ISYE 6402- Time Series Analysis
Final Project

Analyzing Violence and Property Crimes in Atlanta And New York

Group 6 – Samesh Bajaj, Eugene Huang, Siddharth Sen, Yash Veljee

SUMMARY

Crimes and violence have long been an issue in parts of the United States, with Atlanta and New York crime rates always being an issue. This study is aimed at analyzing property and violence crimes in Atlanta and New York between January 2009 and December 2020, to explore whether there is any relationship between the types of crimes in the 2 cities, and whether we can use one to predict the other.

Using exploratory data analysis, we discovered that crimes in Atlanta have trended downwards over the past 12 years, while crimes in New York initially rose through the early-2010s, before then trending downwards till 2020. The crime data had a high degree of seasonality and autocorrelation as well, and we have addressed these using the ANOVA and harmonic seasonality models, along with using a square root transform and differencing the data at varying lags.

We then implemented various time series models such as ARIMA, ARIMAX, SARIMAX, LSTM and VAR to make our predictions. ARIMA was used to try and explain our time series based on its own past values and lags. ARIMAX was implemented to explore temperature and holiday data as exogenous factors to make our predictions. SARIMAX was used to deal with the high degree of seasonality in our data. We also explored LSTM (a method outside the scope of the course) as a deep learning framework to capture any patterns in the data. Finally, VAR was used to explore lead-lag relationships between property and violent crimes for both cities, and to see if crimes in one city lead-lag the other. Prediction accuracy for each of these was validated using metrics such as MAPE and PM.

Future studies may include analyzing the data with other factors outside the scope of this study, to further validate the presence of a lead-lag relationship. One could also look at more models outside the scope of our syllabus to try and capture more patterns in the data and improve prediction accuracy.

TABLE OF CONTENTS

PURPOSE OF STUDY

In 1982, two social scientists introduced a theory that leaving non-violent crimes, such as property crimes or open drug use, unchecked would lead to the proliferation of more violent crimes, such as aggravated assault or rape, as it signaled a breakdown in public order and that criminals could get away with more serious crimes.[1] This theory, known as broken windows, was embraced by elected officials and police in New York City in the 1990s after skyrocketing violent crime rates, and was credited with making New York City the safest big city by 2013. A new mayor was voted into office in 2014 and began restricting police enforcement of non-violent crimes after rising complaints of racial profiling and poor relations with local communities.[2] This shift in policing strategy was followed by a rise in violent crimes.[3]

Atlanta, another city that had implemented broken windows policing, experienced something similar after a public declaration in 2018 by the city's police chief that officers would stop responding to shoplifting calls in parts of the city. Violent crime rose in the years that followed, with murder increasing 58% from 2019 to 2021, and rape increasing 39% from 2020 to 2021. This spike in violent crime made it hot button issue in Atlanta's 2021 mayoral race.[4]

Utilizing time-series crime data from New York City and Atlanta, this study seeks to determine if there is a lead-lag relationship between non-violent crime and violent crime, with property crime serving as a proxy for all non-violent crime

[1] https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/

[2] https://www.npr.org/2016/11/01/500104506/broken-windows-policing-and-the-origins-of-stop-and-frisk-and-how-it-went-wrong

[3] https://nypost.com/2020/08/15/how-nyc-used-then-tore-up-broken-windows-policing-goodwin/

[4] https://www.ajc.com/news/local/bill-torpy-large-broken-windows-racist-needed-enforcement-tool/AkknsLDsGjAKA2kTNM05jO/, https://www.wsbtv.com/news/local/atlanta/apd-will-no-longer-respond-to-shoplifting-calls-in-parts-of-atlanta/720842357/, https://www.wsj.com/articles/atlanta-2021-mayor-election-is-dominated-by-crime-problem-11635678001

based on the data available. This study will also explore if there are exogenous time series, such as temperature and holidays, that have a relationship with violent crime, if violent crime exhibits trend and/or seasonality, and if it's possible to forecast violent crime.

EXPECTED RESULTS

Even though broken windows was credited with reducing homicide in New York City by more than 18%, and shootings by more than 15% in the year after police began vigorously enforcing non-violent violations,[5] studies since have been mixed regarding the effect broken windows policing has had on violent crime, with several claiming that there was no effect.[6] Based on the anecdotal evidence that we have seen in New York City and Atlanta, we expect there to be a lead-lag relationship between property crime and violent crime.

Since we expect there to be a lead-lag relationship between property crime and violent crime, we also expect the trend for violent crime in New York City to be mostly flat, or slightly decreasing, in the period from 2009-2016, as non-violent crime was vigorously policed in the preceding two decades, and then rising in the period from 2016-2020, as a newly elected administration rolled back broken windows policing policies. We then expect violent crime to dip in early 2020 because of COVID before resuming its rise. Similarly in Atlanta, we expect the trend to be mostly flat or decreasing from 2009-2018 before rising in the following years, with a brief decrease in early 2020.

A study conducted by the US Department of Justice examining the seasonality in crime for the period from 1993-2010 concluded that there is seasonality in both property and violent crimes, with summer having the highest rates for most types of property and violent crimes and winter having the lowest rates.[7] This study also noted that previous studies associated temperature change with crime rates throughout the year. Based on the results of this study, we expect temperature to have an exogenous relationship with violent crime and for violent crime to exhibit seasonality.

Another study that was published in the *Journal of Criminal Justice* examined the relationship between different types of crime and holidays discovered that both violent and property crimes were significantly related to major/legal holidays such as New Year's Day and Independence Day, and less related to minor holidays such as St. Patrick's Day or Valentine's Day.[8] Based on the results of this study, we expect holidays to have an exogeneous relationship with violent crime.

Finally, a study that was published in *Security and Communication Networks* looked into forecasting Philadelphia's daily and monthly violent crime utilizing Simple Exponential Smoothing, Holt-Winters Exponential Smoothing, ARIMA, and RNN-LTSM.[9] The study's results showed that it was possible to forecast violent crime with a high degree of accuracy, with RNN-LSTM having higher prediction accuracy than other models with a MAPE score of 13.42. Based on this study, we believe we will be able to forecast violent crime for New York City and Atlanta with a high prediction accuracy as well.

RAW DATA

For Atlanta crime data, we sourced the data from the Atlanta Police Department's website, which was spread out across five CSV files.[10] The combined dataset contained 389,590 rows and 18 variables spanning a date range from January 1, 2009, through March 4, 2022. For New York City crime data, we sourced the data from the city's OpenData portal in CSV format.[11] The dataset contained 7,375,993 rows and 35 variables spanning a date range from January 1, 2006 through December 31, 2020.

[5] https://archive.org/details/YearOfChange/page/n1/mode/2up

[6] https://www.nydailynews.com/opinion/ny-oped-break-the-broken-windows-spell-20190526-ulwcdd7fnjg4fgv6dnskls6vhi-story.html

[7] https://bjs.ojp.gov/content/pub/pdf/spcvt.pdf

[8] https://www.sciencedirect.com/science/article/pii/S0047235203000291

[9] https://www.hindawi.com/journals/scn/2021/5587511/

[10] https://www.atlantapd.org/i-want-to/crime-data-downloads

[11] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

The two datasets contained variables that covered when the crime was reported, when it occurred, the location of the crime, and the type of crime committed. The New York City dataset contained additional variables pertaining to the suspect's and the victim's demographic information.

We created the temperature dataset for Atlanta and New York City by writing a Python script to interface with the NCDC Climate Data Online API to pull NOAA weather station data.[12] We used the NOAA weather station located in Central Park for New York City temperature and the NOAA weather station located at Atlanta Hartsfield Jackson International Airport for Atlanta temperature. We pulled minimum temperature and maximum temperature for both locations, averaged them for each day from January 1, 2009 through December 31, 2020, and exported them into separate CSV files for each city.

