Capstone Project- Battle of Neighborhoods

Understand Chicago Community Areas

Hung Lee
June 2019

Introduction

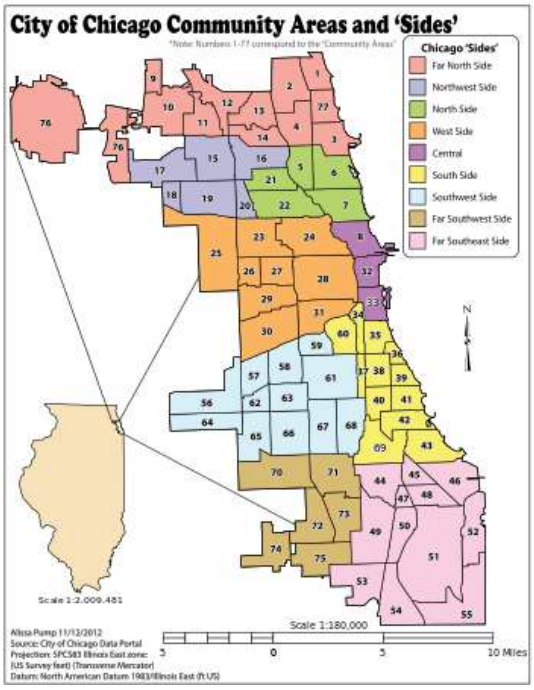
This project studies the ***Chicago Community Areas*** from two different perspectives: social economic indices and neighborhood venues.

Background

As described in https://en.wikipedia.org/wiki/Community_areas_in_Chicago, the ***community areas*** in Chicago are defined by the Social Science Research Committee at the University of Chicago in early 1920s; they have since been used by the City of Chicago for statistical and planning purposes. Census data are tied to the community areas.

Chicago ***neighborhoods*** are loosely mapped to the community areas. In most cases, there is one-to-one mapping between a *neighborhood* and a *community area*. We use the latter since official census data is available for community areas, not for neighborhoods. In this report, we'll use the terms interchangeably as Chicagoans typically do.

Below is a map extracted from the above Wikipedia showing the 77 *community areas, or loosely speaking, the neighborhoods,* in Chicago:



Problem Statement

In this study, we explore the community areas from two different perspectives:

- Social economic indices, and
- Neighborhood venues

The community areas are grouped into different clusters with different characteristics along these two dimensions. In other words,

- first of all, we group the community areas based on their social economic indices,
- secondly, we group the community areas based on the neighborhood venues,
- lastly, we combine the information learned from these two clustering exercises and correlate the above clustering results to see if there are any useful insights.

The study means to promote better understanding on the city of Chicago and hopefully reveal information that may not be as obvious on each community area as well as groups of community areas.

Use of Results

The audience of this report include, but not limited to

- Business owners
- Future residents or home owners

With better understanding of the community areas, business owners can understand their customers better thus have more effective marketing strategies, and the future residents or home owners can make better choices on which area they want to live.

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, and
- 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 but It is adequate for this study.

This dataset includes the following columns (i.e., "features"):

- 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 the FourSquare API.

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

The data is then transformed into a data frame where each community area has a row with scores under each venue categories. This is then used for clustering.

Data Cleansing, Manipulation, and Transformation.

The ***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 in two community names that have prevented the retrieval of their geo location coordinates. Also, the “Community Area Number” is not needed in the study thus is removed.

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

In addition to generate a data frame suitable for clustering, we also create several csv files that can be used directly for various queries:

- "neighborhood_venues.csv": this file lists all the venues for each community area as retrieved via the FourSquare API.
- "neighborhood_venue_count.csv": this file counts the number of venues in each community area; in addition to knowing the venue categories, we also know the number of venues.
- "neighborhood_v_category_count.csv": this file counts the number of venues in each category for each community area; this provide similar but more refined information than the above.

Please refer to the Jupyter Notebook for the step-by-step generation and transformation in the section of "Data Preparation".

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.

Methodology

Algorithm

k-means is used to cluster the community areas into different clusters for both datasets. The ***KMeans*** module in ***sklearn*** library is used in this project.

Features

For the ***social economic indices clustering***, all the features in the dataset are used, namely the following:

- *Percent of housing crowded* (i.e. living condition)
- *Percent household below poverty* (i.e., Poverty level)
- *Percent aged 16+ unemployed* (i.e., unemployment rate)
- *Percent aged 25+ without high school diploma* (i.e., education level)
- *Percent aged under 18 or over 64* (i.e., population not working)
- *Per capita income*
- *Hardship index* (this is an index calculated by the census organization using other indices)

For the ***neighborhood venue clustering***, the features are the venue categories. This is the data frame generated from the "one hot" conversion. All the venue categories are used as features.

