Chicago Crime Analysis

Arjun Ravikumar, Tejal Shanbhag , Utkarsh Havle

1. INTRODUCTION

Crimes are something that happens every day in all parts of the world. Some crimes are organized crimes and others are unorganized. There have been multiple risk factors that have been the underlying reason for these crimes happening. Three major categories of crime risk factors: (1) biological factors; (2) socioeconomic factors; and (3) psychological factors, as a result of the individual value system of those involved in criminal activity. Chicago is a city that has a history of having one of the highest violent crime counts in the country. According to [3] Chicago has had 40,000 homicides from 1957 - 2020. For the first time since 1957, the homicides went below 500 for a year in 2019 in Chicago. This might not seem like a large number but we need to consider that these numbers are of just a city and not a state in the US. From the beginning of the 20th century, the Chicago Police department has been tracking the crimes happening in the city and making them accessible to the general public through their data website [2].

We propose to study these relationships between the socioeconomic factors of the residents of the area to the number of crimes happening in that area. The dataset we are utilizing for the same is of Hardship Index [4] and Crimes: 2001-present [1] from the City of Chicago's data portal.The Crimes: 2001-present dataset[1] consists of 7,279,930 rows of crimes between 2001 - February 18th 2021. This dataset[1] has 22 attributes which are id, case_number, date, block, iucr, primary_type, description, location_description, arrest, domestic, beat, district, ward, community_area, fbi_code, x _ coordinate, y_coordinate, year, updated_on, latitude, longitude and location. The HarshShip Index data [4] consists of 78 rows and has a selection of six socioeconomic indicators of public health significance and a "hardship index". This data[4] has 6 attributes which are hardship_index, community_area_name, percent _ households _below _ poverty, percent_aged_25_without_high_school_diploma, percent_aged_16 _ unemployed, percent_aged_under _18_or _ over _64 and per_capita_income_. In this project, we intend to do data exploration where we try to visualize the data to analyse the trends in the data. We try to visualize the number of crimes against the year, day of the week, the hour of the day, location. We also try to visualise the type of crimes and their count and the severe and not severe crimes against the hardship index and income. Finally, we try to predict the severity of the crime using the Naive Bayes Classifier.

The rest of the study would be conducted as follows: Section 2 provides the motivation for choosing this topic. Section 3 provides information about the design of the project. Section 4 describes the implementation and analysis of the methods and data respectively in detail. Section 5 will describe inferences from the analysis followed by the future work in the project.

2. PHASE 2: EXPLORING THE SHAPE AND STRUCTURE OF THE DATA

2.1 Issues in the data

We have identified mainly two issues with the raw dataset [1, 4] that are missing data and incorrect data. We did data visualization of the fields of both datasets to identify missing values. Missing values were easily identifiable after data visualization but incorrect values were not that evident on data visualization. The data visualization and our findings are explained in detail in further Section 2.1.1 and Section 2.1.2.

2.1.1 Missing Data

Missing values were something that we were expecting when we had considered the raw dataset. Hence this is one of the first things which we looked for and found. Figure 1 shows the missing data marked in yellow in the dataset [1] and complete data marked with purple. We could not notice a pattern in the missing data values and hence suspect these to be a case of missing at random. With the second dataset also found similar results after data visualization in Figure 2. Similar to Figure 1 in Figure 2 missing values are marked in yellow and complete data is marked in purple.

