Finding the best neighbourhood where to open a restaurant in a foreign city

I. Business Problem

a. Introduction

Immigrating to a new country as a foreigner isn't easy, even more when thinking about opening a new business. The number of French restaurants in Chicago is, at first glance, relatively low. But even if Chicago might look like a good choice, Chicago is totally unknown to the future business owner. The aim of this study is to find the best suited neighbourhood for the opening of a French restaurant based on several criteria.

b. Objective

The goal of the study will be to determine which neighbourhoods of Chicago are suitable for the opening of a French restaurant. The neighbourhood will have to fit the following criteria:

- Affordable lease
- Few competitors
- Safe neighbourhood
- Wealthy neighbourhood (middle range)

II. Data acquisition

a. Community Areas of Chicago

The community areas (CA) and names of Chicago will be directly scrapped from the following Wikipedia page https://en.wikipedia.org/wiki/Community_areas_in_Chicago. Other useful information like the population and the area will be used later to create density based features.

Features sample:

	Community Area Number	Community Area Name	Population	Area (km2)
0	1	Rogers Park	55062	4.77
1	2	West Ridge	76215	9.14
2	3	Uptown	57973	6.01
3	4	Lincoln Square	41715	6.63
4	5	North Center	35789	5.31

Figure 1: Imported data for the size and population of the community areas of Chicago

b. Housing

The Chicago.gov website has been used to know the affordability of every neighbourhood. The data has been exported as CSV and will be first manually cleaned in excel (quicker); then imported into our Jupyter for further cleaning. The index named NOAH presented in the table is particularly interesting for our needs because it shows the 'Naturally Occurring Affordable Housing' in percentage.

Link:
https://www.chicago.gov/content/dam/city/depts/dcd/general/CITY_OF_CHICAGO_AFFORDABLE_HOUSING_DATA_TABLE.xlsx

Features sample:

	Community Area Name	% NOAH
0	ALBANY PARK	0.28
1	ARCHER HEIGHTS	0.57
2	ARMOUR SQUARE	0.40
3	ASHBURN	0.39
4	AUBURN GRESHAM	0.47
5	AUSTIN	0.39

Figure 2: Imported data for the affordability of every community area of Chicago

c. Competitors

In order to get an idea about the level of competition (or number of similar restaurants), the Foursquare API "search" query will be used. The number of French restaurants of every area will be calculated by summing up their number in every CA in a radius of 3000m.

Features sample:

	Community Area Name	French Restaurant
0	Albany Park	5
1	Auburn Gresham	1
2	Austin	2
3	Avalon Park	1
4	Avondale	5

Figure 3: Calculated data for the number of French restaurant around the centre of every community area of Chicago

d. Safety

The cityofchicago.org website has been used to calculate the level of safety of every neighbourhood. Only the annual reports from 2018 has been used in our case because the data was recent and complete. The data will be imported as XLSX into our Jupyter Notebook for further cleaning. Then the number of crimes will be counted for every area, without differentiation of the gravity of the crime. A new feature called crime ratio (number of crime divided by population) will also be created for every area.

Link:
https://data.cityofchicago.org/Public-Safety/Crimes-2018/

Features sample:

	Community Area Number	Number of crimes
0	1	3832
1	2	3536
2	3	3660
3	4	1940
4	5	1332

Figure 4: Calculated number of crime for every community area of Chicago

e. Wealth

The cityofchicago.org website has been used to get the per capita income for every neighbourhood. The data may be a bit old (2008-2012) but still good enough to have an idea about the wealth of every neighbourhood.

Link:
https://data.cityofchicago.org/Health-Human-Services/Per-Capita-Income/r6ad-wvtk

Features sample:

	Community Area Number	Community Area Name	Per Capita Income
0	1	Rogers Park	23939
1	2	West Ridge	23040
2	3	Uptown	35787
3	4	Lincoln Square	37524
4	5	North Center	57123

Figure 5: Imported data for the per capita income of every community area of Chicago

III. Methodology

After the import and cleaning of every CA feature, we merge the feature tables showed before into a single one. A new feature called "Crime Ratio (/p)" has been created in order to calculate the crime density of every area.

