Social centers in Rome

Coursera Data Science Capstone Project Course

Objective:

Find neighborhoods in Rome more suitable for opening a social center aimed to prevent social discomfort, in particular among young people

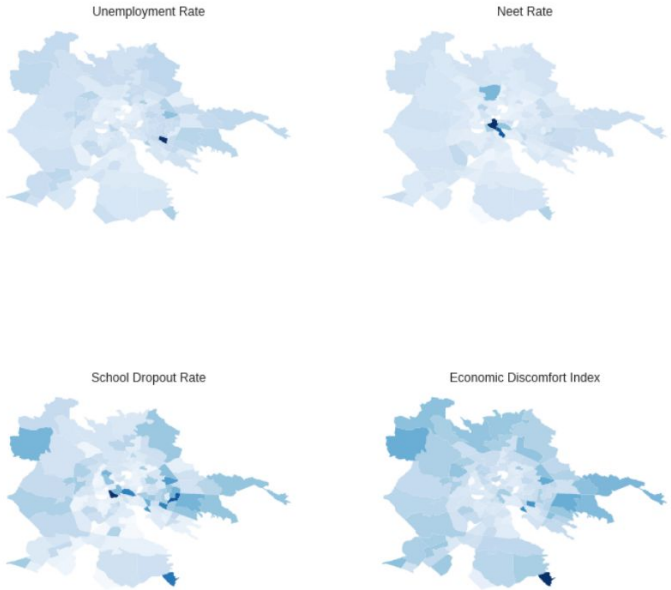
Data acquisition

- Social economic indicators by Urbanistic Zone, from Italian 2011 Census, published by Istat (Italian Institute of Statiscs). Paper link: https://www.istat.it/it/files//2017/01/C-allegato-statistico24_01_17_REV.pdf available as PDF (scaped by tabula-py)
- Venues availability data, retrieved by Foursquare APIs
- Geolocation of Rome's urbanistic zones, retrieved as shapefile from www.mapparoma.info

Exploratory Data Analysis - socio-economic features

	Surface	Population	Pop0-14	Foreigners	Centrality Index	Old Age Index	Foreigner Rate	School Dropout Rate	Unemployment Rate	Neet Rate	Economic Discomfort Index	Cultural Employees Index
mean	8.04	18128.08	137.79	1551.05	3.29	163.38	90.51	2.39	9.33	10.42	2.22	6.07
std	12.07	13982.91	28.84	1665.40	9.00	70.01	58.52	1.38	3.24	5.42	0.89	15.81
min	0.50	240.00	65.18	17.00	0.20	34.90	22.70	0.70	3.00	2.60	0.50	0.10
25%	1.80	7575.75	119.45	486.25	0.40	110.00	59.33	1.48	7.40	7.70	1.60	1.08
50%	3.20	15460.00	132.41	1055.00	0.80	171.30	74.15	2.00	8.90	9.55	2.05	1.80
75%	7.22	25989.25	156.05	2008.00	2.12	213.45	106.38	2.90	10.80	11.22	2.70	3.08
max	56.60	80311.00	235.09	10169.00	88.00	471.40	369.60	8.80	36.80	51.30	7.50	97.30

