IBM Applied Data Science Capstone

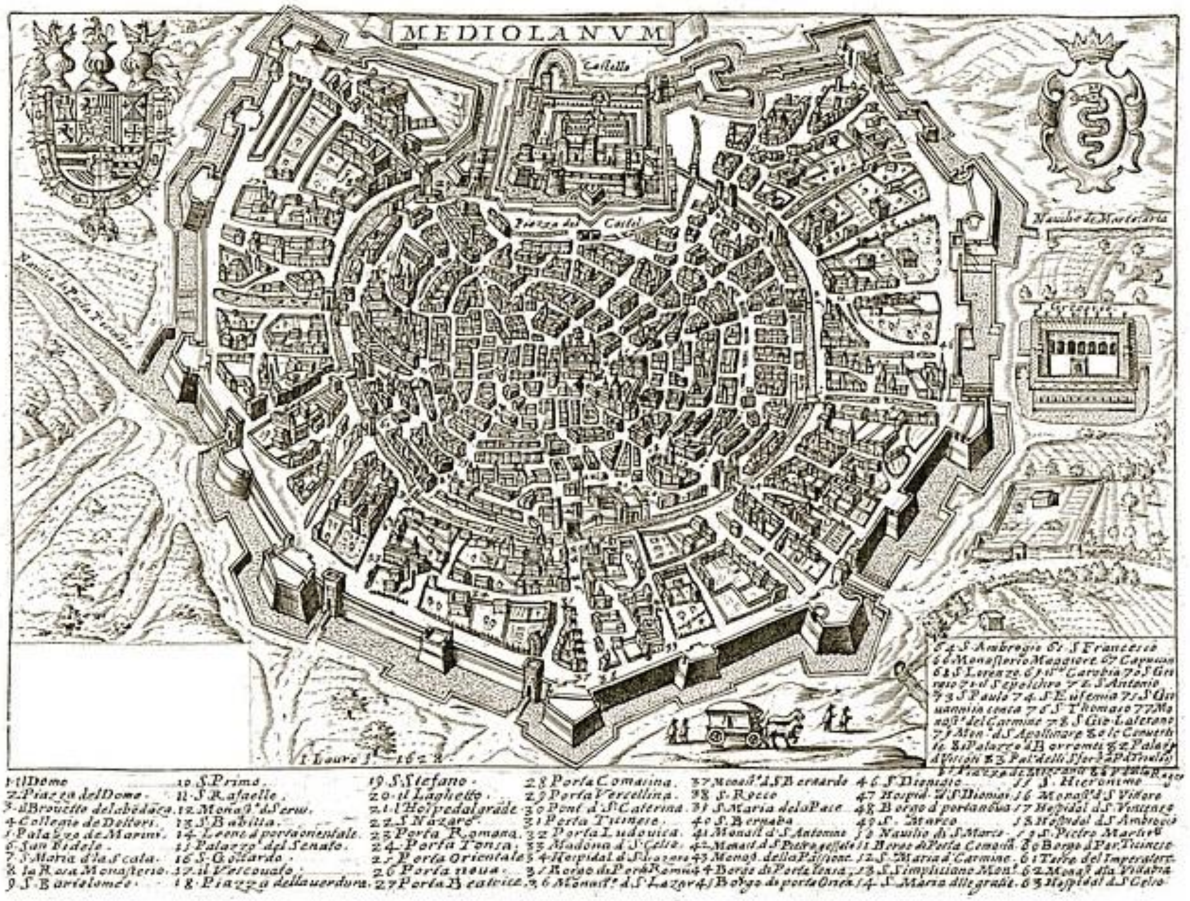
Living in Milano

Marco Berci

May 2019

Introduction

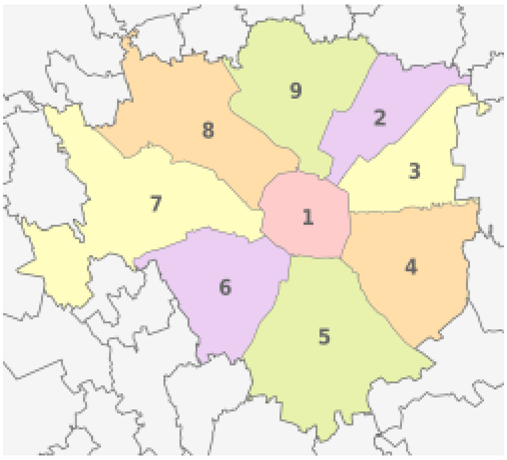
Funded as Medhelan by the Celtic Insubres around 600 BC and then renamed as Mediolanum ("plain in the middle") by the Romans around 220 BC, Milano is the capital of Lombardy in northern Italy. It served as capital of the Western Roman Empire from 286 AC to 402 AC, then became a Duchy during the medieval 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].



