights to maximize earnings. Dynamic Pricing Strategies**

**Higher fare rates could be applied in high-demand periods (weekends, holidays). Improving Service in Low-Demand Areas**

**Consider ride-sharing promotions or discounts in lesser-used boroughs to increase demand. Encourage Digital Payments & Tipping**

**Since tip amount increases with trip distance, promotions on cashless payments may boost overall earnings.**

### Detailed EDA: Insights and Strategies

* + 1. **Identify slow routes by comparing average speeds on different routes Done in Jupyter**

**PULocationID DOLocationID pickup\_hour avg\_trip\_distance \**

**16171 62 62 23 0.550000**

**62517 142 142 5 0.560000**

**123095 261 13 23 0.600000**

**17317 67 122 23 0.600000**

**24765 75 264 22 0.780000**

**1914 12 87 23 0.800000**

**6735 40 65 21 1.120000**

**97252 226 145 18 1.563333**

**8123 42 166 23 1.240000**

**87137 181 65 23 1.250000**

**avg\_trip\_duration avg\_speed\_mph**

**16171 0.400000 1.37500**

**62517 0.400000 1.40000**

**123095 0.400000 1.50000**

**17317 0.400000 1.50000**

**24765 0.400000 1.95000**

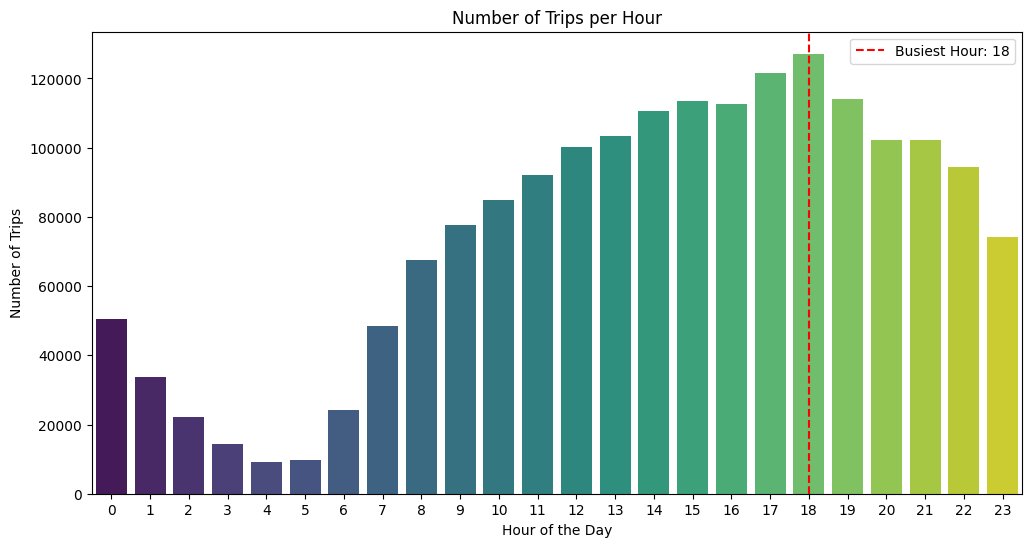
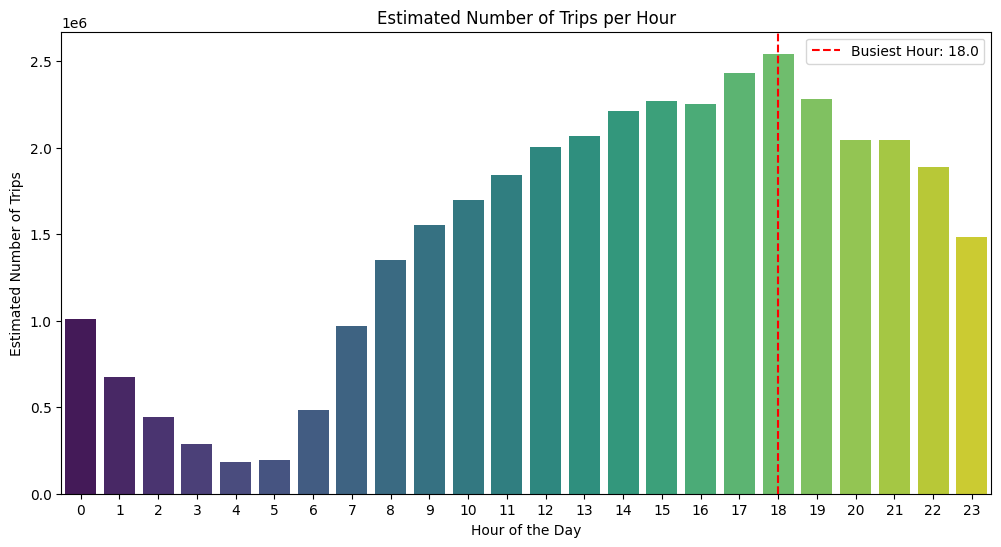
**1914 0.400000 2.00000**

