Aishwarya Subramanya,

Northeastern University

ALY 6110 - Fall 2019

December 13, 2019

**Final Project Report: Chicago Crimes between 1/1/2017 - 11/12/2019**

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data>

**Introduction**

The City of Chicago offers a wide range of datasets relative to their municipal operations. Our group took interest in their crime dataset, which is reflective of crimes reported to the police in the Chicago metropolitan area. The reported crime records are updated constantly to reflect the crime most accurately, and to report whether or not an arrest was made. It is important to note that the crimes dataset entries are sometimes only based upon the preliminary data of a reported crime, so entries are not necessarily legally proven or substantiated by law enforcement (City of Chicago, 2019).

The first step of our analysis was to export the dataset and define the data. All columns in the dataset are defined by descriptions in Appendix I. Next, we proposed our initial questions before beginning the analysis:

1. Which areas have the fastest growing crime rates?
2. In which districts do fatal crimes occur the most?
3. In which districts is theft most prevalent?
4. How can we utilize time series analysis to show trends in crime over time?
5. Can we map out areas where security measures should be prioritized?
6. How have the proportion of arrests relative to non-arrests changed over time?

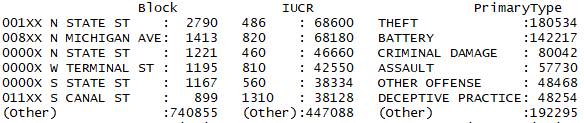
We then targeted a timeframe for our analysis, then reduced and cleaned the dataset. We shared our dataset with the group and took the next steps in our analysis: analyzing our data and generating insights with R and Python, and creating a dashboard visualization with Tableau. With these tools we derive our key insights. Lastly, we produce our recommendations to the city and its police force.

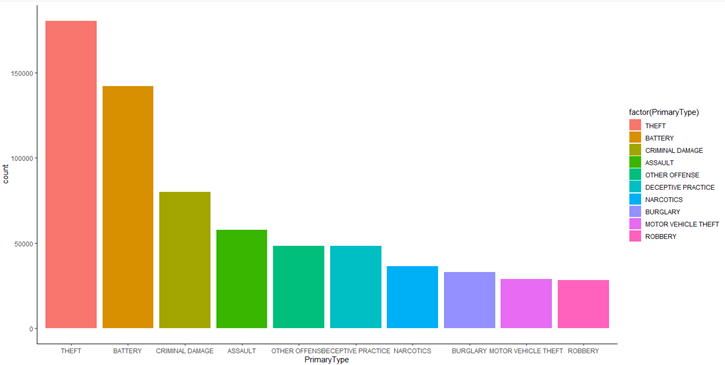
**Cleaning the Dataset**

When we first got the dataset, it had 7 million rows, which is clearly too large and computationally expensive. Since the requirement for data is more than 500,000 rows, we have reduced the data to the past three years. Also, there was a proportion of null values in the data, in columns such as longitude and latitude. Since our analysis will be partly based on geography, we removed all the rows containing null values, which brought the data down to 749540 rows and 22 columns. In the last stage of data cleaning, the column date had the following format of data: 04/10/2019 04:37:00 PM. In order for us to extract the time components, we divided the date columns into two columns, with date and time in their separate columns.

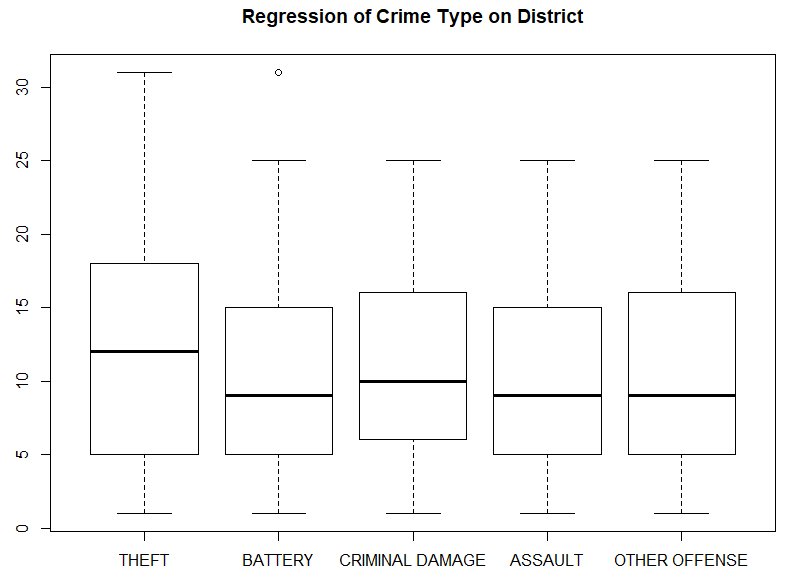
**Analysis with R and Python**

The plan for our analysis with R is to perform an Exploratory Data Analysis - we produce summary stats, intro to data features, analyze the volume of crime by district, type of crime and/or year, and include visual representations using R.

Before we begin to answer some of our initial questions, it is important to perform an Exploratory Data Analysis (EDA) to gain insight on what our dataset can address. By using a few simple commands, we understand how many features and rows we have, and what each column means. “Dim(mydata)” gives the total number of rows and columns (24 and 749,540, respectively). “Summary(mydata)” provides the summary statistics below on the number of occurrences of crime on certain dates, blocks and overall type of crime in general. In our EDA statistics, we see that crimes most commonly occurred on New Years Day (presumably New Years Eve), which could be tied certain types of crime involving heavy drinking. Also of importance is to note the blocks in which the highest volume of crime occurs (N State St and Michigan Ave) and the highest type of crime (theft and battery). With this information, the City of Chicago would likely want to better allocate staff to these areas, and on those dates.

As seen below, a visual interpretation of what types of crimes are most frequently committed (theft and battery) could be added to the Chicago state website. This could be used to heighten the importance of locking doors and learning self defense.

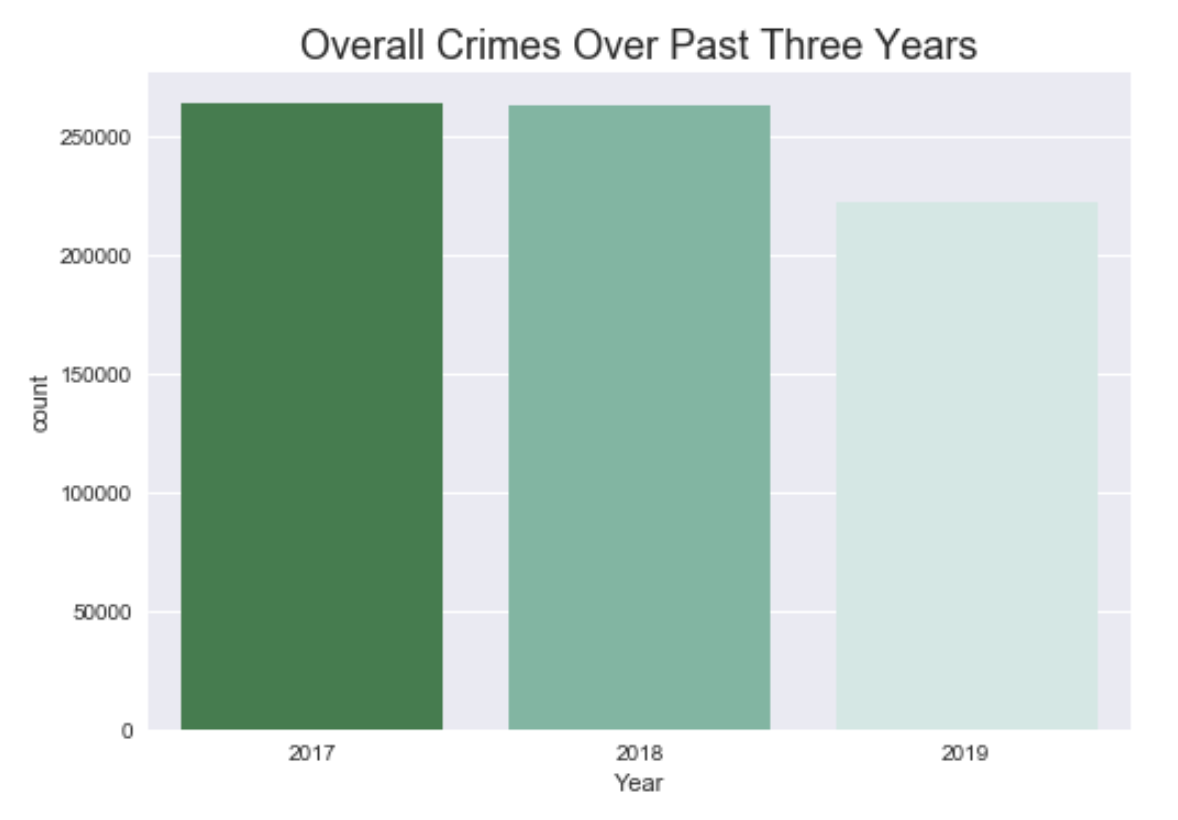
There are over 30 types of crime categories in the dataset. To prioritize, I created a subset of my original data, only taking the category “Primary Type”. I then sorted the top 10 most frequent Primary Types of crime, and plotted them in most frequent order of occurrence. Visual interpretations of data can be used to identify what times of crime are most likely to occur for those who learn better through graphics (i.e. the young or the elderly).

