Site selection for a new Peruvian restaurant

1. Introduction

Investing in a new business opens a series of questions, among them: Where should I locate? Answering this question is critical for the success of any business, and specifically for any entrepreneur interested in opening a new chain of restaurants. When thinking in opening restaurants, entrepreneurs must manage different factors or variables related to social, economic, cultural, business, and logistics issues that will be key for the investment. In that sense, this project aims to identify potential areas to open a new Peruvian restaurant in Chicago city, United States through a single score based on the factors mentioned. The study area of this project includes the Central Business District (CBD) and a 5 km. trade area to the CBD. The CBD characterizes for concentrating most of the commercial and business venues of a city. While a 5 km. trade area to the CBD was included to increase the number of units of analysis in this project and for its proximity to the CBD that might become a new center of attention for entrepreneurs in the future. This project could be of interest to Peruvian entrepreneurs who are looking for suggestions of locations to open a new restaurant in a multicultural city as Chicago based on a single site selection criterion that will help them to make better decisions.

2. Data acquisition and cleaning

2.1 Data sources

To identify potential areas for a new Peruvian restaurant, this project used tabular and geospatial data from different sources. First, tabular data was downloaded from the new website to explore census data at census tract level, our unit of analysis, within Cook county in the state of Illinois. These tables were: 2018 estimated population, 2018 estimated median income household of the last 12 months, 2018 estimated unemployment rate and 2018 estimated Latino population. These variables would be useful for exploring and analyzing them later. The reason of choice of these variables can be seen in table 1.

Table 1. Tabular data

Data	Description	Source
2018 estimated population	Total population in each census tract. This variable was selected to build other variables such as: density population, Latino population rate and crime rate which will be used for modeling.	https://data.census.gov/cedsci/advanced
2018 estimated household median income of the last 12 months	This variable was selected due to its importance to distinguish areas with high income and low income. Entrepreneurs will always be interested in opening a new business where households have high income	
2018 estimated unemployment rate	Entrepreneurs look for areas with low unemployment rates. This is an indicator of stability and a high likelihood that employed people might visit its restaurant	
2018 estimated Latino population	Although it is not a key variable as the 3 above, a restaurant located in a Latino area could accept the Peruvian cuisine faster in comparison to other customs from Europe, Asian, and so on.	

On the other hand, geospatial data as census tracts boundaries were downloaded from The United States Census Bureau. Furthermore, centers of population by census tract were identified from the same website to extract its coordinates. The advantage of using this kind of centers is that population live in those centers and they are not the common geographical centroid of census tracts where might not have people living. Another spatial data was CBD and obtained from Chicago Data Portal in order to select census tracts as our study area. Finally, crime locations represented as point data were extracted from the same website. More details about the use of each data in table 2.

Table 2. Geospatial data

Data	Description	Source
Census tracts	Census tracts boundaries in the state of Illinois. This data was downloaded as shapefile and then exported as 'json' file for mapping purposes	https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
Centers of Population by Census Tract	Real centroids with geographical coordinates where people live in each census tract in the state of Illinois. It provides a link to be read in a Python environment. The coordinates were used to identify food venues around each coordinate using Foursquare API	https://www.census.gov/geographies/reference-files/time-series/geo/centers-population.html
Central Business District	Central Business District boundary was helpful to identify census tracts within it and those census tracts located to 5 km. from the CBD. All these census tracts would form our study area	https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Central-Business-District/tksj-nvsw
Crime	Crime data that contains the following categories: assault, battery, burglary, robbery and theft which took place within and outside of restaurants. This data was processed to build a crime rate by census tract. The higher the rate, the lower the chance of opening a restaurant	https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Map/c4ep-ee5m

2.2 Data cleaning

The data pre-processing and cleaning was made in a Jupyter notebook. First, census data tables were read and only extracted the variables of interest: total population, income, unemployment and Latino population by census tracts. Additionally, a table containing the areas (in hectares) was loaded in the notebook which had been retrieved from the shapefile of census tracts in a GIS software. These tables shared a common key variable which was the unique code of each census tract.

Then, crime data had been downloaded as a CSV file which was geocoded in a GIS environment. Here, crimes that took place inside or outside of a restaurant and in the following most common categories: assault, battery, burglary, robbery and theft, were filtered and imputed the key variable of census tracts to be exported as CSV file and then read in the notebook and aggregated to that key variable to retrieve the number of crimes by census tract.

