




Investigating Causes and Impacts of Air Pollution in Poland

By Group 2




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



CONTEXT AND GOALS


 Notes from Poland







NEWSINSIGHTSPOLSKUTOPICSPODCASTSABOUT USCONTACTNEWSLETTERS



Poland has EU's worst air pollution, shows new report

NOV 25, 2020 | SOCIETY





Poland has the European Union's most polluted air, according to a new report from the European Environment Agency (EEA). The country also recorded among the lowest reductions in air pollution during the pandemic lockdown.

LATEST NEWS

LAW, MEDIA, NEWS

Notary involved in Polish's government public media takeover charged by prosecutors

The development marks another potential blow to the government.

NEWS, SOCIETY

Most dogs temporarily adopted during Polish city's winter freeze find permanent homes

Only 20 of the 123 dogs temporarily adopted three weeks ago have been returned, despite temperatures returning to normal levels.

LAW, NEWS, SOCIETY

Polish government approves bill restoring prescription-free access to morning-after pill

The former conservative government made emergency contraceptives available by prescription only in 2017.

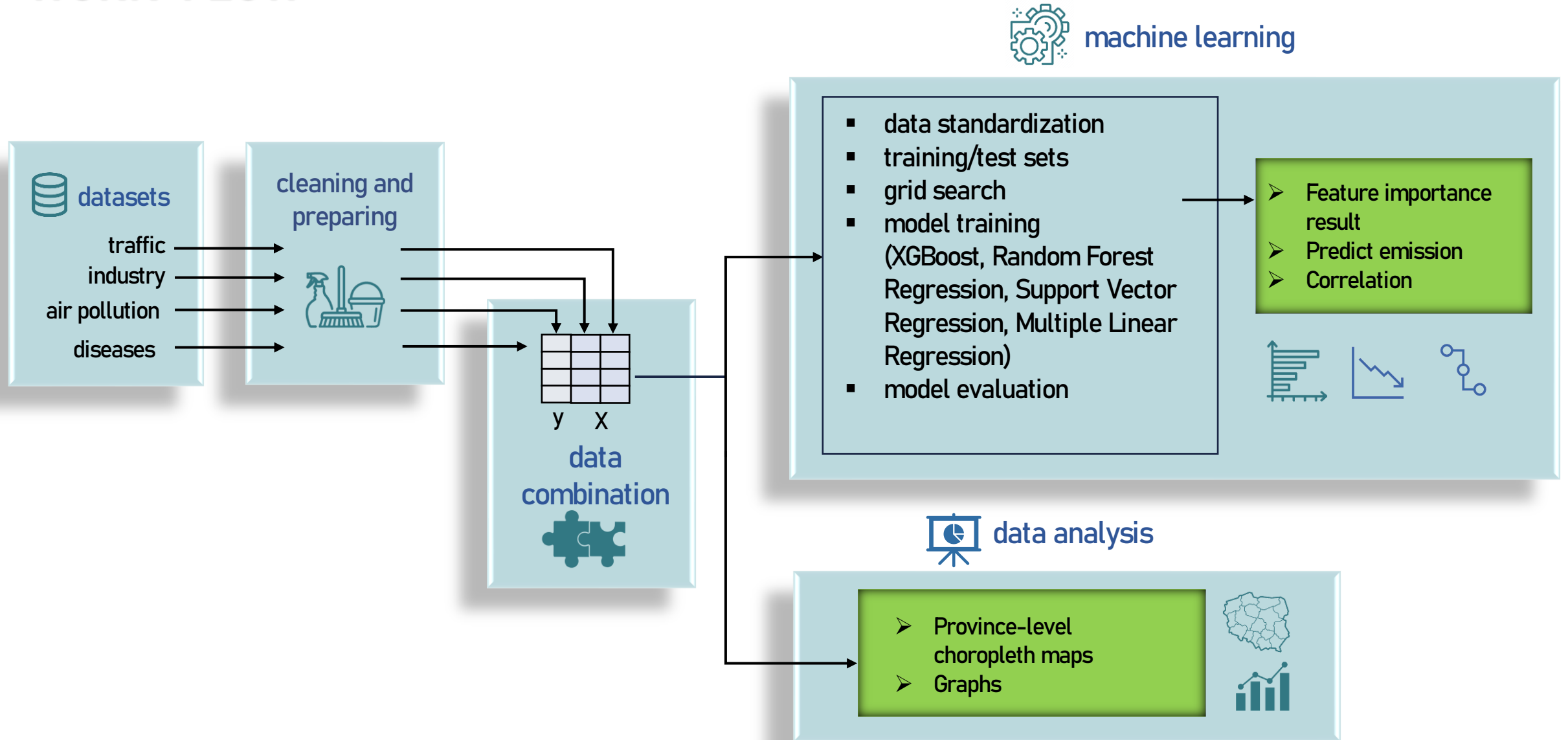
MORE NEWS

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 & Niedozytko, 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



DATA AVAILABILITY

FINDING DATA WITH THE SAME RANGE

Type of pollutant available:

type information	type variable	pollutants available														
Air pollution	Target	BC	CO	NO	NO2	-	O3	-	PM10	PM2.5	SO2	-	-	-	-	-
Air pollution	Target	-	CO	NO	NO2	-	O3	-	PM10	PM2.5	SO2	-	NOx	-	-	-
Traffic	Explanatory	-	-	-	-	N2O	-	-	-	PM2.5	-	NMVOC	NOx	CH4	CO2	-
Seaports	Explanatory	-	-	-	-	-	-	PM	-	-	-	-	NOx	-	CO2	SOx
Industries	Explanatory	-	-	-	-	N2O	-	-	PM10	-	-	NMVOC	NOx	CH4	CO2	SOx

