TIME SERIES ANALYSIS AND FORECASTING ON CHICAGO CRIME DATASET

Group members:

Max Jacobs	mlj102
Mayukh Sen	ms3802
Krish Shah	ks1984
Divya Shah	ds2097
Manan Jagani	mj868

Table of Contents

ABSTRACT

In this project, we aimed to perform Time Series Analysis and accurate forecasting on the Chicago Crime Dataset. We began by identifying and dropping entries with missing data and removing unnecessary columns during the preprocessing stage. Exploratory Data Analysis was conducted to analyze the number of crimes based on crime types, days of the week, times of day, months, and crime locations.

During the preprocessing stage, we performed feature engineering by creating new columns for time, date, weekday, and month. Additionally, we computed the total crime counts per day based on these features. Plots and heatmaps of crime counts were generated to effectively visualize and understand the crime statistics in Chicago. Spatial analysis involved creating a map of Chicago and obtaining heatmaps based on the number of crimes committed in each Police District and Ward.

In the modeling stage, we focused on forecasting in districts with the highest number of crimes. Time Series Models such as ARIMA, SARIMA, TBATS, and Holt-Winters Method were experimented with. The Augmented Dickey-Fuller Test (ADF) was employed to check for stationarity. The Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were studied to analyze trends and seasonality components. To effectively capture seasonality, we ultimately employed Long Short-Term Memory (LSTM), a type of recurrent neural network (RNN) architecture. Various experiments were conducted to identify the best model fit based on the criteria of Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE).

INTRODUCTION

For this project we employed the ARIMA model . A time series $\{Yt\}$ is said to follow an **integrated autoregressive moving average** model if the dth difference $Wt = \nabla dYt$ is a stationary ARMA process. If $\{Wt\}$ follows an ARMA(p,q) model, we say that $\{Yt\}$ is an ARIMA(p,d,q) process.
ARIMA model is used to perform effective forecasting on non-stationary time series. However it does not take into account the seasonality of the time series.

In order to capture the seasonality component, we used **Seasonal ARIMA**. It is an extension of the ARIMA. Its components are **Seasonal Autoregressive (SAR), Seasonal Integrated (SI)**, and **Seasonal Moving Average (SMA)** components. Parameters are (**p, d, q, s**) representing the seasonal order. Given below is an example of a SARIMA Time Series

$$(1-\phi_1 B)\,(1-\Phi_1 B^4)\,(1-B)\,(1-B^4)y_t \;=\; (1+\theta_1 B)\,(1+\Theta_1 B^4)e_t.$$

(Non-seasonal AR(1)) (Seasonal AR(1)) (Non-seasonal difference) (Seasonal difference) (Non-seasonal MA(1)) (Seasonal MA(1))

In order to further improve our predictions we employed the **TBATS** model. It stands for **Trigonometric Seasonal, Box-Cox Transformation, ARMA errors, Trend, and Seasonal components.** It is a Robust time series forecasting method that handles multiple seasonalities with **trigonometric functions**. It employs **Box- Cox Transformation** to reduce variance.
TBATS uses the trigonometric representation of seasonal components based on Fourier series
It is an improvement on the BATS model to handle complex seasonality in the Time Series.
Based on trigonometric functions, it can be used to model non-integer seasonal frequencies

Model:

$$y_t^{(\lambda)} = l_{t-1} + \phi\, b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t$$

$$l_t = l_{t-1} + \phi\, b_{t-1} + \alpha\, d_t$$

$$b_t = \phi\, b_{t-1} + \beta\, d_t$$

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i e_{t-i} + e_t$$

Where:

$y_t^{(\lambda)}$ - time series at moment t (Box-Cox transformed)

$s_t^{(i)}$ - ith seasonal component

l_t - local level

b_t - trend with damping

d_t - ARMA(p,q) process for residuals

e_t - Gaussian white noise

Seasonal part:

$$s_t^{(i)} = \sum_{j=1}^{(k_i)} s_{j,t}^{(i)}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos(\omega_i) + s_{j,t-1}^{*(i)} \sin(\omega_i) + \gamma_1^{(i)} d_t$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin(\omega_i) + s_{j,t-1}^{*(i)} \cos(\omega_i) + \gamma_2^{(i)} d_t$$

$$\omega_i = 2\pi j / m_i$$

Model parameters:

T - Amount of seasonalities

m_i - Length of ith seasonal period

k_i - Amount of harmonics for ith seasonal period

λ - Box-Cox transformation

α, β - Smoothing

ϕ - Trend damping

φ_i, θ_i - ARMA(p, q) coefficients

$\gamma_1^{(i)}, \gamma_2^{(i)}$ - Seasonal smoothing (two for each period)

TBATS considers models with Box-Cox transformation and without it, with and without Trend, with and without Trend Damping, with and without ARMA(p,q) process used to model residuals. It also considers non-seasonal models, and various amounts of harmonics used to model seasonal effects. The final model will be chosen using Akaike information criterion (AIC).

For better forecasting, we implemented the **Holt-Winters** model. It is also known as the **Triple Exponential Smoothing method**. It is very useful when data has a seasonal component. It has
Level (l): Represents the average value in the time series.
Trend (b): Represents the average growth or decline in the time series.
Seasonality (s): Represents the repeating patterns or cycles in the time series.
Holt-Winters uses three smoothing equations to update the level, trend, and seasonality components. The method is adaptive and considers both the recent values and the historical patterns to make predictions.

Naive Method

This is the most primitive forecasting method. The premise of the naive method is that the expected point is equal to the last observed point:

$$\hat{y}_x = \frac{1}{x} \sum_{i=1}^{x} y_i$$

Simple Average

A less primitive method is the arithmetic average of all the previously observed data points. We take all the values we know, calculate the average and bet that that's going to be the next value. Of course it won't be it exactly, but it probably will be somewhere in the ballpark, hopefully you can see the reasoning behind this simplistic approach.

$$\hat{y}_{x+1} = y_x$$

Single Exponential Smoothing

$$\hat{y}_x = \alpha y_x + (1 - \alpha)\hat{y}_{x-1}$$

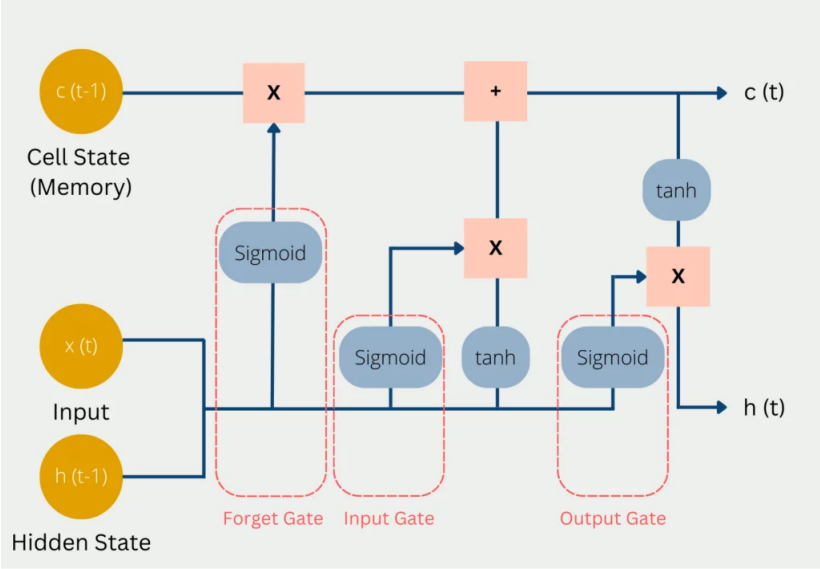
convert the data into a time series object.

Holt-Winters produces forecasts by combining the level, trend, and seasonality components.

The forecasted value is the sum of the current level, the estimated trend, and the adjusted seasonal component.

For the predictions with the least RMSE we employed **LTSM** , **Long Short-Term Memory (LSTM)** is a type of **recurrent neural network** (RNN) architecture designed to overcome the limitations of traditional RNNs in capturing long-term dependencies in sequential data. LSTMs consist of **memory cells, input gates, forget gates and output gates**. LTSMs are very useful when handling sequential data.

LSTMs are trained using **gradient descent** and **backpropagation** through time, where the weights are adjusted to minimize the difference between predicted and actual values. The mechanism of a simple LSTM is given below.



For this project, we employed **two LSTM Layers** followed by a dropout layer. **One Dense Layer** is also added as the **output layer**. The model is compiled using the Adam optimizer and **mean squared error (MSE)** as the loss function.

DATA

Data Description

We used the Chicago Crime Dataset filtered from January 1,2008 to October 31,2023 for our time series forecasting . Important attributes of the database include: Date, Time, Crime type, Description of crime, Block Address, Ward, District, Location type, Arrest, Domestic.
It can be accessed from here.

