IBM Capstone project

Moving house after COVID19

Fabrizio Ruffini

Abstract

After the **COVID-19 outbreak**, the working habits will likely change from an office-centered perspective to a more **home-based** culture with a prevalence of **smart-working approach**. Thus, moving to cities where the life quality is higher, and in particular to green neighborhoods will possibly become important, also to provide your family with a **sane** and **safe** environment where to live.

In this analysis we are going to compare the **four main Italian cities** in terms of **pollution** with a focus on **air quality**, and once we chose the "most interesting one" we will look for an **appealing neighbourhood** in terms of life quality, looking at the availability of green areas (parks, gardens...), small shops and also their crowd-levels.

Introduction

Business problem

Moving from house to house is one of the most interesting business case. After the covid19 outbreak, it is likely that our moving preferences will change; for example, the lockdown experience allowed us to better appreciate large internal spaces and the possibility of having a garden outside. In addition, the working habits will likely change from an office-centered perspective to a more **home-based** culture with a prevalence of **smart-working approach.** In this context, while moving house you will naturally give more and more attention to information about safe and sane neighbour environment. In this analysis we are going to compare the **four main Italian cities** in terms of **pollution** with a focus on **air quality**, and once we chose the "most interesting one" we will look for an **appealing neighbourhood** in terms of life quality, looking at the availability of green areas (parks, gardens...), small shops and also their crowd-levels.

Stakeholders

Estate agents are naturally interested in all added value insights that can represent a competitive advantage with respect to other companies. In particular, more and more interest is currently given to green topics, and we expect that the health topic will become more and more important after the current 2020 situation.

Data acquisition and description

Pollution data

Pollution data are taken from Copernicus Atmosphere dataset. As a representative dataset, we took the whole 2019 year, and we also added the most recent data available at the time of the analysis (up to June 2020). The pollutant variables chosen are NO2, O3, PM2.5, PM10. Note that there are also additional pollutants, such as SO2, CO, NO, toluene, benzene and many harmful others. They are not treated here.

NO2: nitrogen dioxide level measured in $\mu g/m^3$.

- Source: In cities comes from motor vehicle exhaust (about 80%), petrol, metal refining, electricity generation from coal-fired power stations.
- Effect: Long-term exposure is a cause of chronic lung diseases, and are harmful for the vegetation. Nitrogen Dioxide is gaseous air pollutants which is released mainly during fuel combustion from the reaction of nitrogen and oxygen gases. Coughing, wheezing or difficulty breathing

More about NO2 and its effects can be found here https://www.epa.gov/no2-pollution/basic-information-about-no2

O3: Ground-level Ozone measured in $\mu g/m^3$. Ozone is a gas that is formed when nitrogen oxides react with a group of air pollutants known as 'reactive organic substances' in the presence of sunlight.

- Source: Emitted by cars, power plants, industrial boilers, refineries, chemical plants.
- Effect: High levels can produce asthma, bronchitis or other chronic pulmonary diseases in sensitive groups or outdoor workers. Also, chest pain, coughing, throat irritation, and airway inflammation.

More about O3 and its effects can be found here https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics

PM2.5: Besides gaseous pollutants, the atmosphere can also be polluted by particles. These particles have a divergent composition and size and are sometimes called aerosols. They are often catalogued as 'floating dust', but are best known as particulate matter (PM). PM2.5 are particles smaller than 2.5 µm level measured in $\mu g/m^3$. The size of these particles allow them to penetrate into the gas exchange regions of the lungs (alveolus) and even enter the arteries.

- Source: Formed from construction sites, unpaved roads, fields, smokestacks or fires.
- Effect: Irregular heartbeat, aggravated asthma, decreased lung function. Long-term exposure is proven to be related to low birth weight and high blood pressure in new-born babies.

More about PM and its effect here https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

PM10: particles smaller than 10 µm. Even though the cannot penetrate the alveolus, they can still penetrate through the lungs and affect other organs.

- Source: Formed from construction sites, unpaved roads, fields, smokestacks or fires.
- Effect: Irregular heartbeat, aggravated asthma, decreased lung function. Long term exposure can result in lung cancer and cardiovascular complications.

More about PM and its effect here https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

AQI: Using pollutant variables it is possible to evaluate an index called **air quality index (AQI)** used by government agencies to communicate to the public how polluted the air and to easy understand the correlated risks. Over the time, there has been different air quality indices. In the following, we will use the definition based on this document from EPA (Environmental Protection Agency of United States). https://www.epa.gov/sites/production/files/2014-05/documents/zell-aqi.pdf. AQI converts the measured pollutant concentrations in a communities' air to a number on a scale of 0 to 500 and above. Preferred way to communicate is via a color-coded Air Quality Index (AQI) that is easy for the public to understand. It provides indicator of the quality of the air and its health effects as shown in the following tables.

