Mitigating Air Pollution in Poland Through Machine Learning

April 5, 2023

Contributors

Joseph Antony (Task lead)
Adam Sciegaj
Catalin Vulcan
Parnika Nikhil Damle
Shubhankar Sharma (Task lead)
Tim Hayes
Michael Adeyeri
Catalina Valdivia (Task lead)
Vidushi Khanna (Task lead)
Konstantin Skrebunou
Rukshar Alam
Basavaraj Hirebidari
Devarshi Choudhury
Vinod Cherian
Sivarama Krishna Raju Teeparti
Amit Sakar
Paula Montagnana
Kojo Kesse-Amoahene (Task lead)
Maciej Zdonski
Shwetha Mallikarjun Hiregowdar
Shnehi Karki

Chapter Lead

Alexander Lau

Table of Contents

INTRODUCTION **4**
PROBLEM STATEMENT **4**
DATASETS **4**
METHODOLOGY & TOOLS **5**
1. TASK 1 - Data Translation & Data Cleaning 5
    1.1 Air Quality Dataset 5
    1.2 Weather Dataset 6
    1.3 Static Annual Dataset 7
    1.4 Final Merged Dataset 8
2. TASK 2 - Data Pre-processing 11
    2.1 Inferring Powiat and Voivodeship from coordinates 11
    2.2 Air Quality Dataset 11
    2.3 Weather Dataset 12
    2.4 Static Annual Dataset 12
    2.5 Powiat proximity matrix 13
3. TASK 3 - EDA (Exploratory Data Analysis) 13
    3.1. Missing Data Imputation 14
    3.2. Merging Datasets 20
    3.3. Calculation of CAQI Values 22
4. TASK 4 - Modeling 24

RESULTS AND INSIGHTS **25**
1. Analyzing Pollutant Data 25
    1.1. Weekly Seasonality of Pollutants 27
    1.2. Monthly Seasonality of Pollutants 29
    1.3. Annual Mean Pollutant Concentrations 29
2. Analyzing CAQI (Common Air Quality Index) 30
3. CAQI Levels vs Weather 32
4. Impacts of COVID-19 Lockdown 33
5. Impact of Public Holidays and School Holidays 34
6. Modeling 35
    6.1. Feature Engineering 36
    6.2. Supervised Learning 37
    6.3. Feature Importances of Models 37
    6.4. Forecasting Future CAQI Levels 40

CONCLUSION **41**
EXECUTIVE SUMMARY **42**

INTRODUCTION

Air pollution is a growing concern in Poland, with the country consistently ranking among the countries with the worst air quality in Europe. This poor air quality not only affects the health and well-being of the population, but also puts a strain on the healthcare system. In order to address this issue, it is important to identify the main factors contributing to air pollution in Poland and to develop an effective tool for predicting air quality.

PROBLEM STATEMENT

Air pollution is a particular problem in Poland. The annual EEA (European Environment Agency) reports on air quality show that Poland is among the countries with the worst air quality in Europe. Bad air quality affects people's lives and constitutes a considerable health risk. Therefore, mitigating air pollution could improve quality of life and lead to an overall healthier society. At the same time, this would reduce costs for the Polish health care system.

An important step towards mitigation is the identification of the main factors and causes of air pollution specific to Poland. By using local time-series data on air pollutants together with other relevant country-specific data, an AI-assisted approach could yield valuable insights in this matter. In particular, a machine-learning model for air quality prediction could give policy makers a simple but powerful tool to help tackle the issue of air pollution in Poland.

DATASETS

Three datasets were provided, each one containing a different type of data:

- Dataset 1: Daily air quality data from 2015 to 2021, provided by the Chief Inspectorate for Environmental Protection in Poland (https://powietrze.gios.gov.pl/pjp/archives). This dataset contained daily measurements of the NO_2, O_3, PM_{10}, and $PM_{2.5}$ levels, taken by various measurement stations across Poland. The O_3 measurements were reported every hour, while the rest of the measurements reported one value per day. This dataset was complemented by the list of measuring stations along with their addresses, coordinates, and miscellaneous data. This dataset was in Polish.

- Dataset 2: Daily weather data from 1979 to 2021, provided by European Climate Assessment and Dataset (ECAD) project (https://www.ecad.eu/) . This dataset contained daily measurements of the cloud cover, global radiation, humidity, mean temperature, precipitation, sea level pressure, snow depth, sunshine, and wind speed. The data was measured by different weather stations across Poland. For each individual category, a list of stations together with their unique identifiers was also available. This dataset was in English.

- Dataset 3: Static annual data from 2010 to 2021, provided by the Polish Central Statistical Office (https://bdl.stat.gov.pl/bdl/start). This dataset contained information on various features such as animal stock, area by land use, crop production, emission of particles and pollutant gasses, forest area and fires, population density, production of electricity, vehicle types, and air pollution reduction systems. Some categories are given on the Powiat (district/county) level, whereas other categories are aggregated on the Voivodeship (state) level. Although the majority of the data is reported for the period 2010-2021, some categories span different time ranges, i.e., vehicles by type and fuel data is available from 2015. Similarly, the data on air pollution reduction systems and plants of significant nuisance is available from 2017. This dataset was in English.

- Additional datasets pertaining to Polish public holidays & observances, school holidays and list of Powiats with functioning coal plants were gathered from online sources (source 1, source 2, source 3).

METHODOLOGY & TOOLS

1. TASK 1 - Data Translation & Data Cleaning

1.1 Air Quality Dataset

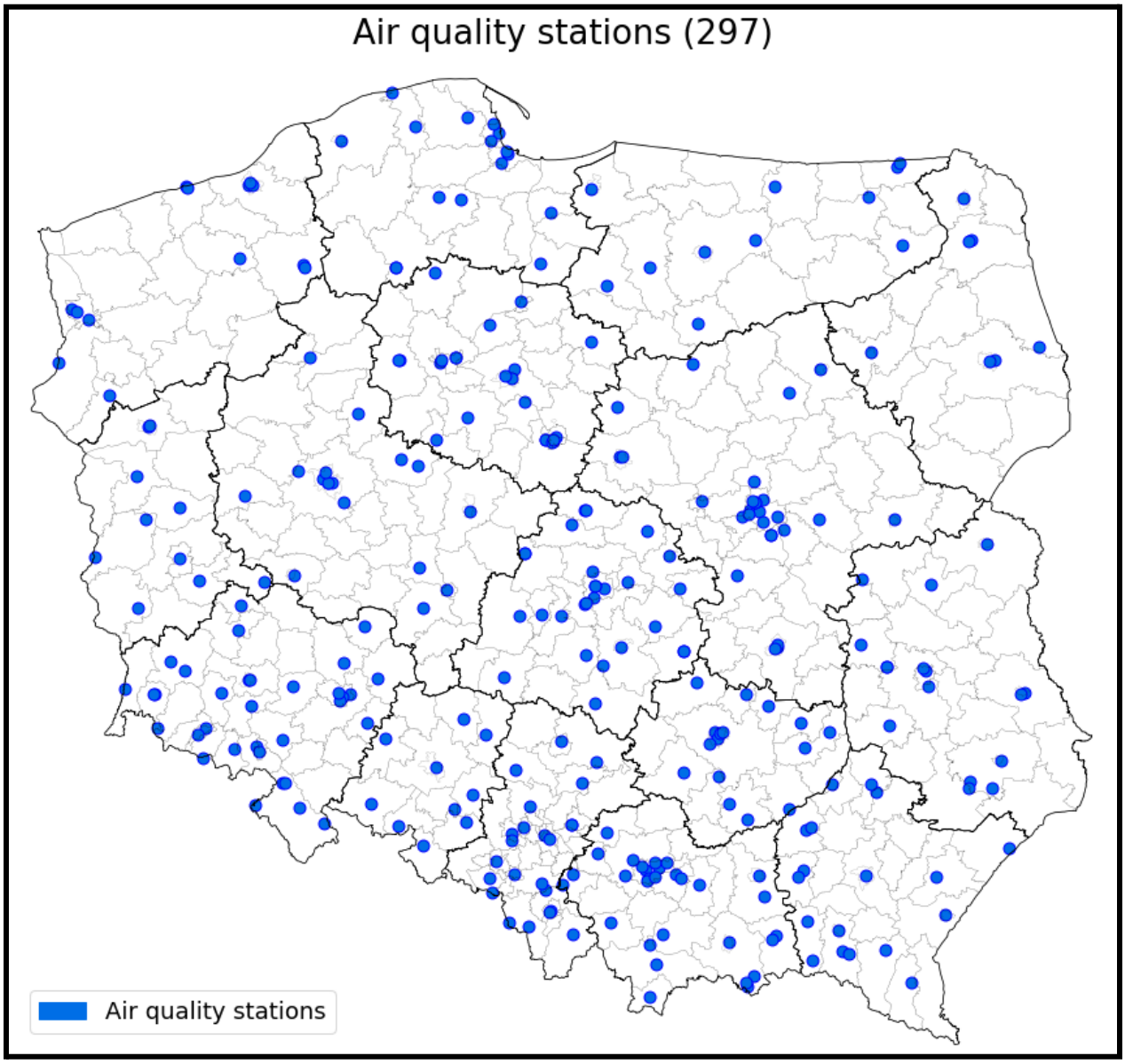
The work started with translating the Air Quality dataset from Polish into English. Apart from translating the column names, the type of decimal separator was changed from comma (used in Poland) to period. Afterwards, the dataset was filtered so that only the data from the period 2017-2021 was left. As the data for every year and pollutant was in separate files, those were combined into a single file. Since some of the pollutants were measured every hour, it was decided to use a daily average value for the pollutants measured at a finer temporal resolution.

Usually, different pollutants were measured by different stations. To take this into account, the station ID (and its coordinates) was added as another column to make sure that all data is present. This resulted in quite much missing data, as e.g., NO_2 was measured by only three stations across the country. Furthermore, it was the case that some stations did not show a complete record over all the years. Similarly in these cases, missing data was inserted into the dataset.

Also, some inconsistency was that some air quality measurement stations had very similar names and there was a suspicion that it is the same station that has a typo in the name or changed the name. At the same time, for such stations, the measurement periods supplemented each other, and after checking that it is indeed one station, their readings were combined into a single time series. The approach was based on searching for stations with similar names with differences of 1-3 characters. The search for such stations is implemented using the Levenshtein library. Thus, about 30 stations with similar names were found. Next, the presence of such stations in metadata was checked and their coordinates were compared.

Further, readings from such stations were combined into one row if their coordinates coincided, and the measurement periods did not intersect, i.e. complemented each other.

Final datafile contained the date, station ID, station coordinates (latitude and longitude), and the averaged pollutant measurement for each of the pollutants.

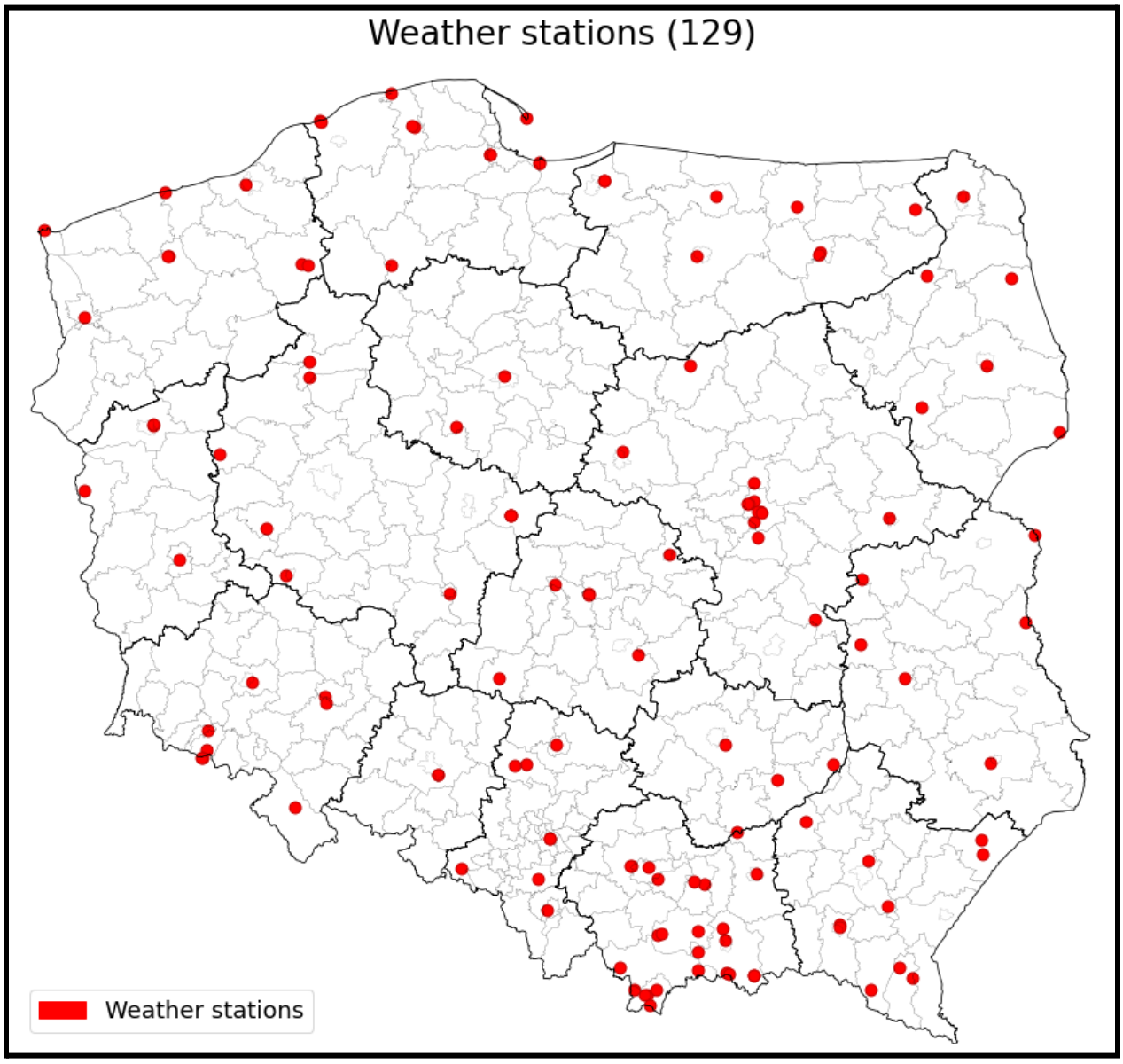


Air Quality Station Coverage

1.2 Weather Dataset

As the data for every variable was given in multiple text files (one file per station), the work with this dataset started with combining and merging the data into a single file. Each text

file included the station ID in the name, which made it possible to include this information as another column.



Weather Station Coverage

1.3 Static Annual Dataset

The static annual dataset did not require much cleaning, as the data was already aggregated on the Powiat and Voivodeship level. Each Powiat and Voivodeship had a unique code, which made it possible to infer the Voivodeship for a given Powiat, as the first two digits encoded the Voivodeship. Filtering the dataset to the period 2017-2021 and merging into one single file was done together with Task 2 (data pre-processing).

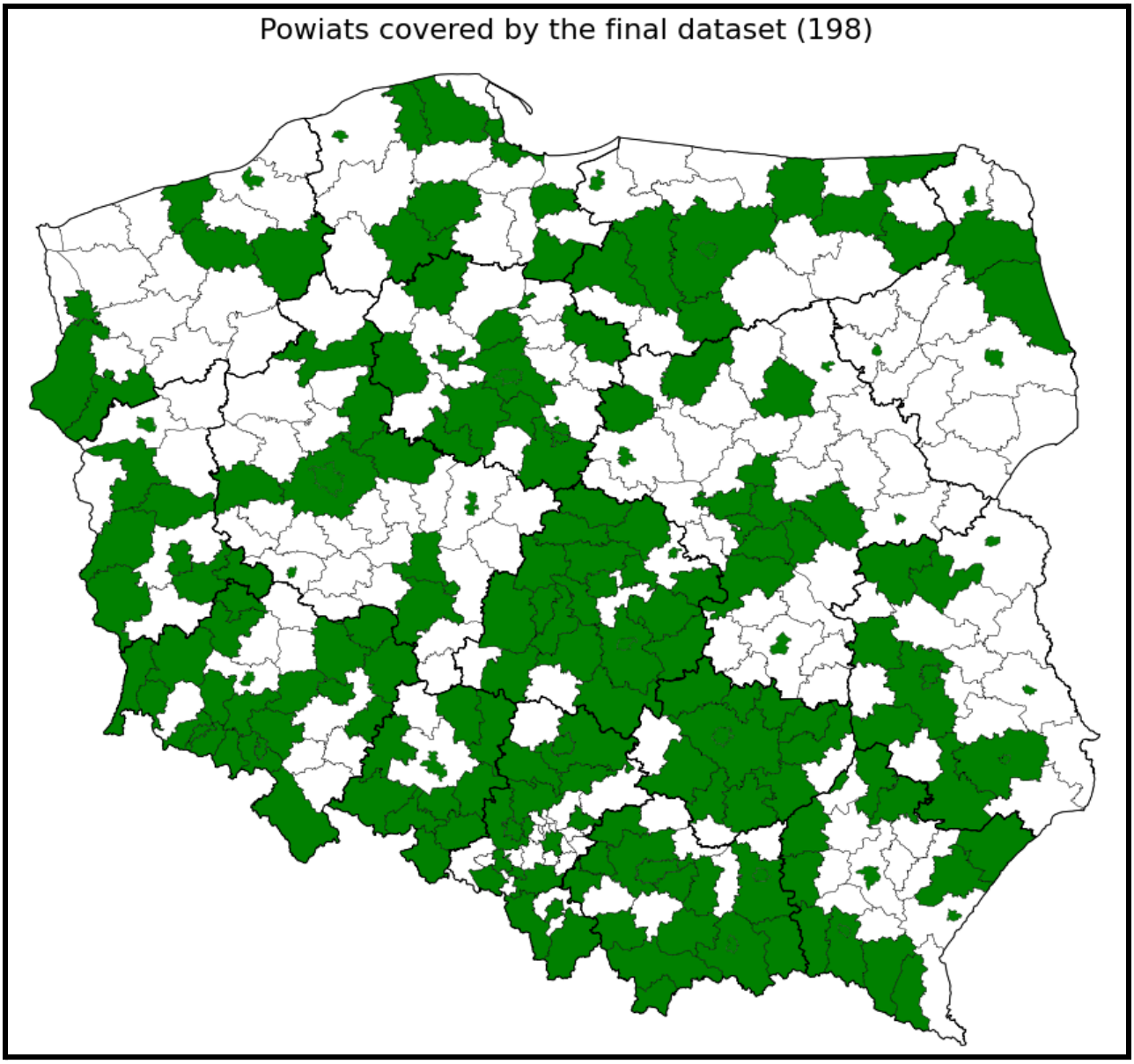
1.4 Final Merged Dataset

Our primary goal regarding the three datasets was to join them meaningfully. As such, we decided to split the process into two stages.

