Santiago de Chile neighborhood similarity

Beltrán Larraín Budge

May 8, 2020

Introduction

When the COVID-19 pandemic ends, a lot of people will feel the rush to go outside, use parks, eat at restaurants, check museums, go to bars with friends and meet people at coffee shops. The thing is, the most popular zones will experience a big influx of people, so getting a table in your favorite bar will be difficult on a Saturday night, or maybe you just want to try something new after the long time quarantined.

Santiago is the biggest city of Chile, there are many places where you could go, you just don't know them. If you can't go to Providencia, one the best places to go get a drink or grab a bite, you could search for similar neighborhoods to that one, so you can fulfill your plans and try something new, and trying new things is always exiting.

The scope of this project is the following: for each neighborhood in Santiago de Chile, check the most common type of places of interest, rank them by occurrence and place a marker on a map with that information. The next step is clustering the different neighborhoods into similar ones based on the type of the places of interest, that way you can choose a substitute based on what you want.

You like Neighborhood X because it has a lot of places to eat and cultural landmarks, but you do not want to go again to the same places you are used to? No problem, check what neighborhoods are similar and give them a try!

The Data

First, we need the Postal Codes of the neighborhoods of Santiago and we will scrap them from https://es.wikipedia.org/wiki/Anexo:C%C3%B3digos_postales_de_Chile, there is a table with every Postal Code of every neighborhood in Chile, so we will need to filter after the scrapping. Postal Codes between 7000000 and 8999999 correspond to neighborhoods inside the Metropolitan Area of Santiago, so we will use those.

Using geolocator library, we will get the latitude and longitude for every neighborhood and add them to the data set, that way we have the neighborhood with the associated coordinates.

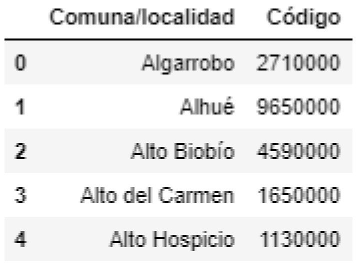
With the Foursquare API, the venues of interest will be requested and added to a new data frame, where they will be one hot encoded, grouped by neighborhood and standardized, so a clustering algorithm may be applied.

Methology

Preparing the data

After the data was scrapped, the data set contained the postal codes of every single neighborhood in the country, so a filter was necessary. In this case was easy, since Chilean postal codes are only numbers, and the ones corresponding to the metropolitan region of the city of Santiago are between 7000000 and 8999999.

Table 1: Result of the web scrapping

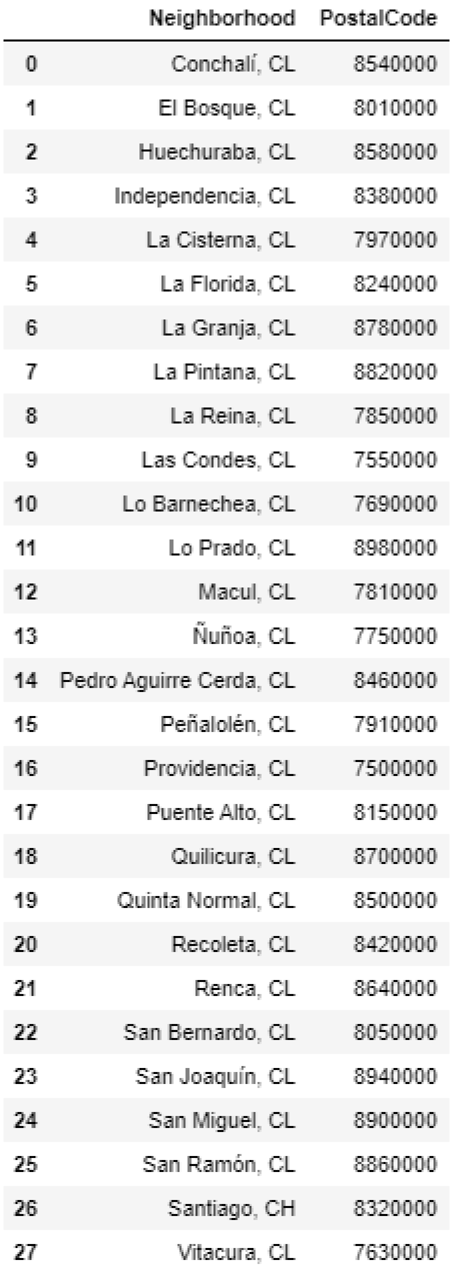


	Comuna/localidad	Código
0	Algarrobo	2710000
1	Alhué	9650000
2	Alto Biobío	4590000
3	Alto del Carmen	1650000
4	Alto Hospicio	1130000

The data must be transformed in a way that can interact with the different libraries, such as folium, so adding the latitude and longitude is necessary.

