Capstone Project- Battle of Neighborhoods

Understand Chicago Community Areas

Hung Lee
June 2019

Data

To support the analysis from two different perspectives on Chicago's community areas, two sets of data are required:

- Census data with social economic indices for the community areas
- Neighborhood venues for the community areas

Census Data

The latest census data is available in Chicago Data Portal (https://data.cityofchicago.org/) at https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2. This data set is generated a few years ago yet It is sufficient for this study. This dataset includes the following columns:

- Community Area Number
- Community Area Name
- Percent of housing crowded
- Percent household below poverty
- Percent aged 16+ unemployed
- Percent aged 25+ without high school diploma
- Percent aged under 18 or over 64
- Per capita income
- Hardship index

Below is a snapshot of this dataset:

Community Area Number	COMMUNITY AREA NAME	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER CAPITA INCOME	HARDSHIP INDEX
1	Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39
2	West Ridge	7.8	17.2	8.8	20.8	38.5	23040	46
3	Uptown	3.8	24	8.9	11.8	22.2	35787	20
4	Lincoln Square	3.4	10.9	8.2	13.4	25.5	37524	17
5	North Center	0.3	7.5	5.2	4.5	26.2	57123	6
6	Lake View	1.1	11.4	4.7	2.6	17	60058	5
7	Lincoln Park	0.8	12.3	5.1	3.6	21.5	71551	2

Neighborhood Venues

This dataset is obtained via FourSquare APIs.

Firstly, the community names from the previous census dataset are used to retrieve the geo coordinates of the community areas. The FourSquare APIs are then invoked using the geo locations to retrieve the neighborhood venues for each community area.

The data is then transformed so that each community area has a row with scores under each venue categories.

Data Cleansing, Manipulation, and Transformation.

This census dataset requires very minimum cleansing as it has been well prepared; it is ready for use. The only updates required are to fix typos of two community names where the wrong names have prevented the retrieval of their geo location coordinates. Also, the “Community Area Number” is not needed in the study thus can be removed.

The dataset for neighborhood venues are obtained as JSON file by calling the FourSquare APIs. The JSON file is manipulated and the required data is extracted and put into a Pandas data frame for further analysis.

Use of Data

These two datasets are the basis of this analysis. Each set is clustered into different number of groups and the results are analyzed in their own dimensions separately. Later, the two sets of clusters are combined and correlated to see if there are additional insights.