**Determining the Ideal Community Area for a Korean Restaurant in Chicago, Illinois**

Sunjae Lim

Applied Data Science Capstone on Coursera by IBM

July 24, 2020

**INTRODUCTION**

Restaurants are popular establishments which serve multiple functions. They act as meeting spots, landmarks, introductions to new cultures and cuisines, and provide nourishment. In addition, the industry itself has an upward trend in growth and continues to flourish.

The National Restaurant Association indicates a projected restaurant industry revenue of nearly $899 billion for 2020, which is a $36 billion increase compared to the industry revenue from 2019 (National Restaurant Association). With growth projections seemingly positive for the future, the health of the industry seems robust. Two of the most important factors for establishing a restaurant are location and population.

Chicago is the third largest city in the United States with over 7,000 restaurants as of 2010 (City of Chicago, 2010 - 2020). The city presents a ripe opportunity for new restauranteurs or franchise owners to leave their mark and expand their business into a densely populated area.

This project will examine neighborhood populations and nearby venues for each neighborhood and determine which spot would be ideal for setting up a Korean restaurant in Chicago, Illinois. The findings may benefit Chicago city planners, Korean restauranteurs aiming to expand into Chicago, or food enthusiasts who are seeking a big picture of which community areas have popular Asian venues.

**DATA**

**Community Areas and Demographics**

A list of neighborhoods and associated community areas were scraped from the Wikipedia page titled ‘List of neighborhoods in Chicago’. The website indicates Chicago neighborhoods have historically changed over time, so the neighborhoods will be grouped into their respective community areas. Defined community areas are not used by residents and may contain several neighborhoods, but their boundaries do not change which allows for comparisons over time (Community areas in Chicago, 2020). Table 1 depicts a portion of the dataframe after the neighborhoods have been grouped.

**A screenshot of a cell phone

Description automatically generated**

***Table 1****. Dataframe of community areas and grouped neighborhoods*

For community area population data, a CSV table from July 2020 posted on the CMAP Data Hub site maintained by the Illinois government will be imported into our notebook. The data contains various fields, but we will be focusing on fields pertaining to community area age groups and total population from the 2014 to 2018 community survey. All other fields will be dropped, as shown in Table 2.

A screenshot of a cell phone

Description automatically generated

***Table 2****. Modified dataframe depicting demographic data for community areas*

**Location Data**

Coordinate data was obtained by running each community area through Geopy. A sample of the code used to obtain latitude and longitude coordinates via Geopy is below:

**for** z **in** locations:

**try**:

query\_ll = "**{}**, Chicago, Illinois"

location = geolocator.geocode(query\_ll.format(z))

latitude\_ca.append(location.latitude)

longitude\_ca.append(location.longitude)

query\_result = "**{}**, latitude **{}**, longitude **{}**"

print(query\_result.format(z, location.latitude, location.longitude))

Foursquare API will be utilized to obtain venue information for locations near the community areas. API credentials will be obtained from my personal Foursquare developer account. Credentials passed into the query are the developer client ID and client secret code. The user can define the radius of search, which is in meters, and limit, which limits the max amount of results returned A snippet of code used to obtain venue data from Foursquare API is below:

url = 'https://api.foursquare.com/v2/venues/explore?&client\_id=**{}**&client\_secret=**{}**&v=**{}**&ll=**{}**,**{}**&radius=**{}**&limit=**{}**'.format(CLIENT\_ID, CLIENT\_SECRET, VERSION, lat, lng, radius, LIMIT)

**METHODOLOGY**

Different community areas were first examined according to their populations and age groups. Given that each community area contained several neighborhoods and held varied population compositions, an analysis of the top community areas in terms of total population vs non-senior population was also conducted. The main visualization techniques used involved bar graphs, both horizontal and stacked variants. This allowed for a quick visual examination and recognition of basic features from the dataset.

The second comparison was performed by using one-hot encoding to preprocess Foursquare API data. By using one-hot encoding, we can find averages for each of the venue locations frequented by the population and order them by the most popular to least popular. By doing so, we can then accurately use K-means clustering to group similar areas together according to popular venues.

After K-means clustering, we can investigate individual clusters to spot any patterns or similarities. Adjusting the k value to obtain a stronger separation will lead to more accurate clustering. Because the values are more categorical, the 'elbow' method was not used to find the optimal K value.

**Comparing by Population**

A larger population may indicate a larger potential customer base. Population data was arranged into descending order based on total population. From the community areas, the top ten most populated areas were visually compared with a stacked bar graph, shown in Figure 1. Older age groups were stacked towards the top and the youngest group, aged 19 and under, were placed at the bottom.

