

**Air Quality in Europe**

W.CIP02: Data Collection, Integration and Preprocessing

TEAM 16

1. Hazimé-Zayour Maha [maha.hazime-zayour@stud.hslu.ch](mailto:maha.hazime-zayour@stud.hslu.ch)
2. Batschelet Jimena [jimena.batschelet@stud.hslu.ch](mailto:jimena.batschelet@stud.hslu.ch)
3. Devdas Samuel [samuel.devdas@stud.hslu.ch](mailto:samuel.devdas@stud.hslu.ch)

Thursday, December 13, 2023

# Air Quality in Europe

In this project, we delve into a multifaceted analysis of air quality across European countries. The primary focus is on understanding how various factors — including air pollution levels, renewable energy consumption, economic growth (GDP), weather conditions, and geographical variables — interact and influence each other. The overarching goal is to provide insights that could inform environmental policy decisions and contribute to the global dialogue on sustainable development and public health.

[Air Quality in Europe 1](#_Toc153473526)

[Introduction 2](#_Toc153473527)

[Motivation 2](#_Toc153473528)

[Project Goals 3](#_Toc153473529)

[Research Questions 3](#_Toc153473530)

[What is the impact on Air Quality if we think about GDP and how many people live in an area? 3](#_Toc153473531)

[How can we predict air pollution levels in Europe? And which of these factors are most influential in affecting air pollution? 3](#_Toc153473532)

[What is the impact of geo location/elevation on average AQI? 3](#_Toc153473533)

[Context Diagram 4](#_Toc153473534)

[Data Sources, Model and Task Allocation 4](#_Toc153473535)

[Tools and Technologies Used 5](#_Toc153473536)

[ETL Process 6](#_Toc153473537)

[Data Loading Process 8](#_Toc153473538)

[Answering the Questions 9](#_Toc153473539)

[What is the impact on Air Quality if we think about GDP and how many people live in an area? 9](#_Toc153473540)

[How can we predict air pollution levels in Europe? And which of these factors are most influential in affecting air pollution? 10](#_Toc153473541)

[What is the impact of geo location/elevation on average AQI? 13](#_Toc153473542)

[Lesson learned and self-reflection: 14](#_Toc153473543)

## Introduction

This document outlines the ETL (Extract, Transform, Load) process for a data-driven investigation into the impact of energy consumption patterns, economic factors, and geographic traits on European countries' air quality and public health.

## Motivation

With Europe's commitment to the Paris Agreement and its varied energy consumption behaviors, there is a valuable opportunity to analyze the link between types of energy consumed and their subsequent environmental and health effects. The study will assess the influence of energy sources and economic conditions on air quality, contributing to the development of sustainable policies.

## Project Goals

* Provide insights to aid the formulation of sustainable energy policies in Europe.
* Enhance the conversation on environmental preservation by examining the correlation between different factors and its ecological consequences.

Our project is centered on improving community health and protecting our collective environment. It aims to delineate the correlation between the energy portfolio of European nations and air quality. The findings will help drive improvements in public health, shape environmental policy, strengthen economic stability, and support thoughtful urban planning. We seek to help Europe shift toward a green economy in line with global climate obligations.