Air Quality Index (AQI)

AQI Value	AQI Category	AQI Color
0 - 50	Good	Green
51 - 100	Moderate	Yellow
101 - 150	Unhealthy for Sensitive Groups	Orange
151 - 200	Unhealthy	Red
201 - 300	Very Unhealthy	Purple
301 - 500	Hazardous	Maroon

Figure 2: AQI categories from ref aq2

AQI Health Messages

AQI Value	Health Message	AQI Color
0 - 50	None	Green
51 - 100	Unusually sensitive people should reduce prolonged or heavy exertion	Yellow
101 - 150	Sensitive groups should reduce prolonged or heavy exertion	Orange
151 - 200	Sensitive groups should avoid prolonged or heavy exertion; general public should reduce prolonged or heavy exertion	Red
201 - 300	Sensitive groups should avoid all physical activity outdoors; general public should avoid prolonged or heavy exertion	Purple
301 - 500	Everyone should avoid all physical activity outdoors	Maroon

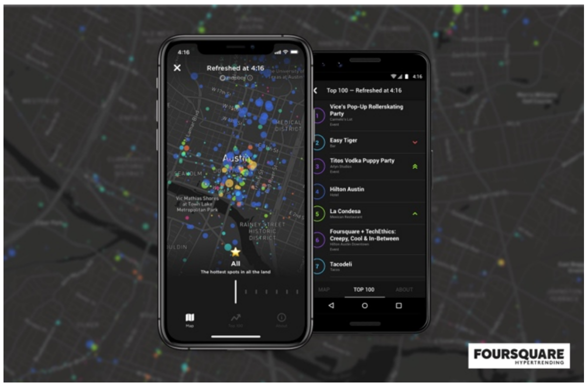
Figure 1: AQI Health messages from ref aq2

Neighbourhoods life quality data

Neighbourhoods life quality data are taken from google and Foursquare opendata APIs.

Venue data

Foursquare Places Database allows you to access precise, up-to-date community-sourced venue data. There is a large selection of rich and firmographic location data with the description of locations and trends info. In particular:



- 60M+ commercial POI including restaurants, shops & services, and more
- 941 Venue Categories [(full list available here)](), updated quarterly
- 1M+ fresh* Tips
- 18.5M+ fresh* Photos
- 1M+ Tastes, popular search terms generated by our consumer app users
- 23K+ Chains, including top QSRs, retailers, auto dealerships, and more
- 7.6M+ Venues with popular hours

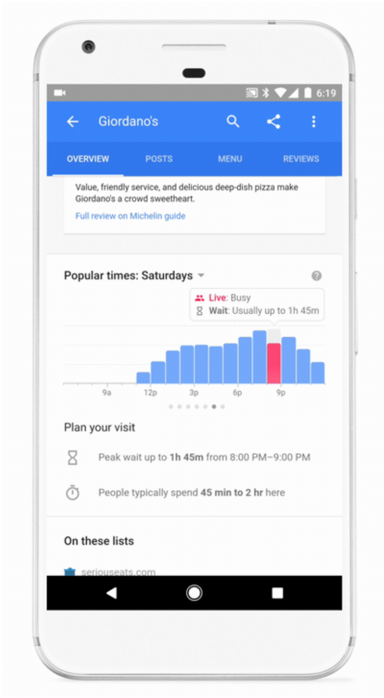
*Added within the past six months from June 2020

Foursquare database details can be found here https://developer.foursquare.com/docs/places-database/details/

Popular times data

Google will be used to find **Popular times, wait times, and visit duration information** of cities' places (ref: https://support.google.com/business/answer/6263531?hl=en). This data is the one that appears below the regular business information on Google Maps and Search.
To determine popular times, wait times, and visit duration, Google uses aggregated and anonymized data from users who have opted in to Google Location History. Popular times, wait times, and visit duration are shown for your business if it gets enough visits from these users.
Visit data may include:

- **Popular times graph**: This graph shows how busy your location typically is during different times of the day. Popular times are based on average popularity over the last few months. Popularity for any given hour is shown relative to the typical peak popularity for the business for the week. For example, in the image below, 8 PM–9 PM on Saturday is one of the more popular times of the week for this business.
- **Live visit data**: This data shows how active your location is right now. Live visit data is updated in real time and overlaid on the popular times graph. For example, in the image below, the highlighted section of the graph represents how active the location is right now compared to its usual level of activity.
- **Visit duration**: This data shows how much time customers typically spend at your location. Visit duration estimates are based on patterns of customer visits over the last several weeks.
- **Wait time estimates**: This data shows how long a customer would have to wait before they receive service during different times of the day. It also shows the peak wait time for each day of the week. The displayed wait time is based on patterns of customer visits over the last several weeks. Wait time estimation differs for different business types. For instance, a sit-down restaurant's wait time reflects how long customers wait before they are seated.



Neighbourhood data

General neighbourhood data such as traffic information, shopping centers, green/entertainment areas and so forth are available from open google api searches, and they will be used to get insights on the neighbourhood life-quality. Google maps api overview can be found here: https://developers.google.com/maps/documentation/javascript/tutorial

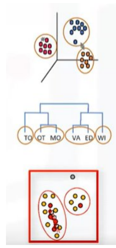
Figure 3: example of info that can be gathered using google API

Methodology

The analysis is composed by two steps.
In the first step, an exploratory analysis comparing pollution data from the four main Italian cities namely Milan, Naples, Rome, and Turin is done, with the goal of understanding which is the less polluted. Histograms, scatter plots, are shown and median and variance calculated to have both a qualitative and a quantitative evaluation. Additionally, the air quality index will be calculated to have the health related summary considerations.
In the second step, once the less polluted city has been found, we will focus on the neighbourhoods's characteristic with the goal of providing information on where to go to live based on life quality parameters. In this analysis, clustering techniques using K-means are involved to cluster similar neighbourhoods. k-means clustering is a method of vector quantization that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. Generally speaking, clustering is the process of grouping together similar non-overlapping data (clusters) with an **unsupervised approach** (in this context, sometimes it is referred to as a similarity approach). Similarity is often calculated using distance metrics (Euclidean, Minkowsky…). Clustering is ypically used for exploratory and preprocessing data analysis, such as summary generation, outlier detection, finding duplicates tasks.

