Ilia Notin

Optimal place to live in Milan

Milan Cathedral

Dataset and motivation. Actual task definition/research question

I plan to move to Italy (hopefully) and to be more precise in Milan. That city is a vibrant multicultural megapolis with all its pros and cons. Therefore I have to choose an optimal place for living (an apartment to rent). I have identified several parameters for choosing one:

- proximity to employers: I will look for existing offers, find locations of the companies and calculate median location. Then I will assess the distance between each apartment and the median location.
- remoteness from disadvantaged areas. I will find locations of disadvantaged areas (based on reviews in the internet) and then calculate distance to the closest disadvantaged are for each apartment.
- vegetation concentration. I will calculate the amount of green areas per km2 for each of 9 zones (districts) in Milan and then define in what zones apartments are located.
- air quality. I will use information on air pollution for the last year and define mean scaled concentration of harmful substances and dust particles which monitoring stations provide. Then I will build a regression model that would predict that value based on location in the city.
- price (and location): I will take into account apartments of certain range (€1000 to €2000) which can be modified in the notebook. I will find existing offers and define their location. Then for each of the apartments I will assess the parameters above

Based on those parameters I calculate the total score based on their weighted sum. Weight of each parameter can be based on priorities of a user (I used all equal).

Then each apartment is assigned this score which helps to choose the best one.

As an addition I will create a model which would predict rental price based on the location, total score and apartment type (1 bedroom, 2 bedrooms, etc.)

Literature review

- https://www.moneycrashers.com/where-should-i-live-decide-best-places/
  - I found a web-site where some factors influencing the choice of a place to live are presented. There are also some web-sites links which allow to choose the best city to live based on some statistical data. My goal is a little bit more precise which is to choose the best apartments from existing ones within one city.

Github repository

https://github.com/inotin/finalProjectCommit/

https://mybinder.org/v2/gh/inotin/finalProjectCommit/HEAD

Data
1. Proximity to employers

Source

https://it.indeed.com/

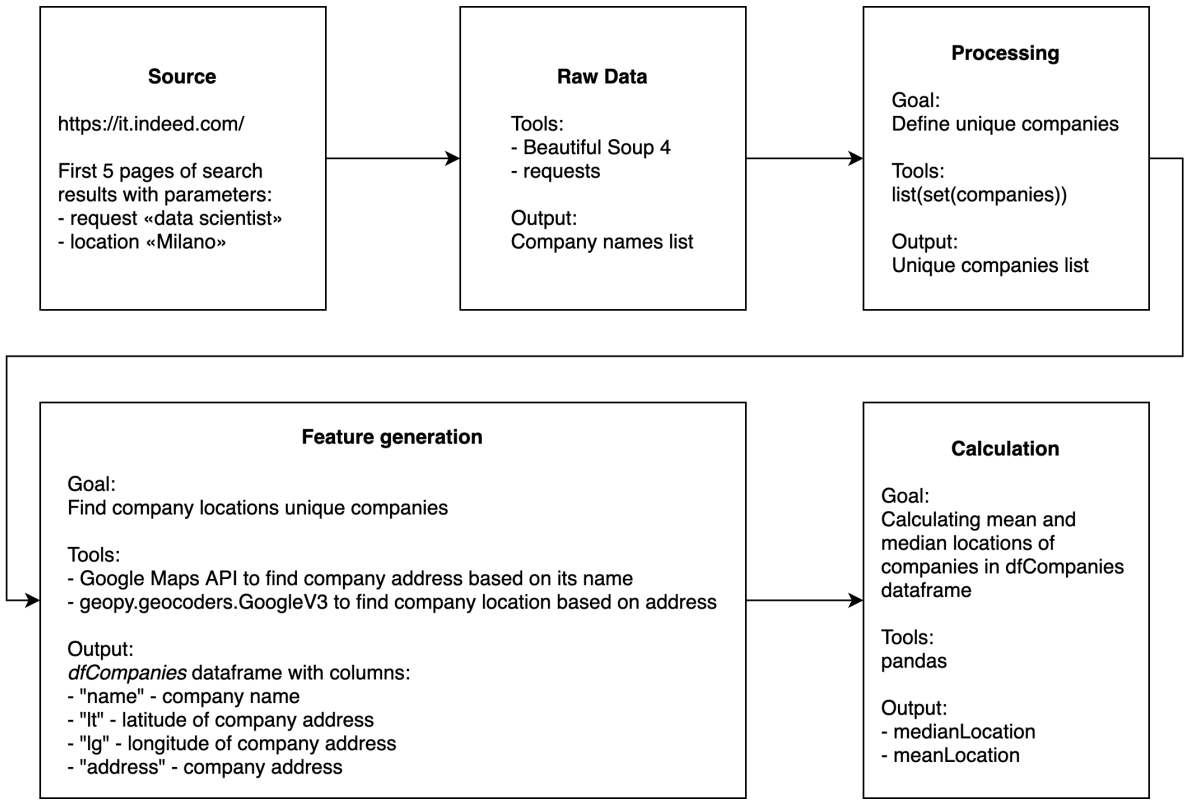
Method of gaining

Web scraping using Beautiful Soup and requests libraries

Raw data contents

Company names from 5 pages of results based on search request «data scientist» and location «Milano»

