How different are the metro stations in the city?

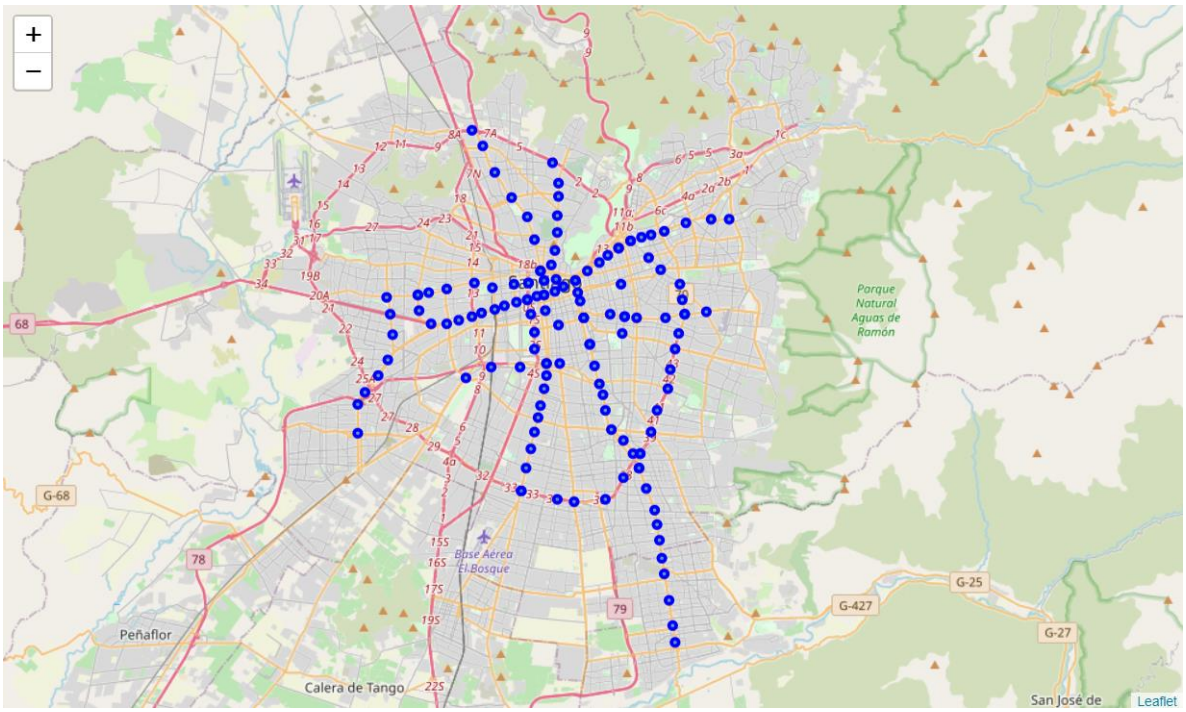
I born and raise in in Mérida, a little city in Venezuela, when I grow up I moved to Caracas and always preferred to travel in the city using the metro, I barely used the bus to get to my destiny. I noticed that outside of many stations looks similarly. Now I live in Santiago de Chile and I still use a lot the metro and still notice the similarities between some stations.

Knowing which are the similar stations it could be used to:

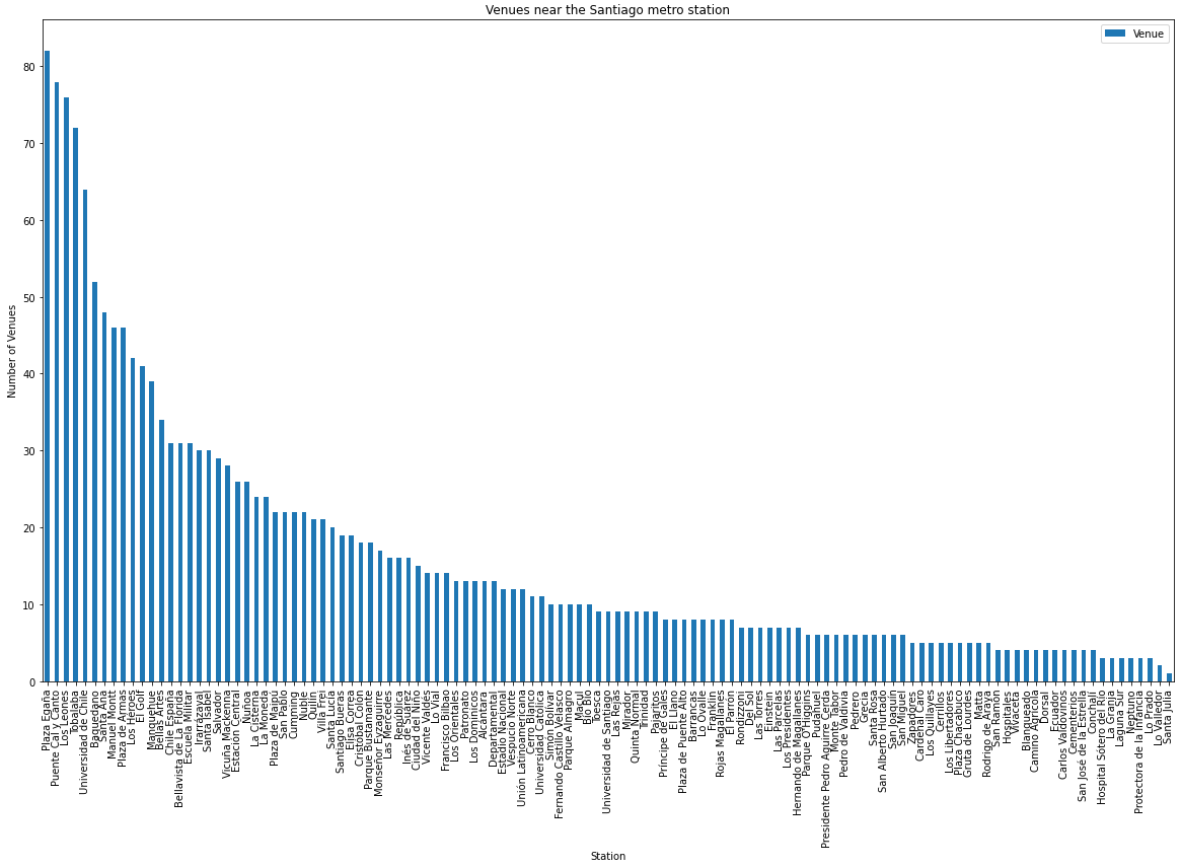
- Establish a profile for the people who use the metro station, with this it could offer a publicity campaign specific for each kind of station.
- Establish what clusters are the best place for a new store.
- The tourist agencies could use this information to offer low-cost routes focusing on the most attractive and comfortable places near metro stations.
- -Compare two cities let to compare urban planning policies, it will be useful to improve it

With foursquare, it is possible to get data and compare the venues near the station and establish clusters and check how similar are them. This clustering could help to classify the stations and improve the location for new stores or classify the zone with this information.

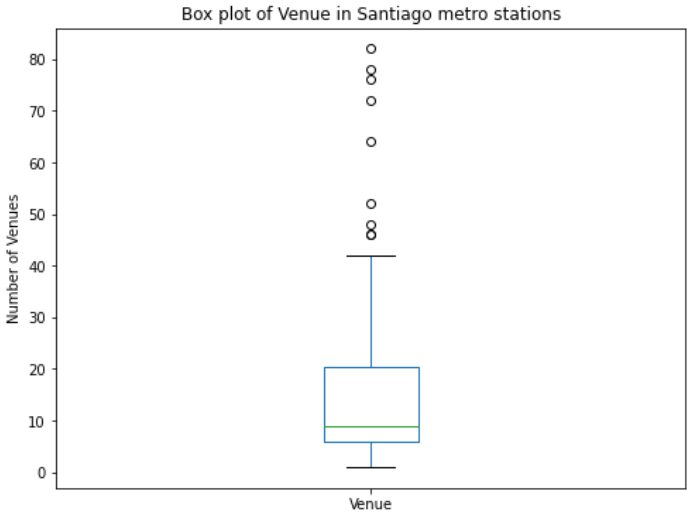
For Metro de Santiago can obtain the list of metro stations from the Wikipedia page: https://es.wikipedia.org/wiki/Anexo:Estaciones_del_Metro_de_Santiago, with this it's possible to use the package geopy and to obtain the coordinates for each metro station. However, some stations are misplaced, thus it’s obtained from the Wikipedia pages of the metro station. With Folium library, it can show up the metro stations on the map of Santiago de Chile



Using foursquare API, It's possible to get the venue near the station. It's stablished 100 as maximum amount of venues and the radius as 260 meters, this is a quarter of the average distance between Santiago metro stations.

