City Crimes and Seasonality

Vincent Xu yx2021@nyu.edu, Tyler Zhang tz2076@nyu.edu

1. Abstract

The purpose of this project was to analyze city crime activities from a time series perspective, focusing on both the macro and micro trends and patterns in crime occurrence. We obtained crime records from New York City and Chicago for the past 20 years and utilized statistical methods and visualizations to identify patterns and stationarity in the data. Our findings revealed a decreasing trend in crime occurrences for both cities over the past 20 years. Furthermore, both cities exhibited seasonal patterns in criminal activities, although the specific patterns differed between them, and the patterns varied across different sub-crime types. In the second part of this project, we built a prediction model using seasonal ARIMA with cross-validation and assessed its performance on the test set. Our model was effective in capturing the seasonal patterns in crime activities for both cities, although it did not incorporate the impact of COVID-19, which is acceptable. This model can be beneficial for predicting future crime peaks and serving as a guide for policymakers and police departments. Additionally, it provides an excellent starting point for further research on seasonal patterns of city crimes across the entire United States. Overall, our research provides insights into the trends and patterns of city crime activities and highlights the potential usefulness of a predictive model for future crime prevention efforts.

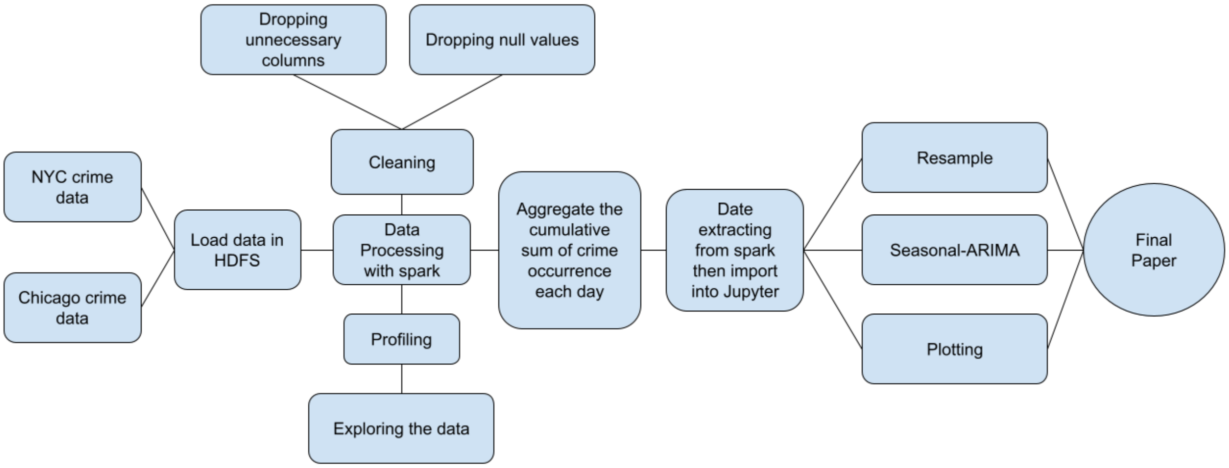
2. Introduction

Crime is a significant social issue that affects individuals, communities, and society as a whole. The pattern of crime occurrences used to be a hot topic in research fields. This project focuses on analyzing crime activities from a time series perspective to identify patterns and trends in crime occurrence over different scales including day, week, and month in urban areas.

The importance of this study lies in its potential to inform crime prevention strategies and improve public safety. Analyzing crime activities from a time series perspective can provide us with a better understanding of the underlying patterns, trends, and fluctuations. By doing so, we can make more accurate predictions of future crime peaks and provide alerts to police departments and policymakers to prepare appropriate adjustments and decrease the damage and loss caused by crime activities preemptively.

Previous studies on crime seasonality have primarily focused on the relationship between criminal activity and weather, particularly temperature. This is largely due to temperature aggression and routine activities theories. According to the temperature aggression theory, rising heat and humidity levels increase irritability in humans (McDowall et al., 2012b). The routine activities theory, on the other hand, emphasizes how temporal cycles structure individual behaviors, particularly changes in the physical environment (McDowall et al., 2012b), and how this influences crime activity. One classic assumption is that more pleasant weather encourages people to spend more time outdoors, leading to longer exposure to potential violent crimes, and an increased likelihood of theft due to absence from home. However, due to the controversial nature of these two theories, many studies have failed to align, resulting in no uniform consensus on the cause of crime peaks.

