Settling Down in Chicago

Aribet21

For IBM Data Science Course

Table of Contents

1. Introduction

Suppose you have a friend who, due to his change of job, will be moving to Chicago but know little of the city. He is of upper-middle class and has a family of four, he and his wife, together with their two young kids.

You are going to give him advice on **which community area to settle down in Chicago**. After a discussion with your friend, you both agree that this community should meet 2 requirements,

1) **Safe**. Since we all know that Chicago is by no means a safe city, so it is the first thing that we would consider.
2) **Relaxed**. As we are looking for a community to live a life, this community should provide a calm and relax environment, and of course with sufficient venues to support daily life, such as dry cleaning, restaurants, playgrounds for kids, etc.

In this project, we are going to use data science knowledge to sort out the ideal community(s) for your friend to settle down in Chicago based on the 2 criteria.

2. Data

- Regarding safety

There are 77 communities in Chicago. In order to get the information of the safety status for each community, we can look for statistics in https://data.cityofchicago.org/ . There is **a dataset recording each incident of crime that occurred in Chicago** from 2001 to present, and for simplicity, We just downloaded the subset for year 2018.

But the above dataset only has community areas in numerical form. In order to get the names for each community, we will have to scape a Wikipedia page to match the numbers with the names. The webpage is https://en.wikipedia.org/wiki/Community_areas_in_Chicago .

With these statistics in hand, we could solve the problem of finding safe communities. **Let's just define that the communities which have crime incidents less than the average of Chicago is safe**.

- Regarding relaxation

As for the second criterion, we'll turn to https://foursquare.com/ to **segment the safe communities into 3 clusters based on the similarity of venues**. For example, the cluster we're looking for should have venue categories like parks, fields, restaurants, dry cleaners, etc in a high ocurrence.

Here, we'll use the **K-Means Clustering** algorithm of machine learning to find out each cluster's characteristics and to decide which community(s) to recommend.

3. Methodology

As discussed in the Data section, we will **divide our analysis into 2 parts**. In the first part, we sort out the safe communities in Chicago use *Pandas*, *BeautifulSoup4* and other Python libraries. In the second part, we cluster these safe communities using *Foursquare API* and *k-means* algorithm. Below, we'll explain in more detail.

3.1 Sorting out the safe communities

Here, we'll combine the Chicago crime records downloaded from https://data.cityofchicago.org/ and Chicago community areas scraped from https://en.wikipedia.org/wiki/Community_areas_in_Chicago to form a *Pandas* dataframe indicating each community's crime occurrence.

We define the safe communities as those have less crimes reported than the average of Chicago.

3.1.1 Data acquisition

Download the records of Chicago Crime incidents of year 2018 and upload it to the console as a *csv* file named 'Crimes_-_2018.csv'.

Table 1

	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	Ward	Community Area	FBI Code	Coordina
0	11556487	JC104662	12/31/2018 11:59:00 PM	112XX S SACRAMENTO AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET	False	False	2211	22	19.0	74.0	14	1158309
1	11561837	JC110056	12/31/2018 11:59:00 PM	013XX W 72ND ST	1153	DECEPTIVE PRACTICE	FINANCIAL IDENTITY THEFT OVER $ 300	NaN	False	False	734	7	6.0	67.0	11	1168573
2	11552699	JC100043	12/31/2018 11:57:00 PM	084XX S SANGAMON ST	1310	CRIMINAL DAMAGE	TO PROPERTY	APARTMENT	False	False	613	6	21.0	71.0	14	1171454
3	11552724	JC100006	12/31/2018 11:56:00 PM	018XX S ALLPORT ST	0440	BATTERY	AGG: HANDS/FIST/FEET NO/MINOR INJURY	OTHER	True	False	1233	12	25.0	31.0	08B	1168327
4	11552731	JC100031	12/31/2018 11:55:00 PM	078XX S SANGAMON ST	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	False	False	621	6	17.0	71.0	08B	1171332

As we can see, each incident is recorded as a row and there are many columns. But actually, what we need here is simply the crime counts of each community. So, we'll do some data wrangling.

