Fabio Marco Monetti

Exploring Milan boroughs:

a simple project of data science.

April 28, 2020

Contents

Chapter 1

Introduction

Taking inspiration from the un-graded laboratories that were offered by the course and from one of the ideas suggested in the last week of this course, I decided to explore and inspect the boroughs of my own city of birth: Milan, in Italy.

In this project I will explore the main boroughs of the city and see what kind of venues are the most popular.

This idea occurred because I want to remember my beautiful city as it was before the spreading of this awful virus called Covid-19 that is affecting our lives and will, for the foreseeable future, change how we conduct daily life and the relationship that everyone has with the city in which they live.

In recent years, after the Global Exposition of 2015, Milan has been one of the most visited cities in Europe; therefore, plenty of data is available on Foresquare about the main venues that one could visit.

Moreover, since the numbers of tourists coming in the city have been steadily growing up and it seemed that they would never going to drop, opening a new restaurant in Milan, or, for what it matters, any kind of venues that one could think of, might seem a bright idea to exploit this growth. It is true that competition is high in Milan, therefore it is advisable to run some data analysis in order to find out which parts of the city are developing and where is the best place to open a new restaurant.

In this project I will go through the process of making such a decision.

Chapter 2

Data

2.1 Data acquisition

I will use “Municipalities of Milan” () wiki page to get information about the boroughs in Milan. This page has a table that displays plenty of data about the boroughs; it also includes:

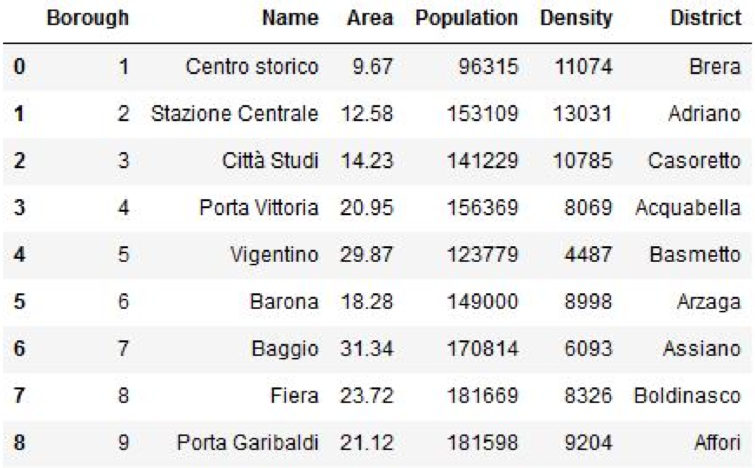
- boroughs;
- name;
- area;
- population;
- population density;
- quartieri (we could translate as "districts").

I will use a query in pgeocode and the geopy package to retrieve all the informations about the coordinates of those boroughs.

To get location and information about various venues in Milan I will use the Foursquare explore API. Using the borough’s postal codes, from Foursquare API (), I retrieved the following informations about the venues:

- name;

Figure 2.1



	Borough	Name	Area	Population	Density	District
0	1	Centro storico	9.67	96315	11074	Brera
1	2	Stazione Centrale	12.58	153109	13031	Adriano
2	3	Città Studi	14.23	141229	10785	Casoretto
3	4	Porta Vittoria	20.95	156369	8069	Acquabella
4	5	Vigentino	29.87	123779	4487	Basmetto
5	6	Barona	18.28	149000	8998	Arzaga
6	7	Baggio	31.34	170814	6093	Assiano
7	8	Fiera	23.72	181669	8326	Boldinasco
8	9	Porta Garibaldi	21.12	181598	9204	Affori

- category;
- latitude;
- longitude.

