Crime Data Analysis and Prediction

City of Philadelphia

Data Source: kaggle.com Dataset

DSC680 – Applied Data Science

Project No. 3 - Milestone 3 – Final Paper

Winter 2021

Manish Kalkar

Final Paper

1. Abstract

In the recent past, crime analyses are required to reveal the complexities in the crime dataset. This process will help the parties that involve in law enforcement in arresting offenders and directing the crime prevention strategies. The ability to predict the future crimes based on the location, pattern and time can serve as a valuable source of knowledge for them either from strategic or tactical perspectives. Nevertheless, to predict future crime accurately with a better performance, it is a challenging task because of the increasing numbers of crime in present days. Therefore, crime prediction method is important to identify the future crime and reduces the numbers of crime. Currently, some researchers have been conducted a study to predict crime based on particular inputs. Crime forecasting refers to the basic process of predicting crimes before they occur. Tools are needed to predict a crime before it occurs. Currently, there are tools used by police to assist in specific tasks such as listening in on a suspect's phone call or using a bodycam to record some unusual illegal activity.

2. Background

Philadelphia consistently ranks above the national average in terms of crime, especially violent offenses. It has the highest violent crime rate of the ten American cities with a population greater than 1 million residents as well as the highest poverty rate among these cities. It has been included in real estate analytics company Neighborhood Scout’s "Top 100 Most Dangerous Cities in America" list every year since it has been compiled. Much of the crime is concentrated in the

North, West, and Southwest sections of the city. The legal entities responsible for maintaining law and order are The Philadelphia Police Department (PPD) is the police department, The Court of Common Pleas of Philadelphia County (1st Judicial Circuit) is the state trial court, The Philadelphia District Attorney is the district attorney, The Defender Association of Philadelphia is the government-funded independent public defender office.

3. Business Problem

Crime is a global concern that impacts individuals and society on a daily basis and negatively affects society. Using the historical data from 2006-2016 within open source Philadelphia Crime dataset to make predictions for the number of monthly violent crimes that will occur in future months. We will use exploratory data analysis techniques to answer questions and make assumptions about the data. Then we will use time series model such as Autoregressive Integrated Moving Average (ARIMA) to make predictions on future data.

4. The Dataset

Source: kaggle.com

Dataset: https://www.kaggle.com/jagannathrk/arima-philly-violent-crime/data

This dataset contains crime incidents from the Philadelphia Police Department. Part I crimes include violent offenses such as aggravated assault, rape, arson, among others. Part II crimes include simple assault, prostitution, gambling, fraud, and other non-violent offenses. The dataset previously had separate endpoints for various years and types of incidents. These have since been consolidated into a single dataset.

Attributes:

1.	Dc_Dist	DC District Code
2.	Psa	Police Service Area Boundary
3.	Dispatch_Date_Time	Dispatch Date and Time
4.	Dispatch_Date	Dispatch Date
5.	Dispatch_Time	Dispatch Time
6.	Hour	Hour
7.	Dc_Key	Police District Key
8.	Location_Block	City Block
9.	UCR_General	Universal Crime Reporting Code
10.	Text_General_Code	Crime Short Description
11.	Police_Districts	Police District Code
12.	Month	Month of the year
13.	Lon	Longitude - Crime Location
14.	Lat	Latitude - Crime Location

5. Exploratory Data Analysis

Exploratory Data Analysis involved Data Preparation, Feature Selection and Feature Engineering.

Feature Selection

List of relevant features available for the analysis:

- Dispatch_Date_Time
- Police_Districts
- Text_General_Code

Feature Engineering

- Remove Null values
- Change data type police district to int
- Sort / Order by based on date
- Remove the year 2017 since it's not a complete year
- New Features are derived from the original feature Dispatch_Date_Time
    - Year_Nr – Year
    - Month_Nr – Month
    - Day_Nr – Day

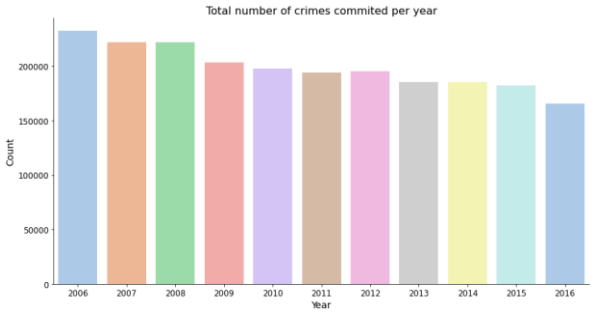
Analysis Results

- Almost every year there is a decrease in the number of crimes happening. In other words, Total number of crimes is decreasing every year.
- Number of crimes happening reaches a peak at 16:00h and is at its lowest point around 06:00h.
- The category Assaults and Theft are crimes that are committed the most within Philadelphia.
- Most crimes happen in District 11.
- Considerably more crimes are made in summer than in the winter.
- Most police stations every year the number of crimes is reduced but there are areas where the number of crimes increased e.g.: 15.0).
- From 3:00 - 7:00 there is less crimes. In 2015, significantly reduced the number of crimes between 23:00 - 2:00 hours than in 2006.

