**Income and Venue Data Analysis of Chicago**

Ting Gong

June 13, 2021

**Introduction**

Chicago is the third largest city in the United States with its nearly 9 million population. The entire city includes 77 community areas and are highly demographically diverse.

I like the community that I live in because of the convenience and the similar lifestyle of other community members. For example, there are many restaurants in the neighborhood, and the grocery stores are in five-minute walking distance. Park and recreational areas also are just a block away. Community members are well-educated and have decent jobs and are at a relatively high-income level.

However, I may need to move to other areas due to career development. I am curious about the similarity of different community areas in Chicago and how income levels distribute across Chicagoland. By knowing the similar communities as my current one, I may better decide the possible communities for me to live in the future.

**Data used in the analysis**

First, I used Requests and BeautifulSoup to webscrape the community areas in Chicago from Wikipedia page. It contains the name of total 77 community areas in Chicago.

Second, I used community name to get the geographic coordinates of each community. Next, I used geographic coordinates to get venue information for each community by accessing Foursquare API.

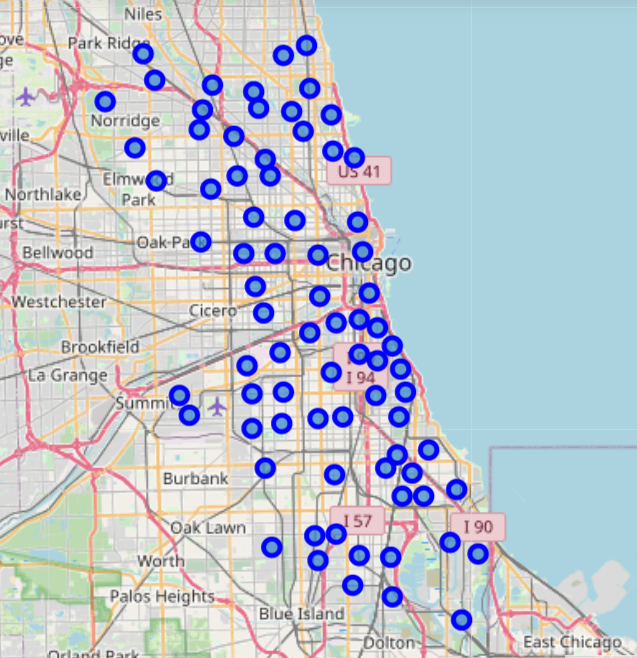
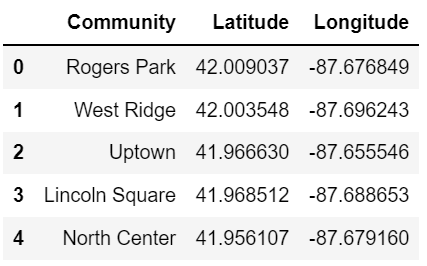
Third, I downloaded “Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012” from City of Chicago data portal. The file includes Per Capita Income from 2008 to 2012 of each community. I used this data to group communities into 4 income levels.

Last, I downloaded "chicago-community-areas.geojson" from kaggle.com. The geojson file contains the geocodes of the boundaries of each community in Chicago. I used the geojson file to create a choropleth map that color-codes the income levels of community.

**Methodology**

**1.Obtain community information and locations**

I created a main data frame that contains the community name and the geographic coordinates. First, webscrape “List of community areas” table from Wikipedia page[1], then convert it to a dataframe. Second, use Geopy liabrary to get the geographic coordinates of each community. Third, clean and process data and concatenate all information into one dataframe, then use Folium to visualize communities on the map.



**2. Explore venues in each community**

I utilize Foursquare API[2] to return venue information with 1500 meters of each geographic coordinates. As we know, the area of Chicago is 234 square mile (606 square km). The city is divided into 77 communities. On average, the area of each community is 3 square mile (7.8 square km). Thus, the radius of each community is approx. 0.98 mile (1.58 km). To maximize the coverage but also reduce overlap, I set the radius to 1500 meters.

Foursquare API returned 5266 venues, of which 987 are duplicated. I removed duplicates and only use 4279 unique venues in the analysis. Next, I group venues by community so I can see the number of venues in each community. Considered the default limit of Foursquare API is 100, I also checked the communities that are caped by the limit, which turned out there are seven communities with the maximum 100 venues. The limit may skew the further analysis.

