Exploratory Data Analysis and Visualization Final Project - NYC 311 Complaints

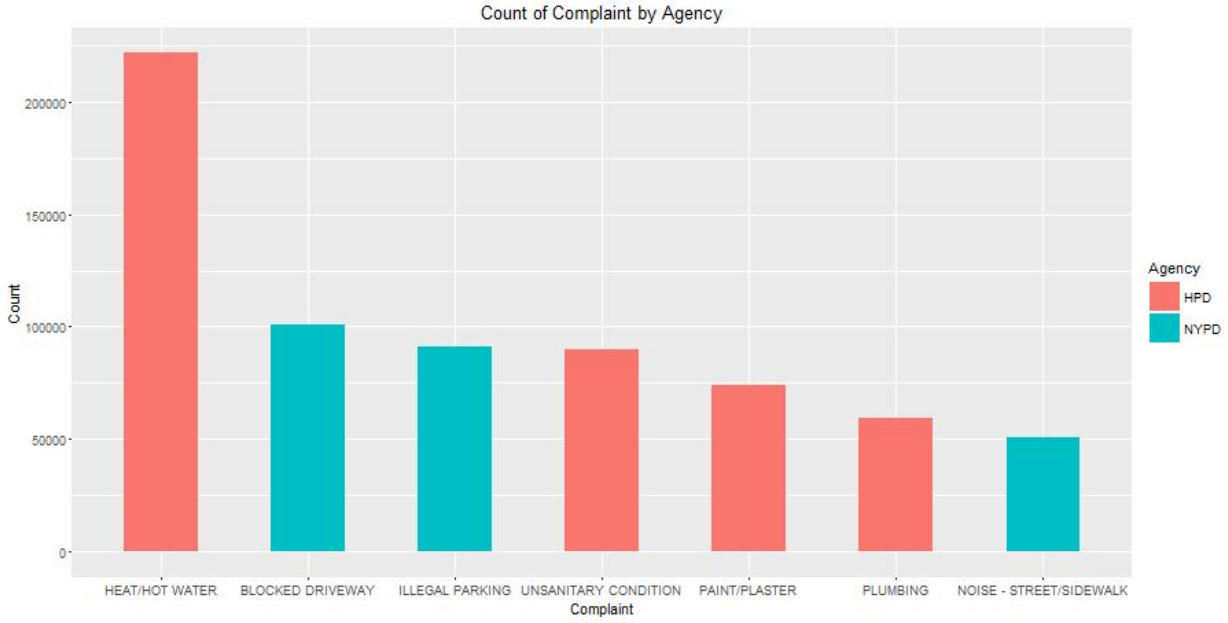
Abstract

Our analysis of the 311 complaints data was focused on determining the factors that played into variance in response time across different complaint types. We were interested in which of these factors might play the largest role in causing an increase in response time, so that we might be able to draw conclusions on how best to handle the negative effects of this variable. To investigate response time we chose the top 7 complaints by count, and looked into several independent variables from the original dataset, as well as from other datasets from alternative sources. We subsequently formed several predictive models that took into account these various factors, and allowed us to predetermine a response time given certain initial conditions for the complaint. Please visit http://complaints.ianjohnson.co for a helpful visualization of our 2015 subset of the NYC 311 complaints data.

Data manipulation

There were several steps that we took to massage the data into a more useful form in order to perform our analyses. First off, we chose to use only data from the year 2015, in order to make sure that our analyses were being performed on a consistent subset of the data, and to give us faster running times for further subsetting and data manipulation.

Next, we subsetted to the top 7 complaint types by count, shown by the following graph:



Our intent was to choose those complaint types for which there were the most data points to work with, and to ensure that our model would have enough data for training and testing using cross validation. Each of these complaint types proved to be a good candidate for

further investigation, as brief preliminary research into the criteria for the complaints showed that there are pressing needs for fast resolution of the complaints.

HPD complaints classes and necessary response times:

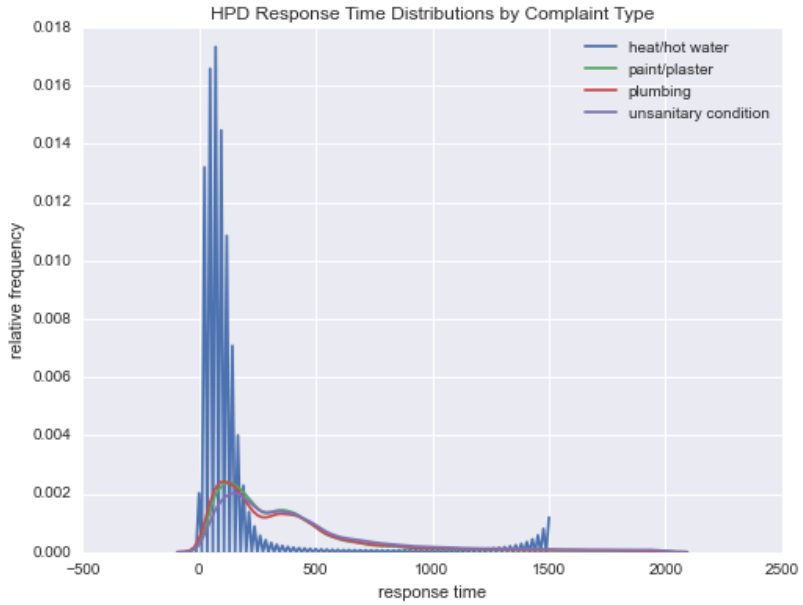
Violation Type	Class	Time owner has to correct
Class A	Non-Hazardous	90 days
Class B	Hazardous	30 days
Class C Lead-Based Paint or Window Guards	Immediately Hazardous	21 days
Class C Heat and Hot Water Violations	Immediately Hazardous	Immediately
Class C (all others)	Immediately Hazardous	24 hours

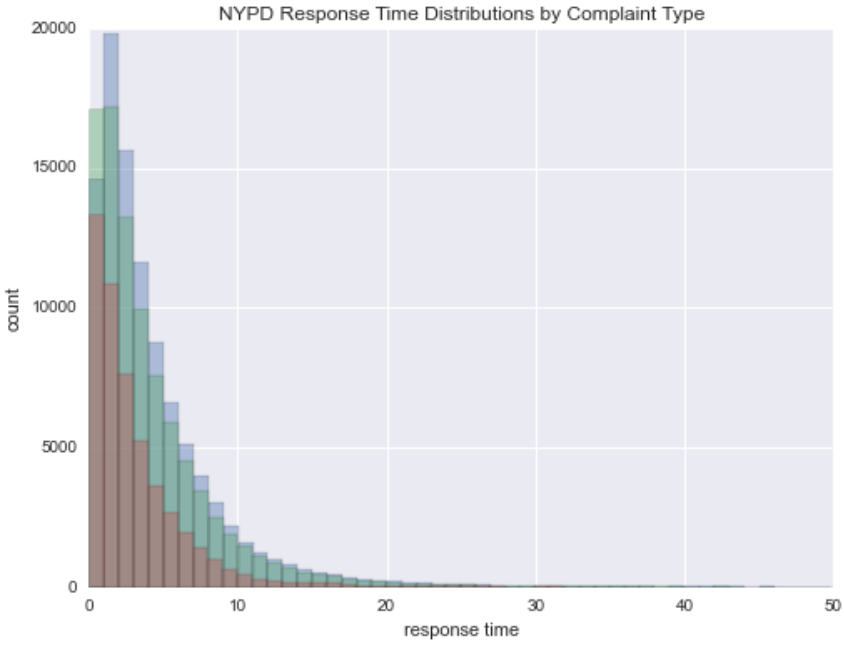
We can see from this table that the HPD has mandated that heat and hot water complaints should be resolved “immediately” by the owner, and thus this complaint type is a good candidate for further investigation. In addition, for noise complaints, because the NYC loud noise ordinance is in effect from 10pm thru 7am, any response time that ends up falling outside of this time window would be ineffectual to the caller, and would necessitate some investigation into how to decrease this response time. This subsetting also left us with data from just two of the seventeen total agencies, namely the Department of Housing Preservation and Development, and the New York City Police Department.

To form the calculation for our response time variable, we subtracted the *Closed.Date* colum from the *Created.Date* column in the dataset, and added a new Response_Time variable for the total response time for the call in minutes. This column would serve as our independent variable for further analyses and modeling. Final data cleanup involved removing rows with NA values, using substring methods on the zipcode column to remove extraneous characters from the end of some zip codes, and leaving out certain complaints with a zero response time, which was usually an indication that a response to these calls was never initiated.

With these subsetting criteria on date, complaint type, and agency, along with our calculated predictor variable and some cleaning of the data, we began our analysis based on factors included in the dataset, along with additional data drawn from outside data sources.

