Social centers in Rome

by Michele Gava

February 2021

Coursera Data Science Capstone Project

Table of contents

Introduction

Suppose a social cooperative providing social services for the city of Rome (Italy)

This organization is in charge by Rome's administration to supply services to prevent youth discomfort.

They decided to open social centers in various locations of the city, to provide places where young people can meet and find informations and support

Due to budget limitations, the organization needs to find the areas of the city where this service is needed the most, in order to concentrate their limited financial rescources.

So they reached our Data Science Team and toghether we decided to work on the following question:

Can we find Rome's neighboroods where a social center is most needed?

As we will see in more details in the Results section, we will be able to:

- identify a big cluster of zones with highest discomfort profile (60 zones out of 155)
- suggest the following 5 zones to start our customer's project:
    i. S. Alessandro
    ii. Magliana
    iii. Barcaccia
    iv. Infernetto
    v. Lucrezia Romana

Background

Rome is the capital and biggest city of Italy, with a population of 2,808,293 at 01.01.2021 (source: www.istat.it) and a territory of 1,285 km²

It has several administrative subdivision; one of them is the subdivion in "Zone urbanistiche di Roma" ("Rome's urbanistic zones"), which can approximate the concept of neighborhood, establised in 1977 for statistical and territorial planning reasons. This divides Rome's territory in 155 Zones (= Neighborhoods), grouped in 15 "Municipi" (=Boroughs)

You can find more on Wikipedia: https://it.wikipedia.org/wiki/Zone_urbanistiche_di_Roma (I apologize for been only in italian) and on google maps: https://www.google.com/maps/d/u/0/viewer?mid=1E4TpQ9oftDIBneEK4KzDFwQfbvQ&ll=41.89062981328977%2C12.346966041866727&z=11

The following image (taken from Wikipedia page) shows the map of Rome's urbanistic zones

For the objective of our analyses, we look for a method to create "profiles" of Rome's neighborhoods, such that we can create groups (*clusters*) of neighborhoods which are similar with respect of:
- socio-demographic factors:
economic and social situation can be seen as a factor of discomfort, so we want a neighborhood's profile to take account of it
- density of venues:
shortage of venues where people can meet (including bars, restaurants, cinemas, theatres, museums, and so on) can motivate the needing of a social center as an alternative for young people

The underlying idea is that if we can find the clusters where the above-mentioned factors are low, than those neighborhoods that belong to these clusters are those who requires our service the most.

Data

As stated in the background section, we need both socio-demographic and location data

1. socio-demographic data

For this category of data, we can rely on a paper from Italian National Institute of Statistics ("Istat")

This paper reports a set of socio-economic indicators for all of the 155 Zones of Rome. It's available only in PDF format from ISTAT's site, but we can leverage tabular-py Python's library to scape the tables of data from the document (https://pypi.org/project/tabula-py/)

The document in only in Italian, but We'll provide description of all relevant features.

In this image we show a preview of the data available in the paper:

Tavola riassuntiva per le Zone urbanistiche del Comune di Roma

DENOMINAZIONE AREA SUBCOMUNALE	Superfice dell'area (Km^2)	Popolazione residente	Popolazione 0-14 anni residente	Stranieri residenti	Indice di centralità	Indice di vecchiaia	Incidenza di residenti stranieri	Indice di non completamento del ciclo di scuola secondaria di primo grado	Tasso di disoccupazione	Incidenza di giovani fuori dal mercato del lavoro e dalla formazione	Incidenza delle famiglie con potenziale disagio economico	Indice degli addetti ad attività creative e culturali sulla popolazione servita	Numero di sezioni di censimento
Acilia Nord	9,2	26.023	4.241	2.008	0,2	100,2	77,2	2,1	12,4	10,9	2,9	0,7	197
Acilia Sud	7,1	23.640	3.797	1.833	0,7	107,3	77,5	2,7	12,0	11,3	3,1	2,5	140
Acqua Vergine	11,3	5.355	1.033	487	0,8	48,2	90,9	2,2	7,8	10,9	1,9	0,4	45
Acquatraversa	1,4	8.689	1.266	678	0,3	129,2	78,0	1,1	6,1	7,2	2,6	2,9	26
Aeroporto dell' Urbe	4,4	1.924	247	172	11,0	172,1	89,4	1,6	11,8	10,0	1,9	3,3	27
Alessandrina	3,1	25.978	3.740	2.838	0,4	138,0	109,2	3,5	11,4	11,5	2,8	1,8	96
Appia Antica Nord	20,5	2.389	306	475									36
Appia Antica Sud	10,6	642	127	135									15
Appio	1,4	26.397	2.897	1.981	0,8	235,7	75,0	1,4	7,8	7,6	1,2	2,3	90
Appio-Claudio	3,5	28.783	3.532	1.726	0,6	211,5	60,0	1,7	8,5	7,8	1,4	2,0	99
Aurelio Nord	1,3	17.645	1.996	1.099	0,9	266,2	62,3	1,4	8,3	7,1	1,5	6,0	51
Aurelio Sud	2,9	23.694	2.852	1.751	1,3	234,7	73,9	1,0	6,9	7,0	1,5	2,9	88
Aventino	1,6	6.963	845	692	3,9	214,8	99,4	1,5	8,1	7,1	2,0	6,3	71
Barcaccia	5,1	10.099	2.227	368	0,3	42,3	36,4	1,3	7,7	7,6	2,6	0,3	73
Boccea	47,9	6.902	1.208	622	1,2	77,4	90,1	2,7	9,4	12,0	2,3	1,8	89
Borghesiana	23,7	43.135	7.593	5.463	0,3	75,9	126,6	3,9	12,7	12,9	3,6	0,9	369
Bufalotta	14,0	6.124	1.043	448	0,3	90,7	73,2	2,4	11,0	9,5	3,3	0,3	57
Buon Pastore	6,8	29.463	3.756	2.022	0,8	189,0	68,6	1,5	8,5	7,9	1,7	1,5	86
Casal Bertone	1,3	15.304	1.619	890	1,1	255,1	58,2	1,8	8,8	9,5	1,4	1,4	53
Casal Boccone	5,7	11.769	1.736	461	1,4	116,7	39,2	2,3	8,8	7,6	2,3	1,2	44
Casal Bruciato	2,5	21.143	2.321	814	0,7	227,4	38,5	2,3	10,5	8,7	1,9	2,6	51
Casal de' Pazzi	4,9	26.123	3.001	1.590	0,8	186,1	60,9	2,0	10,0	8,5	1,9	2,5	102
Casalotti di Boccea	3,1	16.039	2.427	1.846	0,4	112,2	115,1	3,6	11,7	11,4	3,0	1,0	70
Casetta Mistica	3,3	813	129	195	4,9	110,9	239,9	7,5	6,8	16,1	2,5	25,2	26
Casilino	2,0	10.741	1.227	801	0,7	224,7	74,6	3,8	10,1	11,0	1,9	1,8	32

