Air Pollution Across the Globe

Sneha Arya, Greer Moran, and Gabrielle Smeltzer

2025-04-24

Table of contents

Research Topic: Air Pollution Across the Globe 2

Provenance of Data 2
- **FAIR Principles** 3
  - **Findable** 3
  - **Accessible** 3
  - **Interoperable** 3
  - **Reusable** 3
- **CARE Principles** 3
  - **Collective Benefit** 3
  - **Authority to Control** 3
  - **Responsibility** 4
  - **Ethics** 4

An Analysis of Air Quality 5
- **Cases & Attributes:** 5
  - **Cases:** 5
  - **Attributes:** 5

AQI Values by Country 6

AQI & Urban Populations 7

Will Ozone Quality Be Effected by Land Area? 8

Simple Linear Regression for Pollutant on AQI 9

AQI Value by Each Pollutant 11

Multiple Linear Regression for Pollutant on AQI 13

Data Sources and Acknowledgement 14

Code Appendix: 14

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.

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 each variable's relationship 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.

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. Regarding 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. Our analysis focuses on combining both data sources at the country level to maintain consistency in our visualizations.

To maintain further clarity and avoid overcrowding in our visualizations, 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.

FAIR Principles

Findable

Our data sources are openly available. Researchers can find the original datsets through Kaggle under users Hasib Al Muzdadid and Mohamed Fadl.

Accessible

Both datasets used are publicly accessible without restrictions, which supports the transparency and reproducibility aspects of accessibility. All sources are cited enabling direct access for viewers.

Interoperable

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

Reusable

The datasets have open licensing on Kaggle with a Creative Commons license, indicating long term usability. Other people can also replicate our 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.

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 are used for the intended purposes of education and research.

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 aims to avoid 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.

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:

We chose to focus on the values from each pollutant (CO, 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.

The pollutants can be properly explained as:

- Carbon Monoxide (CO), toxic gas from incomplete combustion.

- Ozone, reactive gas that is a large component in smog.

- Nitrogen Dioxide (NO2), pollutant sourced from burning fuel.

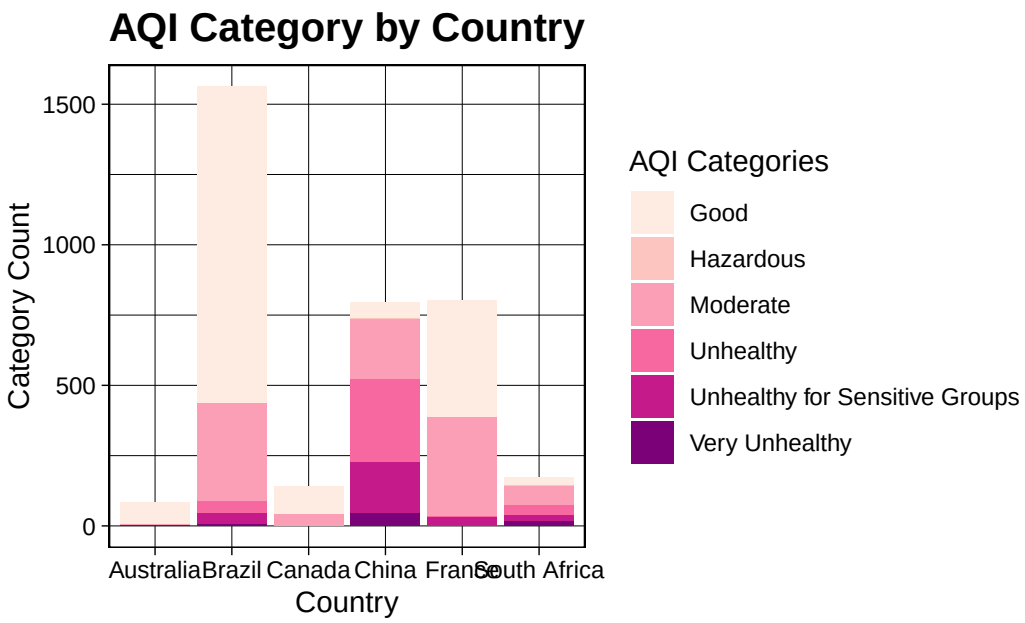
- Particulate Matter with diameter 2.5 micrometers or less (PM), small harmful particles also produced from burning fuel.

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 1, 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 1



A count of each AQI category by country.