Exploratory Analyses
Distribution of Response Times by Agency



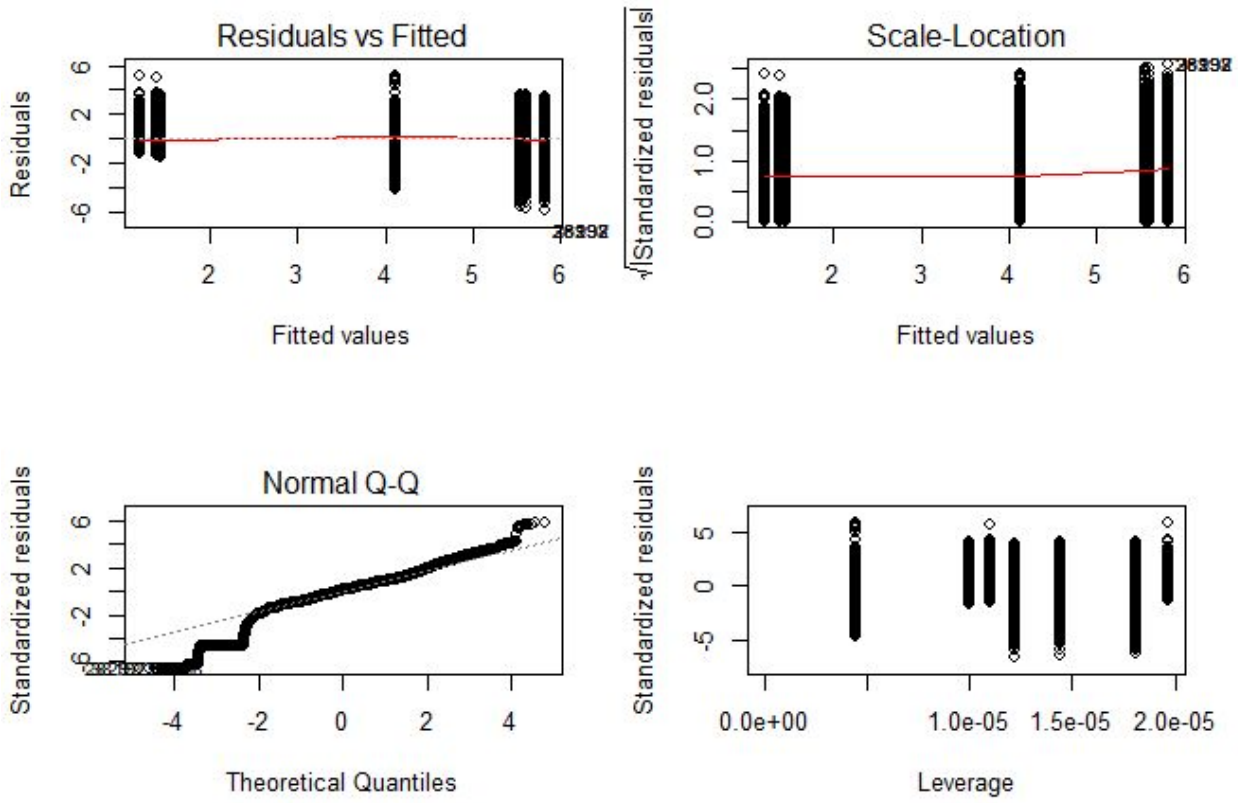


ANOVA for response time by complaint type

We applied one-way Analysis of Variance (ANOVA) to test at a rejection level of 0.05 the null hypothesis that different complaint types have the same mean response time against the alternative hypothesis that at least one pair of complaint types has different mean response times. This test is done on the assumption that the response times are derived from a Gaussian distribution with an unknown but fixed variance. This test results in a p-value of < 2 exp(-16).

Therefore, we can reject the null hypothesis, as the different complaint types likely have different average response times.

Information about the ANOVA model diagnostics is provided below:



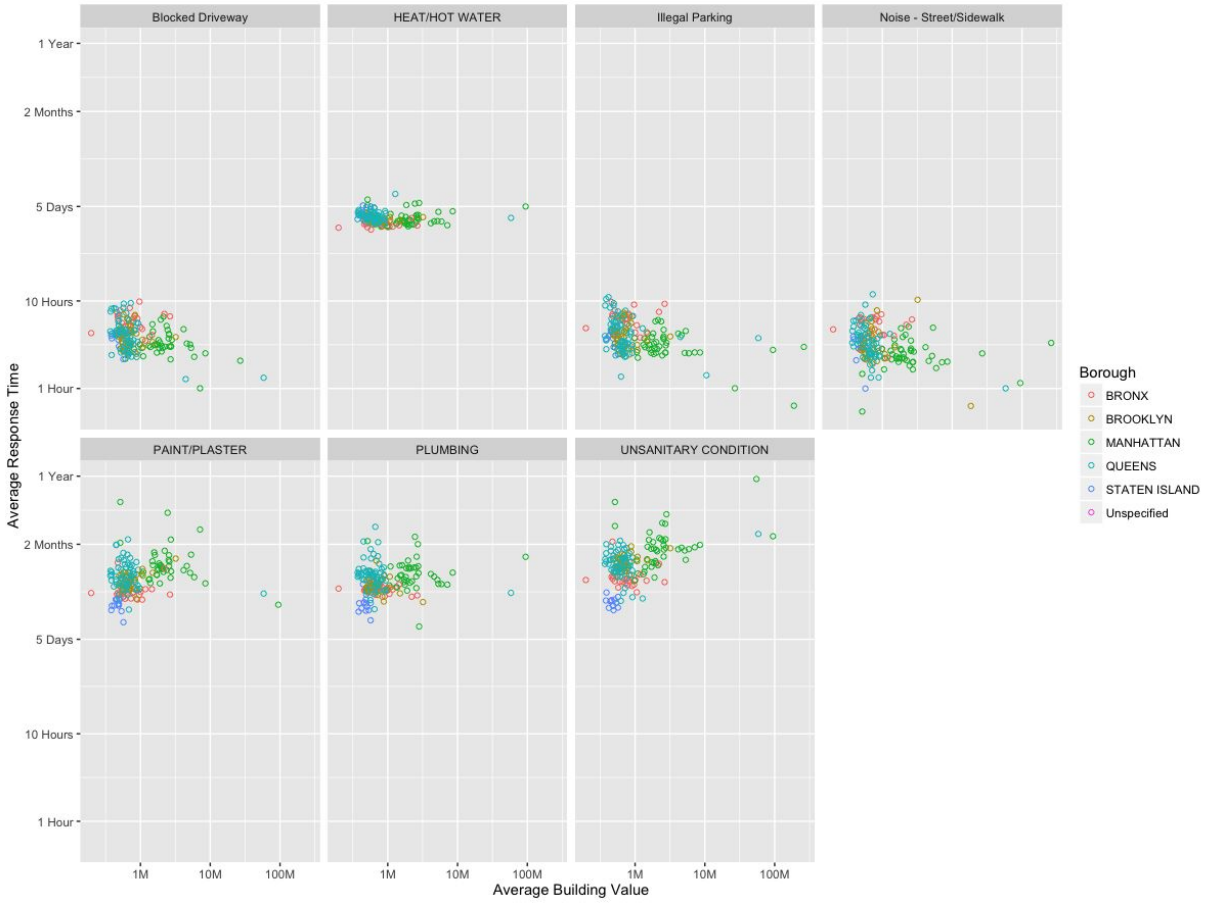
The plot of the residuals against the fitted model values shows that the residual errors stay in the same range as the fitted values change, so that the error terms have a constant or almost-constant variance, and that homoscedasticity is a valid assumption. The normal Q-Q diagram demonstrates that the residual data, when standardized, approximates a standard Gaussian distribution. The visualization of the standardized residuals vs. their leverage values shows that the data points have the same range of residual error for different leverages.

This ANOVA model also gives an Akaike information criterion (AIC) value of 1,741,752 and a Bayesian information criterion (BIC) value of 1,741,844. Even though these values are very high, this result may be due to running ANOVA on one variable only. Nevertheless, the results of the ANOVA modeling demonstrate that there is justification for exploring and modeling on complaint types separately.

