Using Unsupervised Machine Learning Technology to Determine the Best Location for Opening a Steakhouse in Chicago

Applied Data Science Capstone Project

Author: Kui Shen

Date: May 2nd, 2021

Table of Content

Part I – Introduction

§1.1 - Description of a Business Problem and its Background

Since 1865 with the founding of the Union Stock Yards, Chicago has been at the heart of high-quality American meats. Steak aficionados have long acclaimed Chicago beefsteaks being among the top grade nationwide. Tourists are flocking in to landmark steakhouses in the Windy City. Although there are already hundreds of steakhouses in River North alone, launching a competitive modern steakhouse in Chicago could still be lucrative and help enrich the city's culture and history.

While thinking about an idea like that can be exciting even when the COVID-19 Pandemic hasn't really become a thing of the past, implementing it would be a much different story. One particularly vexing problem is – where exactly to open it?

As you may have heard repeatedly, "The three most important words for real estate are - location, location, and location" – Yes, this is 100% true for a steakhouse as well. Location determines volume of customer, cost of space, safety, competition, and much more. Anyone who wants to open a steakhouse will have to take it very seriously.

Using Unsupervised Machine Learning (UML) technologies recently acquired from Coursera's Machine Learning with Python and Applied Data Science Capstone courses, I conducted a thorough study in an effort to provide useful insights to this challenging problem.

§1.2 - Who Will be Interested

For someone who wants to open a steakhouse in Chicago but can't really figure out where to launch the business, this analysis can help. Systematic UML approaches are adopted throughout this project, one can be confident about the results and conclusion we are going to unveil.

Furthermore, if the business problem becomes, for example, where to open a hair salon or a fitness center (i.e., a different venue), this methodology and most of the Python codes in this project can be reused exactly in their entirety.

Part II – Data Description

Data used in this project are mainly derived from three major sources:

- Wikipedia Websites
- Python GeoPy Library
- Foursquare venue database

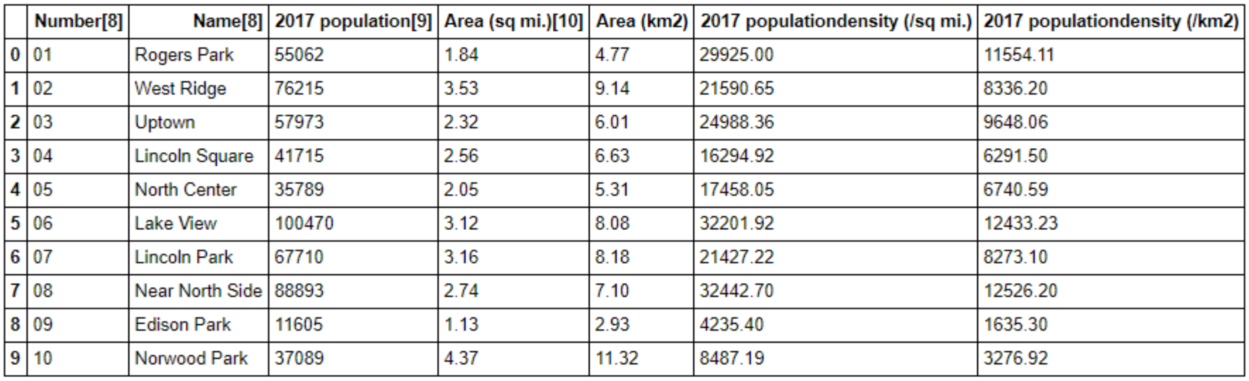
§2.1 - Wikipedia Websites

Tons of Chicago related information can be found on the internet. After careful comparing and selecting, I finally choose to use the Wikipedia webpages for Chicago neighborhood and community information, such as the following:

- https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago
- https://en.wikipedia.org/wiki/Community_areas_in_Chicago

Being the nation's 3rd largest city, Chicago has 284 neighborhoods which combined to form its 77 official communities. While dealing with 284 neighborhoods isn't particular overwhelming for this project, most publicly available Chicago city data are on community level rather than on neighborhood level. Therefore, this analysis is based on community level data.

The 2nd link above presented a table that contains Chicago Community Name, Population, Area Size, Population Density, etc. It can readily be used. I attached an image of this table (Top 10 rows) after it has been read into Jupyter Notebook by Python, see Figure 1 below:



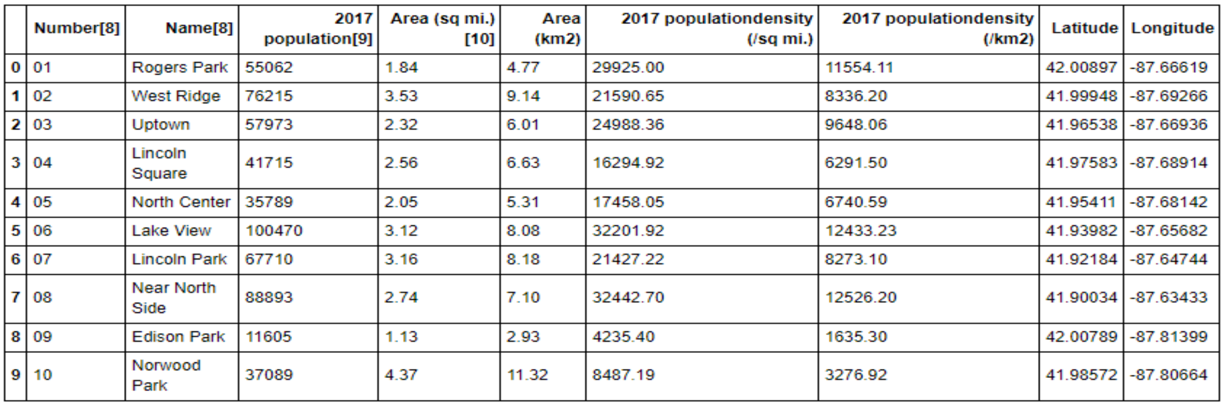
	Number[8]	Name[8]	2017 population[9]	Area (sq mi.)[10]	Area (km2)	2017 populationdensity (/sq mi.)	2017 populationdensity (/km2)
0	01	Rogers Park	55062	1.84	4.77	29925.00	11554.11
1	02	West Ridge	76215	3.53	9.14	21590.65	8336.20
2	03	Uptown	57973	2.32	6.01	24988.36	9648.06
3	04	Lincoln Square	41715	2.56	6.63	16294.92	6291.50
4	05	North Center	35789	2.05	5.31	17458.05	6740.59
5	06	Lake View	100470	3.12	8.08	32201.92	12433.23
6	07	Lincoln Park	67710	3.16	8.18	21427.22	8273.10
7	08	Near North Side	88893	2.74	7.10	32442.70	12526.20
8	09	Edison Park	11605	1.13	2.93	4235.40	1635.30
9	10	Norwood Park	37089	4.37	11.32	8487.19	3276.92

Figure 1: Snapshot of Chicago Community Overview

§2.2 - Python GeoPy Library

As one may have noticed, the community table in 2.1 doesn't include geographical coordinates for the communities. These coordinates are necessary when one needs to fetch venue information later on. Fortunately, the Python GeoPy library (in particular the geocode module) can be employed to obtain the latitude and longitude data for the city of Chicago as well as its 77 official communities. I will then create Python codes to concatenate these coordinates to the community table.

