Geospatial & Temporal Modelling of Hourly NYC Yellow Taxi Demand

James Nakos
Github Repository: Taxi Demand Modelling

September 29, 2025

1 Introduction

In the dynamic and bustling transport environment of New York City (NYC), travel demand patterns can swing within an instant. Reliable forecasts of hourly taxi demand are essential for maintaining operational efficiency throughout the network and ensuring optimal resource allocation. "Accurate predictions of short-term taxi demand"(Schleibaum et al. 2024) patterns are a vital insight for the NYC Taxi and Limousine Commission (TLC), our target audience, as they can be utilised to "strategically position idle vehicles... reducing idle times for taxis and waiting times for passengers"(Schleibaum et al. 2024).

This project focuses on Yellow Taxi trips across NYC during 2024. We begin by analysing higher-level spatial and temporal patterns in hourly demand levels across the city. Analysis focuses on Manhattan, where high trip density makes it the most operationally relevant borough. We aim to reveal insightful trends related to both subway ridership and weather conditions (primarily rainfall) that the TLC should consider as they operate throughout the year.

Two machine learning models will then be introduced, both predicting hourly demand by zone, given a range of engineered features, such as lagged pickup counts per hour and subway ridership 2). Features were chosen to capture short-term fluctuations and spatial diversity. The models also draw on local subway and weather data, allowing us to evaluate the predictability of taxi demand, acting as a useful hourly insight for the TLC to deduce where unmet demand is occurring.

1.1 Datasets

Each dataset used spans all of 2024; they are all free to use for non-commercial analysis. The primary dataset we utilised to explore taxi demand was the **TLC Trip Record Data** (New York City Taxi and Limousine Commission 2025), which contains granular trip-level records including timestamps, pickup zones (with shapefiles), etc. We define our focus: "demand" as trips per "hour" where the passenger is picked up in "zone". We also chose to incorporate subway ridership data, as subway usage likely reflects broader mobility patterns and helps contextualise fluctuations in taxi demand. We used the **MTA Subway Hourly Ridership** dataset (Metropolitan Transportation Authority 2025), which offers station-level hourly entry counts throughout the NYC subway network. Lastly, our NYC weather data for each borough was downloaded from the **Open-Meteo API** (Open-Meteo 2025), which was selected for its free access and reputation. Downloaded features included hourly rainfall, weather codes, snowfall, and cloud cover, as weather conditions are known to influence short-term travel demand in urban settings (Lepage and Morency 2021).

2 Preprocessing

2.1 Yellow Taxi Data

2.1.1 Validation & Cleaning

Original Dataset Size: $\rightarrow 41,169,720$

1. Schema: Ensured datatypes and feature names were synchronised across each month of data.
2. Missing Values: Most missing values were from "Payment Type 0", since these represented app-based payments/reservations (New York City Taxi and Limousine Commission 2025) and were not features of interest, the features were dropped, not the rows. Passenger counts were imputed with the mode, while all null zones were dropped since they are key to our analysis. Imputation was not possible. $\rightarrow 41,169,720$
3. Range Checking/Duplicates: For key features such as distance and fare amount, any records that fell outside defined reasonable ranges likely represented data errors/impossible trips and were dropped. There were no duplicate records. $\rightarrow 38,912,304$

2.1.2 Outlier & Distribution Analysis

Outlier Analysis: Distributions of key features were visualised (by sampling due to the large dataset size) to identify natural outlier cutoffs. A reasonable cutoff percentile for each attribute was then deduced by looking at the density of the data as well as the value for thresholds over the whole dataset.

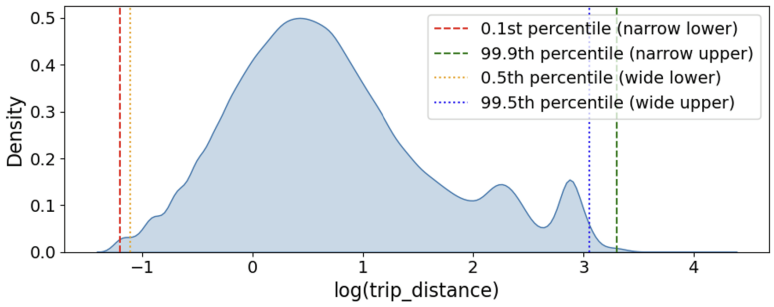

Figure 1: KDE of log Trip Distance

Figure 1 exhibits the cutoff for the 99.5th percentile falling within a dense cluster of valid longer trips; hence, we decided to cut off at the higher percentile for distance, while cutting wider for fares/duration. This ensures we exclude (drop) likely errors and rare anomalies while keeping data that reflects typical taxi trips, improving the reliability of subsequent analysis. $\rightarrow 38,095,566$

2.1.3 Aggregation & Feature Engineering

Records were grouped by (hour & pickup location), aggregating key statistics such as trip count, fare totals/averages, and durations. Aggregation ensured all (hour, zone) combinations were represented and filled with zeroes if no data was given (assuming complete data). $\rightarrow 2,292,624$ Our trip count feature had a strong right skew; however, the larger values still seemed plausible, removing them as outliers would harm future analysis, as we would lose data due to joins and possibly introduce bias.

