Cluster in Rome

Claudio Sanguigni

December 2020

1 Introduction

For many tourist, visiting,italian restaurants are a great way to relax and enjoy themselves during holidays. They can taste the best cousine in the world, eating from pasta to good fish, passing throgh delicious dessert. Property developers are incouraged to open new restaurant since the tourism in Rome is always increasing year after year. Of course, as with any business decision, opening a restaurant requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the it is one of the most important decisions that will determine whether the mall will be a success or a failure.

2 Business Problem

The objective of this capstone project is to analyse and select the best locations in the city of Rome to open a newrestaurant. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In the city of Rome, if a property developer is looking to open a new restaurant, where would you recommend that they open it?

3 Audience

This project is particularly useful to property developers and investors looking to open or invest in restaurant in the capital city of Italy. This project is timely as the city is currently suffering from oversupply of srestaurant. . The local newspaper also reported in March last year that the true occupancy rates in restaurants may be as low as 40 per cent in some areas, quoting a Financial Times (FT) article cataloguing the country's continued obsession with food and bevarage.

4 Data

To solve the problem, we will need the following data:

- List of neighbourhoods in Rome. This defines the scope of this project which is confined to the city of Rome, the capital city of the country of Italy Asia.
- Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to restaurant. We will use this data to perform clustering on the neighbourhoods.

Sources of data and methods to extract them is this Wikipedia page (`https://en.wikipedia.org/wiki/Category:Rome_by_rione`) contains a list of neighbourhoods in Rome, with a total of 21 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautifulsoup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods. After that, we will use Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the restaurants category in order to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

5 Methodology

Firstly, we need to get the list of neighbourhoods in the city of Rome. Fortunately, the list is available in the Wikipedia page . We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Rome. Next, we will use Foursquare API to get the top 100 venues that are within a radius of 1500 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the

venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analysing the "Italian restaurant" data, we will filter the "Italian Restaurant" as venue category for the neighbourhoods. Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 5 clusters based on their frequency of occurrence for "Restaurant". The results will allow us to identify which neighbourhoods have higher concentration of restaurants while which neighbourhoods have fewer number of restaurants. Based on the occurrence of shopping malls in different neighbourhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new restaurants.

6 Results

Cluster 1

```
In [95]: rome_merged.loc[rome_merged['Cluster Labels'] == 0]
```

Out[95]:

	Neighborhoods	Italian Restaurant	Cluster Labels	Latitude	Longitude
21	Rome R. XXII Prati	0.14	0	41.90322	12.49565
19	Rome R. XX Testac	0.14	0	41.90322	12.49565
3	Rome R. IV Campo Marzio	0.12	0	41.90710	12.47786
4	Rome R. IX Pigna	0.14	0	41.90322	12.49565
5	Rome R. V Ponte	0.14	0	41.90322	12.49565
18	Rome R. XVIII Castro Pretorio	0.16	0	41.90551	12.50188
17	Rome R. XVII Sallustiano	0.13	0	41.90783	12.49613
16	Rome R. XVI Ludovisi	0.11	0	41.90776	12.48955
14	Rome R. XIX Celio	0.14	0	41.90322	12.49565
13	Rome R. XIV Borgo	0.14	0	41.90322	12.49565
11	Rome R. XII Ripa	0.14	0	41.90322	12.49565

Figure 1: Cluster 1

Cluster 4

```
In [98]: rome_merged.loc[rome_merged['Cluster Labels'] == 3]
```

Out[98]:

	Neighborhoods	Italian Restaurant	Cluster Labels	Latitude	Longitude
7	Rome R. VII Regola	0.19	3	41.89480	12.47028
12	Rome R. XIII Trastevere	0.23	3	41.88839	12.46621

Cluster 5

```
In [99]: rome_merged.loc[rome_merged['Cluster Labels'] == 4]
```

Out[99]:

	Neighborhoods	Italian Restaurant	Cluster Labels	Latitude	Longitude
20	Rome R. XXI San Saba	0.14	4	41.87985	12.4902
15	Rome R. XV Esquilino	0.11	4	41.89403	12.5060

Figure 3: Cluster 4,5

Cluster 2

```
In [96]: rome_merged.loc[rome_merged['Cluster Labels'] == 1]
```

Out[96]:

	Neighborhoods	Italian Restaurant	Cluster Labels	Latitude	Longitude
2	Rome R. III Colonna	0.25	1	41.836429	12.756938

Cluster 3

```
In [97]: rome_merged.loc[rome_merged['Cluster Labels'] == 2]
```

Out[97]:

	Neighborhoods	Italian Restaurant	Cluster Labels	Latitude	Longitude
0	Rome R. I Monti	0.06	2	41.89315	12.48825
9	Rome R. X Campitelli	0.12	2	41.89325	12.48143
8	Rome R. VIII Sant'Eustachio	0.08	2	41.89965	12.47491
6	Rome R. VI Parione	0.10	2	41.89778	12.47075
1	Rome R. II Trevi	0.05	2	41.90119	12.48446
10	Rome R. XI Sant'Angelo	0.12	2	41.89355	12.47871

Figure 2: Cluster 3,4

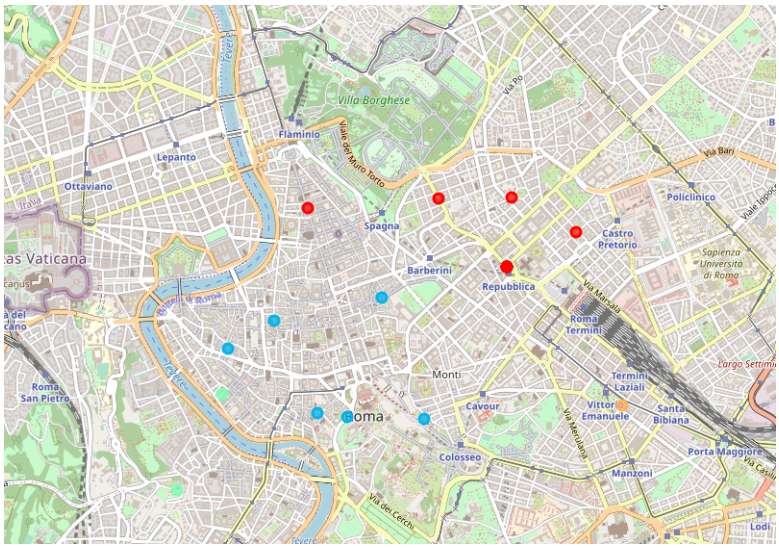

Figure 4: Cluster map

7 Discussion

Most of the restaurant are concentrated in the central area of Rome, with the highest number in cluster 2,3,4 and moderate number in cluster 5. On the other hand, cluster 3 has very low number of restaurant in the neighborhoods. This represents a great opportunity and high potential areas to open new restaurant as there is very little to no competition from existing ones. Meanwhile, restaurant in cluster 3 are likely suffering from intense competition due to oversupply and high concentration of sthem. From another perspective, this also shows that the oversupply of restaurant mostly happened in the central area of the city, where the most tourist are concentred. Therefore, this project recommends property developers to capitalize on these findings to open new restaurants in neighborhoods in cluster 3 with little to no competition.. Lastly, property developers are advised to avoid neighborhoods in cluster 3 which already have high concentration of restaurant and suffering from intense competition