Figure 2 below shows what the community table looks like when concatenated with their corresponding geographical coordinates (Top 10 rows):



	Number[8]	Name[8]	2017 population[9]	Area (sq mi.) [10]	Area (km2)	2017 populationdensity (/sq mi.)	2017 populationdensity (/km2)	Latitude	Longitude
0	01	Rogers Park	55062	1.84	4.77	29925.00	11554.11	42.00897	-87.66619
1	02	West Ridge	76215	3.53	9.14	21590.65	8336.20	41.99948	-87.69266
2	03	Uptown	57973	2.32	6.01	24988.36	9648.06	41.96538	-87.66936
3	04	Lincoln Square	41715	2.56	6.63	16294.92	6291.50	41.97583	-87.68914
4	05	North Center	35789	2.05	5.31	17458.05	6740.59	41.95411	-87.68142
5	06	Lake View	100470	3.12	8.08	32201.92	12433.23	41.93982	-87.65682
6	07	Lincoln Park	67710	3.16	8.18	21427.22	8273.10	41.92184	-87.64744
7	08	Near North Side	88893	2.74	7.10	32442.70	12526.20	41.90034	-87.63433
8	09	Edison Park	11605	1.13	2.93	4235.40	1635.30	42.00789	-87.81399
9	10	Norwood Park	37089	4.37	11.32	8487.19	3276.92	41.98572	-87.80664

Figure 2: Snapshot of Chicago Communities with Geographical Coordinates

Below is a map of Chicago communities (generated based on the obtained data and using Folium Map):

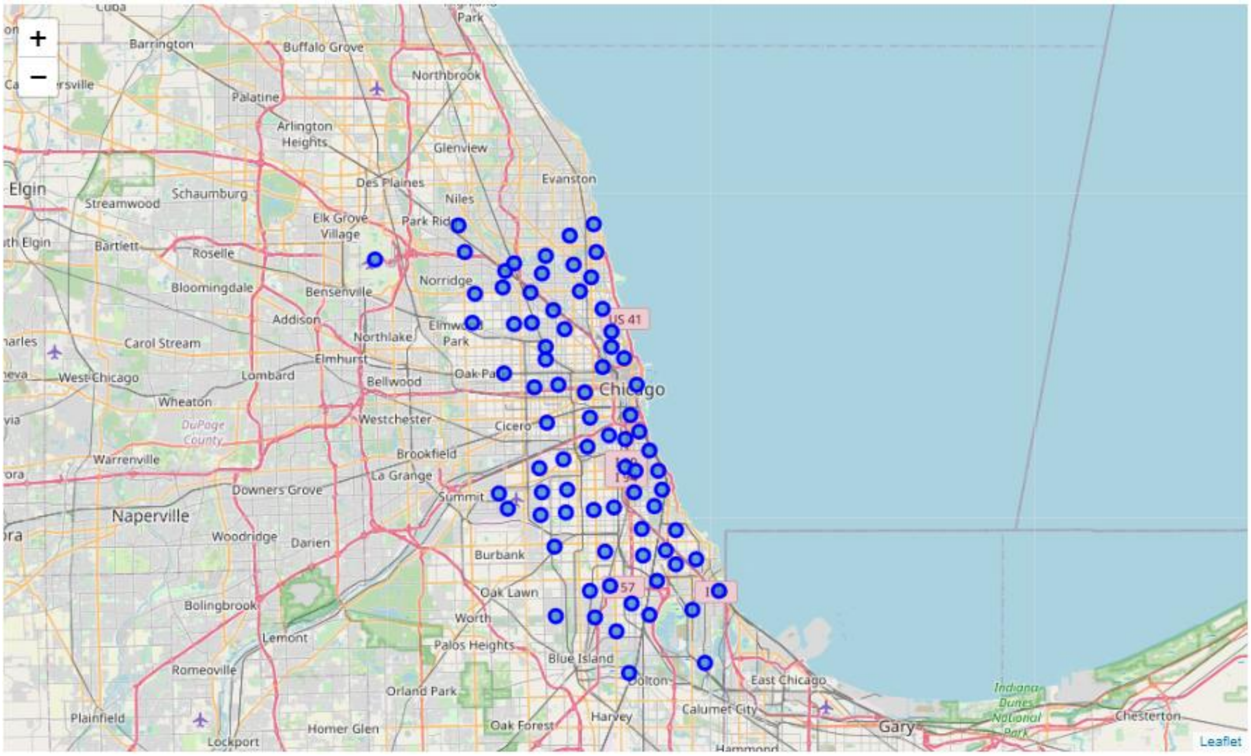


Figure 3: Map of Chicago Communities

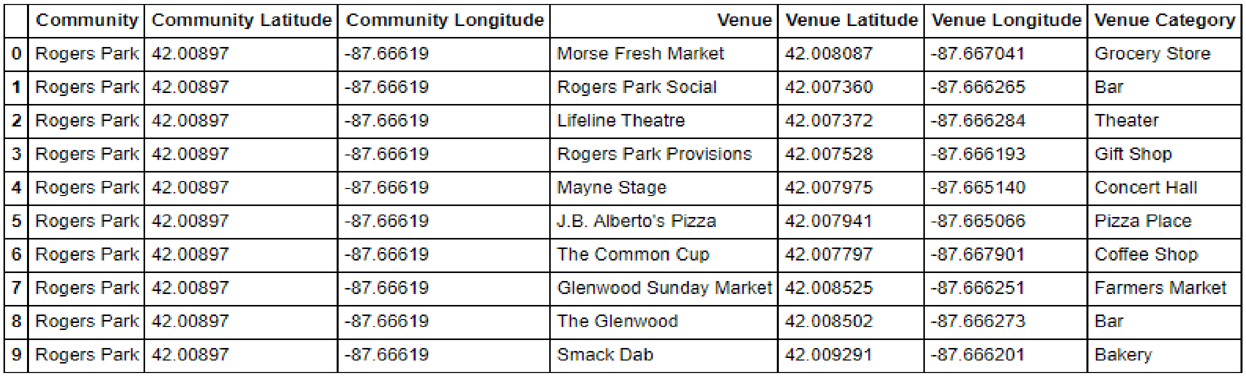
§2.3 - Foursquare API for Venue Information

Venue information can be obtained via several different sources, and each has its own Pros and Cons. Foursquare is becoming increasingly popular and we just learned it in the Applied Data Science Capstone Course. So I decide to use Foursquare API to obtain venue information for each of the 77 Chicago communities. These venues will include basically everything – supermarkets, bars, restaurants, parks, gyms, libraries, etc.

