

NYC Taxi Trips Analysis with Python and Tableau

Zsolt Oláh

2026.02.08

0. Scenario

I am a data analyst on the financial department team at a taxi company in New York. The director of finance wants to gain insights into **what factors drive revenue**. The initial hypothesis is that revenue is primarily influenced by **area- and time-related factors**. Therefore, I was asked to conduct an initial analysis of taxi trips in 2024, focusing on boroughs and different time periods, and examining how these factors affect revenue.

If these assumptions are validated, the director of finance can then collaborate with operational leadership to make data-driven decisions about reallocating the taxi fleet to higher revenue-generating areas in New York.

1. Ask

In this phase, I define the main business needs to be addressed by the analysis. Based on discussions with the director of finance, the analysis is designed to answer **two key questions**:

- **How do boroughs and different time periods affect overall revenue?**
- **Are there identifiable patterns in the data that can support decisions related to taxi fleet reallocation?**

2. Prepare

The data is publicly available and can be accessed via [this link](#). It is delivered in monthly segments, and I downloaded the complete dataset for the year 2024.

To assess data quality and suitability, I conducted a **ROCCC analysis**:

- **R – Reliable:** The dataset is generated by the automated electronic routing system, so it is considered reliable.
- **O – Original:** The data is collected directly by the taxi company, making it an original source.
- **C – Comprehensive:** The dataset contains sufficient information to answer our questions, such as where trips started and ended, when they started and ended, the distance of each trip, the payment method, and the trip amount, etc.
- **C – Current:** The analysis is conducted in 2025, and the data is relevant because it was collected over a full year and includes seasonality (202401–202412).
- **C – Cited:** The dataset originates from the company's own data collection.

After reviewing the structure of the data, I determined that the monthly files needed to be merged into a single, uniform dataset to allow for efficient processing and analysis.

3. Process

The dataset is segmented into monthly files and represents real-world operational data, so it required consolidation, cleaning, and preprocessing before analysis. I used **Python** to perform the following steps:

- Load the required libraries.
- Identify and import all 2024 yellow taxi trip files from local storage. The first file was used to define the column schema, which I enforced across all subsequent files to **ensure consistent formatting** before merging.
- Create and save a small random sample of the full dataset for external investigation (e.g., exploratory analysis or LLM-assisted review of structure, data quality, and integrity).
- Perform exploratory checks on individual columns to assess data types, value distributions, and the presence of missing values across categorical, numeric, and datetime fields.
- Convert datetime columns to proper datetime formats, coercing invalid values to nulls.
- Generate **derived columns**, including trip duration (in seconds and minutes), to enable time-based and efficiency-related analysis.
- Check for missing values and remove rows containing nulls in columns essential to the analysis.
- **Inspect numeric columns for invalid or unrealistic values** (e.g., extremely short or long trip durations) and **filter out entries outside reasonable thresholds**.
- Review datetime columns for anomalies; no critical outliers were identified after validation.
- Examine categorical columns for consistency by reviewing value distributions.
- Check for duplicated rows and **remove duplicates** to ensure each trip record is unique.
- Select and export a cleaned subset of relevant columns and generate additional sampled datasets for downstream visualization and analysis in **Tableau**.

