OREOLUWA ASABA

# Comprehensive Crime Analysis in Washington DC: Integrating Spatial Hotspots Mapping and Temporal Patterns

DATS 6501 – Data Science Capstone

Spring 2024

# INTRODUCTION

## Background

The Comprehensive Crime Analysis project seeks to provide law enforcement agencies in Washington DC with a holistic understanding of crime dynamics by integrating temporal patterns and spatial hotspots analysis.

The project focuses on analyzing crime data in Washington DC over a period of several years to identify patterns and trends in criminal activity. Advanced data analytics techniques, including time series and spatial analysis, are employed to provide insights into the geographical and temporal distribution of crimes.

This analysis can help law enforcement agencies make informed decisions and allocate resources effectively to improve public safety. By merging these approaches, the project aims to enhance proactive crime prevention strategies to help address the when and where of criminal activity in the district.

## Problem Statement

Despite the availability of large datasets on crime incidents, there remains a need for advanced analytical methods to better understand and predict crime patterns. These insights can enable law enforcement agencies to allocate resources efficiently, implement preventive measures, and ultimately enhance public safety. The project aims to address these challenges by utilizing forecasting models, including ARIMA, SARIMA, and Auto ARIMA, for time series analysis, as well as spatial analysis techniques for geographic crime pattern detection.

## Problem Elaboration

Current crime analysis approaches are often limited to basic statistical summaries and visualizations. While these methods provide useful insights, they may not fully capture the complex patterns in crime data over time and across different geographic areas. Machine learning models, such as ARIMA and SARIMA, can help identify and forecast trends in crime data, while spatial analysis can reveal crime hotspots and clusters. This comprehensive approach can lead to more effective law enforcement strategies.

## Motivation

The motivation for this project is to leverage advanced data analytics to improve the understanding of crime patterns in Washington DC. By providing law enforcement agencies with actionable insights, the project aims to enhance their ability to prevent and respond to criminal activity. Additionally, the findings from this project can contribute to ongoing research in the field of crime analysis and may have implications for policy development and resource allocation.

**Project Scope**

The project involves the analysis of crime data across 8 years, with a focus on both time series and mapping analysis. Time series models, including ARIMA, SARIMA, and Auto ARIMA, are used to identify trends and forecast crime counts over time. Mapping, through the use of the Python library, Folium, is employed to detect crime hotspots and patterns across different districts. The project aims to provide a comprehensive view of crime patterns in Washington DC and offer recommendations for law enforcement strategies.

## LITERATURE REVIEW

- Spatial analysis encompasses a diverse set of techniques for analyzing geographic data to understand spatial relationships and patterns.

- Domain-wide applications including urban planning, epidemiology, ecology, and criminology, among others.

- Time series analysis involves the study of data collected over time to identify patterns, trends, and relationships.

- It has wide-ranging applications across various fields, including finance, economics, climate science, epidemiology, and criminology.

- Traditional time series analysis techniques include autoregression (AR), moving average (MA), and exponential smoothing methods.

- The Box-Jenkins methodology, which introduced the Autoregressive Integrated Moving Average (ARIMA) model, revolutionized time series analysis by providing a systematic framework for model identification, estimation, and diagnostics.

- The Seasonal Autoregressive Integrated Moving Average (SARIMA) model extends ARIMA to account for seasonal patterns in the data, making it particularly useful for analyzing time series with recurring seasonal variations.

- In recent years, machine learning approaches, such as neural networks and gradient boosting machines, have gained popularity for time series forecasting tasks due to their ability to capture complex patterns and nonlinear relationships.

## METHODOLOGY

### Dataset Collection/Description

The dataset was obtained from the Metropolitan Police of Washington D.C. (https://mpdc.dc.gov/). Crimes in the dataset range from 2016-2014.

The dataset contains 29 features/variables, with over 510,000 records.

Some of these variables are the date and time of the report, the type of crime, the neighborhood and police district, the longitude and latitude etc. Below are the descriptions of the most relevant variables used in the analysis:

- NEIGHBORHOOD_CLUSTER: Represents the neighborhood cluster where the crime occurred.
- offense-group: Classifies the crime into a group (e.g., violent, property).
- LONGITUDE: The longitude coordinate of the location where the crime occurred.
- offense-text: The specific text description of the offense committed.
- DISTRICT: Represents the police district where the crime occurred.
- SHIFT: Indicates the shift during which the crime occurred (e.g., midnight, evening).
- YEAR: The year the crime occurred.
- offensekey: A combination of the offense group and a specific offense type, separated by a '|'.
- OFFENSE: The type of offense committed.
- REPORT_DAT: The date and time when the crime incident was reported.
- location: The latitude and longitude of the crime location in a single string.
- LATITUDE: The latitude coordinate of the location where the crime occurred.

