Exploring Rome: where to open your restaurant

1. Introduction

The aim of this project is to decide in which neighbourhood of Rome it is best to open a new restaurant. We will use the following as parameters to classify the neighbourhoods:

- activities already present;
- population;
- distance from the city center.

This is a complex problem and many variables can play decisive roles into getting to a solution. For this reason, it is convenient and useful to apply an accurate data analysis and to use appropriate machine learning techniques to solve it.

The present report's target is an audience who might be interested in opening a restaurant in Rome. Nevertheless, it can be used as an informative paper for any person to have a different insight about what the Italian capital currently offers.

2. Data

2.1. Choice of data and sources

We want to decide in which neighbourhood of Rome to open a restaurant. Therefore, we first gathered a list of all the neighbourhoods of Rome from Wikipedia. This resulted in 35 entries.

Always from Wikipedia, we gathered the total population of each neighbourhood.

We then used Google Maps to find the latitude and longitude of each neighbourhood.

The list of neighbourhoods, together with their population, latitude and longitude was finally compiled into a Numbers spreadsheet.

2.2. Initial data processing

The spreadsheet was converted into an .xls format and imported in a Python Jupiter Notebook as a pandas dataframe.

The coordinates of the city center of Rome where found using a Python dedicated library and a custom function was built in order to calculate the distance of each neighbourhood, given its coordinates, from the city center.

The distance from the city center is then added as a column to the dataframe (Figure 1).

The total population of all neighbourhoods is calculated and the Population column of the dataframe is divided by the total population, resulting in a new column containing the fraction of the total population present in each neighbourhood.

	Name	Area	Population	Latitude	Longitude	Distance
0	Flaminio	1.1877	12155	41.92633	12.468281	3.866773
1	Parioli	4.7506	13749	41.93473	12.492440	4.672743
2	Pinciano	3.5662	20854	41.92177	12.486754	3.180231
3	Salario	0.4668	8301	41.91507	12.500687	2.830723
4	Nomentano	3.2611	39245	41.91354	12.515327	3.500043

Figure 1

On the other hand, the Distance column is first normalised using a max-min procedure, so that all distances ranges between 0 and 1 (zero being exactly at the center, while 1 being the farthest away neighbourhood). Finally, a transformation is performed such that 0 indicates the farthest away from the center and 1 the exact center.

2.3. Further feature extraction

The last set of data we need are the existing activities already present on the territory. This is done by making API calls to the Foursquare website. In particular, for each neighbourhood, we feed its coordinates to Foursquare and get all the venues that result in exploring a certain radius from the specified coordinates. This results in a list of 534 venues.

Finally, we focus on the Venue column, which contains the "type" of each venue. This is the feature we are interested in.

Since we look for opening a restaurant, we replaced all the venues containing the word "Restaurant" with "Restaurant" and all the venues containing the word "Café" with "Bar". In this way we will not distinguish an Italian restaurant from a Chinese one. Also, "Bar" is essentially the Italian word for "Café".

Being it a categorical variable, we use a one-hot encoding to turn it into binary variables.

Therefore, we finally obtain the dataframe we use for the analysis (Figure 2).

	Neighborhood	Art Gallery	Art Museum	Athletics & Sports	BBQ Joint	Bakery	Bar	Basketball Stadium	Beach	Bed & Breakfast	...	Train Station	Trattoria/Osteria	University	Video Game Store
0	Flaminio	0	0	0	0	0	1	0	0	0	...	0	0	0	0
1	Flaminio	0	1	0	0	0	0	0	0	0	...	0	0	0	0
2	Flaminio	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	Flaminio	0	0	0	0	0	0	0	0	0	...	0	0	0	0
4	Flaminio	0	0	0	0	0	0	0	0	0	...	0	0	0	0

Figure 2

3. Methodology

First of all we use the folium package to plot each neighbourhood in the city map (Figure 3). This immediately shows the presence of three neighbourhoods (those in the Ostia location) which are well outside the city boundaries.