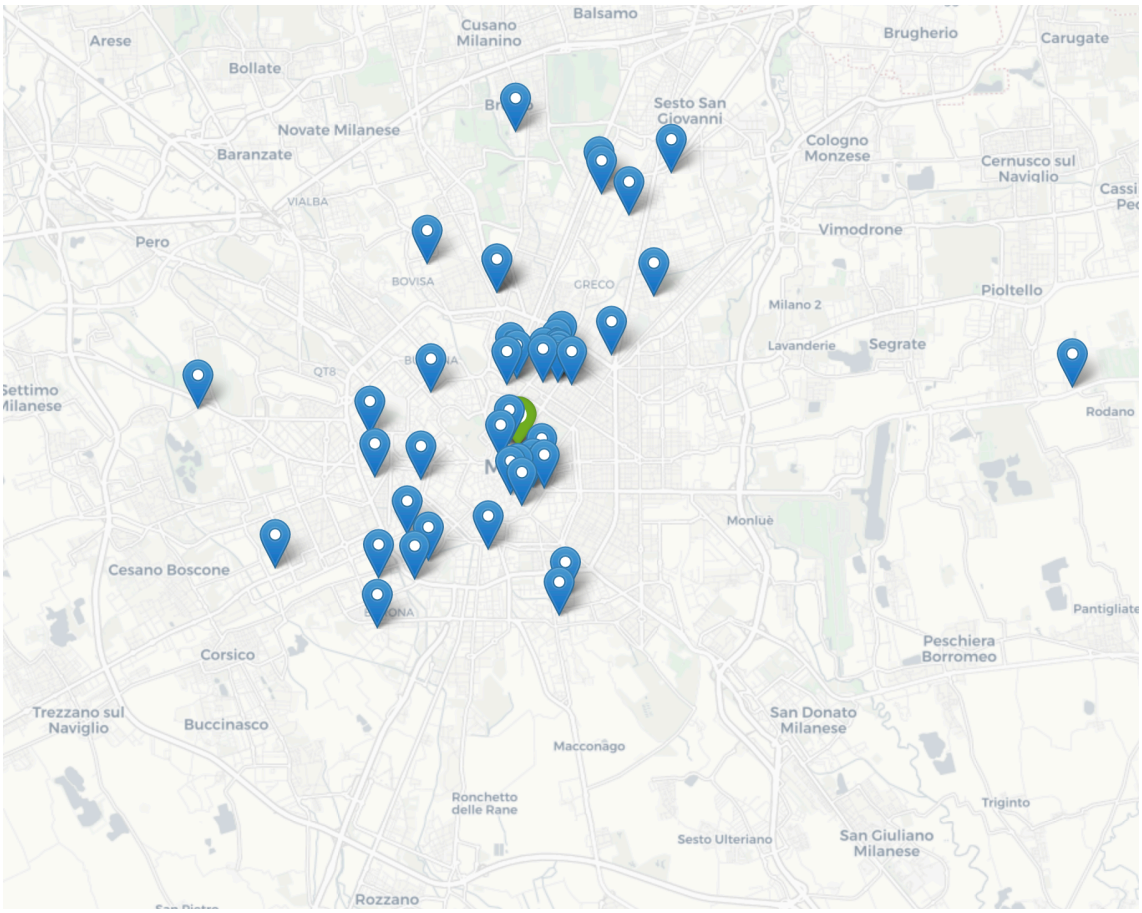
Workflow



Output Dataframe

	name	lt	lg	address
19	AURIGA SPA	45.484071	9.200962	Via Giovanni Battista Pirelli, 7, 20124 Milano...
17	PwC	45.450796	9.181933	Via Pietro Custodi, 12, 20136 Milano MI, Italy
5	Start Hub Consulting	45.500401	9.184232	Via Privata Roberto Bracco, 6, 20159 Milano MI...
57	TK SOLUZIONI SRL	45.623214	9.032359	Via Gaudenzio Ferrari, 21, 21047 Saronno VA, I...
43	Generali Italia	45.481032	9.166182	Corso Sempione, 36, 20154 Milano MI, Italy
59	Abbvie	40.063354	-75.676718	505 Eagleview Blvd, Exton, PA 19341, USA
46	AWS EMEA SARL (Italy Branch)	45.464204	9.189982	Milan, Metropolitan City of Milan, Italy
3	KPMG	45.483653	9.201544	Via Vittor Pisani, 27/31, 20124 Milano MI, Italy
55	Page Personnel	45.465709	9.196654	Galleria Passarella, 2, 20122 Milano MI, Italy
41	Ali Spa	45.462876	9.194306	Via Larga, 2, 20122 Milano MI, Italy

Exploratory data analysis



We can see that companies are mostly located in the central, northern and western parts. Median location (green marker) is almost in the center.

2. Remoteness from disadvantaged areas

Source

https://www.tripadvisor.com/ShowTopic-g187849-i143-k1372098-Bad_areas_to_avoid-Milan_Lombardy.html

https://www.quora.com/What-are-the-most-dangerous-areas-of-Milan

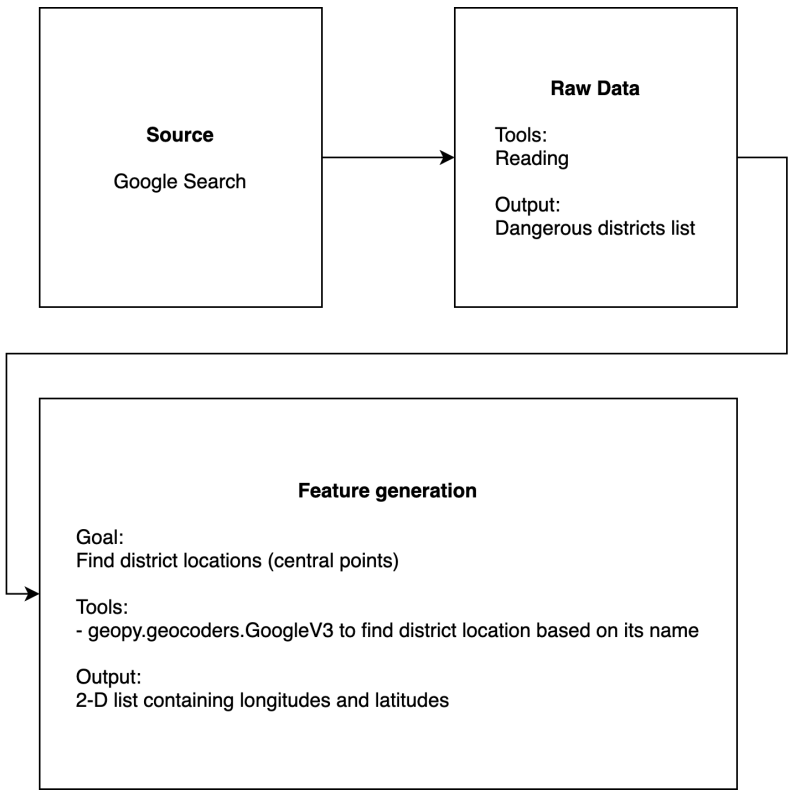
Method of gaining

Reading (as there are not so many disadvantaged districts)

