CHICAGO NEIGHBORHOOD BUSINESS ANALYSIS

Prasanna Vengatesh Venkataraman

July 6, 2020

TABLE OF CONTENTS

1. INTRODUCTION:

1.1 BACKGROUND:

City of Chicago (aka Windy City) is the most populous city in the state of Illinois and third most populous city in The United States of America. Spanning an area of 227.63 square miles Chicago is the home to 247 neighborhoods in 77 community areas. Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. It has one of the highest GDP in the world and is the home to various Fortune 500 companies such as Allstate, Boeing, McDonald's, United Airlines Holdings etc.

1.2 PROBLEM:

Details of different neighborhoods in Chicago can be obtained from Wikipedia. The coordinates of each of these neighborhoods can be obtained using GoogleMaps API. Various types of business in a neighborhood can be obtained from the sites such as Foursquare City Guide site which helps provide details about different businesses within a certain radius of a particular neighborhood location. This project aims to classify neighborhoods based on the similarity of venues in each neighborhood and analyze each classification to determine the most and least common types of venues.

1.3 INTEREST:

People who might be interested in this project are entrepreneurs whom are planning to start a new business in the city of Chicago. Others who might be interested are people whom are re-locating to Chicago can determine which kind of neighborhood they would like to move to. People re-locating within Chicago can check out other neighborhoods which are similar to their current neighborhood for re-location.

2. DATA:

2.1 DATA SOURCES:

Data regarding different neighborhoods in Chicago and the community areas they belong to can be obtained from Wikipedia by web scraping. Once obtained using the combination of "*neighborhood, community area, Chicago, Illinois*" the coordinates of each neighborhood is obtained with GoogleMaps API from a Google Developer account.

Using the coordinates the different venues in each neighborhood along with their venue types are leveraged using Foursquare API from a Foursquare Developer account. This data provides the venue name, business type and the venue's coordinates.

2.2 DATA CLEANING & FEATURE SELECTION:

The Wikipedia page is web scraped using urllib.request library and the desired table is selected using the BeautifulSoup library. The data obtained now comprises of the HTML code for the table. The data in each row of the table is extracted using a loop and all the data from each column is stored in a list. It could be seen that some of the Neighborhoods and Community Areas contain the next line code as text "\n". Hence each element in the list is checked and the next line code is removed using a for loop if present. Data from both lists are combined into a dataframe.

Using the neighborhood and community area details from the dataframe the coordinates are obtained by geocoding using GoogleMaps API. The result is obtained in a JSON format from which only "*Latitude*" and "*Longitude*" are extracted and stored in separate lists. Both these lists are added as a separate columns in the dataframe.

The details regarding different businesses in a neighborhood is extracted in a JSON format using Foursquare API. The resulting JSON file contains details for all businesses in a neighborhood. Using a for loop only the *"Name", "Category",* "*Latitude*" and "*Longitude*" for each individual business is selected and the details are added to the dataframe.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Albany Park	41.968327	-87.728028	Lawrence Fish Market	41.968280	-87.726250	Seafood Restaurant
1	Albany Park	41.968327	-87.728028	Starbucks	41.968911	-87.728817	Coffee Shop
2	Albany Park	41.968327	-87.728028	Marie's Pizza & Liquors	41.968132	-87.731533	Pizza Place
3	Albany Park	41.968327	-87.728028	Ssyal Korean Restaurant and Ginseng House	41.968172	-87.733207	Korean Restaurant

Figure 1 – Dataframe after extracting data and cleaning it

2.3 DATA MANIPULATION:

The data regarding venues obtained and stored in a dataframe is now used to create a new dataframe with the 10 most common venues in each neighborhood to which will be used for the clustering algorithm.

A total of 100 businesses at most is fetched for each neighborhood using the Foursquare API. The resulting data consists of a total of 353 different types of businesses. The frequency of each business type in a particular neighborhood is obtained by taking the mean count for a particular business type in a neighborhood.