This project advanced the knowledge by focusing only on the number of crime occurrences itself as a variable to do a time series analysis, thereby preventing the potential noises introduced by other variables such as temperature and other weather data. Moreover, we chose two city-scale datasets with crime records for more than 20 years, allowing for a more comprehensive examination of crime seasonality at the city level. This approach is more adaptive to modern police systems and can provide valuable information in addressing crime peaks.


Dropping unnecessary columns
Dropping null values
Cleaning
NYC crime data
Chicago crime data
Load data in HDFS
Data Processing with spark
Profiling
Exploring the data
Aggregate the cumulative sum of crime occurrence each day
Date extracting from spark then import into Jupyter
Resample
Seasonal-ARIMA
Plotting
Final Paper

The data flow diagram above presents a comprehensive workflow of our project. We first ingested the crime dataset into HDFS, and used Spake to do Data Profiling, cleaning, and preprocessing. Then, the aggregation is performed on the dataset to turn it into time series format and loaded into Jupyter to do time series analysis and model building with Seasonal-ARIMA.

3. Motivation

The impact of crimes on society is significant. Criminal behavior disregards the protection of individual property, life safety, and personal dignity under the law. This is the root cause of the social disorder and chaos and contradicts the inevitable path of social development and the improvement of people's well-being.

Specifically, our motivation in choosing this topic lies in our past personal experiences. One of our friends was assaulted publicly, and we also had another friend who was robbed at knife-point and forced to withdraw money from an ATM by the robbers.

Furthermore, when we were kids, a parent of one of our friends who worked at a police department in China always had to work overtime during the summer, implying a possible surge in demand for the police during that period.

These real-life stories make us realize the importance of public safety and choose crime seasonality as our research topic. We hope by investigating crime activity patterns and trends in crime occurrence, police departments and policymakers can develop more effective prevention strategies and improve public safety in our communities.

4. Related Work

Related works have been focusing on different approaches to study the seasonality pattern in criminal activities. In the study conducted by David McDowal et al, they use an extensive dataset that includes a monthly panel of 88 U.S cities followed over a twenty-four-year period to be associated with temperature. The model they chose is classical time series decomposition that divides a series into trend, seasonal, and random components. Particularly, in their model, each month is specified as a variable in the calculation. Their analysis found most crimes peak in July or August and fall to low points in February for all cities they investigated as a whole and the temperature variation fails to explain the seasonal fluctuation of crime occurrences (McDowall et al., 2012b).

Despite extensive research on the seasonality of crime, no definitive consensus or explanations have been reached regarding the trend. In 2017, a research team from the University of Toronto, led by Linning, S. J., investigated the seasonal pattern in eight cities in British Columbia, Canada, between 2000 and 2006. They found that changes in weather modify people's routines, which in turn influences the frequency of crime. Additionally, the patterns they discovered in their research are not universal across all cities, so they cannot be generalized. Lastly, they suggest that the seasonality of crime should be studied at the type level to achieve greater accuracy (Linning et al., 2017).

Compared with the two studies mentioned above, our project chose data specifically from the city crimes of New York City and Chicago in the past 20 years and only focused on the number of crime occurrences itself as the target variable. By doing so, we could avoid the noise introduced when measuring other temporal and spatial variables such as temperature and spatial data. Taking one step further on McDowall's research, besides the overall pattern of cities as a whole, our project showed that there is also a city-level seasonal pattern that differs for each city aligning with Linning's finding. For instance, peaks of crime occur around June in Chicago, and peaks occur in February in NYC. Moreover, as suggested by Linning, our project inspected crime types, and we found that the seasonality pattern differs significantly for various crime types and cities.

