Ocean Data Challenge :: Air Quality in Catalonia

1. Challenge description
2. Data structuring
    1. What is the observation?
    2. Visual examples of time series data
    3. Pollutants segmentation
3. Data exploration
    1. Pollutants yearly trends
    2. Pollutants monthly seasonality
    3. Pollutants daily seasonality
    4. Pollutants day of week seasonality
    5. Pollutants concentration in urban, suburban and rural areas
    6. Relationship between altitude and pollutants concentration
    7. Countries rankings by the level of pollution
4. Predicting the monthly average of O3 values
    1. Evaluation
    2. Final predictions
5. Predicting the hourly average of O3 values.
    1. Evaluation
    2. Final predictions
6. Recommendations

ocean

1. Challenge description

Global Analysis (presented in the report) :: 40 points

• Analyze the evolution of pollution in Catalonia over time to determine the best/worst hours and best/worst months of the year in terms of pollution, and explain the periodicity of the rate of certain pollutants in the air. (10 points)
• Analyze the relationship between altitude and concentration of particles in the air, and present your conclusions in graphical form. (10 points)
• Analyze the concentration of pollutants in urban, suburban and rural areas, and present your conclusion in graphical form. (10 points)
• Rank the cities in the dataset according to their level of pollution, and create best-5 and worst-5 lists. (10 points)

Algorithms (published on the OM using compute-to-data feature) :: 40 points

• Build and publish an algorithm to predict the average concentration of one pollutant of your choice per month for the next 24 months - on average for all stations. (20 points)
• Build and publish an algorithm to predict the concentration of one pollutant of your choice for each hour of the day from February 15 to 28 - on average for all stations. (20 points)

Report (published on the Ocean Market for free public consumption) :: 20 points

• Introduction - provide background information on the challenge and the problem being solved. (2 points)
• Methodology - describe the steps taken to analyze the evolution of pollutants and predict their concentration. (5 points)
• Results - present the findings from the analysis of the evolution of pollutants and the performance of the predictive model. (5 points)
• Observations and Conclusions - summarize the key observations and draw conclusions based on the results obtained. (5 points)
• Limitations and Recommendations - identify limitations of the challenge and provide recommendations for future competitions. (3 points)

2.1. Data structuring. What is the observation?

• In the input dataset we have 3,106,374 rows of data and 40 features. Each row have 24 hours daily measurements of one pollutant for one station. The first 3 rows of the dataset:

	CODI EOI	CODI INE	CODI COMARCA	NOM ESTACIO	MUNICIPI	NOM COMARCA	ALTITUD	TIPUS ESTACIO	AREA URBANA	LATITUD	LONGITUD	GEOREFERENCIA	MAGNITUD	CONTAMINANT	UNITATS	DATA	01h
0	43148003	43148	36	Tarragona (Bonavista)	Tarragona	Tarragonès	39	industrial	suburban	41.12	1.19	POINT (1.1919986 41.11591)	10	PM10	µg/m3	25/01/2023	11.00
1	8137001	8137	41	Montseny (La Castanya)	Montseny	Vallès Oriental	693	background	rural	41.78	2.36	POINT (2.358002 41.77928)	12	NOX	µg/m3	25/01/2023	2.00
2	8124009	8124	41	Mollet del Vallès	Mollet del Vallès	Vallès Oriental	90	traffic	suburban	41.55	2.21	POINT (2.2120984 41.549183)	7	NO	µg/m3	25/01/2023	49.00

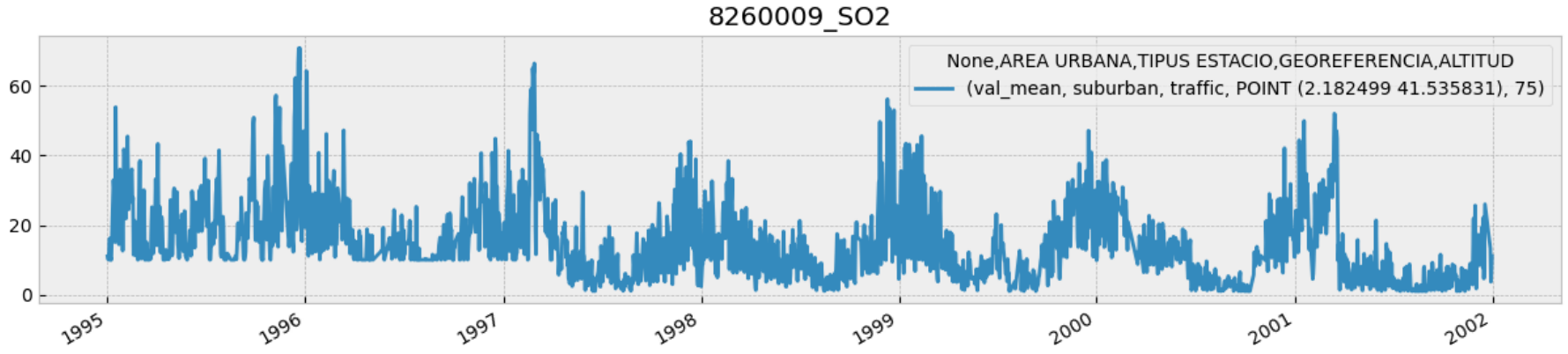
• Each station have identification code **CODI EOI** and the attributes describing it. Green line attributes are static. Blue line attributes are changing, it means that for one **CODI EOI** we could have different values.
• Each pollutant **CONTAMINANT** have unique identification code **MAGNITUD** and unique measurements unit **UNITATS**. One exception here is H2S pollutant that measured in µg/m3 and ug/m3. After visual exploration we decided that they are the same.
• For one day, one station and one pollutant we have one row of measurements. Exception here is two identical rows with different measurements. We took only first values for this case:

	CODI EOI	CODI INE	CODI COMARCA	NOM ESTACIO	MUNICIPI	NOM COMARCA	ALTITUD	TIPUS ESTACIO	AREA URBANA	LATITUD	LONGITUD	GEOREFERENCIA	MAGNITUD	CONTAMINANT	UNITATS	DATA	01h	02h
2365482	43103001	43103	36	Perafort (Puigdelfi)	Perafort	Tarragonès	97	industrial	rural	41.19	1.24	POINT (1.236701 41.193603)	65	H2S	µg/m3	02/07/2000	1.70	2.00
2365543	43103001	43103	36	Perafort (Puigdelfi)	Perafort	Tarragonès	97	industrial	rural	41.19	1.24	POINT (1.236701 41.193603)	65	H2S	µg/m3	02/07/2000	1.70	1.90

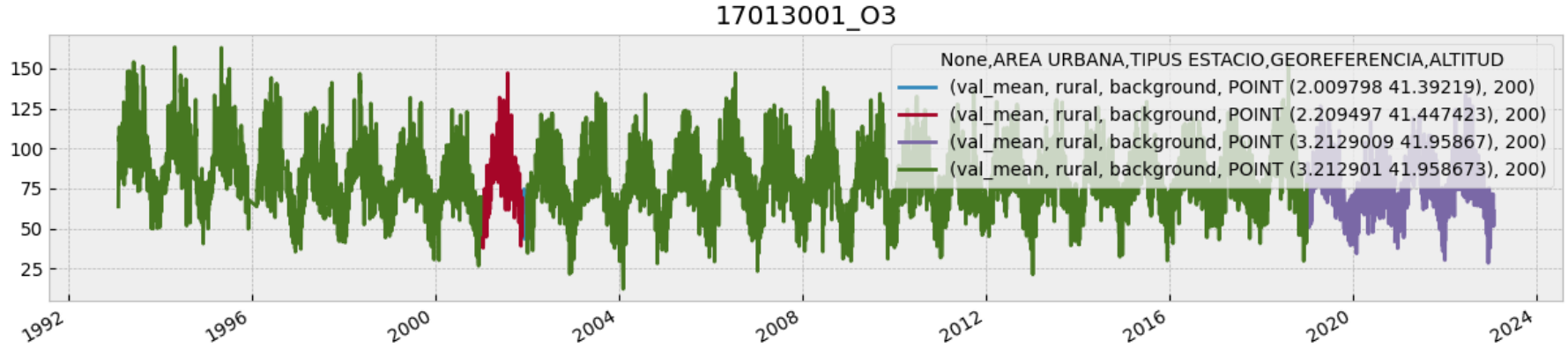
2.2. Data structuring. Visual examples

Based on the above explorations we can show the examples of time series plots for pollutants and stations.

