Chicago Crime Analysis

Group 5: A Data-Driven Approach to Understanding Crime Patterns

William Cole Akers
Student
University of Colorado Boulder
wiak0179@colorado.edu

Steven Delaney
Student
University of Colorado Boulder
steven.delaney@colorado.edu

Problem Statement and Motivation

With this project, we intend to analyze crime trends in Chicago using historical data ranging from the years 2001 to 2023. We will conduct an in-depth analysis of the crime trends present in this data set. Using statistical techniques, we seek to understand how the patterns in crime have changed and further to infer future changes so that law enforcement could use the findings to help mitigate future crimes and plan for changing crime patterns in the years to come. Using the knowledge gained from the analysis, we hope to uncover and present trends that show patterns across different locations and time periods.

Firstly we will identify what areas in Chicago are "hot spots" for violent crimes. We will analyze which areas experience the highest rates of crime and what types of crimes are most common there. Additionally, we will explore how the crime landscape has changed ward to ward. We can apply this knowledge to help law enforcement better understand what areas they need to concentrate on more heavily in the present day. This could be used to help create policing routes and patrolling areas emphasizing areas that have been identified as hot spots.

We also intend to answer the question, "Is crime impacted by seasonality?" That is to say, we want to be able to tell if crime rates increase or decrease during certain months and do these changes persist in the same season year to year. When patterns are identified, law enforcement will know when to expect more or less crime activity. This can help with the preparation of officers on the individual level, allowing them to know to expect more or less crime. It can also help at an administrative level where the number of patrols could be increased or decreased to match the crime rates of the current season.

As one measure of performance, we will investigate which districts and wards have the highest and lowest arrest rates for crimes that warrant arrest. This examination will allow us to see if there are noticeable discrepancies between districts. By identifying districts that do not have a high success rate for arrest, we can compare differences in policies and practices between lower-performing districts and wards and higher-performing ones. This analysis may also bring to light if performance is a factor of policing quality or if a difference in resources plays a bigger role.

A third area of analysis is whether or not there is a correlation between location and type of crime. Using the data we can find out which crimes are most prevalent in certain neighborhoods or under specific environmental conditions. Armed with this knowledge, officers patrolling certain locations can be better outfitted to recognize and mitigate specific criminal activities that are common in the areas they patrol. Lastly, we will conduct a temporal analysis based on time of day. For each case the data provides a time/date stamp that is space delimited and can be easily parsed. Time will be used to illustrate what crimes are most prevalent in a 24-hour period.

Literature Survey

We have identified three previous studies that performed research on crime trends, law enforcement strategies, and data-driven crime prevention.

The University of Chicago Crime Lab studies conduct research on Chicago crime trends with a focus on policy interventions and the impact of law enforcement strategies on crime rates. These studies highlight evidence-based intervention and the resulting reductions in violent crime. Additionally, they evaluate programs like predictive policing and youth crime intervention. The University of Chicago Crime Lab studies give us an inside look at historical crime trends and which intervention strategies worked well to mitigate crime rates.
URL**:**
https://crimelab.uchicago.edu/resources/2024-end-of-year-analysis-chicago-crime-trends/

The Chicago Police Department (CPD) crime Statistics study focuses on annual crime statistics with an official record of reported crimes, arrests corresponding to these crimes, and success rates of law enforcement. These reports examine this data, differentiating by district and category of crime. The key findings include statistics that show patterns in crime occurrence, arrest rates, and law enforcement responses.
URL**:**
https://www.chicagopolice.org/statistics data/crime-statistics/

The third study from the Journal of Data Analysis and Information Processing focuses on long-term crime trends in Chicago. These trends are examined using machine learning techniques and statistical monitoring. This study revealed that crime rates fluctuate depending on socioeconomic conditions, seasonal trends, and neighboring infrastructure. We can use the information found by this study to help us check our findings for seasonal trends and as a model to check which statistics from our data set are needed to conduct this type of evaluation.

