STAT405 Group7 - Crimes in Chicago

AUTHOR
Benjamin Mao, Cecilia Xin, Monica Liu, Jared Boyd

PUBLISHED
April 19, 2024

Abstract

This paper presents a comprehensive study aimed at optimizing police resource allocation in Chicago through a detailed analysis of crime patterns by time and location. Utilizing a primary dataset encompassing all reported crimes in Chicago from 2001 to Feb. 24, 2024, including specifics such as date, time, crime type, and precise location, we integrate a secondary dataset detailing the locations of police stations across the city. Our analysis employs a multifaceted approach, incorporating ggplot to visualize crime distributions, dplyr for data manipulation, and SQL queries to extract relevant information. Our findings reveal that the number of crimes has been decreasing over time, with theft, battery, and criminal damage being the most common crime types. We also identify the street, residences, and apartments as the most frequent crime locations. Furthermore, we pinpoint two areas with high crime densities that lack police stations, suggesting a need for increased police presence in these regions. Our study provides valuable insights for law enforcement agencies seeking to optimize resource allocation and enhance public safety.

Dataset Description

The primary dataset is “Crimes - 2001 to Present”, which reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to the present, minus the most recent seven days. Data is extracted from the Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The dimensions of the dataset are 7998563 rows and 22 columns. Columns include “Case Number”, “Date”, “Primary Type”, “Location”, and much more.

The secondary dataset is “Police_Stations”, which shows the current location of police stations in Chicago. Data is extracted from the Chicago Data Portal. The dimensions of the dataset are 24 rows and 15 columns. The key data we will use from this dataset is the exact location of the police stations through the “Latitude” and “Longitude” columns.

The third dataset is “Chicago Map”, which describes the geological boundaries of areas in Chicago. This data comes from the City of Chicago Data Portal in the form of a shape file. It outlines and labels each of the 25 police districts within Chicago allowing us to easily visualize crime data based on the district.

We then moved the CSV datasets related to crimes and police station locations into an SQLite database, leveraging the RSQLite package for database operations. We first set up the project directory and named the database Crimes_and_Police_Stations, followed by establishing a connection to this newly created SQLite database located in a specified directory. Secondly, we listed existing database tables and loading extension functions to enhance SQLite’s capabilities. We then import crimes_data and police_stations_data datasets into the database. This approach not only facilitates efficient data storage and retrieval within a relational database framework but also capitalizes on the synergies between R's data manipulation strengths and SQLite's reliability and simplicity for local data storage.

Data Cleaning and Preprocessing

For the primary dataset "Crimes - 2001 to Present", several columns are selected for analysis, including "Date", "Primary Type", "Location Description", "Arrest", "Year", "Latitude", and "Longitude". "Date" contains the year, month, date, hour, minute, and second information of when the crime took place. "Primary Type" describes the primary type of crime, such as "Theft", "Battery", "Criminal Damage", etc. "Location Description" describes the location where the crime took place, such as "Street", "Residence", "Apartment", "Sidewalk", etc. "Arrest" represents whether the criminal is arrested or not, with "TRUE" being arrested and "FALSE" being unarrested. "Latitude" and "Longitude" record the specific location of crimes.

Along with this, we will remove any rows with NA values as they will not be useful and may cause problems if not removed.

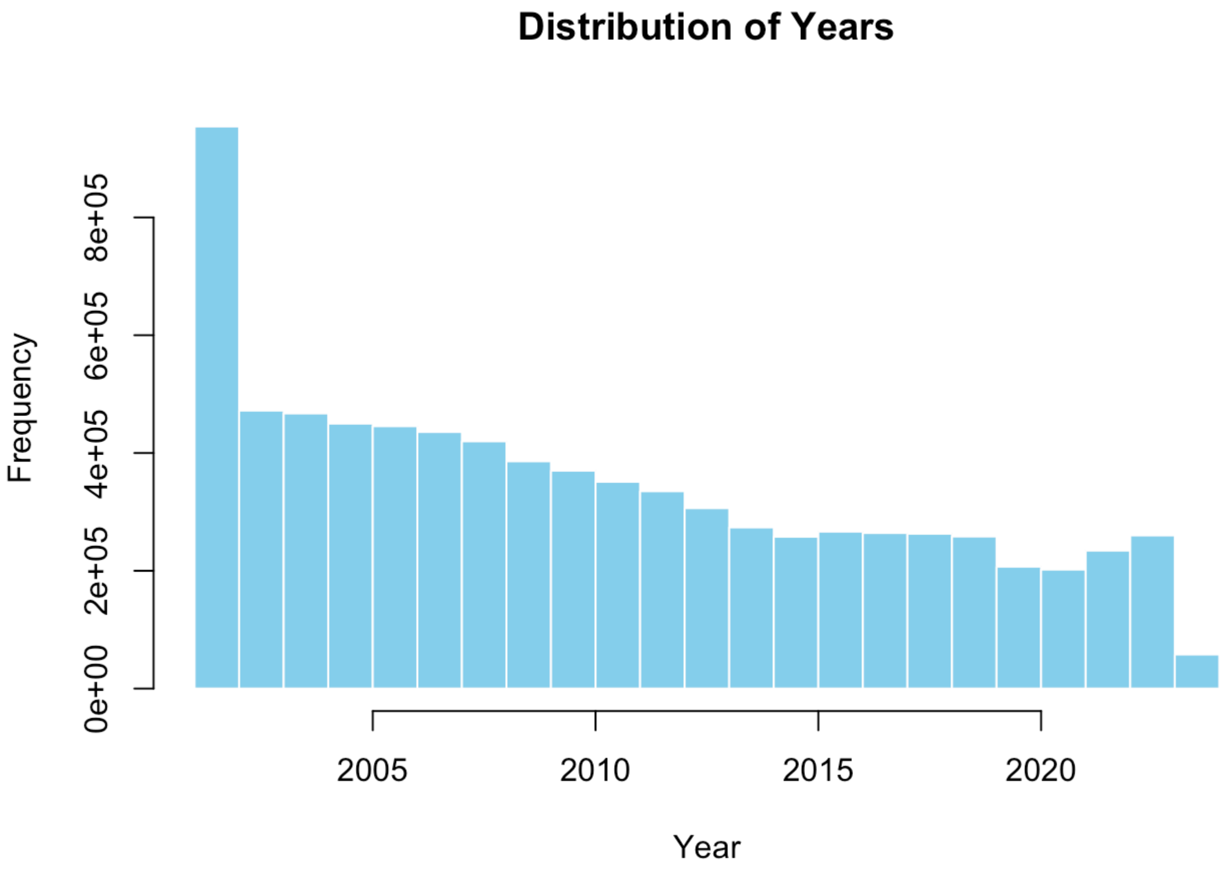
Finally, for simplicity later on the data in column "Date" is disassembled and converted into a standard format. From this, we take the "Month", "DateWithoutTime", and "Hour" all of which we will use in our analysis.

