Introduction

This report analyzes the surroundings of the Metro de Santiago'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)

	est_cod	linea	via	lon	lat	est_name
0	SP	L1	1	-70.723218	-33.445281	San Pablo
1	NP	L1	1	-70.722736	-33.451431	Neptuno
2	PJ	L1	1	-70.715469	-33.457532	Pajaritos
3	LR	L1	1	-70.706766	-33.457639	Las Rejas
4	EC	L1	1	-70.699733	-33.456026	Ecuador

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

	Station	Station Latitude	Station Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	San Pablo	-33.445281	-70.723218	Zen Sushi Ya	-33.444410	-70.724894	Sushi Restaurant
1	San Pablo	-33.445281	-70.723218	Ahí Sushi	-33.443676	-70.723286	Sushi Restaurant
2	San Pablo	-33.445281	-70.723218	Spacio1	-33.444496	-70.723931	Convenience Store
3	San Pablo	-33.445281	-70.723218	Pollos Asados A Las Brasas Las Cañas	-33.443959	-70.723401	Wings Joint
4	San Pablo	-33.445281	-70.723218	China Popular	-33.444355	-70.724579	Chinese Restaurant
...	...	...	...	...	...	...	...
3634	FERNANDO CASTILLO VELASCO	-33.452115	-70.557643	Salcobrand	-33.453642	-70.558655	Pharmacy
3635	FERNANDO CASTILLO VELASCO	-33.452115	-70.557643	Dulceria Don Felipe	-33.451829	-70.555030	Dessert Shop
3636	FERNANDO CASTILLO VELASCO	-33.452115	-70.557643	Plaza Blest Gana	-33.454248	-70.558539	Plaza
3637	FERNANDO CASTILLO VELASCO	-33.452115	-70.557643	Fu Xing	-33.452243	-70.561770	Chinese Restaurant
3638	FERNANDO CASTILLO VELASCO	-33.452115	-70.557643	Dr. Fish	-33.452354	-70.561728	Pet Store

3639 rows × 7 columns

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.

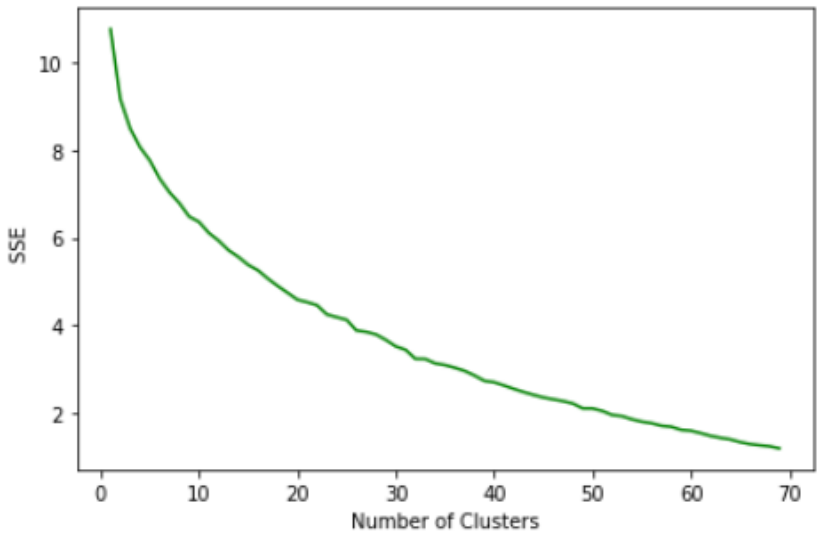
Methodology

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.