URL: https://www.scirp.org/journal/paperinformation?paperid=134329

Figure 2: Chicago skyline
Google maps image

Proposed Work

Data preprocessing and cleaning is the first task for this project. First, we intend to detect outliers and remove any that may skew the results of our analysis. Using statistical techniques like z-score analysis, interquartile range, and clustering to identify data points that could cause skewing

Additionally,we will check for any missing values and, where possible, fill in with either mean, median, or mode. If no replacement can be found for the value, we will handle it case-by-case if very few missing values are found for either data or categories, filling in missing values when possible. Filling in categories as unknown when no other reasonable solution is achievable. We will then use tools to display the data and look for correlations that will help us answer the questions we presented in the problem statement section. Some questions we are attempting to answer, like looking at the crime trends over the years, success rates of law enforcement, and seasonal trends, replicate the studies we observed and summarized in the literature survey section. However, our project adds several unique observations that could be of use to law enforcement. These include looking for a correlation between location and type of crime as well as identifying hot spots for crime.

Dataset

The primary dataset for this project is sourced from Kaggle and was originally collected by the Chicago Police Department as part of the CLEAR (Citizen Law Enforcement Analysis and Reporting) System. The dataset spans from 2001 to the present, containing over 7 million reported crimes with the following attributes: ID, Case Number, Date, Area (location), IUCR, Type of crime, Description, Location (street, residence, or other), Arrest status, and Domestic.

Dataset URL: Chicago Crimes Dataset –

https://www.kaggle.com/datasets/utkarshx27/ crimes-2001-to-present

Evaluation Methods

We intend to evaluate our findings by cross-referencing prior research and comparing our findings to them. By using the existing crime research like the methods mentioned in the literature survey section, we can validate our findings against existing solutions. We will look for correlations that answer our research questions with the hope that we find expected correlations or definitive analysis that disproves correlations we were searching for.

We plan to develop a predictive crime model and compare it against existing crime prediction models to evaluate its accuracy and effectiveness. Our model will use crime type and time of occurrence to predict where crimes are most likely to happen, leveraging historical data to identify patterns and trends. This prediction model will focus on tying locations to crime types and descriptions. It is our aim to create a model that might aid in narrowing in on the location of a crime when a vague description of the area of the crime is given.

To validate our approach, we'll test our model against proven methods like logistic regression, decision trees, and neural networks, assessing its performance based on key metrics like precision, recall, and accuracy. This comparison will help us gauge how well our data handling and methodology hold up against established standards.

By refining our model through this process, we aim to create a reliable tool for crime forecasting, one that can support law enforcement, resource allocation, and crime prevention strategies with data-driven insights. Specifically, we aim to help narrow down the location where a crime may have occurred based on the type of crime committed.

Tools

Visualization: Tableau.

Analysis: Pandas, NumPy, Matplotlib, Alteryx

Programming Language: Python.

Development Environment (IDE): VSCode, Jupyter Notebook.

Collaboration: GitHub, Outlook, Discord.

Milestones & Timeline

Data Cleaning & Preprocessing	Mar (15-16)
Initial Exploratory Analysis	Mar (17-21)
Crime Hotspot & Time-Series Analysis	Mar (24-28)
Arrest Success & Correlation Studies	Mar (24-28)
Model Development	Mar (31-Apr 11)
Final Report & Visualization	Apr (21-26)

Milestones Completed

Targets:

- Data cleaned
  - Address dropped as it was partially censored data, making it unusable
  - Crimes that do not warrant arrest will be separated into their own category of analysis.
  - Rows with null latitude and longitude
- Initial exploratory analysis
  - We have derived a set of initial summary statistics that will lead to further exploration.

To clean the data, we dropped the address column as we found it to be extraneous information. This is because the data set offers longitude and latitude as well as a precinct that are much more useful for plotting on a geographical map. Additionally, we removed from our data sets crimes that do not warrant arrest, and instead, the repercussions are a citation or non-arrest-worthy penalty. Our original approach to find crimes where the punishments do not warrant arrest was to look for primary types that always resulted in a false arrest rate. However, we discovered that no such primary type existed.

