SF 311 Service Requests — Project 2 Paper

Andrew Young

December 10, 2025

Introduction

Street and sidewalk cleanliness is a visible service that residents notice quickly, and San Francisco's 311 system records thousands of requests for pickup of items like furniture, loose garbage, overflowing cans, and other debris. These records offer a look into the responsiveness to these issues as each request has an open time, a close time, and descriptive fields that tell when and why the request was made. Understanding which factors are associated with longer closure times could help set expectations for residents and guide operational choices for the city.

In this project I analyzed 2018–2019 Street & Sidewalk Cleaning (SSC) requests from San Francisco's 311 open data portal in an effort to answer this question: Does the type of request, the day of week, and the time of day when a request is made help explain how long it takes to close said request? The broader context is that city agencies must efficiently use limited crews across multiple tasks that arrive at different times. Some categories, such as those involving large items, may be slower to resolve, and staffing patterns across weekdays and time of day may also further affect these resolution times. While many dashboards report counts of requests, fewer examine closure time as an outcome and how it varies across operationally meaningful predictors. That gap motivates this analysis.

To tackle this question, I used a multiple linear regression that includes three sets of categorical predictors, request type, weekday, and time of day, with the duration of the request from opened to closed being the response variable. The goal is not to build a forecasting tool, but to quantify possible existing patterns in this data and to communicate them with clear graphics.

My analysis found that the estimated differences in closure time brought on by request type, weekday, and time of day are small and imprecise relative to the overall variability in closure times. In simpler words, these features alone offer very limited explanatory power for how long a case remains open. This indicates that closure time delay patterns are not fully captured by these predictive variables alone, and that additional factors not captured here may have a much larger influence on response time, such as travel distance and crew staffing. The rest of the paper then proceeds as follows: the Data section describes variables and cleaning, Methods states the regression model and assumptions, Results presents estimates and the fitted plot, Discussion interprets findings and limitations, and Conclusion summarizes implications and next steps.

Data

The observational unit in this study is a single SSC service request recorded by San Francisco's 311 system during 2018 to 2019. Each row corresponds to one request with an "opened" timestamp, a "closed" timestamp, and descriptive fields.

Variables used:

Closure time (renamed to duration_hours) is the outcome. It is the number of hours between the request's close time and open time computed from the raw timestamps.

Request type (renamed to request_type) is the descriptive category for the service call, such as Loose Garbage, Human or Animal Waste, or Refrigerator. To stabilize estimates, the analysis keeps the ten most frequent categories and collapses all remaining levels into "Other".

Weekday (renamed to weekday) is the calendar day label (Mon–Sun) for when the request was opened.

Time of day (renamed to tod_bin) is derived from the opening timestamp by converting the given time to 24-hour time and binning into four mutually exclusive intervals: Night (0–5), Morning (6–11), Afternoon (12–17), and Evening (18–23).

Only duration_hours and the three explanatory variables enter the regression. Other raw fields in the 311 file, such as location coordinates, agency, and free text notes were left for potential extensions and checks but are not modeled here.

Collection and transformations:

The source is an administrative transaction log, not a sample survey. I standardized column names, parsed the AM/PM timestamps, and explicitly converted opening times to 24 hour clock times before binning into time of day categories. I dropped rows with missing timestamps, negative durations, or extremely long closure times (those exceeding 60 days) as the small handful of points would otherwise dominate linear fits and alter patterns interest. After these filters, I created the final analysis with the outcome and predictors described above.

Key characteristics:

The tables below summarize sample size, the distribution of closure times, and the marginal distribution of the predictors used in the model.

Table 1: Sample size and closure-time summary (hours) after cleaning.

n_requests	dur_min	dur_q1	dur_median	dur_q3	dur_max
4146	0.0013889	1.657014	4.217083	18.66771	1280.179

Table 2: Requests by request type (top 10 types kept, remaining lumped to 'Other').

request_type	n
Other Loose Garbage	1217
Human or Animal Waste	766
Furniture	488
Boxed or Bagged Items	485
City_garbage_can_overflowing	361
Mattress	227
Refrigerator	194
Electronics	127
Other	103
Glass	92
Needles	86

Table 3: Requests by weekday (day of opening).

weekday	n
Fri	577
Mon	766
Sat	414
Sun	473
Thu	599
Tue	657
Wed	660

Table 4: Requests by time of day (based on opening hour).

