IBM Data Science Professional Certificate Capstone Project

Neighborhood Analysis for Real Estate Investing in Rome

Katarzyna Anna Parys

June 2, 2021

Sommario

1. Introduction

1.1 Background

Rome is Italy's capital and largest city, with 2,822,981 residents, as on December 31, 2020[1]. It's also the third most populous city in Europe, after Berlin and Madrid, if the areas within city administrative boundaries are considered. The area of the Municipality of Rome is 1,285 km^2.

Rome is topping also the list of the most popular European tourist cities. In fact, according to the data compiled by the Italian National Institute of Statistics (Istat), Rome reached 24.5 million of overnight stays in 2019. These numbers dropped drastically during the Covid-19 pandemic, but they are bound to increase as soon as the emergency is over.

On the other hand, the Rome's housing market remains steady, despite the pandemic-induced recession. During the year to February 2021, homes prices stood at €2,848 (US$3,404) per sq. m., on average, in February 2021, up 0.6% from a year earlier (unchanged when adjusted for inflation)[2].

The prices of the properties for sale may vary considerably across the city's neighborhoods. As of June 2021, house prices in Piazza di Spagna can fetch 9,100 euros per square meter, while in the Castelverde suburbs the average price for a residential property is around 1,600 euros per square meter. However, the distance from the city center isn't the only factor influencing property prices.

There are 15 districts in Rome (in Italian "municipio") divided into urban zones or neighborhoods, (in Italian "zona urbanistica"), which are 155 in total.

Figure 1 - Rome's City Centre, where apartments are most expensive

[1] Data from the Italian National Institute of Statistics, http://dati.istat.it/Index.aspx?lang=en&SubSessionId=681b71ca-412f-4837-acbf-6d7323786e5f

[2] https://www.idealista.it/en/press-room/property-price-reports/sale/report/

1.2 Business Problem

The project aims at understanding the main factors determining house prices in Rome. The characteristics affecting property prices might include the distance from the city center, the density of population, the presence of different kinds of venues, like such as number of shops, restaurants, professional buildings, museums, universities, cafes, hotels, gyms and more, the traffic, the percentage of the foreign-born population. Such data gathered from Foursquare and other open sources will be used to cluster the 154 neighborhoods of Rome and understand the relationship between the price and the location of a property.

1.3 Interest

Knowing which characteristics of a neighborhood affect property prices in a given area could be useful for all participants in real estate markets, especially investors, to understand and predict property price trends. In particular, the Department of the Treasury of the Ministry of Economy and Finance, which manages the public sector real estate in Italy, could benefit from the project for a better property valuation and more efficient management.

2. Data

The table below reports the type of data needed and their sources.

Data	Source	Last updated
List of prices per square meter to buy apartament in neighborhoods of Rome, Italy	https://www.immobiliare.it/mercato-immobiliare/lazio/roma/	Daily updates
List of the neighborhoods (urban zones)	https://www.comune.roma.it/web-resources/cms/documents/Elenco_Z_Urbanistiche_rg_A.pdf	2019
Area of the urban zones	https://www.comune.roma.it/web/it/roma-statistica-territorio.page	31.12.2019
Population of the urban zones	https://www.comune.roma.it/web/it/roma-statistica-popolazione1.page	31.12.2020
Foreign population of the urban zones	https://www.comune.roma.it/web/it/roma-statistica-popolazione1.page	31.12.2020
Top 100 venues in a 1000 meters range of the centre of each urban zone, categorized by high-level groups	FourSquare API	Daily updates
Pollution levels (as a rough indicator of traffic)	http://www.arpalazio.net/main/aria/sci/basedati/bollettini/2021/BA192021.pdf	16.05.2021
List of the air monitoring stations in Rome with coordinates	http://dati.lazio.it/catalog/it/dataset/rete-di-monitoraggio-della-qualita-dell-aria/resource/0c9d32b8-06ed-4bb0-8727-9954c6d703f2	2020

3. Methodology

3.1 Data loading

First the price data for Rome has been scraped from the website https://www.immobiliare.it/mercato-immobiliare/lazio/roma/. The Python library BeautifulSoup has been used to parse the text and retrieve the table containing purchase and rental data for 50 areas defined by the No 1 real estate portal in Italy. The dataframe has been then cleaned and saved as Excel file.

```
In [2]: #loading real estate data from the web page

        # source url
        url = "https://www.immobiliare.it/mercato-immobiliare/lazio/roma/"

        # performing the request
        file = requests.get(url).text
```