We discovered this is caused by the different scales of crime types. An example of scaling that we encountered is theft. A theft where the stolen item is valued at $100 would not be a crime where arrest is warranted, but when the $ value is increased, the crime warrants an arrest at a certain point of the upscaling. We were able to find crimes that are not arrest-worthy by instead filtering through the descriptions of crimes and then dropping crimes with descriptions that never lead to arrest. We did this to get an acurate look at arrest rates by precinct where crimes that do not warrant arrest may be more common and lead to lower arrest rates.

Milestones To-Do

Still to do:

- Crime hotspot time-series analysis
- Arrest success and correlation studies
- Acquire additional dataset on Chicago property values
- Develop a predictive model
- Final Report & Visualization

We plan to use crime hotspot time series analysis to analyze the shift of crime hotspots over the years our data set takes place. Using arrest success and correlation studies, we will compare the success rates of different wards and precincts by using the total number of crimes that warrant arrest and the number of successful arrests in these predefined divisions. Our final steps will be running a machine learning model through our data, focusing on arrest location and the success of arrest rate. It is our goal to create a model that can project future changes in success rate by ward/precinct.

Results So Far

In our initial exploration, we have derived several summary statistics that will guide further exploration and experimentation.

One major finding is the consistent drop in the number of cases year over year, illustrated in Figure 1 below.

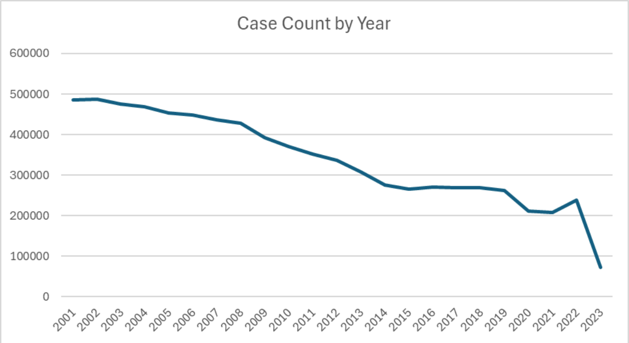

Figure 1 Case count trend by year

The number of cases has dropped at an average rate of 6% each year compared to the previous year. This begs the question, why? Is this a result of better policing or changes in how crimes are reported?

Additionally, we have begun to conduct a geospatial analysis utilizing Tableau. Specifically, crime density by zip code. This allows us to easily identify areas with the highest and lowest volume of reported crimes. We Aimed to take a more granular approach to our refinement of the data by separating the geographical space into more areas building on what previous research projects have already done. See Figure 2 below

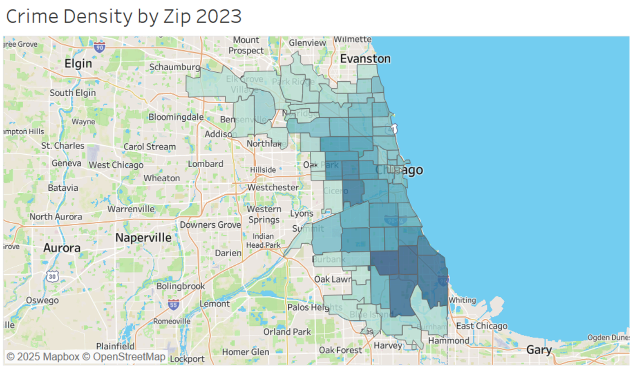

Figure 2

Figure 2 clearly illustrates that in 2023 South Central Chicago was a hot spot for crime while the Upper West Side had a much lower volume. This opens several additional avenues we can go down. One next step would be to run the same analysis for multiple years to see if there is a geographical change in density over time or if hot-spots have remained consistent for several years. We will also explore potential reasons why the identified areas might have higher or lower crime rates. The factors we discover will play a major role in our predictive modeling.

Responses to Feedback

What method of cluster analysis will be used?

