The Battle of Neighborhoods

Real estate in the Italian economic hub, Milan

Alessandro Ruzzone

June 2020

The Battle of Neighborhoods - Applied Data Science Capstone Project

Real estate in the Italian economic hub, Milan

1. Introduction

1.1 Background

Milan has historically been one of the most international cities in Italy, as well as its financial and economic center. Its fashion and design businesses are famous worldwide and have greatly contributed to the success of the city, known as one of the fashion capitals of the world.

The real estate market saw a sharp decline both in number of sales and house prices, following the Global 2008 financial crisis.

Fig. 1 House prices and number of sales for the period 2008-2017. Source Tecnocasa Group

Starting from 2017 there has been a rebound in both prices and number of sales, however the recovery has been slower than that of other main European cities.

House Price Change						
2013	2014	2015	2016	2017	2018	2019
-6,9%	-5,5%	-1,7%	-0,1%	+3,5%	+8,4%	+13%

Fig. 2 House price changes for the period 2013-2019. Source Tecnocasa Group

The last three years have been extremely positive with an average increase of over 8% each year, however the prices are still around 20% below the pre-crisis levels, making the city extremely interesting to those who are looking to invest in real estate.

The growth is attributed by many to the World Expo which took place in Milan in 2015, counting over 20 million visitors and with over 140 participating countries. This event put Milan under the international spotlight and the demand from both national and international buyers has grown since.

1.2 Problem

The goal of this research is to classify the different areas of the city by category of venues and find any correlation with the real estate market. The research evaluates how the presence of subway stations contributes to the diffusion of these venues and to the house prices.

Ultimately, the report should be able to guide a potential house buyer regarding:

- what parts of the city have affordable houses compatibly with their budget
- what areas have a high density of commercial or residential spaces
- what areas are best served by public transport

2. Data

2.1 Data sources

• The data regarding real estate market prices will be scraped from the website of one of the largest real estate advertising websites (https://www.borsinoimmobiliare.it). Please notice that the scraped page contains already the average sale price per square meter for each area. Scraping all individual listings would have brough more detailed results but it is unfortunately not allowed by the T&Cs of the website (a non-free API service is however offered by most websites of this type). This data is useful to understand which areas are affordable for a potential buyer.

• The coordinates of each subway station are provided on the official website of the City Council (https://dati.comune.milano.it/dataset/ds535_atm-fermate-linee-metropolitane/resource/dd6a770a-b321-44f0-b58c-9725d84409bb).
This data is useful to understand what areas are best served by public transportation, by plotting the location of each station on a map of the city.

• The information regarding what venues are present in each area will be collected thanks to Foursquare.
This data can be used to form clusters by types of venues. This should help the potential buyer to understand what areas most suit his interests (i.e. a buyer with small children might prefer to move to a residential area, instead a young couple might prefer one with many bars and restaurants).

2.2 Data cleaning

The data containing house prices was obtained by scraping the website of one of the most used platforms to list house sales in Italy. The T&C did not allow scraping the individual listings therefore only a table containing the average sale price expressed in €/sq meter for each area was scraped. The data was cleaned and a row was formed for each of the listed areas because the website grouped them by price.

The geographic coordinates of each area have been geocoded by using geolocator. Following the first attempt to geocode the areas it became evident that the names of three areas had been misspelled.

After manually correcting them all full addresses but two could be found. These rows were dropped as geolocator assigned to these locations the general coordinates of Milan. The resulting dataframe had 164 rows. For each row it was possible to calculate the distance from the city center (Piazza Duomo) thanks to pyproj (cartographic projections and coordinate transformations library). Only locations within 5000 meters from the city center were taken into consideration. As a result 60 rows were dropped.