The Geolocator library was used to do so, given the neighborhood name, it returned the longitude and latitude, but first some adjustments to the data were done.

Table 2: Data set prepared for Geolocator

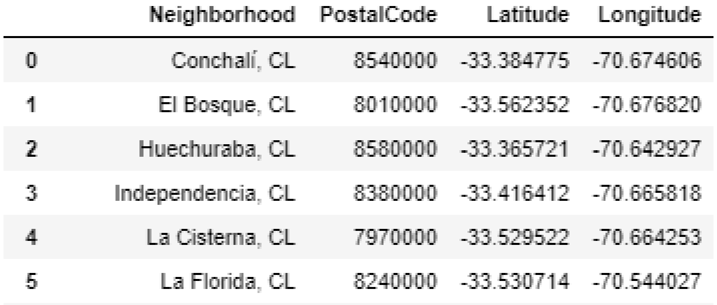


	Neighborhood	PostalCode
0	Conchalí, CL	8540000
1	El Bosque, CL	8010000
2	Huechuraba, CL	8580000
3	Independencia, CL	8380000
4	La Cisterna, CL	7970000
5	La Florida, CL	8240000
6	La Granja, CL	8780000
7	La Pintana, CL	8820000
8	La Reina, CL	7850000
9	Las Condes, CL	7550000
10	Lo Barnechea, CL	7690000
11	Lo Prado, CL	8980000
12	Macul, CL	7810000
13	Ñuñoa, CL	7750000
14	Pedro Aguirre Cerda, CL	8460000
15	Peñalolén, CL	7910000
16	Providencia, CL	7500000
17	Puente Alto, CL	8150000
18	Quilicura, CL	8700000
19	Quinta Normal, CL	8500000
20	Recoleta, CL	8420000
21	Renca, CL	8640000
22	San Bernardo, CL	8050000
23	San Joaquín, CL	8940000
24	San Miguel, CL	8900000
25	San Ramón, CL	8860000
26	Santiago, CH	8320000
27	Vitacura, CL	7630000

As shown in Table 2, there was a string ‘, CL’ added to every name of neighborhood, except for Santiago (neighborhood, not the city), where a ‘, CH’ was added, that is because the geolocator gives the correct coordinates of Santiago only when it is paired with the ‘, CH’ string, while the correct coordinates for the other neighborhoods are only given when paired with the ‘, CL’ string.

The first rows of the resulting data set look like this:

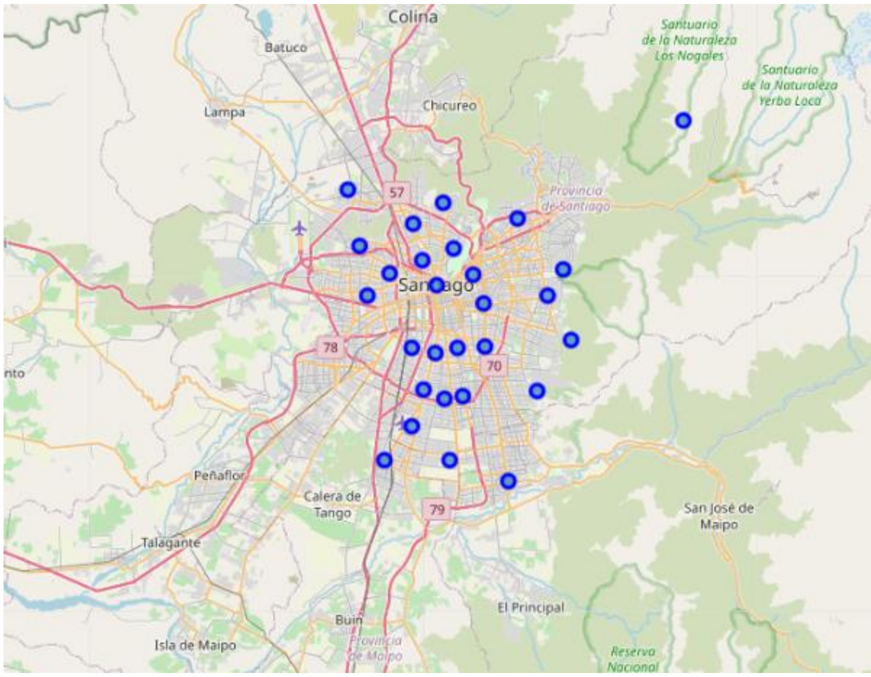
Table 3: Head of the data set with coordinates added