There are ***252*** unique venue categories. Instead of listing all, a snapshot from the Jupyter notebook is shown below. One may refer to the notebook to see the complete list:

```
#chek the types of venues in chicago area
chicago_venues['Venue Category'].unique()
```

```
8]: array(['Mexican Restaurant', 'Grocery Store', 'Pet Store', 'Coffee Shop',
       'Farmers Market', 'Bar', 'Bakery', 'American Restaurant',
       'Deli / Bodega', 'Pizza Place', 'Chinese Restaurant',
       'Asian Restaurant', 'Diner', 'Sandwich Place', 'Sushi Restaurant',
       'Theater', 'Dive Bar', 'Breakfast Spot', 'Discount Store',
       'Train Station', 'Donut Shop', 'Performing Arts Venue', 'Park',
       'Fried Chicken Joint', 'Convenience Store', 'Bus Station',
       'Fast Food Restaurant', 'Climbing Gym', 'General Entertainment',
       'Vietnamese Restaurant', 'Ethiopian Restaurant', 'Tattoo Parlor',
       'Jazz Club', 'Lounge', 'Music Venue', 'Yoga Studio',
       'Big Box Store', 'Thai Restaurant', 'Concert Hall',
       'Mobile Phone Shop', 'Korean Restaurant', 'Salon / Barbershop',
       'Wings Joint', 'Smoke Shop', 'Bank', 'Automotive Shop',
       'Mediterranean Restaurant', 'Pharmacy', 'Bistro', 'Wine Bar',
       'New American Restaurant', 'Speakeasy', 'Hockey Arena',
       'Rock Club', 'Convention Center', 'Pedestrian Plaza', 'Gym',
       'Café', 'Building', 'Beer Store', 'Beer Garden',
       'German Restaurant', 'Pub', 'Brewery', 'Bowling Alley',
       'Restaurant', 'Burger Joint', 'Shipping Store',
       'Turkish Restaurant', 'Dance Studio', 'Office', 'Boutique', 'Spa',
       'Seafood Restaurant', 'Gym / Fitness Center', 'Video Store',
```

Selecting the right k's

Prior to the analysis, it is not clear how many clusters make the best sense for either the datasets. Should the community areas be grouped in to 3 clusters? 4, 5, or even 10? Without diving into the analysis, it is difficult to tell. As a result, different ***k*** values are tried in the Jupyter notebook and the results are analyzed before settling with a reasonable k.

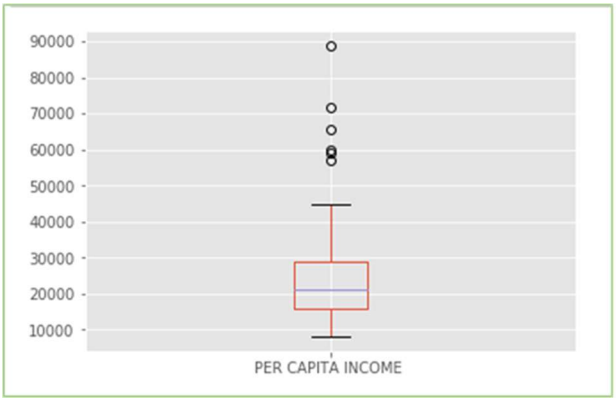
The k value for the two datasets are selected independently.

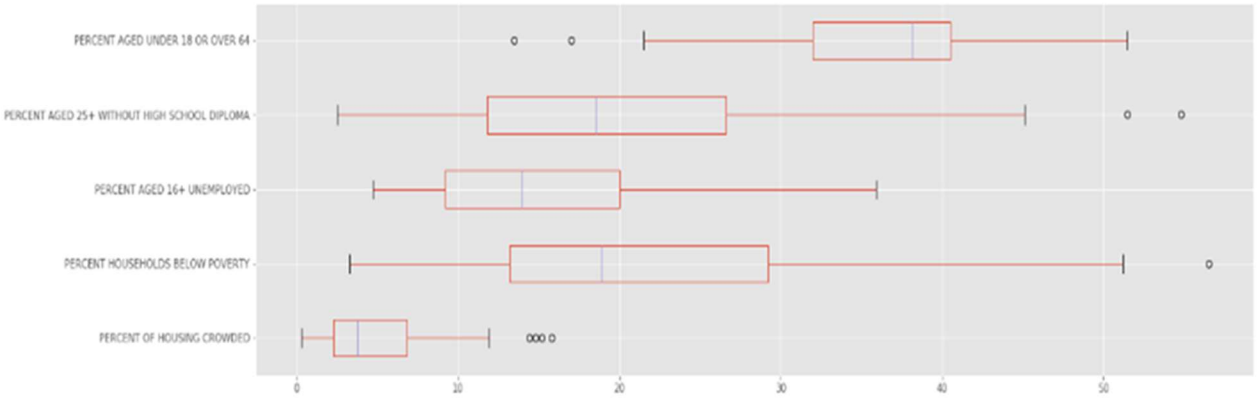
The census data is evaluated first. Below is the results running the ".describe()" method on the data frame:

Statistics of The Entire Dataset

	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
count	77	77	77	77	77	77	77
mean	4.92	21.77	15.37	20.34	35.75	25563	49.51
std	3.68	11.53	7.54	11.82	7.33	15293	28.69
min	0.3	3.3	4.7	2.5	13.5	8201	1
25%	2.3	13.2	9.2	11.8	32	15754	25
50%	3.8	18.9	13.9	18.5	38.1	21323	50
75%	6.8	29.2	20	26.6	40.5	28887	74
max	15.8	56.5	35.9	54.8	51.5	88669	98

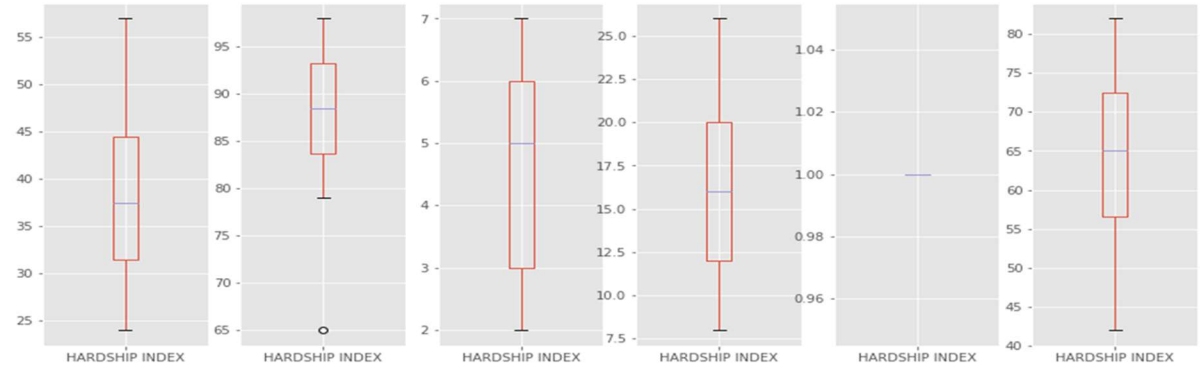
Box plots are also done for all these7 features





Looking at all the above, it seems the hardship index is well balanced with no outliers. It can be served as a tool to evaluate the choices of k along with other statistics.