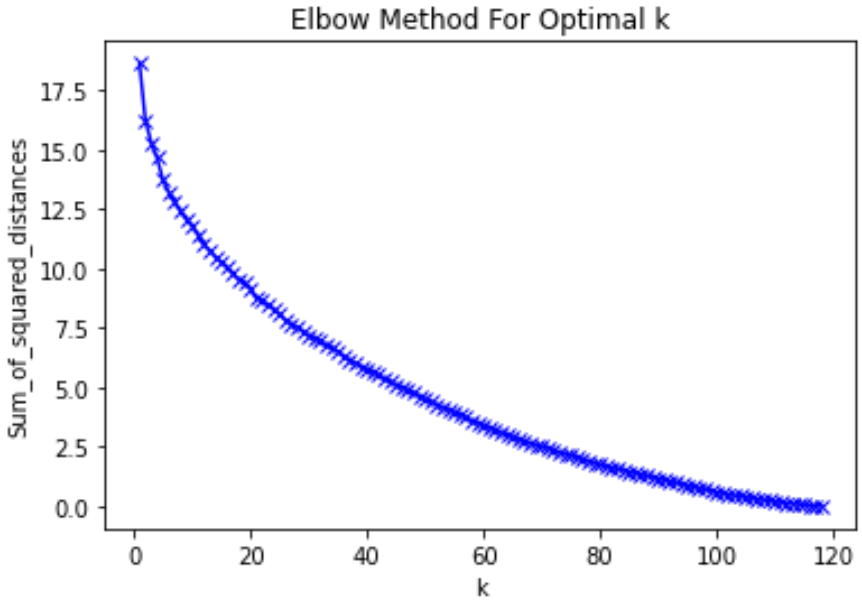


The plot shows the amount of Venue for each station, the average value of venues is 15.974, It's obvious that there are many stations with a number of venues greater. In a box plot, it's observed that there are some outliers stations. It´s listed these outliers

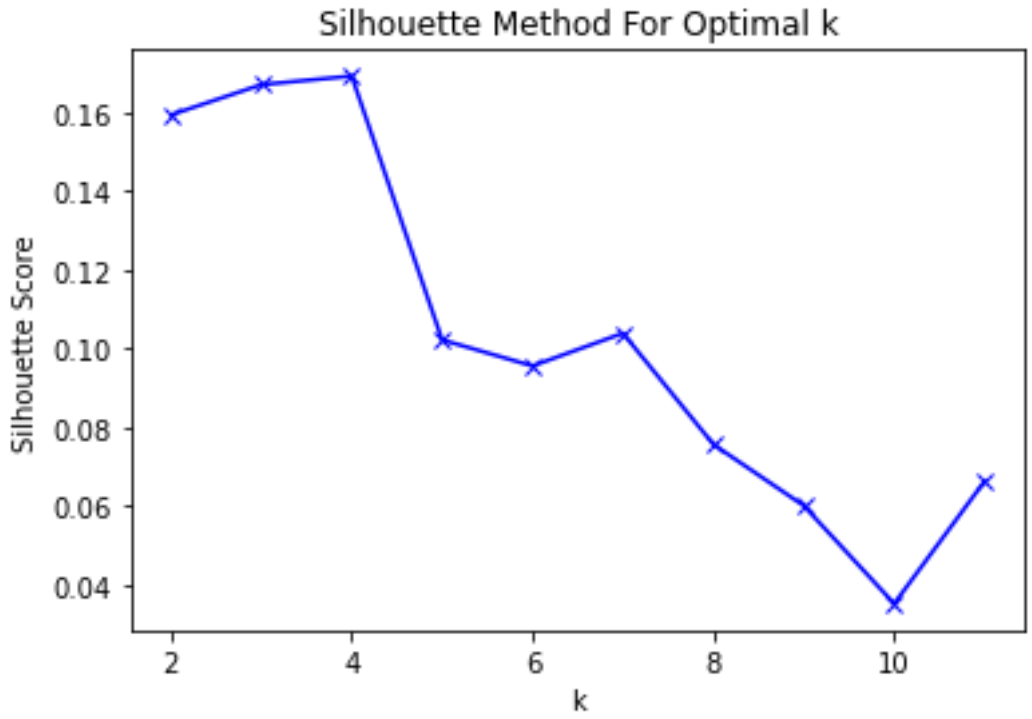


	Metro	Venue
1	Baquedano	52
54	Los Leones	76
61	Manuel Montt	46
75	Plaza Egaña	82
76	Plaza de Armas	46
83	Puente Cal y Canto	78
97	Santa Ana	48
104	Tobalaba	72
108	Universidad de Chile	64

It's done the clustering of the data using the K-means method, for selecting the k value It's used the elbow method.



