# Introduction

A friend of mine is extremely worried about violence. He is moving to another city, which does not provide statistics about the crime rate in its neighborhoods, and is wondering how he will be sure to choose the least violent neighborhood.

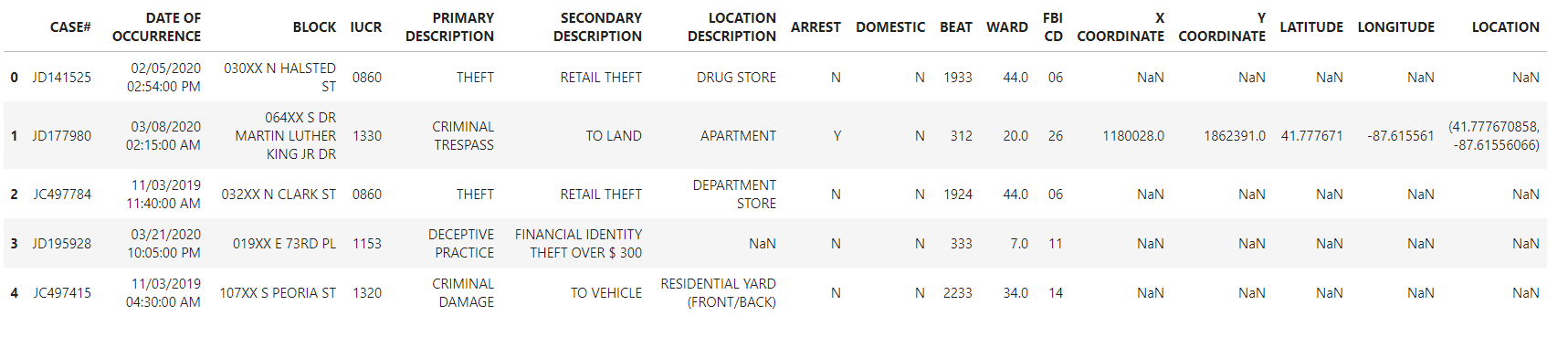
I had an idea. Based on the crime rate of the neighborhoods in the city of Chicago, I will develop a crime index of each neighborhood, use foursquare data to get the number and type of venues in the neighborhoods, train a model to fit the crime index to the number and type of venues in them and apply this model to predict a crime index of each neighborhood in the city my friend is moving to.

# Data

To solve the problem above, I will be using the Crimes - Map dataset from Chicago Data Portal (https://data.cityofchicago.org/Public-Safety/Crimes-Map/dfnk-7re6) in combination with foursquare data. The variable of the Crimes Map that I will use will be CASE#, PRIMARY DESCRIPTION, LATITUDE and LONGITUDE. The case# is a identifier of the crime, the primary description shows the details of the crime (for example: primary description: THEFT), and the location show the coordinates of the crime.

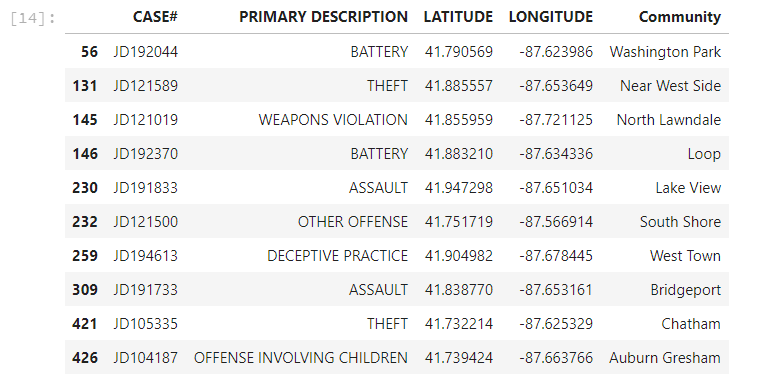
## Data exploration

The crimes map from Chicago have 52,305 observations. The structure of the table is the following

