Crime in Chicago

A Data Visualization Project

Melanie McCord

Write 4-5 paragraphs that expand your previous proposal, answering the following questions.

1. What is the problem?

The goal of this project is to examine the crime rate in the City of Chicago and its relationship with time and location (i.e. zip code and neighborhood). Specifically, I will be visualizing a few key factors:

* How has the crime rate changed over time from January 2001 (the beginning of data collection) to the most recent date?
* How does the crime rate relate to some other variables: - the neighborhood where the incident was reported - the demographics of the neighborhood - the population density of the neighborhood and/or Chicago - the unemployment rate in Chicago (overall) and the specific neighborhood (if available)
* How does the proportion of arrests depend on: - the category of crime (included in the dataset) - crime itself (focusing on the top 5 most and least common)
* (stretch goal) Special focus on gender-related crimes (e.g. domestic violence, sexual assault, stalking, etc.)
* How do the arrests and other factors of these crimes compare with the overall arrests and other factors?
* (additional stretch goal) => more detailed exploration of some related crimes, including a brief page on that as well

1. Why is solving the problem important?

The City of Chicago’s crime dataset [1] is fairly extensive. It covers every reported crime since 2001, updated daily. As a result, many works have made use of this dataset to solve a variety of crime-related problems. This dataset has a variety of uses. It includes variables related to the location of the incident, the category of the crime, as well as the date the crime occurred. It is both a time series dataset and a spatial dataset and can be used for either or both depending on the use case. - What prior work (1-2 papers/projects should be good) has done something similar? Why naively apply prior solution does not work.

Due to the extensive nature of this dataset and its applicability to real-world problems, it has many purposes that are facilitated by visualization. Kim et. al visualized the time series element of violent crimes to assist in the creation of an ARIMA model to compare violent crime instances during and before the COVID-19 pandemic. [2] Another work by Kim et. al. analyzed violent crime areas, predicted safe areas, and generated a visualization to help guide users with finding safe routes in Chicago to travel. [3]

My proposed research question is to visualize the crime rate about variables including arrests, type of crime, and unemployment. A future extension could be to implement crime prediction models and visualize those as well to guide decision-making related to crime prevention efforts. Time permitting, I will also answer a specific research question regarding gender-related violence, such as stalking or sexual assault. Specifically, I would compare the crime rate overall with the rate of gender-related crimes, and compare the proportion of arrests, locations, etc. where gender-related crimes occur.

* What is your data?

The dataset contains roughly 8 million rows as of February 14th. The data is dynamic, and the frequency of the data updates is daily. However, my starting point is the dataset from 01/01/2001 to 02/13/24. Note that the user will not be able to view my visualizations from GitHub Pages. Instead, they will need to view my visualizations by explicitly downloading the dataset itself.

Table 1: Variables and descriptions of the Chicago Crimes Since 2001 dataset

| ID | Unique identifier for the record. |
| --- | --- |
| Case number | Chicago id for the case number |
| Date | Date when the incident occurred, this is sometimes a best estimate. |
| Block | The partially redacted address where the incident occurred. |
| IUCR | The Illinois Uniform Crime Reporting Code. |
| Primary Type | The primary description of the IUCR code. |
| Description | The secondary description of the IUCR code. |
| Location description | The primary description of the location where the incident occurred. |
| Arrest | Whether or not the incident resulted in an arrest. |
| Domestic | Whether or not the incident was a domestic incident. |
| Beat | Indicates the beat where the incident occurred. |
| District | The police district where the incident occurred. |
| Ward | The city council district where the incident occurred. |
| Community Area | The community area where the incident occurred. |
| FBI Code | FBI Code crime classification. |
| X Coordinate | The X coordinate location where the incident occurred. |
| Y coordinate | The Y coordinate where the incident occurred. |
| Year | Year the incident occurred. |
| Updated on | Date and time the record was last updated. |
| Latitude | The latitude where the incident occurred. |
| Longitude | The longitude where the incident occurred. |
| Location | The location of the incident. |

1. How you are likely going to solve the problem.

The purpose of the website is to visualize different questions regarding the Chicago dataset over multiple categories. Note that the primary audience for this is researchers, so some of the time series graphs in particular may not be clear to a general audience with no training in time series analysis. However, there will also be histograms and other common visualizations that are more commonly understood.