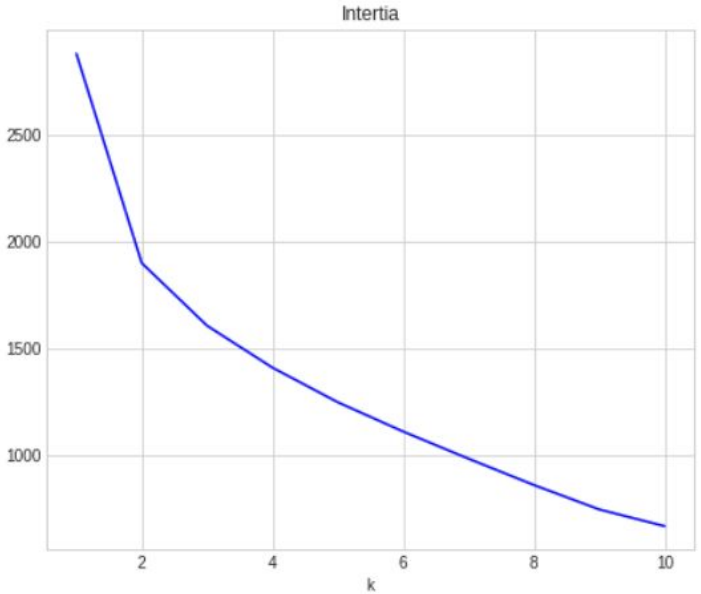
EDA - socio-economic features

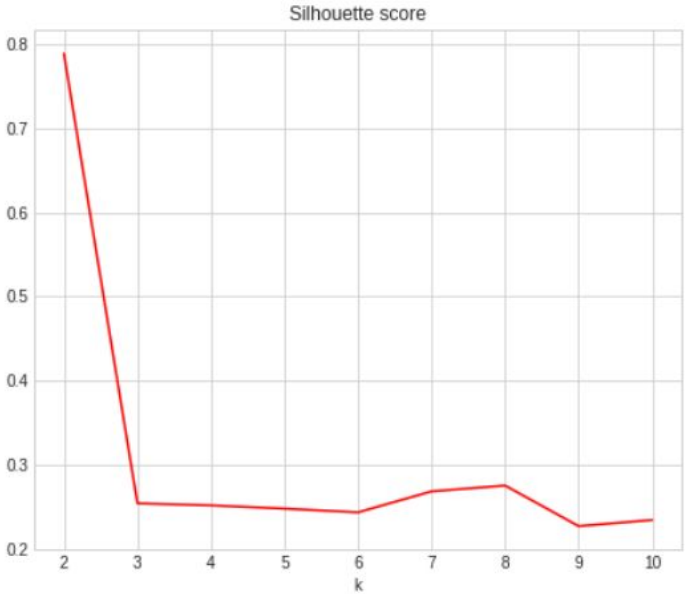


EDA - venues features

	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Total Venues
mean	0.39	0.00	0.00	2.93	0.54	0.69	0.07	0.00	0.48	0.58	5.67
std	1.78	0.00	0.00	10.79	2.18	2.30	0.33	0.01	1.02	2.05	19.14
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	0.20	0.00	0.05	0.00	0.00	0.08	0.00	0.51
50%	0.00	0.00	0.00	0.53	0.04	0.16	0.00	0.00	0.18	0.08	1.15
75%	0.10	0.00	0.00	1.69	0.24	0.38	0.00	0.00	0.38	0.38	3.31
max	14.90	0.05	0.00	108.01	20.48	18.62	3.72	0.07	9.31	20.55	186.22