At that time, there were 6 tables which were merged and cleaned to build 3 new variables: density population, Latino population rate and crime rate (inflated by 1,000 habitants to avoid short numbers). Moreover, income was represented in thousands of dollars for mapping visualization purposes later. These 4 variables plus unemployment were the main variables for modeling and represented the social, economic and cultural aspects that an entrepreneur might consider before opening a Peruvian restaurant. A new data frame was created including the main variables.

Other variables were needed to measure the business dimension and identify competitors in census tracts. To accomplish with this, Foursquare API was used to find food venues by census tract. However, what the API requires are coordinates of census tracts which were read from the web (Centers of Population by Census Tract) and cleaned to have the correct key variable. Then, these coordinates were merged to the data frame with the 5 main variables.

After that, it was possible to use the Foursquare API to find food venues (a limit of 100 venues) within a radius of 300 metres of each census tracts. The choice of the radius is due to census tracts are small and a bigger radius might have resulted in duplicated information. According to the Foursquare website the code for food venues is '4d4b7105d754a06374d81259'. Having obtained all food venues, only restaurants were filtered and aggregated by census tracts to match to the main data frame to get the number of restaurants by census tract. Thus, a new and main variable was built for modeling.

As a final step in data pre-processing, 2 new main variables were built: the first most common restaurant and the second most common restaurant by census tract. They were built implementing code and calling the Foursquare API. The idea was that if one of the top 2 most common restaurants was a Latin American restaurant in a census tract, then a low score would be assigned to the census tract given that an entrepreneur might be more willing to invest money in areas without the presence of Latin American restaurants. In that way, the entrepreneur would compete with other kind of cuisine such as Chinese, Indian, European and others, offering more variety to the consumer.

Finally, these 2 variables were merged to the main data frame and was ready for exploratory data analysis.

2.3 Feature selection

After data pre-processing and cleaning, there were 206 census tracts with the following 8 main variables and 2 variables used in the previous section to call Foursquare API (latitude and longitude):

- Household median income (thousands of dollars)
- Unemployment rate
- Density population (persons/Ha)
- Latino population rate
- Crime rate (crimes by 1,000 people)
- Latitude
- Longitude
- Number of restaurants
- The first most common restaurant
- The second most common restaurant

3. Methodology

3.1 Exploratory Data Analysis

The first step in this section was exploring the main variables of analysis to get a first look about where an entrepreneur might consider opening a Peruvian restaurant. Therefore, the strategy employed here was map visualization made in Jupyter notebook. Figure 1 shows a map of household income. The assumption is the following: the entrepreneur should consider opening a new Peruvian restaurant in those census tracts with high income. This would guarantee a high chance of visiting its restaurant and consuming Peruvian cuisine. Census tracts in the center of CBD and north of the study area have high income (> 76 thousands of dollars).