The holiday dataset that we created utilized both federal holidays and informal holidays.[13] For informal holidays, we selected those with a reputation for increased alcohol usage, as studies have shown that alcohol consumption promotes aggressive behavior and violence.[14] We created a CSV file with a "Holiday" variable that has a 0 for non-holidays and 1 for holidays for the date range from January 1, 2009 through December 31, 2020. The holidays that we selected were New Year's Eve, New Year's Day, Valentine's Day, St. Patrick's Day, Independence Day, Halloween, Veteran's Day, Christmas Eve, Christmas Day, Martin Luther King Jr. Day, President's Day, Labor Day, Columbus Day, and Thanksgiving.

DATA CLEANING

The Atlanta and New York City datasets required significant cleaning, which we performed in Python using pandas. For both datasets, there were nonsensical dates such as "01/14/1019", nonexistent times such as "T", and null values that we had to remove. We also removed rows outside of the date range from January 1, 2009 through December 31, 2020, which is the date range covered by both datasets. We then deleted all the variables that we did not need to conduct time series analysis.

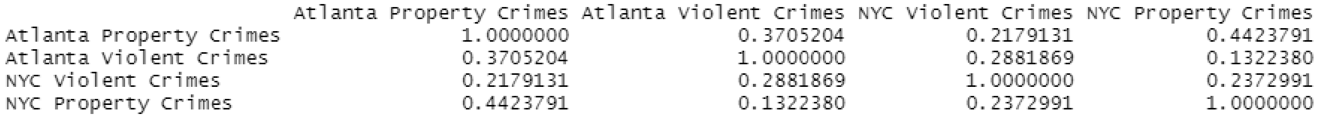
FEATURE ENGINEERING

We conducted feature engineering for the crime type variable in both datasets, 13 types of crime for Atlanta and 72 types of crime for New York City, using pandas in Python to aggregate them into violent and property crime categories as defined by the FBI.[15] Unrelated crime types were dropped. We then used R to count the occurrences of each type of crime, violent or property, for each date of the specified date range for each city, before generating four CSV files, one for each city-crime pairing.

STATISTICAL ANALYSIS

EXPLORATORY DATA ANALYSIS

We initially looked at the correlations between the 4 different datasets and we can see that there is some weak to moderate positive correlation between the different datasets:



```
                        Atlanta Property Crimes Atlanta Violent Crimes NYC Violent Crimes NYC Property Crimes
Atlanta Property Crimes               1.0000000              0.3705204          0.2179131           0.4423791
Atlanta Violent Crimes                0.3705204              1.0000000          0.2881869           0.1322380
NYC Violent Crimes                    0.2179131              0.2881869          1.0000000           0.2372991
NYC Property Crimes                   0.4423791              0.1322380          0.2372991           1.0000000
```

Figure 1 – *Correlation Table Between Cities and Crime Types*

[12] https://www.ncdc.noaa.gov/cdo-web/webservices/v2

[13] https://www.opm.gov/faqs/QA.aspx?fid=e64d74ab-20a3-484c-8682-d2a2b46c22da&pid=c41e6beb-0c14-449d-bde5-355a3a3014cd

[14] https://www.alcohol.org/guides/booziest-holidays/, https://pubs.niaaa.nih.gov/publications/aa38.htm

[15] https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/violent-crime, https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/property-crime

We next look at the time series and ACF plots of the property and violent crime categories in Atlanta and New York City.. We used a ‘square root’ transformation on the 4 datasets and have also done a TS and ACF analysis on the differenced data, with varying lags. We then do a trend and seasonality estimation for each of the 4 datasets:

Atlanta

We observe a decreasing trend in Atlanta violent and property crime over the years with heavy seasonality, which are confirmed by the ACF plot, splines regression for trend, and ANOVA seasonality. Based on the violent crime splines trend line, we a rise in violent crime beginning around 2019 as expected. Differencing by a lag of 7 removes the trend and seasonality for both time series. These observations can be seen in Figures 2 and 3 below:

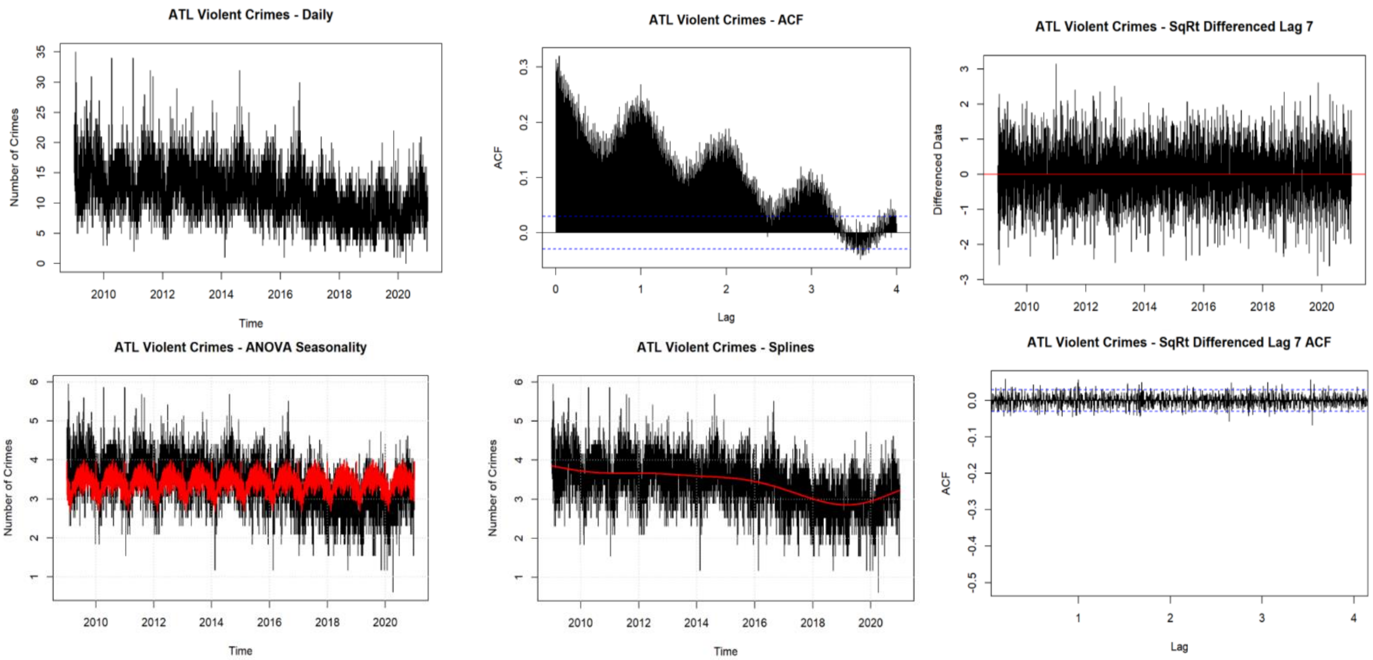


Figure 2 – Atlanta Violent Crime Time Series, ACF, Splines, ANOVA, Difference Lag 7, Difference Lag 7 ACF