Further, we examine the regression on the type of crime and district in Chicago in which it was committed. Districts are translated by a numerical value 1-30 and then plotted against the type of crime. Again, I used a subset of the data, this time using the top 5 most frequently committed crimes Below, we see that a majority of crimes are committed in districts 5-15. For certain crimes such as narcotics related crimes (not pictured in subset), the range of districts in which most crimes are committed is small, mostly between districts 8-11. This information could be used specifically to staff areas where narcotics crimes are prevalent with the drug trafficking force.

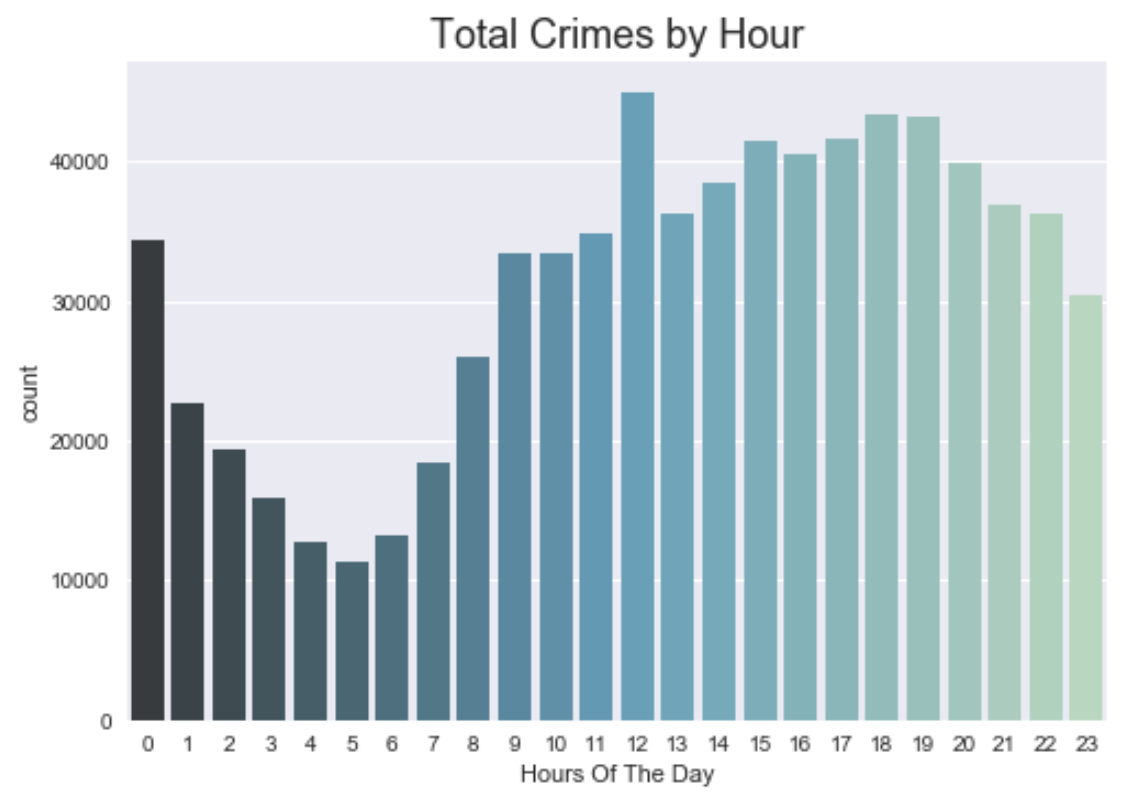
While our dataset only contains crime information logged over the past 3 years, we could compare year-over-year regression analysis to see if certain types of crimes have always existed in the same districts. Addressing these sorts of trends with analytics is the future of crime prevention.

Furthermore, using Python, we conducted some analysis regarding the time of the crimes committed, and tried to gain some insights on the time when the crimes are most likely to occur. We first examined some big pictures regarding crimes in general, and then we dug deeper into the time of specific crime types.

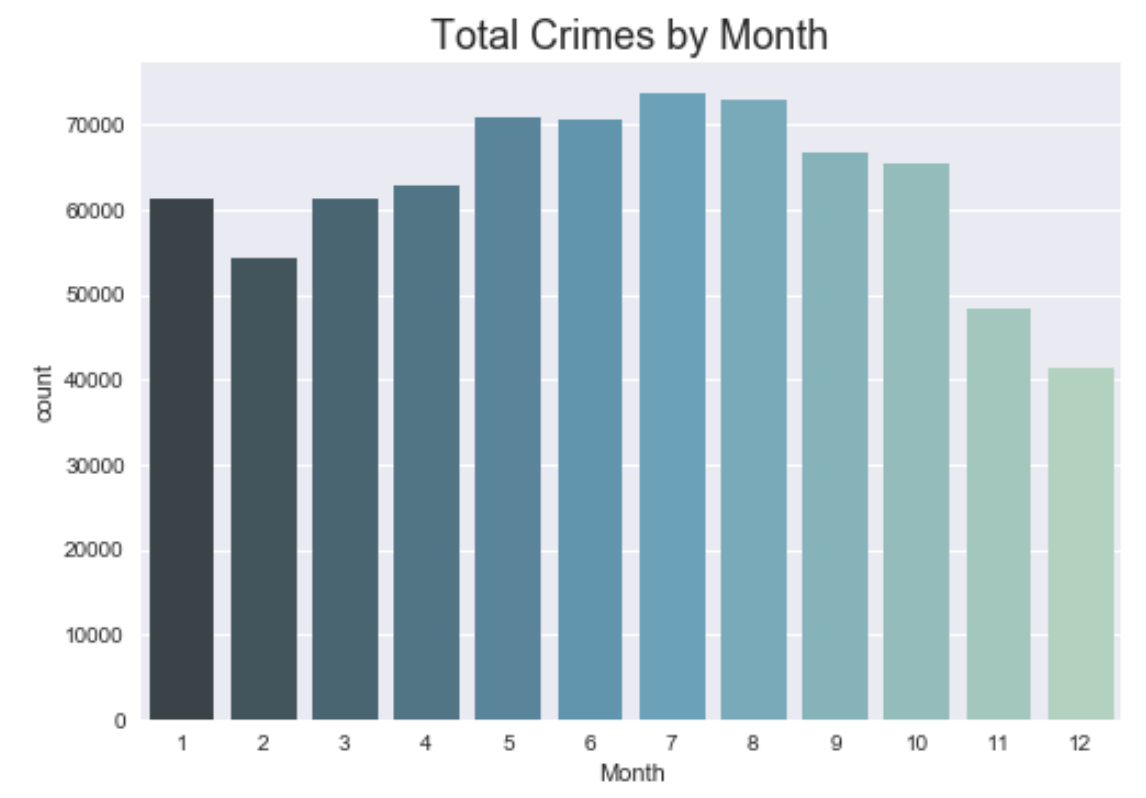
Firstly, we wanted to present a big picture total crimes trend for the past three years:



