NYC Taxi vs Citi Bike Wait- and Travel-Time Estimation

ECE 225A – Fall 2025 Project
Atharv R. N.

December 14, 2025

Abstract

We build a reproducible pipeline for comparing New York City taxi and Citi Bike service quality. Component 1 estimates rider wait times by modeling arrivals as Poisson / Negative-Binomial processes and translating inter-arrival times into exponential wait-time distributions. Component 2 focuses on travel-time estimation: we fit discrete Gamma distributions per distance cohort and a continuous lognormal accelerated failure-time model that removes the need for coarse bins. Both components feed a Streamlit decision-support dashboard. This report describes the data, modeling choices, UX considerations, and future work.

1 Problem Setup

Urban travelers often need a quick answer to "Should I bike or hail a taxi?". The decision hinges on *wait time* (how quickly a vehicle becomes available) and *travel time* (how long the trip takes once underway). We tackle the modeling side by:

1. Estimating arrival counts and wait times for taxis and Citi Bike stations.
2. Predicting trip durations as a function of distance, rush-hour, and weekend flags.
3. Surfacing the results in an interactive web app.

Assumptions include:

- Wait times are proxied by inter-arrival gaps between consecutive pickups (taxis) or bike check-outs at a station.
- Travel-time samples are filtered to 1–120 minutes and capped at 12 km to keep taxi and bike cohorts comparable.
- Weather and roadway disruptions are not modeled (future work).

2 Data Sources & Preprocessing

Taxi. Jan–Jun 2024 Yellow Taxi Parquet files from NYC TLC [1]. Each trip provides pickup/dropoff timestamps and mileage plus zone IDs. We join with TLC zone metadata to bin by neighborhood and compute inter-arrival counts.

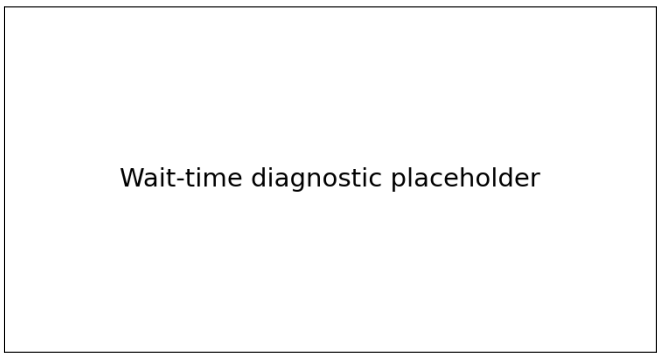

Figure 1: Taxi wait-time diagnostic (arrivals + exponential overlay). Replace with actual PNG export.

Citi Bike. Jan–Jun 2024 Citi Bike trip CSVs from the public S3 mirror [2]. Each record contains start/end timestamps and station metadata. Distances are computed via haversine formulas using station coordinates.

Derived artifacts. Scripts generate:

- Station-level rate summaries (‘taxi_rates.parquet‘, ‘citibike_rates.parquet‘).
- Travel-time bin stats (‘travel_bins.parquet‘) and lognormal GLM coefficients (‘travel_lognormal_glm.json‘).

3 Wait-Time & Arrival Modeling

3.1 Arrivals

For each taxi zone and Citi Bike station we bucket arrivals into 5/15/30/60-minute windows. We require a minimum average rate (e.g., 1 taxi per 15 minutes) and a non-zero proportion threshold to focus on active locations. Method-of-moments Negative-Binomial fits (parameters (r, p)) capture over-dispersion; Poisson curves are plotted as baselines.

3.2 Wait Times

Given an arrival process with mean λ, the expected wait time for a passenger is $\frac{1}{\lambda}$ (memoryless exponential). We validate this by computing empirical inter-arrival gaps and overlaying exponential PDFs. See Fig. 1 for a representative screenshot (placeholder).

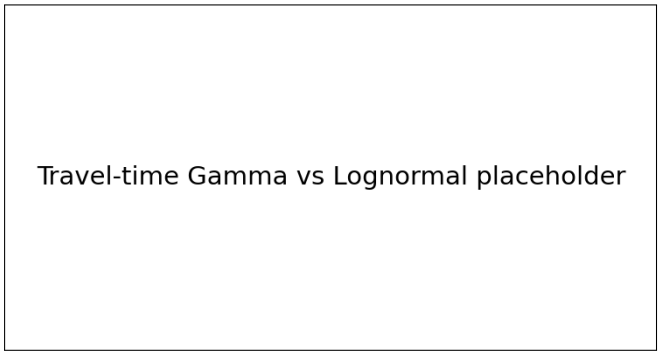

Figure 2: Gamma vs lognormal overlay for a 4–6 km Citi Bike cohort (placeholder).

Figure 3: Streamlit UI mock/screenshot placeholder.

4 Travel-Time Analysis

4.1 Gamma Cohorts

Trips are bucketed by mode, 2 km distance bins (0–2, 2–4, ..., 10–12 km), rush vs off-peak, and weekend vs weekday. For cohorts with ≥ 50 samples we estimate Gamma shape/scale via mean/variance. These bins are easy to cache and interpret but struggle with sparse Citi Bike cohorts.

4.2 Lognormal GLM

To remove binning artifacts we fit a lognormal accelerated failure-time model:

$$\log(\text{travel\_min}) = \beta_0 + \beta_1 d + \beta_2 d^2 + \beta_3 \text{rush} + \beta_4 \text{weekend} + \varepsilon,$$

with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The expected travel time is $\mathbb{E}[T] = \exp(\mu + \frac{1}{2}\sigma^2)$. We evaluate MAE/RMSE per mode (Table 1) and include combined histograms (Gamma vs lognormal) in the notebook (Fig. 2 placeholder).

Table 1: MAE/RMSE (minutes) comparing Gamma bin means vs lognormal GLM. Replace values with notebook outputs.

Mode	Model	MAE	RMSE
Taxi	Gamma	–	–
Taxi	Lognormal	–	–
Bike	Gamma	–	–
Bike	Lognormal	–	–

5 Streamlit Dashboard

The front end (https://github.com/AtharvRN/NYC_Public_Transit) uses Streamlit to let users:

- Pick origin/destination on an interactive map (Folium).
- View taxi vs bike wait times for the selected locations.
- See travel-time estimates using the lognormal GLM (with Gamma fallback).
- Compare total journey time distributions.

5.1 UX Notes

Tabs separate taxi/bike diagnostics; tooltips explain why certain cohorts are disabled (insufficient samples). We plan to add UI screenshots here (placeholder Fig. 3).

6 Limitations & Future Work

- **Modeling assumptions:** Wait-times use inter-arrival proxies; no weather or traffic features; GLM only uses distance/rush/weekend.
- **Data coverage:** Only Jan–Jun 2024; Citi Bike GPS routes or e-bike vs classic distinctions are ignored.
- **Scalability:** Subway data is not integrated yet—downloading GTFS and MTA ridership feeds is planned but time-consuming.
- **Future features:** Add uncertainty bands, enable user-uploaded O/D pairs, integrate subway travel-time estimators, and experiment with probabilistic travel-time percentiles.

7 Conclusion

We delivered a reproducible workflow: raw data ingestion, exploratory notebooks, lognormal travel-time modeling, and a Streamlit app. The lognormal GLM improves Citi Bike estimates substantially over Gamma bins while keeping the pipeline explainable. Next steps include richer covariates, subway integration, and automated CI for the derived artifacts.

References

[1] NYC Taxi & Limousine Commission, "TLC Trip Record Data," https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[2] Citi Bike NYC, "Historical Trip Data," https://s3.amazonaws.com/tripdata/index.html