**6735 0.400000 2.80000**

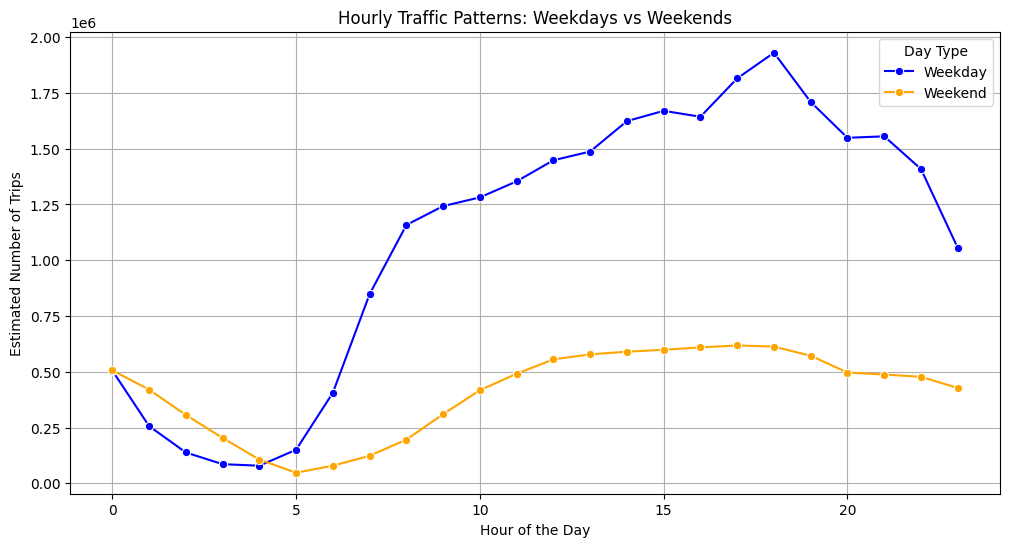
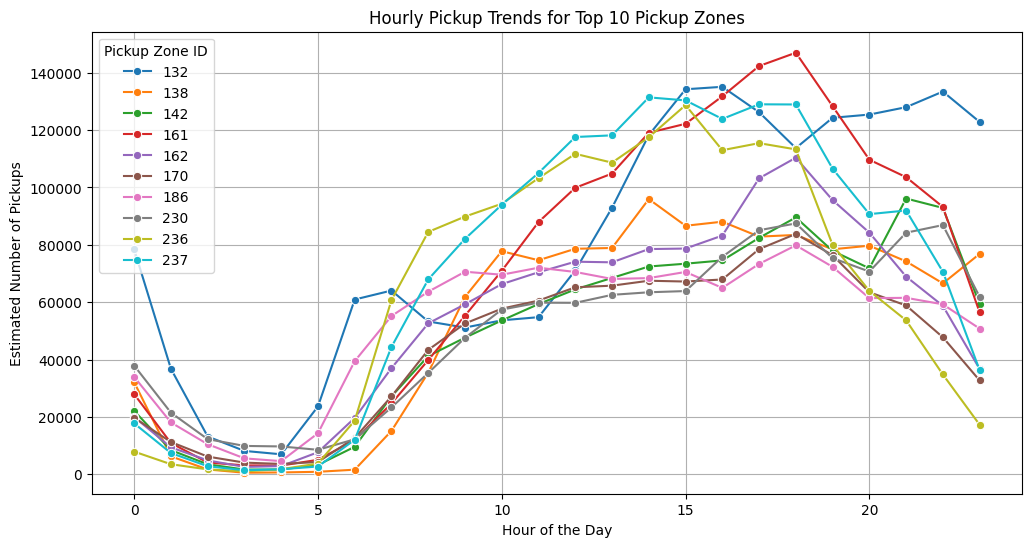
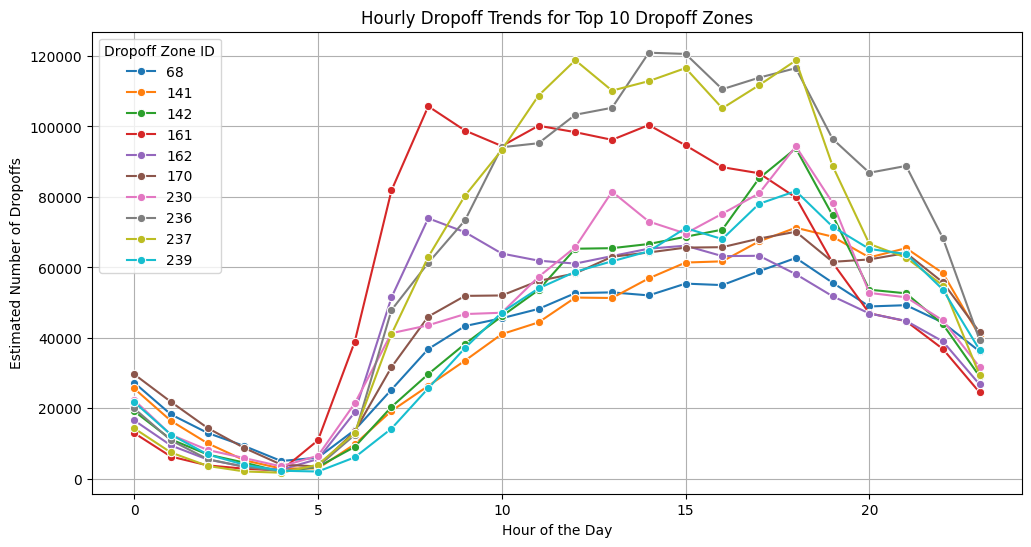
**97252 0.533333 2.93125**

