Accurately Predicting Response Time for Servicing 311 Requests in the City of Boston

Wladimir Nikoluk, Gabriel Walker, Erin St. Peter, Stephane Roux

May 2, 2019

Abstract

We examined 311 reports from 2016–2019 to the city of Boston to create a prediction model for the time it takes to service 311 requests. Although the city provides a generic estimated time for a reported case to be closed, there is no way for individuals reporting issues to know exactly how long it will take for their request to be resolved. The goal of our model is to give transparency to citizens about when the request would be completed given a series of predictors such as location, request type, and month. On average, the model provides a better estimate of completion time than the city does by about 5 days.

Introduction

After a successful pilot program in Baltimore in 1996, the FCC made the 311 phone number available to any police department in the United States wanting to use it to better fulfil requests for city services. Over the years, cities have increased the types of requests able to be serviced through 311 reports and logically, cities are starting to use machine learning and other methods of prediction to improve performance. For example, Chicago is building predictive models with 311 data to find where rat populations are high and where new infestations might emerge. Our approach to prediction with 311 data is more focused on predicting the time it takes for a 311 request to be resolved. At present, many cities provide citizens with a highly general (and often overestimated) time estimate for 311 request fulfillment. Our goal with this project was to create a model that would provide citizens a much more accurate prediction of the time it will take for their request to be fulfilled. To our knowledge, such a tool does not exist for any US cities and it certainly does not yet exist for the city of Boston, our particular area of interest.

The Data

Our data is drawn from a dataset on 311 engagements made public by the city of Boston.[1] It contains 1,493,967 observations, each of which corresponds to a unique request. Most importantly, each observation contains a precise time stamp specifying when it was opened, by what time the city promised to complete it, and when it was completed. There are a range of other variables for each observation, such as the type of issue, or the ZIP code in which the request was made, that serve as predictors. Exhibit 1 in the appendix shows a list of predictors available for each observation with a brief explanation of each.

Data subset selection

Over more than a span of a few years, public service levels change as new departments are created, budgets shift, and political priorities change. To prevent our model from being overly influenced by older observations, we use only data from 2016 onwards to train our model, leaving 729,345 observations. Of these, we drop all observations with missing values and outlier completion-time values, leaving 394,923 observations in total. If data completeness is not random, this will introduce a certain bias into our predictions.[2] Incomplete observations could be more frequent in areas with certain demographic characteristics (e.g., less educated populations with lower civic awareness), which would under-represent those individuals in the dataset. With more time and computing resources we recommend imputing missing values, for example through multivariate imputation by chained equations, and only removing those observations with several missing fields. We also eliminated "exceptional" low volume types of reports (e.g., animal issues, noise disturbances, "abandoned bicycle") that each account for less than 1 percent of the overall reported issues.

We considered outlier completion times to be under 15 minutes or over six months. For the former we assumed that the request was not actually completed, but withdrawn, cancelled, or

[1] Find the original data set here: https://data.boston.gov/dataset/311-service-requests.

[2] For example, claims with incomplete data tend to come from areas in which the processing times are longer, or they tend to be types of claims that are more complex and thus take longer to solve.

immediately referred to a different party for resolution. In this case, we were making an assumption that the request was not truly "fixed" if it was marked resolved in such a short time. For the latter we assumed those were left pending either because staff failed to close them or because they were left pending with no intention of completion.

Data manipulation

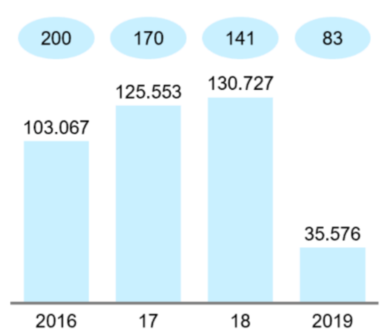
We created completion time variables by subtracting the opening and closing times of engagements to calculate the overall time taken to complete any given request.[3] We then added three additional variables to the dataset to facilitate the analysis:

Variable	Description
score	The percentile rank of the difference between when the city reported the issue would be closed and when the issue actually was closed. A high score corresponds to a large difference between times (e.g., Boston said one week but the issue was completed in a few hours). Note that the data source does not specify whether the requests were actually *completed* or merely *closed*. The score also depends on the city's reported expectation, which is highly variable (ranging from half an hour to 180 days with a mean of 9 days).
month_open	The month when the request occurred. Plausibly, seasonality affects completion time, particularly in a city with harsh winters.
time	We created categorical variables for bins of the time of day at which the request was opened: 6 a.m. to 12 p.m. (morning); 12–6 p.m. (afternoon); 6 p.m. to 6 a.m. (night). We hypothesize that the time of the day might be a significant predictor, as the time of the day at which the request is received by staff might change the way in which it is prioritized and processed.

