Finding the best neighborhood in Chicago

Devi Selvakumar

Data

There are two data sets to be used for this project.

1. Chicago crime data set:

Chicago's current crime data is acquired from the City of Chicago data portal https://data.cityofchicago.org/api/views/qzdf-xmn8/rows.csv . The best part of this data set is: it has the crime records from Jan.1st 2020 to Dec.5th 2020, which is very rare. The CSV file lists the crime type, action taken, the district/ward/neighborhood it happened along with other data. The data set contains 1048576 rows and 24 columns containing the crime data details from Jan.1,2020 to Dec.5,2020.

ID	Case Nu	Date	Block	IUCR	Primary	Descrip	Locatio	Arrest	Domest	Beat	District	Ward	Commu	FBI Cod	X Coorc	Y Coorc	Year	Upda
12240781	JD453981	1/1/2020 0:00	040XX W (	1154	DECEPTIVE	FINANCIAL	MOVIE HC	FALSE	FALSE	1114	11	28	26	11			2020	12/1
12220321	JD430436	1/1/2020 0:00	091XX S D	1752	OFFENSE I	AGGRAVA	RESIDENCI	FALSE	TRUE	413	4	8	47	17			2020	11/1
12191982	JD397743	1/1/2020 0:00	116XX S E(	1153	DECEPTIVE	FINANCIAL	APARTMEI	FALSE	FALSE	522	5	34	53	11	1175385	1827727	2020	10/1
12186705	JD391621	1/1/2020 0:00	058XX W (	1153	DECEPTIVE	FINANCIAL	OTHER (SP	FALSE	FALSE	1511	15	29	25	11	1137369	1904748	2020	10/1
12181533	JD384685	1/1/2020 0:00	044XX S G	1754	OFFENSE I	AGGRAVA	APARTMEI	FALSE	TRUE	222	2	4	39	2	1184334	1875715	2020	10/
12168471	JD370494	1/1/2020 0:00	049XX N K	281	CRIMINAL	NON-AGG	APARTMEI	FALSE	FALSE	1712	17	39	14	2	1145802	1932546	2020	9/2

	Community Area	Date	Primary Type
0	8	2020-04-10	DECEPTIVE PRACTICE
1	47	2020-12-03	OTHER OFFENSE
2	66	2020-12-03	WEAPONS VIOLATION
3	46	2020-12-03	THEFT
4	71	2020-12-03	CRIMINAL DAMAGE
5	30	2020-11-27	DECEPTIVE PRACTICE
6	61	2020-11-30	THEFT
7	29	2020-12-03	ROBBERY
8	32	2020-12-03	THEFT
9	68	2020-12-03	CRIMINAL DAMAGE
10	24	2020-12-03	MOTOR VEHICLE THEFT
11	48	2020-11-25	DECEPTIVE PRACTICE
12	48	2020-12-01	DECEPTIVE PRACTICE

2. Chicago Neighborhood data set:

There is no direct data set available which lists the community name along with the area code and the map coordinates. Upon thorough search, I came across a text file which has the needed data. This text file parsed the Chicago Neighborhood map and listed the map coordinates of all the points within the neighborhood boundaries along with the name. The original text file had 1048576 rows and 4 columns. In total, there are 77 neighborhoods in Chicago.

Community Area	Community Name	Latitude	Longitude
1	Rogers Park	42.003801	-87.657651
1	Rogers Park	42.002439	-87.657809
1	Rogers Park	41.99839	-87.657676
1	Rogers Park	42.009069	-87.661341
1	Rogers Park	42.00568	-87.660129
1	Rogers Park	42.00567	-87.661205
1	Rogers Park	42.005401	-87.661198
1	Rogers Park	42.006879	-87.66174
1	Rogers Park	42.001313	-87.662222
1	Rogers Park	42.001551	-87.663739
1	Rogers Park	41.998281	-87.664583
1	Rogers Park	42.009012	-87.664783
1	Rogers Park	42.00806	-87.665325
1	Rogers Park	42.005597	-87.666079
1	Rogers Park	42.007759	-87.666653
1	Rogers Park	42.019371	-87.667178
1	Rogers Park	41.998205	-87.66811
1	Rogers Park	42.01924	-87.668418