To normalize the data for clustering, I examined the unique venue categories of all returned venues before assigning dummies to each venue category. The API returned 4279 unique venues in 338 unique categories.

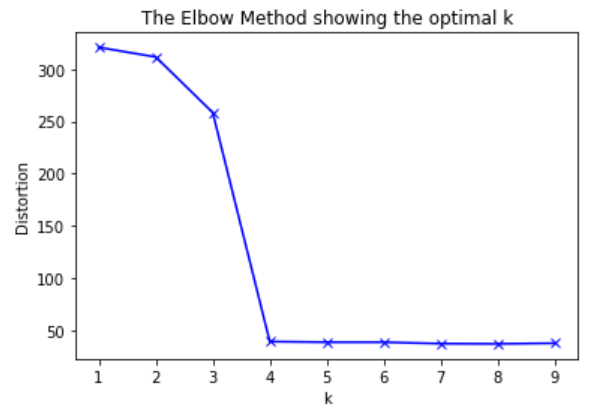
**3. Analyze Each Community**

I assigned dummies to each category, discovered the top 10 venues in each community, and then sort them in descending order.



**4. Cluster communities using K-Means**

I used correlation of Canberra function to see the elbow point to determine the optimal K. The below chart show the optimal K is 4.

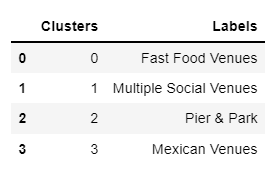


Once the algorithm clustered all communities into 4 clusters, I joined new clustered data frame to the top 10 venues data frame. The new data frame is show as below.



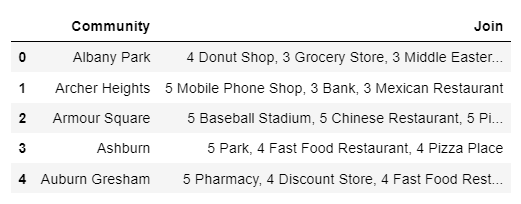
Then, I discovered the top venues in each cluster so I can add cluster names to each cluster.



After observing the results above, I added cluster names as below.

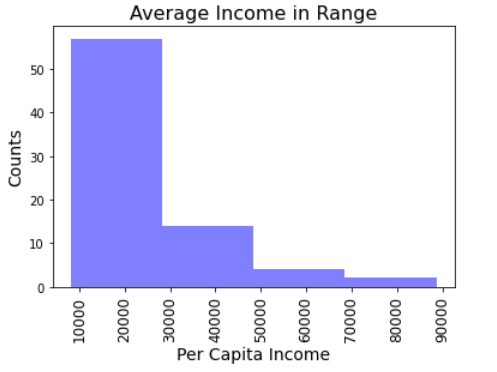
* + Cluster 0 : "Fast Food Venues"
  + Cluster 1 : "Multiple Social Venues"
  + Cluster 2 : "Pier & Park"
  + Cluster 3 : "Mexican Venues“

I also created a label that will show top 3 venue categories in each community for the use on map later.



**5. Obtain Per Capita Income of communities in Chicago**

I used Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012”[3] from City of Chicago data portal, and then read it into pandas dataframe. Then I used histogram to visualize the distribution of income. The histogram clearly shows four groups of income.



* + Below $30k, Low Income
  + $30k-$50k, Medium-Low Income
  + $50k-$70k, Medium Income
  + Above $70k, High Income

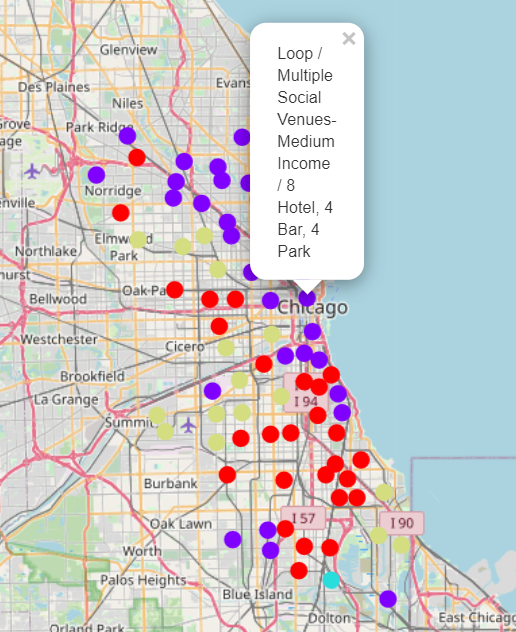
After creating new labels and obtaining the income data, I merged all new variables with the top 10 data frame. The merged data frame is shown as below.



