REPORT

Optimizing 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.

**Objective :** The goal of this EDA was to analyze and understand the NYC Taxi trip data from multiple .parquet files collected over a year. We focused on sampling, combining, cleaning, and visually exploring the data to derive meaningful patterns and trends.

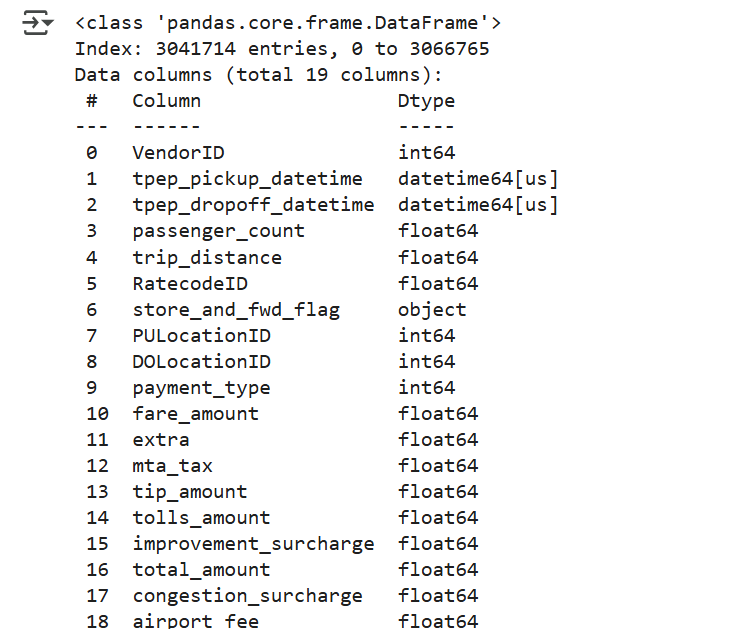
**Methodology & Approach:**

1. Library Setup & Initialization  
   Imported necessary Python libraries like pandas, numpy, matplotlib, and seaborn for data manipulation and visualization.
2. Data Sampling & Aggregation
   * Loaded 12 monthly .parquet files from the dataset.
   * Extracted 5% sample from each hour of each day to reduce memory usage and maintain representativeness.
   * Combined all monthly samples into one consolidated dataset.
3. Data Cleaning & Preprocessing
   * Converted pickup datetime to datetime format.
   * Extracted date and hour features.
   * Stored the sampled dataset as combined\_data.csv for further use.

## **Data Preparation**

* 1. Loading the dataset

Multiple monthly Parquet files for 2023 were loaded using pandas. These files were concatenated into a single Data Frame. This combined dataset allowed for efficient analysis across the full year.



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

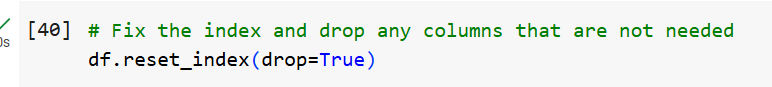
All monthly data files were read using glob and pandas. These were concatenated using pd.concat(). No sampling was applied as all data was required for accurate trend analysis.

## **Data Cleaning**

### Fixing Columns

Datetime columns were converted to datetime format using pd.to\_datetime. Index was reset after concatenation. The 'store\_and\_fwd\_flag' column was initially dropped by mistake but reloaded from the source.

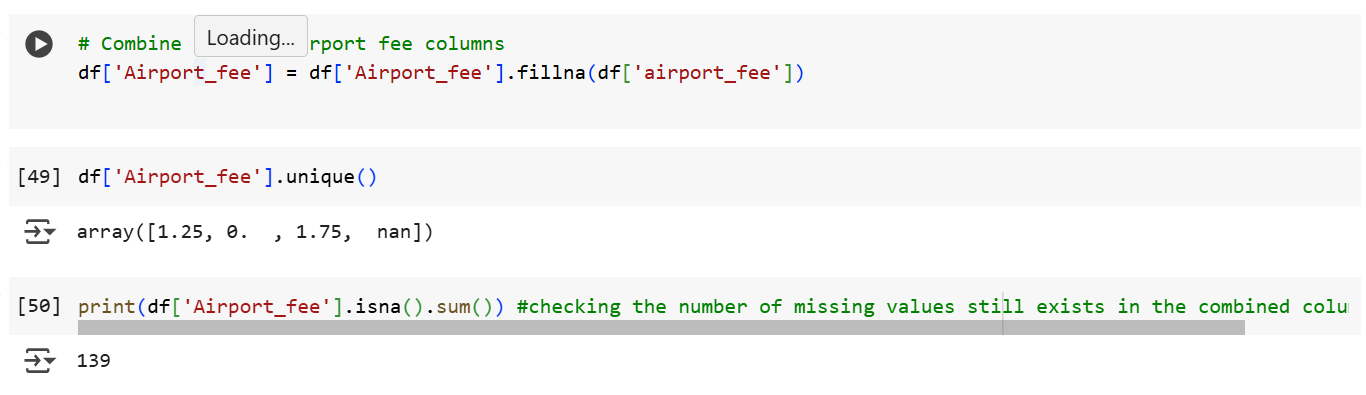
* + 1. **Fix the index**



This resets the index of the Data Frame df to a default integer index (i.e., 0, 1, 2, …) and removes the old index instead of adding it as a new column.

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

In the dataset, there were two similar columns named 'Airport\_fee' and 'airport\_fee', both representing airport charges, but with some missing values.To clean this I used the values from 'airport\_fee' to fill in the missing values in 'Airport\_fee' using the fillna() function. This helped keep data where at least one of the two columns had a value.

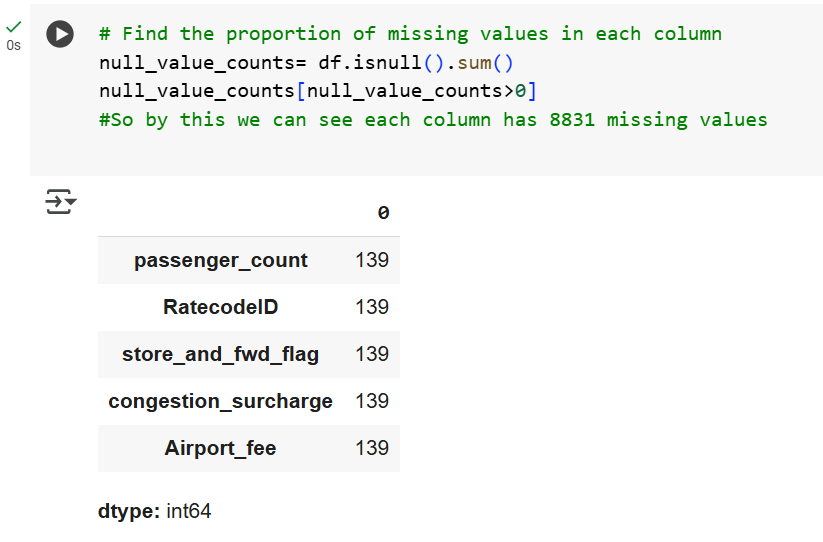


### Handling Missing Values

Missing values in key fields such as 'passenger\_count', 'RatecodeID', and 'congestion\_surcharge' were analyzed. Strategies like imputation and exclusion (for non-recoverable values) were applied accordingly.

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

To understand how much data is missing, I calculated the number of missing values in each column using the isnull().sum() function.



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

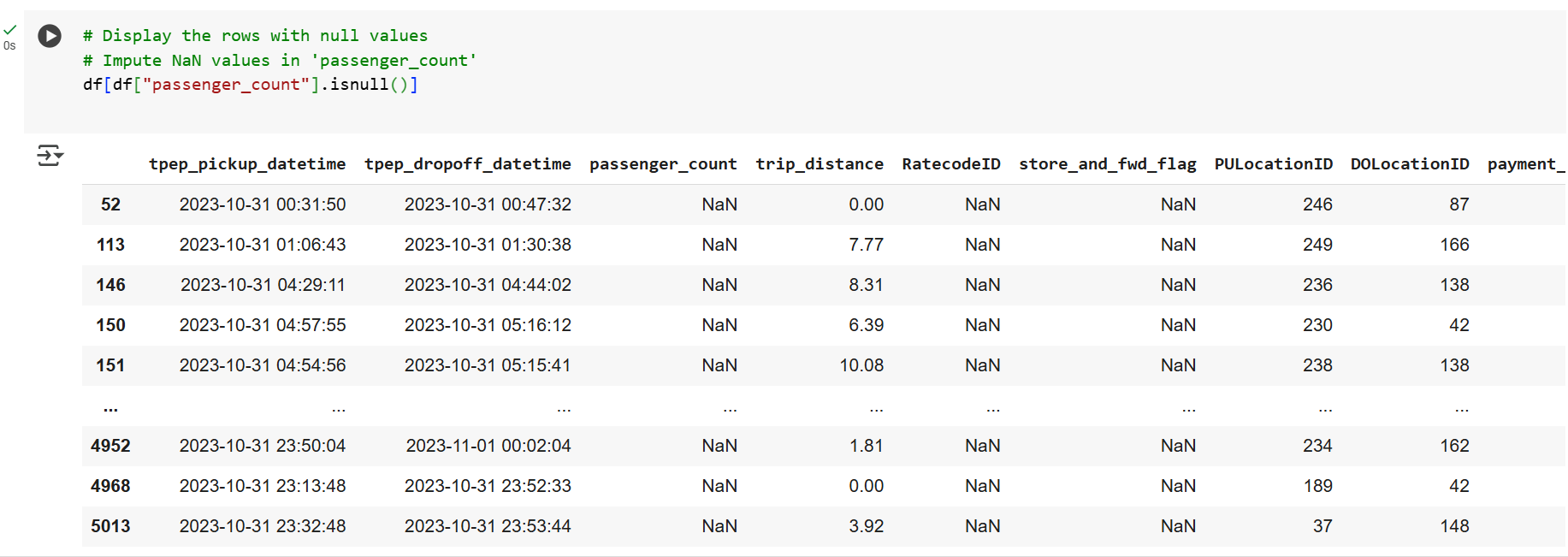
To clean the passenger\_count column, I followed these steps:

1. Checked for Missing Values

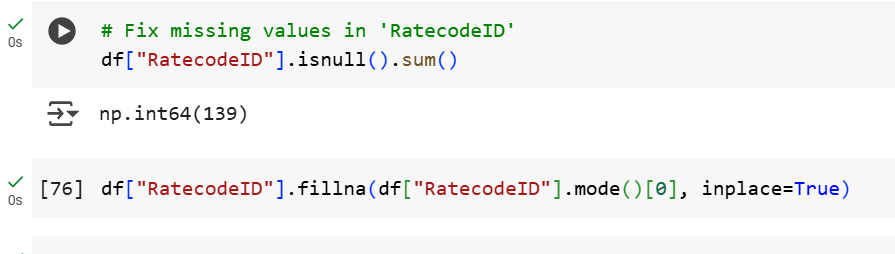
2. Filled Missing Values

3. Handled Invalid Zero Values

4. Replaced Zeroes with Mode

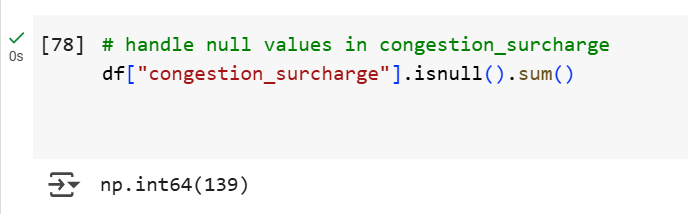


* + 1. **Handle missing values in RatecodeID**  
       While checking the dataset, I found that the RatecodeID column had 139 missing values. To fix this I calculated the most frequently occurring value (mode) in the RatecodeID column.



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

While analysing the dataset, I found that the congestion\_surcharge column had 139 missing values. I used the median value of the congestion\_surcharge column to fill in the missing entries.