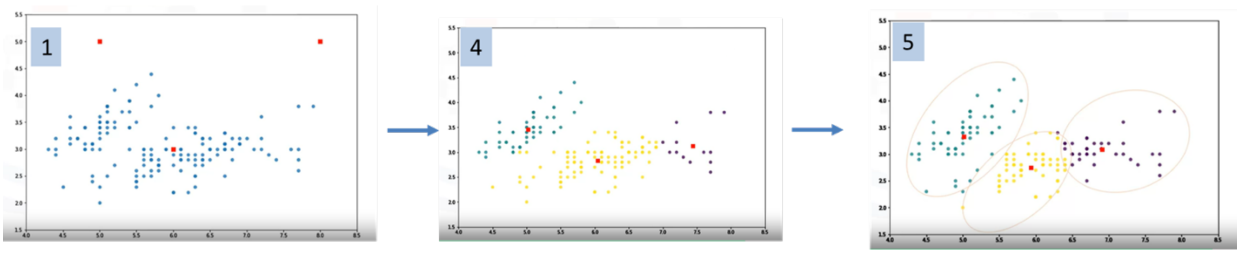
There are different clustering algorithms:

1. **Partition based clustering** (eg. k-Means, k_Median, Fuzzy c-Means)
   They are relatively efficient also with large datasets
2. **Hierarchical clustering** (eg. Agglomerative, divisive)
   They produce trees of clusters and are efficient for small datasets
3. **Density-based clustering** (eg. DB-Scan)
   They produce arbitrary shaped clusters and are typically effective on spatial-datasets.



In particular, **k-Means** works as follows:

0. Fix a number of cluster, based on your a-priori knowledge of the data
1. Randomly place k centroids, one for each cluster
2. Calculate the distance of each point from each centroid
3. Assign each data point (object) to its closest centroid, creating a cluster
4. Recalculate the position of the k centroids
5. Repeat steps 2-4, until the centroids no longer moves.



k-Means PROs:

a. it produces easy to interpret sphere-like clusters
b. it will always find a minimum (but not a global minimum)

k-Means CONs:

a. it is not effective on arbitrary shaped clusters
b. it needs a-priori number of clusters k
c. it assigns all points to a cluster even if they do not belong in any (outlier detection is not possible)

Data analysis I: finding best city

An comparative analysis of pollution data from the four main Italian cities namely Milan, Naples, Rome, and Turin is performed.

Daq (data acquisition)

Pollution data are taken from Copernicus Atmosphere dataset from the following github repository: https://github.com/CopernicusAtmosphere/air-quality-covid19-response. As a representative dataset, we took the whole 2019 year, and we also added the most recent data available at the time of the analysis (up to June 2020). The pollutant variables under investigation are, as described above, NO2, O3, PM2.5, and PM10.

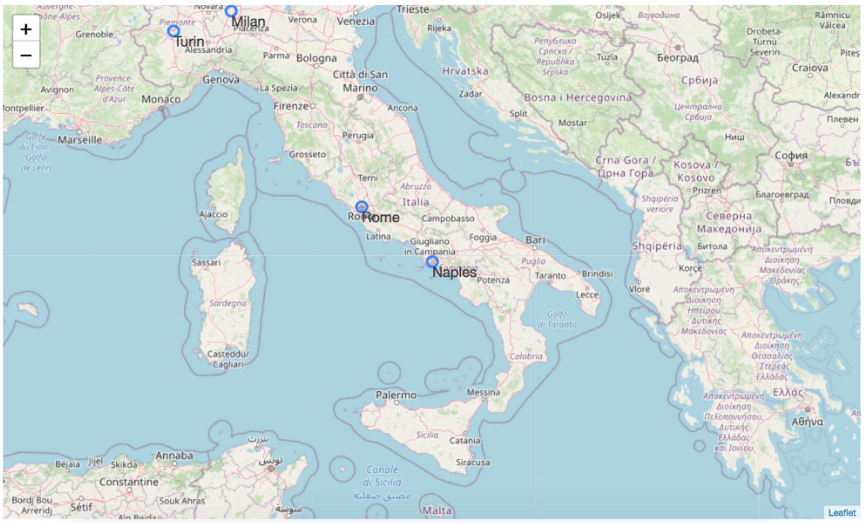


Figure 4: four Italian cities involved in the present analysis

Table 1 reports the full dataset description. Data corresponds to 2080 observations, that is 502 daily observations for 4 cities, from 1st January 2019 to 3rd June 2020.

Table 1: dataset description

	NO2	O3	PM10	PM2.5
count	2080.000000	2080.000000	2080.000000	2080.000000
mean	25.594500	51.164687	22.368207	17.423995
std	13.529827	23.357247	12.392036	10.719891
min	4.460000	0.960000	3.620000	2.980000
25%	15.157500	35.715000	14.030000	10.220000
50%	22.190000	54.175000	19.665000	14.725000
75%	32.797500	69.010000	26.710000	21.130000
max	78.280000	129.070000	78.940000	70.270000

Data exploration

We explored the dataset looking at the different cities, with the goal of understanding which is the more polluted. Time trends and histograms are reported below.

