Choosing best location for a new Coffee shop in Chicago

Juan J. Belaustegui

February 2020

1. Introduction

1.1 Background

Americans consume 400 million cups of **coffee** per day, making USA the leading consumer of **coffee** in the world. To have an idea, that's mean an average of 250 Cups of espresso and coffee drinks are sold per day at almost any Coffee shop with a great visible location.

If we have in mind that market equals to $12 billions in annual sales[1], seems very interesting to analyze if there is space to open a new Coffee Shop.

Chicago is the 3rd major city in US and the statistics says their citizens are above the average of US coffee consumption per capita.

1.2 Problem

We have a customer that is in the Coffee shop business and it is interested to open a Cafeteria in Chicago. They have requested a Data Scient to analyze public data available and provide a recommendation with best option where to implement a new Coffee shop.

2. Data acquisition and cleaning

2.1 Data sources

Getting some insights about coffee consumption, I found there are many cities that are plenty of coffee shops: Seattle, Manhattan, San Francisco and Pittsburg leading the ranking[1] in the US.

The next one in the ranking is Chicago, that shows it in the limit of the “National Saturation Line”.

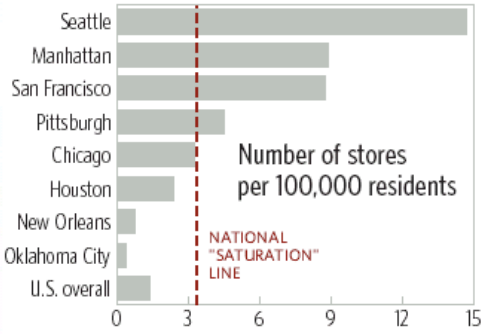

Fig.1 Coffee stores density US

This is an average, so probably we may find Areas with too many coffees and others with no one. So, it will be relevant to check every single Area and neighborhood from Chicago.

The data of the community areas and neighborhoods from Chicago has been taken from Wikipedia[2]

I will be using Geocoder library to get the coordinates from different Neighborhoods and using ***Foursquare*** API to get all the venues from each Neighborhood. Especial attention to all Coffee shops and Cafes that we can find, in order to exclude them and focus on Areas where is a lack of these stores.

Additionally, I’ll get the neighborhoods population, so we can have another information on what to make a decision. This information has been gathered it from last census data[3]

2.2 Data cleaning

The data of the neighborhoods was read as a table from the web. After some cleanup, we got a Pandas dataset like table 1.

	index	Area	Neighbourhood
0	0	Albany Park	Albany Park,Mayfair,North Mayfair,Ravenswood M...
1	1	Archer Heights	Archer Heights
2	2	Armour Square	Armour Square,Chinatown,Wentworth Gardens
3	3	Ashburn	Ashburn,Ashburn Estates,Beverly View,Crestline...
4	4	Auburn Gresham	Auburn Gresham,Gresham
5	5	Austin	Galewood,The Island,North Austin,South Austin
6	6	Austin, Humboldt Park	West Humboldt Park
7	7	Avalon Park	Avalon Park,Marynook,Stony Island Park
8	8	Avondale	Avondale,Jackowo,Wacławowo
9	9	Avondale, Irving Park	Polish Village
10	10	Belmont Cragin	Belmont Central,Brickyard,Cragin,Hanson Park

Table 1: Community Areas and Neighborhoods from Chicago

The information related to neighborhood population was read from a pdf document. I used ***Camelot*** module to read the file. After some parameter's adjustment to have a good presentation of the data, I deleted and disregarded some data not necessary for the analysis, like population in other years and percentage of growing. I got another Pandas dataframe like table2.

	Neighborhood	Population
0	Rogers Park	54,991
1	West Ridge	71,942
2	Uptown	56,362
3	Lincoln Square	39,493
4	North Center	31,867
5	Lake View	94,368
6	Lincoln Park	64,116
7	Near North Side	80,484
8	Edison Park	11,187
9	Norwood Park	37,023

Table 2: Neighborhood Population. Census 2010

After that I'll be using Geocoder to get coordinates and incorporate everything within the same dataframe and plot them in a map.

