Air Pollution Across the Globe

Sneha Arya, Greer Moran, and Gabrielle Smeltzer

2025-04-24

Research Topic: Air Pollution Across the Globe

This research focuses on air pollution across the globe, a topic that directly impacts society's health, the climate, and quality of life. As pollution levels are rising exponentially worldwide, it is vital to understand relationships between population distribution and the severity of air pollution. Using a main dataset regarding Global Air Quality and a secondary dataset on global city-level populations, our research centers on observing pollution and population data to explore these patterns. Since population values in our secondary dataset were provided by city, we aggregated them to generate total population figures by country. We chose to visualize data at the country level to reduce clutter and improve clarity among visualizations, allowing for a more accessible and focused exploration of global air pollution patterns.

Research Questions

To guide our analysis, our team develops a set of research questions focused on identifying key relationships between population and pollution levels. One area of focus is how population contributes to pollution across a select six countries.

Another central focus of our project is investigating how specific pollutant values (Carbon Monoxide, Ozone, Nitrogen Dioxide, and Particulate Matter) contribute to a country's overall Air Quality Index (AQI). To explore this, we apply linear regression modeling and plan to visualize the results through a regression summary table and a slope plot illustrating each pollutant's influence on AQI. In addition, we plan to use faceted scatter plots to compare pollutant values to AQI, offering a clear view of how each variable individually correlates with air quality.

We also aim to examine whether urban population percentage plays a role in pollution severity. To do this, we plan to implement a histogram visualization that compares AQI value distributions in relation to levels of urban population across the six countries, helping us identify patterns linked to urbanization. Furthermore, we use a grouped bar graph to present the frequency of AQI categories by country, enabling straightforward comparisons of pollution severity across regions. Lastly, we analyze ozone AQI values in relation to land area through a boxplot to explore how physical geography may intersect with pollution levels.

Overall, this is not an exhaustive list of the questions we aim to explore, as the dataset contains many overlapping features that may reveal new insights. Our goal is to use structured analysis and

clear, accessible visualizations to better understand global air quality trends and their potential implications for effects on the environment and public health.

FAIR Principles

Findable

Our data sources to ensure that the datasets were openly available and had proper citation information. Researchers can find the original datsets through Kaggle.

Accessible

Both of the datasets used are publicly accessible without restrictions, which supports transparency and reproducibility aspects of accessibility. Additionally, we have cited all sources used in our research to allow for direct access for viewers.

Interoperable

The data is in CSV files, which is a standard dataset form for R. There is also consistent and specific variable naming and data wrangling/cleaning.

Reusable

Since the datasets have open licensing on Kaggle with a Creative Commons liscence, the original data sources have a long term usability. By documenting the data cleaning and analysis steps, other people can replicate the work or conduct further research.

CARE Principles

Collective Benefit

Our team's analysis aims to raise awareness of global air pollution trends in the environment globally so people across the world can benefit from the findings. Additionally, visualizations and insights are concise and explained in depth.

Authority to Control

The original authors for the datasets, Hasib Al Muzdadid and Mohamed Fadl are credited in the "Provenance of Data" section, and the datasets were used fo the intended purposes of education and research. The meaning of the data is not altered or misrepresented further than the original purpose.

Responsibility

To promote inclusivity, our team has selected one country per continent (Australia, Brazil, Canada, China, France, and South Africa) using a random name generator, excluding Antarctica due to its low population.

Ethics

Our analysis avoids misleading visualization or unsupported claims. Though our team's intent is to use global pollution data, we acknowledge the potential limitations of focusing on just six specific countries.

Provenance of Data

For this project, our team uses two datasets to explore global air pollution and population patterns. The main Kaggle dataset, "Global Air Pollution Dataset", contains pollutant-specific AQI values and was originally compiled by Hasib Al Muzdadid. The dataset was created to support research in areas like environmental science, pollution forecasting, and public health. As for cases, each case represents a city, with an association of pollutant measurements and country information.