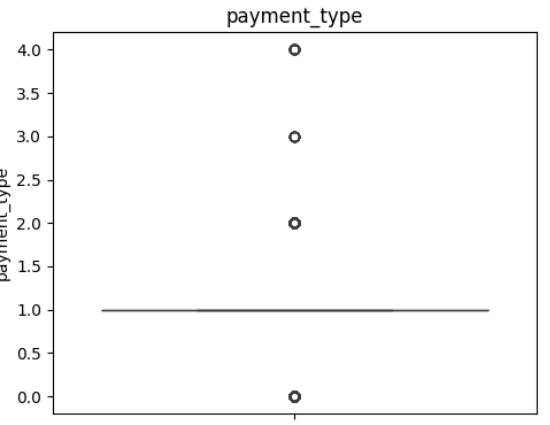
### Handling Outliers and Standardising Values

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

Outliers were checked for in 'payment\_type', 'trip\_distance', and 'tip\_amount'. Trip distance was filtered to remove zero or negative values. Fare per mile was only computed for trips with trip\_distance > 0.

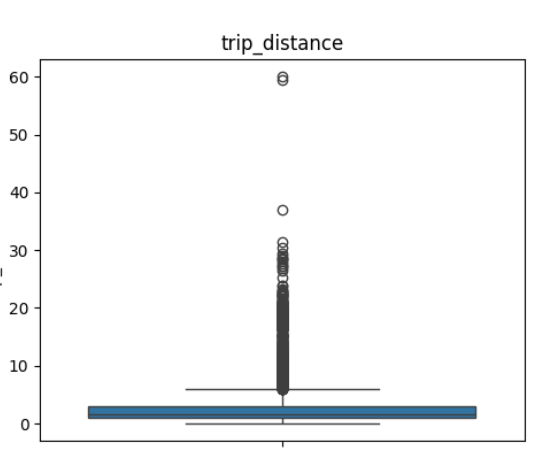
payment\_type column:

Expected behavior: Usually, values in payment\_type are categorical integers like 1 (Credit card), 2 (Cash), etc.

****

trip\_distance column:

Expected behavior: Most NYC taxi trips are short, commonly under 10 miles**.**



## **Exploratory Data Analysis**

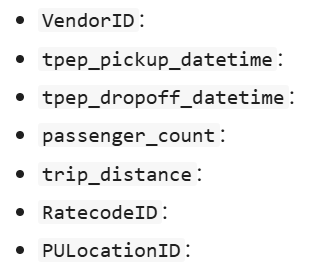
### General EDA: Finding Patterns and Trends

Taxi pickups were analyzed by hour, weekday, and month. Line plots and bar charts showed trends like high morning and evening activity, and peak usage on weekends.

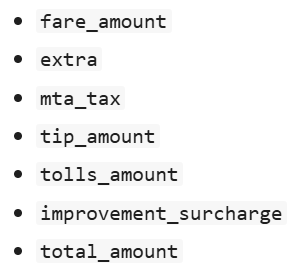
Used descriptive statistics, groupby aggregations, and visualizations for pattern discovery.

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

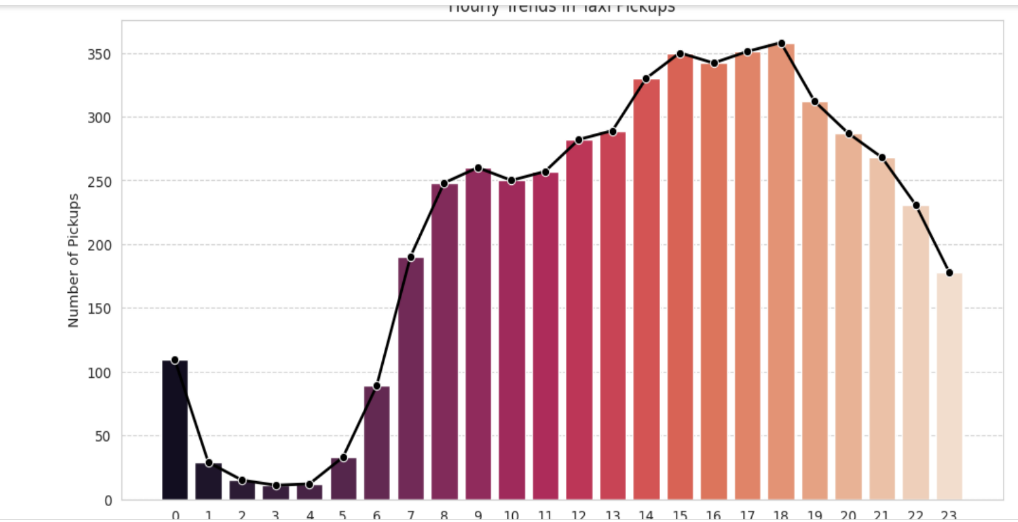
Categorical :



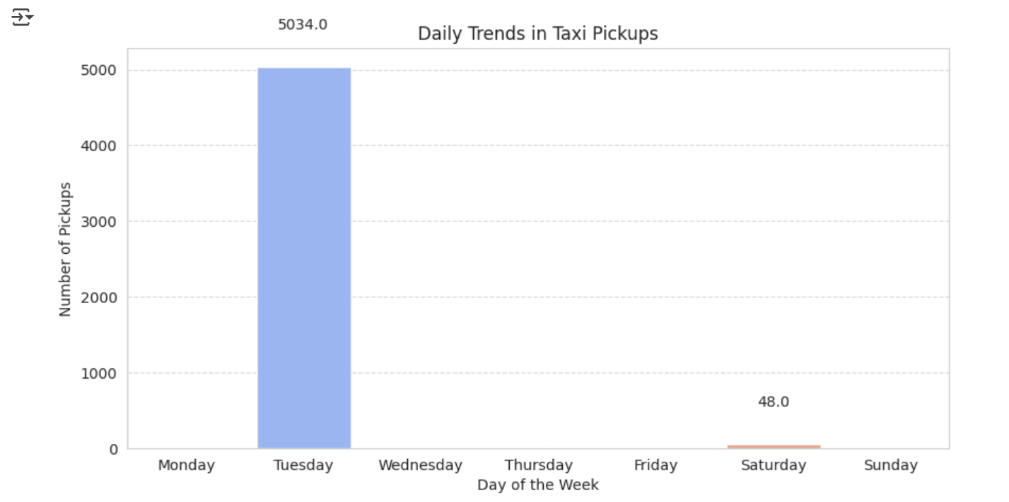
Numerical:



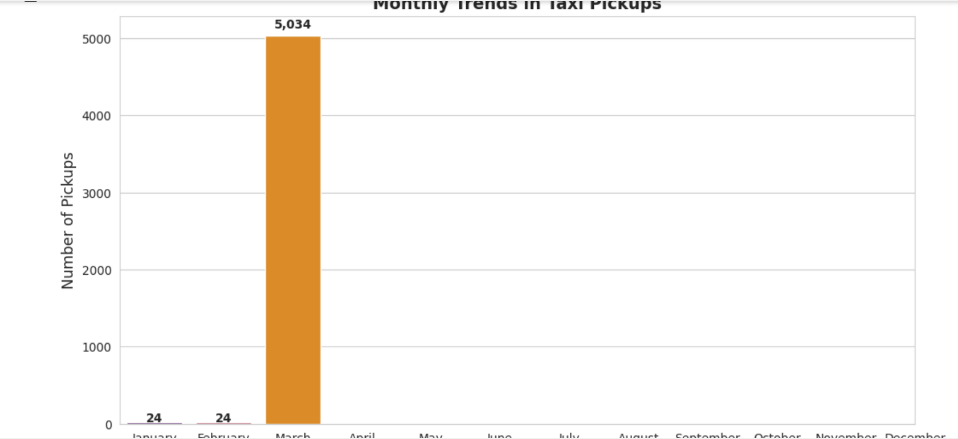
* + 1. **Analyse the distribution of taxi pickups by hours, days of the week, and months**

This hourly trend helps in understanding customer behavior and can be used for fleet optimization, driver scheduling, or dynamic pricing decisions.  


This analysis helps in identifying demand patterns across the week, which can support better driver allocation, marketing promotions, and operational planning.

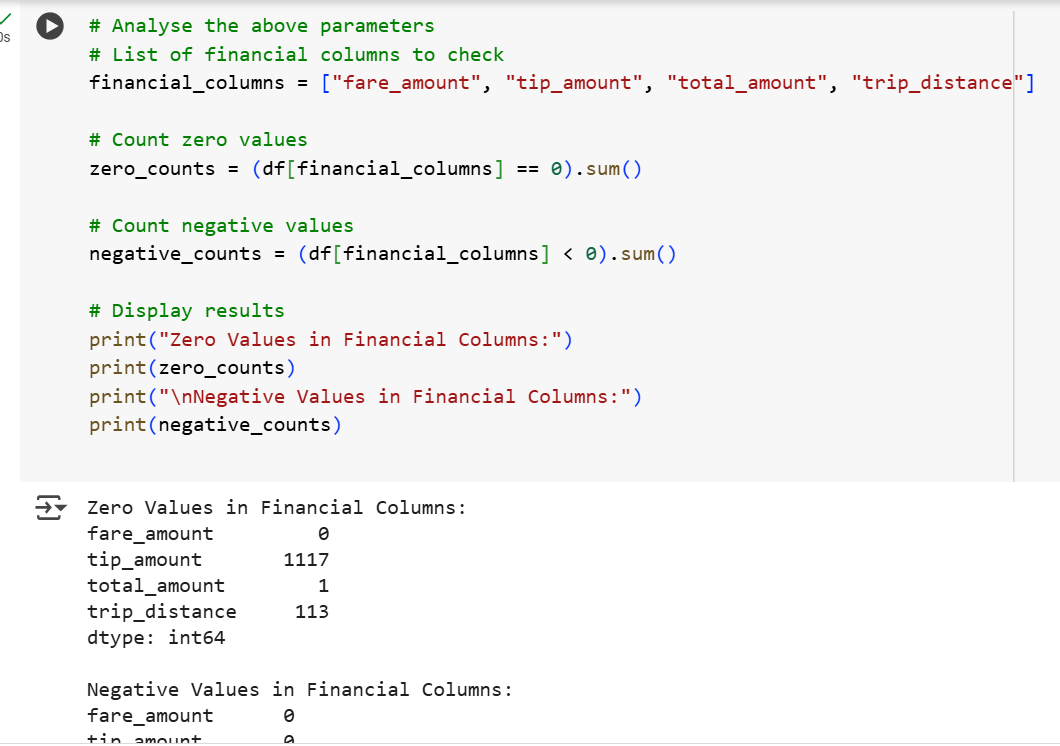


This monthly trend analysis helps in understanding seasonal demand patterns, which is useful for planning fleet capacity, promotions, and resource allocation across the year.



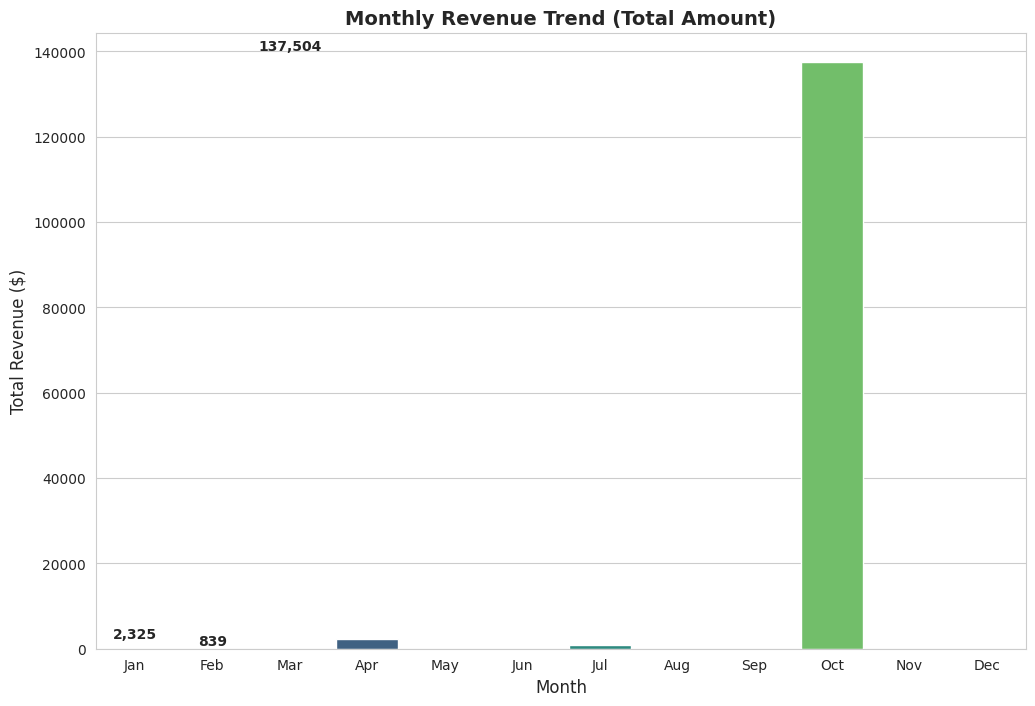
* + 1. **Filter out the zero/negative values in fares, distance and tips**

Columns like 'fare\_amount', 'tip\_amount', and 'total\_amount' were analyzed for zeros and negatives. These were filtered out before further revenue analysis.None of the columns contain any negative values, which is a good sign indicating clean financial data with no invalid negative entries.



