# Analysing crime dataset of Chicago

**Sonali Rajhans**

**June , 2019**

# Table of Contents

•**Introduction**.....

• **Data**

i. Data Acquisition.....

ii. Data Cleaning and Analysis.....

iii. Data Processing.....

iv. Predictive Models.....

- **Results and Discussion**.....
- **Conclusion..**.....
- **Future work**.....

## Table of Figures

**fig 1:** A map showing the location of various police department in each district .

**fig 2:** relation between each district and the number of recorded cases in 2018 for each of them

**fig 3:** Clustering to check relationship between District and Type of Crimes

**fig 4:** relationship between the type of crimes that exist in each district

**fig 5:** top 10 crimes in each district

**fig 6:** Provide relationship between the Location Type of a crime and the districts

# 1 . Introduction

### 1.1 Background

In this project we analysed various distribution of crimes in the Chicago city area. The city being divided into 23 districts became the key driver of the distribution . As we can see there is a strong co-relation of the clusters along with its primary type as well as location of the crime that is taking place . In this project we analysed various distribution of crimes in the Chicago city area. The city being divided into 23 districts became the key driver of the distribution . Today, crime rate is a menace that each country faces. It is said that society has a direct influence in making criminals. Government imposed many laws to reduce crime rate to make world a better place to live in, but majority did not find expected results.

### 1.2 Problem

Violent crimes, thefts, and robberies occur in majority of places. On daily basis, local businesses, houses etc encounter such crimes. Our job as data scientist is to analyse the pattern of crimes. study the distribution of types of crime that occur in various part of a particular place. It is important to know where the crimes are happening, usually areas with police stations farther in distance can result into increase crime count . In this case, I decided to perform statistical analysis and data science on the crimes in the city of Chicago . We will also compare the location of local police station in the area, and their relation with the location of the crimes.

### 1.3 Interest

Obviously , I am interested in predicting the crime rate in Chicago as well as correspondent to that the percentage of cops rate in the city.

# 2. Data Acquistion and Cleaning

### 2.1 Data Sources

I am going to use two different datasets , with one referring to crime rate and the other one referring to cops rate presence in that area , both are present in the csv (comma separated value ) file format .Both of the files are being provided by Chicago Data Portal.

Furthermore we will use the Foursquare API to fetch the venues to see in what specific type of environment or location does these crimes take place.

### 2.2 Data Cleaning and Analysing

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping . There were many such unwanted values which were not at all needed for the completion of project hence , using drop method all those values were dropped from the new data frame.

The crime data set consist of reported incidents of crime that occurred in the City of Chicago from 2018, The particular data set is subset of a bigger crime data with over 6.9 Million fields which would require greater computation power.. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified

| | Case Number | Date | Primary Type | Location Description | District | Community Area | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|
| 1 | JC104662 | 12/31/2018 11:59:00 PM | CRIMINAL DAMAGE | STREET | 22 | 74.0 | 41.689079 | -87.696064 |
| 2 | JC100043 | 12/31/2018 11:57:00 PM | CRIMINAL DAMAGE | APARTMENT | 6 | 71.0 | 41.740521 | -87.647391 |
| 3 | JC100006 | 12/31/2018 11:56:00 PM | BATTERY | OTHER | 12 | 31.0 | 41.857068 | -87.657625 |
| 4 | JC100031 | 12/31/2018 11:55:00 PM | BATTERY | APARTMENT | 6 | 71.0 | 41.751914 | -87.647717 |
| 5 | JC100026 | 12/31/2018 11:49:00 PM | BATTERY | STREET | 15 | 25.0 | 41.875684 | -87.760479 |
| 6 | JC100011 | 12/31/2018 11:48:00 PM | BATTERY | APARTMENT | 6 | 71.0 | 41.750154 | -87.661009 |
| 7 | JC100089 | 12/31/2018 11:47:00 PM | BATTERY | VEHICLE - OTHER RIDE SHARE SERVICE (E.G., UBER... | 19 | 5.0 | 41.939625 | -87.673996 |
| 8 | JC101094 | 12/31/2018 11:45:00 PM | THEFT | BAR OR TAVERN | 19 | 6.0 | 41.940519 | -87.654124 |
| 9 | JC101652 | 12/31/2018 11:45:00 PM | CRIMINAL DAMAGE | APARTMENT | 14 | 23.0 | 41.905562 | -87.707589 |
| 10 | JB574407 | 12/31/2018 11:44:00 PM | CRIMINAL TRESPASS | MOVIE HOUSE/THEATER | 19 | 3.0 | 41.968463 | -87.659670 |