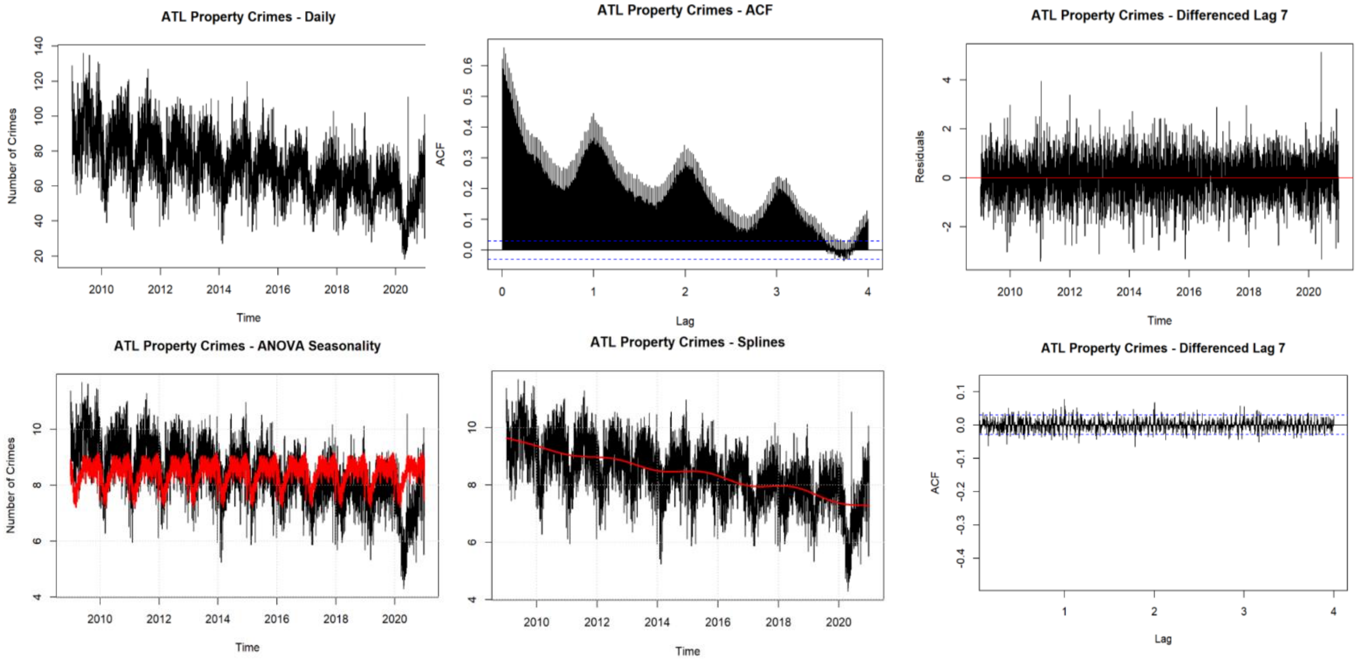


Figure 3 – Atlanta Violent Crime Time Series, ACF, Splines, ANOVA, Difference Lag 7, Difference Lag 7 ACF

New York City

We observe an initially increasing trend in New York City violent crime between 2009 and 2014 before a steady decline between 2014 and 2020. In early 2020, violent crime drops before rebounding towards the later half of 2020. This goes against our expectation that violent crime would be steady, or decreasing, up until 2016. In the case of New York City property crime, there is no clear indication of trend.

There is heavy seasonality for both violent crime and property crime, which are confirmed by the ACF plot and ANOVA seasonality plots. Differencing by a lag of 7 mostly removes the trend and seasonality, however, the end results may not stationary.

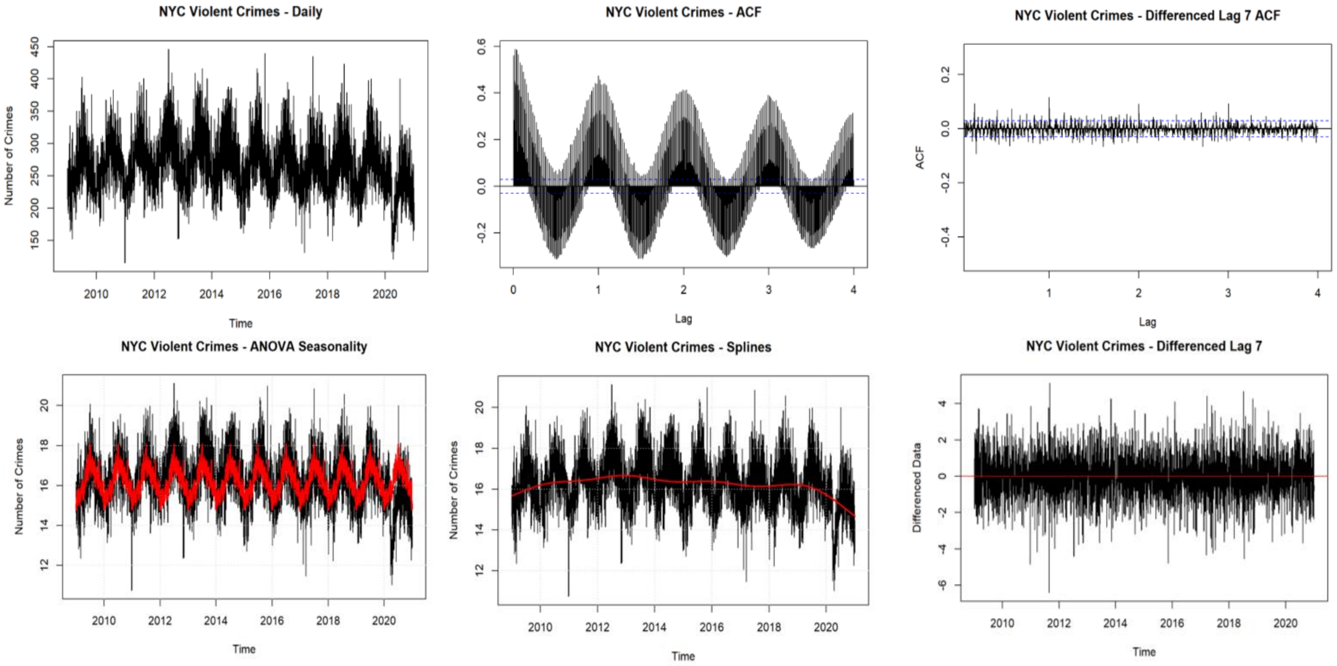


Figure 4 – NYC Violent Crime Time Series, ACF, Splines, ANOVA, Difference Lag 7, Difference Lag 7 ACF

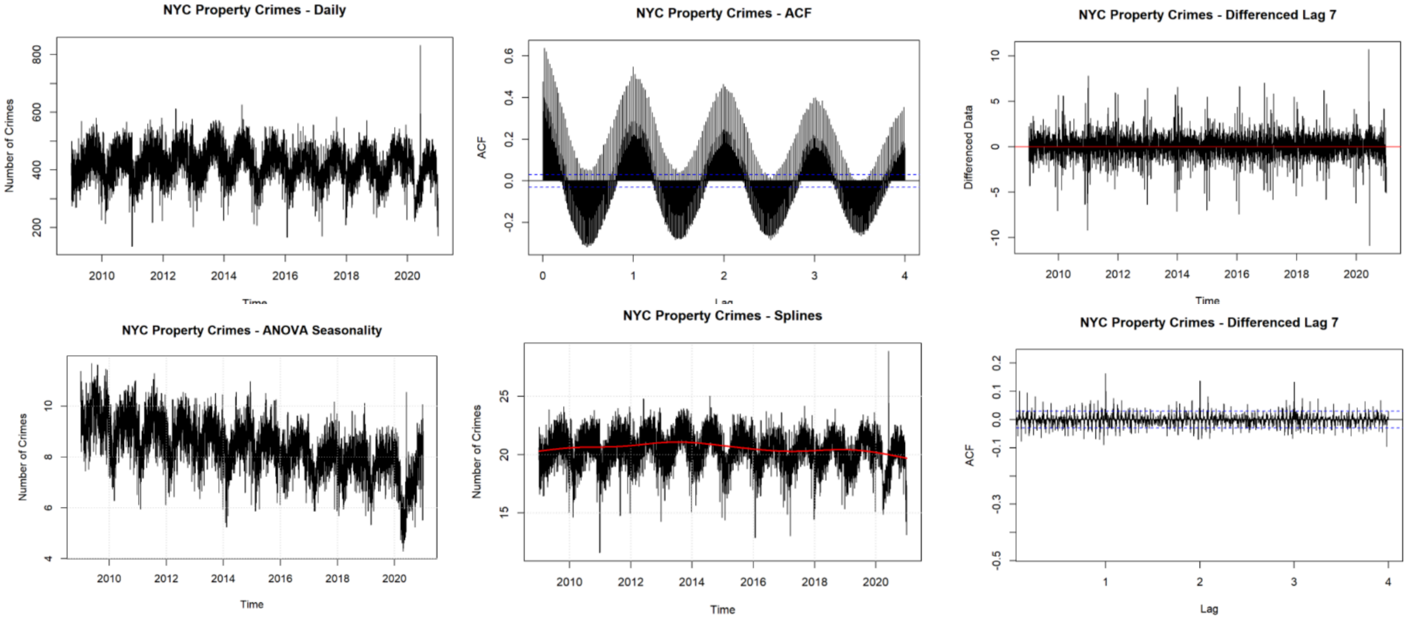


Figure 5 – NYC Violent Crime Time Series, ACF, Splines, ANOVA, Difference Lag 7, Difference Lag 7 ACF

Plotting the splines trendline for property and violent crime on the same plot for both cities does not reveal whether property crime rises and falls prior to violent crime. For New York City, property and violent crime seem to rise and fall in tandem, whereas in Atlanta, we see that property crime has been decreasing at a faster rate and has periods where it rises when violent crime does not.