	Neighborhood	ATM	Accessories Store	African Restaurant	Airport	Airport Service	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	...	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint
0	Albany Park	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0
1	Altgeld Gardens	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0
2	Andersonville	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.022222	0.0	0.0	0.0	0.0
3	Archer Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.0

Figure 2 – Frequency of business type in each neighborhood

From the above dataframe based on the frequency of a business type the top ten most frequent business in a particular neighborhood are extracted and stored in a separate dataframe.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park	Korean Restaurant	Pizza Place	Hookah Bar	Mexican Restaurant	Discount Store	Mobile Phone Shop	Fast Food Restaurant	Taco Place	Fried Chicken Joint	Seafood Restaurant
1	Altgeld Gardens	Park	Zoo Exhibit	Electronics Store	English Restaurant	Entertainment Service	Ethiopian Restaurant	Event Service	Event Space	Eye Doctor	Falafel Restaurant
2	Andersonville	Bakery	Coffee Shop	Café	Shoe Store	Mexican Restaurant	Breakfast Spot	Lounge	New American Restaurant	Jewelry Store	Frozen Yogurt Shop
3	Archer Heights	Mexican Restaurant	Bakery	Video Store	Bar	Gas Station	Nightclub	Seafood Restaurant	Restaurant	Grocery Store	Rental Car Location

Figure 3 – Top ten most common venue in each neighborhood

This data can now be used to perform clustering.

3. CLUSTERING:

With the required data extracted, cleaned and manipulated it is now used as input to perform clustering using K-Means algorithm. The algorithm is run and neighborhood is segregated into 4 clusters based on the top ten common venue types in each neighborhood.

The cluster labels are now added as a column to the dataframe along with each neighborhoods coordinates. Using the coordinates the neighborhoods are plotted on a map using the Folium library and each marker is color coded based on the cluster they belong to.

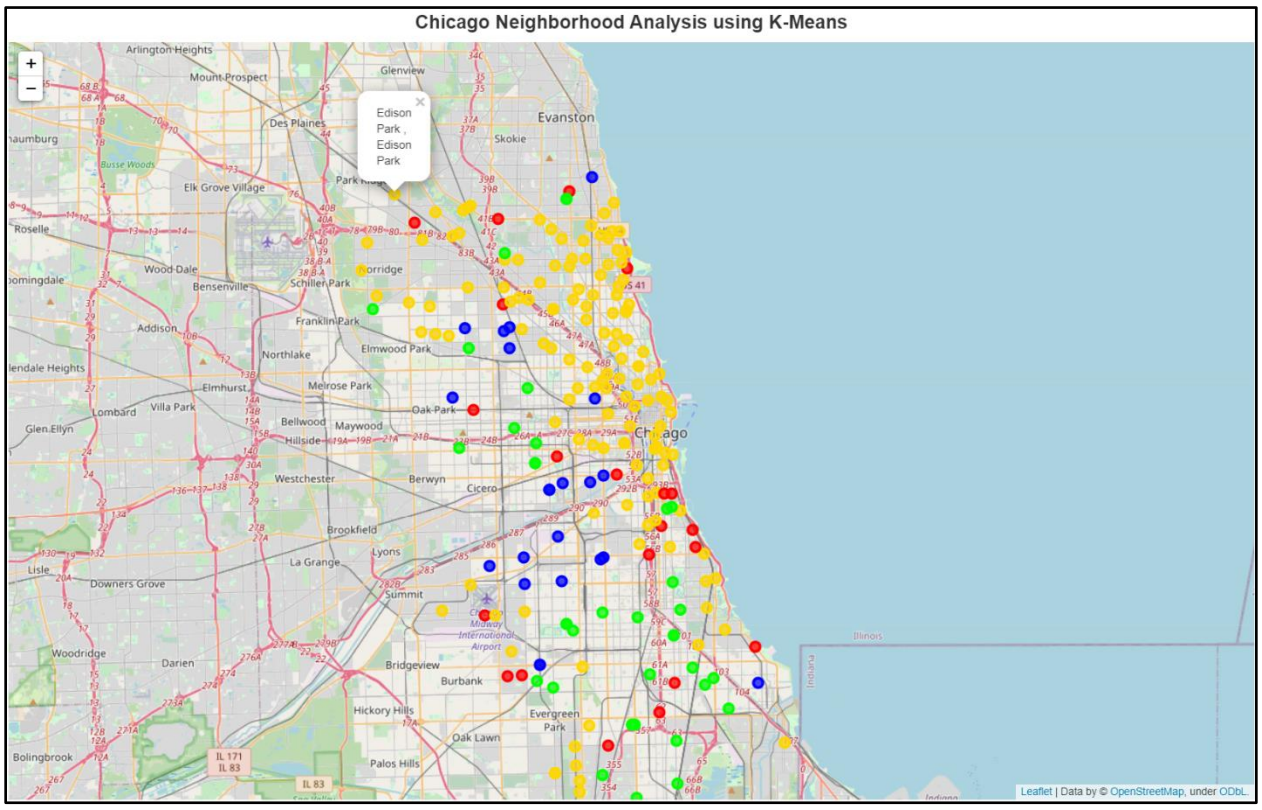

Figure 4 – Chicago Neighborhood using K-Means

Using the results of K-means clustering and the clustering labels, the dataframe can be split into four different DataFrames, one for each cluster and analyzed separately to determine the best businesses to start in each cluster of neighborhoods.

