Investigating Causes and Impacts of Air Pollution in Poland

Dong Liang, Hudson Passos, Intan Pamungkas,
Qin Xu, Sabrina Ramadwiriani

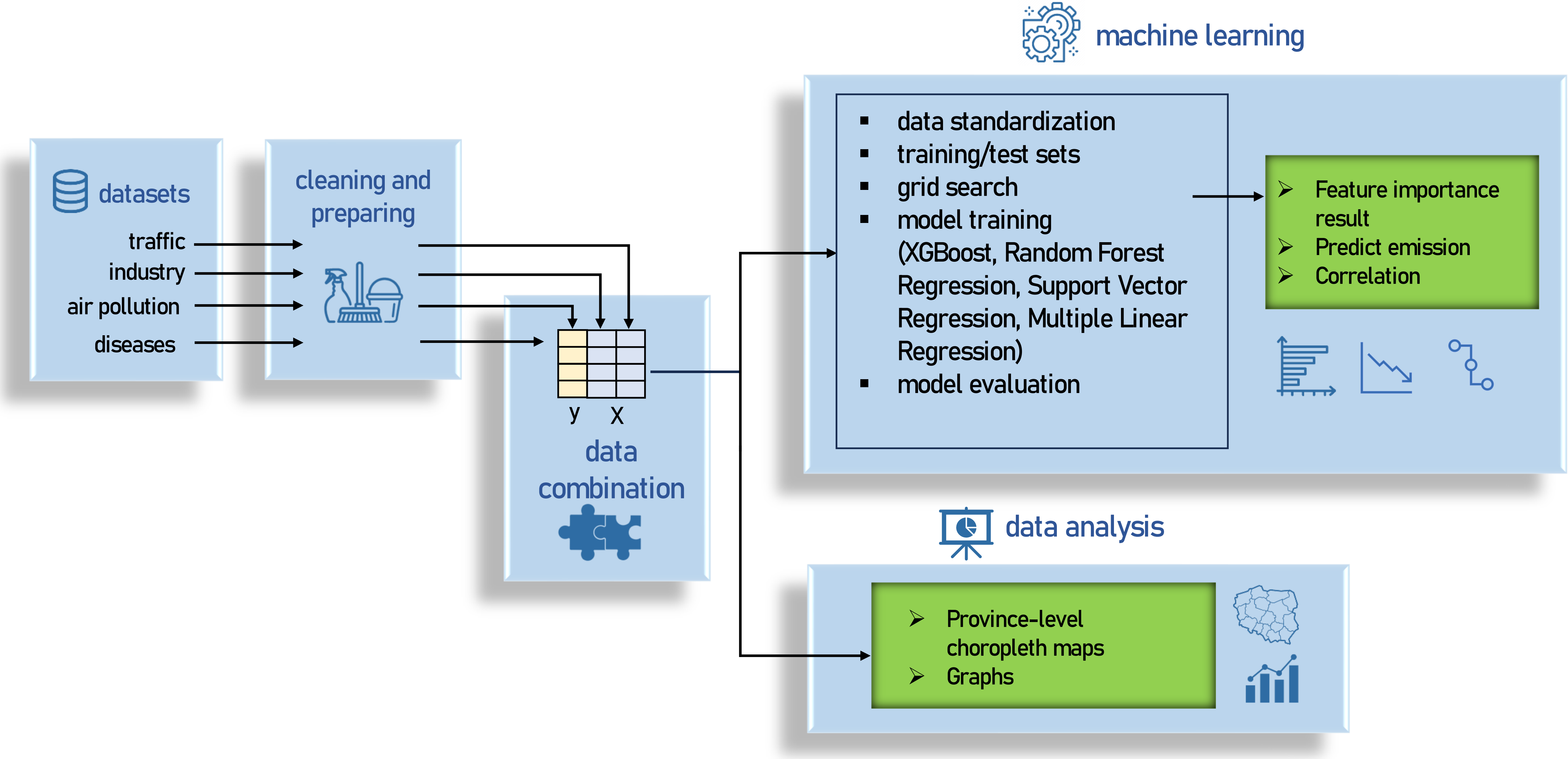
CONTEXT AND GOALS

Poland has been a major contributor to European dust, sulfur dioxide, and nitrogen oxides emissions (Sawicka-Kapusta & Zakrzewska, 1998). The detrimental health effects of air pollution in Poland include increased mortality rates, higher prevalence of respiratory diseases such as asthma and lung cancer, and a higher risk of COVID-19 infections (Nazar & Niedoszytko, 2022).

This project focuses on understanding the causes and impact of the air pollution in Poland through several objectives:

1. Analyze the contribution of sources of Nitrogen Oxides (NOx) to the air pollution in Poland in 2019.
2. Investigate the relationship between traffic and emissions using machine learning and predict the emission in 2023.
3. Investigate the relationship of annual average concentration of PM 2.5 and rate of death caused by respiratory diseases.

WORK-FLOW



machine learning
datasets
traffic
industry
air pollution
diseases
cleaning and preparing
y
X
data combination
data standardization
training/test sets
grid search
model training (XGBoost, Random Forest Regression, Support Vector Regression, Multiple Linear Regression)
model evaluation
Feature importance result
Predict emission
Correlation
data analysis
Province-level choropleth maps
Graphs

DATA AVAILABILITY

FINDING DATA WITH THE SAME RANGE

Type of pollutant available:

type information	type variable	pollutants available														
Air pollution	Target	BC	CO	NO	NO_2	-	O_3	-	PM10	PM2.5	SO_2	-	-	-	-	-
Air pollution	Target	-	CO	NO	NO_2	-	O_3	-	PM10	PM2.5	SO_2	-	NO_x	-	-	-
Traffic	Explanatory	-	-	-	-	N_2O	-	-	-	PM2.5	-	NMVOC	NO_x	CH_4	CO_2	-
Seaports	Explanatory	-	-	-	-	-	-	PM	-	-	-	-	NO_x	-	CO_2	SO_x
Industries	Explanatory	-	-	-	-	N_2O	-	-	PM10	-	-	NMVOC	NO_x	CH_4	CO_2	SO_x

Administrative level:

Air pollution (target): point → interpolation (OK) → mean value for each **province**

Traffic (explanatory): **province**

Industries (explanatory): point → sum of emissions for each **province**

Timeframe:

Air pollution (target): hourly → median of the **year**

Traffic (explanatory): **yearly**

Industries (explanatory): **yearly**

Measure unit:

Air pollution unit: **$\mu g/m^3$ (median per year)**

Industrial emissions: **kg (per year)**

Traffic: tonnes per year → **kg (per year)**

Data Sources:

GIS support

European Environment Agency | Datahub

World Resources Institute

eurostat

GUS | TranStat

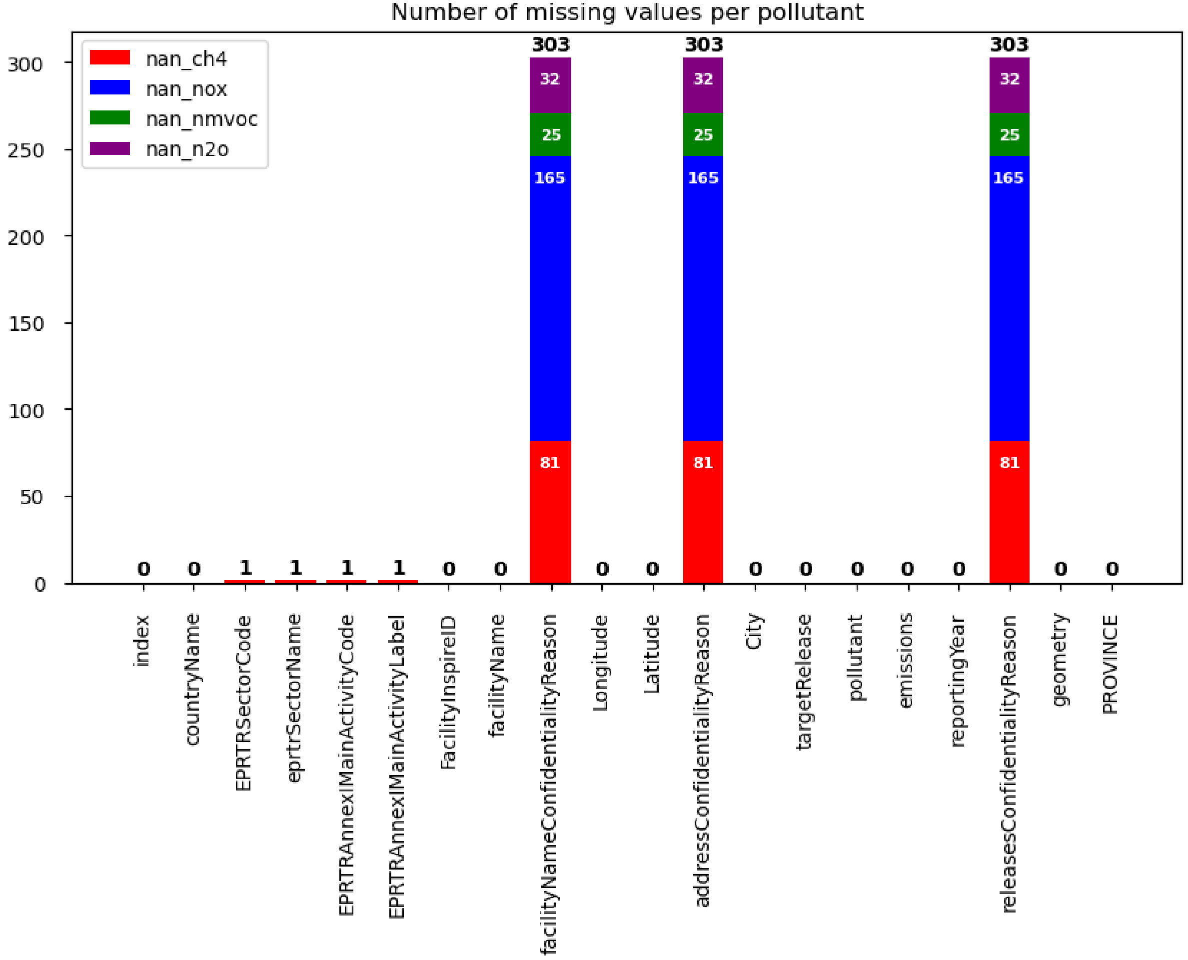
DATA CLEANING

REMOVING “NOT A NUMBER”, INCONSISTENT VALUES AND DATA WITHOUT COORDINATES

Air pollution

	nan	-999	0
Nr	0	0	0
Kod stacji	0	0	0
Kod międzynarodowy	353	0	0
Nazwa stacji	0	0	0
Stary Kod stacji \n(o ile inny od aktualnego)	814	0	0
Data uruchomienia	0	0	0
Data zamknięcia	267	0	0
Typ stacji	0	0	0
Typ obszaru	0	0	0
Rodzaj stacji	0	0	0
Województwo	0	0	0
Miejscowość	0	0	2
Adres	58	0	0
WGS84 φ N	0	76	0
WGS84 λ E	0	76	0

