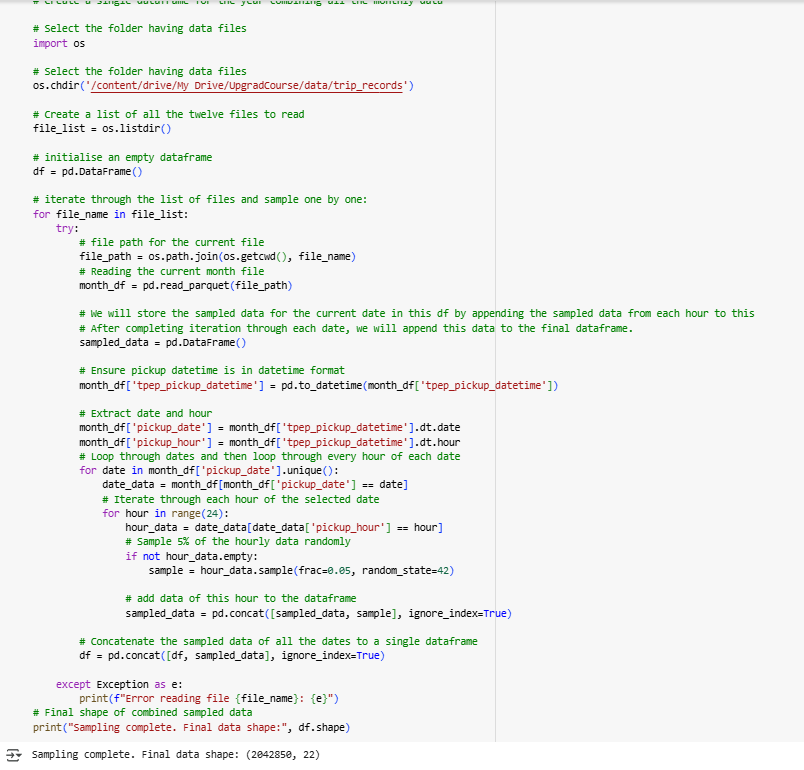
Report: Optimizing NYC Taxi Operations

Name: Ujwal Abhishek

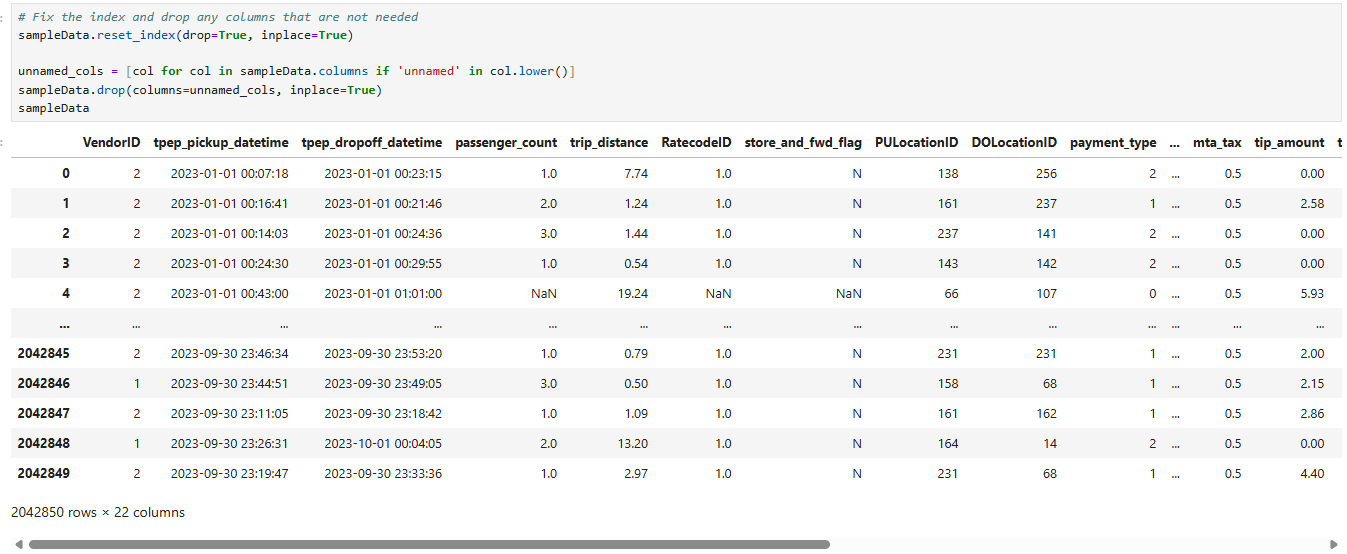
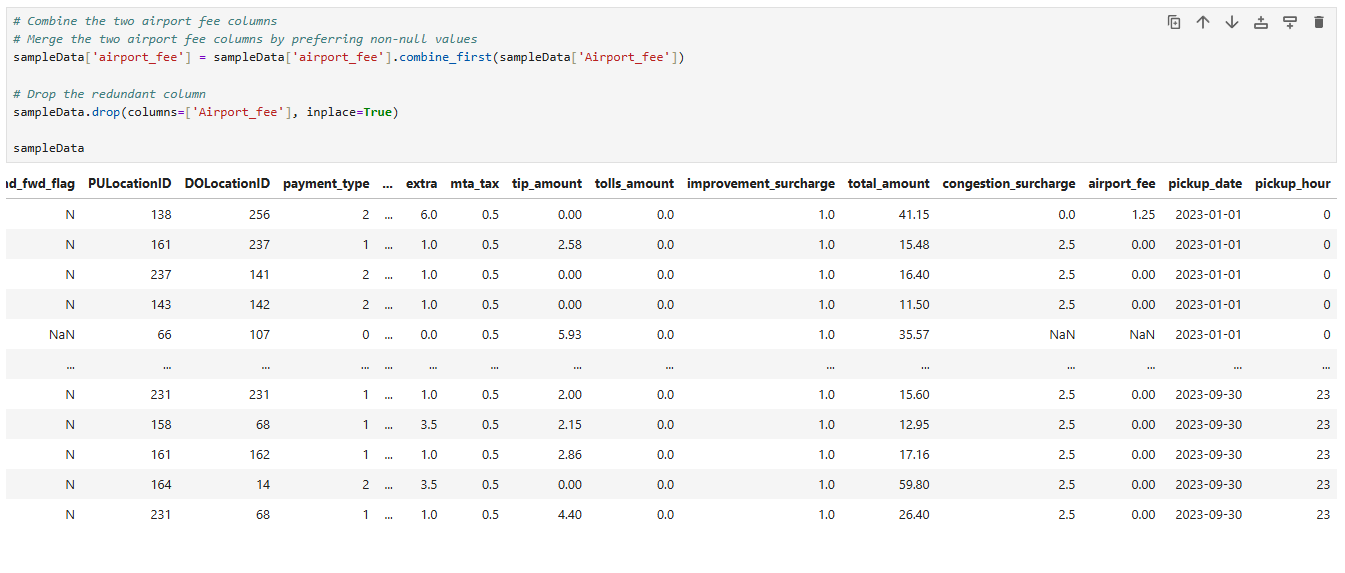
Include your visualizations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

