# *Chicago Crime Data Analysis*

## I. Business Understanding

In the last few years, crime trends in Chicago have intensified dramatically, leading to more doubt and less trust in the Chicago Police Department. Despite all the recent controversy about “de-funding” the police, the majority of credible research still points to the fact that crime will continue to rise by even more dangerous levels if the police department resources decrease.

Over the last 50 years, Chicago’s biggest criminal justice challenges have changed very little – statistically residing with homicide, armed robbery, gang violence, and aggravated battery. By 2010, Chicago's homicide rate had surpassed that of Los Angeles (16.02 per 100,000) and was more than twice that of New York City (7.0 per 100,000). By 2015, Chicago's homicide rate rose to 18.6 per 100,000. By 2016, Chicago had recorded more homicides and shooting victims than New York City and Los Angeles combined. And by the end of 2020, the homicide rate rose to 28 per 100,000.

Despite this upward-sloping crime trend, Chicago’s City Council voted to de-fund the police in 2021. As a result, the city ended last year as one of the most violent on record. Chicago had the most homicides – by far – than any other city in the United States in 2021. After the year-long de-funding stint that left more residents dead than in any other year in a quarter century, Chicago Mayor Lori Lightfoot recognized the need for the city to support the Police Department. She increased the budget for the upcoming year and is backing Chicago PD in their efforts to improve efficiency in allocating resources. Moving forward, the Chicago City Council plans to put more officers on the street than ever before and increase the number of detectives investigating violent crimes by nearly 20 percent.

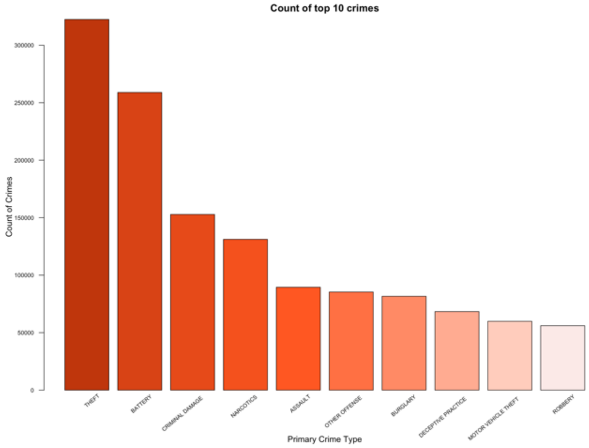
Even though mitigating crime rates surfaces as a societal issue, discovering the most efficient method of allocating police resources to the city is certainly a business problem. Currently, there is a detrimental mismatch between the location of the patrol car and the location of the crime, which contributes to high officer response times – in situations where a few minutes can sometimes be the difference between life or death. Data science plays a huge role in resource deployment decisions for the Chicago PD to maximize efficiency and minimize criminal activity. The next few years will be critical for Chicago PD as they search for efficient allocation of resources amidst a large budget increase.

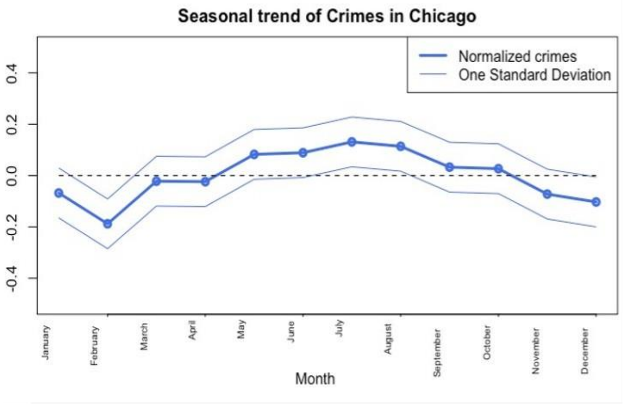
In our analysis of Chicago crime data, we want to address and answer the following questions; in what areas of the city does crime most commonly occur? Are there noticeable trends between city districts and crime intensity? At what time of day do crimes occur? How far from the scene is the closest officer, and how long, on average, does it take the officer to arrive? By analyzing data of 1.5 million crimes (2012-2017) from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system, we aim to identify trends among crimes in the city, utilize predictive modeling techniques (namely, logistic regression) to forecast crime locations and provide actionable solutions to the Chicago PD based on quantitative methods. By doing so, we will recommend the most efficient patrolling locations for Chicago PD vehicles, which minimize the distance between the route and crime-intense areas, as well as the number of officers needed per district.

