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**

Since the data is very large, we will use sampling technique for smoother analysis.

Sampled data will be the 5% of actual dataa.

## Data Cleaning

### Fixing Columns

* + 1. **Fix the index**  
         
       fixing the index using df.reset\_index() code for better visualization of data
    2. **Combine the two airport\_fee columns**

Combining the two airport\_fee columns to avoid error of duplicate columns

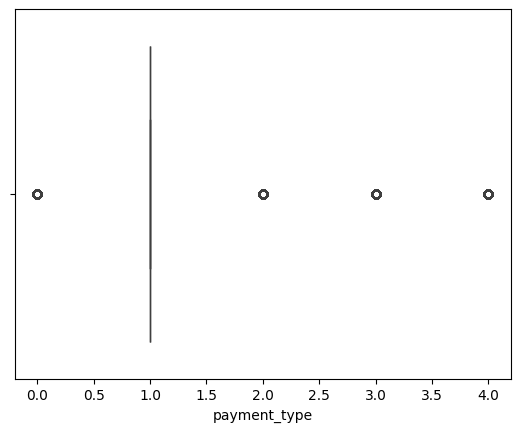
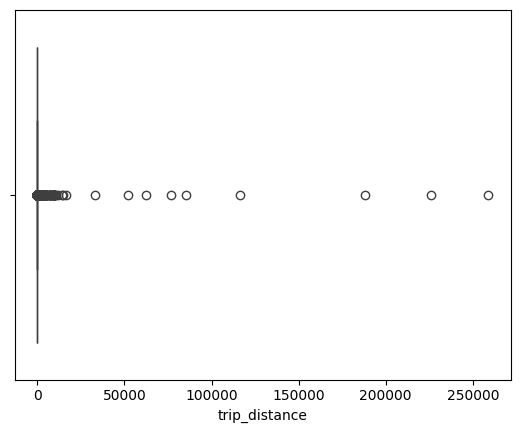
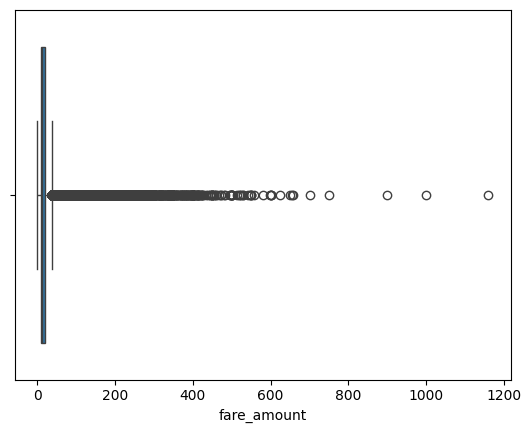
### Handling Missing Values