Raw data contents

I constructed a string list based on the reviews from the web-sites above

Workflow



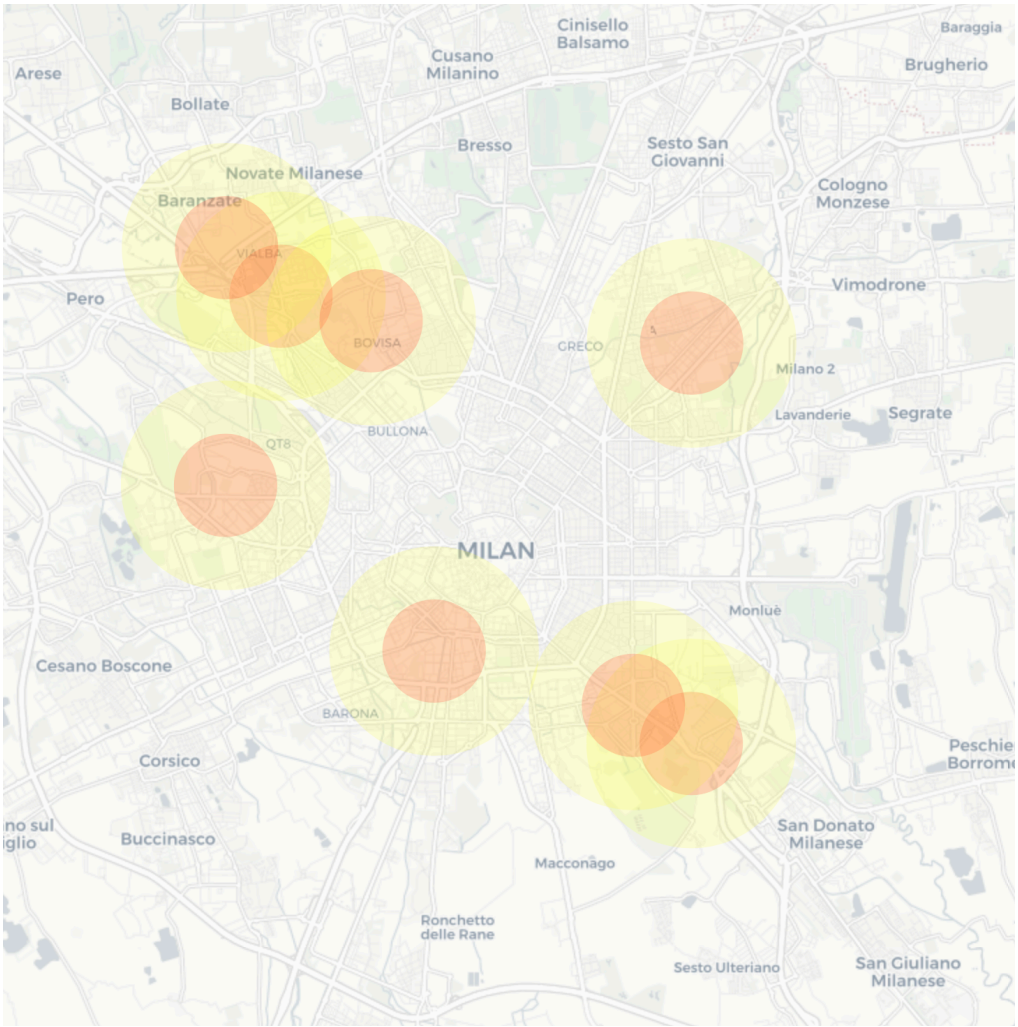
Output List

In [414]:
```
1 blackList = ["Quarto Oggiaro", "Roserio", "viale Padova", "Bovisa", "Rogored", "Barona", "Corvetto", "San Siro",
2 dangerousZones = [getLoc(i) for i in blackList]
3 dangerousZones
```

```
Quarto Oggiaro was succesfully added

Roserio was succesfully added

viale Padova was succesfully added

Bovisa was succesfully added

Rogored was succesfully added

Barona was succesfully added

Corvetto was succesfully added

San Siro was succesfully added

Via Gola was succesfully added
```

Out[414]:
```
[[45.51072300000001, 9.137585399999999],
 [45.51908539999999, 9.124203399999999],
 [45.5026549, 9.238589],
 [45.5065037, 9.1597554],
 [45.433589, 9.2384878],
 [32.9390049, -116.8736109],
 [45.4401682, 9.2242604],
 [45.4781236, 9.123961999999999],
 [45.4495399, 9.1752465]]
```

Exploratory data analysis



I added circle markers with 1 km (red) and 2 km (yellow) radius for each disadvantaged district. We can see that central, northern and east parts of the city are mainly safe.

3. Vegetation concentration

Source

https://dati.comune.milano.it/dataset/ds339-territorioambiente-aree-verdi-zona-superficie-2014

https://en.wikipedia.org/wiki/Municipalities_of_Milan

Method of gaining

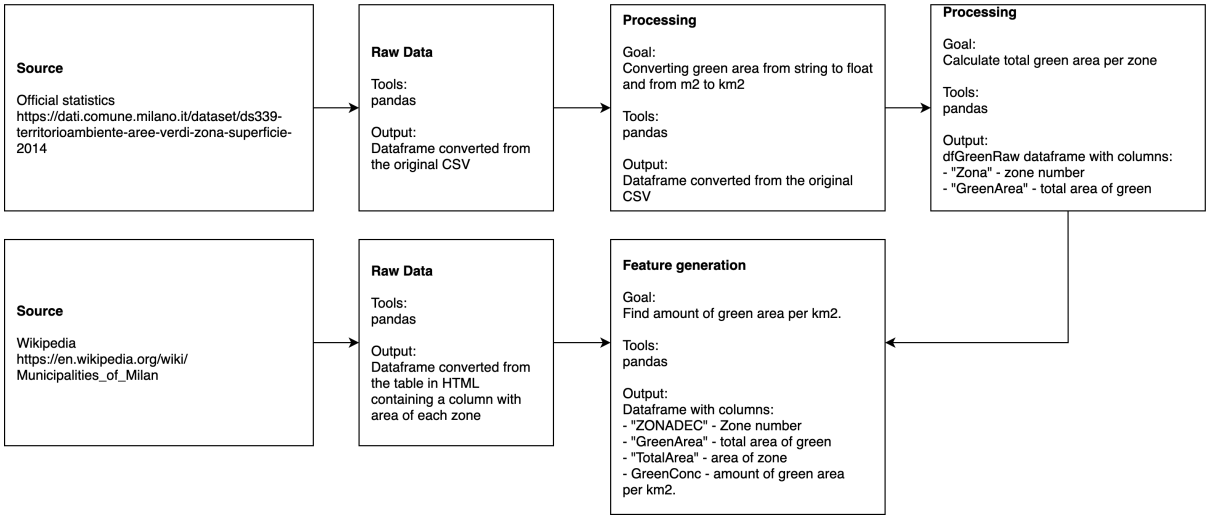
Downloading CSV file and extracting table from HTML using pandas.

Raw data contents

This link contains CSV file which consists of records representing green areas in Milan such as parks, gardens, etc., their locations and area. I transformed it to a dataframe.

