Crime Data Report

Suki Iska, Manoj Bandi, Nikhil Chhatre, Pratyush Pandey

We decided to take a look at Chicago crime data from 2017 through 2022 using a dataset from Chicago PD. Crimes - 2001 to Present | City of Chicago | Data Portal

Motivation:

As we know, Chicago is of the largest cities in the world. It is known as windy city. Not only It is famous for its food and skyline but also it is one of the cities in USA with highest crime rates. Crime is a significant social problem in the nation, affecting public safety, child development, and adult socioeconomic status. Chicago has always been a holiday destination for international tourists and top pick for international students to pursue their studies. Safety is the primary concern while staying in a new city, miles away from home. The motivation behind taking up this topic is to use predictive techniques to assess the safety of a place and feel safe in our neighborhood.

A police officer may know where the dangerous/unsafe areas are according to his experience, but he may not be able to tell about the type of crime that can happen and when it can occur. By performing EDA we will be cleaning the data and make it more feasible to use. We will be performing different kind of visualizations to show how the trends and will derive useful insights. These insights will help the tourists, students as well as the citizens of Chicago to become aware of what type of crimes happening where and at what time, so that they can take precautionary measures.

We will be approaching the analysis using 3W approach. Where we analyze the data and develop the visualizations in the order of what type of crimes are happening, when are those crimes happening and where are they happening.

Exploratory Data Analysis (EDA):
EDA of the dataset includes initial analysis of the underlying raw data to identify any patterns, anomalies and duplicates which may have an influence on the next stage of our visualization project.

We start by formatting the column headers so as to standardize and make the handling easier. This was achieved using the following code to remove any leading or lagging blank spaces, replace commas and blanks and convert all headers into lowercase letters.

```
In [4]: #Handling any inconsistensis of column names

        crimes_data.columns = crimes_data.columns.str.strip()
        crimes_data.columns = crimes_data.columns.str.replace(',', '')
        crimes_data.columns = crimes_data.columns.str.replace(' ', '_')
        crimes_data.columns = crimes_data.columns.str.lower()
```

Dataset characteristics: This gives us a brief summary of the dataset which was extracted from Chicago PD webpage.

- Shape : 1413406 rows and 22 columns
- Size : 31094932 (which is product of rows & columns)

Above attributes are interrelated and help us to measure the volume of the dataset that needs to be processed. Also this helps us to compare in the latter stage after EDA to measure if dataset size was reduced which will help in improving the process performance.

- Info : Summary of column list and associated data type

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1413406 entries, 0 to 1413405
Data columns (total 22 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   id                    1413406 non-null  int64
 1   case_number           1413406 non-null  object
 2   date                  1413406 non-null  object
 3   block                 1413406 non-null  object
 4   iucr                  1413406 non-null  object
 5   primary_type          1413406 non-null  object
 6   description           1413406 non-null  object
 7   location_description  1407133 non-null  object
 8   arrest                1413406 non-null  bool
 9   domestic              1413406 non-null  bool
 10  beat                  1413406 non-null  int64
 11  district              1413405 non-null  float64
 12  ward                  1413356 non-null  float64
 13  community_area        1413405 non-null  float64
 14  fbi_code              1413406 non-null  object
 15  x_coordinate          1392285 non-null  float64
 16  y_coordinate          1392285 non-null  float64
 17  year                  1413406 non-null  int64
 18  updated_on            1413406 non-null  object
 19  latitude              1392285 non-null  float64
 20  longitude             1392285 non-null  float64
 21  location              1392285 non-null  object
dtypes: bool(2), float64(7), int64(3), object(10)
memory usage: 218.4+ MB
```

Next we checked for any data inconsistencies in the data in the form of duplicates, null values and removing extraneous columns in the dataset.

- Duplicate data : we found no duplicates in the dataset. Hence no action required.

```
In [8]: #Check the data forany duplicates

        crimes_data[crimes_data.duplicated(keep=False)]
```

Out[8]:

id	case_number	date	block	iucr	primary_type	description	location_description	ar

0 rows × 22 columns

- Extraneous columns : 12 columns were dropped which have no value addition in the analysis. This will help to reduce the data size. This is based on the data dictionary and listed below.

Columns excluded: 'id', 'case_number', 'block', 'iucr', 'beat', 'ward', 'community_area', 'fbi_code', 'x_coordinate', 'y_coordinate', 'updated_on', 'location'

Column_Name	Include	Comment	Type	Description
ID	No	Detail not requir	Number	Unique identifier for the record.
Case Number	No	Detail not requir	Plain Text	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
Date	Yes		Date & Time	Date when the incident occurred. this is sometimes a best estimate.
Block	No	Detail not requir	Plain Text	The partially redacted address where the incident occurred, placing it on the same block as the actual address.
IUCR	No	Duplicate	Plain Text	The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e.
Primary Type	Yes		Plain Text	The primary description of the IUCR code.
Description	Yes		Plain Text	The secondary description of the IUCR code, a subcategory of the primary description.
Location Description	Yes		Plain Text	Description of the location where the incident occurred.
Arrest	Yes		Checkbox	Indicates whether an arrest was made.
Domestic	Check	Detail not requir	Checkbox	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
Beat	No	Detail not requir	Plain Text	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. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74.
District	Yes	Detail not requir	Plain Text	Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.
Ward	No	Detail not requir	Number	The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.
Community Area	No	Detail not requir	Plain Text	Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.
FBI Code	No	Duplicate	Plain Text	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.
X Coordinate	No	Duplicate	Number	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
Y Coordinate	No	Duplicate	Number	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
Year	Yes		Number	Year the incident occurred.
Updated On	No	Detail not requir	Date & Time	Date and time the record was last updated.
Latitude	Yes		Number	The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Longitude	Yes		Number	The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
Location	No	Duplicate	Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

- Null data : We checked for null data and below is the summary where 3 columns have null values.