## II. Data Understanding

The dataset we used reflects reported incidents of crime that occurred in the city of Chicago from 2001 to present. We decided to focus on a subset of this data, from 2012 to 2017. Data is extracted from the Chicago Police Department’s CLEAR (Citizen Law Enforcement Analysis and Reporting) system. There are a total of 22 variables in the original dataset. The original variables in the dataset are: id, case number, date, block, IUCR, primary type, description, location description, arrest, domestic, beat, district, ward, community area, FBI code, x-coordinate, y-coordinate, year, updated on, latitude, longitude, and location. See appendix for variable definitions. We found many of these variables useful for our analysis.

While initially exploring the data, we created many visualizations, one of which identifies the top 10 crime types in Chicago and compared them via a bar plot. Moreover, we captured trends in the way the crimes were distributed over time. This helped us observe that crimes have a distinct seasonal trend – typically increasing during the summer.





Over the years, Chicago’s criminal activity has remained high, with slight fluctuations among the primary type of the crimes (see appendix). Theft, battery, criminal damage, narcotics and assault are the top 5 crimes in Chicago over the data collection period.

## III. Data Preparation

In order to produce the format required for data mining, we spent significant efforts cleaning and preparing the data. This included our method for adding a necessary independent variable to the dataset through further research, dealing with NA values, organizing and facilitating the date and time variables, and factoring some variables.

We first began with the complete 2012-2017 crime data from the Chicago Police Department, but in order to carry out a robust analysis, we identified an inherent need for an additional variable – one that quantifies the intensity and severity of each crime. Our goal was to create a "crime intensity" measure that identified the severity of committed crimes to better understand criminal trends across districts in Chicago. The Chicago PD's data includes a variable called 'primary type', which is formatted as a character string and identifies the main classification of the crime committed. In other words, these "primary types" are the broad categories that each crime falls into – assault, theft, kidnapping, criminal trespassing, robbery, and homicide are examples of primary types. This data classification is interesting, and we wanted to utilize it to further enhance our quantitative models, but it needed to be quantified – when deciding where the Chicago PD's resources are best used, a trespassing crime should not hold the same weight as a homicide.

The United States Sentencing Commission possesses a crime intensity scale nationally known as "federal base offense levels." In this continuum, each type of crime is assigned a base offense level (ranging from 1 to 43), which is the starting point for determining the seriousness of a particular offense. More serious types of crime have higher base offense levels (for example, a trespass has a base offense level of 4, while kidnapping has a base offense level of 32). Essentially, we used the USSC's continuum of crime seriousness to assign a 'crime intensity' rating to each record in our dataset. We made a few assumptions about the crimes themselves to simplify the process – if there was any discrepancy between the primary type of the crime and the base offense level, we assigned the lowest base offense level in the relevant range. These discrepancies were few and should not have any impact on the analysis, because the method used to quantify crime intensity was consistent across all data observations. We added the column ‘crime intensity’ to the dataset. Lastly, we found the average crime intensity level from this new column, which turned out to be 14, and applied the value 14 to the crimes marked with a primary type of ‘other’.

Next, working in R-Studio, we dealt with the NA values in the data. There is 1 NA in ‘district’, 14 in ‘ward’, 40 in ‘community area’ and 37,083 NAs in ‘coordinates. The data appeared to be missing at random with no pattern to it, none of the NA values overlapped, and the missing data made up less than 2% of the entire data (1.5 million records). Because of this, we decided to omit those records with a NA value.

Another data cleaning task involved formatting the date and time columns. The original data was a timestamp value, with the date, time, and AM/PM. We extracted the year, month, day, and day of week, transforming each of these columns into numerical values. We also transformed the ‘time’ variable into a 24-hour format to remove the need for AM/PM indicators. Finally, we factored these date and time variables, along with ‘district’, ‘primary type’ and ‘crime intensity’, to prepare the data for the analysis.

## IV. Modeling

We used the variable ‘crime intensity’ as the dependent variable for prediction. Several other variables which were irrelevant to the analysis were left out. Others were transformed into useful factors, as mentioned in the Data Preparation section, and combined with the remaining variables to form the independent variables used for predicting crime intensity.