	Zona	Area	ID Localita	Nome localita	Descrizione	Tipo	Classificazione	Classificazione ISTAT	AFFIDATARIO	Superficie totale in mq
0	1	1	1	piazza Castello - Minghetti	aiuola spartitraffico	Filare	Verde di arredo stradale	4 - Aree di arredo urbano	Appaltatore Servizio Manutenzione	573,07
1	1	2	2	piazza Paolo VI	area verde	Parco	Giardino	1 - Verde attrezzato	Appaltatore Servizio Manutenzione	1313,85
2	1	3	3	via San Simpliciano	filare alberato - parcheggio	Filare	Filare alberato	4 - Aree di arredo urbano	Appaltatore Servizio Manutenzione	2,82
3	1	4	4	via De Marchi Marco	filare alberato	Filare	Filare alberato	4 - Aree di arredo urbano	Appaltatore Servizio Manutenzione	212,28
4	1	5	5	via Croce Rossa	filare alberato	Filare	Filare alberato	4 - Aree di arredo urbano	Appaltatore Servizio Manutenzione	6,47

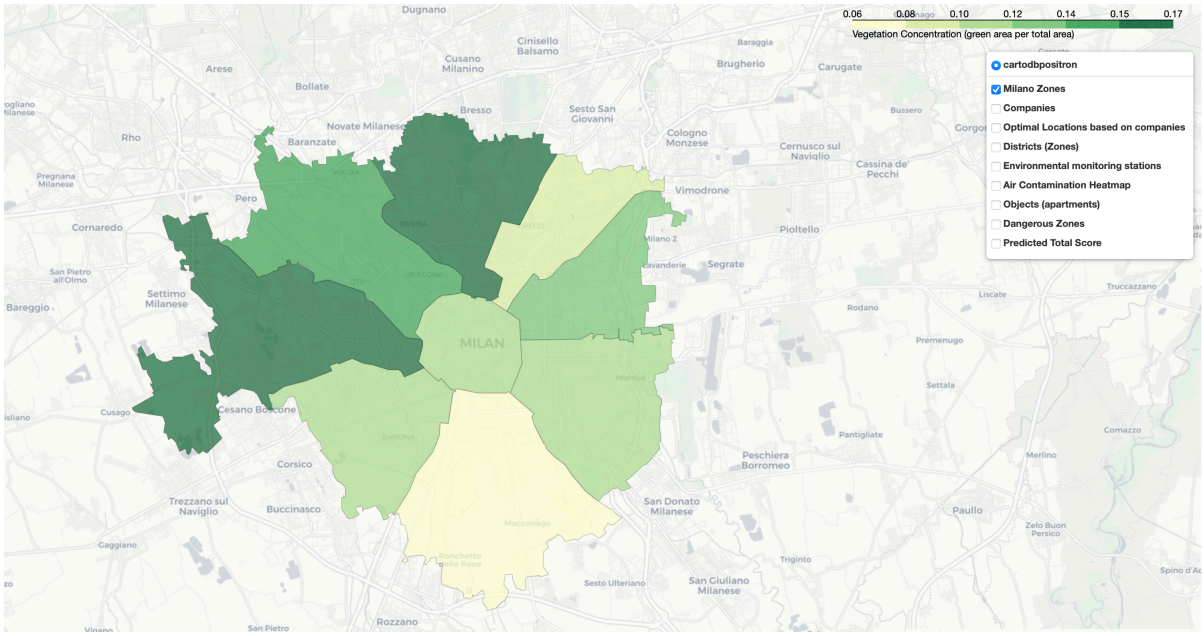
Workflow



Output Dataframe

	ZONADEC	GreenArea	TotalArea	GreenConc
0	1	1.108470	9.67	0.114630
1	2	1.067198	12.58	0.084833
2	3	1.732050	14.23	0.121718
3	4	2.133899	20.95	0.101857
4	5	1.745979	29.87	0.058453
5	6	2.125662	18.28	0.116283
6	7	5.294581	31.34	0.168940
7	8	3.549903	23.72	0.149659
8	9	3.677692	21.12	0.174133

Exploratory data analysis



We can see that northwest part of the city is pretty green.

4. Air Quality

Source

https://dati.comune.milano.it/dataset/ccf8b61d-728f-46e7-bee9-e685c7b6cd35/resource/88c1e729-420e-433f-9397-875b54aa471d/download/qaria_datoariagiornostazione_2020-11-07.csv