In the first column, we find the name of each zone (in the image we see only the first zones in alphabetical). Subsequent columns represent an indicator that we will map to a feature in a Pandas data frame:
- *Surface* (squared km)
- *Population*

- *Pop0-14* (Population with age less or equal 14 years)
- *Foreigners* (Number of)
- *Centrality index* (ratio between exiting commuters and incoming commuters)
- *Old Age Index* (Population 65+ old over total population)
- *School Dropout Rate*
- *Unemployment Rate*
- *Neet Rate* (Percentage of population aged 15-29 not working, not studying, not looking for a job)
- *Economic Discomfort Index* (Percentage of families with sons, with no member with a job)
- *Cultural Employees Index*

(Note: data from 2011 Italian Census)

2. location data

For this category of data, we can leverage Foursquares APIs in order to retrieve:
- venues in every neighborhood, using the explore endpoint and the coords retrieved on previews step
- category of each venue from the response, in order to build the venues profile of each neighborhood
- as we are not intrested in detailed venues categories, we will perform our analysis on macro-categories of venues, as defined on Foursquare categories page
- for geo-localization of the zones, we will use the shapefiles of Rome's territory provided by the web-site #mapparoma: https://www.mapparoma.info/zone-urbanistiche/, witch we will handle through geopandas geo-spatial plotting library and folium

Methodology

Data acquisition

As stated, we extract socio-economic data from the cited paper and store data on a Pandas dataframe. This is a preview of the results:

	Name	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
0	Acilia Nord	9.20	26023	4241	2008	0.20	100.20	77.20	2.10	12.40	10.90	2.90	0.70
1	Acilia Sud	7.10	23640	3797	1833	0.70	107.30	77.50	2.70	12.00	11.30	3.10	2.50
2	Acqua Vergine	11.30	5355	1033	487	0.80	48.20	90.90	2.20	7.80	10.90	1.90	0.40
3	Acquatraversa	1.40	8689	1266	678	0.30	129.20	78.00	1.10	6.10	7.20	2.60	2.90
4	Aeroporto dell' Urbe	4.40	1924	247	172	11.00	172.10	89.40	1.60	11.80	10.00	1.90	3.30

Next, we retrieve geo-localization data (coordinates) for each zone in the data set, by downloading shapefiles from www.mapparoma.info and using geopandas Python library to read the file and calculate centroids for each zone:

	Name	Latitude	Longitude
0	Tuscolano Nord	41.88	12.52
1	Tuscolano Sud	41.87	12.53
2	Tor Fiscale	41.86	12.54
3	Appio	41.88	12.51
4	Latino	41.87	12.51
...	...	...	...
150	Acqua Vergine	41.91	12.64
151	Lunghezza	41.91	12.68
152	Torre Angela	41.88	12.64
153	Borghesiana	41.86	12.69
154	S. Vittorino	41.91	12.75

For venues data, we use Foursquare *explore* API which provides venues near to the centroids of each zone, with category for each venue. Then, we collapse categories to each ancestor "macro category" as defined by Foursquare, as we do not need detailed analysis on sub-categories

Here's a preview of the resulting data frame:

name	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Total Venues
Acilia Nord	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.04	0.00	0.15
Acilia Sud	0.04	0.00	0.00	0.34	0.04	0.04	0.00	0.00	0.04	0.00	0.51
Acqua Vergine	0.19	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.19	0.56
Acquatraversa	0.00	0.00	0.00	0.12	0.00	0.58	0.00	0.00	0.12	0.23	1.04
Aeroporto dell' Urbe	0.00	0.00	0.00	0.52	0.52	0.52	0.00	0.00	0.00	1.56	3.12
...	...	...	...	...	...	...	...	...	...	...	...
Vallerano Castel di Leva	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.04