	Neighborhood	PostalCode	Latitude	Longitude
0	Conchalí, CL	8540000	-33.384775	-70.674606
1	El Bosque, CL	8010000	-33.562352	-70.676820
2	Huechuraba, CL	8580000	-33.365721	-70.642927
3	Independencia, CL	8380000	-33.416412	-70.665818
4	La Cisterna, CL	7970000	-33.529522	-70.664253
5	La Florida, CL	8240000	-33.530714	-70.544027

After mapping those coordinates into a folium map the result is not ok, since there are 4 neighborhoods that are shown outside or at the limit of the city. Las Condes, Lo Barnechea, Peñalolén and La Florida seem very far away, because those are incredibly huge neighborhoods they center will look like outside or on the border of the city. For example, Lo Barnechea reaches until the border with Argentina.

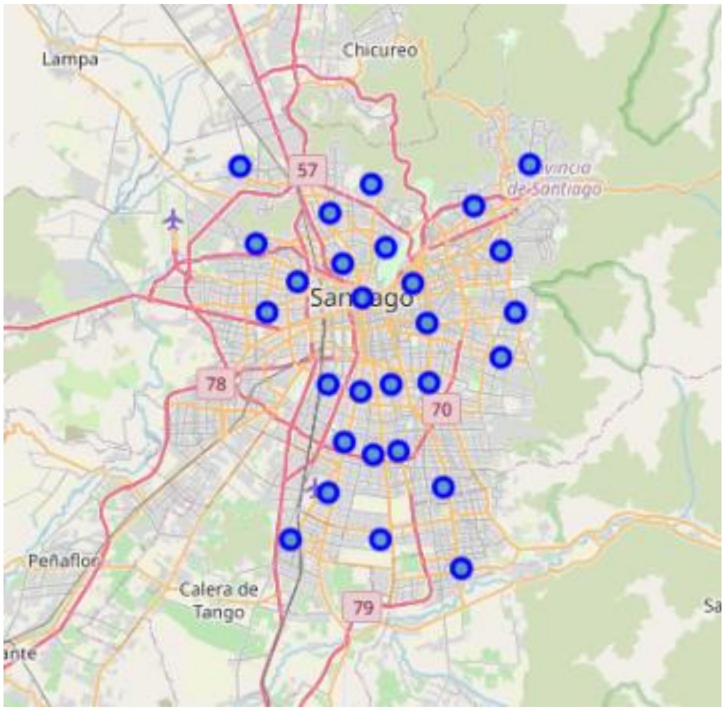
Image 1: First mapping of the Neighborhoods



To fix the issue, the new center for each neighborhood was considered as the coordinates of that neighborhood’s municipality building.

The resulting map has every neighborhood inside of the city.

Image 2: second mapping of the Neighborhoods



With the corrected locations the data was correctly prepared, the next step is getting venues from the Foursquare API.