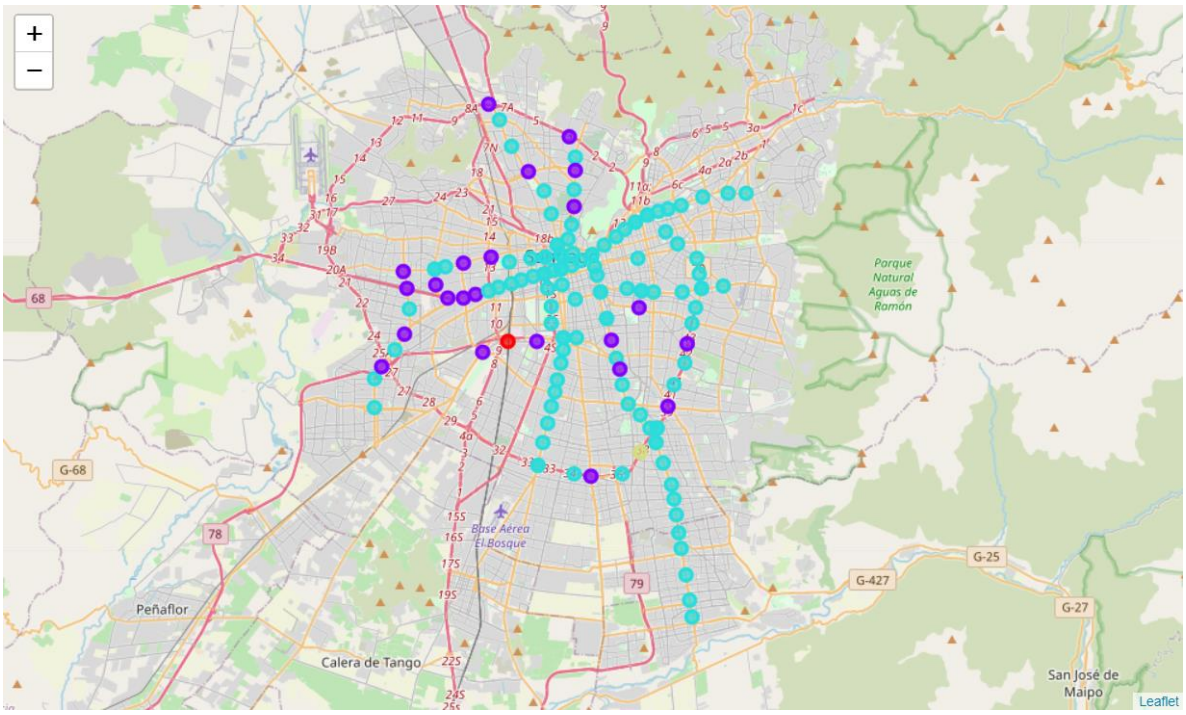
It isn't observed a clear elbow in the plot, then it's used another method: silhouette method, this method quantifies how dissimilar is the point to its own cluster and how good is assigned the point to the near cluster, returning a value *S*, between -1 and 1, when *S* is close to 1 better is the value k for clustering. The silhouette method could be applied using the function *silhouette_score*,



It's selected the greatest value k=4. When the clustering is done, it's obtained the percentage of stations in each cluster.

Cluster Labels	
0	0.735294
1	16.911765
2	81.617647
3	0.735294

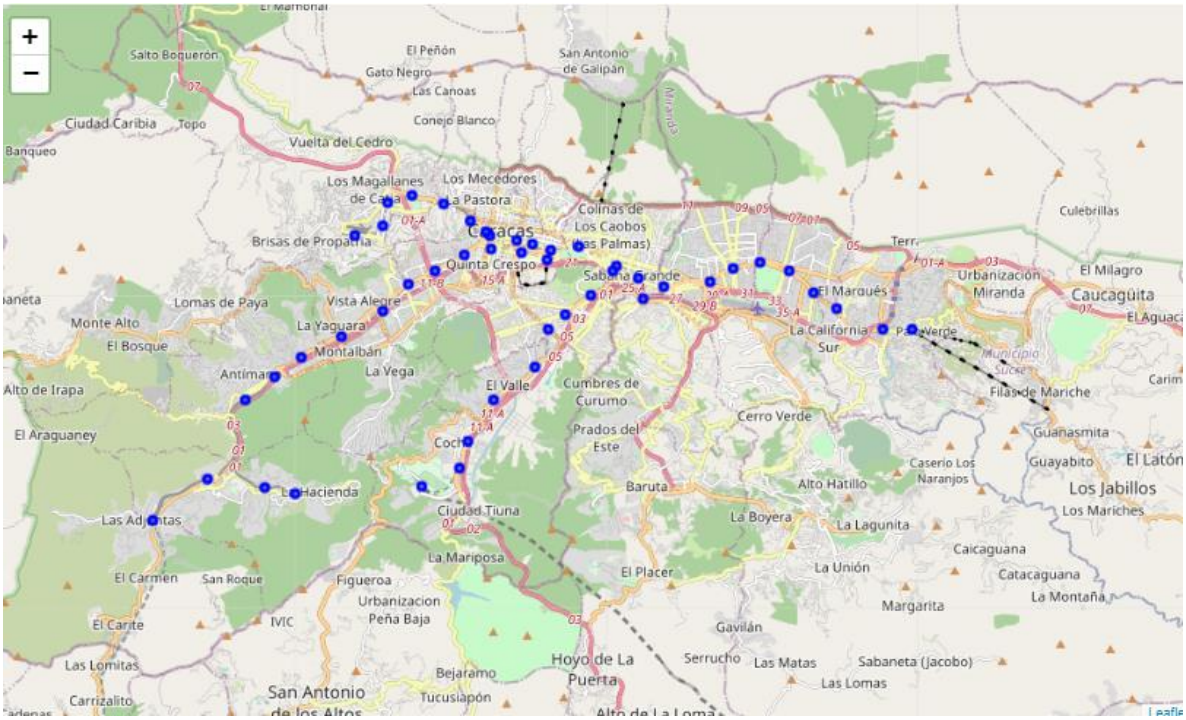
This showed that there is a big cluster. On the map, it's observed than almost all the station belongs to a big cluster