For example, with k=6, we have the following results:



Census_Group	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
0	3.75	14.25	11.50	17.95	37.25	24395.0	37.5
1	7.50	34.15	20.40	27.55	41.60	12343.0	88.5
2	1.10	12.30	5.10	3.60	21.50	60058.0	5.0
3	2.30	14.70	8.40	9.70	26.20	37524.0	16.0
4	1.90	12.90	7.00	2.50	22.60	88669.0	1.0
5	5.20	23.70	19.55	24.55	38.90	16930.5	65.0

As one can see, cluster 2 and 4 have very close hardship index and relatively close per capita income; cluster 4 also have only 1 entry. These probably can be combined into one. Cluster 1 and 5 are also closer and it may make sense to reduce the k value at least by 1 or probably 2.

Following this logic, we tried k = 3, 4, 5, 6, 7, and concluded that 4 is a reasonable k for grouping.

Clustering using different k values are also performed on the neighborhood venue data. Instead of looking at the box plots, the actual categories in each cluster are evaluated to see if we could identify and label a cluster properly. Manual examination of the neighborhood venue data is heavily used to select the right k. It turns out that majority of the neighborhood are put into one cluster with over 50% of the neighborhoods. This cluster doesn't not change much when k is set to any number larger than 3. With a larger number, other neighborhoods are simply refined into smaller clusters. With this observation, we decide that 3 is a reasonable k value for the neighborhood clustering. Also, one of the neighborhoods has no neighborhood venues. It's a cluster by itself. As a result, we also have 4 clusters in this exercise, although k-means uses k=3 to cluster the remaining 76 community areas. (This is shown in the Jupyter notebook.)

Showing results in the map

In both exercises, the community areas are shown on the Chicago maps with different colors. This helps to see where they are located and confirm with one's real experience. The "folium" library is used for the plotting.

Jupyter notebook

The methodology section is incomplete without mentioning about the Jupyter notebook – it is the lab, the scratch pad, and the ultimate tool used for this analysis.

Analysis Results

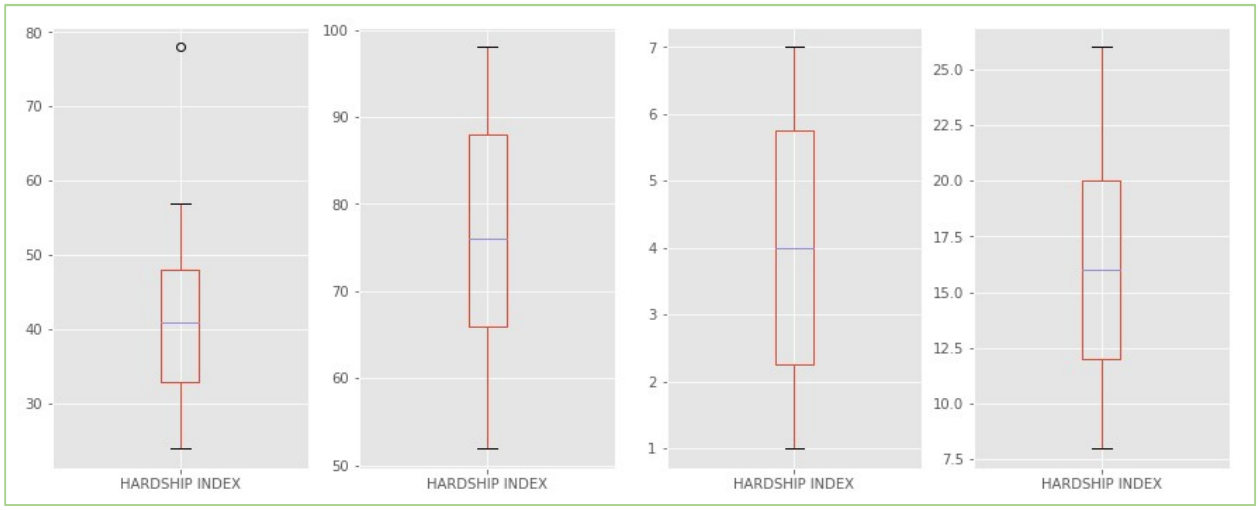
Social economic clustering

The community areas in Chicago are clustered into 4 different groups using the census dataset.

