Predicting Resolution Time of 311 Calls in Minneapolis

By Langhan Dee

The Problem:

How can the City of Minneapolis improve the way they allocate resources and handle non-emergency service requests? Minneapolis, like many cities, has a 311 phone number you can call to find information about the city's non-emergency services, make complaints, or report problems. Questions can be related to almost any of the city's departments or services. For example, residents can ask about trash collection information or zoning permits. They can report graffiti, potholes, stray animals, water pollution, or abandoned cars. Each call (or "case") is tracked by type, location, the date the issue was reported and resolved. This project examines 311 complaint trends and seeks to predict how long each case will take to resolve.

The Client:

The primary client for this project is the City of Minneapolis. More specifically, findings will interest city planners and departments which monitor and resolve 311 Requests. By identifying the most common 311 information requesting calls, the City of Minneapolis can determine key information to make more prominent on their website. The City can use case resolution predictions to decide where to efficiently allocate department resources. This has the potential to reduce the staffing budget in multiple departments and improve response time to residents' concerns.

The Data:

I. Summary

The data used in this project was gathered from a few different sources. Information was accessed with an API or downloaded as CSV files. The data was cleaned and consolidated into a single CSV file, ready for further analysis and modeling.

Minneapolis 311 Open Dataset (2016-2018)

Annual 311 service request data was available from the City of Minneapolis Open Data website. The data for 2017 and 2018 data was collected as a JSON file through an API. The 2016 dataset was downloaded as a CSV file. The 2016-2017 datasets include 49,897-55,164 rows of data per year. The 2018 data is updated daily and will be used as a test dataset.

Columns:

- **OpenDateTime, ClosedDateTime, CaseStatus** - This indicates if and when a service request was resolved and closed
- **SubjectName** - This is closely associated with the department designation (ex. Vehicles and Commuting, Public Safety, Animal Related, etc.)
- **SubjectReason, SubjectType** - 2nd and 3rd level classification and description of the service request
- **XCoord, YCoord** - Coordinates of where the service request occurred

Source: http://opendata.minneapolismn.gov/datasets?t=311

File: 311_Incidents_2016.csv

Historical Climate Data for Minneapolis (2016-2018)

Daily historical climate data was collected as a CSV file from the Minnesota Department of Natural Resources.

Columns:

- Maximum and minimum temperature (degrees Fahrenheit)
- Precipitation (inches)
- Snowfall (inches)
- Snow depth (accumulated inches)

Source: https://www.dnr.state.mn.us/climate/historical/summary.html

File: Climate_Data_16_17.csv

II. Cleaning and Consolidating the Data

In order to consolidate 311 data from different years (and different file formats), column headers and timestamp formats were standardized. Some unnecessary attributes such as Case ID and Enquiry ID were dropped to reduce dimensionality. A new feature, called Days_Open, was engineered by calculating the number of days before each case was closed. The timestamps in the 311 data included the date and time, however, the climate data was recorded per day. The date each 311 case was opened was extracted from the timestamp so that the data frames could be merged on that feature.

III. Missing Values and Outliers

Although neither dataset had missing values, there were a few inconsistencies in the notation that were adjusted for ease of future analysis.

- One 311 case (a parking violation) had an opening time stamp 3 seconds *after* it was closed. The time stamps were left as they were, but the Days_Open value was updated to be 1 (the minimum number of days).
- Some 311 call types tended to have large clusters of cases all closed on the same date. These clusters occurred most notably for Sidewalk Snow & Ice Complaints, Exterior Nuisance Complaints, Private Property - Residential Conditions, and Hoarding Complaints. It's unclear if these close dates are accurate, or were filled in to avoid missing values.
- The climate data contained some entries of "T", indicating there were "trace amounts" of precipitation, snowfall or snow accumulation (depth). To convert "T" to a numerical value, it was replaced with 0.05 or 0.005 (inches).

