IBM Applied Data Science Capstone

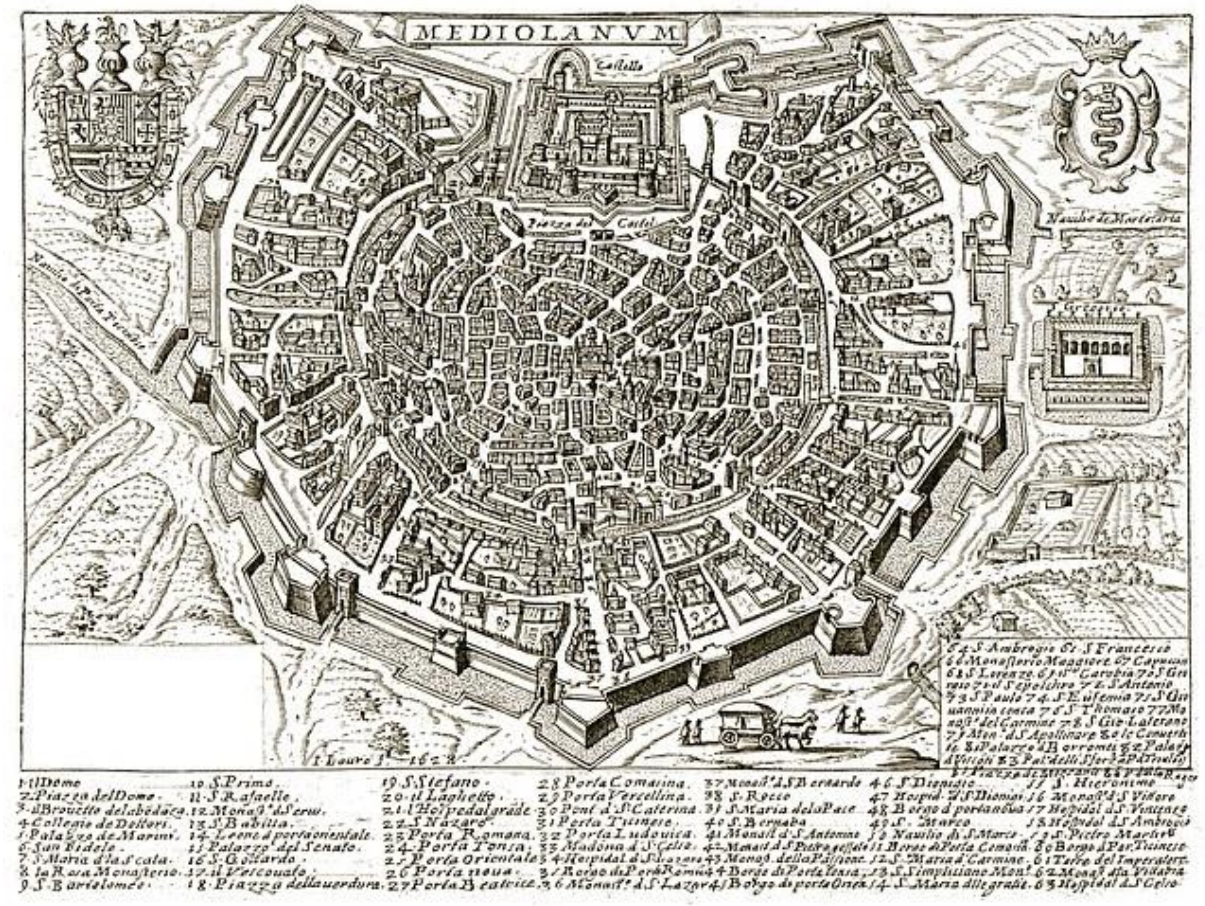
Living in Milano

M.B.

May 2019

Introduction

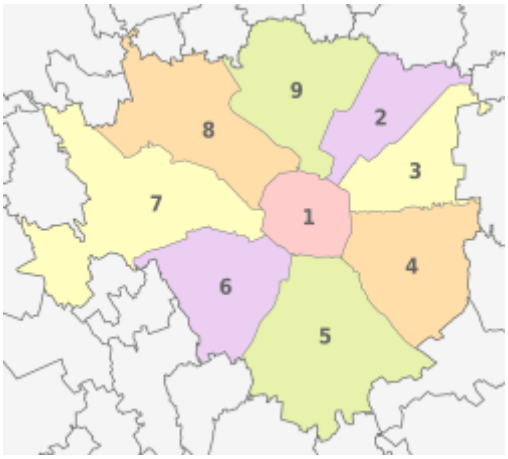
Funded as Medhelan by the Celtic Insubres around 600 BC and then renamed as Mediolanum ("plain in the middle") by the Romans around 220 BC, Milano is the capital of Lombardy in northern Italy. It served as capital of the Western Roman Empire from 286 AC to 402 AC, then became a Duchy during the medieval period (when prospered as a centre of trade, due to its position) and early modern age, being protagonist of the italian Renaissance before serving as the capital of the satellite Napoleonic Kingdom of Italy at the beginning of the 19th century [1].



