Chicago: Is It the Safest City to Live in the U.S.?

Report for the course of "Applied Data Science Capstone"

The Battle of Neighborhoods (Week 4 - 5)

Totka Toneva

Sept 19, 2019

1. Introduction

Background

Chicago is one of biggest cities in the USA. According to the United States Census Bureau estimates as of 2018, Chicago is the third-largest city by population. It is also a major world financial center and the second-largest central business district in the United States. Nonetheless, Chicago is also among the ten largest metropolitan areas in the US by crime rate. According to Wikipedia, the city's overall crime rate is higher than the US average.

Problem

The question to be discussed is whether there is a relationship or dependency between the crime rate and the venues in the city. Arguably, the crime rate determines the number of venues in the city. Here we will test this hypothesis. Moreover, these two indicators inevitably have some bearing on the housing sale price.

Interest

This project is tailored towards a variety of audiences, ranging from researchers and sociologists to anyone, who might be in a search for a new home or is interested in exploring new community areas in Chicago. For instance, researches, studying the social dependencies in the society; real estate agents and people, interested in real estate market, as the current topic deals with the relationship between crime rate and availablity of venues for meetings and other social and cultural gatherings.

2. Data

In the 1920s, the city has been subdivided into 77 distinct community areas. In order to find the areas with the biggest crime activity, the Data Portal of the city of Chicago provides a detailed information pertaining to the crimes since 2001 (https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2). However, since we would like to get a more up-to-date perspective on the matter, we will utilize the date since 2016.

Chicago Data Portal will also furnish us with details for the community areas, including the respective coordinates. This will help us create a choropleth map of the crime data per area.

From FOURSQUARE website, and more precisely by using its API call, we will obtain information about the current venues in these areas. The idea is to check whether the crime activity, accumulated over a period of three and a half years shapes the venues in the current daily life.

3. Methodology

Data collection

We collected all the necessary data from Chicago Data Portal, i.e. crime data as of 2001 and coordinates of the community areas. The dataset was cleaned and narrowed to include only crime data for each area since 2016. The missing data was dropped, since it would not had any affect on the results (only 2 out of more than 900,000 values were missing). Respectively, Numpy and Pandas packages were employed to manipulate the datasets.

Crime Map of Chicago

Geopy.geocoders.Nominatim package was used the obtain the coordinates of Chicago, while a GeoJSON file from Chicago Data Portal provided the necessary data to define the different areas. A choropleth map of Chicago was created to visualize how the crime rate varies across the community areas. In order to do that, we used Python Folium library.

