Objective

Purpose of the project is to analyze the neighborhood of the different metro stations in Milano to split them in different clusters based on the venues profiles.

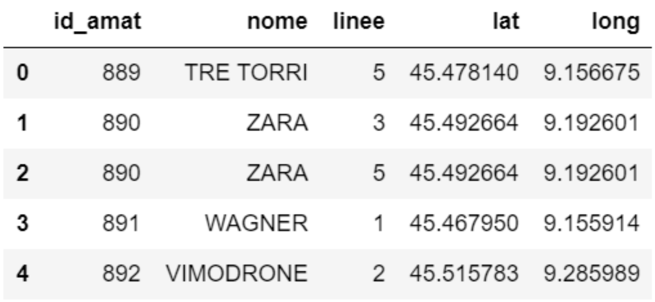
Milano has 4 metro lines (red, green, yellow and purple) with more than 100 stations.

The analysis might be useful for the city of Milano or real estate companies.

Data

Metro Stations Data

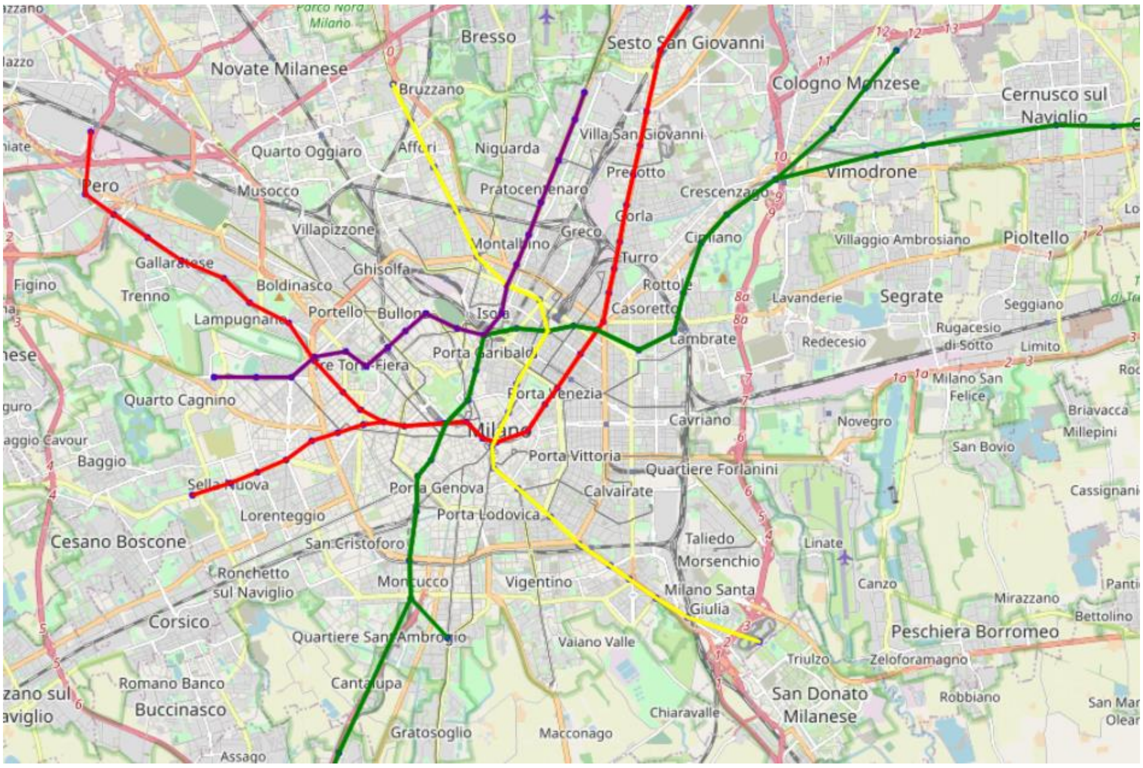
1) List of all metro stations with geo-coordinates which can be downloaded from this site: http://dati.comune.milano.it/it/dataset/ds535_atm-fermate-linee-metropolitane



	id_amat	nome	linee	lat	long
0	889	TRE TORRI	5	45.478140	9.156675
1	890	ZARA	3	45.492664	9.192601
2	890	ZARA	5	45.492664	9.192601
3	891	WAGNER	1	45.467950	9.155914
4	892	VIMODRONE	2	45.515783	9.285989

