# Data Management and Ethics - Individual Integrative Assignment

Thanh Dat Nguyen - 532618tn

2 October 2022

# Project Outline

Ensuring public safety and fighting crime are the main purpose of the police department, but the unpredictability of crimes and increasing public hostility in the US towards the institution can make maintaining this purpose a challenging task. After the death of George Floyd, the US experienced an upsurge of protests against racial injustice and police brutality, which undoubtedly worsened the relationship of the public with the police (Altman, 2020). The following report makes use of analytics to gain further insights into the public opinion on the institution and potential ways to improve it through better resource allocation. To achieve this, the paper is centered around answering a set of questions by using the data at hand:

1. How has the quantity of crimes against police officers developed over the years?

First, in light of the recent incidents involving the cases of police brutality in the US, it may be insightful to study the trend in crimes against police officers. This may be interesting for the police department to keep track of to see the development of its public image and take precautions to protects its employees, the police officers.

2. Which police districts experience the highest amount of index crimes?

These are crimes recorded by the FBI for its uniform crime reports and include more serious crimes such as murder and robbery (CPD, 2021). Answering the question allows the police department to better distribute its forces, with the more experienced officers being assigned to these problematic districts. Index crimes are also high-impact so knowing where they happen can help address them and help the citizens feel safer. It also improves the image of Chicago, as these are the crimes that are accounted in FBI's semi-annual report.

3. Which police districts have the lowest arrest rate?

This is again useful for the police department to know in terms of resource allocation, as arrest rates are one of the standard performance metrics for police departments (Sparrow et al., 2015). A low number of arrests can imply that further investigation of the district by the police department is needed.

4. Which areas experience the highest amount of burglaries?

According to our data, burglaries are the 9th most common criminal offense in Chicago. It also has a major psychological impact on the victim, making them feel unsafe at home (Maguire, 1980). Furthermore, burglaries are only solved also 13% of the time (Covington, 2022). As such, it can be beneficial for the police department to monitor this and have increased patrols in areas with higher rate of burglaries to increase public trust.

# Database Design

As far as entities are concerned, we require knowledge of three: cases (this details information about every single crime that occurred in the past five years - necessary for all questions), crime types (this lists all the different types of crimes and their descriptions - necessary for questions 1, 2 and 4) and locations (this outlines the district as well as the geographic information needed to describe an area, which is required for questions 2 and 4).

Keeping these entities in mind, for cases, CrimeID is required for identification and relational database reasons, while date and arrest are key to answering the posed questions. The variables case number and location description can be left out from the analysis, as case number serves the same purpose as CrimeID (uniquely identifies each crime) and location description is not needed for answering our questions.

Next up, the variables to identify the crimes are required for questions 1, 2 and 4. For this we need variables IUCR, the primary type of crime and the description from the provided data set. Additionally, an external data set has to be consulted to classify the crime types into index or non-index crimes, which is pulled from the Chicago Data Portal (CPD, 2021).

Finally, for variables concerning the geographic area, we require the district number, which is key to answering two of the questions. As the smallest (geographic) unit of analysis in the study are police districts, police beat information is unnecessary in the data set. Furthermore, as the analysis focuses on geographic variables, which can be used for map visualization purposes to answer question 4, longitude and latitude are included. Location is excluded as it is simply a combination of latitude and longitude (and it is also redundant from a normalization standpoint) and block is also excluded, since it is unnecessary for answering any of the questions.

## Normalization

The data set in its current form could be considered to not conform to 1NF, because there are certain values that are not atomic (Watson, 2022). Technically, the date variable is made out of four components - month, day, year and time. Since SQL can work with this type of data as a time-stamp, it can be considered an atomic value (time is discarded though). Therefore, we will proceed with the data set as having conformed to 1NF. The primary key in this case is CrimeID. There are, however, duplicate entries in the data set, which would need to be deleted for CrimeID to be a unique identifier of each row. There are no foreign keys at this stage as there is still only one table at this stage. 2NF is automatically satisfied as the primary key is not a composite key.

