Nile University

CSCI322: Data Analysis

Professor: Mustafa A. Elattar

Chicago Crime Data Analysis & Visualization

Fall 2019

Submitted By:

Ahmed Samir 1510217

Al Shaimaa Samir 1610758

Mohamed Ayman 1510162

i. Contents

ii. List of Figures

iii. Introduction

Crime analysis has become one of the most important topics in the field of data science, since the availability of crime data sets, in addition to the need of analyzing the reasons of those crime activities and the prediction of future crimes, according to the past criminal records. A lot of online crime datasets are available nowadays, like the city of Atlanta Crime 2009-2017 dataset, city of Baltimore Crime 2011-2016 dataset and many other datasets that can be found here: https://data.world/datasets/crime. Thus, the developers decided to take part in such an intriguing field, which is still open for analysis and research. Moreover, the selection choice of a specific crime sector, needs a decent amount of investigation and research. In addition to a comparison of these various sectors will help the developers find a specific preference.

iv. Criminology

As mentioned in the previous section, studying crimes and their causes needs a lot of research. Moreover, a more scientific name is Criminology, which is study of crimes and their underlying factors. Furthermore, Criminology also studies the procedures to prevent these crimes from their roots. In addition to try to detect patterns in these incident and setup strategies that can work as an early prevention from them. Some of the specializations of Criminology are mentioned in Figure (1) and through our project we mainly focus on Criminal Statistics, try to correlate factor to relate it to the Theory Construction and Victimology.

Criminal Statistics	Patterns and trends of criminal behavior
Sociology of Law	Legal Construction and CJ Systems
Theory Construction	Why People Do/Not Commit Crime
Criminal Behavior Systems	Nature and Cause of Specific Crime Types
Penology	Sanction Types & Effectiveness
Victimology	Victim Behavior Patterns & Treatment

Figure 1: Criminology Fields

v. Chicago Crimes Dataset

Since adapting the idea of analyzing crimes and extracting useful insights, the developers had to find a specific sector to focus on. Thus, the chosen sector was crimes in the city of Chicago, which has been the focus of most news channels recently. They even consider it the most gang-infested country in America. Moreover, the city of Chicago and more specifically, Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system has extracted all the criminal data from 2001 to the present day, minus the most recent 7 days. The dataset contains records for various crimes in the last two decades. The dataset includes various crime types, for the exception of murder. Moreover, the Chicago Police department as started, they do not guarantee:

- Accuracy
- Correct Sequencing
- Completeness
- Timeliness

Figure 2: Chicago Police Department Logo

Thus, the dataset might have several flaws and should be preprocessed and cleaned before any analysis of the given data. Furthermore, the dataset includes different attributes to describe each crime record. The main attributes in the dataset are:

- **ID**: Unique identifier for each record.
- **Date:** Date when the crime/incident happened → close estimate.
- **Block**: Estimate block address where the incident happened, the real address of the incident was omitted for privacy concerns.
- **Primary Type**: The primary crime type, which is basically a general name for the type of crime committed. For example: (Theft, Battery and other offences).
- **Description**: Which states the details of the primary type crimes. For example: (Theft: FINANCIAL ID THEFT: OVER $600).
- **Location Description**: The description of the location where the event happened.
- **Arrest**: Indicates if for the given record, an arrest has been made or not.
- **Community Area**: Specifies which community area has this incident happened in. The city of Chicago has exactly 77 community areas.
- **X Coordinate**: The x coordinate of where the event happened in the State Plane.
- **Y Coordinate**: The y coordinate of where the event happened in the State Plane.
- **Domestic**: Specifics if the incident is domestic related or not.
- **Location**: The location of where the incident taken place. It is formatted to be used in maps creation and geographic operations. It consists of a pair which is:
    - **Longitude**: The longitude of the location where the incident happened.
    - **Latitude**: The latitude of the location where the incident happened.

vi. Objective

The objective of the project is to answer some questions regarding criminology. Moreover, answering these questions and reaching a well-established analysis, will give us a chance to use these approaches with crimes in other cities also. The developers aim to explore the dataset and find key points that can help in detecting the major areas of crime, factors behind these crimes and safe spots around the city. Moreover, we explore and answer more questions regarding performance indicators of the police in the city and the correlation between socioeconomic factors and crimes in the city.

