IBM Applied Data Science Capstone

Opening a new Chinese Restaurant in Milan, Italy

Nunzio Logallo

August 2020

Introduction

Nowadays Traditional Cuisine Restaurant are an essential part of Multicultural Cities like Milan. In each big city we have a lot of different restaurant like Japanese, Korean, Mexican, Turkish and so on. Anyway, if we look at the competition part of opening a new restaurant in big cities there is a difference between restaurant of different culture, for example, it would be easier to open a Lebanese restaurant than a Sushi Restaurant because of the number of opened Lebanese restaurants and the consequent low competition. Opening a Chinese Restaurant is even more difficult than Sushi Restaurant in Milan because of the high number of them and the presence of a specific neighborhood called "Chinatown", anyway, there are busy zones without any Chinese Restaurant.

Business Problem

The main goal of this project is to analyze all the neighborhoods in Milan and select the best locations to open a new Chinese Restaurant. Using Data Science Methodology and Machine Learning this project aims to answer a question: "If an entrepreneur is looking to open a new Chinese Restaurant in Milan, where would you recommend to open it?".

Data

Obviously, Data and their sources are very important in this project and more specifically we need these data:

- List of neighborhoods in the city of Milan where we should be careful to select only the main city and not the metropolitan area;
- Latitude and Longitude of the neighborhoods, very important to plot a map and get venues data;
- Venues data, related to Chinese Restaurant, we will use it to perform clustering on neighborhoods.

Sources and extraction

Respectively, the sources and the methods to extract the data above are:

- Wikipedia (https://en.wikipedia.org/wiki/Category:Districts_of_Milan), where there is a list of 77 neighborhoods. We can web scraping technique using the Python libraries "beautifulsoup" and "requests" to extract the names of the neighborhoods and put them inside a Python list;
- Python Geocoder package, to extract the latitude and longitude of each neighborhood in the list, being careful about searching the neighborhoods inside the City of Milan instead of the metropolitan area;

- Foursquare API to extract the venue data of each neighborhood. Foursquare is one of the best sources to extract this type of data, having an extremely large database of 105+ million places and used by 150000 developers. With Foursquare we can filter the 900 available categories of venues and get only one we need: Chinese Restaurant.

Methodology

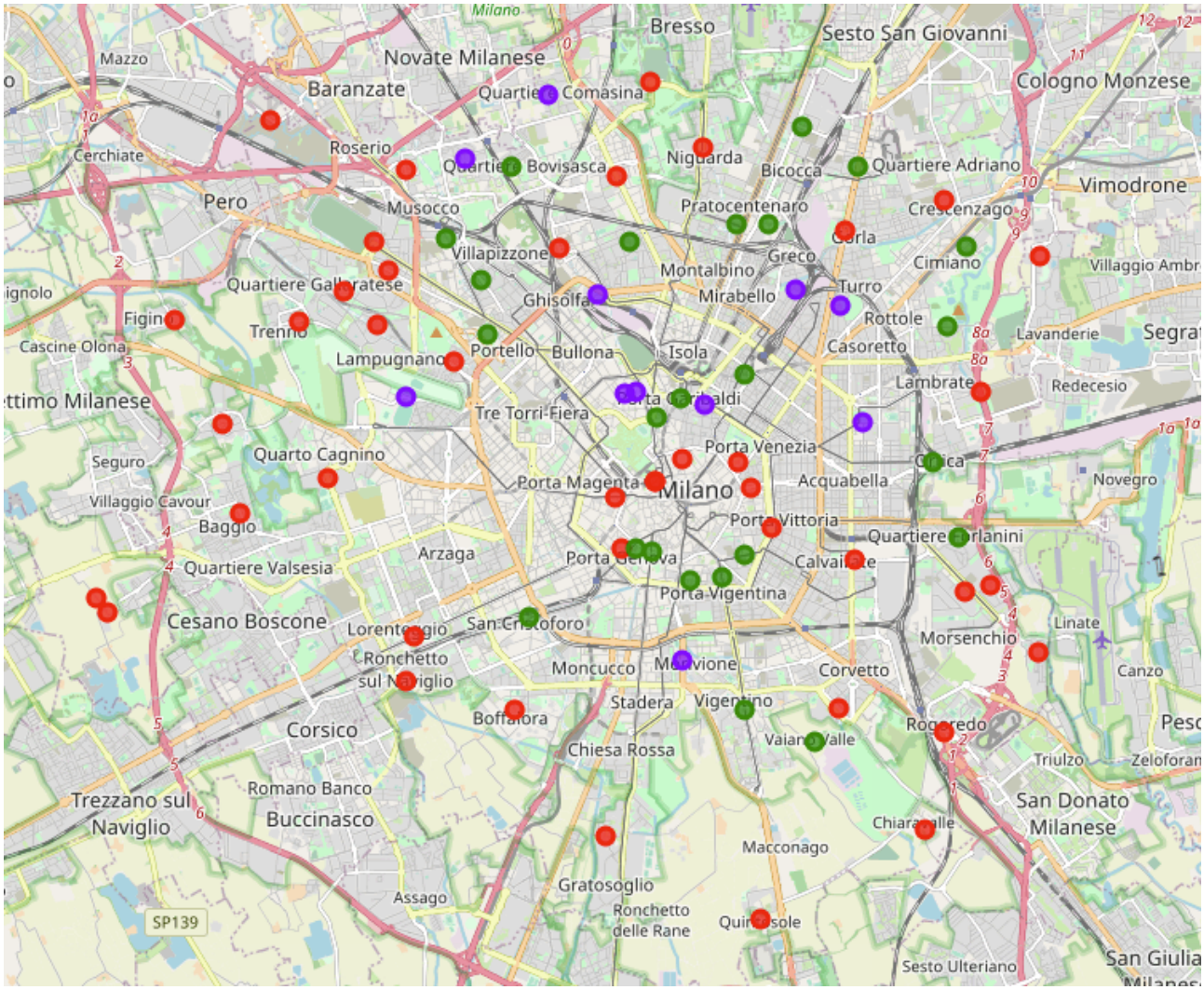
First of all, we need to get the list of all the districts in Milan from Wikipedia and we can easily do it using Python package “requests”, where we can get the full HTML web page. Next, we have use web scraping techniques, for this we can deal with the Python package “beautifulsoup” which allows us to filter each tag of the HTML page in order to extract the names of districts from a HTML list. After this we have to do a bit of data cleaning to remove all that suffix such as “(Milan district)” or “district” from the name of neighborhoods to better visualize them in a table. After this we need to get the geographical coordinates of each district, to create a map of the city; to do this I used “geocode” a python package that allows us to get coordinates of districts giving name, city and province. Now using districts names and coordinates we can create a first map of Milan with all the districts. This step is very important because knowing the city you can notice if the coordinates of some districts are wrong, for example plotting the map I notice that the coordinates of Chinatown were wrong and because of the relevance of this neighborhood in this project I decided to correct them. After this we can finally use Foursquare API to get all the venues (actually the top 350) from each district. At this point we should filter the big amount of venues from Foursquare API to the one needed in this project, Chinese Restaurants, to do this we can use One Hot Encoding to better visualize the venues of each district, than we can merge the table of each district in one containing values for each venue in each district and after this we can create a new dataframe with Chinese Restaurant data only. The next step and the most important of this project is clustering. We will use the k-means clustering algorithm with 3 clusters where the cluster 0 represent districts with very low number or zero Chinese Restaurants; cluster 1 high number of Chinese Restaurant; cluster 2 some Chinese restaurant. By completing clustering, we can make a map showing the distribution of Chinese Restaurant in Milan.