5. Description of Datasets

5.1 NYPD Arrests Data (Historic)

The first source is a dataset provided by the New York City Police Department that contains information on all arrests made in the city since 2006, updated annually. The data contains information about the demographic of the arrested individual, the date and location of the arrest, as well as the digital codes and string descriptions that categorize it. It is a public data source that is used in various fields of research including disparity, local safety, and time series analysis. The size of this dataset is 1.16 GB composited with 5.31 million rows and 19 columns.

5.2 Chicago Crimes

The second source is a dataset from Chicago Data Portal, which contains similar features to the previous source. However, the classification codes within these two departments are different and we have to manually link them together. It contains information about crimes reported from 2001 to the present, updated daily. The size of this dataset is a bit large at size 1.84 GB composited with 7.77 million rows and 22 columns.

5.3 Selected Features from Datasets

All data from two sources are read as String into our project, and further casting and other processes are done on Spark.

Here are the features included in NYPD dataset: ARREST_KEY, ARREST_DATE, PD_CD, PD_DESC, KY_CD, OFNS_DESC, LAW_CODE, LAW_CAT_CD, ARREST_BORO, ARREST_PRECINCT, JURISDICTION_CODE, AGE_GROUP, PERP_SEX, PERP_RACE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat

And following are the features included in Chicago Crime dataset: ID, Case Number, Date, Block, IUCR, Primary Type, Description, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X Coordinate, Y Coordinate, Year, Updated On, Latitude, Longitude, Location.

However, our project focuses on the seasonal patterns for crimes happening in those two cities, not all columns are required. Below is a table describing the variables we select for the project, their descriptions, and value types.

Data Source	Field Name	Description	Value Type
NYPD Crime	ARREST_KEY	Randomly generated persistent ID for each arrest	Integer
	ARREST_DATE	Exact date of arrest for the reported event	Date Object
	KY_CD	Three digit internal classification code (more general category than PD code)	Integer
	OFNS_DESC	Description of internal classification corresponding with KY code (more general category than PD description)	String
Chicago Crime	Date	Date when the incident occurred. this is sometimes a best estimate.	Date Object
	Primary Type	The primary description of the IUCR code.	String
	Description	The secondary description of the IUCR code, a subcategory of the	String

Data Source	Field Name	Description	Value Type
		primary description.	

5.4 Combining Data Sources

The combination of the NYPD Arrests Data (Historic) from NYC Open Data and the Crimes - 2001 to Present from Chicago Open Data can generate a more extensive dataset for researching seasonal crime patterns. The merging of the data can enhance the sample size, minimize the randomness effect, and improve the credibility of the findings. Moreover, examining the crime trends of two cities can help identify commonalities and disparities in the factors that influence criminal activities, including weather, economy, and population. This can offer valuable insights for policymakers and law enforcement agencies aiming to reduce and prevent crime in their respective cities.

6. Analytic Stages, process

6.1 Ingestion

Both datasets are downloaded in CSV format, and uploaded to HDFS on NYU HPC with the command `hdfs dfs -put file.csv final_project/data`.

6.2 Profiling

During the profiling stage, we reviewed the schema of both datasets, counted the occurrences of distinct values, and grouped the data by crime type to determine which specific types of crimes to investigate in addition to the overall patterns. We ultimately decided to focus on Assault(Type 1) and Theft(Type 2), which are prevalent in both datasets and easily recognizable by the general public. However, these crime types are labeled differently in the two datasets. In the NYPD dataset, Assault is named "Assault" while Theft is referred to as "Larceny". On the other hand, in the Chicago dataset, Assault is classified as "Battery", while Theft remains labeled as "Theft".

6.3 Cleaning

In this stage, both datasets undergo cleaning and transformation to produce cleaned files for analysis. Three separate files are generated for each city, including one for all types of crime occurrences on each day, one for the daily occurrence of type 1, and one for the daily occurrence of type 2.

The following procedure was followed for handling the data. Firstly, we read the data as a CSV file, with all fields characterized as strings. Then, we dropped unnecessary columns within the data and removed rows containing NaN values. Next, we constructed two new dataframes containing only the required type of crime. It is worth noting that in the NYPD data, there is no general code for larceny and assault, so we used a method based on regular expressions to locate records that contain the words "Larceny" and "Theft". After the construction process, we used these three datasets to aggregate the occurrence of crime on a daily scale, respectively. The resulting data was saved in new CSV files, which are ready to be sent to the analysis stage.