Parsing the text with BeautifulSoup to retrieve the table

```
In [3]: # parsing data with Beautiful Soup
        parsable_file = BS(file, 'lxml')

        # retrieving the table
        data_table_list = parsable_file.find_all('table')
        data_table = data_table_list
```

Converting the table into a list

```
In [4]: # converting the table into a list
        list = pd.read_html(str(data_table), header=0)
        list
```

Figure 2 - Use of BeautifulSoup to web scrape data

Then a list of 155 urban zones in Rome was obtained from the City of Rome official website, together with the information on each zone's area, population in 2019 and 2020 and the number of immigrants living in each zone. The data has been inserted in Excel's UrbanZonesRome.xlsx, which calculated also the density of population of each zone (Population/Area) and the percentage of foreigners among population (Foreigners/Population*100). For each zone an address has been manually added, indicating the center of each neighborhood. Finally, the table has been then loaded in the Jupyter Notebook.

Later the csv file with air monitoring stations and their coordinates has been loaded in a pandas dataframe. Since it contained the data on all the Province of Rome, the rows non regarding the Municipality of Rome have been dropped and the column "Province" has been removed, to obtain data on the 12 air monitoring stations within Rome.

Another dataframe has been created with the air pollution data for each monitoring station: the average level of the PM10 (Particulated Matter) in the week 10-16.05.2021, preceding the data collection, and

the number of days between January 1, 2021 and May 16, 2021 when the measurements exceeded the level of 50 µg/m3 (overcomings). This dataframe has been merged with the dataframe "stations".

3.2 Finding coordinates and calculating the distance from the city center

In order to use Foursquare, display maps and calculate distances we need the geographical coordinates (latitude and longitude) of the centers of the urban zones.

Hence the location data has been captured from the Nominatim geolocation service provider, using the address of the center of each zone, and added to the data table.

The haversine formula has then been used to calculate the distances from the zone centers to the city center (the location of the Centro Storico neighborhood).

```
# retrieving longitude and latitude of the center of each zone

# instantiating a geolocator
geolocator = Nominatim(user_agent="stat_explorer")
city = 'Roma'
count = 0
# retrieving data for each station
columns_list = []

for address in df_zones['Address']:
    complete_address = address + ',' + city
    # passing the location to the geolocator
    location = geolocator.geocode(complete_address)
    if location is None:
       location = geolocator.geocode(str(city))
       print('city '+city)
    else: print('complete_address '+complete_address)
    # retrieving latitude and longitude from the geolocator


    print(location)
    print(count)
    count = count + 1
    latitude = location.latitude
    longitude = location.longitude
    display_name = location.address
    columns_list.append([latitude, longitude, complete_address])

columns_list
```

Figure 3 - Using Nominatim to retrieve latitude and longitude from address

3.3 Matching zones and pollution data

Having Rome 155 zones and only 12 air monitoring stations, it becomes necessary to match each zone with the nearest monitoring station. The match is done by calculating the geodesic distances between each zone and each station and assigning the pollution data of the nearest station to each zone.

3.4 Adding FourSquare data

FourSquare API is used to retrieve up to 100 top venues in the range of 1000 m from the center of each meighborhood for 10 high-level categories: 'Arts&Entertainment', 'College&University', 'Event', 'Food', 'Nightlife Spot', 'Outdoors&Recreation', 'Professional&Other Places', 'Residence', 'Shop&Service', 'Travel&Transport'. The counts for each zone are stored in the final dataframe.

3.5 Exploratory data analysis

Our dataframe contains now the following data for each of the 155 urban zones in Rome: zone code, zone name, area name, district, population on December 31, 2019, population on December 31, 2020, zone area, population density, number of foreign-born residents, percentage of foreign-born residents, zone address, zone latitude, zone longitude, distance from the city center, PM10 levels, number of days with PM10 threshold level exceeded and the numbers of zone's venues (up to 100) for the ten venue categories.

