Capstone Project

“Santiago, grouping of neighborhoods for tourists”

By Nelson Ibarra Vergara

July 23, 2020

Table of contents

1. Introduction

1.1 Business Problem

Santiago de Chile is a growing city with many opportunities and a great place to visit. But as any other capital, for tourists with little time it could be confusing and complicated to choose where to go and when to go in order to optimize their time.

Regarding to that problem, in this project we will try to create options as recommendations for tourists. The idea is to help them to take decisions in base to characteristics of group of neighborhoods.

With this info the tourists could have more interesting data regarding to the concentration of venues that it could be possible to found in each group of neighborhoods. In addition, as it could be possible to find similar neighborhoods, the tourists could choose more easily in base to other variables like as close as possible from their hostel.

We will use our data science powers to generate a few most promising neighborhoods based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by tourists.

1.2 Data

Based on definition of the problem, factors that will influence our decision are:

- Venues of each neighborhood
- Category of venues
- Distance of neighborhood from city center

Following data sources will be needed to extract/generate the required information:

- Name of the best-known neighborhoods of Santiago, chile: those names will be obtained through scraping to Wikipedia.
- Geographical coordinates of each neighborhood: As there are not formal address to each neighborhood, the coordinates looking by names will be obtained using Google Maps API geocoding.
- Venues around each neighborhood will be obtained using Foursquare API.
- Categories of each venue will be obtained using Foursquare API.

1.3 Methodology

In this project I will limit the analysis considering that each neighborhood has a radius of 200mts from its center point (initially the idea was to consider 300mts but in the analysis it was found that several neighborhoods overlapped).

In first step we have collected the required data scraping Wikipedia and complementing through api query to google maps. Once both requests were done, we will have the name of neighborhoods and their coordinates.

Second step is to integrate the data with foursquare data calling to their api. The first call should get data regarding to venues around each location(neighborhood) and the second call should get the data of categories that foursquare use to tag each venue. As there are a lot of categories available, the decision was to take the most grouped set of categories, getting in this case a top 10 of categories.

In third and final step we will focus in understanding the integrated data using boxplot and maps tools. After that we will create clusters (using k-means clustering) of neighborhoods that meet some basic requirements established before in the business problem and we will explain the difference between each cluster using the most representative categories of each clusters.

2. Data acquisition and cleaning

2.1 Data sources

2.1.1 Wikipedia

Scraping the best known neighborhoods of Santiago (Capital of Chile) according to wikipedia the result is the following:

Neigh_names = ['Barrio Balmaceda', 'Barrio Bellas Artes', 'Barrio Bogotá','Barrio Brasil','Barrio Cívico de Santiago','Barrio Club Hípico','Barrio Concha y Toro','Barrio Copiapó','Barrio Cummings','Barrio Dieciocho','Barrio Ejército','Barrio Franklin','Barrio Huemul','Barrio Judicial','Barrio La Bolsa','Barrio Lastarria','Barrio Mapocho','Barrio Matta Sur', 'Barrio Meiggs', 'Barrio Panamá', 'Barrio París-Londres', 'Barrio Parque Almagro', 'Barrio Parque Forestal', 'Barrio Pedro Montt', 'Barrio República','Barrio San Borja', 'Barrio San Diego','Barrio San Eugenio','Barrio San Isidro','Barrio San Pablo', 'Barrio San Vicente', 'Barrio Santa Ana','Barrio Santa Elena','Barrio Santa Isabel','Barrio Santa Rosa','Barrio Sierra Bella','Barrio Viel','Barrio Yungay']

2.1.2 Google Maps

Making request to Google maps api in order to get the geographical coordinates of each neighborhood give us the following result:

Barrio Balmaceda, santiago, chile - [-33.431778, -70.672386]

Barrio Bellas Artes, santiago, chile - [-33.43634960000001, -70.64360409999999]

Barrio Bogotá, santiago, chile - [-33.46379, -70.62938]

Barrio Brasil, santiago, chile - [-33.4390218, -70.66692909999999]

Barrio Cívico de Santiago, santiago, chile - [-33.4450427, -70.654051]

Barrio Club Hípico, santiago, chile - [-33.46342, -70.66709999999999]

Barrio Concha y Toro, santiago, chile - [-33.446018, -70.6663377]

Barrio Copiapó, santiago, chile - [-33.4544896, -70.6442671]

Barrio Cummings, santiago, chile - [-33.4488897, -70.6692655]

Barrio Dieciocho, santiago, chile - [-33.448326, -70.65823]

Barrio Ejército, santiago, chile - [-33.44909, -70.66244499999999]

Barrio Franklin, santiago, chile - [-33.4728669, -70.6420638]

Barrio Huemul, santiago, chile - [-33.474624, -70.65112979999999]

Barrio Judicial, santiago, chile - [-33.4488897, -70.6692655]

Barrio La Bolsa, santiago, chile - [-33.4424145, -70.65161499999999]

Barrio Lastarria, santiago, chile - [-33.4378263, -70.639476]

Barrio Mapocho, santiago, chile - [-33.4488897, -70.6692655]

Barrio Matta Sur, santiago, chile - [-33.4627623, -70.6342324]

Barrio Meiggs, santiago, chile - [-33.451482, -70.6774673]

Barrio Panamá, santiago, chile - [-33.4580796, -70.62570629999999]

