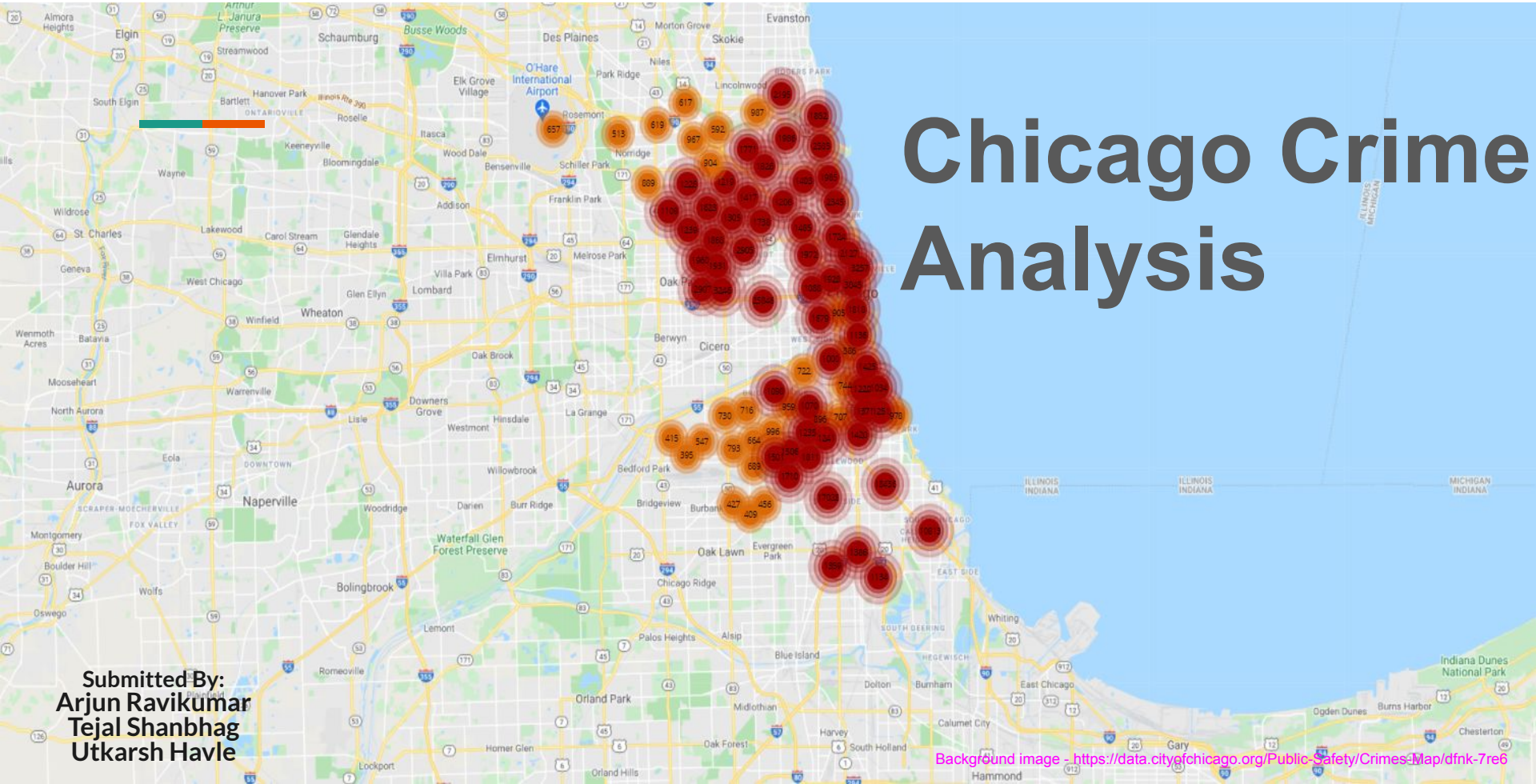

Chicago Crime Analysis
Submitted By:
Arjun Ravikumar
Tejal Shanbhag
Utkarsh Havle
Background image - https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6

Contents

Introduction

- Crimes are something that happen globally.
- There could be multiple reasons having various risk factors for these crimes to happen.
- Three major categories of crime risk factors[6]:
  - Biological factors[6]
  - Socioeconomic factors[6]
  - Psychological factors[6]
- This could also happen as a result of the value of the individual in society involved in the criminal activity.
- Chicago is a city that has history of having one of the highest violent crime in USA.
- From the start 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 portal [5].

Motivation

- In the day and age where armed shooting at a public place has become a regular column in the newspapers , we wanted to find out if there was any pattern in theses crimes to occur.