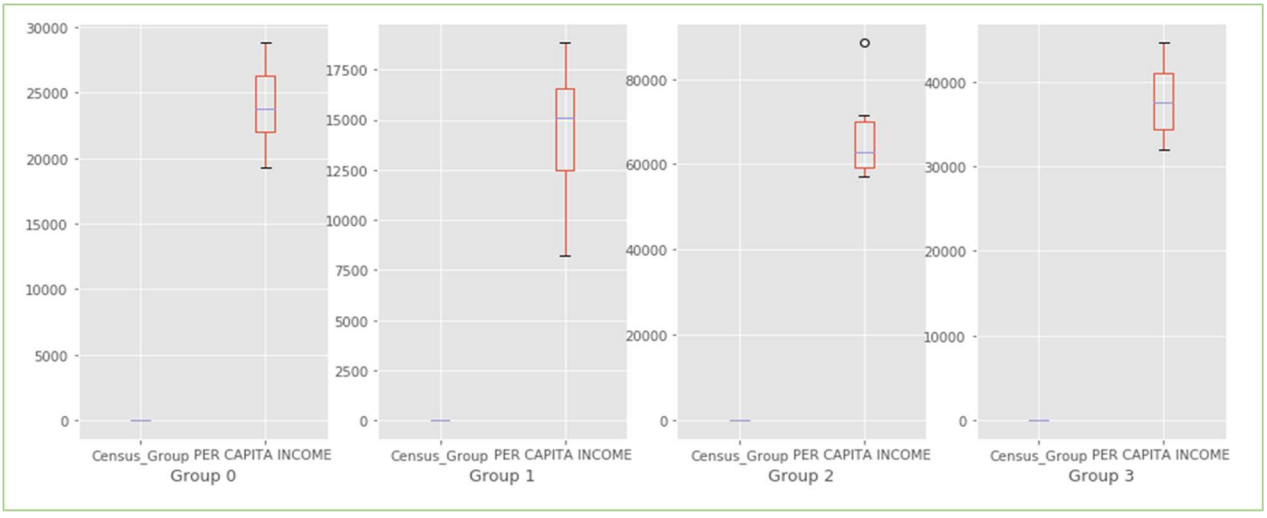
Below are the mean and box plots for the 4 groups (please note that the y-axis' in different plots have different scales/values):

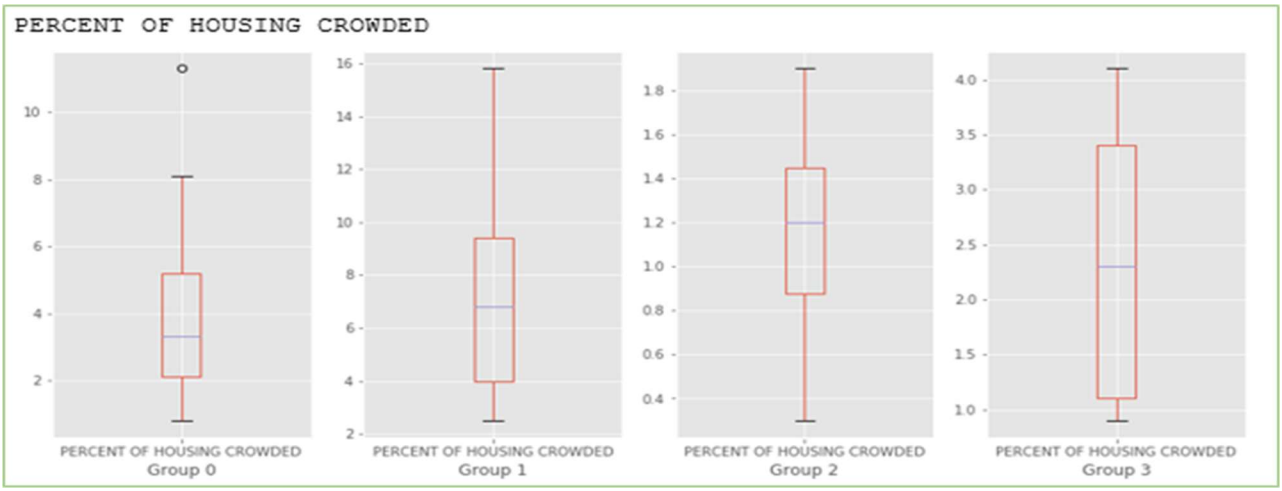
Census_Group	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER CAPITA INCOME	HARDSHIP INDEX
0	3.3	15.4	12.40	17.70	37.60	23791	41
1	6.8	29.0	19.40	27.60	40.30	15089	76
2	1.2	12.6	5.15	3.35	21.65	62792	4
3	2.3	14.7	8.40	9.70	26.20	37524	16

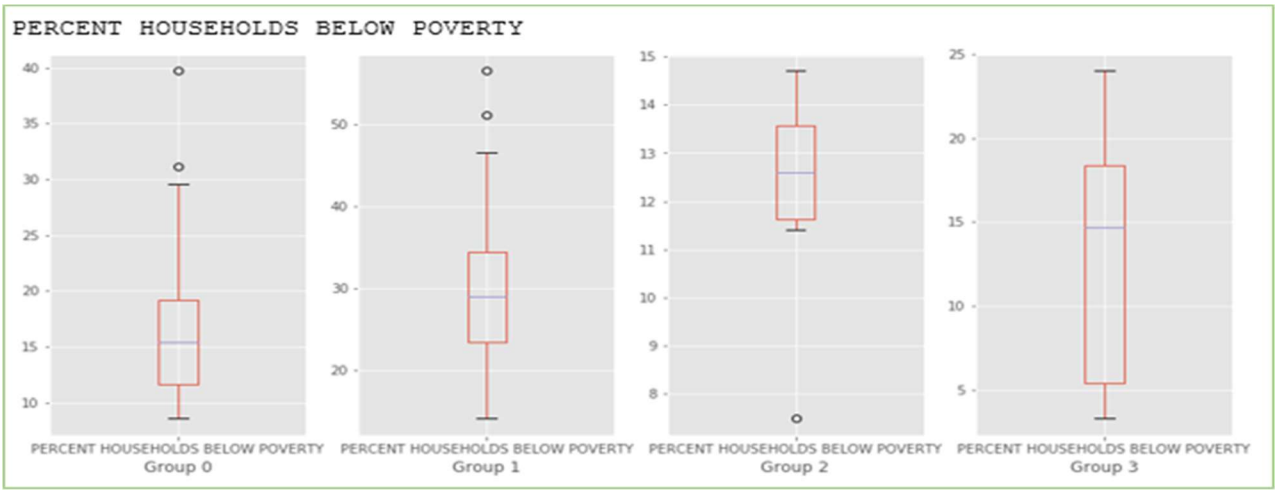
Hardship index:

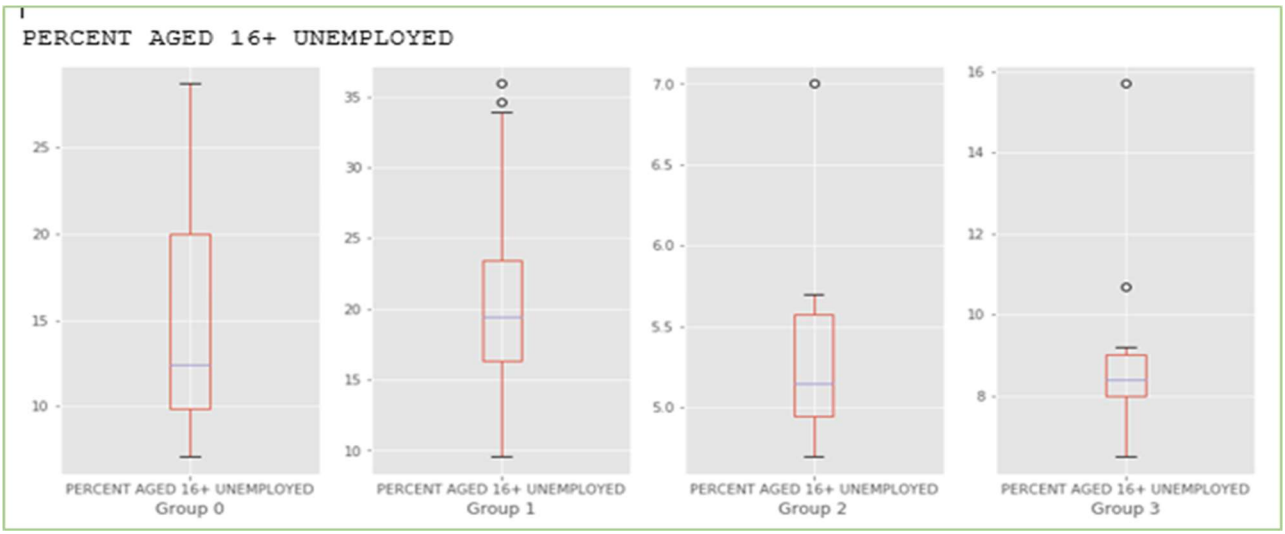


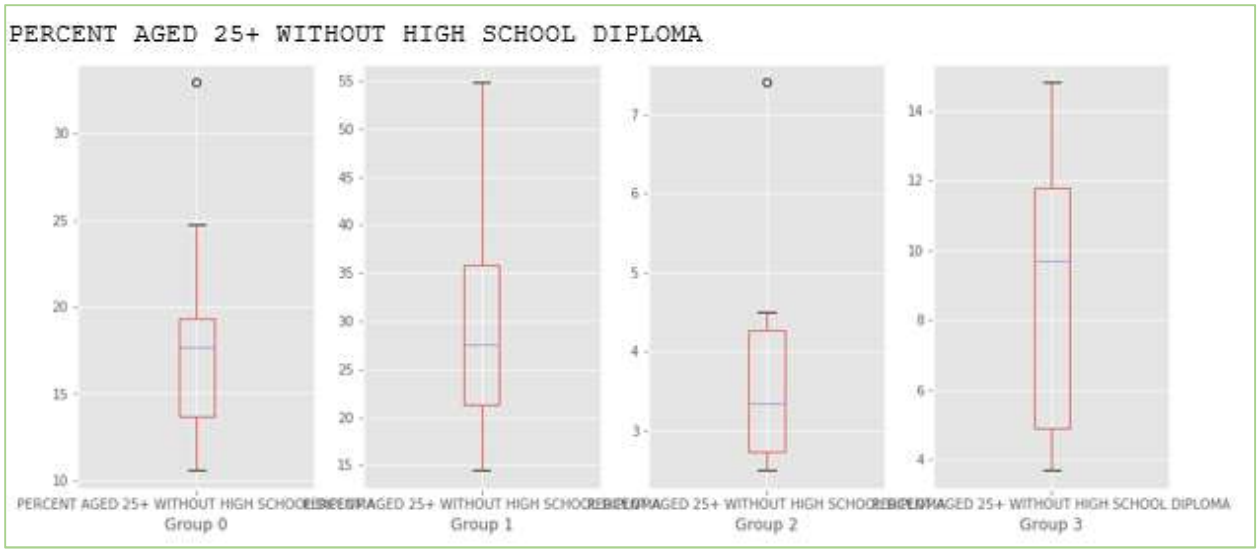
Per Capita Income:

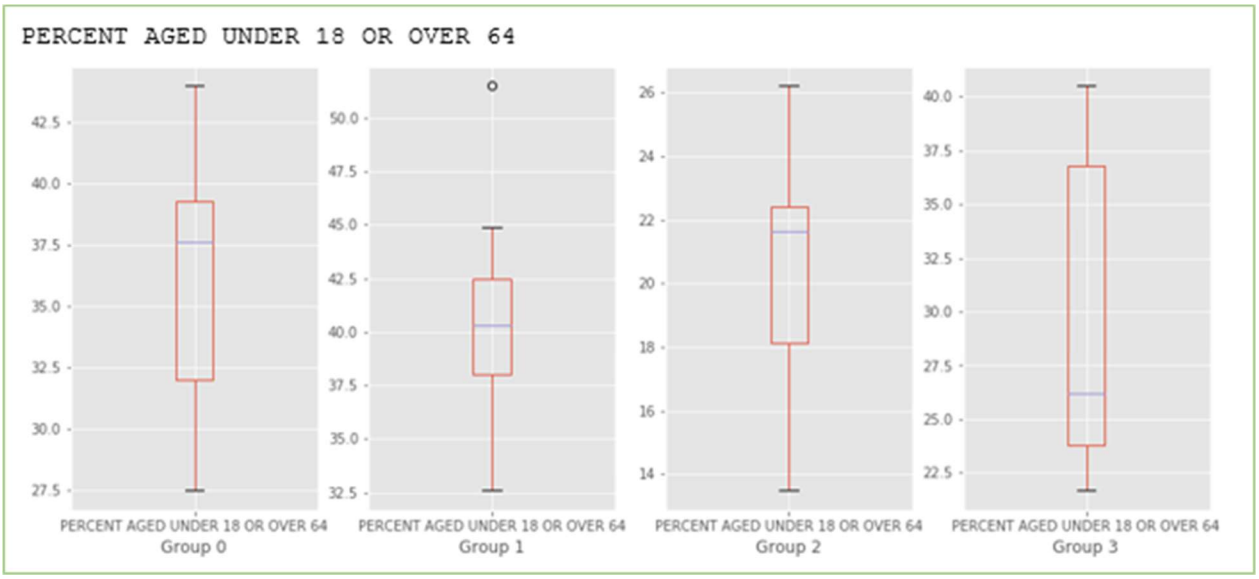



PERCENT OF HOUSING CROWDED
PERCENT OF HOUSING CROWDED
Group 0
PERCENT OF HOUSING CROWDED
Group 1
PERCENT OF HOUSING CROWDED
Group 2
PERCENT OF HOUSING CROWDED
Group 3


PERCENT HOUSEHOLDS BELOW POVERTY
PERCENT HOUSEHOLDS BELOW POVERTY
Group 0
PERCENT HOUSEHOLDS BELOW POVERTY
Group 1
PERCENT HOUSEHOLDS BELOW POVERTY
Group 2
PERCENT HOUSEHOLDS BELOW POVERTY
Group 3


PERCENT AGED 16+ UNEMPLOYED
PERCENT AGED 16+ UNEMPLOYED
Group 0
PERCENT AGED 16+ UNEMPLOYED
Group 1
PERCENT AGED 16+ UNEMPLOYED
Group 2
PERCENT AGED 16+ UNEMPLOYED
Group 3

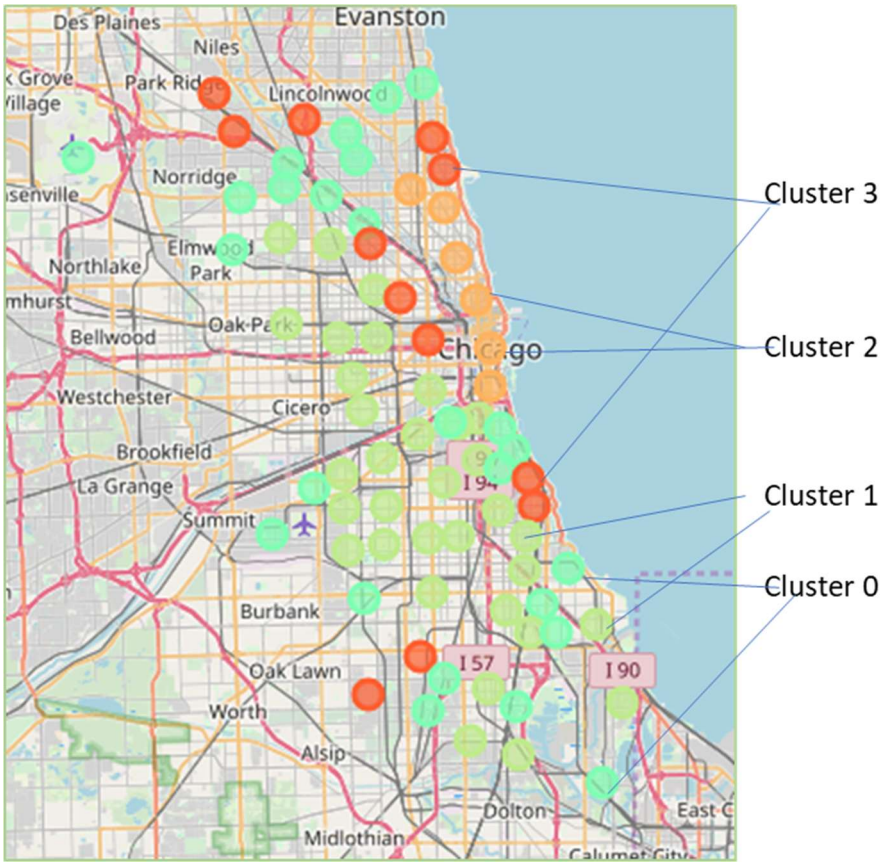




With the above information, we can describe the groups as follows:

- Group 0: **Low income**, high unemployment rate, less educated, moderate hardship index
- Group 1: **Poor** living condition, large number of populations below poverty, every high unemployment rate, low education, very high hardship index
- Group 2: **Highest income**, lowest unemployment rate, highly educated, lowest hardship index
- Group 3: **Moderate income**, low unemployment rate, well educated, low hardship index

The clusters are showing on the map below:



As one can see, the high-income community areas are along the lake side, in downtown and near north area. The green ones for clusters 0 and 1 are more on the south side of Chicago. Cluster 2 community areas are spread more towards north except two community areas on the south. This is consistent with general observations for people living in the Chicago area.