Industry



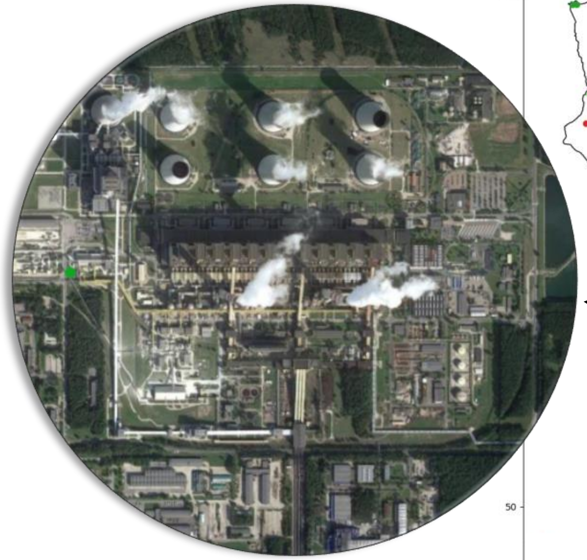
Traffic & Respiratory Diseases

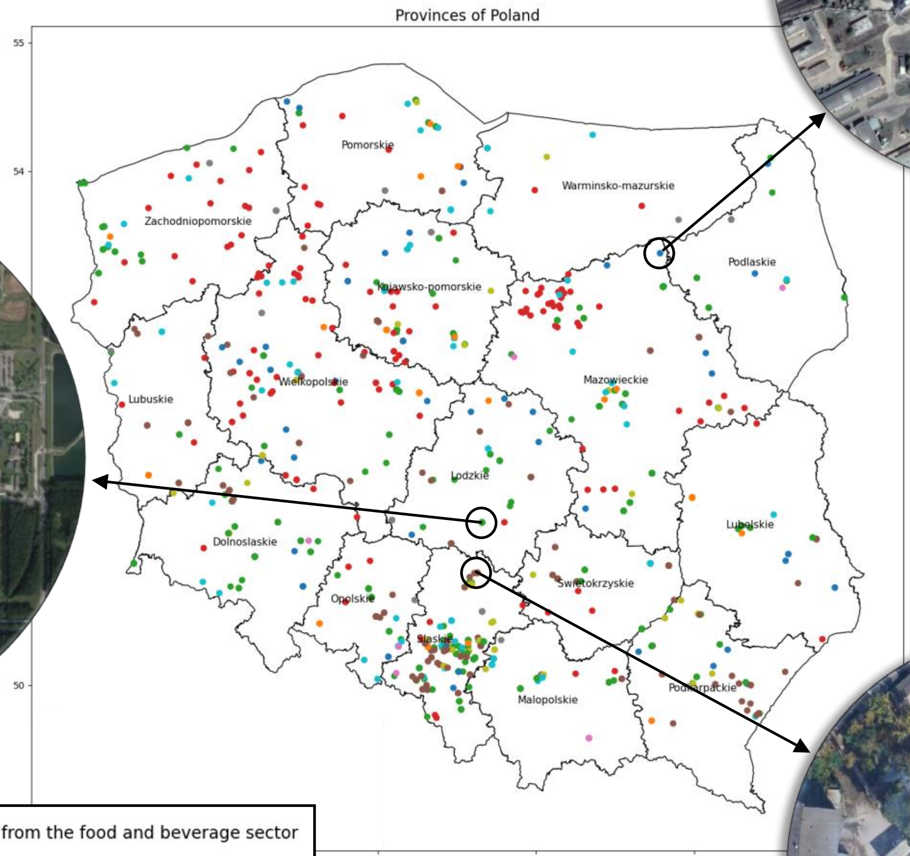
- Small dataset
- No need of cleaning

INDUSTRIES DATASET

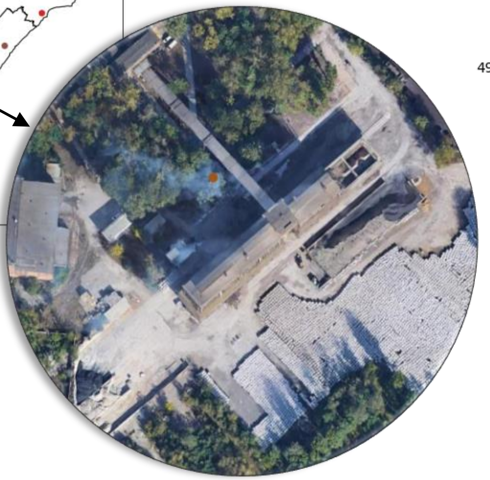
INFORMATION: “INDUSTRIAL SECTOR” AND “EMISSIONS”

Energy sector

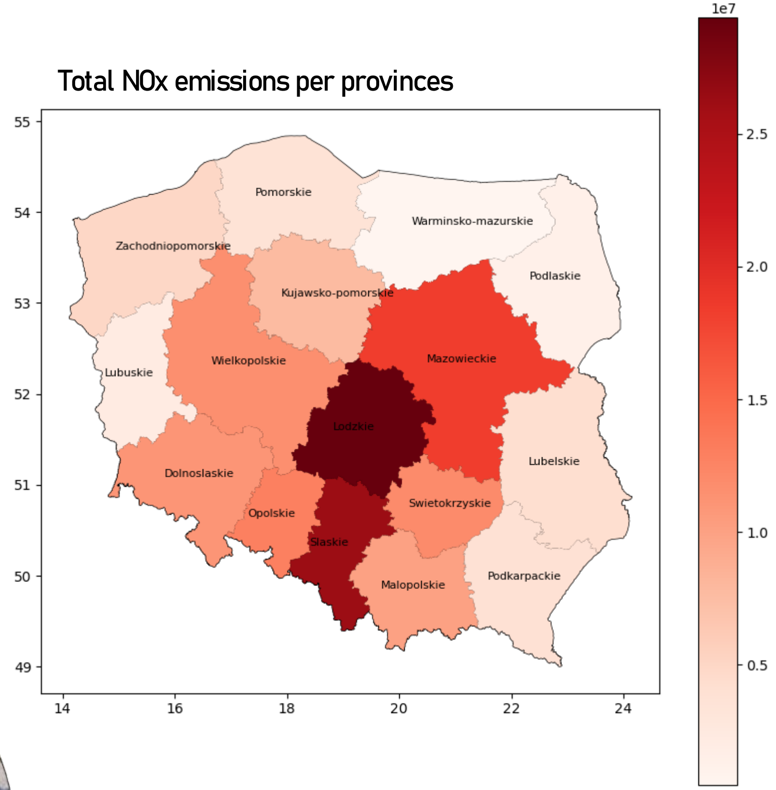




waste management



Mineral industry

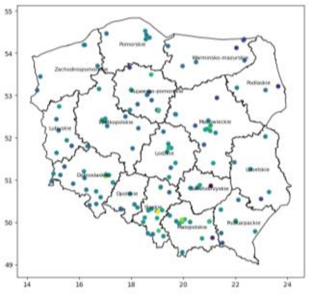


Data to information:

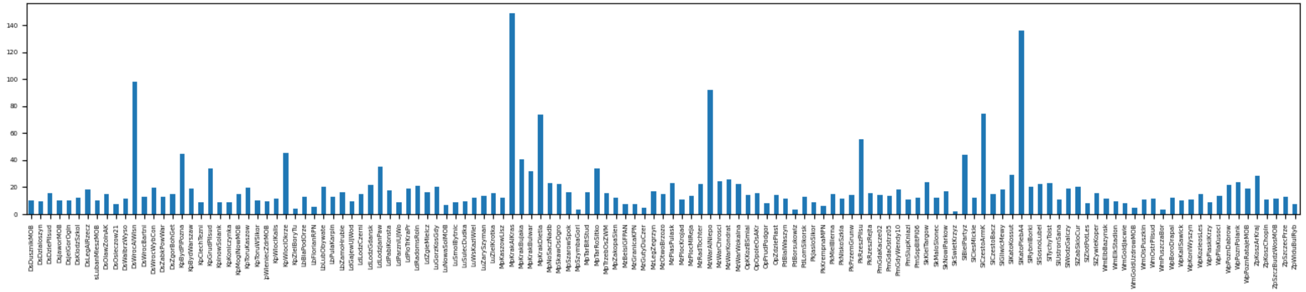
- Sum of emissions for each sector
- Sum of emissions from all sectors

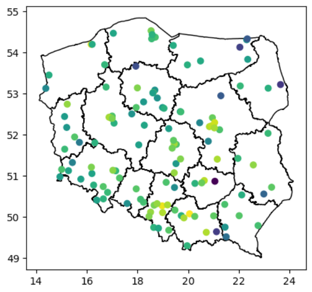
NOX AIR POLLUTION DATASET

INFORMATION FROM DATA POINTS TO PROVINCES

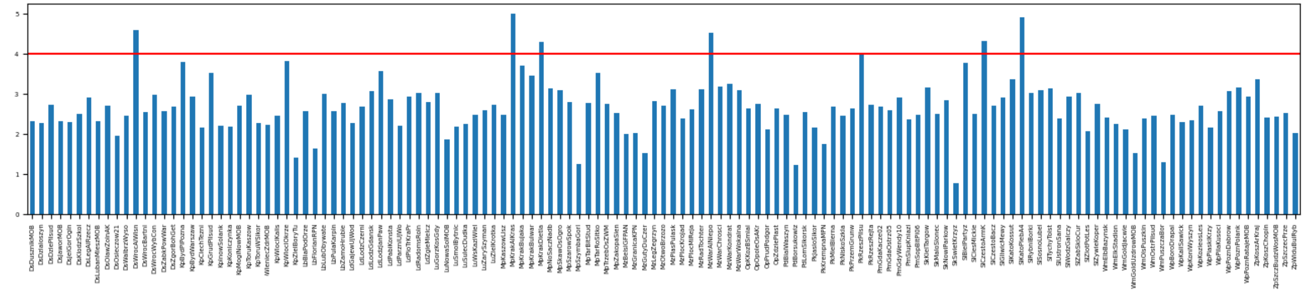


Median NOx in the air measured in the stations

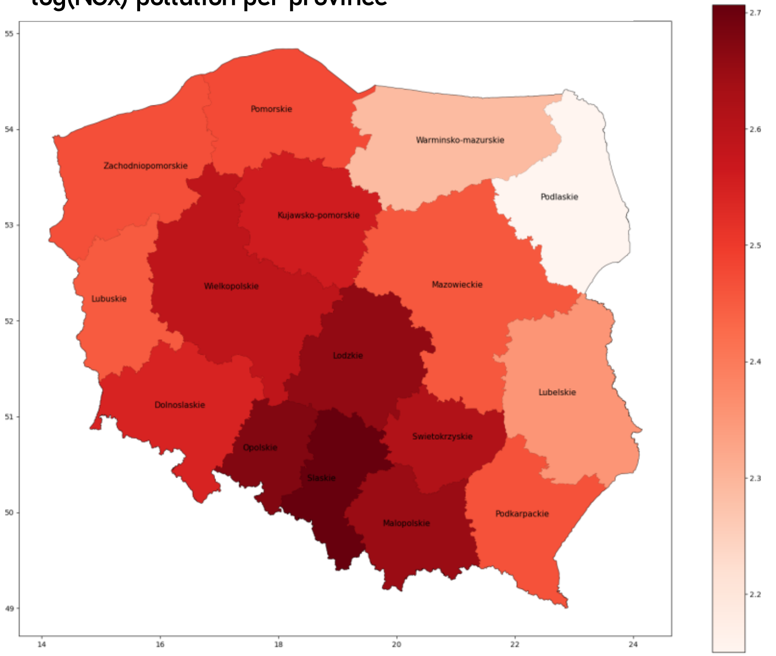




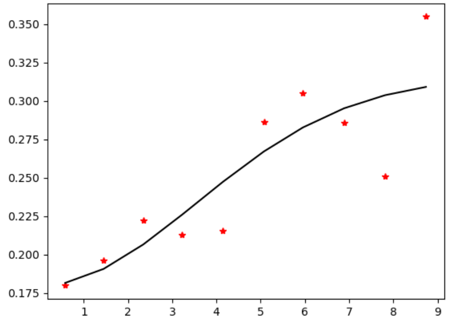
Median log(NOx) in the air measured in the stations



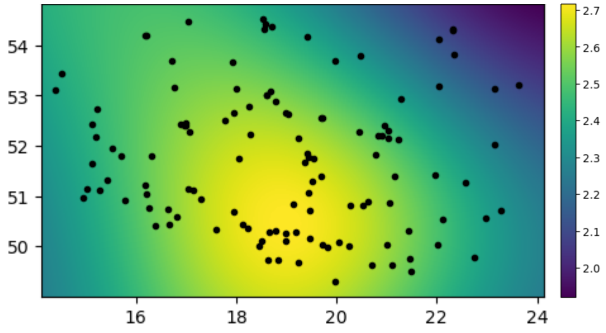
log(NOx) pollution per province



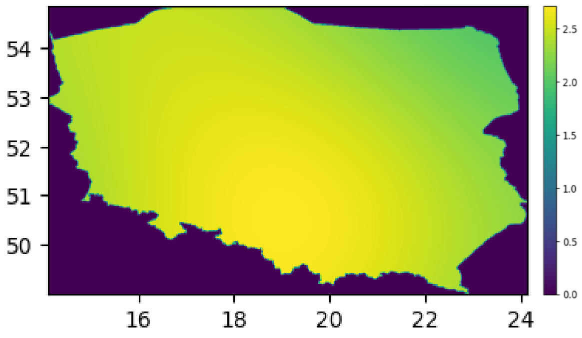
Variogram



log(NOx) interpolation (O.Kriging)