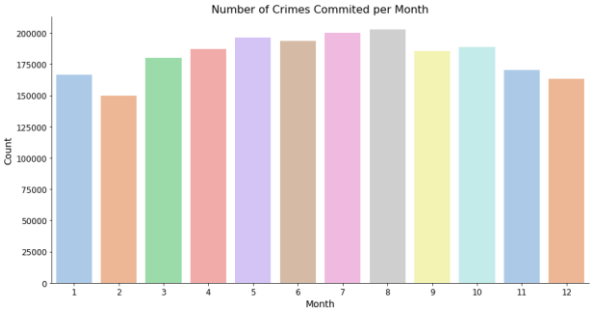
6. Graphical Analysis

Graphical Analysis was performed by plotting the attributes as listed below:

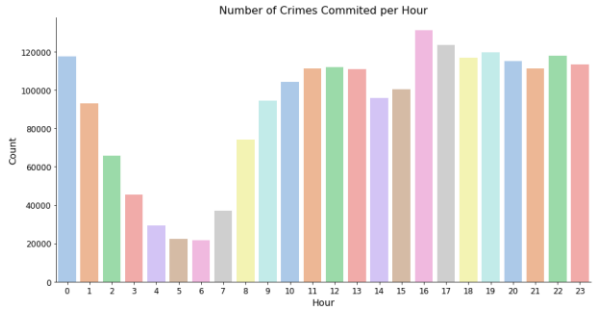
1. Total number of crimes committed per Year



