Coursera Capstone Project

The Battle of the Neighborhoods

1. Introduction

1.1. Background

With a population of more than 2,693,976, Chicago, Illinois, USA, is the third-largest city in the USA. With the highly developed and diversified economy, Chicago is a very attractive city for people looking for work or to start their own business, including small businesses such as coffee shops, bakeries, and restaurants. This also makes entry into the market very competitive. For someone, for example, opening a bakery, choosing the right location for the shop can make or break the business.

1.2. Problem

How someone who wants to open for example a bakery should choose the right location? Naturally, knowing where the competition is located would be important to know, as well as the size of the community they potentially can serve. We can argue that choosing location in communities with no bakeries might be a good choice. However, we also need to look for business who may provide similar products, for example coffee shops sell bakery products too.

1.3. Interest

For purpose of this exercise, I'll will do the analysis and try to predict the best spot to open a bakery. Therefore, this research might be of interest to somebody who wants to open a bakery and needs to know which location might be the best to open such a business.

2. Data Description

For the analysis I'll use **Foursquare API** to get most popular venues in Chicagl. Also, I'll check the location of current bakeries and bakeries within grocery stores.

I'll divide Chicago into communities by using the zip codes of Chicago. This data set can be found on zipatlas.com; this data set also has population numbers by zip code. I'll remove communities with less than 3,000 on the assumption that opening a bakery in this area will not be suitable since foot traffic is insufficient for business to sustain.

Since zipatlas.com data set has only zip codes with no latitude and longitude, I'll use OpenDataSoft to get longitude and latitude for each zip code.

To match community names to their zip code, I'll use unitedstateszipcodes.org data set that has zip codes with corresponding community names.

3. Methodology

3.1. Preparing the Data: Zip Codes, Population, Location and Communities

First, we'll get population by zip code in the state of Illinois. We'll get this data from zapatlas.com. Some of the data we do not need so I removed unnecessary information. Also, I removed zip codes with less than 3000 inhabitants on basses that those communities will not be attractive for opening a bakery.

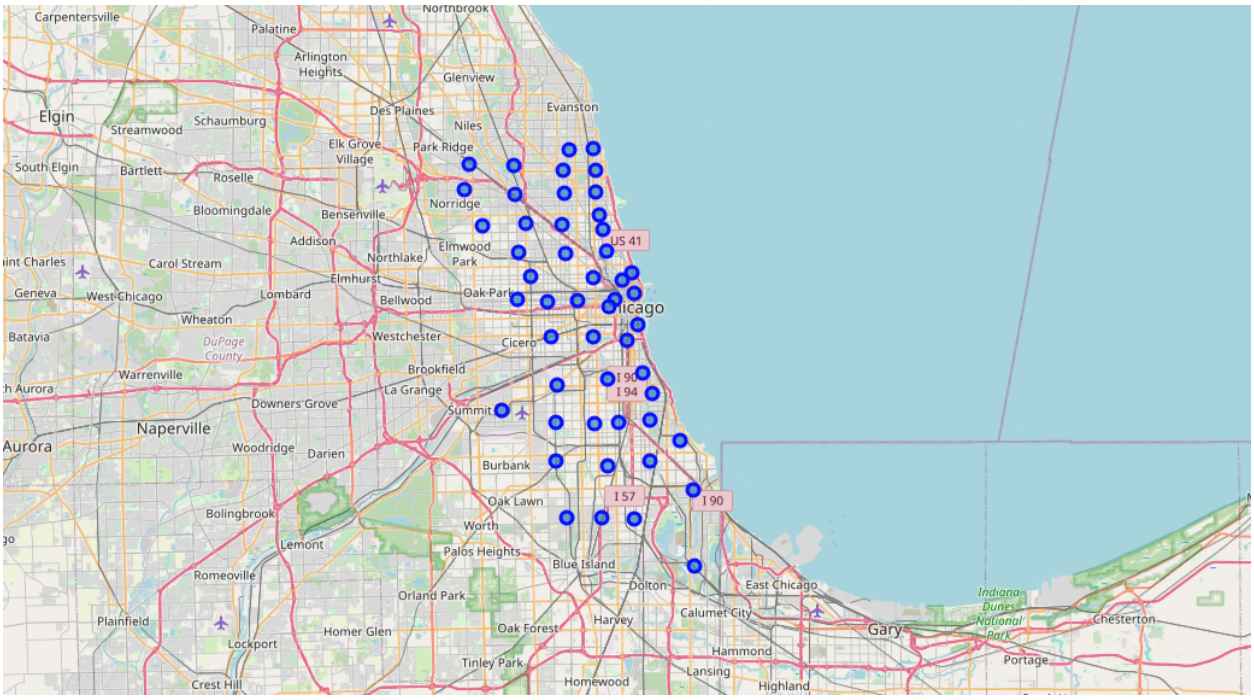
	Zip Code	Population
0	60626	59251
1	60660	47726
2	60640	74030
3	60657	66789
4	60610	47513