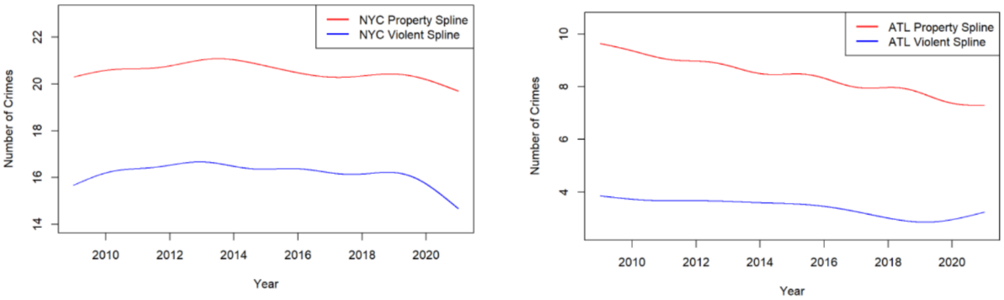


Figure 6 – NYC Property vs. Violent Splines (Left) Atlanta Property vs. Violent Splines (Right)

MODELS

We decided to run the following models: ARIMA, ARIMAX, SARIMAX, LSTM, and VAR. Because we did not see unconditional variance or heteroskedasticity in the data, we decided to skip running any GARCH models, which analyzes volatility within time series data.

ARIMA

In our quest to fit our time series data for violent crime in the cities of Atlanta and New York City, we first decided the to the ARIMA model for fitting the time series and thereby making predictions. ARIMA (Auto Regressive Integrated Moving Average) is a class of models that explains a given time series based on its own past values, its own lags, and the lagged forecast errors. However, any non-seasonal time series that exhibits patterns and is not a random white noise can be modelled with ARIMA models. An ARIMA model is characterized by 3 terms p, d, q where:

- p is the order of the AR term (no. of lags of Y to be used as predictors)
- q is the order of the MA term (no. of lagged forecast errors)
- d is the number of differencing required to make the time series stationary

Model Fitting: To implement the ARIMA model, we first had to split the data into train and test sets. We decided to set the last 7 days of the time series as the test set i.e., from 25th Dec to 31st Dec 2020, whereas the remaining data from 1st Jan 2009 to 24th Dec 2020 was determined to be our training set. To fit our ARIMA model, we need to select the optimal orders for p, d, and q. This was done in an iterative fashion with orders for p and q ranging from 0 to 6 while allowing for a differencing (d) of 1 to 2. We decided to minimize the AIC score to compute the orders.

City	p	d	q	AIC
Atlanta	6	1	6	7444.383
New York City	5	1	6	12777.38

Box-Ljung Test	p-values
Atlanta	0.008418
New York City	2.20E-16

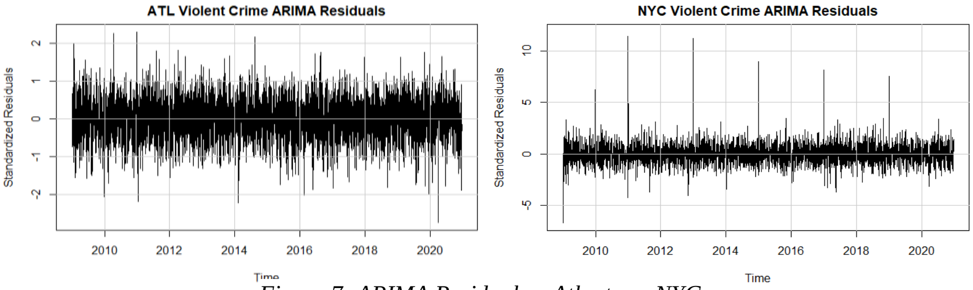


Figure 7: ARIMA Residuals – Atlanta vs NYC

We then proceeded to fit the ARIMA models based on the orders obtained and plotting the residuals from the both the fitted models, we see that the residuals appear to be stationary for both Atlanta and NYC, as shown in the figures below. The ACF plot for the ARIMA model fitted on Atlanta contributes to the same observation of stationarity as there are no significant autocorrelations at any lags after lag 0. The model for NYC does show a few significant spikes at some lags, indicating a possibility of the residuals being correlated and not stationary.

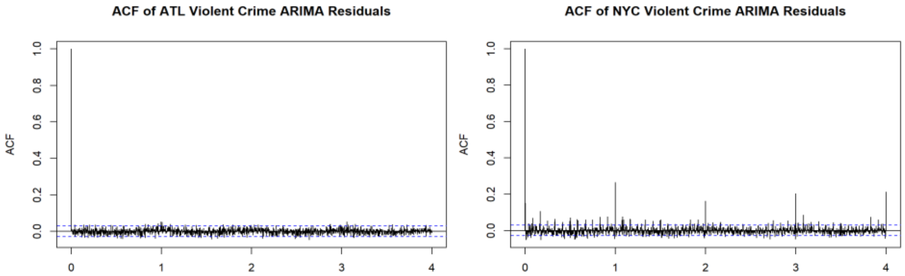


Figure 8: ARIMA Residuals ACF – Atlanta vs NYC

Model Evaluation & Forecasting: The next step is to perform statistical tests to evaluate whether the residuals obtained from the ARIMA model are correlated or not. We conducted the Box-Ljung test to evaluate the same. The p-values obtained from the tests were very low, below the 5% significance level. This indicates that the null hypothesis of uncorrelated residuals needs to be rejected.

This analysis indicates that our ARIMA models may not fully explain the data and perhaps we need to more complex models for our purposes. This makes sense as the ARIMA models are designed to fit any trend present in a time series that is *not seasonal* and stationary.

ARIMA	MAPE	PM
Atlanta	0.09834	1.2142
New York City	0.05693	1.4658

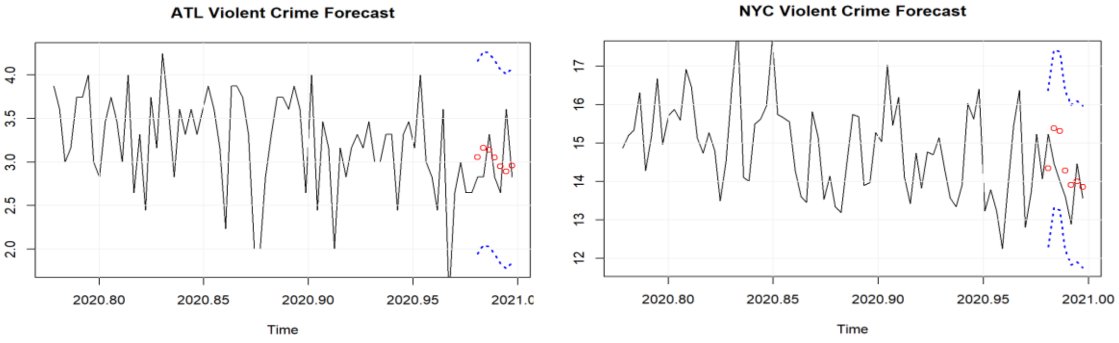


Figure 9: ARIMA Forecasting – Atlanta vs NYC

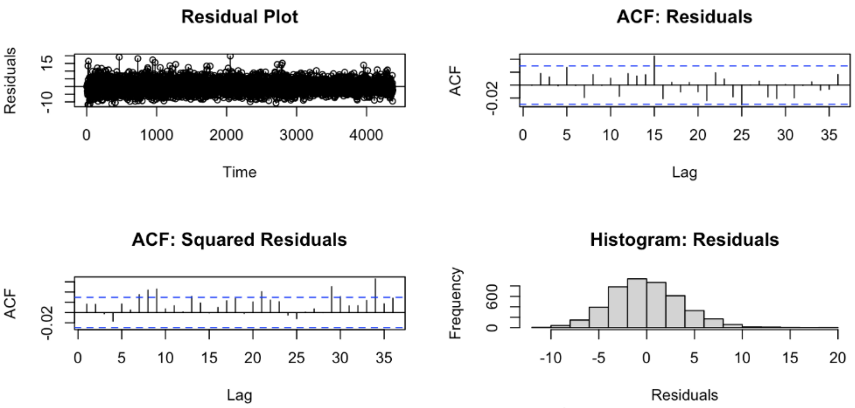
We forecast the last 7 days of our time series data and evaluate it against the test data to obtain a baseline to compare other complex performance where we hope to have better performance. We evaluate the predictions from these models using the Mean Absolute Percentage Error (MAPE) and Precision Measure (PM). The fitted ARIMA models show a PM of 1.21 and 1.46 for ARIMA models for Atlanta and NYC respective. We would like to achieve an PM < 1 or closer to 1 to indicate a good model fit. We see that this is not the case, and therefore intend to improve on our ARIMA models by attempting to fit seasonal ARIMA or SARIMA models.