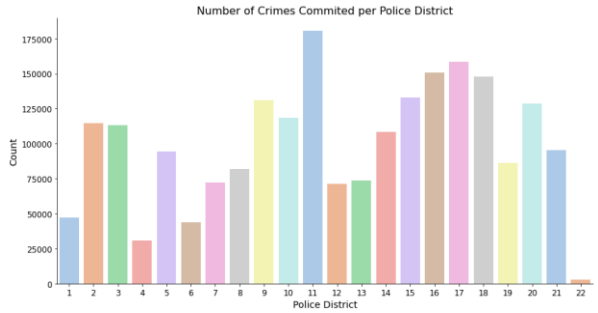
2. Number of crimes committed per Month



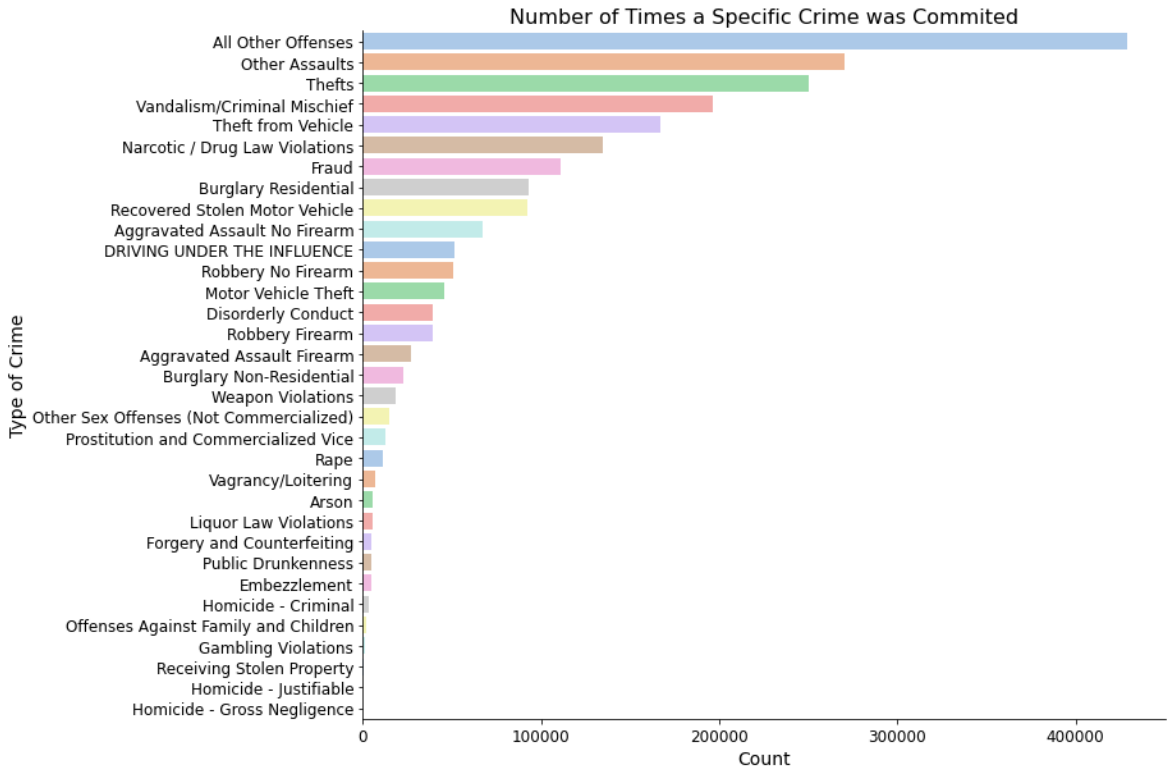
3. Number of crimes committed per Hour



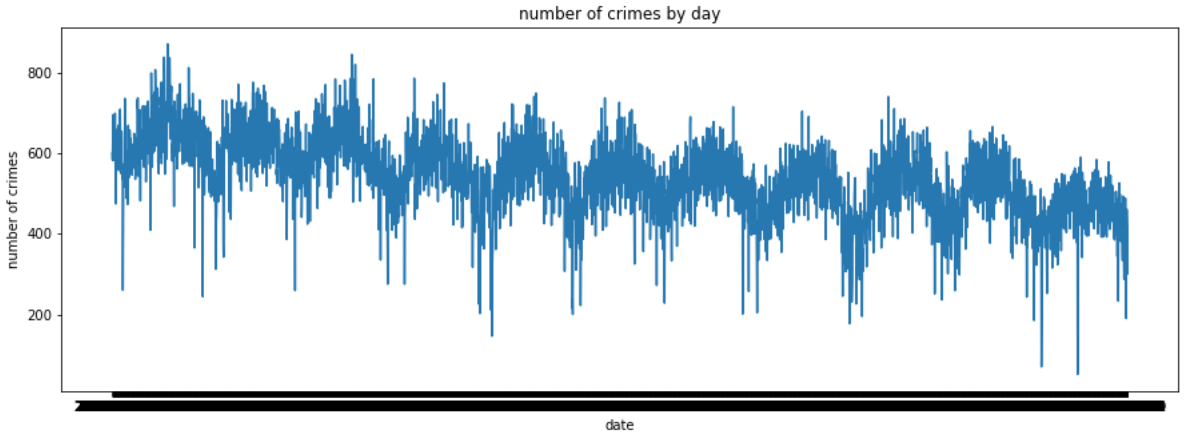
4. Number of Crimes Committed per Police District



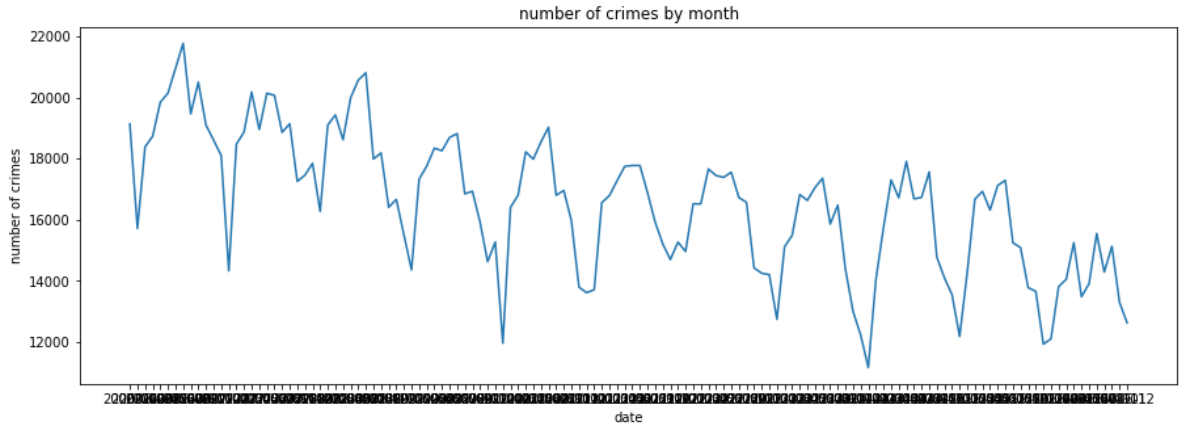
5. Number of Times a Specific Crime was Committed