Data Overview:

1. Our Data consisted of **4674587 rows and 22 columns** initially.

2. A snapshot of the dataset:.

```
crimes.head(100)
```

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...	X Coordinate	Y Coordinate	Year	Updat
0	13213472	JG426663	2008-01-01 00:00:00	043XX W 26TH ST	1752	OFFENSE INVOLVING CHILDREN	AGGRAVATED CRIMINAL SEXUAL ABUSE BY FAMILY MEMBER	APARTMENT	0	1	...	NaN	NaN	2008	09/18/20 03:42:
1	12301302	JE152215	2008-01-01 00:00:00	061XX N JERSEY AVE	0281	CRIMINAL SEXUAL ASSAULT	NON-AGGRAVATED	HOTEL / MOTEL	0	0	...	NaN	NaN	2008	02/26/20 03:40:
2	12954680	JG118873	2008-01-01 00:00:00	065XX S LAFLIN ST	1582	OFFENSE INVOLVING CHILDREN	CHILD PORNOGRAPHY	APARTMENT	0	1	...	NaN	NaN	2008	02/14/20 03:45:
3	13025083	JG203207	2008-01-01 00:00:00	054XX N ASHLAND AVE	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	APARTMENT	0	0	...	NaN	NaN	2008	03/30/20 03:41:
4	6020368	HP117521	2008-01-01 00:00:00	076XX S EAST END AVE	1750	OFFENSE INVOLVING CHILDREN	CHILD ABUSE	APARTMENT	0	1	...	1188862.0	1854636.0	2008	02/28/20 03:56:

Data Preprocessing:

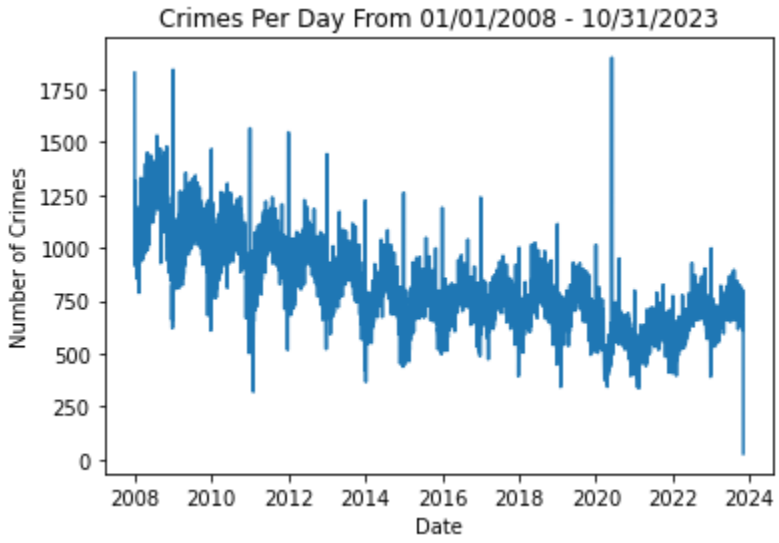
The data preprocessing involved the following steps:

1. Several entries with missing data were identified, and those entries were dropped. Following this step, 4,610,839 entries remained.
2. The date column was converted into the proper date-time format.
3. The dataset was enhanced with new columns—Month, Weekday, and Hour—derived from the date information.
4. The date column was set as the new index of our dataset.
5. Categorical columns, namely Arrest and Domestic, were encoded into 0s and 1s from boolean values.
6. A new column for Crime Type was created, where all crime types were classified based on the US Topological Crime Classification.

Date	ID	Case Number	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	...	Updated On	Latitude	Longitude
2008-01-01 00:00:00	6020368	HP117521	076XX S EAST END AVE	1750	OFFENSE INVOLVING CHILDREN	CHILD ABUSE	APARTMENT	0	1	414	...	02/28/2018 03:56:25 PM	41.756184	-87.583423
2008-01-01 00:00:00	6000402	HP107451	114XX S THROOP ST	1152	DECEPTIVE PRACTICE	ILLEGAL USE CASH CARD	RESIDENCE	0	0	2234	...	02/28/2018 03:56:25 PM	41.685508	-87.654327
2008-01-01 00:00:00	6064706	HP162069	001XX E 124TH PL	0890	THEFT	FROM BUILDING	RESIDENCE	0	0	532	...	02/28/2018 03:56:25 PM	41.667710	-87.619236
2008-01-01 00:00:00	6025760	HP129015	003XX N STATE ST	2820	OTHER OFFENSE	TELEPHONE THREAT	RESIDENCE	0	1	1831	...	09/07/2021 03:41:02 PM	41.887919	-87.628019
2008-01-01 00:00:00	6073869	HP173211	010XX E 101ST ST	0840	THEFT	FINANCIAL ID THEFT: OVER $300	RESIDENCE	0	0	511	...	02/28/2018 03:56:25 PM	41.710692	-87.598884

Exploratory Data Analysis

Distribution of the total number of crimes committed per day from 1/1/2008 to 10/31/2023



- We observe that over the time period the number of crimes committed every day in Chicago remains more or less between **1000 to 1750** in the early years of **2008 to 2010.**
- We observe a **steady decrease** in the number of crimes over the time period. This can be attributed to **better law enforcement** in the City of Chicago .

- We observe that the total number of daily crimes goes down to between **500 and 1250 in mid 2010s** and further down to between **500 and 1000 in late 2010s and post 2020**
- We notice a s**udden decline** in the number of crimes in **early 2020** which can be attributed to **COVID-19 outbreak**. A sudden rise in the number of crimes is observed around mid 2020 . That can be attributed to the **nationwide riots in response to the death of George Floyd.**

Distribution of the type of crimes

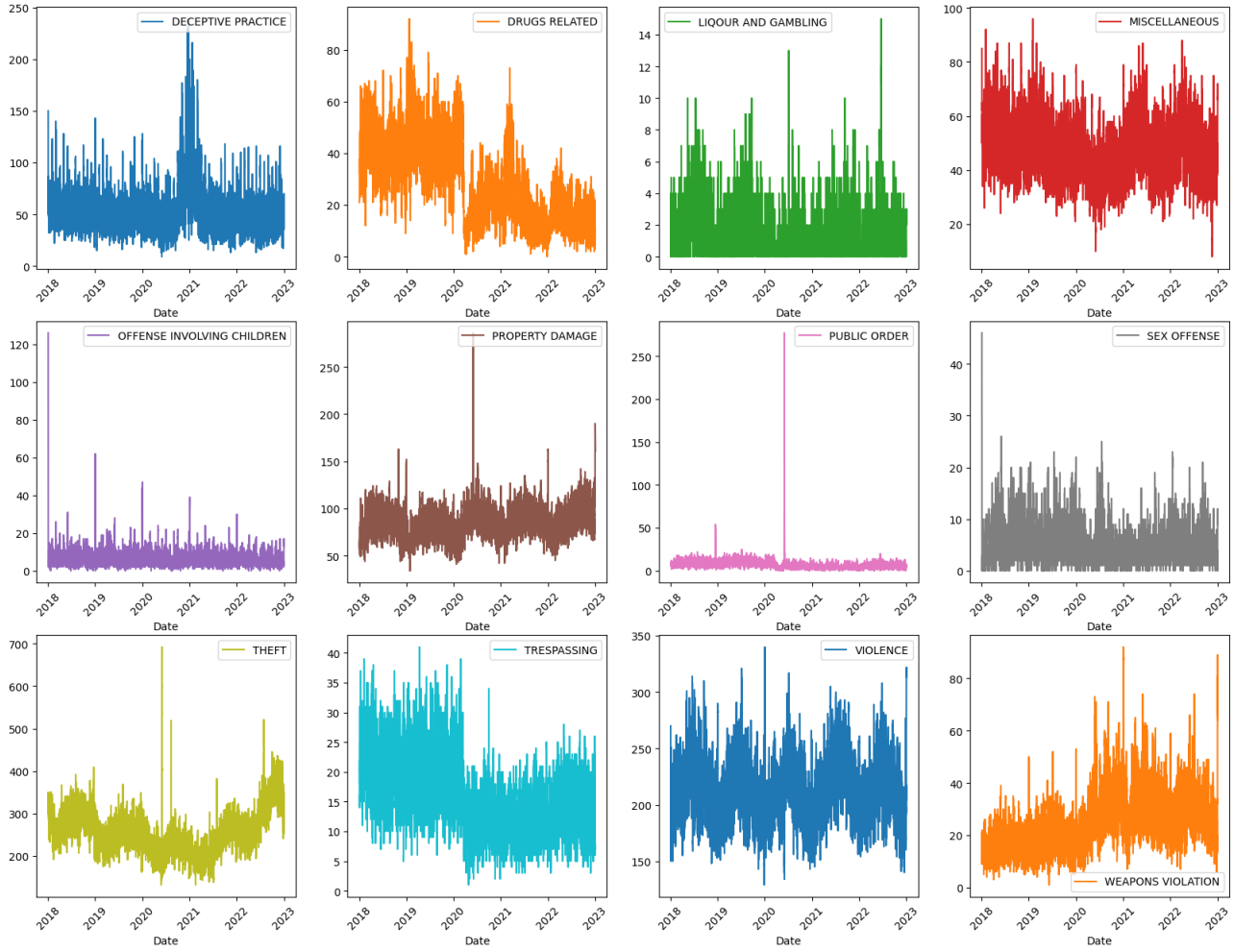
Let us analyze the type of crimes committed in Chicago most frequently.

```
df['Primary Type'].value_counts()
```

```
Primary Type
THEFT                                1012318
BATTERY                               838237
CRIMINAL DAMAGE                       516227
NARCOTICS                             361329
ASSAULT                               318491
OTHER OFFENSE                         284238
BURGLARY                              251250
DECEPTIVE PRACTICE                    234616
MOTOR VEHICLE THEFT                   229245
ROBBERY                               180561
CRIMINAL TRESPASS                     113690
WEAPONS VIOLATION                      82448
OFFENSE INVOLVING CHILDREN             34429
PUBLIC PEACE VIOLATION                 33477
PROSTITUTION                           24475
SEX OFFENSE                            17057
CRIM SEXUAL ASSAULT                    15540
INTERFERENCE WITH PUBLIC OFFICER       14907
HOMICIDE                                9165
ARSON                                   7621
CRIMINAL SEXUAL ASSAULT                 6691
GAMBLING                                6611
LIQUOR LAW VIOLATION                    6309
STALKING                                3554
KIDNAPPING                              3285
INTIMIDATION                            2652
CONCEALED CARRY LICENSE VIOLATION       1180
OBSCENITY                                677
NON-CRIMINAL                             173
PUBLIC INDECENCY                         151
OTHER NARCOTIC VIOLATION                  95
HUMAN TRAFFICKING                         92
NON - CRIMINAL                            38
NON-CRIMINAL (SUBJECT SPECIFIED)           9
RITUALISM                                  1
Name: count, dtype: int64
```

