Report: Optimising NYC Taxi Operations

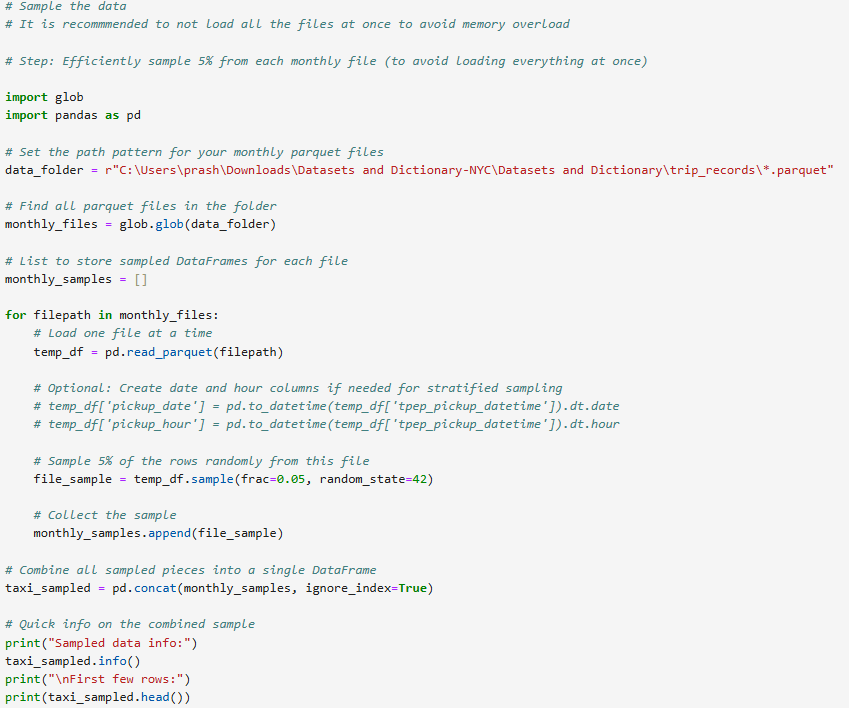
Include your visualizations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

## Data Preparation



* 1. Loading the dataset



* + 1. **Sample the data and combine the files**

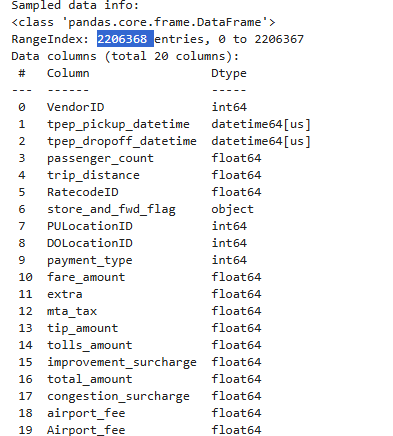
 Iterated over each 2023 monthly Parquet file and parsed pickup timestamps to extract **date** and **hour**.

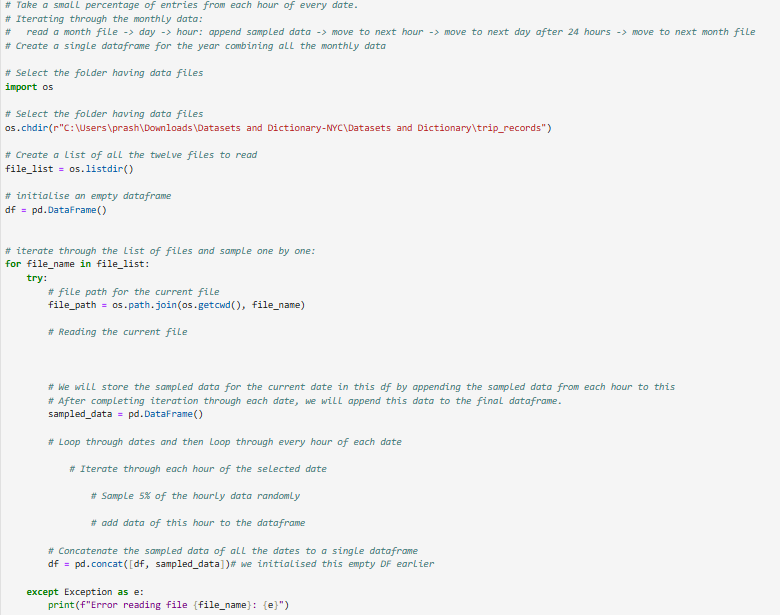
 For every distinct **date** within a month, looped through all **24 hours (0–23)**.

 Whenever an hour contained trips, drew a **5% random sample** of those records (with random state=1 for reproducibility).

 Combined the sampled hourly slices into a **monthly** Data Frame, then concatenated all months into a **single annual** dataset used for the remainder of the assignment.

Total number of records in sampled dataset: **2206368**.

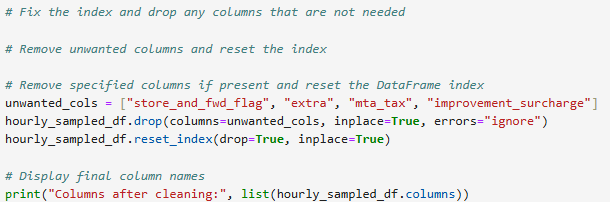
****



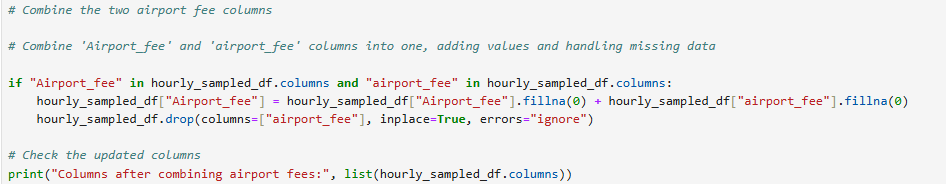
## Data Cleaning

### 2.1 Fixing Columns

**2.11 Fix the index**



1. Reset the Data Frame to the **default integer index**, replacing the previous index.
2. Reviewed **missing values** across all columns. The highest null rates are in the two airport-fee fields (airport fee and Airport fee). I left these untouched here because they’re addressed in the next step.
3. Assessed remaining columns and found store\_and\_fwd\_flag adds little analytical value and contains nulls, so I **dropped** it.

* + 1.  **Combine the two airport\_fee columns**

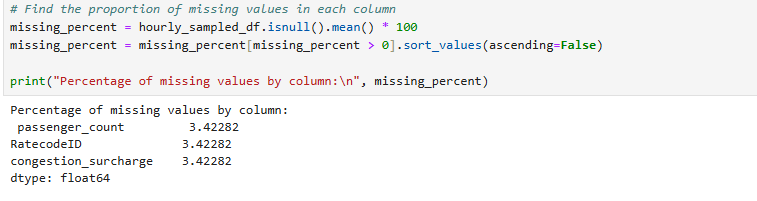
1. Assessed the distribution of both columns by computing the **median** (middle value) and **mode** (most frequent); **both were 0**.
2. Checked for any rows where **both columns had values simultaneously**; the **overlap count was 0**.
3. Based on these findings, **imputed missing values as 0** in both columns.
4. **Merged** the two fields into a single column named **airport\_fee**.

### Handling Missing Values

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

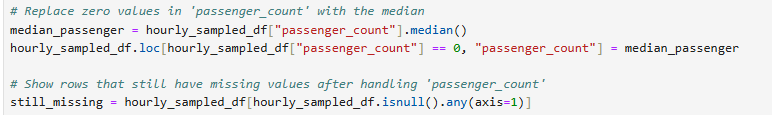
Highest proportion of missing values in columns ‘passenger\_count’,

‘RatecodeID’ and ‘congestion\_surcharge’.



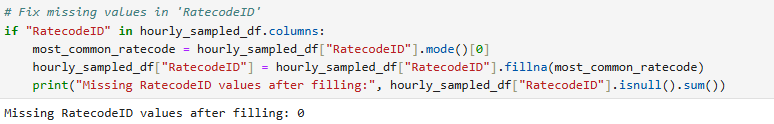
* + 1. **Handling missing values in passenger\_count**

1. Imputed missing values in `passenger\_count` using the (median).

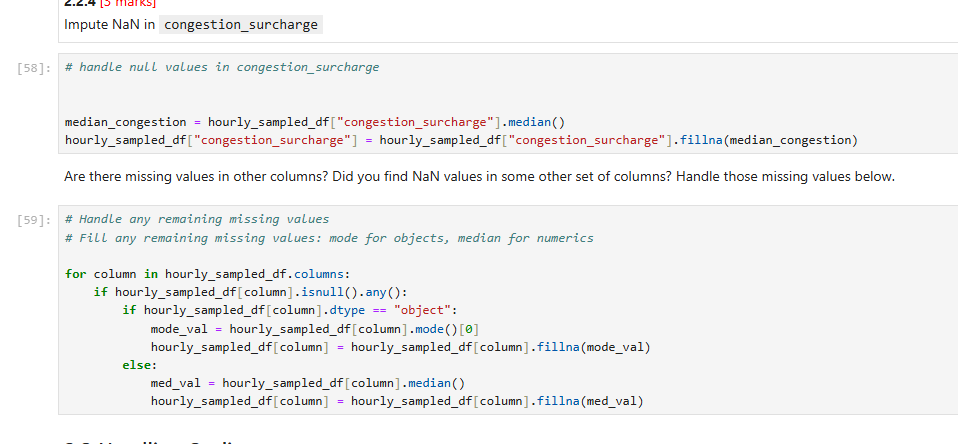
2. Treated zeros as invalid and replaced them with the mode (1), since the majority of trips in the data involve one passenger.

* + 1. **Handle missing values in RatecodeID**

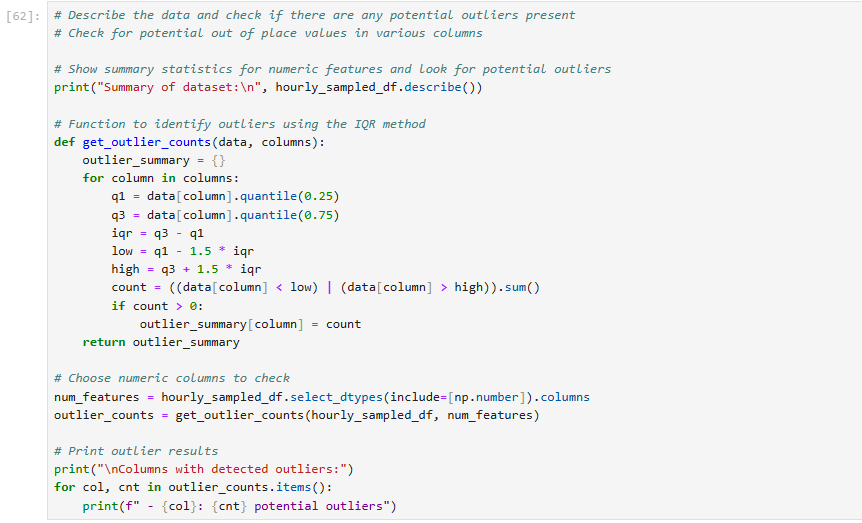
1.Since the **mode is 1** (standard rate) — indicating most trips use this code — imputed all missing RatecodeID values with **1**.