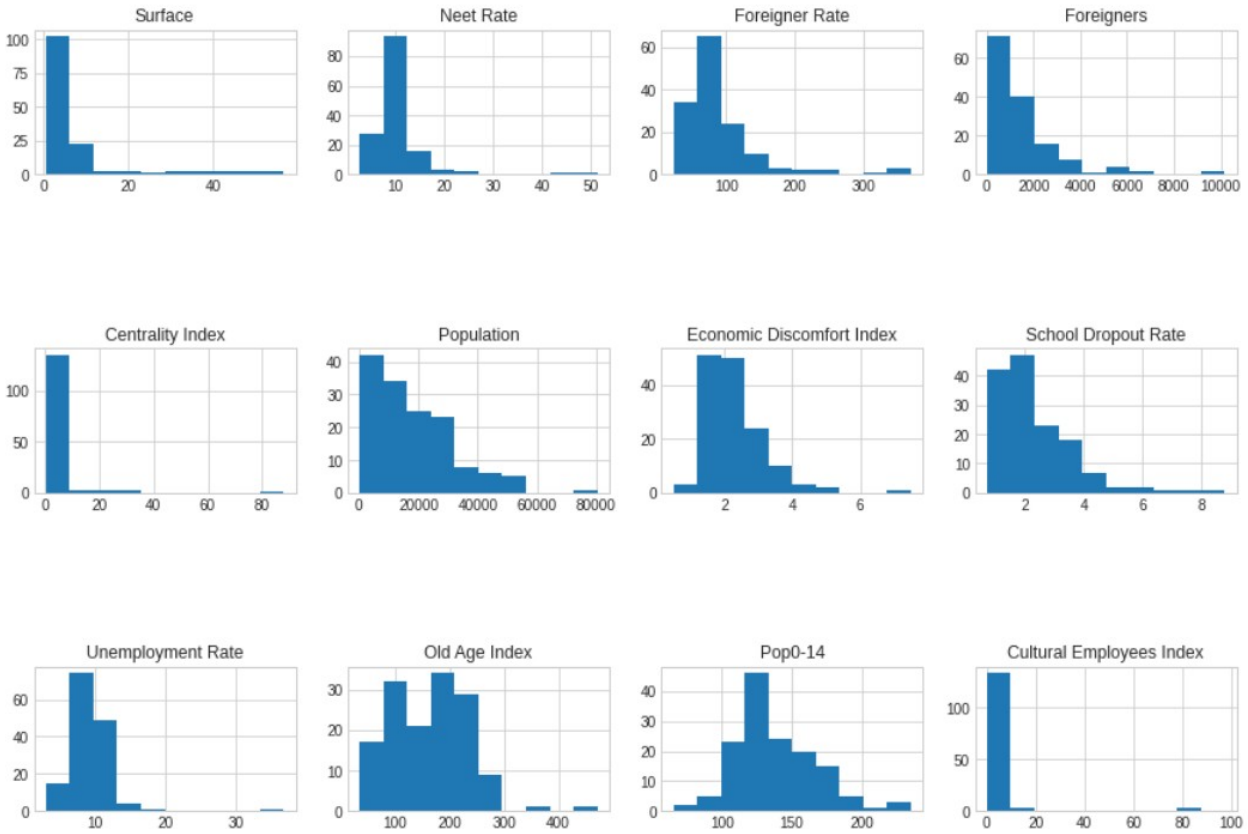
Explorative data analysis

First, we show summary statistics on socio-demographic 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

Summary statistics of socio-demographic features

Next, we display histograms of their distributions:



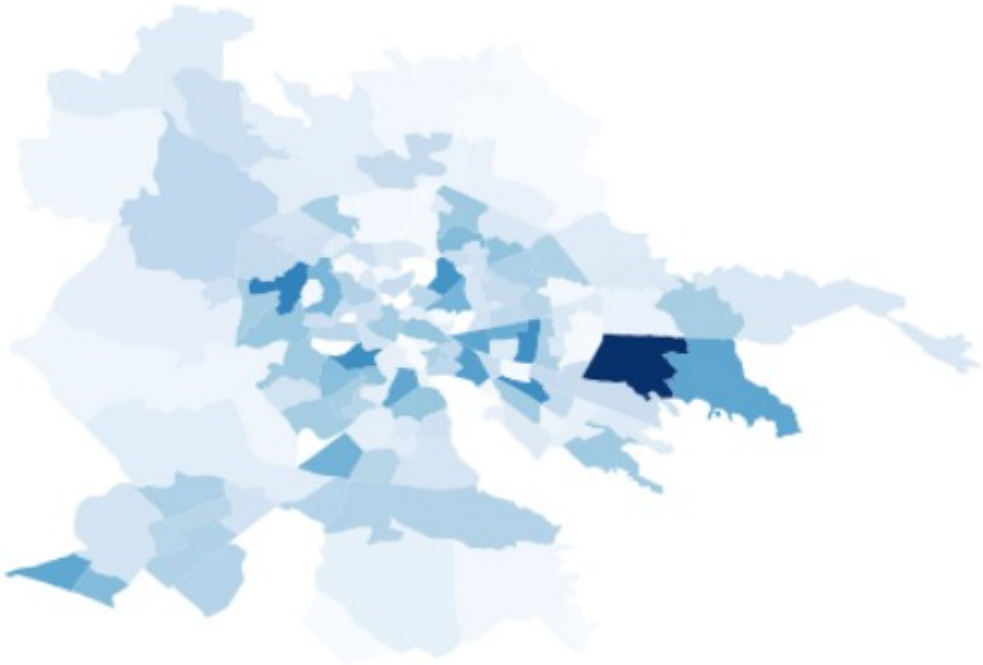
Distributions of socio-demographic features

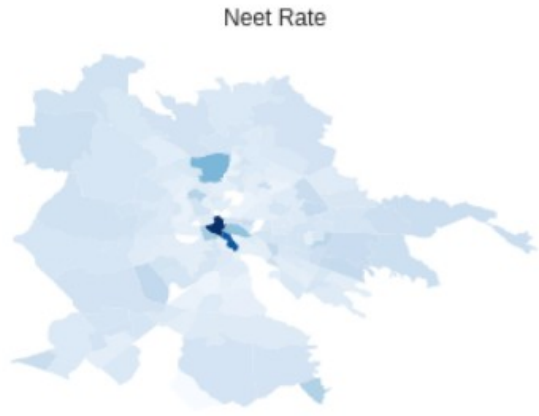
Observations:

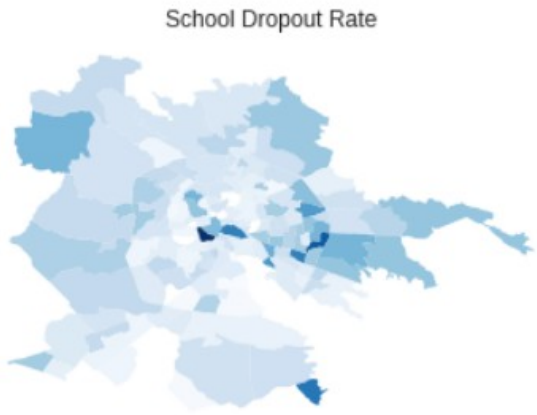
- We see high variability in Size and population
- Age indicators are also very variable, with a median of 132 pop 0-14 per 1,000 residents, but with minimum of 65 through a maximum of 235. This means we have very young zones beside very old ones
- Economic indicators such as Unemployment Rate, Neet Rate, Economic Discomfort Index, tend to be less dispersed, but show very high spikes that would need some further investigation (left for projects to come...)

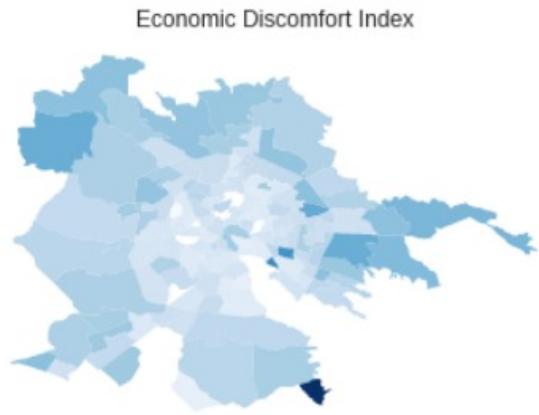
Territorial distribution is shown in figures below:

Population









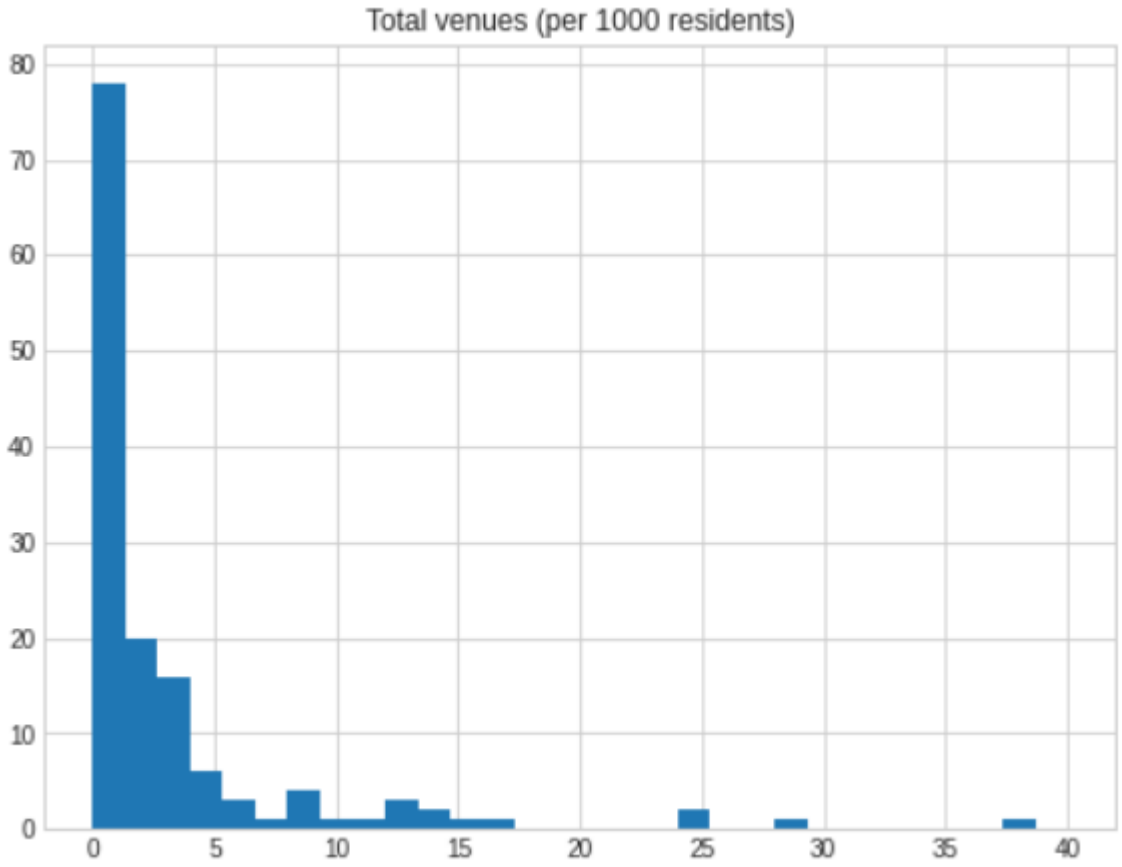
Here we see as economic indicators tend to perform worst on periphery zones, with some interesting exceptions.