For the station = 8260009 and SO2 pollutant the station parameters not changing over time:



For the station = 17013001 and O3 pollutant the station parameters are different:

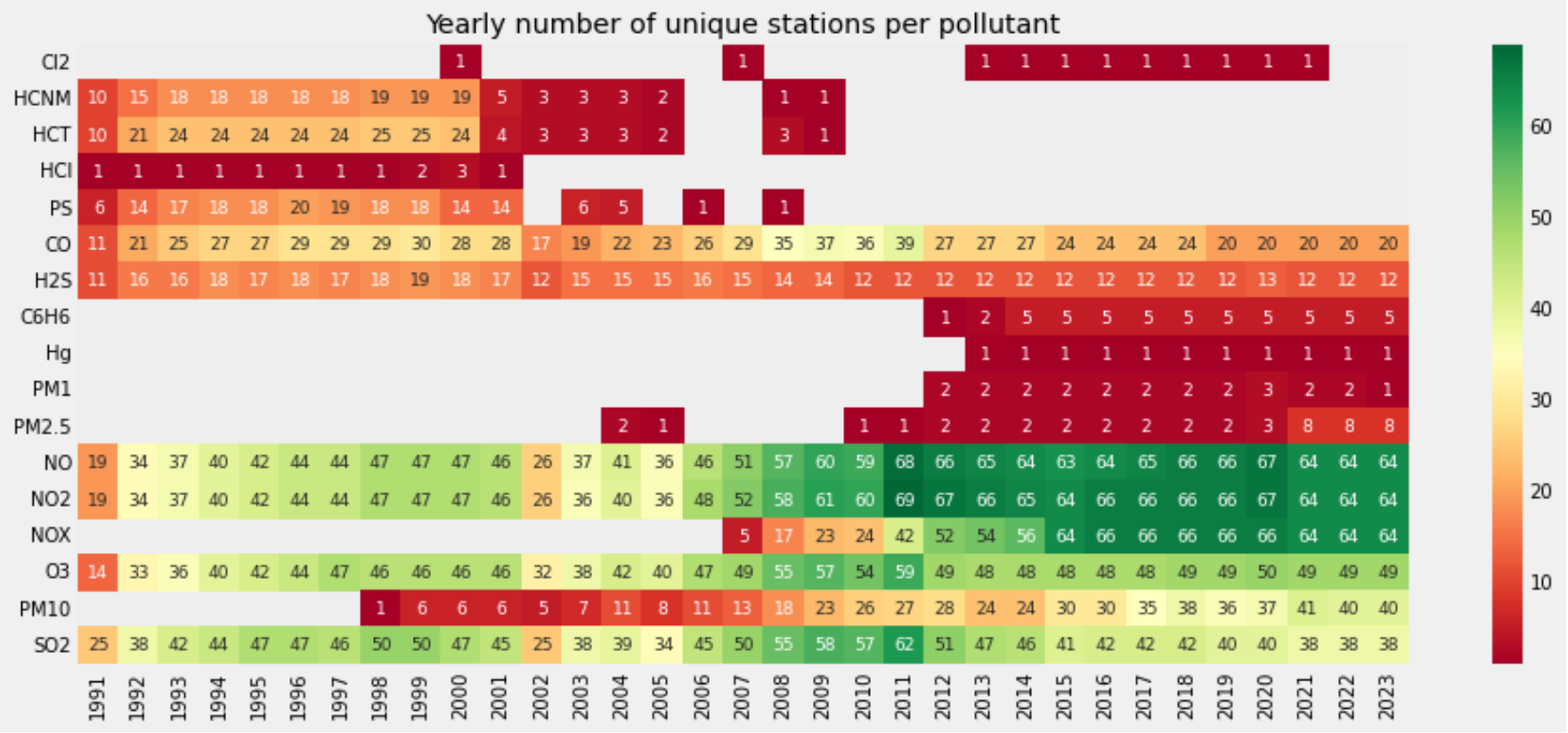


ocean

2.3. Data structuring. Pollutants segmentation

We calculated the number of stations where there were any measurements by each pollutant per year.
Based on that matrix we can create the following groups of pollutants:

- **OLD** (5) – Cl2 HCNM HCT HCl PS
- **RARE** (6) – CO H2S C6H6 Hg PM1 PM2.5
- **TOP** (6) – NO NO2 NOX O3 PM10 SO2



Further analytics we will show separately by these groups, excluding OLD group.

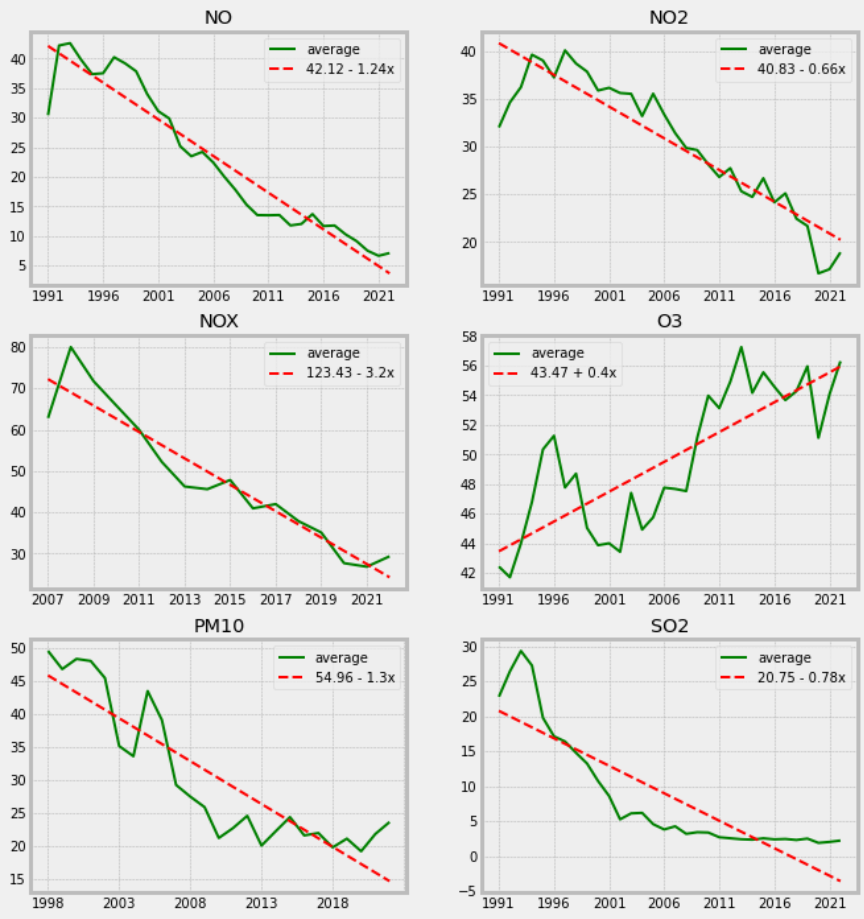
*More details you can find in the following script: *1.data_exploration.ipynb*

ocean

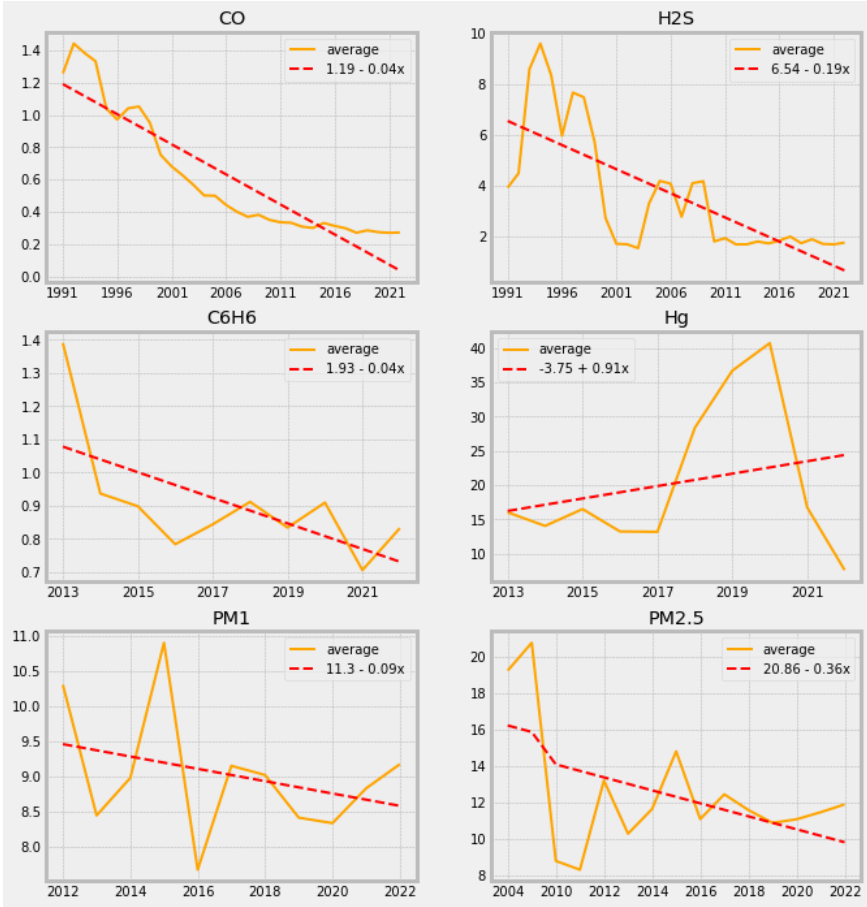
3.1. Data exploration. Pollutants yearly trends

• For all pollutants except O3 (ozone) we are observing the down trending behavior, that means the improving of air quality from year to year.