For 3NF to be satisfied there cannot be any non-key functional dependencies in the table (Watson, 2022). This condition is violated as primary type and description are dependent on IUCR and repeated multiple times in the table. As such a separate table (crime types) with IUCR as a primary key needs to be created. By doing this, the IUCR variable in the original table becomes a foreign key. The newly created table would also have the added variable Index. Another case of this violation would also be the variables district, latitude and longitude. District is dependent on the composite key of latitude and longitude. Thus, it makes sense to create a separate table for location, which includes these attributes. To eliminate issues that come with working with a composite key, a variable LocationID should be created, which serves as a primary key in the new table and foreign key in the old one.

# Data Models

Firstly, the conceptual model is presented. As mentioned, there are three entities in the database: case, crime type and location. The relationships between these are as follows:

1. A specific case can only have one location (and is required to have one), but a particular location can host multiple cases. As a location is registered in the database only if a crime happens there, we will assume a mandatory one to mandatory many relationship.
2. Crime types on the other hand are defined by the IUCR, which is based on an external codebook created by the FBI. As such, not every type defined this book has to be present in the Chicago crime cases, even though it most likely is. Every case, however, needs to have the IUCR and crime type defined. Every case can also only be filed under one type. As such, the relationship is depicted as an optionally many.

Figure 1 shows the conceptual model:

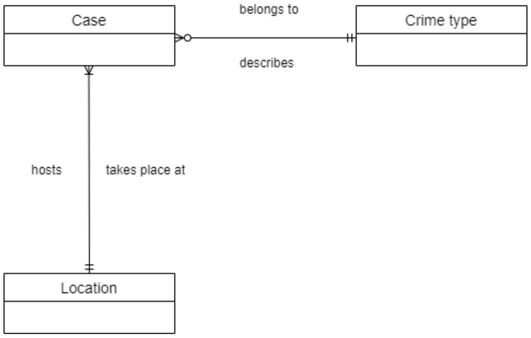


Figure 1: Conceptual Diagram of the Crime Database

Secondly, the logical model is discussed. For the table Case, the variable CrimeID was chosen as the primary key, which is readily provided in the database. It was chosen as it uniquely denotes each case that happened in Chicago between the years 2017 and 2021. Each crime is described with two attributes - the date of when it happened and whether it lead to an arrest (converted into the binary variables 1 for arrest and 0 for no arrest , hence the INT data type).

For the table Crime Type, the IUCR code is chosen as a primary key. It uniquely denotes each crime type and is an alphanumerical variable, hence TEXT. The primary type attribute denotes the general type of crime at hand - e.g. battery, assault, robbery etc. The description attribute specifies the crime with information such as whether the damage was done to a person or property. The index attribute denotes whether the crime is an index crime or not (1 for index, 0 for non-index).

For the table Location, the identifier LocationID is chosen as the primary key - this is for processing reasons as it is easier to work with one key than composite keys. Latitude and longitude are relatively self explanatory and the variable district just describes the police district for the location.

Figure 2 shows the conceptual model:

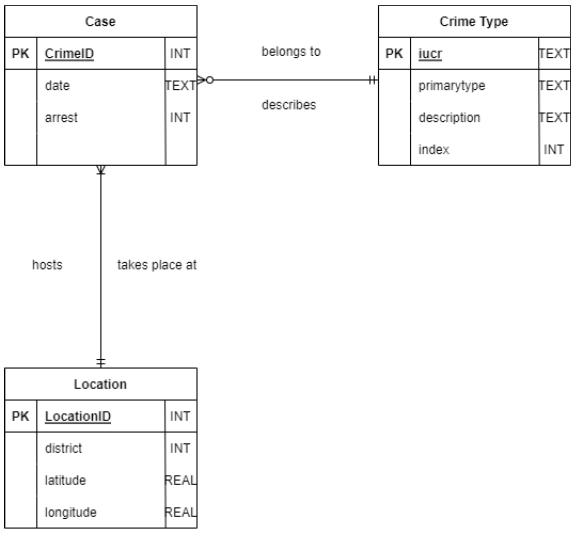


Figure 2: Logical Diagram of the Crime Database

Finally, the physical model is developed. Since there are no many-to-many relationships present, the easiest way is to declare foreign keys in the appropriate entity tables. As both relationships present relate to the table Case, it makes sense to place the foreign keys into this table as well. The final product is presented in Figure 3.