For venues data, unfortunately they tend to be generally very low. So, in the next sections, we'll concentrate on "Total Venues" feature

Here we present some summaries:

	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

Summary statistics on venues data



Distribution of total venues per 1,000 residents

Cluster analysis

For our main purpose, we choose to perform Cluster analysis on our zones data set

As pre-processing step, we decided to:

- remove *Latitude*, *Longitude* and *Surface* features, as they are of no interest for this task

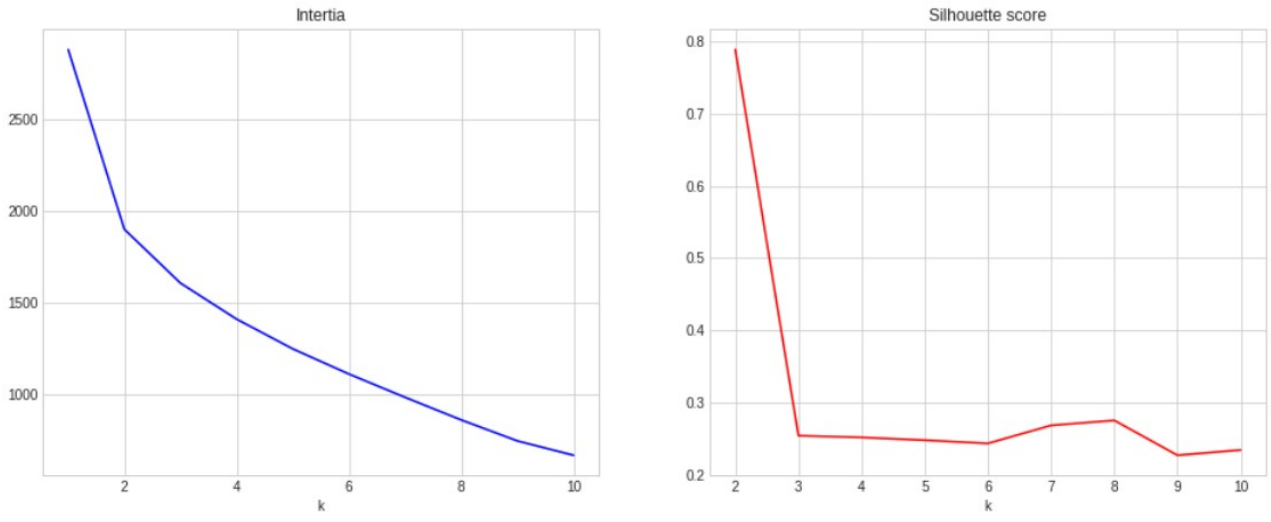
- remove *Foreigners* feature, as this feature is collinear with *Foreigner Rate* and *Population*

- perform standardization on remaning features via scikitlearn's Standard Scaler class ((X - μ) / σ)

As clustering Algorythm, we opted for **K-Means** clustering.

In order to find the best number of clusters, we performed 10 runs of clustering with increasing value of k (1..10), calculated *Inertia* and *Silhouette score* metrics for each k in (1..10) and evaluated results via visual inspection of the resulting plots ("elbow" method)

In the next figure, we show the metrics results (Inertia and Silhouette):



Observations:

- Applying the *elbow* method on Inertia plot doesn't provide istantaneous result, but we can say that the optimal number of cluster lies between 2 and 4

- Silhouette scores show that there are 2 clear clusters, but for our purposes this number is too low

Further experiments (we don't show here these results) showed that we can choose **3** as the optimal number of clusters, so we go on with clustering our zones in 3 clusters

Next, we present results of the clustering procedure:

	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

Summaries of cluster analysis

Observations:

> We can see that our zones are bucked in three distinct groups:

* "Cluster 0": poor socio-economic indicators and shortage of available venus

* "Cluster 1": better socio-economic indicators and venues availability

* "Cluster 2": a very tiny cluster, with only 3 zones with very low population and sufrace and singular behaviour

> For our purposes, we can say that

* Cluster 2 has no interest, as it's probably composed of singular zones that can be left apart for our analysis

* Cluster 0 is what we are looking for, as it's a huge cluster of zones sharing these characteristics:

    * High surface

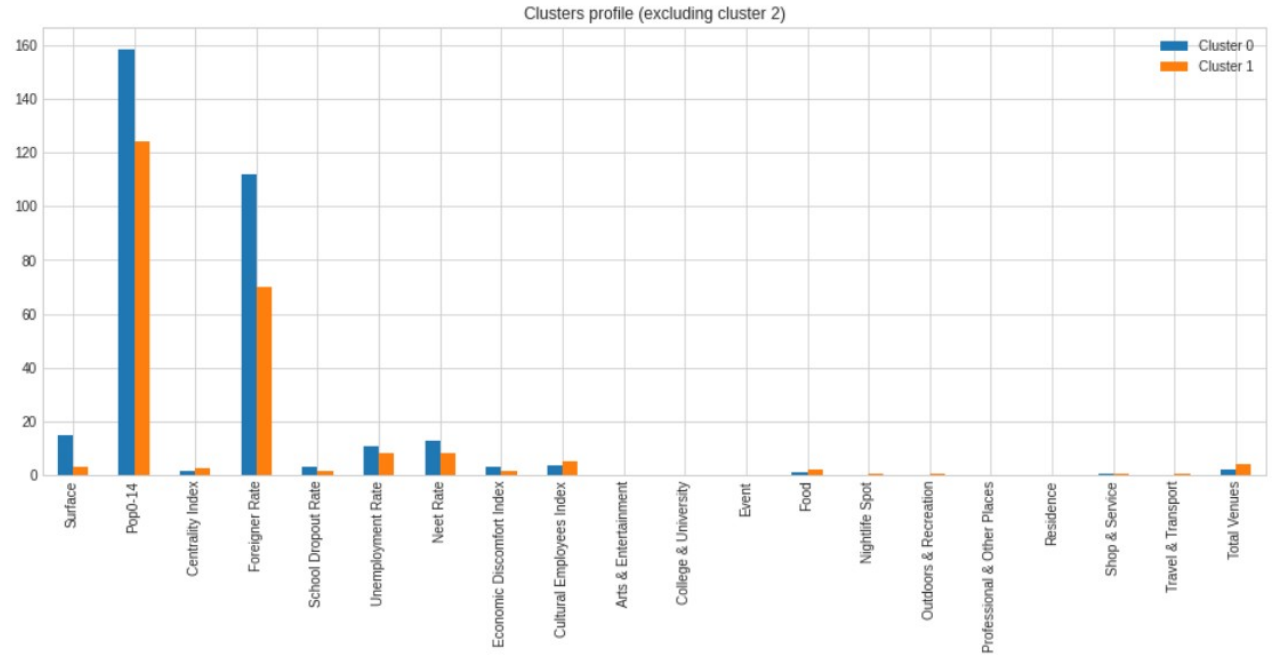
    * Relatively young population

    * Low centrality (periphery zones)