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: **µg/m³ (median per year)**

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

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

Data Sources:



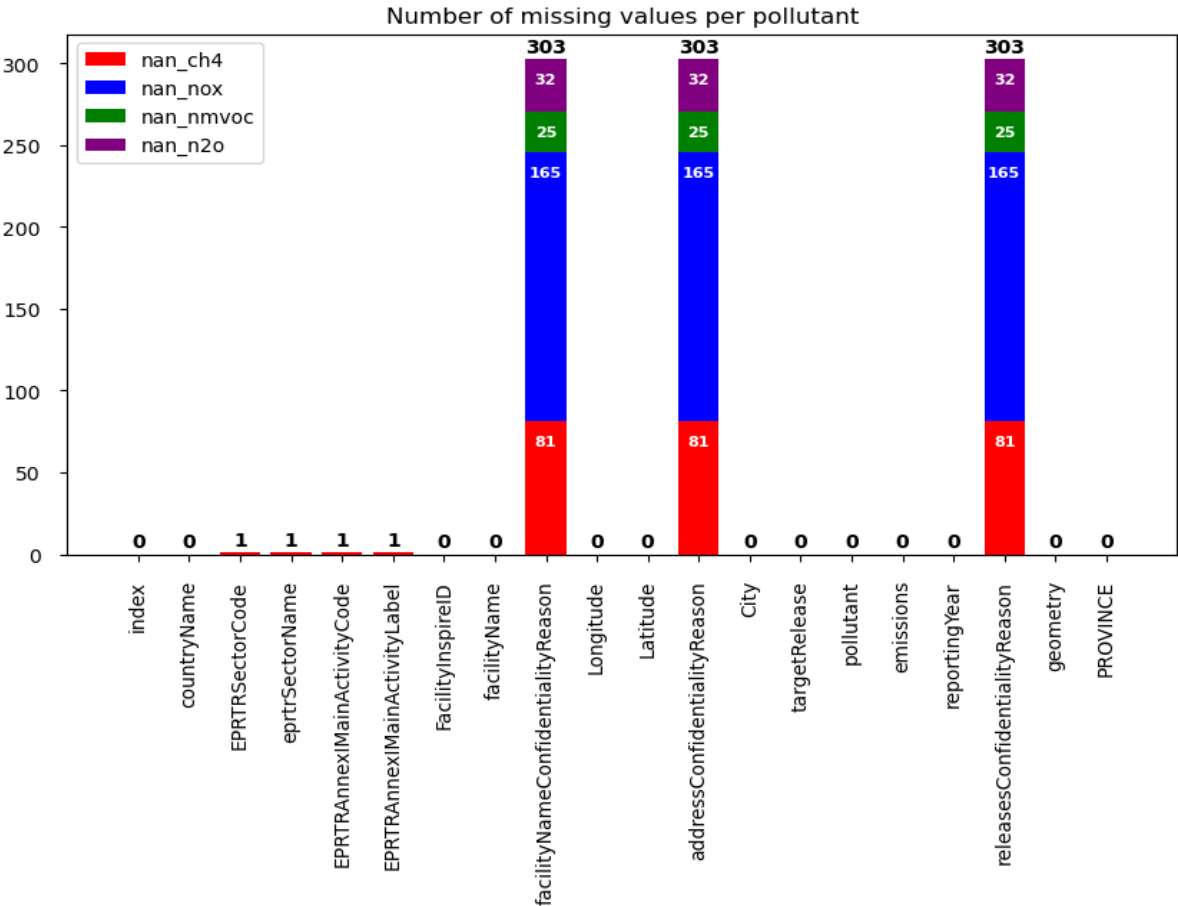
DATA CLEANING

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



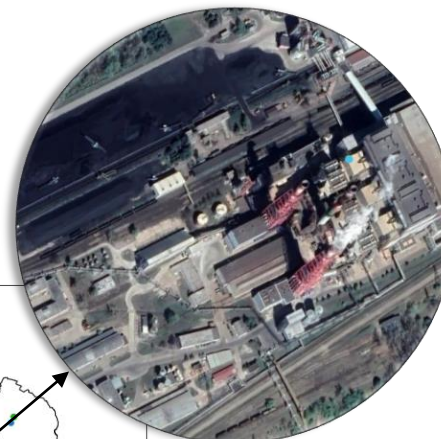
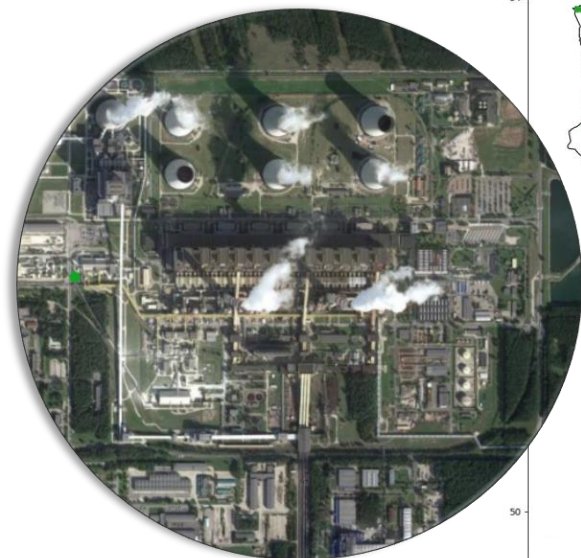
Traffic & Respiratory Diseases