2) Information of the order of stations along the metro ride. The information can be downloaded from this site: http://dati.comune.milano.it/dataset/b8eb04e7-4e99-4ba7-9502-c2aecc4f7e11/resource/0555a0ae-9493-46ca-bc3d-65544f4e99b2

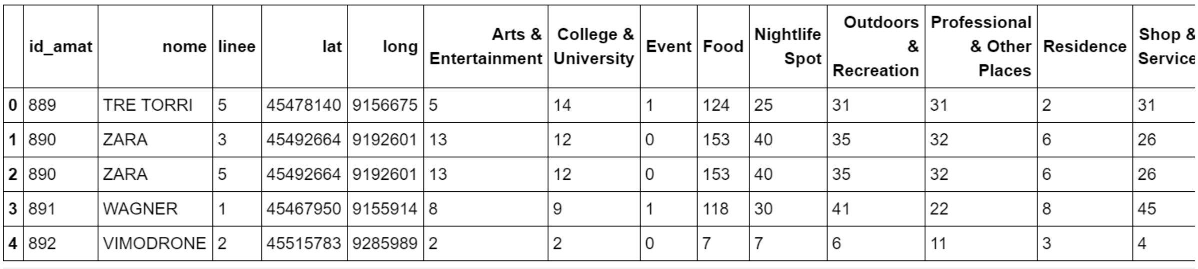
Thanks to the above it will be possible to plot a folium map with all the stations and the different metro lines



Foursquare Data

For each metro stations we will use the Foursquare explore API (https://developer.foursquare.com/docs/api-reference/venues/explore/) to get the number of venues for each of the topline categories

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



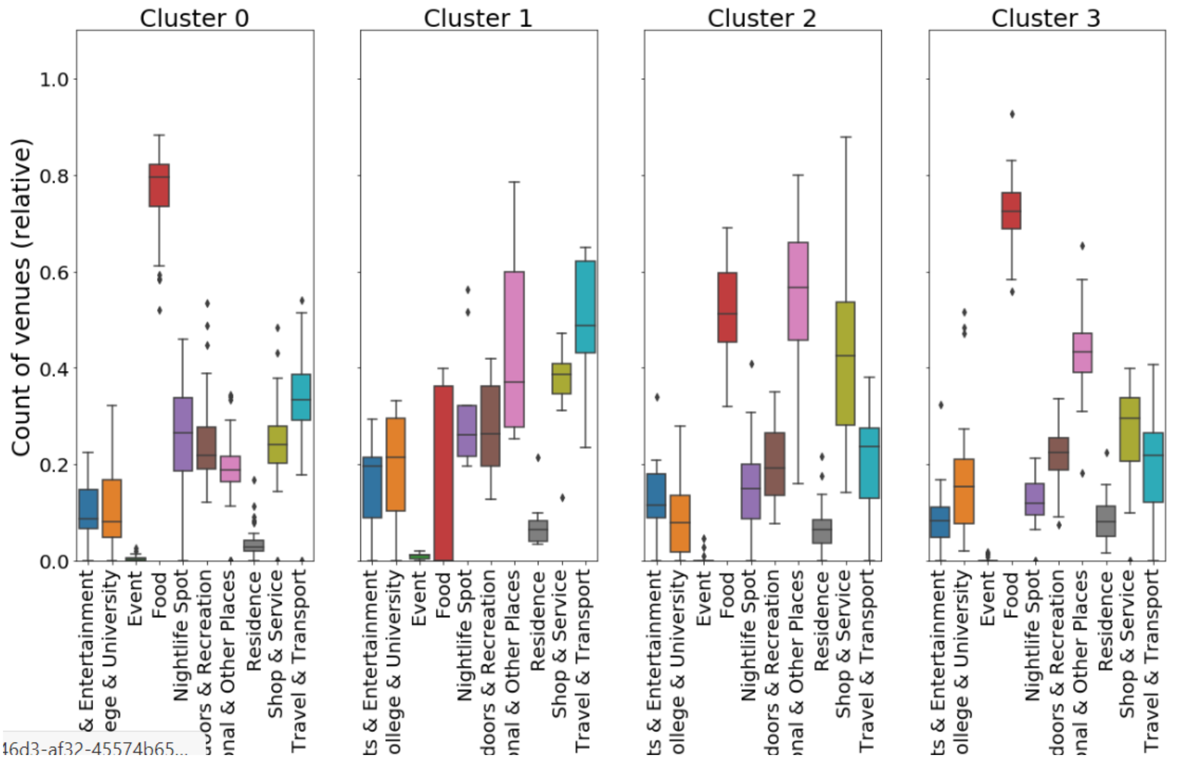
	id_amat	nome	linee	lat	long	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service
0	889	TRE TORRI	5	45478140	9156675	5	14	1	124	25	31	31	2	31
1	890	ZARA	3	45492664	9192601	13	12	0	153	40	35	32	6	26
2	890	ZARA	5	45492664	9192601	13	12	0	153	40	35	32	6	26
3	891	WAGNER	1	45467950	9155914	8	9	1	118	30	41	22	8	45
4	892	VIMODRONE	2	45515783	9285989	2	2	0	7	7	6	11	3	4

