**Team ID:** 14

**Team Name:** Runtime Terror

**Team Member Names:**

Zack Labus (ztlabus2)

Jeffrey Guamán (jguama4)

Jonathan Kang (jk36)

**Group Project Proposal: Using Chicago Crime Data to Detect When and Where Crime is Highest and Lowest**

The problem we are interested in looking into is whether there is an increasing trend of criminal activity in Chicago, and what the most frequent categories of crimes are. We are also attempting to answer when a crime is the highest and lowest in a year span, and where crime is highest and lowest. By answering the problem, we will get an insight into where crime is higher and lower, and by using that, people can make better-influenced decisions on where to stay or visit around Chicago, and the places to be cautious of. Answering the problem will also give us an insight into when crime is at its highest and lowest, which will positively influence when a person wants to visit the city.

One of the biggest challenges we may face is to clearly define our range of study and how to utilize the variables being provided. As there are over 7 million entries starting from 2001, the file size may be too large for us to work with and we will have to truncate it. Next, since we are working with a lot of locational data, we will have to figure out how we can turn the strings into coordinates such that we can plot our findings on a map. Finally, we will also have to figure out how we can make it automated so that we can update the code and visuals frequently. We will also need to figure out how we can split the workload and share our code without many errors.

We wanted to work on this dataset because several months ago, we received an email regarding increased support from our University to the local police department to help fight crime. This leads us to wonder what the current crime rate in Chicago is compared to the past and how we can visualize the information. A successful final product would entail visuals that will show when and where an increase or decrease rate of crime takes place by location, district, time (monthly, seasonally, etc.), type of crime, etc. Essentially, graphs that can clearly tell the story of the current situation in Chicago. We will be using Python as a tool to statistically analyze civilian safety in cities, by generating reproducible code to analyze and visualize a dataset that is being updated frequently.

Each of the 3 members listed in the team member names contributed equally to the proposal, working in a zoom call discussing what data we wanted to use and how to answer each question listed on the proposal page on Canvas.

Jeffrey Guamán - 33%

Jonathan Kang - 33%

Zack Labus - 33%

Work Cited:

Department, Chicago Police. “Crimes - 2001 to Present: City of Chicago: Data Portal.” *Chicago Data Portal*, 13 Oct. 2022, <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2>.

**Abstract (ADDED)**

The objective of this project is to investigate criminal activity within Chicago. This involves analyzing timely trends and geographical locations for crimes since the year of 2019. Through our models, we aim to get an insight into estimating the likelihood of criminal activity, and by using that, people can make better-influenced decisions on where and when we should avoid certain locations around Chicago. We will use a softmax regression model to predict the type of criminal activity based on time and geographical descriptions. Also, we will compare the differences between Community Areas and Beats to attempt on understanding the types of criminal activity.

**Introduction (ADDED)**

Several months ago, we received an email regarding increased support from our University to the local police department to help fight crime. This leads us to wonder what the current crime rate in Chicago is compared to the past and how we can visualize the information. A successful final product would entail visuals that will show when and where an increase or decrease rate of crime takes place by location, district, time (monthly, seasonally, etc.), type of crime, etc. From this project, we hope to gain more insights into locations with higher crime rates and determine if Chicago does have an increasing trend of crime rates during and after Coronavirus.

As most of our variables are categorical, we will have to use certain clustering classifying methods to develop predictive models. We will be using Python as a tool to statistically analyze civilian safety in cities, by generating reproducible code to analyze and visualize a dataset that is being updated frequently. Essentially, during the exploratory phase, we would like to develop graphs that can give a general insight into the current situation in Chicago. Next, we will develop multiple predictive models to understand the relationship between location and crime types.

**Related Work**

1. Journal - “Crime Rate Inference with Big Data” (ADDED)

This research paper focuses on Crime data between 2001 and 2015 and combines this information with Chicago’s Taxi database to understand the correlation between the geographical location of criminal activity and the corresponding area’s point of interest (such as nightlife, residential, etc.). Similar to our project, this paper looks at Chicago’s crime dataset and performs inferential learning that focuses on the geographical location of where the crime occurred. The researchers use linear regression and negative binomial regression as the inference model and then construct edges and nodes to understand their calculated features. Unlike their approach, we will be focusing on both the geographical location and the time when the crime occurs. Moreover, we will not be utilizing taxi data or POIs when we model.

Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. 2016. Crime Rate Inference with Big Data. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 635–644. https://doi.org/10.1145/2939672.2939736

1. Journal - “Marked point process hotspot maps for homicide and gun crime prediction in Chicago” (ADDED)

The objective of this research was to develop a model that can predict the likelihood of gun crimes in Chicago using datasets of gun crime hotspots and homicides. The EM algorithm was built to estimate the parameters such that the predictions were as accurate as possible. Unlike our given datasets, they focus on the longitude and latitude coordinates of the crimes and generate hotspots around the location. From thereon, they predict the likelihood of gun crimes happening by building a hotspot map for these crimes. We will look at all crimes happening in Chicago and will focus more on the predetermined locations and patrol areas (beats) set by the Chicago Police Department.

Mohler, George. “Marked Point Process Hotspot Maps for Homicide and Gun Crime Prediction in Chicago.” International Journal of Forecasting. Elsevier, April 12, 2014. https://www.sciencedirect.com/science/article/pii/S0169207014000284.

1. Journal - “CRIME ANALYSIS IN CHICAGO CITY” (ADDED)

The objective of this research is to generate an algorithm that can quickly output the most prominent hotspots for criminal activity. This research uses the k-means clustering algorithm to create hotspots of criminal activity around Chicago. Moreover, they use SAT SCAN, a spatial clustering method that allows them to identify the hotspots and match the hotspots to the real, geographical location with the help of Google Maps. Though we may use a similar approach in utilizing a clustering algorithm to identify the most concentrated locations, we will not use SAT SCAN methods and will probably focus on the hotspots for more violent crimes.

Alqahtani, Ayidh & Garima, Ajwani & Alaiad, Ahmad. (2019). Crime Analysis in Chicago City. 166-172. 10.1109/IACS.2019.8809142. https://www.researchgate.net/publication/335361962\_Crime\_Analysis\_in\_Chicago\_City

1. Web - “ Violent Crime in Chicago was Down in Summer 2022 Compared to 2021 - Did Police Safety Plans Help?” (ADDED)

‘News article comparing the violent crime rates in Chicago in the summer for the past 4 years. The article detailed how this year, the number of shootings were down from 59 last year to 55, but the shootings were more deadly, with 9 people perishing this year, compared to 5 last year. The article interviews the police department and there is discussion about violent crimes in the summer in Chicago and how the police strategized to decrease violent crimes..

Hickey, Megan. “Violent Crime in Chicago Was down in Summer 2022 Compared with 2021 -- Did Police Safety Plans Help?” *CBS News*, CBS Interactive, 6 Sept. 2022, https://www.cbsnews.com/chicago/news/violent-crime-in-chicago-was-down-in-summer-2022-compared-with-2021-did-police-safety-plans-help/.

…

1. Web - “Chicago sees a drop in homicides and shootings, but carjackings and other crimes are up from year ago”

New article that explains that homicides and the number of people shot are down in Chicago, and details where the changes in violence have been in Chicago. The article also explains how although homicides and shootings are down, carjackings and burglaries among other crimes are up. The article talks about the possible reasoning behind why there is an increase in these crimes. Finally, the article shows the fatal and non-fatal shootings in Chicago by community area on a map of Chicago.

Boyle, Andy. “Chicago Sees a Drop in Homicides and Shootings, but Carjackings and Other Crimes Are up from Year Ago.” *Times*, Chicago Sun-Times, 4 Apr. 2022, https://chicago.suntimes.com/2022/4/1/23006317/chicago-homicides-shootings-increase-carjackings-crimes-crime-statistics.

1. Research Journal - “Property Crime Specialization in Detroit, Michigan”

Felson et al. (2022) researched whether areas in Detroit had a specialized class of crime activity. To research whether the regions can be grouped by categories of criminal activity, the researchers used pairwise comparisons to compare crime hotspots and the corresponding criminal activity with the highest percentage. Both our research project and Felson's project involves analyzing multiple categorical variables. Unlike how we are planning to use binary coding and develop classification trees, they have done deeper analysis in the EDA, developing multiple tables for percentage comparisons. The project concludes that a hot spot with specialized criminal activity is not necessarily a hotspot for another. The idea of the “law of specialization” applies more to urban criminal analysis than a general scope.

