# Exploring Rome: where to open your activity

## 1. Introduction

The aim of this project is to decide in which neighbourhood of Rome it is best to open a new activity, such as a café or a gym. 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.

## 2. Data

### 2.1. Choice of data and sources

We want to decide in which neighbourhood of Rome to open an activity. 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.

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.

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. 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.