Mitigating Air Pollution in Poland Through Machine Learning

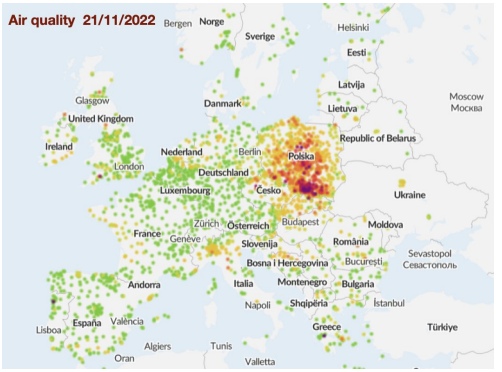
Final Presentation

Agenda

1. Problem statement
2. Learning outcomes
3. Task overview
4. Data cleaning and preprocessing
5. EDA
6. Modelling
7. Results

The Problem

- Air pollution is a problem in
  → among the countries with the worst air quality in Europe.
- Bad air quality affects people's lives and constitutes a considerable health risk
- Mitigating air pollution could improve quality of life and lead to an overall healthier society
- A statistical model based on machine learning could yield valuable insights into main factors and causes of air pollution specific to Poland
- 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.



Learning Outcomes

- Collaboration and Communication skills
- Problem understanding
- Explore the tools required for the project
- Brainstorming sessions and weekly meetings
- Data cleaning, data preprocessing, data analysis, data visualization
- Building different machine-learning models for classification and regression

Task Overview

- **Task 1 - Data Cleaning:** Task leads: Kojo Kesse, Vidushi Khanna
- **Task 2 - Data Preprocessing:** Task leads: Catalina, Chidansh M
- **Task 3 - EDA:** Task leads: Joseph Antony, Malini
- **Task 4 - Modelling:** Task leads: Shubhankar Sharma, Saga
- **Task 5 - Dashboard:** Task lead: Vinod

Data Cleaning and Preprocessing

Data Cleaning and Preprocessing - Provided Raw Datasets

Daily air quality data (NO_2, O_3, $PM_{2.5}$, PM_{10} levels) from 2015 to 2021, provided by the Chief Inspectorate for Environmental Protection in Poland.

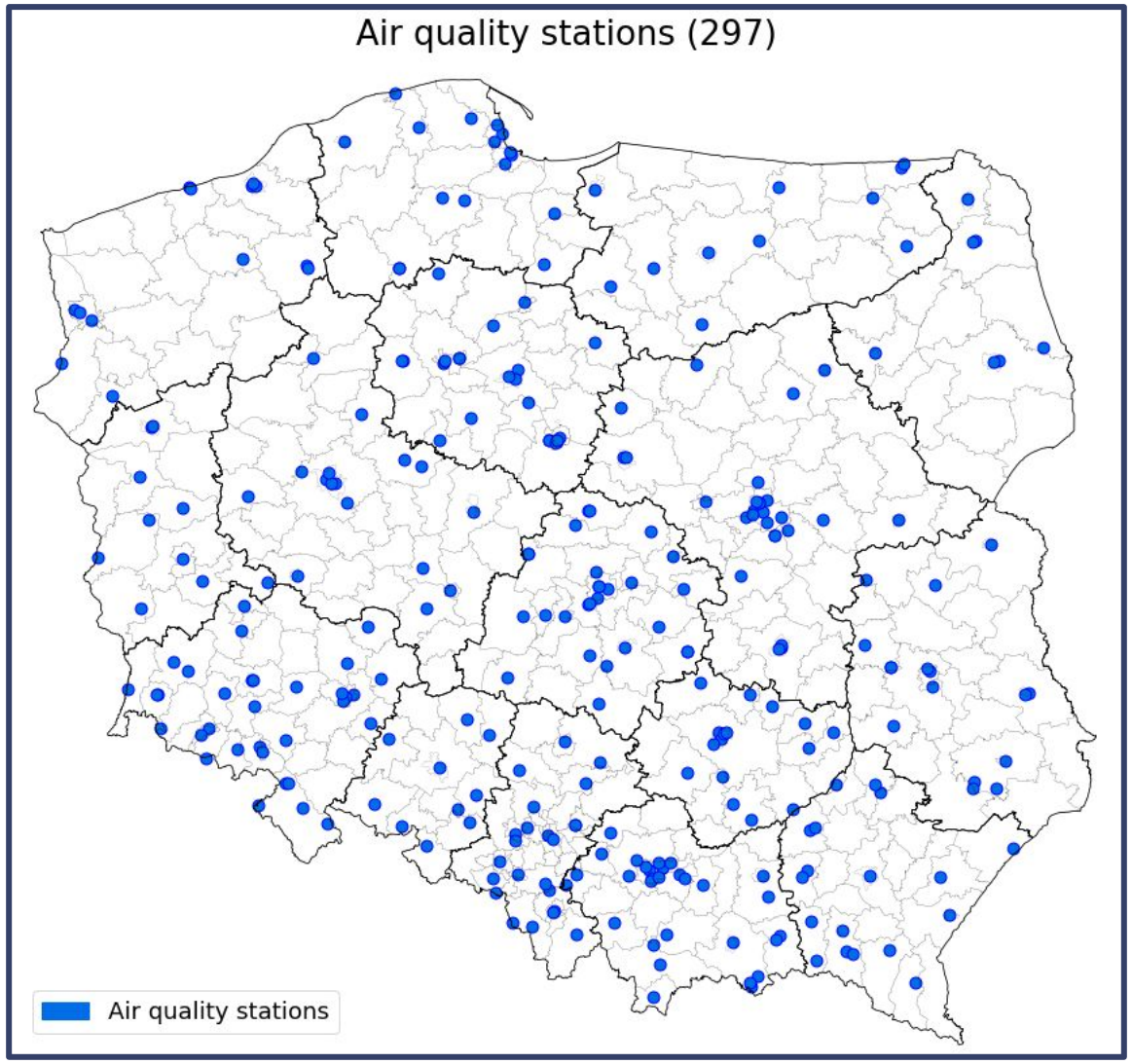
- NO_2
- O_3
- $PM_{2.5}$
- PM_{10}

	A	B	C	D	E
1	Nr	1	2	3	4
2	Kod stacji	DsBoleslaMOB	DsCzerStraza	DsJelGorOgin	DsKlodzSzkol
3	Wskaźnik	O3	O3	O3	O3
4	Czas uśredniania	1g	1g	1g	1g
5	Jednostka	ug/m3	ug/m3	ug/m3	ug/m3
6	Kod stanowiska	DsBoleslaMOB-O3-1g	DsCzerStraza-O3-1g	DsJelGorOgin-O3-1g	DsKlodzSzkol-O3-1g
7	01/01/2017 01:00		92,4583	5,445	7,60278
8	01/01/2017 02:00		94,5167	6,47833	9,27111
9	01/01/2017 03:00		91,6167	5,66278	8,46222
10	01/01/2017 04:00		90,7433	5,09944	9,31833

- one file per pollutant per year
- both hourly and daily measurements
- different coverage of stations for each pollutant
- dataset in Polish
- station data in separate file

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Nr	Kod stacji	Kod międzynaro	Nazwa stacji	Stary Kod stacji	Data uruchomienia	Data zamknięcia	Typ stacji	Typ obszaru	Rodzaj stacji	Województwo	Miejscowość	Adres	WGS84 ϕ N	WGS84 λ E
2	1	DsBialka		Białka	DsBialka	1/3/1990	12/31/2005	przemysłowa	podmiejski	kontenerowa stacjonarna	DOLNOŚLĄSKIE	Białka		51.197783	16.117390
3	2	DsBielGrot		Bielawa - ul. Grota Roweckiego	DsBielGrot	1/2/1994	12/31/2003	tło	miejski	w budynku	DOLNOŚLĄSKIE	Bielawa	ul. Grota Roweckiego 6	50.682510	16.617348
4	3	DsBogatFrancMOB	PL0602A	Bogatynia Mobil	DsBogatMob	1/1/2015	12/31/2015	tło	miejski	mobilna	DOLNOŚLĄSKIE	Bogatynia	ul. Francuska/Kręta	50.940998	14.916790
5	4	DsBogChop	PL0315A	Bogatynia - Chopina	DsBogChop	1/1/1996	12/31/2013	przemysłowa	miejski	kontenerowa stacjonarna	DOLNOŚLĄSKIE	Bogatynia	ul. Chopina 35	50.905856	14.967175
6	5	DsBogZatonieMob	PL0576A	Bogatynia - Mobil	DsBogZatonieMob	1/1/2012	12/31/2012	przemysłowa	miejski	mobilna	DOLNOŚLĄSKIE	Bogatynia	ul. Konrada, Zatonie	50.943245	14.913327
7	6	DsBoleslaMOB	PL0658A	Bolesławiec	DsBoleslaMOB	1/1/2017	1/2/2018	tło	miejski	mobilna	DOLNOŚLĄSKIE	Bolesławiec	Juliusza Słowackiego 2	51.263245	15.570354
8	7	DsBrzegGlog		Brzeg Głogowski	DsBrzegGlog	1/1/1980	12/31/2003	przemysłowa	pozamiejski	w budynku	DOLNOŚLĄSKIE	Brzeg Głogowski		51.691438	15.917840
9	8	DsChojnowKil	PL0185A	Chojnów,ul.Kilińskiego	DsChojnowKil	1/6/2004	12/31/2005	tło	miejski	kontenerowa stacjonarna	DOLNOŚLĄSKIE	Chojnów	ul. Kilińskiego	51.268889	15.940278
10	9	DsCzar07	PL0186A	Czarna Góra	DsCzar07	12/1/1996	9/30/2008	tło	pozamiejski	kontenerowa stacjonarna	DOLNOŚLĄSKIE	Czarna Góra		50.255072	16.801641

Data Cleaning and Preprocessing - Air Quality Station Coverage



Data Cleaning and Preprocessing - Provided Raw Datasets

Daily weather data from 1979 to 2021, provided by European Climate Assessment and Dataset (ECAD) project.

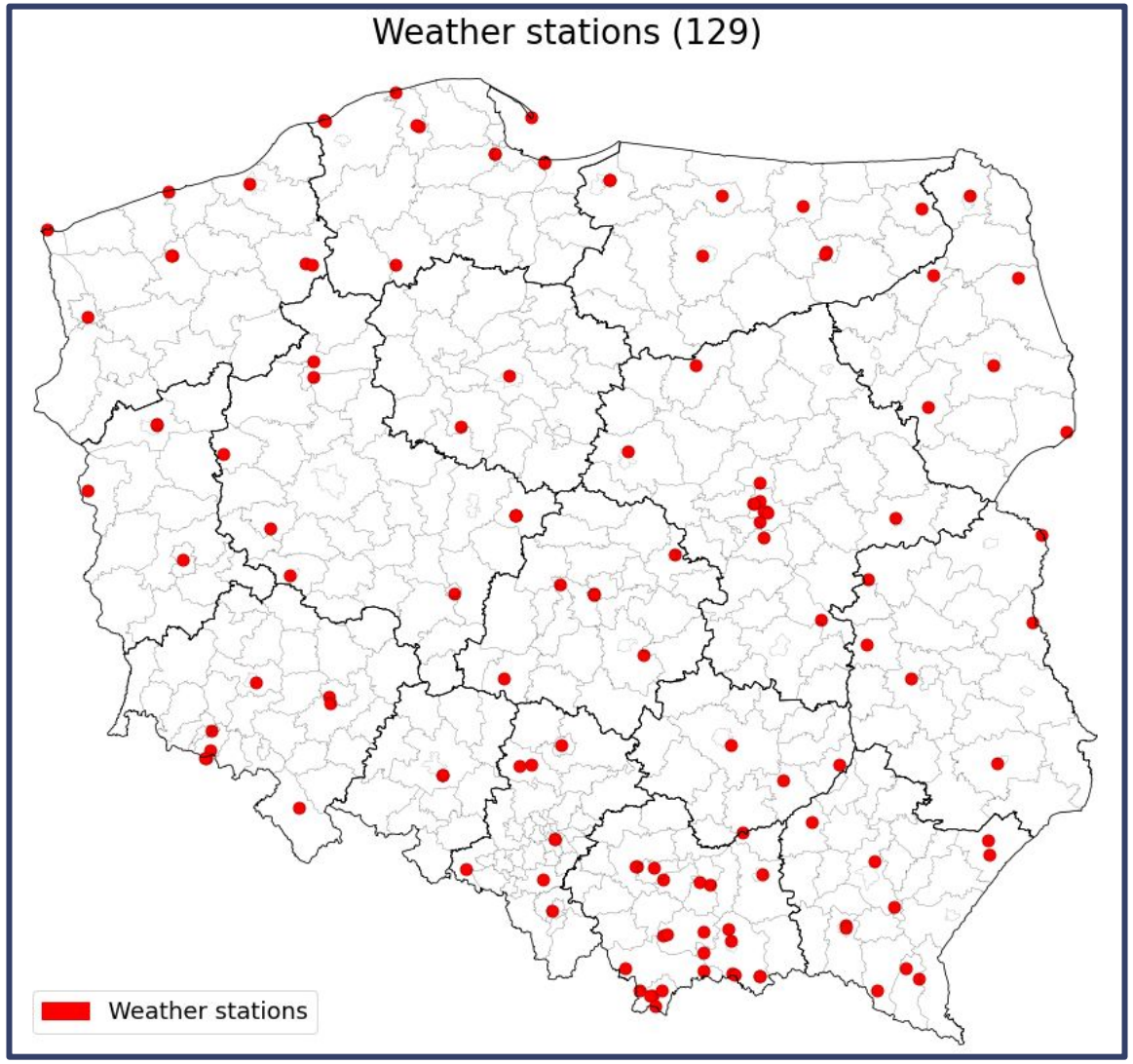
```
01-06 SOUID: Source identifier
08-15 DATE : Date YYYYMMDD
17-21 RR   : precipitation amount in 0.1 mm
23-27 Q_RR : Quality code for RR (0='valid'; 1='suspect'; 9='missing')

This is the blended series of station POLAND, HEL (STAID: 205).
Blended and updated with sources: 100643
See file sources.txt and stations.txt for more info.

 SOUID,    DATE,   RR, Q_RR
100643,19790101,    0,    0
100643,19790102,   16,    0
100643,19790103,    0,    0
100643,19790104,   35,    0
```

```
FILE FORMAT (MISSING VALUE CODE IS -9999):

 01- 05 STAID  : Station identifier
 07- 46 STANAME: Station name
 48- 49 CN     : Country code (ISO3116 country codes)
 51- 59 LAT    : Latitude in degrees:minutes:seconds (+: North, -: South)
 61- 70 LON    : Longitude in degrees:minutes:seconds (+: East, -: West)
 72- 75 HGTH   : Station elevation in meters

STAID,STANAME                                 ,CN,      LAT,       LON,HGHT

  204,BIALYSTOK                               ,PL,+53:06:26,+023:09:44, 148
  205,HEL                                     ,PL,+54:36:13,+018:48:43,   1
  206,POZNAN                                  ,PL,+52:11:59,+018:39:37, 115
```

- cloud cover (CC),
- global radiation (QQ),
- humidity (HU),
- mean temperature (TG),
- precipitation (RR),
- sea level pressure (PP),
- snow depth (SD),
- sunshine (SS),
- wind speed (FG)

- one file per station per measured quantity
- daily measurements
- different coverage of stations for each quantity
- quality codes
- station data in separate files

Data Cleaning and Preprocessing - Weather Station Coverage



Data Cleaning and Preprocessing - Provided Raw Datasets

Static annual data from 2010 to 2021, provided by the Polish Central Statistical Office.