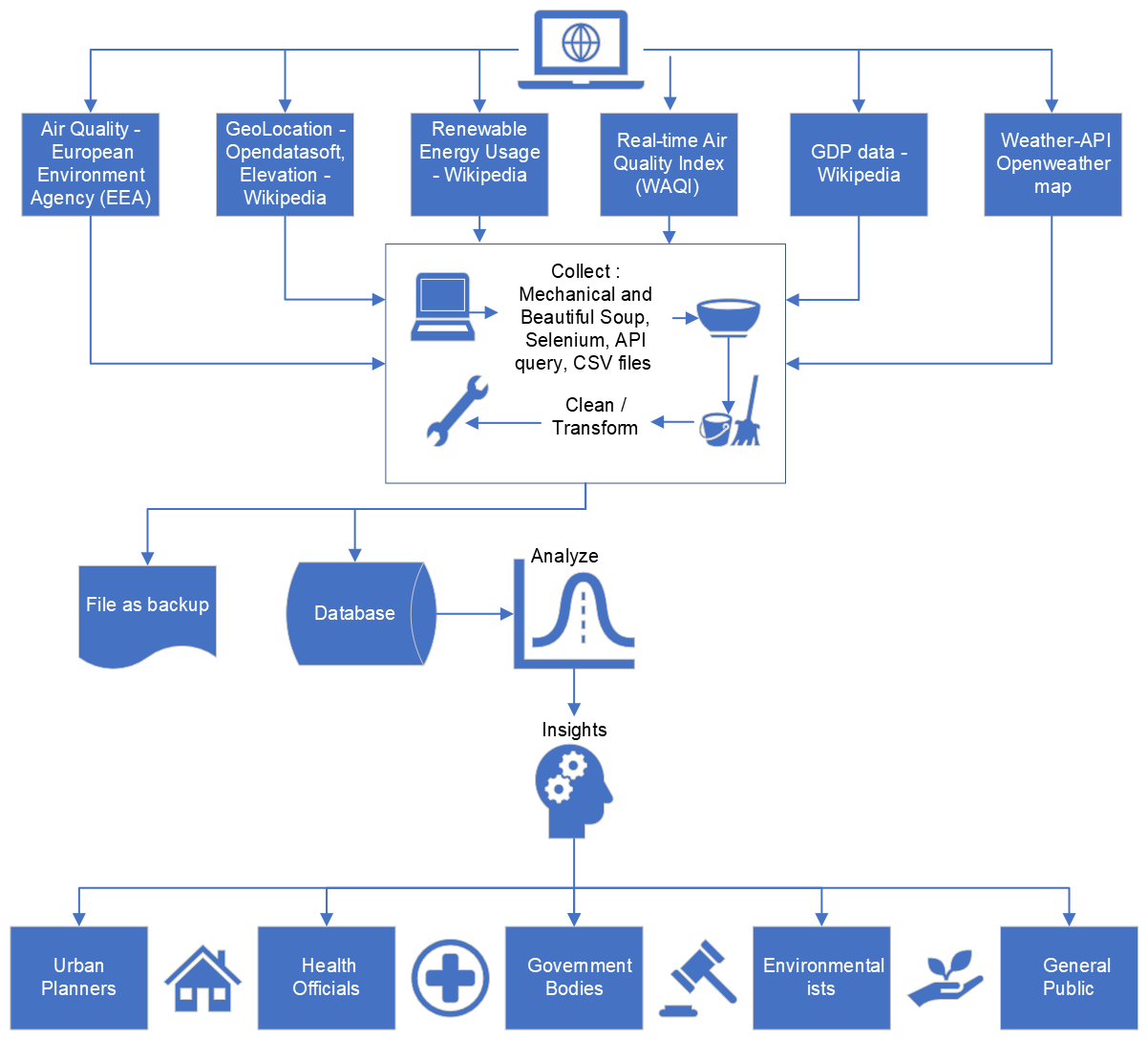
## Research Questions

## What is the impact on Air Quality if we think about GDP and how many people live in an area?

## How can we predict air pollution levels in Europe? And which of these factors are most influential in affecting air pollution?

## What is the impact of geo location/elevation on average AQI?

## Context Diagram



## Data Sources, Model and Task Allocation

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Source | Method | Student |
| COUNTRY (Data Merging Attribute) | [European Environment Agency (EEA). "Air Quality Health Risk Assessments." Retrieved from the EEA Discomap, Air Quality Viewer](https://discomap.eea.europa.eu/App/AQViewer/index.html?fqn=Airquality_Dissem.hra.countries_sel&EUCountries=Yes&ScenarioDescription=WHO_2021_AQG_Scen_Base&AirPollutant=PM2.5&UrbanisationDegree=All%20Areas%20(incl.unclassified)&Year=2020). | Scrape | **Student A:**  **Hazimé-Zayour Maha** |
| Year |
| Total Population |
| Populated Area [km2] |
| Premature Deaths |
| Years Of Life Lost |
| NO2 |
| O3 |
| PM2.5 |
| Renewable Energy Usage | [Wikipedia. "Erneuerbare Energien in Europa"](https://de.wikipedia.org/wiki/Erneuerbare_Energien_in_Europa) | Scrape |
| Temperature and  Wind Speed | [OpenWeatherMap API. Query with a valid API key.](http://api.openweathermap.org/data/2.5/weather) | Direct call to API | **Student B:  Batschelet Jimena** |
| GDP per Country | [Wikipedia. "List of sovereign states in Europe by GDP (nominal) per capita."](https://en.wikipedia.org/wiki/List_of_sovereign_states_in_Europe_by_GDP_(nominal)_per_capita) | Scrape |
| Geographical Data: Coordinates, Altitudes | [Opendatasoft. "Geonames - All Cities with a population > 1000." Retrieved from the public.opendatasoft.com database.](https://public.opendatasoft.com/explore/dataset/geonames-all-cities-with-a-population-1000/table/?disjunctive.cou_name_en&sort=name&refine.timezone=Europe)  [Opendatasoft. "Natural Earth - Countries scale 1:10m" Retrieved from the public.opendatasoft.com database.](https://public.opendatasoft.com/explore/dataset/ne_10m_admin_0_countries/table/?location=3,43.51669,-62.75391&basemap=jawg.light) | Scrape, CSV | **Student CDevdas Samuel** |
| Elevation | [Wikipedia: "List of countries by average elevation"](https://en.m.wikipedia.org/wiki/List_of_countries_by_average_elevation) |

