Citadel Data Open - Team 1 Report

Krishna Sardana, Zhaosen Guo, Yeyubei Zhang, and William Das

March 2022

1 Problem Statement

Crime is a relevant and pressing issue within all large cities in the United States. One of these cities is New York City which has had a surge in crime of about 60% since the same time last year. The broken windows theory is a relevant criminological theory that links disorder and incivility in neighborhoods with subsequent crime: a broken window or an abandoned building can be regarded as an example of disorder within a neighborhood. This disorder, as a result, invites crime into these neighborhoods. Thus, the question arises if there is a relationship between the broken windows theory and crime within a city. While exploring the other datasets provided to us, we investigated businesses that have certain licenses that could *cause* disorder and crime within a community. Liquor licenses emerged as a key category of interest, due to prior research highlighting heightened crime rates in areas with higher densities of liquor stores and retail outlets. Therefore, we sought to answer two key questions in our analysis:

1. What is the relationship between relevant 311 complaints and areas of high crime and what specific complaints contribute the most to the number of crimes within New York City?
2. Can we predict areas of high crime utilizing relevant 311 complaints and number of active liquor-licensed businesses?

Additionally, after noticing that liquor licenses and 311 complaints can help predict crimes, we decided to also explore the internal relationship between liquor licenses and 311 call volume over time. Therefore, we asked an additional question:

1. Is there a causal relationship between the increase of active liquor-licensed businesses and 311 call volume within NYC? If so, can we predict future 311 call volume using the increase of active liquor-licensed businesses?

2 Executive Summary

Using the questions above as guidance, we came upon some interesting insights and relationships within the data. In our first step of analysis, we performed geospatial analysis of the 311 complaints and crime data. We divided the data into a grid system (200x200) where a certain data point belonged to a specific grid. The division of the data allowed us to create a dataset to perform feature importance analysis on. The feature importance revealed that street/sidewalk noise, consumer complaint, homeless person assistance, residential noise, heat/hot water complaints, lost property, drug activity, non-residential heat complaints, face covering violations, and encampment were the top 10 complaint types within the dataset that impacted 311 complaints the most. Many of the these features show signs of disorder and incivility within a community, supporting the broken window theory. Street and sidewalk noise, especially in cities, usually indicates signs of disorderly behavior, which through the broken windows theory, leads to subsequent crime. Additionally, encampment, drug activity, and homeless person assistance are more visible indicators for disorder, inviting more crime to these areas. However, there were also complaints that a city environment is less likely to experience such as improper heating and consumer complaints. These complaints are not as clearly deemed disorderly, but in actuality they fall into this category of behavior. Heating issues and other low standard of living issues drive out residents and drive in crime, as a result. Similarly, consumer complaints regarding businesses is another interesting result from the feature importance analysis. Areas where consumers are not satisfied with the businesses in their neighborhood drive them out and once again, drive crime in. This also supports the broken window theory, which shows that disorder among businesses in the community will once again lead to subsequent crime.

After identifying important features, we developed a model to predict areas of high crime using these relevant 311 complaints: we applied, particularly, Gradiant Boosting methods, training our models on historic 311 complaint and active liquor-permit data, with crime count being the labels. We saw solid results with a 6.34 MAE, 420.2 MSE, and .7 R^2 value.

After our initial analysis, we decided to explore the relationship between active liquor-licenses and 311 call volume over time. Specifically, we investigated how an increase in active liquor-licenses affects the volume of 311 calls in neighborhoods. After performing a multiplicative decomposition analysis on the 311 call time-series dataset, we found a positive trend, as well as a seasonality pattern. Next, after analyzing and visualizing the new active-liquor licenses per day, we noted a peak in license initiations at the beginning of each month. Additionally, in the months between July and October, we noted consistently higher peaks versus the rest of the year. Once we had performed individual analysis, we performed causality testing between the two time-series datasets, using the Granger Causality Test with a time lag of 10 days. The test gave a p-value of .0029 (<.05) meaning that there was enough statistical proof that the data for the 10 previous days of new active liquor-licenses is useful for predicting the 311 call volume for the given day. After evidence of causality, we tried to predict call volume using the increase in active liquor-licensed businesses. Using a VAR (Vector Autoregression) model, we were able to predict call volume 5 days ahead with the lagged call volumes and the active-liquor license data with

10.98% MAPE.

Using these garnered insights, we sought to investigate subsequent actions that the local government in NYC can take: first, focusing on the feature importance results, we note that the local government should continue to focus their efforts on drug control and user-assistance, noise, and homelessness. More programs should be implemented in areas of high crime to assist drug users to seek help and obtain the correct resources in order to overcome their addiction. To address noise complaints in areas of high crime, developing an outreach program educating these neighborhoods on acceptable levels of noise should be implemented. Further, implementing additional programs and developing homeless shelters in areas of high crime to assist the homeless and get them off the streets, will indirectly lower encampment and 311 calls to assist the homeless. These are relevant policy implementations the NYC government can use to mitigate 311 complaints and crime, using our insights.

However, we believe that the NYPD can also greatly benefit from the predictive model we developed. The NYPD is currently greatly outnumbered in terms of crime count and police personnel. This means that the NYPD has to be careful in how they allocate their police resources in order to focus on areas that are in need of the greatest assistance. Therefore, with our predictive model, we can also help the NYPD decide how to properly allocate their police resources across the various boroughs in NYC.

