Clusterizing Santiago neighborhoods with foursquare data

1. Introduction

I am using the Foursquare data to clusterize different neighborhoods in my city, Santiago de Chile. This clusterization could be useful for many purposes, but I am focusing my analysis on its real estate implications, hopefully getting an answer to the following question: if a new real estate appears to market, which would be a good description of its neighborhood?

The features feeding my algorithm are the absolute abundance of the different type of venues cataloged by the Foursquare API.

The potential stakeholders for my analysis could be real estate agents looking for a high-level quantitative look of the neighborhoods they have in their portfolio. It could also serve people looking to rent in Santiago who want a general notion of the types of neighborhoods in the city, which can be useful to narrow and optimize their search.

2. Data

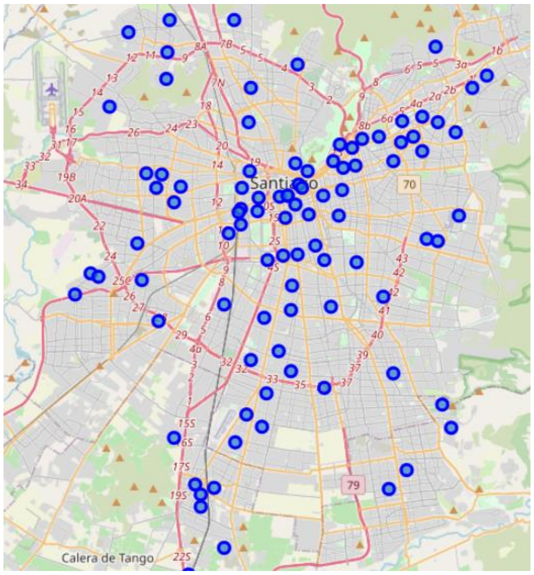
Foursquare provides data about the venues around a specific GPS coordinate such as coffee shops, yoga studios, international restaurants and more, as well as reviews and other features. I will combine the venue data with a database containing the coordinates of the main neighborhoods in Santiago, extracted from google maps.

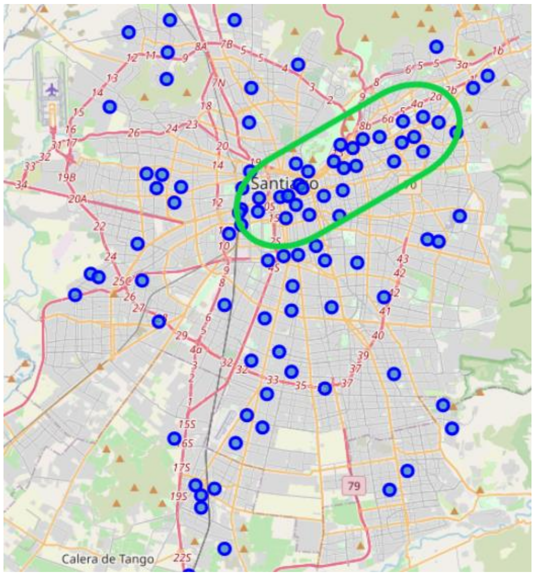
The coordinates of 93 neighborhoods are obtained from this custom google map. Its noteworthy that the coordinates are not readily available to download, and they must be extracted via RegEx from the source code.