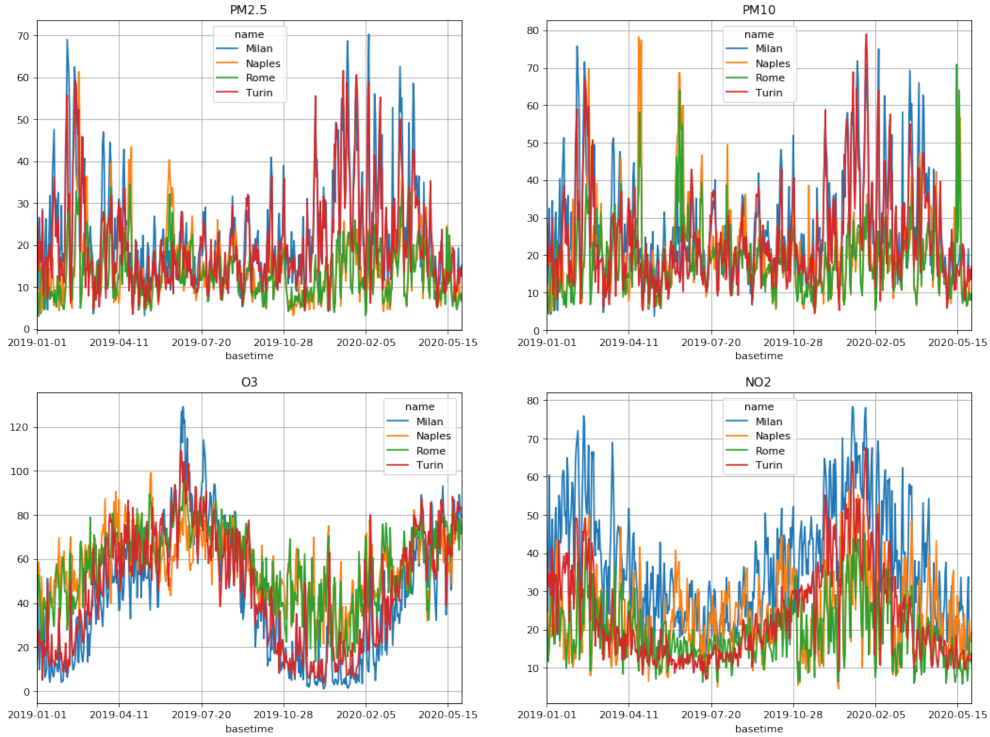


Figure 5: global time-trend for the 4 cities

Figure 5 shows the global time trend for all the variables. A seasonal pattern is clearly visible for O3 and NO2, while it is less evident in PM2.5 and PM10. Generally, Milan and Turin shows peaks and most pronounced behaviours with respect to Naples and Rome.

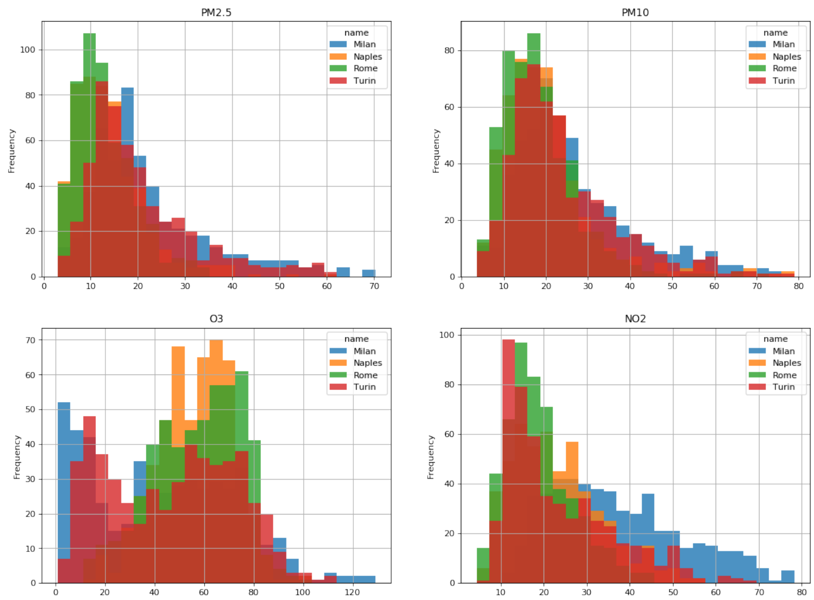


Figure 6: global histograms for the 4 cities

Figure 6 shows the global histograms all the variables. Rome PM2.5 and PM10 histograms are clearly the ones with smaller median value, while it is interesting the two-peaks structure of O3 clearly visible for Turin and Milan.

Data processing

By a simple calculation of the median and the standard deviation of each city subsample, we can see that Rome is the one where PM2.5, PM10, and NO2 levels are smallest, while O3 is the highest one. Median has been chosen as a robust estimator for the mean true value instead of the arithmetic average, to avoid outliers-related issues.

	PM2.5	PM10	NO2	O3
Milan	18.955	23.005	34.545	42.345
Rome	11.610	17.105	17.435	59.815
Turin	17.030	20.790	19.090	49.750
Naples	12.290	17.925	21.650	58.220

Table 2: median values

	PM2.5	PM10	NO2	O3
Milan	12.798048	14.049052	15.051572	28.922493
Rome	6.137670	9.056311	7.822407	17.653542
Turin	11.476924	12.764039	12.419786	25.850945
Naples	7.633185	11.412343	10.237374	14.930984

Table 3: standard deviation values

To provide more user-friendly information, we will also calculate the air quality index for Rome. Since the data are daily-based, we cannot calculate the air quality indices for all of them (in particular, we can only look at PM2.5 and PM10).

