1. Research problem specification and data description

Ensuring public safety is a challenging task, which requires excellent time management and well-thought resource allocation, as well as adequate planning. Given the history of criminal activity in Chicago, this task is even more difficult, with the city holding the 7th position in terms of murder rate in 2020 (*T.Monkovic, J.Ascher, New York Times, 2022*) in the US.

As the usage of advanced analytics plays an play a significant role in many sectors, it is assumed that it can also empower public safety. The goal of this report is to enhance the power of analytics in crime prevention and law enforcement, by providing insights tailored to the needs of the Chicago Police Department.

In order to identify the crucial areas needed for planning the operations of the PD, an explanatory analysis of the data set shared by the Chicago PD was conducted. The data set consists of more than 730 thousand crime records with data on various areas of interest: crime trends over time, geographic location, police units and types of location. On the basis of data analysis, a research problem consisting of two principles was defined:

- The first principle is the crime occurrence over time (regarding both: yearly and monthly perspective, and the average hourly distribution) and specifying the most common types of crimes in Chicago and location types in which they occur.
- The second one is to identify the effectiveness of law enforcement by measuring the ratio of arrest by districts and beats.

To use data provided in a efficient way, that will create added value for the Chicago Police Department in its areas of interest, three groups of research questions were defined:

1. How did the number of crimes changed over the years and over the months? Which types of crimes were the most common and how they have been changing over the years 2017 - 2021?

2. Which districts had the highest number of crimes over the years 2017 - 2021? In which beats and districts the relative numbers of arrests compared to the numbers of crimes (arrest ratio) committed was the highest, and in which - the lowest?

3. How has the number of crimes occurring changed given time of the day on average? In which types of locations do they happen most often? Which type of location is the most common for crimes committed in a given time of the day?

Investigating this questions will help to identify if the focus areas of the PD are aligned with the crime patterns in which way the Chicago PD can effectively allocate resources to successfully prevent and counteract crime. Arguably, these questions not represent the whole substance of problem – one might say that f.ex. the ratio of arrests ratio is not a good indicator of the effectiveness of the police force, because in some cases it's very rare to catch the criminal (e.g. arson), however with a given data set it cannot be represented in this project.

Providing responses for those questions is crucial for understanding the trends related to crime occurrence and detection and therefore – for making law enforcement more effective and- the city of Chicago safer.

2. Necessary requirements for the database

Research problem defined in the section 1 leads to defining key entities of importance of the relationship database developed. To address the questions asked in that section, the attributes have to represent different characteristics of a crime, the time when the crime was committed and location in which it occurred. This division leads to identification of three entities of importance for every crime in the designed database and assignment of variables to them (names of variables are presented in the parentheses):

- **Crime** *(Crime ID, case number, location description, longitude, longitude, location, arrest, date)*
- **Neighbourhood** *(district, beat, block)*
- **Type of crime** *(IUCR code, primary type, description)*

Variables needed for the analysis

Although most of the data is important to the Police Department in general, only some variables can provide valuable insight for this project. According to the Data Minimization good practices, redundant data should be removed to make the databases easier and more efficient. Attributes that are needed to answer the stationed questions by creating queries to are:

For entity <u>Crime:</u>

- ***CrimeID*** – it is an unique identifier for every record and will play a substantial role, as the Primary Key of the entity Crime. Records were given Crime IDs in the chronological order starting from beginning of 2017, which enhances the transparency of the projects.
- ***arrest*** – this variable is a dummy variable that indicates if the arrest was made or not. Including this variable is needed to assess the arrest ratio and answer the question 2.
- ***date*** – time aspect has to be included in this analysis according to questions 1 and 3. Variable date gives information on both: the date when the crime was committed and the time during the day.
- ***locationdescription*** – this variable accounts for type of the place when crime occurred. It is needed for investigating question 3.

For entity <u>Neighbourhood:</u>

- ***beat*** - this variable representing the number of the beat will play a role of the Primary Key of the entity Neighbourhood, as it can connect districts with the unique CrimeID. It is needed to answer the question 2. Beats are smaller area units than districts and have from 3 to 4 digits, where the first two digits are the same as the number of the district.
- ***district*** – this variable represents the number of district and it's needed for answering question 2.

For entity Type_of_crime:

- ***iucr*** - this variable aligned with the IUCR code from the Chicago PD methodology and it's needed to link CrimeID with characteristics of crime – it will play the role of Primary Key of the entity Type_of_crime. It will be used in answering question 1 regarding the type of the crime.
- ***primarytype*** – this variable states the most important characteristics of a crime and will be used to answer the question 1.
- ***description*** – secondary description that will be used to describe the type of the crime on the similar level as iucr code, to answer the question 1.

