PM2.5 East and West Europe

Jusab Aziz (2763065), Angelo Habib (2821154), Abba Ayad (2737337), Chaniro Kocuvie-Tay (2862242),

2025-06-19

Set-up your environment

Title Page

Jusab Aziz, Angelo Habib, Abba Ayad, Chaniro Kocuvie-Tay, Nasir Ali, Mikail Cörüz, Fatima Malik

Tutorial group 2

Chantal Schouwenaar

Part 1 - Identify a Social Problem

1.1 Describe the Social Problem

Air pollution is still one of the biggest environmental risk to human health, especially in the context of urban areas across Europe. Fine particulate matter PM2.5, has been singled out as a significant higher threat and will be given the most attention. PM2.5 particles are defined as airborne particles with diameters less than 2.5 micrometers, and are also small enough to penetrate into the lungs and even into the bloodstream. Long-term exposure to PM2.5 is also related with numerous chronic health conditions, including asthma, lung cancers, cardiovascular disease, and premature death (World Health Organization [WHO], 2024). The WHO too, reinforce that there is no safe level of PM2.5 exposure, especially for the sensitive sub-populations such as children, the elderly, and individuals with pre-existing health conditions.

In European Union, diverse air quality is seen by regions, and PM2.5 remains an air quality concern in the EU overall even though environmental governance in the EU has improved. The European Environment Agency (EEA) stated that while Western Europe has made substantial advancements in emissions reduction, Eastern European countries still exhibit high PM2.5 levels due to high coal use in heating systems, high industrial activity, and inefficient enforcement (EEA, 2024). The EEA's air quality dashboards show that most of Eastern Europe exceeds WHO's annual PM2.5 standard.

This division is substantiated through Eurostat data on urban population exposure to air pollution by PM2.5 across the EU member states. Eurostat shows that people in Eastern European countries are exposed to concentrations of PM2.5 that are substantially higher on average than their counterparts in the West (Eurostat, 2024). These sustainability concerns and inequalities raise additional questions about environmental inequalities and the unequal distribution of the health risks of air pollution.

In this research study, we are focusing on understanding the health consequences of PM2.5 by looking at differences in mortality rates caused by PM2.5 pollution in Eastern and Western Europe. We had two major sources of data. First, the PM2.5 concentrations based on averages from urban stations across each of the countries in Europe from Eurostat. We cleaned and aggregated this data to find an average per country to make the comparisons more reliable. Second, death statistics published by the EEA looked at death figures

per country that were linked to PM2.5 exposure. We used these figures, and divided the deaths per country by the population, calculating deaths per 100,000 people to normalize for population.

In aggregating these sources, we attempt to paint a clearer picture of the impact PM2.5 pollution has on public health across Europe. The juxtaposition of Eastern and Western Europe highlights variability not only in air quality but also on the environmental outcomes that may result in more equal and/or effective policy decisions. In light of the WHO's strict health-based guidelines, these findings justify the continuation of regional (or even local) air pollution reduction initiatives that advocate for health protection policies in the European region.

Part 2 - Data Sourcing

2.1 Load in the data

The countries we use for East are:

Bulgaria, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia, Slovenia

The countries we use for West are:

Austria, Belgium, France, Germany, Ireland, Luxembourg, Netherlands

Merging

Here I will merge the two data sets that I have loaded in.

```
all_data_east <- bind_rows(
  east_2018, east_2019, east_2020,
  east_2021, east_2022, east_2023)

all_data_west <- bind_rows(
  west_2018, west_2019, west_2020,
  west_2021, west_2022, west_2023)
```

