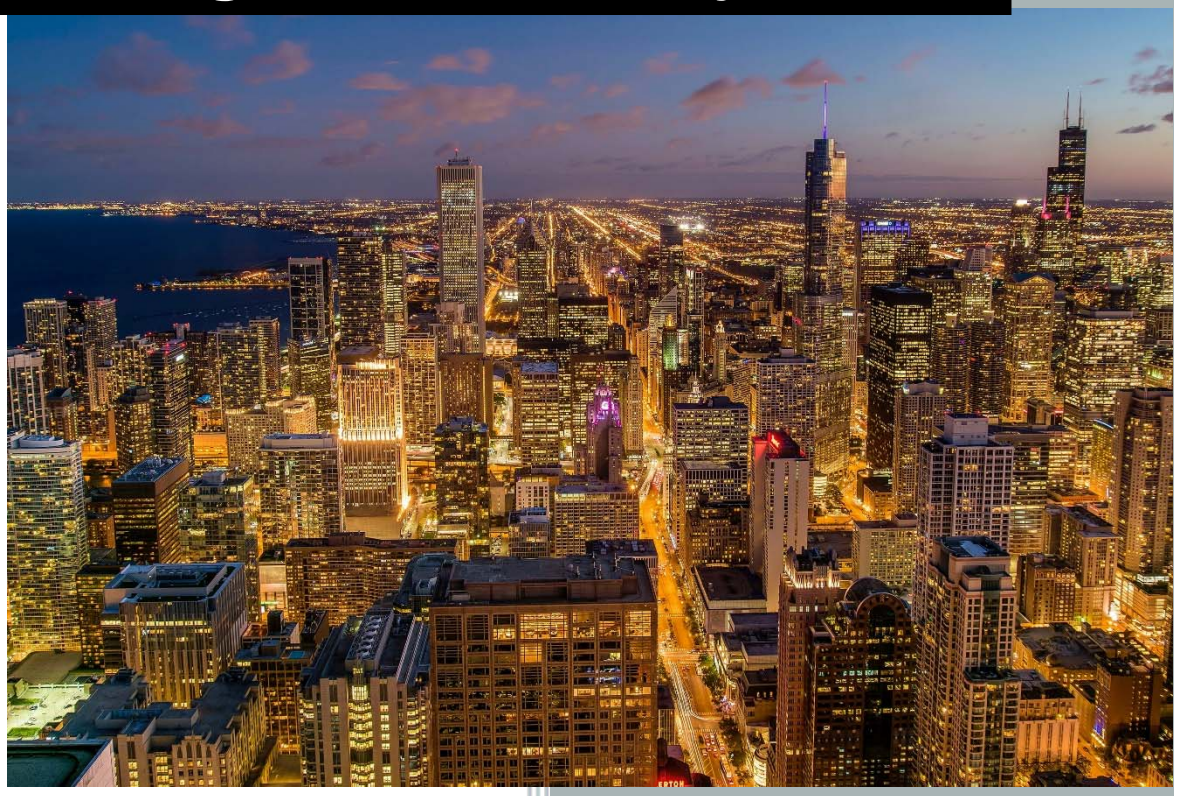
Clustering Chicago's Community Areas



1. Introduction

Chicago is often listed as the food capital of the US. A lot of Chicago based businesses are looking to expand across the city, which has led to the emergence of food based accelerators and related scale up advisory units. Scaling up, however, doesn't always lead to the same amount of success as the original business and one of the main reasons for that is a poor choice of location for the new venture.

This project aims to cluster the 77 different community areas of Chicago and to provide guidance for businesses, mainly restaurants/cafes and bars, looking to scale up across the city of Chicago. If a business is successful in a certain area of Chicago then they can use this clustering to pinpoint similar areas and to identify community areas in which they can project equal success. In addition to the clustering based on neighborhood venue similarity, an additional map is provided segmenting the different neighborhoods based on Population Density. Population Density can play a critical role in deciding where to set up a business as it provides an estimate of the reach of the business.

This clustering can also be used for new businesses that want to set up their first venue in Chicago and it is not just limited to scale ups. To use it in this way a business can look at a community area where their competitors are performing well and based on this mapping they can select a similar community area for the location of their new venue. By doing so they can pick an area where they can expect success while also avoiding getting into direct competition with their competitors.

This clustering alone will not guarantee the success of a scale up, as other factors such as cost also matter when choosing new locations, but it will provide a good indicator of the similarity of neighborhoods and the expected amount of success of launching in a given location.