```
In [10]: # EDA - checking no of null data points in the dataset.

         crimes_data.isnull().sum()
```

```
Out[10]: date                      0
         primary_type              0
         description               0
         location_description   6273
         arrest                    0
         domestic                  0
         district                  1
         year                      0
         latitude              21121
         longitude             21121
         dtype: int64
```

a) latitude and longitude: These columns have about 21k null values. Since we decided not to plot the data on a geographical visualization, this will have no impact on our analysis. Hence no further action required to handle these null values.

b) Location_description: This column has 6273 null values and we further analyzed to identify what types of crimes are associated. Related code is on line 11. Output as below.

Out[11]:

primary_type	description	district
ARSON	BY FIRE	1
BATTERY	AGGRAVATED - HANDGUN	1
BURGLARY	FORCIBLE ENTRY	2
CRIMINAL SEXUAL ASSAULT	NON-AGGRAVATED	2
DECEPTIVE PRACTICE	AGGRAVATED FINANCIAL IDENTITY THEFT	2
	BOGUS CHECK	3
	COMPUTER FRAUD	8
	COUNTERFEIT CHECK	3
	COUNTERFEITING DOCUMENT	3
	CREDIT CARD FRAUD	32
	FINANCIAL IDENTITY THEFT $300 AND UNDER	1034
	FINANCIAL IDENTITY THEFT OVER $ 300	5099
	FORGERY	5
	FRAUD OR CONFIDENCE GAME	40
	ILLEGAL POSSESSION CASH CARD	1
	ILLEGAL USE CASH CARD	22
	UNLAWFUL USE OF A COMPUTER	6
OTHER OFFENSE	OTHER CRIME AGAINST PERSON	1
THEFT	$500 AND UNDER	3
	OVER $500	3
	POCKET-PICKING	1
	RETAIL THEFT	1

6207 are associated with deceptive practice of which about 6000 are associated with Identity theft. Since 6200 rows is immaterial compared to the total dataset of 1.4M we decided to retain the data for the analysis with no additional action to be taken.

Data set after EDA:

- Shape : 1413406 rows and 10 columns