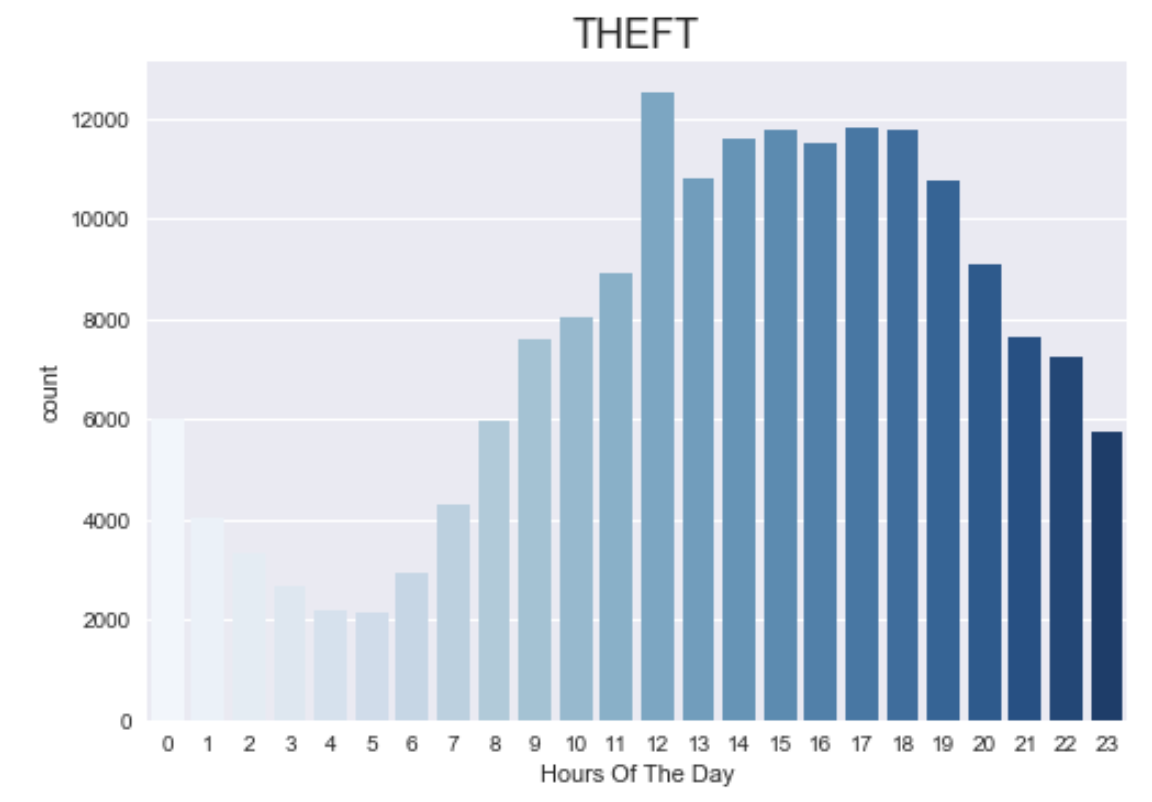
The graph below shows the total number of crimes committed in the past three years, grouped by hour of the day. From this we see a trend that, crime rates goes down from 7pm, and starts to increase rapidly from 5am, peaks at 12 noon, and starts to go down from 7pm.



We also wanted to look at total crimes by month. Evidently, the crime rate is lower during the holiday seasons, and is higher in the middle of the year.

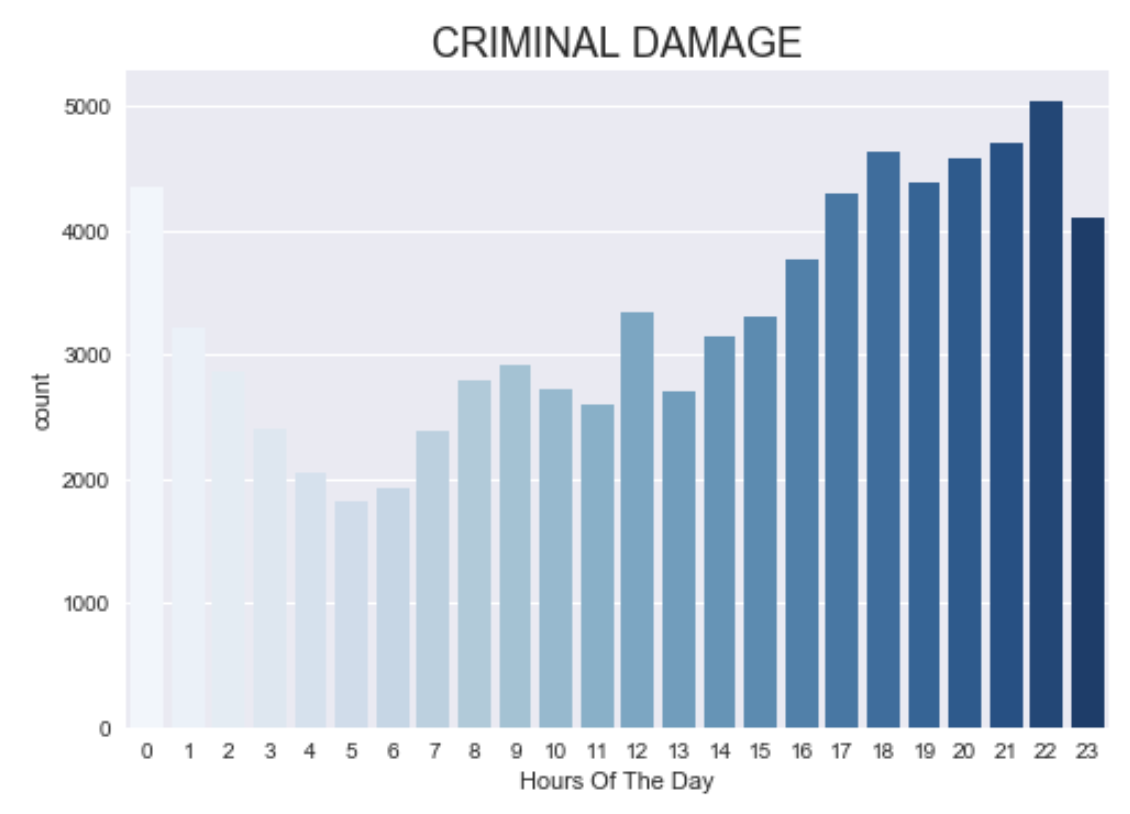


But it wasn’t enough. In order to best allocate the police force, it might be useful to examine each crime type with the same time granularities. Here is a plot of just the theft crimes by hours.



It is similar to the total crimes by hours, and 12 noon seems to be a peak time for theft as well.

Lastly, unlike theft, the crime rate of criminal damage keeps going up all the way until late night at 10pm.