tod_bin	n
Afternoon (12–17)	1546
Evening (18–23)	498
Morning (6–11)	2009
Night (0–5)	93

SSC requests are diverse, some categories, such as those involving bulky items, require equipment or coordination, while others can be addressed quickly. The resulting closure times thus are right skewed with occasional long tails. Times are also operational timestamps, and thus are affected by reporting delays, daylight saving transitions, reopened tickets, and batched data entry. Collapsing request types to a top 10 and "Other" will stabilizes estimates but could mask small categories with unusual means. Time of day bins reflect open time only and do not capture travel delays between opening and first action. Because of this skew and the categorical predictors, I expected wide variability and modest explanatory power from a multiple linear regression model, thus I chose to focus on clear estimates with uncertainty intervals and diagnostics rather than prediction.

Methods

I study how request characteristics relate to case closure time with an ordinary least squares multiple linear regression.

Model and notation:

Let Y_i denote the closure time for request i.

Let Type_i be the request type

Let Day_i be the weekday the request opened

Let Tod_i be a four level time of day bin (Night 0–5, Morning 6–11, Afternoon 12–17, Evening 18–23).

Let β_0 be the mean closure time in hours for the baseline case, a Boxed or Bagged Items request opened at Night (0–5) on Sunday. For any non-baseline level, the coefficient $\beta_j^{(\text{type})}$, $\beta_k^{(\text{day})}$, or $\beta_m^{(\text{tod})}$ is the additive difference in mean hours relative to that baseline while holding the other factors fixed. In other words, a positive $\beta_j^{(\text{type})}$ means that request type t_j takes longer on average than Boxed or Bagged Items (with weekday and time-of-day held constant), while a negative value means it closes faster. The same logic applies to weekday effects relative to Sunday and time-of-day effects relative to Night (0–5). Because all predictors are categorical and enter additively, there is no single "slope per unit", each reported coefficient is a difference in means in hours from the reference level.

With all of these, the model is:

$$Y_i = \beta_0 + \sum_{j=1}^{J-1} \beta_j^{(\text{type})} \{\text{Type}_i = t_j\} + \sum_{k=1}^{K-1} \beta_k^{(\text{day})} \{\text{Day}_i = d_k\} + \sum_{m=1}^{M-1} \beta_m^{(\text{tod})} \{\text{Tod}_i = \tau_m\} + \varepsilon_i, \quad \varepsilon_i \sim (0, \sigma^2).$$

Justification and modeling decisions:

I use a multiple-linear regression because it cleanly answered the question: do request type, weekday, and time of day help explain differences in closure time? All three are categorical and map naturally to how work is organized. Keeping the model additive with no interactions keeps interpretation straightforward and avoids very thin cells that would appear if I fully crossed three factor sets.

On the data side, I made a few practical choices to stabilize the estimates and keep the story readable. I parsed the open/close timestamps, converted opening times to a 24-hour clock, and binned them into four time of day groups (Night, Morning, Afternoon, Evening). I filtered out impossible or overly long durations (negative or >60 days) so a handful of extreme records wouldn't dominate the fit. For request types, I kept the ten most common levels and lumped the rest to "Other," which trades a bit of specification for less noisy coefficients. Overall, the goal is explanatory rather than predictive, to quantify average differences in closure time associated with each factor and communicate those patterns clearly.

Assumptions:

The standard linear-model assumptions apply.

Linearity in the mean, meaning the expected closure time changes linearly with the predictors. With categorical predictors, this means each level shifts the mean by a constant amount relative to its reference.

Independence, meaning requests are treated as independent observational units.

Homoskedasticity, meaning the error variance is constant across fitted values and groups, so the spread of errors does not systematically increase or decrease with the predictors.

Normality of errors, meaning the unobserved error terms ε_i are assumed $\sim \mathcal{N}(0, \sigma^2)$, which justifies f-tests, t-tests, and confidence intervals.

Diagnostics and validation steps

I assessed assumptions using two diagnostic plots. The Residuals vs Fitted plot is used to check linearity of the mean and equality of variance, patterns or increasing spread would signal problems with the linear model or homoskedasticity. The Normal Q–Q plot evaluates tail behavior and the plausibility of normal errors. Taken together, these plots provide a quick screen for violations that could affect inference.