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

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

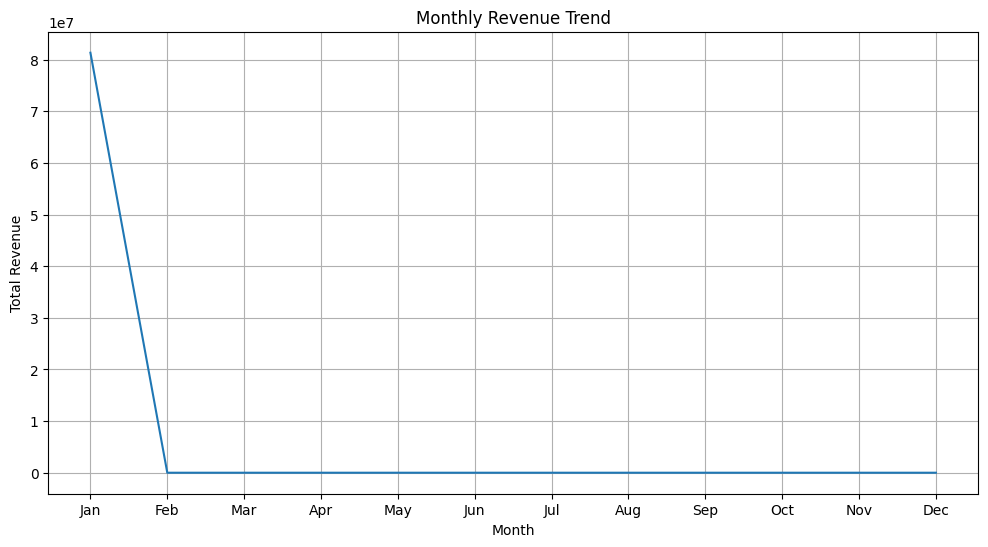
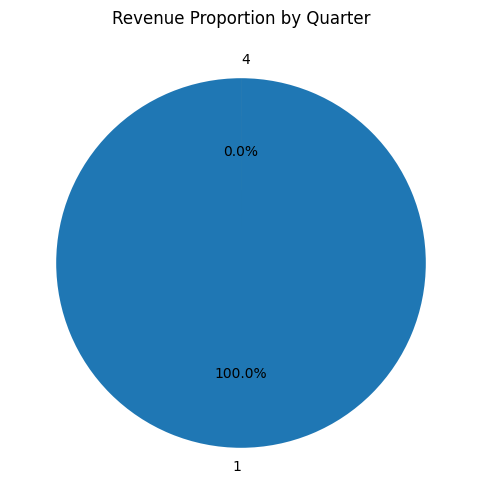
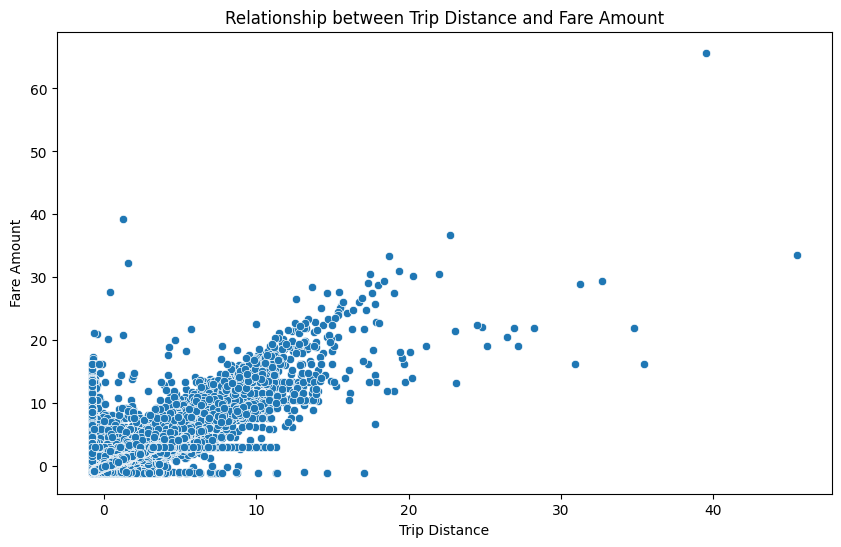
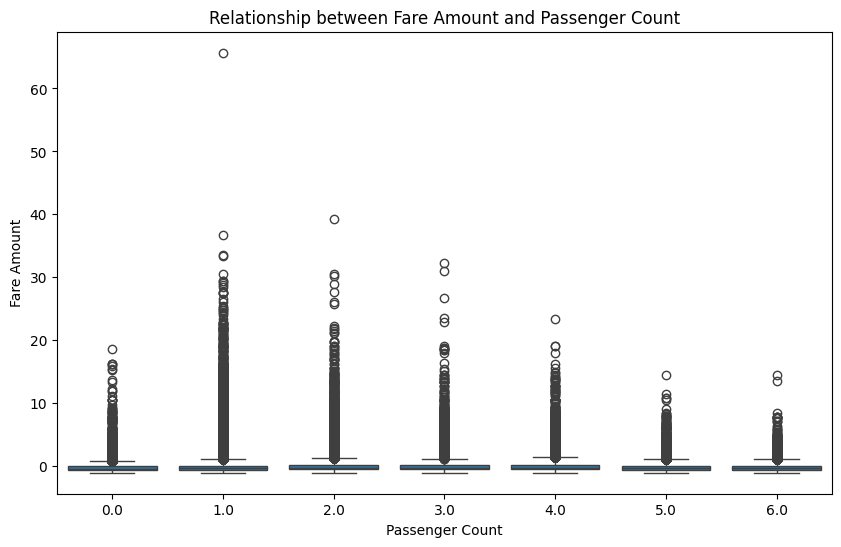
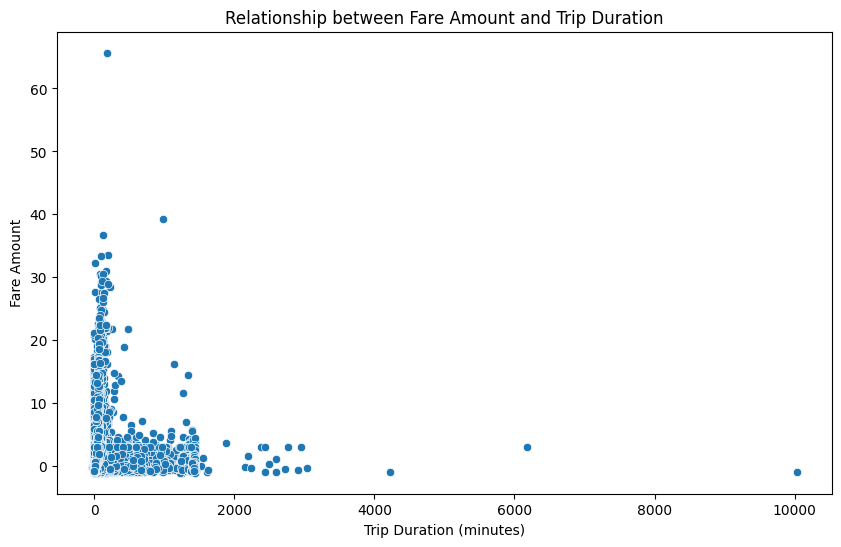
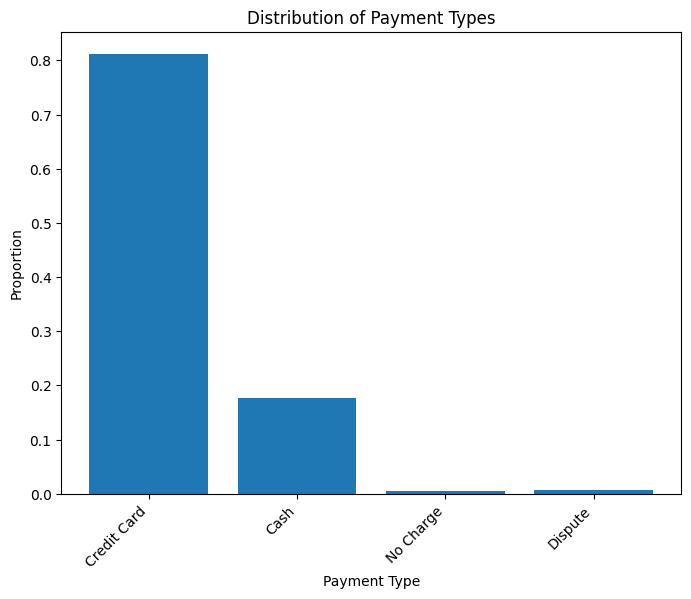
Finding and handling the missing values gives us more appropriate analysis of data and will give better results