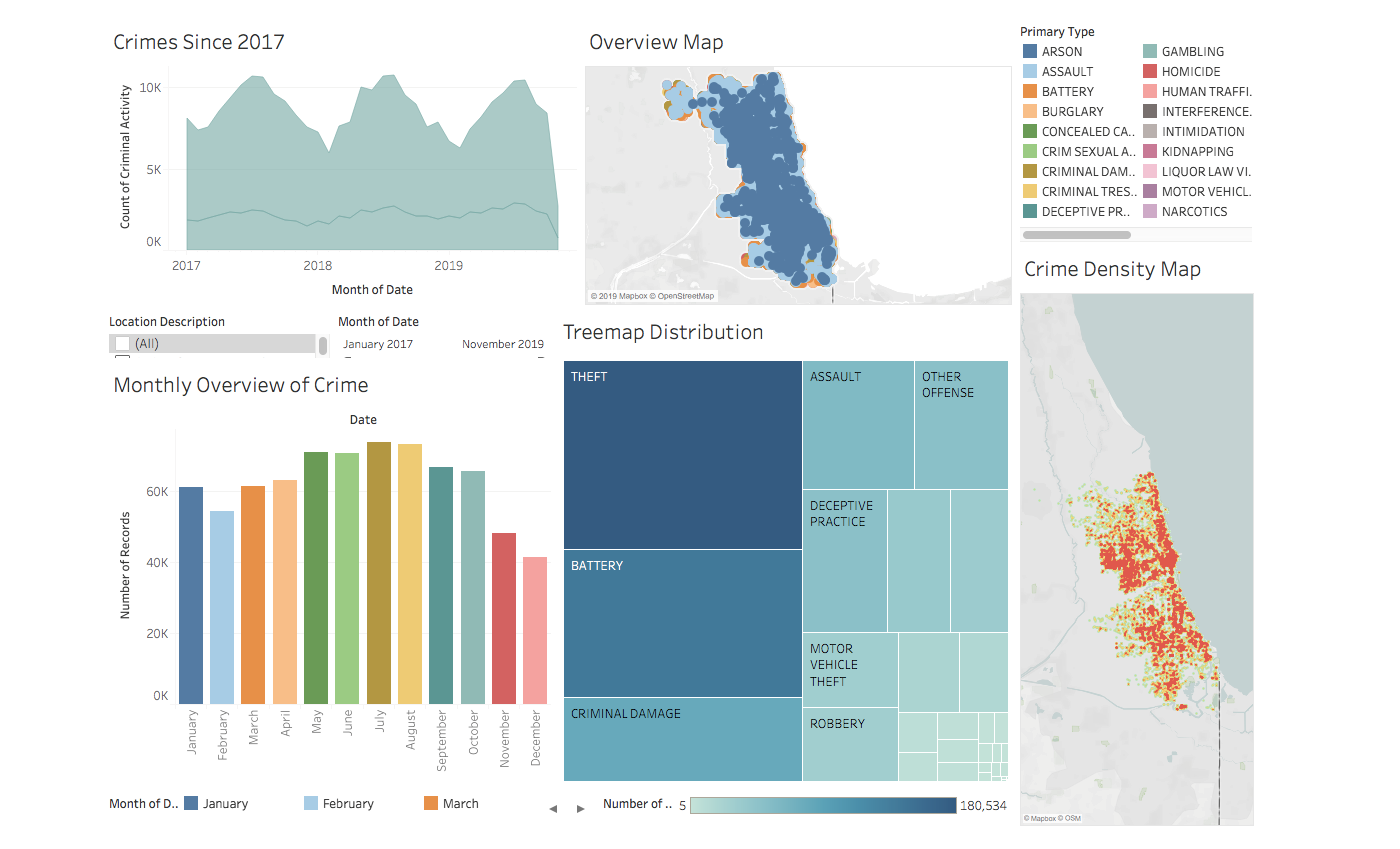


**Visual Analysis with Tableau**

Next, we created our dashboard in Tableau to explore the following real-world business and policy questions:

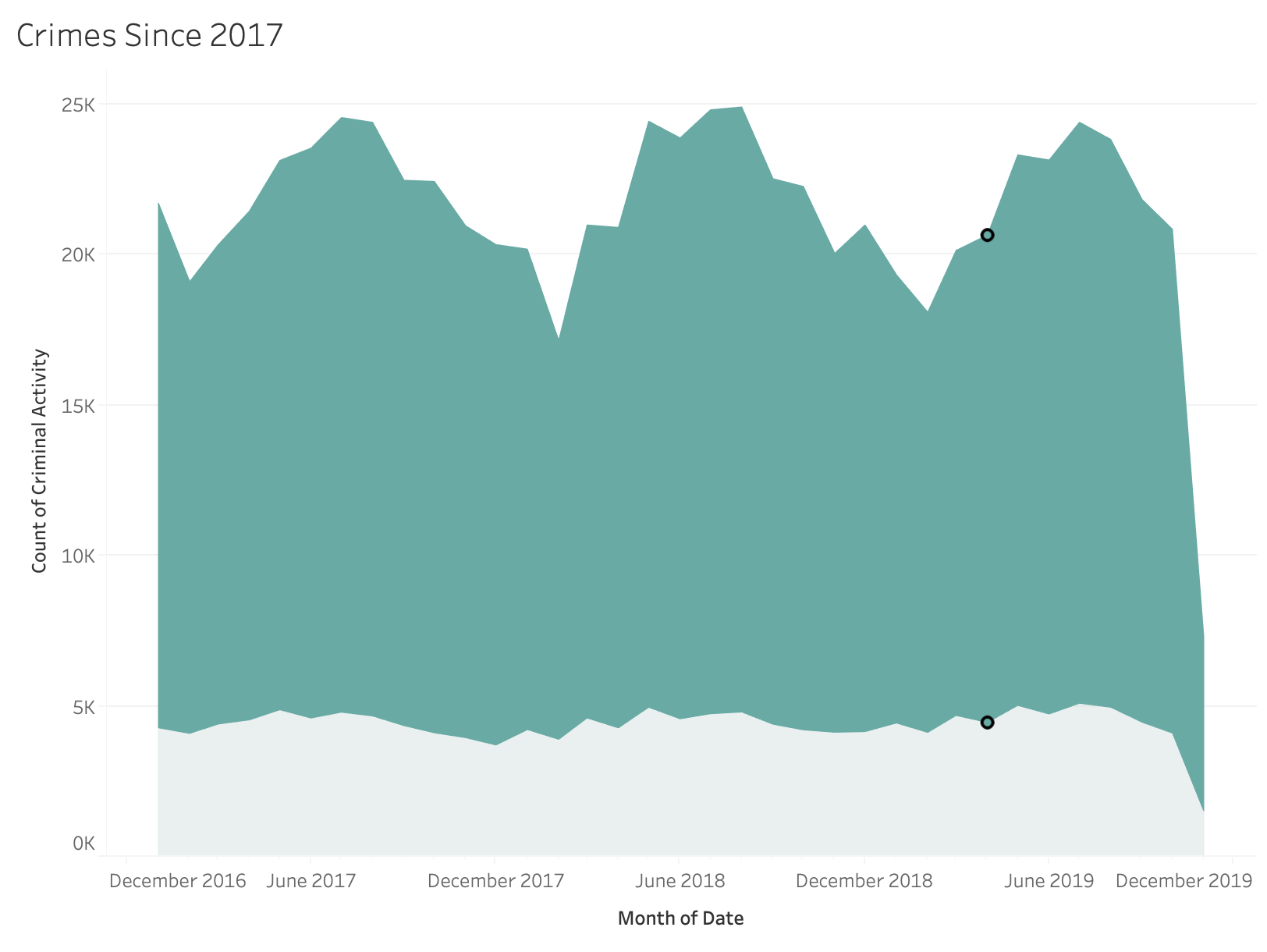
* What are the crimes that occur the most in the city?
* In which month are most of the cases are recorded?
* In which districts is theft most prevalent?
* Are there seasonal trends in criminal activity?
* How have the proportion of arrests changed over time?
* Where are the geographical crime hotspots and do they change throughout the 24hr day?

