Analyzing segregation in Santiago, Chile

Vicente Soto Armijo

June 16, 2019

1. Introduction

1.1 Background: Chile is a South America country, with a population of 18 million. The capital, Santiago, represents 1/3 of its population. Due to multiple factors (economics, politics, among others), nowadays it is really segregated by its geography, where a few neighborhoods concentrate medium-high and high socioeconomic sectors. Although multiple studies have been made to highlight this phenomenon, it is highly necessary to keep digging and find insights, patterns, or analysis with the objective of develop public initiatives that can be done to mitigate this situation

1.2 Problem: Santiago's subway system is used by almost every citizen every day, so the movements and disposition of every station is probably a decisive factor of segregation. This project aims to cluster the activities that take place around each station to highlight the differences between lifestyles only using location as input to the model.

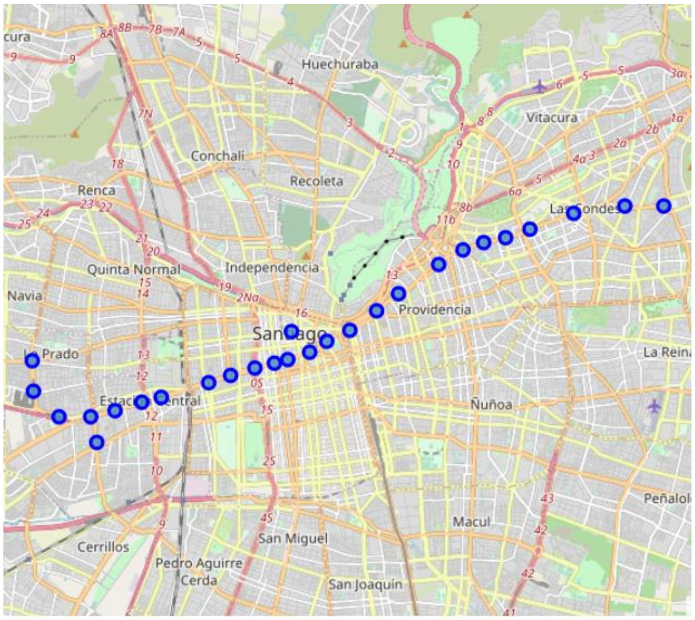
2.2 Interest: in Chile, public sector rarely takes advantage of technologies such as Machine Learning to discover patterns and insights. The realization of this project aims to be a first step to develop more general analysis to improve public sector decision-making.

2. Data acquisition and cleaning

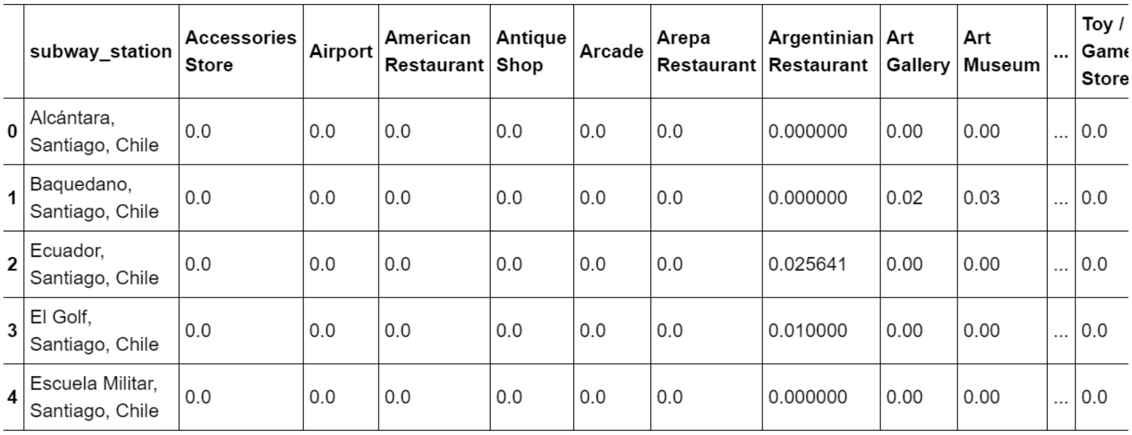
2.1 Data Sources: in the present, extracting data directly from a public repository about subway stations is not possible, mostly because the inexistence of a repository with this information. The datasets used in this project are extracted using the list of stations in Wikipedia (link here) as search input to Geopy package (in Python) and the Foursquare API.

2.2 Data Cleaning: there are multiple routes defined in the subway system. This project focuses in one, "Linea 1" (Spanish translation of Route 1), because the route defined by their corresponding stations are in high-income zones and low-income zones too. The algorithm was adjusted iteratively to ensure every station in this specific route is well defined and accurate in position.

