**Chicago Criminal Case Analysis**

**MScA 31012**

Dec 10th, 2021

Ann Huang

Bill Wang

Boyuan Wang

Daniel Zhang

Rachel Li

**Table of Contents**

1. **Executive Summary - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 2**
2. **Business Case and Objective - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 2**
3. **Data Models - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 3**
4. **Data Profiling - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 4**
5. **Methodology and Various Tools Used - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 5**
6. **Insights - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 5**
7. **Recommendations and Lessons Learned - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 8**
8. **References - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - 10**
9. **Executive Summary**

As Chicago citizens, we are worried about the crime rate in Chicago, and it inspired us to do an investigation on it. The data set is available on Kaggle, a website with millions of data sets from different backgrounds. We decided to analyze the data and provide insights for stakeholders, both the citizens and the police. During the process, various tools and platforms were used, such as MySQL, Excel, and Tableau, to better understand and analyze the data. From the results, we also did data visualization, which is a more straightforward way to show our results and insights, and better help stakeholders to understand what we are doing.

1. **Business Case and Objective**

Crimes are always the main concern of citizens in every city, especially in Chicago, a city that has a higher crime rate than other US cities. As part of the Chicago citizens, our group was also concerned with the crime rates and got inspired—how did crime rates change over the years? What types of crimes are the most frequent ones? Which neighborhoods of Chicago are the most dangerous ones?

With such questions in mind, we explored official websites for data and chose the data from Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. This dataset includes multiple columns such as type of the crime, date, block, coordinate, community area, etc.

The three main stakeholders of our analysis will be the Chicago Police Department (CPD), real estate companies, and ordinary Chicago citizens.

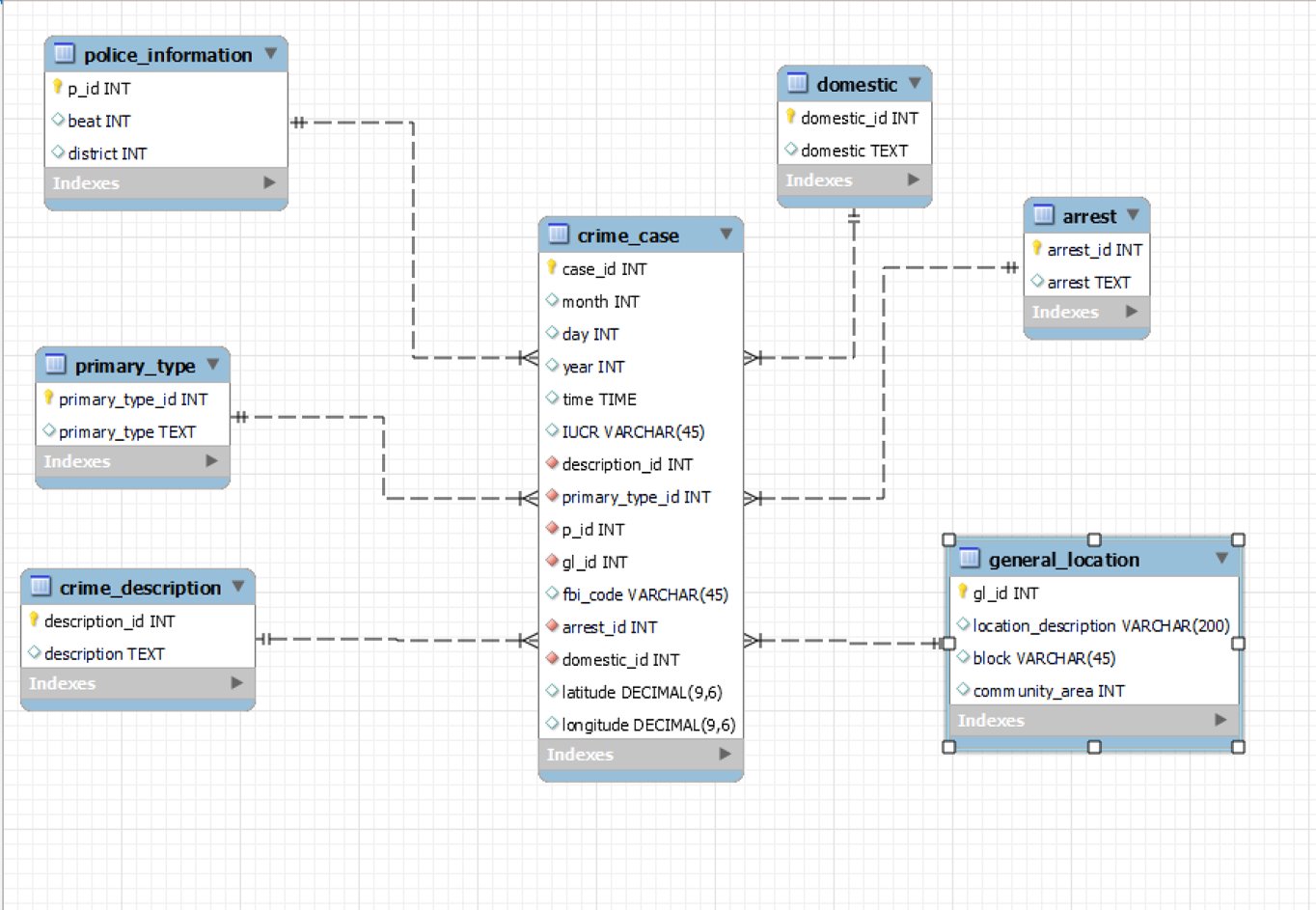
* **CPD**: Results will allow CPD to reference the crime patterns and efficiently allocate the police force to better protect the people.
* **Estate companies**: Results will allow estate companies to know which areas will be safer and to explore the area or adjust the rent and infrastructure.
* **Chicago citizens**: It is always good to know which communities are dangerous during which hours. Our project will help Chicago citizens to have a general sense of how dangerous a particular neighborhood is during a specific period, thus making more informed decisions.