## Data Preprocessing

The bulk of the preprocessing and engineering was focused on and done for the time series section of the project. From the Report_Date column, I was able to derive the following variable information:

- The year
- The time (hh:mm:ss format; datetime data type)
- The month of the year (number 1 through 12)
- The day of the week

The longitude and latitude values also needed some standardization. There is a need for a common coordinate system to ensure that analysis can be carried out on the location data provided in the longitude and latitude columns. I converted all the values to the World Geodetic System (WGS84) to ensure this uniformity across the board, and then checked for null and extreme (outlier) values.

## Data Visualization/Modeling

## EDA - Results & Analysis

The exploratory data analysis section of the report provides a comprehensive overview of the crime data from 2016 to 2024. The highlighted visualizations include:

1. Crime Trend Over Time: The line graph showcasing the entire data set across the period of 2016 to 2024 reveals the overall trend in crime counts over the years. This visual provides insights into the fluctuations in crime occurrences, allowing for the identification of periods with higher or lower crime rates.

2. Crime Counts by Hour, Day, and Month: Bar graphs display the distribution of crime counts by different time intervals.
    - *Crime counts by hour of the day*: This visualization reveals peak crime activity during certain hours, providing insights into potential times when law enforcement might need to allocate additional resources. It is interesting to note that crimes are highest over time during the early afternoon and late evenings. The former speculation is backed up by the idea that these are time periods where there are not a lot of people reporting crimes, as most people are either at work or in school. So crime numbers shoot up and police are unable to curb these. The later afternoon

numbers are potentially enforced because this is around the time a lot of students get out of classes. There is an outbreak in the District of juvenile crimes, so this does not come as a surprise.

- *Crime counts by day of the week*: The bar graph shows crime distribution across days, which can guide law enforcement planning for high-risk periods. The weekend days (Saturday and Sunday) are intuitively the highest here, as people are potentially freer on these days to commit acts of crime. Tuesday is also interestingly higher than the other days of the week.
- *Crime counts by month of the year*: This chart identifies seasonal trends in crime occurrences, offering opportunities for targeted interventions. The summer months are characterized by having the highest crime numbers. Also, somewhat intuitive since a lot of people are 'outside'.

3. Crime Counts by Police District: The bar graph displays aggregated crime counts grouped by police district. District 3 has the highest crime count, while District 7 has the lowest, providing a clear indication of where crime prevention efforts may need to be focused. District 3 contains some of DC's most well-known neighborhoods such as Adams Morgan, Columbia Heights and DuPont Circle.

4. Crime Counts by Crime Type: A grid of line charts illustrates crime counts grouped by offense type through the years. This visual provides a nuanced understanding of how different types of crimes have evolved over time, allowing for more specific and targeted interventions. The more common crimes such as vehicle theft, general thefts, armed robbery, and homicides witnessed significant rises in 2023 while crimes such as burglary and assault with dangerous weapon were on the decline.