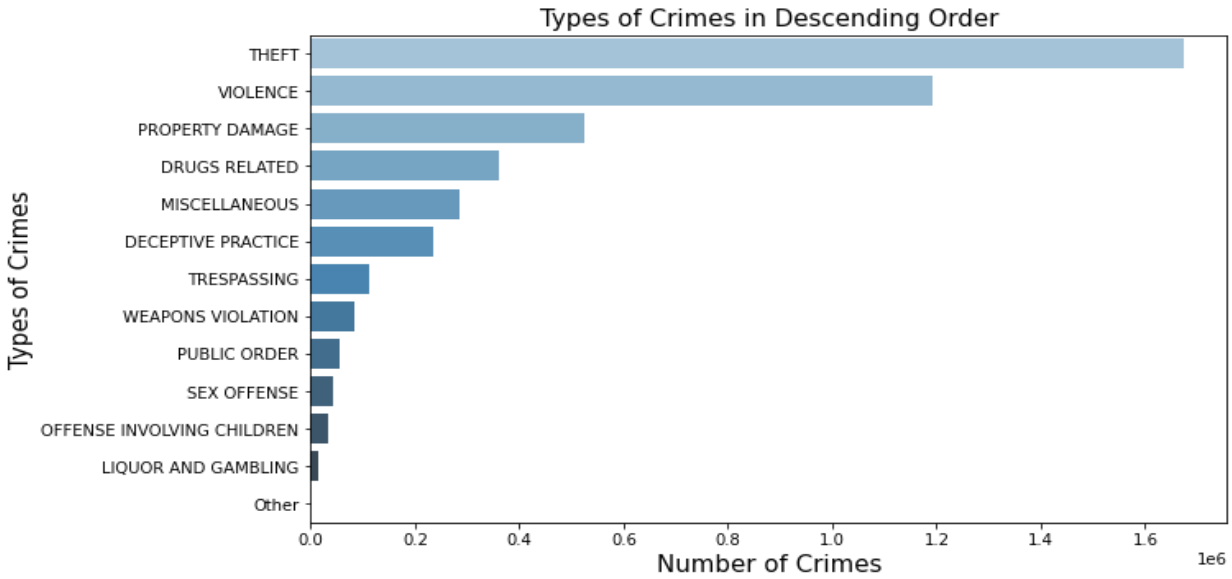
We see that the most prevalent type of crime is **Theft followed by Battery, Criminal Damage , Narcotics and Assault**.

We categorized these crimes into 12 broader classifications based on the US Topological Crime Classification.
These 12 Categories are : THEFT, VIOLENCE, PROPERTY DAMAGE, DECEPTIVE PRACTICE, DRUGS RELATED, WEAPONS VIOLATION, TRESPASSING, OFFENSE INVOLVING CHILDREN, SEX OFFENSE, LIQUOR AND GAMBLING, MISCELLANEOUS. Our analysis will focus on examining the distribution of crime types over the specified time period.

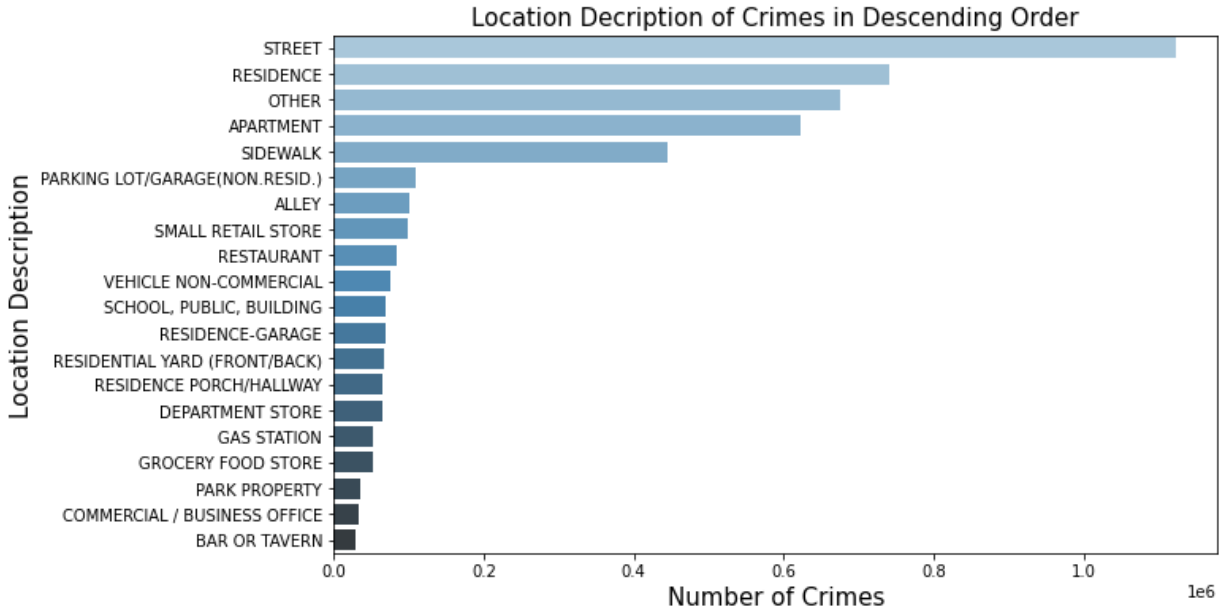


The analysis indicates a general decline in the occurrence of various crime types such as Drugs Related, Liquor and Gambling, Sex Offenses, Trespassing, Violence, and Miscellaneous crimes. Conversely, a noteworthy upward trend is observed in Theft and Weapons Violation over the recent years.

Distribution of Crime Types

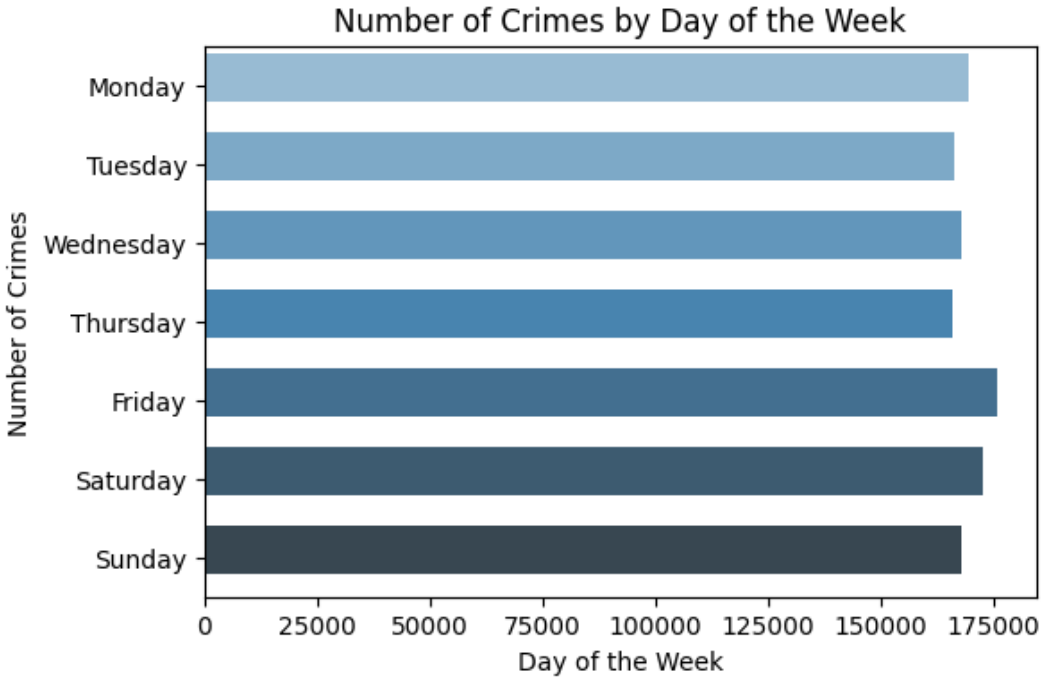


Distribution of the location of crimes



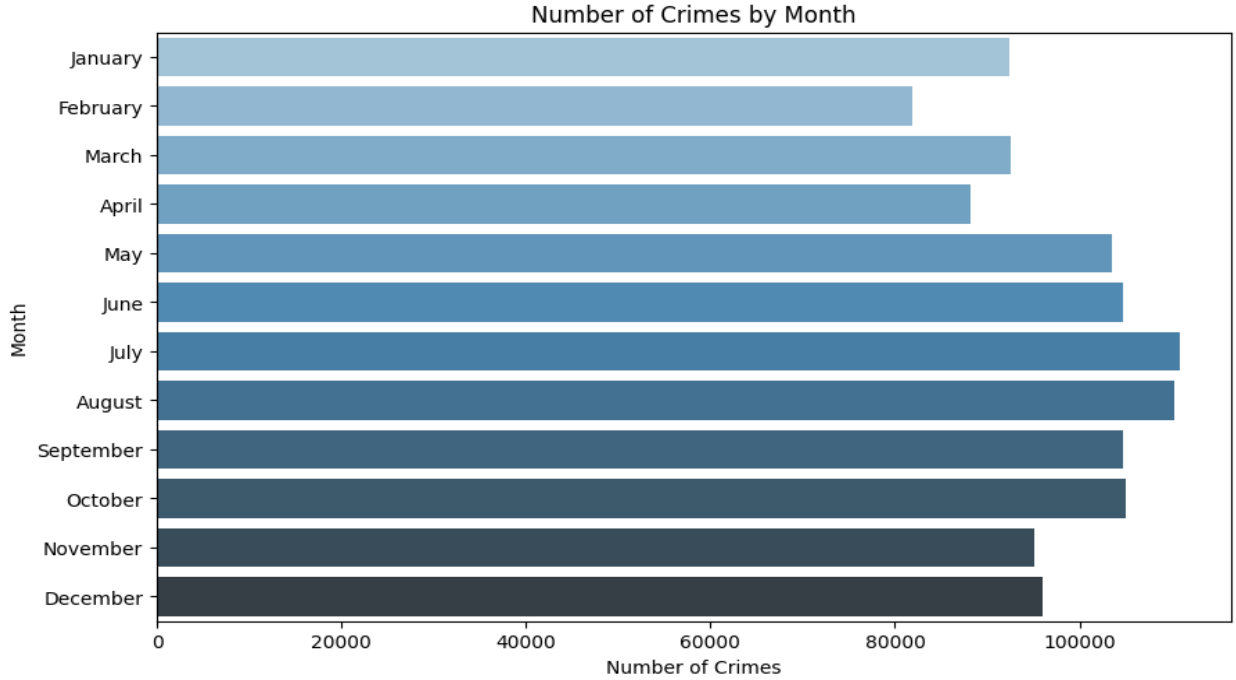
Our observation reveals that the **Streets, Residential Apartments, and Sidewalks** register the highest frequencies of criminal incidents.

Distribution of the number of crimes on each weekday



Observationally, Fridays stand out as the day with the highest incidence of crimes, while the remaining days exhibit a relatively uniform distribution of criminal activities.