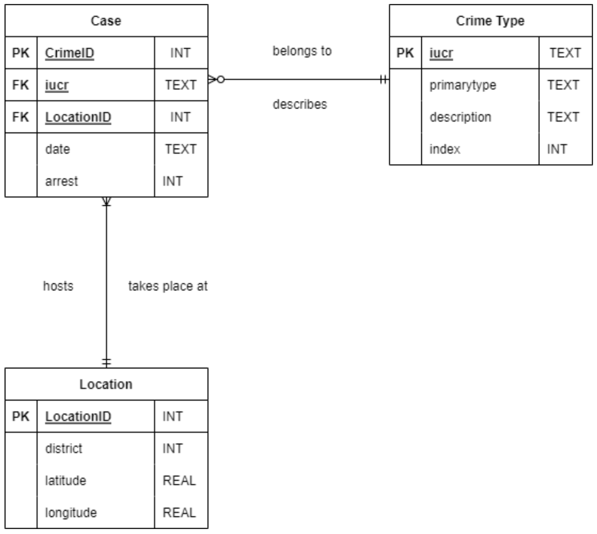


Figure 3: Physical Diagram of the Crime Database

# Data Quality Checks

This task is performed with R. Following standard quality procedure steps (Wickham, 2014), data deduplication is first performed - multiple records of the same crime are removed. A duplicate check is also performed solely on CrimeID (the primary key) and removed if possible (only one such row, which had Date in the wrong format and no other information - removed). An additional duplicate check also needs to be performed on the IUCR codes and the corresponding primary types and descriptions. It is found that there are different primary types and descriptions, which share the same code (they refer to the same crime type, but have differences in descriptions).The next transformation, therefore, unifies these types and descriptions. A similar check is also run on latitude and longitude and a similar problem occurs - there is a number of locations, which, despite having the exact same spatial data, belong to different districts - this happens on boundaries of different districts. It is therefore assumed that the multiple district entries for the same location variables are intentional and as such are kept in the database as is (LocationID differentiates these types of locations).

Next up, the outliers are investigated. A summary table suggests there may be outliers among the CrimeID variable, as there are both non-integer (0.697) and extremely high *1e18* values present. Both these rows are removed from the data, as they do not contain any information. While exploring the data based on CrimeID, it is also noticed that some IDs are stored as real numbers - the variable is therefore converted into integer. The other variables are in the correct format (although arrest is converted into 1 and 0). I also put the date into SQL-legible format: YYYY-MM-DD.

Next, the missing values in the data set are dealt with. It is noticed that there are both NA values and blank values in the data set, so first all the blank values are converted into NA values.

First, let us inspect the missing values for Date. When filtering on these missing dates, it is noted that the corresponding observations also have missing arrest data. Since the crimes are given an ID in a chronological order, we replace the missing dates with the dates of the preceding crime (this is also the same as the date of the following crime). For the missing arrest values, these are replaced with the average arrest rate for that given IUCR and then either rounded up or down to generate a binary value of 1 or 0. This is due to the assumption that similar crime types have a similar arrest rate.

Next, we deal with the variable district, which has a missing value for CrimeID 698092. Here, the NA value is simply replaced with the district number of another crime, which has the same longitude and latitude.

Finally, the missing values for latitude and longitude are resolved. To deal with the issue, it is decided to use the average latitude and longitude of the next smallest geographical object, which in this case would be block. This operation involves making its data format consistent by converting all of its values to uppercase letters (there were a few observations, which had the blocks recorded in lowercase letters). There are still some missing data though, so we perform the same operation using the next smallest geographical object (beat).

After performing these operations, we now have a clean data set. Before exporting the data set as a CSV file, one last transformation is performed, which merges the given data set with the separate IUCR data set, which lists whether a crime is considered index (1) or non-index (0) according to the FBI. When joining the tables, it is found that certain IUCR codes in the original data set have no counterparts in the external data set and, therefore, are assigned to non-index crimes as they are all smaller crimes. Finally, redundant variables identified in task 2 are removed.

## Implementing the Database in DB Browser

Firstly, the table Crime type is created with the CREATE TABLE command. IUCR is set as the primary key with the constraints of being not null and unique. Since IUCR serves as the primary key for the crime type table, it is absolutely vital that it is unique and not null. The rest of the variables are set to have the constraint not null - since during the data cleaning step in Task 4, I either removed or replaced all missing values. In fact, all the variables in the other two tables have the constraint NOT NULL. All the variables also retain the correct data type from importing (so primarytype and description are TEXT, index is INT) except for IUCR, which is changed from the ”Database structure” tab. The exact commands to create the table can be found in the ”Execute SQL” tab ”Crime type”.