- Small dataset
- No need of cleaning

INDUSTRIES DATASET

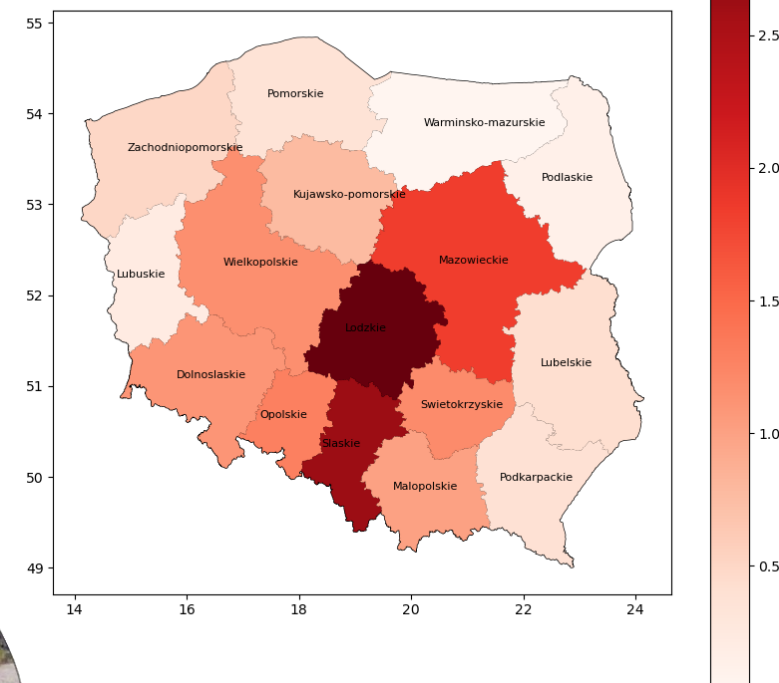
INFORMATION: "INDUSTRIAL SECTOR" AND "EMISSIONS"

Energy sector

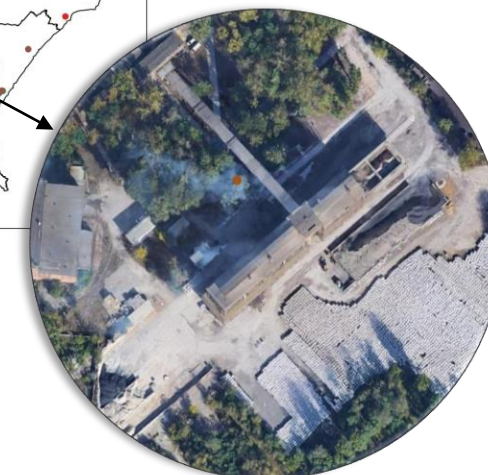


waste
management

Total NOx emissions per provinces



Mineral
industry



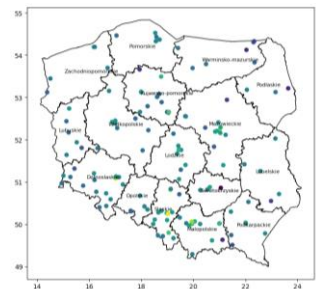
- Animal and vegetable products from the food and beverage sector
- Chemical industry
- Energy sector
- Intensive livestock production and aquaculture
- Mineral industry
- Other activities
- Paper and wood production and processing
- Production and processing of metals
- Waste and wastewater management

Data to information:

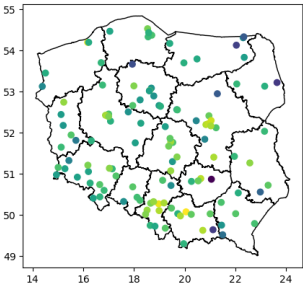
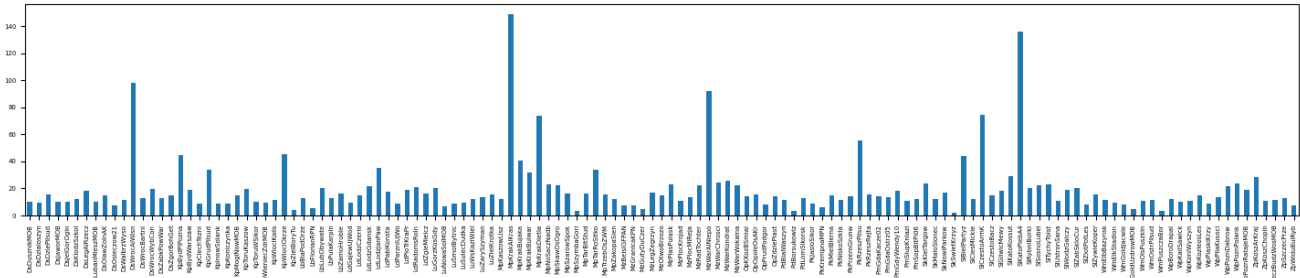
- Sum of emissions for each sector
- Sum of emissions from all sectors