- animal stock,
- area by land use,
- crop production,
- emissions of particles and pollutant gasses,
- forest area and fires,
- population density,
- production of electricity,
- vehicle types,
- air pollution reduction systems

	A	B	C	D	E	F	G	H	I	J	K
1	Code	Name	population per 1 km2								
2			2010	2011	2012	2013	2014	2015	2016	2017	2018
3			[person]	[person]	[person]	[person]	[person]	[person]	[person]	[person]	[person]
4	0201000	Powiat bolesławiecki	69	69	69	69	69	69	69	69	69
5	0202000	Powiat dzierżoniowski	222	221	220	219	217	216	215	213	212
6	0203000	Powiat głogowski	204	204	204	204	204	203	203	203	202
7	0204000	Powiat górowski	50	50	50	49	49	49	48	48	48
8	0205000	Powiat jaworski	91	90	90	90	89	88	88	88	87
9	0206000	Powiat karkonoski	104	104	104	104	104	103	103	102	102
10	0207000	Powiat kamiennogórski	116	116	115	114	113	112	112	111	110
11	0208000	Powiat kłodzki	102	102	101	100	100	99	98	98	97

- some data was given on Powiat level, some data is only given on Voivodeship level
- some categories span different time ranges (e.g., from 2015 or from 2017)
- each Powiat has a unique code, making it possible to identify its Voivodeship

Data Cleaning and Preprocessing - Inferring Powiat and Voivodeship

- Data to be aggregated at the powiat level
- Problems:
  - no Powiat/Voivodeship data for air quality and weather datasets
  - sometimes only voivodeship data in Static Annual Dataset
- Solution:
  - GEOJSON data found online comprising vertex coordinates of individual Powiat and Voivodeships
  - possible to check the Powiat and Voivodeship for a given station (air quality or weather)

Data Cleaning and Preprocessing - Air Quality Dataset

Actions carried out in the following areas:

- translation from Polish to English (both language and decimal separator change)
- filtering the dataset to 2017-2021 and combining into one time-series for each pollutant
- computing daily averages (for the hourly measurements)
- matching stations with similar names (stations at the same location but operating in different periods had different names, although very similar)

	DATE	POLSTID	Voivodeship	City	county	postcode	LAT	LON	NO2_24H_AVG_POLLUTION	O3_24H_AVG_POLLUTION	...
0	2017-01-01	DsBoleslaMOB	DOLNOŚLĄSKIE	Bolesławiec	powiat bolesławiecki	59-700	51.263245	15.570354	NaN	38.285294	...
1	2017-01-01	DsCzerStraza	DOLNOŚLĄSKIE	Czerniawa	powiat lubański	59-850	50.912475	15.312190	NaN	90.390000	...
2	2017-01-01	DsGlogWiStwo	DOLNOŚLĄSKIE	Głogów	powiat głogowski	67-200	51.657022	16.097822	NaN	NaN	...
3	2017-01-01	DsJaworMOB	DOLNOŚLĄSKIE	Jawor	powiat jaworski	59-400	51.049212	16.202317	NaN	NaN	...
4	2017-01-01	DsJelGorOgin	DOLNOŚLĄSKIE	Jelenia Góra	powiat Jelenia Góra	58-506	50.913433	15.765608	NaN	11.980000	...

Data Cleaning and Preprocessing - Air Quality Dataset

Actions carried out in the following areas:

- complementing the dataset with Powiat and Voivodeship based on the extracted station coordinates
- often different stations measured different pollutants, which produced many missing values (to be handled later by imputation)

	powiat_voivod	DATE	LAT	LON	NO2_24H_AVG_POLLUTION	O3_24H_AVG_POLLUTION	PM10_24H_AVG_POLLUTION
0	powiat aleksandrowski, kujawsko-pomorskie	2017-01-01	52.888422	18.780908	NaN	32.22	24.968064
1	powiat aleksandrowski, kujawsko-pomorskie	2017-01-02	52.888422	18.780908	NaN	34.62	17.943745
2	powiat aleksandrowski, kujawsko-pomorskie	2017-01-03	52.888422	18.780908	NaN	42.00	14.477950
3	powiat aleksandrowski, kujawsko-pomorskie	2017-01-04	52.888422	18.780908	NaN	57.46	8.418471
4	powiat aleksandrowski, kujawsko-pomorskie	2017-01-05	52.888422	18.780908	NaN	64.62	13.189740

Data Cleaning and Preprocessing - Weather Dataset

Actions carried out in the following areas:

- converting weather station coordinates to the same format (minutes and seconds to decimal)
- combining and merging weather data from all stations into single dataset
- filtering the dataset to 2017-2021
- inferring Powiat and Voivodeship based on station coordinates.

	county	DATE	CC	FG	HU	PP	QQ	RR	SD	SS	TG	LAT	LON	Voivodeship
0	powiat Białystok	2017-01-01	8.0	0.0	89.0	10151.0	85.859777	0.0	0.0	39.002987	0.4	53.107222	23.162222	województwo podlaskie
1	powiat Białystok	2017-01-02	7.0	35.0	89.0	10069.0	70.528059	15.0	0.0	2.998211	0.6	53.107222	23.162222	województwo podlaskie
2	powiat Białystok	2017-01-03	7.0	0.0	91.0	10090.0	18.000000	0.0	1.0	7.000000	-1.4	53.107222	23.162222	województwo podlaskie
3	powiat Białystok	2017-01-04	8.0	32.0	93.0	9885.0	20.000000	50.0	7.0	0.000000	0.3	53.107222	23.162222	województwo podlaskie
4	powiat Białystok	2017-01-05	6.0	0.0	83.5	10101.0	20.000000	0.0	16.0	9.000000	-10.8	53.107222	23.162222	województwo podlaskie

Data Cleaning and Preprocessing - Static Annual Dataset

Actions carried out in the following areas:

- filtering the dataset to 2017-2021
- distribution of the Voivodeship level data onto Powiat level either by Powiat area or population density
- merging the different data categories for every Powiat
- value for the corresponding year can be used in the time-series

	Unnamed: 0_level_3	Unnamed: 1_level_3	Unnamed: 0_level_3	Unnamed: 1_level_3	2017_by_area	2017_by_pop	2018_by_area	2018_by_pop	2019_by_area	2019_by_pop
	Unnamed: 0_level_4	Unnamed: 1_level_4	Unnamed: 0_level_4	Unnamed: 1_level_4	[head]	[head]	[head]	[head]	[head]	[head]
1	201000.0	powiat bolesławiecki	200000.0	dolnośląskie	13092.0	6210.0	14222.0	6746.0	13746.0	6528.0
2	202000.0	powiat dzierżoniowski	200000.0	dolnośląskie	4809.0	7042.0	5224.0	7613.0	5050.0	7333.0

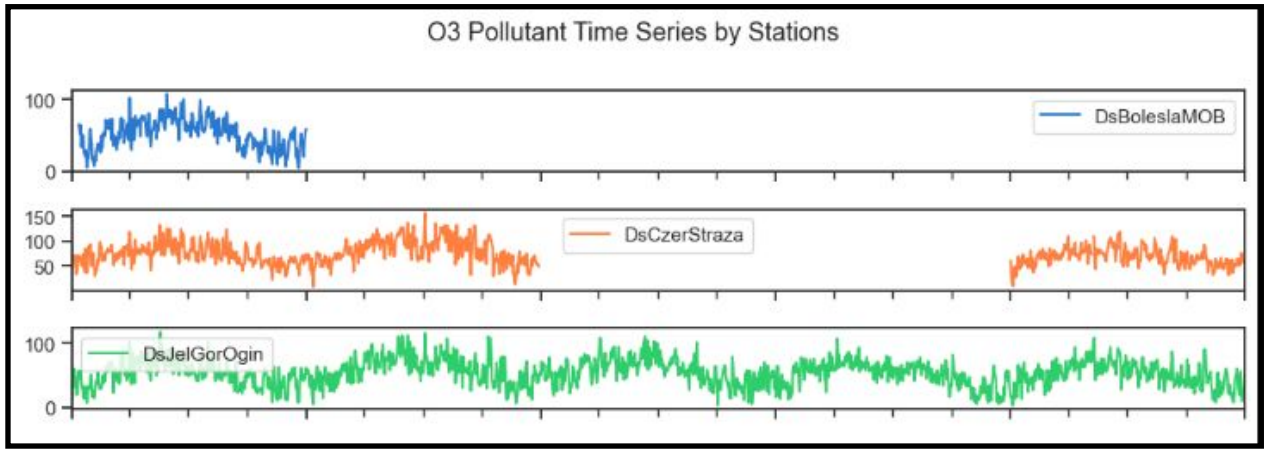
Data Cleaning and Preprocessing - Static Annual Dataset

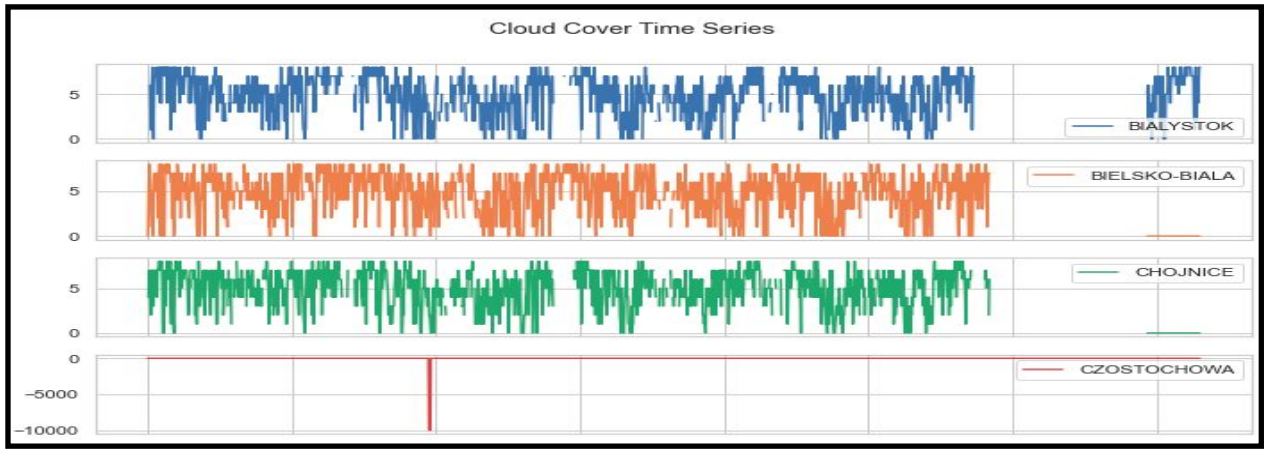
Data (Voivodeship only level)	Distributed by
Animal stock	area
Crop production	area
Forest fires	area
Production of electricity	population
Air pollution reduction systems in plants	population
Plants of significant nuisance to air quality	population

Data Cleaning and Preprocessing - Missing Data Imputation and Handling Outliers

- Large gaps in data and outliers present in both AQ and Weather dataset.

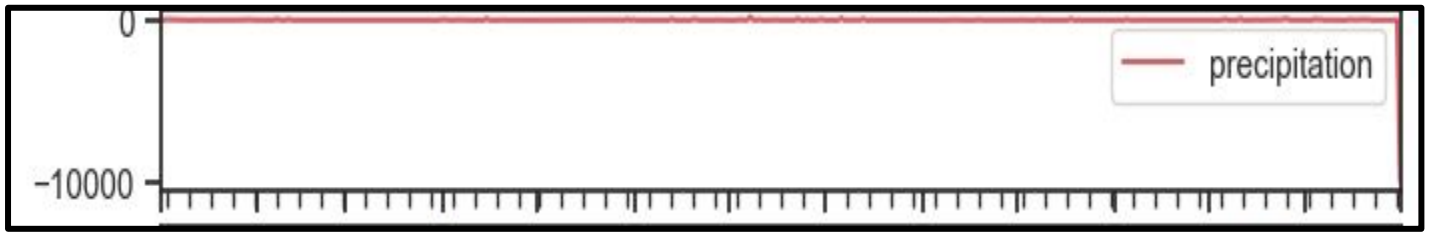
- Need to handle all missing values and outliers at each station level.

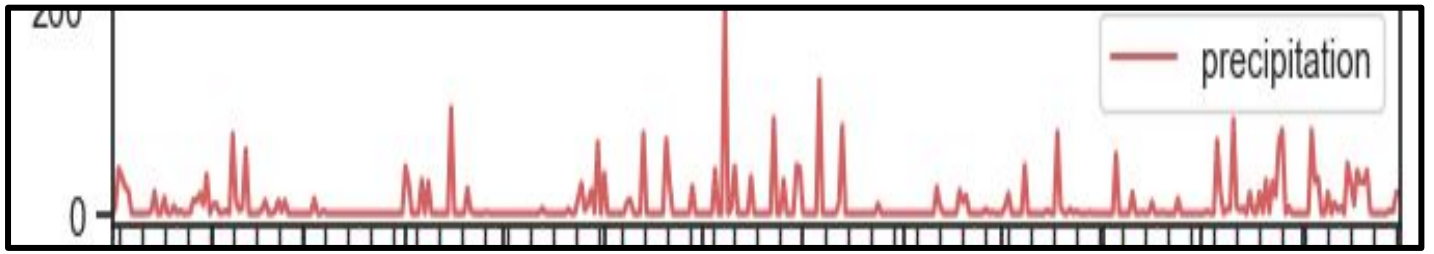




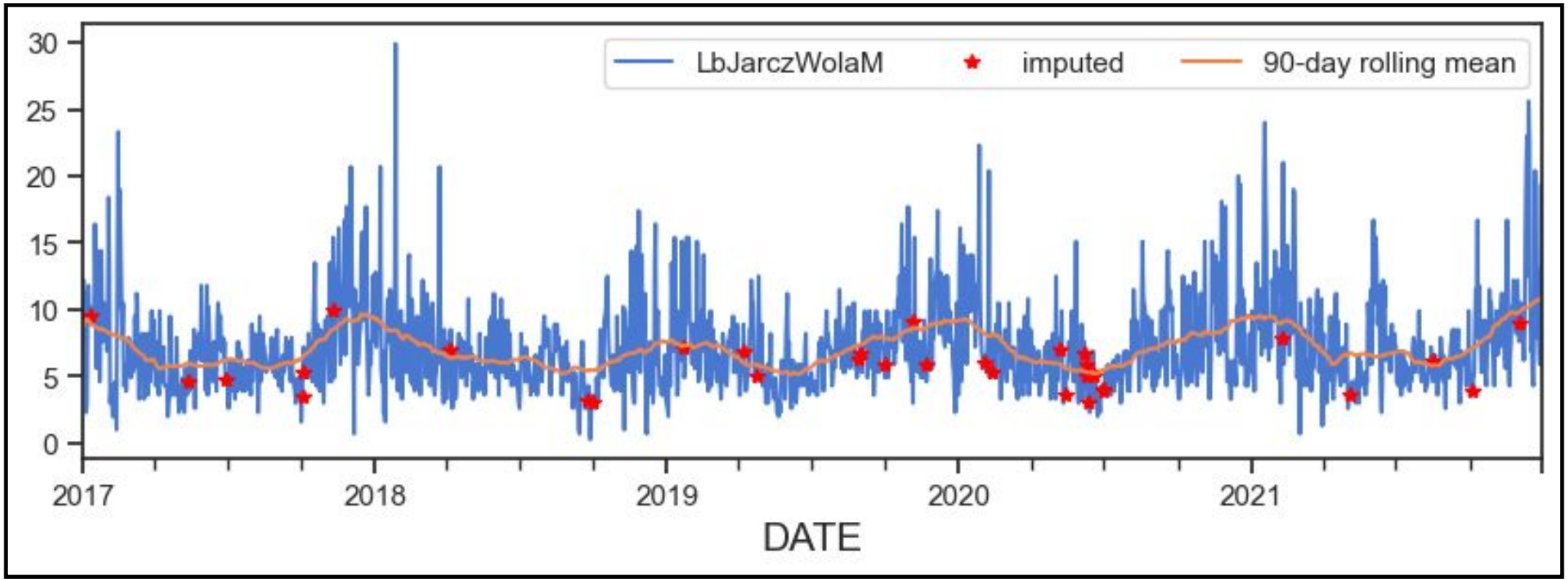
Missing Data Imputation - Handling Outliers in Weather Data

For weather datasets, rows with outliers replaced with null values.