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 2). The x-axis represents the overall AQI Values gathered, and the y-axis shows the count of each AQI value occuring. 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.

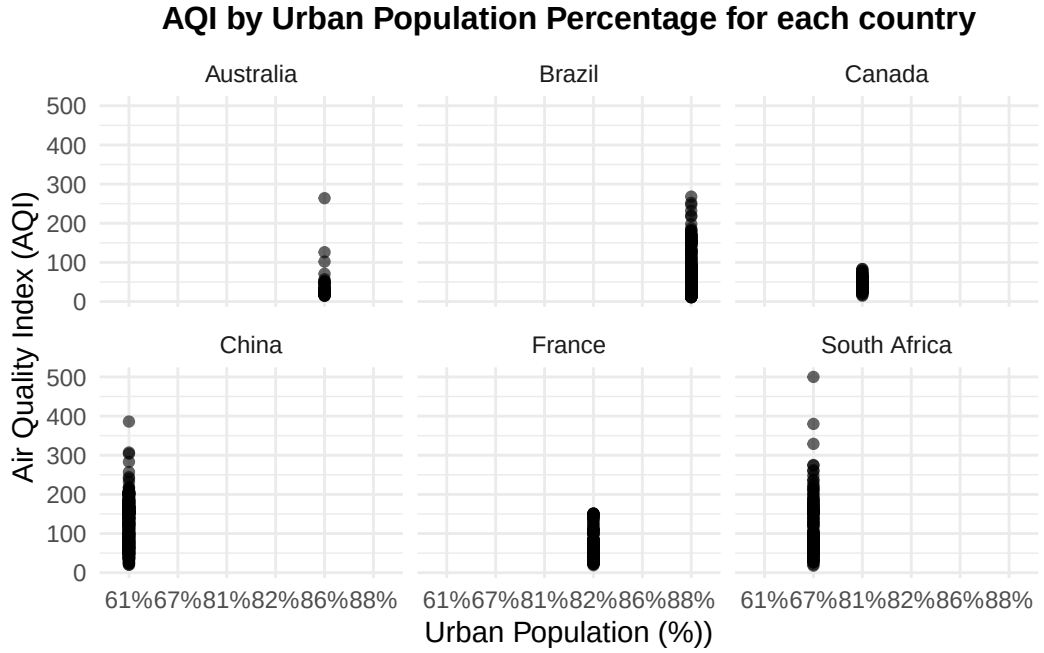
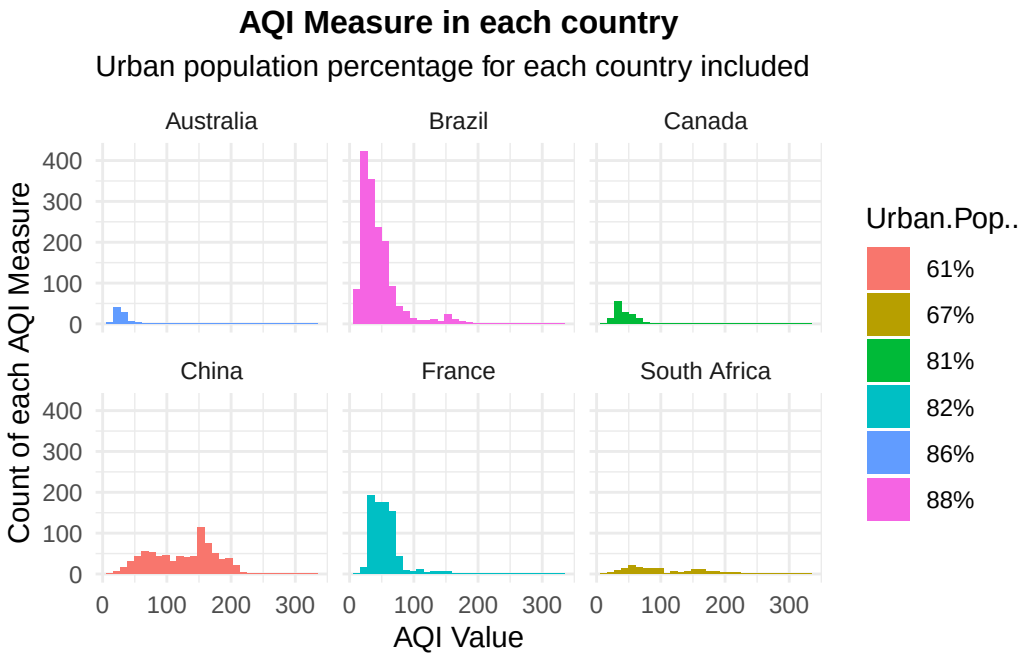