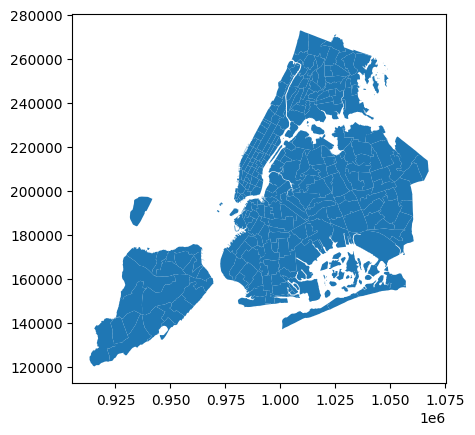
### Handling Outliers and Standardising Values

* + 1. **Check outliers in payment type, trip distance and fare amount columns**
    2. **Payment type outliers**
    3. **trip distance outliers**  
       
    4. **Fare amount outliers**  
       

## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical**
    2. **Analyse the distribution of taxi pickups by hours, days of the week, and months**
    3. **Filter out the zero/negative values in fares, distance and tips**
    4. **Analyse the monthly revenue trends**
    5. **Find the proportion of each quarter’s revenue in the yearly revenue**
    6. **Analyse and visualise the relationship between distance and fare amount**
    7. **Analyse the relationship between fare/tips and trips/passengers  
         
       ** 
    8. **Analyse the distribution of different payment types**
    9. **Load the taxi zones shapefile and display it***import geopandas as gpd*

*Read the shapefile using geopandas*

* + 1. **Merge the zone data with trips data**
    2. **Find the number of trips for each zone/location ID** LocationID trip\_count

