

# Finding the best community area in Chicago to live in

Aparna Subhakari R S

## 1 Introduction/Business Problem

### 1.1 Background

Chicago is the most populous and popular city in Illinois. With an estimated population of 2,716,450 (2017), it is the third most populous city in the United States, the third-largest in the United States, and the fourth largest in North America. Chicago is an international hub for finance, culture, commerce, industry, technology, telecommunications, and transportation. Chicago has one of the highest GDP in the world, generating \$ 680 billion in 2017. The city lies beside huge freshwater Lake Michigan, and two rivers— the Chicago River in downtown and the Calumet River in the industrial far South Side—flow entirely or partially through Chicago making it most beautiful place to live in. One of the main considerations of a move to Chicago is affordability.

### 1.2 Description of the problem

According to the details mentioned above , Chicago seems to be a great place to move in and to live. But the question is which community area or which community area neighbourhood we choose as the city is divided into 77 community areas and around 100 neighbourhoods. Factors that determine the best place for living include 'Housing Availability', 'Employment rate', 'House Holds below poverty line'. The goal of this project is to find the best community area and neighbourhoods to live in based on the factors mentioned above. The questions can be framed as :

1. What is the best community area to live in
2. How community areas are categorized according to crowded housing or unemployment rate etc. features?
3. Does correlation exist between these features selected?

## 1.3 Target Audience

Students who want to pursue higher education as the universities here are ranked in top among the top universities is USA. The city is popular with college students because they can enjoy student life without worrying too much about covering the bills as cost of living is reasonable. Job aspirants who want to get into good jobs with high salaries and also the comfortable life in every means because Chicago has steady job market. There are many big businesses here and plenty of work opportunities. It offers a vibrant city life without the hefty property price tags of New York or Los Angeles and the cost of living is lower here. This makes a very good reason to live in Chicago with families as well.

## 2 Data Description

### 2.1 Data Acquisition

1. I scraped list of neighbour community areas and neighbourhoods from the page : [https://en.wikipedia.org/wiki/Community\\_areas\\_in\\_Chicago](https://en.wikipedia.org/wiki/Community_areas_in_Chicago)

#### Sample Screen Shot of List of Community Areas and Neighbourhoods

Number	Community area	Neighborhoods
08	Near North Side	<ul style="list-style-type: none"><li>• Cabrini–Green</li><li>• The Gold Coast</li><li>• Goose Island</li><li>• Magnificent Mile</li><li>• Old Town</li><li>• River North</li><li>• River West</li><li>• Streeterville</li></ul>
32	Loop	<ul style="list-style-type: none"><li>• Loop</li><li>• New Eastside</li><li>• South Loop</li><li>• West Loop Gate</li></ul>
33	Near South Side	<ul style="list-style-type: none"><li>• Dearborn Park</li><li>• Printer's Row</li><li>• South Loop</li><li>• Prairie Avenue Historic District</li></ul>

2. I have obtained the required census data regarding housing, unemployment etc. from the page : <https://ibm.box.com/shared/static/05c3415cbf9b2f94atenb2sd361ze.csv>

#### Sample Screen Shot of Census Data

COMMUNITY_AREA_NAME	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED
1 Rogers Park	7.7	23.6	8.7
2 West Ridge	7.8	17.2	8.8
3 Uptown	3.8	24	8.9
4 Lincoln Square	3.4	10.9	8.2
5 North Center	0.3	7.5	5.2
6 Lake View	1.1	11.4	4.7
7 Lincoln Park	0.8	12.3	5.1
8 Near North Side	1.9	12.9	7
9 Edison Park	1.1	3.3	6.8
10 Norwood Park	2	5.4	9
11 Jefferson Park	2.7	8.6	12.4
12 Forest Glen	1.1	7.5	6.8
13 North Park	3.9	13.2	9.9
14 Albany Park	11.3	19.2	10
15 Portage Park	4.1	11.6	12.8
16 Irving Park	6.3	13.1	10
17 Dunning	5.2	10.6	10
18 Montclair	8.1	15.3	13.8
19 Belmont Cragin	10.8	18.7	14.8
20 Hermosa	6.9	20.5	13.1
21 Avondale	6	15.3	9.2
22 Logan Square	3.2	16.8	8.2

