Global Air Pollution Data Analysis

Udbhaav Mudgil

2025-10-22

Introduction

This report provides an exploratory analysis of a **global air pollution dataset**.
We will use R to examine trends in Air Quality Index (AQI) across countries and regions, visualize relationships between pollutants and AQI, and generate a world map displaying average AQI by country.

The analysis uses several R libraries such as **tidyverse**, **ggplot2**, **sf**, and **rnaturalearth** for data manipulation, visualization, and geospatial mapping.

Load and Inspect the Dataset

We start by loading the cleaned global air pollution dataset using `read_csv()`.
Next, we explore its structure, summary statistics, and check for missing values to understand data quality.

```
pollution <- read_csv("global_air_pollution_dataset_cleaned.csv")
```

```
## Rows: 23035 Columns: 15
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (9): Country, Region, City, AQI Category, CO AQI Category, Ozone AQI Cat...
## dbl (6): AQI Value, AQI Category Score, CO AQI Value, Ozone AQI Value, NO2 A...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(pollution)
```

```
## Rows: 23,035
## Columns: 15
## $ Country              <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afg…
## $ Region               <chr> "Asia", "Asia", "Asia", "Asia", "Asia", "Asia", "…
## $ City                 <chr> "Kuhestan", "Qunduz", "Rostaq", "Tokzar", "Carika…
## $ `AQI Value`          <dbl> 151, 117, 113, 77, 67, 57, 83, 72, 104, 99, 84, 1…
## $ `AQI Category`       <chr> "Unhealthy", "Unhealthy for Sensitive Groups", "U…
## $ `AQI Category Score` <dbl> 4, 3, 3, 2, 2, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 3, 2…
## $ `CO AQI Value`       <dbl> 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0…
## $ `CO AQI Category`    <chr> "Good", "Good", "Good", "Good", "Good", "Good", "…
## $ `Ozone AQI Value`    <dbl> 41, 44, 42, 40, 37, 38, 41, 44, 34, 49, 64, 29, 4…
## $ `Ozone AQI Category` <chr> "Good", "Good", "Good", "Good", "Good", "Good", "…
## $ `NO2 AQI Value`      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ `NO2 AQI Category`   <chr> "Good", "Good", "Good", "Good", "Good", "Good", "…
## $ `PM2.5 AQI Value`    <dbl> 151, 117, 113, 77, 67, 57, 83, 72, 104, 99, 84, 1…
## $ `PM2.5 AQI Category` <chr> "Unhealthy", "Unhealthy for Sensitive Groups", "U…
## $ `Primary Pollutant`  <chr> "PM2.5", "PM2.5", "PM2.5", "PM2.5", "PM2.5", "PM2…
```

```
summary(pollution)
```

```
##    Country             Region              City             AQI Value
##  Length:23035       Length:23035       Length:23035       Min.   :  6.00
##  Class :character   Class :character   Class :character   1st Qu.: 39.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 55.00
##                                                           Mean   : 72.34
##                                                           3rd Qu.: 80.00
##                                                           Max.   :500.00
##  AQI Category       AQI Category Score  CO AQI Value     CO AQI Category
##  Length:23035       Min.   :1.00       Min.   :  0.000   Length:23035
##  Class :character   1st Qu.:1.00       1st Qu.:  1.000   Class :character
##  Mode  :character   Median :2.00       Median :  1.000   Mode  :character
##                     Mean   :1.91       Mean   :  1.376
##                     3rd Qu.:2.00       3rd Qu.:  1.000
##                     Max.   :6.00       Max.   :133.000
##  Ozone AQI Value  Ozone AQI Category NO2 AQI Value    NO2 AQI Category
##  Min.   :  0.00   Length:23035       Min.   : 0.000   Length:23035
##  1st Qu.: 21.00   Class :character   1st Qu.: 0.000   Class :character
##  Median : 31.00   Mode  :character   Median : 1.000   Mode  :character
##  Mean   : 35.23                      Mean   : 3.085
##  3rd Qu.: 40.00                      3rd Qu.: 4.000
##  Max.   :235.00                      Max.   :91.000
##  PM2.5 AQI Value  PM2.5 AQI Category Primary Pollutant
##  Min.   :  0.00   Length:23035       Length:23035
##  1st Qu.: 35.00   Class :character   Class :character
##  Median : 54.00   Mode  :character   Mode  :character
##  Mean   : 68.88
##  3rd Qu.: 79.00
##  Max.   :500.00
```

```
colSums(is.na(pollution))
```

```
##            Country             Region               City          AQI Value
##                  0                  0                  0                  0
##       AQI Category AQI Category Score       CO AQI Value    CO AQI Category
##                  0                  0                  0                  0
##    Ozone AQI Value Ozone AQI Category      NO2 AQI Value   NO2 AQI Category
##                  0                  0                  0                  0
##    PM2.5 AQI Value PM2.5 AQI Category  Primary Pollutant
##                  0                  0                  0
```