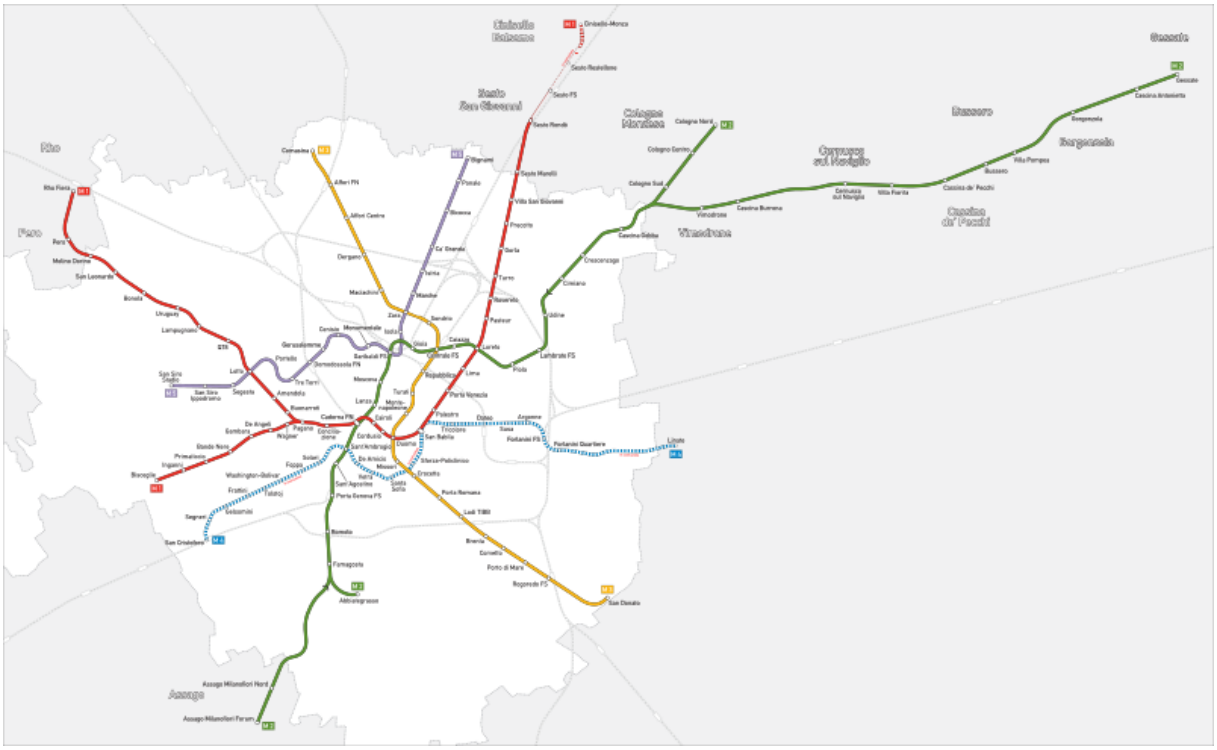
With its long history, Milano is regarded a leading global city in the eclectic fields of finance, commerce, services, healthcare, tourism, cuisine, art, design, fashion, entertainment, media, research and education, with numerous museums, galleries and theaters as well as cultural institutions, academies and universities. Hosting the Italian stock exchange ("Borsa") as well as the headquarters of both national and international banks and companies since the early 20th century, Milano is the wealthiest among european non-capital cities and has the fastest economic growth; it is also part of the "Blue Banana", due to the post-war economic boom and the recent technological advancements. The city hosted the Universal Exposition in 1906 and 2015 and has been recognized as one of the world's four fashion capitals since the 1980s, with several international high-revenue events and fairs attracting visitors and investments.

Home to successful football and basketball teams as well as famous artists and orchestras, Milano is a modern and dynamic city continuously driven by scientific and business innovation putting peoples and their creativity at its core. Sustainable green architecture is developing to favor a passionate and healthy life in a fully-integrated multicultural environment very rich of opportunities and social aggregation (culture, music, performances, sport, food, night life, ...), surrounded by a belt of green areas and featuring several gardens/parks even in the centre.

Milano has a humid subtropical climate with hot humid summers and cold foggy winters, as the close Alps and Apennines form a natural barrier that protects it from the major circulations coming from northern Europe and the Mediterranean sea, which is less than 150km south. The city is located approximately halfway between the river Po to the south, the foothills of the Alps with the great lakes (Como, Maggiore, Lugano) to the north, the Ticino river to the west and the Adda to the east; its land is flat, the highest point being at about 120m altitude. The urban area of Milano keeps growing and includes a large metropolitan area, which counts an estimated population of 3200000 over about 1600km^2 while the administrative commune itself covers an area of about 180km^2 with an estimated population of 1400000 (with 20% of foreign residents); the concentric layout of the centre reflects the Navigli, an ancient system of navigable and canals now mostly covered. Milano is divided into administrative boroughs and several districts, which are connected via a system of radial streets and circular rings over ground (for peoples, cars, buses and trams) and a metro network under ground (for peoples).



With five major railway stations and three international airports (with almost fifty millions passengers per year), the city is a key transport node and hub in southern Europe as well as the core of Lombardy's regional network. Metropolitana Milanese is the rapid transit system that consists of four lines (yet a fifth one is under construction) identified by different numbers and colors, with a wide network of almost 100km and 113 stations (mostly underground); it connects boroughs and main transport nodes (including suburban railway services and Linate city airport in the near future), with a daily ridership exceeding 1.5 million passengers per day.



Business Problem

While Roma is the political capital, Milano is Italy's industrial and financial heart generating approximately 10% of the national GDP; it contains Europe's most expensive street and its vast province is home to about 45% of businesses in the Lombardy region (which generates approximately 25% of the national GDP), including three famous Fortune 500 companies. The city was an important industrial and manufacturing centre (especially for the automotive, but also for health and pharmaceuticals, chemicals and biotechnologies, food and beverage) and is home to internet and telecommunication companies, national newspapers and publishing companies as well as a large number of media and advertising agencies. As national financial hub, many headquarters of insurance companies and banking groups are located in the city, like most asset management companies, research institutions and professional associations. As a major world fashion centre, Milano is also a global hub for design and trade, especially with the brand new business district; tourism is increasingly important too, with almost ten millions international arrivals per year enjoying events, fairs, shopping, cuisine and art (from the figurative one of the Middle Ages, to Gothic, Renaissance, Baroque, Liberty and Futurism).

