Predicting the best location in Rome

Alessandro Bonelli

27th May 2020

1. Introduction

1.1 Background

The city of Rome is the capital of Italy, an unparalleled historical and cultural patrimony. Among stable citizens (just under 3 million in 2017) and tourists throughout the year for its mild climate (29 million according to 2018 statistics), it is an absolute *must visit* among the cities of Europe, if not of the world.

Therefore, considering the enormous potential basin between citizens and tourists, commercial activities that are currently unsatisfactory (such as quality and / or quantity) could be an interesting target for new investments.

1.2 Problem

It's interesting to evaluate the possibility of creating new commercial premises such as restaurants or pizzerias (or other types not sufficiently widespread to serve users), especially in parts of the city that are deficient in this type of structure (both in quantity and quality).

In any case, the objectives can be qualified in the short term (identification of the best placement) and in the medium term (define whether to take over a business or request permits to build a new one). The search for the best placement is not however based exclusively on the distribution of the population, but also on the actual presence of commercial activities in the areas of the city and their type. The logical combination of statistical information on the number of citizens and existing activities (also with an assessment of their degree of satisfaction) makes it possible to have an information pool of sure interest for the appropriate assessments.

In this project we will try to find an optimal location for a **new pizzeria in Rome**.

Since there are lots of restaurants/pizzeria in Rome we will try to detect locations that are not already crowded with them, but however with high density of residents.

We will use our data science powers to generate a few most promising neighbourhoods based on this criterion. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

1.3 Interest

This project is therefore aimed at entrepreneurs in the restaurant sector who have the aim of identifying placements for new structures with prudent choices on where to create them. The benefits of a prudent choice are in maximizing the ROI in the shortest possible time, therefore also being able to employ staff (reception/cooks/waiters), giving also workplaces to citizens.

2. Data acquisition and cleaning

2.1 Data sources

From the urban point of view, Rome is divided in a decidedly particular way with four toponymic groups: the first are the districts (so called "*rioni*") that make up the historic centre established since the Middle Ages (22 of limited territorial extension); then there are the neighbourhoods (so called "*quartieri*") that surround the historic centre, reaching as far as Ostia, that is, the sea (35 in all); finally, secondary groups are the suburbs and zones. In this analysis we will focus on the first two (*rioni* and *quartieri*) which comprise the majority of the residents and are of greater value as cleaning, transport, security and office presence.

Based on definition of our problem, factors that will influence our decision are:

- number of existing restaurants/pizzerias in the neighbourhood (any type of restaurant)
- population and its density for each neighbourhood

We decided to use regularly defined neighbourhoods in Rome (*rioni* and *quartieri*) as areas to check for the best location.

Following data sources will be needed to extract/generate the required information:

- demographic information taken from Wikipedia (https://it.wikipedia.org/wiki/Quartieri_di_Roma and https://it.wikipedia.org/wiki/Rioni_di_Roma)

- centres of candidate areas will be generated algorithmically and approximate addresses of centres of those areas will be obtained using **Google Maps API reverse geocoding**
- number of restaurants/pizzeria and location in every neighbourhood will be obtained using **Foursquare API**
- coordinate of Rome centre will be obtained using **Google Maps API geocoding** of well-known Rome location (*Coliseum monument*).

