Ocean Data Challenge :: Air Quality in Catalonia

Introduction

Examining open data on particulate matter can help improve our understanding of air pollution and inform strategies to reduce it. By using the provided air quality data measured from various places in Catalonia, we analyze the global trends for the various pollutants over hourly and monthly periods of time, and employ algorithms to predict future hourly and monthly values. Insights and predictions would ideally be used to make more informed decisions on mitigating damage caused by poor air quality, and identify potential factors that lead to seemingly recurrent air quality patterns in Catalonia.

The following are the more precise requirements I attempt to tackle:

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

1. Analyze the evolution of pollution in Catalunya 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)
2. Analyze the relationship between altitude and concentration of particles in the air, and present your conclusions in graphical form. (10 points)
3. Analyze the concentration of pollutants in urban, suburban and rural areas, and present your conclusion in graphical form. (10 points)
4. 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

1. 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)
2. 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)

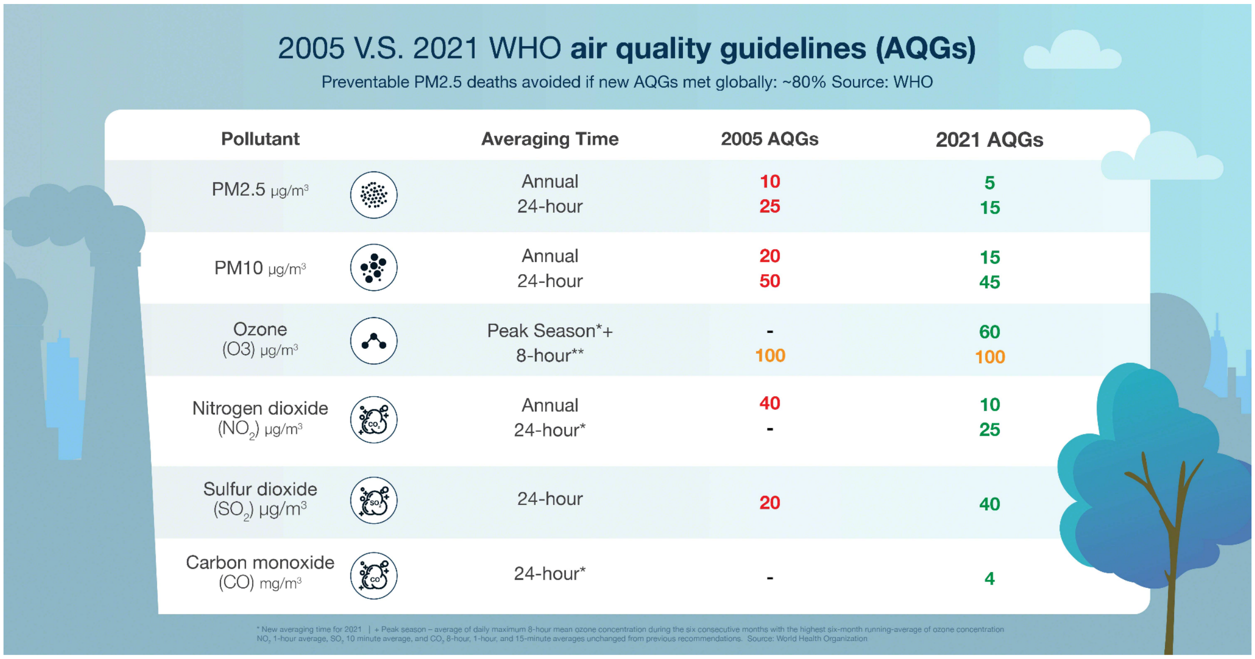
Methodology

Do the following feature engineering to prepare the dataset for the 4 points to analyze:

- Get mean, max and min of the pollutants value across the day

- Index and group dataframe by dates
- Extract pollutant specific dataframes

When tackling the global analysis section, I would focus on the following 6 pollutants as they are sufficient to check the air quality. It could have been used to compute the Air Quality Index (AQI) but due to lack of time I didn't look into it.