External Demographics Data

In exploring outside datasets to for a predictive model for response time, we hypothesized that there might be a decrease in response time as general measures of wealth in the location where the call originated increased. Our initial assessment of the effects of income and general wealth on response time was based on information gathered from the Department

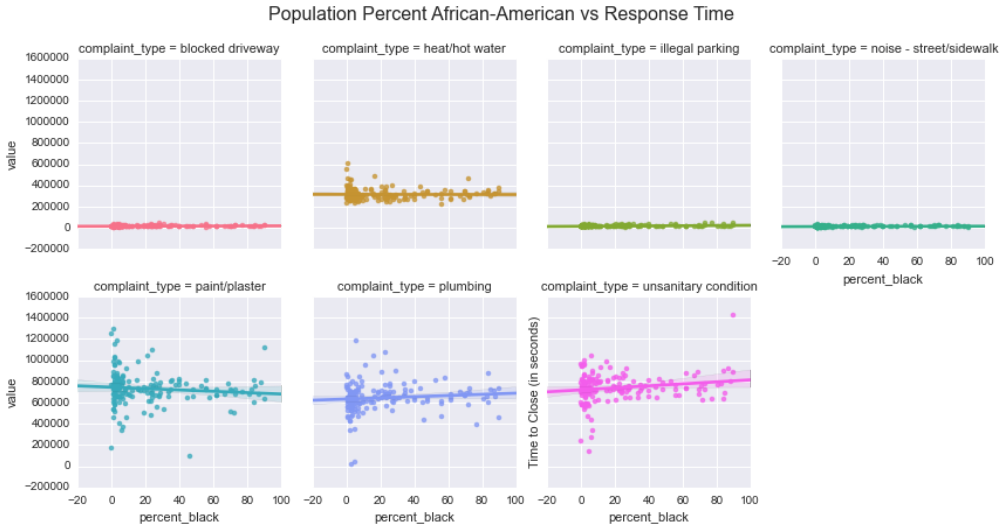
of Finance on building values in NYC. We averaged these building values across zip codes and boroughs and included these results in the following graph:

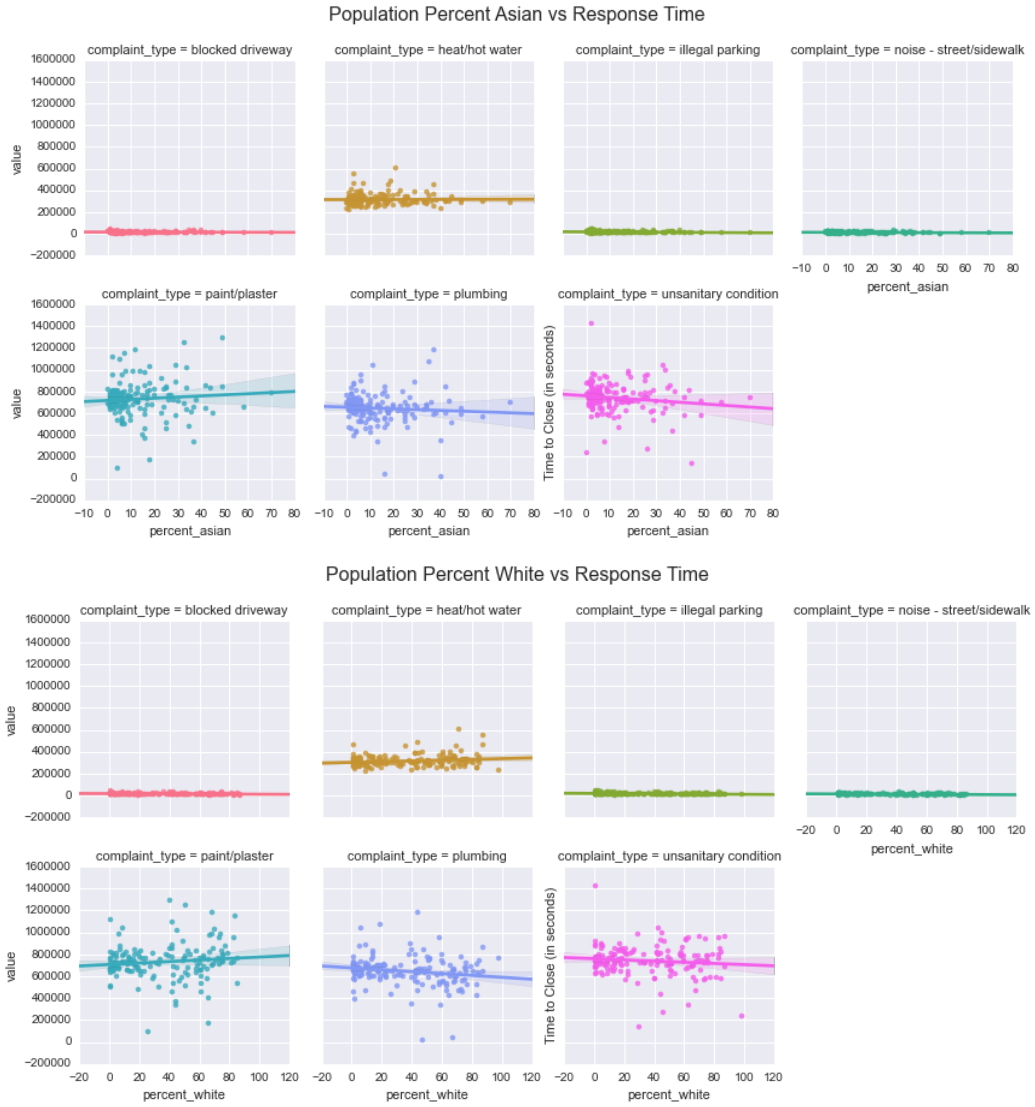


Our results showed a small but plausible correlation between average building value and response time across zipcodes. In particular, there was a slight downward trend for outliers in manhattan in the Paint/Plaster, Illegal Parking, and Blocked Driveway complaints. This relationship may have been influenced by most of these zipcodes being in the Manhattan borough, which is indicated by the color of the points on the graph. This suggests a relationship between borough and zip code, and is further supported by the lower response times across the board for all complaint types in the borough of Staten Island. We also noted that there was a large discrepancy between the four HPD complaint types and the three NYPD complaint types in terms of response time, which is likely due to the relative difficulty of the complete resolution of these complaints.

Because of the lack of significant correlation between these variables from the DOF dataset, as well as the possibility that average building value is not an accurate predictor for general wealth of an area, we sought information from additional data sources including Zillow and data from the US Census Bureau. Zillow did not have enough housing price information for NYC zipcodes, and so we use the average income data from the US Census bureau and came up with the following relationships between average income and response time:

From this new data, we notice a slight correlation between average income and response time across several of the complaint types including Paint/Plaster, Plumbing, and Unsanitary Condition. This slight correlation is an overall increasing relationship, which is contrary to our initial hypothesis that this would be an inverse relationship. One complaint type, Noise - Street/Sidewalk, did show this negative correlation, however it was only very slight, and we were not convinced that predictions from this independent variable would be accurate. We nonetheless took note of this variable for inclusion in our predictive model, and decided to pursue further analyses on other outside datasets, including other NYC demographics data such as race and gender.

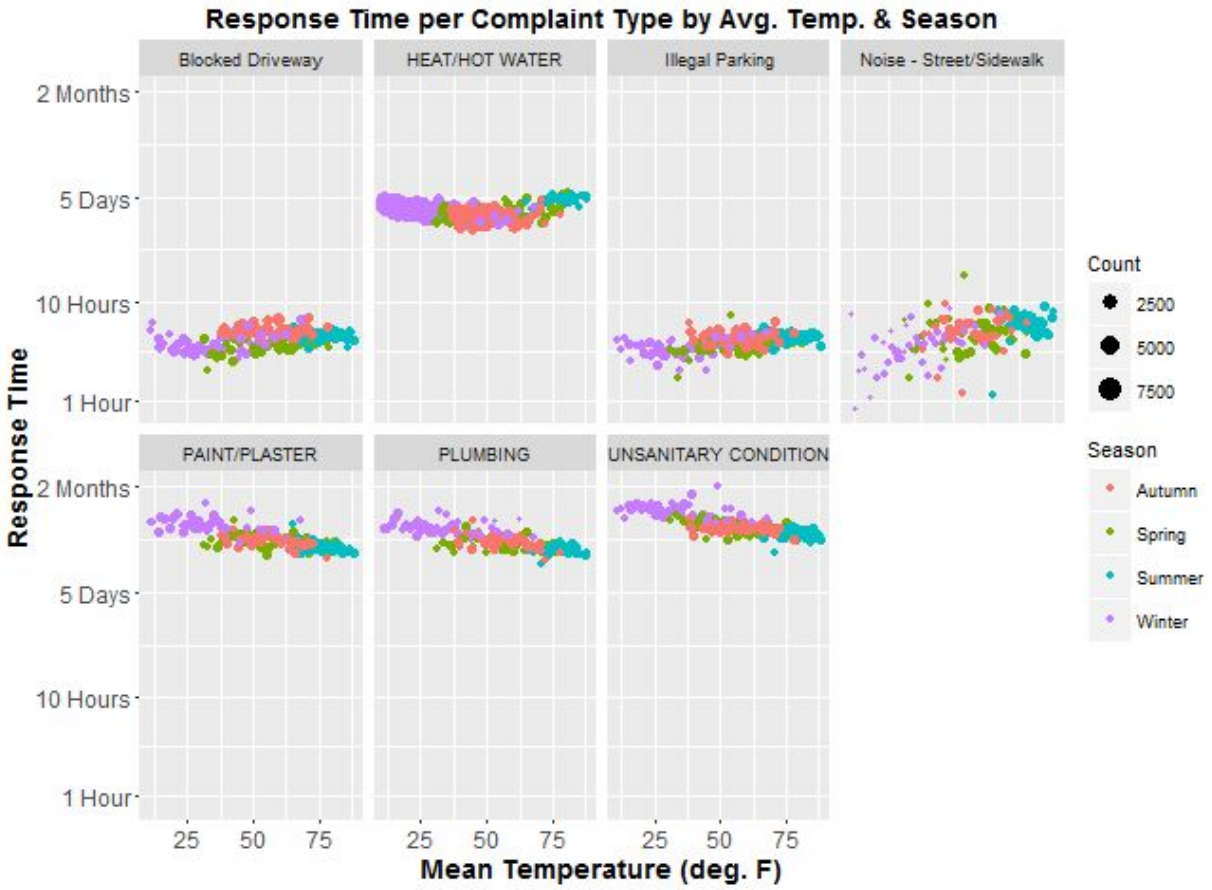




As the above charts indicate, the racial characteristics of the zip codes where complaints originate from appeared to have little effect on complaint response time. In these charts response time for each complaint type is plotted against the proportion of each race in the zip code from which the complaint originated. Though some of the HPD-related complaints have slightly non-zero slope, it's nearly flat and there's significant uncertainty. Interestingly though, the sign of the slope for HPD complaints does flip across races, hinting at the possibility that race could be a proxy for some other meaningful feature.