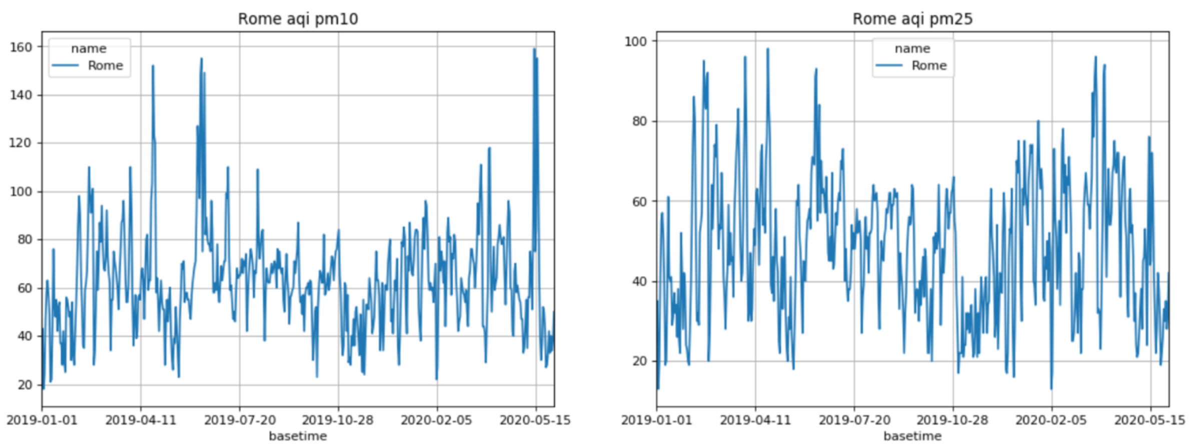


Figure 7: Rome AQI time trend

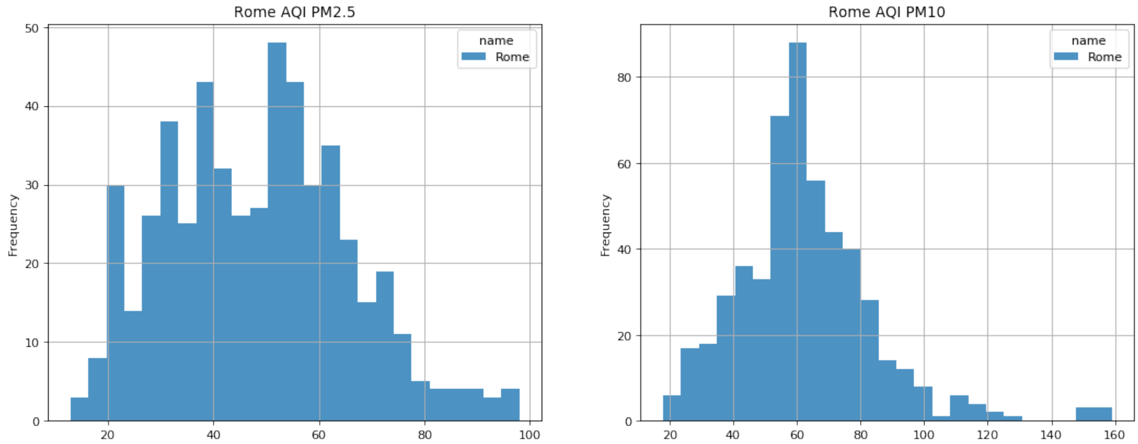


Figure 8: Rome AQI time trend

From the histograms, it can be observed that AQI is typically under the health warning orange-red thresholds.

Air Quality Index (AQI)

AQI Value	AQI Category	AQI Color
0 - 50	Good	Green
51 - 100	Moderate	Yellow
101 - 150	Unhealthy for Sensitive Groups	Orange
151 - 200	Unhealthy	Red
201 - 300	Very Unhealthy	Purple
301 - 500	Hazardous	Maroon

Figure 10: AQI categories from ref aq2

AQI Health Messages

AQI Value	Health Message	AQI Color
0 - 50	None	Green
51 - 100	Unusually sensitive people should reduce prolonged or heavy exertion	Yellow
101 - 150	Sensitive groups should reduce prolonged or heavy exertion	Orange
151 - 200	Sensitive groups should avoid prolonged or heavy exertion; general public should reduce prolonged or heavy exertion	Red
201 - 300	Sensitive groups should avoid all physical activity outdoors; general public should avoid prolonged or heavy exertion	Purple
301 - 500	Everyone should avoid all physical activity outdoors	Maroon

Figure 9: AQI Health messages from ref aq2

Data evaluation and interpretation

Thus, we can suggest that based on the 2019-2020 pollution available data, the most promising city where to move in is **Rome**. Now we shall see in which neighbourhood you should move in.

Data analysis II: finding lovely neighbourhood

Daq (data acquisition)

1. **Neighbourhood location**: Rome is currently divided into 15 administrative division called municipalities. The municipalities’ center coordinates were taken from the relative wikipages.
2. **Venues**: Forsquare API is used to get the most common venues of given location
3. **Popular times**: google api used to get info on specific venues
4. **Traffic**: google traffic layer was used to get info on traffic
5. **Historical monuments**: openstreetmap was used to get info on historical monuments
6. **Parking lots**: openstreetmap was used to get info on parking
7. **Green parks**: openstreetmap was used to get info on green parks