Moving on to the less transparent complaints, heat and consumer complaints, the government will need to take extra initiatives to monitor living conditions within areas of high crime and ensure that proper standards are being met, especially in terms of necessities such as heat or hot water. The Department of Consumer Affairs are in charge of addressing consumer complaints that are reported. The department needs to focus their efforts on addressing the consumer complaints within areas of high crime over areas of lower crime. As our analysis shows, this can assist with the high levels of crime in such areas.

After seeing the strong relationship that exists between liquor licenses and 311 call volume, we believe that the government needs to limit the increase of active liquor-license permits within areas of high crime. It is extremely important that the government works to address this problem. One suggestion is to develop a working group to investigate new active liquor-licenses in areas of high crime. If this group believes that these new liquor businesses are contributing to the crime rate in that area, then they can give a recommendation for removal to the appropriate beverage control personnel.

3 Technical Exposition

3.1 Data and Preprocessing

The data is pulled from three main sources: the NYC 311 Complaints, the NYC Crime Reports, and the Liquor Authority's Active Licenses in NYC. All three datasets have a similar timeframe covering over the past two years.

The 311 dataset has 5,907,009 instances of complaints, and there are a total of 228 different types of complaints within. The distribution of the complaints are extremely skewed (Figure 1), as some of the complaints were reported more than 100,000 times in the past two years (Figure 2), while some complaints only have single digit counts (Figure 3).

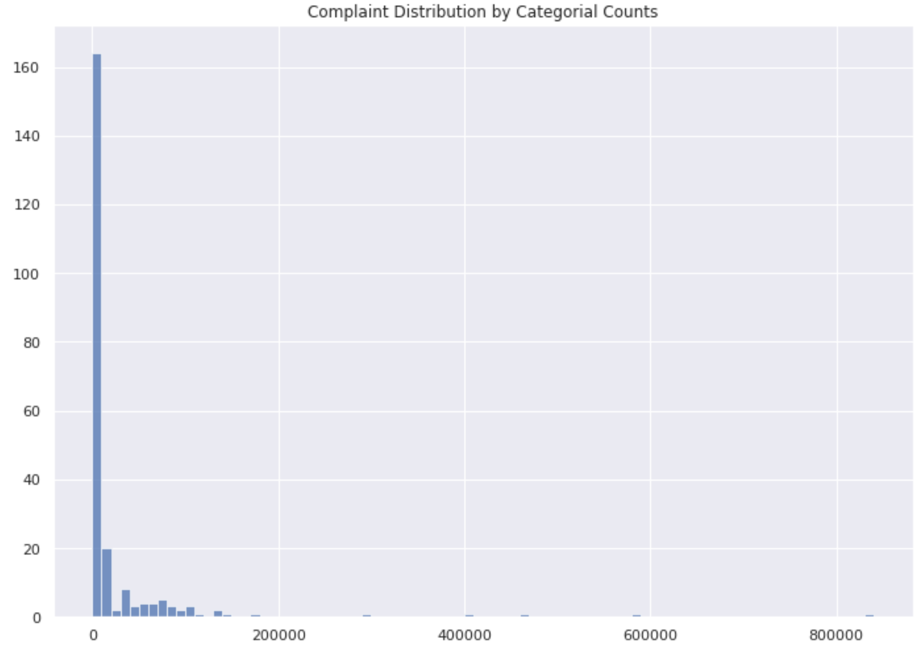


Figure 1: Skewed Distribution - Histogram of Different Complaint Types

The crime dataset consisted of 449,506 records of crimes, including violation, misdemeanor, and felony cases handled by the NYPD. The liquor license dataset contains the current 18,290 businesses with an active liquor license. Additionally, there are dates associated with it, corresponding to when the liquor license became active and will expire. All the data entries among all datasets have corresponding geographic coordinates (latitude and longitude). The total entries of among the complaints, crimes, and liquor license datasets sums up to 6,374,805.

In order to capture the regional representation of the data, an approximated bound of NYC was applied upon all of the datasets. Any datapoints with a longitude outside of [-74.2555913, -73.70000906] or with a latitude outside [40.91553278, 40.4961154] were eliminated from the dataset. There were exactly 22 of these certain points (outliers) in the dataset. Next, a grid was applied to the map and thus, dividing the map into 30,351 cells. Each occurrence of the 311 complaints, crime incidents, and liquor-license businesses are associated with a certain cell in the grid and then aggregated into a new dataset based on their