Excluded variables

- ***Casenumber -*** plays the similar role as the *CrimeID* as a unique identifier of a crime, however is not numerical and there are no chronological patterns. Due to that issues and redundancy with *CrimeID it will be removed from the data set.*
- ***Longitude and latitude*** – those two variables could be useful in presenting crime trends in different places of the city, however visualizing the point values of each crime is not the goal of none of the questions (high granularity is not useful as they are stated on high level) and it is not feasible to present crime locations with such granularity in this report.
- ***Location*** - presents the best estimate of the place where the crime occurred and can be useful in general, however it is redundant with the longitude and longitude variables and - like those variables - it is not in scope of this report.
- ***Block*** – block presents many data on location – street, zipcode etc., however the granularity is too high, which means that its usage will not be useful to answer any of the stated questions.

Normalization issues

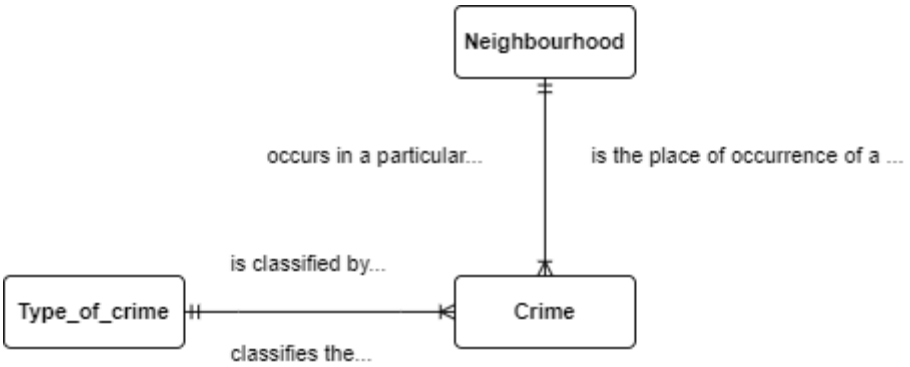
The conceptual model of the database will be describe in the next section of the report, however some of the normalization issues can be foreseen. Normalization in the process of developing the database will done to the extent allowing to provide possibly highest data consistency and to allow correct and efficient operation of the database, however it is important to mentioned if the tables are aligned to the normal forms.

- ***Crime*** conforms the 1st normal form as all domains of the attributes contain only atomic values and the value of each attribute only contains a single value from that domain. It may be unclear in case of *date*, because it contains both date and time values, however in the database it will be used as a timestamp, which is an atomic value. As there is no composite key and partial dependencies the 2nd is also fulfilled. It can be also concluded that the 3rd normal form, as there are no transitive dependencies.
- ***Neighbourhood*** as well conforms the 1st normal form as all domains of the attributes contain only atomic values and the value of each attribute only contains a single value from that domain. As there is no composite key and partial dependencies the 2nd is also fulfilled. It can be also concluded that the 3rd normal form, as there are no transitive dependencies.

- ***Type_of_Crime*** also conforms the 1st normal form as all domains of the attributes contain only atomic values and the value of each attribute only contains a single value from that domain. As there is no composite key and partial dependencies the 2nd is also fulfilled. It can be also concluded that the 3rd normal form, as there are no transitive dependencies.

3. Data Models designed for efficient operation of the database

Diagram 3.1 - Conceptual Entity Relationship Diagram

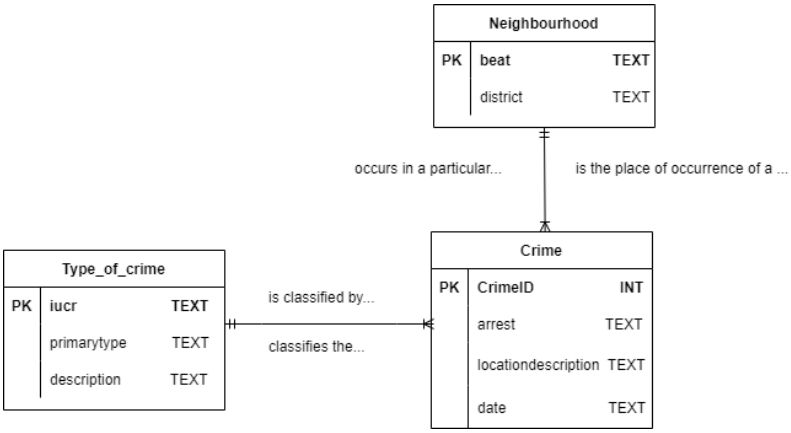


Conceptual data model was developed to represent the relationships between entities of importance specified in the section 2. Cardinalities chosen represent the characteristics of the relationship – all of which are one-to many relationships in this case. Those relationships are described below:

Crime – Type of Crime many-to-1 relationship
When the *Crime* is committed, it is always **classified by** a type with set of characteristics. It is assumed that no crime can have no characteristics. Based on this, both the lower and upper cardinalities from site of the *Type_of_*crime are equal to 1. One set of crime characteristics has to **classify** at least one crime (lower cardinality is 1), but can characterize more (upper cardinality is many).

Crime – Neighbourhood many-to-1 relationship
When the *Crime* is committed, it always **occurs in a particular** (and only one) *Neighbourhood*, which is represented by a set of a characteristics. It is assumed that no crime can occur without a location. Based on this, both the lower and upper cardinalities from site of the *Type_of_*crime are equal to 1. One **Neighbourhood** has to be a place of occurrence of at least one crime (lower cardinality is 1), but more crimes occur there (upper cardinality is many).