Description of the Data & Summary Statistics

Exhibit 2.2 shows the number of observations by year in our reduced dataset. The observations are distributed broadly equally across years, with a slight year-on-year increase. Observations for 2019 are obviously fewer given

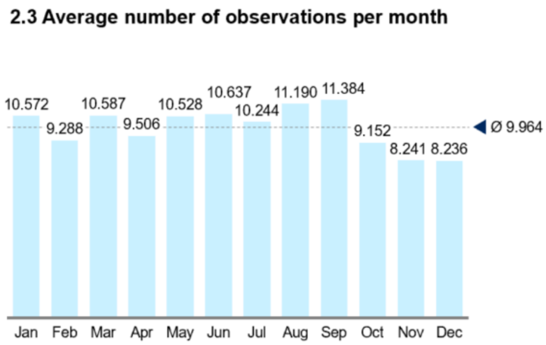
2.2 Number of observations per year (bar chart)
Average duration until completion (in hours)

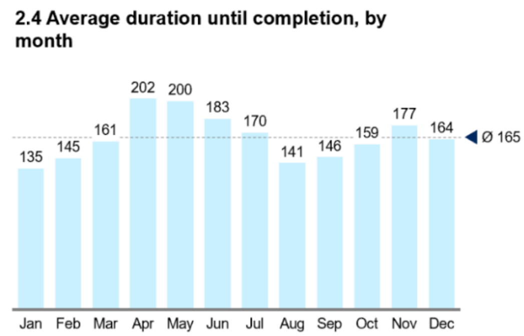


(Exhibit 2.2 shows observations in units of 1000)

[3] Some cases are referred to other agencies or "closed" due to lack of information. Currently, cases that were closed without being completed are indistinguishable from successfully completed cases in our dataset. Thus, our predictive model will strictly speaking predict time until case closure (whether actually addressed by the city or not) rather than time until the issue was resolved. Ideally, 311 data would differentiate between these types of resolution.

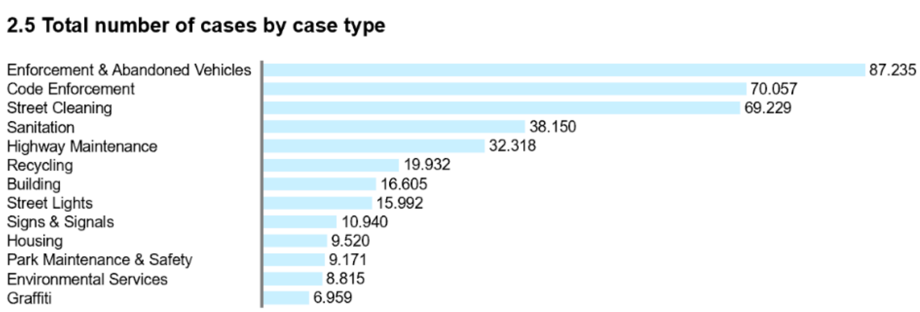
that the year is incomplete. The average monthly number of cases for 2019 is similar to the overall average. The chart also shows that the average duration until completion of cases has gone down substantially since 2016, potentially due to improvements in efficiency.[4]





(Exhibit 2.3 shows observations in units of 1000 observations and 2.4 shows duration in units of hours)