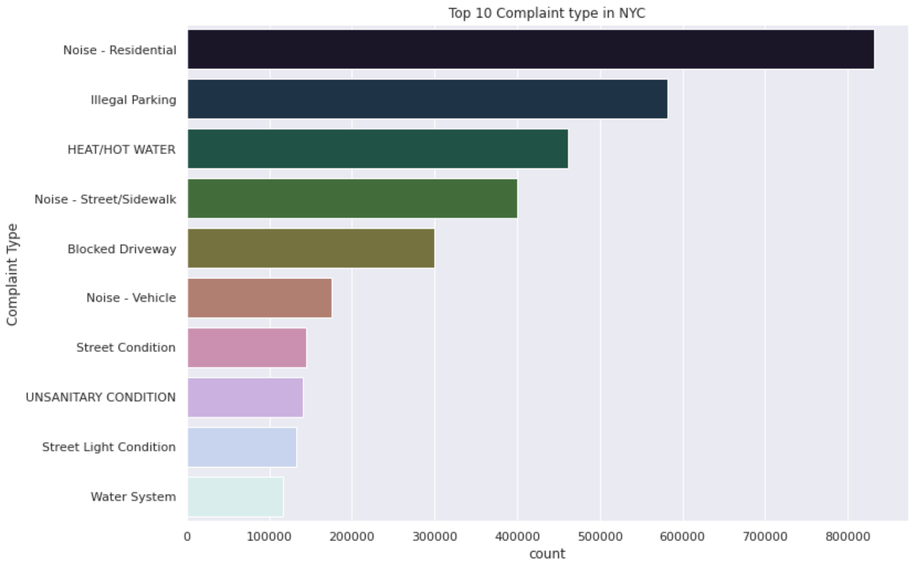


Figure 2: Complaints with High Counts

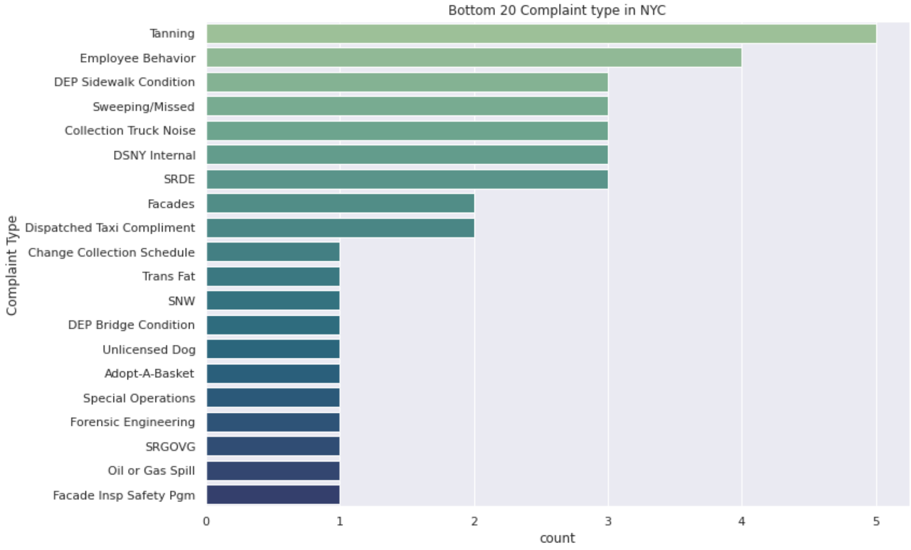
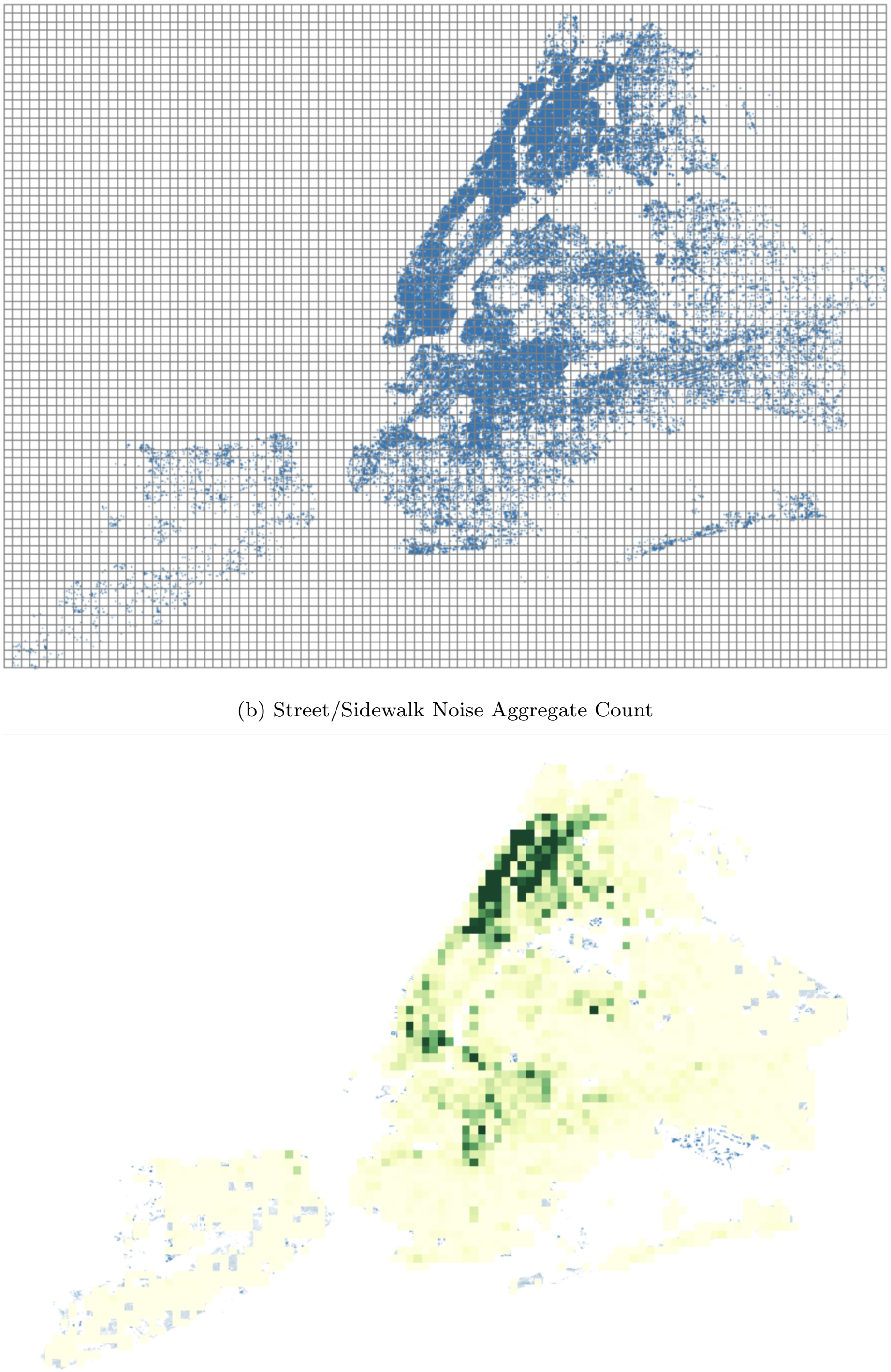


Figure 3: Complaints with Low Counts

respective longitude and latitude information.