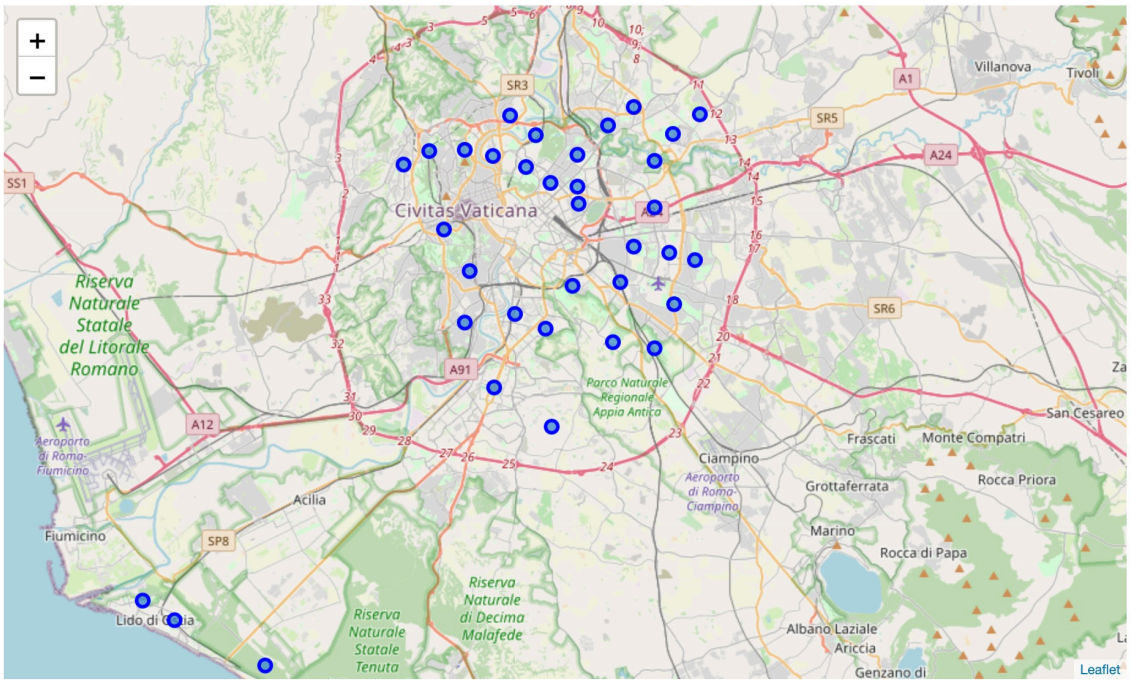

Figure 3

Then, we identify the 10 most important features of each neighbourhood (i.e. the feature with the highest value, including population and distance - remember that a distance is 1 if we are exactly in the center!). This is shown in Figure 4. By looking at the full table we realise that the distance from the center is the main feature for all the neighbourhoods except the Ostia ones.

	Neighborhood	1st Most Relevant Feature	2nd Most Relevant Feature	3rd Most Relevant Feature	4th Most Relevant Feature	5th Most Relevant Feature	6th Most Relevant Feature	7th Most Relevant Feature	8th Most Relevant Feature	9th Most Relevant Feature	10th Most Relevant Feature
0	Alessandrino	Distance	Restaurant	Park	Athletics & Sports	Gym / Fitness Center	Population	Donut Shop	Electronics Store	Farmers Market	Flea Market
1	Appio Claudio	Distance	BBQ Joint	Football Stadium	Tennis Court	Population	Bakery	Donut Shop	Farmers Market	Flea Market	Food
2	Appio-Latino	Distance	Restaurant	Trattoria/Osteria	Gym	Gastropub	Pizza Place	Pub	Supermarket	Hotel	Population
3	Appio-Pignatelli	Distance	Garden	IT Services	Population	BBQ Joint	General Entertainment	Farmers Market	Flea Market	Food	Food Court
4	Ardeatino	Distance	Restaurant	Bar	Supermarket	Bistro	Hotel	Gym Pool	Pizza Place	Plaza	Diner

Figure 4

We choose to perform a K-Means clustering algorithm on the data in Figure 2 in order to cluster the neighbourhoods by venues, distance from the city center and population.