	Station	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
0	Agricola	Hardware Store	Sandwich Place	Motel	Football Stadium	Plaza	Food Truck	Furniture / Home Store	Bakery	Cafeteria	BBQ Joint
1	Alcantara	Restaurant	Coffee Shop	Bakery	Pharmacy	Cosmetics Shop	Café	Plaza	Sandwich Place	Salad Place	Hotel
2	Baquedano	Restaurant	Café	Sandwich Place	Coffee Shop	Park	Plaza	Gift Shop	Hotel	Hostel	Rental Car Location
3	Barrancas	Soccer Field	Food & Drink Shop	Restaurant	Food	Soccer Stadium	Bus Station	Pizza Place	Speakeasy	Tunnel	Food Truck
4	Bellas Artes	Restaurant	Yoga Studio	Sandwich Place	Hotel	Bookstore	Coffee Shop	Café	Burger Joint	Hostel	Pizza Place

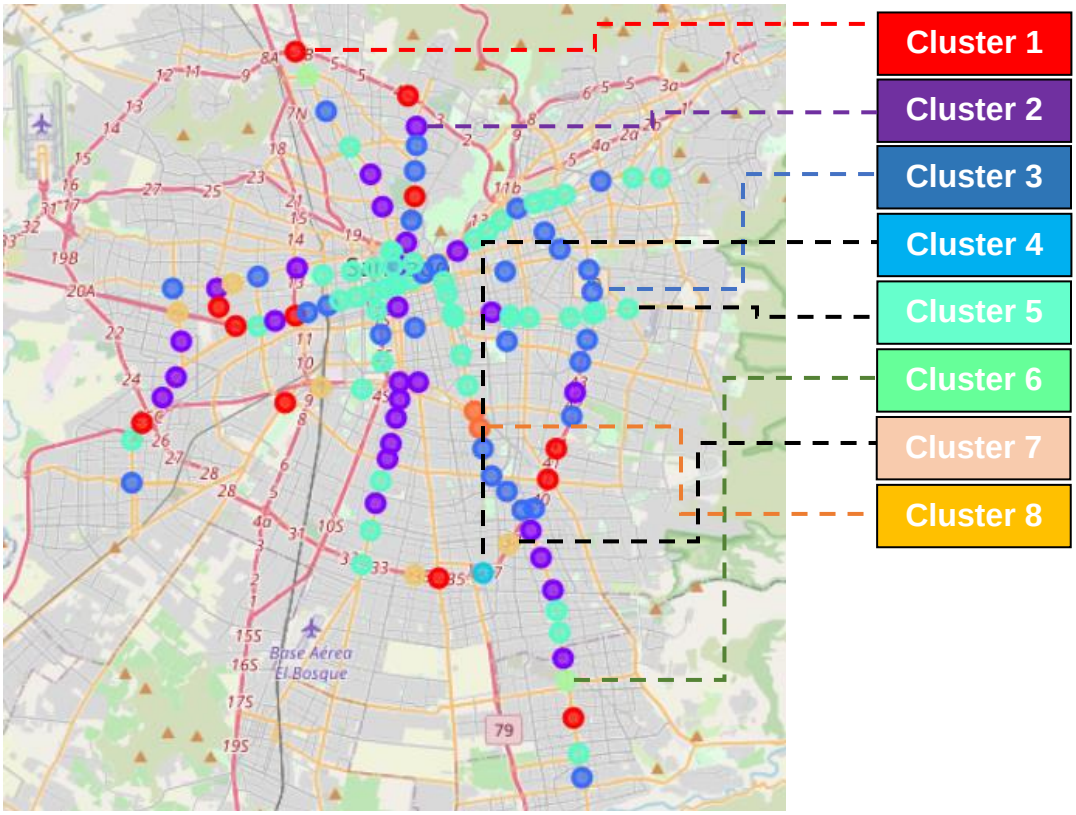
With the numeric table we test different cluster number to see how the sum square error vary. This plot show the result:



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:



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

Cluster 1

	index	Mean	Qtl 0.25	Median	Qtl 0.75	Stand. Desv.
0	Bus Station	0.227080	0.17402	0.225000	0.250000	0.080951
1	Restaurant	0.086404	0.00000	0.079412	0.143939	0.089630
2	Park	0.056291	0.00000	0.000000	0.085784	0.093281
3	Pharmacy	0.054144	0.00000	0.000000	0.085227	0.081891
4	Food Truck	0.045139	0.00000	0.000000	0.031250	0.086051
5	Bakery	0.040530	0.00000	0.000000	0.087500	0.062961
6	Dessert Shop	0.038235	0.00000	0.000000	0.014706	0.077411
7	Hot Dog Joint	0.032680	0.00000	0.000000	0.014706	0.064808
8	Furniture / Home Store	0.030556	0.00000	0.000000	0.000000	0.071715
9	Grocery Store	0.023611	0.00000	0.000000	0.000000	0.060494
10	Ice Cream Shop	0.021569	0.00000	0.000000	0.000000	0.058681

Cluster 2

	index	Mean	Qtl 0.25	Median	Qtl 0.75	Stand. Desv.
0	Restaurant	0.405022	0.37282	0.400000	0.425595	0.050729
1	Pharmacy	0.054362	0.00000	0.037241	0.073317	0.067696
2	Bakery	0.036492	0.00000	0.000000	0.048864	0.061042
3	Park	0.031095	0.00000	0.000000	0.000000	0.099835
4	Food Truck	0.026842	0.00000	0.000000	0.030000	0.057517
5	Farmers Market	0.026363	0.00000	0.000000	0.041810	0.054802
6	Plaza	0.026030	0.00000	0.000000	0.044960	0.041447
7	Soccer Field	0.020055	0.00000	0.000000	0.000000	0.062336
8	Pizza Place	0.018740	0.00000	0.000000	0.030000	0.035820
9	Gym	0.013028	0.00000	0.000000	0.000000	0.030505
10	Café	0.012959	0.00000	0.000000	0.028309	0.020898

Cluster 3

	index	Mean	Qtl 0.25	Median	Qtl 0.75	Stand. Desv.
0	Restaurant	0.136894	0.116368	0.139610	0.160740	0.048428
1	Pharmacy	0.084369	0.000000	0.054887	0.095308	0.105733
2	Bakery	0.040840	0.000000	0.000000	0.051669	0.070431
3	Sandwich Place	0.034460	0.000000	0.026688	0.066667	0.038026
4	Gym	0.029566	0.000000	0.005882	0.035024	0.048905
5	Pizza Place	0.028644	0.000000	0.013129	0.051669	0.036802
6	Grocery Store	0.026036	0.000000	0.000000	0.017045	0.059634
7	Coffee Shop	0.019894	0.000000	0.000000	0.027679	0.028872
8	Bar	0.016643	0.000000	0.000000	0.027421	0.030687
9	Ice Cream Shop	0.016499	0.000000	0.000000	0.024432	0.029035
10	Plaza	0.016219	0.000000	0.000000	0.032112	0.025549

Cluster 4

	index	Mean	Qtl 0.25	Median	Qtl 0.75	Stand. Desv.
0	Music Venue	0.5	0.5	0.5	0.5	NaN
1	Shopping Mall	0.5	0.5	0.5	0.5	NaN
2	Accessories Store	0.0	0.0	0.0	0.0	NaN
3	Nightclub	0.0	0.0	0.0	0.0	NaN
4	Noodle House	0.0	0.0	0.0	0.0	NaN
5	Office	0.0	0.0	0.0	0.0	NaN
6	Optical Shop	0.0	0.0	0.0	0.0	NaN
7	Organic Grocery	0.0	0.0	0.0	0.0	NaN
8	Other Great Outdoors	0.0	0.0	0.0	0.0	NaN
9	Other Nightlife	0.0	0.0	0.0	0.0	NaN
10	Other Repair Shop	0.0	0.0	0.0	0.0	NaN

