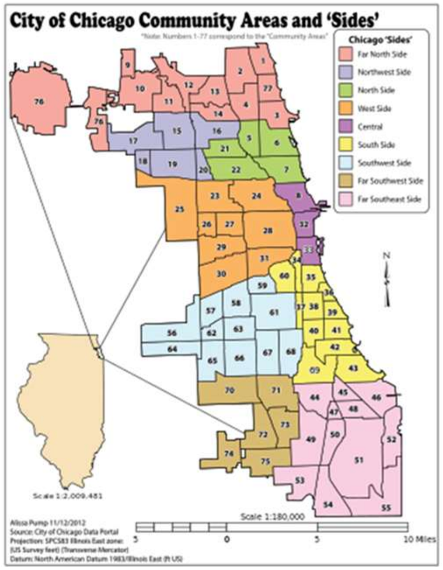
Capstone Project – Battle of Neighborhoods

Understand Chicago Community Areas

Introduction

- This project studies the ***Chicago Community Areas*** from two different perspectives:
  - social economic indices, and
  - neighborhood venues
- About Community Area
  - "Community areas" are loosely the same as "neighborhoods", and in most cases, there is a one-to-one mapping between a neighborhood and a community area.
  - We use "community area" in this study since official census data are collected for community areas but not for neighborhoods
  - These two terms are used interchangeable in the rest of the package.


City of Chicago Community Areas and 'Sides'
*Note: Numbers 1-77 correspond to the "Community Areas"
Chicago 'Sides'
Far North Side
Northwest Side
North Side
West Side
Central
South Side
Southwest Side
Far Southwest Side
Far Southeast Side
N
Scale 1:2,009,481
Scale 1:180,000
5
0
5
10 Miles
Alissa Pump 11/12/2012
Source: City of Chicago Data Portal
Projection: SPCS83 Illinois East zone: (US Survey feet) (Transverse Mercator)
Datum: North American Datum 1983/Illinois East (ft US)

- In this study, the community areas are clustered into different groups using the census data as well as the neighborhood venue data; the results of the two clustering exercises are also combined and correlated.

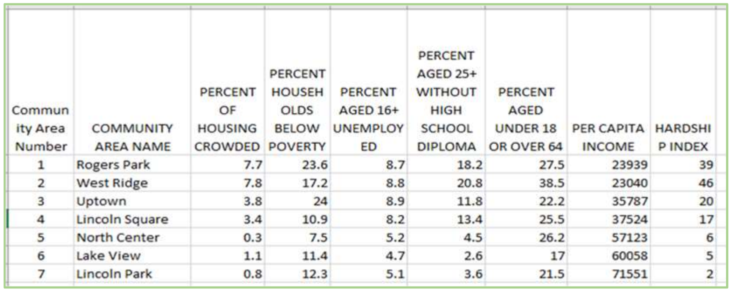
Data Sets

Two sets of data are required for this study:

Dataset	Data Source
Census data with social economic indices for the community areas	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.
Neighborhood venues for the community areas	*This dataset is obtained via the FourSquare API*. We use the ***community area names*** from the previous census dataset 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.

Data Preparation - Social Economic Indices from Census

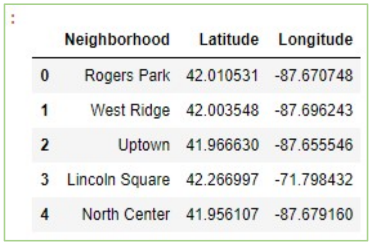
- The census data requires very minimum cleansing as it has been well prepared; it's ready for use. The only updates are to fix typos in two community areas.
- The features in this dataset include 7 different social economic indices
    - *Percent of housing crowded* (i.e. living condition)
    - *Percent 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)



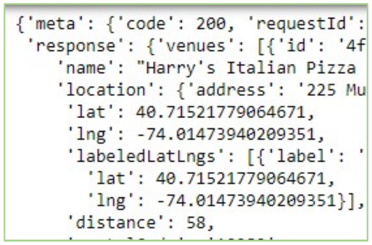
Community Area Number	COMMUNITY AREA NAME	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER CAPITA INCOME	HARDSHIP INDEX
1	Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39
2	West Ridge	7.8	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