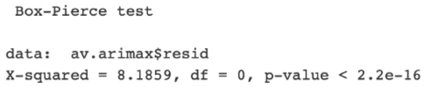
ARIMAX

Crime stems from many exogenous factors, most importantly the macro-economic factors. The macro-economic factors such as unemployment levels, increased income disparity, political rifts etc. These factors change the levels of crime and may cause a jump at yearly levels, on a lag from changes in one of the mentions factors. Our exploratory data analysis didn't show us any such major shifts in crime levels for the scope of our dataset (Atlanta & NYC Violent Crime data from 2009 –

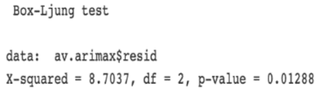
2020) and we saw constant levels of crime with year around seasonality, except some increased variance around covid time. Hence, we ruled out above factors for the scope of our ARIMAX analysis and decided to make forecast based on temperature and the holiday data as our exogenous factors. ARIMAX is an extension of ARIMA, and X represents exogenous variables. ARIMAX takes knowledge from exogenous variable to enhance the prediction accuracy our Y variable (Violent Crime in our case).

Atlanta Violent Crime: We used the untransformed data to run this model as the ARIMAX would take care of the differencing order based on the exogenous factors. The holiday data was changed as a dummy variable. All the selected orders for the model appeared to be statistically significant at 0.1% significance levels, based on the coefficient test below.
The Residual analysis for this model indicates stationarity, which indicates a correct order choice for achieving constant mean & variance. The residuals also seem constant from the time series plot. We don't see any serial correlation based on our squared residuals ACF Plot.

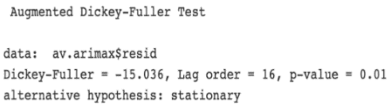




```
Box-Pierce test

data:  av.arimax$resid
X-squared = 8.1859, df = 0, p-value < 2.2e-16
```



```
Box-Ljung test

data:  av.arimax$resid
X-squared = 8.7037, df = 2, p-value = 0.01288
```

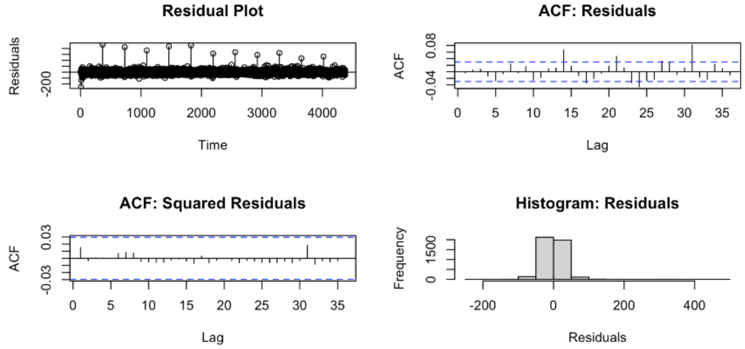


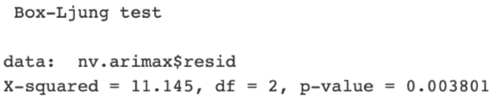
```
 Augmented Dickey-Fuller Test

data:  av.arimax$resid
Dickey-Fuller = -15.036, Lag order = 16, p-value = 0.01
alternative hypothesis: stationary
```

Figure 10: ARIMAX Residual Analysis (Atlanta)

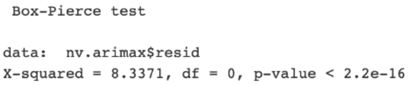
The P-value of Box-Ljung test leads us to rejection of null hypothesis that the residuals from this model are serially correlated & we also reject the null hypothesis of non- stationarity based on P-value for the Dicky Fuller Test.
Both significance tests conclude the residuals of ARIMAX for Atlanta Violent Crime as Stationary.
Moving forward, we used ARIMAX to forecast last 7 days and rolled on each day, to increase the accuracy of predictions. The MAPE (0.188) and PM (0.953) for Atlanta violent crime are both in optimal desired ranges, hence we can confidently believe our predictions to be satisfactorily accurate.

NYC Violent Crime: Like Atlanta Violent Crime, we used untransformed data of NYC Violent Crime to run this model as the ARIMAX would take care of the differencing order based on the exogenous factors. The temperature data was not transformed as well but the holiday data was changed as a dummy variable. All the selected orders for the model appeared to be statistically significant at 0.1% significance levels, based on the coefficient test below.

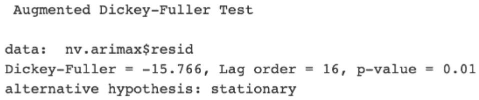




```
Box-Ljung test

data:  nv.arimax$resid
X-squared = 11.145, df = 2, p-value = 0.003801
```



```
Box-Pierce test

data:  nv.arimax$resid
X-squared = 8.3371, df = 0, p-value < 2.2e-16
```



```
 Augmented Dickey-Fuller Test

data:  nv.arimax$resid
Dickey-Fuller = -15.766, Lag order = 16, p-value = 0.01
alternative hypothesis: stationary
```

Figure 11: ARIMAX Residual Analysis (NYC)

The Residual analysis for this model indicates stationarity, which indicates a correct order choice for achieving constant mean and variance. The residuals also seem constant from the time series plot. We don't see any kind of strong serial correlation based on our squared residuals ACF Plot. The P-value of Box-Ljung test leads us to the rejection of null hypothesis that the residuals from this model are serially correlated and we also reject the null hypothesis of non-stationarity based on P-value for the Dicky Fuller Test. Both significance tests conclude the residuals of ARIMAX for Atlanta Violent Crime as Stationary.

Moving forward, we used ARIMAX to forecast last 7 days and rolled on each day, to increase the accuracy of predictions. The MAPE (0.1098) for NYC violent crime is in optimal desired ranges of 0.2 but our PM (1.4812) is a little higher than 1.

