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

ECE 225A – Project
Atharv Nair (A69042035), Nitin Shreyes (A69041917)

December 14, 2025

Abstract

We build a reproducible pipeline for comparing New York City taxi and Citi Bike service quality. The first pillar estimates rider wait times by modeling quarter-hour arrivals (split by rush/off-peak and weekday/weekend) as Poisson / Negative-Binomial processes and translating inter-arrival times into exponential wait-time distributions. The second pillar focuses on travel-time estimation: we fit discrete Gamma distributions per distance cohort and a continuous lognormal accelerated failure-time model—augmented with an e-bike indicator—that removes the need for coarse bins. Both tracks feed a Streamlit decision-support dashboard backed by cached Parquet/JSON artifacts. This report describes the data, modeling choices, UX considerations, and future work.

1 Introduction

Even a short Manhattan hop raises the question: *Is it quicker to unlock a Citi Bike or flag a yellow cab?* Taxi supply, dock availability, and demand swing minute by minute, so rule-of-thumb advice ("taxis are faster at night") quickly breaks down. Instead of mimicking a full routing engine we zoom in on the probability distributions that control two door-to-door components:

- **Wait time.** How long until a taxi arrives or a bike becomes available in the nearby dock cluster?
- **Travel time.** Once a rider is moving, how long will the trip take as a function of distance and traffic patterns?

To answer these questions we fit probability distributions directly to the NYC Taxi and Citi Bike datasets (Jan–Jun 2024). Arrivals are modeled with Poisson/Negative-Binomial processes and converted to exponential wait distributions, while trip durations are handled with Gamma cohorts and a lognormal accelerated failure-time model. Both estimates feed a public Streamlit app (https://nycpublictransit.streamlit.app/) that lets users click origin/destination pairs and instantly compare modes (Fig. 1).

Assumptions and design choices:

- Wait times rely on inter-arrival gaps as a proxy for rider experience. Stations within 200 m are pooled into "catchments" to reflect a rider's willingness to walk to the next dock.

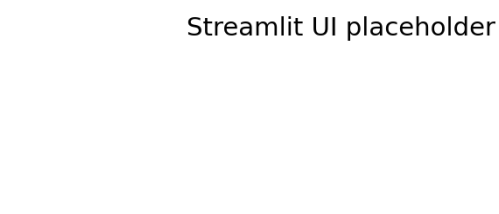


Figure 1: Public dashboard (https://nycpublictransit.streamlit.app/) showing taxi vs Citi Bike recommendations for arbitrary O/D pairs.

- Travel-time samples are restricted to 1–120 minutes and within 12 km to keep Citi Bike and taxi cohorts comparable. Forecasts outside this range fall back to coarser heuristics.
- We only use features available historically (distance, rush/weekend indicators, e-bike/classic flag). Weather, incidents, dynamic pricing, and subway usage are left for future work.

2 Dataset

Taxi

Citibike:

- The dataset contains 18M entries spread across six months,01/2024 - 06/2024.
- Start Station and End Station along with their location coordinates were provided.
- Travel time was provided in 'i64' format.
- Member/Casual, Electric/Classic bike information was also given
- Around 2.2k unique start and stop locations were identified
- Over 1.14M unique routes were identified, with top routes situated around Central Park
- Top routes had high variance , denoting high density in paricular areas
- Manhattan accounted for more than half the trips, followed by Bronx.
- Over 450k entries have the same start and end stations.

3 Wait-Time Estimation

Our starting point was the textbook assumption that arrivals follow a Poisson process, meaning the hourly count N_t for a zone or station has mean and variance both equal to λ_h. This worked in the busiest taxi zones but failed for bikes: weather swings or rebalancing trucks often sent the variance far above the mean. We therefore relaxed the model to a Negative-Binomial, which keeps the same mean λ_h but adds a dispersion term so that quieter locations are no longer forced to look Poisson. The NB fit is purely about understanding how many pickups happen in each hour.

Once an arrival rate is available we need an approximate wait time. We cannot observe actual rider waits, so we use the *inter-arrival gap*—the time difference between two consecutive pickups—as a proxy. This is imperfect: it assumes riders show up uniformly at random and that bikes/taxis are immediately available once a previous customer leaves. In reality someone could arrive during a lull or a bike dock might be empty even though the last checkout just happened. Still, the proxy aligns well with intuition and gives us a measurable quantity.

Plotting the inter-arrival gaps revealed a clear exponential shape for both taxis and bikes (after pooling nearby docks into catchment areas). Figure 2 shows two representative cohorts in the *top row*: the empirical histogram (blue) decays at roughly the same rate as the exponential curve (green), while the Poisson and Negative-Binomial fits describe the arrival counts themselves. Because the exponential mean equals $1/\lambda$ we can turn each hourly arrival rate into a wait-time estimate with a single formula. These exponentials become the default wait model inside the Streamlit app, while the NB parameters remain in the cache for diagnostics and future improvements.