Felson, Marcus, Yanqing Xu, and Shanhe Jiang. “Property Crime Specialization in Detroit, Michigan.” *Journal of Criminal Justice* 82 (June 22, 2022): 101953. https://doi.org/10.1016/j.jcrimjus.2022.101953.

…

**Data (ADDED)**

What type of data is it?

Who collected the data?

How many records and features exist?

What is the update frequency of the data set?

Where is the data set available?

Were any special steps required to import the data?

Show examples from the data. For example:

If the data is in CSV form, show the first 5 observations.

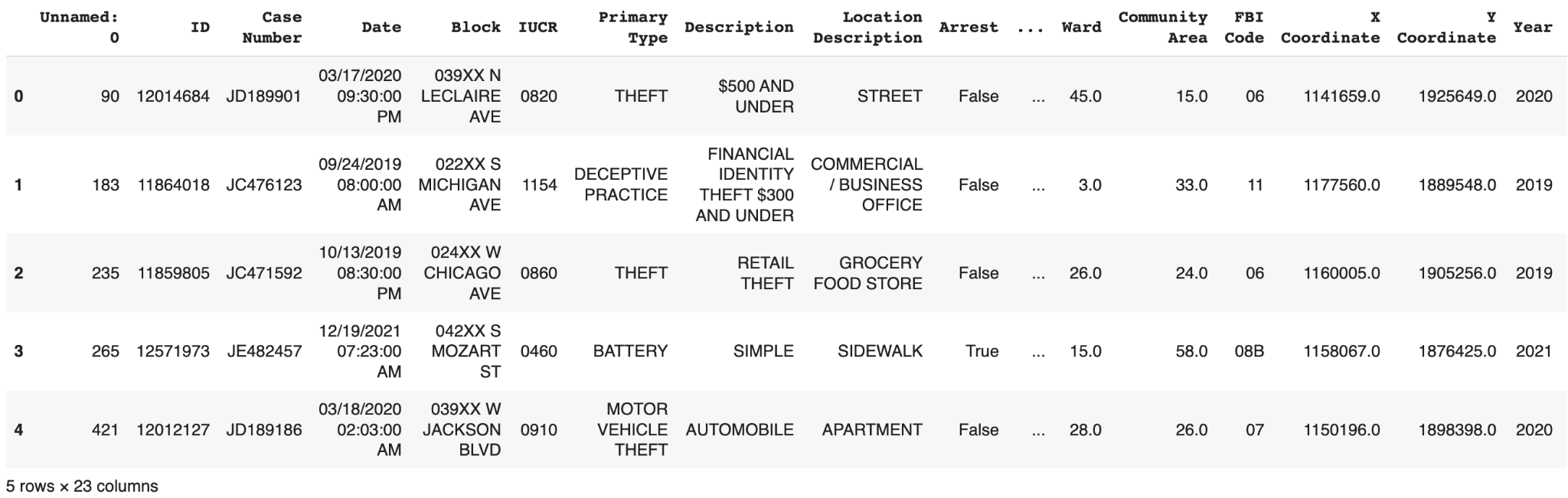
For data in the JSON form, show at least 2 observations.

Please include a reference to where the data set can be found.

This does not count toward the minimum of 5 works cited. (Already included in the citation[1])

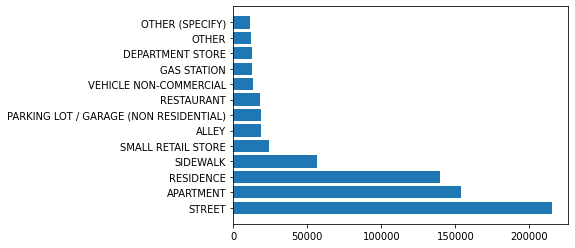
Ensure the code to load the data set is available within the code notebook turned in.

To restate, our dataset contains information about what the criminal activity is, where it occurred, and the time when it happened. The data recorded is a data table containing over 882,000 data entries with 22 columns of data, most of which are categorical and ordinal variables. Data is being gathered by Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The dataset is updated daily, updating the crimes that happened 7 days ago. This dataset can be found in the link “<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2>”. As this dataset contains over 7 million rows, it is extremely time extensive to download and read the CSV file. Thus, for our research purpose, we will focus on analyzing crimes between 2019 and the present day. By doing so can we officially begin exploring the data. The following figure shows the first 5 observations of some variables in our dataset.

