Navigating Complexity: Analyzing Passenger Traffic and System Performance in the NYC Subway

Andrew Chung, hc893

BTRY 4100 – Final Project

Dr. Dana Yang

Introduction

As the largest and busiest of its kind in the U.S. and the Western Hemisphere, the **New York City subway system**, operated by the Metropolitan Transportation Authority (MTA), is unparalleled in its scale of service, comprising 36 lines (28 services – previous and present), operating 472 stations, serving a city of 8 million, and boasting a daily ridership of 3.6 million passengers.

Its vast scope of operations poses **challenges** towards optimizing transportation systems, mitigating delays, improving service efficiency/quality, and managing large-scale infrastructure. Understanding tendencies in **passenger traffic** and its relationship with **systemic performance features** is vital for sensible MTA policies that help maintain such an intricate transit system.

Research Question

I seek to identify and address key predictors of ridership trends in the NYC Subway system during **weekday peak hours in January-February 2025** by developing a **linear regression model** to quantify the relationship between **service performance features** and **subway ridership**, accompanying it also a potential for future trend prediction.

Data Acquisition

All data files used for this project were sourced from the New York State's Open Data Portal API, accessible at data.ny.gov. Eight (8) data sets were acquired, all up to date effective March 24, 2025:

1. MTA Subway Hourly Ridership, Beginning 2025
2. Customer Journey-Focused Metrics, Wait Assessment, Service Delivered
3. Terminal On-Time Performance, 4-5 Minute Late Arrivals
4. Major Incidents Log, Train Delays Log

The Hourly Ridership Data set tracks the **volume of passengers** entering the turnstiles or completing a revenue transaction throughout all public transit points (subway stations, bus stops, tram stops) or transportation complexes (railroad terminals, bus terminals, ferry ports) throughout New York City's five boroughs on an **hourly** basis. All possible permutations of unique station identifiers (ID) and date-time representations of hourly timestamps since January 1, 2025 are stored as unique data points (rows).

All other data sets concern the **relative performances** of each subway line, whose rows, representing each of the 24 lines in the system, comprise the month, division (A/B, not useful), line ID, day type (weekday vs. weekend), and the pertinent performance metric(s).

Through a rigorous data cleaning procedure, I extracted information on **22 subway lines and 424 viable subway stations** – compiling a matrix of **11 distinct performance metrics** and an aggregate station-level

log of monthly ridership, ultimately integrating them into a single data set compatible with my regression modeling objective.

Data Cleaning

Though publicly compiled and vetted, my acquired files were certainly not without their "flaws." Through a rigorous data cleaning process I sought to reduce raw performance files into a single matrix that concisely captures all available performance metrics for each of the 22 viable lines. Wrangling the hourly ridership log into a monthly summary of station ridership was more involved due to the constraints imposed by my research question and the size of the data set.

For **performance data** files,

1. I first removed any lines that were either (a) defunct, or (b) smaller, non-independent lines (e.g. the Rockaway Shuttle).
2. I dropped unimportant features such as division and day type (the scope of the project is confined to weekdays).
3. Noticing slight discrepancies in line ID, such as the separation of J and Z lines (which run in pairs) and the naming of shuttle services, I manually compared the distinct identifiers and created simplified encodings (e.g. "S Fkln" = "SF", "J" and "Z" = "J/Z").
4. Once all files were formatted under a unified set of 22 subway lines as indexes, I merged the seven datasets under the levels of Line ID (22) and Month (1, 2) such that the final matrix would comprise 44 rows (22 lines × 2 months) and 11 columns.

Major Incidents and **4-5 minutes Train Delays** data sets were organized not in percentages or ratios but rather as counts of distinct categories of incidents, ranging from infrastructural ones like track maintenance, brake activation, staffing/station issues, or signal malfunctions, to non-infrastructural ones including medical emergencies or law enforcement intervention. I extracted **4 incident count features** – critical infrastructure, critical non-infrastructure, non-critical infrastructure, and non-critical non-infrastructure, depending on the source dataset and nature of the event.

I thus collected **11 performance-related metrics** to be incorporated as covariates for my linear model:

1. Monthly Passenger Volume by Line
2. Additional Wait (Platform) Time, Additional Train Time
3. Wait Assessment – measures how regularly trains are spaced during peak hours
4. % Over 5 Minutes, % Service Delivered, % Late, % Terminal On-Time Performance
5. Critical/Non-Critical Infrastructure/Non-Infrastructure Incident Counts

For the **hourly ridership** dataset,

1. I first removed nodes that didn't correspond to subway stations, that is, bus/tram stops, standalone train terminals, and ferry ports
2. Then, I used the datetime encodings to filter out hour timestamps corresponding to:
    a. Off-Peak Hours *****
    b. Weekend Dates
    c. Holidays and Special Occasions (e.g. New Year's, MLK Day, Presidents' Day)
    d. ***** The MTA defines **Peak Hours** as 06:30 AM-09:30 AM and 03:30 PM-08:00 PM, during which trains run at increased frequencies (and express services) to accommodate heightened traffic.

3. I aggregated hourly ridership figures into monthly numbers by station ID and month, reducing millions of data points into ~800 rows.
4. Each station ID contained the set of lines serving it in the format "station_name (*lines)", which I extracted through the use of regular expressions – Regex (e.g. "36 St (D,N,R)" = "D,N,R").

The compiled ridership data set thus encompasses rows containing each unique station ID, month, ridership figures (distinguished by January vs. February), borough, and lines serving (encoded in comma-separated strings).

