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

	CASE#	DATE OF OCCURRENCE	BLOCK	IUCR	PRIMARY DESCRIPTION	SECONDARY DESCRIPTION	LOCATION DESCRIPTION	ARREST	DOMESTIC	BEAT	WARD	FBI CD	X COORDINATE	Y COORDINATE	LATITUDE	LONGITUDE	LOCATION
0	JD141525	02/05/2020 02:54:00 PM	030XX N HALSTED ST	0860	THEFT	RETAIL THEFT	DRUG STORE	N	N	1933	44.0	06	NaN	NaN	NaN	NaN	NaN
1	JD177980	03/08/2020 02:15:00 AM	064XX S DR MARTIN LUTHER KING JR DR	1330	CRIMINAL TRESPASS	TO LAND	APARTMENT	Y	N	312	20.0	26	1180028.0	1862391.0	41.777671	-87.615561	(41.777670858, -87.61556066)
2	JC497784	11/03/2019 11:40:00 AM	032XX N CLARK ST	0860	THEFT	RETAIL THEFT	DEPARTMENT STORE	N	N	1924	44.0	06	NaN	NaN	NaN	NaN	NaN
3	JD195928	03/21/2020 10:05:00 PM	019XX E 73RD PL	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	NaN	N	N	333	7.0	11	NaN	NaN	NaN	NaN	NaN
4	JC497415	11/03/2019 04:30:00 AM	107XX S PEORIA ST	1320	CRIMINAL DAMAGE	TO VEHICLE	RESIDENTIAL YARD (FRONT/BACK)	N	N	2233	34.0	14	NaN	NaN	NaN	NaN	NaN

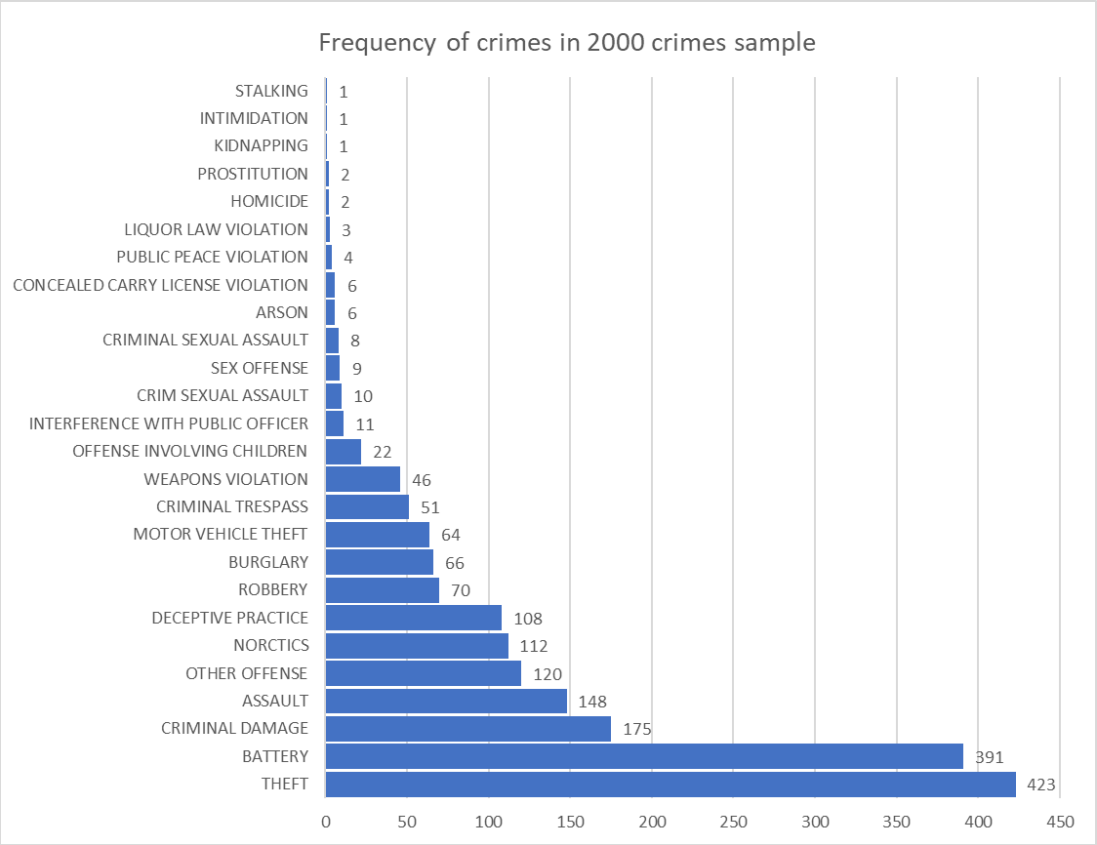
There were 329 lines with empty cells in the table and these were removed.