**8123 0.400000 3.10000**

**87137 0.400000 3.12500**

* + 1. **Calculate the hourly number of trips and identify the busy hours  
         
       Busiest Hour: 18 with 127151 trips.**
    2. **Scale up the number of trips from above to find the actual number of trips  
         
       **

**Busiest Hour: 18.0 with approximately 2543020 trips.**

* + 1. **Compare hourly traffic on weekdays and weekends  
         
       **
    2. **Identify the top 10 zones with high hourly pickups and drops  
       **
    3. **Find the ratio of pickups and dropoffs in each zone  
         
        Top 10 Pickup/Dropoff Ratio Zones:**

**pickup\_count dropoff\_count pickup\_dropoff\_ratio**

**70 8318.0 941.0 8.830149**

**132 96523.0 20854.0 4.628291**

**138 63943.0 22171.0 2.883953**

**199 2.0 0.0 2.000000**

**186 62929.0 39602.0 1.588996**

**43 30446.0 22103.0 1.377398**

**114 23865.0 17351.0 1.375346**

**249 39981.0 30047.0 1.330571**

**162 64946.0 51656.0 1.257255**

**161 85156.0 70917.0 1.200767**

**Bottom 10 Pickup/Dropoff Ratio Zones:**

**pickup\_count dropoff\_count pickup\_dropoff\_ratio**

**30 0.0 18.0 0.000000**

**99 0.0 3.0 0.000000**

**176 0.0 12.0 0.000000**

**245 0.0 30.0 0.000000**

**27 1.0 38.0 0.025641**

**221 1.0 34.0 0.028571**

**257 28.0 756.0 0.036988**

**1 208.0 5303.0 0.039216**

**115 1.0 23.0 0.041667**

**198 52.0 986.0 0.052685**

* + 1. **Identify the top zones with high traffic during night hours  
         
       Top 10 Nighttime Pickup Zones:**

**PULocationID**

**79 15379**

**132 14500**

**249 12362**

**48 10277**

**148 9545**

**114 8669**

**230 8084**

**186 6891**

**164 6072**

**68 5977**

**Name: night\_pickup\_count, dtype: int64**

**Top 10 Nighttime Dropoff Zones:**

**DOLocationID**

**79 8194**

**48 6741**

**170 6192**

**68 5758**

**107 5693**

**141 5194**

**263 4933**

**249 4859**

**230 4539**

**229 4324**

**Name: night\_dropoff\_count, dtype: int64**

* + 1. **Find the revenue share for nighttime and daytime hours  
         
       Nighttime Revenue: $6,336,169.24 (12.08%)**

**Daytime Revenue: $46,120,883.83 (87.92%**

* + 1. **For the different passenger counts, find the average fare per mile per passenger  
         
       passenger\_count fare\_per\_mile\_per\_passenger**

**0 1.0 10.859230**

**1 2.0 6.447111**

**2 3.0 3.911747**

**3 4.0 4.394022**

**4 5.0 1.705503**

**5 6.0 1.346712**

* + 1. **Find the average fare per mile by hours of the day and by days of the week  
         
