Predicting Crime in Chicago

Aritra Chowdhury
Allie Touchstone
Yash Warty
Serena Wu

Project Goals

Our dataset consists of all crimes in Chicago from 2018 to 2020. We looked at the most common crimes in the city based on different income levels, locations, and anomalies. The patterns in crimes gave a unique insight into how the police should distribute their resources, as well as things civilians should look out for to help avoid becoming a victim of a crime.

Project Importance

In most of the nation's 20 largest local law-enforcement agencies, city and county leaders want funding increases where next year's budgets have already been proposed. While these are concrete steps to address rising crimes, optimizing resource usage is the need of the hour.

Chicago specifically is an interesting case study in analyzing crimes since it is the 3rd most populous city in the United States after New York and Los Angeles. The Chicago PD is also chronically understaffed, which has historically led to a low arrest rate and more unsolved crimes. Thus, it is important to study criminal activities so we can foster better policing, more community involvement and improve overall safety for all stakeholders.

As a focal point of the George Floyd protests, Chicago provided the ideal location for our analysis since it has the perfect distribution of general, holiday, and anomalous events. By specifically analyzing crime data from the city, we were able to discover interesting annual patterns as well as unique data points such as the spike in crimes after George Floyd's murder. Using our insights as a foundation, we hope to provide actionable recommendations for the Chicago PD and similar organizations to effectively combat rising crime.

Exploratory Analysis

The initial data set of crime statistics from the Chicago Data Portal contains 7,367,466 rows. To investigate the most recent occurrences of crime, we extracted crime data from 2018 to 2020. Next, we removed several columns such as case number and FBI code, since they were not helpful in this analysis and were merely adding noise to the data. We then proceeded to remove 73,578 rows of crime records that did not contain X and Y coordinates. Finally, we merged our crime data with socioeconomic indicators from the

Chicago Data Portal by using the community area number. Our final data set consists of 730,532 rows and 23 columns.

In the correlation matrix [Fig 1.1], we found that per capita income is negatively correlated with the other socioeconomic indicators, including hardship index, housing crowded (%), household below poverty (%), aged 16+ unemployed (%), aged 25+ without high school diploma (%), and aged under 18 or over 64 (%). On taking a closer look at the correlation plots, we discovered a couple of trends: the higher per capita income in a community, the lower the unemployment percentage; the higher the percentage of unemployment, the higher the hardship index; the higher the percentage of people who don't have a high school diploma over 25, the lower the per capita income [Fig 1.2 - 1.4].

Additionally, we determined that the average person is most likely to be a victim of crime in a residence or on the street, with most other locations being relatively safe [Fig 2.1]. Another issue specific to Chicago is the dismal percentage of arrests across all crime types, with about a fifth of criminals ever getting arrested [Fig 2.2].

Solution and Insights

We looked at the use of data wrangling to identify crime patterns and predict plausible crime types using Classification models.

We first plotted a trend of crimes for the three years of data [Fig 3.1], followed by the dates and hours with the highest crime count [Fig 3.2 and 3.3]. We see that there was a substantial dip in crimes in March 2020 and another dip in November 2020; this can be attributed to the lockdown enforced due to Covid-19. We can also see that the most crimes happened in the first hour of New Year's Day (mainly due to underage drinking and sale of tobacco products [Fig 4.1]) and the day of the main George Floyd protests (burglary, criminal damage, and public peace violation being the main reasons here [Fig 4.2]). These observations gave us important data points for holidays and anomaly instances of high crimes. The most frequent crimes were related to theft, battery, and criminal damage [Fig 4.3].

For further exploratory data analysis, we created some basic plots summarizing total crimes across types, locations, and years in order to understand these trends better. Using the analysis of top crime data from our analysis above, we created three samples for comparison: General Crimes (2018-20), Midnight of New Year's Day (Holidays), and the George Floyd protest (Anomalies).

Focusing on income-related variables, we found a negative correlation between crimes committed and per capita income. As anticipated, we noticed a tendency for there to be more crimes in low-income areas. To cement our hypothesis further, we plotted a map of the city with the distribution of crime

across districts for all three samples of data [Fig 5.1-5.3]. The distribution of crime remained similar on all 3 times.