We used K-Means clustering for our analysis, specifically applying it to latitude and longitude data to identify geographical outliers in reported crimes. This method helps us group crimes based on relative location, allowing us to see patterns that might not be obvious at a glance.

We've organized the crimes into five distinct clusters, giving us a view of crime distribution that is purely data driven across the city. By analyzing these clusters, we can pinpoint high-crime areas, detect anomalies, and compare crime density across different regions. This approach helps us understand how crime is spatially distributed and whether certain areas exhibit unique trends or outlier behavior.

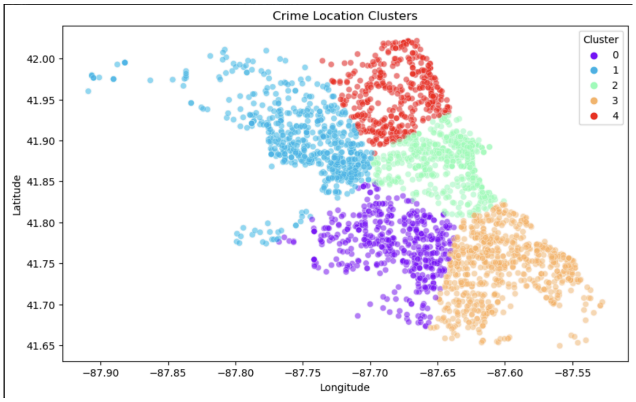

Figure 3 K-means clustering

Through cluster analysis, we grouped crime-prone areas into five distinct clusters based on the longitude and latitude of where the crime occurred. This allows us to analyze crime geographically from an overarching viewpoint. This allows us to double-check our findings for high-density crime areas that are larger in scale and do not conform to pre-defined borders that are present in the data set.

Each of the five clusters represents a grouping of crimes that occur in geographically similar areas as determined by k-means. Using these larger groups to look at arrest percentages will help us discover if there are larger trends that exceed precinct/ward or zip code limits.

Moving forward, we plan to analyze each cluster to determine common characteristics, evaluate potential influencing factors, and explore how crime shifts within and between clusters over time.

How does this build on previous work, and what differentiates it from these studies?

Our work will most closely mirror that of *"The Windy City's Dark Side: A Statistical Exploration of Crime in the City of Chicago"* (Odooh et al., 2024.) Not only was the same data set used in this research, but many of the same questions were asked and explored. Particularly correlations in location and crime type, the impact of seasonality on crime, and crime density based on location. We will differentiate our work by taking economic factors into account, namely correlations between property values and how this impacts arrest rates and crime density. Economic impact is not a factor considered in the previously referenced work.

Additionally, we also consider how departments have analyzed crime throughout the years and what they consider to be their measure of success. Tableau allows us a unique opportunity to easily create geographic maps that these studies do not include in their analysis. If they do have the map, Tableau allows us to add more distinct borders to examine the data with a finer-toothed comb. This allows us more granularity than previous studies, which we hope will lead to better accuracy when pinpointing success rates. The more granular approach is inspired by the suggestion for a continuation in the future research directions section of the Journal of Data Analysis and Information Processing study mentioned in the literature review section of this proposal.

Our graphs help visualize the changes in time and arrest rates in a more readily apparent way for consumers of the report to get a more complete understanding of the studies in a much shorter time frame. We also include a more expansive scope of crimes committed, whereas the studies before focused on violent crimes. This differentiation leads to different hypotheses and conclusions for example, the previous studies focused on the lethality of shootings. Previous studies also picked the years they would include with an aspect of randomness. We plan to analyze in three-year chunks, which will give us a more complete look at the changes in the attributes we are observing.

Finally, our most clear-cut distinction from other studies is the development of a machine learning model that will help to predict the location of a crime based on the type of crime reported. While previous studies have graphically illustrated the data, none of our references have attempted to use the data to make a predictive model. This model could be applied by law enforcement to increase response times to crimes by helping to predict the area of the crime when an emergency line caller gives vague or nondescript locations. In terms of building off of previous studies and differentiating our work from what has been done previously, this model sets our exploration apart from the work covered in our literature review.