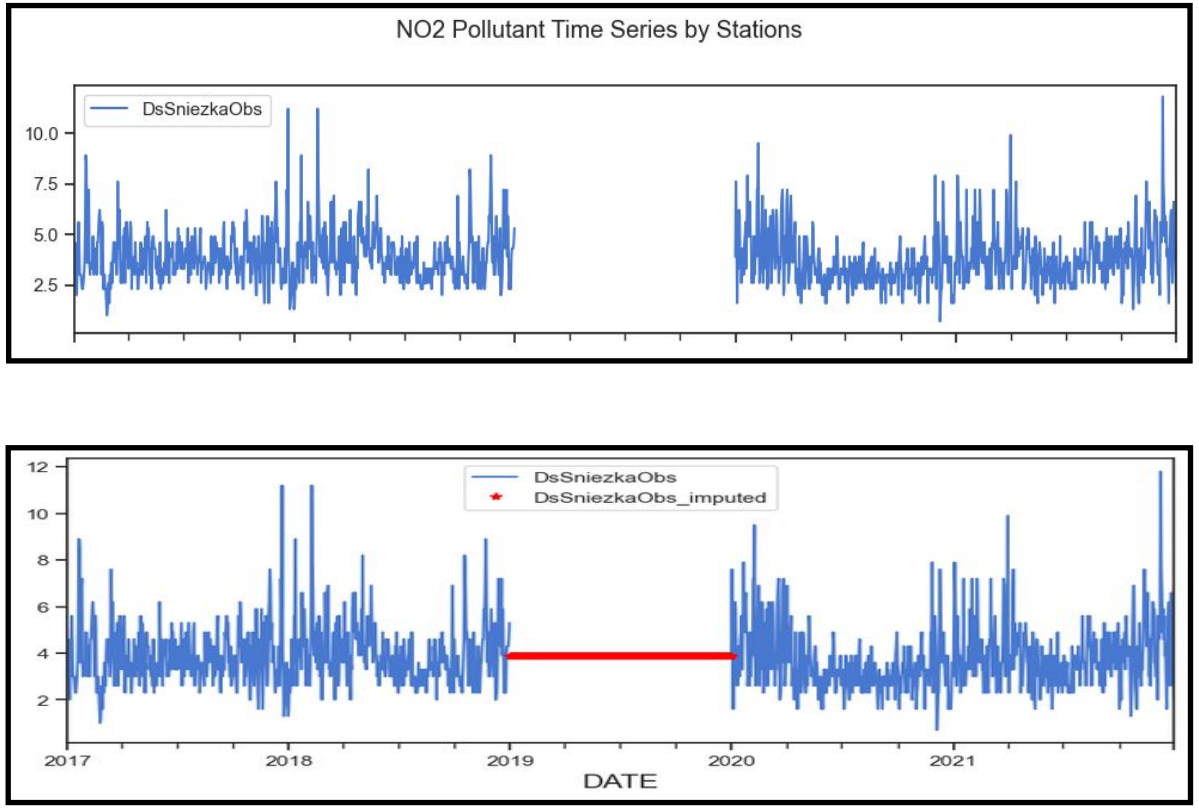
Missing Data Imputation - Linear Interpolation



For both datasets, station measurements with few missing data were imputed using linear interpolation.

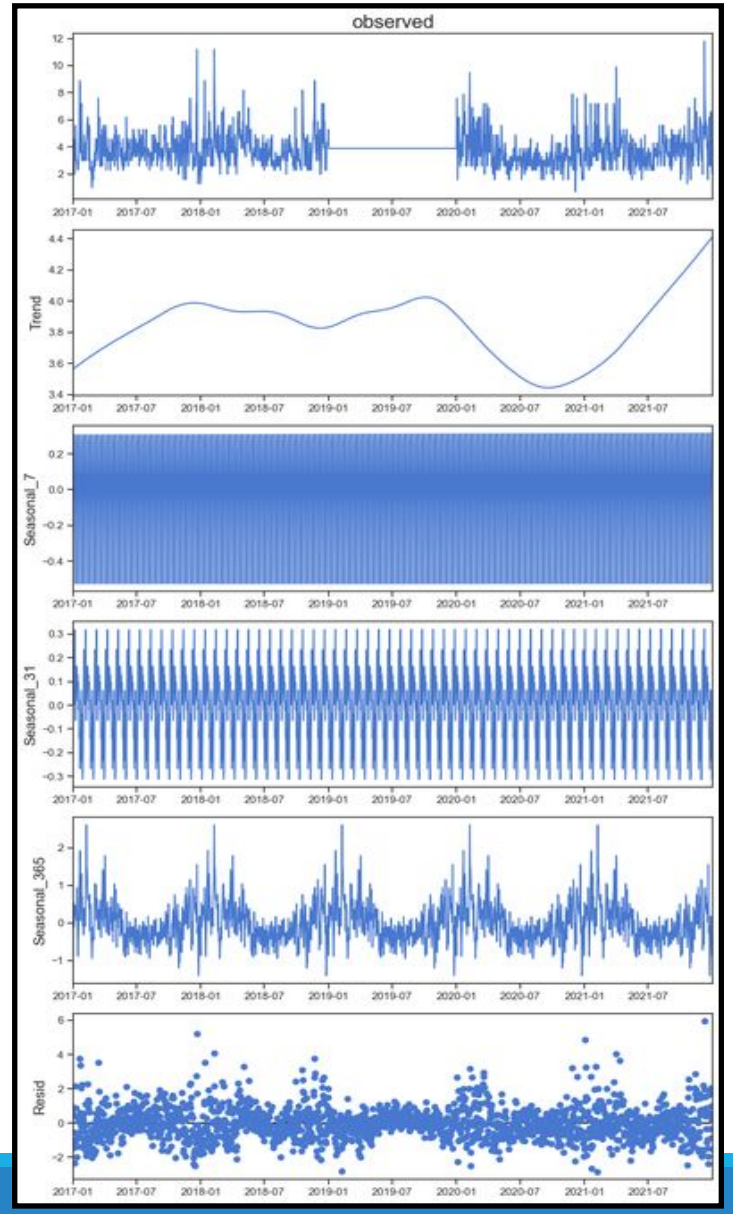
Missing Data Imputation
MSTL (Multiple Seasonal-Trend decomposition using LOESS) - Part 1

- Stations with large gaps was imputed using MSTL.
- Decomposing the data into trend and seasonal components, and reconstruct the time series by adding the estimated components together.
- First the missing data is imputed using simple linear interpolation.

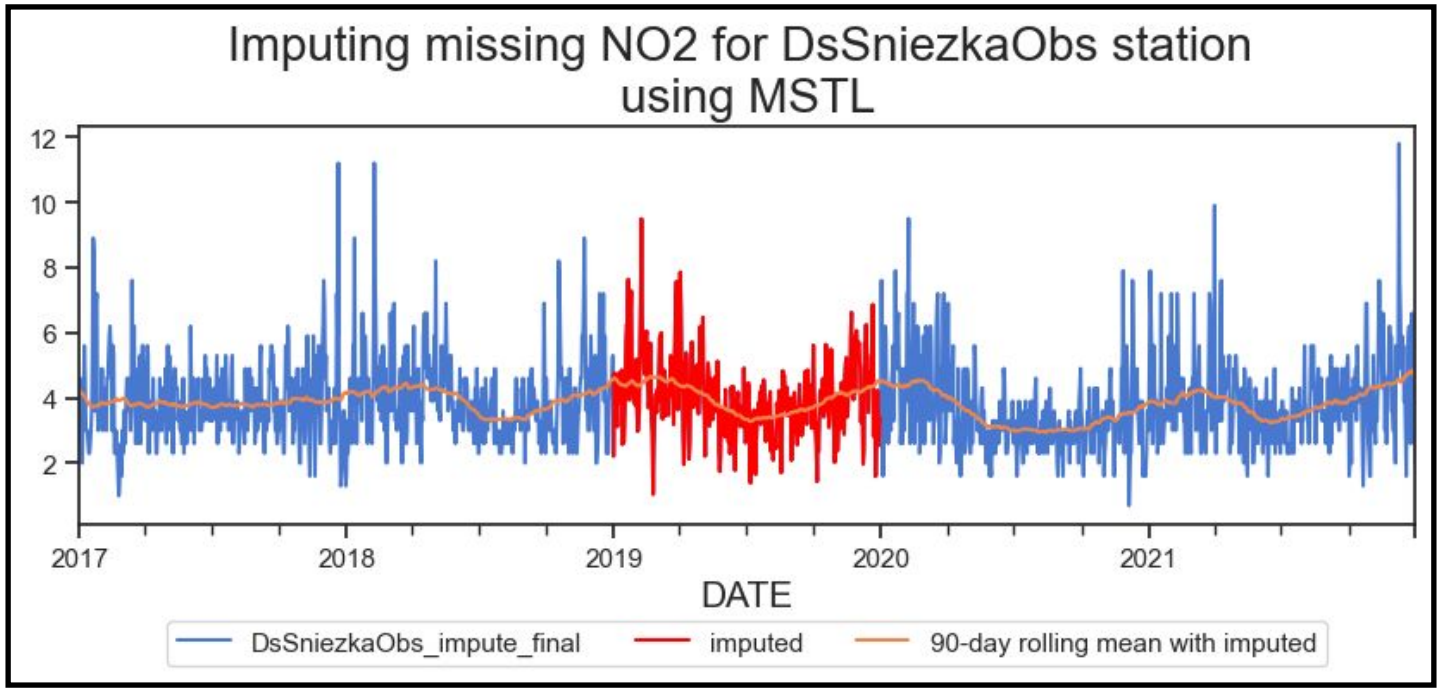


Missing Data Imputation
MSTL (Multiple Seasonal-Trend decomposition using LOESS) - Part 2

The imputed time series is then decomposed to Trend, Seasonal and Residual components.



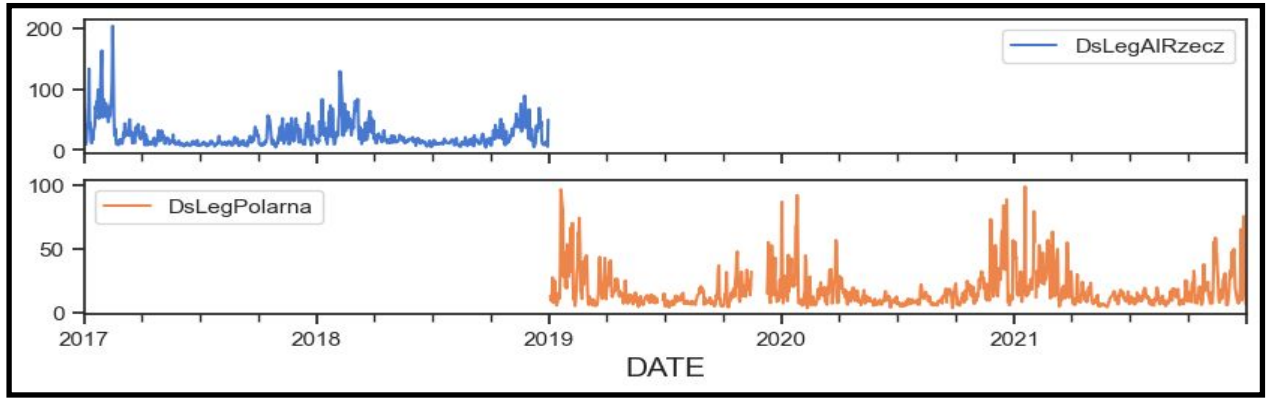
Missing Data Imputation
MSTL (Multiple Seasonal-Trend decomposition using LOESS) - Part 3

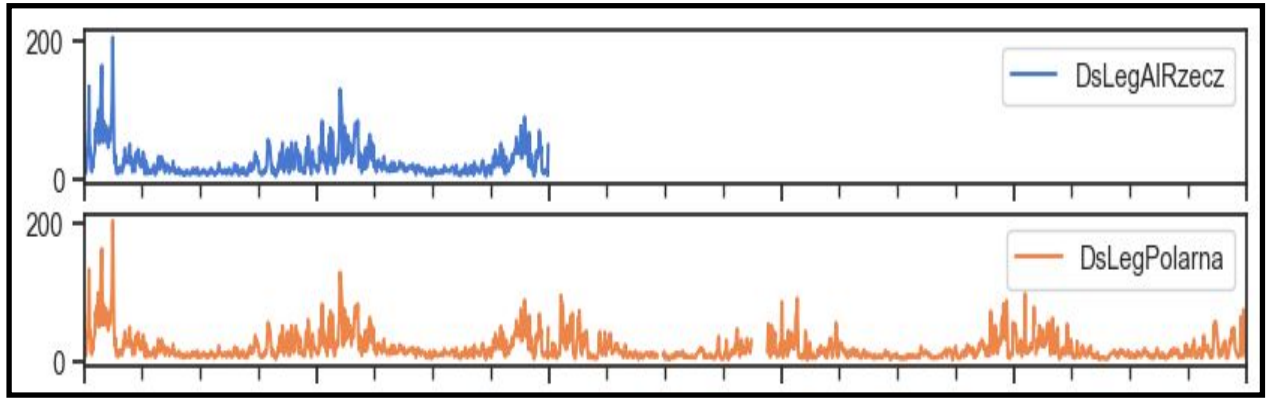


Time stamps containing missing values were then imputed by replacing them with the sum of corresponding seasonal and trend components.

Missing Data Imputation - Combining Measurements from Nearby Stations

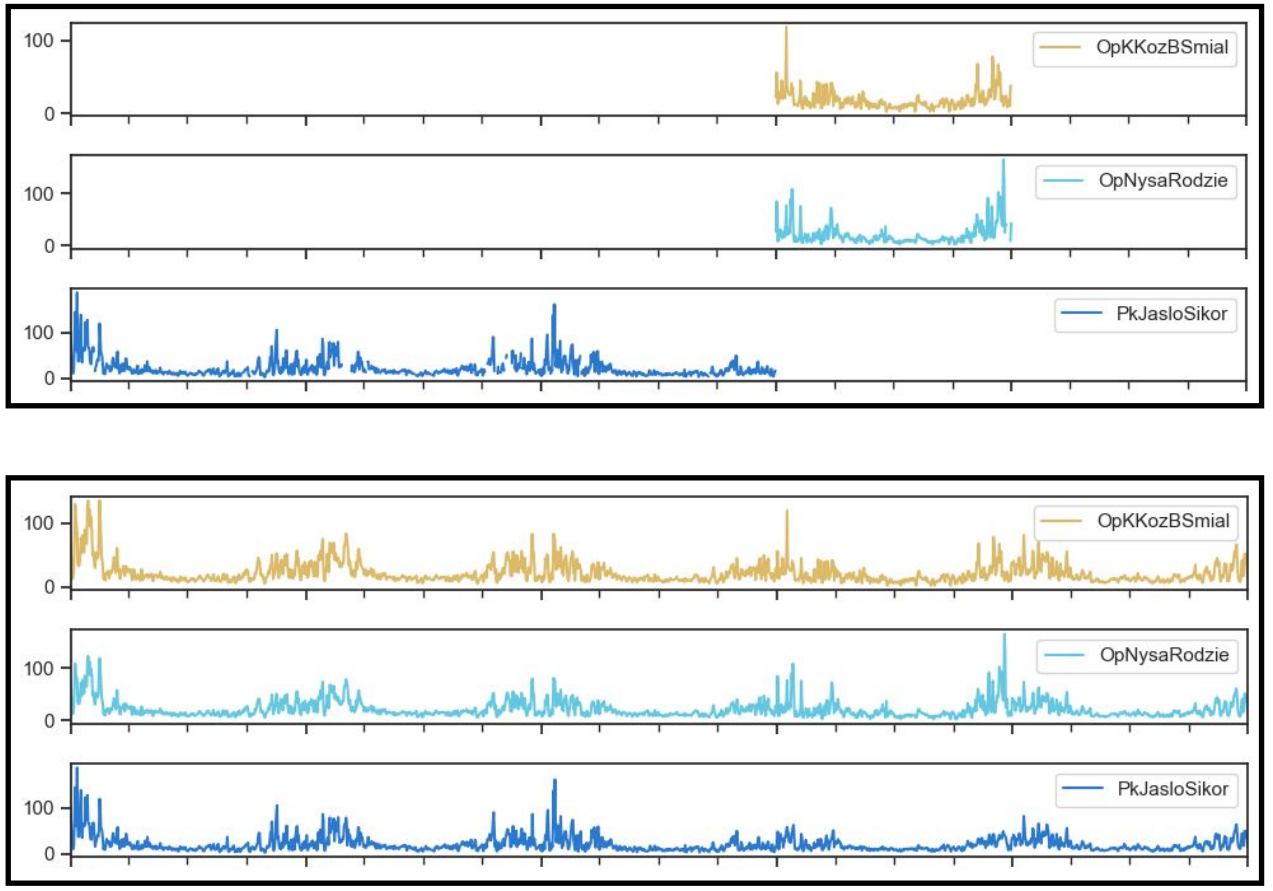
- Some stations observed to stop measuring for the rest of the year. Subsequently, another station starts measuring exactly at the time where the former station stopped measuring data.
- Nearby stations grouped using station name abbreviations.
- Imputed missing values of the station containing the latest data with the station that stopped measuring data early.