When comparing regular days and high-alert days like the George Floyd protests and New Year's Day, we noticed that the probability of being a victim of a crime significantly increases (+75% and +30% respectively). Since the types of crime also change drastically, we built our models and predictions on the top 3 for each sample, as shown in the below table.

All Crimes	Theft, Battery, Criminal Damage
Holidays	Offense Involving Children, Deceptive Practice, Battery
Anomalies	Burglary, Criminal Damage, Public Peace Violation

Table 1: Top 3 entries for each sample

We now went ahead and built classification models to understand the most important features for the above crimes. We built 2 models - Decision Trees [Fig 6] and Logistic Regression. Both models gave us similar accuracies, with the Trees providing an accuracy of 83% and Logistic Regression providing us an accuracy of 83.2%. We moved forward with the latter to analyze feature importance and predict types of crime in the future.

We built a total of 9 models (3 samples; 3 crime types for safest and most dangerous locations for the types of crime above, allowing citizens an assurance or a warning. The analysis of those models gave us the results outlined in the tables below.

All Crimes [Fig 7]

Crime Type	*Safe Areas*	*Dangerous Areas*
Theft	Sidewalks, Residences	Drug Stores, Porches
Battery	Retail Stores, Residences	CHA Apartments, Public Buildings
Damage	Sideways, Grocery Stores	Residence Yards, Streets

Table 2: Public Areas and CHA (Chicago Housing Authority) buildings are more dangerous

Holidays [Fig 8]

Crime Type	*Safe Areas*	*Dangerous Areas*
Underage	Streets, Ward 12	Ward 37, Ward 20
Deception	Ward 2, Streets	Ward 11, District 14
Battery	Residences, Ward 6	Ward 38, District 9

Table 3: Lists of Wards/Districts to avoid during Holidays like New Year's

Anomalies [Fig 9]

Crime Type	*Safe Areas*	*Dangerous Areas*
Burglary	Streets, Sidewalks	Grocery Stores, District 25
Damage	Grocery Stores, Sidewalks	Ward 21, Ward 44
Public Peace	District 25, Residences	Retail Stores, Restaurants

Table 4: We can see that Stores are the most dangerous during anomalies/threats

The above analysis allowed us to propose the following solutions:

The one-word solution to solving Chicago's crime problem; efficiency. The Chicago PD has historically been a short-staffed, overburdened police department. It is time for them to start effectively allocating resources. For example, our data confirmed the stereotype of higher crime rates in lower-income areas. So, the department needs to assign more officers to these areas and run a higher proportion of citizen education programs in these localities to reduce the probability of crimes occurring.

Another interesting trend was the spike in crimes on New Year and during protests. To combat such events, the police must double down on its current strategy of deploying additional officers. To enhance outreach and conserve finances, there could be volunteer corps created to handle non-violent crimes such as underage drinking, tobacco consumption, and drug use allowing licensed officers to focus on only combating more dangerous reports.

Lastly, our data showed that most crimes occur in residential locations. The city of Chicago must engage with stakeholders across demographics and backgrounds to create a robust framework for combating domestic violence, a large proportion of residential crimes. An anonymous helpline must be set up for assistance, and community volunteers should be educated regarding telltale signs of abuse. To combat other residential crimes like burglary/forced entry, the city should formulate standardized guidelines for neighborhood watch programs and outline clear procedures for escalation to law enforcement agencies in case of suspicious activities.

We acknowledge that criminal activity is a complicated and challenging phenomenon to accurately understand and predict. Our hope is for this exploratory analysis and modeling to serve as an effective starting point in initiating a broader change. Communities and local governments must aim to embrace data to solve the longstanding crime problem to improve living standards for their citizens and enhance their overall quality of life.

References

Elinson, Zusha, Dan Frosch and Joshua Jamerson. *Cities Reverse Defunding the Police Amid Rising Crime*. New York, 26 May 2021. Article.

Kaste, Martin. *How Data Analysis Is Driving Policing*. Los Angeles, 25 June 2018. Article.

Mitchell, Chip. *Here's Why Chicago Police Solve So Few Of The City's Murders*. Chicago, 30 October 2019. Article.

