### **A Chicago Crisis: An Analysis of Crimes in Chicago 2012-2016**

#### Team: Ashourrina Henshaw, Noreen Moeed, Meenu Ravi, Austin Villegas, Sara Wang

#### **Executive Summary**

There is a fear that crime has been increasing rapidly in various cities including Chicago. Therefore, there is a strong need to identify patterns in crimes and analyze different factors involved. Specifically, the Chicago Police department has been tracking crimes in the city since the turn of the century. Looking at data from 2000 to present day, it is evident that crime runs rampant in the streets of Chicago. The Chicago Police Department has collected a significant amount of data in regard to location of the crime, whether an arrest was made, crime categories and more. This project will help analyze trends in criminal activity in the city of Chicago and hope to make recommendations to the Police Department to understand, and in turn curb, crime rates. Our goal is to present this preliminary analysis of historical data to the Chicago Police Department and to provide them with suggested next steps in their goal of reducing crime.

#### **Business Use Case**

The Chicago Police Department would like our team to analyze the data between the years 2012 and 2016. Their goal is to identify historical data trends to effectively and equally police the varying regions of Chicago. The department wants to look at annual data and data throughout the year to evaluate if there is a time component to consider when identifying which departments may need additional resources allocated and where cutbacks can be made to offset this. Additionally, The Chicago Police Department would like to know in which years there were overall decreases in crime or in crime categories to understand what factors may have contributed to such decreases.

#### **Data Sources**

The data sources are retrieved from Kaggle and Chicago Data Portal. Additional research may be conducted online regarding the Chicago Police Department or Federal Crime Information, as well as any additional information about Chicago that can be used for analysis.

See [References](#_heading=h.2xcytpi) for source links.

**Tools Used**

Table

Description automatically generated

#### **Database Model Considerations**

We decided to choose a relational model to represent our data. Relational models aim to model data regarding transactional systems.

##### **Purpose**

This crime data shows a bigger picture of the crimes in Chicago that happened over time and is stored as records and is collected from the real world, in Chicago as the crimes occur. This resembles a transactional model.

##### **Data Format**

Structured publicly available data recorded as each crime occurs.

##### **Requirements and Users**

Only an end user is required to input information regarding crime information. As years progress, the data will continue to grow. However, since only a specific set amount of data is being looked at: crimes in the years 2012-2016, the amount of data is fixed.

**Data redundancy**

There is minimal data redundancy because the data was normalized to 3NF form when we created the EER diagram. There are many tables within the database that each have a particular purpose and that are related to one or more other tables using primary and foreign keys to make certain that there are unique records within each table. Also, the fields are unique to each table and not replicated in multiple tables

##### **Efficiency and Maintenance**

A relational model is optimized for fast read and fast write so that the ETL processing is efficient. Also, a dimensional model would require ongoing maintenance to add new data sources and maintain any changes that may occur in existing data sources. Instead, a relational model is a better fit for this crime data because it stores current data: precinct's current location, etc. and more data sources can be easily added by creating a new table and adding a relation and relations can be modified easily too.

**ETL Flow**

Graphical user interface, application

Description automatically generated

**Extract:** Assembled data from multiple sources including: The City of Chicago Data Portal, Chicagopolice.org, and Kaggle.com. Downloaded the data into a CSV format. If the file was not available, we manually copied the data from the online data source into a CSV file.

**Transform:** Read the csv file using python, used Pandas library for data cleansing.

Dropped columns(fields) to match the EER diagram, checked the data type for each column to check for anomalies (which were present), altered the data types to match the data types in MySQL. Data was normalized into 3NF during the EER diagram creation. We added extra keys and columns and segregated certain columns so that the model would be normalized.

During the data cleansing process, we checked for null values using the pandas library functions and replaced them with appropriate values so we wouldn’t have issues when loading into the database. We checked for anomalies in the data using the min() and max() functions. We transformed the crime ID by removing the “C” in front of the numbers to match the database type integer. We checked the data type for each column to ensure it matched the schema that had been created for proper loading. We removed the time from the date column and re-formatted to yyyy-mm-dd so the column could properly load into the database.