Some basic statistical details are displayed for each variable, through the pandas describe() method and the boxplots are plotted to visualize some of the numerical variables. The variables 'Zone', 'Name', 'Zona Immobiliare', 'District' and and 'Address' are dropped as non-numerical, while some others, like 'Population 31.12.2019', 'Population 31.12.2020', 'Density', 'Foreigners 31.12.2020', 'latitude', 'longitude' and 'Area', are not plotted as having the spread much larger than the others.

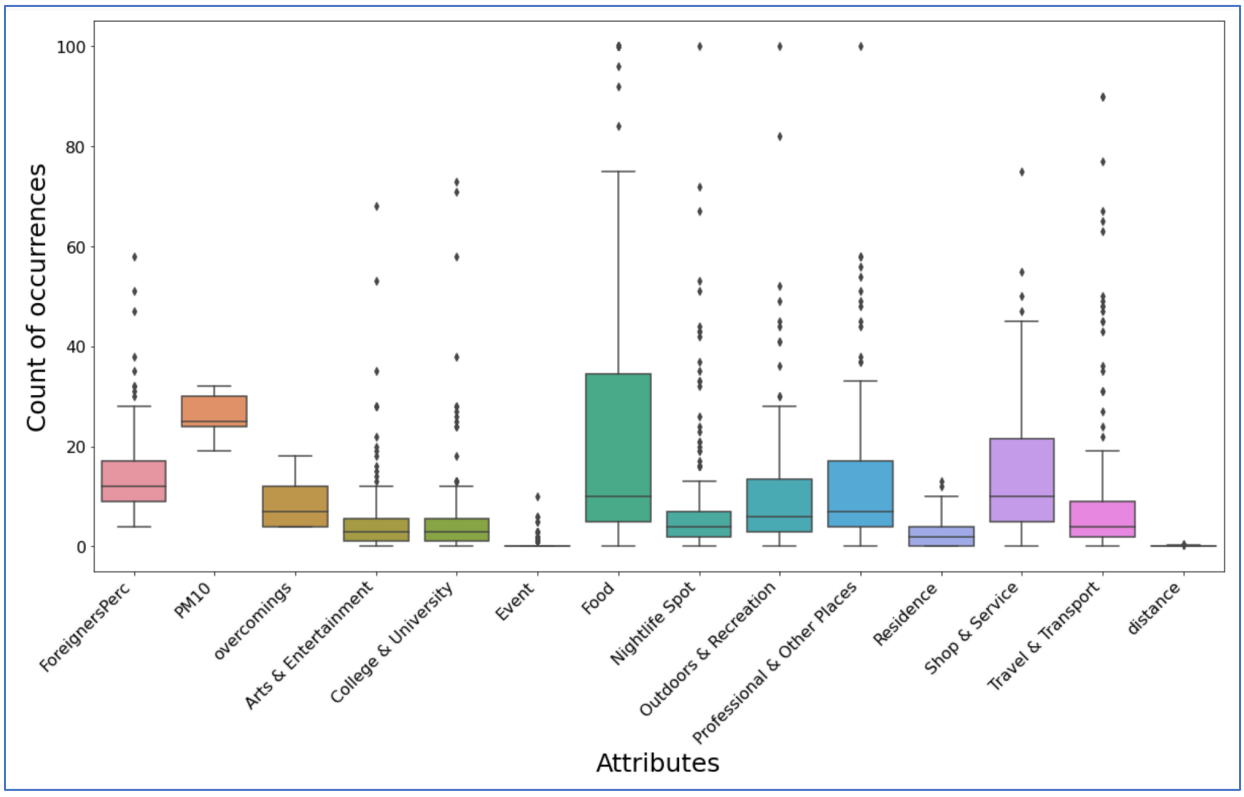

Figure 3 - Boxplots of some explanatory variables before normalization

The most represented venues are Food, Shops and Professional and all venues present significant upper outliers. PM10 and overcomings are more evenly distributed.

3.6 Data normalization

The data has been normalized using MinMaxScaler. Each feature has been scaled to the range 0 - 1, making the values consistent.