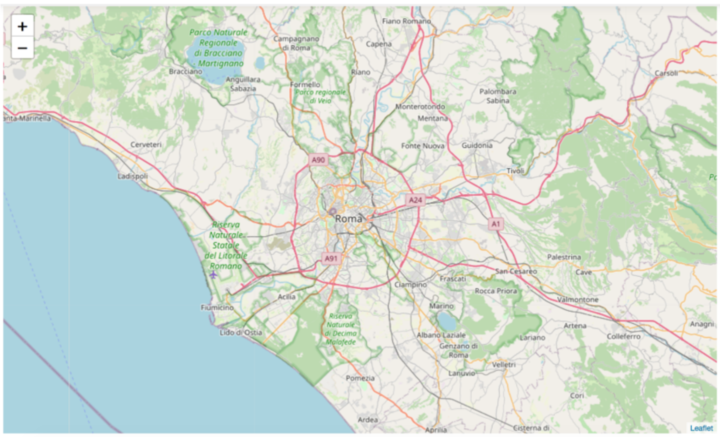
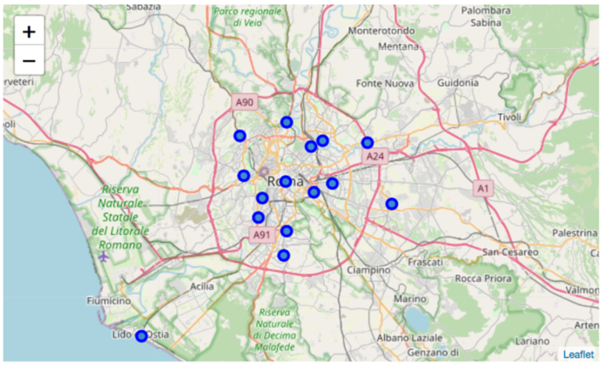


Figure 11: Rome map

Data exploration

The plot below helps in visualizing the geographic details of Rome municipalities.

Municipio	lat	lon
municipio 1	41.893056	12.482778
municipio 2	41.929958	12.518931
municipio 3	41.936080	12.535116
municipio 4	41.933491	12.598746
municipio 5	41.890665	12.548488
municipio 6	41.869657	12.632731
municipio 7	41.881700	12.522800
municipio 8	41.841228	12.484290
municipio 9	41.814879	12.479980
municipio 10	41.730660	12.280531
municipio 11	41.855282	12.444762
municipio 12	41.876099	12.450100
municipio 13	41.899141	12.424158
municipio 14	41.940965	12.418627
municipio 15	41.955435	12.484851



The Foursquare API allows to explore the municipalities and cluster them. The query limit is **100 venues** and the radius is set to **500 meters** for each municipality geographical center.
As a preliminary check that the query is working, I looked for the 5 venues closest to the Rome center

name	categories	lat	lng
Palazzo Merulana	Art Museum	41.889889	12.503289
RADICI - Pizzicheria Salentina	Deli / Bodega	41.892456	12.505773
civico 31	Bed & Breakfast	41.892445	12.505911
Bar Trani	Café	41.893034	12.505343
Caronte	Restaurant	41.892508	12.503160

Data processing

Everything seems to work fine, so I processed the Foursquare query for all municipalities with a radius of 1000 m (each municipality has about 2-3 km range) and made 5 clusters using a k-means approach, one of the most used unsupervised approach. I already discussed in some detail the k-means approach in the methodology section of this document. Here, let's just say that this approach cluster together elements with similar features.

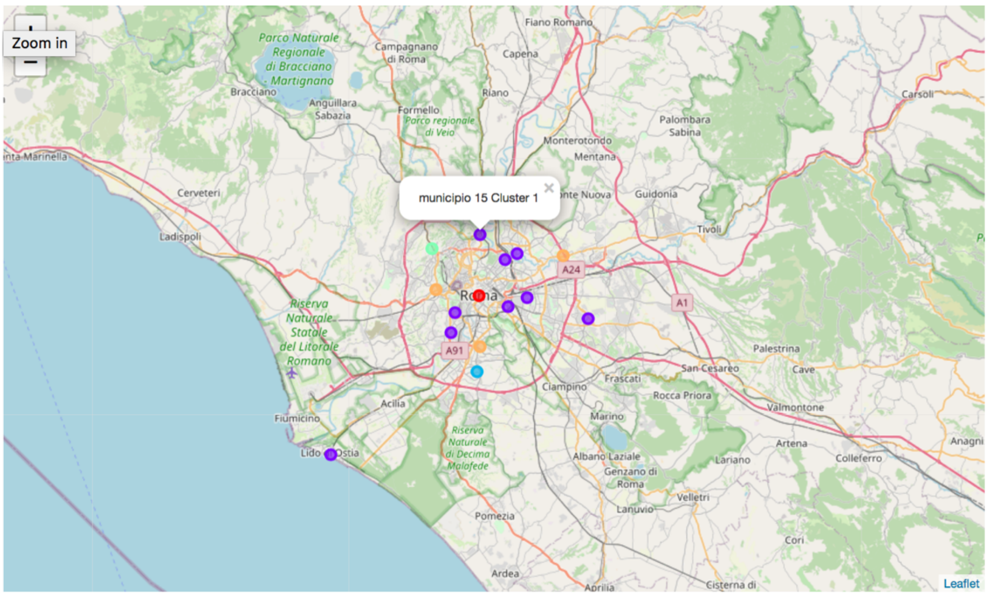


Figure 12: clusters map

Data evaluation and interpretation

As can be observed from the map, the clustering evidenced a lot of similar neighbourhoods (purple, cluster 1) and some other lesser populated clusters. Let's see them in detail:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
municipio 1	Historic Site	Hotel	Monument / Landmark	Temple	Pub	History Museum	Scenic Lookout	Italian Restaurant	Plaza	Fountain