- Size : 14134060 (which is product of rows & columns)

We see a significant reduction in the data size from 31094932 to 14124060, which is a 54% reduction (16960872).

Analysis of the current year 2022:

In first part of analysis, we look at 2022 dataset to identify any trends and major observations which will help the citizens. For this we created a dataframe with data limited only to the year 2022 from the entire dataset.

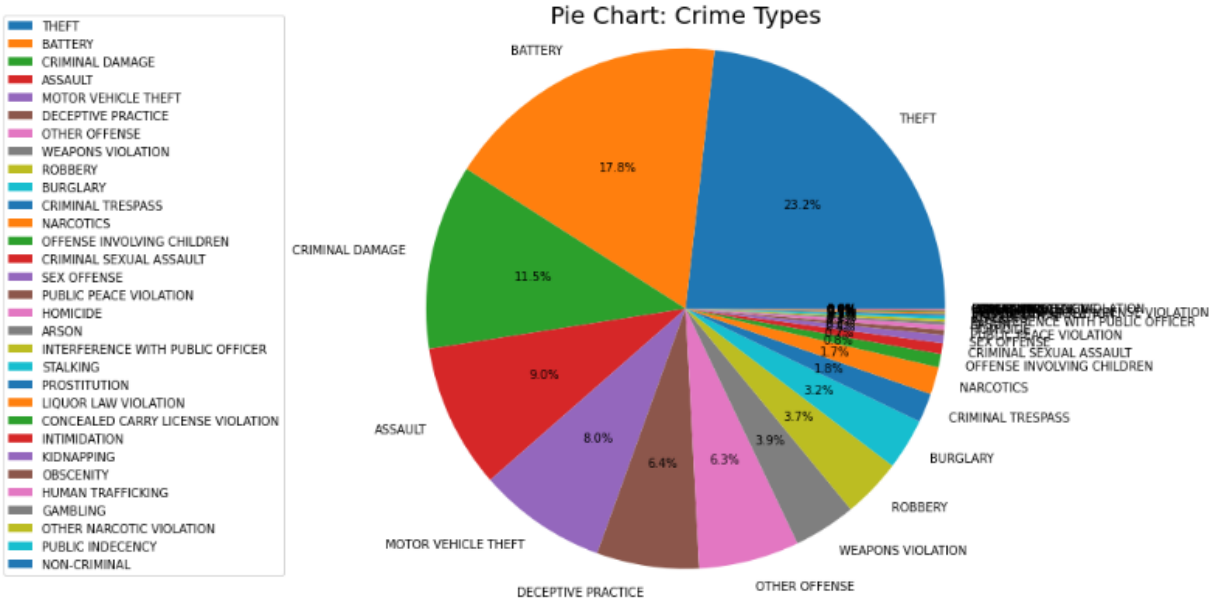
As stated in the introduction, we have performed analysis in 3 parts:

1) What : what are the types of crimes in Chicago
2) Where : where are the crimes taking place, are there any particular districts, locations which are riskier compared to others
3) When : Is there any specific time or season when crimes are more prevalent than others

1) **What**: Types of Crimes that commonly occur across Chicago, what proportion of the toal they are. This helps us to educate the readers of what type of crimes citizens should be vigilant of and accordingly they can prepare themselves.

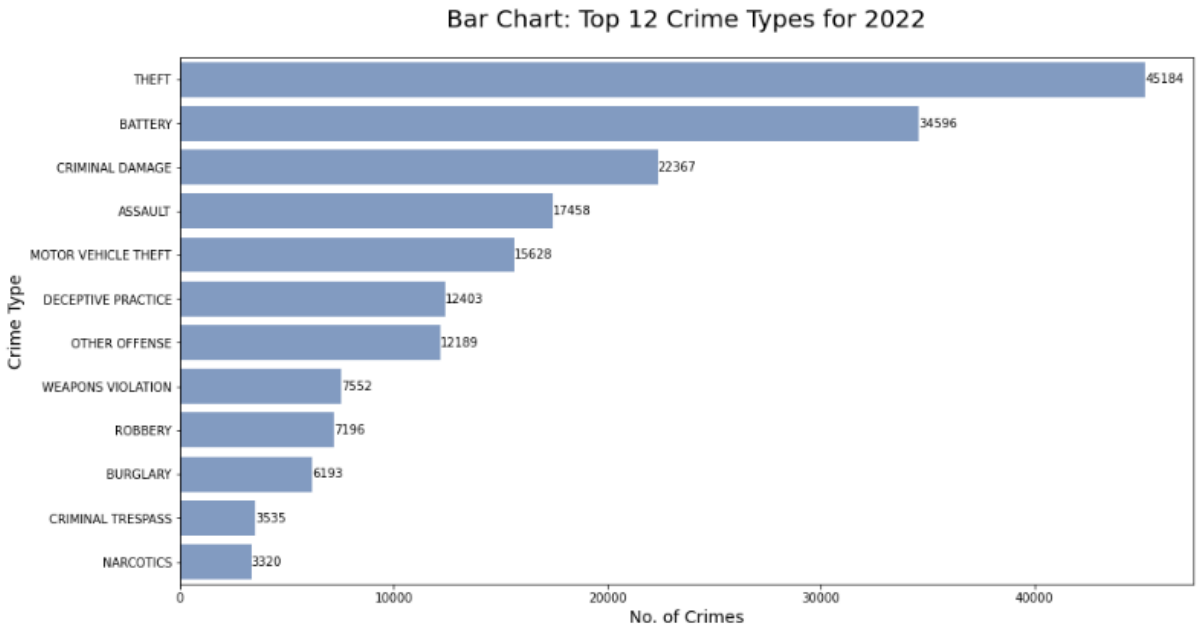
Initially we looked at all the types of crimes and we see that crimes are broadly classified into 31 types. But this resulted in a poor visualization as only 12 crime types have percentages > 1% while the rest of 19 represent insignificant proportion of the total crime data for 2022.

Related code for the below pie chart is on line 16.



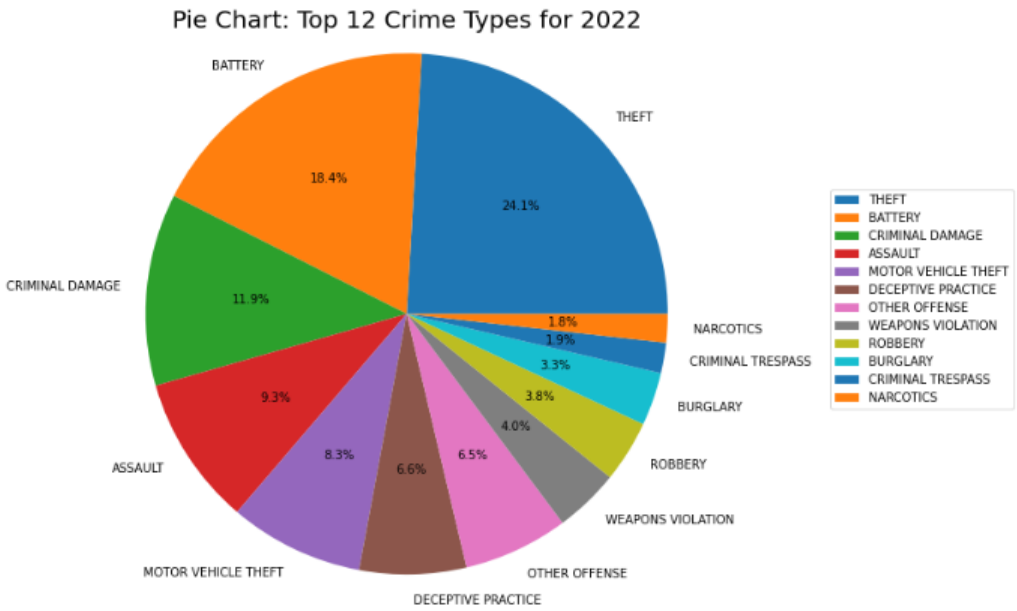
As seen in this pie chart crimes with percentages lower than 1% are overlapping and this is a poor visualization. We need to be cognizant that while visualization is a powerful tool, if we don't apply the right controls it may result in a poor outcome

Hence we decided to apply a threshold of 1% for “Types of Crimes” for clear and meaningful visualization.



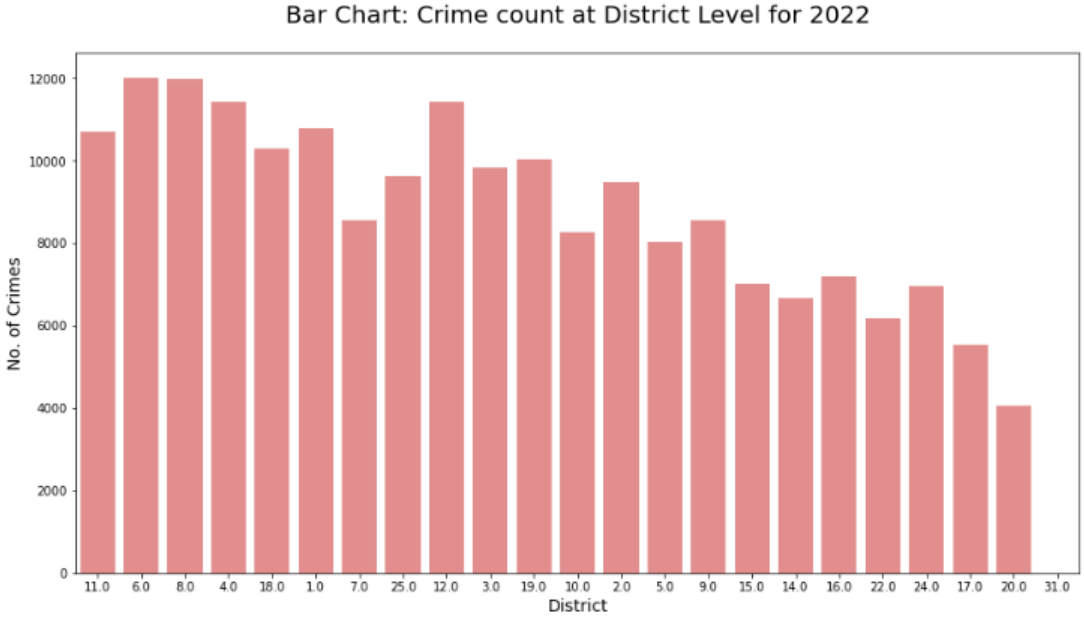