No data cleaning is needed in the second and third data sets as the Police Stations are all needed and consistent data. As for the Chicago Map, we can not clean this data as it is a shape file and we need all data given to outline the districts.

Data Analysis

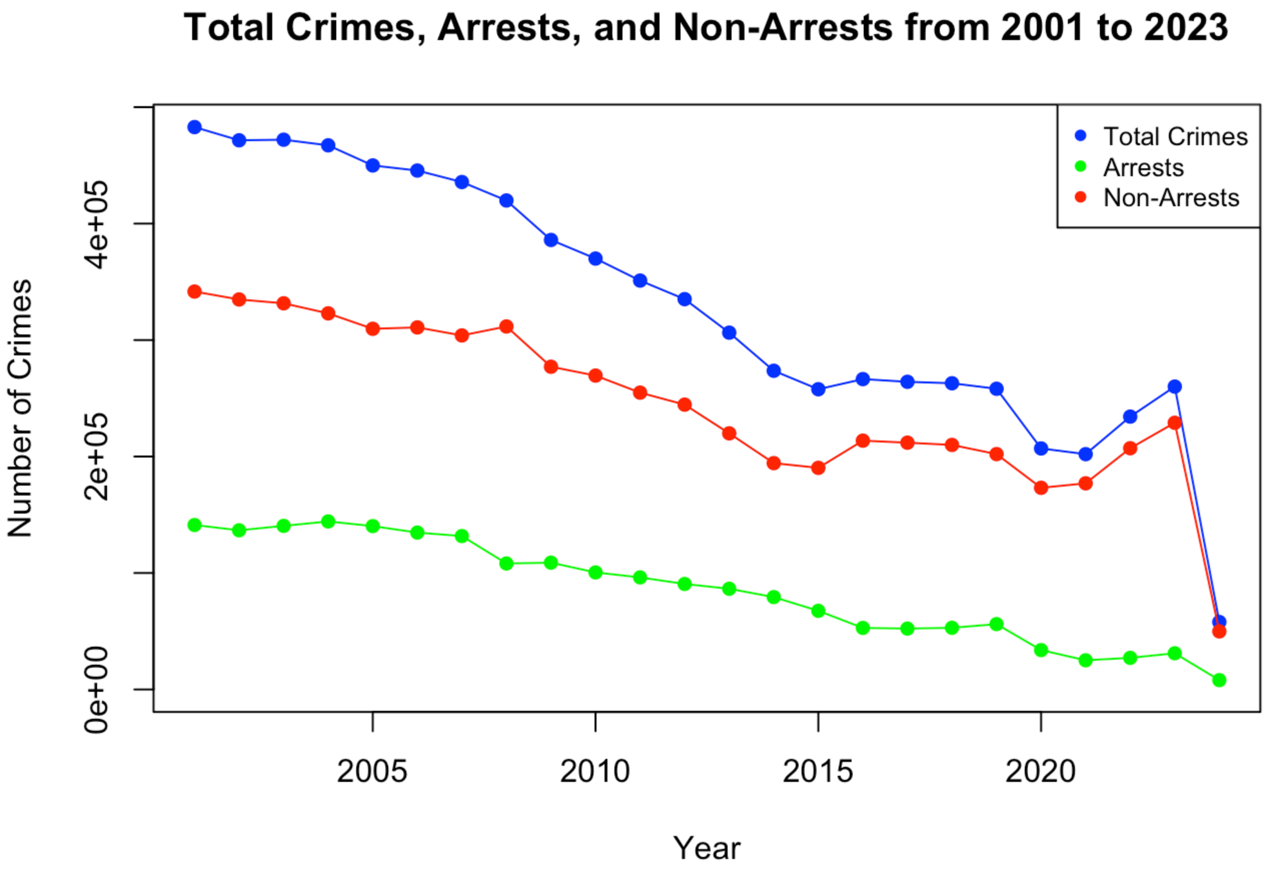
1. When Crimes Are Committed

Throughout this first section, the main focus will be to see when crimes are committed most commonly. This will help us start narrowing down when police stations should be provided more resources. The first graph is a hisogram of the number of crimes based on the year they were committed. This plot will help us see if there are any trends in the number of crimes over time.