The secondary Kaggle dataset, "Countries Population by Year 2020", provides country-level population data and was originally created by Mohamed Fadl. This dataset provides statistics based on demographics in countries, including total population, population density, land area, fertility rate, median age, and percentage of urban population. Each case in this dataset represents a country or dependency. While the population dataset is organized by country, our analysis focuses on combining both sources at the country level to maintain consistency in our visualizations.

To maintain further clarity in our visualizations and avoid overcrowding, our team chose to not include individual cities in our graphics. Instead, we aggregated the population values by country, using the sum of the city-level populations to represent each country's total. This approach allows us to incorporate population context into our pollution analysis without overwhelming the visualizations. Both of the datasets are publicly available and are intended for educational and research use.

An Analysis of Air Quality

Cases & Attributes:

Cases:

A case is a specific global location.

Our cleaned and tidied data:

The more in-depth cases include the countries: Australia, Brazil, Canada, China, France, and South Asia.

These cases contain multiple cities worth of population data of varying counts.

Attributes:

The attributes from within the data include: city, AQI Value, AQI Category, CO2 AQI Value, CO2 AQI Category, Ozone AQI Value, Ozone AQI Category, NO2 AQI Value, NO2 AQI Category, PM2.5 AQI Value, PM2.5 AQI Value, Population during 2020, Yearly Population Change, Net Change, Density per Km, Land Area (Km), Migrants Count, Fertile Rate, Median Age, Urban Population Percentage, and World Share Percentage.

We chose to focus on the values from each pollutant (CO2, Ozone, NO2, & PM2.5), the overall AQI Value and Category, Land Area (Km), and Urban Population. These attributes were parts of our original datasets. The datasets were then combined and cleaned to represent a smaller case set so that comparisons would be easier attained.

Table 1: Linear Regression Summary for Pollutants impact on AQI

Pollutant	Est. Slope	Std. Error	t value	p-value
(Intercept)	-0.60391	0.11144	-5.41918	0.00000
CO	0.01707	0.03970	0.43007	0.66715
Ozone	0.15751	0.00234	67.37098	0.00000
NO2	-0.03620	0.01350	-2.68260	0.00731
PM	0.98014	0.00126	775.38517	0.00000