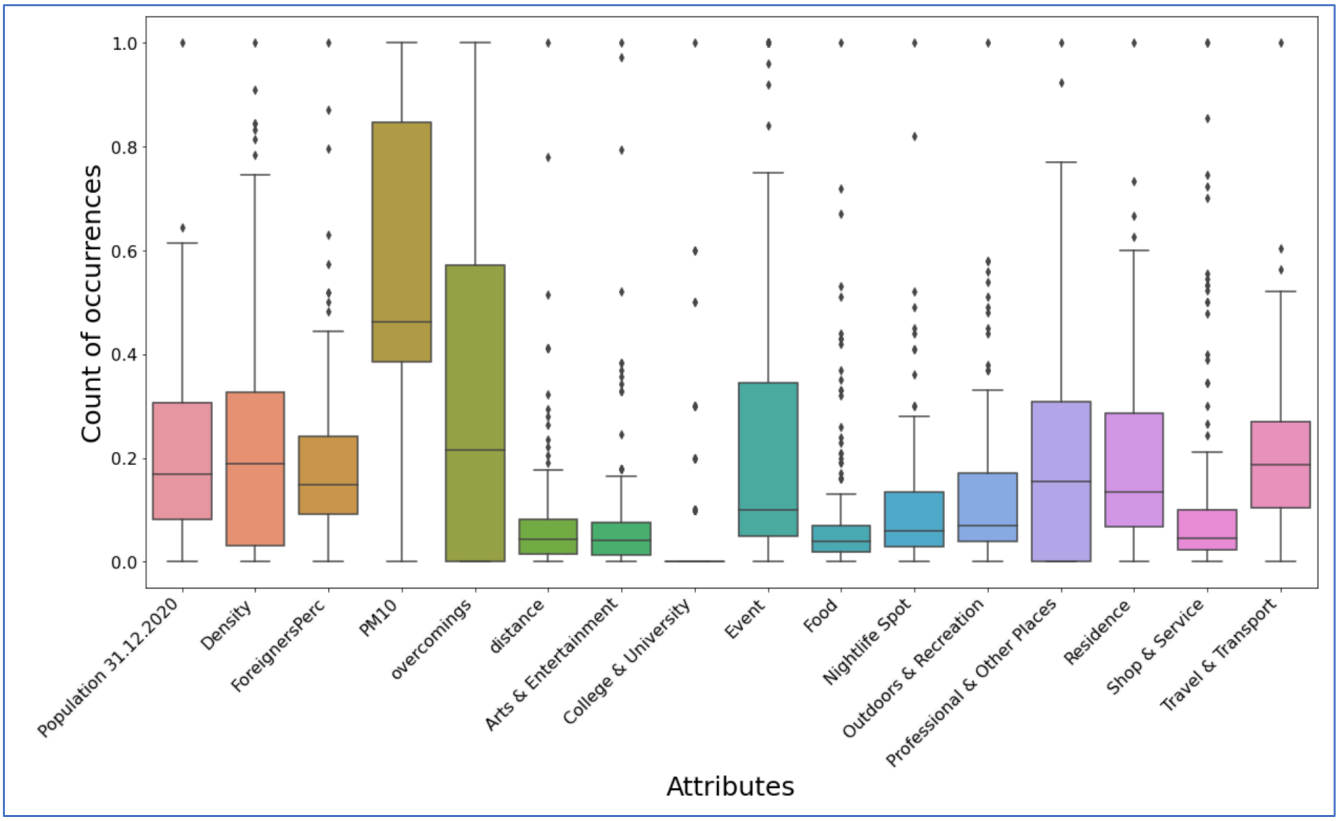

Figure 4 - Boxplots of all explanatory variables after normalization

3.7 K-means clustering

K-means clustering, an unsupervised machine learning algorithm, has been used to group similar data points together and discover underlying patterns.

The elbow method has been applied to find the optimal number of clusters, by fitting the model with a range of values for *k* from 1 to 9. The following line chart shows "elbow" (the point of inflection on the curve) when the underlying model is fit with 4 clusters, which is a good indication that the model fits best with $k = 4$.