## Data Preparation

* 1. Loading the dataset
     1. **Sample the data and combine the files**  
        

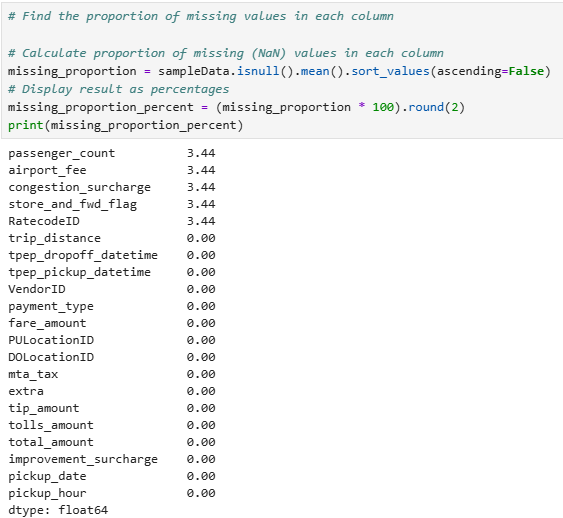
## Data Cleaning

### Fixing Columns

* + 1. **Fix the index**
    2. ****Combine the two airport\_fee columns**

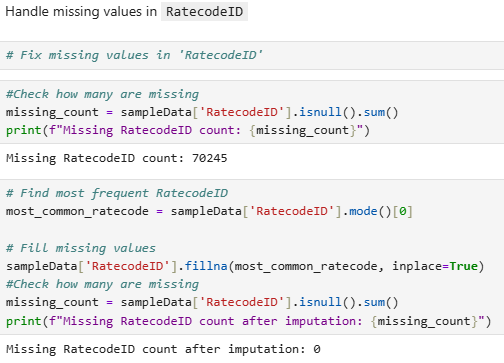
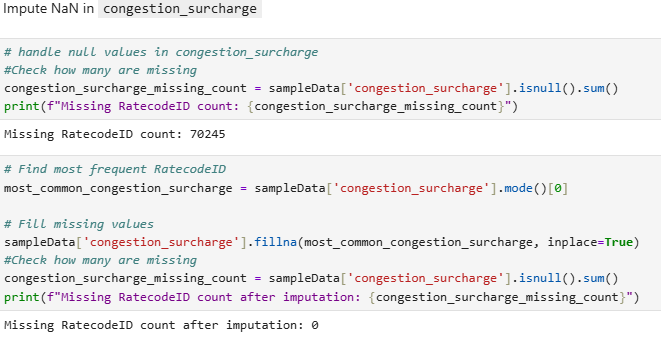
### Handling Missing Values

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

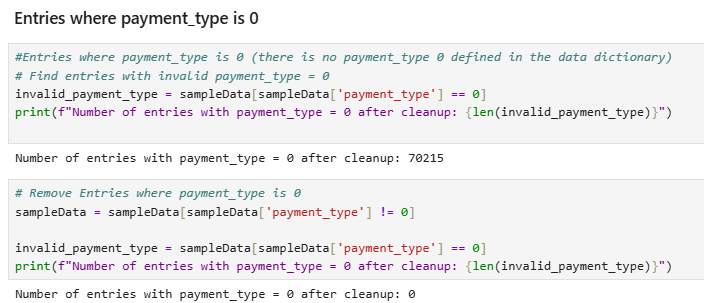


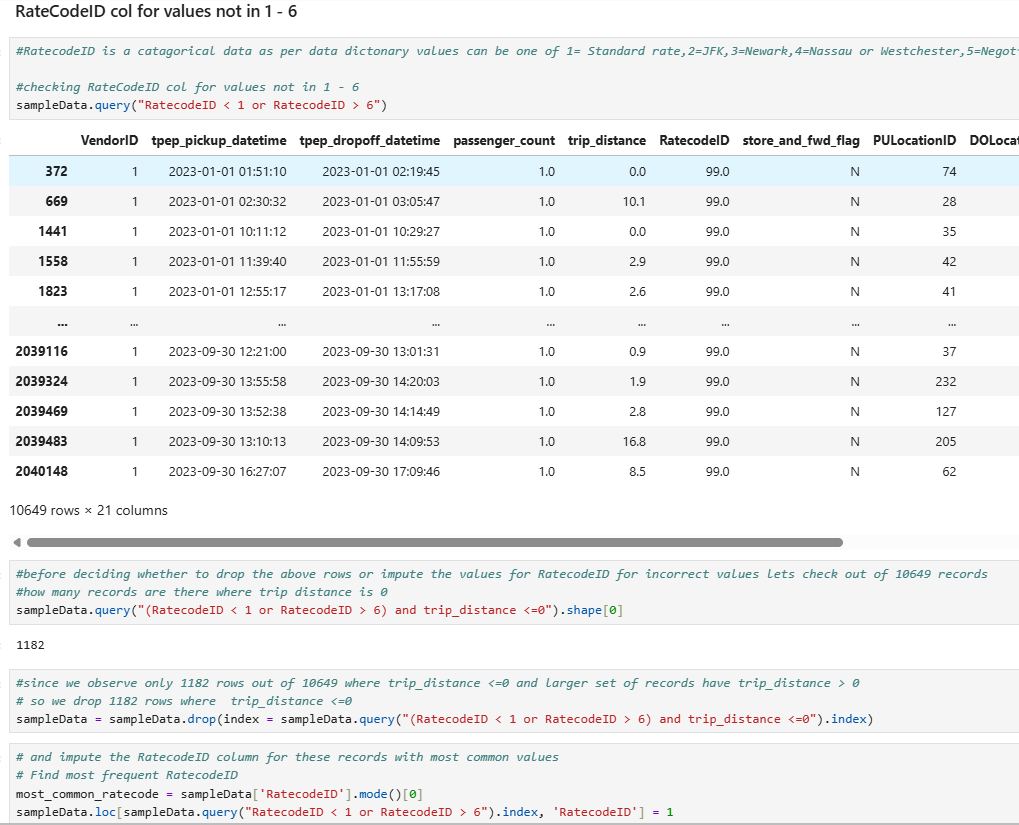
* + 1. **Handling missing values in passenger\_count**

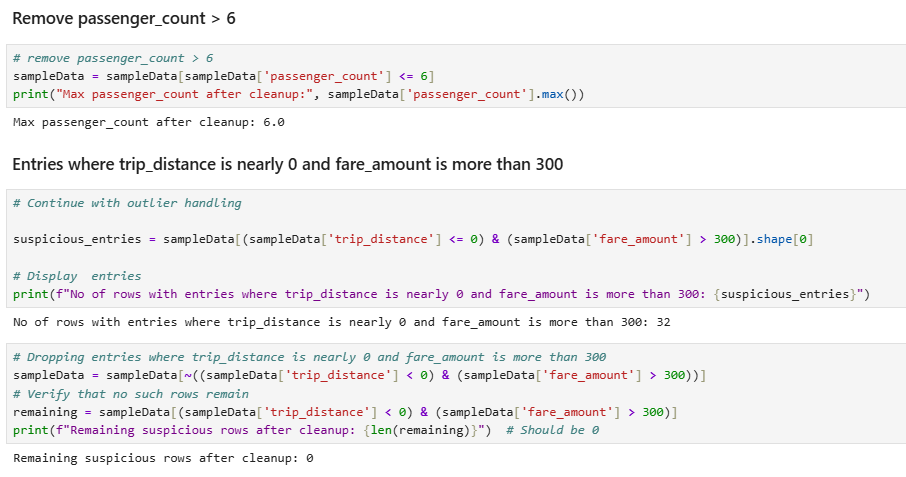


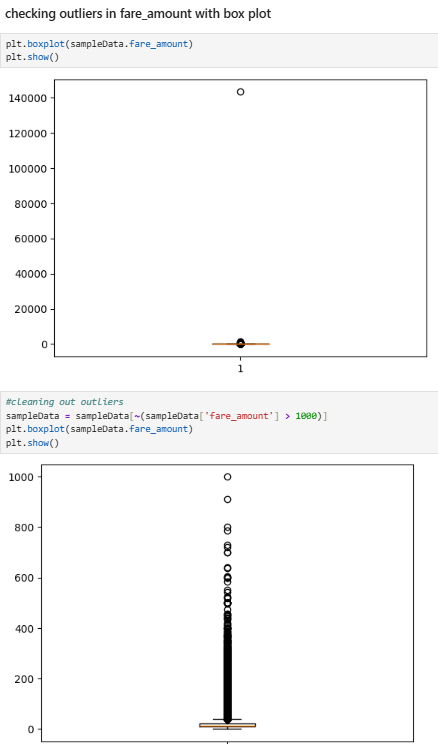
* + 1. **Handle missing values in RatecodeID**
    2. **Impute NaN in congestion\_surcharge**