crime data after cleaning

The dataset consist of 22 fields and over 226K rows. The fields include attributes related to crime such as Case number, block, IUCR codes, Description, Type of Crime, location, Lattitude, Longitude, district . Since there are many fields that are not required in our analysis, we will require a lot of cleaning to do.

### 3. Data Processing

relation between each district and the number of recorded cases in 2018 for each of them . it can be clearly be seen that district no. 12 of Chicago has the highest rate of occurrence of crime ,according to the records of 2018 .

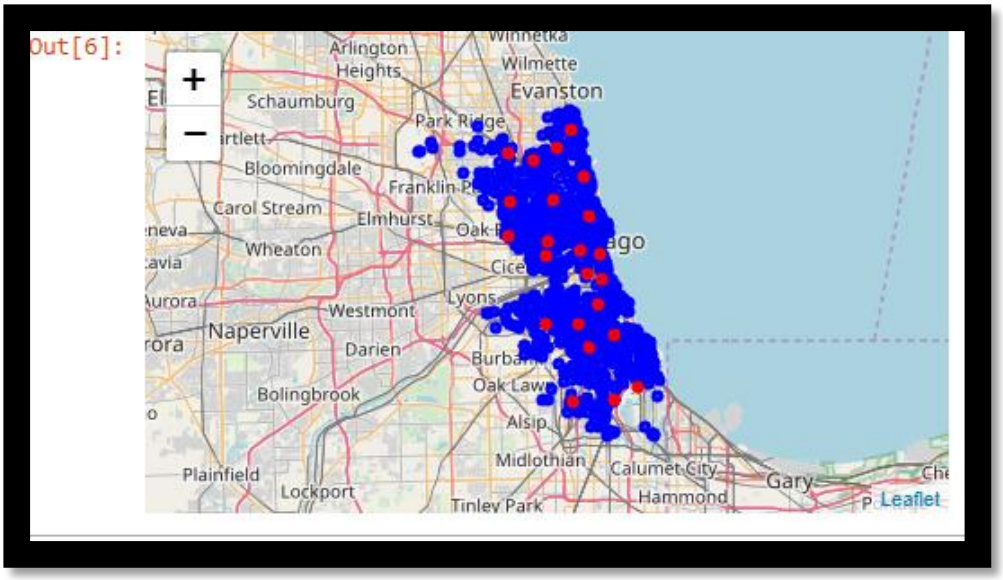
Getting top 10 crimes in each district

| | District | 1st Most Common Crime | 2nd Most Common Crime | 3rd Most Common Crime | 4th Most Common Crime | 5th Most Common Crime | 6th Most Common Crime | 7th Most Common Crime | 8th Most Common Crime | 9th Most Common Crime | 10th Most Common Crime | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | THEFT | DECEPTIVE PRACTICE | BATTERY | CRIMINAL DAMAGE | ASSAULT | OTHER OFFENSE | ROBBERY | CRIMINAL TRESPASS | MOTOR VEHICLE THEFT | BURGLARY | N |
| 1 | 2 | THEFT | BATTERY | CRIMINAL DAMAGE | ASSAULT | OTHER OFFENSE | DECEPTIVE PRACTICE | ROBBERY | MOTOR VEHICLE THEFT | BURGLARY | CRIMINAL TRESPASS | N |
| 2 | 3 | BATTERY | THEFT | CRIMINAL DAMAGE | ASSAULT | OTHER OFFENSE | BURGLARY | DECEPTIVE PRACTICE | ROBBERY | NARCOTICS | MOTOR VEHICLE THEFT | |
| 3 | 4 | BATTERY | THEFT | CRIMINAL DAMAGE | ASSAULT | OTHER OFFENSE | BURGLARY | DECEPTIVE PRACTICE | MOTOR VEHICLE THEFT | NARCOTICS | ROBBERY | |
| 4 | 5 | BATTERY | THEFT | CRIMINAL DAMAGE | OTHER OFFENSE | ASSAULT | BURGLARY | DECEPTIVE PRACTICE | NARCOTICS | WEAPONS VIOLATION | ROBBERY | |
| 5 | 6 | BATTERY | THEFT | CRIMINAL DAMAGE | ASSAULT | OTHER OFFENSE | NARCOTICS | DECEPTIVE PRACTICE | BURGLARY | ROBBERY | WEAPONS VIOLATION | |
| 6 | 7 | BATTERY | THEFT | CRIMINAL DAMAGE | ASSAULT | OTHER OFFENSE | NARCOTICS | WEAPONS VIOLATION | BURGLARY | ROBBERY | DECEPTIVE PRACTICE | |

