Clustering Districts in Downtown Milan

Francesco Bellogi

April 2020

Contents

1 Introduction

1.1 Background

Milan is Italy financial capital and it's by far the most economically lively city in Italy. Milan is also particularly famous for its *fashion weeks* and for its *lyfestyle*.
With around 1,3 millions inhabitants it's the second most populous city in Italy and also one of the most densely populated. People from all around the country move to Milan to chase a career and in the last year Milan has attracted a lot of people from all around the world.
Milan is a fast pace growing city in which people enjoy going out for dinner, for *aperitivo*, for shopping and that makes it an optimal place to open a restaurant, a shop, a club. On the other hand, not all districts are the same, some of them have been growing over the years and others have fallen behind. Many businesses are opening and many are closing. Having a more clear picture of the current situation of the various in districts in downtown Milan can be extremely useful for anybody who is looking for business opportunities, for a family wanting to buy a house, for the local government having to distribute finances.

1.2 Description of the data

I built the dataset for this analysis from scratch, collecting pieces from different sources. I obtained the list of names of districts is Milan by scraping the relative Wikipedia page. The resulting list of names needed a lot of cleaning. Once cleaned I was able to obtain the geo-spatial coordinates for each districts using the python geopy package.
After that I collected for each district the first 100 venues in a radius of 300 miles through the Foursquare API. Following some more data wrangling I obtained my dataset ready for clustering, containing all relevant venues for each districts. For example it will display Brera's Art Galleries, Cafes and Restaurants.
As an example, in the table below are reported 5 venues found with the Forsquare API

in Cordusio's district.

Table 1: 5 Places in Cordusio

Neighborhood	Venue	Venue Category	Venue Latitude	Venue Longitude
Cordusio	Starbucks Reserve Roastery	Coffee Shop	45.464920	9.186153
Cordusio	Venchi	Ice Cream Shop	45.465214	9.187340
Cordusio	Palazzo della Ragione	Monument / Landmark	45.464792	9.187785
Cordusio	Park Hyatt Milan	Hotel	45.465532	9.188911
Cordusio	Bialetti Store	Kitchen Supply Store	45.464775	9.188343

2 Analysis

2.1 Web Scraping and Data Wrangling

A data set containing district level segmentation with geographic coordinates for the city of Milan is not readily available. Actually it is impossible to find even a clean list of postal codes for the different districts of Milan. In order to overcome this lack of available data, I used the *pandas* function *read_html* to scrape the Wikipedia page containing the list of names of Milan's districts at `https://it.wikipedia.org/wiki/Municipi_di_Milano`. After some cleaning I obtained the below table.