Next, the table Location is created. LocationID is the primary key for the table and it was created with the AUTOINCREMENT command in SQL. Finally, the table Cases is created (see tab ”Cases”) - this table is left as last to be created as it contains two foreign keys (IUCR, LocationID) and the foreign key constraint would therefore not work if the other tables were not created earlier. CrimeID here is defined as the primary key with the NOT NULL and UNIQUE constraints.

Now that the tables have been created, we can populate them with a combination of the INSERT INTO (Table) and SELECT DISTINCT command from the table $tidycrimes_fin$. This process is relatively straightforward, except for the Cases table - since the LocationID variable is not present in the original CSV file (as I had decided to create this variable in SQL during Task 5), it is necessary to first use the JOIN command, which merges the $tidycrimes_fin$ and Location tables.

# Results

## How has the quantity of crimes against police officers developed over the years?

To answer this question (see Question 1 tab for all the queries), first the SELECT command is again used to extract the year, month (with strftime) and the total amount of crimes against the police (with count function). The JOIN command is used to merge the tables Cases and Crime type since these are the entities of interest in the question. The WHERE command is used to condition the table to only show crimes, which contained the word OFFICER in either the primary type or the description (using the LIKE command). Results are then grouped by Year and Month. The full table can be generated in the project file but here only the query grouped by Year is shown for the sake of clarity.

Table 1: Yearly Police Crimes

| Year | Month | Crimes against police |
|---|---|---|
| 2017 | 01 | 697 |
| 2018 | 01 | 834 |
| 2019 | 01 | 991 |
| 2020 | 01 | 477 |
| 2021 | 01 | 242 |

As can be seen the overall trend is actually a decreasing one. A plot created in DB Browser can also be used to visualize the entire development (created with the help of the query on line 9 and then using the view Plot functionality).

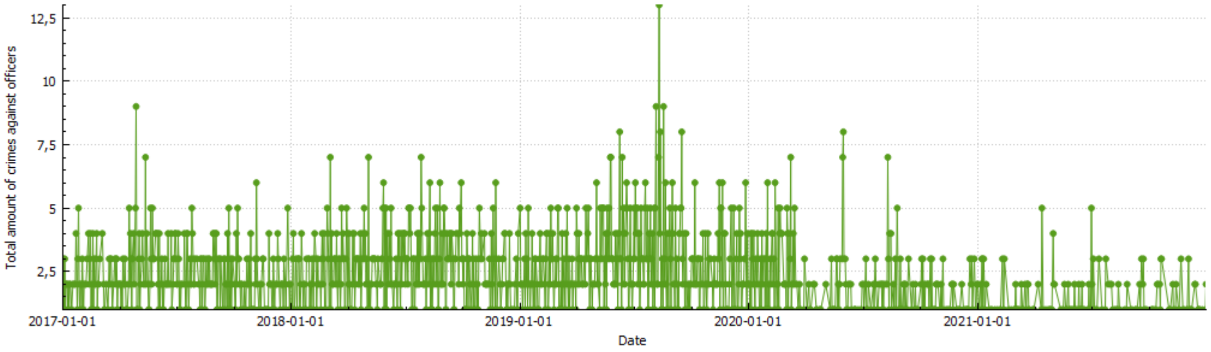


Figure 4: Development of Crimes against Police

The plot confirms the previous findings, wherein the crimes against police officers are generally decreasing. This could, however, be due to the pandemic. It is also possible to zoom in on the last two years and see how transgressions against officers developed then (figure 5). Overall, this type of crimes decreased, but there were several spikes throughout year 2020.

