Data Science for Better 311 Service

Analysis of Patterns in 311 Call Volumes of Major Metropolis

1. Executive Summary

1.1. Background

In recent years, as more government leaders realize the importance of citizens' access to government activities, many cities across the United States began the undertaking of making public data more accessible to their citizens. The datasets published cover a wide range of topics, including but not limited to education, business, transportation, and public health. This undoubtedly grants everyone the freedom to explore the datasets and motivates data analysis on these data.

This week, our team explored data of 311 requests in New York City and Washington D.C. With the help of other auxiliary datasets from OpenData and technical tools, we aimed to answer the following questions:

a) Are there any patterns, spikes, or temporal distribution shifts in the volume of 311 requests?
b) What are the factors that influence 311 request distributions? Specifically, how do COVID-19 and crime incidents influence 311 requests?
c) How can we predict the volume of 311 requests in order to help the governments optimize their resource allocations?

Our analysis mainly focused on two indicators from the 311 request dataset. The first indicator is the daily amount of 311 requests, including the total number of requests in a city, and the amount of different request types in different city sub-areas. This indicator functions as a profile of the citizens' major complaints as well as an indirect indication of their living quality. The second indicator is the average time needed for the cities to resolve 311 requests. This indicator is a measurement of the working efficiency of government agencies and can give us an idea of whether they lack necessary resources.

Based on the two indicators, we conducted data analysis, time series prediction, and linear regression to achieve our aim. By the end of this report, we would like to give several suggestions to the city governments, in hope of enabling them to better serve the citizens in the future.

1.2. Summary of Findings

From our initial EDA on a daily volume of 311 calls, we found that seasonality had a significant trend effect. This seasonality covariate indicated typically more calls on Sundays and the fewest on Thursdays. By assuming an additive decomposition of the time series, we also found that there was a slightly positive trend with potential peaks that jumped up for a period of time. Given what happened in both NYC and DC during those jumps we believe these were due to a new wave of Covid-19 and presidential election events for NYC and Omicron hitting DC.

From our analysis of crime, we proceed similarly and decompose the daily crime data. We see some potential seasonality trends not being fully accounted for within the data as we see that there are cosine-like fluctuations within the trend post decomposition. For DC when comparing the call data and the crime data through determining the correlation between the two series with TLCC, we found that there is a 21-day lag difference between the two. And between

resolution time vs crime, we found that there is a 6-day lag difference between the two. For NYC, we ran the same correlation tests and found that there was nearly no significant effect of any time lag.

2. Technical Exposition

2.1. Data Preprocessing

We delved into the 311 service request datasets for New York City and Washington D.C. To further empower analysis, we also made use of datasets on COVID-19 and crime incidents for both cities. All of the datasets are downloaded from the OpenData websites. Detailed citations can be found at the end of this report.

To prepare the data for analysis and modeling we completed the following tasks on the 311 request datasets:

a) data cleaning, which includes filling in or removing null values, filtering out entries in the dataset whose location suggests that they happened outside of the city, etc.
b) conflated similar request types. For instance, we categorized “noise-street” and “noise-road” into the same category “noise-traffic”
c) filtered out the less frequent 311 request types. We defined “less frequent” to be the request types that take up less than 1% of the total number of requests. Eventually, we were left with the top 23 to 24 request types for each dataset

While we were preparing the datasets, we noticed that there are actually request types directly related to covid, such as “face covering” and “covid19 testing”. We conflated them to the category “covid” and included them in the cleaned set. However, there are very few entries of these request types.

2.2. Exploratory Data Analysis

2.2.1. 311 Requests - Request Types in New York City

To begin with, we visualized the distribution of 311 requests across different request types and government agencies.

For 311 requests in New York City, we can see from the plot that the issues related to noise stand out as the clear majority category. In addition, the agency column is similarly imbalanced. Since NYPD handles the noise complaints, it responds to the vast majority of requests from 311. The Department of Housing Preservation and Development (HPD) and the Department of Parks and Recreation (DPR) also respond to a significant proportion of calls.

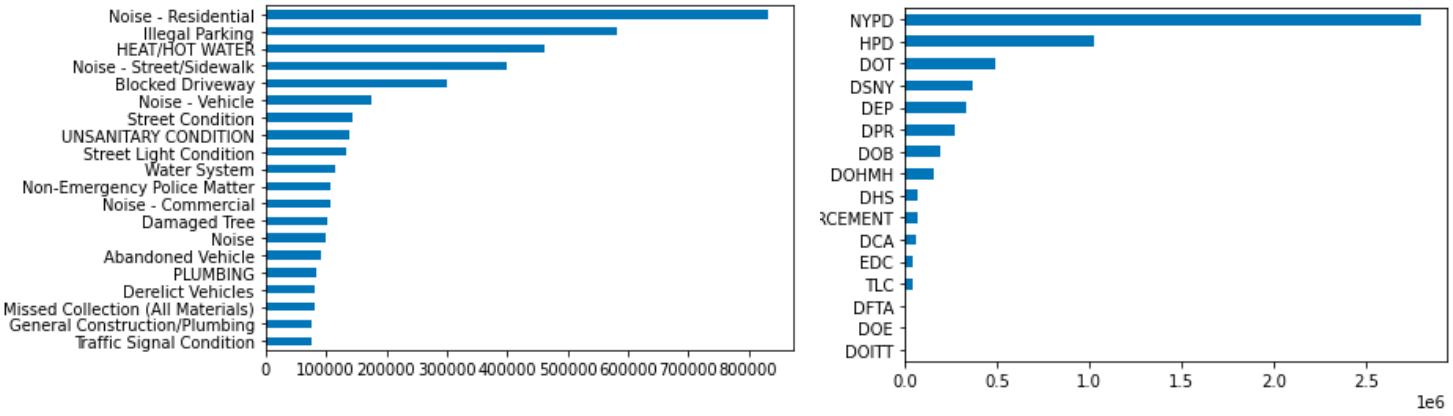

Fig 1. Distribution of 311 Requests by Request Type and Agency, New York City

After ranking each dataset by 311 call volume, we noticed that some complaint types are quite similar and could be classified as a broader category. For instance, there are 5 subcategories of noise appearing in the top 20 incident types for New York City. A plot of the monthly cumulative volume of requests regarding noise is shown below. Although the absolute frequencies differ from specific noise complaint types, we can see a prominent pattern across all subtypes – the call volumes peak for the first time from May 2020 to September 2020, and the second wave starts from May 2021 till September 2021.

In order to visualize the overall trend of call volumes in a broader view, we decided to conflate similar incident types. Details about data wrangling and preprocessing are discussed above in section 2.1.

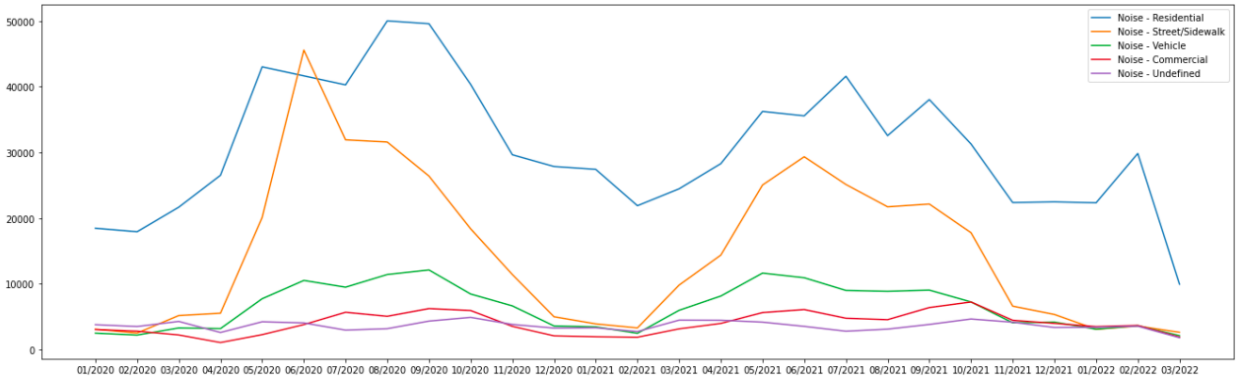

Fig 2. Trend of 311 Request Volume, New York City

In addition to the volume of calls over time, we also visualized the distribution geographically. Taking NYC as an example, we plotted the location data of each request, seeing how volume varied across the 5 boroughs. The yellow areas below indicate high call volume and appear predominantly in the Bronx, Manhattan, and part of Brooklyn. We can see that the call volume is higher in June around the Brox area compared with August.