Software and implementation:

All analyses are implemented in R. I use readr and janitor for import/cleaning, lubridate and stringr for time parsing, dplyr/forcats for wrangling and factor handling, stats::lm for estimation, broom for tidy summaries, and ggplot2 for coefficient and diagnostic plots.

Results and Discussion

Model fit and key estimates:

I fit an OLS multiple linear regression with indicator variables for each non-reference level of request type, weekday, and time of day. Because the predictors are categorical, each coefficient represents a difference in mean hours relative to a baseline level.

Global test:

The primary hypothesis asks whether these factors, taken together, improve fit over an intercept-only model:

$$H_0 : \ \beta_j^{(\text{type})} = \beta_k^{(\text{day})} = \beta_m^{(\text{tod})} = 0 \text{ for all } j, k, m \quad \text{vs} \quad H_A : \ \text{at least one } \beta \neq 0.$$

This is evaluated by the model's F-test. In our fit, $F \approx 1.17$ with $p \approx 0.27$, so I fail to reject H_0. Interpreted plainly: with these predictors and a linear mean, the model does not explain more variability than an intercept alone.

Per-coefficient tests:

Each dot in the coefficient plots corresponds to a t-test of

$$H_0 : \ \beta = 0 \quad \text{vs} \quad H_A : \ \beta \neq 0,$$

with a 95% confidence interval shown as whiskers. Intervals crossing 0 indicate results not statistically distinguishable from the reference at $\alpha = 0.05$. Given many comparisons, heavy tailed residuals, and very small R^2, isolated small p-values were interpreted cautiously, and effect size and precision are the focus.

After fitting an OLS multiple linear regression of closure time in hours on request type, weekday, and time of day, the overall fit is very little. From `summary(lm_ext)`, the residual standard error is about 71.1 hours on 4126 df. The model explains very little variation, with an $R^2 \approx 0.005$, Adj. $R^2 \approx 0.001$, and the overall F-test that all non-intercept coefficients are zero is not significant, with p-value = 0.27. In short, with these three predictors and a linear mean, most variation in time to close remains unexplained.

Plain-English takeaway.

Estimated mean differences associated with request category, weekday, and time-of-day are small and imprecise relative to the overall spread in closure times.

Coefficient plots:

Below I visualize the estimated mean differences in hours relative to the reference levels used in the model. Positive values indicate longer average time to close than the baseline; negative values indicate faster closure.