	Community Area Number	Community Area Name	Per Capita Income	Number of crimes	Population	Area (km2)	% NOAH	Crime Ratio (/p)
0	1	Rogers Park	23939	3832	55062	4.77	0.20	0.07
1	2	West Ridge	23040	3536	76215	9.14	0.19	0.05
2	3	Uptown	35787	3660	57973	6.01	0.19	0.06
3	4	Lincoln Square	37524	1940	41715	6.63	0.10	0.05
4	5	North Center	57123	1332	35789	5.31	0.06	0.04

Figure 6: Merged features table

The latitude and longitude will be retrieved thanks to the « Geopy » library.

	Community Area Number	Community Area Name	Per Capita Income	Number of crimes	Population	Area (km2)	% NOAH	Crime Ratio (/p)	Latitude	Longitude
0	1	Rogers Park	23939	3832	55062	4.77	0.20	0.07	42.010531	-87.670748
1	2	West Ridge	23040	3536	76215	9.14	0.19	0.05	42.003548	-87.696243
2	3	Uptown	35787	3660	57973	6.01	0.19	0.06	41.966630	-87.655546
3	4	Lincoln Square	37524	1940	41715	6.63	0.10	0.05	41.975990	-87.689616
4	5	North Center	57123	1332	35789	5.31	0.06	0.04	41.956107	-87.679160

Figure 7: Features table with latitudes and longitudes

The number of French restaurants in every CA radius of 3km is calculated based on the number of venues returned by the Foursquare Search query. This number will be then divided by the CA population and multiplied by 1000 to get a density of French venues per 1000 inhabitants.

	Community Area Number	Community Area Name	Per Capita Income	Number of crimes	Population	Area (km2)	% NOAH	Crime Ratio (/p)	Latitude	Longitude	French Restaurant	FR per pop
0	1	Rogers Park	23939	3832	55062	4.77	0.20	0.07	42.010531	-87.670748	4.0	0.072645
1	2	West Ridge	23040	3536	76215	9.14	0.19	0.05	42.003548	-87.696243	3.0	0.039362
2	3	Uptown	35787	3660	57973	6.01	0.19	0.06	41.966630	-87.655546	7.0	0.120746
3	4	Lincoln Square	37524	1940	41715	6.63	0.10	0.05	41.975990	-87.689616	5.0	0.119861
4	5	North Center	57123	1332	35789	5.31	0.06	0.04	41.956107	-87.679160	9.0	0.251474

Figure 8: Features table with density of French restaurant

We'll finally slice the table to get the necessary inputs for the K-Means clustering:

	Per Capita Income	% NOAH	Crime Ratio (/p)	FR per pop
0	23939	0.20	0.07	0.072645
1	23040	0.19	0.05	0.039362
2	35787	0.19	0.06	0.120746
3	37524	0.10	0.05	0.119861
4	57123	0.06	0.04	0.251474

Figure 9: Features passed to the K-Means algorithm

Given that every feature has another scale, we'll first normalize the features before using a clustering algorithm. A K-Means algorithm will be used in our case.

The plotting of the K-Means inertia shows us, thanks to the elbow method, that the optimal number of clusters lies around 4 or 5 clusters. After analyzing both cases, 5 clusters will be chosen.