We created a predictive model for the crime data. To help Chicago Police allocate police units to districts, it may be difficult to arrive at an all-inclusive metric to pinpoint the need for officers in each area. Our method broadly involved two stages; the first stage was to predict the probabilities of each type of 'crime intensity' occurring in each district. Then, the probabilities can be used as the weights for the crime intensity variable, combined with crime frequency rate (number of crimes for each district) to obtain a 'severity ratio', as a fraction of all crime in Chicago, for each district. The second stage was to cluster all the crimes occurring in each district to obtain geographical locations for placing police units.

**Stage 1:** From our data, there were 19 unique 'crime intensities'. As the dataset contained almost 1.5 million records, we faced major memory space constraints, and therefore created 10 partitions while maintaining the ratio of 'crime intensities. The 'crime intensities' depended on the geographical location, month, day, weekday, and time of the crime. We modeled this using two regression analyses: multinomial regression and logistic regression. The data was split into a training and testing in a 9:1 ratio. Since our data is in chronological order and we would like to predict future crime intensities, we left out the most recent 10% of the data to form our holdout sample.

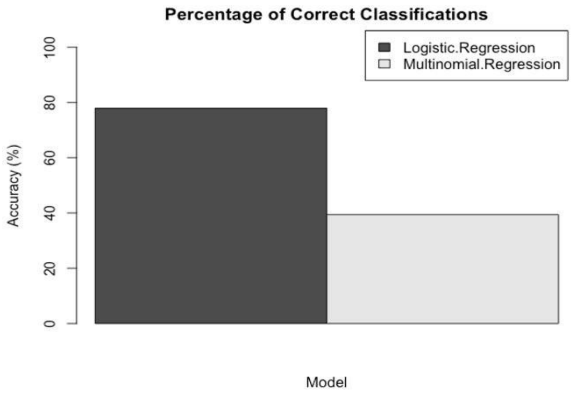
Multinomial regression was the ideal model, since we are predicting the probabilities of multiple levels (more than 2 factors). However, the model could not handle as many features as we wanted to include to improve its fit. We then ran a logistic regression for each of the 19 types of 'crime intensities' to get their respective probabilities. The primary advantage of using this method was the ability to use more significant features in the model. As these were 19 individual logistic regressions coming together, individually they required less computing power.

**Stage 2:** For each district, we calculated the "crime frequency rate" as the number of crimes occurring in the district as a fraction of the total crimes reported. Our final predictive model produces a matrix of outputs that reports the probability of a crime of each intensity level occurring in each district. Based on these likelihoods, we were able to calculate a weighted average crime intensity rating for each district – which we dubbed its "severity ratio." Based on this ratio, we can identify and deploy a number of police units for each district from the total police units (12,000 at the time of the dataset preparation, but likely to increase with the larger budget next year) proportional to the severity ratio of each district.

For each individual district, we ran k-means clustering weighted by 'crime intensities' with the number of police cars in the district as *k,* and only using latitude and longitude of the crimes reported. The output of this clustering gave us the centers, or coordinates, of where the Chicago PD should station each police unit within each district in order to minimize the response time to a crime.

## V. Evaluation

For each model, we removed 10% of the data as our "holdout" sample to allow us to test the accuracy of each model. In order to gauge which model performed better, we calculated accuracy as the number of correct predictions of the model divided by the total number of predictions.  From our findings, the accuracy of logistic regression is more than that of multinomial regression. It is important to note, however, that we can't include as many significant features in multinomial regression as compared to logistic regression because of a lack of computing power. Because of these limitations, we cannot be completely certain as to what the potential of the multinomial regression could have been in an ideal setting.



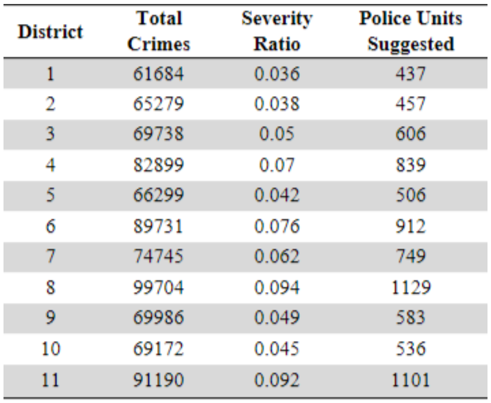
## VI. Deployment

Based on our findings, we can recommend a plan of action for the Chicago Police Department – one that would mitigate crime extensivity in the city by placing officers as close as possible to crime-intense areas. We suggest that the Chicago PD deploy resources proportional to the standardized results of our 'severity ratio'.