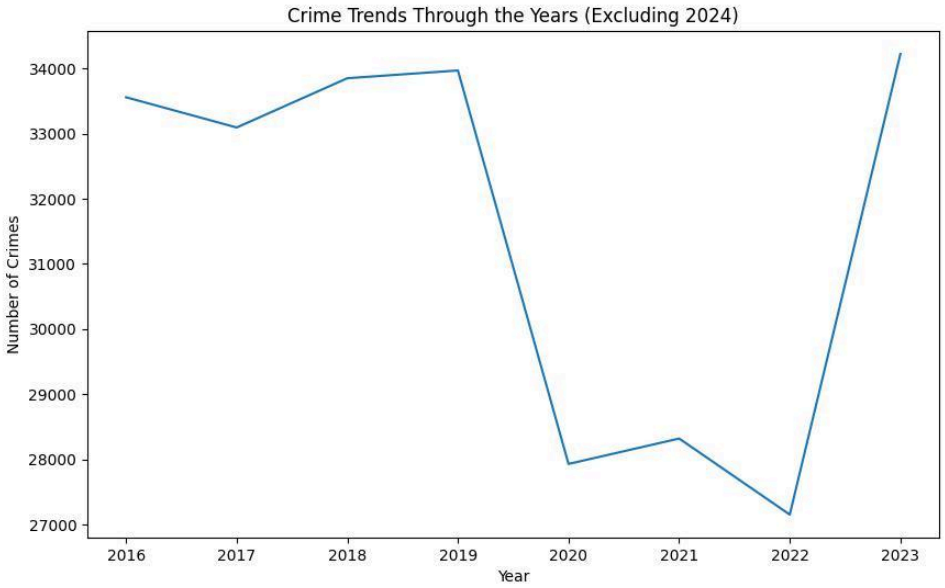
**Potential Reasons -**

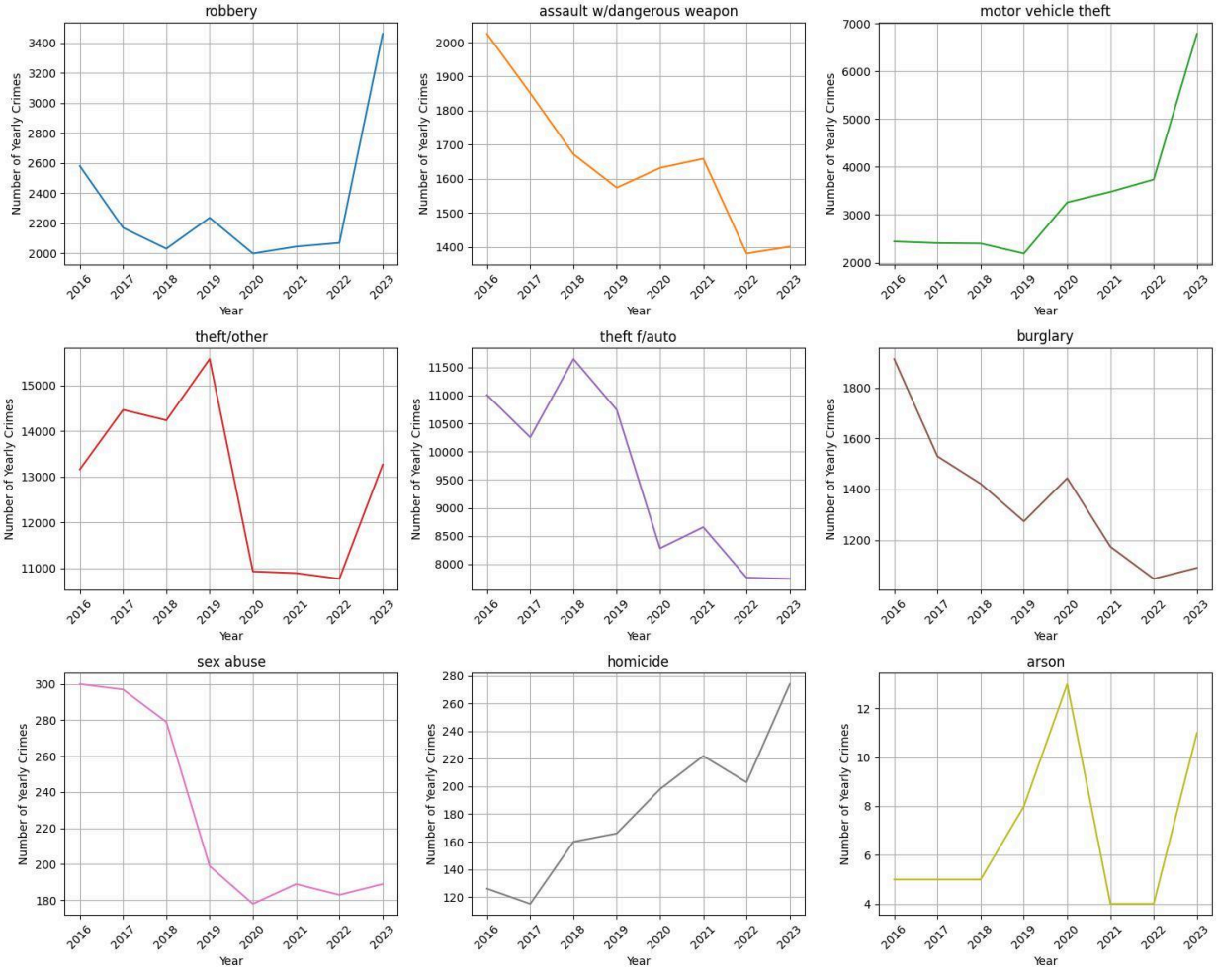
The crime rate in the District of Columbia is on the rise due to several interrelated factors, but I will touch on a few here:

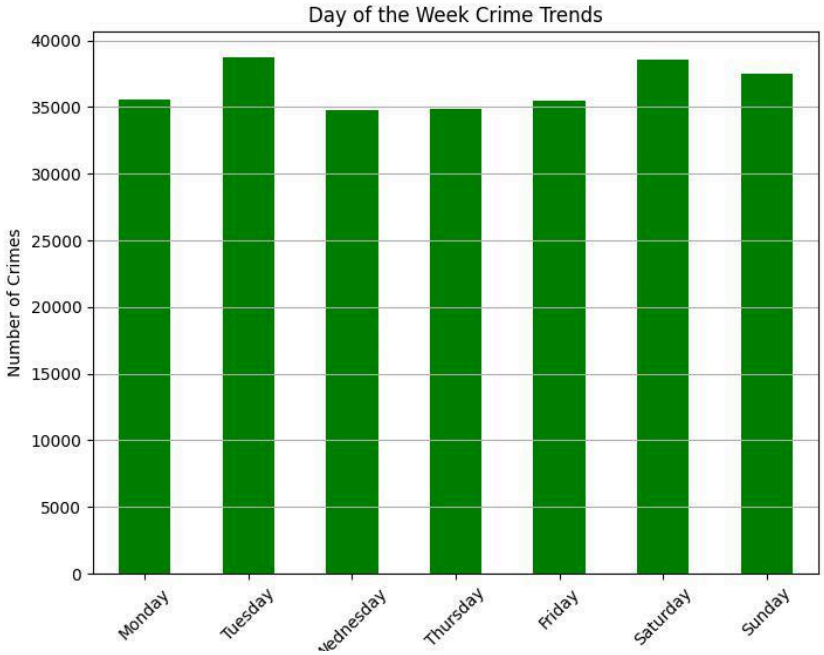
- Shrinking police force and higher workload: The department has faced staffing shortages and leadership changes, affecting response times and community engagement. Officers are handling more calls and slower response times, impacting their ability to build community relationships and trust. Officers are responding to 23% more calls on average in 2023 than previously, since a huge number of these officers are tasked with safeguarding monuments and buildings within the district.