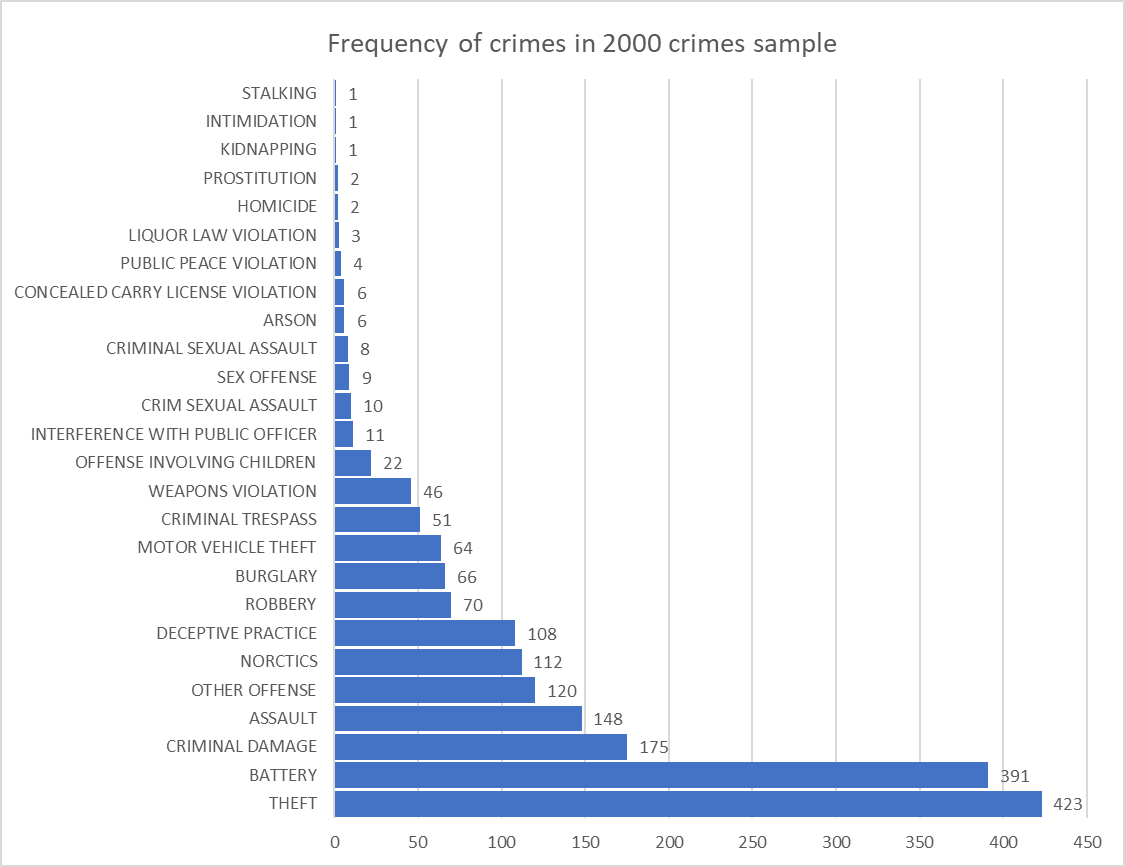


There were 329 lines with empty cells in the table and these were removed.

We will need to assign a neighborhood to each location that a crime has occurred. In order to do that, the [Boundary Service API](http://boundaries.tribapps.com/api/) will was used. A query with the coordinates of each crime must que made using this API, but to this process takes a bit of time. To assign a neighborhood to 2000 crimes took approximately 20 minutes. To use the whole table would take several hours. Therefore, a random sample of 2000 crimes was taken. The result was the following table:



The types of crimes in the samples have the distribution below.



# Methodology

Based on the coordinates of the crime, I will assign it to a neighborhood of Chicago. A crime index will be created based on the number and type of crimes in each neighborhood. Then, Foursquare data will be used to describe the venues in each neighborhood. Foursquare data will also be used to get the number and type of venues in each neighborhood in the city my friend is moving to.

A crime index will be assigned to the neighborhoods of this new city based on the similarities between the number and type of venues in them and the number and type of venues in the neighborhoods in Chicago. My friend would probably choose the neighborhood with the smallest crime index.

## The Crime Index

The methodology to construct the crime index will be subjective but tailored to the needs of my friend: I simply asked him to rank every type of crime in table, giving the highest scores to the worst crimes, according to him. He gave me the following list:

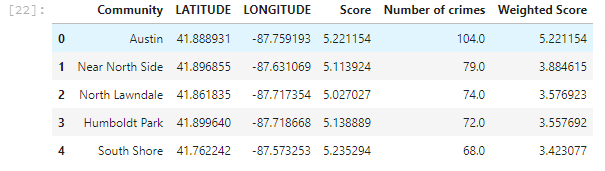


The average score of crimes for each neighborhood was calculated, with the following top ranking neighborhoods:



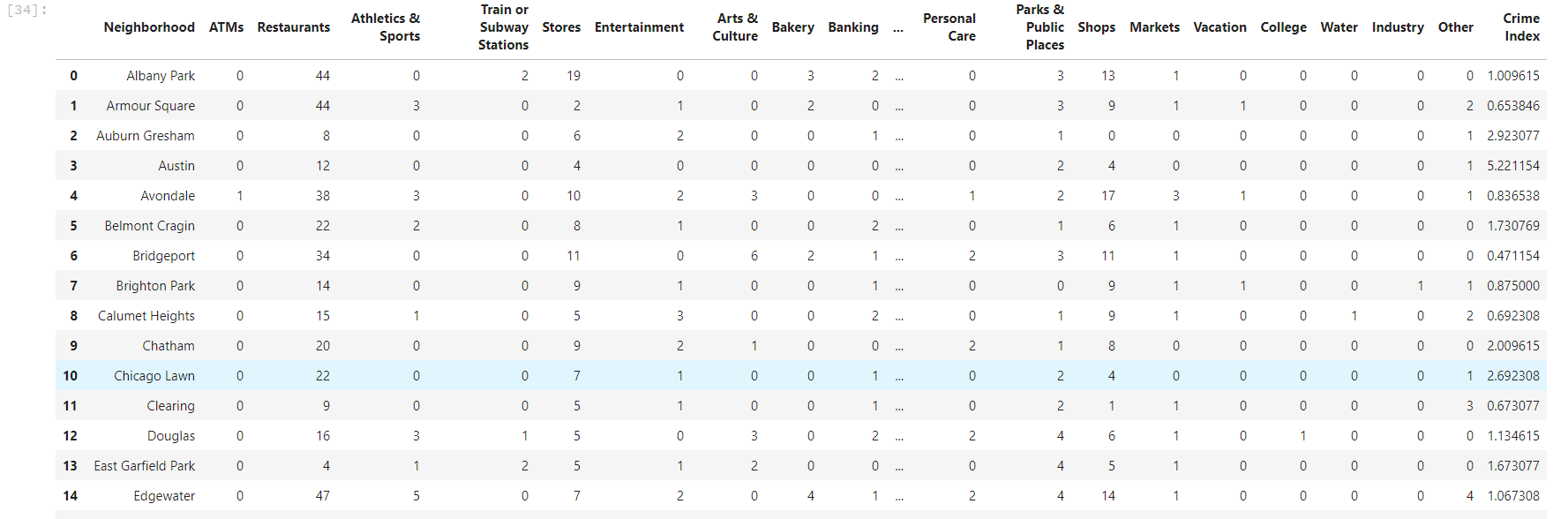
But looking at the table, we see that the number of crimes should be important in a reliable crime index. Therefore, the average score will be multiplied by a weight. The weight given to the index of the neighborhood with the most crimes will be 1, and the weight given to the index of other neighborhoods will be proportional to their number of crimes divided by the maximum number of crimes in the column “Number of crimes”. The result will be a weighted score, the true crime index used in this exercise.

The top 5 rank is:



The Foursquare API was used to get a sample of venues in each of these neighborhoods. The venue’s categories were classified into one of the following classes:

ATMs, Restaurants, Athletics & Sports, Train or Subway Stations, Stores, Entertainment, Arts & Culture, Bakery, Banking, Bars, Gymnasiums and Courts, Hotels and Hospitality, Services, Bus Lines and Stations, Gas Stations, Stadiums and Concert Halls, Hospitals and Clinics, Laundry, Personal Care, Parks & Public Places, Shops, Markets, Vacation, College, Water, Industry and Other. A partial view of the result is below.



My friend is moving to the city of Santo André, State of São Paulo, Brazil. He has to choose between the following neighborhoods:



The Foursquare API was used again to get a sample of venues in each neighborhood and they were classified into the different categories mentioned above. These neighborhoods in the city of Santo André were added to the table containing the neighborhoods in the city of Chicago. The neighborhoods were clustered together using k-means clustering.

The neighborhoods in the city of Santo André were clustered together with the neighborhoods in Chicago. A crime index was assigned to the neighborhoods in Santo André. The crime index assigned was the average crime index of the Chicago neighborhoods in the same cluster.