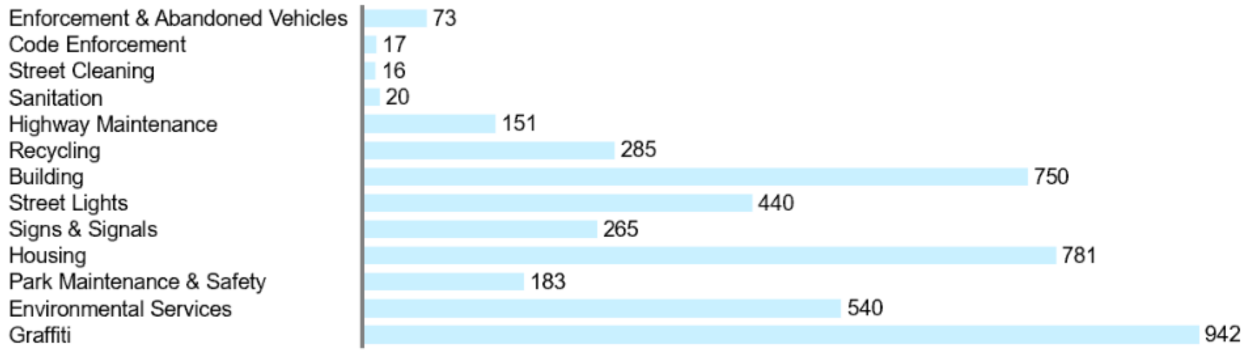
Exhibits 2.3 and 2.4 show the variation in caseload by month, as well as completion time by month. There appears to be some seasonality in the number of cases across months, but a substantive amount of seasonality with respect to the duration until completion. This could be due to the type of cases that are associated with particular times of the year, or some other factors about the months that make it more time-consuming to complete cases (e.g., more absences, holidays and illnesses of staff, or other responsibilities that coincide with April/May).



(Exhibit 2.5 shows cases in units of 1000 cases)

[4] The 2019 figure, however, looks suspiciously low, perhaps due to the fact that the recent months do not contain some of the cases with the longest duration.

2.6 Average duration until case completion by case type

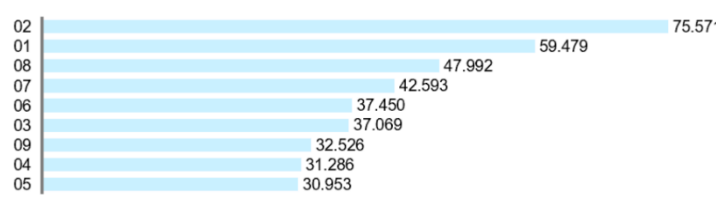


(Exhibit 2.6 shows duration in hours)

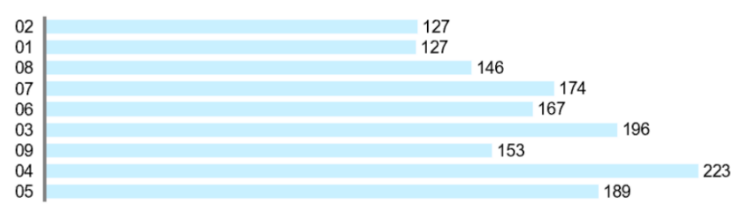
Exhibits 2.5 and 2.6 show the total number of cases by case type and the average duration until case completion by case type. While some case types are much more common than others (the top 3 case types make up around 55 percent of the overall case volume), the distribution is not as uneven as often observed in traditional 20-80 settings, with a long tail of exceedingly rare cases, because we eliminated cases constituting less than 1 percent of the overall observations.

The duration of request fulfillment varies substantially by type. For example, average completion times for “code enforcement” are 55 times faster than for graffiti. Exhibits 2.7 and 2.8 give insights into the distribution of cases by city council (each number stands for one of 9 city councils). The cases are fairly evenly distributed across councils, with councils 01 and 02 carrying a higher share of the total. Completion time varies substantially, with the fastest councils (01 and 02) almost twice as fast as the slowest councils. Interestingly, the councils with the highest numbers of cases also appear to be the fastest. This could be due to the fact that there are scale

2.7 Total number of cases by city council



2.8 Average duration until completion by city council



(Exhibit 2.7 shows cases in units of 1000 cases and Exhibit 2.8 shows duration in hours)