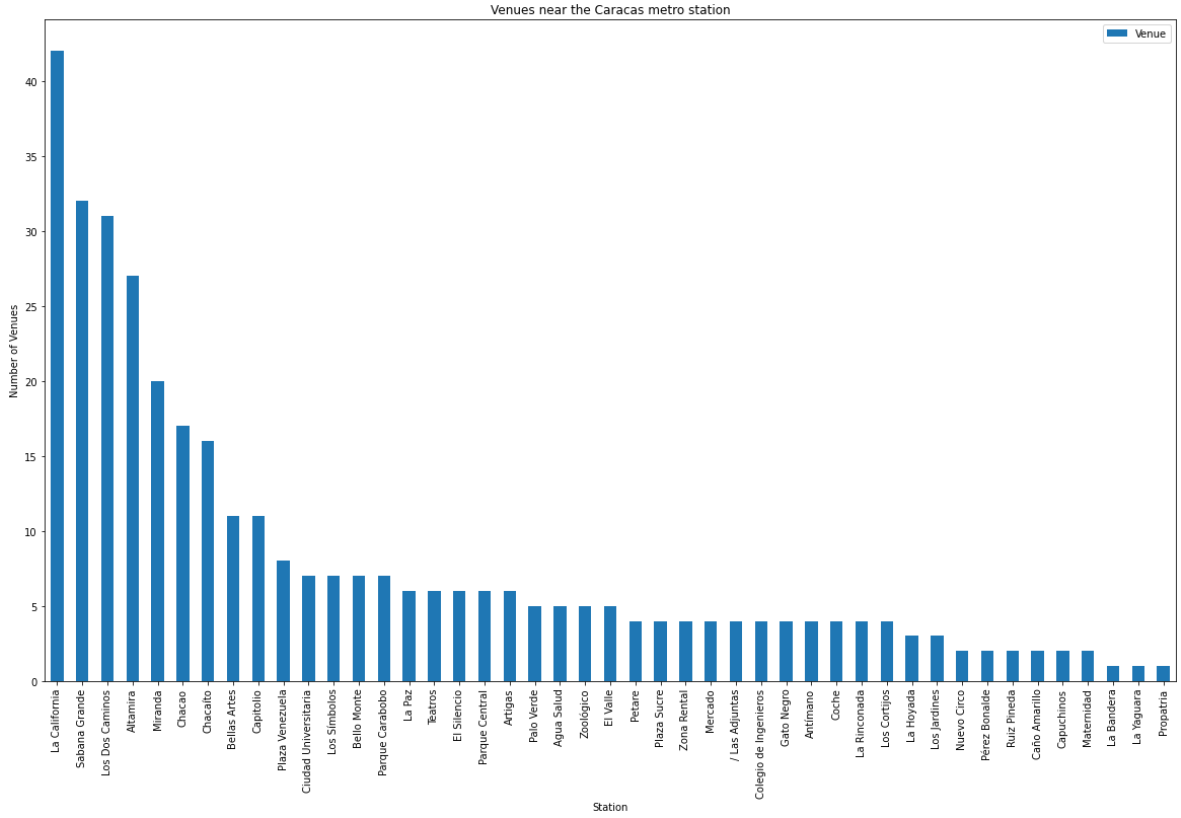
It’s remarked two clusters made of one cluster. Santa Julia station has just one venue (Food Court) and Lo Valledor station has just two (Metro Station and Train Station). It’s checked that the outliers stations belong to the big cluster, as it’s seems

