DATS 6401 Final Project - Chicago Crime Analysis

Kyle Hall, Aishwarya Maddula
12/11/2024

Table of Contents

Introduction

- **Topic Selection:**
  We chose this topic and dataset because it provides a rich source of information with thousands of records capturing various crime types, their locations, spatial features and temporal characteristics. The level of detail within the dataset, (especially the geographical latitude and longitude data) also allowed us to experiment with new visualizations in Tableau we hadn't tried before.

  This topic is particularly compelling because crime is a pervasive issue that impacts communities worldwide, influencing public safety, economic conditions, and policymaking. By analyzing this dataset, we aimed to generate actionable insights that could assist a wide variety of stakeholders, including law enforcement, policymakers, and community leaders. These insights could support efforts to create safer neighborhoods, allocate resources more effectively, and improve Chicago's overall quality of life

- **Research Questions:**

  1. What are the overall trends in crime types and their occurrences over time?
  2. How do crime patterns vary spatially and temporally across different regions or neighborhoods?
  3. What are the key factors and trends associated with homicide occurrences, and how do they differ from other types of crimes?

- **Team Members Contribution:**

  - Kyle's tasks included: Visualizations (Seasonal, Homicide), formatting, editing, dashboard creation and written report writing. Work was split evenly across the team.

  - Aishwarya's tasks included: Visualizations (Spatial, Temporal), formatting, editing, dashboard creation and written report writing. Work was split evenly across the team.

Description of the Data

Data Source:

The original dataset can be found here at the Chicago Data Portal. Before our EDA, the original dataset contained 8.21M rows and 22 columns, capturing crime records from 2001 to 2024.

Data Collection Process:

The data was downloaded from and accessed via this link. It contains recorded incidents of crime, including attributes such as type, date, location, and arrest status.

Data Dictionary:

- **Summary**: The original downloaded version contained 8,199,424 rows with 22 unique columns (bool(2), float64(7), int64(3), object(10)). Highlighted names in this Data dictionary are columns we chose to keep for out analysis.
- **ID**: Unique identifier for the record.
- **Case Number**: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
- `Date` : Date when the incident occurred. This is sometimes a best estimate.
- **Block**: The partially redacted address where the incident occurred, placing it on the same block as the actual address.

- **IUCR**: The Illinois Uniform Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes here.
- `Primary Type`: The primary description of the IUCR code.
- `Description`: The secondary description of the IUCR code, a subcategory of the primary description.
- `Location Description`: Description of the location where the incident occurred.
- `Arrest`: Indicates whether an arrest was made.
- `Domestic`: Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
- **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. The Chicago Police Department has 22 police districts. See the beats here.
- **District**: Indicates the police district where the incident occurred. See the districts here.
- **Ward**: The ward (City Council district) where the incident occurred. See the wards here.
- `Community Area`: Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas here.
- **FBI Code**: 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 here.
- **X Coordinate**: 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**: 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`: Year the incident occurred.
- **Updated On**: Date and time the record was last updated.
- `Latitude`: 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`: 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**: 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.

Noteworthy Features:

- This data is open source and free to download, it requires no permissions to access and is downloaded to CSV
- We considered the data source to have a high level of credibility, and to be an "authoritative" dataset for crime as it resides publicly on the City of Chicago Open Data Portal and is compiled from police reports and updated regularly by the Chicago Police Department
- The data is partially redacted for privacy, with block-level precision for addresses

---

Analysis of Data Quality

We downloaded the entire dataset, totaling 1.80GB, to ensure we had a comprehensive source for our data selection and pre-processing. From the outset, our goal was to reduce the dataset to a more manageable size while maintaining its relevance regarding our research questions. This process involved careful consideration of the dataset's completeness, accuracy, and consistency:

- The original dataset spans over two decades and includes over 8 million records, providing an extensive and detailed view of crime trends in Chicago.

- To maintain completeness, we focused on retaining data that aligned with our research objectives, such as crime type, date, and geographic location. We excluded columns that were less relevant or redundant, ensuring we preserved key dimensions of the dataset. The columns we dropped include: `ID`, `Case Number`, `Block`, `IUCR`, `Beat`, `District`, `Ward`, `FBI Code`, `X Coordinate`, `Y Coordinate`, `Updated On`, and `Location`.

- The dataset is sourced from the Chicago Data Portal, a credible source managed by the city’s government. However, certain fields, such as `Latitude` and `Longitude`, are partially redacted to protect privacy, which introduces a minor margin of error for precise spatial analyses.

- During pre-processing, we checked for inconsistencies such as missing values, unexpected data types, and discrepancies in categorical values.

- We dropped rows with null values in critical columns like `Location Description`, `Community Area`, `Latitude`, and `Longitude` to ensure the final dataset was consistent and ready for analysis.

Dataset Parameters

To focus our analysis, we scoped the final dataset using the following parameters:

- **Timeframe**: Limited to the years 2014–2023 to include recent data and visualize the effects of the COVID-19 pandemic on crime patterns.
- **Columns Retained**:
  - `Date` : To analyze temporal trends in crime.
  - `Primary Type` : To study crime categories.
  - `Description` : To add context to crime types.
  - `Location Description` : To understand where crimes occurred.
  - `Arrest` : To examine law enforcement responses.
  - `Domestic` : To identify incidents involving domestic violence.
  - `Community Area` , `Latitude` , and `Longitude` : To perform spatial analyses.
  - `Year` : To filter and analyze data across years.

This scoping resulted in a final dataset size of approximately ~266MB, enabling a much more managable experience, while retaining what we felt was the most important features within the data.

In [1]:
```
#Libraries

import pandas as pd
from IPython.display import Image, display

# Step 1: Loading the dataset
#crimes = pd.read_csv(r'C:\\Users\\17572\Desktop\\Mid Term\\Visualization Wr
```

In [2]:
```
# Step 2: Previewing the dataset and summary statistics
#print("Dataset preview:")
#print(crimes.head())
#print(f"Initial dataset shape: {crimes.shape}")
#print("Dataset information:")
#crimes.info()
```

As you can see from the above, our original dataset has 8,199,424 rows with 22 unique columns (bool(2), float64(7), int64(3), object(10)).

In [3]:
```
# Step 3: Dropping unused columns
# columns_to_drop = [
#    'ID', 'Case Number', 'Block', 'IUCR', 'Beat',
#    'District', 'Ward', 'FBI Code', 'X Coordinate',
#    'Y Coordinate', 'Updated On', 'Location'
#]
#print(f"Dropping columns: {columns_to_drop}")
#crimes.drop(columns=columns_to_drop, inplace=True)
```

In [4]:
```
# Step 4: Filtering the dataset for years between 2014 and 2023

# crimes = crimes[(crimes['Year'] >= 2014) & (crimes['Year'] < 2024)]
```

```
# print(f"The data is from {crimes['Date'].min()} to {crimes['Date'].max()}"
```

In [5]:
```
# Step 5: print("\nChecking for null values in each column:")
# for column in crimes.columns:
#    print(f'Sum of null values in {column} are: {sum(crimes[column].isnull(
```

As you can see from the above the dataset is fairly clean, with ~11,495 null values in Location Description, 2 in Community Area and 40,202 in both latitude and longitude. Because our dataset will remain quite large after cleaning and processing we chose to drop all null values.