Plots

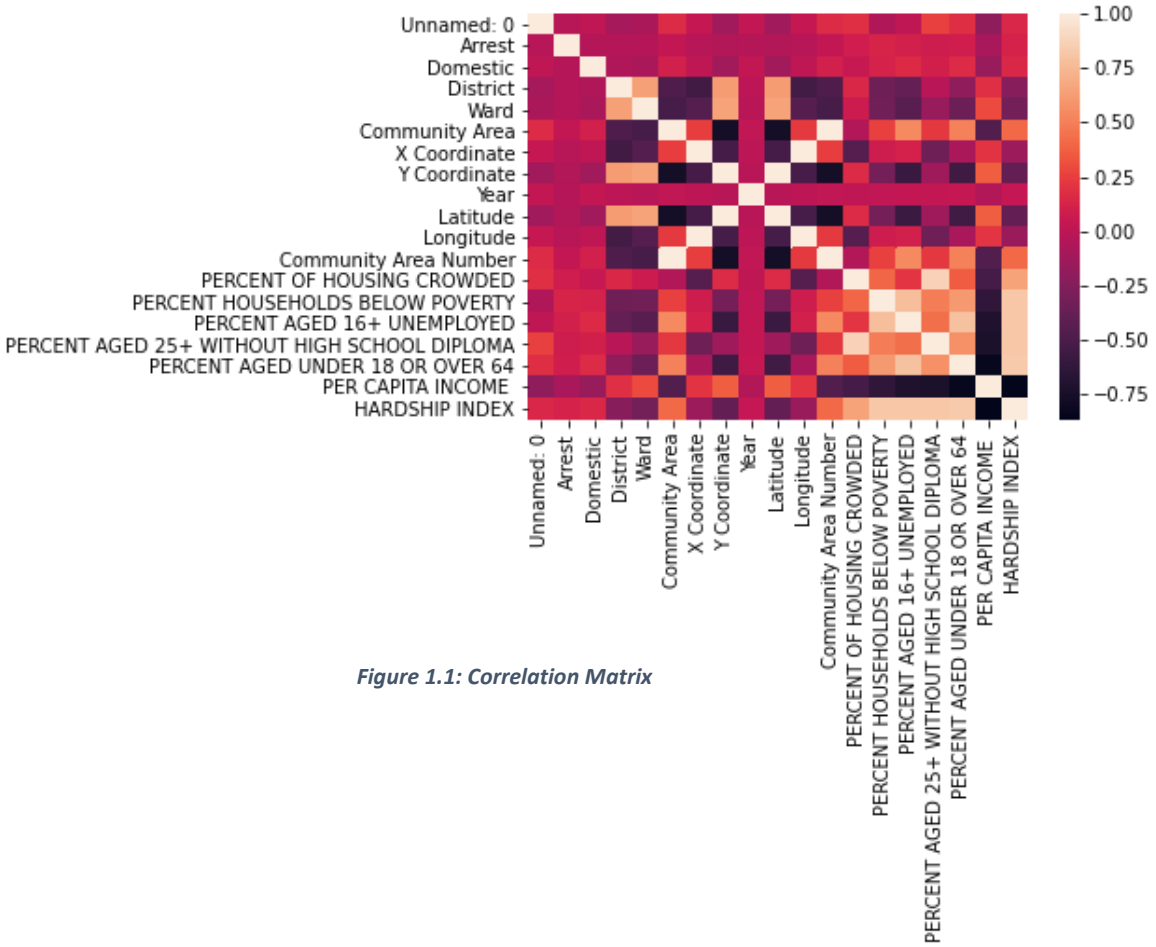

Figure 1.1: Correlation Matrix

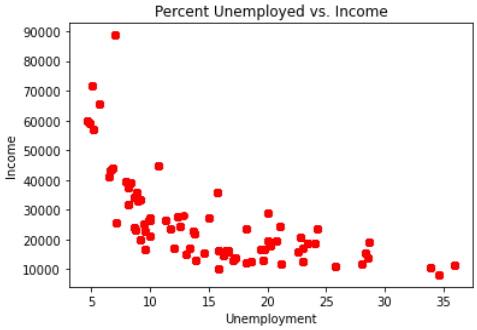

Figure 1.2: Unemployment vs. Income

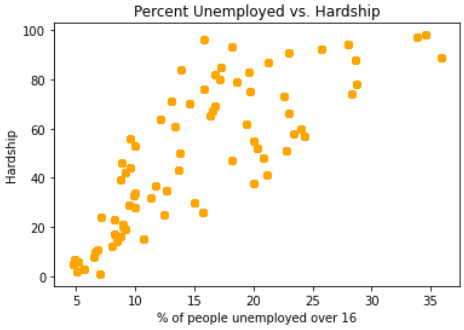

Figure 1.3: Unemployment vs. Hardship

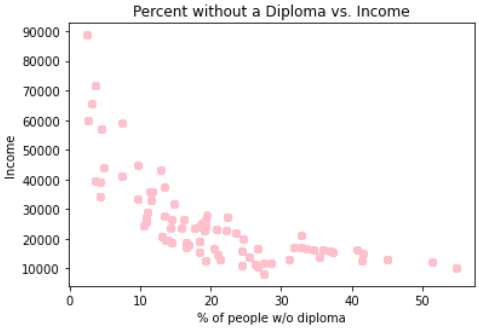

Figure 1.4: Percent with Diploma vs. Income

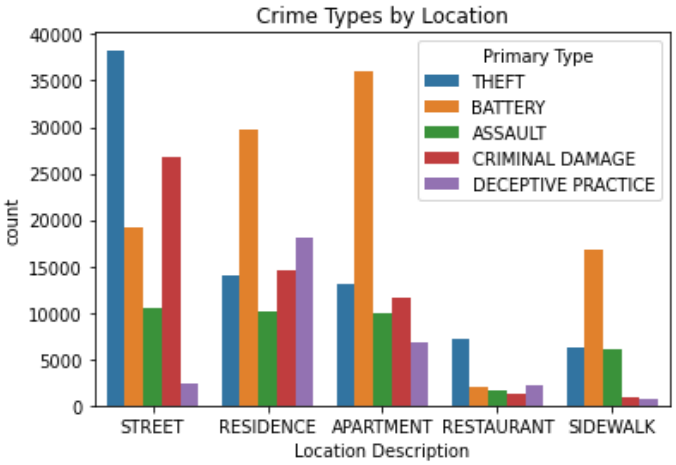