2. Data

The data used for this project will be from the following sources:

- Web Scraping will be performed on the Wikipedia page about the "Community areas in Chicago", to retrieve the names of the Community areas and their respective population densities.
- Foursquare Location Data will be used to retrieve the top venues in each of the different community areas of Chicago.
- The Python Geocoding Library will be used to retrieve the coordinates of the different community areas in Chicago.

3. Methodology

The names of the different community areas in Chicago and their respective population sizes were scraped from the Wikipedia page about the Community areas in Chicago using the BeautifulSoup library in python. This data needed to then be cleaned and modified to later be used in creating the model. The respective latitudes and longitudes of the different Community Areas were then generated using the Geocoder library in python. The data was then combined to create the Dataframe shown in the image below.

	Community Name	Population Desnity(/sq mi)	Community Latitude	Community Longitude
0	Rogers Park	29,925.00	42.010531	-87.670748
1	West Ridge	21,590.65	42.003548	-87.696243
2	Uptown	24,988.36	41.966630	-87.655546
3	Lincoln Square	16,294.92	41.975990	-87.689616
4	North Center	17,458.05	41.956107	-87.679160
...	...	...	...	...
72	Washington Heights	9,598.95	41.706423	-87.656160
73	Mount Greenwood	7,113.28	41.698089	-87.708662
74	Morgan Park	6,786.06	41.690312	-87.666716
75	O'Hare	927.81	41.973101	-87.906768
76	Edgewater	32,163.79	41.983369	-87.663952

The Foursquare API was then used to retrieve the top performing venues in the different community areas. The API that was used was based on the names of the different community areas(the API deciphered the locations using Geocoders) and the limit of the venue search was set to 100. A snapshot of the Dataframe can be seen below.

	Community Name	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rogers Park	Rowlett Park	28.025015	-82.431707	Park
1	Rogers Park	Rogers Park Golf Course	28.019893	-82.426683	Golf Course
2	Rogers Park	Temple Crest Park	28.025126	-82.419121	Park
3	Rogers Park	RiverHills Dr & N 40th St Roundabout	28.021548	-82.414320	Intersection
4	Rogers Park	Rowlett Park (Dog park)	28.026378	-82.429982	Dog Run

The Get Dummies Pandas command was then used to break down each venue category into a subsequent column as is shown in the below snapshot

	Community Name	ATM	Accessories Store	Adult Boutique	African Restaurant	Airport	Airport Food Court	Airport Service	American Restaurant	Antique Shop	...	Water Park	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint	Women's Store	Yoga Studio	Zoo
0	Rogers Park	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	0
2	Rogers Park	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
4	Rogers Park	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

The data was then grouped by the Community Name column based on the mean value.

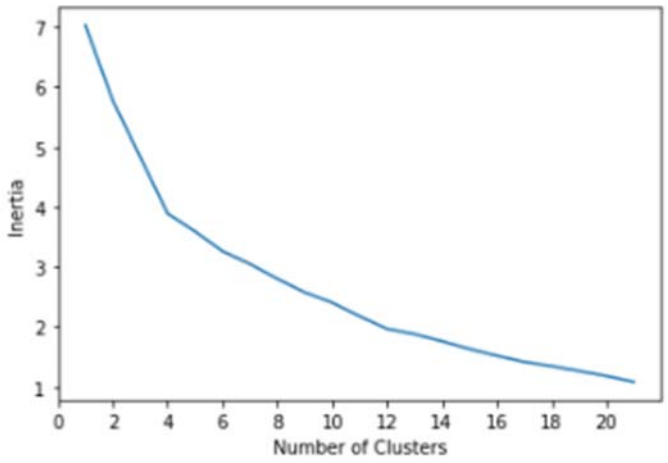
	Community Name	ATM	Accessories Store	Adult Boutique	African Restaurant	Airport	Airport Food Court	Airport Service	American Restaurant	Antique Shop	...	Water Park	Weight Loss Center	Whisky Bar	Wine Bar	Wine Shop	Winery	Wings Joint	Women's Store	Yoga Studio	Zoo
0	Albany Park	0.00	0.0	0.0	0.00	0.000000	0.0	0.00	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0
1	Archer Heights	0.00	0.0	0.0	0.00	0.000000	0.0	0.02	0.010000	0.0	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.010000	0.0	0.000000	0.0
2	Armour Square	0.00	0.0	0.0	0.01	0.000000	0.0	0.00	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.010000	0.0	0.000000	0.0
3	Ashburn	0.00	0.0	0.0	0.00	0.000000	0.0	0.00	0.050505	0.0	...	0.0	0.0	0.0	0.010101	0.010101	0.0	0.010101	0.0	0.000000	0.0
4	Auburn Gresham	0.00	0.0	0.0	0.00	0.000000	0.0	0.00	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0
5	Austin	0.00	0.0	0.0	0.00	0.000000	0.0	0.00	0.000000	0.0	...	0.0	0.0	0.0	0.000000	0.010000	0.0	0.000000	0.0	0.020000	0.0

The top 10 most common venues for each of the Community Areas was then found as is shown in the below snapshot.