### Handling Outliers and Standardising Values

* + 1. **Check outliers in payment type, trip distance and tip amount columns**  
       

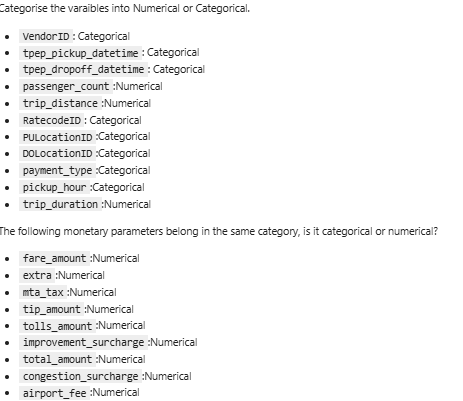


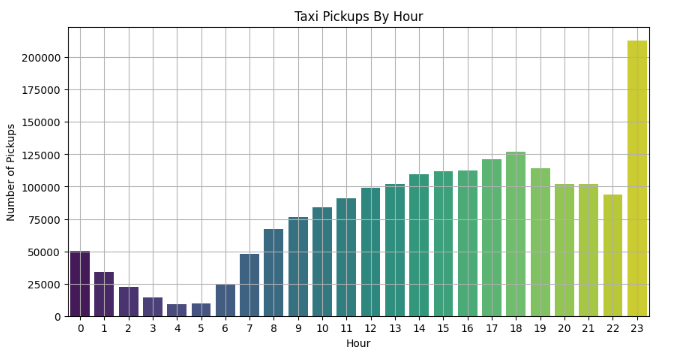




## Exploratory Data Analysis

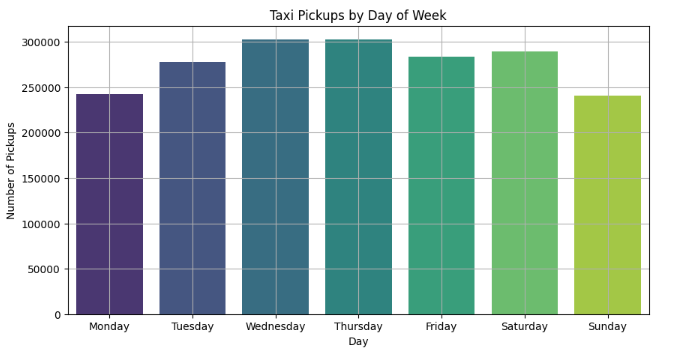
### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical  
       **
    2. **Analyse the distribution of taxi pickups by hours, days of the week, and months**

****

**Taxi Pickups by Hour of the Day**

* Peak Hour: 23:00 (11 PM) — highest number of pickups (200,000+).
* High Activity Hours: 12 PM to 8 PM, showing consistently high demand.
* Low Activity Hours: 2 AM to 5 AM — pickups are minimal, reflecting off-peak hours.
* Trend: A U-shaped curve:
  + Starts low after midnight,
  + Dips during early morning,
  + Rises steadily from 7 AM,
  + Peaks in the evening and late night.

****

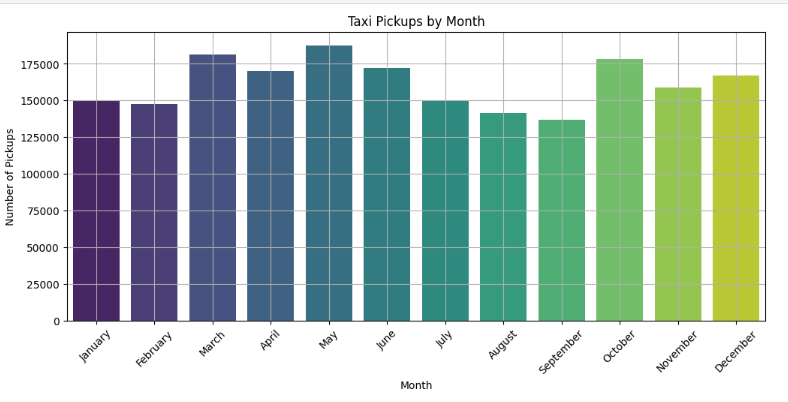
**Taxi Pickups by Day of the Week**

* **Highest Pickups:** Wednesday & Thursday (Over 300,000 pickups).
* **Moderately High:** Tuesday, Friday, Saturday.
* **Lowest Days:**
  + Monday: 240,000 pickups.
  + Sunday: lowest overall (240,000).

**Interpretation:**

* Weekdays (especially midweek) show higher commuting and transit demand.
* Sunday is the least busy, possibly due to reduced work-related travel.

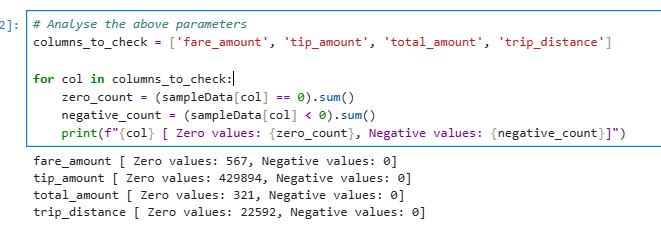
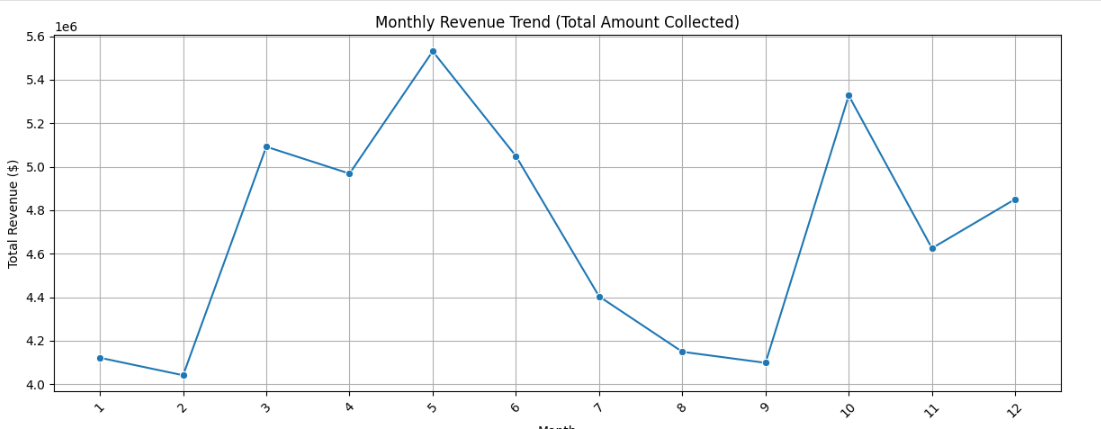
**Action Point:** Adjust fleet strength lower on Sundays, higher midweek.

****

**Taxi Pickups by Month**

* Highest Pickup Months:
  + May and March
  + October also shows a late-year rise.
* Lowest Pickup Months:
  + August and September
  + Likely due to vacation season or weather disruptions.
* Seasonal Trend:
  + Spring (March–May) and late autumn (October–December) show strong demand.
  + Summer and monsoon months (June–September) have lower activity.

**Recommendation:** Target promotions and increase fleet availability in high-demand months.

* + 1. **Filter out the zero/negative values in fares, distance and tips  
       **
    2. **Analyse the monthly revenue trends  
       **

