Air Pollution Report

Nayananshu Garai

2025-08-09

```
# This chunk sets up the global options for the R Markdown document.
# include=FALSE prevents the code and its output from appearing in the final report.

# Load necessary libraries
# Ensure these packages are installed using install.packages("package_name")
library(knitr)      # For creating dynamic reports
library(readr)      # For fast and friendly CSV reading
library(dplyr)      # For data manipulation and transformation
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)    # For creating elegant and complex plots
library(lubridate)  # For working with dates and times
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(corrplot)   # For visualizing correlation matrices
```

```
## corrplot 0.95 loaded
```

```
library(tidyr)      # For tidying data

# Set default chunk options
opts_chunk$set(
  echo = TRUE,       # Display the code chunks in the output
  warning = FALSE,   # Suppress warnings
  message = FALSE,   # Suppress messages
  fig.align = "center" # Center-align figures
)
```

2. Data Loading and Preparation

First, we load the main dataset, which contains the combined air quality information for all cities. We will then proceed with cleaning and preprocessing steps.

```
# Load the main dataset
# We use the consolidated file as it contains data for all cities.
air_quality <- read_csv("Dataset/Air_Quality.csv")

# Display the first few rows and the structure of the data
head(air_quality)

## # A tibble: 6 × 10
##   Date                City        CO   CO2   NO2   SO2    O3 PM2.5  PM10   AQI
##   <dttm>              <chr>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 2024-01-01 00:00:00 Brasilia   323    NA  23.8   2.8    42  12    17.1  16.8
## 2 2024-01-01 01:00:00 Brasilia   318    NA  21.9   2.7    40  12.5  17.9  16
## 3 2024-01-01 02:00:00 Brasilia   309    NA  19.2   2.6    39  12.1  17.3  15.6
## 4 2024-01-01 03:00:00 Brasilia   295    NA  16.3   2.4    38  11.4  16.2  15.2
## 5 2024-01-01 04:00:00 Brasilia   270    NA  13     2.1    40  10.2  14.6  16
## 6 2024-01-01 05:00:00 Brasilia   239    NA   9.4   1.9    44   8.7  12.4  17.6

str(air_quality)

## spc_tbl_ [52,704 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Date : POSIXct[1:52704], format: "2024-01-01 00:00:00" "2024-01-01 01:00:00" ...
##  $ City : chr [1:52704] "Brasilia" "Brasilia" "Brasilia" "Brasilia" ...
##  $ CO   : num [1:52704] 323 318 309 295 270 239 215 205 201 199 ...
##  $ CO2  : num [1:52704] NA NA NA NA NA NA NA NA NA NA ...
##  $ NO2  : num [1:52704] 23.8 21.9 19.2 16.3 13 9.4 6.8 6 6.1 5.9 ...
##  $ SO2  : num [1:52704] 2.8 2.7 2.6 2.4 2.1 1.9 1.7 1.8 2.1 2.2 ...
##  $ O3   : num [1:52704] 42 40 39 38 40 44 47 46 45 46 ...
##  $ PM2.5: num [1:52704] 12 12.5 12.1 11.4 10.2 8.7 7.5 6.1 5.7 5.7 ...
##  $ PM10 : num [1:52704] 17.1 17.9 17.3 16.2 14.6 12.4 10.7 8.7 8.2 8.2 ...
##  $ AQI  : num [1:52704] 16.8 16 15.6 15.2 16 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Date = col_datetime(format = ""),
##   ..   City = col_character(),
##   ..   CO = col_double(),
##   ..   CO2 = col_double(),
##   ..   NO2 = col_double(),
##   ..   SO2 = col_double(),
##   ..   O3 = col_double(),
##   ..   PM2.5 = col_double(),
##   ..   PM10 = col_double(),
##   ..   AQI = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

2.1. Data Cleaning

The initial data inspection reveals a few areas for cleanup: - The `Date` column should be converted to a proper datetime format. - The `CO2` column contains many missing values (`NA`), which we need to address. - We will check for any other missing values across the dataset.