	Community Name	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	Park	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1	Archer Heights	Mexican Restaurant	Rental Car Location	Donut Shop	Pizza Place	Bar	Sandwich Place	Taco Place	Seafood Restaurant	Gas Station	Fast Food Restaurant
2	Armour Square	Chinese Restaurant	Bar	Pizza Place	Mexican Restaurant	Hot Dog Joint	Bakery	Korean Restaurant	Historic Site	Coffee Shop	Asian Restaurant
3	Ashburn	Coffee Shop	American Restaurant	Fast Food Restaurant	Seafood Restaurant	Grocery Store	Brewery	Bar	Thai Restaurant	Sushi Restaurant	Supermarket
4	Auburn Gresham	Sandwich Place	Discount Store	Bar	Seafood Restaurant	Grocery Store	Lounge	Cosmetics Shop	Donut Shop	Fried Chicken Joint	Fast Food Restaurant
5	Austin	Taco Place	Coffee Shop	Pizza Place	Park	BBQ Joint	Trail	Grocery Store	Bar	Movie Theater	Hotel

The different community areas were then clustered using the K-means Clustering, an unsupervised Machine Learning algorithm, based on the similarities and differences of their venues. The number of clusters to be used must first be decided on and this was found using the Elbow Point method.

This method works by first computing the inertia, the within-cluster sum-of-squares, for a range of clusters, in this case it was from 1 to 21. The ultimate goal is to minimize this inertia, however, the inertia will always go down as the number of clusters increases and so the most efficient way is to choose an elbow point, a point where the absolute value of the slope of the chart shifts and is reduced significantly. The plot of the different number of clusters and the respective inertia is shown below. In this model, there seems to be a number of elbow points, such as at k=4 and at k=12. I decided to take the number of clusters to be 12 as the inertia at k=4 is quite high.



Using the Sci-Kit Learn library, the cluster Label for each of the different Community Areas was found. The data frames showing the most popular venue categories, the coordinates of the different community areas and their respective cluster labels were then combined as is shown in the below snapshot.

	Community Name	Population Desnity(/sq mi)	Community Latitude	Community Longitude	Cluster Label	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	Rogers Park	29,925.00	42.010531	-87.670748	7	Park	Yoga Studio	Intersection	Dog Run	Golf Course	N/A	N/A	N/A	N/A	N/A
1	West Ridge	21,590.65	42.003548	-87.696243	0	Pizza Place	College Cafeteria	Brewery	Liquor Store	Gas Station	Park	Grocery Store	Laundromat	Sandwich Place	Performing Arts Venue
2	Uptown	24,988.36	41.966630	-87.655546	0	Coffee Shop	Vietnamese Restaurant	Grocery Store	Breakfast Spot	Chinese Restaurant	Beach	Mexican Restaurant	Bakery	Thai Restaurant	Middle Eastern Restaurant
3	Lincoln Square	16,294.92	41.975990	-87.689616	9	Mexican Restaurant	Chinese Restaurant	Pizza Place	American Restaurant	Fast Food Restaurant	Discount Store	Pharmacy	Ice Cream Shop	BBQ Joint	Bar
4	North Center	17,458.05	41.956107	-87.679160	0	Bar	Coffee Shop	Pizza Place	Dive Bar	Pub	Italian Restaurant	Brewery	Café	Chinese Restaurant	Salon / Barbershop

Further segmentation was carried out based on the Population Density of each of the Community Areas. After cleaning the data, the different Community Areas were split into 5 Bins of varying Population Density ranges.