In [6]:
```
# Step 6: Removing null values in the following columns
# columns_to_check = ['Location Description', 'Community Area', 'Latitude',
# print(f"Removing rows with null values in the following columns: {columns_
# crimes.dropna(subset=columns_to_check, inplace=True)
```

In [7]:
```
# Step 7: Saving the cleaned and filtered dataset to a new CSV file
# output_file = 'crimes_2014to2023.csv'
# print(f"Saving the cleaned dataset to {output_file}...")
# crimes.to_csv(output_file, index=False)
```

In [9]:
```
# Step 8: Loading the cleaned dataset and displaying the information
file_path = r'C:\Users\olear\Desktop\(Wednesday) Data Viz\Final\Final Data S
Final_Crimes = pd.read_csv(file_path)

print("Dataset Information:")
Final_Crimes.info()

# Generate a summary of the dataset
print("\nDataset Summary Statistics:")
print(Final_Crimes.describe(include='all'))

# Double checking for null values in each column
print("\nChecking for null values in each column:")
for column in Final_Crimes.columns:
    null_count = Final_Crimes[column].isnull().sum()
    print(f"Sum of null values in {column}: {null_count}")
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2486988 entries, 0 to 2486987
Data columns (total 10 columns):
 #   Column                Dtype
---  ------                -----
 0   Date                  object
 1   Primary Type          object
 2   Description           object
 3   Location Description  object
 4   Arrest                bool
 5   Domestic              bool
 6   Community Area        float64
```

```
 7   Year                    int64
 8   Latitude                float64
 9   Longitude               float64
dtypes: bool(2), float64(3), int64(1), object(4)
memory usage: 156.5+ MB

Dataset Summary Statistics:
                          Date Primary Type Description Location Description
\
count                  2486988      2486988     2486988              2486988
unique                 1147483           35         514                  200
top     01/01/2016 12:01:00 AM        THEFT      SIMPLE               STREET
freq                       132       558807      277145               605854
mean                       NaN          NaN         NaN                  NaN
std                        NaN          NaN         NaN                  NaN
min                        NaN          NaN         NaN                  NaN
25%                        NaN          NaN         NaN                  NaN
50%                        NaN          NaN         NaN                  NaN
75%                        NaN          NaN         NaN                  NaN
max                        NaN          NaN         NaN                  NaN

         Arrest Domestic  Community Area          Year      Latitude  \
count   2486988  2486988    2.486988e+06  2.486988e+06  2.486988e+06
unique        2        2             NaN           NaN           NaN
top       False    False             NaN           NaN           NaN
freq    2008306  2004818             NaN           NaN           NaN
mean        NaN      NaN    3.692568e+01  2.018343e+03  4.184334e+01
std         NaN      NaN    2.145785e+01  2.899099e+00  8.703221e-02
min         NaN      NaN    1.000000e+00  2.014000e+03  3.661945e+01
25%         NaN      NaN    2.300000e+01  2.016000e+03  4.176820e+01
50%         NaN      NaN    3.200000e+01  2.018000e+03  4.186160e+01
75%         NaN      NaN    5.400000e+01  2.021000e+03  4.190598e+01
max         NaN      NaN    7.700000e+01  2.023000e+03  4.202267e+01

           Longitude
count   2.486988e+06
unique           NaN
top              NaN
freq             NaN
mean   -8.767047e+01
std     5.993882e-02
min    -9.168657e+01
25%    -8.771305e+01
50%    -8.766405e+01
75%    -8.762770e+01
max    -8.752453e+01

Checking for null values in each column:
Sum of null values in Date: 0
Sum of null values in Primary Type: 0
Sum of null values in Description: 0
Sum of null values in Location Description: 0
Sum of null values in Arrest: 0
```

```
Sum of null values in Domestic: 0
Sum of null values in Community Area: 0
Sum of null values in Year: 0
Sum of null values in Latitude: 0
Sum of null values in Longitude: 0
```

Our final cleaned data set is now 2,486,988 entries, with 10 columns (bool(2), float64(3), int64(1), object(4)). The dataset is from 01/01/2014 01:00:00 AM to 12/31/2023 12:55:00 AM,a full 10 years of data with all null values having been removed.

In [ ]:

```python
# Step 9: Descriptive Statistics for our selected variables
# Numeric variables
centroid_lat = Final_Crimes['Latitude'].mean()
centroid_long = Final_Crimes['Longitude'].mean()
print(f"Geographic center (centroid): ({centroid_lat}, {centroid_long})")


# Categorical variables
categorical_columns = ['Primary Type', 'Description', 'Location Description'
for column in categorical_columns:
    print(f"\n{column} Value Counts:")
    print(Final_Crimes[column].value_counts())

# Unique value counts
print("\nUnique Values for Each Column:")
for column in Final_Crimes.columns:
    print(f"{column}: {Final_Crimes[column].nunique()} unique values")
```

Main Analysis

Approach and Methodology

We knew based on the quality of our dataset that we could focus on multiple areas, especially with the geolocation columns. During the EDA process, we decided early on that we wanted to devote at least a portion of the analysis to homicides within Chicago. The other two questions, involving the overall nature of crimes committed in Chicago and the spatial and temporal trends of those crimes, seemed fitting given the additional column types. Due to the high number of Crime types at 35 we also chose to narrow the focus with most of our visualizations. When using crime type as a visualization we focused on the top ten crimes as a way to narrow the focus and provide cleaner visuals.

Once we scoped our research questions and cleaned the dataset, our visualization journey began by creating draft visualizations in Python. While we eventually made the executive decision to create a Tableau dashboard, most of our draft visualizations were performed in Python. For the sake of this assignment, we will be providing only our finalized visuals from our dashboard in Tableau.

Overall Crime Analysis

Our overall crime analysis began with simple plots in Python to visualize crime types. This dataset contains **35 different "Primary Types"** (crime categories). Given the substantial number of categories, we hypothesized that there would be a few dominant crime types and many smaller, less common crimes.

To explore this, we visualized a "Top 10" list of crimes shown below. Additionally, we discussed and analyzed nuanced interactions within the dataset, such as:

- Arrest rates for different crime categories.
- The proportion of domestic versus non-domestic violence.
- Potential trends the Seasonality of crimes.

Ultimately, we decided on the below four visualizations to highlight the most interesting interactions within the dataset which showcase the results of our analysis and provide insights into the overall crime landscape in Chicago.

Spatial and Temporal Analysis:

For spatial and temporal trends, we wanted to focus our analysis on when and where the most crimes were committed. We also wanted to try and visualize seasonal trends based on stereotypes we had encountered previously involving when the propensity of crime is committed each year. We again chose four visualizations; each one focusing on different subsets of the data to try and provide us with further trend analysis.

Homicide Analysis:

For our homicide analysis we wanted to integrate a visual depiction of the most dangerous areas in Chicago. Additionally, we wanted to visualize the arrest rate of homicide and the location where most homicides take place. We chose three different visualizations for the section, with the highlight being the interactive choropleth map which plots homicides by community area.

Tableau Visualization Process:

We wanted to add some thematic enhancements to the overall dashboard while also including some reference to the city of Chicago. The image in the dashboard itself is a shot of some of the skyscrapers in the Chicago skyline. We also chose a bold red on black theme to harken back to the crime "theme" them of the dashboard. We tried to keep the color schema the same throughout most of the visualizations, while make some small edits for readability or clarity.

Visualizations and Insights:

We have leveraged most to the Tableau functionalities required to create visualizations and dashboard like

- creating calculated field
- editing the tiles display
- adding customized shapes to visualizations
- using spatial file for map visualizations
- additing actions, filters, coloring etc

Question 1:

What are the overall trends in crime types and their occurrences over time?