Features engineered at this stage include:

- Temporal features such as hour of day, month, season, day of week, etc.
- Percentage of trips within borough for each (hour,zone) pair (short trips).

2.2 Subway Data

2.2.1 Validation & Cleaning

Original Dataset Size: $\rightarrow 27,023,937$

1. Filtered to only subway data (exclude trams). $\rightarrow 26,803,065$
2. No nulls, duplicates or values out of range, so we kept each record.
3. Each record was grouped by (hour,station), ridership was aggregated. $\rightarrow 3,649,615$
4. Added each subway station's taxi zone (the zone it is within) as an attribute for joining.
5. Saturated all (hour, zone) combinations present in our subway data with records of 0 ridership to ensure a complete join and no loss of taxi records.

2.2.2 Outlier & Distribution Analysis

The distribution of ridership was strongly right-skewed; however, the values of potential outliers were still reasonable for areas of high demand. It would be more harmful to remove them, as we would be unable to join with aggregated taxi data for these (hour,station) pairs, so we kept them to avoid biasing our analysis to off-peak periods.

2.3 Weather Data

2.3.1 Validation & Cleaning

1. Converted attributes to their correct data type (e.g., timestamp from string to date/time).
2. No nulls, out of range values or duplicates.

2.3.2 Outlier & Distribution Analysis

Data was clean; most attributes looked normally distributed (except precipitation); however, outliers were all reasonable values, so they were kept to preserve future joins.

2.3.3 Feature Engineering

- Flags for several certain interpretable weather conditions made by thresholding continuous attributes. E.g., flagging extreme cold hours, hours with rain, binning rain intensity, etc.

2.4 Further Feature Engineering & Joining

Since both taxi and subway data now had complete grids (even hours with no activity are represented), the datasets were then left joined, so all taxi records were retained. The resulting dataset was then joined with the complete weather data to obtain a final dataset with $2,292,885$ rows. Lastly, 4,437 rows were dropped due to Spark timezone mismatch issues, creating nulls. $\rightarrow 2,288,448$.

For Analysis

- **Taxi-Subway Ratio Residual:** Measures how unusual the taxi/subway trip ratio is for a zone and hour, compared to its typical value (z-score).
- **Is High Anomaly:** A binary flag set to 1 if the ratio residual is greater than 1, indicating a significant anomaly. Used to identify and count periods of unusually high taxi demand relative to subway usage.
- **Swing Ratio:** Shows the relative change in mean (or median) taxi demand between rainy and dry conditions for a zone.

For Modelling

- **Taxi Subway Ratio:** Captures how heavily taxis are used compared to subway ridership per given zone and hour.
- **Lag Features:** Includes past values of all variables, such as trips 1 hour ago, or ridership 1 day ago, to help models learn from recent trends and prevent leakage.
- **Rolling Averages:** Smooths out short-term fluctuations by averaging features over a given time window, establishing what's normal within that period.
- **Growth Features:** Demand change indicators comparing last hour's values to previous ones.

Table 1: Features engineered for analysis & modelling

Since lag and average features required past data to compute, they created nulls, which we imputed with zeroes. Median imputation was not possible as it was too computationally expensive while modelling.

Our **attributes of interest** are hence the attributes used to create and visualise our analysis features, these are: trips, timestamp (hour), hour of day, pickup zone, subway ridership, temperature, and weather code.

3 Analysis and Geospatial Visualisation

3.1 Taxi, Subway Analysis

To explore the effect of subway ridership on taxi demand measured as the hourly pickup count in each zone, we created the aforementioned "is anomaly" metric, which helps identify when and where unusual swings in taxi and subway demand patterns occur. This was done by calculating the number of hours with unusual demand swings (grouped by hour of day, zone) and graphing it as a proportion of all hours (with the same hour of day) in that zone. This ensured that we identified anomalies regardless of time of day, creating a more even distribution and a more insightful, applicable measure that is resistant to peak-hour skews. We set this up as a unidirectional measure, only showing the proportion of hours in which we experience swings that increase taxi demand relative to subway ridership.