A picture containing drawing

Description automatically generated

***Figure 1.*** *Comparing the top ten most populated community areas*

To obtain another perspective on the top ten community areas, age groups for individuals considered senior or older were excluded and new top areas were obtained. A horizontal bar graph, Figure 2, was used to visually compare which neighborhoods held the youngest population.

A screenshot of a cell phone

Description automatically generated

***Figure 2.*** *Comparison of the top ten youngest community areas*

Given that the top ten community areas appear to have changed after removing the older population from the comparison, a new graph was created to visually distinguish younger vs older populations in the top fifteen neighborhoods. The top fifteen areas were viewed to include areas excluded previously due to their older populace. Community areas were sorted in descending order to produce a horizontal bar graph titled Figure 3, below.

A screenshot of a cell phone

Description automatically generated

***Figure 3.*** *Top fifteen community areas with the youngest overall populations by count*

**Obtaining and Examining Venues**

Foursquare API was used to obtain venue data for locations near the 77 community areas. The initial radius limit of 500 meters did not return any venue results for Riverdale and South Deering. The radius was increased to 1000 meters and results were obtained for all 77 area. Given that the lowest venue result was 3 and the second lowest was 4, categorization for the most popular venues was limited to the top 4.

Due to venues being categorical entries, one hot encoding was used to convert the data, which was then averaged per community area to find the most popular venues.

A screenshot of a cell phone

Description automatically generated

***Table 3.*** *Dataframe depicting community areas with their top four most popular venues*

Following our venue discovery, each community area was sorted into a cluster using the scikitlearn library. A snippet of code is shown below:

kclusters = 5

chicago\_clusters = chicago\_onehot\_grouped.drop(['CommunityArea'], 1)

kmeans = KMeans(n\_clusters = kclusters, init='k-means++',random\_state=0).fit(chicago\_clusters)

Initially a cluster value of 5 was used; however, many of the community areas remained in large groups with mixed features. The cluster value was increased to 10 and reevaluated to obtain a better yield. Each cluster was examined for consistency or patterns before moving onto mapping the data using folium and matplotlib. Below is an image of cluster 3, which held areas with Mexican restaurants as their top venues.

A screenshot of a cell phone

Description automatically generated

***Table 4.*** *Cluster 3 community areas with their top four venues*

A side examination of clusters containing specific venues such as Korean, Chinese, Japanese, Vietnamese, Asian, Indian or Pakistani food was performed. Venues with train stations, bus stops and taxis were also queried. A potential list of candidates for a Korean restaurant location was formed from this list.

**Mapping Clusters**

Folium and matplotlib libraries were used to visualize the result of our clusters. Different colors ranging from yellow to purple were used to plot different clusters. Please view Figure 4 below.

A close up of a map

Description automatically generated

**Figure 4**. Plotted clusters for the 77 community areas

The candidates derived from the queries earlier were also plotted (Figure 5). The candidates were ordered according to total population count and compared via a horizontal bar graph (Figure 6).

A close up of a map

Description automatically generated

**Figure 5**. Mapped clusters for possible candidates which have Asian restaurants or transportation.

A screenshot of a cell phone

Description automatically generated

**Figure 6**. Selected candidates graphed according to total population count

**RESULTS**

**Population and Venue Results**

Starting with the largest population count, the community area of Lake View has the largest population in general with 100547, as well as the largest non-senior population. Within the top ten community areas with the highest populations, the populations with the largest senior population included Austin, Near North Side, West Ridge and Portage Park. The community area in tenth place was home to over 60,000 individuals.

From our venues query, The Loop presented the largest number of venues within a 500 meters. Riverdale and South Deering had no venues within a 500 meter radius. Upon expanding the search radius to 1000 meters, Riverdale and South Deering produced venues, but only a small amount.

**Clustering Results**

The clusters from the k-means produced interesting results. Cluster 2 contained many areas which had popular venues involving transportation, such as train stations, intersections, or bus stops. Cluster 2 also contained the only community area with a Korean restaurant, North Park. Cluster 3 held mainly Mexican Restaurants as the 1st venue picks. Cluster 4 contained areas with popular fast food restaurants. This cluster could have been more accurate if it contained several areas in Cluster 2, such as Calumet Heights, Douglas and North Lawndale. Clusters with single areas included 0, 5, 6, 7, and 8. Clusters 6 and 7 are both well known for their BBQ joints. Cluster 9 presented a variety of locations which usually had a Mexican or fast food restaurant in either the 1st or 2nd venue spot.