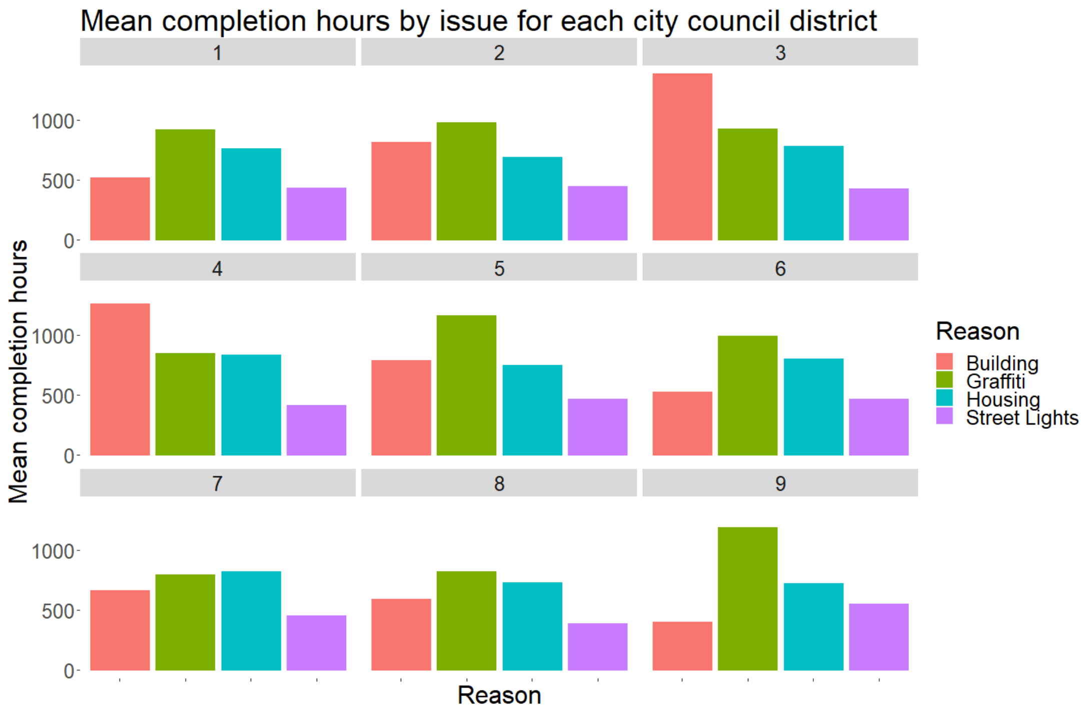
effects, or that citizens are more likely to report cases where response rates are fast.

Variable Selection

Our dataset provided numerous geographic predictor variables. Of these, we selected city council as our geographic demarcation variable for our basic analysis for reasons further described in the appendix. After eliminating alternative location variables (e.g., ward) we plotted the remaining variables to identify the correlations that might exist in the data (see table 1 in the appendix). The plots suggest that correlation is generally low between variables—no clear patterns emerge in most binary relationships. This was interesting because some relationships could have been reasonably expected. For example, we were assuming that promised hours (the estimate provided during the call for how long it would take to service the request) would be strongly correlated with completion hours (the actual time it took to complete the request). The fact that the correlation was only 0.37 and that the mean of promised hours was higher than actual completion (211 vs 159 hours) suggests that the 311 estimation might consciously overstate the time it will take to service a request, thereby reducing accuracy but also ensuring that citizens don't get frustrated if some requests are serviced later than expected. This reinforced the importance of creating a better prediction model for better city planning purposes as well as for increasing transparency in communications to citizens.

Before building a more robust prediction tool beyond linear regression, we were also interested in exploring service time across reasons *and* districts. We find some heterogeneity in how mean completion hours by request reason vary across districts. As shown in the graph below, housing and street light requests are fulfilled fairly uniformly across districts while graffiti and building requests are not. Generic city estimates produced from request reason without considering geography could miss this variation.[5]

[5] Of course, it could also be true that the distribution of different types of requests within each "reason" category also differ by district.



