Criminal Activities in Chicago

Jonathan Kang[#1] (jk36), Zack Labus[*2] (ztlabus2)

The University of Illinois at Urbana-Champaign
STAT 430 - Python in Data Science Programming - Group Runtime Terror

[1]jk36@illinois.edu

[2]ztlabus2@illinois.edu

***Abstract*— 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 decision tree model to predict the type of criminal activity based on time and geographical descriptions and use time series to forecast future criminal activity counts. Moreover, we will also use clustering algorithms and attempt to identify locations with the most violent cases.**

***Keywords*— Chicago Crime Rates, Data Analysis, Time Series, Decision Tree**

I. Introduction

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. For this research project, we will be working with Chicago's public criminal activity dataset [1], specifically between the years of 2019 and the present day. 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.

For our modeling and analysis, we will first build a decision tree model to classify the different criminal activities. We are wondering if we are given a time, location, and type of location, how accurately can we predict the resulting type of crime? The next model is a time series model built using exponential smoothing methods. This method allows us to build the time series by analyzing the trends and seasonality from past observations during training. We are looking to see if it is possible to forecast the number of crimes that will occur.

II. Related Works

As similar projects have been done by many other scholars, we will mention several similar kinds of research that have been conducted and how ours will differ:

A. Academic Journal - "Crime Rate Inference with Big Data"

This research, conducted by Hongjian Wang and others, 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. [2]

B. Academic Journal - "Marked point process hotspot maps for homicide and gun crime prediction in Chicago"

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. [3]

C. Academic Journal - "Crime Analysis in Chicago City"

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. [4]

D. Website - "Violent Crime in Chicago was Down in Summer 2022 Compared to 2021 - Did Police Safety Plans Help?"

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 was 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 a discussion about violent crimes in the summer in Chicago and how the police strategized to decrease violent crimes. In our research, we will gain further insight into the number of crimes in different areas. [5]

E. Website - "Chicago sees a drop in homicides and shootings, but carjackings and other crimes are up from a year ago"

This news article 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 shows the fatal and non-fatal shootings in Chicago by community area on a map of Chicago. Our research objective will focus beyond shootings, also including crimes that do not involve human injury. [6]

F. Academic 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 involve 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 hotspot 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. [7]

G. Academic Journal - "A Proposed Framework for Analyzing Crime Data Set Using Decision Tree and Simple K-Means Mining Algorithms"

The objective of this research was to discover patterns and trends and make forecasts of crimes by grouping the crimes according to type, location, time, and other attributes. We are interested in this paper because it presents a proposed framework for data analysis of crime data using decision tree algorithms as well as a simple K-means algorithm. After all, we will be attempting to use both of these methods in our analysis. This paper gives insights into the methods used and can provide help if needed. [8]

III. Data

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.[1] As this dataset contains over 7.5 million rows, it is extremely time extensive to read and analyze the entire dataset. Thus, for our research purpose, we will focus on analyzing crimes between 2019 and Dec 12, 2022.

	Unnamed: 0	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	...	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year
0	90	12014684	JD188901	03/17/2020 09:30:00 PM	039XX N LECLAIRE AVE	0820	THEFT	$500 AND UNDER	STREET	False	...	45.0	15.0	06	1141659.0	1925649.0	2020
1	183	11864018	JC476123	09/04/2019 08:00:00 AM	022XX S MICHIGAN AVE	1154	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT $300 AND UNDER	COMMERCIAL / BUSINESS OFFICE	False	...	3.0	33.0	11	1177560.0	1889548.0	2019
2	235	11859805	JC471592	10/13/2019 08:30:00 PM	024XX W CHICAGO AVE	0860	THEFT	RETAIL THEFT	GROCERY FOOD STORE	False	...	26.0	24.0	06	1160035.0	1905256.0	2019
3	265	12571973	JE462457	12/19/2021 07:25:00 AM	042XX S MOZART ST	0460	BATTERY	SIMPLE	SIDEWALK	True	...	15.0	58.0	08B	1158087.0	1876425.0	2021
4	421	12012127	JD189186	03/18/2020 02:05:00 AM	039XX W JACKSON BLVD	0910	MOTOR VEHICLE THEFT	AUTOMOBILE	APARTMENT	False	...	28.0	26.0	07	1150196.0	1898398.0	2020

5 rows × 23 columns

Fig. 1 The first 5 observations of our reduced data set for some column variables.

