Clustering Districts in Downtown Milan

Coursera Capstone Project

Francesco Bellogi - Apr/2020

Clustering districts can be a very useful exercise for many different purposes

- There is a lack of this kind of studies for Italian cities.
- Geo-spatial datasets are not publicly available even for a city like Milan
- 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.
- Not all districts are the same, some of them have been growing over the years and others have fallen behind.
- Having a more clear picture of the current situation of the various in districts in downtown Milan can be useful for anybody who is looking for **business opportunities**, for a f**amily wanting to buy a house**, for the l**ocal government having to distribute finances**.

Data acquisition and cleaning

- No ready available dataset containing Milan's districts.
- The name list of districts has been obtained by scraping the Wikipedia page https://it.wikipedia.org/wiki/Municipi_di_Milano
- Some data wrangling was needed in order to obtain a clean full list of districts name.
- I created an address variable for each district and used the geopy python package to obtain the geographical coordinates for each address.
- I selected a subsample of districts belonging to the Municipio 1 area, which is the downtown area of Milan, obtaining a new sample of 20 districts.
- For those 20 districts I used the Forsquare API to get all the venues in a radius of 300 feet for each district.

Downtown districts

An extract of the data frame with all the venues per each district

10 venues in Cordusio, Milan

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Cordusio	45.465832	9.186094	Starbucks Reserve Roastery	45.464920	9.186153	Coffee Shop
1	Cordusio	45.465832	9.186094	Venchi	45.465214	9.187340	Ice Cream Shop
2	Cordusio	45.465832	9.186094	Palazzo della Ragione	45.464792	9.187785	Monument / Landmark
3	Cordusio	45.465832	9.186094	Park Hyatt Milan	45.465532	9.188911	Hotel
4	Cordusio	45.465832	9.186094	Bialetti Store	45.464775	9.188343	Kitchen Supply Store
5	Cordusio	45.465832	9.186094	Panini Durini	45.465238	9.188590	Sandwich Place
6	Cordusio	45.465832	9.186094	Loggia dei Mercanti	45.464946	9.187911	Plaza
7	Cordusio	45.465832	9.186094	Andry	45.467240	9.184450	Italian Restaurant
8	Cordusio	45.465832	9.186094	Ciacco. Gelato senz'altro	45.463704	9.186796	Ice Cream Shop
9	Cordusio	45.465832	9.186094	Città del Sole	45.464870	9.186752	Toy / Game Store

