New York City 311 Call Data

Habib Alvi | Chris Lindeman | Vikram Makker | Saif Qureshi

Motivation/Introduction

Government spending is always considered an important area of focus regardless of the political party in power. It is a vital performance indicator that can be used to hold the government officials accountable during, and outside of an election cycle. As the world wrestles to return to normalcy in the post pandemic world, the responsibilities of the government and how its resources are allocated are more important than ever. The team proposes to use 311 Calls data made by the residents of New York City so that the city government may have additional tools to make strategic tools pertaining to agency budgets at its disposal. The motivation of this project to create a user friendly, and a data driven framework, that will not only provide a visual representation of the city's most widespread concerns to the city officials, but also allow its citizens to be more engaged and informed.

Data

The data came from multiple sources, and was dealt with accordingly:

NYC 311 Data: 41 fields, 27.2M records, 15.4 GB

The data was obtained from NYC Open Data. The city has maintained the data since 2011. The team stored the clean data in a .db file, which was queried in the analysis via SQLite.

Agency Budgets

Part of the data was obtained from NYC Open Data, however that dataset was created in 2016. The websites of agencies, and other repositories were found for budget information pre-2016. The data was saved in a CSV, and then accessed via Pandas in the analysis.

Los Angeles Benchmarking Data

The data was obtained from data.lacity.org, and queried via Sodapy.

To ensure all members were following a uniform approach, the team used free credits provided by Google Cloud Platform during initial stages of data gathering, cleaning, and analysis.

Experiments/Results

Complaint analysis by volume

The team created visualization charts on all complaints data to understand following aspects:

- Borough with highest number of complaints
- Government agency tackling most number of the complaints
- The type of complaint that is most reported
- The areas where the complaints originated

This analysis led to the finding that noise complaints, specifically residential noise complaints are most prevalent amongst the residents. Brooklyn has the most number of complaints, which is not at all surprising considering it is the largest borough with the highest number of residents. Furthermore, the affluent neighborhoods have the lowest noise complaints but less affluent neighborhoods across all boroughs have a consistent numbers of noise complaints.

Complaint analysis by response time

Response time is the best indicator of the performance of the city departments and following aspects were analyzed:

- Boroughs with lowest and highest response times
- Response time by agency
- Changes in response time over time
- Neighbourhood response time heatmaps by every agency

Brooklyn has the highest complaints but the response time for Queens is longest. Interestingly, Staten Island has the least number of complaints, but it has the second highest response time.

Benchmarking

Los Angeles' performance was used to benchmark against New York City, however, given storage constraints, only 1 million records were used for analysis. Even though Los Angeles is quicker to respond to complaints, the city does have half the population of New York City, and is not as densely populated. To truly establish effective benchmarks, a more thorough analysis of Los Angeles 311 data will be required.

Regression (budget and performance)

The team established a relationship between budget allocation as a response to predictors: complaints logged, complaints closed, and the ratio of complaints closed to complaints logged. Multiple models were created with a combination of said predictors, and in the end, the model with all three predictors had the best R^2 value.

Time series (complaint forecast)

Application of seasonal ARIMA model to time series data of noise complaints, was performed twice:

I. On data that included complaints logged during the pandemic.

II. On data that excluded complaints logged during the pandemic.

Comparing the forecast charts, one can see that the confidence interval for the years 2022-2023 is significantly wider than that provided by fitting the model to the data ending with March 2020. However, the model without 2020-2021 data will fail to take into account any fundamental changes in society as a result of the pandemic.

Overall, the team's approach is better than current literature since it provides a way of prioritizing the most prevalent issues, establishing relationships between financial obligations, and predicting the concerns of the residents that will lead to actionable insights. Furthermore, the combination of the analysis with interactive visualizations would lead to many more discoveries that the data has to offer.

Approach

Summary

Data cleaning

Prior to any analysis, it was necessary to clean the data intricately. As a result, the team reduced the size of the data by converting the type of fields (datetime to integer, creating dictionaries to map complaint and agency data to integer values) to more memory efficient formats, improved upon categorization, and rid of all columns that were redundant to the project.

Static visual analysis

A visual analysis of the raw data was done to narrow the scope of the problem. Since 311 calls document complaints of all of New York's residents, and intent of the project is to provide guidance to the municipality to allocate funding more effectively, it was prudent to establish most important concerns voiced by the citizens and the neighbourhoods they lived in.

Budgetary regression analysis

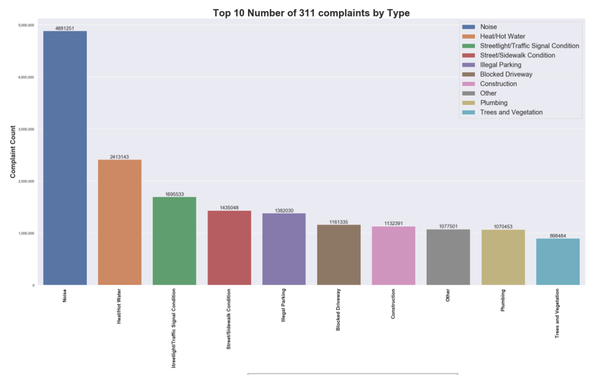
In the interest of effective budget allocation, it was necessary to establish relationships between complaints and budget allocated. The team hypothesized that for any agency, its performance level (defined as "closed complaints/total complaints" for a particular year) must be correlated with its operating budget. Given its simplicity and ease of explanation, linear regression was used, where predictors were open complaints, closed complaints, and performance level; the predictor was budget allocated. It was assumed that fraction of the budget spent on closing certain categories of complaints was equivalent to fraction of the complaints in that category over total complaints for the year.

