Start-up recommendation: Province of Rome, Italy

(Capstone project for IBM Data Science Professional Certification)

# Introduction

My first preference would have been to solve a problem relating to my home country (South Africa) but data on the platforms that we were introduced to in the course, is limited for this region. My second preference is Italy given my previous life in the region of Lazio, Rome and surroundings. Looking at the data available, it appears as if there is enough to make this a viable region outside of the US, to target.

The city of Rome has a population of 2856133 but the province of Rome has 1 486079 distributed in about 121 towns around Rome (radius of 50km).

Some of these towns around Rome has large populations that could make a business viable. Office space and taxes in Rome itself is high and capital needed for start-up would be prohibitive for a small business.

We therefor explore the possibility of opening a business in the province of Rome.

# Business question and approach:

The business question that attempt to answer is:

*What would be the best business to start in the province of Rome (outside the capital) and which towns would be most appropriate?*

The approach will be:

1. Find data for all the towns in the province of Rome
2. Analyse all the towns for the following aspects:
   1. Population
   2. Venues already recorded
3. find all the data on Foursquare for those towns
4. Use K-means to cluster the business venues
5. Select the towns with the highest population and propose as business the types that are the least common in those towns.

# Data used for project

1. Get the geospatial co-ordinates for the towns surrounding Rome:

Publicly available at: <http://www.dossier.net/utilities/coordinate-geografiche/provincia-roma.htm>'

Scrape this website with beautifulsoup to get the data.

2. Get the demographical of each town:

Publicly available at: <https://www.tuttitalia.it/lazio/provincia-di-roma/36-comuni/popolazione/>'

Scrape this website with beautifulsoup to get the data.

3. Combine the **geospatial and demographical** data in a pandas dataframe

4. Draw a map of the Province

5. Get the **Foursquare data** for all towns in the province

6. Explore the venues in the Rome Province

7. Transform the data so that **kmeans clustering** can be applied to it.

8. Select the value of k most appropriate using the elbow method.

9. Apply **kmeans clustering** to get the most common venues in each town.

10. Map the clusters

11. Select the **least common venue** in the towns with the **highest population**. These will be the towns and businesses proposed to a prospective start-up.