Methodology

Once organized the data into a table they are normalized by using the preprocessing package from sklearn.

We will then use the k-cluster approach to build groups of metro stations based on the similarity of the venues.

I tried different number of clusers and went for 4 as final number as providing the best trade-off between segmentation and relevancy of data (e.g. k=2 was simply splitting center to suburbs stations)



Results

Cluser 0 (red): high scores for all venues (especially food). It’s related to the richest areas in the centre of Milano with high footfall from workers and tourists

Cluster 1 (blue): similar to cluster 0 but with a stronger results of Professional places. It’s mainly related to business areas (especially closest to the two train stations of Centrale and Garibaldi)

Cluster 2(yellow): low scores for all values but high for professional places. Areas not really developed but with lot of companies.

Cluster 3 (orange): low scores for all venues. Less developed areas

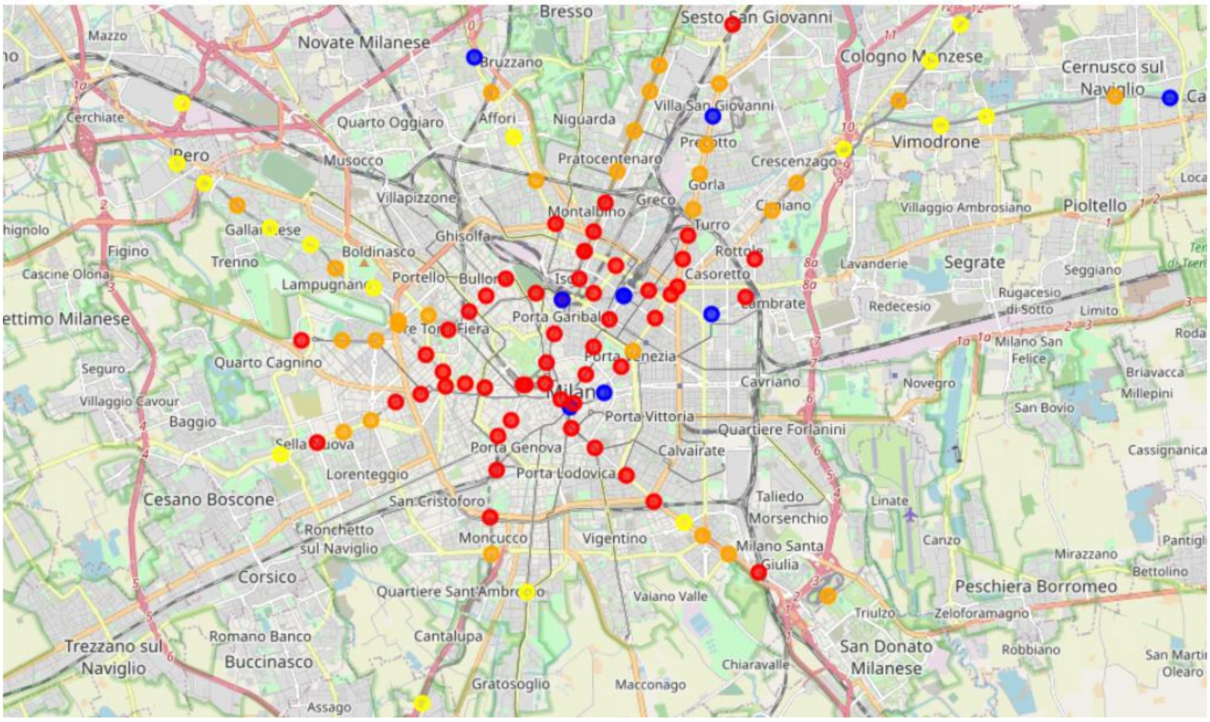
Comments

Foursquare data are good in terms of mapping (especially in the city centre) but they are not able to distinguish between different elements falling in the same category (e.g. a school or a university and a 5 start restaurant versus a coffee place).

Choosing a radius of 1000 meters for collecting venues has been requested to get enough places for suburbs areas but it might create overlaps between metro stations in the city centre (as they are not really far one from the other).