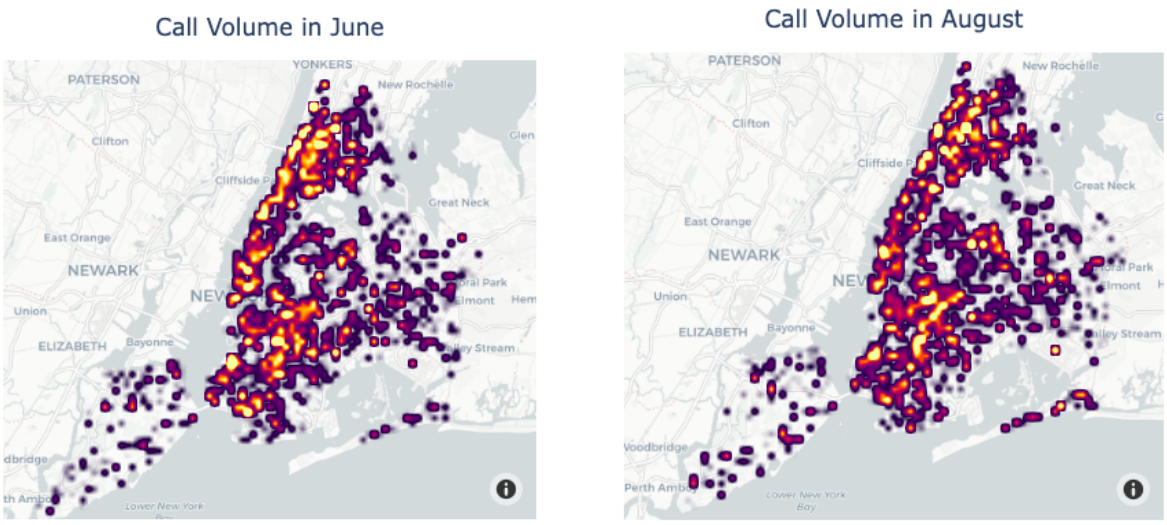

Fig 3. Request Volume on Map, New York City

2.2.1. 311 Requests - Request Volume Time Series Decomposition

Just by looking at the daily volume of 311 calls, we can see a clear pattern, with trends and repetitive seasonality. Hence, to grasp the patterns, we applied additive time series decomposition to extract seasonal, trend, and residual (noise) data from each volume series.

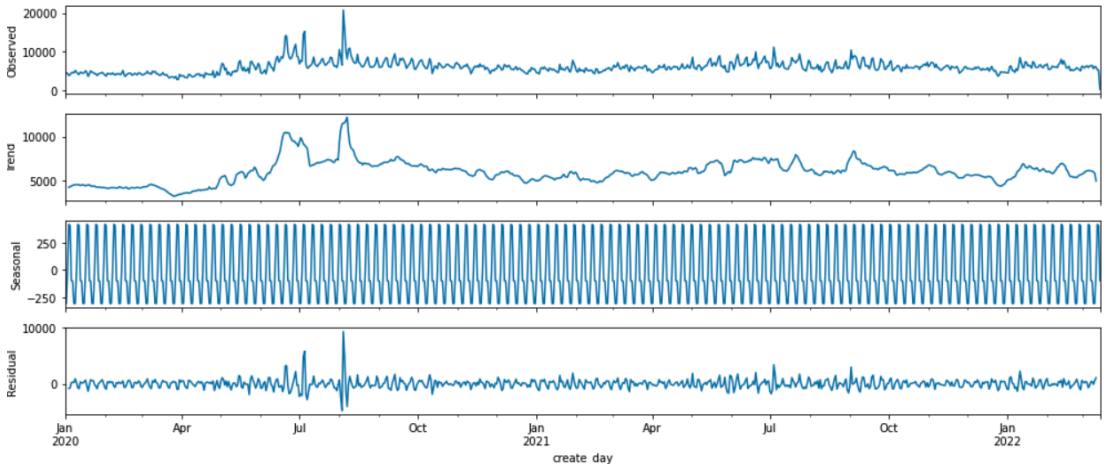

Fig 4. Time series Decomposition of 311 Calls, New York City

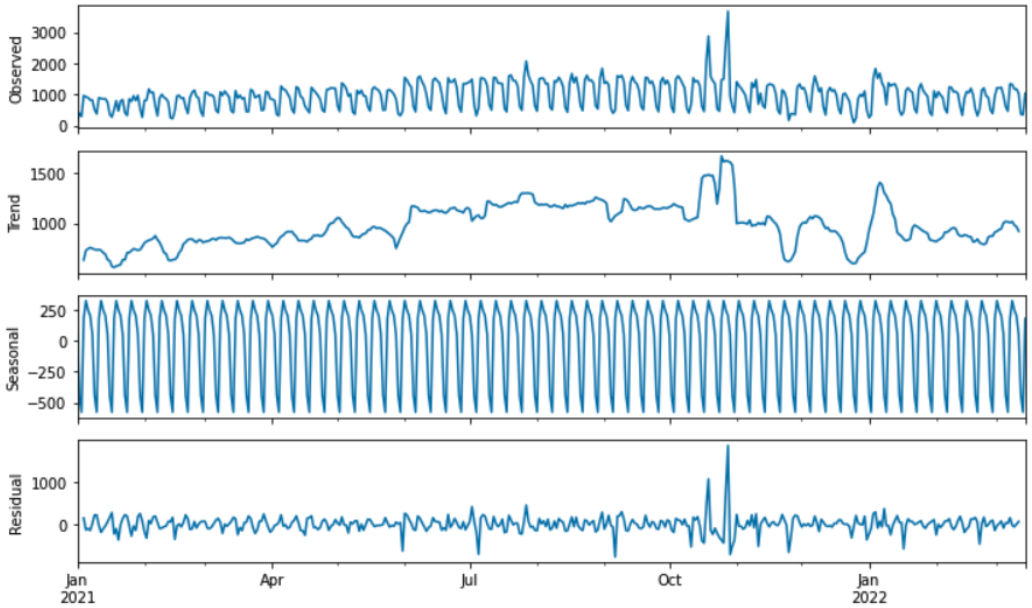

Fig 5. Time series Decomposition of 311 Calls, Washington D.C.

The additive model that we used for time series decomposition uses the following formula:

$$y_t = T_t + S_t + I_t$$

In the formula, T_t represents the trend component at time t, which reflects the long-term progression of the time series data. S_t represents the seasonal component at time t, which reflects seasonal variation. I_t represents the irregular component (or "noise") at time t, which describes random, irregular influences.

The decomposition gives us the following insights:

a) 311 request follows a weekly seasonality. Sunday usually has the most requests, and Thursday the least.
b) Most of the time, the range and variance of the noise are trivial compared to the total volume. Therefore, the amount of 311 requests is stably following the trend and seasonality pattern and is potentially predictable. To be noted, the seasonality of the 311 volume in Washington DC is significantly stronger than that of NYC.
c) There are spikes. For NYC, the spike happened around August 2020. Possible reasons are the new wave of Covid19 and presidential election events. For DC, the spike happened around the end of October 2021. A potential reason is the spread of the Covid19-Omicron. We further explored the correlation between Covid and 311 Volumes in the following sections.
d) There are distribution shifts. After May 2020, volume in NYC is in general higher. After May 2021, the volume in DC is higher as well.