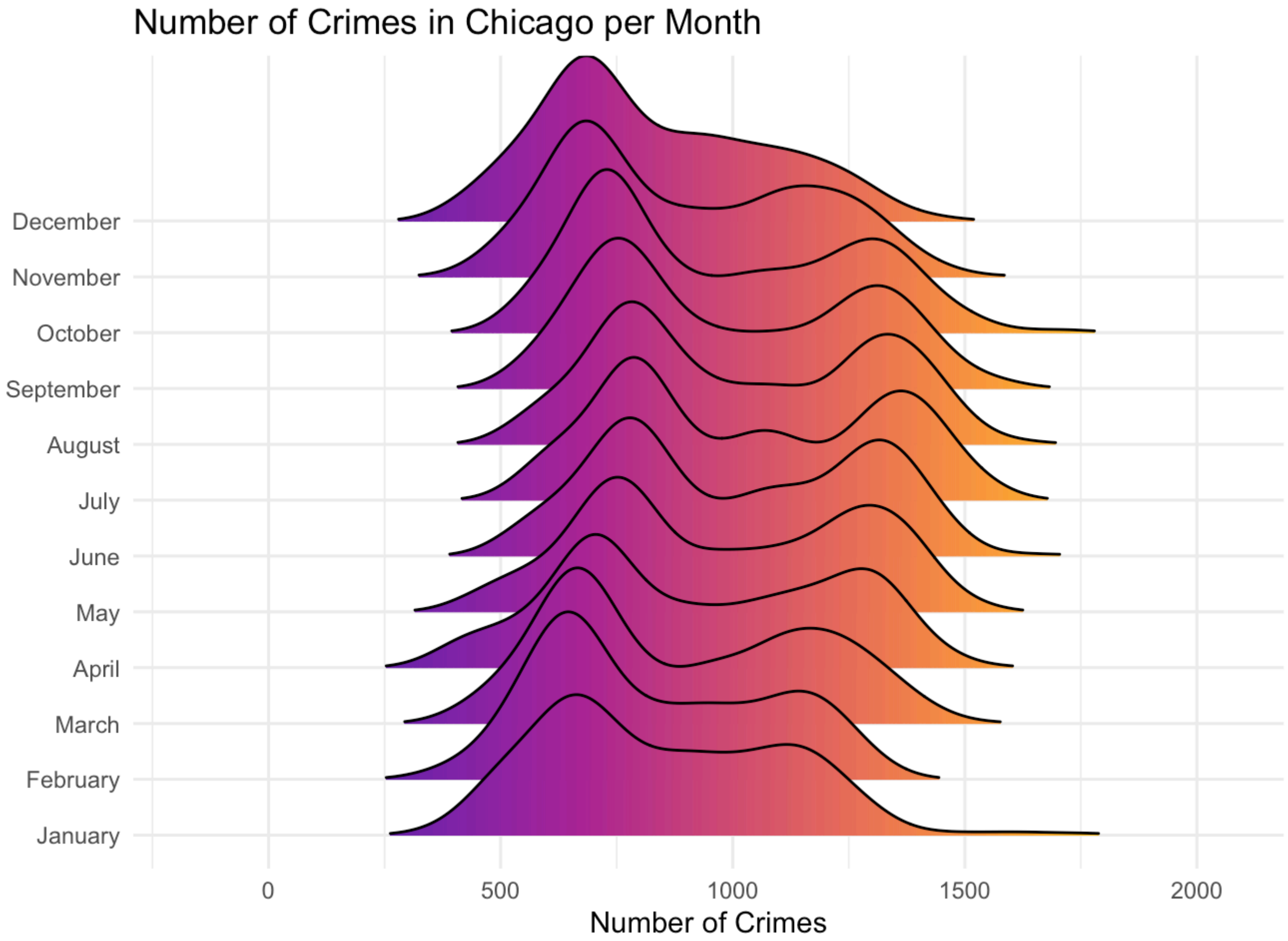
In this histogram, we can see that the number of crimes has been decreasing over time. This is a good sign as it shows that the police are doing a good job in controlling crime. However, we need to look at the number of crimes that resulted in an arrest to see if the police are doing a good job in catching criminals.

Thus, we plot the second graph, which is a line plot of the number of crimes based on the year separated by whether or not they resulted in an arrest:



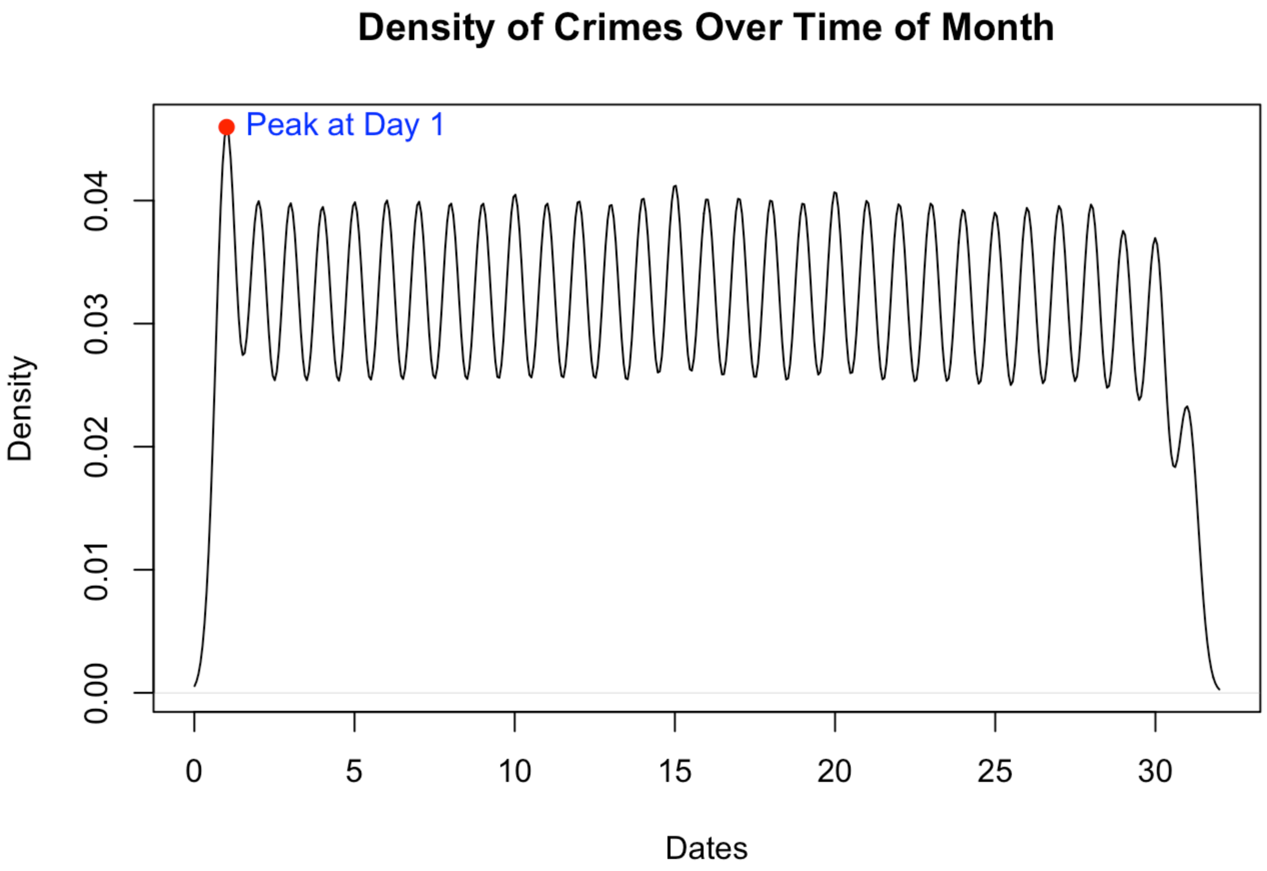
With a LINE PLOT, we analyzed the total number of crimes, the number of crimes arrested, and the number of crimes not arrested over time from 2001 to 2023. The graph indicates a significant decrease in the total number of crimes from 2001 to 2015. The disparity in the slopes of the blue line (Total Crimes) and the green line (Arrests) suggests a declining arrest rate over time. One thing to note is since we are only a small part of the way through 2024 the last data points are not from a full year and can not be compared to the others yet. Additionally, whatever steps Chicago is currently taking against crime, whether that be through the police or social outreach programs it is effective.