Below is an example of the process mentioned above for the 'Noise - Street/Sidewalk' complaint:



(a) Street/Sidewalk Noise Grid

(b) Street/Sidewalk Noise Aggregate Count

3.2 Relationship Between 311 Complaints and Crime

Based on the broken-window theory, we first looked for correlations between complaint types and crime counts in each of the geographical cell that were previously defined. The aggregated data in the end had 30,351 rows for each cell, and 229 columns, which included 228 columns for the counts of each type of the 311 complaint, and one last column for the crime count. Due to the data only spanning over 2 years, there were many columns of complaints types that were 0. With a sparse data set and a high number of features, we standardized the high variance data where for instance, some cells have 6,000 noise complaint, while others have 5. After applying a few different models such as Principal Component Analysis (PCA) and Decision Trees, the Gradient Boosting algorithm provided the best performance as it is robust to sparse data, and is able to optimally select important features to use. With additional hyperparameter tuning, the top 10 most important features that correlates to crimes were extracted. They were as follows: Noise - Street/Sidewalk, Consumer Complaint, Homeless Person Assistance, Noise - Residential, HEAT/HOT WATER, Lost Property, Drug Activity, Non-Residential Heat, Face Covering Violation, Encampment (see Figure 6). This result is meaningful because the complaint types are a direct indicator of areas of high crime and are associated with disorder and incivility within neighborhoods. This directly supports the broken window theory.

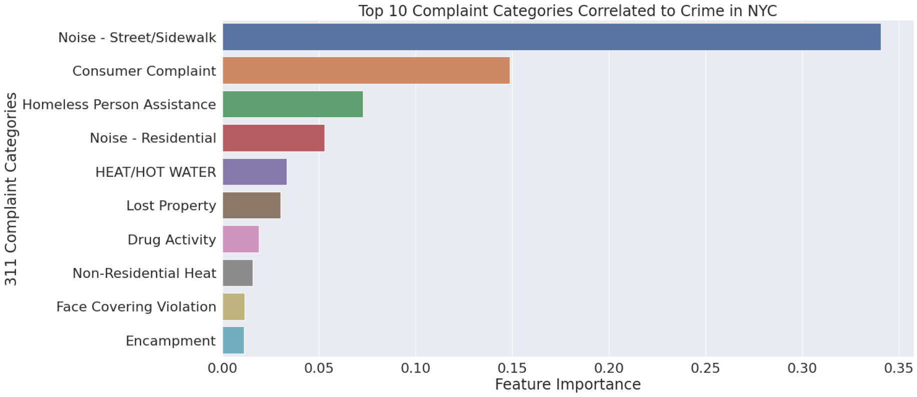
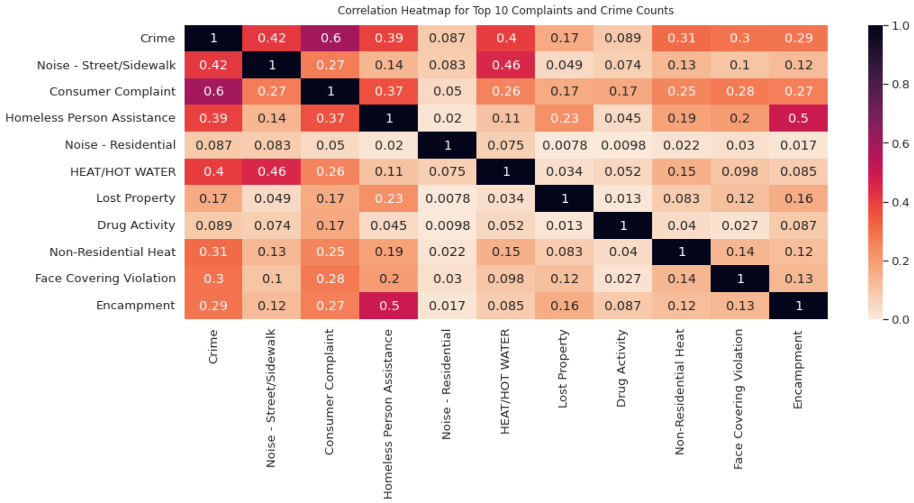