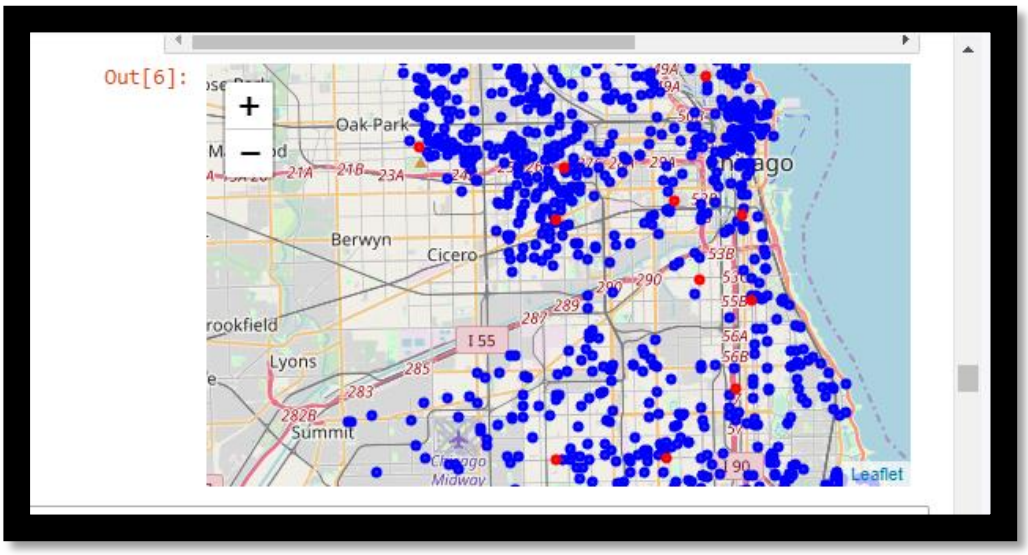
## 4. Predictive Modeling

There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can provide additional information on the amount of improvement, while classification models focus on the probabilities a player might improve. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other.