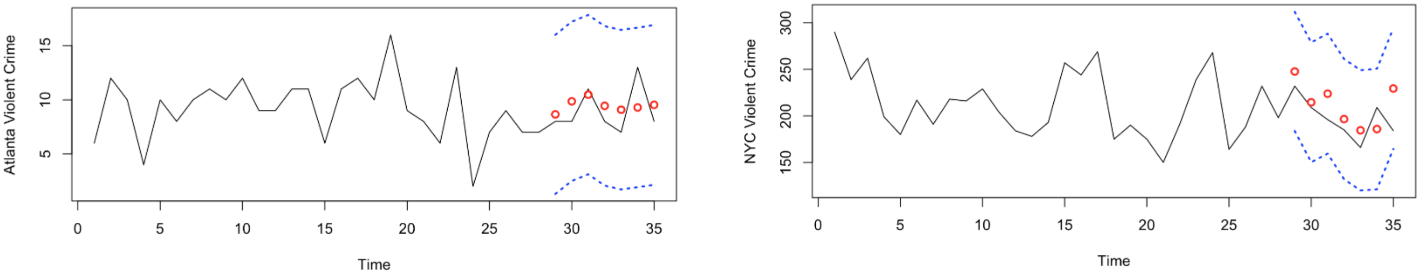
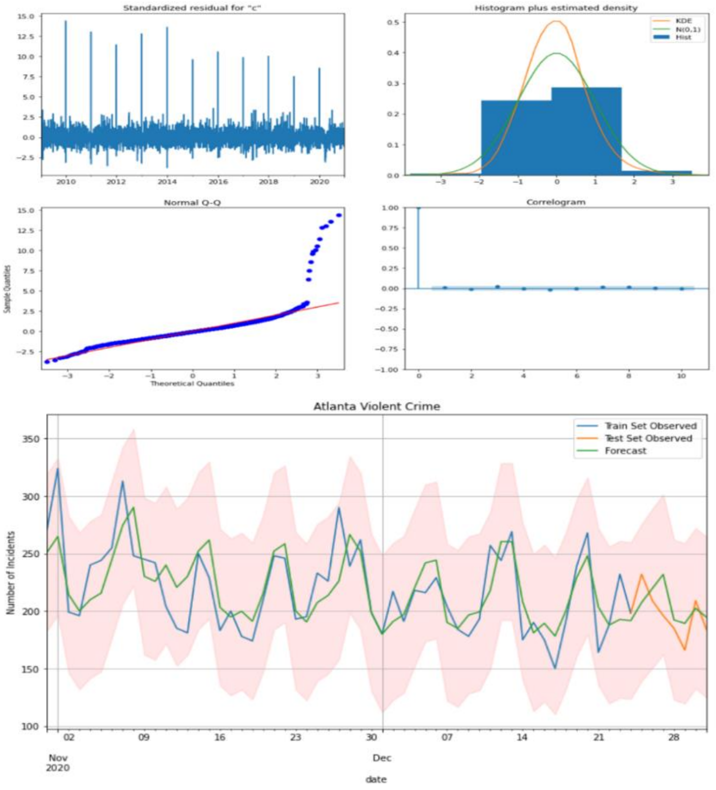


Figure 12: ARIMAX Forecasts – Atlanta vs NYC

SARIMAX



Having a seasonal data and low prediction accuracy scores, we went a step ahead to bringing the SARIMAX as one of the models for this project. Compared to the ARIMAX, the SARIMAX [16]requires 4 additional orders. This is like what a SARIMA is to ARIMA, where we specify 4 additional orders to cater for seasonality. So SARIMAX is just an extension ARIMAX, using exogenous variables to increase the prediction accuracy of Y variable, while forecast the trend and seasonality at the same time. We have used same exogenous variables as ARIMAX, but we used both variables in isolation and together to stress test the forecasting accuracy. We found that Temperature alone does a better job at predicting the violent crime in both cities and adding holidays as a second exogenous variable increases the noise and just ends up decreasing our forecasting accuracy. Therefore, moving forward, we would be sharing the statistics pertaining to the temperature as our only exogenous variable for predictions.

Atlanta Violent Crime: We used the untransformed data to run this model as the SARIMAX would take care of the differencing order for the Atlanta violent crime, Atlanta temperature and holidays datasets. The holiday data is being treated as a dummy variable. The Residual analysis for this model indicates stationarity, which indicates a correct order choice for achieving constant mean and variance. The residuals also seem constant from the time series plot. The P-value of Dicky Fuller test leads us to the rejection of null hypothesis that the residuals from this model are non-stationary. Hence, we conclude the residuals of SARIMAX for Atlanta Violent Crime as Stationary.

Figure 13: SARIMAX Residual Analysis & Forecasts – Atlanta

Moving forward, we used SARIMAX to forecast last 7 days. The MAPE (0.08) being below 0.2 and PM (1) being at the verge for Atlanta violent crime to optimal desired ranges, hence we can confidently believe our predictions to be satisfactorily accurate. The motivation to run SARIMAX served its purpose because the forecast accuracy is way better than the forecast accuracy for ARIMAX.

[16] SARIMAX Code: https://quan-possible.github.io/energy-demand-prediction/daily

NYC Violent Crime: All the data specifications have been exactly like Atlanta Violent Crime and only NY Temperature data is being taken as an exogenous variable for running SARIMAX. The Residual analysis for this model indicates stationarity, which indicates a correct order choice for achieving constant mean and variance. The residuals also seem constant from the time series plot. We don't see any kind of strong serial correlation based on our squared residuals ACF Plot. The P-value of Box-Ljung test leads us to the rejection of null hypothesis that the residuals from this model are serially correlated and we also reject the null hypothesis of non-stationarity based on P-value for the Dicky Fuller Test. Both significance tests conclude the residuals of ARIMAX for Atlanta Violent Crime as Stationary.

Moving forward, we used SARIMAX to forecast last 7 days. The MAPE (0.09) being below 0.2 and PM (1.27) being a little over 1 for NYC violent crime, hence we can't confidently believe our predictions to be satisfactorily accurate. The motivation to run SARIMAX served its purpose because the forecast accuracy is way better than the forecast accuracy for ARIMAX.

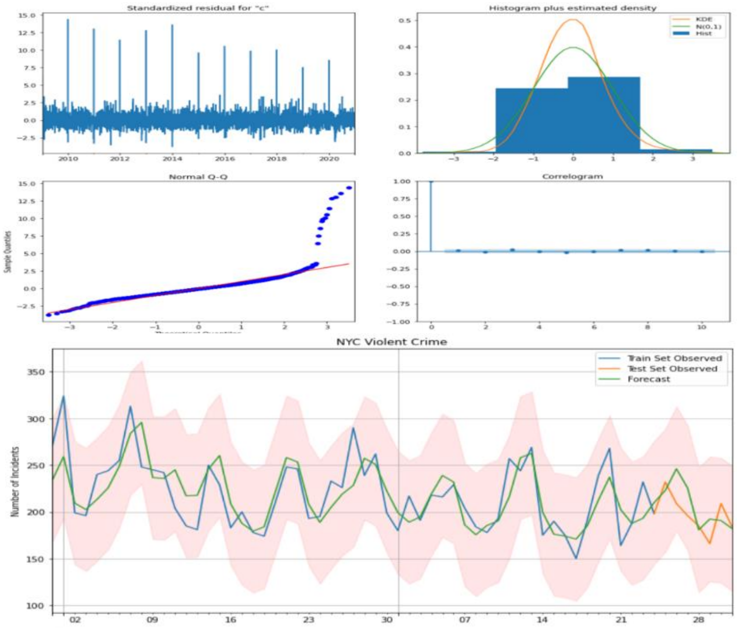


Figure 14: SARIMAX Residual Analysis & Forecasts – NYC

LSTM

We decided to try out and a few implement methods that have not been covered in the scope of the course. One such method for modelling time series data is to use deep learning frameworks such as LSTMs (Long Short-Term Memory) to capture patterns the data and make future predictions and forecasts. LSTM networks, which are a form of recurrent neural networks (RNNs), were designed specifically to overcome the long-term dependency problem faced by RNNs (due to the vanishing gradient problem). This enables LSTMs to process entire sequences of data without treating each point in the sequence independently, while retaining useful information about previous data in the sequence to help with the processing of new data points. As a result, LSTMs are particularly good at processing sequence data such as text, speech, and time-series.

The output of an LSTM at a particular point in time is dependent on three things:

- Cell State – the current long-term memory of the network
- Hidden State – the output at the previous point in time
- Input Data – at the current time step

LSTMs use a series of gates which control how the information in a sequence of data comes into, is stored in, and leaves the network. There are three gates in a typical LSTM; forget gate, input gate and output gate. These gates can be thought of as filters and are each their own neural network. This combination of gates comprises an LSTM cell, as visualized in the following diagram[17].

[17] LSTM Networks | A Detailed Explanation: https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9

Instead of coding the LSTM network by hand, we shall leverage an existing python deep learning library, PyTorch developed by Meta (Facebook) for the purposes of modelling our time series data.

Data Pre-processing: We use the same train-test split as we have used for all our previous analyses of the time series data covered in the report thus far. Thus, our test data consists of the final 7 days in the time series for both Atlanta and New York City.