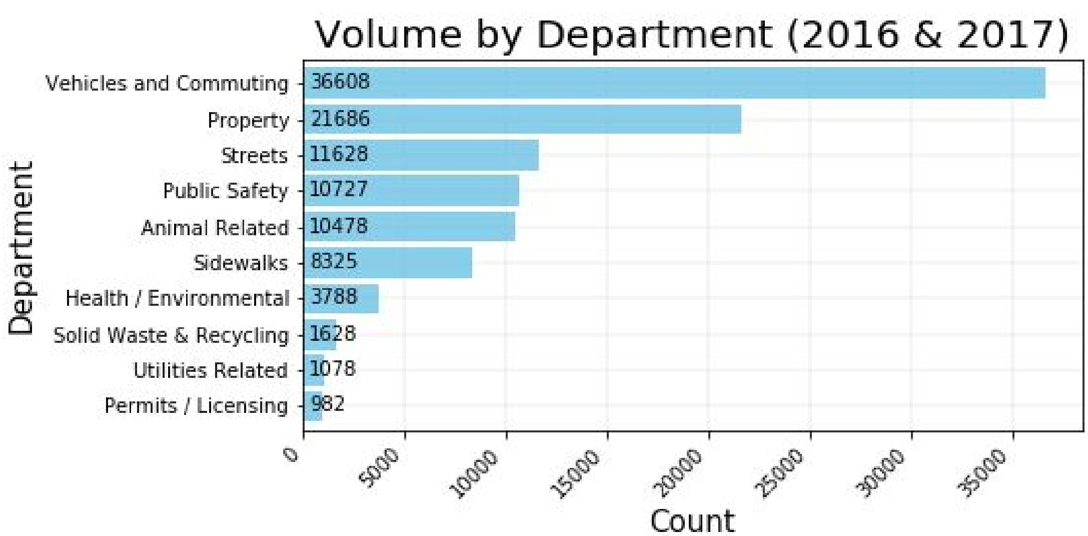
After cleaning and consolidating the data, it was converted to the CSV file listed below.

- 311_clean.csv

Exploratory Data Analysis:

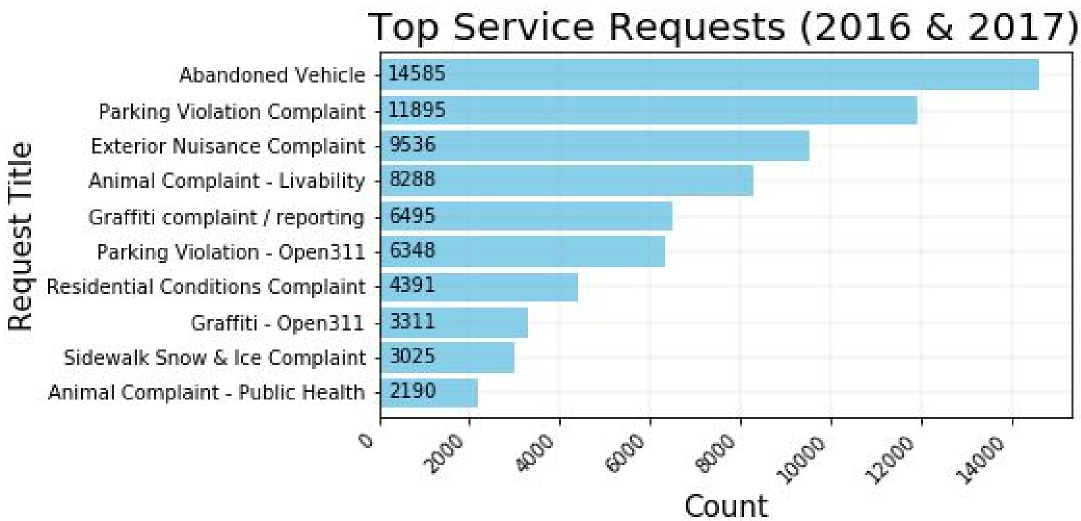
1. The busiest departments

Vehicles and Commuting and **Property**, make up **55% of call volume**. Not surprisingly, 5 of the top 7 case titles are from these two departments



2. Ten most common 311 service requests

Minneapolis residents care about where other people leave their cars. The two most common call types were reporting **abandoned vehicles** and **parking violations.** In 2017 alone, there were 8,700 calls regarding abandoned vehicles, and 9,775 reports of parking violations. That's an average of 26.78 parking violations per day.