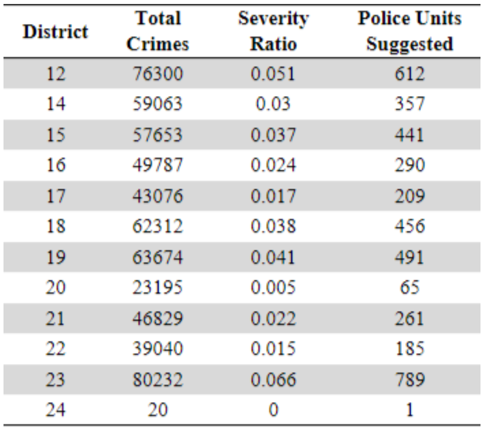
The methods described in Stage 2 detail the process by which we assigned a "severity ratio" to each of the 24 districts that the Chicago Police Department is responsible for. Crime in Chicago is, unfortunately, quite ubiquitous, and it's impossible for the officers to be at the location of a crime when it happens. The goal is to station the officers as close to the predicted crimes as possible to minimize response times and maximize arrests.

Currently, the Chicago PD has officers stationed in districts fairly equally – they designate exactly one patrol vehicle to each beat in each district. However, it's inherently inefficient to delegate resources by physical geographical area – rather, the organization needs to consider neighborhood crime frequency and severity when assigning resources to the city. Our analysis quantifies crime severity by geographical location. Subsequently, we recommend the

Chicago PD to allocate resources (officers and vehicles) proportionately to each district based on the 'severity ratio' of each district. This is where intense crime is most likely to occur. For example, district 8 has a severity rating of 0.094, so we advise the Chicago PD to send 9.4% of police units to this area. The tables below show the breakdown of our proposed resource distribution for each district.

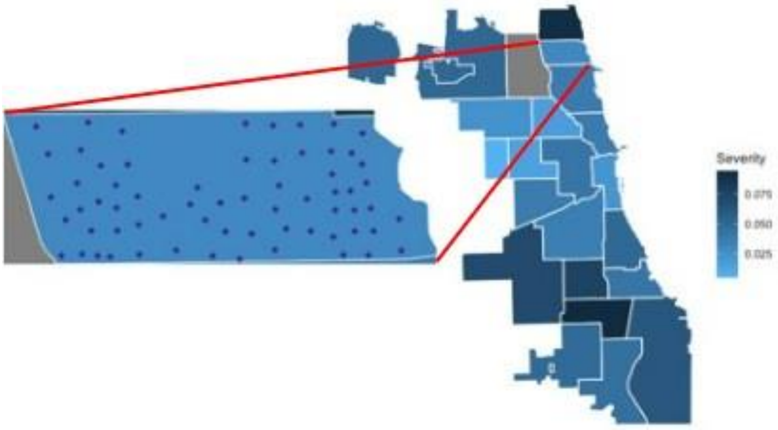


| District | Total Crimes | Severity Ratio | Police Units Suggested |
|---|---|---|---|
| 1 | 61684 | 0.036 | 437 |
| 2 | 65279 | 0.038 | 457 |
| 3 | 69738 | 0.05 | 606 |
| 4 | 82899 | 0.07 | 839 |
| 5 | 66299 | 0.042 | 506 |
| 6 | 89731 | 0.076 | 912 |
| 7 | 74745 | 0.062 | 749 |
| 8 | 99704 | 0.094 | 1129 |
| 9 | 69986 | 0.049 | 583 |
| 10 | 69172 | 0.045 | 536 |
| 11 | 91190 | 0.092 | 1101 |



| District | Total Crimes | Severity Ratio | Police Units Suggested |
|---|---|---|---|
| 12 | 76300 | 0.051 | 612 |
| 14 | 59063 | 0.03 | 357 |
| 15 | 57653 | 0.037 | 441 |
| 16 | 49787 | 0.024 | 290 |
| 17 | 43076 | 0.017 | 209 |
| 18 | 62312 | 0.038 | 456 |
| 19 | 63674 | 0.041 | 491 |
| 20 | 23195 | 0.005 | 65 |
| 21 | 46829 | 0.022 | 261 |
| 22 | 39040 | 0.015 | 185 |
| 23 | 80232 | 0.066 | 789 |
| 24 | 20 | 0 | 1 |

We translated all of the severity ratios into a map of Chicago, color-coded by crime intensity (shown below). The darker regions represent more frequent and severe criminal activity. The enlarged section of the map details an example of the comprehensive analysis that we ran in R to determine the most efficient locations for police units.



