Analysis of Chicago city for selecting optimal location for a cocktail bar

1. Introduction

Chicago, officially the City of Chicago, is the most populous city in the U.S. state of Illinois, and the third-most-populous city in the United States. With an estimated population of 2.7 million, it is also the most populous city in the Midwestern United States. Chicago is an international hub for finance, culture, commerce, industry, education, technology, telecommunications, and transportation. Chicago has one of the world's largest and most diversified economies, with more than four million employees and generating an annual gross regional product (GRP) of over $609 billion. The city is an efficient economic powerhouse, home to more than 400 major corporate headquarters, including 36 in the Fortune 500. Among the most diverse economies in the nation, Chicago is a key player in every sector from risk management innovation to manufacturing to information technology to health services. It has attracted many different players into the market. It is a global hub of business and commerce. This also means that the market is highly competitive. As it is a highly developed city, the cost of doing business is also extremely high. Thus, any new business venture or expansion needs to be analyzed carefully. The insights derived from the analysis will give a good understanding of the business environment which would help in strategically targeting the market. This will reduce risk as well as reap generous returns on the investment.

2. Business problem

The city of Chicago is famous for its flashy restaurants, bars and nightclubs at flamboyant neighborhoods. Cocktails bars have always been a go-to spot for most of the residents after a hard week at work to blow off steam. Starting a cocktail bar presents a really promising business opportunity but you have to distinguish yourself from the competition to expect long-term success.

Opening a cocktail bar would require a large capital, however, strategically selecting the best location for the venture would definitely result in better returns. My client is willing to open the cocktail bar in those community areas of Chicago whose residents have a high standard of living, the crime rates are comparatively low and the competition is less. My client is the owner of several medium sized businesses all across the globe and has ample of financial resources. He has hired me to present a comprehensive analysis of the city of Chicago and present to him a list of three neighborhoods which have great potential for opening a cocktail bar.

3. Data requirement and description

For the analysis I am leveraging four separate datasets which would provide me information about the coordinates of community areas in Chicago, the socioeconomic factors of the neighborhoods, crime rates of each neighborhood and fetch nearby points of interests using foursquare API.

Dataset 1: I am scraping a list of all 236 community areas of Chicago from Wikipedia(https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago) and then using geocoder I stored the latitude and longitude coordinates of each area in a pandas data frame.

	Community area	Latitude	Longitude
0	Albany Park	41.968290	-87.723380
1	Riverdale	41.654410	-87.602250
2	Edgewater	41.980460	-87.668340
3	Archer Heights	41.811540	-87.725560
4	Armour Square	41.834580	-87.631890
5	Ashburn	41.747850	-87.709950
6	Ashburn	41.941674	-88.198809
7	Auburn Gresham	41.743190	-87.655040
8	Avalon Park	41.745070	-87.588160
9	Avondale	41.939250	-87.711250
10	Irving Park	41.948461	-87.723240

Dataset 2: The second dataset is the socioeconomic factors of Chicago data which is available on the Chicago city data portal. Since the dataset is extremely large(> 1.5gb), I am using a snippet of the dataset available at https://ibm.box.com/shared/static/05c3415cbfbtfnr2fx4atenb2sd361ze.csv. I am only using the hardship index as a parameter for my analysis.

	Community area	Latitude	Longitude	COMMUNITY_AREA_NUMBER	hardship index
0	Albany Park	41.96829	-87.72338	14.0	0.536082
4	Riverdale	41.65441	-87.60225	54.0	1.000000
8	Edgewater	41.98046	-87.66834	77.0	0.185567
13	Archer Heights	41.81154	-87.72556	57.0	0.680412
14	Armour Square	41.83458	-87.63189	34.0	0.835052

Dataset 3: For the crime data in Chicago, I am leveraging the Chicago city crime dataset on the portal. Again, since the dataset is extremely large, I will be using a snippet of the dataset available at https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv.