TOP segment of pollutants



RARE segment of pollutants



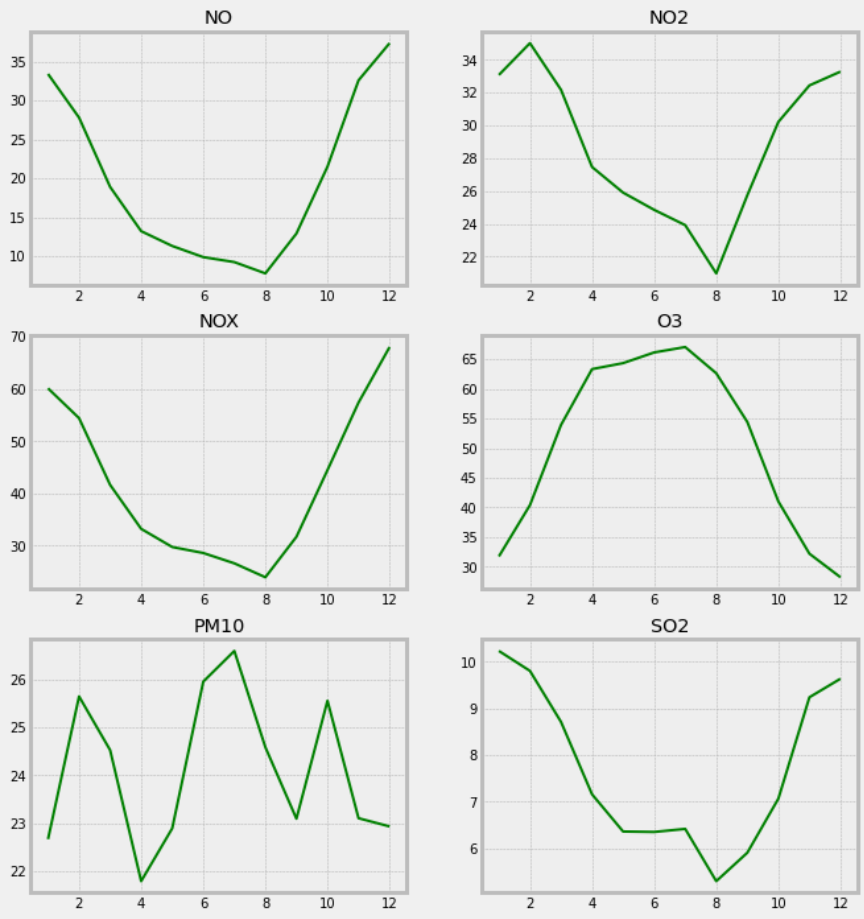
*More details you can find in the following script: *1.data_exploration.ipynb*

ocean

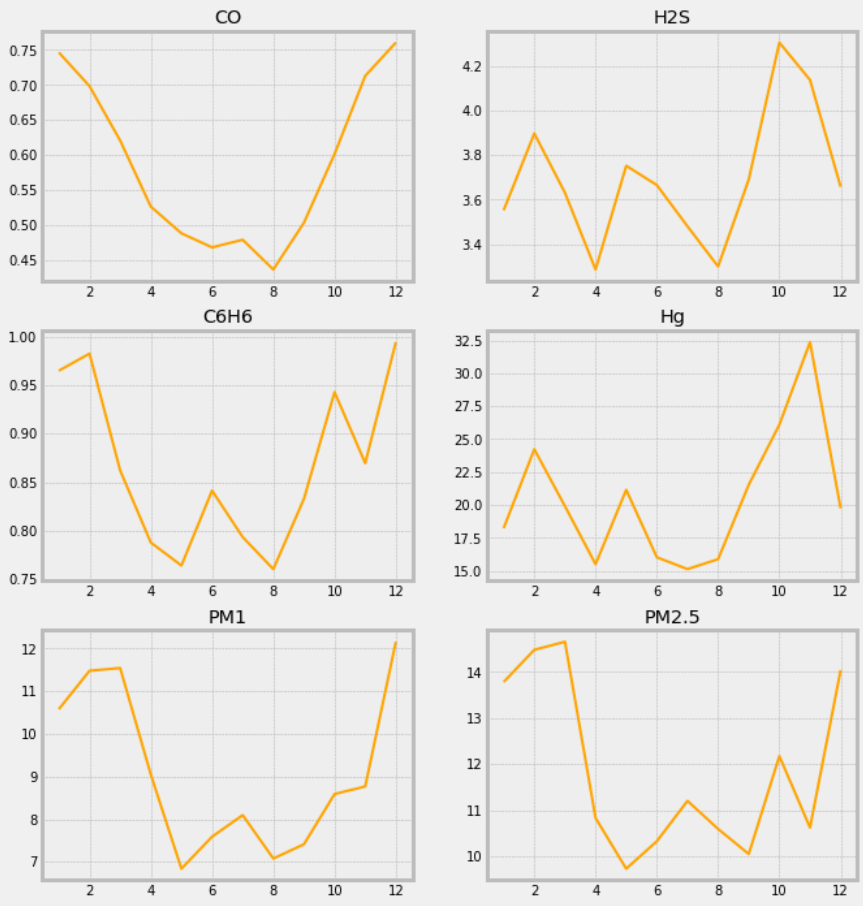
3.2. Data exploration. Pollutants monthly seasonality

• For most of pollutants we can see the picture that in the summer time (months 6,7,8) are the lowest values. The best month is August. The highest values are in the winter time.
• O3 (ozone) have the opposite dependancies – the lowest values in the winter time and highest in the summer.

TOP segment of pollutants



RARE segment of pollutants



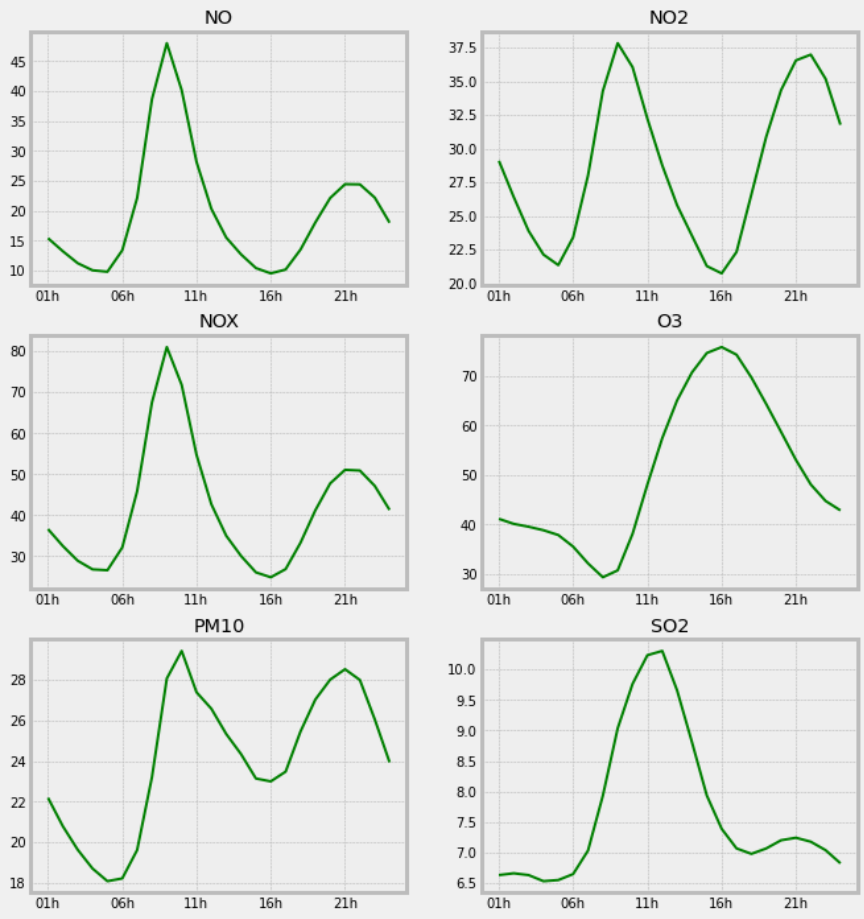
*More details you can find in the following script: *1.data_exploration.ipynb*