3.

Distribution of complaints by reporting platform (Phone, App, Online)

- **The Open311 mobile app accounted for 17% of cases.** This platform was heavily used for issues residents noticed while not at home: Parking Violation complaints, Graffiti, Abandoned Vehicles, Sidewalk Snow and Ice, Potholes, and other road-related complaints.

- **The App topics were well chosen to reduce phone staff.** All are high-volume service drivers, while none require immediate assistance (as Animal-Related complaints might). This seems to be an efficient use of technology.

- **Online reporting made up only 2% of cases.** The self-service online portal was only used for 3 case types: Sidewalk Snow and Ice, Exterior Nuisance, and Unpermitted Work (work requiring a City permit).

4. How many days does it take to close a case?

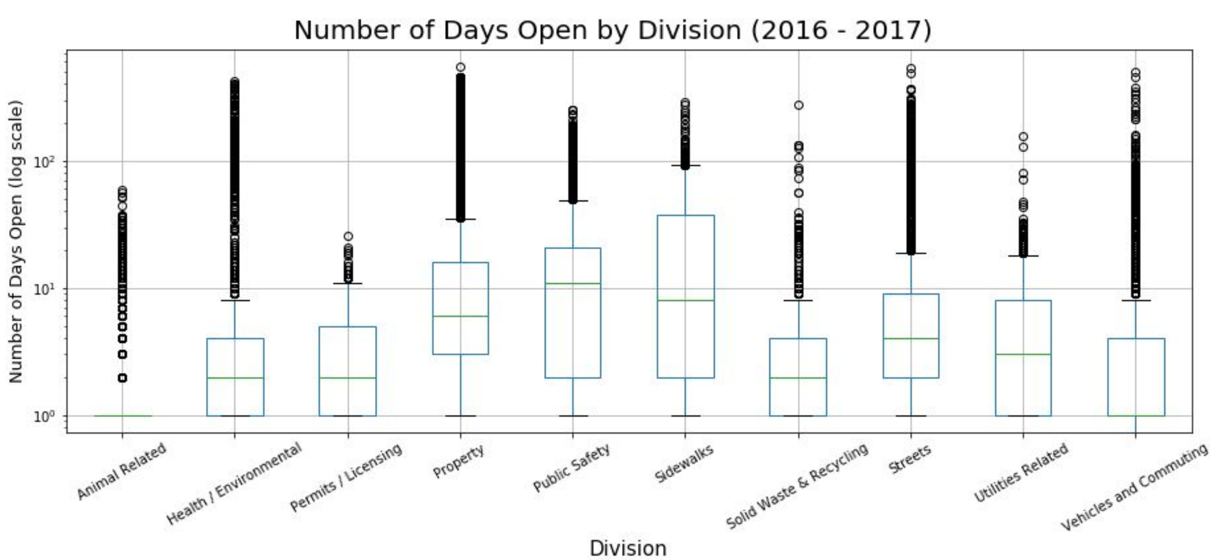
- **On average, cases were closed within 3 days.** Within a week (7 days), 75% of cases are closed.

5. Average number of days cases are open - per division

- **Animal-Related** cases had the fastest response time. More than 75% of cases were closed within 1 day of being opened. As these cases included reports of abandoned, stray, or mistreated animals, they were the only call types affecting the immediate safety of living creatures.