2.2.2. Crime Incidents

We hypothesized that the number of crimes in a city is another factor that potentially influences 311 requests. Thus, to see the correlation between the two, for all three cities, we

looked at numbers of crime incidents per day. Data came from OpenData websites of the cities. For Washington D.C., we used the yearly Crime Incidents datasets. For New York City, we looked at NYPD Calls for Service datasets and filtered out crime-related entries using the "CIP_JOBS" (flag indicating if the call relates to a Crime In Progress) column.

2.2.2.1 Washington D.C.

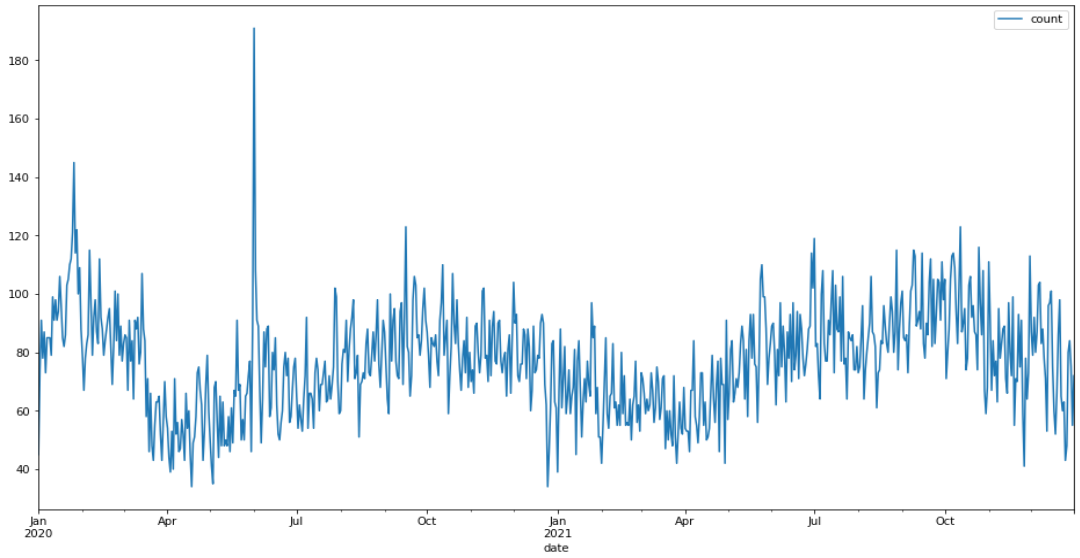

Fig 6. Number of Crime Incidents Per Day, Washington D.C.

The plot above demonstrates how the number of crime incidents fluctuates over time in Washington D.C. To eliminate the obvious seasonality shown in the plot, similar as what we did for 311 request datasets, using time series decomposition, we are able to isolate the trend of this data after removing seasonality and noise.

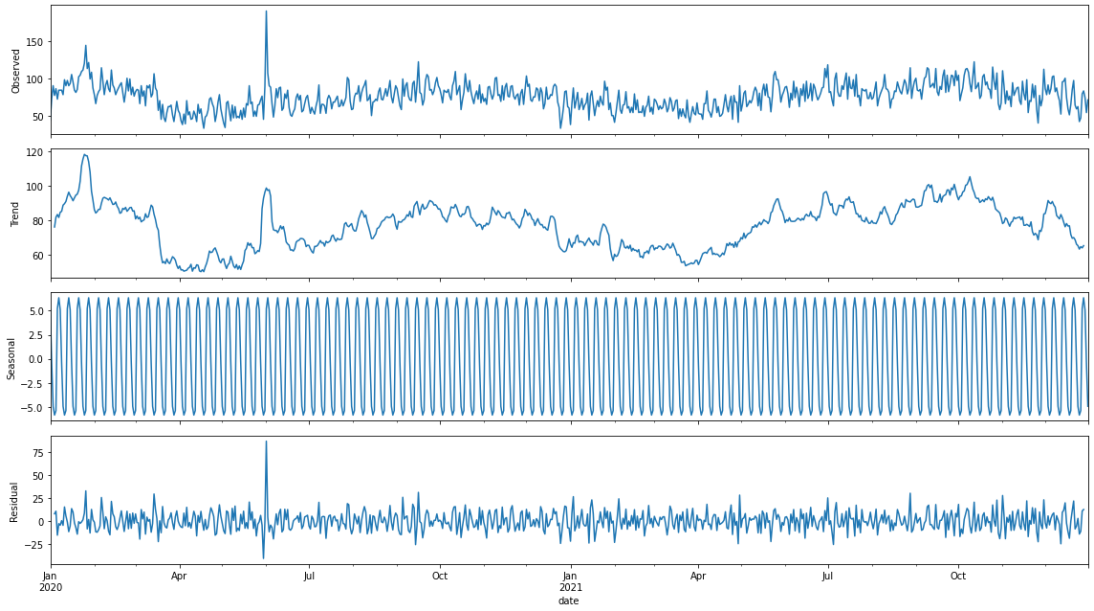

Fig .7 Time Series Decomposition of Crime Incidents Data, Washington D.C.

From the trend subplot above we can see that there were drops in the number of crime incidents from the start of the year to April, after which the number increases until it drops again in October. This pattern is consistent across 2020 and 2021.

Next, we calculated the correlation between the trend of crime incidents and the volume of 311 requests. Given the nature of the datasets, we decided to calculate the Time Lagged Cross-Correlation (TLCC) in order to make the best use of their time-series feature. TLCC examines the existence of a “leader-follower” relationship between two-time series data. In other words, it demonstrates if one of the data initiates a response in the other. In real-life contexts, especially when looking at the city as a whole, the impact of one factor on another is unlikely to take place right away. TLCC captures and quantifies this time lag.

TLCC is measured by shifting one of the time series data and repeatedly calculating correlations with another time series data. Doing this on the number of crime incidents and trend of 311 requests volume after decomposition gives us the following result:

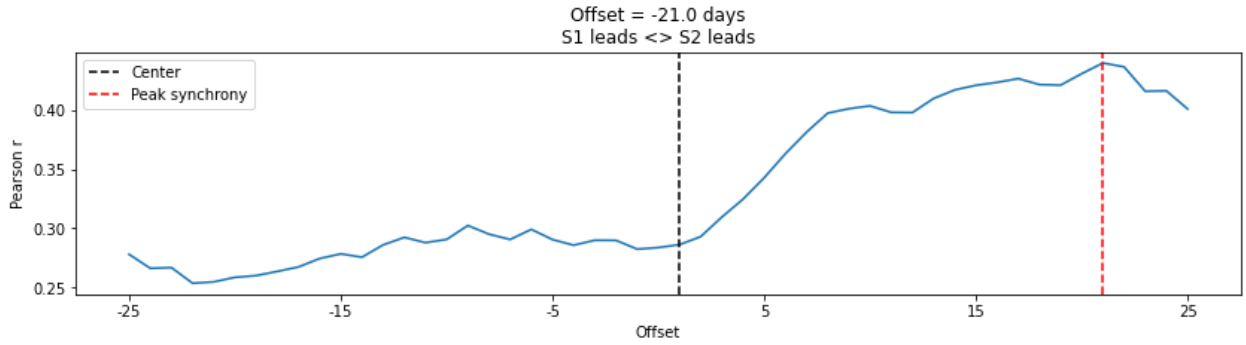

Fig 8. Time Lagged Cross-Correlation Between Number of 311 Requests (S1) and Crime Incidents (S2), Washington D.C.