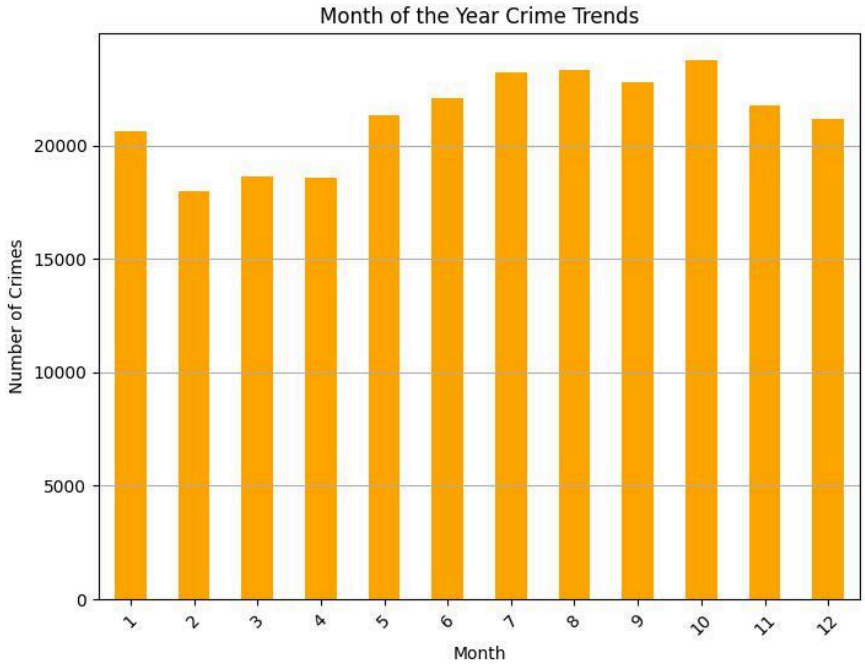
- Challenges with the crime lab: The closure of the crime lab in 2021 led to delays in forensic testing, which has slowed the judicial process and resulted in some case dismissals. The crime lab is tasked with handling forensics, and its closure led to a substantial increase in the backlog of cases.

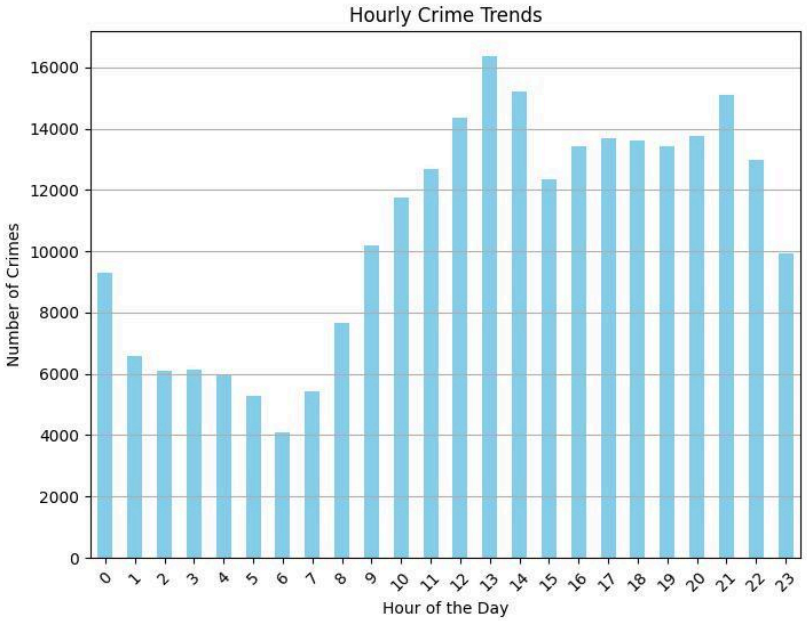
- Pandemic-related challenges: A large proportion of remote workers and closed schools and courts have disrupted community structures, potentially emboldening criminals due to fewer witnesses. High levels of absenteeism and truancy among youth may contribute to an increase in gang activity and criminal behavior. School attendance numbers in the District are still very low since the pandemic rules were eased off, compared to the national average (43% for DC, 27% national average. The district is also yet to fully recover from the pandemic's effects on adult workers. DC has the highest rate of remote workers and this has potential to reduce the number of crime witnesses.


