Coursera Capstone Project: Clustering of Rome Subway stations

Sergei Rastompakhov

Rome, Italy

Introduction

The Rome Metro (Metropolitana di Roma) started operation in 1955. It has three lines: A (orange), B (blue) and C (green). It has 73 stations, the length of it is about 60 km. The lines A and B intersect at Termini Station, which is also the main train station in Rome. Annually it serves for 279 million of people (2012).

We want to look at the areas surrounding metro stations and cluster them. The areas in the historical center of Rome have a lot of touristic attractions (Vatican, Colosseum, Trevi Fountain, Spanish Steps) and hotels. Some neighborhoods are in residential areas, other in commercial. The venues near to the station determine how people use it.

Analysis of data can show the primary purpose of the station. This data is useful for commercial business (location to open a new business), for local government (can help them in city planning), and for subway development in the future (where open new stations).

Data

We need several pieces of data to start our study.

1. List of all subway stations and their geo data (latitude, longitude). The list of station is easy to get from [Wikipedia](https://it.wikipedia.org/wiki/Stazioni_della_metropolitana_di_Roma). However, this table does not contain geo data. To get geo data we can use prepared csv file [github](https://raw.githubusercontent.com/riccione/Coursera_Capstone/master/Rome_Metro.csv), which contains list of the stations and geo data. By merging two sources of data we can get this dataframe. Also we need to do some data preparation: clean it, remove duplicates and sort it by name of the Station in lexicographical order.

	Station	Opened	Type	Line	Latitude	Longitude
0	Alessandrino	2014	underground	C	41.871349	12.578701
1	Anagnina	1980	underground	A	41.842778	12.586111
2	Arco di Travertino	1980	underground	A	41.866705	12.535070
3	Baldo degli Ubaldi	2000	underground	A	41.898889	12.432778
4	Barberini - Fontana di Trevi	1980	underground	A	41.903889	12.488889
5	Basilica San Paolo	1955	overground	B	41.856111	12.478194
6	Battistini	2000	underground	A	41.906461	12.414722
7	Bologna	1990	underground	B	41.913333	12.520556
8	Bolognetta	2014	overground	C	41.865134	12.680883
9	Borghesiana	2014	in trincea	C	41.864707	12.667300

2. Foursquare API to explore venue types surrounding each station. Foursquare outlines these high-level venue categories with more sub-categories.
    - Arts & Entertainment (4d4b7104d754a06370d81259)
    - College & University (4d4b7105d754a06372d81259)
    - Event (4d4b7105d754a06373d81259)
    - Food (4d4b7105d754a06374d81259)
    - Nightlife Spot (4d4b7105d754a06376d81259)
    - Outdoors & Recreation (4d4b7105d754a06377d81259)
    - Professional & Other Places (4d4b7105d754a06375d81259)
    - Residence (4e67e38e036454776db1fb3a)
    - Shop & Service (4d4b7105d754a06378d81259)
    - Travel & Transport (4d4b7105d754a06379d81259)