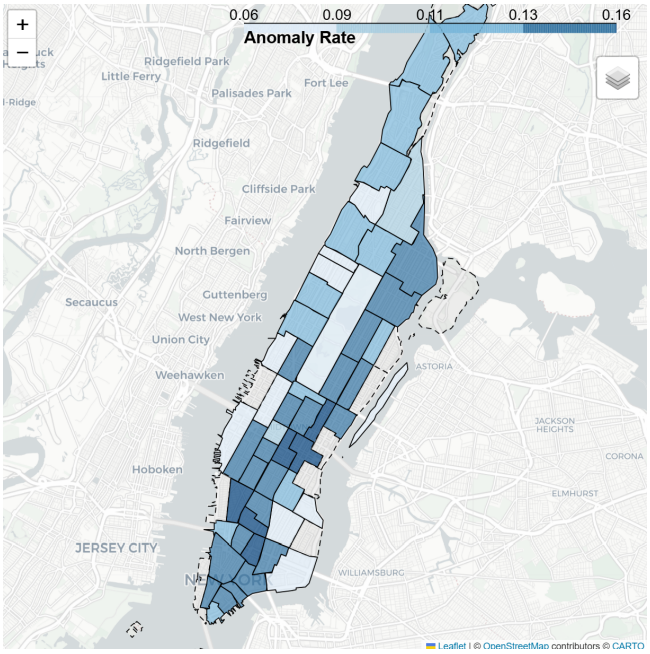

Figure 2: Positive Taxi Zone Anomaly Rate. Manhattan, 2024

Figure 2 reveals that high anomaly taxi demand rates are not evenly distributed across Manhattan's taxi zones, but instead cluster around certain zones such as Midtown, East Harlem and West Village, while other areas such as Central Park experience more steady demand. This suggests that taxi-subway demand interaction is stronger in these high anomaly areas, where factors such as disruptions or ridership swings are more likely to translate into spikes in taxi demand. This is perhaps because these areas have dense, subway-reliant populations, and there exists limited redundancy in alternative transport, so when subway services falter, people are more likely to switch to taxis. Although this doesn't fully capture where unserved demand lies, it acts as a useful indicator that the TLC can factor in when **planning and reallocating fleet during subway outages or anomalous days**. This anomaly rate could be further exploited alongside other data, such as recent pickup counts, for the TLC to maximise operational efficiency on anomalous days.

3.2 Taxi, Weather Analysis

Next, we analyse the effect of the presence of rainfall on taxi demand patterns. We utilised our (is rain) feature that flagged an hour as rainy if any form of rain was present to do so.

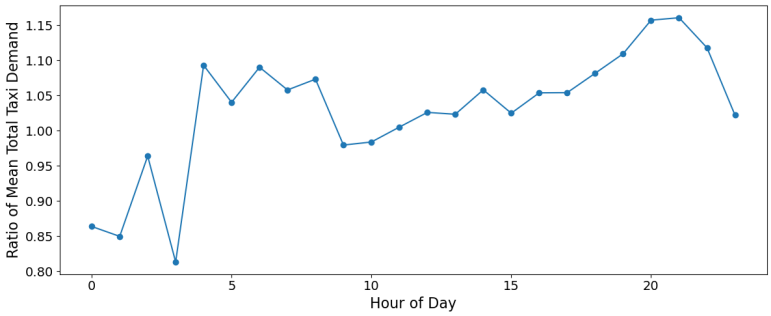

Figure 3: Ratio of hourly mean taxi demand (rainy vs dry hours) in Manhattan

Figure 3 illustrates that taxi demand in Manhattan is noticeably higher during rainy hours as opposed to their dry average for most of the day. There are two clear periods of heightened taxi demand during rainy hours: a sharp morning peak that occurs around 6 AM, and a broader evening surge appearing from 6-10 PM, where we have 15% more taxi trips than our dry average. Therefore, rain appears to amplify existing commute and nightlife demand surges, perhaps because commuters are less likely to seek other options such as walking or cycling during these periods. Midday ratios remain steady, suggesting rain has little effect on demand, possibly due to greater availability of other transport & flexible trip timing, hence customers may want to wait for rain to pass as they do not need to travel instantly. Demand ratios, however, drop below 1 after midnight, although this is likely due to a lack of data rather than a true behavioural drop. We can utilise these findings to advise the TLC to **encourage more drivers to work in Manhattan during rainy peak periods and early mornings, and avoid overcompensating with fleet increases around midday**.