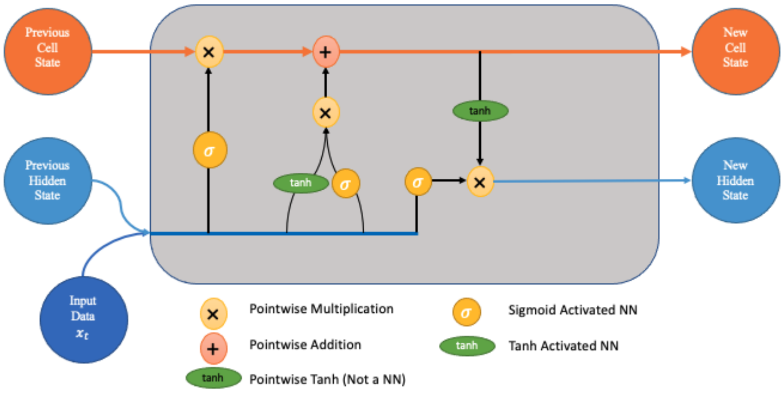


Figure 15: LSTM Unit Cell Schematic Diagram

For fitting our LSTM model, we first need to scale our data into a standard scale (from 0 to 1). This has been done using the MinMaxScaler function available in the sklearn library in python. We then divide our data into sliding windows keeping a window of 7 days. Thus our 1st datapoint is the window from Day 1 to Day 7, 2nd data point as Day 2 to Day 8 and so on. This dataset of sliding windows works as our X variable in the model input. As for the Y variable, we use the n+1-th day's data for every window, i.e., for the 1st window from Day 1 to Day 7, the Y variable will be the value for Day 8. With these data pre-processing steps implemented, we then moved on to define our LSTM architecture and train the model.

Model Fitting: We directly used the torch.nn.LSTM class from PyTorch library to define our LSTM model. Our model contains 1 input layer, 1 hidden layer and finally 1 output layer. The size of our hidden layer was set at 2. We also specified the sequence length to be the same as the length of our windows which was equal to 7.
We used the Mean Squared Error as our loss function, since our prediction task is regression (predicting the actual values). Further, we used a learning rate of 0.01 with 300 epochs for running the gradient descent and training the model.
The configurations were set identical for both the time series for Atlanta and New York City. With our model defined, we proceeded to train two models, one for each time series. Following were the training loss curves obtained from both the models.

LSTM	Hyper-Parameter
# hidden layers	1
Size of hidden layer	2
Sequence length	7
# classes	1
Loss	MSE
Optimizer	Adam
Learning Rate	0.01
# Epochs	300

Figure 16: LSTM Model Hyperparameters

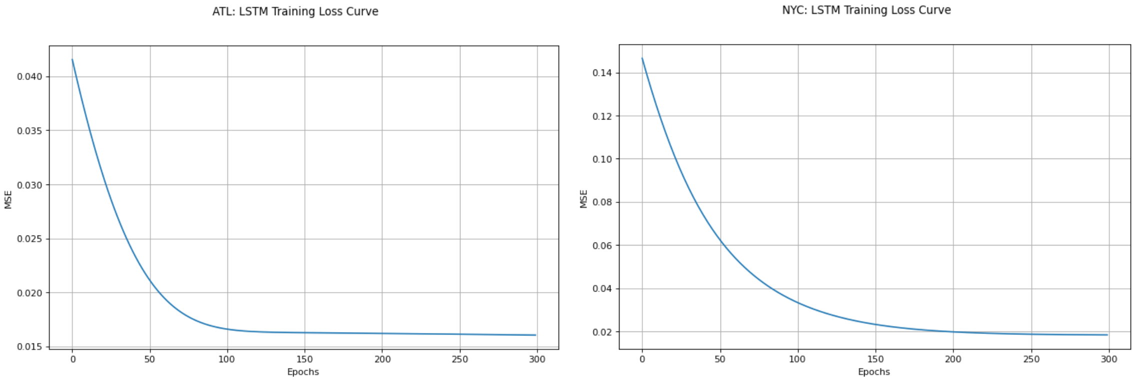


Figure 17: Training Loss Curves for Atlanta and NYC

We see that the loss in both models converge to become asymptotic at later epochs. One thing to note is that the loss took longer to converge, at about 200 epochs for NYC when compared to Atlanta which converged at around 100 epochs. This could indicate that there are more irregularities in NYC data, which caused the model to take more iterations to converge.

Model Forecasting & Evaluation: From the trained models, we generate the 7-day forecast using the training data for both the time series. The predicted values were first inverse transformed to obtain the original scale of each time series. Then the forecasts were plotted against the original time series. The plots for the same are shown below:

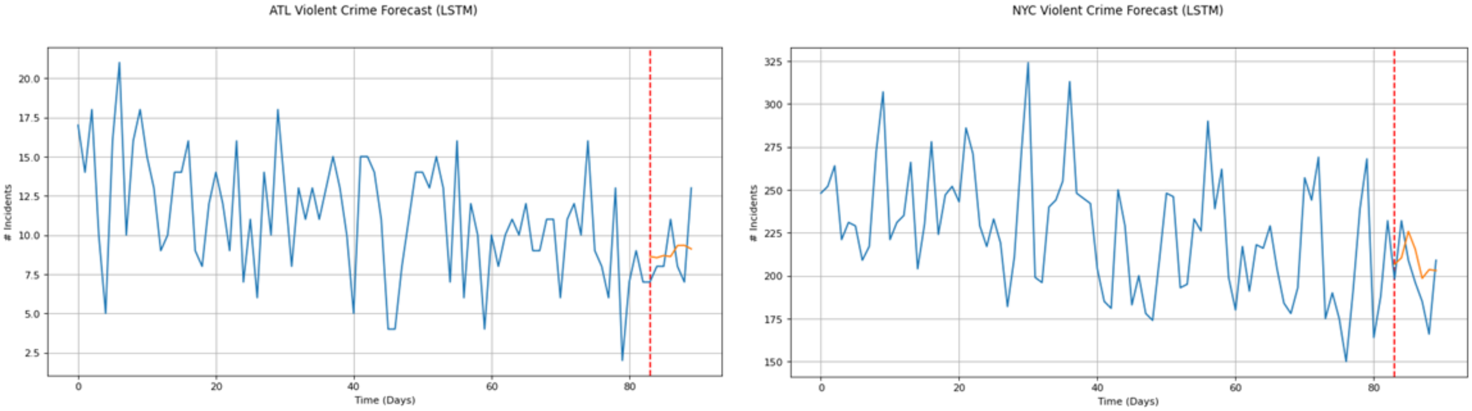


Figure 18: LSTM Forecasts for Atlanta and NYC

We evaluate the predictions from these models using the Mean Absolute Percentage Error (MAPE) and Precision Measure (PM), for both the training data as well as the test data. The observations are as follows:

LSTM Model	MAPE (Train)	MAPE (Test)	PM (Train)	PM (Test)
Atlanta	0.351	0.2	0.792	1.018
New York City	0.101	0.092	0.583	1.093

Figure 19: Model evaluation output

From the metrics above, we can see that the LSTM does a very good job at fitting the training data. The training PMs are considerable below 1 for both the time series indicating a very good training fit. The forecasted data also compared well to the original time series data as the PMs are approximately equal to one for both Atlanta and New York City.
Thus, we can conclude that our LSTM model performs well in our analysis for crime data time series forecasting for the two cities, to an extent that it is able to even surpass the performance of most of the time series models conducted in our analysis thus far. We can thus consider using LSTM in applications for time series forecasting and expect reasonably good predictions without the need to explicitly model trend, seasonality, or conditional variance, unlike the other models we've built up till now.

VAR

We wanted to see if there is a lead-lag relationship between property and violent crime for both cities, and also between Atlanta violent crime and NYC violent crime. We initially plotted the cross-correlation plots between the different time series. However, these plots provided inconclusive evidence regarding any possible lead-lag relationships.