6.4 Analytics

6.4.1 Setting Up

All analytic works are done on Jupyter Notebook, the instruction for setting up the environment can be found in the readme file.

6.4.2 Time Series Analysis

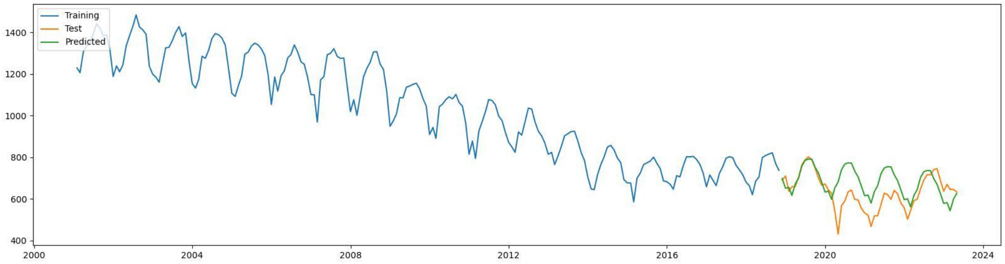
We have a total of nine notebooks, comprising three cities (Chicago, NYC, and a merged dataset) and three crime types (all crime types, type 1 which includes theft/larceny, and type 2 which includes battery/assault).

All notebooks follow the same analysis pipeline which includes the following steps:

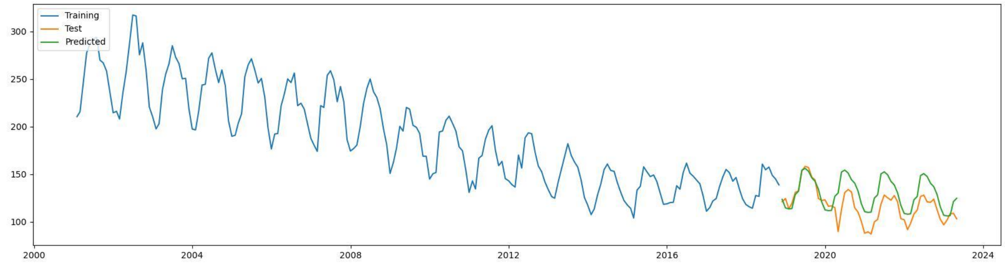
1. Import necessary packages and resolve dependencies.
2. Load datasets, and initialize dataset-specific variables.
3. Set up the data frame by converting it to a date object and setting it as an index.
4. Visualize crime frequency at different scales (daily, weekly, monthly).
5. Check if the dataset is stationary using the augmented Dickey-Fuller test.
6. Perform differencing by shifting the data by 12 months to remove macro trends from the data.
7. Perform the augmented Dickey-Fuller test again on the differenced data to check stationarity.
8. Create ACF and PACF plots to identify potential lag values with high auto-correlation and partial auto-correlation.
9. Split the dataset into training and testing sets.
10. Train the seasonal ARIMA model and optimize it with parameter tuning.
11. Show the model summary.
12. Use the model to make predictions on the test set and visualize the true values versus the predicted values.
13. Output the plots to the "output" folder.
14. Evaluate each model on the test set using the root mean square error (RMSE).