Figure 2.1: Crime Types by Location

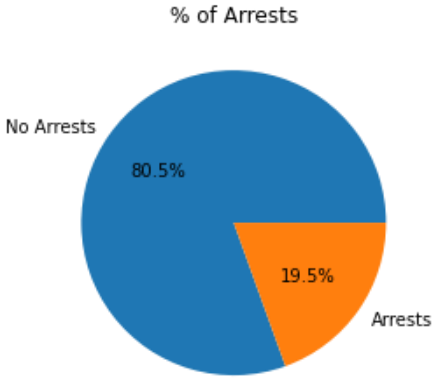

Figure 2.2: Percentage of Arrests for All Crimes

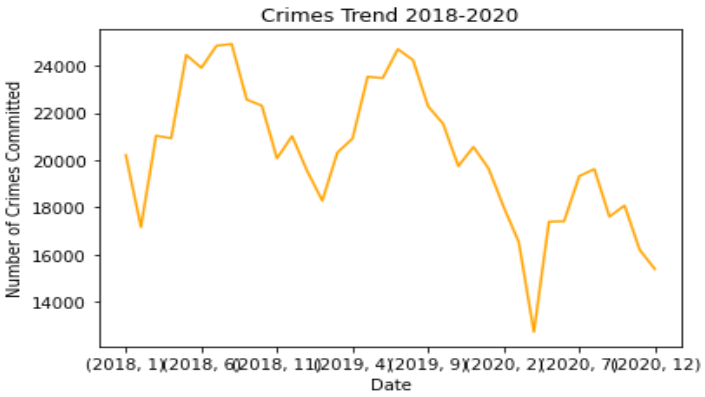

Figure 3.1: Crimes Trend 2018-2020

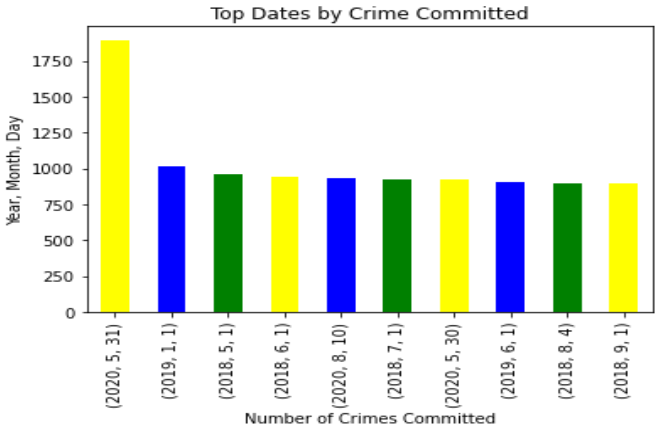

Figure 3.2: Top Dates by Crime Committed

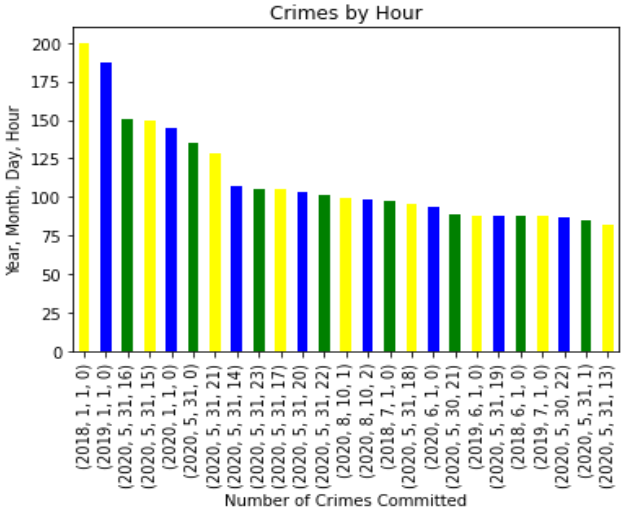

Figure 3.3: Top Hours by Crime Committed