The following map, generated by the data extracted and cleaned, represents the stations that are going to be analyzed:

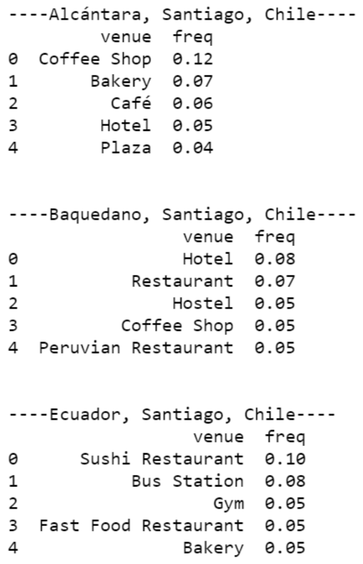


2.3 Feature Selection: the dataset, after extract and process the data from the mentioned data sources, is represented by the following table:



	subway_station	Accessories Store	Airport	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	...	Toy / Game Store
0	Alcántara, Santiago, Chile	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00	0.00	...	0.0
1	Baquedano, Santiago, Chile	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.02	0.03	...	0.0
2	Ecuador, Santiago, Chile	0.0	0.0	0.0	0.0	0.0	0.0	0.025641	0.00	0.00	...	0.0
3	El Golf, Santiago, Chile	0.0	0.0	0.0	0.0	0.0	0.0	0.010000	0.00	0.00	...	0.0
4	Escuela Militar, Santiago, Chile	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.00	0.00	...	0.0

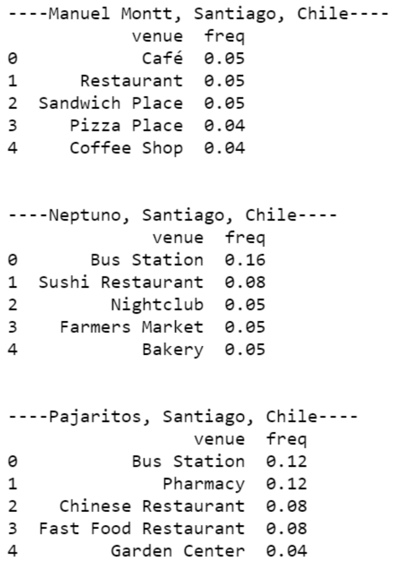
where 217 columns represent every possible venue from all the subway stations in a radius of 500 meters from each one. With the proper transformation, we can extract the information shown in the tables below, where for every metro station, the top 5 most present venue category is displayed. This information is then used for the clustering process



```
----Alcántara, Santiago, Chile----
         venue  freq
0  Coffee Shop  0.12
1       Bakery  0.07
2         Café  0.06
3        Hotel  0.05
4        Plaza  0.04


----Baquedano, Santiago, Chile----
                venue  freq
0               Hotel  0.08
1          Restaurant  0.07
2              Hostel  0.05
3         Coffee Shop  0.05
4  Peruvian Restaurant  0.05


----Ecuador, Santiago, Chile----
                 venue  freq
0      Sushi Restaurant  0.10
1           Bus Station  0.08
2                   Gym  0.05
3  Fast Food Restaurant  0.05
4                Bakery  0.05
```



```
----Manuel Montt, Santiago, Chile----
            venue  freq
0            Café  0.05
1      Restaurant  0.05
2  Sandwich Place  0.05
3     Pizza Place  0.04
4     Coffee Shop  0.04


----Neptuno, Santiago, Chile----
              venue  freq
0       Bus Station  0.16
1  Sushi Restaurant  0.08
2         Nightclub  0.05
3    Farmers Market  0.05
4            Bakery  0.05


----Pajaritos, Santiago, Chile----
                  venue  freq
0           Bus Station  0.12
1              Pharmacy  0.12
2    Chinese Restaurant  0.08
3  Fast Food Restaurant  0.08
4        Garden Center  0.04
```