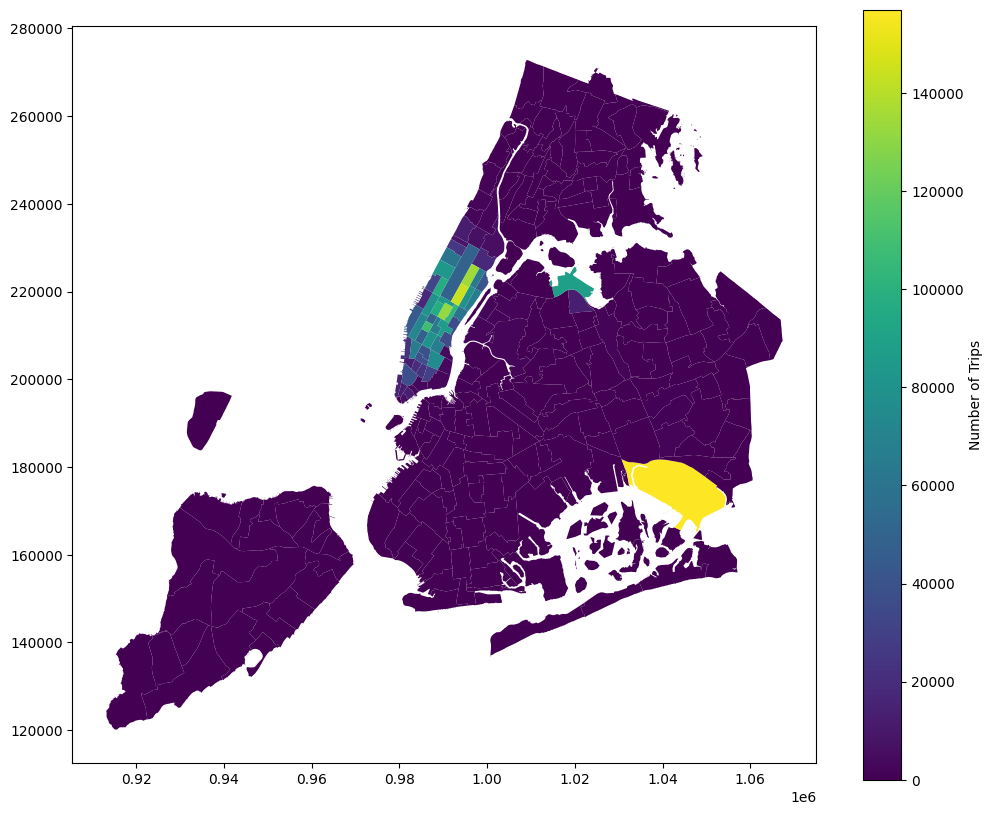
0 1.0 398

1 2.0 1

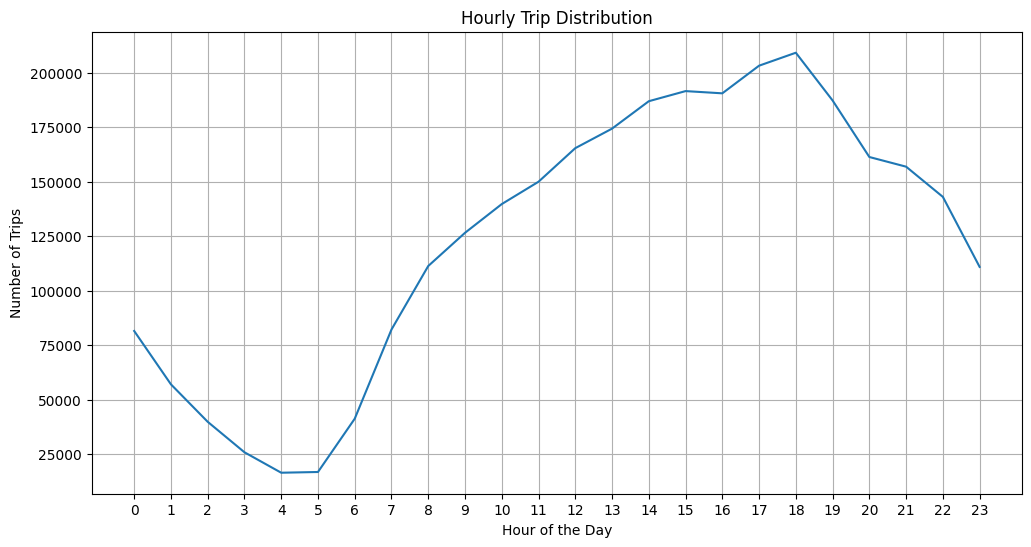
2 3.0 38

3 4.0 3088

4 5.0 56

* + 1. **Add the number of trips for each zone to the zones dataframe**
    2. **Plot a map of the zones showing number of trips**
    3. **Conclude with results**

### 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 busiest hour is 18:00 with 209154 trips.*

* + 1. **Scale up the number of trips from above to find the actual number of trips**

Actual Number of Trips in the 5 Busiest Hours (Scaled):