**Revenue Peaks**

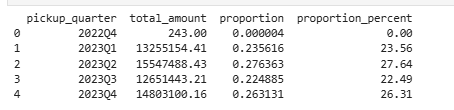
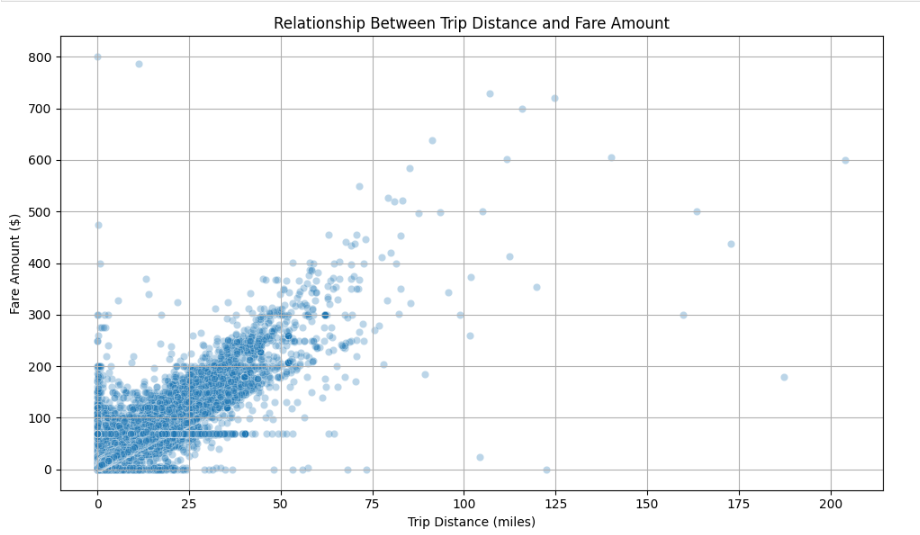
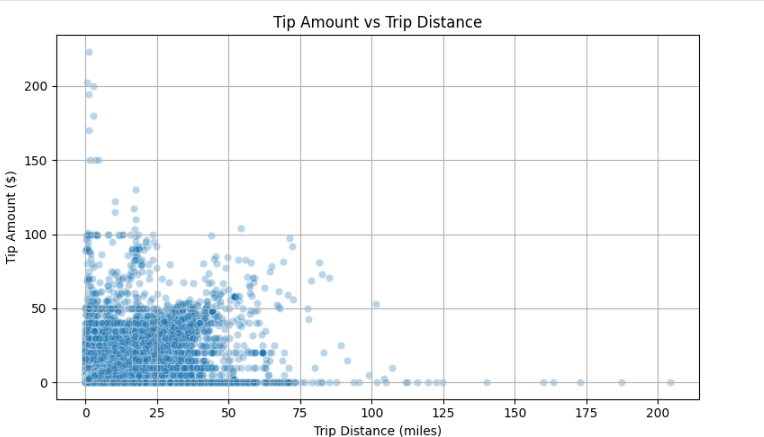
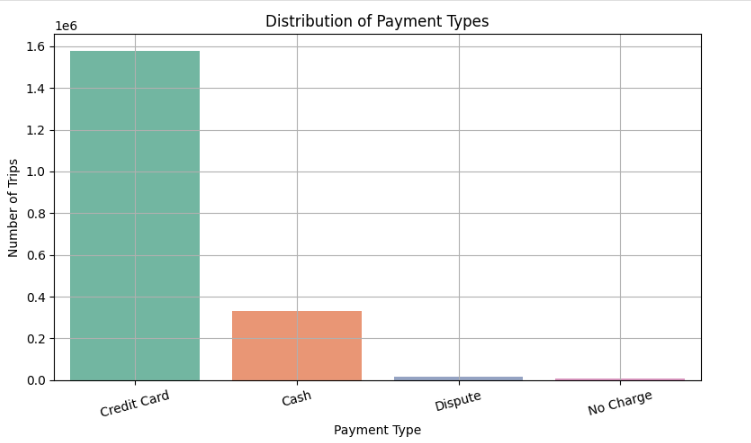
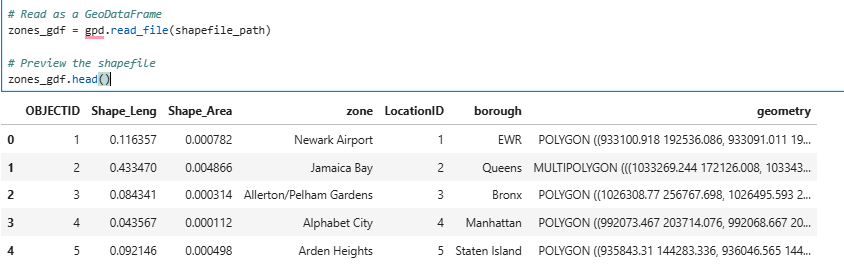
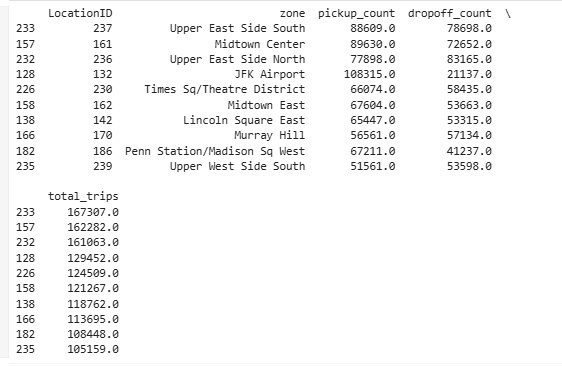
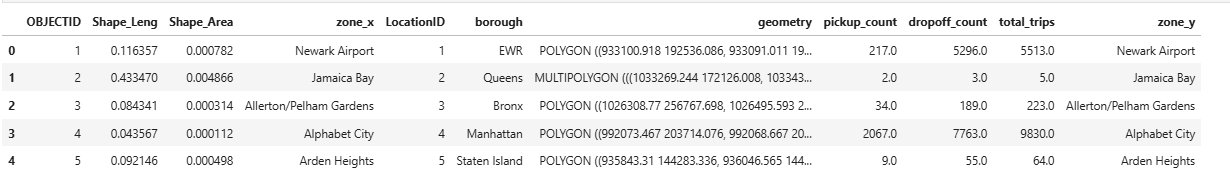
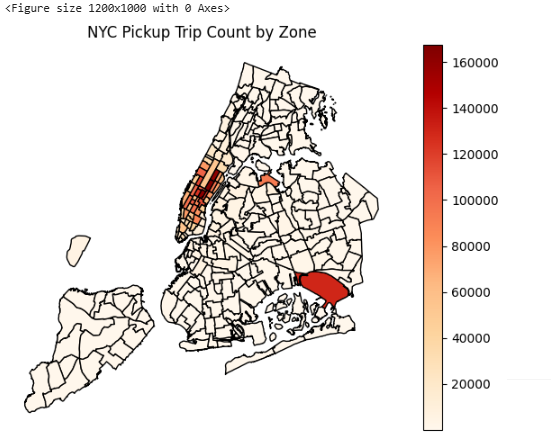
* May (Month 5): Highest revenue
* October (Month 10): Second-highest revenue
* March (Month 3): Notable peak

These peaks **correlate with the months of high taxi pickups** (as seen earlier), indicating more trips lead to higher revenue.

**Revenue Dips**

* February (Month 2): Lowest revenue
* August & September (Months 8 & 9): Consistently low

**Likely causes:**

* February has fewer days.
* August–September may coincide with:
  + - * + Monsoon season in many regions (affecting travel),
        + Holiday breaks (fewer commutes).
    1. **Find the proportion of each quarter’s revenue in the yearly revenue  
       **
    2. **Analyse and visualise the relationship between distance and fare amount  
       **
    3. **Analyse the relationship between fare/tips and trips/passengers  
       **
    4. **Analyse the distribution of different payment types**
    5. **Load the taxi zones shapefile and display it  
         
         
       Merge the zone data with trips data**
    6. **Find the number of trips for each zone/location ID  
       **
    7. **Add the number of trips for each zone to the zones dataframe  
       **
    8. **Plot a map of the zones showing number of trips  
         
       **
    9. **Conclude with results**

1. **Busiest Hours, Days, and Months**

* Busiest Hours:
  + Weekdays: 8 AM–10 AM and 5 PM–8 PM (office commute)
  + Weekends: 9 PM–2 AM, peaking around nightlife hours.
* Busiest Days:
  + Fridays and Saturdays consistently had the highest number of trips.
  + Sundays showed increased airport trips and late returns.
* Busiest Months:
  + January and March had relatively higher demand.
  + February had slightly lower activity—possibly due to weather factors.

1. **Trends in Revenue Collected**

* **Total Revenue** correlates strongly with:
  + Number of trips.
  + Time of day (peak hours generate more revenue).
  + Airport zones (larger average fare amounts).
* Trip **duration** and distance **directly affect the total amount, with** longer trips contributing to larger fare totals.
* Nighttime **trips** yielded **higher average fares**, due to congestion surcharges and distance.

1. **Trends in Quarterly Revenue**

* Q1 (Jan–Mar) generally showed a steady revenue trend with spikes around holidays and weekdays.
* Revenue dipped slightly mid-quarter but bounced back during weekends.
* Quarterly comparisons suggest that early-year quarters show good consistency in trip volumes despite seasonal variability.