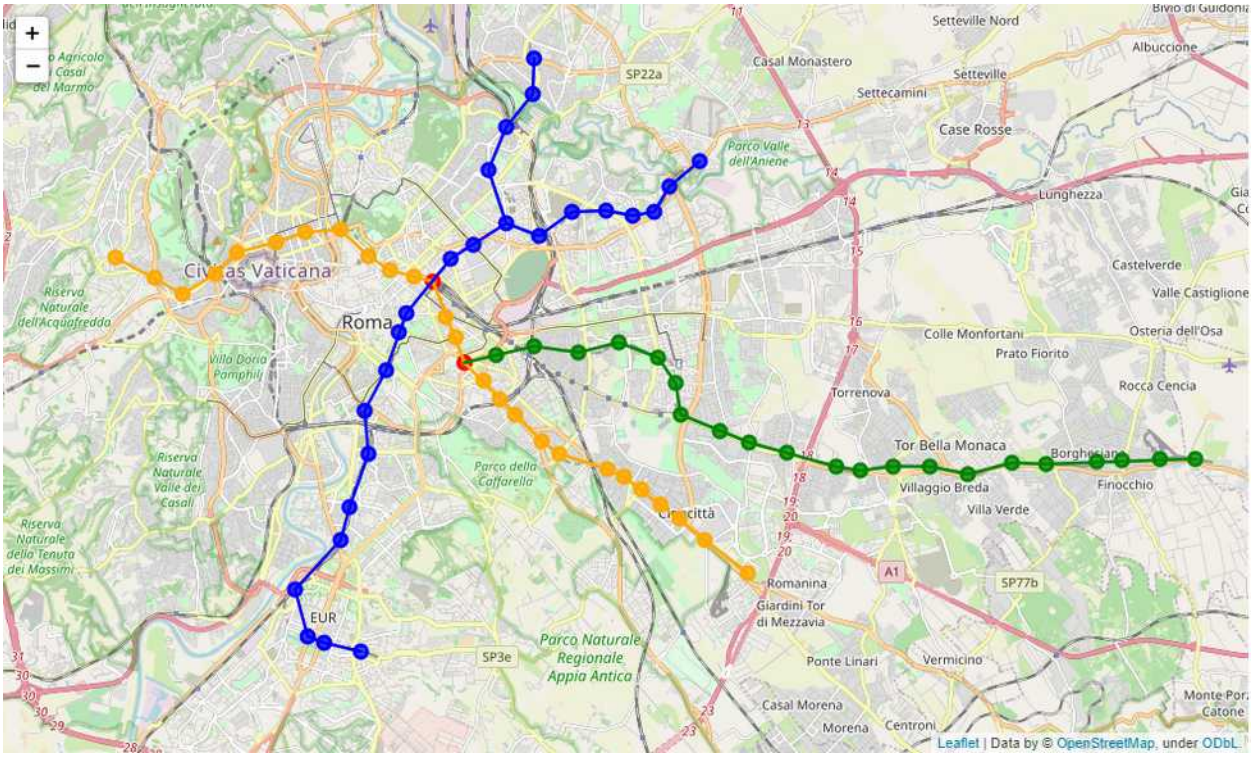
    We will query the number of venues in each category in a 1000m (walking distance) radius around each station. This will be the first way to analysis data, because I would like to have comparable data with similar project related to the Moscow subway [Classification of Moscow Metro stations using Foursquare data](https://towardsdatascience.com/classification-of-moscow-metro-stations-using-foursquare-data-fb8aad3e0e4). Thank you for creation study against Moscow metro, it inspired me.

For limitations of free Foursquare API account all data will be saved in csv files, for future use. All data necessary for the notebook is available on github.

Methodology