I have decided to take the average of latitude/longitude for each community area to use for our purpose. To locate the community's name with latitude / longitudes, the map of Chicago neighborhoods is processed and the various points in the neighborhoods is listed as a text file in : https://raw.githubusercontent.com/jkgiesler/parse-chicago-neighborhoods/master/community_to_gps.txt

	Community Area	Community Name	Latitude	Longitude
0	1	Rogers Park	42.009120	-87.668648
1	2	West Ridge	41.999316	-87.692394
2	3	Uptown	41.966222	-87.658792
3	4	Lincoln Square	41.968844	-87.685397
4	5	North Center	41.950503	-87.681029
...	...	...	...	...
72	73	Washington Heights	41.717396	-87.650298
73	74	Mount Greenwood	41.695023	-87.706722
74	75	Morgan Park	41.691034	-87.667514
75	76	O'Hare	41.975927	-87.843830
76	77	Edgewater	41.984299	-87.663061

Exploratory data analysis

The total crimes per community area is calculated.

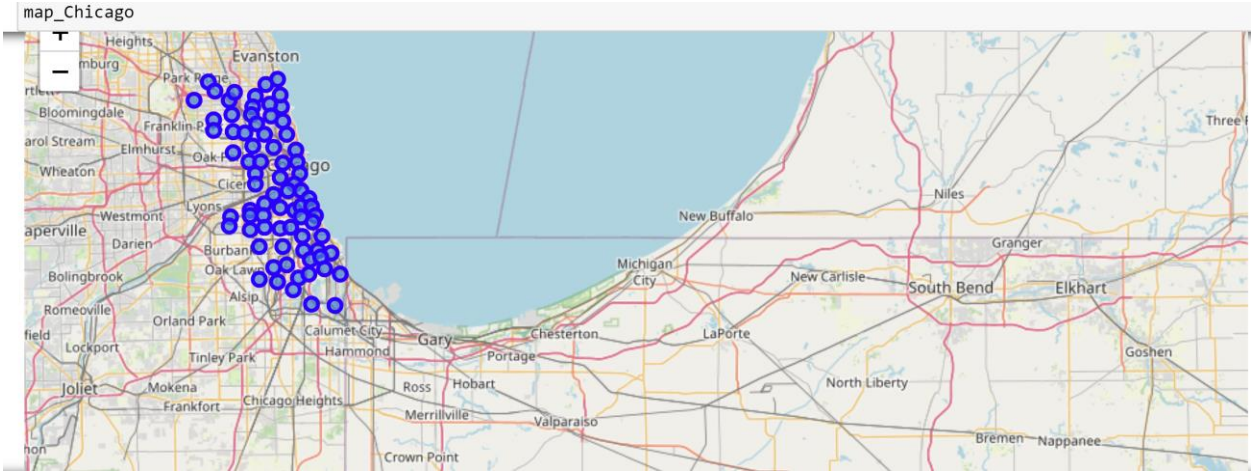
	Community Area	Community Name	Latitude	Longitude	Total crimes
24	25	Austin	41.893204	-87.762589	11709
42	43	South Shore	41.763049	-87.574549	7188
7	8	Near North Side	41.897838	-87.630686	6739
28	29	North Lawndale	41.860835	-87.715818	6503
22	23	Humboldt Park	41.903889	-87.719205	6174
...	...	...	...	...	...
11	12	Forest Glen	41.988860	-87.757509	475
54	55	Hegewisch	41.655467	-87.546453	455
73	74	Mount Greenwood	41.695023	-87.706722	436
46	47	Burnside	41.725473	-87.600361	321
8	9	Edison Park	42.005335	-87.813459	222

77 rows × 5 columns

Data visualization

Using Folium, a map is plotted to see how the different neighborhoods are spread out in Chicago.