In general results are quite good and they are reflecting the reality of the city.

For more advanced analysis it might be useful to drill on a more granular level of analysis with Foursquare data (I used top line categories) and blend them with other sources (e.g. house market price).


Bresso
Sesto San Giovanni
Mazzo
Novate Milanese
Bruzzano
Cologno Monzese
Cernusco sul Naviglio
Cerchiate
Quarto Oggiaro
Affori
Niguarda
Villa San Giovanni
Vimodrone
Pero
Musocco
Pratocentenaro
Precotto
Crescenzago
Villapizzone
Montalbino
Greco
Gorla
Cimiano
Villaggio Ambrosiano
Pioltello
Gallaratese
Ghisolfa
Turro
Rottole
Figino
Trenno
Boldinasco
Cascine Olona
Portello
Bullona
Isola
Casoretto
Lavanderie
Segrate
Seggiano
Lampugnano
Rugacesio di Sotto
Lambrate
Redecesio
Limito
Porta Garibaldi
Settimo Milanese
Tre Torri Fiera
Milano San Felice
Porta Venezia
Quarto Cagnino
Seguro
Briavacca
Cavriano
Novegro
Millepini
Villaggio Cavour
Milano
Baggio
San Bovio
Porta Vittoria
Quartiere Forlanini
Sella Nuova
Porta Genova
Calvairate
Cassignanica
Lorenteggio
Taliedo
Porta Lodovica
Cesano Boscone
Linate
Morsenchio
San Cristoforo
Ronchetto sul Naviglio
Canzo
Moncucco
Vigentino
Milano Santa Giulia
Pantigliate
Mirazzano
Bettolino
Corsico
Quartiere Sant'Ambrogio
Vaiano Valle
Peschiera Borromeo
Triulzo
Zeloforamagno
Romano Banco
Cantalupa
San Donato Milanese
Trezzano sul Naviglio
Buccinasco
Robbiano
San Martino Olearo
Chiaravalle
Gratosoglio
Macconago
Assago