Barrio París-Londres, santiago, chile - [-33.4448446, -70.6484199]

Barrio Parque Almagro, santiago, chile - [-33.4521121, -70.6535998]

Barrio Parque Forestal, santiago, chile - [-33.4356565, -70.6412692]

Barrio Pedro Montt, santiago, chile - [-33.4732494, -70.65957879999999]

Barrio República, santiago, chile - [-33.4529256, -70.6686261]

Barrio San Borja, santiago, chile - [-33.4590693, -70.6807922]

Barrio San Diego, santiago, chile - [-33.45413, -70.64739999999999]

Barrio San Eugenio, santiago, chile - [-33.47018, -70.67327]

Barrio San Isidro, santiago, chile - [-33.4502812, -70.6429674]

Barrio San Pablo, santiago, chile - [-33.4431667, -70.7157647]

Barrio San Vicente, santiago, chile - [-33.46264, -70.67519]

Barrio Santa Ana, santiago, chile - [None, None]

Barrio Santa Elena, santiago, chile - [-33.47789, -70.62608]

Barrio Santa Isabel, santiago, chile - [-33.4449679, -70.6253737]

Barrio Santa Rosa, santiago, chile - [-33.4448446, -70.6484199]

Barrio Sierra Bella, santiago, chile - [-33.4722601, -70.6329839]

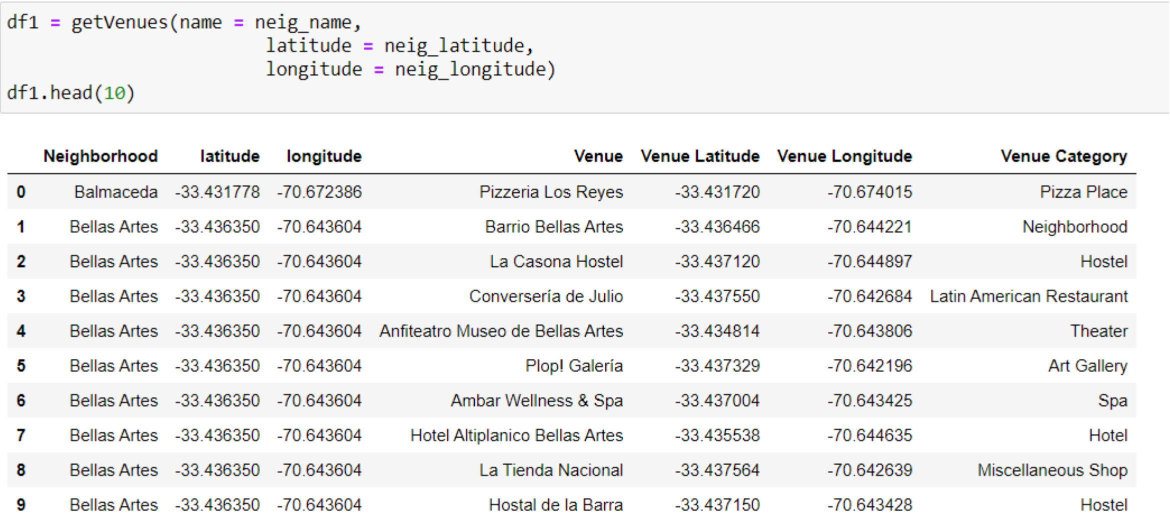
Barrio Viel, santiago, chile - [-33.465534, -70.65556]

Barrio Yungay, santiago, chile - [-33.4429112, -70.6731231]

2.1.3 Foursquare

Two different requests are necessary to make to foursquare. The first one will be the request the top 50 venues surrounding each neighborhood. The second request ask for getting all the categories in which foursquare classifies the venues in order to look the most aggregated set of categories. The results are as following:

Venues for each neighborhood:



```
df1 = getVenues(name = neig_name,
                        latitude = neig_latitude,
                        longitude = neig_longitude)
df1.head(10)
```

	Neighborhood	latitude	longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Balmaceda	-33.431778	-70.672386	Pizzeria Los Reyes	-33.431720	-70.674015	Pizza Place
1	Bellas Artes	-33.436350	-70.643604	Barrio Bellas Artes	-33.436466	-70.644221	Neighborhood
2	Bellas Artes	-33.436350	-70.643604	La Casona Hostel	-33.437120	-70.644897	Hostel
3	Bellas Artes	-33.436350	-70.643604	Conversería de Julio	-33.437550	-70.642684	Latin American Restaurant
4	Bellas Artes	-33.436350	-70.643604	Anfiteatro Museo de Bellas Artes	-33.434814	-70.643806	Theater
5	Bellas Artes	-33.436350	-70.643604	Plop! Galería	-33.437329	-70.642196	Art Gallery
6	Bellas Artes	-33.436350	-70.643604	Ambar Wellness & Spa	-33.437004	-70.643425	Spa
7	Bellas Artes	-33.436350	-70.643604	Hotel Altiplanico Bellas Artes	-33.435538	-70.644635	Hotel
8	Bellas Artes	-33.436350	-70.643604	La Tienda Nacional	-33.437564	-70.642639	Miscellaneous Shop
9	Bellas Artes	-33.436350	-70.643604	Hostal de la Barra	-33.437150	-70.643428	Hostel

10 top level categories:

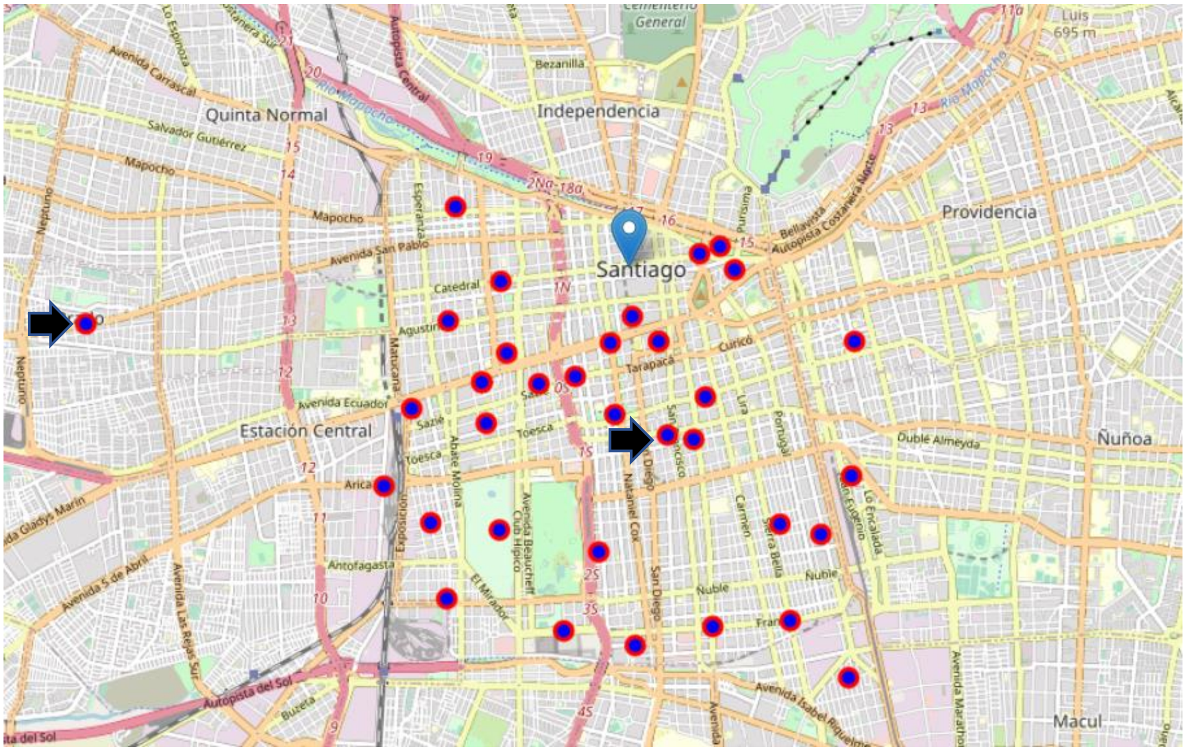
- Arts & Entertainment
- College & University
- Event
- Food
- Nightlife Spot
- Outdoors & Recreation
- Professional & Other Places
- Residence
- Shop & Service
- Travel & Transport

All the foursquare categories can be seen here:

https://developer.foursquare.com/docs/build-with-foursquare/categories

2.2 Data cleaning

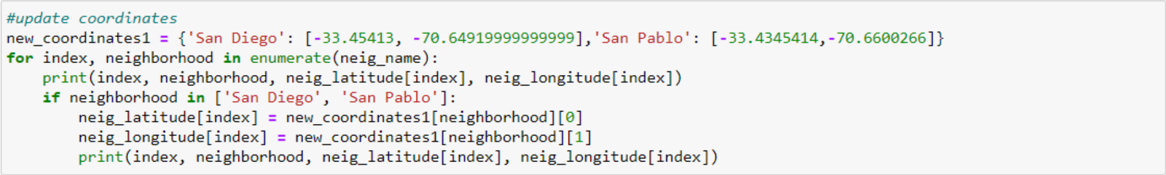
To be able to see how the data is working, I would rather check it through plotting a map with the coordinates of each neighborhood.



After doing a zoom it is possible to see two point (black arrow) that are not correctly assigned. These points are the neighborhood of ‘San Pablo’ and ‘San Diego’.