Table 2: Initial Dataframe

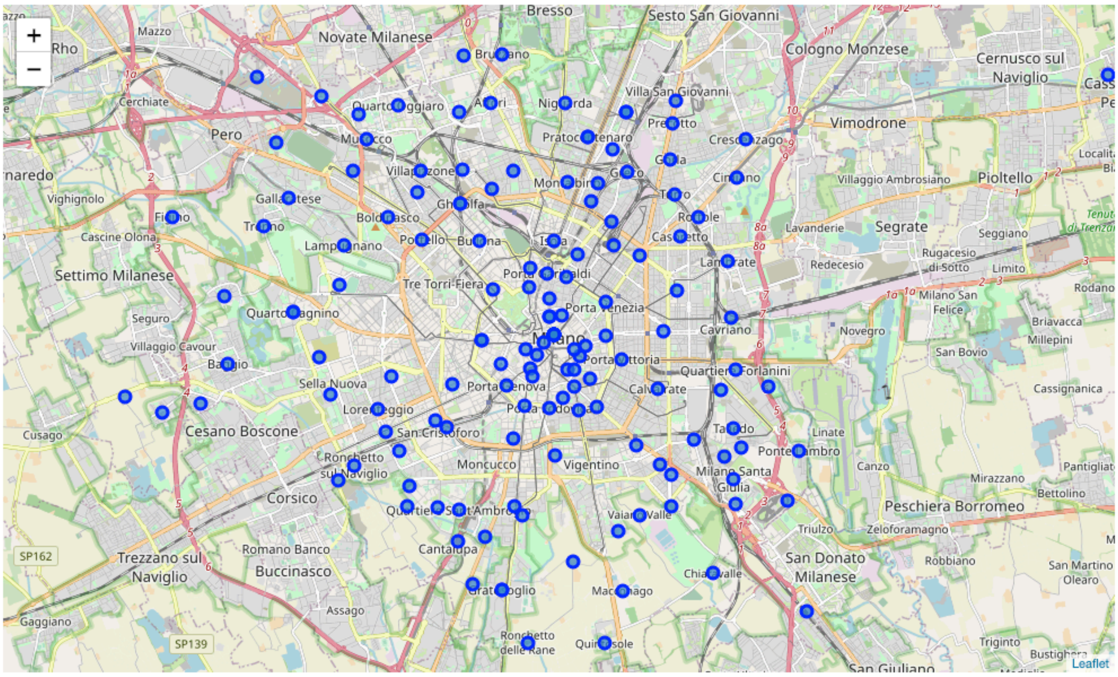
Municipio	Denominazione	Districts
Municipio 1	Centro storico	Cordusio, Cinque Vie, Brisa, Brera, Case Rotte...
Municipio 2	Stazione Centrale, Gorla, Turro, Greco, Cresce...	Stazione Centrale, Loreto, Turro, Crescenzago,...
Municipio 3	Città Studi, Lambrate, Venezia	Porta Venezia, Porta Monforte, Casoretto, Rott...
Municipio 4	Vittoria, Forlanini	Porta Vittoria, Porta Romana, Acquabella, Sena...
Municipio 5	Vigentino, Chiaravalle, Gratosoglio	Porta Vigentina, Porta Lodovica, Porta Ticines...
Municipio 6	Barona, Lorenteggio	Porta Genova, Conchetta, Moncucco, Barona, Qua...
Municipio 7	Baggio, De Angeli, San Siro	Vepra, Quartiere De Angeli - Frua, San Siro, Q...
Municipio 8	Fiera, Gallaratese, Quarto Oggiaro	Porta Volta, Bullona, Ghisolfa, Portello, Cagn...
Municipio 9	Stazione Garibaldi, Niguarda	Porta Garibaldi, Porta Nuova, Centro Direziona...

As you can see in the above table district were stored a lists inside the variable district. I had to perform some further wrangling to obtain a clean list of areas and districts that was fitted to be fed to geocoding algorithms.

2.2 Geocoding

Once I had a dataframe with a single row for every one of the 161 district I created a new variable *address* as "district_name, Milano, Lombardia" and then I was able to apply the gecode function from the geopy's python package Nominatim and retrieve geo-spatial coordinates for each district.
Thanks to the folium package we can plot a map of Milan and put markers on it in correspondence of every district location as shown in the figure below.