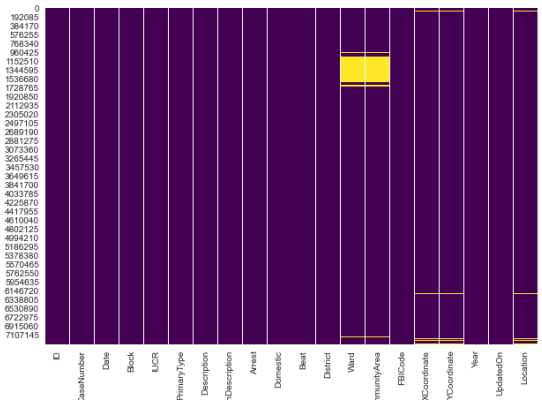

Figure 1: Crimes dataset [1] missing values

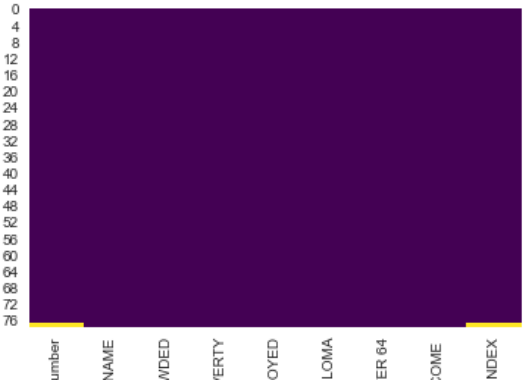

Figure 2: Hardship Index dataset [4] missing values

2.1.2 *Incorrect Data*

Incorrect values are something that were really tough for us to identify. These were not as evident to us as the missing values were but visualization helped us to identify some of these issues. In Figure 3 we can notice that some of the crime districts are marked in locations of another district.

2.1.3 *Cleaning data plans*

We plan to analyze the data some more and get some more information about the dataset [1, 4] before finding a way to handle the issues in the dataset.

2.1.4 *Other data findings*

While exploring the data we found that there was a sudden peak in crimes like sex offense, obscenity, criminal sexual assault, concealed carry license, and weapons violations spike after 2017. Also, some crimes have reduced or are on a downward trajectory after 2020. Which might be due to COVID-19.

2.2 Future plans

We plan to clean the data first by gaining some more domain knowledge. After which we plan to visualize the

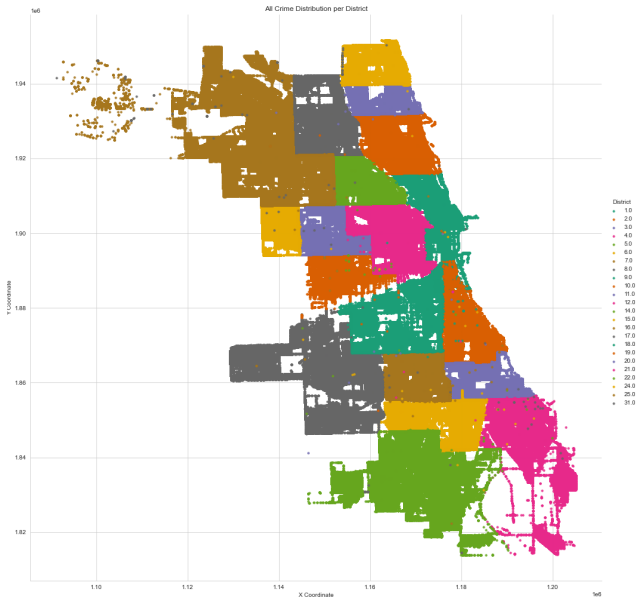

Figure 3: Incorrect districts in [1] dataset

data again and identify any patterns in the data. Then we plan to use the Naive Bayes Classifier to predict the severity of the crimes and also find a pattern between severe and not severe crimes against the hardship index and income. We also plan to further investigate the patterns of data which was explained in Section 2.1.4.

3. REFERENCES

[1] Chicago Police Department. Crimes - 2001 to present. `https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2`. Accessed: 2021-02-19.

[2] City Of Chicago. Chicago data portal. `"https://data.cityofchicago.org"`. Accessed: 2021-02-19.

[3] Kyle Bentle & Jonathon Berlin & Ryan Marx & Kori Rumore. 40,000 homicides: Retracing 63 years of murder in chicago. `https://www.chicagotribune.com/news/breaking/ct-history-of-chicago-homicides-htmlstory.html`. Accessed: 2021-02-20.

[4] U.S Census Bureau. hardship index: Census data - selected socioeconomic indicators in chicago, 2008 – 2012. `https://data.cityofchicago.org/Health-Human-Services/hardship-index/792q-4jtu`. Accessed: 2021-02-15.