The enlarged section shown on the map is District 20, and each of the 65 dots in the district represent the recommended location for a police unit to be stationed, in order to minimize the distance between their location and the predicted crime locations. In this visual, District 20 simply functions as an example of our complete geolocation analysis. For each of the 24 police districts in Chicago, we provide the optimal geolocation to station each one of Chicago PD's 12,000 police units. The darker colored districts will be more concentrated with dots (police units), implying a stronger police presence needed in crime-intense areas. These suggested values, by district, are also shown in the table above.

The largest issue that may arise with this resource allocation technique is that crime is not fully independent of officer location. As officers relocate to crime-intense areas, crime levels may respond to this over time and the locations of criminal activity will shift further from officer location. For this reason, we recommend that the Chicago PD reruns this analysis on a bi-annual basis to predict the ebb and flow of crime trends.

During the implementation of this new resource allocation, the Chicago PD should be mindful of the ethical implications of these decisions. Utilizing resources efficiently is a business problem, but preventing crime is a social issue. Removing officers from neighborhoods to relocate them to other areas can be potentially detrimental to resident and neighborhood safety. We recommend that the Chicago PD maintains at least 1 patrol vehicle and 2 officers in each sector (for full geographical coverage) but reconsiders deployment within beats and districts.

## VII. Conclusion

Overall, we found that a slight adjustment to the Chicago PD's resource allocation decision-making process can significantly enhance the efficiency of police units. Using quantitative methods, namely logistic regression, we can capitalize on the extensive amount of data available on crimes in Chicago and improve the deployment of officers in each area, proportional to crime intensity and crime frequency.

## VIII. References

“CLEAR Application | Chicago Police Department.” *Office of Public Safety Administration*, Sept. 2022, home.chicagopolice.org/services/clearmap-application/.

NPSF. “Chicago Mayor Lightfoot Proposes More Police Funding as Crime Continues to Spiral.” *National Police Support Fund*, 21 Nov. 2021, nationalpolicesupportfund.com/mayor-lightfoot-proposes-more-police-funding-as-crime-continues-to-spiral/.

“Myths and Realities: Understanding Recent Trends in Violent Crime.” *Brennan Center for Justice*, 4 Oct. 2022, www.brennancenter.org/our-work/research-reports/myths-and-realities-understanding-recent-trends-violent-crime.

Ramsden, Daryn. *Making Maps with GGPLOT2*. 28 Mar. 2020, thisisdaryn.netlify.app/post/intro-to-making-maps-with-ggplot2/.

Rushin, Stephen, and Griffin Edwards. “De-Policing.” *Cornell Law Review*, 2018, doi:10.31228/osf.io/sc37v.

“2021 Ends as Chicago's Deadliest Year in a Quarter Century.” *NBC Chicago*, NBC Chicago, 1 Jan. 2022, www.nbcchicago.com/news/local/2021-ends-as-chicagos-deadliest-year-in-a-quarter-century/2719307/.