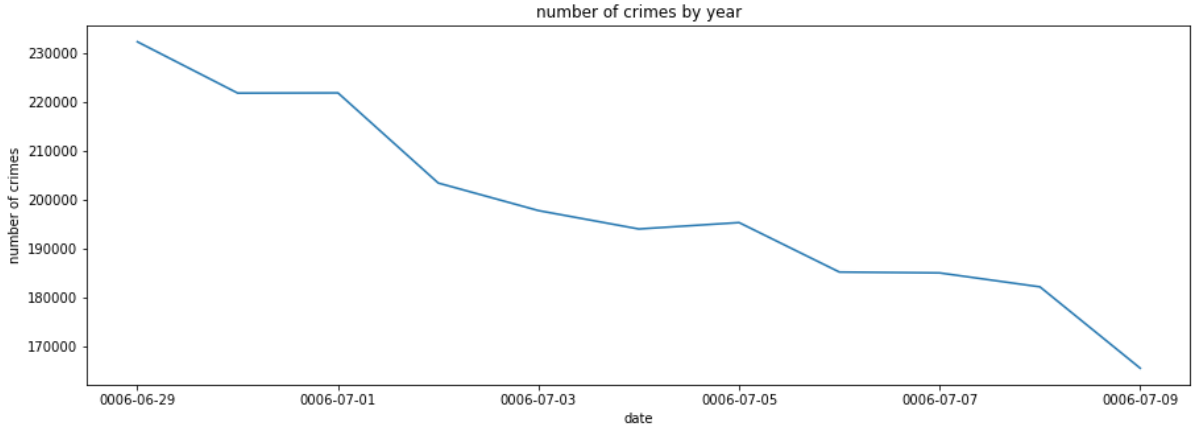
6. Number of crimes by Day



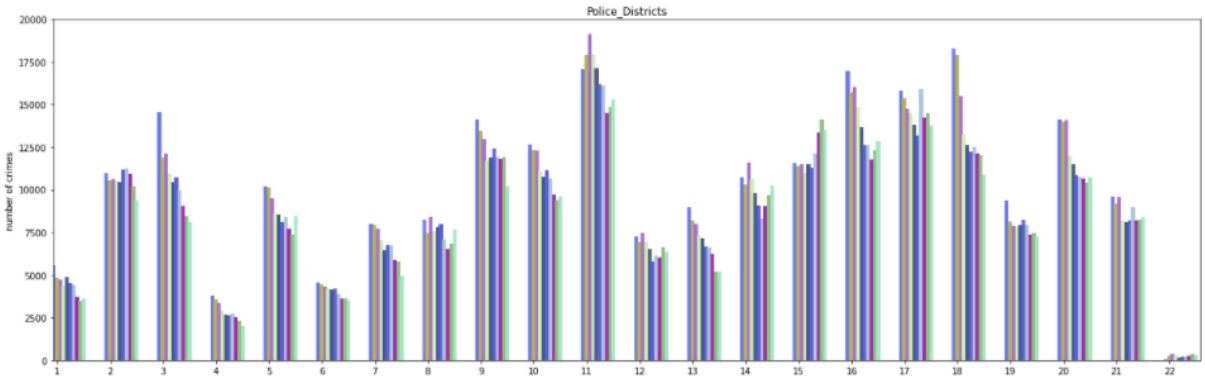
7. Number of crimes by Month



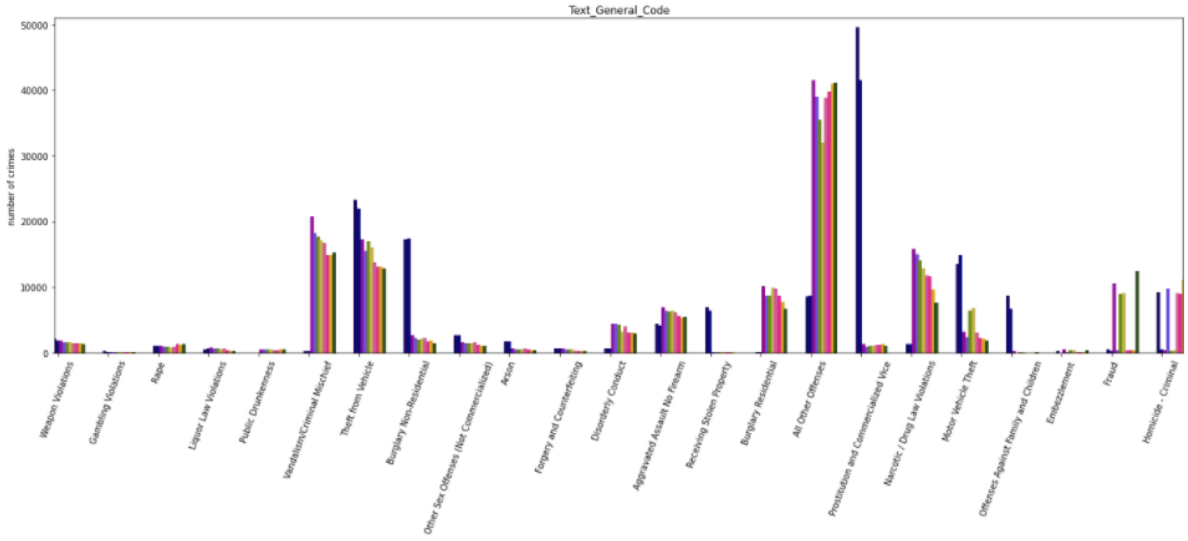
8. Number of crimes by Year



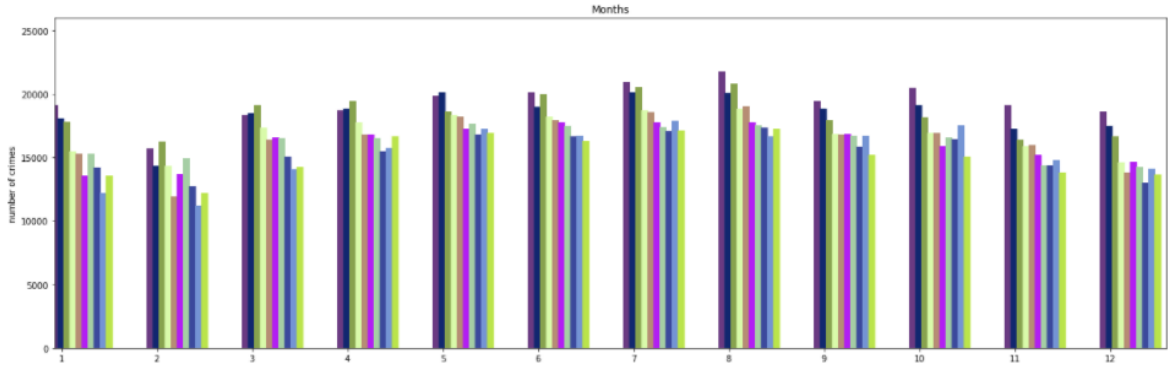
9. Total number of crimes per year in each of the Police_Districts



