**Data acquisition and Cleaning**

2.1 Data Acquisition

The data acquired for this project is from three sources. The first data source uses a Chicago crime that shows all the crimes from 2001 till 2020 in Chicago. The dataset contains the following columns:

1. ID

2. Case number

3. Date

4. Primary type

5. Beat

6. District

7. Community area

8. FBI code

The second source is scraped from a Wikipedia page that contains the list of community areas. This page also contains additional information about each community area.

1. Serial number

2. Name

3. Area (sq. kms)

4. 2017 population density (per square km)

The third data source is a list of all the neighbourhoods in each community area as found on a Wikipedia page. The dataset is created from scratch using the data from the website. The following are the columns:

1. Neighbourhood

2. Community area

3. Latitude

4. Longitude

## 2.2 Data Cleaning

The data preparation for each of the three sources was done separately. From the crime in Chicago dataset the crimes of year 2017 are only selected.

The second data is scraped from a Wikipedia page .We make sure that the community area numbers are same so that we can merge the dataframes using these numbers(1-77). To identify the community area with the least crimes in the year 2017.

After visualising the crime in each community area we can find the community with the least crime rate and hence tag the safest community area.

The third source of data is created from the list of neighbourhoods from scratch using the list available on Wikipedia. The new dataset is used to generate the venues for each neighbourhood using the Foursquare API.