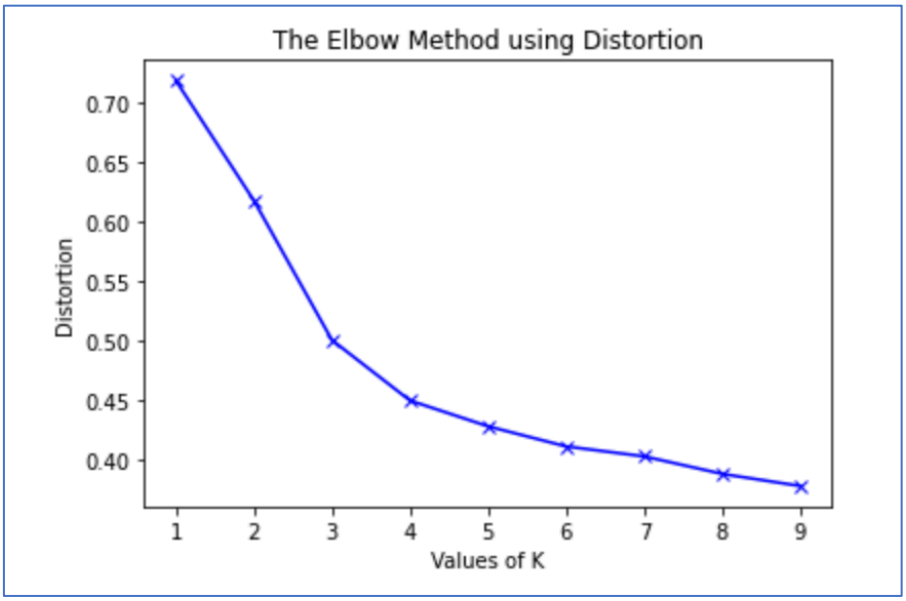

Figure 5 - The elbow method to find k

4. Results

Based on the available data, Rome's 155 neighborhoods have been divided into 4 clusters.

Cluster 0 contains 44 neighborhoods: all neighborhoods in the south-west part of the city, both inside and outside the GRA, an orbital motorway that encircles Rome, 68,2 km long, and some neighborhoods in the east, inside the GRA. They are marked in light blue on the map that follows.

Cluster 1 contains 69 neighborhoods in the north, east and south east of Rome, both inside and outside the GRA, plus the neighborhood of Ostia Nord in south west, on the coast. Cluster 1 is marked in dark blue on the map.

Cluster 2 numbers only 13 neighborhoods, all located in the central part of the city. They are marked in red on the map.

Cluster 3 numbers 29 neighborhoods, all in the area surrounding the city center, within the GRA. They are marked in yellow on the map.

