

# 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 datasets 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 license, 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 for 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.

## Graphics Table of Contents

- Pollutant Influence on AQI:
  - [Figure 1](#)
- Pollutant Values on Air Quality Measure:
  - [Figure 2](#)
- Selected Countries & their Air Quality Measurements:
  - [Figure 3](#)
- Urban Population Rate Influence on Air Quality per Generated Countries:
  - [Figure 4](#)
- Land Area Influence on Ozone Layer Pollutants per Generated Countries:
  - [Figure 5](#)

Table 1

Table 2: 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

## Pollutant Influence on Air Quality Index

Figure 1 - Figure 1

Linear Regression Summary of Pollutant on AQI

This table, Table 1 provides an overview of the regression summary each pollutant (CO, Ozone, NO2, and PM) has on the measured AQI value along different cities across the globe. This offers valuable context about how Air Quality Index (AQI) is measured, how the pollutants influence the strength of AQI. The regression summary includes statistics of estimated slope, standard error, t-value, and p-value.

Key Insights:

- The estimated slope between Particle Matter with diameter 2.5 micrometers or less have a near 1.0 slope of regression line, at 0.98014, indicating they have a very strong relationship.
- The p-value of carbon monoxide (CO) is considerably above the average significance level of 5%, with a value of 0.66715. Despite lack of running a hypothesis test, CO would be considered a statistically insignificant pollutant on AQI measure since the statistic suggests there is a 50% chance that the observed results are due to chance.
- Both nitrogen dioxide and carbon monoxide, NO2 and CO, have near zero estimated slope values of -0.01350 and 0.01707 respectively. Further, the pollutants could have little impact on the air quality measure.

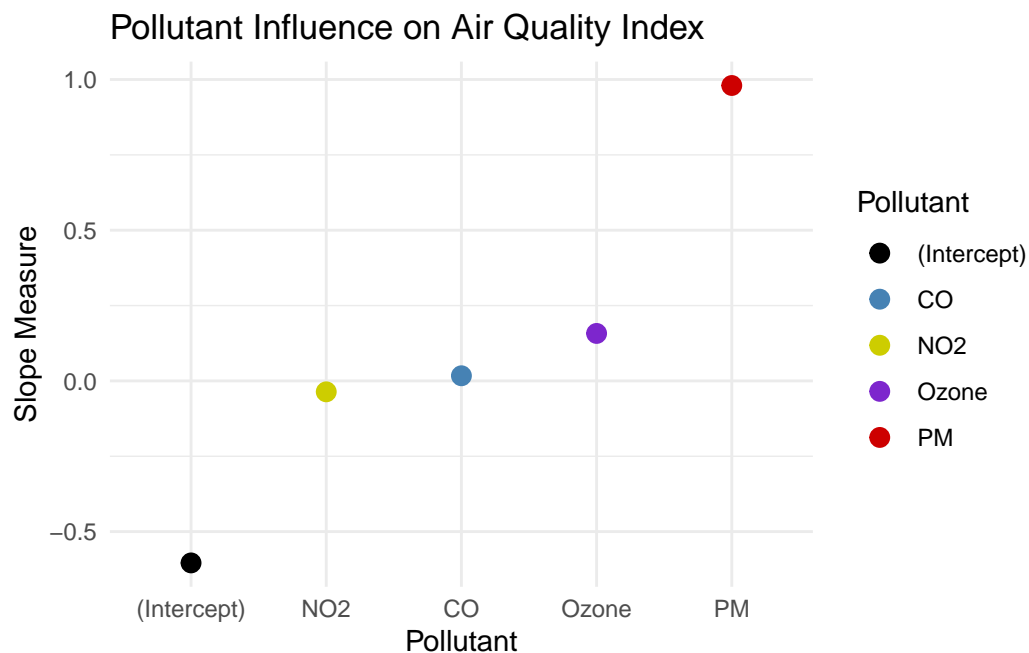
## Linear Regression Slope Evaluation for each Pollutant

The figure(Figure 1) provides the estimated slope of each pollutant's linear regression relationship with air quality index graphed as a single point. This plot of five points helps identify the difference among pollutants estimated slope measured outputted in Table 1.