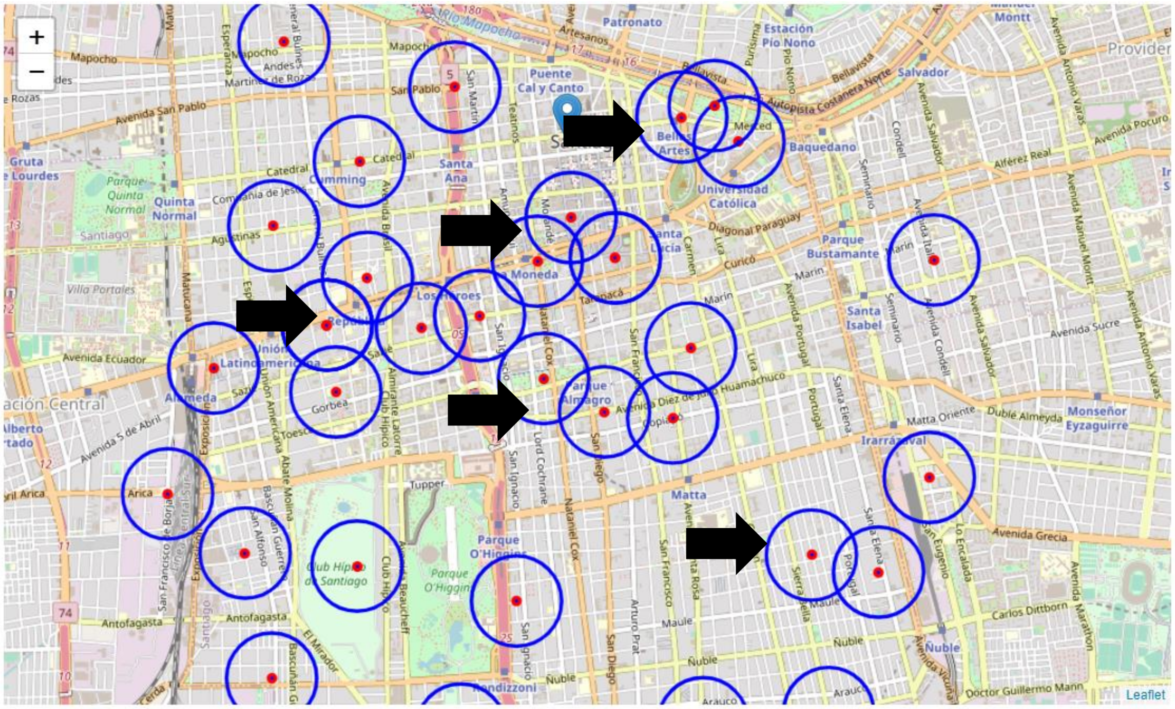
To amend it, I have used google maps web-app to get both, latitude, and longitude values for each neighborhood.

The code to update the coordinates were the following:

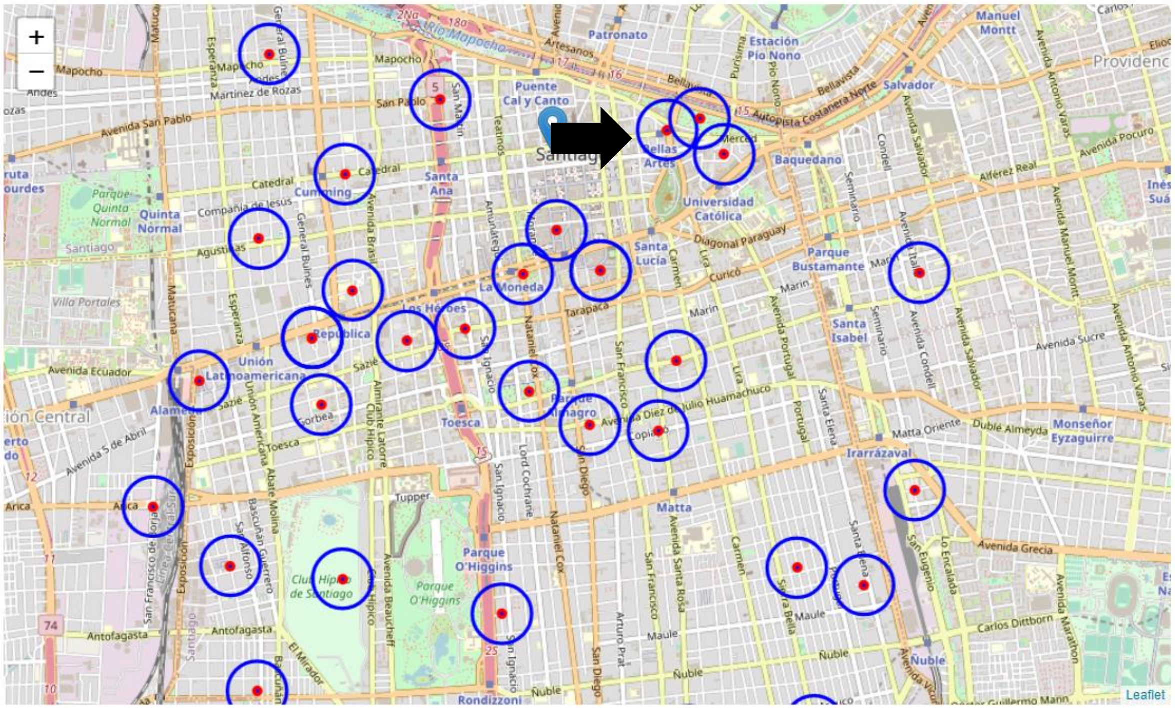


```
#update coordinates
new_coordinates1 = {'San Diego': [-33.45413, -70.64919999999999],'San Pablo': [-33.4345414,-70.6600266]}
for index, neighborhood in enumerate(neig_name):
    print(index, neighborhood, neig_latitude[index], neig_longitude[index])
    if neighborhood in ['San Diego', 'San Pablo']:
        neig_latitude[index] = new_coordinates1[neighborhood][0]
        neig_longitude[index] = new_coordinates1[neighborhood][1]
        print(index, neighborhood, neig_latitude[index], neig_longitude[index])
```

As I was thinking about to use a radius of 300 meters to look for venues inside of each neighborhood, I wanted to check whether it could be possible. Therefore I plotted the following map simulating a radius of 300 meters from the coordinates captured for each neighborhood.