Above visualization we are looking at the count of each of the types of crimes in 2022. Theft, Battery and Criminal Damage are top 3 crimes along with the number of crimes recorded in 2022. This shows that Theft and Battery are our highest, almost competing with one another. Battery is the unlawful application of force on another person or their belongings causing bodily injuries or offensive contact. Essentially this is when someone has committed real physical harm to someone. Going back to our previous visualization, we noticed that most crimes occur in the daytime, and battery and theft are both crimes that can be frequently committed during the daytime, as such is shown in the news or in our phone's emergency alerts.

But this does not provide a comparison across crimes and their proportion in the totality. This can be better viewed using the below Pie chart, which provides a percentage proportion in the total number of crimes reported.

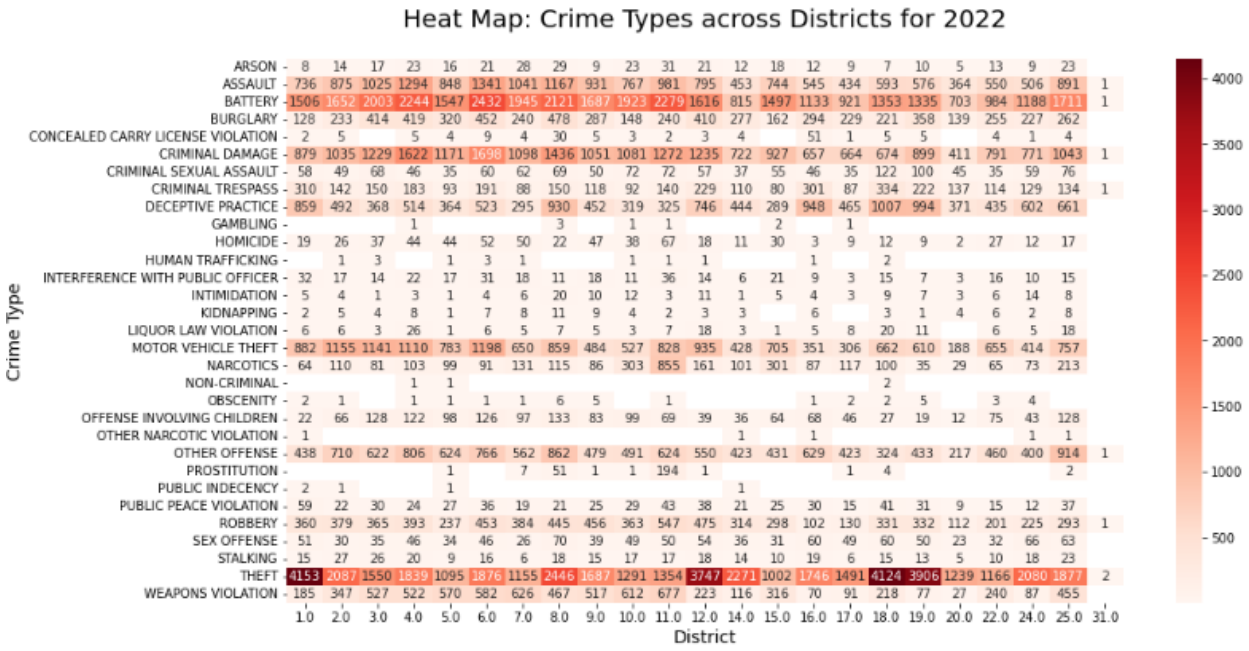


2) **Where**: At what locations are the crimes taking place. We can uncover this looking at the data at District level, locations and a heatmap of Districts and Crime types.

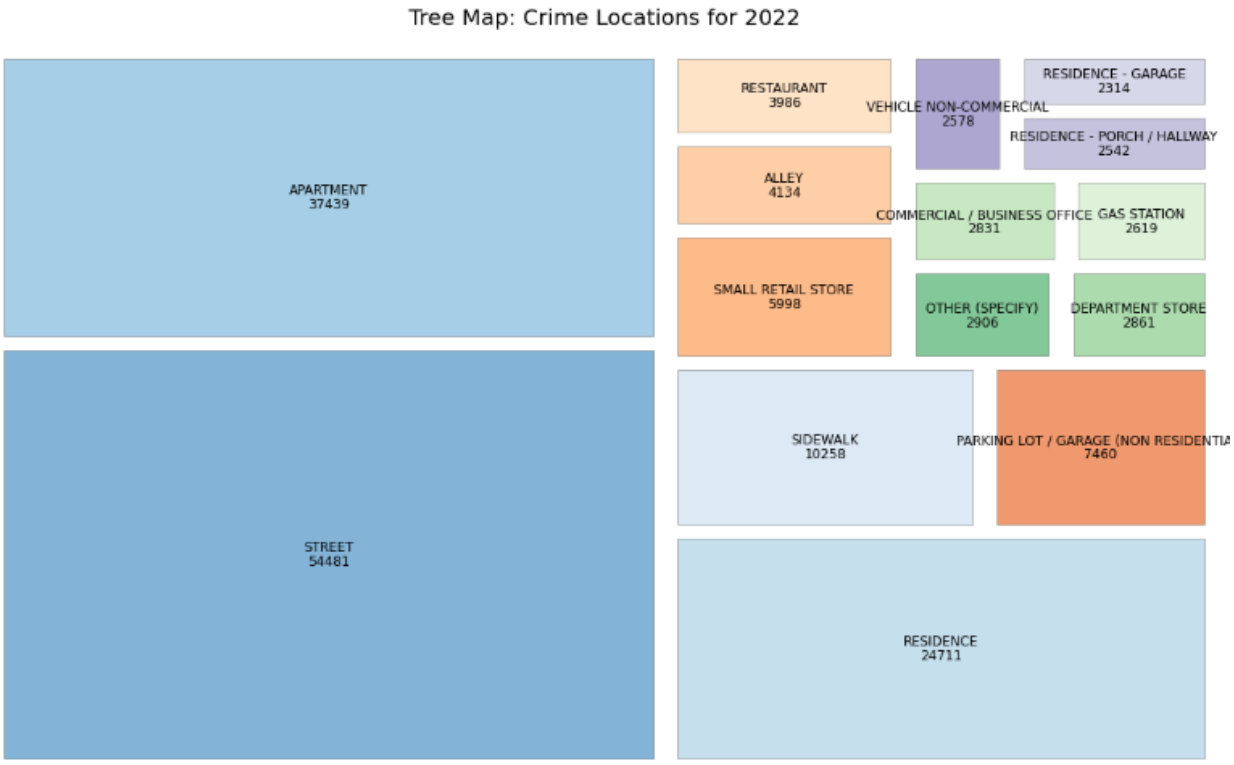
Below visualization shows us the number of crimes per district in Chicago. 6 and 8 are the most affected with 18 and 12 being the next highest. Anyone should be well aware of their surroundings no matter what district they are in but they should be informed of what districts may have more crimes. Below for our sixth visualization we have a heat map of a more detailed view of crimes across all districts.



Barchart is helpful to compare crimes across districts in totality, but it doesn't provide a breakdown of crimes. Breakdown of crimes across districts will help the citizens to identify which districts are prone to what type of crime to take the required precautions. For this purpose we plotted a Heatmap as shown