	Metro	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
12	Universidad de Chile	2	Bookstore	Coffee Shop	Restaurant	Gym	Theater	Dessert Shop	Chinese Restaurant	Memorial Site	Mediterranean Restaurant	Sandwich Place
15	Baquedano	2	Plaza	Sandwich Place	Coffee Shop	Convenience Store	Theater	Pizza Place	Event Space	Farmers Market	Ski Area	Snack Place
17	Manuel Montt	2	Pizza Place	Café	Restaurant	Coffee Shop	Deli / Bodega	Bar	French Restaurant	Burger Joint	School	Sandwich Place
19	Los Leones	2	Hotel	Coffee Shop	French Restaurant	Café	Clothing Store	Chinese Restaurant	Tattoo Parlor	Record Shop	Gourmet Shop	Restaurant
20	Tobalaba	2	Coffee Shop	Restaurant	Pharmacy	Deli / Bodega	Hotel	Bakery	Dessert Shop	Diner	Indian Restaurant	Electronics Store
34	Puente Cal y Canto	2	Seafood Restaurant	Peruvian Restaurant	Latin American Restaurant	Flower Shop	Farmers Market	Bar	Camera Store	Greek Restaurant	Boutique	Food
35	Santa Ana	2	Japanese Restaurant	Coffee Shop	Restaurant	Yoga Studio	Dance Studio	Breakfast Spot	Burger Joint	Cafeteria	Café	Deli / Bodega
63	Plaza de Armas	2	Coffee Shop	Peruvian Restaurant	Sandwich Place	History Museum	Fast Food Restaurant	Breakfast Spot	Boutique	Café	Cantonese Restaurant	Salad Place
72	Plaza Egaña	2	Coffee Shop	Multiplex	Electronics Store	Fast Food Restaurant	Italian Restaurant	Clothing Store	Bakery	Café	Bike Rental / Bike Share	Bookstore

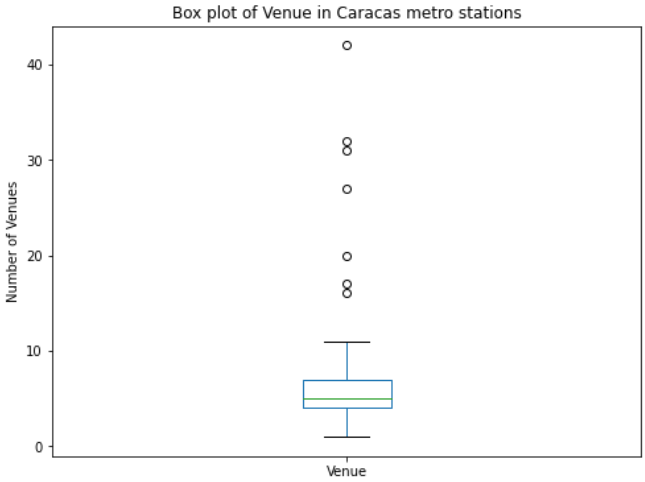
For Metro de Caracas, almost all the stations are misplaced with geopy, therefore it’s gotten the coordinates for these stations from these Wikipedia pages It’s rejected the metro stations that still are not operative. With Folium library, it can show up the metro stations on the map of Caracas



Using foursquare API, It's possible to get the venue near the station. It's established 100 as maximum amount of venues and the radius as 260 meters, this is a quarter of the average distance between Santiago metro stations.