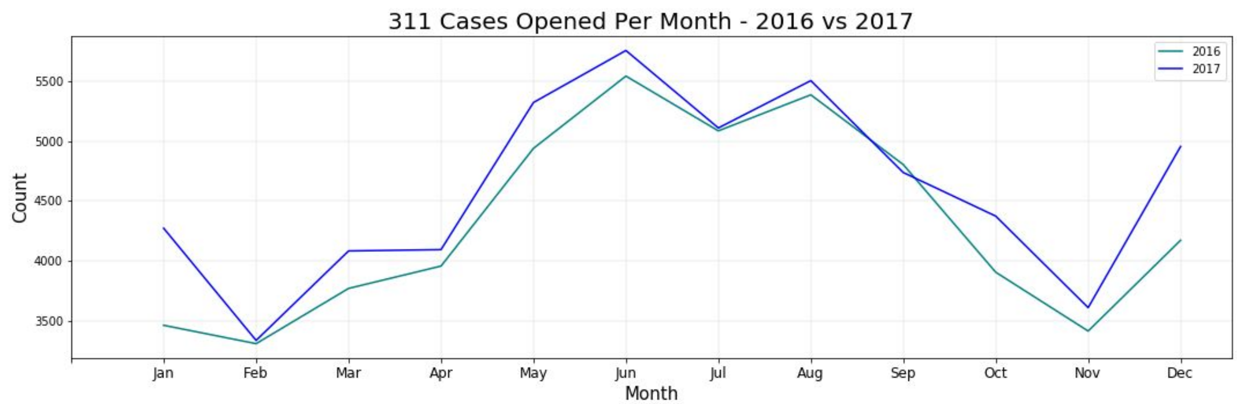
- **Public Safety** had the highest median of 11 days open. This subject included one of the top ten volume drivers, Graffiti complaints.

- **Sidewalks** had the second highest median days open at 8 days, however, 75% of these complaints took 38 days to close. Sidewalk Snow and Ice complaints built up each winter and were closed as a group when the weather warmed up. The volume of these complaints increased dramatically in 2017. They were so problematic that in 2018 the City plans to enforce inspections and fines to encourage people to keep their sidewalks clear.



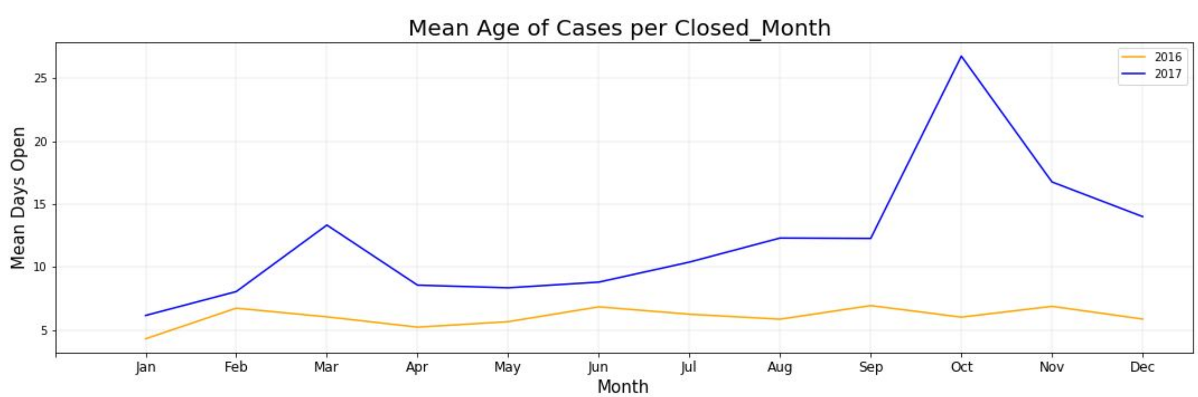
6. Busiest months for opening cases

- **May - September ('16) and October ('17)**: Volume during the summer months is driven by Abandoned Vehicles and Exterior Nuisance cases. The peak season lasted a month longer in 2017 and was slightly higher than 2016 every month except September.
- **December**: Another spike happens in December when snow accumulates and Sidewalk Snow and Ice complaints are reported.