IV. Exploratory Analysis

Our dataset has a total of 899484 recorded entries with 22 variables dated between Jan 01, 2001, and Dec 12, 2022, prior filter.

For our exploratory analysis, we have split our dataset into 3 different sections. Specifically, we will look at the geographical location of the crime, the type of crime being committed, and the time when the crime was done.

A. Geographical Location

This dataset provides us with a description of where the crime happened. This information will be useful in classifying the type of crime that could happen based on the type of location. There are a total of 184 unique location descriptions. The top three types of locations happen in the streets, apartment buildings, then residential areas, with 221177, 157229, and 141773 counts respectively. The average number of crimes per location description is 4866.51 cases.

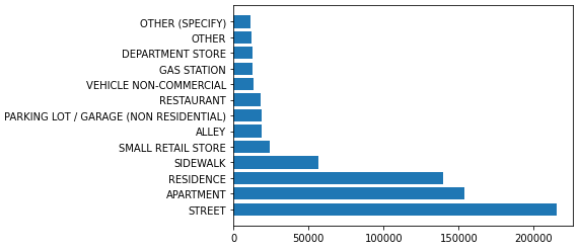

Fig. 2 This horizontal bar chart describes the most common types of locations where criminal activities have occurred in.

Along with the description, the dataset also provides us with the specific community area, district, and even the Beat of where the crime took place. We will utilize geographical data to build a model to determine the severity of the criminal activity. We can see that in Austin, there are a total of 50393 recorded criminal activity, with the second highest being Near North Side with 37783 recorded. Edison Park has the least number of recorded criminal activity, which is 1039.

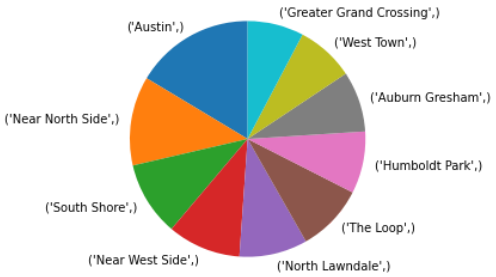

Fig. 3 A pie chart exhibiting the top 10 community areas with the most recorded crimes.

B. Time

The figures below 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 2020.

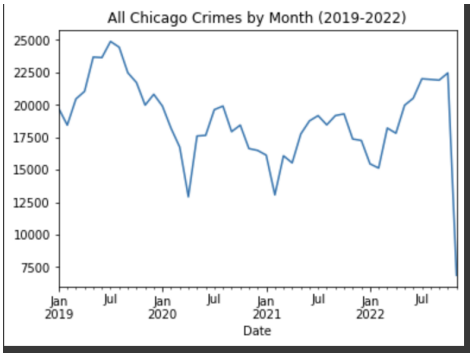

Fig. 4 Time Series Graph of Crimes by Month

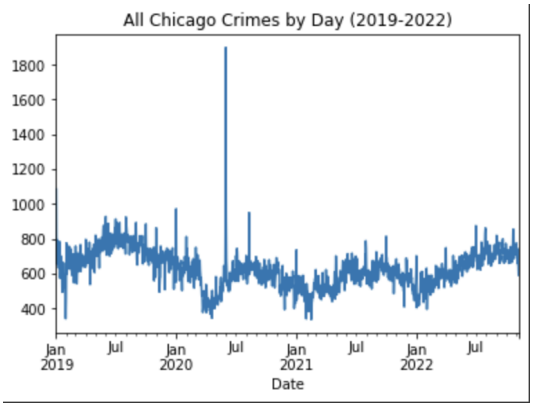

Fig. 5 Time Series Graph of Crimes by Day

Since we are also interested in the frequency of crimes happening depending on the time of the day, we will look at the distribution of times using a violin plot. Most crimes happen at 12 am with 26432 cases, and the next most common is at 12 pm, with 23425 criminal cases.