Key Insights:

- Particle matter with diameter 2.5 micrometers or less (PM) has a considerably stronger relationship with air quality index compared to the other pollutants.
- Both glyphs of nitrogen dioxide and carbon monoxide, NO2 and CO, lie closely to the zero line. Indicating a near-zero relationship with little to no impact on AQI.

Figure 1



## AQI Value by Each Pollutant

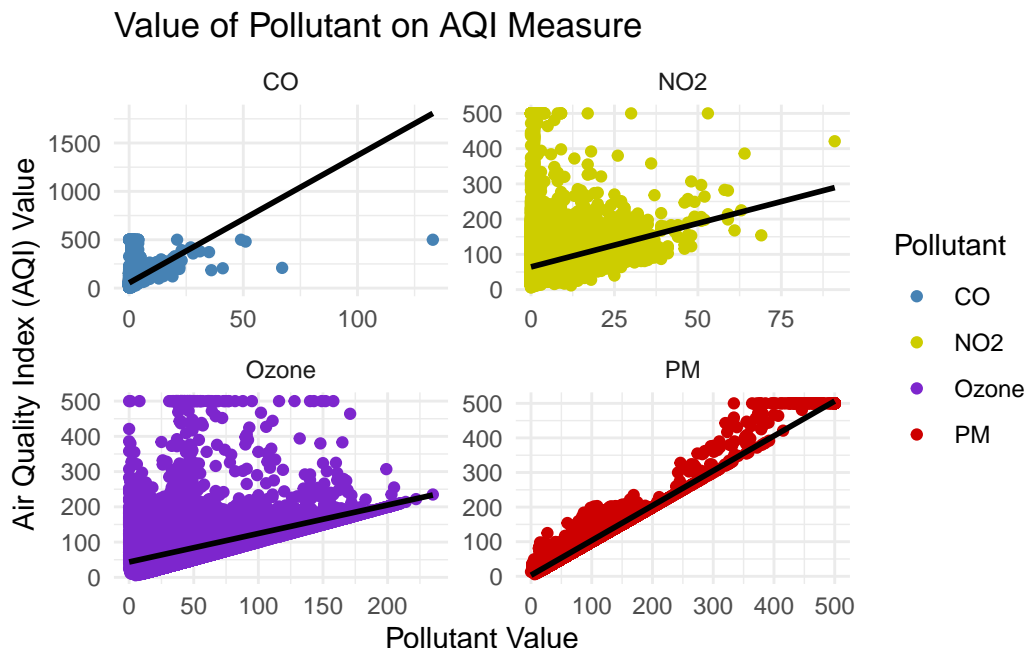
Figure 2 - Scatterplot of Each Pollutant on AQI Measure

This figure (Figure 2) displays each air quality index (AQI) by the value of each pollutant. This scatterplot identifies the spread that creates the slope of each Pollutant on AQI measure.

Key Insights:

- There is a strong positive relationship between particle matter with diameter 2.5 micrometers or less (PM) and air quality index (AQI). A higher particle matter consistently increases the measure of air quality index.
- There is strong variation within Ozone above slope line reflecting a weak-to-moderate relationship with air quality index. This indicates that because the relationship is not strong, air quality index can still have a high value despite having a low Ozone value.
- Both nitrogen dioxide and carbon monoxide, NO<sub>2</sub> and CO, display a near cone shaped spread of variation. The glyphs come to a slightly triangular shape, indicating that there could be heteroscedasticity. Further, the line of slope shown in black for each scatterplot has observed values closer to it with higher values of the pollutant and AQI because the variance of its residuals is most likely not constant. With this observed, a linear regression model is not best to explain the relationship between air quality index and each of these variables.