3. Methodology

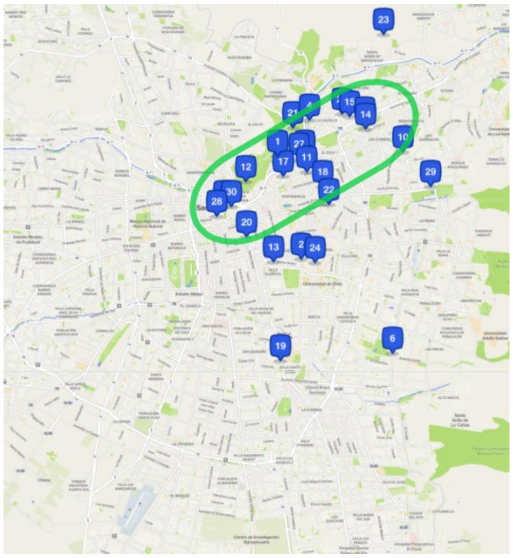
We begin our analysis with a map with all the identified neighborhoods of Santiago. We can observe that the zone with the denser accumulation of points correspond to the continuation of the Alameda-Providencia central avenues. This area is both upper-class and the sector with the most data available in Foursquare. This could be explained because Foursquare is not a well-known platform in chile, and its users tend to be people who travel abroad frequently, hence wealthier.