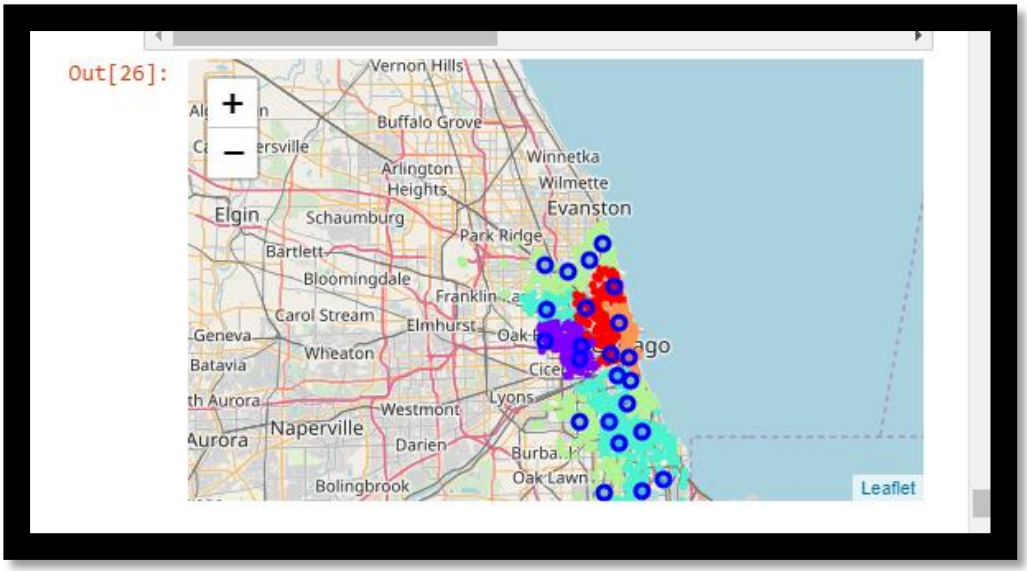
4.1


Out[6]:
Arlington Heights
Wilmette
Evanston
Schaumburg
Park Ridge
Bloomingdale
Carol Stream
Elmhurst
Wheaton
Naperville
Westmont
Darien
Lyons
Bolingbrook
Oak Lawn
Alsip
Midlothian
Calumet City
Gary
Hammond
Plainfield
Lockport
Tinley Park
Leaflet

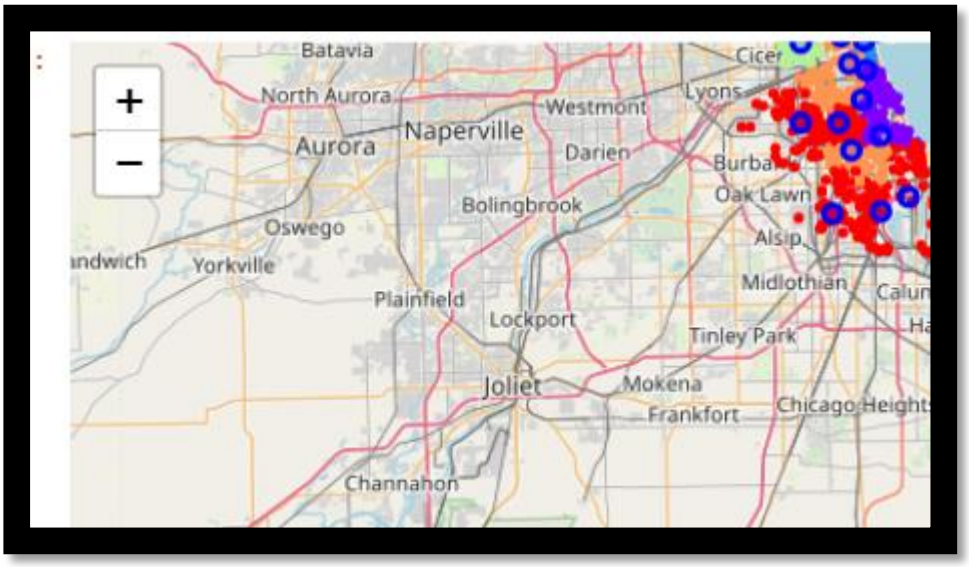
4.2


Out[6]:
Oak Park
Berwyn
Cicero
Lyons
Summit
I 55
I 90
Leaflet

4.3




4.4




## Provide relationship between the Location Type of a crime and the districts

each district is different and similar in terms of different types of building that are in the area. Some are residential area, some are commercial area, this clustering will give us insight on how the crimes distributes itself based on the location.

## 5. Conclusions

In this project we analysed various distribution of crimes in the Chicago city area. The city being divided into 23 districts became the key driver of the distribution.

As we can see there is a strong co-relation of the clusters along with its primary type as well as location of the crime that is taking place

In this study , I analysed the relationship between crime types and the districts . I identified the models predicting the highly populated and crime affected areas also the areas involving maximum police stations. I believe that having greater number of can enable for complex data science such as density based clustering and segmentation.

Further more , it will help people in :

1. Avoid unnessacary outings

2.Avoid travelling those places of Chicago which are not crime prone.

## 6.Future Work

Include more dependent variables and collect more data sets to create more accurate predictive model .