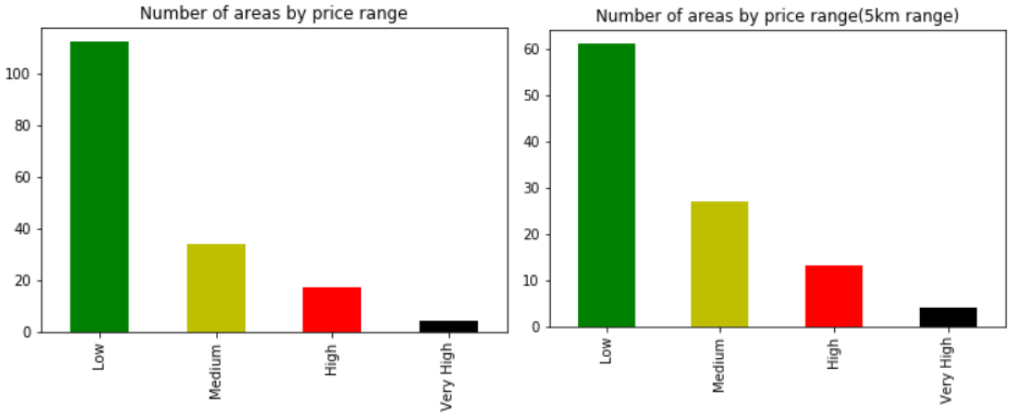
The location of subway stops and subway lines were contained in GeoJSON files published by the city of Milan itself and no data cleaning was required.

The data regarding venues was obtained by using Foursquare's API function. Due to the limitations of its free service the research was limited to 500 meters and a limit of 50 venues for each location. The resulting dataframe contained 3920 rows and 253 unique categories.

3. Methodology – Exploratory Data Analysis

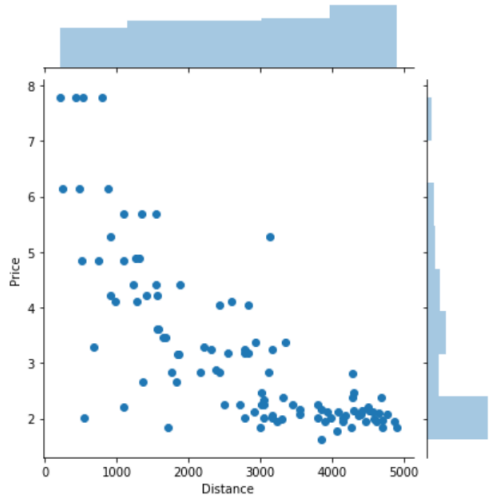
3.1 Price ranges

The prices ranged from 1589 €/sqm to 7781 €/sqm. In order to visualize the distribution of prices four bins were formed. The resulting plots show that the vast majority of the areas have low average prices however, keeping into consideration only neighborhoods within a 5km range from the city center (right plot) the trend is less marked.

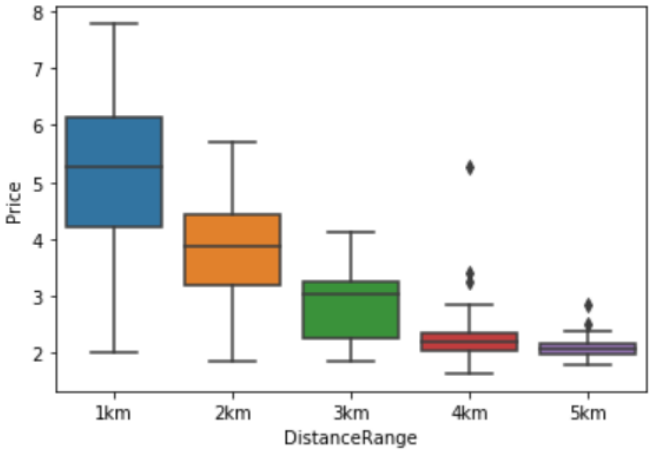


3.2 Distance from the city center

The distance from the city center was plotted against the price to visualize how the most central areas are also the most expensive. In order to do so 5 bins were formed, corresponding to an increase of 1km from the center for each.