Figure 4: Feature Importance Results

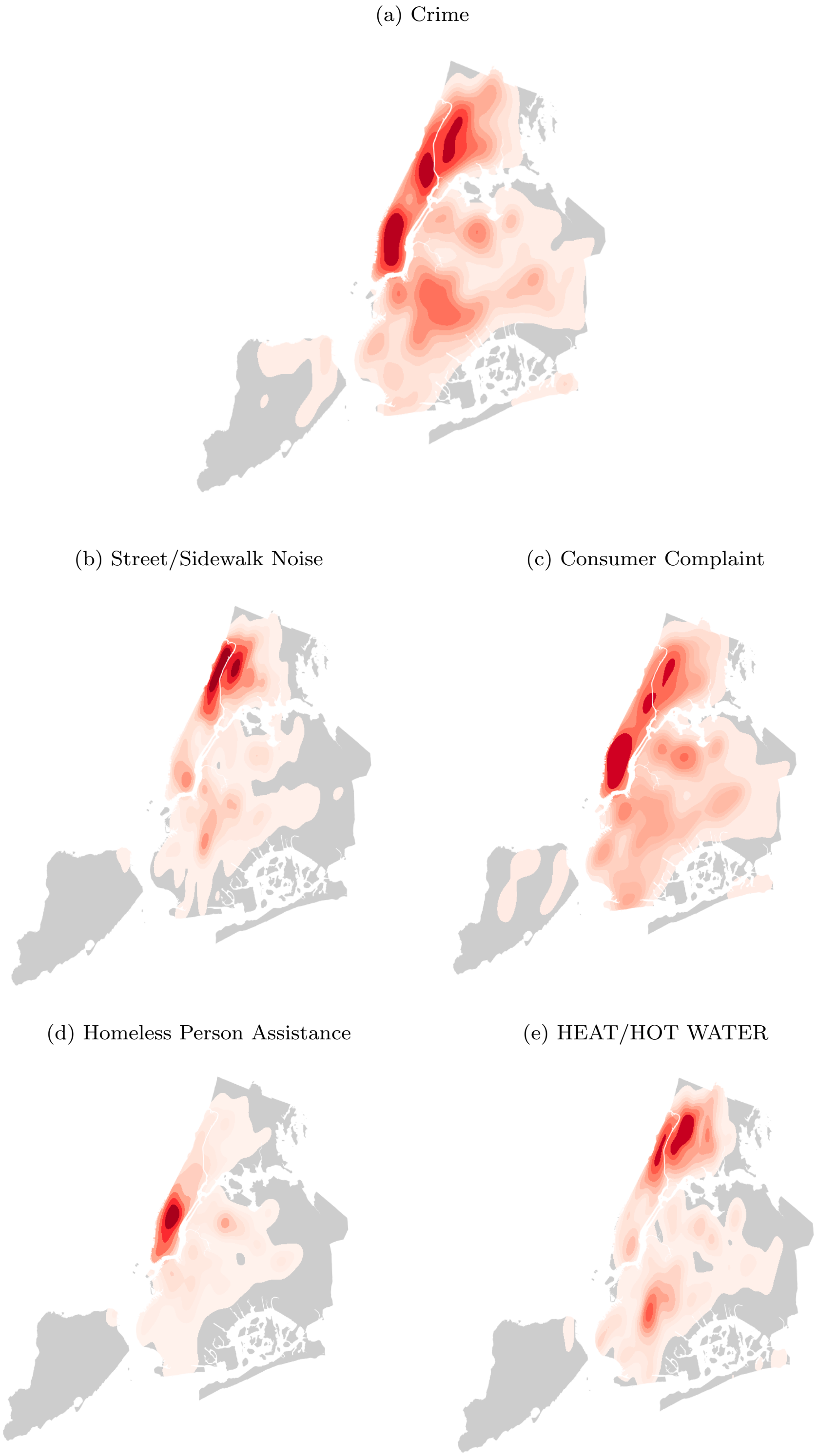
The model employed in this analysis achieved a Mean Absolute Error of 6.3 and R^2 of 0.674 on the testing set. For the complexity and size of the dataset, the performance is good. To further assist the understanding, an annotated correlation heatmap was constructed, with the number of Crime and the top 10 important features, in order of their respective importance. High correlations between Crime and its predictors and can be seen. In the following section is more detailed analysis on how the 311 complaint features, in conjunction with the liquor store density, can be used to predict the number of crimes in a specific



	Crime	Noise - Street/Sidewalk	Consumer Complaint	Homeless Person Assistance	Noise - Residential	HEAT/HOT WATER	Lost Property	Drug Activity	Non-Residential Heat	Face Covering Violation	Encampment
Crime	1	0.42	0.6	0.39	0.087	0.4	0.17	0.089	0.31	0.3	0.29
Noise - Street/Sidewalk	0.42	1	0.27	0.14	0.083	0.46	0.049	0.074	0.13	0.1	0.12
Consumer Complaint	0.6	0.27	1	0.37	0.05	0.26	0.17	0.17	0.25	0.28	0.27
Homeless Person Assistance	0.39	0.14	0.37	1	0.02	0.11	0.23	0.045	0.19	0.2	0.5
Noise - Residential	0.087	0.083	0.05	0.02	1	0.075	0.0078	0.0098	0.022	0.03	0.017
HEAT/HOT WATER	0.4	0.46	0.26	0.11	0.075	1	0.034	0.052	0.15	0.098	0.085
Lost Property	0.17	0.049	0.17	0.23	0.0078	0.034	1	0.013	0.083	0.12	0.16
Drug Activity	0.089	0.074	0.17	0.045	0.0098	0.052	0.013	1	0.04	0.027	0.087
Non-Residential Heat	0.31	0.13	0.25	0.19	0.022	0.15	0.083	0.04	1	0.14	0.12
Face Covering Violation	0.3	0.1	0.28	0.2	0.03	0.098	0.12	0.027	0.14	1	0.13
Encampment	0.29	0.12	0.27	0.5	0.017	0.085	0.16	0.087	0.12	0.13	1

Figure 5: Correlation Heatmap

area. Additionally, in order to visually understand the relationship, we included heatmap visualizations of the crime dataset in addition to important complaint types below:



(a) Crime

(b) Street/Sidewalk Noise

(c) Consumer Complaint

(d) Homeless Person Assistance

(e) HEAT/HOT WATER

3.3 Crime Prediction

After the feature importance analysis, we combined the active liquor-license business counts into our current aggregated data. Then, we used all 229 features to model predicting the number of crimes. After being re-scaled, the data was split into a training set of 20,335 observations and testing set of 10,016 observations. The ratio of testing versus training is a bit higher because the model should be tested more rigorously against sparsity.
We applied **GradientBoostingRegressor** from the *scikit-learn* library once again because of its performance and scalability. The first round of model fitting and predicting resulted in a relatively satisfactory result. The model training/testing was repeated four times for the averaged performance. We provide the following learning curve, scalability metrics, and runtime performance:

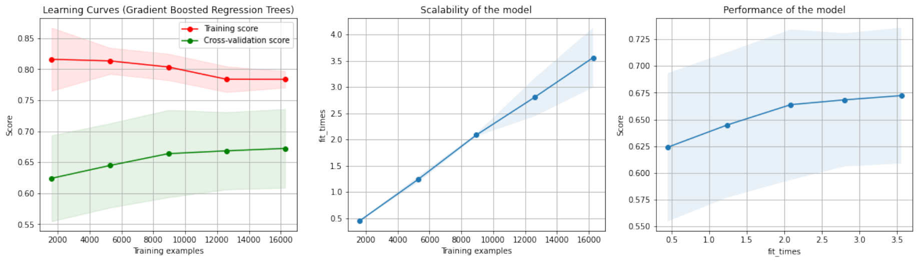


Figure 6: Training and CV, Scalability, Run-time Efficiency of Original Model

As the cross-validation (or testing) score increases, the training curve decreases. Additionally, the rest of the two metrics are not doing so well. Therefore, after hyperparameter tuning, the new model increased the tree depth for a more complex model, increases constraints on making branches/leaves due to the total number of observations, implements sub-sampling to reduce variance, and lowers the learning rate. Fortunately, we saw improved scores all across the board as seen below:

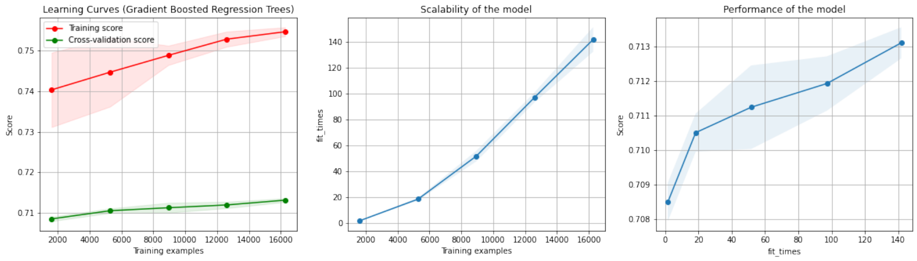


Figure 7: Updated Metrics

The better model's performance on the randomized testing set is as follows: Mean Absolute Error (MAE) 6.34, Mean Squared Error (MSE) 420.20, and R^2

= 0.70.
Another important insight from the finalized gradient boost model is the impact of introducing the liquor-license data as an additional feature in the dataset. The final model, compared to the original model that determined the important features in the dataset, had a decreased MSE value of 230 and decreased R^2 value of 0.03. This means that the additional data helped to explain the variability in the crime distribution in addition to making the tree models more efficient. Furthermore, after the liquor-license data was added, it became the second-most important feature in the model's feature importance list proving that there exists some relationship between the number of liquor licensed businesses and crime.

3.4 Liquor Licenses and 311 Call Volume

3.4.1 Time Series Visualizations

To explore the additional question previously mentioned, we performed time-series analysis between the increase of active liquor-licensed businesses and the total 311 call volume on a daily basis.
Starting with the 311 call volume time-series, Figure 8 shows a 30-day rolling window plot of daily 311 call counts from Jan 2020 to Mar 2022. The rolling window is helpful because it is able to smooth the noisy call volume data. Based on this plot, we observe a slightly upward trend and some seasonality patterns within each year like peaks around July.

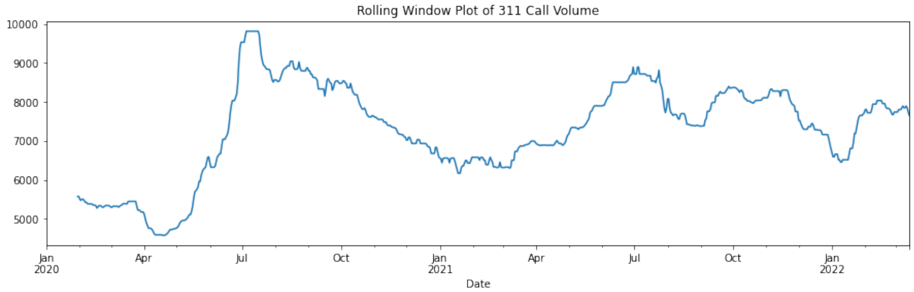


Figure 8: Rolling Window Plot of 311 Call Volume

Next, we performed a multiplicative decomposition by month of this time series, which suggests a constant seasonal pattern within each month and confirmed the upward trend along the time series. (Figure 9)