Figure 1: Map of Milan's Districts

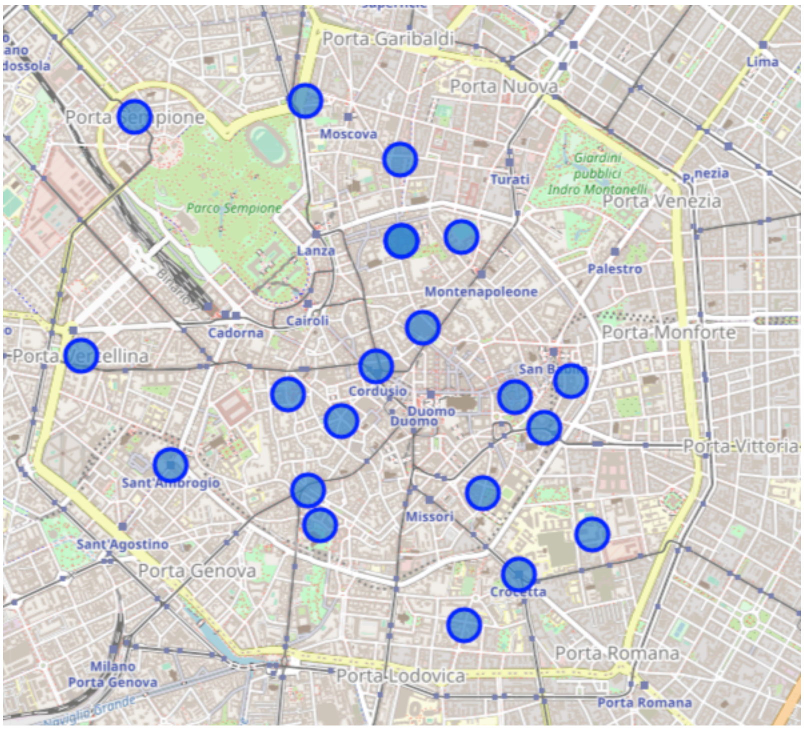


The objective of this analysis is to cluster districts of Downtown Milan, hence I selected a sub-sample of the 161 total clusters keeping only districts located in the most central part of Milan, namely the ones belonging to the *Municipio 1* area. I obtained a sub-sample of 20 districts listed in the table below and plotted on the map below.

Table 3: Dataframe with Coordinates

area	district	latitude	longitude
Municipio 1	Cordusio	45.465832	9.186094
Municipio 1	Cinque Vie	45.463354	9.183900
Municipio 1	Brisa	45.464559	9.180482
Municipio 1	Brera	45.471519	9.187735
Municipio 1	Scala	45.467605	9.189120
Municipio 1	Sant'Ambrogio	45.461391	9.172917
Municipio 1	Carrobbio	45.460262	9.181695
Municipio 1	Verziere	45.463094	9.196925
Municipio 1	Pasquirolo	45.464445	9.195079
Municipio 1	Borgonuovo	45.471675	9.191571
Municipio 1	Brolo – Pantano	45.460137	9.193010
Municipio 1	Crocetta	45.456475	9.195268
Municipio 1	Quadronno	45.454182	9.191743
Municipio 1	Vetra	45.458709	9.182518
Municipio 1	Brera	45.471519	9.187735
Municipio 1	Porta Tenaglia	45.477821	9.181593
Municipio 1	Porta Sempione	45.477128	9.170598
Municipio 1	Porta Magenta	45.466327	9.167105
Municipio 1	San Marco	45.475203	9.187694
Municipio 1	Guastalla	45.458252	9.200023
Municipio 1	Borgogna	45.465188	9.198664

Figure 2: Map of Municipio 1's Districts



In the map above we can see the heart of Milan, inside the ancient walls that are still clearly visible on the map forming a polygon inside of which we have all the historic districts of *Municipio 1.* In the upcoming part of the analysis I am going to collect all the main venues of these districts using the Forsquare API. Then I will try to cluster them based of those venues.

2.3 Foursquare

Forsquare is a local search-and-discovery mobile app developed by Foursquare Labs Inc. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history (Wikiepedia definition).