2.2 Data cleaning

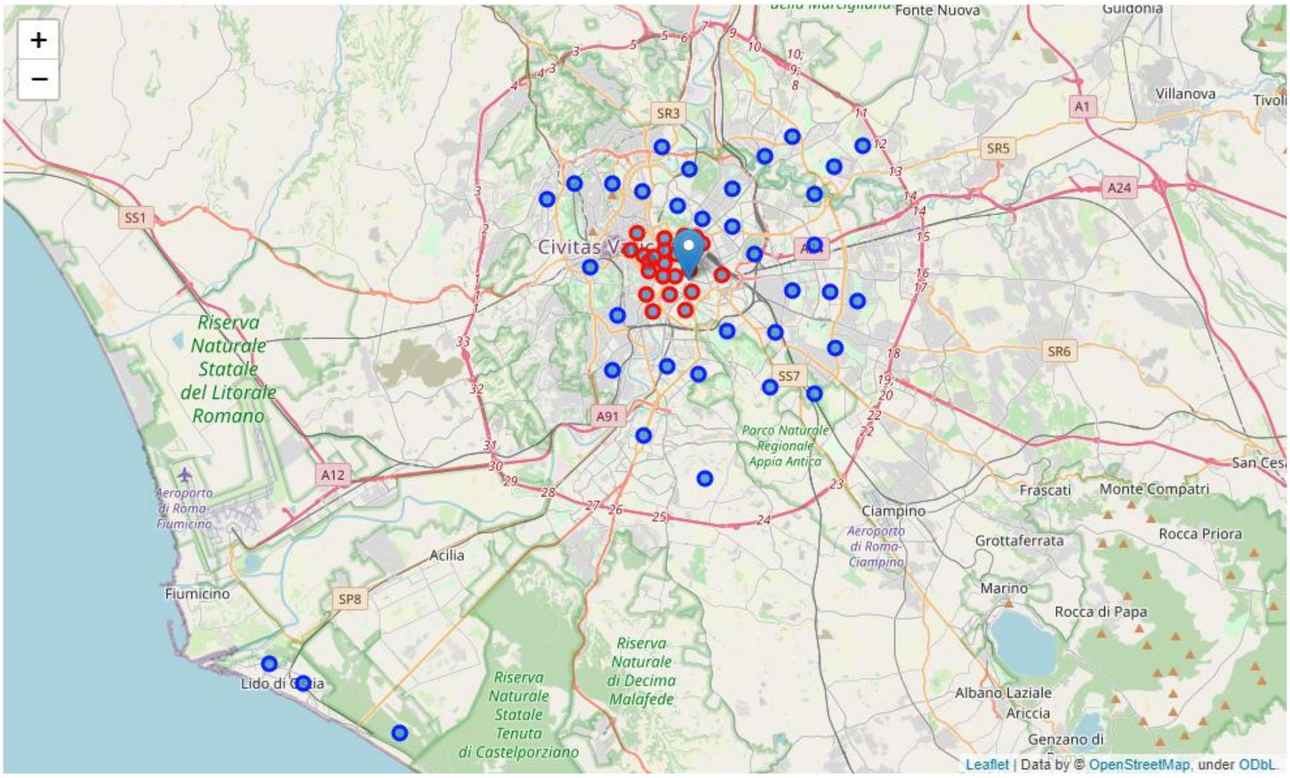
Data downloaded or scraped from different location on Wikipedia were combined into one table. There are no problems with missing data or outliers.

2.3 Feature selection

After data cleaning (just for points and commas for numeric values), we obtain the following features:

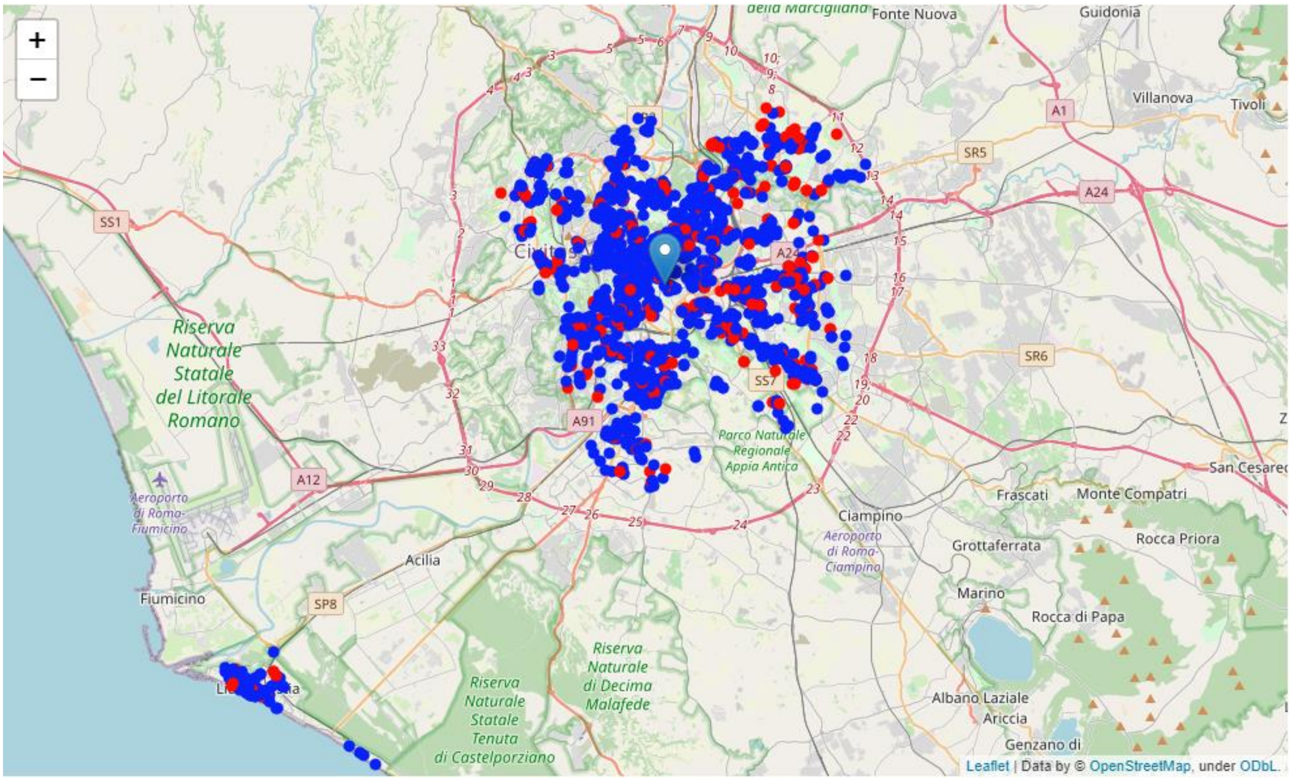
- Type of borough (*quartiere* or *rione*)
- Progressive code for borough (it's a Roman numeral, as I, IV…)
- Name of borough
- Number of inhabitants
- Population density

