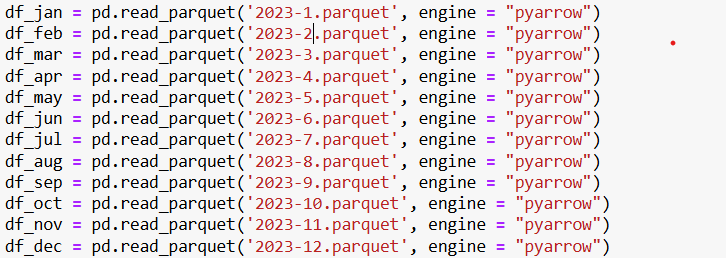
Report: Optimising NYC Taxi Operations

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

## Data Preparation

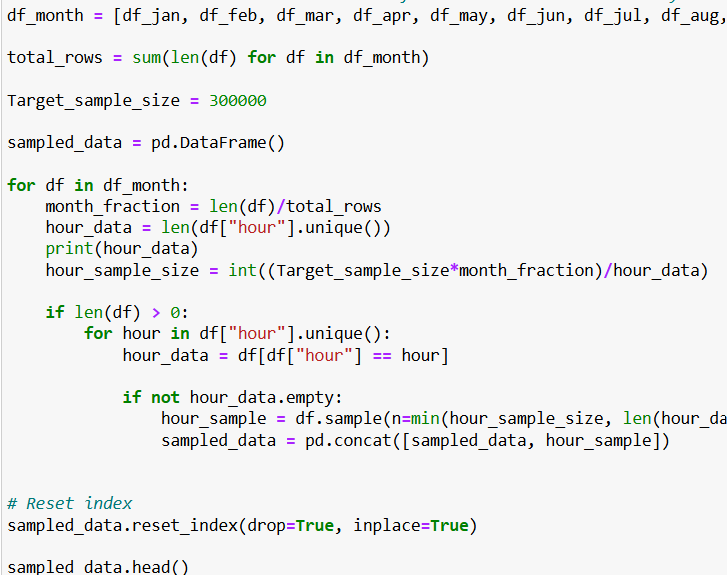
* 1. Loading the dataset



In above script, I get data from all 12 month using read\_parquet and consolidate multiple dataset files into single CSV file, ensuring all trips records are stored in a unified format. This allows for seamless data processing and analysis by combining different sources into one structured dataset.

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

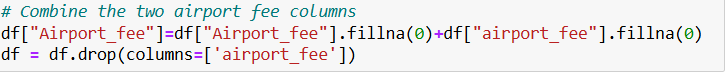
To ensure efficiency, a sample of the data was taken instead of using the entire dataset. Thae files were merged into single DataFrame for further processing.



## Data Cleaning

### Fixing Columns

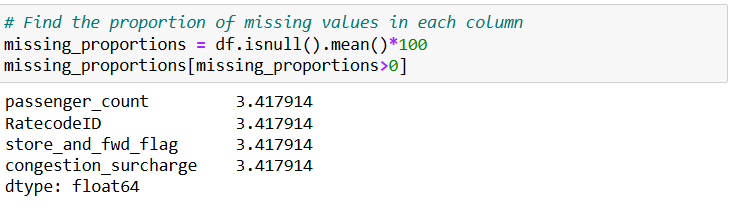
* + 1. **Fix the index**  
       The index was reset and standardized to maintain consistency.
    2. **Combine the two airport\_fee columns**The dataset ahd duplicate airport\_fee columns, which were merged to avoid redundancy, Ensuring accurate fare calculation was necessary for financial analysis

**

### Handling Missing Values

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

We checked for missing values, found the columns with NaNs. And filled them in to keep the dataset accurate.

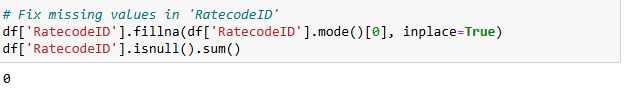


* + 1. **Handling missing values in passenger\_count**

Mode-based imputation was applied to missing values, leveraging the predominance of single-passenger rides to maintain dataset coherence.  
A screenshot of a computer code

AI-generated content may be incorrect.