       Average Fare Per Mile by Day of the Week:**

**pickup\_day avg\_fare\_per\_mile**

**0 0 10.998590**

**1 1 11.392215**

**2 2 11.112610**

**3 3 11.274470**

**4 4 10.906168**

**5 5 10.923219**

**6 6 12.548744**

**Average Fare Per Mile by Hour of the Day:**

**pickup\_hour avg\_fare\_per\_mile**

**0 0 10.532263**

**1 1 11.268722**

**2 2 9.899217**

**3 3 10.873669**

**4 4 14.097316**

**5 5 14.310415**

**6 6 11.278070**

**7 7 10.305061**

**8 8 10.388318**

**9 9 10.379959**

**10 10 10.689135**

**11 11 10.967440**

**12 12 12.079546**

**13 13 12.015500**

**14 14 11.626542**

**15 15 12.500803**

**16 16 13.946956**

**17 17 11.967907**

**18 18 11.556238**

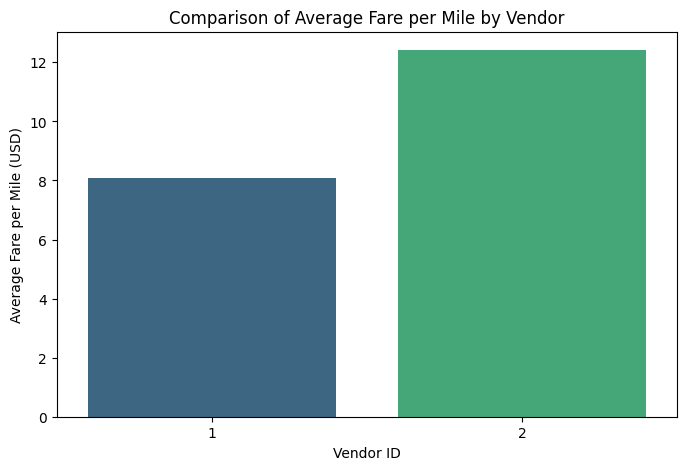