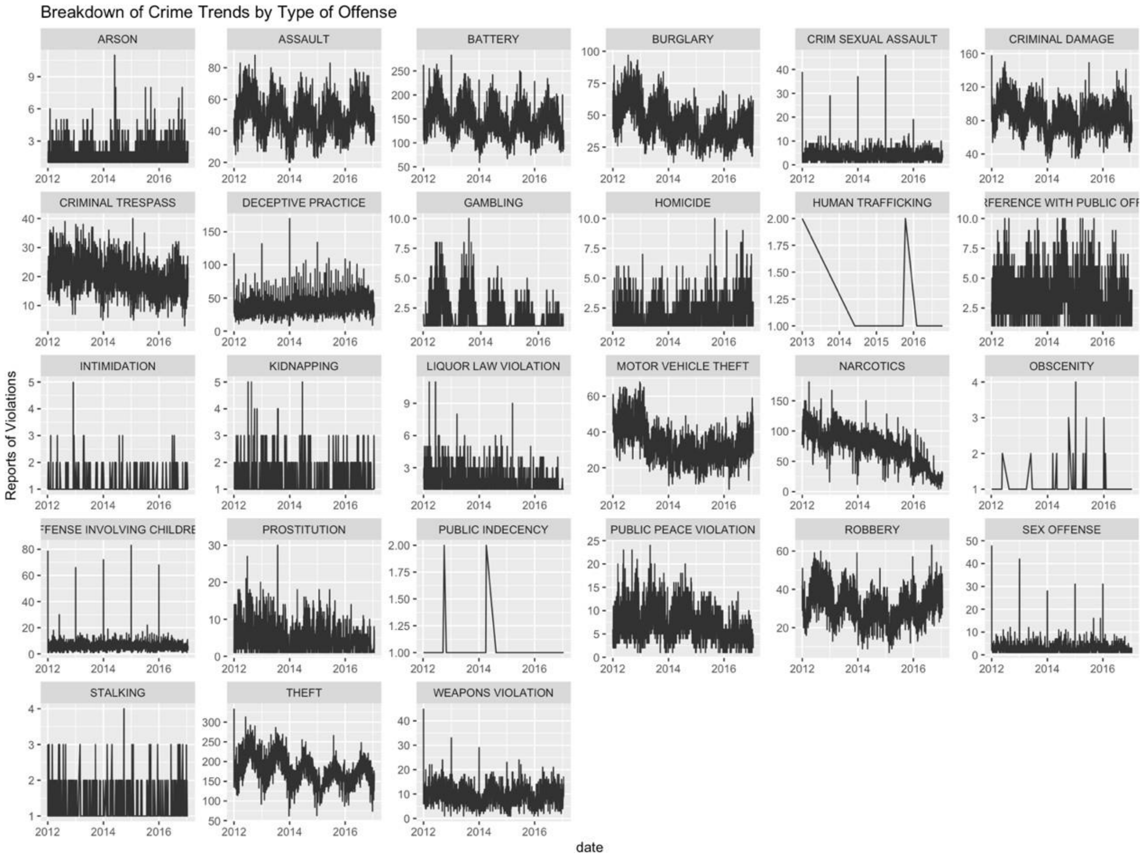
## IX. Appendix

**Variable Definitions:**

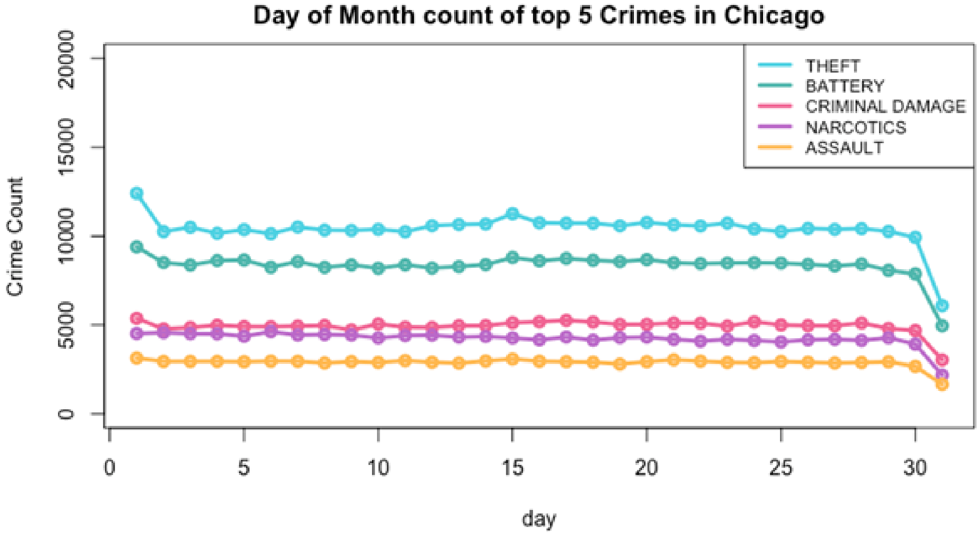
a. ID: the unique identifier for the record
b. Case number: the Chicago Police Department Record Division number, which is unique to the incident
c. Date: when the incident occurred, which is sometimes the best estimate
d. Block: the partially redacted address where the incident occurred, placing it on the same block as the actual address
e. IUCR: the Illinois Uniform Crime Reporting code, which is directly linked to the ‘primary type’ and ‘description’ variables
f. Primary type: the primary description of the IUCR code
g. Description: the secondary description of the IUCR code, a subcategory of the primary description
h. Location description: the description of the location where the incident occurred
i. Arrest: indicates whether an arrest was made
j. Domestic: indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act
k. Beat: indicates the beat where the incident occurred; a beat is the smallest police geographic area – each beat has a dedicated police beat car – three to five beats make up a police sector and three sectors make up a police district
l. District: indicates the police district where the incident occurred; the Chicago Police Department has 22 police districts
m. Ward: indicates the ward (City Council district) where the incident occurred
n. Community area: indicates the community area where the incident occurred; Chicago has 77 community areas
o. FBI code: indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS)
p. X-coordinate: represents the x coordinate of the location where the incident occurred
q. Y-coordinate: represents the y coordinate of the location where the incident occurred
r. Year: represents the year the incident occurred
s. Latitude: represents the latitude of the location where the incident occurred
t. Longitude: represents the longitude of the location where the incident occurred
u. Location: represents the location where the incident occurred in a format that allows for creation of maps