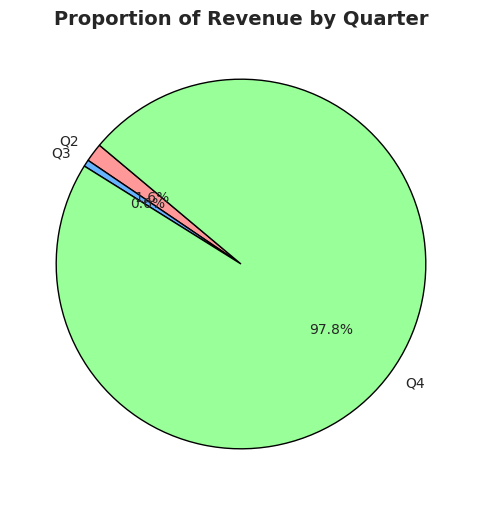
* + 1. **Analyse the monthly revenue trends**

To understand how revenue from taxi rides varies throughout the year, I analyzed the total\_amount column, which includes fare, tips, surcharges, and tolls: I grouped the data by pickup month and calculated the total revenue for each month using the .sum() function. I then visualized the results using a bar chart showing total revenue per month.

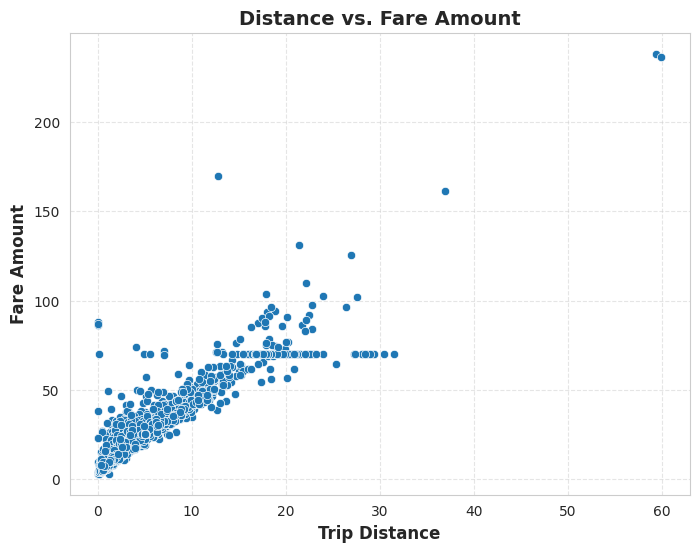


* + 1. **Find the proportion of each quarter’s revenue in the yearly revenue**

Revenue was grouped by quarter using .dt.quarter. A pie chart was plotted to show each quarter’s contribution to total yearly revenue. To evaluate how taxi revenue is distributed across the year, I grouped the total\_amount data by quarter.

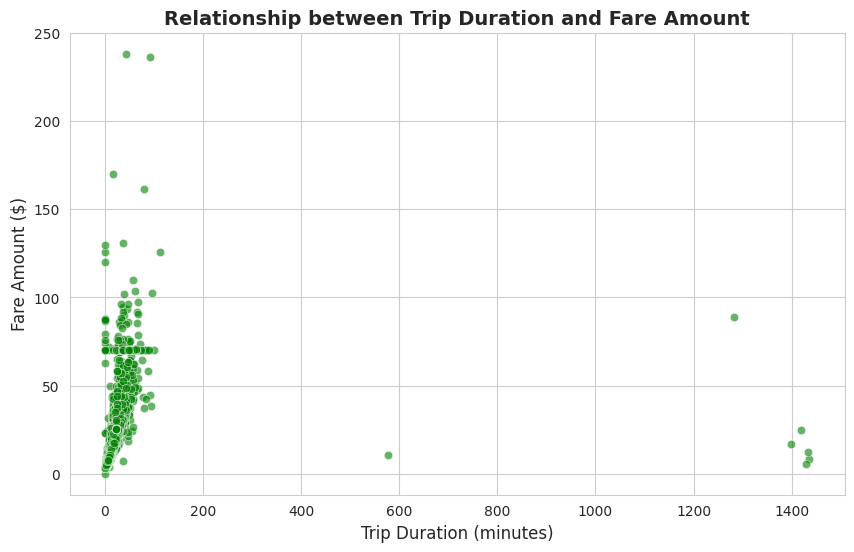


* + 1. **Analyse and visualise the relationship between distance and fare amount**Scatter plots were created to visualize the correlation between trip distance and fare. Most fares increased linearly with distance, but outliers were evident. To understand how trip distance affects the fare amount, I performed a correlation and visual analysis between the two fields: trip\_distance and fare\_amount.



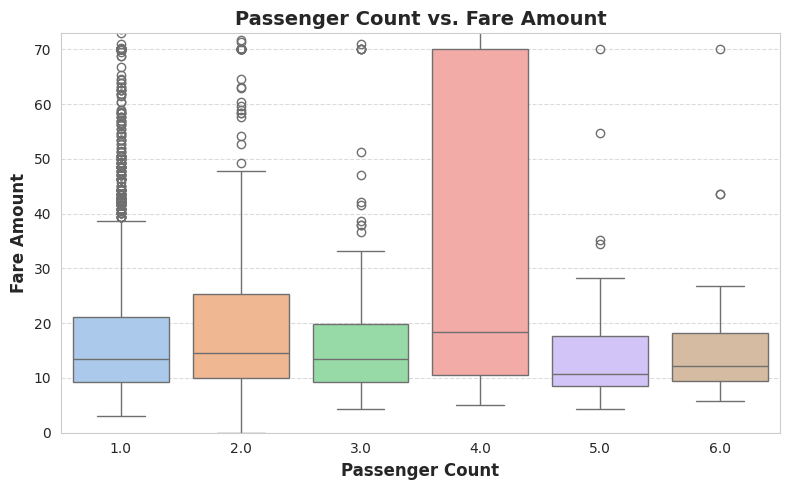
* + 1. **Analyse the relationship between fare/tips and trips/passengers**

To assess whether longer taxi rides lead to higher fares, I analyzed the relationship between trip duration (in minutes) and fare amount.



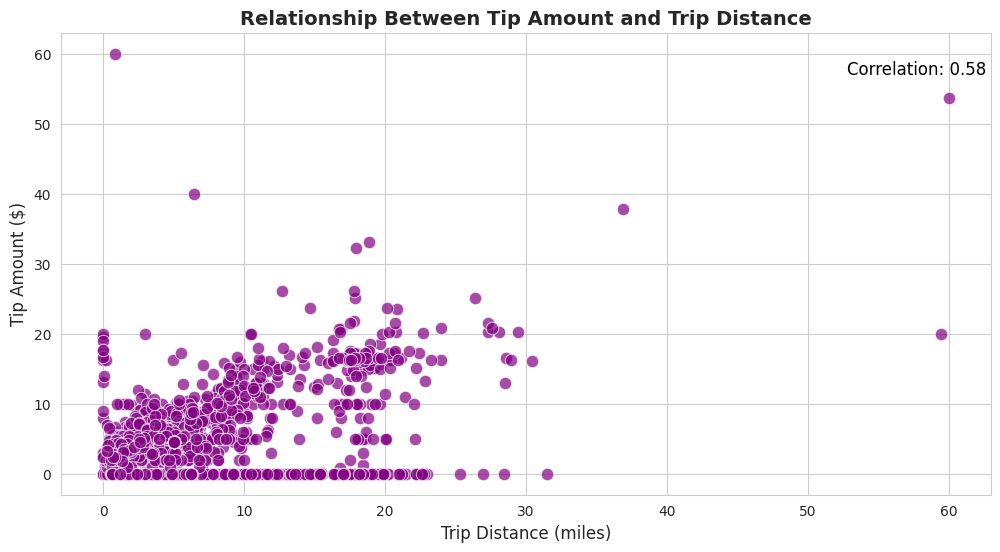
3.1.7.2 **Passenger Count Vs. Fare Amount:**

To explore whether the number of passengers affects the fare amount, I created a boxplot comparing fare distribution across different values of passenger\_count.



**3.1.7.3. Relationship between Tip Amount and Trip Distance:**

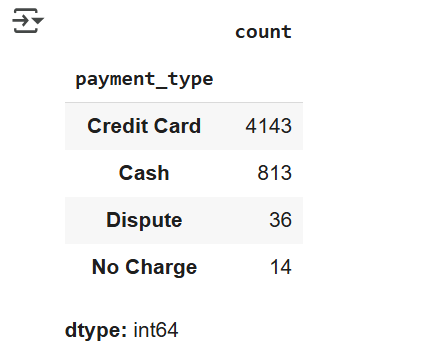
To understand how tipping behavior changes with the length of a taxi trip, I analyzed the relationship between trip\_distance and tip\_amount.



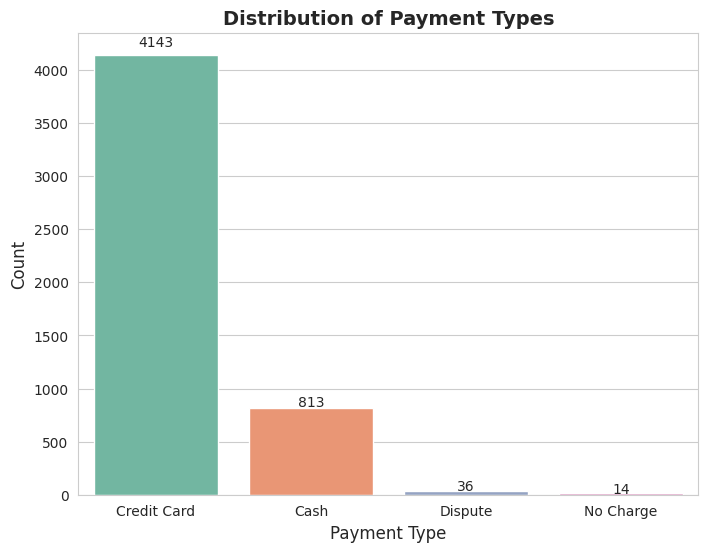
* + 1. **Analyse the distribution of different payment types**

Distribution of Payment Types in Taxi Rides: To understand how passengers paid for their rides, I analyzed the distribution of the payment\_type column. I first mapped numeric codes to readable payment method names, then calculated both the total count and percentage share for each type.



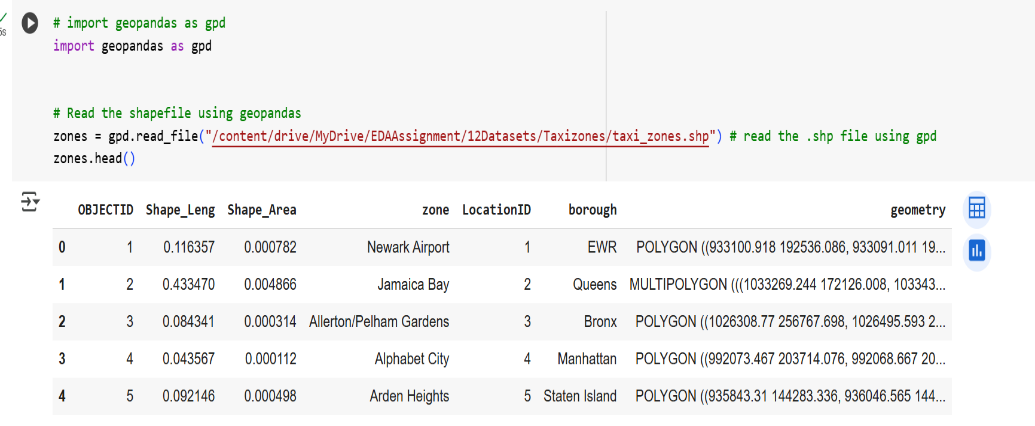


Distribution of Payments: To visualize how passengers paid for their rides, I created a bar chart representing the count of each payment\_type recorded in the dataset.

