IBM Applied Data Science Capstone

Living in Milano

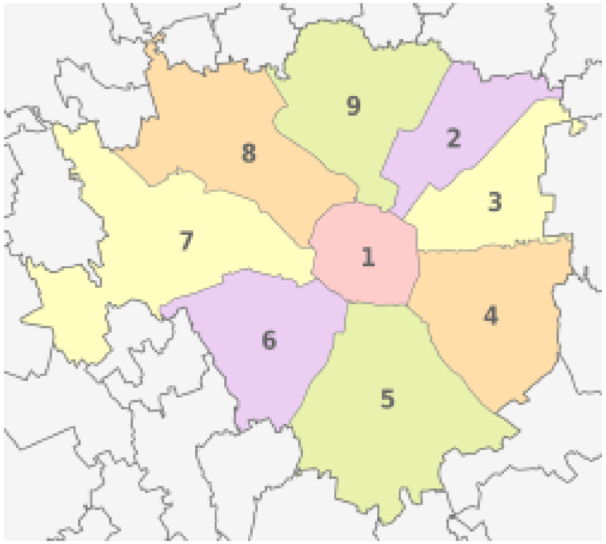
M.B.

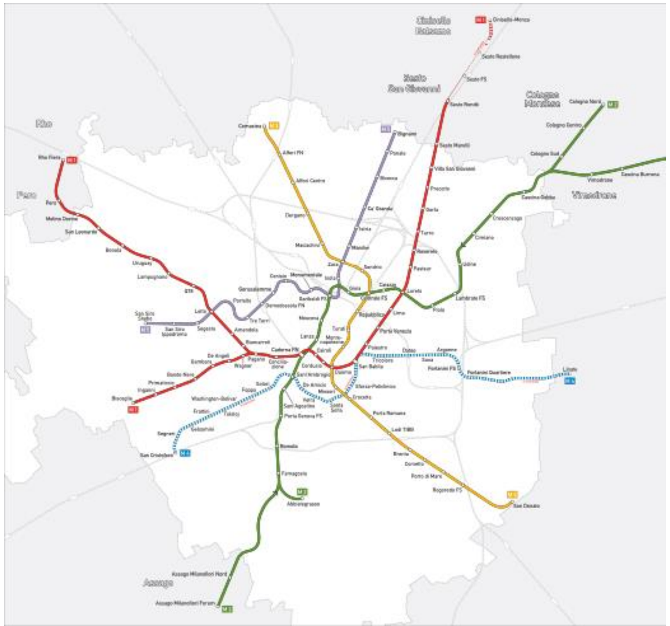
May 2019

Milano: past and present

- Funded as Medhelan by the Celtic Insubres around 600 BC in the north of Italy and then renamed as Mediolanum ("plain in the middle") by the Romans around 220 BC
- Italy's industrial and financial heart, generating approximately 10% of the national GDP
- The wealthiest among european non-capital cities and has the fastest economic growth (it is part of the "Blue Banana“), key transport node in southern Europe
- Global hub for design, art, education, innovation, ITC, industry, business, trade, finance
- Modern and dynamic city continuously driven by scientific and business innovation putting peoples and their creativity at its core, with its long history and tradition
- The urban cosmopolitan area keeps growing, the administrative commune covers an area of about 180km^2 with a population of 1400000 (20% of which foreign residents)
- Divided into administrative boroughs and several informal districts, connected via a system of radial streets and circular rings as well as Metro stations (underground)

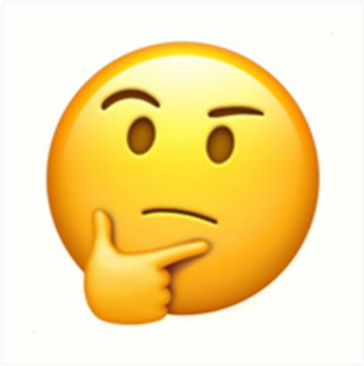
Milano: area and plan





Business Problem

- Preamble: despite the general economic stagnation, many jobs are available and attracts new residents in the cosmopolitan Milano
- Objective: taking advantage of IBM Data Science methodology and Machine Learning techniques, 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
- Audience: 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: sources

