Begum Babur and Kerem Tuncer

Wayne Lee

Applied Data Mining

02/15/2021

Project 1

The Relationship of 311 Complaints to the NYPD with Demographics and Poverty Status

Introduction

Providing connection with local authorities in times of need is an obligation of a local government that cares about its citizens' welfare. In New York City, NYC311 exists to undertake that duty. It is a widely-used platform that provides quick and easy access to all NYC government services and information. In 2018 alone, NYC311 had exactly 44,023,630 customer requests for services or information[1]. The interactions took place through various channels, such as phone calls, the mobile app, the website, and social media channels. It is also important to note that there has been a constant upsurge in the number of complaints filed throughout the years. Between 2017 and 2018, the use of the platform showed a 10% increase[2]. A similar increase was observed in 2020, mostly due to COVID-related requests.

Considering this rise in demand, we were curious to see whether factors such as gender, race, age, and poverty status could be used to predict the number of 311 complaints. Specifically, we were interested in requests that required immediate action and had a certain degree of importance. Consequently, we decided to concentrate specifically on 311 complaints to the NYPD, which primarily consisted of requests related to noise, illegal parking, blocked driveways, and derelict vehicles.

Data Collection

The 311 Service Requests dataset was obtained from the NYC OpenData platform. The dataset includes complaints from 2010 to the present. There are over 24 million observations in it. Hence, it was necessary to pull a subset given that downloading "all data" is, in general, a bad practice with large datasets. The first filtering condition was to include only complaints that the NYPD received. Then, we only acquired rows with a creation date between the beginning of 2015 and the end of 2019. This is because we used demographic and poverty indicators from the

[1] https://www1.nyc.gov/311/311-sets-new-record-in-2018.page

[2] ibid.

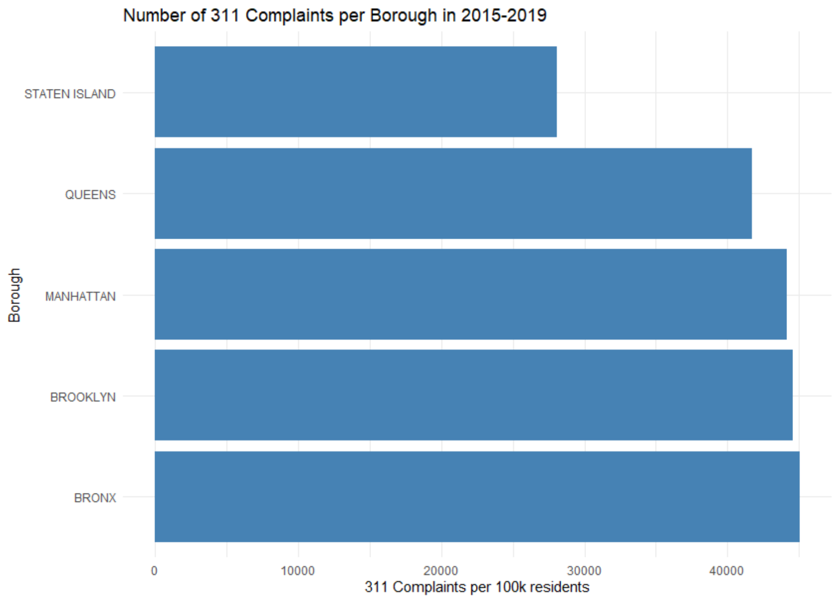
5-year estimates of the American Community Survey, which is best representative of 2015-2019. The filtering process – done through Socrata API – resulted in a total of 3,586,505 observations.

To investigate the demographics and the poverty status of New York City, we consulted the Census Bureau's website. We used data from the American Community Survey (ACS) 5-year Estimates. We downloaded a total of two CSV files: Demographic and Housing Estimates and Poverty Status, which was measured over the last 12 months but the data are shared to represent the 5 year estimates. They reported the race, gender, age, and poverty status breakdowns by NY State ZIP codes.