tpep\_pickup\_datetime

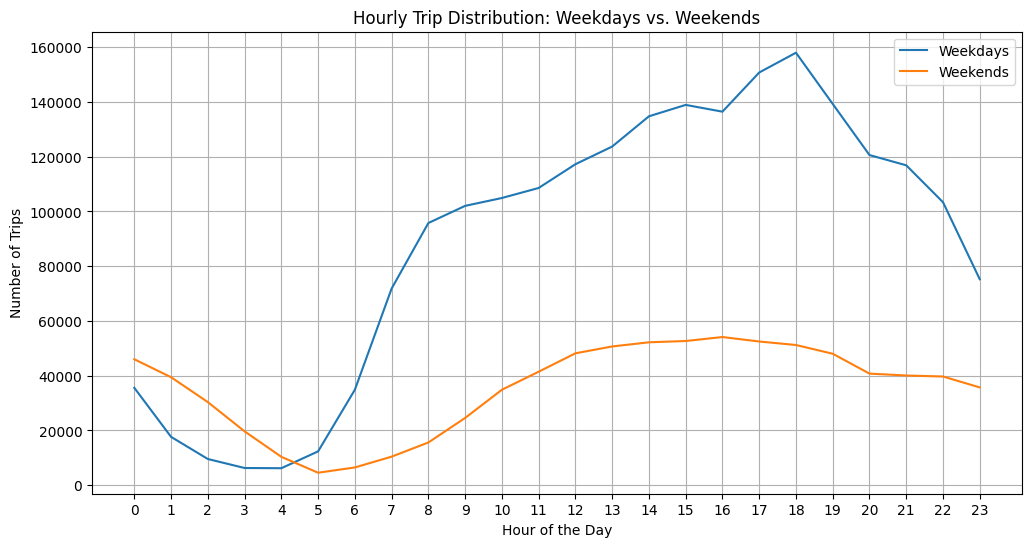
18 4183080.0

17 4064160.0

15 3831120.0

16 3810700.0

1. 3745140.0

**1.2.4 Compare hourly traffic on weekdays and weekends**

**1.2.5 Identify the top 10 zones with high hourly pickups and drops**Top 10 Pickup Zones:

PULocationID

132 157034

237 144668

236 134405

161 132492

186 107455

162 103285

142 97422

230 96559

138 88231

170 86043

Top 10 Dropoff Zones:

DOLocationID

236 142237

237 128967

161 112630

230 87194

170 86506

142 85364

239 85317

141 85290

162 80474

1. 75088

**1.2.6 Find the ratio of pickups and dropoffs in each zone**Top 10 Pickup/Dropoff Ratios:

pickup\_dropoff\_ratio

PULocationID

1 NaN

2 NaN

3 NaN

4 NaN

5 NaN

6 NaN

7 NaN

8 NaN

9 NaN

10 NaN

Bottom 10 Pickup/Dropoff Ratios:

pickup\_dropoff\_ratio

PULocationID

1 NaN

2 NaN

3 NaN

4 NaN

5 NaN

6 NaN

7 NaN

8 NaN

9 NaN

1. NaN

**1.2.7 Identify the top zones with high traffic during night hours**

Top 10 Pickup Zones during Night Hours (11 PM - 5 AM):

79 27746

249 20127

132 19818

48 15908

148 15652

114 13814

230 11888

107 9769

138 9487

186 8927

Top 10 Dropoff Zones during Night Hours (11 PM - 5 AM):