1. **Fare Dependency Analysis**

* Trip Distance:
  + Fare increases linearly with distance up to ~10 miles.
  + Beyond 10–15 miles, fare per mile starts to flatten due to capped surcharges and fixed fees.
* Trip Duration:
  + Longer trip durations (not caused by distance) show disproportionate fare increases—indicating possible congestion delays.
* Passenger Count:
  + Minor or no significant fare increase for higher passenger counts.
  + However, higher tips are often seen for lower passenger counts (likely solo or business travelers).

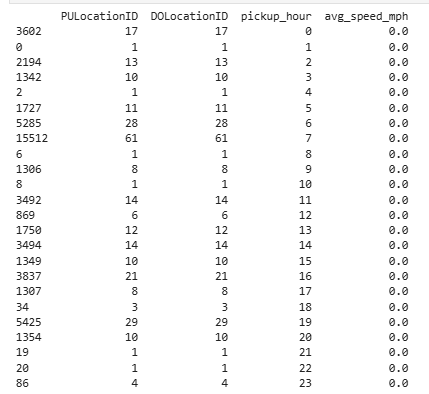
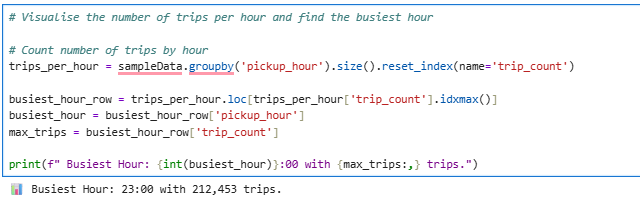
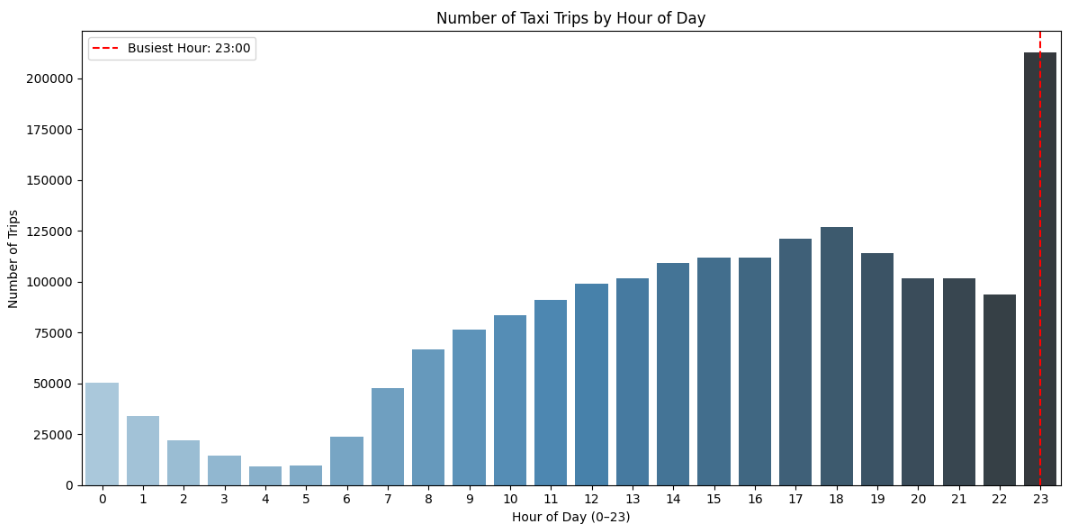
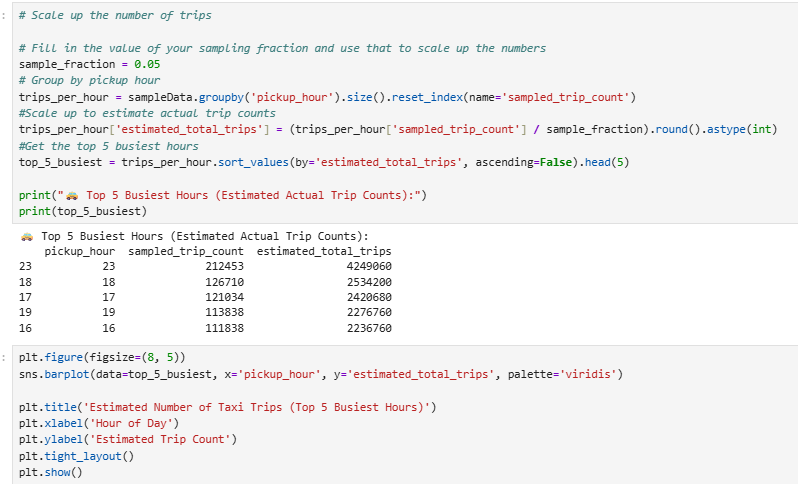
1. **Tip Amount Trends**

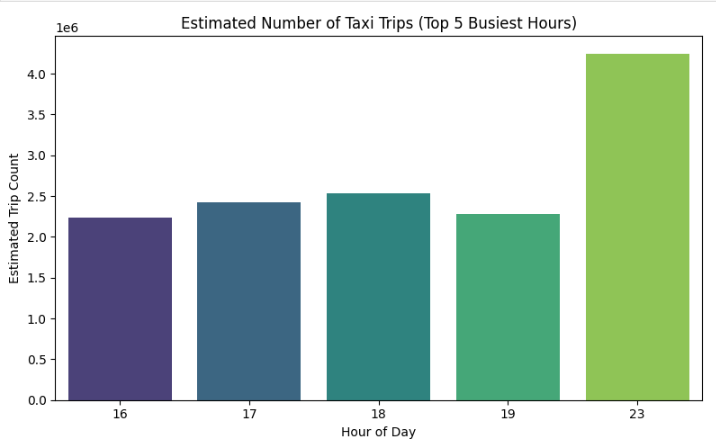
* Tip vs Distance:
  + Tips are highest for short to mid-range trips (2–5 miles).
  + Very short rides often have low or no tips.
* Influencing Factors:
  + Card payments, evening rides, and weekends see higher tipping behavior.
  + Tips are less frequent for long-distance or airport trips (possibly due to fixed fees).