Distribution of the number of crimes on each Month

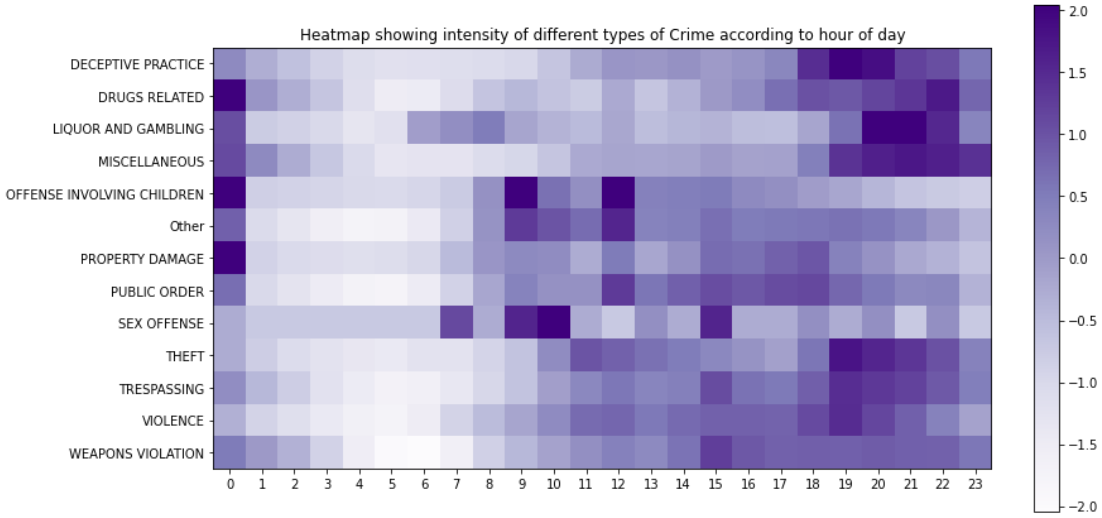


The graphical representation illustrates a discernible **decline** in crime rates during the **winter months**, coinciding with lower temperatures. Notably, **February** exhibits the lowest frequency of crimes, with December, November, and January also reflecting comparatively reduced crime numbers. Conversely, a notable surge in crime incidents is observed during the Summer months of July and August.

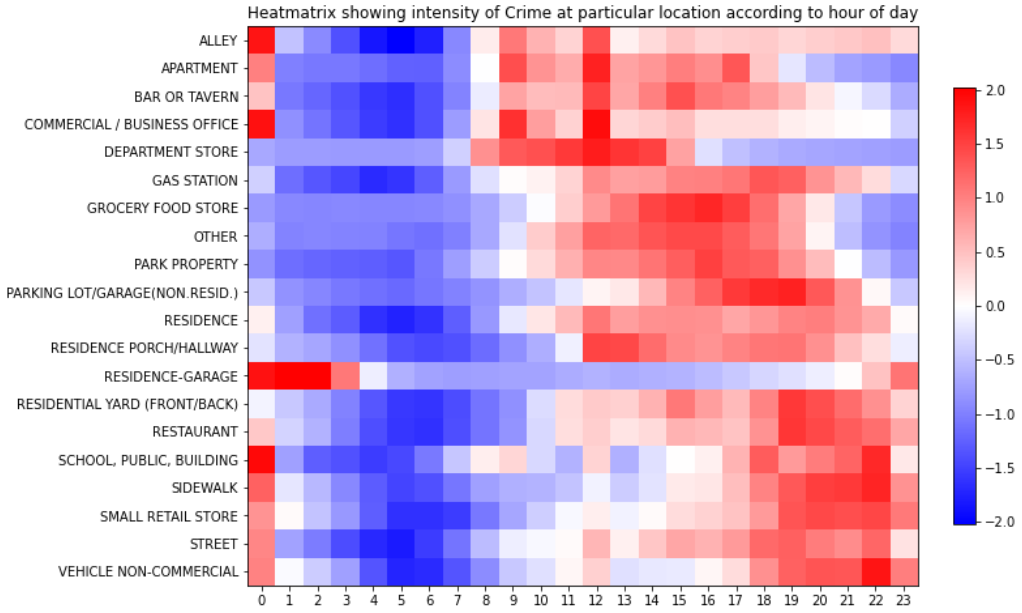
Correlation Heat Matrices

We plotted some heat matrices to understand the crime statistics of Chicago.

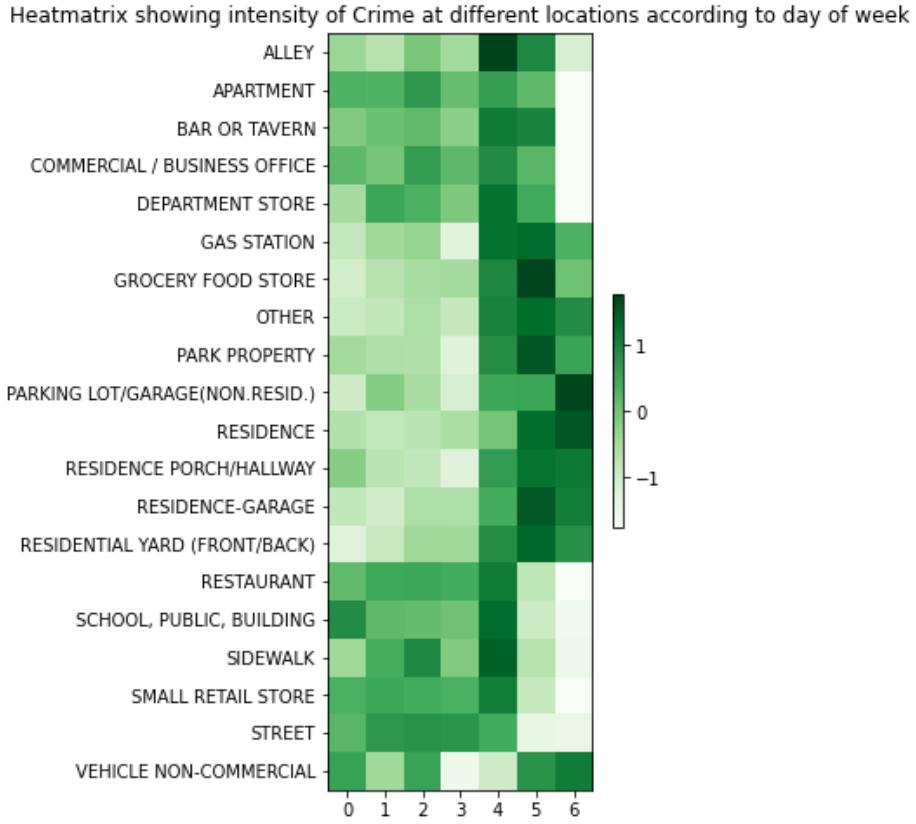
1. Correlation Heat Matrix showing the correlation between **different types of crimes committed with the hour of the day.**



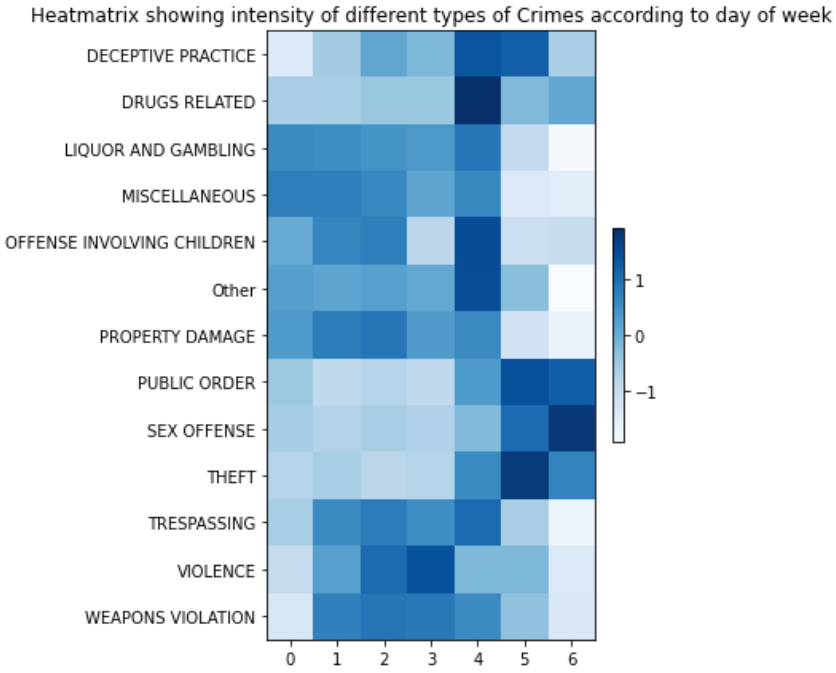
2. Correlation Heat Matrix showing the correlation between **intensity of crimes committed at a particular location with the hour of the day.**



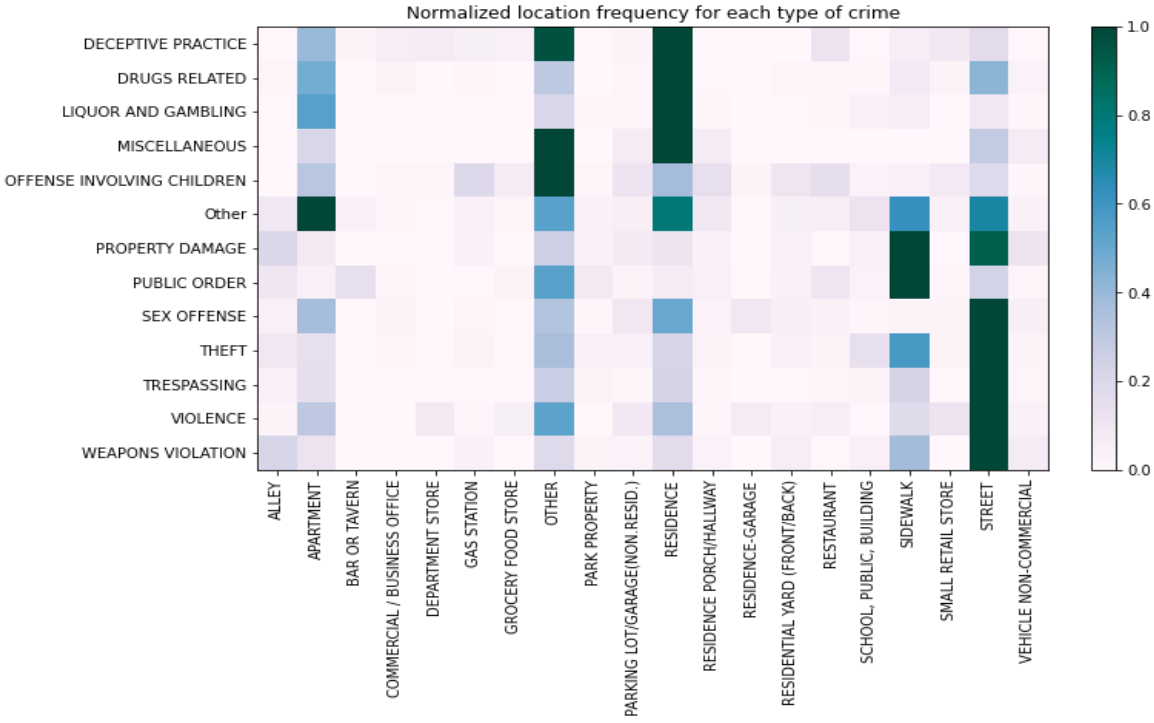
3. Correlation Heat Matrix showing the correlation between **intensity of Crime at different locations according to day of week.**