The data are then enriched with the latitude and longitude values of the neighbourhood centres through the Google API.


+
−
Fonte Nuova
Guidonia
Villanova
Tivoli
SR3
SR5
A1
A24
SS1
Civitas Vaticana
Riserva Naturale Statale del Litorale Romano
Aeroporto di Roma-Fiumicino
A12
A91
SS7
SR6
Parco Naturale Regionale Appia Antica
San Cesareo
Frascati
Monte Compatri
Ciampino
Aeroporto di Roma-Ciampino
Grottaferrata
Rocca Priora
Acilia
Fiumicino
SP8
Marino
Rocca di Papa
Riserva Naturale di Decima Malafede
Lido di Ostia
Riserva Naturale Statale Tenuta di Castelporziano
Albano Laziale
Ariccia
Genzano di
Leaflet | Data by © OpenStreetMap, under ODbL.

Finally, we asked Foursquare to retrieve the commercial premises related to food (therefore restaurants, pubs, pizzerias, ...) within a predetermined distance from the centre of the neighbourhoods (differentiating between *quartieri* and *rioni* for their different extension); the collected data are the name of the venue, its coordinates and the category.

In the following figure the blue dots indicate the places related to food, while the red ones indicate the pizzerias.


+
−
Fonte Nuova
Guidonia
Villanova
Tivoli
A1
SR5
A24
SS1
Riserva Naturale Statale del Litorale Romano
A24
SR6
SS7
A91
Parco Naturale Regionale Appia Antica
A12
Aeroporto di Roma-Fiumicino
San Cesareo
Frascati
Monte Compatri
Ciampino
Aeroporto di Roma-Ciampino
Grottaferrata
Rocca Priora
Marino
Rocca di Papa
Acilia
Fiumicino
SP8
Riserva Naturale di Decima Malafede
Riserva Naturale Statale Tenuta di Castelporziano
Albano Laziale
Ariccia
Genzano di
Leaflet | Data by © OpenStreetMap, under ODbL.

3. Exploratory Data Analysis

With the complete and clean database, we made a distinction between the number of pizzerias and other places related to food, so as to see their representativeness within the neighbourhoods. We have identified a neighbourhood without pizzerias ("Lido di Castel Fusano", so far from the city centre) which is perfectly logical being a stretch of coast with few houses in the Ostia area.

4. Predictive Modelling

In this project we will direct our efforts on detecting areas of Rome that have low foods-related places/pizzeria density, but with high population density.
In first step we have collected the required data: location and type (category) of every food-related places / pizzeria in Rome. We have also identified pizzeria (according to Foursquare categorization → '*pizza place*').

Therefore, we will focus on most promising neighbourhoods, according to previous "low/high" rules.

We will present map of all such locations but also create clusters of those locations (using **k-means clustering** for similar features) to identify general zones/neighbourhoods which should be a starting point for final exploration and search for optimal venue location by stakeholders.

We're using an unsupervised machine learning algorithm for classification (*k-means*) because we haven't a prior defined target, therefore we wish to discover the relationship existing (hopefully) between homogeneous features.