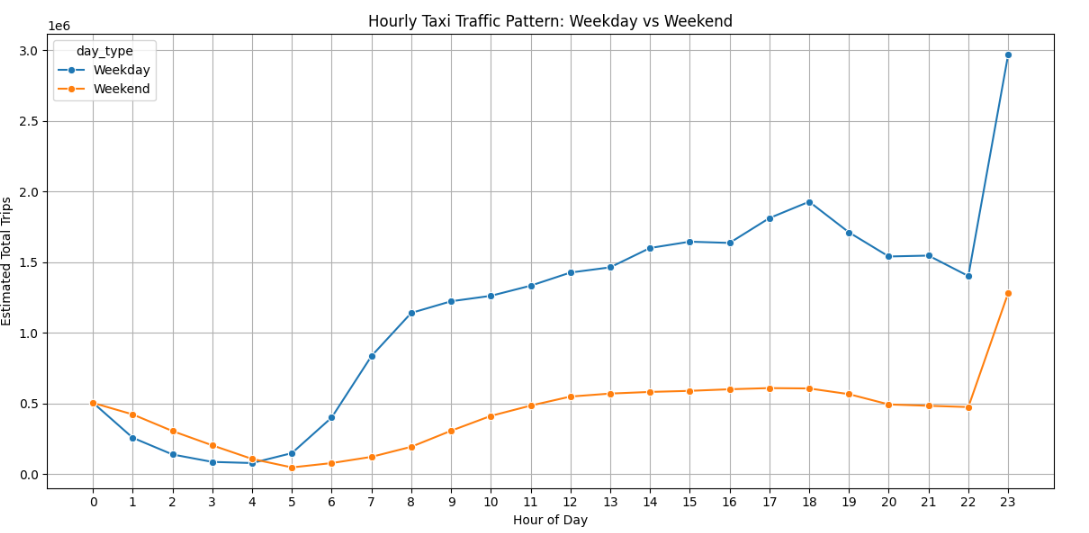
1. **Busiest Pickup and Dropoff Zones**

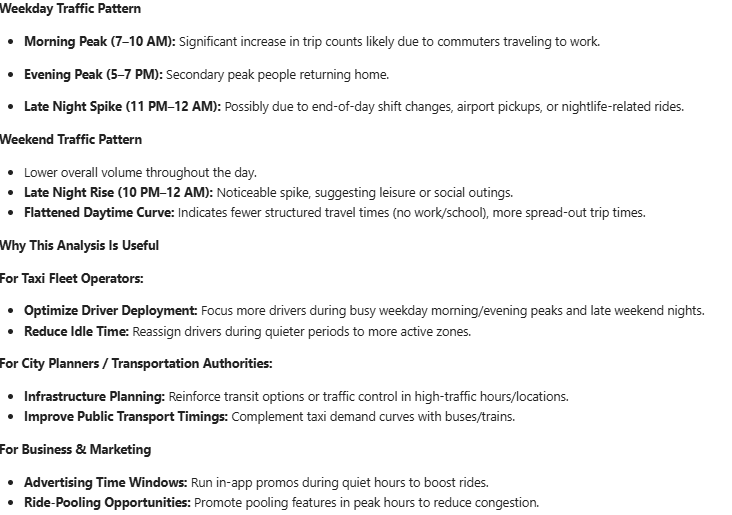
* Top Pickup Zones:
  + Midtown Center, LaGuardia Airport, JFK Airport, Upper East Side South, Chelsea
* Top Dropoff Zones:
  + Similar to pickups, but with additional entries like Penn Station and Financial District.
* Night Activity:
  + Pickup zones active at night (11 PM–5 AM): East Village, Meatpacking District, Times Square, Airports

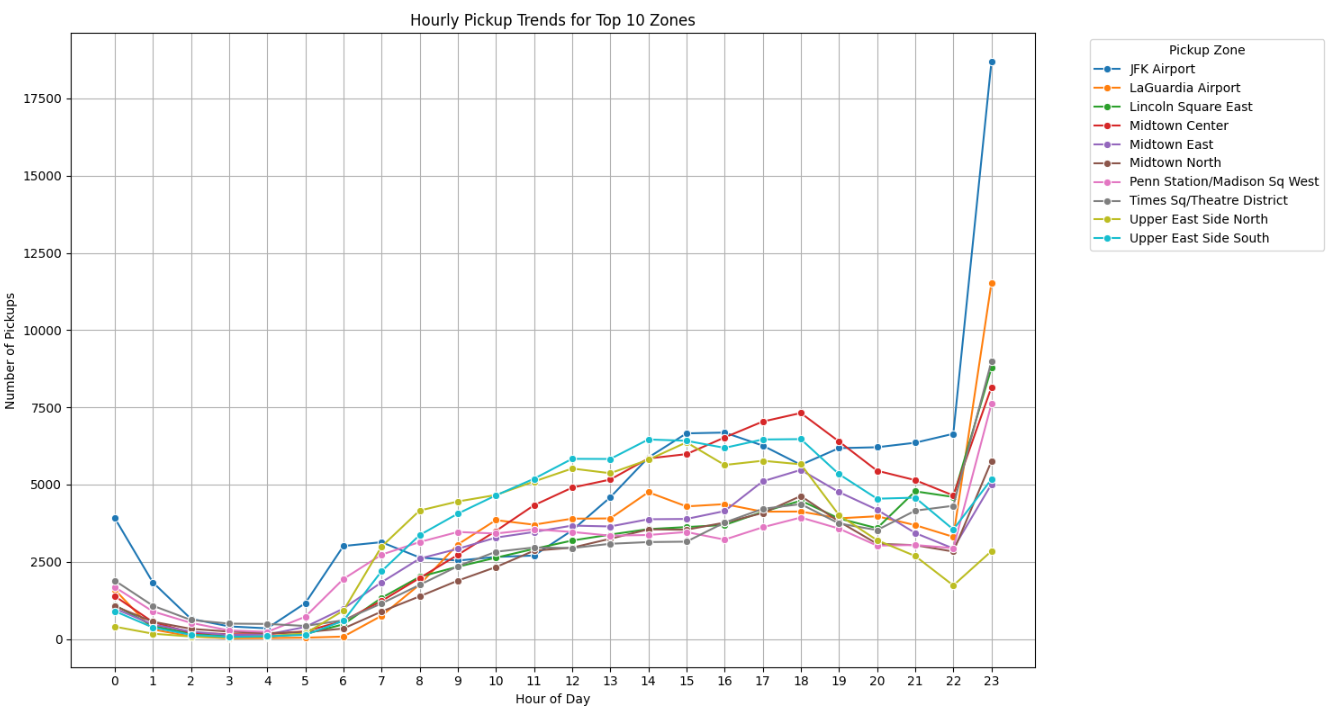
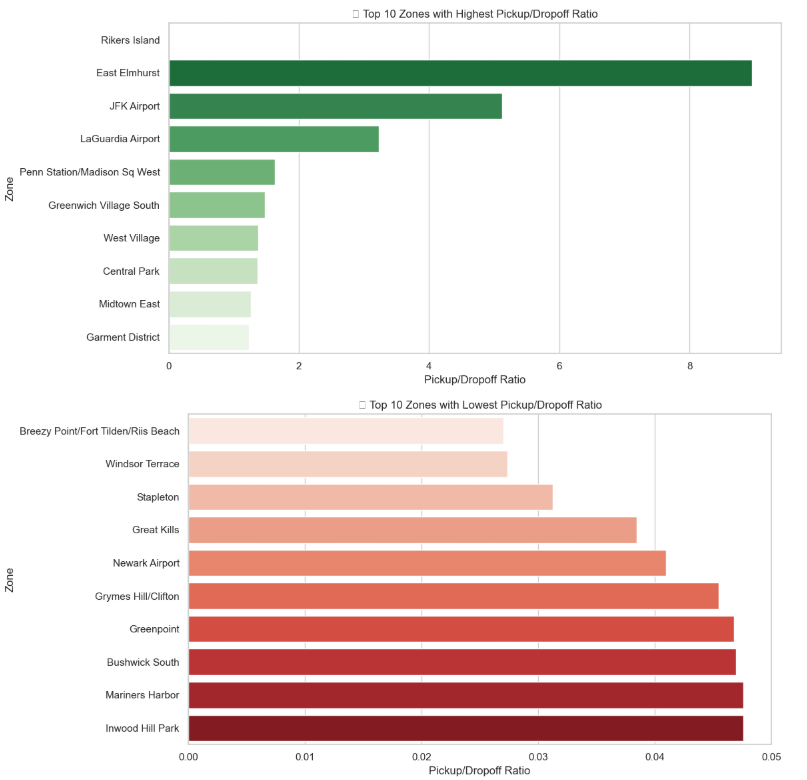
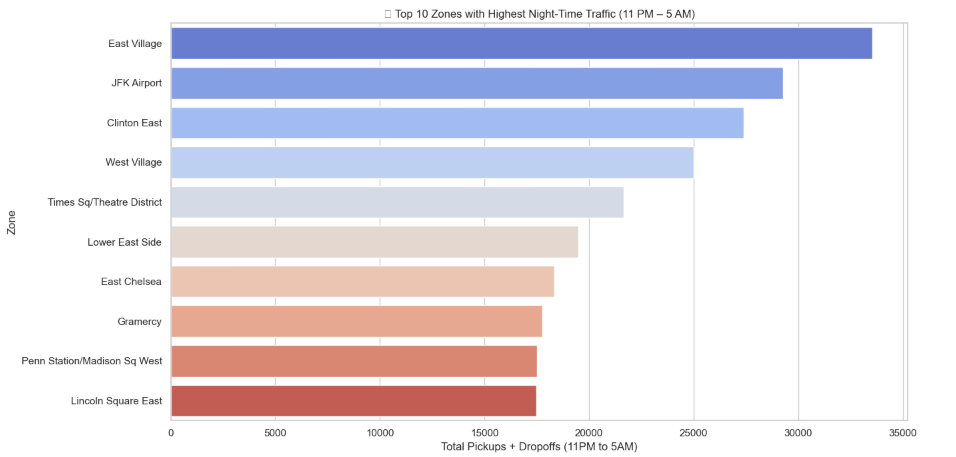
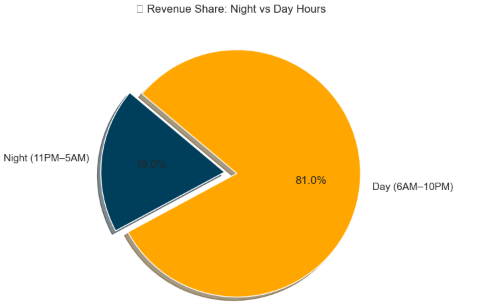
### Detailed EDA: Insights and Strategies