We assume the distance from the center to be an important variable, so its values (at the moment between 0 and 1) are multiplied by a factor 1.5. In this way, the distance feature

will be more relevant than the Venue and Population feature, as those range between 0 and 1.

We choose to cluster the neighbourhoods in 5 groups, and the result is shown in Figure 5.

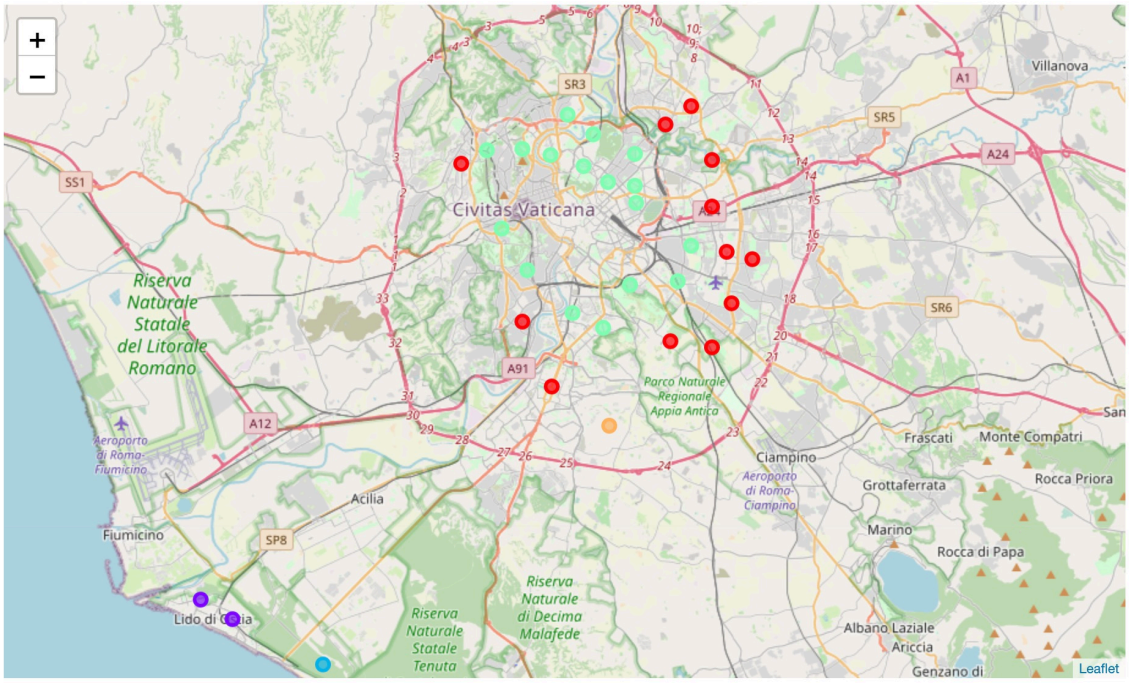

Figure 5

The cluster labels are to be read from the picture as follows:

- Cluster 0
- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4

4. Results

As expected, the Ostia neighbourhoods are isolated into Clusters 1 and 2, and yet they are not in the same cluster, showing that the distance is not the only feature that made a difference.

Cluster 4 consists in a single neighbourhood (Giuliano-Dalmata), which is quite far from the city center.

Cluster 1, 2 and 4 are too far from the center, thus we focus on Clusters 0 and 3. All neighbourhoods in Clusters 0 and 3 have the distance as most important feature. Thus, we focus on the 2nd most important feature, i.e. the most important venue.

For each neighbourhood we compute the number of venues which correspond to the most important venue, and the results are displayed using seaborn in Figures 6 and 7.