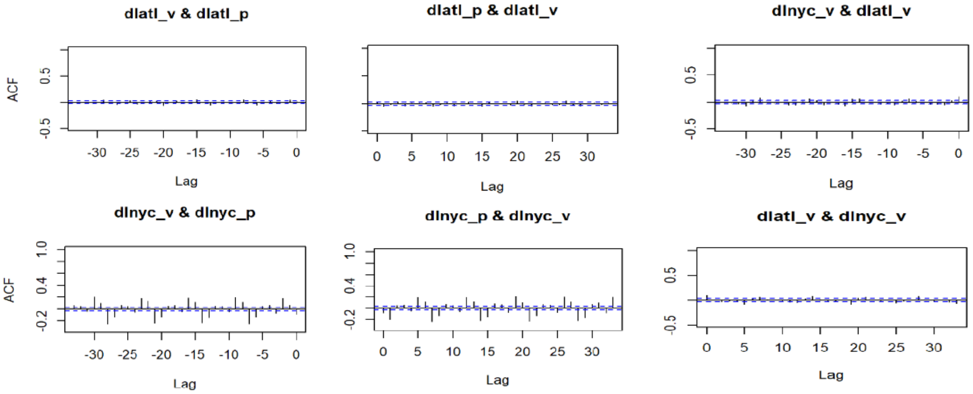


Figure 20: Cross-correlation plots between Atlanta Property-Violent, NYC Property-Violent, Atlanta-NYC Violent

We then proceeded onto the VAR model, which is used for multivariate time series. The structure is that each variable is a linear function of past lags of itself and past lags of the other variables, allowing for feedback to occur between the variables in the model. We decided to use VAR is because we believe that crimes in Atlanta and New York City (both property and violent) may exhibit a lead-lag relationship. We wanted to see if we could capture this relationship and use it for prediction.

Most coefficients for both violent and property crime for both cities were statistically significant, indicating that there was some sort of relationship between these time series. Furthermore, the low p-values in the respective Wald tests allow us to conclude that there is an endogenous relationship between property and violent crime in both cities, and also between Atlanta violent crime and New York City violent crime.

WALD TEST P-VALUES

Influencing Variables (Down) vs Influencee (Right)	Atlanta Property	Atlanta Violent	NYC Property	NYC Violent
Atlanta Property	X	0	N.A	N.A
Atlanta Violent	0.000015	X	N.A	0.00031
NYC Property	N.A	N.A	X	0
NYC Violent	N.A	0	0	X

Figure 21: Wald test p-values for all VAR model runs

Atlanta Violent Crimes Rolling 1-day VAR predictions using ATL property: The MAPE we get for our VAR prediction of Atlanta Violent crimes is: 0.203 and the Precision Measure (PM) is: 1.196.

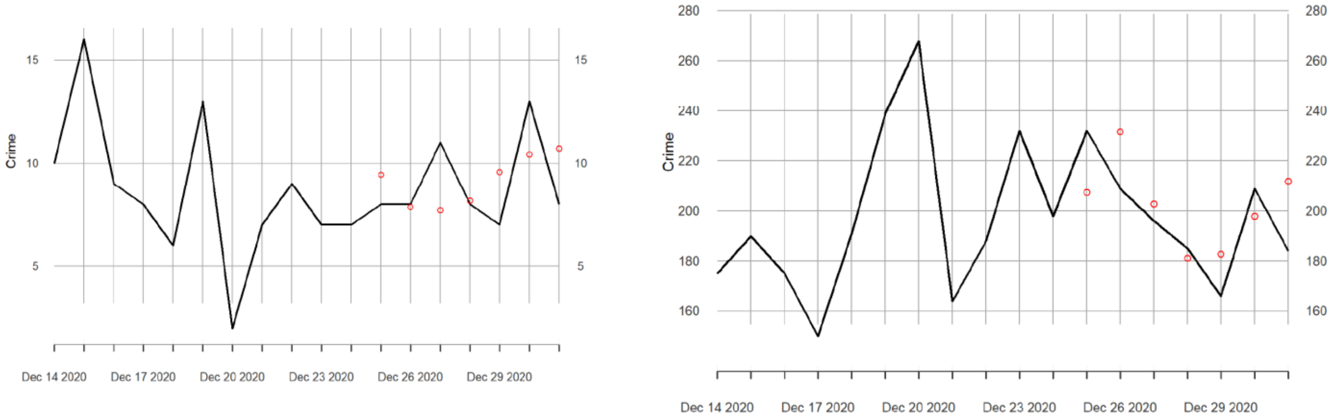


Figure 22: Rolling 1-day VAR prediction for Atlanta & NYC Violent Crime

NYC Violent Crimes Rolling 1-day VAR predictions using NYC property: The MAPE we get for our VAR prediction of NYC Violent crimes is: 0.082 and the Precision Measure (PM) is: 0.839

CONCLUSION FROM ANALYSIS

We undertook multiple approaches to model our time series data for violent crimes in Atlanta and New York City. Our initial approach was to model the data using an ARIMA model. However, the model was not able to capture the seasonality in the data efficiently as ARIMA is traditionally designed only to fit trends. Further, we felt the need to include exogenous factors such as temperature, to see if that had any impact on crime rates in both cities.

Therefore, we attempted to model the data using ARIMAX and SARIMAX. These models offered an improved from the obtained ARIMA results, with an increase in the PM scores. We further modelled our data using VAR analysis to establish any lead-lag relationship between Property Crime and Violent Crime for each respective city. However, we are unable to conclude if the relationship is just between these 2 datasets or other factors at play. In our opinion, there could be other variables that affect violent crimes in both cities, which are not a part of this study. Due to this, we are unable to conclude which way the lead-lag relationship goes in this case.

We also tried deep-learning approaches via LSTM which does a very good job at fitting the training data. The training PMs are considerable below 1 for both the time series indicating a very good training fit. The forecasted data also compared well to the original time series data as the PMs are approximately equal to one for both Atlanta and New York City.

SUBJECT MATTER IMPLICATIONS

We sought to understand if broken windows policing, or the enforcement of non-violent crime, would lead to a reduction in violent crime in this study. We did this through a visual inspection of the splines trend plots, and through VAR analysis, which models the relationship between multiple time series. Both our exploratory data analysis and our VAR model proved inconclusive. While the VAR model proved that there was an endogenous relationship between property crime and violent crime, we were unable to determine if there was a lead-lag relationship between the two. Even though our VAR model results were inconclusive, we can see that temperature and holidays have an impact on violent crime though our ARIMAX and SARIMAX model results. We were also able to forecast violent crime with a great deal of accuracy with LSTM and VAR performing the best, depending on the city and the type of crime.

Based on our study, we believe that police departments across the country, and perhaps the world, can use LSTM and VAR for forecasting violent crime citywide, and in troubled neighborhoods, to adjust their staffing levels accordingly and for budget requests.

As for the debate whether broken windows policing leads to reduced violent crime, our study proved inconclusive, and more studies will need to be completed regarding the matter.

TOPICS FOR FUTURE STUDY

Based on our results, we believe that there are many questions that could be investigated further. For instance, we only explored the time series data at the daily level. Given that timestamped data was provided, future analysis could either scale down to the hourly level or up to weekly and monthly levels for forecasting and to analyze how that impacts the relationships other time series may have with violent crime.

We also only explored whether temperature and holidays have an exogenous relationship with violent crime. There are many other possible exogenous time series that would be worth exploring to see if a relationship with violent crime exists, such as violent movie theatrical releases at the daily and weekly levels, and school/work hours at the hourly level.

Furthermore, we only analyzed the time series data utilizing ARIMA, ARIMAX, GARCH, SARIMAX, VAR, and LSTM. Future studies could apply other models to the datasets in our study, from simpler models such as exponential smoothing, to more complex models such as TFT, and how that would impact prediction. There may also be models that could better examine the relationships between time series and are worth exploring.

Finally, the VAR model provided inconclusive results. It revealed an endogenous relationship between property and violent crime; however, the lead-lag relationship was unclear. The VAR model revealed an endogenous relationship between Atlanta and New York City for violent crime. These results indicate that there may be other factors that we did not or study that impact both cities. A future study could analyze crime in a city with a different climate pattern, such as Honolulu, to validate our results.