External Weather Data

Several features associated with weather were also analyzed. Some key variables used included temperature, humidity, wind speed and visibility. We hypothesized that response time may be higher for more inclement weather (hot/cold temperatures, high humidity, more precipitation, etc.).

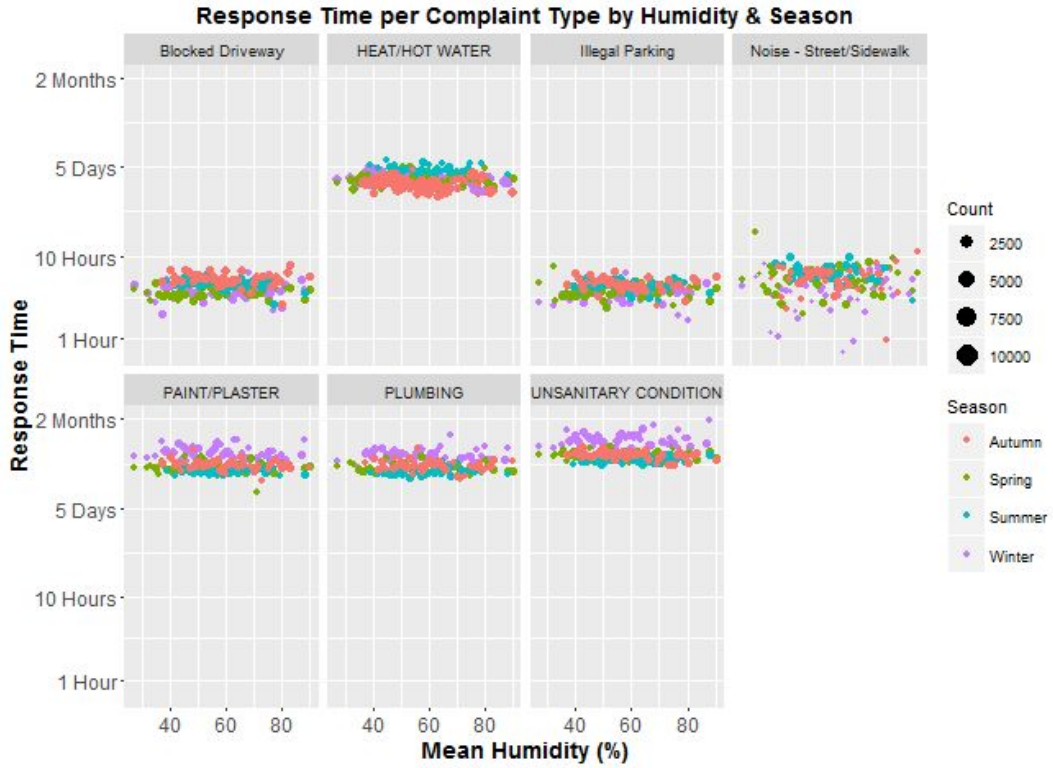


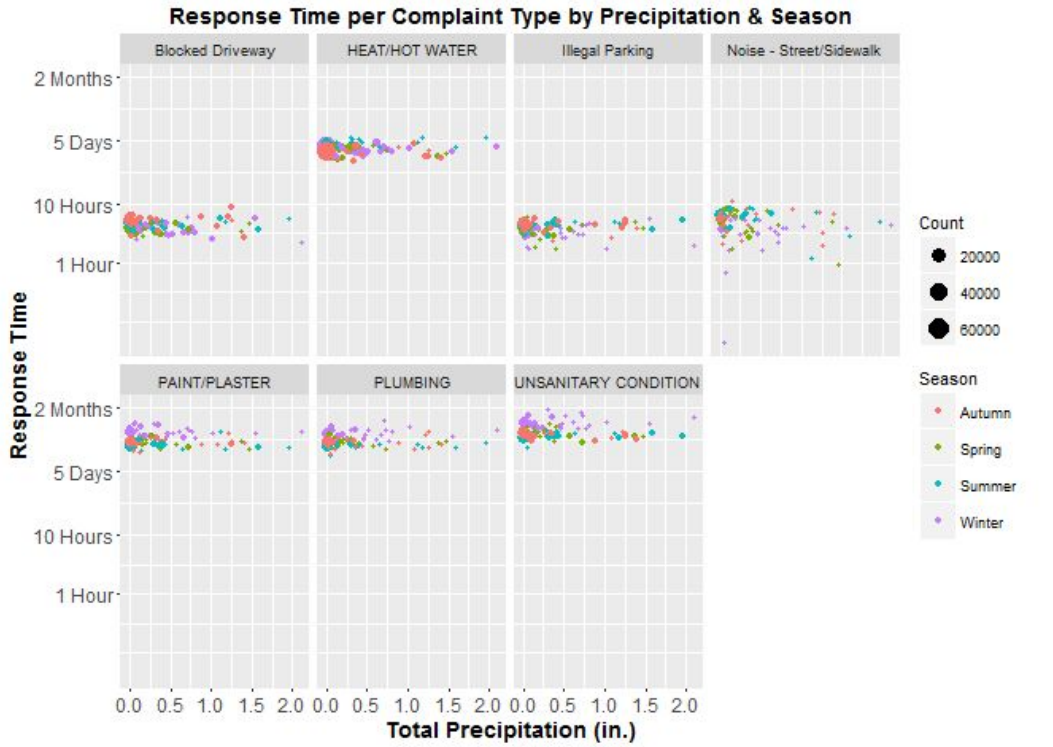
For each given temperature value, we plotted the average response time over all complaints. Issues for which the NYPD was responsible took more time with higher temperatures, while housing-related complaints (except for heat and hot water) were solved more quickly as the temperature increased. The NYPD's response time may be due to more violent crimes occurring during warmer months, possibly reaching or exceeding the capacity of the police's resources. Less serious cases of noise complaints and illegal parking would then have a lower priority to resolve. Heat and hot water complaints are given highest priority during the winter months, when the impact of not having heat and hot water will be the most severe. As such, most of these cases are resolved within about five days. Other complaints are given a lower priority in the winter, so that they can be resolved more quickly during the warmer months.

Response time does not have any apparent relationship with either humidity or precipitation amount, even when controlling for the season of the year. This finding is a good sign because ideally public agencies can respond and resolve complaints at the same rate, independent of (non-extreme) weather events or the time of the year.

Note that most complaints in 2015 occur on days with little to no precipitation. According to the Weather Underground, that year had 250 days with zero or trace precipitation in Central Park (67 in Autumn, 61 in Spring, 74 in Summer, and 48 in Winter), and only 40.93 total inches of precipitation. This amount is below the average of 54.56 inches for the previous 10 years (2005-2014) in Central Park, and the average of 52.16 inches for the previous 5 years

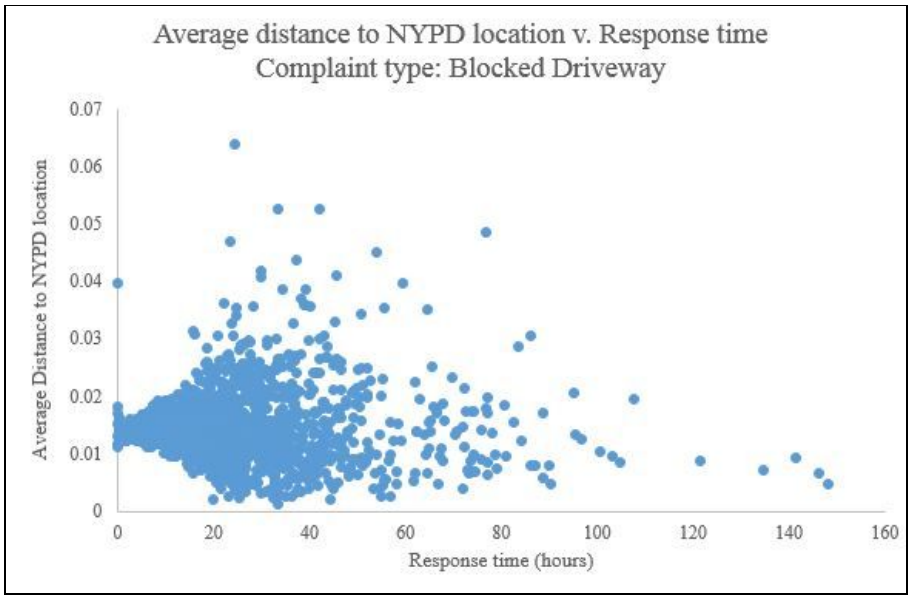
(2010-2015), according to Weather.gov (http://www.weather.gov/media/okx/Climate/CentralPark/monthlyannualprecip.pdf). Since 2015 was a particularly dry year, its weather patterns are not the most representative for the time frame of the 311 dataset.