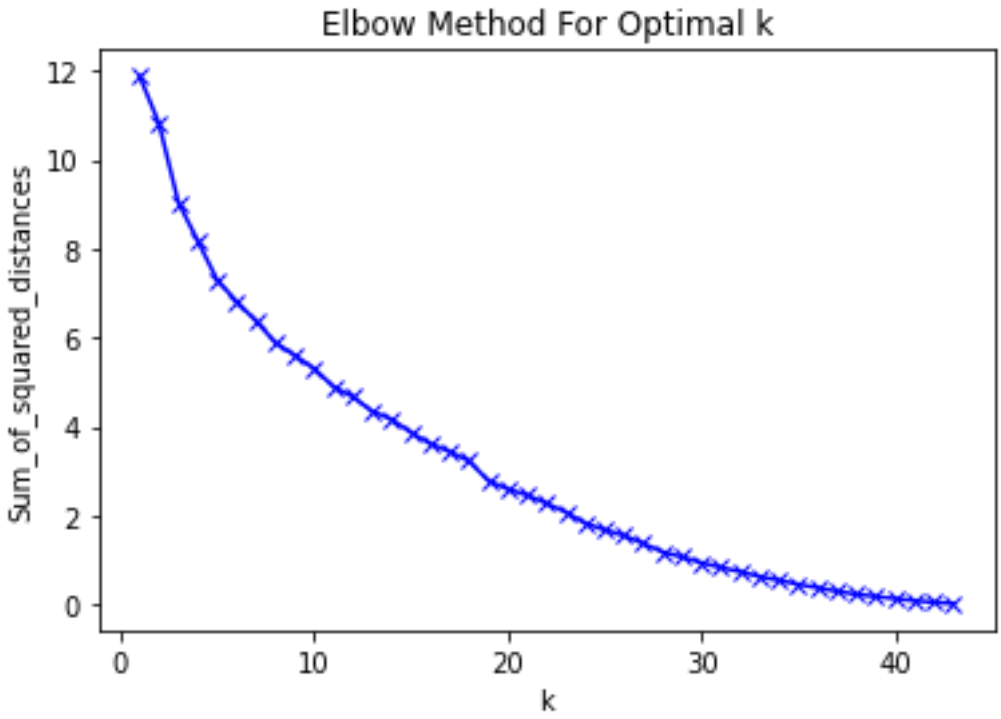


The plot shows the amount of Venue for each station, the average value of venues is 7.955, It's obvious that there are many stations with a number of venues greater. In a box plot, it's observed that there are some outliers stations. It´s listed these outliers

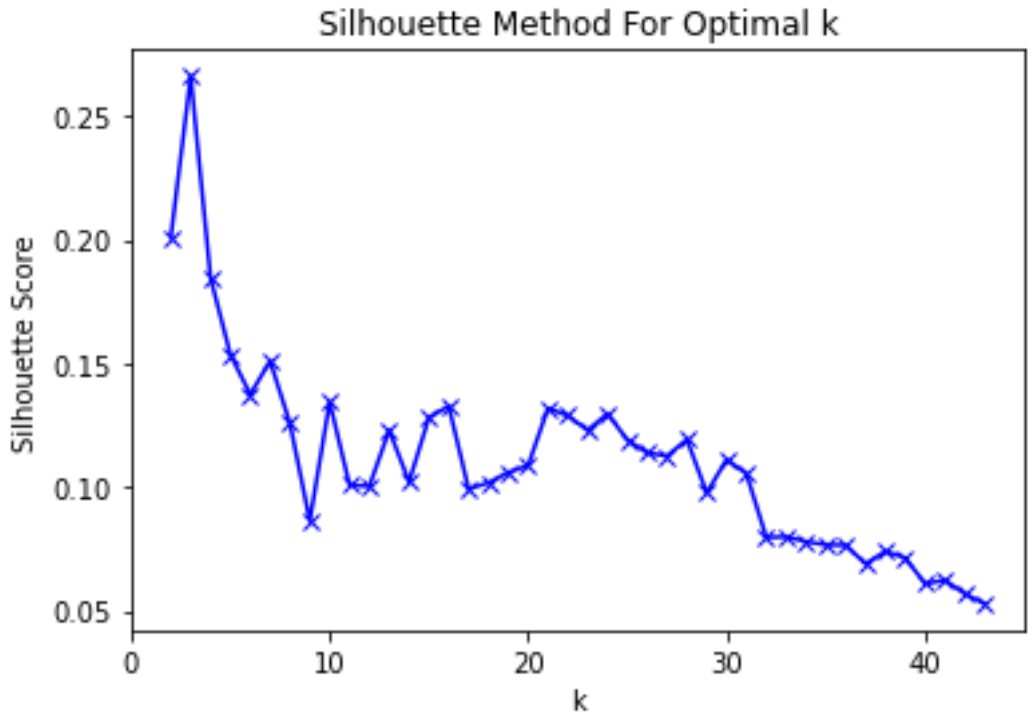


	Metro	Venue
2	Altamira	27
10	Chacao	17
11	Chacaíto	16
19	La California	42
25	Los Dos Caminos	31
30	Miranda	20
41	Sabana Grande	32

It's done the clustering of the data using the K-means method, for selecting the k value It's used the elbow method.



