Ocean Data Challenge :
Air Quality in Catalonia

By : 0xCChan

Table of Contents

Section 1 : Introduction

The challenge examines the air quality in Catalonia and how the level of air pollution has changed over time. Environmental pollution has grown to be a pressing, urgent and rising concern today. As such, this will provide researchers and policymakers with an informed perspective of trends and patterns over time, whilst have the predictive algorithms to forecast future levels of prediction to adjust policies and regulations accordingly.

The data provided was the hourly data for pollutants measured at the automatic measurement points of the Air Pollution Monitoring and Forecasting Network in Catalunya from 1991 till 2023, providing a wide range of sample data points to analyze and predict with.

Section 2 : Global Analysis

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)

Best Hour (Lowest Pollution): 05h
Worst Hour (Highest Pollution): 09h

Best Month (Lowest Pollution): 8 (August)
Worst Month (Highest Pollution): 12 (December)

Periodicity of pollution

In general, pollution levels have increased over the years since 1991 on an aggregate view. Fig 1 shows a line graph of year vs mean daily air pollution level. A visible upward trend is seen till before 2000 where a slight dip was then observed. Since 2003, pollution levels began to rise once again with a fall observed from 2019, possibly due to COVID-19 where general consumption of electricity and industrial production slowed. The huge fall later is due to incomplete data from 2023 given that this is the current year.

Fig 1: Year vs Pollution Level

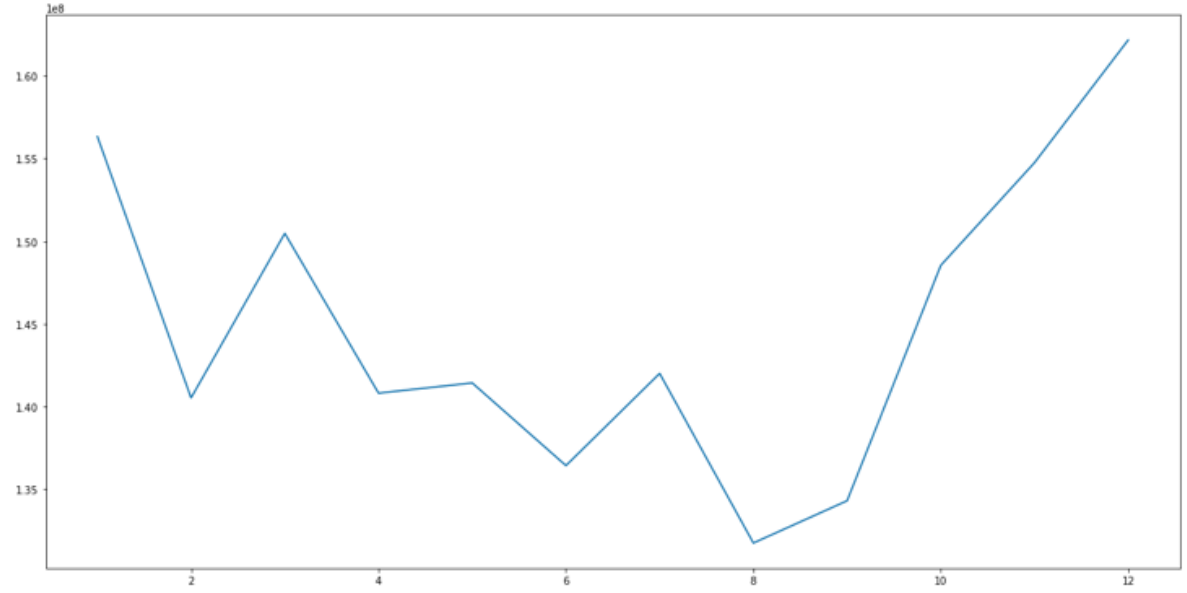

Fig 2: Month vs Pollution Level

From a monthly perspective, pollution levels peak near end of the year and is highest in December, while lowest in August.