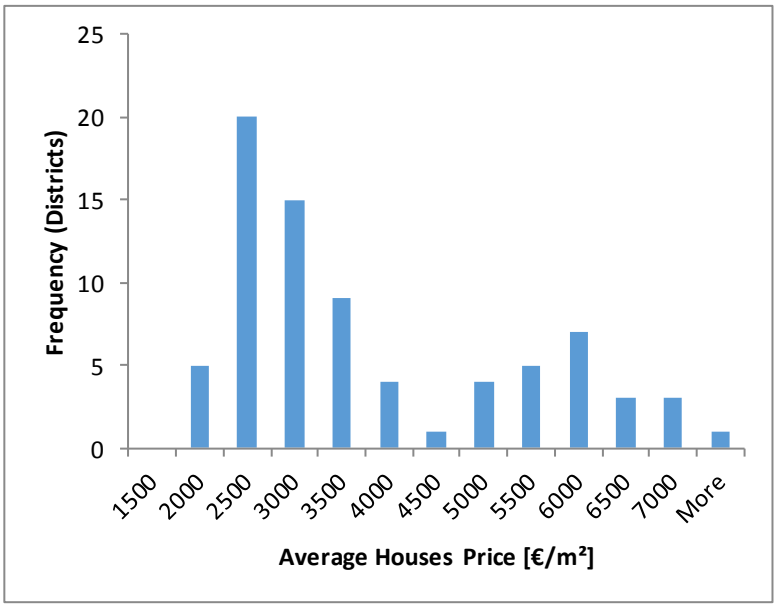
Despite the general economic stagnation, many jobs are available and attracts new residents in the cosmopolitan Milano. Taking advantage of IBM Data Science methodology and Machine Learning techniques, the aim of this project is to identify the best affordable district to live in based on personal preferences for the zone characteristics (venues, transportation, parking, services, schools, parks, …) and perspectives on its development, also from an investor point of view. This project is hence useful for people either relocating to Milano or selling their own house in Milano, property developers and investors looking for remunerative opportunities in real estate, public administrators planning and managing the urbanization of the entire city.

Data

The lists of Milano administrative boroughs, traditional (informal) districts and metro stations are found in [2], [3] and [4], respectively. The average houses price in the districts is found in [5] and refinements based on the close-by metro stations are given in [6] (see the Appendix). A first glance to the data available from [2], [5], [6] is given in the tables and statistics below. Note that the average houses price €/m² considers all residential types without distinction and increases rapidly (although not always monotonically; see M1 and M5) toward the city center.

Borough	Area [km²]	Population	Houses Price [€/m²]	Range
#1	9.7	97403	7045	upper
#2	12.6	159134	3085	lower
#3	14.2	142939	3080	lower
#4	21.0	15975	2700	lower
#5	29.9	124903	2060	lower
#6	18.3	150356	3240	lower
#7	31.3	173643	4325	upper
#8	23.7	186179	3685	lower
#9	21.1	186566	3170	lower

Descriptor	*Price [€/m²]*
Mean	3591
Median	2950
Mode	2250
Minimum	1600
Maximum	7100



Also note that the distribution of the average houses price in the Milano's districts is bimodal and hence suggests ranking the latter as "lower" or "upper" if lower or higher than 4000€/m², respectively. Out of the 77 districts considered here, 53 fall in the "cheap" range (with a sharp frequency peak around 2500€/m²), while the remaining 24 fall in the "expensive" range (with a smooth frequency peak around 6000€/m²). The former class of districts is hence much more common than the latter and the distinction between the two is very neat; this is true also for the administrative boroughs, with #1 standing out as the very rich historical centre of the city while #5 covering one of the most rural areas (which gradually fades into the Pianura Padana). Recall that informal districts will be the focus, since boroughs are too large while metro areas are too small: the former cover the entire city but also a rather vast and heterogeneous own area, the latter do not cover the entire city and their mutual distance does vary considerably; marginal overlap of the districts surfaces is reasonably needed to grant geographic contiguity. As for the Appendix, bear in mind that the metro line M4 is currently still under construction; finally, all metro stations outside the municipality of Milano have coherently been discarded. The outlined raw web data in [1-6] are first collected in the *Milano_rawebdata.xlsx* Excel file, containing the list of administrative boroughs, informal districts and metro stations along with the relative information about average houses price; demographics is also available for the former only. As the "popular" nature of the districts brings variability in their definition from different sources, clean consolidated data are eventually provided in *Milano_dataframe.xlsx*, already containing average houses price range as well as latitude and longitude (grasped via Python's Geocoder) of all locations; Foursquare API will then provide venues information too.