2.2 Provide a short summary of the dataset(s)

```
head(df_mortality_raw)
```

```
## # A tibble: 6 x 35
##   TIME   `2005` ...3  `2007` ...5  `2008` ...7  `2009` ...9  `2010` ...11 `2011`
##   <chr>  <chr>  <lgl> <chr>  <lgl> <chr>  <lgl> <chr>  <lgl> <chr>  <lgl> <chr>
## 1 GEO (~ <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA    <NA>   NA    <NA>
## 2 Europ~ 431202 NA    349335 NA    354004 NA    362672 NA    367509 NA    391884
## 3 Europ~ :      NA    :      NA    :      NA    :      NA    :      NA    :
## 4 Belgi~ 9333   NA    7197   NA    8628   NA    9500   NA    9829   NA    9064
## 5 Bulga~ 19288  NA    17315  NA    19064  NA    16173  NA    15923  NA    20014
## 6 Czech~ 13121  NA    8613   NA    9119   NA    10233  NA    11846  NA    11180
## # i 23 more variables: ...13 <lgl>, `2012` <chr>, ...15 <lgl>, `2013` <chr>,
## #   ...17 <lgl>, `2014` <chr>, ...19 <lgl>, `2015` <chr>, ...21 <lgl>,
## #   `2016` <chr>, ...23 <lgl>, `2017` <chr>, ...25 <lgl>, `2018` <chr>,
## #   ...27 <lgl>, `2019` <chr>, ...29 <lgl>, `2020` <chr>, ...31 <lgl>,
## #   `2021` <chr>, ...33 <lgl>, `2022` <chr>, ...35 <lgl>
```

The "all_data_east" and "all_data_west" datasets contain yearly air pollution data for each country. Each row represents one country in a specific year (from 2018 to 2023). The variables include PM2.5 concentration, number of monitoring stations, threshold exceedances, and other environmental indicators. A "Year" column was added manually to keep track of the time dimension.

The "df_mortality_raw" dataset shows the estimated number of premature deaths attributed to PM2.5 pollution. It is structured in wide format, with countries in the rows and years (e.g., 2016–2021) as column headers. Each value represents the total number of premature deaths for that country in a given year.

2.3 Describe the type of variables included

1. Air Pollution Datasets (all_data_east, all_data_west) The air pollution datasets provide annual, country-level environmental data for the years 2018 to 2023. Most variables are numerical and describe air quality levels, monitoring activity, and regulatory exceedances.

Key variables include:

- **PM2.5 concentration**: The annual average level of fine particulate matter in the air, measured in micrograms per cubic meter ($\mu g/m^3$).
- **Number of monitoring stations**: The total number of stations reporting valid data per country and year.
- **Threshold exceedances**: The number of instances where measured pollution levels exceeded legal EU air quality thresholds.
- **Population exposure**: In some years, indicators report the proportion of the population exposed to levels above legal limits.
- **Country**: The name of the reporting country.
- **Year**: The calendar year of the observation (manually added).

Data source and collection method:
These data are collected through official **air quality monitoring stations** managed by national environmental authorities. The values are compiled and standardized by the **European Environment Agency (EEA)**. All data are obtained from administrative or sensor-based sources, not from individuals or surveys.

2. Mortality Dataset (df_mortality_raw) The mortality data set contains health outcome estimates related to air pollution exposure. It presents the number of premature deaths attributable to PM2.5 per country and year, based on modelled estimates.

Key variables include:

- **TIME**: Country name.
- **Year columns**: Each year column represents the total number of estimated premature deaths for that country and year.

Data source and collection method:
This data set originates from the United Nations Sustainable Development Goals (SDG) indicator database, specifically SDG 11.6.2. The figures are not directly observed but are produced using statistical models that combine environmental exposure, population demographics, and epidemiological data. As such, these are administrative and model-based estimates, not derived from individual-level surveys or medical records.

Part 3 - Quantifying

3.1 Data cleaning

Merging

```
# Combine
combined_data <- bind_rows(all_data_east, all_data_west)
```

Data cleaning

Here I will data clean the combined data file because it included a lot of weird numbers and rows.

