Poverty versus services in Santiago's communes

Osvaldo Gac Pabst

December 20, 2019

1.- Introduction

1.1.- Backgound

Santiago of Chile, is the capital and largest city of Chile as well as one of the largest cities in the Americas. It is the center of Chile's largest and most densely populated conurbation, the Santiago Metropolitan Region, whose total population is 7 million. The city is entirely located in the country's central valley. Most of the city lies between 500 m (1,640 ft) and 650 m (2,133 ft) above mean sea level. Santiago is the cultural, political and financial center of Chile and is home to the regional headquarters of many multinational corporations. The Chilean executive and judiciary are located in Santiago, but Congress meets mostly in nearby Valparaíso.

The whole of Greater Santiago does not fit perfectly into any administrative division, as it extends into four different provinces and 37 communes. The majority of its 641.4 km2 (247.65 sq mi) (as of 2002) lie within Santiago Province, with some peripheral areas contained in the provinces of Cordillera, Maipo and Talagante.

Chile is into the 10 countries most inequality in the world, reflected in Gini's index 0.47 in 2018, in Santiago you can find poor communes and rich communes, nowadays difficult to understand in term of what kind of services and shops there are in each commune.

1.2.- Problem

the objective of the project is to explore the possibility of, taking advantage of state of the art Machine Learning Algorithms, to segment the communes using Foursquare API to get all the venues and services and compare with the poverty index of each commune, to resolve the questions:

"A more poor the commune less services?"

“What type of services characterizes the poor communes and the rich communes”

This will help to stakeholder to decide the services they want to develop in a commune considering the lack of services per commune and the level of richness

2.- Methodology

The methodology used in the project consist in: Data Collection, Principally is a secondary source get from searching on Internet and using Foursquare API Search for poverty index by communes and map it to observe the different zones in Santiago. To get the data, I scraping the Wikipedia page where I found a table with the information for each commune and the respectively poverty index.

I used python folium library to visualize geographically details of Santiago and how looks each communes with different index: poverty and number of services. Using Foursquare get the venues per communes

To get the segmentation is used k-means algorithm, the reason to use this algorithm is because segmentation is the practice of partitioning a communes base into groups that have similar characteristics.

To determine the better number of segments is used: Elbow Point, Silhouette score and Davies Bouldin score and giving us the possibility to compare all of this methods, get the number of clusters and finally analyzes each group or segment of venues for each commune and the relation with the poverty index.

3.- Data acquisition and cleaning

3.1.- Data source

To consider the problem we can list the data as below:

I found the index of poverty of each commune to 2015 [reference 1].

I used Foursquare API to get the most common venues of given communes of Santiago [Reference 2] .

There are not too many maps of Santiago in geopandas or shapefile map, but not in geojson, so I built geojson file, including the most important and populated communes in Santiago, using the information of the border point from Carto Maps [Reference 3].

3.2.- Data cleaning

The data from the Wikipedia where near complete and good formatting, only was necessary the follow process:

a) Eliminate all the latin characters like: “í”, “ó”, “ñ”, etc, because where load the geojson all of these characters were not interpreted rightly, so this will complicate the using of the information to create a chroropleth map.

b) Correct the number format, eliminating the coma like decimal separator and the percentage, where I eliminate the symbol “%” creating another column without it.
c) The information of the foursquare was clean and ready to use

3.3.- Feature selection

The feature selection was done as follow:

a) The Wikipedia only I need the poverty index and the name of each commune, for this reason I delete the extra information to get the follow dataframe:

	Comuna	Location	Population_(2017)	Population_density_(2002)	Population_growth_(2002-2017)	Poverty_(2015)
0	Cerrillos	surponiente	80832	4329.08	12.9%	19.7
1	Cerro Navia	norponiente	132622	13482.91	-10.7%	35.6
2	Conchalí	norte	126955	12070.29	-4.4%	21.6
3	El Bosque	sur	162505	12270.72	-7.3%	27.0
4	Estación Central	surponiente	147041	9036.31	16.6%	14.5

b) Inspecting the dataframe, I realize that Peñaflor commune doesn’t have any value including the poverty index, so I decided to delete, because for the purpose of this work, there are enough communes to find the relation between poverty and the amount and kind of the venues.