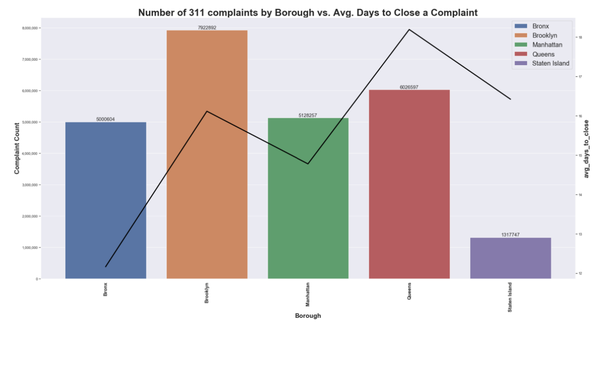
Complaint forecast

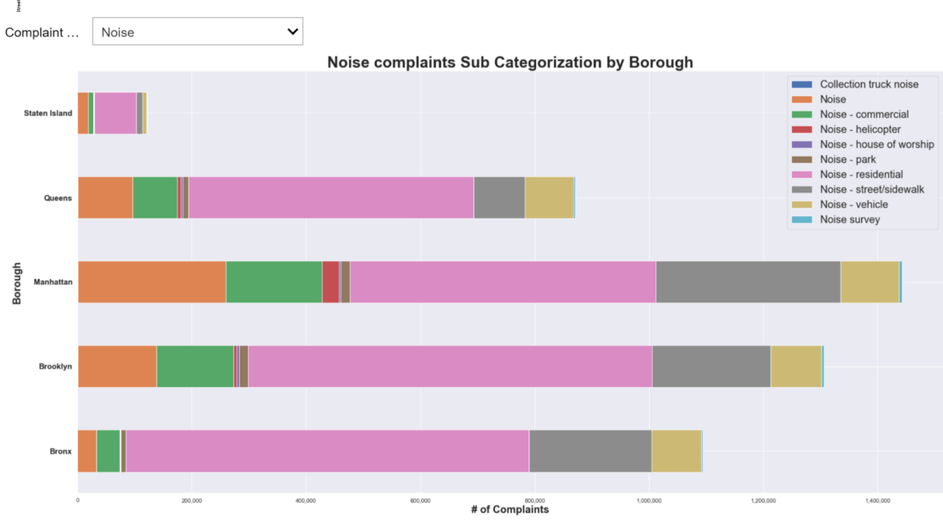
After establishing a relationship between complaints and budget, another piece of the puzzle was giving agencies the ability to predict what the next years may look like, providing them with a strong justification as they ask for funding from the government. As per the static visual analysis, New York City's residents are most bothered by noise. As a result, the team chose to implement a seasonal ARIMA model to forecast the amount of noise complaints agencies may receive in the years that follow. Two forecasts were made, one where the entire dataset was used, and another where data after March 2020 was ignored; the team wanted to differentiate between the behaviour of New Yorkers when the pandemic began.

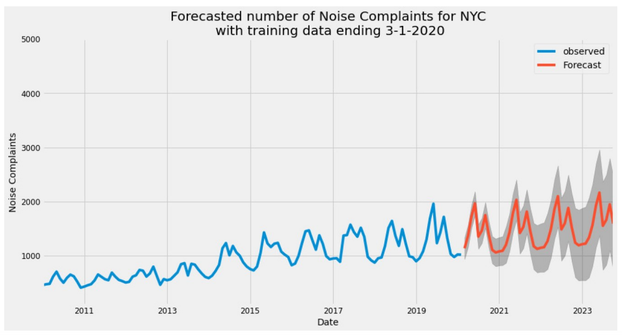
Benchmarking

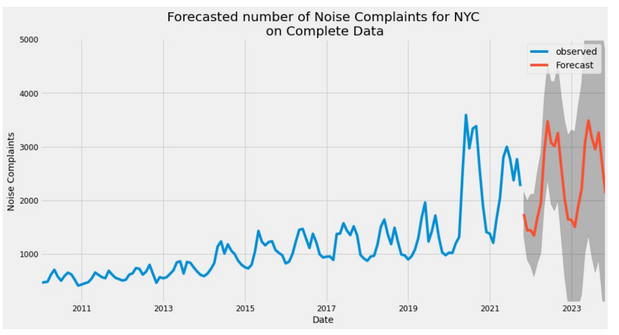
Finally, to prove that the City of New York is, in fact, a great place to live, one would need to measure its performance against another great American city. The City of Los Angeles was used as benchmark and the team compared how quick were the agencies when it came to closing open complaints

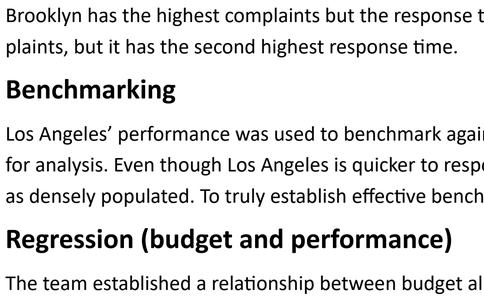


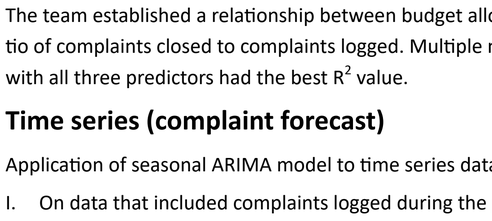












Created	Closed	Performance	R^2
X			0.57
	X		0.57
		X	0.21
X	X		0.62
X		X	0.59
	X	X	0.59
X	X	X	0.72