Data cleaning (average of cities PM2.5)

```
# Compute average PM2.5 by country, region, and year
avg_pm25_by_country_year <- df_combined_clean %>%
  group_by(country, region, year_of_statistics) %>%
  summarise(average_pm25 = mean(air_pollution_level, na.rm = TRUE), .groups = "drop") %>%
  arrange(year_of_statistics, region, country)
```

Data cleaning (mortality)

Here I will data clean the mortality data file because it included a lot of weird numbers.

3.2 Generate necessary variables

Merging final

```
merged_data <- full_join(df_mortality_clean, avg_pm25_by_country_year, by = c("year_of_statistics", "co
```

Merging to get the filtered data to use further

```
merged_data$premature_deaths <- as.numeric(gsub(",", "", merged_data$premature_deaths))
filtered_data <- merged_data %>%
  filter(!is.na(premature_deaths), !is.na(average_pm25))
countries_with_data <- unique(filtered_data$country)
```

Variable 1: Population

```
population_region_df <- tibble::tibble(
  country = c(
    "Austria", "Belgium", "Bulgaria", "Estonia", "France", "Germany", "Hungary",
    "Ireland", "Latvia", "Lithuania", "Luxembourg", "Netherlands", "Poland",
    "Romania", "Slovakia", "Slovenia"
```

```
  ),
  population = c(
    9, 11.5, 6.5, 1.3, 67, 83, 9.6, 5.1, 1.9, 2.8, 0.65, 17, 38, 19, 5.4, 2.1
  ) * 1e6,
  region = c(
    "Western", "Western", "Eastern", "Eastern", "Western", "Western", "Eastern",
    "Western", "Eastern", "Eastern", "Western", "Western", "Eastern", "Eastern",
    "Eastern", "Eastern"))
```

Variable 2: Average deaths

```
avg_deaths <- filtered_data %>%
  group_by(country) %>%
  summarise(avg_deaths = mean(premature_deaths, na.rm = TRUE)) %>%
  inner_join(population_region_df, by = "country") %>%
  mutate(
    deaths_per_100k = (avg_deaths / population) * 100000)
```

Variable 3: deaths per 100.000 people

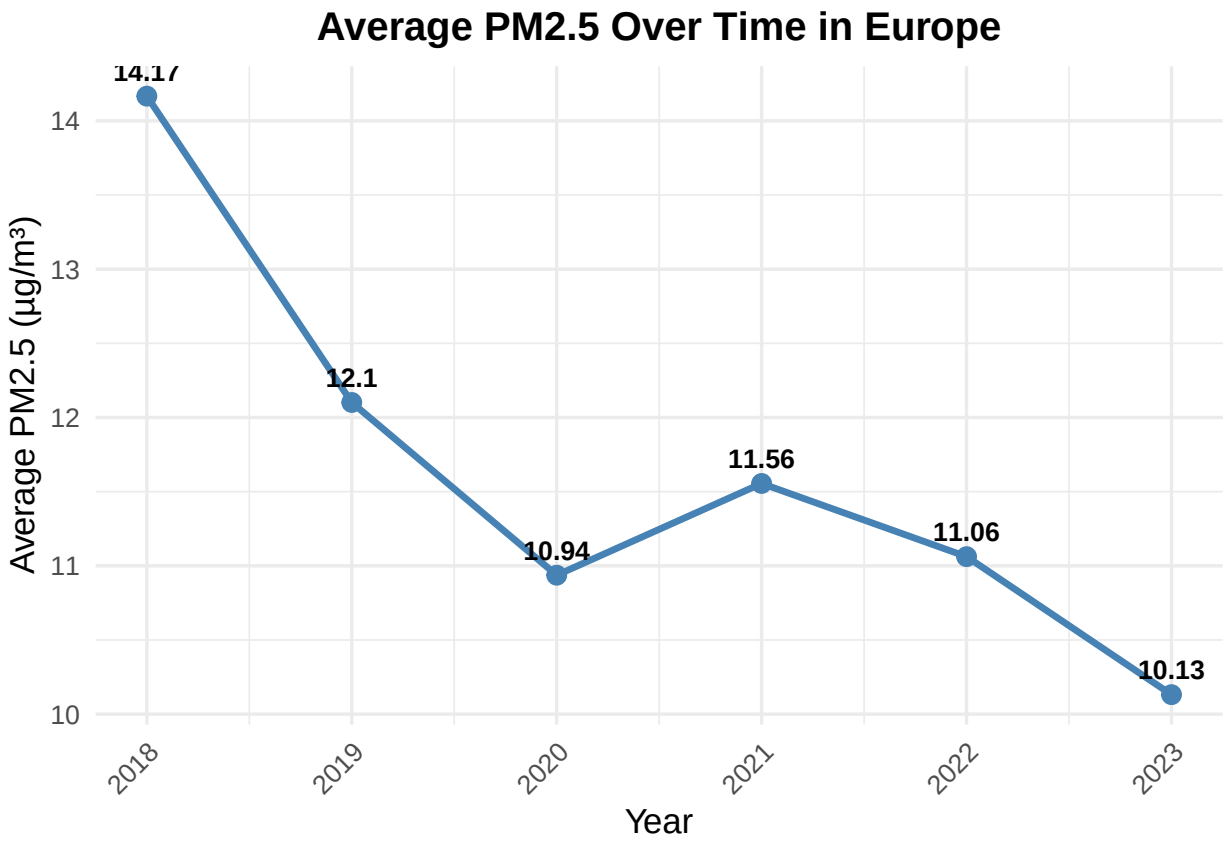
```
region_summary <- avg_deaths %>%
  group_by(region) %>%
  summarise(
    avg_deaths_per_100k = mean(deaths_per_100k),
    .groups = "drop"
  ) %>%
  mutate(region = fct_reorder(region, avg_deaths_per_100k))
```