One Hot Encoding

- 20 districts
- 137 unique catgories

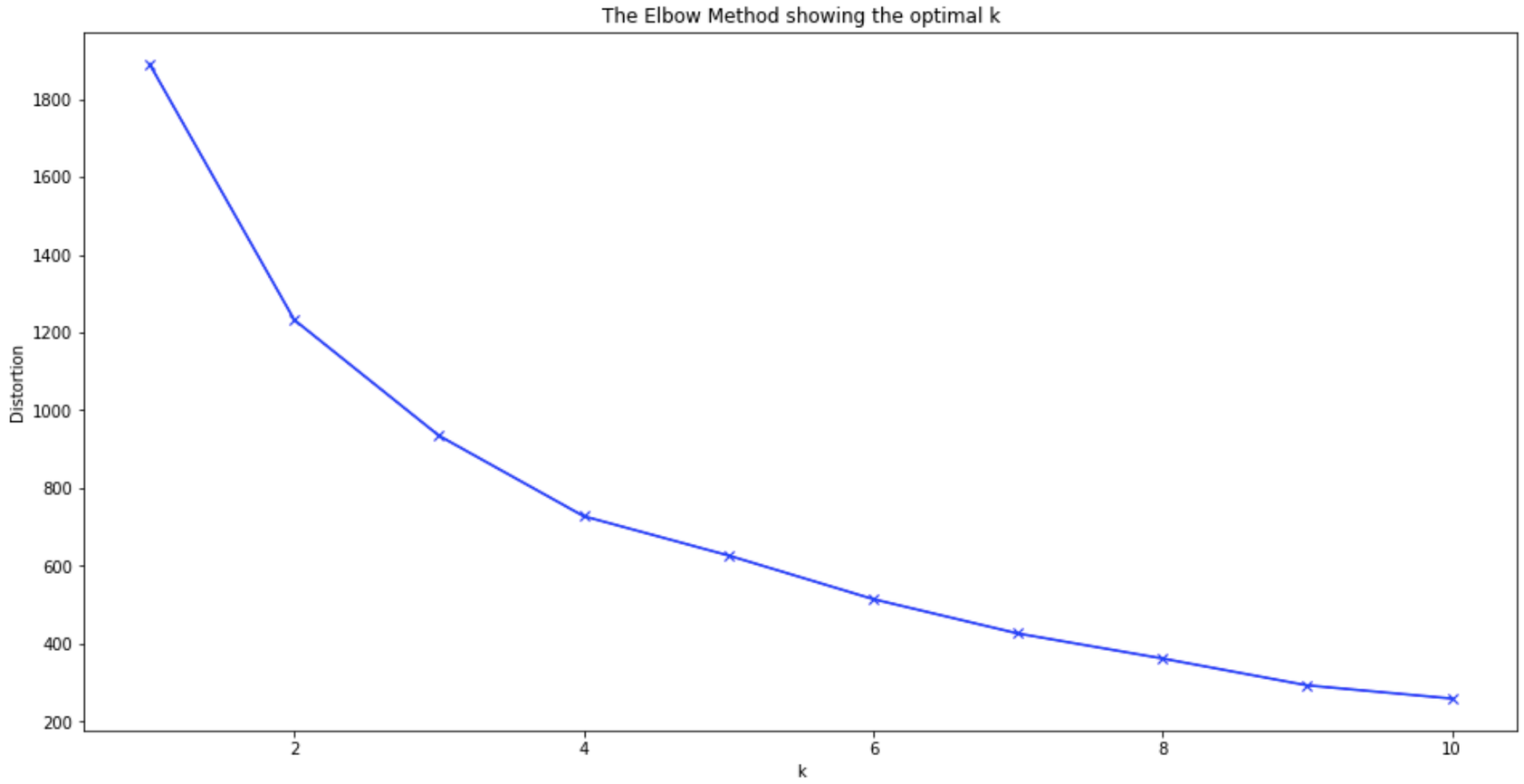
	Neighborhood	Accessories Store	American Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Auditorium	Bakery	Bar	...	Thea
0	Borgogna	1	0	0	0	0	0	0	0	0	...	
1	Borgonuovo	0	0	0	1	0	0	0	0	0	...	
2	Brera	0	0	0	4	4	0	0	2	4	...	
3	Brisa	0	0	0	0	0	0	0	0	0	...	
4	Brolo – Pantano	0	1	0	0	0	1	0	2	1	...	
5	Carrobbio	0	0	0	1	0	0	0	0	1	...	
6	Cinque Vie	0	0	1	1	0	0	0	0	2	...	
7	Crocetta	0	0	0	0	0	0	0	2	1	...	
8	Guastalla	0	0	0	0	0	0	0	2	0	...	
9	Pasquirolo	1	0	0	0	0	2	0	0	2	...	
10	Porta Magenta	0	0	0	1	0	0	0	0	0	...	
11	Porta Sempione	0	0	0	0	0	0	0	1	0	...	
12	Porta Tenaglia	0	0	0	0	0	0	0	2	0	...	
13	Quadronno	0	0	0	0	0	0	0	0	0	...	
14	San Marco	0	0	1	1	0	0	0	1	3	...	
15	Sant'Ambrogio	0	1	0	0	0	0	0	0	0	...	
16	Verziere	1	0	0	0	0	0	0	0	1	...	
17	Vetra	0	0	0	1	0	0	0	0	2	...	
18	Cordusio	0	0	1	1	0	0	0	2	0	...	
19	Scala	0	0	1	1	1	0	1	1	2	...	