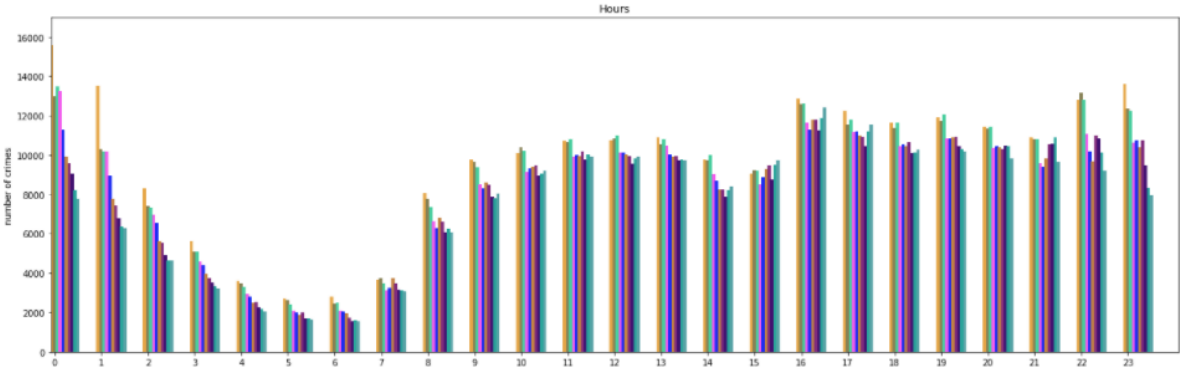
10. Total number of crimes per year for each Text_General_Code