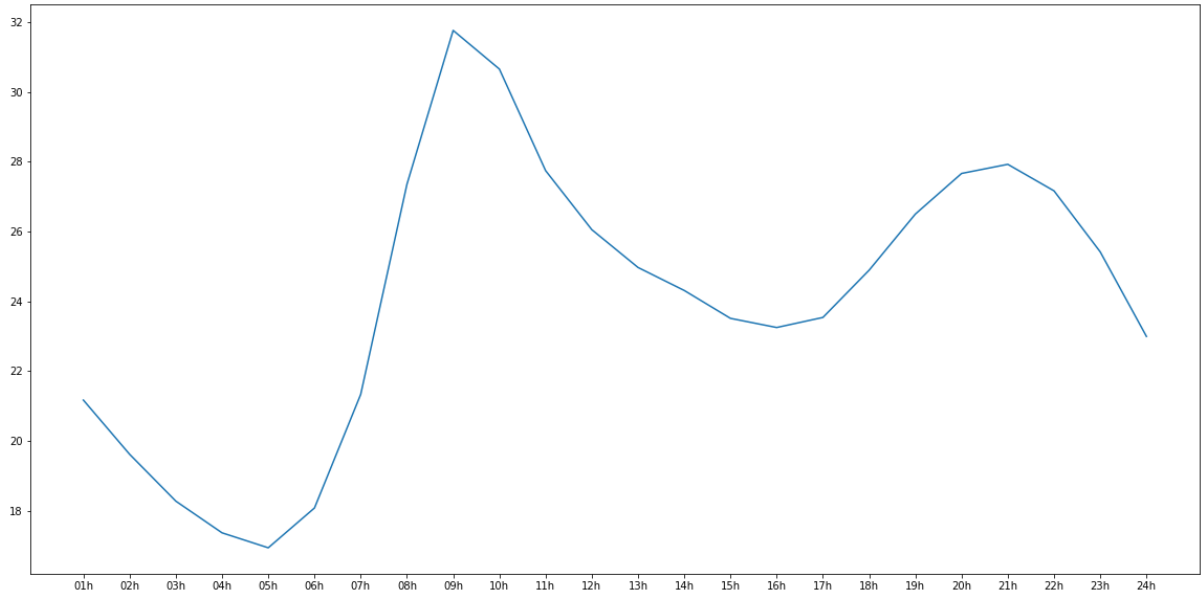

Fig 3: Time of Day vs Pollution Level

It is interesting to note that throughout the day, pollution levels are lowest in the early hours of the morning (0100 – 0500) before rising and peaking at 0900. This is likely due to the morning rush hour where more households and industries are awake and begin the workday. This then tapers off through the day till 1800 which is when work ends for majority of the workforce and hence, slightly rising once again. This then gradually drops entering late in the night as most offices and industries power down.

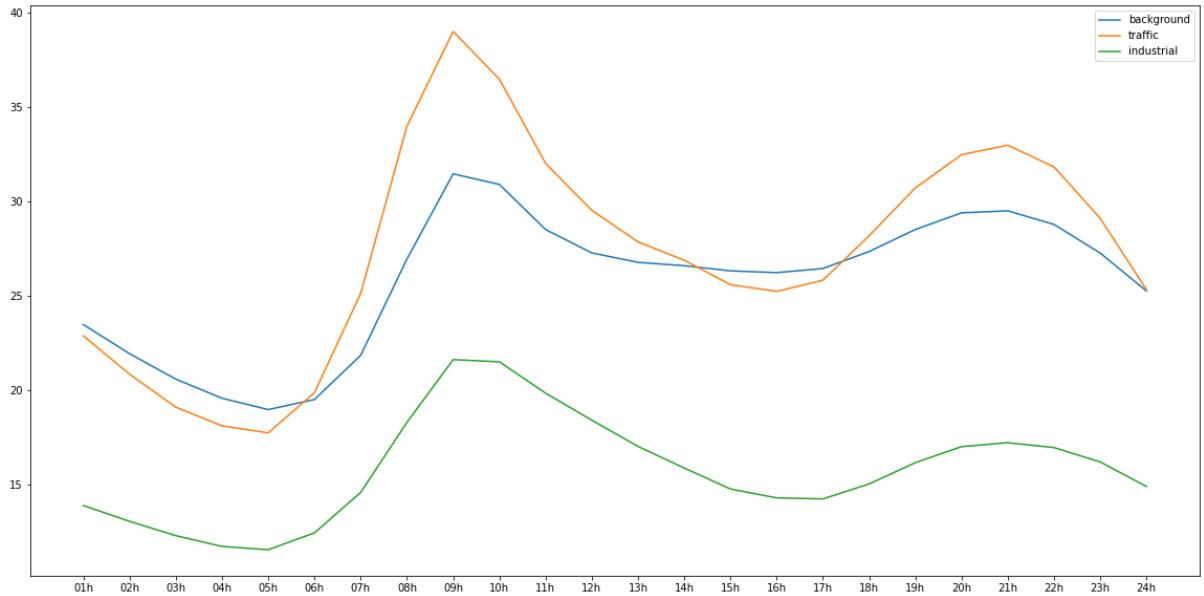

Fig 4: Time of Day vs Pollution Level, based on type of activity

This observation is consistent when I focused specifically on each activity – background, traffic, industrial. Surprisingly, traffic contributed the most out of the 3 categories, followed by background and industrial.

For the various contaminants, the fluctuations across time can be seen below, showing visible peaks between 0700 and 0900 during the morning rush hour and between 2000 and 2300.