Figure 2

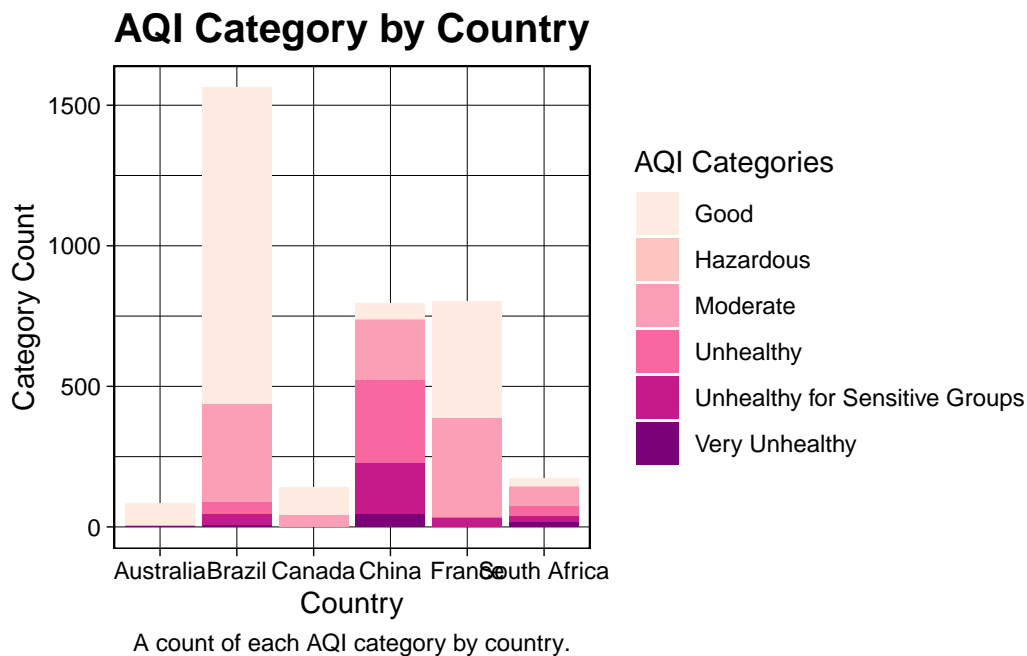


## AQI Values by Country

Countries Selected: Australia, Brazil, Canada, China, France, & South Africa

In researching the relationship between Pollution/Air Quality and Population, there was an excess of data. In order to create visuals that would be readable and clear we randomly generated (Australia, Brazil, Canada, China, France, and South Africa) one from each continent (excluding Antarctica due to its low population). These selected countries are a subset of the overall data and present the global rates of pollution. They display a potential relationship for comparison in each country per continent. The visual, Figure 3, explains the overall relationship between each AQI Category and its relevance to each country. The x-axis contains each country while the y-axis displays the count of each category displayed. Each color within the bars represents the AQI Categories; Good, Hazardous, Moderate, Unhealthy, Unhealthy for Sensitive Groups, and Very Unhealthy. Based on the data, we can note that Brazil had the most data representing a Good AQI while China had a range of all AQI values. This suggests that Brazil, Australia, Canada and France could have the lowest rates of pollution while China and South Africa may have higher rates. It may also be skewed due to a lower collection of data from each city within the countries selected. The AQI data was taken from each city within our selected countries and there may be less cities represented in the countries of Australia, Canada, and South Africa. The shorter bar graph does not necessarily mean that it has less or more pollution, but is a representation of the data that had been previously recorded.