```
# Convert 'Date' column to datetime objects using lubridate
air_quality$Date <- ymd_hms(air_quality$Date)

# Check for missing values in each column
missing_values <- colSums(is.na(air_quality))
print("Missing values per column:")

## [1] "Missing values per column:"

print(missing_values)

##  Date  City    CO   CO2   NO2   SO2    O3 PM2.5  PM10   AQI
##  2196     0     0 43056     0     0     0     0     0     0

# The CO2 column has a significant number of NAs. For this analysis,
# we will exclude it from correlation and some plots, but keep it for now.
# For other columns with few NAs, we can choose to omit them for simplicity.
air_quality_clean <- air_quality %>%
  na.omit() # Omitting rows with NA for robust analysis

# Verify the structure of the cleaned data
str(air_quality_clean)

## tibble [9,246 × 10] (S3: tbl_df/tbl/data.frame)
##  $ Date : POSIXct[1:9246], format: "2024-10-26 01:00:00" "2024-10-26 02:00:00" ...
##  $ City : chr [1:9246] "Brasilia" "Brasilia" "Brasilia" "Brasilia" ...
##  $ CO   : num [1:9246] 918 851 772 669 554 469 438 438 441 445 ...
##  $ CO2  : num [1:9246] 471 472 472 472 472 472 475 478 478 471 ...
##  $ NO2  : num [1:9246] 24.4 23.7 22.3 19.5 16 13.2 12.2 12 11.1 8.7 ...
##  $ SO2  : num [1:9246] 2.5 2.7 2.8 2.9 2.9 2.9 2.9 2.9 2.8 2.6 ...
##  $ O3   : num [1:9246] 35 32 31 32 34 36 34 31 35 54 ...
##  $ PM2.5: num [1:9246] 14.9 15.1 15 14.8 14.8 14.8 14.8 14.7 15.2 15.7 ...
##  $ PM10 : num [1:9246] 21.4 21.6 21.4 21.1 21.1 21.1 21.1 21 21.7 22.4 ...
##  $ AQI  : num [1:9246] 27 27.1 27.2 27.3 27.3 ...
##  - attr(*, "na.action")= 'omit' Named int [1:43458] 1 2 3 4 5 6 7 8 9 10 ...
##   ..- attr(*, "names")= chr [1:43458] "1" "2" "3" "4" ...
```

.

3. Exploratory Data Analysis (EDA)

With the data cleaned, we can now explore it to uncover insights.

3.1. Summary Statistics

Let’s start by calculating summary statistics for the key numerical columns, grouped by city. This gives us a high-level overview of the pollution levels in each location.

```r
# Calculate summary statistics for pollutants and AQI by city
summary_by_city <- air_quality_clean %>%
 group_by(City) %>%
 summarise(
  Avg_AQI = mean(AQI, na.rm = TRUE),
  Avg_PM2.5 = mean(`PM2.5`, na.rm = TRUE),
  Avg_PM10 = mean(PM10, na.rm = TRUE),
  Avg_NO2 = mean(NO2, na.rm = TRUE),
  Avg_SO2 = mean(SO2, na.rm = TRUE),
  Avg_CO = mean(CO, na.rm = TRUE),
  Avg_O3 = mean(O3, na.rm = TRUE)
 ) %>%
 arrange(desc(Avg_AQI)) # Arrange by highest average AQI