Data Pre-processing

Kerem was responsible for readying up the 311 complaints dataset. After looking at the columns' data types, he realized that R did not register the NA cells in the ZIP code column. Hence, he converted cells with whitespace ("") to NA. Then, he deleted all the rows where the ZIP code column was NA, as imputation would not be possible in this instance. This got the number of rows down to 3,573,033. He then counted the number of rows per unique ZIP code and converted this table into a new data frame with two columns: ZIP code and the number of complaints. Figure 1 shows the distribution of 311 complaints per borough in 2015-2019. The boroughs of Bronx and Brooklyn had the highest number of 311 complaints per 100k residents.

Figure 1



Begum was responsible for pre-processing the two ACS survey datasets. The dataset that included the demographics had over 300 columns that consisted of IDs, ZIP codes, estimates, the margin of error for the estimates, the percentages to the total population, and the margin of errors for the percentages. She began by deleting all the columns except the ZIP code and the ones that showed the percentages to the total population. From those, she manually selected columns that provided interesting demographic information. In the end, the variables of interest were the percentage of males and females, estimate of the sex ratio, percentages of the age breakdowns of the population (< 5, 5-9, 9-14, 15-19, 20-24, 25-34, 35-44, 45-54, 55-59, 60-64, 65-74, 75-84, 85 <), median age, % under age 18, % 16+, % 18+, % 21+ , % 62+, % 65+, % 18+ and male, % 18+ and female, % 65+ and male, and %65+ and female. For the poverty status dataset, Begum excluded all columns except the ones showing the ZIP code, the total population for which poverty status was determined, and the population living in poverty. Finally, she created a new column that was the result of dividing the total population for which poverty status was determined by the population living in poverty. This column made it possible to compare areas of different populations. Also, given that the ZIP codes were for all of New York State, she downloaded a list of all NYC ZIP codes and excluded all rows with a ZIP code indicating a non-NYC location.

Afterward, we merged the pre-processed 311 and ACS datasets by ZIP code. However, there was a small caveat as the ZIP codes in the ACS had whitespace. After trimming the whitespace, the merging was successful. We ended up with 181 observations and 37 columns without any NAs.

Lastly, we renamed the columns from the ACS dataset because they were named as specific technical labels. The new names of the columns were indicative of what the column was describing. For example, DP05_0002PE became "TotalPercentMale." Several numeric columns were registered as factors in R. As a result, we converted them to numeric data types.

Feature Engineering

As discussed in the Data Pre-Processing section, the primary feature we engineered was the "Complaints" feature, which showed the number of total complaints to the NYPD between 2015-2019 per NYC ZIP code. This feature's calculation was done by summarizing the total number of complaints per NYC ZIP codes in the 311 dataset. Engineering this feature allowed us to have a dependent variable, with which we were able to look for factors that could have affected the number of complaints made. Furthermore, summarizing by the count of complaints allowed us to keep the ZIP code attribute, which was paramount for merging the datasets. Finally, this feature was more than useful for the project as it served as the response variable.

Algorithm

Given that our dependent variable (number of complaints) was a numeric value, we were interested in using a regression algorithm. For the purposes of our project, we decided to use a more interpretable yet less flexible algorithm. Hence, we chose a multiple linear regression model, for which we could examine the RMSE, the R-squared, and the p-values of the coefficients.

We had a relatively long process of feature selection, where we combined trial-and-error, the variable importance function by caret (which investigates the relationship between each predictor and the outcome), and the Boruta library (which provides an automated feature selection method).

In deciding our optimal model, we tried to balance these three factors: (1) a high R-squared, (2) a low RMSE, and (3) statistically significant coefficient estimates. Before starting the trial-and-error procedure, we took a look at the importance rankings that both Boruta and the variable importance functions produced. Below are the visualizations of the importance rankings:

Figure 2

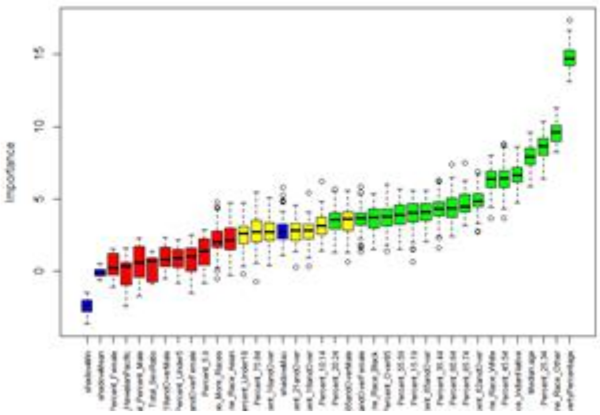
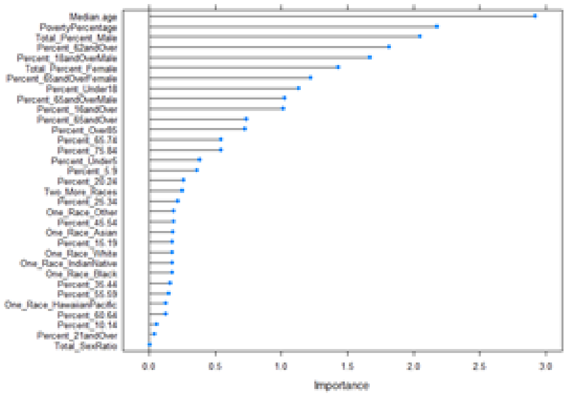


Figure 3

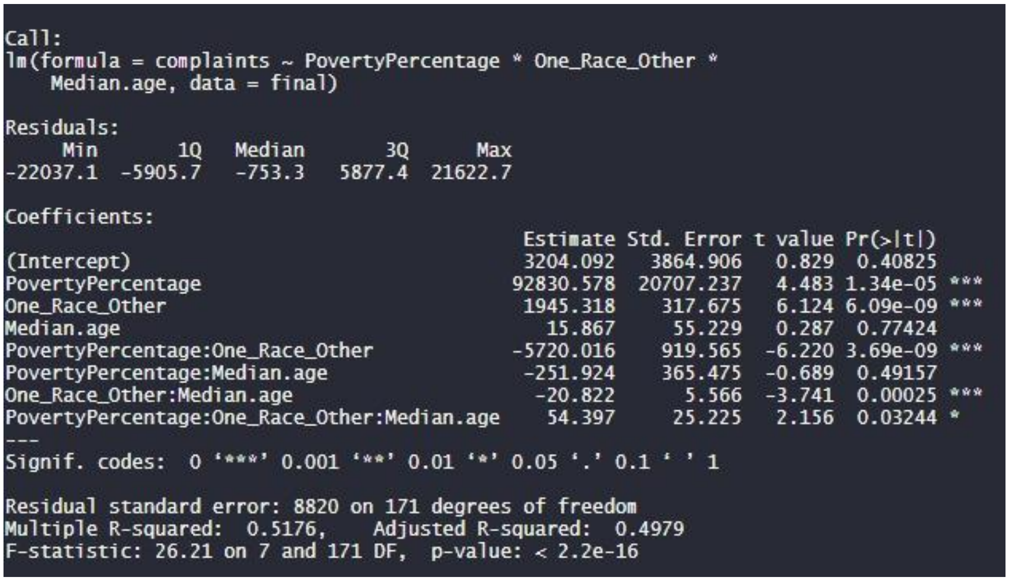


These two functions allowed us to understand which variables were more likely to be essential for the model. Subsequently, we tried a lot of different combinations of features and examined their output. In the process, we also made use of interaction terms and polynomial terms. Throughout our experimentation, we learned that we could get a really high R-squared and a low RMSE value by including crazy polynomial terms and increasing feature amounts. However, we were aware that this phenomenon was due to overfitting and required the employment of a validation method. We decided on repeating K-fold cross-validation with k = 10. Suddenly, a complex model with 0.70+ R-squared turned into one with less than 0.10 R-squared.