Data Preparation – Neighborhood Venue Categories

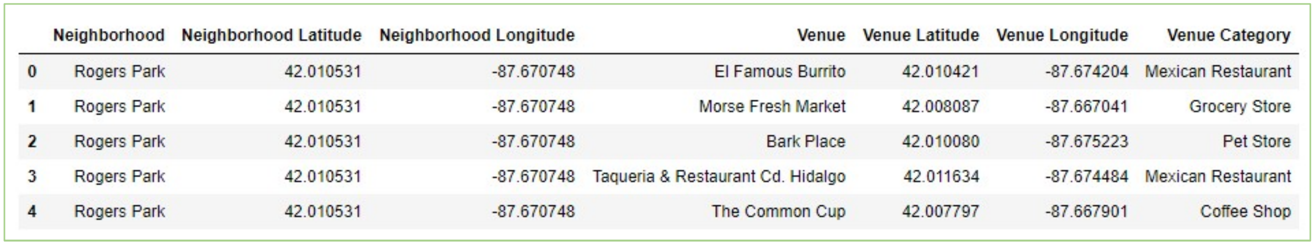
The neighborhood venue data is generated by firstly getting the geo-locations of the community areas using the names in the previous data set, and then calling the FourSquare API to retrieve the data. The resulting data is also transformed to a data frame with all the venue categories as the feature for clustering:



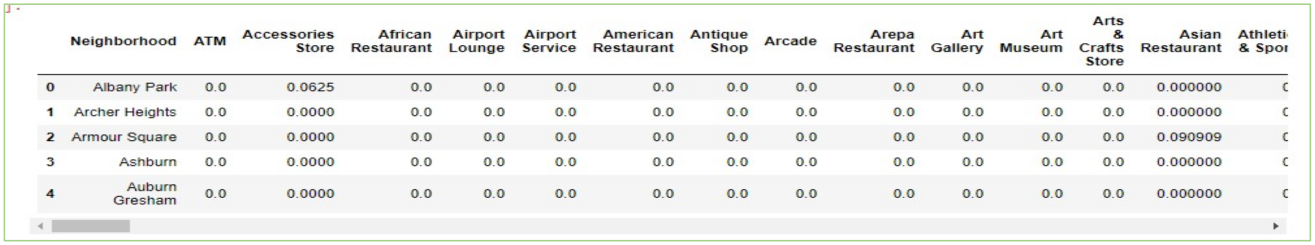
	Neighborhood	Latitude	Longitude
0	Rogers Park	42.010531	-87.670748
1	West Ridge	42.003548	-87.696243
2	Uptown	41.966630	-87.655546
3	Lincoln Square	42.266997	-71.798432
4	North Center	41.956107	-87.679160



```
{'meta': {'code': 200, 'requestId':
 'response': {'venues': [{'id': '4f
    'name': "Harry's Italian Pizza
    'location': {'address': '225 Mu
     'lat': 40.71521779064671,
     'lng': -74.01473940209351,
     'labeledLatLngs': [{'label': '
       'lat': 40.71521779064671,
       'lng': -74.01473940209351}],
     'distance': 58,
```



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park	42.010531	-87.670748	El Famous Burrito	42.010421	-87.674204	Mexican Restaurant
1	Rogers Park	42.010531	-87.670748	Morse Fresh Market	42.008087	-87.667041	Grocery Store
2	Rogers Park	42.010531	-87.670748	Bark Place	42.010080	-87.675223	Pet Store
3	Rogers Park	42.010531	-87.670748	Taqueria & Restaurant Cd. Hidalgo	42.011634	-87.674484	Mexican Restaurant
4	Rogers Park	42.010531	-87.670748	The Common Cup	42.007797	-87.667901	Coffee Shop



	Neighborhood	ATM	Accessories Store	African Restaurant	Airport Lounge	Airport Service	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athleti & Spor
0	Albany Park	0.0	0.0625	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0
1	Archer Heights	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0
2	Armour Square	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.090909	0
3	Ashburn	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0
4	Auburn Gresham	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0