The Chicago Crimes dataset can be visualized to answer various business questions that are stated above and to have an insight to the audience to understand the dataset. Here, a dashboard created for the dataset is as shown below.

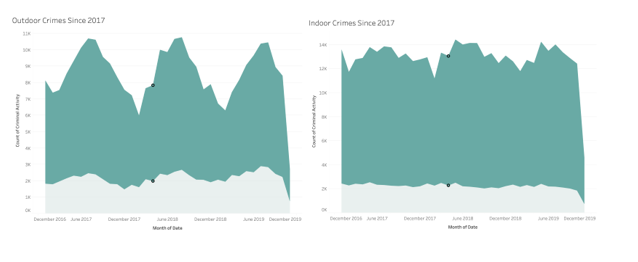
****

This dashboard consists of various types of visualization to get an insight and solve the problem statements.

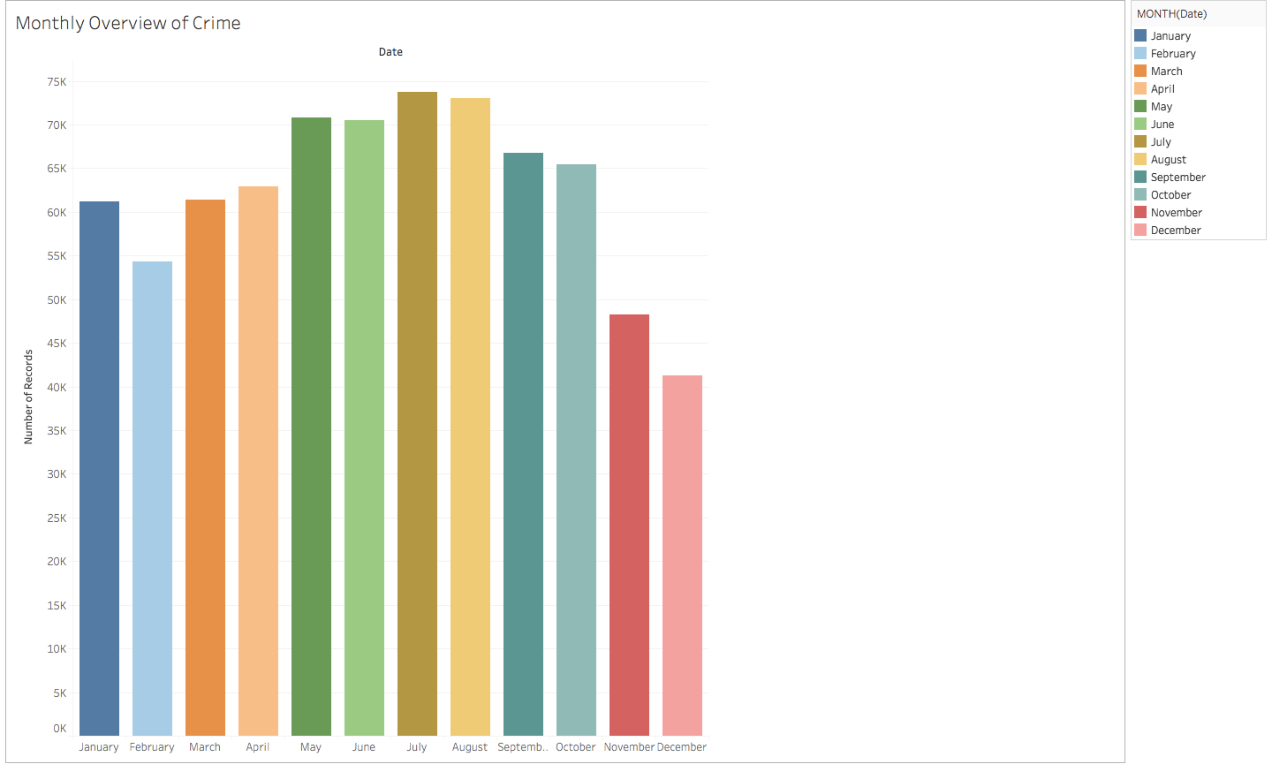
First, our Crimes Since 2017 graph shows a time series of count of crimes committed each month since 2017. The light green portion of the graph represents the proportion of crimes resulting in arrest. The dark green portion represents non-arrests. We can see the overall crimes reported peaking in the warmer months, July and August, and overall crimes hitting a low point each February across the almost 3-year span.



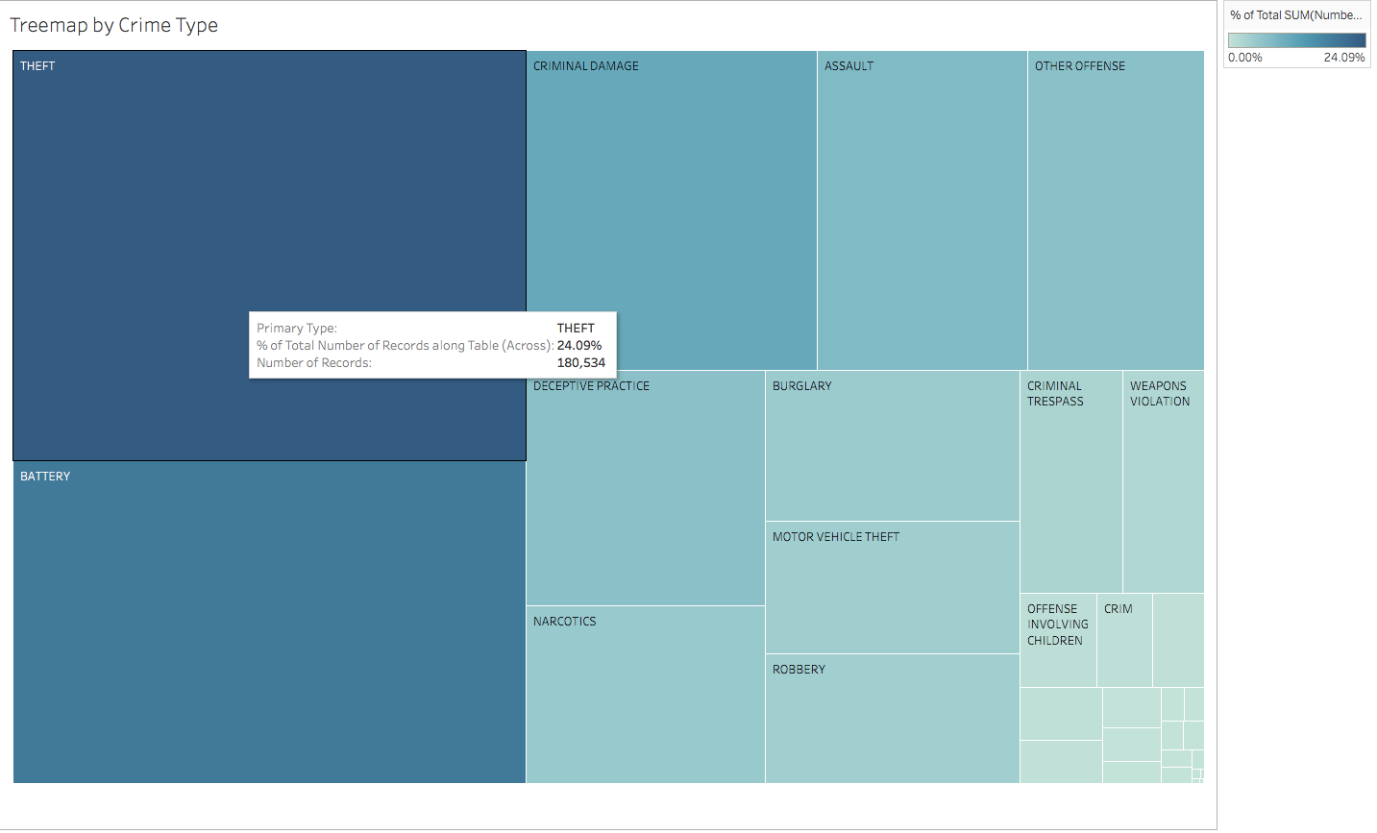
We can see that the number of arrests monthly hover around 4,000 and do not fluctuate seasonally. This is not necessarily a bad thing; we anticipate that this could be due to the set number of officers active in the police force. With this graph, we can also filter on several dimensions of the dataset. Using the Location Description filter, we look at a time series representation of outdoor and indoor crimes to exacerbate the effects that weather can have on criminal activity and arrests.