ocean

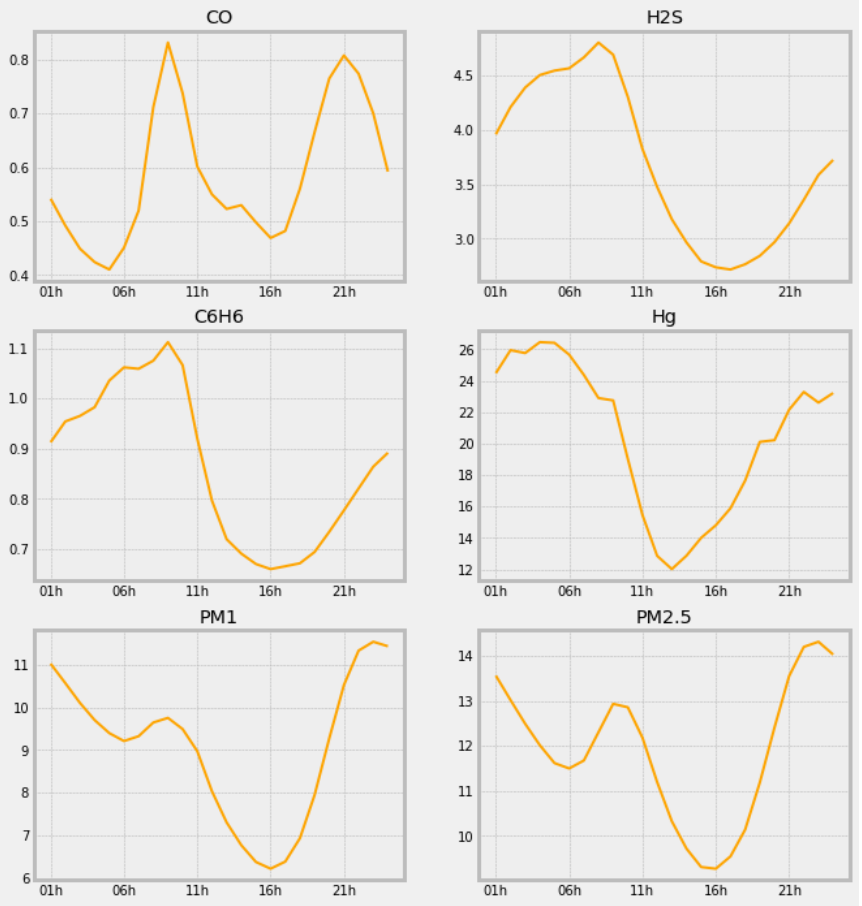
3.3. Data exploration. Pollutants daily seasonality

• For most of pollutants we can see the picture there are two peaks – in the morning time when people are going to the work and in the evening when people are coming back.
• O3 has the peak at 16:00 and SO2 at 12:00.

TOP segment of pollutants



RARE segment of pollutants



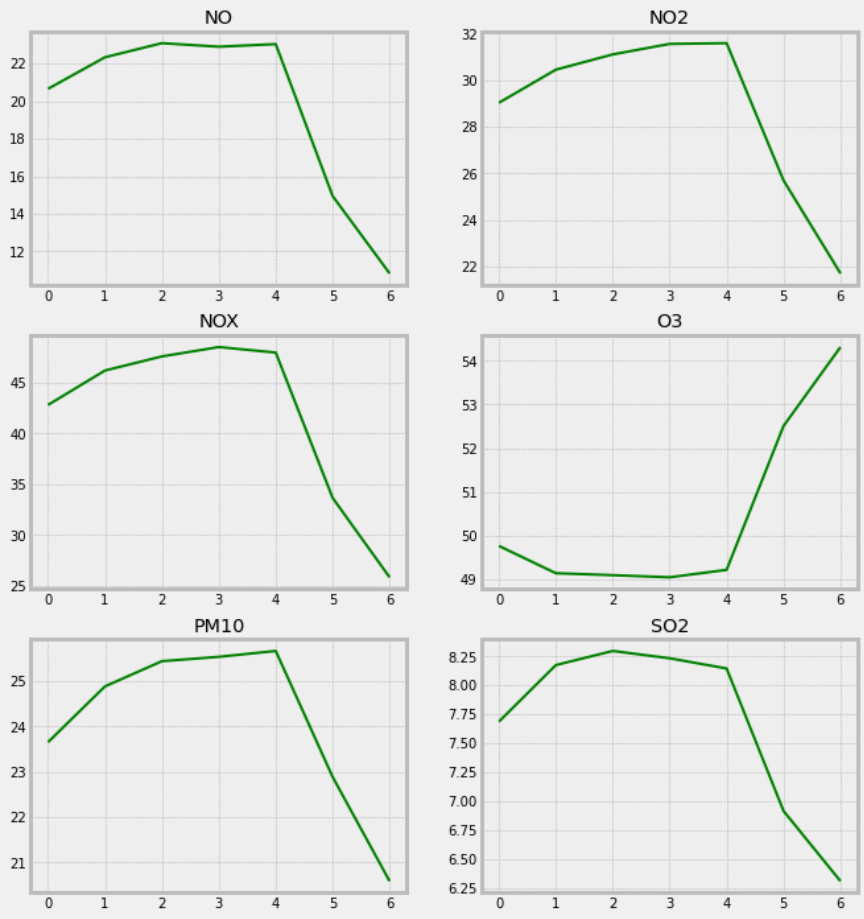
*More details you can find in the following script: *1.data_exploration.ipynb*

ocean

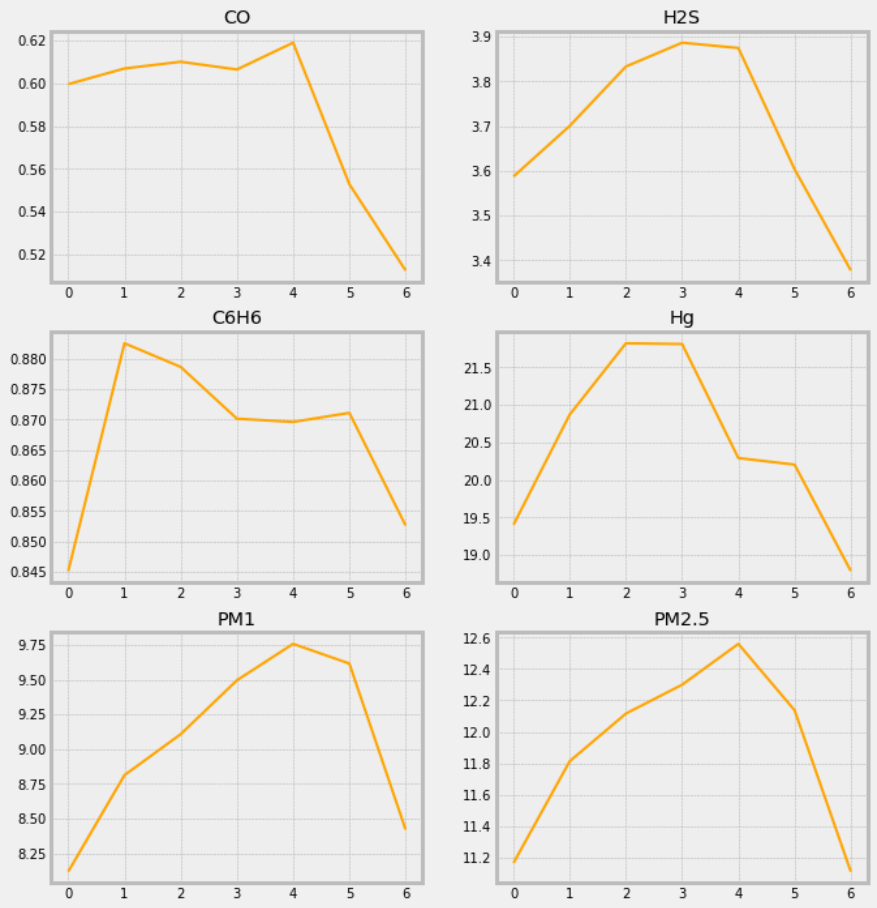
3.4. Data exploration. Pollutants day of week seasonality

• For all pollutants from the TOP segment, except O3, we are observing that on the weekend (Saturday and Sunday – 5 and 6) the air quality is better.