4 Travel-Time Estimation

4.1 Gamma cohorts

Travel time is inherently positive and skewed; the Gamma distribution is a natural parametric family on $(0, \infty)$ with shape-scale parameters determined by the first two moments. By binning trips into distance slices (2 km wide) and rush/weekend strata we obtain cohorts that are approximately homoscedastic, so the Gamma assumption provides a quick descriptive summary: the mean captures central tendency, while the shape parameter indicates how peaked vs. heavy-tailed the cohort is. This is conceptually analogous to fitting an Erlang distribution for service times, which is widely used in queueing theory. Gamma cohorts therefore serve as a pedagogical baseline and a sanity check for the more flexible model.

4.2 Lognormal accelerated failure-time model

While Gamma bins are interpretable, they suffer from two theoretical limitations: (i) boundaries introduce discontinuities as a rider crosses 3.99 km vs. 4.01 km, and (ii) they ignore the smooth nonlinear effect of distance. A continuous accelerated failure-time (AFT) model addresses both shortcomings. The lognormal AFT assumption posits that $\log T$ is Gaussian, meaning T is positive with multiplicative noise—precisely the pattern seen in empirical travel-time distributions where relative errors remain roughly constant across trip lengths. The quadratic distance terms (d and d^2) allow the mean log-time to curve, capturing the diminishing marginal penalty of distance once cruising speed stabilizes. Binary covariates for rush hours, weekends, and e-bike usage enter additively in log space, so they scale trip times multiplicatively in the original units (e.g., $\beta_{\text{ebike}} < 0$ implies a constant percentage speed-up for e-bikes regardless of distance). Once coefficients are learned via maximum likelihood (equivalent to OLS on $\log T$), the expected travel time is $\mathbb{E}[T] = \exp(\mu + \sigma^2/2)$, inheriting the full lognormal distribution for uncertainty quantification. Table 1 lists the fitted parameters and demonstrates that the model respects known physics: distances dominate, rush hours inflate taxi durations, and e-bikes enjoy a sizable negative shift.

Table 1: Lognormal GLM parameters extracted from `travel_lognormal_glm.json`.

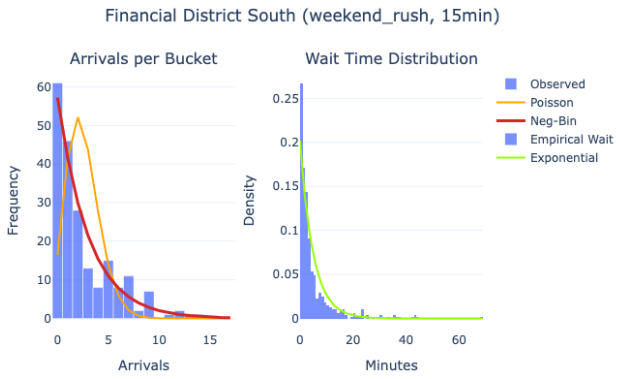
Mode	σ	β_d	β_{d^2}	β_{rush}	β_{ebike}
Taxi	0.391	0.499	−0.031	0.059	0.000
Bike	0.485	0.702	−0.049	−0.025	−0.193

Quality metrics (log-space $R^2 \approx 0.62$ and MAE$\approx$ 0.33 for bikes) indicate the quadratic distance term handles most variance, while rush/weekend dummies provide smaller adjustments. When the GLM fails (e.g., long-distance inputs outside the 12 km cap) the system falls back to the Gamma bin mean, and finally to a heuristic constant-speed estimate. The bottom row of Fig. 2 shows how the GLM smooths the Gamma histogram for taxi and Citi Bike 0–2 km rush-hour cohorts without losing interpretability.