* + 1. **Handle missing values in RatecodeID**The missing RatecodeIDs were imputed with the most common value, preserving the accuracy of trip classification and fare calculations.



* + 1. **Impute NaN in congestion\_surcharge**The missing congestion surcharge values were filled using median imputation, ensuring consistency in the pricing calculation throughout the dataset.

A computer screen shot of a computer code

AI-generated content may be incorrect.

### Handling Outliers and Standardising Values

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

I used box plots to spot outliers and the IQR methid to handle them. Really high trip distances and tips were double -checked to make sure they weren’t typos.

A group of graphs with numbers

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.

## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical  
       Categorical Variables :** VendorID,passnger\_count,RatecodeID, store\_and\_fwd\_flag, PULocationID, DOLocationID ,payment\_type

Numerical columns : trip\_distance, fare\_amount, extra, mta\_tax, tip\_amount, tolls\_amount, improvement\_surcharge, total\_amount, congestion\_surcharge, airport\_fee

**A computer code with red text

AI-generated content may be incorrect.**

* + 1. **Analyse the distribution of taxi pickups by hours, days of the week, and months**Visualizations showed peak demand during rush hours (morning and evening) and weekends. This was observed through trend of trip counts by time.

**A graph showing the amount of tax pickups

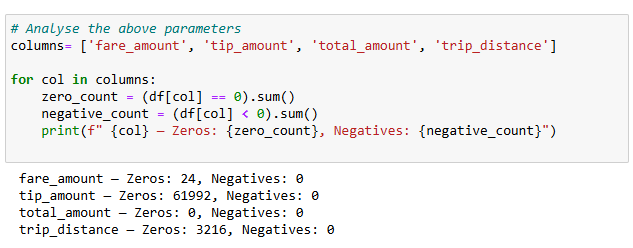
AI-generated content may be incorrect.**

**A graph showing the daily trend in taxi pickups

AI-generated content may be incorrect.**

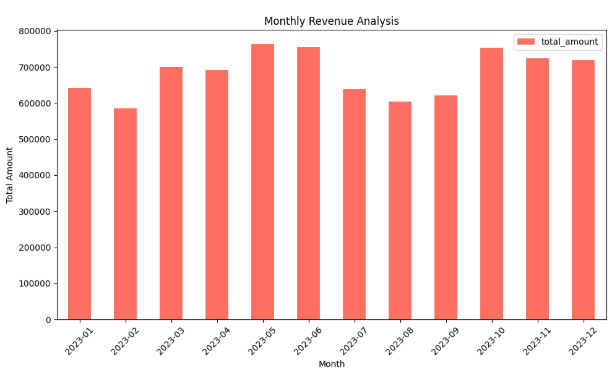
**A graph showing the number of pickups

AI-generated content may be incorrect.**

* + 1. **Filter out the zero/negative values in fares, distance and tips  
       **we got rid of entries with imposssible values, like zero or negative fares, distances, and tips, to keep the data accurate.