Methodology

In practice, web-scraping techniques will be used whenever most convenient to extract data about the municipality of Milano from Wikipedia pages [1-4], with the help of Python [7] and its Beautiful Soup library [8]. Latitude and longitude of Milano's boroughs, districts and metro stations will be obtained via Geocoder library [9] and allow relating areas and locations based on their geographical coordinates as well as associating the pertinent average houses price

loaded from a *.xls* file of preprocessed web sources [5-6], so to get the most information with the least effort. Note that the Geocoder is not very reliable; thus, all geolocation data will need consolidation (in terms of consistency checks and possible corrections using Google Maps, as the case here). After appropriate data cleaning and wrangling in a structured dataframe form, Foursquare API [10] will be used (with free Sandbox account, subject to limitations) to get the venues in each district along with their category. Scikit-learn library [11] will then be employed to cluster the districts based on their venues frequency (not the average houses price, since we want to find the minimum one among similar areas) and Folium library [12] will be adopted to visualize the result on Milano's map; data standardization and one hot encoding [13] will help clustering via *k*-means algorithm [14]. Looking at both venues information and average houses price, the nature of the derived clusters will hence be investigated and suggestions will be drawn with respect to living or investing there. Although the surroundings typically affect houses price, no explicit functional relation (e.g., linear regression) will be sought between the latter and the area characteristics, due to abundance of variables and lack of data (for which dimension reduction methods such as PCA may help [15] but still without granting sufficient accuracy and robustness). However, clear qualitative correlations will arise and be assessed on quantitative grounds, exploiting data science methodology and artificial intelligence tools to support the decision making process about relocating or investing in a district of Milano.

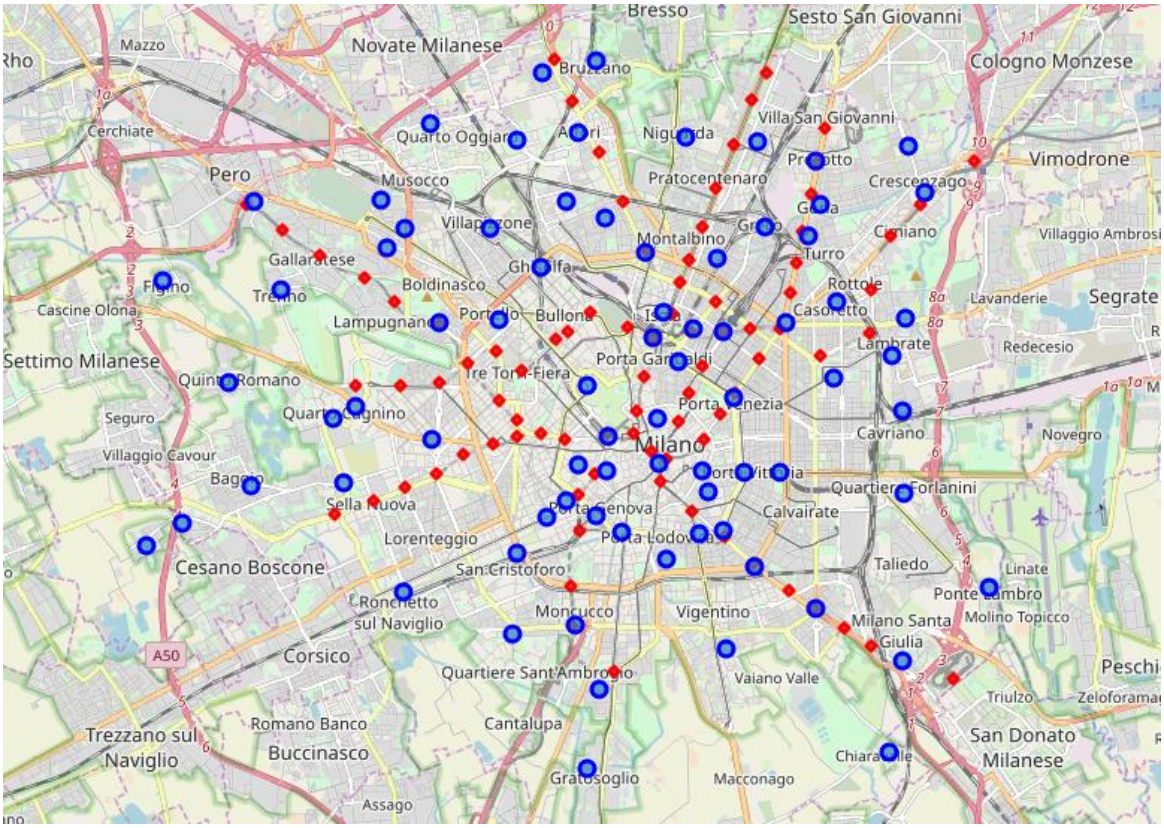
The outlined procedure is fully implemented in the Jupyter Notebook *Living_in_Milano.ipynb*:

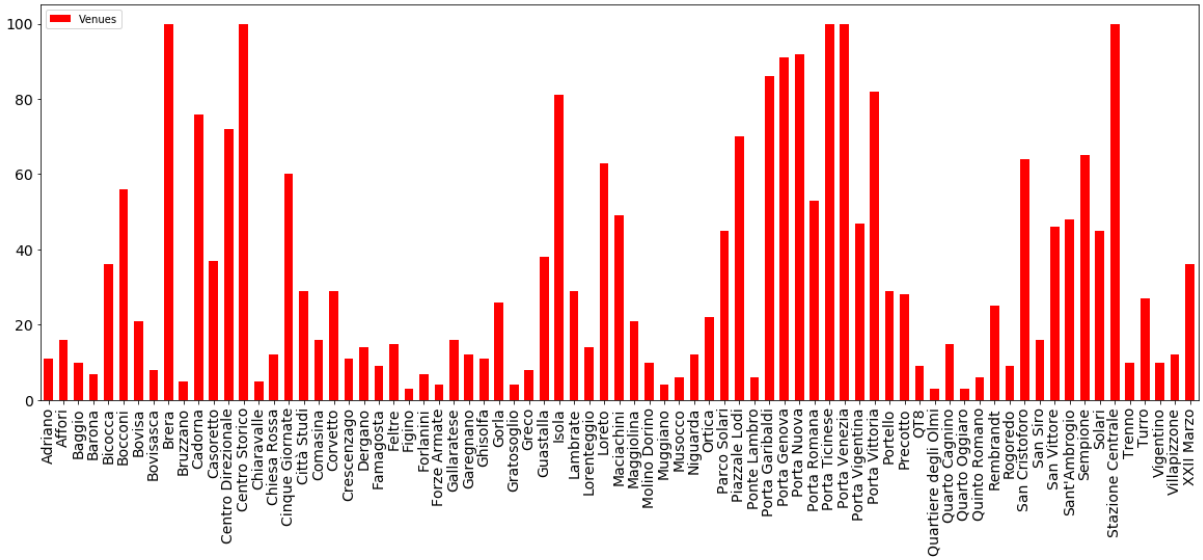
1. Install and import relevant Python libraries
   (pandas, numpy, matplotlib, geopy, geocoder, folium, sklearn, beautifulsoup)
2. Scrape Wikipedia pages for districts and Metro stations in Milano
   (https://en.wikipedia.org/wiki/Category:Districts_of_Milan via beautifulsoup)
   (https://en.wikipedia.org/wiki/Category:Milan_Metro_stations via beautifulsoup)
   a. *Create a (clean) dataframe for Milano's districts and metro stations*
3. Load and analyse raw web data with average houses price
   (*Milano_rawebdata.xlsx* Excel file with boroughs, districts and metro stations information)
   a. *Derive average houses price distributions and rank the class range*
   b. *Merge (clean) information with dataframe for Milano's districts and metro stations*
4. Get the geographical coordinates of Milano's districts and Metro stations
   (latitude and longitude of all considered locations via geocoder)
   a. *Add (clean) information to the dataframe for Milano's districts and metro stations*
5. Load and refine consolidated dataframe of Milano's districts and Metro stations
   (*Milano_dataframe.xlsx* Excel file with boroughs, districts and metro stations information)
   a. *Focus on Milano's districts and Metro stations within the city's outer ring*
6. Create a map of Milano with districts and Metro stations superimposed
   (informative interactive markers are added for all considered locations via folium)
7. Use Foursquare API to explore Milano's districts
   (a personal free Sandbox account is used subject to limitations)
   a. *Get a dataframe of the top 100 venues per district that are within a radius of 500m*
   b. *Check how many venues were returned per district*
   c. *Analyze the venues per district using one hot encoding*
   d. *Get the mean frequency of occurrence of each venue category per district*
   e. *Get the 5 most common venues per district*

8. <u>Group Milano's districts into clusters per venue category</u>
   (in analogy with Milano's boroughs, 10 clusters are sought via *k*-means algorithm)
   a. *Create a map of Milano with Metro stations and all districts' clusters superimposed*
   b. *Show the most relevant clusters of Milano's districts and their average houses price*