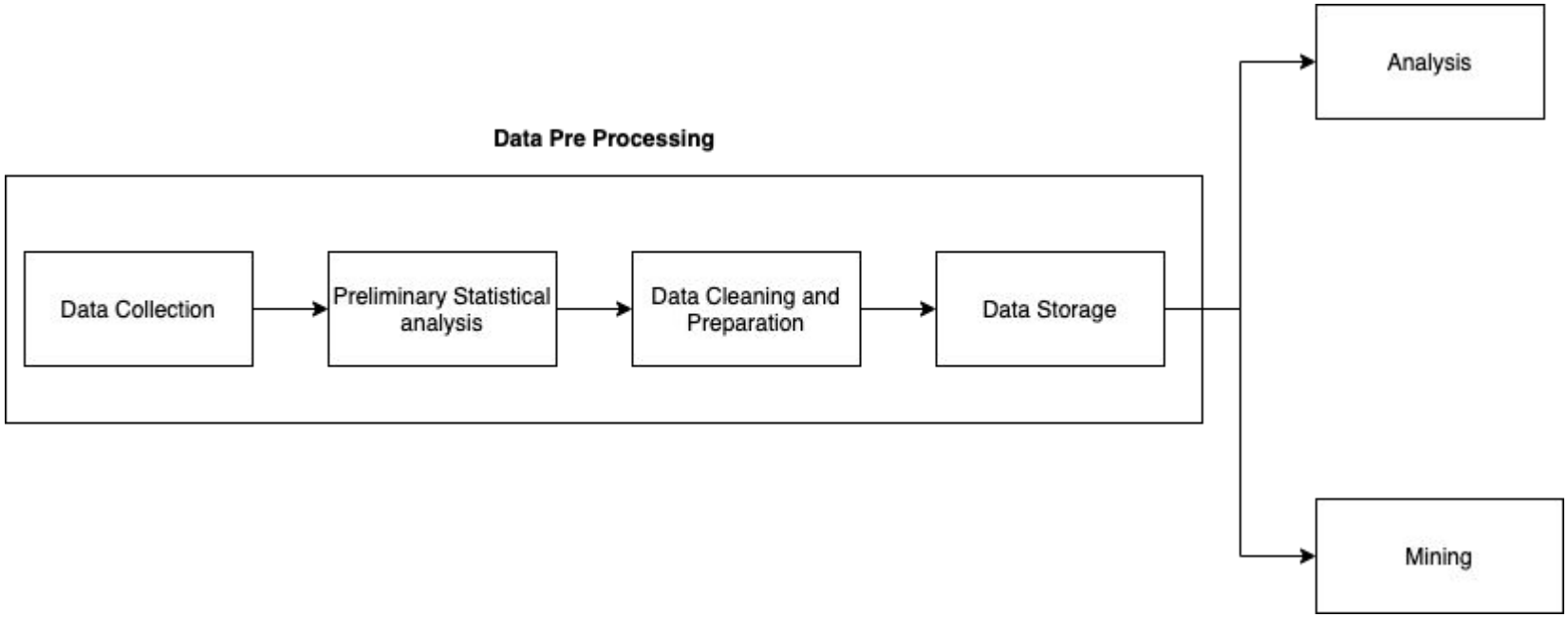
- We wanted to study if society and its factors like unemployment, lack of proper education, or crowded households can affect a community of people to commit crimes.

- We have selected the city of Chicago for this study mainly because .Chicago has always had a very bad record against the prevention of crimes.

- Due to these number of crimes , Chicago is never considered to be one of the unsafe cities in the US.

- Chicago also keeps a record of all the individual cases that occurred and also the number of attributes provided for each crime is what made Chicago the only candidate for this study.

Project Phases



I. Data Collection

- The data for this project was collected from the Chicago city data portal [4] in the CSV format.
- The datasets used for this project are:
    - Crimes in Chicago from 2001 to 2021 [1]
    - U.S Census Bureau hardship index [2]
    - City of Chicago. Boundaries - community areas [3]
- The Crimes dataset [1] has in all 7.2 million records with 22 attributes.
- The Hardship index dataset [2] has 78 records with 9 attributes.
- The community areas dataset [3] was used to fill in the missing values of community attribute in the Crimes dataset [1].

II. Preliminary Statistical Analysis

- In order to understand the collected datasets [1][2] better , we performed this step.
- During this step we encountered following discrepancies in the data.
  - Missing Values
  - Incorrect Values
- Along with the above discrepancies there were many attributes which were very unique to the case.
- There were also many categorical values for location where the crime took place and type of crime.

III. Data Cleaning and Preparation

- In this stage of the project we have handled the issues we found with the datasets[1][2].
- Following is the detailed explanation of how the issues were handled:

    - **Unique primary key attributes**
      Elimination of attributes in the datasets[1] [5] which were unique values and were used for documentation purpose.(eg. Case number)

    - **Redundant attributes**
        - The Crime dataset had duplicate columns of Longitude and Latitude as a tuple named as Location.
        - We eliminated this tuple data as it was repetitive values and was accessible through the Longitude and Latitude columns.

    - **Missing Values**
        - Almost 8% of the records had missing community and ward values. As this chunk of records had valuable information on the crimes we decided to fill these missing values by using the community area dataset[3].
        - The correlation between the location attributes Longitude ,Latitude , X Coordinate,Y Coordinate was indicator that if one of the values missing then other 3 also missing.
        - So we eliminated this subset of records.

III. Data Cleaning and Preparation

- Following is the detailed explanation of how the issues were handled
  - **Grouping to broader categories:**
    There was 38 different crime types and 212 different locations types where crimes occurred. Due to which we had to reduce the dimensionality of these categories by grouping them into broader categories.
    - **The crimes were grouped into the below categories:**
      1. **Crimes involving physical harm to humans**
      2. **Crimes involving monetary benefit**
      3. **Crimes involving safety concerns to humans**
      4. **Crimes involving violations**
      5. **Other non-criminal offenses**

    - **The location of crimes were grouped into the below categories**
      1. **Residential Area**
      2. **Business Area**
      3. **Vehicle**
      4. **Public Buildings**
      5. **Public Area**
      6. **Government Locations**
      7. **Public Transportation**
      8. **Other**

III. Data Cleaning and Preparation

- Following is the detailed explanation of how the issues were handled
  - **Subdividing attributes**
    - The date of when the crime occurred was in the format **month-day-year hour:minute:second.**
    - In order to extract useful information from the date attribute, we divided it into multiple attributes. Therefore we created the subdivided attributes hour, day, month and weekday.
  - **Encoding nominal data**
    - After the data analysis, numeric records were encoded for the classification.
    - Crime type was categorized into two classes severe and non-severe as we wanted to perform binary classification on the data.