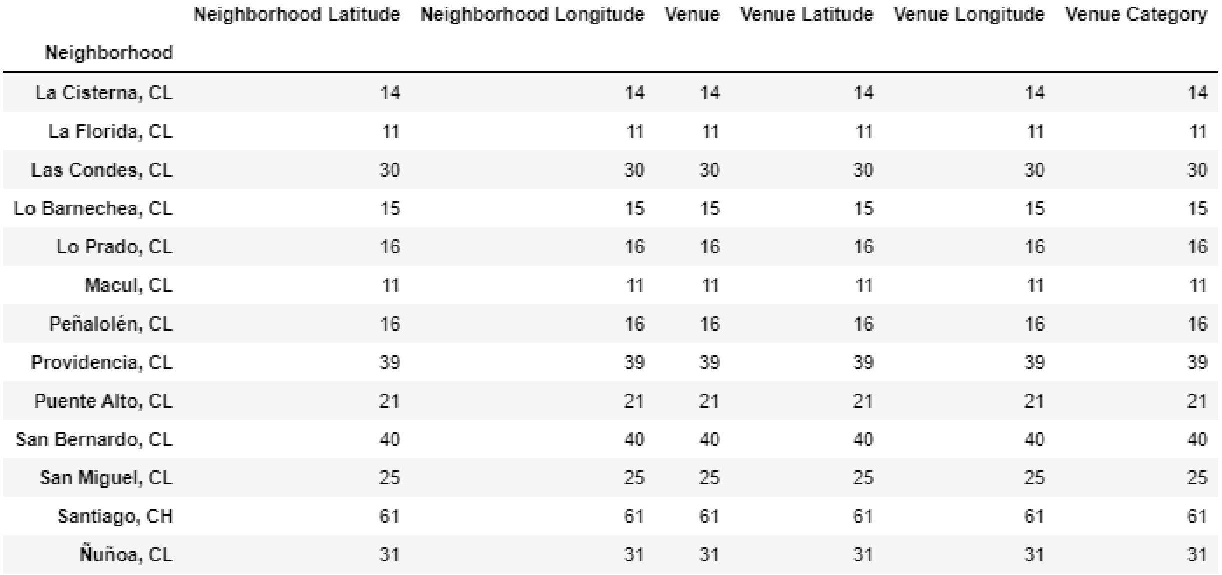
Foursquare API

Using a GET request, a new data set containing every nearby venue for each neighborhood was created.

After grouping by neighborhood and counting how many nearby venues there are for each there is a problem, some of them have no venues, other have 2 or 4. Since the information about venues in Santiago in Foursquare is so limited, the rows having inconsistent or incomplete data will be removed. Any neighborhood with less than 10 nearby venues will not be considered and deleted.

The remaining neighborhood, and the ones that will be considered for the analysis are the following:

Table 4: Counted nearby venues by Neighborhood

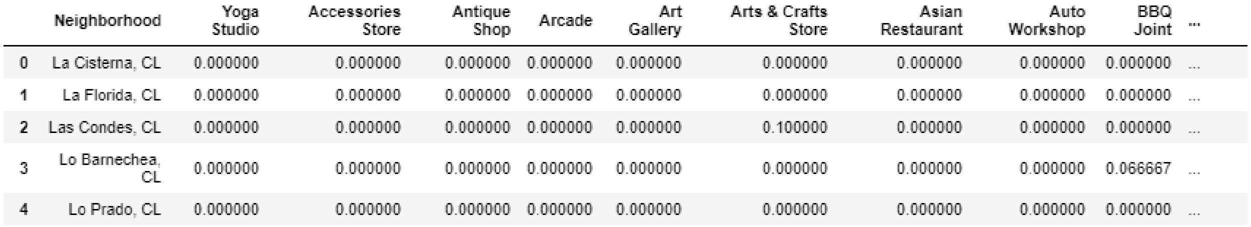


	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
La Cisterna, CL	14	14	14	14	14	14
La Florida, CL	11	11	11	11	11	11
Las Condes, CL	30	30	30	30	30	30
Lo Barnechea, CL	15	15	15	15	15	15
Lo Prado, CL	16	16	16	16	16	16
Macul, CL	11	11	11	11	11	11
Peñalolén, CL	16	16	16	16	16	16
Providencia, CL	39	39	39	39	39	39
Puente Alto, CL	21	21	21	21	21	21
San Bernardo, CL	40	40	40	40	40	40
San Miguel, CL	25	25	25	25	25	25
Santiago, CH	61	61	61	61	61	61
Ñuñoa, CL	31	31	31	31	31	31

One Hot Encoding

Every venue type was one hot encoded, and then they were grouped by neighborhood using the average method, that way we have the data standardized, which is mandatory for clustering.

Table 5: data frame with venue type one hot encoded

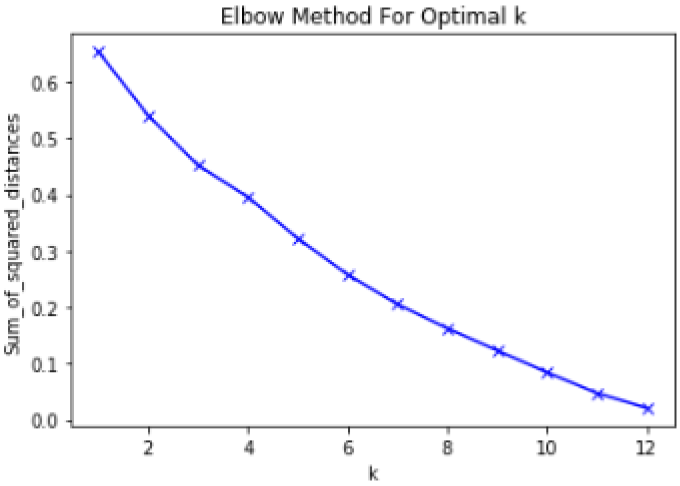


	Neighborhood	Yoga Studio	Accessories Store	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Auto Workshop	BBQ Joint	...
0	La Cisterna, CL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
1	La Florida, CL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
2	Las Condes, CL	0.000000	0.000000	0.000000	0.000000	0.000000	0.100000	0.000000	0.000000	0.000000	...
3	Lo Barnechea, CL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.066667	...
4	Lo Prado, CL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...