Data Preprocessing

We will start data preprocessing with converting Region, Country, City, and AQI Category to factors as this ensures that R treats them as categorical variables, which is essential for grouping operations, summaries, and color-coded plots. This step also helps to avoid errors in aggregation functions that require proper factor levels for grouping.

```
pollution <- pollution %>%
mutate(
Region = as.factor(Region),
Country = as.factor(Country),
City = as.factor(City),
`AQI Category` = as.factor(`AQI Category`)
)
```

Exploratory Analysis

Average AQI by Country

We calculate the average AQI for each country by using `group_by()` and `summarise()` functions. This provides a high-level view of air quality at the national level and allows comparisons across countries.

Key Insights:

- Countries like Bahrain and Pakistan have the highest average AQI.
- Countries, such as New Zealand and Australia, have low average AQI.

```
aqi_country <- pollution %>%
group_by(Country) %>%
summarise(Avg_AQI = mean(`AQI Value`, na.rm = TRUE))
print(aqi_country)
```

```
## # A tibble: 174 × 2
##    Country     Avg_AQI
##    <fct>         <dbl>
##  1 Afghanistan    96.0
##  2 Albania        68.2
##  3 Algeria        88.2
##  4 Andorra        29.3
##  5 Angola         83.9
##  6 Argentina      28.2
##  7 Armenia        53.6
##  8 Aruba         163
##  9 Australia      33.6
## 10 Austria        53.7
## # ℹ 164 more rows
```

Average AQI by Region

Aggregating AQI by region helps identify which geographic areas experience the worst air quality. Sorting by descending average AQI highlights the most affected regions.

Key Insights:

- Asia and Africa have the highest regional AQI averages.
- South America and Oceania have significantly lower average AQI, indicating cleaner air.

```
mean_aqi_region <- pollution %>%
group_by(Region) %>%
summarise(Avg_AQI = mean(`AQI Value`, na.rm = TRUE)) %>%
arrange(desc(Avg_AQI))
print(mean_aqi_region)
```

```
## # A tibble: 6 × 2
##   Region        Avg_AQI
##   <fct>           <dbl>
## 1 Asia           114.  
## 2 Africa          73.2
## 3 North America   65.3
## 4 Europe          49.4
## 5 South America   48.2
## 6 Oceania         31.8
```

Top 15 Most Polluted Cities

We sort cities by their AQI values to identify urban pollution hotspots. This helps pinpoint specific cities that may need urgent intervention.

Key Insights:

- Most top polluted cities are in India.
- Urban density, industrial activity, and traffic contribute significantly to high AQI.

```
top15_cities <- pollution %>%
arrange(desc(`AQI Value`)) %>%
select(Country, City, `AQI Value`, `AQI Category`) %>%
head(15)
print(top15_cities)
```

```
## # A tibble: 15 × 4
##    Country City         `AQI Value` `AQI Category`
##    <fct>   <fct>              <dbl> <fct>
##  1 India   Rania                500 Hazardous
##  2 India   Gohana               500 Hazardous
##  3 India   Gunnaur              500 Hazardous
##  4 India   Khetri               500 Hazardous
##  5 India   Jahangirpur          500 Hazardous
##  6 India   Kakrala              500 Hazardous
##  7 India   Kandhla              500 Hazardous
##  8 India   Mahendragarh         500 Hazardous
##  9 India   Gajraula             500 Hazardous
## 10 India   Nagaur               500 Hazardous
## 11 India   Dataganj             500 Hazardous
## 12 India   Pilkhuwa             500 Hazardous
## 13 India   Siwani               500 Hazardous
## 14 India   Shamsabad            500 Hazardous
## 15 India   Phalodi              500 Hazardous
```

Visualizations

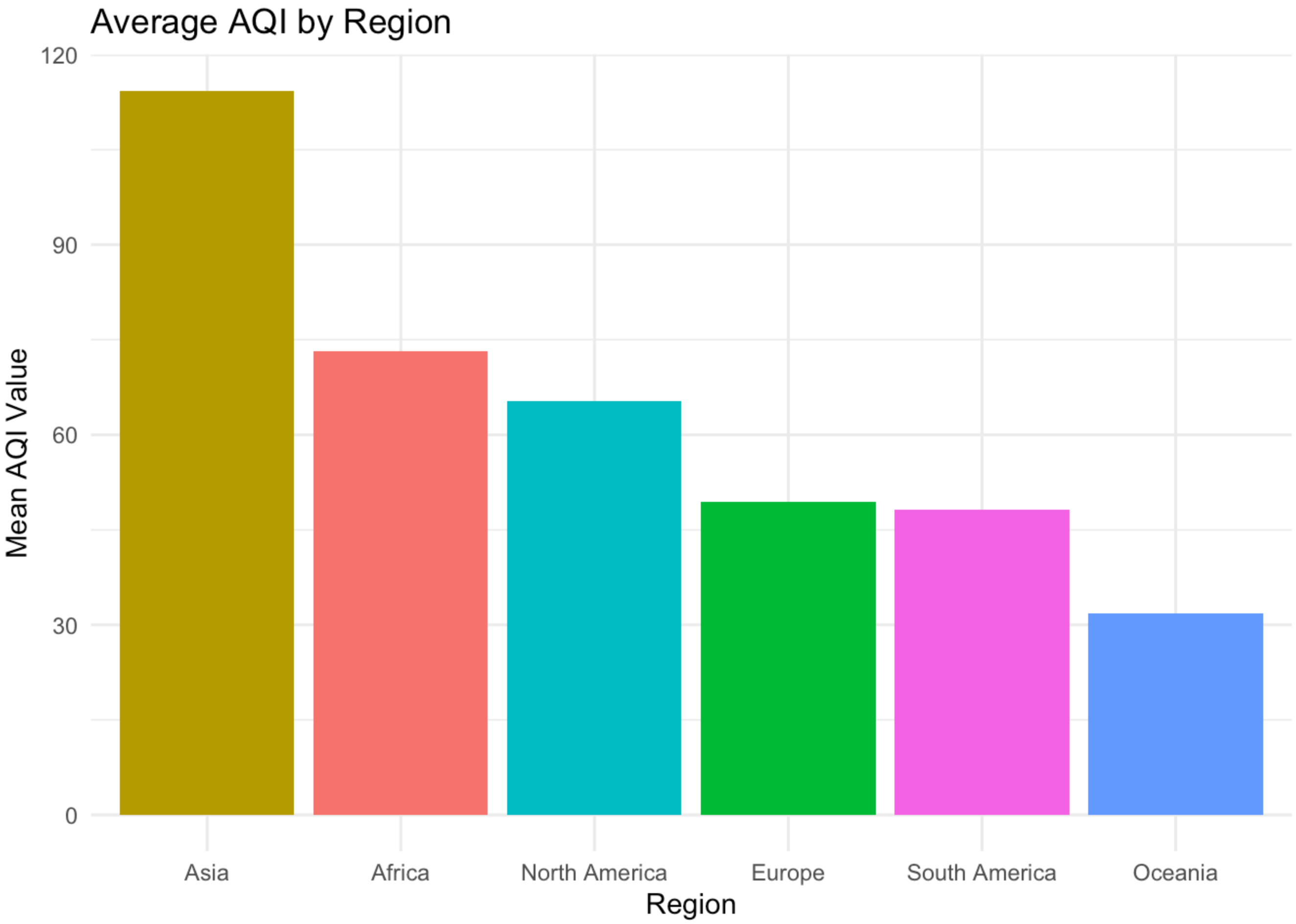
Average AQI by Region

We will use a bar chart to visualize average AQI by region. Reordering bars by descending AQI makes it easier to compare regions visually.

Key Insights:

- Asia shows the highest average AQI, followed by Africa.
- Oceania and South America appear cleaner in comparison.

```
ggplot(mean_aqi_region, aes(x = reorder(Region, -Avg_AQI), y = Avg_AQI, fill = Region)) +
geom_col(show.legend = FALSE) +
labs(
title = "Average AQI by Region",
x = "Region",
y = "Mean AQI Value"
) +
theme_minimal()
```



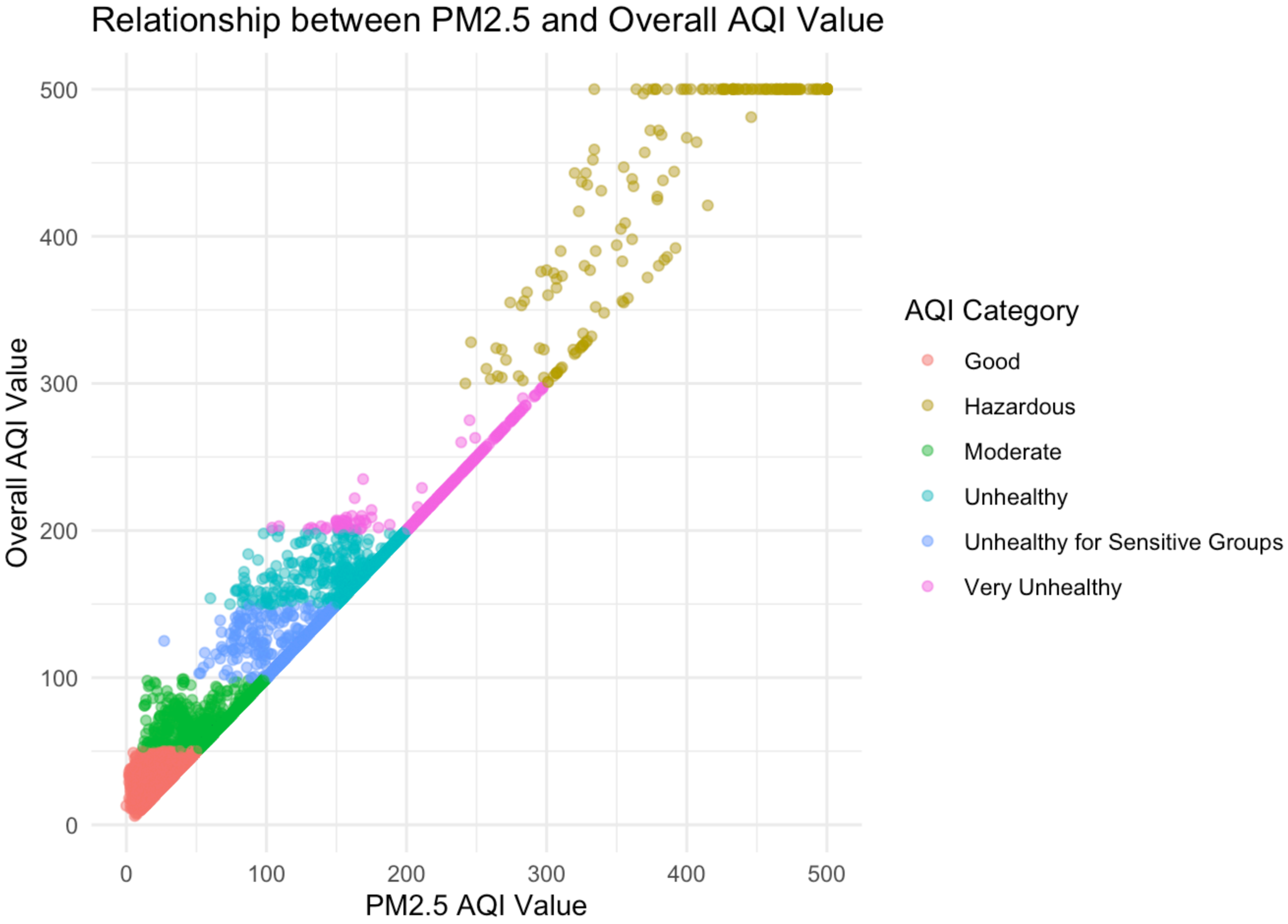
Relationship between PM2.5 and Overall AQI