4. Correlation Heat Matrix showing the correlation between **intensity of different types of Crimes according to day of week.**



5. Correlation Heat Matrix showing Normalized location frequency for each type of crime.



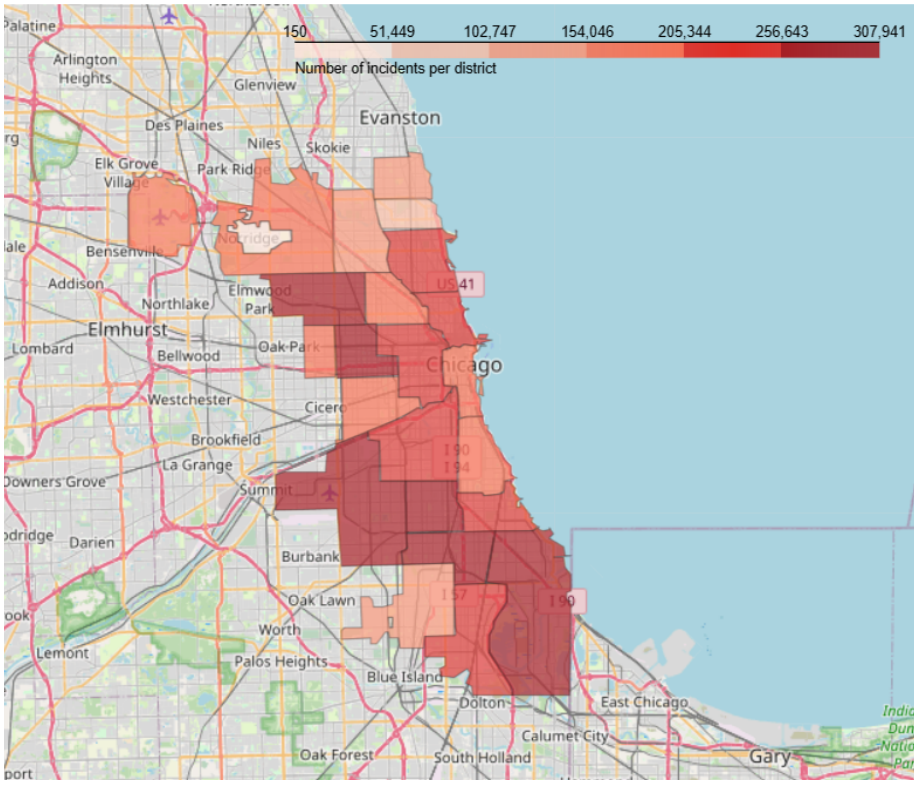
Insights from Heatmatrices:-

1. Crime rates spike in the evening, particularly between 7 PM and midnight, with an observable decline after midnight until 8 AM. Offenses involving children and sex crimes are more prevalent around 9 AM. There is an increased risk for offenses involving children around 1 PM, suggesting danger around school hours.
2. Crime rates surge in residential garages and offices after midnight, while departmental stores, grocery stores, offices, and apartments experience higher crime rates during the day. In the evening, streets, restaurants, sidewalks, schools, residential yards, and parking lots see increased criminal activity.
3. Crime rates escalate from Monday through Thursday in schools, public buildings, sidewalks, streets, restaurants, and department stores. Fridays witness higher crime rates across all areas. Weekends experience increased criminal activity in bars, alleys, residential areas, parking lots, and gas stations.
4. Fridays and weekends exhibit a higher occurrence of drug-related crimes, deceptive practices, sex offenses, theft, and public order violations. In contrast, trespassing, violence, and weapons violations are more prevalent from Monday through Thursday.
5. In residences, drug-related crimes, deceptive practices, and liquor and gambling offenses are common. On the streets of Chicago, theft, trespassing, weapons violations, and violence are prevalent. Sidewalks often witness property damage and public order disruptions.

Spatial Analysis

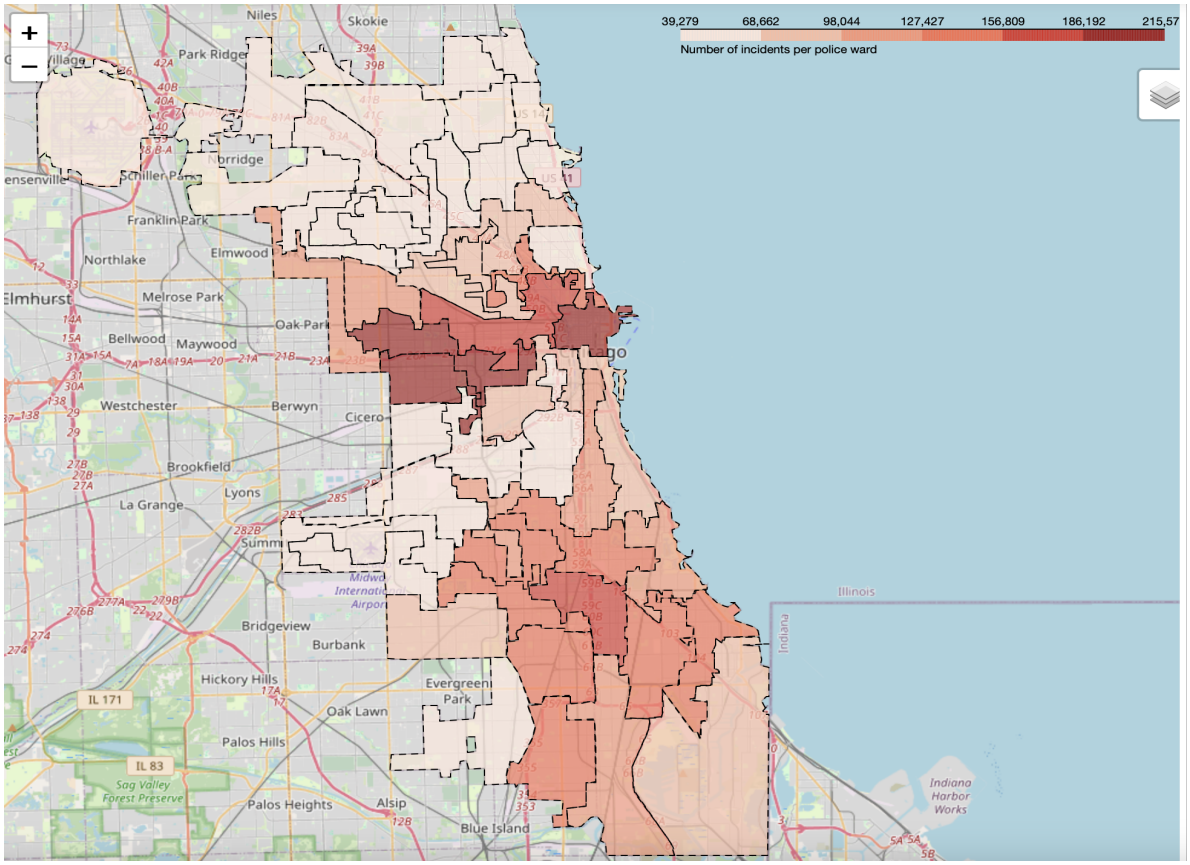
We performed Spatial Analysis on the map of Chicago. We plotted heatmaps of the intensity of crimes on the map of Chicago divided into districts and wards. We generated these heatmaps using the folium library in Python.

Heatmap of crimes committed across different districts of Chicago



This map highlights the districts in Chicago with the highest reported crime rates (250k-300k) from 2008 to 2023 in Districts 25, 11, 8, 7, 6, and 4. Additionally, Districts 5, 3, 9, 12, 18, and 19 exhibit substantial crime numbers, ranging from approximately 200k to 250k. In contrast, District 31 has the lowest crime rate, reporting fewer than 50k crimes over the 15 years. Districts 20, 24, 17, and 22 are relatively safer, with reported crime numbers ranging from 50k to 100k during the same timeframe.

Heatmap of crimes committed across different wards of Chicago



It has been noted that Wards 42, 24, and 28 experience the highest crime rates, ranging from 185k to 214k reported crimes between 2008 and 2023. Additionally, Wards 27 and 6 exhibit elevated crime rates, with reported crimes falling between 150k and 185k.
Wards 37, 29, 34, 21, 16, and 20 observe moderate crime rates, ranging from 90k to 150k during the same period. In contrast, all other police districts in North Chicago report fewer than 70k crimes within the given timeframe.

Notably, Wards 42, 27, 28 and 24 primarily cover downtown Chicago. Ward 42 features iconic districts like The Loop, the central business district boasting skyscrapers, financial institutions, and cultural attractions. It also includes River North, renowned for its art galleries, restaurants, and nightlife, and Streeterville, home to Navy Pier and Northwestern University's Chicago campus. Ward 42 also encompasses the Near North Side, including the famous Magnificent Mile shopping district. Wards 27 and 28 cover the West Loop, characterized by its trendy atmosphere with a vibrant restaurant scene and loft-style living spaces.