Crime Trends Through the Years (Excluding 2024)
Number of Crimes
Year
34000
33000
32000
31000
30000
29000
28000
27000
2016
2017
2018
2019
2020
2021
2022
2023


robbery
assault w/dangerous weapon
motor vehicle theft
theft/other
theft f/auto
burglary
sex abuse
homicide
arson
Number of Yearly Crimes
Year

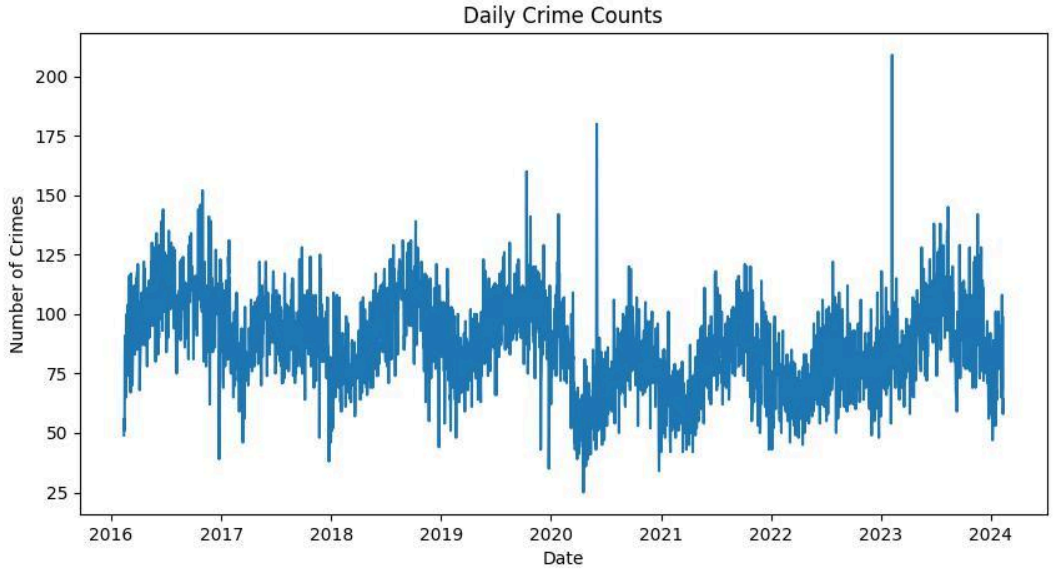

Day of the Week Crime Trends
Number of Crimes
40000
35000
30000
25000
20000
15000
10000
5000
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday


Month of the Year Crime Trends
Number of Crimes
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
9
10
11
12
Month




## Time Series - Results & Analysis

Initial Time Series: The initial time series visual represents the daily crime count in Washington DC over a given period - 2016-2024.




This chart serves as a foundation for understanding the data and identifying potential patterns, trends, and seasonal variations in crime occurrences.

**Splitting the dataset.**

Training data - 60% - (2016-02-13 - 2020-11-29)

Validation data - 20% - (2020-11-30 - 2022-07-06)

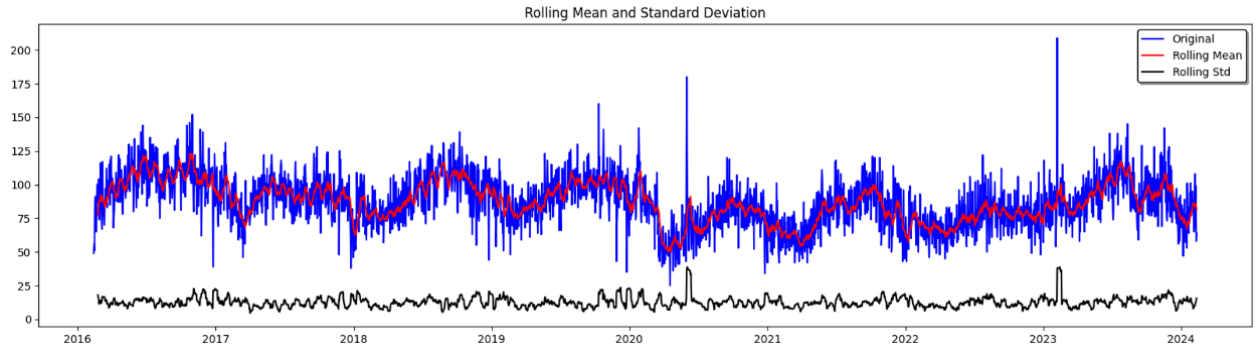
Test data - 20% - (2022-07-07 - 2024-02-11

**Stationarity**:

Stationarity refers to a time series whose statistical properties, such as mean, variance, and autocorrelation, remain constant over time. A stationary series is more predictable and easier to model, making it a condition for time series analysis. Visuals of the time series with and without differencing can illustrate whether the data exhibit stationarity. There are a few other ways to check for stationarity besides just visualizing however.