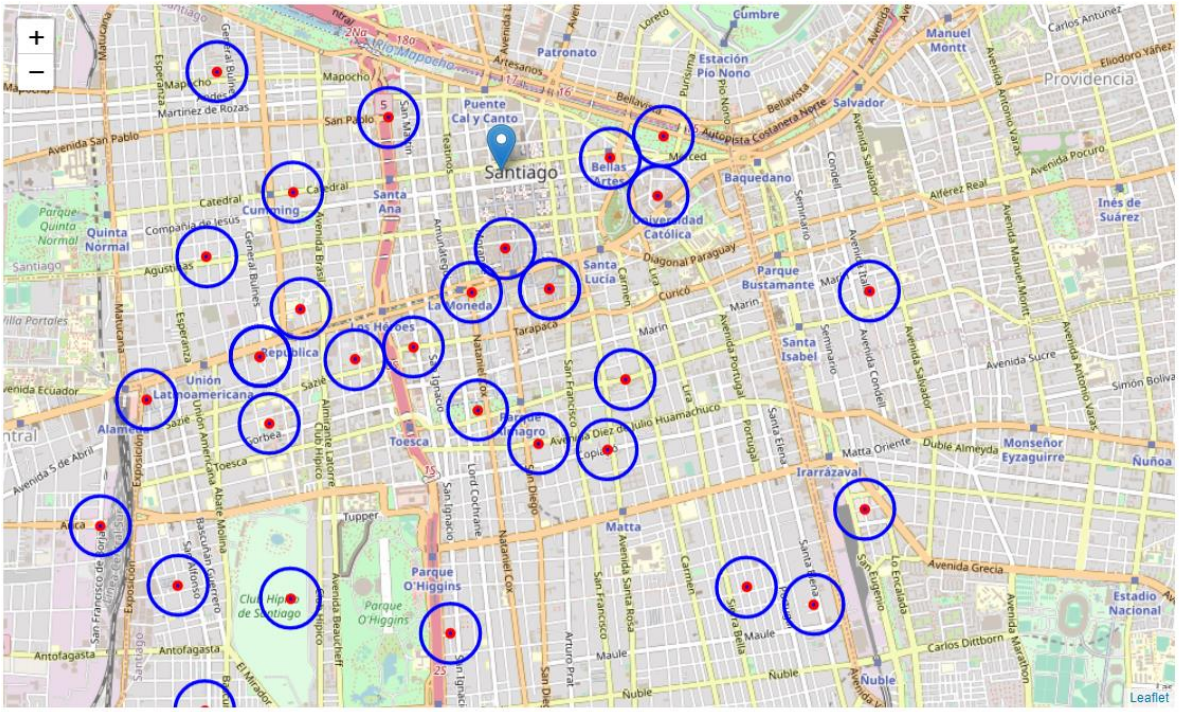


With that, I had to change my mind and use a radius of 200 meters.



After plotting a 200 meters radius, the only problem that I got was the points showing by the black arrow. To solve it, I used the same procedure before, and I amended the three point to fit better (Bellas artes, Lastarria and Parque forestall).

Finally, the result of cleaning data was as follow:



2.3 Feature selection

The first neighborhood to take off from the Wikipedia list was “Barrio Santa Ana”, this because I was not able to get the coordinates from google maps api and google maps web-app.

After the cleaning and wrangling the data, I finally got 31 neighborhoods, all of them will be use to analyze.

Regarding to the features to used, these are the following:

- Neighborhood
- Latitude
- Longitude
- Arts & Entertainment (category)
- Food (category)
- Nightlife Spot (category)
- Outdoors & Recreation (category)
- Professional & Other Places (category)
- Shop & Service Travel & Transport (category)

The others top 10 categories like College & University, Event and Residence were no found as part of the neighborhoods.

3. Exploratory Data Analysis

As the data of geographical coordinates was analyzed while it was cleaned and wrangled, the data captured from foursquare according to the top 50 of venues for each point, need more analyses since the venues categories were grouped into main categories to do more easy the data analysis.

The tool utilized to do this work is a box plot due to this tool is a type of chart often used in explanatory data analysis to visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.