Once obtained, the information is used to rank the popular venues for each community. Based on that, an unsupervised machine learning technology known as K-Means Clustering can be utilized to group the communities into several clusters. The clusters are expected to contain indicative information as to whether or not a steakhouse is a likely fit here, and I will be able to eliminate all but one cluster.

Finally, I will use some other criteria (such as the ratio of population over # of steakhouse and published community reviews) to determine which community is the best candidate for opening a steakhouse.

Figure 3 below shows an example of some venues in one community (Top 10 rows):



	Community	Community Latitude	Community Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park	42.00897	-87.66619	Morse Fresh Market	42.008087	-87.667041	Grocery Store
1	Rogers Park	42.00897	-87.66619	Rogers Park Social	42.007360	-87.666265	Bar
2	Rogers Park	42.00897	-87.66619	Lifeline Theatre	42.007372	-87.666284	Theater
3	Rogers Park	42.00897	-87.66619	Rogers Park Provisions	42.007528	-87.666193	Gift Shop
4	Rogers Park	42.00897	-87.66619	Mayne Stage	42.007975	-87.665140	Concert Hall
5	Rogers Park	42.00897	-87.66619	J.B. Alberto's Pizza	42.007941	-87.665066	Pizza Place
6	Rogers Park	42.00897	-87.66619	The Common Cup	42.007797	-87.667901	Coffee Shop
7	Rogers Park	42.00897	-87.66619	Glenwood Sunday Market	42.008525	-87.666251	Farmers Market
8	Rogers Park	42.00897	-87.66619	The Glenwood	42.008502	-87.666273	Bar
9	Rogers Park	42.00897	-87.66619	Smack Dab	42.009291	-87.666201	Bakery

Figure 4: Snapshot of Chicago Community Venues Obtained via Foursquare API

Part III – Methodology

§3.1 – Methodology Overview

This project utilizes unsupervised machine Learning K-Means clustering methodology to address a practical business problem – where to open a steakhouse in Chicago. In order to do that, a systematic approach is adopted to achieve sound results. The major steps involved are outlined as follows:

1) Obtain necessary data (discussed in Part II);
2) Analyze target area (e.g., the city of Chicago) by feature extraction, discover specific venue distributions within each community using a data analysis technique called segmentation;
3) Based on segmentation results, categorize the 77 communities into several mutually exclusive groups called clusters. This will be performed using a data science unsupervised machine learning technique known as clustering. Then based on clustering results we'll determine the most suitable cluster for the targeted venue (e.g., a steakhouse).

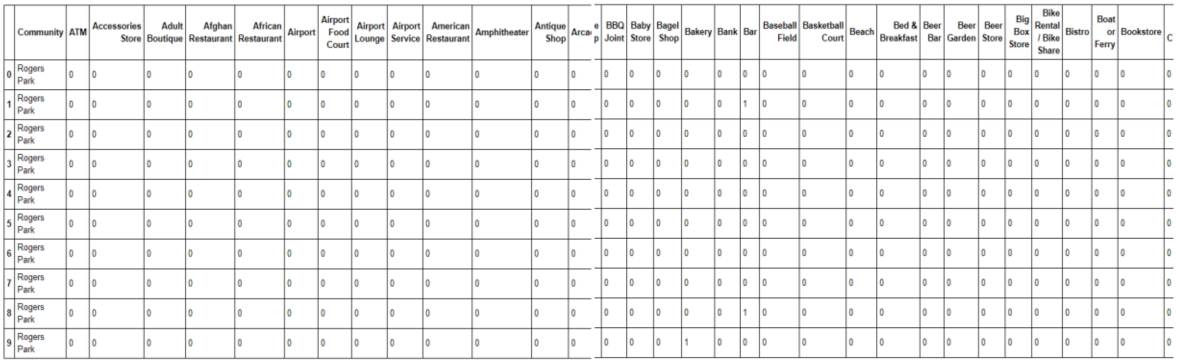
4) Finally, select the best community in that cluster based on a set of self-defined venue-specific criteria (will be discussed in Part IV).

In this chapter we will mainly discuss segmentation and clustering.

§3.2 – Segmentation Process

For the Chicago community data augmented with geographical coordinates, Foursquare API (along with my account credentials) was used to get the community venue information (shown in Figure 3 above).

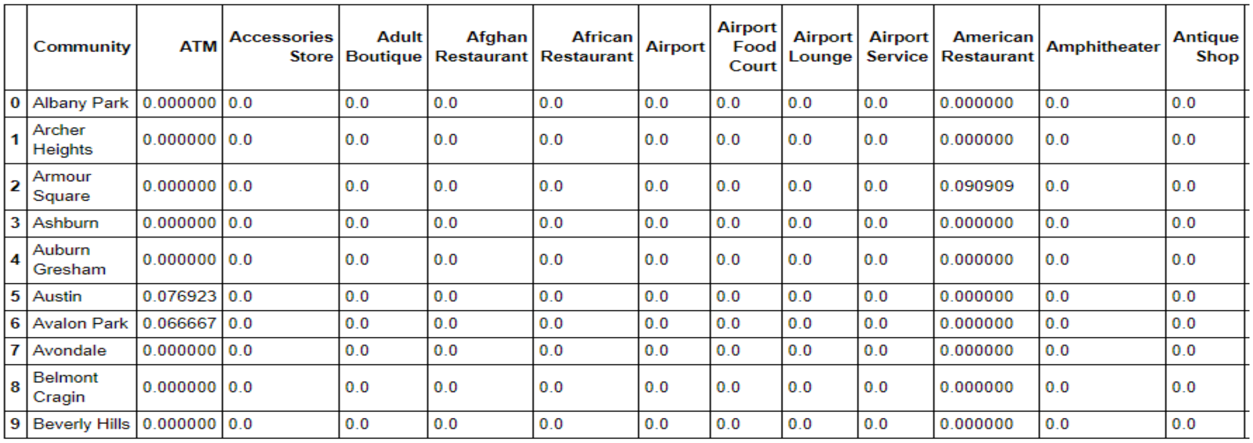
For segmentation, we need to study venue distributions in each community and I used the onehot encoding method to achieve that. Specifically, the Pandas get_dummies function can be used to obtain venue category - marking 1 under a particular venue if the community has this venue and 0 if it doesn't. Result is a big sparse matrix with many 0's and some 1's. Like this (showing only a tiny portion):