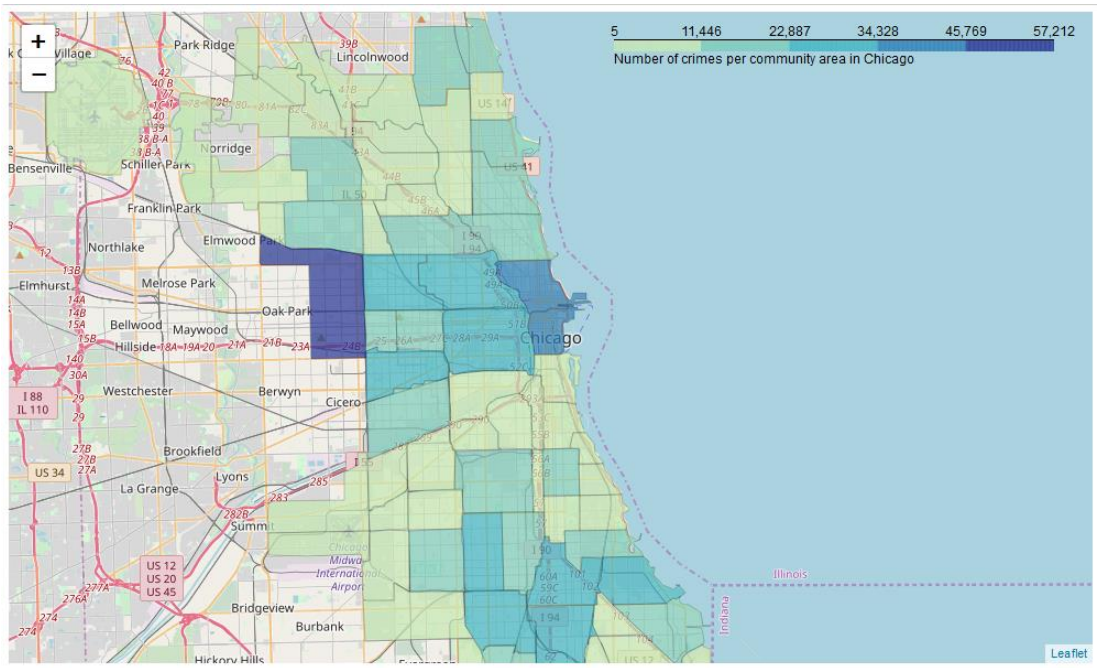

Figure 1: Crime Map of Chicago

Foursquare

The usage of Foursquare API was a prerequisite for this project. For this reason it was used to explore the venues in the community areas of Chicago. We limited our request to 100 venues per area and wrote all the results in a dataframe. Foursquare returned 1,893 venues. Here is a snapshot of the table of the venues in the different areas:

	Community_Name	Community_Latitude	Community_Longitude	Venue	Venue_Latitude	Venue_Longitude	Venue_Category
0	Rogers Park	42.00912	-87.668648	Morse Fresh Market	42.008087	-87.667041	Grocery Store
1	Rogers Park	42.00912	-87.668648	The Common Cup	42.007797	-87.667901	Coffee Shop
2	Rogers Park	42.00912	-87.668648	Glenwood Sunday Market	42.008525	-87.666251	Farmers Market
3	Rogers Park	42.00912	-87.668648	The Glenwood	42.008502	-87.666273	Bar
4	Rogers Park	42.00912	-87.668648	Rogers Park Social	42.007360	-87.666265	Bar

Figure 2: The main venues dataset

The following chart will gives us a better understanding of the number of venues across the areas and how distinctive the different community areas are:

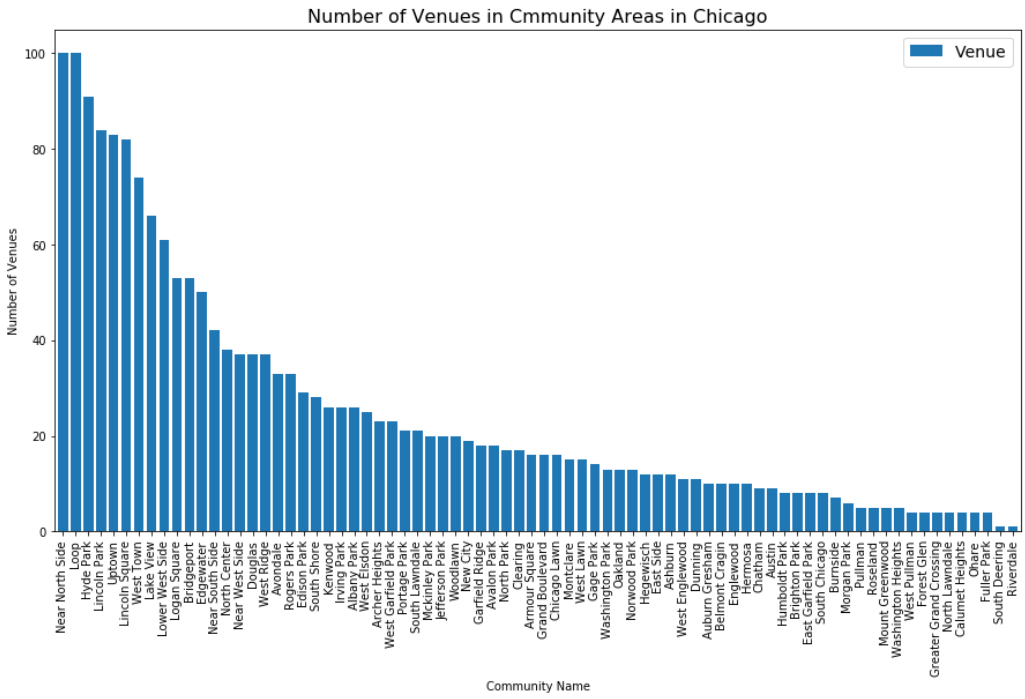

Figure 3: Number of venues in each community area in Chicago

Afterwards, we analyzed all venue category for each community area and created a new dataframe, containing information about the top 10 most common venues in each area.