IV. Data Storage

- We chose MySQL for storing the datasets after the cleaning stage.
- After performing the Preprocessing of the data, the datasets were stored in the database for performing analysis on the data.
- After the Data analysis, the nominal data was encoded and used for mining purpose.

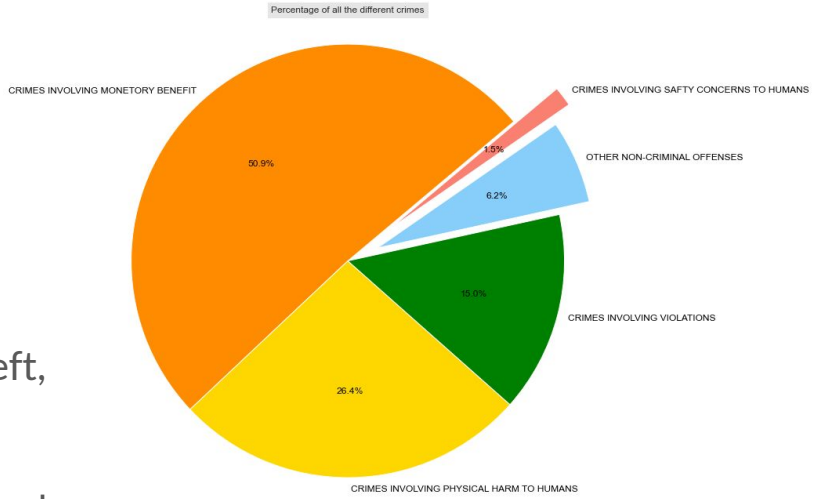
V. Data Analysis

1. Crime Based Analysis

Top Crimes happening in Chicago

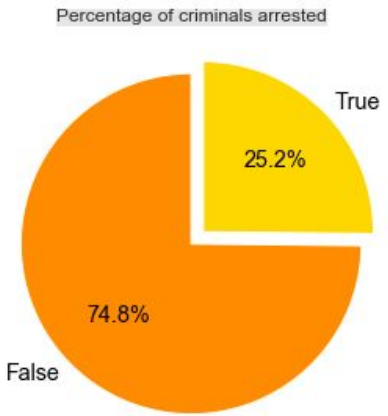
1. Crimes involving monetary benefit like theft, burglary etc are the highest in the city.

2. Then crimes causing physical harm is the 2nd highest in the city of Chicago.



Are these Criminals arrested ?

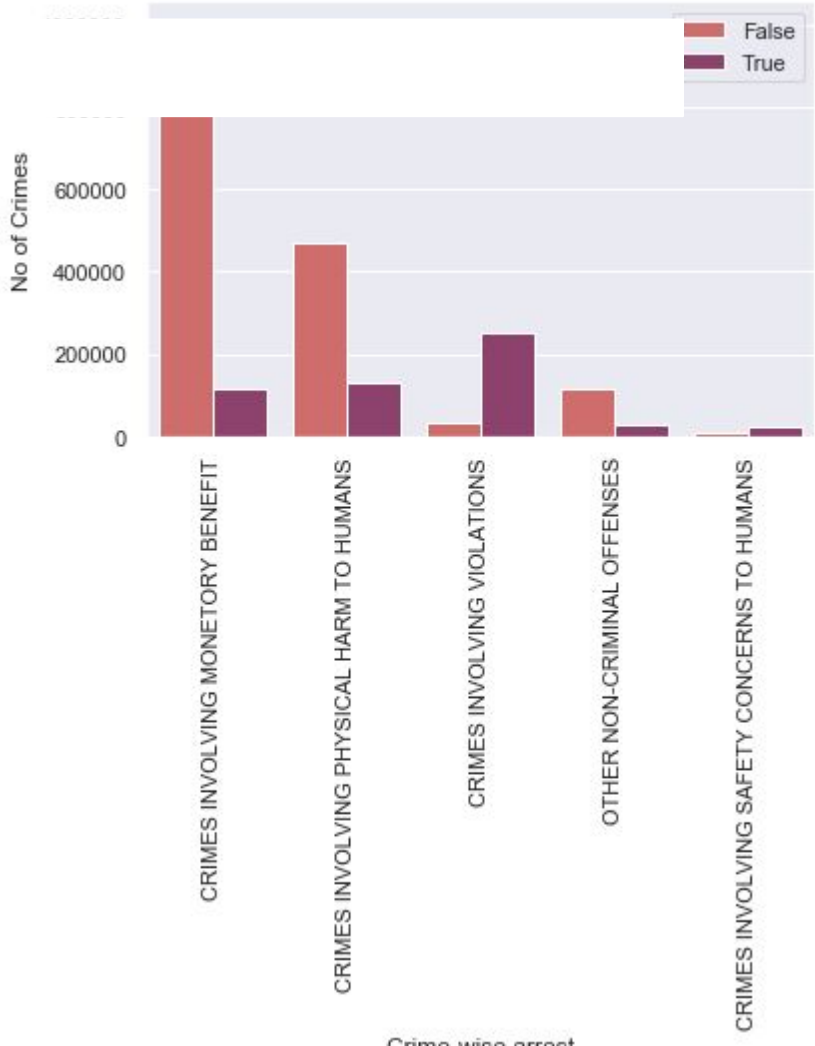
Almost 75% of the criminals remain unarrested.