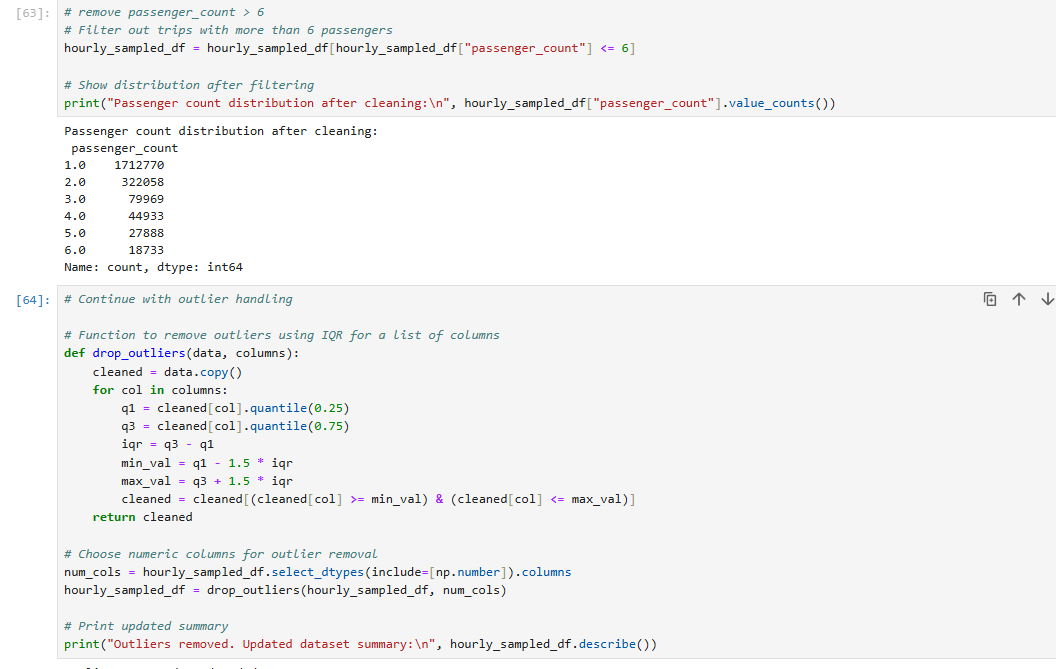
* + 1. **Impute NaN in congestion\_surcharge**



### Handling Outliers and Standardising Values



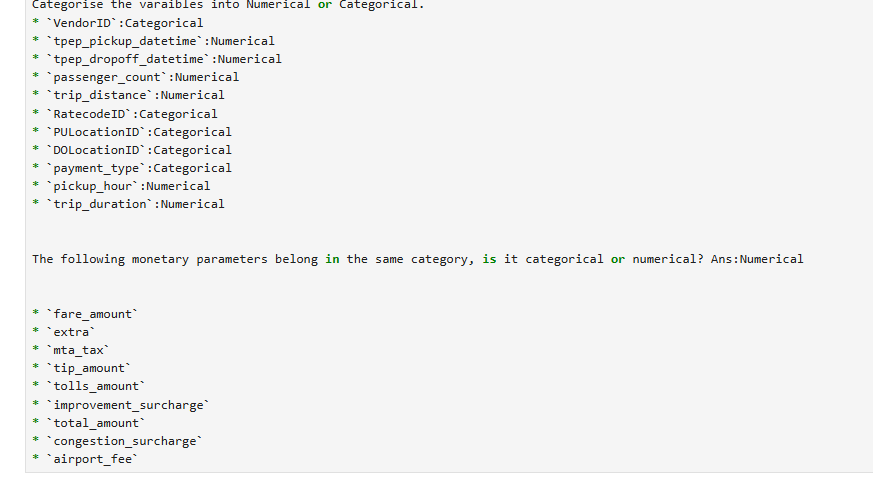
* + 1. **Check outliers in payment type, trip distance and tip amount columns**

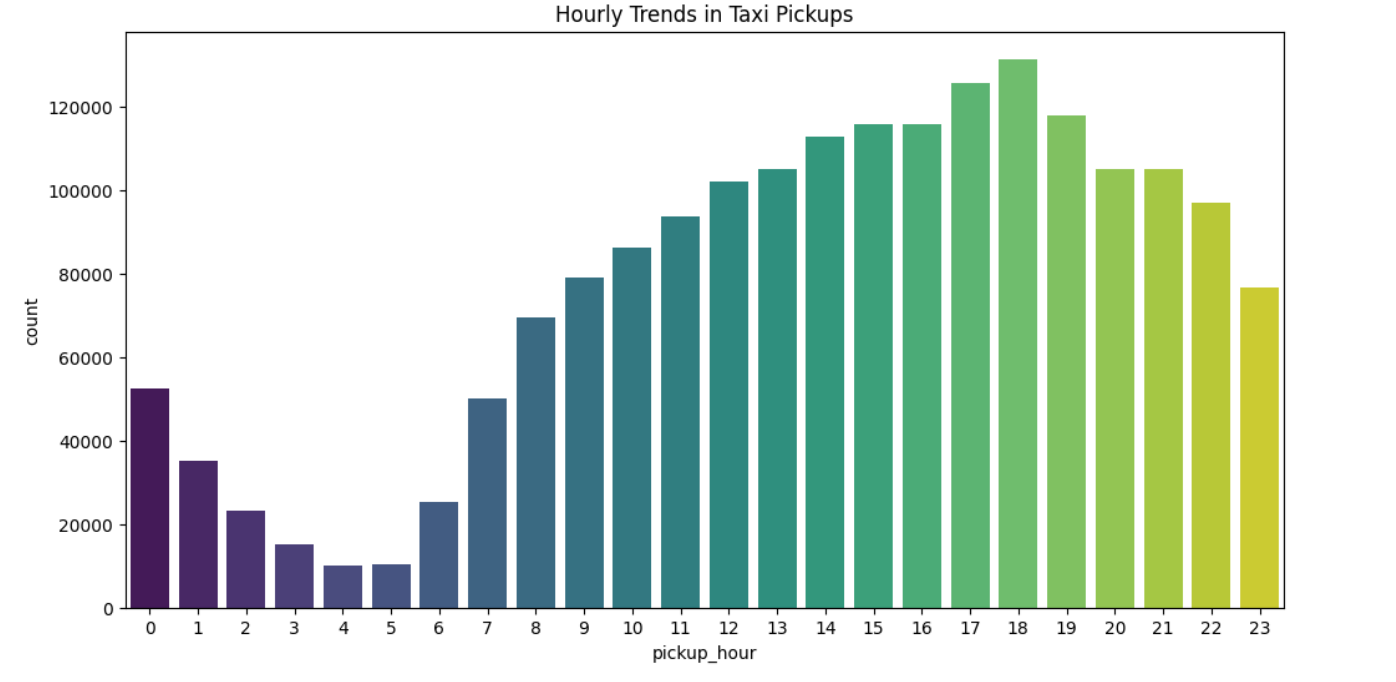


## Exploratory Data Analysis

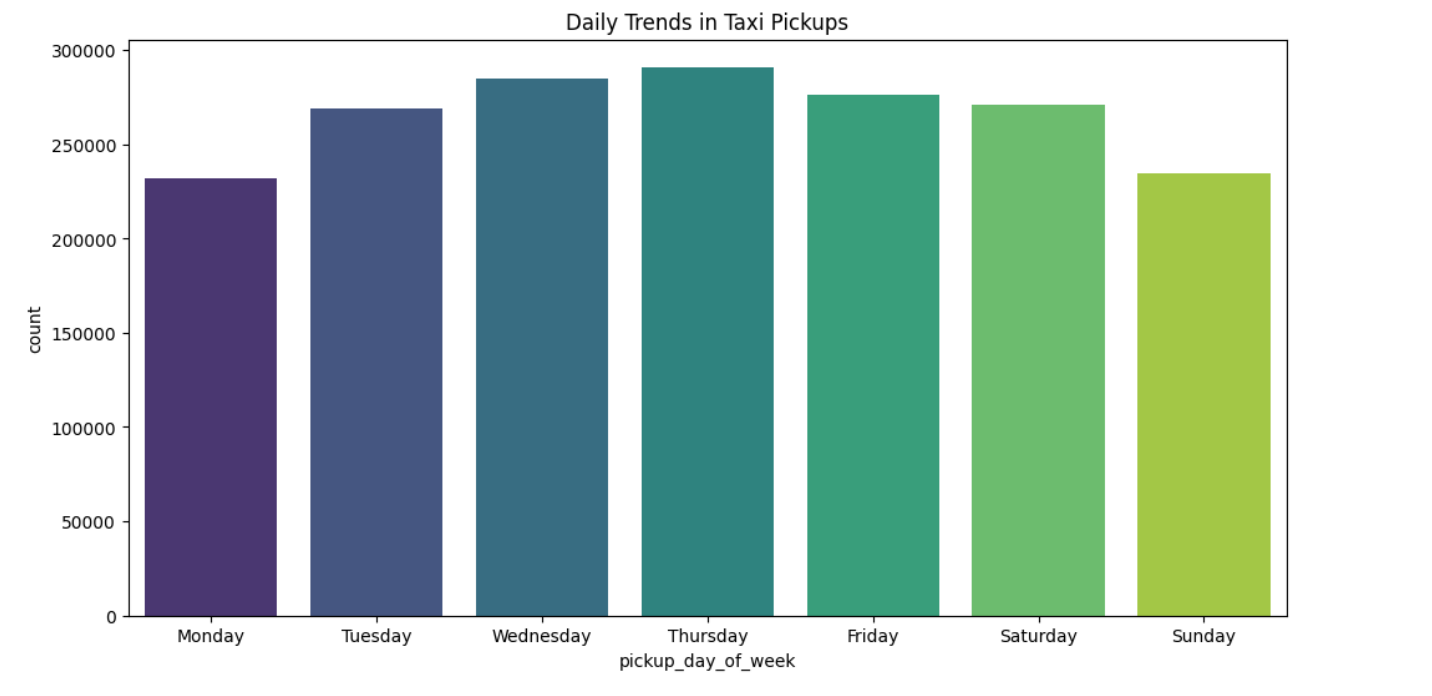
### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical**

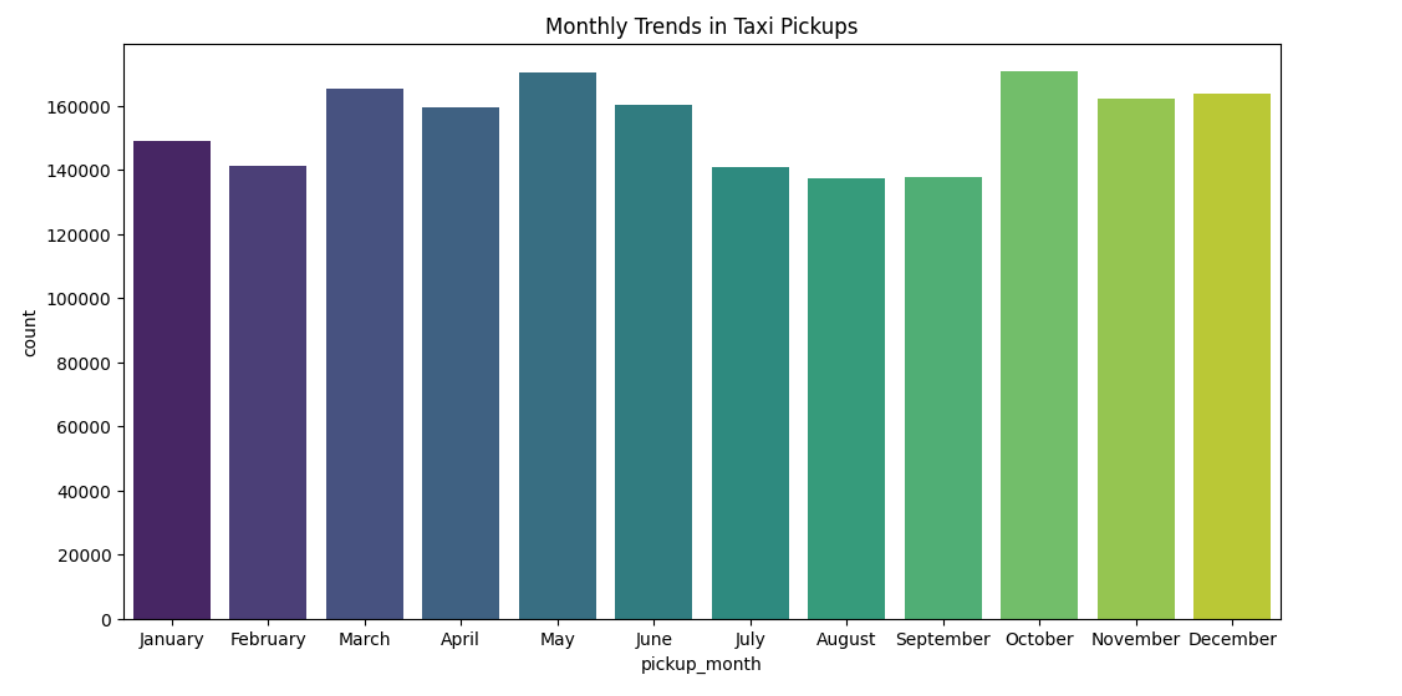


**3.1.2 Analyse the distribution of taxi pickups by hours, days of the week, and months**

**Taxi pickups rise from 6am and the peak reaches at 6pm, which indicates people travel during these hours for work.**

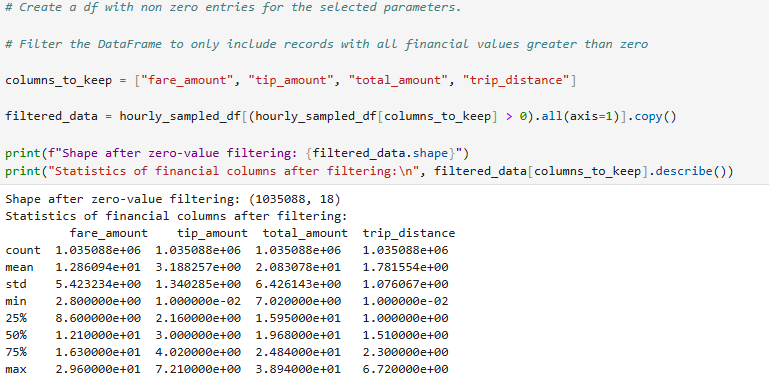
****

**Pickups are at highest from Tuesday to Friday, showing pattern of people traveling within the city, possibly to work or airport.**