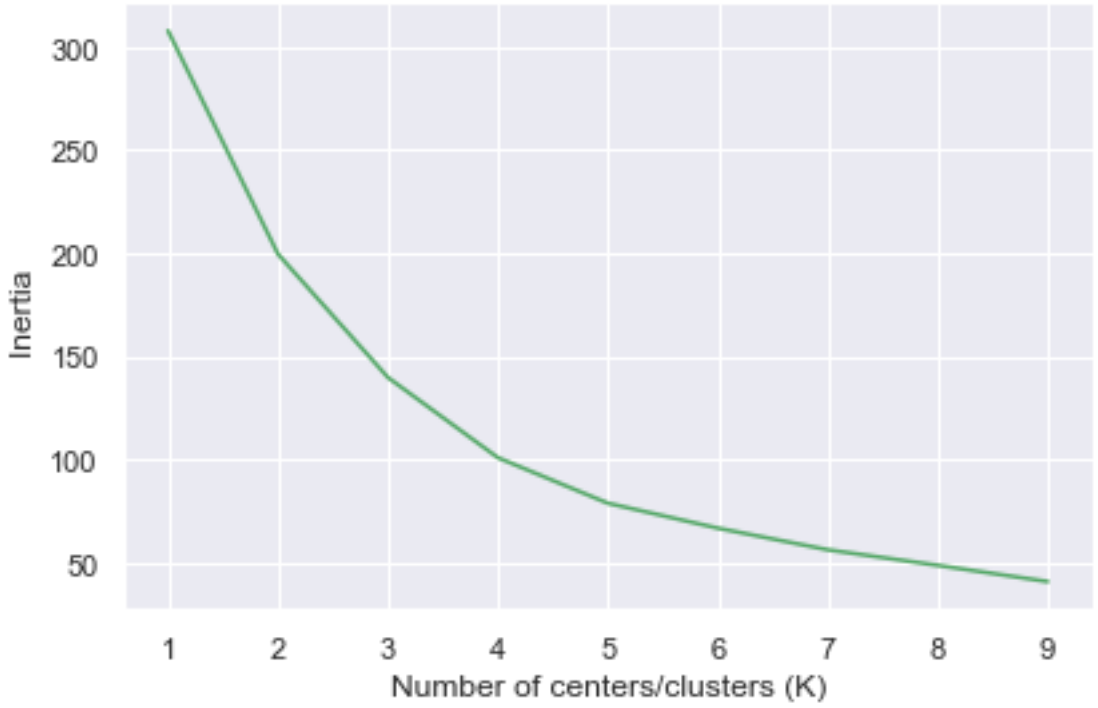

Figure 10: Plotting of the K-Means residuals as a function of the number of clusters

IV. Results

The five clusters are represented below on both maps. It can be noted that the clustering is also geographical; most of the CA are located in bigger cluster of the same label. Interesting is also the fact that Chicago's centre, the Loop, has its own cluster. These previous observations can be considered as meaningful and reassure us about the pertinence of the clustering.

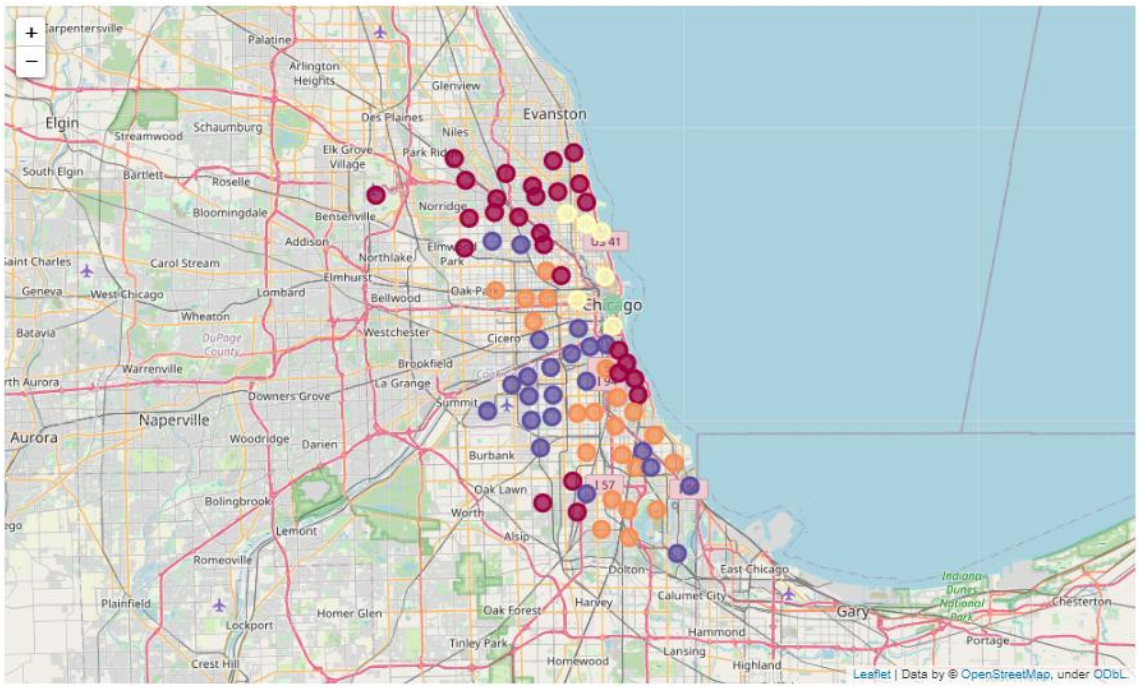

Figure 11: Chicago map showing the clustering of Chicago's community areas with coloured markers