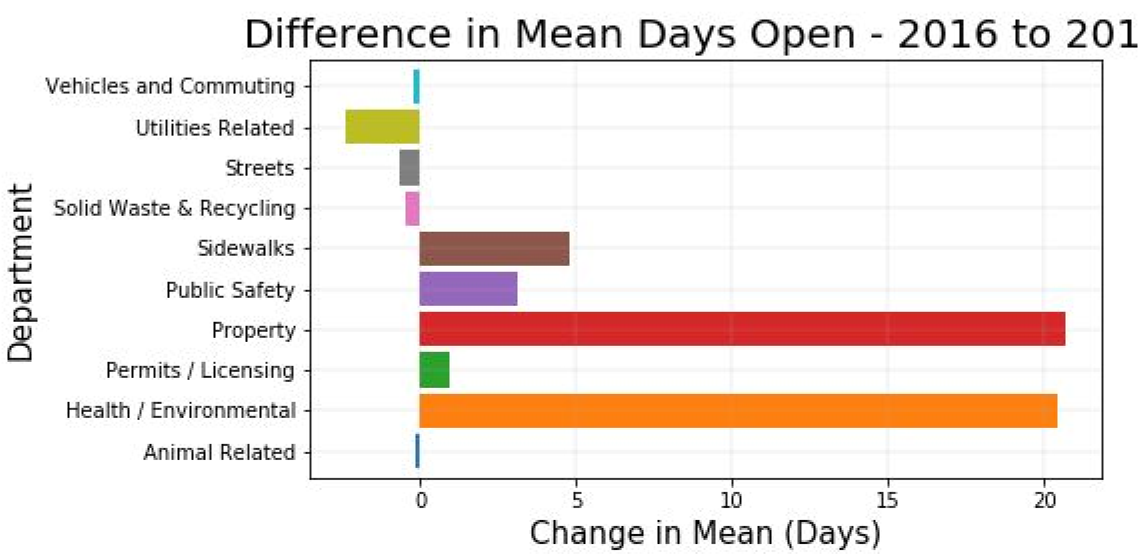
7. How old are cases when they're closed?

- Spikes in case age show when batches of old cases were closed. On 2017-03-31, the 2016 **Sidewalk Snow and Ice complaints** were closed en masse. The following year, Minneapolis didn't wait as long to close similar cases. The group of Sidewalk cases was closed on 2018-01-29.
- A cluster of **Residential Conditions Complaints** was closed on 2017-10-31. Some of these complaints were opened as early as November 2016.
- It should be noted that the 2016 line is relatively flat because 2015 cases were not included in the data, but were still being handled by the city of Minneapolis.



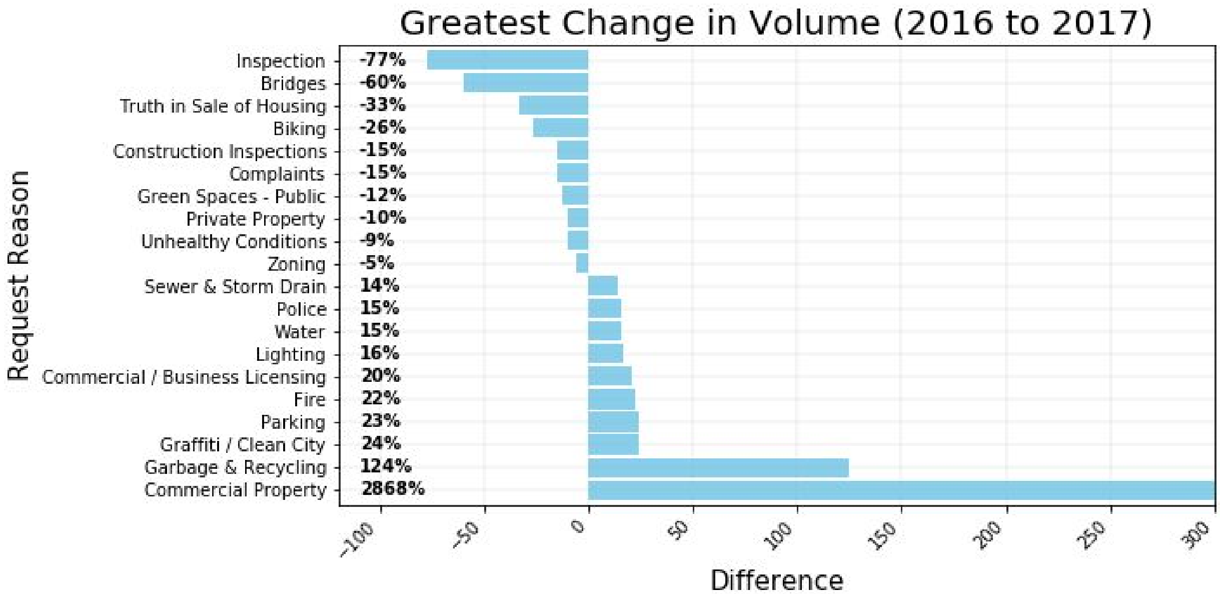
8. Complaint types with the largest change in the number of days open from 2016 to 2017