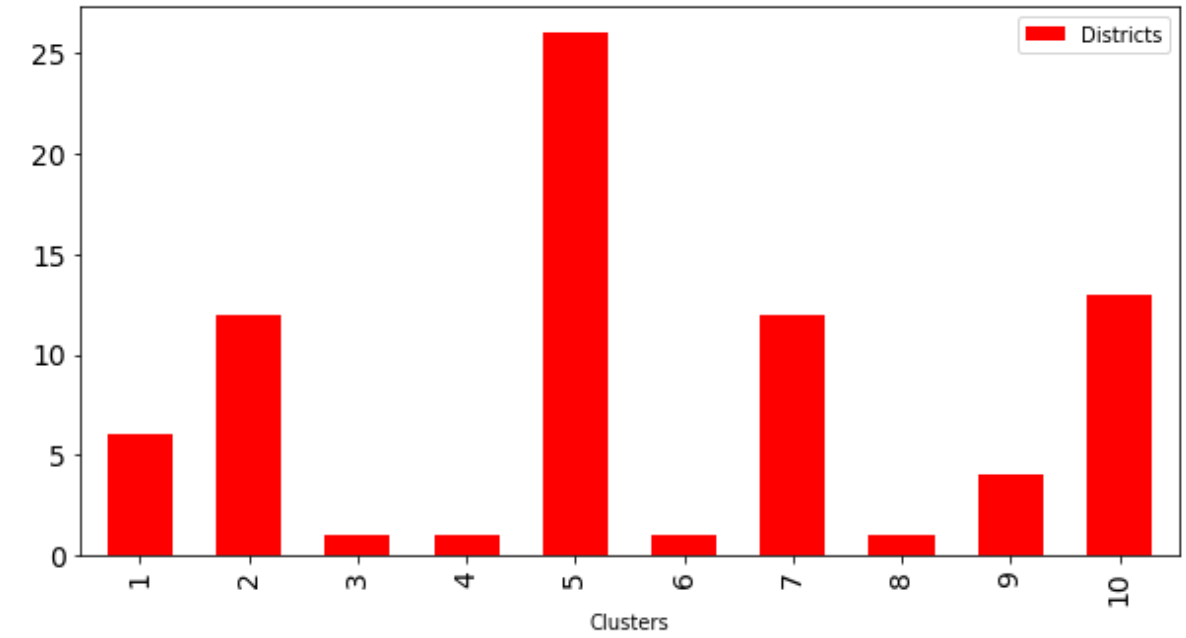
Results

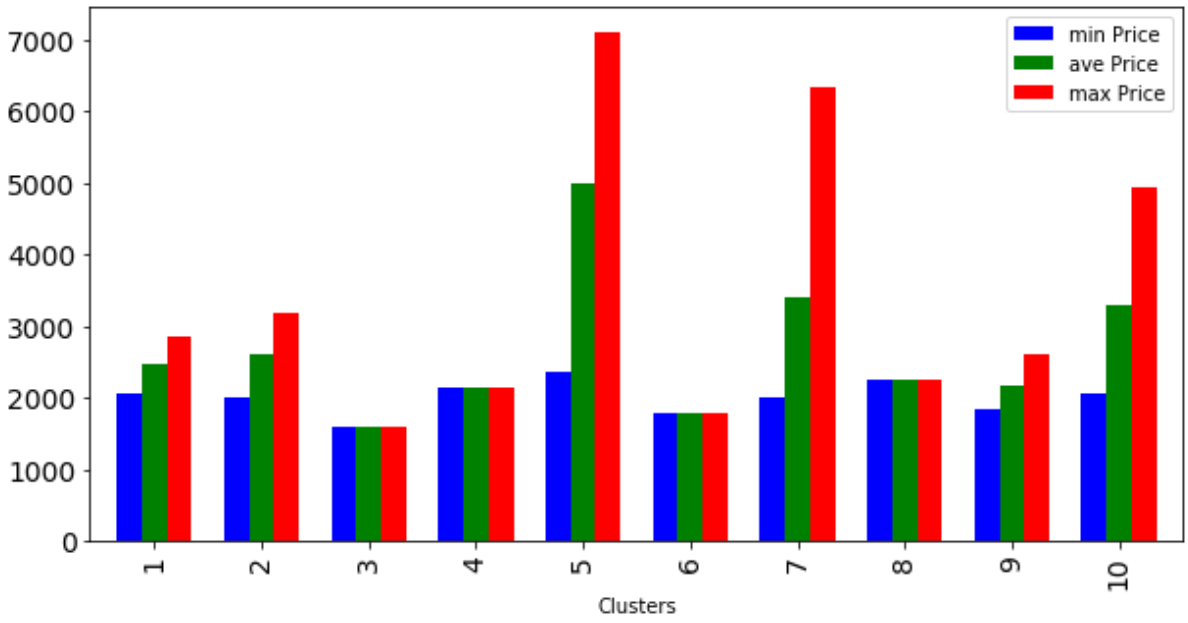
The geographical coordinates of Milano resulted in 45.46679 latitude and 9.19049 longitude; the map below shows the geolocation of all districts and metro stations that were considered.



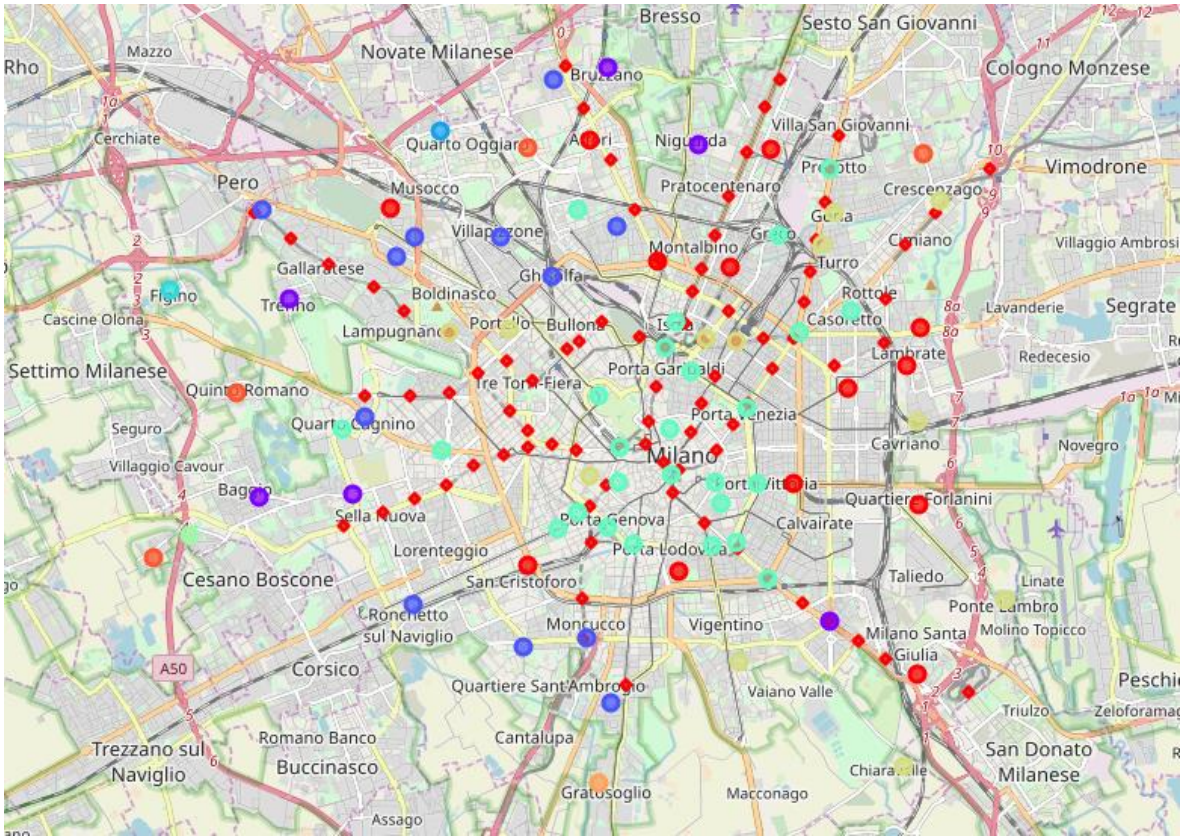


Using Foursquare API, 2573 venues of 254 categories were found in Milano's districts within a radius of 500 from their centres; a limit of 100 venues was also imposed but rarely reached, as evident from the bar chart above. 10 clusters were then sought and found (by the *k*-means algorithm), based on the categories of the 5 most common venues in each district; the bar charts below show the number of districts in each cluster and resulting average houses price (in terms of minimum, maximum and mean values, without size/population-based weighting).