Method of gaining

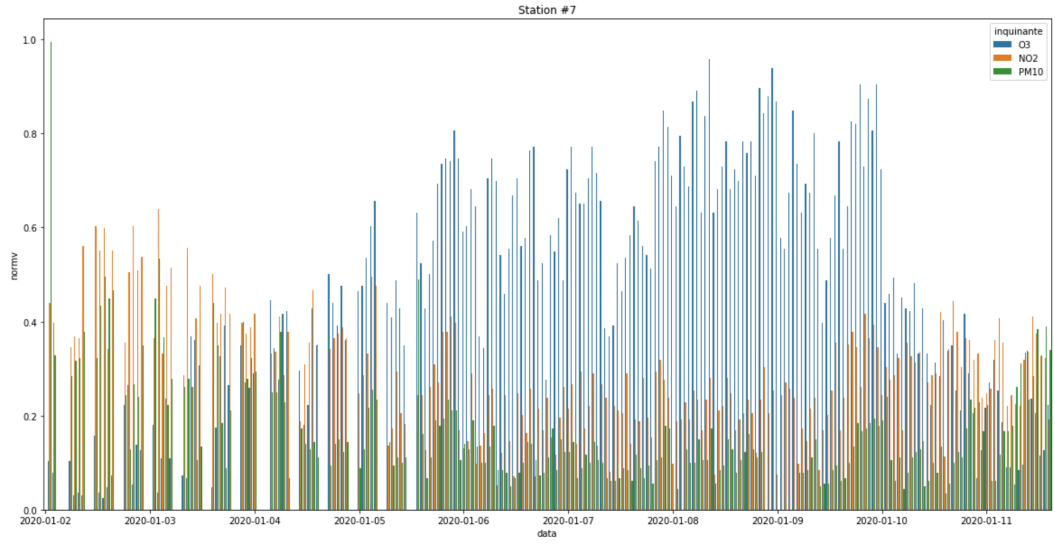
Downloading CSV file

Raw data contents

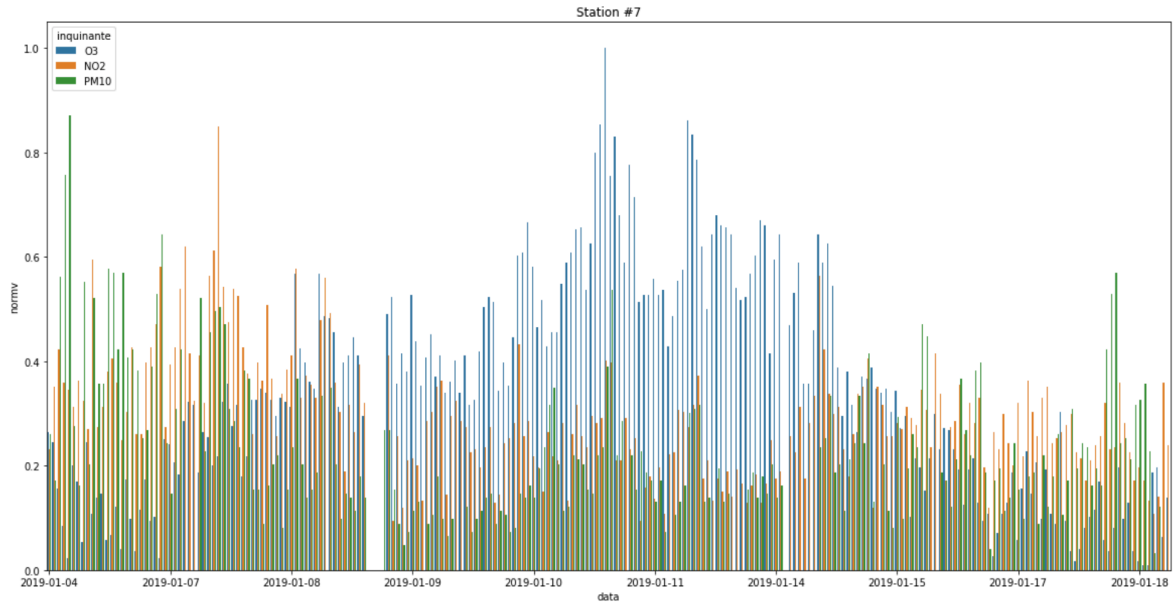
This link contains CSV file which consists of official contamination records for several monitoring stations for each day of the year.

	stazione_id	data	inquinante	valore
4127	7	2020/03/25	O3	81.0
2932	2	2020/05/29	SO2	16.0
1925	6	2020/07/23	NO2	57.0
93	4	2020/11/11	NO2	75.0
1736	6	2020/08/03	CO_8h	0.9
2801	6	2020/06/08	C6H6	0.5
5579	4	2020/01/10	C6H6	5.4
743	4	2020/10/07	NO2	80.0
5340	6	2020/01/23	PM10	64.0
655	7	2020/10/12	NO2	24.0

The chart below shows the values for one of the stations. We can see that the variation during the year is significant probably due to the seasonality.

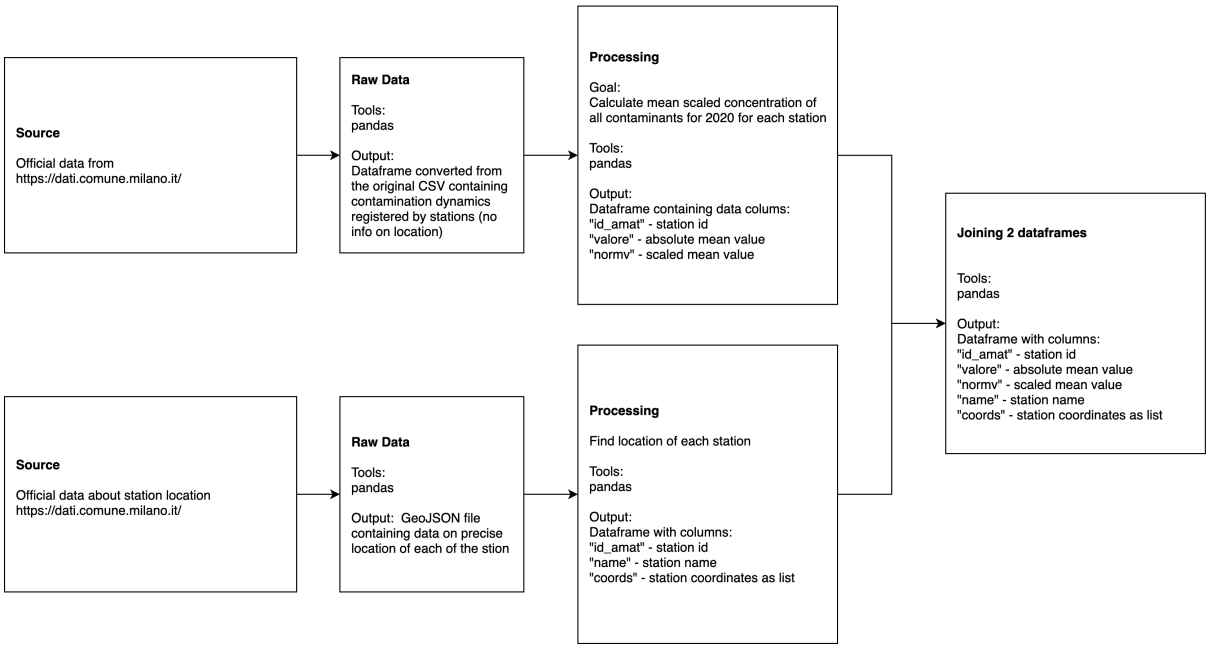


Here is the chart for the same station but for the last year.



We can see that in autumn there's an increase of O3 concentration, whereas in winter the values of NO2 and PM10 (particles) is slightly higher. But I am interested not in absolute values but how median concentrations for all contaminants varies for different stations.

