Finding Similar Neighborhood to Denver in Chicago

Michael Gray

12/20/2020

1. Introduction

1.1 Background

Billy is living in the neighborhood Lincoln Park in Denver Colorado and while working on a project in Chicago Illinois he met some contacts and got a great job offer in Chicago. After thinking about the offer Billy decided to accept. He is excited about moving to a large city like Chicago with all the activities, however in the news Billy has learned that Chicago is having problems with gun violence and he doesn't know where to look for an apartment. In addition, Billy is very happy with his current neighborhood in Denver, and wants to find a similar neighborhood in Chicago.

1.2 Problem

The available venues in Denver is a big reason Billy is happy with his neighborhood, and it is relatively safe. As a result, the venues from Billy's neighborhood in Denver, will be compared to the venues around the 55+ 'postal codes' in Chicago to find an area that is most similar to his neighborhood in Denver. Then, the crime will be analyzed in Chicago to find an area where Billy will feel safe.

1.3 Interest

This analysis will be very important to Billy because starting a new job is a big decision. In addition, it can be difficult moving to a new city not knowing anybody or any of the nearby venues. As a result, if Billy can find an apartment in an area he is comfortable it will ease the transition starting a new job in a new city.

2. Data Acquisition and Cleaning

2.1 Data sources

In the project the Foursquare and Geocoders APIs' are used. The Foursquare API is used to find the venues around a given location (in latitude & longitude), more information can be found here https://developer.foursquare.com/developer/. The Geocoders API will be used to find the 'latitude & longitude coordinates for the desired locations, more information can be found here

https://geocoder.readthedocs.io/. Then, the ‘postal codes’, ‘population’, and ‘crime’ data in Chicago was downloaded from https://data.cityofchicago.org/.

2.2 Collecting & Cleaning data

First thing to do is get the coordinates in (latitude & longitude) for the neighborhood in Denver and for all the postal codes in Chicago. To get the coordinates for the neighborhood in Denver the Geocoder API is used (39.7331384, -105.0052409729397). Then, to get the coordinates for each postal code in Chicago all the postal codes need to be collected. for the postal codes in Chicago the dataset “Chicago Population Counts” found at https://data.cityofchicago.org/. In the dataset there are 60 entries and 20 features, there are two features of importance; the ‘Population – Total’, and ‘Record ID’. The postal codes will be used to get the coordinates for the areas in Chicago, and the population will be used later in the analysis.

Population and Postal Codes in Chicago

	Geography Type	Year	Geography	Population - Total	Population - Age 0-17	Population - Age 18-29	Population - Age 30-39	Population - Age 40-49	Population - Age 50-59	Population - Age 60-69	Popul - Ag
0	Citywide	2018	Chicago	2705988	548999.0	552935.0	456321.0	336457.0	312965.0	262991.0	155
1	ZIP Code	2018	60601	14675	NaN	NaN	NaN	NaN	NaN	NaN	
2	ZIP Code	2018	60602	1244	NaN	NaN	NaN	NaN	NaN	NaN	
3	ZIP Code	2018	60603	1174	NaN	NaN	NaN	NaN	NaN	NaN	
4	ZIP Code	2018	60604	782	NaN	NaN	NaN	NaN	NaN	NaN	

Then using the Geocoders the coordinates for each postal code was found, and a new data frame was built:

Chicago Postal Codes, Population and Coordinates

	Postal Code	Population	Latitude	Longitude
0	60601-2018	14675	41.885967	-87.624265
1	60602-2018	1244	41.882993	-87.629251
2	60603-2018	1174	41.880925	-87.628176
3	60604-2018	782	41.878284	-87.628010
4	60605-2018	27519	41.870417	-87.627331

Next, to find the venues around each location the Foursquare API is used. By running the Foursquare query all the most popular venues are returned, however the free version of the API is used so that a maximum of 100 venues can be returned for each query. The API returns a .json file which is parsed to create a dataframe that combined with the dataframe above with the features to

describe the center of the location with the coordinates, the venue with the coordinate, and the distance from the venue and the center location.

Venue Dataframe

	Center	Lat_center	Lng_center	Category	Name	Lat_venue	Lng_venue	Distance (m)
0	Lincoln Park	39.733138	-105.005241	Brewery	Renegade Brewing Company	39.730616	-104.999292	581.849369
1	Lincoln Park	39.733138	-105.005241	Japanese Restaurant	Domo Japanese Country Foods Restaurant	39.738100	-105.005650	551.998957
2	Lincoln Park	39.733138	-105.005241	Arts & Entertainment	Santa Fe Art District	39.730636	-104.998669	628.149538
3	Lincoln Park	39.733138	-105.005241	Steakhouse	The Buckhorn Exchange	39.732205	-105.005067	104.710877
4	Lincoln Park	39.733138	-105.005241	Café	The Molecule Effect	39.735386	-104.998781	607.382310

Lastly, the crime data in Chicago is used to find the crimes around each location. The crime data “Crimes _ 2019” was downloaded from https://data.cityofchicago.org/, and the dataset contains 260,000 entries with 22 rows.