TOP segment of pollutants



RARE segment of pollutants



*More details you can find in the following script: *1.data_exploration.ipynb*

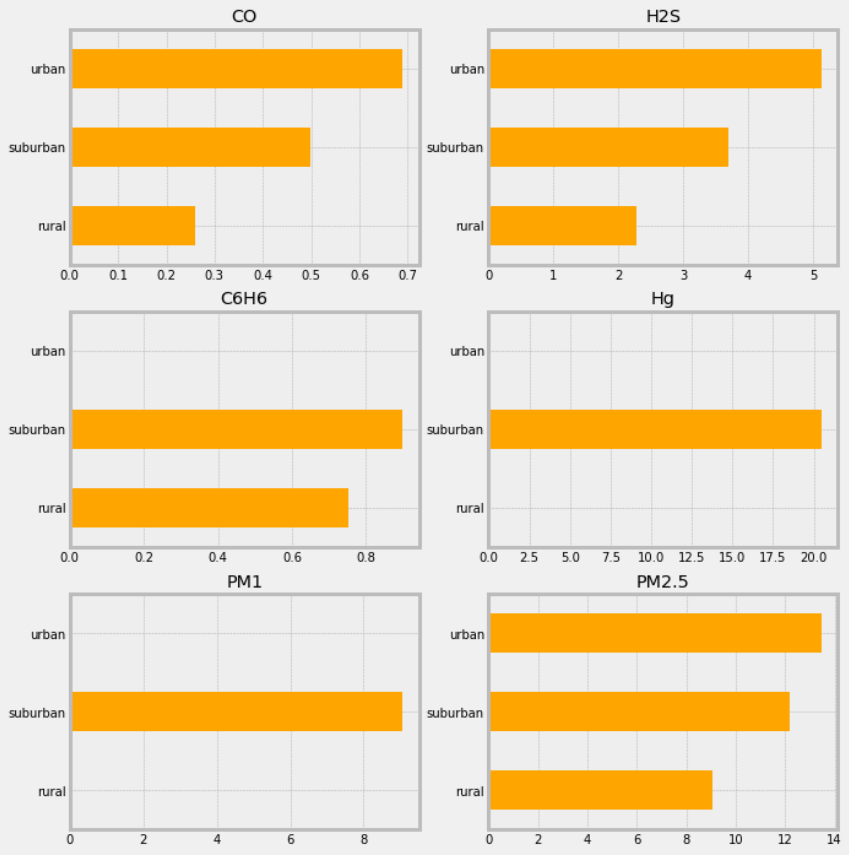
ocean

3.5. Data exploration. Pollutants in urban, suburban and rural areas

• For all pollutants except O3 and SO2 we are observing the in rural areas the air quality is better.
• Ground-level ozone (O3) pollution comes from cars, power plants, industrial boilers, refineries, and chemical plants. For this reason, levels of ground-level ozone tend to be the highest near urban centers as opposed to rural areas. But in this picture we have and opposite dependency that is surprising.

TOP segment of pollutants

RARE segment of pollutants



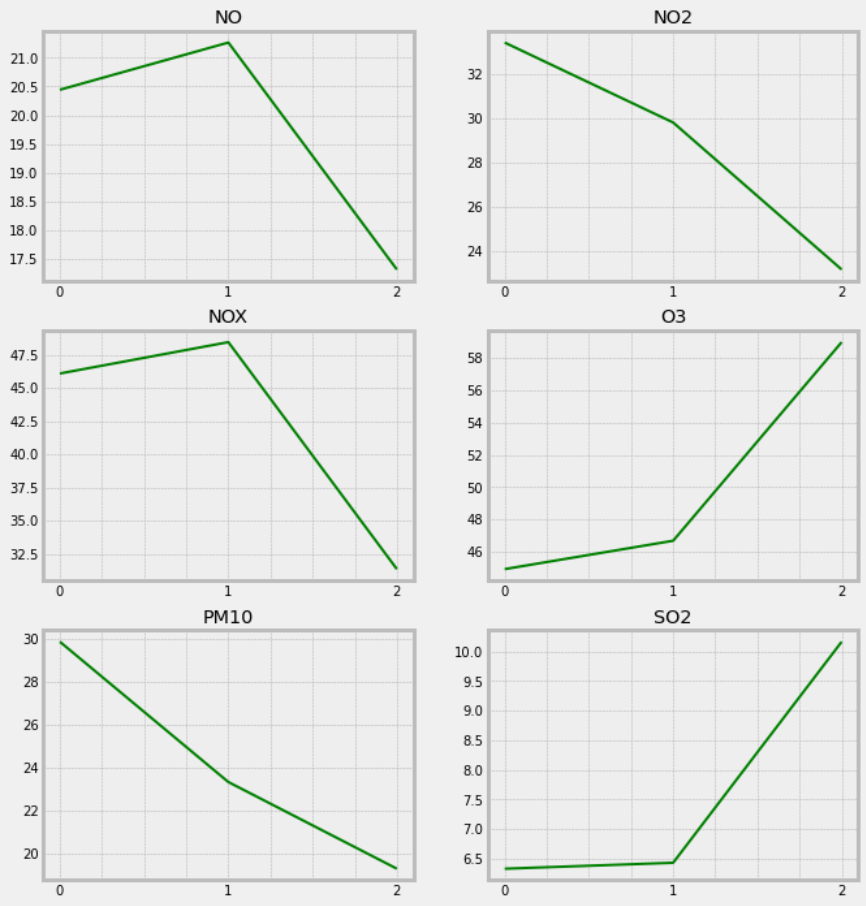
*More details you can find in the following script: *1.data_exploration.ipynb*

ocean

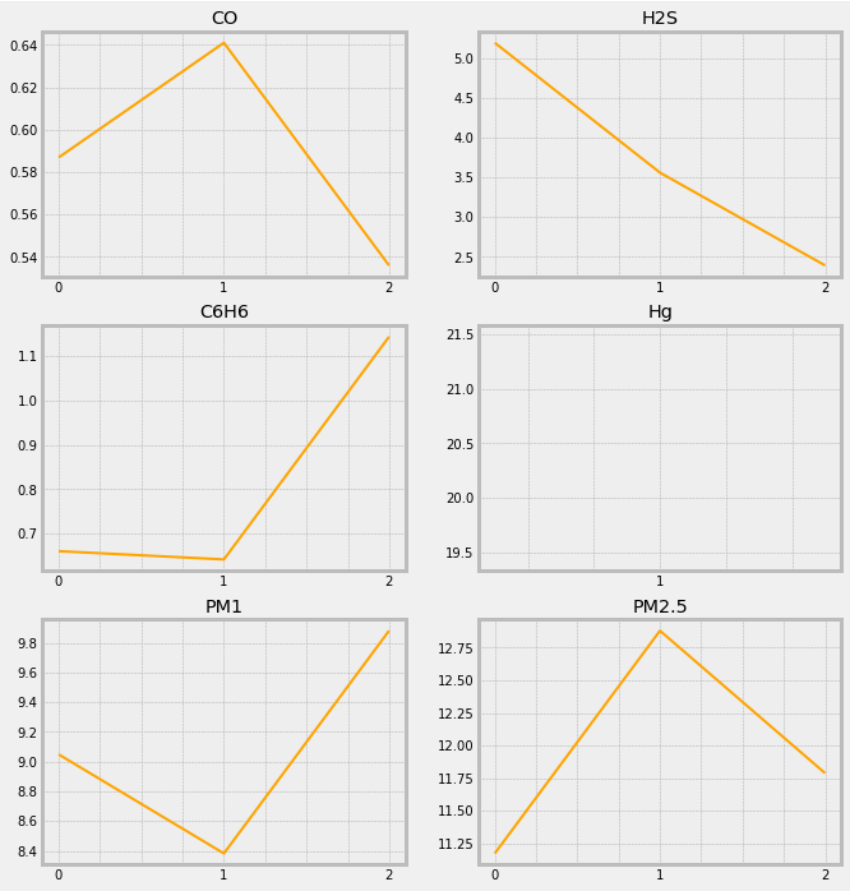
3.6. Data exploration. Relationship between altitude and pollutants concentration

- We split the altitude feature into 3 almost equally buckets (0,1,2) in each of the pollutant.
- Then we calculated the average values of every pollutant in the buckets.
- From the TOP pollutants we can see the picture that O3 and SO2 have dependency that lower altitude is better. For other pollutants the higher altitude is better.