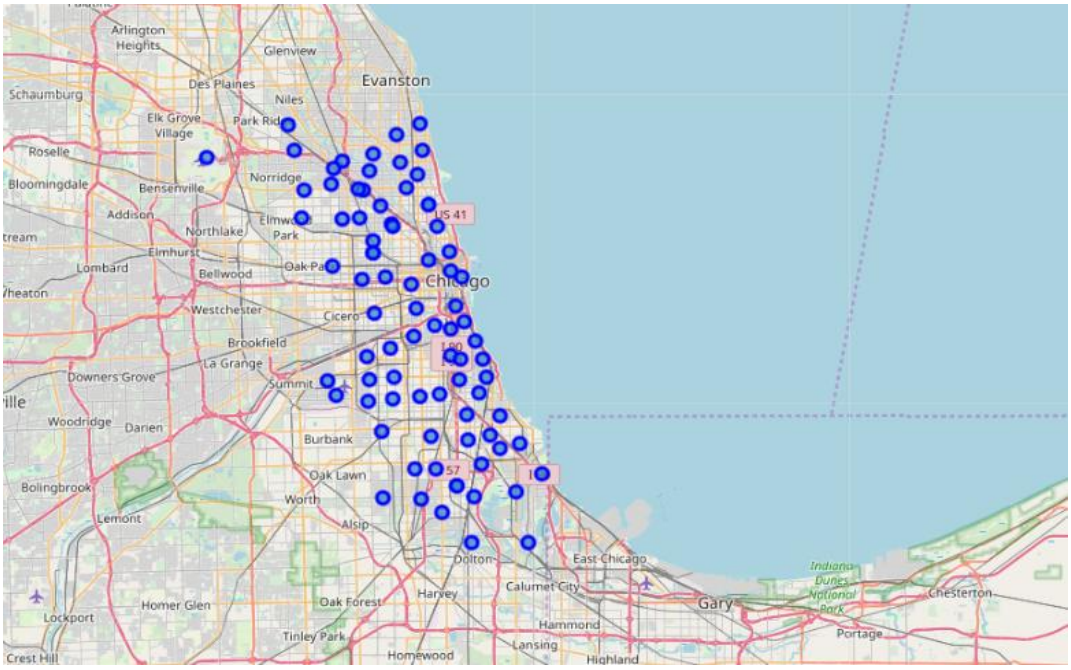

Fig 2: Neighborhoods from Chicago

2.3 Data selection

Once I have all the data into the same dataframe, I'll be using ***Foursquare*** API to explore the venues around the different areas. I'm going to exclude the neighborhoods where I find relevant presence of Coffee shops.

With a more concentrated data I'll be running a clustering with K means and will choose the one with presence of venues that can makes attractive for a coffee, taking in consideration of course, the amount of people that lives close in the Area, in order to increase the probability to reach a consumption of at least 250 cups of coffee per day.

3. Exploratory Data Analysis

3.1 Neighborhood population

The information extracted from a pdf file was read using ***Camelot*** , after get the pandas dataframe, I found some problems to convert *string* to *int* type, due to the value was presented with a *comma* as a decimal separator and it didn't work, so I had to delete it. After that I ordered the dataframe in descending order per population and presented for plot. I defined a subset with the Neighborhoods bigger than 50,000 people.

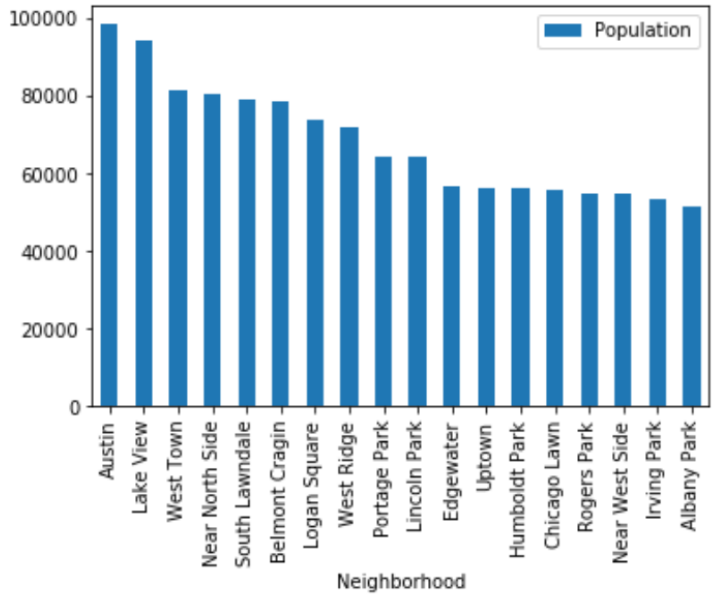

Fig 3: Population per Neighborhoods - Chicago

3.2 Neighborhood exploration

Having already the neighborhood coordinates with Geocoder, I started to explore the different venues using ***Foursquare*** API. I defined a limit of 100 venues in a radius of 1 Km from each point. A total of 4,282 venues were found with 343 unique categories. Going into detail for each neighborhood, it was found that 15 of them has at least 100 venues. For all the rest, ***Foursquare*** API gathered less venues for the radius defined.