4. RESULTS & DISCUSSION:

After segregating into individual dataframes each cluster can be analyzed separately to determine the most common venues (business categories) in each cluster. A frequency of one indicates that at least one shop of corresponding business type is available in each neighborhood of the cluster while a value of 0.1 indicates that on average only one store of corresponding business type is available in every ten neighborhoods of the cluster.

The most frequent venues in each cluster when analyzed provides the below result:

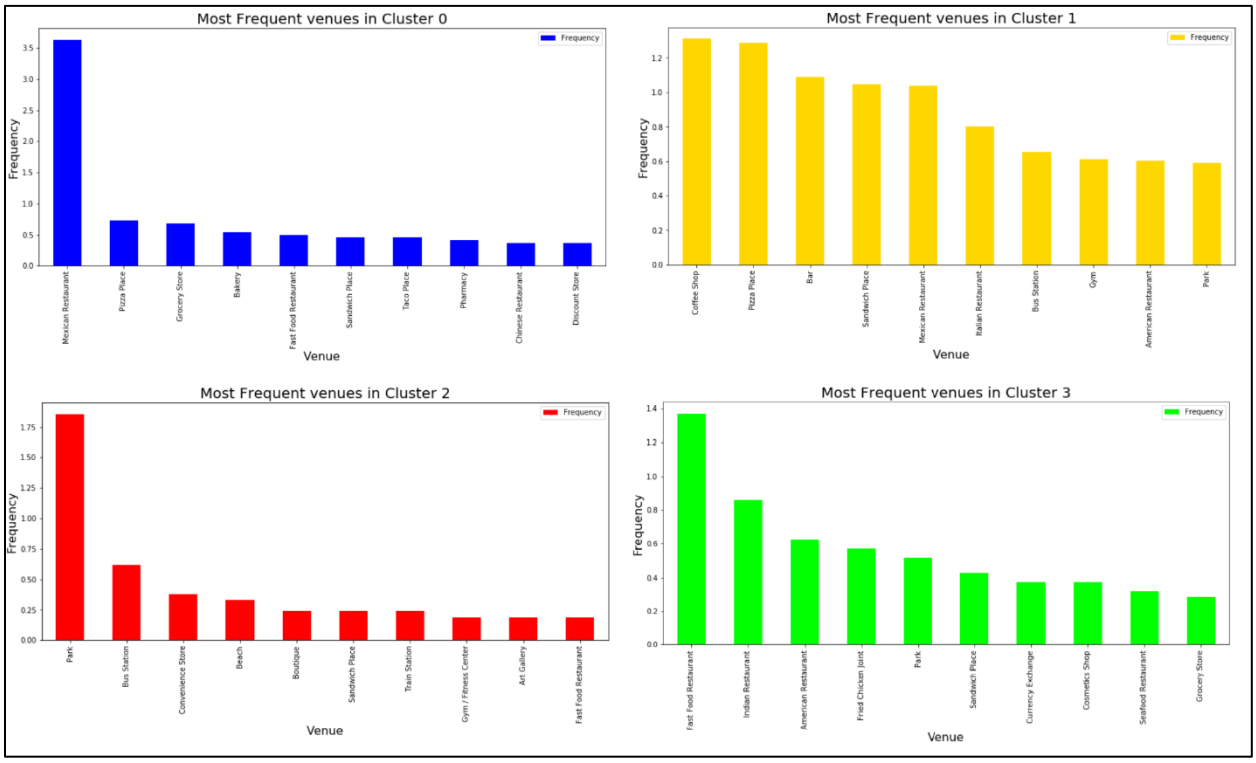

Figure 5 – Most frequent venues

Places such as Mexican restaurants & Pizza Place are quite frequent in clusters 0 and 1. Parks are quite frequent in all clusters except for cluster 0. All clusters have at least one food place or a restaurant among the top ten frequent places.

