Chicago Crime

An Analysis of Chicago Crime Trends

Graham Dominick
Computer Science
University of Colorado Boulder
Boulder, CO USA
Graham.Dominick@colorado.edu

Isaac Hames
Computer Science
University of Colorado Boulder
Boulder, CO USA
Isaac.Hames@colorado.edu

Amanda Killeen
Computer Science
University of Colorado Boulder
Boulder, CO USA
Amanda.Killeen@colorado.edu

Daniel Kingsley
Computer Science
University of Colorado Boulder
Boulder, CO USA
Daniel.Kingsley@colorado.edu

PROBLEM STATEMENT

The United States is currently enjoying a period of all time low crime rates, according to the New York University Law School. The number of crimes per 100,000 people in the US's 30 largest cities has dropped from ~9,300 in 1990 to less than 4,000 in 2016, and violent crime has dropped from ~730 to ~386 instances per 100,000.[5] Despite this, Chicago stands out as a city that has experienced a disturbing rise in crime, particularly violent crime, in recent years, with a rate of over 900 violent crimes per 100,000 people. A 2016 article in Time Magazine showed how Chicago responsible for for almost half the increase in the national murder rate in 2016 (13%).

Chicago's spike in crime has received a lot of publicity, and has made the city somewhat of an example city for those interested in talking about or studying crime rates. In January 2017 President Trump tweeted about Chicago crime and suggested that he may "send in the feds" in response to a 24% increase in killings from 2016.

Crime trends in any city can be difficult to interpret, especially in such a large one like Chicago, and it's unclear whether tougher policing would have more positive effects than other potential strategies, like investing in social services such as education, mental health, and drug rehabilitation. An article published by the Atlantic in 2017 discussed how the causes are essentially unknown, despite lots of research.

MOTIVATION

Our goal for this project is to use data mining techniques on a dataset of Chicago crime information in order to uncover novel trends about crime in the city. We hope to realize trends that could potentially help understand the factors causing crime in Chicago, and provide insight leading to possible solutions.

1 Literature Survey

As we began work on this project we explored various articles, notebooks and repositories to understand the existing work that has been conducted on the topic of Chicago Crime using the Chicago Crime Portal data.

The first existing work we evaluated was from the City of Chicago itself and their Crimes - 2001- to Present Dashboard. This dashboard is a comprehensive resource of 13 visualizations covering the entire data set and has an emphasis on the locality of reported crimes, which is a topic we are planning to explore in our research. Their geographical visualizations however tend to implement a regional visualization technique and there is an

opportunity to explore other geographic visualization methods, such as a heatmap or point map to better express volume of crimes and where the literal hotspots are.

In addition to the City of Chicago's visualizations, we explored Kaggle**,** an online machine learning and data science community, where we found several projects on the use of the Chicago crime dataset that we are utilizing in our analysis. We mainly focused on the work conducted by Djona Fegmen, in his Chicago Crime Data Analysis Jupyter notebook. Fegmen uses Python along with the pandas, seaborn and matplotlib.pyplot libraries to create a number of visualizations on overall counts of primary crime type, arrest volume over time, top 5 crimes and trends over time, and volume and trends in domestic violence arrests.

On Github we found a repository by Tejal Behangle, where she conducted a similar project at the Missouri University of Science and Technology, but her project attempted to predict the location and time of day where crimes and when crimes were likely to be committed. Using both the Chicago crimes dataset and a second dataset of socioeconomic data to explore the impact of socioeconomic factors (such as literacy rate and employment) on the rate of crime occurrence.

Finally, we reviewed an article published on Medium.com by Sadaf Tafazoli. In his article, Tafazoli explains how following preprocessing, he used Spark SQL to explore the data and answer questions. This is an interesting approach and we may consider also utilizing the Spark SQL module, however we'll likely primarily use python for our analysis. This article also sought to answer multiple questions and created visualizations to support their questions, but did not have a cohesive or particularly interesting outcome or story. It was simply a lot of count queries that were later graphed in some manner.