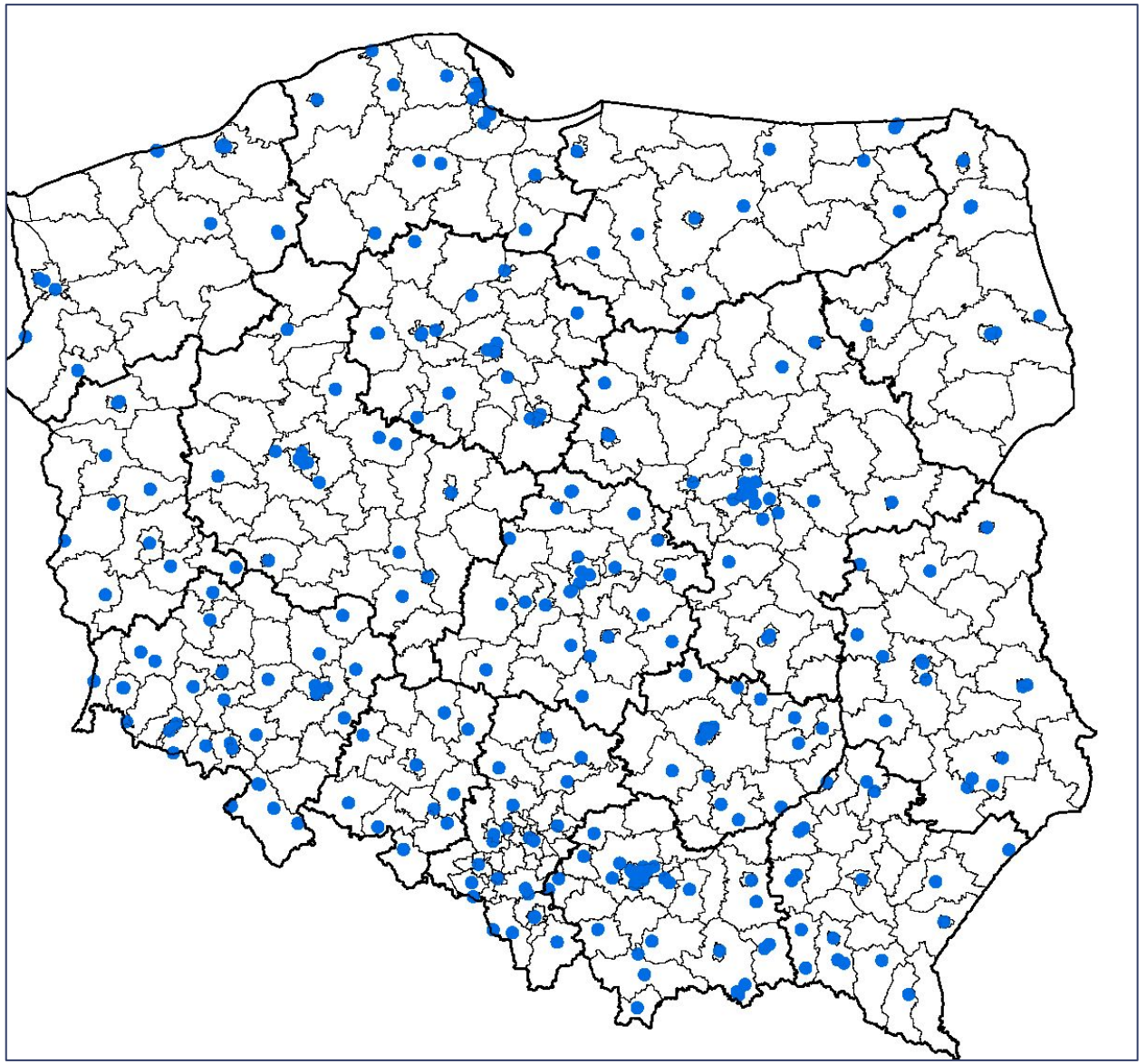
Missing Data Imputation - Imputation using Rolling Mean

Remaining few stations filled 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 data



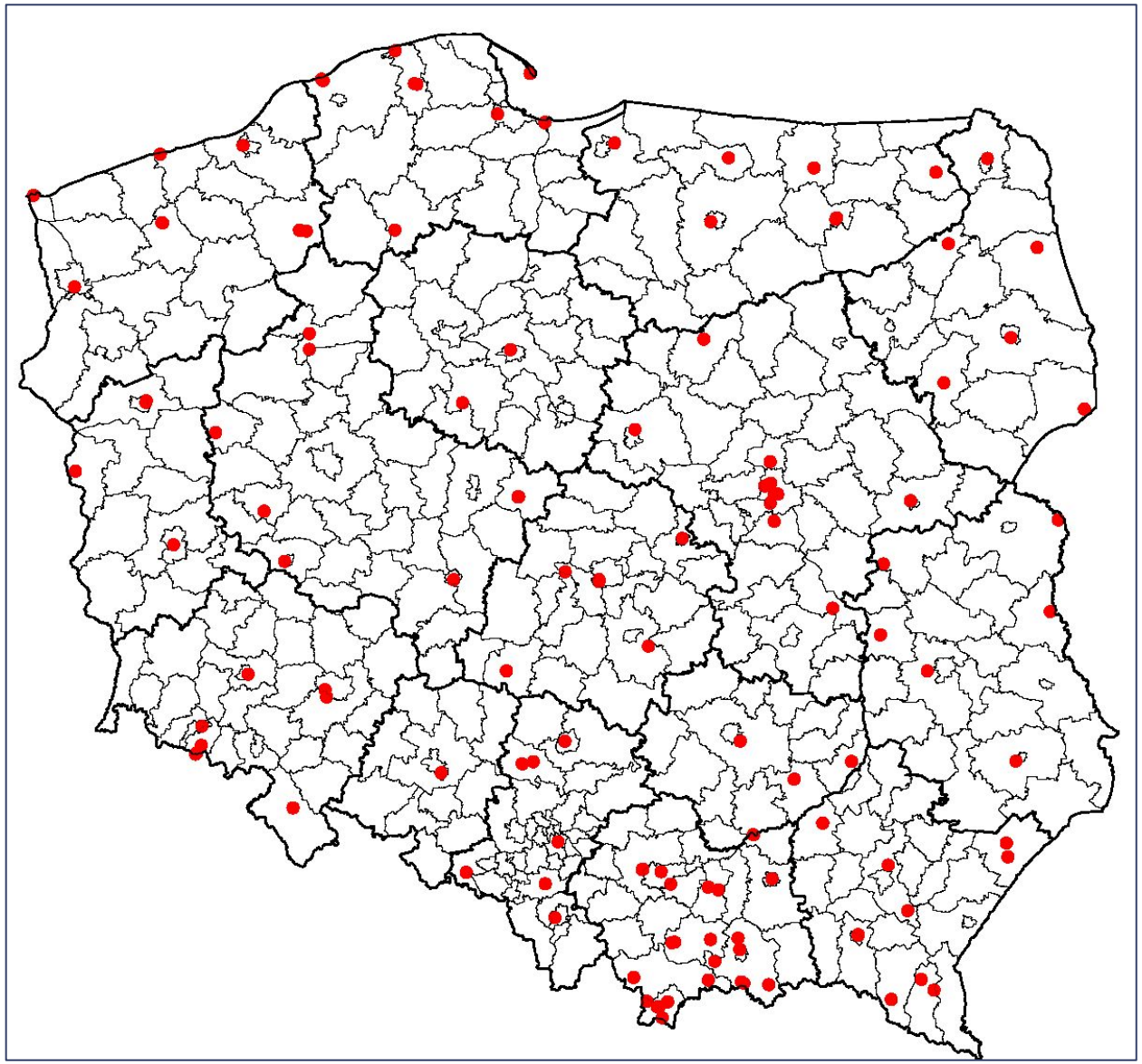
Merging Datasets - Air Quality Dataset

- We aggregated Air Quality dataset at the **Powiat level** and took the **mean of pollutant measurements**.
- **Reduced the hierarchy level of the time series datasets from 297 stations to 198 Powiats.**



Merging Datasets - Weather Dataset

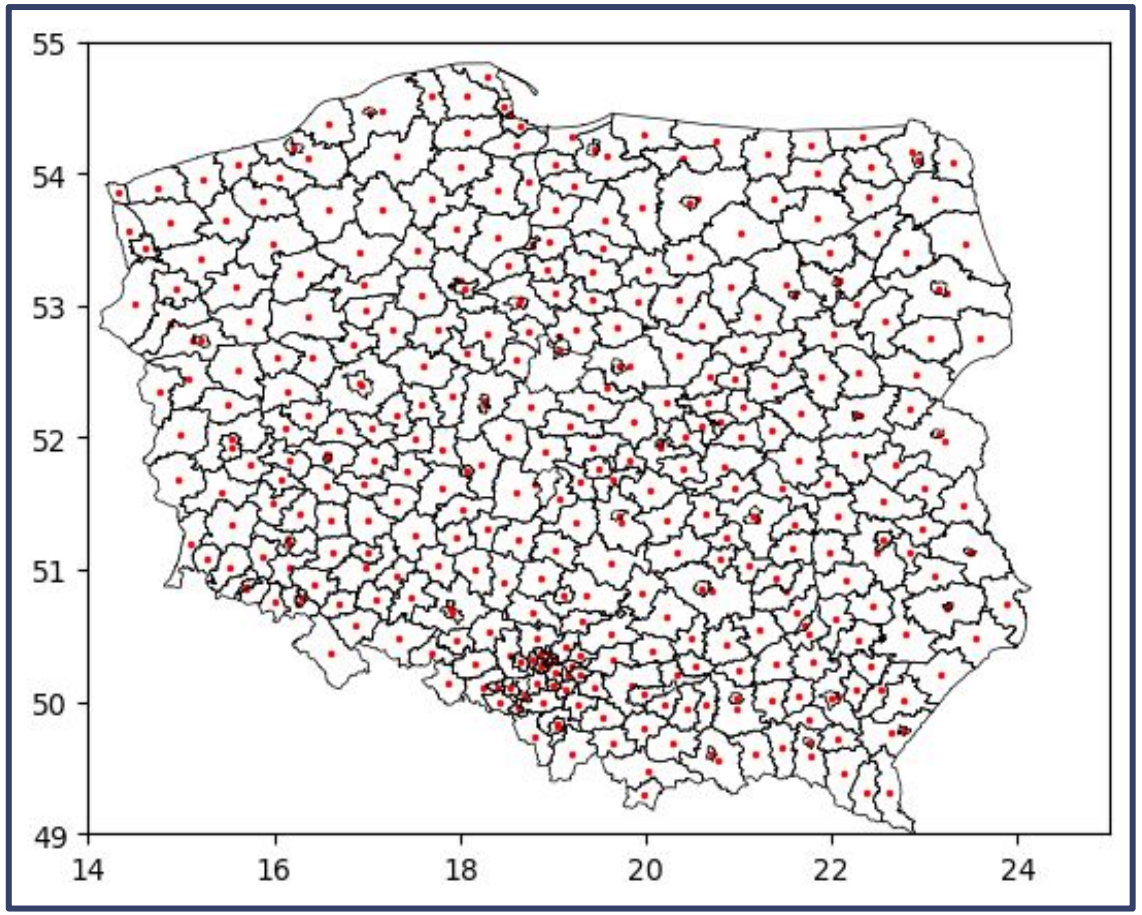
- Similarly, we aggregated weather dataset at the **Powiat level** and took the **mean weather measurements**.
- **Reduced the hierarchy level of the time series datasets from 129 stations to 86 Powiats.**



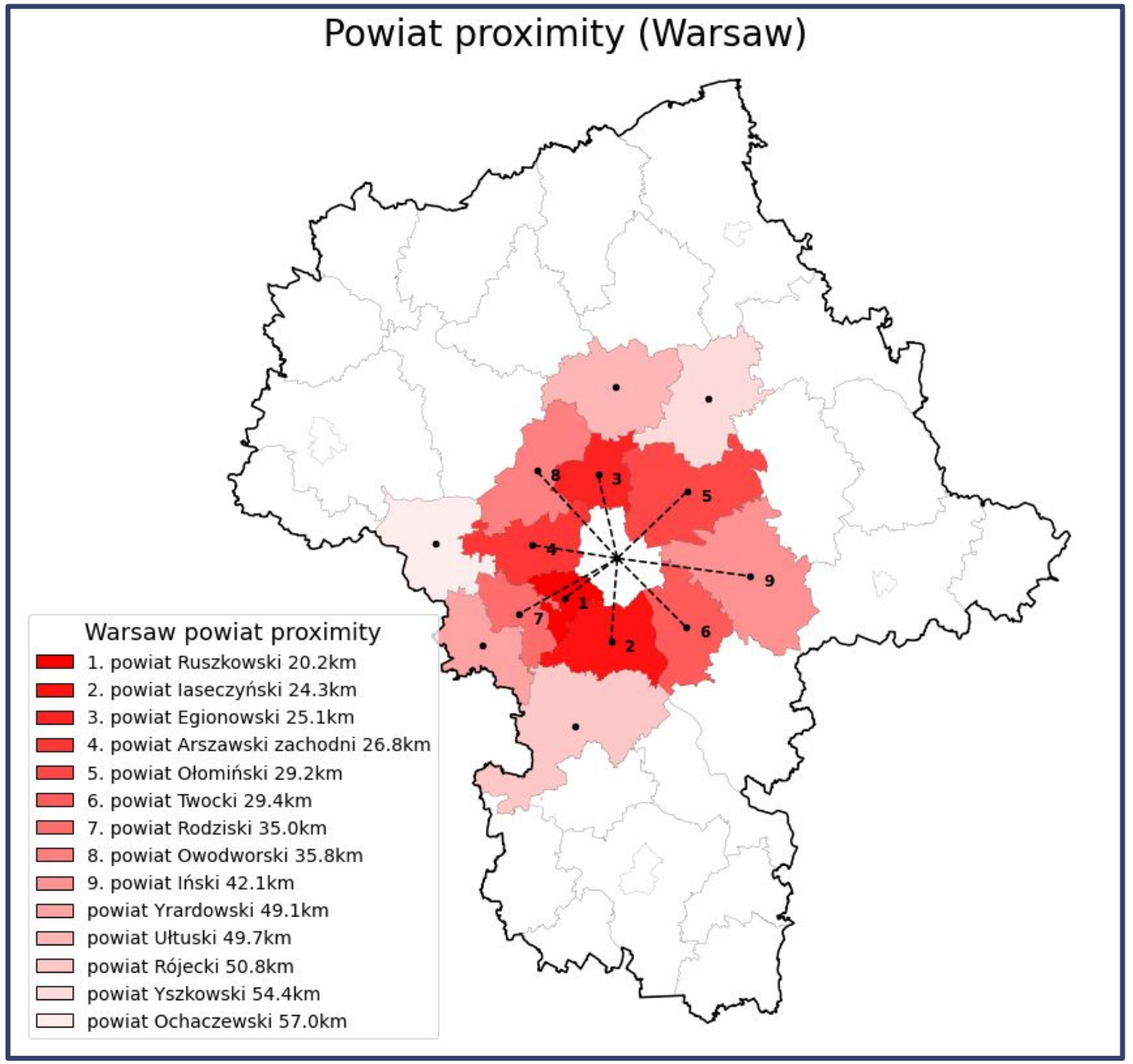
Data Cleaning and Preprocessing - Powiat proximity matrix

- Some Powiats (from AQ dataset) are not present in the Weather dataset, resulting in missing values.
- Idea:
  - use weather data from neighbour Powiats to impute the missing values
- Problem:
  - no information about the neighbours
- Solution:
  - use the GEOJSON data to compute centroids of all powiats
  - calculate the distances between every centroid
  - sort in ascending order to create a proximity matrix (column = base Powiat, rows = nearest Powiats in sorted in ascending order)

Data Cleaning and Preprocessing - Powiat proximity matrix



	powiat ropczycko-sędziszowski, podkarpackie	powiat łosicki, mazowieckie	powiat piaseczyński, mazowieckie	powiat radomski, mazowieckie	powiat sierpecki, mazowieckie
0	['powiat dębicki, podkarpackie', 21.2190122881...	['powiat siemiatycki, podlaskie', 29.002313749...	['powiat pruszkowski, mazowieckie', 19.2449686...	['powiat Radom, mazowieckie', 4.133242414485163]	['powiat żuromiński, mazowieckie', 26.35726136...
1	['powiat strzyżowski, podkarpackie', 21.379793...	['powiat Biała Podlaska, lubelskie', 29.869142...	['powiat otwocki, mazowieckie', 23.92159964549...	['powiat szydłowiecki, mazowieckie', 26.473805...	['powiat rypiński, kujawsko-pomorskie', 30.644...
2	['powiat Rzeszów, podkarpackie', 25.2138446464...	['powiat siedlecki, mazowieckie', 36.149599601...	['powiat Warszawa, mazowieckie', 24.3264518842...	['powiat zwoleński, mazowieckie', 28.631095044...	['powiat lipnowski, kujawsko-pomorskie', 30.78...
3	['powiat rzeszowski, podkarpackie', 29.1282279...	['powiat bialski, lubelskie', 37.444938021970735]	['powiat grójecki, mazowieckie', 27.3339090243...	['powiat białobrzeski, mazowieckie', 31.833877...	['powiat Płock, mazowieckie', 32.437683147101914]
4	['powiat kolbuszowski, podkarpackie', 29.71115...	['powiat Siedlce, mazowieckie', 39.3763810158149]	['powiat grodziski, mazowieckie', 30.115810250...	['powiat kozienicki, mazowieckie', 33.28111004...	['powiat płocki, mazowieckie', 32.475861191601...