The plot above visualizes correlations with different time lags. The black dashed line shows the center and the red dashed line shows the peak synchrony. The offset (position of the peak synchrony) tells us that in Washington D.C. the number of crime incidents initiates responses in the other and that there is a 21-day difference (the correlation between S1 and S2 is maximized when S1 is pulled forward 21 days). If we see the number of crime incidents as an indicator of residents’ safety (more crime means less safety), we can conclude from here that safety is positively correlated with the number of 311 requests. Stated differently, the less safe residents are, the more likely they are to make 311 requests. From the graph, we can also see that the correlation increases from 0-day lag to 21-day lag. Especially within the first week, the rate of increase is the highest. This implies that the impact of crime, whether a sense of safety or a sense of panic, lasts for a long time in Washington D.C. and is cumulative over time before it eventually cools down.

In the same manner, we looked at the Time Lagged Cross-Correlation between crime incidents and average resolution times of 311 requests. Since there is no seasonality detected in the average resolution time data, we used it without decomposition.

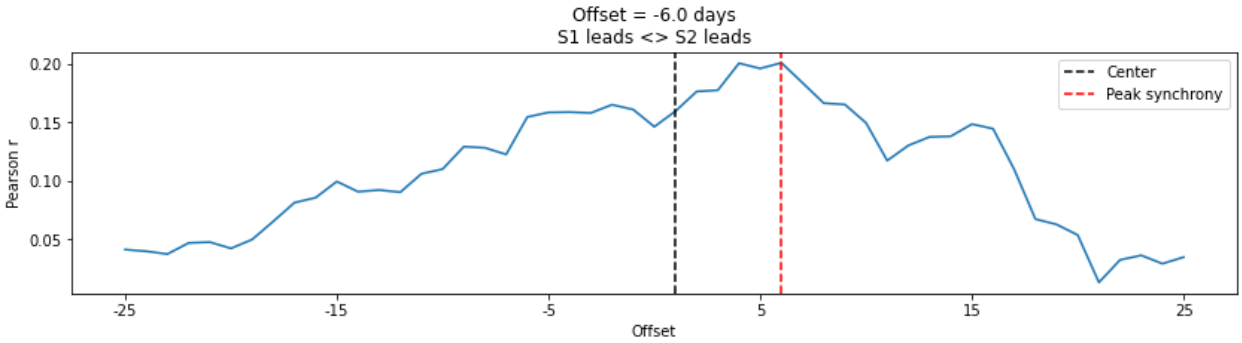

Fig 9. Time Lagged Cross-Correlation Between Average 311 Requests Resolution Time (S1) and Crime Incidents (S2), Washington D.C.

There are two conclusions that we can obtain from the graph above. Firstly, the number of crime incidents leads to the responses of an average of 311 requests resolution time. The positive correlation suggests that more crime incidents are associated with a longer average resolution time. Secondly, the time lag is 6 days. Compared to the volume of 311 requests, it takes a much shorter time for the influence of crime incidents to take place on the resolution time of 311 requests. One possible way of interpreting this difference is that fluctuations in the number of crime incidents directly affect the number of resources that the city government needs to spend on investigating the crimes. This then immediately affects resources available for resolving 311 requests, given that the amount of total resources available to the government is very unlikely to change across relatively short periods of time. However, since 311 requests come from residents in the city, it is more closely related to social factors such as living conditions and public safety. Changes in the number of crime incidents will have an influence on these aspects, but it takes time both for the influence to happen and for the city residents to sense them and react. Thus, it is reasonable to assume that this takes a longer time compared to crime incidents' influence on 311 request resolution time given the complexity behind it.

2.2.2.2 New York City

We took a similar approach when analyzing crime data in New York City. The time series plot of the number of crimes over time is demonstrated in the plot below:

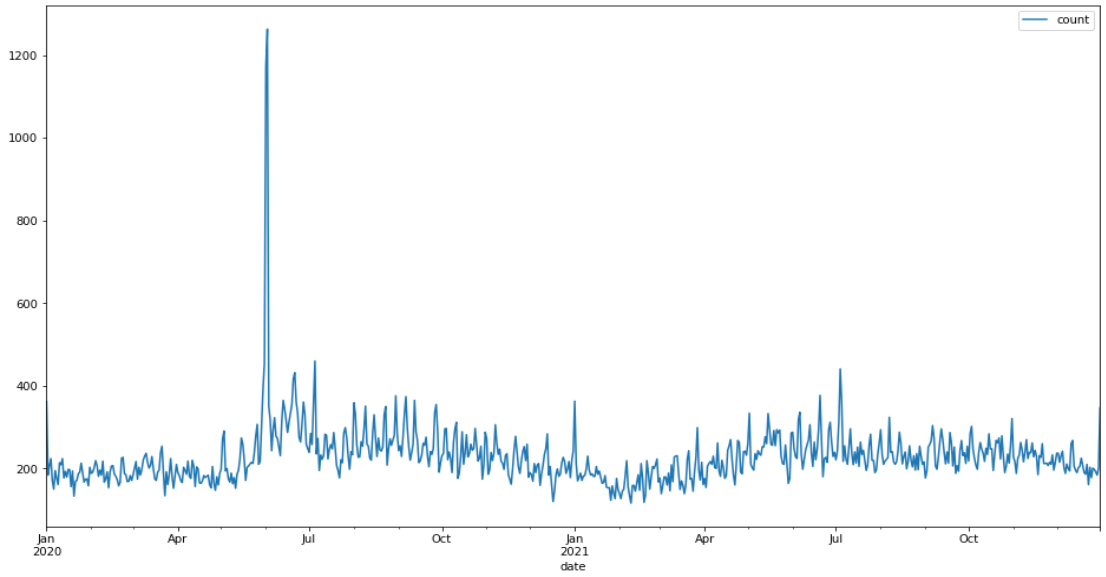

Fig 10. Number of Crime Incidents Per Day, New York City

To better visualize the trend of this data, we used time series decomposition to separate trend, seasonality and noise:

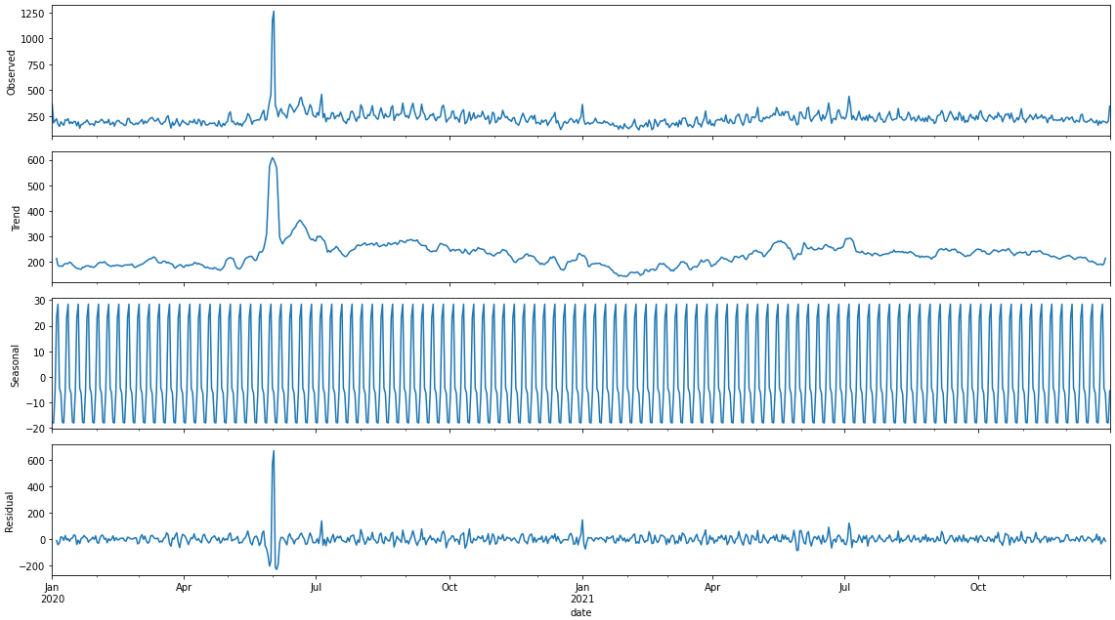

Fig 11. Time Series Decomposition of Crime Incidents Data, New York City

There is a sudden spike in June 2020. However, other than this, we can see that the data has a different pattern than the data for Washington D.C. In New York City, the number of incidents experiences a gradual increase in the first half of the year, and months from July to October generally see more crime incidents than other times of the year.