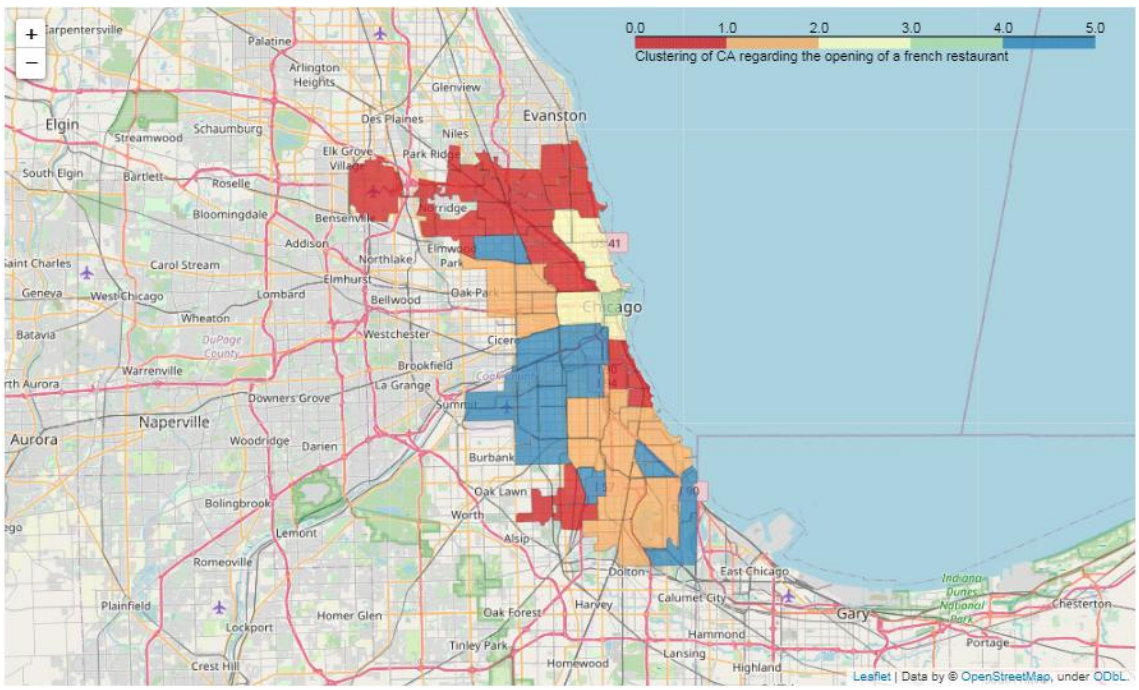

Figure 12: Chicago map showing the clustering of Chicago's community areas with filled areas

The median value of the normalized features has been plotted for every CA, see Figure 13. This helps us to characterize every cluster in the following manner:

- Cluster 0: Very safe middle class CAs with low competition and affordable homes
- Cluster 1 : Lower class CAs, rather unsafe but with little competition and affordable housing
- Cluster 2: Safe upper class CAs with higher competition and fewer affordable places
- Cluster 3: The CBD of Chicago, very high incomes but unsafe, unaffordable and with high competition
- Cluster 4: Safe lower class CAs with low competition and affordable housing