1. **Data Models**

To analyze the crime cases in Chicago, we built a relational data model by using MySQL Workbench. In the original dataset, all columns were put in one table. We then normalized our original dataset into the third normal form and this model contains seven tables. We created an entity-relationship diagram (ER diagram, **Figure 3.1**) in MySQL, and the diagram shows the relationship between each table.

In the normalization process, we first separated the columns based on the information of stakeholders, such as *police\_information* for CPD and *general\_location* for citizens. Then, we normalized the data based on the type of crime cases. This step of normalization separates one table into six different tables and each of them is independent of the other. Next, we created the ID for each table, and these IDs are the primary keys and foreign keys in our relational data model. Finally, we created the *crime\_case* table and connected it with the other six tables. This table contains the most information and was used for the analysis to achieve our objectives. The following is a brief description of the entities in the model:

* *crime\_case*: contains all the cases in the dataset
* *police\_information*: includes the information of CPD, such as beat and district
* *domestic* & *arrest*: whether the case is domestic and whether the suspects are arrested
* *primary\_type*: the primary type of the case
* *crime\_description*: the secondary, more specific description about the type of the case
* *general\_location*: the location, blocks, and the community where the case happened



*Figure 3.1 ER Diagram*

1. **Data Profiling**

We downloaded our dataset from Kaggle, a publicly available website. The original dataset has 23 columns, but some of the columns are repeatable or unnecessary. To better analyze the data, we removed columns such as *ID*, *Ward*, *X Coordinate*, *Y Coordinate*, and *Year*. After cleaning the dataset, we normalized the dataset into a third normal form which contains seven tables. There are no outliers or data quality anomalies noticed, but some of the variables, such as *Beat*, *Community*, and *FBI Code*, are connected to other outside information.

Then, we imported the tables and data into MySQL Workbench for management and analysis. Since we have three stakeholders, we plan to perform crime case count and distributions under different situations. We counted the number of cases under different situations by using the aggregation functions in MySQL queries.