We will use a scatter plot to examine the relationship between PM2.5 AQI values and overall AQI. Points are colored by AQI category to highlight pollution severity.

Key Insights:

- PM2.5 strongly correlates with overall AQI.
- Cities with higher PM2.5 are often in the “Hazardous” or “Very Unhealthy” AQI categories.

```
ggplot(pollution, aes(x = `PM2.5 AQI Value`, y = `AQI Value`, color = `AQI Category`)) +
geom_point(alpha = 0.5) +
labs(
title = "Relationship between PM2.5 and Overall AQI Value",
x = "PM2.5 AQI Value",
y = "Overall AQI Value"
) +
theme_minimal()
```



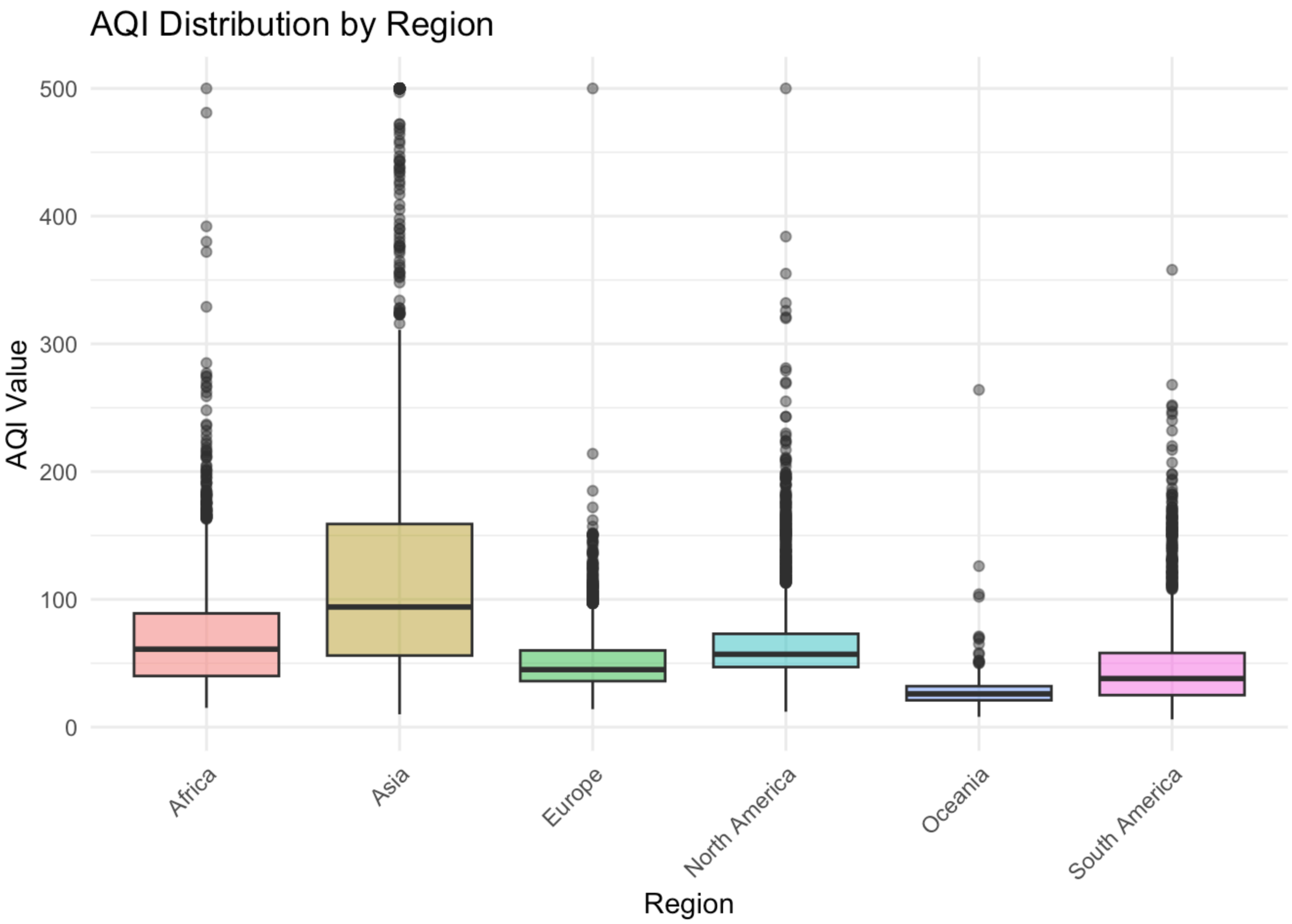
AQI Distribution by Region

We will use boxplots show median, quartiles, and outliers for AQI within each region, providing insights into variability and extreme values in each region.

Key Insights:

- Asia and Africa have both high medians whereas Africa and North America have large variability in AQI.
- Apart from South America and Ocenia all other regions have at least 1 city with 500 AQI.
- Europe and Oceania show tighter distributions with lower median AQI.

```
ggplot(pollution, aes(x = Region, y = `AQI Value`, fill = Region)) +
geom_boxplot(show.legend = FALSE, alpha = 0.5) +
labs(
title = "AQI Distribution by Region",
x = "Region",
y = "AQI Value"
) +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



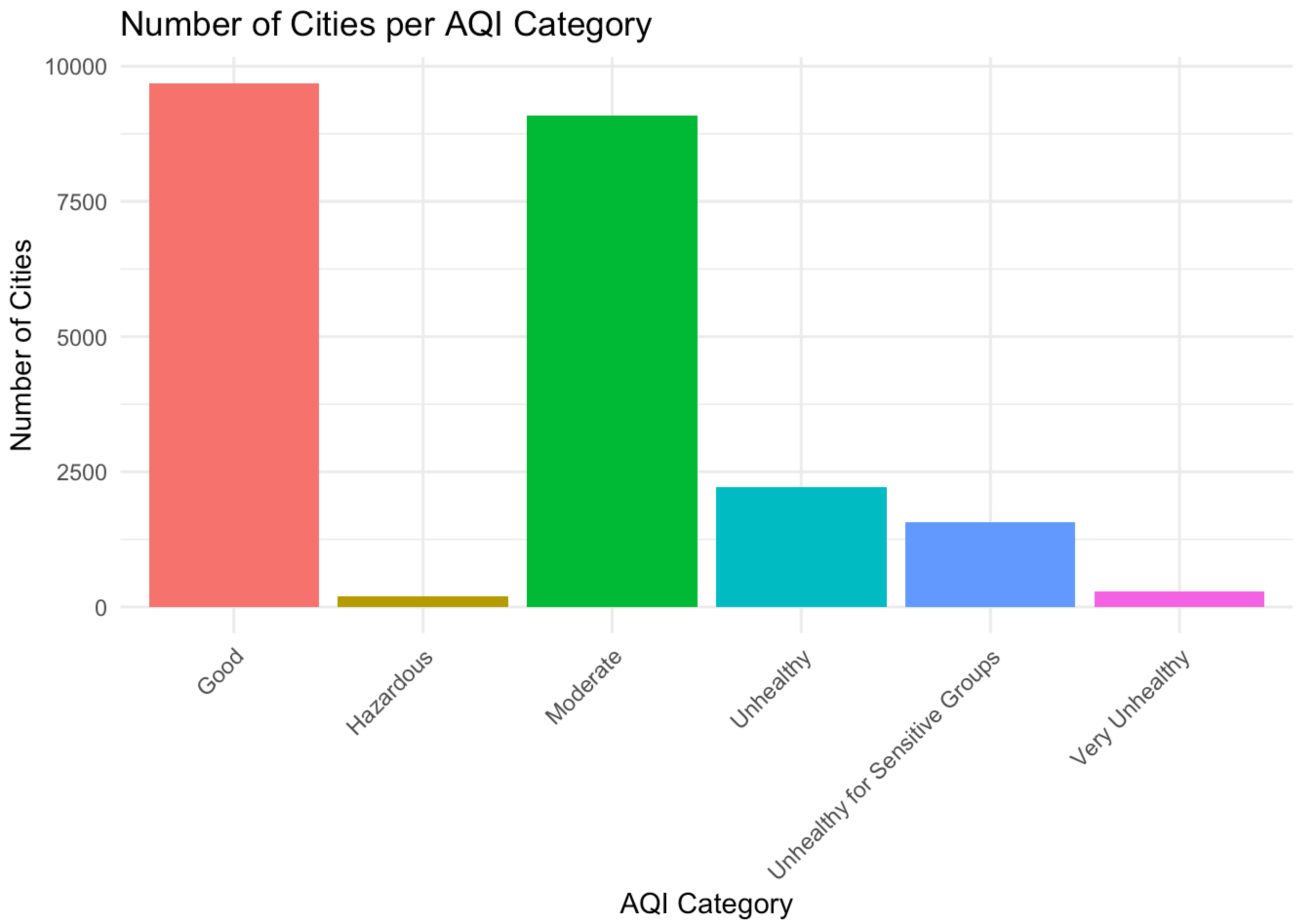
Number of Cities per AQI Category

This bar chart shows how many cities fall into each AQI category, indicating global exposure to different pollution levels.

Key Insights:

- Most cities fall into Good or Moderate for Sensitive Groups categories.
- Few cities are in “Very Unhealthy” or “Hazardous” extremes.

```
pollution %>%
group_by(`AQI Category`) %>%
summarise(City_Count = n_distinct(City)) %>%
ggplot(aes(x = `AQI Category`, y = City_Count, fill = `AQI Category`)) +
geom_col(show.legend = FALSE) +
labs(
title = "Number of Cities per AQI Category",
x = "AQI Category",
y = "Number of Cities"
) +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Global Air Pollution Map

Load World Map Data

We load country boundary data using `rnaturalearth` to create a base map for plotting AQI values geographically.

```