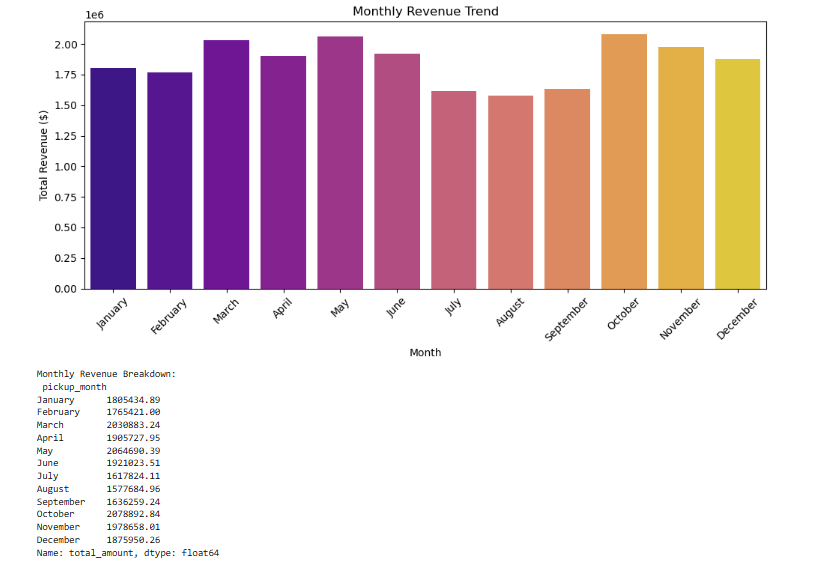
****

Highest pickups reported from May and October, with March, November and December close by. This indicates people traveling more during holiday season (starting from late Oct to December). On the other hand, March to May being spring break and summer vacation time, taxi pickups see a spike in usage.

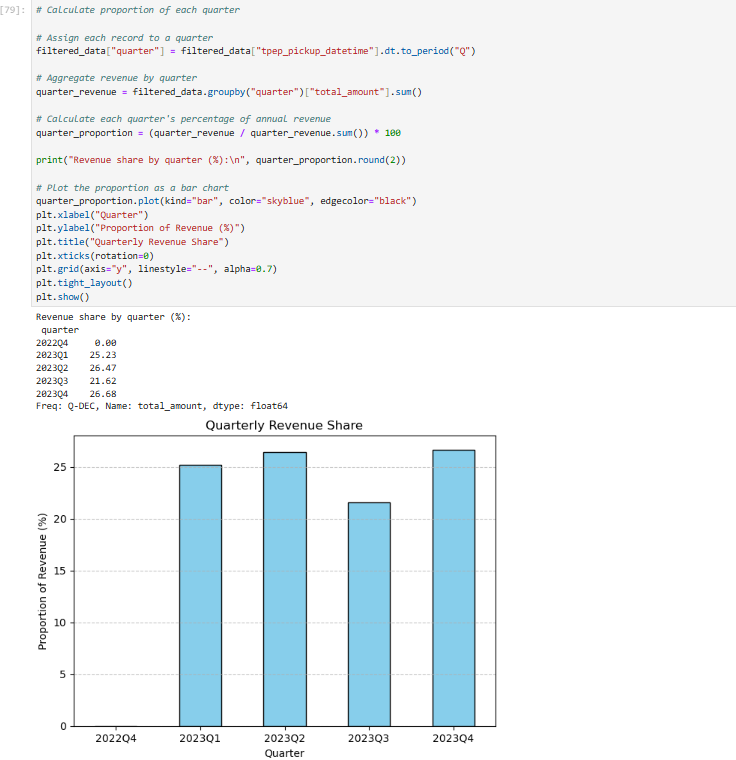
* + 1. **Filter out the zero/negative values in fares, distance and tips**

****

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

****

**Monthly revenue increases are at highest during May and October, directly proportional to the trend we saw in above when analysing monthly trends of taxi pickups.**

****

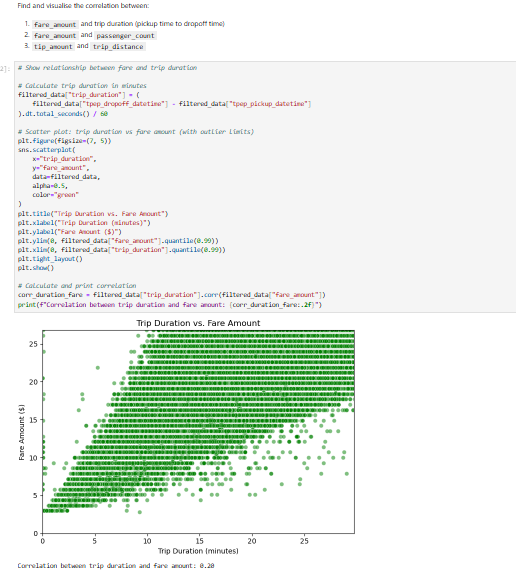
* + 1. **Find the proportion of each quarter’s revenue in the yearly revenue**

* + 1. **Analyse and visualise the relationship between distance and fare amount**



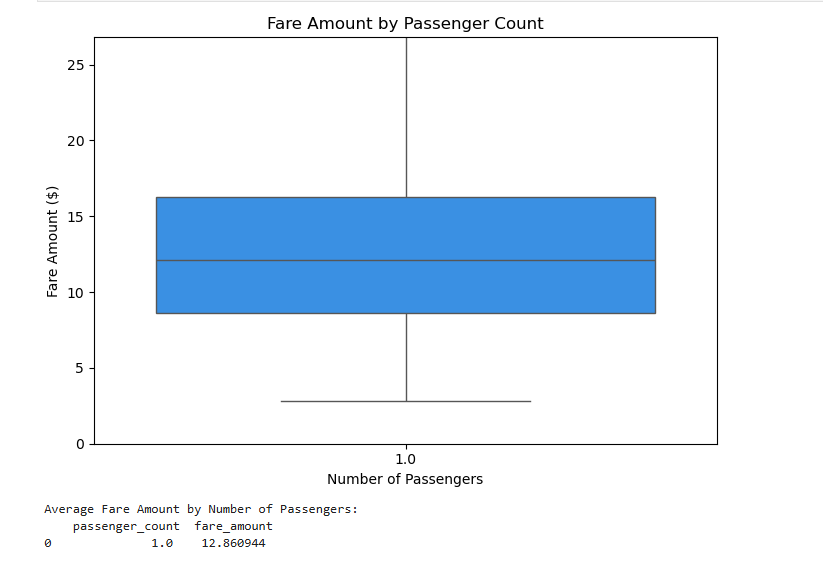
**Correlation between trip\_distance and fare\_amount:**

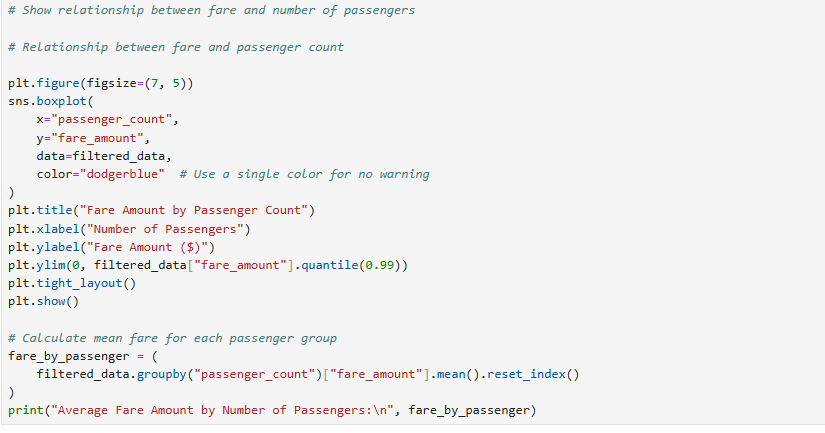
**strong positive correlation (0.87). This means that as the trip distance increases, the fare amount also tends to increase proportionally.**

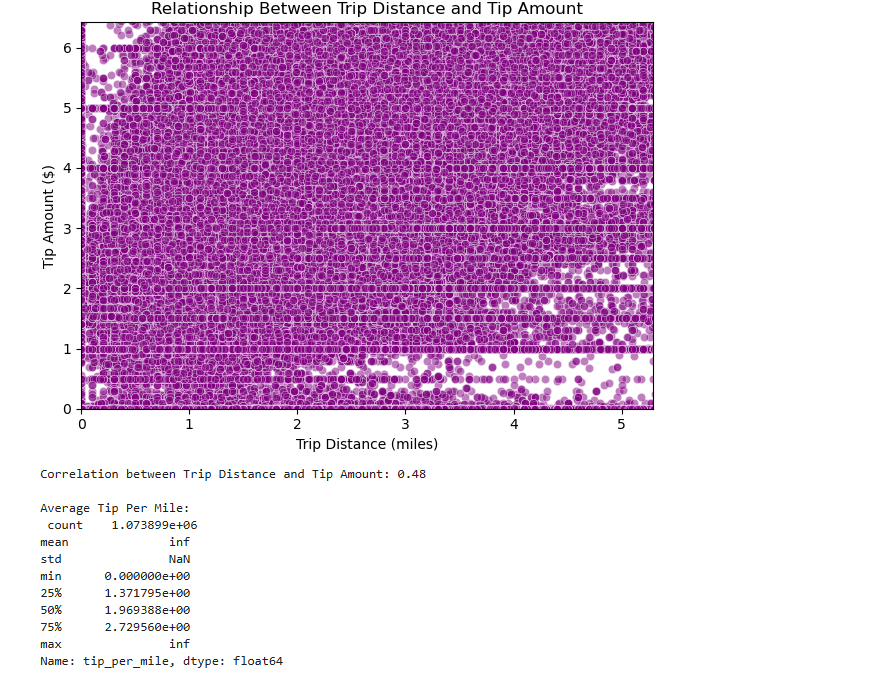
* + 1. **Analyse the relationship between fare/tips and trips/passengers**

**Observation: Trip duration is not directly proportional to fare amount. The correlation is also towards lower side (0.20). There are few outliers with high fare amount and low trip duration.**