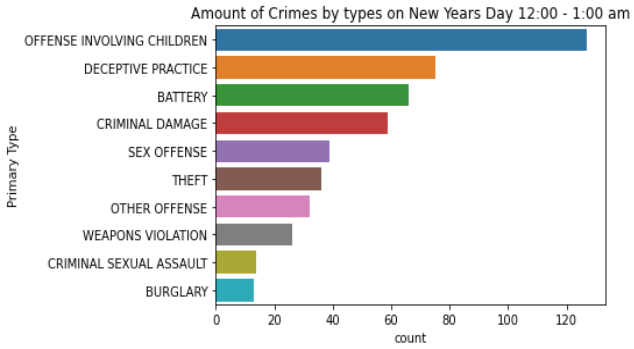

Figure 4.1.1: Crime Split on New Year's

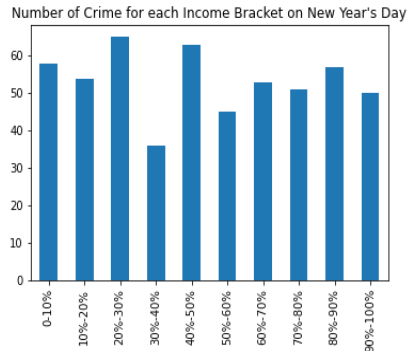

Figure 4.1.2: Crime Totals by Income (New

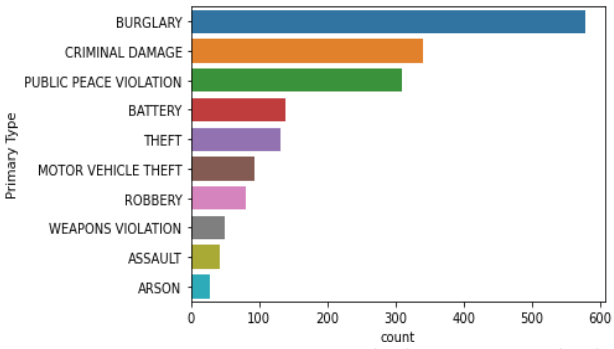

Figure 4.2.1: Crime Split during George Floyd Protests

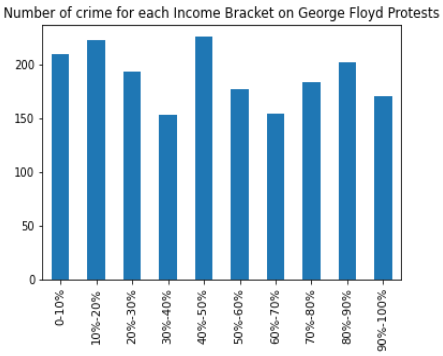

Figure 4.2.2: Crime Totals by Income (George