Next we'll have to get community names that belong to each zip code as well as longitude and latitude per zip code. To achieve this we need to data sets one drop OpenDataSoft and the other one from unitedstateszipcodes.org.

We'll merge those data frames together with population data frame. The final data frame will conation zip codes, city and community names, population, longitude and latitude. We'll drop communities that do not belong to the city of Chicago.

	Zip Code	Community	Population	City	State	Latitude	Longitude
0	60626	Rogers Park	59251	Chicago	IL	42.009731	-87.66938
1	60660	Edgewater	47726	Chicago	IL	41.990631	-87.66670
2	60640	Uptown	74030	Chicago	IL	41.973181	-87.66650
3	60657	Lake View	66789	Chicago	IL	41.940832	-87.65852
4	60610	Uptown	47513	Chicago	IL	41.898582	-87.63710

Finally, we'll plot Chicago communities on the map using **folium** library.

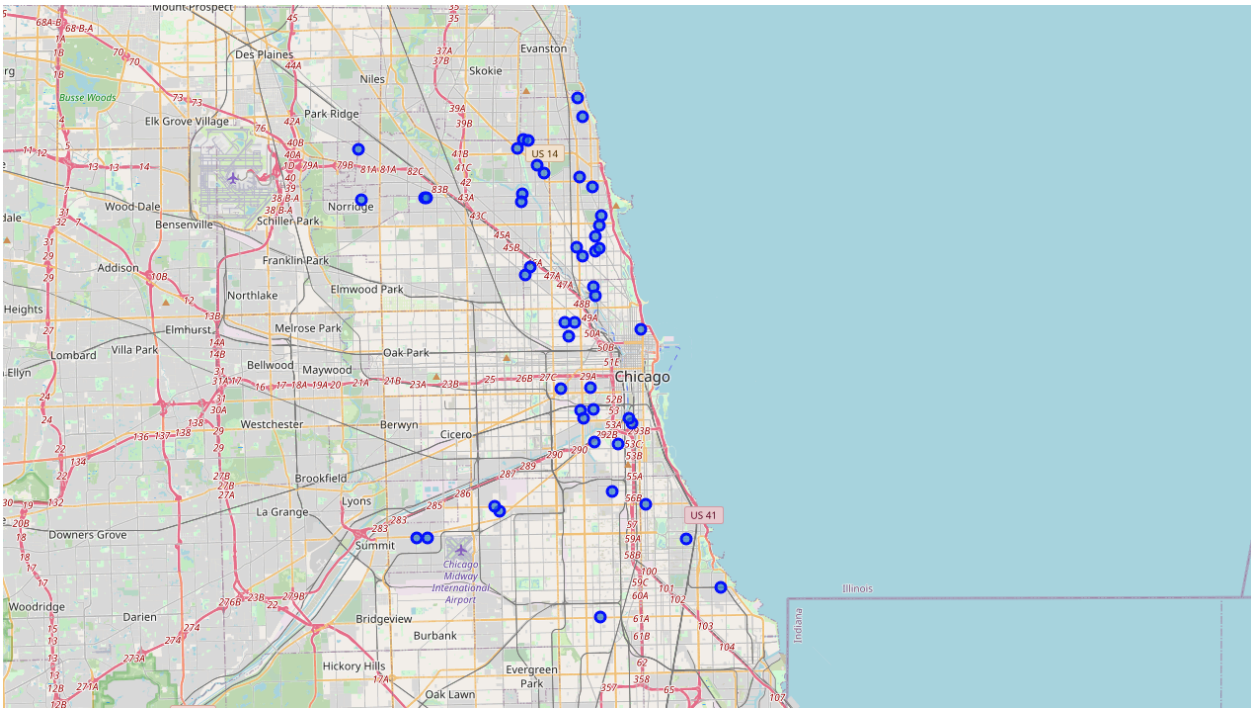


3.2. Popular Venues and Population

Using Foursquare we'll get the list of the most popular venue in the radius of 1500 meters. We'll remove any duplicates that we may have due to the query overlap. Also, we need to merge venues data set with populations and community data frame.