Data Integration

With the ultimate objective of building a **regression model**, I proceeded to **integrate** the two datasets obtained through the data cleaning process. Initially, the discrepancy in the structural nature of the two datasets proved it a challenging task; I thus chose to leverage the line sets contained in each station as a bridge for this task.

1. In the ridership dataset, I identified all unique existing **combination of lines** (e.g. [1], [4,5,6], [E,J/Z], [6,E,M], [2,3,4,5,B,D,N,Q,R]) to roughly represent distinct line segments (e.g. Grand Concourse Line, Nostrand Avenue Line) or unique train terminals.
2. Within each cluster, the monthly ridership figures across all constituent stations were **averaged** to represent the mean traffic across a specific line segment.
3. Once all line segments were identified, I imputed the 11 performance metrics for each segment:
    a. For single-line segments (e.g. Canarsie Line [L]), the set of metrics for the unique line was accepted.
    b. For multi-line segments, the metrics were taken as a **weighted sum** of lines making up the segment, with weights computed according to relative line ridership. For example, for a line segment [A,B], with w_A = 0.2 and w_B = 0.1 (i.e. A ridership is twice that of B), the segment-wise metrics would be summed with respect to the weights 0.2 and 0.1.
4. As always, ridership figures and aggregate metrics would be separated by month.

The final data set contains **184 rows** (92 distinct line segments × 2 months) and **11 columns**.

Exploratory Data Analysis (EDA)

EDA of Station Ridership Data

First and foremost, I analyzed the station-level ridership data set. Each row represents the monthly ridership levels tallied at each station in the New York City subway for the months of January and February. To optimize for predictive accuracy in my linear model, I systematically eliminated stations that saw less than 1,000 monthly passengers, yielding 835 rows and 5 columns.

I first plotted a **histogram of station-level ridership** (Fig 1), which was heavily **right-skewed**, which is logical considering the presence of a select few stations (e.g. Times Square, Penn Station, Grand Central) in touristy or otherwise hectic spots in the city that drive up subway traffic. Next, I attempted a **log-transformation** on the response (Fig 1), which gave it a much more **symmetrical** and approximately normal distribution. (Though discussed in greater depth later, the log transformation of the response would prove vital for the validity of my linear model.)

Additionally, density plots corresponding to January and February riderships (log-transformed) were superimposed (Fig 1); barring a miniscule leftward shift (decline) in overall traffic, no significant deviations were detected in usage patterns, further confirmed by the near-perfect linear shape in a January vs. February by-Station scatter plot (Fig 2).

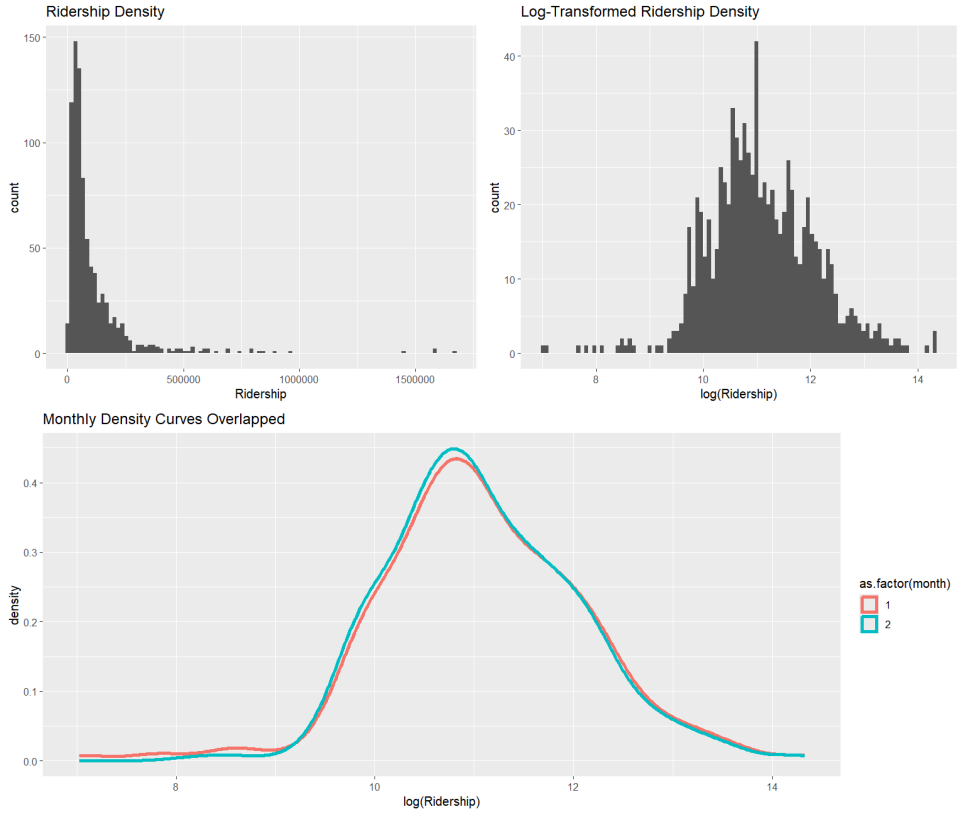

Figure 1. Histogram of Monthly Ridership Before and After Log Transformation; Jan vs. Feb Ridership