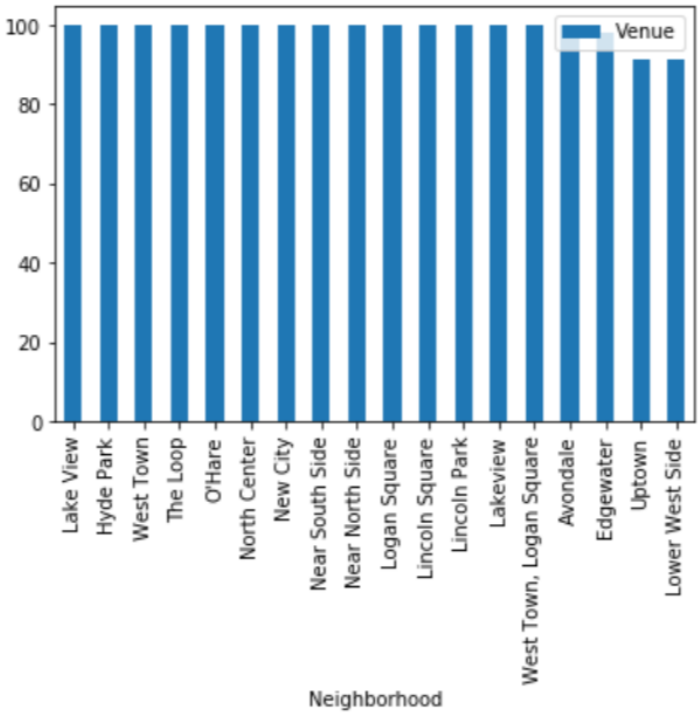

Fig 4: Venues detected per Neighborhood (amount >90)

A search of the neighborhoods with Coffee shops and Cafes was done. It was detected there are 31 Neighborhoods with No Coffee store within 1 Km area.

Plotting them in a map, as was expected, we see they are locations out from center of city.

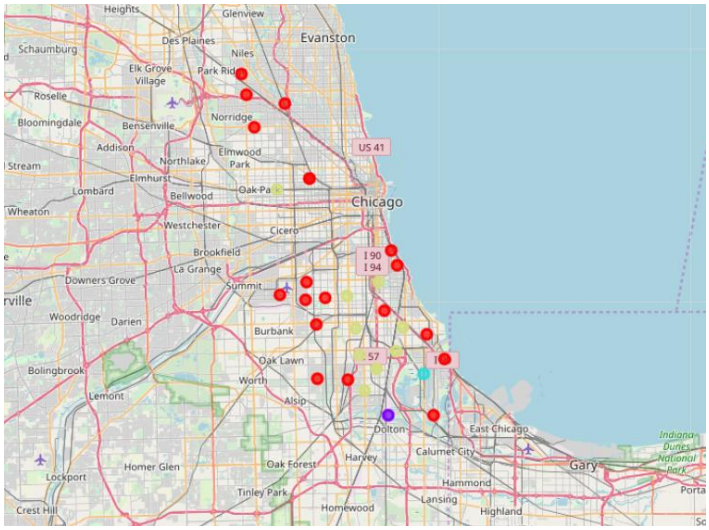

Fig 5: No Coffee Areas.

Now we run a K means to clustering the neighborhoods, this will allow to match different areas knowing how similar they are, in terms of the most common venues they have. But first we need to determine the optimal number of clusters (K). To do that, I check the Elbow Method and found K should be equals to 4.

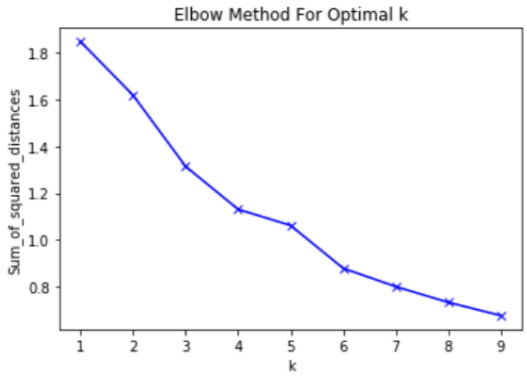

Fig 6: Elbow Method

Now we run K Means and the result is merged with population data.

Neighborhood	Neighbourhood	Latitude	Longitude	Cluster	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
Ashburn	Ashburn,Ashburn Estates,Beverly View,Crestline...	41.74785	-87.70995	0	Park	Pizza Place	American Restaurant	Liquor Store	Discount Store	Locksmith	Furniture / Home Store	Seafood Restaurant	Fried Chicken Joint	Train Station
Auburn Gresham	Auburn Gresham,Gresham	41.74319	-87.65504	3	Fast Food Restaurant	Lounge	Park	Currency Exchange	Discount Store	Pharmacy	Southern / Soul Food Restaurant	Boutique	Mexican Restaurant	Seafood Restaurant
Austin	Galewood,The Island,North Austin,South Austin	41.88774	-87.76392	3	Fried Chicken Joint	Food Court	Southern / Soul Food Restaurant	Gymnastics Gym	Train Station	Seafood Restaurant	Park	Discount Store	Grocery Store	BBQ Joint
Austin, Humboldt Park	West Humboldt Park	41.89907	-87.71947	0	Park	Donut Shop	Latin American Restaurant	Grocery Store	Mexican Restaurant	Music Venue	Liquor Store	Gas Station	BBQ Joint	Discount Store
Avalon Park	Avalon Park,Marynook,Stony Island Park	41.74507	-87.58816	3	Fast Food Restaurant	Sandwich Place	Pharmacy	Park	Grocery Store	Dance Studio	Bank	Cajun / Creole Restaurant	Supermarket	Automotive Shop