Diagram 3.2 - Logical Entity Relationship Diagram



Crime

CrimeID has been chosen as an the Primary Key of the entity Crime, because it is an unique identifier for every record in the database, it has been assigned in the chronological order and is in the form of integer, which enhances transparency.

Meaning of Crime attributes and data types set:
CrimeID – primary key, unique identifier for every crime, numerical form - INTEGER
Arrest – indicator if the arrest was made, TRUE or FALSE values – TEXT for the purposes of querying
Locationdescription – describes type of the place when crime occurred, text string, - TEXT
Date – timestamp representing time and date when the crime occurred – TEXT for the purposes of using datetime functions

Neighbourhood

The attribute *beat* has been chosen has been chosen as an the Primary Key of the entity Crime, because there is unique set of beats for each district (structure of beat number described in the section 2.)

Meaning of Neighbourhood attributes and data types set:
Beat – number of beat unit, primary key of the table, numerical form, TEXT for processing purposes (to avoid errors)
District – number of the district, numerical form, TEXT for processing purposes (to avoid errors)

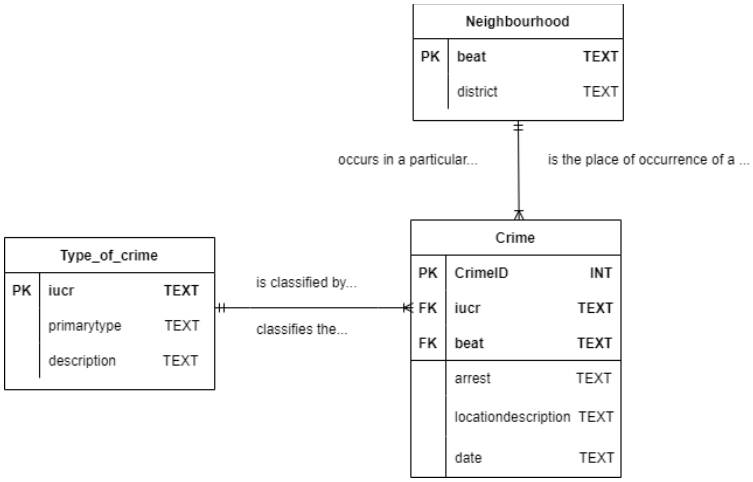
Type_of_Crime

The attribute *iucr has* been chosen as an the Primary Key of the entity Type_of_crime because there is unique set of iucr codes for each primarytype, and each iucr code has got one, or more secondary descriptions.

Meaning of Type_of_crime attributes and data types set:
iucr – IUCR code of crime used by Chciago PD, primary key of the table, form of strings with numbers and text, TEXT
primarytype – main description of the crime typed used by Chicago PD, text string form, TEXT

description – secondary description of the crime typed used by Chicago PD (similar level of granularity as iucr, however many descriptions may be assigned to one iucr, text string form, TEXT

Diagram 3.3 – Physical Entity Relationship Diagram



By setting *beat* as a foreign key in *Crime* a 1:M relationship between Neighbourhood and Crime can be created, which is supposed to allow creation of join queries showing crime numbers over time in given districts and beats.

The second foreign key in *Crime* is *iucr,* which allows to create a 1:M relationship between Type_of_crime and Crime. This is supposed to give a possibility of (join) querying showing crime numbers over time in given districts and beats.

As no normalization issues were in this and previous stages, no normalization issues occurred. The database will be implemented accordingly in the section 5.

4. Summary of Data quality checks performed as a part of Data Processing

The original dataset consisted of 730,900 records of 14 variables. To stay in line with data minimization practices mentioned earlier, the variables that will not be used in the database (*casenumber, block, location, latitude, longtitude*) part were removed from the dataset.

The data quality checks were performed by using RStudio and R language. As well all of the procedures mentioned in this section were executed in RStudio, using R packages *janitor*, *tidyr* and *tidyverse*. The whole process has been divided into sub-sections, that describe data quality dimensions (as defined in class, based on Eryuek et al 2021) and CESSDA (2020)

Completeness

The first step in the process of performing data quality check was to check the completeness of the data by identifying missing values. The process was started with indicating the empty

spaces and space-values as “NA” values in R. It was done by attaching an appropriate condition to the read.csv function in R. In total 3482 missing values were found, of which 3430 were found for location description, in case of which, it was not considered a problem in terms of data completeness, given the scale of the dataset. Count of N/A values after the removal is presented below - 19 rows containing 62 NA’s were removed.

Accuracy
The dataset was investigated as well for duplicate rows, and 148 of these rows were found. Variable *date* was transformed into the universal timestamp, which can be processed by SQLite in text formats. All duplicated rows have been deleted, as they would not allow CrimeID to be a primary key for the developed relationship database. 7 types of non-existing IUCR codes (according to the Chicago PD website) were found : “585”, “1581", “3961", “5073", “5093", “5113" ,“5114". All rows containing this codes were removed.