The second graph will be the number of crimes in Chicago based on the month they were committed. The goal of this graph is to see if there is any seasonal correlation or a specific month that police attention should increase.



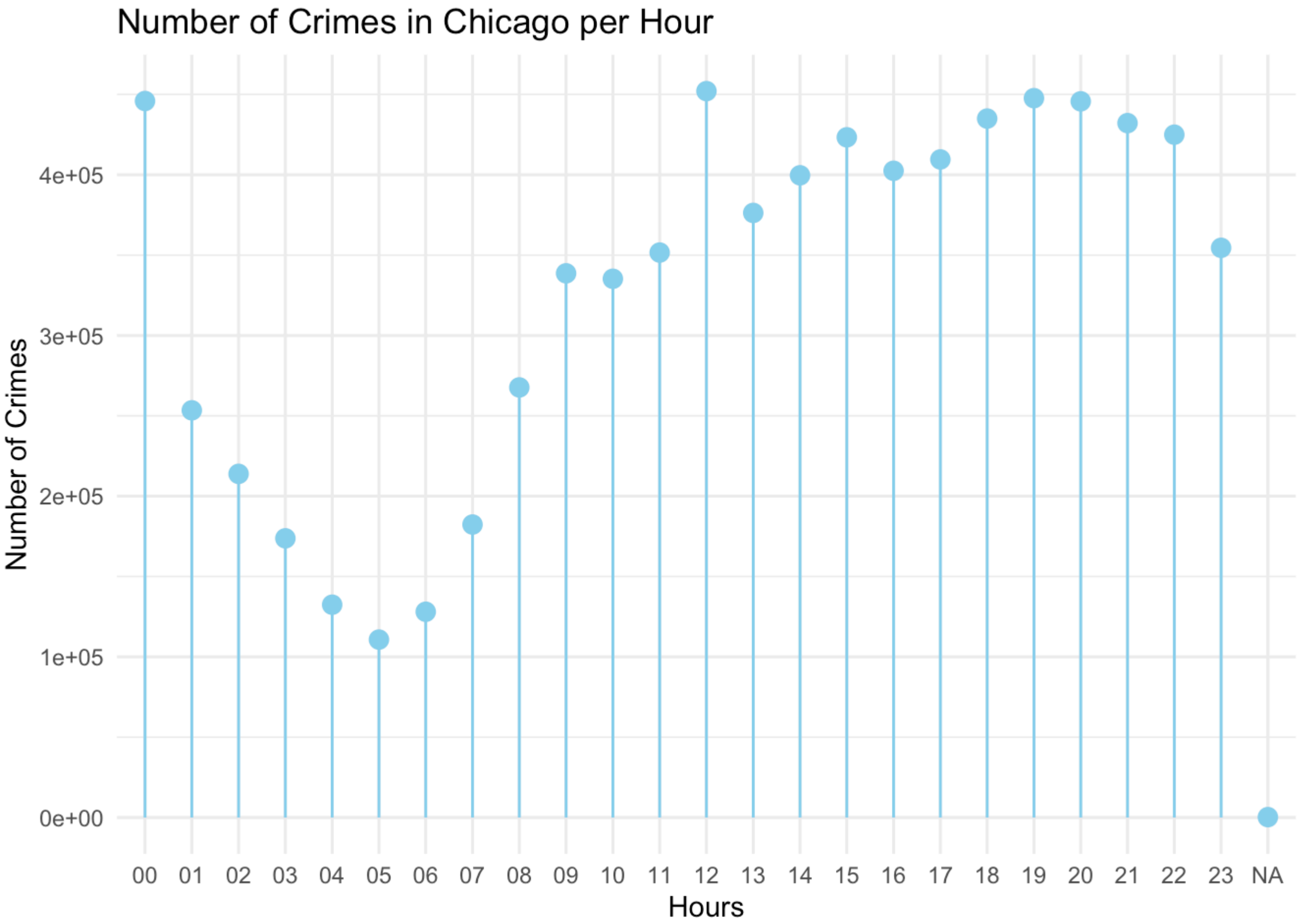
We used a RIDGELINE PLOT to show the number of crimes in Chicago per month. The plot shows that the number of crimes is highest in the summer months, with a peak in July. The reason for the peak could be due to the warmer weather and longer days, which may lead to more people being outside and more opportunities for crime. The plot also shows a smaller peak in December, which could be due to the holiday season. The number of crimes is lowest in the winter months, with a trough in February. The lowest amount could be due to the colder weather and shorter days, which may lead to fewer people being outside and fewer opportunities for crime.

So far, we have looked at the frequency of crime in Chicago based on the year and the month. Here we will be looking at the density of crime based on the day that it was committed continuing the search for patterns for when to.