in the next visualization. Heatmap helps to both breakdown the numbers in a grid format and more importantly identify the major crimes based on the color hue. In this case the higher intensity of red hue, the higher is the number of crimes for the given crime type and district combination.

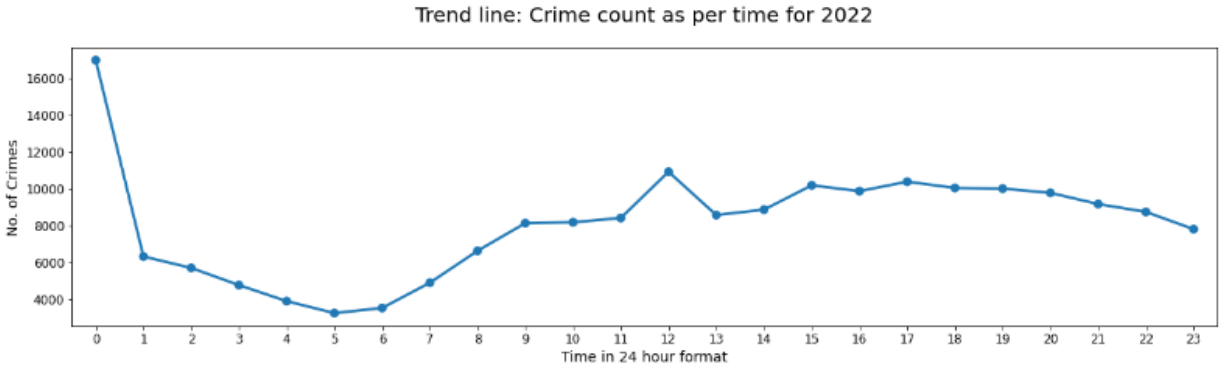


While the above 2 visualizations help to uncover the districts and crime types, we would like to also check on the locations. Which locations have the highest reported crimes so this prepares the reader and citizens to be aware and be on higher alert levels. While there are more than 280 location descriptions reported, we focus on the top 15 in the Treemap to ensure we focus on the locations with the highest frequency.



Notice how the majority of these crimes occur primarily in common places where people tend to be more relaxed and not aware of their surroundings. So it makes sense why people with bad intentions would take advantage of us when we are at our most relaxed state of mind. This tells us to be aware of anything that can seem suspicious. Make sure to observe if anyone is trying to follow you home, or always look through the peephole before opening the door to any visitors.

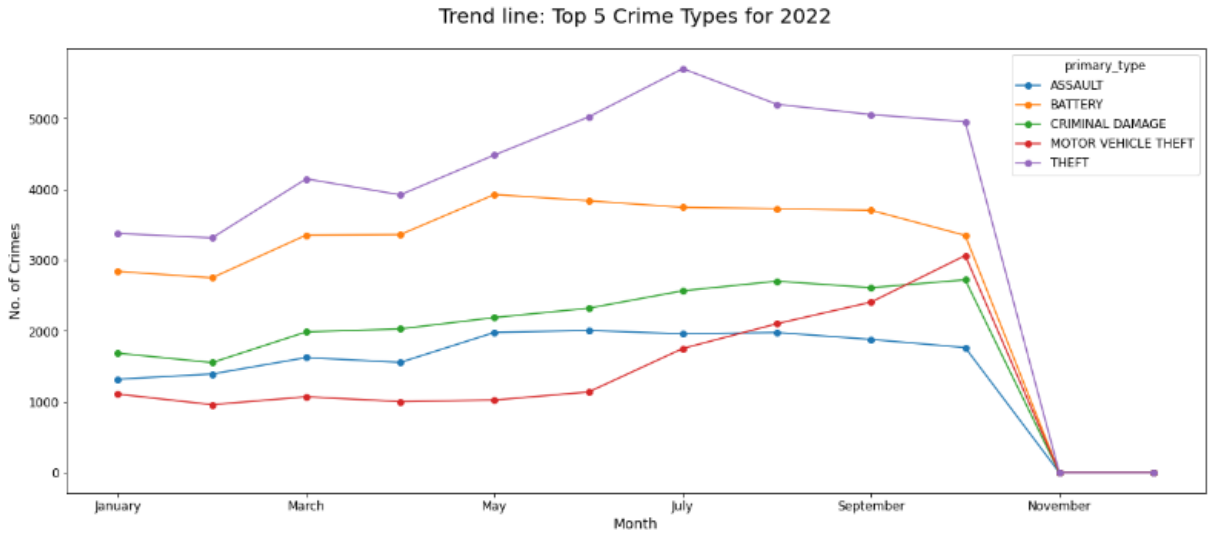
3) **<u>When</u>**: In this part we check if there is any particular time of the day which has higher crimes reported and also if there is any seasonality in the crimes.



Above visualization shows us the number of crimes per district in Chicago. 6 and 8 are the most affected with 18 and 12 being the next highest. The time is formatted in military time so 1200 = 12am, and 1300 = 1pm. We can see that surprisingly, the majority of crimes occur from around