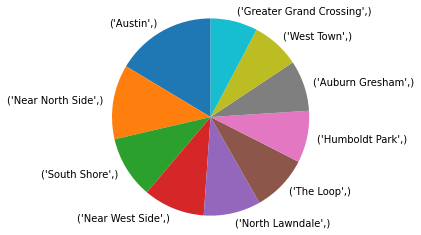


**Exploratory Data Analysis**

Locations and Community Area Top Results: (ADDED)

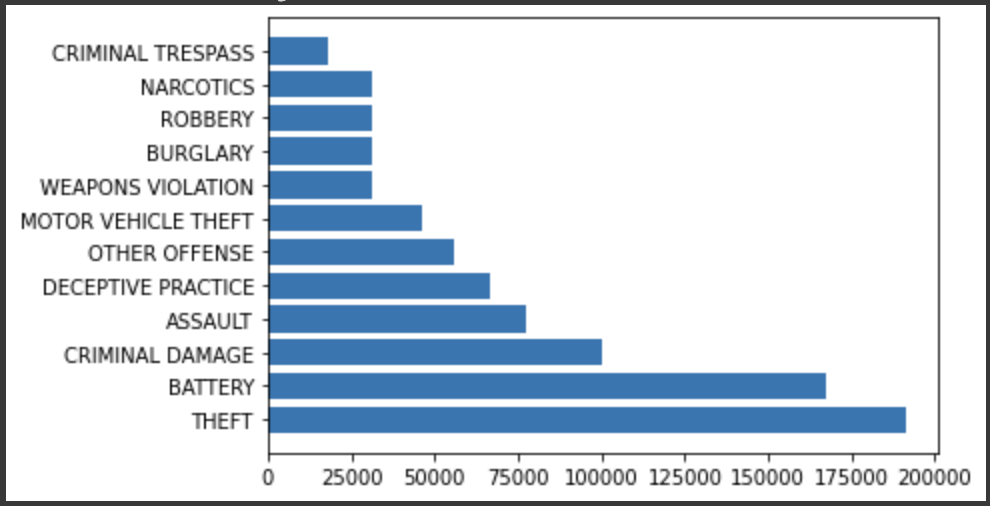


This graph shows the most common types of locations where criminal activity happened. The three most common location descriptions were the street, apartments, and residential areas.

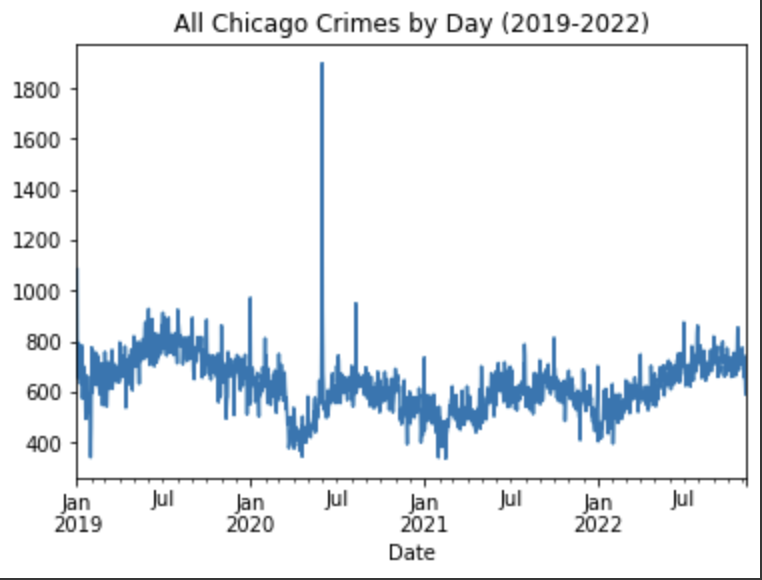
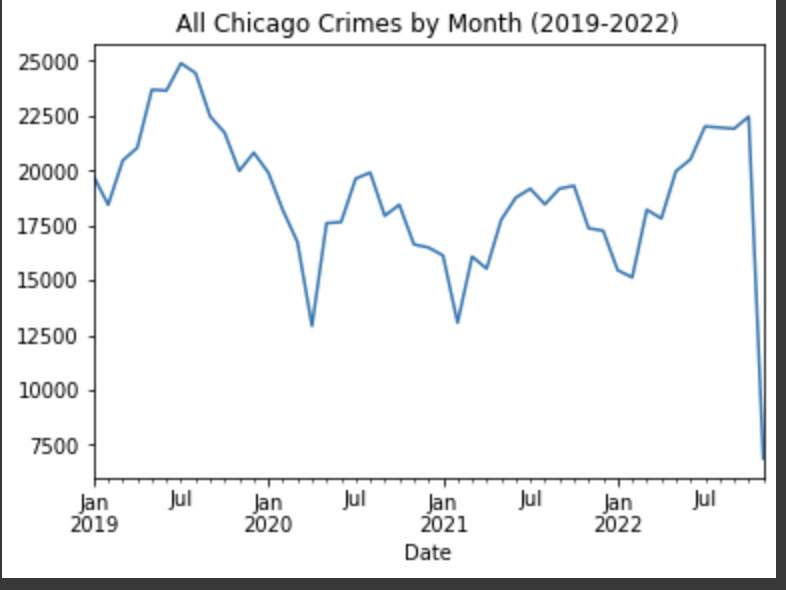