	Community area	Latitude	Longitude	COMMUNITY_AREA_NUMBER	hardship index	Total crimes
0	Albany Park	41.96829	-87.72338	14.0	0.536082	5
1	Riverdale	41.65441	-87.60225	54.0	1.000000	2
2	Edgewater	41.98046	-87.66834	77.0	0.185567	2
3	Ashburn	41.74785	-87.70995	70.0	0.371134	8
4	Auburn Gresham	41.74319	-87.65504	71.0	0.752577	14

Dataset 4: For the final dataset I am using the foursquare API to fetch the details of all the bars at a distance of 10 km around all community areas. The data fetched consists of coordinates and type of the venue.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Albany Park	41.96829	-87.72338	Luxe Hookah Lounge	41.968836	-87.727922	Hookah Bar
1	Albany Park	41.96829	-87.72338	Rotana Cafe	41.968806	-87.728004	Hookah Bar
2	Albany Park	41.96829	-87.72338	240 Lounge	41.968283	-87.727558	Dive Bar
3	Albany Park	41.96829	-87.72338	Alpha Delta Dyke	41.961101	-87.717582	Fraternity House
4	Albany Park	41.96829	-87.72338	Albany Place	41.964036	-87.727688	American Restaurant

4. Methodology

For this project, I am using the community area hardship index, total crimes in the area and a list of nearby venues as parameters to cluster the data. Once the data has been clustered I am analyzing each cluster and choosing the best three neighborhoods to open a cocktail bar for my client.

- **Data gathering:** First, I scraped the table containing a list of community areas in Chicago from the Wikipedia page(https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago). The table contained a "neighborhood" column which is redundant in this case so I dropped the column from the dataframe. Using geocoder I stored the latitude and longitude coordinates for each community area into a dataframe.

	Community area	Latitude	Longitude
0	Albany Park	41.968290	-87.723380
1	Riverdale	41.654410	-87.602250
2	Edgewater	41.980460	-87.668340
3	Archer Heights	41.811540	-87.725560
4	Armour Square	41.834580	-87.631890
5	Ashburn	41.747850	-87.709950
6	Ashburn	41.941674	-88.198809
7	Auburn Gresham	41.743190	-87.655040
8	Avalon Park	41.745070	-87.588160
9	Avondale	41.939250	-87.711250
10	Irving Park	41.948461	-87.723240

For hardship index of each area I used the socioeconomic factors dataset available on Chicago data portal. They dataset originally consisted of 7 columns but for this analysis I require only the "hardship index" column so I dropped the other 6 columns.

	COMMUNITY_AREA_NUMBER	COMMUNITY_AREA_NAME	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER_CAPITA_INCOME	HARDSHIP_INDEX
0	1.0	Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39.0
1	2.0	West Ridge	7.8	17.2	8.8	20.8	38.5	23040	46.0
2	3.0	Uptown	3.8	24.0	8.9	11.8	22.2	35787	20.0
3	4.0	Lincoln Square	3.4	10.9	8.2	13.4	25.5	37524	17.0
4	5.0	North Center	0.3	7.5	5.2	4.5	26.2	57123	6.0

I used the crime data index dataset available on the Chicago data portal to store the number of total crimes per community area. However, the crime dataset consisted of 22 columns but for this analysis we require only the count of total crimes in each community area. So, I grouped the dataset on community areas and stored the total count of each area.

	Community area	Latitude	Longitude	COMMUNITY_AREA_NUMBER	hardship index	Total crimes
0	Albany Park	41.96829	-87.72338	14.0	0.536082	5
1	Riverdale	41.65441	-87.60225	54.0	1.000000	2
2	Edgewater	41.98046	-87.66834	77.0	0.185567	2
3	Ashburn	41.74785	-87.70995	70.0	0.371134	8
4	Auburn Gresham	41.74319	-87.65504	71.0	0.752577	14