Raw Chicago Crime Data

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	...
0	11937949	JC566932	12/31/2019 08:35:00 PM	064XX S WESTERN AVE	031B	ROBBERY	ARMED - OTHER FIREARM	COMMERCIAL / BUSINESS OFFICE	False	False	...
1	11666065	JC237357	04/25/2019 01:49:00 AM	068XX S JUSTINE ST	051A	ASSAULT	AGGRAVATED - HANDGUN	RESIDENCE - PORCH / HALLWAY	False	False	...
2	11613029	JC173331	02/28/2019 04:00:00 PM	035XX W 55TH ST	1582	OFFENSE INVOLVING CHILDREN	CHILD PORNOGRAPHY	SCHOOL - PUBLIC BUILDING	False	False	...
3	11596162	JC152837	02/14/2019 09:03:00 PM	072XX S UNIVERSITY AVE	0498	BATTERY	AGG. DOMESTIC BATTERY - HANDS, FISTS, FEET, SE...	RESIDENCE	False	True	...
4	11936658	JC565320	12/29/2019 04:00:00 PM	097XX S INGLESIDE AVE	1020	ARSON	BY FIRE	RESIDENCE - GARAGE	False	False	...

The crime data was then used to count the crimes around each location.

3. Methodology

The method for analyzing the data in this project is to use logistic regression. First, the dataframe of the venues.

Example of Venue Dataframe

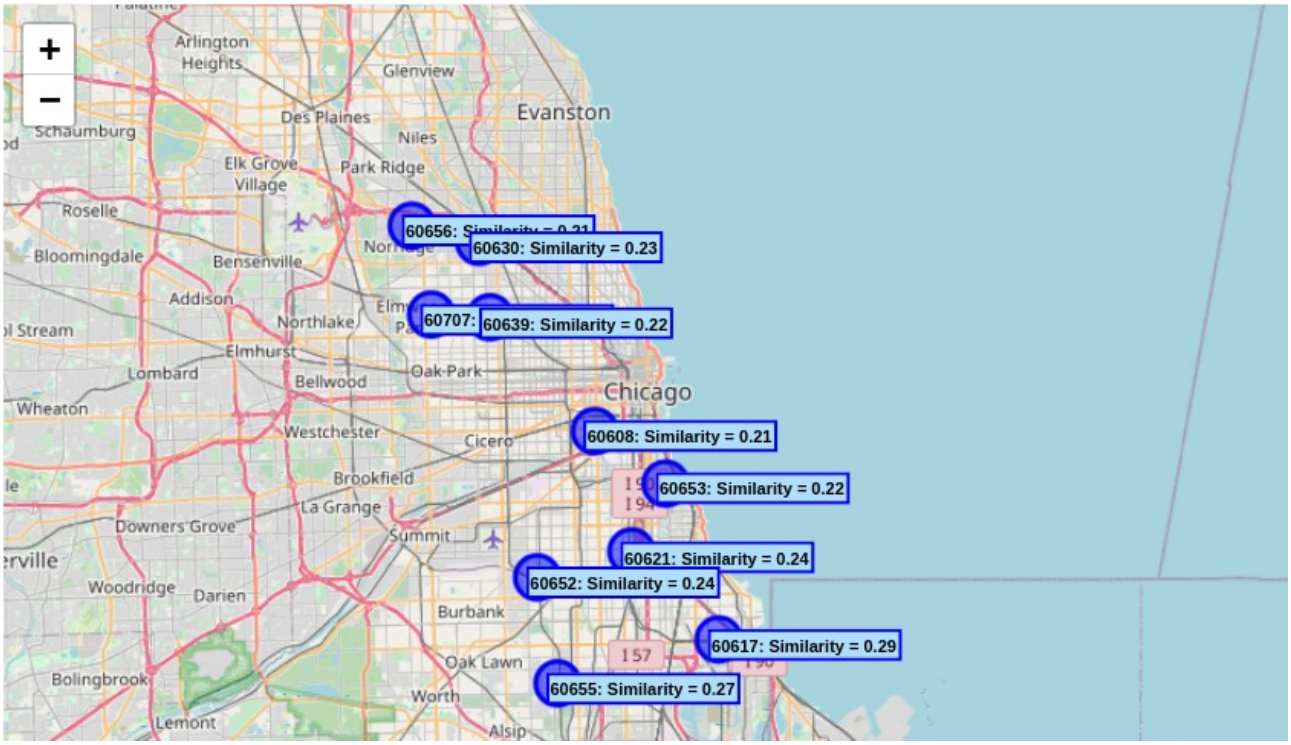
	Center	Lat_center	Lng_center	Category	Name	Lat_venue	Lng_venue	Distance (m)
0	Lincoln Park	39.733138	-105.005241	Brewery	Renegade Brewing Company	39.730616	-104.999292	581.849369
1	Lincoln Park	39.733138	-105.005241	Japanese Restaurant	Domo Japanese Country Foods Restaurant	39.738100	-105.005650	551.998957

In the datafame the only the features ‘Center’, ‘Category’, and ‘Distance (m)’ will be used. The feature ‘Category’ is of type object, so it will need to be expanded to integer features.