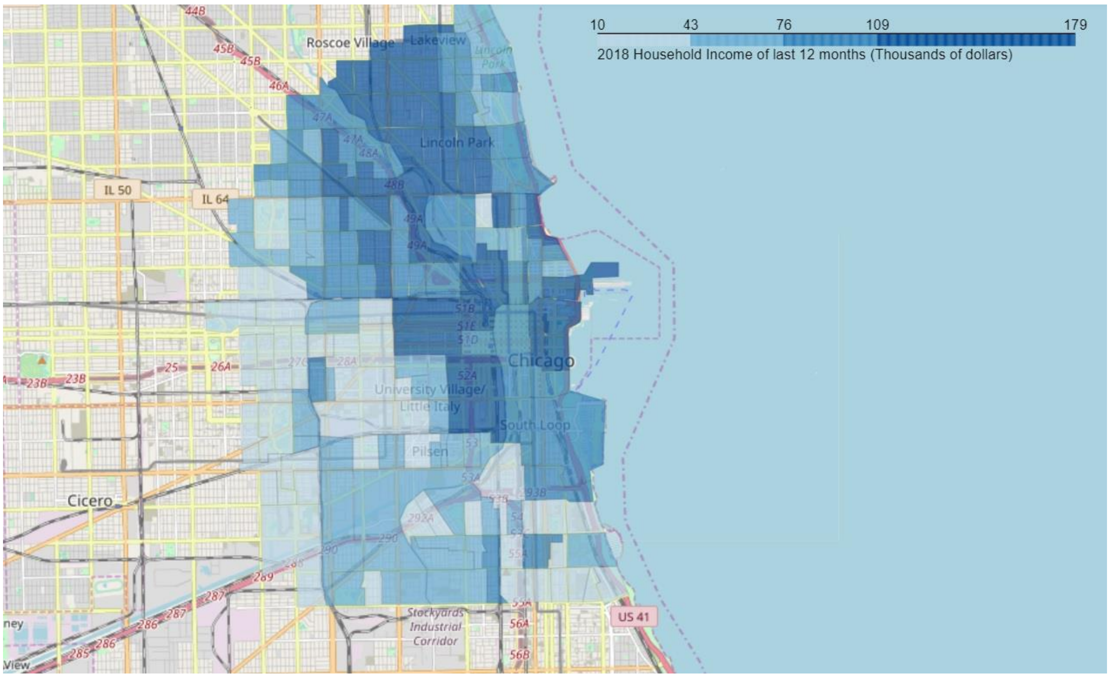

Figure 1. Household income

The second variable was unemployment. This is an important variable since the entrepreneur looks for areas where people are working and have a stable employment. So, they want to avoid areas with high unemployment rates as they can be mainly seen in the west and south (Figure 2).

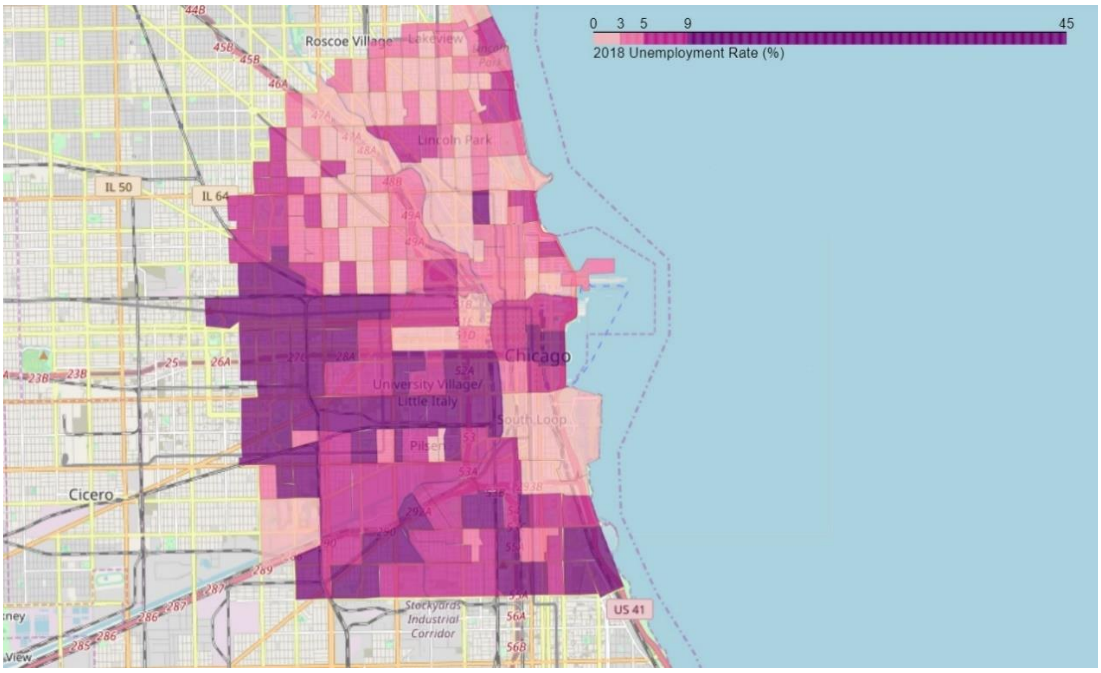

Figure 2. Unemployment rate

The third variable was density population. Populated areas might guarantee a high chance of visiting a commercial store as restaurants. Thus, entrepreneurs are always looking for populated areas which were identified in the north and center (Figure 3).