Figure 3

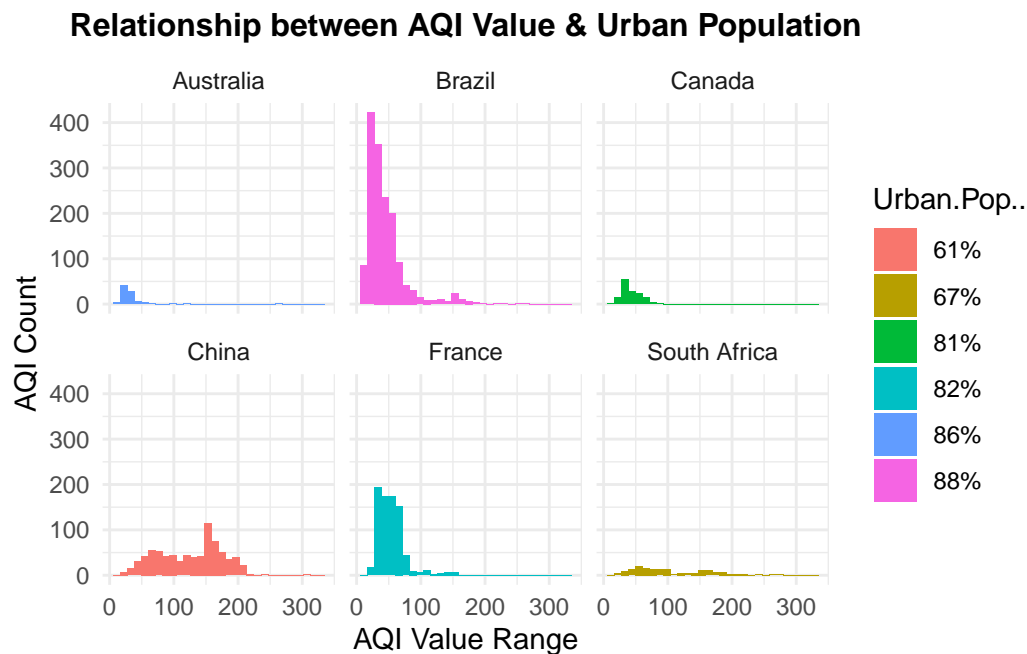




## AQI & Urban Populations

When looking at the relationship between AQI Value and Urban Populations it was observed that the countries with a higher urban population rate had a large number of lower AQI values. This illustrates that there is a potential relationship between the Urban rate of living and pollution. The increased count of low AQI values indicates that the more people that live within the Urban Population, the lower the pollution rate will likely be (Figure 4). The x-axis represents the overall AQI Values gathered, and the y-axis shows the count of each AQI value occurring. Additionally, the visual is faceted into our six randomly generated countries, and the counts displayed are color coded to represent the Urban Populations for each country. As noted above, our team acknowledges that the datasets used have less data representing certain countries. Thus, Australia, Canada and South Africa have less data representing their case, whereas Brazil, China, and France have more data. So in conclusion, while the data may lack full representation of all major cities within each country, the visualization depicts that countries with more relevant urban population have less pollution.