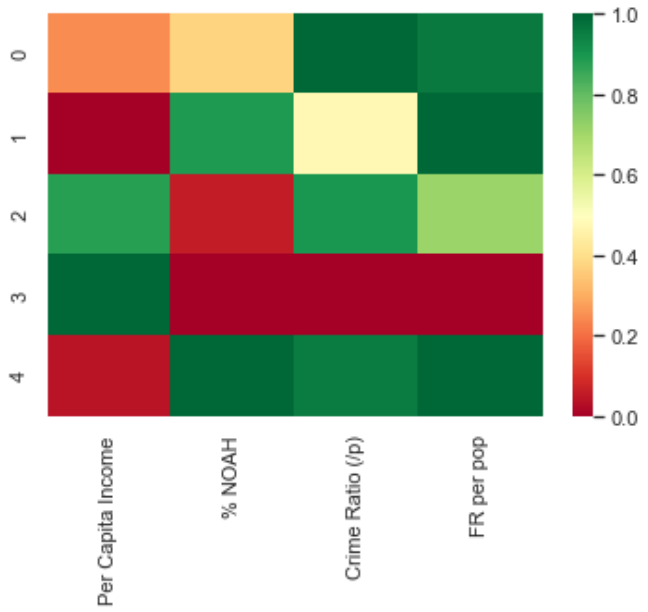

Figure 13 Heatmap comparing the normalized features of every cluster

The cluster 0 seems to be a good candidate for the safe and affordable middle class neighbourhoods we are looking for. We'll further analyze this cluster in the same way for knowing which CA could be the perfect candidate for the opening of a French restaurant. The Figure 14 shows again the normalized median value of every feature, this time comparing the CA together. After analysis it seems that Beverly or Forest Glen could be perfect places for the opening of the business.

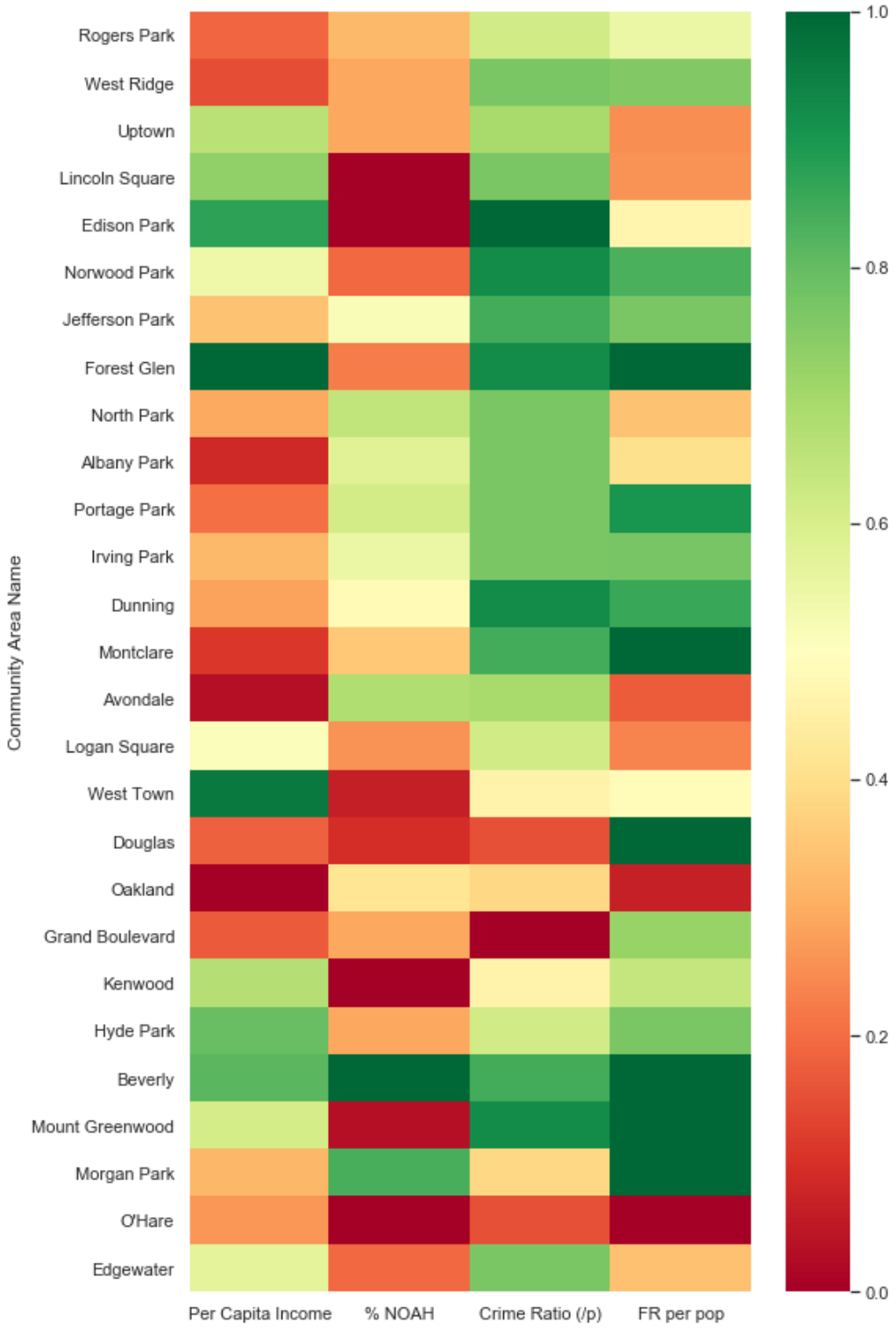

Figure 14 Heatmap comparing the normalized features of the Cluster 0 Community Areas

We'll create our own rating of the CAs of cluster 0 by adding up every normalized feature together and then plot them with a choropleth map (see Figure 15). The other clusters will be attributed the score of 0 because they do not have the features the business is targeting (safe and affordable middle class CA).