	Community	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	American Restaurant	Amphitheater	Antique Shop	Arca	e p	BBQ Joint	Baby Store	Bagel Shop	Bakery	Bank	Bar	Baseball Field	Basketball Court	Beach	Bed & Breakfast	Beer Bar	Beer Garden	Beer Store	Big Box Store	Bike Rental / Bike Share	Bistro	Boat or Ferry	Bookstore	C
0	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Rogers Park	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5: Snapshot of Chicago Community Venues Onehot Encoding

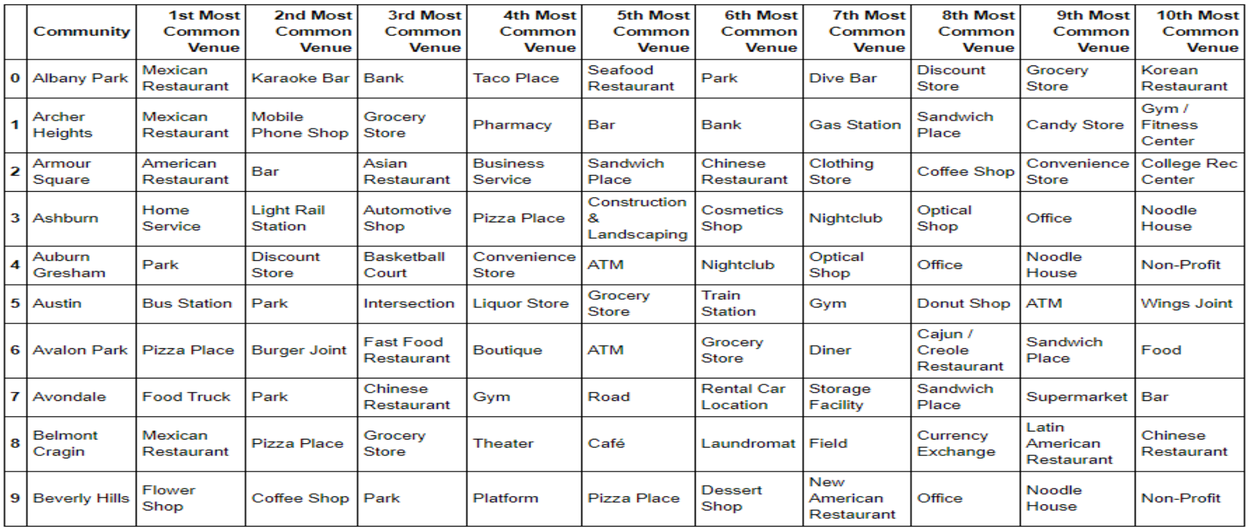
Notice that each row must have one "1" and can only have one "1". Then for each community, calculate the venue mean for all existing venues. For example, community XYZ has a total of 200 venues of many kinds. Of the 200 venues 17 are bars, then the bar mean for community XYZ is 17/200 = 0.085000. After performing venue mean calculations we got a community venue mean summary like this:



	Community	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	American Restaurant	Amphitheater	Antique Shop
0	Albany Park	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	Archer Heights	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	Armour Square	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.090909	0.0	0.0
3	Ashburn	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
4	Auburn Gresham	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
5	Austin	0.076923	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
6	Avalon Park	0.066667	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
7	Avondale	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
8	Belmont Cragin	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
9	Beverly Hills	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

Figure 6: Snapshot of Chicago Community Venues Means

Now that we have obtained this critical information about venue distribution, we can rank the venues and come up with a Top-10 venue list in each community:



	Community	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	Mexican Restaurant	Karaoke Bar	Bank	Taco Place	Seafood Restaurant	Park	Dive Bar	Discount Store	Grocery Store	Korean Restaurant
1	Archer Heights	Mexican Restaurant	Mobile Phone Shop	Grocery Store	Pharmacy	Bar	Bank	Gas Station	Sandwich Place	Candy Store	Gym / Fitness Center
2	Armour Square	American Restaurant	Bar	Asian Restaurant	Business Service	Sandwich Place	Chinese Restaurant	Clothing Store	Coffee Shop	Convenience Store	College Rec Center
3	Ashburn	Home Service	Light Rail Station	Automotive Shop	Pizza Place	Construction & Landscaping	Cosmetics Shop	Nightclub	Optical Shop	Office	Noodle House
4	Auburn Gresham	Park	Discount Store	Basketball Court	Convenience Store	ATM	Nightclub	Optical Shop	Office	Noodle House	Non-Profit
5	Austin	Bus Station	Park	Intersection	Liquor Store	Grocery Store	Train Station	Gym	Donut Shop	ATM	Wings Joint
6	Avalon Park	Pizza Place	Burger Joint	Fast Food Restaurant	Boutique	ATM	Grocery Store	Diner	Cajun / Creole Restaurant	Sandwich Place	Food
7	Avondale	Food Truck	Park	Chinese Restaurant	Gym	Road	Rental Car Location	Storage Facility	Sandwich Place	Supermarket	Bar
8	Belmont Cragin	Mexican Restaurant	Pizza Place	Grocery Store	Theater	Café	Laundromat	Field	Currency Exchange	Latin American Restaurant	Chinese Restaurant
9	Beverly Hills	Flower Shop	Coffee Shop	Park	Platform	Pizza Place	Dessert Shop	New American Restaurant	Office	Noodle House	Non-Profit

Figure 7: Snapshot of Chicago Community Top-10 Venues

This would be the end of the segmentation process. Clustering will start next.

§3.3 – Clustering Process

Clustering is a popular unsupervised machine learning technology. The main idea is to partition a given data set into several mutually exclusive groups called clusters, having data points in the same cluster be close enough to their center (aka centroid). Note that "close" here doesn't always mean physical distance. It often refers to a generalized distance in multi-dimensional space. There are several ways of performing clustering and I used the UML KMeans Clustering for this project (see reasons in Part V & VI).