TOP segment of pollutants



RARE segment of pollutants



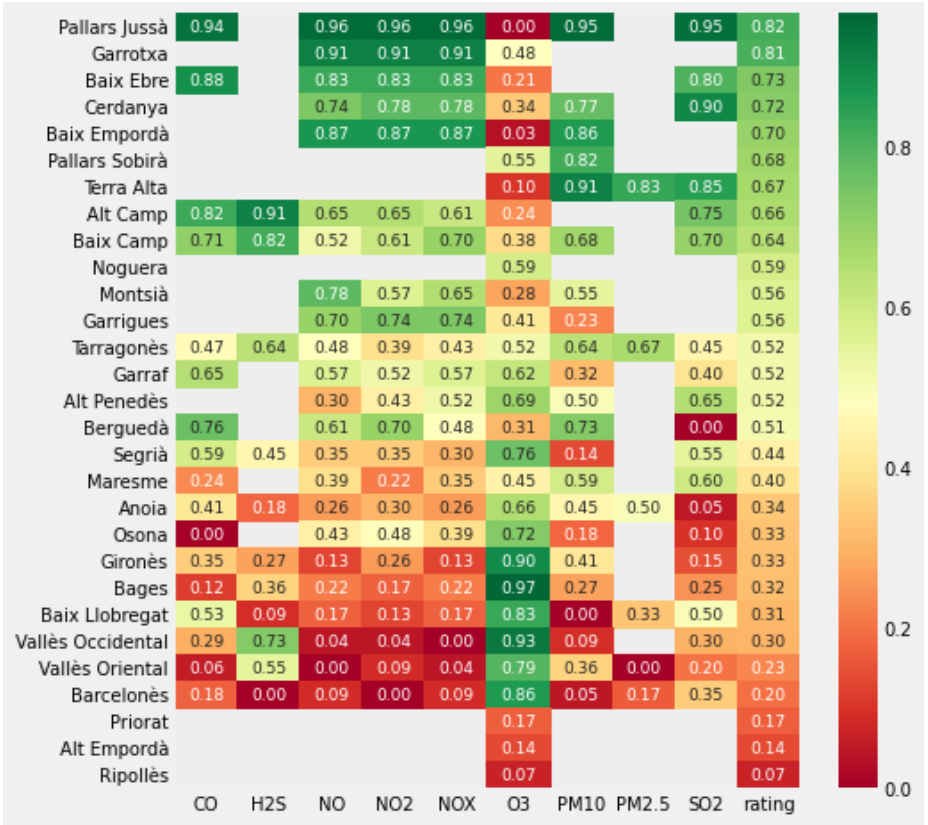
*More details you can find in the following script: *1.data_exploration.ipynb*

ocean

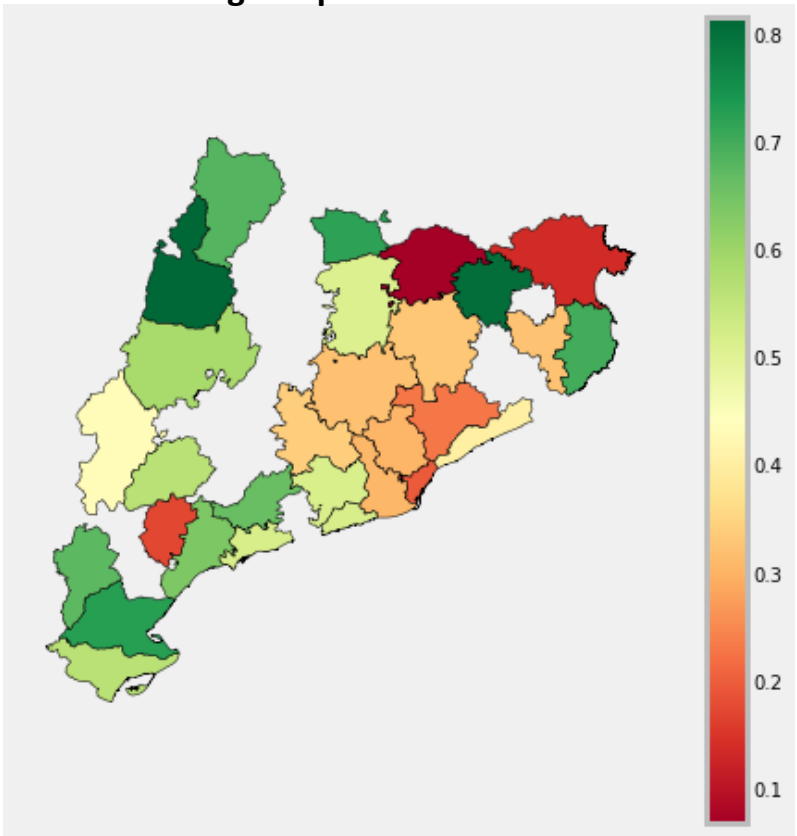
3.7. Data exploration. Countries rankings by the level of pollution (30 years)

• We are using only pollutants from the TOP and RARE segments. Also dropping pollutants with less then 5 unique countries. Here is the dropped list: C6H6, Hg, PM1.
• Calculating average pollutants values for the period and convert them to percent rankings from 0 to 1.
• Calculating the inverse: 1 –x , because if the pollution is lower then it's better.
• Rating is the average by all pollutant ratings:

Countries ratings heatmap



Countries ratings map



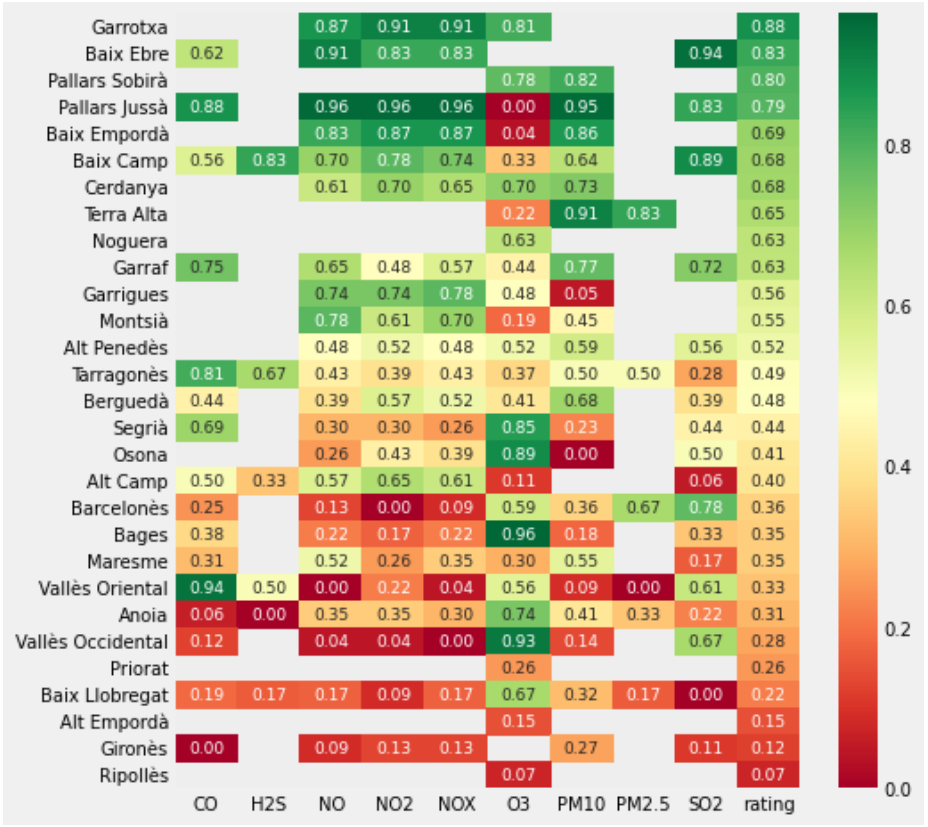
*More details you can find in the following script: *1.data_exploration.ipynb*