Clustering by K-Means - metrics



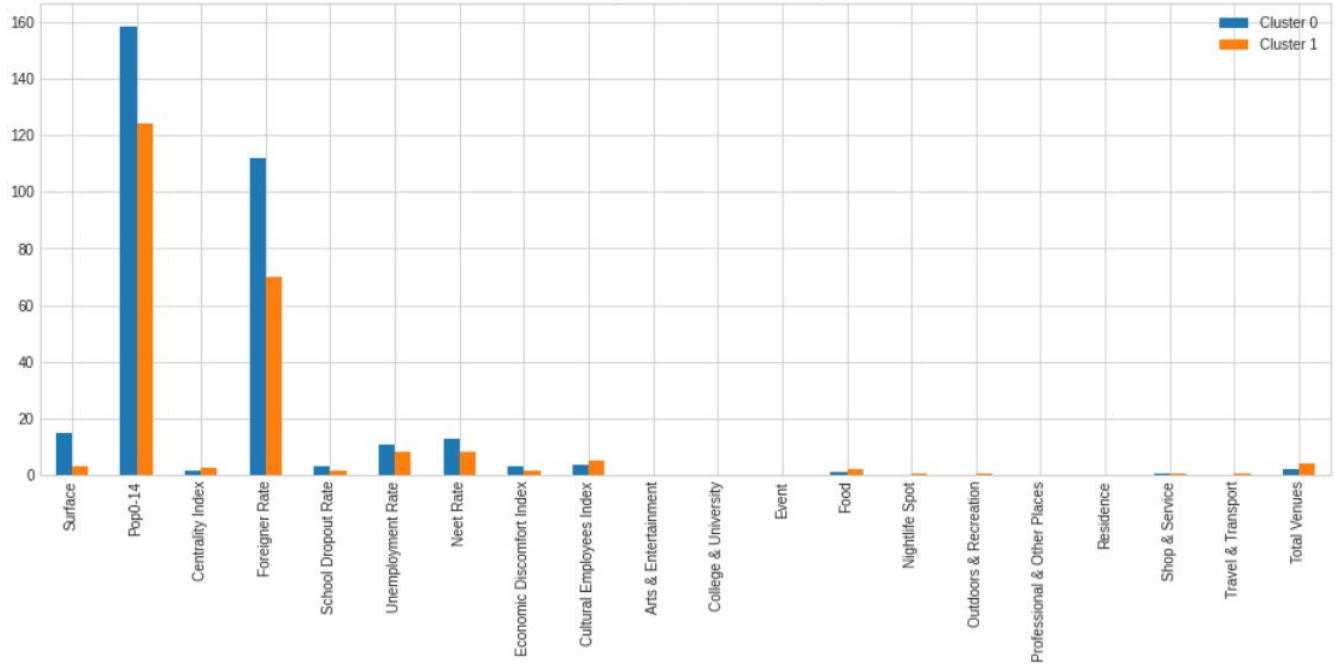


Cluster results - statistics

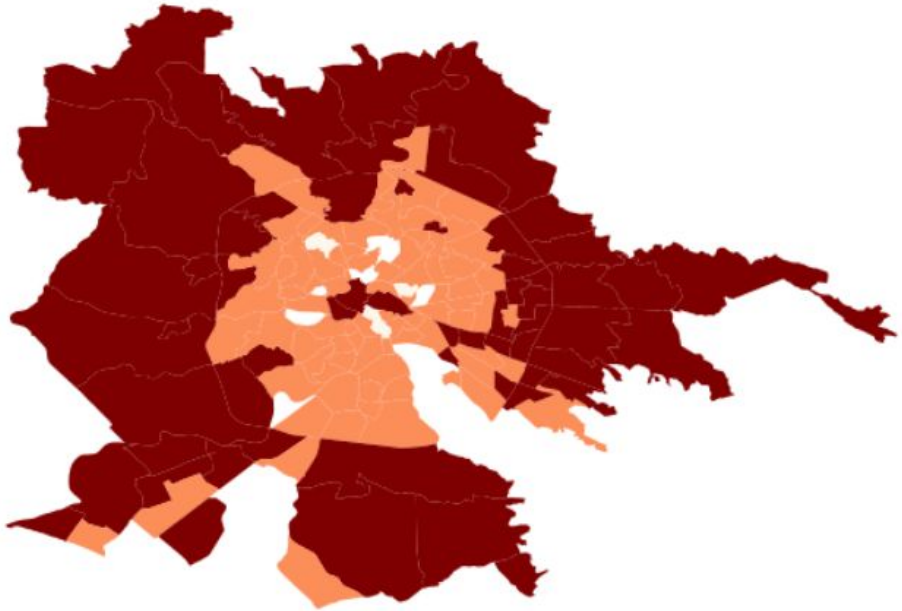
	Surface	Population	Pop0-14	Foreigners	Centrality Index	Old Age Index	Foreigner Rate	School Dropout Rate	Unemployment Rate	Neet Rate	Economic Discomfort Index	Cultural Employees Index	Total Venues	Nr of Zones
0	14.671667	14715.450000	158.542113	1705.100000	1.855000	107.175000	112.001667	3.256667	10.990000	12.666667	2.898333	3.471667	2.110949	60
1	3.369136	21302.160494	124.350715	1489.506173	2.603704	199.029630	70.206173	1.729630	8.222222	8.213580	1.704938	5.035802	4.075663	81
2	1.633333	680.666667	85.840113	131.666667	50.700000	324.800000	209.066667	2.766667	5.833333	24.866667	2.533333	86.066667	119.732840	3