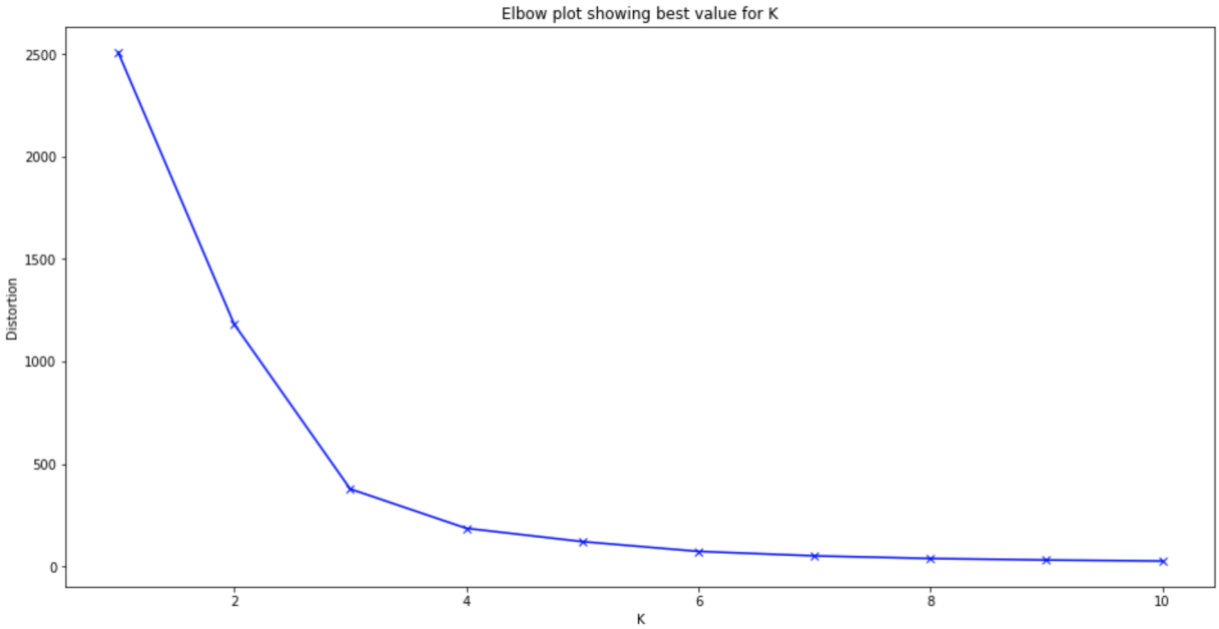
For the last dataset, I used the foursquare API to access all the bars around a 10km radius of each community area and store them. The access id for bars can be found at the foursquare website. I am taking the mean of total number of each type of venue and store them in a dataframe. Sorting the mean of each type on venue in descending order gives the list of most common venues.

	Community area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	hardship index	Total crimes
0	Albany Park	Hookah Bar	Dive Bar	Speakeasy	Karaoke Bar	Bar	0.536082	5
1	Ashburn	Sports Bar	Karaoke Bar	Liquor Store	Bar	Eastern European Restaurant	0.371134	8
2	Auburn Gresham	Bar	Wine Bar	Cocktail Bar	Lounge	Dive Bar	0.752577	14
3	Austin	Bar	Dive Bar	Pub	Gay Bar	Eastern European Restaurant	0.742268	43
4	Avalon Park	Wine Bar	Wings Joint	General Entertainment	Eastern European Restaurant	English Restaurant	0.412371	4