# Results

The results of the clustering are shown below

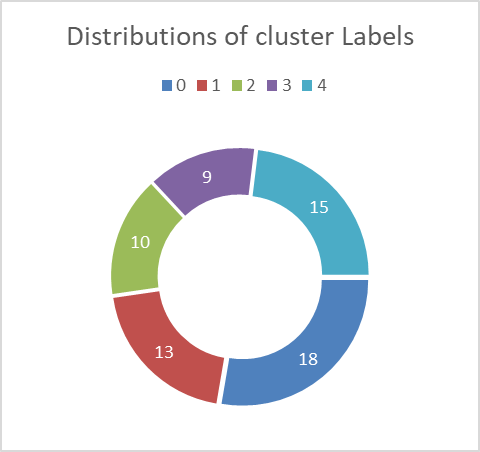
|  |  |  |
| --- | --- | --- |
| **Neighborhood** | **Cluster Labels** | **Crime Index** |
| North Lawndale | 0 | 3,576923077 |
| Auburn Gresham | 0 | 2,923076923 |
| Englewood | 0 | 2,740384615 |
| Roseland | 0 | 2,644230769 |
| West Englewood | 0 | 2,567307692 |
| South Chicago | 0 | 1,894230769 |
| Jardim Irene | 0 | 1,768887363 |
| Jd Alvorada | 0 | 1,768887363 |
| Utinga | 0 | 1,768887363 |
| Vila Luzita | 0 | 1,768887363 |
| East Garfield Park | 0 | 1,673076923 |
| West Pullman | 0 | 1,620192308 |
| Woodlawn | 0 | 1,346153846 |
| Washington Park | 0 | 1,028846154 |
| Washington Heights | 0 | 0,923076923 |
| Clearing | 0 | 0,673076923 |
| South Deering | 0 | 0,615384615 |
| Morgan Park | 0 | 0,538461538 |

|  |  |  |
| --- | --- | --- |
| **Neighborhood** | **Cluster Labels** | **Crime Index** |
| West Town | 1 | 3,365384615 |
| Loop | 1 | 3,278846154 |
| Logan Square | 1 | 2,028846154 |
| Lake View | 1 | 1,951923077 |
| Lincoln Park | 1 | 1,682692308 |
| Bairro Jardim | 1 | 1,61451049 |
| Centro | 1 | 1,61451049 |
| Rogers Park | 1 | 1,365384615 |
| Irving Park | 1 | 1,288461538 |
| Avondale | 1 | 0,836538462 |
| Lincoln Square | 1 | 0,817307692 |
| Hyde Park | 1 | 0,673076923 |
| Bridgeport | 1 | 0,471153846 |

|  |  |  |
| --- | --- | --- |
| **Neighborhood** | **Cluster Labels** | **Crime Index** |
| Greater Grand Crossing | 2 | 3,125 |
| Chicago Lawn | 2 | 2,692307692 |
| Chatham | 2 | 2,009615385 |
| South Lawndale | 2 | 1,788461538 |
| Belmont Cragin | 2 | 1,730769231 |
| Grand Boulevard | 2 | 1,586538462 |
| Gage Park | 2 | 1,211538462 |
| Portage Park | 2 | 0,769230769 |
| Kenwood | 2 | 0,557692308 |
| West Lawn | 2 | 0,538461538 |

|  |  |  |
| --- | --- | --- |
| **Neighborhood** | **Cluster Labels** | **Crime Index** |
| Near North Side | 3 | 3,884615385 |
| Near West Side | 3 | 3,259615385 |
| West Ridge | 3 | 1,423076923 |
| Uptown | 3 | 1,403846154 |
| Edgewater | 3 | 1,067307692 |
| Lower West Side | 3 | 1,028846154 |
| Albany Park | 3 | 1,009615385 |
| Near South Side | 3 | 0,682692308 |
| Armour Square | 3 | 0,653846154 |

|  |  |  |
| --- | --- | --- |
| **Neighborhood** | **Cluster Labels** | **Crime Index** |
| Austin | 4 | 5,221153846 |
| Humboldt Park | 4 | 3,557692308 |
| South Shore | 4 | 3,423076923 |
| West Garfield Park | 4 | 2,355769231 |
| Bairro Campestre | 4 | 2,163461538 |
| Santa Terezinha | 4 | 2,163461538 |
| Vila Assuncao | 4 | 2,163461538 |
| Vila Humaita | 4 | 2,163461538 |
| Vila Metalurgica | 4 | 2,163461538 |
| Vila Pires | 4 | 2,163461538 |
| New City | 4 | 1,471153846 |
| Douglas | 4 | 1,134615385 |
| Brighton Park | 4 | 0,875 |
| Garfield Ridge | 4 | 0,740384615 |
| Calumet Heights | 4 | 0,692307692 |



# Discussion

This project was just an exercise and the crime index developed here does not reflect the reality. In fact, the correlation is very small between the number of each venue's category in a neighborhood and the crime index of that neighborhood. The construction of the crime index itself was very subjective and the weights and averaging process could be improved.

# Conclusion

Below is the crime index for the neighborhoods of the city my friend is moving to.



Based on the results of this exercise, my friend would probably choose one of the first two neighborhoods.