With this density plot we will analyze the distribution of crime incidents across different dates within a month, with the x-axis representing the dates since the first day of the month and the y-axis representing the density of crime occurrences. The graph reveals that crime is most likely to occur, on average, between 0.025 and 0.04, with a peak observed on day 1.

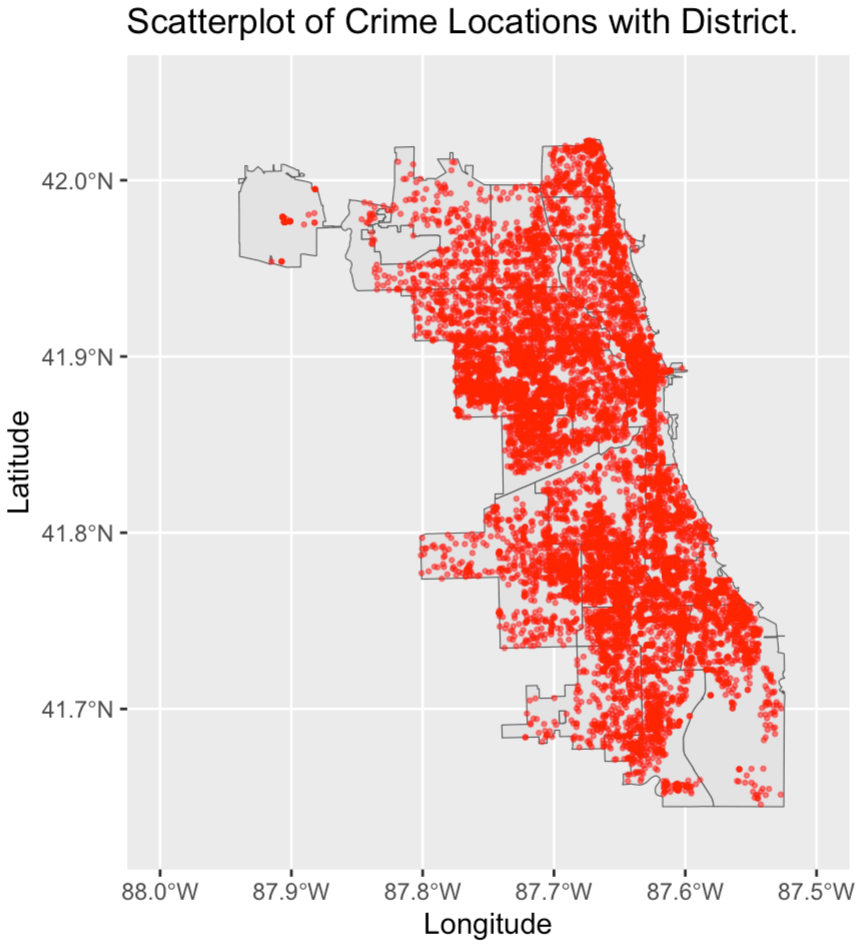
The next graph determines when crimes are being committed in Chicago is the frequency of crimes based on the time during a day that they were committed.



We used a LOLLIPOP PLOT to see how crime incidents are distributed across different times of the day, with the x-axis representing hours since midnight and the y-axis representing the number of crimes in Chicago. The graph shows that the number of crimes is highest in the evening, with a peak around 8:00 PM. This could be due to the warmer weather and longer days, which may lead to more people being outside and more opportunities for crime. The graph also shows that the number of crimes is lowest in the early morning, with a trough around 5:00 AM. This could be due to the minimal number of people outside, or even the decrease in police presence so fewer people are caught. We must stay weary that this data does not represent all crimes that are committed in Chicago and instead, only crimes that were caught.

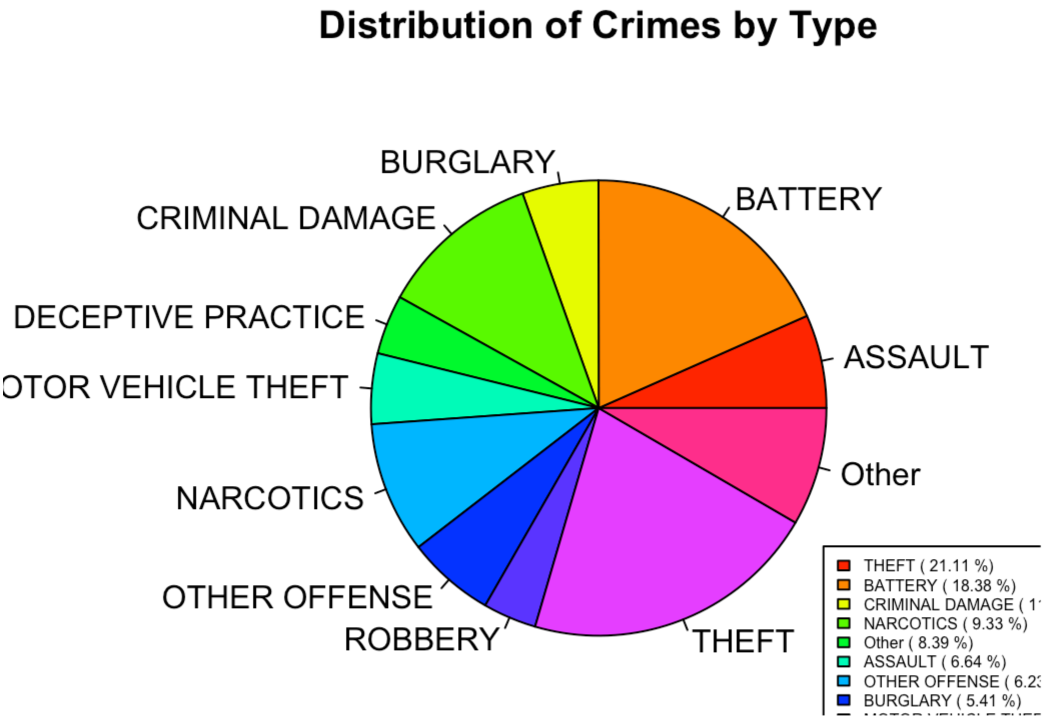
2. Where crimes are committed

Now that we have data on the frequency of when crimes in Chicago are committed, the next step is to figure out the frequency of where they are happening. The first graph for this is a scatter plot. Each dot represents a crime committed in the Chicago area based on its longitude and latitude. Along with this, using the shape data from the Chicago Data Portal we are able to outline the police districts in Chicago.