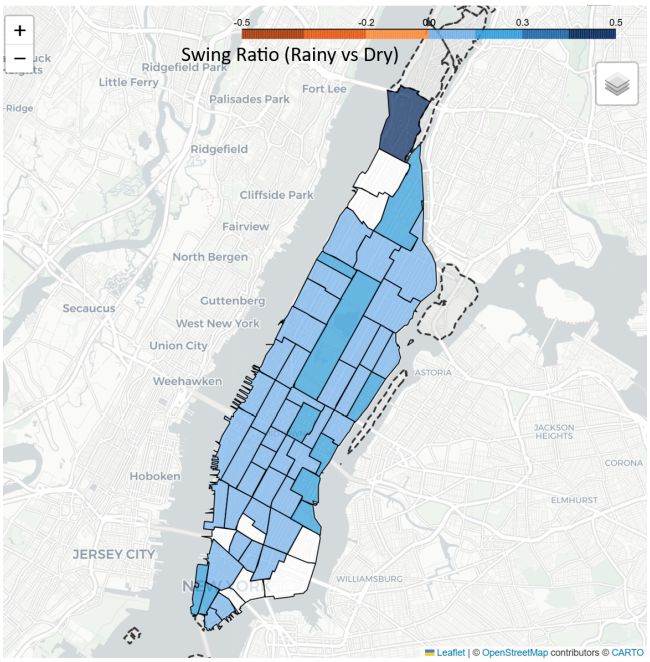

Figure 4: "Swing ratio" by Manhattan taxi zone (rainy vs dry)

Our "swing ratio" map 4, displays the relative change in median trip demand between rainy and dry hours across Manhattan. It exhibits clear spatial heterogeneity, with certain zones in Upper Manhattan, such as Washington Heights South and Central Harlem North, experiencing disproportionately higher increases, while parts of lower Manhattan, like Chinatown, exhibit little to no shift in demand. This suggests that the effect of rain on taxi usage is not uniform, but rather it amplifies demand in certain more residential peripheral zones, perhaps due to reduced walking options and less subway accessibility. As opposed to lower Manhattan, which is primarily "commercial districts" (New York City Department of City Planning 2025), where demand remains steadier for the adverse. Our utilisation of the median rather than the mean for trip counts helps address the data's natural right skew and helps capture a more typical effect of rain on demand. Thus, these findings could guide the TLC in dynamically reallocating supply during hours with forecast rain.

4 Modelling

4.1 Features

Response: Trip count per hour in each zone. This is a non-negative, discrete occurrence count over a fixed interval. Each model predicted this count.

Inputs:

- Temporal features such as hour of day, season, day of week.
- Spatial features such as zone & pickup borough.
- All available hourly weather features such as precipitation, weather codes, etc.
- All lagged as well as rolling averaged trip counts, ridership, and associated data such as lagged average distance, etc. No hourly features, as they introduce leakage.

While modelling, we are assuming that the historical weather data we have is a suitable proxy for forecast data since we do not have actual forecasts per hour, and that taxi and subway data are complete and leakage-free.

Before modelling, all categorical features, such as binary flags and our temporal features, were converted into integer indices and one-hot encoded, ensuring they were properly represented in our models.

4.2 Poisson Regression

We first employ a Poisson regression model (a generalised linear model (GLM)), since our response consists of count data, which is naturally modelled by the Poisson distribution (Shim 2025). Our main assumption for this model is that the count's mean is equal to its variance. The model expresses the log of the expected count (demand) as a function of our predictors through the canonical log link: Shim 2025

$$\log(\mathbb{E}[Y_i \mid X_i]) = X_i^\top \beta$$

4.3 Gradient Boosted Trees (GBTs)

We then chose to build a supervised GBT model, which predicts hourly taxi demand by sequentially building an ensemble of decision trees, with each tree trained to correct the errors (residuals) of the previous trees. GBTs capture complex nonlinear relationships between features and trips and are robust to overdispersion (a weakness of our GLM).

4.4 Training & Analysis

Both models were trained on our 2024 dataset, split into January to August for training, September to October for validation, and November to December for testing to avoid leakage.

- To increase predictive performance of the GLM and reduce overfitting, L2 (Ridge Regression) was implemented, which shrinks parameters towards 0 at a given rate (λ) while training. We experimented with different values and found $\lambda = 0.01$ as a value that maximised performance on our validation set, reducing Root Mean Square Error (RMSE) from $\approx 30 \rightarrow 24.16$.
- Time constraints limited tuning of our GBT model; however, the model reduces error as it learns, so parameter tuning was not extremely necessary with our model still yielding strong results: Validation RMSE = 10.92.