2005 V.S. 2021 WHO **air quality guidelines (AQGs)**

Preventable PM2.5 deaths avoided if new AQGs met globally: ~80% Source: WHO

Pollutant	Averaging Time	2005 AQGs	2021 AQGs
PM2.5 μg/m³	Annual 24-hour	10 25	5 15
PM10 μg/m³	Annual 24-hour	20 50	15 45
Ozone (O3) μg/m³	Peak Season*+ 8-hour**	- 100	60 100
Nitrogen dioxide (NO_2) μg/m³	Annual 24-hour*	40 -	10 25
Sulfur dioxide (SO_2) μg/m³	24-hour	20	40
Carbon monoxide (CO) mg/m³	24-hour*	-	4

* New averaging time for 2021 | + Peak season – average of daily maximum 8-hour mean ozone concentration during the six consecutive months with the highest six-month running-average of ozone concentration
NO_2 1-hour average, SO_2 10 minute average, and CO_2 8-hour, 1-hour, and 15-minute averages unchanged from previous recommendations. Source: World Health Organization

These are the following periodities I found:

MONTHLY

PM10

- early years have massive spikes of 50+ microgram/m3 (and usually in the middle of the year)
- Used to be generally above 100 microgram/m3
- Monthly mean and max PM10 starts to drop from 2006 onwards from slightly above 100 to about 50 for max PM10 value for the month
- Over most of the years, there are still significant spikes in PM10 mid year relative to the general level of PM10 throughout the year, so best times are start/end of year

PM 2.5

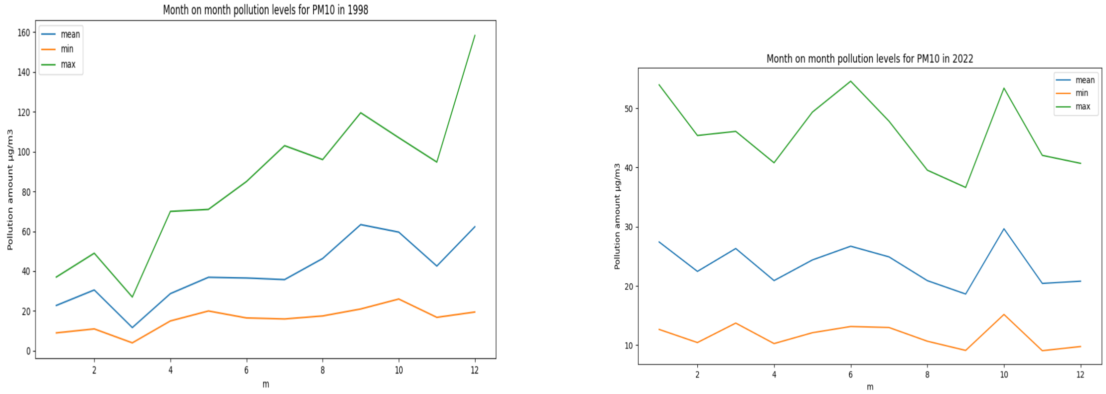
- Most years have spikes at the start/end of year and thus these are bad times
- General decrease of pm2.5 levels from mean 20 to mean 10, max 80+ to max 40

Ozone

- For all years from 1991-2022 except 2023, generally logartihmic increase and exponential decrease, peaking in the middle of the year, so worst in mid year
- Generally the mean o3 levels doubled from 1991 to 2022, with a significant decrease back to about 25+ mean in 2023

NO2

- Generally increase at the start/end of year, so best mid year
- N02 max levels have generally been above 70, and gradually decreased to about 50 in 2023