Consistency
In order to investigate data consistency, various aspects were checked. 78 mismatches between beat and district were found and removed. For district 31 (which is a very small area), all of the beats were mismatched and they were removed, as it supposedly was a human mistake.

According to the list on the Chicago PD Internet site, each iucr has exactly one description. In the dataset, the number of different versions of unique descriptions was higher by 123 than the number of iucr codes. Redundant, similar and misspelled descriptions were removed by matching data with the list published on the Chicago PD site. In the cleaned data set, 336 unique IUCR codes and 312 unique descriptions can be found.

No significant violations of other relevant areas of data quality dimensions were found.
In total, 682 rows were deleted, which is less than 0.01% of total observation, which is very low and should not reduce the value of the data set.
Cleaned dataset, containing 730 218 records of 9 variables was imported to DB Browser.

5. Implementation of the database in DB Browser

Cleaned dataset was imported to DB Browser as table “CRIMES”. Tables Crime and Type of Neighbourhood were created by using CREATE TABLE command, and primary keys and data types set as described in the section 3. Table Crimes was also created using this command.

Appropriate constraints have been set for some variables in the data set, in cases when it was needed:
Type of Crime - iucr TEXT PRIMARY KEY NOT NULL UNIQUE constraints were set for iucr. With iucr being the Primary Key NOT NULL and UNIQUE are needed by default, however they were set just set to ensure that data complies to it.
Neighbourhood – beat TEXT PRIMARY KEY NOT NULL UNIQUE constraints were set for beat with the same motivation as above.

Crime – Crime ID INTEGER PRIMARY KEY NOT NULL UNIQUE constraints were set for CrimeID, the same motivation as above.
In table Crime, iucr and beat were set as foreign keys by using command FOREIGN KEY (..) REFERENCES.
For all other attributes from other tables, except *locationdescription* which has (and can have) no value, NOT NULL constraints were set to ensure that no empty values are present.

After that, the two tables without foreign keys were populated by using command INSERT INTO and SELECT, and, in the end table Crime was populated.

6. Insights and conclusions from data analysis conducted on data queried from the database

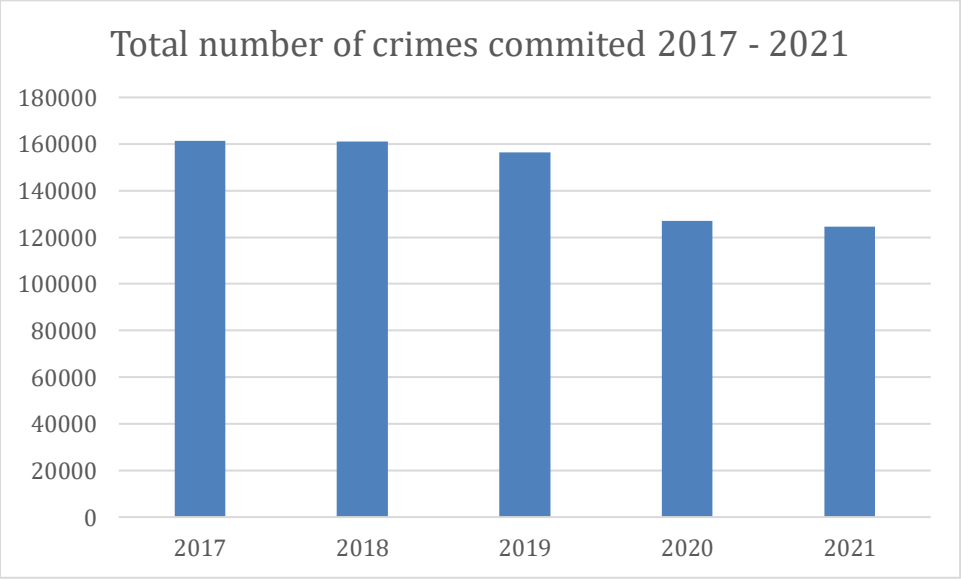
Operational commands

Trigger
A trigger that prevents updates on columns *CrimeID* in the table *Crime* was implemented to prevent changes of CrimeID, which is the primary key of the table and is assigned to crime records chronologically. Change of the values might cause serious problems of functioning of the database.

Index
An index that represents Top 10 crimes was implemented for querying for the section dedicated to the question 1. By using index, SQLite uses less memory it to find the top 10 Crime types, which makes the database more efficient.

