Data Science Boot Camp: Project 1
Jessica Andras
Saber Garibi
Andrew Rexford
Dennys Urdiales
Jennifer Alvarez

Unveiling Patterns: An In-depth Data Analysis of Arrests and Crime in Chicago, IL

1.   **Purpose and Background**

The purpose of this project is to give us a chance to interact with a strong issue Chicago and Chicagoans face; though there are plenty, our issue was…Crime. As cities around the U.S. deal with this issue, we wanted to do our due diligence as well. An attempt to fact check popular beliefs, which examples won't be given as it is different to another individuals' biases. The last point to make regarding crime data in Chicago, luckily there is enough data to work with (perhaps even more than enough) due to city initiatives to be more data-driven. By examining the available datasets, the project aims to provide valuable insights into patterns, trends, and correlations related to arrests and crimes within the city.

a. **Finding and Using Public Data**

The data used in the analysis came from public data sources of the Chicago Police Department, Chicago's Data Portal, and a publicly available dataset by the Washington Post. Sources are credible and were expected to be accurate, complete, and consistent.

The direct links to the sources we used in this project can be found in the References sections.

b. **Questions Worth Asking**

After viewing the datasets with Excel to understand the data and formulate the questions. The questions that came up were:

Which districts have the highest number of arrests?

What are the top 5 Zip Codes with the highest number of fatalities in the City of Chicago, and how have fatalities trends in these areas changed from 2015-2021?

What is the breakdown of the most/least frequently violated laws and statutes, and which charges are the most commonly filed?

What are the Common Causes of Arrest?

The list of questions that can be asked will need a longer timeline, from that long list. The above four were chosen to be shared in this project, because they were interesting to the group and more discoveries had come up through seeking answers for those questions.
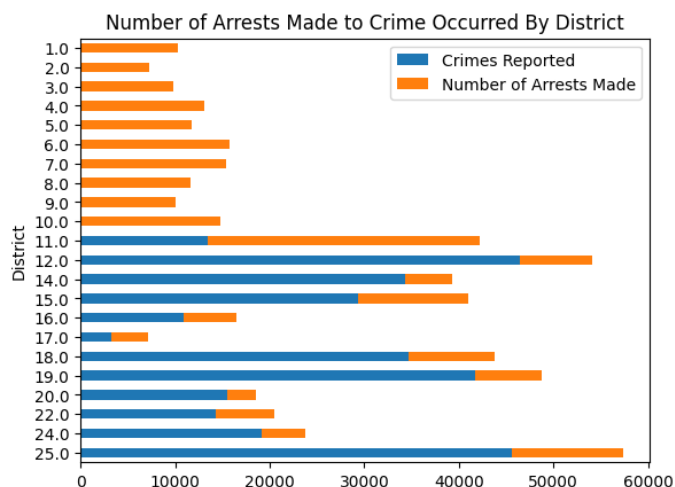
2. Data Exploration and Clean Up

The City of Chicago Data Portal's dataset ranged from 2001 to 2023 with 30 columns. This dataset caused problems with git as it was a large file. What was best is to keep the data more recent and a line was drawn from 2015 (beginning) to 2022 (ending) due to high CPU usage. Then extra columns were removed like FBI Code, Beat, etc.

The same concept of shortening the range and removing extra columns was done on other two datasets. There were times when null values caused an error, which dropping the null values solved it.

3. Conclusion from Analyses 1, 2 and 3

In Figure 1 when analyzing the district with the highest number of arrests, it is shown that district 11 resulted in the most total number of arrests with over 50,000 arrests made but having the least reported crimes in total and district 25 being the district with the most reported crimes but lower arrests made.