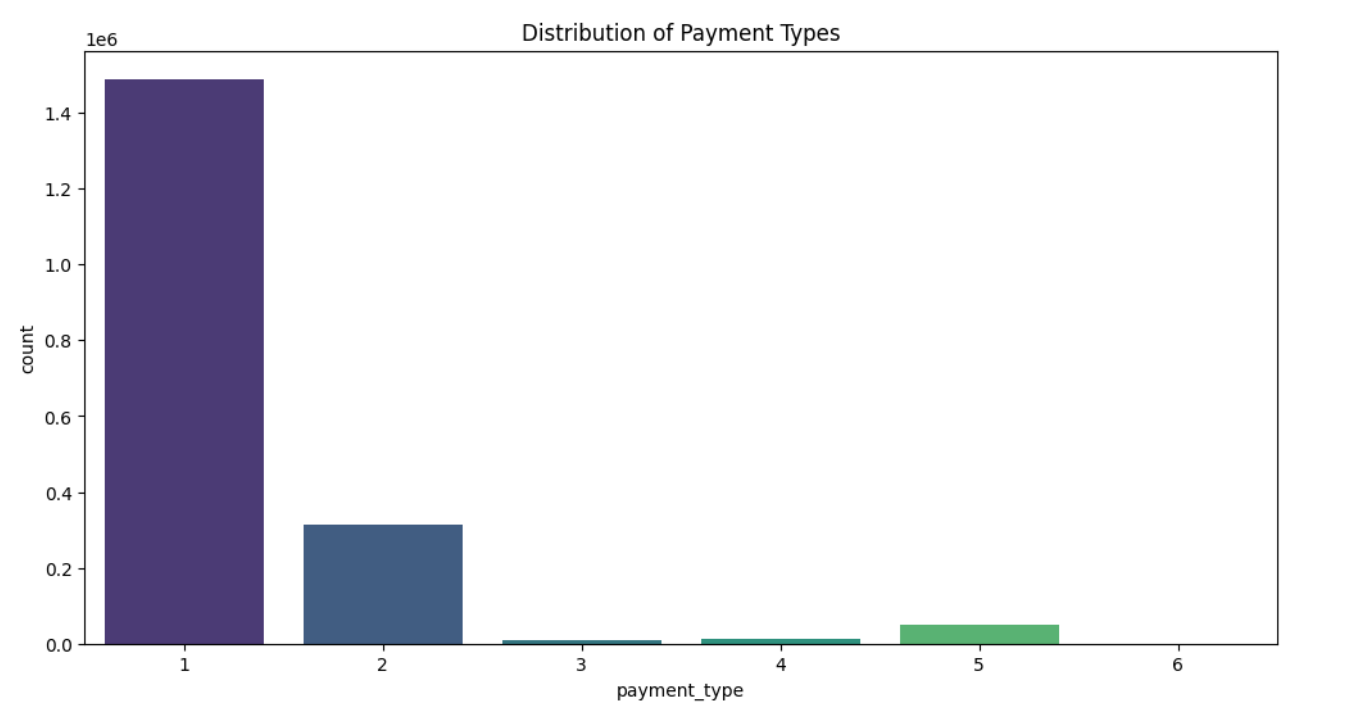
**This could be due to surge pricing during peak hours or other factors such as negotiation between driver and passenger.**

****

****



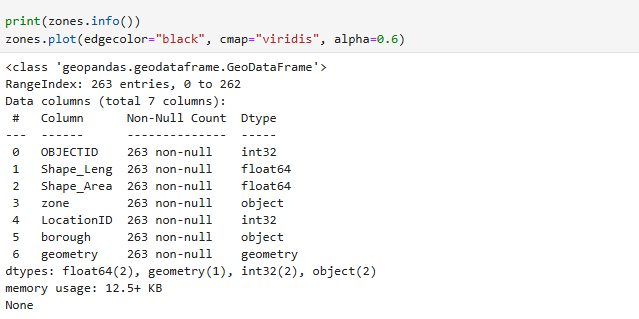
**Observation:** The tip amount is not directly proportional to the trip distance. Tip amount for few short distances is higher than the tip amount for longer distances. This could be due to external factors such as time of the day, negotiation between driver and passenger.

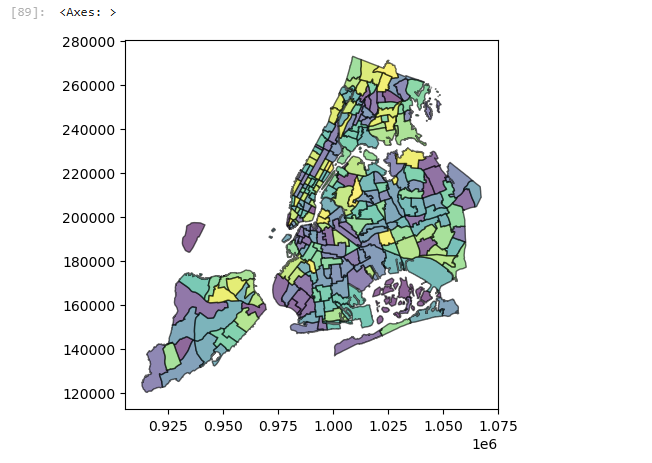
* + 1. **Analyse the distribution of different payment types**

**Observation:** The most common payment type is credit card (1), followed by cash (2). Number of voided trips (6) are zero.

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

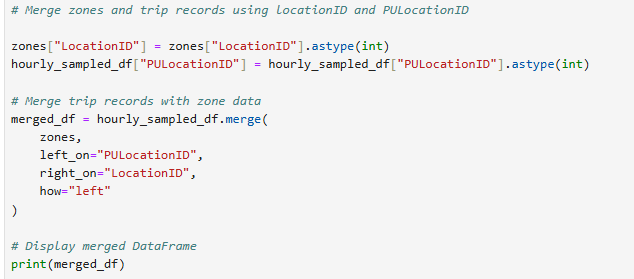
****

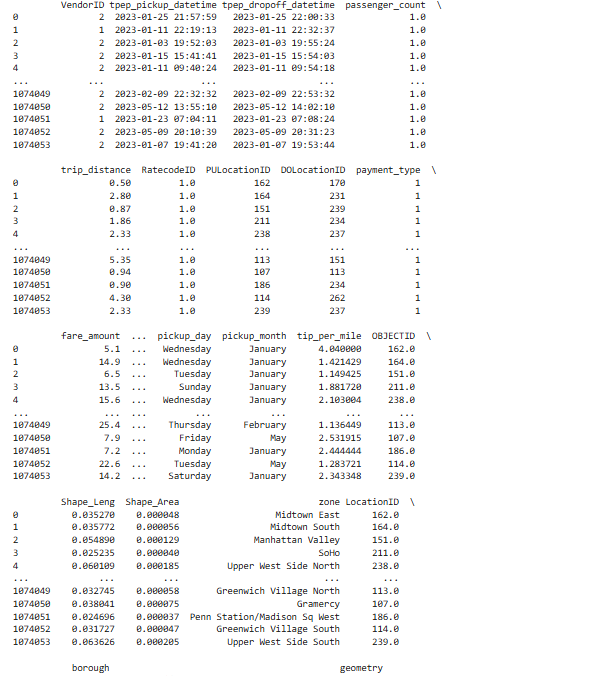
****

****

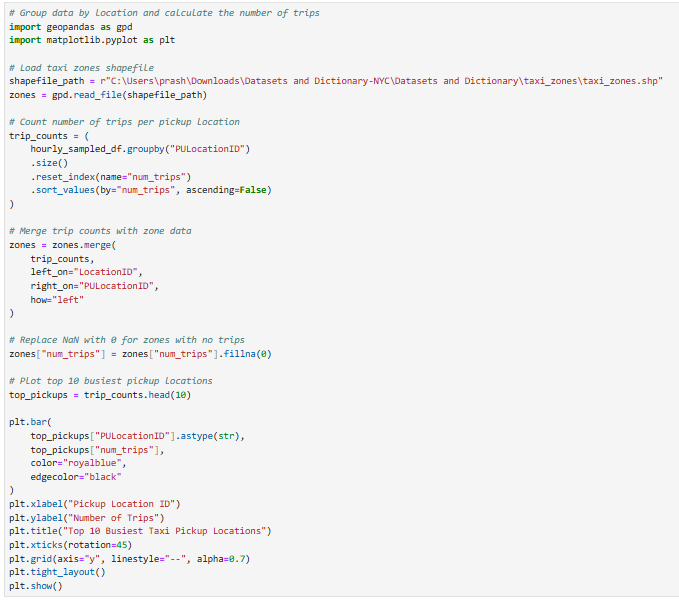
**1.1.4 Merge the zone data with trips data**

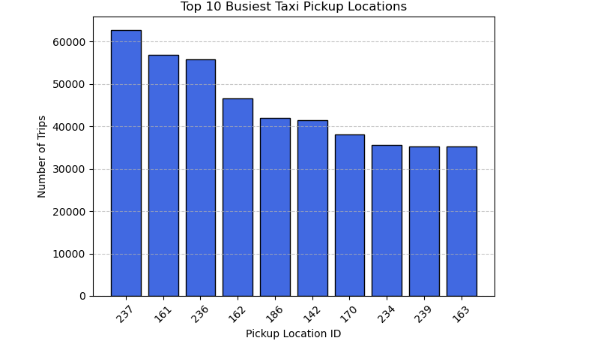
**Verified if ‘PULocationId’ column has any empty value, otherwise merging will result into missing zone info in those records.**

****

****

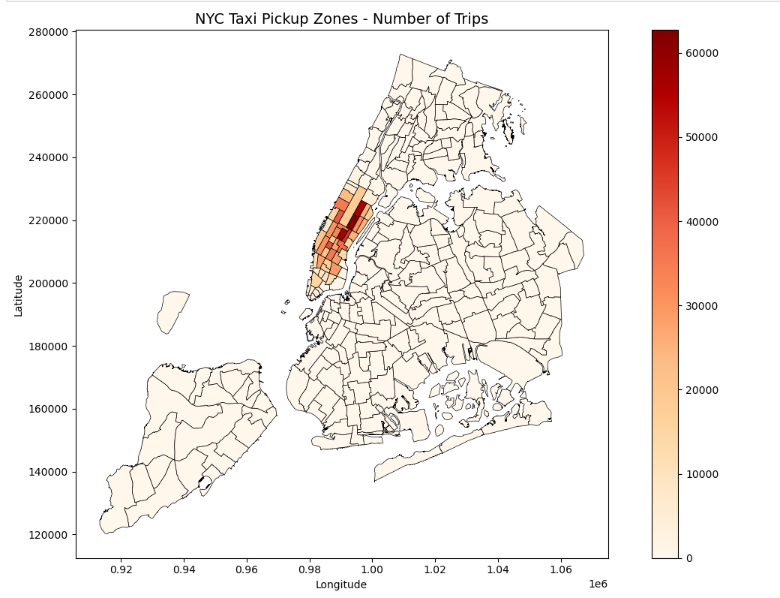
* + 1. **Find the number of trips for each zone/location ID**