Table 3: Final Dataframe

4. Results

The final *dataframe.head* is showed in table 3.

Analyzing the different clusters, we can find the following:

Cluster 0: Many Parks and Mexican Restaurants

Cluster 1: Food and Park. But only one Neighborhood is in.

Cluster 2: Mexican Restaurants and Pier. Only 1 Neighborhood.

Cluster 3: Fast Food , Parks and Train Station.

Within the alternatives, Cluster 3 seems good option to install a new coffee, because having in mind, there is no coffee shop known in the area, sounds nice to be installed close to a Bus or Train Station, where people use to drink coffee waiting for the scheduled bus/Train, or waiting for a passenger. Also, there are Parks in the area, where people can drink a cup of coffee after some walk.

Within cluster 3, Neighborhood of Austin seems very attractive, due to it additionally has Gyms as the 4th most common venue. People usually takes coffee drinks after Gym or between classes. Also, Austin has a very high population (98,514), it doubles to next one, so it sounds like the best option.

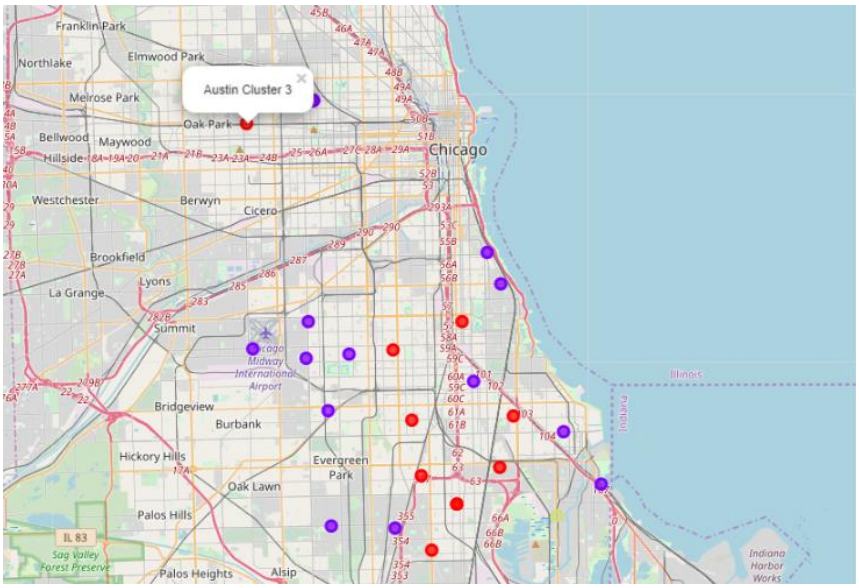

Fig 6: Austin Neighborhood

5. Discussion

The analysis has been done basically with the data gathered from ***Foursquare.*** For a more detailed information, the data should be crossed with a survey on the field, specially in Austin where the recommendation is provided.

I used the Kmeans algorithm as part of this study. When I tested the Elbow method, I set the optimum k value to 4. However, in the final Dataframe only 31 Area coordinates were used (The areas where no Coffee shop presence). For a more detailed and accurate guidance, the data set could be expanded to the level of neighborhood or streets.

6. Conclusion

In this study, I analyzed the best option where to open a new Coffee Shop in Chicago. I took the different neighborhoods from city, got their coordinates and explored the different venues recognized by the Foursquare within 1,000 meters. I selected the Areas where no coffee shop where identified and run a clustering with KMeans. Using the information of the clusters, plus the level of population of the different areas, I determined the best option to install a new Coffee shop is in neighborhood of Austin.

This study can be replicated to do the same with any kind of store in any place around the world. It could be a nice primarily tool to analyze alternatives, before going into a more detailed work or *in situ* surveys.

7. Bibliography

1. http://www.e-importz.com/coffee-statistics.php
2. https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago
3. www.chicago.gov/content/dam/city/depts/zlup/Zoning_Main_Page/Publications/Census_2010_Community_Area_Profiles/Census_2010_and_2000_CA_Populations.pdf