A heat map is created to show the most arrests made per district based on the latitude and longitude and the map showed the most activity for district 11. When looking at the arrests per year it showed that the most arrests were made in 2018 with a total of 52,314, with 2016 being the second highest of total arrests made, followed by 2017. The most crimes committed in district 11 by description of crime was possession of heroin, with over twice as many arrests compared to other causes of arrests per description.
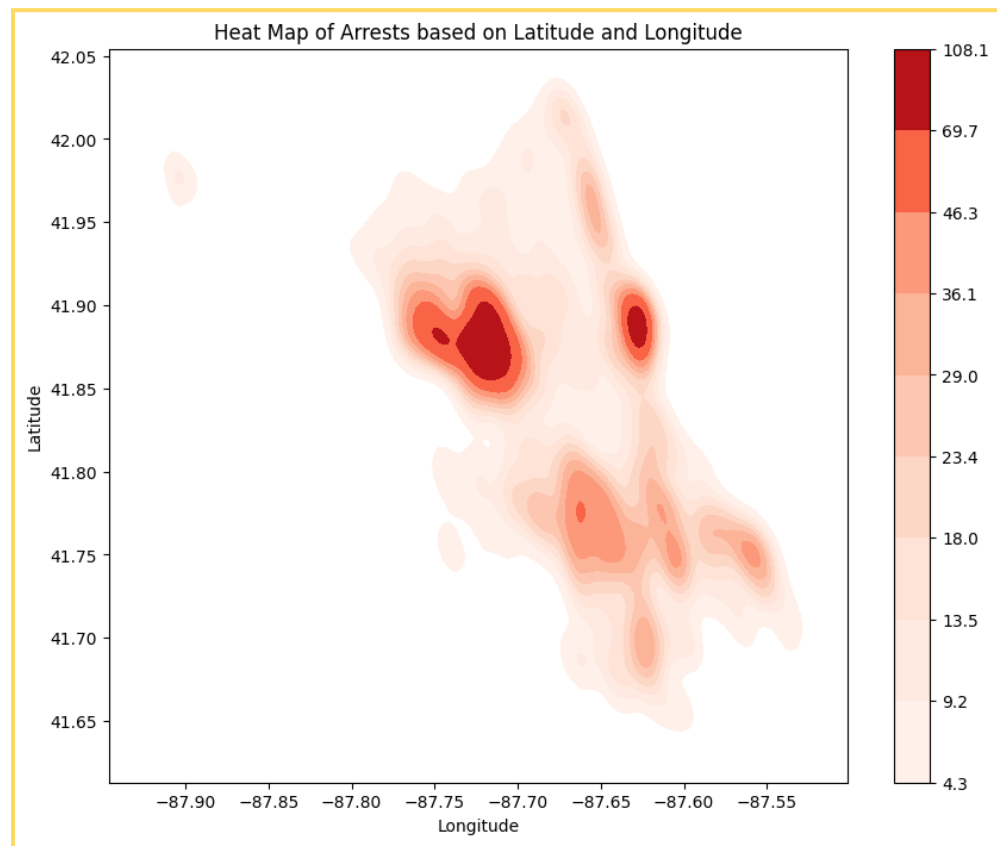

Figure 2 shows the most intense color red where more arrests are made.

When analyzing the data by looking at the relationship between the ethnic demographic and police-involved fatalities it is very evident that

African-American/Black individuals have had the most fatalities overall at 64% more

than other races. Hispanic/Latino people come in second at 14%. It is clear that there is a

huge disadvantage when being arrested or reporting a crime by/to CPD if you are an

African-American/Black individual. The arrests with the highest fatality rate for

African-American/Black individuals was the highest in 2015 followed by 2016. 2020 had

the highest fatality rate when looking at the overall rate with hispanic/latino and race

unspecified being equally as much as African-American/Black deaths.
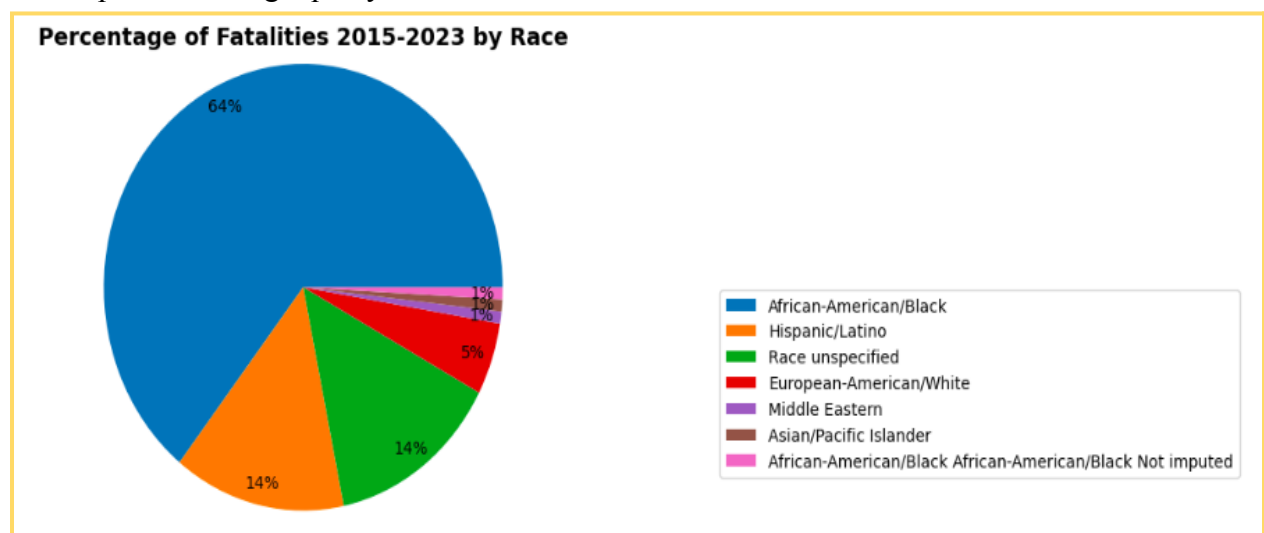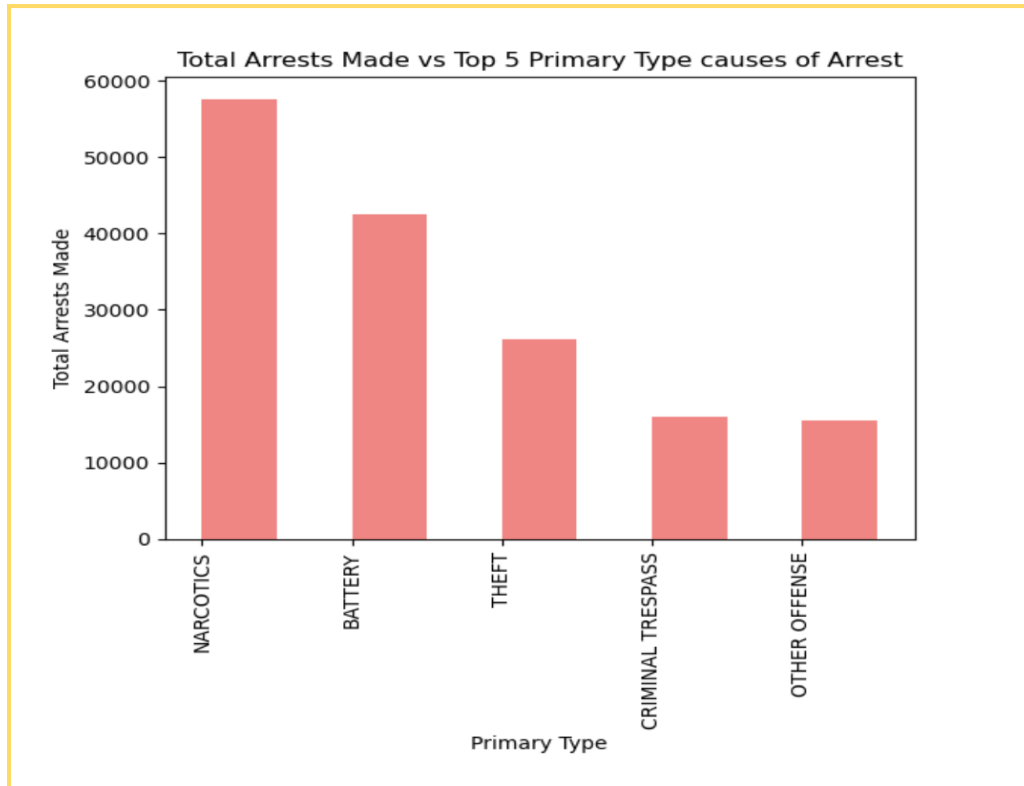