Cluster “0” is our target

Comparison: cluster 0 vs. cluster 1



Clusters distribution

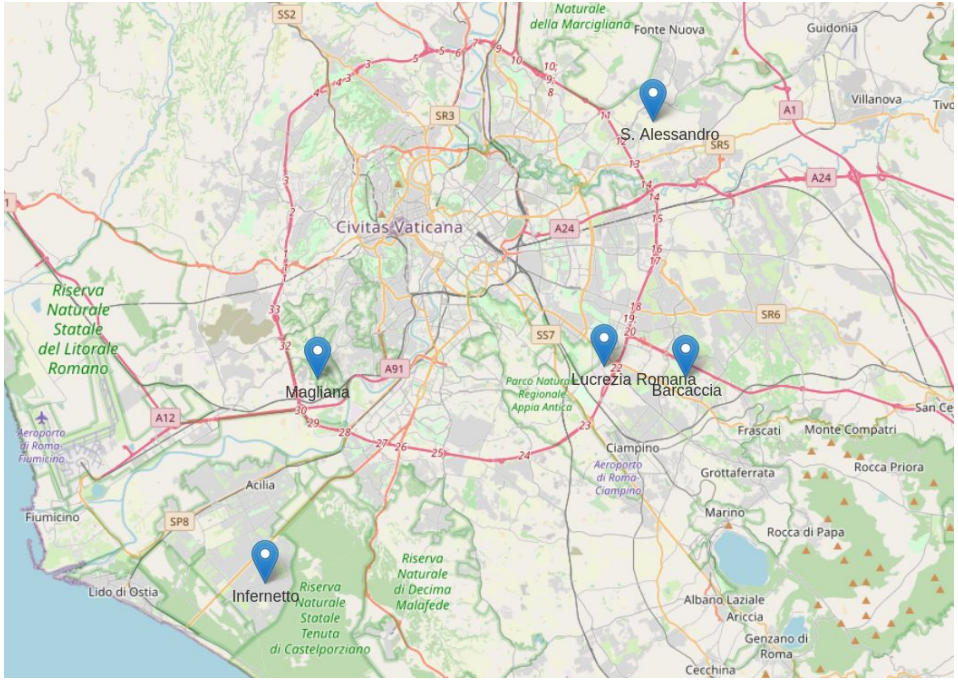


Let's start from here...

Five zones with highest youth rate

name	Surface	Population	Pop0-14	Foreigners	Centrality Index	Old Age Index	Foreigner Rate	School Dropout Rate	Unemployment Rate	Neet Rate	Economic Discomfort Index
S. Alessandro	11.40	9856	235.09	474	0.40	36.20	48.10	1.40	6.60	8.50	2.40
Magliana	11.50	3803	223.24	195	8.30	34.90	51.30	1.70	5.60	15.40	1.80
Barcaccia	5.10	10099	220.52	368	0.30	42.30	36.40	1.30	7.70	7.60	2.60
Infernetto	11.50	24356	210.22	1750	0.20	63.60	71.90	1.10	8.00	9.60	3.10
Lucrezia Romana	1.70	4451	198.38	228	1.40	56.10	51.20	1.30	7.20	7.10	1.60

Where are them?



Conclusions

- One big cluster of 60 zones, suitable for our purpose
- Within this cluster, 5 zones to begin with due to youngest population
- Improvements:
    - Try hierarchical agglomerative clustering to split further Cluster 0
    - Retrieve more data from economic indicators or other sources of location data
    - Try Factor Analysis to identify hidden dimensions