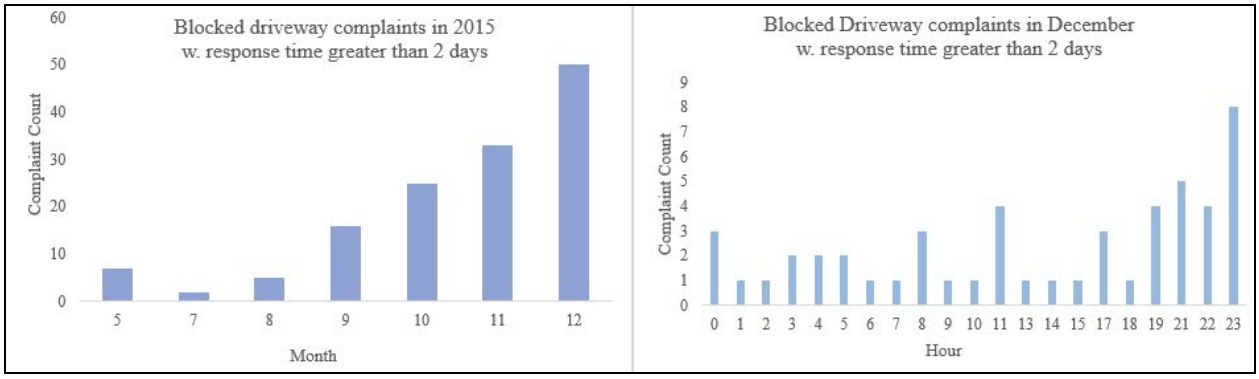


NYPD Blocked Driveway Complaints and Distance to NYPD Precincts

One hypothesis we had was that complaints closer to agency locations would be resolved faster. To explore this hypothesis, we calculated the nearest distance from complaint location to NYPD precincts. We also calculated the response time for each complaint, i.e. the time it was opened to the time it was closed. The results were plotted to get a sense of the distribution.



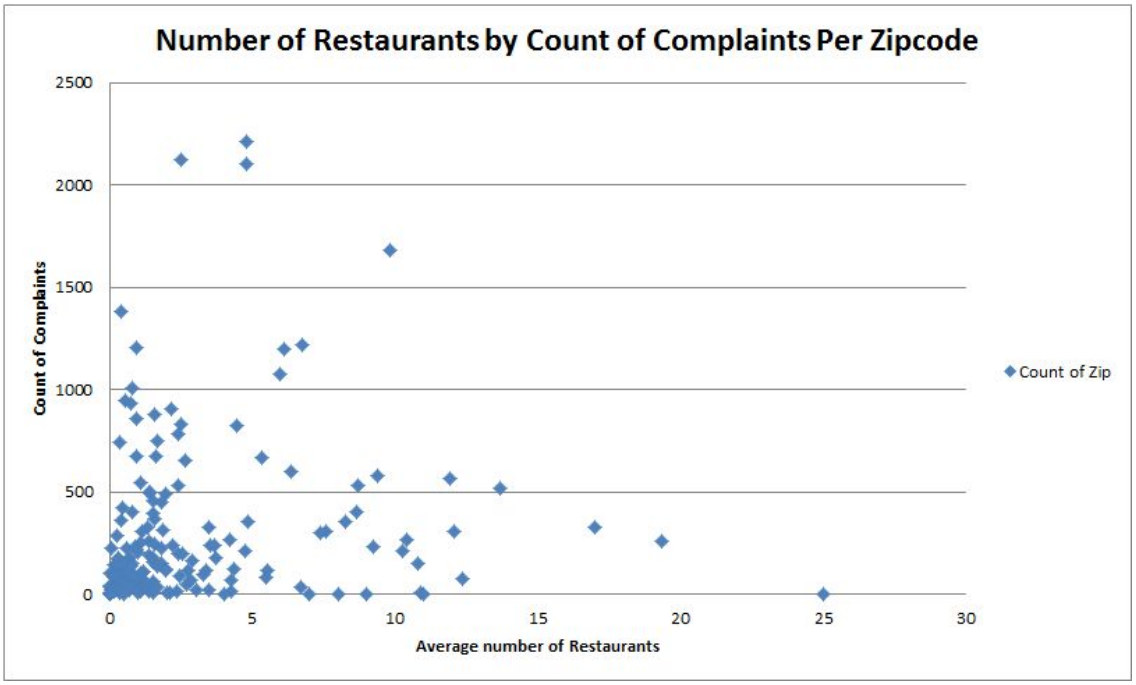
Most NYPD complaints were closed in a short amount of time. We discovered that most of these complaints for 'Blocked Driveway' therefore do not depend on the distance to the police station. However, a counterintuitive observation was that a number of complaints taking greater than 24 hours to resolve, were less than 0.03 units from the police station. This could be because action for these complaints might be driven by mobile units that are in reality further away than the office location. Similar results were observed for 'Illegal Parking' complaint type. Further analysis showed that most of the complaints with response time greater than 2 days were associated with the time when the complaint was registered. These complaints are mostly in the later part of the year, especially in December. Moreover, these complaints are mostly registered at night time. This reveals that the time of complaint, not distance is a driver of response time.



Since there seems to be no immediate relationship between the response time and distance, this feature will not be included in the model.

NYPD Noise Complaints and Distance to a Bar

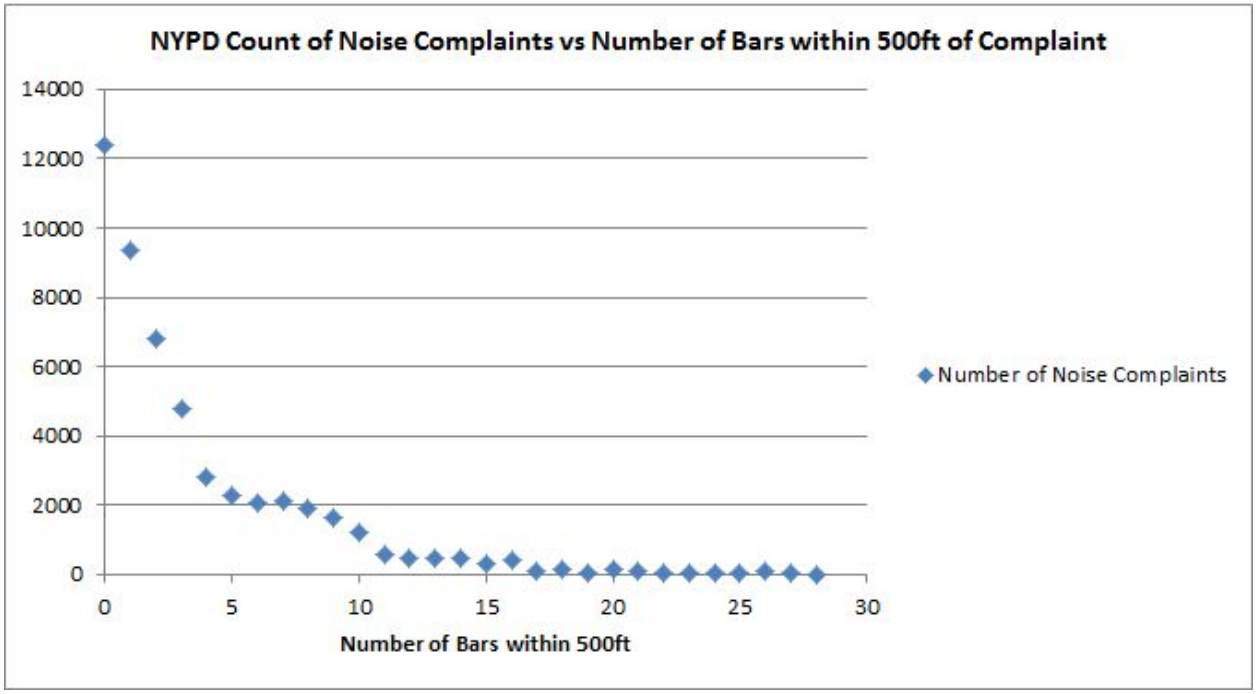
Another hypothesis we wanted to consider was to what effect does the distance to the nearest bar contribute to the existence and frequency of noise complaints. Using information from the NYC Health Grades dataset we were able to compile a list of bars in NYC. Along with locations classified as serving primarily “bottled beverages,” we used a string search to compile a list of locations that contain the following words: bar, pub, tavern, beer, shot, whiskey, wine. We then geocoded each location to obtain longitude and latitudes and grouped the bars by zip code. The following graph shows the relationship between the number of bars per zip code and the frequency of noise complaints within the zip code.