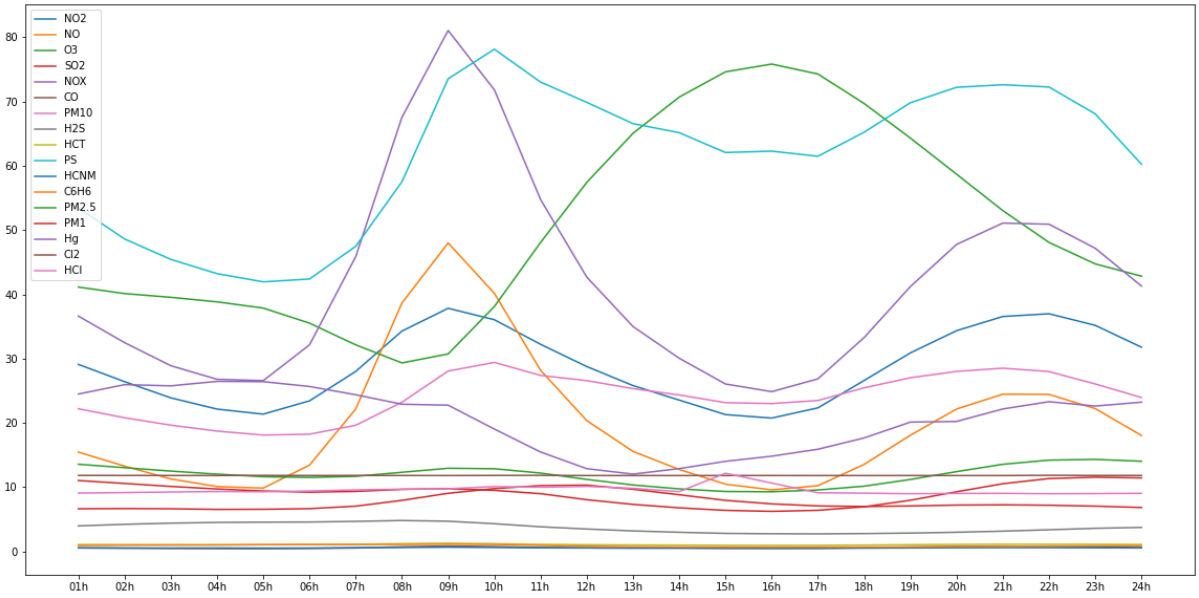

Fig 5: Time of Day vs Pollution Level, based on individual contaminants

In fact, when the mean levels of pollution for these contaminants were obtained, I noticed the following:

```
{'NO2': 28.726972450086652,
 'NO': 19.67604398930696,
 'O3': 50.44755462832006,
 'SO2': 7.679463006200341,
 'NOX': 41.77924053191837,
 'CO': 0.5867206946028865,
 'PM10': 24.08944333241132,
 'H2S': 3.6951940319250896,
 'HCT': 1.0584592538351025,
 'PS': 61.3644527172767,
 'HCNM': 0.534444459895515,
 'C6H6': 0.867045850332035,
 'PM2.5': 11.889098498921326,
 'PM1': 9.053263421854398,
 'Hg': 20.456384538515575,
 'Cl2': 11.806028663691471,
 'HCl': 9.526350766907232}
```

Fig 6: Mean of pollution level for individual contaminants

PS was the highest, followed by O3 and NOX.

From Fig 5, PS and NOX are noticeably elevated starting from 0800-0900, and NOX levels dip from 0900 before rising again end of the day, at 1800. Meanwhile, PS remains at elevated levels since then.
On the other hand, O3 levels also rose at the same time before gradually diminishing from 1700.

Focusing specifically on these 3 pollutants,

Fig 7: Year vs Pollution Level, based on Top 3 pollutants

PS levels dropped drastically throughout the years between 1991 and 2008 while NOX levels fell between 2009 and 2020. O3 levels remained relatively constant with a slight increase throughout the time frame of 1991 – 2023.

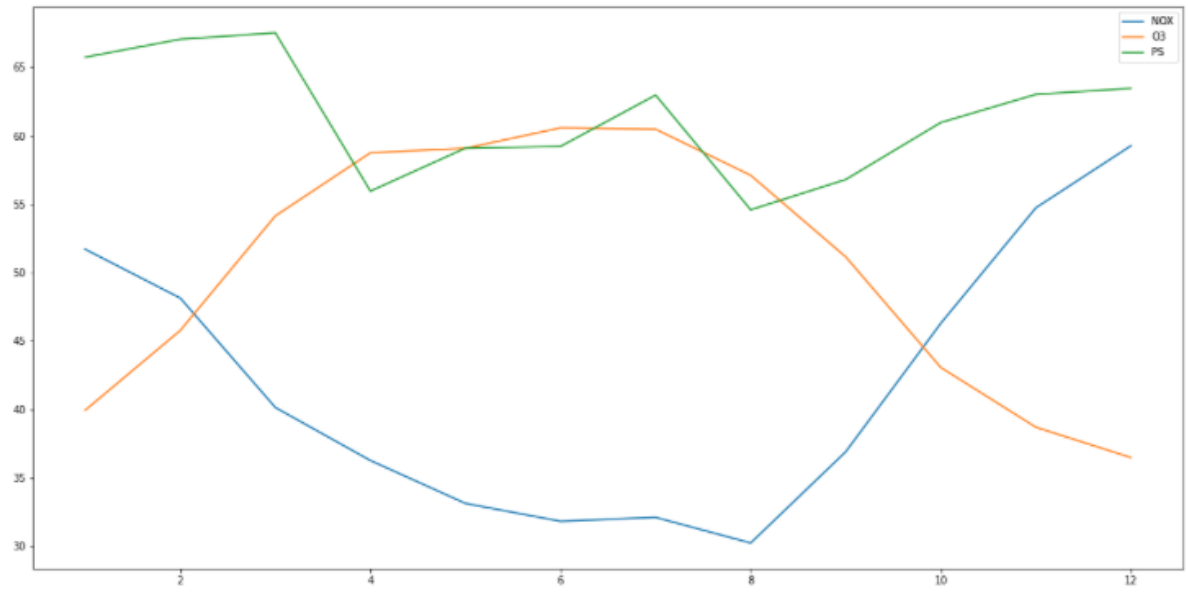

Fig 8: Month vs Pollution Level, based on Top 3 pollutants

On a monthly basis, PS pollution levels are the highest, peaking in March before gradually dipping till August and finally recovers heading into year end.