Next, we also plotted the Time Lagged Cross-Correlation between the number of crimes and trend of the volume of 311 requests after decomposition.

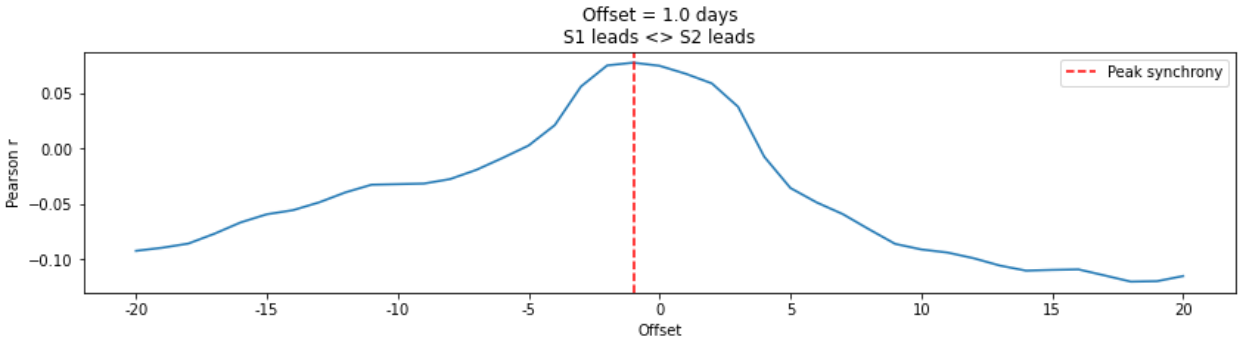

Fig 12. Time Lagged Cross-Correlation Between Number of 311 Requests (S1) and Crime Incidents (S2), New York City

It is noticeable that the magnitudes of correlations are low (below 0.1), which shows that in New York City, the influence of crimes on 311 request volume is very limited. This might be a result of the insensitivity to crimes among NYC residents, unique policies of the city, or the existence of other stronger factors that influence 311 request volume. In either case, we decided that it is more reasonable for us to avoid interpreting information from the correlation graph in order to avoid misleading conclusions.

We then looked at the correlation between crime and 311 requests resolution time:

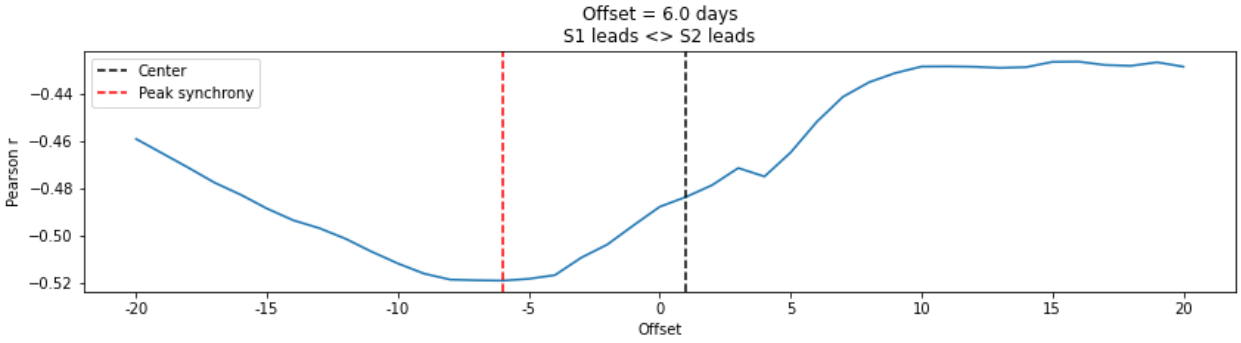

Fig 13. Time Lagged Cross-Correlation Between Average 311 Requests Resolution Time (S1) and Crime Incidents (S2), New York City

The conclusions are very different from the ones for Washington D.C. Firstly, in New York City, the average 311 request resolution time leads to the interaction between itself and crime incidents. Secondly, the correlation is negative, meaning that a longer resolution time is associated with less crime. Combining these two points, we suspected that the increase in crime would warn related government agencies in NYC about problems in the city and make them prepared for future complaints. In this case, the resolution time for future 311 requests is likely to be shorter. On the contrary, if the agencies are not able to obtain warnings because of the low volume of 311 requests, the resolution time can be expected to be longer.

2.2.3. Covid-19

Covid-19 has been influencing almost every part of people's lives since the start of the pandemics. We are interested in how it affects 311 requests and resolution time. We obtained data from OpenData websites and focused on the number of lives lost and positive cases in the cities. Since we only got cumulative data for Washington D.C. our first step was to calculate the number of lives lost and test positives per day. Following this, similar to how we analyzed crime data, we plotted the Time Lagged Cross-Correlation between covid and 311 request data.

2.2.3.1 Washington D.C.

We expected that among all four combinations that we are interested in, *(lives lost, 311 volume), (lives lost, resolution time), (positive cases, 311 volume),* and *(positive cases, resolution time)*, not all correlations will be significant. But surprisingly, the magnitudes of all four correlations are between 2% to 4%, indicating that none of the correlations is strong enough for us to do analysis. We interpreted this as a demonstration of the good performances of Washington D.C. city government in terms of dealing with Covid-19 - they managed to minimize the impact of cthe ovid on both the volume of 311 requests and the average resolution time.

Here are the correlation plots for Washington D.C.:

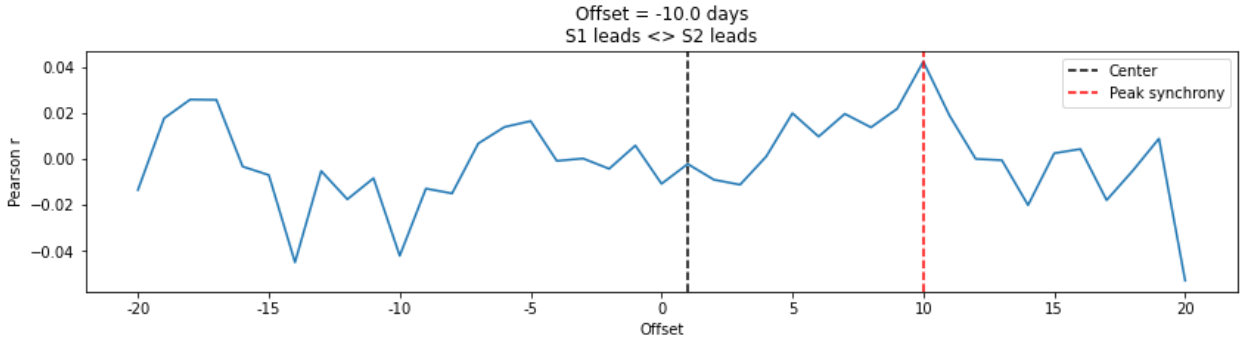

Fig 14. Time Lagged Cross Correlation Between Number of 311 Requests (S1) and Positive Cases (S2), Washington D.C.