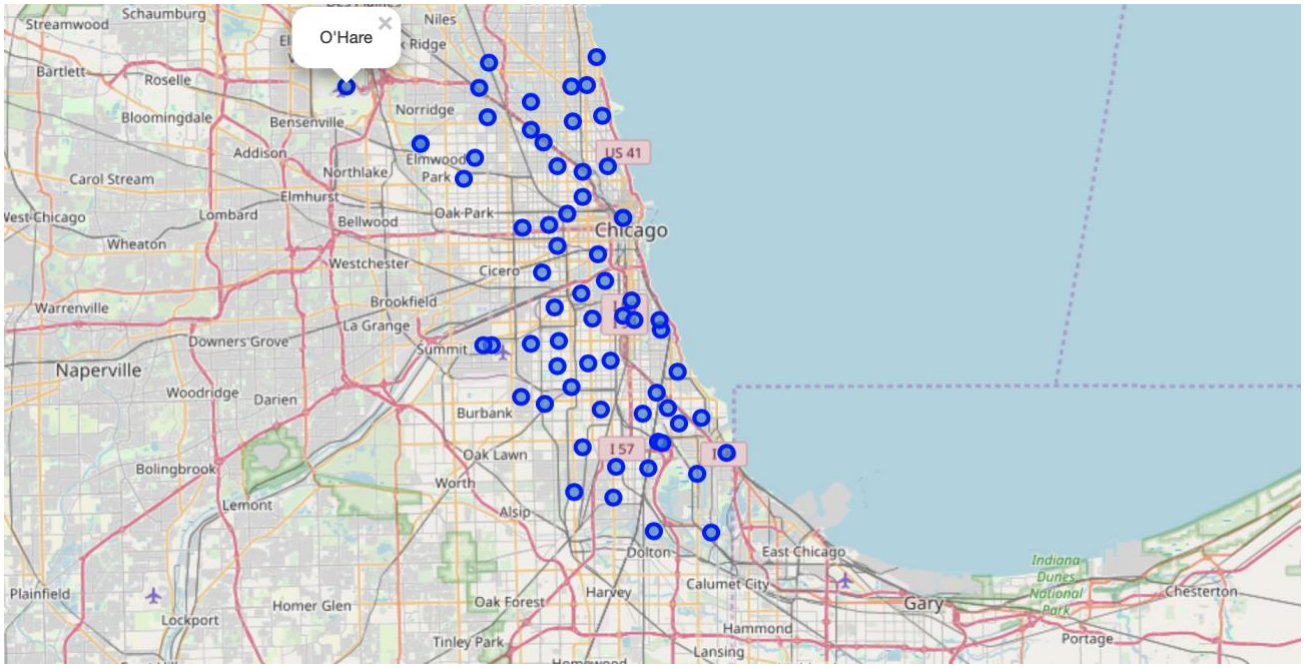
- **Data exploration:** Since I am using a clustering algorithm as my primary means for data analysis, I have to normalize the data so as to get the most accurate results. The hardship index provides a more complete, multidimensional measure of community socioeconomic conditions than individual measures such as income or employment alone. A community with a high hardship index score has worse social and/or economic conditions than a community with a low or medium hardship score. The hardship index score values were originally ranging from 0-100, so in order to get better results from the algorithm I normalized the values using max-min normalization.
  Once all the datasets were complete I merged them together to form a single dataset to carry out further analysis.

	Community area	Latitude	Longitude	COMMUNITY_AREA_NUMBER	hardship index	Total crimes	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Albany Park	41.96829	-87.72338	14.0	0.536082	5	2.0	Hookah Bar	Dive Bar	Speakeasy	Karaoke Bar	Bar
2	Edgewater	41.98046	-87.66834	77.0	0.185567	2	2.0	Bar	Gay Bar	Speakeasy	Wine Bar	Lounge
3	Ashburn	41.74785	-87.70995	70.0	0.371134	8	0.0	Sports Bar	Karaoke Bar	Liquor Store	Bar	Eastern European Restaurant
4	Auburn Gresham	41.74319	-87.65504	71.0	0.752577	14	3.0	Bar	Wine Bar	Cocktail Bar	Lounge	Dive Bar
5	Avalon Park	41.74507	-87.58816	45.0	0.412371	4	2.0	Wine Bar	Wings Joint	General Entertainment	Eastern European Restaurant	English Restaurant