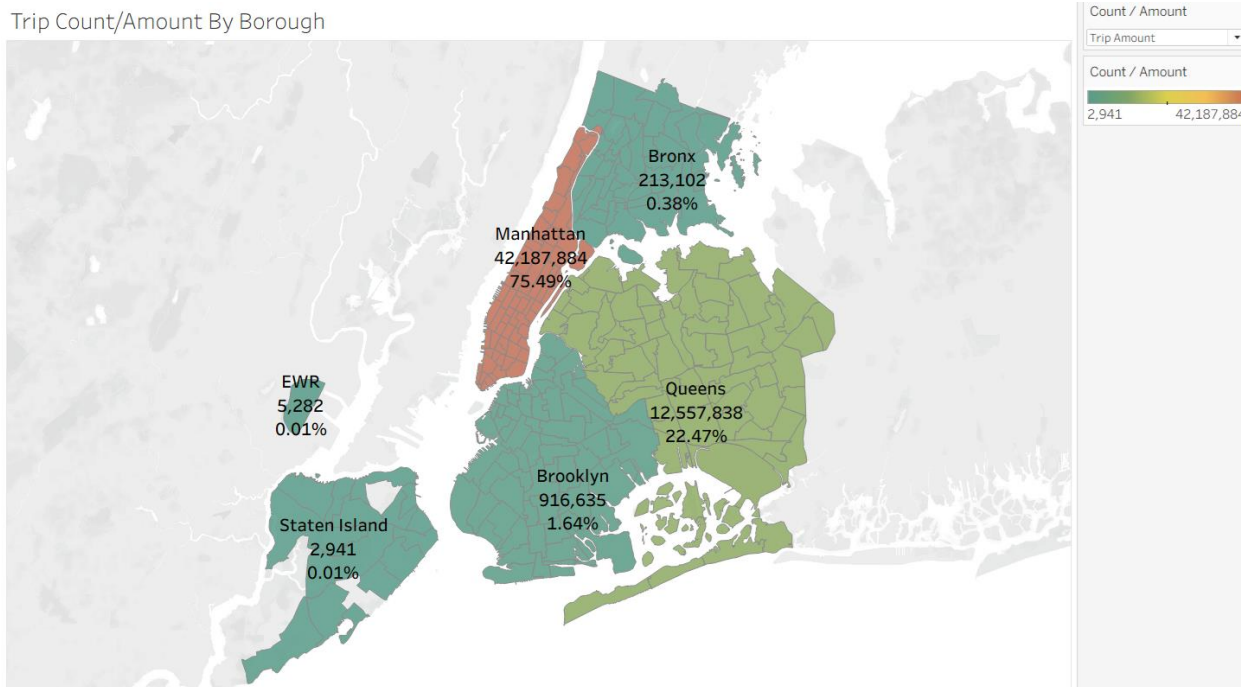
During preliminary exploratory analysis, I **identified an issue in an auxiliary shapefile** used to enrich the data with territorial zone information. **One taxi zone ID appeared twice**, which caused row duplication and misleading results during table joins. I **resolved this issue locally** by correcting the shapefile **and informed the data provider** of the inconsistency via email.

4. Analysis and Share

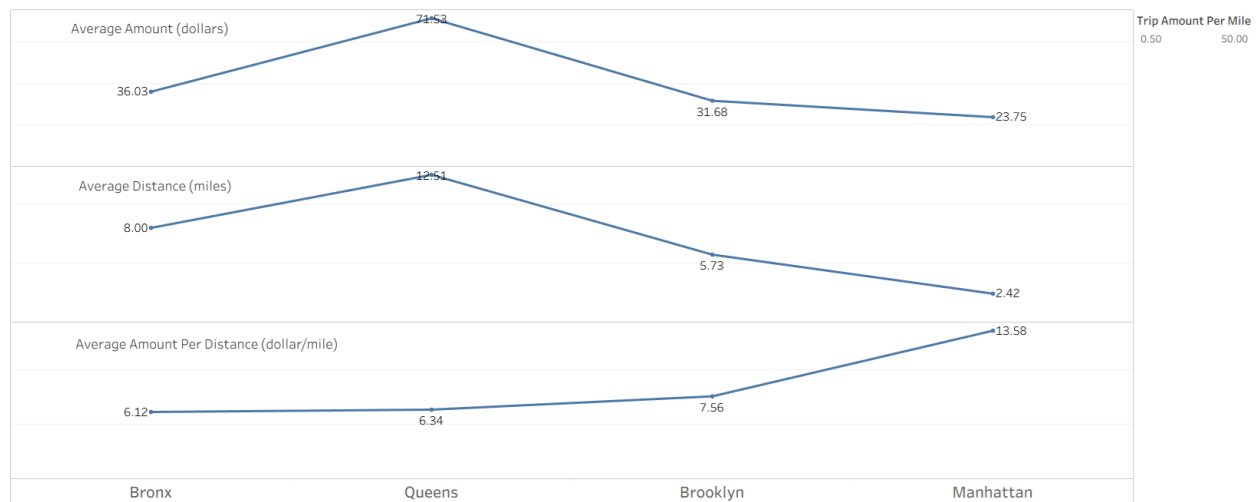
[You can view the story visualization on Tableau Public using this link.](#)