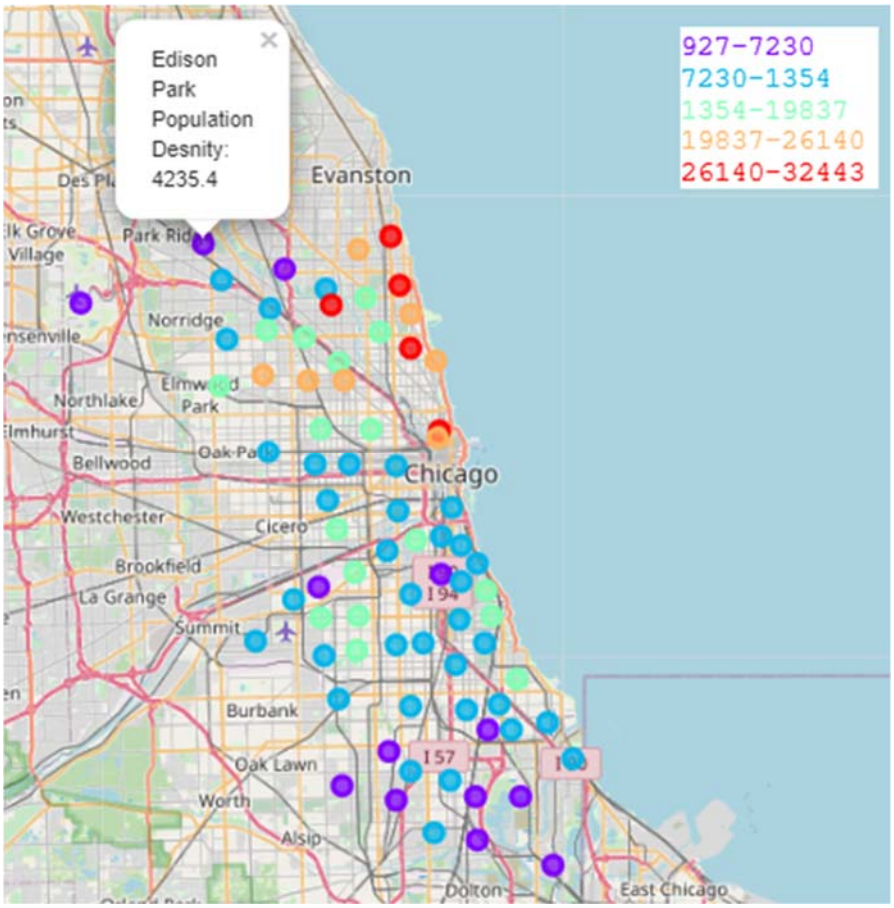
Using the Folium Library, 2 separate maps were generated from the data sets based on the coordinates that were retrieved using the Geocoder library. The first map highlights the different clusters that were generated from the K-means clustering algorithm based on the venue data. The second map highlights the different Population Density Bins of the different Community Areas. These maps will be shown in the Results Section below.

4. Results

The below Chart shows the Clustering of the different Chicago Community Areas based on the Venue data retrieved from the Foursquare API. The Color codes represent the different Clusters and by clicking on each circle, you are able to display the name of the area and its respective cluster number. The map shows that the Chicago community areas display significant differences, with the majority of the community areas falling into two separate clusters.

Clustering Based on Venue Type

The below chart shows the segmenting of the different Community Areas based on the Population Density. The Populations Densities are segmented into 5 Bin ranges and the color coding of the circles reflects the respective Bin Label. The legend displays the Population Density range for each bin (/sq mi). The population densities vary significantly between the different neighborhoods with the minimum value being 927 and the maximum value being 32443.



Segmenting Based on Population Density

5. Discussion

The above maps and findings add an additional tool in the toolkit of entrepreneurs looking to scale up or set up new businesses in Chicago. The maps show an amount of heterogeneity between the 77 Community Areas of Chicago. The results of the K-Means Clustering show that the best way to cluster the Chicago communities is into 12 Clusters, however, the majority of clusters fall into 2 main clusters.

Community areas that belong to the same cluster exhibit similarities between their venues and this should be an indicator of where to open new ventures based on past success. Community areas that have similar venues tend to attract similar customer groups and being able to understand what customer groups flock to certain areas should be a good predictor of success for a new venture.

Besides the venue types and in turn the expected customer types, population density can be a good indicator of the expected foot traffic at a given area. The second map, segmenting the

community areas based on population density, can be a good predictor of foot traffic. Chicago displays a wide variety of population densities across its different community areas, as is shown in the map.

The generated cluster maps can provide valuable indicators for new businesses and scale ups on deciding which location to open up their ventures, mainly based on the venues in a given community area and the types and amounts of customers in that area. This, however, should not be the sole tool that is used in deciding where to set up shop as other factors such as rent cost play an important role in the success of a business.

6. Conclusion

Chicago has long attracted new businesses, mainly restaurants, bars and cafes. This paper aims to be a guide for new businesses and scale ups in deciding which community area to open their new venture. A big percentage of businesses fail and often it’s a case of choosing the wrong location. This report should help guide businesses in deciding where to set up their venture but it should not be the sole tool that businesses use in making this decision. Understanding the expected customer type and the foot traffic in a given area is of extreme importance but businesses should not overlook other factors, such as rent and subsequent costs when making their decision.