79 13605

48 11093

170 9780

141 9063

107 9041

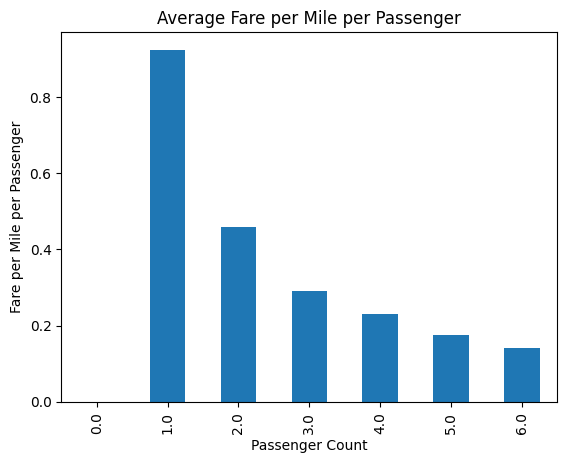
263 8769

68 8573

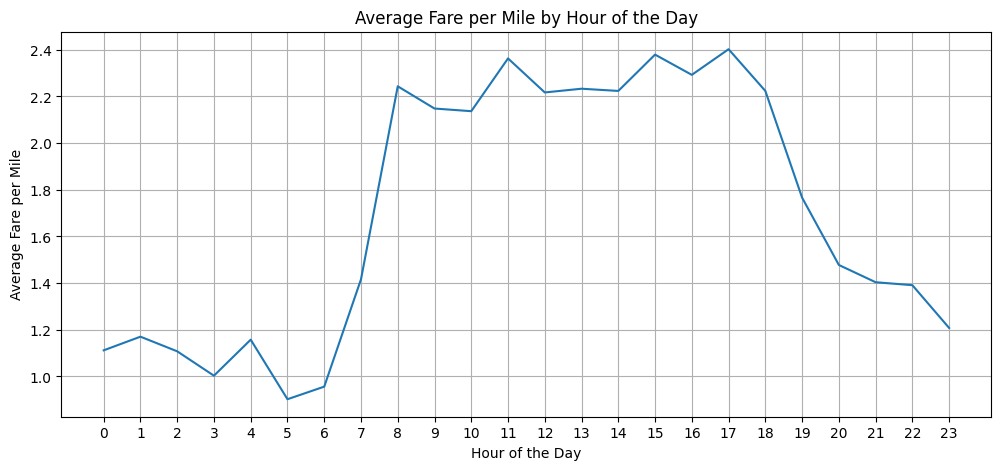
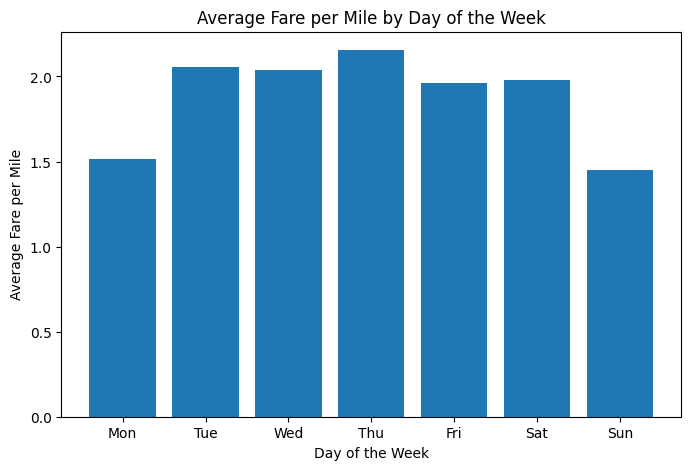
249 7786

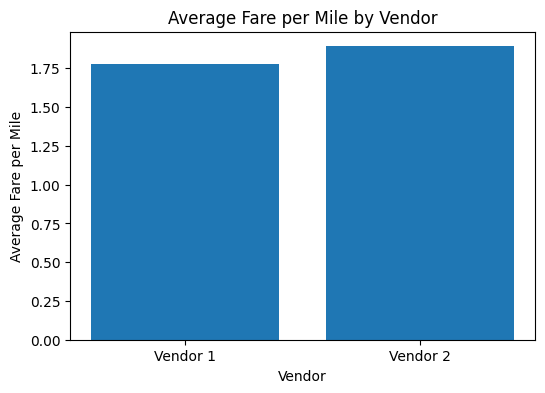
229 7265

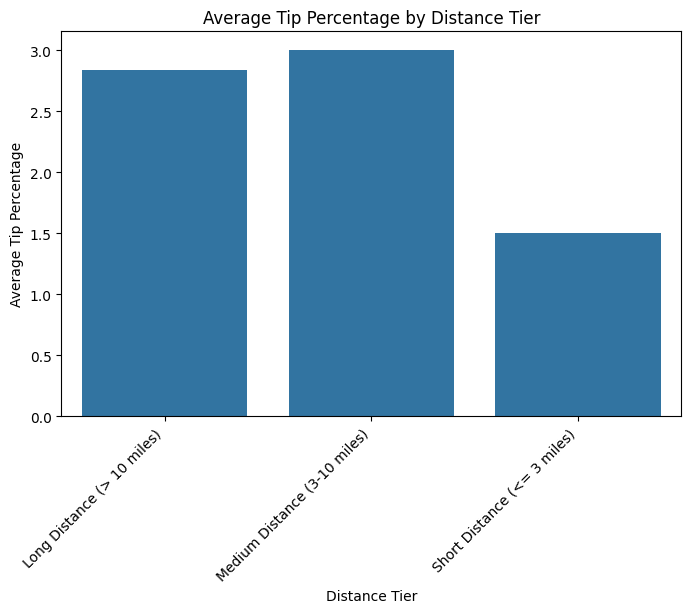
1. 6793  
   * 1. **Find the revenue share for nighttime and daytime hours**
     2. **For the different passenger counts, find the average fare per mile per passenger**

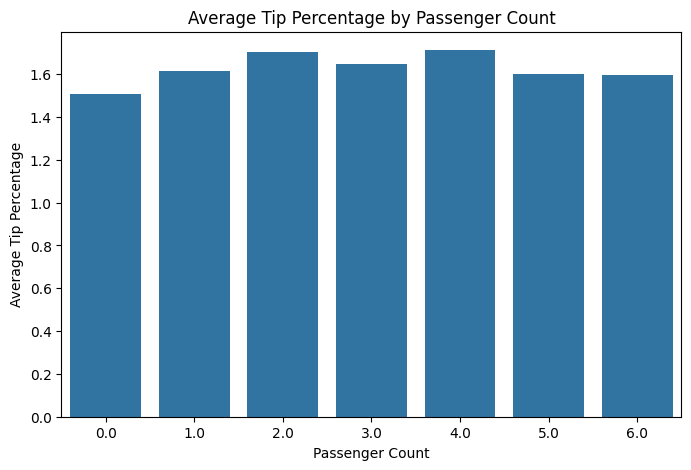


**1.2.10 Find the average fare per mile by hours of the day and by days of the week**



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



* + 1. **Analyse the trends in passenger count**
    2. **Analyse the variation of passenger counts across zones**