# Print the summary table using kable for better formatting
kable(summary_by_city, caption = "Average Pollutant Levels and AQI by City")
```

Average Pollutant Levels and AQI by City

City	Avg_AQI	Avg_PM2.5	Avg_PM10	Avg_NO2	Avg_SO2	Avg_CO	Avg_O3
Dubai	66.06656	33.014406	56.532836	41.46515	24.926022	457.3446	74.23491
Cairo	53.53991	23.981506	41.456716	39.64536	52.602661	340.7080	41.70019
London	31.1755	12.28384	16.91044	31.9306	4.654640	207.141	33.1155

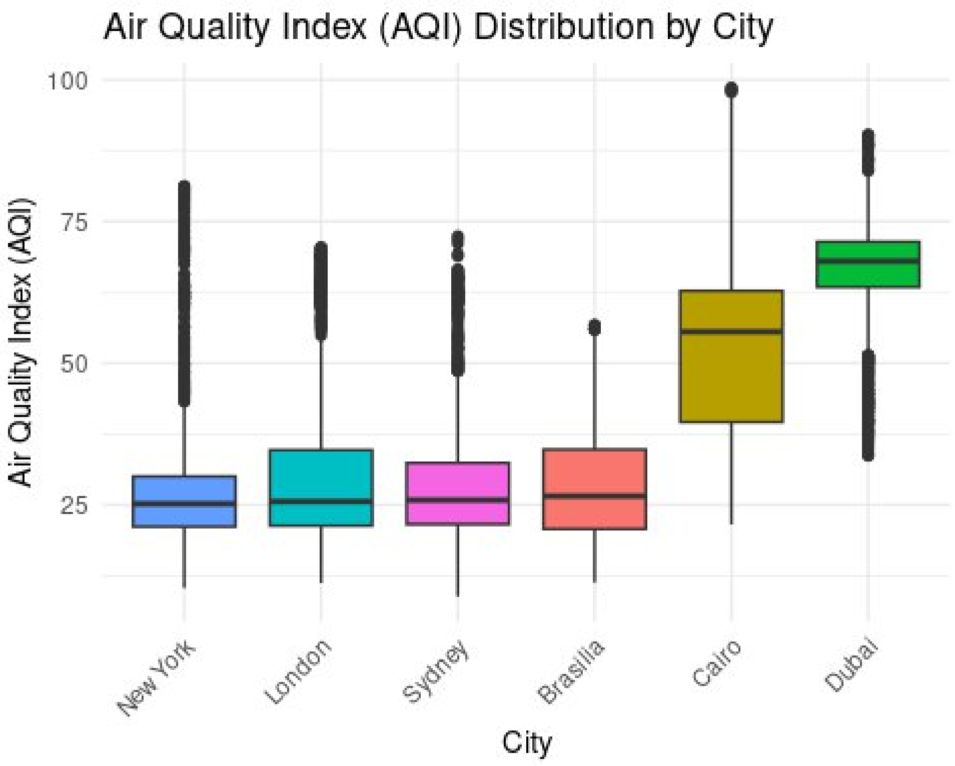
City	Avg_AQI	Avg_PM2.5	Avg_PM10	Avg_NO2	Avg_SO2	Avg_CO	Avg_O3
	8	2	8	9		5	1
Sydney	28.75797	11.086372	15.383387	10.62239	5.120247	123.7969	61.29137
Brasilia	28.13014	8.619338	9.560156	10.40318	1.948735	291.5964	64.45879
New York	27.86988	11.508241	13.100779	32.07852	6.919143	305.6042	35.78196

The summary table clearly shows that **Dubai** and **Cairo** have the highest average AQI, while **London** and **Sydney** have the lowest among the cities in this dataset.

3.2. Visualizing AQI Distribution by City

A boxplot is an excellent way to visualize the distribution of AQI values for each city, showing the median, quartiles, and potential outliers.