vii. Hypothesis

The developers propose that the factors combined from different aspect, should highlight interesting facts about them and the crime rate. Moreover, the exploration of districts in the City of Chicago will give us a better view on where exactly most of the crimes happen and might conclude locations where they are safe enough to stay in. In addition to investigating the locations where the crimes take place and the relation between this place and how they correlate with the living standards of the block.

viii. Workflow Pipeline

The work flow of a data analysis program is typically described of a pipeline of stages and each stage of the pipeline is responsible for a specific task. Moreover, the Chicago crime analysis program is like the data science pipeline. This pipeline has a simple acronym, which is O.S.E.M.N.

a. Obtaining Data

This is the first stage of the pipeline and it makes sense that the first objective is to fetch the required dataset. Furthermore, the identification of the dataset source and the downloading procedure is the focus of this data. Thus, the obtained data is stored in the project's directory in a suitable format, to be read and manipulated with ease.

b. Scrubbing Data

This stage is very crucial in our pipeline, to obtain well typed data with no anomalies or wrong data. Moreover, through this stage every missing value is filled with a suitable value, the records with wrong data are dropped, checking errors generated by this dataset and removing unused attributes, which add to the dimensionality of the current data frame.

c. EDA

Exploratory data analysis -abbreviated EDA- is where we are recognizing patterns in our data. Furthermore, finding correlation between various attributes and if there are any compound attributes that could enhance the correlation. Additionally, visualizations by all sorts and techniques are used to help provide beneficial visuals that will help us identify more meanings and relations in our data. This stage uses statistical measures also to test correlations between variables (Chi Square, Correlation Coefficient, Pearson correlation).

d. Modeling

After extracting insights and patterns from our data, we then need to predict futuristic events. Moreover, this stage is responsible for enhancing the **decision making** by far. In the Chicago crime analysis project, a predictive model can help in forecasting crimes futuristically and upon that prediction, a counter measure could be taken. This stage will be embedded in our project soon.

e. Interpreting

This stage is all about explaining our findings. Moreover, visualizations are a key method for providing the findings in a neat and understandable manner. The interpretation of the data from the Chicago crimes analysis are presented in our results section with several graphs and plot that shows the relation mentioned earlier in our hypothesis. Moreover, the interpretation of the data and suggestions are quite astonishing.

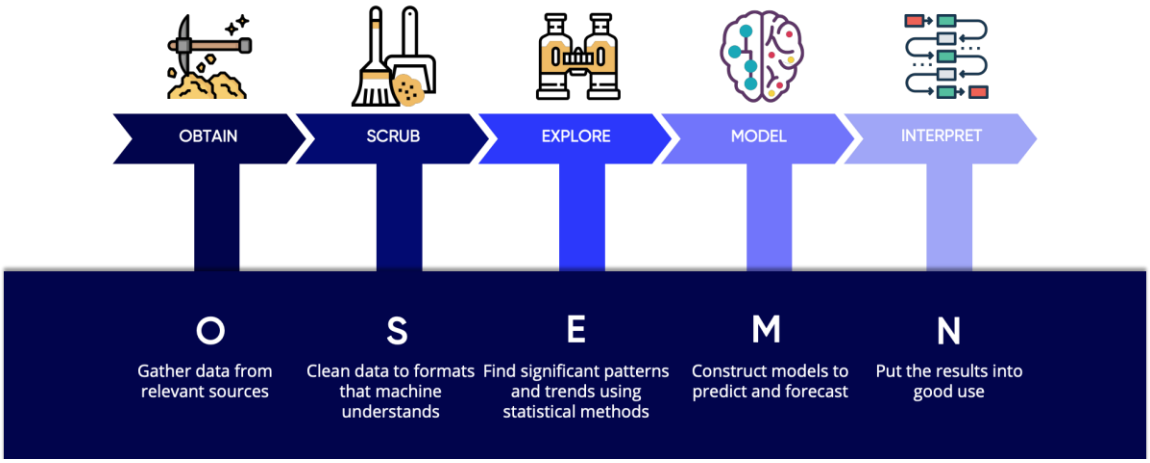

Figure 3: O.S.E.M.N Pipeline

ix. Results

a.Number of Crimes

We were trying to validate the assumption that the total number of crimes is decreasing by time. For this purpose, we use the information of the number of crimes reported and rolling them on scale of 365. The below figure the Graph shows an obvious decrease in the crimes in the time between 2001 and 2016.