	Community	Population	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park	59251	42.009731	-87.66938	Morse Fresh Market	42.008087	-87.667041	Grocery Store
1	Rogers Park	59251	42.009731	-87.66938	Rogers Park Social	42.007360	-87.666265	Bar
2	Rogers Park	59251	42.009731	-87.66938	The Common Cup	42.007797	-87.667901	Coffee Shop
3	Rogers Park	59251	42.009731	-87.66938	Lifeline Theatre	42.007372	-87.666284	Theater
4	Rogers Park	59251	42.009731	-87.66938	Glenwood Sunday Market	42.008525	-87.666251	Farmers Market

There are 3,796 venues returned by Foursquare of those there are 52 bakeries in the city of Chicago. We'll plot the bakeries on the map to get an idea how they are dispersed around the city.



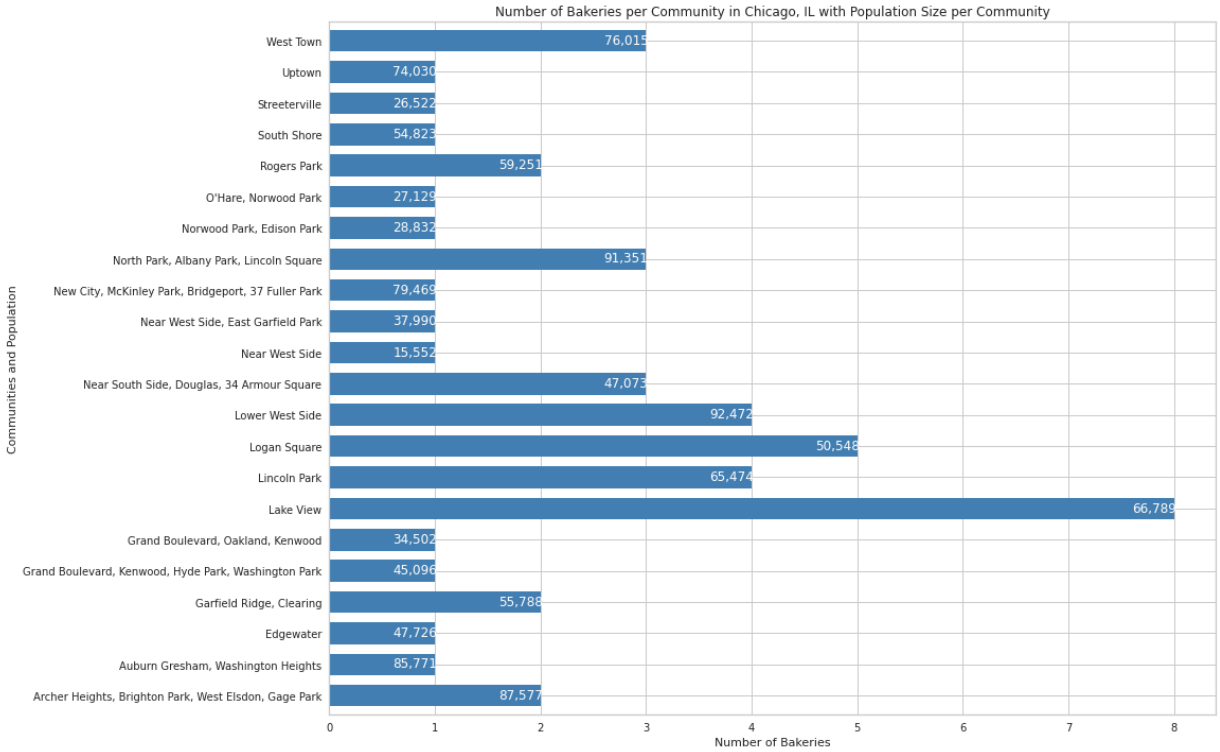
There are more bakeries in the norther comparing to southern part of the city. Also, they are much closer to each other.

3.3. People per Bakery

Next, we'll take a look at how many people each bakery may serve. We'll create a new table that will group bakeries per community.