ocean

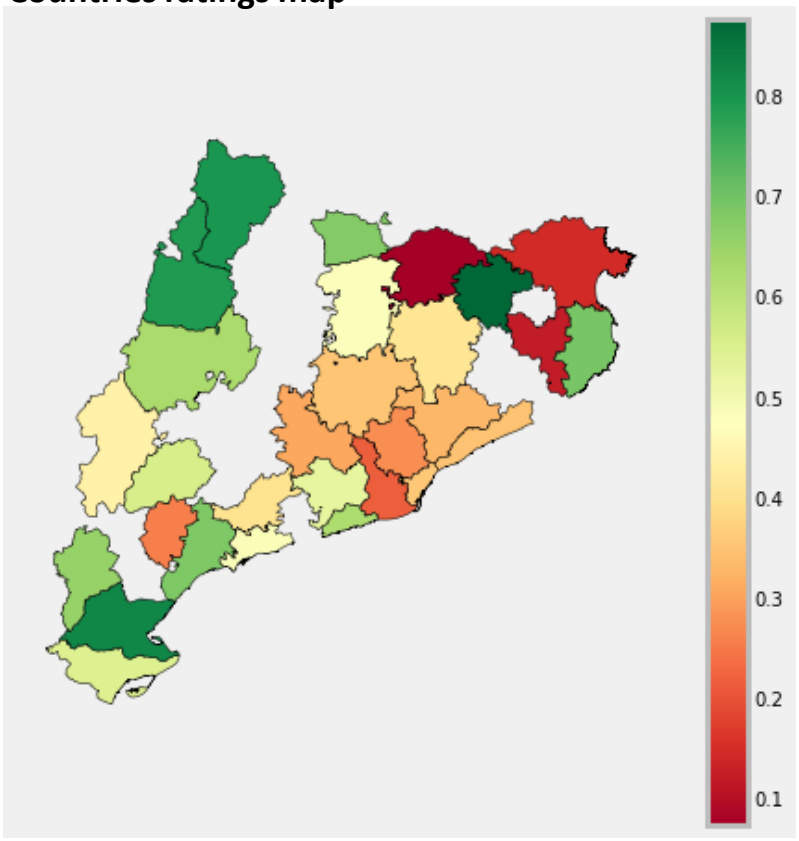
3.8. Data exploration. Countries rankings by the level of pollution (5 years)

Here is the similar picture but the calculations for the last 5 years.
We can see that for the most of the regions the air quality position stays the same except of the couple of western regions where they positions becomes lower.

Countries ratings heatmap



Countries ratings map



*More details you can find in the following script: *1.data_exploration.ipynb*

ocean

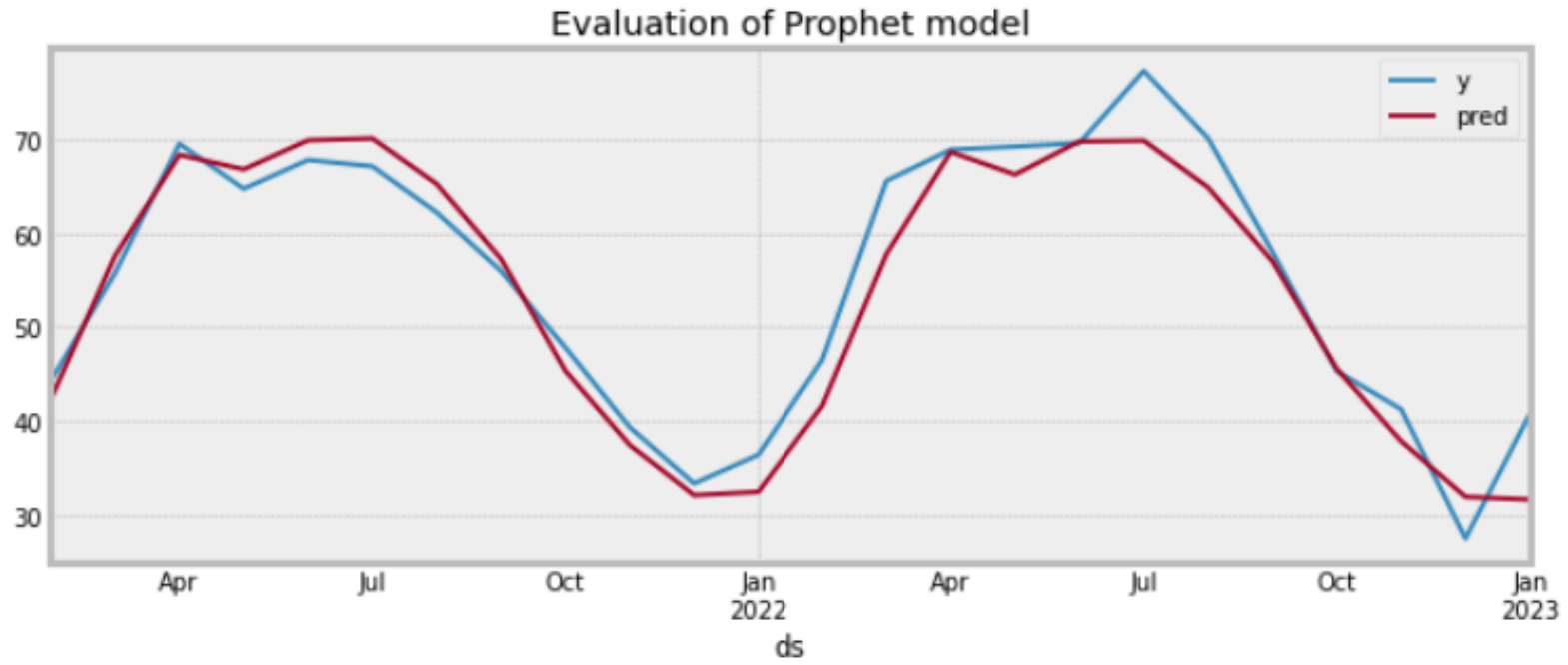
4.1. Predicting the monthly average of O3 values. Evaluation

We have used the last available 24 months for evaluation set, the data before that was used for training.
Evaluation was doing by the different metrics:

- mse – mean squared error
- mae – mean absolute error
- mape – mean absolute percentage error

The results of evaluation:

```
{'mae': 3.030144338598429,
 'mse': 14.715596518003013,
 'mape': 0.06032912725917581,
```



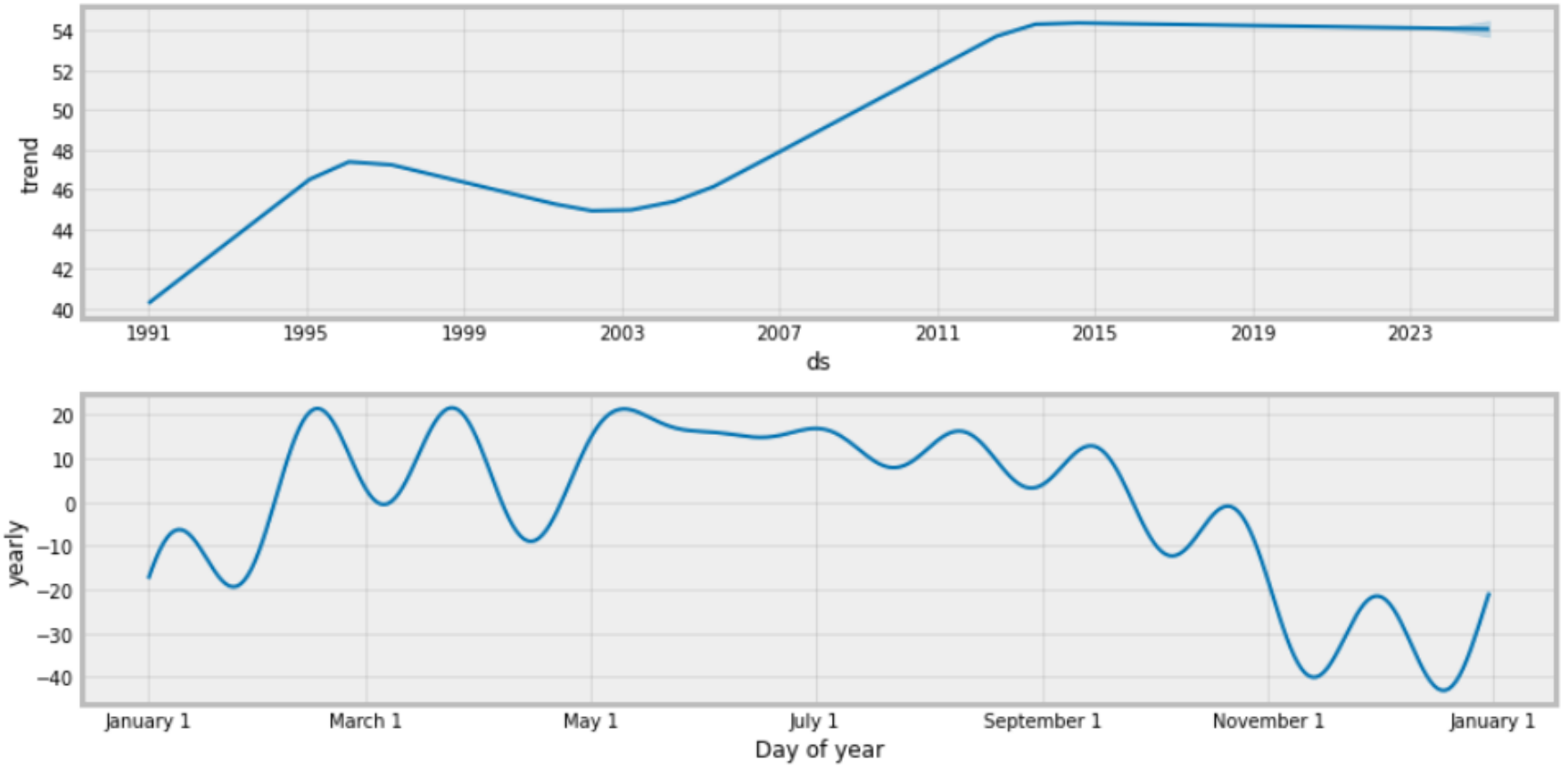
*More details you can find in the following script: *2.modelling_monthly.ipynb*