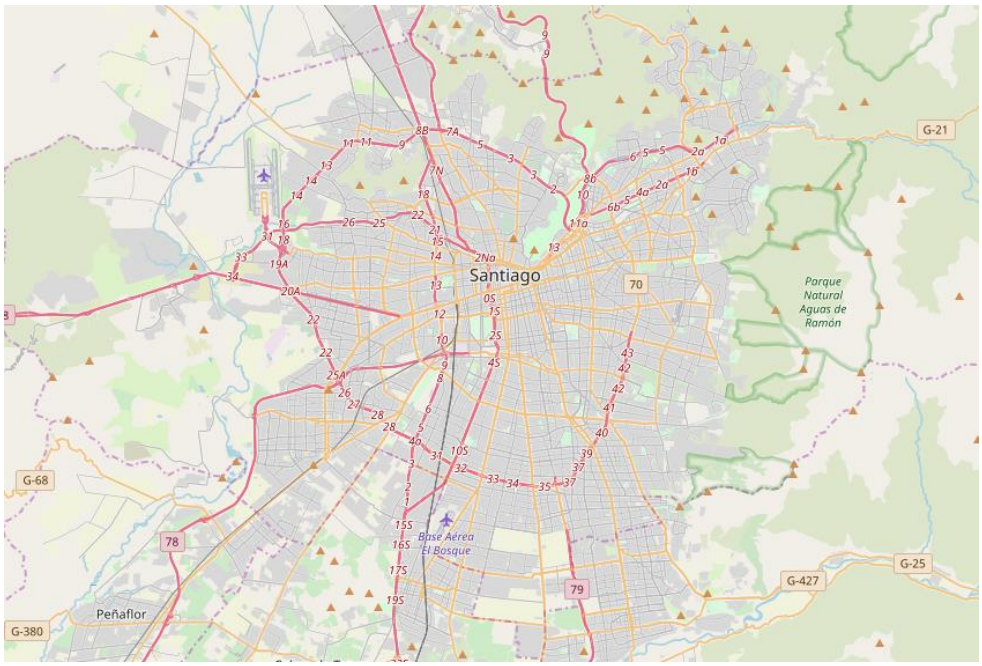
c) To map and obtain the venues from Foursquare, was necessary to get the geographical location using the Nominatim platform, where I get the latitude and longitude for each commune, this information was added to the original dataframe to get the following

	Comuna	Location	Population_(2017)	Population_density_(2002)	Population_growth_(2002-2017)	Poverty_(2015)	Population_growth_(2002-2017)_%	Latitude	Longitude
0	Cerrillos	surponiente	80832	4329.08	12.9%	19.7	12.9	-33.502503	-70.715918
1	Cerro Navia	norponiente	132622	13482.91	-10.7%	35.6	-10.7	-33.425145	-70.743954
2	Conchalí	norte	126955	12070.29	-4.4%	21.6	-4.4	-33.385096	-70.674491
3	El Bosque	sur	162505	12270.72	-7.3%	27.0	-7.3	-33.562352	-70.676820
4	Estación Central	surponiente	147041	9036.31	16.6%	14.5	16.6	-33.463658	-70.704966

d) From Foursquare I get all the venues from each communes using the limits: Radius = 2000 and Limit 100. In total ware 1.788 venues, ready to use in the analysis.

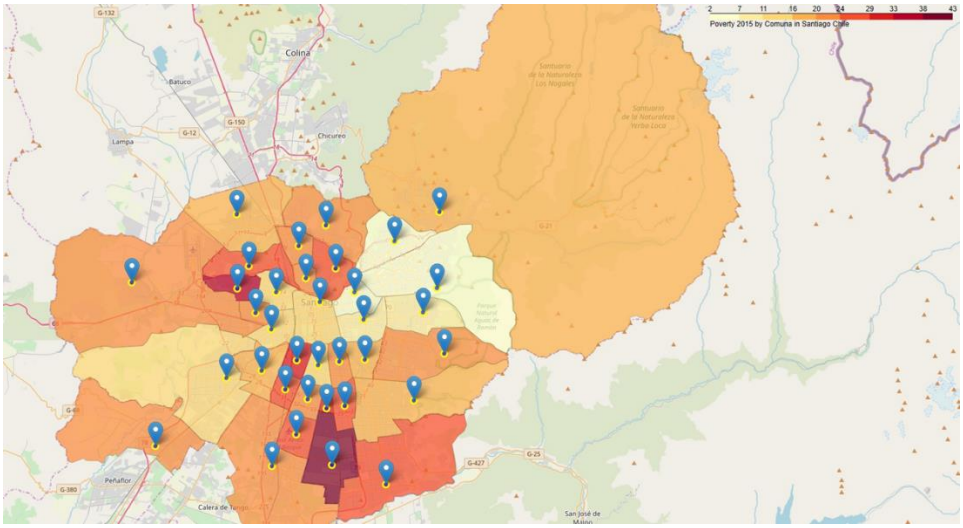
4.- Analysis

I uses different maps to visualize the information and concluded in a easy way, the next is a Santiago's map clean (map 1)