- Administrative boroughs, legacy districts and Metro stations of Milano available from [2], [3] and [4], respectively, geographical coordinates obtained via Python Geocoder
- *Informal districts are the focus, since boroughs are too vast and heterogeneous while metro stations areas are too small and do not cover the entire city surface*
- Average houses price (€/m²) in the districts available from [5] (and from [6] in the Metro stations area), venues information obtained via Foursquare API https://foursquare.com/
- *The average houses price considers all residential types without distinction and increases rapidly (although not always monotonically) toward the city center [1], it is ranked as "lower" or "upper" when respectively lower or higher than 4000€/m²*

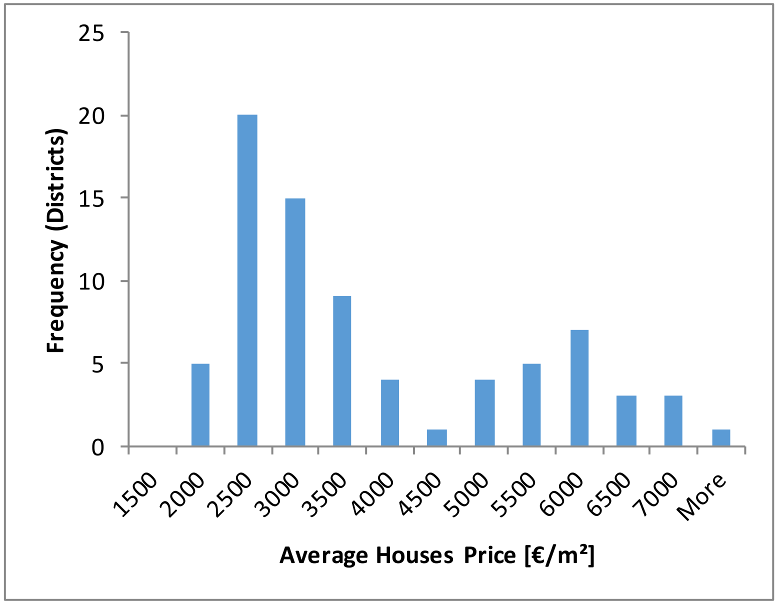
- [1] https://en.wikipedia.org/wiki/Milan
- [2] https://en.wikipedia.org/wiki/Zones_of_Milan
- [3] https://en.wikipedia.org/wiki/Category:Districts_of_Milan
- [4] https://en.wikipedia.org/wiki/Category:Milan_Metro_stations
- [5] https://www.idealista.it/news/statistiche/prezzo-linea-metro/milano
- [6] https://www.mercato-immobiliare.info/lombardia/milano/milano.html