	Community_Name	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park	Mexican Restaurant	Bakery	Sushi Restaurant	Hot Dog Joint	Park	Korean Restaurant	Bus Station	Taco Place	Fast Food Restaurant	Diner
1	Archer Heights	Mexican Restaurant	Discount Store	Bank	Pizza Place	Sandwich Place	Seafood Restaurant	Bakery	Mobile Phone Shop	Bus Station	Hotel
2	Armour Square	Chinese Restaurant	Pizza Place	Event Service	Business Service	Mobile Phone Shop	Mexican Restaurant	Print Shop	Breakfast Spot	Light Rail Station	Park
3	Ashburn	Pizza Place	Train Station	Construction & Landscaping	Cosmetics Shop	Park	Fast Food Restaurant	Liquor Store	Food	Italian Restaurant	Automotive Shop
4	Auburn Gresham	Pharmacy	BBQ Joint	Fast Food Restaurant	Cosmetics Shop	Discount Store	Video Store	Bakery	Dim Sum Restaurant	Seafood Restaurant	English Restaurant

Figure 4: Top 10 most common venues in each community area

As there are some common venue categories in the community areas, we need to cluster them using the K-Means clustering algorithm and the package Scikit-learn. First, we had to determine how many clusters we should use. 10 different values of K (i.e. from 1 to 10) were tried and "elbow" method helped to find the optimal k, which in this case turned to be 4.

For visualization of the clusters map we used again Folium package. For the charts throughout this project Matplitlib package is utilized.

4. Results and Discussion

We created a choropleth map of the crime rate, combined with the map of clusters. The resulting map looks as follows:

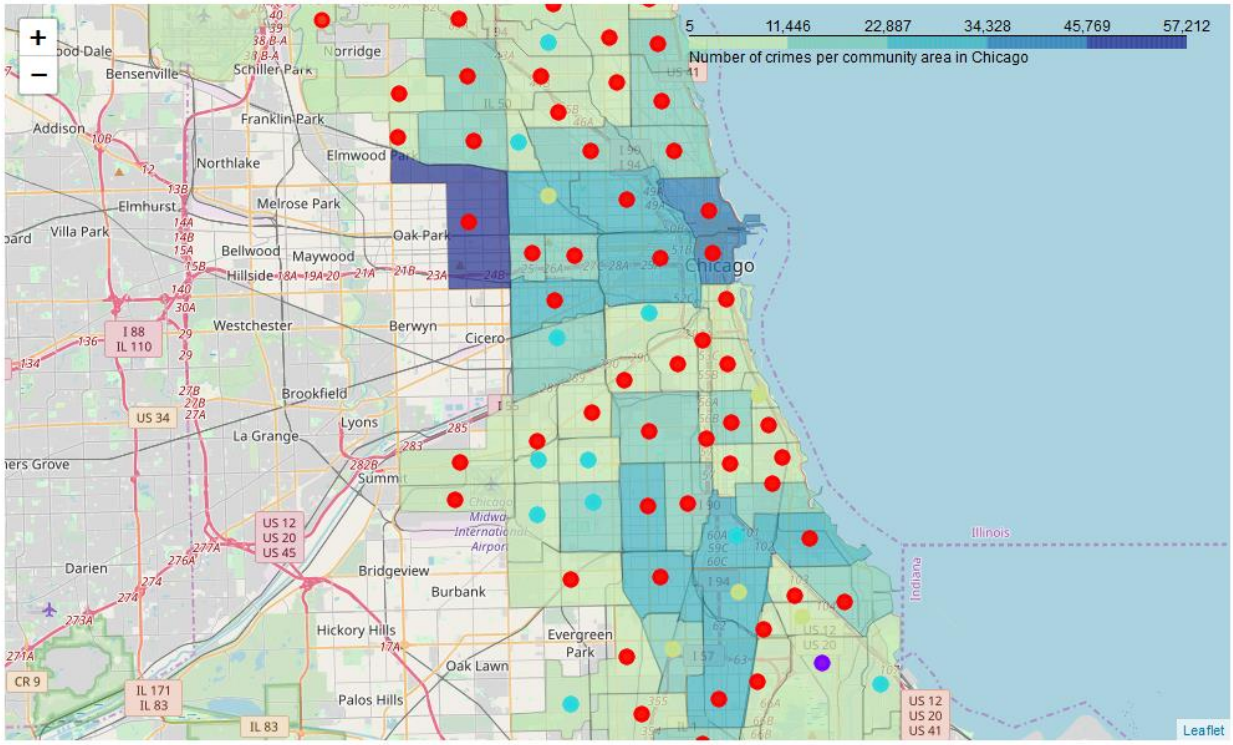

Figure 5: Crime map of Chicago, combined with the map of the clusters

It is evident from the map above that community areas with higher crime rate enjoy similar venues for events and gathering as areas with lower crime rate. Seemingly there is no relationship between them. To verify the this conclusion we performed exploratory data analysis of the dataset.