Methodology

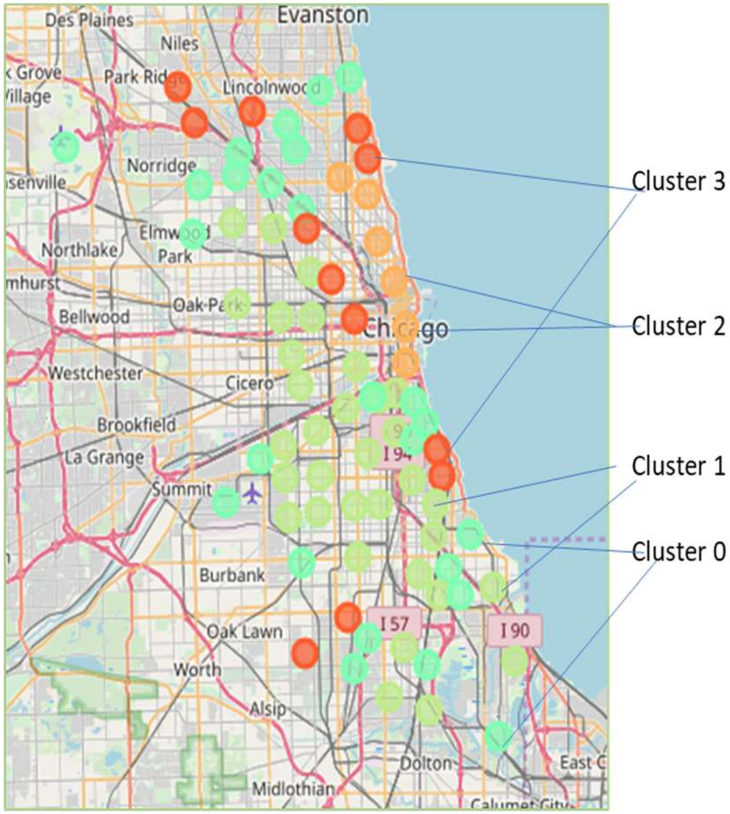
The table below summarizes the methods and consideration factors in this study

Consideration factors	Description
Algorithm/Model	***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	The social economic clustering uses all the features in the census data. The neighborhood venues clustering uses all the venue categories as the features
Selecting the right K's	Prior to the analysis, it is not clear how many clusters make the best sense for the clustering exercise. We iterate the k-means using different k values and select the k that allow the best description of the clusters. The k values are selected independently for the two data sets.
Tools	Jupyter notebook is the ultimate environment for this exercise. We used the multiple libraries for this study including pandas, numpy, json, geocoders, requests, matplotlib, sklearn, and folium

Analysis Results – grouping community areas by census data

We used k=4 to cluster the community areas using the census data resulting with the following groups/clusters:

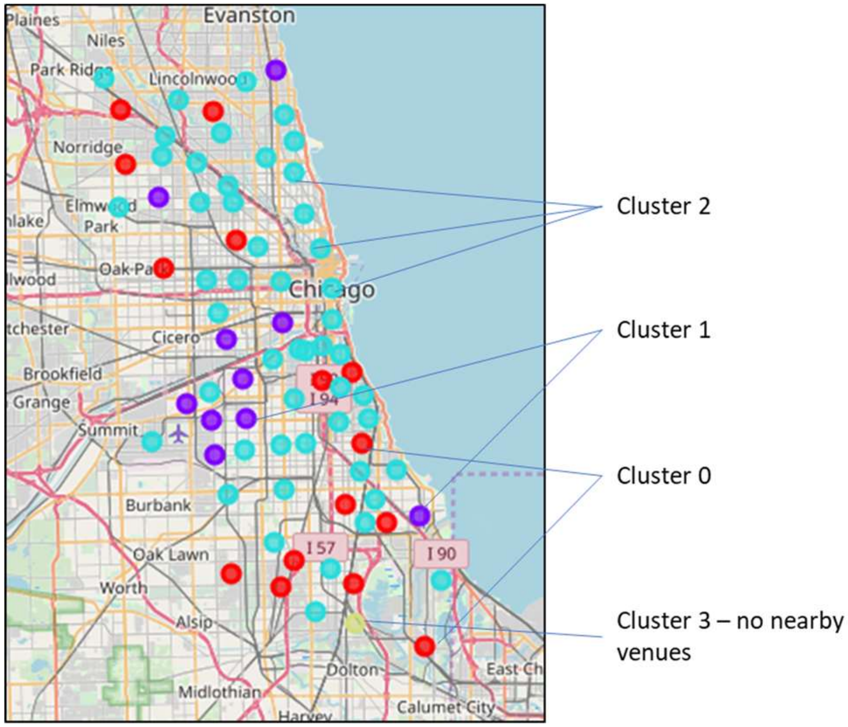
- Group 0: Low income, high unemployment rate, less educated, moderate hardship index
- Group 1: Poor living condition, large number of populations below poverty, every high unemployment rate, low education, very high hardship index
- Group 2: Highest income, lowest unemployment rate, highly educated,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