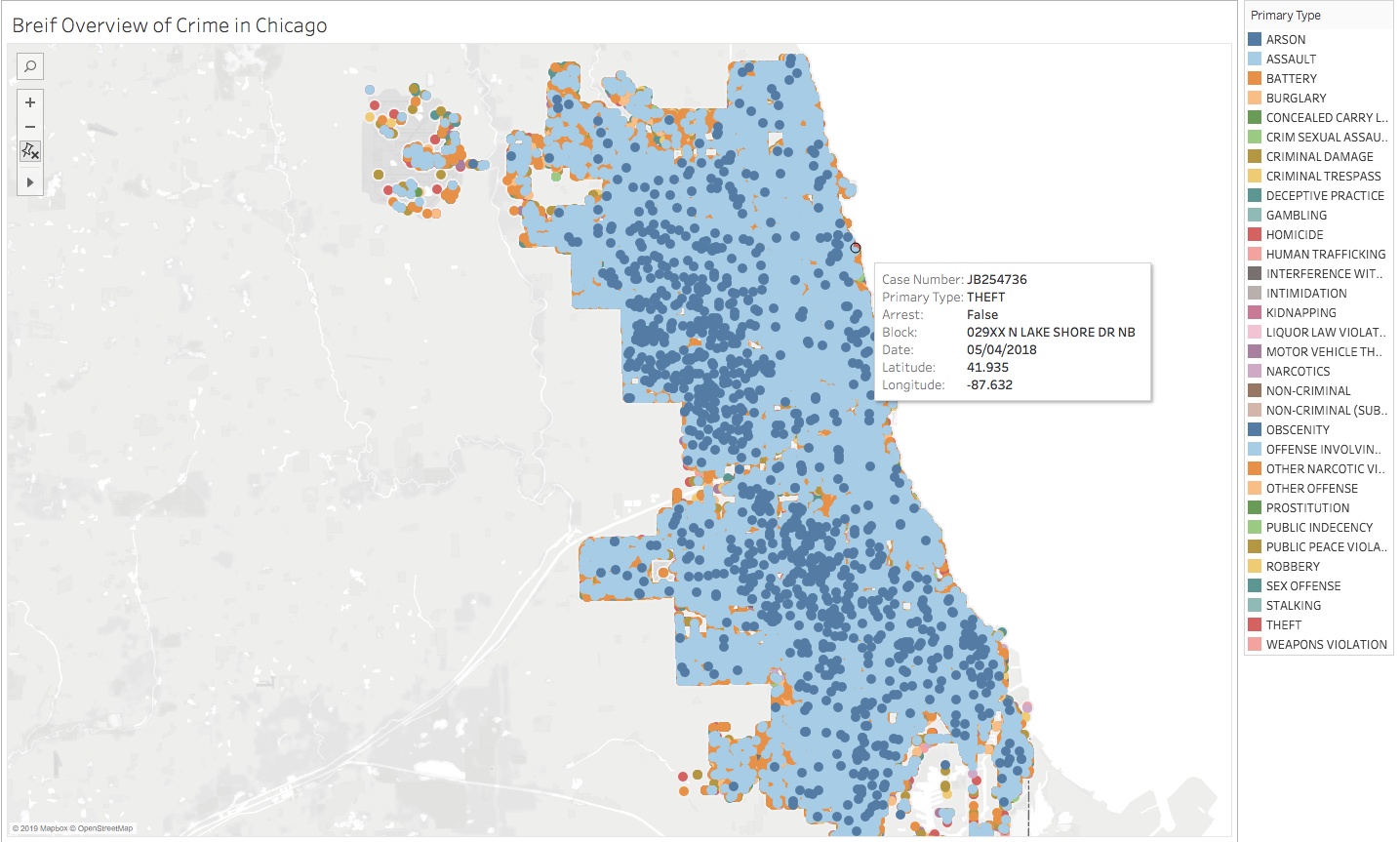
In the next visualization, we depict the total monthly overview of crimes for the years 2017, 2018 and 2019. It can be clearly observed that during the month of July, the number of crime incidents appear to be higher and the crime incidents appear to be more during the mid-year rather than holidays.



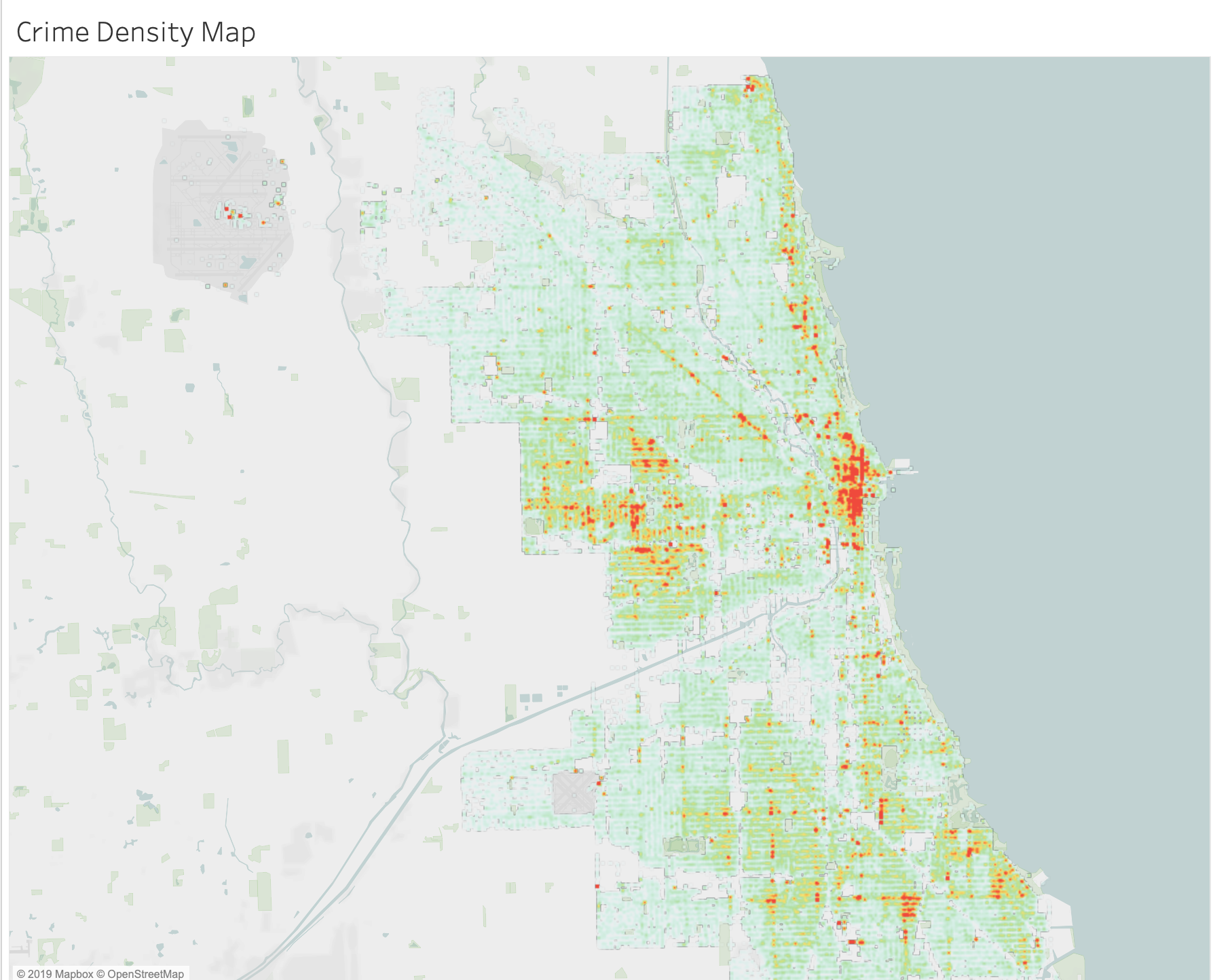
Next, the tree map distribution by crime type depicts the primary type of the crime along with its distribution and percentage of the incident. It can be seen that crime type “Theft” is found to be around 25% which is distributed more than other associated crime types in the city.