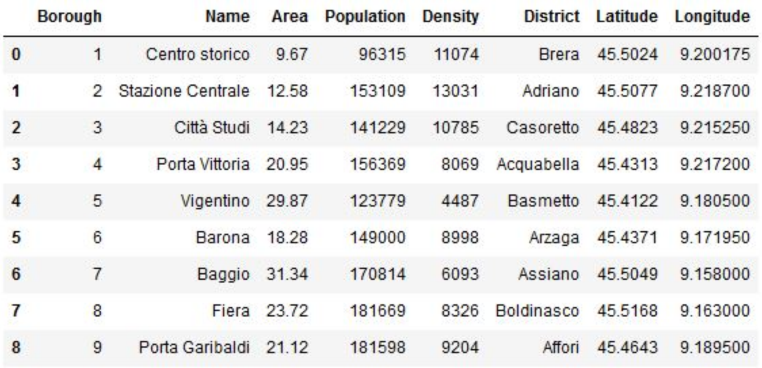
2.2 Data cleaning

Firstly, I scraped the Wikipedia page in order to obtain the data about the municipalities (or boroughs) of Milan. I used the BeautifulSoup package to scrape the data.

In extracting data from the table I retrieved also data that were not considered useful for the purpose of this project (though they can turn out to be of the most importance if this study is to be followed and implemented with further studies - so of course nothing has been permanently deleted). Therefore, at the end of the scraping project, the extracted dataframe presented nine rows (one for each borough) and five columns, as you can see in 2.1. These data needed to be cleansed, too, because there were inconsistencies about the separator for decimals and thousands (sometimes a comma, sometimes a point, for both). The table you see is the rightly formatted one.

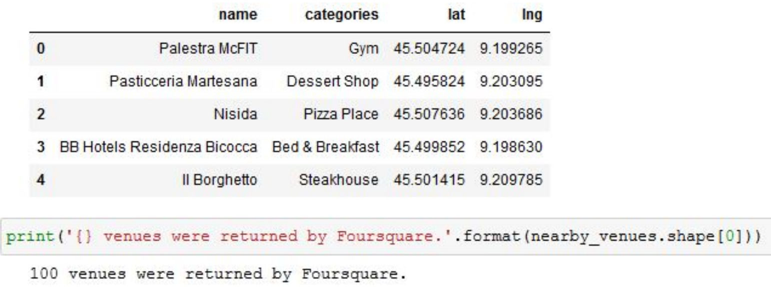
Secondly, I imported pgeocode to retrieve data about the latitude and longitude of the boroughs. In order to do so, I needed the postal code of those boroughs, which unfortunately was not part of the Wikipedia table; so, I went on the webpage of "Comune di Milano" (), and

Figure 2.2



	Borough	Name	Area	Population	Density	District	Latitude	Longitude
0	1	Centro storico	9.67	96315	11074	Brera	45.5024	9.200175
1	2	Stazione Centrale	12.58	153109	13031	Adriano	45.5077	9.218700
2	3	Città Studi	14.23	141229	10785	Casoretto	45.4823	9.215250
3	4	Porta Vittoria	20.95	156369	8069	Acquabella	45.4313	9.217200
4	5	Vigentino	29.87	123779	4487	Basmetto	45.4122	9.180500
5	6	Barona	18.28	149000	8998	Arzaga	45.4371	9.171950
6	7	Baggio	31.34	170814	6093	Assiano	45.5049	9.158000
7	8	Fiera	23.72	181669	8326	Boldinasco	45.5168	9.163000
8	9	Porta Garibaldi	21.12	181598	9204	Affori	45.4643	9.189500

Figure 2.3



	name	categories	lat	lng
0	Palestra McFIT	Gym	45.504724	9.199265
1	Pasticceria Martesana	Dessert Shop	45.495824	9.203095
2	Nisida	Pizza Place	45.507636	9.203686
3	BB Hotels Residenza Bicocca	Bed & Breakfast	45.499852	9.198630
4	Il Borghetto	Steakhouse	45.501415	9.209785