9 am to the afternoon around 1 or 2 pm, which is strange since we assume that most crimes occur at night. However, this doesn't mean that you should let your guard down, instead make sure to be aware of your surroundings during all times of the day. Anyone should be well aware of their surroundings no matter what district they are in but they should be informed of what districts may have more crimes.

Below we focus on top 5 crimes and if there is any seasonality. It can be seen that crimes increased during Summer and peaked in July and then tapered down gradually. There is a sudden drop after October, as we have taken crime data reported until October end for our analysis.



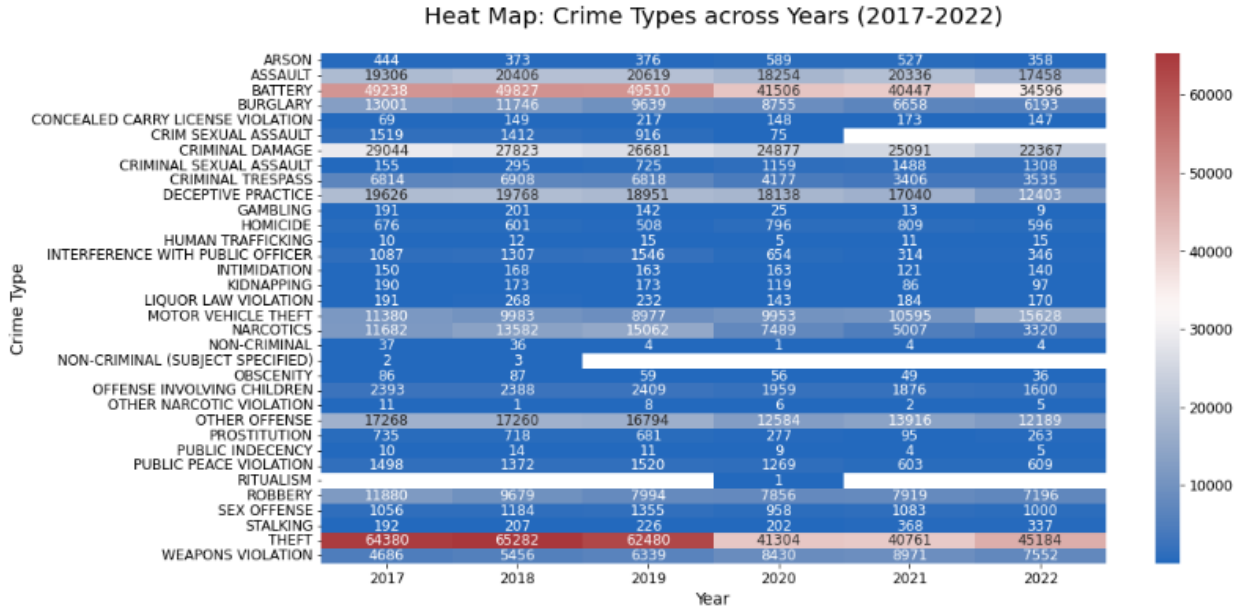
Year on Year Analysis across 6 years:

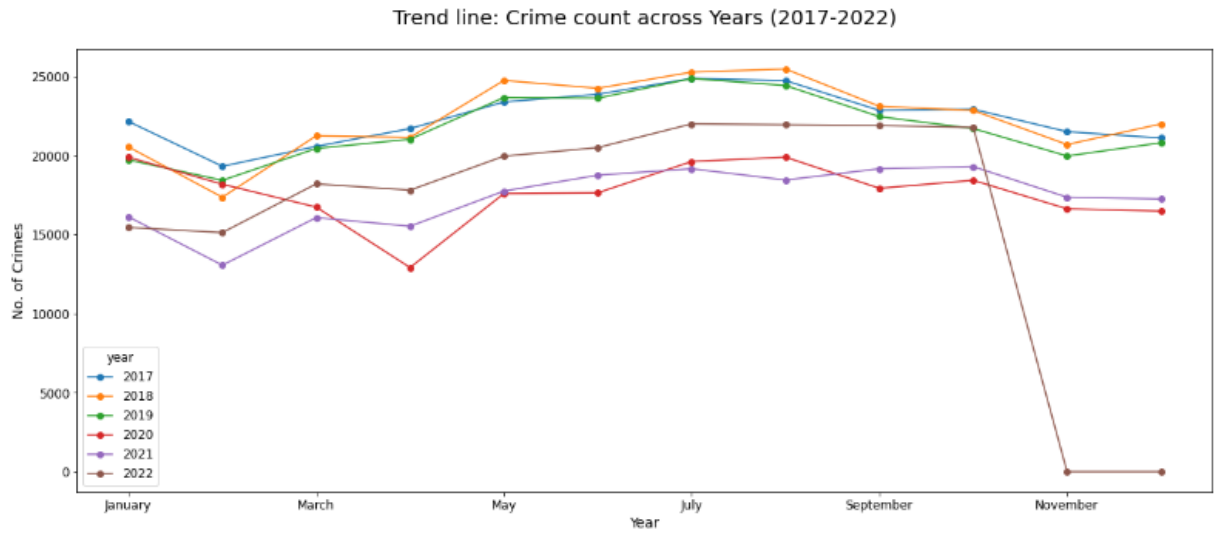
For the second part of the analysis,we have considered data for 6 years instead of the usual 5 years as we wanted to cover 3 periods which are:

- Pre Covid : 2017 to 2019 (3 years)
- Covid : 2020 and 2021 (2 years)
- Post Covid : 2022 (current year)

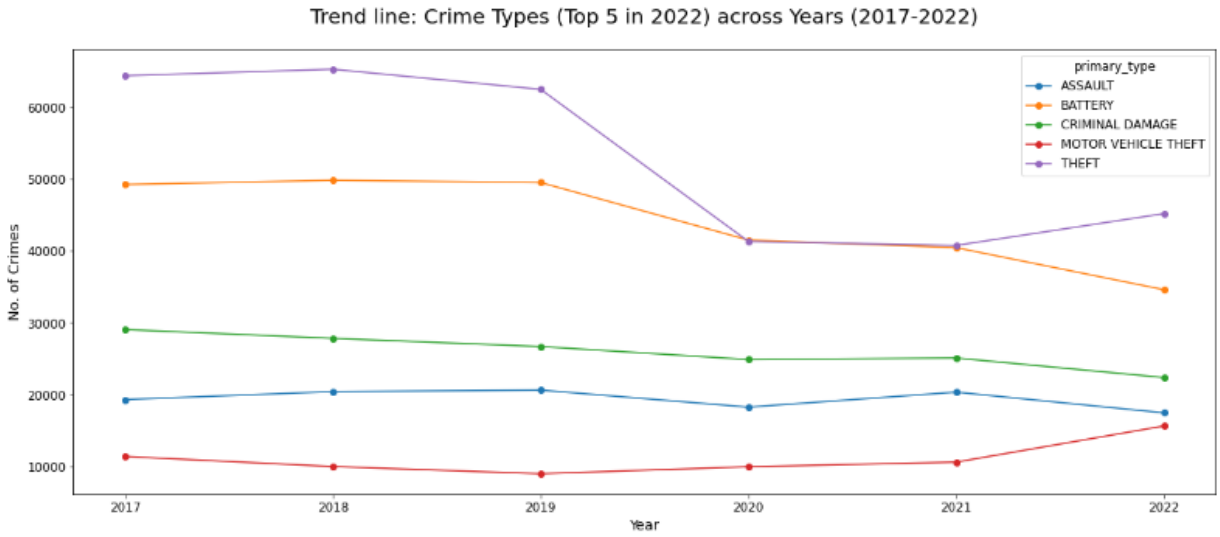
This will help us to identify if there are any trends in terms of total crimes reported across years, any change in terms of increase or decrease in any particular crime type and if seasonal variations are similar or any changes took place across these 3 periods.

We begin with a Heatmap of Crime Types across the entire period of 6 years. It shows that Theft has the highest value across the years with a dark red hue. Though it steadily decreased from 2017 to 2019 and was lowest during the COVID period (almost 50% of the Pre Covid period), it is gradually increasing with 45K crimes reported in 10 months of 2022. On the other hand, 2nd highest crime in 2022, Battery has shown consistent decrease across 2017 to 2022. Crimes like Heatmap is a powerful tool to compare entire data set but it is not useful to compare total numbers across Crimes or Years.

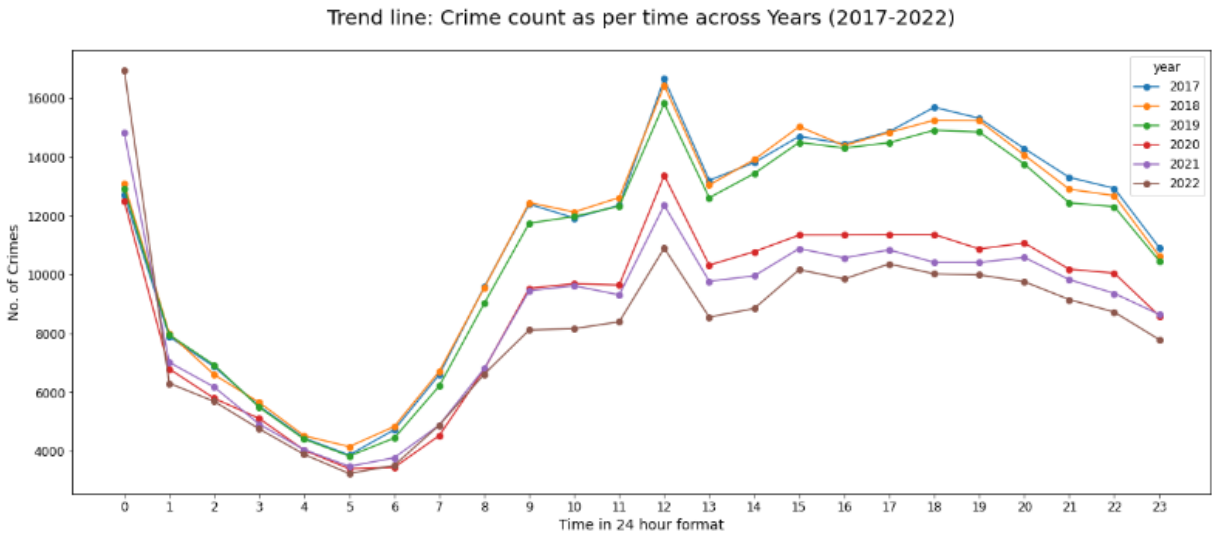