Neighborhood Venue Clustering

The community areas are clustered into three different groups using the neighborhood venue category data.

Out of the 77 community areas,

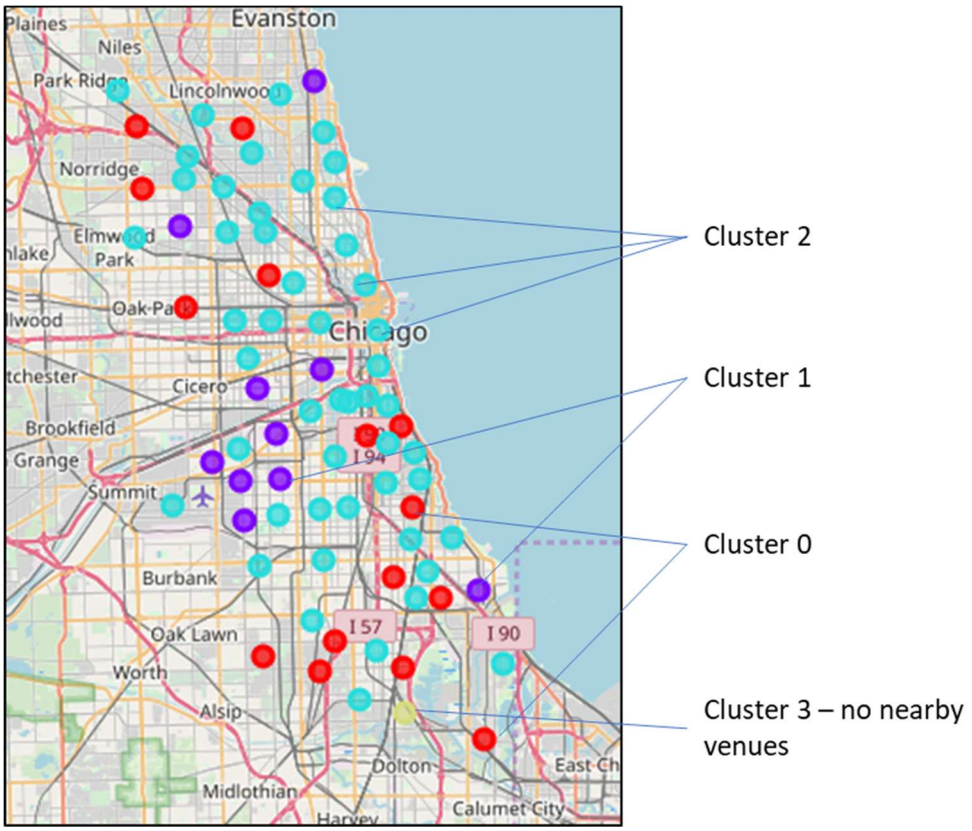
- 15 (~20%) are in Cluster 0,
- 10 (~13%) are in Cluster 1,
- 51(66%) are in cluster 2. And
- 1 community area has no neighborhood venues and it's in its own cluster.

By examine the venues in each cluster, we can label the clusters as follows:

- Cluster 0: (older) residential areas with various ethnic restaurants, possibly with residents who were immigrants from similar regions many years ago

- Cluster 1: quieter residential areas with bakeries, glossary stores, pharmacies, banks, etc.
- Cluster 2: busy city regions with all types of stores, restaurants, bars, gyms, entertainments, etc.

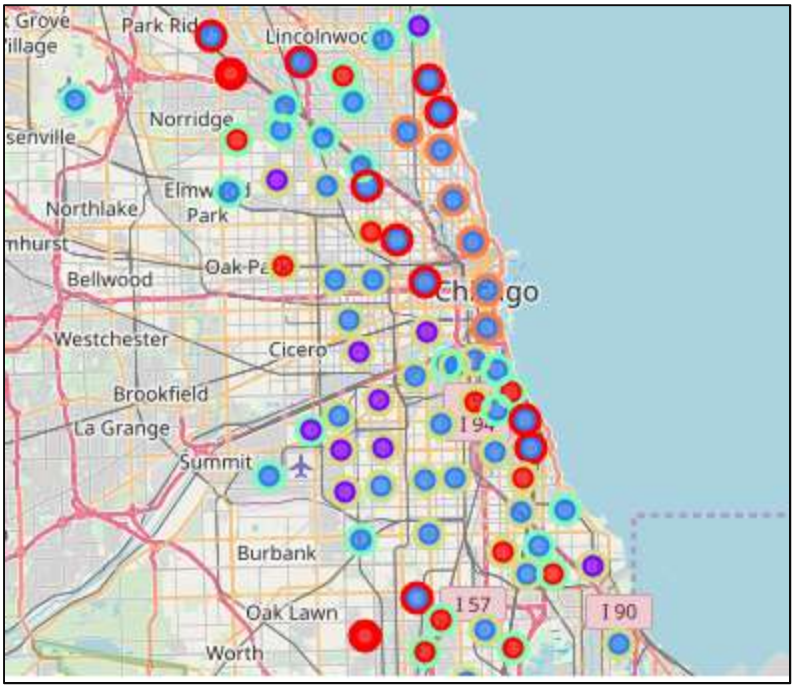
These clusters are shown in the map below:



As one can see, that majority of the community areas are in blue – these are busy city regions. Cluster 1 are mostly on the west side wile Cluster 0 are on either north or south sides of the city.

Correlating the two cluster sets

With the understanding of the above two clusters, one may ask, are there any relationships between the two clusters? Are certain neighborhood venues related to the social economic indices? Below we plot the two clusters on the same map, with the census grouping using larger circles on top of the neighborhood clustering. Unfortunately, it is difficult to see an obvious pattern.