**Load:** Inserted clean data into MySQL workbench using two different processes.

10 tables are joined to one another with unique keys and relationships. Since some tables have a parent/child relationship, it was very important to load the tables in the proper order and ensure all data was loaded properly to prevent errors. We used the sql script to create the schema. We checked the ChicagoCrime schema was created, including all the tables, indexes, and foreign keys. Imported data into the tables through the sql script and used the data import wizard along with sql scripts to import the data for the 2 largest tables (Location and Case\_Information)

\*Scripts relating to ETL process are located in our attached zip folder

#### **Conceptual Model**Diagram Description automatically generated

**Logical Model**A picture containing text, sign, screenshot, attached

Description automatically generated

#### 

#### 

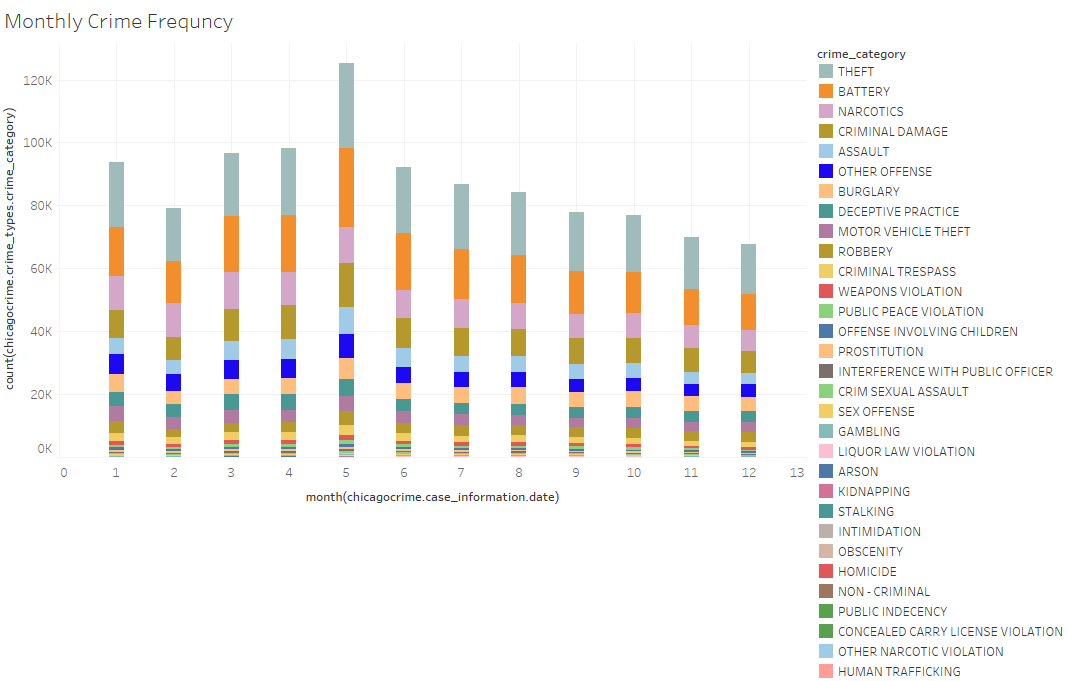
#### **Physical Model**

**Visualizations**

##### **Visual 1**

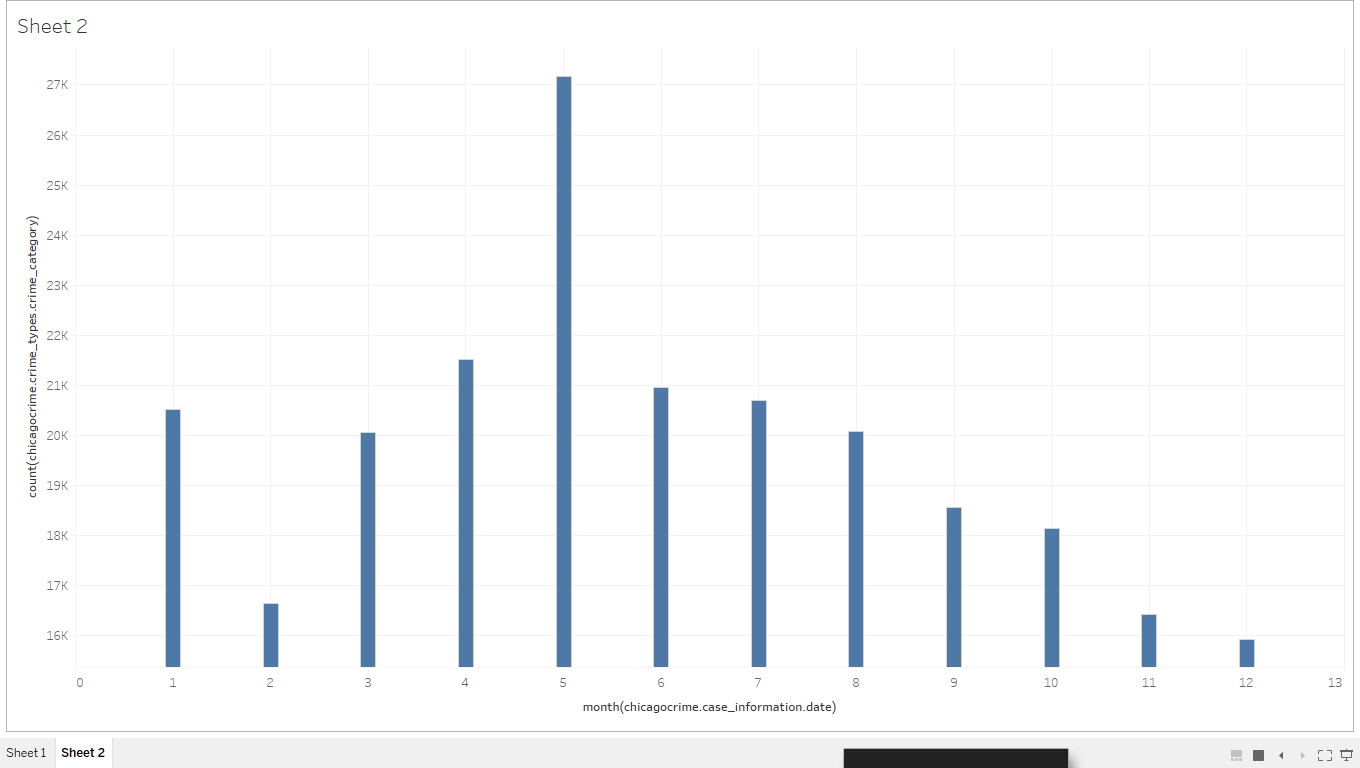
As seen by the visualization, the most cases of crime occurred in the zip code 60612 with around 89,000 cases. The demographic for this zip code is 59% black, 24% white, 10% other race, and 6% Asian.