The least frequent venues in each cluster when analyzed provides the below result:

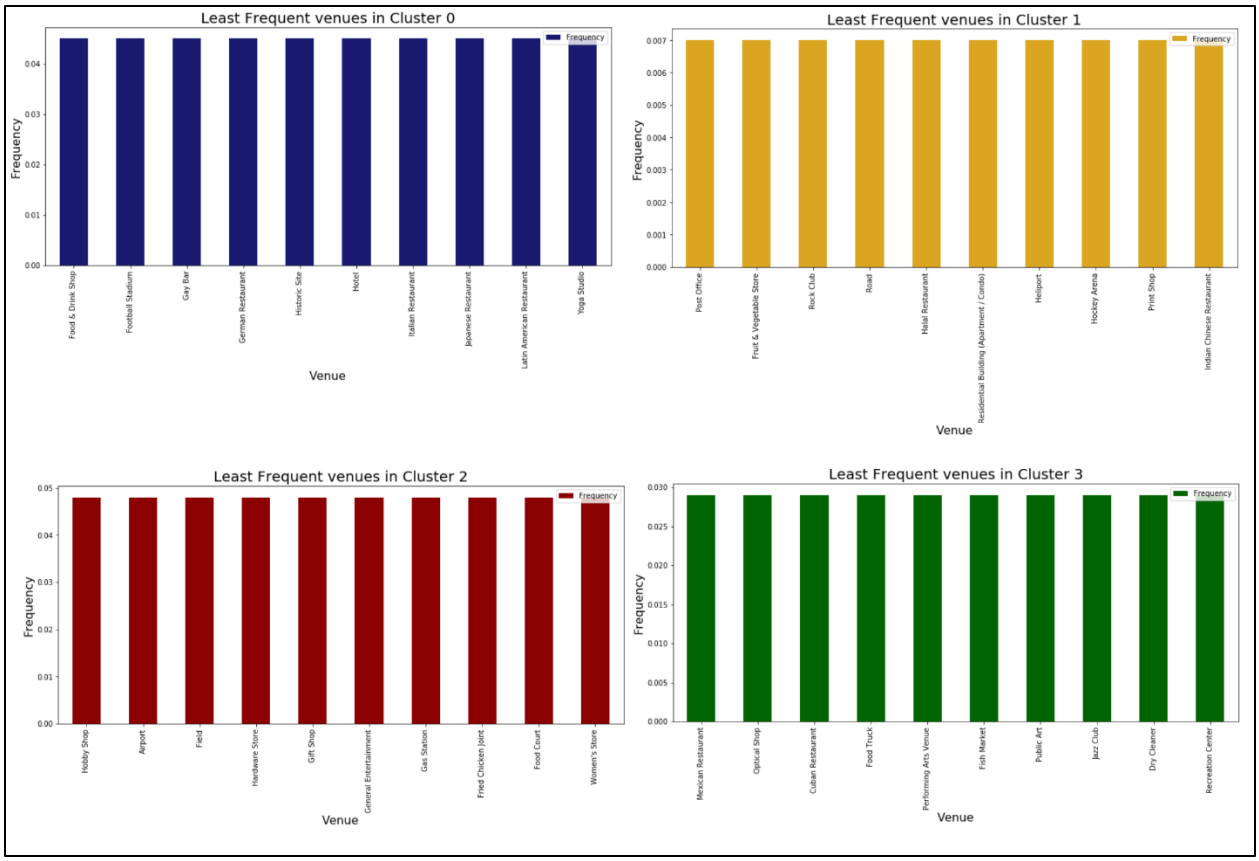

Figure 6 – Least frequent venues

Venues such as Hotel are less frequent in cluster 0, while Halal restaurants are less frequent in Cluster 1.. These businesses displayed in above visualization can be considered to be started in corresponding cluster of neighborhoods due to their scarcity.

Apart from business point of view this project can also be used by people moving to (or) relocating inside Chicago. People moving to Chicago can analyze each cluster to determine which kind of neighborhood they would prefer moving to based on their requirements. For example, people who prefer Fast food joints and parks can move to a neighborhood in cluster 0 while those who prefer buying coffee on their way to work in morning can move to neighborhoods in cluster 1 which have a high number of Coffee shops. People relocating inside Chicago can determine to which cluster their current neighborhood belongs to and look for other neighborhoods in the same cluster which would help them relocate to similar neighborhoods.

5. CONCLUSION:

The information regarding each neighborhood was obtained using GoogleMaps and Foursquare API. The K-Means algorithm was used on this data used to determine similarities of the neighborhood and segregate them into 4 clusters based on the same. This project is useful for any aspiring entrepreneurs whom are looking to start any particular business type in Chicago and also for people whom are relocation to/within Chicago.