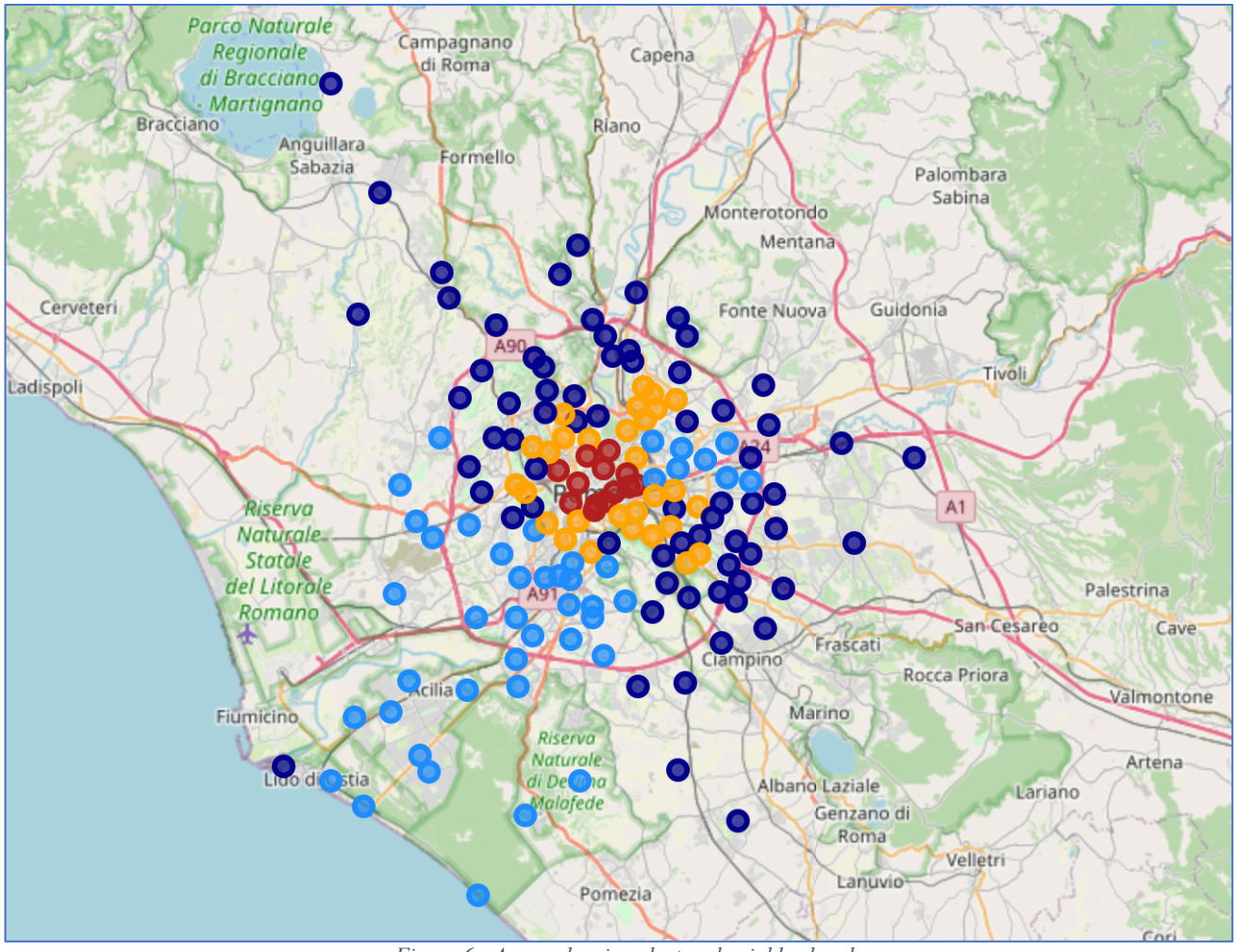

Figure 6 - A map showing clustered neighborhoods

The boxplots of the clusters show differences and similarities between them.