3.3 Visualize temporal variation

```
pm25_trend <- merged_data %>%
  filter(!is.na(average_pm25), !is.na(year_of_statistics)) %>%
  group_by(year_of_statistics) %>%
  summarise(avg_pm25 = mean(average_pm25, na.rm = TRUE)) %>%
  ungroup()

# Plot
ggplot(pm25_trend, aes(x = year_of_statistics, y = avg_pm25)) +
  geom_line(color = "steelblue", size = 1.2) +
  geom_point(color = "steelblue", size = 3) +
  geom_text(aes(label = round(avg_pm25, 2)), vjust = -0.8, size = 3.5, fontface = "bold") +
  theme_minimal(base_size = 13) +
  labs(
    title = "Average PM2.5 Over Time in Europe",
    x = "Year",
    y = "Average PM2.5 (µg/m³)"
  ) +
  theme(
    plot.title = element_text(face = "bold", size = 15, hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1))
```

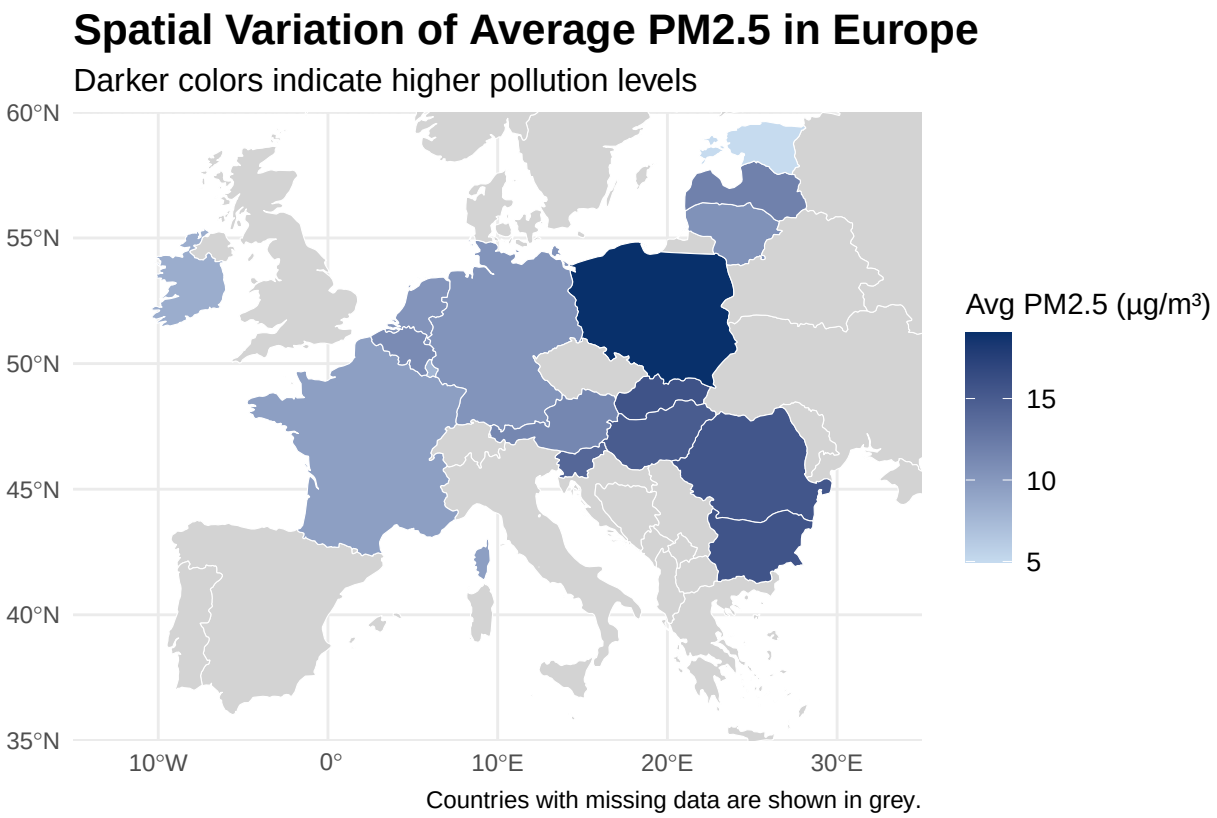
```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



The plot shows average PM2.5 concentrations in Europe from 2018 to 2022. PM2.5 is fine particulate matter that can pose health risks when above dangerous levels. The average concentration was 14.17 µg/m³ in 2018, the highest point in the time series dataset. The concentrations dropped sharply in the following two years, to 12.1 µg/m³ in 2019, and 10.94 µg/m³ in 2020. The significant decline of PM2.5 for the year 2020 may be attributed to the COVID-19 pandemic lockdowns and lower industrial outputs.