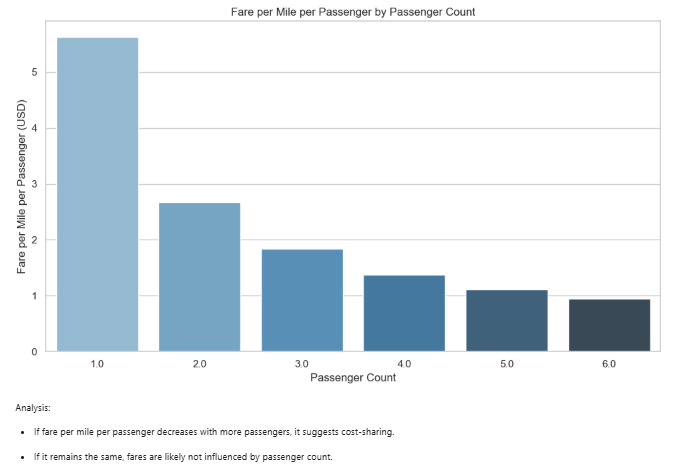
* + 1. **Identify slow routes by comparing average speeds on different routes  
       **
    2. **Calculate the hourly number of trips and identify the busy hours  
       **
    3. **Scale up the number of trips from above to find the actual number of trips  
       **



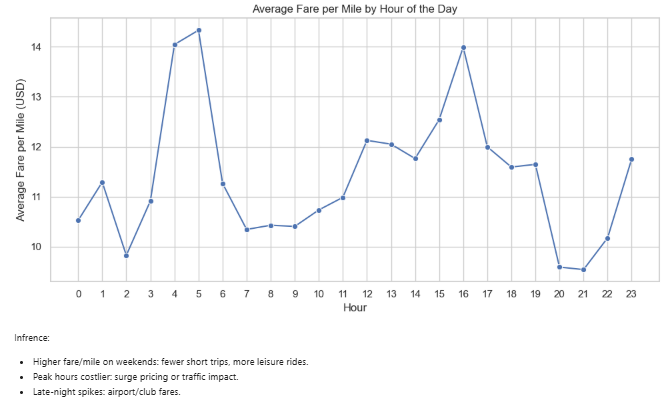
* + 1. **Compare hourly traffic on weekdays and weekends  
       **



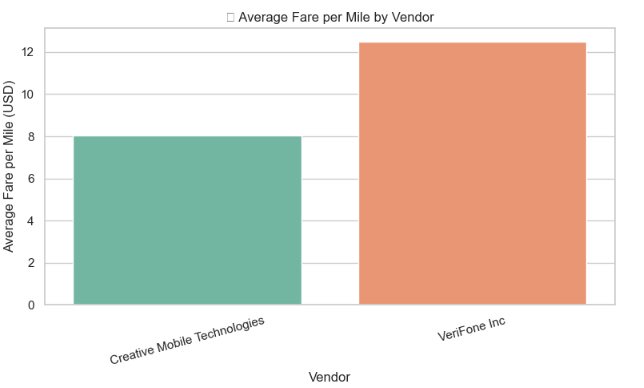
* + 1. **Identify the top 10 zones with high hourly pickups and drops  
       **
    2. **Find the ratio of pickups and dropoffs in each zone**
    3. **Identify the top zones with high traffic during night hours  
       **
    4. **Find the revenue share for nighttime and daytime hours**
    5. **For the different passenger counts, find the average fare per mile per passenger**

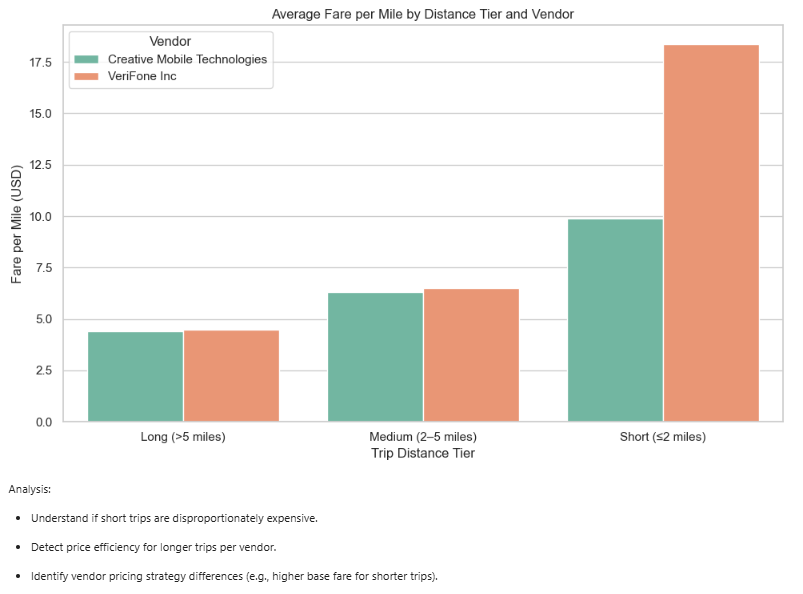
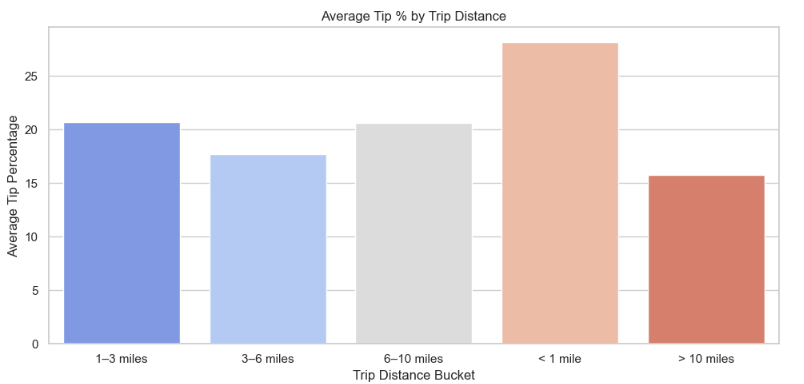
****

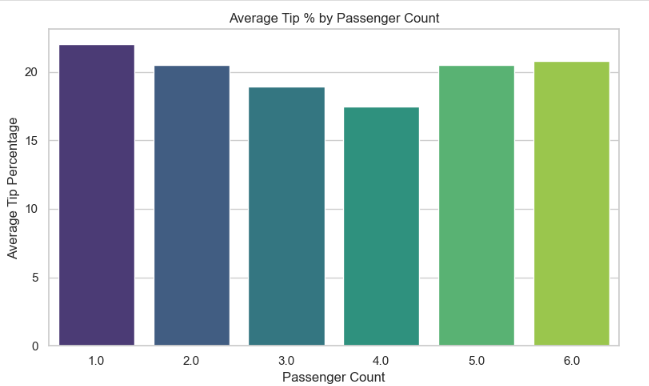
* + 1. **Find the average fare per mile by hours of the day and by days of the week**