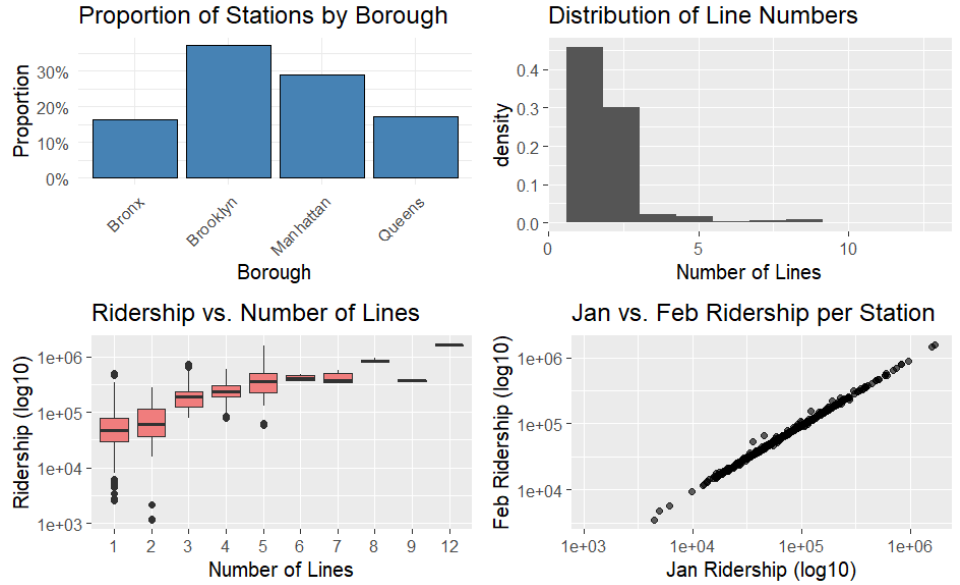

Figure 2. Proportion of Stations by Borough; Distribution of # of Lines Serving Stations; Ridership Box Plots Blocked by # of Lines Served; January vs. February Ridership by Station Scatter Plot

Further exploration of the station data yielded conventional yet interesting results. Despite the subway's reputation among non-New Yorkers and intricacy of service in Manhattan, Brooklyn is home to the highest proportion of subway stations (Fig 2), reflecting its status as the most populous borough. Manhattan comes in second thanks to its population density, economic/commercial activity, and tourism; Queens and the Bronx, owing to their suburban nature, see lower levels of train service. Next, the distribution of line density (# of lines serving a station) and their respective ridership levels closely parallel the trends seen in Figure 1 — representing both the high prevalence of isolated one-line stations (serving mostly residential neighborhoods) and the "outlier" major transfer hubs.

EDA of Line Performance Data

Next, I sought to analyze the **individual predictors** (performance metrics). Primarily, I discovered that critical incidents happen at much smaller frequencies compared to non-critical incidents (several lines did not suffer from any), which inspired me to merge the critical and non-critical incidents (4 columns) into infrastructural vs. non-infrastructural incidents (2 columns).

Figure 3 illustrates the distributions of the three (3) categories of covariates, distinguished by measurement type (Raw Count, Time (min), Ratio (%)).

- **Additional times** spent on the platform waiting or on board a train were largely symmetrically distributed around 0-1 minutes (trains occasionally skip select stops or proceed at faster speeds, hence the negative values).
- **Ratio (percent)-based metrics** show similar patterns, except with indicators of smooth service (Terminal Performance, Wait Assessment, Service Delivered) concentrated above 0.6 and those of inconsistency (% over 5 minutes delayed, % Late) clustered close to 0.0, a testimony to the remarkable efficiency and precision underlying such a complex subway system.
- **Incident counts (critical + noncritical)** are no different; as raw counts, infrastructural incidents occur to be slightly more frequent, as it encompasses a broader range of incidents directly relevant to service fulfillment.

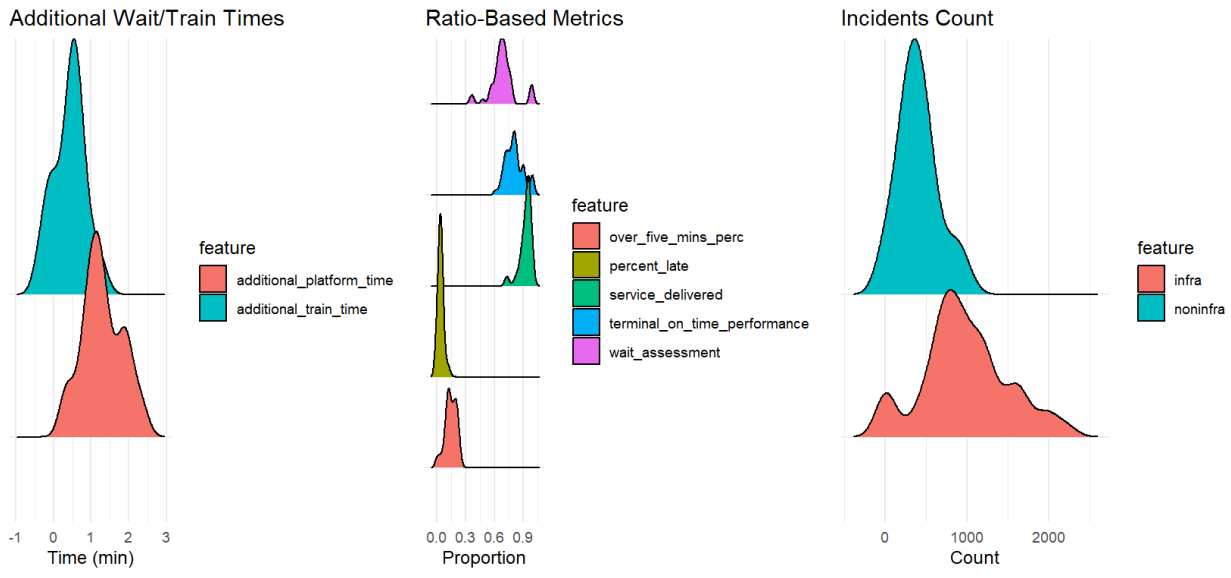

Figure 3. Ridgeline Plots that Represent the Densities of Distinct Types of Performance Metrics, Paneled by Measurement Type