There does not seem to be any significant relationship between these variables as we see a number of zip codes with a significant number bars in the area but very few complaints as well as locations with few bars in the area but many complaints.

Next we calculated the number of bars within 500ft of each noise complaint to see if locations with a higher number of surrounding bars have more noise complaints. Based on the graph below it does not seem as if noise complaints are generated in locations with a significant number of bars as locations with zero bars or one bar had the highest frequency of complaints. The trend looks negative but this is most likely due to the fact that many locations are not surrounded by more than a few bars and there is most likely some other variable contributing to the frequency of complaints. Also, many of the bar locations are located in manhattan while the

noise complaints are dispersed throughout the boroughs, so the frequency of noise complaints with zero or one bar is going to be significantly higher.



As a result of these exploratory analyses, we concluded that some features influence response time more than others. We decided to separate the predictive modeling based on complaint type, and for each complaint we tried various combinations of features to try to predict response time.

NYPD Prediction Models:

Prediction Model for Blocked Driveway

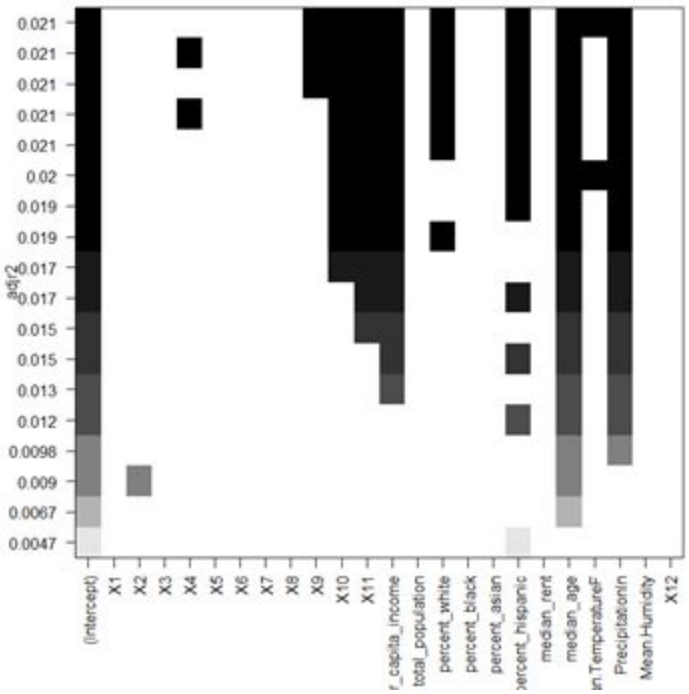
Based on the exploratory analyses conducted earlier, we understand that the month in which the complaints are opened have an impact on the response time. As a result, it is important to use these variables in the prediction model. Dummy variables corresponding to the 12 months of the year are made using one-hot encoding from the ‘Created Date’ features in the 311 data set. These dummy variables will be used in the subsequent prediction model for this complaint type. Additionally demographic and weather data will also be used to test for any possible relationship.

First, a linear regression model was applied. The results indicate that the model has a high deviance, suggesting that the model is not a very good fit. However, most of the features tested are significant.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
 -6.349   -2.864   -1.321    1.272  142.308

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.596e+00  4.604e-01  18.673  < 2e-16 ***
X1                -1.403e+00  9.529e-02 -14.720  < 2e-16 ***
X2                -1.517e+00  1.000e-01 -15.167  < 2e-16 ***
X3                -1.502e+00  8.276e-02 -18.149  < 2e-16 ***
X4                -1.577e+00  8.513e-02 -18.522  < 2e-16 ***
X5                -1.160e+00  9.053e-02 -12.814  < 2e-16 ***
X6                -9.409e-01  9.093e-02 -10.348  < 2e-16 ***
X7                -9.391e-01  1.062e-01  -8.844  < 2e-16 ***
X8                -9.452e-01  1.074e-01  -8.802  < 2e-16 ***
X9                -6.470e-01  9.710e-02  -6.663 2.69e-11 ***
X10               -2.781e-01  7.818e-02  -3.557 0.000375 ***
X11               -2.393e-01  7.459e-02  -3.209 0.001333 **
X12                       NA         NA      NA       NA
per_capita_income -6.695e-06  2.611e-06  -2.564 0.010337 *
total_population  -5.756e-06  7.220e-07  -7.973 1.56e-15 ***
percent_white     -2.171e-02  3.556e-03  -6.106 1.02e-09 ***
percent_black     -2.147e-02  3.744e-03  -5.733 9.90e-09 ***
percent_asian     -4.321e-02  4.122e-03 -10.481  < 2e-16 ***
percent_hispanic  -1.704e-02  3.642e-03  -4.677 2.91e-06 ***
median_rent       -1.899e-03  1.496e-04 -12.694  < 2e-16 ***
median_age         4.850e-02  5.747e-03   8.439  < 2e-16 ***
Mean.TemperatureF  7.467e-03  2.451e-03   3.047 0.002311 **
PrecipitationIn    1.370e-01  6.455e-02   2.123 0.033787 *
Mean.Humidity     -7.419e-03  1.603e-03  -4.629 3.67e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 26.15521)

    Null deviance: 2679299  on 100113  degrees of freedom
Residual deviance: 2617901  on 100091  degrees of freedom
  (12 observations deleted due to missingness)
AIC: 610913

Number of Fisher Scoring iterations: 2
```

In order to improve the fit of the model, forward stepwise regression was applied. From the figure below, we can see that 10 out of the 23 features should be included in the model.

Figure: Adjusted R-squared forward stepwise selection



The linear regression model was applied to the 10 features identified above. The results indicate that out of the 10 variables shown above, only X9, X10, X11, per capita income, median age and mean temperature are significant features.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
 -5.365   -2.909   -1.404    1.226  143.781

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.618e+00  2.073e-01  17.453  < 2e-16 ***
X9                2.887e-01  6.133e-02   4.707 2.52e-06 ***
X10               7.798e-01  5.668e-02  13.758  < 2e-16 ***
X11               8.208e-01  5.603e-02  14.649  < 2e-16 ***
per_capita_income -2.088e-05  1.745e-06 -11.967  < 2e-16 ***
percent_white     -1.223e-03  8.552e-04  -1.431   0.1525
percent_hispanic   1.399e-03  1.027e-03   1.363   0.1730
median_age         1.086e-02  5.308e-03   2.046   0.0407 *
Mean.TemperatureF  1.472e-02  9.083e-04  16.200  < 2e-16 ***
PrecipitationIn    1.355e-02  5.438e-02   0.249   0.8032
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 26.51482)

    Null deviance: 2679309  on 100114  degrees of freedom
Residual deviance: 2654266  on 100105  degrees of freedom
  (11 observations deleted due to missingness)
AIC: 612273

Number of Fisher Scoring iterations: 2
```

The model is not a good fit and the variables have a weak effect on the response time given the small coefficients. However the model uncovers general trends for resolution of blocked driveway complaints. If a complaint is made in September, October or November, the resolution time is 12 to 48 minutes longer than for other months. Greater income levels correspond to quicker response time - for an increase of $1,000 in per capita income, the response time improves by 1.2 minutes. Higher temperatures lead to slower response times as well. For a 10 degree Fahrenheit increase in temperature, the response time declines by 8 minutes.

The prediction model can be further improved by incorporating variables indicating the day complaints were registered in addition to the month. From data exploration we know that there are more complaints with longer resolution time that are registered at the end of December. However this would complicate the model by adding dummy variable for each combination of date and month.

Prediction Model for Illegal Parking

The following explanatory variables describing demographic and weather conditions were used in modeling response time for illegal parking complaints:

'per_capita_income', 'total_population', 'percent_white', 'percent_black', 'percent_asian', 'percent_hispanic', 'median_rent', 'median_age', 'Mean.TemperatureF', 'Mean.Humidity', 'PrecipitationIn'

We first ran a correlation analysis to see which of the variables exhibited collinearity.

