Santiago de Chile’s food venues Clusterization

Camilo Fuentes Moenne

January 2020

1. Introduction

1.1. Background

After the events on October 2019 in Chile, little and medium size companies have suffered from a high decrease in their sales. Nevertheless, a crisis might represent an opportunity. It is from common knowledge that a bad economic scenario tends to be beneficial to lower prices food venues such as fast food and sandwiches places in the long run. In Santiago de Chile, a city divided in communes, there are plenty of independent food venues option that were created by entrepreneurs and that now face new chances for expansion or relocation as part of business strategy.

1.2. Problem

There has never been an accessible categorization of how the different communes behave in terms of food venues preferences. This investigation will propose a method to group and assign preferences to the group of communes.

1.3. Interest

The results of this investigation can be used to suggest possible expansion or relocation for food venues based on the popularity of their category in a group of communes.

2. Data Acquisition

2.1. Data Sources

For the venue information, I used the Foursquare API to collect data from different locations in Santiago given the geolocation of the communes with a certain ratio heuristically determined. For this investigation, the standard free account was utilized so the number of requests per day was also limitated.

The geolocation of the commune is public information that can be found at the SINIM (“Sistema Nacional de Información Municipal” or National System of Communal Information) or extracted from Wikipedia article about Chile’s communes that has now the exact same information after performing some corrections about some communes.

2.2. Data Cleaning

Geolocated data can come in many formats (named geotagging):

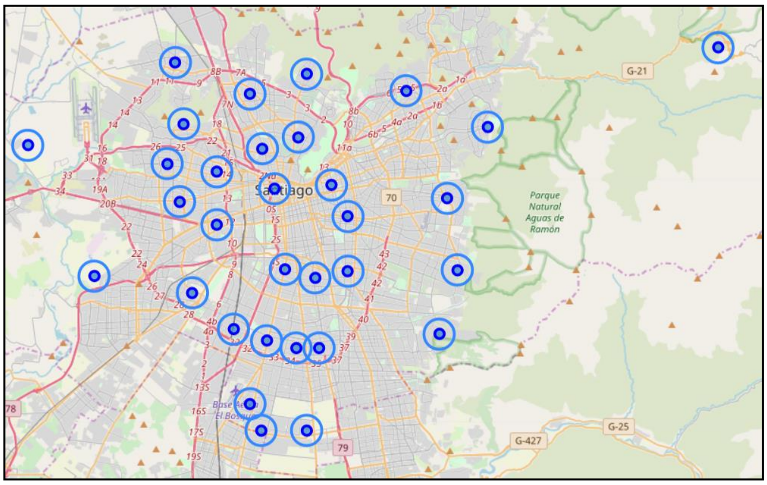
Template	Description	Example
[-]d.d, [-]d.d	Decimal degrees with negative numbers for South and West.	12.3456, –98.7654
d° m.m′ {N|S}, d° m.m′ {E|W}	Degrees and decimal minutes with N, S, E or W suffix for North, South, East, West	12° 20.736′ N, 98° 45.924′ W
{N|S} d° m.m′ {E|W} d° m.m′	Degrees and decimal minutes with N, S, E or W prefix for North, South, East, West	N 12° 20.736′, W 98° 45.924′
d° m' s" {N|S}, d° m' s" {E|W}	Degrees, minutes and seconds with N, S, E or W suffix for North, South, East, West	12° 20' 44" N, 98° 45' 55" W
{N|S} d° m' s", {E|W} d° m' s"	Degrees, minutes and seconds with N, S, E or W prefix for North, South, East, West	N 12° 20' 44", W 98° 45' 55"

Since the data used by the API is decimal degrees and the collected data was degrees, minutes and seconds I created a function to format the data.

The API also needs to receive a diameter to make an exploratory search given a geographical location, I chose an initial radius of 1 km around the center of the communes to start. This radius will change individually for each commune based on a minimum amount of venues that will be set to perform the analysis.

2.3. Feature selection

Since data from all Chile was collected, I filtered by the Santiago province in order to get the location I needed to work on. After everything was set, I performed a first view of what my starting exploration would be like using the Folium library where the communal points were represented surrounded by their initial radius.