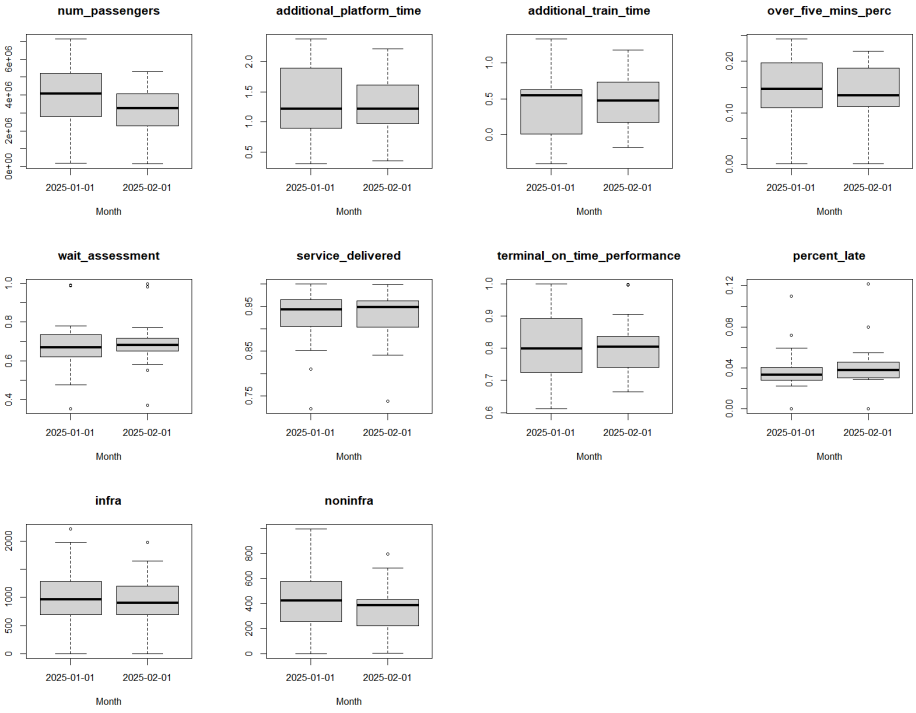

Figure 4. Month-by-Month Boxplot Comparison of Nine Performance Metrics as well as Line Usage (num_passengers)

The distribution of performance metrics and line usage are additionally **juxtaposed by month** (January vs. February) (Fig 4); though most metrics are similarly centered and distributed (often, sharing common outliers) subtle shifts are evident throughout as well. In addition to the increased availability of data points, these shifts signal the benefit of segmenting the regression data by month, potentially enabling the model to capture hard-to-notice, month wise pattern shifts.

Though not documented in this report, the EDA code file (RMD) also contains bar charts ranking lines by ridership — analogous to the weights assigned to each line during the data integration step — as well as incident counts, documenting lines that (1) cater to high passenger volumes and (2) are prone to infrastructural and non-infrastructural defects.

Particularly noticeable was the observed **multicollinearity between predictors**, exhibited by several high pairwise correlation values (Fig 5). Though many predictor pairs showed modest or occasionally low Pearson's correlation coefficients ($|r| \leq 0.6$), select pairs, such as % Over 5 Minutes/Terminal Performance (-0.742) and Wait Assessment/Service Delivered (0.883) demonstrated unusually high covariance, further corroborated by **linear patterns** embedded in corresponding paired scatterplots. I have largely attributed this phenomena to the following (with pertinent examples):

- **Measurement of Similar Attributes** — Both service delivered and wait assessment examine line adherence to posted train schedules.
- **Complementary Nature** — % Over 5 minutes is essentially a subset of % late trains, distinguished primarily by the magnitude of the delay (>5 min vs. >0 min).
- **Non-Independent (Causal) Events** — Accrued delay minutes (additional platform time) can not only translate directly to late (over 5 minutes) train arrivals but affect downstream services as well (amplification).