Our Time Series Analysis focused on data from these wards (42, 28, 27, and 24) as they collectively represent downtown Chicago and are considered the city's most notorious areas.

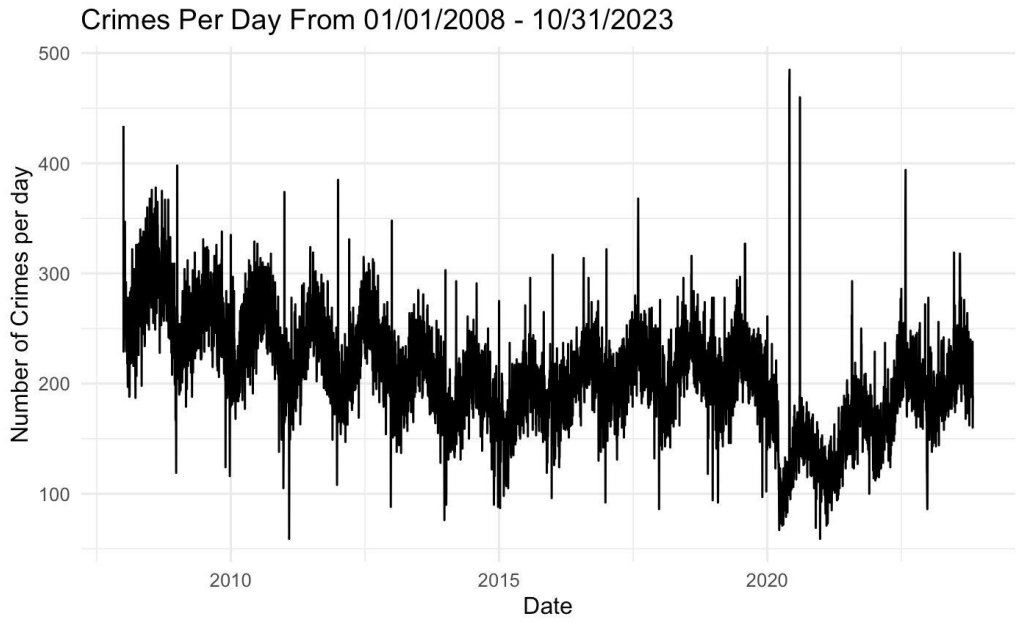
METHODOLOGY

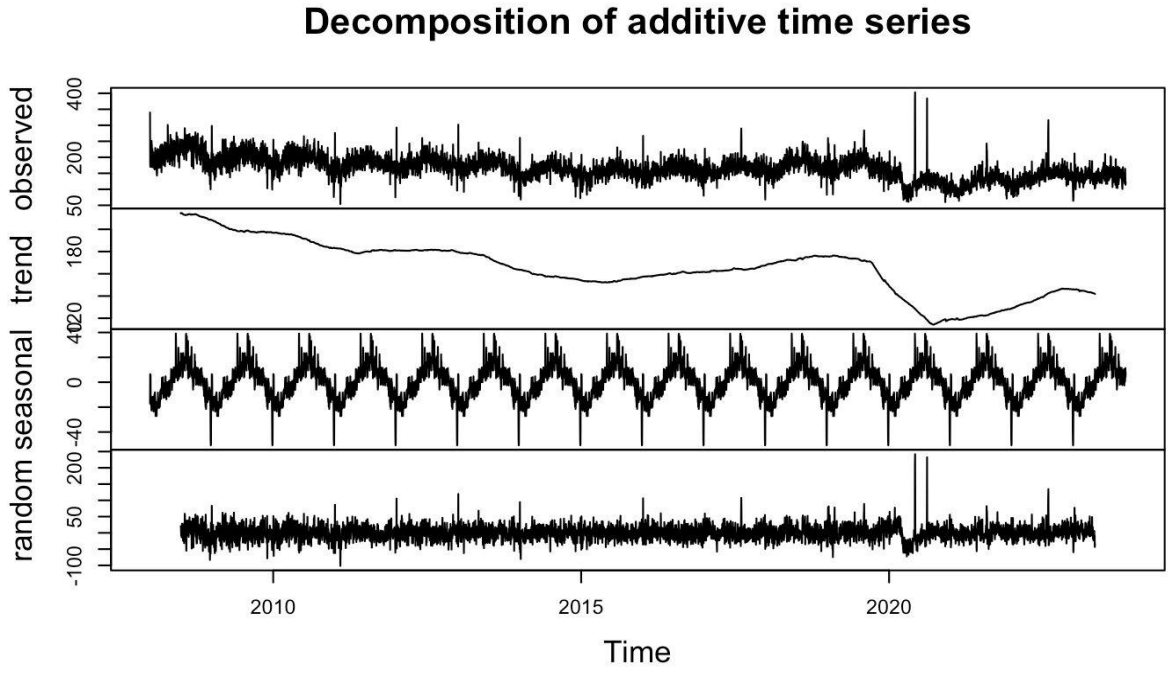
Filtering the Data based on wards 42, 28, 27, and 24

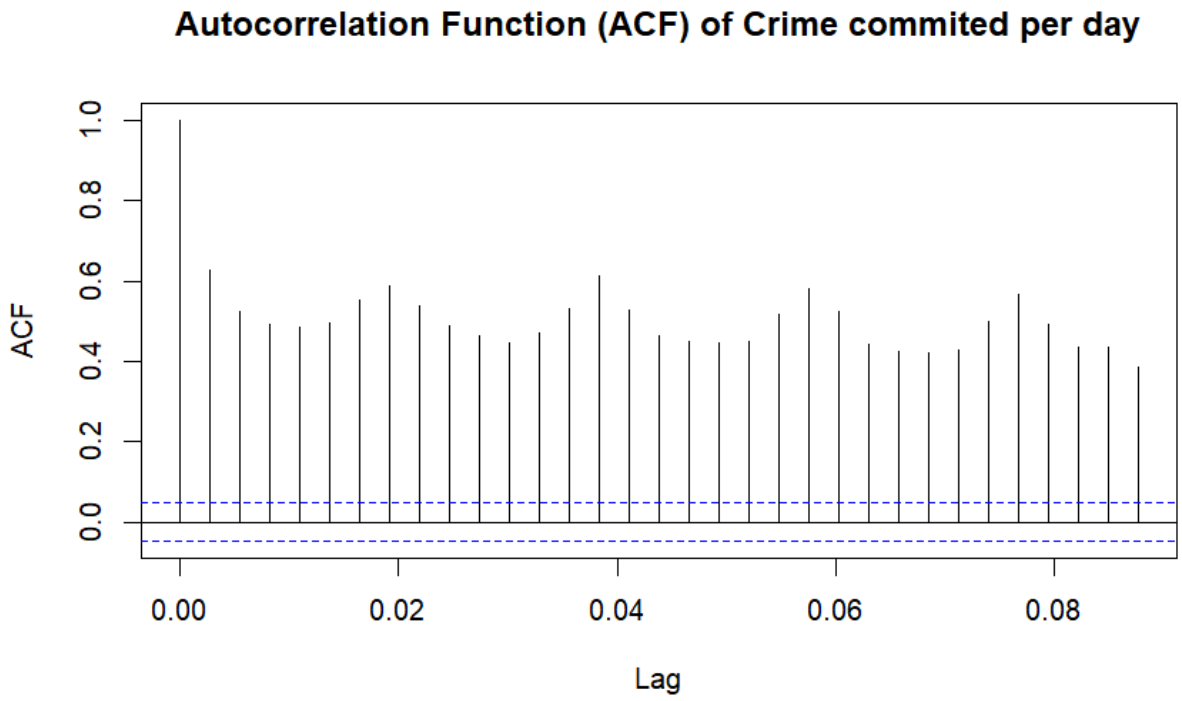
We performed our Time Series Analysis focused on data from these wards (42, 28, 27, and 24) as they collectively represent downtown Chicago and are considered the city's most notorious areas.

Time Series Plot on the Filtered Data

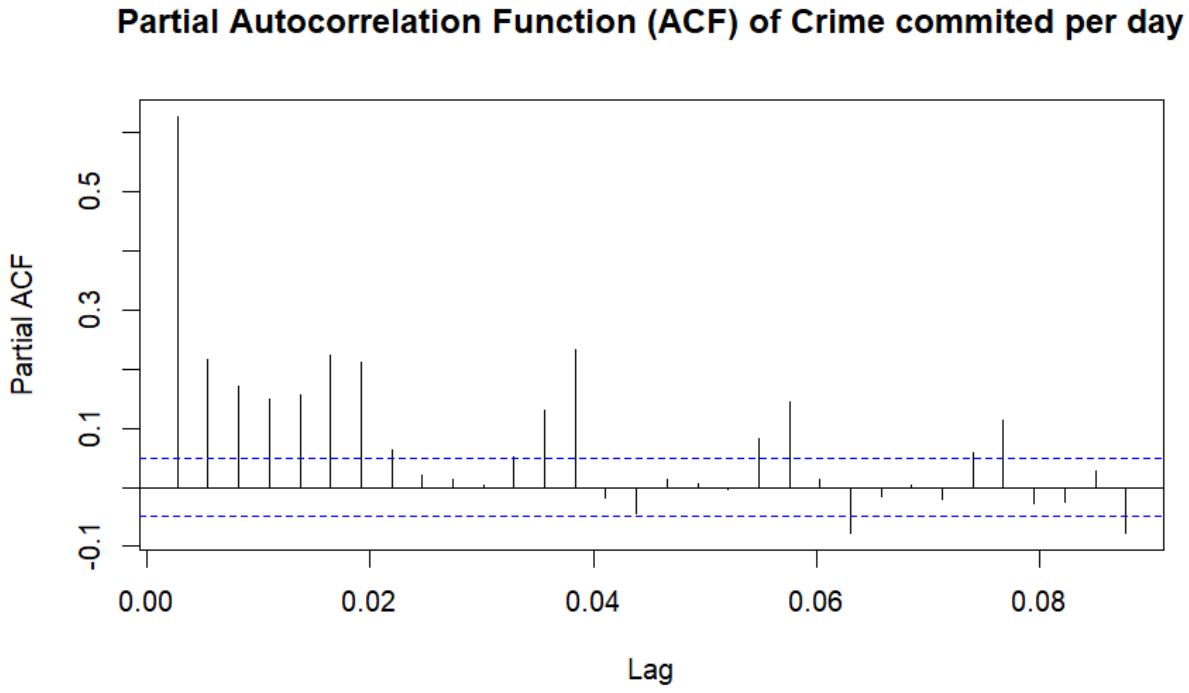
We obtained a time series plot from the filtered data in R to plot the daily number of crimes.