Data: glance

- Structured data were collected and consolidated in convenient Excel files

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



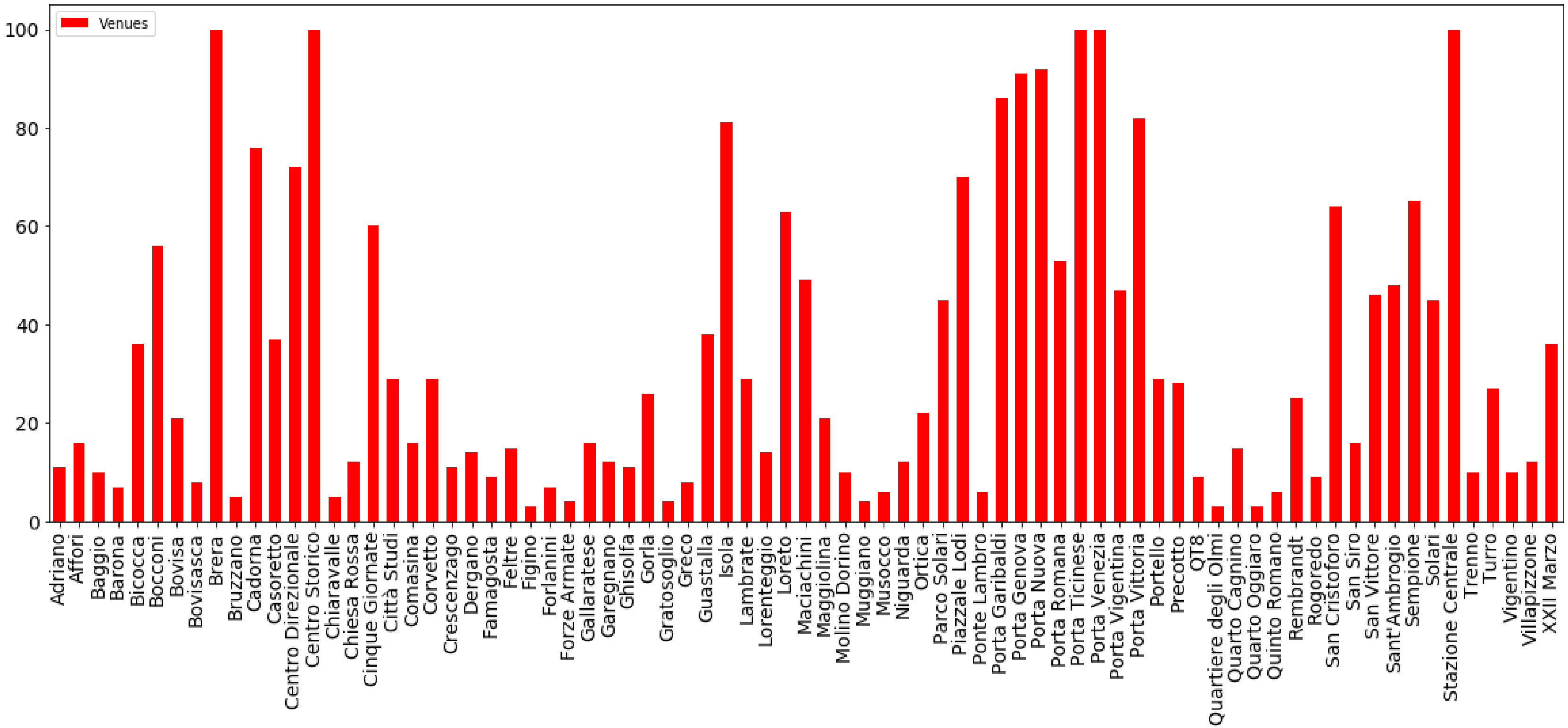
Methodology

- The following procedure was implemented in a convenient Jupyter Notebook

1. Install and import relevant Python libraries
2. Scrape Wikipedia pages for districts and Metro stations in Milano
    - *Create a (clean) dataframe for Milano's districts and metro stations*
3. Load and analyse raw web data with average houses price
    - *Derive average houses price distributions and rank the class range*
    - *Merge (clean) information with dataframe for Milano's districts and metro stations*
4. Get the geographical coordinates of Milano's districts and Metro stations
    - *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
    - *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
7. Use Foursquare API to explore Milano's districts
    - *Get a dataframe of the top 100 venues per district that are within a radius of 500m*
    - *Check how many venues were returned per district*
    - *Analyze the venues per district using one hot encoding*
    - *Get the mean frequency of occurrence of each venue category per district*
    - *Get the 5 most common venues per district*
8. Group Milano's districts into clusters per venue category
    - *Create a map of Milano with Metro stations and all districts' clusters superimposed*
    - *Show the most relevant clusters of Milano's districts and their average houses price*