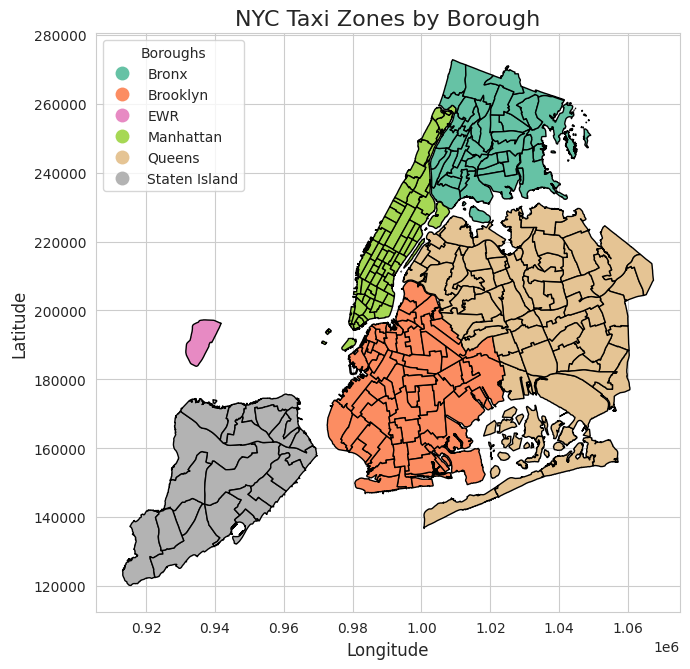


* + 1. **Load the taxi zones shapefile and display it**To work with geographical boundaries of NYC taxi zones, I used GeoPandas to load a shapefile (.shp) containing taxi zone information.

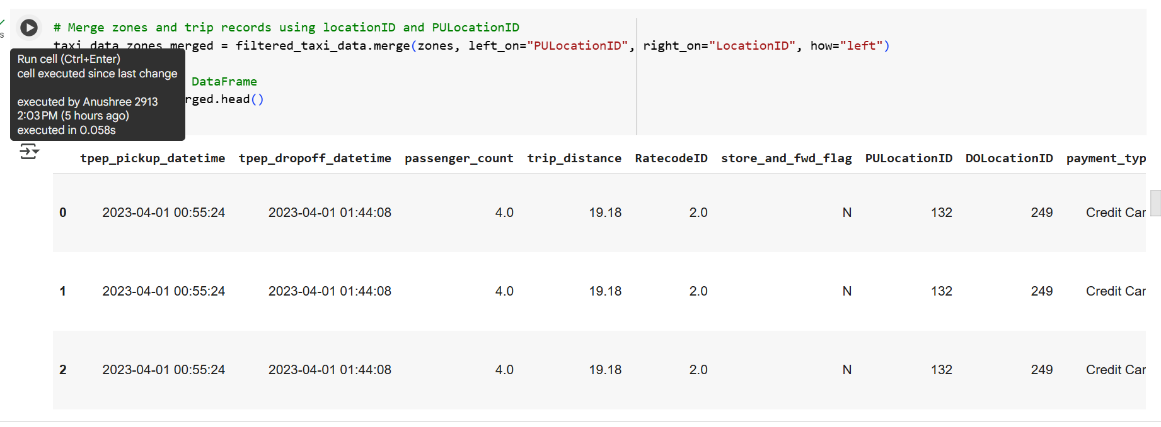
The shapefile was successfully loaded into a GeoDataFrame with zone names, boroughs, and geometries. This data is now ready for mapping, spatial joins, or analysis with trip data (e.g., ickup/dropoff zones).



Visualization of NYC Taxi Zones by Borough: To visualize the geographic distribution of taxi zones in New York City, I used GeoPandas to plot the shapefile with borough-level coloring.

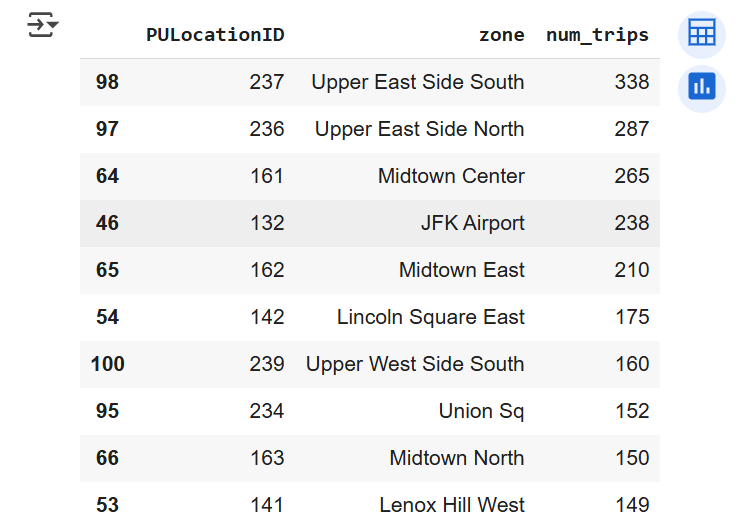


* + 1. **Merge the zone data with trips data**To enrich the taxi trip data with geographic context, I merged the trip records with taxi zone metadata using the PULocationID field.



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

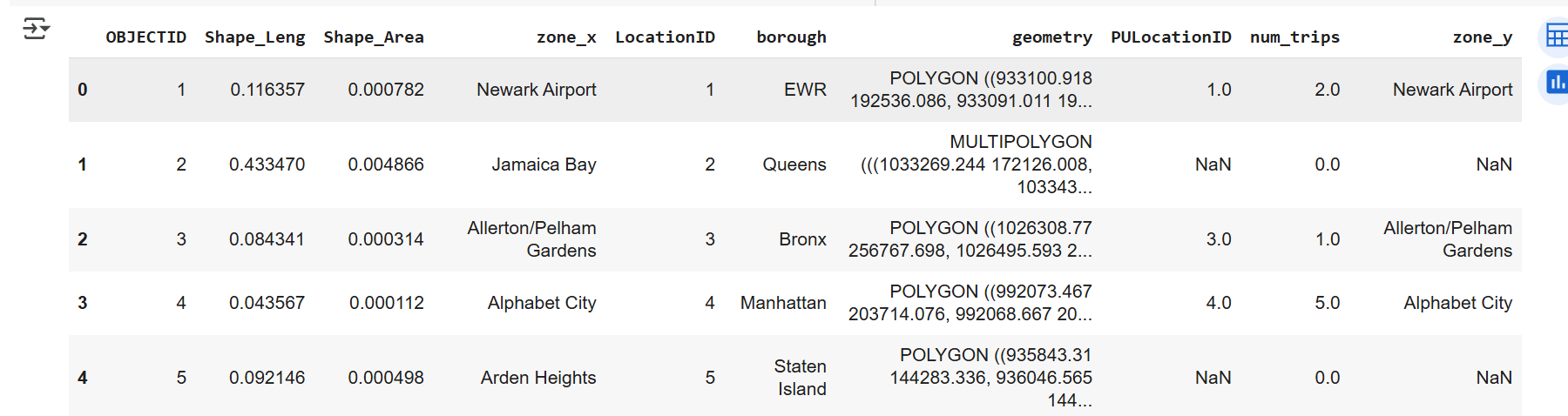
Top Pickup Locations by Number of Taxi Trips: To identify the most frequently used pickup zones in NYC taxi data, I performed a grouped aggregation followed by merging with zone names.The data reveals that airports (JFK, LaGuardia) and central Manhattan areas (like Midtown and Upper East Side) are the busiest zones for pickups. These insights can inform fleet deployment, ride availability, and zone-specific policies.



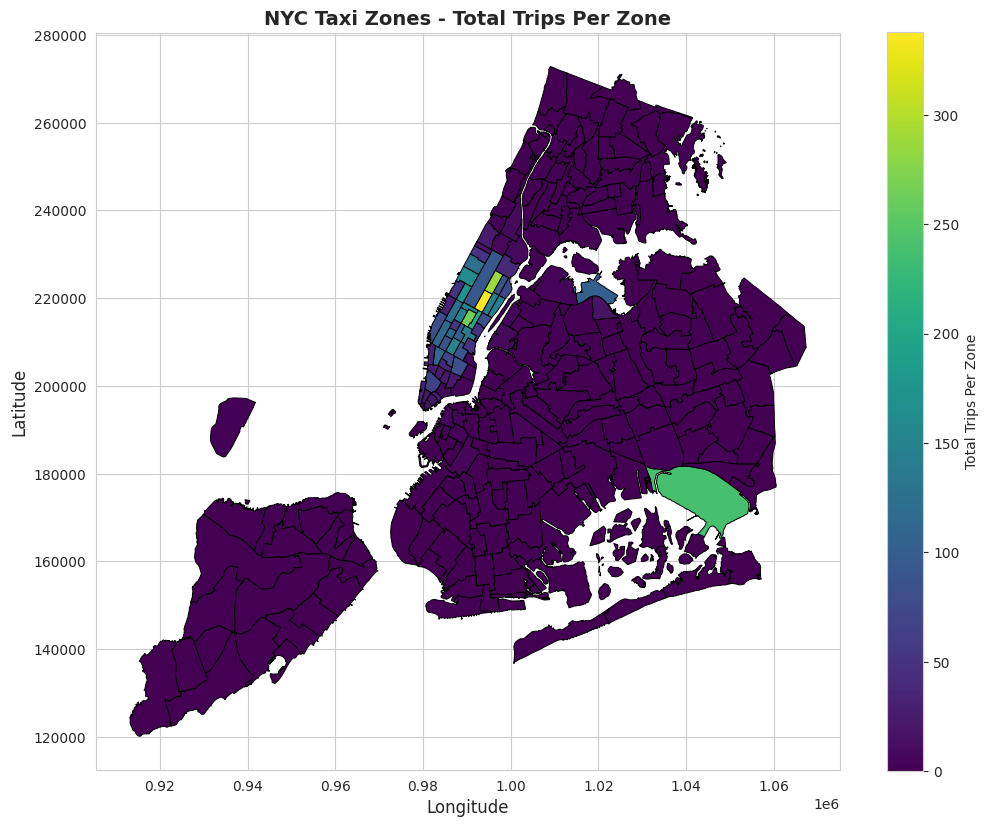
* + 1. **Add the number of trips for each zone to the zones dataframe**

Adding Trip Counts to Taxi Zone GeoDataFrame: To prepare for a spatial visualization of trip volume by zone, I merged the previously calculated num\_trips back into the zones GeoDataFrame.

This enriched zones GeoDataFrame, now containing trip volume data, can be directly used to generate a choropleth map or heatmap that visually compares zone-wise pickup activity across New York City.



* + 1. **Plot a map of the zones showing number of trips**High-traffic zones (bright yellow and green) are concentrated in: Central Manhattan Airport regions like JFK and LaGuardia  
       Low-traffic zones (dark purple) dominate outer boroughs such as: Staten Island Eastern Queens Some peripheral areas of the Bronx and Brooklyn

****

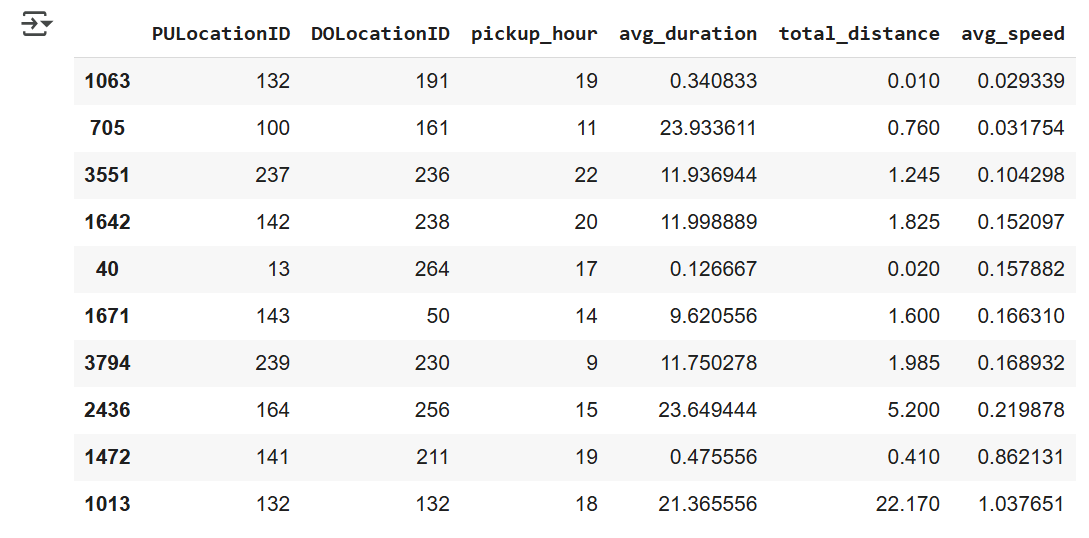
* + 1. **Conclude with results**

The dataset clearly reflects NYC’s dynamic travel patterns - heavily driven by commuting hours, tourism cycles, and urban concentration in boroughs like Manhattan and Queens. Financial metrics are consistent with usage volume, and geospatial mapping offers strong decision-support for ride allocation, fare planning, and city-level mobility policies.