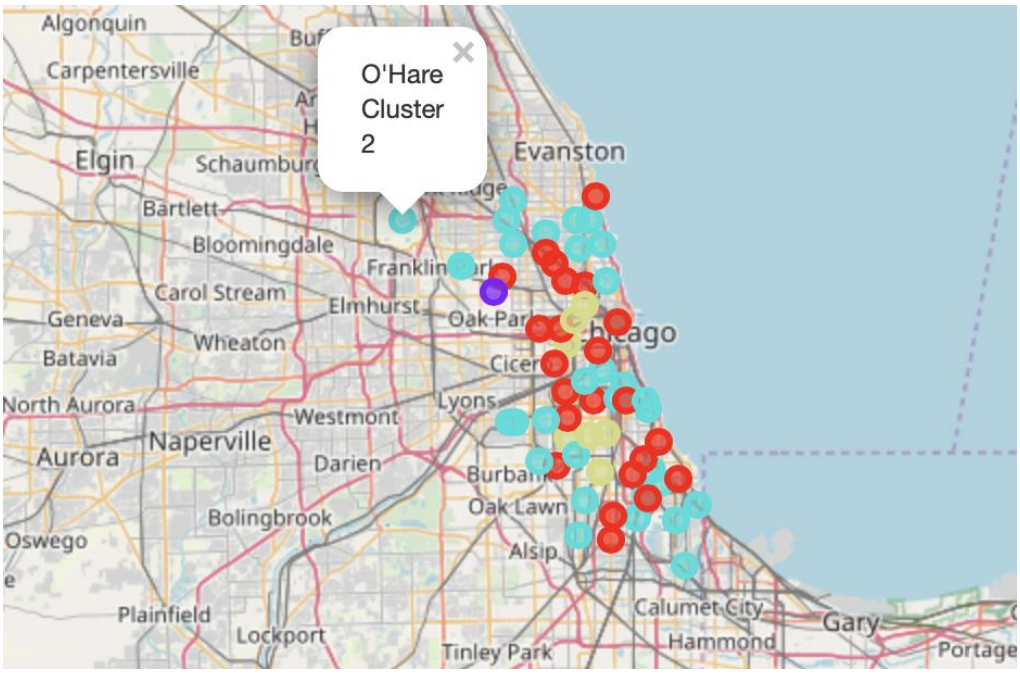
- **Data clustering and visualization:** For this project the clustering algorithm used is K-Means clustering. Once the dataset was ready after data wrangling, I used the elbow plot to determine the best value of K for the number of clusters. I had initially decided to group the data into four different clusters and the elbow plot shows that it is possible to group the data into four clusters as it has very little error value.



Before clustering the data I used to location data of community areas and plotted them on the map of Chicago using folium.



Then using the K-Means clustering algorithm, I cluster the data into 4 unique clusters. The clusters are then visualized using the folium library on the Chicago city map.



5. Results

Using the K-Means clustering algorithm with K=4, I got a total of 4 clusters.

Cluster 0

	Community area	hardship index	Total crimes	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	Ashburn	0.371134	8	0.0	Sports Bar	Karaoke Bar	Liquor Store	Bar	Eastern European Restaurant
6	Avondale	0.422680	6	0.0	Bar	Dive Bar	Pub	Tiki Bar	German Restaurant
7	Irving Park	0.340206	8	0.0	Bar	Karaoke Bar	Pub	Cocktail Bar	Speakeasy
8	New City	0.927835	10	0.0	Bar	Food Truck	Brewery	Wings Joint	General Entertainment
9	Belmont Cragin	0.711340	8	0.0	Bar	Cocktail Bar	Gay Bar	Wings Joint	General Entertainment
15	Lake View	0.041237	11	0.0	Bar	Cocktail Bar	Dive Bar	Speakeasy	Arcade
18	Brighton Park	0.855670	10	0.0	Bar	Dive Bar	General Entertainment	Eastern European Restaurant	English Restaurant
20	Logan Square	0.226804	9	0.0	Bar	Dive Bar	Cocktail Bar	Pub	Beer Bar
25	Chatham	0.608247	8	0.0	Bar	Cocktail Bar	Lounge	Wings Joint	General Entertainment
30	East Garfield Park	0.845361	8	0.0	Bar	Sports Bar	Wings Joint	Eastern European Restaurant	English Restaurant
32	Lower West Side	0.773196	9	0.0	Bar	Dive Bar	Cocktail Bar	Speakeasy	Beer Bar
37	Roseland	0.525773	11	0.0	Dive Bar	Speakeasy	Gay Bar	Eastern European Restaurant	English Restaurant
41	Gage Park	0.948454	7	0.0	Speakeasy	Dive Bar	Hotel Bar	Gay Bar	Eastern European Restaurant
45	Grand Boulevard	0.577320	8	0.0	Bar	Speakeasy	Karaoke Bar	Cocktail Bar	Caribbean Restaurant
46	Greater Grand Crossing	0.670103	11	0.0	Bar	Wine Bar	Speakeasy	Wings Joint	Gay Bar