Finally, the clustered districts have been superimposed (with all interactive markers providing the district name, range of the average houses price and cluster number) to the map of Milano shown below; the Metro stations are also visualized, to highlight the public transportation.



The Jupyter Notebook *Living_in_Milano.ipynb* contains all details for each cluster of districts; extracts of consolidated dataframe and Cluster #1 are given below to show the data structure.

	District	Price	Range	Latitude	Longitude
0	Adriano	2600	lower	45.514684	9.245257
1	Affori	2250	lower	45.516972	9.169389
2	Baggio	2250	lower	45.460108	9.094160
3	Barona	2900	lower	45.436325	9.154339
4	Bicocca	2600	lower	45.515597	9.210627

	Station	Line	Borough	Price	Range	Latitude	Longitude
0	Cairoli	1	1	7970	upper	45.468508	9.182230
1	Conciliazione	1	1	6950	upper	45.467686	9.166304
2	Cordusio	1	1	8100	upper	45.465675	9.186362
3	Palestro	1	1	7310	upper	45.471644	9.201917
4	Porta Venezia	1	1	5270	upper	45.475052	9.205523

	District	Price	Range	Cluster	1st Venue	2nd Venue	3rd Venue	4th Venue	5th Venue
26	Forze Armate	2550	lower	1	Pizza Place	Bakery	Supermarket	Fish Market	Falafel Restaurant
19	Corvetto	2750	lower	1	Pizza Place	Gym / Fitness Center	Café	Italian Restaurant	Food Truck
2	Baggio	2250	lower	1	Pizza Place	Café	Japanese Restaurant	Bar	Supermarket
43	Niguarda	2400	lower	1	Pizza Place	Café	Beer Bar	Pub	Hotel
72	Trenno	2850	lower	1	Pizza Place	Adult Education Center	Pub	Park	Bakery
9	Bruzzano	2050	lower	1	Bakery	Italian Restaurant	Gym / Fitness Center	Train Station	Pizza Place

Discussion

The districts distribution granted a rather homogenous coverage of Milano's surface enclosed within both the inner ring (where most of the upper class area are) and the outer ring (where most of the lower class area are), spanning all boroughs; this is not true (not yet, at least) for the Metro stations, as all Metro lines extend along radial directions crossing the city centre. Yet, the density of both districts and Metro stations does gradually increase toward the latter.

Note that many districts have just few venues; the limit of 100 venues was actually reached only in the touristic districts of the city centre as well as in the university and business districts (which attract many services of different kinds), whereas it is typical for the most peripheral districts to have progressively less venues of logistic-like nature (e.g., public stations, parks, parking, stores). Moreover, there are many more venues categories than the districts; this is indeed not ideal for seeking regressions between the average houses price and area features. Yet, the latter successfully drove the clusters identification. In particular, Clusters #3, #4, #6 and #8 contain just one peripheral district (note that minimum, maximum and mean values of the average houses price coincide), characterized by common popular venues like Metro and rail stations, parks, sport and performance sites, markets, bars; note that their average houses prices are in the lower range and do represent the lowest maximum values among all districts. Clusters #1, #2 and #9 also have similar characteristics but include a few districts between the outer and inner rings of the city, thus showing some (small) variability of the average houses price; however, the latter stays rather constant and always within the lower range. Cluster #5 and #7 have similar venues such as restaurants, hotels, café, shops, boutiques and small parks; however, the former includes mostly upper-range districts located within the inner ring (i.e., Milano's historic centre with most of the lucrative business and touristic places), whereas the latter includes mostly lower-range districts located within Milano's outer and inner rings (i.e., where most of the residential places are). It is worth noting that the only three upper-range districts in Cluster #7 include the regional public offices, the biggest rail station and the prison, respectively, whereas the few lower-range districts in Cluster #5 are essentially old areas soon to be renewed (works are already in progress for most of them, in fact). Finally, Cluster #10 includes all major university districts with student aggregation places and residences, where the significant variability of the average houses price is due to the campus type (i.e., public or private) and location (i.e., closer or farer to the city centre and nicest neighborhoods). Note that increasing the number of clusters creates more single-district ones, whereas reducing it still does not group the latter as they were developed in different times (at the opposite sides of the city, rather isolated) and maintain their own specificity, which results in a different order of the most common venues. On the other hand, the closer proximity of the central districts (the exact bounds of which are not considered or even defined in the first place) may lead to venues sharing between neighbor districts that might then easier them clustering together.