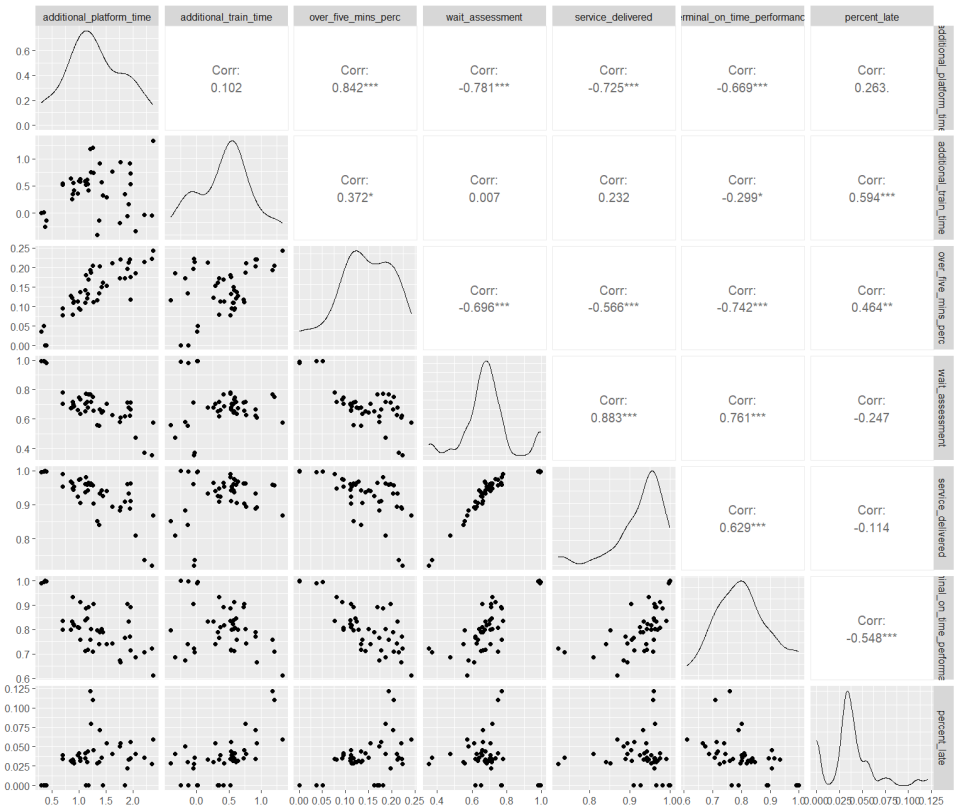

Figure 5. Pair Plot Illustrating the Covariance Levels Between Continuous Predictors

Acknowledging the presence of multicollinearity is a critical precursor to linear regression modeling, as it **inflates the variances** of the estimated regression coefficients, hinders the **stability/validity** of standard errors, p-values, and confidence intervals, reinforces **sensitivity** to small changes in the data, and corrodes **model interpretability**. (The regression dataset additionally takes weighted sums (linear combinations) of the above predictors, amplifying predictor linearity especially among similar data points.) Later, I take diagnostic measures such as principal component analysis (PCA) and variable transformation to remedy these effects.

Linear Regression Modeling

In this section, I implement a **multiple linear regression model** to characterize subway ridership behavior as a function of line performance. The regression data set was encoded to contain **184 rows** that represent unique segments within the subway system as line combinations, each containing the mean ridership across all constituent stations, as well as **11 imported performance covariates**. (That is, each unique station is represented by the line segment it belongs to, dependent on the lines serving it.) Before fitting a linear model, I undertook the following pre-processing steps:

1. Merged Incident Counts Under Infrastructural/Non-Infrastructural Tallies
2. Standardized Ridership Figures and Predictors by **Line Count** — the original encoding of the dataset incurs a huge risk of multicollinearity due to the proportional nature between ridership/aggregated metrics and line count (e.g. line segments with 3 lines correspond to **roughly** 3 times as many passengers and 3 times the magnitude of predictors as a 1-line segment), so

standardizing it was key to ensure a reliable linear model. To capture the effect of running multiple lines, I instead encoded the number of lines (n_lines) as an additional predictor.

3. **Re-configured** metrics — "Timeliness" = 1 - % Late, "% Under 5 mins" = 1 - % Over 5 mins

I additionally re-plotted the distribution of ridership levels before and after log transformation, which remained largely analogous to those in Figure 1.

Ordinary Least Squares (OLS) Linear Model

Initially, I ran an **OLS regression model** with the 11 performance features as covariates and the raw ridership figures of each station-cluster as the response. A glance of its characteristic plots (Fig 6), however, immediately hinted towards glaring issues detracting from the model's validity.

- Though not overt, a **slight convexity** is visible in the **residual scatter plot's** LOESS line, suggesting that the predictor-response relationship may not be strictly linear.
- Heavy, pronounced **upward tails** on the **normal Q-Q plot** strongly indicate that the assumption of residual normality is violated; t-tests, p-values, and confidence intervals are thus not reliable.
- An increase in residual variance at higher fitted values, though subtle and confined to a few outliers.

Furthermore, the **Box-Cox plot** (Regression Code file) yielded an optimal lambda of $\lambda = 0$ with a narrow 95% confidence interval, signaling that a **log-transformation of the response** (ridership) is most conducive towards normalizing the data/residuals.