As a result, we examined each model's performance both on the total dataset and under 10-fold cross-validation. During the process, we discovered that there were four outliers that had a significant negative impact on our model's performance. Our discovery took place when we were taking a look at the residuals vs. leverage and the Cook's distance plots. For exploration purposes, we decided to remove these four points that either had an extremely large Cook's distance or exceed three standard deviations in the residuals vs. leverage plot. There was a significant increase in our R-squared and a decrease in our RMSE for all models tried.

After hours of trial-and-error, we settled on a model using the variables median age, the proportion of people living in poverty, the proportion of people from other races in the population (someone who is not Asian, Black, White, Native American, Native Hawaiian, or Pacific Islander), the interaction term between each variable, and the interaction term of all three variables. Figure 4 shows the model formula and the output:

Figure 4



```
Call:
lm(formula = complaints ~ PovertyPercentage * One_Race_Other *
    Median.age, data = final)

Residuals:
     Min       1Q   Median       3Q      Max
-22037.1  -5905.7   -753.3   5877.4  21622.7

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 3204.092   3864.906   0.829  0.40825
PovertyPercentage                          92830.578  20707.237   4.483 1.34e-05 ***
One_Race_Other                              1945.318    317.675   6.124 6.09e-09 ***
Median.age                                    15.867     55.229   0.287  0.77424
PovertyPercentage:One_Race_Other           -5720.016    919.565  -6.220 3.69e-09 ***
PovertyPercentage:Median.age                -251.924    365.475  -0.689  0.49157
One_Race_Other:Median.age                    -20.822      5.566  -3.741  0.00025 ***
PovertyPercentage:One_Race_Other:Median.age   54.397     25.225   2.156  0.03244 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8820 on 171 degrees of freedom
Multiple R-squared:  0.5176,    Adjusted R-squared:  0.4979
F-statistic: 26.21 on 7 and 171 DF,  p-value: < 2.2e-16
```

The model’s output in the 10-fold cross-validation produced an R-squared that went up to 0.53 and an RMSE value that went as low as 8715 (number of complaints ranged between 13,000 and 49,000). Additionally, the following coefficients had statistically significant p-value ($\alpha < 0.05$): poverty status, the proportion of people from other races in the population, the interaction term between poverty status and being from the “other” race, the interaction term between median age and being from “other” race, and the interaction term between all three variables.

Interpretation of Results

Next, we investigate the relationship between these variables and the dependent variable. The predicted number of complaints, when the rest of the variables are 0, is 3204. This indicates the average number of complaints when the percentage of poor population and people who belong to the race category “Other” is 0, as well as when the median age is at 0 in the ZIP code area, the number of predicted complaints in the area is 3204. Considering that these conditions are fairly unlikely, it is essential to examine the relationship between our strongest suggested independent variables on the number of complaints.

The poverty percentage is found to be highly significant in predicting the number of complaints. When the poverty percentage in a given area increases by 0.01, the number of complaints is also predicted to increase by 928.30, on the condition that the percentage of “Other” race and median age stays constant. Although we want to err on the side of not making causal remarks, this points out that the higher the poverty percentage in a given ZIP code area, the more likely people are filing complaints to 311.

When the percentage of people who belong to the race “Other” increases by 0.01, the number of complaints is predicted to increase by 19.45, when the poverty percentage and median age stays constant. The percentage of “Other” races in a given ZIP code area was significant in predicting the increase in the number of complaints made to 311.

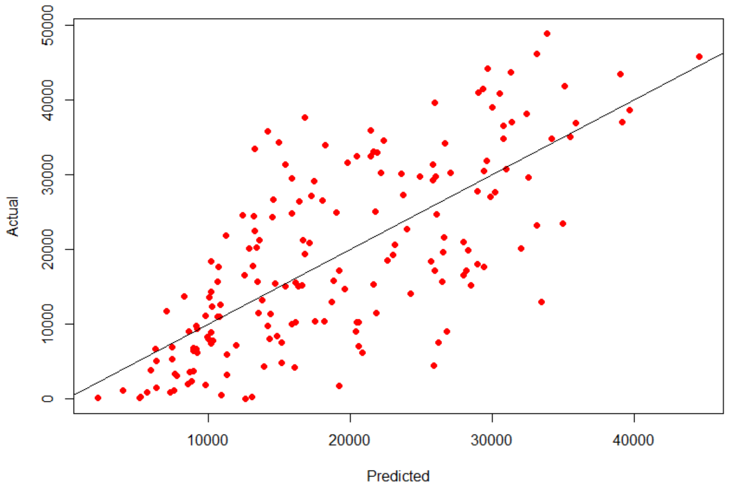
The significant interaction term between poverty percentage and the “Other” race percentage indicates that the effect of poverty percentage on the number of complaints is different depending on the percentage of people who belong to the race category “Other.” More precisely, poverty percentage points are predicted to reduce the number of complaints by 57.20 for every 1% increase in the number of people who belong to the race category “Other” within the ZIP code areas.