4.4.1 Comparison

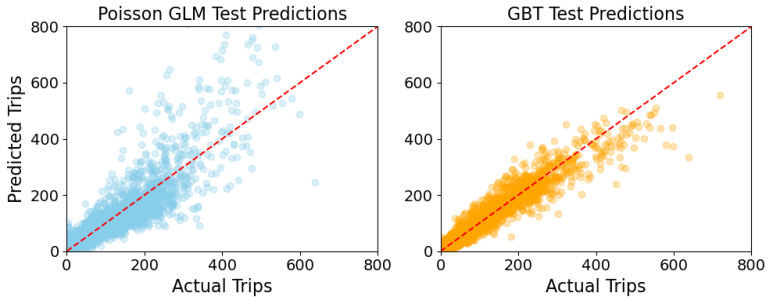

Figure 5: Sampled Model Test Predictions Comparison (GLM on Left), (GBT on right)

As illustrated in Figure 5, clear performative differences emerged between the GLM and GBT models. On the test set, the GLM achieved RMSE = 26.31 & R^2 = 0.77, while the GBT achieved RMSE = 11.76 & R^2 = 0.95. The GLM shows a systematic under-prediction bias at higher demand levels 5, likely attributable to overdispersion, where variance in our data exceeds the mean, violating the Poisson assumption.

In contrast, the GBT model produces predictions that cluster tightly around the ideal prediction line across different demand levels; this reveals its superiority for taxi demand prediction, as it is less affected by increased natural variation during high demand periods, a time the TLC would require accurate predictions.

Although our Poisson GLM tends to underpredict sharp spikes, it offers benefits in interpretability; its coefficients/weights can be helpful for the TLC to quantify how much changing conditions, such as time of day, rainfall, and temperature, shift baseline demand across the city. However, the GBT model with half the RMSE is much more trustworthy; it can be used to flag possible large demand increases and provide the TLC with actionable information to improve taxi service efficiency across NYC. Accurate forecasts of hourly trip demand per zone could be used by the TLC in conjunction with fleet data to **advise or incentivise drivers to position themselves in specific zones when and where supply gaps are expected**. This would offer the TLC a way to increase potential revenue and fleet utilisation, all while reducing passenger wait times.

5 Conclusion & Recommendations

Although recommendations have been highlighted/justified throughout the analysis and modelling sections, we will now consolidate the key findings into actionable guidance for the TLC. Our analysis highlights that taxi demand in NYC is strongly shaped by both subway disruptions and rainfall-driven surges, with effects more prevalent in specific zones during certain time periods. The Poisson GLM offers interpretable insights on how conditions shift baseline demand, while the GBT's higher accuracy makes it a reliable indicator for forecasting demand surges. We recommend that TLC advise and incentivise drivers to position themselves in high anomaly zones during subway outages, and increase fleet availability in Manhattan during rainy peak hours, especially in residential areas such as Upper Manhattan. By utilising the GLM for high-level policy insights and the GBT for operations, the TLC can reduce wait times, improve fleet usage, and increase driver revenue.

References

Schleibaum, Sören et al. (2024). "A systematic analysis of design choices in short-term taxi demand prediction models". In: *Transportation Research Procedia* 78. Accessed: 2025-08-27, pp. 554–561. ISSN: 2352-1465. DOI: https://doi.org/10.1016/j.trpro.2024.02.069. URL: https://www.sciencedirect.com/science/article/pii/S2352146524001236.

New York City Taxi and Limousine Commission (2025). *TLC Trip Record Data*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2025-08-15. URL: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

Metropolitan Transportation Authority (2025). *MTA Subway Hourly Ridership: 2020–2024*. https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-2020-2024/wujg-7c2s/about_data. Accessed: 2025-08-15. URL: https://data.ny.gov/Transportation/MTA-Subway-Hourly-Ridership-2020-2024/wujg-7c2s/about_data.

Open-Meteo (2025). *Open-Meteo: Free Open-Source Weather API*. https://open-meteo.com/. Accessed: 2025-08-15. URL: https://open-meteo.com/.

Lepage, Simon and Catherine Morency (2021). "Impact of Weather, Activities, and Service Disruptions on Transportation Demand". In: *Transportation Research Record* 2675.1. Accessed: 2025-08-28, pp. 294–304. DOI: 10.1177/0361198120966326.

New York City Department of City Planning (2025). *Commercial Districts Guide*. Accessed August 30, 2025. URL: https://www.nyc.gov/content/planning/pages/zoning/zoning-districts-guide/commercial-districts.

Shim, Dr Heejung (2025). *Modern Applied Statistics (MAST30027) Lecture Notes*. Accessed: 2025-08-25.