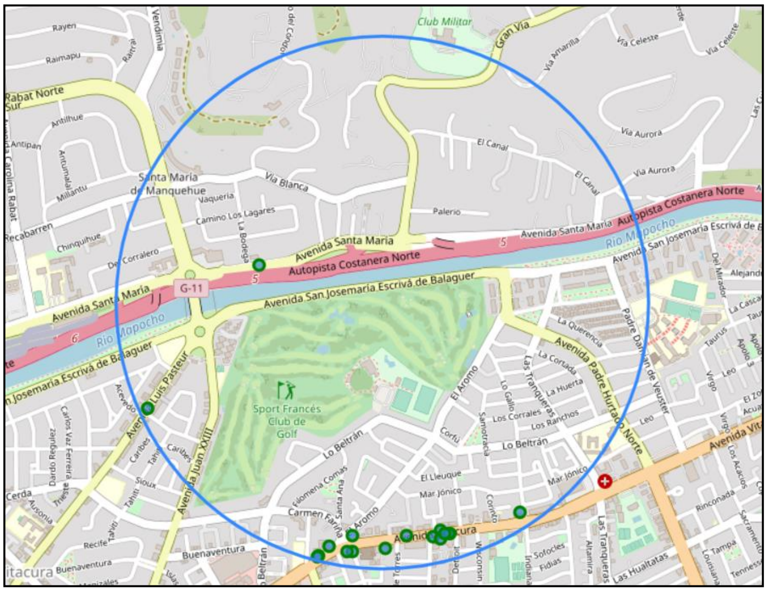


Picture 1: first exploration dataset of Santiago de Chile's communes.

3. Exploratory data analysis and methodology

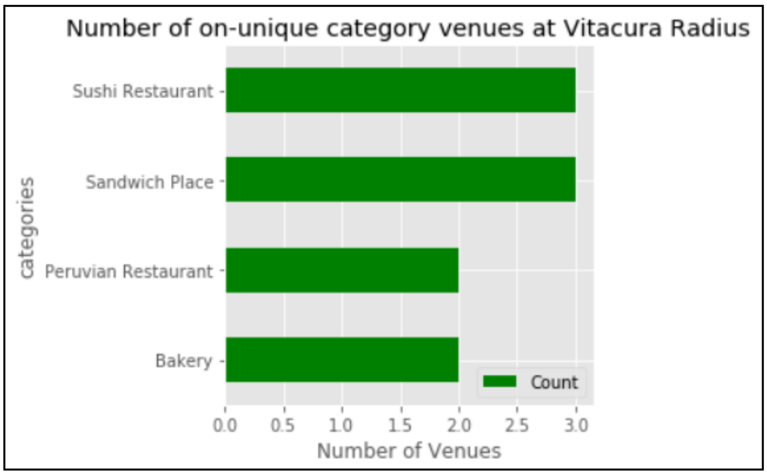
3.1. Test Sample

Before extracting data from every commune, I performed a test scan in an arbitrary commune to know what to expect. I chose the commune of Vitacura for this testing and I plotted the different results for the food section in Foursquare API:



Picture 2: communal testing results for food Venues at Vitacura.

The green dots represent the venues given by the API (17 in total) and the blue circle represents Vitacura’s 1-kilometer initial radius. I then plotted the non-unique categories of venues to see which the initial communal preferences were.