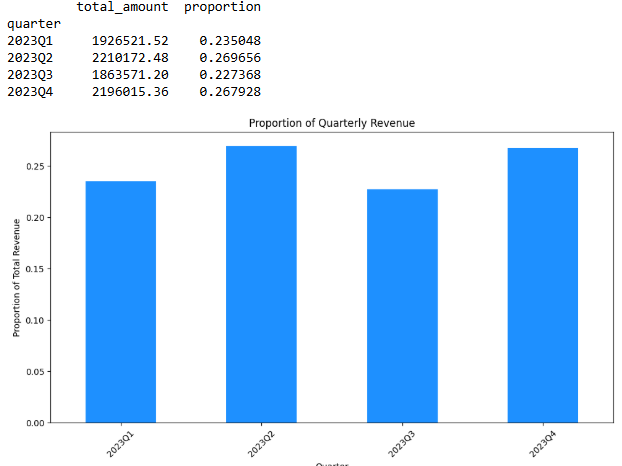
**A white box with black text

AI-generated content may be incorrect.**

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

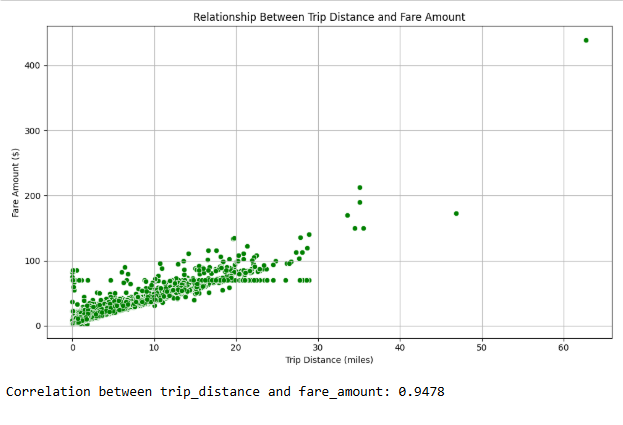
From above imag, Revenue peaks in May-June and Octomber-December, with dip in July – september, likely due to seasonality, promotions, or economic factors. Understanding these trends can help optimize buisness strategies.

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

****

Q2 (26.97%) and Q4(26.79%) contribute the highest revenue, while Q3 has the lowest share (22.74%). This suggests stronger business activity in late spring/early summer and the holiday season, likely driven by seasonal demand and promotions

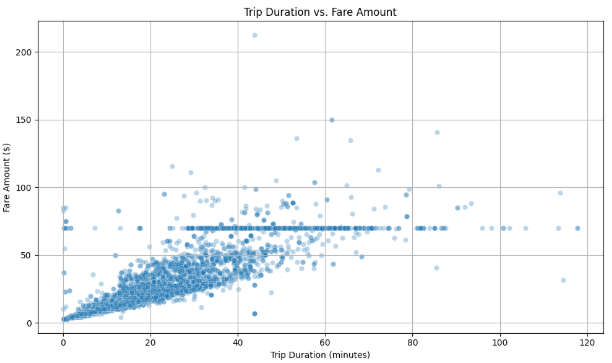
* + 1. **Analyse and visualise the relationship between distance and fare amount**

****

The high correlation suggests that fare pricing is primarily distance-based, possibly with additional charges like tolls, or peak – hour pricing

Outliers might indiacte premium rides, surge pricing, or exceptional cases.

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

1. **Trip Duration vs Fare Amount  
   **

Fare is mostly time-based, but outliers suggest factors like surge pricing, toll, or special services

Fixed fares might indicate common routes with standard pricing, such as airport transfers.

1. **Average Fare by Number of passenger**

A graph of red bars

AI-generated content may be incorrect.

Sared rides(2-4 Passengers) generate more revenue per trip, likely due to longer distance or higher demand.

Larger Groups(5-6 passengers) pay less on average, possibly due to fare splitting or discounts

1. **Tip Amount VS Trip Distance  
   A graph with green dots

   AI-generated content may be incorrect.**

Encouraging tipping for short trips through incemtives or service improvments could boost earnings.

Premium service for longer trips might justify higher tips.

* + 1. **Analyse the distribution of different payment types**cash and credit card payments were the most frequesnt. Difgital payment adoption was also analyzed.  
       A graph with numbers and a bar

       AI-generated content may be incorrect.
    2. **Load the taxi zones shapefile and display it**

A geographic representation of taxi zones was generated, providing insights into trip origin and destinations

**A screenshot of a computer code

AI-generated content may be incorrect.  
  