3.1.2 Data wrangling

After dropping these unnecessary columns and group the dataset by 'Community Areas' and Convert the data type of 'Community Areas' to integer, we get this:

Table 2

	Community Areas	Crime Counts
0	1	2706
1	2	646
2	3	723
3	4	3082
4	5	1505

Table 3

	Community Areas	Crime Counts
count	77.000000	77.000000
mean	39.000000	3411.272727
std	22.371857	3003.399708
min	1.000000	233.000000
25%	20.000000	1145.000000
50%	39.000000	2251.000000
75%	58.000000	4835.000000
max	77.000000	14750.000000

Notice there's a problem within this dataframe, it only provides each community's code but don't provide its name. That’s why we’re going to scrape the Wikipedia page and fill in each community code's corresponding name.

3.1.3 Web scraping

We use *BeautifulSoup4* to scrape the tables from Wikipedia page and here's the result:

Table 4

	Community Areas	Community Names	Neighborhoods
0	None	None	None
1	08	Near North Side	Cabrini–Green\nThe Gold Coast\nGoose Island\nM...
2	32	Loop	Loop\nNew Eastside\nSouth Loop\nWest Loop Gate
3	33	Near South Side	Dearborn Park\nPrinter's Row\nSouth Loop\nPrai...
4	None	None	None

Because this dataframe is formed from several tables on the webpage, it contains some blank rows. Also, we do not need the column 'Neighborhoods' in our case, again, we'll do some cleaning and merge it with the Table 2.

Table 5

	Community Areas	Crime Counts	Community Names
0	1	2706	Rogers Park
1	2	646	West Ridge
2	3	723	Uptown
3	4	3082	Lincoln Square
4	5	1505	North Center
5	6	2017	Lake View
6	7	2718	Lincoln Park
7	8	1971	Near North Side
8	9	3658	Edison Park
9	10	3569	Norwood Park

3.1.4 Adding geocode

Now that this dataframe is neat and tidy, the next thing we're going to do is to get each community's geocode, so that we can point them on a map. Here we use

geocoder to get each community's geocode, and add it to Table 5, the results are as below.

Table 6

	Community Areas	Crime Counts	Community Names	Latitude	Longitude
0	1	2706	Rogers Park	42.008820	-87.666180
1	2	646	West Ridge	41.999480	-87.692660
2	3	723	Uptown	41.981230	-87.660000
3	4	3082	Lincoln Square	41.975700	-87.689140
4	5	1505	North Center	41.954110	-87.681420
5	6	2017	Lake View	41.939820	-87.656820

3.1.5 Sort out the safe communities

As we discussed before, in this case, we define the safe communities as those have less crimes reported than the average of Chicago. So, from the above dataframe, we'll narrow down the community candidates to those who fit our definition.

Table 7

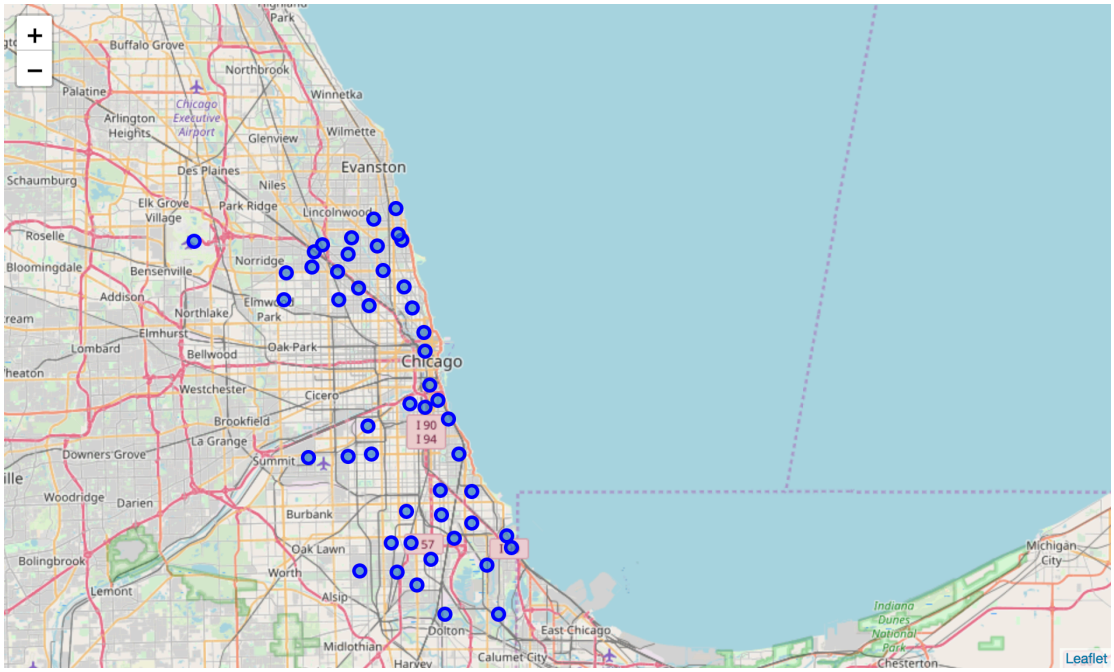
```
Community Areas    49
Crime Counts       49
Community Names    49
Latitude           49
Longitude          49
dtype: int64
```

It turns out **there are 49 communities**.