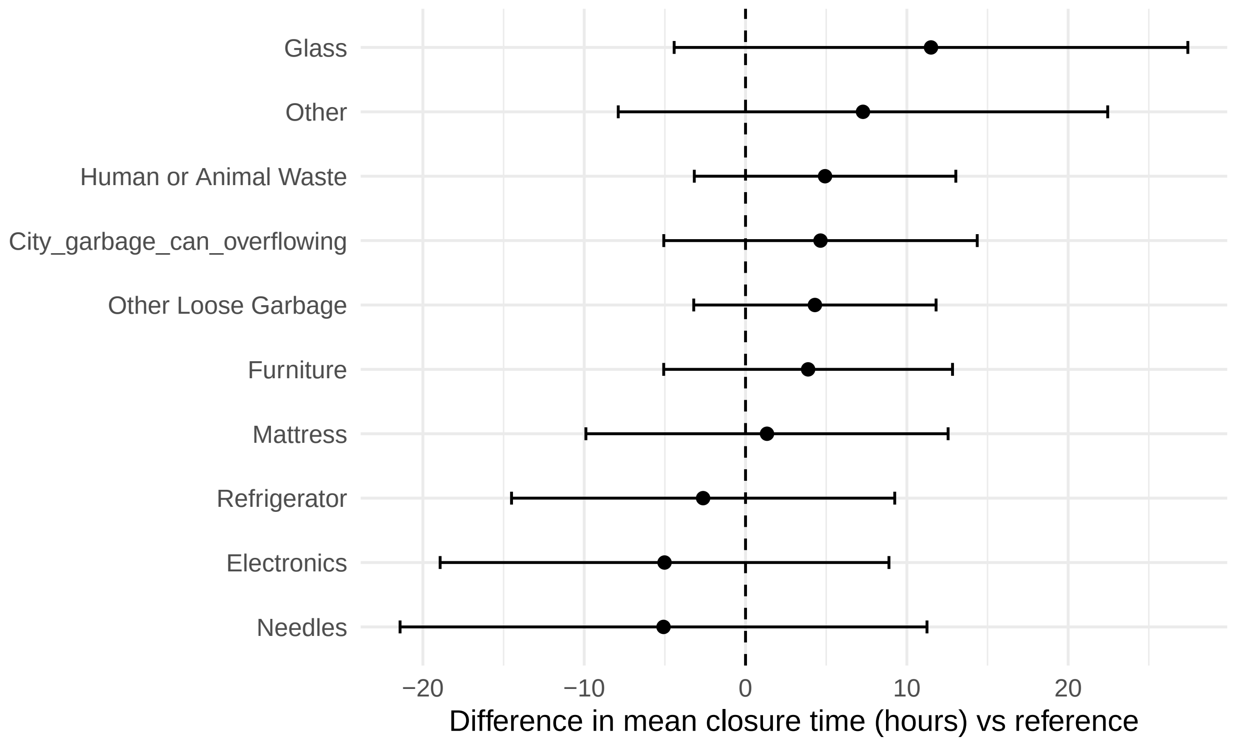

Figure 1: Request type coefficient plot. Each dot is the estimated difference in mean closure time in hours for a given request type compared to the reference type used in the model. Horizontal whiskers show the 95% confidence interval for that difference. The vertical dashed line at 0 marks 'no difference' relative to the reference, intervals crossing 0 are not statistically distinguishable from the reference at the 5% level.

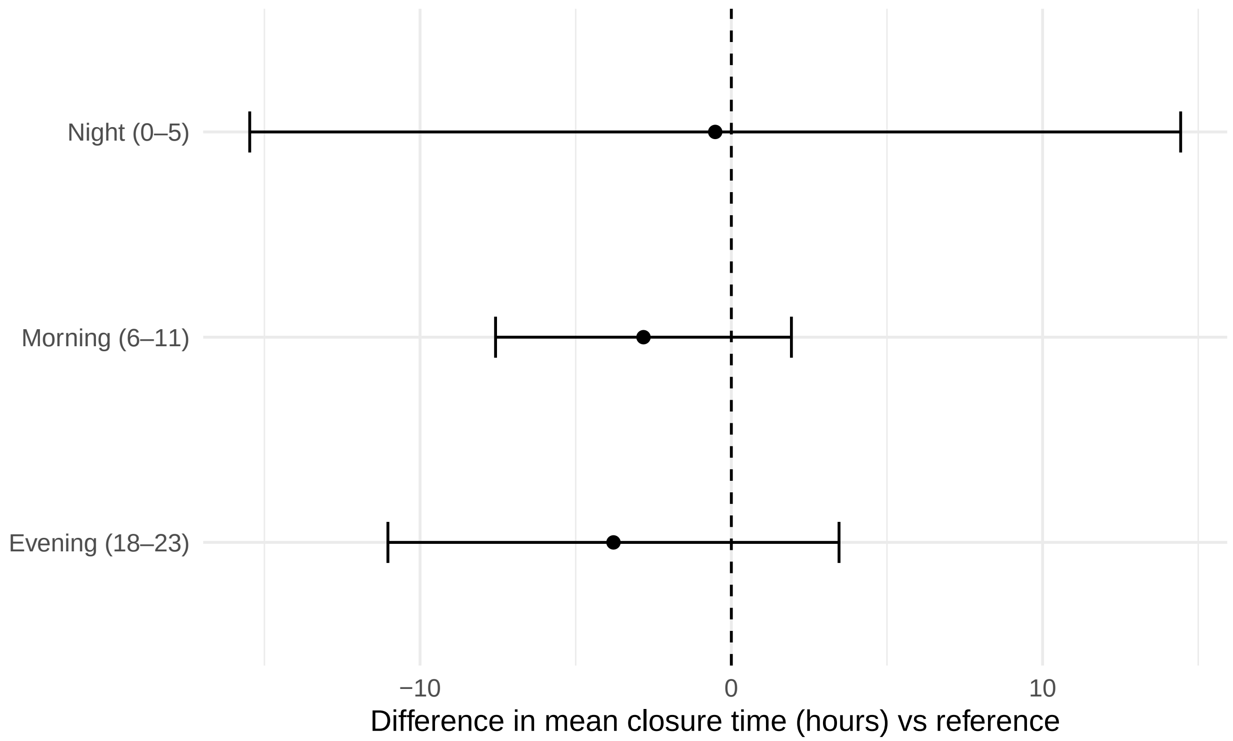

Figure 2: Time-of-day coefficient plot. Each dot is the estimated difference in mean closure time in hours for the indicated time-of-day bin (Night 0–5, Morning 6–11, Afternoon 12–17, Evening 18–23) relative to the reference bin. Whiskers are 95% confidence intervals. The vertical dashed line at 0 is the 'no difference' reference, intervals crossing 0 are not statistically distinguishable from the reference at the 5% level.