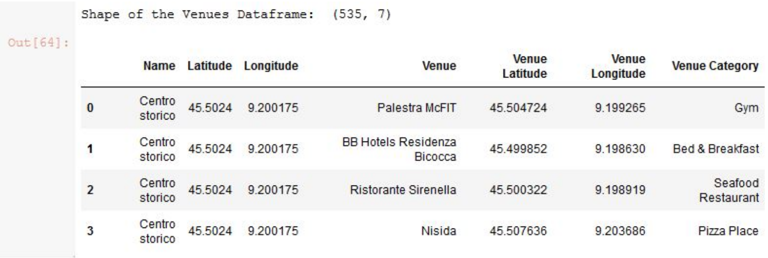
```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
```

```
100 venues were returned by Foursquare.
```

I was able to use those data to run a query on pgeocode with the postal codes and retrieve the coordinates of the boroughs.

I then added those data to the previous dataframe, see 2.2. Lastly, thanks to the Foursquare API, I retrieved information about the venues present in the boroughs of Milan. I looked for at most a hundred places in a radius of 1 km around every center of borough. The API returned a JSON file that was then formatted to become a data frame. For example, for the first borough, you can see what I did in 2.3. And the complete data frame looked like you can see in 2.4.

Figure 2.4



```
Out[64]:
Shape of the Venues Dataframe:  (535, 7)
```

	Name	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Centro storico	45.5024	9.200175	Palestra McFIT	45.504724	9.199265	Gym
1	Centro storico	45.5024	9.200175	BB Hotels Residenza Bicocca	45.499852	9.198630	Bed & Breakfast
2	Centro storico	45.5024	9.200175	Ristorante Sirenella	45.500322	9.198919	Seafood Restaurant
3	Centro storico	45.5024	9.200175	Nisida	45.507636	9.203686	Pizza Place

Chapter 3

Data Analysis

3.1 Interactive map

I will use Folium to draw an interactive leaflet map of the boroughs of Milan (see 3.1). Unfortunately, Milan has only nine main boroughs, so data cannot be so dense and thorough as for a bigger city like New York or Toronto; anyway, it is the city that I hold most dear and therefore I will go on with the analysis nevertheless.

3.2 Boroughs, venues and restaurants

Then I will draw an insightful map about venues around the cores of these boroughs to show the density of amenities one could find in these places (see 3.2). As you can see, there is quite