V. Data Analysis

1. Crime Based Analysis

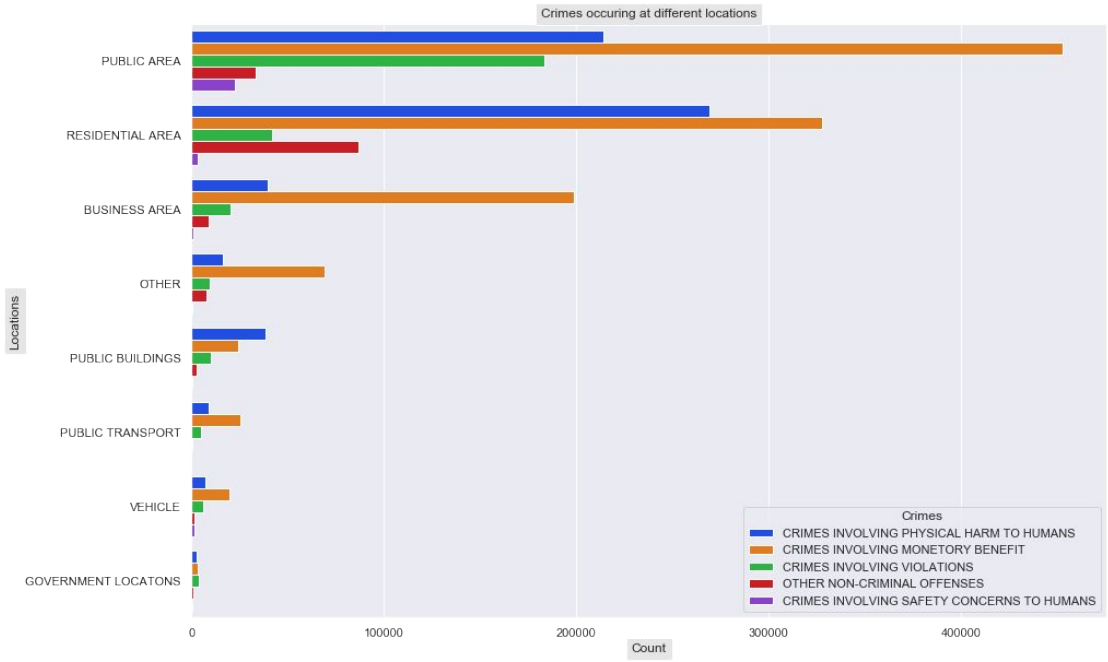
- Though only 25% criminals are caught.
- Checking further which crimes have the highest arrests.
- Criminals involved violations such as weapons violation ,public peace violation,narcotics etc are mostly caught.



V. Data Analysis

2. Location Based Analysis

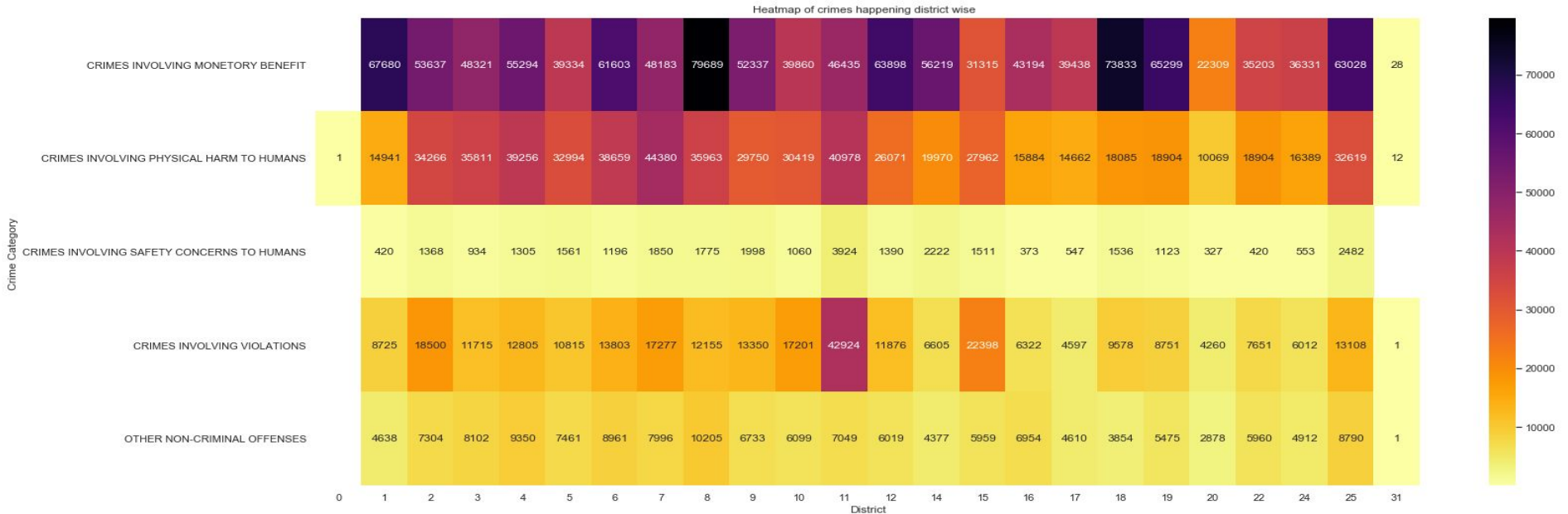
- Analysing the crimes happening at various locations.
- Public area and residential areas were the highest to have all categories of crime
- Also another observation was business areas had a spike in crimes with monetary benefits.
- This clearly indicates that crimes such as theft, motor vehicle theft etc are higher in business areas which involves locations like departmental stores, commercial offices, grocery stores etc.