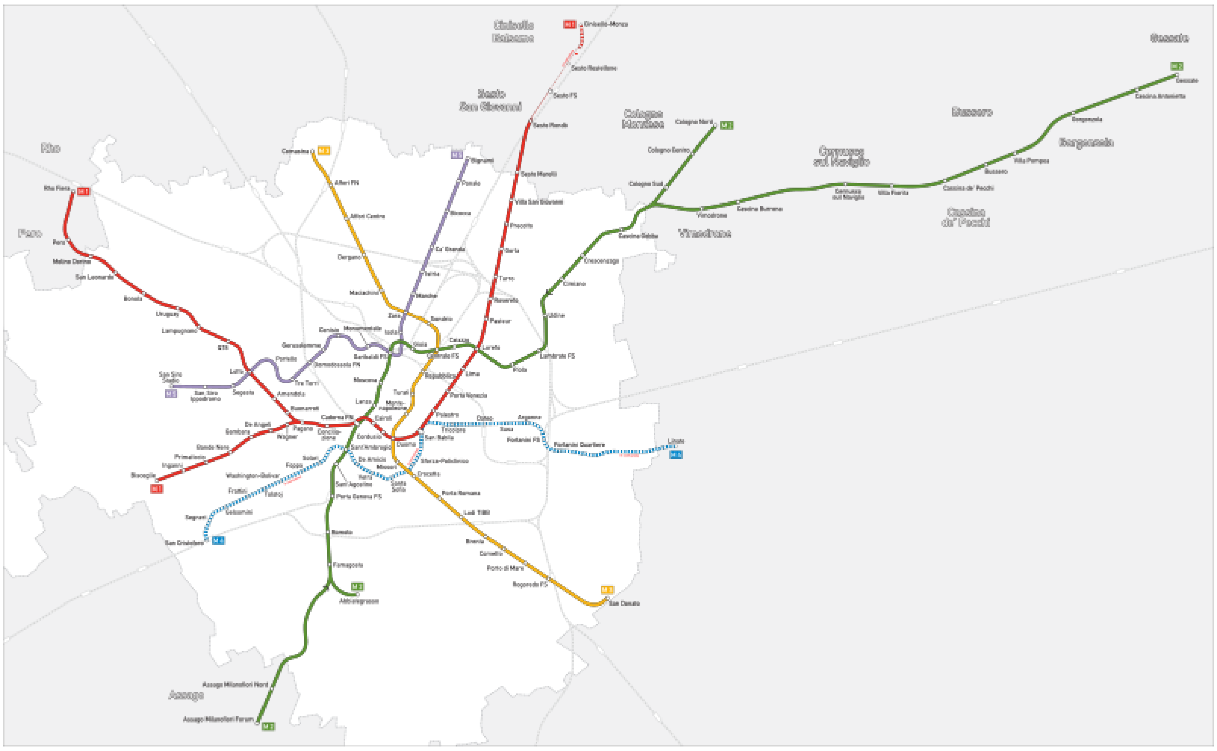
With its long history, Milano is regarded a leading global city in the eclectic fields of finance, commerce, services, healthcare, tourism, cuisine, art, design, fashion, entertainment, media, research and education, with numerous museums, galleries and theaters as well as cultural institutions, academies 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 also increasingly important, 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).