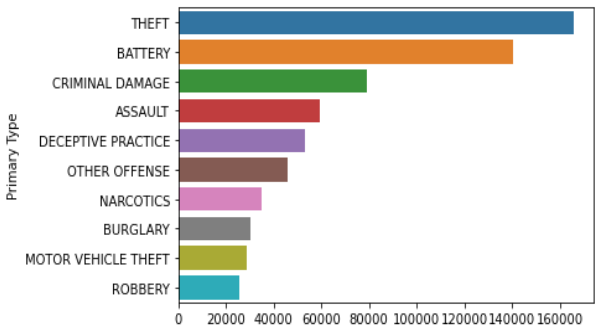

Figure 4.3.1: Crime Split on All Days

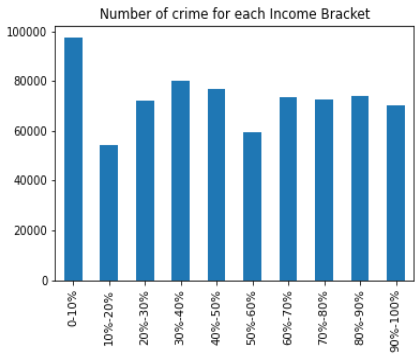

Figure 4.3.2: Crime Split by Income (All

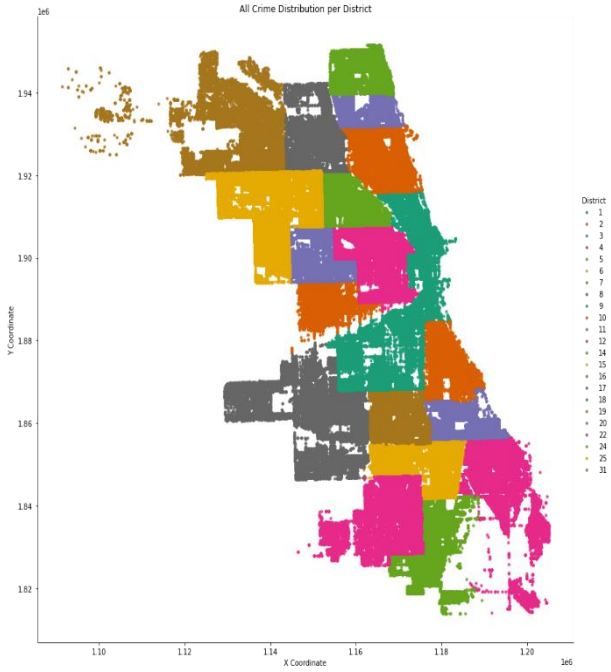

Figure 5.1: All Crimes Distribution by District

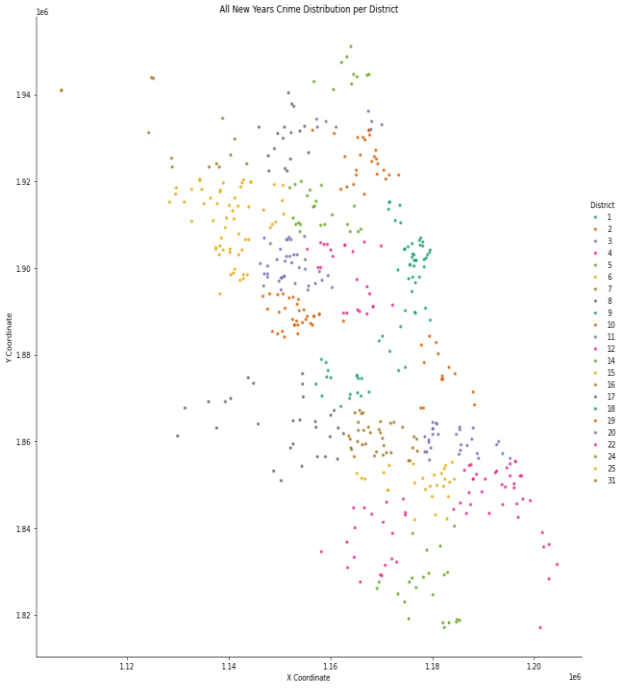

Figure 5.2: All Crimes Distribution by District (New Year's)