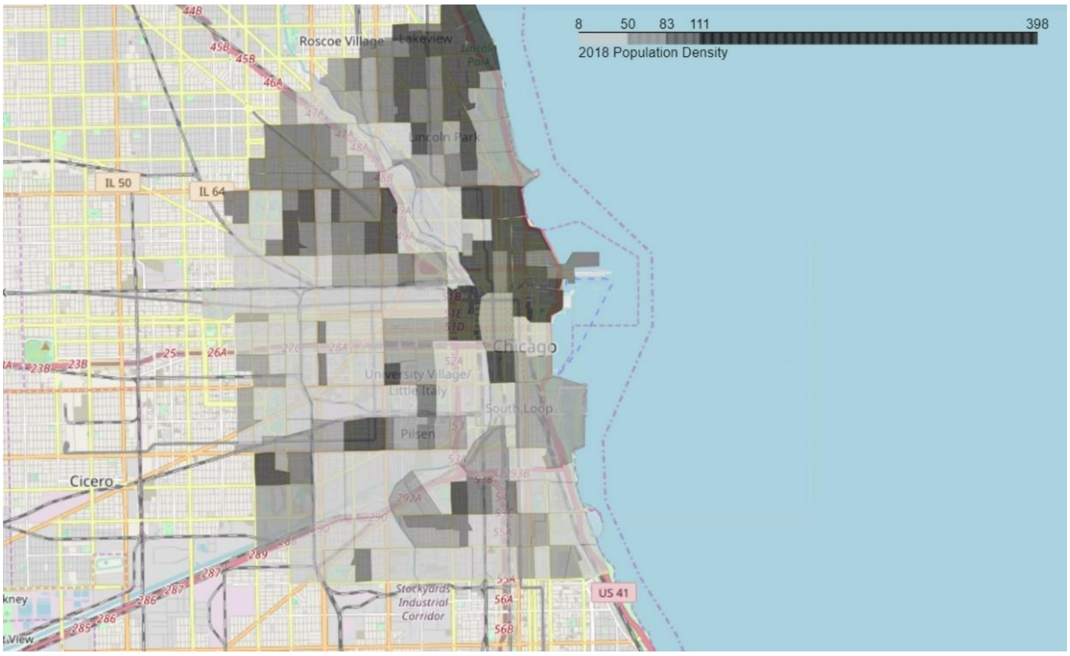

Figure 3. Density population

Figure 4 illustrated the Latino population rate. Here the northwest and southwest have a high proportion of Latinos. The assumption was that a restaurant located in a populated Latino area might accept the Peruvian cuisine faster in comparison to other people with their own customs from Europe and Asian as examples.

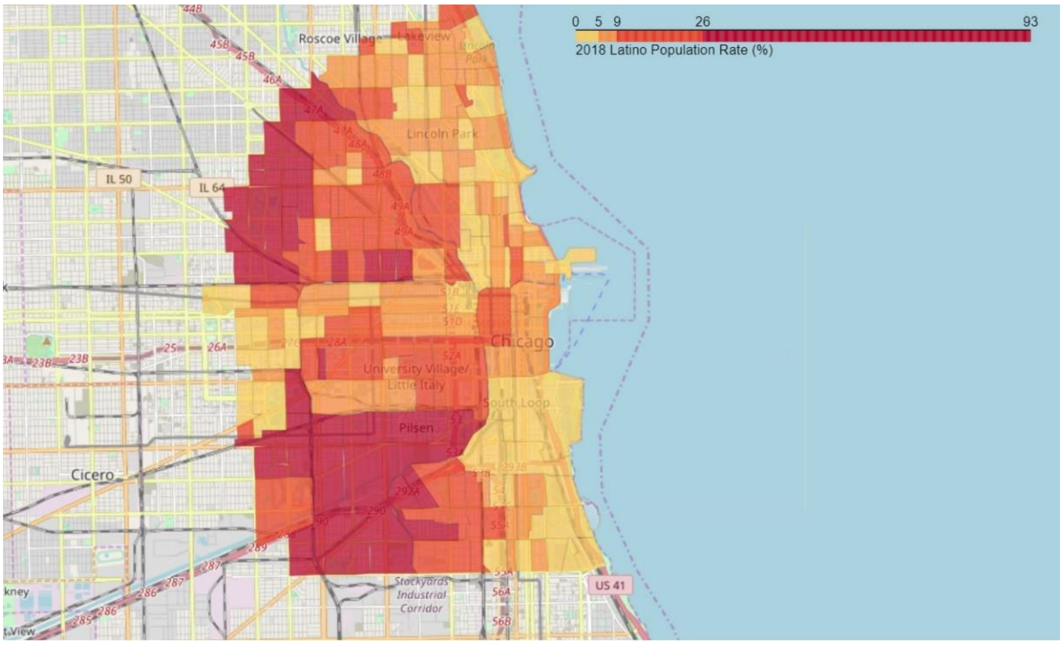
0 5 9 26 93
2018 Latino Population Rate (%)
Roscoe Village
Lakeview
Lincoln Park
IL 50
IL 64
Chicago
University Village/ Little Italy
South Loop
Pilsen
Cicero
Stockyards Industrial Corridor
US 41

Figure 4. Latino population rate

On the other hand, crime is another critical variable. Since the Chicago Data Portal provides crimes of different types and where they happened, it was important to identify the most common crimes as assaults, battery, burglary, robbery and thefts in restaurants. So, it tells the entrepreneur to think 2 times if it would be a good idea to invest in an area with high crime rates in restaurants. Clearly from the map, the center of the study area is the most affected by those kinds of crime (Figure 5).

