Individual Integrated Assignment

Eran Kan Kohen

2nd October 2022

Word Count: 3762

1 Introduction

Chicago Police Department has been one of the pioneers of utilizing computing and data to tackle crime. CPD has been using such information systems since the 1980s. (Buslik & Maltz, 1998) The data management unit of CPD has decided on establishing a brand new database to keep track of crime data. Later, the data will be analyzed to detect key areas to be improved upon. In order to efficiently improve the performance of CPD and gather insights, three key inquiries were highlighted.

1.1 Arrest Rates of Types of Crimes in Different Police Districts

In 2006 it was suggested that the organizational structure in which the police officers are performing can affect their arrest rates (Chappell et al., 2006). The data will be checked to search if such an effect is present in CPD. Random types of crimes will be selected such as ”Theft”, ”Sexual Assault”, and ”Narcotics”. Subsequently, the arrest rates of different police districts will be compared for the selected crime types. For example the arrest rate for "Theft" will be calculated as:

$$\text{Arrest Rate} = \frac{\text{Total Arrested for Theft}}{\text{Total Recorded Cases of Theft}}$$

By comparing the arrest rates of different districts, the CPD could learn which police officers are making unusual amounts of arrest, either higher or lower. A more in-depth version of this analysis could determine which type of crimes have the highest and lowest arrest rates per districts.

1.2 New Mayor, Same PD?

On 20^{th} of May 2019 Lori Lightfoot was elected as the new mayor of Chicago. Due to her experience as a federal prosecutor and close contacts with the police department, police accountability was one of her agenda items (Simpson et al., 2022). Hence, to visualize the difference, crime trends before and after Lightfoot's election will be compared. Different parameters of CPD such as the location description of the crimes and the number of arrests will be highlighted to see if a specific area changed drastically.

1.3 Safest Place to be?

The final inquiry will shed light on the location of the crimes. The data set provides information on the types of area in which the crime occurred (such as street, residence, and sports arena). An analysis into the amount of crimes and arrest can indicate which types of places are the safest and where police influence is the most. By highlighting location types which require more attention, the CPD can pinpoint the types of training they should conduct and where to place their patrols.

2 Design and Organization

2.1 Relevant Entities

In order to construct a functioning database which can also be used to answer the desired questions, only specific variables from the data set should be recorded. Hence, some variables from the initial data are removed. casenumber is not used since CrimeID will be used to identify the rows. Since, specific addressed will be not analyzed, variables block, latitude, longitude, and location will be removed as well.

The remaining variables will be used to create entities which concern the type of the crime, specific details regarding the specific crime, and information regarding the police beat which reported it. The normalization procedure will outline how the entities are created.

2.2 Normalization

It has been stated that database normalization is an important part of establishing a database as it clears unnecessary redundancy, inconsistent data, and improves efficiency (Wang et al., 2010). Moreover, a normalized database will not experience issues with anomalies and has the possibility to lose valuable information (Lee, 1995). Finally, properly normalizing a database to the third normal form ensures that future insertion, deletion, or updates to the database will be easy to implement (Wang et al., 2010).

Hence, it is assumed that after properly establishing the physical model of the database in section 3 and cleaning the data in section 4 according to the normalization features, setting up the database will be very simple. Thus, the model will be checked to align with the initial three forms of normalization.

2.2.1 1NF

The first normal form should make sure that each attribute contains atomic values. In the data set, the variable Date has both the time and the date inside the cell. In order to achieve the smallest possible value, the two values were separated into two different columns. Later, it was decided that, time was not going to be used in the analyses, it is redundant to include it in the database. Hence, the time column was deleted and the first normal form was achieved.

2.2.2 2NF

The second normal form concerns partial dependencies. This is relevant to the database due to the variables concerning the crime type and the location of the crime.

Variables district and beat depend on each other. Each district has specific beats under it. So, the district of a crime depends on the beat. In order to fulfill the second normal form, the two variables will be used to generate a second table with beat as the primary key.