The first step of any clustering is to determine the # of cluster. This is usually not an easy task and some arbitrariness will inevitably be involved. I used Silhouette Score and Elbow Method to help determine the optimal # of cluster. Python codes generated the following plotting:

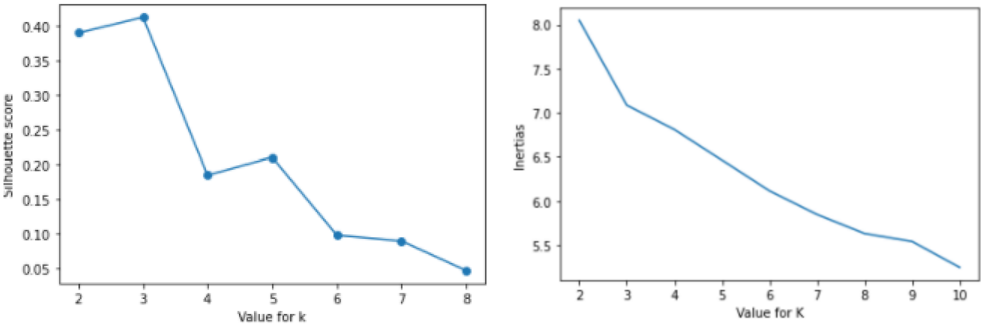
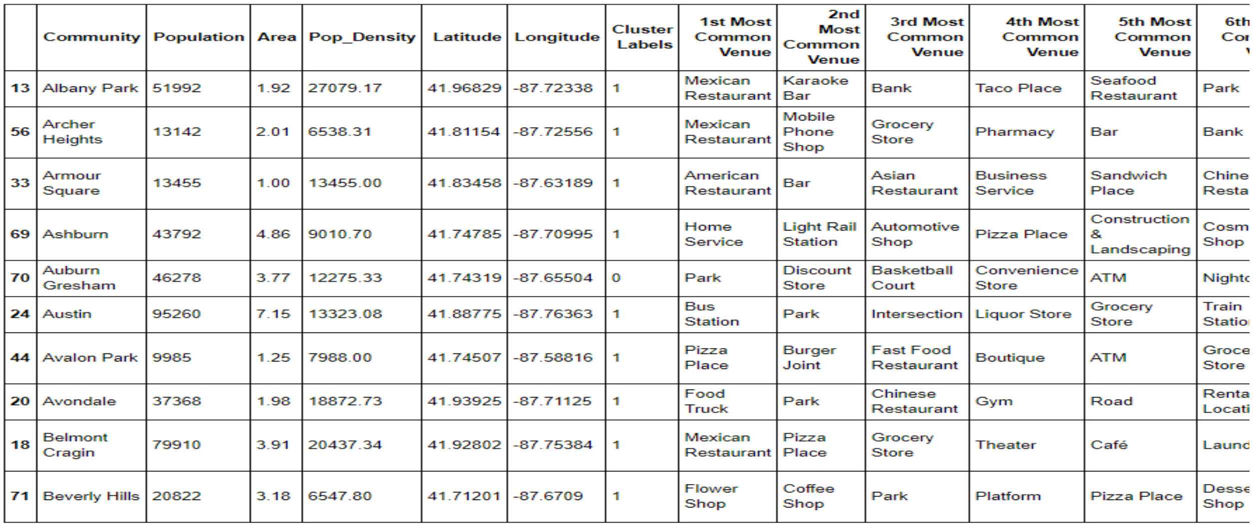


Figure 8: Silhouette Score Plotting (left) and Elbow Method Plotting (right)

As we can see, k=3 has the highest silhouette score and k=3 is also where elbow occurs (elbow is the point where steepest decrease in slope occurs). Let's note that the silhouette is a randomized process and results can be different between runs, while the elbow is not randomized, it is a function of the underlying dataset and result is fixed. Therefore, elbow method is considered more reliable. My submitted Jupyter Notebook has detailed explanations on how to gauge the two different methods.

We can now proceed to KMeans Clustering using 3-cluster. Python sklearn.cluster library contains the KMeans model to fit the data. After fitting and attaching cluster labels, Chicago community clustering is generated: (showing a portion of it):



	Community	Population	Area	Pop_Density	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Co
13	Albany Park	51992	1.92	27079.17	41.96829	-87.72338	1	Mexican Restaurant	Karaoke Bar	Bank	Taco Place	Seafood Restaurant	Park
56	Archer Heights	13142	2.01	6538.31	41.81154	-87.72556	1	Mexican Restaurant	Mobile Phone Shop	Grocery Store	Pharmacy	Bar	Bank
33	Armour Square	13455	1.00	13455.00	41.83458	-87.63189	1	American Restaurant	Bar	Asian Restaurant	Business Service	Sandwich Place	Chine Resta
69	Ashburn	43792	4.86	9010.70	41.74785	-87.70995	1	Home Service	Light Rail Station	Automotive Shop	Pizza Place	Construction & Landscaping	Cosm Shop
70	Auburn Gresham	46278	3.77	12275.33	41.74319	-87.65504	0	Park	Discount Store	Basketball Court	Convenience Store	ATM	Nightc
24	Austin	95260	7.15	13323.08	41.88775	-87.76363	1	Bus Station	Park	Intersection	Liquor Store	Grocery Store	Train Statio
44	Avalon Park	9985	1.25	7988.00	41.74507	-87.58816	1	Pizza Place	Burger Joint	Fast Food Restaurant	Boutique	ATM	Groce Store
20	Avondale	37368	1.98	18872.73	41.93925	-87.71125	1	Food Truck	Park	Chinese Restaurant	Gym	Road	Renta Locati
18	Belmont Cragin	79910	3.91	20437.34	41.92802	-87.75384	1	Mexican Restaurant	Pizza Place	Grocery Store	Theater	Café	Laund
71	Beverly Hills	20822	3.18	6547.80	41.71201	-87.6709	1	Flower Shop	Coffee Shop	Park	Platform	Pizza Place	Desse Shop

Figure 9: Snapshot of Chicago Community Clusters Table

We can plot a graph of the Chicago community clusters using Folium Map:

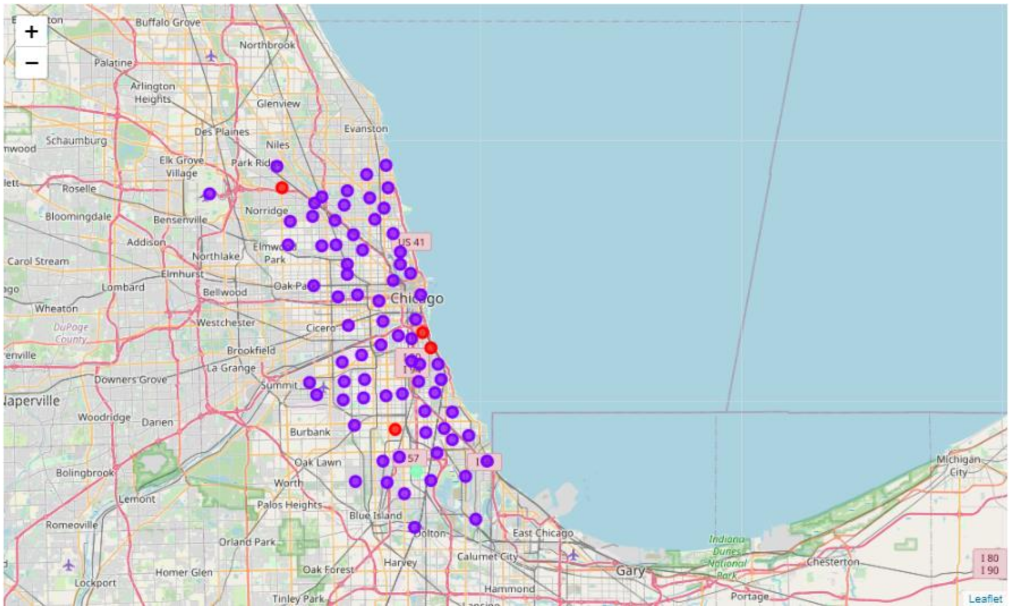
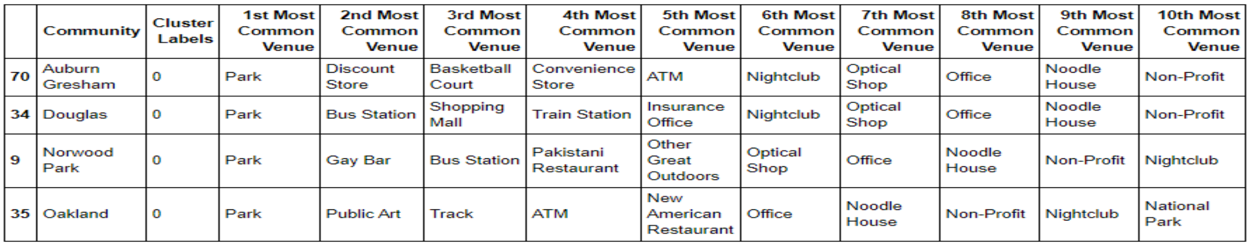


Figure 10: Map of Chicago Community Clusters

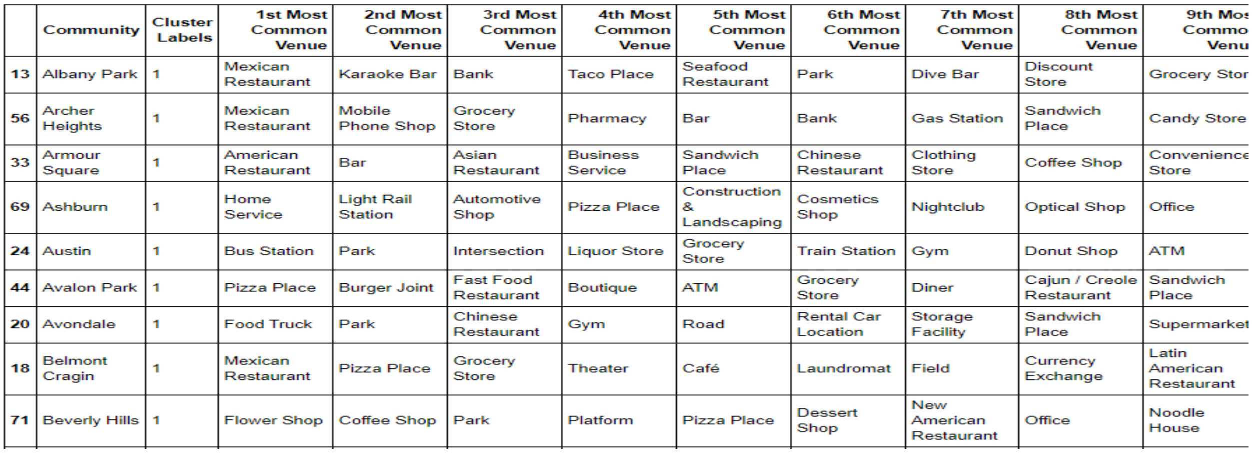
Since we have successfully partitioned the 77 Chicago communities into 3 clusters, we can now analyze each cluster and see if a steakhouse can be a good fit there. Detailed analysis and coding can be found in the notebook. Here is a quick summary:

- Cluster 0 doesn't seem to be a good fit as the most popular venue in all communities are parks. It is much like a big recreation area rather than somewhere people would go for a formal dinner.
- Cluster 1 seems to be a good fit, as it has basically everything a decent city life entitles.
- Cluster 2 doesn't seem to be a good fit either, as it contains only one community and its most popular venue doesn't seem relevant. We don't know what that is and don't want to risk our investment there.



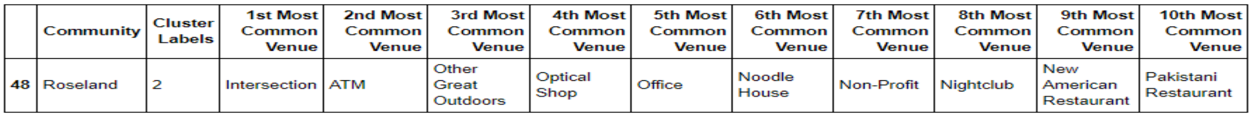
	Community	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
70	Auburn Gresham	0	Park	Discount Store	Basketball Court	Convenience Store	ATM	Nightclub	Optical Shop	Office	Noodle House	Non-Profit
34	Douglas	0	Park	Bus Station	Shopping Mall	Train Station	Insurance Office	Nightclub	Optical Shop	Office	Noodle House	Non-Profit
9	Norwood Park	0	Park	Gay Bar	Bus Station	Pakistani Restaurant	Other Great Outdoors	Optical Shop	Office	Noodle House	Non-Profit	Nightclub
35	Oakland	0	Park	Public Art	Track	ATM	New American Restaurant	Office	Noodle House	Non-Profit	Nightclub	National Park

Figure 11: Chicago Community Cluster # 0 – Recreation Paradise



	Community	Cluster Labels	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 Mos Commo Venu
13	Albany Park	1	Mexican Restaurant	Karaoke Bar	Bank	Taco Place	Seafood Restaurant	Park	Dive Bar	Discount Store	Grocery Stor
56	Archer Heights	1	Mexican Restaurant	Mobile Phone Shop	Grocery Store	Pharmacy	Bar	Bank	Gas Station	Sandwich Place	Candy Store
33	Armour Square	1	American Restaurant	Bar	Asian Restaurant	Business Service	Sandwich Place	Chinese Restaurant	Clothing Store	Coffee Shop	Convenience Store
69	Ashburn	1	Home Service	Light Rail Station	Automotive Shop	Pizza Place	Construction & Landscaping	Cosmetics Shop	Nightclub	Optical Shop	Office
24	Austin	1	Bus Station	Park	Intersection	Liquor Store	Grocery Store	Train Station	Gym	Donut Shop	ATM
44	Avalon Park	1	Pizza Place	Burger Joint	Fast Food Restaurant	Boutique	ATM	Grocery Store	Diner	Cajun / Creole Restaurant	Sandwich Place
20	Avondale	1	Food Truck	Park	Chinese Restaurant	Gym	Road	Rental Car Location	Storage Facility	Sandwich Place	Supermarket
18	Belmont Cragin	1	Mexican Restaurant	Pizza Place	Grocery Store	Theater	Café	Laundromat	Field	Currency Exchange	Latin American Restaurant
71	Beverly Hills	1	Flower Shop	Coffee Shop	Park	Platform	Pizza Place	Dessert Shop	New American Restaurant	Office	Noodle House