**1.2.16 Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.**Surcharge Application Frequency:

Frequency Percentage

extra 1805386 60.791215

mta\_tax 2947163 99.237292

tip\_amount 1082200 36.439993

tolls\_amount 220558 7.42666

improvement\_surcharge 2968820 99.96653

congestion\_surcharge 2744240 92.40444

airport\_fee 260945 8.786577

## Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**

**Routing Optimization:**

1. **Identify and Address Slow Routes:** Analyze trip data to pinpoint routes with consistently slower average speeds, particularly during peak hours. This could indicate traffic congestion, road closures, or inefficient route choices.
   * **Possible solutions:**
     + **Suggest alternative routes to drivers**: Provide drivers with real-time traffic updates and alternative route suggestions through navigation apps or dispatch systems.
     + **Collaborate with traffic authorities**: Work with city officials to address traffic bottlenecks or implement traffic management strategies in high-congestion areas.
     + **Invest in predictive models**: Develop predictive models to anticipate traffic patterns and proactively adjust routes.
2. **Route Optimization based on Pickup/Dropoff Ratios**: Utilize the pickup/dropoff ratio analysis to identify zones with imbalanced taxi supply and demand.
   * **Possible solutions:**
     + Redirect taxis from areas with excess supply: Encourage drivers to move from zones with low demand to areas with higher pickup requests.
     + **Incentivize drivers to accept trips to high-demand zones**: Offer bonuses or incentives for drivers who complete trips in areas with high pickup-to-dropoff ratios.
     + **Implement dynamic pricing**: Adjust fares dynamically based on demand and supply in different zones to encourage drivers to serve high-demand areas.
3. **Leverage Real-Time Traffic Data**: Integrate real-time traffic information into routing algorithms to ensure drivers are taking the most efficient routes.
   * **Possible solutions**:
     + **Partner with navigation providers:** Collaborate with companies like Google Maps or Waze to access real-time traffic data and integrate it into dispatch systems.
     + **Utilize GPS tracking**: Track taxi locations and speeds to monitor traffic conditions and identify potential delays.

**Dispatching Optimization:**

1. **Prioritize Trips to High-Demand Zones**: When dispatching taxis, prioritize trips originating in zones with high pickup-to-dropoff ratios to ensure efficient allocation of resources.

**Possible solutions:**

* + **Implement a queueing system**: Create a system that prioritizes dispatching taxis to areas with the longest wait times or highest demand.
  + **Utilize predictive models:** Develop predictive models to forecast demand in different zones and preemptively dispatch taxis to those areas.

1. **Dynamically Adjust Dispatch Zones:** Continuously monitor demand patterns and adjust dispatch zones to optimize taxi coverage and response times.

**Possible solutions:**

* + **Create flexible dispatch zones**: Define dispatch zones that can be dynamically resized or relocated based on real-time demand.
  + **Utilize geofencing**: Use geofencing technology to track taxi locations and automatically assign them to the most appropriate dispatch zone.

1. **Improve Communication with Drivers**: Enhance communication between dispatchers and drivers to provide real-time updates on demand, traffic conditions, and route changes.

**Possible solutions**:

* + **Implement a mobile dispatch app**: Develop a mobile app that allows drivers to receive trip requests, communicate with dispatchers, and access real-time information.
  + **Utilize two-way communication systems**: Equip taxis with two-way radios or messaging systems to facilitate seamless communication between dispatchers and drivers.