    * School Droput, Unemployment and Economic Discomfort above the mean

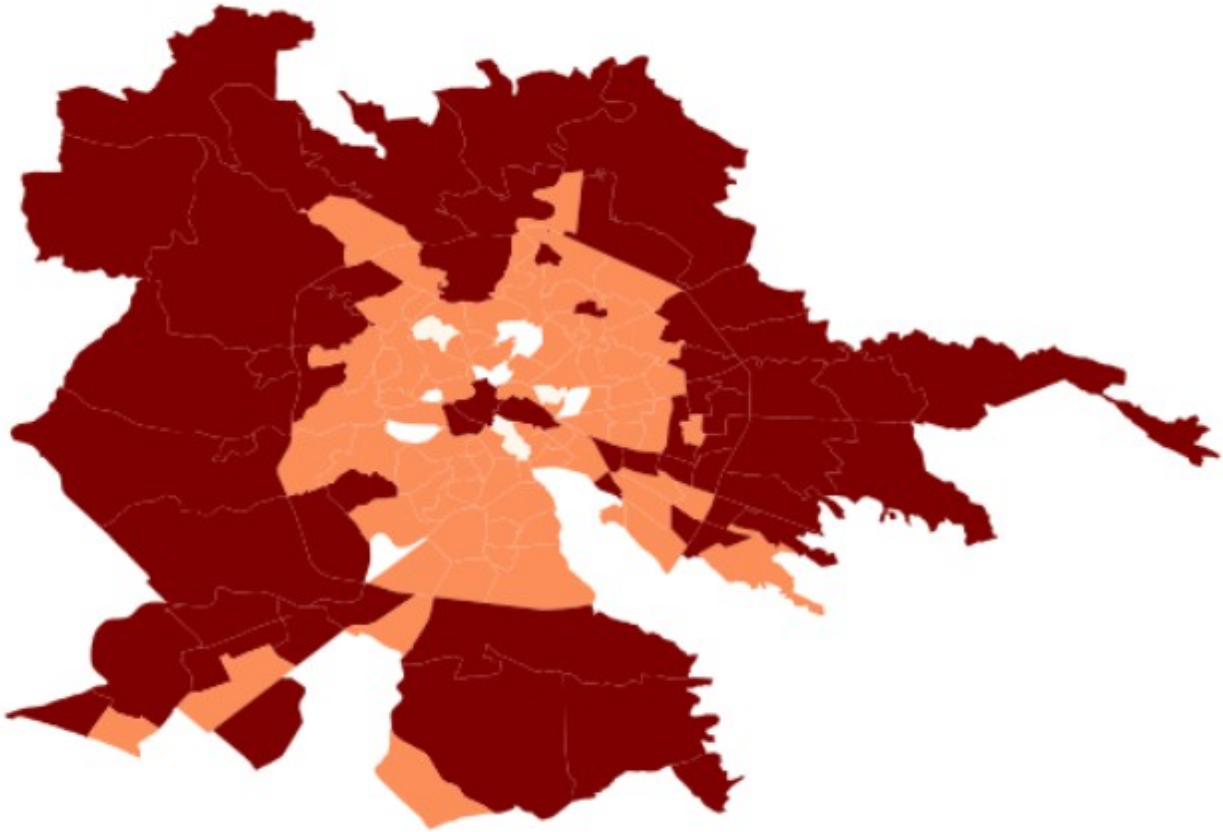
    * Venues availability below the mean

So, considering only clusters 0 and 1, we can plot their respective profiles:



We see that the plot confirms our first impression

Now, let's take a look of the territorial distribution of our clusters



Territorial distribution of clusters

There's a quite strong evidence that zones in cluster 0 (**dark colored**) are mostly localized at the periphery of the city

Having decided to point our attention on Cluster 0, it has to be sad that this is still a huge cluster with 60 zones.