Figure 13: cluster 0

The cluster 0 is characterized by monuments, hotels and historic sites. This is reasonable, since it is the very center of Rome, "municipio I" with il Foro romano and il Colosseo, near the Vatican state.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	municipio 2	Ice Cream Shop	Dessert Shop	Plaza	Italian Restaurant	Pastry Shop	Bookstore	Café	Women's Store	Park	Pub
2	municipio 3	Pizza Place	Cocktail Bar	Italian Restaurant	Café	Hotel	Turkish Restaurant	Mexican Restaurant	Gastropub	Plaza	Pub
4	municipio 5	African Restaurant	Supermarket	Pizza Place	Gym	Noodle House	Café	Sandwich Place	Market	Italian Restaurant	Light Rail Station
5	municipio 6	Shopping Mall	Theater	Plaza	Bus Station	Supermarket	Pizza Place	Hotel	Falafel Restaurant	German Restaurant	Cocktail Bar
6	municipio 7	Plaza	Trattoria/Osteria	Italian Restaurant	Hotel	Hostel	Fish Market	Bistro	Photography Lab	Pizza Place	Thai Restaurant
9	municipio 10	Pizza Place	Beach	Italian Restaurant	Café	Seafood Restaurant	Plaza	Restaurant	Ice Cream Shop	Fast Food Restaurant	Hotel
10	municipio 11	Pizza Place	Gym / Fitness Center	Café	Supermarket	Clothing Store	Bistro	Italian Restaurant	Fast Food Restaurant	Park	Plaza
11	municipio 12	Pizza Place	Café	Restaurant	Italian Restaurant	Ice Cream Shop	Plaza	Supermarket	Bus Station	Steakhouse	Food
14	municipio 15	Bookstore	Trattoria/Osteria	Gym / Fitness Center	Café	Restaurant	Basketball Court	Light Rail Station	Brewery	Sandwich Place	Fish Market

Figure 14: cluster 1

The cluster 1 is characterized by Pizza places, bars and restaurant. This is reasonable, since Rome is a very touristic place.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
8	municipio 9	Pizza Place	Restaurant	Hotel	Dessert Shop	Food	Convention Center	Department Store	Diner	Falafel Restaurant	Fast Food Restaurant

Figure 15: cluster 2

The cluster 2 is characterized by just Municipio 9, where we can see restaurants, convention centers and food venues. It is reasonable that this municipio has been treated by itself, because it is the one in south Rome with the area called EUR, that is an office and convention-like place.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
13	municipio 14	Chinese Restaurant	Italian Restaurant	Park	Supermarket	Women's Store	Food	Department Store	Dessert Shop	Diner	Falafel Restaurant

Figure 16: cluster 3

The cluster 3 is characterized by restaurants and parks. This is reasonable, since this municipio go to the very north of Rome and has a lot of parks close, and also a natural reserve.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	municipio 4	Café	American Restaurant	Rental Car Location	Hotel	Italian Restaurant	Fast Food Restaurant	Motorcycle Shop	Shoe Store	Women's Store	Food
7	municipio 8	Café	Italian Restaurant	Plaza	Japanese Restaurant	Salad Place	Bistro	Mexican Restaurant	Food	Department Store	Dessert Shop
12	municipio 13	Café	Hotel	Italian Restaurant	Supermarket	Steakhouse	Plaza	Gourmet Shop	Gastropub	Food Truck	Fast Food Restaurant

Figure 17: cluster 4

The cluster 4 is characterized by café, restaurants, plazas. They are not so touristic areas, because they are not very close to the center of Rome.

Now, let's assume that you evaluated the different cluster and went to live in the first one, in the center of Rome, because you are very interested in history, monuments and so forth. SO you moved in, unpacked your stuff and now you got hungry. You saw that there is a very interesting food-place called "pizzicheria salentina" with regional delicacies from the southern Italy, and you want to know more about it (for example, when is the best time to go). You can access those information using the google popular times api, for example using the python package populartimes https://github.com/m-wrzr/populartimes and you get these results (here you have an example of query made on Friday 2020-06-05):

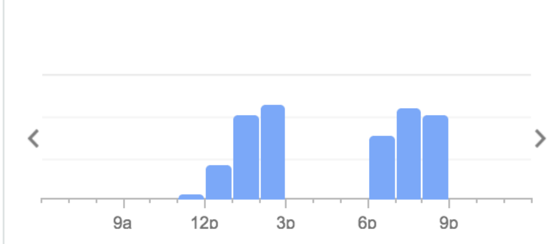
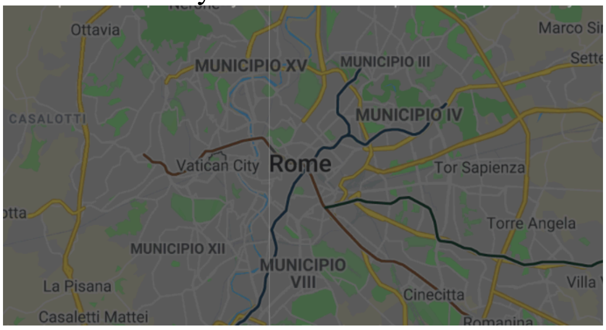


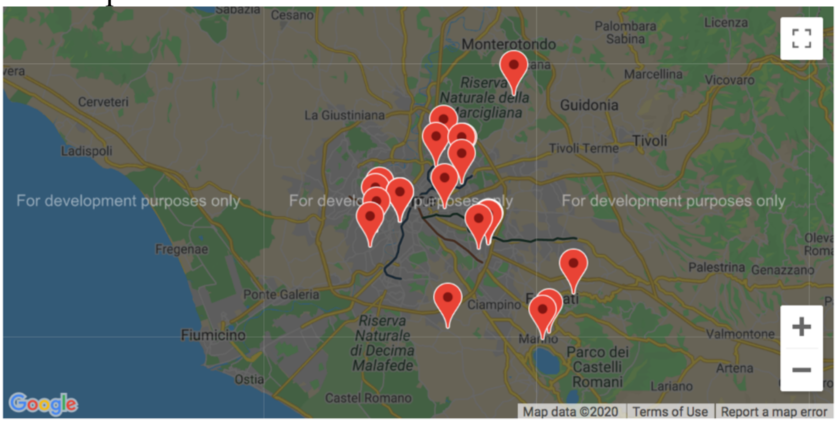
Figure 18: Pizzicheria salentina most frequent hours on Friday

Google has a lot of interesting additional informative layers, so you can look at different life quality information such as traffic (from gmaps.transit_layer() python gmap package https://pypi.org/project/gmaps/), parks or other places from openstreetmap (www.openstreetmap.org using the convenience download website https://overpass-turbo.eu). They are downloaded as geojson and can be easily plotted in maps.

