# Introduction

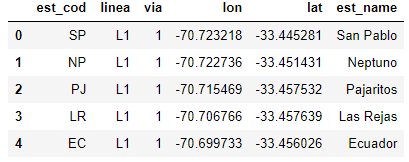
This report analyzes the surroundings of the [Metro de Santiago](https://www.metro.cl/)'s subway stations. For Metro, it is important to understand the behavior of its customers. Much of this behavior comes from the characteristics of the origin and destination of the trip, therefore describing the stations regarding the services and products of the surrounding environment can provide Metro with valuable information to understand certain indicators of demand.

The aim of this report will be to get information of the surrounding of each station from the Foursquare API, then with this data create clusters of stations that allows to summary the pattern observed in the data and finally visualize this results to conclude with the findings of the models.

# Data

The data sets used for this analysis are:

* Station information: Station's identifier, geospatial information (latitude and longitude), line and route it belongs to. Table shape: (125, 6)

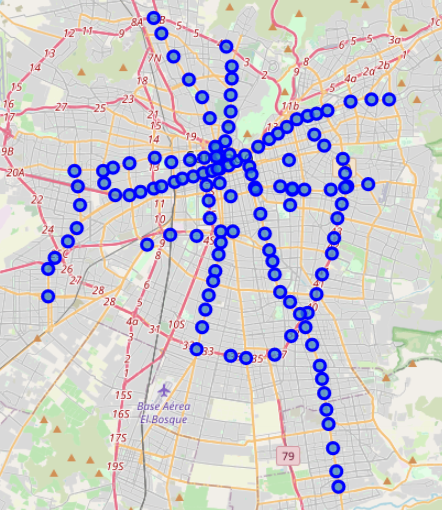


* Surrounding information: The venues and categories of the surrounding that we get from Foursquare API.



We will transform the venues data set in a wide format that allows us to use each category like a feature, then we will calculate the representative percentage for each category and use it to define clusters of stations.

Let’s plot the stations of the Metros network.



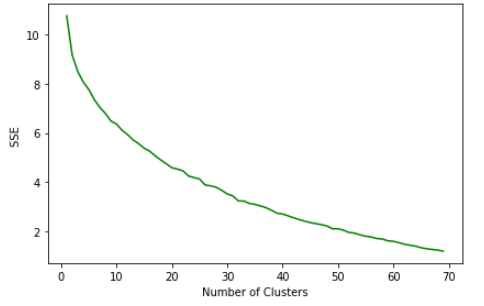
# Methodology

For this project we will fit a clustering model of the stations. For that we will use the frequency of the different categories of venues. First we create a wide format and then we group each station and aggregating the frequency by the mean, with this we obtain how important is that category in the station surrounding.

In this next table we transform the data to show the most common venues for some stations for better understanding.



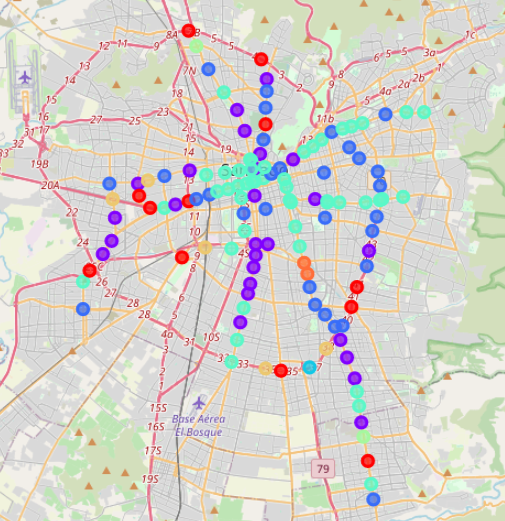
With the numeric table we test different cluster number to see how the sum square error vary. This plot show the result:



With the criteria of the elbow, and observing the plot, a good number of clusters to use is 8 clusters to fit our model.

# Result and discussion

Let´s plot the result of the model:



**Cluster 8**

**Cluster 7**

**Cluster 6**

**Cluster 5**

**Cluster 4**

**Cluster 3**

**Cluster 2**

**Cluster 1**

Now let’s present the structure of each cluster with the venue frequency for the 10 most common venue for each one.