	Center	Distance (m)	Category_ATM	Category_Afghan Restaurant	Category_African Restaurant	Category_Airport	Category_Airport Service	Category_Airport Terminal	Category_American Restaurant	Categc
0	Lincoln Park	581.849369	0	0	0	0	0	0	0	
1	Lincoln Park	551.998957	0	0	0	0	0	0	0	
2	Lincoln Park	628.149538	0	0	0	0	0	0	0	
3	Lincoln Park	104.710877	0	0	0	0	0	0	0	
4	Lincoln Park	607.382310	0	0	0	0	0	0	0	

The feature ‘Distance (m)’ is currently in a floating point number from 0-1500m which is cannot be used for the analysis, as a result the feature will be converted to an integer with step sizes of 150m. The machine learning method used to analyze the data is linear regression. This method is appropriate because the objective is to find venues that are a close distance from the center location. For the analysis linear regression analysis was performed on the venues around Lincoln Park, Denver, Colorado which had an accuracy of 0.73. Then, the model was used to predict the venues found around the postal codes in Chicago

Comparison to locations in Chicago

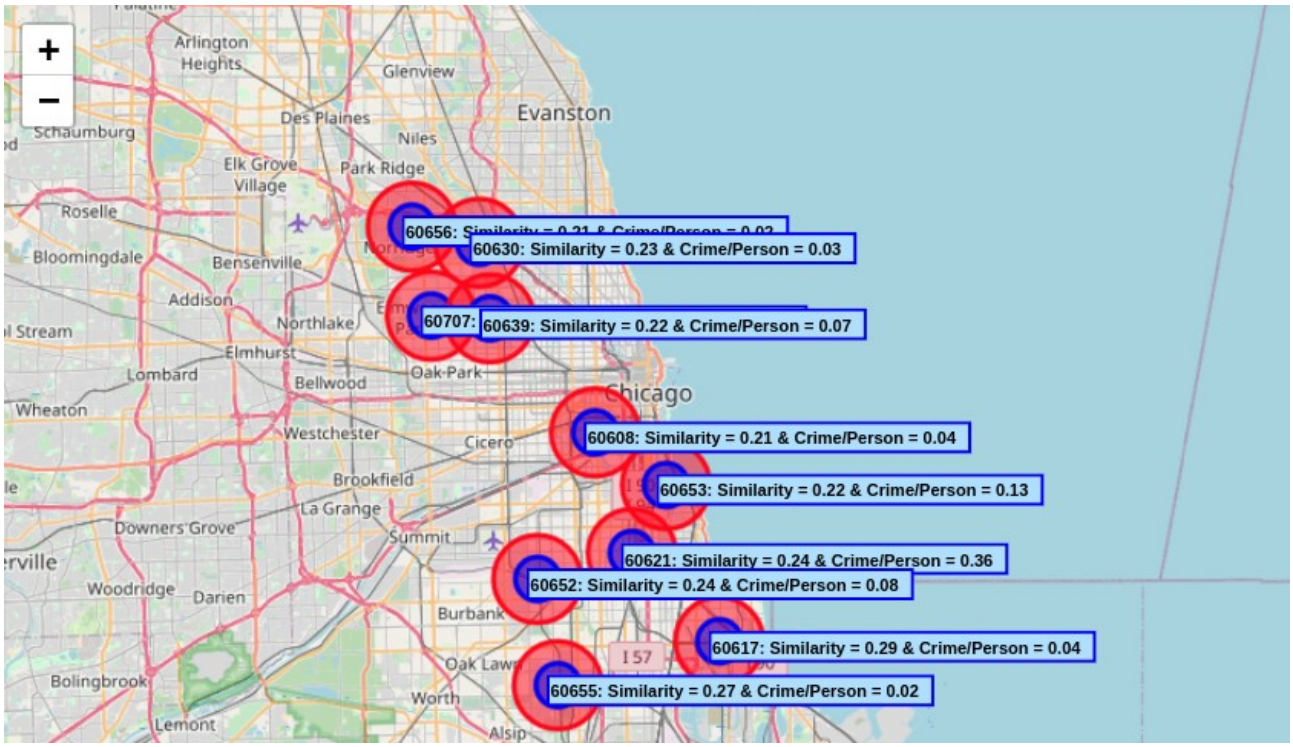


In addition, the crimes in Chicago were analyzed to find a location that is similar to the neighborhood in Denver and a relatively low crime rate. Since Chicago is such a large city there is a large amount of crimes in the city. As a result, only crimes considered to be the most sever were counted.