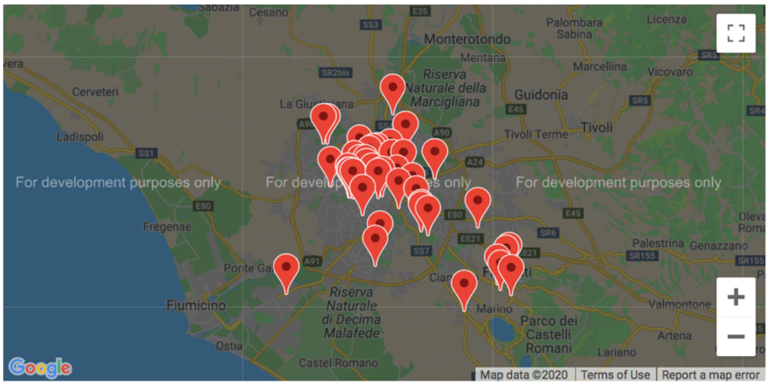
Here some example of additional info that can be acquired from the web.
Rome traffic layer:



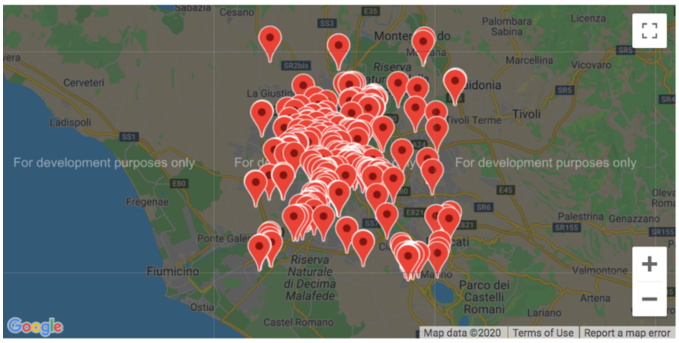
Leisure parks:



Historical monuments



And, of course, parking lots (Rome has lots of them!)



Results

The results obtained are interesting (and it was also very **funny**). We were able to acquire and take a deep look at different information from different point of view, and this could potentially be very beneficial for estate agency to acquire a data science-powered **mind-set** representing a fantastic competitive advantage. This is very important in this period, where you need to be prepared to the new crisis and opportunities that may come after the **COVID19**. We do not have a detailed understanding on how the world will change, so having **additional knowledge** and **expertise** with respect to your competitors will be probably **essential**.

Discussion

To further improve our results, we could:

1. For the air quality index, the data available allowed to calculate only the aqi related to PM2.5 and PM10. Acquiring data with ad-hoc time-resolution would be beneficial. Also, maps of aqi are produced by national agencies, with typical spatial resolution of 10 km. It would be interesting to look at the relative maps to find new insights.

2. As a different approach, the use of another kind of clustering technique (DB-Scan or hierarchical) to see if there are different results. K-Means assign all points to a cluster even if they do not belong to any while density-based clustering locates region of high density and separate outliers. Additionally, k-Means create spherical-shape clusters, while density-based clustering can face arbitrary shaped clusters.

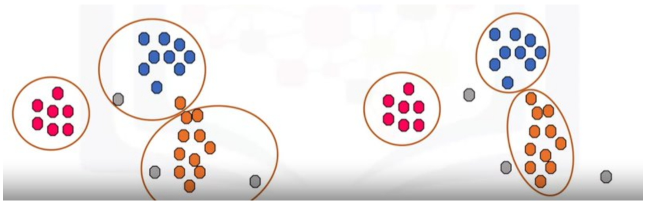


Figure 19: k-Means approach (left), and density approach (right)

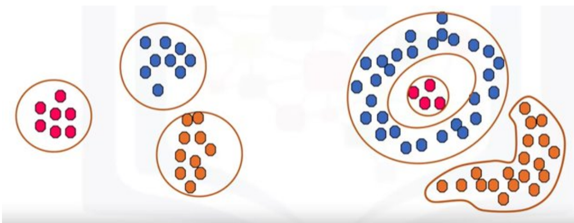


Figure 20: k-Means approach (left), and density approach (right)

3. The data acquired in a map-like style of the life quality could be organized as tables or plots, to better quantify their relevance.

Conclusion

In conclusion, the capability of using data analytics is very important and can be applied where numbers are available.

In this uncertain times, it’s good to enlarge your expertise and polish your skills, and this is the main reason I took this capstone project activity.

Thanks for reading up to here,

Fabrizio Ruffini

Bibliography

Info on air quality:

- Aq1: https://towardsdatascience.com/sensing-the-air-quality-5ed5320f7a56
- Aq2: https://www.epa.gov/sites/production/files/2014-05/documents/zell-aqi.pdf

Info on NO2 https://www.epa.gov/no2-pollution/basic-information-about-no2

Info on O3 https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics

Info on PM https://www.epa.gov/pm-pollution/particulate-matter-pm-basics

Foursquare data: https://developer.foursquare.com/docs/places-database/details/

Google maps data: https://developers.google.com/maps/documentation/javascript/tutorial