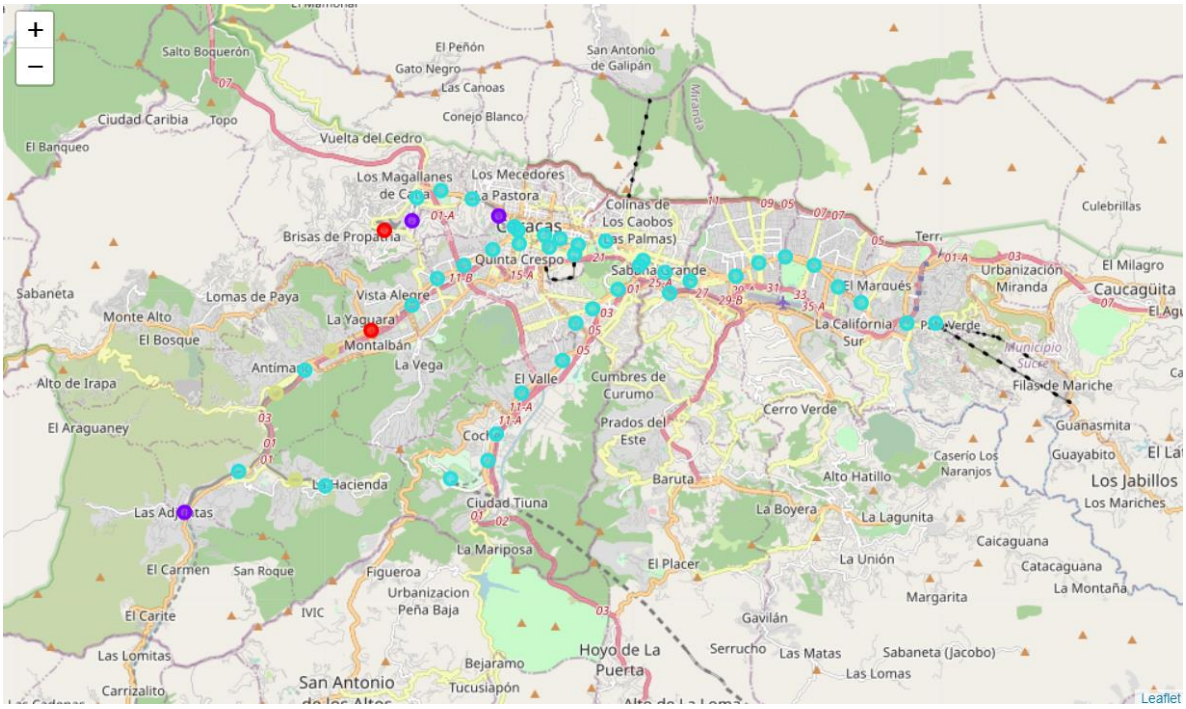
Again, It is not observed a clear elbow in the plot, then it's used the silhouette method, explained above. The silhouette method could be applied using the function *silhouette_score*,



It’s selected the greatest value k=3. When the clustering is done, it’s obtained the percentage of stations in each cluster. When it’s made the clustering, it’s found three stations (Carapita, Mamera y Caricuao) without a venue, then they are grouped in a new cluster

Cluster Labels	
0	4.166667
1	6.250000
2	83.333333
3	6.250000

As shown for Santiago metro stations, there is a big cluster. In the map, it’s observed than almost all the station belongs to a big cluster



It’s checked that the outliers stations belong to the big cluster, as it seems below:

	Metro	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
12	Sabana Grande	2	Italian Restaurant	Spanish Restaurant	Gym	Bakery	Café	Fast Food Restaurant	Other Great Outdoors	Burger Joint	Coffee Shop	Mediterranean Restaurant
13	Chacaíto	2	Italian Restaurant	Gym	Cupcake Shop	Department Store	Mobile Phone Shop	Chinese Restaurant	Café	Bus Station	Sandwich Place	Breakfast Spot
14	Chacao	2	Sandwich Place	Business Service	Café	Cajun / Creole Restaurant	Mediterranean Restaurant	Pizza Place	Japanese Restaurant	Pharmacy	Fast Food Restaurant	Gastropub
15	Altamira	2	Coffee Shop	Chinese Restaurant	Arepa Restaurant	Pizza Place	Plaza	Falafel Restaurant	Gastropub	Breakfast Spot	Gay Bar	Pub
16	Miranda	2	Sandwich Place	Café	Yoga Studio	Gym	Hotel	Hotel Bar	Falafel Restaurant	Latin American Restaurant	Donut Shop	Concert Hall
17	Los Dos Caminos	2	Fast Food Restaurant	Electronics Store	Pizza Place	Café	Sushi Restaurant	Sandwich Place	Ice Cream Shop	Clothing Store	Shopping Mall	Cupcake Shop
19	La California	2	Ice Cream Shop	Fast Food Restaurant	Italian Restaurant	Clothing Store	Coffee Shop	Sandwich Place	Seafood Restaurant	Café	Pharmacy	Pizza Place

Comparing both cities, it is observed that there is a great difference in the average number of venues near the stations. However, it's remarked that in both cities it's found stations that have a number of places far above the average, denoting that there are stations that have a greater socio-economic interest. Another similarity between the cities is that in both we find a large cluster, which indicates that although they are in different parts of the city, the stations are very similar. The analysis presented can be extended to other cities, as well as it can be deepened to find much deeper similarities.