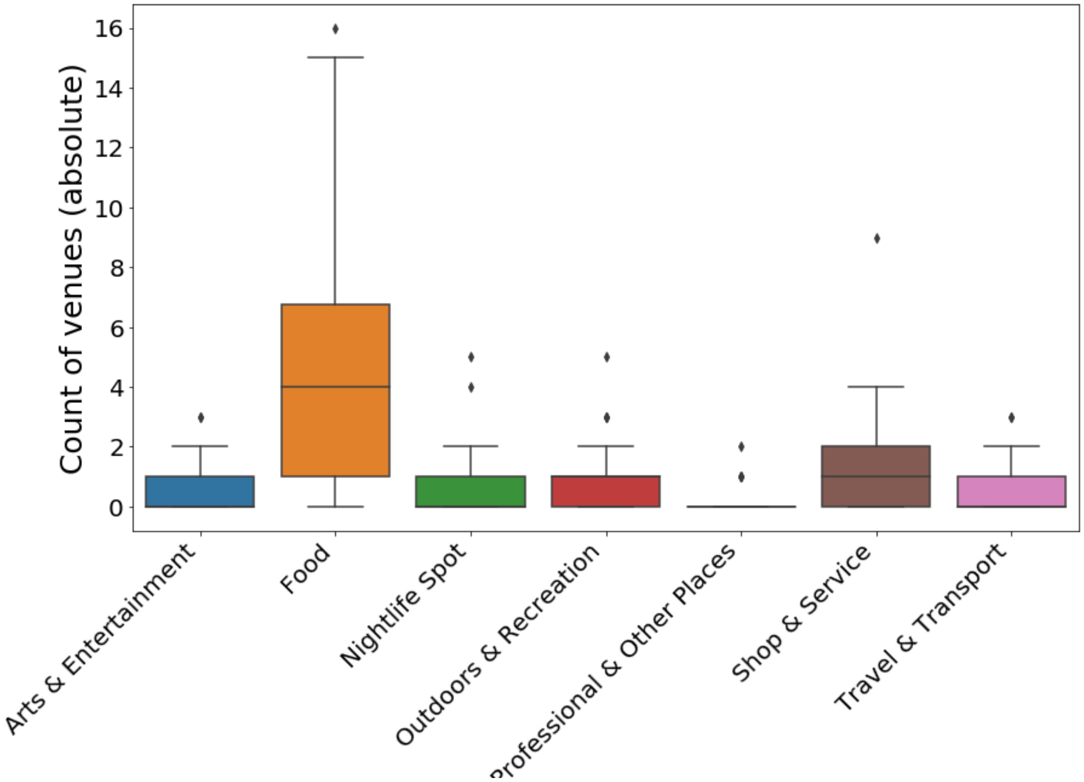
Some advantages of using boxplot to analyze are:

- Box plots are useful as they show the average score of a data set.
- Box plots are useful as they show the skewness of a data set.
- Box plots are useful as they show the dispersion of a data set.
- Box plots are useful as they show outliers within a data set.

The five-number summary of a set of data showing in box plot are: minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.

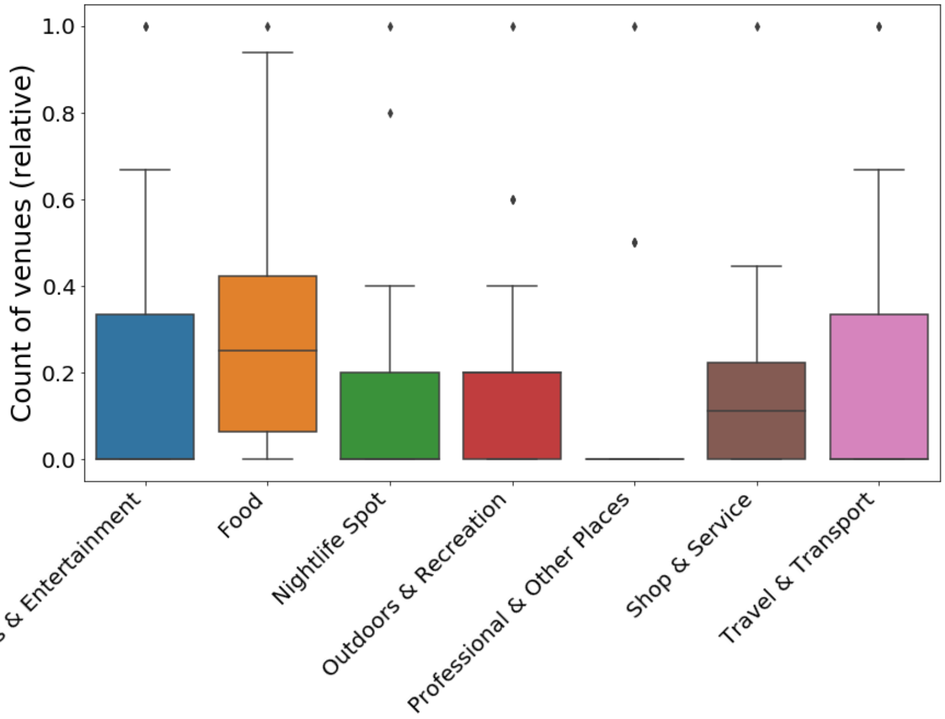
3.1 Venues by category (without normalization)

Check the distributions of venues by main categories without normalization (absolute values).



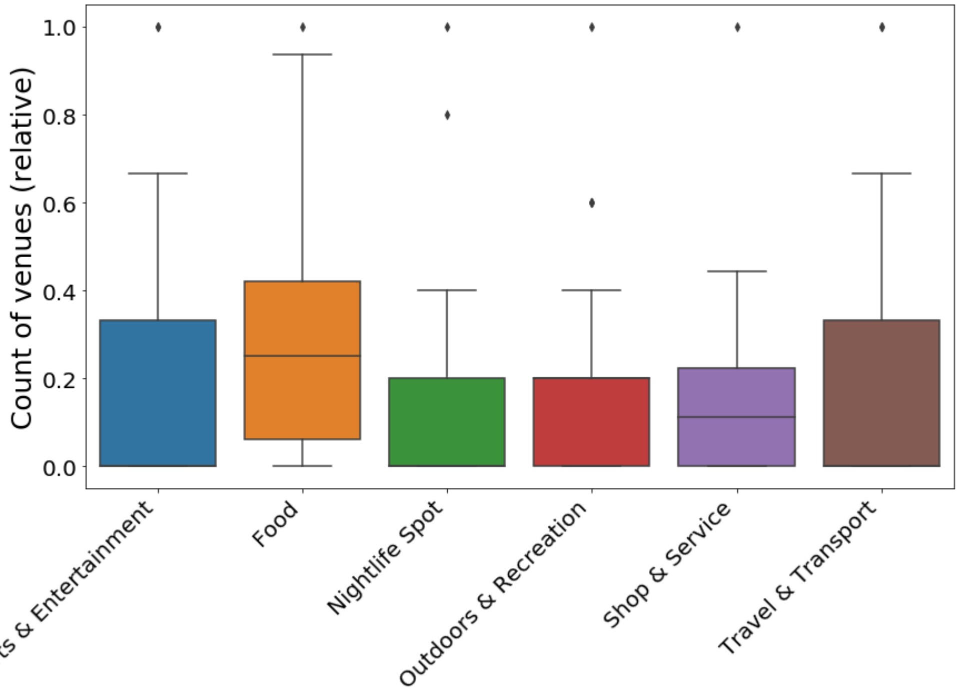
Too many venues related to food than others categories and very poor venues related to Professional & other places. The neighborhoods are very concentrated.