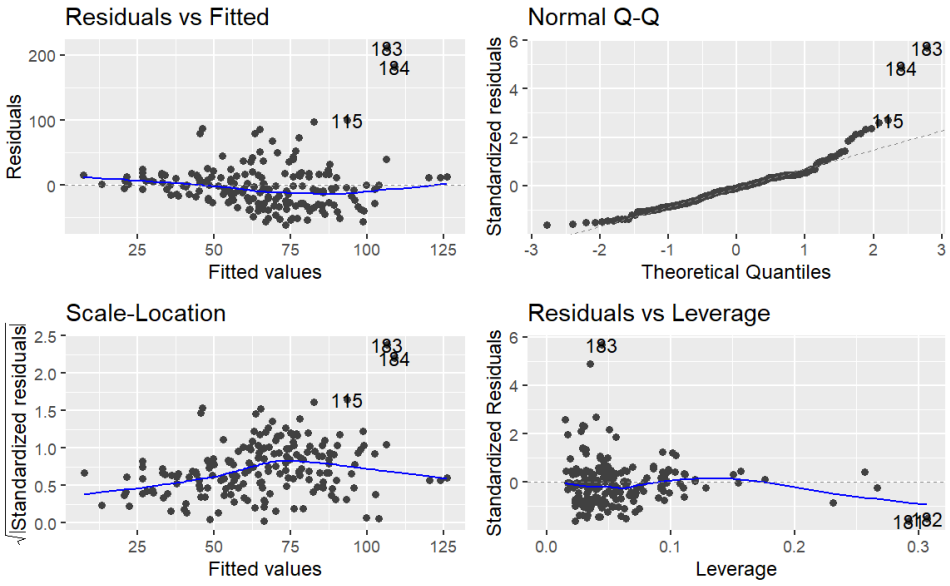

Figure 6. Characteristic Plots of the OLS Model

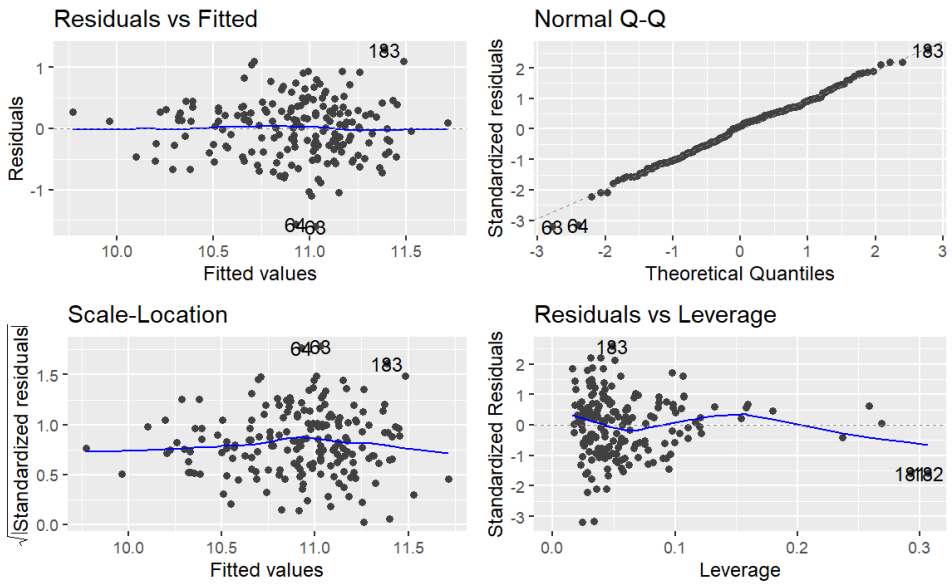

Figure 7. Characteristic Plots of the Log-Transformed OLS Model

OLS Linear Model with Log-Transformed Response

Earlier, I discovered that the log-transformed response had a much more symmetrical, approximately normal distribution than the raw response, suggesting that the log transform could successfully draw the residuals closer to meeting the OLS assumption of normality and introduce a truly linear model fit.

Previously, in standardizing the predictor space and ridership figures, I opted to regress ridership levels and service quality on a **per-line basis** (e.g. a 3-line station is treated as three equal units when assessing average traffic and quality) thereby isolating pure service-quality effects. This helps break the spurious collinearity arising from a simple summation of metrics (conflating the number of lines with higher ridership and thus service excellence). Consistent with that objective, I **log-transformed the line count predictor (n_line)** as a mechanism to capture the multiplicative scale effect of adding lines (more in the code file), as well as the infrastructural/non-infrastructural **incident counts** under a similar logic.

To that end, I trained a new linear regression model with log-transformations of the response and a portion of the predictors. Examining the characteristic plots (Fig 7) in a similar fashion demonstrates:

- An essentially flat residual scatter LOESS line, giving faith to a **linear relationship**.
- Aside from small far-tail departures, the normal Q-Q plot closely hugs the 45° line, strongly suggesting that **the residual normality assumption is fulfilled**.
    - This is further corroborated by the Shapiro-Wilk test ($p = 5.488 \times 10^{-12} \rightarrow p = 0.3332$) (Regression Code file).
- A reduced (but nonetheless eminent) heteroskedastic effect on the residual variance, as well as a more **robust fit** against high-leverage points.

Compared to the original model, therefore, the log-transformation step helps restore approximate linearity, constant variance, and near-normal errors, uplifting the validity of related statistical tests and inference.

Partial Principal Component Analysis (PCA)

Although log-transformed regression vastly improved the legitimacy of my linear model, the inherent issue of **predictor multicollinearity** persisted to a degree.

Specifically, three predictors — wait assessment (23.864338), service delivered (47.415712), and terminal performance (13.283) — had VIF > 10 and were flagged for severe multicollinearity. More surprisingly, a paired correlation matrix showed **near-perfect correlation** between the three, prompting me to perform **principal component analysis (PCA)** only on the three variables; seeing that the first principal component (PC1) explained 99.7% of the variance, I concluded that replacing the three predictors with PC1 would largely crush the latent collinearity in the predictor space and simplify the model with virtually no loss of information. (Regression Code file)

Having aptly named the new variable "Reliability PC" — wait assessment (regular train spacing), % service delivered, and terminal on-time performance measure how reliably trains run on schedule at regular intervals — I re-fitted the linear model, achieving lower VIF values across the board as well as a modest drop in $adj(R^2)$, a small cost for model parsimony.

Addition of Interaction Term

Lastly, I tested **paired interaction terms** to capture how the impact of one service metric might **vary** depending on another (if at all), ensuring the model reflects real-world drivers of ridership sensitivity rather than assuming uniform effects across all lines.

To begin, I identified every possible interaction pair from the 8-dimensional predictor space (including the newly added PC) and, for each of the $\binom{8}{2} = 28$ pairs, compared the full interaction model with the base model (log-transform + PC) using ANOVA. Having compiled the p-values and newly resulting $adj(R^2)$ values, I identified 4 viable pairs, and for the sake of model parsimony, settled on a single interaction term (**additional_train_time:reliability_PC**; $p = 0.005332992$) based on the p-value and increase in $adj(R^2)$.