The second highest number of crimes which is around 71000 occurred in the zip code of 60609 where the demographic is 37% white, 34% other race, 25% black, and 5% Asian.



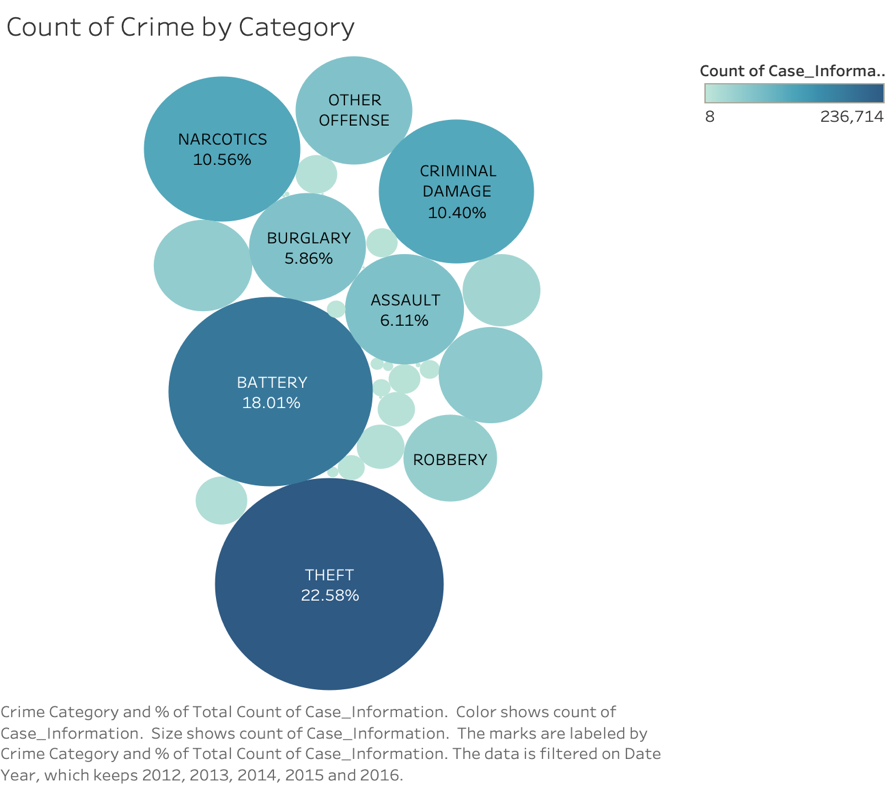
**Visual 2**

When looking at the amount of crime that occurred between 2012 and 2016 by focusing on the months, we see that most crimes occur during the summer months. We further divided the crimes by color to see the division of crime categories. Crimes like stalking and obscenity were not as frequent. However, we see that in all the months, theft was very frequent. As you can see, in all the larger rectangles, the teal blue grey color dominates. Shortly after, Battery, criminal damage, and narcotics follow.

**Visual 3**

Since theft was most frequently, we have shown a bar graph here to clearly visualize during which months theft was most frequent.

We see that theft is greatest during the summer months and gradually decrease as we head into the autumn and winter months. Then, after the holidays, we see a steep increase.



**Visual 4**

This visual represents the crime types that have been reported between 2012-2016. The percentage shown is a percentage of all crimes reported.

Nearly one quarter of all crimes that were reported were crimes of theft.

The interactive visual allows the user to filter by year to see changes although it stayed pretty consistent.