From this graph, we interpret that there are a few districts where crime is extremely sparse and others where it is extremely dense. Moving forward we will be able to look at the differences between these districts and try to suggest changes to promote limiting crime. There is a lot to take into consideration for this including that these dots only represent crimes that were caught. In addition, this graph only takes into account a sample of the data because otherwise the entire city would be covered in red.

Moving onto the next graph to visualize where crime in Chicago is happening we have a graph of each common crime and its frequency in each district. Once again we are trying to focus on which districts need to be allocated more resources. First, we need to find the most common crimes by the primary type.



With this pie chart we will analyze the frequency of the different crime types. The graph shows that theft is the most common crime type, followed by battery and criminal damage. We have ordered the legend in decreasing amounts to make a clear view of the most common to least common crimes.The result is a that the nine most common crimes are theft, battery, criminal damage, narcotics, assault, other offenses, burglary, deceptive practice, and motor vehicle theft.

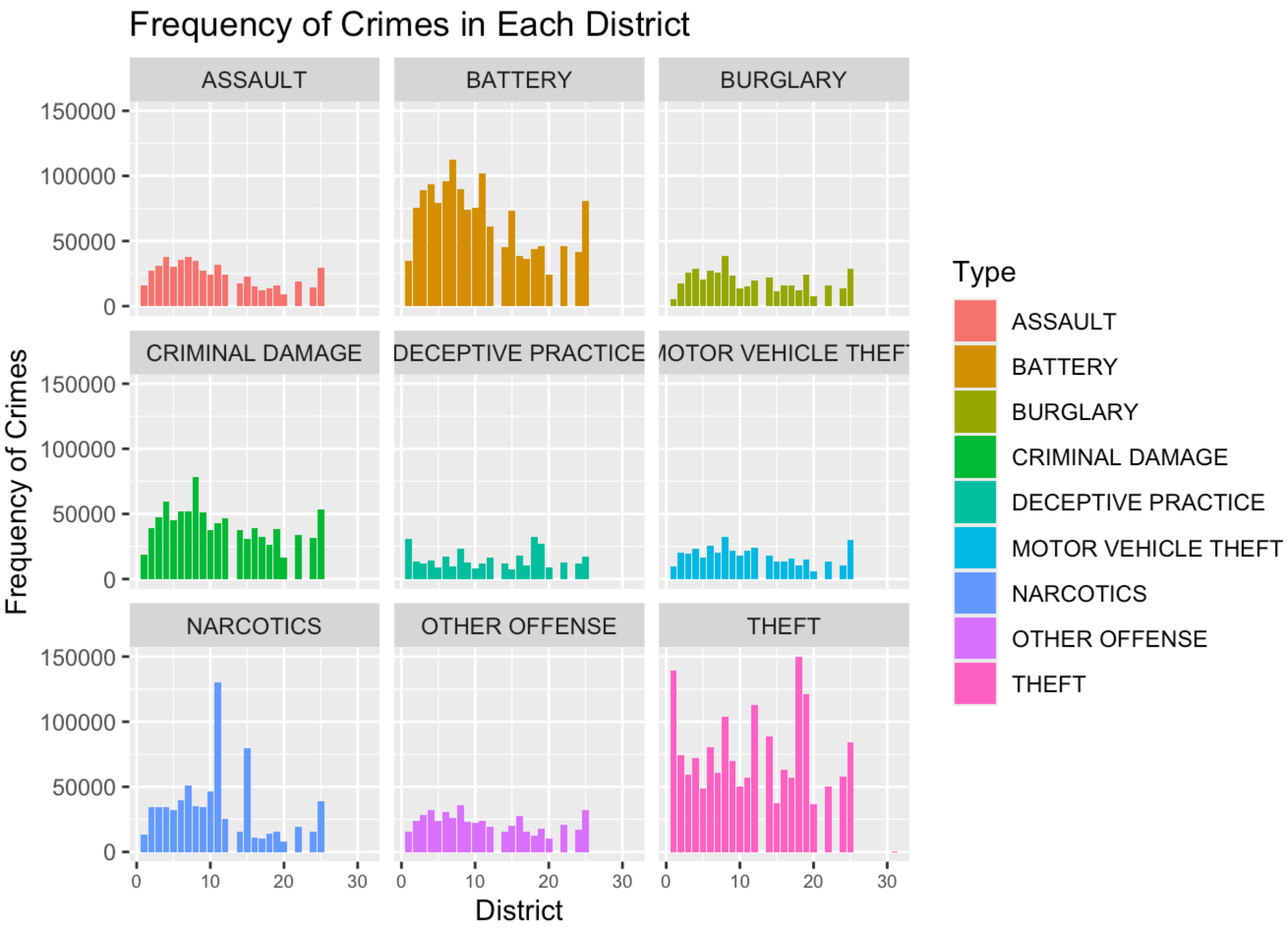
Also, we can replace the pie graph with SQL queries and achieve the same results.

Number of Crimes Based on Primary Type

Primary Type	count
THEFT	1675712
BATTERY	1458987
CRIMINAL DAMAGE	911085
NARCOTICS	740406
ASSAULT	526718
OTHER OFFENSE	494621
BURGLARY	429535
MOTOR VEHICLE THEFT	397843
DECEPTIVE PRACTICE	335123