20 rows × 137 columns

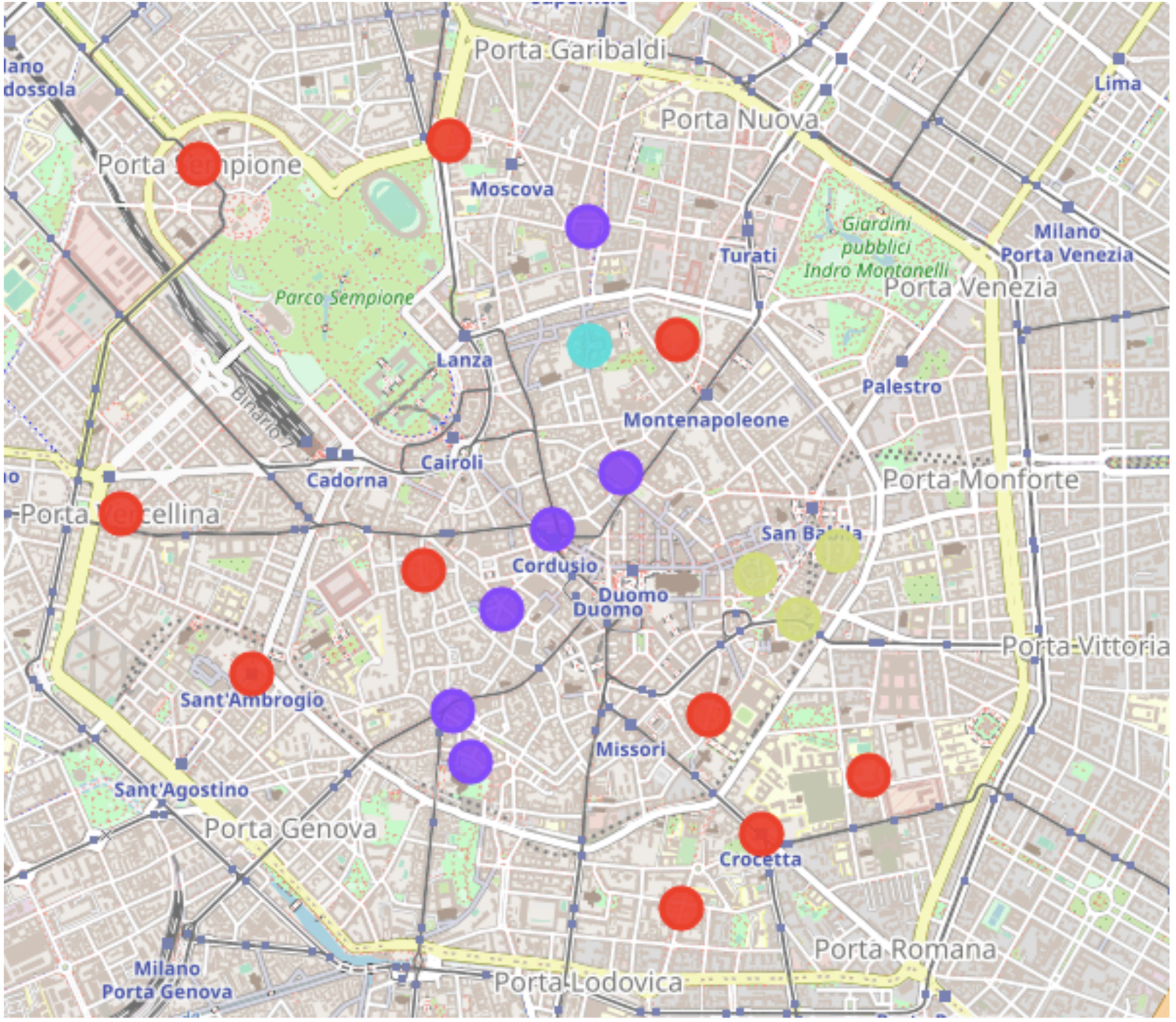
Clustering with k-means

Choice of k

- Plot a measure of distortion on the value of k
- A possible optimal value of k is a the point in which the slope of the curve shows an “elbow2, becoming suddenly flatter
- No clear elbow in this case, **k=4 was the final choice**



Final clusters



Residential

Cluster 0

	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

Brera, Unparalleled

Cluster 1

	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

Brera Design District

Shopping

Cluster 2

	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

Tourists Eating Italian

Cluster 3

	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

Results Discussion

- From a dataset containing the number of venues for each different category (Restaurant, Store, Museum, Hotel, etc.) for **20 districts** of the most **central area of Milan** the **k-means** algorithm 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 in fact unique among its peers in terms of both total number of venues and most popular categories of venues.
- A **third cluster** made of districts very close to each others and characterized by a relatively high presence of **shopping** 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.

Conclusions and Future Recommendations

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 has 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.**