Similarly, the significant interaction term between the “Other” race and the median age indicates that the percentage of “Other” races in an area is predicted to decrease the number of complaints by 20, for every one year increase in the median age of the population within the ZIP code areas.

Our three-way interaction term indicates that the interaction between the poverty percentage and the other race percentage is different across different median ages within the ZIP code areas. We observe a positive relationship between these variables. This shows that the effect of interaction between poverty percentage and the percentage of people from the “Other” races increases the number of complaints by 54, when the median age of the population within the ZIP code area increases by 1.

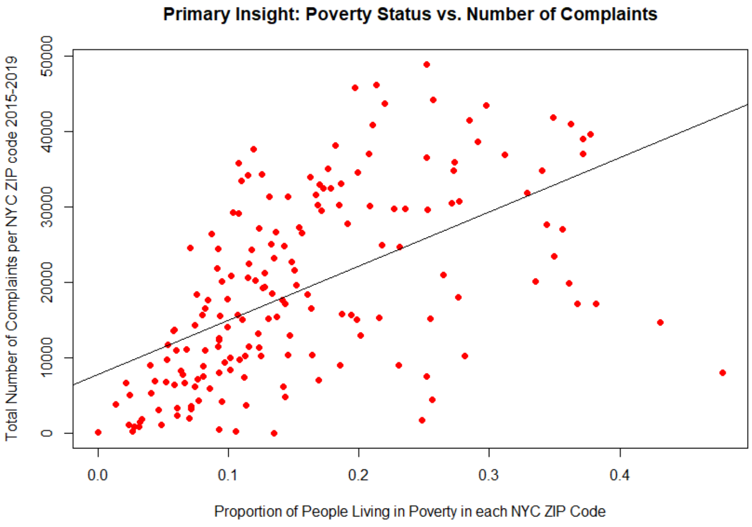
Although we are not really looking for causal effects, we also investigated the residual plot and tested some linear regression assumptions. As expected, the residuals vs. fitted plot looked more or less random, and there were no patterns. Additionally, we used the gvlma package - which tests several criteria of linear model assumptions by inspecting skewness, kurtosis, heteroscedasticity, and the link function. According to gvlma, the assumptions were acceptable and not violated in the case of our multiple regression model. Figure 5 represents a plot of our predicted values and the actual values.

Figure 5



Given our coefficients and their statistical significance, the primary insight we acquired from our model was that the proportion of people living in poverty had a positive correlation with the number of 311 complaints to the NYPD in NYC, even when accounting for the median age and the "other race" variable. It is also important to note that the 311 complaints to the NYPD do not include anything that we would consider a serious crime; therefore, this is not an insight related to the relationship between crime and poverty. Figure 6 visually depicts the insight, where we show the univariate association between complaints and poverty status.