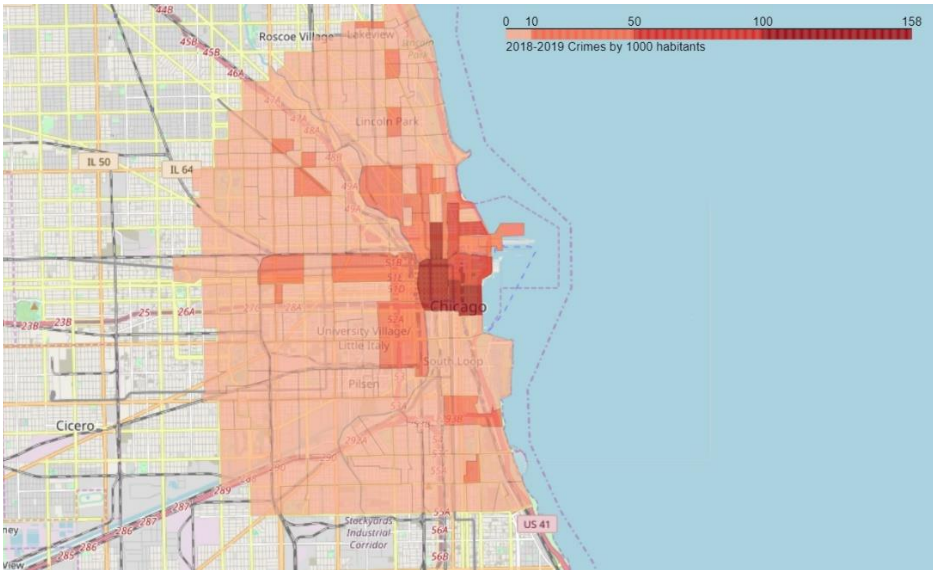
0 10 50 100 158
2018-2019 Crimes by 1000 habitants
Roscoe Village
Lakeview
Lincoln Park
IL 50
IL 64
Chicago
University Village/ Little Italy
South Loop
Pilsen
Cicero
Stockyards Industrial Corridor
US 41

Figure 5. Crime rate

Finally, the last variable about the number of restaurants in figure 6. The hypothesis is the following: A high number of restaurants might reduce the chance of visiting the new restaurant. The idea is to avoid census tracts with high number of restaurants as it is observed in the north and center of the study area.

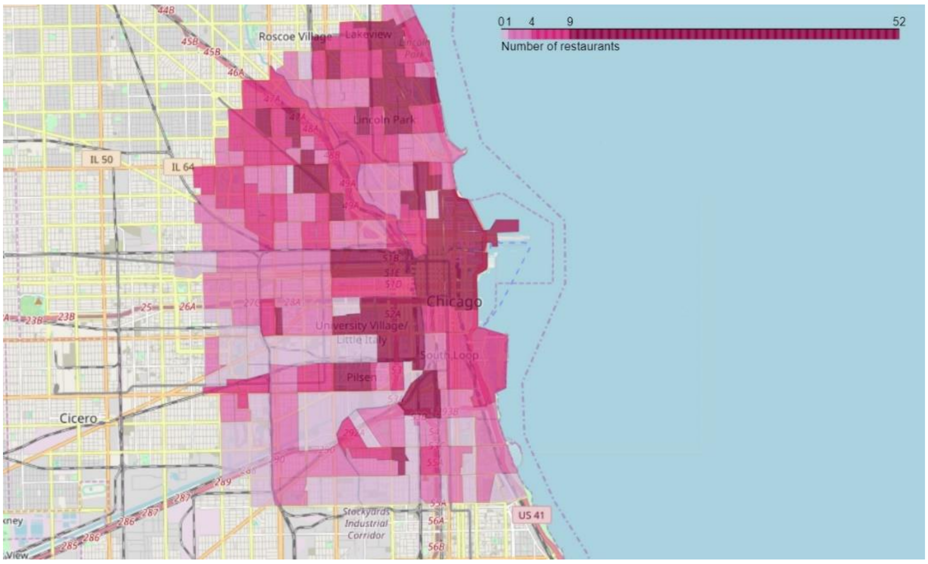

Figure 6. Number of restaurants

By visualizing these variables through maps, an entrepreneur can have a first impression about where he/she can open a new Peruvian restaurant based on social, economic and cultural aspects. However, it is not enough since he/she should consider what kind of competitors are present in each census tract. Next section takes into consideration this characteristic for modeling.