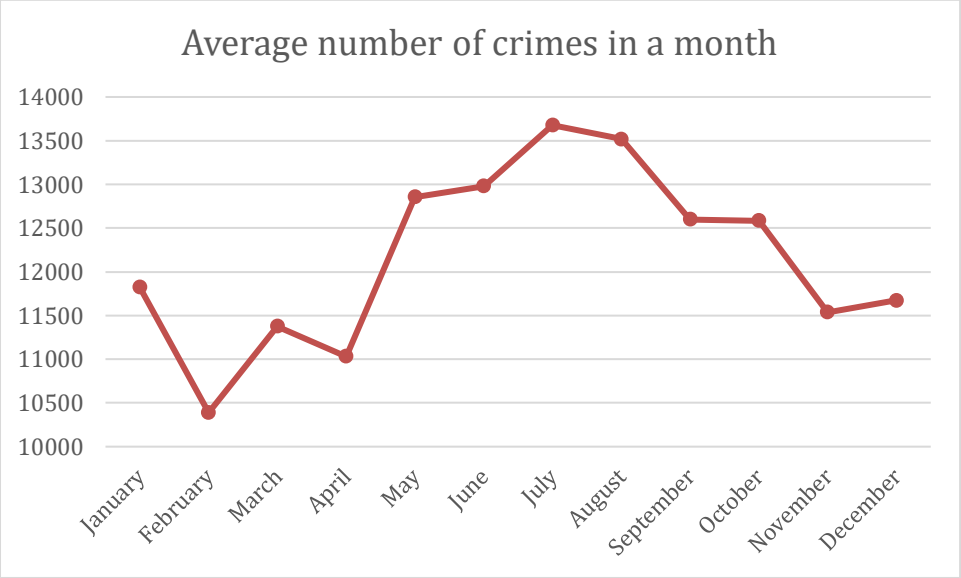
Question 1

Plot 6.1 Crimes committed in Chicago in years 2017-2021, by year:



By using windows functions SELECT with OVER ORDER BY to get information from table Crime about the changes of number of crimes in Chicago was found. Here and among most of the queries, function COUNT is used to count the number of crimes, and function strftime is used to manipulate the date and time in the form of timestamp. Function LAG was used to calculate the differences between the years to better understand the data. Plot 6.1 shows that the number of crimes in the city has been declining over the years, which is a positive tendence. Function ROUND was used to round the decimals of the results.

Plot 6.2 Crimes committed in Chicago in years 2017-2021, by monthly average:



By using similar window function as in the example before, information from table Crime about occurrence of crime by monthly average was found. Plot 6.2 shows that there is a seasonal pattern – most of the crimes occurred in the summer months and the values were up to 30% higher than in winter months.

Table 6.1 10 most common crimes committed in years 2017-2021, grouped by *primarytype*:

primarytype	Crimes commited	Yearly average	% of all crimes
THEFT	164,072	32,814	22.5%
BATTERY	138,142	27,628	18.9%
CRIMINAL DAMAGE	80,116	16,023	11.0%
ASSAULT	59,267	11,853	8.1%
DECEPTIVE PRACTICE	56,330	11,266	7.7%
OTHER OFFENSE	46,880	9,376	6.4%
NARCOTICS	31,548	6,310	4.3%
MOTOR VEHICLE THEFT	30,662	6,132	4.2%
BURGLARY	29,707	5,941	4.1%
ROBBERY	27,347	5,469	3.7%

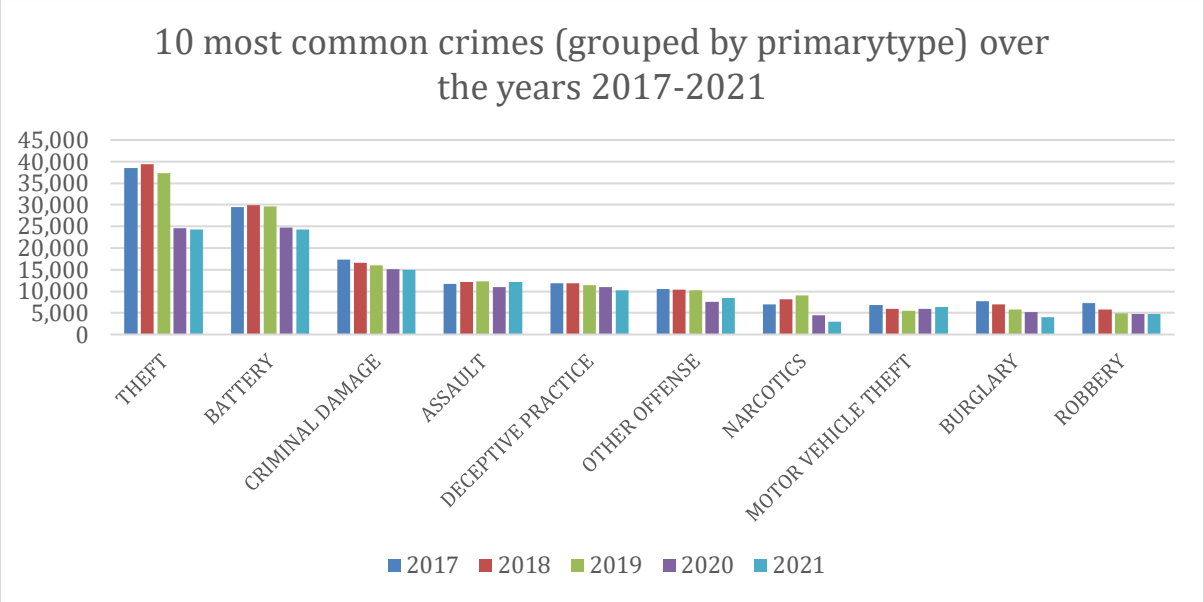
Query with SELECT DISTINCT and JOIN of tables Crime and Type_of_Crime were used to identify top 10 crimes by *primarytype* over the analyzed period. LIMIT 10 was used to limit the results to top 10 records. Records were grouped (GROUP BY) primarytype (ORDER BY) by year. Table 6.1 presents that in this period, many of the most common crimes are the serious and violent crimes. 50% of all Crimes in Chicago in this period have been Theft, Battery and Criminal damage.