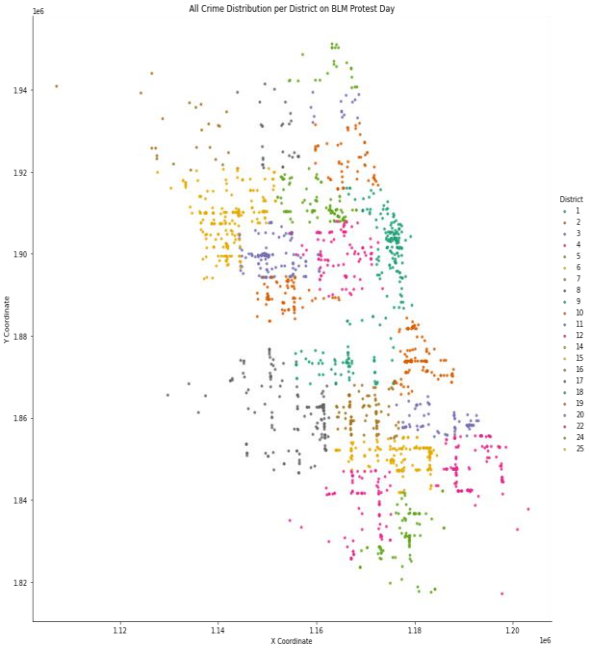

Figure 5.3: All Crimes Distribution by District (George Floyd Protests)