7. Graph(s) - a visual representation of your analytics

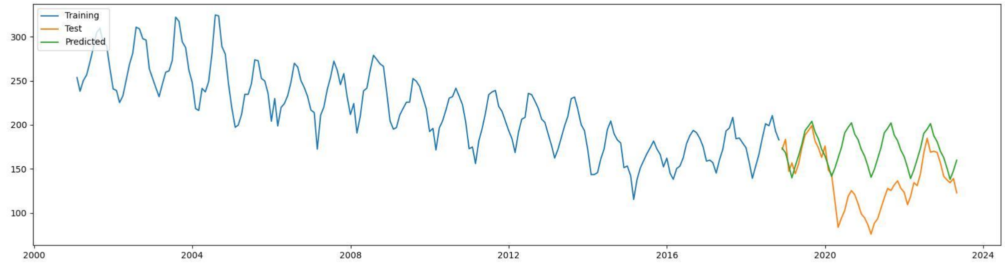
7.1 City: Chicago — Crime: All



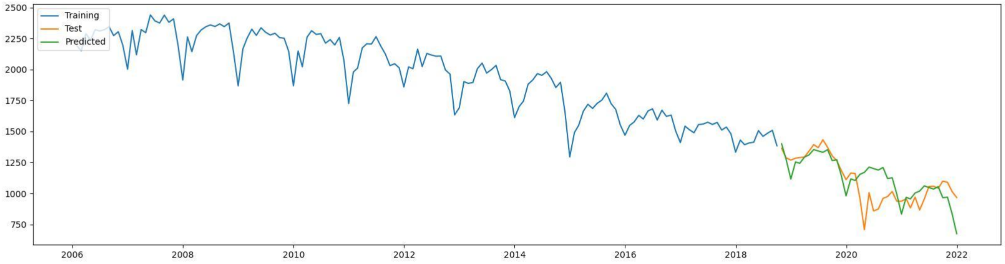
7.2 City: Chicago — Crime: Type 1



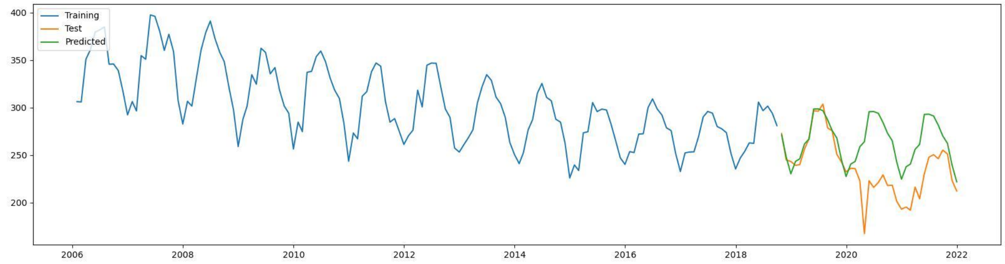
7.3 City: Chicago — Crime: Type 2



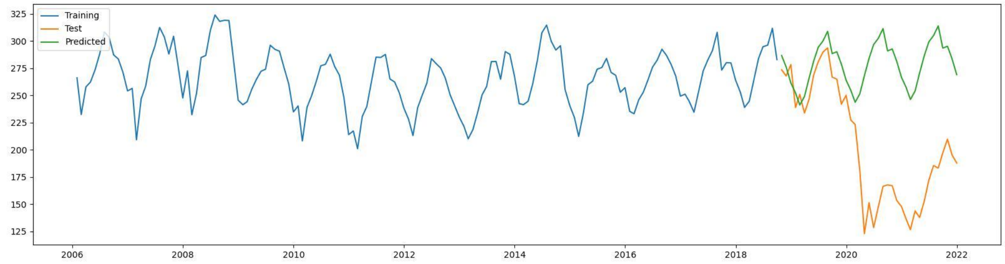
7.4 City: Merged — Crime: All



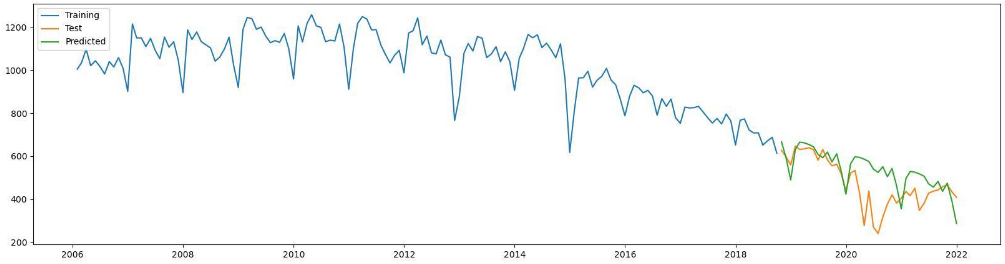
7.5 City: Merged — Crime: Type 1



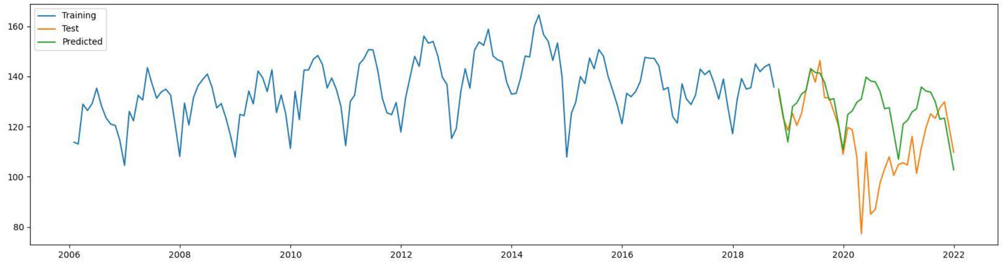
7.6 City: Merged — Crime: Type 2



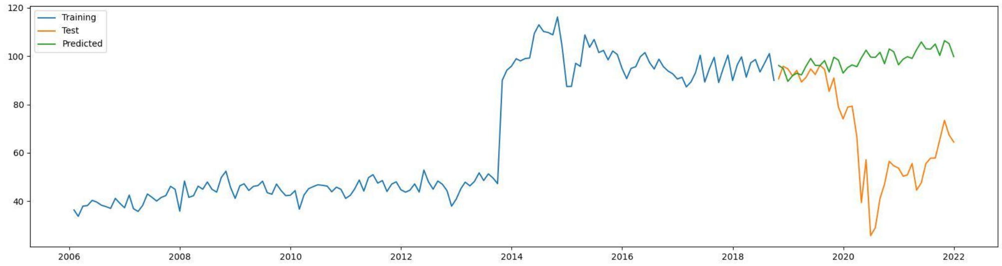
7.7 City: NYC — Crime: All



7.8 City: NYC — Crime: Type 1



7.9 City: NYC — Crime: Type 2



8. Conclusion

Overall, most of the models performed well in predicting the occurrence of the following time series. It is evident that both cities show a clear trend of decreasing occurrence and follow a seasonal pattern over 20 years. However, similar to previous findings, the peaks and drops are located differently in these two cities. In Chicago, the peaks occur around the middle of the year, while the drops occur at the beginning and end of the year. Conversely, NYC's peak is in February and gradually decreases throughout the year. It is worth noting that significant drops are visible after 2020, which is most likely due to the pandemic. The models we developed emphasize the significance of seasonality in crime analysis, and the forecasts could aid in predicting future crime highs and lows in specific cities. This is valuable for policy-makers and law enforcement agencies in