V. Data Analysis

2. Location Based Analysis

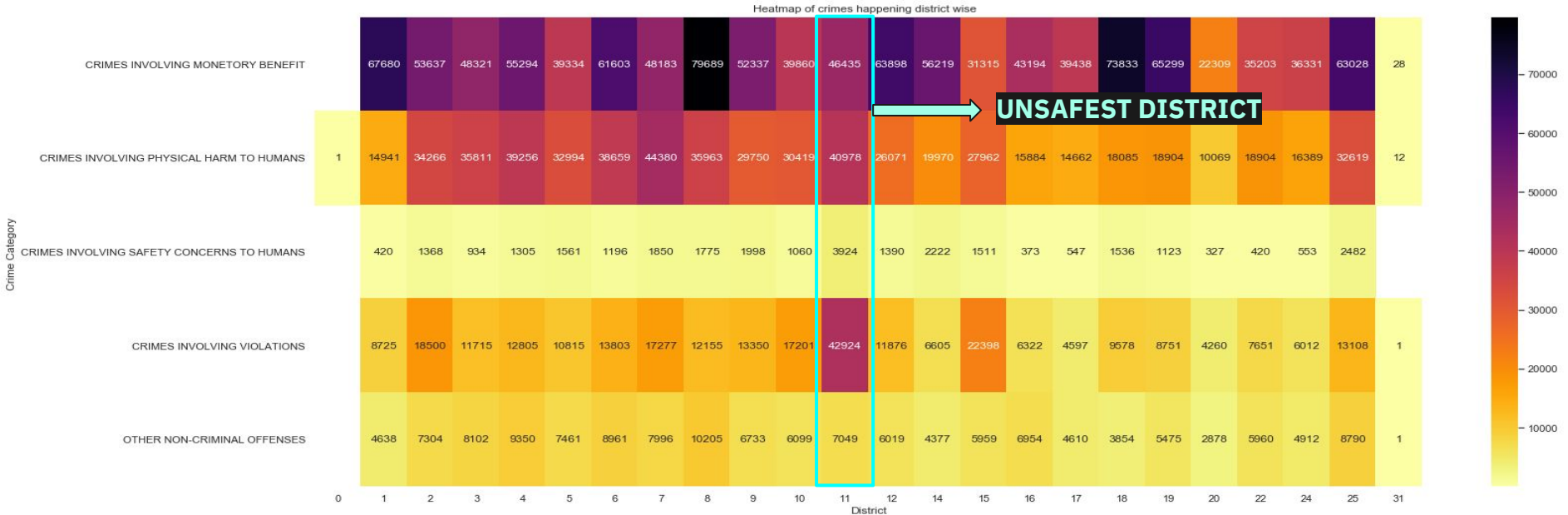
- From the below heatmap of the crimes happening in various districts, it can be observed that crimes involving monetary benefits is highest in all the districts with maximum happening in district 8 and district 18.
- Also the crimes involving physical harm to humans is second highest among all districts.



V. Data Analysis

2. Location Based Analysis

- We also observed that almost all categories of crime is higher in district 11.
- So this district can clearly be termed as **UNSAFEST DISTRICT.**

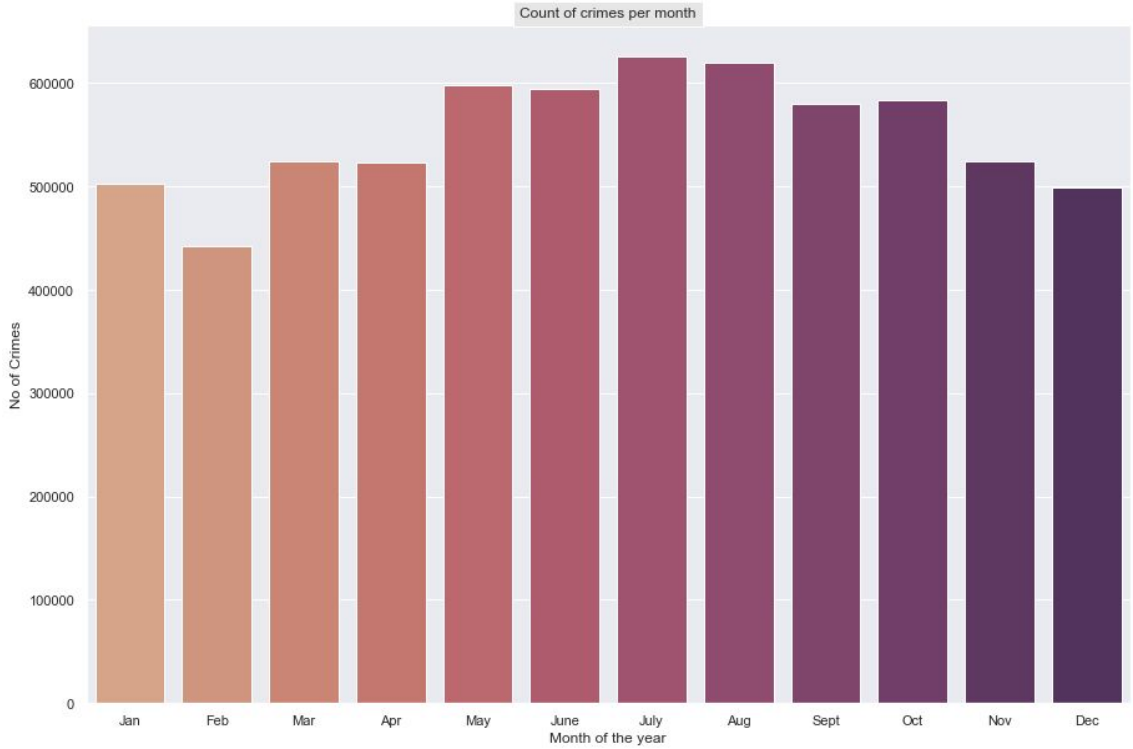


V. Data Analysis

3. Time Based Analysis

- Checking the crimes occurring month-wise.