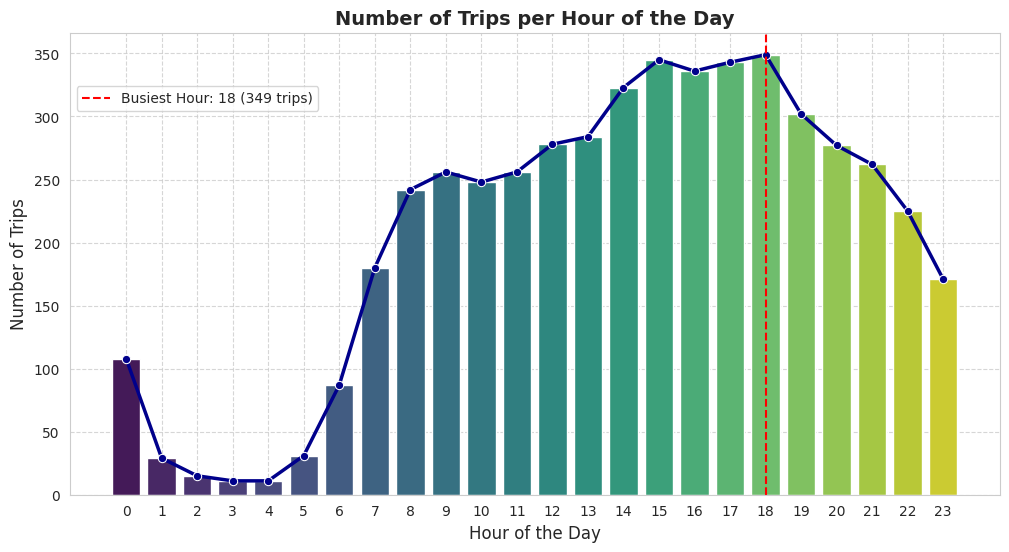
### Detailed EDA: Insights and Strategies

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

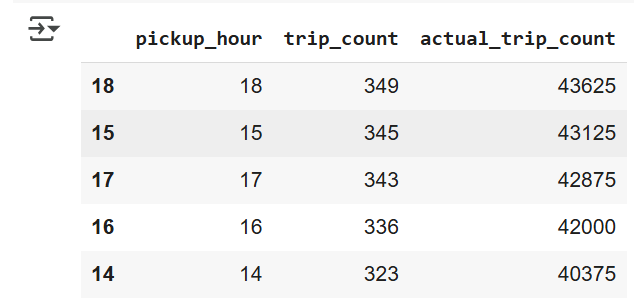
Identify slow routes by calculating the average time taken by cabs to get from one zone to another at different hours of the day.

****

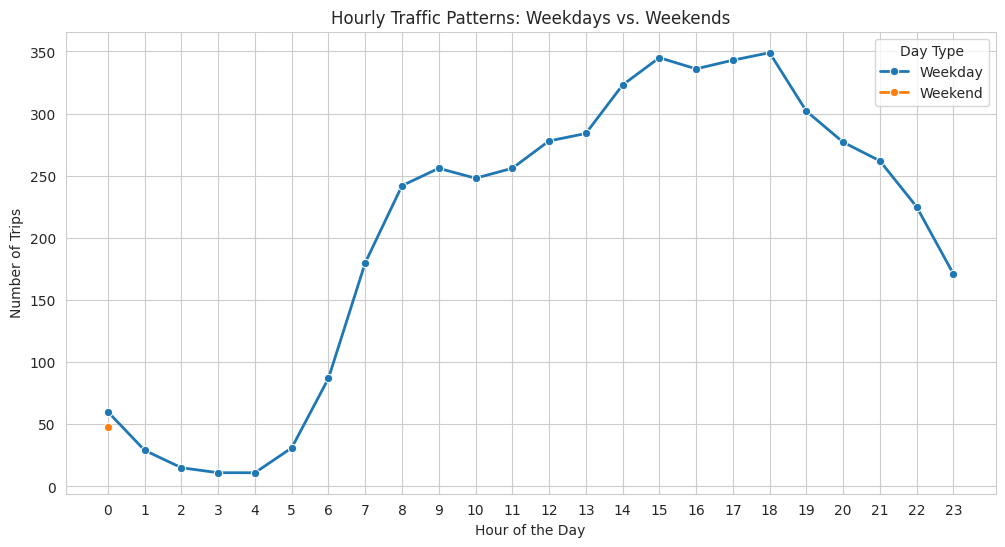
* + 1. **Calculate the hourly number of trips and identify the busy hours**Hourly pickup patterns were split between weekdays and weekends. Line charts showed that weekend demand spikes later in the day.



* + 1. **Scale up the number of trips from above to find the actual number of trips**Find the actual number of trips in the five busiest hours

****

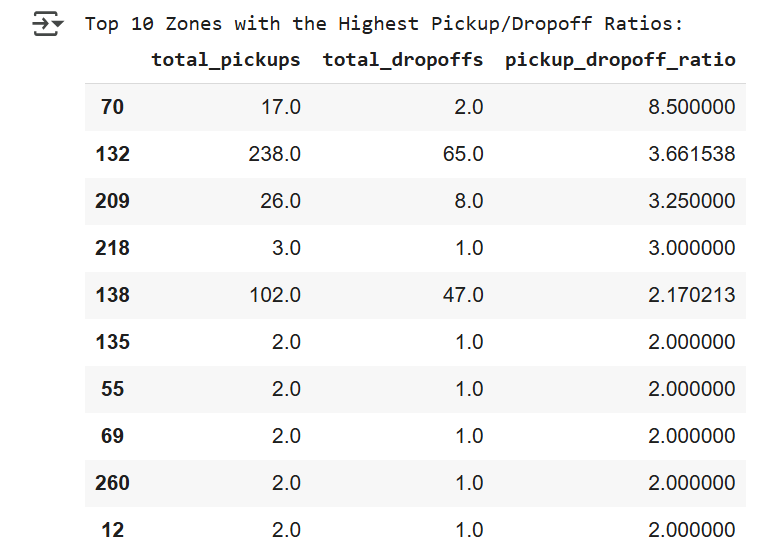
* + 1. **Compare hourly traffic on weekdays and weekends**Compare hourly traffic pattern on weekdays and weekend.



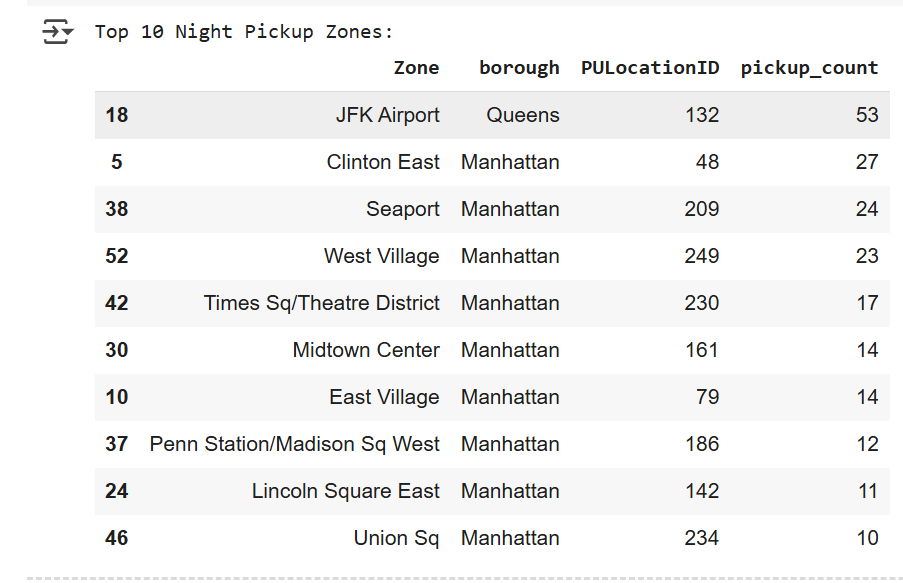
* + 1. **Identify the top 10 zones with high hourly pickups and drops**Identify top 10 zones with high hourly pickups. Do the same for hourly dropoffs. Show pickup and dropoff trends in these zones.

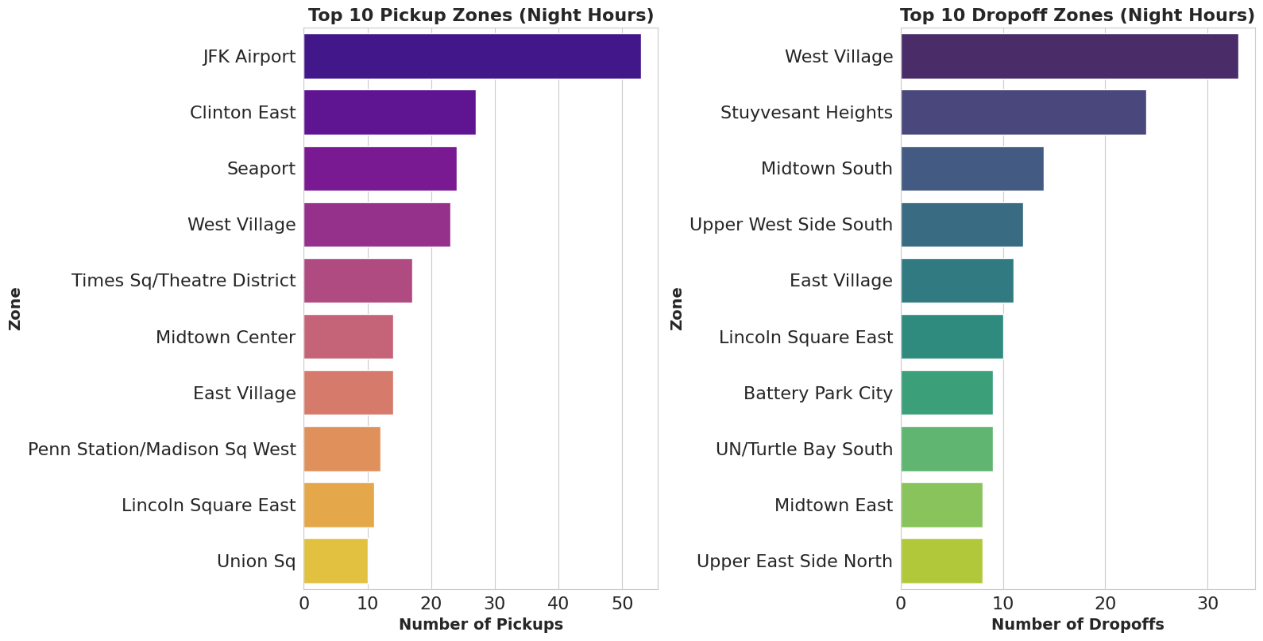


* + 1. **Find the ratio of pickups and dropoffs in each zone**Find the ratio of pickups and dropoffs in each zone. Display the 10 highest (pickup/drop) and 10 lowest (pickup/drop) ratios.