Before we start to a specific analysis, lets visualize our data and get some basic knowledge about it. Using Folium we build the map of Rome with metro stations on it, the center of the map we choose the location of Barberini - Fontana di Trevi. This is very nice picture of it, because usually it is crowded with tourists.

Our map looks like this.



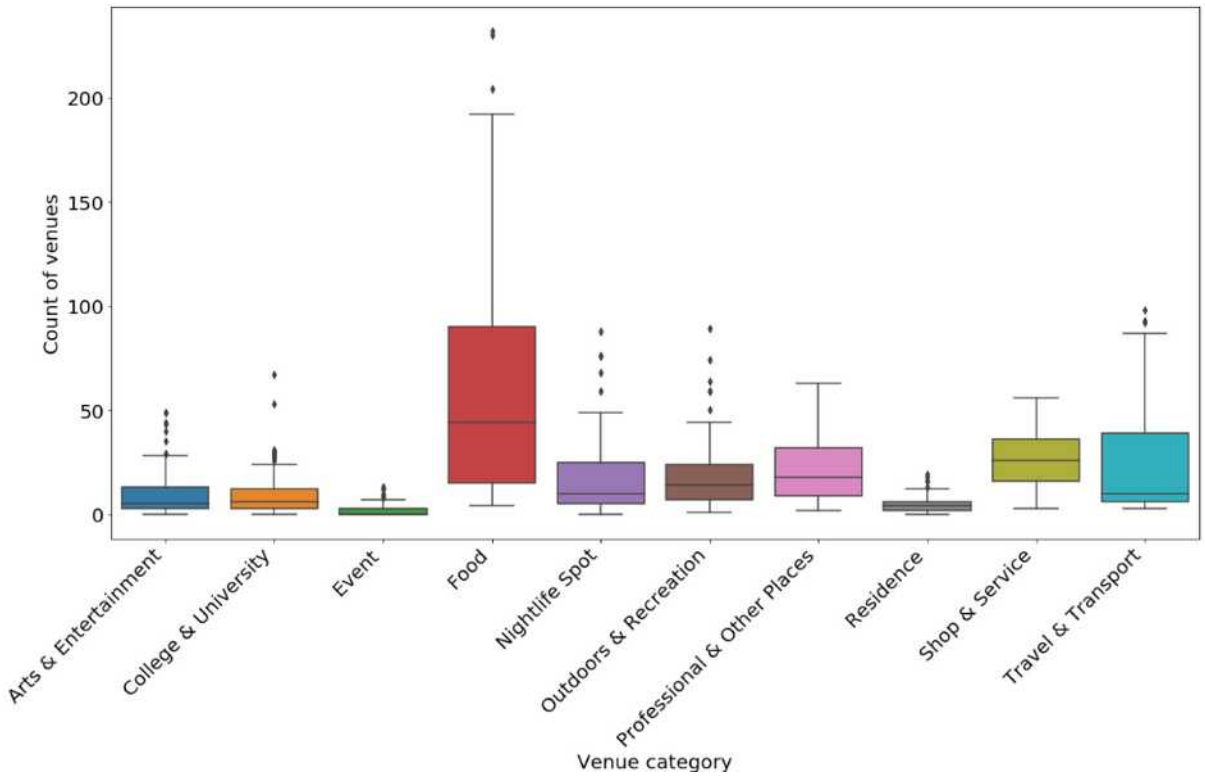
First, we use Foursquare API to get venues by categories for each station. The list of categories we can get using API call or hard coded them. As discussed above we select radius equal to 1000 meters. The API call straight forward and well documented on the official website. It returns the necessary data for us, which is total number of venues by a specific category. We need to parse JSON and extract this value and update our dataframe.