3.1.6 Show the community candidates on the map

Here we use the *Folium* library to visualize the safe communities in Chicago.

Map 1



So far, we have finished our first part of data analysis: sorting out the safe communities.

3.2 Clustering these safe communities

In the second part, we want to cluster our community candidates based on their similarity of environment. **We define the environment by the categories of venues each community has. And we can further assume that the communities falling to the same cluster share the similar categories of venues**.

3.2.1 Explore each community candidate

Here, we will use the *Foursquare API* to explore community candidates in Chicago, to get the most common venue categories in each community.

First, we would like to get the top 100 venues that are in each community candidates within a radius of 500 meters. And then group them by Community Names, to check how many venues are returned for each community candidate, the results are shown below.

Table 8

Community Names	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Albany Park	16	16	16	16	16	16
Armour Square	32	32	32	32	32	32
Auburn Gresham	20	20	20	20	20	20
Avondale	4	4	4	4	4	4
Beverly	8	8	8	8	8	8
Bridgeport	15	15	15	15	15	15
Brighton Park	12	12	12	12	12	12

Second, we'll check out the frequency of occurrence of each venue category in each community candidate.

Table 9

	Community Names	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Amphitheater	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store
0	Albany Park	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	Armour Square	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	Auburn Gresham	0.000000	0.00	0.000000	0.000000	0.000000	0.050000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Avondale	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	Beverly	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Then, get the top 10 venues for each community candidate.

Table 10

	Community Names	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	Liquor Store	Asian Restaurant	Cosmetics Shop	Dance Studio	Park	Paper / Office Supplies Store	Bus Stop	Light Rail Station	Bar	Bank
1	Armour Square	Bus Station	Chinese Restaurant	Food Truck	Park	Diner	Dance Studio	Discount Store	Donut Shop	Rental Car Location	Road
2	Auburn Gresham	Pizza Place	Deli / Bodega	Intersection	Playground	Pharmacy	Park	Record Shop	Optical Shop	Chinese Restaurant	Sandwich Place
3	Avondale	Wings Joint	Mexican Restaurant	Financial or Legal Service	Pizza Place	Donut Shop	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Exhibit
4	Beverly	Park	Boutique	Donut Shop	Lounge	Fast Food Restaurant	Dry Cleaner	Financial or Legal Service	Filipino Restaurant	Field	Farmers Market

3.2.2 Cluster communities

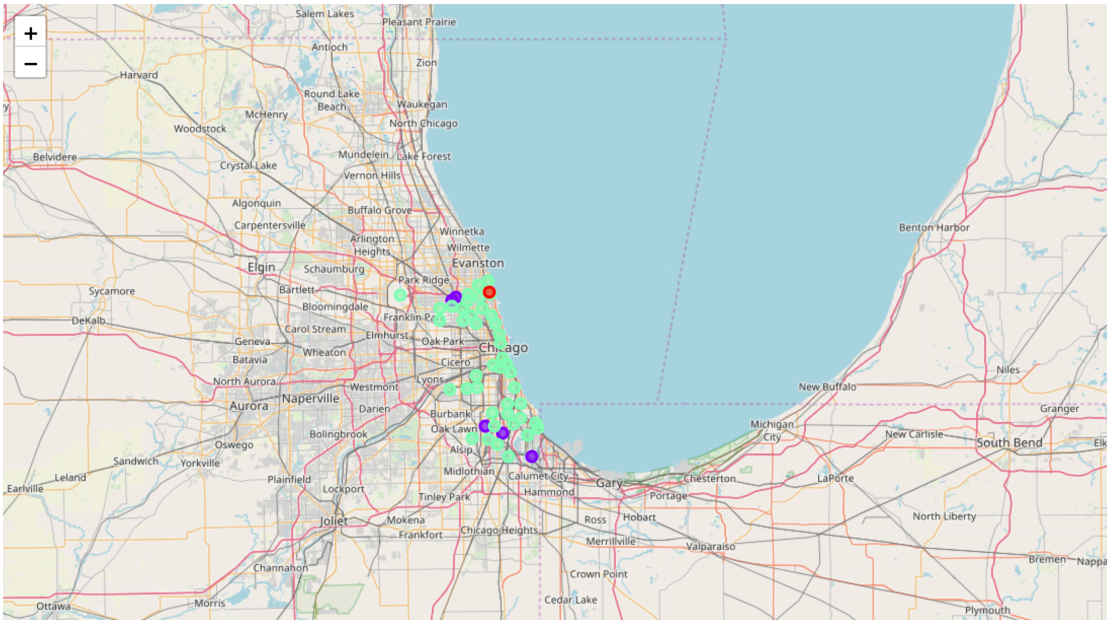
We use *k-means* algorithm to segment the community candidates, with the feature of most common venue categories in each community, and based on our sample size (49), we decide to **set the number of clusters to 3**.