Figure 2



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 3). 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 3

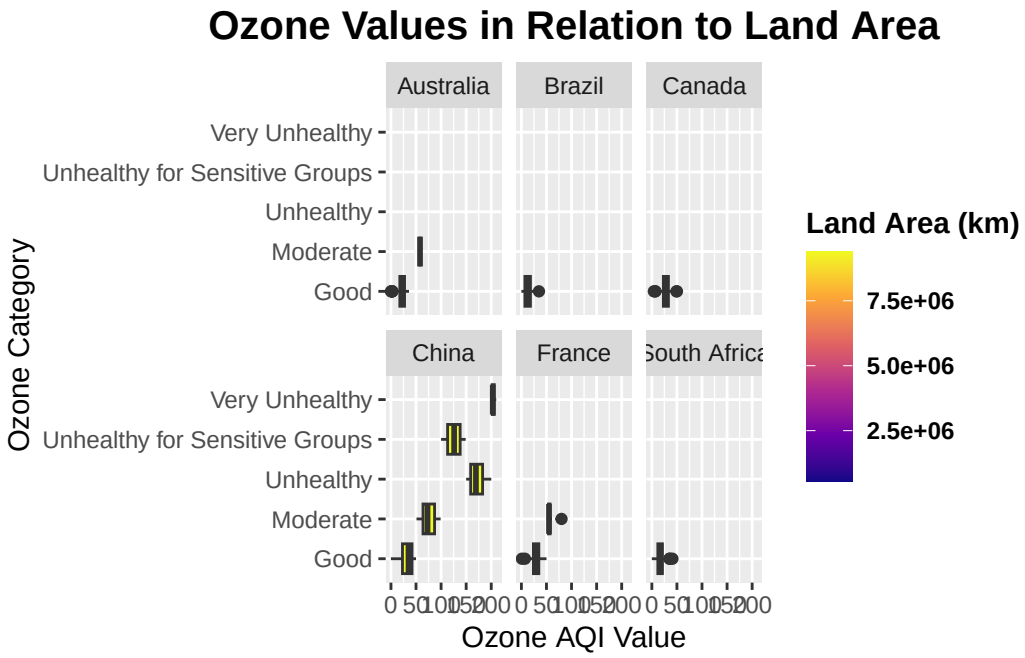


Table 1: Simple Linear Regression Table: Pollutant on AQI

	Pollutant_Name	Coefficient	p_value	R_squared
CO	Carbon Monoxide	13.1750	0	0.1854
Ozone	Ozone	0.8086	0	0.1643
NO2	Nitrogen Dioxide	2.4726	0	0.0537
PM	Particulate Matter	1.0069	0	0.9689