O3 levels are especially high in the months from April to August while NOX levels dip from January to August before increasingly sharply going into the year end.

Analyze the relationship between altitude and concentration of particles in the air and present your conclusions in graphical form. (10 points)

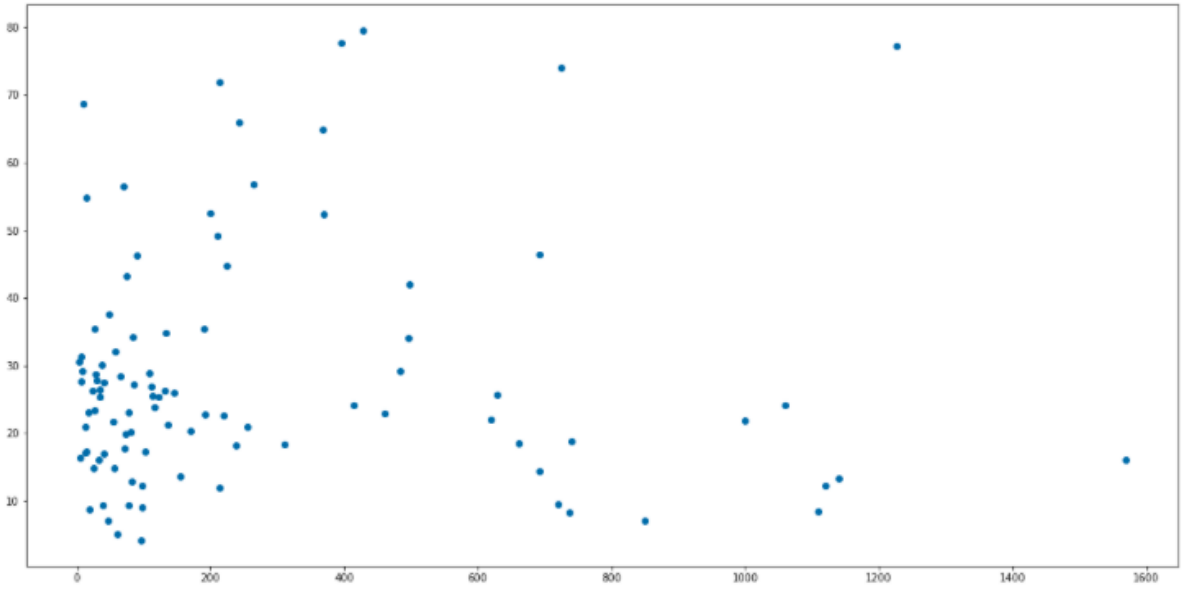

Fig 9: Altitude vs Mean Concentration of particles

From the above graph, the higher the altitude, the lower the concentration of particles in the air. In fact, it is particularly more concentrated between 0 and 200.