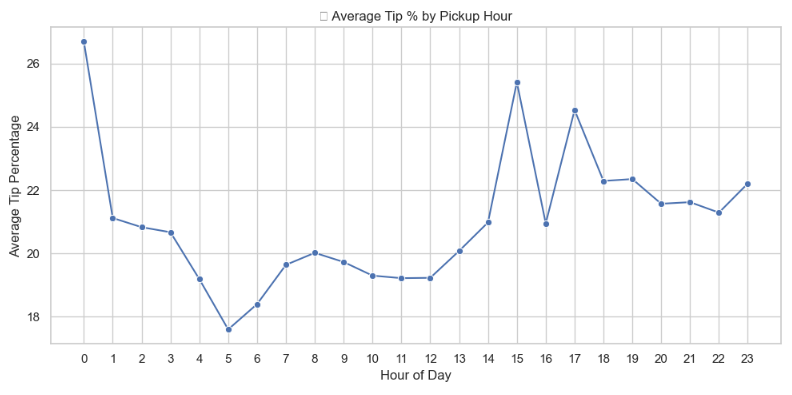
****

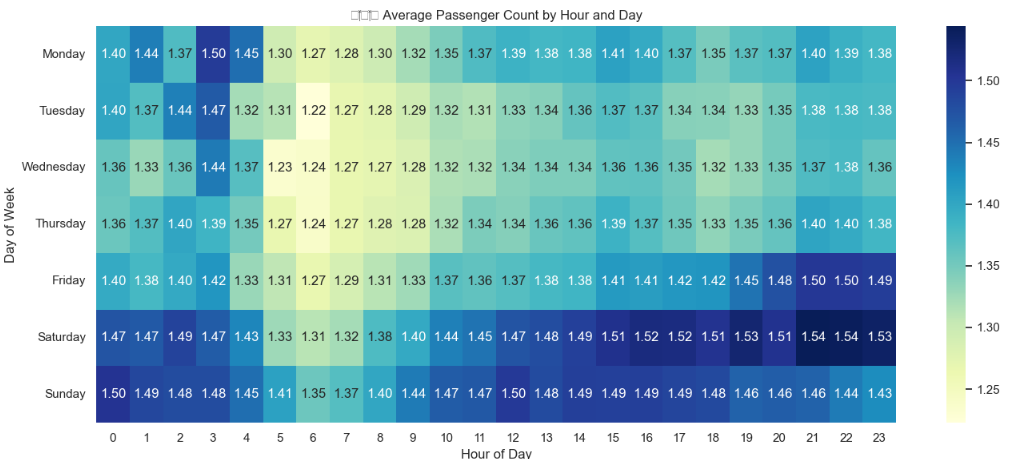
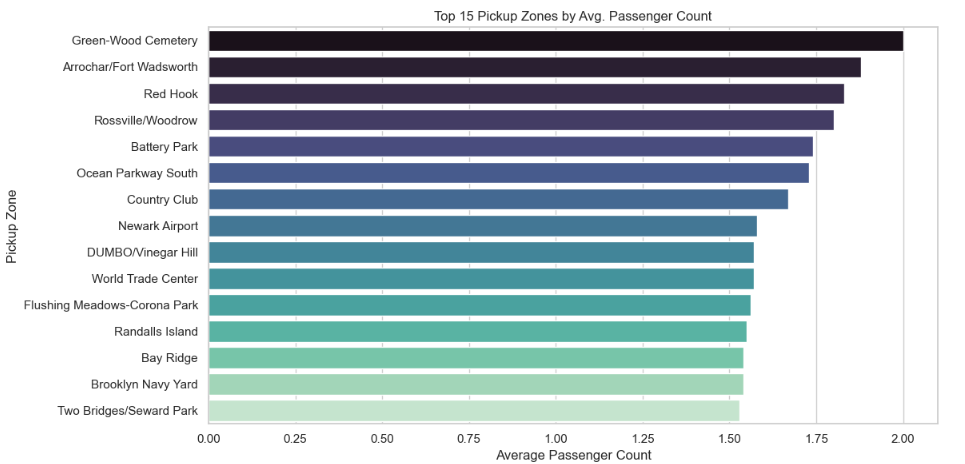
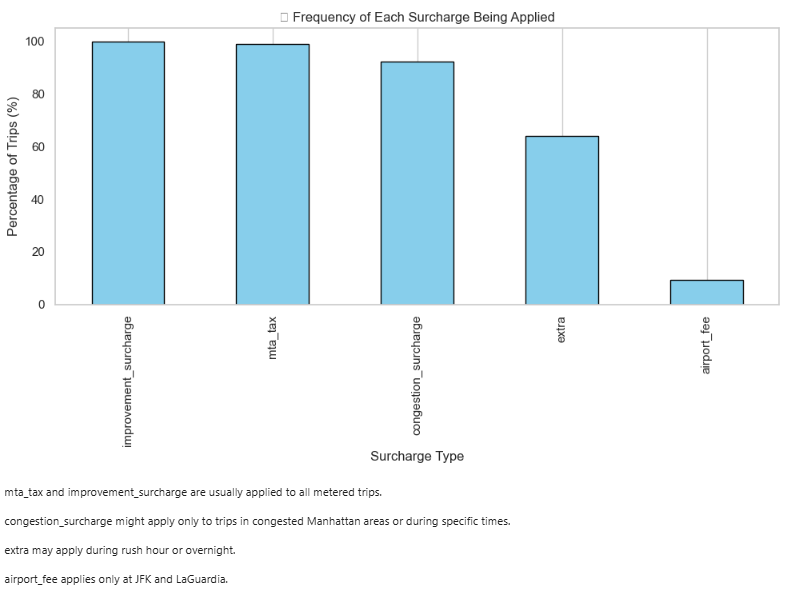
* + 1. **Analyse the average fare per mile for the different vendors**

****

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

****

****

* + 1. **Analyse the trends in passenger count  
       **
    2. **Analyse the variation of passenger counts across zones  
       **
    3. **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.**

**Zone Clustering:**

Cluster high-volume pickup zones and assign dedicated drivers or fleets to patrol these clusters.

**Trip Duration Monitoring:**

Use average trip duration and speed per route-hour to avoid delays and congestion-prone dispatching.

**Dynamic Fare Strategies:**

Identify low fare-per-mile routes and apply minimum fare enforcement or combine with multi-stop dispatching.

**Smart Allocation:** Prioritize dispatching based on:

High tipping zones.

Repeat demand during specific time slots (using historical data).

**Shared Ride Optimization:**

Use high passenger-count zones to suggest pooled rides.

Encourage shared rides during airport peak hours.

**Idle Time Reduction:**

Reposition vehicles to adjacent zones where demand is predicted to spike within 15–30 minutes (based on pickup trends).

**Alert System:**

Flag trips with unusually low fare per mile or duration >1 hour for inspection or dynamic rerouting.

* + 1. **Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analyzing trip trends across time, days and months.**

**Segment Zones Based on Peak Demand Windows :**

**Commercial Zones:** Weekdays, 7–10 AM & 4–8 PM due to Office commute patterns (e.g., Midtown, Downtown)

**Airport Zones:** All days, 6 AM–10 AM & 6 PM–12 AM due to Regular flights, business & tourist travel

**Nightlife/Hotels:** Weekends & Fridays, 9 PM–2 AM due to Bars, restaurants, clubs (Chelsea, East Village, Soho)

**Tourist Hotspots:** Weekends and holidays, 10 AM–6 PM due to Statue of Liberty, Central Park, Times Square

**Residential Zones:** Weekdays, 6–9 AM & 6–9 PM due to To cater to home-to-office and return trips

**Weather Forecasting:** Position more cabs near transit hubs, shopping malls, and residential areas during rain or snow.

**Event Calendar Integration:** For zones near stadiums, parks, concert halls, dynamically increase fleet during events.

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

From this data analysis following is the current status:

* ***Short trips (< 2 miles):*** *High base fare impact → high fare per mile*
* ***Medium trips (2–5 miles):*** *Most frequent → optimize this band for volume*
* ***Long trips (> 5 miles):*** *Lower fare per mile → risk of underpricing*

**Adjustment Strategy:**

* **0–2 miles:** Slight increase in base fare, reduce per-mile rate slightly to remain competitive
* **2–5 miles:** Keep fare per mile slightly above average vendor fare
* **5 miles**: Introduce a minimum fare floor or a long-trip surcharge to ensure profitability

**Time-of-Day Dynamic Surcharges From trip trends:**

* Rush hours and late nights see peak demand
* These hours offer willingness to pay higher

**Dynamic Pricing Proposal:**

**Time Slot & Adjustment:**

* **7–10 AM & 5–8 PM:** Add peak hour surcharge (₹5–10)
* **11 PM–3 AM (Fri/Sat)**: Add nightlife demand surcharge
* **2–6 AM**: Consider incentive-based fare cuts to boost usage

**Zone-Based Pricing Differentiation From zone analysis:**

* Airport zones have higher demand but fixed fees.
* Tourist zones (Times Square, Central Park) can bear higher rates.

**Proposal:**

* **Zone Type**: Pricing Recommendation
* **Airport pickups:**  Enforce fixed base + dynamic zone surcharge during rush
* **Tourist zones**: Slightly higher base fare with transparent reasoning
* **Low-demand zones**: Offer off-peak discounts to increase usage