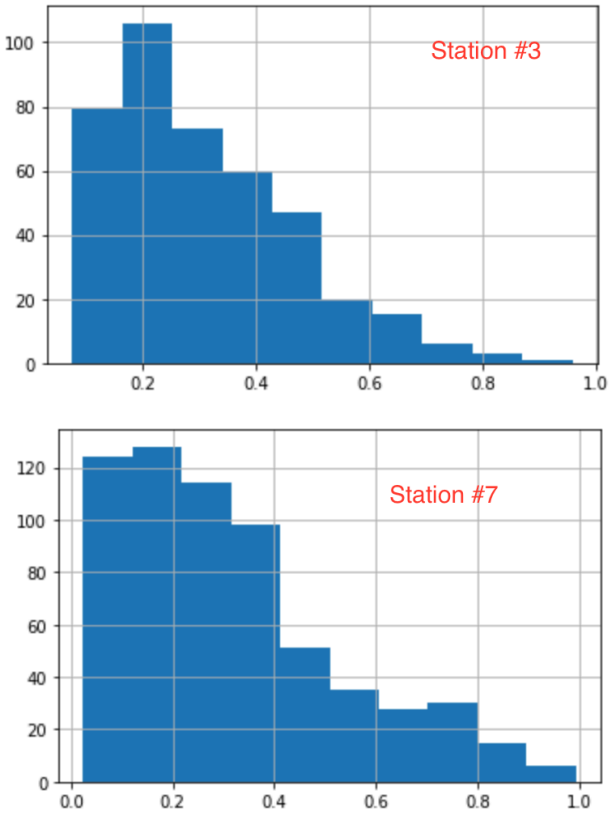
Workflow



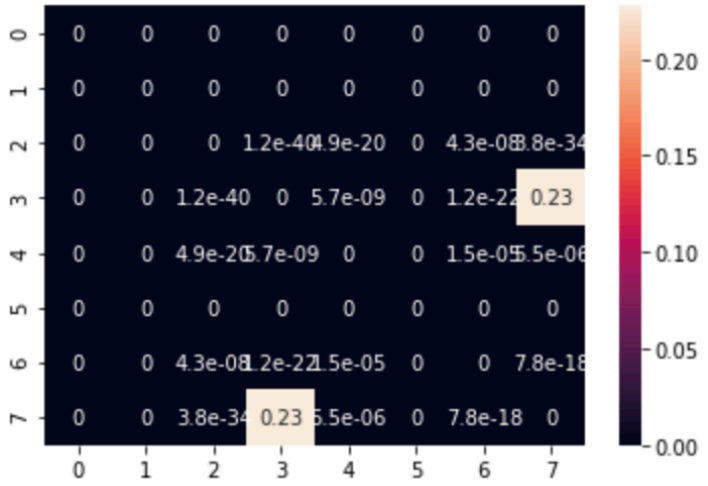
Output Dataframe

	id_amat	valore	normv	name	coords
0	2	12.0	0.150000	via Pascal *	[9.23478031158447, 45.4740982055664]
1	3	2.2	0.280812	viale Liguria	[9.16944026947021, 45.4441986083984]
2	4	8.0	0.218750	viale Marche	[9.19083976745605, 45.4962997436523]
3	6	14.0	0.196262	via Senato *	[9.19791984558105, 45.4705009460449]
4	7	52.0	0.261682	Verziere	[9.19534015655518, 45.4635009765625]

Here we can see that 2 of the stations have close values of norm variable. Here are the distributions for stations 3 and 7.



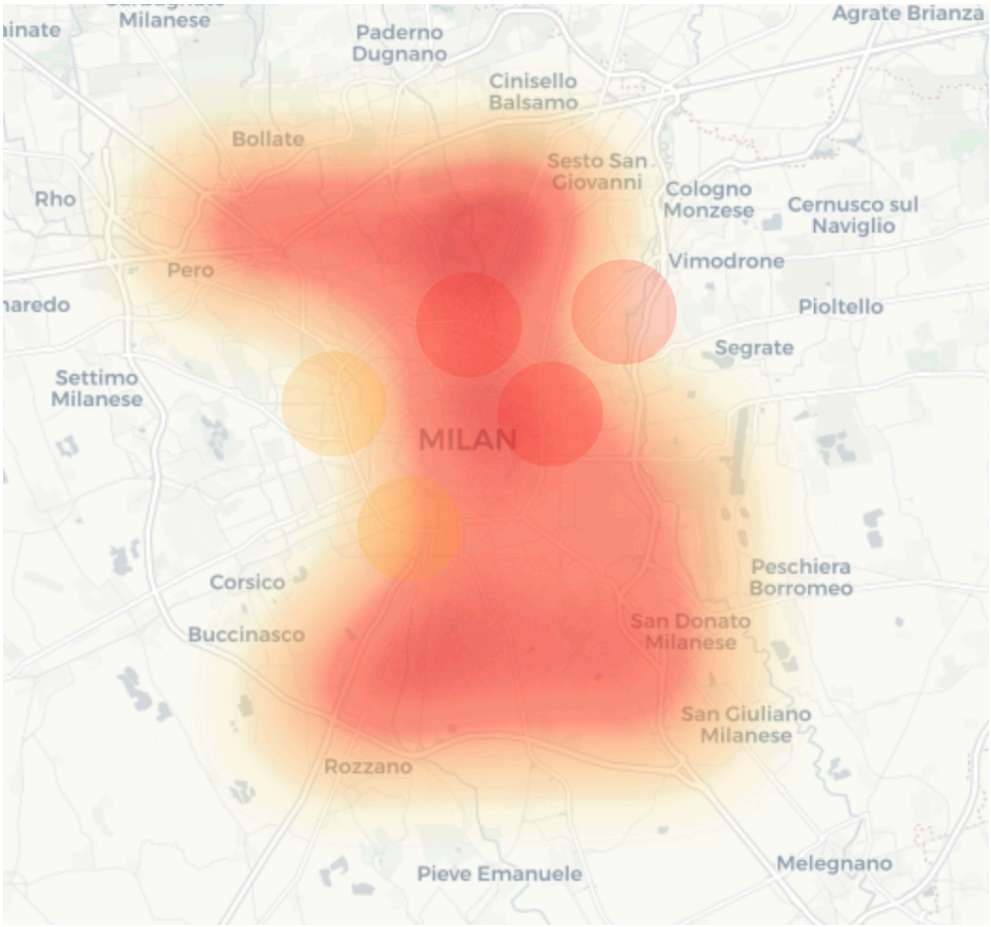
I decided to conduct a pairwise Kruskal-Wallis test (similar to ANOVA but for medians) to find out if the difference is significant between the stations. Below is the heat map of the obtained p-values. Some of the stations are not active and absent in the dastaset. It can be concluded that stations 3 and 7 do not have significantly different values of normv.



Model and visualization

I used a linear regression model applied to polynomial features with degree of 3 to predict normv values across Milan.

Here is the heat map based on this model. Cirlces on the map represent the monitoring stations.



I am sure that the results are not the most accurate for the outer parts of the city as the stations are located primarily in the center.