- We observed that the count of crimes occurring rises in summer months i.e. June, July, August.
- We suspect this could be because summer months are the months when people go out on holidays and also tend to be out more than the winter months as Chicago has harsh winter weather.



V. Data Analysis

3. Time Based Analysis

- Checking the crimes occurring on hourly basis.

- We observed that the count of crimes is low in the early morning but has a spike in the peak afternoon when people tend to visit restaurants or stores to grab lunch
- The count also rises in the evening when people tend to visit more of the stores after work or are on the streets more.
- This count is higher in the night as well and with peak in the midnight.

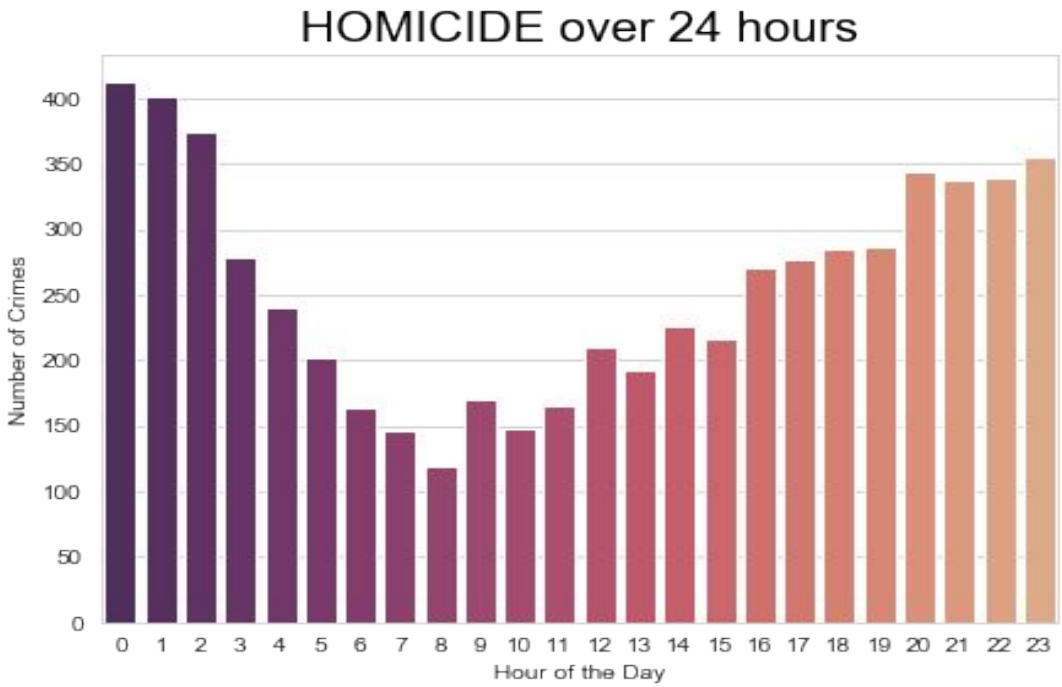
Unsafe Hours in Chicago

V. Data Analysis

4. Understanding the pattern of various crimes over 24 hrs

HOMICIDES

- Homicides have the peak at midnight and remaining up till
- 2am - 3 am.
- Homicide also rises in the evening after 7 pm.

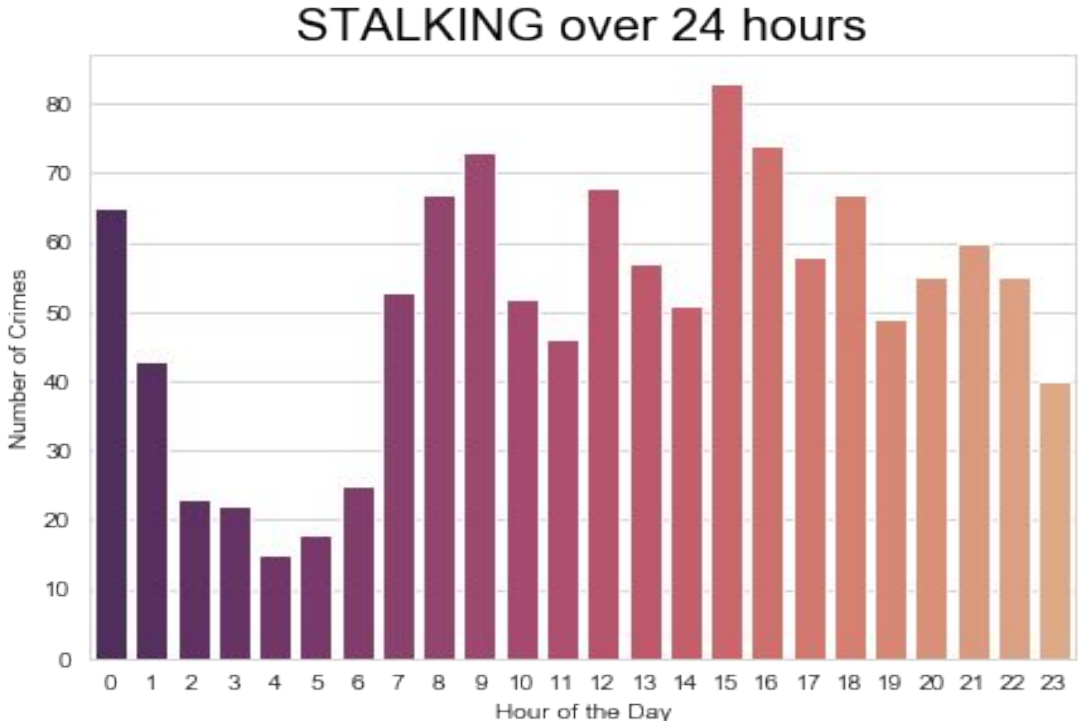


V. Data Analysis

4. Understanding the pattern of various crimes over 24 hrs

STALKING

- Stalking seems to start rising at 7 am.
- The trend seems to be high low at various times, and it tends to be high after 3pm till midnight.

