Coursera Capstone Project

IBM Applied Data Science Capstone

Opening a truly Italian restaurant in Milan

Giuseppe De Nittis

1 Introduction

Food is one of the most important resources for mankind. For Italian people, it becomes even more important, bringing emotions and feelings with it. In fact, in Italy, having a meal is not just feeding ourselves with what we need, but it becomes a fundamental moment to speak, discuss and share ideas (often, about food different w.r.t. the one we are eating).

Thus, deciding to open a restaurant in Italy can result in a very complex decision. Location, type of food and style affect significantly the future of the restaurant, making these decisions crucial. The objective of this capstone project is to decide the best location to open a new restaurant in the city of Milan (Italy). Using data science methodology and machine learning techniques learned during the Specialization, the question I want to answer is the following: where would you open a new truly Italian restaurant in the city of Milan? Which district would you choose?

1.1 Audience

The audience is constituted by investors willing to open a new activity in one of the most important Italian cities, namely Milan. Despite the high number of restaurants already present in the city, unfortunately, a lot of them is just driven by the latest fashion, and recipes are very simple and immediate, just trying to attract non-experienced tourists. An investor would probably decide to open a new and classy place, a unique spot in the city, with high-level chefs and sommeliers. This project can provide her the answer with the best location for her new truly Italian restaurant.

2 Data

To solve the problem, I will need (and use) the following data:

- list of districts in Milan. This defines the boundaries of the project, which is restricted to one of the most important cities in Italy;
- latitude and longitude coordinates of those neighborhoods;
- venue data, specifically the ones related to Italian restaurants, needed to perform the neighborhood clustering.

2.1 Sources of data and methods to extract them

This Wikipedia page (`https://en.wikipedia.org/wiki/Category:Districts_of_Milan`) contains a list of districts in Milan, with a total of 76 districts. I will extract the data from the Wikipedia page by exploiting Python requests and the BeautifulSoup. Then, I will get latitude and longitude coordinates of

the districts using the Geocoder package. Then, I will use Foursquare APIs to retrieve information on the venues present in those districts. Such APIs provide many categories of the venue data: I am interested in the Italian restaurant category.

3 Methodology

The first step consist in retrieving the districts of the city of Milan. Such information can be gathered from the Wikipedia page `https://en.wikipedia.org/wiki/Category:Districts_of_Milan`. To do this, I perform web scraping by exploiting Python libraries *requests* and *beautifulsoup*. This way, I got the list of names, which is not enough since I have to add the geographical coordinates, i.e., latitude and longitude, to call the Foursquare APIs. To this aim, I use the *geocoder* library, which allows to convert address into geographical coordinates in the form of latitude and longitude. Once this data have been collected, I populate a DataFrame and show an interactive map plotted using *folium* library.

Next, I employ the Foursquare APIs to get the top 500 venues that are within a radius of 5000 meters. To perform this operation, I registered as a developer on Foursquare to get a Foursquare ID and Foursquare a secret key. Then, calling the APIs, Foursquare provides the venue list in JSON format, from which I extract name, category, latitude and longitude of the venues.

The next step is analyzing the districts by grouping the rows by district and taking the mean of the frequency of occurrence of each venue category. By doing so, I am already preparing the data for the subsequent clustering phase. Since I want to find the best location to open a truly Italian restaurant, I will filter *Italian restaurant* as venue category.

Now, everything is ready for the final phase, clustering. I adopt the k-means algorithm, which works as follows:

1. Choose some points, called *centroids*, that will constitute the center of the clusters;
2. compute the Euclidean distance of each venue from the centroids, and assign such venues to the closest centroid;
3. update the centroids by computing the mean of the points of each cluster;
4. repeat steps 2 and 3 until the algorithm converges, i.e., a certain number of iterations has been executed or some percentage of clusters do not change the cluster they belong to between two subsequent iterations.