- From 2016 to 2017, **Vehicles and Commuting**, **Streets**, **Solid Waste & Recycling**, **Permits / Licensing**, and **Animal-Related** cases all had minimal change in the amount of time it took to close cases. The mean close time differed by less than 1 day.
- At the other end of the spectrum, **Property** and **Health / Environmental** cases increased by slightly over **20 days**.
- There are over 10 times as many **Private Property** cases than any other Property sub-type. These cases, specifically Exterior Nuisance complaints, overwhelmingly drive the Property group's metrics. Although the number of Private Property calls decreased by around 10% from 2016 to 2017, **the average days open increased from 14.5 to 37.1 days**.



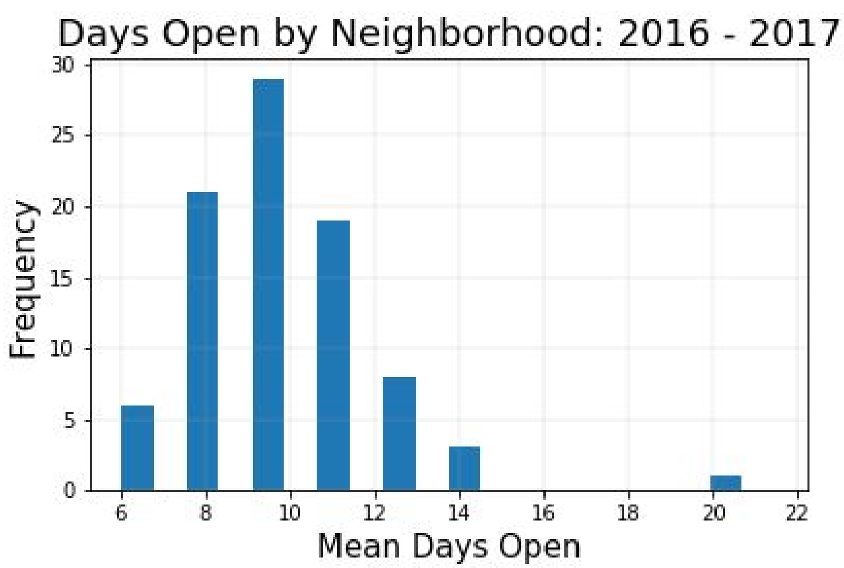
9. Complaint types with the largest change in volume from 2016 to 2017

- Commercial Property: 2868.97%
- Garbage & Recycling: 124.95%
- Graffiti / Clean City: 24.14%
- Parking: 23.80%
- Fire (Fire Rig Visit Request): 22.71%



10. Do case types correlate with neighborhood location?

- The chi-squared test was used to examine the relationship between the Case Type and Neighborhood variables. The test produced a p-value near 0.0, indicating a strong correlation.
- The McKinley neighborhood is an outlier for average case age at 21 days.