Google maps data



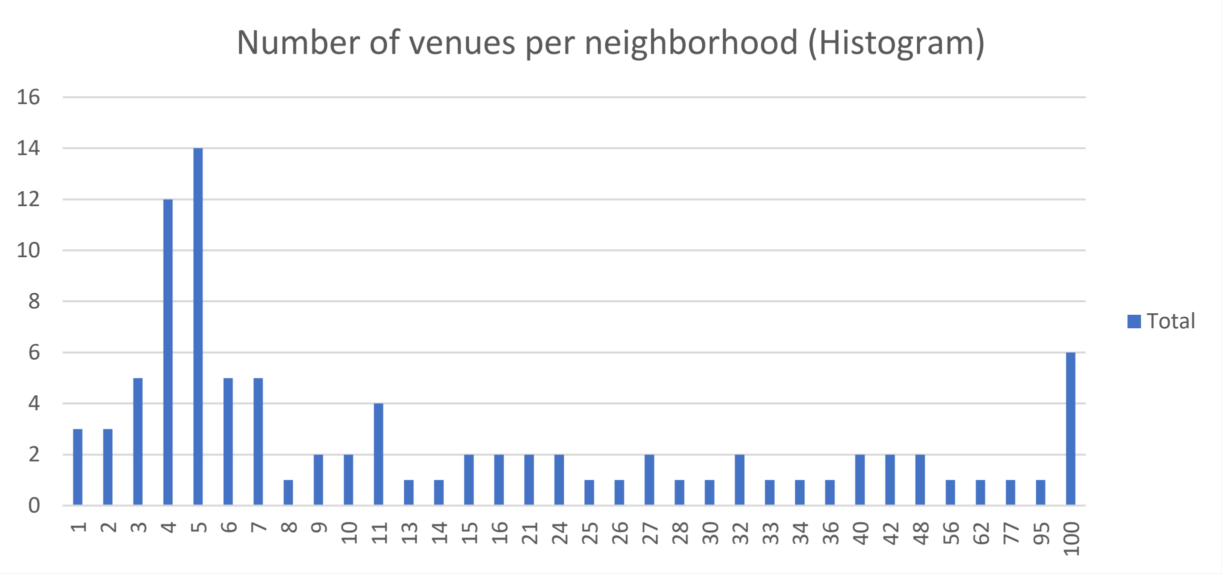


Foursquare data



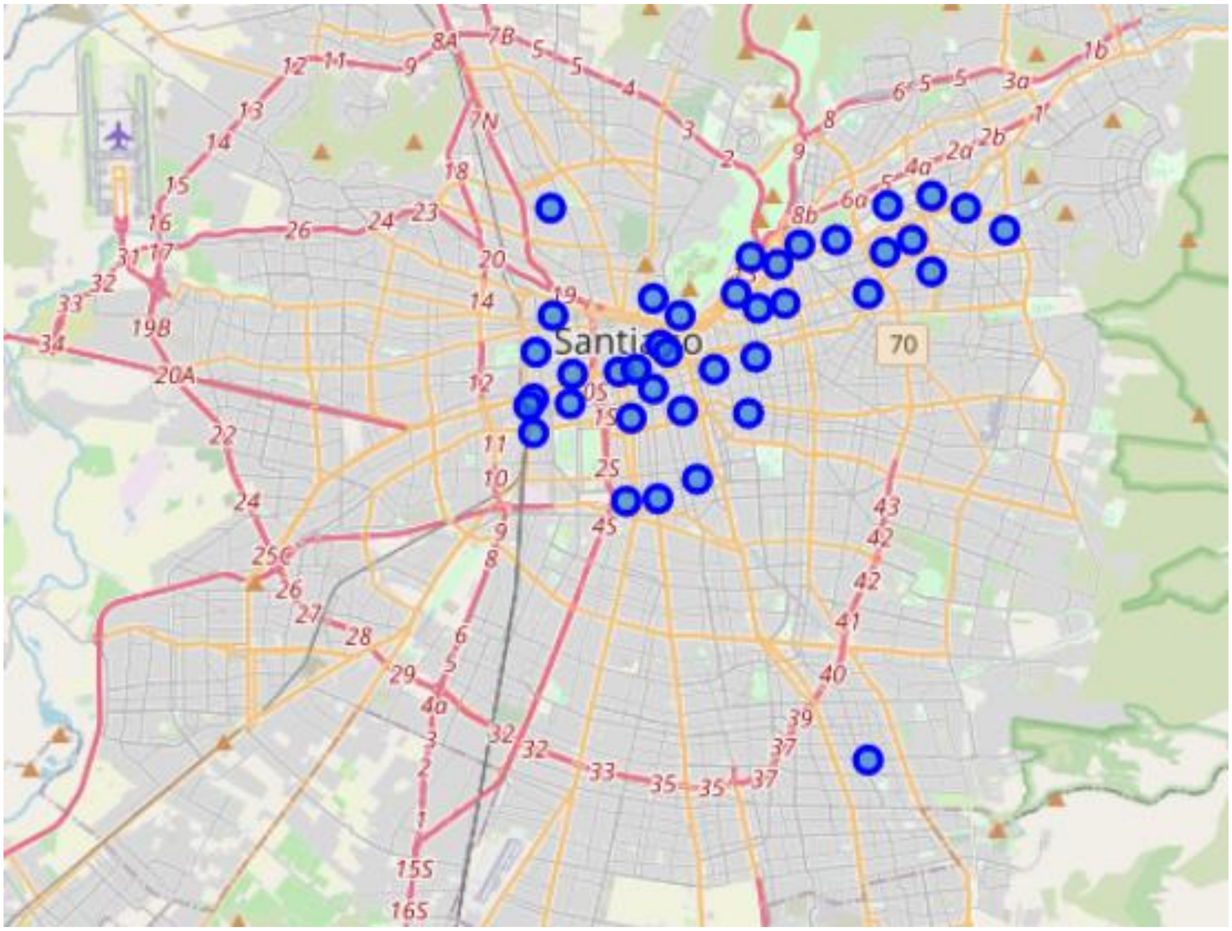
3.1 Problems with the data

This brings problems to our analysis. First, there are many neighborhoods that have five or less venues in their vicinity, which results in a poor characterization of them. Second, as our algorithm clusterize according to the similarity of the venues between neighborhoods. Those who have no information are similarly venue-less, so they all end in the same cluster when we know by experience that they could be potentially very different from each other. Given this conundrum I am filtering for neighborhoods that have 10 venues or more to assure some degree of characterization.



3.2 Filtered data

The new map is shown below. As expected, it mainly contains neighborhoods within the center and north-east sector of the city, which are also wealthier sectors. Our analysis is therefore limited to these areas but given new venue information the extension of our analysis is straight-forward.



3.3 Descriptive analysis

Below there is a list with the most common venues found in our neighborhoods. This can give a general sense of the venues in our dataset and a benchmark to compare our final clusters with.