Table 11

	Community Areas	Crime Counts	Community Names	Latitude	Longitude	Cluster Labels	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
0	1	2706	Rogers Park	42.00882	-87.66618	2	Theater	Pizza Place	American Restaurant	Bar	ATM	Gas Station	Mexican Restaurant	Sandwich Place	Café
1	2	646	West Ridge	41.99948	-87.69266	2	Indian Restaurant	Pakistani Restaurant	Grocery Store	Fast Food Restaurant	Football Stadium	Donut Shop	Market	Fruit & Vegetable Store	Clothing Store
2	3	723	Uptown	41.98123	-87.66000	2	Pizza Place	Sandwich Place	Sushi Restaurant	Asian Restaurant	Vietnamese Restaurant	Bus Station	Chinese Restaurant	Coffee Shop	Theater
3	4	3082	Lincoln Square	41.97570	-87.68914	2	Bus Station	Bar	Café	Convenience Store	Pizza Place	Korean Restaurant	Sandwich Place	Liquor Store	Food & Drink Shop
4	5	1505	North Center	41.95411	-87.68142	2	Bar	Coffee Shop	Bank	Mobile Phone Shop	Boutique	American Restaurant	Dive Bar	Pub	Pharmacy

As we can see, the dataframe now has a Cluster Label to each community. we'll display them on a map.

Map 2



As the map above shows, we now have 3 clusters of communities. In the next part, we will examine each cluster.

4. Results

Let's now take a look at the 3 clusters.

4.1 Cluster 1

Table 12

	Crime Counts	Community Names	Cluster Labels	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
76	233	Edgewater	0	Intersection	Chinese Restaurant	Yoga Studio	Donut Shop	Financial or Legal Service	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Exhibit

As we can see, cluster 1 only contains 1 community. Based on its Top 10 Most Common Venue Categories, it looks like a community with a slow-paced life, as venues like Yoga Studio, Farmers Market, Exhibit are very common, although the 1st common venue is intersection, which may seem a little awkward.

4.2 Cluster 2

Table 13

	Crime Counts	Community Names	Cluster Labels	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
10	1093	Jefferson Park	1	Neighborhood	Park	Theater	Yoga Studio	Donut Shop	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Exhibit
11	474	Forest Glen	1	Park	Mexican Restaurant	Bar	Yoga Studio	Dry Cleaner	Financial or Legal Service	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market
48	1085	Roseland	1	Park	Gas Station	Liquor Store	Clothing Store	Seafood Restaurant	Dog Run	Field	Fast Food Restaurant	Farmers Market	Exhibit
54	2224	Hegewisch	1	Park	Food & Drink Shop	Bus Station	Yoga Studio	Donut Shop	Filipino Restaurant	Field	Fast Food Restaurant	Farmers Market	Exhibit
71	3113	Beverly	1	Park	Boutique	Donut Shop	Lounge	Fast Food Restaurant	Dry Cleaner	Financial or Legal Service	Filipino Restaurant	Field	Farmers Market

Cluster 2 looks exactly the type of community we are looking for. There are venues like parks, dry cleaners, fields, farmers markets, exhibits and of course restaurants, which perfectly meet our requirement of a relaxed, laid-back environment, and the parks and fields are 'must-have's for family with kids.