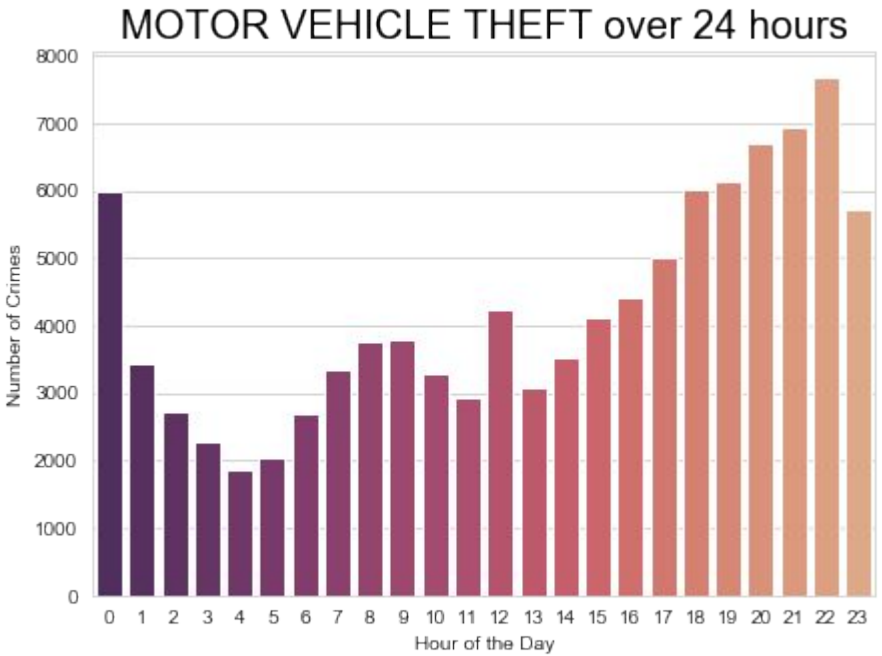


V. Data Analysis

4. Understanding the pattern of various crimes over 24 hrs

MOTOR VEHICLE THEFT

- This seems to occur mostly in the late evening hours.
- It happens highest at 11 am.
- We suspect it when people return back home and park, this seems to occur more.

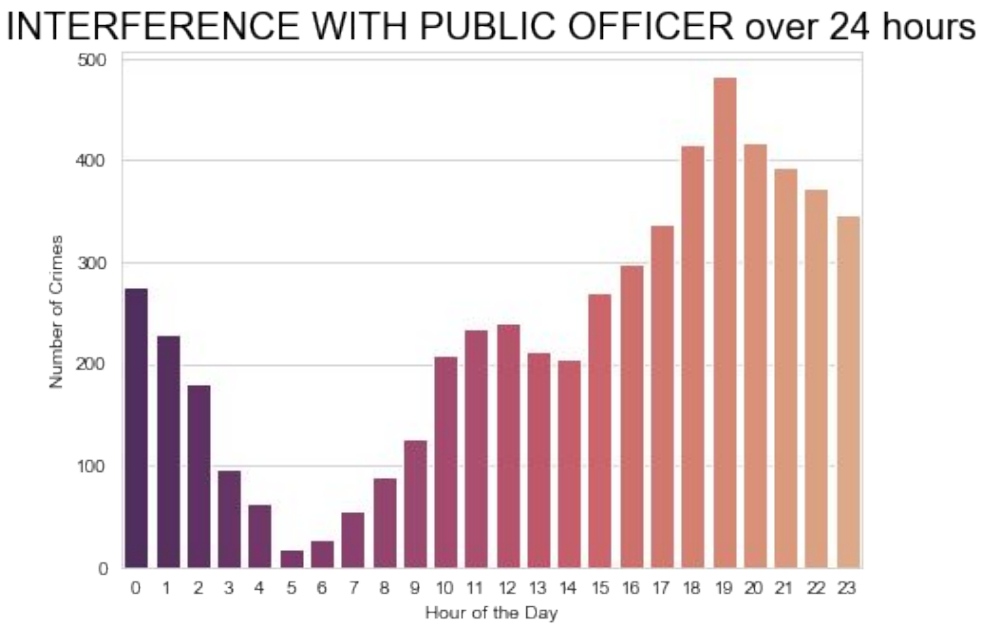


V. Data Analysis

4. Understanding the pattern of various crimes over 24 hrs

INTERFERENCE WITH PUBLIC OFFICERS

- This starts to rise in the morning between 8am -10 am.
- This has its peak at 7 pm.
- We suspect this is the time when most vehicles are on the streets.