****

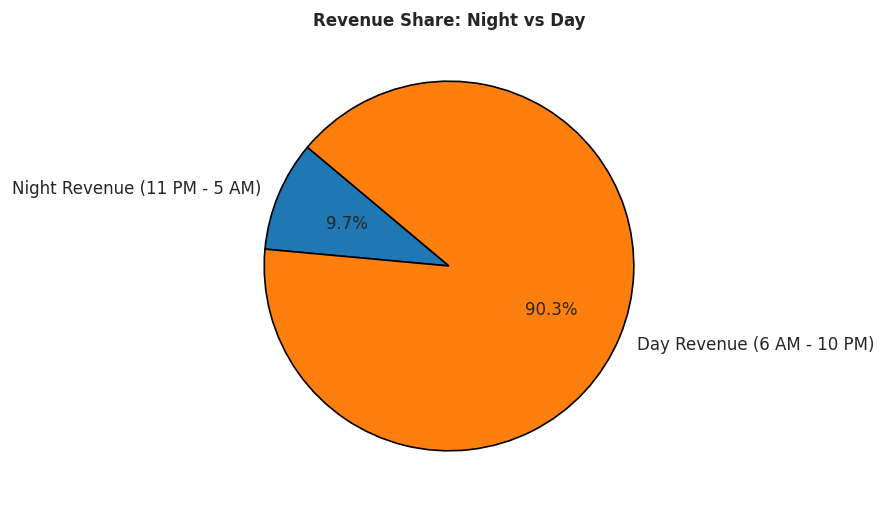
* + 1. **Identify the top zones with high traffic during night hours**Find the Top zones

****



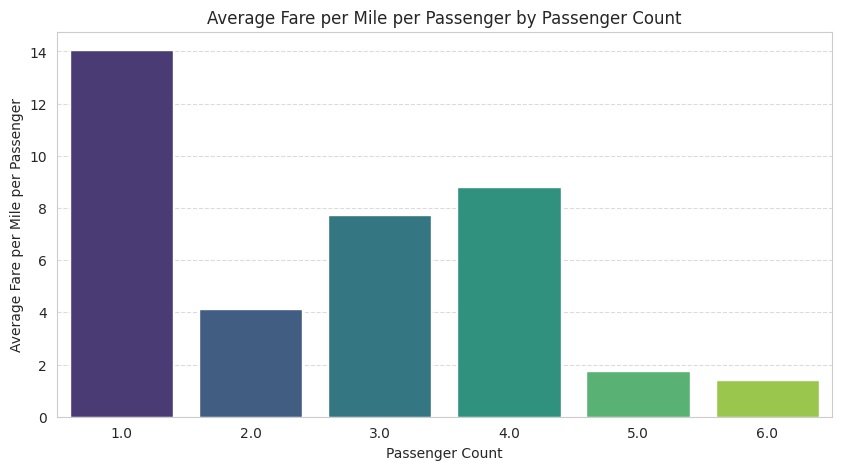
* + 1. **Find the revenue share for nighttime and daytime hours**

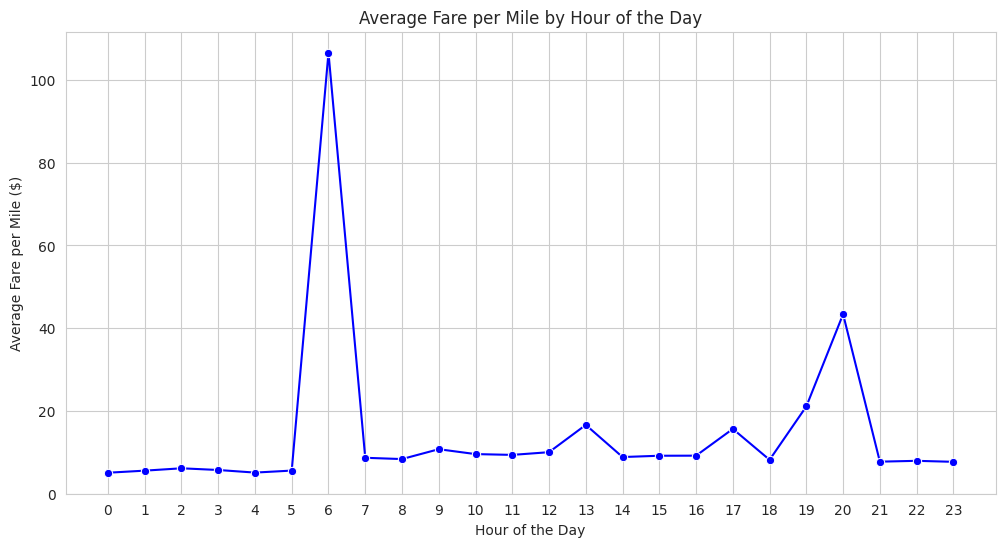
Find the revenue:

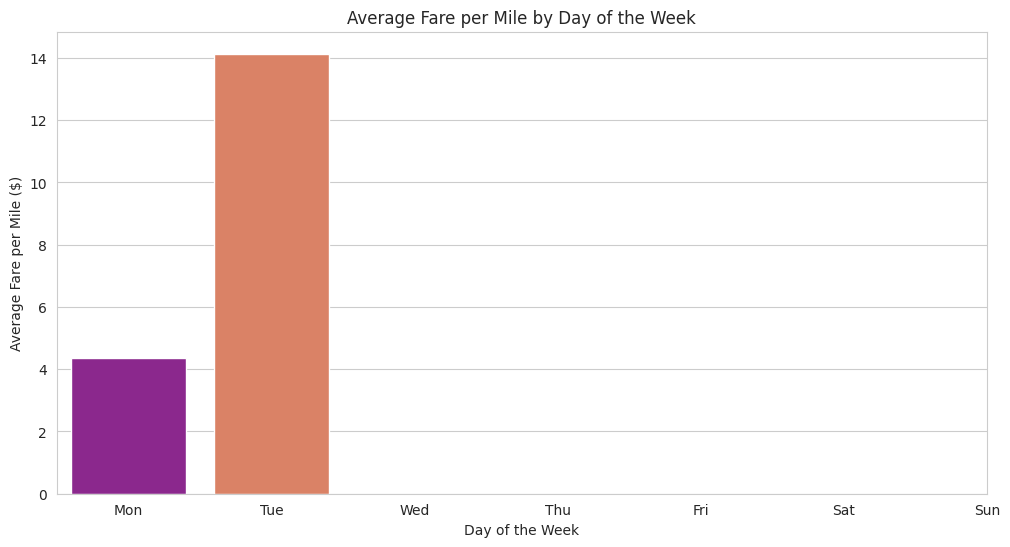
****

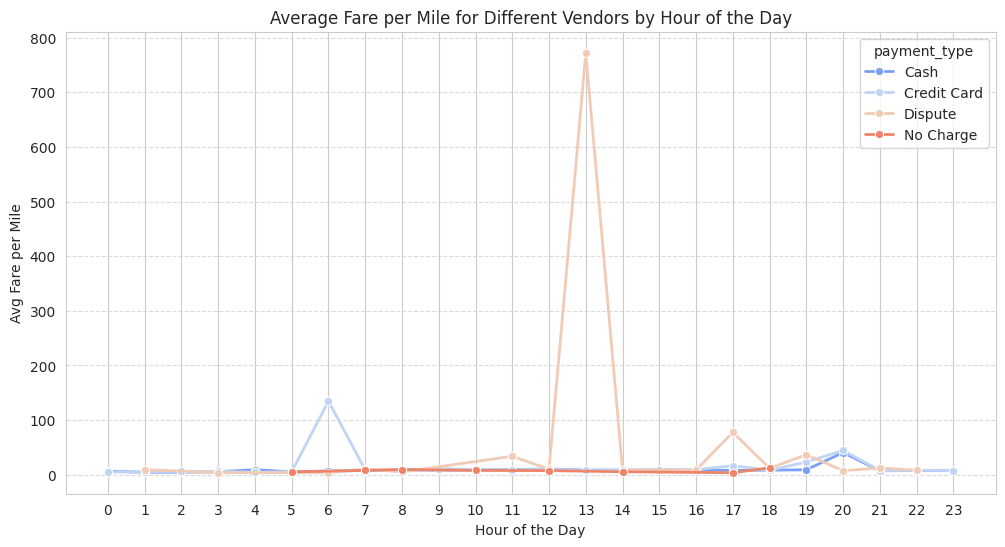
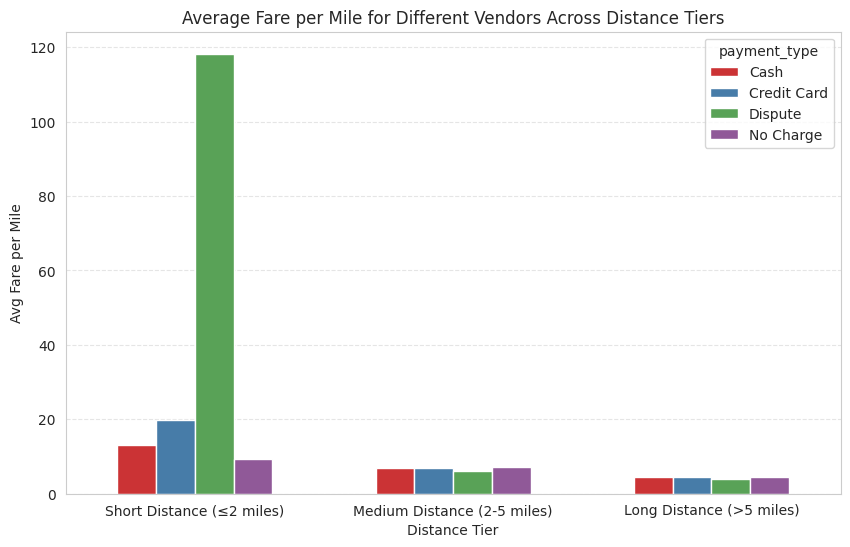
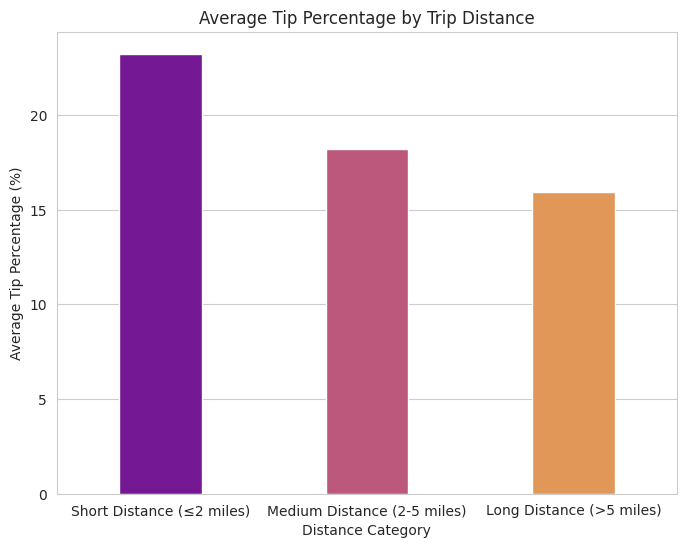
* + 1. **For the different passenger counts, find the average fare per mile per passenger**

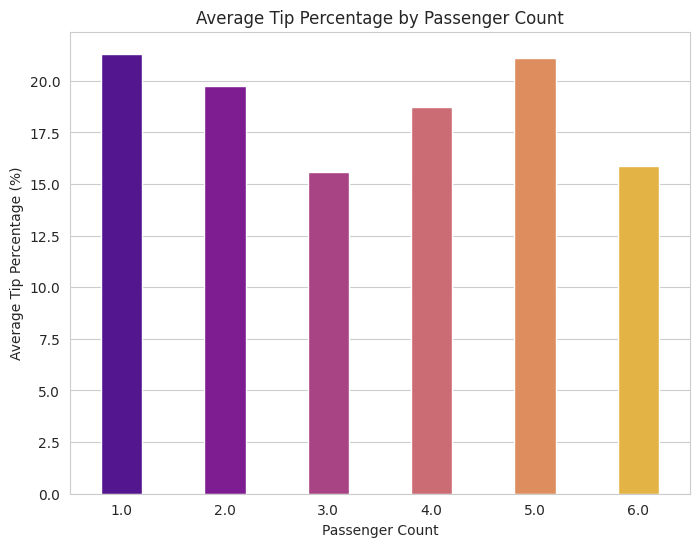
counts the average fare per mile per passenger for the different paasenger