Above visualization is a snapshot of the total crimes each year (2017 to 2022) reported across the months. This gives us an insight on the seasonality and if there is a common theme. It can be seen that there is an upward trend in crimes across years during holiday seasons namely Summer and Winter break. Also this shows that during COVID total crimes dipped substantially and are the lowest during the 6 years. With 2022 when people have started to step out more for work, recreation and vacations the total number of crimes for each month are higher than COVID period (2020 and 2021).



Above visualization shows us the crime count trend line for top 5 crimes across the years. The trend line for theft is decreasing during covid and rising again after covid. Battery is decreasing and motor vehicle theft is rising in 2022.



Above visualization shows us the crime rate in 24 hour fromat for different years. If we can observe , we can see that crime is at its peak in mid noon and has a high rate from 3pm to 8 pm and after that there is gradual decrease in crime rate till early morning.Crime is at its lowest at 5AM. Here we can see that the trend line for 2022 is below the trend lines of every other year. But in the previous visualizations, the 2022 crime rate is higher than both 2020,2021. This might be because of the previously explained anomaly of data entry where data which has no time stamp directly goes to 0000 time stamp. So, in 2022 the no.of data entries with no time stamos be high and so most of the count is distributed to 00:00 timestamp and only remaining data is distributed across the other time stamps.

By this we can say that we cannot derive a conclusion just by looking at a single visualization. Both multiple visualizations and context plays to derive a conclusion accurately.

Conclusion:

By looking at all the visualizations and analysis of the data, we can derive following conclusions

For the most recent data i.e., for the year 2022,