Statistical Inference of the Completed Linear Model

The complete final model takes the form

$$\begin{aligned}\log(\text{ridership})_i = {} & \beta_0 + \beta_1 \log(\text{n.lines})_i + \\ & \beta_2(\text{Additional Platform Time})_i + \\ & \beta_3(\text{Additional Train Time})_i + \\ & \beta_4(\%\ \text{Under 5 Mins})_i + \\ & \beta_5(\text{Timeliness})_i + \\ & \beta_6 \log(\text{Infrastructural Incident})_i + \\ & \beta_7 \log(\text{Non-Infrastructural Incident})_i + \\ & \beta_8(\text{Reliability PC})_i + \\ & \beta_9(\text{Additional Train Time} \times \text{Reliability PC})_i + \epsilon_i\end{aligned}$$

with a summary of the fitted model below.

```
Call:
lm(formula = ridership ~ . + additional_train_time:reliability_PC,
    data = data.log.pc)

Residuals:
     Min       1Q   Median       3Q      Max
-1.58167 -0.34272  0.01105  0.30272  1.34146

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                          2.75522    3.85175   0.715 0.475374
additional_platform_time             0.62559    0.23368   2.677 0.008137 **
additional_train_time               -4.70853    1.66831  -2.822 0.005323 **
under_five_mins_perc                 8.76396    2.38765   3.671 0.000322 ***
timeliness                          -2.34391    3.19395  -0.734 0.464021
infra                                0.28246    0.14931   1.892 0.060176 .
noninfra                             0.01586    0.11547   0.137 0.890902
n_lines                              0.29246    0.06819   4.289 2.97e-05 ***
reliability_PC                       0.12196    0.27407   0.445 0.656873
additional_train_time:reliability_PC -2.64694    0.93807  -2.822 0.005333 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5089 on 174 degrees of freedom
Multiple R-squared:  0.2999,    Adjusted R-squared:  0.2637
F-statistic: 8.282 on 9 and 174 DF,  p-value: 3.165e-10
```

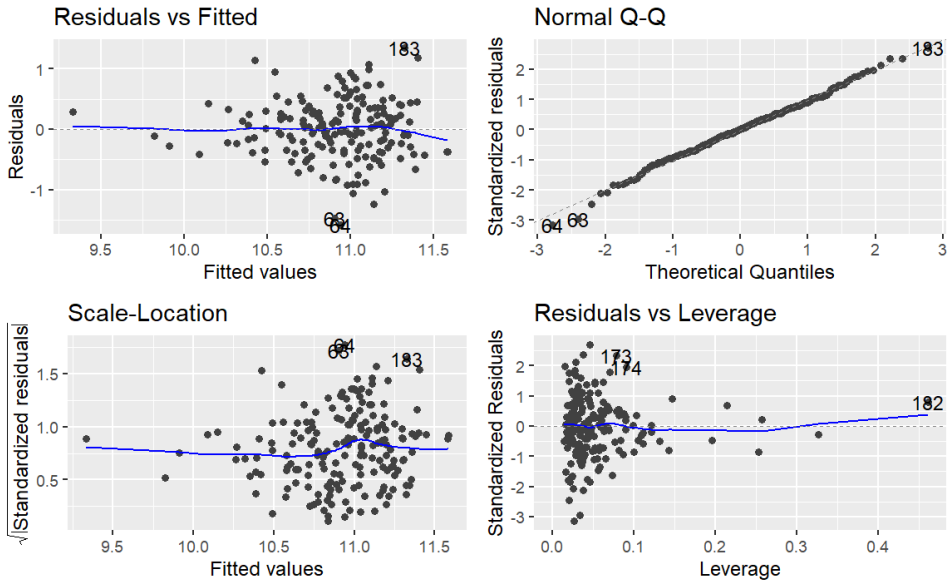

Figure 8. Summary of Final OLS Linear Model and Characteristic Plots

Initially, the p-value of 3.165×10^{-10} signals that the model as a whole is **useful** in predicting $log(ridership)$. Among individual predictors, the following were found to be significant:

- **Additional Platform Time** ($p = 0.008137$)
- **Additional Train Time** ($p = 0.005323$)
- **% Under 5 Mins** ($p = 0.000322$)
- **Number of Lines** ($p = 2.97 \times 10^{-5}$)
- **Additional Train Time × Reliability PC** ($p = 0.005333$)

Additional Model Insights (Fig 8):

- The **Shapiro-Wilk Test** of the model ($p = 0.6835$) verifies the unwavering adherence of the model to the assumption of **residual normality**.

- The highest **Cook's Distance** among high-influence points is 0.05221665, a testament to the general **robustness of the model against extreme outliers** (e.g. Times Square Station, served by 12 lines and sees millions of riders monthly).
- Some multicollinearity is still present, but at much more manageable levels.
- A consistent theme across all linear models trained is the rather miniscule values of R^2 and $adj(R^2)$. Though this does not have much bearing on the model fit or significance of our findings, it does highlight the intrinsic complexity underlying subway ridership patterns.
- **Heteroskedascity** of the residuals persists to some degree (Breusch-Pagan Test: $p = 0.026$); to correct for this I instituted **Heteroskedastic-Consistent (HC3) Robust Standard Errors**. [1]

To improve my **fidelity of the model** and optimize the validity of observed p-values, I implemented and compared the results of HC3 SE-adjusted parameter inference to its original (Fig 9).