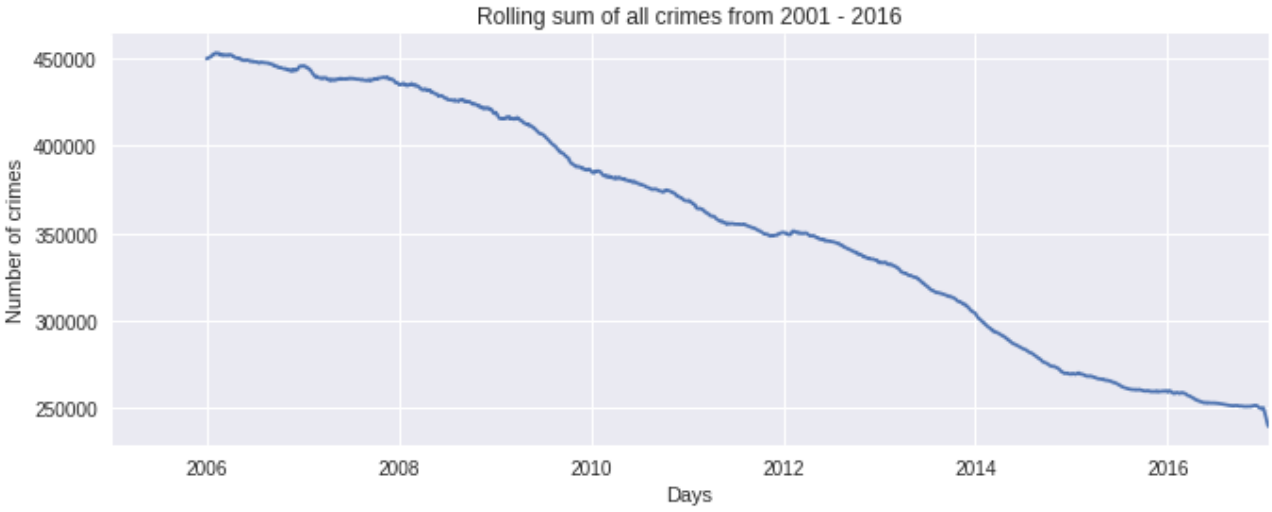

Figure 4: Rolling Sum of Crimes (2001 ->2016)

b. Investigating the performance of the Police station

Investigating the performance of the Police station yielded an interesting result. Moreover, the below graph shows the number of reported cases and the number of cases at which arrest has really occurred. We assumed that the percentage of the cases at which the police could arrest the suspect with respect to the total number of cases is an indicator for the police performance.

As can be noted the total number of crimes and arrests both decreases, but arresting cases decrease by a larger factor. This indicate that the police performance is degrading with respect to arresting the suspects of the cases.