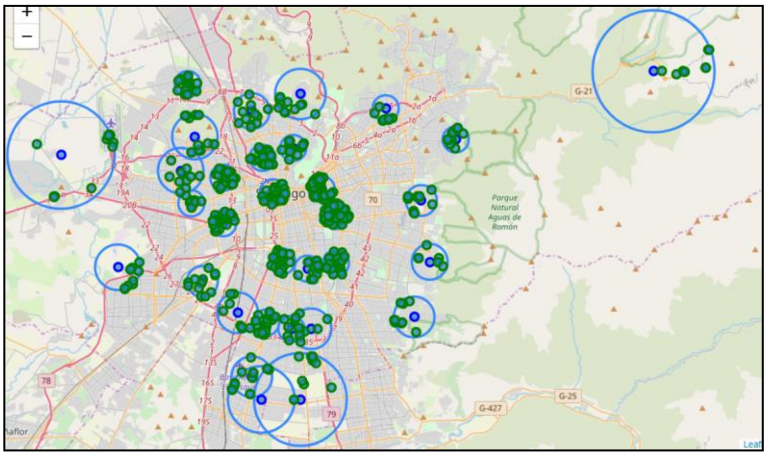


Picture 3: Vitacura’s more common venues.

Sushi restaurants and sushi-places represent approximately 35% of the categories given by Foursquare within the radius of analysis.

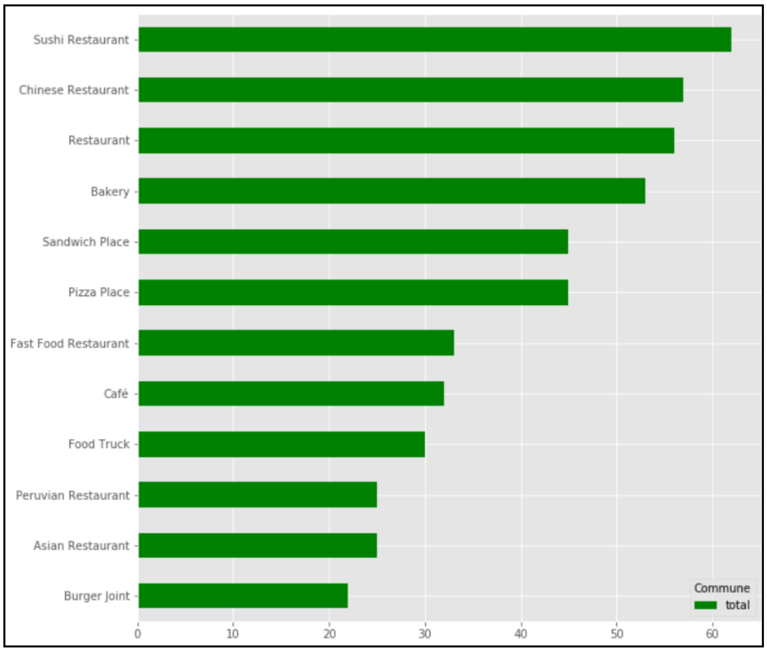
3.2. Whole City

Then I did the same exercise for the whole city, I imposed a radius incrementation of 200 meters on each commune with fewer than 10 venues associated with it until we reached this value, here is the result:



Picture 4: Map of Santiago's Foursquare venues.

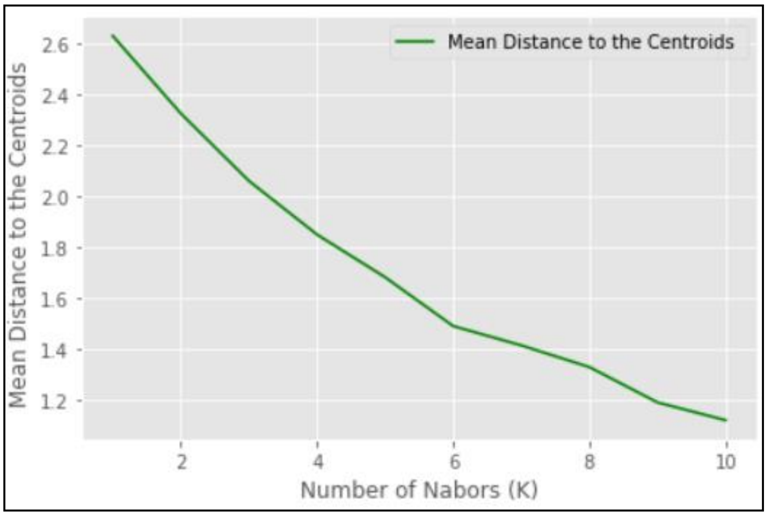
Then I plotted what the most common venues categories in the city are given our coordinates:



Picture 5: Santiago's Foursquare most common venues.