Selecting areas which contained Korean, Chinese, Japanese, Vietnamese, Asian, Indian, or Pakistani restaurants yielded the following areas: North Park, Armour Square, Bridgeport, Chatham, Near South Side, Hegewisch, Uptown, West Lawn, Edgewater, West Ridge. Searching for community areas with train, bus or taxi stops yielded the following: East Garfield Park, Fuller Park, Pullman, West Garfield Park, Riverdale, Grand Boulevard, North Park, & Morgan Park.

**Map Results**

From the folium maps, distinct divisions can be seen among some areas. Cluster 1, represented in purple ran along the Kennedy Expressway, but was localized to the north side of the city. Cluster 2, represented by dark blue, ran along the Kennedy Expressway, but was mainly on the south side of the city. Cluster 9, represented by red-orange, was mainly towards the inner west side of Chicago. Clusters 3 and 4 represented by tints of blue were also located mainly towards the left side of the expressway. Clusters 5, 6 and 8 were all located near the southernmost side of the city.

**Candidate Comparisons**

Assuming 60,000 population is the cut off for community area population to be within the top ten, only West Ridge qualifies. Edgewater and Uptown are just below with 57022 and 58424 population, respectively. Upon further examination, community areas with train stations or bus stops were not within the top 10 most populated areas. In fact, community areas with train stations or intersections were the least populated areas.

**CONCLUSION**

**Recommendations**

Overall, my recommendations for setting up a Korean restaurant would be in the community areas of West Ridge, Uptown or Edgewater. My top choice would be West Ridge. Despite the most popular venues being Indian and Pakistani restaurants, it holds the sixth largest population and is nearby North Park, which has a Korean restaurant as one of its top venues. Being near established Asian restaurants will draw in customers who are familiar with or seeking Asian cuisine. Neighborhoods in West Ridge include Nortown, Peterson Park, Rosehill, West Ridge, & West Rogers Park.

Uptown is my second choice due to being further away from a community area with a well-known Korean restaurant but having popular Vietnamese and Chinese restaurants. Edgewater has a popular Asian restaurant as a venue, but slightly less total population compared to Uptown, so it is my third choice.

**Observations**

The Loop held one of the highest venue counts. It easily had 100 venues nearby within 500 meters. The Loop is a bustling business district centrally located in Chicago, so the results seem accurate (Chicago Loop, 2020). South Deering and Riverdale both produced no venues when searching within 500 meters, but they produced results once the radius was expanded over that range. Even at 1000 meters, only 2 to 4 venues were nearby. It is safe to assume these areas are located towards the edges of the city.

Cluster 4 contained areas with popular fast food restaurants. This cluster could have been more accurate if it contained several areas in Cluster 2, such as Calumet Heights, Douglas and North Lawndale. Clusters 6 and 7 both held BBQ joints as one of their top venues; however Cluster 6 was located to the outer edges of Chicago, near Beverly and Cluster 7 was located near the center, east of Chicago. Objective popularity may be debatable due to the neighborhood disparity and population difference of nearly 15,000.

**Limitations**

The first limitation is the use of community areas, which incorporate multiple neighborhoods. Due to the malleable boundaries of individual neighborhoods, community areas were used. Further investigation into specific neighborhoods should be performed for accuracy. Second, the population data used is possibly outdated by a few years. The latest census data for 2020 is not available; however, once it is, it would provide a more accurate view of demographics. Third, the venues yield from the Foursquare API may change over time. This would most likely yield different results from the ones shown here.

**References**

Chicago Metropolitan Agency for Planning. (2020). *Chicago Community Area (CCA) CDS data*. [Data set]. CMAP Data Hub. https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data/resource/8c4e096e-c90c-4bef-9cf1-9028d094296e

Chicago Loop. (2020, July 20). In *Wikipedia*. https://en.wikipedia.org/wiki/Chicago\_Loop

City of Chicago (2010 – 2020). *Facts & Statistics*. Chicago.gov. Retrieved July 18, 2020, from https://www.chicago.gov/city/en/about/facts.html

Community areas in Chicago. (2020, May 30). In *Wikipedia*. https://en.wikipedia.org/wiki/Community\_areas\_in\_Chicago

List of neighborhoods in Chicago. (2020, June 4). In *Wikipedia.* https://en.wikipedia.org/wiki/List\_of\_neighborhoods\_in\_Chicago

National Restaurant Association. *2019 Restaurant Industry Factbook*. Restaurant.org. https://www.restaurant.org/downloads/pdfs/research/soi/restaurant\_industry\_fact\_sheet\_2019.pdf

National Restaurant Association. *National Statistics*. Restaurant.org. Retrieved July 18, 2020 from https://restaurant.org/research/restaurant-statistics/restaurant-industry-facts-at-a-glance