4.1 Solution to the problem

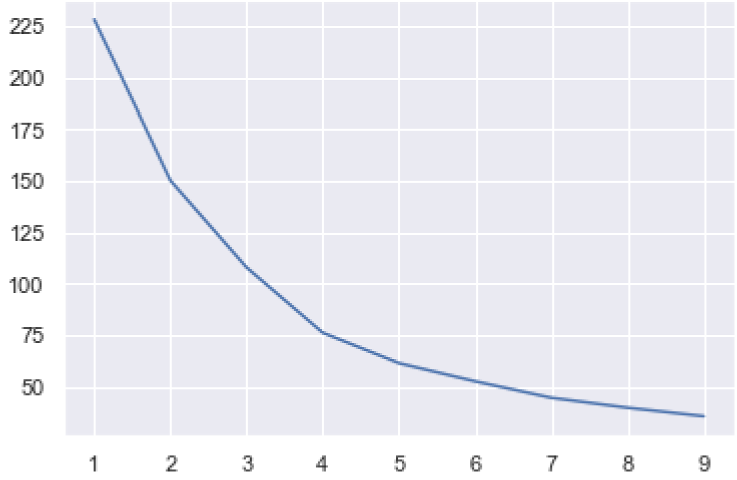
In order to correctly apply the k-means algorithm, we proceed with the selection of the features of interest: number of inhabitants, population density, number of pizzerias, number of other places related to food; regarding the other categorical variables (such as the name or type of neighbourhood) we do not take anything as non-discriminating.

Then we carry out the standardization of the variables (zero mean and unit variance), since for the correct functioning of the algorithm (based on Euclidean distances in the clusters) it is necessary that the features are of the same magnitude.

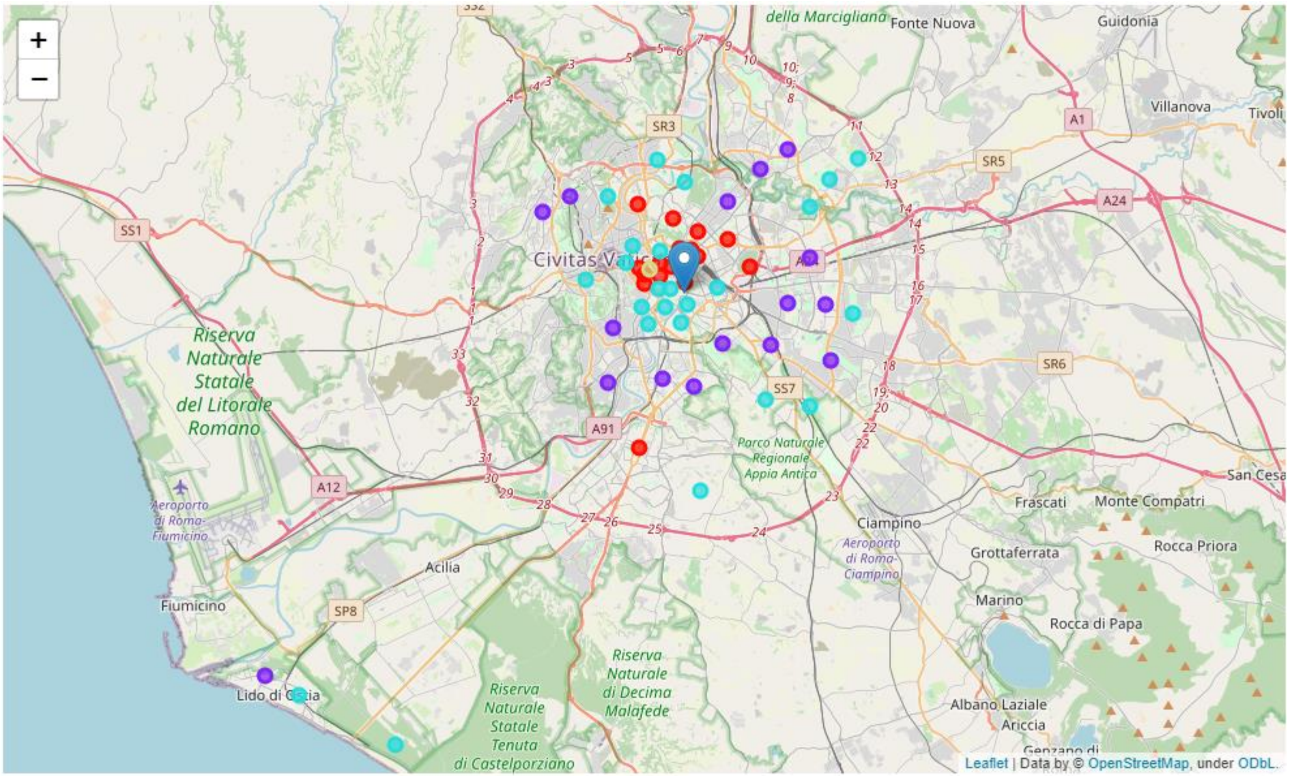
Because we're searching for high population density and low density for food-related places, we need to invert the sign of standardized data for the latter to give a "high rate" to the four variables together.