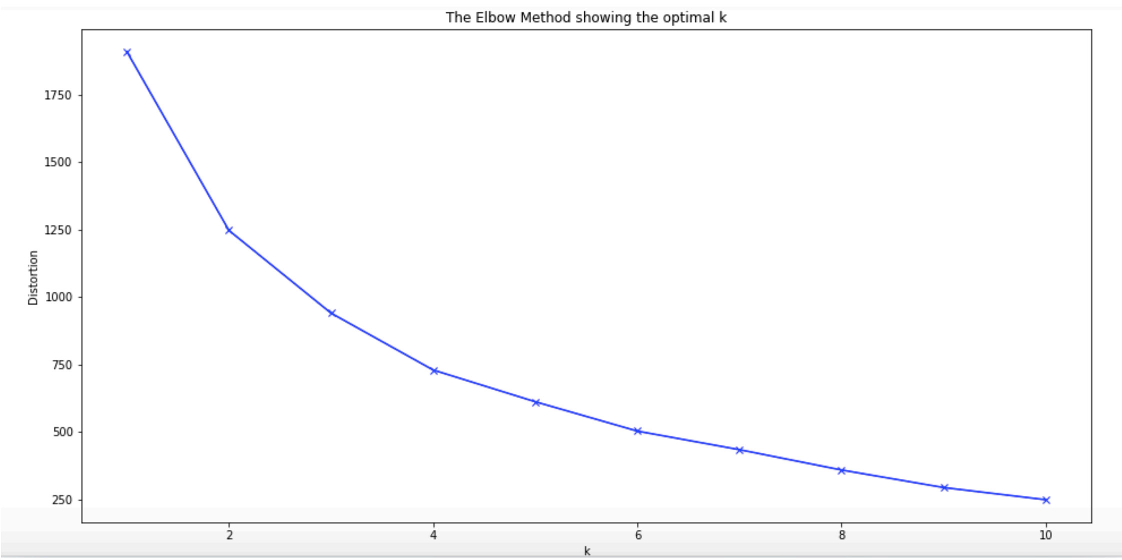
Having the spatial coordinates for each of the districts of interest, Through the Forsquare API in python I can retrieve all the *places to go* in a radius of 300 feet around the location of the district. Now I can characterize every districts based on the number and types of venues it contains. I found that on average a district contains around 45 venues, with a maximum of 130 in Brea (no surprise) and a minimum of 4 in Quadronno (where I live). Furthermore, I found 137 unique venue categories in Milan' Municipio 1.

Before proceeding with cluster analysis I had to re-shape my dataframe in order to have one row for each district and a columns for each one of the 237 unique categories, containing the number of venues of that category in that district.

2.4 Cluster Analysis

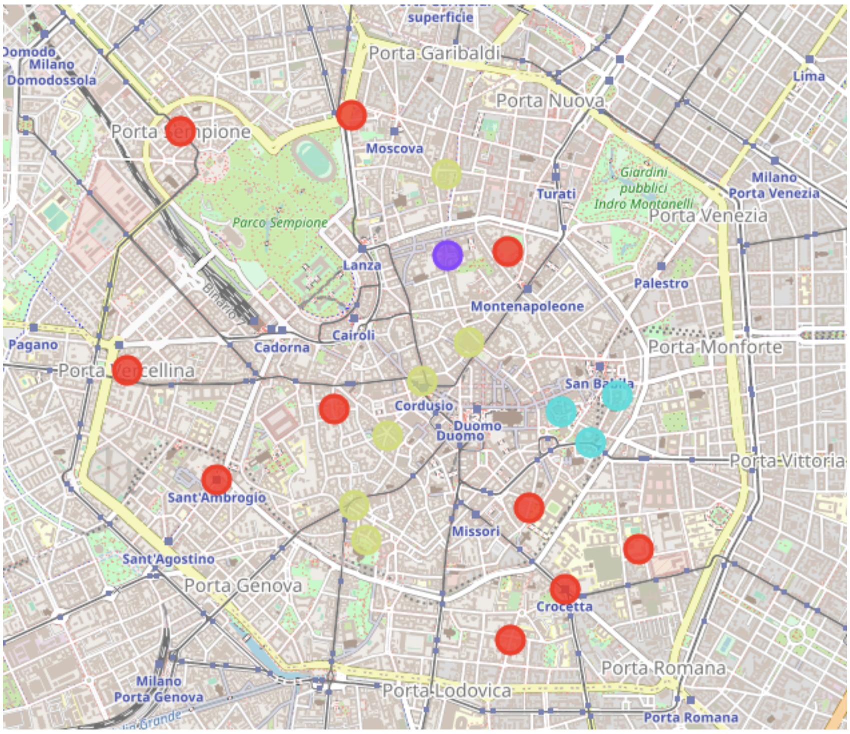
I chose the k-means algorith to cluster the districts in my dataframe. When using k-means, one is suppose to choose the parameter k being the number of cluster one wants to obtain. There is no official rule regarding how to best choose the value of k. One good practice is to choose the value of k that minimizes a measure of distortion. Distortion always decreases as the value of k increases, hence I chose k by applying the *elbow rule*. This rule roughly consists in looking at the distortion measure plotted on the values of k and choose the value of k corresponding to the point where the marginal decrease of distortion becomes significantly flatter, forming an elbow shape.
I tried out 10 values of k, from one to ten, and the results are plotted below.