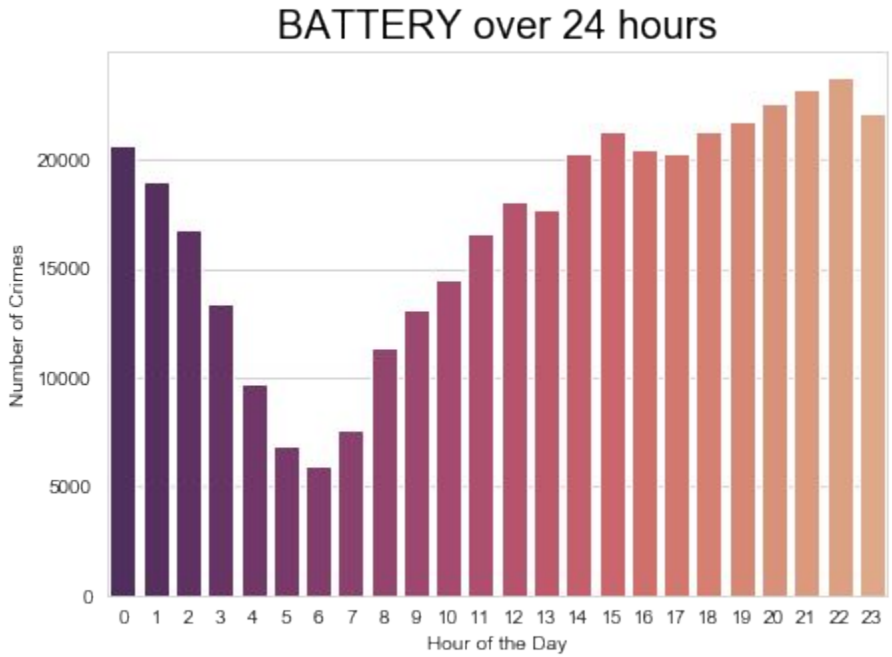


V. Data Analysis

4. Understanding the pattern of various crimes over 24 hrs

BATTERY

- This has a gradual rise from 6 AM to 10 PM.
- After 10 PM the crimes reduce till 6 AM.
- Like most crimes the battery rises after 6PM which is usually the time when most people are outdoors.



VI. Data Mining

- The data mining technique used was classification.
- The type of crime was split into severe and non severe techniques.
- Classification of severe vs non severe crimes was conducted with two datasets.

Class	Types of crimes
Severe	homicide, kidnapping, human trafficking, offense involving children, battery, crim sexual assault,sex offense, assault,criminal sexual assault, domestic violence, stalking, prostitution, intimidation, interference with public officer, obscenity, public indecency, arson
Non Severe	criminal damage, deceptive practice, burglary, motor vehicle theft, theft, robbery, gambling, weapons violation, concealed carry license violation, public peace violation, liquor law violation, narcotics, other narcotic violation, criminal trespass, non-criminal, non - criminal, non-criminal (subject specified), other offense, ritualism

VI. Data Mining

Attributes used for classification

- Crimes dataset without the hardship index
  Attributes which are used for classification:
  - Arrest -If an arrest was recorded or not
  - Domestic - If the case is considered domestic-related or not
  - District - Police district where the crime occurred
  - Ward - The ward where the crime occurred
  - Communityarea - The community where the crime occurred
  - Year - The year where the crime occurred
  - Latitude- The latitude of location where the crime occurred
  - Longitude - The longitude of location where the crime occurred
  - Hour - The hour in 24 hours at which the crime occurred
  - Month - The month at which the crime occurred
  - Severity- If the crime was severe or not (class on which the classification is performed)
  - Location - The location at which the crime occurred
  - Weekday - The day of the week in which the crime occurred
  - Day - The day of the month in which the crime occurred
- Crimes dataset with the hardship index
  Attributes which will be used in addition to above attributes:
  - Croudedhousing - Percentage of citizens living in a crowded home
  - Belowpoverty - Percentage of citizens who are below the poverty line
  - Unemployed - Percentage of citizens who are above 16 and below 64 who are unemployed
  - Uneducated - Percentage of citizens who do not have a diploma
  - Nonworkingage - Percentage of citizens who are above 16 and below 64

VI. Data Mining

Accuracies of different classification models on both the datasets

Model	With hardship index	Without hardship index
Naive Bayes	67.34%	77.56%
Random Forest	78.52%	78.51%
Decision Tree	68.73%	68.83%
Multi-layer Perceptron	77.53%	77.54%

The accuracies of the model clearly show that addition of hardship index did not provide any significant benefit towards identifying the severity of the crimes.

Future Work

Crimes in a place are not only caused by one factor but multiple factors some of them may be political or international. We plan to explore more events that might have triggered the crimes in those area. We also plan to further elaborate our study into other states and explore if there are any patterns of crimes that exist across multiple states. Crimes or violations in one state can very often spread across the rest of the country fast. The correlation of crimes among different neighboring and non-neighboring states is something which we plan to study further.

Conclusion

- We improved the accuracy of the model by 8% after imputing the missing values in the model with accurate values from the another dataset.

- Although we couldn't find any direct correlation between hardship index of the are with the crimes there are several important patterns we noticed while working with the data.

- The importance of data cleaning and preparation step in the process of data mining is very important as we noticed proper dimensionality reduction and removing unnecessary attributes can improve the speed of the data mining process.

- Even though throughout this project we criticize the Chicago city and it's safety record we understand that prevention of crime is not an easy task, if was easy everyone would do it. We would like to convey our respect for the police department of Chicago. As without them the project of this wouldn't have been possible.

References

1. Chicago Police Department. Crimes - 2001 to present https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.%20Accessed:2021-02-19

2. 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

3. City of Chicago. Boundaries - community areas. https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6.Accessed: 2021-02-19

4. City Of Chicago. Chicago data portal. "https://data.cityofchicago.org". Accessed:2021-02-19.

5. Title page background image - https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6

6. Lobonţ, O. R., Nicolescu, A. C., Moldovan, N. C., & Kuloğlu, A. (2017). The effect of socioeconomic factors on crime rates in Romania: A macro-level analysis. Economic Research-Ekonomska Istrazivanja, 30(1), 91–111. doi:10.1080/1331677X.2017.1305790

THANK YOU