****

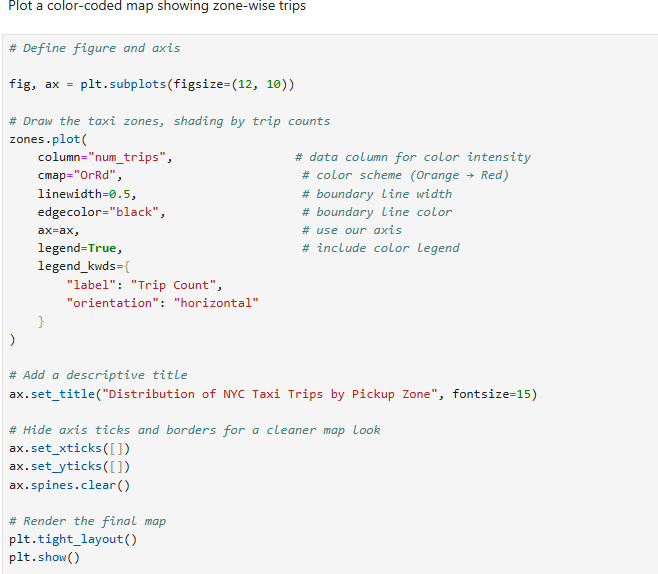
****

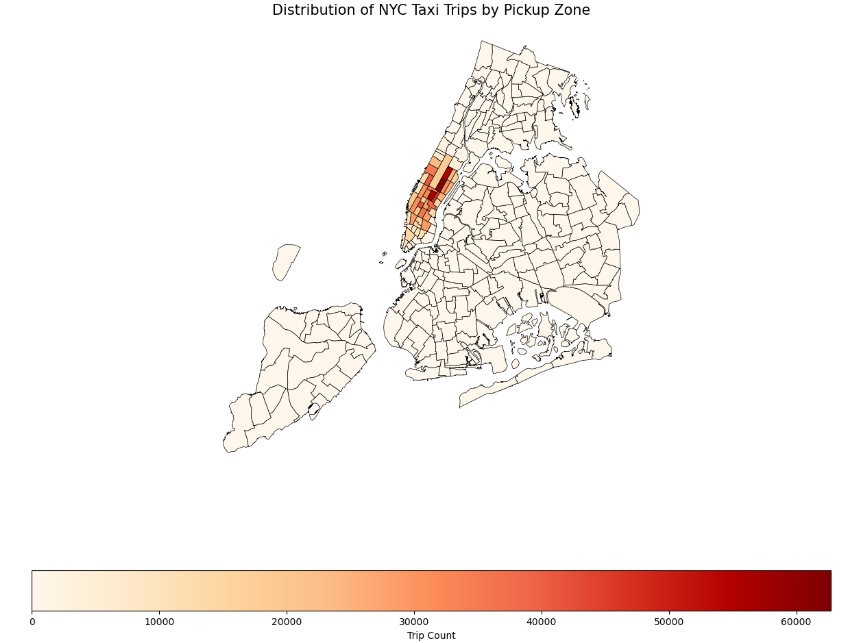
* + 1. **Add the number of trips for each zone to the zones dataframe**

****

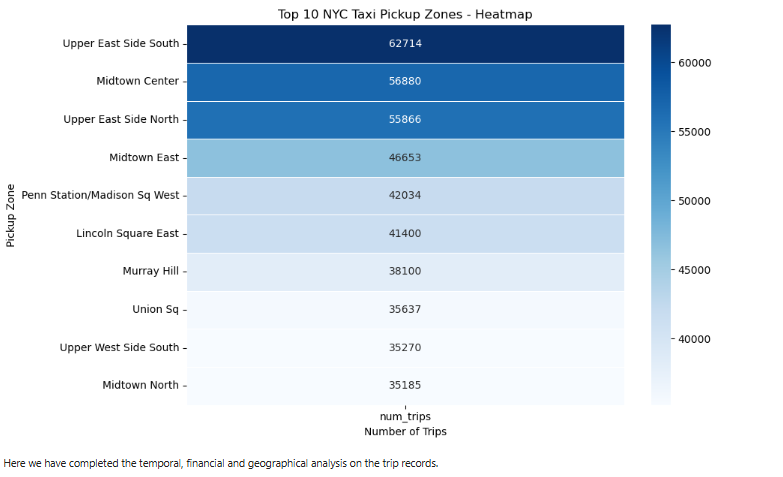
****

* + 1. **Plot a map of the zones showing number of trips**









* + - **Conclude with results**

1. **Taxi service usage:** The peak hours for taxi pickups is from 5:00 pm - 7:00 pm. The reason being people avail taxi at this time to travel from office to home during weekdays, and go out for dinner during weekends.

Weekdays, especially Wednesday, Thursday and Friday, have higher taxi pickups/drop-offs. Possible reason being most people work from office during mid-week.

Spring Break (March), Summer vacation (May) and Holiday season (Oct-Dec) see the

highest taxi service usage.

2. **Trends in revenue collection trend:**

Revenue collection is at its highest in Q2 (April, May, June) and Q4 (Oct, Nov, Dec),

aligning with the higher taxi activity during these periods, aligning with the higher taxi service usage during these periods, with Q4 being the peak due to the holiday season.

**Taxi fare:**

Fare amount vs Trip distance: There is a strong positive correlation between trip distance and fare amount. Longer trips result in higher fares.

- Fare amount vs Trip duration: There is also a positive correlation between trip duration and fare amount, although the correlation is less compared to above. In some instances, shorter trips have resulted in higher fare amount, indicating surge pricing and/or negotiation between driver and passenger.

- Fare amount vs Passenger count: Passenger count = 4 usually results in highest fare, however more than 4 passengers have shown lower fare trend.

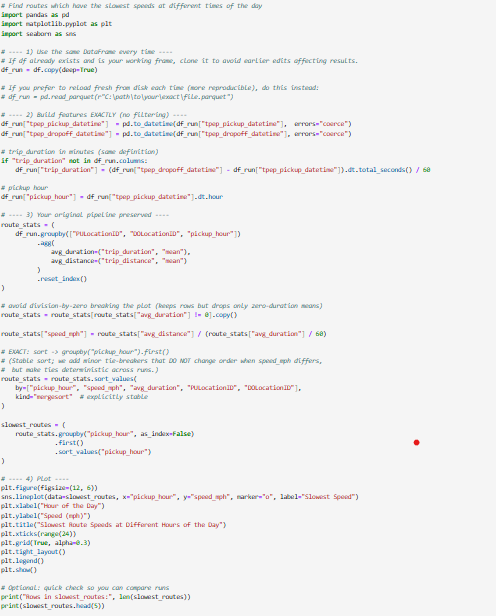
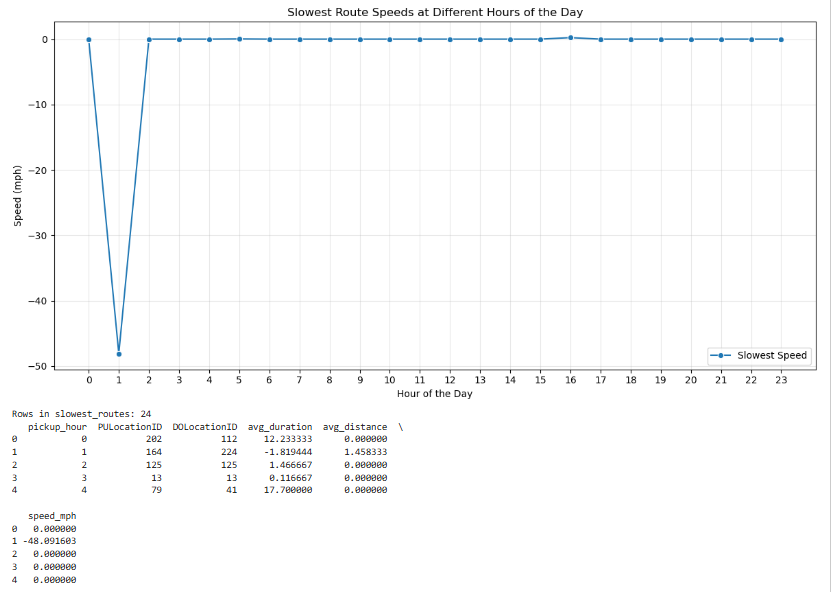
- Tip amount vs Trip distance: The correlation is positive, however as per the visualization, shorter trips also resulted in higher tip amount. Possible reason being time of the day/night and negotiation between passenger and driver.

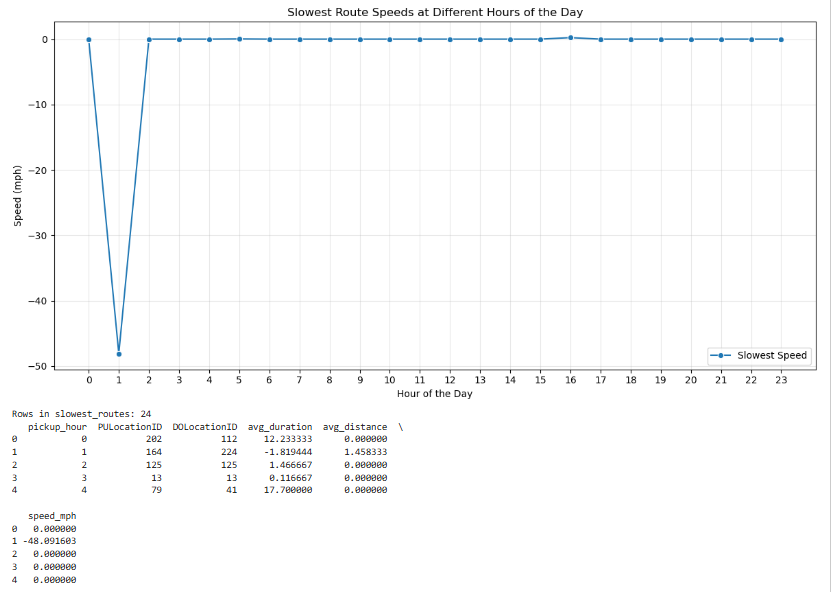
**Busiest Zones:**

- Top pickup location is JFK Airport, followed closely by Upper East Side South, Midtown canter and Upper East Side North.

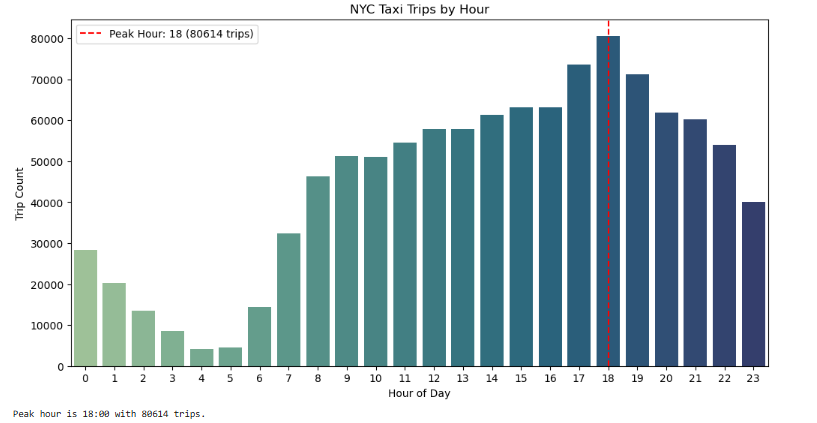
### Detailed EDA: Insights and Strategies

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

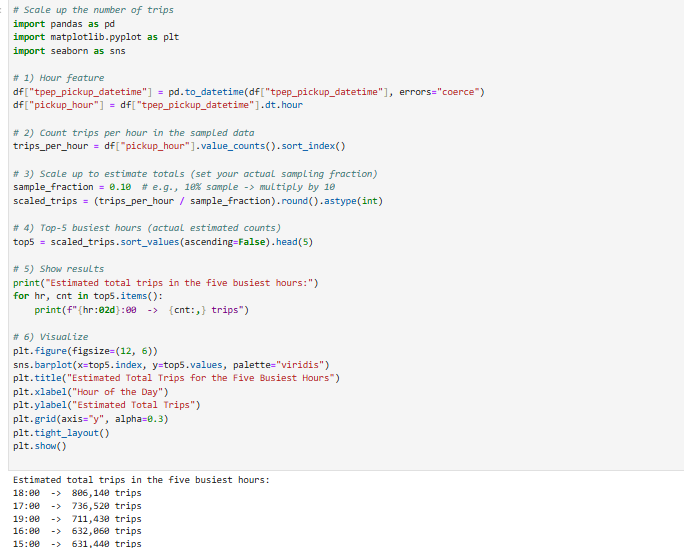


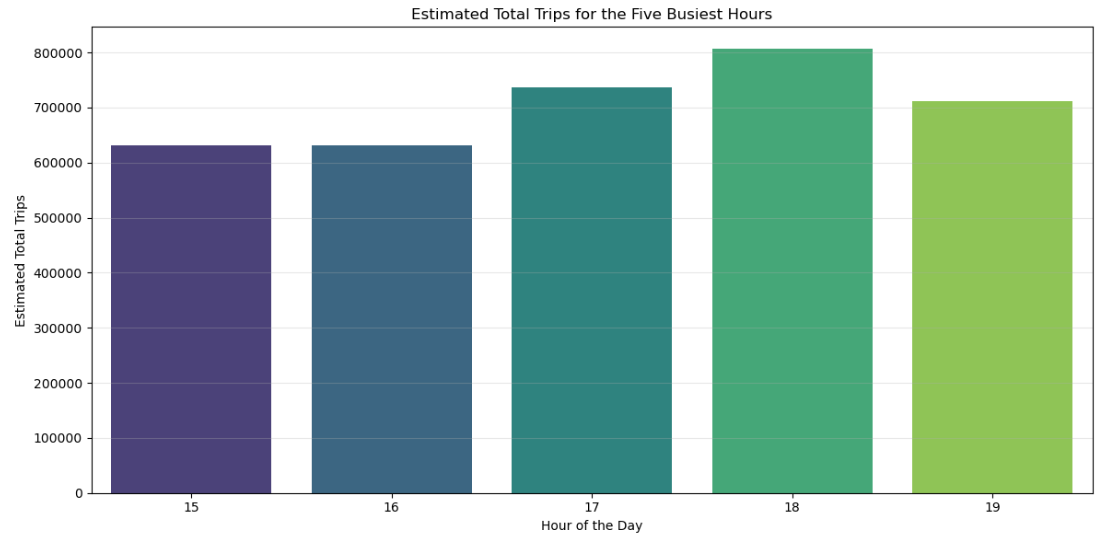
* + 1. **Calculate the hourly number of trips and identify the busy hours**