## Tools and Technologies Used

**Jupyter Notebooks**

Our primary tool for conducting data analysis was Jupyter Notebooks. This interactive computing environment allowed us to write and execute Python code in a modular and readable format. The use of Jupyter Notebooks facilitated exploratory data analysis, visualization, and statistical modeling in an integrated manner. Its ability to combine code, output, visualizations, and narrative text made it an ideal platform for both performing the analysis and documenting our process and findings.

**GitHub**

For version control and collaborative coding, we utilized GitHub. This platform was instrumental in managing our codebase, tracking changes, and facilitating collaboration among team members. GitHub's robust version control capabilities ensured that our project's development was seamless, even as multiple contributors worked on different aspects of the analysis simultaneously. Moreover, GitHub provided a centralized repository for our project, making it accessible and maintainable over time.

**OneDrive/ Switch Drive**

OneDrive served as our primary tool for storing and sharing non-code related project documents and data files. Its cloud-based storage capability allowed team members to access and share project-related materials conveniently and securely. The integration of OneDrive with our other tools streamlined the workflow, ensuring that all team members had access to the latest versions of all project documents and datasets.

**WhatsApp**

These platforms were critical in ensuring constant and efficient information sharing. WhatsApp facilitated immediate communication and quick sharing of insights

**Zoom**

Zoom was crucial for facilitating virtual meetings and discussions. It allowed our team members to connect and engage in real-time, fostering a collaborative environment. Its video conferencing capabilities made it possible to conduct detailed discussions, brainstorming sessions, and progress reviews, ensuring that all team members were aligned and informed about the project's developments.

**Office 365 Teams**

offered a comprehensive environment for seamless communication, efficient document sharing, and collaborative work, perfectly aligning with our other tools to boost the project's overall productivity and cohesion.

**MechanicalSoup, BeautifulSoup and Selenium**

To gather the necessary data for our analysis, we employed web scraping techniques using MechanicalSoup, BeautifulSoupand Selenium. MechanicalSoup was used for simple web scraping tasks, where data could be extracted from static web pages. Its straightforward interface allowed us to quickly and efficiently scrape data without the need for extensive programming. For more complex tasks, involving dynamic web pages and the need for browser automation, Selenium was our tool of choice. Selenium enabled us to programmatically navigate web pages, fill out forms, and interact with web elements that required a higher level of interaction than what MechanicalSoup could offer. The combination of MechanicalSoup and Selenium provided us with a powerful and flexible suite of tools for extracting the data we needed from various online sources.

## ETL Process

Our project's data pipeline was structured around a comprehensive ETL process, involving data extraction from various web sources, data cleansing and enrichment, and loading the processed data into a MariaDB database.

### Data Extraction Process

The data extraction phase was a critical foundation for our project, requiring precise coordination and a strategic approach. Each team member focused on gathering specific types of data from various online sources. We targeted air quality metrics, weather information, geolocation data, renewable energy usage, and Gross Domestic Product (GDP) statistics. The diversity of these data types demanded a multifaceted extraction strategy, employing a range of tools suitable for different web environments.

### Approach and Tools

To address the variety and complexity of our data sources, we utilized a combination of Python-based tools and command-line utilities for web scraping.