	Station	Opened	Type	Line	Latitude	Longitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence
0	Alessandrino	2014	underground	C	41.871349	12.578701	4	7	1	7	5	8	6	3
1	Anagnina	1980	underground	A	41.842778	12.586111	3	3	0	9	4	1	3	0
2	Arco di Travertino	1980	underground	A	41.866705	12.535070	5	5	0	37	16	17	16	6
3	Baldo degli Ubaldi	2000	underground	A	41.898889	12.432778	13	8	1	49	5	19	28	4
4	Barberini - Fontana di Trevi	1980	underground	A	41.903889	12.488889	49	27	9	230	88	74	63	5

Using describe method helps us to understand the data better.

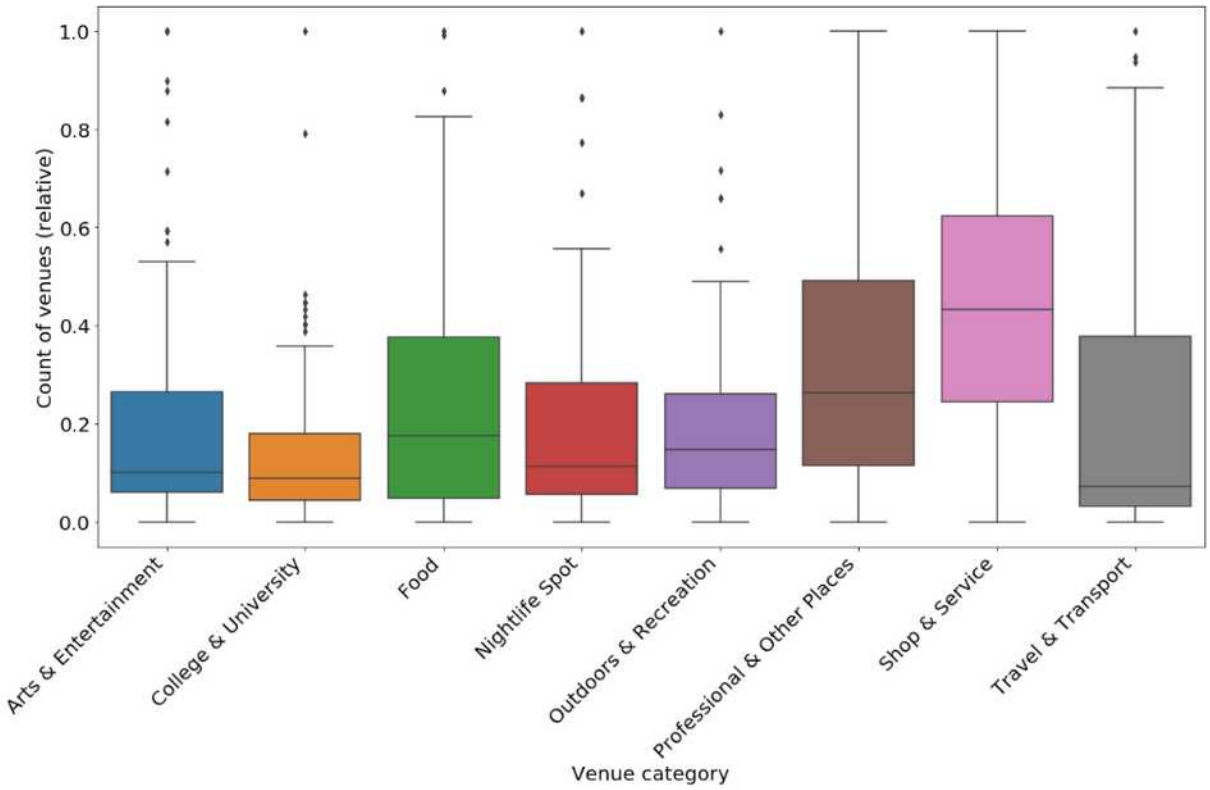
	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
count	73.000000	73.000000	73.000000	73.000000	73.000000	73.000000	73.000000	73.000000	73.000000	73.000000
mean	10.438356	10.260274	1.986301	60.767123	18.643836	19.794521	21.945205	4.904110	26.273973	24.191781
std	12.225868	12.225914	3.255308	58.152224	19.896517	17.933487	15.730270	4.308146	13.507467	26.187177
min	0.000000	0.000000	0.000000	4.000000	0.000000	1.000000	2.000000	0.000000	3.000000	3.000000
25%	3.000000	3.000000	0.000000	15.000000	5.000000	7.000000	9.000000	2.000000	16.000000	6.000000
50%	5.000000	6.000000	0.000000	44.000000	10.000000	14.000000	18.000000	4.000000	26.000000	10.000000
75%	13.000000	12.000000	3.000000	90.000000	25.000000	24.000000	32.000000	6.000000	36.000000	39.000000
max	49.000000	67.000000	13.000000	232.000000	88.000000	89.000000	63.000000	19.000000	56.000000	98.000000

Using matplotlib and seaborn we can create boxplot for our data.



As you see, the number of Event and Residence categories are not so high. Also by studying what exactly included in each of these categories it is obvious that we can drop them. The contents of the categories is possible to study here https://developer.foursquare.com/docs/build-with-foursquare/categories/

It is better to normalize our data and plot it again. Normalization will keep the same scaling and won't change the data relation. We use Min-Max Scaling.



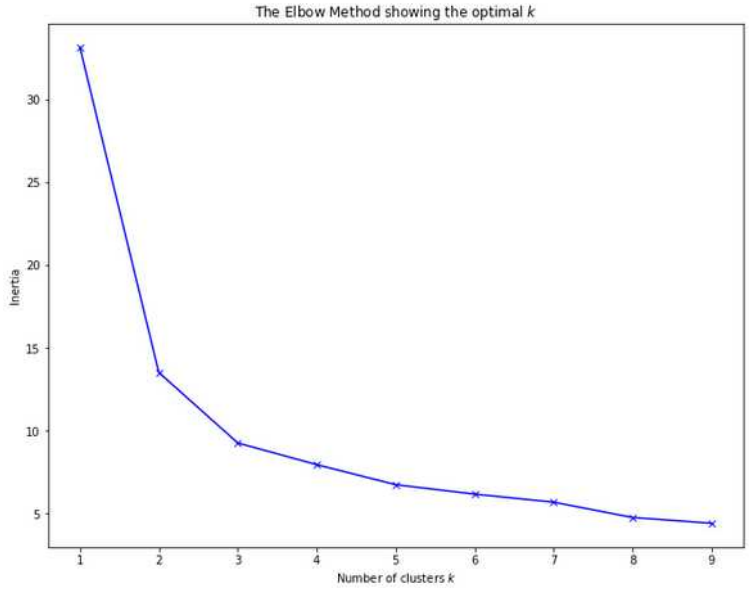
Clustering

Let's cluster our data. How many clusters to select? This is a good question, it is possible to use just a visualization and common sense. But we choose different approach.