Key insights from the visual analysis include:

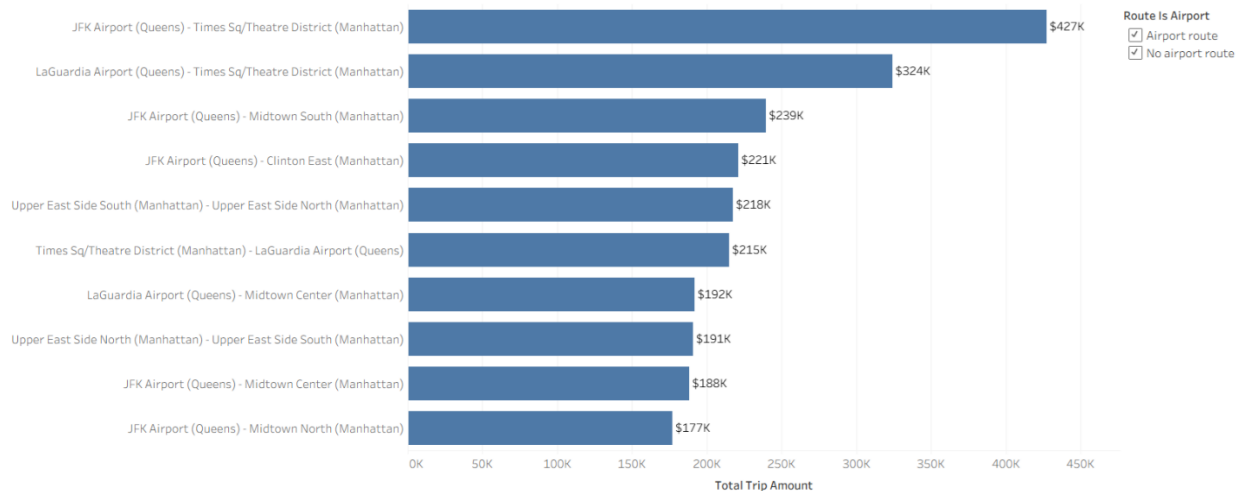
The first visualization is a map chart showing the spatial distribution of taxi trips and revenue. Taxi activity is **heavily concentrated in Manhattan**, which consistently leads in both trip volume and total revenue. However, **Queens emerges as a strong secondary contributor when revenue is considered**, suggesting higher-value trips relative to volume in certain areas. A filter allows users to switch between trip count and total revenue to explore this dynamic.



The second visualization is a line chart displaying average metrics by borough. Although Manhattan has the lowest average trip distance and total trip amount, it records the **highest revenue per mile**, indicating shorter but higher-value trips compared to outer boroughs. The chart is pre-filtered to exclude average trip amounts below \$0.50 and above \$50, with flexibility for users to adjust thresholds.

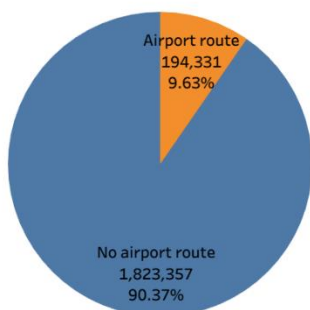


Next, a bar chart highlights the top ten most frequent pickup–drop-off zone pairs by total revenue. **Airport-related routes dominate** this ranking, prompting further analysis of high-value non-airport routes when airport trips are filtered out.