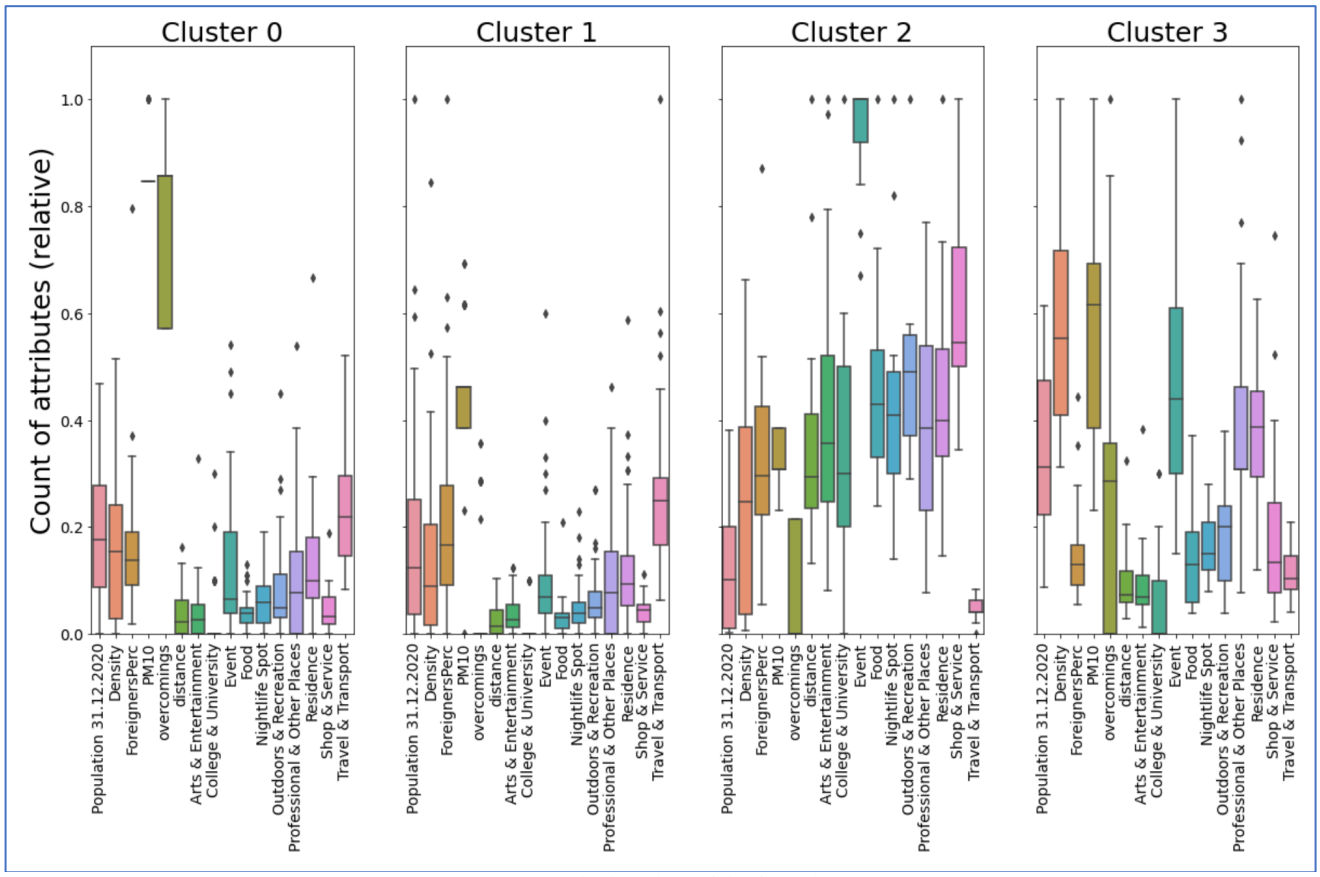

Figure 7 - The boxplots of the four clusters

Also, the average value of most important features has been calculated for each cluster, including the average house price in each cluster, that had not been used for clustering. The means are stored in the following table:

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Mean Population	16085	15086	11144	30150
Mean Density	4206	3276	6496	14653
Mean Foreigners (%)	12,4	15,4	22,2	12,5
Mean PM10	30	25	23	26
Mean Overcomings	15	5	5	8
Mean distance from center (km)	12	10	2	7
Shop&Service	10	9	34	27

Price (euro)	2786	2782	5467	3758

The visual analysis of the clusters (see boxplots above) and the comparison of the mean values of features for all clusters showed the first two clusters having similar distributions, with Cluster 1 denoting slightly lower values for all features but the percentage of foreign-born residents. The only significant difference between Clusters 0 and 1, in addition to the percentage of foreign-born residents, seems to be air pollution: the neighborhoods in Cluster 0 had higher levels of PM10 and more days with PM10 level exceeding the safety threshold of 50 $\mu g/m^3$.

Clusters 2 and 3 present definitely higher values for almost all examined features, than the first two clusters. The density of population is much higher in these neighborhoods, with the mean of 6496 residents/square km for Cluster 2 and 14653 for Cluster 3. There are also more venues in all categories, especially in Cluster 2. Also, Cluster 2 neighborhoods have better air quality and a higher mean percentage of foreigners (22,2% compared with 12,5% for Cluster 3, 12,4% for Cluster 0 and 15,4% for Cluster 1).

5. Discussion

The K-Means algorithm applied to our dataset divided Rome's neighborhoods into 4 clusters as described in the Results section. The mean house price per square meter was very similar for the first two clusters, 2786 euros for Cluster 0 and 2782 for Cluster 1. The average prices increase significantly for Cluster 3 (3758 euro/sqm), and even more for Cluster 2 (5467 euro/sqm).

Hence, we can observe that the highest house prices per square meter can be found in the neighborhoods of the strict center of the city that form Cluster 2. The mid-range prices can be found in the neighborhoods adjacent the city center that form Cluster 3. Finally, the lowest house price averages belong to clusters 0 and 1, whose neighborhoods are located further away from the city center.

Therefore, notwithstanding the fact that FourSquare data is unbalanced, with some categories, such as food, over-represented, the clustering made quite sense, with the neighborhood distribution matching broadly the price ranges given at www.immobiliare.it, as shown on the maps below. We can see that the areas marked in red and yellow roughly correspond to each other and that the areas marked in blue on our map (both light and dark blue) correspond to the areas colored green on the web-sourced pricing map. In fact, the mean house price for the two blue clusters is very similar. This means that some of their features balance each other, leading to similar pricing.