world <- ne_countries(scale = "medium", returnclass = "sf")
```

Harmonize Country Names

We use mutate function to convert country names in AQI dataset so that they align with the map data to ensure correct merging and avoid missing matches.

```
aqi_country <- aqi_country %>%
mutate(Country = case_when(
Country == "Bolivia (Plurinational State of)" ~ "Bolivia",
Country == "Bosnia and Herzegovina" ~ "Bosnia and Herz.",
Country == "Central African Republic" ~ "Central African Rep.",
Country == "Democratic Republic of the Congo" ~ "Dem. Rep. Congo",
Country == "Dominican Republic" ~ "Dominican Rep.",
Country == "Equatorial Guinea" ~ "Eq. Guinea",
Country == "Iran (Islamic Republic of)" ~ "Iran",
Country == "Kingdom of Eswatini" ~ "eSwatini",
Country == "Lao People's Democratic Republic" ~ "Laos",
Country == "Republic of Korea" ~ "South Korea",
Country == "Republic of Moldova" ~ "Moldova",
Country == "Republic of North Macedonia" ~ "North Macedonia",
Country == "Russian Federation" ~ "Russia",
Country == "Saint Kitts and Nevis" ~ "St. Kitts and Nevis",
Country == "Solomon Islands" ~ "Solomon Is.",
Country == "South Sudan" ~ "S. Sudan",
Country == "Syrian Arab Republic" ~ "Syria",
Country == "United Kingdom of Great Britain and Northern Ireland" ~ "United Kingdom",
Country == "United Republic of Tanzania" ~ "Tanzania",
Country == "Venezuela (Bolivarian Republic of)" ~ "Venezuela",
TRUE ~ Country
))
```