Cluster 1

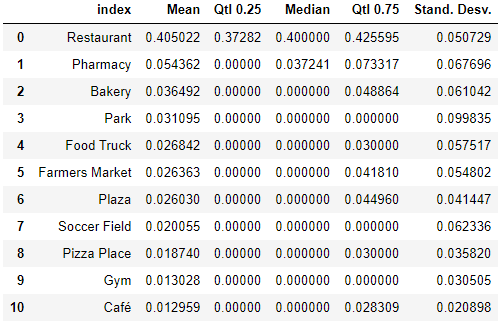
Cluster 4

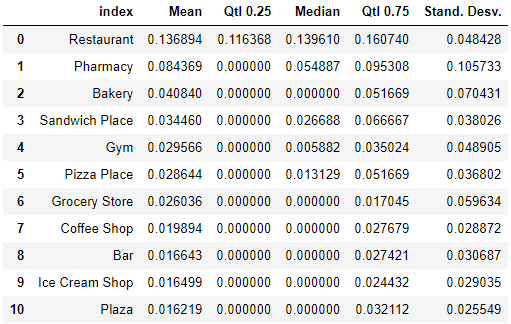
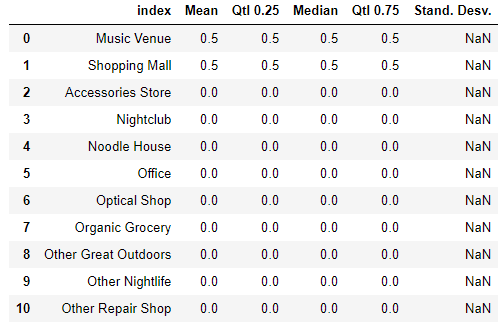
Cluster 5

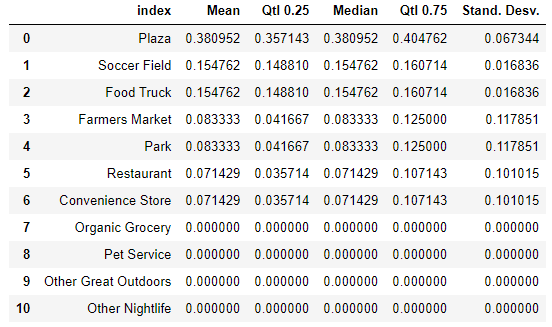
Cluster 6

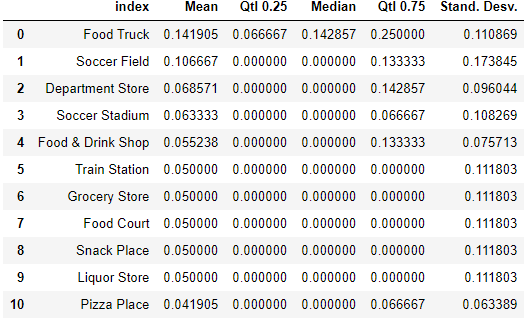
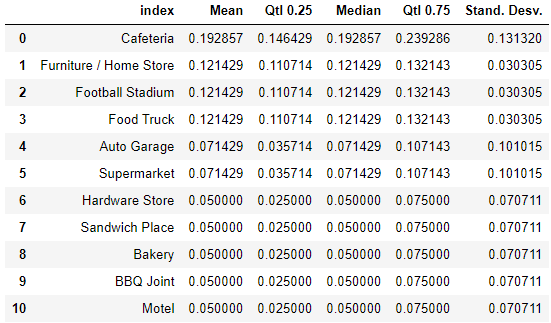
Cluster 3

Cluster 2









Cluster 8

Cluster 7

Then with this tables let’s try to describe the different clusters:

* Cluster 1: The surroundings of this stations are characterized with a bus station near them, with some other venues like restaurants, parks and pharmacy. Probably in this group are the intermodal stations.
* Cluster 2: This stations are near a restaurants zone, with this venue having the higher percentage of the surrounding (37% to 42% interquartile distance)
* Cluster 3: This stations have more variety of venues in the surround with a medium percentage of restaurant, but also pharmacy and a lot of low percentage venues like gyms, bakeries and others.
* Cluster 4: This is a cluster with one station, that only have a mall and music venue in the surrounding.
* Cluster 5: This stations have a high percentage of the surrounding with restaurants, greater than cluster 3, but it also have a good variety of venues like bars, plazas, coffee shops and others.
* Cluster 6: This group stations have a surrounding environment with a high percentage plaza, soccer field and food truck. So is a zone for sports and outdoors activities.
* Cluster 7: This stations are near food trucks, soccer fields and stadium.
* Cluster 8: This cluster has high percentage of coffee stores and furniture/homes stores.

# Conclusion

In this analysis we merge data from Santiago’s subways stations with the venues of their surroundings downloaded from the Foursquare’s API. This information allows to fit a cluster model using k-means methods and found segmentation for the 125 stations of the network. We obtain 8 clusters that represent specific surrounding structure and allows us to have a better and summarized understanding of this system.

For next step we can use the pattern we find in this process and test if it can helps in other models of customer behavior.