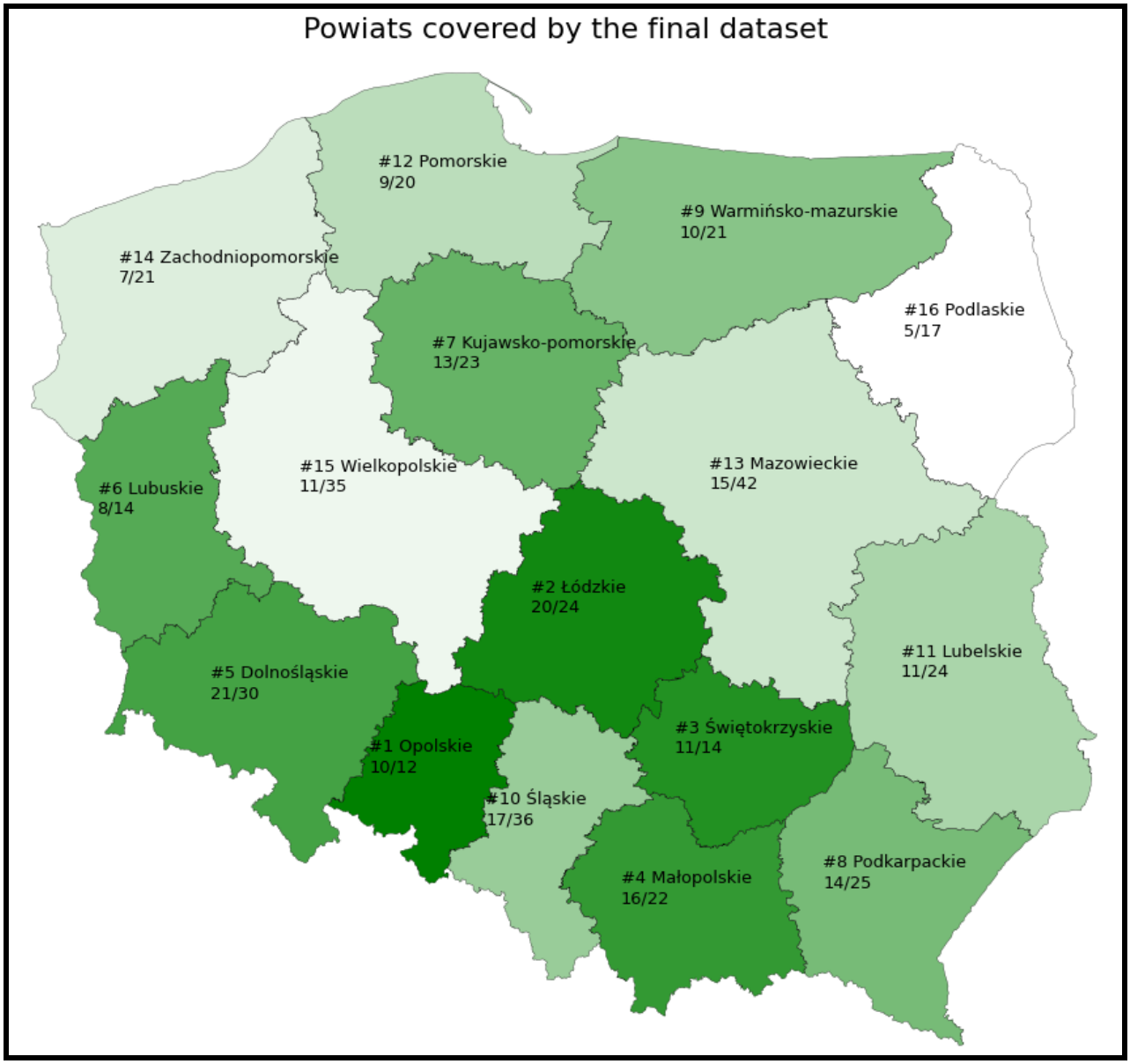
In the first stage, we joined the pollutant stations dataset with the weather stations dataset. The pollutant and weather stations datasets have been aggregated at powiat level, then merged together using powiat as the binding key between the two datasets.

In the second stage, we joined the resulting dataset from the first stage with the static annual dataset. For this, we assigned the corresponding powiat to all pollutant station locations and calculated all annual static values at the powiat level. Then, we assigned the derived static annual values per powiat to the pollutant station locations described above.

The bulk of joining the three datasets was performed and detailed in TASK2 - data preprocessing and was finalized in TASK 3 - EDA (Exploratory Data Analysis). The final dataset contains data for 198 of the 380 powiats in Poland. We highlighted the coverage of the final dataset reported to the total number of powiats in the plots below.



Powiats (districts) of Poland for which Air pollution data is available in the dataset



Voivodeships (Provinces) of Poland with number of Powiats with available data

2. TASK 2 - Data Pre-processing

2.1 Inferring Powiat and Voivodeship from coordinates

In order to merge the air quality and weather data, it is necessary to gather stations which are close to each other. This can be done e.g., with the *haversine* library, which computes the distance between two coordinates. This would require to look for the nearest (or few nearest) station for each station in the primary dataset. However, it was decided that a better approach would be to aggregate the data on Powiat level, due to several reasons. First, the Powiat was considered to be small enough so that any variation in air quality and weather data could be disregarded. Second, this key is present in all three datasets and could be potentially used for merging the complete dataset.

In Poland there are 16 Voivodeships and 380 Powiats, whereas 66 of the latter are City Powiats. Moreover, some Powiats share the name and could be only distinguished when paired with the parent Voivodeship. However, neither the Powiat nor Voivodeship information was given for weather stations. For air quality stations only the Voivodeship was given. However, for both types of stations we had access to the coordinates. For this reason, it was decided to be necessary to be able to determine the Powiat and Voivodeship given the coordinates.

To do this, an approach based on GEOJSON data was used. The GEOJSON data found online comprised vertex coordinates of individual Powiats and Voivodeships in Poland. Using the *shapely* library, it was possible to check whether a point with given coordinates lies inside the polygon defined by the vertices present in GEOJSON files. Moreover, the format of the returned Powiat and Voivodeship matched with the format given in the Static Annual Dataset. Thanks to this, it was possible to use a unified format in the separate datasets.

2.2 Air Quality Dataset

Using the merged air quality data, station coordinates were used to infer the corresponding Powiat and Voivodeship for each air quality station.

The approach to finding the relevant Powiats and Voivodeships was to reverse geocode with the Geopy client. For those stations that failed to find the corresponding Powiat, its status as a city with the Powiat status was checked in open access sources. For all other stations for which it was not possible to determine their administrative units, information about Powiats and Voivodeships was obtained using POST requests via API to the geonames.org service.

Furthermore, a binary feature (isUrban) indicating whether the Powiat is a City Powiat or not, was added to the station list. In this form, the dataset was later used for imputation of missing data.

2.3 Weather Dataset

It was found that the coordinates of the weather measuring stations were given in degrees, minutes and seconds, whereas the coordinates of air quality measurement stations were provided with decimal values. To remedy that, the coordinates of the weather stations were converted to the decimal format. Furthermore, a binary feature (isUrban) indicating whether the Powiat is a City Powiat or not, was added. Using the list of weather stations, a database containing the unique station ID along with its coordinates was created. Using the GEOJSON approach, the corresponding Powiat and Voivodeship for each station was found.In this form, the dataset was later used for imputation of missing data.

2.4 Static Annual Dataset

The static annual dataset has already cleaned data, aggregated on the Powiat and Voivodeship level. However, some data was available only on the Voivodeship level. For these categories, it was decided to distribute the data to the Powiat level. The distribution could be made based on either the area of the Powiat or the population density, as both information was available. For each Voivodeship, the area and population ratios of every constituent Powiat were calculated. Subsequently, the data given on the Voivodeship levels was multiplied with the ratios, so that every Powiat was present in the dataset. In the end, the data from all separate files was merged into one dataset, in which each row represented a distinct Powiat, and the data from all files was placed in subsequent columns.

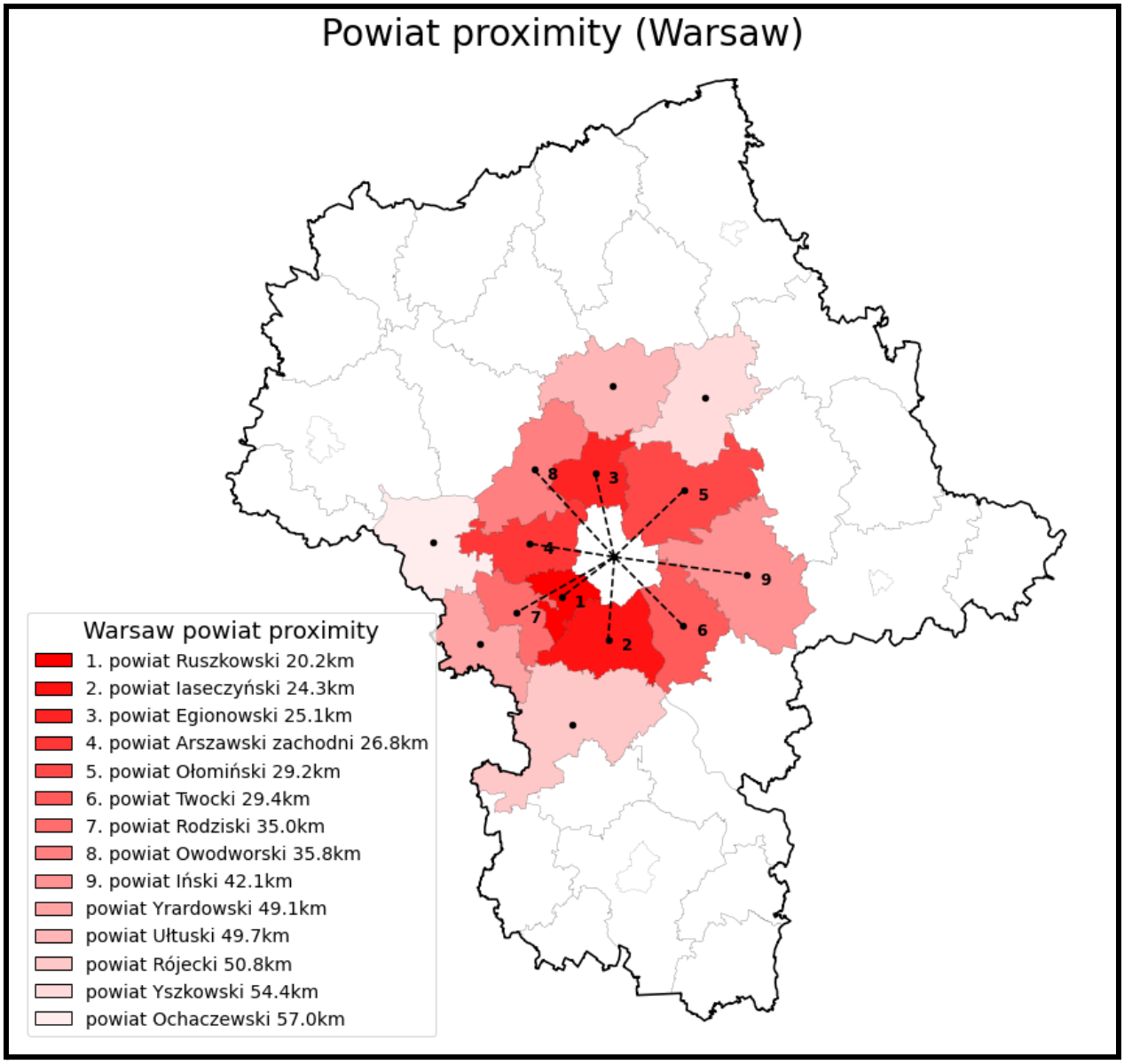
As the data was distributed by both area and population, it was decided to keep only one distribution for merging with the remaining datasets, depending on the nature of the underlying data. Following choices were made for the affected columns:

Data	Distributed by
Animal stock	area
Crop production	area
Forest fires	area
Production of electricity	population
Air pollution reduction systems in plants	population
Plants of significant nuisance to air quality	population

The final merged dataset contained columns representing the different types of data (previously located in separate files) and rows containing unique Powiats. In this form, this dataset was used for merging with the remaining datasets after imputation of the missing data was performed (Task 3).

2.5 Powiat proximity matrix

In order to perform imputation of missing data (as described further in Section 3), it was necessary to find the nearest (or a few nearest) stations for a given primary station. Since it was decided to work on the Powiat level, this task translated to finding the nearest Powiat(s) to a given Powiat. To do this, it was chosen that the distance between the Powiats could be represented by the distance between Powiat centroids. Thus, after calculating the centers of each Powiat, it was possible to calculate the distances to every other Powiat.



Neighbouring Powiats of Warsaw by distance

From the GEOJSON data, the vertices of the polygon constituting the Powiats were available. It was assumed that all Powiats are non self-intersecting polygons, i.e., the centroid lies inside the polygon. For each Powiat with *n* vertices (x_0, y_0), (x_1, y_1), … , (x_{n-1}, y_{n-1}), the centroid coordinates (C_x, C_y) can be calculated as:

$$C_x = \frac{1}{6A}\sum_{i=0}^{n-1}(x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

$$C_y = \frac{1}{6A}\sum_{i=0}^{n-1}(y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i)$$

where the area *A* is calculated as:

$$A = \frac{1}{2}\sum_{i=0}^{n-1}(x_i y_{i+1} - x_{i+1} y_i)$$

Having computed the centroids for each Powiat, the distances between the given Powiat and the remaining 379 powiats were computed using the haversine library. Afterwards, the list of distances was sorted in ascending order. As a result, a distance matrix was created, in which each column contained a list of Powiast from closest to furthest. This information was later used during data imputation for finding the closest powiat which contained air quality/weather data.

3. TASK 3 - EDA (Exploratory Data Analysis)

The pre-processed Air Quality datasets contained daily mean measurements of all four pollutants (NO_2, O_3, PM_{10}, $PM_{2.5}$) from stations located throughout various Powiats (districts) in Poland. However, most stations only measured a single pollutant and very few measured multiple pollutants. None of the stations measured all four pollutants (Air Quality Stations Dashboard).

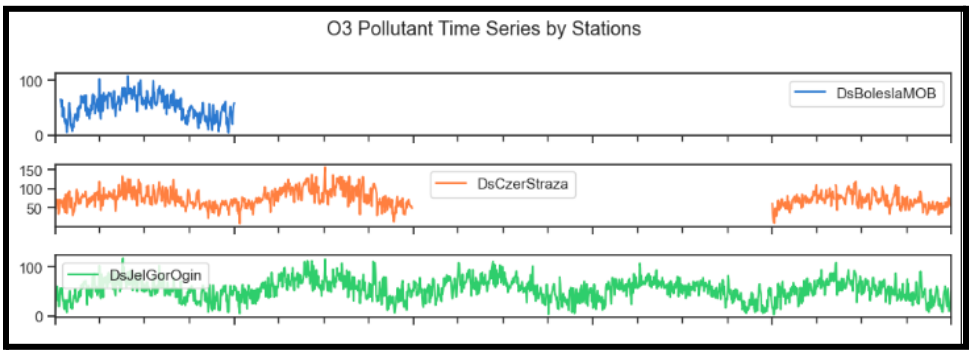
Similarly, the pre-processed weather datasets contained mean daily measurements of various weather variables, such as Cloud Cover (CC), Wind Speed (FG), Humidity (HU), Sea Level Pressure (PP), Global Radiation (QQ), Precipitation (RR), Snow Depth (SD), Sunshine (SS), and Mean Temperature (TG). Only a few weather stations measured multiple weather conditions, and several outliers were present in some measurement readings.