* **Selenium for Dynamic Websites**: Selenium was our primary tool for scraping dynamic website where the content changes based on user interactions or is dynamically loaded using JavaScript. Selenium, being a powerful tool for browser automation, allowed us to programmatically navigate these complex web pages, interact with web elements, and extract the required data. Its ability to mimic human browsing behavior made it possible to access and scrape data that would otherwise be challenging to obtain using traditional scraping methods.
* **MechanicalSoup and BeautifulSoup for Static Websites**: MechanicalSoup was used for straightforward web scraping tasks. For more complex static websites, we also employed BeautifulSoup. This Python library enabled us to parse HTML and XML documents, extracting data with more nuanced control over the parsing process. BeautifulSoup's flexibility in navigating and searching the document tree was essential for accurately extracting data from intricately structured web pages.
* **Pandas read.csv(url):** For direct file downloads we used Pandas read.csv(url) method for Student C’s CSV file.

### Data Transformation Process

Following the meticulousextraction of diverse datasets, the transformation phase was pivotal in refining and preparing our data for in-depth analysis. This stage involved a series of processes to cleanse, enrich, and structure the data, ensuring its relevance, accuracy, and usability for our project objectives.

* Data 'Dirtying' and Cleansing
* Introducing Controlled Impurities: In a unique approach to simulate real-world data scenarios, we initially introduced controlled 'impurities' or anomalies into our datasets. This step was crucial for testing the robustness of our cleansing methods and preparing us to handle actual imperfect data.
* Data Cleansing: After the 'dirtying' process, we embarked on a comprehensive data cleansing operation. This involved identifying and correcting errors and inconsistencies, such as missing values, data types, and text suffixes in numeric columns, removing spaces. We employed Python and Pandas for their robust data manipulation capabilities, enabling us to efficiently clean and standardize the data.

### Data Enrichment and Transformation

* Enrichment Techniques: Post-cleansing, the datasets underwent enrichment to enhance their analytical value. This involved deriving new variables, aggregating data, and performing necessary transformations. We calculated additional metrics from existing variables such as Country Codes, Mean GDM and Rank by GDP, AQI, integrated external data for deeper insights, and converted data into formats more suitable for analysis.
* Structuring for Analysis: The final step in the transformation process was structuring the data into a form optimized for analysis. This included organizing data into meaningful categories, normalizing numerical values, and aligning different datasets to ensure compatibility for combined analysis.

### Technical Approach

* Python and Pandas: Our primary tools for data transformation were Python and Pandas. Python's versatility and Pandas' powerful data manipulation features allowed us to perform complex transformations with ease. Functions like merge, and pivot\_table in Pandas were particularly useful in reshaping the data and extracting meaningful patterns.
* Custom Scripts and Functions: To handle specific transformation tasks, we developed custom scripts and functions. These tailored solutions addressed unique data challenges, ensuring that our transformation processes were as effective and precise as possible.

## Data Loading Process

### Database Connection and Setup

The loading phase of our project began with each team member establishing a connection to the MariaDB database using Python. This step was essential for enabling direct interaction with the database for data insertion and manipulation. We utilized the **mariadb** module for establishing these connections, ensuring secure and authenticated access with appropriate credentials and database details.

### Individual Table Creation and Schema Design

In this phase, each team member was responsible for creating tables within the database tailored to their specific dataset. These tables were designed with schemas that aligned with the format and types of the transformed data, ranging across different domains such as environmental metrics, geolocation data, economic figures, and more. The creation of these tables was a critical step, ensuring that each dataset had a suitable structure for storage in the MariaDB database.

### Data Loading by Individual Team Members

Each member then proceeded to load their transformed data into their respective tables in the MariaDB database. This process was facilitated by Python scripts, which were developed to efficiently transfer data from Pandas DataFrames to the database. These scripts included comprehensive error handling and data integrity checks to ensure the accuracy and reliability of the data loading process.

### Merging and Loading the Final Dataset

In addition to loading individual datasets, Student C undertook the task of merging all datasets into a final, cohesive dataset. This merged dataset was then loaded into the MariaDB database, representing the collective effort of the entire team. The final dataset was stored in a specifically designed table, encompassing all the diverse data elements, and providing a comprehensive view of the overall data.