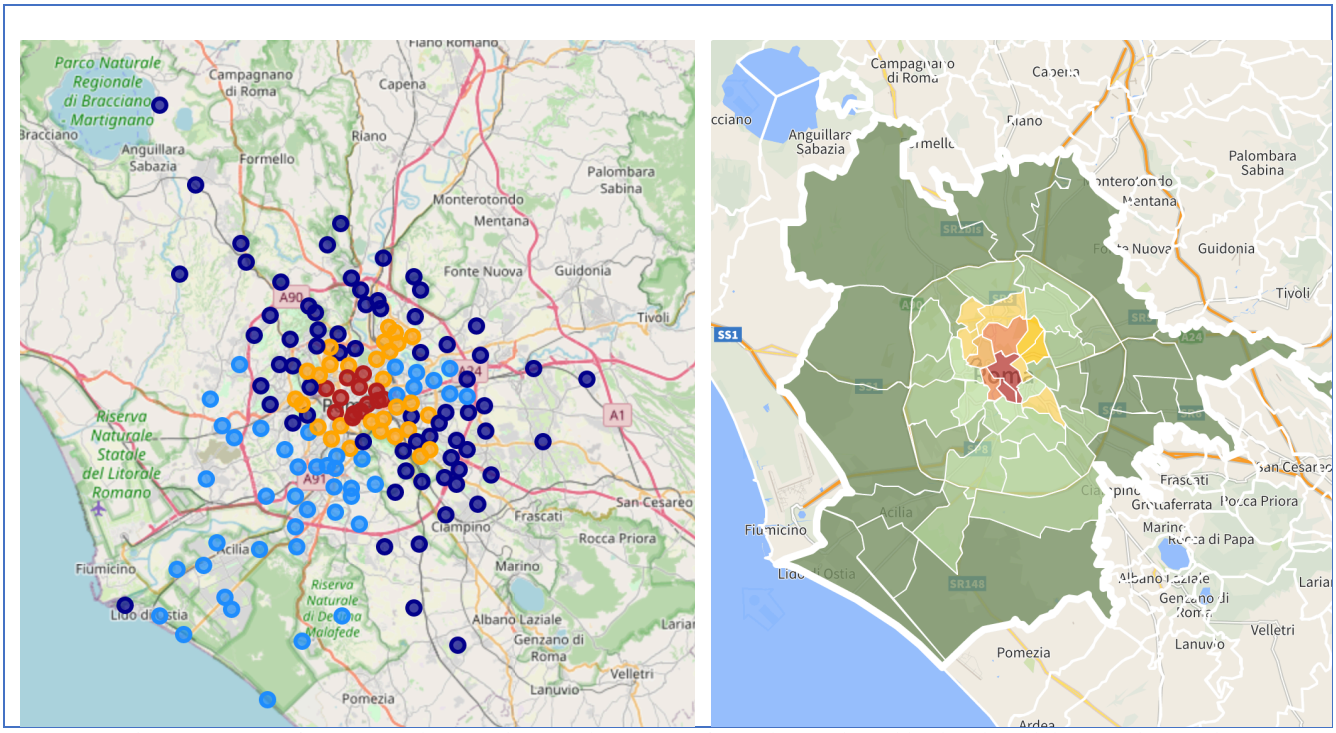

Figure 8 - The comparison of two maps: the map showing the centers of 155 clustered neighborhoods and the map showing house price areas at www.immobiliare.it

6. Conclusions

The goal of the project was to use data gathered from FourSquare (free developer account) and other open sources to cluster the neighborhoods in Rome and understand in what way the characteristics of each area affect house prices.

The clustering model showed that house prices depend on the density of population, the traffic (air pollution), the presence of venues as shops, restaurants, cinemas etc. and the distance from the city center. Observing all the discussed indicators and their trends over time could greatly enhance property valuation and help investors predict in which neighborhoods the prices are likely to increase and in which to decrease.

To make the model more accurate, some other features could be added, like public transport availability, income statistics or crime statistics.

An interesting development would be also the use of Google Street View images and deep learning models for image recognition to assess the structural condition of buildings in each area.