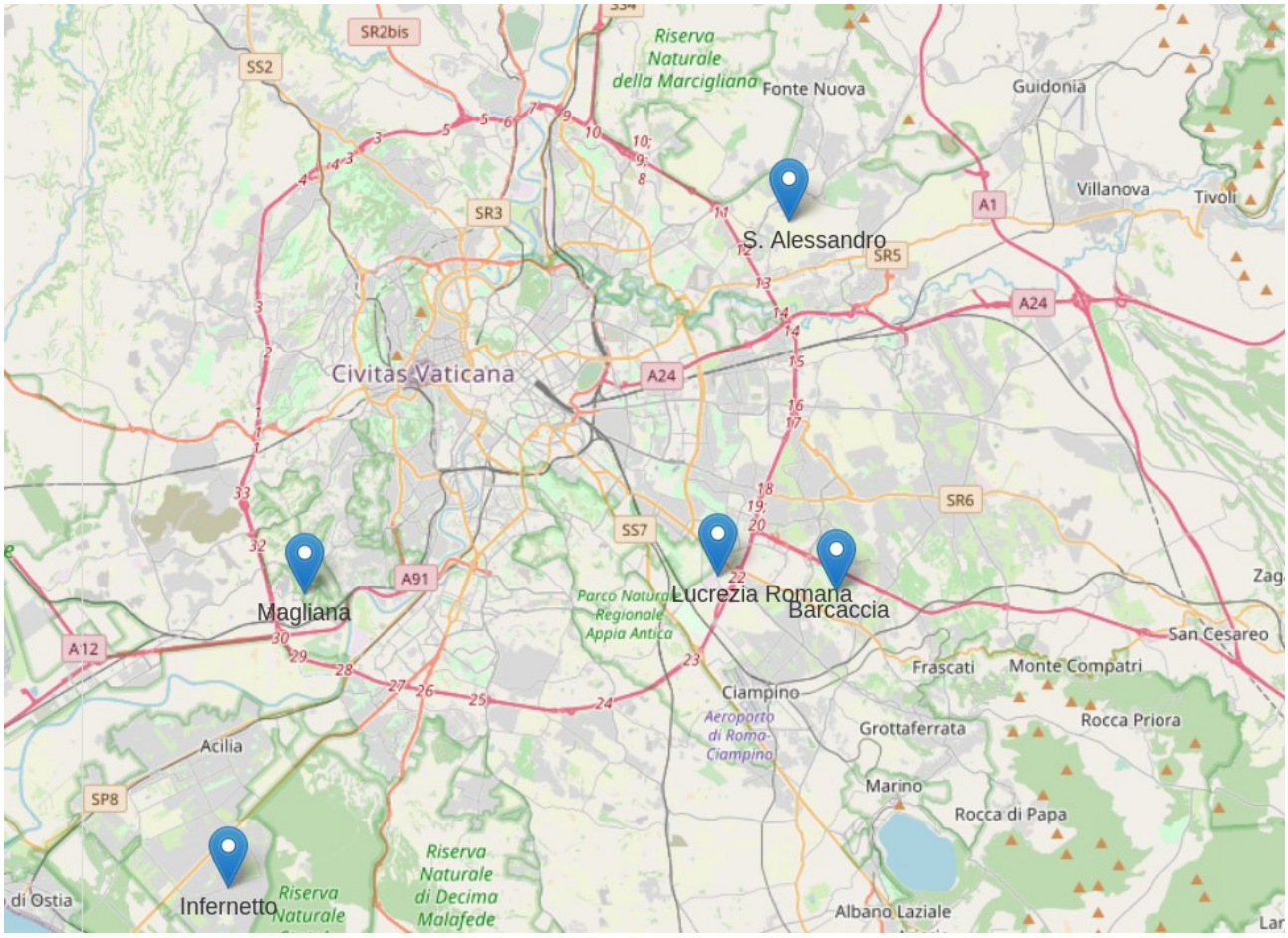
So, it's of much interest to find a further criterion to sub-divide cluster 0 and find most interesting zones inside it

This can be a task for other clustering algorythms, such as Hyerarchical (agglomerative) clustering. But for now, we keep it simple: as our customer's target is younger people, we simply sort our cluster on Pop0-14 rate and conclude suggesting the 5 youngest zones, as shown below:

name	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
S. Alessandro	11.40	9856	235.09	474	0.40	36.20	48.10	1.40	6.60	8.50	2.40	2.40
Magliana	11.50	3803	223.24	195	8.30	34.90	51.30	1.70	5.60	15.40	1.80	1.70
Barcaccia	5.10	10099	220.52	368	0.30	42.30	36.40	1.30	7.70	7.60	2.60	0.30
Infernetto	11.50	24356	210.22	1750	0.20	63.60	71.90	1.10	8.00	9.60	3.10	1.50
Lucrezia Romana	1.70	4451	198.38	228	1.40	56.10	51.20	1.30	7.20	7.10	1.60	0.80

Five zones in cluster 0 with highest youth index

Geo-location of these five zones is shown in the map below:



Geolocation of the 5 youngest zone in cluster 0

Results

As primary result, we identified a main cluster (*Cluster "0"*) of Rome's urbanistic zones ("neighborhoods") which we find particulary suitable for a social center

The 60 zones in this cluster are shown below:

Acilia Nord, Acilia Sud, Acqua Vergine, Alessandrina, Barcaccia, Boccea, Borghesiana, Bufalotta, Casalotti di Boccea, Casetta Mistica, Castelluccia, Centro Direzionale Centocelle, Centro Storico, Cesano, Corviale, Decima, Esquilino, Fidene, Fogaccia, Giardinetti-Tor Vergata, Gregna, Grottarossa Ovest, Infernetto, La Rustica, La Storta, Labaro, Lucrezia Romana, Lunghezza, Magliana, Malafede, Massimina, Mezzocammino, Omo, Ostia Antica, Ostia Nord, Ottavia, Pantano di Grano, Ponte Galeria, Porta Medaglia, Prima Porta, Quadraro, Romanina, S. Alessandro, S. Basilio, S. Cornelia, S. Maria della Pietà, S. Maria di Galeria, S. Vittorino, Santa Palomba, Settecamini, Tor Cervara, Tor Fiscale, Tor S. Giovanni, Tor Sapienza, Torre Angela, Torre Maura, Torrespaccata, Trastevere, Tufello, Vallerano Castel di Leva