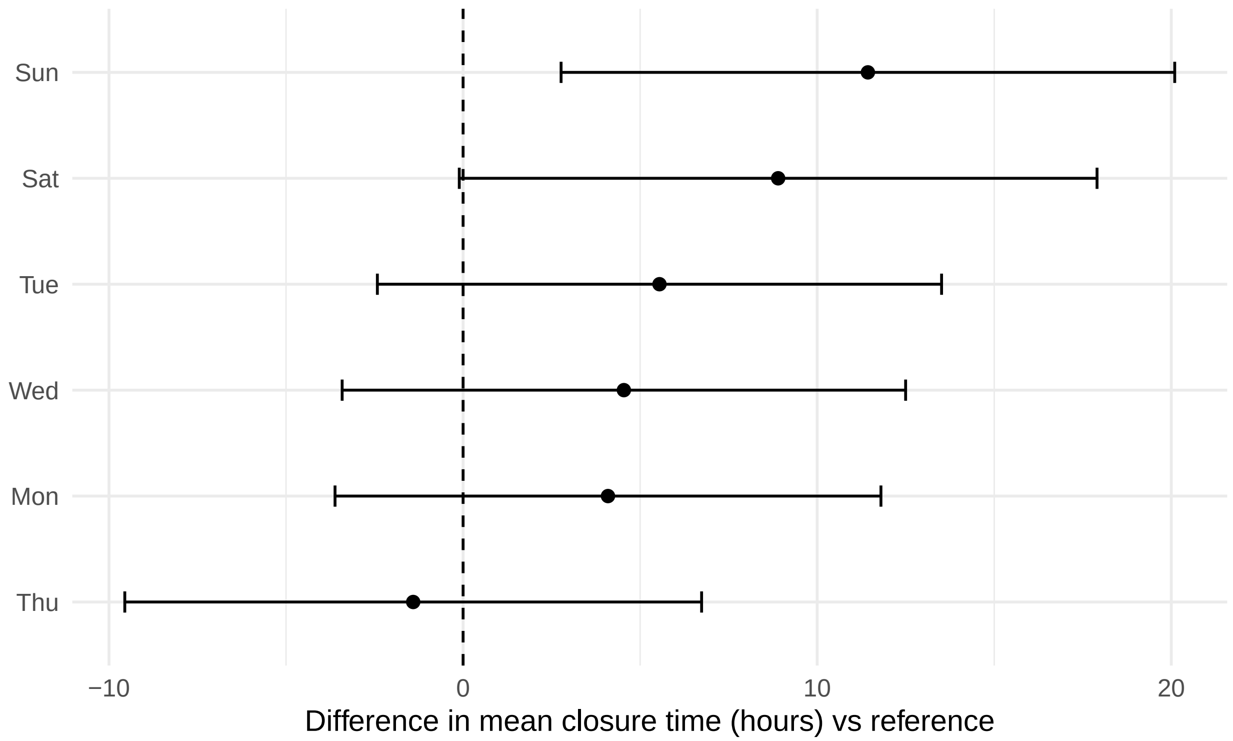

Figure 3: Weekday coefficient plot. Dots are estimated differences in mean closure time in hours by weekday relative to the model's reference day. Whiskers are 95% confidence intervals. The vertical dashed line at 0 indicates no difference from the reference, intervals crossing 0 suggest no clear evidence of a weekday effect at the 5% level.

What these estimates show:

Request type: Most categories' intervals cross zero; none stand out as materially different from the baseline at the 95% level given our sample and noise.

Time of day: Morning, Afternoon, and Evening effects are close to zero with wide intervals.

Weekday: A few weekdays, such as Thu and Fri, show small negative estimates, meaning slightly faster average closure than the baseline day, but effect sizes are only of a few hours and should be interpreted cautiously.