- 
3. I have extracted geo location data for every community area from the page : <https://ibm.box.com/shared/static/f9gjj1gjmxxzycdhplzt01qtz0s7ew7.csv>

### Sample Screen Shot of School Data

1	Latitude	Longitude	COMMUNITY_AREA_NAME	COMMUNITY_AREA_NUMBER
2	41.92449696	-87.64452163	LINCOLN PARK	7
3	41.76032435	-87.55673627	SOUTH SHORE	43
4	41.74711093	-87.73170248	ASHBURN	70
5	41.8097569	-87.6721446	NEW CITY	61
6	41.82814609	-87.63279369	ARMOUR SQUARE	34
7	41.9711433	-87.70962725	ALBANY PARK	14
8	41.94661693	-87.69105603	NORTH CENTER	5
9	41.90684338	-87.68304259	WEST TOWN	24
10	41.93576106	-87.68052441	NORTH CENTER	5
11	41.97160605	-87.72464139	ALBANY PARK	14
12	41.94952795	-87.68605496	NORTH CENTER	5
13	41.81329195	-87.64338051	NEW CITY	61
14	41.94524725	-87.67085683	LAKE VIEW	6
15	41.90672216	-87.69300325	WEST TOWN	24
16	41.68663682	-87.63359476	ROSELAND	49
17	41.70485072	-87.62360199	ROSELAND	49
18	41.90612797	-87.72993324	HUMBOLDT PARK	23
19	41.70514024	-87.65811642	BEVERLY	72
20	41.886183	-87.71539705	EAST GARFIELD PARK	27

## 2.2 Data Cleaning

The data is extracted from the sources mentioned above and loaded into data frames for further processing.

1. Data extracted from the wikipedia has no problems. It is a structured tabular data with no null values and empty rows.
2. Data extracted from the 'Census Data Set' is also a well structured tabular data with many features. I eliminated empty rows i.e null values for features as they cant be included in the calculations.
3. The above process is repeated for the 'School Data Set' and prepared it for further processing.
4. Longitude and Latitude values of every community area in Chicago are extracted too from 'School Data set'.

Every data frame has same number of rows which is equal to the total number of unique community areas i.e 77, after the data pre-processing and data preparation. All the data frames are merged on a single column 'Community area' as the aim of this project is to find the best community area, into a single data frame for better understanding of the relation among features and it is also very easy to apply clustering algorithm on a single data frame for the better visualization of clusters and features.

## 2.3 Feature Selection

Only certain features from the data frame are needed for the processing. Rest of the features are redundant for this project. Selecting the required features for the next data analysis is the crucial part. These features actually become 'the factors' that affect our results and the main goal of the project.

### Required Features

	Features Selected	Reasons for Selection
Neighbourhood Data Frame	Community Area Name, Neighbourhoods	Result feature
Census Data Frame	Crowded Housing, Households below poverty line, Unemployment Rate	To estimate the quality of living
School Data Frame	Type of Schools available	Education purpose
School data Frame	Longitude, Latitude	Geo location

### Final Data Frame

	Community_Area_Name	Neighbourhoods	Housing_Crowded_Percent	Below_Poverty_Percent	Unemployment_Percent	School_Type	Latitude	Longitude
0	ROGERS PARK	East Rogers Park	7.7	23.6	8.7	ES, ES, ES, ES, HS, ES	42.009665	-87.669949
1	WEST RIDGE	Arcadia Terrace, Peterson Park, West Rogers Park	7.8	17.2	8.8	ES, ES, ES, ES, ES, ES, HS, ES, ES	42.003435	-87.697500
2	UPTOWN	Buena Park, Argyle Street, Margate Park, Sheridan...	3.8	24.0	8.9	ES, ES, ES, ES, HS, ES, ES	41.964346	-87.656053
3	LINCOLN SQUARE	Ravenswood, Ravenswood Gardens, Rockwell Crossing	3.4	10.9	8.2	ES, ES, ES, HS, ES	41.975867	-87.683254
4	NORTH CENTER	Horner Park, Roscoe Village	0.3	7.5	5.2	HS, HS, ES, HS, ES, ES, ES	41.946617	-87.691056
5	LAKE VIEW	Boystown, Lake View East, Greeland West South	1.1	11.4	4.7	ES, ES, ES, ES, ES, ES, ES	41.945247	-87.670057