3.2 Modeling

This study proposes a model to build a single score based on the 8 main variables of analysis in Jupyter notebook. Since all census tracts in the study area might be potential areas for opening a new restaurant because they are within the CBD or close to it, assigning a score to each census tract could be convenient and thus decide where to invest first. Hence, 2 scores were obtained separately given the characteristics of data types. A first score was calculated as a weighted mean of normalized variables. Normalization is the process of transforming values of several variables into a similar range. There are several types of normalizations, but this project considers scaling variable, so the variable values range from 0 to 1 by dividing each value of a variable by its maximum value. After normalization, it was defined a weight to each variable according to its relative importance.

The formula of the first score is as follows:

$$Score\ 1 = \frac{(5 * N.inc) - (3 * N.une) + (5 * N.dpo) + (2 * N.lat) - (3 * N.cri) - (2 * N.rest)}{20}$$

Where:

N. inc: Normalized variable for income

N. une: Normalized variable for unemployment rate

N. dpo: Normalized variable for density population

N. lat: Normalized variable for Latino population rate

N. cri: Normalized variable for crime rate

N. rest: Normalized variable for the number of restaurants

On the other hand, the second score comes from string variables. Hence, a simple condition was made to assign a second score (score 2):

- If Latin American restaurants become the first or second most common restaurant in a census tract, then a value of 1 is given.
- Otherwise, a value of 3 is given.

In this study, by observing the category of restaurants in the data frame, it was defined as Latin American restaurants those from Argentina, Brazil, the Caribbean, Cuba, the Latin American, Mexico, Peru and the South American. Other countries known as Latin American were not found. Then, a final score is the simple sum of both scores as follows.

$$Final\ score = score\ 1 + score\ 2$$

Thus, the higher the score, the more potential an area is to open a Peruvian restaurant in Chicago.

4. Results

A first map (divided by 4 quantiles) showing scores 1 based on social, economic and cultural factors is seen in figure 7. To the light of the results, the map clearly illustrates a geographical pattern where high scores (the highest quantiles >0.16) were mainly located in the north and close to the center of CBD and therefore might be considered as potential areas to open a new Peruvian restaurant.

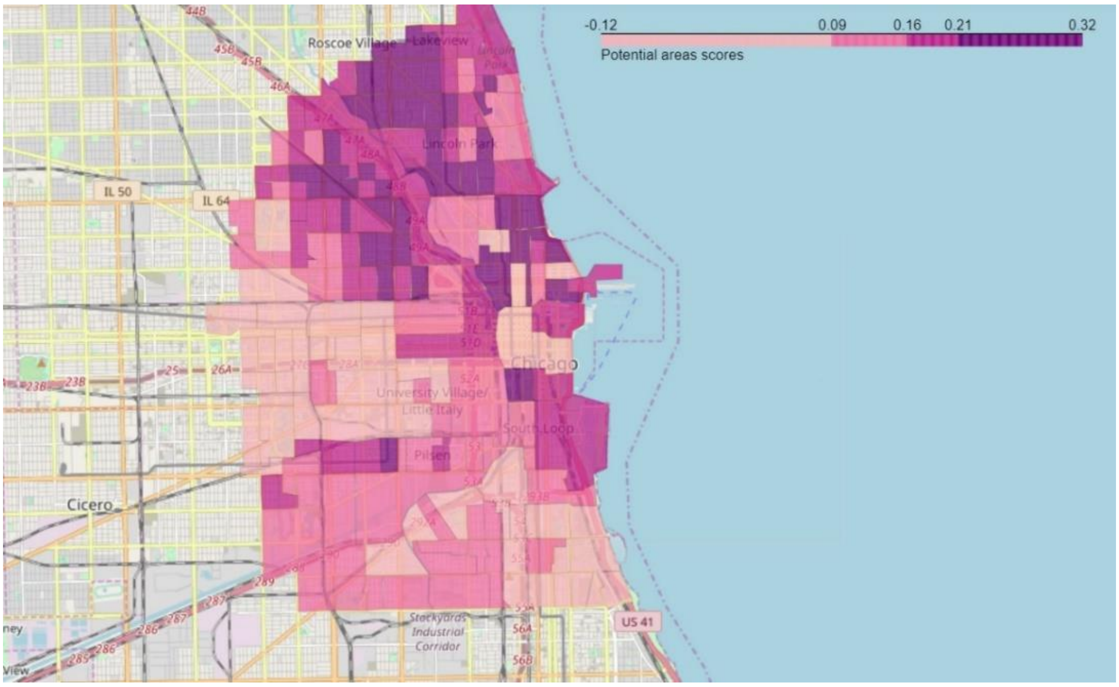

Figure 7. Map with scores 1

However, it is also meaningful to research what kind of food predominates in these areas. If Latin American cuisine is the most common, an entrepreneur would not be very interested in investing in those areas. The likelihood of getting revenues is lower because there will be more competitors with similar characteristics to the Peruvian cuisine. For that reason, it was created score 2 and add it to the first score to obtain a final score which can be seen in figure 8.