Correlation Matrix	per_cap_inc	tot_pop	%_white	%_black	%_asian	%_hispanic	med_rent	med_age	Mean.Temp	Mean.Humid	Precip
per_cap_inc	1.0000	-0.3774	0.5425	-0.2539	-0.0669	-0.4038	0.8560	0.2819	0.0012	-0.0033	-0.0036
tot_pop	-0.3774	1.0000	-0.1825	0.1016	0.0541	0.1270	-0.3435	-0.2805	0.0062	0.0113	0.0046
%_white	0.5425	-0.1825	1.0000	-0.6192	-0.0196	-0.5831	0.4685	0.4367	-0.0255	-0.0048	-0.0018
%_black	-0.2539	0.1016	-0.6192	1.0000	-0.4535	-0.0398	-0.3260	-0.2525	0.0317	0.0087	0.0051
%_asian	-0.0669	0.0541	-0.0196	-0.4535	1.0000	-0.1750	0.1242	0.2114	-0.0160	-0.0015	-0.0087
%_hispanic	-0.4038	0.1270	-0.5831	-0.0398	-0.1750	1.0000	-0.3700	-0.4540	0.0074	-0.0021	0.0029
med_rent	0.8560	-0.3435	0.4685	-0.3260	0.1242	-0.3700	1.0000	0.2243	-0.0122	-0.0041	-0.0013
med_age	0.2819	-0.2805	0.4367	-0.2525	0.2114	-0.4540	0.2243	1.0000	-0.0077	-0.0108	-0.0123
Mean.Temp	0.0012	0.0062	-0.0255	0.0317	-0.0160	0.0074	-0.0122	-0.0077	1.0000	0.1627	0.0073
Mean.Humid	-0.0033	0.0113	-0.0048	0.0087	-0.0015	-0.0021	-0.0041	-0.0108	0.1627	1.0000	0.5167
Precip	-0.0036	0.0046	-0.0018	0.0051	-0.0087	0.0029	-0.0013	-0.0123	0.0073	0.5167	1.0000

From this correlation matrix we decided to exclude %_white, median_rent, and Precipitation. %_white is naturally calculated based on the %_black, %_asian, %_hispanic numbers and is also moderately correlated with per_capita_income. Median_rent is highly correlated with per_capita_income and humidity and precipitation show a moderate correlation to each other.

Using this subset of factors we ran a generalized linear model, fitting a Poisson distribution to the response time variable.

Generalized Linear Model Regression Results

Dep. Variable:	y	No. Observations:	77060
Model:	GLM	Df Residuals:	77052
Model Family:	Poisson	Df Model:	7
Link Function:	log	Scale:	1.0
Method:	IRLS	Log-Likelihood:	-1.8491e+05
Date:	Sun, 08 May 2016	Deviance:	1.5673e+05
Time:	15:24:48	Pearson chi2:	1.62e+05
No. Iterations:	8		

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
x1	-1.561e-06	1.17e-07	-13.288	0.000	-1.79e-06	-1.33e-06
x2	1.212e-06	7.7e-08	15.747	0.000	1.06e-06	1.36e-06
x3	0.0016	0.000	15.790	0.000	0.001	0.002
x4	-0.0032	0.000	-17.949	0.000	-0.004	-0.003
x5	0.0044	0.000	39.619	0.000	0.004	0.005
x6	0.0233	0.000	78.535	0.000	0.023	0.024
x7	0.0028	0.000	25.379	0.000	0.003	0.003
x8	0.0015	0.000	9.996	0.000	0.001	0.002

x1 = 'per_capita_income', x2 = 'total_population', x3 = 'percent_black', x4 = 'percent_asian', x5 = 'percent_hispanic', x6 = 'median_age', x7 = 'Mean.TemperatureF', x8 = 'Mean.Humidity'

Based on the output from the linear model all the features are deemed to be significant, however the the model as a whole does not help to explain a significant portion of the variability in the data as the R^2 value is near zero. Additionally, a holdout set was used to test the model and the mean residual value was calculated and compared to residuals from predicting average response time. The model performed only slightly better than using the average. Variations to the generalized linear model with different sets of features were performed and these too did not perform better than the base model described above.

We next used a Decision Tree Regression model on the same set of features. Using 100 estimators, with splitting decisions based on mean squared error minimization we were able to come up with a model with an R^2 of approximately 51%, meaning that 50% of the variability in the data could now be explained using this model. This was a significant improvement from the linear regression previously observed, and a 5 minute improvement from using the average approximation. Based on the feature scores humidity and temperature seem to be the main drivers of response time.

Features sorted by their score:
(0.3612, 'Mean.Humidity'),
(0.3469, 'Mean.TemperatureF'),
(0.0673, 'per_capita_income'),
(0.0485, 'median_age'),
(0.0481, 'percent_asian'),
(0.0449, 'percent_hispanic'),
(0.0438, 'total_population'),
(0.0393, 'percent_black')

Predicting Response Time for Noise Complaints

To make predictions for the response time of the NYPD for noise complaints, we used a random forest prediction model on our aggregated data from each of the outsourced data sets. This data set consisted of demographic features including ethnic information, average income data, and weather data for each complaint. Our response variable was the response_time column computed from the original dataset.

Our first prediction for noise complaints response time was made using a model trained on 10% of the Noise complaints data, so we could iterate faster with our analysis, and gave us an average error of 2 hours and 2 minutes when tested on the remaining 90% of the data. Analysis of the feature importances for this predictor showed that the features were ordered as follows in terms of largest reduction in mean squared error:

```
> round(importance(fit), 2)
                      %IncMSE  IncNodePurity
Mean.Humidity           40.61      101814.74
Mean.TemperatureF       36.58      113946.05
Precipitation_inches    35.80       31428.92
total_population        24.85       18611.39
median_age              23.39       11880.55
per_capita_income       23.37        32299.42
median_rent             21.78       19004.96
percent_black           20.79       13393.24
percent_asian           18.97       11766.41
percent_hispanic        17.90       12761.56
percent_white           16.90       15838.12
```

The average error when simply always predicting the mean response time for this complaint was 2 hours 37 minutes, and thus our prediction model was more accurate on average than just predicting the mean response time for any data point by 35 minutes. Still, an error of 2 hours was large considering the actual values of the response times, most of which were less than five hours, so we pursued further feature combinations to improve our predictions.

Our next step was to include one-hot encoded columns for the zip codes as new features for the random forest. Our average test error with these columns was 2.49 hours when using only 10% of the training data, 2.41 hours with 50%, and 2.31 hours with 90%. These error rates were not an improvement over our previous predictor when leaving out the zip code, which was confirmed by the presence of several negative reductions in mean squared error corresponding to the one-hot encoded zip code features:

```
> round(importance(rf), 2)
                      %IncMSE  IncNodePurity
per_capita_income       6.67        4745.25
total_population        6.42        3232.92
percent_white           5.25        2350.14
percent_black           5.92        1954.24
percent_asian           6.69         3079.45
percent_hispanic        7.22        2045.08
```

median_rent	7.21	2951.51
median_age	6.18	2111.53
Mean.TemperatureF	9.30	17647.89
Mean.Humidity	4.04	17326.23
Precipitation_inches	5.70	7283.18
pred_data_noise.f10002	3.08	10.89
pred_data_noise.f10003	-1.84	13.90
pred_data_noise.f10004	-2.32	3.35
pred_data_noise.f10005	3.15	2.64
pred_data_noise.f10006	0.00	0.06

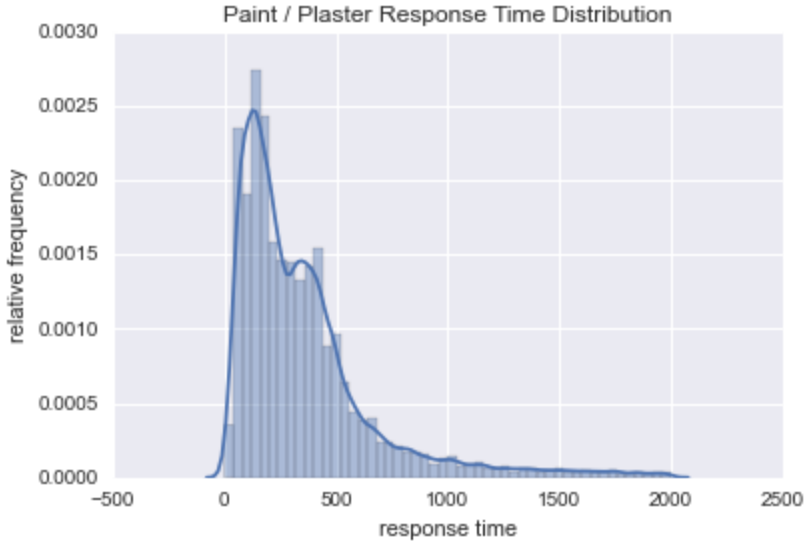
… remaining zip code features truncated ...

Thus, our model for Noise - Street/Sidewalk complaints was able to predict the response time of the NYPD to within 2 hours and 2 minutes on average. We would improve on this by adding additional data to the model, including seasonality, day of the week, time of day, or features from additional outside datasets.