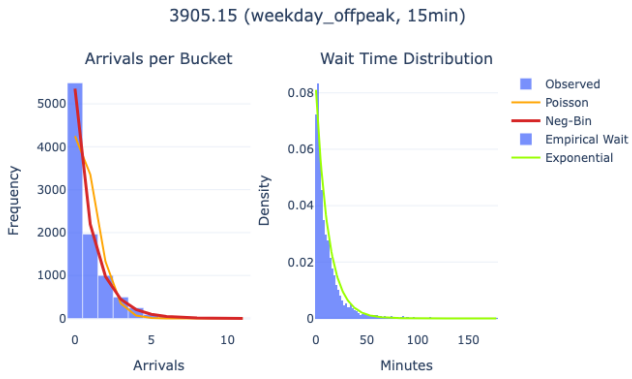
5 Streamlit Dashboard

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

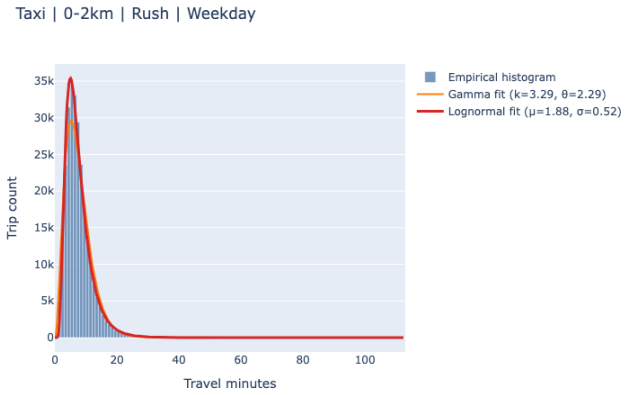
- Pick origin/destination on an interactive map (Folium).
- View taxi vs bike wait times for the selected locations, with automatic fallbacks to the nearest station if the exact cohort is missing.
- See travel-time estimates using the lognormal GLM (with Gamma fallback).



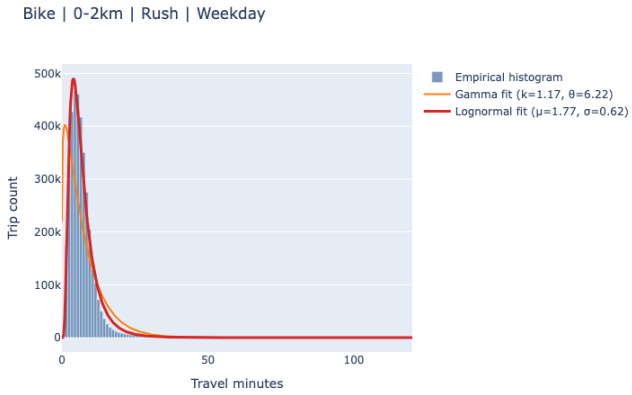
Taxi wait: zone 3905.15, weekday off-peak



Citi Bike wait: Financial District, weekend rush



Taxi travel: 0–2 km, rush-hour weekday



Citi Bike travel: 0–2 km, rush-hour weekday

Figure 2: Model diagnostics spanning the two modeling tasks: top row shows arrival/wait fits (Section 3), bottom row shows travel-time Gamma vs lognormal fits (Section 4).

Figure 3: Streamlit UI mock/screenshot placeholder.

- 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/e-bike.
- **Distance calculation:** Haversine distance does not account to the roads and turns involved in reaching the end location
- **Data coverage:** Only Jan–Jun 2024; Citi Bike GPS routes or real-time inventory (docks per station) are not modeled.
- **Costs:** Trip fares are not surfaced; taxi Parquet files contain the needed columns but Citi Bike pricing must be inferred from membership tiers.
- **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 Discussion

Speed modelling was the other approach used to estimate travel time. Similar to earlier methods, the data were grouped according to rush hours and weekends. However, due to the limited number of entries for many unique routes, this grouping scheme was not able to properly distinguish between different routes. Furthermore, across all four categorizations, the observations were heavily concentrated around the mean, leading to underdispersion. As a result, Poisson and negative binomial distributions were not suitable for capturing the underlying variability in travel times. Instead, a Weibull distribution was adopted, as it provided a higher AIC score compared to the gamma and lognormal alternatives.

In contrast to the taxi dataset, not all entries in the Citi Bike data contributed meaningfully to modelling New York's traffic conditions. More than 450,000 entries had identical start and end stations. Many of these trips were concentrated around peak tourist locations, such as Central Park 6 Ave, where over 12% of rides began and ended at the same station. While such usage patterns are reasonable in practice—for example, casual tourist rides that start and end at the same landmark—they reduce the effective amount of information available for our traffic-focused modelling, since the perceived volume of data is much larger than the subset that is actually informative for our purposes.

Arrival-count modelling for both datasets also did not conform well to a Poisson distribution. A likely explanation is the dependence of arrival counts on time-varying and context-specific factors such as time of day, day of the week, public holidays, and city events. These factors introduce strong temporal patterns and clustering in arrivals, violating the Poisson assumptions of independence and a constant arrival rate over the counting interval. While our datasets did not include explicit information on these finer-grained covariates, incorporating them would have required a more complex model and substantially more data than were available. For instance, in the Citi Bike dataset, although the total number of rows was large, the number of rides within a specific hour on a given weekday (e.g., a Wednesday) was in the single digits for many stations in cases, limiting our ability to reliably model hourly arrival patterns.

8 Conclusion

We delivered a reproducible workflow: raw data ingestion, exploratory notebooks, lognormal travel-time modeling with e-bike adjustments, cached wait-time summaries, and a Streamlit app that gracefully degrades to nearby stations. The lognormal GLM improves Citi Bike estimates substantially over Gamma bins while keeping the pipeline explainable. Next steps include richer covariates, explicit cost modeling, 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