Methods

We first used a simple linear regression approach to build an initial exploratory model. To this end, we created dummy variables for the council districts and created additional dummies for each of our categorical values (e.g., request reason, month, source of request and time of day), which increased the number of features in our dataset. For our categorical dummy variables, we deleted one value, to ensure that the regression could work properly. We then fit a linear regression model to the data. The "naive" model resulted in a modest r-square of 0.251. Our table of coefficients for the linear regression model is provided in the appendix and its broad insights are provided below:

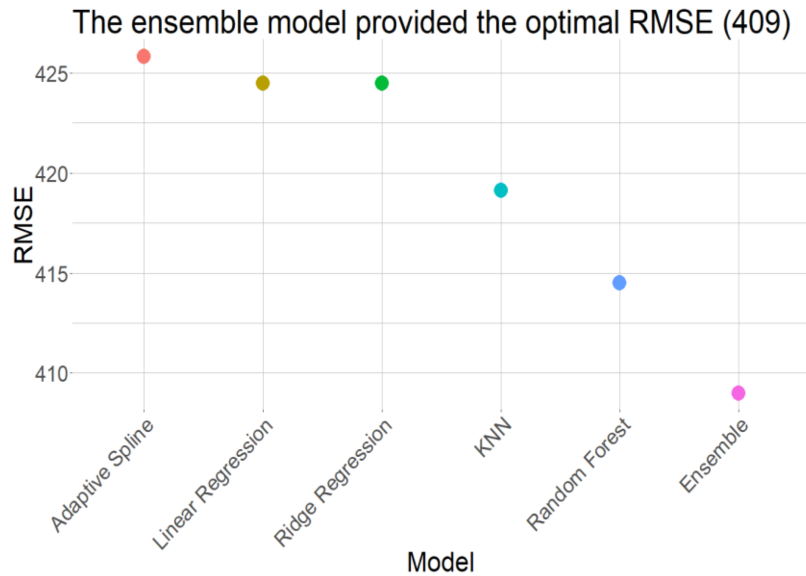
- *None of the districts alone were statistically significant, indicating no overall difference in service time between them (though creating interaction variables between the type of request and geographic area may have produced different results).*
- *The source of the request (city worker call, constituent, app, etc) does not seem to matter either since none of the associated variables turned out to be statistically significant.*
- *Unsurprisingly, the key determinants of completion time were request reason and timing. Winter months such as November and December added on average around 20 hours to a request. Surprisingly, daytime requests were correlated with longer completion times than evening ones, perhaps because the staff is less busy at night or because requests sent at different times are different in ways unobservable in our dataset.*

We then built a stacked ensemble model that attempted to maximize the accuracy of predicted completion hours.[6] Our goal was not interpretability but to provide the best estimate for city residents—a feature that could be included in the Boston 311 mobile app, for example. First, we selected features with a random forest on 1 percent of the training data and settled on the top five (reason, department, month_open, source, fire_district). We then took a random subset of 20 percent of our observations to build training and test sets (90 and 10 percent of the subset, respectively) to save on computational time. The ensemble consisted of a KNN model (optimized with $k = 9$), linear regression, adaptive spline, ridge regression, and random forest. This combination of different "types" of models (regressions, trees, etc.) has proven effective for many other machine learning applications and is preferred to using an individual model as an alternative. Lastly, we created meta-model (an adaptive spline) that created a final prediction based on the predictions of each of the sub-models. For those not familiar with machine learning, the ensemble model is essentially a collection of a group of low-level learners that provided predictions to one high-level learner that synthesizes them for an even more accurate prediction.

Results

Of all the individual models, the ensemble model produced the lowest RMSE of approximately 409 hours, thus providing the best performance as a predictive tool for estimating actual completion time. Although this figure, around 17 days, is not a precise value for Boston citizens, it is an improvement on the current RMSE of 530 hours, or 22 days, for the city's predicted completion time. But if predictive accuracy is paramount, this model is objectively better than the base-level individual learners or the city's current estimate.

[6] Find the complete final model at https://github.com/gbwalker/boston_311.



Discussion

The choice of an ensemble model, though providing the best RMSE, involves considerable computational time and sacrifices the transparency and interpretability of simpler models. Additionally, while the ensemble model produced the lowest RMSE, the level of RMSE for all models was still rather high. This may be because limitations on the number of categorical variables in the random forest model caused us to drop the "type" variable which was a more detailed description of the "reason" variable that coded the subject of the request. Within the umbrella "reason" categories, much of the actual performance time may be determined significantly by this more precise description of the request. For example, requests for "highway maintenance" (a reason) could involve a variety of types of requests that naturally would have different response times (e.g., debris blocking lanes vs. a small pothole). Additionally, this ensemble model was trained on only 5 percent of the *total* public data (about 1.5 million observations). If the computational resources were available to add more base learners and process the other 95 percent of the data the out-of-sample performance could be dramatically decreased from a RMSE of 409 hours.