This pie chart shows the 10 community areas with the most criminal activity. Austin has the greatest proportion of cases, then next is the Near North Side, and so on.

Type of Crime: (ADDED)

****

This graph shows the types of crimes committed most frequently, with theft being the most frequent, then battery, and so on.

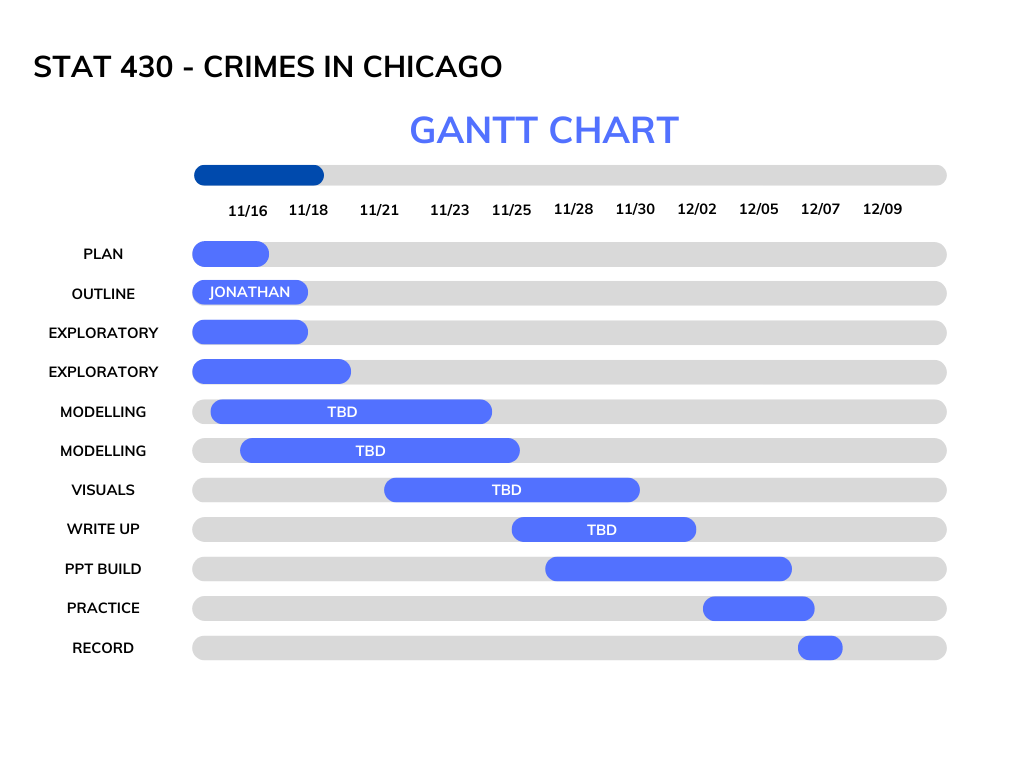


The graphs above show the distribution of when crimes are committed. From the data, we can see that crime typically falls in the winter months, and rises in the summer. We also see from the chart that tracks the daily number of crimes that there was a very large outlier around June of 2020.