Chicago Neighborhood map



Exploring nearby venues using Foursquare API

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	1	42.009120	-87.668648	Morse Fresh Market	42.008087	-87.667041	Grocery Store
1	1	42.009120	-87.668648	The Common Cup	42.007797	-87.667901	Coffee Shop
2	1	42.009120	-87.668648	Smack Dab	42.009291	-87.666201	Bakery
3	1	42.009120	-87.668648	Glenwood Sunday Market	42.008525	-87.666251	Farmers Market
4	1	42.009120	-87.668648	Lifeline Theatre	42.007372	-87.666284	Theater
...	...	...	...	...	...	...	...
1812	77	41.984299	-87.663061	Edgewater Mexican Cafe	41.983495	-87.657809	Mexican Restaurant
1813	77	41.984299	-87.663061	Uvae	41.983364	-87.668650	Wine Bar
1814	77	41.984299	-87.663061	House 5863 Bed & Breakfast	41.988497	-87.664767	Hotel
1815	77	41.984299	-87.663061	Continuum Studio	41.985280	-87.668904	Spa
1816	77	41.984299	-87.663061	Blu 57 Seafood & Small Plates	41.985570	-87.668810	Seafood Restaurant

1817 rows × 7 columns

Top 10 venues for each neighborhood

	Neighborhood	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	1	Mexican Restaurant	Theater	American Restaurant	Bar	Donut Shop	Bakery	Pizza Place	Diner	Greek Restaurant	Breakfast Spot
1	2	Indian Restaurant	Pakistani Restaurant	Grocery Store	Donut Shop	Burmese Restaurant	Fruit & Vegetable Store	Market	Chinese Restaurant	Fast Food Restaurant	Bar
2	3	Coffee Shop	Pizza Place	Bar	Mexican Restaurant	Chinese Restaurant	Sandwich Place	Diner	Bus Station	Sushi Restaurant	Convenience Store
3	4	Bar	Sandwich Place	Café	Thai Restaurant	Sushi Restaurant	American Restaurant	Art Gallery	Pizza Place	Mobile Phone Shop	Chinese Restaurant
4	5	Pub	Mobile Phone Shop	Bar	Pizza Place	Coffee Shop	Boutique	Beer Garden	Bank	German Restaurant	Sushi Restaurant

Methodology

The way the neighborhood data will be used is by finding the following scores for each neighborhood:

1. Crime score –>
   = 0 if the total number of crimes in that neighborhood > the mean of total number of crimes in all neighborhoods
   = 1 if the total number of crimes in that neighborhood < the mean of total number of crimes in all neighborhoods.

2. Shopping score->
   = 0 if the total number of shopping places in that neighborhood < the mean of total number of shopping centers in all neighborhoods
   = 1 if the total number of shopping places in that neighborhood >= the mean of total number of shopping centers in all neighborhoods

3. Eatery score –>
   = 0 if the total number of restaurant/eating places in that neighborhood < the mean of total number of eatery places in all neighborhoods

   = 1 if the total number of restaurant/eating places in that neighborhood >=the mean of total number of eatery places in all neighborhoods
4. Total score = Crime score + Shopping score + Eatery score

So the best score for each neighborhood is determined as:

- If the total score = 3, then it is the best choice, as
  - the # of crimes in that neighborhood is less than the average
  - The # of shopping centers is more than the average of all shopping places included

    - The # of eatery places are more than the average of all restaurants and other fast-food places.
- If the total score = 2 then it is moderate, 2 of the scores are good and one of them may be because of:
    - Either the # of crimes in that neighborhood is more than the average
    - Or the # of shopping places is less than the average number of shopping centers
    - Or the # of eatery places may be less than the average.
- If the total score = 1, then it is the least choice, as only one of the scores is looking good and other scores are bad.
- If the total score = 0, then it should be avoided, as the total # of crimes is more than the average and the shopping and restaurants are less than the average available in Chicago.

Unsupervised learning

K means clustering will be used to cluster the neighborhoods into various clusters and then the data analysis and visualization is done per cluster. Silhouette coefficient will be used to find the number of clusters to be used. With the help of histograms and bar charts, the data is modelled per cluster to find the best options of selecting the neighborhoods in Chicago.