Figures showing the change in PM-10 levels on a year-on-year basis

For more monthly plots, you can view it in my github repo here

Hourly

PM2.5: In the past decade, generally best time is 3pm, worse is 8am and midnight

PM10: Best: 5am, Worst, 10am

Ozone: Best 8-9am, Worst 4-5pm

NO2: Best 5am and 3-4pm , Worst 10am and 10pm

S02: Best 11pm - 6am, worst is 11am-2pm

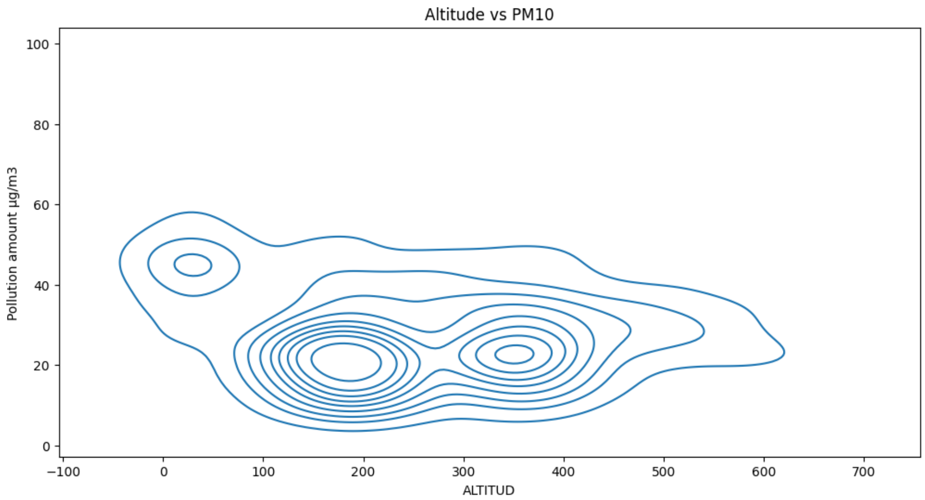
CO: 5am and 3-4pm, worst 8am and 8-10 pm

For more monthly plots, you can view it in my github repo here

Conclusion for challenge 1: Catalonia technically has no “best/worst” month or hour, because at any one time, the best times for one pollutant tend to coincide with the worst for another, and that’s just among the 6 important pollutants.

Global Analysis 2

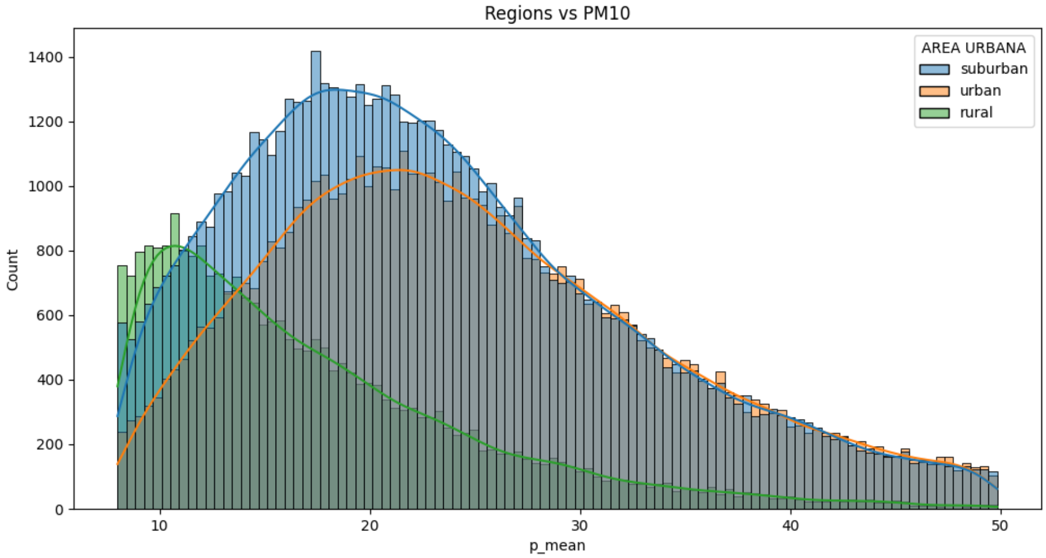
The rest of the plots can be found here



I've used a kernel density plot to visualise the distribution of the pollutant concentrations at various altitudes. E,g, for PM10, measurements done at about 200m altitude are likely to find concentrations at 20. For PM 2.5, there's concentrations measured are usually found around 10 and 40, with 10 likely being found at low and high altitudes.

Global Analysis 3

The rest of the plots can be found here



I trimmed the 5% and 95% percentiles away so that I could visualise the overall distributions better. For most of the data that has measurements from 3 regions, most of the time the mean of the concentration mean is concentrated at higher values in urban, suburban and rural area lowest in descending order

Global Analysis 4

I derived a city's rank by getting their ranks for each pollutant type and averaging the rank value. A shortcoming of this method is the pollutants are all equally weighted in terms of importance to "level of pollution". Again, lack of time to address this.