Figure 3: Elbow Rule



Looking at the above figure, there is no clear elbow. Nonetheless, $k = 4$ seemed to me a good enough educated guess.So I proceeded with $k = 4$, I obtained my 4 clusters and plotted them in the map below.

Figure 4: Clusters



As a next step, I looked into each one of them and tried to gave them meaning by assigning a meaningful label to each one of them.

In the below table wee see districts belonging to Cluster 0. This cluster is the most numerous and also the most difficul to characterize. If we look on the map the red markers corresponding to Cluster 0's districts are scattered around the map, mostly near the edges. The number of venues in each districts is smaller than average (44.95) and the most common venues are pretty various across districts.
Being far from the the most touristic areas and having a smaller number of venues, we can label this cluster as Residential.

Figure 5: Clusters 0: Residential

	district	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Venue
2	Brisa	Ice Cream Shop	Café	Theater	Pizza Place	Monument / Landmark	Chocolate Shop	Sandwich Place	Dessert Shop	Fabric Shop	Falafel Restaurant	10
5	Sant'Ambrogio	Café	Italian Restaurant	Pizza Place	Science Museum	Ice Cream Shop	Supermarket	Emilia Restaurant	Monument / Landmark	Furniture / Home Store	Spanish Restaurant	25
9	Borgonuovo	Hotel	Boutique	Cocktail Bar	Japanese Restaurant	Bookstore	Lounge	Park	Restaurant	College Arts Building	Spa	18
10	Brolo – Pantano	Café	Coffee Shop	Tram Station	Bistro	Bakery	Burger Joint	Pizza Place	Italian Restaurant	Hotel	Lounge	38
11	Crocetta	Café	Bistro	Pizza Place	Italian Restaurant	Hotel	Tram Station	Bakery	Falafel Restaurant	Restaurant	Salad Place	26
12	Quadronno	Burger Joint	Restaurant	Gym	Café	Dessert Shop	Diner	Electronics Store	Emilia Restaurant	Fabric Shop	Falafel Restaurant	4
15	Porta Tenaglia	Italian Restaurant	Wine Bar	Japanese Restaurant	Bakery	Cocktail Bar	Café	Pizza Place	Hotel	Tram Station	Korean Restaurant	34
16	Porta Sempione	Cocktail Bar	Italian Restaurant	Pizza Place	Japanese Restaurant	Tram Station	Lounge	Noodle House	Sandwich Place	Plaza	Pharmacy	43
17	Porta Magenta	Italian Restaurant	Pharmacy	Plaza	Sushi Restaurant	Ice Cream Shop	Pastry Shop	Salon / Barbershop	Cocktail Bar	Church	Design Studio	20
19	Guastalla	Bakery	Restaurant	Pub	Park	Clothing Store	Farmers Market	Food Truck	Pizza Place	Tram Station	Italian Restaurant	12

Figure 6: Clusters 1: Brera, Unparalleled

	district	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Venue
3	Brera	Italian Restaurant	Ice Cream Shop	Cocktail Bar	Pizza Place	Hotel	Art Museum	Arts & Crafts Store	Plaza	Lounge	Wine Bar	136

As shown in Figure 6 Brera ended up havig its own personal cluster. This is not surpising given the well known uniqueness of Brera. With 136 venues it has more than two times the average number of venues and in addition to the various food and drink places Brera is characterize by the presence of a number of Art Museums and Art Stores, indeed in Brera we find one most historic and worldwide famous art academy: the Brera Academy of fine-arts. It is mostly the presence of art and artists to have made Brera unparalleled among Milanese districts.