11. Total number of crimes per year in each Month



12. Total number of crimes per year in each Hour



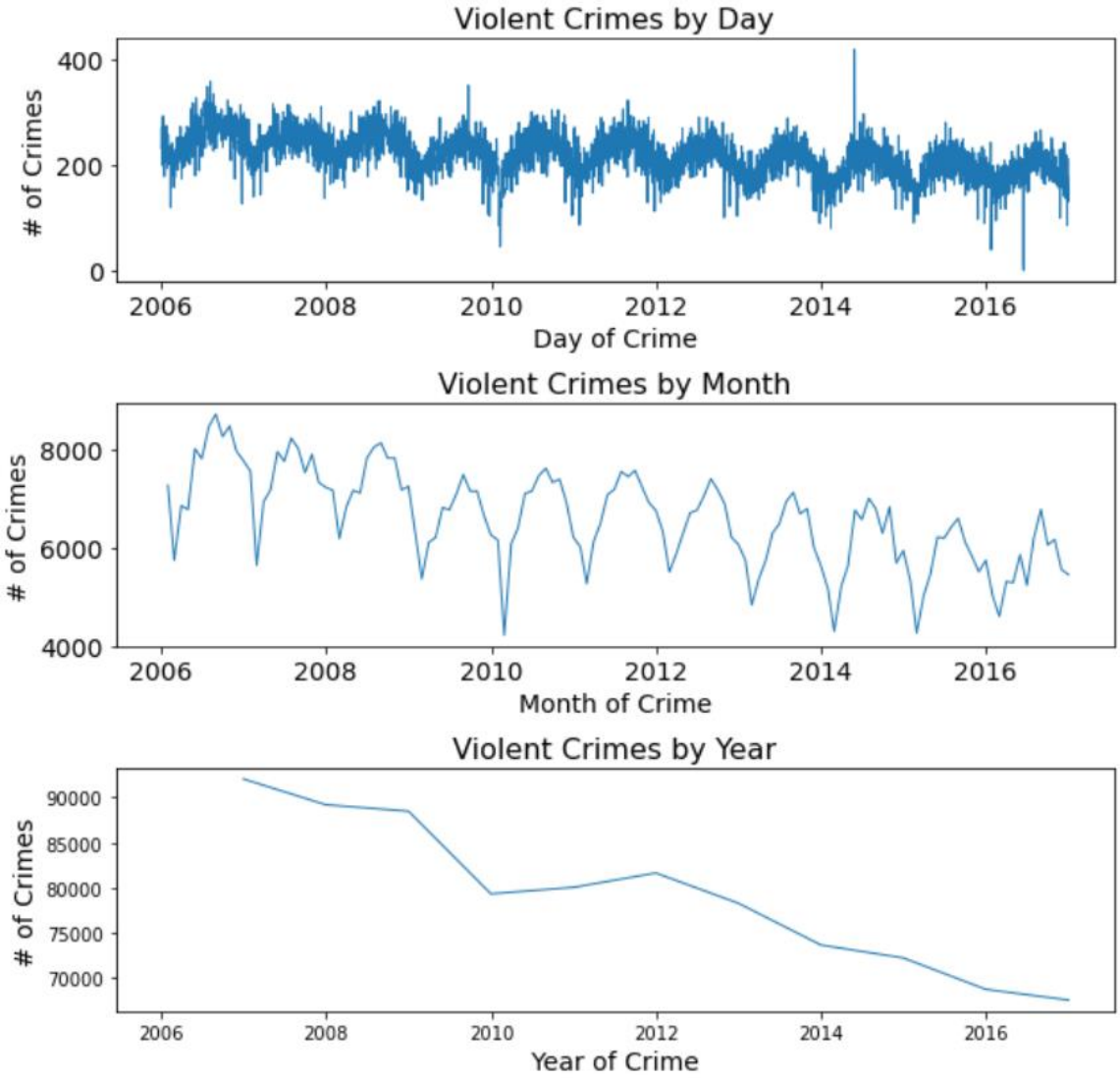
6. Modeling

Model Selection

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. A statistical model is autoregressive if it predicts future values based on past values.

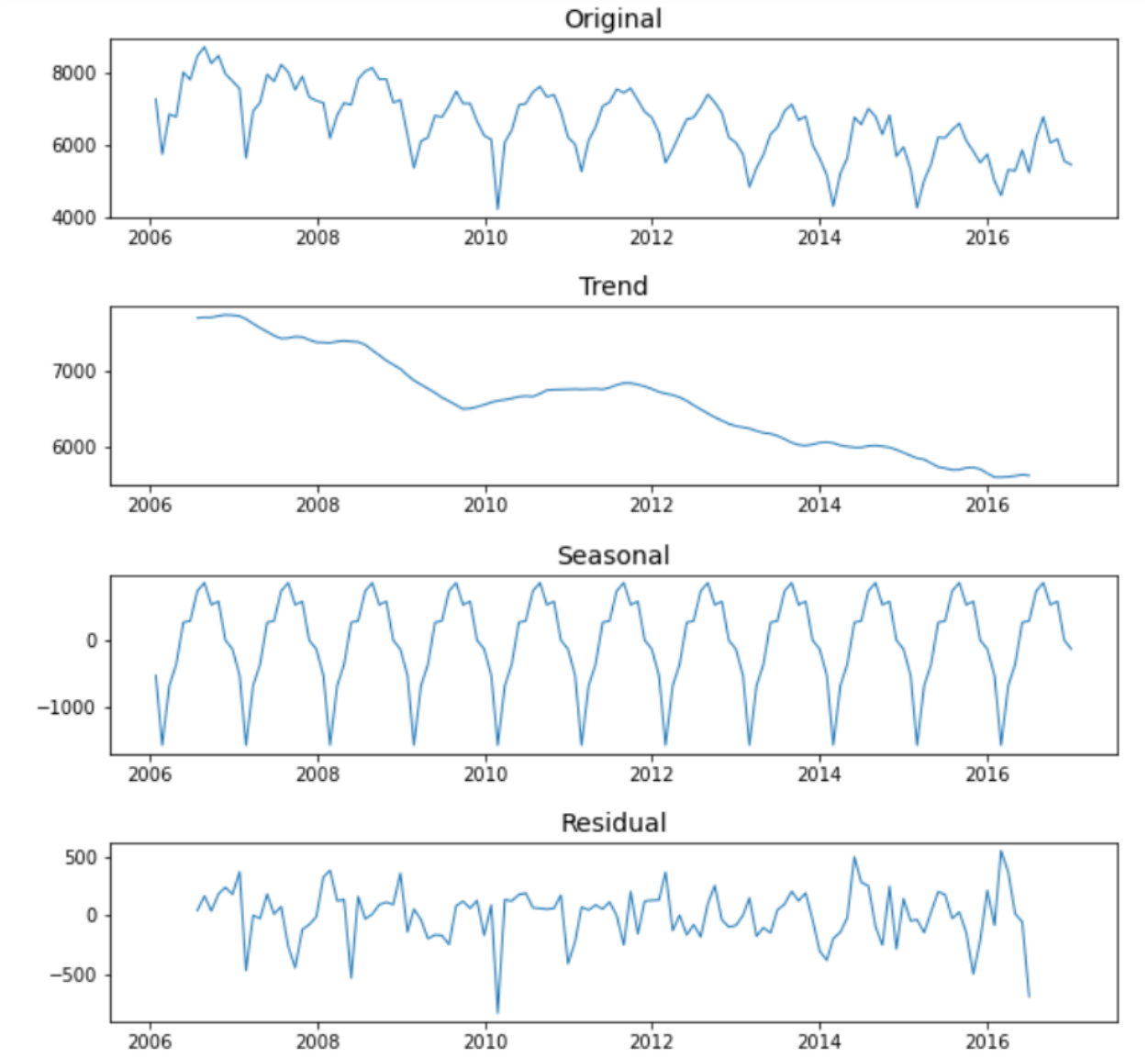
7. Model Evaluation

Evaluate Violent Crimes by Day, Month, Year



Visualize the Seasonal Decomposition

- Trends - What is the overall trend in the data?
- Seasonality - How does crimes fluctuate between seasons?
- Residuals - When removing trends and seasonality what does the data look like?