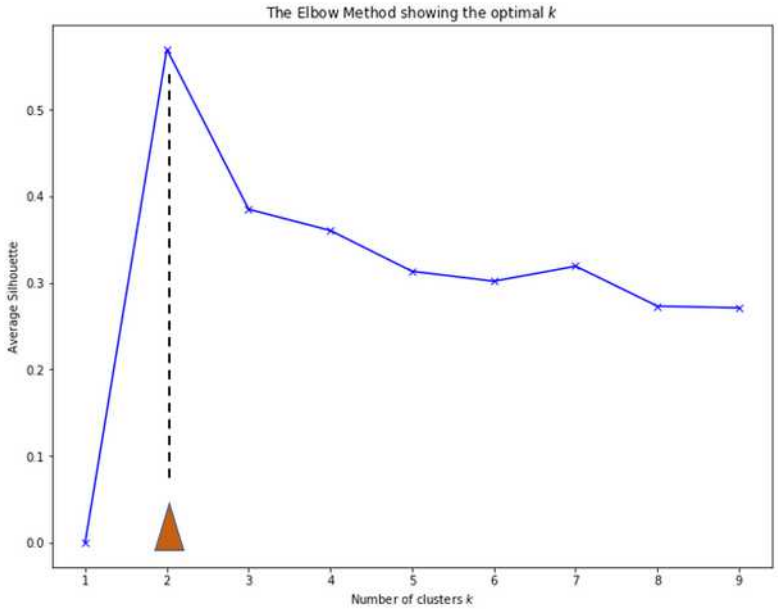
There are two popular methods, which helps to select the optimal number of clusters:

1. The Elbow method. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.

So, optimal number of clusters based on 'Elbow method' is 3 for our dataframe. Note that, the elbow method is sometimes ambiguous.