It should also be noted that the data itself could be biased: wealthier, more educated individuals might be more frequent reporters of 311 issues (thus comprising a greater proportion of the dataset). If this is true, then the model would theoretically be trained better for predicting response times for their issues. Overall, the contribution of this model seems to be that it is a better predictor of response time than currently exists in Boston, but the representation issue should not be ignored.

Recommendations

The ensemble model, even with limited training data, provided a more accurate estimate of completion time than the city's estimates. First, besides using a greater portion of the data, we recommend that future work be done to attempt to include some of the information that was dropped from excluding "type" from our model. Although there exist too many different types of requests to fit certain standard models, a bundling of types into certain categories (e.g., emergency vs. non-emergency) might provide for more robust predictions. Our dataset also did not provide information on whether or not a photo was included with a request. One way to report 311 requests to Boston is through a mobile app where you can attach a photo along with the request. One might imagine that photos play some role in city response time (for example, more detail provided in photos might lead to greater accuracy of city predictions on fulfillment time). If photos tended to be correlated with reduced completion time, an app could even nudge submitters to remember this and include a photo. Finally, we would recommend further attempts by cities to label data in such a way to better distinguish between requests that are actually fulfilled by the city and requests that are resolved due to lack of information or referral to an outside agency.

Appendix

Exhibit 1: List of all Variables in Dataset

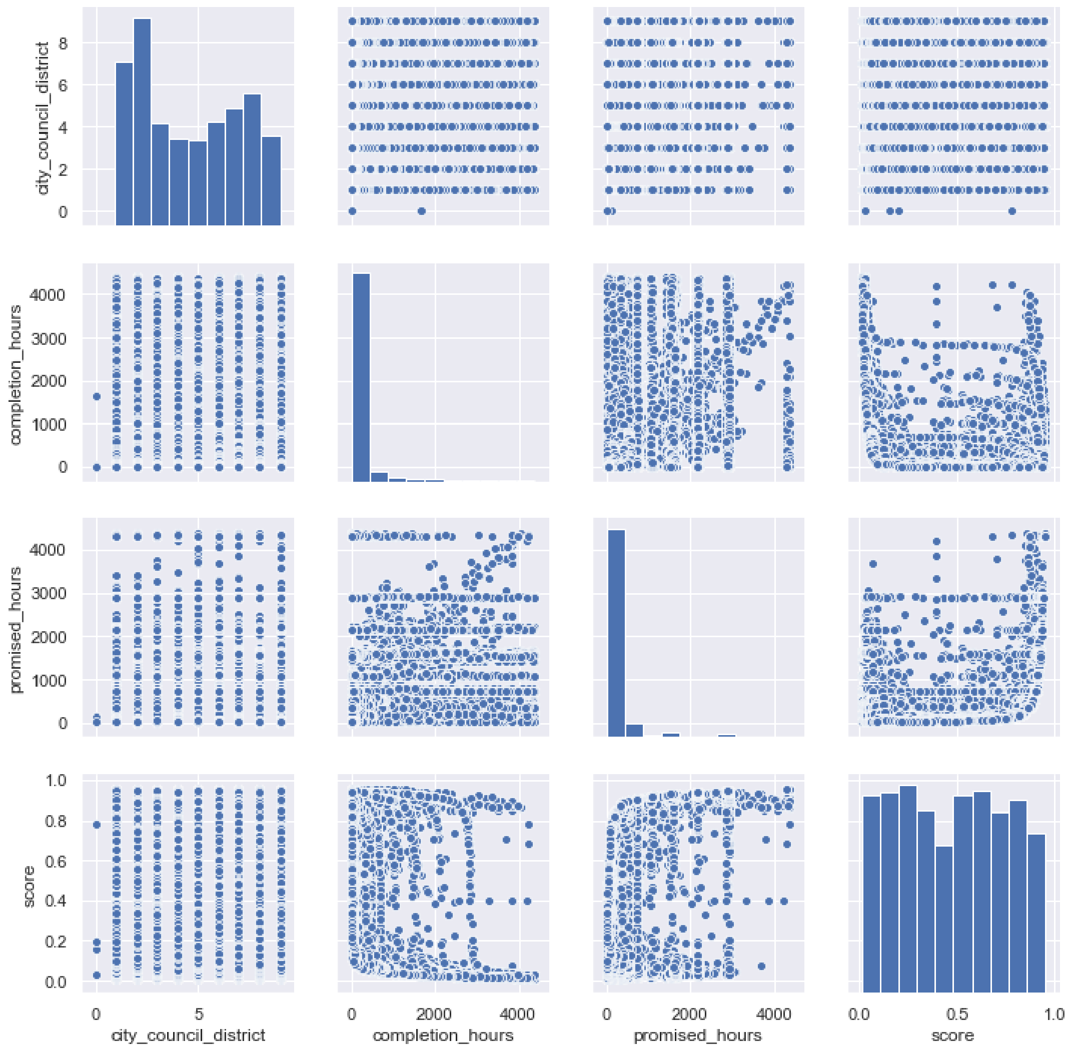
open_dt	Date the issue was reported.
target_dt	Date the city planned to handle the issue.
closed_dt	Date the issue was closed.
reason	General issue area (graffiti, highway maintenance, housing, etc.).
type	Specific issue.
department	Whose responsibility the issue is.
fire_district	Fire district the issue was reported in.
pwd_district	Public Works district.
city_council_district	City Council district.
police_district	Police district.
neighborhood	Neighborhood.
ward	Ward.
location_zipcode	ZIP code.
source	Who reported the issue (constituent call, city worker, etc).
latitude	Latitude.
longitude	Longitude.
month_open*	Month the issue was reported.
completion_time*	Time it took to handle the issue.
completion_hours*	Hours it took to handle the issue.
promised_time*	Time the city planned to take to handle the issue.
promised_hours*	Hours the city planned to take to handle the issue.
time*	Categorical variable for morning, afternoon, or night reporting.
score*	Percentile score for the difference between how long the city thought it would take and how long it actually took (promised - completion). Higher is better, since it means it took shorter than expected.