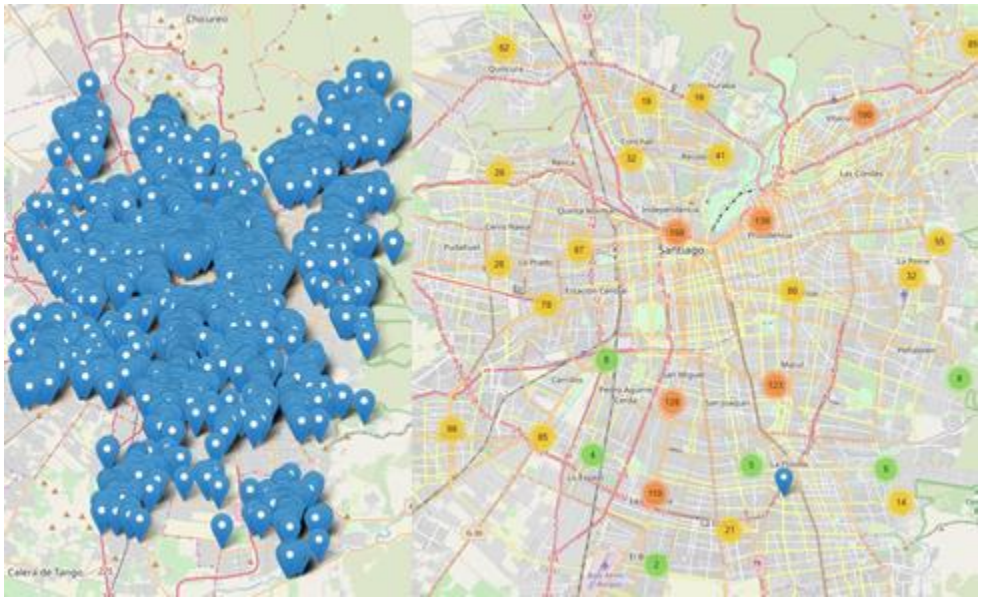
Map 1: Clean map of Santiago of Chile

When is uses the information of the poverty index is clear to see the poor communes are located to the west and the rich to the east (map 2)



Map 2: Choropleth map of Santiago, while more poor more dark

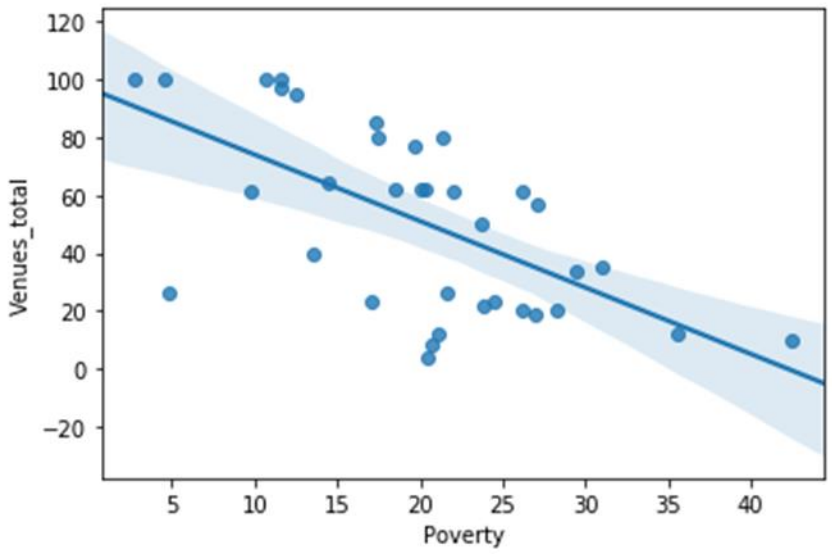
When I analyze the venues, it is possible to observe the density that Foursquare have in the platform, so to have a view more sorted I create other map that include a venue cluster



