**Let’s find a new house for Novae Spes Company**

Alberto Bortolotti

August 02, 2021

**1. Introduction**

**1.1 Context**The "Novae Spes Company", an emerging society working with renewables and big data, has grown quickly in employees number in the last years. Actually the society is located in a small building in the suburbs of Milan, which is way too small and it does not give the image of an uprising company.  
The company's board finally decided to move to a new house and they commit us to find a proper location.

**1.2 Problem**In order to find the best location for the new office some key points has been found out in preliminary meetings with board. In particolar they want a location that is easy to reach for clients coming from other cities and which is enjoyable for workers, either in terms of public transport and attractions close to the new office. Obviously they want to pick the best option from an economical point of view but this aspect must not prevail with respects to the others. In shorts, hereby we summarize the points we must take into account into our analysis:

1- Presence of Train station nearby the new office  
2- Presence of Metro or bus station nearby the new office  
3- Meter square office cost (€/M2)  
4- Level of pub/resteraunts around the office  
5- Presence of parks around the office  
6- Distance from city center

**1.3 Approach**Our approach in the solution of this task will be to compute a peculiar score for each of the keypoints defined in the previous paragraph. Then we are going to compute a “Suitability score” as the average of them and we will sort the borough according to it.

**2 Data analysis and managing**

**2.1 Sources**In order to exploit the results requested we are going to use two different database, from which we will extract all the information we need.

**A- List of adresses in the city of Milan**

<https://dati.comune.milano.it/dataset/ds634-numeri-civici-coordinate/resource/533b4e63-3d78-4bb5-aeb4-6c5f648f7f21>

From this database we will extract the list of neighborhoods and their geo-location features (latitude and longitude). This dataset contains (among many other informations) all the addresses in the cities, their latitude and longitude and their neighborhood name. Starting from this raw dataset we are going to create the neighborhoods dataset, grouping the adressess according to the corresponding neighborhood. We are going to compute the neighborhood latitude and longitude by making the average of all the addresses latitude and longitude within that specific neighborhood. This means that we will not have the exact center geographically speaking but it will be the most accurate according to building density in the area.

You can fin the .csv version in our github repository, following this link:

https://github.com/ReTurbolento/Coursera\_Capstone/blob/3697ab1fc4bb8a91beb2b1b6b220cd885de5ef40/ds634\_civici\_coordinategeografiche\_20210701\_final.csv

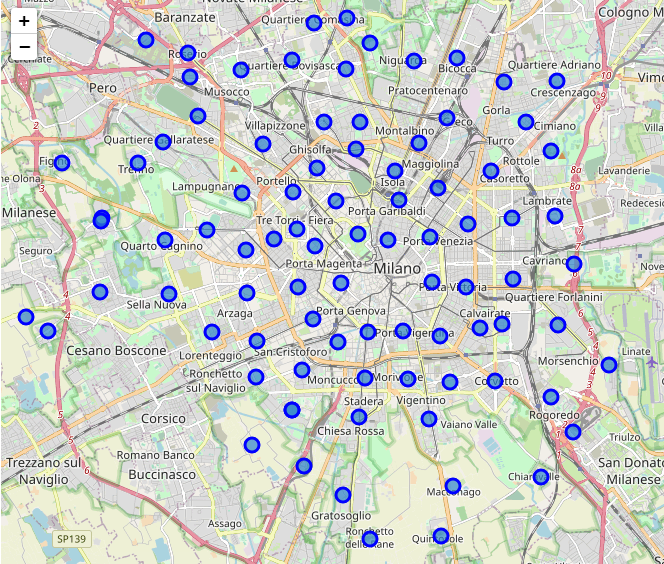
**B- House price in Milan**

<https://www.milanotoday.it/economia/prezzi-case-2019.html>

In this case we have a table giving the meter square pricing for building in Milan. Actually different prices are contained referring to different houses cluster, we will compute an average between new house in the mid range economically (Column "Medio nuovo") and already built house in mid range economically (column "Medio usato") Since the name of neighborhoods are slightly different in the two databases we have add a column to the original .csv reporting the neighborhood name according to the first database. In our github repository you can find the pre-processed file, at this link:

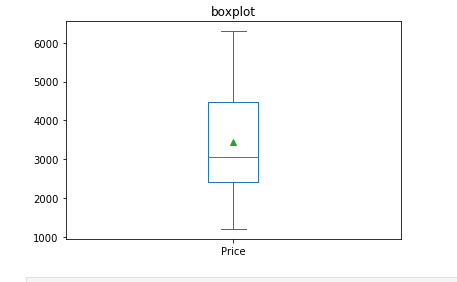
<https://github.com/ReTurbolento/Coursera_Capstone/blob/f8eb926fda88f643c841aa71f3406e991b5874cb/Milan_House_Prices.csv>

**2.2 Borough definition and dataset analysis**As described before we are going to define borough according to the ufficial addresses of the Milan municipality. This is the final results: in the following map a circle is printed in the center of each borough:



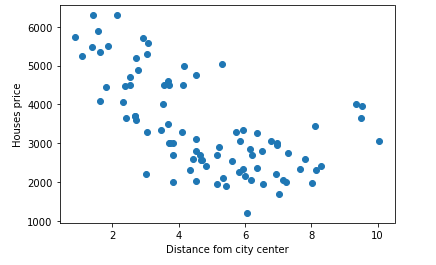
We will use the borough feature to query Foursquare engine and extract information about each of them.

From the second database instead we can extract the price of houses and office in each borough. Let’s see their distribution:



As you can see median and average value are around 3000€/M2 and most of the borough price are between 2500 and 4500 €/M2. Interesting: no outliers has been found out.

Another interesting feature to check is if there is a relation between price and distance from the center. As you can see in the following figure there is a general decreasing relation between price and distance from the center but the deviation from the trend line is quite strong for some borough.

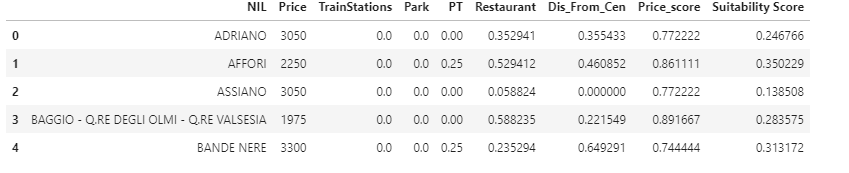


**3 Model definition**

**3.1 Dataframe descritpion**Startingfrom original database we used Pandas library in order to get to a final dataframe which will have all the information needed to query Foursquare. This is the structure of this merged dataframe:



All the other information are going to be extracted by foursquare and group by categories for each borough. This is instead the final form of the dataframe after the integration of foursquare data:



In the following paragraph we are going to define how each score has been defined.

**3.2 Scores computation**

In previous paragraphs we build different dataframes containing various information for each borough. In particolar we are going to compute for each borough a specific score, given by each of the main feature requested by the company board about the neighborhood. More specifically:

1- Presence of Train station nearby the new office   
2- Presence of Metro or bus station nearby the new office   
3- Meter square office cost (€/M2)   
4- Number of pub/resteraunts around the office   
5- Presence of parks around the office

For each of the above points exluded the third we are going to compute a normalized score by dividing the number of train station, metro station, pub/resteraunts and park with the number of this venues found in the best borough. For example if borough 'Affori' has 3 parks and the borough with most park has 5 parks, the parks score will be 3/5= 0,6. Hereby the math equation used:

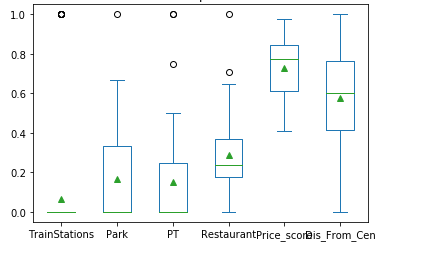
For price score instead we want to penalyze the most expensive borough but we want to keep the score between 0 and 1 (or 0% and 100%). For this reason the normalisation equation will be the difference between 10000 €/M2 and the actual cost for the borough divided between the difference between 10000 €/M2 and 1000 €/M2, which is a default value chosen considering price of houses/offices outside the city. Speaking mathematically:

Similarly the distance from city center should have the best values close to the city center and the worst values for peripherical borough, thus:

Once all this data will be computed we will make an average and express the final score as a percentage. Each borough will thus be X percent suitable for finding the new office. If ideally one borough would have the best score in each of the 5 categories it would be 100% suitable for finding the company new office.

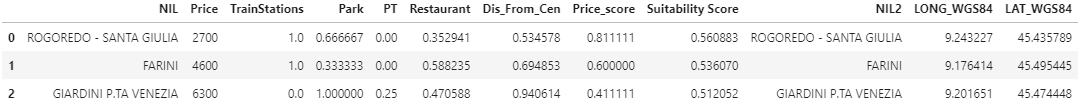
**3 Results**

**3.1 Some consideration about scores distribution**



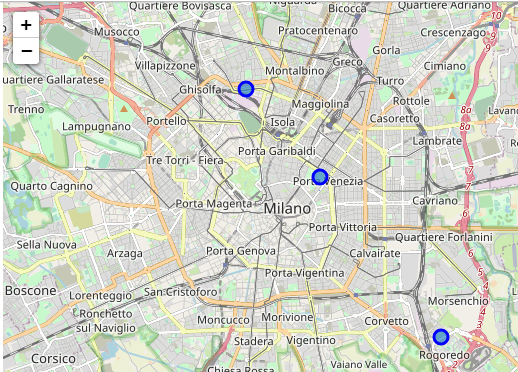
Hereby some consideration about insights that can be derivede from the above boxplots:  
1- Trainstation distribution is actually a binary variable: some of the borough has a station and most not. This means that we only have scores equal to 1 or equal to 0: this is actually good for our model since the train station is a strong requirement from company's board so it's correct to strongly penalyse borough without a train station.   
2- The distribution about public transport is quite poor and this is not so coherent with public transport distribution in the city. This is probably due to foursquare venue search which penalyse transport station with respect to other venues. Actually we can't do much about it, we will eventually consider this in comparing borough with similar suitability score.   
3- Price score is bias towards high values. This is actually normal since we consider an high value as maximum value for normaliisation, since the cost point was not the most important for the choice.

**3.2 Winner borough**In the following table you can find the individual scores and suitability scores for the three most suitable borough within the city:



As you can see in the following map, the three borough are very different between each other. Let's analyze them quickly:

1- ROGOREDO - SANTA GIULIA has the best score: it's very easy to reach thanks to its train station and the cost is very low. It has also quite a few parks but restaurant is very low and is very far from the city center.   
2 - FARINI is an intermediate situation with all the scores around 50%, its best quality is a good price score for a borough not so far from city center.   
3 - GIARDINI P.TA VENEZIA is a very central borough with great attraction but it has no train station and price is very high. For this reason the suistanibility score is quite good but well above the score of the previous two borough.



**4 Conclusions**

Comparing the different Milan’s borough according to keypoints requested by the company’s board we have found out the three most interesting borough for locate the new offices of the company. The result are quite consistent and we are confident enough that some good accomodations can be found in these boroughs.

Nevertheless, some more steps must be followed before taking the final decision. In fact the main quality of our model is that it can be replicated to different dataset, as long as geographical information about different localities and their price are provided.

For this reason our suggestion is to use the insights of this analysis to chose some possible location within these borough. Once we have a set of option we could run again the model, so that suitability score is computed for each location. This is important because borough can be very wide so scores within the same borough could also change.

Another strong suggestion is not to take into account the only suitability score as key for the choice: in our opinion the best solution is the one with good scores for each keypoint instead a solution with very high scores in 3 or 4 keypoints but nihil scores for the others.