One of the challenges we faced was figuring out how to merge the air quality, weather, and static datasets into a single dataset. The air quality and weather datasets had the data at a station level, but a single powiat (district) could have multiple monitoring stations. The static dataset mostly contained annual data at the Powiat or Voivodeship level.

To merge the datasets, we decided to aggregate all the data at the Powiat level and then join all three datasets using the column indicating the respective Powiats in all three datasets. By aggregating and joining on this column, we limited the final dataset to 198 Powiats and predictions would be made at the Powiat level rather than the individual station level, which

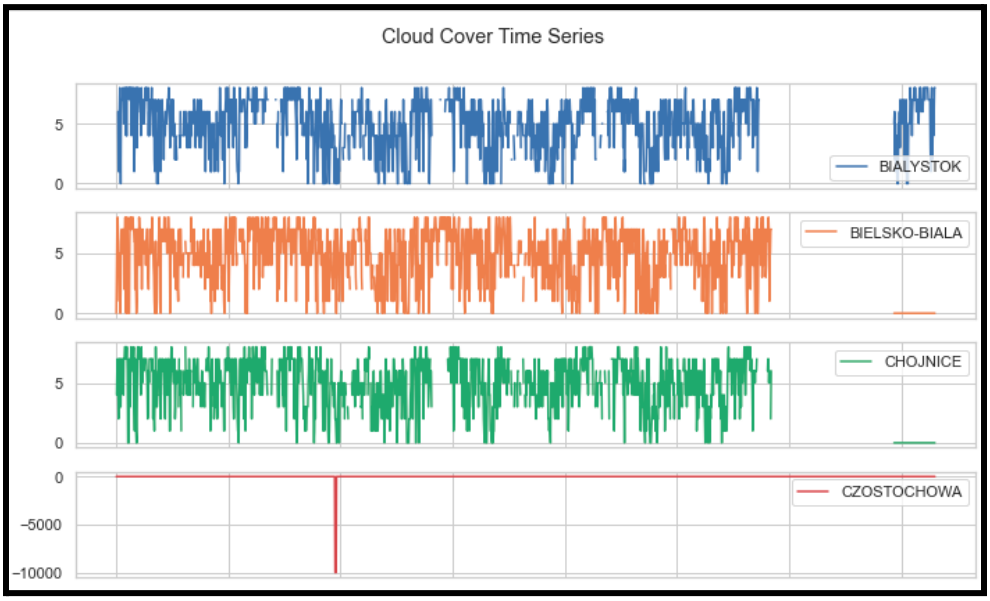
made more sense (explained further in section 3.2). Before joining the datasets, we had to impute missing data in both air quality and weather datasets and aggregate the data from the station level to the Powiat level by taking the mean measurements of the stations.

3.1. Missing Data Imputation



Daily measurements of O_3 from 2017 to 2021 for each station. Many stations have large gaps of missing data.

During the preliminary exploratory data analysis (EDA), we identified a significant data quality issue: many stations, both Air Quality and Weather stations, had missing data for long periods of time. To derive meaningful insights from the data and advance to the modeling phase of the project, we needed to handle these missing values appropriately. Consequently, we created two subtasks: one to address missing values in the Air Quality dataset and the other to handle missing values and outliers in the Weather dataset.



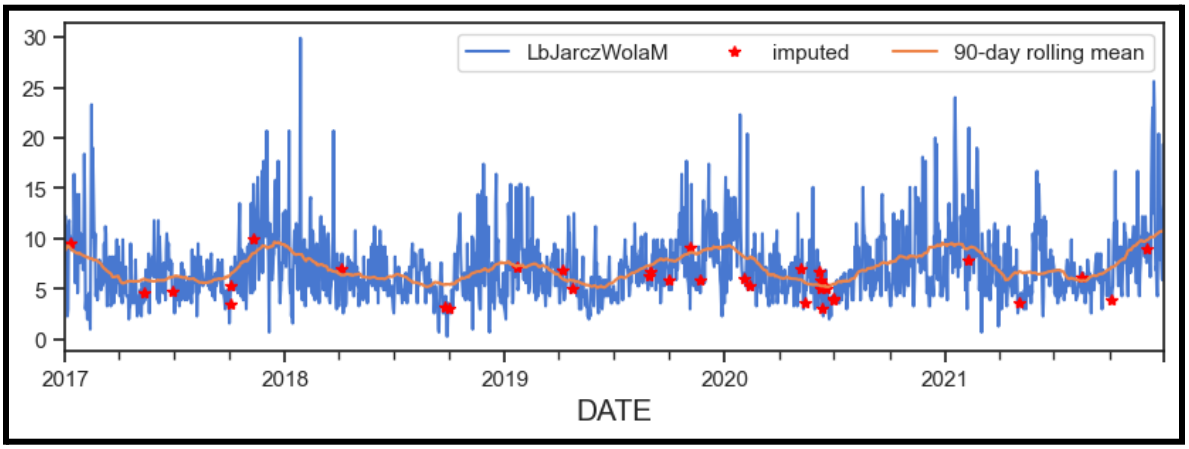
Daily measurements of Cloud Cover from 2017 to 2021 for each weather station. Many stations have large gaps of missing data, especially for the year 2021. Also, the station at the bottom has extreme outliers.

a. Subtask 1 - Pollutant Dataset

For the Air quality datasets, depending on the scale of missing values for each station, several imputation techniques were utilized.

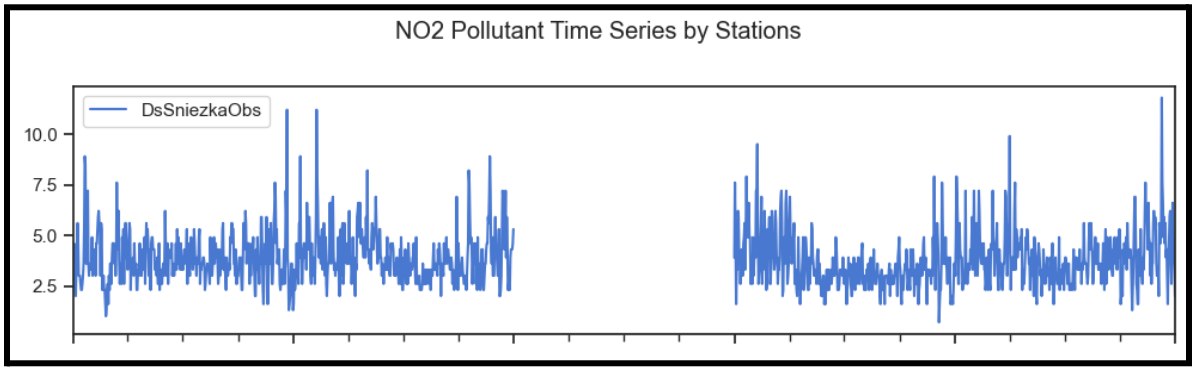
I. Linear Interpolation

The stations measurements with very few missing data were imputed using Pandas interpolation function, where the “linear” method is selected. The result is as shown in the plot below. The red points represent the imputed data point.



Example of Imputed NO_2 measurement from a air quality monitoring station.

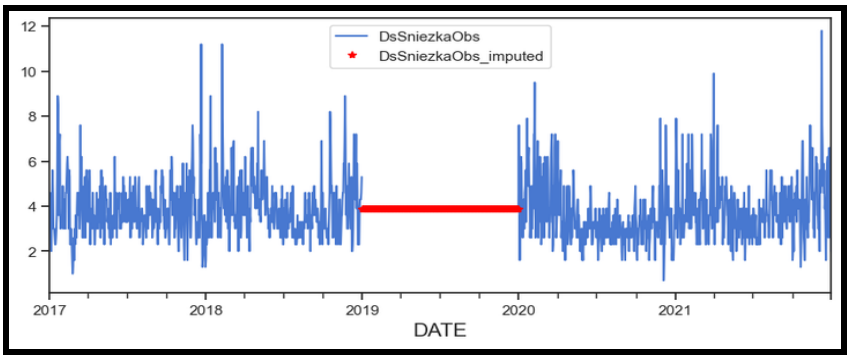
II. Imputation using MSTL (Multiple Seasonal-Trend decomposition using LOESS)



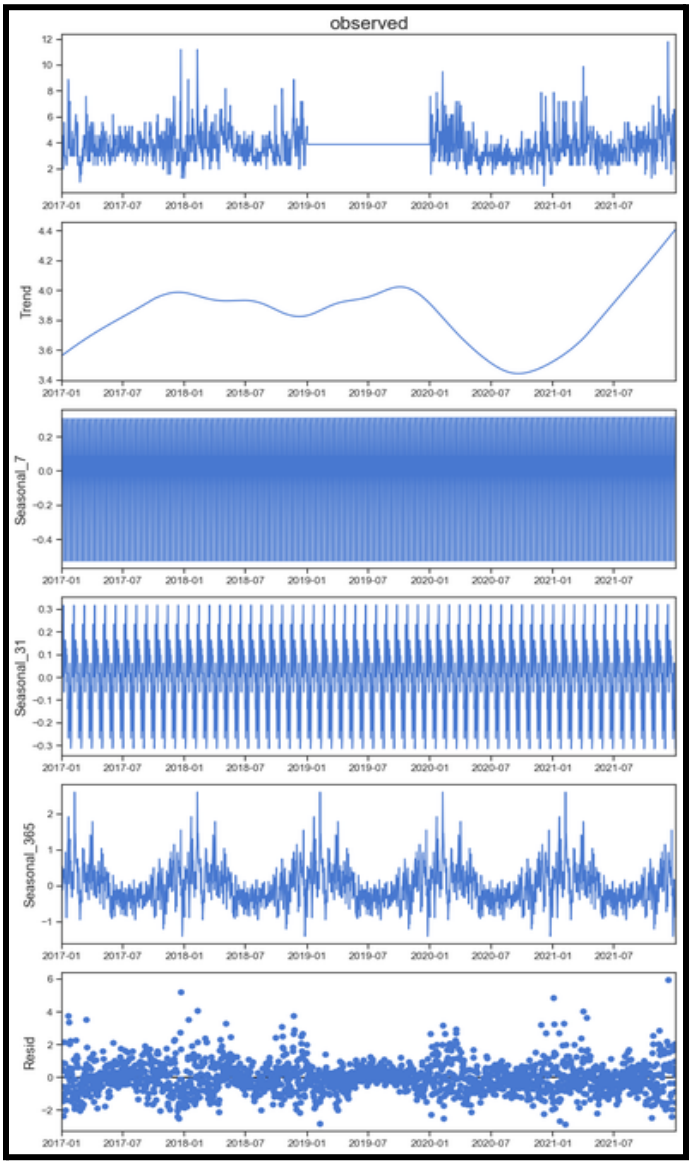
Example of a NO_2 measurement from an air quality monitoring station. The missing data is for all of 2019.

Station measurements with longer time periods of missing data (approximately 1 year) between 2017 and 2021 were imputed through the following steps:

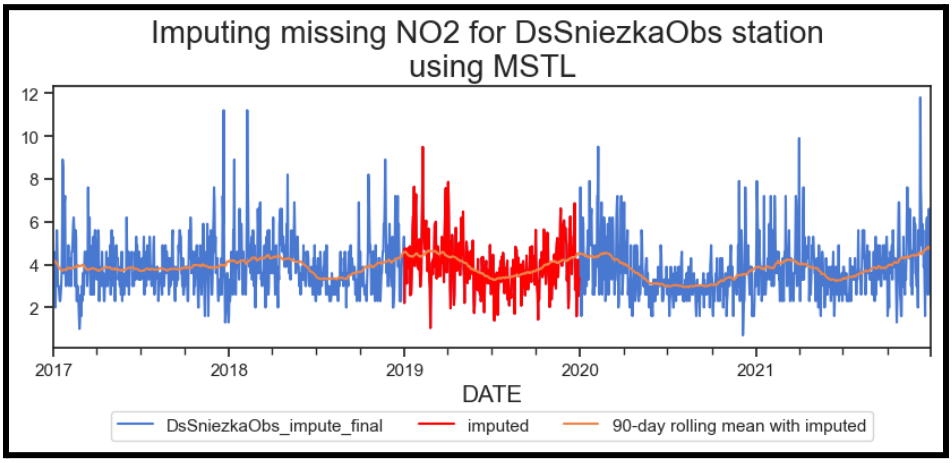
- First, the missing values were simply imputed using linear interpolation, as shown in the plot below.



- Then, the time series is decomposed to its trend, multiple seasonal (weekly, monthly and yearly seasonality) and residual components using MSTL package from statsmodels, as shown in the plot below.

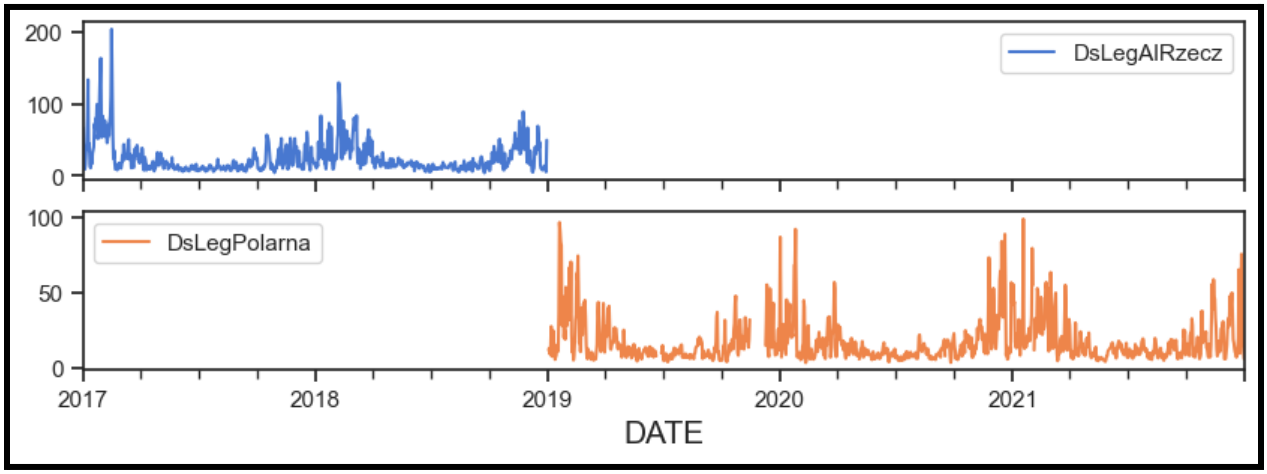


- Finally, those time periods containing missing values were imputed by replacing them with the sum of corresponding seasonal and trend components. The resulting imputed data is as shown in the plot below.



III. Combining nearby station measurements

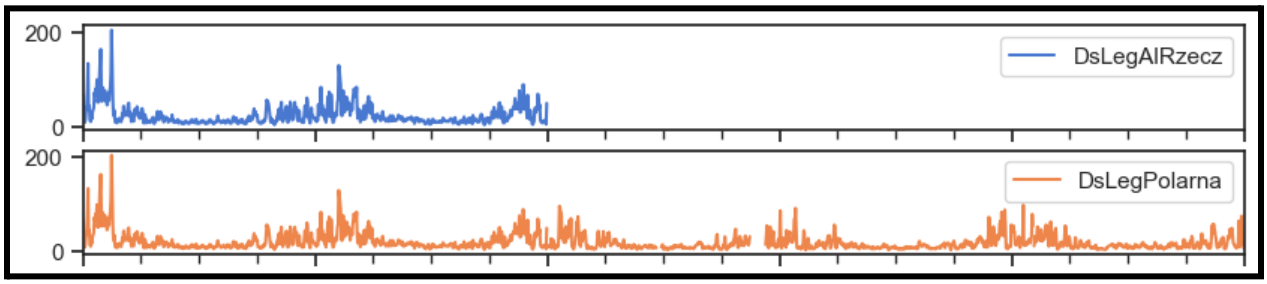
Some stations were observed to have measurements for a certain time period and then stop measuring for the rest of the year. Subsequently, another station starts measuring exactly at the time where the former station stopped measuring data, as illustrated below.



Example of a $PM_{2.5}$ measurement from two air quality monitoring stations located close to each other. This station is located in Powiat Legnica from the DOLNOŚLĄSKIE Voivodeship in Poland.