This plot implies that there is high autocorrelation in the data. There is high correlation in every lag of the time series.



From the PACF plots we see that there are at least 5 significant terms. We can expect an AR order of at least 5 in the Time Series.

Augmented Dickey-Fuller Test for Stationarity

The null hypothesis of the test is that the time series has a unit root and is non-stationary. If the p-value is less than or equal to the chosen significance level (commonly 0.05), you reject the null hypothesis and conclude that the time series is stationary.

Augmented Dickey-Fuller (ADF) Test for Stationarity

```
adf_result = adfuller(daily_data['Number_of_Crimes'])

print('ADF Statistic:', adf_result[0])
print('p-value:', adf_result[1])
```

```
ADF Statistic: -2.1786418373882284
p-value: 0.21408834964279555
```

Null hypothesis *H***0** : A unit root exists; the series is non-stationary
Alternate hypothesis *H***1** : The series is stationary
The p-value is **greater than the significance level of 0.05**, so we **fail to reject the null hypothesis.**
Thus, the ADF test suggests the time series is non-stationary.

Data Split

For the purpose of modeling, we split the dataset into 90% training dataset and 10% test dataset. Other combinations such as 80% training data and 20% testing data were also experimented with, but the prior combination yielded the best results as it was able to properly capture the sudden fall in crime rates due to the COVID-19 outbreak and the rise in crime rates once again post COVID-19

Loss Functions

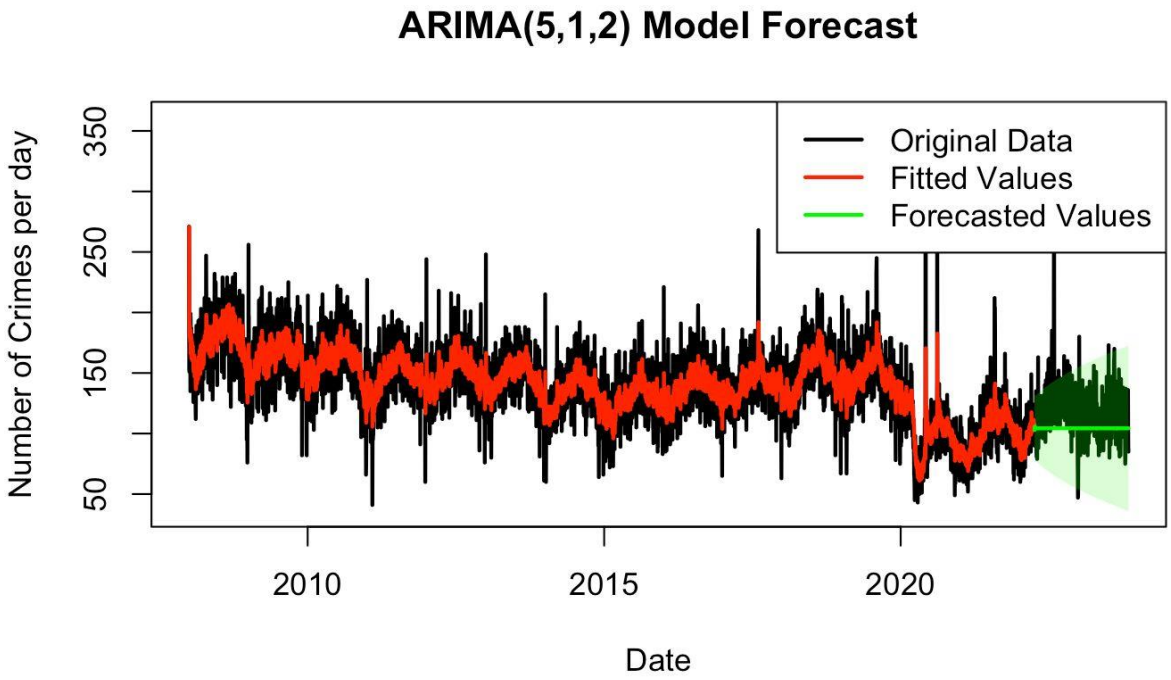
A loss function (also known as a cost function) is a crucial component in the training and optimization of machine learning models. It evaluates the efficiency of a particular model. For this project, we used Root Mean Squared Error(RMSE) and Mean Average Percentage Error(MAPE).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad \text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right|$$

Modeling

1. ARIMA(Autoregressive Integrated Moving Average)

A time series $\{Yt\}$ is said to follow an **integrated autoregressive moving average** model if the dth difference $Wt = \nabla dYt$ is a stationary ARMA process. If $\{Wt\}$ follows an ARMA(p,q) model, we say that $\{Yt\}$ is an ARIMA(p,d,q) process. ARIMA model is used to perform effective forecasting on non-stationary time series.

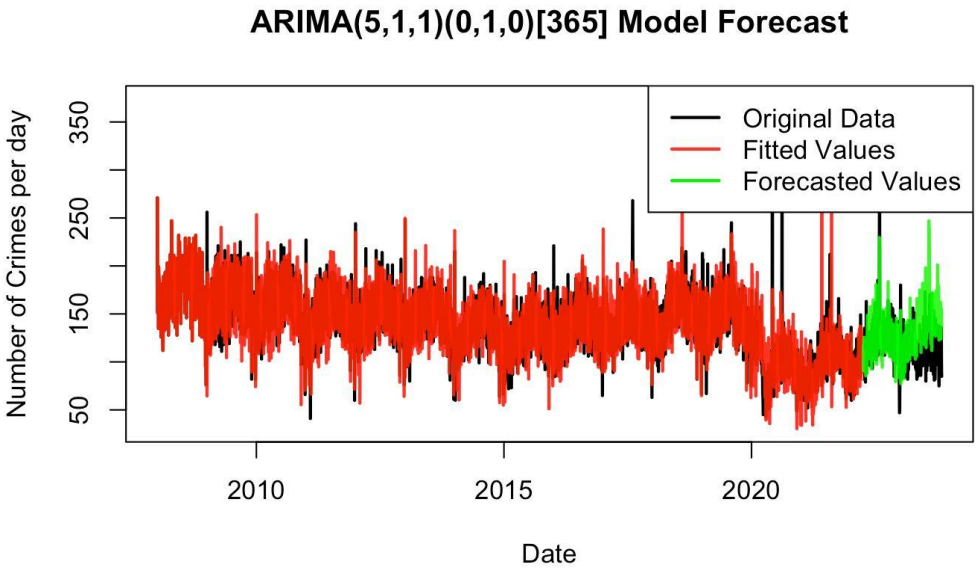


The model was chosen using the Auto-Arima function which chooses the best values of (p,d,q) based on AIC. ARIMA does not consider seasonality component so it does not give accurate predictions. It gives a straight line as forecast.

Observed RMSE:24.43 Observed MAPE: 14.36%

2. Seasonal ARIMA

It is an extension of the ARIMA. Its components are Seasonal Autoregressive (SAR), Seasonal Integrated (SI), and Seasonal Moving Average (SMA) components. Additional parameters are (p, d, q, s) representing the seasonal order in addition to the components of ARIMA. It effectively captures the seasonality of a non-stationary time series.



We noticed that SARIMA effectively captures the seasonality and gives accurate predictions in the beginning but predicts higher crime rates toward the end of the test set.

Observed RMSE:29.30 **Observed MAPE:20.08%**

3. Holt-Winters Model

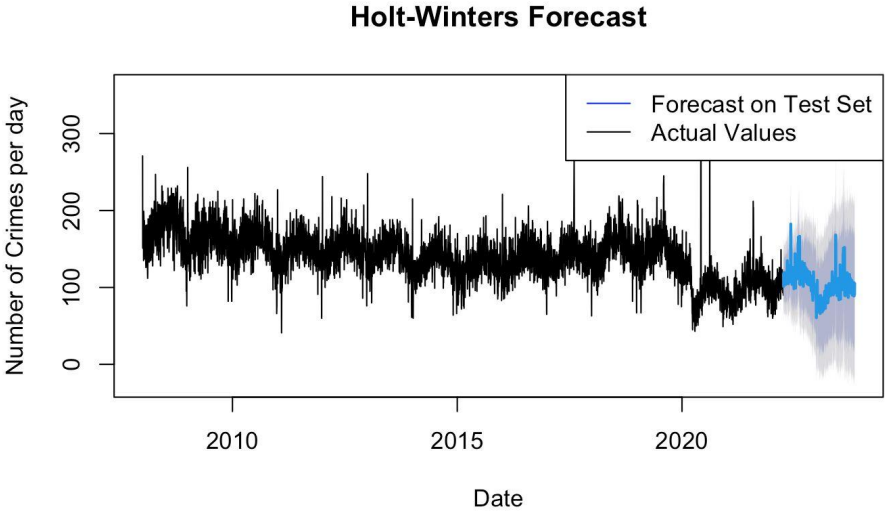
It is also known as the Triple Exponential Smoothing method. It is very useful when data has a seasonal component. It has

Level (l): Represents the average value in the time series.

Trend (b): Represents the average growth or decline in the time series.