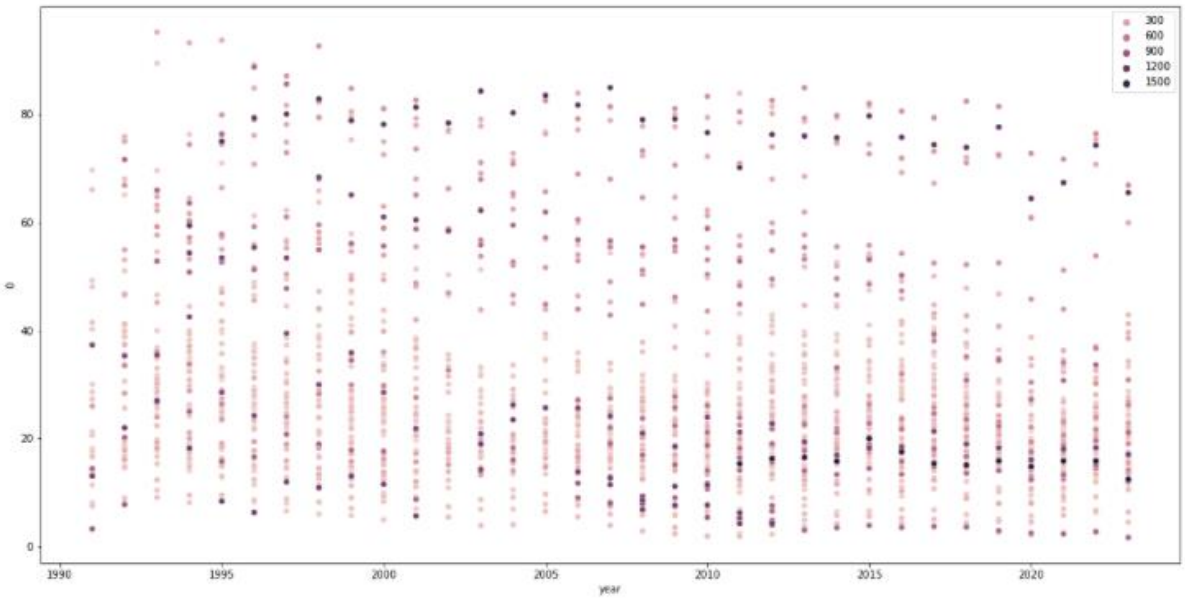

Fig 10: Altitude vs Concentration of particles in the day, hue = year

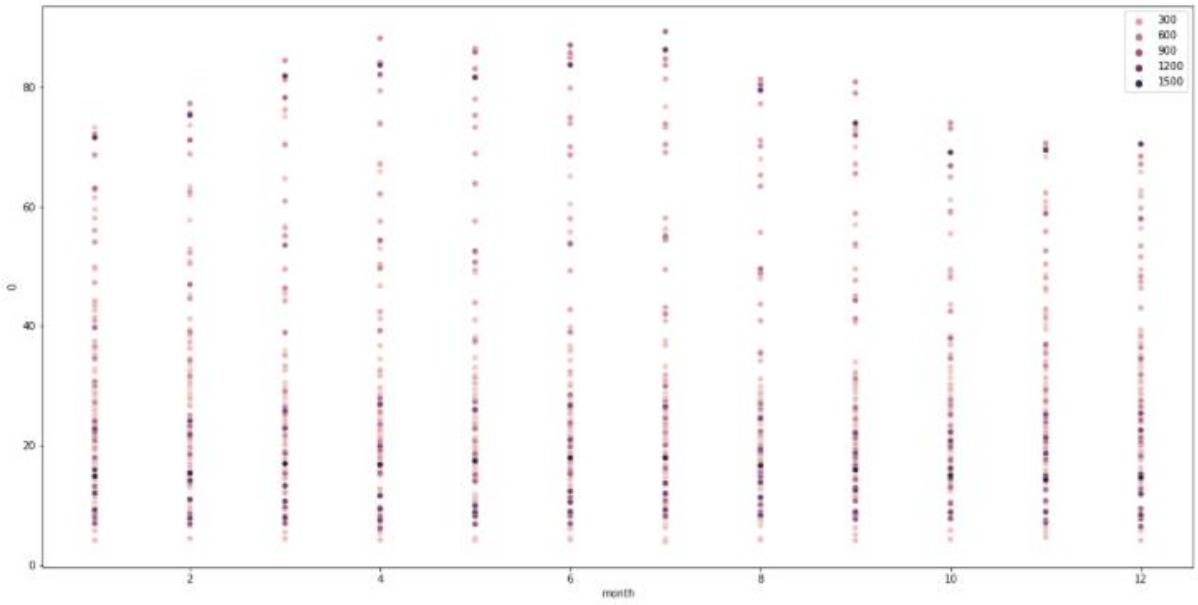

Fig 11: Altitude vs Concentration of particles, hue = month

On closer examination, this phenomenon is consistent throughout all months and years based on the scatterplot. A clearer view of Fig 10 and Fig 11 can be seen in the notebook where it's visualized with 1 year and 1 month at each iteration, respectively.

Interestingly, when I observed the level of pollution based on the altitude and hour of the day, the concentration of particles is the highest at 0900 for altitudes below 200m.

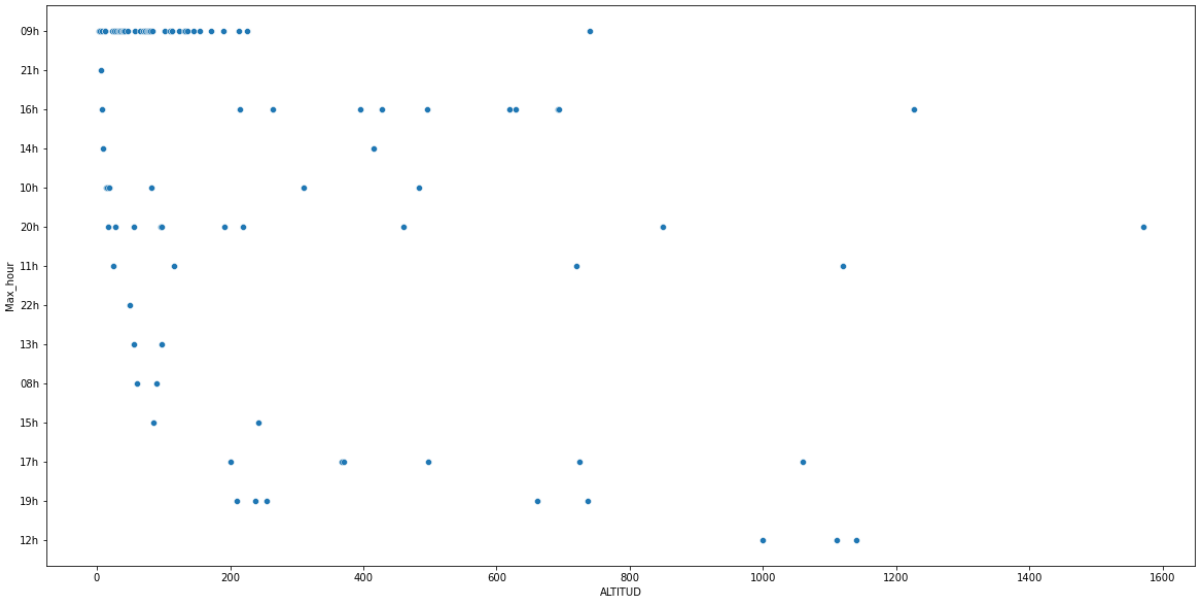

Fig 12: Altitude vs Max_hour, concentration of particles

Analyze the concentration of pollutants in urban, suburban and rural areas, and present your conclusion in graphical form (10 points).