NOX AIR POLLUTION DATASET

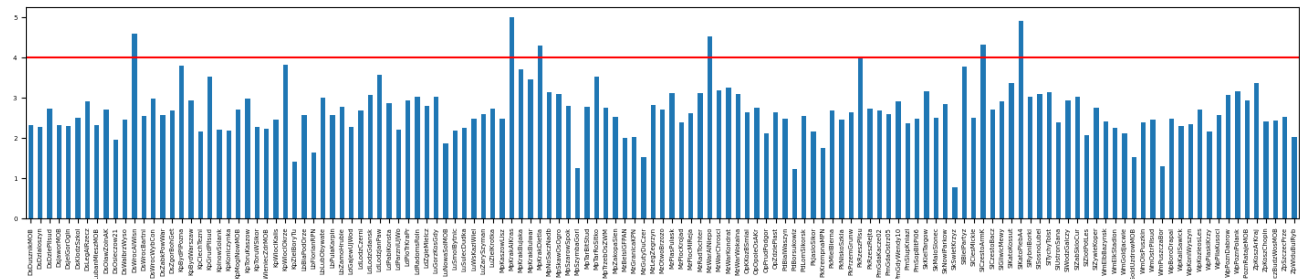
INFORMATION FROM DATA POINTS TO PROVINCES



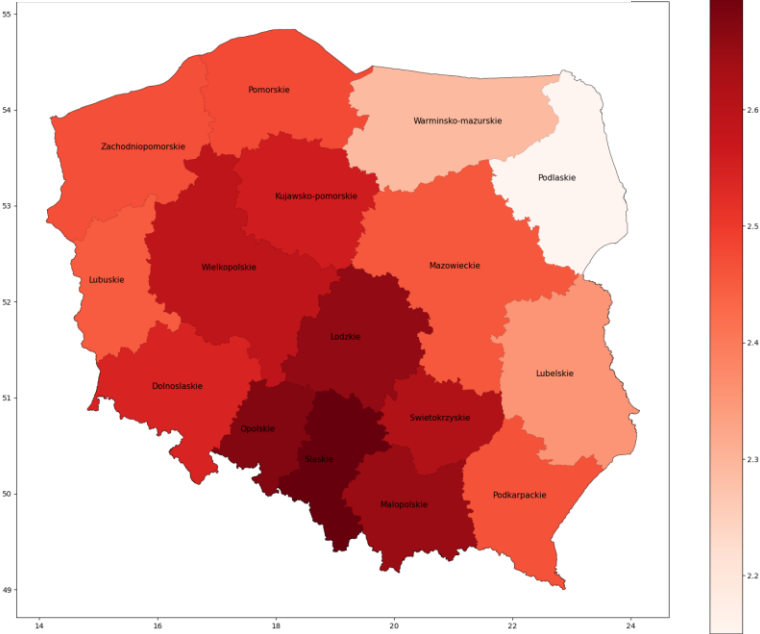
Median NOx in the air measured in the stations



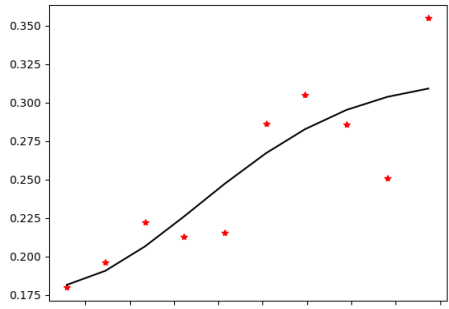
Median log(NOx) in the air measured in the stations