K-means clustering

To select the optimal K, the elbow method is used. That is calculating, for each K, the sum of squared distance error, and select the K where the decrease in error is not significant enough compared with the last K.

Graph 1: Sum of Squared Distances vs K

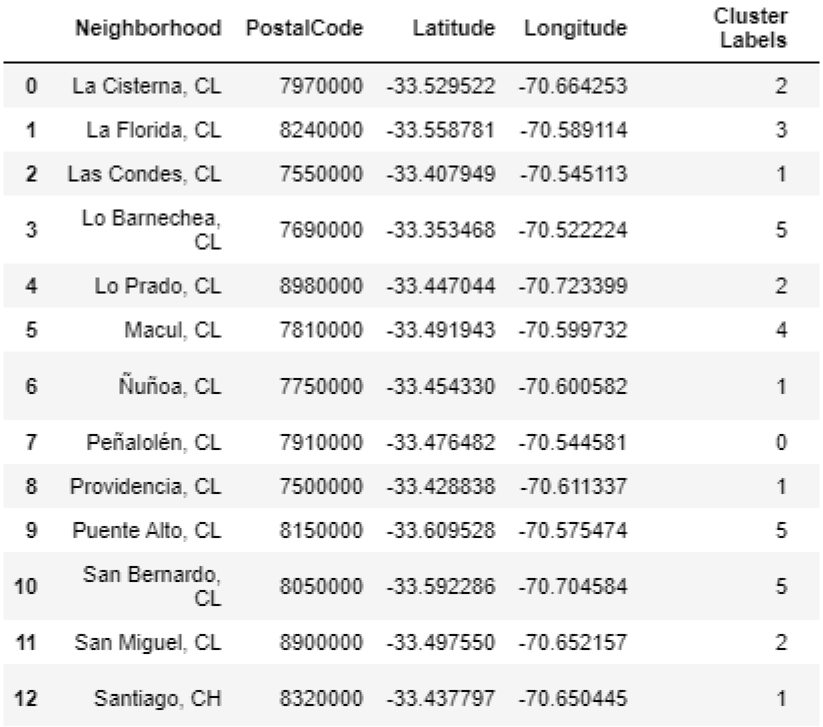


In this case, it's hard to see where that happens, since the graph is almost a straight line, but for this case it was considered that K=6 would yield the best result having in mind the graph and the characteristics of the data.

Results

The clusters assigned were matched with each neighborhood.

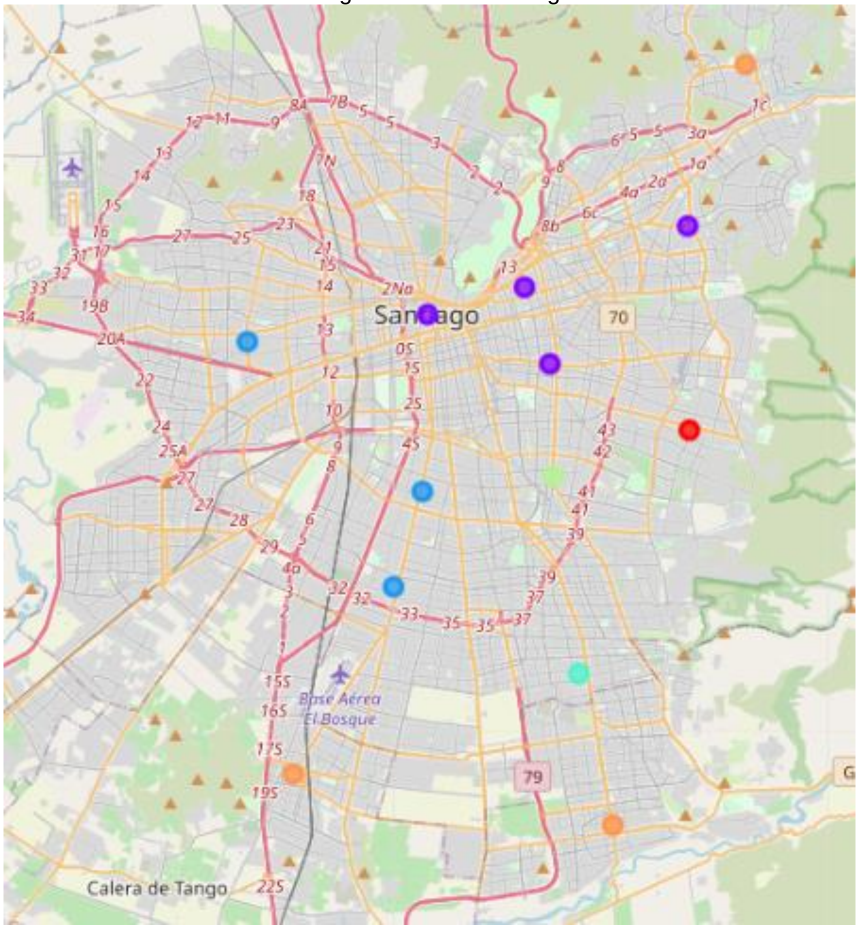
Table 6: Neighborhood, Postal Code, Coordinates and Cluster