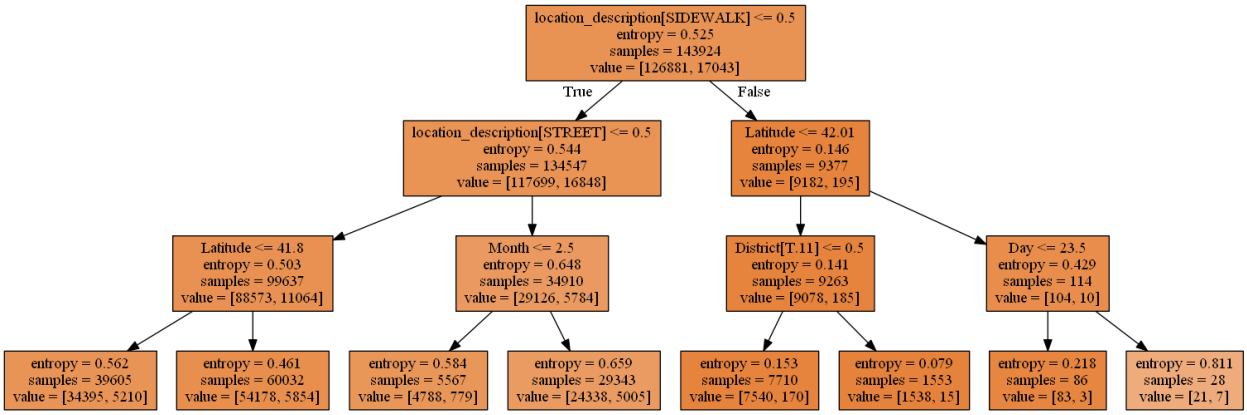

Figure 6: Sample Decision Tree for Criminal Damage

```
Theft Accuracy: 0.8152327221438646

Features with the most negative weights for Theft:

location_description[SIDEWALK]
location_description[RESIDENCE]
location_description[ALLEY]
location_description[APARTMENT]
location_description[OTHER (SPECIFY)]

Features with the most positive weights for Theft:

location_description[DRUG STORE]
location_description[RESIDENCE - PORCH / HALLWAY]
location_description[SMALL RETAIL STORE]
location_description[GROCERY FOOD STORE]
location_description[DEPARTMENT STORE]
```

```
Battery Accuracy: 0.8007554755767391

Features with the most negative weights for Battery:

location_description[SMALL RETAIL STORE]
location_description[RESIDENCE - GARAGE]
location_description[DEPARTMENT STORE]
location_description[COMMERCIAL / BUSINESS OFFICE]
location_description[OTHER (SPECIFY)]

Features with the most positive weights for Battery:

location_description[CHA APARTMENT]
location_description[SCHOOL, PUBLIC, BUILDING]
location_description[RESIDENCE]
location_description[APARTMENT]
location_description[SIDEWALK]
```

```
Criminal Damage Accuracy: 0.8798858680673768

Features with the most negative weights for Criminal Damage:

location_description[SIDEWALK]
location_description[GROCERY FOOD STORE]
location_description[DEPARTMENT STORE]
location_description[RESIDENCE - PORCH / HALLWAY]
location_description[SMALL RETAIL STORE]

Features with the most positive weights for Criminal Damage:

location_description[RESIDENCE - YARD (FRONT / BACK)]
location_description[DRIVEWAY - RESIDENTIAL]
location_description[STREET]
location_description[PARKING LOT / GARAGE (NON RESIDENTIAL)]
location_description[RESIDENCE - GARAGE]
```

Figure 7: Logistic Regression Results for All Crimes

Offense Involving Children Accuracy: 0.8837209302325582

Features with the most negative weights for Offense Involving Children:

location_description[STREET]
Ward[T.12]
Ward[T.2]
location_description[SIDEWALK]
Ward[T.11]

Features with the most positive weights for Offense Involving Children:

Ward[T.37]
Ward[T.20]
location_description[APARTMENT]
District[T.11]
location_description[RESIDENCE]

Deceptive Practice Accuracy: 0.813953488372093

Features with the most negative weights for Deceptive Practice:

Ward[T.2]
location_description[STREET]
District[T.10]
Ward[T.20]
District[T.3]

Features with the most positive weights for Deceptive Practice:

Ward[T.11]
District[T.14]
Ward[T.34]
location_description[MOVIE HOUSE / THEATER]
Ward[T.7]

Battery Accuracy: 0.7441860465116279

Features with the most negative weights for Battery:

location_description[RESIDENCE]
location_description[OTHER (SPECIFY)]
Ward[T.6]
Ward[T.27]
Ward[T.21]

Features with the most positive weights for Battery:

Ward[T.38]
District[T.9]
Ward[T.12]
Ward[T.28]
location_description[HOTEL/MOTEL]

Figure 8: Logistic Regression Results for Crimes on Holidays

```
Burglary Accuracy: 0.7741935483870968

Features with the most negative weights for Burglary:

location_description[STREET]
location_description[SIDEWALK]
location_description[PARKING LOT / GARAGE (NON RESIDENTIAL)]
location_description[RESIDENCE]
District[T.2]

Features with the most positive weights for Burglary:

location_description[OTHER (SPECIFY)]
location_description[GROCERY FOOD STORE]
District[T.25]
location_description[TAVERN / LIQUOR STORE]
location_description[DEPARTMENT STORE]
```

```
Criminal Damage Accuracy: 0.8405017921146953

Features with the most negative weights for Criminal Damage:

location_description[GROCERY FOOD STORE]
location_description[SIDEWALK]
Ward[T.2]
location_description[TAVERN / LIQUOR STORE]
Ward[T.18]

Features with the most positive weights for Criminal Damage:

Ward[T.21]
Ward[T.44]
location_description[VEHICLE NON-COMMERCIAL]
location_description[BAR OR TAVERN]
location_description[BANK]
```

```
Public Peace Violation Accuracy: 0.8673835125448028

Features with the most negative weights for Public Peace Violation:

District[T.25]
location_description[RESIDENCE]
location_description[APARTMENT]
location_description[STREET]
location_description[ALLEY]

Features with the most positive weights for Public Peace Violation:

location_description[SMALL RETAIL STORE]
location_description[RESTAURANT]
location_description[CONVENIENCE STORE]
District[T.11]
location_description[GROCERY FOOD STORE]
```

Figure 9: Logistic Regression Results for Anomalous Crimes