2 **Proposed Work**

Data collection: Data was collected by the Chicago Police Department and the data set was made available by the city of Chicago

Preprocessing:

- Remove unneeded columns and attributes
- Clean string data (ensure uniformity between descriptions, locations, etc.)
- Bin by location
- Filter by crime type/arrest

Process for derived data: We may be able to derive a boolean "Likely gang related" or "Not" value with the following process: One area of obvious interest for this project is gang-related crime. While our dataset provides minimal information on the gang-affiliation of certain crimes, perhaps it could be inferred from the series in which crimes of a type occur in. For example, if a murder in a neighborhood known for gang activity is closely followed by another murder, it may likely be gang-related. We propose to look for neighborhoods or "beats" with serial episodes of violent crimes, to identify areas with potentially higher violent gang activity. We can contrast these information with supplemental information on the geographical distribution of gangs. Information gained from this work could inform law enforcement of areas that need increased response after an instance of a violent crime.

Design: We also propose to create an arrest prediction model for crimes in Chicago, and infer why certain crimes result in an arrest made, while others do not. We anticipate that several of the attributes in our dataset will correlate with arrest, for instance: crime type, neighborhood, date, and whether or not a crime was domestic in nature. By using logistic regression we can

examine these relationships and extrapolate factors that lead to unsolved crimes.
How it is different from previous work: Our work differs from the previous work surveyed in that we are looking to explore a relationship between violent crime and gang activity. The closest work we surveyed related to this topic was Tajal Behangle's analysis which looked at predicting location and time of crimes being committed and socioeconomic factors.

3 Data Set

The data set which we are using for our project comes from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2, which has been made available by the Chicago Police Department. It includes crimes that are reported in Chicago, IL from 2001 to present. There are 7.08 million instances and 22 attributes for the data. Here is a list of all of the attributes:

- ID (Number) - Unique key for tuple
- Case Number (Text) - Unique number for the case
- Date (Floating Timestamp) - Date and time when the crime occured.
- Block (Text) - Partial address of crime location.
- IUCR (Text) - Illinois Uniform Crime Reporting Code which is based on crime type and description.
- Primary Type (Text) - Primary description from IUCR code.
- Description (Text) - Secondary description of the IUCR code.
- Location Description (Text) - Description of the location where the crime occured.
- Arrest (Boolean) - Whether an arrest was made for the crime or not.
- Domestic (Boolean) - Whether the crime was domestic or not.
- Beat (Text) - Beat where crime occured. A beat is the smallest geographic area that police coverage is divided into.
- District (Text) - Police district where crime occured.
- Ward (Number) - Ward where crime occured.
- Community Area (Text) - Community area where the crime occured.
- FBI Code (Text) - FBI's crime classification.
- X coordinate (Number) - X coordinate where crime occured.
- Y coordinate (Number) - Y coordinate where crime occured.
- Year (Number) - Year crime occured.
- Updated On (Floating Timestamp) - Date and time record was updated.
- Latitude (Number) - Latitude where crime occured.
- Longitude (Number) - Longitude where crime occured.
- Location (Location) - Coordinates of where crime occured.

4 Evaluation Methods

We will make use of metrics such as accuracy, error rate, and tools like confusion matrices for classification tasks. We also plan to use mean squared error for inference/prediction tasks and the holdout method to train and test sections of data.

5 Tools

- Python
- Jupyter Notebook
- Sklearn (KNN)
- Python statsmodels.api (Linear/Logistic Regression)
- Numpy
- Pandas
- Tableau

6 Milestones

- Data collection: Complete
- Preprocessing of data: March 30
- Analysis of data: April 13
- Write up Final Report: 1 May

REFERENCES

[1] Djona Fegnem. 2017. Chicago Crime Analysis. Retrieved March 10, 2020 from https://www.kaggle.com/djonafegnem/chicago-crime-data-analysis

[2] Sadaf Tafazoli. 2018. My notes on Chicago Crime data analysis. Retrieved March 10, 2020 from https://medium.com/@stafa002/my-notes-on-chicago-crime-data-analysis-ed66915dbb20

[3] Chicago Crime Portal. 2020. Crimes - 2001 to present - Dashboard. Retrieved March 10, 2020 from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g

[4] Tejal Bhangale. 2017. Chicago Crime Analysis. Retrieved March 10, 2020 from https://github.com/Tbhangale/Chicago-Crime-Analysis

[5] FBI Crime in the United States Report 1995-2019 https://www.fbi.gov/services/cjis/ucr/

[6] Josh Sanburn. 2016. Chicago is responsible for almost half of the increase in U.S. homicides. Time Magazine online https://time.com/4497814/chicago-murder-rate-u-s-crime/