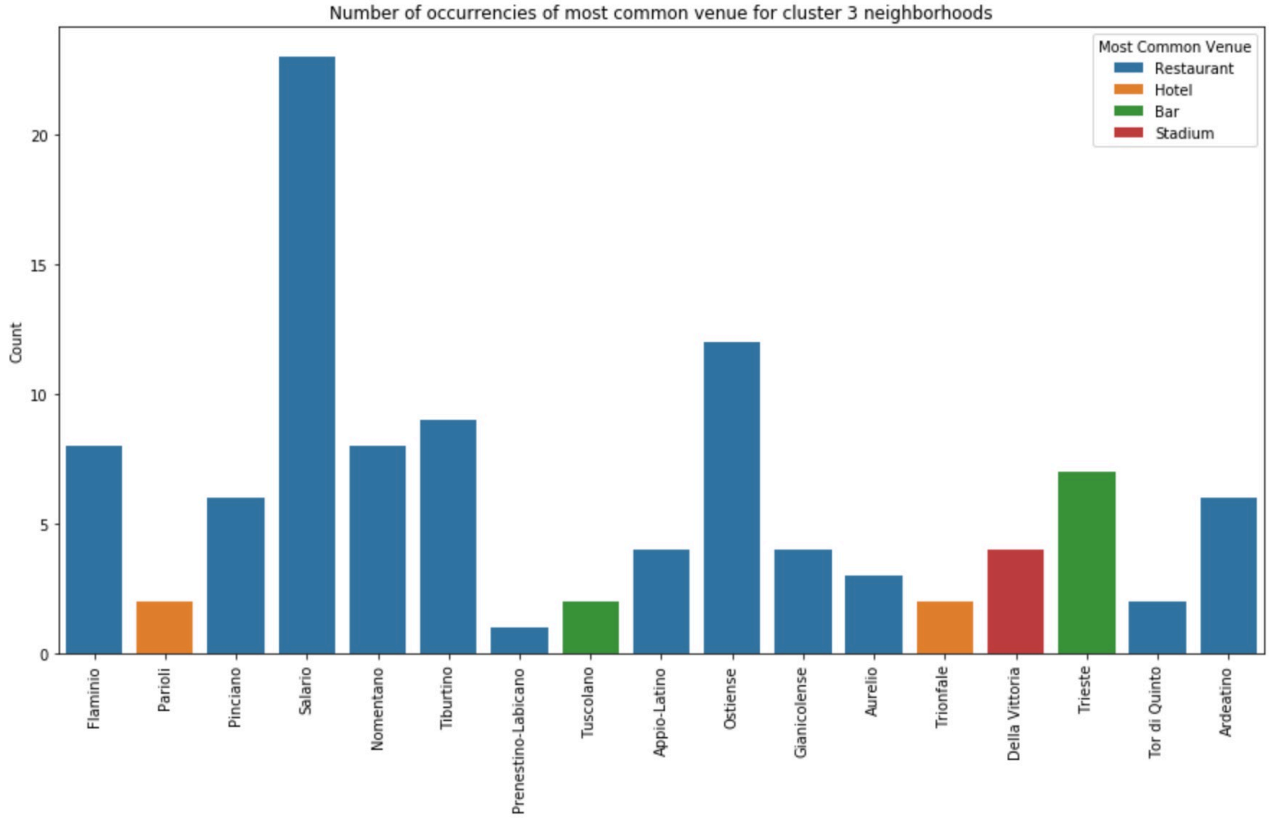

Figure 6

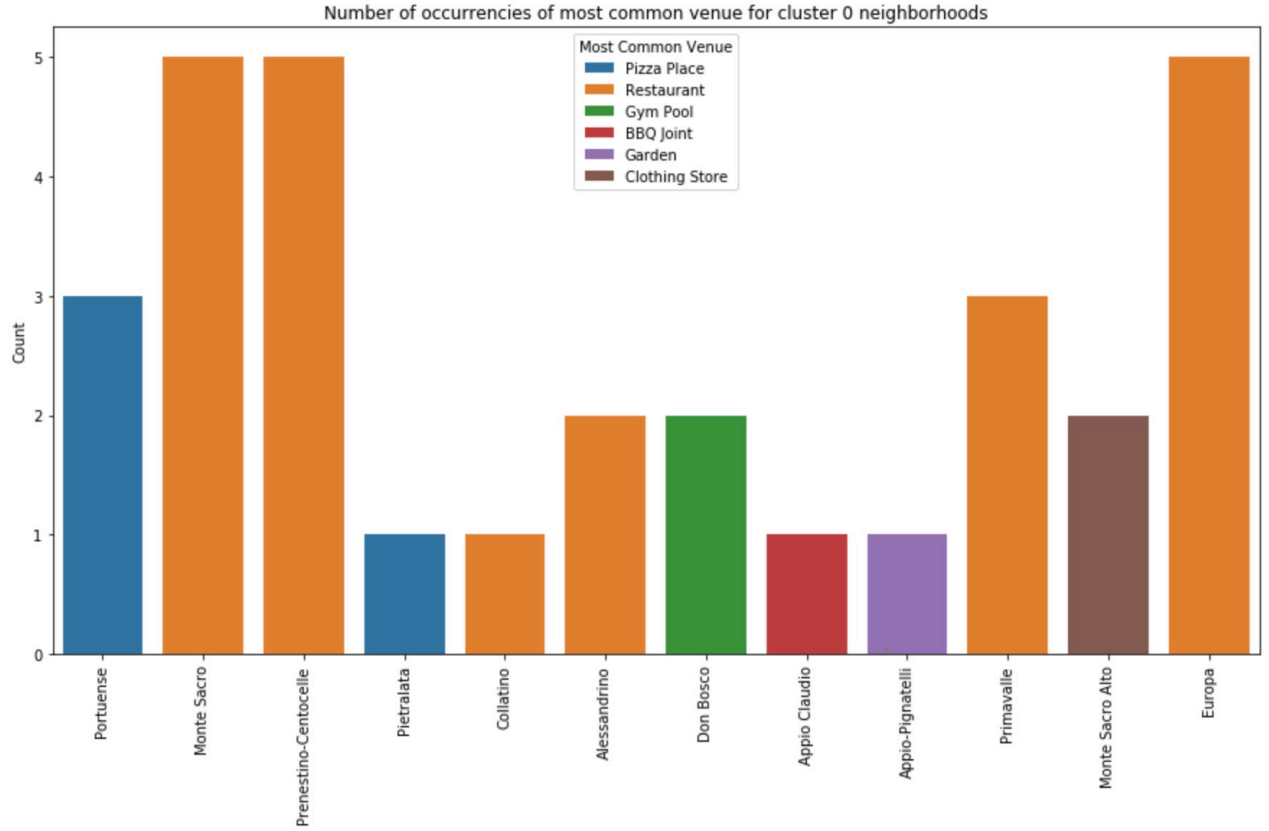

Figure 7

We assume that distance from the center takes priority over other features, so we examine Figure 6, first. This shows that there are three neighbourhoods, Parioli and Della Vittoria, where the main venue is neither a bar nor a restaurant.

Finally, it turns out that Parioli is, among the three, the closest to the center, and thus it is the one we recommend.

5. Discussion and future projects

The analysis we carried out is heavily based on the data obtained from Foursquare. They seem to be consistent with the intuition: in fact we have more data of venues for the most central neighbourhoods (Cluster 3).

Nevertheless, this may indicate a bias of the data towards the city center, as it is likely that Foursquare's users (especially if tourists) would visit the city center rather than the other areas of the city.

Therefore, a more accurate analysis would be possible by ensuring that venues data are collected uniformly from each part of the city.

Furthermore, we presented an analysis which did not take into account any business detail. For example, it would be useful to repeat this analysis with, for instance, data about the income of each restaurant, as well as the rent that each activity pays. In this way we would also be able to construct a predictive, as well.

6. Conclusion

We performed an analysis to locate the neighbourhood of Rome in which it is most convenient to open a restaurant.

This has been done by collecting data from Wikipedia and Foursquare.

The data have been analysed using a Jupiter Notebook (in Python) and the main Machine Learning tool we used was the K-Means clustering algorithm.

This resulted in recommending Parioli as the neighbourhood in which to open the new restaurant.