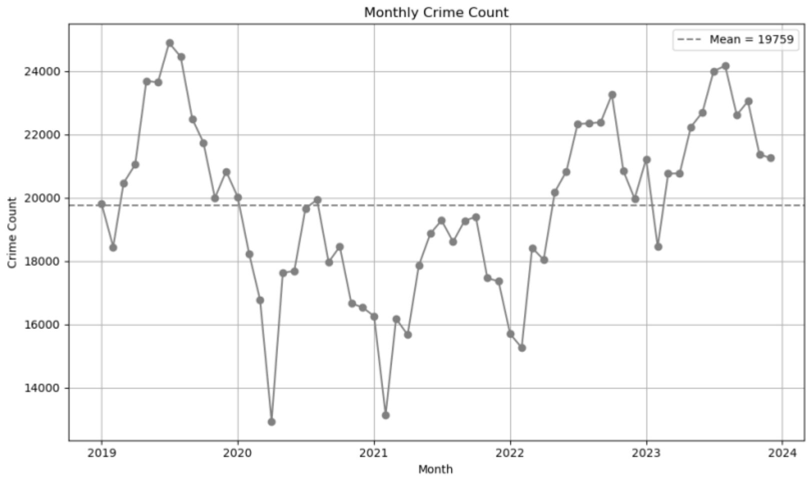
Course Project Part 1

Yu-Hsuan (Monica) Ko

Overview of Data Set – Chicago Crimes (2019-2023)

In the high-crime city of Chicago, this project aimed to analyze crime trends in the area in the previous five years. The dataset contains 1.18 million crime records from 2019 to 2023, obtained from the Chicago Data Portal. It includes details about each crime, such as the date, location, type of crime, and whether an arrest was made.

According to Figure 1, crime in Chicago has fluctuated over the past five years. Crime counts were high in 2019, then dropped sharply in early 2020, possibly due to COVID-19. From 2021 to 2023, crime counts show an upward trend with seasonal fluctuations. In Figure 2, the map on the right shows crime concentrated in downtown Chicago and nearby areas, likely due to higher population density and activity. Areas with lower density, such as around the University of Chicago, see fewer incidents.



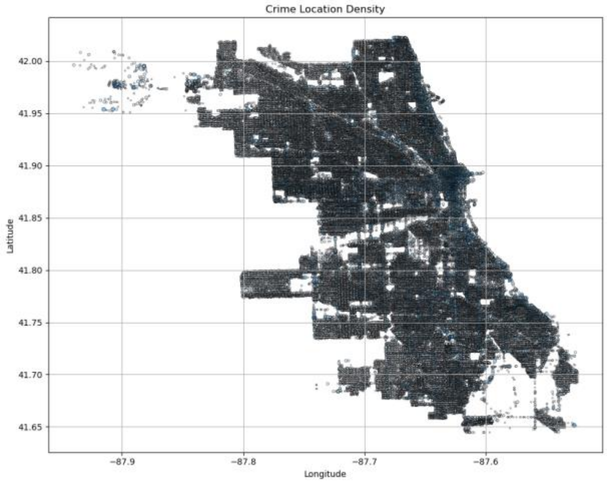
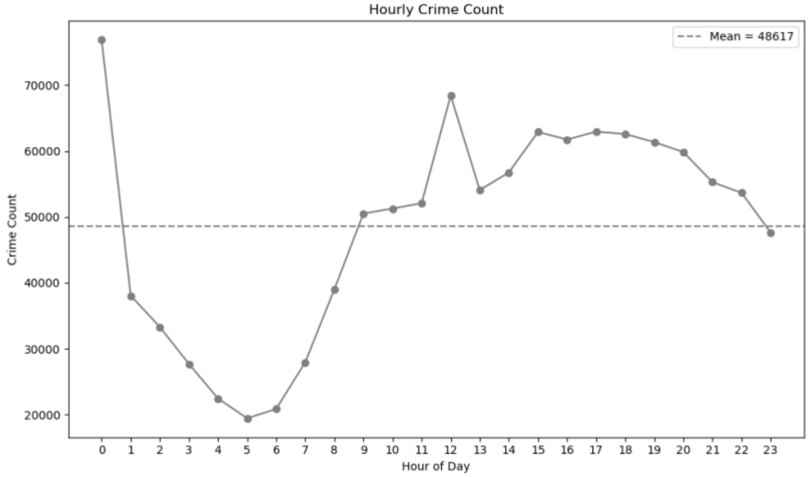


Figure 1: The number of crimes in 2019-2023 (per month)

Figure 2: The crime incident location

In Figure 3, the left plot shows hourly crime counts in Chicago, with peaks around midnight and low points from 4 to 6 AM. Crime rates rise again during the day, following daily activity patterns. This indicates that late-night hours have higher crime risks. Based on Figure 4, the plot breaks down crime by primary type, showing theft as the most common, followed by battery, criminal damage, and assault.