Success Criteria

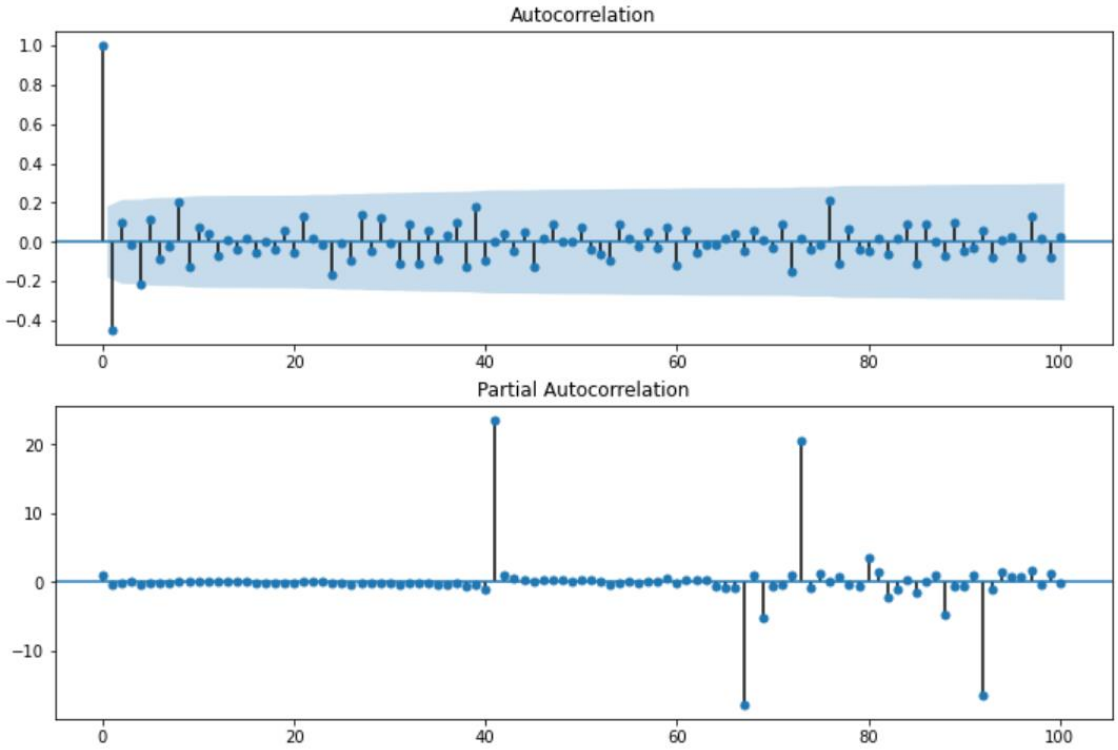
We will evaluate our model with the metrics Mean Absolute Percentage Error (MAPE). Aim for a model with a MAPE < 10%

- Outcome Variable: number of violent crimes
- Predictors: Time
- Relevant Timeframe: January 2006 – Present

Find optimal parameters for ARIMA

Determine optimal number or AR terms (p), MA terms (q)

Plot Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)



- p – The lag value where the PACF chart crosses the upper confidence interval for the first time. p=1.
- q – The lag value where the ACF chart crosses the upper confidence interval for the first time. q=1.