All of the air quality monitoring stations follows the same naming convention where the first two letters is the Voivodeship abbreviation, followed by the city/powiat abbreviation until the next capital letter appears. From the above plot, both of these stations have the same starting abbreviation of “DsLeg”, denoting these stations are located in Powiat Legnica from the DOLNOŚLĄSKIE Voivodeship. Comparing the station coordinates revealed that these stations were only apart by less than 5 km. Therefore, it made sense to impute the missing values of the

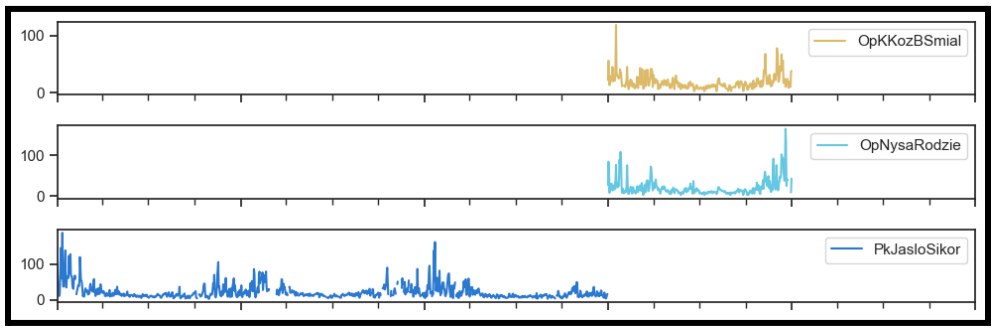
station containing the latest data with the station that stopped measuring data early, as shown in the plot below.



The station below that contained latest measurement is imputed using the measurement from the station above.

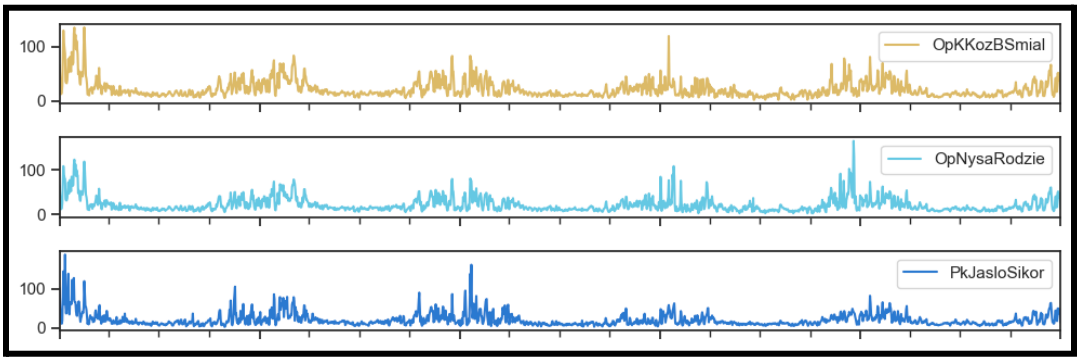
IV. Imputation using Rolling Mean

After implementing all of the above imputation methods, there were few more pollutant measurements with significantly large missing values during all periods of 2017, 2021 or more than one year of missing data, as shown in the plot below.



Remaining stations measuring $PM_{2.5}$ having large missing values.

All of these stations with large missing data were imputed by taking all of the imputed and non-imputed station pollutant data and then filling the missing values by taking 19-day rolling mean of the pollutant data and backfilling missing values, if any. The resulting imputed plots are as below.



Remaining stations measuring $PM_{2.5}$ having large missing values after imputation using 19-day rolling mean.

b. Subtask 2 - Weather Dataset

For the weather datasets, the first step taken to clean the data was replacing all of the outliers with null values. All the individual weather data column were accompanied by another column with the starting abbreviation “Q_” that indicated the quality of the data at row level. This column has three categories:

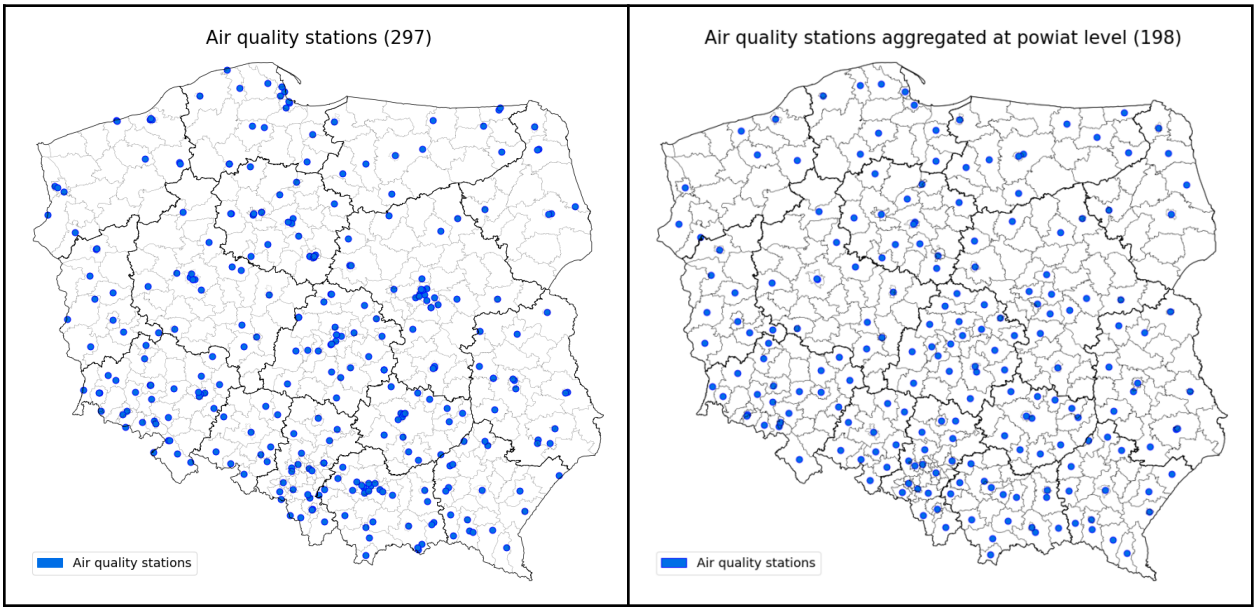
- **0**: Indicates the data for that row is correct.
- **1**: Indicates the data for that row is missing.
- **9**: Indicates the data for that row is a potential outlier.

For instance, the column for cloud cover is indicated as “CC”. Right next to this column is the quality column for cloud cover, indicated as “Q_CC”. All the row values in “CC” that are either 1 or 9 for the corresponding rows in “Q_CC” are replaced with null values.

Next, similar to the Air Quality dataset, the weather data missing values were either imputed using linear interpolation or using MSTL, depending on the size of missing values. There were some stations with no measurements for any years. These stations were ignored.

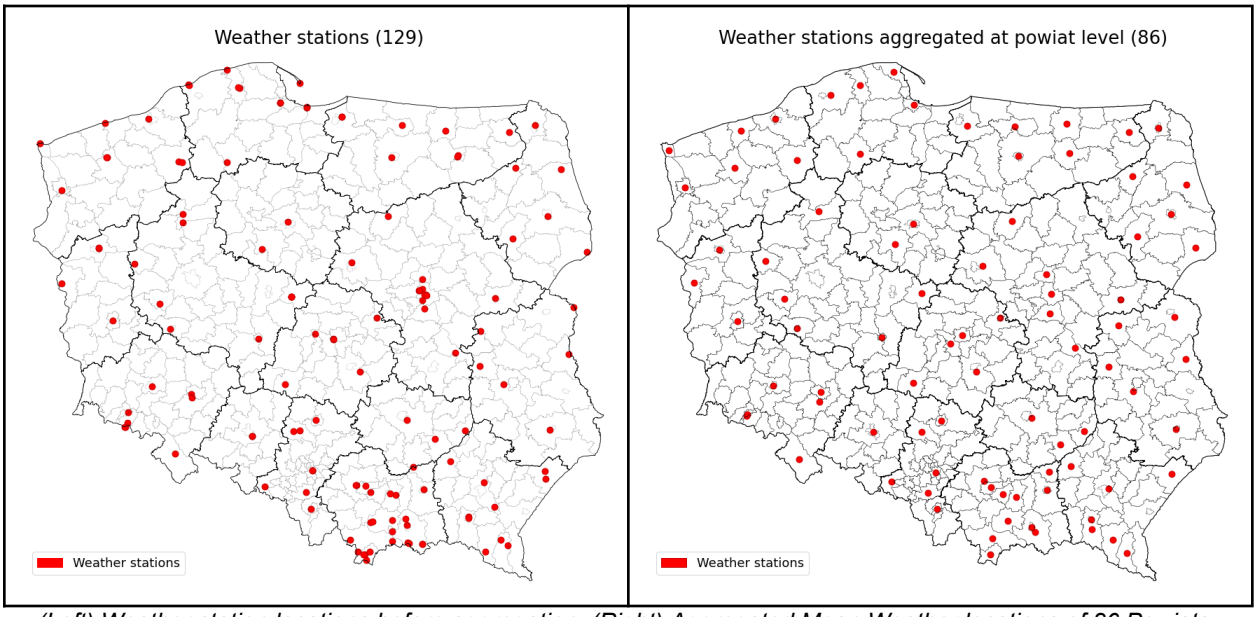
3.2. Merging Datasets

Following imputation, we aggregated both datasets at the Powiat level and took the mean of each corresponding pollutants. This way, we managed to reduce the hierarchy level of the time series datasets. For example, originally, the air quality dataset has time series data for 297 stations. After aggregating at the powiat level, the the number of time series data has been reduced to 198, as shown below.



(Left) Air Quality station locations before aggregation. (Right) Aggregated Mean Air Quality locations of 198 Powiats.

Similarly, the original weather dataset had time series data for 129 weather stations. After aggregating at the Powiat level and taking the mean weather measurement values, the the number of time series data has been reduced to 86, as shown below.



(Left) Weather station locations before aggregation. (Right) Aggregated Mean Weather locations of 86 Powiats.

However, we found a discrepancy in the number of Powiats between the two datasets. Specifically, the Air Quality dataset contained 198 Powiats, while the Weather dataset only had 86 Powiats. This mismatch presented a challenge when attempting to merge the datasets using a left join, as many Powiats would be left without any weather data. To address this, we decided to impute the missing weather data for these Powiats using data from the closest neighboring Powiats before joining.

To determine the closest neighboring Powiats, a separate Powiat-proximity dataset was created, as explained in section 2.5 under TASK 2 - Data Pre-processing. This included all 380 Powiats in Poland as columns, with all Powiats under each column sorted in order of closest distance to that particular Powiat. Using this dataset, we were able to retrieve the closest neighboring Powiat for each Powiat containing missing weather data and impute the weather data from the neighboring Powiat. Despite this approach, there were still some Powiats without any weather data. For these Powiats, we imputed the missing data using the average daily weather data of their corresponding voivodeships.

After imputation and aggregating, weather data was joined with the Air Quality dataset. After merging, the resulting dataset has data for 198 powiats. All weather datasets were filled. Finally, the static datasets were merged with the merged Air Quality and Weather dataset, creating a final single dataset.

3.3. Calculation of CAQI Values

The CAQI or Common Air Quality Index was proposed to facilitate the comparison of air quality in European cities in real-time. There are a number of compounds that constitute air pollution, including particulate matter (PM), which is mixture of solid and liquid particles suspended in air, as well as common gasses such as NO_2, CO_2 and O_3 to name a few. CAQI is a number on a scale from 0 to 100, and the higher the number the worse the air quality is. The overall air quality index for a certain day is based on the worst air quality index rating for the individual pollutants. CAQI index can be calculated if at least one pollutant measurement is available. Below reference table is used for calculating individual pollutant's sub-index and CAQI index values.

Qualitative name	Index or sub-index	Pollutant (hourly) concentration in µg/m^3			
		NO_2	PM_{10}	O_3	$PM_{2.5}$(optional)
Very low	0–25	0–50	0–25	0–60	0–15
Low	25–50	50–100	25–50	60–120	15–30
Medium	50–75	100–200	50–90	120–180	30–55
High	75–100	200–400	90–180	180–240	55–110
Very high	>100	>400	>180	>240	>110

Common Air Quality Index and pollutant's sub-indices (airly.org)

Here are the calculation steps. For each pollutant measurement data in each row:

- Obtain the range (a1, b1) of the corresponding pollutant from the table which contains the measurements.
- For the range (a1, b1) of the pollutant, check what is the range of Index or sub-index of CAQI from the table. Let this range be (a2, b2).
- Map the measurement 's' in the range (a1, a2) to the sub-index value 't' in the new range (b1, b2) using linear interpolation technique as follows:

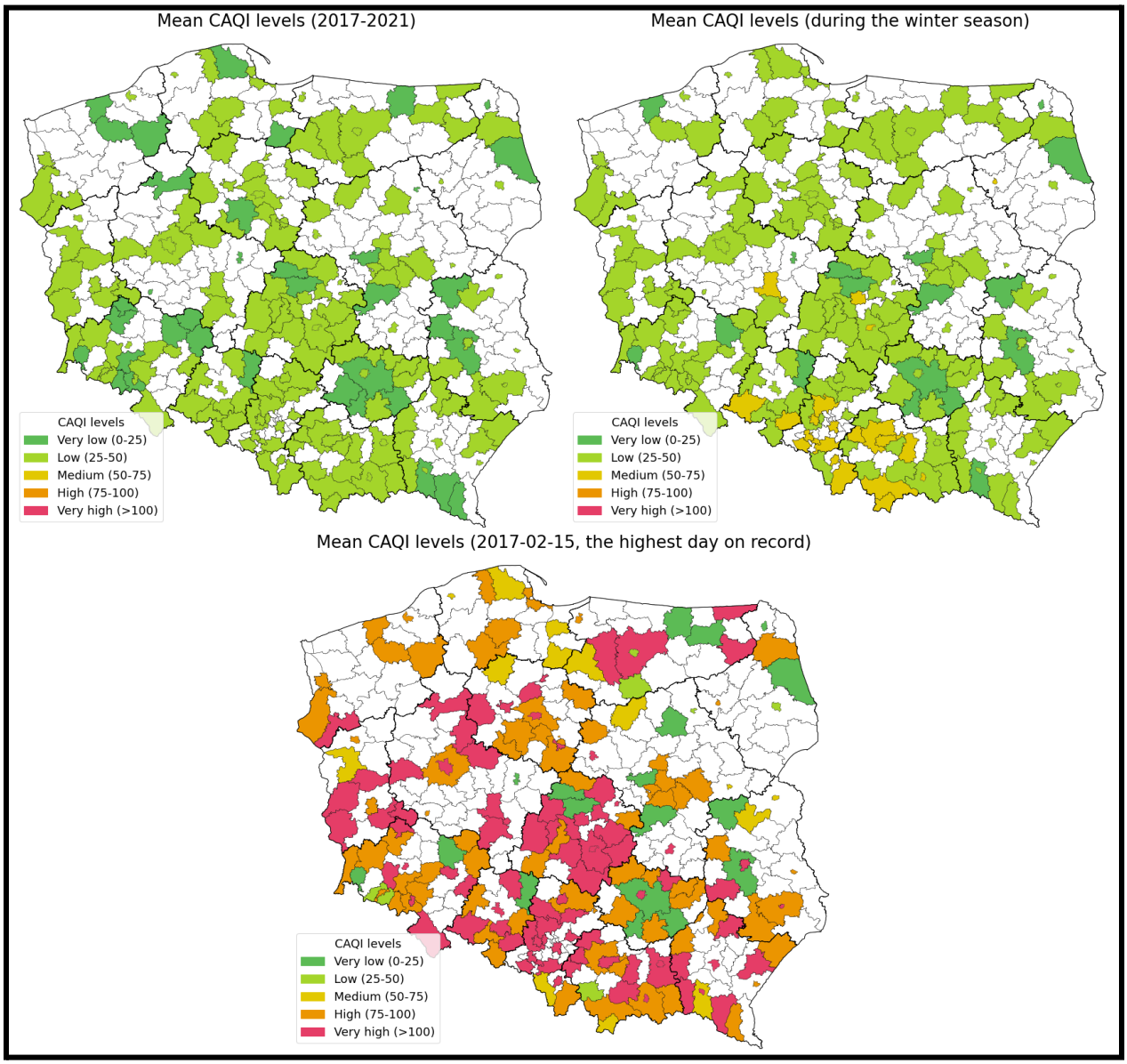
$$t = a2 + (s - a1) * ((b2 - a2) / (b1 - a1))$$

- Store the new value 't' in list 'L'
- After mapping each pollutant measurement from 's' to 't', calculating final CAQI index as follows:

$$CAQI\_index = max(L)$$ ***(Obtain maximum sub-index value from list 'L')***

- Store CAQI_index and all sub-indices in list 'L' in the data record.

For comparison, we plotted the mean CAQI levels for the entire studied period, only for the winter seasons and the day with the highest CAQI values (15-02-2017).



(Left): Mean CAQI levels from 2017 to 2021. (Right): Mean CAQI levels during Winter. (Bottom) Mean CAQI levels for 15-02-2017, the highest day on record

4. TASK 4 - Modeling

The following step after preprocessing and EDA is training and testing of various machine learning models to predict future CAQI level and indexes. Machine learning can be described as the development and utilization of a computer that can learn and adapt to data without providing explicit instructions.
Supervised Learning was chosen. This can be regarded as a machine learning technique in which a computer is provided an input or a set of inputs (often regarded as features) and an output (often regarded as labels) after which it would use a new set of unseen input to predict an output. The two main types of supervised machine learning paradigm are ***Classification***, which contains a categorical types label such as gender (male and female), grades (A, B, C, D, F) and so on and ***Regression***, which utilizes a continuous data type label such as housing prices, temperature and so on.