In 2021, average PM2.5 concentrations rose slightly to 11.56 µg/m³, perhaps because of renewed economic and social activities. In 2022, PM2.5 concentrations declined again to 11.06 µg/m³, while remaining slightly above the 2020 low average PM2.5 concentrations. In summary, the data trend indicates a general decline of PM2.5 pollution for the time series data, with a slight increase for the year 2021 due to temporary changes in human activity.

3.4 Visualize spatial variation



The code produces a choropleth map that shows the spatial pattern of average PM2.5 concentrations in Europe. Specifically, the countries in which the shading represents the average PM2.5 concentration, shading mean that higher PM2.5 concentrations appear in darker shades of blue, while lower concentrations appear in lighter shades of blue. Countries with no data appear in grey.

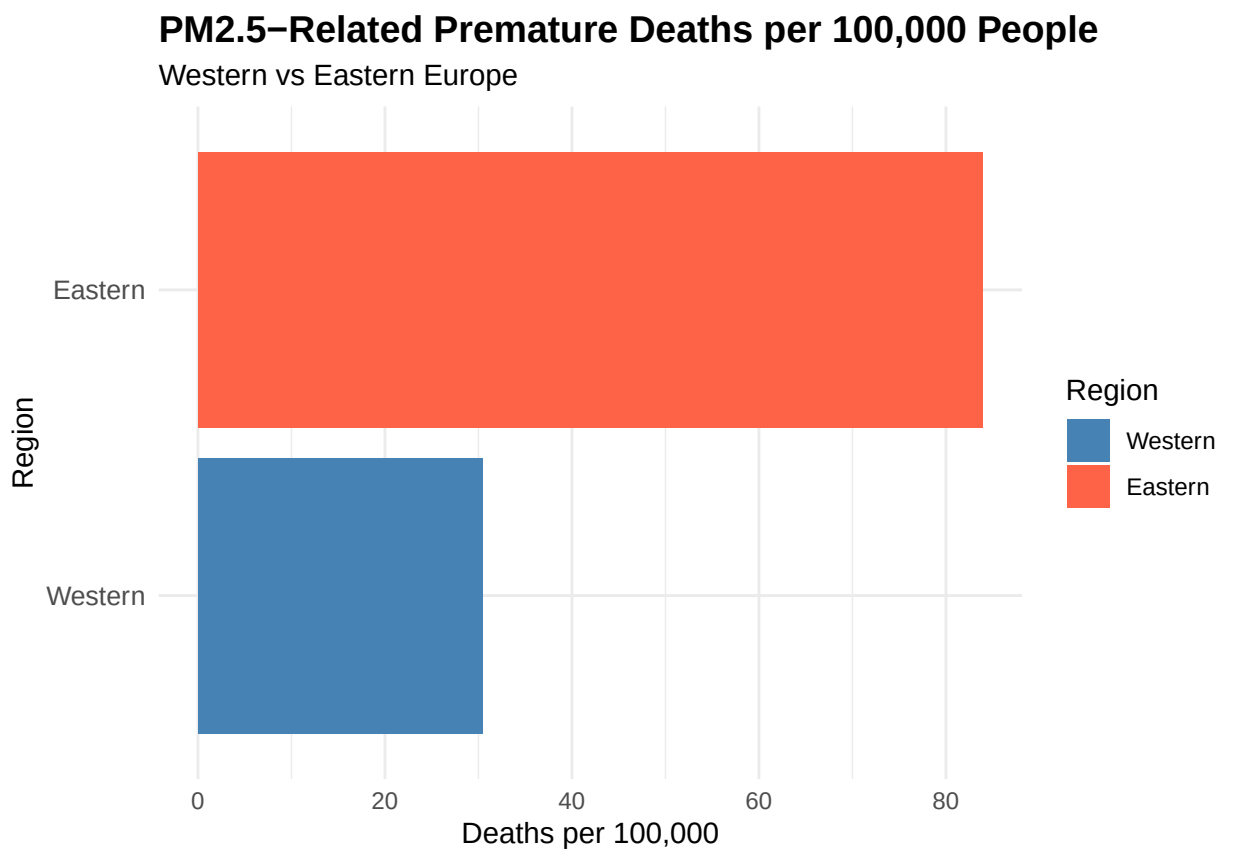
The map visually captures the regionalization of air concentrations. For example, if particular areas of Eastern or Southern Europe are shaded darker than the rest of Europe, then these areas have a higher average PM2.5 concentrations than elsewhere in Europe. This exercise is important because it shows where the geographic hotspots for air pollution are in Europe, and it may determined whether targeted environmental, or public health policy will need to be enacted.

3.5 Visualize sub-population variation

What is the poverty rate by state?