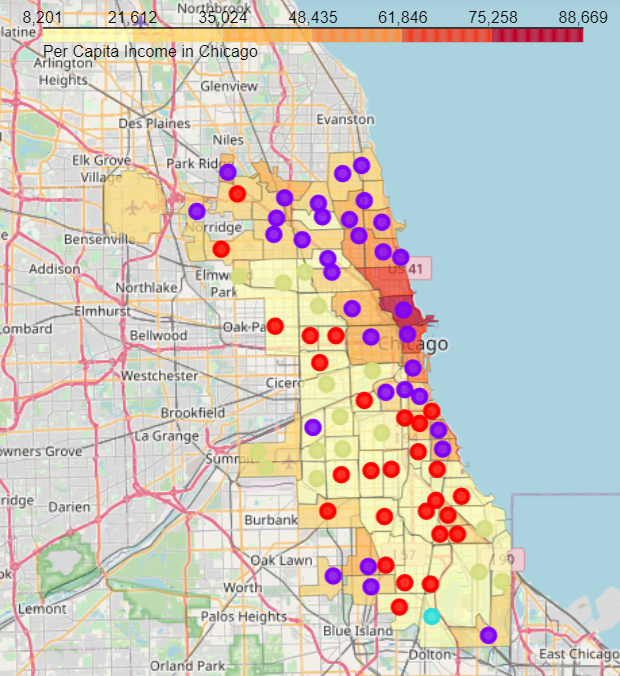
Now, we can visualize all clusters with income level on the map.

The label of each community contains:

* + Community name
  + Cluster name
  + Income level
  + Top 3 venue category in the community



To better present the correlation between income and community clusters, I create a choropleth map indicating income levels in different colors. “chicago-community-areas.geojson”[4] downloaded from Kaggle.com is used as the map layer.



**Discussion**

Chicago is a large city with residents with diverse background. The analysis only used venue information to cluster communities regardless the demographic background. On the other hand, Foursquare API, as default, only returned 100 venues within the radius of each coordinates, which excluded many venues in seven communities including Loop, Near North Side, Near South Side, Near West Side, Rogers Park, Uptown, and West Town. Some of these communities are highly influenced by their demographic and economical factors. The limit of data may skew the accuracy of clustering.

Per Capita Income from City of Chicago represents the income levels from 2008 – 2012. Income has increased nationwide in the past years. Per Capita Income for Chicago in 2019 was $40,144, which has increased $6,664 (19.5%) from $33,590 in 2012[5]. The data used in the analysis may not accurately present the current income level in Chicago. However, the distribution of income in Chicago remains reliable.

To get more accurate results, further analysis maybe needed to take demographics, house value, and crime into consideration.

**Conclusion**

Communities with multiple social venues are mainly located on the northside of Chicago.

Fast food venues are mainly located on the southside of Chicago.

Mexican venues are on the west of Chicago.

Communities with medium to high income are on the northeast of Chicago along the lake. The income decreases from east to west.

Fast food venues and Mexican venues are in the low-income communities.

To conclude, I will remain in the communities with multiple social venues. In addition, the house price and crime data will be further considered.

**Reference**

[1] Wikipedia: List of community areas

<https://en.wikipedia.org/wiki/Community_areas_in_Chicago>

[2]Foursquare API

https://developer.foursquare.com

[3] “Census Data - Selected socioeconomic indicators in Chicago, 2008 – 2012” from City of Chicago data portal.

<https://data.cityofchicago.org/Health-Human-Services/Per-Capita-Income/r6ad-wvtk>

[4] "chicago-community-areas.geojson" from kaggle.com.

<https://www.kaggle.com/doyouevendata/chicago-community-areas-geojson>

[5] Historical Real Per Capita Income for Chicago

https://www.deptofnumbers.com/income/illinois/chicago