A similar issue occurs with the variables primarytype, description, and iucr. primarytype explains the general classification of the crime, and description further explains the specifications of the crime. Also, the same description could be used for different types of crimes. For example, both robbery and sexual assault can have knife in the description. In order to specify the combination, the iucr is used. The iucr (Illinois Uniform Crime Reporting) codes are used to classify the combination of both the type and the description of the crime. Hence, every iucr has a specific primarytype and description under it.

Since this is a violation under the second normal form, these variables are moved to a tertiary table with the iucr as the primary key. The remaining variables arrest, locationdescription, date, and CrimeID remain in the main table with CrimeID acting as the primary key.

2.2.3 3NF

There are no further dependencies remaining in the data. The third normal form would require that the database lacks any column depending on a non-key value. However, after making sure the database satisfies the initial normalization requirements, it automatically satisfied the third normalization form as well.

3 Database Models

With the normalization processes in mind, the models of the database was created. Initially, a conceptual model was created to show the entities and their relationships.

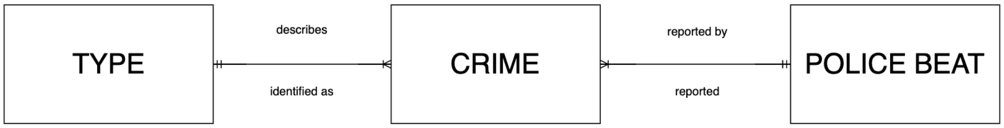


Figure 1: Conceptual Model

As seen above, the conceptual model has three entities called Type, Crime, and Police Beat. The Type entity includes information on regarding the type of the crime and the description of it. It describes specific crimes. The Crime entity has specific information regarding each data point. Finally, the last entity has information on the police beat which reported the crime.

To highlight the variables under each entity, a logical model was created as well.

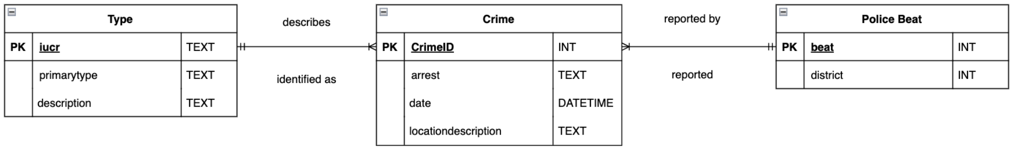


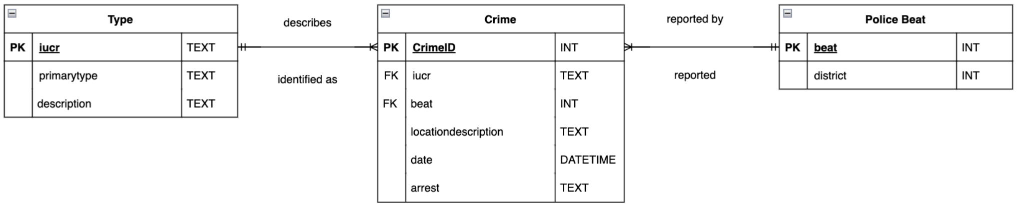
Figure 2: Logical Model

As described under the normalization section, the logical model has three tables with various variables. The table in the middle describes the main aspects of the crime with the date and the type of location being present. The arrest variable indicates whether the suspect was arrested, and the CrimeID variable is used to identify each case.

The table on the left indicates the type of the crime with the iucr as the primary key. The primarytype and descrioption variables are used to further explain the contents of the type of the crime. There is a one-to-many relationship present between the aforementioned tables since a iucr can be used to classify a single crime but a single crime can only have a specific iucr.

The final table on the right concerns the specific police beat in which the crime occurred. The beat is the primary key of the table and the specific district in which the beat belongs to can be found in the table. The Crime table in the middle has a one-to-many relationship with the Poice Beat table since a crime has only a beat recorded but multiple crimes can occur in a single beat.

Finally, before the database is implemented in DB Browser SQLite, a physical model was created to highlight the foreign keys.