Figure 3.1

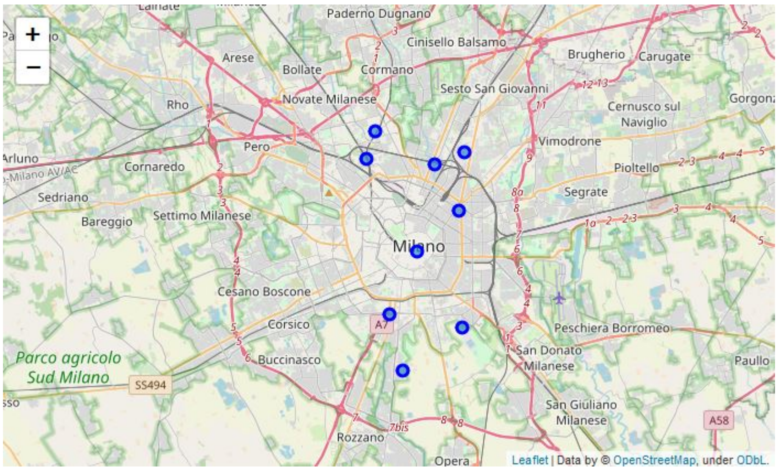


Figure 3.2

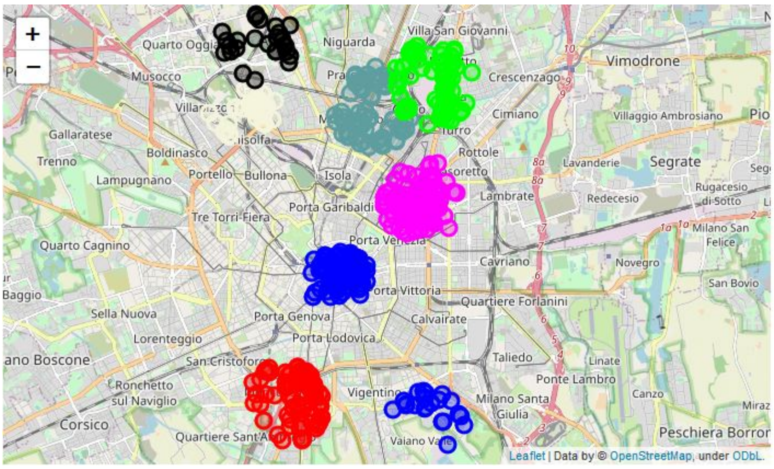
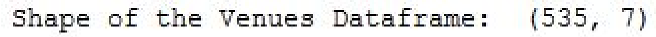


Figure 3.3



```
Shape of the Venues Dataframe:  (535, 7)
```

some density of venues (at least those that Foursquare was able to retrieve - living here, I can assure you that there are many more!), the number has been counted - see 3.3.

As an example for what this analysis could be useful for, I will extract from the data frame of venues, only those that contain the word "Restaurant", to see where anyone interested in opening a dining place might have the most success. Bare in mind that this kind of analysis could be carried on for every type of venues one could imagine: i.e. for a café, bar, even a gym! Therefore, the interest in this research is understandable.

For more informations about restaurants around the center of the boroughs of Milan, see the images shown in 3.5 for the number of restaurants retrieved by Foursquare and 3.4 to see the actual map: please note how different this is from the previous map; it means that there are many different venues that could be analysed, as previously stated.

Figure 3.4

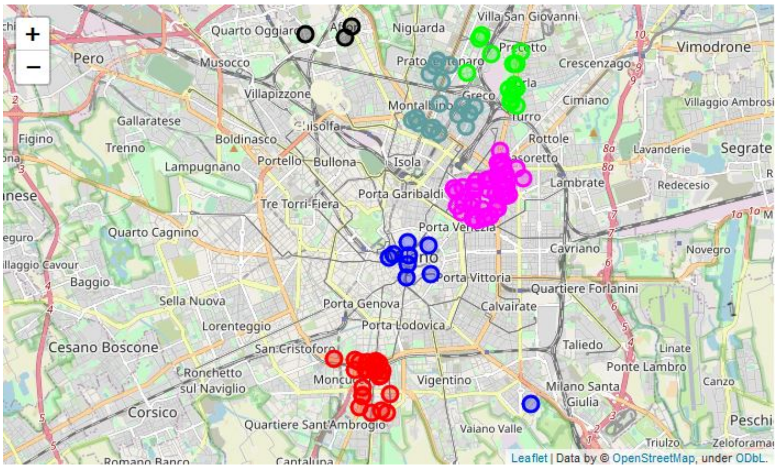
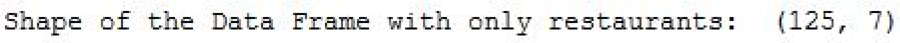


Figure 3.5



```
Shape of the Data Frame with only restaurants:  (125, 7)
```

Chapter 4

Modeling

4.1 Clustering Milan Boroughs

The first step of the K-means clustering that I will use is to identify the best K value, meaning the number of clusters in a given dataset. In order to choose the best values of K, I will use the elbow method on the data frame with the venues extracted 4.1. As you can see, and as it was highly predictable, the method doesn't precisely show an elbow because the number of boroughs that I am analysing is not very large; therefore, I would use some indications of this graph and a bit of logic in order to choose the value K.