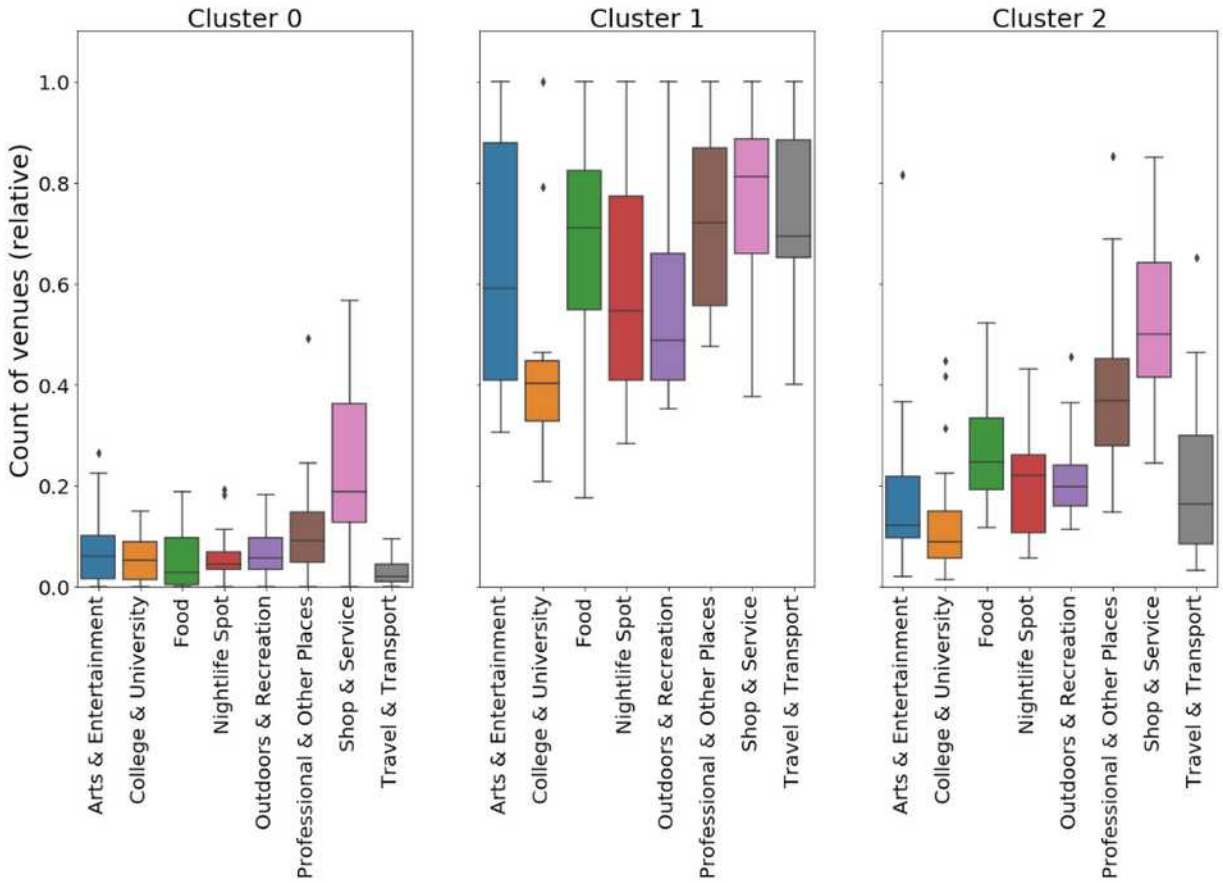


2. The Silhouette method. An alternative is the average silhouette method (Kaufman and Rousseeuw, 1990) which can be also used with any clustering approach. Let's try to use different approach to find optimal number of clusters – Silhouette method. The Silhouette Score reaches its global maximum at the optimal k.

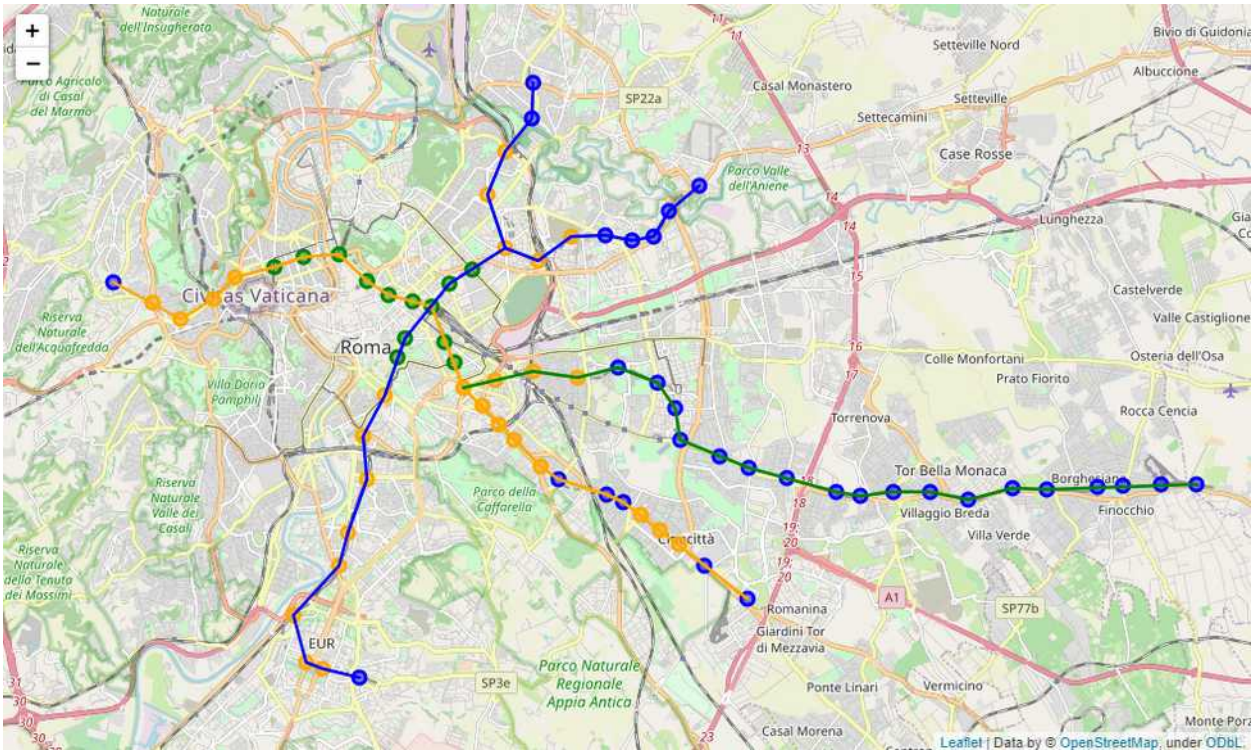


The Silhouette method shows, that 2 is optimal choice for number of clusters. However observation and analysis showed that 2 is not enough, so we are going to use value – 3.

Finally, we cluster our data to 3 clusters. Lets, visualize them with boxplots.