A map of the united states

AI-generated content may be incorrect.**

* + 1. **Merge the zone data with trips data  
       we added zone info to trip data to analyze locations and find the buisness areas.**

**A screenshot of a computer code

AI-generated content may be incorrect.**

* + 1. **Find the number of trips for each zone/location ID**the zone with the highest trip columes were identified, notably including JFK Airports, Midtown Center, Midtown East, Upper East Side North, Upper East Side South (Location ID : 132, 161, 162, 236, 237) respectively.  
       A screenshot of a computer

       AI-generated content may be incorrect.
    2. **Add the number of trips for each zone to the zones dataframe**Trip frequencies were added to the zones dataset for mapping demand visually.
    3. **Plot a map of the zones showing number of trips**A heatmap was created to visualize taxi demand across NYC, with hot zones indicating high pickup volumes **A map of new york with a number of locations

       AI-generated content may be incorrect.**
    4. **Conclude with results**
       1. Buisest Hours, Days, Months
* Peak Hours : Highest trip volumes occur during rush hours(morning & evening)
* Friday and weekeneds show increased demand, likely due to nightlife and leisure travel
* May & June see peak activity, while July -September show a decline before recovering in q4
  + - 1. Trends in Revenue Collected

Revenue is highest in Q2 (april -June) and Q4 (octomber – December) due to seasonal factors.

Late summer (July -September) see a dip, possibly due to reduced commuting and vacations.

### Detailed EDA: Insights and Strategies

* + 1. **Identify slow routes by comparing average speeds on different routes**Slow routes were identified by analyzing trip time and distance data, allowing for the detection of high-congestion areas

**A graph with different colored bars

AI-generated content may be incorrect.**

* + 1. **Calculate the hourly number of trips and identify the busy hours  
       A graph of a number of taxi trips

       AI-generated content may be incorrect.**From the chart, there is a significant increase in trips during the morning and evening, coinciding with peak commuting hours for work, school, and airport travel. Demand gradually declines at night, with lower trip volumes compared to the morning and evening peaks.
    2. **Scale up the number of trips from above to find the actual number of trips**

Hourly trip counts were extrapolated to estimate total trip volumes per day and month.

**A screenshot of a computer

AI-generated content may be incorrect.**

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

Weekday rush hours saw higher demand, while weekends had more leisure trips and night time activity **A graph with a line graph

AI-generated content may be incorrect.**

* + 1. **Identify the top 10 zones with high hourly pickups and drops  
       A graph of different colored bars

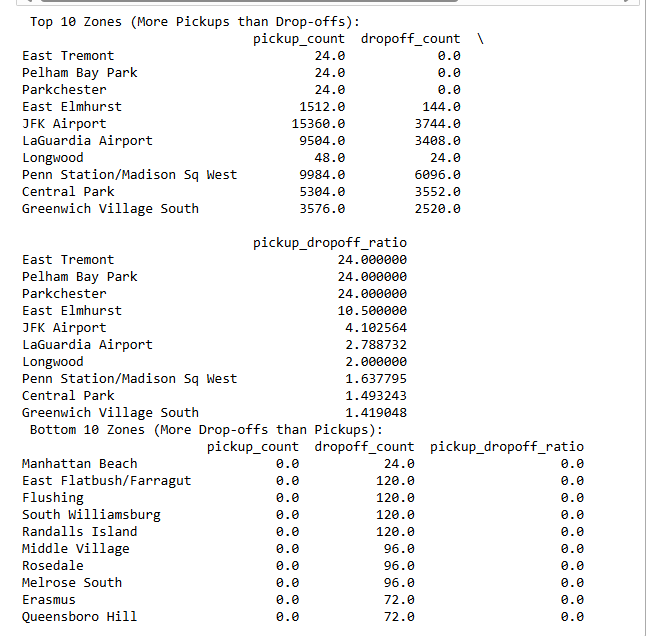
       AI-generated content may be incorrect.**

In the above chart, Top 10 zones where high hourly pickups.

**A graph of different colored bars

AI-generated content may be incorrect.**

In the above chart, top 10 zones where high hourly drops

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

Imbalnced zones were identified, indicating areas where taxis may be underutilized.