Table 6.2 10 most common crimes committed in years 2017-2021, grouped by *iucr* code:

iucr	primarytype	description	Total crimes commited	% of all crimes
486	BATTERY	DOMESTIC BATTERY SIMPLE	67,125	9.2%
820	THEFT	$500 AND UNDER	63,631	8.7%
460	BATTERY	SIMPLE	43,109	5.9%
810	THEFT	OVER $500	41,344	5.7%
1310	CRIMINAL DAMAGE	TO PROPERTY	38,905	5.3%
560	ASSAULT	SIMPLE	38,542	5.3%
1320	CRIMINAL DAMAGE	TO VEHICLE	36,793	5.0%
910	MOTOR VEHICLE THEFT	AUTOMOBILE	27,173	3.7%
860	THEFT	RETAIL THEFT	26,402	3.6%
890	THEFT	FROM BUILDING	24,704	3.4%

Similar query as above with SELECT DISTINCT and JOIN of tables Crime and Type_of_Crime was used to provide a more granular list of 10 most common crimes committed in years 2017-2021 grouped by (GROUP BY) iucr code and ordered (ORDER BY) by year. The insights are similar to the previous table – theft and battery are the most common. Interesting observation is that the more common occurrence on thefts is the one under $500, which has almost a half larger occurrence.

Plot 6.3 Changes in the amount of 10 most common crimes (grouped by *primarytype*) over the years 2017-2021:

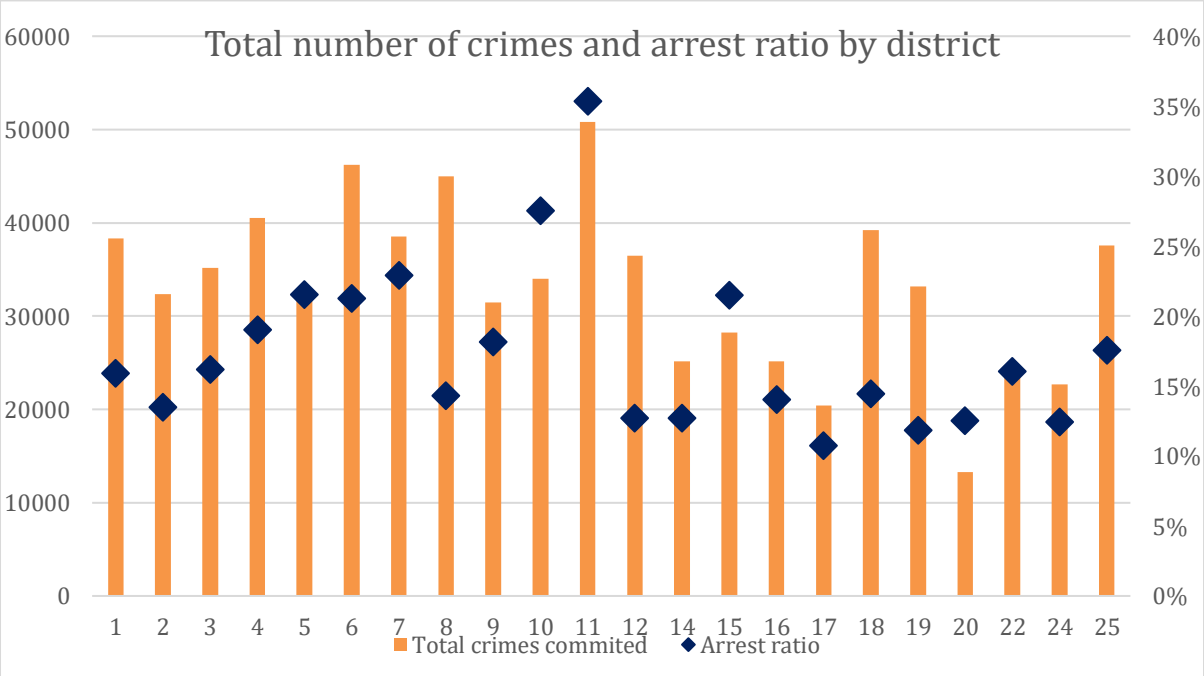


Similar query (joint of Crime and Type_of_Crime) that was used to define the Top 10 crimes by primarytype (Table 6.1) was now used for getting the values of those crimes over the years. An index that represents Top 10 crimes was implemented for querying to make it more efficient.

In conclusion, it can be observed that the number of thefts has dropped between 2019 and 2020 by more than 30%. The number of batteries has dropped by nearly 20% between those years, and the other types of crimes also dropped by a similar level, except assault, which has stayed around the same levels.