* + 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.**

**Strategic Cab Positioning:**

1. **High-Demand Zones & Peak Hours:**
   * **Increase cab density**: During peak hours (4 PM to 8 PM) on weekdays, concentrate a higher number of cabs in zones with historically high pickup and dropoff activity, such as Manhattan's Midtown and Lower Manhattan. This ensures quick response times and reduces passenger wait times during periods of high demand.
   * **Utilize predictive models**: Develop predictive models to forecast demand in different zones based on historical data, time of day, day of the week, and other factors. Use these models to proactively position cabs in anticipation of surges in demand, optimizing resource allocation.
2. **Weekday vs. Weekend Variations:**
   * **Adjust cab distribution**: Recognize the differences in traffic patterns between weekdays and weekends. On weekdays, focus on high-demand zones during rush hour, while on weekends, distribute cabs more evenly across the city to cater to the more dispersed demand.
   * **Target specific zones**: Identify zones that experience higher demand during weekends, such as entertainment districts or tourist attractions, and position cabs strategically in those areas to capture the weekend traffic.
3. **Monthly & Seasonal Trends:**
   * **Adapt to seasonal changes:** Consider seasonal variations in demand, such as increased tourism during summer months or holiday seasons. Adjust cab positioning accordingly to accommodate the influx of visitors and cater to specific event-driven demand.
   * **Monitor and adjust:** Continuously monitor trip data to identify emerging trends or shifts in demand patterns. Use this information to refine cab positioning strategies and ensure resources are allocated effectively throughout the year.
4. **Pickup/Dropoff Ratio Imbalances:**
   * **Redirect cabs to high-pickup zones:** Encourage drivers to move from zones with low demand or high dropoff-to-pickup ratios to areas with higher pickup requests. This helps balance supply and demand across different zones and reduces the number of empty taxi trips.
   * **Incentivize drivers:** Offer bonuses or incentives for drivers who complete trips in zones with high pickup-to-dropoff ratios to encourage them to serve those areas.
5. **Dynamic Dispatch Zones:**
   * **Create flexible zones:** Define dispatch zones that can be dynamically resized or relocated based on real-time demand and traffic conditions. This allows for more agile and responsive cab positioning.
   * **Utilize geofencing:** Employ geofencing technology to track taxi locations and automatically assign them to the most appropriate dispatch zone based on demand and proximity.
     1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**

**Data-Driven Pricing Adjustments**

The goal is to optimize pricing strategies based on the insights derived from the data analysis. The adjustments are aimed at maximizing revenue while ensuring the rates remain competitive with other vendors in the market. Here's a breakdown of the proposed strategies:

1. **Time-Based Pricing**
   * **Rationale**: Trips during peak hours or on weekends typically have higher demand. Adjusting prices based on these temporal variations allows for maximizing revenue during high-demand periods while offering competitive rates during off-peak hours to attract more customers.
   * **Implementation:**
     + **Peak Hour Surcharges**: Implement dynamic surcharges during peak hours (e.g., morning and evening rush hours, Friday and Saturday nights) when demand is highest. Consider a surge pricing model where prices increase gradually as demand exceeds supply, incentivizing drivers to service high-demand areas and balancing the market.
     + **Off-Peak Discounts**: Offer discounts or promotions during off-peak hours or weekdays to encourage more trips and optimize resource utilization.
     + **Weekend/Holiday Rates**: Adjust rates on weekends and holidays to reflect increased demand.
2. **Location-Based Pricing**
   * **Rationale**: Certain pickup and drop-off zones have higher trip volumes or longer distances traveled. This strategy helps in capturing the value of these popular or longer routes while maintaining competitive pricing for trips within less busy areas.
   * **Implementation:**
     + **Zone-Based Pricing**: Implement different rates for trips based on pickup and drop-off zones. High-demand zones or those with longer average trip distances could have slightly higher fares, while less popular zones could have lower fares.
     + **Airport Surcharges:** Consider implementing an airport surcharge for trips originating or ending at airports.
     + **Minimum Fare:** Establish a minimum fare for short trips to ensure profitability, particularly in areas with lower demand.
3. **Demand-Based Pricing**
   * **Rationale**: Real-time demand can fluctuate based on events, weather, and other external factors. Adjusting prices in response to these factors allows for capturing the increased demand during peak periods.
     + **Implementation:**
       - **Dynamic Pricing**: Explore dynamic pricing algorithms that adjust prices in real-time based on supply and demand levels. Consider using machine learning to predict demand and optimize pricing accordingly.
       - **Event-Based Pricing**: For major events or holidays, implement surge pricing to manage increased demand.
4. **Competitive Benchmarking**
   * **Rationale**: Maintaining competitive rates with other taxi vendors is crucial to avoid losing market share. This strategy involves continuous monitoring and comparing of rates with competitors.
   * **Implementation:**
     + **Market Research**: Regularly track pricing offered by competitors, particularly during peak and off-peak periods, weekends, and in popular zones.
     + **Price Adjustments**: Adjust pricing to stay competitive, taking into consideration factors such as service quality and passenger experience.
5. **Passenger-Based Pricing**
   * **Rationale**: Adjusting prices based on the number of passengers can provide higher revenue per trip without significantly impacting passenger cost per person.
   * **Implementation:**
     + **Passenger Surcharge**: Explore implementing a small surcharge for each additional passenger, starting from a certain threshold (e.g., 3 or 4 passengers).
     + **Group Discounts:** Offer discounts for larger groups or shared rides, promoting group bookings and higher revenues per trip