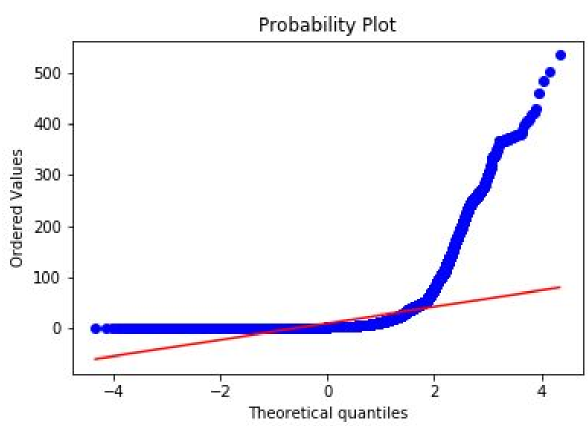
Machine Learning Process:

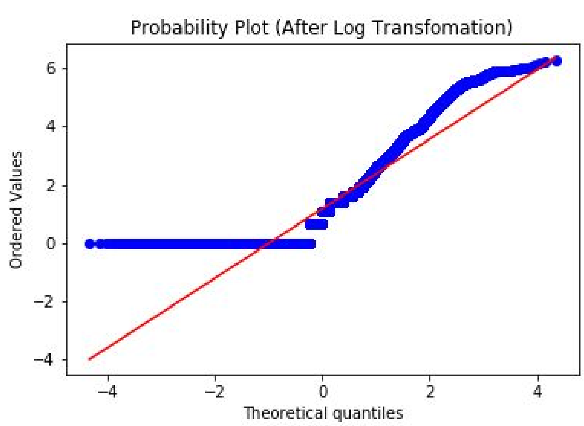
To predict the number of days each case would stay open, three models were tested: **Linear Regression, Ridge Regression,** and **Random Forest Regression**. Model tuning was done on data from 2016 and 2017. When the best algorithm and parameters were determined, the 2018 data was introduced as the final test set.

Pre-Processing

Several steps were taken to prepare the data before it was ready for predictions. The first was to **remove any open cases**. There were only 306 cases (out of 98k) still open from the 2016-2017 data sets. Even if these were left in and given a closed-date, they would likely be outliers. Removing this small percentage of rows seemed more practical.

One major assumption in Linear Regression is that the target variable (days_open) has a normal distribution. After examination with a Q–Q (quantile-quantile) plot, it was apparent that a **log transformation** was needed. This helped, however the high percentage of cases closed in 1 day continues to skew distribution.





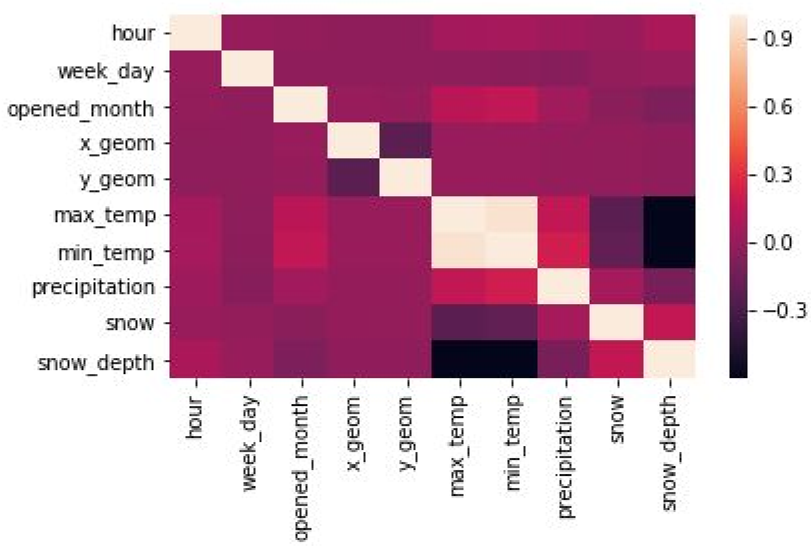
Feature Engineering

Regression models require numerical variables. This data, however, had **categorical features** (case type descriptions) and **temporal features** (Date and Time).