# Answering the Questions

Our analysis begins with a thorough Exploratory Data Analysis (EDA) to deeply understand the data at hand, laying a strong foundation for further inquiry. While the insights from the EDA are crucial for our analysis, this report will specifically focus on addressing and presenting findings related to the three main questions identified as central to our study.

## What is the impact on Air Quality if we think about GDP and how many people live in an area?

To study how GDP and population affects air quality, we calculated Pearson correlation coefficients. This helped us see the link between air quality indicators and factors like GDP and population. The first image showed a matrix with these results. We saw that **higher population usually means more pollution.** However, the link between GDP and pollution wasn't as clear.

A screenshot of a graph

Description automatically generated

Next, looking at trends: Since the GDP-pollution connection was weak, we needed to dig deeper. We used scatter plots. By looking at data points and trend lines (regression analysis), we could see how GDP related to specific pollutants and overall AQI.

A graph of a number of blue dots

Description automatically generated with medium confidence

The trend lines showed that higher GDP usually meant lower levels of O3, PM10, PM2.5, and better AQI. Even though the initial connection was weak, this told us that better air quality was linked to higher GDP.

Our findings showed that GDP wasn't strongly connected to air quality in the correlation matrix. But when we looked at trends, there was a negative link. This might mean that **richer countries with higher GDPs could better control pollution and improve air quality**. Higher GDP could lead to better infrastructure, technology, and rules for controlling pollution.

In the end, the conclusion is that there isn't a strong direct link between GDP and air quality. But there's a trend showing that countries with higher GDPs usually have cleaner air.

Better air quality is found in smaller areas, while larger populations usually have worse air quality.

## How can we predict air pollution levels in Europe? And which of these factors are most influential in affecting air pollution?

To tackle the question of predicting air pollution levels in Europe, we decided to use a Random Forest model. This approach involves analyzing various factors such as weather conditions, air quality metrics, economic indicators, and renewable energy usage. The key question we aim to answer is: **How effectively can a Random Forest model, using a combination of environmental factors (O3 and NO2 levels, temperature, wind speed), socioeconomic indicators (GDP, population), and renewable energy trends, predict PM10 air pollution levels in European regions?**

Additionally, we want to understand **which of these factors are most significant in influencing air pollution**.

**Analysis of Random Forest Model Predictions for PM10 Air Pollution Levels in European Regions**

### 1. Model Optimization and Performance

The optimization of the Random Forest model for predicting PM10 air pollution levels resulted in identifying the best hyperparameters as follows: a maximum depth of None, a minimum sample split of 2, and 300 estimators. This configuration suggests a comprehensive learning process due to the unrestricted growth of trees and a large number of estimators.

Upon evaluation, the model demonstrated a notable performance. The Mean Squared Error (MSE) was recorded at 1.54, indicating the average squared difference between the predicted and actual PM10 levels. The model's Coefficient of Determination (R²) was found to be 0.91, which implies that a significant portion of the variance in PM10 levels is effectively captured by the model.

A screenshot of a computer

Description automatically generated

Figure 1: Output of the Random Forest model showing hyperparameters, performance metrics, and feature importance."

### 2. Cross-Validation Insights

The Cross-Validation R² scores exhibited variability, ranging from 0.35 to 0.94 with a mean score of approximately 0.69. This variation suggests differing levels of model efficacy across various data subsets, highlighting the model's potential strengths and limitations in diverse data scenarios.

### 3. Feature Importance and Environmental Implications

The analysis of feature importance revealed that O3 levels were the most significant predictor, followed by NO2 levels. This finding underscores the critical impact of these pollutants on PM10 air pollution levels. Additionally, socioeconomic indicators such as GDP and renewable energy trends also held moderate importance, indicating the influence of economic and energy policies on air quality. Temperature, population size, and wind speed, while less impactful, still contributed to the model's predictive accuracy.