Results: venues

- 2573 venues of 254 categories were found in Milano's districts

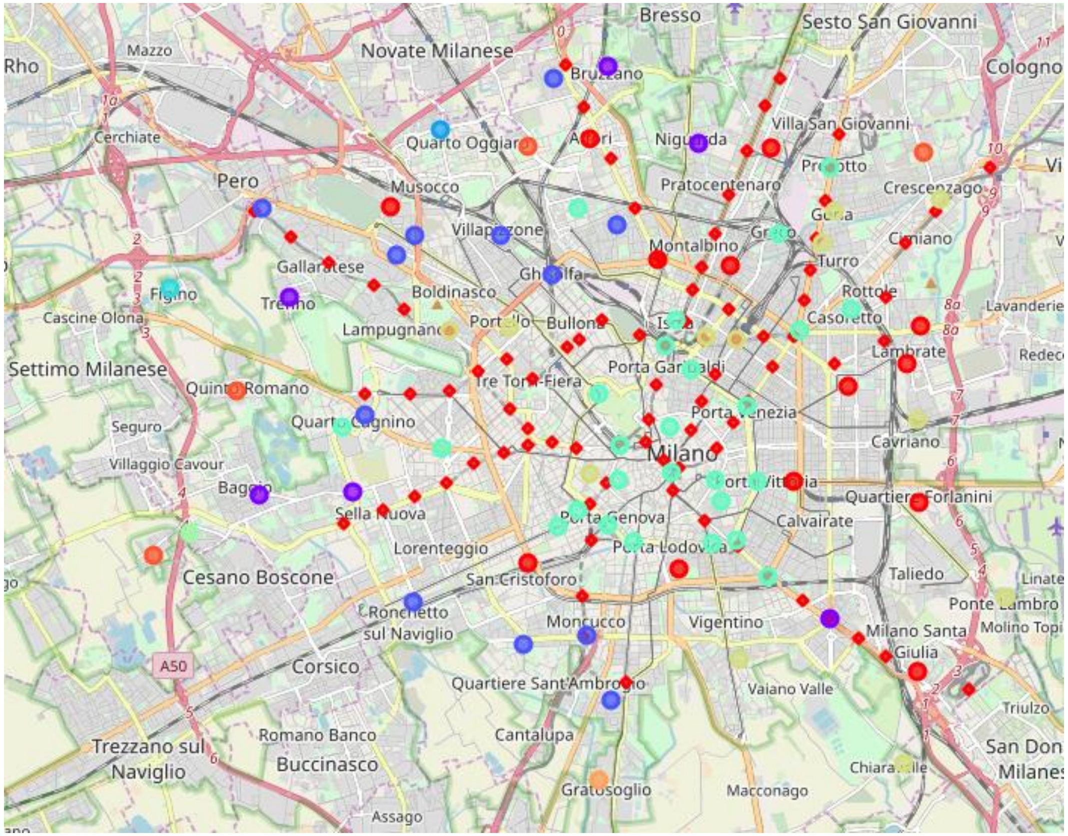


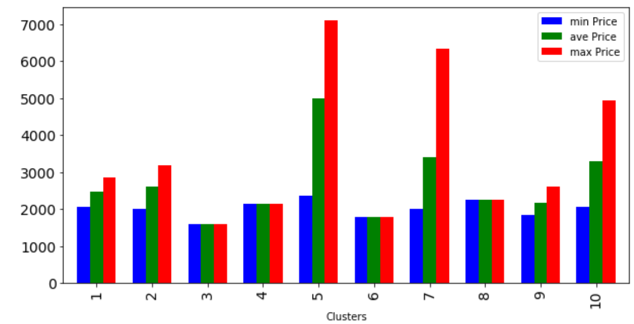
	District	Price	Range	Latitude	Longitude
0	Adriano	2600	lower	45.514684	9.245257
1	Affori	2250	lower	45.516972	9.169389
2	Baggio	2250	lower	45.460108	9.094160
3	Barona	2900	lower	45.436325	9.154339
4	Bicocca	2600	lower	45.515597	9.210627

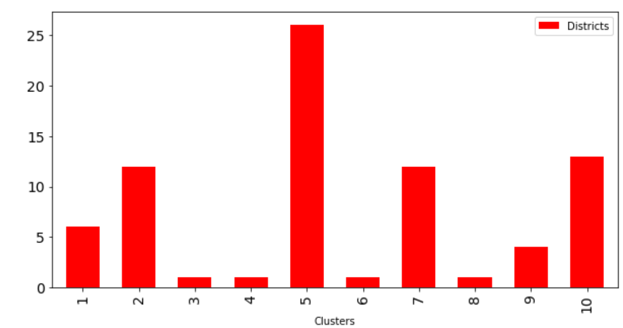
	Station	Line	Borough	Price	Range	Latitude	Longitude
0	Cairoli	1	1	7970	upper	45.468508	9.182230
1	Conciliazione	1	1	6950	upper	45.467686	9.166304
2	Cordusio	1	1	8100	upper	45.465675	9.186362
3	Palestro	1	1	7310	upper	45.471644	9.201917
4	Porta Venezia	1	1	5270	upper	45.475052	9.205523