1. **Methodology and Various Tools Used**

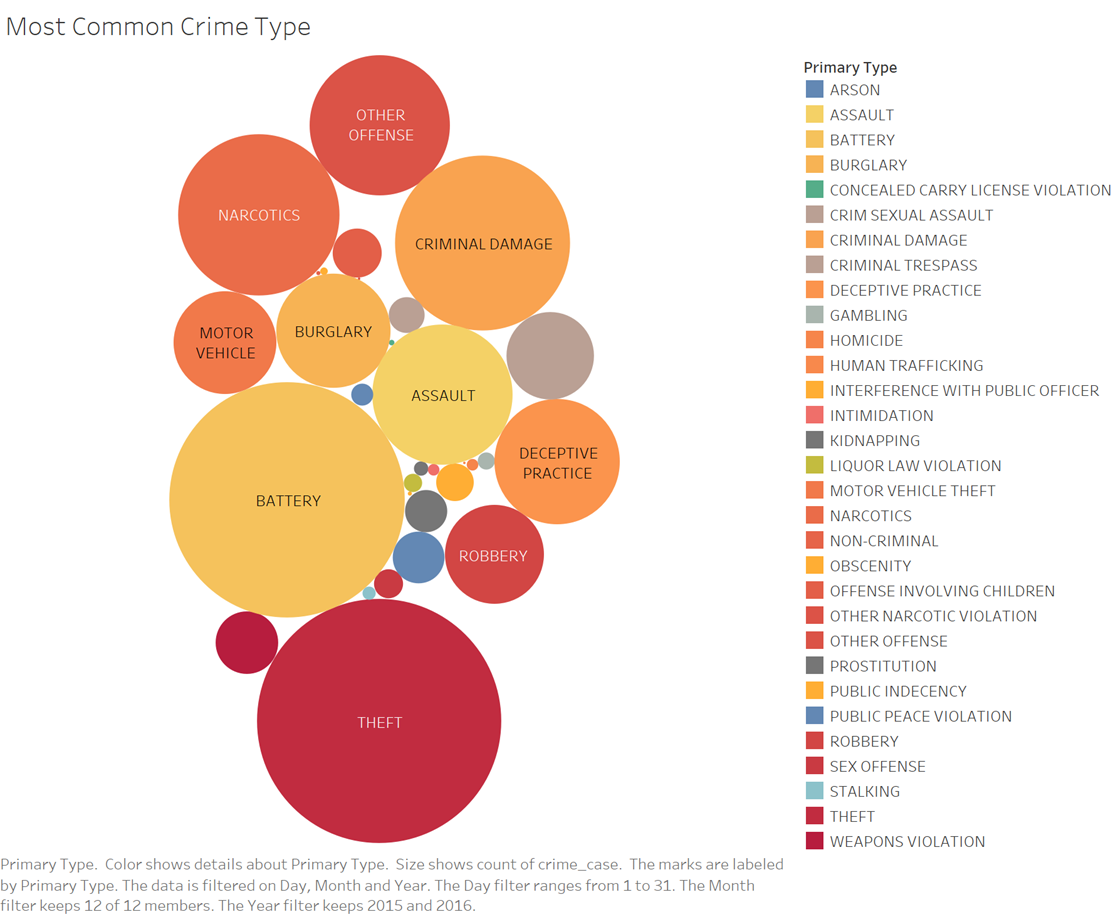
First, we downloaded the data as Excel files and concatenated multiple files into one Excel file. Then we transformed the file into the format of comma-separated values. Using normalization, we divided the original file into different tables, each with a unique id to identify. Excel is the main tool for data management. Then we imported the data into MySQL workbench using the command line because the Table Data Import Wizard in MySQL workbench imports data so slowly and will take hundreds of hours to import all the data. MySQL Workbench is the main tool for us to do data manipulation. At last, using Tableau, we created several different graphs and tables for the data set. Tableau is the main tool to do data visualization.

1. **Insights**

The main objective of our project is to help the people of Chicago understand the occurrence of criminal cases, and make data-driven decisions in protecting themselves. For the major stakeholders including the Chicago Police Department and the everyday citizens, our team has utilized the valuable data to create interactive dashboards.

Our dashboards are categorized into two different sets: one includes tables, lists, and graphs that inspect criminal cases in terms of several key variables like primary type, time, description; the other one focuses on geographic data and visualizes them on the Chicago map.