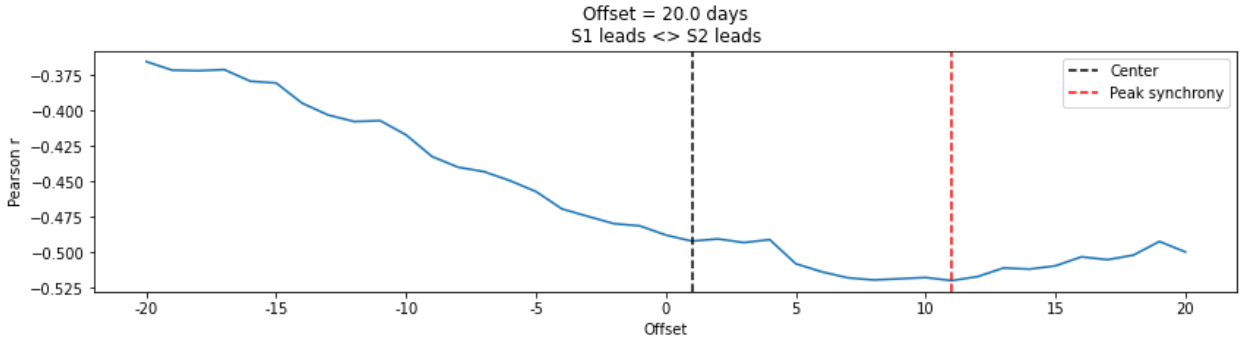

Fig 15. Time Lagged Cross-Correlation Between Average 311 Requests Resolution Time (S1) and Positive Cases (S2), Washington D.C.

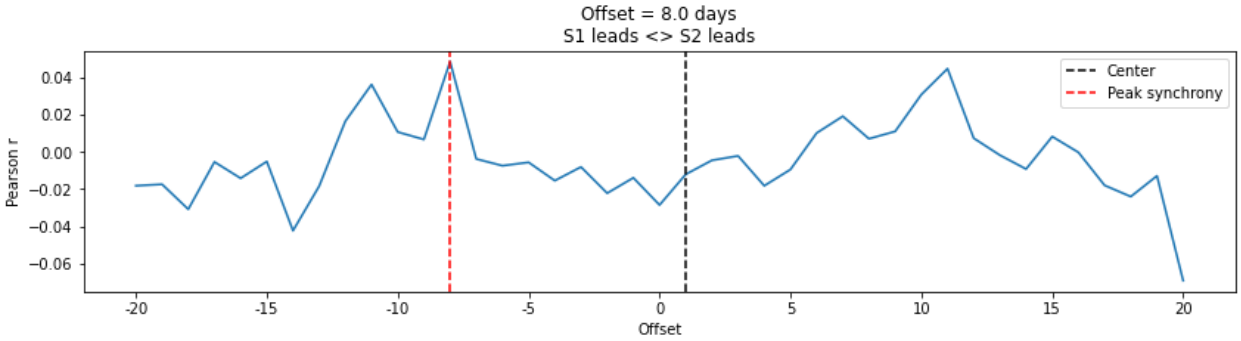

Fig 16. Time Lagged Cross Correlation Between Number of 311 Requests (S1) and Lives Lost (S2), Washington D.C.

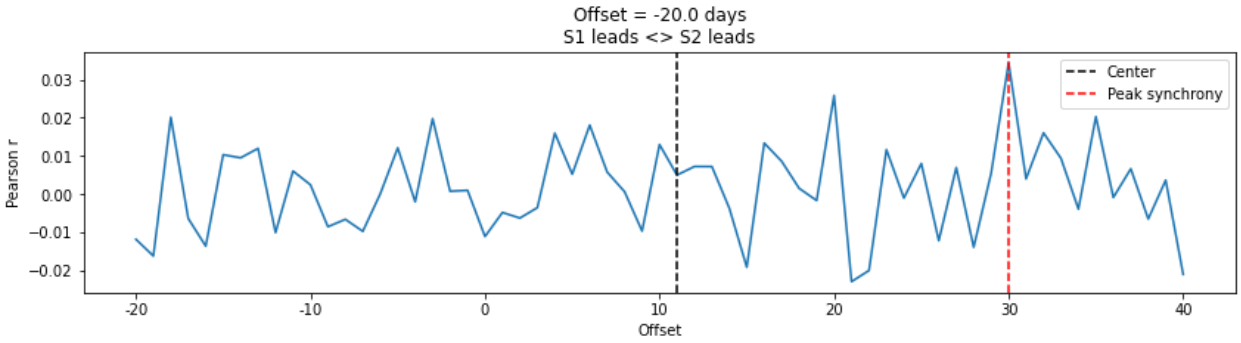

Fig 17. Time Lagged Cross-Correlation Between Average 311 Requests Resolution Time(S1) and Lives Lost (S2), Washington D.C.

2.2.3.2 New York City

We used a similar approach for New York City. Different from the data available for Washington D.C, OpenData NYC provides daily death and test positive counts. We used this data to calculate the correlations with the number of 311 requests and average resolution time. Our findings are listed below.

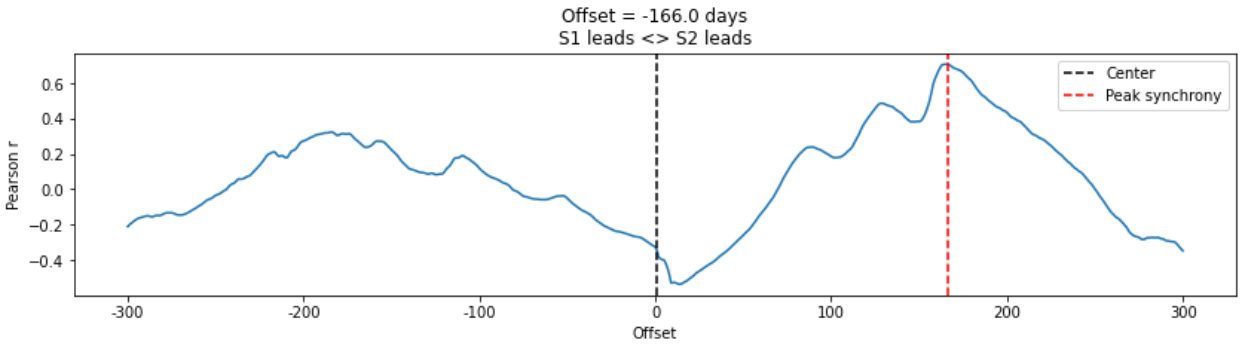

Fig 18. Time Lagged Cross Correlation Between Number of 311 Requests (S1) and Positive Cases (S2), New York City

The plot above indicates that the number of positive cases is the “leader” as well as a positive correlation between the number of 311 requests and the positive cases. It is noticeable that going forward from 0-day lag, the correlation first becomes more negative, meaning that at first, more positive cases are associated with fewer 311 requests. Only after around 50 days does the correlation become positive. During the time in between, the impact of positive cases on 311 requests gradually wears off. Connecting to real-life context, immediately after a surge of positive cases in the city, it is very likely that people redirect their attention on covid protection rather than other minor aspects of their lives, such as noise and garbage collection, which are common types of 311 requests in New York City. As time goes by, people are valuing their living environment again, and this explains why the magnitude of correlation decreases. It is also interesting to notice that the day lags for covid (both positive cases and death counts) are relatively long. We suggested that the government agencies should see this as a warning that the influence of covid lasts much longer than they may possibly expect, and that it is important to keep cautious of this influence for a period of time as long as possible.

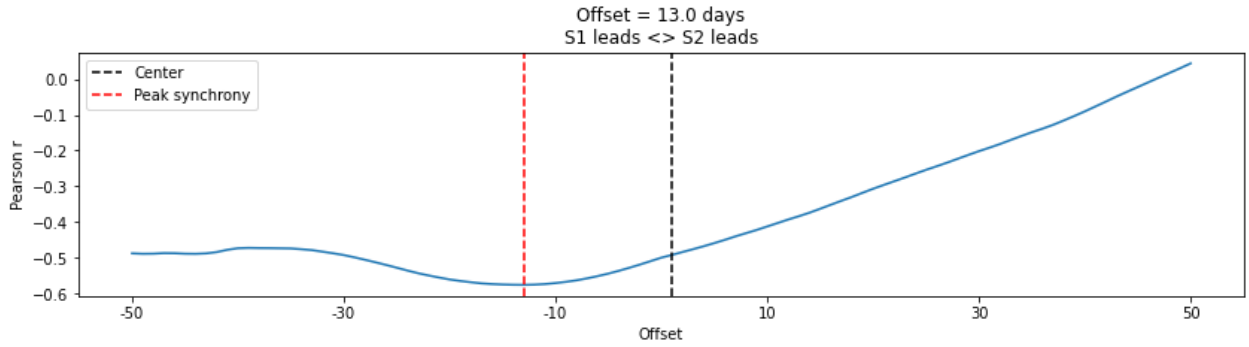

Fig 19. Time Lagged Cross Correlation Between Number of 311 Requests (S1) and Lives Lost (S2), New York City

From this plot above we can get a surprising conclusion that the number of 311 requests can to an extent “predict” the lives lost since it leads to responses in the lives lost time series. Although there is more work needed to validate this assumption (this is not the focus of our report), it is an interesting area to dig into and can potentially offer insights for the city government to reduce lives lost due to covid in the future.