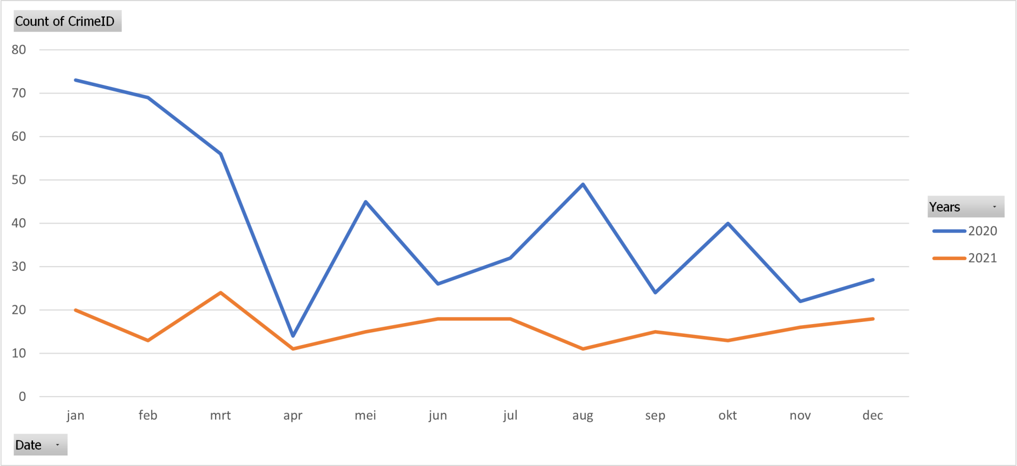


Figure 5: Development of Crimes against Police - 2020/2021

## Which police districts experience the highest amount of index crimes?

To answer the above question, the SELECT command was used to extract the variable district and to create a new variable ”Total number of index crimes” with the help of a COUNT function (see tab Question 2). Since the question uses all three entity tables to answer the question (district from Location, index crime from Crime Type and amount from Cases), it is necessary to use the JOIN operator to merge the first two tables to the Cases table on their primary and foreign keys, which are LocationID and IUCR. While I am at it, I will save this merged table with the CREATE VIEW command to avoid having to replicate the JOIN query every time (see line 13 of tab Question 2). After merging the tables I also use the GROUP BY command to group the results per district and the ORDER BY command to order them based on the total number of index crimes per district. I also used the HAVING command to limit the results to only districts with more than 15000 index crimes total. The results are presented in Table 3 below.

Table 2: Simple Index Crime Table

| District | Total number of index crimes |
|---|---|
| 18 | 22101 |
| 1 | 21434 |
| 12 | 18571 |
| 8 | 17803 |
| 6 | 17595 |
| 19 | 17316 |
| 11 | 15682 |

Table 3: Index Crime Table Including Arrest

| Ranking | District | Index Crimes | Non-arrest rate |
|---|---|---|---|
| 1 | 18 | 22101 | 0.89 |
| 2 | 1 | 21434 | 0.87 |
| 3 | 12 | 18571 | 0.93 |
| 4 | 8 | 17803 | 0.92 |
| 5 | 19 | 17316 | 0.91 |
| 6 | 6 | 17595 | 0.89 |
| 7 | 11 | 15682 | 0.92 |

As can be seen the highest amount of index crimes takes place in districts 18, 1 and 12. It may be more interesting to study how the districts compare when it comes to the amount of unsolved crimes. This question is answered by adding a RANK query to the original one, which ranks the districts based on the number of unsolved index crimes (or rather crimes that did not lead to an arrest, but for sake of simplicity we will assume these two are the same). For additional information, I also add a non-arrest rate column to the table. I use LIMIT to only show 7 observations to make it comparable to the original table - see line 21 in tab Question 2 for the entire query.

Not much has changed, except that in terms of crimes that did not lead to an arrest, district 19 is now above 6. It is also staggering to see that almost 90% of all the serious/violent crimes go unsolved or do not lead to an arrest in all of the districts.

## Which police districts have the lowest arrest rate?

For this question, I first created an index based on the arrest column of the Cases table, since in the following queries I use this column to separate the database into cases with and without arrests. I use the PRAGMA command to check, whether my index for the Cases table was created and following that, I extract the information required. Here I present two tables - one ranking showing the top 10 districts (line 6, tab Question 3) with the highest arrest numbers and then a similar table but with rates (line 24, tab Question 3). Results are both grouped by district and ordered by the corresponding metric. In the second table the RANK command is used twice to create and compare rankings based on total arrest and arrest rates. I also use the EXPLAIN QUERY PLAN to check whether the index was used, which is indeed the case.