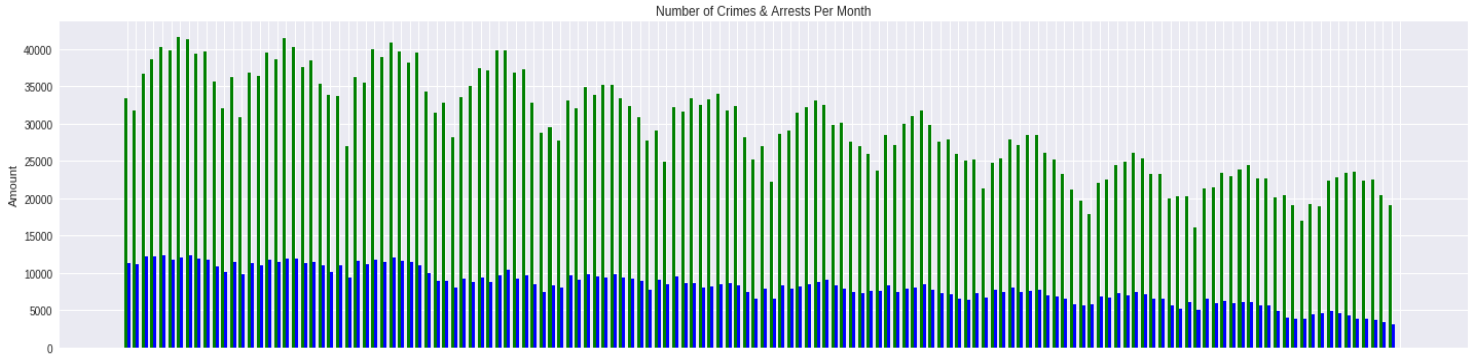

Figure 5: Number of Crimes & Arrests per month

Studying the Pearson correlation coefficient between the number of crimes and some socioeconomic Indicators. We will use another dataset from Kaggle that has the socioeconomic indicators of Chicago, listed by community areas. We will explore the correlation between different indicators and number of crimes.

More visualization and interesting results can be found in the detailed in the notebook

c. Crimes Per Location

This graph visualizes the density of crime better. It shows the map of Chicago, and the color codes indicate which crime type is most common in this area of Chicago.

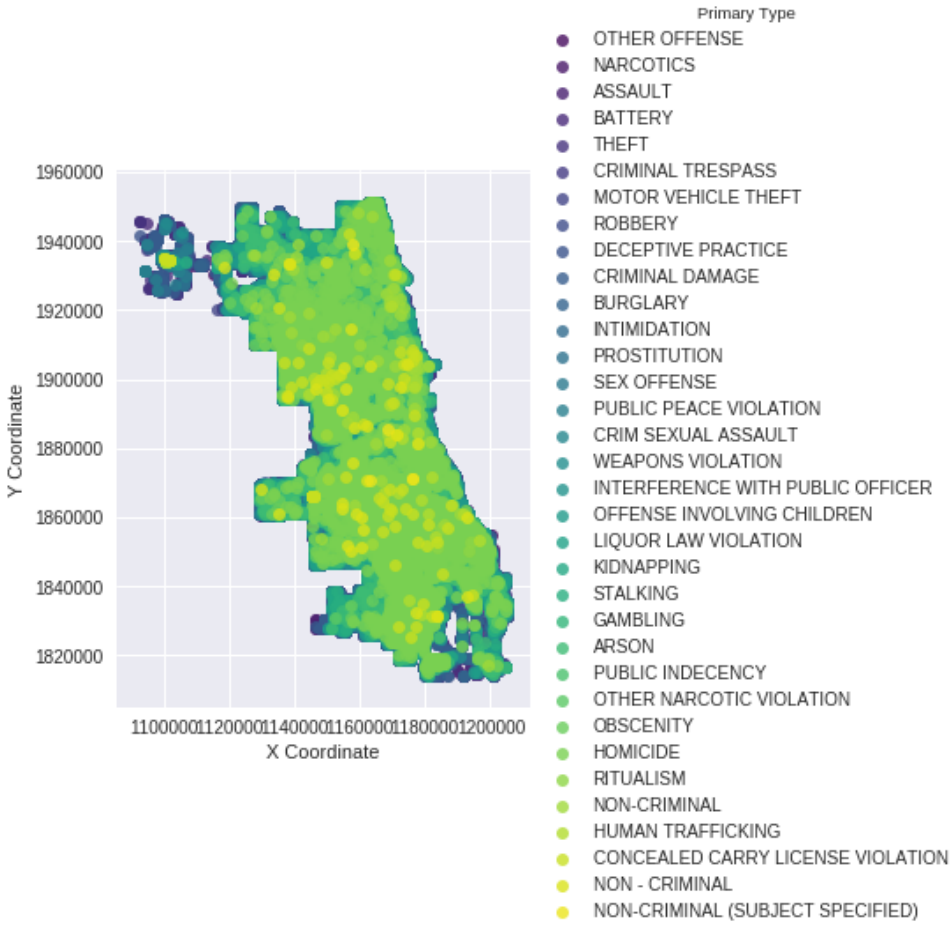

Figure 6: Locations of each Crime

d. Correlation between Percent of Housing Crowded and CRIME Number

The correlation between crowd level of the houses and the Crime numbers was around 0.07 which show that those two variables are not correlated.