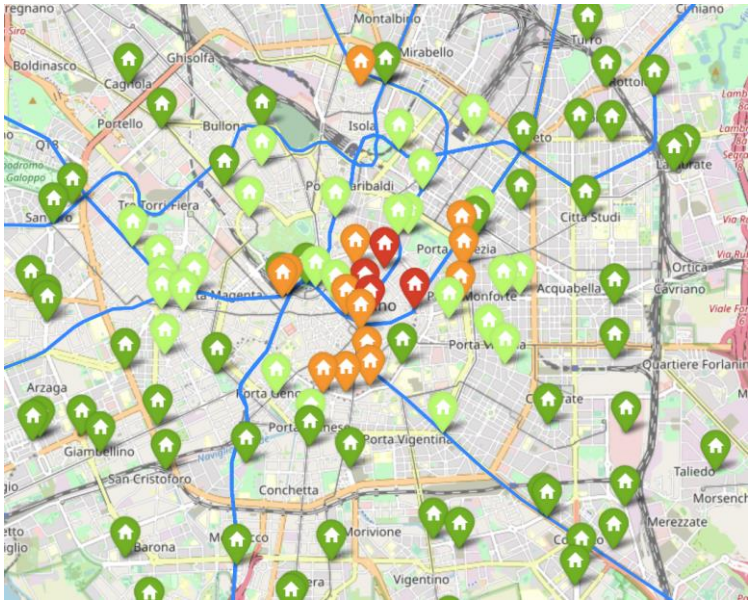


The following boxplot clearly shows a rather large difference in the average price of properties located within 1000 meters and the others clearly indicating buyers are willing to pay a premium to live in very central spots. A less wide difference is seen among other categories while almost no difference is present between the price of those located in the 4-5km and over 5km range.



3.3 Mapping

In order to produce a map which shows location and average price for each area we will use the coordinates found through Geocoder. In order to show price differences 4 bins will be created, associating a colour to each of them. Green colour will mark low prices, lighgreen will mark medium prices, orange will mark high prices and red very high prices. The map below shows that several lines (blue lines on the map) pass through the city center and the most expensive. Several areas located in proximity of a line belong to the cheapest category suggesting therefore that proximity to a subway line might not have a big influence on house prices. All maps were plotted by using folium.



3.4 The fashion district

The dataframe created with data downloaded from Foursquare was filtered by searching venues identifies as boutiques. The names of many of the most famous fashion brands were returned.

Venue	Venue Latitude	Venue Longitude	Category
Armani	45.470443	9.192732	Boutique
Ermenegildo Zegna Boutique	45.469819	9.192989	Boutique
Kenzo	45.469763	9.192539	Boutique
Prada	45.465600	9.189968	Boutique
Louis Vuitton	45.465224	9.191796	Boutique

These venues (black icons) were added to the map previously created to verify weather the most expensive areas of the city (red icons) are near the notorious fashion square district. The map below confirms that the areas with the highest average house prices are also those with the highest number of fashion boutiques for which the city is so famous.

3.5 k-means Clustering

The Foursquare API was used to explore neighborhoods in Milan. The explore function was used to get the most common venue categories in each neighborhood. The *k*-means clustering algorithm with k=3 was used to group the neighborhoods into clusters.

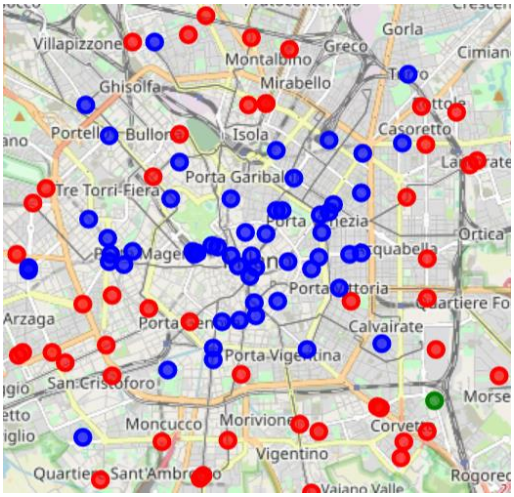
By analyzing each cluster in detail it became evident that :

Cluster 1 is characterized by a high density of bars, restaurants, hotels and venues such as museums and theatres. This cluster includes the boutiques of the famous fashion district and includes the most expensive area of the city.

Cluster 2 is a more residential area, yet shows large presence of italian food restaurants, which is quite normal in Italy.