Let's create map to visualize clusters, where cluster 0 is blue, cluster 1 is green, cluster 2 is orange.



Results

Cluster 1 is a historical and touristic part of the city. Cluster 0 is mainly residential area, and cluster 2 is borderline between cluster 0 and cluster 1.

Cluster 0 has many venues in category Shop & Service.

Cluster 1 has a mix of venues in different categories.

Cluster 2 is very similar to cluster 0, it has many venues in category Shop & Service.

Let's create data sets for each cluster and show the top 3 categories for each station.

So, basically, we can have just two clusters, as silhouette method showed us. However I used 3 clusters, because with 2 clusters stations close to Vatican clustered incorrectly (they should be clustered as a historical part of the city).

Everything important for foreigner/tourist is located in the cluster 1.

Cluster 0

	Station	Max1	Max2	Max3
0	Alessandrino	Shop & Service	College & University	Arts & Entertainment
1	Anagnina	Shop & Service	Travel & Transport	Arts & Entertainment
2	Arco di Travertino	Shop & Service	Professional & Other Places	Nightlife Spot
6	Battistini	Shop & Service	Outdoors & Recreation	Professional & Other Places
8	Bolognetta	Shop & Service	Professional & Other Places	Nightlife Spot
9	Borghesiana	Shop & Service	Professional & Other Places	Nightlife Spot
12	Cinecittà	Shop & Service	Professional & Other Places	College & University
17	Conca d'Oro	Shop & Service	Nightlife Spot	Food
19	Due Leoni-Fontana Candida	Professional & Other Places	Shop & Service	Outdoors & Recreation
23	Finocchio	Shop & Service	Professional & Other Places	Travel & Transport
27	Gardenie	Shop & Service	Arts & Entertainment	Food
28	Giardinetti	Shop & Service	Arts & Entertainment	Outdoors & Recreation
30	Graniti	Shop & Service	Travel & Transport	Outdoors & Recreation
31	Grotte Celoni	Professional & Other Places	Outdoors & Recreation	Shop & Service
32	Jonio	Shop & Service	Outdoors & Recreation	Food
33	Laurentina	Professional & Other Places	Shop & Service	College & University
41	Mirti	Shop & Service	Arts & Entertainment	Food
42	Monte Compatri-Pantano	Outdoors & Recreation	Travel & Transport	Food
43	Monti Tiburtini	Shop & Service	Professional & Other Places	Outdoors & Recreation
44	Numidio Quadrato	Shop & Service	Food	Outdoors & Recreation
46	Parco di Centocelle	Arts & Entertainment	Shop & Service	Nightlife Spot
47	Pietralata	Shop & Service	Professional & Other Places	College & University
52	Ponte Mammolo	Professional & Other Places	College & University	Arts & Entertainment
53	Porta Furba - Quadraro	Shop & Service	Food	Outdoors & Recreation
56	Rebibbia	Arts & Entertainment	Professional & Other Places	College & University
59	Santa Maria del Soccorso	Shop & Service	Professional & Other Places	College & University
63	Teano	Shop & Service	Professional & Other Places	Outdoors & Recreation
66	Torre Angela	Shop & Service	Professional & Other Places	Arts & Entertainment
67	Torre Gaia	Shop & Service	Professional & Other Places	College & University
68	Torre Maura	Shop & Service	College & University	Professional & Other Places
69	Torre Spaccata	Shop & Service	Professional & Other Places	Arts & Entertainment
70	Torrenova	Shop & Service	Nightlife Spot	Arts & Entertainment

Cluster 1

	Station	Max1	Max2	Max3
4	Barberini - Fontana di Trevi	Nightlife Spot	Professional & Other Places	Shop & Service
10	Castro Pretorio	Shop & Service	Travel & Transport	College & University
11	Cavour	Arts & Entertainment	Shop & Service	Professional & Other Places
16	Colosseo	Professional & Other Places	Food	Arts & Entertainment
24	Flaminio - Piazza del Popolo	Shop & Service	Travel & Transport	Arts & Entertainment
34	Lepanto	Food	Shop & Service	Nightlife Spot
39	Manzoni	Food	Shop & Service	Professional & Other Places
45	Ottaviano - San Pietro - Musei Vaticani	Arts & Entertainment	Shop & Service	Professional & Other Places
50	Policlinico	College & University	Travel & Transport	Professional & Other Places
57	Repubblica - Teatro dell'Opera	Food	Travel & Transport	Professional & Other Places
61	Spagna	Outdoors & Recreation	Arts & Entertainment	Professional & Other Places
64	Termini	Travel & Transport	Shop & Service	Professional & Other Places
72	Vittorio Emanuele	Nightlife Spot	Travel & Transport	Shop & Service