Question 2

Plot 6.3 Total number of crimes and arrest ratio by district



Query using SELECT and LEFT JOIN of tables Crime and Neighbourhood was used to provide information about the number of crimes per district and the arrest ratio (calculated by using function AVG) per district. The results were grouped by districts and ordered by arrest ratio. On the plot 6.3 it can be seen that districts 11 and 10 have the highest arrest ratio, and the district 11 has got also the highest number of crimes committed on its premises. It can be also observed that districts with lower numbers of arrests have also lower arrests ratio.

Table 6.3 10 beats with the highest arrest ratio

beat	Crimes_commited	% of all crimes committed	Arrest ratio
1653	550	0.08%	0.48%
1112	5,611	0.77%	0.47%
1652	189	0.03%	0.47%
1134	3,998	0.55%	0.45%
1131	3,435	0.47%	0.44%
1121	4,372	0.60%	0.43%
1132	5,049	0.69%	0.41%
1115	2,680	0.37%	0.39%
1011	5,088	0.70%	0.37%
1122	4,252	0.58%	0.37%

Queries with functions SELECT DISTINCT joining tables Crime and Neighbourhood (JOIN) were used to create tables 6.3 and 6.4. The results were grouped by (GROUP BY) beats and ordered by arrest ratio. LIMIT 10 was used to limit the results to top 10 records.
It can be seen that mostly the beats from 16th and 11th districts are the least effective in terms of arrests ratio.

Table 6.4

beat	Crimes_commited	% of all crimes commited	Arrest ratio
235	1,356	0.19%	0.04%
1214	3,253	0.45%	0.05%
1932	1,631	0.22%	0.06%
233	1,638	0.22%	0.06%
1814	2,164	0.30%	0.06%
1813	1,143	0.16%	0.06%
1935	2,595	0.36%	0.07%
1811	1,786	0.24%	0.07%
1215	2,304	0.32%	0.07%
1912	1,740	0.24%	0.07%

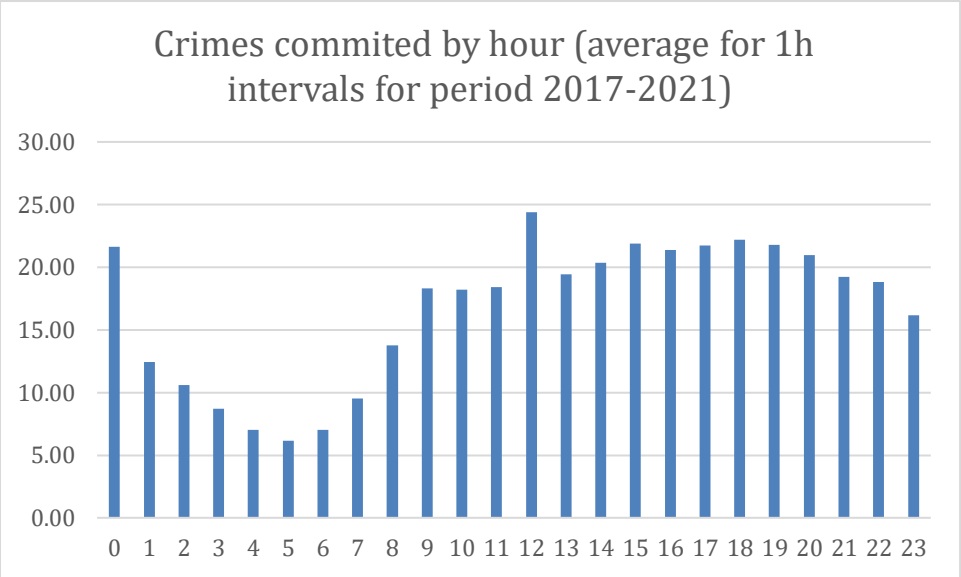
It can be seen that mostly the beats from 12th and 19th and 18th district are the least effective in terms of arrests ratio.