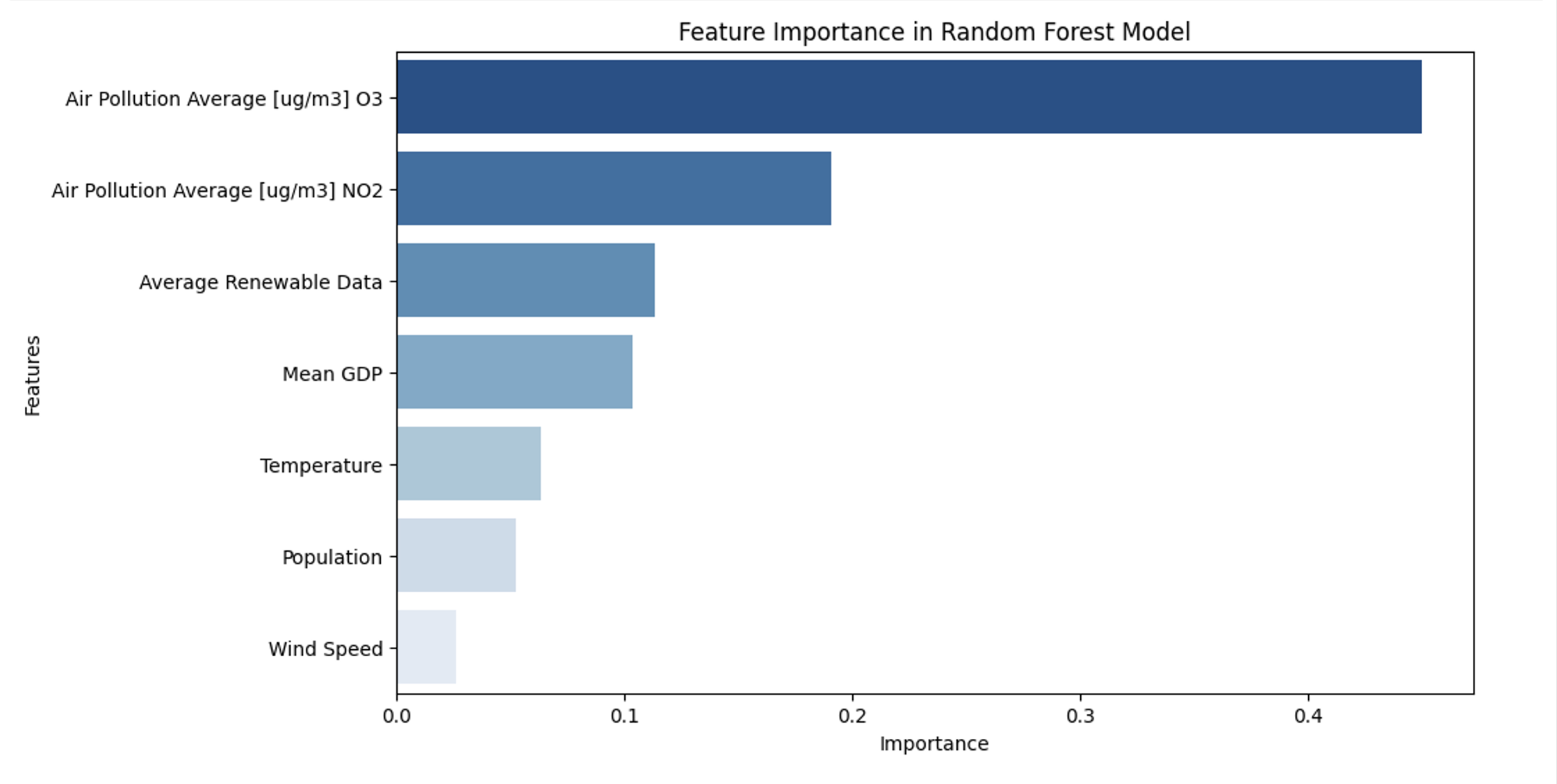


Figure 2: Feature Importance in Random Forest Model

### 4. Policy and Strategic Implications

The results indicate that measures aimed at reducing O3 and NO2 levels could be highly effective in managing PM10 air pollution. The model also highlights the role of economic development and renewable energy in influencing air quality, suggesting that integrated environmental and economic policies could be beneficial.

### 5. Model Reliability and Further Research

Despite the model's strong fit indicated by the high R² value, the variability in cross-validation scores points towards a potential overfitting issue or varying generalizability across different datasets. This necessitates further validation with external datasets to assess the model's applicability in diverse scenarios.

Future research should also explore the reasons behind the variability in cross-validation scores, possibly by examining data subsets or feature interactions more closely. Investigating temporal dynamics, if time-series data is available, can provide insights into how these relationships evolve over time. Additionally, understanding the interaction effects between different predictors could further enhance the model's explanatory power.