Supervised Machine learning is used for training and predicting CAQI levels and indexes in this project because the dataset contains features and labels of which the labels are the CAQI labels and indexes. Two supervised machine learning techniques which are classification and regression were used for training and prediction in this project because this modeling aims to predict two variants of labels which are

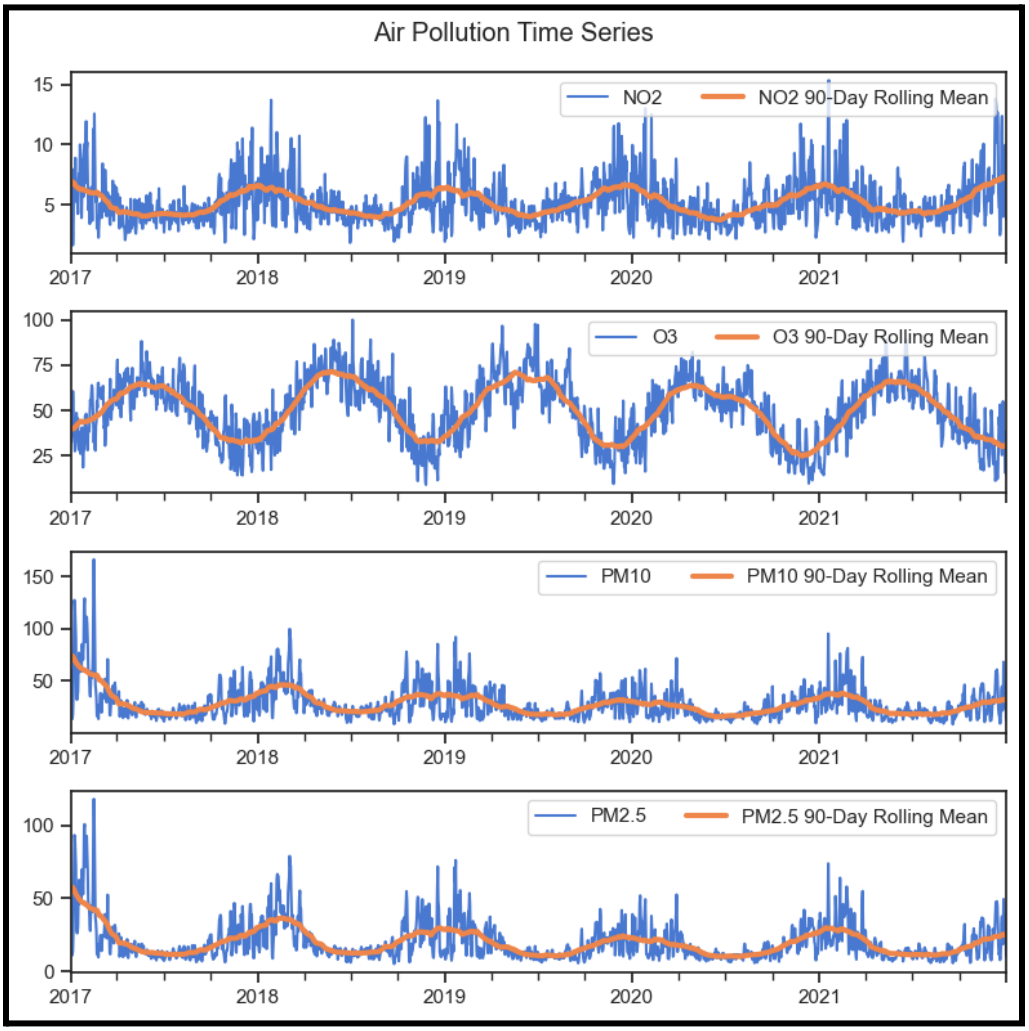
- CAQI labels: - which is a label with categorical data type which includes the following categorical variables: **vlow, low, medium, high** and **vhigh**.
- CAQI indexes: - this is a continuous data type label that indicates the CAQI value in a numerical manner.

From the EDA, we observed there is a temporal dependency and pattern in the dataset. Therefore for better generalization in training machine learning models, splitting of dataset for model training has to be carried out by taking the temporal dependency into account. The dataset was splitted based on specific time frames to ensure uniformity in performing train, test and validation split.

A variety of classification models were trained and tested in this project such as Random Forest Classifier, XGBoost Classifier, Logistic Regression, K-Nearest Neighbors, Support Vector Classifier, LGBM Classifier and so on to compare respective performance in the quest of identifying an optimal model. In addition, a variety of regression models were also trained and tested to predict CAQI indexes in this project such as XGBoost and so on to compare performances for optimal model identification.

RESULTS AND INSIGHTS

1. Analyzing Pollutant Data



Aggregated Daily Mean Time Series of Pollutant concentration from 2017 to 2021.

The above plot represents the daily aggregated mean time series of all four pollutant concentrations from 2017 to 2021. There is a clear repeating pattern for all pollutants every year where pollutants are higher during certain parts of the year. The formation of the four pollutants are briefly explained below.

- **Particulate Matter (PM_{10} & $PM_{2.5}$)**

  According to the United States Environmental Protection Agency (EPA.gov), "*PM_{10} are inhalable particles, with diameters that are generally 10 micrometers and smaller; and $PM_{2.5}$ are fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller.*"

Sources of both PM_{10} and $PM_{2.5}$ emissions include combustion of solid fuels such as coal and wood, road traffic which leads to tyre abrasion and street dust, dust from construction sites or landfills, and agricultural activities primarily related to livestock.

The formation of low emissions smog is mainly due to the use of outdated stoves, low-quality fuel, and waste incineration. Internal combustion engines are responsible for a certain percentage of suspended dust and PAH emissions.

Industrial plants, including coking plants, steel mills, oil refineries, and coal-fired power plants, as well as chipboard plants, remain a significant source of air pollution in many places in Poland. These plants emit particulate matter, PAHs, nitrogen and sulfur oxides, as well as toxic heavy metals like arsenic, mercury, cadmium, and lead.

The severity of symptoms depends largely on the concentration of dust in the air, exposure time, additional exposure to environmental factors and increased individual susceptibility (children and the elderly, coexistence of chronic heart and lung diseases). As some dust components can enter the bloodstream, prolonged exposure to high concentrations of dust can have a significant impact on the course of heart disease (hypertension, heart attack) or even increase the risk of cancer, especially of the lungs. (IQAir.com).

- **Nitrogen dioxide (NO_2)**

The main contributors to nitrogen oxides (as NO_2) emission to air in Poland are combustion processes in energy production, manufacturing industry and road transport. Due to a fast growing number of cars and problems with traffic fluency in urban areas, the road transport has a significant and growing influence on NO_x concentrations in air in Polish cities (EMEP). The use of diesel engines in transportation contributes significantly to the emission of nitrogen oxides. NO_x contributes to the formation of ozone and particulate matter.

NO_2 is associated with adverse effects on health: it can affect the liver, lung, spleen and blood. It can also aggravate lung diseases leading to respiratory symptoms and increased susceptibility to respiratory infection. As with SO_2, NO_x contributes to acid deposition but also to eutrophication of soil and water (europa.eu).

- **Ozone (O_3)**

Ozone is a molecule made up of three oxygen atoms. Ozone is formed when heat and sunlight cause chemical reactions between oxides of nitrogen (NO_x) and Volatile Organic Compounds (VOC), which are also known as Hydrocarbons. This reaction can occur both near the ground and high in the atmosphere.

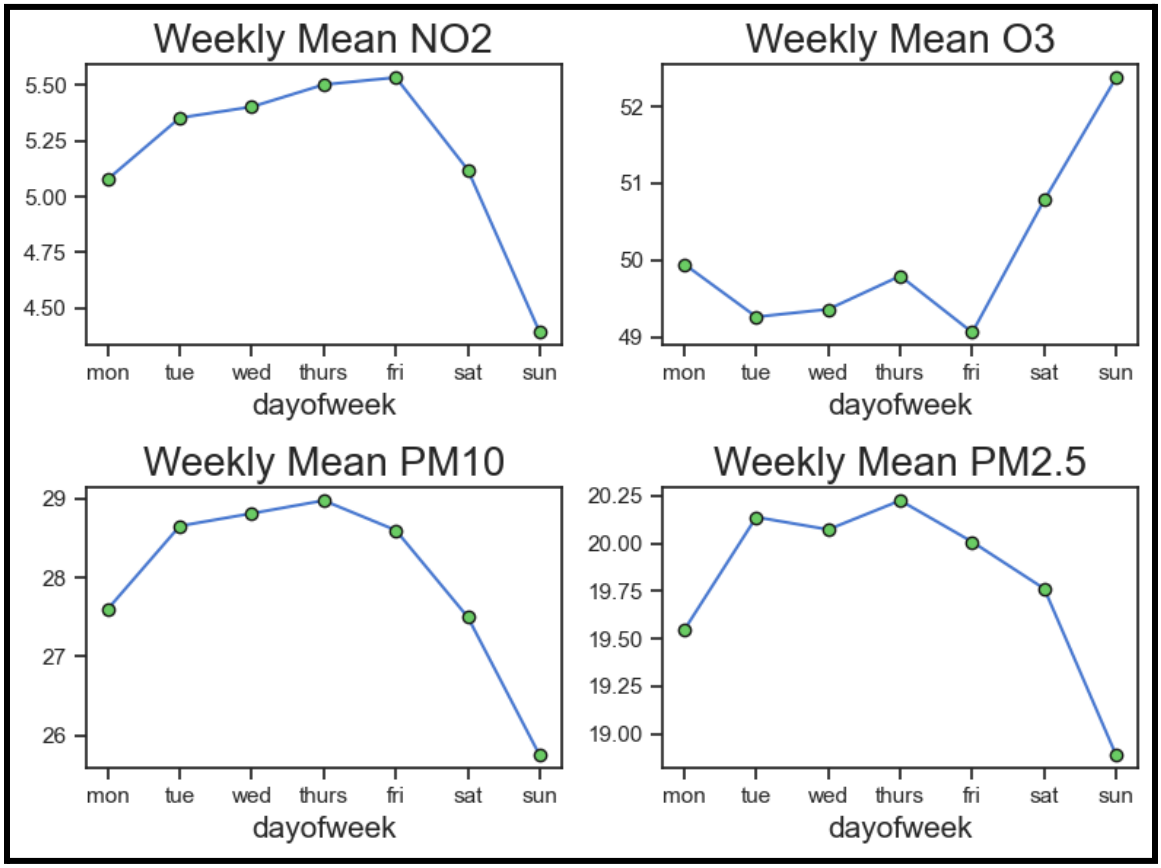
It is important to note that Stratospheric Ozone, which is formed about 10-30 miles above the Earth's surface, is good Ozone and it forms a protective layer, called the ozone layer, that shields us from too much of the sun's harmful ultraviolet radiation (UV).

Conversely, ground level Ozone harms human health and the environment. The most significant factors that form ground-level ozone are NO_x and VOCs (from mobile source emissions and industrial processes), and UV radiation (from sunlight) (scdhec.com).

Elevated levels of ozone can cause respiratory health problems, including decreased lung function, aggravation of asthma, and other lung diseases. It can also lead to premature mortality. Ozone is also a greenhouse gas contributing to warming of the atmosphere (europa.eu).

1.1. Weekly Seasonality of Pollutants

The plot below show weekly seasonality of pollutant concentration. O_3 concentrations are higher during the weekends and lower during weekdays. Conversely, other pollutants shows the opposite patterns. This potentially indicates weekly pollutant concentrations depends on human activities.



Aggregated Weekly Mean Pollutant Concentrations

More than 85% of Polish people identify themselves as catholics and most people attends weekly church service on Sunday. Most shopping malls, supermarkets and smaller shops are closed on Sundays, which could potentially result in less traffic. This could explain lower weekly pollutant concentrations for NO_2, PM_{10} and $PM_{2.5}$.

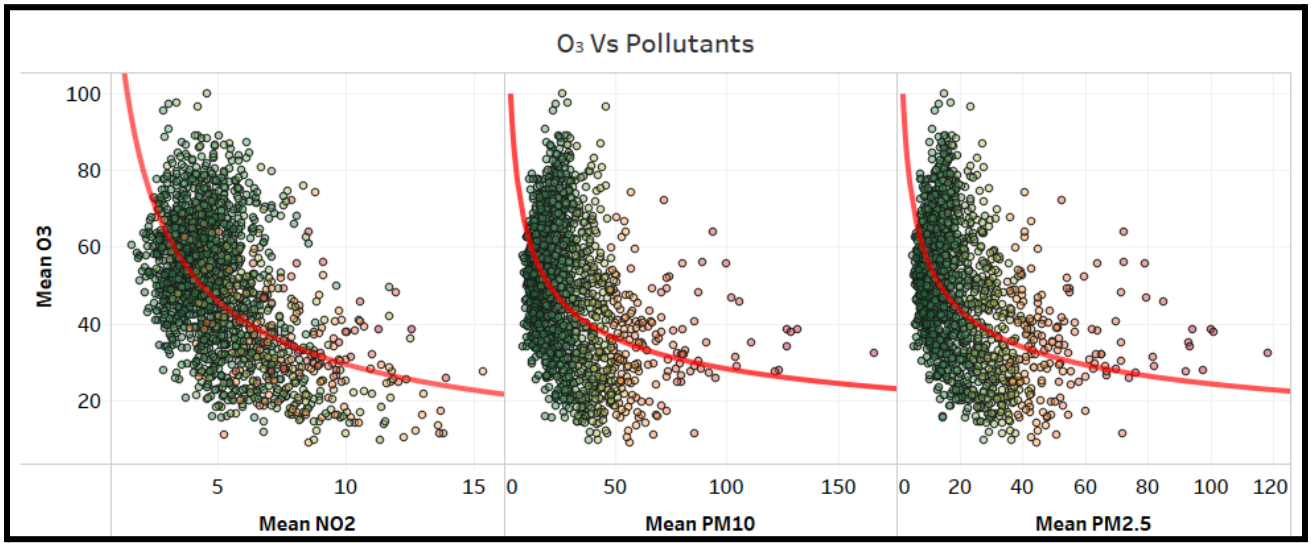
Conversely, O_3 concentrations are higher during the weekend. This phenomenon is known as "weekend effect". Ozone is a secondary pollutant, which means it is not directly emitted by traffic, industries, etc. but is formed on warm summer days by the influence of solar radiation on a cocktail of airborne pollutants. These Ozone precursors are Nitrogen Oxides (NO_x) and volatile organic compounds (VOC). Smog is formed when pollutants such as nitrogen oxides and VOCs are emitted into the lower atmosphere and react with sunlight to create ozone.

$$\boldsymbol{NO_2 + O_2 + (solar\ UV\text{-}light, + heat) \rightarrow NO + O_3}\ (\textit{irceline.be})$$

According to various studies, NO_x at higher concentrations tends to degrade O_3 through the process called NO_x titration. This consists of the removal of O_3 through reaction with NO.

$$\boldsymbol{NO + O_3 \rightarrow NO_2 + O_2}\ (\textit{irceline.be})$$

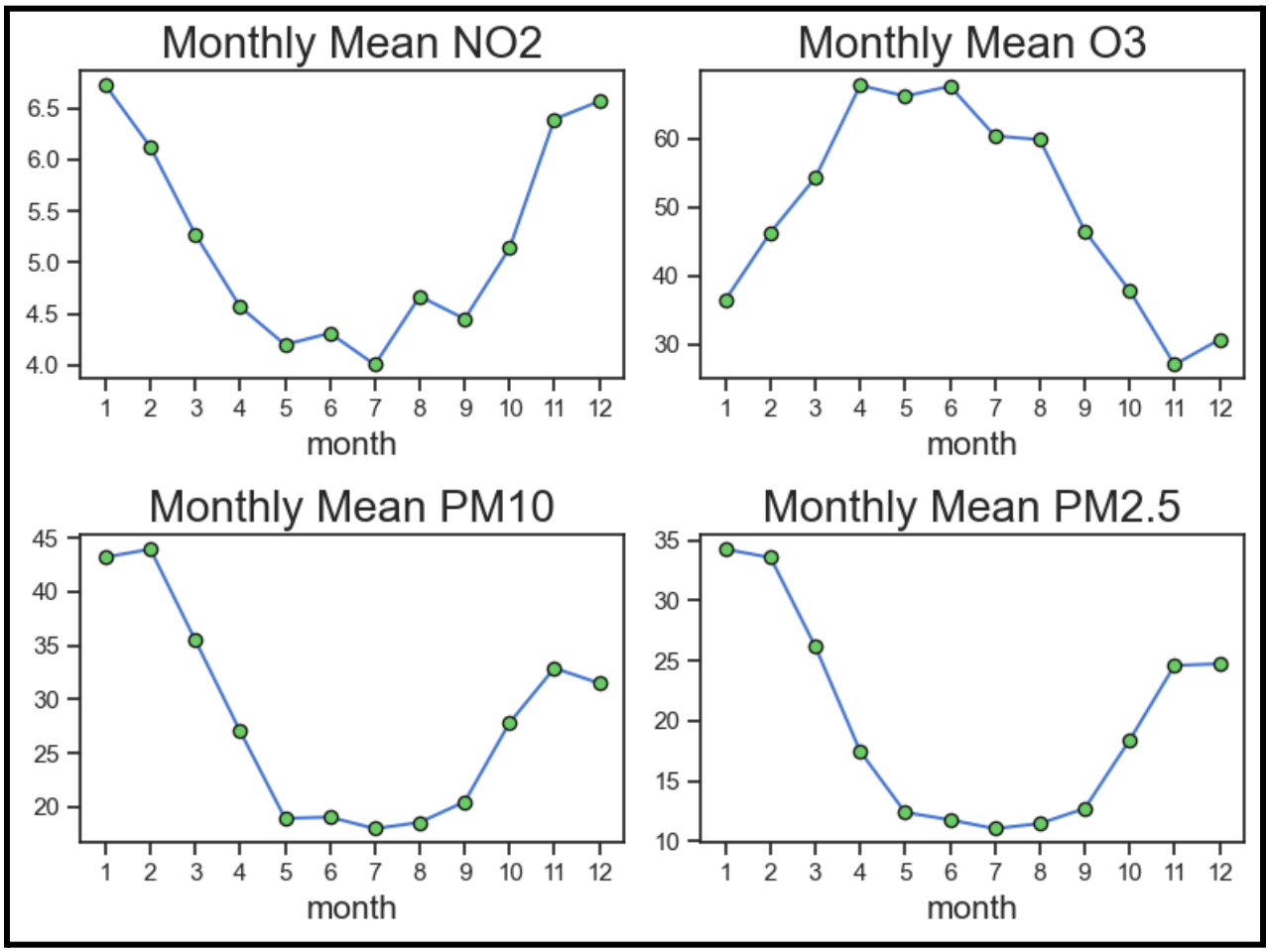
Conversely, when there are less traffic during weekends, there is a reduction of NO_x emissions. This subsequently does not suppress O_3 as shown above, hence leading to higher concentration of O_3 during weekends.