```
ggplot(region_summary, aes(x = region, y = avg_deaths_per_100k, fill = region)) +
  geom_col() +
  coord_flip() +
  scale_fill_manual(values = c("Western" = "steelblue", "Eastern" = "tomato")) +
  labs(
    title = "PM2.5-Related Premature Deaths per 100,000 People",
    subtitle = "Western vs Eastern Europe",
    x = "Region",
    y = "Deaths per 100,000",
```

```
    fill = "Region"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    axis.text.y = element_text(size = 10))
```



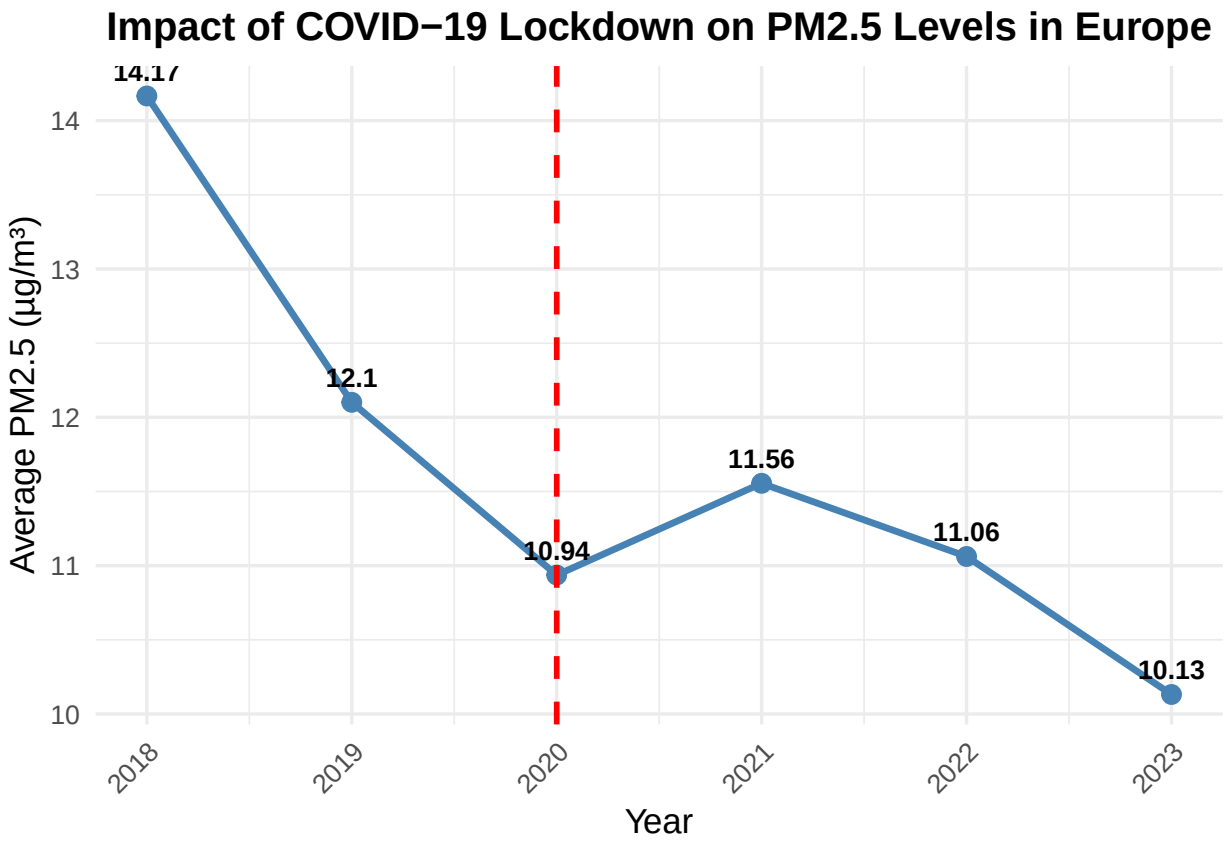
The bar chart compares the total average rate of PM2.5-related premature deaths per 100,000 people in Western and Eastern Europe. The horizontal layout of the bar chart clearly shows that Eastern Europe had a much higher average rate of PM2.5-related premature deaths attributed to air pollution than Western Europe.

More specifically, the rate of PM2.5-related premature deaths in Eastern Europe was over 80 per 100,000 while in Western Europe it was closer to 30 per 100,000. The greater number of PM2.5-related premature deaths in Eastern Europe compared to Western Europe is significant. One interpretation of this finding is that people in the Eastern part of Europe may be at increased risk to the health hazards of PM2.5 mortality, possibly due to living in a more polluted environment, a greater share of heavy economic activities, less stringent policies and regulations of air pollution, or differences in the regions' healthcare systems and public health interventions.

In conclusion, the bar chart highlights the very clear inter-regional inequality in health impacts from air pollution in Europe.

3.6 Event analysis

Analyze the relationship between two variables.



This plot visualizes the effect of the COVID-19 lockdown in 2020 on PM2.5 in Europe over 2018 - 2022. The dashed vertical red line is the year 2020 - when most of Europe had lockdowns and restrictions due to the pandemic.

Before 2020, PM2.5 declined, from 14.17 µg/m³ in 2018 down to 12.1 µg/m³ in 2019. During 2020, however, it dropped sharply down to 10.94 µg/m³, the lowest measurement of the entire time frame, likely due to a simultaneous decline in transportation, industrial activity, and mobility during the lockdown. This provides strong visual proof of what is possible when humankind shifts its behavior (and economy!) on a wide scale, impacting the environment positively.

In 2021, PM2.5 began to increase again to 11.56 µg/m³, indicating some normalization of activity and in 2022 PM2.5 dropped slightly to 11.06 µg/m³. Overall, the plot indicates a clear temporary environmental improvement related to the COVID-19 event, which was followed by a partial rebound.

Part 4 - Discussion

4.1 Discuss your findings

Our analysis reveals a clear regional divide in air quality and its impact on public health across Europe, specifically when comparing Eastern and Western Europe.

