Air Pollution Across the Globe

Sneha Arya, Greer Moran, and Gabrielle Smeltzer

2025-04-24

Research Topic: Air Pollution Across the Globe

Research Questions

In order to analyze air pollution across the globe, we first need to examine how air pollution is impacted by pollutants. Further, how does pollutant air quality index values (e.g. Carbon Dioxide, Ozone, Nitrogen Dioxide, and Particle Matter) impact the overall air quality index, which are most contributing to poor air quality index? Understanding the strength of each pollutant's relationship with air quality will give better insight to how AQI is related by each of these pollutants.

An Analysis of Air Quality

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

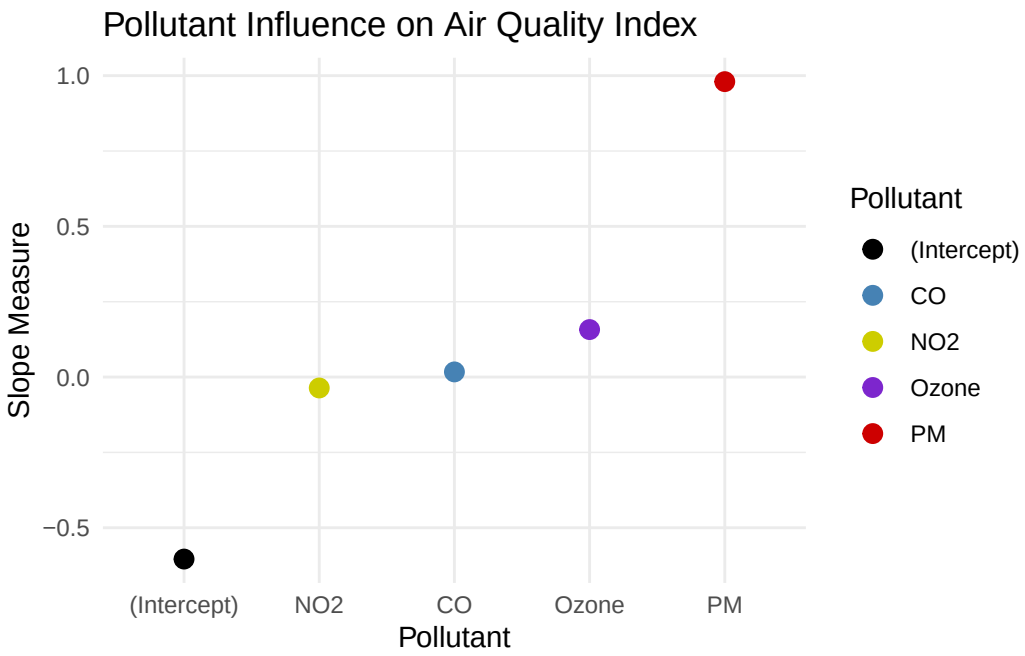


Figure 2

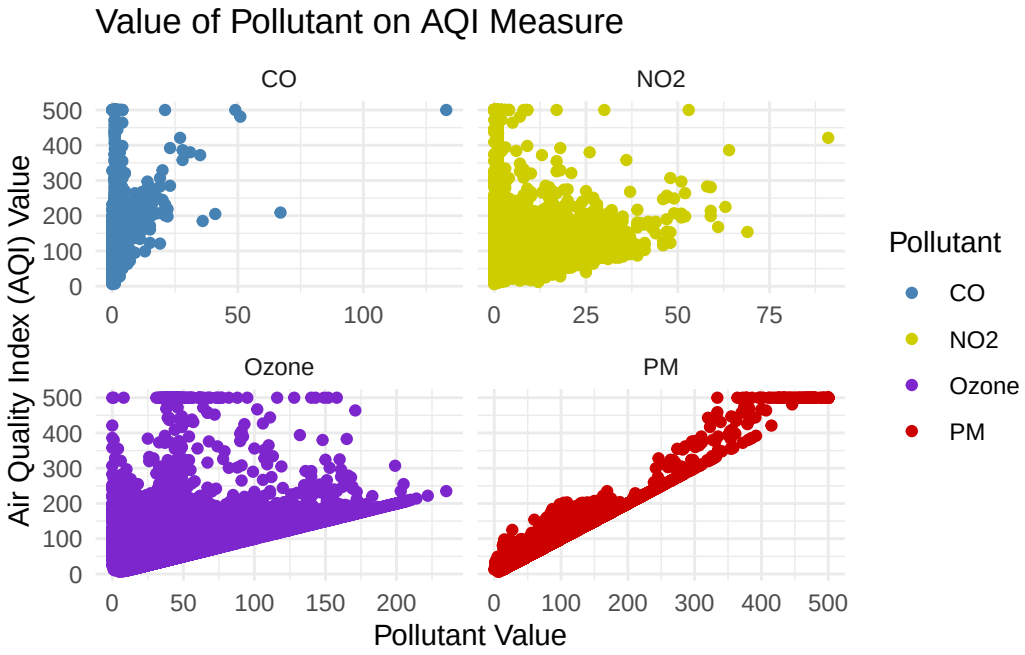
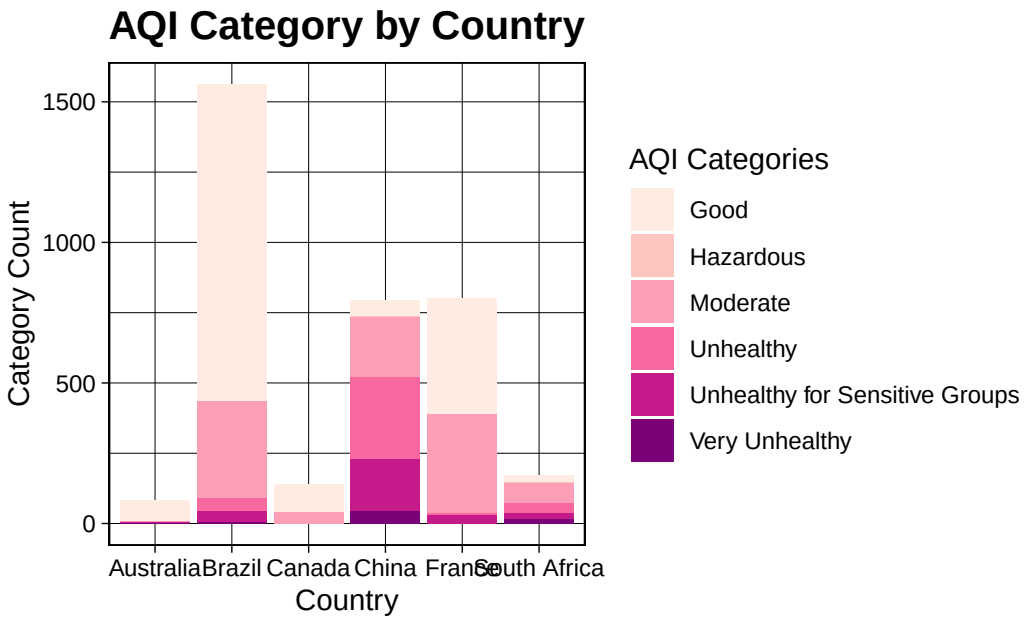


Figure 3



A count of each AQI category by country.