3.2 Venues by category (with normalization)



As it was expected, the conclusions do not change too much but the data fit better to be analyzed.

3.3 Venues by category final version (with normalization)

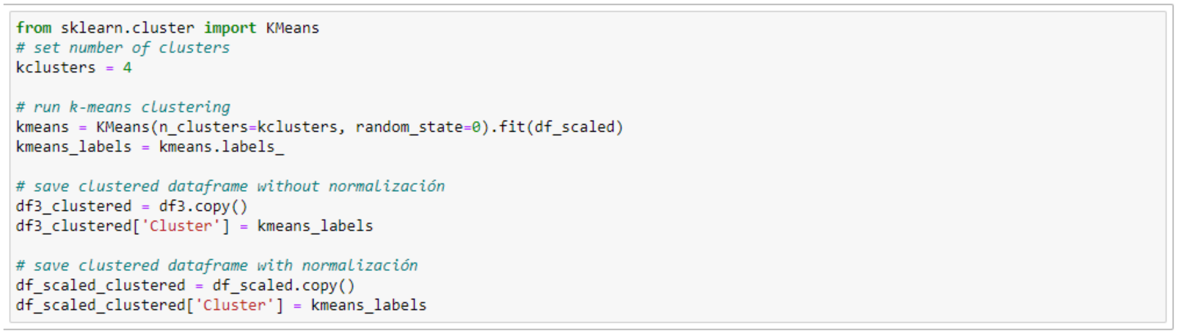


The final version of data analyzed take off the feature “Professional & other places” due this is not data useful for the next step.

4. Clustering

The following code was used to create data frames that cluster the data captured before using the method named Kmeans with examples of 1, 2, 3, 4 and 5 clusters.

The code utilized was the following.



```
from sklearn.cluster import KMeans
# set number of clusters
kclusters = 4

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(df_scaled)
kmeans_labels = kmeans.labels_

# save clustered dataframe without normalización
df3_clustered = df3.copy()
df3_clustered['Cluster'] = kmeans_labels

# save clustered dataframe with normalización
df_scaled_clustered = df_scaled.copy()
df_scaled_clustered['Cluster'] = kmeans_labels
```

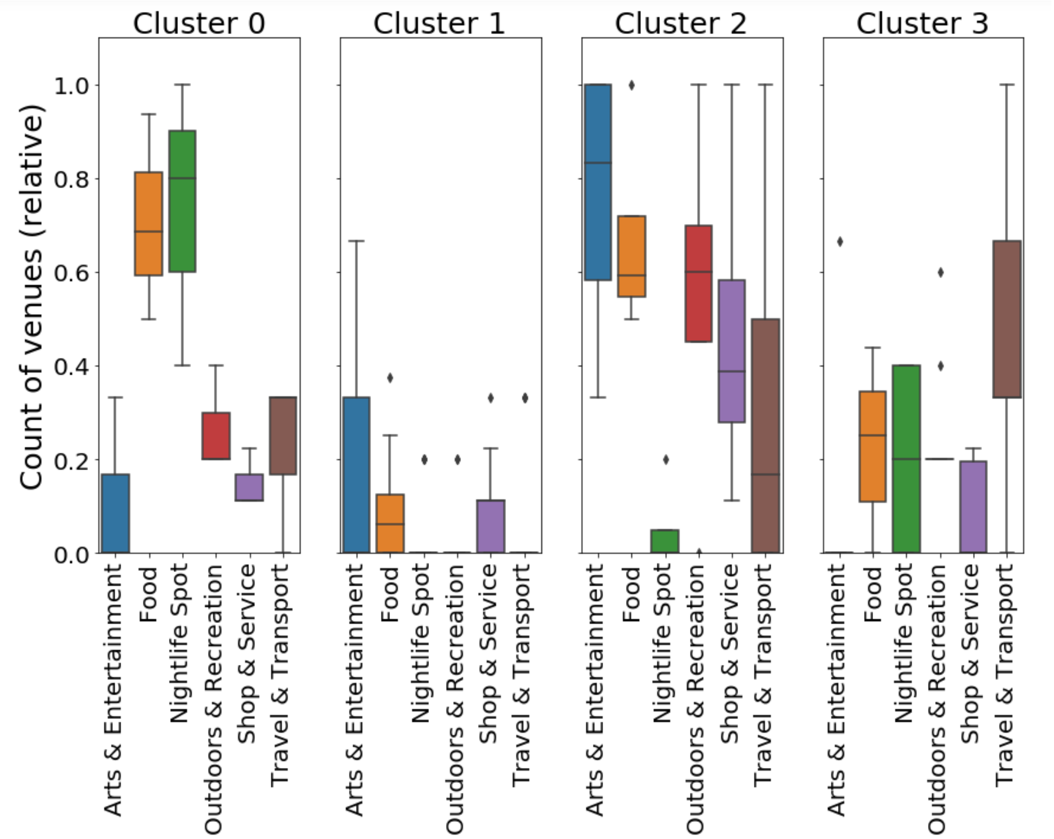
The preliminary results with different number of clusters as follow:

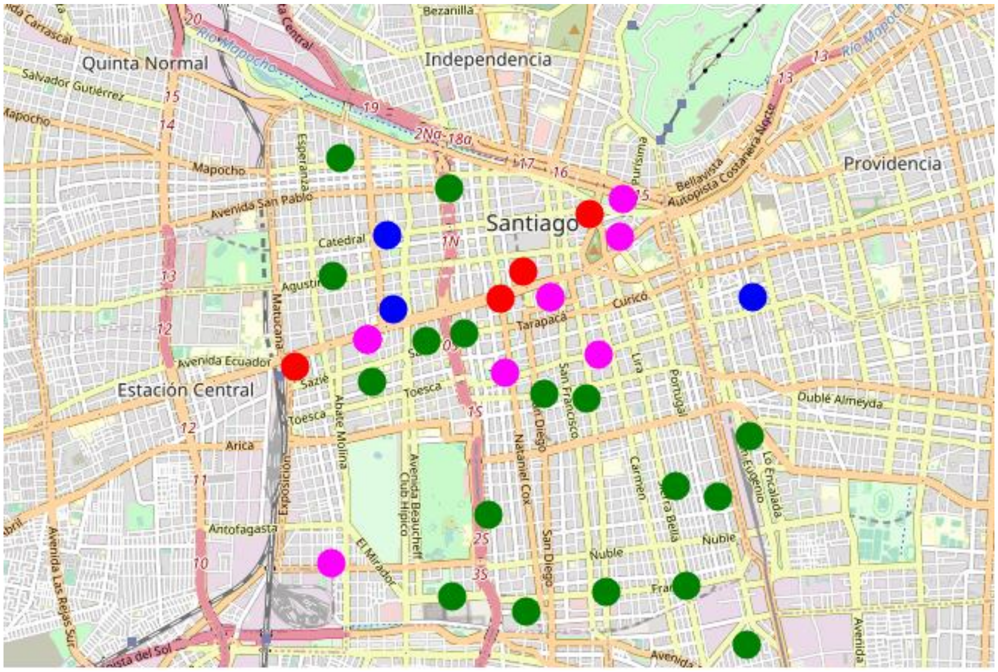
- 2 clusters: Only show the difference between neighborhood with many venues vs neighborhood with low venues (ex. plaza).
- 3 clusters: Decompose the low venues neighborhood in 2 clusters and keep the high neighborhood with many venues.
- 4 clusters: The most miscellaneous and representative clusters. Showing diversity and concentration.
- 5 clusters: Difficult to interpret.

So, for the final analysis I settle on 4 clusters the results and conclusions.

5. Results

To do a good interpretation of the data I would plot both, boxplot of clusters and maps of cluster.





Once the visualizations shows the results, it is easy to see 4 cluster with enough differences and good distribution of neighborhood for each group.

My interpretation in order to make recommendations to tourist are the following :

Cluster 1: Neighborhoods as nice spot to go out for food and bars (nightlife).

Cluster 2: Neighborhoods with low venues, their spots (few) are most related to arts & ei, food and shop and service.

Cluster 3: Most diverse Neighborhoods, great for arts & ei, food, outdoor & recreation, shop & service, and travel & transport.

Cluster 4: Neighborhoods with travel and transport to any point of the city, 2nd place as a nightlife spot, some food and shop & service

5.1 Discussions

From my point of view, there are two point that could change the results and analysis:

- Foursquare did not have too much success in Chile, therefore there are not enough venues assigned to each neighborhood. Anyway, as the assignment was asking to use foursquare-api and I wanted to explore my city I continue it despite of lack of some data.
- Using the same radius space for each neighborhood do not add precision to the analysis because each neighborhood have different distribution of it. With the boundaries of each neighborhood the analysis could improve (not available).

6. Conclusions

Despite of the difficulty found related to get info of Chile using foursquare api, the results make a lot of sense for me. I think that without doing a k test to find the best "k", just doing a manually k-testing from 2 to 5 k's, each result was easy to interpret. Finally, the object of this work is completed, So if you came to Santiago (Chile), you could use this work to see how many options do you have according to your interest, time and money.

The options are as follows:

Cluster #1 : Neighborhoods as nice spot to go out for food and bars (nightlife). 3 Neighborhoods

Cluster #2 : Neighborhoods with low venues, their spots (few) are most related to arts & ei, food and shop and service. 17 Neighborhoods

Cluster #3 : Most diverse Neighborhoods, great for arts & ei, food, outdoor & recreation, shop & service and travel & transport. 4 Neighborhoods

Cluster #4 : Neighborhoods with travel and transport to any point of the city, 2nd place as a nightlife spot, some food and shop & service. 7 Neighborhoods