Scatterplot of O_3 vs NO_2, PM_{10} and $PM_{2.5}$

The above plot illustrates the non-linear relation between O_3 and other pollutants. As the mean NO_2 increases, mean O_3 levels remain low and vice-versa.

1.2. Monthly Seasonality of Pollutants



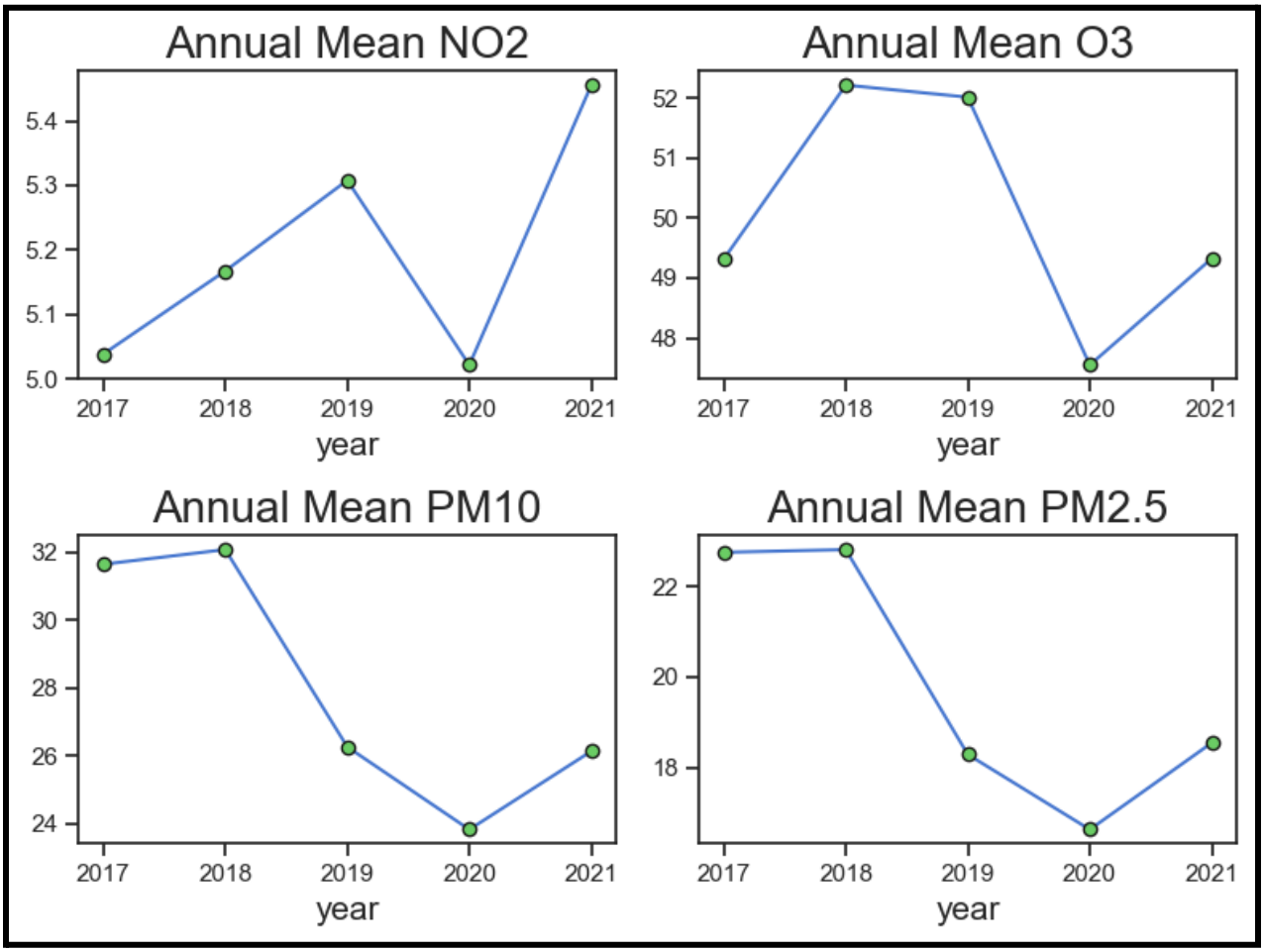
Aggregated Monthly Mean Pollutant Concentrations

The above plots illustrate the aggregated monthly mean pollutant concentrations. O_3 levels are higher during spring and summer months. Conversely, the rest of the pollutants are higher during Autumn and Winter months. Higher concentrations of O_3 during spring and summer is a result of effective photochemical ozone production favored by high temperatures and intensive solar radiation during those months. On the other hand, high levels of NO_2 and particular matter emissions during Autumn and Winter months is mostly attributed to intensive burning of low-quality coal in coal furnaces for heating (Izabela Pawlak).

1.3. Annual Mean Pollutant Concentrations

The plots below depict the annual mean concentration of pollutants from 2017 to 2021. The pollutant matter concentrations were highest during 2017 and gradually decreased from 2018 onwards. NO_2 levels were steadily increasing from 2017 to 2019. O_3 levels were highest during 2018, followed by 2019. Interestingly, all four pollutants were lowest during 2020 before picking up again in 2021. This could indicate the effects of nationwide lockdown in Poland during the

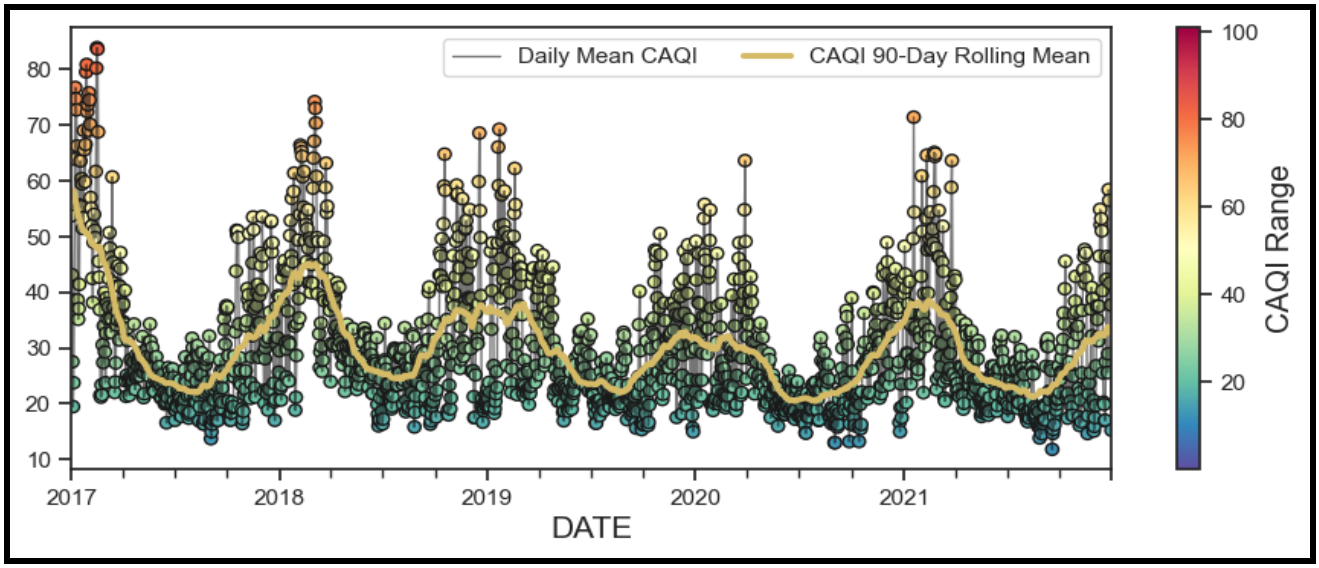
COVID-19 pandemic. The highest reduction in concentrations were observed during April and May 2020. Less economic activity, closure of high power-consuming plants, suspension of air and railway traffic, reduction of car traffic, decrease of power production all led to a reduction of emissions into the atmosphere resulting in a marked improvement in air quality (ncbi.nlm.nih.gov).



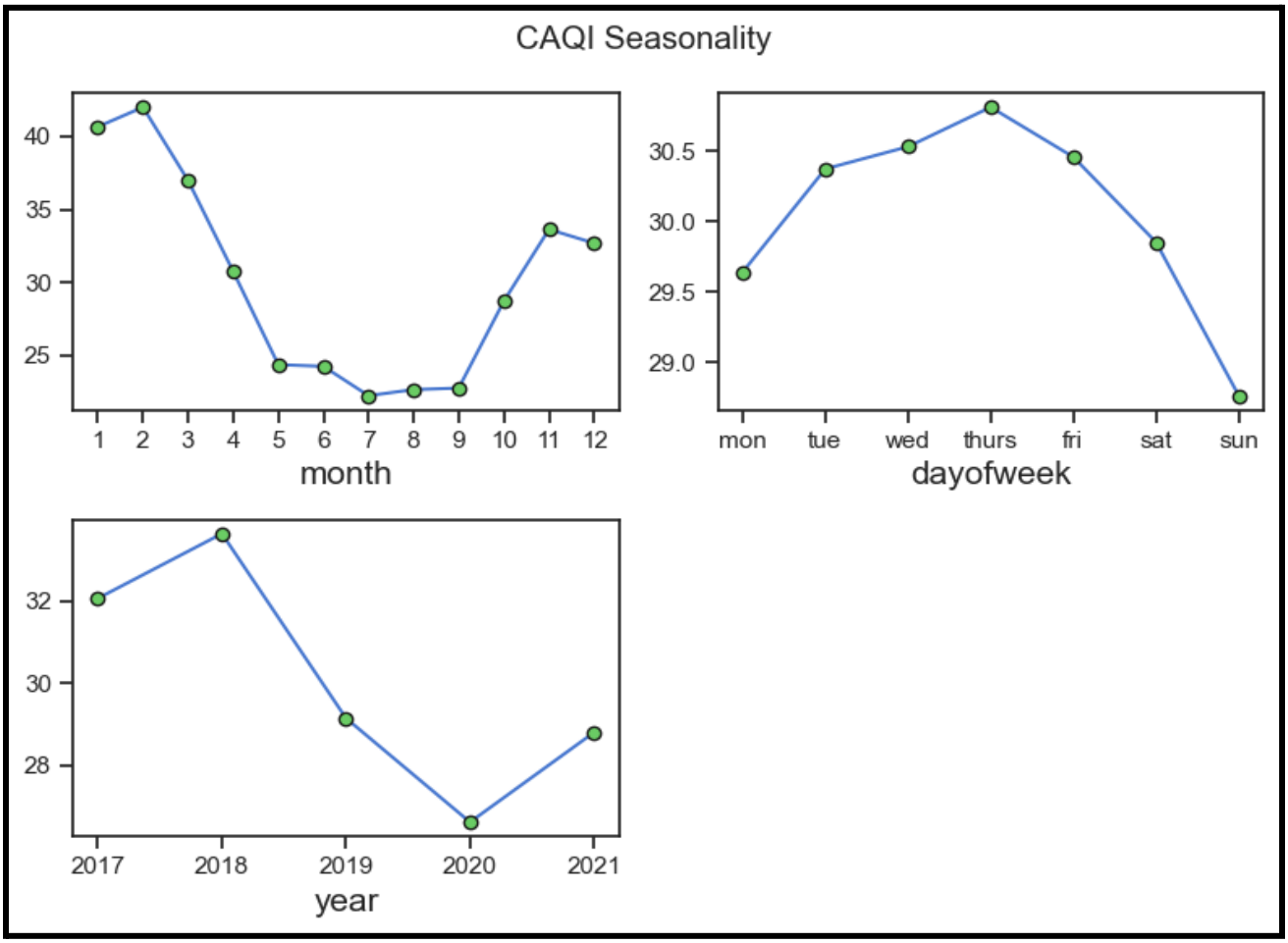
Annual Mean Pollutant Concentrations

2. Analyzing CAQI (Common Air Quality Index)

CAQI is used for easily understanding the general Air Quality in Europe based on all of the above pollutants. CAQI calculation is elaborated in section 3.2 under METHODOLOGY & TOOLS. The plot above illustrates daily mean CAQI in Poland from 2017 to 2021. The 90-Day Rolling Mean of CAQI trends shows that the CAQI has a downward trend in general, the lowest being during 2020. The plots below depict the weekly, monthly and yearly trends of CAQI values.



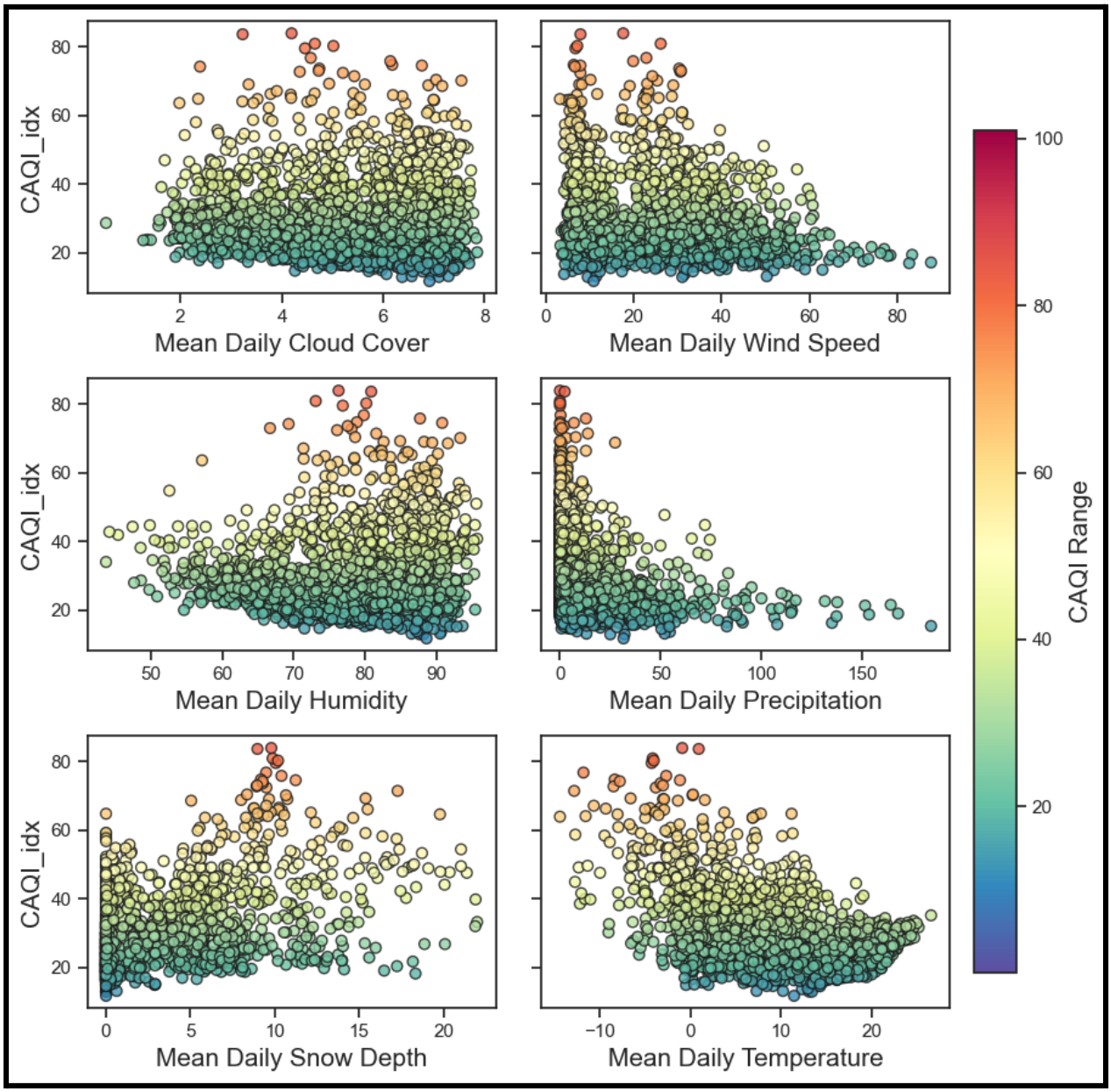
Aggregated Daily Mean CAQI values from 2017 to 2021



Weekly & Monthly Seasonality of CAQI levels, followed by Annual Trend.