HPD Prediction Models

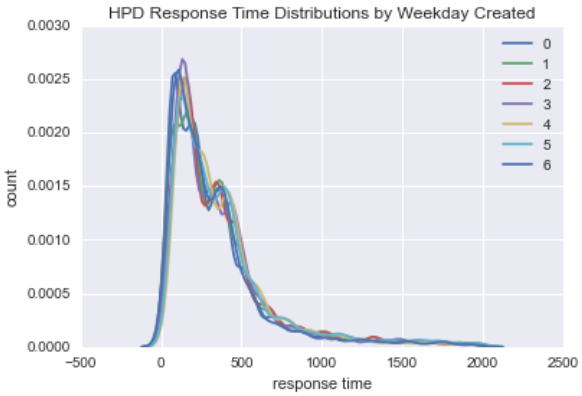
Predicting Paint/Plaster Response Times

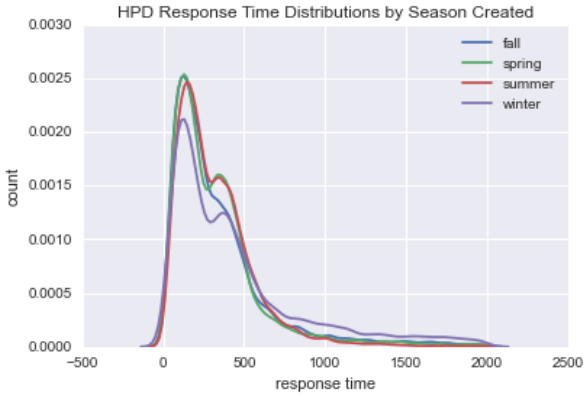
Like the other non-heat/hot water complaints handled by HPD, the distribution of response times the paint and plaster complaint displayed a bimodal distribution with peaks at (approximately) 5 and 15 days and with a long tail of responses times longer than than.

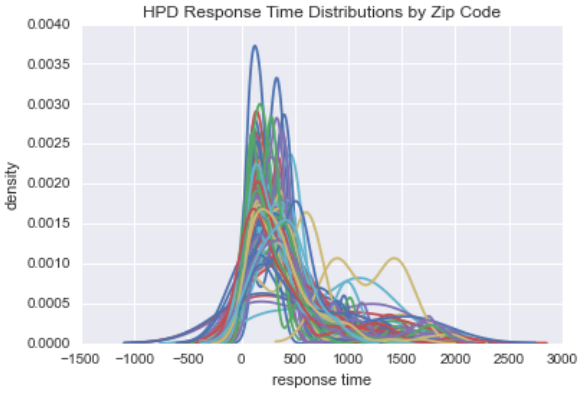


This bimodality hinted at the presence of subpopulations and so prior to conducting the regression analysis we sought to identify whether such populations could be accounted for by features available to us in the dataset. Some hypotheses included: day of the week created (those created near the end of the week could take longer to address), season created (seasonal variation in conditions could affect time), whether distinct complaint types were being lumped into a single category, geographic differences (some locations could be harder-to-reach

and therefore take longer to respond to), etc. Finding features that could account for this variation would be helpful in identifying the features to include in subsequent model.







As evidenced by the above charts, faceting by weekday and season-created changed the shape of the distributions very little, indicating they don't much affect response time. Grouping by zip code on the other hand changed the distributions significantly. Though this might be expected given the smaller size of the groups, upon close examination there seem to be clusters of zip codes at the two modes. This indicated zip could be a useful feature to include.

Our final pre-modeling step was to prepare the final features. We did this two ways. First, we looked for correlations between candidate continuous features to identify columns to exclude from the analysis. Median rent and per capita income were (unsurprisingly) highly correlated, so we decided to exclude the former from our analyses. The proportion of whites in a zip code was also correlated with that of other races, so we excluded that as well. After removing features we then one-hot encoded the zip codes so they'd be represented as continuous variables, which our code was execting.

We tried two different types of regression models: Random Forest Regressor and multivariate OLS. The Random Forest Regressor performed best with a mean error of only 251 hours vs the mean-prediction error (which was 314 hours). Interestingly, the temperature and and humidity were most important features in the model and zip codes had very little importance. Both the RF feature importances and OLS regression gave little significance to the

zip codes. Moreover, the OLS diagnostics showed that population and precipitation did not have a statistically significant effect on response time and were also removed in subsequent iterations.

6	Mean.Temperature F	0.262960
7	Mean.Humidity	0.244822
8	PrecipitationIn	0.105231
4	percent_hispanic	0.087266
0	per_capita_income	0.071364
5	median_age	0.070508
3	percent_asian	0.060407
2	percent_black	0.050580
1	total_population	0.046861

OLS Regression Results

Dep. Variable:	y	R-squared:	0.393
Model:	OLS	Adj. R-squared:	0.393
Method:	Least Squares	F-statistic:	6412.
Date:	Sun, 08 May 2016	Prob (F-statistic):	0.00
Time:	19:04:46	Log-Likelihood:	-5.3522e+05
No. Observations:	69231	AIC:	1.070e+06
Df Residuals:	69224	BIC:	1.071e+06
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
x1	0.0045	0.000	24.943	0.000	0.004	0.005
x2	1.3153	0.129	10.207	0.000	1.063	1.568
x3	2.0061	0.299	6.708	0.000	1.420	2.592
x4	2.1748	0.125	17.401	0.000	1.930	2.420
x5	11.8462	0.435	27.246	0.000	10.994	12.698
x6	-4.5384	0.108	-41.845	0.000	-4.751	-4.326
x7	0.6167	0.157	3.923	0.000	0.309	0.925

Omnibus:	62019.968	Durbin-Watson:	1.950
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2581470.546
Skew:	4.272	Prob(JB):	0.00
Kurtosis:	31.669	Cond. No.	7.18e+03

(using features: 'per_capita_income', 'percent_black', 'percent_asian', 'percent_hispanic', 'median_age', 'Mean.TemperatureF', 'Mean.Humidity')

Predicting Heat/Hot Water Response Times

There are over 220,000 heat or hot water 311 cases in 2015. That data was partitioned into three sets: 20% for training the predictive model, 40% for holdout validation, and 40% for testing. This breakdown was applied due to memory issues when attempting to build the model in R with a larger training dataset size. The features used for training the model were:

- Total_population
- Median_rent
- Median_age
- Mean.TemperatureF
- Mean.Humidity
- PrecipitationIn

The model applied was a random forest regression with 200 component trees, sampling with replacement of data points, at least 5 data points per terminal node, and about d/3 features per tree (where d = total number of features for training set).

On the validation set, the model has a mean absolute error (MAE) of 36.65 hours and a median absolute error (median of the sums of the residuals) of 29 hours.

The baseline average error - the mean absolute deviation between the response times and the average response time for heat/hot water complaints - is about 39.79 hours. The baseline median absolute error (MAD) is 25.93 hours.

As shown below, the top two important features for this complaint type, as shown by the random forest model are humidity and temperature, as they have the highest increase in mean squared error (MSE) when replaced by random noise. Because these complaints have a very high priority during colder seasons, when temperature is at its lowest and humidity is very high, it makes sense that those features are significant for predicting when such cases would be resolved.

```
> round(importance(rf), 2)
                %IncMSE IncNodePurity
total_population    30.67       8970040
median_rent         33.78       8549945
median_age          35.20       8238997
Mean.TemperatureF   53.17      13587575
Mean.Humidity       56.38      10265547
PrecipitationIn     34.29       4666266
```

However, because the model predictions are still off by about 1-1.5 days, it may be beneficial to include additional weather data fields for the model such as barometric pressure and the weather season. It may also be helpful to include other time variables such as the month, hour, day of the week, and time of day (AM/PM). In addition, expanding the weather

dataset to include all observations in a given day, as opposed to averages and totals (ex. Average daily temperature, total daily precipitation, etc.), may also provide a more realistic relationship of these weather variables to response time.

Conclusion

After exploratory analysis of several datasets, we determined that several of of our original features were notable contributors to changes in the response time. Humidity appeared to stand out as the largest contributor to the variance in response time across several complaint types, while the precipitation and income features also provided us with features to increase the accuracy of our predictions. However, while there was some contribution to predictive accuracy by these features, our end results were not as accurate as we had previously hoped.

This was not entirely unexpected, as during our initial analysis several of our predictive features had unsatisfactory correlations with our response variable. In future analyses we would like to incorporate further datasets that might contribute to more successful predictive models. Datasets might include staffing info for the HPD and NYPD agencies, as well as hourly schedules of employee hours. Further analysis that might contribute to more successful models might involve determining the causes and trends of some of the specific outliers in the data, or further parameter tuning for our random forest and linear models.






Peter Darche, Asad Hassan, Ian Johnson, Jeff Khasin, Rajesh Madala, and Chris Rusnak

Sources:

- https://nycopendata.socrata.com/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/xx67-kt59
- https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8
- http://www1.nyc.gov/site/planning/data-maps/nyc-population.page
- http://www.nyc.gov/html/nypd/html/home/precincts.shtml
- https://nycopendata.socrata.com/data
- http://www.zillow.com/research/data/
- https://www.wunderground.com/history/
- http://www.weather.gov/media/okx/Climate/CentralPark/monthlyannualprecip.pdf