Connection to the question

The research question asked whether request type, weekday, and time-of-day help explain how long SSC requests remain open. Based on the plots and the global F-test, the answer is "only weakly, if at all." Whatever drives long closures (e.g., crew assignment, travel distance, queue backlogs, multi-agency coordination) likely isn't well captured by these three fields alone.

Diagnostics

I assessed model assumptions with two standard plots. The Residuals vs Fitted plot shows no strong curvature—the loss line stays close to 0—so a linear mean is a reasonable first pass. The spread of residuals is fairly flat across fitted values, with a few very large positive residuals that dominate the vertical scale, this suggests mild heteroskedasticity driven mostly by outliers rather than a clear funnel pattern. The Normal Q–Q plot, however, departs sharply from the reference line in the upper tail, indicating heavy right tails and several extreme observations; the left tail lies closer to the line. Taken together, these diagnostics support using the linear form but caution that inference relying on normal errors is fragile in the tails. With a large sample the t/F tests are asymptotically robust, but effect sizes and confidence intervals should be interpreted conservatively and with attention to the outliers.

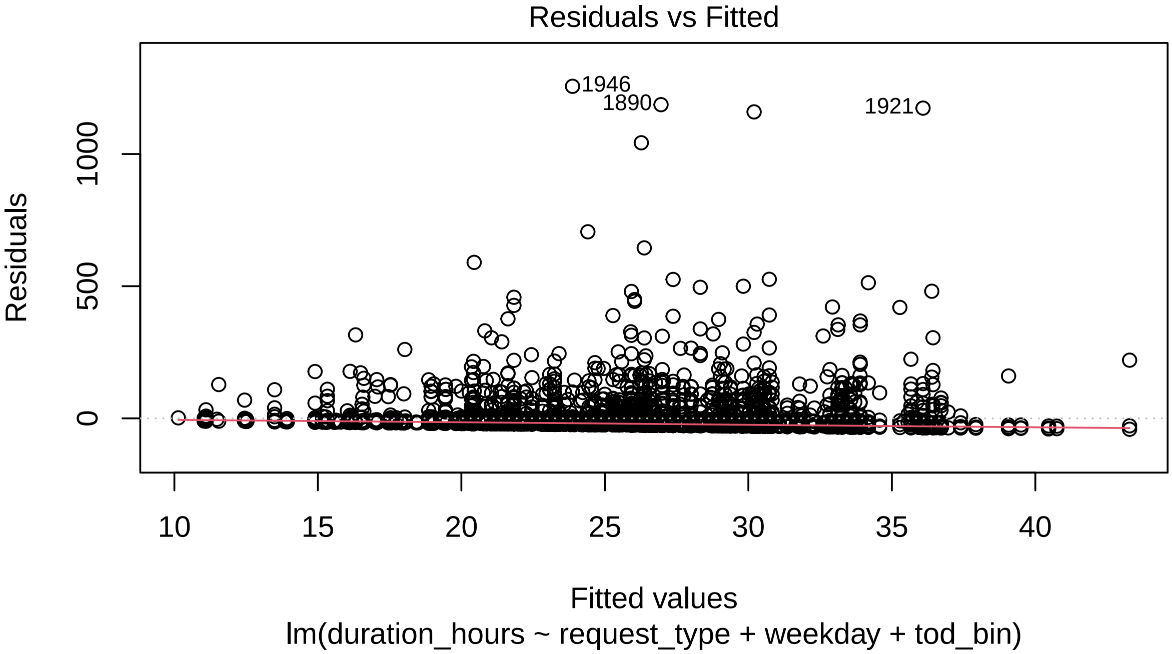

Figure 4: Diagnostics: Residuals vs Fitted. Points are model residuals plotted against fitted values. The horizontal line at 0 marks zero residual. The red smooth curve shows the average trend; noticeable curvature would suggest a non-linear mean, and widening/narrowing scatter indicates heteroskedasticity.