ocean

4.2. Predicting the monthly average of O3 values. Final predictions

• For the final predictions we used Prophet model trained on all dataset.
• The script of the final model on the Ocean Market is here: https://market.oceanprotocol.com/asset/did:op:6f975ec3b9064f98f6c672d070913b10b7bbcbc53a79d2c00e356ba4db719298
• By running the command *m.plot_components()* we can see the following description of the model. We have the trend and yearly components of the model.

Prophet model components:



*More details you can find in the following script: *2.modelling_monthly.ipynb*

ocean

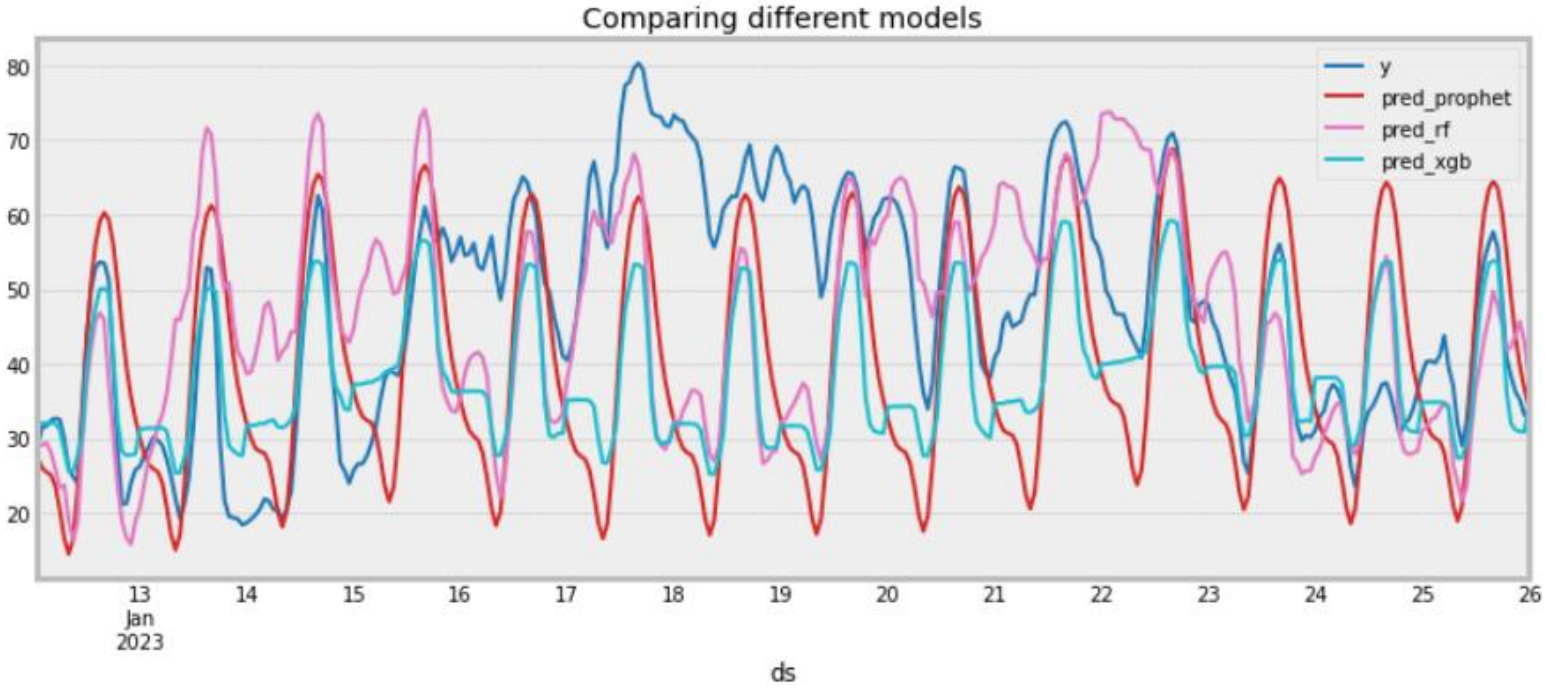
5.1. Predicting the hourly average of O3 values. Evaluation

We have used the last available 24*14 hours (14 days) for the evaluation set, the data before that was used for training. We have tested 3 different models: Prophet, Random Forest and Xgboost on the date-time features.

Evaluation have done by the different metrics:

- mse – mean squared error
- mae – mean absolute error
- mape – mean absolute percentage error

	mae	mse	mape
pred_prophet	13.85	318.67	0.30
pred_rf	13.42	281.86	0.32
pred_xgb	12.49	275.53	0.25

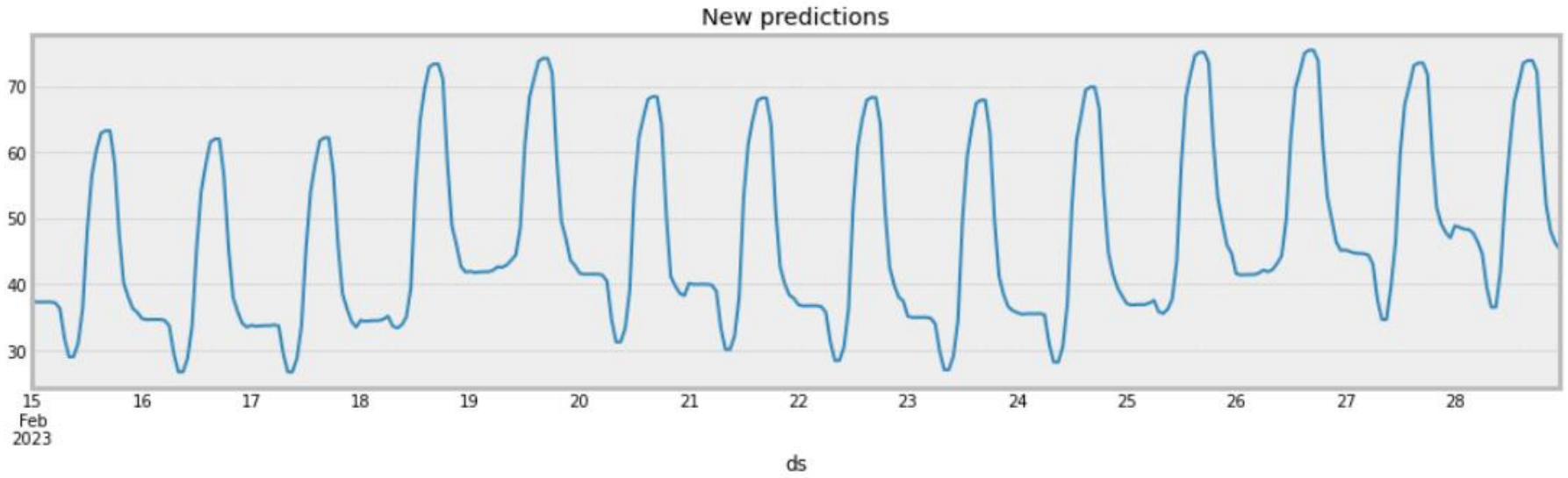


*More details you can find in the following script: *3.modelling_hourly.ipynb*

ocean

5.2. Predicting the hourly average of O3 values. Final predictions

- For the final predictions we decided to use Xgboost model.
- The script of the final model on the Ocean Market is here: https://market.oceanprotocol.com/asset/did:op:7e0bc5156f8cea8d4e752f2f3c4402da58753b5b10ee44a347ff5da7f56bb3d1



*More details you can find in the following script: *3.modelling_hourly.ipynb*

ocean