Merge AQI Data with Map

Merging AQI data with map polygons allows us to create a choropleth map showing global air pollution.

```
world_aqi <- left_join(world, aqi_country, by = c("name" = "Country"))
```

Plot World Map of Average AQI

We generate a choropleth map by using `geom_sf()` function, with color gradients representing average AQI levels. Darker reds indicate higher pollution, while light blues indicate cleaner air.

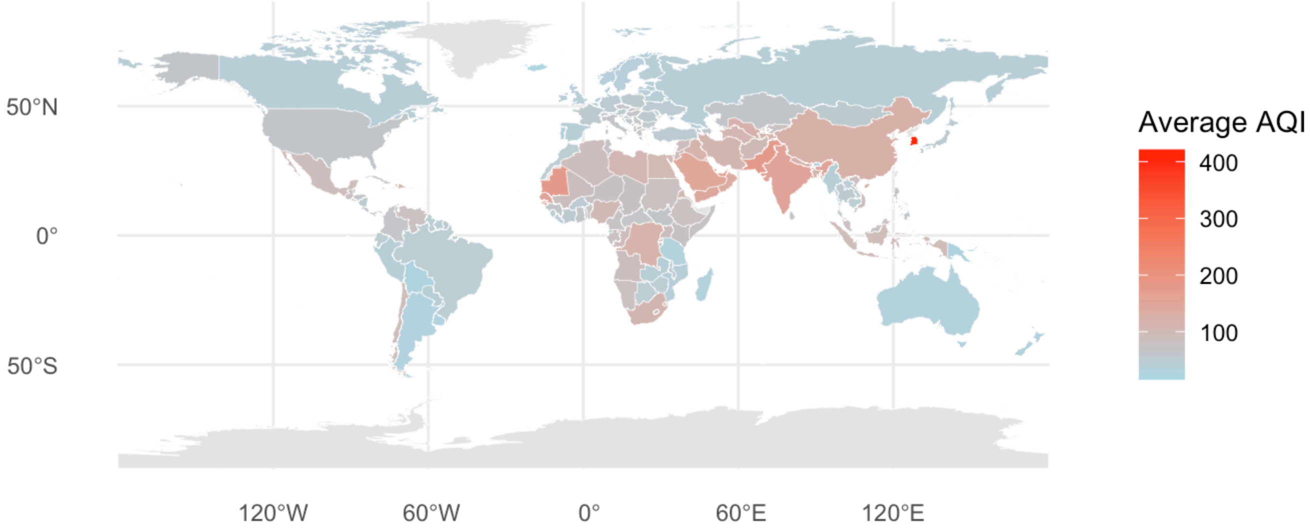
Key Insights:

- South Asia, the Middle East, and parts of Africa show the worst air quality.
- North America, South America, Oceania, and parts of Europe have relatively cleaner air.

```
ggplot(data = world_aqi) +
geom_sf(aes(fill = Avg_AQI), color = "white", size = 0.1) +
scale_fill_gradient(
name = "Average AQI",
low = "lightblue",
high = "red",
na.value = "grey90"
) +
labs(
title = "Average Air Quality Index (AQI) by Country",
subtitle = "Darker red indicates higher pollution levels",
caption = "Source: Global Air Pollution Dataset"
) +
theme_minimal()
```

Average Air Quality Index (AQI) by Country

Darker red indicates higher pollution levels



Source: Global Air Pollution Dataset