Type		
PK	iucr	TEXT
	primarytype	TEXT
	description	TEXT

Crime		
PK	CrimeID	INT
FK	iucr	TEXT
FK	beat	INT
	locationdescription	TEXT
	date	DATETIME
	arrest	TEXT

Police Beat		
PK	beat	INT
	district	INT

Figure 3: Physical Model

The model above is the most detailed form of visualization of the database. It includes the relationship between the entities, the individual variables, primary keys, and the foreign keys. The foreign keys will be used to connect the tables in SQLite and will assist the latter part of this analysis when conducting JOIN operations.

4 Data Cleaning

Before the database was set on SQL, the data set was cleaned with R. The data wrangling consisted of different sections which ensured that the data which was going to be used later was in ideal condition to be analyzed. Moreover, all redundant or misleading parts of the data was tried to removed.

Finally, there are several data quality dimensions. There dimensions are accuracy, completeness, consistency, timelessness, traceability, and accessibility (Eryurek et al., 2021), (Team, 2020) . The dimensions will act as a road map when processing the data.

4.1 Initial Checks

When the data was imported into R, it was made sure that empty cells and cells which only have a space were classified as "NA". This standardization was useful later on when dealing with missing values.

The data was later checked for duplicate row. Out of 730900 rows, 149 rows were completely identical. Hence, the rows were removed from the data. The removal of duplicate rows made sure that data fulfilled the accuracy dimension.

Moreover, since the variable CrimeID was going to be used to identify each row, it was made sure that every value of CrimeID was a unique number. The check showed that there were no duplicates, hence, nothing was affected.

Furthermore, the variables which were not going to be used were removed from the data. With casenumber, block, latitude, longitude, and location removed, the data set consisted of 9 variables.

Finally, since time was not going to be used in the analyses, the variable consisting of the date and time was adjusted and time time was removed from it. Investigating the data also showed that the date was reported in different ways for different ways. In order to achieve complete consistency, the dates were recorded in the format of YYYY-MM-DD.

4.2 Dealing with Missing Values

After the initial check, frequency of missing values was investigated. In order to make sure the data is precise as possible, the rows with more than 1 missing values were removed from the data. This resulted in 3 rows being removed. The remaining rows were checked for missing values.

The check indicated that the only remaining missing values were under the column of locationdescription. Possible ways to remedy the missing values were considered, such as replacing the values with the mode, however since there were only 3393 missing values, the rows were removed.

The collective the removal of the duplicate rows and all of the rows with missing values resulted in the deletion of 3545 data points. It is highly unlikely that this removal deeply affected the data since 99.5% of the remained intact. Moreover, with the missing values removed from the data, the completeness dimension was fulfilled.

4.3 Logical Errors

This section will look at logical inconsistencies and aim to remove or remedy the issues. It will aim to fulfill the consistency dimension so that the logical rules established in the database are not violated.

4.3.1 Aligning with Description

After cleaning the data set, the data was checked for any logical contradictions. For example, the description stated that the data held record from 2017 to 2021. This was checked to make sure the data was not contaminated. Thankfully, the data aligned with the description.

Furthermore, the district numbers were checked to make sure the numbers are within the range of the Chicago Police Department. The data description stated that there are 22 police districts in Chicago. This was surprising considering that the data had a district range from 1 to 31. Since this was a substantial difference, the official Chicago Police website was checked. This revealed that the description was outdated and the department actually had 25 districts ('Police Districts — Chicago Police Department', n.d.).

Following this finding, the data was checked again. It was found that the data actually had districts between 1 and 25 and only a few outliers with the value of 31. Hence, the 5 logically inconsistent rows were deleted. Finally, a pattern among the beat and the district was detected. Each beat number consisted of 3 or 4 digits, and each district number consisted of 1 or 2 digits. The beats corresponding to a specific district started with the district number. For example if the district number is 5, the beat number is 5XX.

This was checked in the data using a for and if loop. It was revealed that there were several rows which did not fit this logical consistency. Hence those 53 rows were deleted from the data.