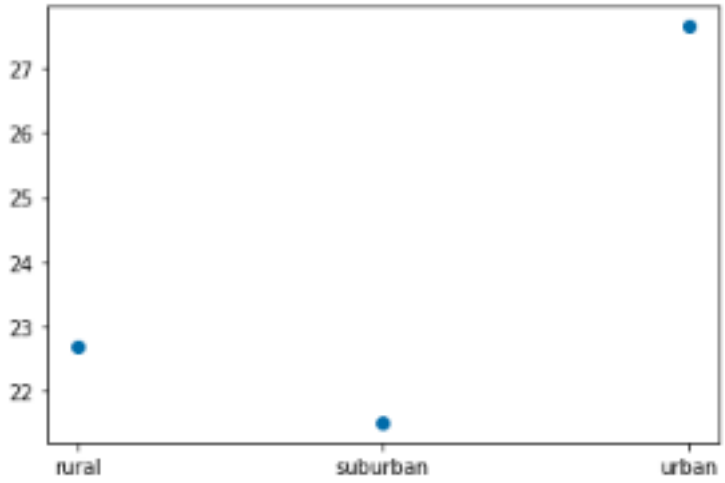

Fig 13: Mean concentration of pollutants in urban, suburban, and rural areas

On average, pollution levels are much higher in urban areas, followed by rural areas and suburban areas.

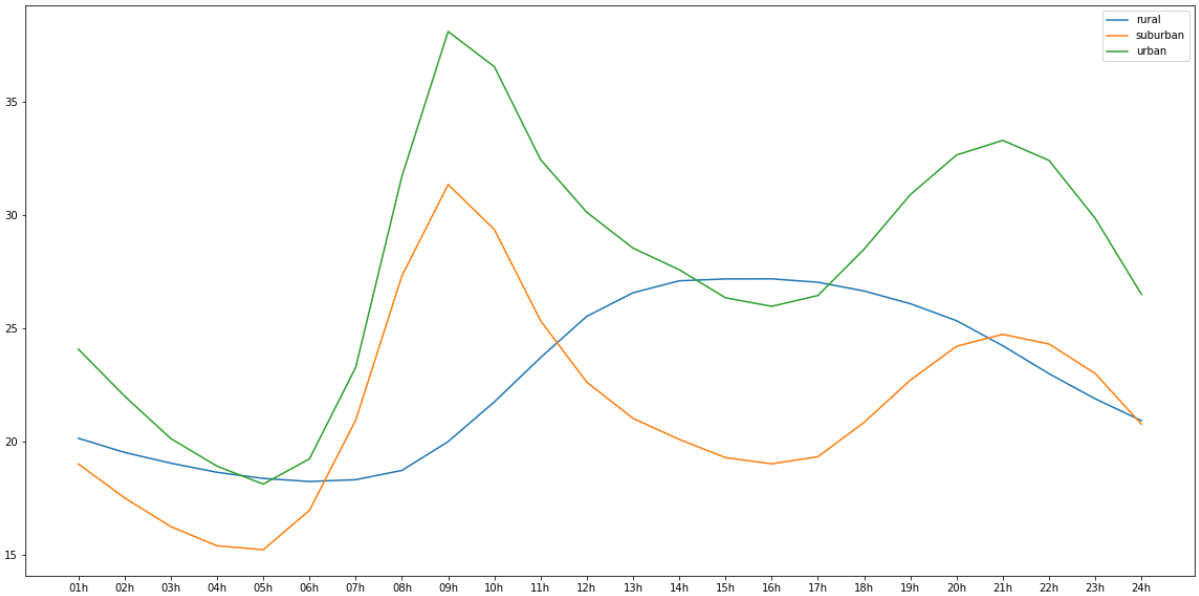

Fig 14: Area vs Concentration of particles in the day

From Fig 14, it can be observed that the pollution levels of both suburban and urban areas follow a similar trend peaking between 0800 and 1000, suggesting the start of the workday

(as suggested in the initial question where my analysis revealed the largest contribution was from traffic). It peaks once again near end of the day around 2100 and 2200.
On the other hand, rural areas seem to have a higher pollution level between 1400 and 1800 hours, coinciding with the valley observed for urban and suburban areas.