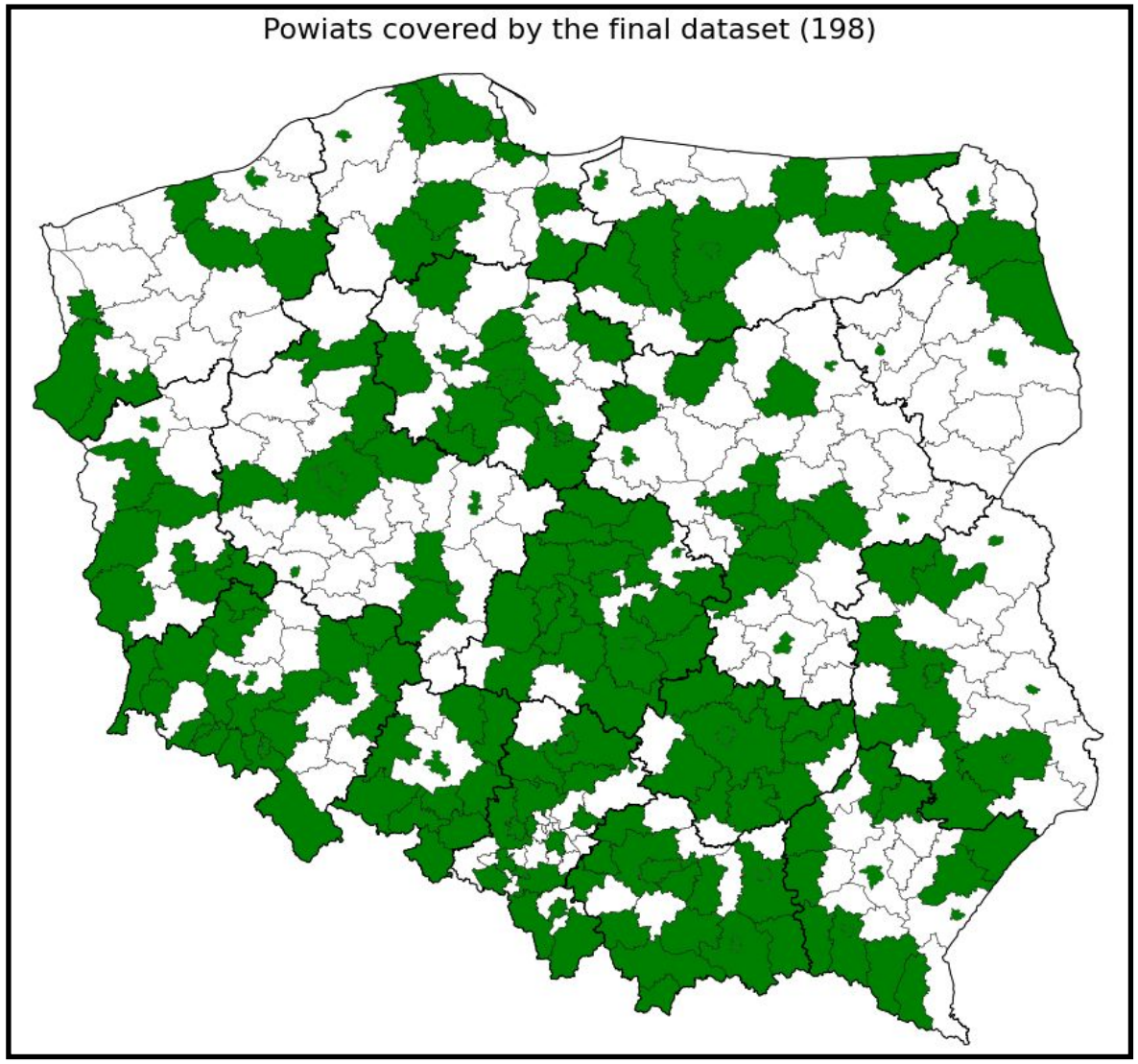

Powiat proximity (Warsaw)
Warsaw powiat proximity
1. powiat Ruszkowski 20.2km
2. powiat Iaseczyński 24.3km
3. powiat Egionowski 25.1km
4. powiat Arszawski zachodni 26.8km
5. powiat Ołomiński 29.2km
6. powiat Twocki 29.4km
7. powiat Rodziski 35.0km
8. powiat Owodworski 35.8km
9. powiat Iński 42.1km
powiat Yrardowski 49.1km
powiat Ułtuski 49.7km
powiat Rójecki 50.8km
powiat Yszkowski 54.4km
powiat Ochaczewski 57.0km

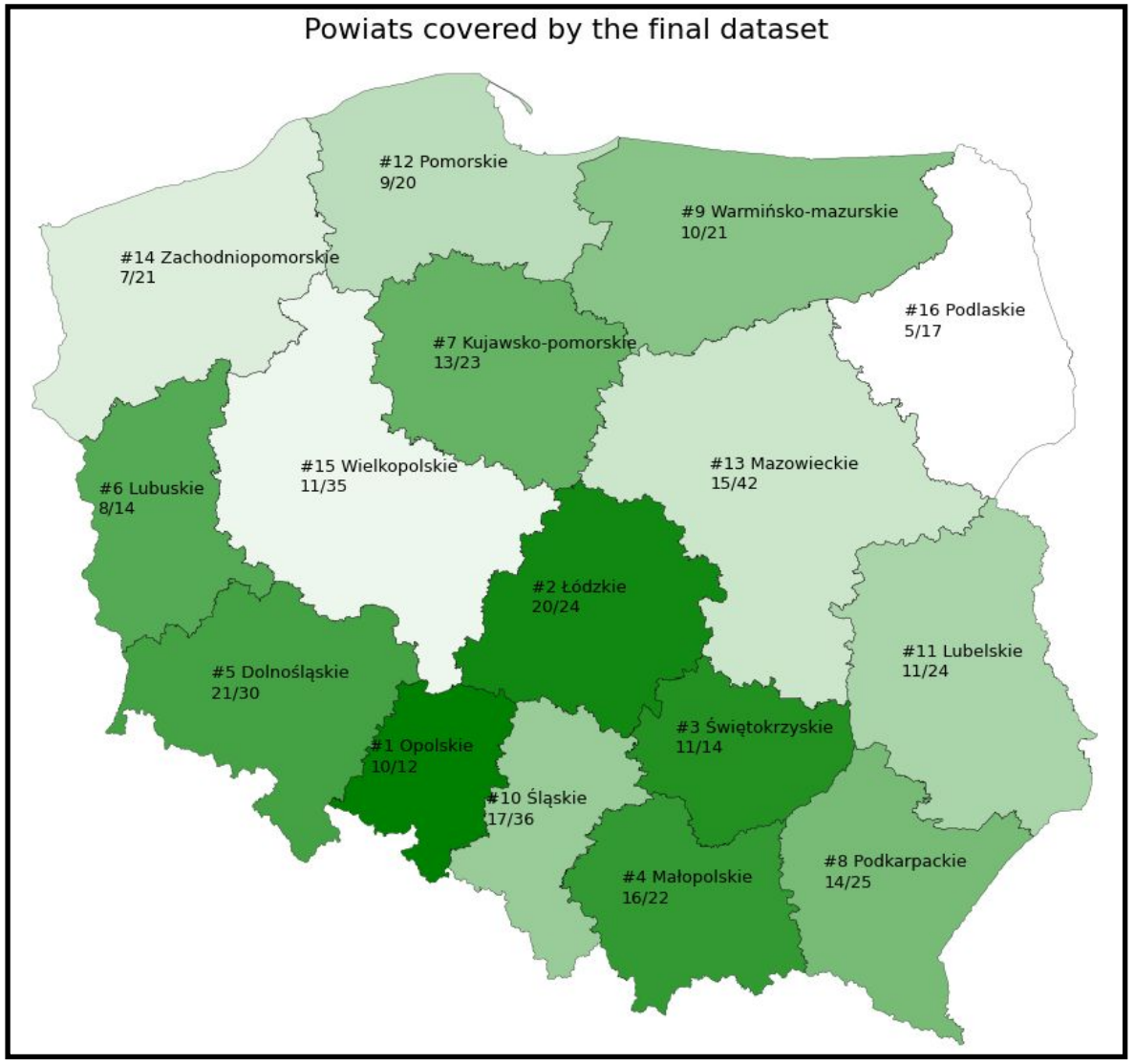
Merging All Datasets

- 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**.
- Decided to **impute the missing weather data** for these Powiats using data **from the closest neighboring Powiats** before joining.
- **Powiat-proximity matrix** used to **get closest neighboring Powiat and impute the weather data from the neighboring Powiat**.
- **Remaining Powiats** with missing weather data imputed **using the average daily weather data of their corresponding voivodeships**.

After imputation and aggregating, weather data was joined with the Air Quality dataset, followed by **merging static datasets to create a final single dataset.**

Data Cleaning and Preprocessing - Merging Datasets

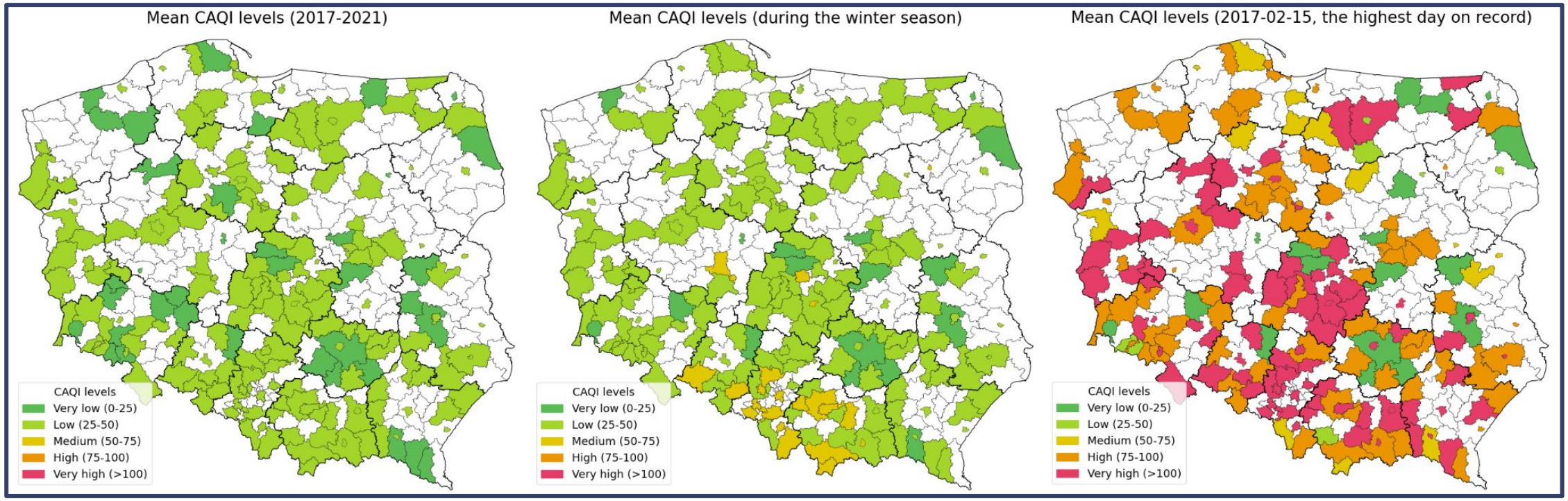




Calculation of CAQI Values

- The CAQI or Common Air Quality Index was proposed to facilitate the comparison of air quality in European cities in real-time.
- 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.

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



Exploratory Data Analysis

Exploratory Data Analysis (EDA) - Pollutants

- **Particulate Matter (PM10 & PM2.5)**

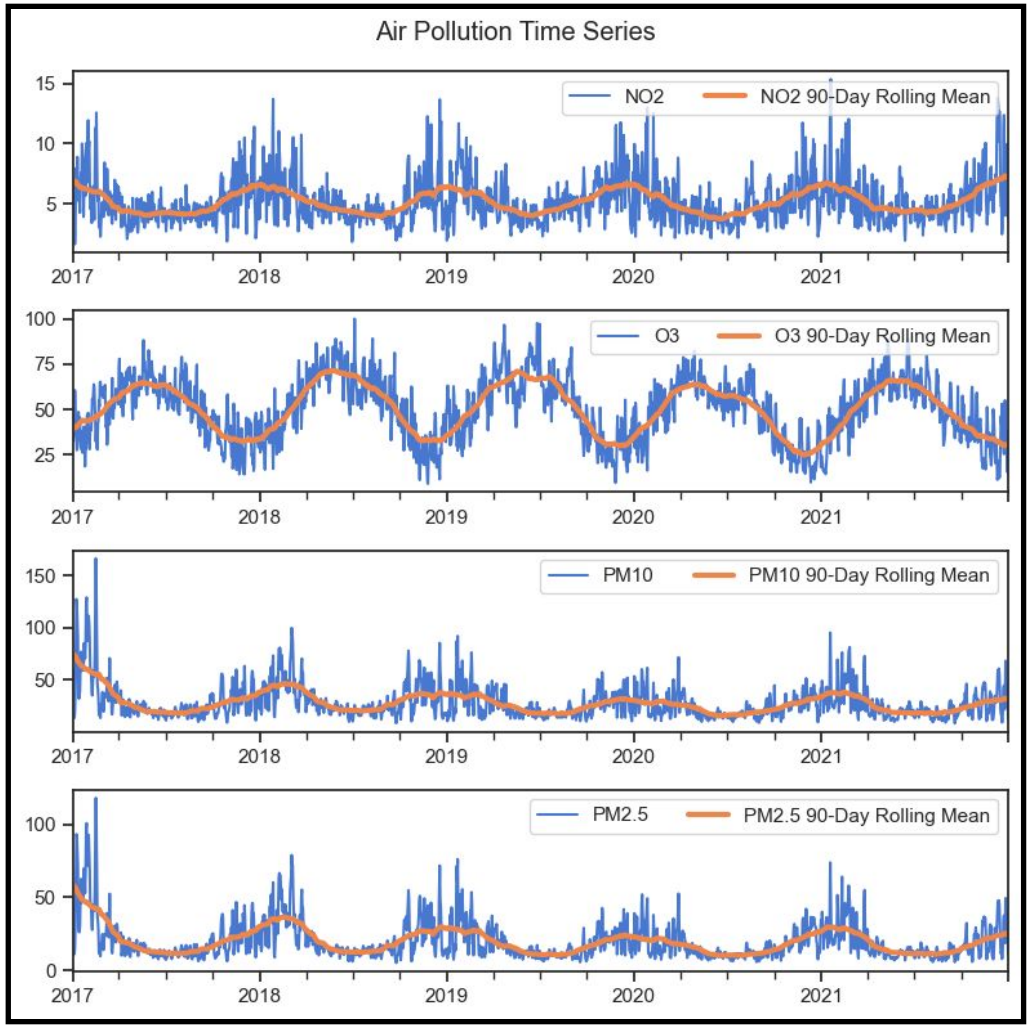
  Inhalable particles, with diameters that are generally 10 micrometers and smaller; or 2.5 micrometers and smaller respectively.

- **Nitrogen dioxide (NO2)**

  Main contributors are combustion processes in energy production, manufacturing industry and road transport.

- **Ozone (O3)**

  Formed when heat and sunlight cause chemical reactions between oxides of nitrogen (NOx) and Volatile Organic Compounds (VOC)



Exploratory Data Analysis (EDA) - Pollutant Seasonality

- **NO2, PM10 and PM2.5 lower during weekends.** But **O3 higher during weekends**. This is called **"weekend effect"**.