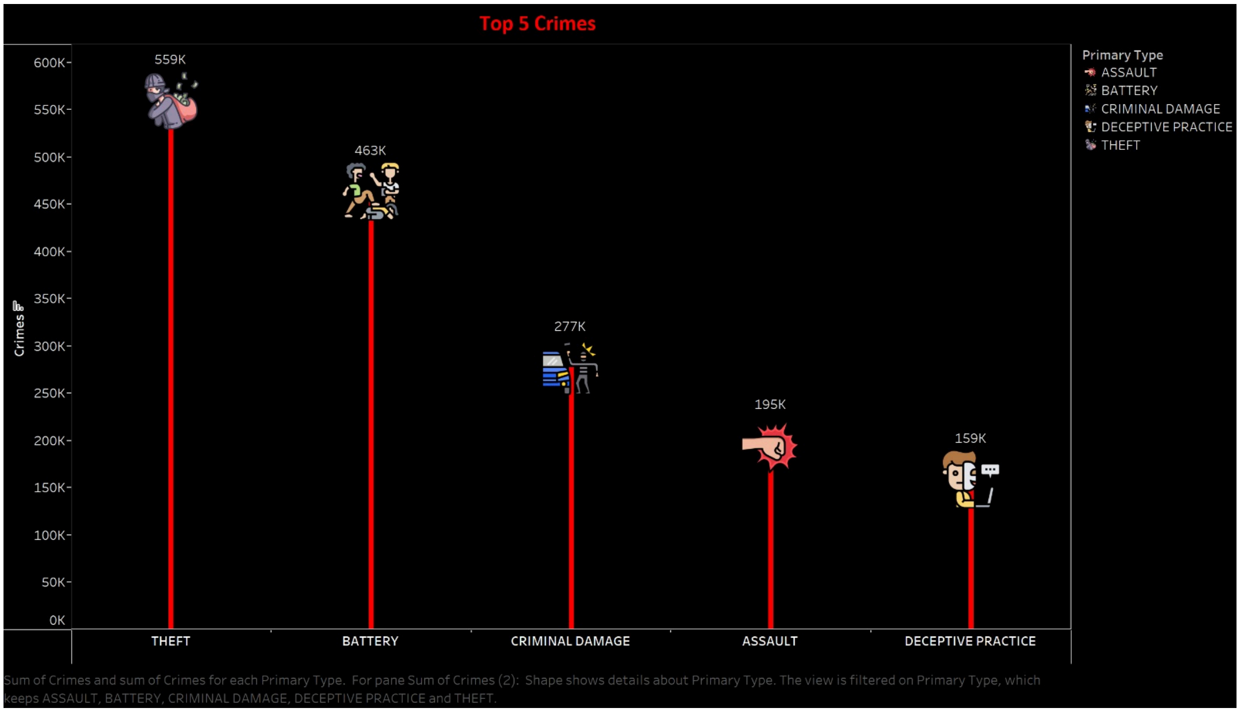
Top 5 Crimes:

- Having noticed other crimes like assualt, battery and criminal damage we also within the top 10, we decided that our next visualization should highlight domestic crimes within Chicago. Instead of plotting bar chart in the common way, we thought of utilising shapes feature of tableau here to add a little more flair to the image. Eventhough it was a bit challenging, the final visualization turned out as intended.

- Deceptive Practice seems to be one among the Top-5 crimes which means that it is a significant issue that requires immediate attention from law enforcement and policymakers.

In [21]:
```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[21]:



Top 10 crimes :

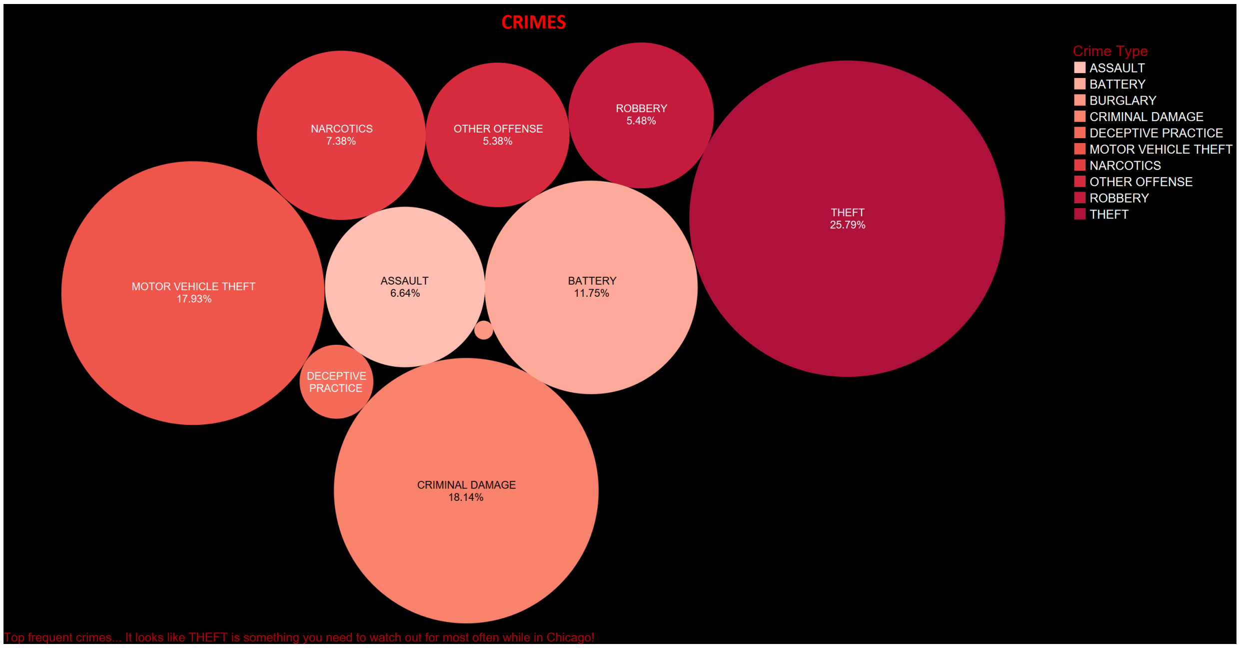
- The first visualization we agreed on was a bubble chart showing the top 10 most committed crimes. This simple visualization gives a **quick understanding of top crimes** based on the bubble size.
- Theft in Chicago is incredibly prevalent and looks like one has to be careful for while in Chicago!

In [10]:

```
from IPython.display import Image

Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[10]:

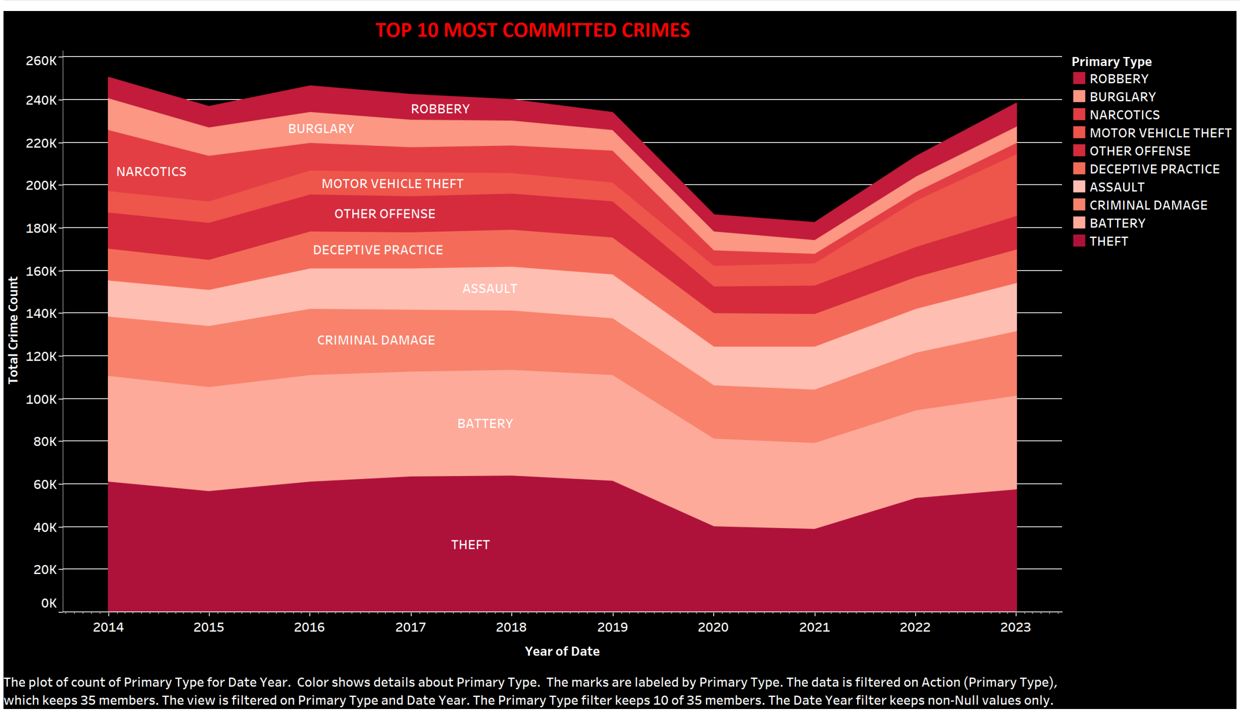


Top 10 Crime Types Trend:

- Next we wanted to visualize the **dominant trends for top frequent crime categories** and so selected area chart.
- This visual shows the drop in overall crimes during Covid(2020-2022). Generally though, crime trends have appeared to remain the same over the last 10 years. Even the pandemic couldn't effect the trend of top 10 crimes.

In [12]:
```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[12]:

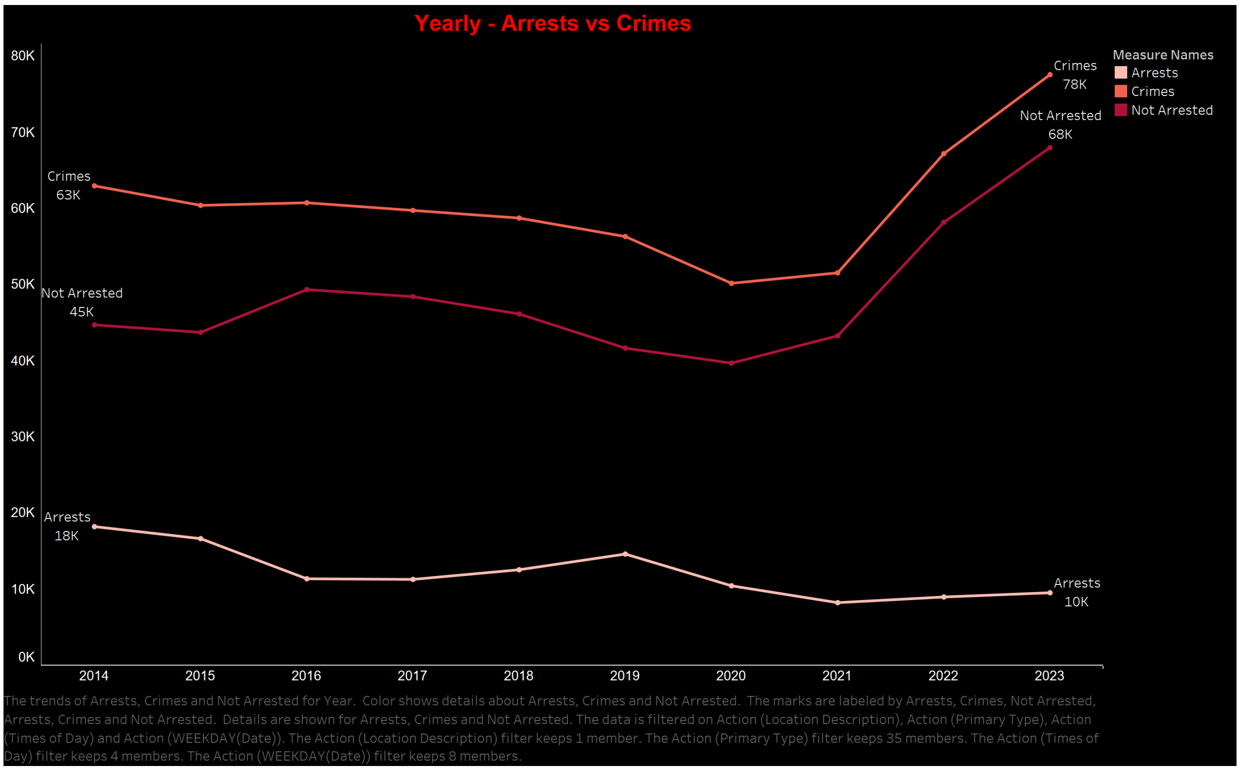


Yearly Arrests vs Crimes:

- We wanted to **highlight the discrepancies in arrest rates** for overall crimes and so decided to use line chart here.
- From the visual we can see, very few crimes end in arrest comparatively.
- This is also our first visualization showing the impacts of covid, both on crime rates and arrests.

In [13]:
```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[13]:



Question 2:

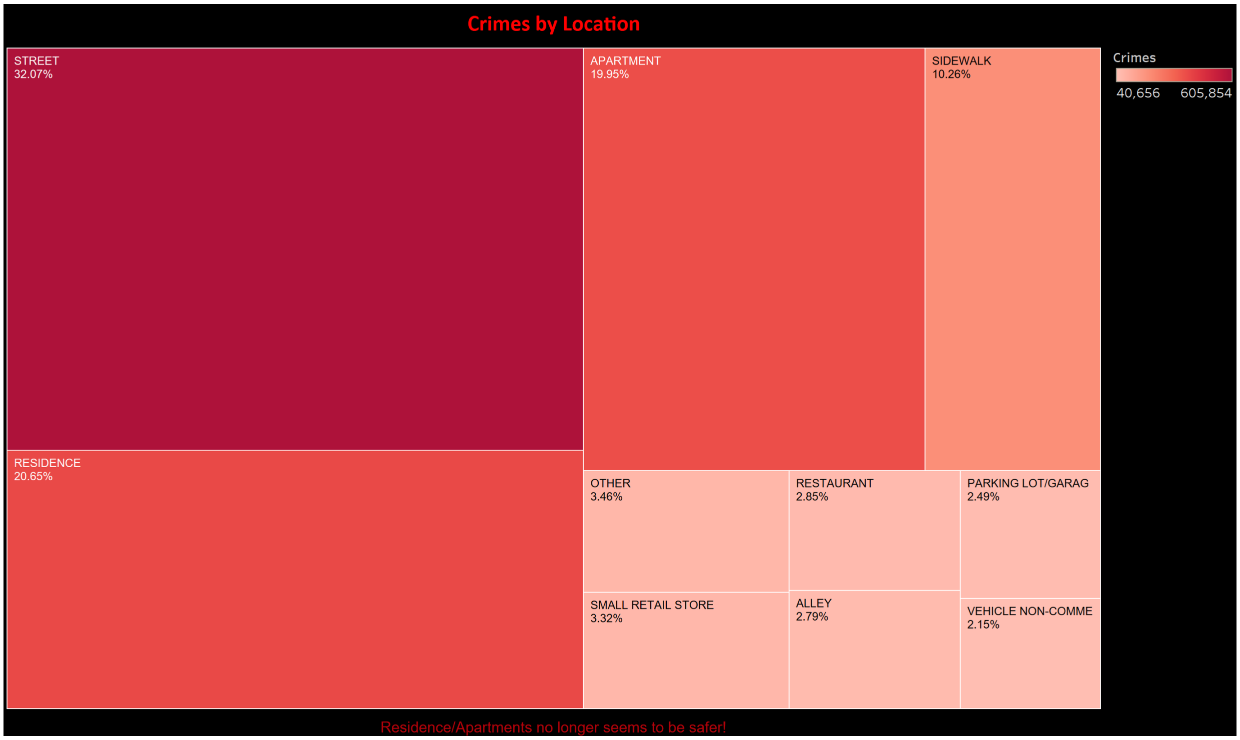
How do crime patterns vary spatially and temporally across different regions or neighborhoods?