Figure 1

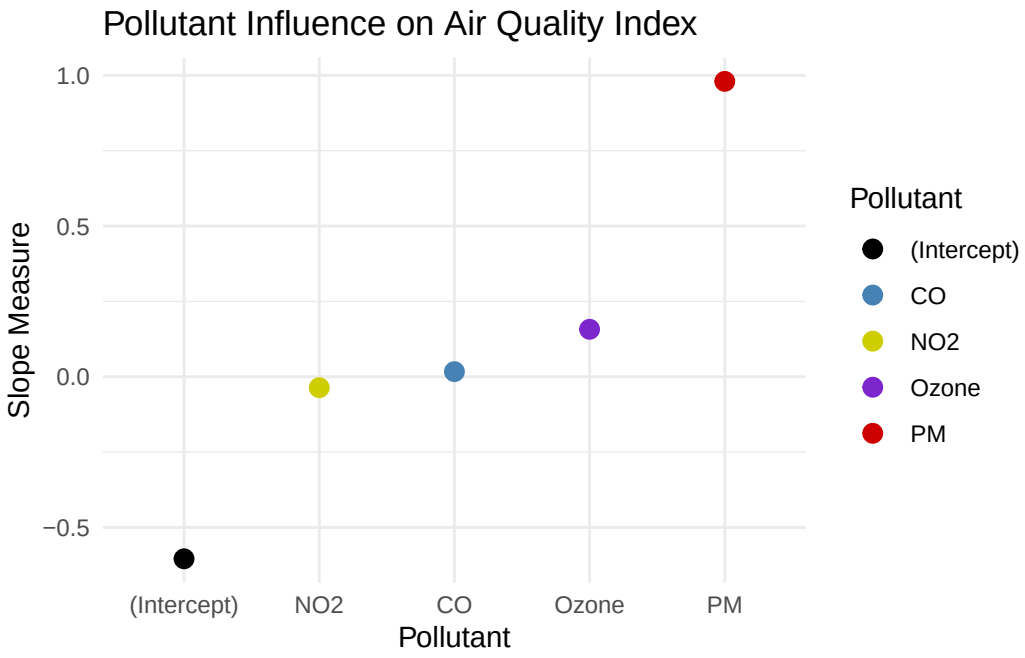
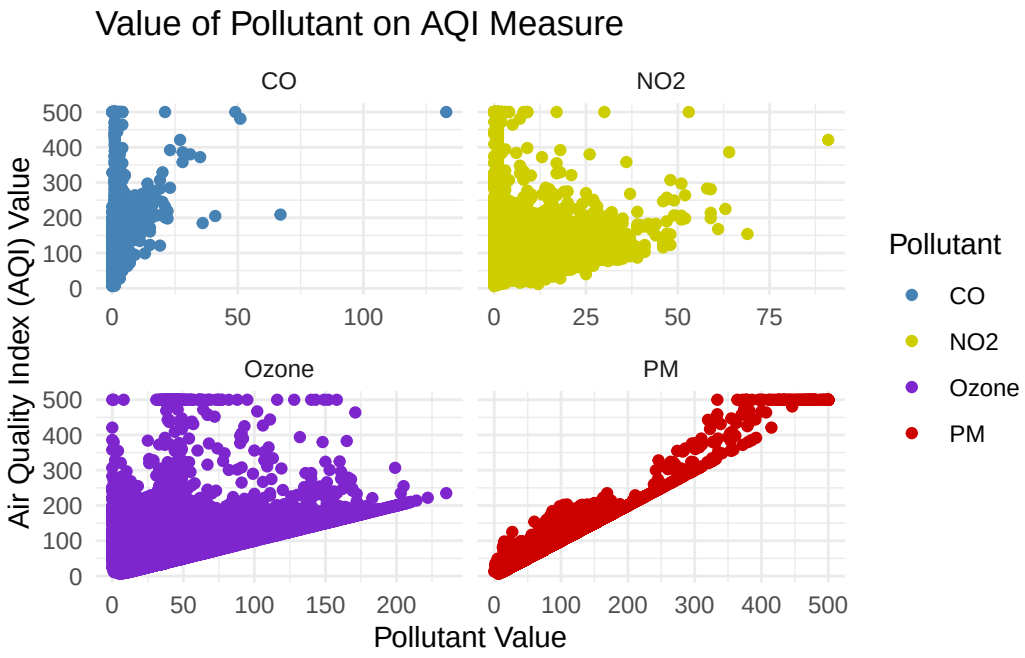
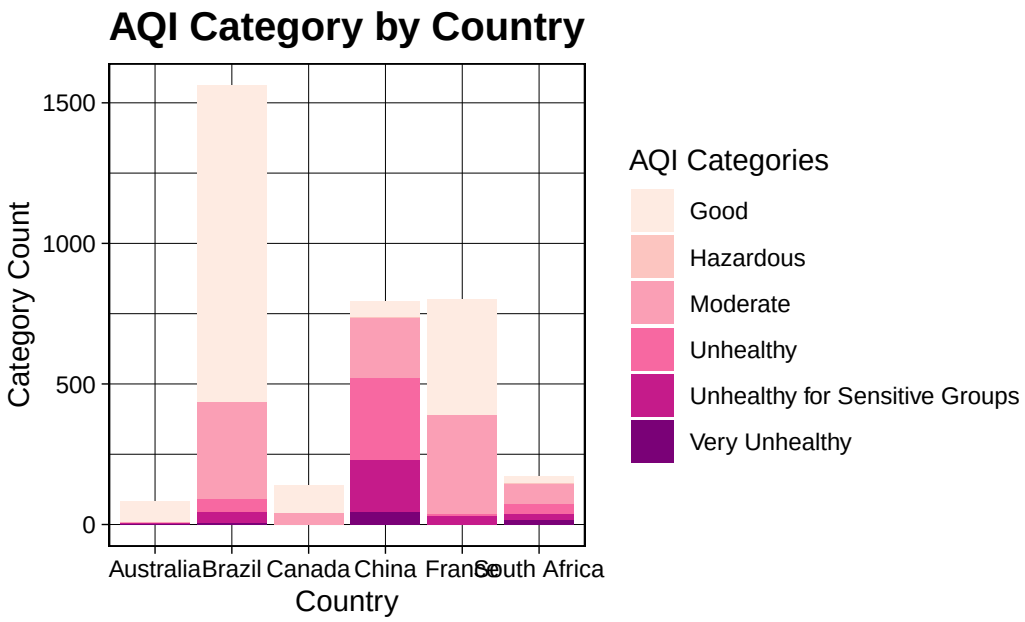