Map 3: Venues in Santiago

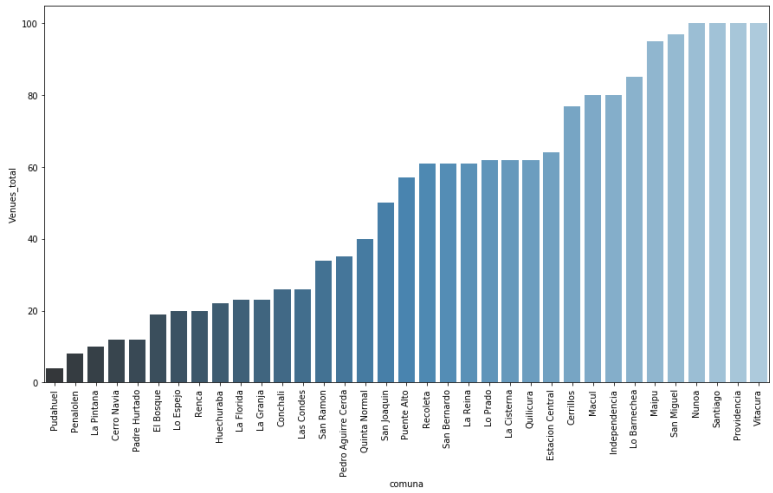
Map: 4: Santiago with venues cluster

To observe the relationship between the poverty and the amount of the venues for each commune, I used a scatter plot (Plot 1)



Plot 1: Scatter Plot show the relation between poverty and amount of venues

We can observe there is a big difference between the rich and poor communes in term of the amount of venues, the follow plot is clear (Plot 2)



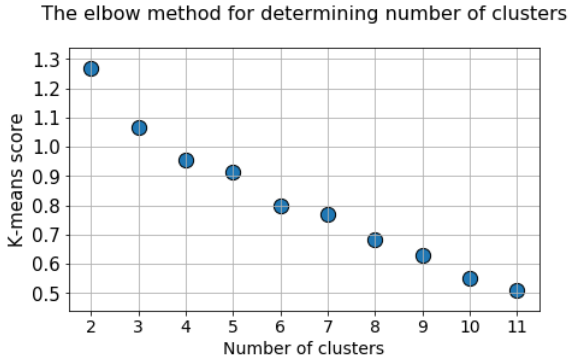
Plot 2: Amount of venues versus commune

While Pudahuel has a few venues, Vitacura, Providencia, Santiago and Ñuñoa has the maximum (100 venues)

4.- Modeling

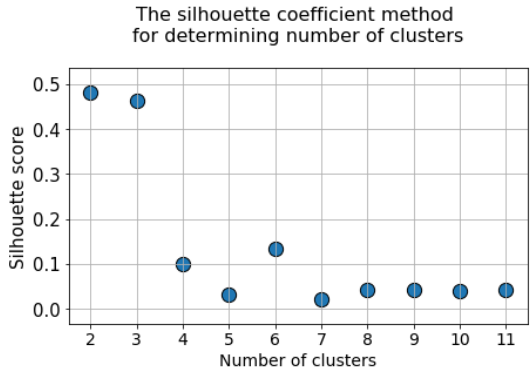
The algorithm used to model is k-means, since the problem to resolve is clustering. First the data must be normalize, for this case I take the mean of the frequency of occurrence of each category and I use 3 methods to determine the better cluster number: Elbow point, Silhouette and Davies Bouldin score. In particular, plotting Silhouette score for different k cluster show a maximum value where the clusters have the minimum distance between samples that belong to a cluster and the maximum distance with samples that belong to another clusters.

Running k-means from 2 to 12 clusters I get the following points (Plot 2:



Plot 2: Scatter plot K-means score versus k cluster

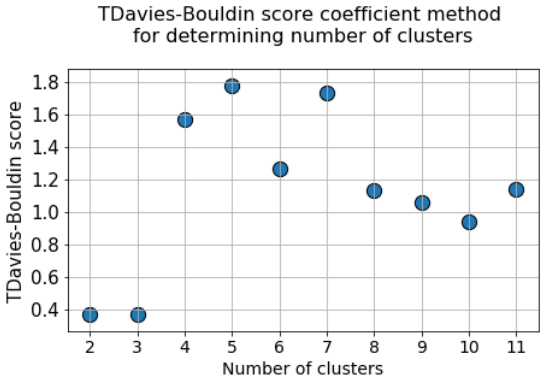
It is not clear to know which is the Elbow point, there is a little change in 4 or 6, but it is not sure. Let's see the plot for Silhouette score (Plot 3):



Plot 3: Scatter plot Silhouette score versus k cluster

It is simple to observe there are 2 point with maximum, 2 and 3, so the better cluster number could be one of them.