Venue	Total
Coffee Shop	80
Restaurant	76
Sandwich Place	66
Café	58
Pizza Place	53
Bakery	53
Bar	50
Hotel	45
Peruvian Restaurant	37
Sushi Restaurant	34
Chinese Restaurant	34
Plaza	32
Pharmacy	31
Bookstore	29
Clothing Store	28
Burger Joint	27
Theater	25
Gym	22
South American Restaurant	22
Park	22
Latin American Restaurant	20

The resulting dataset has 239 different types of venues. This can become problematic since many venues are actually very similar in nature from one another, for example coffee shop and Café or South American Restaurant and Latin American Restaurant. This homologation of different but similar venues should be done in future studies and it could reduce artificially large distances between neighborhoods, resulting in more accurate clusters.

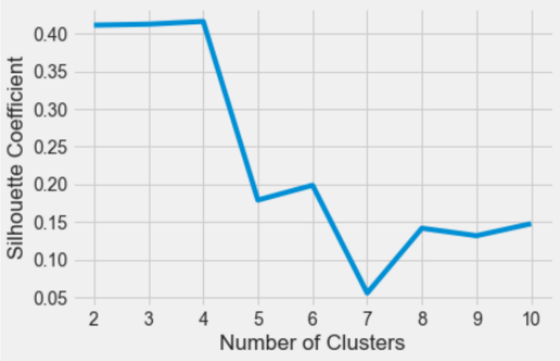
3.4 Algorithm and fit metrics

For this analysis I am using k-means algorithm. K-means is a popular machine learning algorithm because it scales well with the number of samples in the data, and it is easy to implement. But it has short comings for instance it can become stuck in an optimal minimum. On the other hand, it cannot detect arbitrary shapes in the data, and will create linear separation boundaries, furthermore, it is sensitive to outliers. Finally, it depends crucially on the numbers of clusters chosen a-priori, since k-means will always fit the input number of clusters into the data, regardless of if the data fit the clusters.

To choose the best number of clusters I complement 2 metrics usually used for this task, the silhouette coefficient and the elbow method.

Silhouette metric

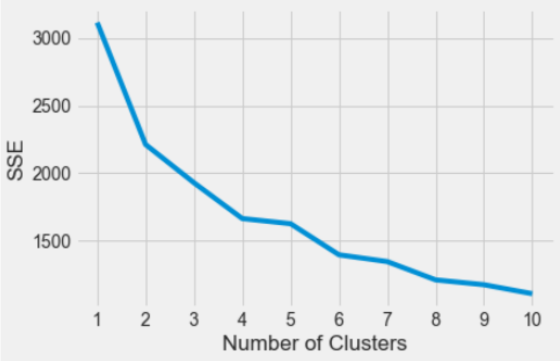
The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b)[1]. Values near 1 indicate that the difference in variances are minimal with respect to the inter-cluster variance. The highest Silhouette Coefficient for my data is at n=4.



The elbow method

The elbow method plots the sum of squared errors (SSE) left in each clusterization done with varying number of clusters. This curve will always be descending as more clusters allows for more flexibility and therefore a better fit. The elbow method consists in finding the point where the SSE significantly slows its descent and thus adding more clusters would be less and less significant.

The major elbow point in the following plot is at n=6, though at n=4 we can observe a similar descent in the SSE thus making it another suitable elbow point.



Ultimately, the decision of the number of clusters depends on an arbitrary choice that considers these 2 scores and the data knowledge of the investigator. Given the prior I choose to model with n=4 clusters, since it has the highest Silhouette and is a reasonable elbow point.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