We will need to assign a neighborhood to each location that a crime has occurred. In order to do that, the Boundary Service API 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:

[14]:

	CASE#	PRIMARY DESCRIPTION	LATITUDE	LONGITUDE	Community
56	JD192044	BATTERY	41.790569	-87.623986	Washington Park
131	JD121589	THEFT	41.885557	-87.653649	Near West Side
145	JD121019	WEAPONS VIOLATION	41.855959	-87.721125	North Lawndale
146	JD192370	BATTERY	41.883210	-87.634336	Loop
230	JD191833	ASSAULT	41.947298	-87.651034	Lake View
232	JD121500	OTHER OFFENSE	41.751719	-87.566914	South Shore
259	JD194613	DECEPTIVE PRACTICE	41.904982	-87.678445	West Town
309	JD191733	ASSAULT	41.838770	-87.653161	Bridgeport
421	JD105335	THEFT	41.732214	-87.625329	Chatham
426	JD104187	OFFENSE INVOLVING CHILDREN	41.739424	-87.663766	Auburn Gresham

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:

Crime	Score
HOMICIDE	10
KIDNAPPING	9,5
OFFENSE INVOLVING CHILDREN	8
CRIM SEXUAL ASSAULT	8
CRIMINAL SEXUAL ASSAULT	8
BATTERY	7
BURGLARY	7
SEX OFFENSE	7
INTIMIDATION	7
CRIMINAL DAMAGE	6
ASSAULT	6
ROBBERY	6
ARSON	6
THEFT	5
MOTOR VEHICLE THEFT	5
NORCTICS	4
INTERFERENCE WITH PUBLIC OFFICER	4
PROSTITUTION	4
STALKING	4
CRIMINAL TRESPASS	3
WEAPONS VIOLATION	3
CONCEALED CARRY LICENSE VIOLATION	3
PUBLIC PEACE VIOLATION	3
DECEPTIVE PRACTICE	2
OTHER OFFENSE	1
LIQUOR LAW VIOLATION	1

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

[20]:

	Community	LATITUDE	LONGITUDE	Score	Number of crimes
0	South Deering	41.715227	-87.573128	6.400000	10.0
1	Clearing	41.777899	-87.765665	6.363636	11.0
2	Irving Park	41.952855	-87.716304	6.090909	22.0
3	Calumet Heights	41.730932	-87.574564	6.000000	12.0
4	Lower West Side	41.852999	-87.667109	5.631579	19.0
5	South Chicago	41.741429	-87.554962	5.628571	35.0
6	Woodlawn	41.778553	-87.602723	5.600000	25.0
7	Chicago Lawn	41.774654	-87.693067	5.490196	51.0
8	West Ridge	41.999191	-87.693149	5.481481	27.0
9	South Lawndale	41.846049	-87.709297	5.470588	34.0

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:

[22]:

	Community	LATITUDE	LONGITUDE	Score	Number of crimes	Weighted Score
0	Austin	41.888931	-87.759193	5.221154	104.0	5.221154
1	Near North Side	41.896855	-87.631069	5.113924	79.0	3.884615
2	North Lawndale	41.861835	-87.717354	5.027027	74.0	3.576923
3	Humboldt Park	41.899640	-87.718668	5.138889	72.0	3.557692
4	South Shore	41.762242	-87.573253	5.235294	68.0	3.423077

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.

[34]:

	Neighborhood	ATMs	Restaurants	Athletics & Sports	Train or Subway Stations	Stores	Entertainment	Arts & Culture	Bakery	Banking	...	Personal Care	Parks & Public Places	Shops	Markets	Vacation	College	Water	Industry	Other	Crime Index
0	Albany Park	0	44	0	2	19	0	0	3	2	...	0	3	13	1	0	0	0	0	0	1.009615
1	Armour Square	0	44	3	0	2	1	0	2	0	...	0	3	9	1	1	0	0	0	2	0.653846
2	Auburn Gresham	0	8	0	0	6	2	0	0	1	...	0	1	0	0	0	0	0	0	1	2.923077
3	Austin	0	12	0	0	4	0	0	0	0	...	0	2	4	0	0	0	0	0	1	5.221154
4	Avondale	1	38	3	0	10	2	3	0	0	...	1	2	17	3	1	0	0	0	1	0.836538
5	Belmont Cragin	0	22	2	0	8	1	0	0	2	...	0	1	6	1	0	0	0	0	0	1.730769
6	Bridgeport	0	34	0	0	11	0	6	2	1	...	2	3	11	1	0	0	0	0	0	0.471154
7	Brighton Park	0	14	0	0	9	1	0	0	1	...	0	0	9	1	1	0	0	1	1	0.875000
8	Calumet Heights	0	15	1	0	5	3	0	0	2	...	0	1	9	1	0	0	1	0	2	0.692308
9	Chatham	0	20	0	0	9	2	1	0	0	...	2	1	8	0	0	0	0	0	0	2.009615
10	Chicago Lawn	0	22	0	0	7	1	0	0	1	...	0	2	4	0	0	0	0	0	1	2.692308
11	Clearing	0	9	0	0	5	1	0	0	1	...	0	2	1	1	0	0	0	0	3	0.673077
12	Douglas	0	16	3	1	5	0	3	0	2	...	2	4	6	1	0	1	0	0	0	1.134615
13	East Garfield Park	0	4	1	2	5	1	2	0	0	...	0	4	5	1	0	0	0	0	0	1.673077
14	Edgewater	0	47	5	0	7	2	0	4	1	...	2	4	14	1	0	0	0	0	4	1.067308

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

[36]:

	Neighborhood	LATITUDE	LONGITUDE
0	Bairro Campestre	-23.637909	-46.542631
1	Centro	-23.661471	-46.529092
2	Vila Assuncao	-23.672044	-46.526948
3	Vila Humaita	-23.673853	-46.503658
4	Bairro Jardim	-23.652431	-46.536477
5	Vila Pires	-23.676149	-46.511200
6	Vila Luzita	-23.701109	-46.507170
7	Jd Alvorada	-23.692945	-46.522974
8	Jardim Irene	-23.709475	-46.510108
9	Vila Metalurgica	-23.627966	-46.534318
10	Utinga	-23.617370	-46.538667
11	Santa Terezinha	-23.635531	-46.532421

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

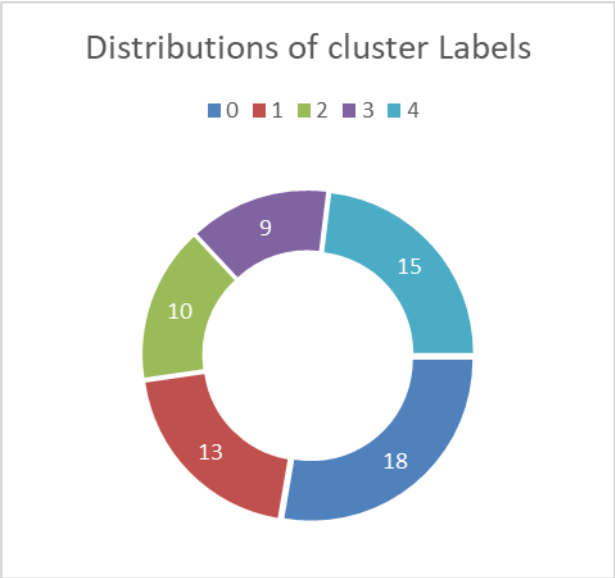
Neighborhood	Cluster Labels	Crime Index
North Lawndale	0	3,576923077
Auburn Gresham	0	2,923076923
Englewood	0	2,740384615
Roseland	0	2,644230769
West Englewood	0	2,567307692
South Chicago	0	1,894230769
Jardim Irene	0	1,768887363
Jd Alvorada	0	1,768887363
Utinga	0	1,768887363
Vila Luzita	0	1,768887363
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.

[20]:

	Neighborhood	Crime Index
1	Bairro Jardim	1.614510
2	Centro	1.614510
3	Jardim Irene	1.768887
4	Jd Alvorada	1.768887
6	Utinga	1.768887
9	Vila Luzita	1.768887
0	Bairro Campestre	2.163462
5	Santa Terezinha	2.163462
7	Vila Assuncao	2.163462
8	Vila Humaita	2.163462
10	Vila Metalurgica	2.163462
11	Vila Pires	2.163462

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