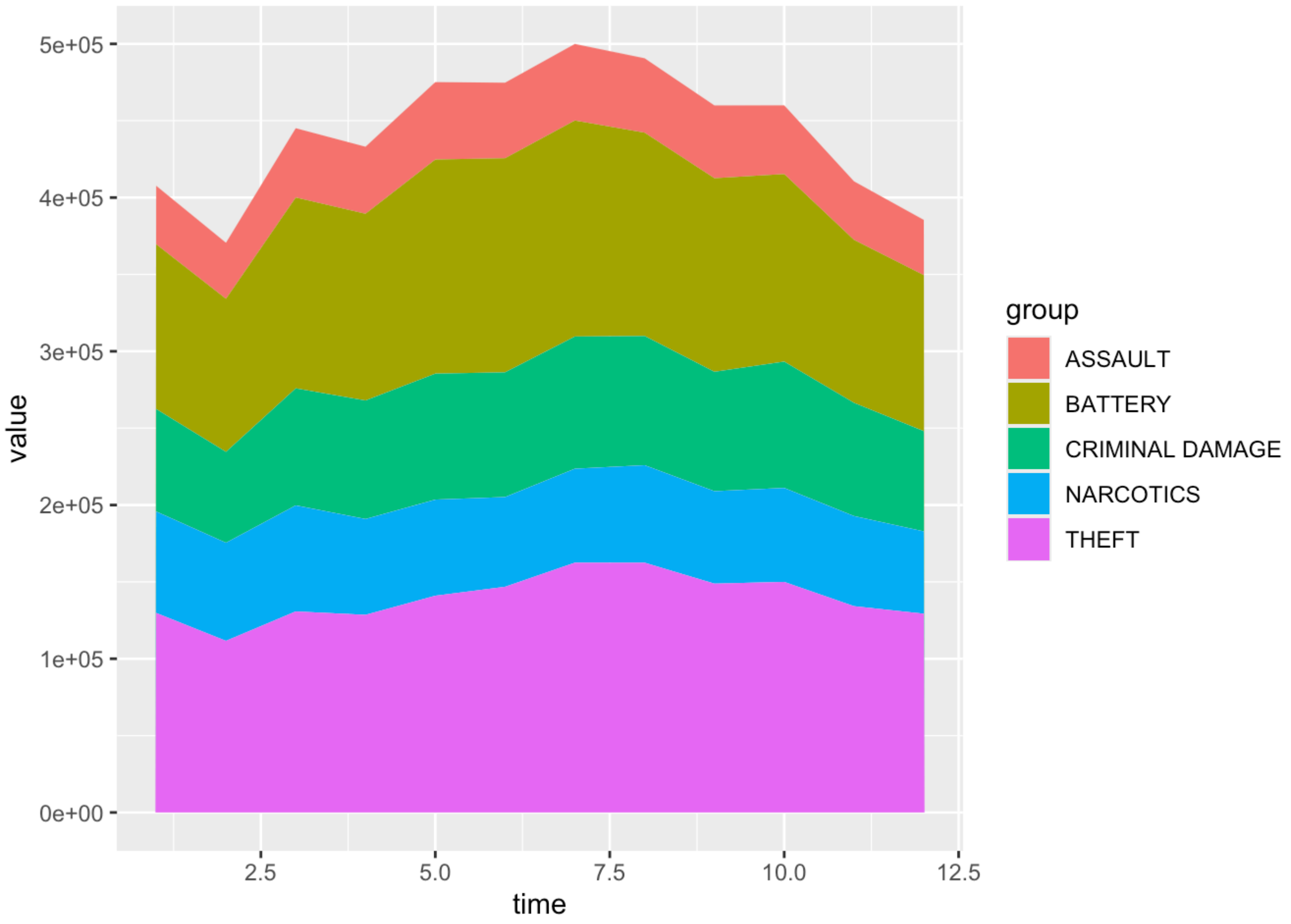
This is the frequency of the nine most common crimes found through an SQLite query put within a Kable for show. We can see that he results in the table is consistent with the pie chart.

Now that we have the most common primary types of crime committed we are able to plot the number of those crimes in each district.



This is a facet wrap graph and there are a couple of key factors visible here that are important to note. For one, assault, battery, burglary, criminal damage, motor vehicle theft, and other offenses all follow a similar curve when comparing the districts. They all almost look like a skewed right normal distribution with a spike on the tail. Between all of these crimes, this means that some of the safest districts are between districts 15-25 and some of the most dangerous are 5-10. From another perspective, focusing on the narcotics graph we see that there is a massive spike in district 11. There could be multiple reasons for this including that there are significantly more people in this district using and distributing narcotics or that the narcotics unit in this police district is much more effective at catching those who are connected to narcotics.

Moreover, we can also use a stacked area chart to show the number of crimes in Chicago per month for the top 5 crime types.



The chart shows that theft is the most common crime type, followed by battery and criminal damage. It chart also shows that the number of thefts is highest in the summer months, with a peak in July. This could be due to the warmer weather and longer days, which may lead to more people being outside and more opportunities for theft. In addition, the chart shows a smaller peak in December, which could be due to the holiday season. Furthermore, the number of thefts is lowest in the winter months, with a trough in February. This could be due to the colder weather and shorter days, which may lead to fewer people being outside and fewer opportunities for theft.

Text Mining

In the text mining part, we filter useful data according to our pie graph and scatter plot in order to make it more convenient for us to modify the graphs in the future.

First, we incorporate text mining into our project, particularly through the use of the dplyr package in R, facilitating a novel approach to data manipulation and enhancement, enabling the generation of additional data columns and innovative filtering techniques. Specifically, we replace a pie chart visualization with a dplyr pipeline to extract and display the nine most common primary crime types from a dataset. This process involves selecting the relevant column (Primary Type), grouping the data by this column, summarizing to count the number of occurrences of each type, and then arranging the types in descending order of frequency. The result is a concise table output using knitr::kable(), which lists the top nine crime types by frequency.

