# **Chicago Crime Detection using Predictive Analysis**

## **Introduction**

## **Background**

### Chicago’s overall crime rate, especially the [violent crime](https://en.wikipedia.org/wiki/Violent_crime) rate, is higher than the US average. [Chicago](https://en.wikipedia.org/wiki/Chicago) was responsible for nearly half of 2016's increase in homicides in the US. Crime in Chicago has been tracked by the [Chicago Police Department](https://en.wikipedia.org/wiki/Chicago_Police_Department)'s Bureau of Records since the beginning of the 20th century. In order to protect the city, Chicago police department has started using artificial intelligence to predict where crimes will happen before they takes place.

## **Business understanding**

Chicago crime data is available on the police department data portal. Existing crime data can be correlated with relevant location data based on the zip code or ward number to detect the patterns of crime behavior. Chicago Police Department has detailed crime data stored in their website. This data contains many variables that can be used for prediction analysis.

## **Defining the target variable**

The target variables for this prediction are the type of crime, the location of crime and the time of crime. These three target variables are used to detect when and where the next crime could happen.

## **Data Understanding**

### The dataset is huge with 1048500 rows of data. Each data row consists of Case number, street address, and primary crime type, description of the crime, FBI code, district, beat, ward, community information, and other related values. The location description gives whether the location is a business, apartment, residence, hotel or sidewalk. The Latitude and longitude information is not available for records prior to 2019. There are some additional fields such as case number and ID which may not be used for this study.

## **Data preparation and Analysis**

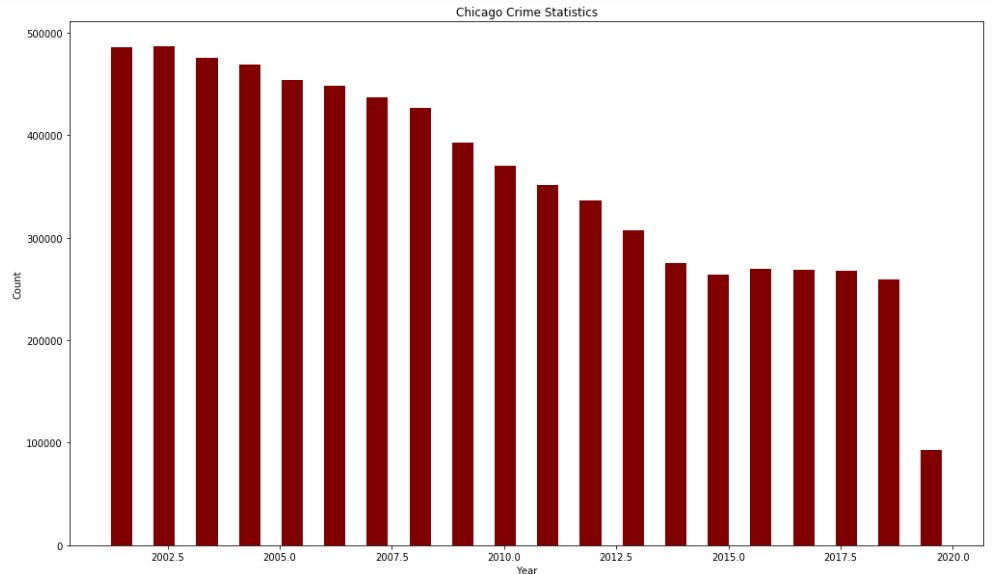
The source of the data:

https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

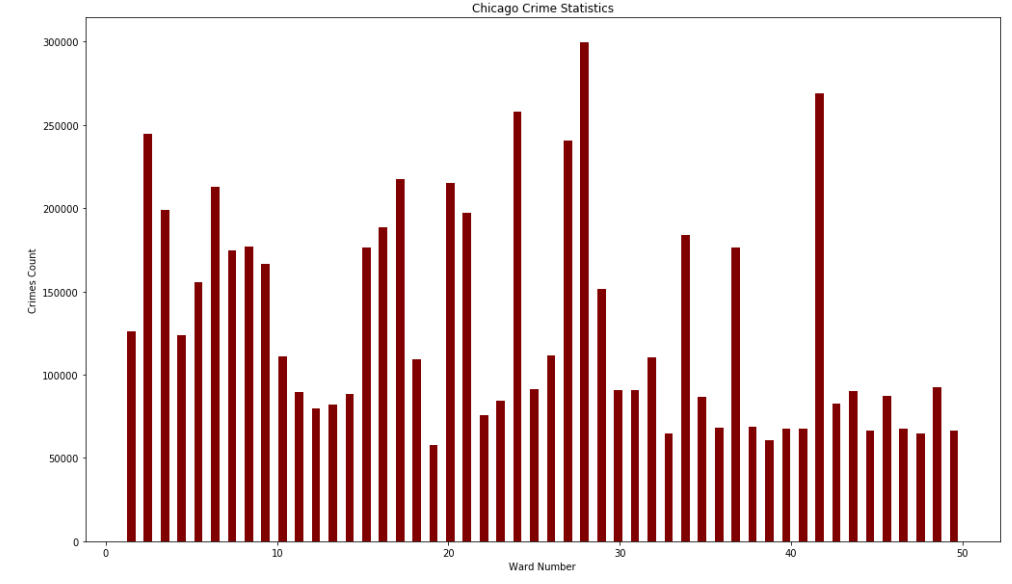
This dataset has 6886 NULL values in location description. Many NULL values are alsio found in the ward and community area variables. Upon further research, it is found that ward data is not available for records of years 2001 and 2002. For this study, we can ignore this data as it is too old to include in the model. Location data is also not available for these older records. Data is imported and cleaned. Old data records (< 2003) are removed where the ward data and community information is not available. Around 6000 of 6800 records where the location information is NULL are marked as fraud or forgery as the crime description. So the physical location of these records can be updated as ONLINE and keep the records in the model development. Any type of duplicate information can also be deleted from the original dataset.

Histograms are developed to determine the number of crimes per year, per crime type and per ward. Also a scatter plot is developed to see the location of different crimes per ward.

1. Crimes per Year



1. Crimes per Ward



### **Modeling:**

Graphs and charts can be developed targeting the following questions.

* The number of crimes happened in last 17 years
* How the number of certain crime types changed over last 17 years?
* What time of the day had maximum number of crimes occurred?
* Are there any correlation between different crime types and location?
* Are the time of crime occurrences and crime types are correlated?
* Have similar crimes occurred in different location at the same time?
* What is the effect of crime type and time on the location of crime?

K-means clustering can be applied to this dataset to classify incidents based on the location, time, ward, and crime type. The grouping is done by minimizing the sum of squares of distances between the data and corresponding cluster centroids. This method will determine the boundary or nearest location where incidents occurred over a period. Also, the time of occurrences of different incidents in a particular location can be determined by using the same method. The accuracy of k-means clustering can be found by Silhouette coefficient and the number of clusters can be found using Elbow method. Linear regression can be done to find out the correlation between location, ward, type of crime and time of incidents with location as the target variable. Twenty percent of the data can be used for the training and remaining data can be used for testing the model.

### **Summary and Conclusion:**

### After the models are trained and tested the time and location of frequent crimes can be determined. Also, the frequently occurring crimes can be found out using the model. These values then can be used to find the location and time of future incidents. Since the geolocation is available for recent records, the regression models can be run with the data from years 2018, 2019, and 2020 to find the target geolocation values, where future crimes will likely happen.

### **References:**

1 https://en.wikipedia.org/wiki/Crime\_in\_Chicago

2https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2