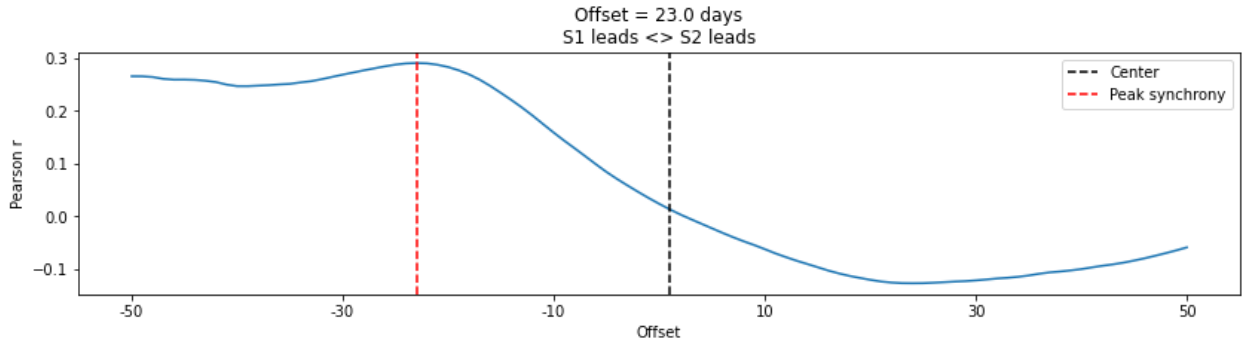

Fig 20. Time Lagged Cross Correlation Between Average 311 Requests Resolution Time (S1) and Lives Lost (S2), New York City

This plot gives us an equally surprising insight that average resolution time can also function as a predictor to the number of lives lost. Even though the correlation is less than the one above, the maximum correlation here happens earlier, which suggests that potentially the government can make use of both variables to forecast the future lives lost. Again, this is only an assumption that we get from the correlation graph, and it is definitely necessary to explore more before putting this method into practice.

Apart from the three correlations above, we didn't see very significant correlation between positive cases and average resolution time. One information that is potentially useful is that the magnitude of the correlation reaches maximum more than 50 days later, which reinforces our earlier point about the length of time during which covid has an influence on 311 related fields.

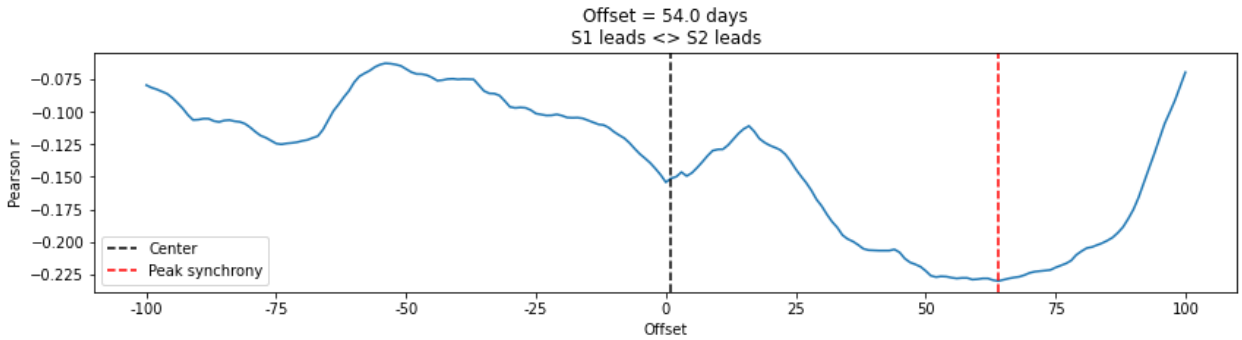

Fig 21. Time Lagged Cross Correlation Between Average 311 Requests Resolution Time (S1) and Positive Cases (S2), New York City

2.4. Modeling and Prediction

Next we draw our attention to predicting future volume of calls on a daily basis. The time series forecasting is considered the most accurate as it is trained on the past 2-year historical data while taking Covid and crime data into account as exogenous variables. On the contrary, given the fact that we do not have a great number of variables to train the regressors, the linear regression and random forest model may be less robust. In this case they are performed as supplemental tools for our time series solution.

2.4.1. Time Series Forecasting

Based on the time series decomposition we had in section 2.2.1, we found that 311 volume calls follow strong seasonality and stable trend. Therefore, we considered employing a univariate time series forecasting model, auto-regressive integrated moving average (ARIMA), to predict the next day's 311 volumes. For more rigorous prediction, we used decomposition to eliminate the seasonality of data during modeling and predicting. On the next page is a plot of the forecasting on NYC 311 volume.

As we can see from the plot, the model fits the time series pretty well and made a reasonable prediction in the end.

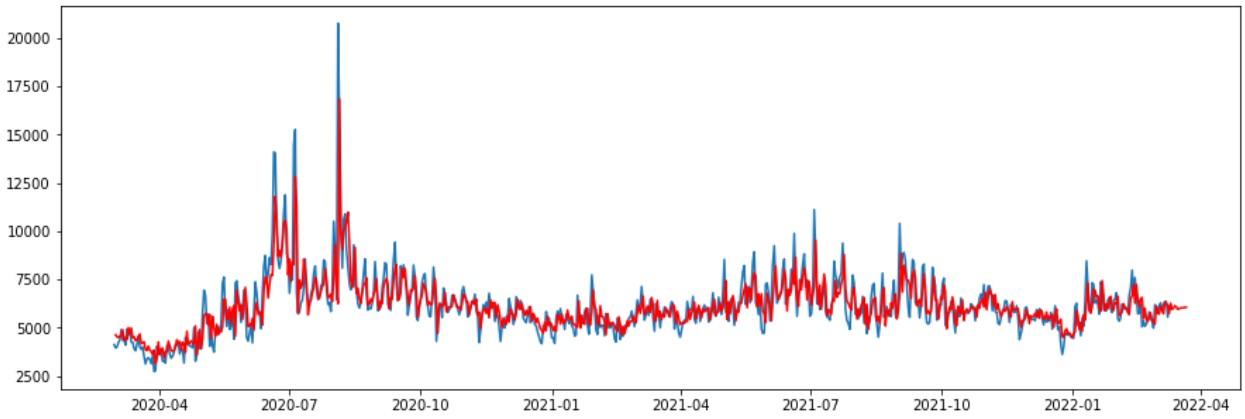

Fig 22. Time Series Model

ARIMA predicts the next-day volume by fitting the distribution of future data conditioned on the lag of the past day, combined with the average window. Here are some statistics about our forecasting:

- **`The next day NYC 311 volume: 5946.3`**
- **`95% Confidence Interval: (5540.4, 6019.7)`**

As a result, the forecast with a confidence interval will give the related department a concrete sense of how much 311 traffic they can expect the next day and whether they can pre-allocate any resources accordingly.

2.4.2. Bayesian Linear Regression

To dig deeper into the 311 request service time with machine learning models, we implemented the most straightforward and commonly used Bayesian Linear Regression model on the full dataset. Regularization is generally required for linear regression to reduce the risk of overfitting. First, we incorporated L1 (Lasso) regularization, penalizing the model by its absolute weight coefficients.