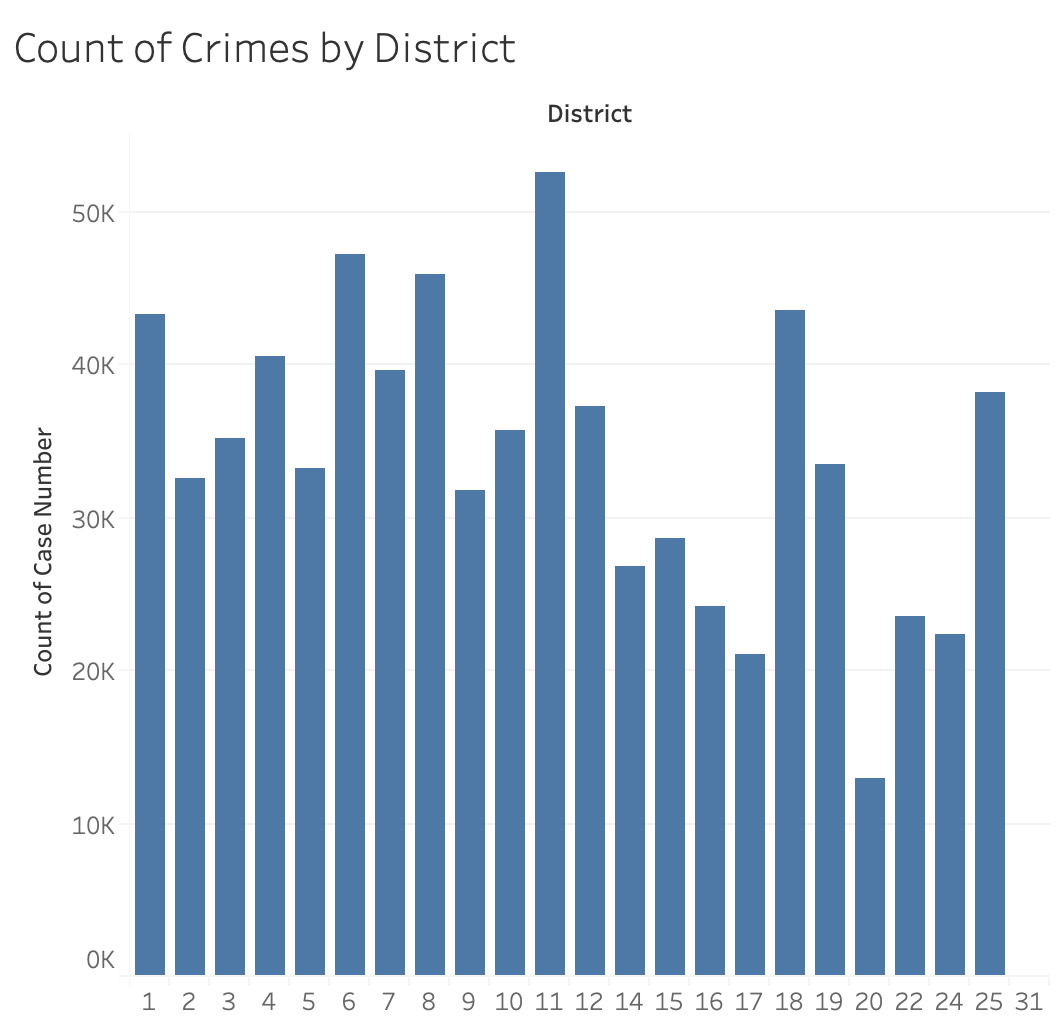
The next visualization depicts the “Brief overview of Crime in Chicago” which is a map visualization which actually provides a brief data regarding the crime type and its location with the date of incident and the arrest status which helps to understand the proportion of arrest for different types of crimes.



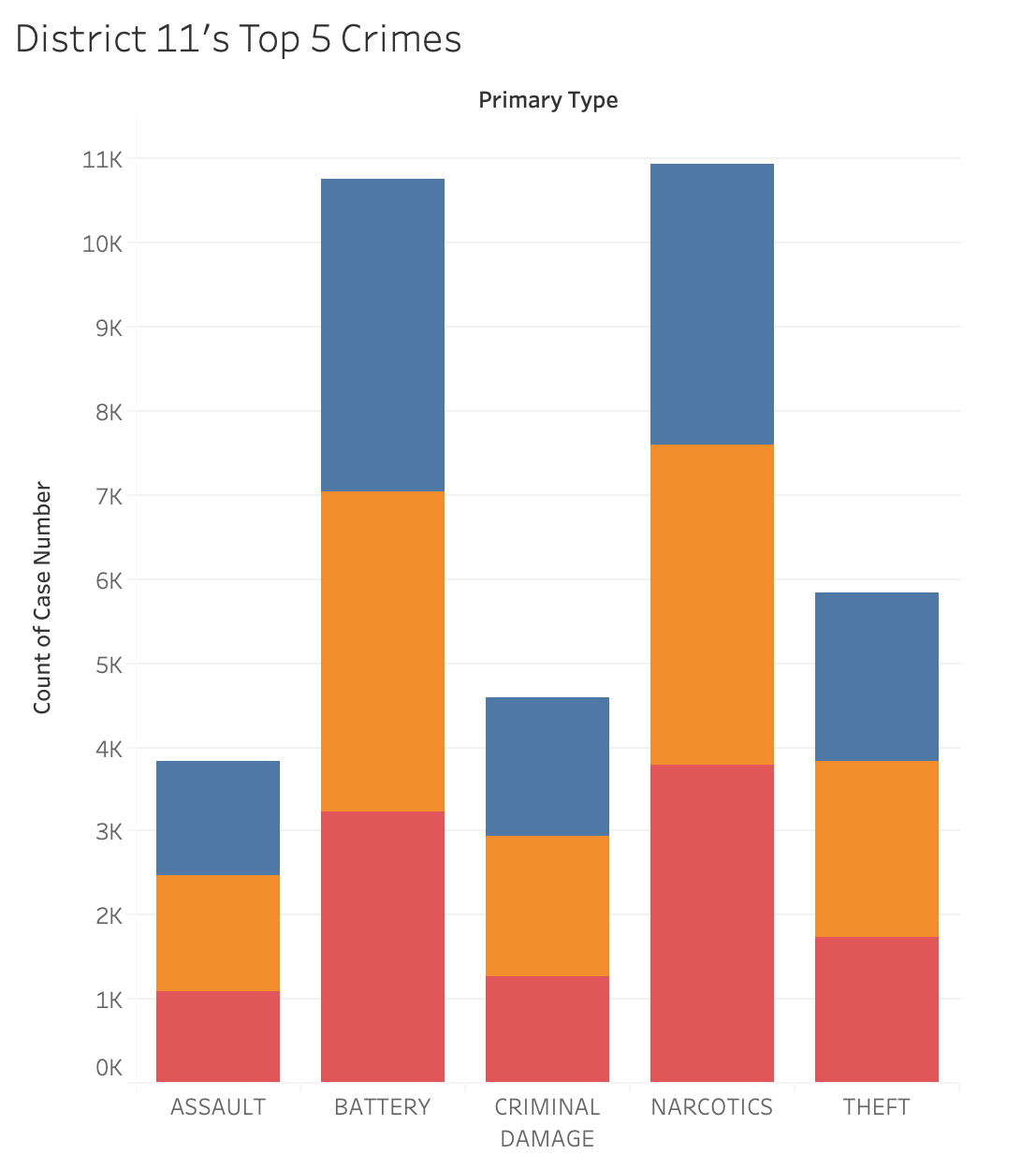
From our analysis in Tableau, we then look at the overall density map taking into account all crimes.