4.3.2 Outliers

The data was checked for possible outliers, however none were found. The CrimeID numbers aligned with the data size, and as mentioned above, the date, beat, and district values matched the logical ranges. The uicr formatting also matched the ones listed in the official Chicago Police website (‘Chicago Police Department - Illinois Uniform Crime Reporting (IUCR) Codes — City of Chicago — Data Portal’, n.d.). Thus, it was certain that the data did not contain any sorts of outliers or logical contradictions.

4.4 Merging Similar Descriptions

The final check was highly tedious, yet, very useful. An in-depth analysis of the data revealed that the columns for description had several cells with different descriptions explaining the same issue. For example, ”VIOLENT OFFENDER - DUTY TO REGISTER” and ”VIOLENT OFFENDER: DUTY TO REGISTER” were two separate descriptions. Such issues were detected and they were combined into a single description. This process was done manually and more than 20 descriptions were combined.

The same investigation was done for primarytype and the same issue was detected. Hence, the same procedure was conducted for primarytype to merge similar cells. Moreover, this was much easier compared to the previous once since there were more than 450 unique descriptions, however, only 93 unique primarytype classifications.

It should be mentioned that there are hundreds of descriptions so, it is likely that several issues still remain in the data. This issue will likely go on until the Chicago Police Department tidies up the achieve or several hours are put in to comparing more than 450 unique descriptions.

Later in the analysis, it was discovered that several IUCR values are connected to descriptions with different notations similar to the example given above. Since this damaged the primary key role of IUCR, each row was checked automatically and the descriptions were aligned. As stated above, the descriptions were not checked due to the abundance of data points, however, it is assumed that this process merged most, if not all, of the similar descriptions. This made the data ready to be used in the database.

4.5 Referring Back to the Dimensions

After the cleaning procedure, it is safe to assume that positive progress was achieved in achieving the dimensions of completeness, consistency, and accuracy. However, several dimensions were not referred to.

First of all, the traceability and accessibility dimensions are completely in the hands of the Chicago Police Department. The data was pulled from the official website, and, as of now, the data is publicly accessible. However, it is possible that the data will be modified or be locked behind limited access. Such issues could deteriorate the traceability or the accessibility of the data. Moreover, it is possible that the Chicago Police Department will move the data to another website which could hurt the traceability of the data.

Finally, the timelessness dimension is not applicable to the current analysis. The aim of this study is to analyze the crime data between the years of 2017 and 2021. The current data is not late or unusable due to an issue. The only possible issue regarding this dimension could occur if the Chicago Police Department changes the borders of their beats or districts, or the borders of the city of Chicago changes. If a border change were to occur, the beats and districts referred in the data may concern outdated allocations. Hence, it is important to remember that the database deals with the beat and district borders which were used in October 2022.

5 Implementing the Database

The database was implemented rather easily. The Physical Model (Figure 3) shown in Section 3 was directly created by populating three tables.

There were no errors of primary keys not being unique, probably due to the extensive data cleaning, and all of the desired data types matched the values imported.

Moreover, when creating the database, it was seen that the date recording method was not constant in the data set. This was handled in the data cleaning part. In order to make sure such an error would not occur in the future, a trigger was set in the database which raises an error when a new row is inserted with an invalid date format.

6 Results

6.1 Question 1

The first question focused on the arrest rate in different districts. Initially, the arrest rates for each type of crime was checked. Tables which show the highest and lowest arrest rates were generated.

Table 1: Arrest Rates (High)

Primary Type	Arrest Rate
Public Indecency	100.0
Liquor Law Violation	100.0
Prostitution	99.8
Narcotics	99.76
Gambling	99.44
Concealed Carry License Violation	97.77
Interference With Public Officer	94.86
Obscenity	79.7
Weapons Violation	66.69
Other Narcotic Violation	61.9

As seen above, crimes such as public indecency and gambling have the highest arrest rates in all of Chicago.

Table 2: Arrest Rates (Low)

Primary Type	Arrest Rate
Human Trafficking	0.0
Deceptive Practice	3.65
Burglary	5.03
Criminal Damage	5.4
Intimidation	5.53
Motor Vehicle Theft	5.61
Kidnapping	6.84
Criminal Sexual Assault	7.65
Non-Criminal	7.69
Robbery	7.73

The table above shows the crimes with the lowest arrest rates. It is very interesting that human trafficking has a 0% arrest rate, either the CPD is terrible at detecting human trafficking cases or the possible suspects are always let go by the judiciary department. Moreover, other crimes such as burglary and robbery have less than 10% arrest rates.