Best 5 overall
Flix at rank 0.0
Nou de Berguedà, la at rank 0.2236842105263158
Alcanar at rank 0.2894736842105263
Santa Maria de Palautordera at rank 0.3026315789473684
Ponts at rank 0.3815789473684211

Worst 5 overall
Hospitalet de Llobregat, l' at rank 3.9473684210526314
Granollers at rank 4.052631578947368
Mollet del Vallès at rank 4.131578947368421
Barcelona at rank 4.2631578947368425
Sabadell at rank 4.2894736842105265

A quick google map search shows that the best 5 are either suburban or rural, one of which seems like a beach-side resort. Meanwhile the worst 5 are all urban city areas. Corroborating global analysis 3.

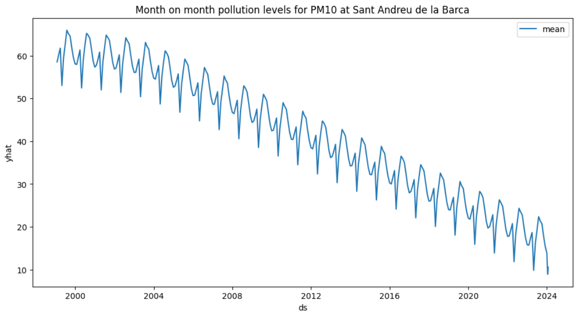
Algorithms

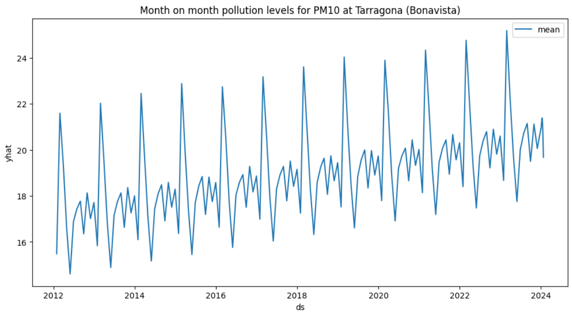
My core methodology is to just implement a model that captures (simple) seasonality well enough. I came to this conclusion because many of the plots have shown a really strong monthly trend that stays rather constant throughout each year.

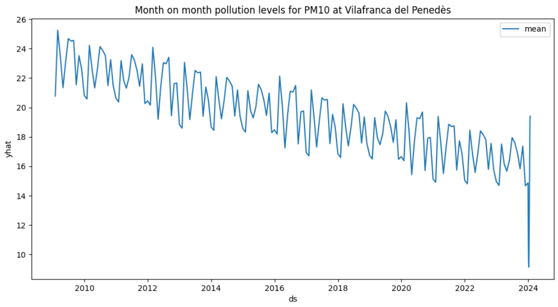
Further investigation into station-based pollution concentration trends however revealed that there aren't such strong monthly trends, but I believe the model of choice would handle another seasonality, which may include lagged features. Lagged features is possible because Spain being significantly big, pollutants from one region may travel to another region and take weeks or months to reach.

Model of choice for quick iteration: Prophet

After testing Prophet on existing data points, we can see prophet not only identifies both the year on year general decrease in pollutants, but also monthly seasonality too for several cities. Some cities have an increase in pollutants year on year and that's reflected as well. However, some sharp drops can be observed which do not fit the trend







Observations and Conclusions

- Prophet may have performed decently (unfortunately, i did not calculate the the error costs)
- Prophet would work very well for the global analysis portion as the seasonality is very clear cut.
- Across Catalonia, it seems that air quality has improved vastly, however it’s still borderline unhealthy based on WHO air quality guidelines.

Limitations and Recommendations

- I would prefer if this was more open-ended, such as allowing challengers to employ a more self-directed approach to gaining insights
- I would prefer more time since, at least from my perspective, the additional time that early attempters get may not give them much advantage since machine learning is hardly the best approach. Though with a bit more time, I could have at least tried the ML approaches
- I’m thankful that the slides were given in the discord, but the link to it, and I feel for many of the resources to get onboard Ocean, are pretty scattered. It would be great if a sanity check script or even a CLI be used to validate and submit our datasets/algo.