Results: clusters

- 10 clusters of districts were identified, based on their most common venues







	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

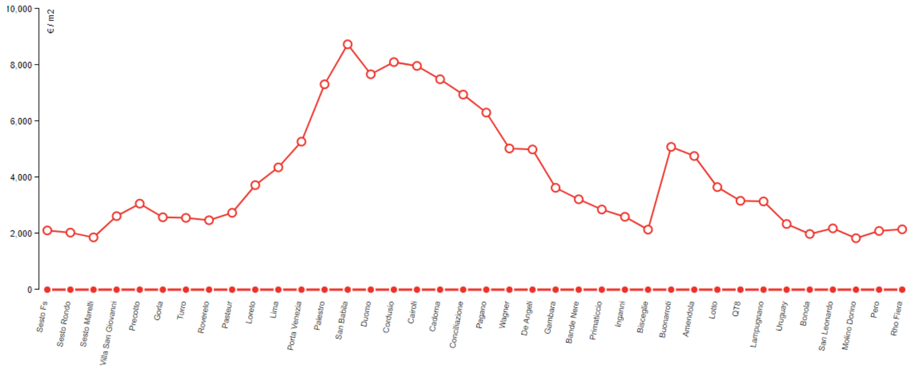
- The districts distribution granted a rather homogenous coverage of Milano’s surface
- Many districts have just few venues, especially the most isolated peripheral ones
- Clusters #3, #4, #6 and #8 contain just one peripheral district characterized by common popular venues (e.g., Metro and rail stations, parks, sport sites, markets) and average houses prices in the lower range with 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, always within the lower price range still
- Cluster #5 and #7 have similar venues (e.g., restaurants, hotels, café, shops, boutiques and small parks): 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)
- Cluster #10 includes all major university districts with student aggregation places and residences, where the variability of the average houses price depends on the campus type (i.e., public or private) and location (i.e., closer or farer to the city centre)
- 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 and maintain their own specificity (which results in a different order of their most common venues)

Conclusion

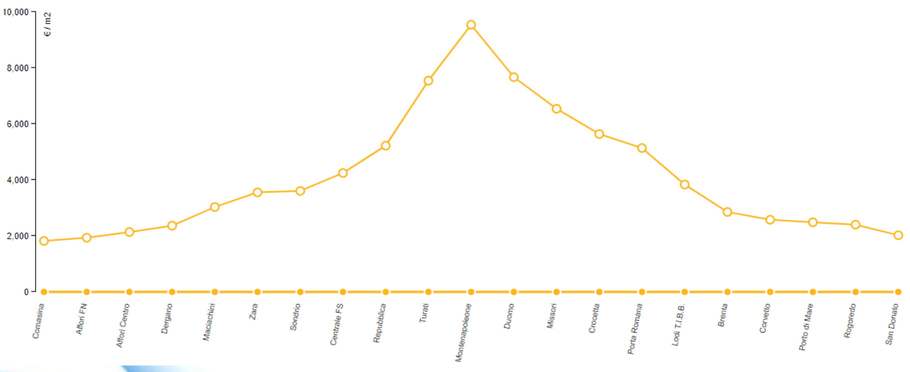
- Taking full advantage of IBM Data Science methodology and Machine Learning techniques, this project derived 10 clusters of Milano's districts based on the categories of their most common venues
- Technically, building consolidated databases of the required inputs was fundamental and took most of the effort, especially to structure and harmonize different data sources in order to answer the business problem consistently
- The unsupervised *k*-means algorithm demonstrated effective and well suited, while interactive maps, histograms and bar charts served as ideal to show data and results
- It is suggested to collect districts demographics and borders data to refine the results
- The identified clusters of districts 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 more data and additional complexity would be necessary to attempt deriving accurate quantitative models
- 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
- Finally, it is recommended to invest in the legacy lower-range areas within cluster #5

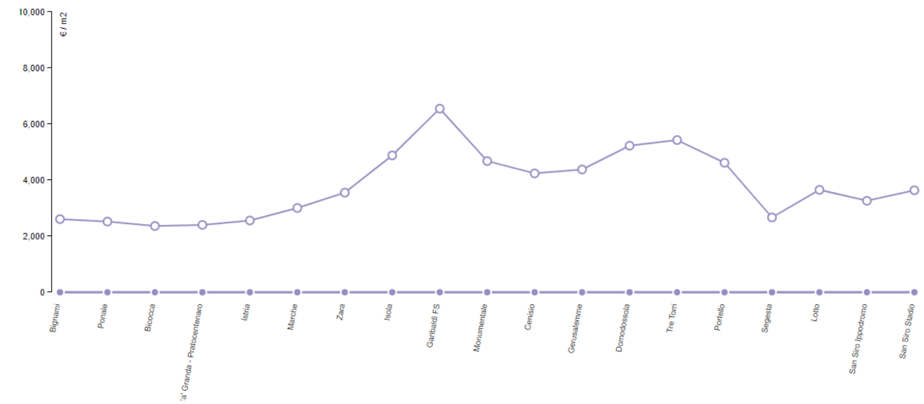
Appendix: Metro lines

M1



M2





M3

M5