Cluster 1

	Community area	hardship index	Total crimes	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
42	Austin	0.742268	43	1.0	Bar	Dive Bar	Pub	Gay Bar	Eastern European Restaurant

Cluster 2

	Community area	hardship index	Total crimes	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Albany Park	0.536082	5	2.0	Hookah Bar	Dive Bar	Speakeasy	Karaoke Bar	Bar
2	Edgewater	0.185567	2	2.0	Bar	Gay Bar	Speakeasy	Wine Bar	Lounge
5	Avalon Park	0.412371	4	2.0	Wine Bar	Wings Joint	General Entertainment	Eastern European Restaurant	English Restaurant
10	Hermosa	0.721649	3	2.0	Bar	Pub	Cocktail Bar	Gay Bar	Karaoke Bar
11	Dunning	0.278351	3	2.0	Bar	Pub	Cocktail Bar	Gay Bar	Karaoke Bar
12	Beverly	0.113402	4	2.0	Bar	Lounge	Beer Garden	Pub	Sports Bar
13	Morgan Park	0.298969	1	2.0	Bar	Karaoke Bar	Wine Bar	Cocktail Bar	Beer Garden
14	Norwood Park	0.206186	3	2.0	Bar	Cocktail Bar	New American Restaurant	Steakhouse	American Restaurant
16	Lincoln Square	0.164948	3	2.0	Bar	Karaoke Bar	Speakeasy	Dive Bar	Pub
17	Bridgeport	0.432990	1	2.0	Bar	Speakeasy	Wings Joint	Korean Restaurant	Sports Bar
19	Douglas	0.474227	5	2.0	Sports Bar	Cocktail Bar	Bar	Dive Bar	Wings Joint
21	Uptown	0.195876	4	2.0	Bar	Gay Bar	Speakeasy	Hookah Bar	American Restaurant
22	Burnside	0.804124	1	2.0	Wine Bar	Bar	Speakeasy	Wings Joint	Gay Bar
23	Calumet Heights	0.381443	5	2.0	Nightclub	Music Venue	Wings Joint	Eastern European Restaurant	English Restaurant
24	Near South Side	0.061856	1	2.0	Bar	Dive Bar	Eastern European Restaurant	General Entertainment	English Restaurant
27	Clearing	0.288660	2	2.0	Wine Bar	Bar	Wings Joint	General Entertainment	Eastern European Restaurant
28	Pullman	0.515464	3	2.0	Bar	Wings Joint	General Entertainment	Eastern European Restaurant	English Restaurant
31	Hyde Park	0.134021	2	2.0	Dive Bar	Gastropub	Wine Bar	Asian Restaurant	Hookah Bar
33	East Side	0.649485	2	2.0	Bar	Wine Bar	Wings Joint	General Entertainment	Eastern European Restaurant
35	Forest Glen	0.103093	1	2.0	Sports Bar	Beer Garden	Speakeasy	Wings Joint	Eastern European Restaurant
38	West Lawn	0.567010	3	2.0	Bar	Wings Joint	General Entertainment	Eastern European Restaurant	English Restaurant

