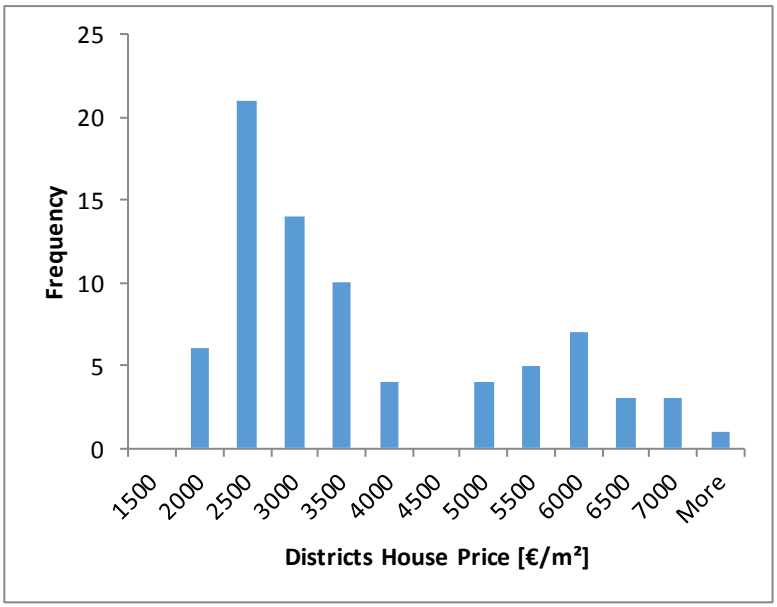
Data

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 €/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	House Price [€/m²]
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 [€/m²]*
Mean	3565
Median	2975
Mode	2500
Minimum	1650
Maximum	7050

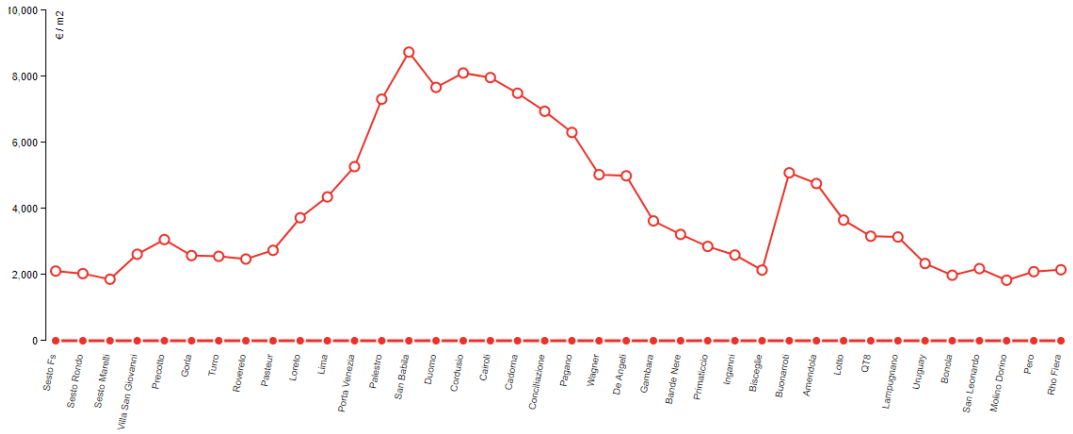


Recall that districts will be the focus, since boroughs are too large while metro areas too small. In particular, web-scraping techniques will be used whenever most convenient to extract data about the municipality of Milano from Wikipedia pages [1-4], with the help of Python [7] and its Beautiful Soup library [8]. Latitude and longitude of Milano's boroughs, districts and metro stations will be obtained via Geocoder library [9] and allow relating areas and locations based on their geographical coordinates as well as associating the pertinent average houses price loaded from a .csv file of preprocessed web sources [5-6], so 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

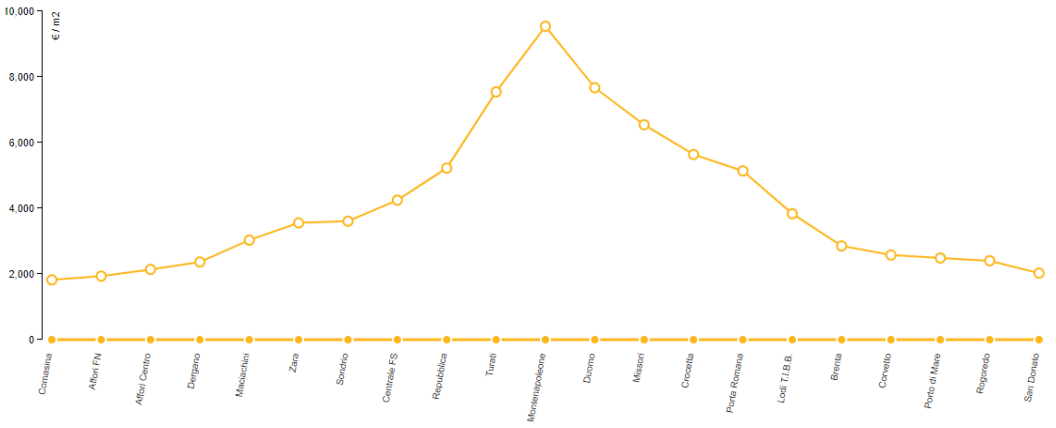
venues in each district along with their category. Scikit-learn library [11] will then be employed to cluster the districts based on their venues frequency (not the average houses price, since we want to find the minimum one among similar areas) and Folium library [12] will be adopted to visualize the 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 to support the decision making process about relocating or investing in a district of Milano.

Appendix: Average Houses Price along Metro Lines [5]

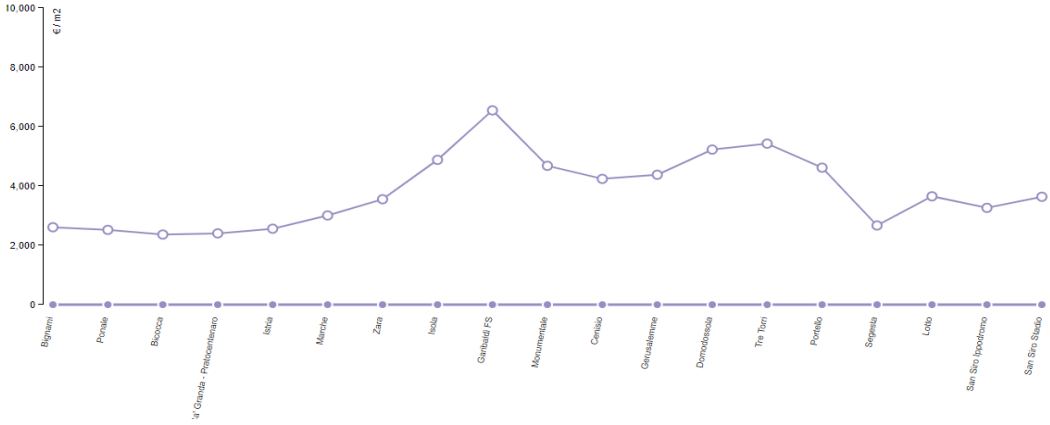
M1



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/
[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