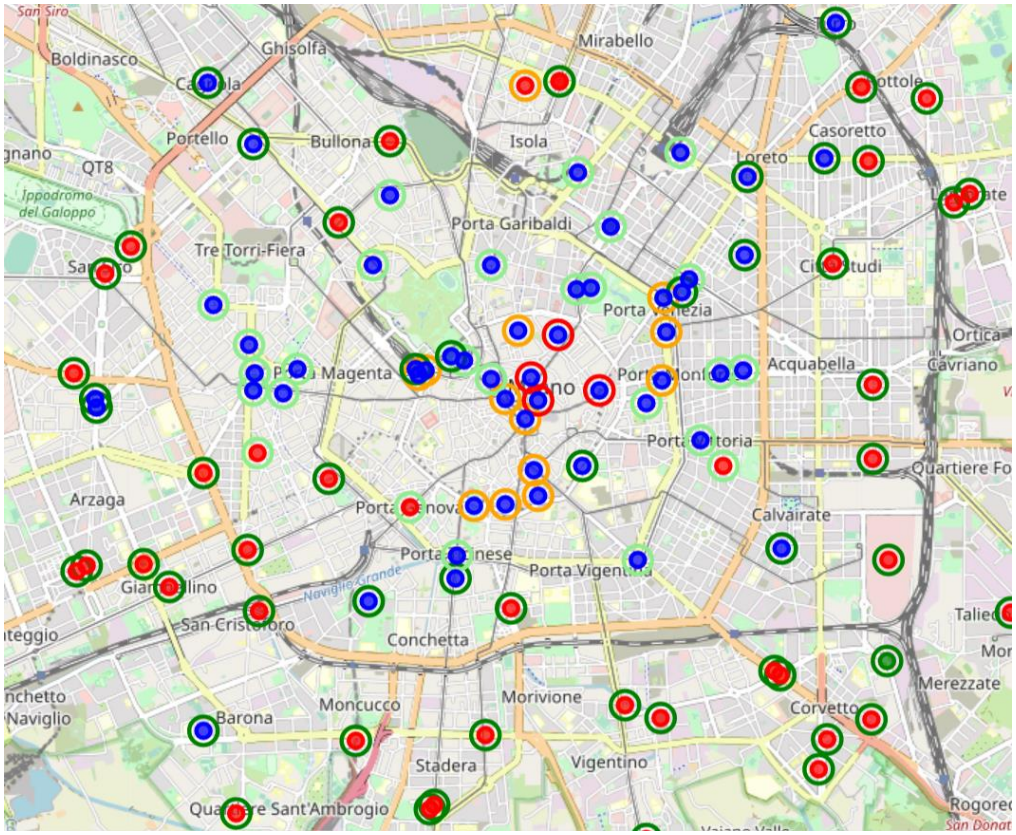
Cluster 3 includes more periferal areas, characterized by markets and shops less likely to be found in large numbers in the city center. It is most likely a periferal residential area.

The map below confirms that the core of the commercial activities is in the center of the city, with more residential areas all around it.



4. Results

In this study I analyzed the factors which most commonly influence the selling price of a house. Location and availability of services are like in most cities the most important variables. In the case of Milan, known as one of the fashion capitals of the world, proximity to the fashion district puts the biggest premium on house prices. This can be explained by the fact that most wealthy foreign buyers or tourists in search of luxury rental accommodation are mostly attracted by this famous district and less by others. In the map below clusters and house prices(circle marker) are reported.



5. Discussion

In order to produce more accurate results, it would be advisable to collect house sale prices directly from individual listings, if allowed (several websites offer non-free API services). Some degree of limitation is also posed by Foursquare's API which in its free version only offers a limited amount of

calls. Furthermore, Foursquare's database seems to be largely focused on restaurants, bars and entertainment sites. For a more comprehensive analysis it would be beneficial to add different data sources.

6. Conclusion

This study should help the target audience to easily identify the neighborhoods of the city which offer housing solutions compatible with their budget. As in most Italian cities which have a center rich of historic monuments the price decreases with distance from it therefore a buyer on a low budget should probably look to find a house located 3 to 5 kilometers away from the city center but near a subway line.

For glamorous buyers with deep pockets the choice is obvious. The fashion district occupies a small area compared to the size of the city and the strong demand has made the surrounding areas the only ones with an average price of over 7000€/sqm.