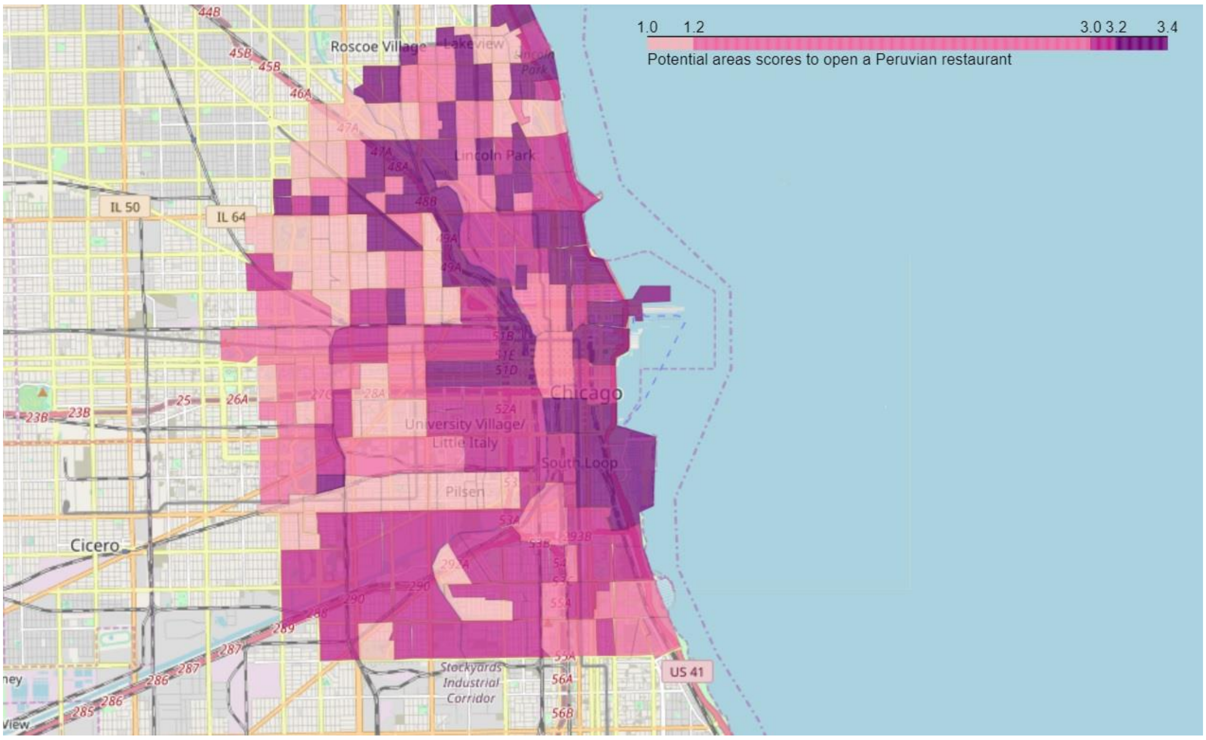
1.0 1.2 3.0 3.2 3.4
Potential areas scores to open a Peruvian restaurant
Roscoe Village
Lakeview
Lincoln Park
IL 50
IL 64
Chicago
University Village/ Little Italy
South Loop
Pilsen
Cicero
Stockyards Industrial Corridor
US 41

Figure 8. Map with final scores (4 quantiles)

The results suggest that an entrepreneur interested in investing in a new Peruvian restaurant, should mainly locate in some areas of the north or close to the CBD of Chicago (scores above 3.2). However, areas at the south seem to be interesting as well given that there are no enough restaurants, crime rate is low and there is a high percentage of Latino population. Additionally, figure 9 tries to validate the fact that an entrepreneur should open a Peruvian restaurant in areas with high scores since there are no enough Latin American restaurants there. Hence, a Peruvian restaurant in those areas might be a great business opportunity and above all in the north where the income is high, unemployment rate is low, density population is somewhat high and crime rate is low.