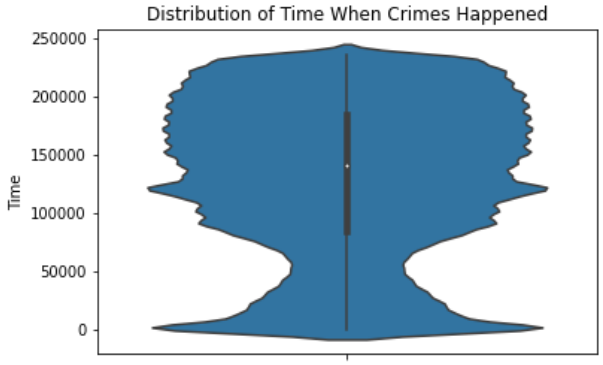

Fig. 6 Violin Plot of the Distribution of Time for Criminal Activity

C. *Types of Criminal Activity*

Each crime has been classified into different categories, and within each category follows a column where an accurate description of the specific crime is provided.

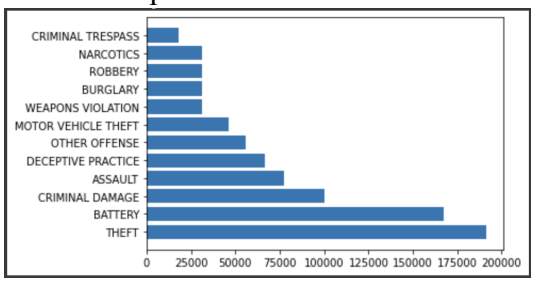

Fig. 7 This figure shows the types of crimes committed most frequently, with theft being the most frequent, then battery, and so on

Overall, we can see that there is not an area with an outstanding number of criminal activities, and most of the crimes done are lighter crimes, which aligns with our original ideas.

V. Models

The next step is to utilize statistical learning methods to analyze this dataset. We have built three different models:

A. *Determining Criminal Activity*

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.

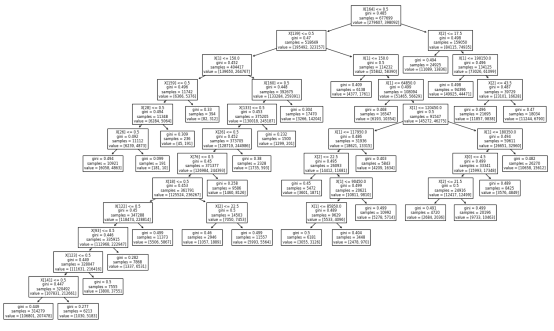

Fig. 8 Resulting Decision Tree Model

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.

B. Predicting Criminal Activity Counts

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.

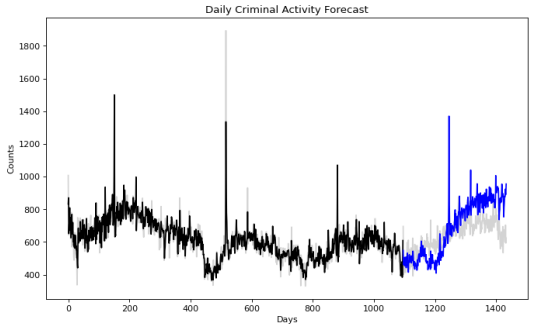

Fig. 9 Time Series Model with Forecast

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.