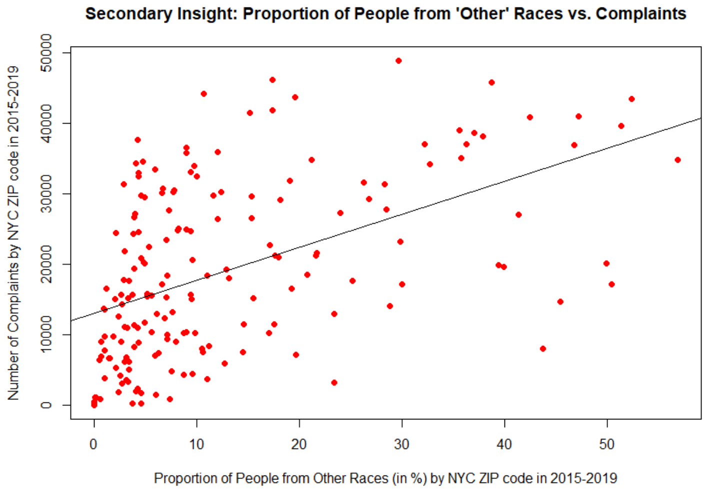
Figure 6



To see whether this pattern was "by chance," we conducted a literature review on the topic. According to Professor Robert Sampson of Harvard University, there is evidence to suggest that 311 calls are increasing in NYC areas that are going through gentrification (Fayyad, 2017). Gentrification usually occurs in more impoverished neighborhoods to change the area's character by improving housing, attracting new business, and making wealthier people move in. Hence, it makes sense that 311 complaints increase as the poverty rate of a neighborhood increases. Likewise, a 2017 study published in the Environmental Health Perspectives (EHP) Journal argues that poorer neighborhoods experience more noise pollution. Given that a majority of the NYPD 311 complaints are related to noise levels, our insight does not seem to be by chance (Weaver, 2020). The insight we found regarding poverty level and complaint amount is a counterintuitive one. We initially thought that a negative relationship would exist as filing complaints require platforms like phones, smartphones, laptops, and internet access, which would probably be less accessible in poorer neighborhoods.

A secondary yet significant insight is the positive correlation between the proportion of "other" race people in an NYC neighborhood and the number of 311 complaints, even when accounting for median age and poverty status. Although the proportion of "other" race people is not a great indicator, we still believe that it reveals enough about an NYC ZIP code's diversity. The same EHP paper we discussed above also holds that there is more noise pollution in more diverse neighborhoods (Weaver, 2020). Consequently, this insight also seems to show a real pattern.

Figure 7



Conclusion

In this project, we looked into the relationship between demographic factors and the neighborhoods' poverty status and its residents' likelihood of filing complaints to NYC311. We specifically focused on non-violent complaints made to the NYPD throughout 2015 and 2019. Our analysis revealed that certain demographic and socioeconomic factors are, in fact, influential in making 311 requests. We found out that more complaints were filed in the neighborhoods where there was a higher percentage of people living in poverty and a higher percentage of people who belonged to races other than White, Black, Asian, Native American, Hawaiian, or Pacific Islander.

Interestingly, the increase in complaint numbers was only predicted when these two demographic factors occurred on their own. The co-occurrence of these factors resulted in a predicted decrease in complaint numbers. When the poverty percentage was increasing, the percent increase in the number of people who belonged to the "Other" race, resulted in a decrease in the predicted number of filed complaints.

Overall, this project provides insights into the widely used NYC311 service by analyzing its use demographically. The insights can help NYPD, or broadly the NYC government, to allocate their resources efficiently to the neighborhoods where the conditions to predict high 311 complaint numbers are met. This would allow NYC to provide quick and easy solutions to their citizens' problems and efficient access to local services for increased welfare across neighborhoods.

Works Cited

311 Sets New Record with 44 Million Customer Interactions in 2018. (2019, February 19). Retrieved February 15, 2021, from https://www1.nyc.gov/311/311-sets-new-record-in-2018.page

Fayad, Abdallah. "The Criminalization of Gentrifying Neighborhoods." *The Atlantic*, Emerson Collective, 20 Dec. 2017, www.theatlantic.com/politics/archive/2017/12/the-criminalization-of-gentrifying-neighborhoods/548837/.

Weaver, Shaye. "Noise Complaints in NYC Have Increased Almost 300 Percent since February." *Timeout*, Time Out Group, 27 July 2020, www.timeout.com/newyork/news/noise-complaints-in-nyc-have-increased-almost-300-percent-since-february-072720.