Clip to Poland border



ML MODEL TRAINING

MODEL USED: XGBOOST

1 Data training preparation

- data standardization
- Training 70% ; test 30%

2 Grid Search

```
from sklearn.model_selection import GridSearchCV
# set up our search grid
param_grid = {"max_depth":    [4, 5, 6, 7, 8, 9], # 7
              "n_estimators": [10, 20, 30, 40, 50, 60], # 22
              "learning_rate": [0.015, 0.02, 0.05, 0.1, 0.2, 0.3]} # 0.1

# try out every combination of the above values
search = GridSearchCV(regressor, param_grid, cv=5).fit(X_train, y_train)

print("The best hyperparameters are ",search.best_params_)
```

```
The best hyperparameters are  {'learning_rate': 0.02, 'max_depth': 4, 'n_estimators': 10}
```

3 Model training: XGBoost

Extreme Gradient Boosting (XGBoost) is an extension to gradient boosted decision trees (GBM) and specially designed to improve speed and performance (Tadakaluru, 2022).

4 Model evaluation

```
y_pred = regressor.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
print("MSE: %.2f" % mse)
print("RMSE: %.2f" % (mse**(1/2.0)))
```

```
MSE: 1.83
RMSE: 1.35
```

5 Feature importance:

Feature importance is a step in building a machine learning model that involves calculating the score for all input features in a model to establish the importance of each feature in the decision-making process. The higher the score for a feature, the larger effect it has on the model to predict a certain variable.

RESULTS

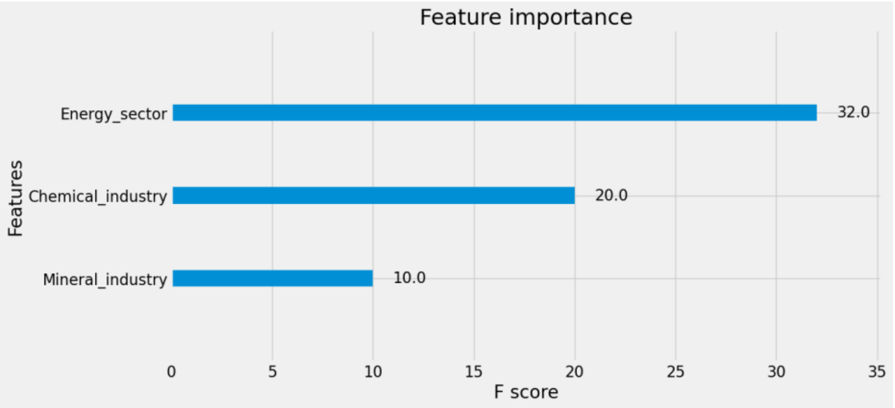
The highest NOx emissions by sector:

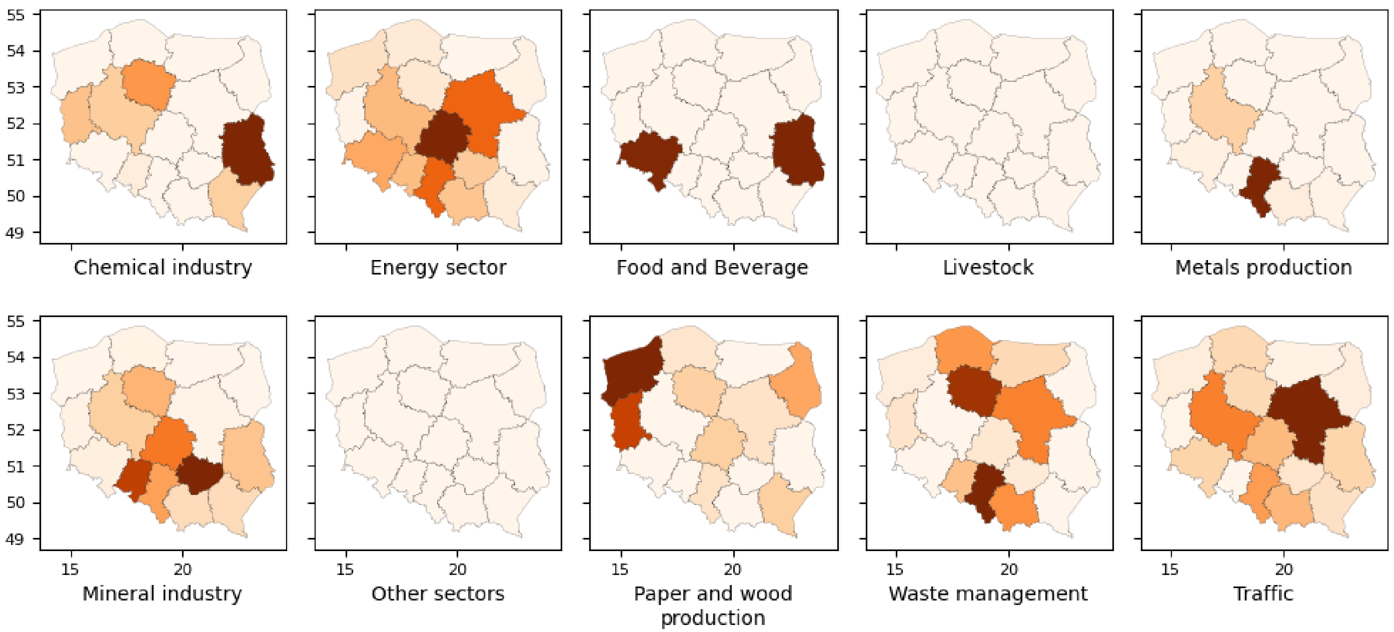
1. Energy sector in the total contribution of NOx emissions, especially in the provinces of Slaskie, Lodzkie, and Mazowieckie.
2. Traffic (urban buses, lorries, and road tractors) are present in all provinces, with the highest level in the province of Mazowieckie, which contains the city of Warsaw, the capital and largest city of Poland.
3. Mining industry has a relatively extensive presence in the Polish territory.