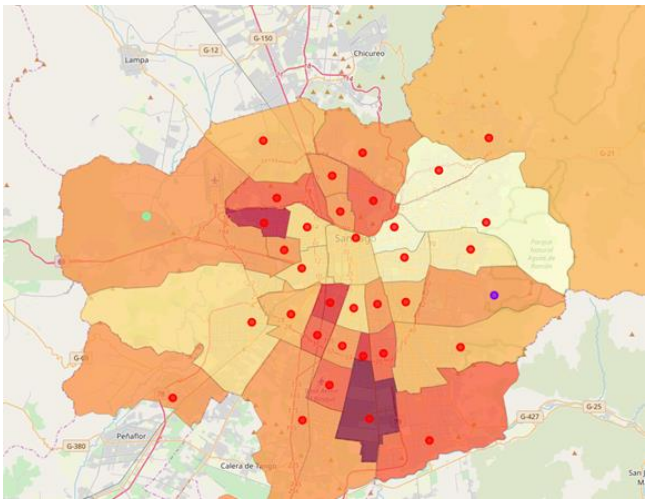
To be sure I use a third method Davies Bouldin score, the plot is the following (Plot 4):



Plot 4: Scatter plot Davies Bouldin score versus k cluster

As Silhouette score, Davies Bouldin score shows 2 and 3 like best numbers of cluster for our data segmentation, so I take 3 like the cluster number to continue the study.

After run the k-means to k = 3, I create a map of Santiago where include the poverty score and the segmentation (Map 5)



Map 5: Map of Santiago, where the communes are painted with the poverty score, dark is poorer and the points represent the segment

The map shows there are two segments with only 1 commune and one segment with the rest of the communes. This suggest there are not difference of the types of venues in different communes, since the commune is rich or poor.

The cluster 1 show that contain a lot of foo venues, while the cluster 2 include park and Garden center, and the third cluster include Zoo exhibit and werehouses.

5.- Result and discussion

Santiago of Chile is a big city and very populated, when travel along the city is very clear what are the poor communes and the rich communes, because the size of houses, gardens and the cars, but it seem when compare the communes with the most common venues, there is no differences between them.

There is a downward trend between poverty and the total venues in each commune, so it is expected that there are more venues in rich communes, where people can afford more restaurant dinner, go shopping and so on.

I used K-means to do the segmentation of the communes, but there are others algorithm to develop this task and I also included 35 communes of Santiago, but it could be included in the study the farthest communes, since those maintain rural life.

When analyze the different score, the elbow point is not clear to determine the number of clusters, probably the cluster are not clearly separated, but when is analyze the Silhouette or Davies Bouldin score is more clear to conclude there are 2 or 3 clusters. I took 3 cluster, however the difference is little

It is very known the difference between the poor and rich communes, to determine the difference is necessary to understand the behavior of the people, so if the analysis is focus in the services like bank's branch offices, Cars shops and Mall is possible to find more difference.

Other important problem is determine the exact border of each commune, to avoid include a venue in the commune that does not belong to. I took 2 km, but is very probable there are some venues in 2 communes and others that were included, because the border is bigger than 2 km.

6.- Conclusions

Using the actual Data Analytics and ML tools to get and analyze information make the life easier and allow to take better decisions.

In this study I found a downward in the amount of venues there are in the poorest versus the richest communes but there is no difference comparing the kind of venues, when the comparison is having done taking the common venues, but for my surprise the food venues are a lot in al the communes.

7.- References

7.1 .- Information from the web pages:

[1] https://es.wikipedia.org/wiki/Anexo:Comunas_de_Santiago_de_Chile

[2] https://en.wikipedia.org/wiki/Santiago

[3] Forsquare API

[4] https://carto.com/

7.2.- Articles

7.2.1.- Article written by by Tirthajyoti Sarkar “Clustering metrics better than elbow-method“ (https://towardsdatascience.com/clustering-metrics-better-than-the-elbow-method-6926e1f723a6PhD)