allocating resources more effectively. However, our project is not perfect. The method of aggregating the monthly average occurrences of crimes is not weighted, such that we had not accounted for the effect of holidays, which may introduce some biases. Moreover, additional data that we have dropped, such as race and location, could be explored to identify any confounding factors.

9. Reference

Andresen, M. A., & Malleson, N. (2013). Crime seasonality and its variations across space. Applied Geography, 43, 25–35. https://doi.org/10.1016/j.apgeog.2013.06.007

Linning, S. J., Andresen, M. A., Ghaseminejad, A. H., & Brantingham, P. J. (2017). Crime Seasonality across Multiple Jurisdictions in British Columbia, Canada. Canadian Journal of Criminology and Criminal Justice, 59(2), 251–280.

McDowall, D., Loftin, C., & Pate, M. (2012). Seasonal Cycles in Crime, and Their Variability. Journal of Quantitative Criminology, 28(3), 389–410. https://doi.org/10.1007/s10940-011-9145-7

Schinasi, L. H., & Hamra, G. B. (2017). A Time Series Analysis of Associations between Daily Temperature and Crime Events in Philadelphia, Pennsylvania. Journal of Urban Health-bulletin of the New York Academy of Medicine, 94(6), 892–900. https://doi.org/10.1007/s11524-017-0181-y
A Time Series Analysis of Associations between Daily Temperature and Crime Events in Philadelphia, Pennsylvania | SpringerLink

Chicago Crime Data: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

NYC Crime Data: https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u