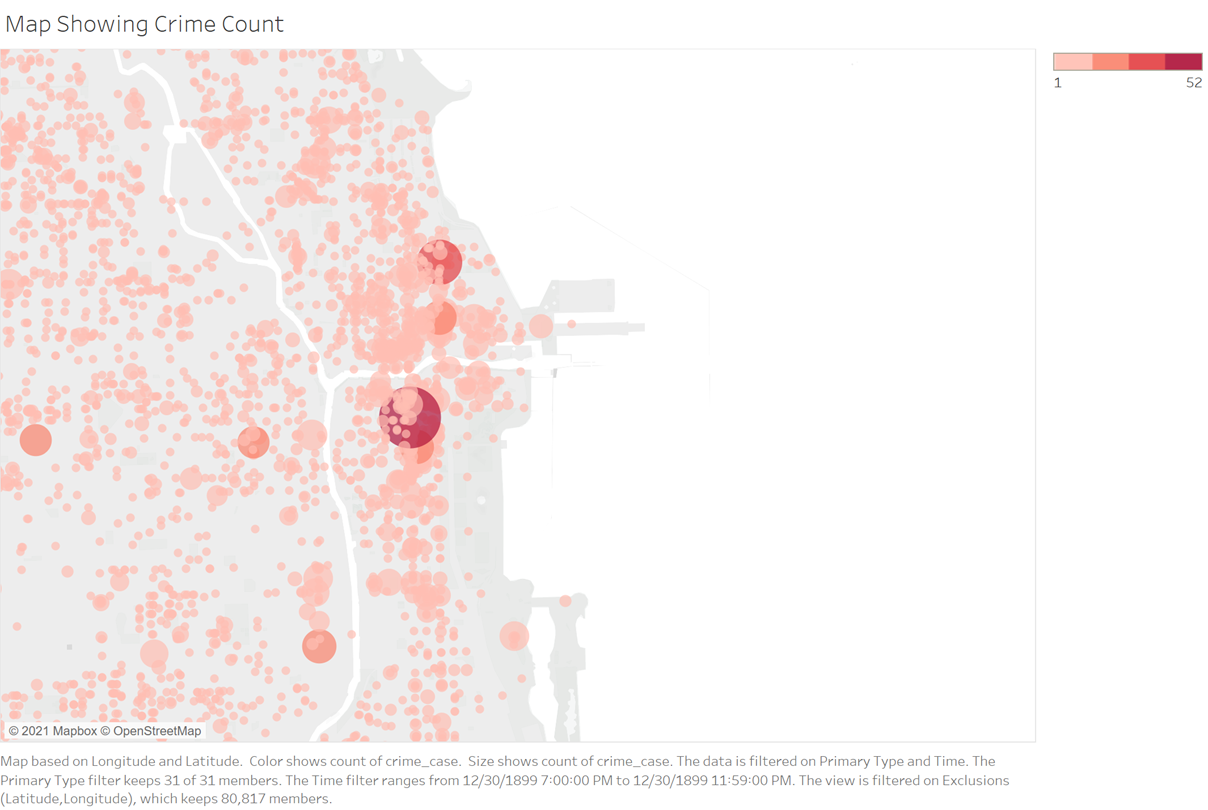
Here is a dashboard example of the first set:



*Figure 6.1. Most Common Criminal Case Type*

Figure 6.1 shows the most common criminal case type in general, providing users an intuitive illustration of what cases one should care about. With a larger circle indicating a higher frequency in the years, users can easily spot the largest circles and draw their own insights from the graph. When combined with time values like specific “hour-minute-second” time, day, month, and year, users can provide their interested timeframe and inspect the most common cases in recent days.

We also include an example from the second set:



*Figure 6.2. Map Showing Crime Count*

The figure above shows how many criminal cases have taken place in the selected area, or more specifically, the particular block. In the figure, we see that different dots have a distinct level of size, indicating how many cases have taken place there. The larger the dot, the more cases. Just by providing the dashboard, we enable users to have a general idea of where cases are taking place and help them make decisions, whether they are about choosing a new apartment or deploying police forces. Furthermore, when combined with time-series values and other variables like case type and description, users can toggle on the filters and obtain the desired result. For example, if a user is interested in the number of theft cases in a certain area, he/she can check the boxes in filters to indicate the interests, and results will be displayed on the map.

A real-life example came from a friend of our team members. The friend planned to move to Chicago next month and approach one of our team members for suggestions. Our visualization maps played a significant role when deciding his apartment. The friend looked at the map and checked several key variables he was most worried about, and he then avoid the areas with big red dots on the map.

1. **Recommendations and Lessons Learned**

For this crime case analysis project, we have some limitations. For example, our dataset only contains data ranging from 2015 to 2016, and it did not provide the motivations for different types of crimes. It would be better to have a larger amount of data with more years and crime motivations shown, as it would help us to develop a model for crime prediction in a certain time with a more accurate result. In terms of data visualization, geographical data such as districts and beats are hard to be shown on the map with each area being colored and labeled rather than only having dots across the map. Also, we are lacking a constructive tool such as a website or system that helps the users to make their residential choices. Thus, we need to consider implementing a tool that has the ability to recommend through data visualization while providing suggestions for the users. Last but not the least, we recommend citizens view more past and current crime information on websites and TVs before making their decisions.

For lessons learned, it is essential for us to understand what each variable in the dataset represents and how it is beneficial for the prediction. In addition to that, we need to stand on the user's perspective for data visualization, as it provides better insights for them to make future decisions. In this project, we had hands-on experience on importing data into MySQL Command-Line and found out that this speed for importing is a lot faster than using MySQL Workbench. Most importantly, we understood and learned that MySQL is a very powerful language for database design and data storage, whereas Tableau is a strong data visualization tool.

1. **References**

* Chicago Crime Data Analysis. (2021). Retrieved 11 December 2021, from https://www.kaggle.com/djonafegnem/chicago-crime-data-analysis/data?select=Chicago\_Crimes\_2012\_to\_2017.csv