Crimes by Location:

- For this visualization, we wanted to f**ocus on where(type of place) crime is committed** in Chicago and so have chosen TreeMap.
- We hypothesized that the propensity of crime would be committed in the home, especially with so many domestic crimes being highlighted in our previous visualizations. In actuality most Chicago crime is committed on the street! With a 32% overall rate, this trend also prevails in our homicide analysis.

```
In [14]: Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[14]:



Top 10 Seasonal Crimes:

- We wanted to visualise the **seasonality of top ten crimes** and so ended up with clustered bar chart.
- The horizontal bar graph shows that more crimes are committed in warmer months, with the Summer and Fall seasons being the high points for crime throughout the year and theft being the most frequent one for all the times!

```
In [22]: Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[22]:

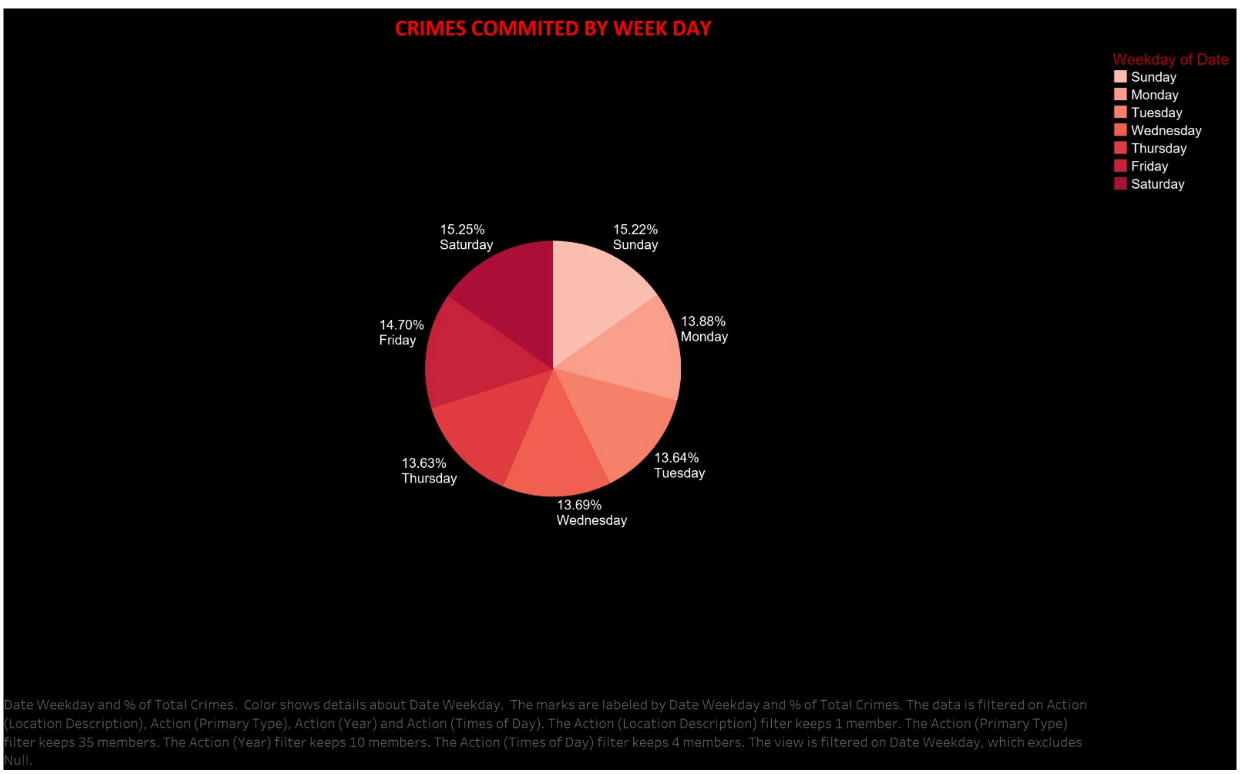
Crimes Committed Seasonality check by Day of the Week:

- We wanted to use pie chart for focusing on **crimes committed by day of the week**.
- The overall trend seems to be split evenly across the week, with a small uptick in crime over weekends.

In [23]:

```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[23]:

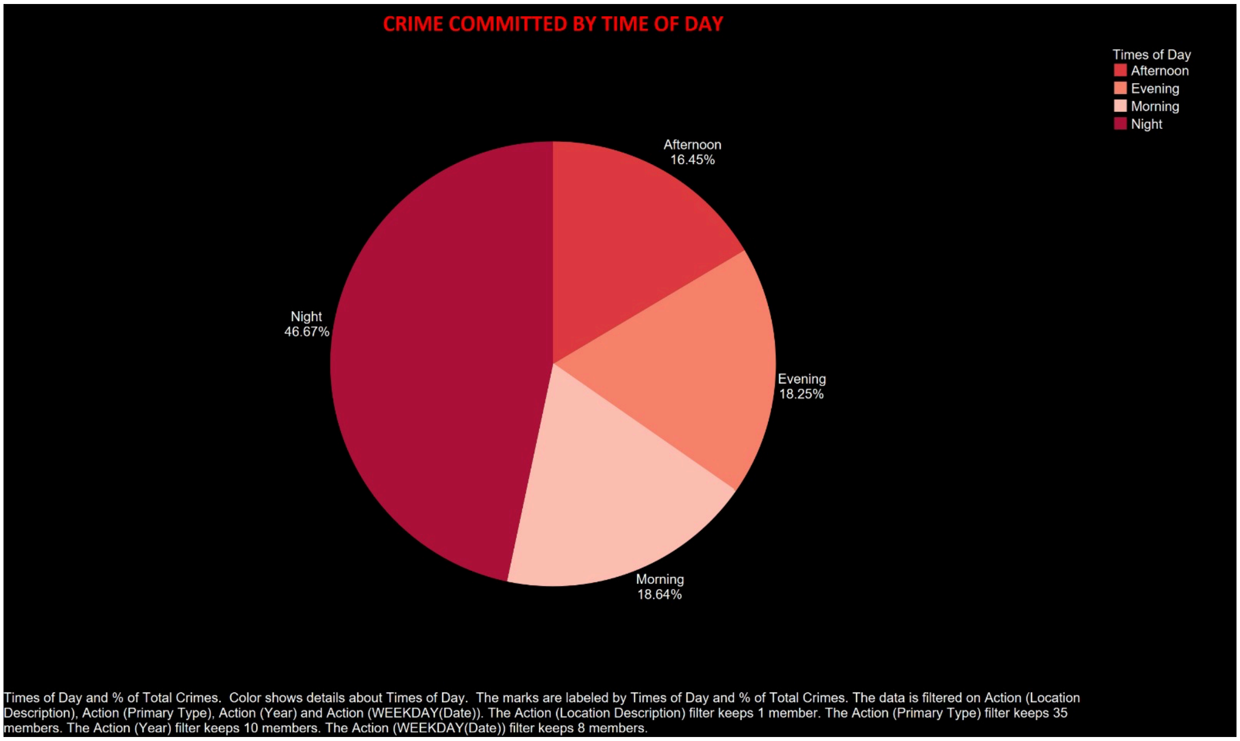


Crimes Committed Seasonality check by Time of Day:

- In this visualization, we wanted to focus on when **crime committed during a 24 hour day cycle** and have chosen pie chart as well for the same.
- Night accounts for over 46% of crimes committed, with a generally even split across other periods of the day.

In [16]:
```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[16]:



Question 3:

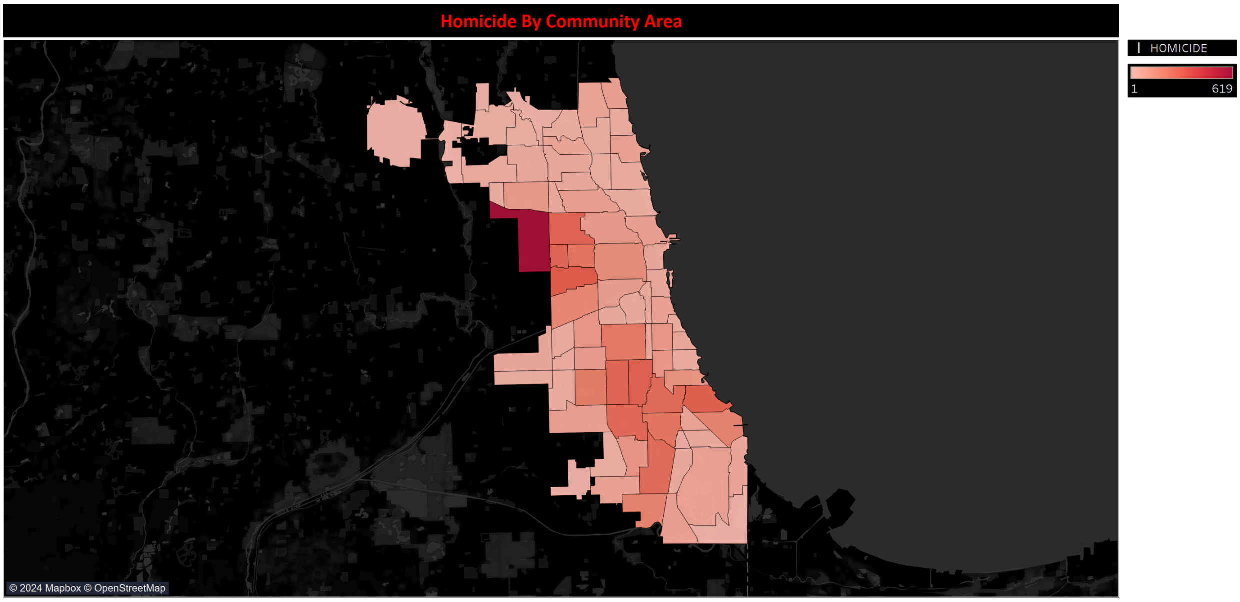
What are the key factors and trends associated with homicide occurrences, and how do they differ from other types of crimes?

Homicide by Community Area:

- In this visualization, we wanted to p**lot the number of total homicides in** our dataset to their corresponding **Community Area** and so have chosen map.
- The most dangerous area by far is a community area named Austin, highlighted in dark red. Of the 6,473 total homicides in our data set, Austin accounted for nearly 10% with 619 homicides across our 10-year period.
- We have used spatial data inorder to color the community area based on the community area number present in our dataset.

In [17]:
```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[17]:



Map based on Longitude (generated) and Latitude (generated). Color shows count of Primary Type. Size shows details about Primary Type as an attribute. The data is filtered on Primary Type, which keeps HOMICIDE.

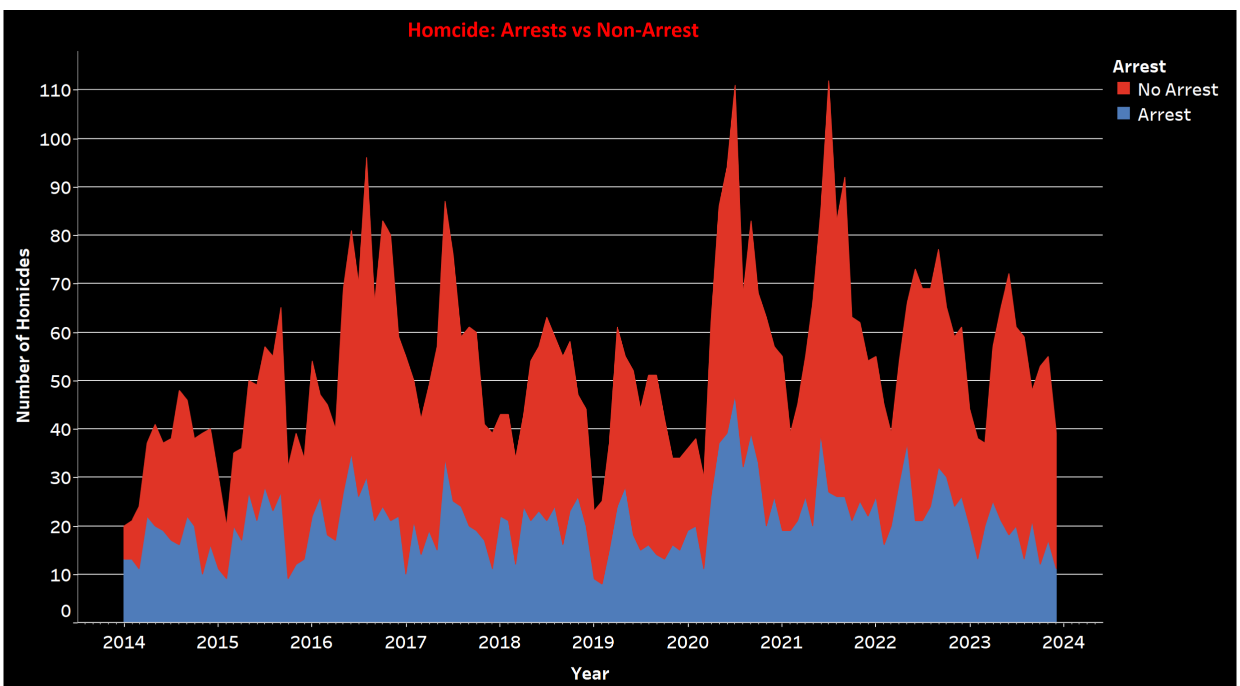
Homicides involving arrests and non-arrests:

- We wanted to visualize the **monthly count of homicides for arrests and non-arrests** and so have selected area chart.
- The number of homicides without arrests is consistently higher than those with arrests, with noticeable spikes at regular intervals. This indicates that there are challenges in resolving a significant portion of homicide cases.

In [18]:

```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[18]:



The plot of count of Primary Type for Date Month. Color shows details about Arrest. The data is filtered on Primary Type, which keeps HOMICIDE.