****

**Observations: Busy hours are during evening 5pm to 7pm, indicating people returning to home**

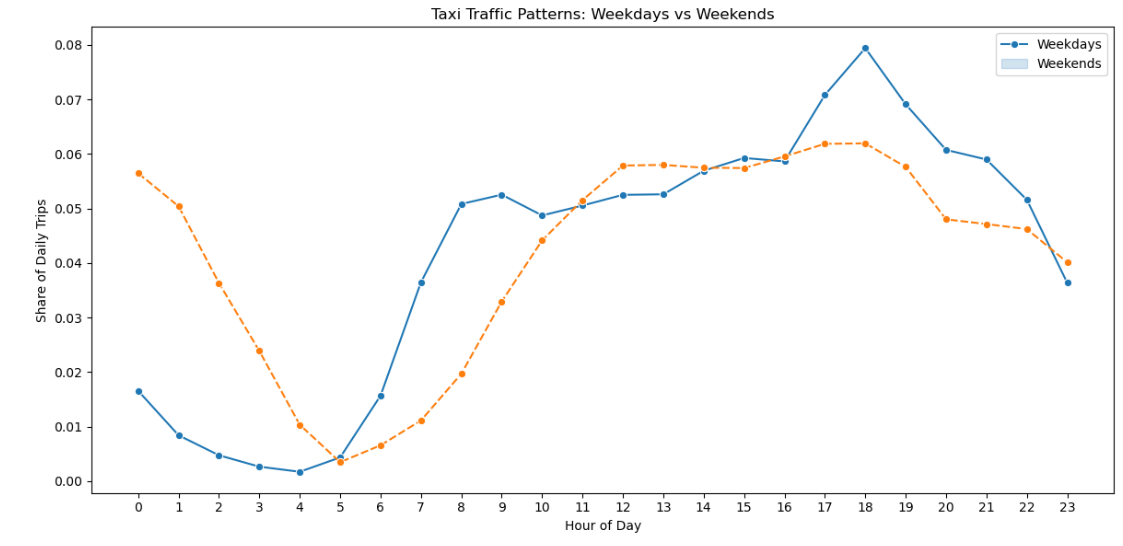
* + 1. **Scale up the number of trips from above to find the actual number of trips**

****

****

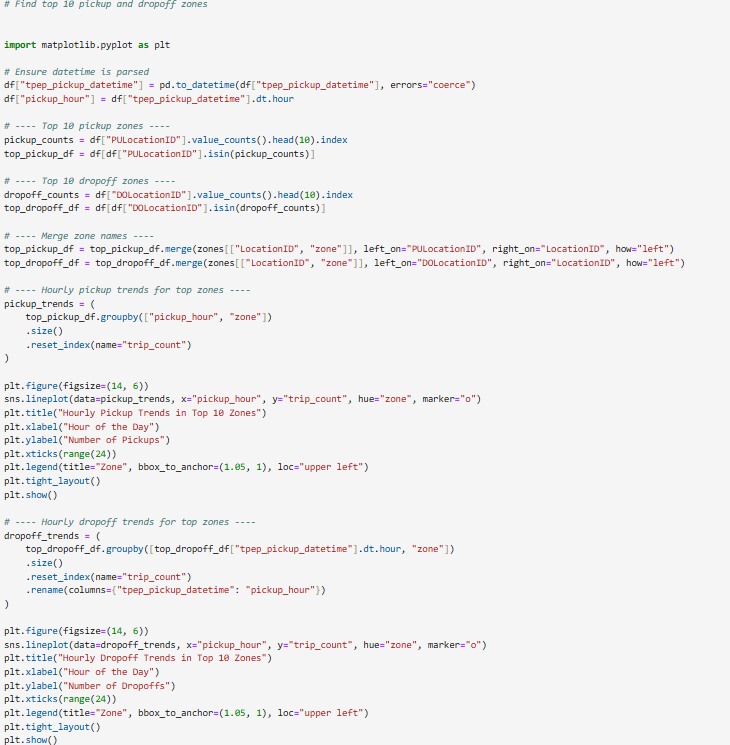
* + 1. **Compare hourly traffic on weekdays and weekends**

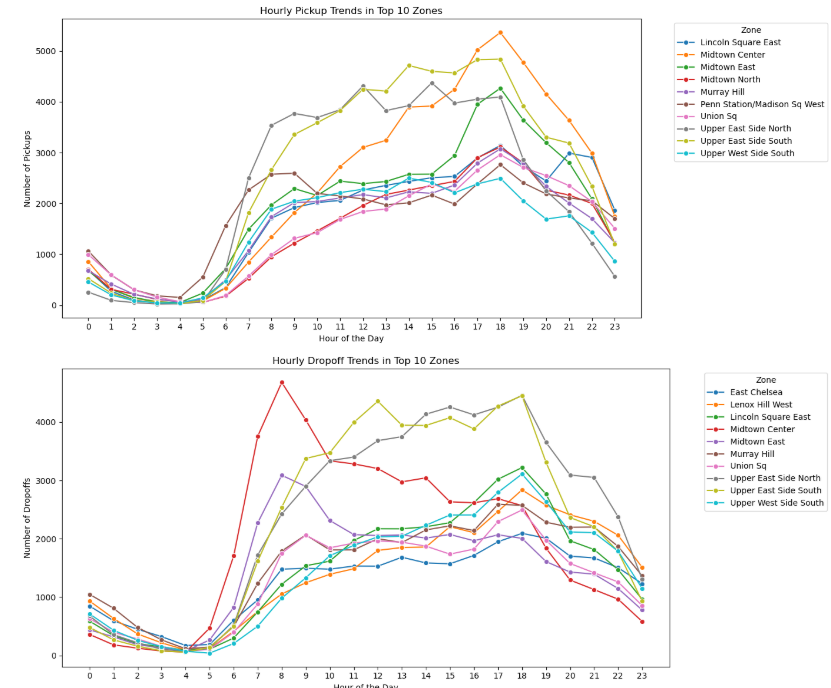


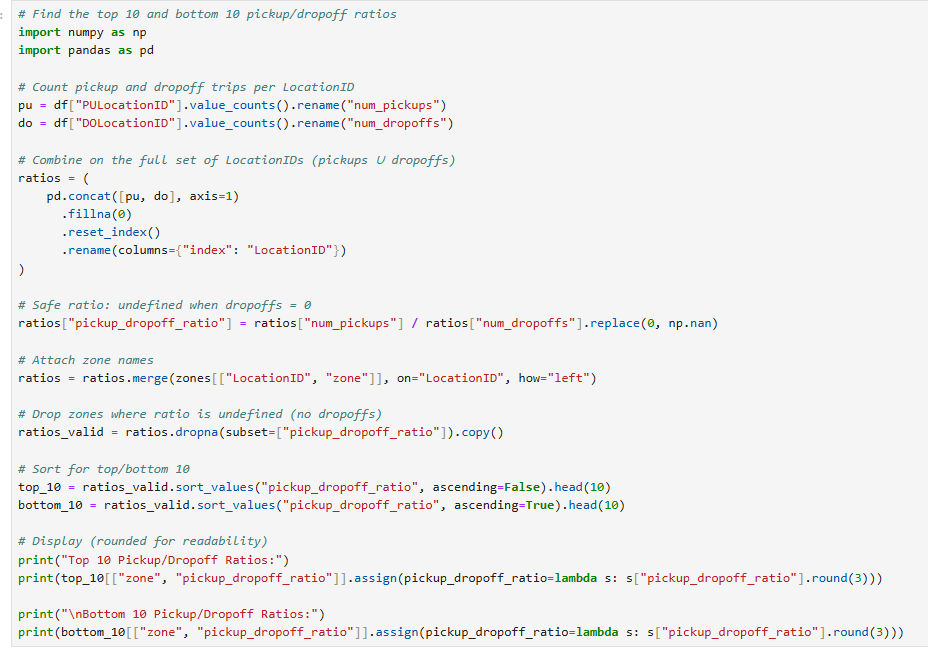
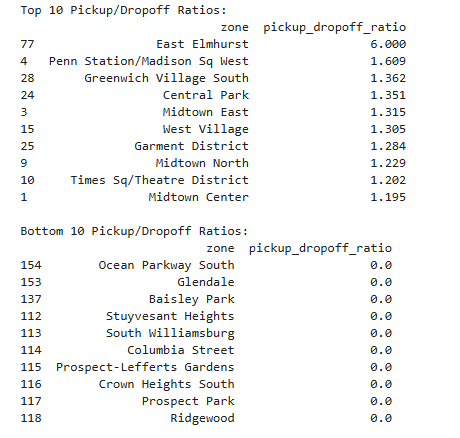


**Observation:** Weekday trips tend to go higher during evening hours, while weekends mostly

see a flat trend in number of trips throughout the day

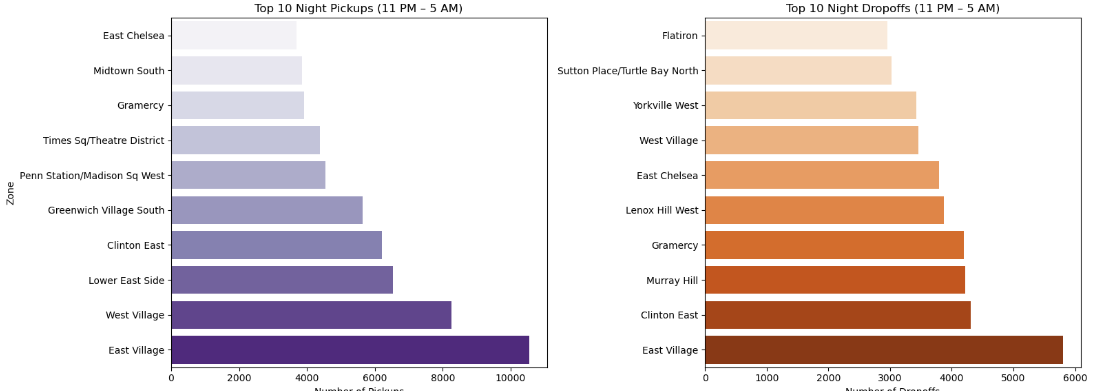
* + 1. **Identify the top 10 zones with high hourly pickups and drops**



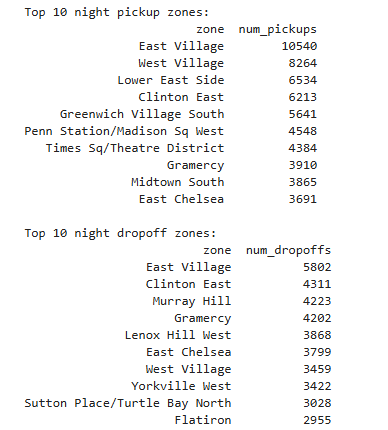
* + 1. **Find the ratio of pickups and dropoffs in each zone**