Figure 2


Value of Pollutant on AQI Measure
CO
NO2
Ozone
PM
Air Quality Index (AQI) Value
Pollutant Value
Pollutant
CO
NO2
Ozone
PM

AQI Values

Figure 3



A count of each AQI category by country.

Figure 4

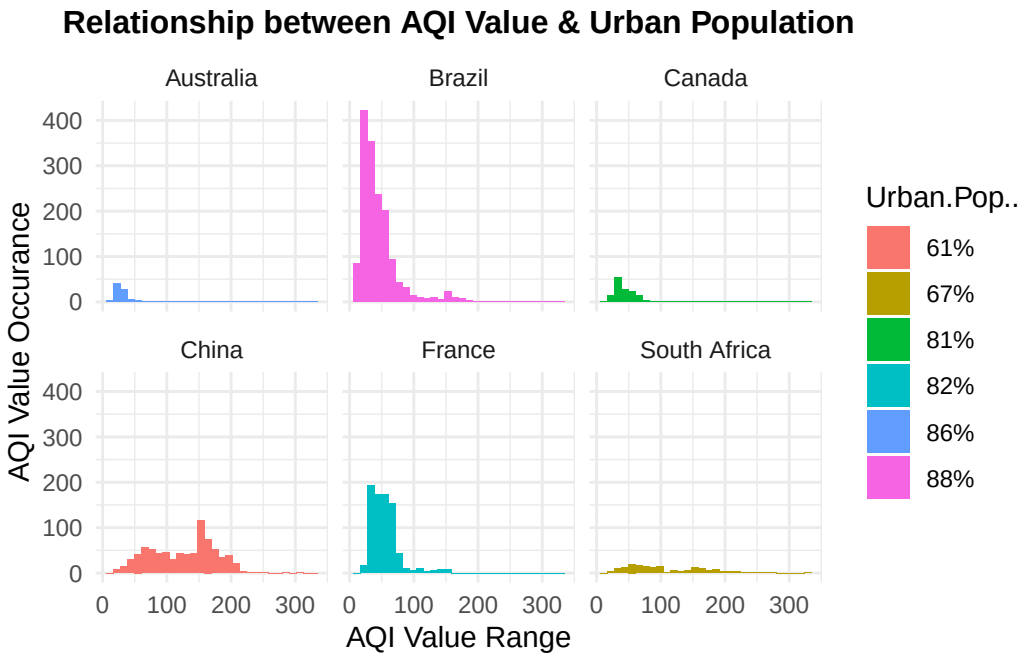
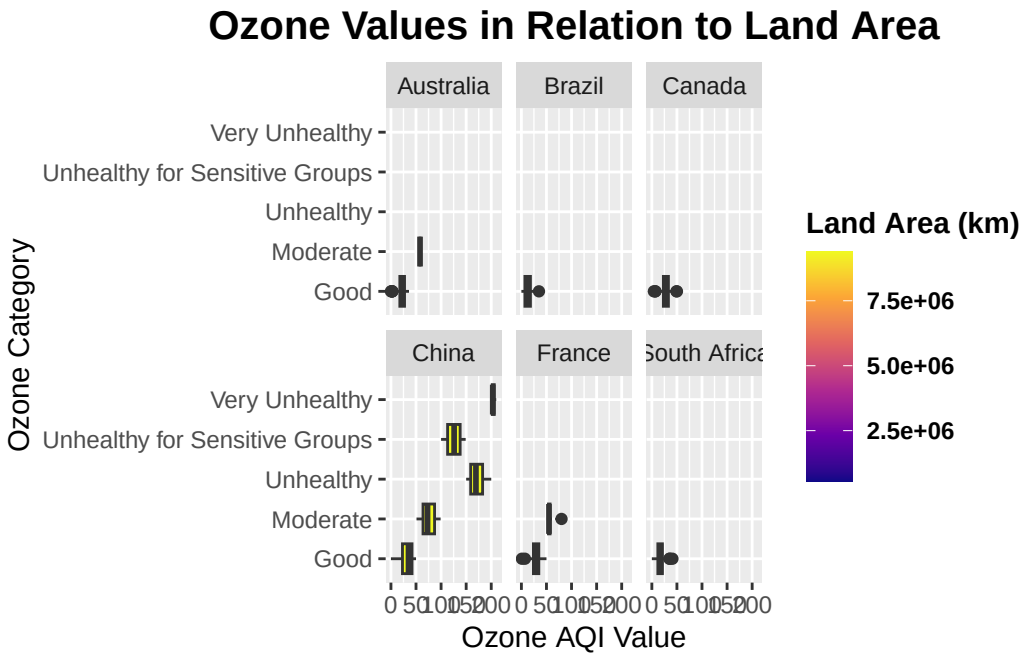