Figure 3 shows the fatalities by race in a pie chart.

The most common causes of arrest for the safest district 17 was analyzed by the

description of the arrest. It showed that retail theft was the highest leading to arrests

followed by domestic battery and a category marked as "simple." The common causes of

arrest for the Primary Type was also analyzed from 2015- present time collectively and it

showed Narcotics being the number one reason for arrest with a total of 57,632 arrests,

Battery came in second with a total of 42,556 arrests, and Theft came in third with a total

of 26,156 arrests. An analysis was done to see the top 5 common causes of arrests per

year for 2015-2022 and it showed Narcotics being the number one reason for arrests

between 2015-2019 and Weapons and Battery being the number one cause for arrests in 2020 and 2021. The information obtained for 2022 is considered to be missing some information that didn't make the data obtained reliable.

In Figure 4, the top 5 arrests are shown by Primary Type

When filtering the data by the most common cause of arrest by description for 2015-2022, it is shown that in 2015 cannabis possession was the highest reason for arrest, domestic battery was the most common for 2016-2019, unlawful possession of handgun for 2020-2021, and manufacture/deliver of cannabis of 10 grams for 2022. Figure 5 shows the most common description per year:

	Year	Description
0	2015	POSS: CANNABIS 30GMS OR LESS
1	2016	DOMESTIC BATTERY SIMPLE
2	2017	DOMESTIC BATTERY SIMPLE
3	2018	DOMESTIC BATTERY SIMPLE
4	2019	DOMESTIC BATTERY SIMPLE
5	2020	UNLAWFUL POSSESSION - HANDGUN
6	2021	UNLAWFUL POSSESSION - HANDGUN
7	2022	MANUFACTURE / DELIVER - CANNABIS OVER 10 GRAMS

An Independent T-Test was performed to find out if the average crime count between years 2015, 2016 and 2020 are statistically significant. An analysis was performed however our data was not normally distributed which affected the accuracy of outcome.

4. **Interesting Takeaways & Things to Consider for Future Work**
   ● The Highest Recorded Crime District was District 25.
   ● Most Common Charges are Damage of $500 and Under, Domestic Battery and Retail Theft.
   ● The quality over quantity of data makes a huge difference.
   ● Narcotics was the most common cause of arrest until the legalization of cannabis happened in 2019.
   ● There is an overwhelmingly high number of police-related fatalities when interacting with perceived African-Americans.
   ● District 11 has the lowest crimes reported but the highest number of arrests.
   ● Found out locations where there are a lot of crimes and arrests.

5. **References**

https://home.chicagopolice.org/statistics-data/public-arrest-data/

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

https://github.com/washingtonpost/data-police-shootings

https://fatalencounters.org