Cluster 3

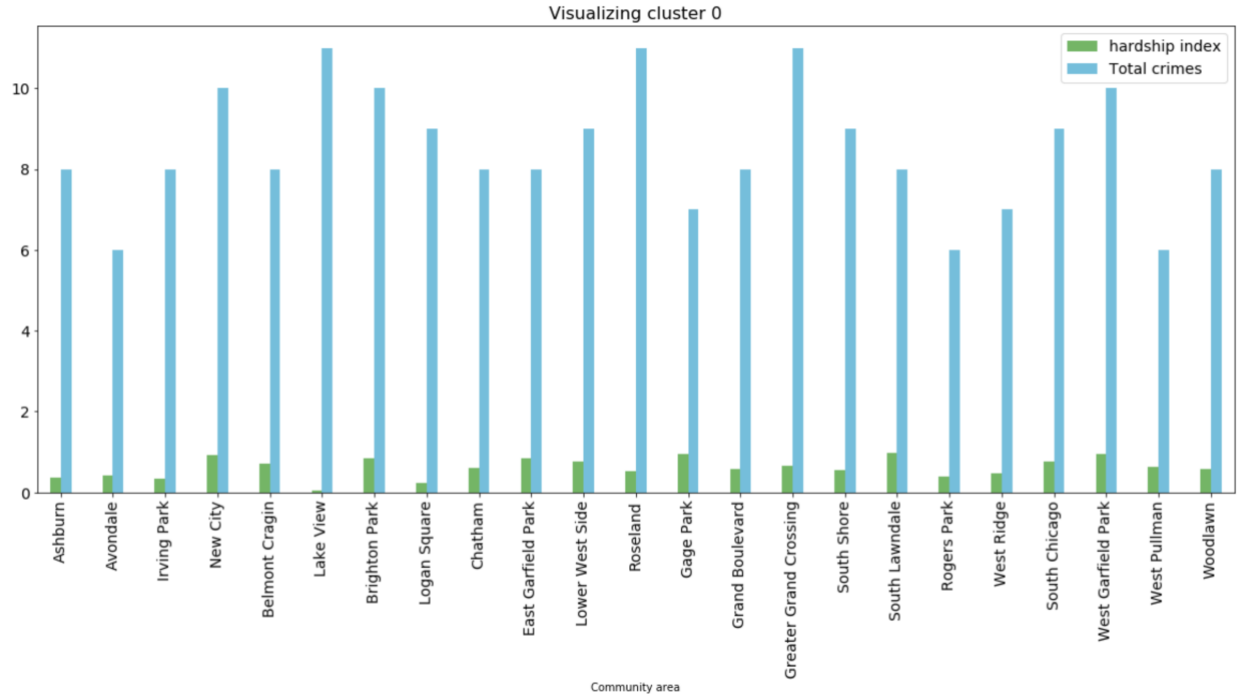
	Community area	hardship index	Total crimes	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Auburn Gresham	0.752577	14	3.0	Bar	Wine Bar	Cocktail Bar	Lounge	Dive Bar
26	Chicago Lawn	0.814433	12	3.0	Bar	Beer Garden	Dive Bar	Wings Joint	General Entertainment
29	North Lawndale	0.886598	16	3.0	Speakeasy	Wine Bar	Cocktail Bar	Bar	Brewery
34	West Town	0.092784	13	3.0	Bar	Dive Bar	Cocktail Bar	Speakeasy	Pub
36	Englewood	0.958763	21	3.0	Bar	Wine Bar	Wings Joint	General Entertainment	Eastern European Restaurant
40	Near West Side	0.144330	16	3.0	Bar	Speakeasy	Wine Bar	Dive Bar	Distillery
61	West Englewood	0.907216	12	3.0	Bar	Wine Bar	Wings Joint	General Entertainment	Eastern European Restaurant