* + 1. **Identify the top zones with high traffic during night hours**



****

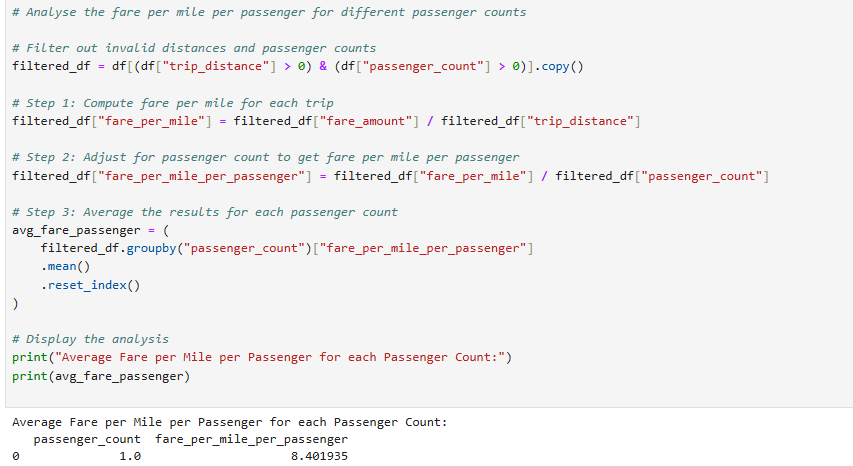
**Observation: People tend to avail taxi service more during night hours from airports, with some pick-up locations residential areas like East village. Drop-off locations are all mostly residential during night, with exception of times square.**

****

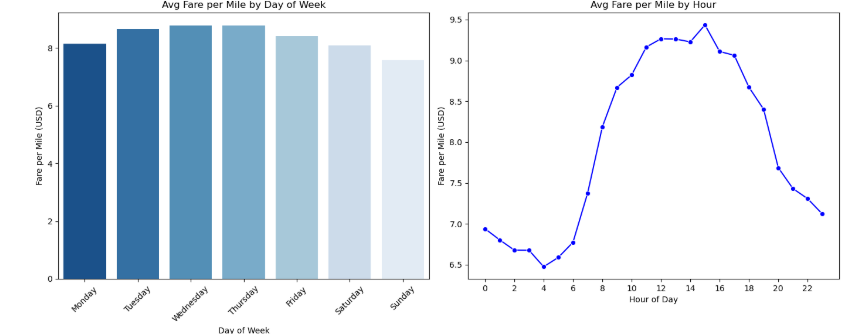
* + 1. **Find the revenue share for nighttime and daytime hours**

**   
Observation: Revenue tends to be highest during early morning (5am), and evening hours. The explanation for night time revenue can be higher tip amount given to driver.**

* + 1. **For the different passenger counts, find the average fare per mile per passenger**

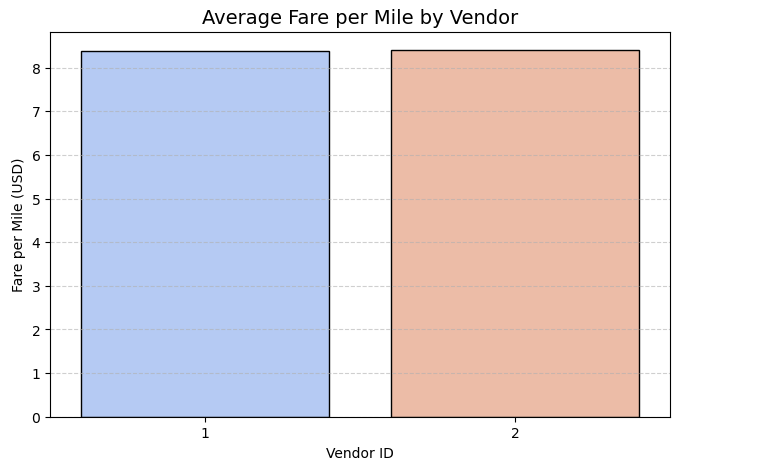
****

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

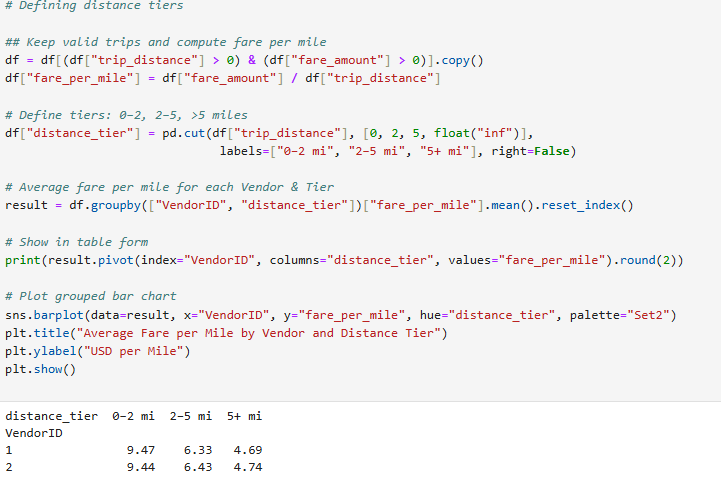
****

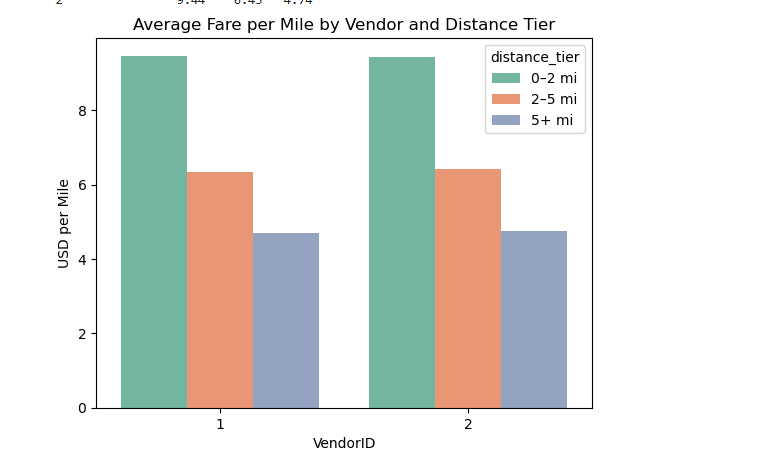
* + 1. **Analyze the average fare per mile for the different vendors**