Note that * corresponds to a user-created variable that was not in the original dataset.

Choice of Geographic Variable

Our dataset provided numerous geographic predictor variables. Of these, we selected city council as our geographic demarcation variable for our analysis because of the largely even distribution of 311 requests across the individual councils as opposed to, for example, the fire districts. Fire districts 5 and 10 only had one 311 request each, which left us wondering whether there was an error in the dataset or whether these districts might have been absorbed into other districts. Furthermore, city council district seemed more relevant for efforts to make improvements in response time given that there is a clear political path for change using this geographic variable as opposed to others.

Appendix Table 1: Scatterplots of district, completion hours, promised hours, and score



Initial Linear Regression Model Output (numbers indicate city council district)

	coef	std err	t	P>|t|	[0.025	0.975]
const	856.9891	209.662	4.087	0.000	446.057	1267.921
promised_hours	0.3452	0.003	102.322	0.000	0.339	0.352
reason_Code Enforcement	-699.0157	3.781	-184.858	0.000	-706.427	-691.604
reason_Enforcement & Abandoned Vehicles	-643.6515	3.831	-168.030	0.000	-651.159	-636.144
reason_Environmental Services	-382.5416	5.818	-65.754	0.000	-393.944	-371.139
reason_Graffiti	-267.6351	7.842	-34.128	0.000	-283.006	-252.265
reason_Highway Maintenance	-555.6255	4.201	-132.249	0.000	-563.860	-547.391
reason_Housing	-758.4584	9.371	-80.936	0.000	-776.825	-740.091
reason_Park Maintenance & Safety	-566.7258	5.606	-101.087	0.000	-577.714	-555.738
reason_Recycling	-554.1273	4.551	-121.751	0.000	-563.048	-545.207
reason_Sanitation	-703.4136	3.946	-178.243	0.000	-711.148	-695.679
reason_Signs & Signals	-559.3135	5.421	-103.179	0.000	-569.938	-548.689
reason_Street Cleaning	-684.4358	3.916	-174.777	0.000	-692.111	-676.760
reason_Street Lights	-404.7691	5.011	-80.775	0.000	-414.591	-394.948
source_Citizens Connect App	2.6253	17.995	0.146	0.884	-32.645	37.895
source_City Worker App	2.8722	18.144	0.158	0.874	-32.689	38.433
source_Constituent Call	33.9715	18.006	1.887	0.059	-1.319	69.262
source_Employee Generated	-34.2123	18.513	-1.848	0.065	-70.498	2.073
source_Maximo Integration	-289.4088	295.911	-0.978	0.328	-869.386	290.569
source_Self Service	-0.3031	18.383	-0.016	0.987	-36.333	35.727
month_open_April	-8.4430	3.050	-2.768	0.006	-14.420	-2.465
month_open_August	-16.6668	3.049	-5.467	0.000	-22.642	-10.691
month_open_December	26.0016	3.347	7.768	0.000	19.441	32.562