Simple Linear Regression for Pollutant on AQI

Simple Linear Regression Summary of each individual Pollutant on AQI

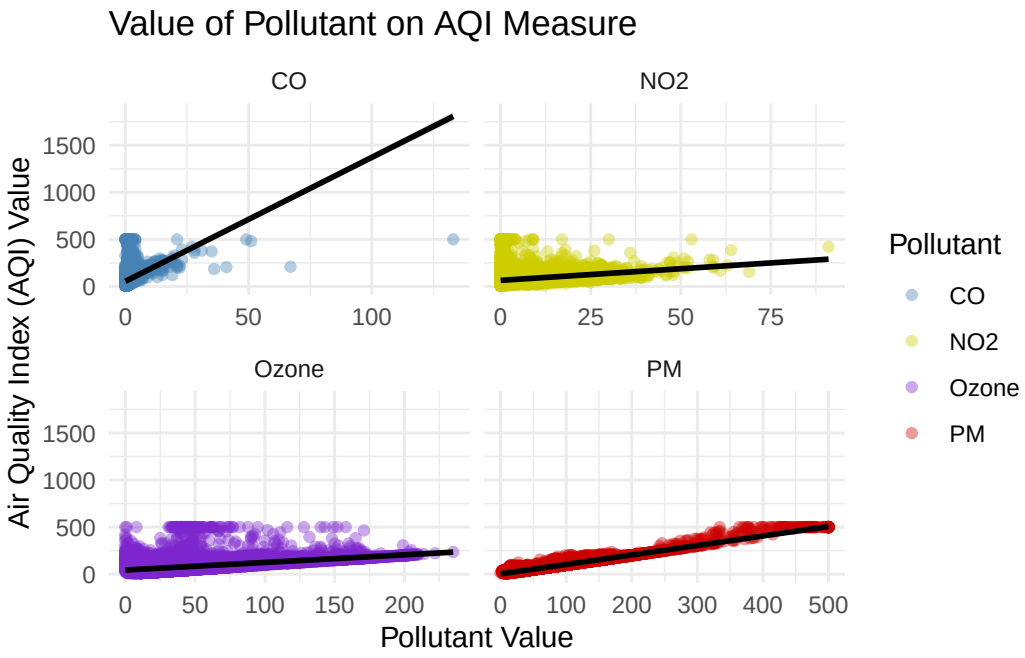
This table, Table 1, provides an overview of the simple linear regression summary each pollutant (CO, Ozone, NO2, and PM) has on the measured AQI value along different cities across the globe. This offers valuable context abut how air quality index (AQI) is measured, how the pollutants influence the strength of AQI on their own. The regression summary includes statistics of estimated coefficient, p-value, and r-squared.

Key Insights:

- From first glance, the strongest relationship in terms of increasing pollutant value is the relationship between Carbon Monoxide (CO) and AQI with a positive magnitude of ~13.175
- However, the most useful aspect of the table on its own is the R-squared value. A near 1 value means that air quality index is best explained by the pollutant one its own. This is true for particulate matter with diameter 2.5 micrometers or less (PM) that has an r-squared of ~0.969

- The r-squared values for the other pollutants are low, near-zero which means they individually have little impact on air quality index. However, since they have p-values much lower than the average significance level of 5% the pollutants could still affect air quality index but not by itself.

Figure 4



AQI Value by Each Pollutant

Figure 4 - Scatterplot of Each Pollutant on AQI Measure

This figure (Figure 4) 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. The regression line in black amounts to the same slope given from the Coefficient column in Table 1.

Key Insights:

- The tight, constant cluster along the relationship between particulate matter with diameter 2.5 micrometers or less (PM) and air quality index (AQI) indicates a strong relationship with little variance. Air quality index consistently increases as particulate matter increases within a strong linear relationship.
- There is strong variation within ozone above the slope line reflecting a weak linear relationship with air quality index. This indicates that because the linear relationship is not strong, their relationship is not a two-way streak and can even underestimate air quality index if based on only ozone. Further, this is not a good predictor of AQI on it's own.
- Most carbon monoxide (CO) values are small, hence the cluster in the bottom left-hand corner of the CO facet. Countries within this data set have low carbon monoxide values, making a linear regression model not ideal. However, despite a low value the regression line appears close to 1. This indicates that an overall relationship is still present in a minor way, with non-constant error variance held in concern.

- Nitrogen dioxide (NO2) has a near zero regression line. Further, there is little impact on the measured air quality index. Additionally, a cone-shaped spread of the glyphs representing each country would indicate a violation of the equal variance (heteroscedasticity). In this case, nitrogen dioxide is not properly predictive of air quality index until nitrogen dioxide is measured at a higher value.

Table 2: Linear Regression Summary for Pollutants impact on AQI

	Pollutant	Regr. Coefficient	p-value
2	CO	0.01707	0.66715
3	Ozone	0.15751	0.00000
4	NO2	-0.03620	0.00731
5	PM	0.98014	0.00000

Multiple Linear Regression for Pollutant on AQI

This table, Table 2, displays the coefficients and p-values for each pollutant's impact on air quality index with the other pollutants held constant. While the model does individually evaluate the pollutant's individual impact on air quality measure, this multiple regression model uses a new intercept gathered from factoring in all the other pollutants. The new intercept causes the slope/regression coefficient to change from its simple linear regression value in order to account for all of the new attributes included.

Key Insights:

- Carbon monoxide (CO), now has a p-value that is statistically insignificant at a 5% significance level. While we are not running a hypothesis test this is important to note. Since further surveying the data from different cities in different countries, this leads us to observe that with other pollutants factored in, air quality index cannot be properly predicted by carbon monoxide.
- The most influential pollutant impacting air quality index given multiple pollutants is particulate matter with diameter 2.5 micrometers or less (PM), which has a regression coefficient of 0.980

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.