In order to to analyse the districts, similar tables were generated by adding district as the grouping variable. Moreover, it was likely that a specific crime had occurred only a handful of times in a specific district hence it could have a 100% or 0% arrest rate with a small sample size. To prevent this from happening, the tables below only include crime types which have occured in a district more than 300 times.

Table 3: Arrest Rates in Districts(High)

District	Primarytype	Arrest Rate
11	Prostitution	100.0
8	Narcotics	99.92
10	Narcotics	99.92
9	Narcotics	99.9
3	Narcotics	99.89
2	Narcotics	99.87
7	Narcotics	99.86
11	Narcotics	99.85
16	Narcotics	99.83
22	Narcotics	99.83
4	Narcotics	99.77
15	Narcotics	99.73
12	Narcotics	99.71
5	Narcotics	99.64
6	Narcotics	99.61
24	Narcotics	99.44
1	Narcotics	99.33
25	Narcotics	99.31
18	Narcotics	98.98
4	Interference With Public Officer	97.05

As very clearly seen above, every district seems to unanimously have a very high narcotics arrest rate. This is reassuring since it indicates that the districts are working harmoniously and following a clear guideline. In order to make sure this is consistent, the lowest arrest rates were checked as well.

Table 4: Arrest Rates in Districts(Low)

District	Primarytype	Arrest Rate
22	Deceptive Practice	1.81
15	Deceptive Practice	1.92
17	Deceptive Practice	1.94
20	Deceptive Practice	2.01
5	Burglary	2.2
24	Deceptive Practice	2.22
25	Deceptive Practice	2.4
4	Burglary	2.48
8	Deceptive Practice	2.49
2	Deceptive Practice	2.62
3	Burglary	2.68
9	Deceptive Practice	2.85
12	Deceptive Practice	2.91
15	Burglary	2.94
4	Deceptive Practice	3.03
16	Deceptive Practice	3.03
19	Deceptive Practice	3.04
8	Burglary	3.1
2	Burglary	3.14
5	Motor Vehicle Theft	3.14

The second table mirrors the initial one since it also has deceptive practice and burglary as the crimes with the lowest arrest rate for many districts. Again, it seems that the CPD is working harmoniously and treating the same kinds of criminals in the same way.

6.2 Question 2

The second question primarily focused on the new mayor of Chicago, Lori Lightfoot, getting elected. Initially, the dates on which the most crimes occured before and after her election was checked. It looked like the busiest days before her election had around 600 crimes being recorded per day whereas after her election the busiest days had around 500 crimes recorded per day, except a scary day when 1172 crimes were recorded. Also, looking at the arrest data showed that, on a the busiest days, the arrest numbers are very similar in both periods. Furthermore, when investigating the arrest numbers of before and after, indices were created with the columns data and arrest which were filtered by the election of Lightfoot. Hence, the analyses took much less time since the code only checked the desired columns in the desired time frame.

Secondly, the investigation was expanded as it analyzed the difference of crimes recorded per location types before and after her election. It should be mentioned that she was elected on 20^{th} of May 2019 and the database has crimes recorded between 2017 and 2021. Hence the dates are very evenly divided with her election date in the middle. There are 29 months and 20 days before her election, and 30 months and 10 days after her election. Since these time frames are very similar, comparing the total amount of crimes recorded can be justified.

The 5 types of locations which had the highest and lowest number of crimes recorded were checked. It revealed that apartments had much more crimes recorded compared after Lightfoot's election. This makes perfect sense since her time in the office happened to be in the same time as the COVID-19 pandemic. It is expected that more crimes occurred at private property than public places. Similarly, much less crimes were recorded on the sidewalk after her election, aligning with the COVID-19 theory. Moreover, the results showed that the location type "Other" has much less crimes recorded whereas "Other (Specify)" had much more crimes recorded after Lightfoot's election. It is possible that the CPD police department changed how they recorded "Other" cases sometime around 2019.