These are the main statistics on this cluster

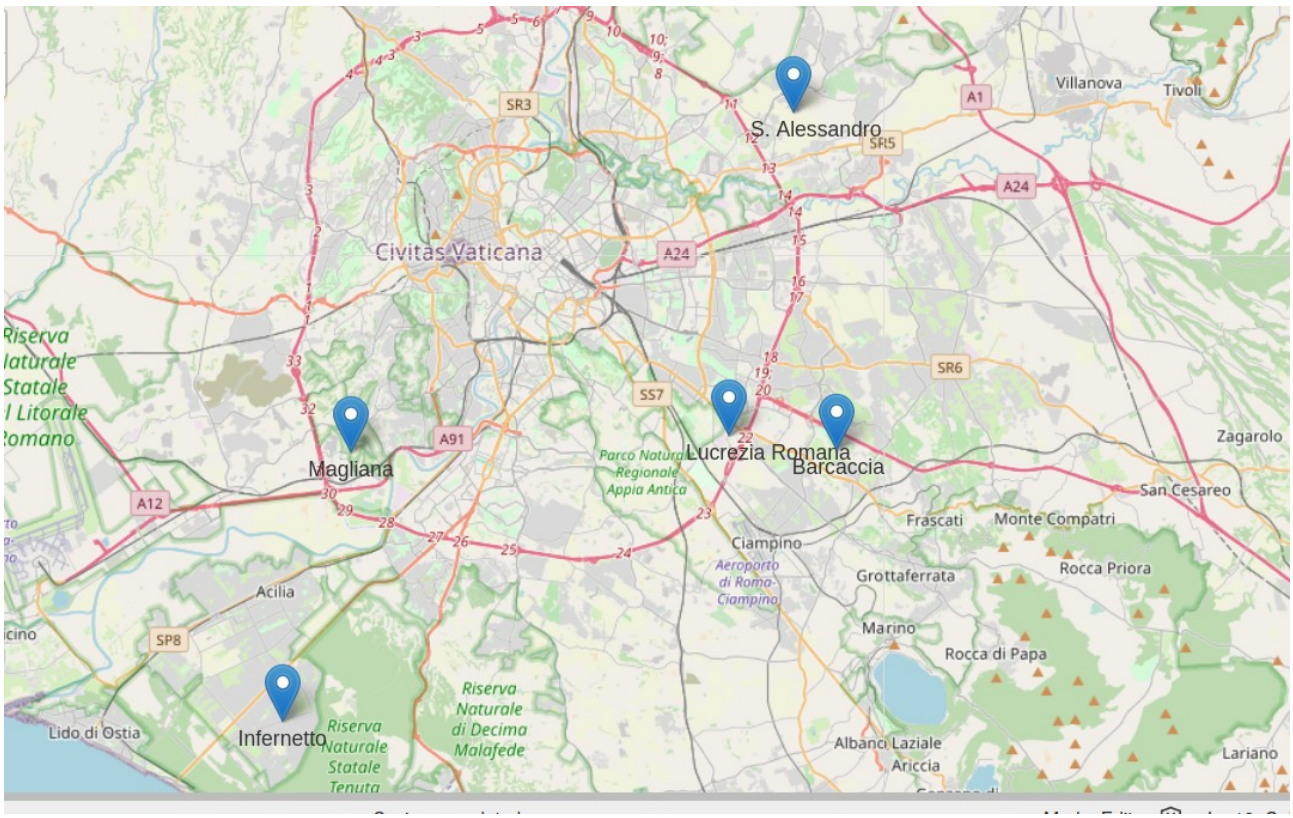
	Surface	Population	Pop0-14	Centrality Index	Foreigner Rate	School Dropout Rate	Unemployment Rate	Neet Rate	Economic Discomfort Index	Total Venues
mean	14.67	14715.45	158.54	1.85	112.00	3.26	10.99	12.67	2.90	2.11

Mean values for cluster 0

Among these 60 zones, we further recommend to begin opening social centers in these 5 zones, which show the highest rate of young population:

- S. Alessandro
- Magliana
- Barcaccia
- Infernetto
- Lucrezia Romana

And this is the location of these zones



Discussion

There are some points of improvement that are wortn-noting:
- First of all, we can retrieve more data on economic factors. For example, scraping from www.mercato-immobiliare.info (a web-site providing house prices for Italy at a very fine-grained territorial level) we can retrive sales price and rent price for houses and apartements.
- As stated in Methodology section, our *Cluster 0* is still a very big cluster, with 60 zones. We can try an alternative clustering approach, such as *Hyerarchical agglomerative clustering* in order to find more narrow clusters where to point our attention
- The choice of top 5 recommended zones was done based on arbitraty sorting by youth rate. We can think about a better criterion
- Clustering algorythms don't tell us the whole story. It would be very interesting (especially when more features will become available) to implement some analytic technique which can provide more insights. A good candidate is *Factor Analysis*, which can be useful in finding hidden factors describing our data
- scraping venues data from Foursquare hasn't provided great results. We can try with other sources of venues data such as Google's

Conclusions

In this project we were able to identify a big cluster of neighborhoods in the city of Rome where the need for a social center is particularly high.

We scraped data from a paper published by Italian Institute of Statistics (Istat) to retrieve socio-demographic information on the 155 Urbanistic zones, using *tabula-py* Python library for pdf scraping, and retrieved informations on venues available in each zone using Fousquare *explore* API. Furthermore, we used shapefiles from www.mapparoma.info, *geopandas* and *folium* Python libraries to geo-locate the zones and display geo-located results

We applied *k-means* cluster analysis using scikitlearn's *cluster* module as clustering algorythm and grouped Rome's neighbohoods in 3 clusters. One of these clusters (*"Cluster 0"*) was identified as the most suited for our customer's purpose

Finally, we selected the 5 most suitable zones among *Cluster 0* by simply sorting by youth rate, and provide this zones as recommended zones where to start-up our customer's project

In the discussion section, we also provided suggestions for further development and investigation

References

- Istituto nazionale di statistica: www.istat.it
- Mapparoma: https://www.mapparoma.info/
- On cluster analyses: https://en.wikipedia.org/wiki/Cluster_analysis
- gitHub repo of this project: https://github.com/MicheleGava/battle_of_neighborhoods_w1