## Exploratory Data Analysis Visualizations:

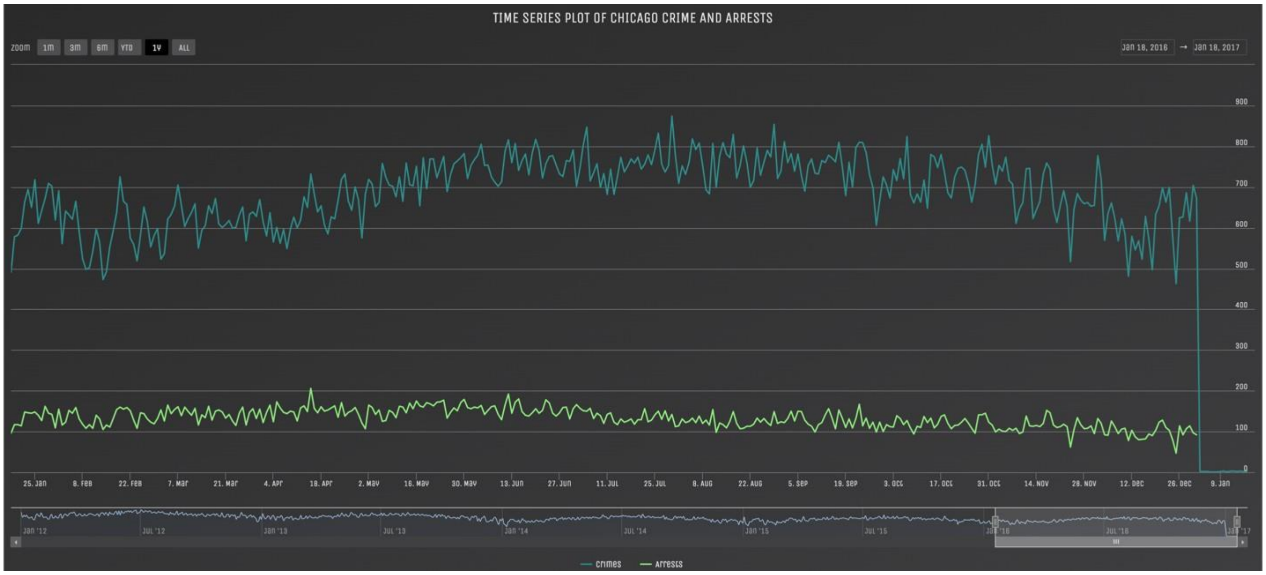
a. Frequency Distribution of each crime type over the years 2012 to 2017