The case type descriptions (title, reason, and subject) are hierarchical and would cause collinearity issues if we fit a model to more than one. With the get_dummies function, we can create **binary columns** for each unique category in the feature. While the title descriptor gives the most detail, it also leaves us with 97 columns after binarization. The reason descriptor (with only 27 binary columns) was also tried, but the increased variance within each category made the model less effective.

From the date and time each case was opened, numerical features were created for **hour**, **week_day** (0-6 represent Monday-Sunday), and **opened_month**.

To avoid collinearity among the **numerical features**, a correlation matrix was examined. The minimum daily temperature (min_temp) had a strong correlation with max_temp and a moderate correlation with opened_month. Similarly, snow_depth had a moderate correlation with min_temp and max_temp. Neither snow_depth or min_temp were used in the final model. The best results were found when modeling with the following features: **title, hour, week_day, opened_month, x_geom, and y_geom**. These features can all be found in or created from the original datasets pulled from the Minneapolis 311 API.



After feature engineering, the training and test variables were aggregated up to the hour (creating 11412 rows). A better performance was seen with aggregation using the sum rather than with the mean.

Model Evaluation

To predict how long cases would stay open, three models were tested: **Linear Regression**, **Ridge Regression**, and **Random Forest Regressor**. The best score for each algorithm was compiled after tuning parameters, trying different standardization functions and feature combinations.

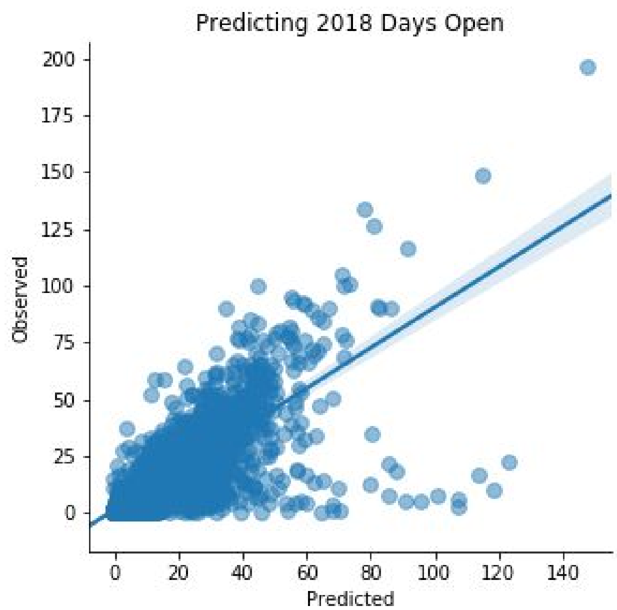
Model	R-2	Parameters
Linear Regression	0.848	RobustScalar was used
Ridge Regression	0.85	Alpha = 10
Random Forest Regressor	0.833	N_Estimators = 500

The results show that Ridge Regression performed best, with an accuracy of 85% (R-2 value). Ridge Regression utilized a ‘regularization’ technique to adjust for over-fitting which often occurs when there is a large number of features. Indeed, it appeared to help model the 97 columns of title features.

Final Predictions

The 2018 data was held out while validating and tuning the data. Each year, there’s a need for some case types to be introduced or retired. Before modeling with the 2018 data, cases with a handful of non-overlapping titles were removed.

Finally, the Ridge Regression model was fit to the combined 2016-2017 data. **Predictions made on 2018 cases had a 63.4% accuracy rate.** We can see there was still some over-fitting present.



Client Recommendations

- Expand the range of cases that can be reported online and through the mobile app and build public awareness of these platforms. These channels can be more automated and require fewer staff hours to monitor.
- High volume/impact case types to address and improve process around: Exterior Nuisance Complaints, Hoarding, Sidewalks - Snow and Ice complaints.

Future Considerations

To improve the predictive model, additional parameter tuning could be performed with the Random Forest Regression. Outliers and influential points would be examined and possibly removed. Other models to try would be Lasso Regression and Extreme Gradient Boosting.

At the neighborhood level, further research could explore which calls are more common in each area. It would also be interesting to examine how average income or other factors correlate with differences in how quickly calls are resolved per neighborhood.