Table 4: Total Arrests by District

| District | Total arrests |
|---|---|
| 31 | 2 |
| 20 | 1649 |
| 17 | 2177 |
| 24 | 2800 |
| 14 | 3215 |
| 16 | 3566 |
| 22 | 3666 |
| 19 | 3986 |
| 2 | 4344 |
| 12 | 4501 |

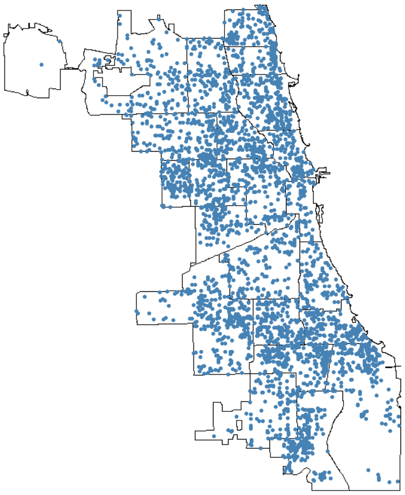
Table 5: Arrest Rates by District

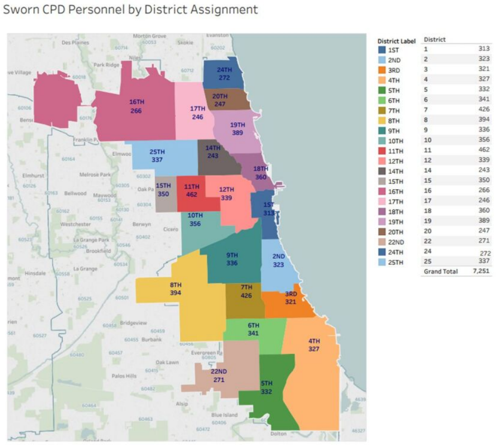
| Ranking - rate | Ranking - total | District | Arrest rate |
|---|---|---|---|
| 23 | 23 | 31 | 0.08 |
| 22 | 21 | 17 | 0.106 |
| 21 | 12 | 19 | 0.12 |
| 20 | 9 | 12 | 0.123 |
| 18 | 22 | 20 | 0.124 |
| 19 | 20 | 24 | 0.124 |
| 17 | 17 | 14 | 0.128 |
| 16 | 14 | 2 | 0.134 |
| 14 | 18 | 16 | 0.144 |
| 15 | 3 | 8 | 0.144 |

As can be seen, the district rankings can vary wildly based on whether we are looking at the arrest rates or the arrest numbers - districts 12 and 8 are both in the top 10, when it comes to the total number of arrests, but are lacking when it comes to arrest rates. When comparing them to results from the section above, there is an overlap between low arrest rates and the number of index crimes - this suggests that index crimes may be harder to resolve. Ignoring these two districts and district 31, the districts that majorly underperform on both the number of arrests as well as the arrest rate include district 17, 20, 24, 14, 2 and 16.

## Which areas experience the highest amount of burglaries?

Before answering the question, I ran SELECT queries to extract the CrimeID, latitude, longitude, primary type and description from the merged table of all three entities, where the primary type was burglary (see Question 4 tab for full queries). This gives a quick overview of the data and informs us that there have 29942 in Chicago over the last 5 years. The next query generates a view which outlines unsolved burglaries (a total of 28375 - an overwhelming majority). This view is then extracted as a CSV table, so we can create a visualization presented in the below figure. The figure only includes burglaries, which took place in 2021 to limit the number of observations and to make the map legible.





| District Label | District | |
|---|---|---|
| 1ST | 1 | 313 |
| 2ND | 2 | 323 |
| 3RD | 3 | 321 |
| 4TH | 4 | 327 |
| 5TH | 5 | 332 |
| 6TH | 6 | 341 |
| 7TH | 7 | 426 |
| 8TH | 8 | 394 |
| 9TH | 9 | 336 |
| 10TH | 10 | 356 |
| 11TH | 11 | 462 |
| 12TH | 12 | 339 |
| 14TH | 14 | 243 |
| 15TH | 15 | 350 |
| 16TH | 16 | 266 |
| 17TH | 17 | 246 |
| 18TH | 18 | 360 |
| 19TH | 19 | 389 |
| 20TH | 20 | 247 |
| 22ND | 22 | 271 |
| 24TH | 24 | 272 |
| 25TH | 25 | 337 |
| | Grand Total | 7,251 |

Figure 7: Police Districts (Spielman, 2019)