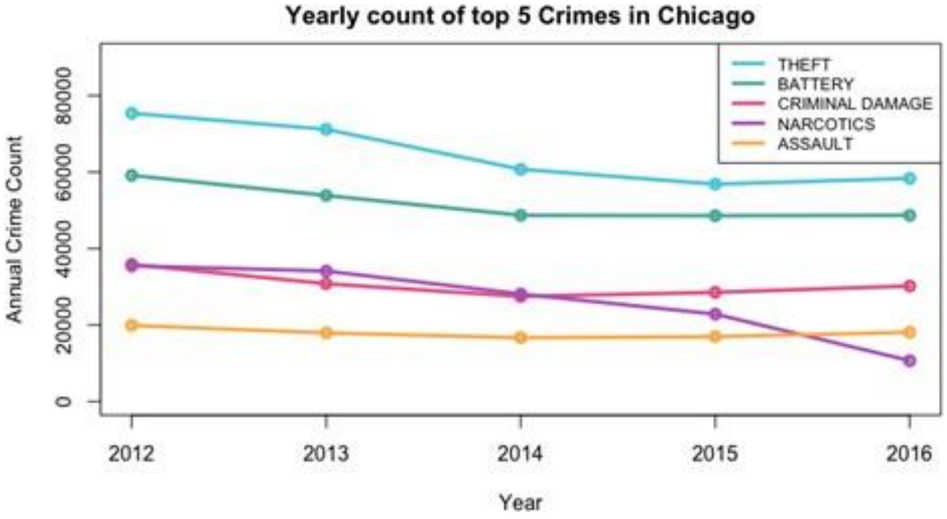
b. Total count of the Top 5 crimes on each day of the month over the years 2012 to 2017



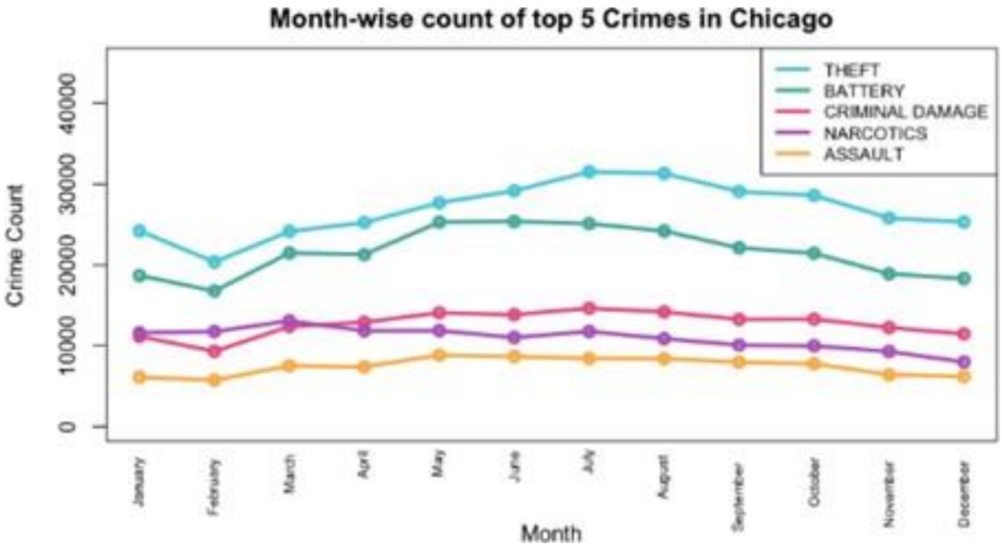
c. Time Series Plot of crimes vs arrests over the period 2012 to 2017



c. Top 5 crimes committed across the years by count



d. Patterns of the top 5 crimes across the months in a year; it can be noticed that most of them have a peak around June and July



**Model Constraints:**

With our goal in mind, we considered modeling our problem statement with three different methods:

1. Random Forest was our first choice. This would be extremely helpful in classifying the crimes into their different crime intensities. However, while implementing this method, we faced a major hurdle: computing power. Our computers were not capable of handling the random forest method even after trying to increase the RAM allocated to the R environment.
2. Multinomial Regression was our second choice. The primary issue we faced was once again computing power and the drawback of using this model was that it was not able to include all the significant features required for the fit. We had to reduce to just 5 independent variables as R was not capable of handing the amount of variables involved for this classification problem. While this compromised the accuracy of our model, we were able to obtain results.
3. Logistic Regression was our third choice. For this, we ran 19 regressions, one for each Crime Intensity Level. This became the model we chose for our final result.