	Neighborhood	PostalCode	Latitude	Longitude	Cluster Labels
0	La Cisterna, CL	7970000	-33.529522	-70.664253	2
1	La Florida, CL	8240000	-33.558781	-70.589114	3
2	Las Condes, CL	7550000	-33.407949	-70.545113	1
3	Lo Barnechea, CL	7690000	-33.353468	-70.522224	5
4	Lo Prado, CL	8980000	-33.447044	-70.723399	2
5	Macul, CL	7810000	-33.491943	-70.599732	4
6	Ñuñoa, CL	7750000	-33.454330	-70.600582	1
7	Peñalolén, CL	7910000	-33.476482	-70.544581	0
8	Providencia, CL	7500000	-33.428838	-70.611337	1
9	Puente Alto, CL	8150000	-33.609528	-70.575474	5
10	San Bernardo, CL	8050000	-33.592286	-70.704584	5
11	San Miguel, CL	8900000	-33.497550	-70.652157	2
12	Santiago, CH	8320000	-33.437797	-70.650445	1

With the results of the clustering, it’s time to create a Folium map to display them.

Image 2: Final clustering



Cluster 0 has only Peñalolén assigned.

Cluster 1 has Las Condes, Ñuñoa, Providencia and Santiago.

Cluster 2 has La Cisterna, Lo Prado and San Miguel.

Cluster 3 has La Florida.

Cluster 4 has Macul.

Cluster 5 has Lo Barnechea, Puente Alto and San Bernardo.

Discussion

Here we can that there are some relationships between some of those neighborhoods.

For example, cluster 1 has Las Condes, Ñuñoa, Providencia and Santiago, which the 4 are known for being the best places to go to restaurants, and the data shows that for all of them, the top 10 venue types are almost exclusively eating spots. This shows that the most concentration of restaurants is there.

Another good insight here is cluster 2, with La Cisterna, Lo Prado and San Miguel which are shown in the data as more rounded places with a lot to offer, there are places to eat but also a lot of pharmacies, bakeries, shops, offices and metro stations. This shows that their neighborhoods are less specialized and have many different venues to offer, making them great for living since you have almost anything nearby and the areas are less comercial.

Cluster 5 shows Lo Barnechea, Puente Alto and San Bernardo, the 3 most peripherical neighborhoods of the city, so having them in the same cluster makes a lot of sense. Since they are the most far away neighborhoods from the center, they need to be self-sufficient, and the data shows the greatest variety in venues, with gyms, jewelry stores, pet shops, outdoor activities, and plazas.

Peñalolén, La Florida and Macul are each in a separate cluster alone, that could be because they are transition neighborhoods, just between centric and peripheric.

Conclusion

In this study, I analyzed the different neighborhoods and nearby venues they have, to make comparisons and see if there were any similarities. I took into consideration the neighborhoods location (latitude and longitude) to request nearby venues from the Foursquare API and with that information use a clustering approach with K-means, since the data I used was unlabeled.

The conclusion discussed in the past section are a success, I was able to find similarities in different neighborhoods and explain them.

To answer the question proposed in the Introduction, after the quarantine, where will I go to grab a drink and a bite with my friends? Since Providencia, our favorite choice is likely to be overrun by people at that time, we will go for sure to Ñuñoa, Las Condes or Santiago, since they all have a very similar venue type distribution, just the ones we are looking for.