6. Apartments prices and locations

Source

One of the most popular real estate resources in Italy

https://www.immobiliare.it/

Method of gaining

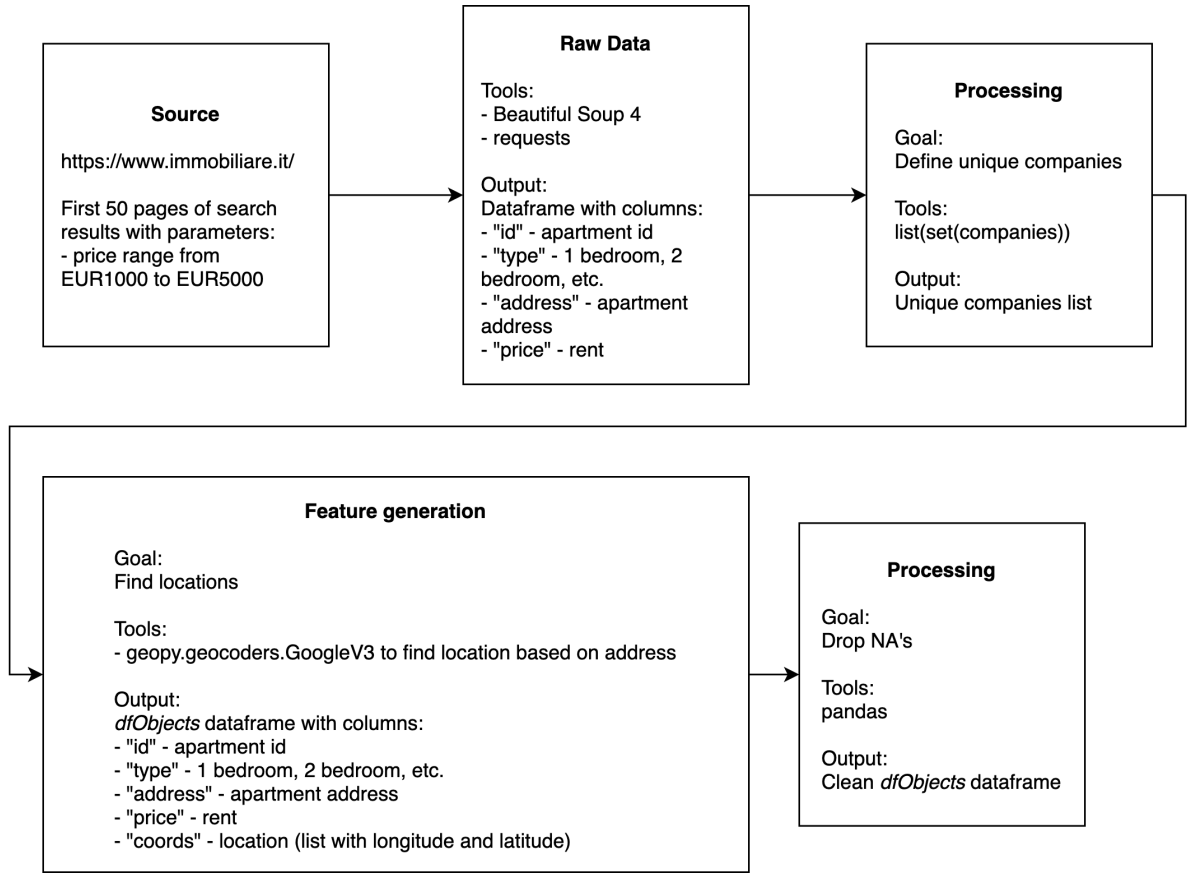
Web scraping using Beautiful Soup and requests libraries

Raw data contents

Using a dictionary I have iterated through 50 pages and formed the data frame below:

	id	type	address	price
0	link_ad_83897872	Monolocale	piazzale Biancamano 2, Moscova, Milano	1000
1	link_ad_83535941	Appartamento	corso Magenta, San Vittore, Milano	5000
2	link_ad_79920811	Bilocale	corso Indipendenza 18, Indipendenza, Milano	1350
3	link_ad_83588673	Appartamento	via Archimede, Plebisciti - Susa, Milano	3200
4	link_ad_81088197	Trilocale	via Sansovino 3, Città Studi, Milano	1550

Workflow



Output Dataframe

	id	type	address	price	coords
0	link_ad_83897872	Monolocale	piazzale Biancamano 2, Moscova, Milano	1000	[45.47823899999999, 9.1819484]
1	link_ad_83535941	Appartamento	corso Magenta, San Vittore, Milano	5000	[45.4657096, 9.1717594]
2	link_ad_79920811	Bilocale	corso Indipendenza 18, Indipendenza, Milano	1350	[45.467451, 9.214442499999999]
3	link_ad_83588673	Appartamento	via Archimede, Plebisciti - Susa, Milano	3200	[45.4681442, 9.2225773]
4	link_ad_81088197	Trilocale	via Sansovino 3, Città Studi, Milano	1550	[45.4793396, 9.217790599999999]

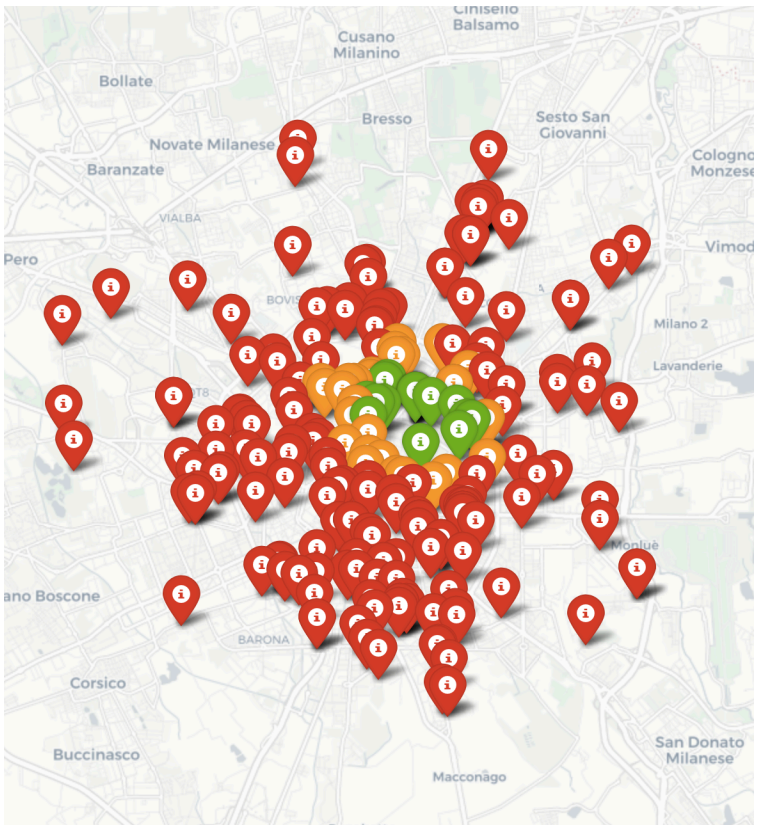
7. Final Dataframe

Then I merged all the data above into one dataframe. I used created model and locations of the apartments to predict corresponding values of contamination, defined zone for each apartment based on GeoJSON data on districts polygons to get vegetation concentration. I calculated distances to median location of companies, to the dangerous zones.