**Checking for Stationarity.**

1. Visualization.

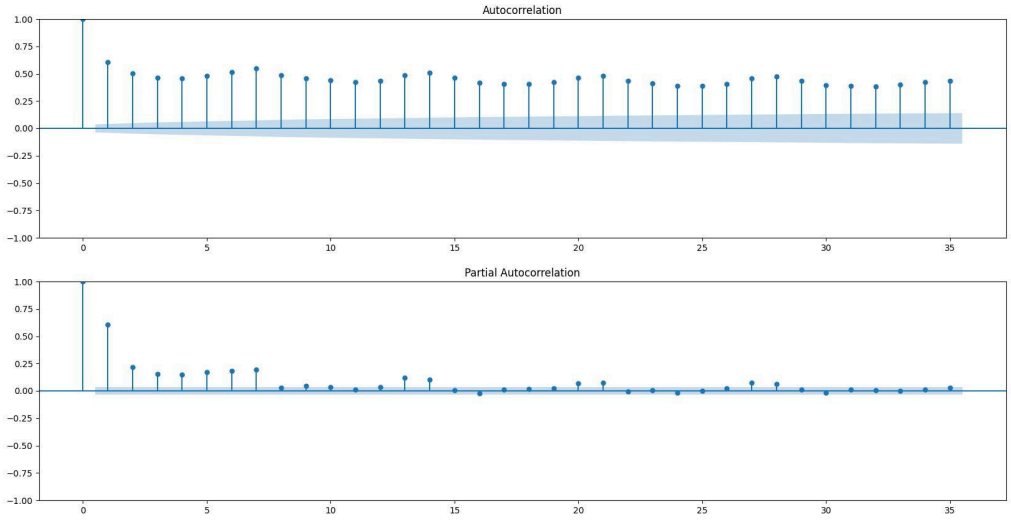



The time series here clearly exhibits stationarity. I have plotted alongside the original observations, its rolling average and rolling standard deviation. We see that those do not change too much over time.

2. ACF/PACF Plots.

Autocorrelation function (ACF) and partial autocorrelation function (PACF) plots help assess the relationship between a data point and its lags.

- ACF: The ACF plot shows the correlation between the current value and its lags. High and significant ACF values at multiple lags indicate patterns in the data, while a gradual decline suggests potential seasonal patterns.
- PACF: The PACF plot shows the partial correlation between the current value and its lags, eliminating the effects of intervening lags. Spikes at specific lags in the PACF plot can guide the choice of AR terms in the models.
  Both graphs are shown below:




- ACF: The ACF plot shows the correlation between the current value and its lags. High and significant ACF values at multiple lags indicate patterns in the data, while a gradual decline suggests potential seasonal patterns. Essentially, we are asking how similar the number of crimes this month is, compared to last month (lag of 30). In the case of this data set, the ACF value was high all the way through, indicating strong autocorrelation. There seems to be a strong relationship or dependency of the current day's crime count on the previous day's count. There might be some seasonality.

- PACF: The PACF plot shows the partial correlation between the current value and its lags, eliminating the effects of intervening lags. Spikes at specific lags in the PACF plot can guide the choice of AR terms in the models. Lag drops around lag 7; indicating most likely seasonality in the data – strong influence from previous observations every 7 days; indicates seasonality. It is very probable that weekly data does not change too drastically.

**ARIMA Models**

1. ARIMA

These are a type of statistical model used to analyze and forecast time series. It stands for AutoRegressive Integrated Moving Average. The ARIMA (Autoregressive Integrated Moving Average) model combines three components: AR (Autoregression), I (Integration, or differencing), and MA (Moving Average). ARIMA models capture patterns in the data and forecast future values.

- Autoregression (AR) – using past values of time series to predict future
- Moving Average (MA) – uses average of recent observations to clean/smooth out fluctuations in the trend.
- Integrated (I) – uses differences in past consecutive observations to ensure time series properties do not change over time.

Choosing the order of ARIMA.

The ARIMA model is typically characterized by 3 terms: p, d and q, where:

- p is the order of the AR term. The number of lag observations included in the model, also called the lag order.
- q is the size of the moving average window, also called the order of moving average.
- d is the number of differencing required to make the time series stationary.

In order to get the right order, I used the Akaike's Information Criterion (AIC) on a set of models and investigated the models with the lowest root mean square deviation (RMSE). For ARIMA, I used order (1,1,0). The summary is below:

```
Validation RMSE: 15.95
Test RMSE: 18.23
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 2336
Model:                 ARIMA(1, 1, 0)   Log Likelihood               -9678.573
Date:                Wed, 01 May 2024   AIC                          19361.146
Time:                        10:49:21   BIC                          19372.658
Sample:                             0   HQIC                         19365.340
                                - 2336
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.3554      0.018    -20.266      0.000      -0.390      -0.321
sigma2       233.3200      5.667     41.173      0.000     222.213     244.427
===================================================================================
Ljung-Box (L1) (Q):                  19.25   Jarque-Bera (JB):               133.80
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.04   Skew:                             0.15
```

2. SARIMA

The Seasonal ARIMA model extends the ARIMA model by incorporating seasonal terms. This model accounts for seasonality in the data, making it more suitable for data with repeating patterns over specific periods.