Cluster 2

	Station	Max1	Max2	Max3
3	Baldo degli Ubaldi	Shop & Service	Professional & Other Places	Travel & Transport
5	Basilica San Paolo	College & University	Professional & Other Places	Shop & Service
7	Bologna	Travel & Transport	Shop & Service	College & University
13	Cipro	Arts & Entertainment	Shop & Service	Professional & Other Places
14	Circo Massimo	Professional & Other Places	Outdoors & Recreation	Food
15	Colli Albani	Shop & Service	Professional & Other Places	Nightlife Spot
18	Cornelia	Shop & Service	Professional & Other Places	Arts & Entertainment
20	EUR Fermi	Professional & Other Places	Shop & Service	Food
21	EUR Magliana	Professional & Other Places	Shop & Service	Nightlife Spot
22	EUR Palasport	Professional & Other Places	Shop & Service	Food
25	Furio Camillo	Shop & Service	Professional & Other Places	Outdoors & Recreation
26	Garbatella	Professional & Other Places	Nightlife Spot	Shop & Service
29	Giulio Agricola	Shop & Service	Professional & Other Places	College & University
35	Libia	Shop & Service	Professional & Other Places	Outdoors & Recreation
36	Lodi	Shop & Service	Travel & Transport	Nightlife Spot
37	Lucio Sestio	Shop & Service	Food	Outdoors & Recreation
38	Malatesta	Shop & Service	Professional & Other Places	Food
40	Marconi	Shop & Service	College & University	Professional & Other Places
48	Pigneto	Shop & Service	Nightlife Spot	Food
49	Piramide	Professional & Other Places	Shop & Service	Food
51	Ponte Lungo	Shop & Service	Professional & Other Places	Food
54	Quintiliani	Shop & Service	Professional & Other Places	Food
55	Re di Roma	Shop & Service	Food	Outdoors & Recreation
58	San Giovanni	Shop & Service	Food	Professional & Other Places
60	Sant Agnese - Annibaliano	Shop & Service	Professional & Other Places	Nightlife Spot
62	Subaugusta	Shop & Service	Professional & Other Places	Food
65	Tiburtina	Shop & Service	Travel & Transport	Nightlife Spot
71	Valle Aurelia	Shop & Service	Professional & Other Places	Travel & Transport

Conclusion

Due to limitations of free data of Foursquare API we can study only basic information about the city. It can be interesting to study city based on attendance of public places and their price range. In this case we can find expensive and cheap places, compare price and attendance relationship. Also it could be

interesting to have monthly statistical data of using each metro station. This information would help in the future development of the metro system in Rome.

Reference:

1. Glickman, Mark, and Pavlos Protopapas. “CS109B Data Science 2: Advanced Topics in Data Science.” CS109B - Lab 7: Clustering, 2019, harvard-iacs.github.io/2019-CS109B/labs/lab7/solutions/.
2. Godfrey, Kate, et al. “Determining The Optimal Number Of Clusters: 3 Must Know Methods.” Datanovia, 21 Oct. 2018, www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/.
3. “Stazioni Della Metropolitana Di Roma.” Wikipedia, Wikimedia Foundation, 6 Jan. 2020, it.wikipedia.org/wiki/Stazioni_della_metropolitana_di_Roma.
4. Rogozhin, Stanislav. “Classification of Moscow Metro Stations Using Foursquare Data.” Medium, Towards Data Science, 24 June 2019, towardsdatascience.com/classification-of-moscow-metro-stations-using-foursquare-data-fb8aad3e0e4.

Links:

1. Capstone_final jupyter notebook https://github.com/riccione/Coursera_Capstone/blob/master/Capstone_final.ipynb
2. Rome_Metro https://raw.githubusercontent.com/riccione/Coursera_Capstone/master/Rome_Metro.csv
3. Stations geo data https://raw.githubusercontent.com/riccione/Coursera_Capstone/master/Stations.csv
4. Stations_venues https://raw.githubusercontent.com/riccione/Coursera_Capstone/master/Stations_venues.csv