Results

K-means clustering gave us a result set of 3 clusters based on the concentration of Chinese Restaurants:

- **Cluster 0:** Low number of Chinese restaurants in the neighborhood;
- **Cluster 1:** High number of Chinese restaurants in the neighborhood;
- **Cluster 2:** Moderate number of Chinese restaurants in the neighborhood.

And this is the results map about August 2020 based on the top 300 venues from Foursquare API for each neighborhood in 1500 meters of radius:



Cluster 0: Affori, QT8, Lampugnano, Milano Santa Giulia, Monluè, Muggiano, Niguarda, Nosedo, Ponte Lambro, Porta Genova, Roserio, Porta Magenta, Porta Monforte, Ronchetto sul Naviglio, Porta Sempione, Rogoredo, Quintosole, Porta Venezia, Quinto Romano, Porta Vittoria, Quarto Cagnino, Quartiere Musocco, Gratosoglio, Gorla, Lambrate, Quadrilatero della moda, Chiaravalle, Giambellino-Lorenteggio, Bruzzano, Brera, Bovisa, Vialba, Barona, Centro Direzionale di Milano, Trenno, Taliedo, Figino, Gallaratese, Baggio, Assiano, Crescenzago, Calvairate;

Cluster 1: Porta Volta, Quarto Oggiaro, Ghisolfa, Turro, Città Studi, San Siro, Comasina, Morivione, Greco, Porta Nuova, Chinatown;

Cluster 2: San Cristoforo sul Naviglio, Vaiano Valle, Stazione di Milano Centrale, Segnano, Porta Garibaldi, Precotto, Bicocca, Bovisasca, Cimiano, Conca del Naviglio, Dergano, Forlanini, Garegnano, Quartiere Feltre, Ortica, Porta Lodovica, Porta Romana, Porta Tenaglia, Porta Ticinese, Porta Vigentina, Portello, Prato Centenaro, Vigentino, Villapizzone.

Discussion

As we can see from the cluster map, there is a high concentration of Chinese restaurant in the northern part of the city center, where there also is Chinatown, a crucial district for this project. If we choose to open a new Chinese restaurant in cluster 1 zones, we should have a famous restaurant with an already known name to compete with others restaurant in the zone. Cluster 2 zones does not have a specific location because as we can see there are some in the city center, in the northern part of the city, and all along the perimeter. These zones are very interesting because there we can compete with other restaurants making good offers in the beginning, and with new and quality products just to start our activity; If we choose to open in cluster 0 zones we can easily compete, but we should be careful at the zones because to open a restaurant we have to be sure to have customers in the zone. Here there are some zones in the city center but usually there we have a higher rent or no place to open.

Conclusion

The best choice depends on the budget we have and, on the products, we are capable to offer to our customers. In my opinion the best cluster of zones is number 2, because if we have a high budget and we are capable to compete with other restaurants, we can choose the southern part of the city center, near Porta Vigentina. If we have a normal Chinese restaurant with traditional and classic plate, as we can see from the map there is place in the north-western area of the city, near Gallaratese, a developing neighborhood of the city where we will not have much competition and because of the neighboring juvenile centers like the university residence, we can have customers. I do not recommend to open a new restaurant in the cluster 1 zones it is too much difficult to compete with the high number of restaurants and famous restaurants there.

References

Category:Districts of Milan - Wikipedia:
https://en.wikipedia.org/wiki/Category:Districts_of_Milan

Foursquare Developers Documentation - Foursquare:
https://developer.foursquare.com/docs