The most sudden change of steepness can be seen for $K = [2, 4, 6]$, respectively, therefore the choice is limited to those values. Using a bit of logic, since I have a total number of boroughs equal to nine, selecting 6 clusters in which to divide those boroughs might seem a

Figure 4.1

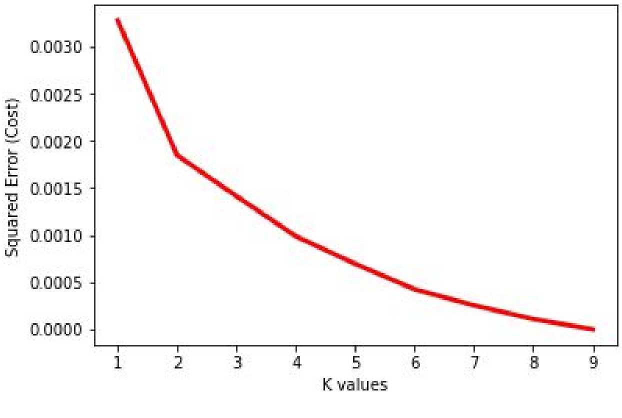
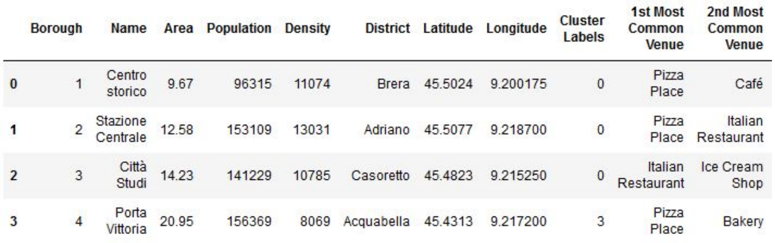
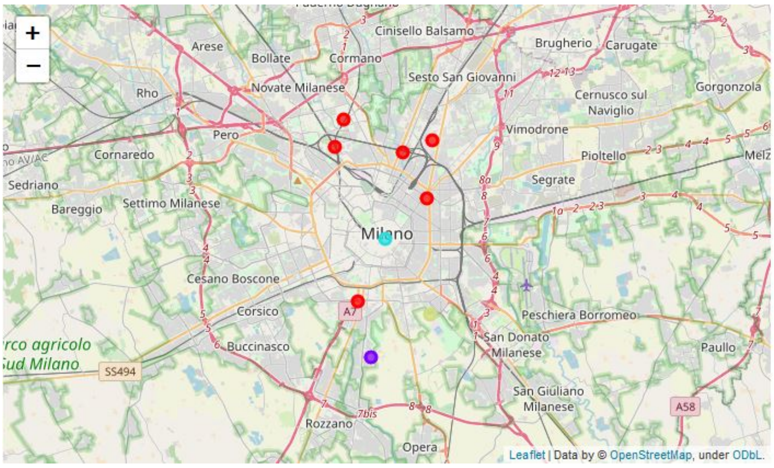


Figure 4.2



	Borough	Name	Area	Population	Density	District	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue
0	1	Centro storico	9.67	96315	11074	Brera	45.5024	9.200175	0	Pizza Place	Café
1	2	Stazione Centrale	12.58	153109	13031	Adriano	45.5077	9.218700	0	Pizza Place	Italian Restaurant
2	3	Città Studi	14.23	141229	10785	Casoretto	45.4823	9.215250	0	Italian Restaurant	Ice Cream Shop
3	4	Porta Vittoria	20.95	156369	8069	Acquabella	45.4313	9.217200	3	Pizza Place	Bakery

Figure 4.3



little hazardous and could bring to the overfitting of the model.

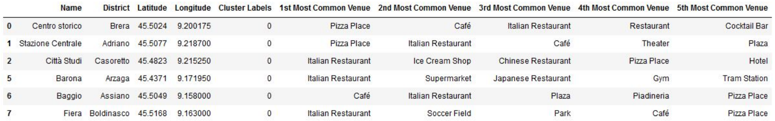
On the other hand, selecting a value for K of just 2 might seem conversely too low because it wouldn't appreciate at all the possible differencies between boroughs; moreover, selecting a low value brings a higher error on the model, so it is advisable to avoid such a choice.