* + 1. **Identify the top zones with high traffic during night hours  
       A graph of different colored bars

       AI-generated content may be incorrect.**

As you see, nighttime high zone will be East village, west village and JFK Airports.

**A graph of different colored bars

AI-generated content may be incorrect.**

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

**A number on a white background

AI-generated content may be incorrect.**

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

**A graph with green bars

AI-generated content may be incorrect.**from the above image, we can say that as the number of passenger increases, the fare per mile per person decreases.

* + 1. **Find the average fare per mile by hours of the day and by days of the week**As you can see, on Wednesday and Sunday have high fare compare to others **A graph of a number of green bars

       AI-generated content may be incorrect.**
    2. **Analyse the average fare per mile for the different vendors  
       A graph of green bars

       AI-generated content may be incorrect.**
    3. **Compare the fare rates of different vendors in a distance-tiered fashion  
       A graph of a graph with a green bar

       AI-generated content may be incorrect.**
    4. **Analyse the tip percentage**

**A graph with a number of different colored bars

AI-generated content may be incorrect.**

A graph of a number of green rectangular bars

AI-generated content may be incorrect.

A graph with red lines and numbers

AI-generated content may be incorrect.

* + 1. **Analyse the trends in passenger count  
       A graph with lines and numbers

       AI-generated content may be incorrect.**
    2. **Analyse the variation of passenger counts across zones  
         
       A graph with blue and purple bars

       AI-generated content may be incorrect.**
    3. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.  
         
       A graph with blue bars

       AI-generated content may be incorrect.**

## Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**
* Increase taxi availability during peak hours, especially in high-demand locations such as **JFK Airport, Upper East Side South, and Midtown Center**, where pickups surge, particularly late at night.
* **Recommendation:** Station more taxis in these areas during **morning and evening rush hours**, as well as late-night weekends, to minimize passenger wait times.
* The **busiest period** for taxi trips occurs between **5 PM and 8 PM**, reflecting peak commuter demand.
* **Recommendation:** Expand the taxi fleet during these hours to maximize ride volume and service efficiency.
* Weekends witness a rise in **late-night travel and group bookings**, with larger passenger counts per trip.
* **Recommendation:** Prioritize coverage in nightlife hotspots such as **Greenwich Village and East Village** on **Friday and Saturday nights** to cater to increased demand.  
  + 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.**
* Boost taxi presence near **airports and nightlife districts** during late-night hours to accommodate increased demand.
* **Morning rush (6–10 AM):** Focus on residential neighborhoods such as the **Upper East/West Side and Queens** to serve commuters heading to work.
* **Afternoon hours (12–4 PM):** Shift deployment toward **business centers, shopping areas, and hospitals**, where midday travel is more frequent.
* **Evening and late-night (8 PM–2 AM):** Prioritize entertainment districts like **SoHo and the West Village**, which experience higher activity during these hours.
* Optimize taxi distribution by balancing fleets between **key pickup and drop-off locations**.
* **Weekday strategy:** Position taxis around **corporate hubs and transit stations** to cater to office workers
* **Weekend approach:** Focus on **tourist attractions, parks, and nightlife zones** to capture leisure travelers
  + 1. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**
* Introduce surge pricing during peak hours to align fares with increased demand.
* Short trips (under 2 miles) generate the highest cost per mile.  
  Recommendation: Adjust base fares slightly to optimize revenue from these trips.
* Long-distance rides (over 10 miles) face lower demand due to high costs.  
  Recommendation: Implement capped fares or promotional discounts to encourage airport and cross-borough travel.
* Peak demand is observed during morning (7–10 AM), evening (5–8 PM), and late-night weekends (9 PM–2 AM).
* Encourage shared rides with discounted fares to improve efficiency and reduce congestion.
* Flat rates for airport trips may not always yield optimal revenue.  
  Recommendation: Conduct an annual review of fixed fares based on fare-per-mile data to ensure they remain competitive yet profitable.

This analysis offers a data-driven framework to enhance NYC taxi operations, maximize revenue, and improve service efficiency.