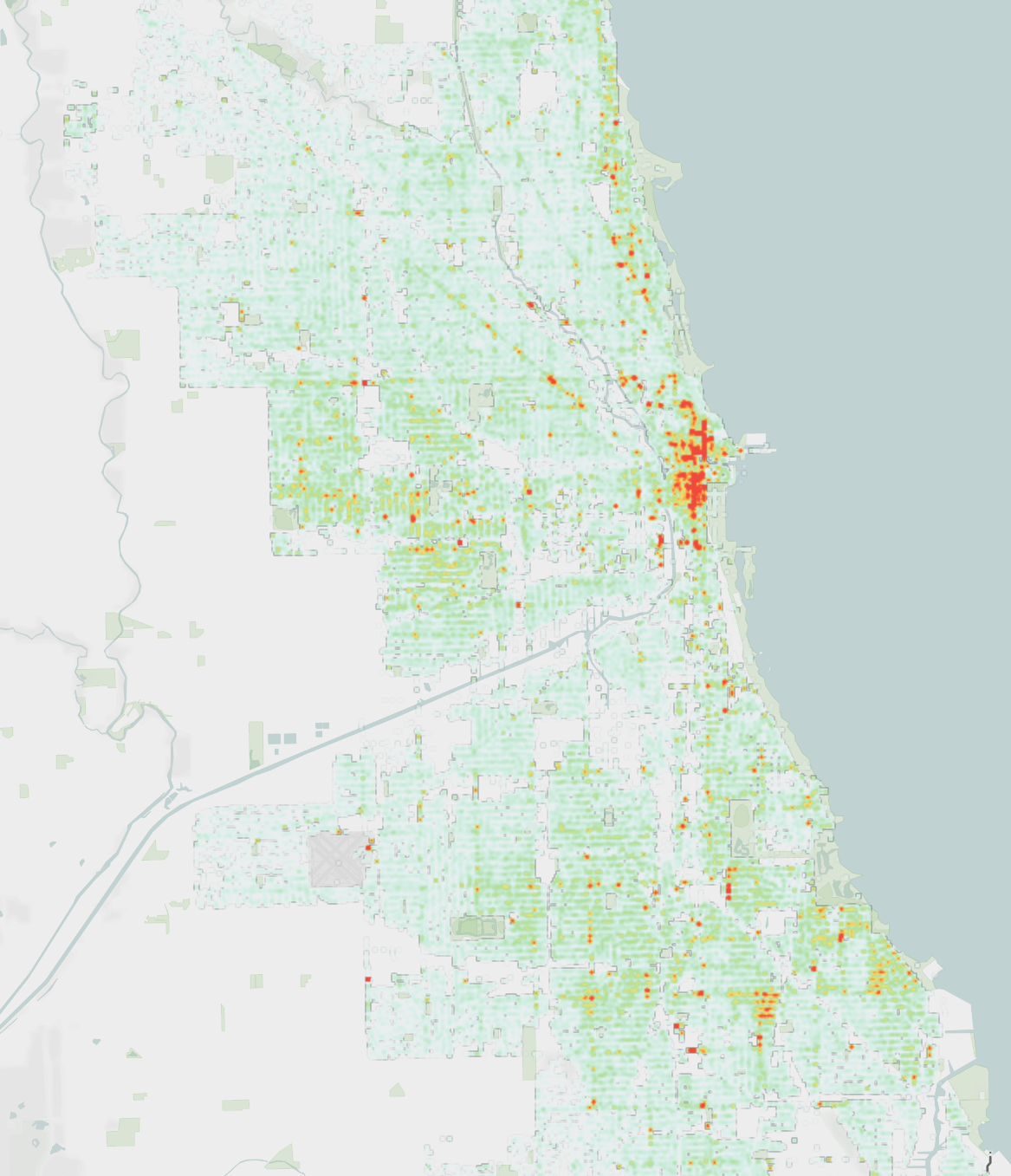
With the Crime Density Map, we see consistency in theft density compared to the density of overall crimes in districts 11, 10, 15 and 25, but we do not see dense overall criminal activity in districts 19, 20, 22, and 24, areas where theft is extremely prevalent. The below histogram represents the counts by district to support the density map. Now, we look at the overall crimes reported by district.



Knowing that district 11 is our top hit for count of overall crimes, we should break down this area’s top 5 crimes in order to review what should be addressed.



Last, we compare outdoor (left) and indoor (right) crime densities, we see a large disparity in the areas crimes are taking place. There is consistent density in the city center, districts 8 and 1, but an alarming rate of outdoor crimes are taking place in districts 11, 10, 15 and 25.



**Insights and Recommendations**

We answered our proposed questions through extensive exploratory statistics, time series, and geographical analysis. We derive our key insights and recommendations based upon our findings.

Insights:

* Crimes without an arrest are fluctuating seasonally, but crimes with arrests are remaining steady. Outdoor crimes fluctuate with great seasonality compared to indoor crimes.
* Crime type “Theft” is the most occurring crime in Chicago, out of which one third of the incidents have resulted in arrest.
* Crime incidents occur more in the Summer season, as it is one of the best times to visit Chicago to enjoy the events and festivals, this is one of the statistics that need to be analyzed and need to be considered.
* The most popular hour for any crime is 12pm. This is also true for the most frequent crime, theft.
* Narcotics crime reports in district 11 have been increasing in numbers since 2017

Recommendations:

* Since the counts of reported crimes are peaking in the warmer months, and the proportion of arrests are lowest in the warmer months, we want to make recommendations about tackling the peaking outdoor crimes. Since the police force will not likely increase the number of officers in the summer, they can distribute their force better based on the density of outdoor crimes. They should also implement more emergency phone booths (blue lights) in the outdoor crime-dense areas, districts 1, 8, 11, 10, 15 and 25. Here is one emergency phone booth and security producer from nearby Wisconson: Rath Security <https://www.rathsecurity.com/emergencyphone.html>.
* Since district 11 is seeing an increase in narcotics related crimes, the city should invest in drug prevention and rehabilitation programs in this district, or at least review to see if their programs have lost effectiveness over the past 3 years.
* Since the frequency of crimes peaks at 12pm central, the police force should research if this is something they can control or if the time is coincidental. For example, they can check to see if 12pm is a common time for police shifts ending and beginning, and verify that they are fully staffed at this time.

**References**

Crimes - 2001 to present. (n.d.). City of Chicago. Retrieved November 12, 2019, from <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2?category=Public-Safety&view_name=Crimes-2001-to-present>.

**Appendix I**