Therefore, to conclude, a value of $K = 4$ has been chosen for the clustering.

Then, the clustering method was carried out using this value and you can see in 4.2 how the venues were divided by most common choice and that cluster labels were added to the dataframe. Furthermore, a map showing the clustering of the boroughs is presented in 4.3.

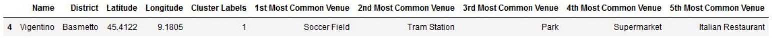
Again, since Milan has only nine main boroughs, it is a little difficult to appreciate the clustering, however - since I selected a value of 4 and not 2 for K - an idea of clustering is still visible here; anyway, the colors clearly highlight that there is one main cluster that includes most of the boroughs, plus other three clusters that differs from one another. In the next section I will try to analyse what possible differences there are and what conclusions can be drawn from this clustering.

Figure 4.4



	Name	District	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Centro storico	Brera	45.5024	9.200175	0	Pizza Place	Café	Italian Restaurant	Restaurant	Cocktail Bar
1	Stazione Centrale	Adriano	45.5077	9.218700	0	Pizza Place	Italian Restaurant	Café	Theater	Plaza
2	Città Studi	Casoretto	45.4823	9.215250	0	Italian Restaurant	Ice Cream Shop	Chinese Restaurant	Pizza Place	Hotel
5	Barona	Arzaga	45.4371	9.171950	0	Italian Restaurant	Supermarket	Japanese Restaurant	Gym	Tram Station
6	Baggio	Assiano	45.5049	9.158000	0	Café	Italian Restaurant	Plaza	Piadineria	Pizza Place
7	Fiera	Boldinasco	45.5168	9.163000	0	Italian Restaurant	Soccer Field	Park	Café	Pizza Place

Figure 4.5



	Name	District	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
4	Vigentino	Basmetto	45.4122	9.1805	1	Soccer Field	Tram Station	Park	Supermarket	Italian Restaurant

4.2 Examining the clusters

From a total of four clusters I will draw my conclusions about this brief project. I will show what is inside those clusters first.

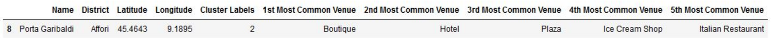
In the first cluster, the one identified by $index = 0$, from now on called cluster 0, there is the majority of the boroughs, i.e. six of them. They represent a group of similar boroughs that namely have as most common venues the likes of restaurants (italian, japanese, chinese), and cafés. You can see in 4.4 why these boroughs could be related. They all have Cluster Label 0 and are shown in red on the map 4.3.

Cluster 1 4.5 is the one that includes borough "Vigentino" and it has as the most common venue Soccer Field. In case you were wondering, Milan isn't exactly the most renowned place where professional footballers (oops, sorry: for american people, soccer player!) are grown up, even though two of the most famous clubs of italy are based here. However, we still love playing football (damn, I keep doing that! Playing. . . soccer!).

Cluster 2 4.6 includes "Porta Garibaldi" borough and has boutiques and hotels as the most common venues.

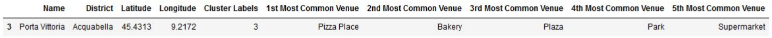
The last one, cluster 3 4.7, includes "Porta Vittoria" borough and shows that people love going to pizza places and bakeries here (who doesn't?).

Figure 4.6



	Name	District	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
8	Porta Garibaldi	Affori	45.4643	9.1895	2	Boutique	Hotel	Plaza	Ice Cream Shop	Italian Restaurant

Figure 4.7



	Name	District	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	Porta Vittoria	Acquabella	45.4313	9.2172	3	Pizza Place	Bakery	Plaza	Park	Supermarket

Chapter 5

Results and Discussion

In this small project I analysed the nine main boroughs of Milan in Lombardy, Italy, with regards to the most common venues that people in this city visit regularly.