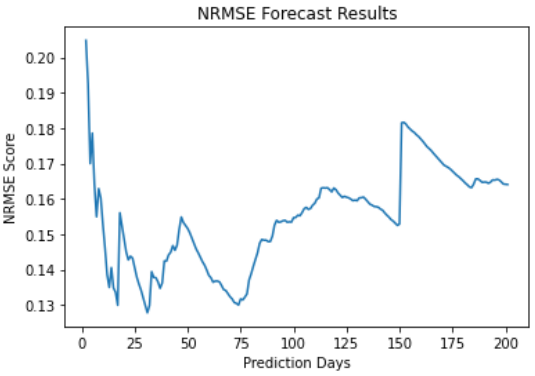

Fig. 10 NRMSE vs Forecast Days Line Plot

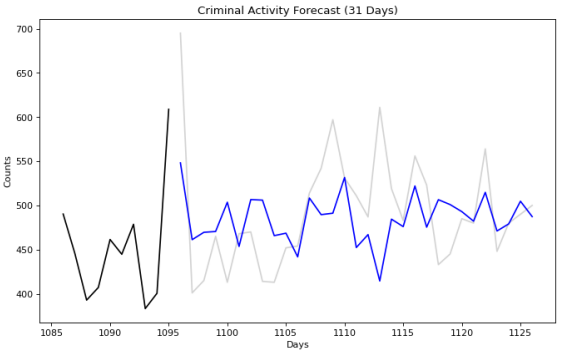

Fig. 11 Forecast Model with the Lowest NRMSE (31 days)

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.

C. Finding Similarities in Community Areas with Clusters

The third model we made was a K-means cluster model. The goal was to group community areas into separate clusters, with those clusters ultimately giving insight into crime frequency in each cluster. We will use the elbow method to determine the optimal K. After determining the optimal K, we will group the data by 'Community Area,' a qualitative variable that describes what area of Chicago the crime took place in. Doing so will give us counts of each crime in each community area. We will normalize that data and from the sklearn

package import Kmeans and PCA. Using the optimal K which was found to be 3, and the normalized data, we fit a model using the Kmeans function. In order to use all the variables in determining the clusters, we had to reduce the dimensionality using the PCA function, and after doing so we received the scatterplot below. The scatterplot alone does not give much insight, but we can use these clusters to determine similarities in community areas in the same group, and looking at the mean crime frequency of each crime in each cluster helps determine which crimes are more or less frequent, and where.

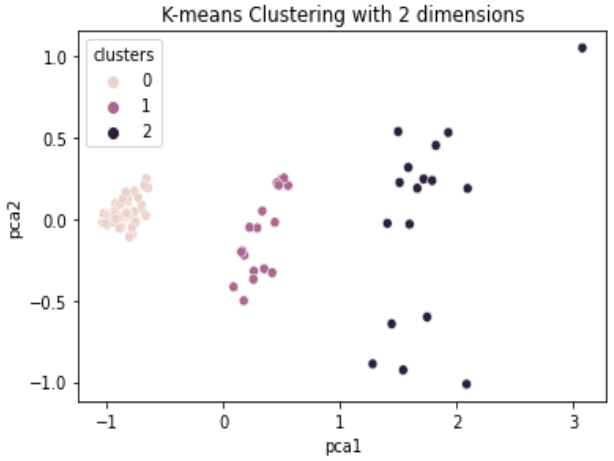

Fig. 12 Scatterplot of Clustered Data

VI. Results and Findings

We created a Shiny app that allows users to customize the entry to the exploratory graphs. Below are two examples of exploratory graphs on the Shiny: one outputs the counts of location description given the range the user is interested in, and the other outputs the daily criminal activity count given the user's date time input.