****

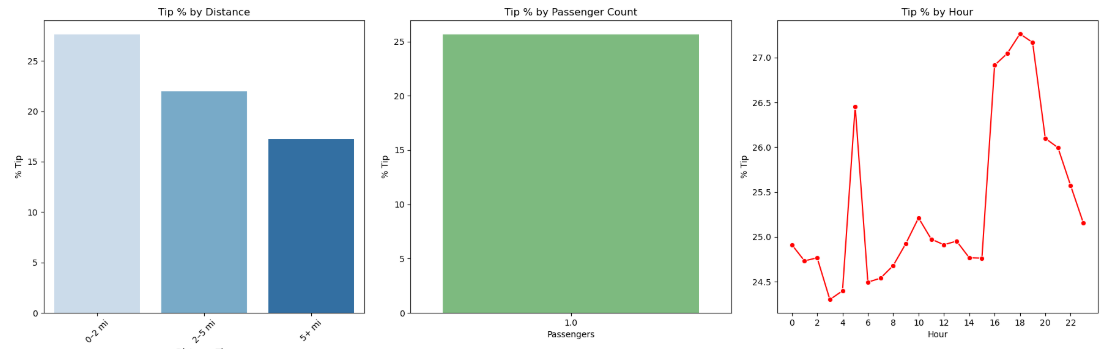
****

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

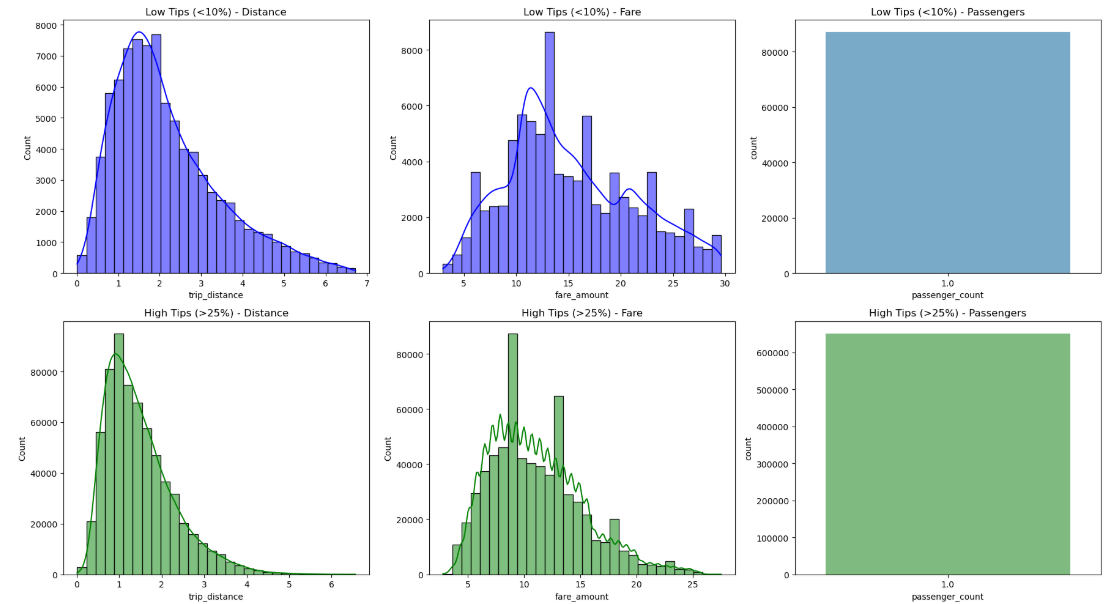
****

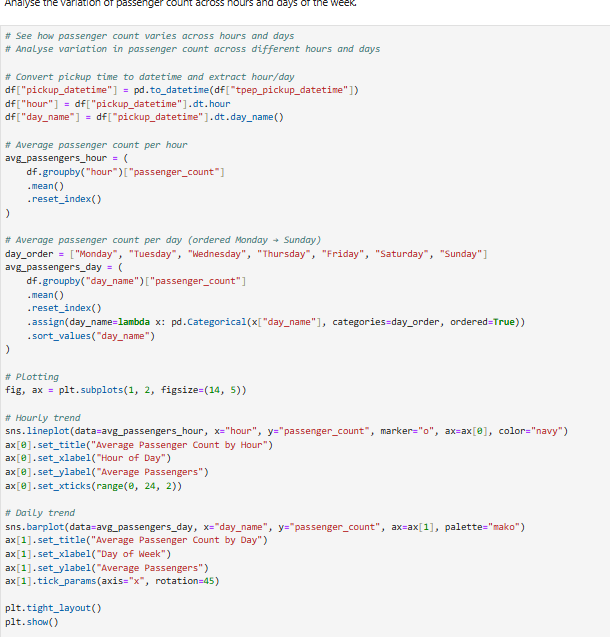


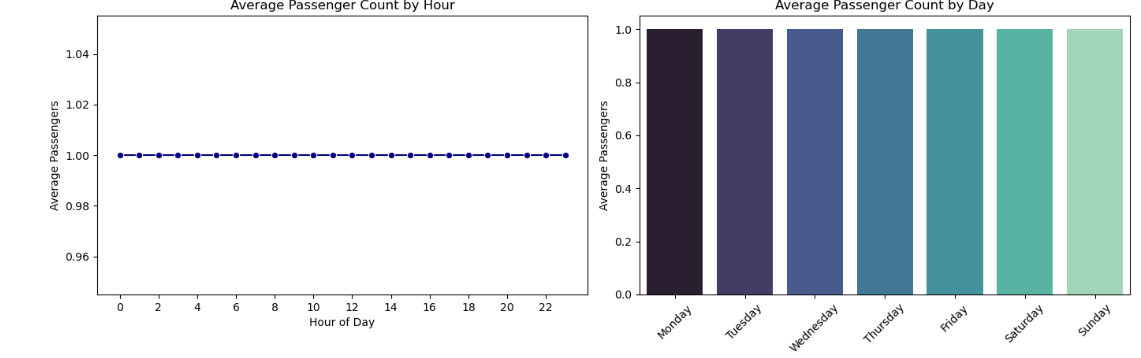
* + 1. **Analyse the tip percentages**

****

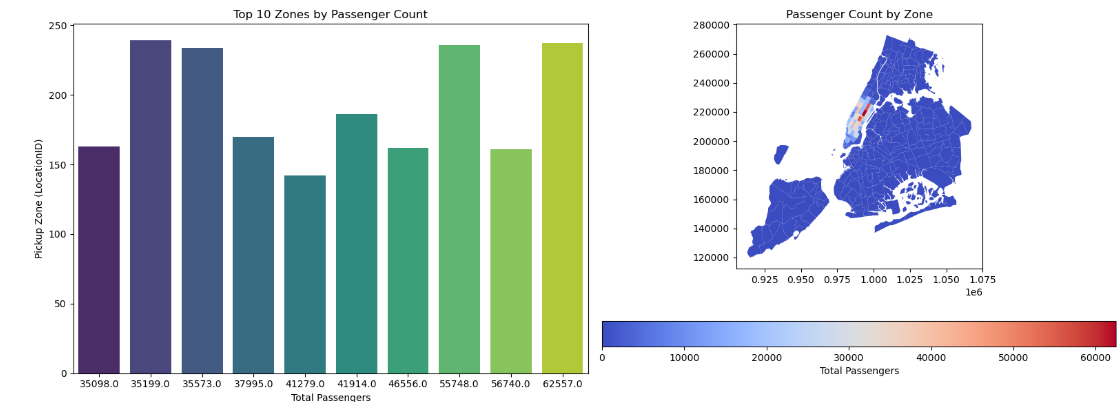
**Observation: 1. Tip amount increases directly with distance, especially from 0 to 20 miles bracket. 2. Drivers get more tips for passenger\_count > 2, however above 5 passengers, it shows a pattern of decrease in tips. 3. Drivers get more tips during night hours (2am – 6am) and during office timing (10am–15pm).**

****

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

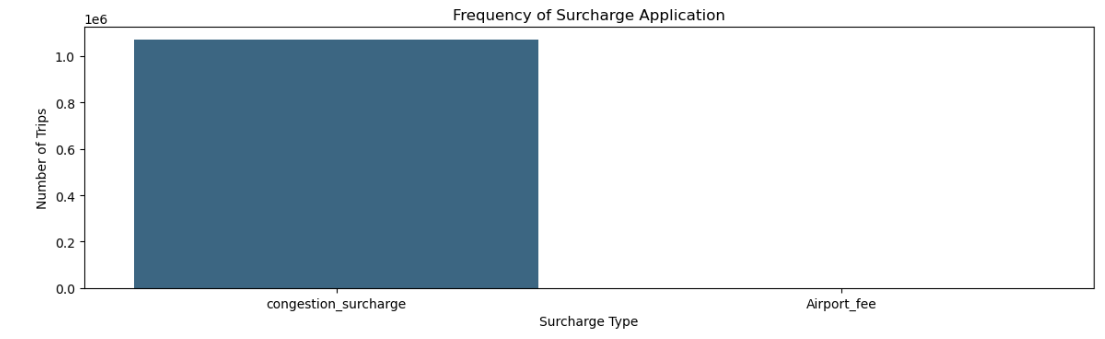


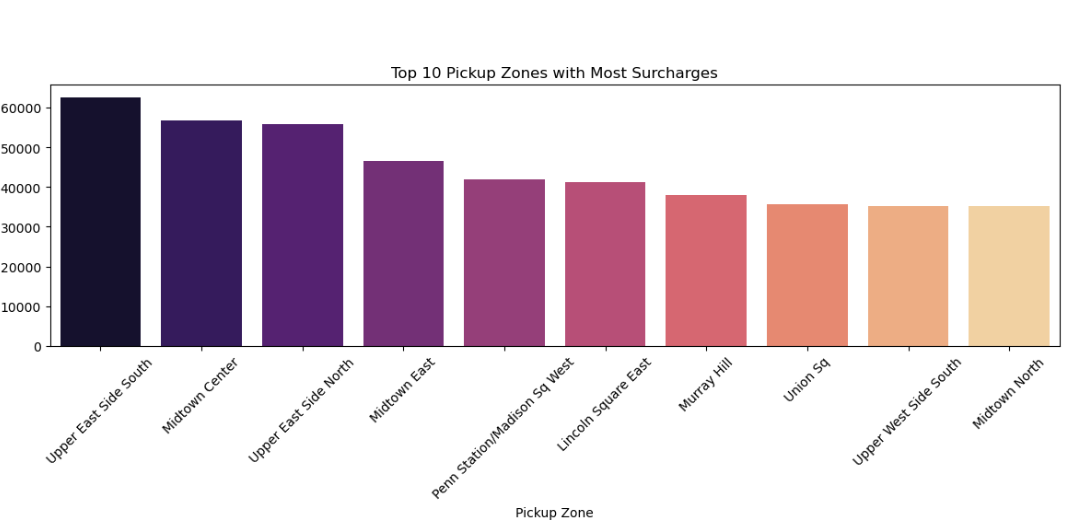
* + 1. **Analyse the variation of passenger counts across zones**

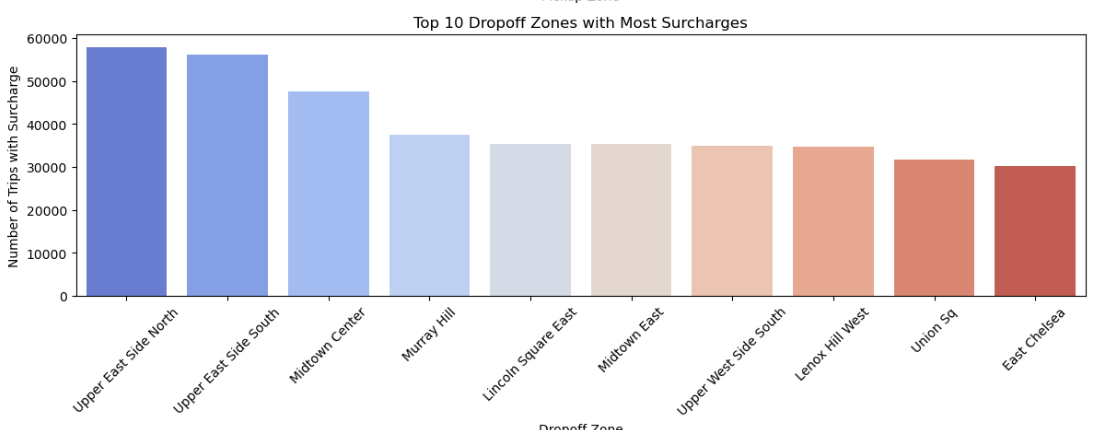


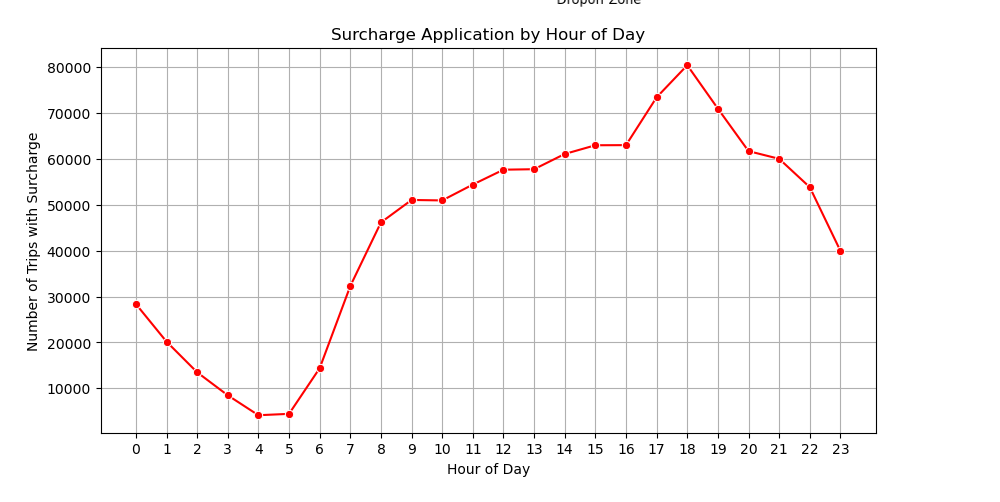
Took only top 10 records so that we don’t make a crowded chart.

* + 1. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.  
       **









## Conclusions

### Final Insights and Recommendations

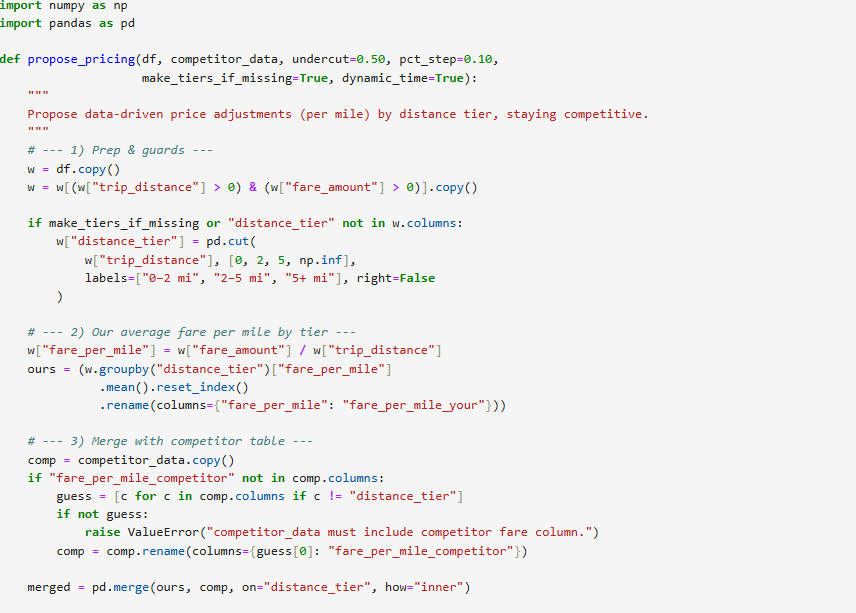
* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**

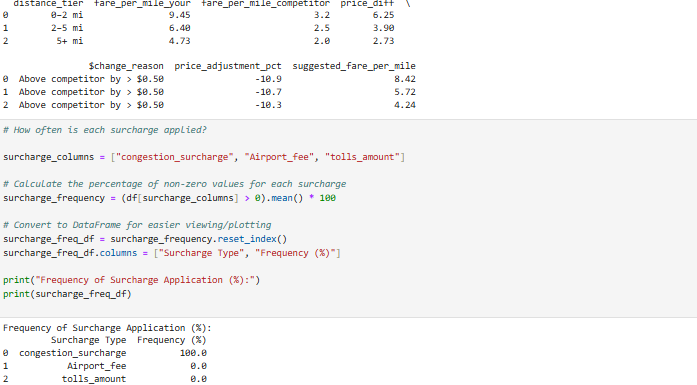
Boost system efficiency by combining dynamic pricing, forward-looking demand analytics, predictive planning, smarter routing, and tailored driver incentives. Use targeted driver allocation informed by external signals and user behavior (e.g., seats booked/occupied), address surcharge fairness with multi-objective optimization, and keep algorithms adaptable through continuous monitoring and A/B testing.

* + 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.**

Position the fleet strategically by aligning taxi supply with fare-per-mile trends, weekday vs. weekend patterns, and peak-hour demand. Use incentives and surge pricing to balance supply and demand, accounting for party size, tipping behavior, trip length, and surcharge hotspots. Adjust dynamically with real-time traffic, weather, and event data aided by predictive models, and continually refine zone-level distribution by tracking KPIs and learning from historical performance.

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





**THANK YOU!!!**