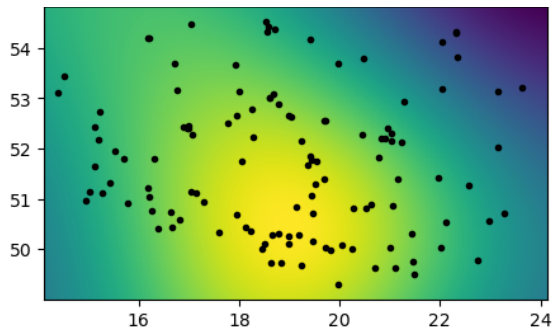
log(NOx) pollution per province



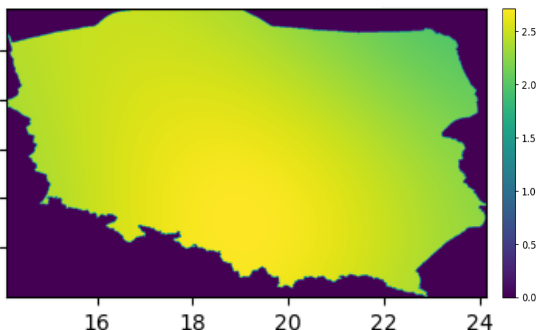
Variogram



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

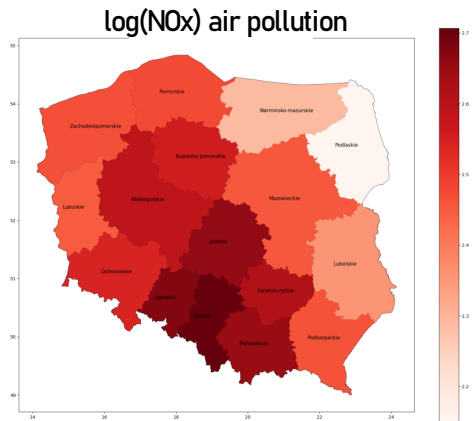
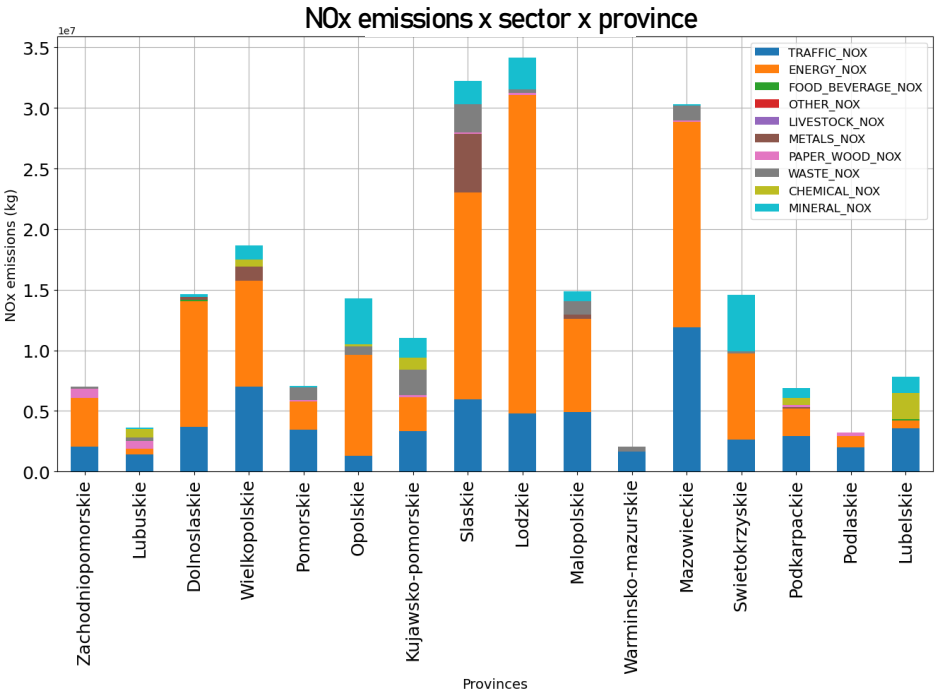
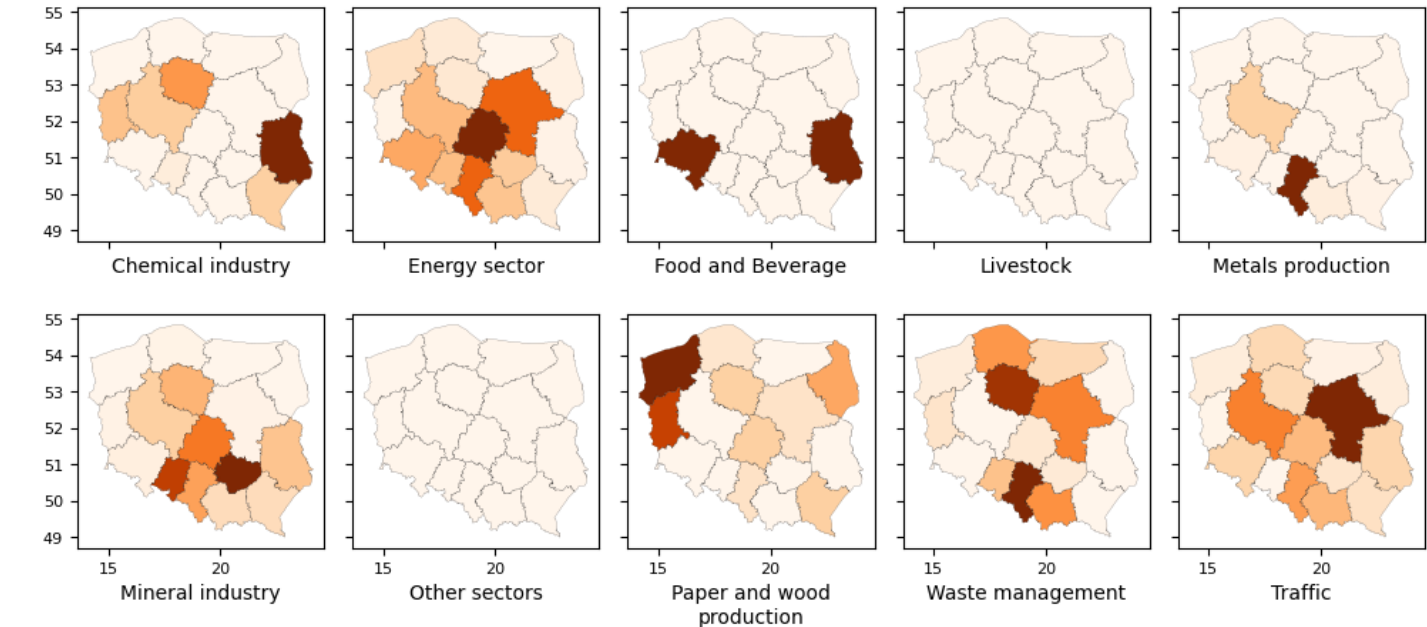
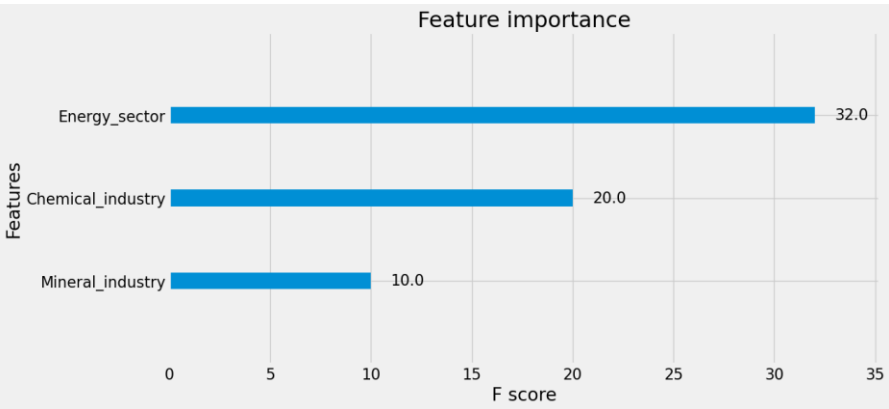
The highest NOx emissions by sector:

- 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.