##### 

##### 

##### **Visual 5**

This crime rating heat map illustrates where the worst rated crimes take place most often. The zip codes with the highest averages are: 60619, 60612, and 60607.

##### 

##### 

##### **Visual 6**

This map shows the arrest rate for each zip code in the top 2 and bottom 2 districts by total number of arrests. The 2 red zip codes are 60628 in district 5 and 60612 in district 11.

#### **Insights**

We saw that the maximum number of crimes occurred in southern Chicago than in northern Chicago as we analyzed crime count using zip codes. As demographics in various locations across Chicago using zip codes was studied, we discovered which demographic areas had high crime rates ([visual 1](#_heading=h.35nkun2)).

When looking at frequency of crimes over the years specifically looking at months, we discovered most crimes occur in the summer months of March through May and there is a slight increase after holidays in January ([visual 2](#_heading=h.1ksv4uv)). Among the various crime categories recorded, it is apparent from the visualizations that theft is most frequent, composing nearly a quarter of the crimes committed ([visual 4](#_heading=h.gjdgxs)). Theft has occurred most frequently during May ([visual 2](#_heading=h.1ksv4uv)).

It is evident that crimes with higher ratings (more despicable crimes), take place in area codes 60619, 60612, and 60607 (visual 5). The districts of zip codes 60628 in district 5 and 60612 in district 11 have had the highest arrest rates (visual 6). We would like to look further at these areas and identify possible reasons why these areas are hot spots for such crimes so that we can provide better and more accurate recommendations.

#### **Recommendations**

Based on the analysis and visualizations we have performed in this project, we have decided it will be useful to consider allocating more resources to district 11 and to zip code 60612, where crime is most prevalent. The districts of zip codes 60628 in district 5 and 60612 in district 11 also have had the highest arrest rates. Therefore, allocating more units and resources to these areas will be helpful. We would like for the CPD to consider dedicating more resources to theft and criminal damage and allocating seasonal units during summer months to curb the influxes of crimes during these periods. During Summer, the CPD can spread awareness about having security systems to neighborhoods and local businesses by having educational programs, community outreach and advertisements.

We would recommend the CPD to record and publish more data for analysis so that with more data on a more granular scale, we could identify better patterns and make more accurate predictions and stronger recommendations.

#### **Next Steps**

We would like to look at patterns in crime in Chicago during a wider time period to further analyze other factors’ contributions to crime rate. Specifically, we would look at times where there was economic depression like in 2007-2008, during election years, and during periods of social unrest like 2001 and 2020. By looking at such factors, we hope to use methods like predictive analysis to decide how and when to allocate resources and help the CPD plan ahead. From this analysis, we discovered crime has been gradually decreasing each year in Chicago between 2012 and 2016 and looking at a wider time frame can help us understand reasons for such crime.

We can do statistical analysis, like a T test, to understand the pattern and distribution of crime. On the other hand, we wanted to look beyond reported crime statistics and see how bias can affect the pattern of crime and understand how collider bias affects this analysis.

We have looked at demographics and community areas thus far and, in the future, we would like to look at immigration and emigration rates from areas of Chicago where crime rates are high to see how much of an effect crime rates have on people’s decision to live in Chicago. Data regarding demographics of people involved in which crimes, weapons used, and gender of perpetrators/accessors are also very useful in order to make further observations and predictions as well as social media analytics to predict crimes. This analysis can help encourage CPD to take further efforts to allocate resources and make changes to help reduce crime and keep the streets of Chicago safer.

**Automation Methodology**

Automated data capture uses technologies that allow machines to capture data and then transform it into meaningful insights. As more data regarding crimes is collected, we expect more users to use the data and add records. We also hope to use End To End Testing to validate and simulate real-time settings to identify system dependencies and to ensure that the data integrity is maintained between various systems.

Tools such as Microsoft Azure and GitHub can be used to simultaneously work on code, testing, and version control.

#### **References**

<https://home.chicagopolice.org/statistics-data/data-dashboards/sentiment-dashboard/>

<https://www.sas.com/en_us/insights/data-management/what-is-etl.html>

<https://stackoverflow.com/questions/2578194/what-are-ddl-and-dml>

<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop_duplicates.html>

**Data Source Links**

<https://www.kaggle.com/currie32/crimes-in-chicago> 2012-2017 data ONLY

<https://data.cityofchicago.org/Public-Safety/Police-Stations/z8bn-74gv> Police Stations

<https://guides.lib.uchicago.edu/ChicagoStudies/Data> Community Data Snapshots

<https://data.cityofchicago.org/widgets/htai-wnw4> Chicago Ward Information

<https://data.cityofchicago.org/widgets/c7ck-438e> Illinois Uniform Crime Reporting Codes (IUCR)

<http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html> FBI Codes

Demographics obtained from ArcGIS (ESRI Business Analyst)