	Estimate	Naive SE	t	Naive p	Robust SE	Robust t	Robust p
(Intercept)	2.755	3.852	0.715	0.475	3.934	0.700	0.485
additional_platform_time	0.626	0.234	2.677	0.008	0.225	2.785	0.006
additional_train_time	-4.709	1.668	-2.822	0.005	1.491	-3.158	0.002
under_five_mins_perc	8.764	2.388	3.671	0.000	2.513	3.488	0.001
timeliness	-2.344	3.194	-0.734	0.464	2.815	-0.833	0.406
infra	0.282	0.149	1.892	0.060	0.144	1.957	0.052
noninfra	0.016	0.115	0.137	0.891	0.114	0.139	0.889
n_lines	0.292	0.068	4.289	0.000	0.066	4.405	0.000
reliability_PC	0.122	0.274	0.445	0.657	0.223	0.546	0.586
additional_train_time:reliability_PC	-2.647	0.938	-2.822	0.005	0.835	-3.170	0.002

Figure 9. Naive SE vs. HC3 Robust SE-adjusted Inference Results

Evidently, significance levels across the board seem to have remained largely constant, with **all previously significant terms remaining highly significant**.

Overall, since the model verifies presumptive conditions for linearity (residual normality, lack of high-influence points, thoroughly mitigated multicollinearity), is supported by fit statistics (F-test), and the effects of significant predictors for key service-quality and network-scale levers remain stable under HC3-robust estimation, **I trust these p-values as reliable indicators of powerful, actionable drivers of peak-hour ridership**. **In addition,** the magnitude of each effect is both **statistically precise** (narrow robust confidence intervals) and **practically meaningful**. Taken together, this gives us both **confidence in the statistical validity** and **clarity on the policy levers** the MTA should pull first.

Below are **real-world interpretations** for significant service features under my linear model (consistent assumption for each covariate: all other predictors fixed), in terms of monthly, per-line station ridership:

- **log(Number of Lines)**: $\beta = 0.292$; a 1-unit increase in log(Number of Lines) will lead to an increase of log(Ridership) by 0.292. Roughly, a 1% increase in the number of lines of a station will lead to a 0.292% rise in per-line ridership.
- **Additional Platform Time**: $\beta = 0.626$; an additional minute of platform wait time is linked to the per-line ridership at that station being multiplied by $e^{0.626} \approx 1.87$-fold.

- **Additional Train Time**: $\beta = -4.709$; an additional minute spent on a train serving a station multiplies per-line ridership at that station by $e^{-4.709} \approx 0.009$-fold.
- **% Under 5 Mins**: $\beta = 8.764$; an additional 1% (pp) improvement in the average % of trains arriving within 5 minutes multiplies per-line ridership at a station by $e^{0.08764} \approx 1.0916$-fold (9.16% increase).
- **Additional Train Time × Reliability PC**: $\beta = -2.647$; a 1-SD increase in average reliability multiplies the multiplication factor of additional train time by $e^{-2.647} = 0.071$.

[1] Kranz, S. (2024) From Replications to Revelations: Heteroskedasticity-Robust Inference. https://mpra.ub.uni-muenchen.de/id/eprint/122724

Potential Real-World Implications on Sensible MTA Policy-Making, and Future Directions

My final OLS Linear Model reveals four high-leverage service levers for the MTA:

1. **On-Time Performance**: Our model elucidates that even a one-percentage-point gain in train timeliness drives a 9.16% increase in per-line train ridership.
    a. An effective transit policy could address current shortcomings by implementing real-time dispatch adjustments and targeted signal precision upgrades system-wise, especially on lines with the largest scheduling gaps.
2. **Travel Time Delay**: In a city where every minute, every second matters during peak hours, even subtle hikes in intercity travel time can prove fatal in customer retention; in traditionally reliable lines, this effect can be amplified due to high expectations.
    a. The MTA should keep maintenance, rapid-response teams on stand-by, and run monitoring on top-performing corridors to protect high-expectation riders.
3. **Platform Traffic Regulation**: The model implies a link between additional platform waiting times (min) and station traffic in general.
    a. To mitigate the effects of public congestion, policies can predict traffic flow, explore crowd-management protocols, and adjust service frequency as needed.
4. **Network and System Connectivity**: Though beyond the scope of grassroots policy-making, our model confirms a logical tendency that, even after controlling for line counts, multi-line stations and especially large transit hubs see higher volumes of per-line traffic.
    a. Prioritize new transfers and line extensions at major hubs to unlock compounded growth.

Raising subway ridership not only boosts farebox revenue and reduces per-rider operating costs, but also advances MTA's goals of sustainability and congestion relief, incentivizing greater public adoption of the subway (and mass transit in general). By concentrating on punctuality, delay reduction, strategic connectivity, and real-time crowd management and fostering sensible, data-driven policy making, the MTA can deliver a more reliable, efficient, and attractive transit system for millions of New Yorkers during peak hours.

Additionally, a pre-eminent limiting factor throughout my analysis has been the **limited range of predictive variables and service quality metrics** made available in the NYS Open Data Portal, which in turn limited the explanatory power of my model. To overcome this constraint, future work should integrate richer public data sources such as those pertaining to city planning, demographic/census analytics, or weather and special‑event schedules to capture the broader socioeconomic and environmental drivers of ridership. Incorporating these variables can not only raise the model's explanatory power but also allow the MTA to tailor service adjustments to neighborhood context, seasonal travel patterns, and fare‑sensitivity, thereby delivering more precise, equity-focused, and demand-responsive transit planning.

Contributors

Andrew Chung (hc893) was the sole contributor of the project.