Figure 4

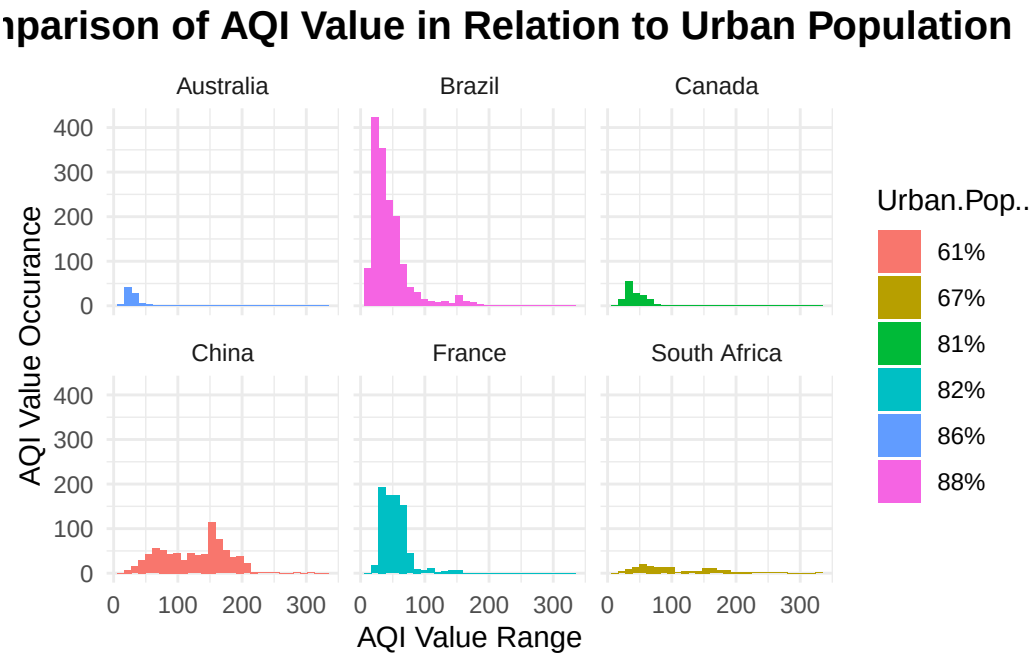
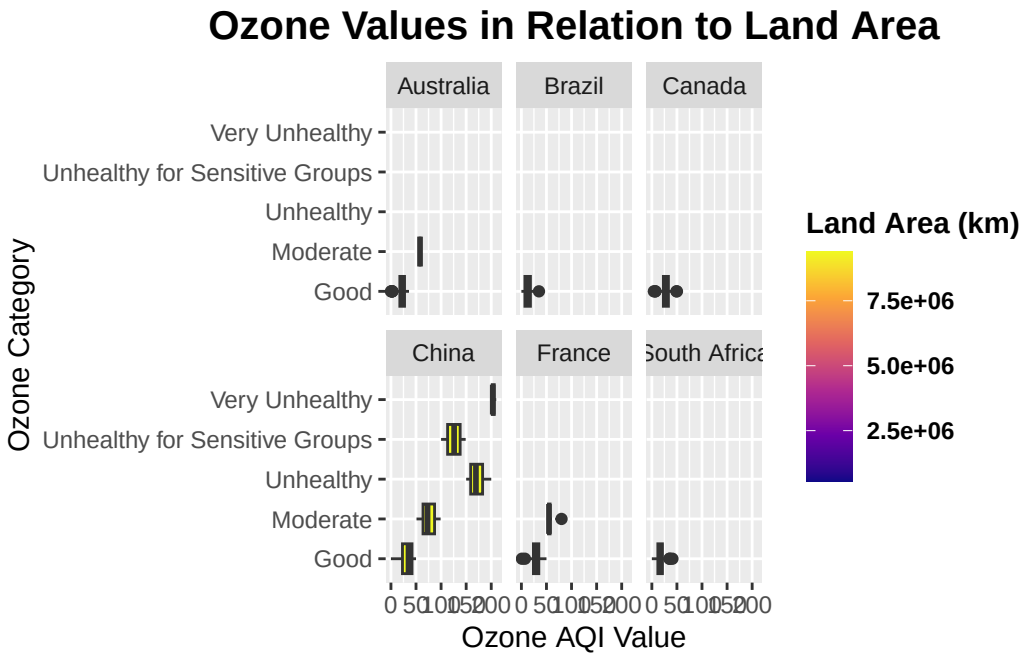

ıparison of AQI Value in Relation to Urban Population
Australia
Brazil
Canada
China
France
South Africa
AQI Value Occurance
AQI Value Range
0
100
200
300
400
Urban.Pop..
61%
67%
81%
82%
86%
88%

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

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(kableExtra)
library(dcData)
library(tinytex)
library(dplyr)
library(ggplot2)

#| 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"
    )
    )

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


# Create Table Summary Visualization ---
```

```r
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()
)

# 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 %>%
```

```r
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 = "Comparison of AQI Value in Relation to Urban Population"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 15L,
    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))
```