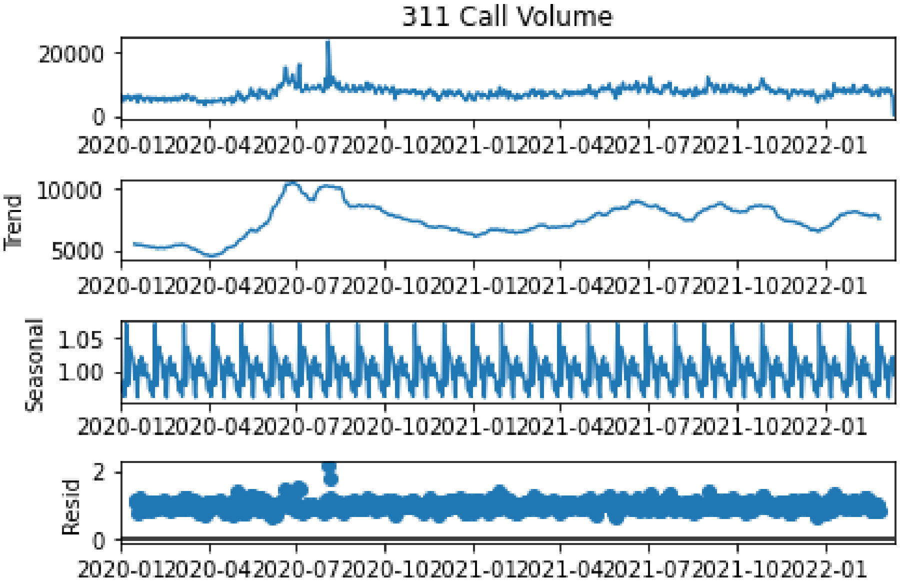


Figure 9: 311 Call Volume Time Series Decomposition

The second time series variable of interest is the increase in daily active liquor-licensed businesses. Figure 10 shows the increase of active liquor-licensed business each day. From this plot, clearly there's a peak at the first day of each month. We can also observe an upward trend and some seasonal patterns such yearly highs between the months of July and October.

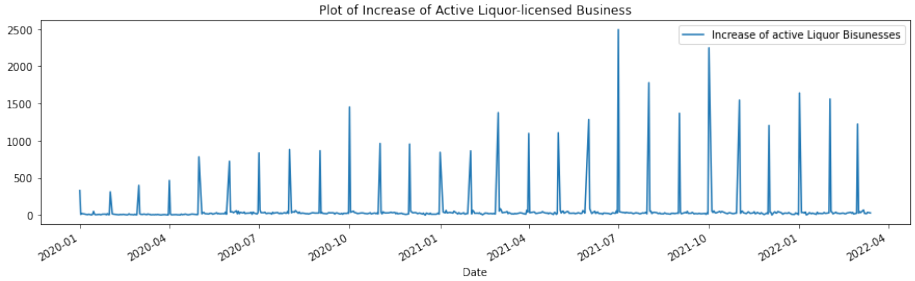


Figure 10: Increase of Liquor-licensed Businesses

To further visualize the two time-series' potential relationship, we visualized synchronous patterns of those two time series (Figure 11). This plot suggests that a rational next step is to test causality and develop a predictive model.

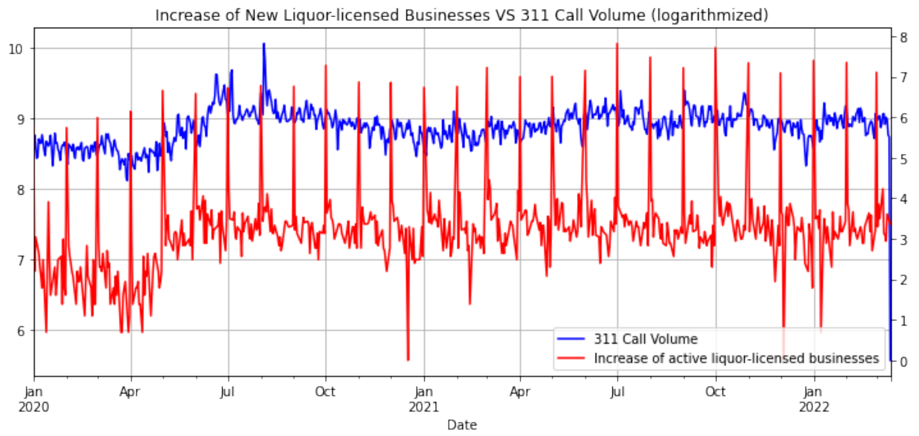


Figure 11: Increase of Liquor-licensed Businesses VS 311 Call Volume

3.4.2 Granger Causality Test

The next logical step is to test for the statistical causality of these two time-series datasets. Specifically, to see if the increase in active liquor-licensed business is a strong predictor of the volume of future 311 calls. To test for causality, we used the Granger Causality Test:

H_0: The increase of active liquor-licensed businesses does not Granger-cause the time-series of 311 call volume.

H_A: The increase of active liquor-licensed businesses does Granger-cause the time-series of 311 call volume.

P-value: 0.05

After running the test with a maximum lag of 10 days, the F test statistic was 2.6987 and the corresponding p-value was 0.0029. Therefore, we reject the null hypothesis and conclude that the increase of liquor-licensed businesses does Granger-cause the time series of 311 call volume. This means that the daily increases of active liquor-licensed businesses within the past 10 days is predictive of the 311 call volume on a given day. For analytical rigor, we also setup this Granger causality test in reverse and failed to reject the null hypothesis. We can then rule out the possibility of reverse causation occurring and confirmed our conclusion above.

3.4.3 Predicting 311 Call Volume

With the confirmed causality, we can build a predictive model for 311 call volume using the increase of liquor-licensed businesses time-series. We developed a

Vector Autoregression (VAR) model to capture the relationship between these two time-series as they change over time.