	Community	Number of Bakeries per Community	Population	Number of People per Bakery
0	Archer Heights, Brighton Park, West Elsdon, Ga...	2	87577	43788.0
1	Auburn Gresham, Washington Heights	1	85771	85771.0
2	Edgewater	1	47726	47726.0
3	Garfield Ridge, Clearing	2	55788	27894.0
4	Grand Boulevard, Kenwood, Hyde Park, Washingto...	1	45096	45096.0

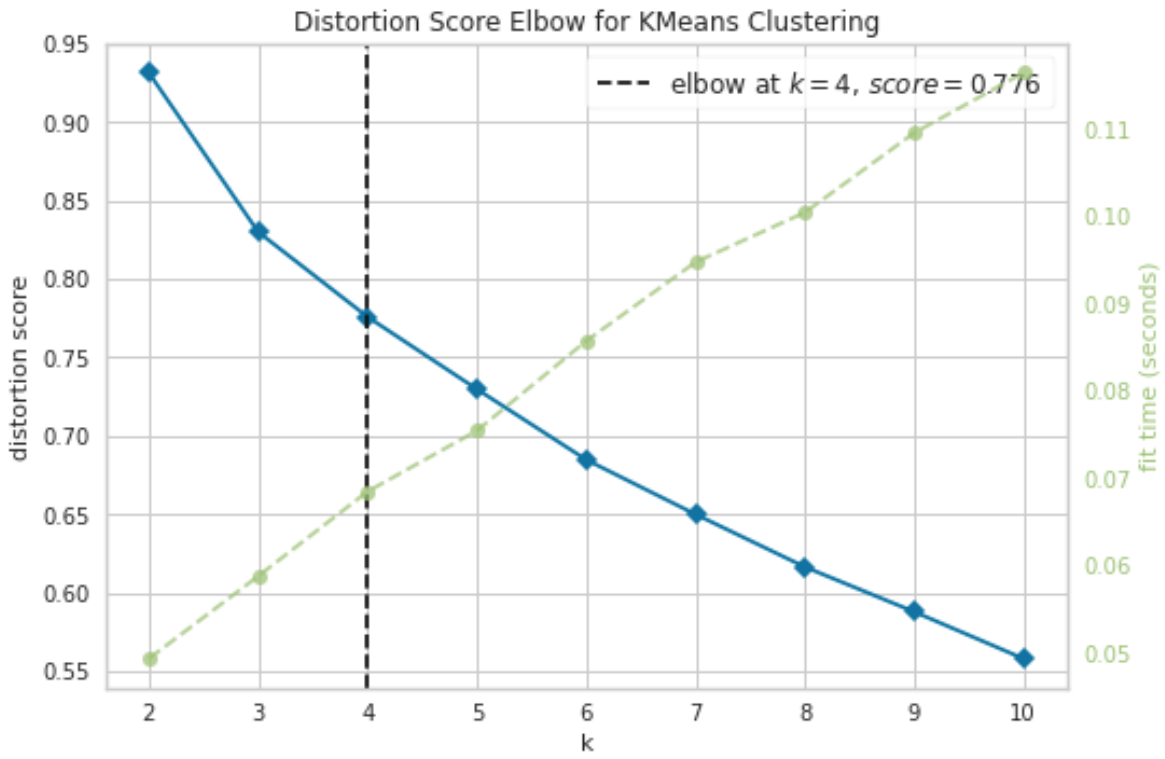
The minimum number of people that the bakery may serve is 8,349 (Lake View community) and a maximum of 85,771 (Auburn Gresham, Washington Heights). On the graph below we can see the relationship between number of bakeries per community and population per community. Lake View has 8 bakeries in total while there are 12 communities with only 1 bakery.