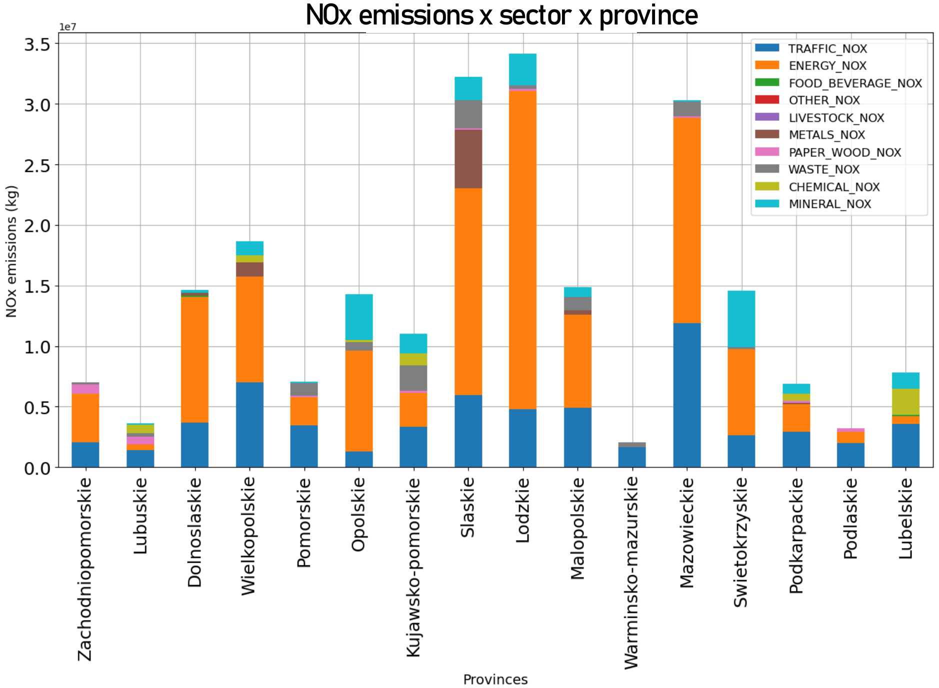
The features that contributes most to the final prediction of NOx air pollution:

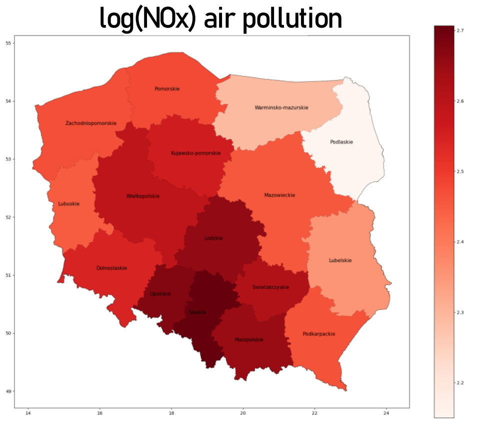
1. Energy sector
2. Chemical industry
3. Mineral industry

Livestock and aquaculture: 80% NH_3, 10% CH_4, 4% N_2O, and 4% PM10.