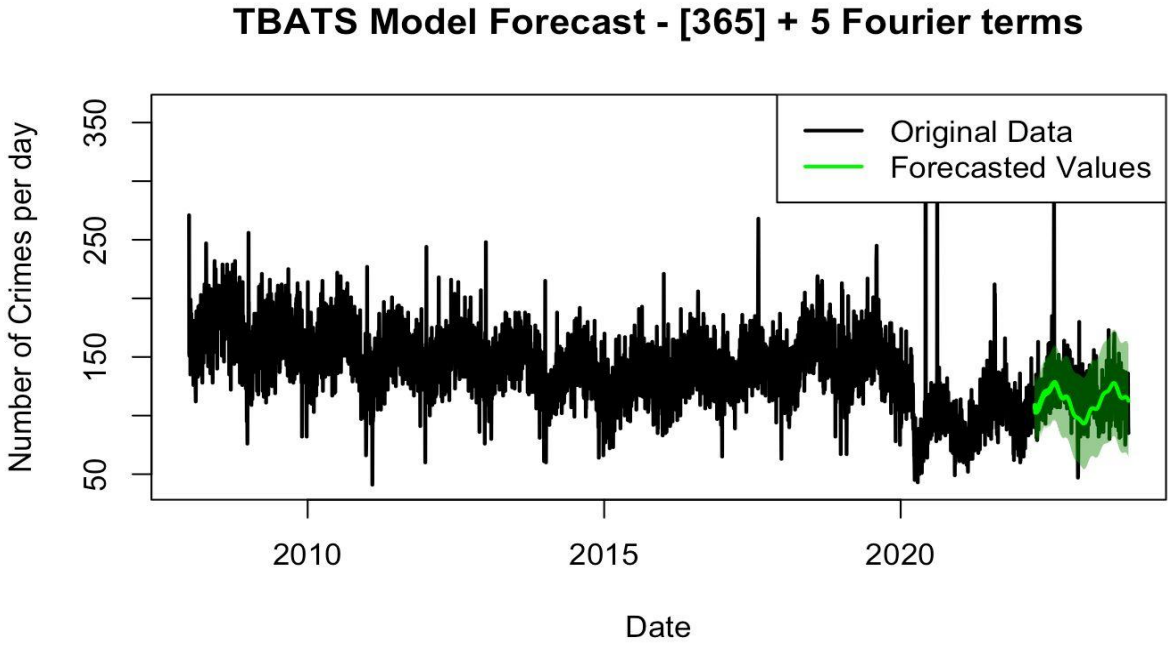
Seasonality (s): Represents the repeating patterns or cycles in the time series.

Holt-Winters uses three smoothing equations to update the level, trend, and seasonality components.



Observed RMSE:26.68 **Observed MAPE:16.66%**

4. **TBATS Model(Trigonometric Seasonal, Box-Cox Transformation, ARMA errors, Trend, and Seasonal components):**
   It is a Robust time series forecasting method that handles multiple seasonalities with **trigonometric functions**. It employs **Box-Cox Transformation** to reduce variance. TBATS uses the trigonometric representation of seasonal components based on the **Fourier series.** The Fourier Series elements follow a sinusoidal form and is used to capture the seasonality component of the Time Series which often follows a similar pattern.



Observed RMSE:20.61 **Observed MAPE:12.53%**

5. **Long Short-Term Memory (LSTM)**
   LSTM, Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to overcome the limitations of traditional RNNs in capturing long-term dependencies in sequential data. LSTMs consist of memory cells, input gates, forget gates and output gates. LTSMs are very useful when handling sequential data.

   For this project, we employed two LSTM Layers of 50 units followed by a dropout layer. One Dense Layer is also added as the output layer. The model is compiled using the Adam optimizer in 30 epochs. We used a Sequential model for creating this model
   A Sequential model is a type of neural network architecture in deep learning, particularly popular in frameworks like Keras. It's called "Sequential" because it allows you to create a neural network model by stacking layers sequentially, one after the other.

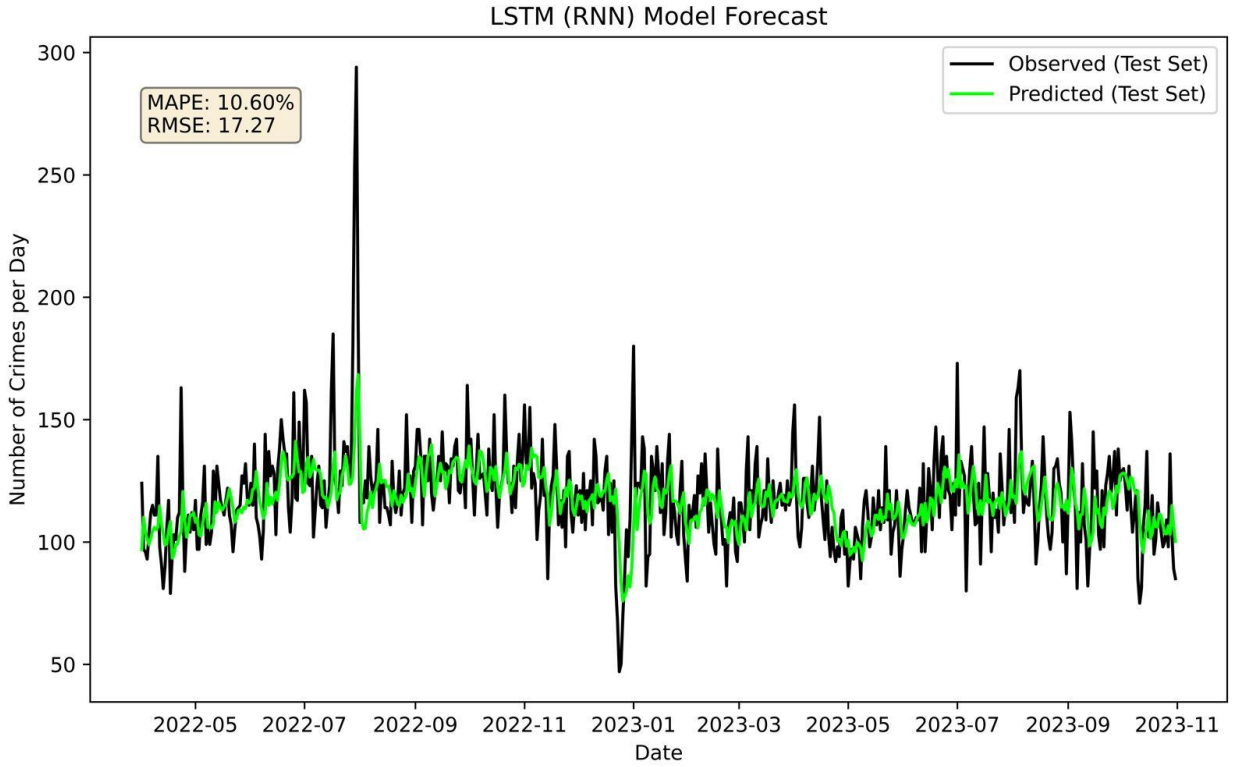
```
# Model creation
model = Sequential()
```

```
model.add(LSTM(units=50, return_sequences=True, input_shape=(X_train.shape[1], 1)))
model.add(Dropout(0.2))
```

```
model.add(LSTM(units=50))
model.add(Dropout(0.2))
```

```
model.add(Dense(1))
model.compile(optimizer='adam', loss='mean_squared_error')
```

```
model.fit(X_train,y_train,epochs=30,batch_size=32)
```



Observed RMSE:17.27 **Observed MAPE:10.60%**

Summary

Model	MAPE	RMSE
ARIMA	14.36%	24.43
SARIMA	20.08%	29.30
Holt-Winters	16.66%	26.68
TBATS	12.53%	20.61
LSTM	10.60%	17.27

After training and evaluating these models on the training and test datasets, we find that the LSTM model outperforms the others in terms of both RMSE and MAPE. The TBATS model also shows strong performance. In contrast, the SARIMA model performs poorly when considering MAPE and RMSE, although it manages to capture the data's seasonality. ARIMA, despite having lower RMSE and MAPE values, generates a constant crime rate in its forecasts, which is unrealistic in the context of real-world data.

CONCLUSION

Model Improvement:

- **Adding More LSTM Layers**: To enhance predictive performance, you can experiment with adding more LSTM layers to the neural network. This increased depth may allow the model to capture more intricate patterns in the crime data.
- **Increasing Complexity of LSTM:** Consider using more complex LSTM variants or architectures, such as stacked LSTM layers, bidirectional LSTM, or attention mechanisms, to further improve the model's ability to learn and generalize from the data.
- **Increasing Granularity:** To provide more detailed insights, you can extend the models to predict the number of specific types of crimes individually. This would allow for more targeted law enforcement strategies.

Future Scope:

- **Operational Use by Chicago Police**: Deploy the improved models for real-time crime forecasting and planning by the Chicago Police Department. Law enforcement agencies can use these predictions to allocate resources more effectively and proactively address crime hotspots.
- **Resource Allocation**: Based on the model's forecasts, allocate more patrol cars and officers to areas predicted to have higher crime rates. This proactive approach can help deter criminal activity.
- **Optimal Timing for Police Checkpoints**: Utilize the forecasts to determine the optimal timing for police checkpoints or surveillance operations, allowing law enforcement to be more strategic in crime prevention efforts.
- **Public Awareness via Heatmaps**: Develop a user-friendly interface, such as a crime heatmap, that provides the public with information about potential crime hotspots and times. This can empower individuals to take precautions and be more mindful.
- **Feedback Loop:** Implement a feedback mechanism where law enforcement can validate the model's predictions and continuously improve its accuracy based on real-world outcomes.
- **Integration with Other Data Sources**: Consider incorporating additional data sources, such as weather data, social media trends, or economic indicators, to enhance the model's predictive capabilities.
- **Emergency Response**: Extend the use of forecasts beyond just law enforcement. Emergency services and first responders can also benefit from crime forecasts to optimize their response times and resource allocation during emergencies.

  By continually improving and expanding the scope of these models, law enforcement agencies can work more efficiently, reduce crime rates, and enhance public safety while being mindful of ethical considerations and community involvement.

REFERENCES

- https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data - Official website for accessing the Chicago Crime Dataset
- https://github.com/mayukhsen13/Chicago-Crime-Analysis-16-954-596-03-Project - Our Github Repository where we created all plots and relevant codes.
- https://otexts.com/fpp2/holt-winters.html - Information and theory of Holt-Winters Model
- https://rpubs.com/chenx/tbats_notes - Information on TBATS model
- Yong Yu, Xiaosheng Si, Changhua Hu, Jianxun Zhang; A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput* 2019; 31 (7): 1235–1270. - Research Paper to develop an understanding of LSTM
- Jian Cao, Zhi Li, Jian Li: Financial time series forecasting model based on CEEMDAN and LSTM - Publication to analyse the use of LSTM in Time Series Forecasting
- https://www.tensorflow.org/tutorials/structured_data/time_series - Tutorial to perform time series analysis in Python
- Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras - MachineLearningMastery.com - Implementation of LSTM in Python with Keras