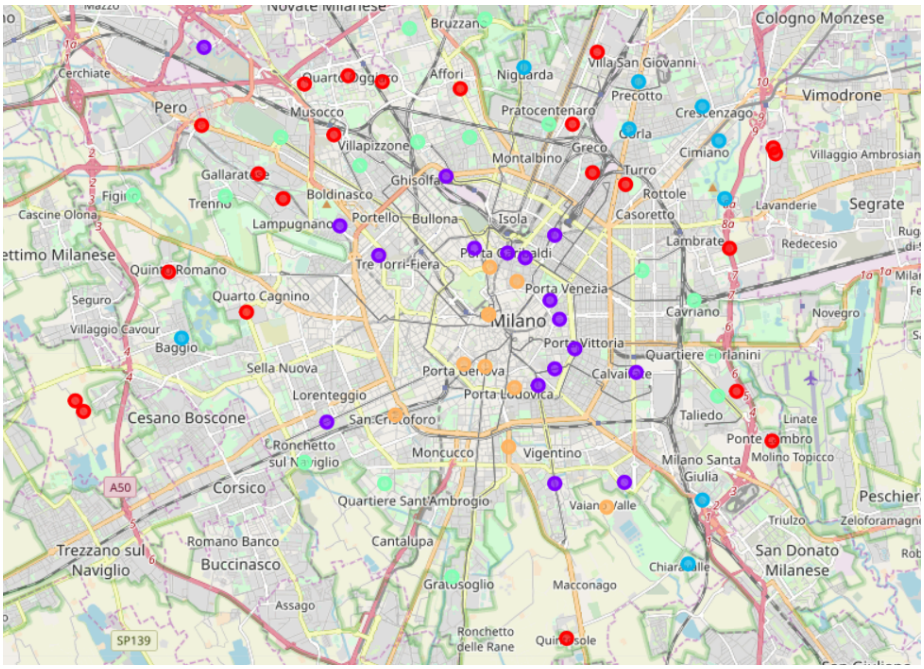
With this approach, I cluster the districts into five clusters depending on the frequency of the occurrence of Italian restaurants.

4 Results

The results of the above methodology are displayed in the following figure.

Specifically:

- Cluster 0 is shown in red;



- Cluster 1 is shown in blue;
- Cluster 2 is shown in dark green;
- Cluster 3 is shown in light green;
- Cluster 4 is shown in orange.

5 Discussion

Looking at the map above, we observe that Clusters 0, 2 and 3 (red, dark and light green, respectively) contain the highest percentage of restaurants. this is due to the fact that most of these districts are big in terms of area and outside from the very center of the city. Actually, they could be considered small towns on their own.

Conversely, Clusters 1 and 4 (blue and orange, respectively), despite being in the center, present values that are significantly smaller if compared to the ones of the previous clusters. This means that the presence if Italian restaurants is not high: this is due to the fact that there are a lot of different types of restaurants that serve all the different kinds of people that may visit the city.

This rationale also explains why in the outer districts the presence of Italian restaurants is higher. Being more residential areas, where commuters and mostly Italians live, such restaurants can be found easier. Thus, the best thing to do is to open a restaurant in the central area of Milan. Moreover, looking at the map, we notice that in the eastern area there are no points. In fact, this area have a very important museum and a University, thus there are more bars. Moreover,

it is also a chic district, so opening a fancy and classy restaurant here would be a good choice.

6 Conclusions and future research

This project has been developed with the goal of answering to the following question: *where would you open a new truly Italian restaurant in the city of Milan?* To answer, first I defined the business problem underlying such a question, specifying the data required and preparing the necessary data. Then, I performed the analysis by clustering the districts of the city into five clusters based on their similarities. According to the obtained results, I provided useful insights to potential investors, adding to the number also more informal data related to the various districts.

The results I found are interesting: opening an Italian restaurant in the center of one of the biggest Italian cities may seem counterintuitive. However, the project could be expanded. Here, I just used Foursquare data to determine the best location, but more factors should be included, e.g., information on the people living in the different districts.