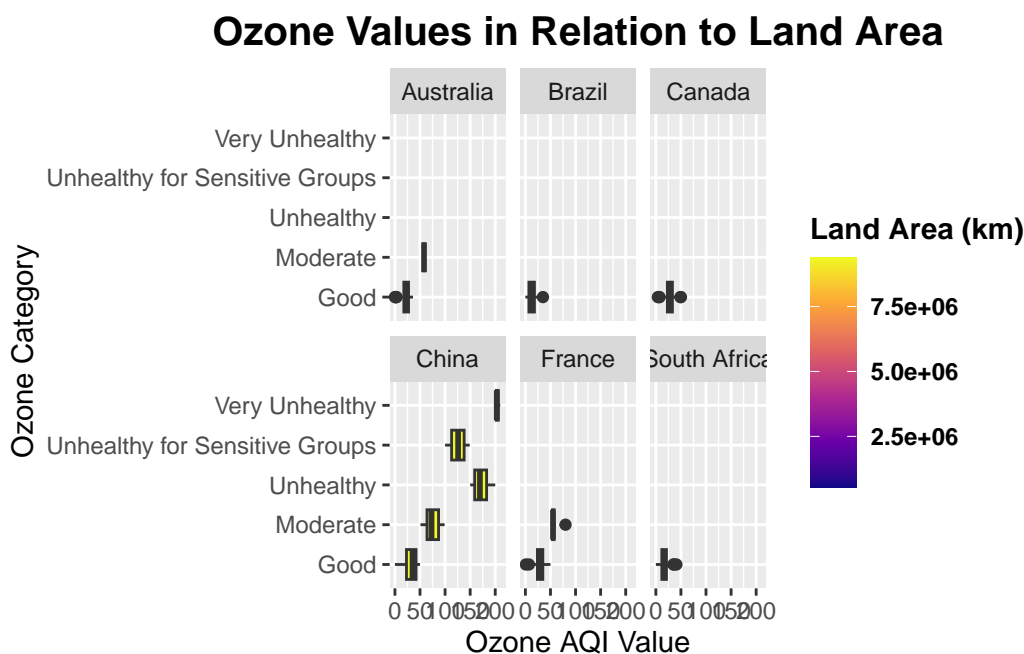
Figure 4



## Will Ozone Quality Be Effected by Land Area?

To further evaluate the relationship between air quality and pollution, it is vital to think about how land area may influence ozone rates. Ozone has an important relationship regarding global warming and pollution. The ozone layer is often talked about in Earth Science courses, on the news, and in any atmospheric/environmental regard. Lately it is important to note that Ozone rates have been on the rise, and this may negatively impact how the lungs function. When intake is high, ozone can cause inflammation and other permanent damage to the lungs (Reyes Becerra, 2025). This relationship between overall air quality and ozone created questions about how certain factors, such as land area, may have effect on ozone and air quality rates. The measured land area includes China, Canada, Brazil, France, and South Africa from highest to lowest, respectively (Figure 5). Most of the data shows that all countries have either moderate or good ozone values, with the exception being China. This indicates that land area has little effect on ozone health, and there may need to be a more specific study conducted to draw further conclusions on the relationship between land area and ozone quality. The data appears to be collected in an observational way leaving way for confounding variables to hinder an overall conclusion for a relationship between ozone and land area. In order to follow F.A.I.R and C.A.R.E conventions, it is appropriate to conclude that a more extensive experiment will need to be done for more definite conclusions.

Figure 5



## Data Sources and Acknowledgement

Fadl, Mohamed. "Countries Population by Year 2020." Kaggle, 19 June 2020, [www.kaggle.com/datasets/eng0mohamed0nabil/population-by-country-2020](https://www.kaggle.com/datasets/eng0mohamed0nabil/population-by-country-2020).

Muzdadid, Hasib Al. "Global Air Pollution Dataset." Kaggle, 8 Nov. 2022, [www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset?resource=download](https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset?resource=download).

Reyes Becerra, Natalia. "Ozone Pollution Is Getting Worse and Increased Wildfires May Be to Blame." *American Lung Association*, American Lung Association, 5 May 2025, [www.lung.org/blog/ozone-pollution-caused-by-wildfires](https://www.lung.org/blog/ozone-pollution-caused-by-wildfires).

"Wheel of Names." Wheel of Names, [wheelofnames.com/](https://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(google sheets4)  
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"
  )
)

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()
)

# 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 faceted scatterplots using ggplot2

ggplot(
  data = tidyAQIdf,
  mapping = aes(
    x = Pollutant_Value,
    y = AQI,
    color = Pollutant,
    group = Pollutant
  )
)+
  geom_point() +
  geom_smooth(
    formula = y ~ x,
    method = "lm",
    color = "black",
    se = FALSE
  ) +
  # geom_abline() +
  facet_wrap(~ Pollutant, scales = "free") +
  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 Count",
      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))

```