Number of Crimes Based on Primary Type

Primary Type	count

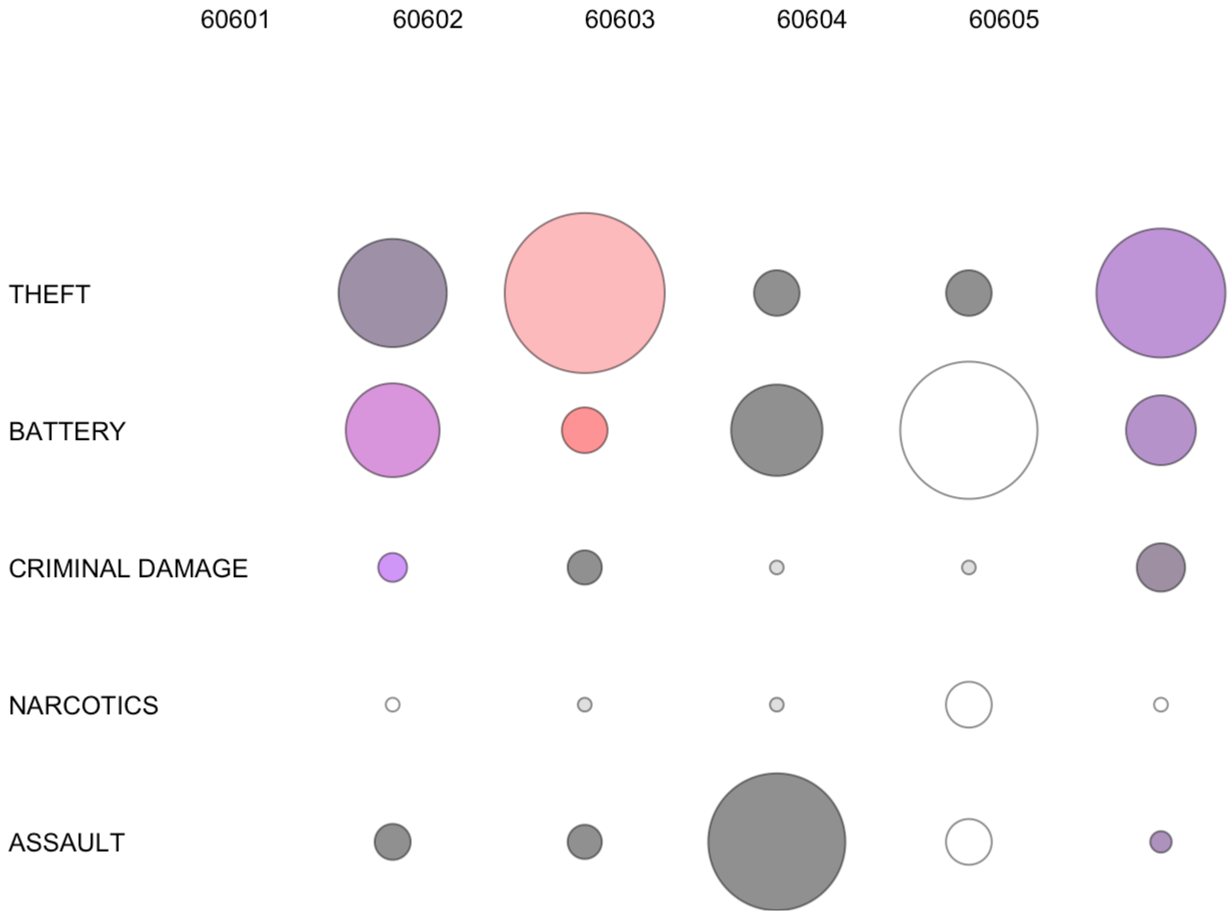
Primary Type	count
THEFT	1675712
BATTERY	1458987
CRIMINAL DAMAGE	911085
NARCOTICS	740406
ASSAULT	526718
OTHER OFFENSE	494621
BURGLARY	429535
MOTOR VEHICLE THEFT	397843
DECEPTIVE PRACTICE	335123

Additionally, we’re interested in determining the number of police stations located in areas with the highest crime densities. Based on the scatterplot analysis (please see the plot below), we’ve identified two such areas: one with a latitude ranging from 41.85 to 41.90 and longitude from -87.78 to -87.60, and another with a latitude ranging from 41.75 to 41.80 and longitude from -87.70 to -87.61. After querying for these areas, the results indicate that out of 25 police stations, a total of 7 police stations are situated near these areas. Therefore, we recommend establishing additional police stations in locations such as (41.85837, -87.62736) and (41.75214, -87.64423), which are the areas with the highest crime densities.

```
  LATITUDE LONGITUDE num_police_stations
1 41.85837 -87.62736                   5
```

```
  LATITUDE LONGITUDE num_police_stations
1 41.75214 -87.64423                   2
```

Killer Plot(2021 March)



The killer plot is a comprehensive and interactive plot of our project. It is made up of a matrix, where each row represents a district in Chicago, represented by different zip codes, and each column represents a type of crime. We listed the top five crime types, which are THEFT, BATTERY, CRIMINAL DAMAGE, NARCOTICS, and ASSAULT. In each column, there will be a drop-down list so that the reader can select different districts to view. In this plot, we select 60601, 60602, 60603, 60604, 60605 as the zip codes. The contents of the matrix are circles. The size of circles represents the number of crimes of that column type in that row district. Larger circle means more crimes. The color of circles represents the arrest rate of crimes of that column type in that row district. As you can see on the legend, the color approaches white and pink as the arrest rate rises and the color approaches dark purple and black as the arrest rate falls. It is very intuitive: the larger and darker the more dangerous. We can also use the slider to search for crime data in a specific month between 2001 and 2024. We can see that for most of the time, theft is the most common crime, and it is way more common than other crimes. Also, many of the circles for theft are dark purple, which means that theft has a relatively low arrest rate. For example, in this plot, we select 2021, March. The circles in the fourth column are very small and white. That means the police in 60604 did a great job in that month. However, the circles in the third column are very large and dark grey That means the police in 60603 did a terrible job in that month.

Conclusion

In conclusion, we used three datasets to analyze the crimes in Chicago. Our primary dataset records the crime information, which consists of over eight million rows and about 20 columns. After cleaning and analyzing the data, we are able to make some suggestions as to where and when more crimes are being caught so that we can help the Chicago police to control crime in a better way. This would include focusing more on the summer months, on the evenings of most days, and increasing support in districts 5-10 with more specifics being available on the map. However, the situation is much more complicated than just the data that we are looking at. We are not aware of the strength of specific police districts within the different units. In addition, catching crimes is much more complex than just having more officers in a specific district. These suggestions may help slightly, but in order to see the results, specific changes would have to be made over years and the resulting data would have to be compared. That’s the end of our presentation. Thanks for your time.