It incorporates extra parameters to account for seasonality. Just like in the ARIMA, I went with the model with the lowest RMSE, and came up with the model SARIMA (2,0,0)x(2,0,1,7).

- 2: The seasonal autoregressive (SAR) order, indicating that the model uses the last two seasonal periods' data (e.g last two weeks if the seasonal period is weekly) to predict the current value.
- 0: The seasonal differencing (D) order, indicating that no seasonal differencing is applied to the data.
- [1]: The seasonal moving average (SMA) order, indicating that one seasonal moving average term is used in the model.
- 7: The seasonal period (m), indicating a seasonal pattern every 7 time periods (e.g, weekly seasonality).

Below is the summary:

```
Final SARIMA model RMSE on test set: 25.388
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                               Count   No. Observations:                 2336
Model:             SARIMAX(2, 0, 0)x(2, 0, [1], 7)   Log Likelihood               -9414.454
Date:                             Wed, 01 May 2024   AIC                          18840.907
Time:                                     10:59:23   BIC                          18875.445
Sample:                                 02-13-2016   HQIC                         18853.489
                                      - 07-06-2022
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.3733      0.019     19.606      0.000       0.336       0.411
ar.L2          0.1434      0.019      7.380      0.000       0.105       0.182
ar.S.L7        1.0829      0.024     44.763      0.000       1.035       1.130
ar.S.L14      -0.0836      0.024     -3.463      0.001      -0.131      -0.036
ma.S.L7       -0.8901      0.012    -76.503      0.000      -0.913      -0.867
sigma2       183.3395      4.269     42.944      0.000     174.972     191.707
==============================================================================
```

3. Auto ARIMA

Auto ARIMA is a technique that automatically selects the best ARIMA model for the data based on predefined criteria (such as AIC or BIC). This method saves time and effort by automating the model selection process, identifying the best combination of AR, I, and MA terms. It is a part of the pmdarima library in Python and I used the lowest AIC value to help determine the Auto ARIMA order here. The best order for this dataset is (2,1,2).