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.

"Wheel of Names." Wheel of Names, www.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)


## 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

CountryAQI <- 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)
  )

CountryAQI

# HI GABI THIS WAS THE EXAMPLE CODE FOR IF IT WAS PARSED NUMERIC, FEEL FREE TO DELETE

# TO SEE NUMERIC, MAKE NEW CODE AT LINE 352
## (or just inbetween tidypopPoll and specificCountry in your first code chunk)
### do: tidypopPoll$Urban.Pop.. <- parse_number(tidypopPoll$Urban.Pop..)

### I like the other visual u have tho, I'm using it for the readMe. If you make both, just ch
specificCountry %>%
  ggplot(
   aes(
     x = Urban.Pop..,
     y = AQI.Value
     )
   ) +
```

```
  geom_point(alpha = 0.6, fill = "darkgreen") +
  facet_wrap(~ Country) +
  labs(
    x = "Urban Population (%))",
    y = "Air Quality Index (AQI)",
    fill = "Urban Population (%)",
    title = "AQI by Urban Population Percentage for each country"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12L,
    face = "bold",
    hjust = 0.5)
  )

UrbanPopulationAQI <- 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",
    y = "Count of each AQI Measure",
    title = "AQI Measure in each country",
    subtitle = "Urban population percentage for each country included"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12L,
    face = "bold",
    hjust = 0.5)
  ) +
  facet_wrap(vars(Country))

UrbanPopulationAQI
OzoneLandAreaFig <- 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))

OzoneLandAreaFig

#| 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)


model_CO <- lm(AQI ~ CO, data = cleanAQIdata)
co_coeff <- coef(model_CO)[2]
co_pValue <- summary(model_CO)$coefficients[2,4]
co_r_squared <- summary(model_CO)$r.squared
```

```r
model_ozone <- lm(AQI ~ Ozone, data = cleanAQIdata)
ozone_coeff <- coef(model_ozone)[2]
ozone_pValue <- summary(model_ozone)$coefficients[2,4]
ozone_r_squared <- summary(model_ozone)$r.squared

model_NO2 <- lm(AQI ~ NO2, data = cleanAQIdata)
no2_coeff <- coef(model_NO2)[2]
no2_pValue <- summary(model_NO2)$coefficients[2,4]
no2_r_squared <- summary(model_NO2)$r.squared

model_PM <- lm(AQI ~ PM, data = cleanAQIdata)
pm_coeff <- coef(model_PM)[2]
pm_pValue <- summary(model_PM)$coefficients[2,4]
pm_r_squared <- summary(model_PM)$r.squared

# Create table

simple_regr_table <-data.frame(
  Pollutant_Name = c("Carbon Monoxide", "Ozone", "Nitrogen Dioxide", "Particulate Matter"),
  Coefficient = c(co_coeff, ozone_coeff, no2_coeff, pm_coeff),
  p_value = c(co_pValue, ozone_pValue, no2_pValue, pm_pValue),
  R_squared = c(co_r_squared, ozone_r_squared, no2_r_squared, pm_r_squared)
)

kable(
  simple_regr_table,
  digits = 4,
  format = "latex") %>%
  kable_styling(position = "center")

# 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
```

```r
PollutantLMPlot <- ggplot(
  data = tidyAQIdf,
  mapping = aes(
    x = Pollutant_Value,
    y = AQI,
    color = Pollutant
    )
  )+
    geom_point(alpha = 0.4) +
   geom_smooth(
     method = "lm",
     color = "black",
     se = TRUE
     ) +
    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()

PollutantLMPlot
# Create Table Summary Visualization ---

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

aqiModel <- aqiModel[-1, ]

kable(
  aqiModel,
  digits = 5,
  format = "latex"
  ) %>%
  kable_styling(
    position = "center"
  )



ggsave("CountryAQI.png", plot = CountryAQI, width = 6, height = 4, dpi = 300)
```

```r
ggsave("UrbanPopulationAQI.png", plot = UrbanPopulationAQI, width = 6, height = 4, dpi = 300)
ggsave("OzoneLandAreaFig.png", plot = OzoneLandAreaFig, width = 6, height = 4, dpi = 300)
ggsave("PollutantLMPlot.png", plot = PollutantLMPlot, width = 6, height = 4, dpi = 300)
```