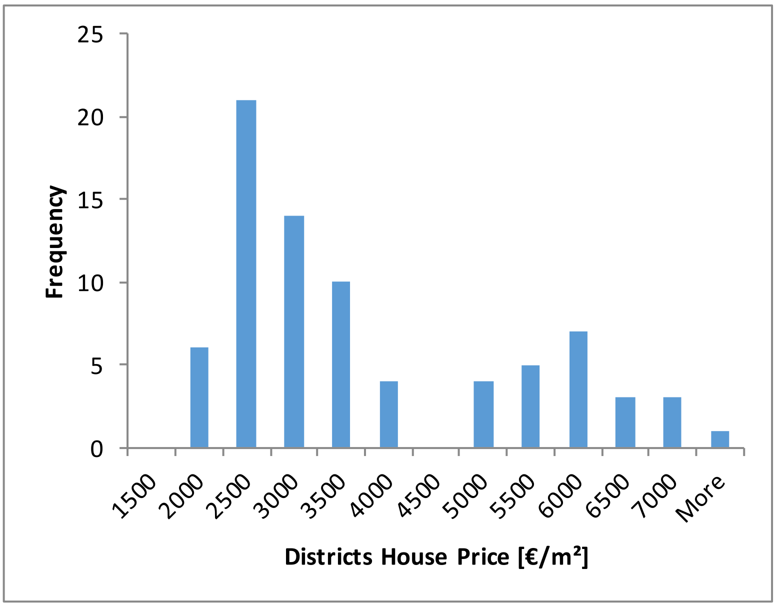
Despite the general economic stagnation, many jobs are available and attracts new residents in the cosmopolitan Milano. Taking advantage of IBM Data Science methodology and Machine Learning techniques, the aim of this project is to find the best affordable district to live, based on personal preferences for the characteristics of the area (e.g., parks, transportation, parking, services, schools, venues, …) and the perspective on their development, from an investor point of view. This project is hence useful for peoples 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 administrative boroughs, traditional (informal) districts and metro stations of Milano 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 €/m^2 considers all residential types without distinction and increases rapidly (although not always monotonically; see M1 and M5) toward the city center.

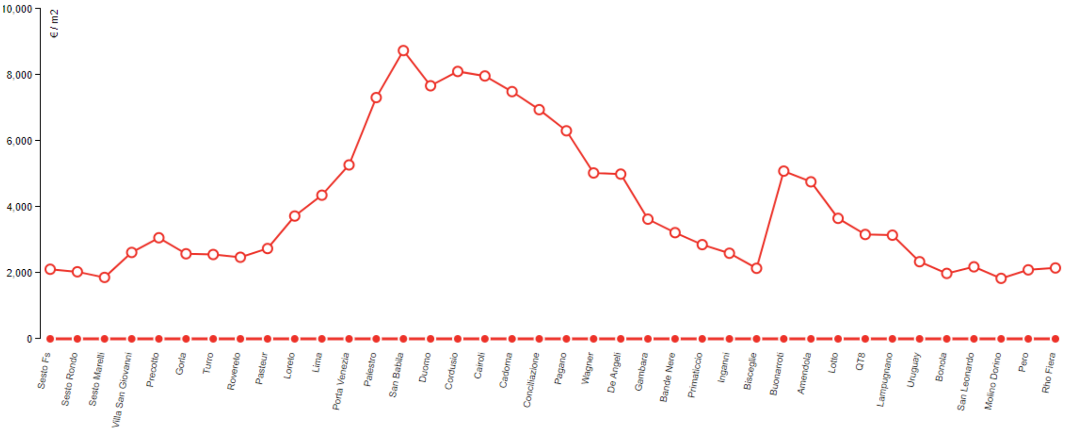
Borough	Area [km^2]	Population	House Price [€/m^2]
1	9.7	97403	7045
2	12.6	159134	3085
3	14.2	142939	3080
4	21.0	15975	2700
5	29.9	124903	2060
6	18.3	150356	3240
7	31.3	173643	4325
8	23.7	186179	3685
9	21.1	186566	3170

Descriptor	*Price [€/m^2]*
Mean	3565
Median	2975
Mode	2500
Minimum	1650
Maximum	7050