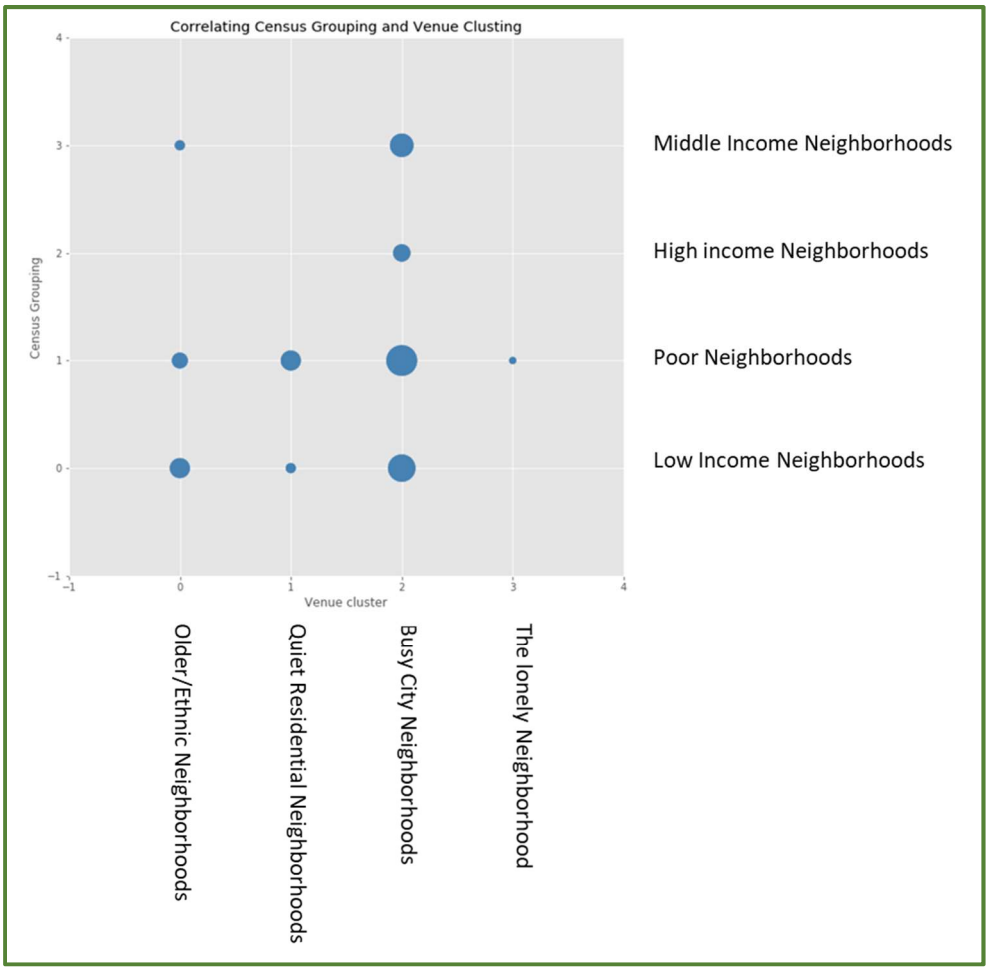


To approach this from a different angle, we merged the two clusters into one data frame that has the name of the community areas along with the two clustering labels. Below shows the head of the data frame:

```
1  combined_data = venue_clust.copy()
2  combined_data = combined_data.join(
3      census_group.set_index('Neighborhood'),
4      on='Neighborhood')
5  combined_data.head(10)
```

	Neighborhood	Cluster_Labels	Census_Group
0	Rogers Park	1	0
1	West Ridge	2	0
2	Uptown	2	3
3	Lincoln Square	2	3
4	North Center	2	2
5	Lake View	2	2
6	Lincoln Park	2	2
7	Near North Side	2	2

This data frame is then used to plot a bubble chart – the axis are the two clusters respectively, while the bubbles reflect the number of items with the same value pairs.



Discussions

From the bubble chart above, one can conclude that cluster 2 per neighborhood type, i.e., the busy city region, exist in all types of community areas; census group 2 has a smaller circle since this particular high-income group has a relatively smaller membership.

Also, the high-income neighborhoods only exist in cluster 2, not in other clusters. These are indeed popular areas that attract businesses, younger generations, and those who love city life.

Middle-income community areas are either in busy city regions or in quiet residential areas, but not in older residential areas.

The one community area that doesn't have any venues (Cluster 3) belongs to the poorest cluster.

From the above analysis, we can see that businesses exist in all types of community areas regardless their social economic indices. However, business owners need to be aware that even for the same business categories, they're dealing with different types of customers.

A set of community areas near downtown Chicago are the wealthiest areas. Older communities have more ethnic foods and are likely to have residents who have been there for long time with older buildings/houses.

Conclusion

In this report, we used two sets of data to cluster the community areas in Chicago into different clusters. Each cluster is examined and labeled with its specific characteristics that will allow users to understand better of the community areas.

Based on the social economic indices, we grouped the community areas into 4 different groups:

- Low income
- Poor
- High income
- Moderate income

Based on the neighborhood venues, we cluster the community areas into also 4 different clusters:

- Quite residential areas
- Older residential areas with specific ethnic groups
- Busy city regions
- Area with no nearby venues (1 community area only).

The relationships between the two clusters are not as obvious but we're able observe a few points using the bubble chart.

It's hoped that the analysis provides useful information to anyone who would like to understand better the Chicago community areas either for setting business strategies or choosing a future neighborhood to stay.