First, the spatial analysis shows that Eastern European countries consistently have higher average PM2.5 levels than Western European countries. Countries like Poland, Romania, Bulgaria, and Hungary showed the most concerning pollution levels. In contrast, France, Germany, and Ireland had much lower PM2.5 concentrations. This supports the hypothesis that air pollution is geographically uneven across Europe,

and suggests that factors such as outdated industrial infrastructure, energy production methods, and less stringent environmental policies in the East contribute to this discrepancy.

Second, the temporal analysis from 2018 to 2022 shows a general decline in PM2.5 levels, reaching a low point in 2020. This dip likely correlates with the COVID-19 pandemic, during which lockdowns reduced transportation and industrial emissions. However, the rebound in 2021 indicates that these reductions were temporary, and that without long-term policy changes, improvements in air quality may not be sustained. This emphasizes the need for permanent structural interventions rather than relying on short-term events.

Third, the subgroup analysis highlights the public health consequences of this pollution. Eastern European countries not only have higher PM2.5 levels but also show significantly higher mortality rates related to air pollution. For instance, Bulgaria, Romania, and Poland exhibit more than 150 deaths per 100,000 people, whereas Western European countries like France, Belgium, and Austria report less than half that amount. This strongly suggests that higher pollution directly correlates with higher health risks, and it highlights the urgent need for targeted health and environmental policies in Eastern Europe.

In conclusion, our findings support the idea that PM2.5 pollution is both an environmental and a social issue, with unequal health outcomes across Europe. Tackling this problem requires coordinated policy efforts, especially in Eastern Europe, focusing on clean energy, industrial modernization, and public health investment. Without addressing these disparities, the health impacts of air pollution will continue to disproportionately affect populations in the East.

Part 5 - Reproducibility

5.1 Github repository link

Provide the link to your PUBLIC repository here: ...

5.2 Reference list

European Environment Agency. (2024). *Air quality statistics dashboards*. https://www.eea.europa.eu/en/analysis/maps-and-charts/air-quality-statistics-dashboards

Eurostat. (2024). *Urban population exposure to air pollution by particulate matter (PM2.5)*. https://ec.europa.eu/eurostat/databrowser/view/sdg_11_52/default/table

World Health Organization. (2024). *Air pollution*. https://www.who.int/health-topics/air-pollution#tab=tab_1

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. https://www.who.int/publications/i/item/9789240034228

European Environment Agency. (2023). *Air quality in Europe — 2023 report*. https://www.eea.europa.eu/publications/air-quality-in-europe-2023