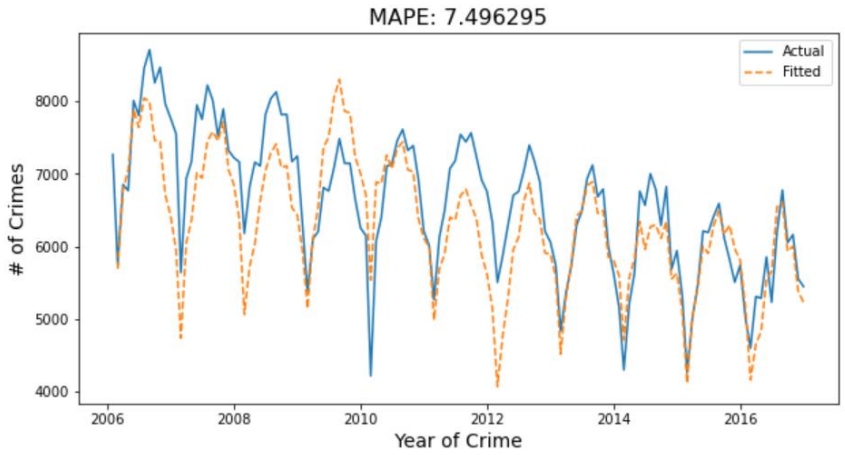
8. Model – Outcome / Results

Use ACF and PACF results to fit our ARIMA model

Fit the ARIMA model

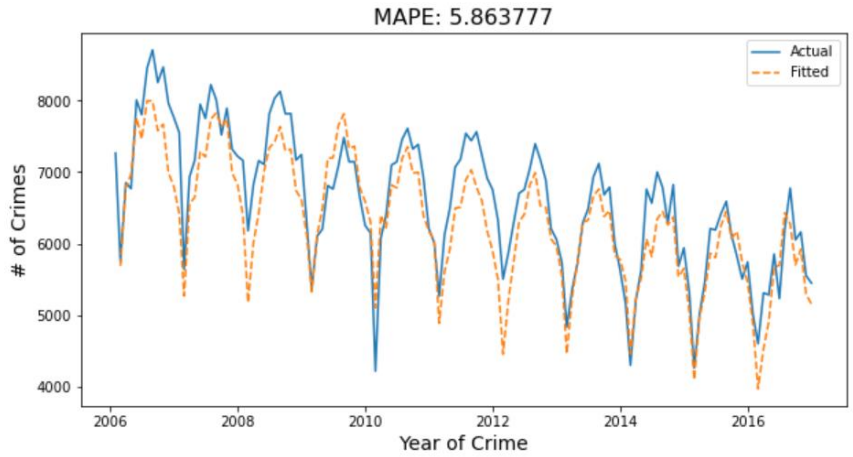
Determine the optimal model by measuring Mean Absolute Percentage Error (MAPE)

MA (1) Model



MAPE = 7.496295

AR (1) Model



MAPE = 5.863777

Based on the Success Criteria, AR (1) Model fit with MAPE = 5.863777 looks great.

9. Assumptions

- There are significant trends in the data
- Seasonality largely effects the data
- The data is not stationary
- Potential outliers will need to be investigated/removed

10. Limitations

- The predictive accuracy seemed to be dependent on demographic, criminological and psychopathological characteristics of the offenders.
- Certain types of crimes including Domestic Violence is not easily amenable to predictive models since they are seldom concentrated in specific locations and cannot be attributed to specific profiles of the victims.
- Predictive policing is often costly owing to data storage, lack of transparency in relation to the underlying algorithms and occasionally can lead to violation of basic rights and civil liberties.
- Identifying biases into the datasets is complex requiring deep knowledge in statistics, mathematics, and programming

11. Conclusion

The City of Philadelphia saw around a 20% decrease in the number of crime incidents in the City of Philadelphia from 2006 through 2015. The number of incidents fell in most police districts but held steady in the 15th district, the district with the most incidents. The fewer number of

Vandalism, Motor Vehicle Thefts, Burglaries, and Drug Violations helped drive down the crime rate. A decline in the catch-all category 'All Other Offenses' helps explain the larger declines in some police districts. Philadelphia has made positive progress in reducing crime over the 2006 through 2015 period.

12. Challenges

The main challenge is to create a solution that conforms to the ethical standards but also should be able to implement. On the other hand, one of the most challenging issues of police departments is to have accurate crime forecasts to dynamically deploy patrols and other resources to improve deterring of crime occurrence and police response times. The effectiveness of the outcome of the predictive analytics is remains to be seen.

13. Implementation Plan

Below six steps were considered while implementing the effort end to end:

- Step 1: Define Problem Statement
- Step 2: Data Collection
- Step 3: Data Cleaning
- Step 4: Data Analysis and Exploration
- Step 5: Data Modelling
- Step 6: Optimization and Deployment

14. Ethical Considerations

The use of software and statistics in order to track crime has been around for over 20 years. This software or computer algorithm would be used for tracking a police department's performance and attacking spikes in crime rates at specific locations where crime was likely to occur. These newer methods of predictive policing were created with the dual intent of reducing crime and making police officers' jobs easier. Despite the lack of malicious intent, this use of training data to find patterns has garnered results that many deem racially biased. While a computer algorithm may simply be the sum of what others put into it, it raises the question: is it ethical to use predictive policing analytics when the data may introduce racial biases? When we look at the consequences, it becomes clear that tools used in predictive policing need to follow stricter guidelines, and security measures need to be implemented against biased data if we wish to use it as an ethical means of crime prevention.

15. References

- **Dataset:** https://www.kaggle.com/jagannathrk/arima-philly-violent-crime/data
- https://www.frontiersin.org/articles/10.3389/fpsyg.2021.587943/full
- https://vce.usc.edu/volume-2-issue-2/ethics-of-predictive-policing/
- https://www.kaggle.com/jagannathrk/arima-philly-violent-crime
- https://www.kaggle.com/suijket/philadelphia-crime-data-visualization
- https://www.kaggle.com/sikayena/visualization-and-analysis
- https://ieeexplore.ieee.org/abstract/document/8075335

- https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z
- https://en.wikipedia.org/wiki/Crime_in_Philadelphia
- https://www.jstor.org/stable/resrep20640.6?seq=1#metadata_info_tab_contents