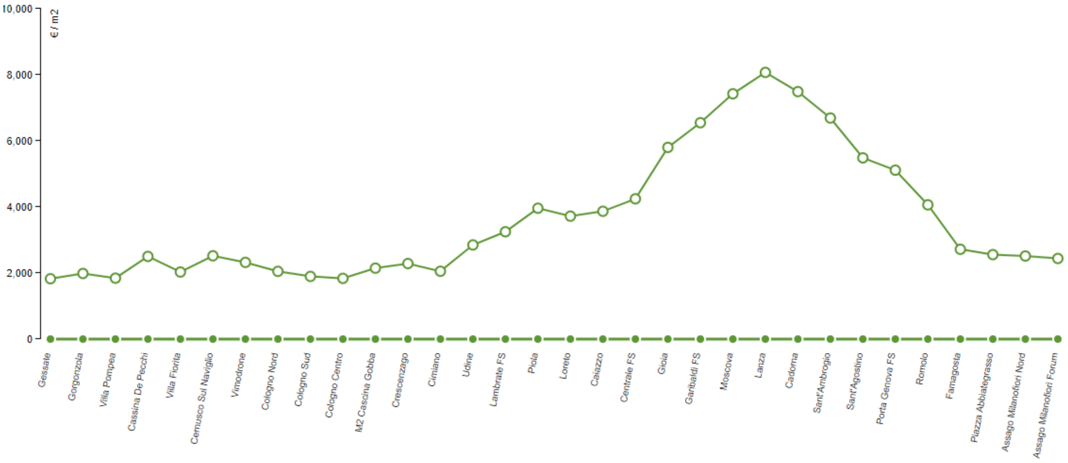


Note that districts are considered in favor of boroughs (too large) and metro area (too small). In particular, web-scraping techniques will be used whenever most convenient to extract data about the municipality of Milano from Wikipedia pages [1-4], with the help of Python [7] and its Beautiful Soup library [8]. Latitude and longitude of Milano's boroughs, districts and metro stations will be obtained via Geocoder library [9], allowing to relate areas and locations based on their geographical coordinates as well as to associate the relevant average houses price loaded from preprocessed web sources (as best practice to get the most information with the least effort). 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.

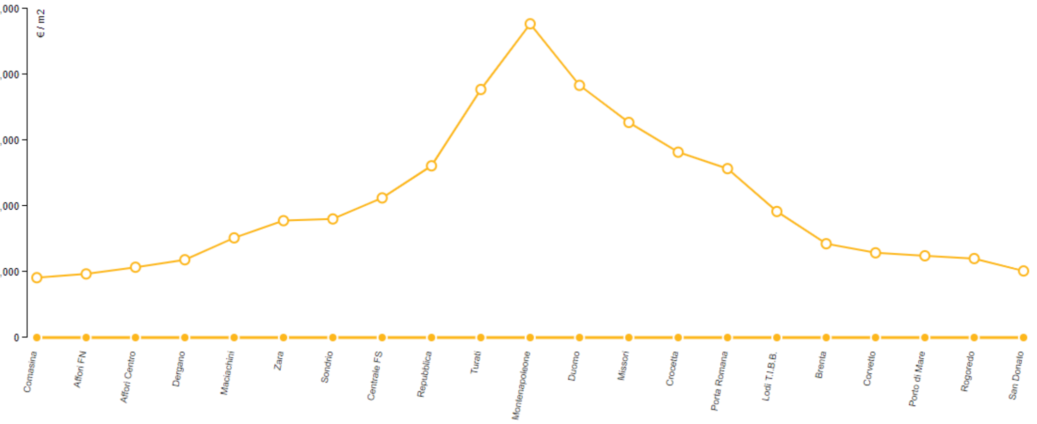
Appendix: House Price along Metro Lines [5]



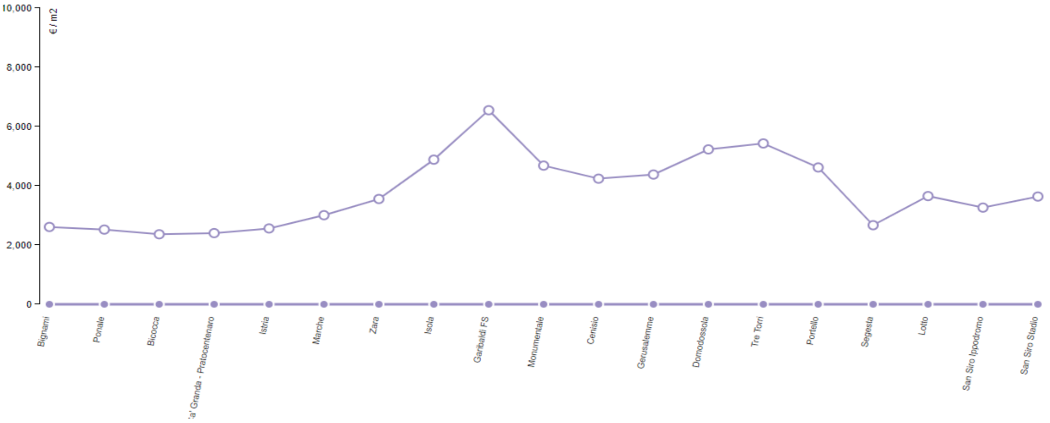
M1



M2



M3



M5

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/ (https://en.wikipedia.org/wiki/Foursquare_(company))
[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