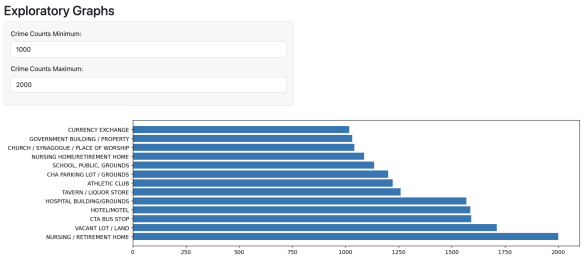

Fig. 13 Shiny App - Location Description Counts Based on User Input

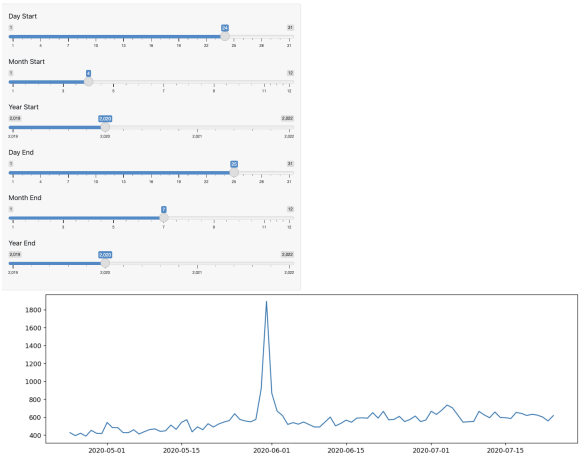

Fig. 14 Shiny App - Line Plot of Crime Counts Based on User Input

After exploring our dataset and building our models, we realized that our predictive models are somewhat adequate in determining a crime and forecasting the number of criminal activities. To demonstrate, we can take one observation as an example.

Given that a crime has taken place on a Wednesday at 1 pm at a Fire Station in Community Area 14, our decision tree model tells us that the criminal activity is most likely a crime against a person. This model can be simulated by running our Shiny App.

Fig. 15 Shiny App - Decision Tree Classification Given User Input

The Shiny App of the model will predict whether the crime is a crime against a person with an accuracy of 61.39%.

VII. Conclusion

In conclusion, the crime rate trends from the EDA show that COVID has actually slightly decreased the number of criminal activities. After 2022 observations, and as shown by the time series forecast model, there is a larger increasing trend of

crime rates in 2022. EDA also tells us that there is no evident hotspot of criminal activity, and we can achieve a somewhat decent prediction model given the time and type of location. Most criminal activity happens either at midnight or at noon time.

Our group has learned that analyzing massive data using typical Python methods was extremely time-consuming. Many other researchers have worked on similar projects, so we had to discuss models and methods that make our research unique. Building models with mainly categorical data was slightly challenging, but this project has taught us the strengths and weaknesses of different methods.

For future projects, we believe that it will be interesting to investigate the different beats defined by police departments and the categorizations of various crime categories. This may be helpful in assigning police departments and specialists more efficiently. For instance, we can have more patrol in areas with higher crimes against people or increased security cameras around certain types of locations.

VIII. Appendix

Throughout our project we have followed our Gantt Chart:

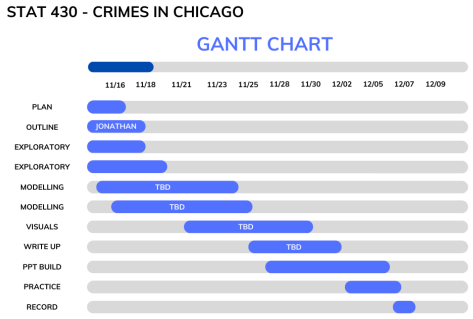

Fig. 16 Gantt chart for splitting the workload

We have split the process into several parts, starting from the exploratory section, which was completed early on the project. Next, to efficiently work on the models, Jonathan will work on two different models while Zack will work on one. Next, Zack will work on and complete the Shiny app while Jonathan will work on formatting the paper and preparing the presentation. Delays were met due to one of our members leaving halfway through the project. Thus, we had to change the scope of our project. This plan allowed us to split the workload evenly between Jonathan and Zack.

In the end, Jonathan worked on half of the exploratory graphs, 2 models, most of the writing for the research paper, and the Shiny App Zack was unable to complete.

Zack has worked on the other half of the exploratory graphs, 1 model, and some of the writing for the research paper.

Jonathan Kang - 60%; Zack - 40%

Works Cited

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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.

[6] 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/.

[7] 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.

[8] Al-Janabi, Kadhim B. Swadi. "A Proposed Framework for Analyzing Crime Data Set Using Decision Tree." Journal of Kufa for Mathematics and Computer vol. 1 (2011): 8-24. https://www.iasj.net/iasj/download/5f87d6c1822ce319