Analysis Results – grouping community areas by venues

We used k=3 to cluster the community areas using the neighborhood venue data obtained via FourSquare API. The clusters are labeled 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.



Cluster 2

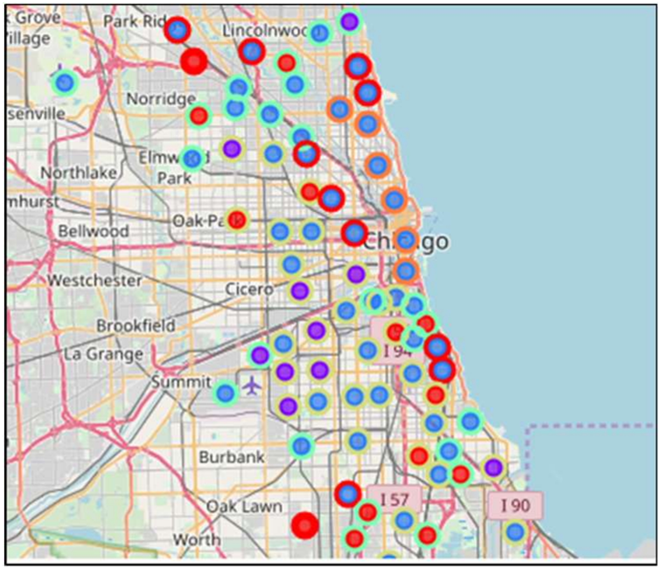
Cluster 1

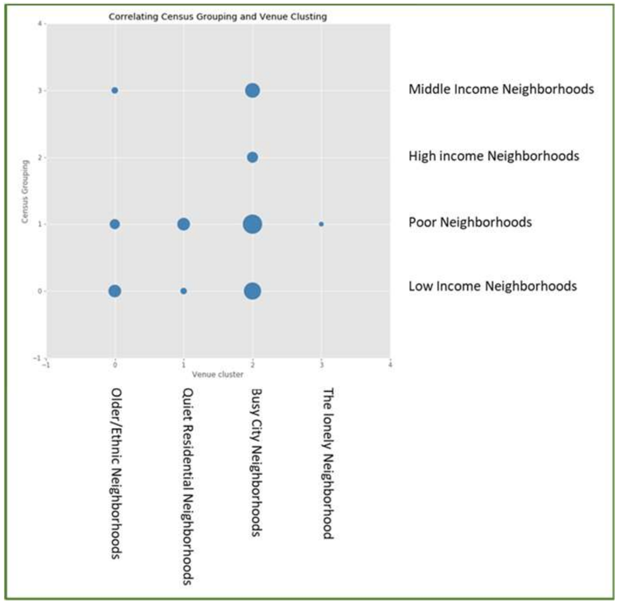
Cluster 0

Cluster 3 – no nearby venues

Analysis Results – Correlating the clusters

To correlate the groupings from the two exercise, we firstly plot the clustering results on the same map; however, it doesn't reveal must insights. We then plot a bubble charts to see the correlation between the groupings:





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 the community areas better.
- Based on the social economic indices, we grouped the community areas into 4 different groups:
  - Low income
  - Poor
  - High income
  - Moderate income

We hope 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.

- 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.