In general, CAQI values are lowest on the weekends compared to weekdays. On a monthly level, CAQI levels are lower during Spring and Summer. Then, CAQI levels gradually increase during Autumn, before reaching highest levels during Winter. CAQI level Annual trends are similar to those we have seen before for particulate matter, where CAQI levels are seen to be at lowest during 2020.

3. CAQI Levels vs Weather

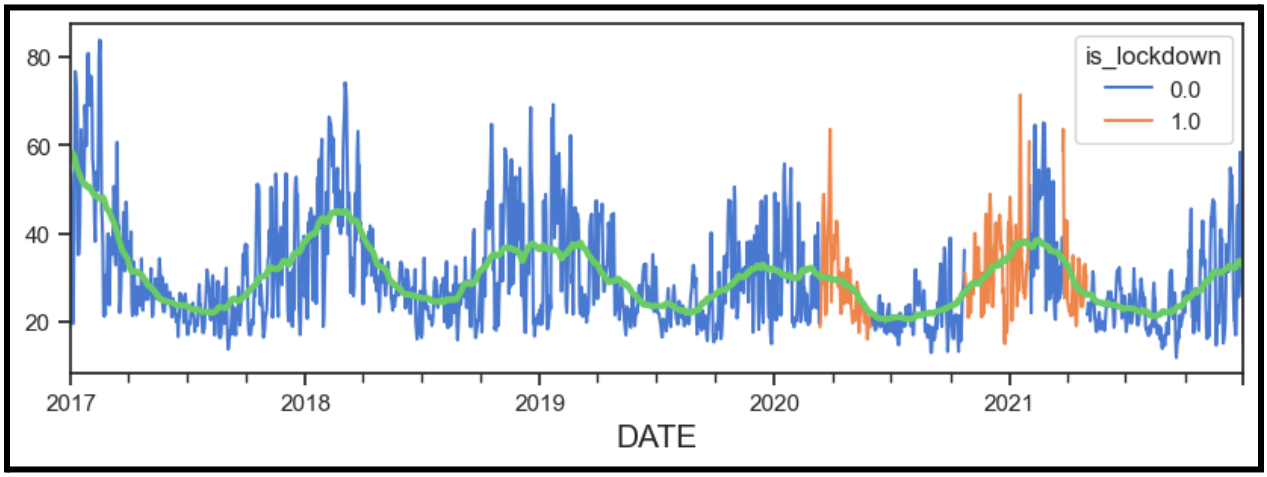


Scatterplot of daily mean CAQI levels and various weather conditions

The above plots depict how various weather conditions affect CAQI levels in Poland. In general:

- Higher CAQI levels indicate worser Air Quality. CAQI levels are evenly distributed from low to high levels of Cloud Cover, which suggests, Cloud Cover alone may not have a huge impact on CAQI levels.
- Groups of high CAQI levels are located at the lower Wind Speed ranges. The higher the Wind Speed, the lower the CAQI.
- CAQI levels are noticeably higher at higher Humidity levels.
- CAQI levels are observed to be higher at only lower Precipitation levels.
- CAQI levels are comparatively higher at higher Snow Depth levels. But temperature could be a confounding variable here as snow only appears when the temperatures are extremely low.
- There is a clear non-linear relation between Temperature and CAQI values, where CAQI levels are higher at lower Temperatures. This could be attributed to the use of coal burners for heating during Winter in Poland. At higher Temperatures, CAQI levels are significantly lower, indicating better Air Quality in Spring and Summers.

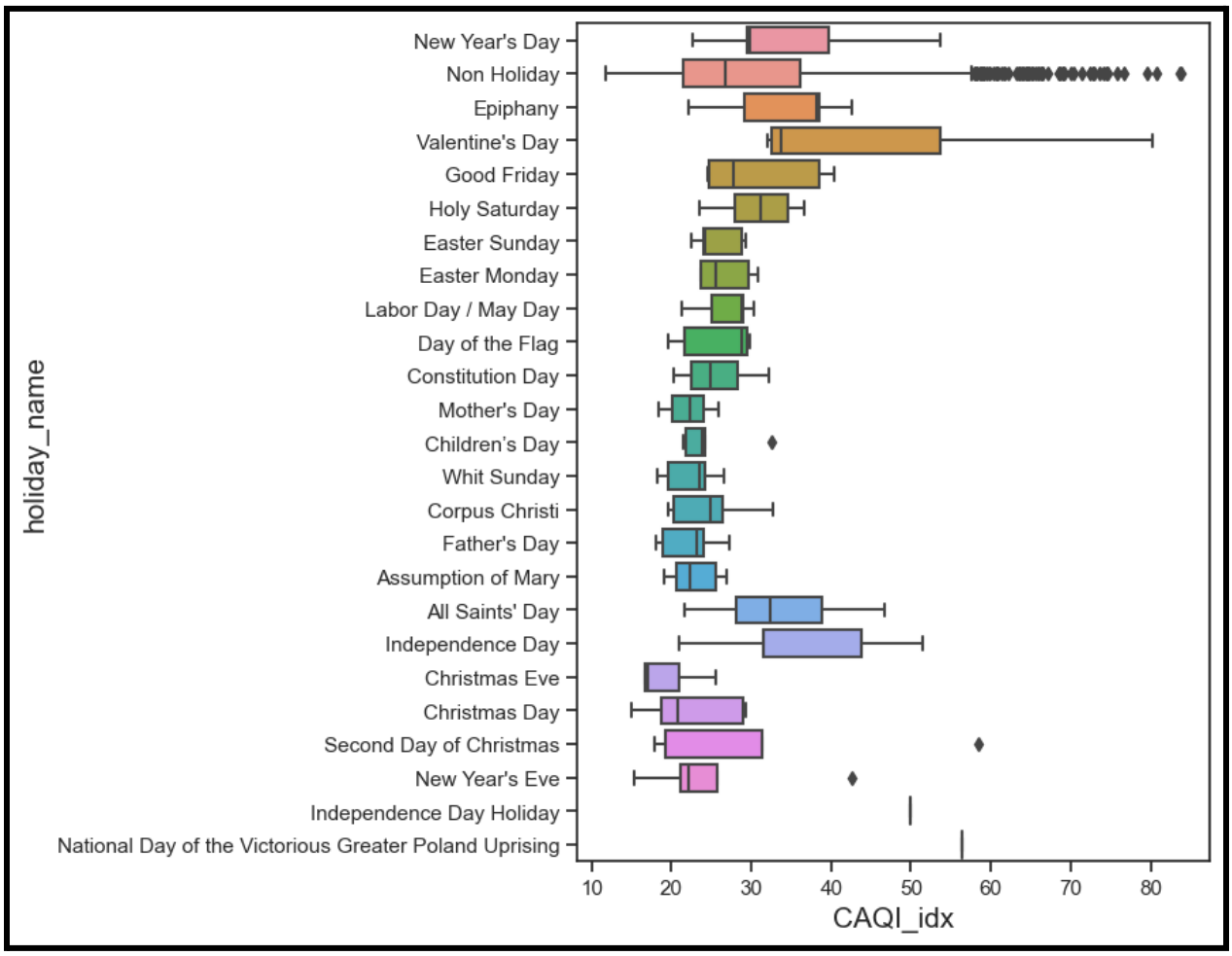
4. Impacts of COVID-19 Lockdown



Daily CAQI Timeseries. COVID-19 Lockdown periods are indicted in orange. The green line represents 90-Day roling mean of CAQI

The first case of COVID-19 in Poland was registered on March 4th, 2020. Governmental measures significantly restricted social and economic activities. As we have seen in the previous section, 2020 saw the lowest CAQI levels in Poland. The orange shaded parts in the above plot depicts the time periods where Poland was under lockdown. One thing to note is that the first lockdown fell during the Spring season, which historically has lower CAQI levels. When observing the 90-Day rolling mean of CAQI levels, we can see there is already a drop in CAQI levels way before the COVID-19 lockdown during 2020. Subsequent lockdowns later in year were during Autumn and Winter season, which historically sees an increase in CAQI levels as more people use coal stoves for heating.

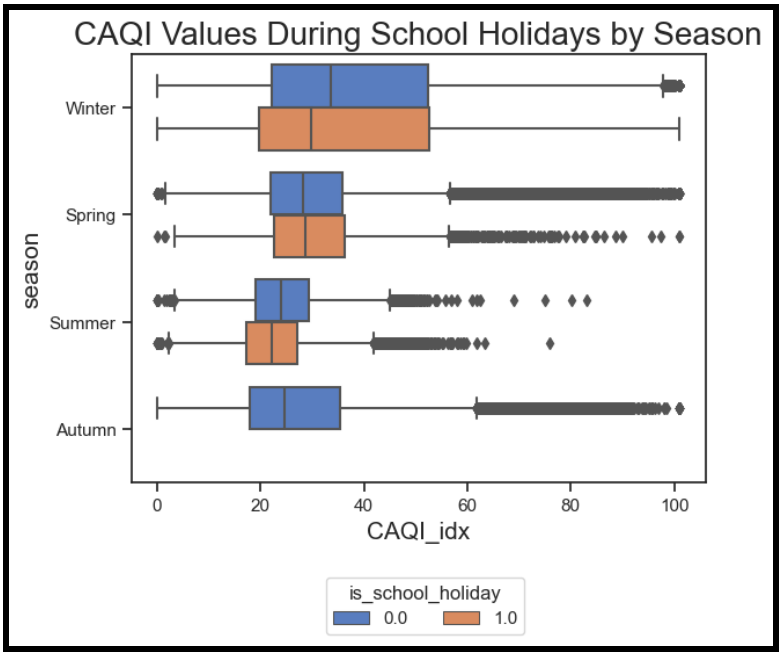
5. Impact of Public Holidays and School Holidays



Boxplot distribution of CAQI levels for various Public Holidays and observances in Poland

The above boxplot illustrates how CAQI levels vary by different holidays in Poland. Non-Holiday category are those days that are not a public holiday. Certain holidays such as *Valentine's Day, All Saint's Day, Independence Day* and *National Day of the Victorious Great Poland Uprising* see a higher median level of CAQI levels compared to other days.

The times of year when there is a school holiday also has an effect on overall CAQI levels in Poland, as depicted in the boxplots below. During Summer and Winter, median CAQI levels are comparatively lower during school holidays.



Boxplot distribution of CAQI levels during School Holidays by season. Orange boxplots are school holidays and blue boxplots are Non-school holidays

6. Modeling

Upon preprocessing of the dataset and gaining some insights into existing data through EDA, the entire dataset was used to generate two main datasets which are for training machine learning models for both classification and regression models. The classification dataset was used to predict a class of CAQI levels while the regression dataset was used to predict CAQI indexes as continuous variables. To achieve optimal models performance and to ensure consistency in splitting training and testing datasets, a structure was decided by the team for uniformity in splitting datasets which can be found below: -

- Train: 1st Jan 2017 to 28th Feb 2021
- Test: 2nd Mar 2021 to 31st Dec 2021

To improve model performance, a variety of techniques were utilized such as: -

- Feature Selection: - for identification and selection of important features
- Feature engineering: - can be described as a process of transforming existing features into more important datasets important for current task
- Oversampling: - for tackling feature imbalance in classification dataset
- Cross validation: - to test and visualize how various models perform under various conditions on the same dataset
- Hyperparameter optimization: - for identification of optimized training parameters for selected models before retraining.

- Feature importance: - this is used to identify the importance of various features used in training a model.

6.1. Feature Engineering

Feature engineering involves leveraging data to create new features in the dataset, with the goal of simplifying and speeding up data transformations while also enhancing model accuracy. The following features were created:

I. **Lag Feature**: Lag features are values at prior timesteps. They are created on the assumption that what happened in the past can influence or contain a sort of intrinsic information about the future. For this dataset, a lag 1 of CAQI index (CAQI values of the previous day) was created.

II. **Rolling window Mean**: Computed Rolling Mean of CAQI index from previous 30 days.

III. **DateTime features**: Extracted additional features from the 'DATE' column sich as Day, Month, year and day of week. Additionally, a new feature called 'is_weekend' was created that denotes Saturday and Sunday as True and the rest of the days as False.

IV. **Cyclical Features**: Day, Month and Day of Week features are cyclical in nature. In these cases, the higher values of the variable are closer to the lower values. For example, December (12) is closer to January (1) than to June (6). By applying a cyclical transformations, that is, with the sine and cosine transformations of the original variables, we can capture the cyclic nature and obtain a better representation of the proximity between values (Feature Engine).

V. **Combining Trasnformations**: Many static features were combined by taking the sum of their values in order to represents the feature in a parse manner.

VI. **Categorical Encoding**: Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the models to give and improve the predictions.

Additionally 2 sets of datasets were created:

- Temporal Dataset - These dataset contains the temporal elements of the datasets.
- Non-Temporal Datasets - These datasets does not contain any temporal elements such as the date features, lag features, rolling mean features, etc. The main aim of this is to check how the model generalizes to features when time components are removed.

6.2. Supervised Learning

The chosen metrics to evaluate model performances were:

- **F1 Score (Classification)** - Interpreted as a harmonic mean of the precision (ratio of true positive to the sum of true positive and false positive) and recall (ratio of true positive to the sum of true positive and false negative), where an F1 score reaches its best value at 1 and worst score at 0.
- **Root Mean Squarred Error (Regression)** - RMSE is a quadratic scoring rule which measures the average magnitude of the error (i.e, the difference between the actual value and the predicted value).

For predicting CAQI classes, a variety of classification models were trained and tested by various collaborators such as: -

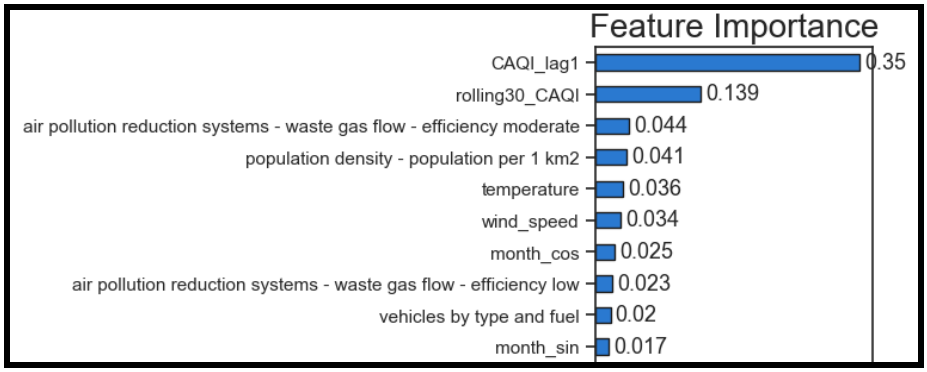
S/N	Model	Dataset Type	F1 Score
1	RandomForest Classifier	Temporal	0.70
2	Logistic Regression	Non-Temporal	0.613
3	XGB Classifier	Non-Temporal	0.675
4	LGBM Classifier	Non-Temporal	0.685
5	LGBM Classifier	Temporal	0.75

For predicting CAQI indexes, a variety of regression models were trained and tested by various collaborators such as: -