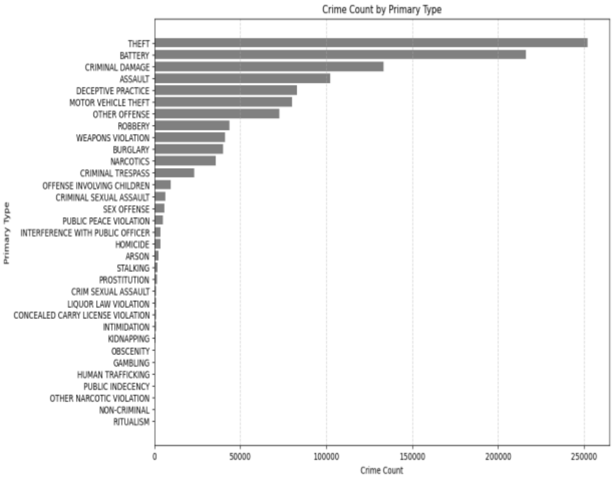


Figure 3: The number of crimes in 2019-2023 (per hour)

Figure 4: The number of primary crime type

Data Table Schema

The data set consists of 1,185,534 rows and 22 columns. Each row represents one crime record.

Column Name	Description	Data Type	Example Value	Notes
ID	Unique identifier for the record.	integer	12045583	Each row has unique ID
Case Number	Records Division Number.	string	"JD226426"	152 rows duplicates (134 unique)
Date	Date when the incident occurred.	timestamp	05/07/2020 10:24:00 AM	Converted to datetime type
Block	Partially redacted address where incident occurred.	string	"035XX S INDIANA AVE"	
IUCR	Illinois Uniform Crime Reporting code.	string	"0820"	Consider mapping its code table to know what it means.
Primary Type	Primary description of the IUCR code.	string	"THEFT"	
Description	Secondary description of the IUCR code.	string	"$500 AND UNDER"	
Location Description	Description of location where the incident occurred.	string	"APARTMENT"	6,693 rows of missing data
Arrest	Whether an arrest was made.	string	false	True/False
Domestic	Whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	string	false	True/False
Beat	Beat where the incident occurred.	string	"0212"	
District	Indicates the police district where the incident occurred.	string	"002"	Consider mapping its code table to know what it means.
Ward	The ward (City Council district) where the incident occurred.	string	"3"	48 rows of missing data
Community Area	Indicates the community area where the incident occurred.	string	"35"	2 rows of missing data.
FBI Code	The crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).	string	"04A"	Consider mapping its code table to know what it mean.
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.	string	1178180	18,715 rows of missing data
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.	string	1881621	18,715 rows of missing data
Year	Year the incident occurred.	integer	2020	2019-2023
Updated On	Date and time the record was last updated.	timestamp	05/14/2020 08:47:15 AM	Converted to datetime type
Latitude	The latitude of the location where the incident occurred.	decimal	41.830481843	18,715 rows of missing data
Longitude	The longitude of the location where the incident occurred.	decimal	-87.621751752	18,715 rows of missing data
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal.	string	(41.830481843, -87.621751752)	18,715 rows of missing data

Data Cleaning and Processing

The dataset contains a few data quality issues that need attention to ensure accurate and reliable analysis. The primary issues identified include missing location data. Additionally, some data are duplicated, requiring careful filtering to retain only relevant entries.

Duplicated data problems:

While each record has a unique ID, some case numbers are duplicated. Upon review, we found that these duplicate entries share identical data values, differing in ID and date. To resolve this issue, we will retain the first instance of each duplicated entry and discard the rest. After this de-duplication process, there are 1,166,663 unique records.

Missing data problems:

Several columns related to location contain missing data, with the missingness pattern appearing random. The following table summarizes the extent of missing values in each relevant column:

Column Name	The number of missing data
Location Description	6,693
Ward	48
Community Area	2
X Coordinate	18,715
Y Coordinate	18,715
Latitude	18,715
Longitude	18,715
Location	18,715

Columns such as Location Description, Ward, and Community Area are not essential for this analysis and should be maintained as missing values. Furthermore, the X Coordinate, Y Coordinate, Latitude, and Longitude columns exhibit the same level of missing data, which may suggest a systemic issue in the data collection process.

Outlier problems:

One record contains a location outside of Chicago, which will be removed from the dataset to maintain geographical consistency.

Column renaming and transformation:

To enhance usability, columns will be renamed to replace spaces with underscores. The "Date" column will be split into separate columns for the year, month, day, and hour, enabling more detailed analysis. Necessary categorical variables will also be converted into dummy variables. The Boolean variables will be changed to 0 and 1.

Summary:

After data cleaning and processing, the remaining dataset contains 1,166,663 rows and 264 columns for further use. Additionally, the dummy variables have many columns, so we will consider addressing or grouping them in further analysis. Other variables like X Ordinate might drop off because they are not the variables we focused on.