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 1752
Model:               SARIMAX(2, 1, 2)   Log Likelihood               -7064.587
Date:                Wed, 01 May 2024   AIC                          14139.175
Time:                        10:59:32   BIC                          14166.515
Sample:                  02-13-2016   HQIC                         14149.281
                       - 11-29-2020
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.6658      0.047    -14.206      0.000      -0.758      -0.574
ar.L2          0.2388      0.032      7.402      0.000       0.176       0.302
ma.L1          0.0277      0.041      0.678      0.498      -0.052       0.108
ma.L2         -0.8730      0.038    -22.847      0.000      -0.948      -0.798
sigma2       186.8857      4.507     41.464      0.000     178.052     195.720
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               478.67
Prob(Q):                              0.98   Prob(JB):                         0.00
```

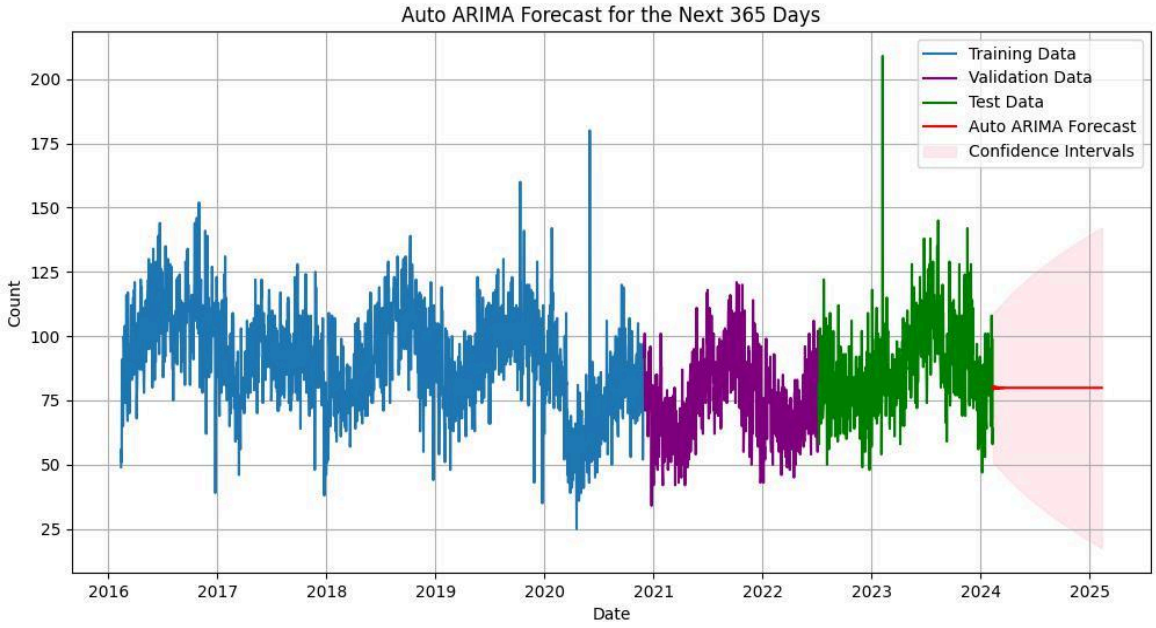
**Model Evaluation**

| | Model | MAE | MSE | AIC |
|---|---|---|---|---|
| 0 | ARIMA | 16.03 | 454.36 | 19361.15 |
| 1 | SARIMA | 19.96 | 644.55 | 18840.91 |
| 2 | Auto ARIMA | 14.09 | 361.01 | 14139.18 |

- Lowest Mean Absolute Error (MAE): Auto ARIMA achieved the lowest MAE among the three models, indicating that its forecasts were closer to the actual values and had smaller average error.
- Lowest Mean Squared Error (MSE): Auto ARIMA had the lowest MSE, suggesting that its forecasts were more accurate and had less variance compared to the other models.
- Lowest Akaike Information Criterion (AIC): Auto ARIMA's AIC was significantly lower than the ARIMA and SARIMA models, implying that it was a more efficient and simpler model.
- Consistent Performance: Overall, the combination of lower error metrics and AIC values demonstrates that Auto ARIMA consistently provided better forecasts and was the most reliable choice for this project.

In reality, when forecasts were made, I believe that this model can definitely be improved.

## Conclusion




Right off the bat, we can tell that this forecast is **not** going to prove to be overly accurate over the course of the next 365 days. It is almost impossible that it stays stagnant, and is quite likely that the model does not entirely capture the patterns in the dataset (albeit it does so better than the other models).

The good thing however is that it mirrors both what the Metropolitan Police of the District has reported so far to start the year off, and what they project as the year goes by. It is believed that the crime rate will begin to drop. Per NBC Washington, the first 3 months of 2024 witnessed significant drops already. Homicides were down 32%, Assault with dangerous weapons down 34%, and overall violent crime was down 17%. They are also beginning to put plans in place to clamp down on the steady increases of crime in the summer months with the goal of increasing the police force from about 3.3k to 4k this year.

### Data Limitations

Two important things that I noticed throughout this task regarding the data are as follows:

- As clean as the data is, I believe that it is not granular enough. For a district/city which has a high rate of juvenile crime, I believe the reported cases dataset

should incorporate the age/age group of perpetrators. This helps provide an even deeper angle to help analyze where crime is taking place and when. Putting this in place can help the city focus better on what to do and allocate its resources better. Groups committed to helping youth delinquency will also be provided with enough granular information to help their efforts improve even more.

- Missing exogenous information: I believe that forecasting with this data record can be further improved if we begin to collect certain extra information. For example, I believe noting public holidays as well as even the weather conditions can be very useful in making more precise predictions regarding how, where and when crime is carried out in the District. Furthermore, economic indicators for the 7 districts within DC could be made available to help dig even deeper into the weeds regarding how crime is perpetuated in the area.

- More time needed?

The data plainly might struggle to capture differences in pre and post pandemic periods. In fact, there is almost no way to split the dataset in a way that does not segment training and testing into these 2 extremely distinct time periods. As a result, the forecast might struggle. The world has changed, and even stronger models might be needed to attempt to capture these changes. The post-pandemic trends most likely tell a different story, and it is quite difficult to extrapolate these stories from pre-pandemic data.

## **Future Research**

1. Advanced Time Series Techniques.

The use of more robust machine learning should be incorporated into crime analysis. Investigate other advanced time series forecasting models such as deep learning models (e.g., Long Short-Term Memory (LSTM) networks, Prophet) or hybrid models that combine classical approaches with machine learning techniques.

2. Exogenous Data and Scenario Analysis

   The implementation of datasets that contain the exogenous variables alluded to above. Exploring the inclusion of these variables like the economic indicators in the different areas, weather data, and public holiday/event data could help us do better with forecasting crime analysis. Some scenario analysis could be carried out to assess the potential impact of different policy changes, interventions, or other external factors on future crime trends.

3. Improved Validation Techniques

Researching and implementing advanced time series cross-validation methods that better account for the temporal nature of the data and lead to more reliable model evaluations. Improving validation techniques for time series data involves using advanced methods that respect the temporal order of the data to provide more reliable model evaluations. Traditional cross-validation methods, such as k-fold, may not be suitable for time series data due to the interdependence of observations over time. Advanced techniques, such as rolling-origin validation and time series cross-validation (e.g., walk-forward validation), allow for model evaluation that respects the chronological sequence of the data.