month_open_February	1.2303	2.972	0.414	0.679	-4.595	7.056
month_open_July	-8.0594	3.143	-2.565	0.010	-14.219	-1.900
month_open_June	0.1397	3.111	0.045	0.964	-5.958	6.237
month_open_March	-6.8484	2.878	-2.380	0.017	-12.489	-1.207
month_open_May	5.7144	3.121	1.831	0.067	-0.403	11.832
month_open_November	19.5533	3.353	5.831	0.000	12.981	26.125
month_open_October	-8.0664	3.245	-2.486	0.013	-14.426	-1.706
month_open_September	-11.7131	3.061	-3.826	0.000	-17.713	-5.713
time_afternoon	24.7170	1.900	13.007	0.000	20.992	28.442
time_morning	11.8400	1.892	6.257	0.000	8.131	15.549
1	-211.9899	208.860	-1.015	0.310	-621.350	197.370
2	-200.3288	208.859	-0.959	0.337	-609.687	209.029
3	-162.5672	208.865	-0.778	0.436	-571.936	246.802
4	-157.7647	208.867	-0.755	0.450	-567.137	251.608
5	-175.4124	208.867	-0.840	0.401	-584.785	233.960
6	-198.4556	208.865	-0.950	0.342	-607.824	210.913
7	-187.2809	208.863	-0.897	0.370	-596.647	222.085
8	-209.5806	208.862	-1.003	0.316	-618.945	199.784
9	-197.7726	208.866	-0.947	0.344	-607.145	211.599

Dep. Variable:	completion_hours	R-squared:	0.251
Model:	OLS	Adj. R-squared:	0.251
Method:	Least Squares	F-statistic:	3234.
Date:	Wed, 24 Apr 2019	Prob (F-statistic):	0.00
Time:	19:28:05	Log-Likelihood:	-2.9436e+06
No. Observations:	394923	AIC:	5.887e+06
Df Residuals:	394881	BIC:	5.888e+06
Df Model:	41		
Covariance Type:	nonrobust		

Sample of Boston 311 Requests
[sourced from live report stream https://mayors24.cityofboston.gov/]

Other at 301 Chestnut Ave, 1, Jamaica Plain

Power outage at Rockview & St John's in JP!?

CLOSED Case Referred to External Agency. Eversource made aware of issue (transformer failure) at time of report and was working to restore asap. - about 3 hours ago #101002895111

Damaged Sign at 411 413 Washington St, Dorchester

Bus stop sign down

OPENED about 4 hours ago #101002895146

Rodent Sighting at 10 Fenwood Rd

There were multiple giant rats seen on the same day for multiple days. Is there anything the city can do?

OPENED 39 minutes ago #101002895218

Other at 315 Mount Vernon St, Dorchester

Baby bunnies in planter on school grounds

OPENED about 20 hours ago #101002887190

Other at 144 Beech St, Roslindale

Loud car. Red sports car. Very loud exhaust.

CLOSED Case Invalid. insufficient information to process; if vehicle has remained stationary for an extended period, please notify Boston Police with details. - 1 day ago #101002886476

Other at 73 Iroquois St, Mission Hill

Tree limb in the middle of the street.

CLOSED Case Resolved. Done dr jd. - about 24 hours ago #101002885083