**Preliminary Technical Details and Results**

…

**Appendix (ADDED)**

****

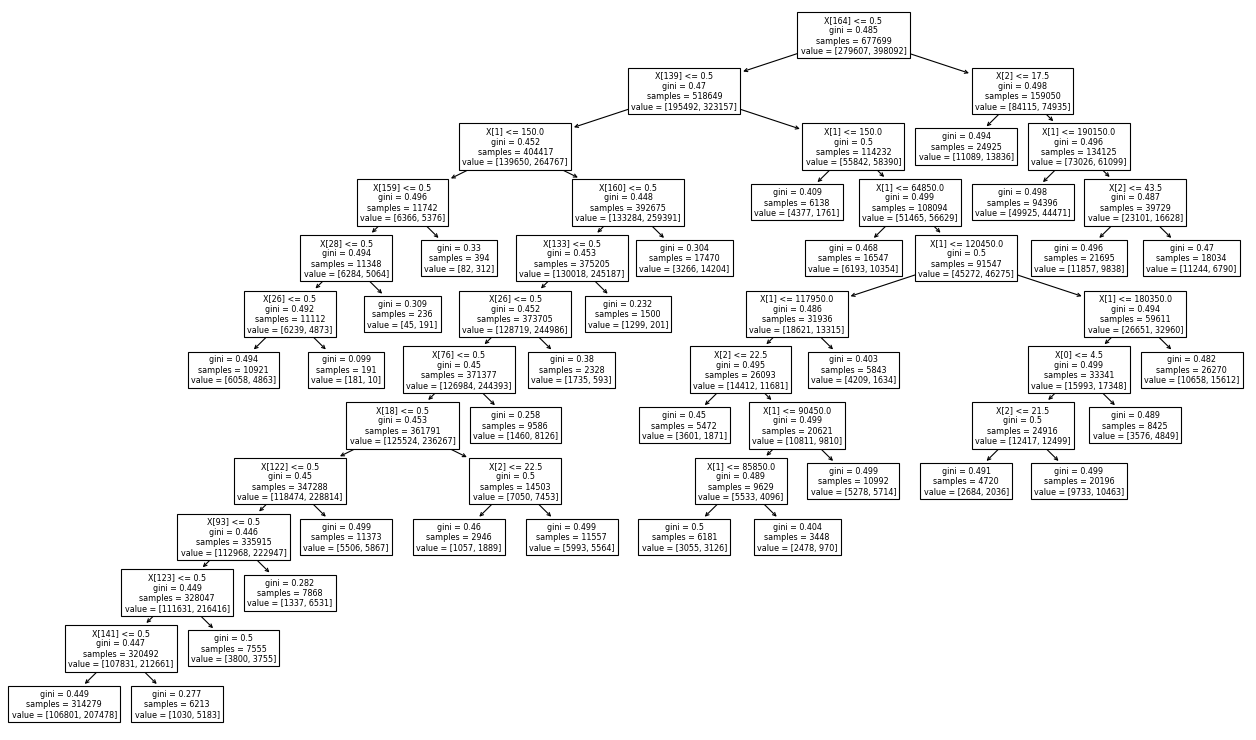
**MODELS:**

* **Why is the approach taken appropriate and correct for the problem?**
* **Are there any alternative approaches that could be taken?** 
  + Make sure to incorporate results from **at least 3 different models**.
* Describe in detail the proposed models.
  + Make sure to have some depth to the explanation for those who might not be familiar with the underlying model.
  + Please use appropriate notation when explaining the model.
* What kinds of metrics will be used to assess performance? Why are these metrics being used?

1. **Decision Tree**

One of our goals was to see if we could classify what category of criminal activity given a time, day of the week, description of the location, and the specific community area in Chicago. Our original goal was to focus on crimes committed against a person and build a learning model to classify the different crimes against a person. However, due to the large and inaccurate model that resulted from this, we decided to focus solely on categorizing whether the crime was against a person or not against a person. Though our study uses a decision tree, it is also possible to use random forests to classify and it may provide different results from the random factor.