**19 19 11.629026**

**20 20 9.575345**

**21 21 9.512710**

**22 22 10.151658**

**23 23 10.738061**

* + 1. **Analyze the average fare per mile for the different vendors  
         
       **
    2. **Compare the fare rates of different vendors in a distance-tiered fashion**

**Average Fare Per Mile by Vendor and Distance Tier:**

**VendorID distance\_tier avg\_fare\_per\_mile**

**0 1 0-2 miles 9.899049**

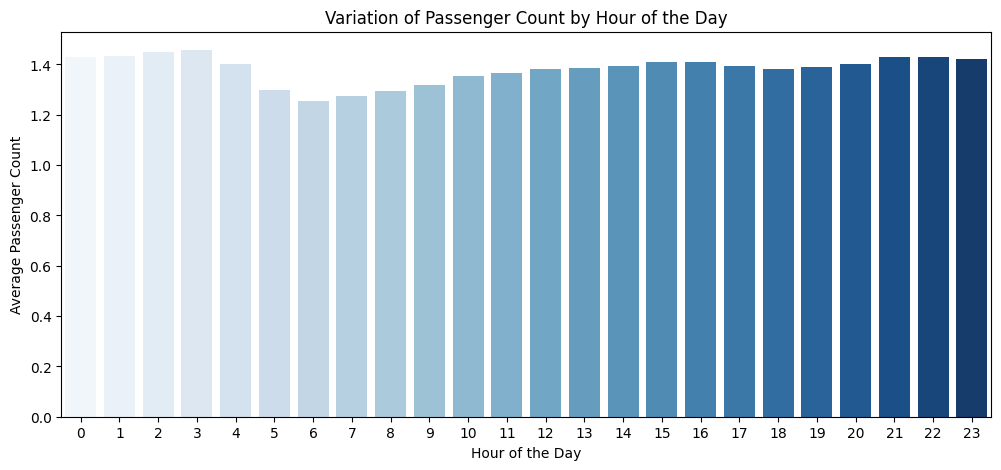
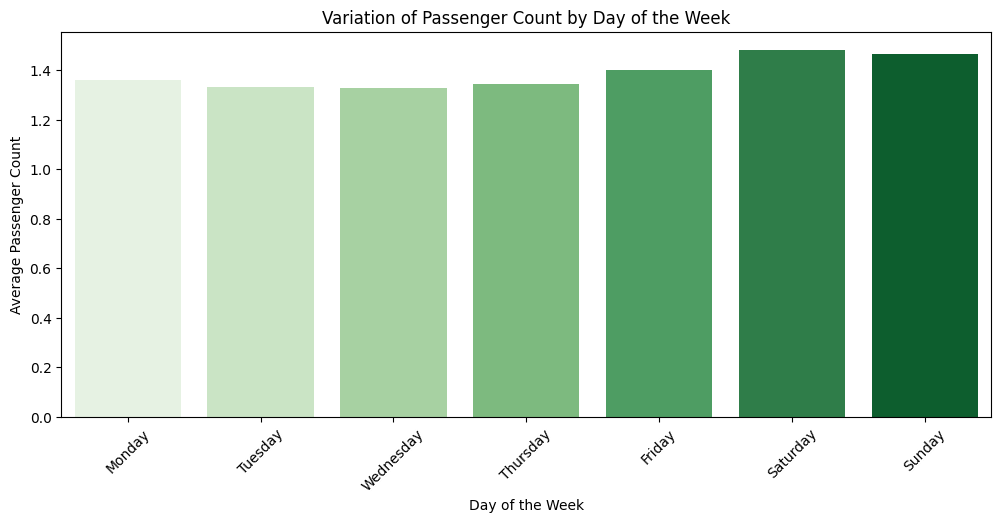
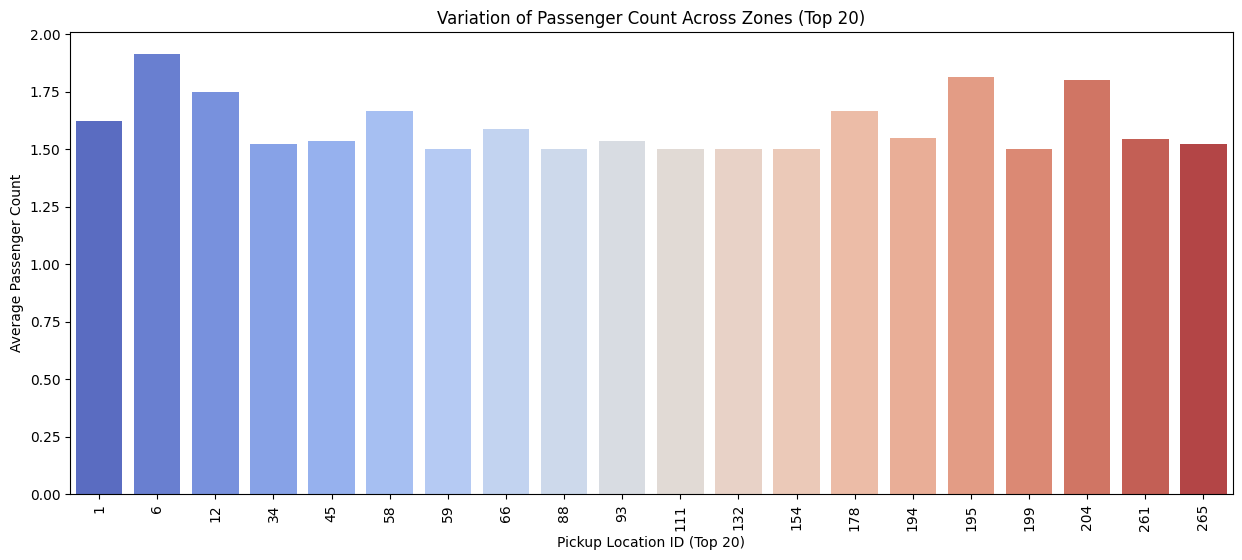
**1 1 2-5 miles 6.382780**