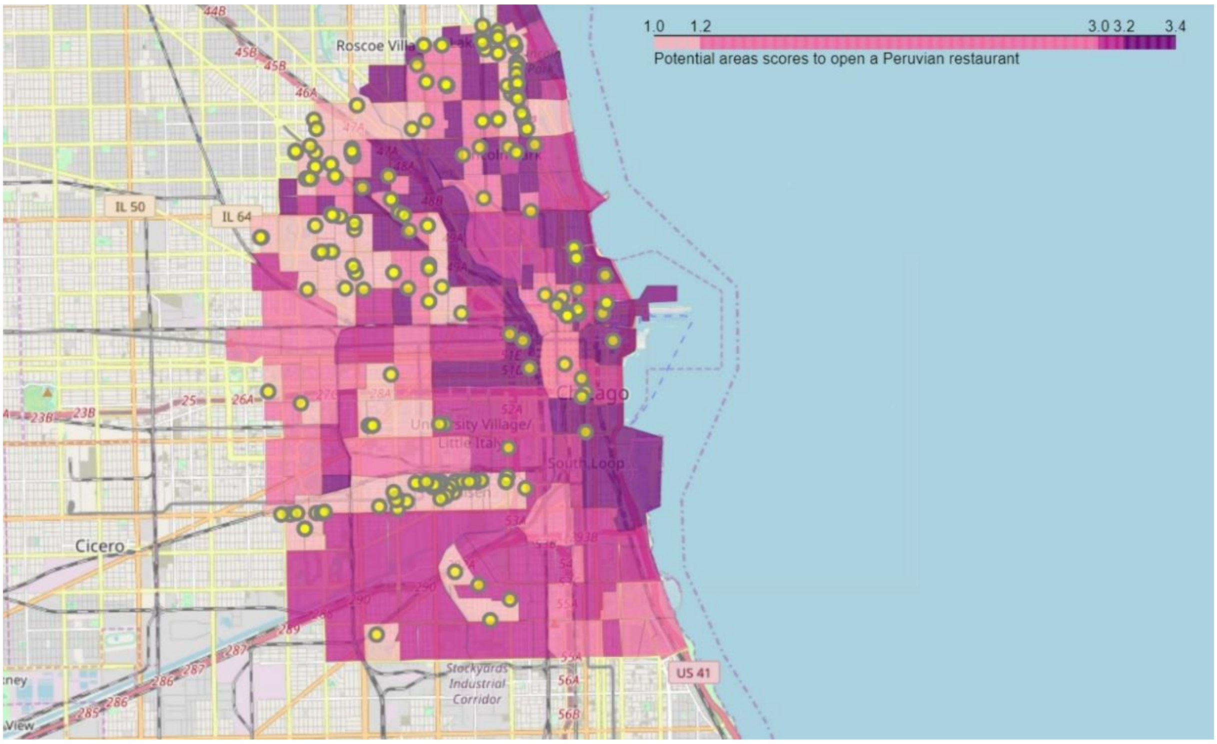
1.0 1.2 3.0 3.2 3.4
Potential areas scores to open a Peruvian restaurant
Roscoe Village
IL 50
IL 64
Chicago
University Village/ Little Italy
South Loop
Cicero
Stockyards Industrial Corridor
US 41

Figure 9. Latin American restaurants on potential areas

5. Discussion

This study used map visualization techniques and a model to build a score to decide where would be suitable to open a new Peruvian restaurant. This study has showed potential areas to open a new Peruvian restaurant in Chicago city, above all in the north or close to the CBD. In these areas income and density population are high, while unemployment and crime are low. To arrive to this recommendation, this study has demonstrated the importance of using a single score based on social, economic, cultural, and business factors derived from different sources as Census Bureau, Foursquare and Chicago Data Portal.

The final score was built using 2 previous scores. Some of them based on social, economic and cultural aspect and another based on business issues. Regarding to the last, this study used Foursquare API to build 3 variables: number of restaurants, the first and second most common restaurant in each census tract. This allowed to research whether an area presented a high number of competitors and moreover knows the type of food offered by restaurants which were very diverse: European, Asian, Indian, Latin American, etc. It helped to have a more confident score to decide where to invest first.

Since this study has made use of only some variables as components of the final score, it must be considered as a limitation. Hence, this study must be considered as a first approximation to find potential areas to open a new restaurant and might be improved by considering other variables that can be missing here. For example: food expenses, nearness to Farm markets or supermarkets.

6. Conclusion

This project proposed a simple model to build a score on census tracts to decide where to open a new Peruvian restaurant in Chicago. This project makes a useful finding by identifying the areas with the highest potential to open a new Peruvian restaurant in the north or close to the CBD of the study area. These areas characterize by having high income and density, and low unemployment and crime.

These findings provide insights to any Peruvian entrepreneurs who are interested in opening a new restaurant business in the CBD of Chicago or close to it. This study has provided a first approximation for the exploration of opening restaurants by using mapping visualizations techniques and a simple model for scoring and will serve as a base for future studies.