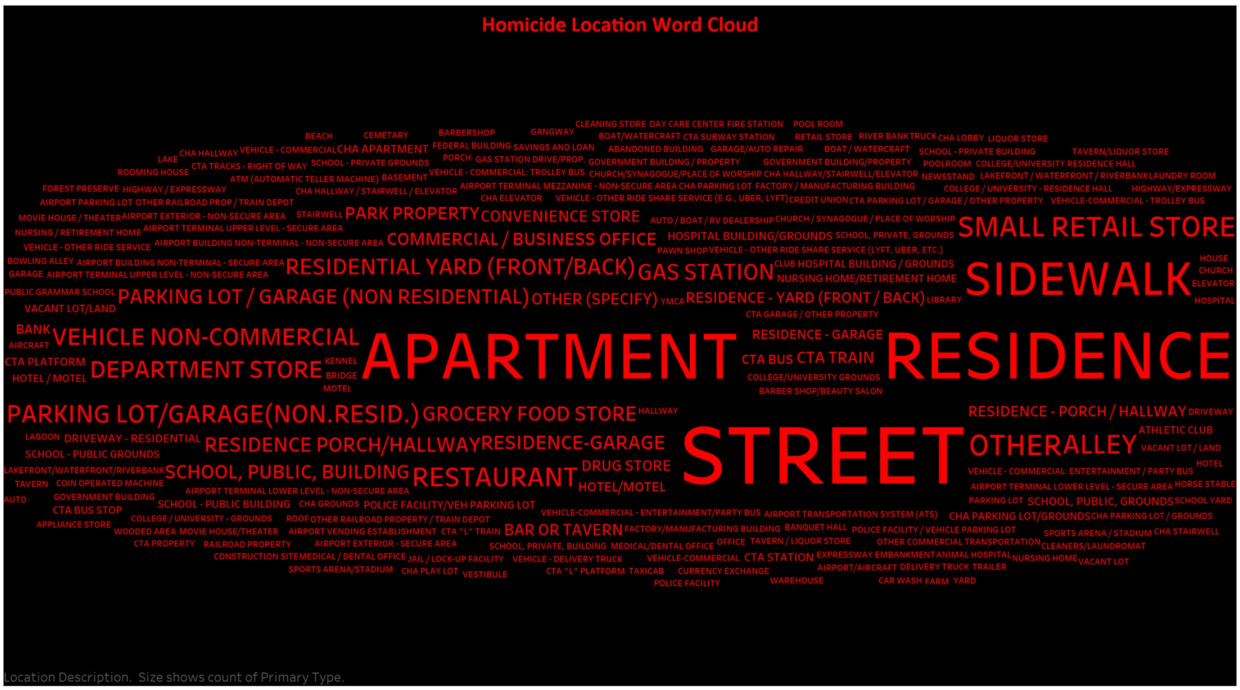
Most Common Location of Homicide:

- We wanted to **highlight the frequency of locations where homicides occurred** in our visualization and so select a bubble chart and converted it into a word cloud chart.

- Data again proved that our misconception about public areas being more common sites for homicides as wrong. The most frequent locations included "Residence" and "Apartment" as few of top most locations