**2 1 5+ miles 4.425634**

**3 2 0-2 miles 18.029016**

**4 2 2-5 miles 6.549727**

**5 2 5+ miles 4.503693**

* + 1. **Analyze the tip percentages  
         
       Done in Jupyter Notebook, but infeasible to paste here because there are too many rows.**
    2. **Analyse the trends in passenger count  
         
       **
    3. **Analyze the variation of passenger counts across zones  
       **
    4. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.  
         
        Count Percentage**

**extra 1119710 61.813902**

**mta\_tax 1794860 99.085745**

**congestion\_surcharge 1671992 92.302783**

**airport\_fee 160459 8.858184**

**improvement\_surcharge 1811015 99.977587**

## Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**
* As can be seen in the visualization above the top 10 pickup zones have the highest passenger counts, we recommend having more yellow cabs deployed in these zones compared to other zones.
* The top ten slowest routes have been identified above, we should identify and recommend alternate routes to get to the drop-off locations to ease the traffic and seep up operational inefficiencies.
  + 1. **Suggestions on strategically positioning cabs across different zones to make the best use of insights uncovered by analyzing trip trends across time, days, and months.**
* There are more cab requests during weekdays compared to weekends, so there should be more cabs deployed during the weekdays in comparison to weekends.
* The busiest hours are 6 pm, 5 pm, 7 pm, 3 pm, and 4 pm. So, there should be more cabs deployed during these times to maximize revenue.
* The highest range of hours where trips are requested is from 10:00 hours to 23:00 hours, so there should be more cabs dispatched during these hours.
* As seen in the map above the highest trips are in the north-west and south-east parts of NYC, so more cabs should be dispatched in these zones.
* Most trips have been requested during months of May and October, so there should be more cabs available during these months to meet the demands.
* Thursdays, Wednesdays, and Fridays are when most trips have been recorded, so there should be more cabs deployed during these days.
  + 1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**
* The highest fare per mile per passenger is the highest when there is only 1 passenger so this category of trips should be prioritized by the cabs.
* Highest average fare per mile is highest on Saturdays so more cabs should be made available during these days.
* Average fares per mile are the lowest during night times, so extra charges should be imposed for trips during these hours.
* The average fare per mile for Vendor 2 is significantly higher than Vendor 1 so priority of trips should be given to Vendor 2.
* 0-2 mile trips have the highest average fare per mile and even more so in the case of Vendor 2, so trips of these distances should be prioritized.