1.Out of all the crimes that are happening in Chicago in the year 2022, only Theft and Battery make up to 40% of total crimes. So, the citizens should be careful and not walk alone at any time of the day.

2. Out of all the districts in Chicago, District 6 & 8 has the highest crime rates, followed by districts 4 & 12. Unfortunately, Our UIC campus and the little Italy where most of the student live at falls under District 12. So, Students must be very cautious all the time.

3. Crime rate in the months June to September, which is summer, has the crime rate. So, we can assume that during the winters, due to the extreme cold and snow there won't be many people roaming outside which further decreases the chances of crimes.

4. Ironically, the probability of crimes happening at mid noon i.e., 12 noon is high. After midnoon, crime rates remain high from 3pm to 8 pm. Surprisingly, crime rate is comparatively low at night and is lowest at early mornings.

For the data across years from 2017 – 2022.

1. Before covid, the crime is a significant crime rate. But during covid there is a drop-in crime rate same as every other field. But after covid not only the people and businesses are getting back to the pre covid times but also the crime rate which slowly rising to pre covid levels.

2. After analyzing different crimes across the years from 2017-2022, we can conclude that crime type theft has upward trajectory. Police should investigate that and implement required protocols to control the rising theft levels.

3. Across the years, the crime rate for crime type 'Battery' has a downward trend. We can assume that that police are successful in controlling the battery.

4. Most surprising trend is of motor vehicle theft. It is at its peak in the year 2022 compared to the previous years. People should be aware of that and should make use of antitheft precautions in their vehicles.