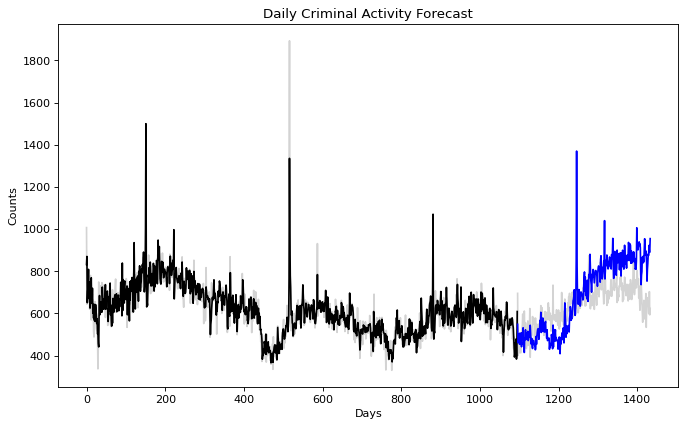
Our decision tree will choose a parameter and a condition that can cleanly split the observed samples, and then repeats this step of splitting to obtain leaves with the minimum impurity until it reaches a limit. We wanted to use a decision tree as the variables we are working with are categorical and it was the simplest to implement and understand if it is valid to classify by considering variables such as entropy. To find the best decision tree and avoid overfitting, we will utilize pruning techniques. Our best model uses the Gini criterion with a max depth of 12 layers and a minimum sample leaf of 20.



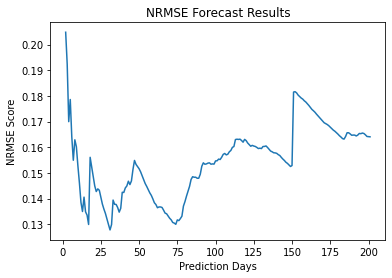
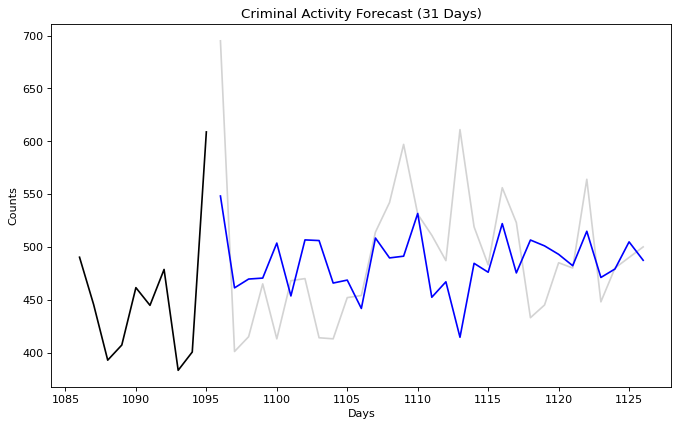
We have used the data from 2019-2021 as our training data, and criminal activity from 2022 as our testing data. By comparing the actual results of 2022 and the predicted results using our decision tree model using sklearn.metrics library, we can see that our model has an accuracy of 61.39%. This accuracy is considered an average machine-learning result. To have improved accuracy, we may modify our search range, excluding certain rows, redefine the crime categories, or instead investigate the beats determined by the Chicago police department instead of community areas.

1. **Time Series**

While we were looking at our EDA for this dataset, we noticed that there seems to be a trend in daily crime rates. With an assumption that the crime rates may increase or decrease given the day of the week, we wanted to generate a time series model that will generate forecasts. Understanding that the trend and seasonal components may be important in this time series model, we will use Holt’s exponential smoothing model. As a note, approaching this analysis with a triple exponential smoothing may be better for its ability to modify the weights and significance of past observations. However, due to the time constraints and modification complexity, we will not use triple exponential smoothing. We will split the dataset into testing and training sets similar to the decision tree’s process. Using 365 days as the seasonal period and damped trends setting, we end up with a relatively decent time series model. The figure below represents the time series model, where black is the trained, blue is the forecast, and grey is the actual data.



We can notice that although the trend is somewhat similar in the first half, the trend heavily deviates after the second half. To determine the best forecast range and model accuracy, we will check the different resulting root mean squared errors and the normalized score. Ultimately, we end up with the conclusion that forecasting for 31 days results in the lowest RMSE and NRMSE of 62.905 and 0.128.



The scores and forecast ability may be enhanced by using other STL decomposition time series methods. However, we are able to conclude that there is perhaps an underlying seasonality trend within this dataset.

3.