4.3 Cluster 3

Table 14

	Crime Counts	Community Names	Cluster Labels	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	2706	Rogers Park	2	Theater	Pizza Place	American Restaurant	Bar	ATM	Gas Station	Mexican Restaurant	Sandwich Place	Café	Donut Shop
1	646	West Ridge	2	Indian Restaurant	Pakistani Restaurant	Grocery Store	Fast Food Restaurant	Football Stadium	Donut Shop	Market	Fruit & Vegetable Store	Clothing Store	Juice Bar
2	723	Uptown	2	Pizza Place	Sandwich Place	Sushi Restaurant	Asian Restaurant	Vietnamese Restaurant	Bus Station	Chinese Restaurant	Coffee Shop	Theater	Mexican Restaurant
3	3082	Lincoln Square	2	Bus Station	Bar	Café	Convenience Store	Pizza Place	Korean Restaurant	Sandwich Place	Liquor Store	Food & Drink Shop	Karaoke Bar
4	1505	North Center	2	Bar	Coffee Shop	Bank	Mobile Phone Shop	Boutique	American Restaurant	Dive Bar	Pub	Pharmacy	Yoga Studio
5	2017	Lake View	2	Café	Japanese Restaurant	Bakery	Gym / Fitness Center	Pizza Place	Bagel Shop	Coffee Shop	Performing Arts Venue	Clothing Store	Sports Bar
6	2718	Lincoln Park	2	Pizza Place	Sandwich Place	Coffee Shop	Bar	Taco Place	Breakfast Spot	Fast Food Restaurant	Art Gallery	Mexican Restaurant	American Restaurant
7	1971	Near North Side	2	Gym / Fitness Center	Gym	Restaurant	American Restaurant	Coffee Shop	Cycle Studio	Breakfast Spot	Sandwich Place	Café	Pub
12	989	North Park	2	Coffee Shop	Pharmacy	Theater	Video Store	Bar	Food Truck	Sushi Restaurant	Fried Chicken Joint	Supermarket	Park
13	2395	Albany Park	2	Liquor Store	Asian Restaurant	Cosmetics Shop	Dance Studio	Park	Paper / Office Supplies Store	Bus Stop	Light Rail Station	Bar	Bank
14	3183	Portage Park	2	Bus Station	Breakfast Spot	Bike Rental / Bike Share	Bakery	Sandwich Place	Asian Restaurant	Coffee Shop	Food Truck	Radio Station	Convenience Store
15	2794	Irving Park	2	Discount Store	Accessories Store	Mexican Restaurant	Taco Place	Fried Chicken	Snack Place	Fast Food Restaurant	Pet Store	Seafood Restaurant	Mobile Phone Shop

The remaining 43 communities all belong to Cluster 3 according to *k-means* algorithm. Table 14 just takes a snap of these communities. By first look, it is a bit chaotic, but if we look closely, we could still find some common features in these communities. Most of them give us the impression of a fast-paced city life. For example, the most common venues include coffee shops, fast food restaurants, pizza places, bus stations, etc. They all remind me of the hustle bustle of downtown areas and CBDs. So, I don't think we'll recommend the communities in Cluster 3 to our friend.

5. Discussion

Based on the observations above, **we'll recommend a community from Cluster 2**. These communities perfectly meet his requirements of safety and relaxation.

But exactly which one to choose is still up to the friend's preference. Say, if he thinks safety is the first priority, then Forest Glen with the least crime records should be his choice. Or if he works in south Chicago and would like to settle his family close by, then he should choose from Roseland, Hegewisch or Beverly.

But as we discussed in the previous part, Cluster 1 which only includes 1 community called Edgewater could also be an option. Especially if the friend want to live near lake.

6. Conclusion

In this project, we firstly narrow down our choices by select those safe communities based on crime records. Secondly, we use the *Foursquare API* to explore these communities and then use *k-means* algorithm to group the communities into clusters based on the feature of most common venues.

We obtain 3 clusters and according to our observations, we recommend Cluster 2 to our friend, which meet his requirements the most.

But, as we can see, there are still some issues regarding to this segmentation. For instance, Cluster 1 and 2 seems alike in terms of their most common venues and within Cluster 3 we can see that there are some communities more similar to Cluster 2. I think this is because the venue categories in *Foursquare* do not always correctly reflect our intension.

So maybe for future improvement, we should explicitly hand-pick the venue categories that we need to solve the problem and try different clustering algorithms as well to get a better result.