Question 3

View

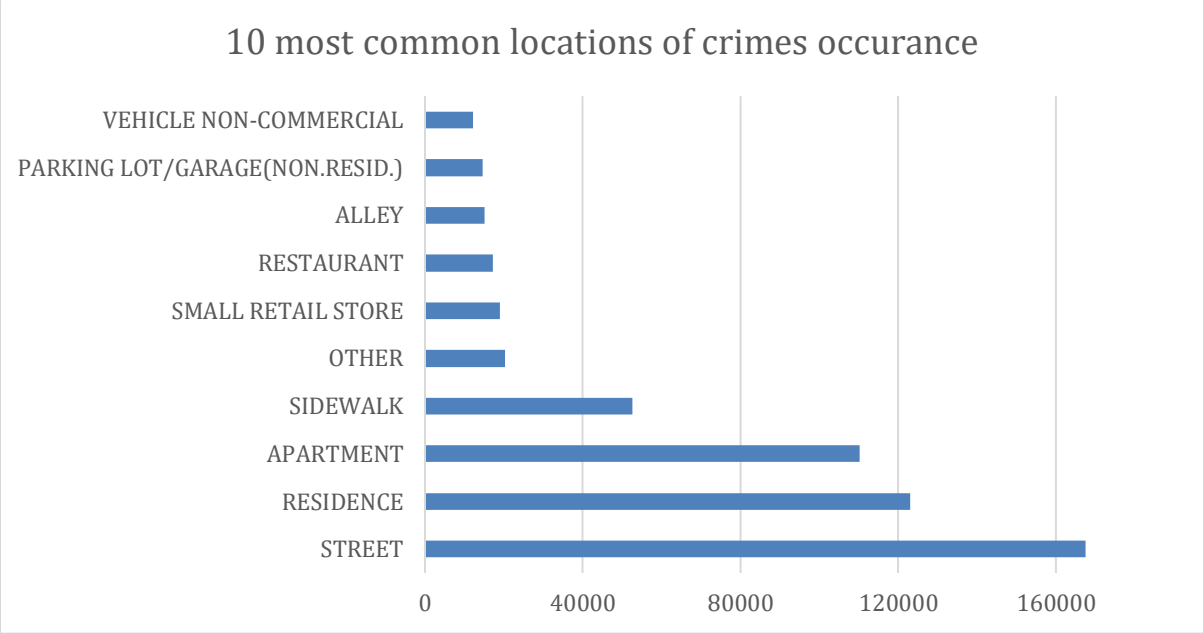
A view that presents the locations in which crimes happen most often in the given hourly interval was implemented for answering the question 3, in order to enhance the database efficiency and making the syntax less complicated and more transparent. Window function SELECT FROM was used to create a ranking (RANK ()) and choose (AS tempor) the most common locationdescription of crimes happening in given hour interval.

Plot 6.4 presents Crimes committed by hour (average for 1h intervals for period 2017-2021, the hour of the beginning of interval presented on x-axis)



Query SELECT FROM was used to query the data from table Crime and present the distribution of the crimes committed by the hour of the day, which was then used to create plot 6.4. In the plot in can be seen, that most of the crimes happened in the afternoon hours and evening hours, however the highest number of crimes in 1 hour interval was observed between 12 AM and 1 PM.

Plot 6.5 presents the 10 most common locations of crimes occurrence



Query selecting data from Crime (SELECT) was used to get the information of 10 most common locations of crimes occurrence. Results were grouped (GROUP BY) locationdescription and ordered by number of crimes, descending COUNT(CrimeID), DESC On the plot it can be observed that the most common locationtypes are streets, residences and apartments, in which nearly half of the crimes occurred.

Table 6.5 presents the most common location type of crimes committed by hour (average for 1h intervals for period 2017-2021)

locationdescription	hour	% of all crimes committed within interval in this location type
RESIDENCE	0:00 - 0:59	22.63%
STREET	1:00 - 1:59	27.86%
STREET	2:00 - 2:59	26.79%
STREET	3:00 - 3:59	26.36%
STREET	4:00 - 4:59	25.52%
STREET	5:00 - 5:59	25.85%
STREET	6:00 - 6:59	23.41%
STREET	7:00 - 7:59	22.25%
RESIDENCE	8:00 - 8:59	20.93%
RESIDENCE	9:00 - 9:59	24.60%
RESIDENCE	10:00 - 10:59	19.09%
STREET	11:00 - 11:59	17.61%
RESIDENCE	12:00 - 12:59	21.72%
STREET	13:00 - 13:59	16.92%
STREET	14:00 - 14:59	17.80%
STREET	15:00 - 15:59	19.14%
STREET	16:00 - 16:59	20.89%
STREET	17:00 - 17:59	22.86%
STREET	18:00 - 18:59	26.04%
STREET	19:00 -19:59	28.16%
STREET	20:00 - 20:59	29.81%
STREET	21:00 - 21:59	30.82%
STREET	22:00 - 22:59	32.20%
STREET	23:00 - 23:59	30.37%

A simple function SELECT FROM was used to obtain data from the View containing window function described above. Table 6.5 presents that only 2 types of crimes dominate in the hour average intervals and most of the crimes that happened during the night and afternoon happened on the street, and crimes happened most commonly in the residences at the afternoon.

Short conclusion

As a result of the analyses conducted three questions defined in the section 1 of the report were answered. Summing up the number of crimes has been decreasing in the period 2017-2021, and on average was the highest in the summer months. Serious and violent crime types constitute the majority of the crimes recorded in Chicago. Districts 11 and 10 have the highest arrest ratio, and the district 11 has got also the highest number of crimes committed on its premises. Beats from 12th and 19th and 18th districts are the least effective in terms of arrests ratio. Most of the crimes occurred in the afternoon and the evening, with peak at between 12 AM and 1 PM. Crimes at afternoon and night occur mostly on the streets, and during the mornigs – in residences.

Hopefully, provided insights can help plan the operations of Chicago PD more effectively and make the city safer, protecting property and human lives.

7. References:

1. (T.Monkovic, J.Ascher, New York Times, 2022)
   https://www.nytimes.com/2021/06/16/upshot/murder-crime-trends-chicago.html
   Access 02.10.2022
2. FAIR Principles
   https://www.go-fair.org/fair-principles/
   Access 02.10.2022

3. Data Quality Standards https://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf
   Access 02.10.2022