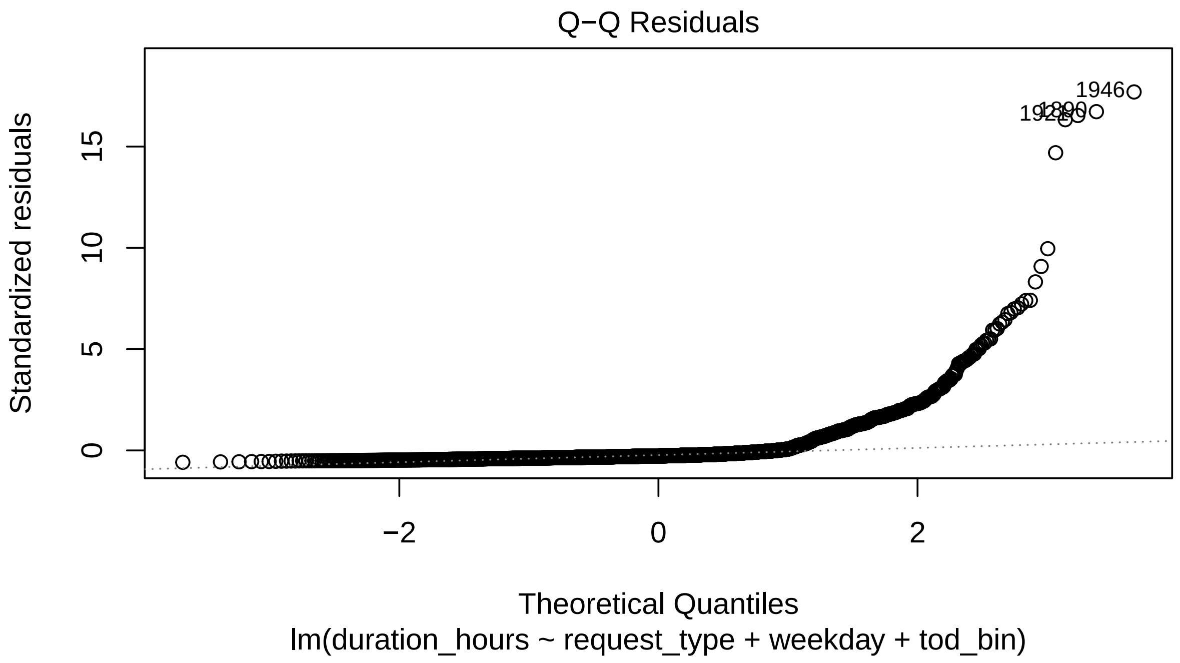

Figure 5: Diagnostics: Normal Q–Q. Points compare standardized residuals to a Normal(0,1) reference. If residuals are approximately normal, points follow the dashed reference line. Systematic tail departures indicate heavy tails or outliers and can affect t-tests and interval accuracy.

Potential limitations and pitfalls.

This analysis uses administrative 311 request logs and a simple linear mean structure with three categorical predictors. That choice keeps the story clear, but it also leaves out a lot of detail. First, closure times are heavily tailed with a few extremely long jobs, thus these points can increase standard errors and pull fitted means around. Second, I removed requests with negative durations and those exceeding 60 days, thus this necessary cleaning could have cut the right tail and slightly changed averages. Third, timing and location matter, jobs may cluster in neighborhoods and time, such as season, special events, and storms, so the independence assumption is only approximately true. Fourth, the three predictors (type, weekday, time of day) likely miss key drivers such as crew assignment, backlogs, or travel distance. With those removed, estimated differences for the included factors could be small and imprecise. Finally, because the model encodes factors via reference levels, coefficients are relative to a baseline and can shift with different coding, and thus interpretation should always be difference from the reference.

Possible extensions

First, address the heavy-tailed outcome by re-fitting with a log or log1p transform of duration_hours. This deals with the long right tail that inflates standard errors and can mask small mean differences. I would then back-transform fitted values for plain-English interpretation and compare residual diagnostics to the current OLS. Second, add a few operational predictors that plausibly drive delays, neighborhood or police district, and month, and one or two interpretable interactions such as request type × time of day to capture staffing or equipment constraints at night. These variables speak directly to scheduling and logistics and are more likely to explain variance than calendar labels alone. Third, account for dependence and check generalization by using cluster robust standard errors to temper over confidence from correlation, and report a simple test or 5-fold cross-validation MAE so readers can see whether any added complexity yields honest results out of sample improvement.

References

City and County of San Francisco. (2019). NBC FY2018–2019 311 SC Calls [Data set]. San Francisco Open Data. https://data.sfgov.org/City-Infrastructure/NBC-FY2018-2019-311-SC-Calls/pk4y-mgyw/about_data

RDocumentation. (2025). RDocumentation. https://www.rdocumentation.org/

Stack Overflow. (2025). Stack Overflow — Questions. https://stackoverflow.com/questions

R Core Team (2025). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.