['ARSON', 'ASSAULT', 'BATTERY', 'BURGLARY', 'CRIM SEXUAL ASSAULT', 'CRIMINAL SEXUAL ASSAULT', 'HOMICIDE', 'KIDNAPPING', 'OFFENSE INVOLVING CHILDREN', 'ROBBERY', 'SEX OFFENSE']

The crimes above were counted for a radius of $3000m$ around each location in Chicago. Then, from the dataframe created with the population the $\frac{crime}{population}$ was computed to find the local crime rate for each location.

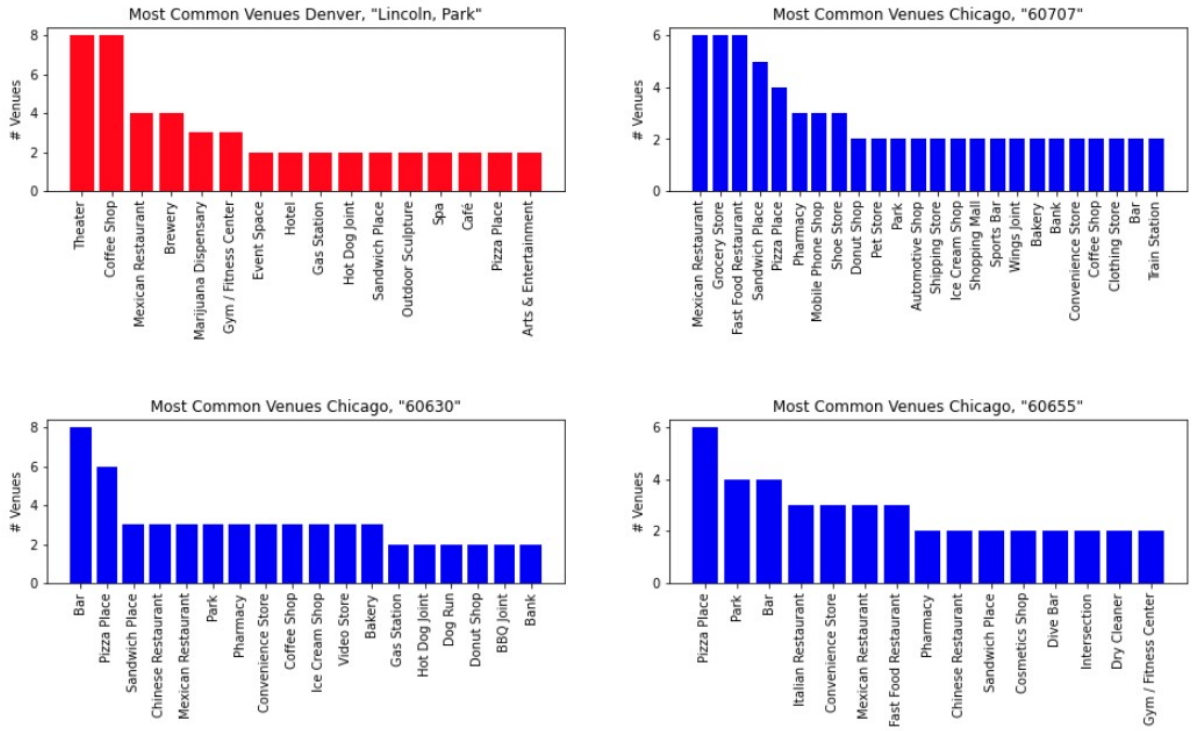
City Similarity (10 most similar) and Crime/Person in Chicago



4. Results

In the analysis the similarity among the neighborhood is not very strong. For example, the most similar location in Chicago was 60607 with a accuracy score of 0.3 and a $\frac{crime}{population}$ = 0.03.

Venues in Most Common Locations



For the 3 most common Chicago location to the neighborhood in Denver, the analysis seems to show that they all have 3 or more Mexican restaurants and sandwich places.

5. Discussion

In the analysis it appears that the locations in Chicago are not very similar to the neighborhood in Denver. In addition, the best place for Billy to look for an apartment is in the northwest neighborhoods of Chicago. The northwest part of Chicago has relatively low crime and the area will be the most similar to his old neighborhood.

6. Conclusion

In the analysis it is shown that the northwest area of Chicago is the most similar to Lincoln Park, Denver, Colorado. In future analysis could be improved by analyzing the Chicago neighborhood in more detail. For example, the analysis could include traffic, property prices, education, ect. However, for this project I got to see how machine learning can be used to solve a real-world problem.