Figure 12: Chicago Community Cluster #1 (showing a portion) – City Dwelling Wander Land



	Community	Cluster Labels	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
48	Roseland	2	Intersection	ATM	Other Great Outdoors	Optical Shop	Office	Noodle House	Non-Profit	Nightclub	New American Restaurant	Pakistani Restaurant

Figure 13: Chicago Community Cluster #2 – Wild Wild West

We have determined the most suitable cluster (#1), and will address the business problem in Part IV.

Part IV – Result

In the beginning of the project, a business problem was raised – where to open a steakhouse in Chicago. The answer should have reference to a specific community, rather than a broader area (i.e., a cluster). We have determined the most suitable cluster (Cluster 1) using UML KMeans clustering technology. However, there are still quite a few communities in Cluster 1. What to do next? A good practice is to come up with a way to continue to narrow down the number of communities. I did the following:

1. Calculated the ratio of population over # of steakhouse for each community (i.e., added a column "Pop_per_Steakhouse"). By examining the min, max, and average of these ratios, I framed an ideal ratio range (500, 2000). Under 500 may indicate fierce competition, over 2000 may signify low popularity and/or low affordability. See stats in Figure 14:

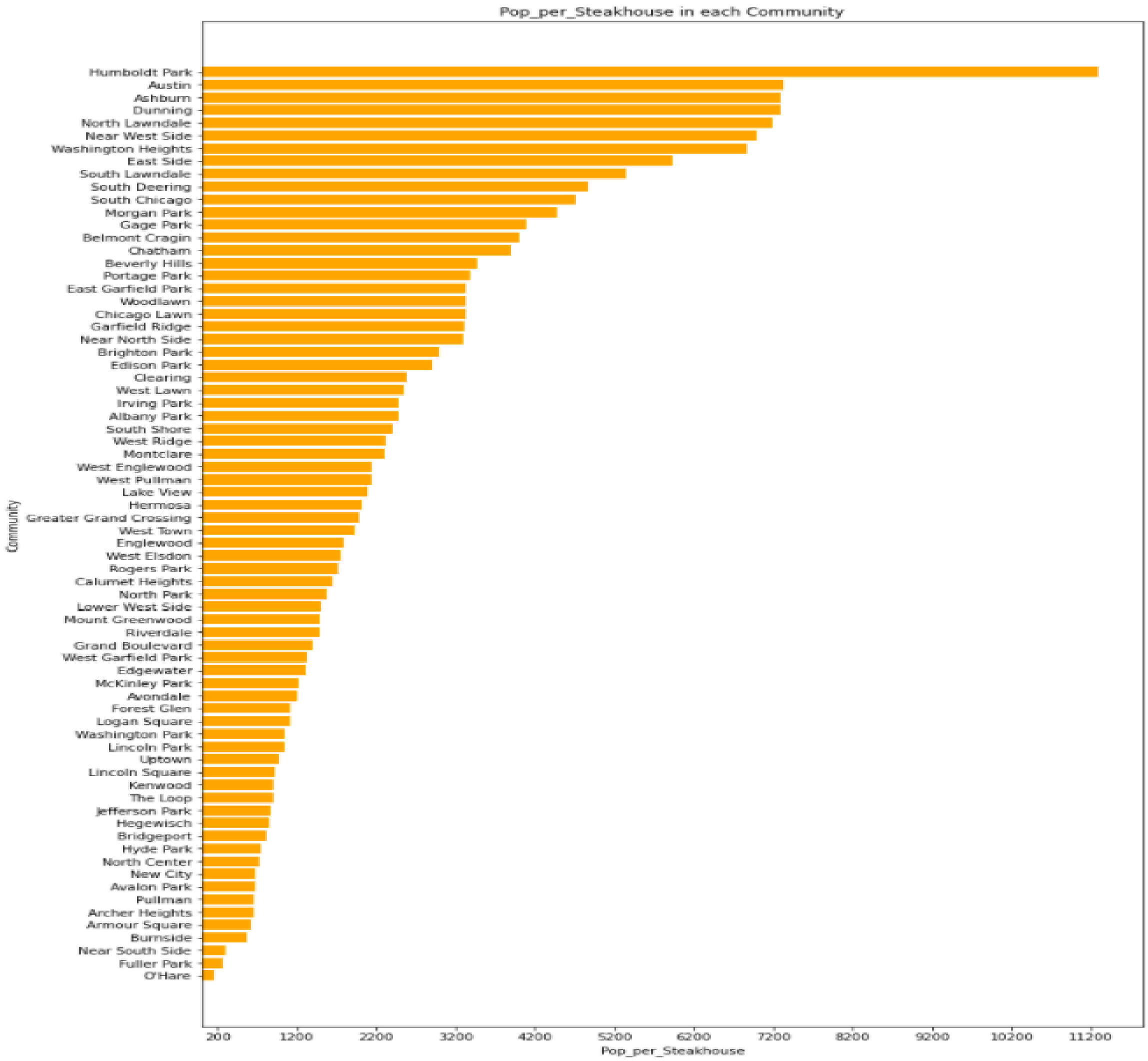
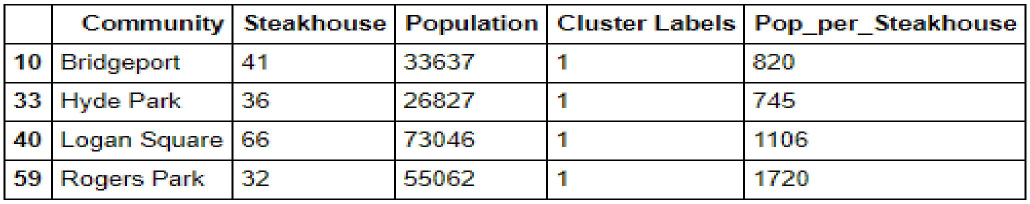


Figure 14: Chicago Community Cluster 1 with Population per Steakhouse stats.