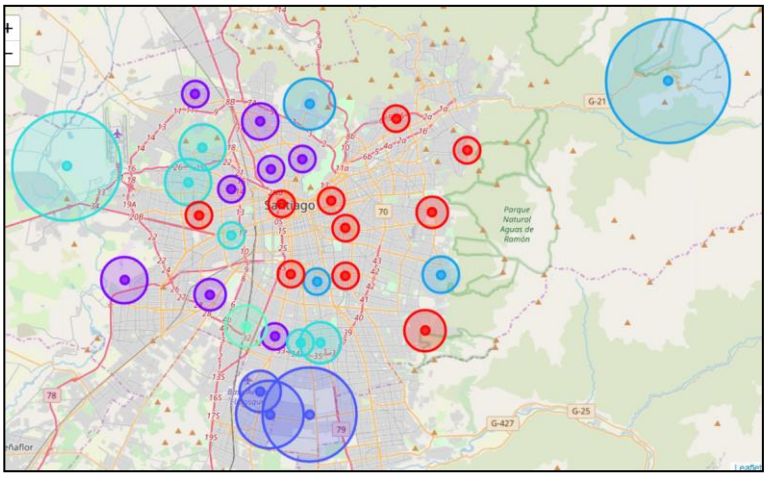
3.3. Clusterization

So now we will take the 3 most popular categories for each commune and use a K-Mean algorithm so seek a good clusterization of the sample given those 3 categories. In order to get a good fit, using the Euclidian distance I plotted the mean distance from the centroids given a certain k (number of neighbors to consider in the algorithm) which I iterated between 1 and 10, here is the result:



Picture 6: Mean distance from centroids vs number of k neighbors.

Using the elbow method, I determined that the optimal k number was 6. Then I plotted the points with colors according to their respective clusters:



Picture 7: Food venues clusterization of Santiago's commune given the 3 more popular venue categories on Foursquare.

4. Results, discussions and considerations

Looking inside the clusters, I could define the next names:

Cluster	Name	Color
1	Sushi and Sandwiches	🟥
2	Chinese and Fast Food Restaurants	🟪
3	Food Truck	🟦
4	Food Truck and Snacks	🟦
5	Bakery and Pizza	🟦
6	Not Defined	🟩

The fact that the sixth cluster results in a none-defined group and that some items are repeated is not a good sign of clusterization but helps with an overview of the whole situation given certain geographic points. It is safe to say that cluster 6 does not bring useful information about se communes behavior.

What can be improved?

- The number of iterations the algorithm could do to find new venues given new radiuses was given by the account type at Foursquare, perhaps with more tries better results can show up. Also, with an improved account it will be possible to include the ranking of Foursquare's users as parameter to the analysis.
- The geolocations of communes seem not to be the best of starting points for a radius, there is much terrain that wasn't covered here. It might be very interesting to locate the neighborhoods of Santiago instead of the communes and have a best shape of the terrain.
- Some venues might not be rated in Foursquare application since it has not yet penetrated into the market enough, perhaps another application or metric could be considered.

What was accomplished through this investigation?

- We have little portions of terrain from Santiago the Chile that might not be representative of their communes but are correlated between the nature of business in food venues in the surroundings.
- We have now delimited certain areas based on Foursquare showing the different venues and clustered them by their categories.
- We have taken a first step in understanding how the food market is segmented in a poorly explored area.

5. Conclusion

Even if this investigation didn’t achieve the optimal segmentation for the city, it throws highlights in how to cover this subject and what results are expected in some specific areas. Five of six clusters can be used to suggest similar behaviors and potential in some areas from Santiago de Chile.

6. Future Discursions

There is a lot that can be done starting with realize a similar analysis with other types of venues, other geospatial locations and other algorithms. Also, there is a chance that popularity between a given space doesn’t guarantee the survival of a business, it would be a good subject of investigation to look for the components that make the success of new business in the city such as the here described category given a commune.