Relationship between AQI Value & Urban Population
Australia
Brazil
Canada
China
France
South Africa
AQI Value Occurance
AQI Value Range
0
100
200
300
400
Urban.Pop..
61%
67%
81%
82%
86%
88%

Figure 5


Ozone Values in Relation to Land Area
Australia
Brazil
Canada
China
France
South Africa
Very Unhealthy
Unhealthy for Sensitive Groups
Unhealthy
Moderate
Good
Ozone Category
0 50 100 150 200
Ozone AQI Value
Land Area (km)
7.5e+06
5.0e+06
2.5e+06

Data Sources and Acknowledgement

Fadl, Mohamed. "Countries Population by Year 2020." Kaggle, 19 June 2020, www.kaggle.com/datasets/eng0moha
by-country-2020.

Muzdadid, Hasib Al. "Global Air Pollution Dataset." Kaggle, 8 Nov. 2022, www.kaggle.com/datasets/hasibalmuzda
air-pollution-dataset?resource=download.

"Wheel of Names." Wheel of Names, wheelofnames.com/. Accessed 4 May 2025.

Code Appendix:

```
# Load needed packages ---
## Must determine which are used in our code so that it is executed


library(tidyverse)
library(googlesheets4)
library(knitr)
library(dcData)
library(tinytex)
library(dplyr)
library(ggplot2)
library(kableExtra)

#| label: Data Wrangling Code for Global Air Pollution
#| lst-label: Global Air Pollution, Data Wrangling
#| lst-cap: "Rename Needed Columns"


# Wrangle Air Pollution Data ---

## Get Correct Column Names ---
### Get Specific Data Frame of Correct Column Names


globalData <- read.csv("statproject.csv")

cleanAQIdata <- globalData %>%
  rename(c(
    AQI = "AQI.Value",
    CO = "CO.AQI.Value",
    Ozone = "Ozone.AQI.Value",
    NO2 = "NO2.AQI.Value",
    PM = "PM2.5.AQI.Value"
    )
```

```r
    )

aqiModel <- lm(AQI ~ CO + Ozone + NO2 + PM, data = cleanAQIdata)
aqiSummary <- summary(aqiModel)


# Create Table Summary Visualization ---

aqiModel <- as.data.frame(aqiSummary$coefficients) ## create data frame
aqiModel$Pollutant <- rownames(aqiModel)
aqiModel <- aqiModel[, c("Pollutant", "Estimate", "Std. Error", "t value", "Pr(>|t|)")]
colnames(aqiModel) <- c("Pollutant", "Est. Slope", "Std. Error", "t value", "p-value")
rownames(aqiModel) <- NULL

kable(
  aqiModel,
  caption = "Linear Regression Summary for Pollutants impact on AQI",
  digits = 5,
  format = "latex"
  ) %>%
  kable_styling(
    position = "center"
  )

#
print(
ggplot(
  data = aqiModel,
  mapping = aes(
    x = reorder(Pollutant, `Est. Slope`),
    y = `Est. Slope`,
    color = Pollutant
    )
  ) +
  geom_point(
    size = 3
    ) +
  labs(
    title = "Pollutant Influence on Air Quality Index",
    x = "Pollutant",
    y = "Slope Measure"
  ) +
  scale_color_manual(
    values = c("black", "steelblue", "yellow3", "purple3", "red3")
  )
  + theme_minimal()
)
```

```r
# AQI Value by Pollutant Value Data Frame
## Tidy data frame so that Pollutant is a column

tidyAQIdf <- cleanAQIdata %>%
  pivot_longer(
    cols = c(
      CO,
      Ozone,
      NO2,
      PM
    ),