The correlation between Crime numbers and multiple other factors were also studies in detailed in the notebook and the factor which showed the most correlation was **the**

percent of house hold below poverty. The correlation was about 0.230. this indicate that the socioeconomic factors don't affect the number of crimes.

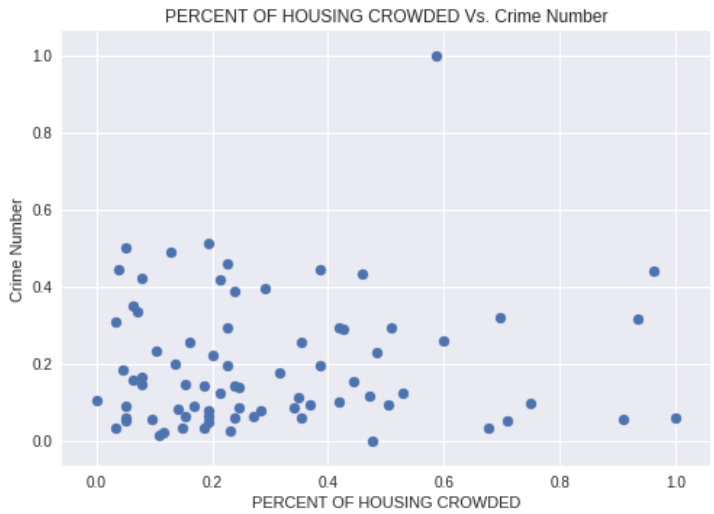

Figure 7: Percentage of HC vs Crime Number

We tried to make this section as short as possible because everything is explained in a well formatted form in the notebook. Please refer to the notebook.

x. Future Work

The exploration of crime data is never over, and endless factors can still be explored. Moreover, the developers set some future work that could be embedded to their project soon. As mentioned earlier, the developers tried different method to evaluate the Chicago police department performance, through extracting the Arrest rate, which is a new attribute that evaluates the number of arrests made per month, relative to the total number of reported crimes in Chicago. Moreover, a module will be added to calculate the real time needed to move from the nearest police station to the crime location, by using time calculations from the google maps API. In addition to trying to minimize the time between the police department and location, by suggesting more beats (moving police cars) in closer locations, which will help increase the arrest rate optimistically. Given the O.S.E.M.N pipeline mentioned earlier, the modeling stage was not implemented yet in the current project version. Thus, the insertion of the prediction model, which will help in predicting crimes in the future by completing data pattern from the previous criminal records data.

Figure 8: Future Modules

xi. Conclusion

In conclusion, exploring Chicago Crimes, police stations in Chicago and Socioeconomic indicators data sets we concluded a set of observations. Firstly, we explored police performance and number of arrests. Moreover, we used clustering to find the position in need for police station increase. Another point to explore was the correlation between crime rates and other factors. Lastly, we showed the density of each crime on the map of Chicago and the what is the ration of each type of crime with respect to each other. Those insights are very informative for both individuals and organization to monitor the state security and come up with solutions for problems that can be found through the research. For individuals, it can demonstrate the general security state of the Chicago along with a safety indicator of each point in the map.

Figure 9: Crimes Decrease

xii.References

[1] R. Lao, "A Beginner's Guide to the Data Science Pipeline," towardsdatascience, 16 January 2018. [Online]. Available: https://towardsdatascience.com/a-beginners-guide-to-the-data-science-pipeline-a4904b2d8ad3. [Accessed 3 january 2020].

[2] C. D. Portal, "Boundaries - Police Districts (current)," Chicago Data Portal, [Online]. Available: https://data.cityofchicago.org/Public-Safety/Boundaries-Police-Districts-current-/fthy-xz3r. [Accessed 3 January 2020].

[3] C. P. Department, Chicago Data Portal, [Online]. Available: https://data.cityofchicago.org/Public-Safety/Crimes-2019/w98m-zvie. [Accessed 3 Januart 2020].