2. Checked online Chicago community reviews. There are tons of them and I selected 2:
    a. Bad communities - https://usaestaonline.com/most-dangerous-neighborhoods-in-chicago
    b. Good communities - https://theculturetrip.com/north-america/usa/illinois/articles/the-10-coolest-neighborhoods-in-chicago/

It is wise to avoid bad communities. Finally, I was able to narrow down to 4 communities:



	Community	Steakhouse	Population	Cluster Labels	Pop_per_Steakhouse
10	Bridgeport	41	33637	1	820
33	Hyde Park	36	26827	1	745
40	Logan Square	66	73046	1	1106
59	Rogers Park	32	55062	1	1720

Figure 15: Ideal Chicago Communities for Launching a New Steakhouse Business

Part V – Discussion

§5.1 – Recommendation Based on the Result

As we can see, Rogers Park has the highest Pop_per_Steakhouse (i.e., 1720) - It has the least number of steakhouse, but the second largest population among the four communities. For a steakhouse business, these facts most likely mean:

- Less competition in Rogers Park than the other 3 communities;
- Larger growth potential in Rogers Park than the other 3 communities.

Both are the ideal traits for a winner candidate. Therefore, according to the analysis, I would recommend Rogers Park Community to be the best place for opening a steakhouse in Chicago.

§5.2 – Pros and Cons of UML KMeans Clustering Methodology

§5.2.1 – Pros

Pros #1 – UML KMeans Clustering represents a systematic approach that has a relatively high degree of reliability. It is guaranteed to converge. Even if the final answer is not the global maximum/minimum, it is at least a local maximum/minimum.

Pros #2 – UML KMeans Clustering is relatively simple and is equipped to handle large data sets. It can be particularly useful and efficient when you have very little or no knowledge about the problem being solved. In this case you can rely on its built-in robust mechanism and won't have to worry about all the unknowns and the unforeseeable, and the clustering results often offer good insights.

Pros #3 – UML KMeans Clustering is easy to adapt to new objectives. The approach and coding in this project can readily be reused for other venues. For example, if the business problem changed to where would be the best community to launch a hair salon in Chicago, then you can reuse everything exactly up to section 4 in the Jupyter Notebook. That is, about 75% of the project work is readily reusable.

§5.2.2 – Cons

Cons #1 – It is difficult to determine the optimal number of clusters. Sometimes there is no optimal at all. Although as determined by the silhouette score and elbow method, a 3-cluster clustering was used in the project, is it really the best? How about other number of clusters (particularly the second best, 5-cluster, according to silhouette score)? I did run a 5-cluster clustering for comparison. The final result is the same in terms of the best community selected. However, what if the results weren't the same? Which one is better? Would it also mean that the chosen number of clusters (i.e., 3 or 5) were only slightly better or not better at all than the unchosen ones (i.e., 4, 6, etc.)?

Cons #2 – Using 3-cluster KMeans I stepped into a situation where a large percentage (i.e., >80%) of the city communities ended up being clustered into one single cluster. This situation wouldn't get significantly better when the number of clusters increases. It is a known inherited issue for KMeans. Often time the KMeans model can only take you so far, there are still many data points in the chosen cluster. What are you going to do next? I developed a set of self-define, venue-specific criteria for selecting the best community. However, we should question ourselves that, if this set of criteria works, then why can't we use them to select the best community and bypass running the KMeans Clustering all together?

In this project, I chose to use the UML KMeans Clustering Methodology because its advantages seem to overweight its disadvantages in tackling this particular business problem.

§5.3 – Choice of Methodology

We must realize that in addition to unsupervised machine learning technologies, one can rely on other creative ways to address this business problem - some may be surprisingly simple and effective. For example, you can ask your friends who happen to be locals of Chicago for they likely know the communities first hand and have a more realistic view; you can patronize a few steakhouses in different neighborhoods in Chicago, which could help you develop a reasonable expectation and improve your judgement; or you can simply post this question on Quora.com, etc. and wait for others to answer - even if there wouldn't be a definitive answer, you will still likely to get some valuable information… In fact, a combination of these approaches could be a really effective way to uncover an answer.

§5.4 – When Not to Use KMeans Clustering Methodology

Let's not to forget that Python sklearn.cluster KMeans isn't the only UML clustering methodology we have in our toolbox.

In some cases, particularly when people have unusual expectations on the shape and size of the partitioned clusters, other methods such as R's

- pam (partitioning around medoids) function; and
- clara (clustering large applications in R) function - an extension of pam to handle large datasets

would work better.

In many other situations, exploring alternatives of KMeans is necessary. It's out of the scope of this project to fully discuss these alternative options. One can obtain relevant information and knowledge by searching them online.

Part VI – Conclusion and Future Directions

We came a long way to the end of this project. To quickly recap, we started with collecting data of Chicago communities, enriched data using Python GeoPy and Foursquare API, performed segmentation and clustering on Chicago communities using UML clustering methodology KMeans, further analyzed the results using a set of self-defined, venue-specific criteria.

At this point, we can answer the business problem raised in the beginning:

Question: where would be the best place to open a steakhouse in Chicago?

Answer: According to this analysis, Rogers Park Community is the best location for opening a steakhouse in Chicago.

Given a systematic approach coupled with careful evaluation and execution of one of the most popular and time tested UML clustering methodologies - KMeans, there should be sufficient credibility in the final results and conclusion. On the other hand, please also note that:

- This analysis is more focused on demonstrating a systematic approach of solving a business problem by selecting and utilizing the appropriate technology (i.e., tool). In reality, when determining where to open a steakhouse, people will have to take many other factors into consideration, such as local demographics, religions, traditions, regulations, and many more.

- In solving a business problem, all options, possibilities, and alternatives should not be easily dismissed only because one model is readily available (i.e., in this project the KMeans clustering methodology). We would better develop a holistic view, evaluate other options, and strive for the best solution possible.

Part VII – References

1. Chicago Community Data: https://en.wikipedia.org/wiki/Community_areas_in_Chicago

2. Foursquare API website: Foursquare

3. Project Jupyter Notebook: https://github.com/SKS9001/Coursera_Capstone/blob/main/Applied_Data_Science_Capstone_Project_Notebook--Kui.Shen.ipynb

4. Project Final Report (this report): https://github.com/SKS9001/Coursera_Capstone/blob/main/Applied%20Data%20Science%20Capstone%20Project%20-%20Final%20Report%20-%20Kui.Shen.pdf

5. Project Presentation: https://github.com/SKS9001/Coursera_Capstone/blob/main/Applied%20Data%20Science%20Capstone%20Project%20-%20Presentation%20-%20Kui.Shen.pdf