Figure 7: Clusters 2: Shopping

	district	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Venue
7	Verziere	Sporting Goods Shop	Italian Restaurant	Bistro	Japanese Restaurant	Hotel	Cosmetics Shop	Plaza	Pizza Place	Furniture / Home Store	Clothing Store	46
8	Pasquirolo	Clothing Store	Sporting Goods Shop	Cocktail Bar	Plaza	Italian Restaurant	Bistro	Furniture / Home Store	Hotel	Asian Restaurant	Bar	68
20	Borgogna	Boutique	Clothing Store	Furniture / Home Store	Italian Restaurant	Sporting Goods Shop	Cocktail Bar	Cosmetics Shop	Sandwich Place	Plaza	Shoe Store	68

The above cluster is mainly characterized by the presence of shops, stores and butiques. District in this cluster have also a relatively high total number of venues and are located all very close to each other on the map. That is why I labeled this cluster as Shopping.

Figure 8: Clusters 3: Tourists Eating Italian

	district	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Venue
0	Cordusio	Italian Restaurant	Hotel	Plaza	Ice Cream Shop	Monument / Landmark	Sandwich Place	Cosmetics Shop	Bakery	Café	Food Court	53
1	Cinque Vie	Italian Restaurant	Plaza	Cosmetics Shop	Ice Cream Shop	Café	Sandwich Place	Gift Shop	Hotel	Furniture / Home Store	Coffee Shop	51
4	Scala	Italian Restaurant	Hotel	Ice Cream Shop	Lounge	Bar	Clothing Store	Coffee Shop	Bookstore	Monument / Landmark	Pastry Shop	57
6	Carrobbio	Italian Restaurant	Café	Ice Cream Shop	Cocktail Bar	Salad Place	Gift Shop	Fast Food Restaurant	Thrift / Vintage Store	Historic Site	Pizza Place	63
13	Vetra	Italian Restaurant	Ice Cream Shop	Cocktail Bar	Café	Bistro	Historic Site	Pizza Place	Gift Shop	Boutique	Hotel	78
18	San Marco	Italian Restaurant	Café	Diner	Restaurant	Bar	Burger Joint	Japanese Restaurant	Convenience Store	Plaza	Peruvian Restaurant	53

If we take a look at the green markers on the map in Figure 4, we notice they are placed along an quasi-straight line. That line is the most inflated itinerary for tourists. The other thing we can immediately notice by looking at the above table is that the 1st most common venue for all 6 districts of this cluster is Italian Restaurant. Another peculiar characteristic of this cluster is the presence of Monuments, Landmark, Historic Sites and also Hotels. Hence, it is quite straightforward to name this cluster Tourists Eating Italian.

3 Results Discussion

Feeding a dataframe containing the number of venues for each different category (Restaurant, Store, Museum, Hotel, etc.) for 20 districts of the most central area of Milan to the k-mean algorithm, I found 4 very clear and distinct clusters.

The first cluster made of mostly residential districts, with relatively low level of economic activities (in terms of total number of venues. The second cluster made of Brera alone

and we saw how Brera is infact unique among its peers in terms of both total number of venues and most populare categories of venues.
A third cluster made of districts very close to each others and characterized by a relatively high presence of shooping oriented categories of venues. The forth and last cluster is made of districts that very well represent a typical guided tour for tourists in Milan with a very high concentration of Italian restaurants.
Even without any previous knowledge about the city of Milan, I think these cluster make sense. The k-means unsupervised algorithm was able to put together districts in a smart way.

4 Conclusions and Future Directions

Analyses like this one are not performed very often on the Italian territory. Indeed it is very hard to find publicly available shapefiles for Italy, or datasets with sub-regional coordinates. It hase been impossible to even find a list of Milan's postal codes by looking on the internet, and Milan is by far the most relevant city in Italy.
This study can work as a positive example of the great potential of simple techniques like the k-mean algorithm. A cluster analysis like this is very useful for different purposes, wether a private investor is scouting the area to exploit untapped business opportunities or wether the local government wants to have a clearer view of the current situation of the different districts. It can also be easily replicated in a different area.

5 References

- Milan - Wikipedia
- Forsquare
- Google Maps