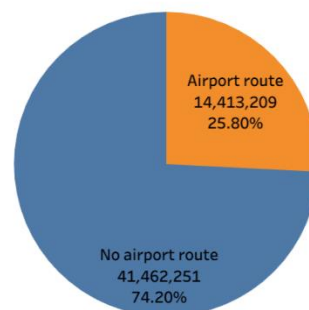


Two pie charts then compare airport-related and non–airport-related trips by trip count and total revenue. While airport-related trips represent **only about 10% of total rides**, they account for **more than 25% of overall revenue**, demonstrating a disproportionate revenue contribution.

Total Trip Count

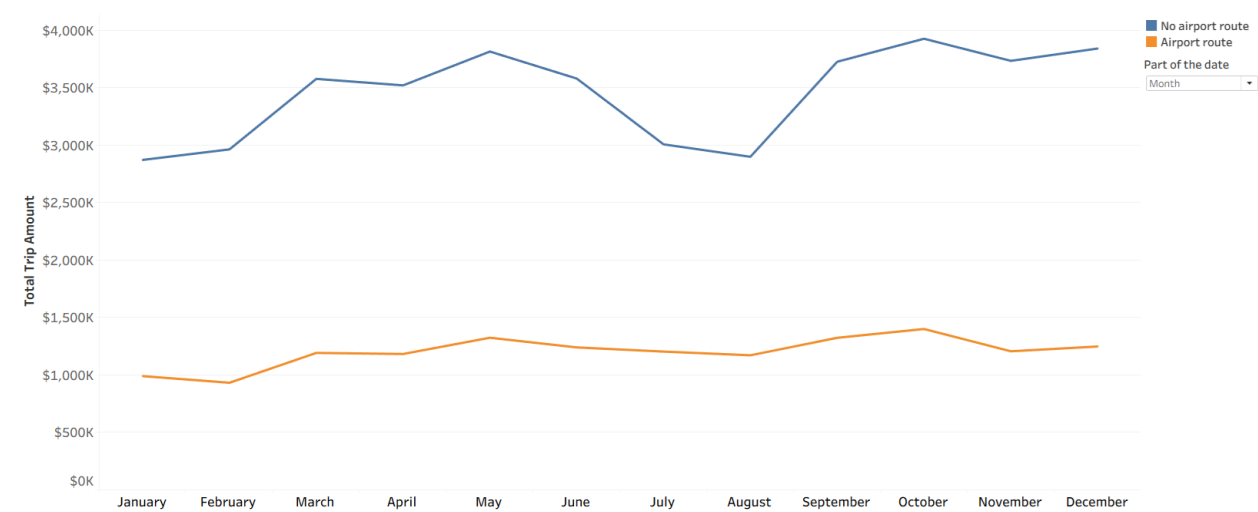


Total Trip Amount

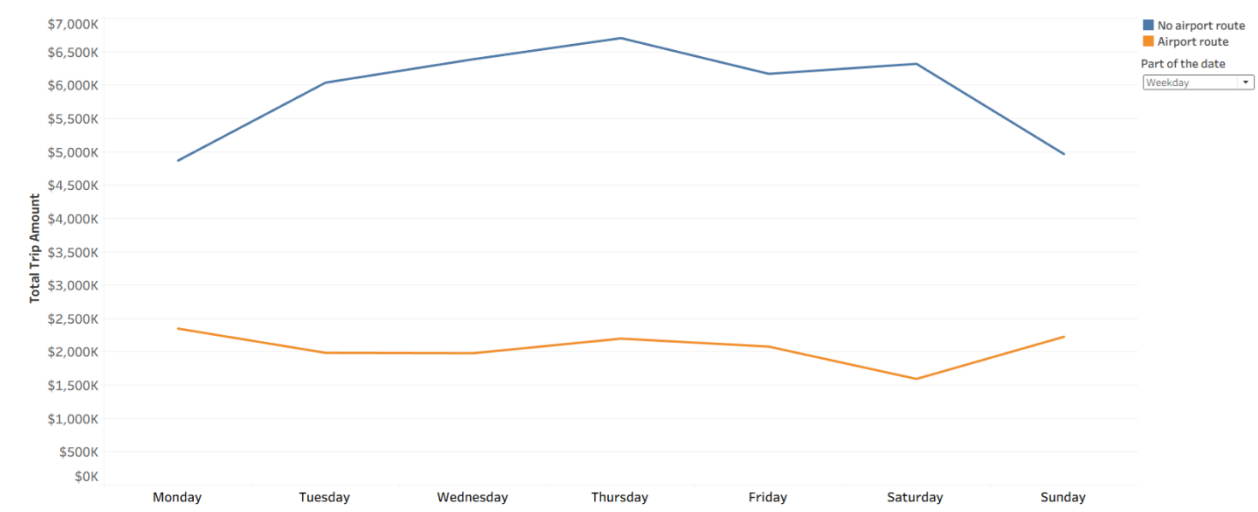


Finally, line charts illustrate revenue trends for airport and non-airport trips across monthly, weekday, and hourly dimensions. Although both trip types follow similar seasonal and intraday patterns, their **weekday trends diverge significantly**, with opposite revenue dynamics observed across the week.

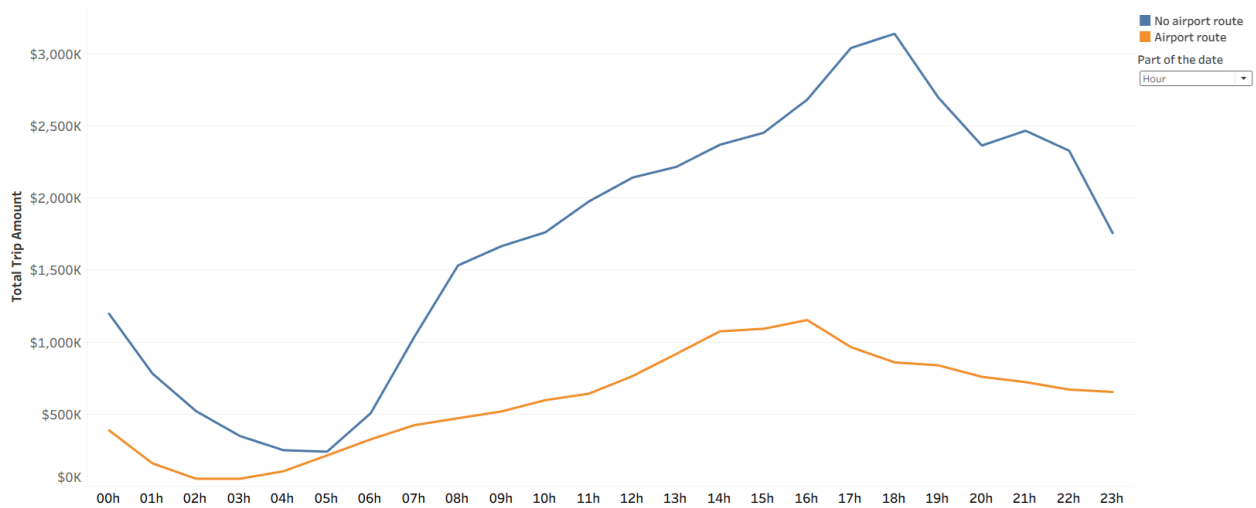
Monthly:



Weekdays:



Hourly:



5. Act

Based on these findings, I **developed three data-driven recommendations** for potential fleet reallocation, focusing on area- and time-based patterns:

- Given Manhattan's dominance in trip volume and revenue, alongside Queens' strong revenue performance relative to volume, it is recommended to further **analyze untapped revenue potential in Queens**. If validated, **increasing fleet capacity** in targeted areas of Queens could improve overall revenue.
- A similar imbalance exists between airport-related and non-airport-related trips. Airport trips generate a disproportionately high share of revenue relative to their volume, suggesting that **increasing capacity for airport-related trips** could yield meaningful revenue gains if demand allows.
- Weekday analysis reveals a notable decline in non-airport trip demand on Sundays and Mondays, while airport-related demand increases on these same days. **Temporarily reallocating non-airport drivers to airport-focused routes** during these periods may improve fleet utilization and drive incremental revenue.