In conclusion, the Random Forest model serves as a robust tool for understanding and predicting PM10 air pollution levels, offering significant insights for environmental management and policy formulation.

## What is the impact of geo location/elevation on average AQI?

A map of europe with different colored circles

Description automatically generated

* **Northern Europe**: Displays low AQI, indicating clean air.
* **Central Europe**: Shows mixed AQI levels; certain areas, including Czech Republic and Poland, exhibit higher AQI, suggesting poorer air quality.
* **Southern Europe**: Generally moderate AQI levels, with variations among countries.
* **Western Europe**: Ranges from good to moderate AQI, with smaller countries exhibiting varying air quality.
* **Eastern Europe**: Predominantly low AQI, reflecting cleaner air, with some exceptions like Bulgaria showing moderate levels.

The AQI varies across Europe, with Northern and Eastern Europe enjoying cleaner air overall. Central Europe has areas of concern with higher AQI, whereas Southern and Western Europe have moderate air quality, necessitating region-specific air quality management strategies.

A comparison of a graph

Description automatically generated

**Data Distribution Analysis**

Elevation data is skewed, with a histogram showing a tail, indicating a non-normal distribution. Average AQI data is more normally distributed but exhibits some skewness.

**Correlation Analysis**

A Pearson Correlation Coefficient of 0.278 indicates a weak positive correlation between elevation and average AQI. A p-value of 0.0217 suggests the correlation is mildly statistically significant and should be cautiously interpreted due to the skewness in the elevation data.

**Linear Regression Analysis**

 An R-value of 0.278 reaffirms the weak positive relationship between elevation and average AQI. The matching p-value of 0.0217 from the regression analysis confirms the statistical significance of the results.

**Conclusion** The statistical analysis indicates a weak but significant positive correlation between elevation and average AQI. The non-normal distribution of elevation data advises caution in interpreting these results and suggests the potential benefit of further analysis using alternative statistical models.

# Lesson learned and self-reflection:

Reflecting on this extensive data science project, several key lessons and insights have emerged, profoundly shaping our understanding and approach to data-driven tasks. Firstly, the integration of diverse web scraping tools like Selenium, MechanicalSoup, and BeautifulSoup highlighted the importance of tool versatility in data extraction. Each tool has its strengths – Selenium excels in interacting with dynamic content, while BeautifulSoup and MechanicalSoup are more suited for static content parsing. The ability to select the right tool for the task is crucial.

The deliberate addition of impurities to the dataset and the subsequent data cleaning process was a valuable exercise in understanding real-world data challenges. It taught us that data rarely comes in a clean, ready-to-use format. Handling missing values, inconsistencies, and outliers is not just a routine task but a critical step that can significantly impact the outcomes of data analysis and modeling.

The project’s focus on data cleansing before loading it into MariaDB reinforced the principle that data integrity and accuracy are paramount. This step is often overlooked in the haste to move to analysis but is essential for reliable results.

Conducting comprehensive exploratory data analysis (EDA) and developing a Random Forest model deepened our appreciation for the intricacies of data analysis and predictive modeling. EDA is a powerful process for uncovering underlying patterns and insights, which are vital for informed model building and decision-making.

Creating map visualizations was not only an exercise in technical skill but also in conveying information effectively. It taught us the importance of visual storytelling in data science – the ability to translate complex data into understandable and insightful visual formats.

Overall, this project was a testament to the multifaceted nature of data science. It required a blend of technical skills, critical thinking, and creativity. It emphasized the need for a multidisciplinary approach, where knowledge of various tools and techniques must be complemented with an understanding of the data and its context.

We also learned that this project required considerable time and organization. Teamwork and communication were key elements to its success. Our team worked well together, leveraging individual skills to produce a cohesive and effective outcome. We completed our project and documentation on schedule, utilizing various tools to enhance our productivity and maintain focus.

The project was a clear reminder that data science is a field of continuous learning and adaptation. Staying updated with the latest tools and techniques, being adaptable in problem-solving, and maintaining a keen eye for detail are essential for success in this dynamic and evolving field. The ability to draw actionable insights from complex datasets is not just a skill but an art that develops with experience and persistent learning.