Cluster 5

	index	Mean	Qtl 0.25	Median	Qtl 0.75	Stand. Desv.
0	Restaurant	0.269116	0.231456	0.272727	0.304073	0.046701
1	Bar	0.035700	0.000000	0.022727	0.052778	0.045367
2	Bakery	0.035510	0.005495	0.022727	0.053571	0.034277
3	Pizza Place	0.033383	0.000000	0.032967	0.050000	0.034494
4	Coffee Shop	0.033352	0.000000	0.027027	0.051286	0.036849
5	Sandwich Place	0.030870	0.000000	0.022727	0.050424	0.031376
6	Pharmacy	0.028291	0.000000	0.013158	0.049754	0.035593
7	Café	0.023181	0.000000	0.011765	0.038462	0.029115
8	Plaza	0.021571	0.000000	0.016949	0.032866	0.024839
9	Burger Joint	0.020149	0.000000	0.016949	0.027212	0.027556
10	Gym	0.016798	0.000000	0.000000	0.026671	0.024956

Cluster 6

	index	Mean	Qtl 0.25	Median	Qtl 0.75	Stand. Desv.
0	Plaza	0.380952	0.357143	0.380952	0.404762	0.067344
1	Soccer Field	0.154762	0.148810	0.154762	0.160714	0.016836
2	Food Truck	0.154762	0.148810	0.154762	0.160714	0.016836
3	Farmers Market	0.083333	0.041667	0.083333	0.125000	0.117851
4	Park	0.083333	0.041667	0.083333	0.125000	0.117851
5	Restaurant	0.071429	0.035714	0.071429	0.107143	0.101015
6	Convenience Store	0.071429	0.035714	0.071429	0.107143	0.101015
7	Organic Grocery	0.000000	0.000000	0.000000	0.000000	0.000000
8	Pet Service	0.000000	0.000000	0.000000	0.000000	0.000000
9	Other Great Outdoors	0.000000	0.000000	0.000000	0.000000	0.000000
10	Other Nightlife	0.000000	0.000000	0.000000	0.000000	0.000000

Cluster 7

	index	Mean	Qtl 0.25	Median	Qtl 0.75	Stand. Desv.
0	Food Truck	0.141905	0.066667	0.142857	0.250000	0.110869
1	Soccer Field	0.106667	0.000000	0.000000	0.133333	0.173845
2	Department Store	0.068571	0.000000	0.000000	0.142857	0.096044
3	Soccer Stadium	0.063333	0.000000	0.000000	0.066667	0.108269
4	Food & Drink Shop	0.055238	0.000000	0.000000	0.133333	0.075713
5	Train Station	0.050000	0.000000	0.000000	0.000000	0.111803
6	Grocery Store	0.050000	0.000000	0.000000	0.000000	0.111803
7	Food Court	0.050000	0.000000	0.000000	0.000000	0.111803
8	Snack Place	0.050000	0.000000	0.000000	0.000000	0.111803
9	Liquor Store	0.050000	0.000000	0.000000	0.000000	0.111803
10	Pizza Place	0.041905	0.000000	0.000000	0.066667	0.063389

Cluster 8

	index	Mean	Qtl 0.25	Median	Qtl 0.75	Stand. Desv.
0	Cafeteria	0.192857	0.146429	0.192857	0.239286	0.131320
1	Furniture / Home Store	0.121429	0.110714	0.121429	0.132143	0.030305
2	Football Stadium	0.121429	0.110714	0.121429	0.132143	0.030305
3	Food Truck	0.121429	0.110714	0.121429	0.132143	0.030305
4	Auto Garage	0.071429	0.035714	0.071429	0.107143	0.101015
5	Supermarket	0.071429	0.035714	0.071429	0.107143	0.101015
6	Hardware Store	0.050000	0.025000	0.050000	0.075000	0.070711
7	Sandwich Place	0.050000	0.025000	0.050000	0.075000	0.070711
8	Bakery	0.050000	0.025000	0.050000	0.075000	0.070711
9	BBQ Joint	0.050000	0.025000	0.050000	0.075000	0.070711
10	Motel	0.050000	0.025000	0.050000	0.075000	0.070711

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.