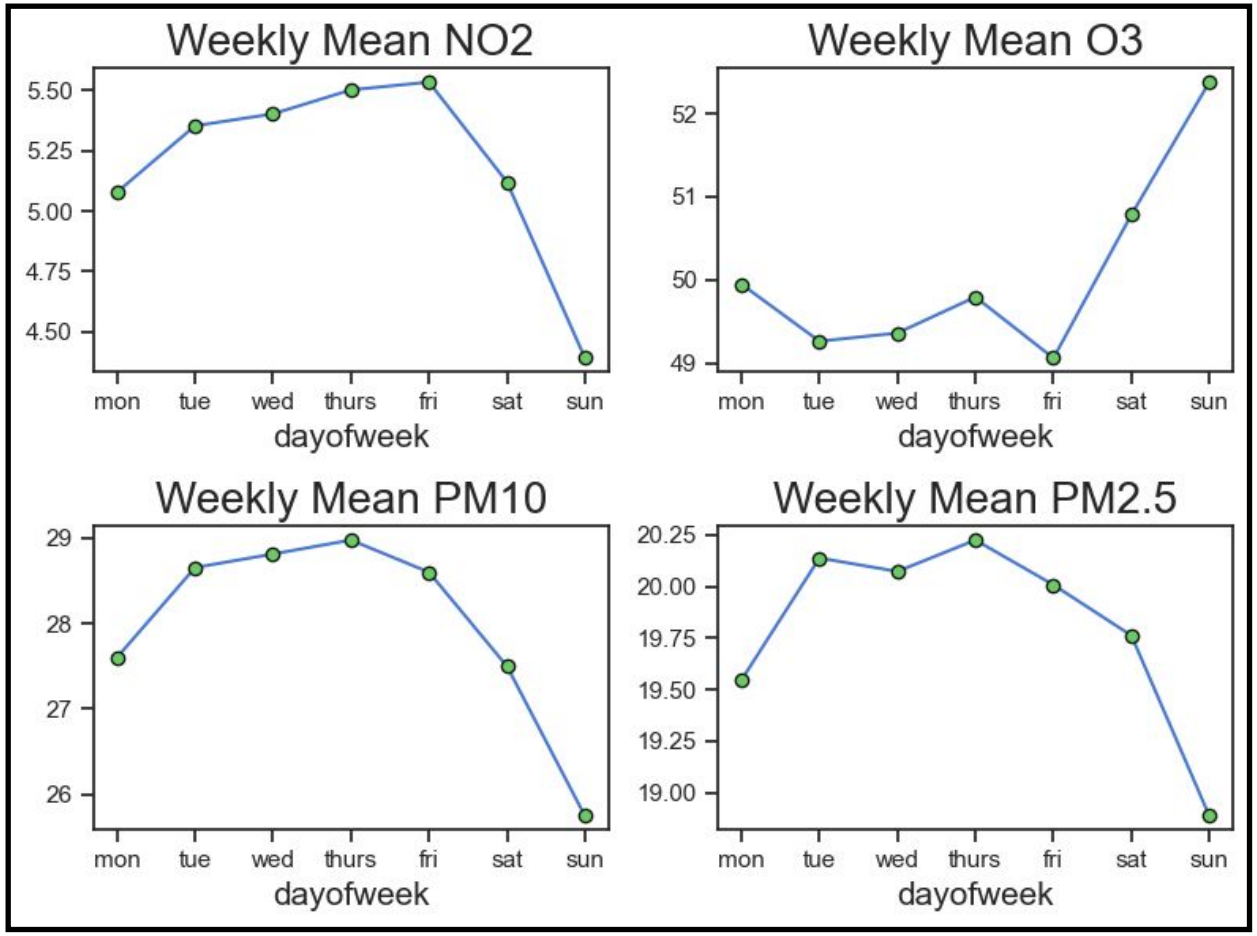
- Ozone is a secondary pollutant - it is not directly emitted by traffic, industries, etc. but is formed on warm summer days by the influence of sunlight and pollutants such as NOx emissions.

$$NO_2 + O_2 + (\text{solar UV-light, + heat}) \rightarrow NO + O_3$$

- NOx at higher concentrations tends to degrade O3.

$$NO + O_3 \rightarrow NO_2 + O_2$$

- During weekends less traffic means reduction of NOx emissions. This subsequently does not suppress O3. Hence leading to higher concentration of O3 during weekends.

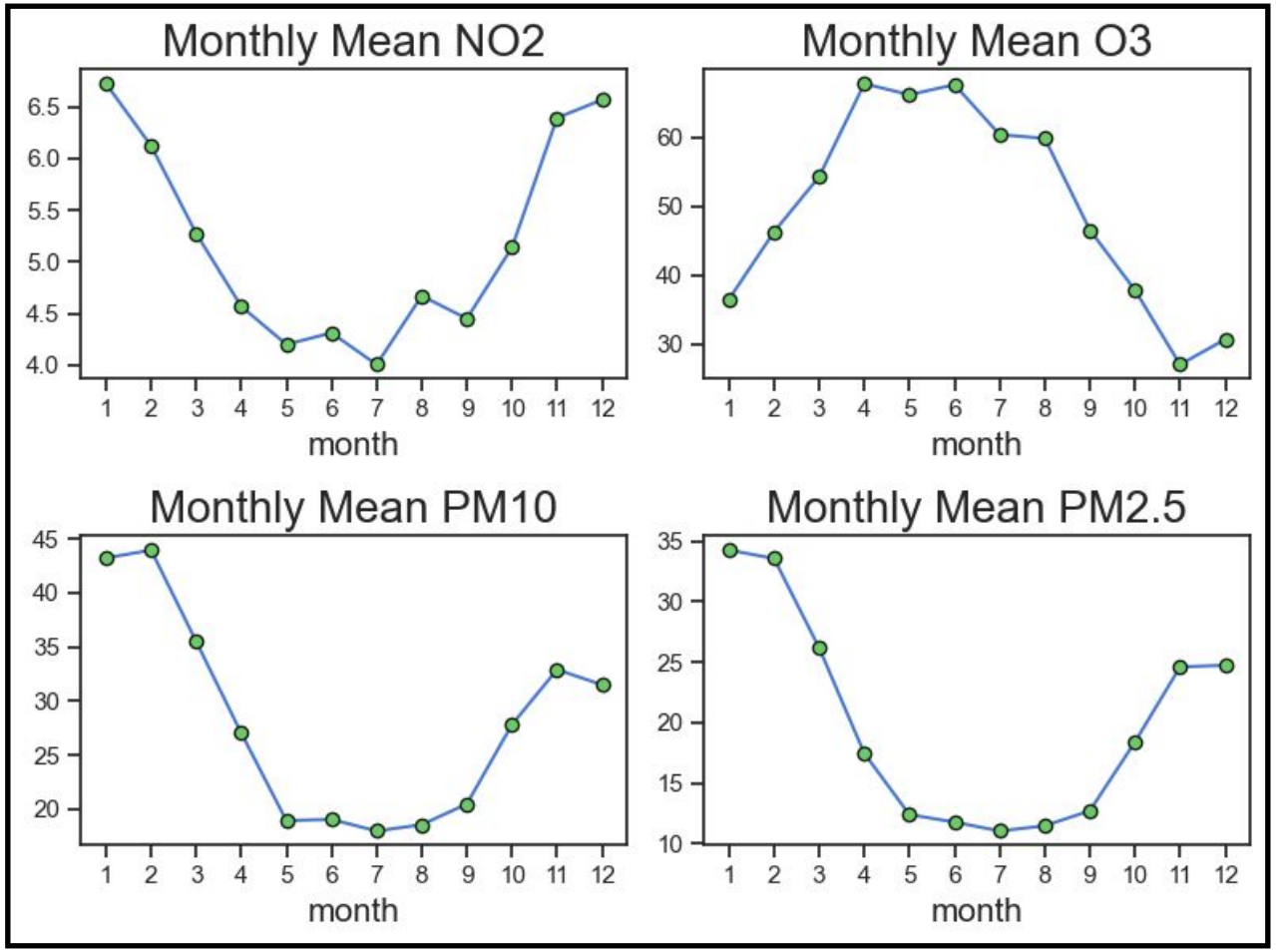


Exploratory Data Analysis (EDA) - Pollutant Monthly Seasonality

- O3 levels higher during spring and summer. Conversely, the rest of the pollutants are higher during Autumn and Winter..

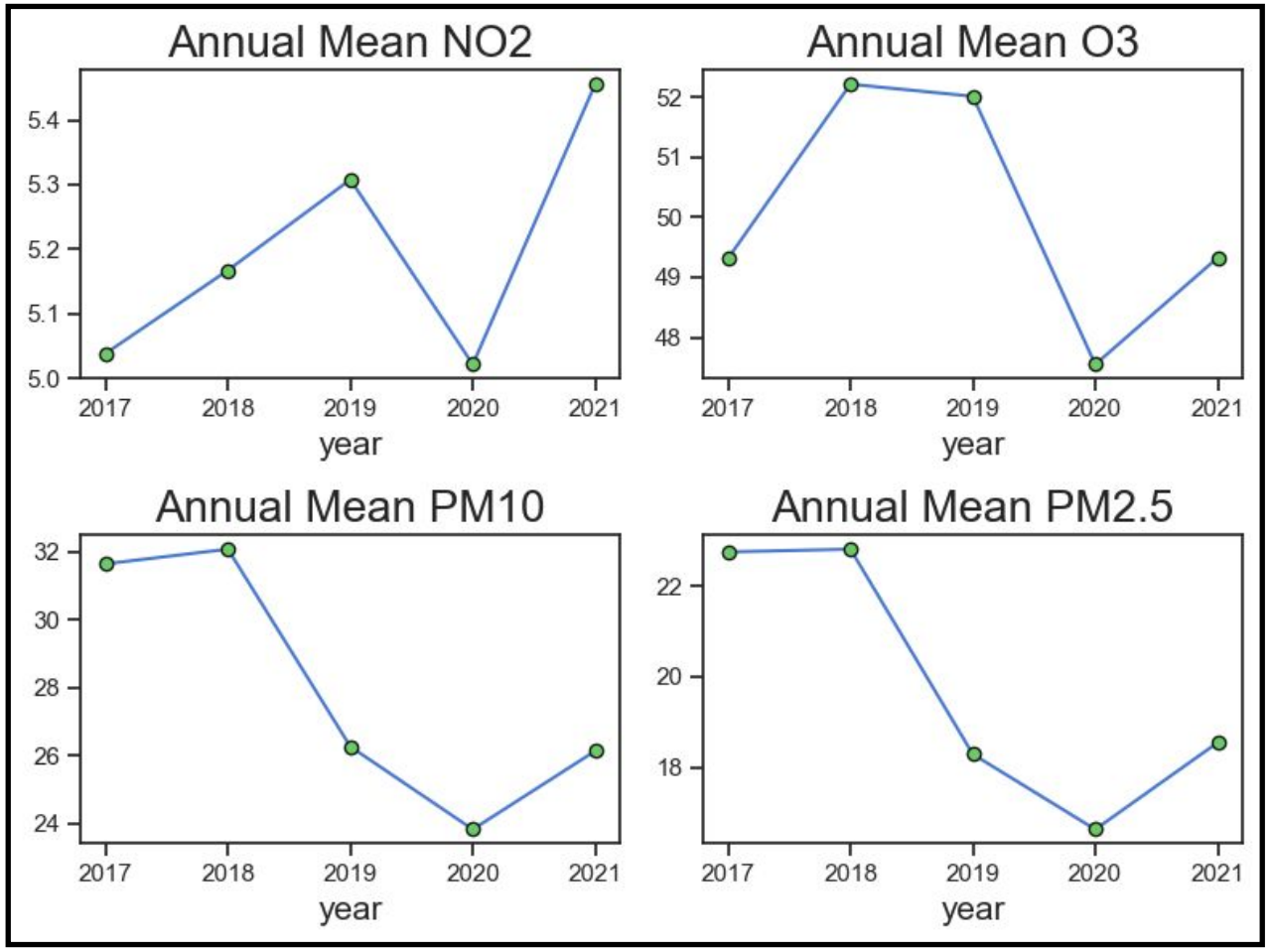
- Higher concentrations of O3 during spring and summer is due high temperatures and intensive solar radiation during those months.

- High levels of NO2 and particular matter emissions during Autumn and Winter is mostly attributed to intensive burning of low-quality coal in coal furnaces for heating.

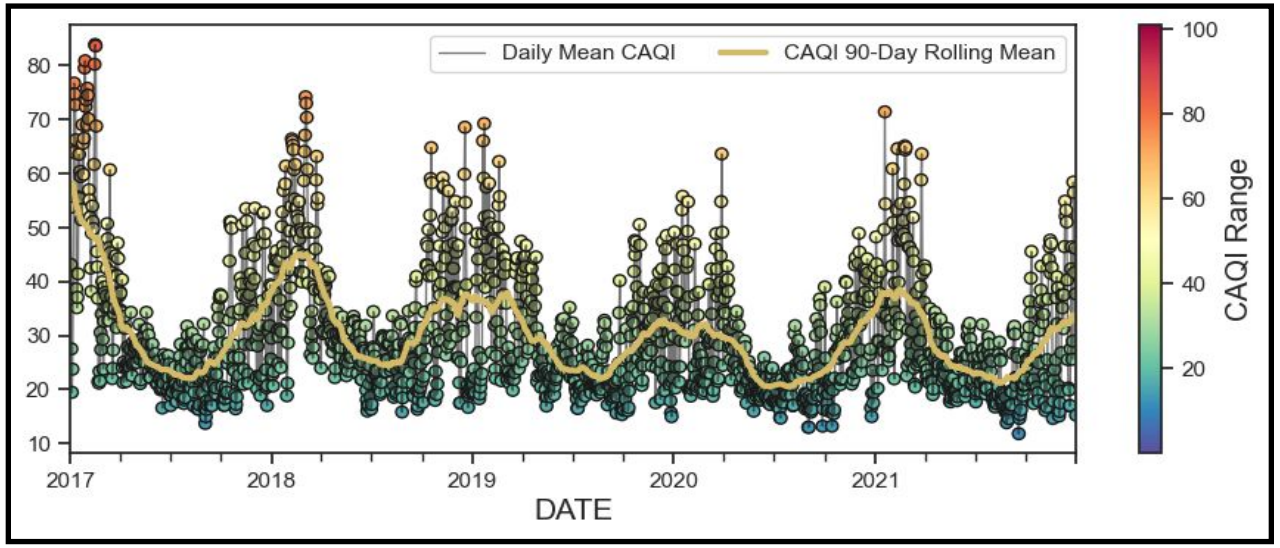


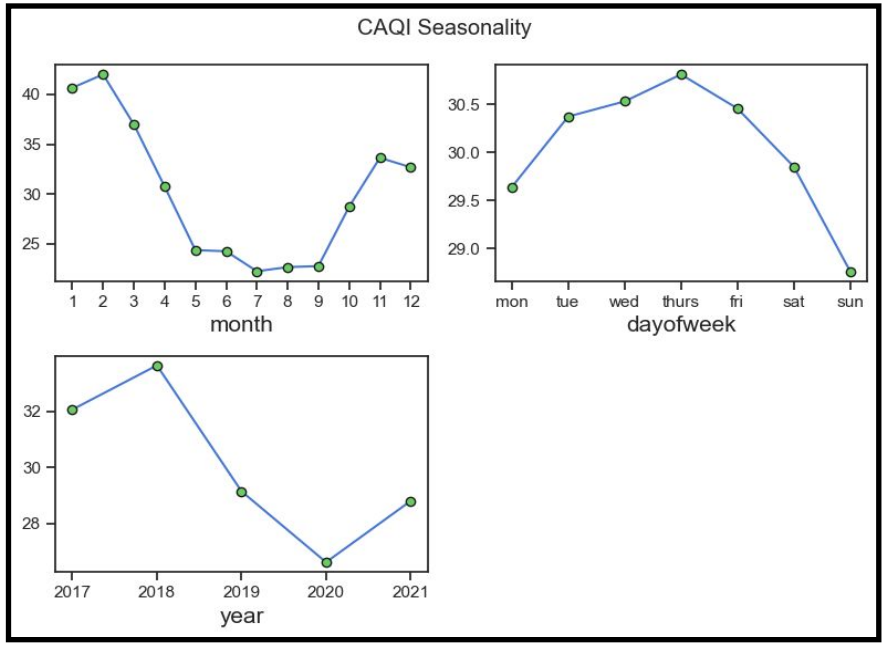
Exploratory Data Analysis (EDA) - Pollutant Annual Pattern

All pollutants observed lowest levels on 2020. This is mainly attributed to COVID19 lockdowns.