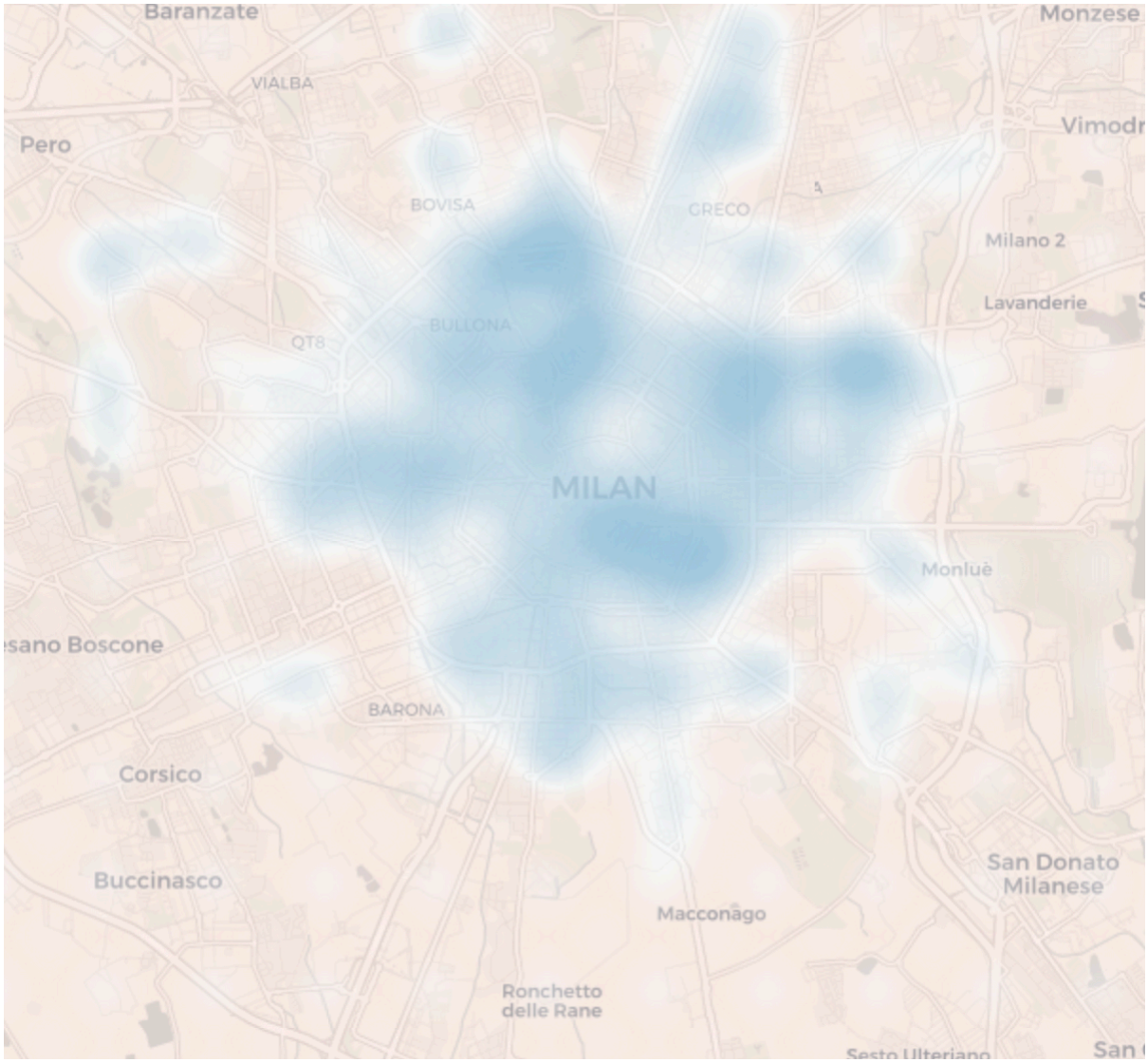
	id	type	address	price	coords	contamination	distanceToDangerZone	distanceToMedianJobLocation	Zone	GreenConc	lt	lg	totalScore	priceScaled	priceLogScaled
0	link_ad_83897872	Monolocale	piazzale Biancamano 2, Moscova, Milano	1000	[45.47823899999999, 9.1819484]	0.134814	3.233713	1.007914	1.0	0.114630	45.478239	9.181948	10.916964	0.111111	0.301030
1	link_ad_83535941	Appartamento	corso Magenta, San Vittore, Milano	5000	[45.4657096, 9.1717594]	0.102206	1.818443	1.543893	1.0	0.114630	45.465710	9.171759	0.434870	1.000000	1.000000
2	link_ad_79920811	Bilocale	corso Indipendenza 18, Indipendenza, Milano	1350	[45.467451, 9.214442499999999]	0.273175	3.128875	1.950986	4.0	0.101857	45.467451	9.214442	4.843964	0.188889	0.431364
3	link_ad_83588673	Appartamento	via Archimede, Plebisciti - Susa, Milano	3200	[45.4681442, 9.2225773]	0.272856	3.113558	2.563470	3.0	0.121718	45.468144	9.222577	1.394753	0.600000	0.806180
4	link_ad_81088197	Trilocale	via Sansovino 3, Città Studi, Milano	1550	[45.4793396, 9.217790599999999]	0.112334	3.057725	2.352282	3.0	0.121718	45.479340	9.217791	3.340804	0.233333	0.491362

I also log scaled prices for the further model creation for price prediction.

As a result I have an interactive map (attached map2.html) with objects which are colored red if their totalScore below 50% of maximum total score and green if it's more than 75% of maximum total score.



I implemented a regression model to predict a total score base on location similar to air pollution model. Here is the result. Blue is better, red is worse.



I also made a draft web application based on Plotly Dash (see the visualization notebook) but it needs some more time to be implemented.

8. Predicting the price

In the last minute I decided to check whether my custom total score can be used to predict the prices. I used Random Forest regressor with features of longitude, latitude, apartment type (dummy variables) and total score. Price was the target variable. I managed to get mean absolute error of less than 200 EUR which doesn't seem bad. Feature important analysis showed that totalScore feature is the most important one.

{'Mansarda': 2.423880935560713e-05,
 'Open': 3.592507503023473e-05,
 'Terratetto': 0.00039482672550974487,
 'Villa': 0.002844627721024826,
 'Loft': 0.0038345084523889535,
 'Trilocale': 0.014185815083260632,
 'Quadrilocale': 0.017961606811265088,
 'Attico': 0.02007921443385385,
 'Bilocale': 0.05108951595119663,
 'Monolocale': 0.05670068718435202,
 'lg': 0.1364950945029547,
 'Appartamento': 0.14049890796886821,
 'lt': 0.22550031447858324,
 'totalScore': 0.330354716802356 3}