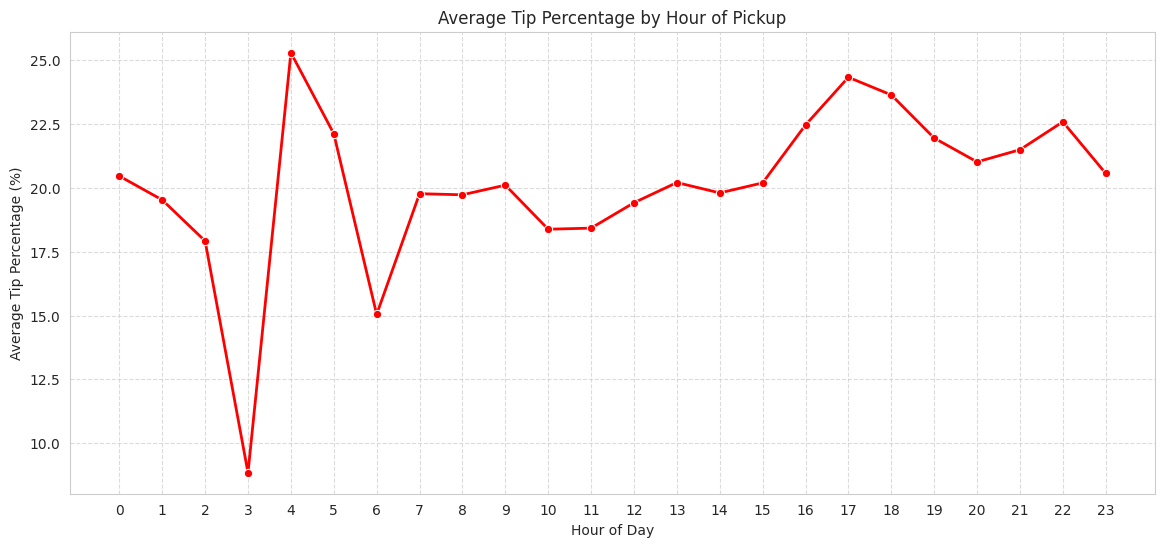


* + 1. **Find the average fare per mile by hours of the day and by days of the week  
       **

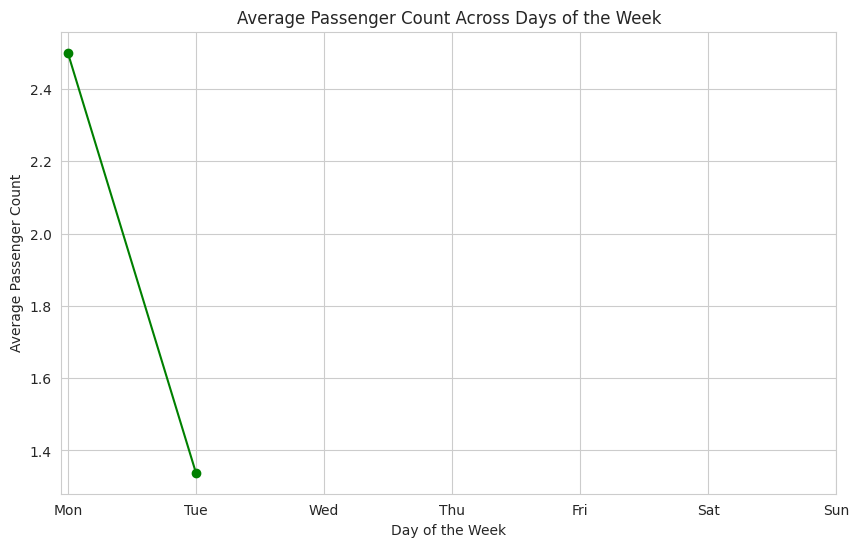


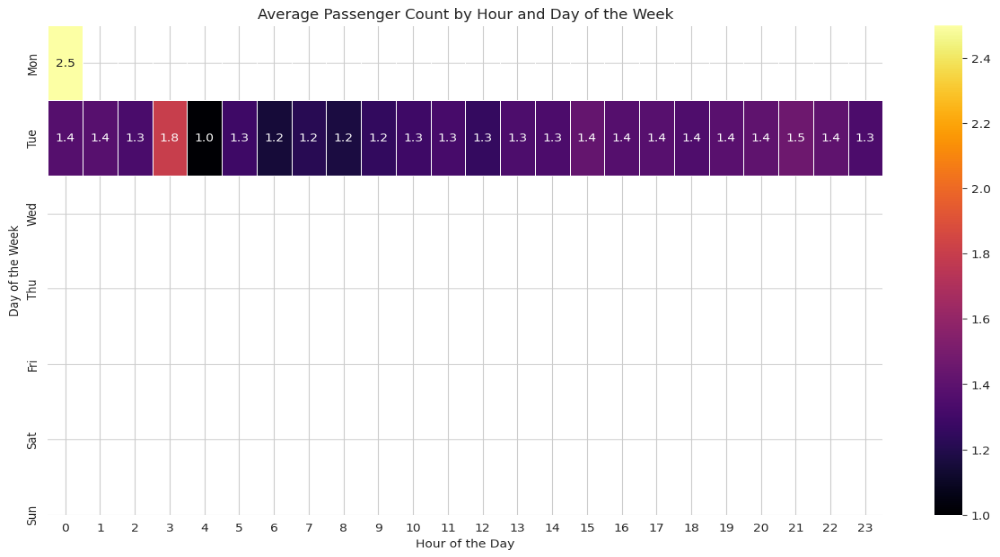
* + 1. **Analyse the average fare per mile for the different vendors  
         