Exploratory Data Analysis (EDA) - Analyzing CAQI (Common Air Quality Index)

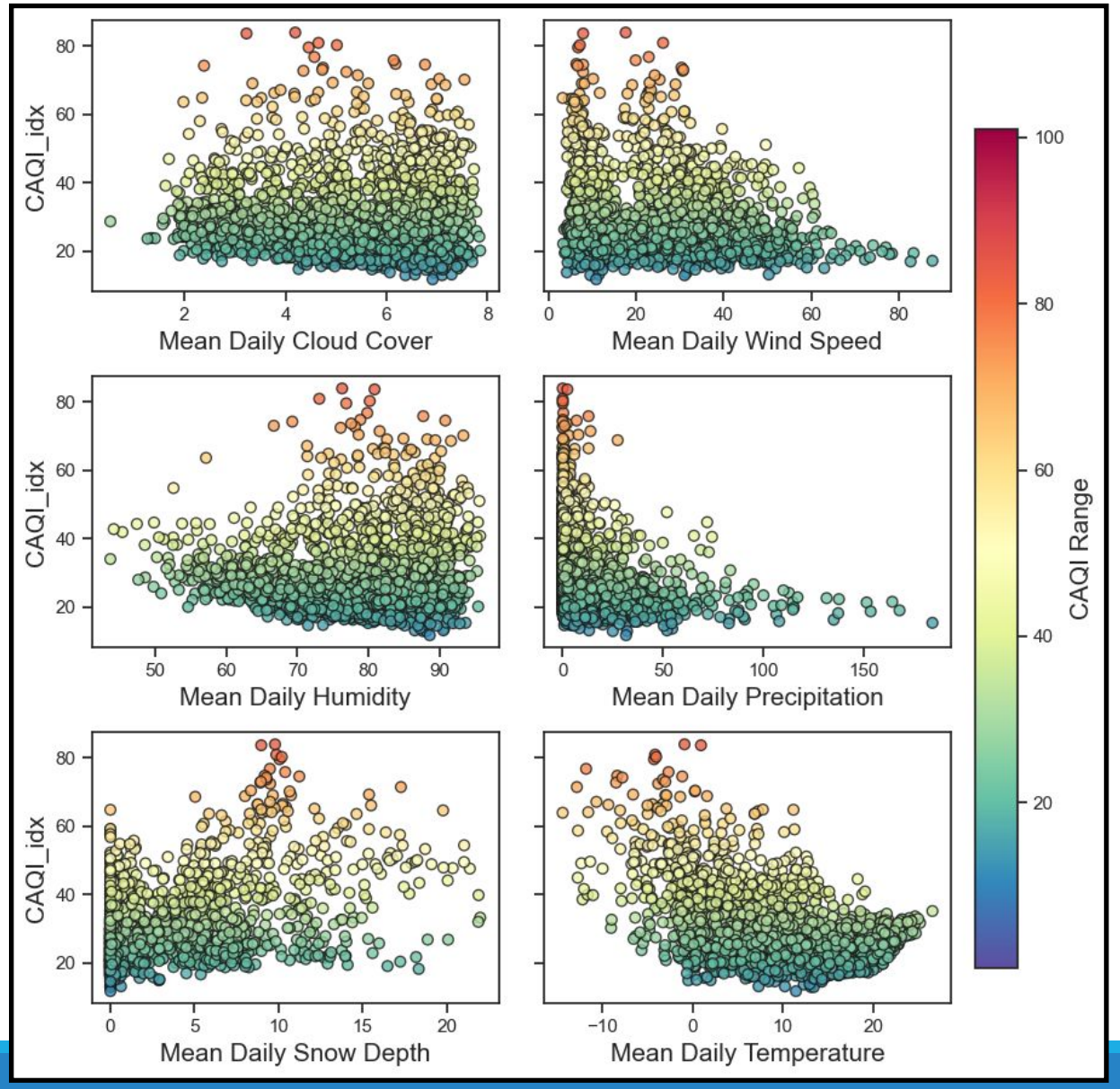




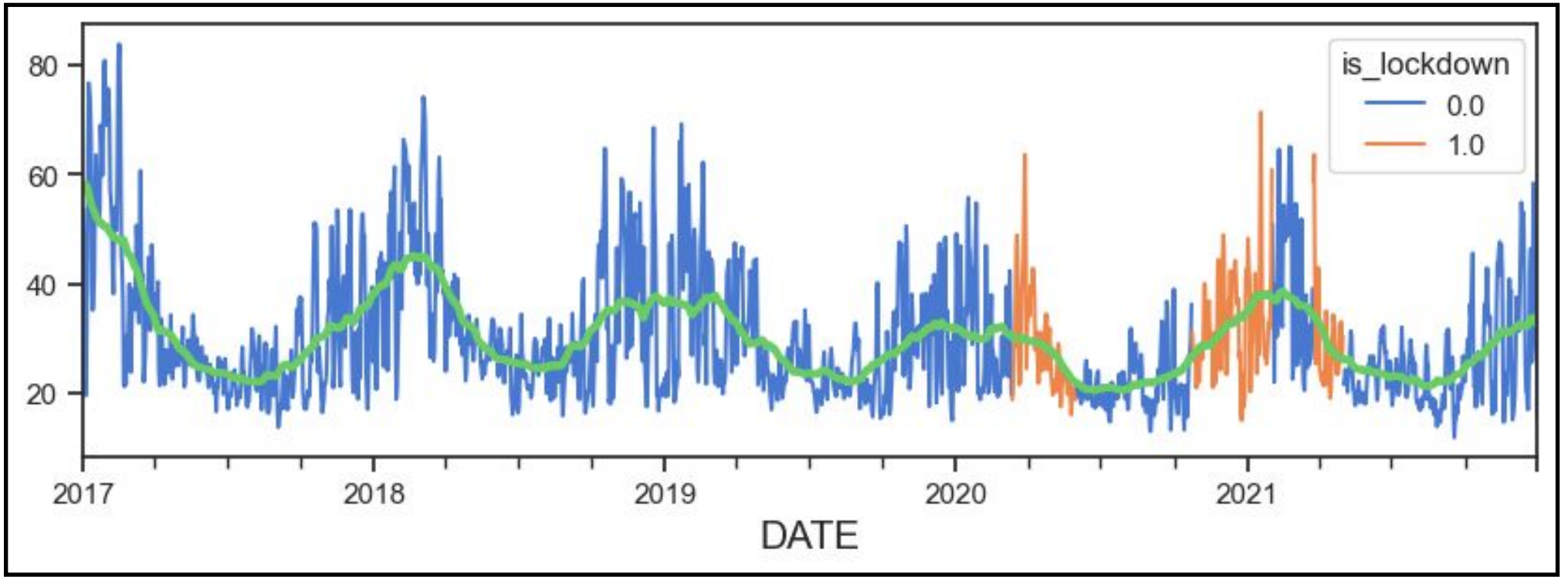
- CAQI used to understand the general Air Quality in Europe based on all mentioned pollutants.
- CAQI levels are lowest in the weekends.
- CAQI levels are lower during Spring and Summer.
- CAQI levels are seen to be at lowest during 2020.

Exploratory Data Analysis (EDA) - CAQI Levels vs Weather

- Higher CAQI levels indicate worser Air Quality.
- CAQI levels evenly distributed from low to high levels of Cloud Cover.
- Groups of high CAQI levels are located at the lower Wind Speed ranges.
- 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.
- There is a clear non-linear relation between Temperature and CAQI values, where CAQI levels are higher at lower Temperatures.

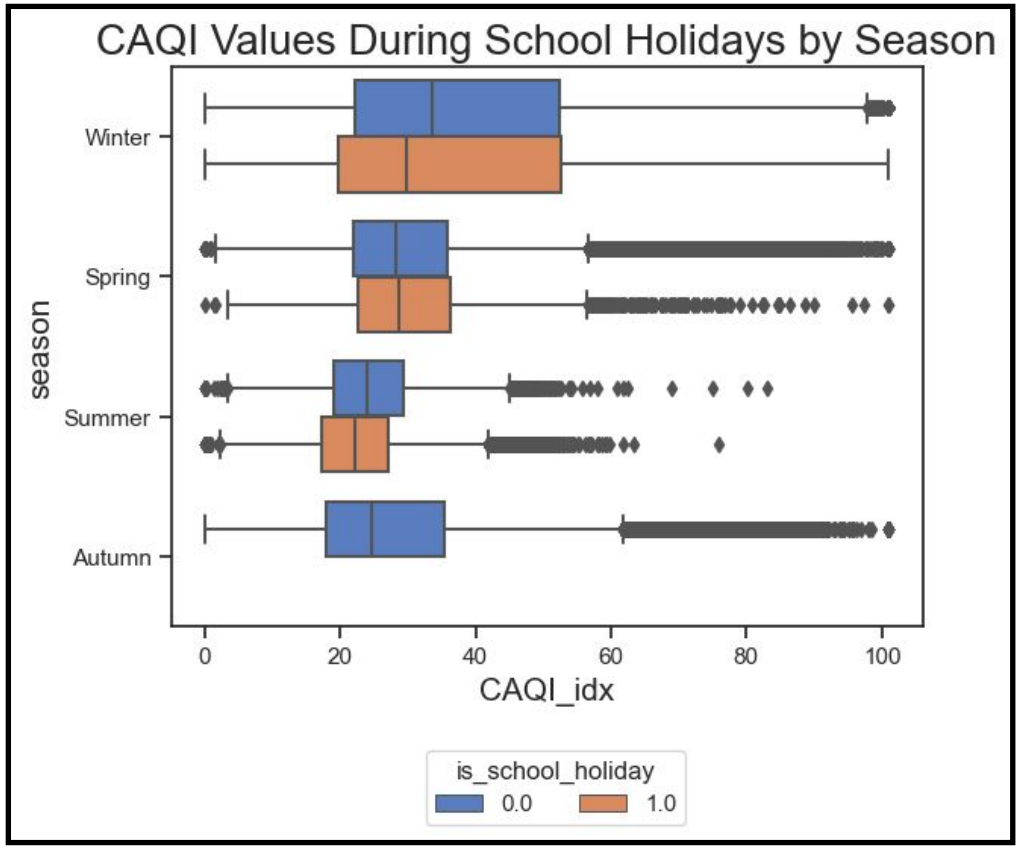


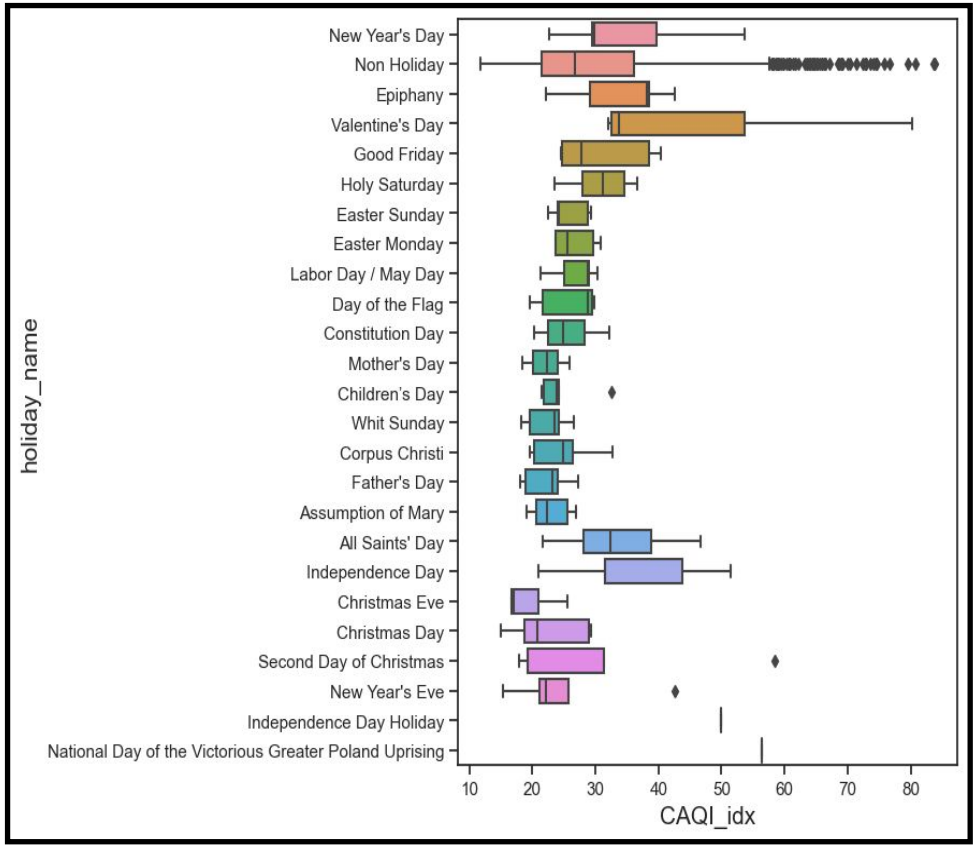
Exploratory Data Analysis (EDA) - Impacts of COVID-19 Lockdown



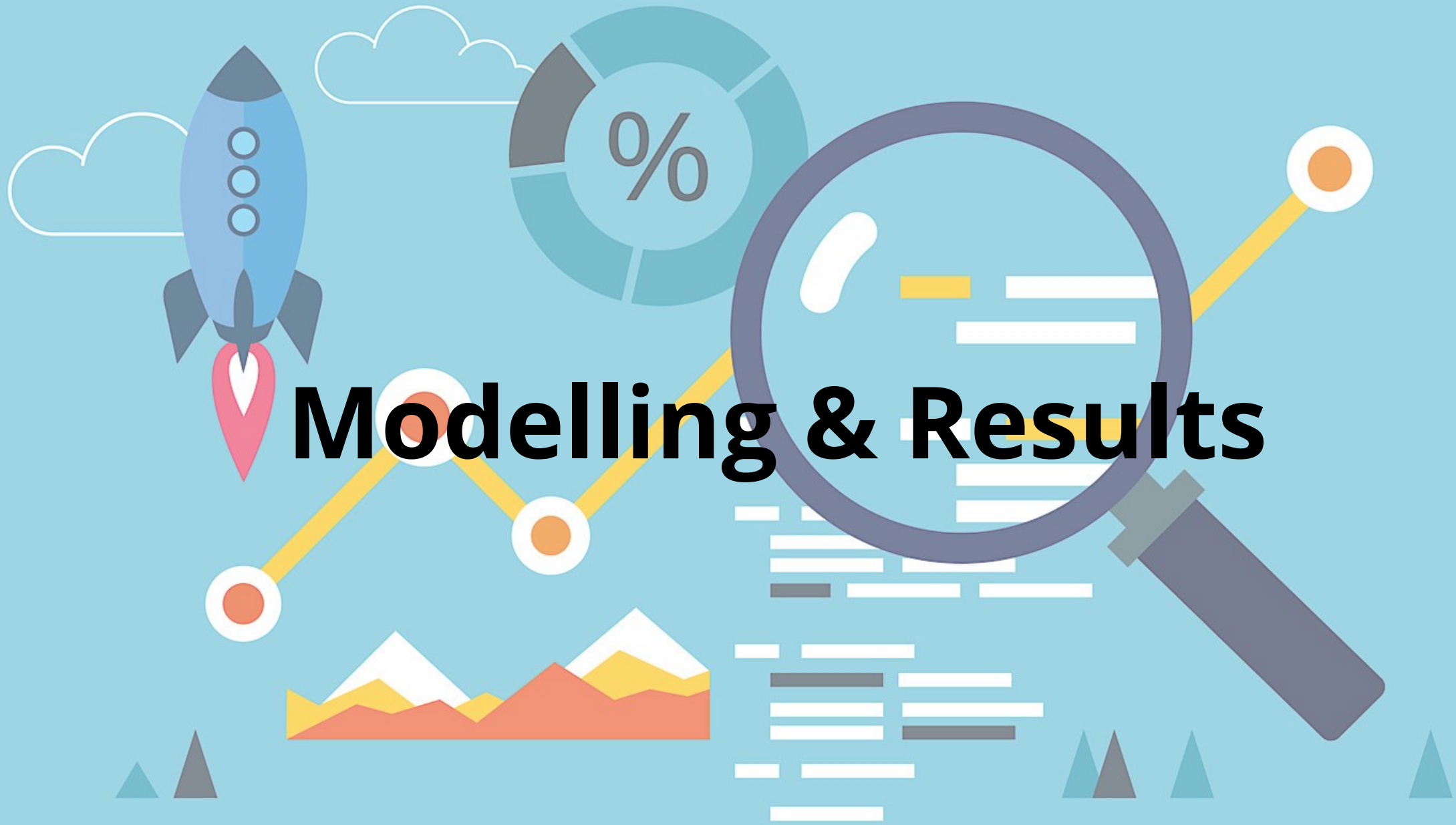
- Orange shaded parts denote lockdown periods.
- Noticable drop in CAQI levels way before the COVID-19 lockdown during 2020

Exploratory Data Analysis (EDA) - Impact of Public Holidays and School Holidays





- Valentines Day, All Saint's Day, Independence Day and National Day of the Victorious Great Poland Uprising sees a higher median level of CAQI levels.
- School holiday also has effect on overall CAQI levels.