Conclusion

Taking advantage of IBM Data Science methodology and Machine Learning techniques, this project derived 10 clusters of Milano's districts based on the categories of their most common venues. From a technical point of view, building consolidated databases of the required inputs was fundamental and took most of the effort, especially to structure and harmonize different data sources in order to answer the business problem consistently; the unsupervised *k*-means algorithm then demonstrated effective and well suited for the job. Finally, interactive maps, histograms and bar charts served as ideal to show data and results. To refine the results in the absence of exact district borders, it is suggested either to relate the latter with the local radius of the venues search or to look for a formal association (possibly already available) based on the venue address, to get the most information and work done with the least effort (generally by all means: few assumptions and clear goals, few constraints and robust approximations,

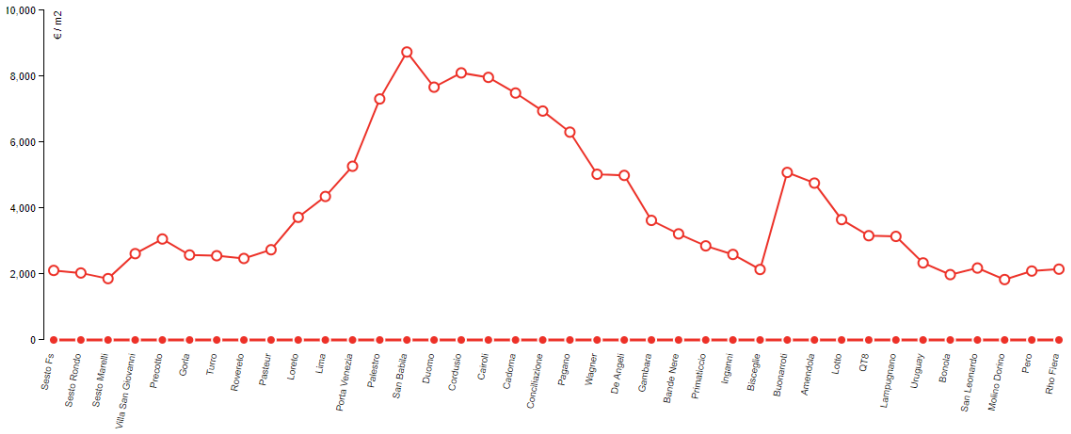
few parameters and meaningful metrics, clean structured data, straightforward logics, smooth slim solution process, simple methods and tools, efficient dimensional reduction techniques). In addition, it is recommended to replace the (free but rather unreliable) Geocoder in Python with Google Maps Geocoding API (unfortunately not free anymore); upgrading the Foursquare account would then remove the limitations on calls and returned information for the venues (note that the information is continuously updated and will hence be different at a later time). It is also suggested to collect districts demographic data and then compute the average houses price for every cluster by weighting that of each district based on its own surface/population. From the business point of view, the identified clusters of districts do reveal clear qualitative indications on the correlation between average houses price and venues categories that may reliably support the decision about where to relocate or invest in Milano, although way more data and additional complexity would be necessary to attempt deriving accurate quantitative models. As for finding the best affordable district to live in based on personal preferences for the zone characteristics: cluster #10 is suggested for students and young researchers, cluster #7 for high-income singles or small families, cluster #5 for medium-income singles or families, clusters #1, #2 and #9 for low-income singles or large families, all remaining clusters for very low-income singles or large families, especially when a spacious house with garden (very rare and expensive within Milano's the inner ring) is wanted. It is often desirable to live close to a Metro station in order to exploit public transportation, but it is also advisable not to live too close to a Metro or railway station, especially in the clusters of lower-range districts. Finally, as for finding the best place to invest in Milano, the legacy lower-range areas in cluster #5 are strongly recommended due to their expected prosperous development in the near future…

References

[1] https://en.wikipedia.org/wiki/Milan
[2] https://en.wikipedia.org/wiki/Zones_of_Milan
[3] https://en.wikipedia.org/wiki/Category:Districts_of_Milan
[4] https://en.wikipedia.org/wiki/Category:Milan_Metro_stations
[5] https://www.idealista.it/news/statistiche/prezzo-linea-metro/milano
[6] https://www.mercato-immobiliare.info/lombardia/milano/milano.html
[7] https://www.python.org/
[8] https://pypi.org/project/beautifulsoup4/
[9] https://pypi.org/project/geocoder/
[10] https://foursquare.com/
[11] https://pypi.org/project/scikit-learn/
[12] https://pypi.org/project/folium/
[13] https://en.wikipedia.org/wiki/One-hot
[14] https://en.wikipedia.org/wiki/K-means_clustering
[15] https://en.wikipedia.org/wiki/Principal_component_regression

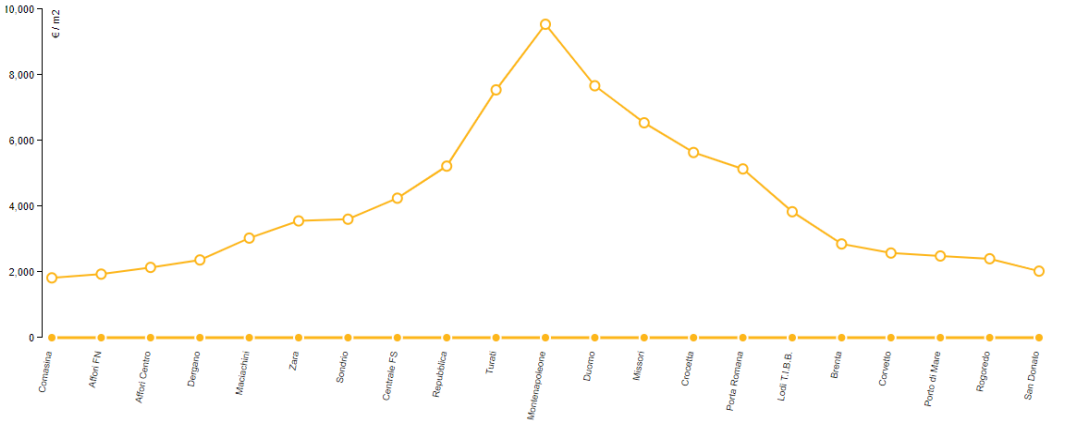
Appendix: Average Houses Price along Metro Lines [5]

M1



M2

M3



M5