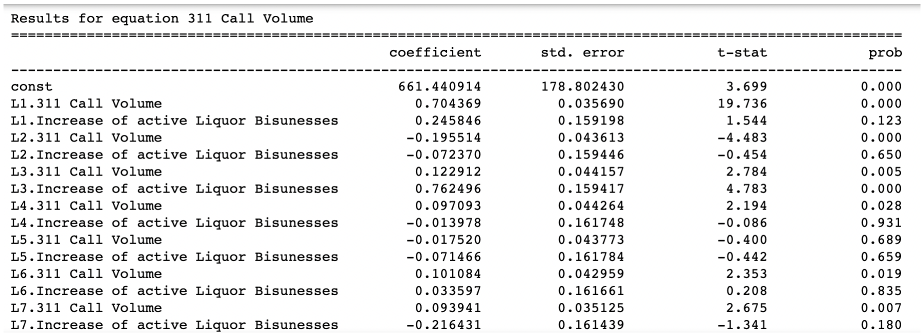
Before fitting the VAR model, we ensured the two time series passed the Johansen Cointegration Test at a 95% significance level, and concluded that they have a long run statistically significant relationship. We also ran an Augmented Dickey-Fuller (ADF) Test for each series and confirmed that they are stationary along the time, indicating that no differencing is needed.

With all of these conditions satisfied, we chose an order level of seven based on the AIC, BIC and FPE scores and developed the VAR model predicting 5 days ahead. Utilizing a training and testing split, the model trained on 798 days and then, was tested on the next 5 days. Figure 12 shows the regression results of the equation for 311 call volume. Through this we were able to obtain the regression equation below using a significance level of 95%:

B_t: Increases of active liquor-licensed business series at day t
V_t: 311 call volume at day t

$$\mathbf{V_t} = 661.44 + 0.7\mathbf{V_{t-1}} - 0.19\mathbf{V_{t-2}} + 0.12\mathbf{V_{t-3}} + 0.76\mathbf{B_{t-3}} + 0.1\mathbf{V_{t-4}} + 0.1\mathbf{V_{t-6}} + 0.1\mathbf{V_{t-7}}$$

From the resulting regression equation, we observed that the increase of the liquor-license time-series on day t-3 is statistically significant in predicting 311 call volume on day t with a coefficient of 0.76. The Durbin Watson's Statistic for this fitted model is 2.0, which indicates that there is no serial correlation of residuals.



```
Results for equation 311 Call Volume
===========================================================================================================
                                             coefficient       std. error           t-stat             prob
-----------------------------------------------------------------------------------------------------------
const                                         661.440914       178.802430            3.699            0.000
L1.311 Call Volume                              0.704369         0.035690           19.736            0.000
L1.Increase of active Liquor Bisunesses         0.245846         0.159198            1.544            0.123
L2.311 Call Volume                             -0.195514         0.043613           -4.483            0.000
L2.Increase of active Liquor Bisunesses        -0.072370         0.159446           -0.454            0.650
L3.311 Call Volume                              0.122912         0.044157            2.784            0.005
L3.Increase of active Liquor Bisunesses         0.762496         0.159417            4.783            0.000
L4.311 Call Volume                              0.097093         0.044264            2.194            0.028
L4.Increase of active Liquor Bisunesses        -0.013978         0.161748           -0.086            0.931
L5.311 Call Volume                             -0.017520         0.043773           -0.400            0.689
L5.Increase of active Liquor Bisunesses        -0.071466         0.161784           -0.442            0.659
L6.311 Call Volume                              0.101084         0.042959            2.353            0.019
L6.Increase of active Liquor Bisunesses         0.033597         0.161661            0.208            0.835
L7.311 Call Volume                              0.093941         0.035125            2.675            0.007
L7.Increase of active Liquor Bisunesses        -0.216431         0.161439           -1.341            0.180
```

Figure 12: VAR model summary report for 311 Call Volume Equation

Using the model and testing dataset to predict 5 days ahead, the performance was as follows: 10.98% Mean absolute percentage error(MAPE) and 810.492 Root-mean-square deviation(RMSE). Figure 13 visualizes the actual versus predicted 311 call volume for days March 9th 2022 to March 13th 2022.

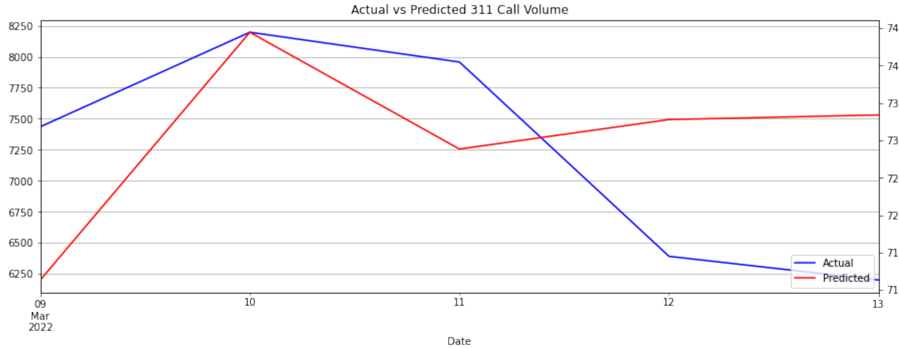


Figure 13: VAR predicting 311 Call Volume 5 days head

4 External Datasets

1. Crime Dataset: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data

2. Active Liquor-License Dataset: https://data.ny.gov/Economic-Development/Liquor-Authority-Current-List-of-Active-Licenses/hrvs-fxs2/data