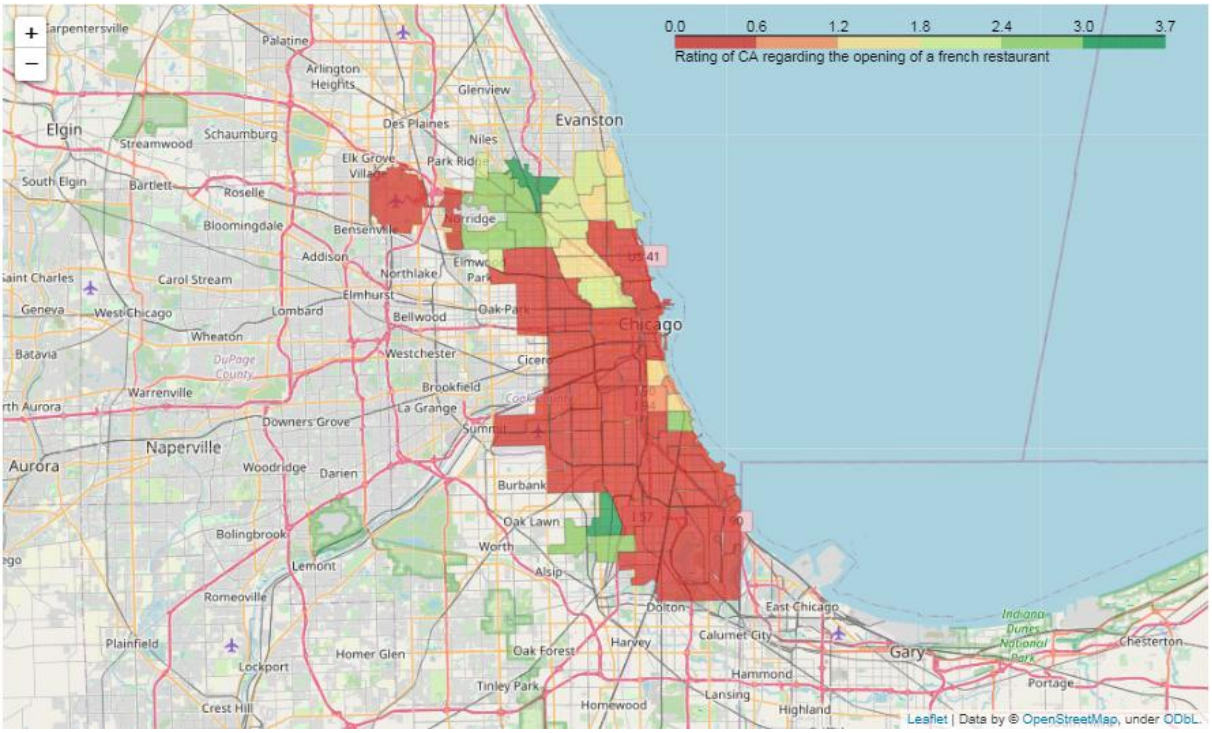

Figure 15: Chicago map showing our rating of Chicago's community areas regarding the opening of the business (high is good, low is bad)

V. Discussion

The results of the clustering process showed some meaningful results. The clustering is based on four features so far, but could also be extended to other features in the future like the average rating of the restaurants nearby or the price category. We can however also notice that there are some outliers in the clusters (ex: O'Hare and Oakland in Cluster 0), meaning that the clustering process could be improved.

VI. Conclusion

Finally, our clustering analysis showed that Beverly and Forest Glen are both very suitable community areas for the opening of a French restaurant. They are both very safe and affordable middle class CA with low competition and above average incomes. The clustering process could give an answer to our initial question by only inputting four features easily available on the internet and on the Foursquare API. The next step could be to refine our analysis by further exploring the selected CAs with other features like the average rating and price category of the restaurants.