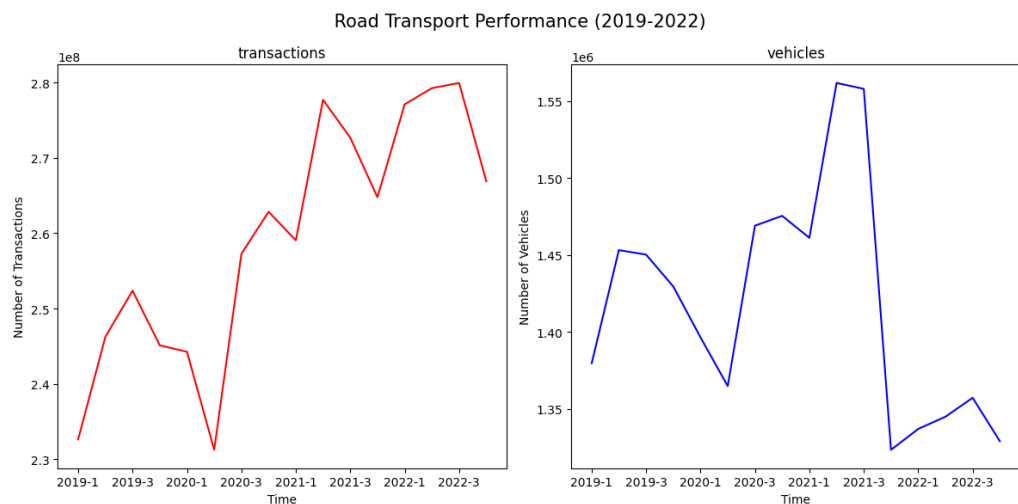
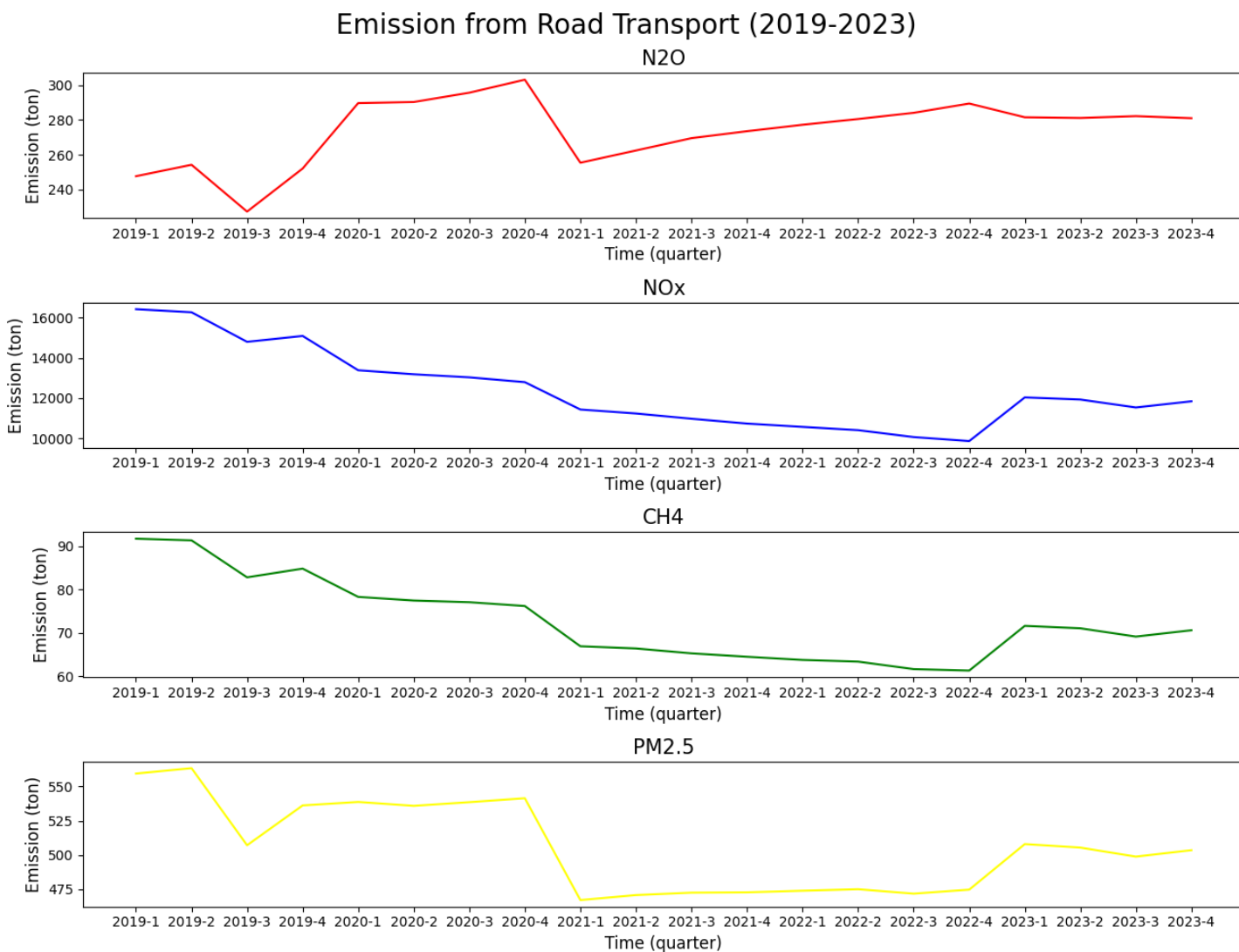
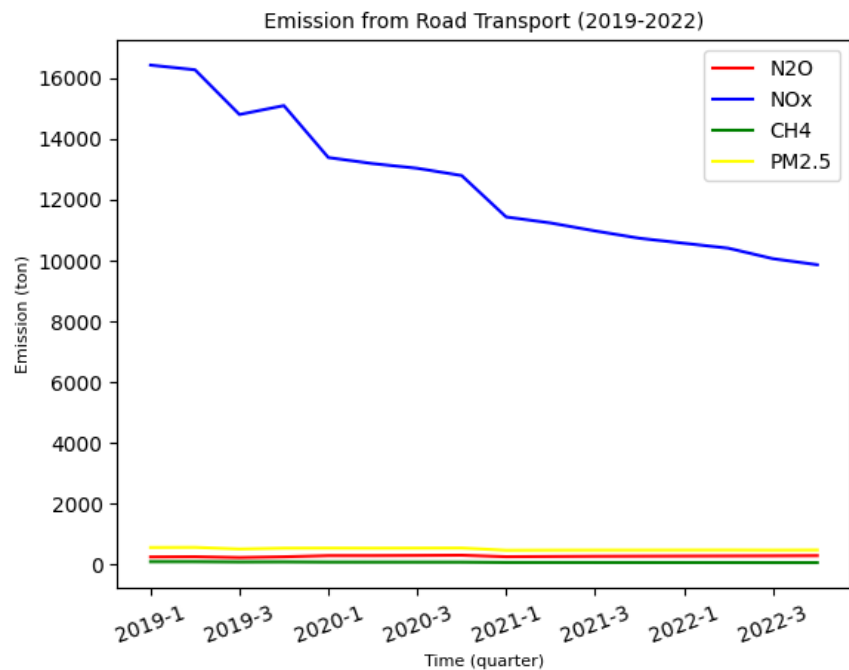
The features that contributes most to the final prediction of NOx air pollution:

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

Livestock and aquaculture: 80% NH3, 10% CH4, 4% N2O, 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

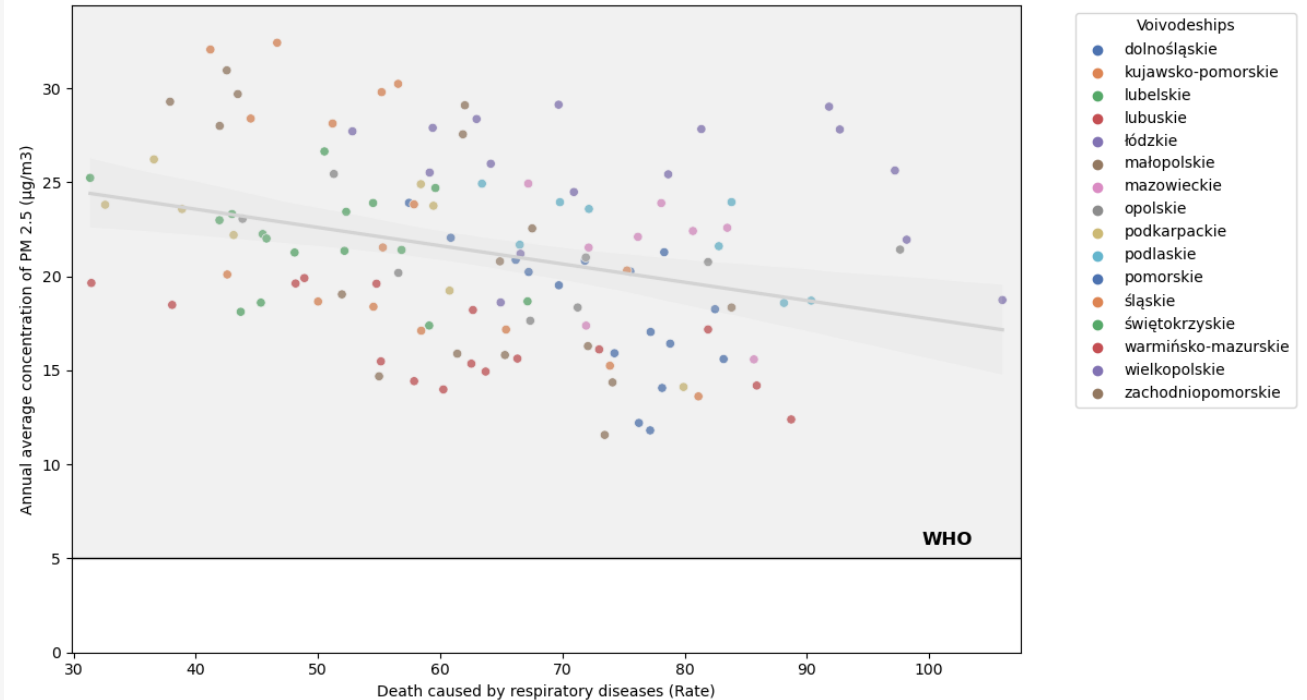
Important things from the result:

- 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**.