I retrieved the data from reliable sources around the web and merged them to obtain a data frame that could be used to perform the analysis and to implement a Machine Learning algorithm in order to clusterize those boroughs.

Then I implemented the actual analysis method to show where the majority of people usually go in different boroughs; after that, it was possible to clusterize the boroughs based on what kind of venues were the most common.

The results are shown, in the previous chapter, in 4.4, 4.5, 4.6, 4.7.

Some finidings have been uncovered.

- In Cluster 0 there are the majority of the boroughs, and namely those were people go out for dinner. It is possible to see that most common venues include restaurants from different cuisines: italian, chinese, japanese, and so on.

- In Cluster 1 there is only one borough, the one where probably people go for practicing sport, because here two of the most common venues are soccer fields and parks.

- In Cluster 2 there is the borough of "Porta Garibaldi", with common venues like boutiques and hotels.

- In Cluster 3, which is the last one, there is a borough where people go probably for lunch to have a pizza or some panini, because here the most common venues are pizza places and bakeries.

From what I have discovered there are many things that could be of use for people in search of a business and especially for those who are thinking about opening a new venue in Milan.

Let's take the example of a restaurateur who would like to come to Milan and open his own restaurant: as you probably may know, here the competition is very high, both in terms of quantity and in terms of quality of food. Therefore it would be important to wisely choose where would be optimal to open the dining place.

Firstly, by looking at the concentration of restaurants in 3.4, compared to the number of total venues 3.2, one could see that there is a big differential especially for the boroughs highlighted in black and in dark blue (the part near the center).

Then, one could examine the clusters 4.3, and clearly see that it would not be advisable to open a restaurant in those boroughs of Cluster 0 (red dots), because here the concentration of restaurant is high and there are many famous and appreciated restaurants of many different cuisines: to stand out here would be an impossible mission and would require a huge amount of resources and effort.

Moreover, looking at Cluster 3, it is another borough where there are many places to eat and therefore it would not be advisable to go there.

Cluster 1 is the one that presents the many places for sport, so I don't think that this would be the best place for a restaurant.

Looking at Cluster 2 4.6, one could easily see that here the most common venues are hotels and boutiques, i.e. places where people go on their spare time and to relax, or where tourists mostly reside. Moreover, but this comes from direct experience, the borough of "Porta Garibaldi" is one renowned for its night life with many pubs and discos, and it is also growing rapidly because of the construction of the new square with skyscrapers commissioned by a bank; therefore, you can expect this growth to go on for many years.

Putting toghether all these discussions, I came to the conclusion that opening a restaurant in the borough "Porta Garibaldi" would be the best idea if one wanted to start a business in Milan, because there the competition is lower than in other boroughs and it has all the features one would want for his business to succeed.

Chapter 6

Conclusion

To conclude this project, I would like to point out how this brief analysis of the clusters has been brought on for someone who would want to open a restaurant in Milan, but it really could be used for any kind of venues that one could imagine; it would just require to acquire the right data from Foursquare and then perform the analysis again having in mind what one would like to show. Therefore, it is very adaptable.

Unfortunately for the visual part of this project, the number of main boroughs of Milan is very little, just nine; moreover, data on these boroughs are easy to find online, but data on smallest samples and a higher number of places (for example the districts of Milan are thirty-seven, if I remember correctly) are not that easy to come by.

It would be very interesting to look for those data and perform a more thorough analysis about Milan and its venues. It could give a more complete vision of what I wanted to highlight with this project and could be useful to many people and businesses.

Maybe it is something on which I will work in the future, meanwhile I hope it could be of inspiration to someone who loves Italy as I do or that maybe would want to know some more about the beautiful city of Milan.

It has been a real pleasure to work on this project and I drew inspiration from the work of many people here, I hope that in my little I could do my part to inspire people with my work.

Thank you to anyone who will take the burden of reading all this for me.

Best of luck and, in this moment of crisis, stay home and stay strong.