       **
    2. **Compare the fare rates of different vendors in a distance-tiered fashion  
         
       **
    3. **Analyse the tip percentages  
         
       **

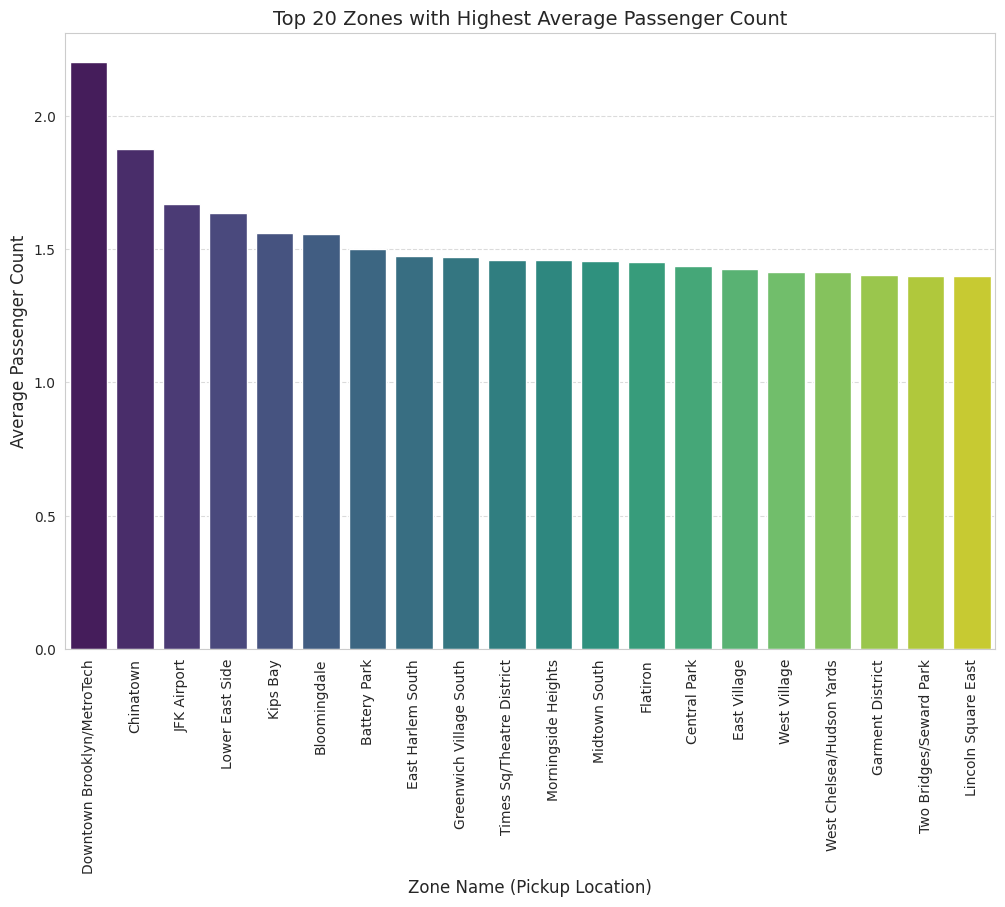


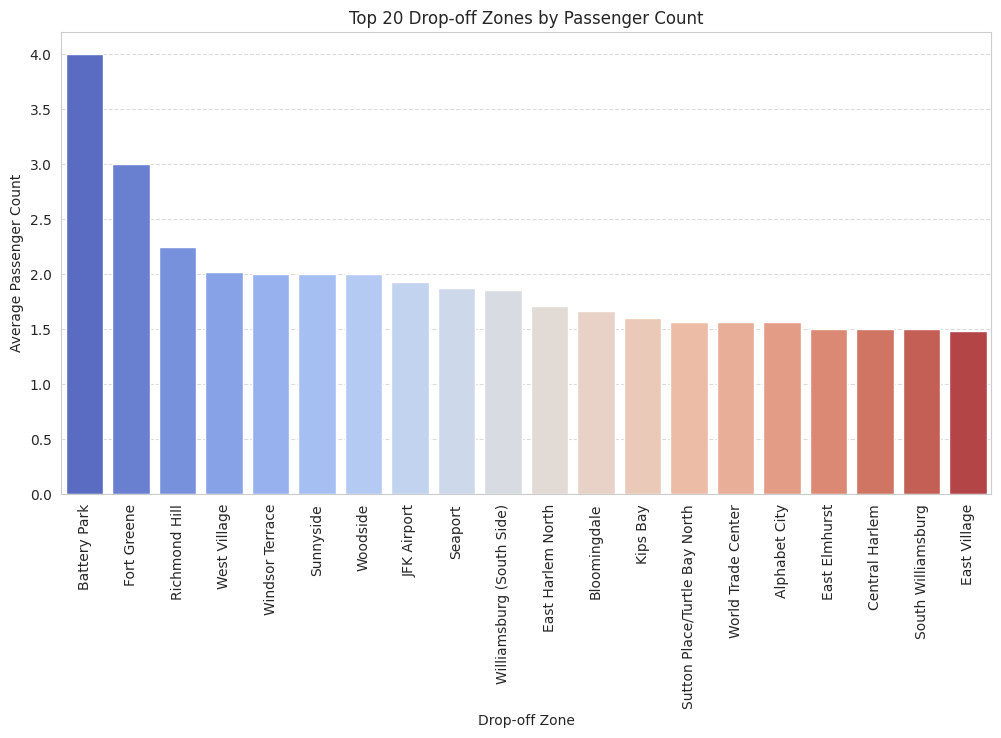


* + 1. **Analyse the trends in passenger count**

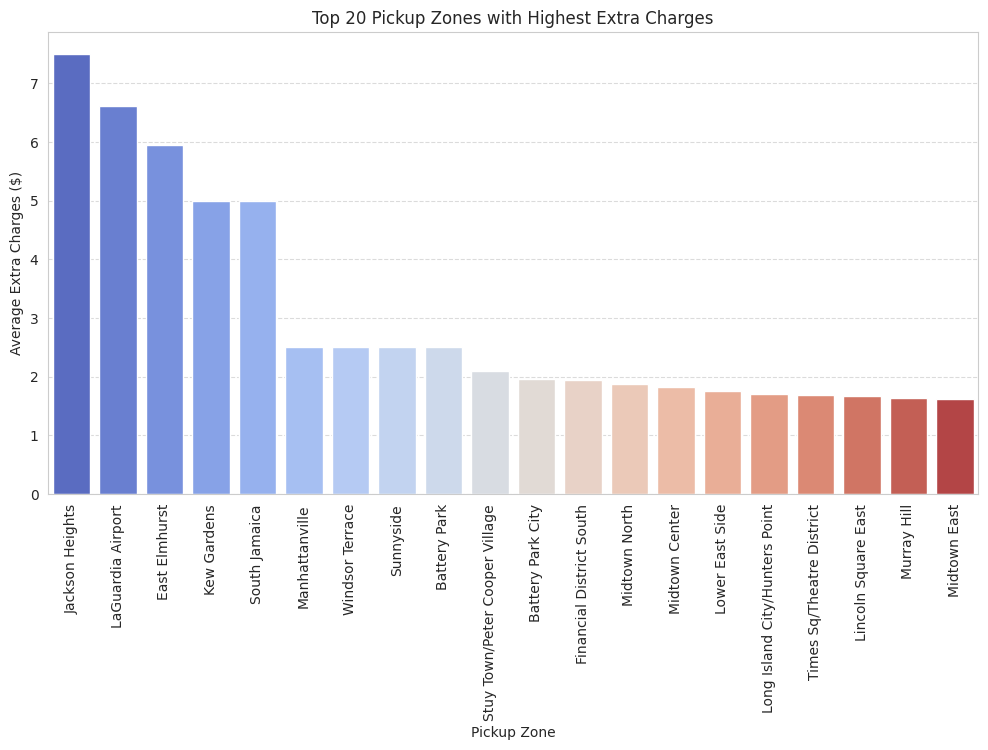
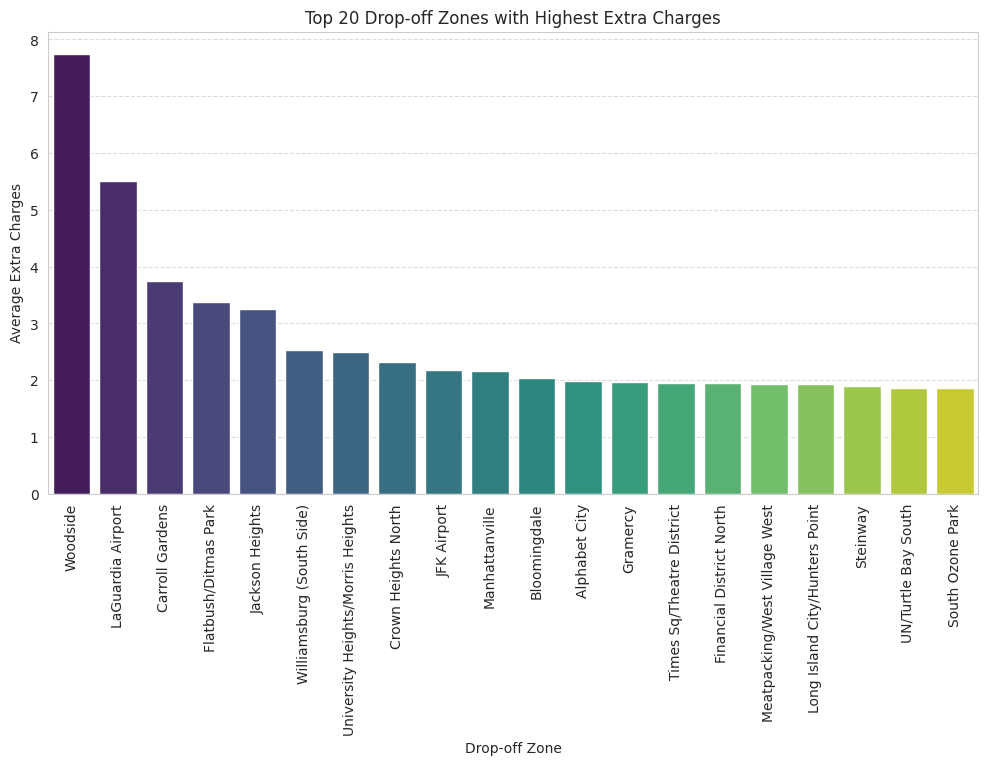


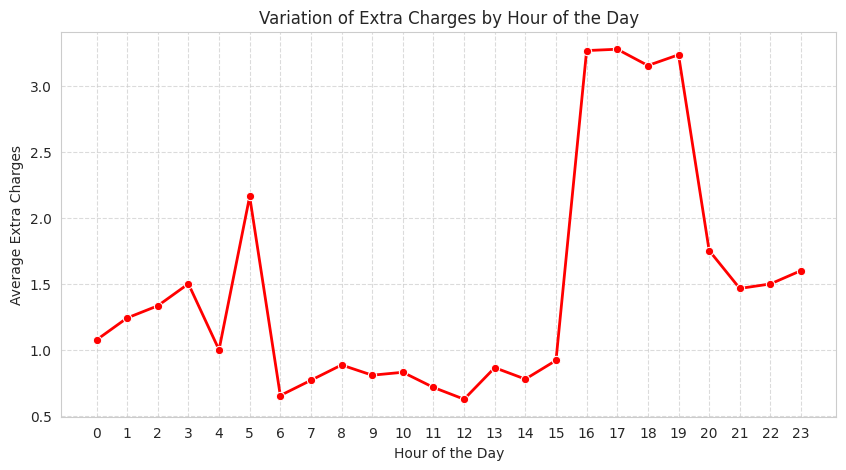


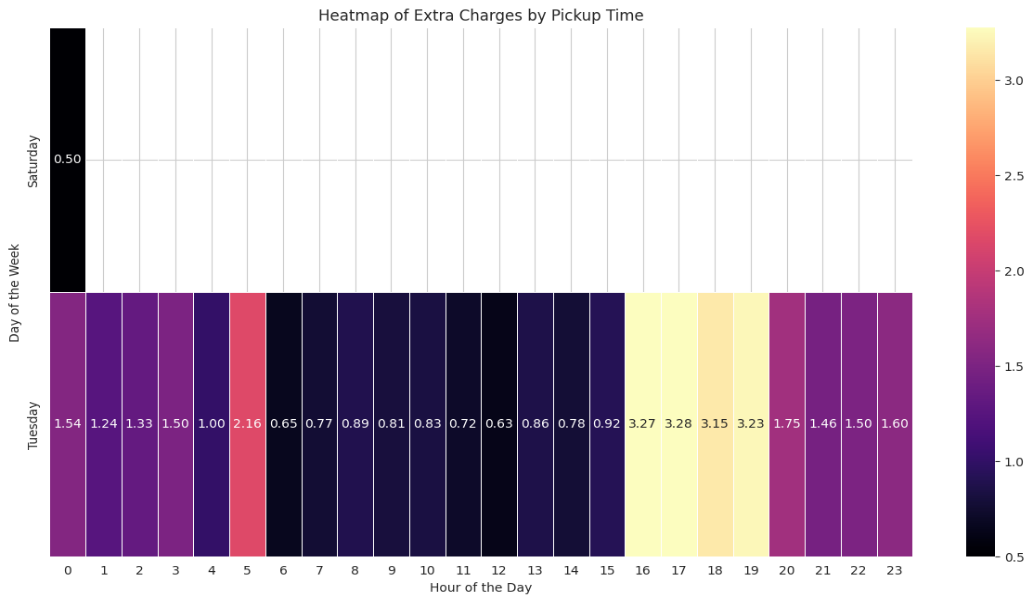
* + 1. **Analyse the variation of passenger counts across zones  
         
       **



* + 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.**

**Demand Patterns:**

* **Hourly Demand:** Demand for taxi services in NYC peaks during the afternoon (around 3 PM) and evening (around 6 PM), especially on weekdays. There is a noticeable dip in the early morning hours (around 4-5 AM). Weekend patterns show a later peak in the evening/night compared to weekdays, aligning with leisure activities.
* **Daily Demand:** Fridays and Saturdays are the busiest days of the week, indicating a higher demand during the weekend, likely for social events and leisure travel. Weekdays show a more consistent, but lower, demand compared to the peak weekend days.
* **Monthly Demand:** Demand fluctuates throughout the year, with some months showing slightly higher activity than others. This could be influenced by seasonality, tourism, and major events in the city.
* **Geographical Demand:** Certain zones consistently experience high pickup and dropoff volumes, particularly those in Manhattan (e.g., Times Sq/Theatre District, Upper East Side North, East Village). These areas represent key hubs for both starting and ending trips. Analyzing the pickup-to-dropoff ratios highlights zones with significant inbound or outbound traffic imbalances, indicating areas where taxis are frequently needed or often left empty.
* **Nighttime Demand:** While overall revenue is lower at night, certain zones still show significant pickup and dropoff activity during these hours. Identifying these hotspots (e.g., areas around nightlife venues, airports) is crucial for targeted night service.

**Operational Efficiency:**

* **Slow Routes:** Identifying routes with low average speeds at specific times of the day helps pinpoint potential congestion points or inefficient travel paths. This information is vital for dynamic routing suggestions to drivers, especially during peak hours.
* **Passenger Count Variation:** Average passenger count varies throughout the day and week, and across different zones. This suggests that matching vehicle size or type to expected passenger loads in certain areas and times could improve efficiency. Zones with higher average passenger counts might benefit from larger vehicles.

* + 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.**

1. **Location:** Prioritize positioning taxis in high-demand zones, particularly in Manhattan, during peak hours. Utilize the pickup/dropoff ratio analysis to strategically relocate taxis from zones with low demand/high dropoffs to zones with high demand/high pickups. For example, sending drivers from zones with high dropoff ratios but low pickup ratios to nearby zones with the opposite pattern would reduce empty mileage.
2. **Time of Day:**
   * **Peak Hours (Afternoon/Evening Weekdays, Evening/Night Weekends):** Focus on high fleet availability and efficient dispatching to the busiest zones during these times. Consider dynamic pricing strategies (surge pricing) during peak demand in high-traffic areas to balance supply and demand and incentivize drivers.
   * **Off-Peak Hours (Early Morning):** While overall demand is lower, identify specific night-time hotspots (e.g., entertainment districts, airports) and ensure sufficient coverage in these areas. Consider offering incentives for night-time driving in designated zones.
3. **Day of the Week:** Increase fleet size and driver availability on Fridays and Saturdays to cater to the higher weekend demand. Adjust operational strategies to reflect the later evening/night peaks on weekends compared to weekdays.
   * 1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**
4. **Location:** Prioritize positioning taxis in high-demand zones, particularly in Manhattan, during peak hours. Utilize the pickup/dropoff ratio analysis to strategically relocate taxis from zones with low demand/high dropoffs to zones with high demand/high pickups. For example, sending drivers from zones with high dropoff ratios but low pickup ratios to nearby zones with the opposite pattern would reduce empty mileage.
5. **Time of Day:**
   * **Peak Hours (Afternoon/Evening Weekdays, Evening/Night Weekends):** Focus on high fleet availability and efficient dispatching to the busiest zones during these times. Consider dynamic pricing strategies (surge pricing) during peak demand in high-traffic areas to balance supply and demand and incentivize drivers.
   * **Off-Peak Hours (Early Morning):** While overall demand is lower, identify specific night-time hotspots (e.g., entertainment districts, airports) and ensure sufficient coverage in these areas. Consider offering incentives for night-time driving in designated zones.
6. **Day of the Week:** Increase fleet size and driver availability on Fridays and Saturdays to cater to the higher weekend demand. Adjust operational strategies to reflect the later evening/night peaks on weekends compared to weekdays.
7. **Routing and Dispatching:**
   * Implement dynamic routing that considers real-time traffic data and historical "slow route" information, especially during peak hours, to minimize trip duration and improve efficiency.
   * Optimize dispatching algorithms to proactively send drivers to areas where high demand is anticipated based on historical hourly and daily patterns for specific zones.
   * Utilize the passenger count analysis to potentially recommend or prioritize larger vehicles in zones and times with historically higher average passenger counts.
8. **Pricing Strategies:**
   * Review and potentially adjust fare structures based on distance tiers and time of day, aligning with average fare per mile patterns to remain competitive and profitable.
   * Investigate factors influencing low tip percentages and explore ways to enhance the customer experience or driver service to encourage higher tips, which can also benefit driver retention.
9. **Payment and Customer Service:** Ensure seamless and reliable digital payment options, as credit card is the dominant payment type. Address any reported issues with surcharges or extra fees to maintain customer trust.

In essence, a successful strategy involves a continuous feedback loop of analyzing demand and operational data (location, time, day, passenger count, etc.), strategically positioning the fleet, implementing dynamic pricing and routing, and optimizing driver incentives to match supply with demand in real-time across the diverse landscape of NYC taxi operations.