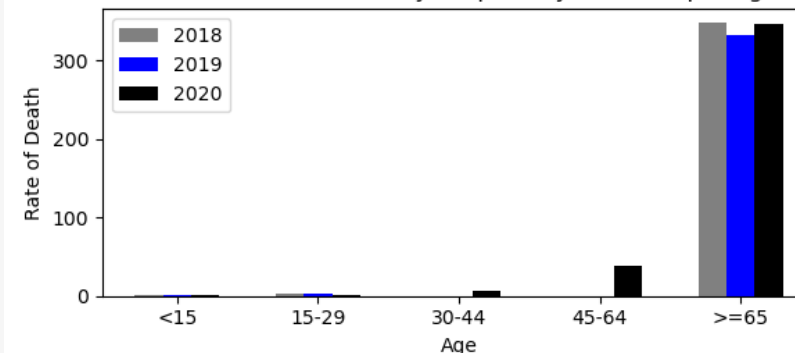
Suggestion:

- 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 as the time lag between the exposure of air pollution and mortality is long.
- Previous studies on the association between LC mortality and exposure to ambient air pollution have been limited and results have been inconsistent (Chung et al., 2021). Further research could **analyze the suggestion** from those studies to develop better model.

Scatterplot of Average Concentration of PM 2.5 and Death Caused by Respiratory Diseases from 2013 to 2020



Rate of Death Caused by Respiratory Diseases per Age

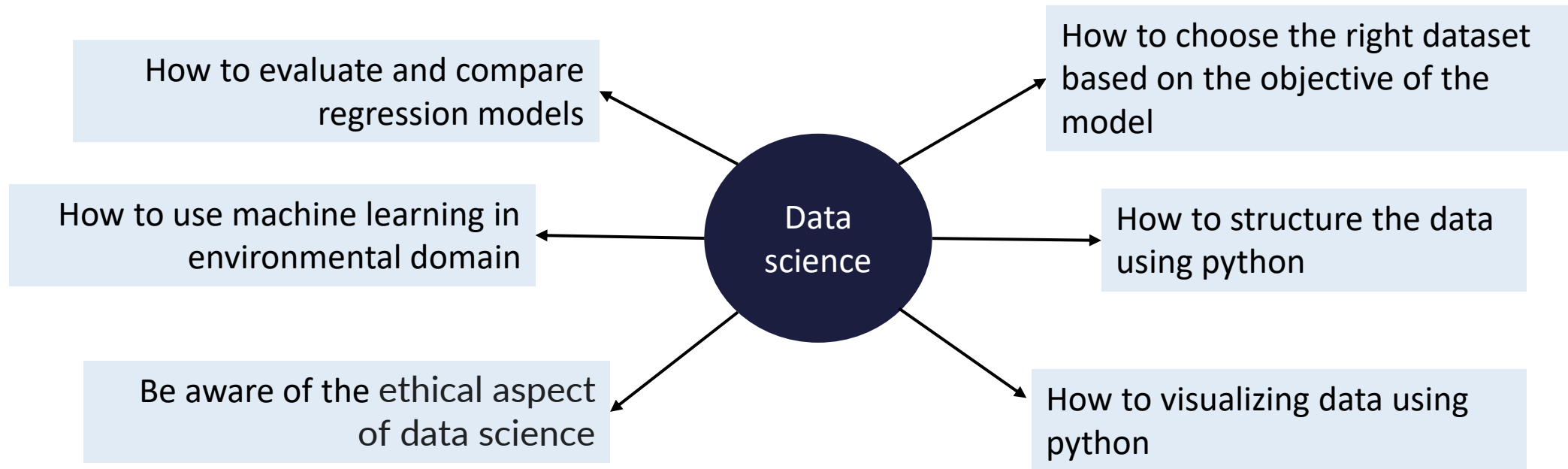


Voivodeship	Correlation
dolnośląskie	-0,77007346
kujawsko-pomorskie	-0,617880397
lubelskie	0,178605767
lubuskie	-0,541824201
łódzkie	-0,702776121
małopolskie	-0,726877269
mazowieckie	-0,367647221
opolskie	-0,372317468
podkarpackie	-0,766783058
podlaskie	-0,732648464
pomorskie	-0,413624106
śląskie	-0,587283826
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 NO_x emissions. The **most significant** factors influencing NO_x air pollution predictions are emissions from the **energy sector, chemical industry, and mineral industry**.
- **Objective 2:** The **positive relationship** between traffic performance and emissions can be fitted well using multiple linear regression. The result shows that emissions of NO_x, CH₄ and PM_{2.5} from transport have been decreasing in recent years, while emissions of N₂O have been increasing.
- **Objective 3:** The correlation between annual average concentration of PM 2.5 and death caused by respiratory diseases is **weak negative correlation**. The death caused by respiratory diseases might not be a right variable for modelling the relation of PM 2.5 exposure and respiratory diseases as the current time horizon is 7 years which is more fit to analyze the effect of the short-term exposure of PM 2.5 to chronic respiratory diseases.

REFLECTION ON WHAT WE HAVE LEARNED ABOUT DATA-SCIENCE





Ethical and privacy aspects

Data collection with ethical consideration:

- Emphasizing free and open access policies while maintaining data quality standards.
- To reuse the data information obliged to clearly indicate the source and creation time of the information, specifying whether it is used in whole or in fragments and users must provide relevant information about the changes.

Privacy:

- Privacy considerations are addressed by allowing users to disable location information in device settings, demonstrating a proactive approach to safeguarding user privacy.
- Specific guidelines and information for reusing data underscore a commitment to protecting user privacy and ensuring responsible data disclosure.

Transparent data ethic: most of them have transparent data ethics in their declaration such as 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 data air pollution in Poland from the Poland government platform are project adheres to Polish law, setting conditions for content usage and re-use, ensuring privacy and ethical standards are maintained.

Digital Accessibility:

- Websites used provide open access to the public, along with features for users with disabilities in both web and mobile versions, highlights a commitment to inclusivity beyond the specific website.
- The provision for individuals to request digital accessibility reflects a proactive approach to societal concerns, ensuring inclusivity in accessing project benefits.