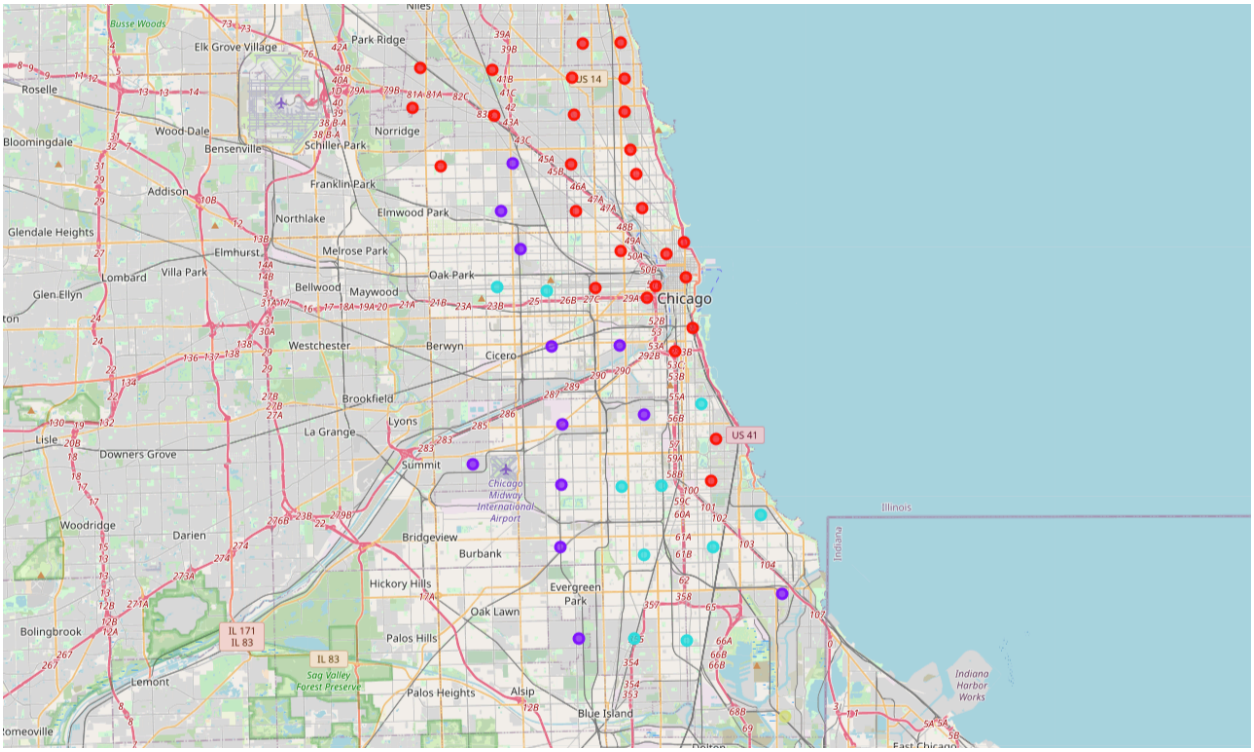
3.4. K-Means Clustering

Finally, we'll use the information we have regarding venues and their popularity to group Chicago's communities. This may help understand shopping behaviours that people in the community may have.

To determine the number of clusters we'll run score elbow for KMeans clustering.



Next, we run KMeans machine learning algorithm, visualize and analyze the clusters.



Of 52 bakeries, 11 show on the tip 10 list of the most popular venues per community. Of those 11, 9 are in the Cluster 1, represented with red color. The most popular venue in this cluster is Coffee Shop. In Cluster 2 (purple), there are 2 bakeries, and Cluster 3 and 4 has none. The most common type of venue in Cluster 2 are restaurants.

4. Results and Discussions

There are 52 bakeries in Chicago and 11 of those show as popular venues in cluster of Chicago communities.

Communities Auber Gresham, Washington Heights, Uptown and New City has only 1 bakery and population with around 80,000. The highest concentration of bakeries is in Lake View, 8 and with population of around 66,000 those bakeries are in the business where they may serve around 8,000 people each.

Analysing Cluster, Cluster 1 ha significant number of bakeries comparing with other clusters, 9; with Cluster 2 of 2 and Cluster 3 and 4 has no bakeries as poplar venues. It is unclear what is the reason behind this. Cluster 1 has Coffee Shop as the most popular venue. Coffee shops may be indirect competitor to the bakeries since some of the complimentary products to the coffee they are selling are bakery products. On the other hand, depending on the ownership of those coffee shops, they can also be potential buyers of bakery products, so it may help new bakery grow the business. More studies are required to understand this phenomenon.

5. Conclusion

It is difficult to conclude where would be the best location to open a bakery in Chicago. We can argue that going to the communities that don’t have many bakeries would a good choice since there is less of direct competition.

However, more analysis is needed to understand not just popular venues but also business in the area and how they may affect a new bakery business. Also, understanding cost to rent or buy the venue for the new bakery is very important since it can affect the cost of the products the bakery is selling. Last but not least bakery product cost analysis is needed to understand profit margins and what affect they may have on the new bakery business.