To identify the correct number of clusters, we apply the ***elbow method*** that calculates the Within-Cluster-Sum-of-Squares (WCSS) for each chosen size, building a graph that allows us to evaluate the best number of clusters.

In this case, the most marked slope variation occurs with **4 clusters**.

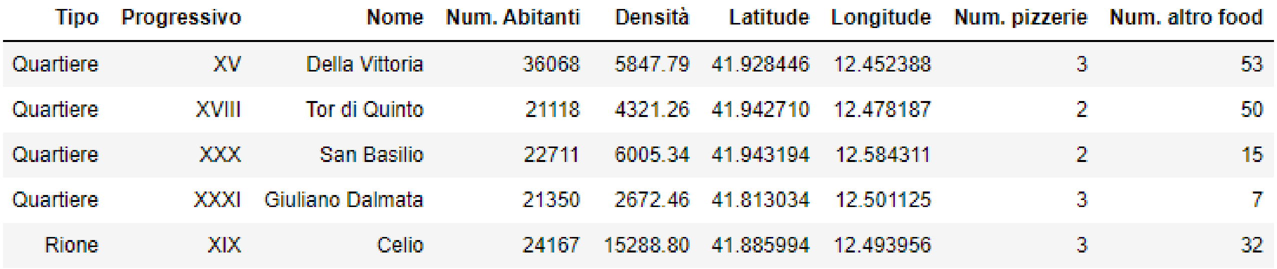


Below is the map of the city of Rome in which the identified clusters are highlighted in different colours.


+
−
della Marcigliana
Fonte Nuova
Guidonia
Villanova
Tivoli
SR3
SR5
A1
A24
SS1
Civitas Vaticana
Riserva Naturale Statale del Litorale Romano
SR6
SS7
A91
A12
Parco Naturale Regionale Appia Antica
Aeroporto di Roma-Fiumicino
Fiumicino
Acilia
SP8
Lido di Ostia
Riserva Naturale Statale Tenuta di Castelporziano
Riserva Naturale di Decima Malafede
Ciampino
Aeroporto di Roma-Ciampino
Frascati
Monte Compatri
San Cesareo
Grottaferrata
Rocca Priora
Marino
Rocca di Papa
Albano Laziale
Ariccia
Genzano di
Leaflet | Data by © OpenStreetMap, under ODbL.

Looking at the average values of the features for the identified clusters, we have found that the cluster labelled as number 1 seems to be a good candidate for our new pizzeria: it has a high population mean density but a low mean number of pizzerias.

Going on to search further for specific neighbourhoods that have the minimum amount of pizzeria among them, trying to limit the output further narrowing the set of pizzerias (less than the mean) and residents (more than the mean).



Tipo	Progressivo	Nome	Num. Abitanti	Densità	Latitude	Longitude	Num. pizzerie	Num. altro food
Quartiere	XV	Della Vittoria	36068	5847.79	41.928446	12.452388	3	53
Quartiere	XVIII	Tor di Quinto	21118	4321.26	41.942710	12.478187	2	50
Quartiere	XXX	San Basilio	22711	6005.34	41.943194	12.584311	2	15
Quartiere	XXXI	Giuliano Dalmata	21350	2672.46	41.813034	12.501125	3	7
Rione	XIX	Celio	24167	15288.80	41.885994	12.493956	3	32

5. Conclusions

We've found a set of quartieri/rioni as good candidates for our project. They seem to be the best choice for our stated rules. Personally, I suggest to avoid *San Basilio* just for a matter of relative degradation and not-so-high security.

Our analysis shows that although there is a great number of food-related places in Rome, there are pockets of low pizzeria density. Our attention was focused on borough which offer a combination of low restaurant density and high population density.

Those location candidates (*quartieri* and *rioni*) were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centres of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

Result of all are zones containing the best fit between population and pizzeria/food-related places density. This, of course, does not imply that these zones are actually optimal locations for a new pizzeria. Purpose of this analysis was to only provide info on areas in Rome not crowded with existing restaurants - it is entirely possible that there is a very good reason for small number of restaurants, reasons which would make them unsuitable for a new pizzeria regardless of lack of competition in the area.