## 3 Methodology

### 3.1 K-means Clustering

k-means clustering is a method of vector quantization aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

### 3.2 Applying the Clustering Algorithm

K-means Clustering Algorithm is applied to partition the final data frame into k(=4) partitions (clusters). The data was normalized using the MinMaxScaler from the sklearn library in Python and then fed to the KMeans cluster object to be fitted.

### Cluster Labels

Out[25]:

	Cluster Labels	Housing_Crowded_Percent	Below_Poverty_Percent	Unemployment_Percent	Community_Area_Name
0	0	3.514286	24.628571	19.035714	SOUTH SHORE, CHATHAM, AVALON PARK, SOUTH CHICA...
1	1	5.864286	18.764286	13.464286	EDISON PARK, NORWOOD PARK, JEFFERSON PARK, FOR...
2	2	5.641667	17.900000	15.883333	GARFIELD RIDGE, ARCHER HEIGHTS, WEST ELSDON, G...
3	3	4.184615	15.415385	7.807692	ROGERS PARK, WEST RIDGE, UPTOWN, LINCOLN SQUAR...
4	4	5.376190	28.842857	18.804762	NEAR WEST SIDE, NORTH LAWNSDALE, SOUTH LAWNSDALE...

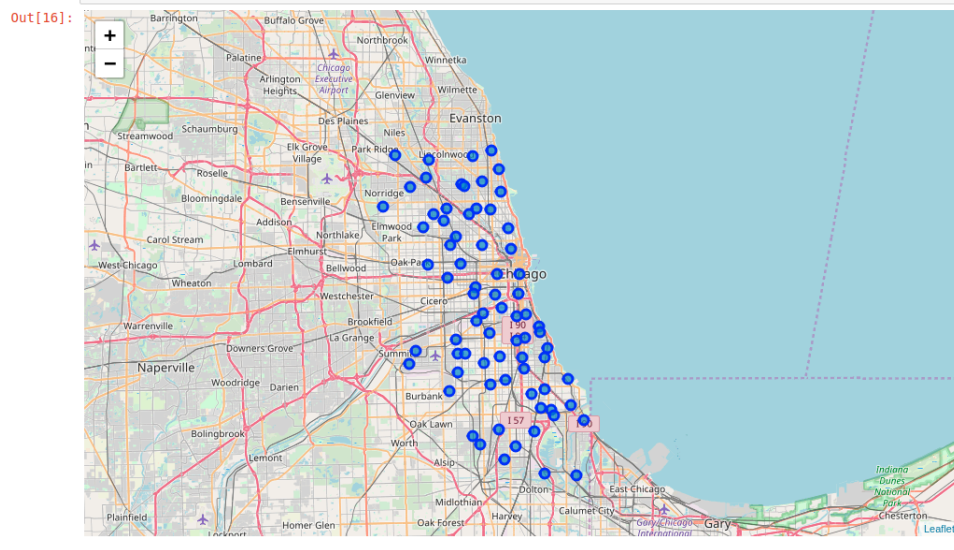
## 4 Results

### 4.1 K-means Clustering Results

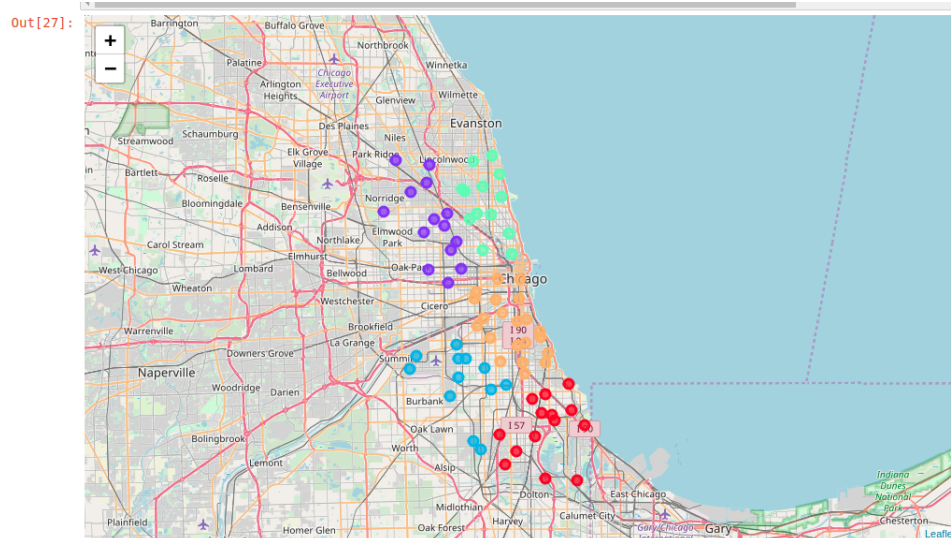
- Cluster 0
  - Features : Low percentage of crowded housing, High percentage of population below poverty, High percentage of unemployment rate
- Cluster 1
  - Features : High percentage of crowded housing , Medium percentage of population below poverty, Low percentage of unemployment rate
- Cluster 2
  - Features : High percentage of crowded housing , Low percentage of population below poverty, Medium percentage of unemployment rate
- Cluster 3
  - Features : Low percentage of crowded housing, Low percentage of population below poverty, Low percentage of unemployment rate
- Cluster 4
  - Features : Low percentage of crowded housing, High percentage of population below poverty, High percentage of unemployment rate

## 4.2 Visualization Maps

- Community Areas



- Clustered Community Areas



## 5 Discussion and Observations

- Correlation between features

Out[26]:

	Cluster Labels	Housing_Crowded_Percent	Below_Poverty_Percent	Unemployment_Percent
Cluster Labels	1.000000	0.317797	0.146386	-0.209088
Housing_Crowded_Percent	0.317797	1.000000	-0.078955	0.028639
Below_Poverty_Percent	0.146386	-0.078955	1.000000	0.840487
Unemployment_Percent	-0.209088	0.028639	0.840487	1.000000

- There is a negative correlation between percentage of crowded housing and percentage of population below poverty.

- 
- Highest correlation between percentage of unemployment and percentage of population below poverty can be observed from the above table.

## 6 Conclusion

In this project, I analyzed the factors i.e. that decide the best place in Chicago to live in. I implemented K-means clustering and clustered community areas into 5 different clusters according the factors mentioned above. After analyzing the best cluster is comprised of the below Community areas and their respective Neighbourhoods. So, the best places to live in Chicago according to the factors considered are :

- Community Areas - Rogers Park, West Ridge, Up Town, Lincoln Square, North Center, Lake View, Lincoln Park, Near North Side, North Park, Albany Park, Avondale, West Town, Edgewater
- Neighbourhoods - East Rogers Park, Arcadia Terrace, Peterson Park, West Rogers Park, Buena Park, Argyle Street, Margate Park, Sheridan Park, Ravenswood, Ravenswood Gardens, Rockwell Crossing, Horner Park, Roscoe Village, Boystown, Lake View East, Graceland West, South East Ravenswood, Wrigleyville, Old Town Triangle, Park West, Ranch Triangle, Wrightwood Neighbors, Sheffield Neighbors, Cabrini–Green, The Gold Coast, Goose Island, Magnificent Mile, Old Town, River North, River West, Streeterville, Brynford Park, Hollywood Park, River's Edge, Sauganash Woods, Mayfair, North Mayfair, Ravenswood Manor, Belmont Gardens, Chicago's Polish Village, Kosciuszko Park, East Village, Noble Square, Polish Downtown, Pulaski Park, Smith Park, Ukrainian Village, Wicker Park, Andersonville, Edgewater Glen, Edgewater Beach, Magnolia Glen, Lakewood/Balmoral