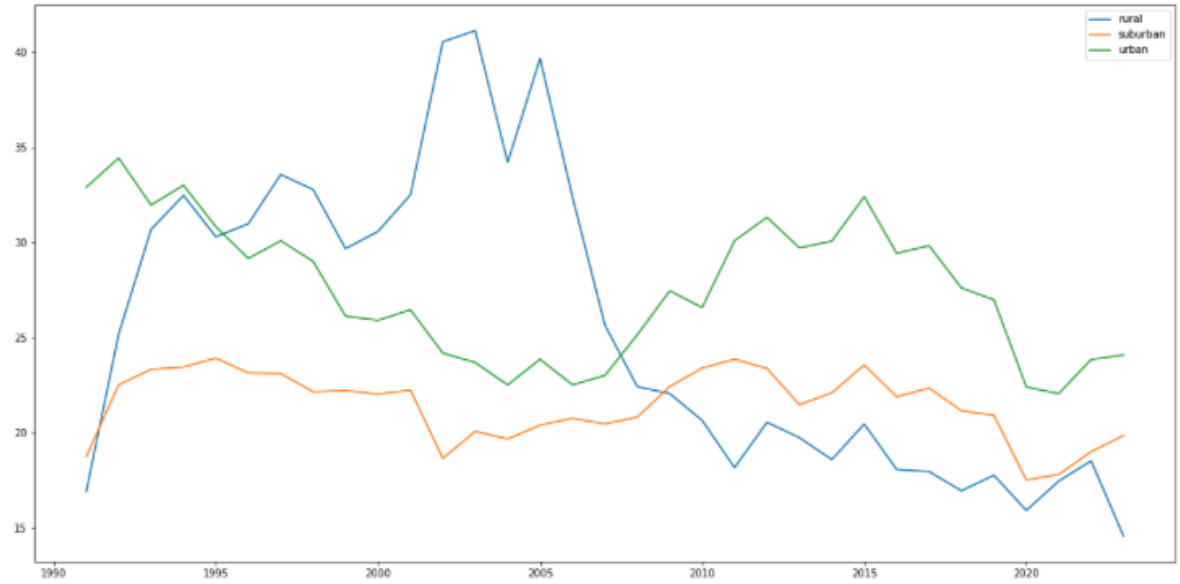

Fig 15: Area vs Concentration of particles in the day, by Year

From Fig 15, pollution levels in rural areas rose drastically between 1991 and 2003 before falling off drastically from 2005. It suggests some form of rapid urbanization during the period and potential large industrial projects, contributing to the sudden spike and anomaly relative to urban and suburban areas within the same period.
Meanwhile, pollution levels for both urban and suburban periods appear to follow the same trend, with urban pollution being higher than the latter.

Rank the cities in the dataset according to their level of pollution, and create best-5 and worst-5 lists (10 points)

Based on the definition in the glossary, cities refer to urban areas.

	Cities	Pollution Level
1 (Least polluted)	'Vilanova i la Geltrú'	17.100073558129107
2	'Lleida'	18.312305052552528
3	'Rubí'	19.491064043491786
4	'Girona’	19.940439887245905
5	'Manresa'	20.071313279676296
6	'Tarragona'	20.562939566006584
7	'Sarrià de Ter'	21.69918096454433
8	'Cornellà de Llobregat’	23.376554732781248
9	'Mataró'	23.464747769114197
10	'Sant Cugat del Vallès'	25.545633654077434
11	'Barberà del Vallès'	26.380800189600972
12	"Hospitalet de Llobregat, l'"	28.05613602146046
123	'Santa Coloma de Gramenet'	28.53543910519406
14	'Terrassa'	28.854513127950053
15	'Vic'	29.268897097485663
16	'Badalona’	29.72637154526249
17	'Barcelona'	29.944528742613794
18	'Granollers'	33.96033151750067
19	'Sant Adrià de Besòs'	34.07701323959278
20	'Sabadell'	44.81476276781273
21	'Vilafranca del Penedès'	44.81677900416478
22	'Gavà'	54.803378772577524
23	'Ripollet'	56.50251247640842
24 (Most polluted)	'Viladecans'	68.65718772670249

1: least polluted

5: most polluted

Best 5 Cities (Least Polluted)

```
Vilanova i la Geltrú
Lleida
Rubí
Girona
Manresa
```

Worst 5 Cities (Most Polluted)

```
Sabadell
Vilafranca del Penedès
Gavà
Ripollet
Viladecans
```

Section 3: Algorithmic Prediction of Pollution

The contaminant chosen was PM10. This is solely random based on :

- contaminant = data['CONTAMINANT'][0]

Observations and Conclusion

a. Per month for the next 24 months

The data was organized into monthly time periods, yielding a total of 312 data points.

Methdology

Data Cleaning / Preparation

- The dataset was split into the following: Training Data (60%), Validation Data (20%) and Test Data (20%). These different datasets were split and ordered by time to avoid the look-ahead bias which could lead to inaccurate results

- In order to impute missing data for variables, KNN Imputation is implemented with an initial value of 5 for n_neighbors.

- Finally, I performed feature scaling through the MinMax Scalar. This involves rescaling and shrinking the data of the features into a given range. It is useful for gradient descent algorithms since a huge difference in ranges of features will cause vastly different step sizes for each feature. Therefore, having features on a similar scale will help the gradient descent converge more quickly towards the minima. Since the distributions of the various features are not Gaussian, StandardScalar was thus omitted as an option.