In [19]:
```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

Out[19]:

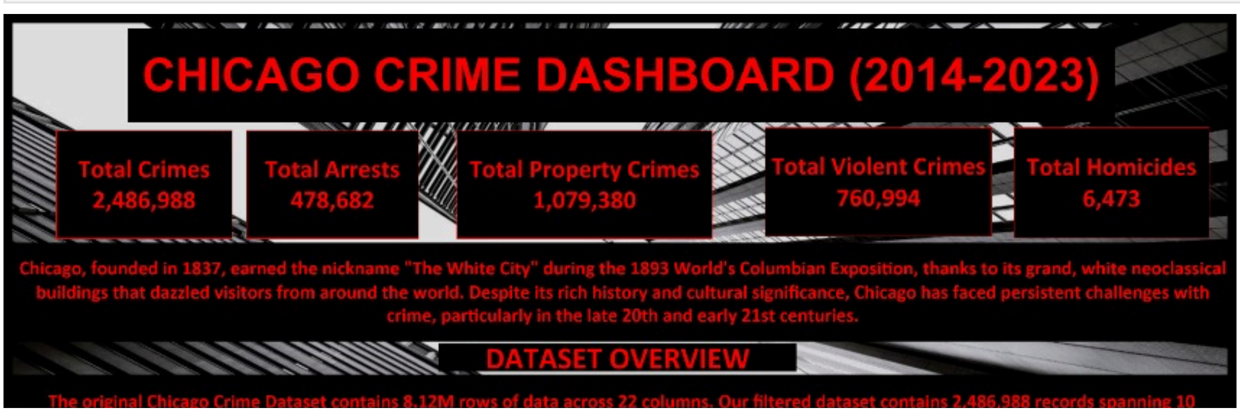


Final Tableau Dashboard

Below is a screen capture of our finalized, finished Tableau Dashboard; Chicago Crime Dashboard (2014-2023)

In [24]:
```
Image(url='https://github.com/AishwaryaMaddula/DataVizProject/blob/main/Imag
```

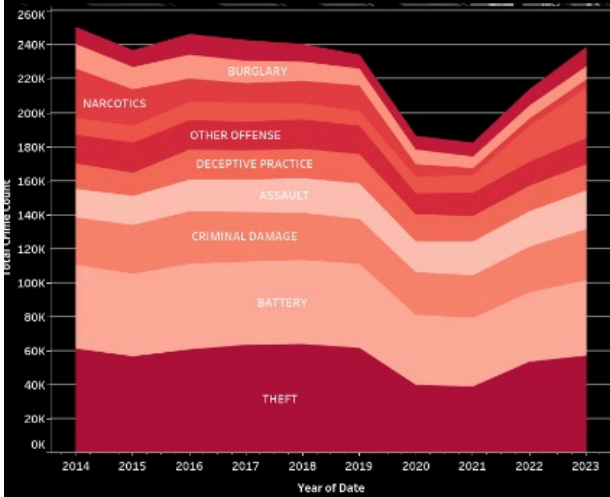
Out[24]:



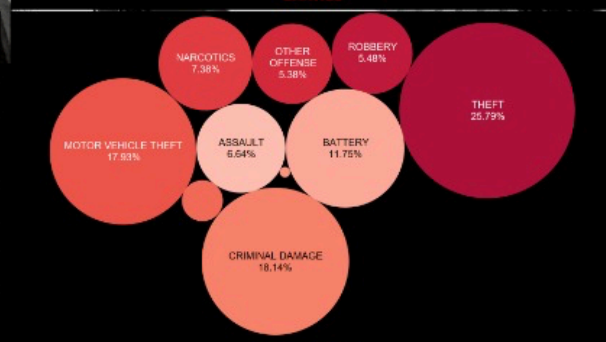
columns, capturing reported crimes in Chicago from 2014 to 2023. It includes 35 distinct crime types and covers Chicago's 77 community areas. Our data consists of 10 total columns, with two time-related columns (Date, Year), 4 detailing crime (Primary Type, Description, Arrest, Domestic), and 4 providing location data (Location Description, Community Area, Latitude, Longitude), providing a comprehensive view of crime patterns across the city.

CRIME TYPE ANALYSIS

TOP 10 MOST COMMITTED CRIMES



CRIMES

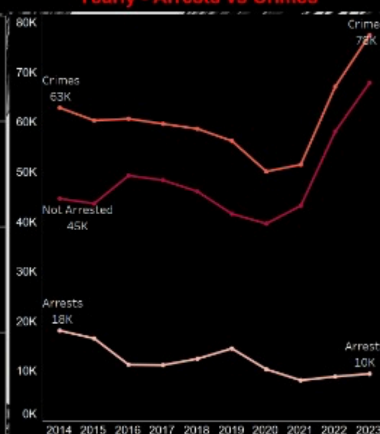


Top 5 Crimes

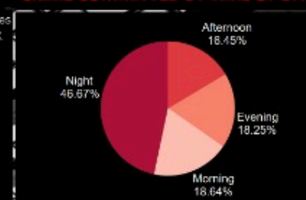
Temporal Crime Analysis

Top 10 Seasonal Crimes

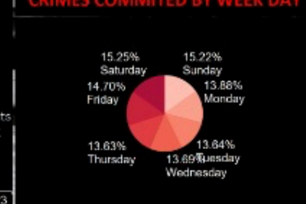
Yearly - Arrests vs Crimes



CRIME COMMITTED BY TIME OF DAY



CRIMES COMMITED BY WEEK DAY



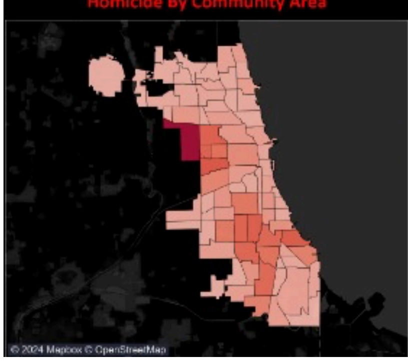
SPATIAL CRIME ANALYSIS

Crimes by Location

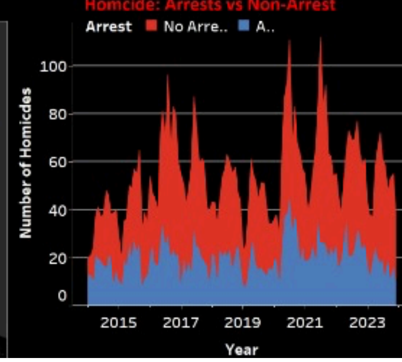
HOMICIDE

In 1992, Chicago recorded its highest number of murders in a single year: 943, during a period marked by gang violence and the crack cocaine epidemic. More recently, in 2023, the city reported 617 homicides, a 50% increase from 2013. This figure cemented Chicago's position as the city with the highest number of homicides in the United States for the 12th consecutive year, though its per capita murder rate is surpassed by several other cities.

Homicide By Community Area



Homcide: Arrests vs Non-Arrest



Homicide Location Word Cloud

Challenges and Iterations:

1. Faced issue initilly with data upload to tableau. Since the file size was around 2 GB it took almost hours and the visualizations also took time to be responsive and we encountered errors most of the time with "maximum shelf limit" reached while displaying crime description or types. File size was reduced as a fix.

2. Obtaining desired visualization of map took time for us. We had to generate spatial data inrorder to get the desired coverage of community area coloured by sum of crimes.

Key Findings

- SMART Question-1: Overall Crime Analysis

From the visualisations, we found out that theft was the most frequent crime overall. Even during the covid, theft was the most frequent crime while the others also seeems to be in the same order of frequency. The count of crimes seemed to be decreased but gradual increase in crime count is observed post pandemic.

- SMART Question-2: Spacial and Temporal Analysis

For spatial and temporal analysis, we noticed common patterns across the entire dataset, with the main takeaways being that crime is mostly committed at night, on the street, in the summer and likely won’t be followed with an arrest. We also visualized season trend crime trends with crime the hotter months host to more crime, and we also noted the drop in overall crime in 2020 and 2021 due to covid. Finally, weekends were host to the most crime, but with only a small uptick in overall percentage compared to the other week days.

- Smart Question-3: Homicide Analysis

Our three visualizations involving homicide generally parallel other crimes within the dataset. We did focus on the Community Area’s and by visualizing homicide within the confines we noticed certain areas had much higher homicide rates than others, especially in Austin and its neighboring areas. Where homicides we committed also showed similar results to other more common crimes, with the propensity being on the streets and in residencies. Arrest rate and seasonality also mirrored typical crime trends, with the arrest rate remaining low and homicide numbers spiking each summer mostly in July and August. Although numbers are lower than say it’s peak in 1992, Chicago remains a dangerous city, leading the nation.

Following are the key findings from python statistical analysis:

- *Community Areas*:
    - Chicago is broken into **77 Community Areas**, and all are represented in the dataset.
- *Crime Descriptions:*
    - There are **514 unique descriptions**, reflecting the detailed categorization within each primary crime type.
- *Domestic Incidents*:
    - Approximately **19.4% of incidents** (**482,170**) are flagged as domestic.

Following are the key findings from tableau analysis:

- *Geographic Center (Centroid)*:

- The centroid of the dataset is located at **(41.8433, -87.6705)**, which is within Chicago.

- *Crime Types*:
    - **Theft** is the most frequently recorded crime (**558,807 occurrences**), followed by **Battery** (**463,348**) and **Criminal Damage** (**277,154**).
    - Rare categories include:
        - **Ritualism** (1 occurrence)
        - **Human Trafficking** (87 occurrences)
- *Crime Locations*:
    - Most crimes occur on:
        - **Streets** (605,854 occurrences)
        - **Residences** (390,166 occurrences)
        - **Apartments** (376,918 occurrences)
- *Arrests*:
    - Only **19.2% of crimes** (**478,682**) resulted in an arrest, while **80.8%** (**2,008,306**) did not.

Conclusion

Limitations Encountered

We encountered several limitations while working with Tableau to create and share our visualizations:

1. **File Sharing Challenges**:
   Our initial plan was to use Tableau Cloud to seamlessly build and share the dashboard. Unfortunately, due to account mismatches and login duplications, we were unable to collaborate effectively using the cloud platform. As a result, we resorted to manually sharing Tableau files, which was less efficient.

2. **Limited Experience with Tableau**:
   Our limited familiarity with Tableau proved to be another limitation. While pre-made dashboards showcase the polished potential of Tableau, the learning curve for crafting nuanced and effective dashboards is steep. This hindered our ability to fully utilize the tool's advanced features and create a more refined final product.

3. **Size of the Original Dataset**:
   The overall size of the original dataset was daunting and was the first challenge we set out to tackle. While it would have been ideal to export a refined CSV directly from Chicago's data portal, we instead performed the necessary transformations in Python to make the dataset manageable for our analysis.

Future Directions

Because this dataset has such a large scope and includes various types of information, there are numerous future directions we could take. As mentioned previously, this dataset—or ones like it—could be highly relevant to law enforcement, policymakers, and community leaders. Research could be performed to provide actionable insights tailored to these groups. Some examples include:

1. **For Law Enforcement**:
   - **Resource Allocation**: Use spatial and temporal crime patterns to optimize patrol routes and allocate police resources more efficiently.
2. **For Policymakers**:
   - **Policy Impact Analysis**: Study the effects of new policies (e.g., gun control measures, community policing) on crime trends.
3. **For Community Leaders**:
   - **Community Outreach**: Identify neighborhoods with high crime rates to focus on community-building initiatives.
4. **General Research Opportunities**:

- **COVID-19 Impact**: Explore how the pandemic influenced crime trends, particularly in areas like domestic violence or theft.

Lessons Learned

1. **Scoping the Dataset**:

    - In hindsight, we likely would have scoped the dataset even further if we had fully realized how large it was going to be after deciding to analyze 10 years of data. A smaller, more focused dataset might have allowed us to streamline the analysis process and reduce computation time.

2. **Tool Selection**:

    - We initially considered using Python as our final visualization tool. However, after working through this assignment in Jupyter, we realized that a more thorough discussion about the benefits of using Tableau might have been valuable early on.
    - That being said, we are much more proficient with Tableau, and choosing it as our visualization platform ultimately allowed us to present our insights more effectively and efficiently.

References

Following sources were referred for tableau visualizations:

1. Bar chart with Custom shapes - Custom Shapes
2. Word Cloud - Tableau Community