    names_to = "Pollutant",
    values_to = "Pollutant_Value"
  )

#AQI Value by Pollutant Value Scatterplots
## Create facetted scatterplots using ggplot2

ggplot(
  data = tidyAQIdf,
  mapping = aes(
    x = Pollutant_Value,
    y = AQI,
    color = Pollutant
    )
  )+
    geom_point() +
    facet_wrap(~ Pollutant, scales = "free_x") +
    labs(
      title = "Value of Pollutant on AQI Measure",
      x = "Pollutant Value",
      y = "Air Quality Index (AQI) Value"
    ) +
    scale_color_manual(
      values = c("steelblue", "yellow3", "purple3", "red3")) +
    theme_minimal()

## Create space for population and pollution data
pollutionData <- read.csv("statproject.csv")
populationData <-read.csv("statproject1.csv")

## Wrangle the data for calling and combination
## Combine the data for population
populationPollution <-left_join(
  x = pollutionData,
  y = populationData,
  by = join_by(Country == 'Country..or.dependency.')
```

```
)


##Filter out all cities without population

tidypopPoll <- populationPollution %>%

filter(if_all(Population..2020., ~ !is.na(.)))

##Filter our data to a country from each continent, excluding Antarctica

filteredCountry <- c("Australia", "Brazil", "Canada", "China","France", "South Africa"  )
specificCountry <- tidypopPoll %>%
  filter(Country %in% filteredCountry)

##Create Comparison Visuals

ggplot(specificCountry) +
  aes(x = Country, fill = AQI.Category) +
  geom_bar() +
  scale_fill_brewer(palette = "RdPu", direction = 1) +
  labs(
    x = "Country",
    y = "Category Count",
    title = "AQI Category by Country",
    caption = "A count of each AQI category by country.",
    fill = "AQI Categories"
  ) +
  theme_linedraw() +
  theme(
    plot.title = element_text(size = 15L,
    face = "bold",
    hjust = 0.5),
    plot.caption = element_text(hjust = 0.5)
  )


specificCountry %>%
 filter(AQI.Value >= 10L & AQI.Value <= 350L) %>%
 ggplot() +
  aes(x = AQI.Value, fill = Urban.Pop..) +
  geom_histogram(bins = 30L) +
  scale_fill_hue(direction = 1) +
  labs(
    x = "AQI Value Range",
    y = "AQI Value Occurance",
    title = "Relationship between AQI Value & Urban Population"
```

```
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12L,
    face = "bold",
    hjust = 0.5)
  ) +
  facet_wrap(vars(Country))
ggplot(specificCountry) +
  aes(
    x = Ozone.AQI.Value,
    y = Ozone.AQI.Category,
    fill = Land.Area..Km..
  ) +
  geom_boxplot() +
  scale_fill_viridis_c(option = "plasma", direction = 1) +
  labs(
    x = "Ozone AQI Value",
    y = "Ozone Category",
    title = "Ozone Values in Relation to Land Area",
    fill = "Land Area (km)"
  ) +
  theme_gray() +
  theme(
    plot.title = element_text(size = 15L,
    face = "bold",
    hjust = 0.5),
    legend.text = element_text(face = "bold"),
    legend.title = element_text(face = "bold")
  ) +
  facet_wrap(vars(Country))
```