Exploratory Data Analysis

In our attempt to find out if there is a relationship between the crime rate and the venues, we built the scatterplot and measured the correlation coefficient.

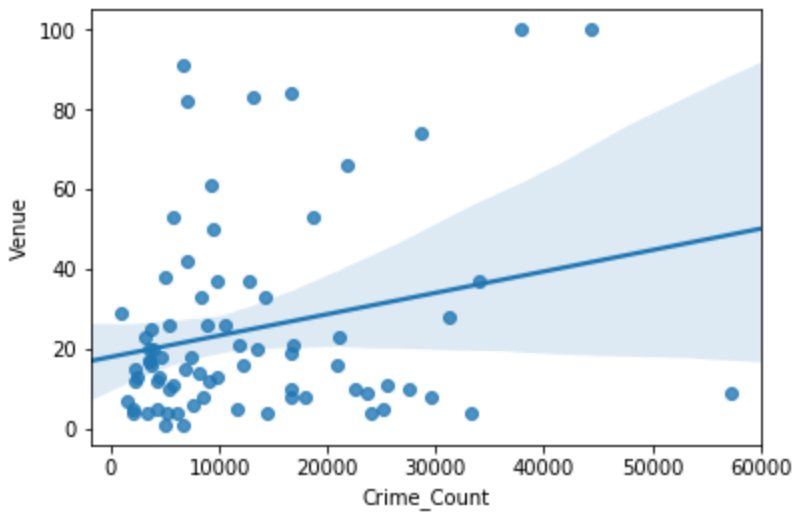

Figure 6: Scatterplot of Venues and Crime rate

Since the p-value is less than 0.05, the correlation between crime rate and the number of venues is statistically significant, but the linear relationship is small (~0.24).

We took a closer look at the 15 community areas with highest crime rate. For the visualization we normalize the data in order to bring it into a similar range for comparison. This enabled us to perform a fair comparison between the two variables.

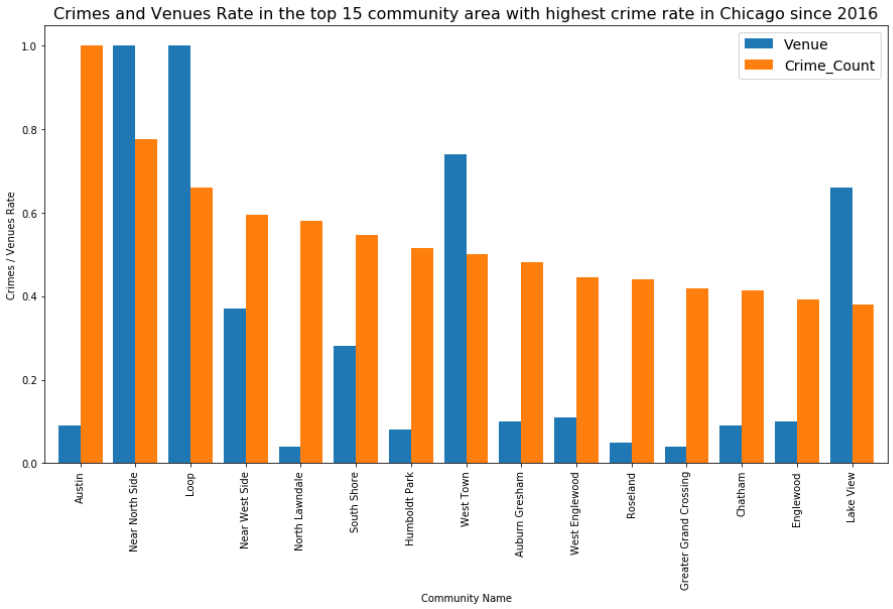

Figure 7: Crimes and Venues Rate in the top 15 community area with highest crime rate in Chicago

5. Conclusion

The results, shown within this research project, do not give us the liberty to assert that there is strong relationship and dependency between the crime rate and the venues in a certain geographic area. Although we cannot confirm to what extent these two factors shape other social and economic indicators as it was beyond the scope this project, our study suggests that their influence should not be underestimated. Additional and more in-depth research on the matter is advisable