4. Results

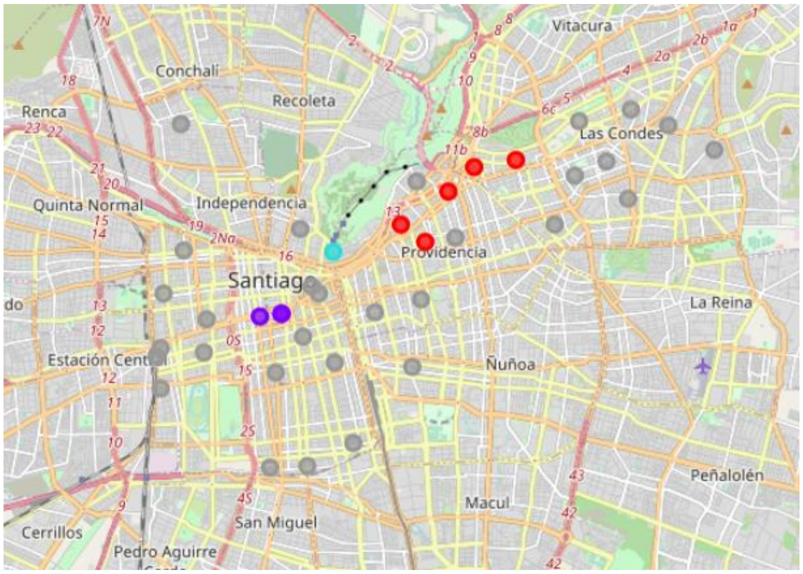
I obtained the following map, with each color representing a different cluster. The clusters obtained seams reasonable as they correspond to know neighborhoods in the city, such as business districts, the civic center or the discotheque/bar zone, and therefore we can tell our methodology is going in the right direction.

Gray Cluster: Represent the cluster with “everything else”, or the points that did not get classified into a particular cluster. Here I have the biggest improvement opportunity, since I recognize several different neighborhoods within this clusters that should be separated, such as the points near “Estacion Central” which is a commercial district, or the points near “Plaza Dignidad” which is a university-tourist district. A better dimensionality reduction could help me better discriminate these neighborhoods.

Purple Cluster: Represents the civic center, which contains the government palace, many government agencies, historic and religious buildings. The most common venues for this cluster are Coffee shops, Sandwich places, Pizza places and Bookstores.

Cian cluster: Barrio Bellavista, this point is an outlier and rightly so. It represents the neighborhood with the most nightclubs and bar density in the capital, being an icon of the Santiago nightlife. Its most common venues are Bars and Nightclubs.

Red cluster: Represent the business district. Most of the financial and engineering firms are in this area. It is the most uptown non strictly residential part of the city. Its most common venues are Coffee shops, Restaurants and Hotels.



5. Discussion

My analysis gives back reasonable clusters, which any Santiago resident can identify, with the advantage that these clusters can now be described quantitatively and not only qualitatively. Also, we can describe clearer borders of the different neighborhoods relying not on intuition but on the results of our algorithm.

On the other hand, we have 2 main problems. First, the lack of data for many neighborhoods which meant that I end up clustering only 32 of the 93 points I originally intended. I tested different values for minimum venues required (5, 10 and 15) but this did not seem to affect signicantly the clusters obtained and almost every neighborhood with few venues ended up classified in the gray “everything else” cluster, meaning that lowering the standard for admission do not necessarily bring value to the analysis. Second, there needs to be a better classification of the different venues that foursquare gives back, since there are many that are classified as different but in fact are very similar (such as Caffe and coffee shop). Since there are 219 types of venues, this work will be left for future iterations of my work.

If better dimensionality reduction is achieved, I will explore using db-scan algorithm instead of k-means, since the former can fit arbitrary shapes in the data and have a better management of outliers. Its downfall is that it does not behave well with high number of dimensions, which is the reason I choose k-means.

6. Conclusion

My study achieved reasonable results, which is a good indication, but there is a lot of work left to do, specially in dimensionality reduction and inclusion of new information, such as socio-economic, lifestyle quality, urban zoning or other that can contribute to make a finer analysis.

As a starting point it gives a useful and quantitative notion of some of the neighborhoods in the north-east sector of the capital which can be used for tourists, real state agents and many more.