Modelling & Results

Modeling - Feature Engineering

- **Lag Feature**: 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.
- **Rolling window Mean**: Computed Rolling Mean of CAQI index from previous 30 days.
- **DateTime features**: Extracted 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.
- **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).
- **Combining Trasnformations**: Many static features were combined by taking the sum of their values in order to represents the feature in a parse manner.
- **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

Modeling Results

- Entire dataset was used to generate two main datasets which are for training machine learning models for both classification and regression models.
  - Train: 1st Jan 2017 to 28th Feb 2021
  - Test: 2nd Mar 2021 to 31st Dec 2021

- Following techniques were utilized to identify optimal model and improve models' performance:
  - Feature Selection
  - Feature engineering
  - Oversampling
  - Cross validation
  - Hyperparameter optimization
  - Feature importance

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

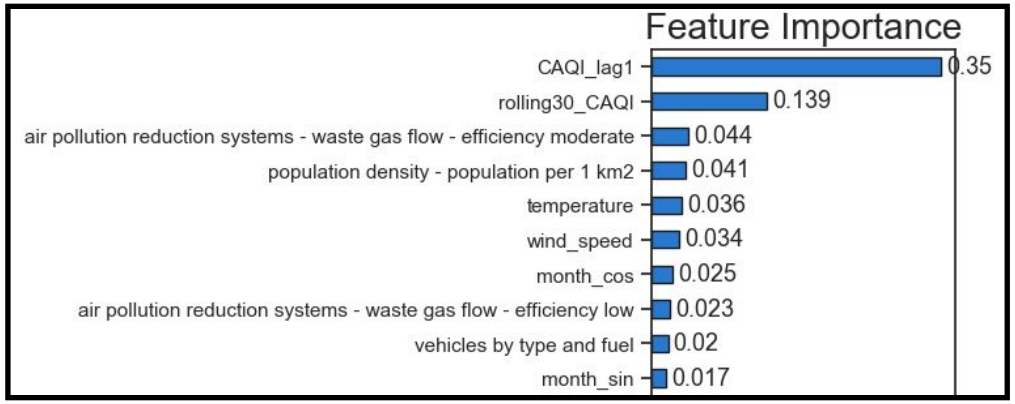
Classification models for predicting CAQI levels

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

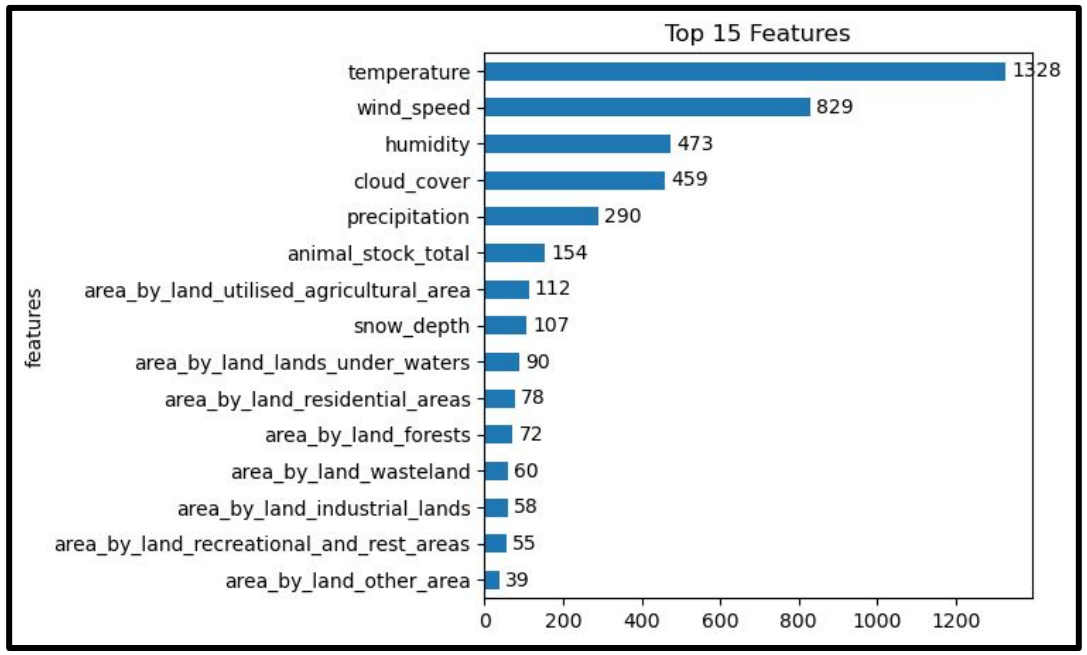
Regression models for predicting CAQI indexes

Feature Importances of various Models

- Feature Importance plots of best performing models (Classification/Regression)..



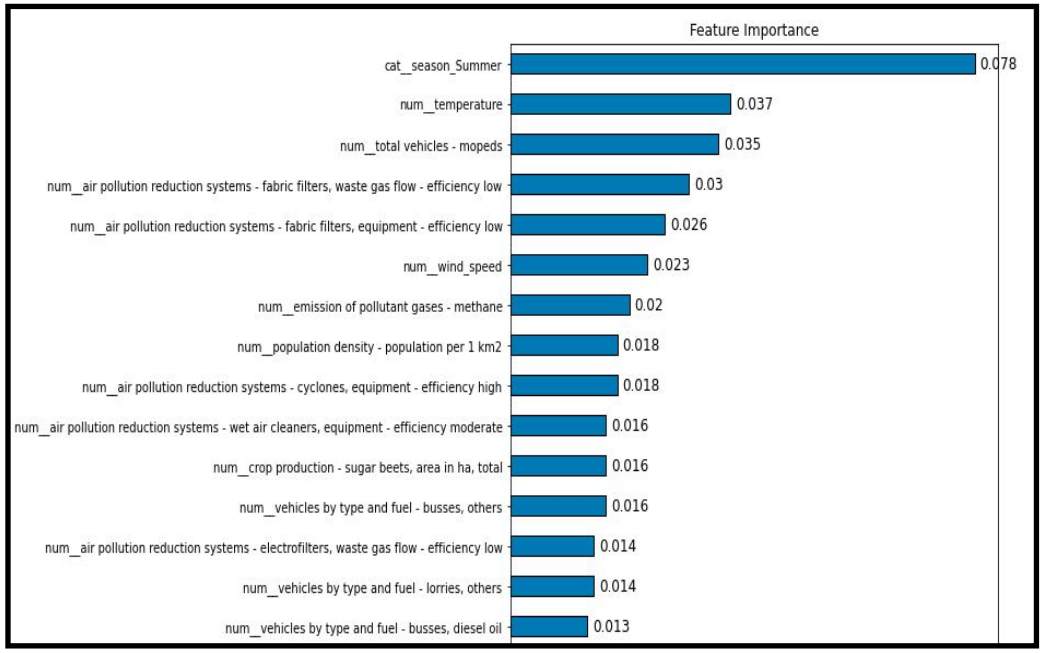
XGBoost Regressor feature importance plots on Temporal Dataset



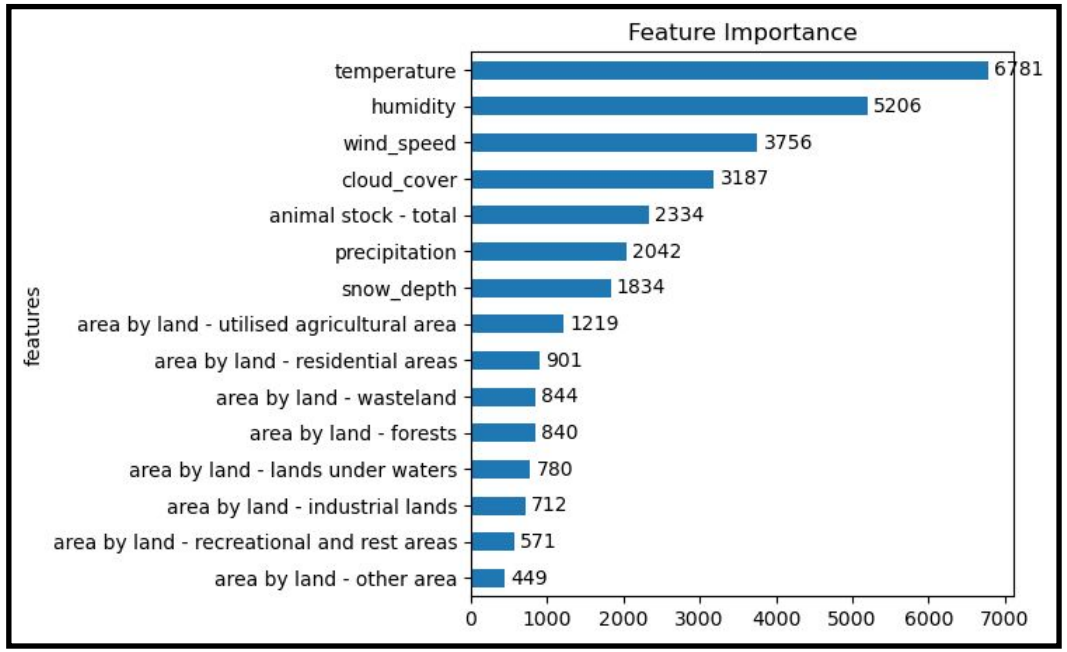
LightGBM Classifier feature importance plot on Temporal Dataset

Feature Importances of various Models

- Feature Importance plots of best performing models (Classification/Regression)..



XGBoost Regressor feature importance plots on Non-Temporal Dataset

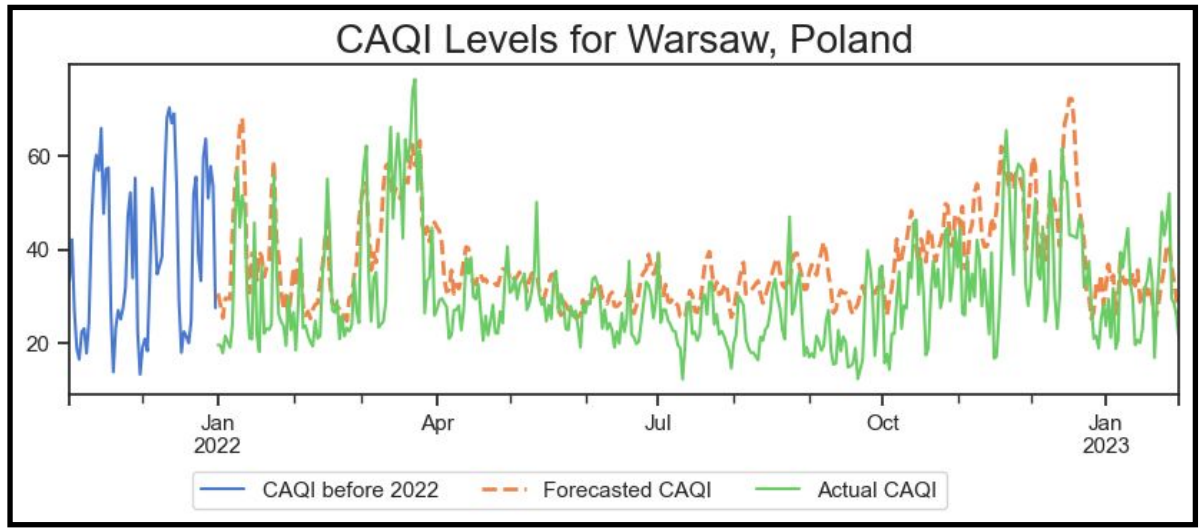


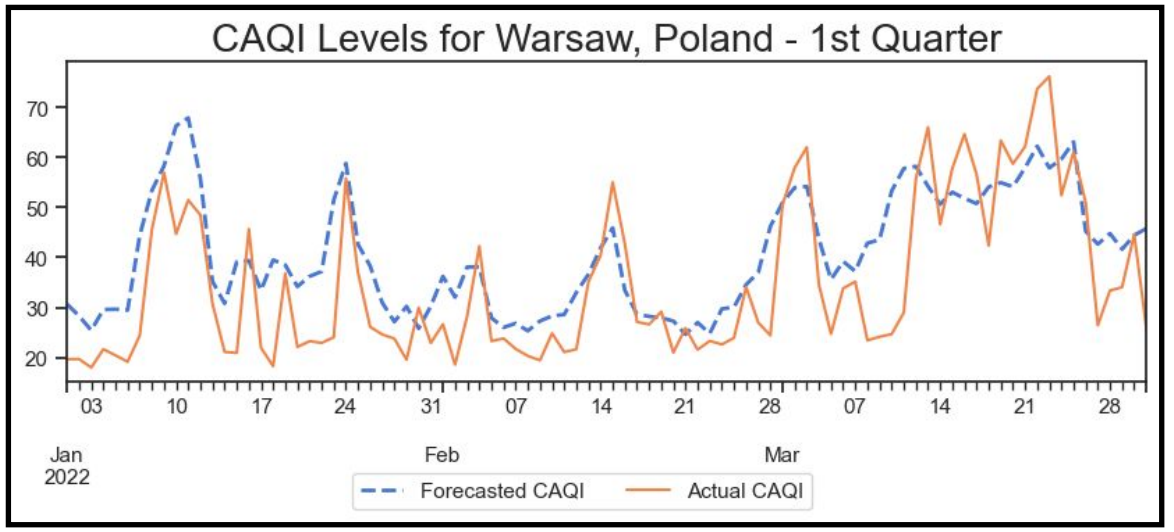
LightGBM Classifier feature importance plot on Non-Temporal Dataset

Observations

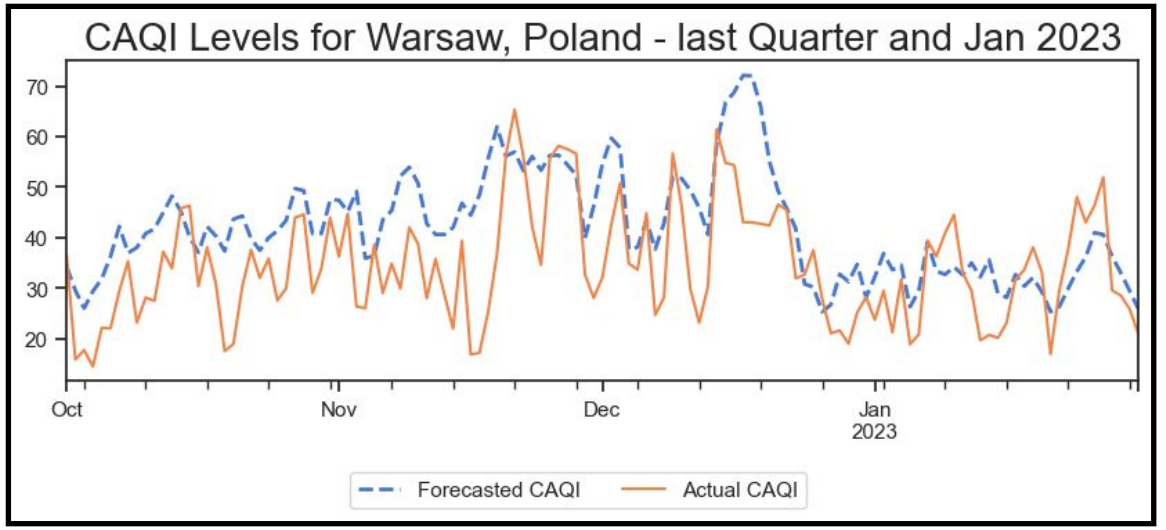
- **Feature importance varies across all models** and type of dataset.
- Weather features such as **Temperature and Wind speed are commonly ranked high** in all models.
- **Features related to human activities** such as air pollution reduction systems, land area, etc. are also ranked high, but these 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 are **only at an annual level, not at a daily level**.
- Recording data at a daily level is valuable in identifying AQ levels in Poland and assisting policy makers in taking appropriate measures to mitigate any negative impacts.
- For eg, 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.

Results - Forecasting Future CAQI Levels





- **XGBoost regressor** model trained on both train & test data to forecast CAQI levels of Warsaw from Jan 2022 to Jan 2023.
- The model generalized well to the data. Successful captured and trend and seasonality of the CAQI levels.
- As the forecasting horizon increases, we can see the model tends to over predict CAQI levels



Summary

- Project objective: **investigate factors responsible for air pollution in Poland** and develop a tool to predict air quality.
- Key factors contributing to air pollution: **PM_{10} and $PM_{2.5}$ from solid fuel combustion, road traffic, construction sites**, etc. **NO_2 from combustion processes in energy production**, manufacturing industry, and road transport. **O_3 is a secondary pollutant** formed by sunlight and combination of airborne pollutants, including NOx and VOCs.
- **Common Air Quality Index (CAQI)** used to standardize the air quality measurement .
- Using appropriate ML techniques, a suitable model was chosen to forecast future CAQI levels of Warsaw, Poland accurately.
- Overall, **Air Quality levels are heavily influenced by human activities, followed indirectly by prevailing weather conditions.**
- Daily monitoring and rigorous data collection related to various features affecting air quality can help policymakers and stakeholders identify areas of concern and take appropriate actions to mitigate the impact of air pollution on public health and the environment.


THE END