Finally, the total amount of crimes were divided by the months before and after Lightfoot's election and compared.

Table 5: Crimes per Month Comparison

Crimes per month Before	Crimes per month After	Difference
13008.0	11276.0	-114.0

As seen above, there are, on average, 114 less crimes recorded per month since Lightfoot's election. This indicates that the crimes recorded have stayed pretty much the same since it is less than a 1% decrease. However, considering the findings outlined above, it is possible that the Lightfoot government focused on securing different types of locations.

6.3 Question 3

Finally, the last question aimed to find the safest type of location in Chicago. It aimed to look at which types of locations host the greatest number of crimes and compare their arrest rates. It also looked at which types of location have the greatest number of murders recorded.

First of all, it was discovered that the most amount of crimes occur on outdoor places like streets, alleys, or restaurant as well as residential places like apartments and residences. From these places, the arrest rate is much higher in public places like alleys (29%) and sidewalks (35%) compared to apartments (12%) and residences (10%). However, things are much different for murder cases. In private places like apartments (63%) and houses (50%) the rates are much higher compared to public areas like streets (26%) and alleys (27%). A great deal of murders in Chicago also occur on the streets and not in private properties.

Secondly, the reverse was checked to find out the types of locations with the least number of crimes recorded. To eliminate descriptions which are very specific, only descriptions with more than 5 crimes committed were viewed. It showed that places like a party bus, a trolley bus, or a gas station has the lowest number of crimes recorded. Of course it can never be sure if this is due to less crimes occurring there or no police officers being present and able to record the crimes. The number of murders were checked as well to find out that the least amount of murders are in vehicle parking sports like garages, parking lots, and park properties. However, when all the values combined, it could actually indicate that a significant proportion of the murders occur in parking ares.

Finally, a window function was used to compare the arrest rates for every location to the total arrest rate in Chicago. It showed that places like banks (8%) and commercial offices (8%) have a much lower arrest rates than the average (25%). On the other hand, grocery food stores (41%) and liquor stores (39%) have much higher arrest rates. A view of the table was created to more clearly visualise the differences. This view is accessible through the database.

References

Buslik, M., & Maltz, M. (1998). Power to the people: Mapping and information sharing in the Chicago Police Department [Publisher: Citeseer]. *Crime Mapping and Crime Prevention. Crime Prevention Studies*, *8*, 113–130.

Chappell, A. T., MacDonald, J. M., & Manz, P. W. (2006). The organizational determinants of police arrest decisions [Publisher: Sage Publications Sage CA: Thousand Oaks, CA]. *Crime & Delinquency*, *52*(2), 287–306.

Chicago Police Department - Illinois Uniform Crime Reporting (IUCR) Codes — City of Chicago — Data Portal. (n.d.). Retrieved September 25, 2022, from https://data.cityofchicago.org/widgets/c7ck-438e

Eryurek, E., Gilad, U., Lakshmanan, V., Kibunguchy-Grant, A., & Ashdown, J. (2021). *Data Governance: The Definitive Guide* [Google-Books-ID: jQYiEAAAQBAJ]. "O'Reilly Media, Inc."

Lee, H. (1995). Justifying database normalization: A cost/benefit model. *Information Processing & Management*, 59–67.

Police Districts — Chicago Police Department. (n.d.). Retrieved September 25, 2022, from https://home.chicagopolice.org/about/police-districts/

Simpson, D., Rossi, M. R., & Gradel, T. J. (2022). City Council Buries its Rubber Stamp Chicago City Council Report# 13 June 12, 2019–March 23, 2022.

Team, C. T. (2020). CESSDA Data Management Expert Guide [Place: Bergen, Norway Publisher: CESSDA ERIC]. https://doi.org/10.5281/zenodo.3820473

Wang, T. J. (, Du, H., & Lehmann, C. M. (2010). Accounting For The Benefits Of Database Normalization. *American Journal of Business Education (AJBE)*, *3*(1), 41–52. https://doi.org/10.19030/ajbe.v3i1.371