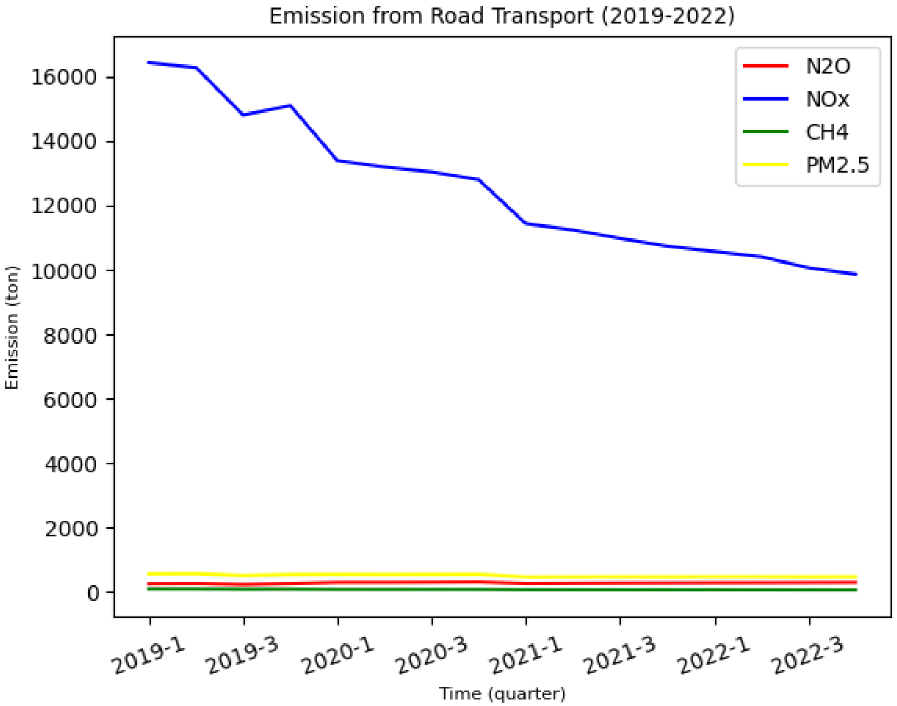


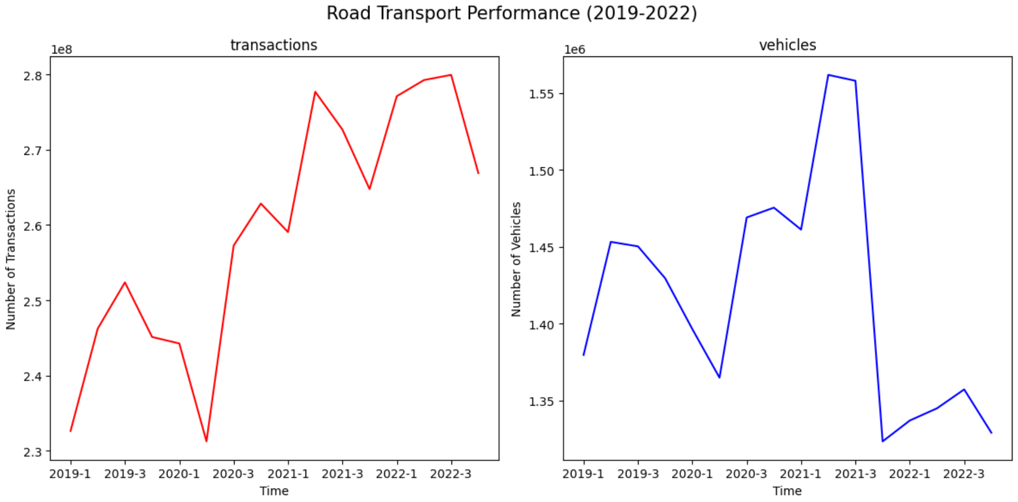


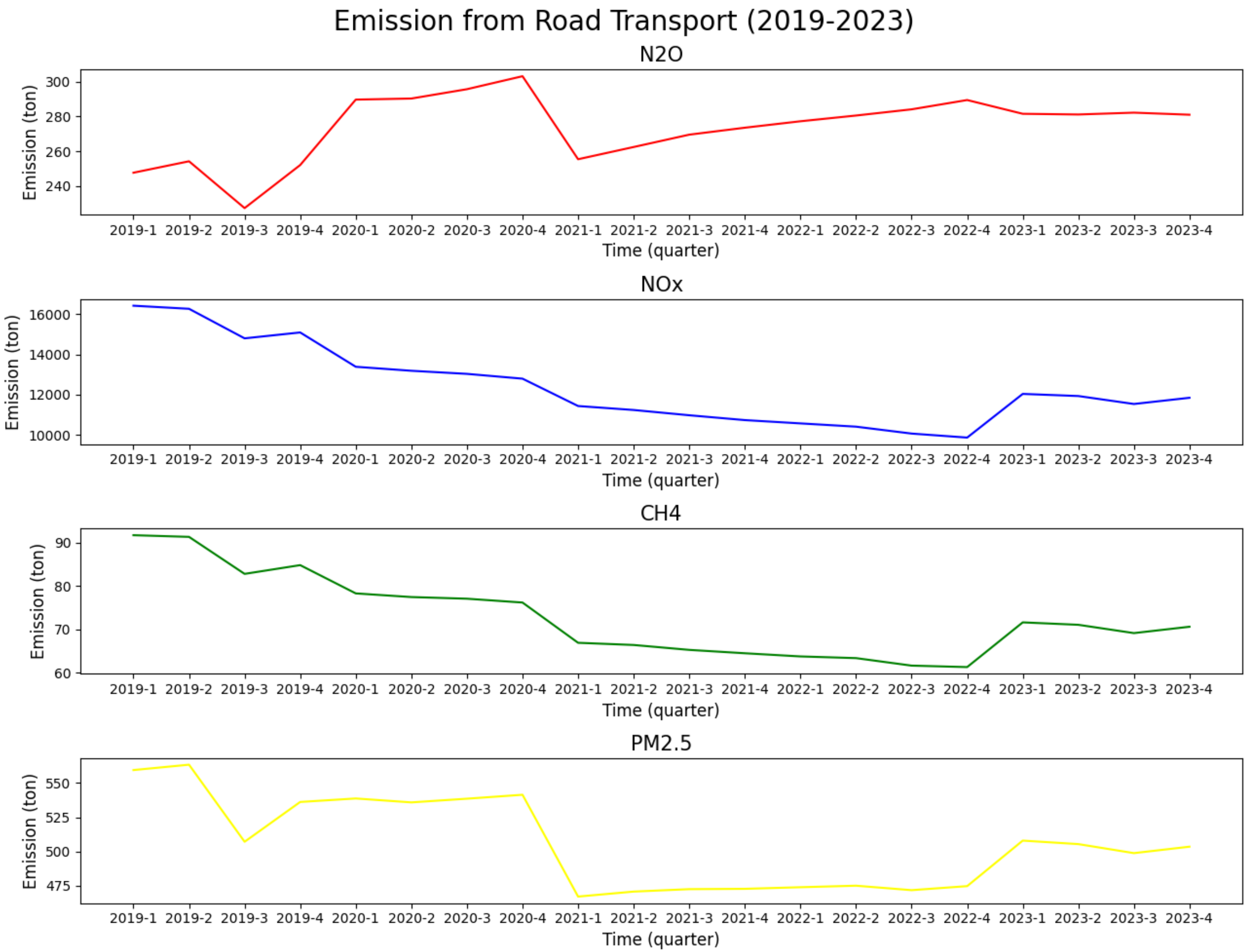




TREND & PREDICTION







MODELS COMPARING

Random Forest Regression

- *Mean Squared Error: 438904.8219667969*
- *R-squared: 0.4412701690718347*

Support Vector Regression (LinearSVR)

- *Mean Squared Error: 39477647.71463582*
- *R-squared: -139.28346233627212 (not a good fit for data?)*

Multiple Linear Regression

- *Mean Squared Error: 188641.41005873526*
- *R-squared: 0.7248884565462297*

Discussion on the negative R-squared

1. There could be over-fitting in our model. It can be caused by various reasons like small dataset and noise in the dataset. Our traffic emission dataset is indeed small(16 rows, 6 columns), so this can be the main reason.

2. R-squared is for least squares regression and not usually for SVR. R-squared is not commonly used to evaluate the SVR model. Metrics such as Mean Squared Error (MSE) or Mean Absolute Error (MAE) are more typical for assessing the accuracy of predictions in SVR.