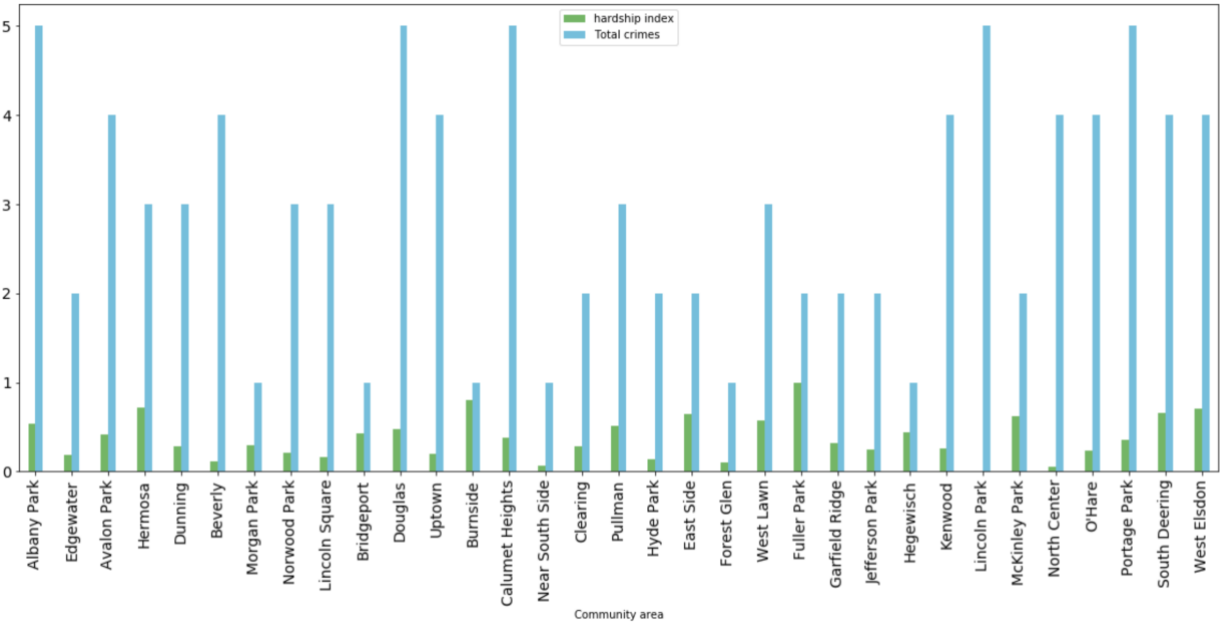
6. Discussion

In this section I will be discussing the observation I have noted and make recommendations based on the observations.
By analyzing the clusters, I observed that cluster 1 has an extremely high crime rate and hardship index. It contains only the neighborhood Austin so that is definitely not a good choice for opening a cocktail bar.

Cluster 3 has comparatively higher crime rates and it already has some cocktail bars as the 3rd common venue. So, this cluster as well is not a good fit my client's need.
The best location for opening a cocktail bar can be found in cluster 0 and cluster 2.

I used bargraphs to visualize the community areas in these clusters:





The location near south side (cluster 0) has extremely low crime rates and has a very low hardship index. The competition at this area is also less than other areas. Lake view and Lincoln Park are also potentially good areas due to their low crime rates and hardship index however the competition is high in these areas.
So, my recommendation for top 3 community areas for my client are:

1. Near South side (cluster 2)
2. Lake view (cluster 0)
3. Lincoln Park (cluster 2).

7. Conclusion

Business owners are always looking for potential locations around the world to expand their business. For this reason, the presence of a platform such as the one developed in this project presents the owner with an edge over their competitors.
Although the goals of this project were met but it can definitely be improved by adding more parameters in the analysis such as parking availability, traffic density, transportation facilities, real time data in and around the locations etc. This project can be augmented by creating a web or mobile phone app which can be accessed remotely. This analysis could be as simple as feeding the type of business, range of capital investment and location into the app and getting a detailed analysis with recommendations based on the data analysis carried out in the project.