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

We looped through different alpha parameters to find the model optimal predictor. Lasso basically forces the sum of the absolute value of the coefficients to be less than a constant, which forces some of the coefficients to be zero and results in a simpler model.

On the contrary, the Ridge regression takes a step further and penalizes the model for the sum of the squared value of the weights. Therefore, the weights not only tend to have smaller absolute values but also penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed. We repeated the above iterative steps on Ridge as well, and its loss function becomes:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} w_j^2$$

What's more, to combine the pros of both L1 and L2 regularizations, we also implemented a Elastic-Net framework, which basically includes both absolute value penalization and squared penalization.

We utilized root mean squared error (RMSE) as our evaluation metric, which measures the average distance between the predicted values from the regression model and the actual values. After the iterative computation, we noticed that the Ridge model gives the best results with alpha = 0.0005, R-squared = 0.175, and RMSE = 19.72. Considering the fact that the number of predictors we have, this result is quite satisfactory.

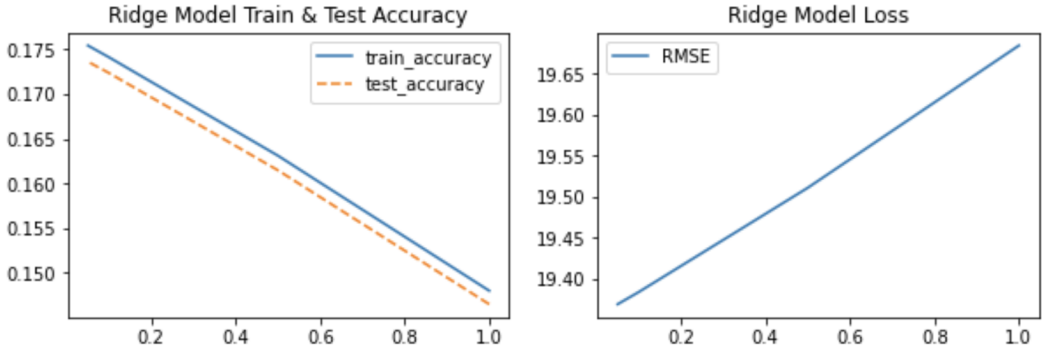

Fig 23. Model Accuracies and Loss

With our best linear regression model, we also plotted the feature coefficients to find what are the important factors that affect resolution latency. We can find in the plot on the next page that most features that are related to different request types and government agencies have larger significance. More specifically:

a) Two features with the largest positive weight are EDC and DOB. These two features represent government agencies, meaning that these two departments typically spend more time on each request. We suggest the governments allocate more resources to handle 311 requests regarding traffic issues.

b) Two features with the largest negative weight are rodent, and covid. These two are request types, meaning that these two types of requests might be easier to resolve in general. To be noticed, although covid-19 brings about a larger volume of 311 calls to some extent, there are not many exact complaints about covid-19, and they are easier to handle as well.

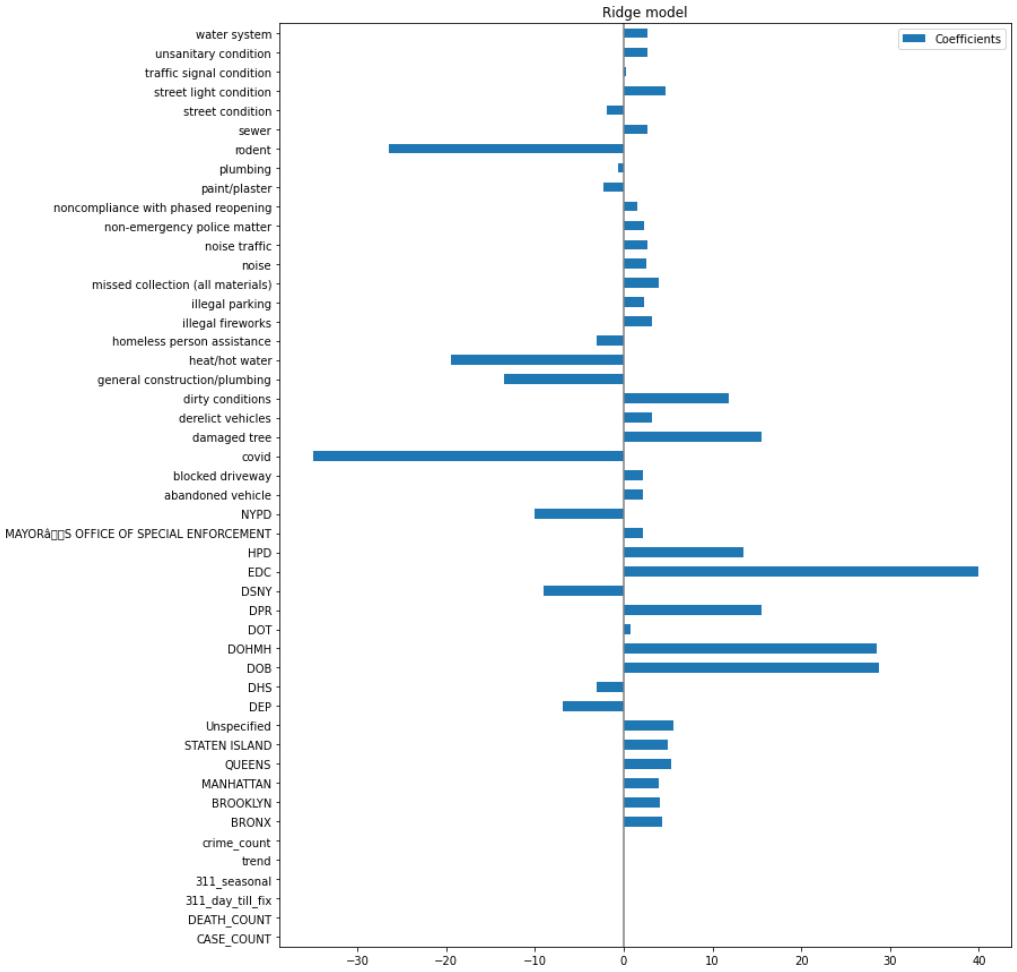

Fig 24. Feature Coefficients

3. Conclusion

Based on all the data exploratory analysis and model prediction performed in previous sections, our time series forecasting ARIMA model appears to be the most robust tool that could potentially assist government agencies to reallocate resources. As it takes in historical data for prediction, the government would be able to input more recent 311 request data to obtain a more accurate prediction for the near future. Our project lends itself to further inquiry into how non-emergency service requests are handled in metropolitan cities like NYC.

4. Data Sources

(NYPD), P. D. (2021, September 13). NYPD calls for Service (historic): NYC Open Data. NYPD Calls for Service (Historic) | NYC Open Data. Retrieved March, 2022, from https://data.cityofnewyork.us/Public-Safety/NYPD-Calls-for-Service-Historic-/d6zx-ckhd

(NYPD), P. D. (2022, February 3). NYPD calls for service (year to date): NYC Open Data. NYPD Calls for Service (Year to Date) | NYC Open Data. Retrieved March, 2022, from https://data.cityofnewyork.us/Public-Safety/NYPD-Calls-for-Service-Year-to-Date-/n2zq-pubd

Crime incidents in 2021. Open Data DC. (n.d.). Retrieved March, 2022, from https://opendata.dc.gov/datasets/crime-incidents-in-2021/explore

DC covid-19 cases by Ward. Open Data DC. (n.d.). Retrieved March, 2022, from https://opendata.dc.gov/datasets/DCGIS::dc-covid-19-cases-by-ward/explore

Department of Health and Mental Hygiene (DOHMH). Covid-19 daily counts of cases, hospitalizations, and Deaths: NYC Open Data. COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths | NYC Open Data. Retrieved March, 2022, from https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3