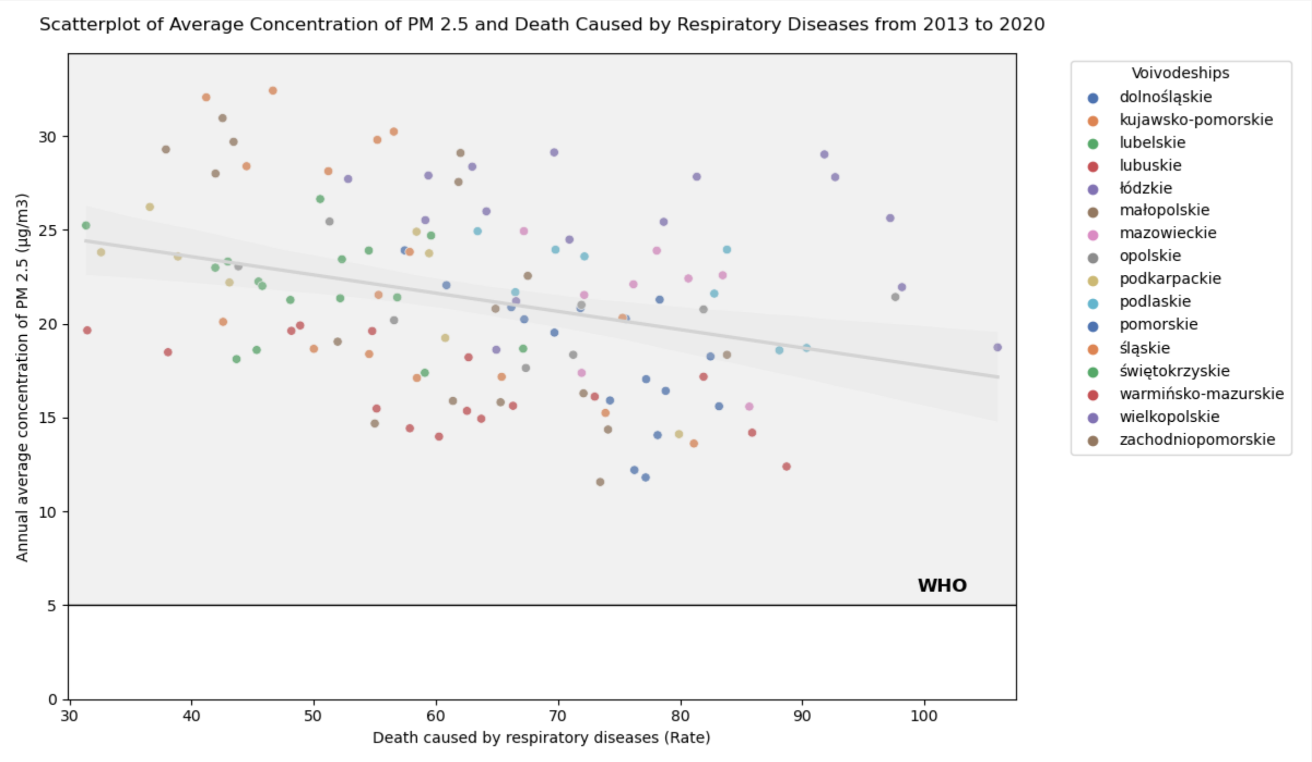
INVESTIGATING CORRELATION

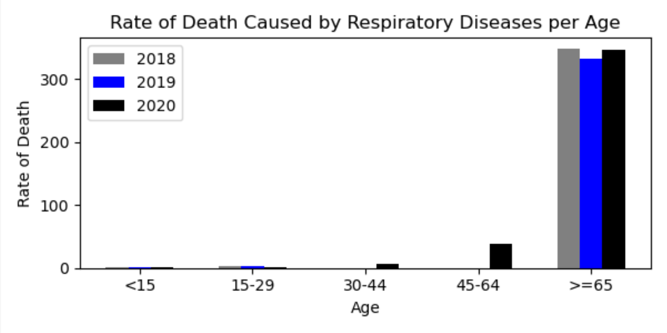
Important things from the result:

- The correlation between annual average concentration of PM 2.5 and death caused by respiratory diseases is weak negative correlation.
- All voivodeships have been exposed to dangerous concentration levels of PM 2.5 at all times between 2013 and 2020.
- The highest rates of death caused by respiratory diseases in Poland from 2018-2020 is happened with people aged 65 y.o or more.

Suggestion:

- There are only two variables considered in the model. In the future, the model could also include another confounding variables such as temperature, and relative humidity (Oo et al., 2020; Qiu et al., 2018; Zhu et al., 2019b).
- Explore another approach and longer time horizon to better understand the correlation between these variables.
- previous studies on the association between LC mortality and exposure to ambient air pollution have been limited and results have been inconsistent.





Voivodeship	Correlation
dolnośląskie	-0,77007346
kujawsko-pomorskie	-0,617880397
lubelskie	0,178605767
lubuskie	-0,541824201
łódzkie	-0,702776121
małopolskie	-0,726877269
mazowieckie	-0,367647221
opolskie	-0,372317468
podkarpackie	-0,766783058
podlaskie	-0,732648464
pomorskie	-0,413624106
śląskie	-0,587283826
świętokrzyskie	-0,538815131
warmińsko-mazurskie	-0,386860839
wielkopolskie	-0,520023508
zachodniopomorskie	-0,159537218

Conclusion

- Objective 1: the energy sector, traffic, and mining industry are the primary sources of NOx emissions. The most significant factors influencing NOx air pollution predictions are emissions from the energy sector, chemical industry, and mineral industry.

a short reflection on what your have learned about data-science

Ethical and privacy aspects

- **Data collection with ethical consideration:**
  - The ethical considerations for data sources, such as administrative boundaries and industry-related data from the European Environment Agency (EEA), are explicitly outlined, emphasizing free and open access policies while maintaining data quality standards.

- **Transparent data ethic:** The EEA's commitment to providing data as a public good with open access policies and quality metadata underscores a transparent and ethical data framework
- **Adherence to Legal and Privacy Standards:**
  - The project adheres to Polish law, setting conditions for content usage and re-use, ensuring privacy and ethical standards are maintained.
  - Specific guidelines for information reuse from *dziennik.gios.gov.pl* underscore a commitment to protecting user privacy and ensuring responsible data disclosure.
- **Digital Accessibility:** digital accessibility, ensuring that the website https://transtat.stat.gov.pl for extracted the traffic data is accessible to all users.