XGBoost

XGBoost is a decision tree based ensemble model that adopts a gradient boosting framework. Presented by Chen and Guestrin in 2016, the algorithm mainly enhances existing boosting models through:

a. **Regularisation** : Penalises more complex models through L1 and L2 regularisation to prevent overfitting and smoothens the final learnt weights

b. **Shrinkage Estimates** : Scales newly added weights after each step of tree boosting through column sub-sampling. It reduces the influence of each individual tree and leaves space for future trees to improve the model

Using the scaled dataset, I iteratively applied the XGBRegressor on the scaled training dataset to train the model and tune the parameters. This was then predicted on the scaled

validation dataset and scaled back to the original range to obtain the predicted price of Bitcoin. To choose the optimum value for each parameter at each iteration, I picked the value which corresponds with the minimum RMSE.

Iteration 1 : Default Parameters

I initiated the model using an XGBRegressor with default values of each variable and the objective function as 'reg:squared error' - regression with squared loss.

Iteration 2 : n_estimators, max_depth

- n_estimators : Range of 10 to 100, with a step of 5
- max_depth : Range of 1 to 10, with a step of 1

Iteration 3 : learning_rate, min_child_weight

- learning_rate : Range of 0.01 to 1 with a step of 0.01
- min_child_weight : Range of 1 to 21 with a step of 1

Iteration 4 : subsample, gamma

- subsample : range of 0.1 to 1 with a step of 0.1
- gamma : range of 0.1 to 1 with a step of 0.1

Iteration 5 : colsample_bytree, colsample_bylevel

- colsample_bytree : range of 0.5 to 1 with a step of 0.1
- colsample_bylevel : range of 0.5 to 1 with a step of 0.1

To evaluate the performance of the model, the following metrics were used:

Accuracy

a. Mean Absolute Percentage Error (MAPE)

In the time series forecasting model, I used MAPE to determine the absolute error and avoid any model or parameters with a high absolute error. However, MAPE would not work well alone as it cannot detect the error of zeros and extreme values. Hence, I need to use RMSE along with MAPE.

b. Root Mean Square Error (RMSE)

RMSE measures the standard deviation between the predicted values from the model and the actual values of a dataset. The higher the standard deviation, the larger the error. Similar to MAPE, RMSE will always be positive and a lower value indicates higher performance.

**For all the hyperparameters tuning, only RMSE was used as the minimum. While both MAPE and RMSE summarise the variability of the observations around the mean, they are of different scale, hence contributing to the vastly different values. RMSE was therefore chosen

since it is the basis for how the model fits the data. RMSE is a representation of model error and hence, a more complete representation.

XGBoost Hyperparameter Tuning Summary Statistics

After training the model,

Parameter	Original	After Tuning
n_estimators	100	95
max_depth	3	1
learning_rate	0.1	0.66
min_child_weights	1	8
subsample	1	0.1
colsample_bytree	1	1
colsample_bylevel	1	1
gamma	0	0.07
RMSE	**3.219**	**3.087**
MAPE	**0.1127**	**0.1078**

Prediction for the next 24 months:

Month (t onwards)	Pollution Level
1	22.250833
2	23.186556
3	23.676156
4	24.478715
5	24.638324
6	24.980042

7	24.562629
8	25.199952
9	23.782390
10	23.686079
11	25.679644
12	24.148466
13	22.945832
14	22.672966
15	31.078597
16	29.637695
17	27.439085
18	25.828588
19	26.812440
20	21.972796
21	18.438138
22	16.255232
23	15.875347
24	19.375509

b. Each hour of the day from February 15 to 28

Methodology

The Prophet model is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with seasonality.

As such, I applied the model with both weekly and daily seasonalities set to True given the patterns observed for air pollution over time. It was trained on the dataset, starting from April 17 2019, comprising of 500,000 data points. This was truncated because of the high computational complexity with over 3 million data points should I use the original dataset.

Predictions : Attached as ‘algo_2.csv’ given the large file size.

Section 4: Observations and Conclusions

Contrary to popular belief, air pollution in Catalunya appears to be widely contributed by traffic and background activities as opposed to industrial activities. These coincide with the early morning rush hour and end of office hours.

In fact, going to the year end, air pollution levels spike significantly on average, peaking in December, which could signal the holiday season as a possible reason. Nonetheless, it is heartwarming to see that air pollution levels have decreased over the years, especially plummeting during the COVID-19 pandemic period.

Section 5: Limitations and Recommendations

Given the urgency of this project, I have only used 2 machine learning algorithms (XGBoost and Prophet Model) and did not manage to leverage many other machine learning algorithms which could possibly produce better results. Since XGBoost is a Gradient Boosting algorithm that works with the concept of predecessor learning, it would be extremely sensitive to outliers. Each classifier will be forced to fix the previous error, which may lead to a huge deviation in the final model. Furthermore, for the Prophet model, it was only applied on 500,000 data points due to the computational intensity required to train should I use the whole data set. However, this can be circumvented with higher computing power.

As for the analysis of data, seasonality was not accounted for in unravelling the patterns observed in similar time periods in the past for Algorithm 1. I could have analysed trends in February over the past few years to check if the recent history might reveal a seasonal pattern as a guide for future predictions. This may involve layering a Prophet model for instance as used in Algorithm 2.

In the future, the competition can consider hosting a workshop / video prior to the event to show users how to access the data on the Ocean Marketplace and steps required to submit the information. This walkthrough will help to clarify any doubts and guide us along the way.