Although each page will represent different research questions, there are some stretch features for the ones that explore the dataset generically:

* Dropdown menu to adapt the color according to the value of some of the categorical variables
* Domestic or not
* One or more of the reporting codes (IUCR, FBI Code) and/or Primary Type
* Filter by date range (potentially only by year)

Pages of the Actual Website with Specific Purposes

* Home Page: an overview of the dataset, links to the original dataset, quick overview of its documentation.
  + Columns in the Dataset
  + An introduction to the various research questions and purposes of the dataset, citation to the dataset.
* Time Series Analysis: How has the crime rate per 1000 people in Chicago changed over time?
  + Visualizations
  + ACF plots
  + PACF Plots
  + Line graphs
  + Seasonal plots
  + Lag plots
  + STL decomposition of crime rate over time
* Spatial Data Analysis
* - pick 2-3 from the following:  
    
   - proportional symbol map  
   - cluster map  
   - choropleth map  
   - cartogram map  
   - hexagon binning map  
   - heat map  
   - contour plot  
   - time-space distribution map  
   - option to visualize this over time
  + Categorical Variable Analysis
    - Top 5 most common Primary Codes
    - Arrests: 5 crimes with highest proportion and lowest proportion (bar plot)
    - Top 5 least common Primary Codes
    - Distribution of frequency of primary codes (histogram)
* Gender-Related Crimes (stretch goal)
  + Pick most readily understandable visualizations from each category
  + Aggregate the ones that are considered gender-based (e.g. domestic violence, sexual assault, etc)
  + Look at each of these crimes individually and compare reporting frequency by arrests
  + Compare side-by-side with distribution overall of crimes (e.g. colored lines marking the frequency on the histogram, side by side comparison of the crime rate)
* Income Related Variables (requires external dataset)
  + Acquiring data for as many years as available from federal agencies. Example datasets:
  + [Federal Bureau of Economic Analysis - 2022 Personal Income](https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas)
  + [Department of Labor - Raw Unemployment Data in the City of Chicago - from this list](https://data.bls.gov/cgi-bin/surveymost)
  + Picking 1 or more employment datasets by zip code from [here](https://beta.bls.gov/dataQuery/find?st=0&s=popularity%3AD&fq=survey:%5Boe%5D&fq=cc:%5BMetropolitan+Statistical+Areas%5D&more=0&q=chicago&r=100&st=0)
* Demographics to Predict Crime
  + Demographics of Chicago (overall and by zip code) (filter by year). See [Census Data API](https://data.census.gov) (note that this data is estimated yearly, census collects true data every 10 years; may not be available for all years 2001-2023)
  + Visualize this data over time (most generally interpretable graphs only)
* Both Demographics and Income as Predictors of Crime
* 3D visualizations: will decide between these options
  + color (map median income into an ordinal variable)
  + default overall, options to filter by specific zip code
  + transparency (use unemployment rate)
  + side by side with categories of demographics
* line chart of crime rate with separate lines depending on proportion of demographic categories over time (default: race and gender, user can pick whatever subset they choose)

# Summary

My proposed research goal is to identify multiple predictors of the crime rate, as well as data related to arrests, locations, and demographics of these areas. I hope that this project can serve as a guide for further research into crime prevention.

[1] “Crimes - 2001 to Present City of Chicago Data Portal.” Accessed: Feb. 14, 2024. [Online]. Available: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about_data>

[2] D.-Y. Kim and W. P. McCarty, “Exploring violent crimes in Chicago during the COVID-19 pandemic: Do location, crime type, and social distancing type matter?” *Journal of Crime and Justice*, vol. 45, no. 4, pp. 522–537, Aug. 2022, doi: [10.1080/0735648X.2021.2014934](https://doi.org/10.1080/0735648X.2021.2014934).

[3] “Safe Route Recommendation based on Crime Risk Prediction with Urban and Crime Data IEEE Conference Publication IEEE Xplore.” Accessed: Feb. 14, 2024. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/10233968?casa_token=B0fa7VMrC1UAAAAA:3S98bupMk0MfzwJCYc8tMJEbAEyUWpse6KnKT-IY17oVgQzlLFqRO7Ayel2Fxbvi-eSuJ0dQXg>