Figure 6: Map of Burglaries Overlaid on Districts

As can be seen from the figure, the burglaries are relatively evenly distributed, although there is a significantly smaller concentration of burglaries in the South-Eastern and North-Western parts of Chicago (districts 4 and 16 respectively), although this is mostly due to the fact that these are the outskirts of the city and both districts largely cover non-residential areas (an airport and a city park). Surprisingly, there seems to be a lower concentration in the central part of Chicago. Besides these areas, however, the burglaries are spread out evenly. Let us also generate a table to accompany these visualizations, as there is a possibility a high concentration of crimes in one area may mask the total number of crimes that is happening within a district. This is generated with a standard SELECT query (see line 15) of district and the number of arrests, although we limit the observations to unsolved burglaries in 2021 with the WHERE command and the strftime command (this lets us separate the year in the Date variable). The observations are grouped by district and ordered in the number of unsolved burglaries.

We can see from Table 6 that our initial insights were not entirely correct - district 4, for example, has a high number of burglaries, but the majority of them happen in a small part of district 4. Other areas that have an overlap of high unsolved burglaries and small area are for example the 5th district and the 6th district).

After running all the queries, the CREATE TRIGGER command was used to create a trigger that would notes down into a separate log table whether there was a query added or deleted.

Table 6: Unsolved Burglaries by District

| District | Unsolved burglaries |
|---|---|
| 8 | 312 |
| 6 | 276 |
| 5 | 271 |
| 19 | 227 |
| 4 | 222 |
| 25 | 206 |
| 3 | 197 |
| 14 | 193 |
| 12 | 192 |
| 7 | 189 |
| 9 | 174 |
| 17 | 165 |
| 11 | 155 |
| 16 | 146 |
| 24 | 136 |
| 2 | 135 |
| 22 | 134 |
| 15 | 132 |
| 18 | 116 |
| 10 | 111 |
| 20 | 88 |
| 1 | 65 |

## Conclusion

Lastly, we can summarize the results. In terms of the crimes against police officers, there was a general downwards trend in the last two years, although this may be attributed to a lower number of crimes in general, which could have been caused by the COVID-19 pandemic. This, however, does not entirely explain the higher level of these crimes in 2020 and the several spikes that occurred during the year - the May spike in particular would be in line with the case of George Floyd (Altman, 2020). It may still be worthwhile for the Chicago Police Department to investigate what the population's stance towards the police further.

When it comes to the districts with the largest number of index crimes, these are districts 18, 1, 12, 8, 6, 19 and 11. The Chicago Police Department should inspect these districts further to uncover why there is such a high rate of serious crimes in the district and possibly think about re-assigning more seasoned officers into the area.

Regarding arrest rates, it is clear that the lowest performing districts would be 17, 20, 24, 14, 2 and 16. The department may want to run a more detailed performance review with these districts and see whether an intervention in the form of additional training or lay-offs is necessary.

Finally, when inspecting the areas where there is a high number of burglaries, there are a number of districts, which have a high concentration in a smaller area like district 4, 5 and 6. The department can cross-reference the created map with the table to identify similar areas, which can be further investigated and the patrols around those areas could potentially be increased.

# References

Altman, A. (2020, Jun). Why the killing of george floyd sparked an american uprising. *Time*.

Covington, T. (2022, Jul). *Burglary statistics amp; research year (from bsj and fbi) — the zebra.* Retrieved from https://www.thezebra.com/resources/research/burglary-statistics/

CPD. (2021, Dec). *Chicago police department - illinois uniform crime reporting (iucr) codes: City of chicago: Data portal.* Chicago Police Department. Retrieved from https://data.cityofchicago.org/Public-Safety/Chicago-Police-Department-Illinois-Uniform-Crime-R/c7ck-438e

Maguire, M. (1980). The impact of burglary upon victims. *Brit. J. Criminology*, *20*, 261.

Sparrow, M. K., et al. (2015). Measuring performance in a modern police organization. *Psychosociological Issues in Human Resource Management*, *3*(2), 17–52.

Spielman, F. (2019, Apr). *Alderman demands reopening of wood street police district.* Chicago Sun Times. Retrieved from https://chicago.suntimes.com/2017/12/5/18391175/alderman-demands-reopening-of-wood-street-police-district

Watson, R. T. (2022). (Open ed.). Retrieved from https://www.richardtwatson.com/open/Reader/_book/data-management-databases-and-analytics.html

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1–23. Retrieved from https://www.jstatsoft.org/index.php/jss/article/view/v059i10 doi: 10.18637/jss.v059.i10