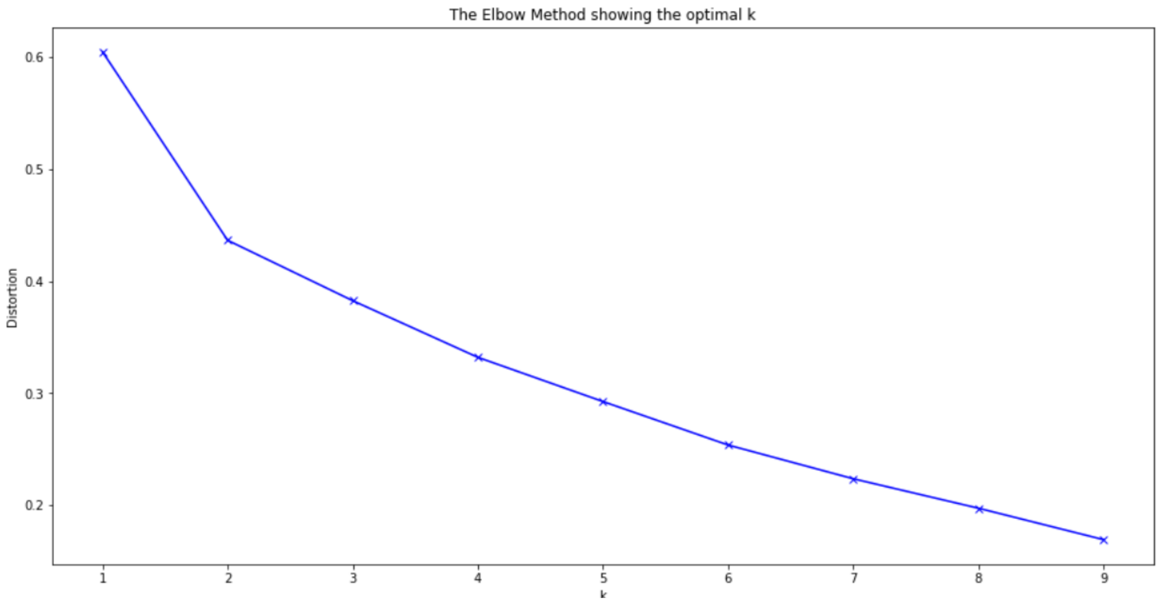
3. Clustering Modeling

3.1 K-Means Clustering: due to the nature of the problem and the few number of stations analyzed, k-means was selected. Mostly, because the scaled nature of the dataset and to ensure that every point is grouped in a cluster (this is not ensured in algorithms like density-based ones).

3.2 Number of Clusters selection: using the easy-to-implement elbow method we can decide the number of clusters to use. Is a good exercise to, first, analyze the number of clusters that could make sense.

As stated in 3.1, the problem needs to cluster a reduced amount of subway stations (27 to be more precise). So, it does not make much sense to use more than 3 clusters, to ensure that every cluster the necessary number of clusters to be able to define a well-defined characteristic.

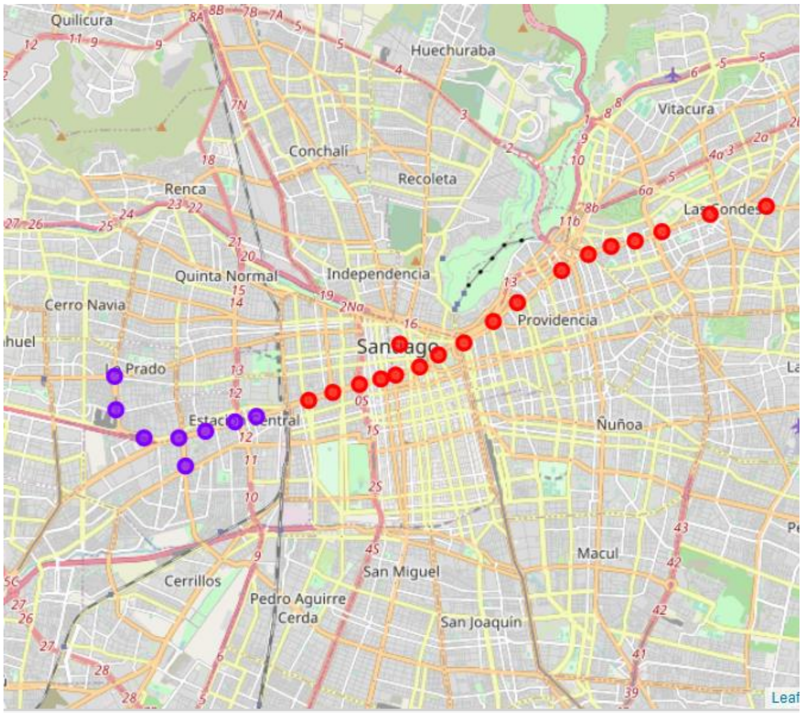
The elbow method is illustrated in the graph below.



In accordance of the previous statement, is clear that 2-3 clusters can be appropriate. For this exercise and because using the knowledge of the author about socioeconomic behaviors in Santiago, 2 clusters are going to be used.

4. Results and Conclusion

4.1 Results: a map with the resulting clusters is presented in the following figure:



- Red Cluster: Coffee Shops, Hotels, Sushi restaurants. Geographic sector: business district and high-income families.
- Purple Cluster: Pharmacies, Fast Food restaurants, Sushi restaurants. Geographic sector: small businesses and low-income families

The results show exactly the hypothesis that was stated. The activities (venues) are clearly defined by geography, phenomenon attributed to the socioeconomic sectors of Santiago. The “barrier” between red and purple dots it’s a well known sector where low income sectors are the majority (if not all) of the population.