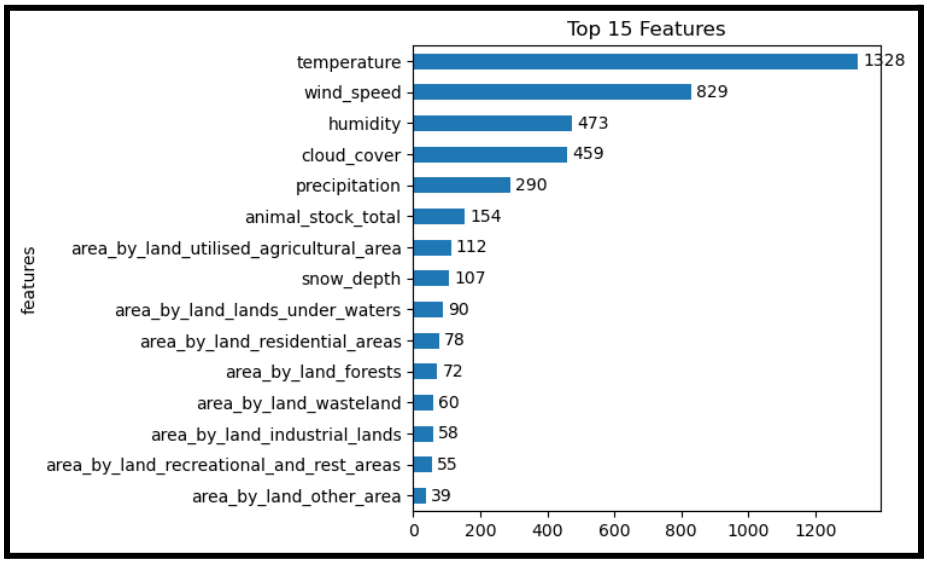
S/N	Model	Dataset Type	RMSE
1	XGBoost	Temporal	7.718
2	XGBoost	Non-Temporal	10.789
3	Random Forest	Non-Temporal	12.390

6.3. Feature Importances of Models

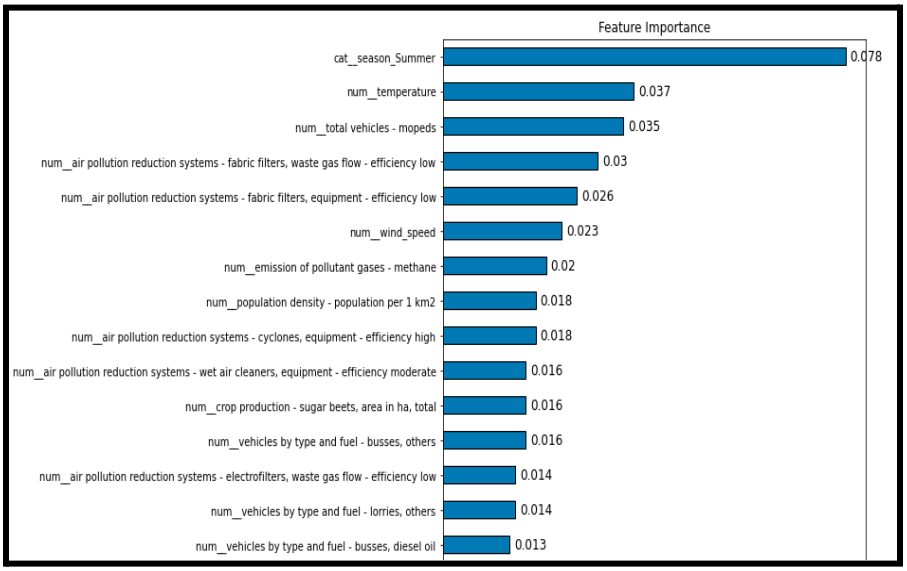
Tree based machine learning algorithms such as Random Forest, XGBoost, etc. come with a feature importance attribute for each feature representing how useful the model found each feature in trying to predict the target. Below are the feature importance plots of various models on both Temporal and Non-Temporal datasets.



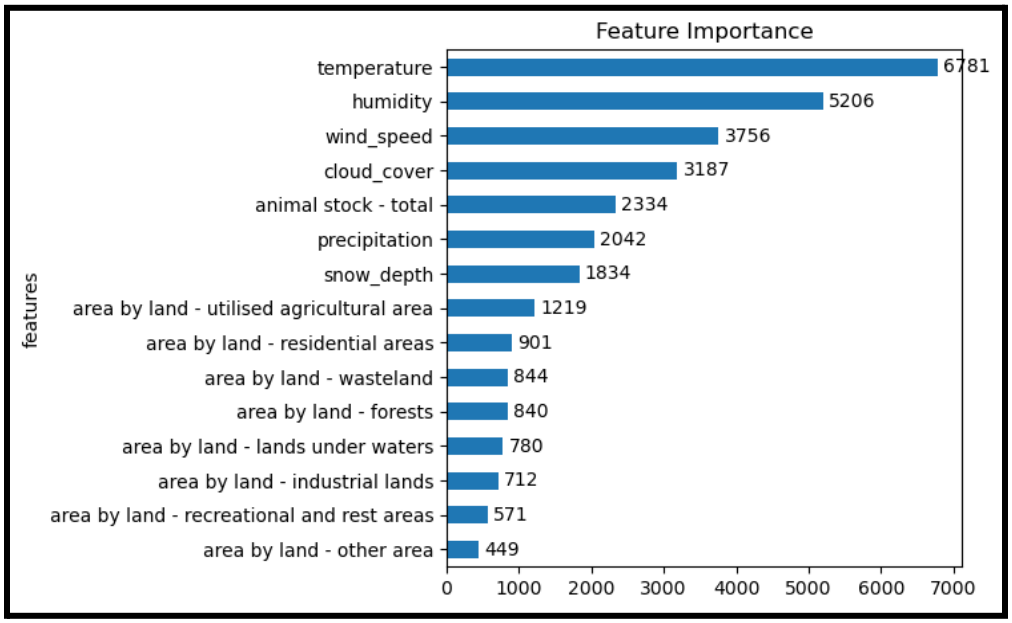
Top 10 Features of XGBoost regressor on Temporal Regression dataset



Top 15 Features of LGBM Classifier on Temporal Classification dataset



Top 15 Features of XGBoost regressor on Non-Temporal Regression dataset



Top 15 Features of LGBM Classifier on Non-Temporal Classification dataset

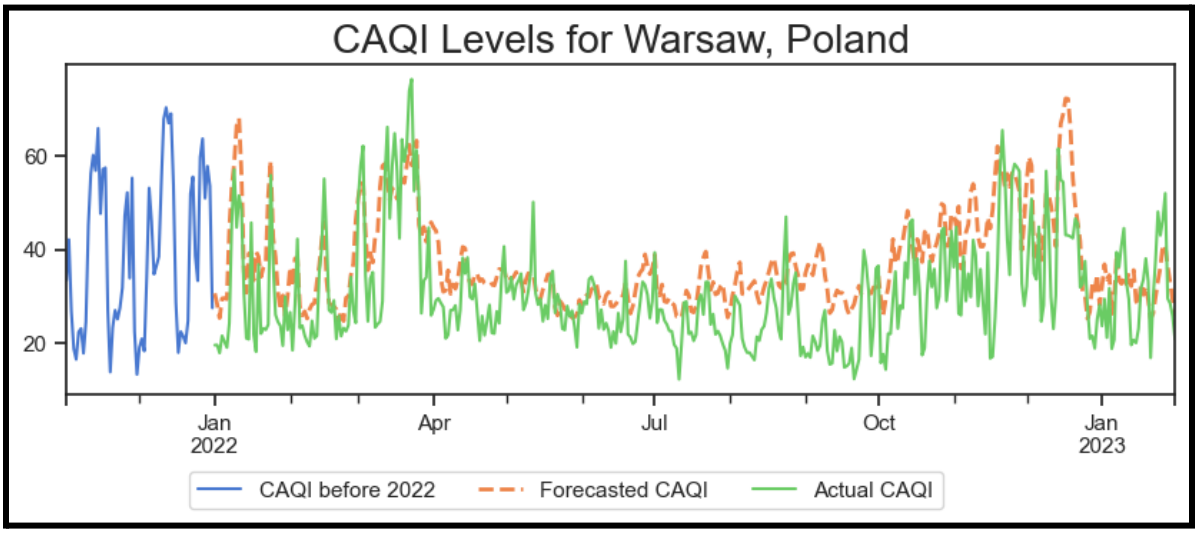
It is important to note that feature importance varies across all models and type of dataset. From the above plots, the features that are commonly ranked the highest seem to be the weather variables, especially Temperature and Wind Speed. As we have seen from earlier EDA, weather do play an active role in CAQI levels. Other variables related to human activities such as air pollution reduction systems, land area, etc. are also ranked high, but these do vary across models.

The EDA has revealed that man-made activities are the main source of pollution in Poland. Weather conditions only have an indirect effect on the overall air quality. Data related to human activities in Poland such as land use, crop production, emission of particles and pollutant gasses, forest area and fires, population density, production of electricity, vehicle types, and air pollution reduction systems are only at an annual level, not at a daily level.

Recording data of these features at a daily level could be a valuable tool in identifying air quality levels in Poland and assisting policy makers in taking appropriate measures to mitigate any negative impacts. By consistently tracking and monitoring the air quality indicators, trends and patterns can be identified and analyzed, allowing for targeted interventions and solutions to be developed. For example, if data consistently shows high levels of $PM_{2.5}$ in a certain region, policy makers can implement measures to reduce emissions from industrial facilities or transportation in that area. This information can be used to inform policies and regulations aimed at improving air quality and protecting public health.

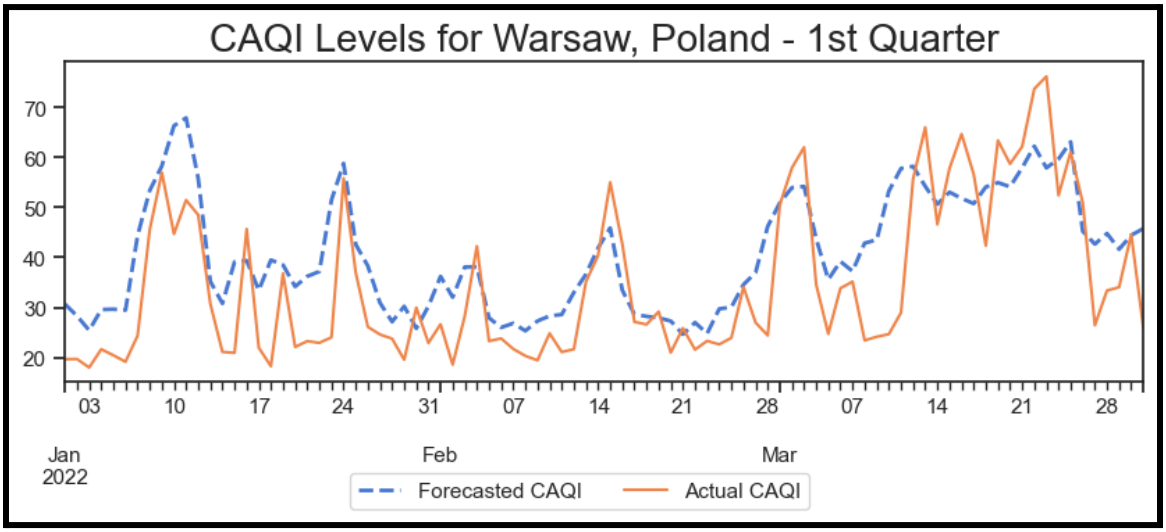
6.4. Forecasting Future CAQI Levels

In order to assess the trained models forecasting accuracy, the trained XGBoost regressor model with the best RMSE score was trained on both train and test data to forecast future CAQI levels of Warsaw, the capital of Poland, from Jan 2022 to Jan 2023. The forecasted CAQI levels and Actual CAQI levels are shown below.

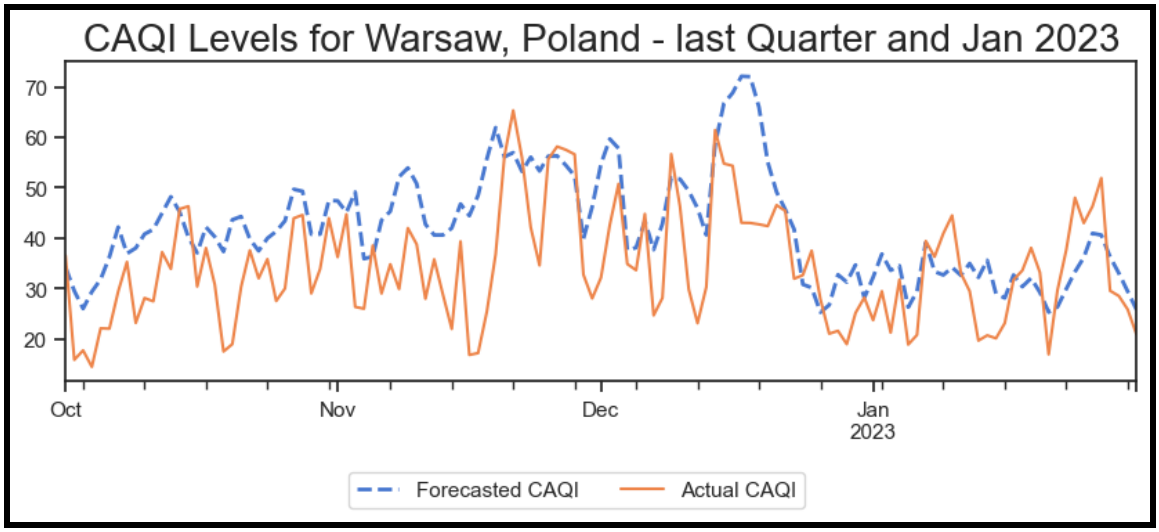


Forecasted vs Actual Warsaw CAQI levels from Jan 2022 to Jan 2023

The model has generalized very well to the data. The model has successful captured the trend and seasonality of the CAQI levels. The model forecasts for the first 2 months are very close to the actual CAQI levels. As the forecasting horizon increases, we can see the model tends to over predict CAQI levels, but this is not unusual. It is difficult to make long-term forecasting compared to short-term forecasting.



Forecasted vs Actual Warsaw CAQI levels for First Quarter of 2022



Forecasted vs Actual Warsaw CAQI levels for Last Quarter of 2022 and Jan 2023

CONCLUSION

In conclusion, our study offers valuable insights into the main causes of high air pollution levels in Poland and highlights the potential of machine learning for addressing this critical environmental issue. By analyzing available air quality, weather, and other relevant data, we were able to identify key trends and patterns that shed light on the sources and impacts of various pollutants. Our findings indicate that human activities are the main drivers of pollution in Poland, with traffic and industrial emissions being among the most significant contributors.

Our machine learning model successfully generalized to past data and demonstrated good forecasting performance for up to 2 months. The model considers weather variables to be one of the important important features to predict CAQI levels. However, we also identified challenges related to data availability, particularly the lack of information on daily activities that are directly linked to pollution. Incorporating more rigorous data collection into pollution monitoring and mitigation strategies is essential for developing robust real-time pollution prediction models and identifying targeted interventions to reduce pollution levels.

Moving forward, we recommend that policymakers and environmental agencies prioritize the development of more comprehensive and frequent data collection systems to support data-driven decision-making. By investing in cutting-edge technologies and data-driven solutions, we can make significant strides towards mitigating air pollution in Poland and creating a healthier and more sustainable future for all.

EXECUTIVE SUMMARY

The aim of this project was to investigate the primary factors responsible for air pollution in Poland and develop an effective tool to predict air quality. To achieve this, three datasets were utilized: daily air quality data, daily weather data, and Static annual data from the Polish Central Statistical Office, all from 2017 to 2021. The annual data provided information on various features related to human acitivities, such as area by land use, crop production, emission of particulates and pollutant gases, forest area and fires, population density, production of electricity, vehicle types, and air pollution reduction systems. To achieve the project's objectives, a methodology was adopted, which involved translating and cleaning the data, combining data from different years and pollutants, followed by data analysis and machine learning. The cleaned datasets were analyzed to identify the key factors contributing to air pollution in Poland.

The results revealed that the concentration of pollutants varies during different times of the year. PM_{10} and $PM_{2.5}$ are primarily emitted from sources such as solid fuel combustion, road traffic, dust from construction sites, and agricultural activities. Industrial plants such as coking plants, steel mills, and coal-fired power plants are also significant sources of air pollution in many places in Poland. NO_2 is primarily emitted from combustion processes in energy production, manufacturing industry, and road transport, with diesel engines being a significant contributor to nitrogen oxide emissions. O_3 is a secondary pollutant formed by the influence of solar radiation on a combination of airborne pollutants, including NO_x and VOCs.

To standardize the measurement of air quality and communicate it to the public, the Common Air Quality Index (CAQI) was used. It considers various pollutants, including PM_{10}, $PM_{2.5}$, O_3, NO_2, and SO_2, and assigns a value to indicate the level of air pollution. The CAQI levels are higher during weekdays and lower during weekends, potentially indicating that weekly pollutant concentrations depend on human activities.

The dataset was preprocessed, and two datasets were created for training classification and regression models. Various techniques, including feature engineering, feature selection, cross-validation, and hyperparameter optimization, were utilized. A range of models were trained and tested to predict CAQI classes and indexes, and their performance was evaluated using metrics such as F1 score and RMSE. Finally, an XGBoost regressor model was used to forecast future CAQI levels of Warsaw, and the model’s short-term forecasts was really close to the actual CAQI levels for the forecasting period.

Daily monitoring and recording of the various features that affect air quality in Poland can play a significant role in identifying the overall air quality levels. By analyzing and interpreting such data, policymakers and stakeholders can gain valuable insights into the trends and patterns of air pollution in the region. This information can help them identify areas of concern and take appropriate actions to mitigate the impact of air pollution on public health and the environment. Overall, a data-driven approach to air quality monitoring and policymaking can contribute significantly to creating a healthier and sustainable environment in Poland.