```
# Create a boxplot to compare AQI distributions across cities
ggplot(air_quality_clean, aes(x = reorder(City, AQI, FUN = median), y = AQI, fill = City)) +
  geom_boxplot() +
  labs(
    title = "Air Quality Index (AQI) Distribution by City",
    x = "City",
    y = "Air Quality Index (AQI)"
  ) +
  theme_minimal() +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1))
```



Distribution of Air Quality Index (AQI) by City

The boxplot confirms our findings from the summary statistics. Dubai and Cairo not only have higher median AQI values but also a wider range of pollution events, indicating greater variability and periods of very poor air quality.

.

4. Pollutant Correlation Analysis

To understand which pollutants have the most significant impact on the AQI, we can calculate the correlation matrix for the numeric variables.

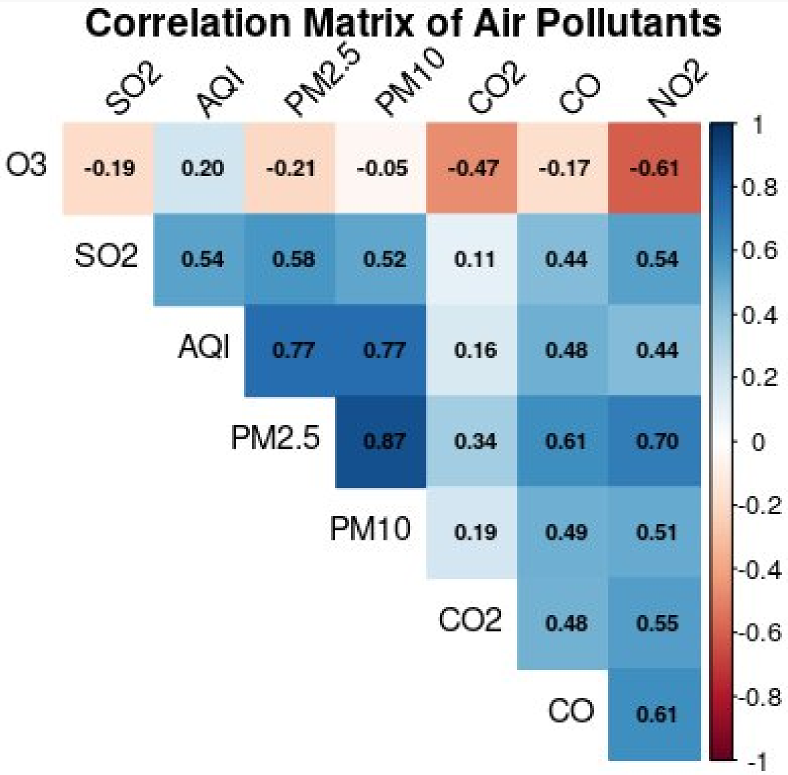
```
# Select only numeric columns for correlation analysis (excluding CO2 due to NAs)
numeric_data <- air_quality_clean %>%
  select(where(is.numeric))

# Calculate the correlation matrix
cor_matrix <- cor(numeric_data, use = "complete.obs")

# Print the correlation matrix for AQI with other pollutants
print("Correlation of pollutants with AQI:")

## [1] "Correlation of pollutants with AQI:"

print(cor_matrix["AQI", ])
```

```
##        CO       CO2       NO2       SO2        O3     PM2.5      PM10       AQI
## 0.4823113 0.1609228 0.4370484 0.5363346 0.1969600 0.7670189 0.7673494 1.0000000
```

4.1. Visualizing the Correlation Matrix

A heatmap provides an intuitive visualization of the correlation matrix, making it easy to spot strong relationships.

```
# Create a correlation heatmap
corrplot(cor_matrix,
        method = "color",       # Use color to represent correlation
        type = "upper",         # Show the upper triangle of the matrix
        order = "hclust",       # Reorder based on hierarchical clustering
        tl.col = "black",       # Text label color
        tl.srt = 45,            # Text label rotation
        addCoef.col = "black",  # Add correlation coefficients to the plot
        number.cex = 0.7,       # Size of the coefficient numbers
        diag = FALSE,           # Don't show the diagonal
        title = "Correlation Matrix of Air Pollutants",
        mar=c(0,0,1,0))         # Adjust margins
```



Correlation Heatmap of Air Pollutants and AQI

Key Observations from the Heatmap:

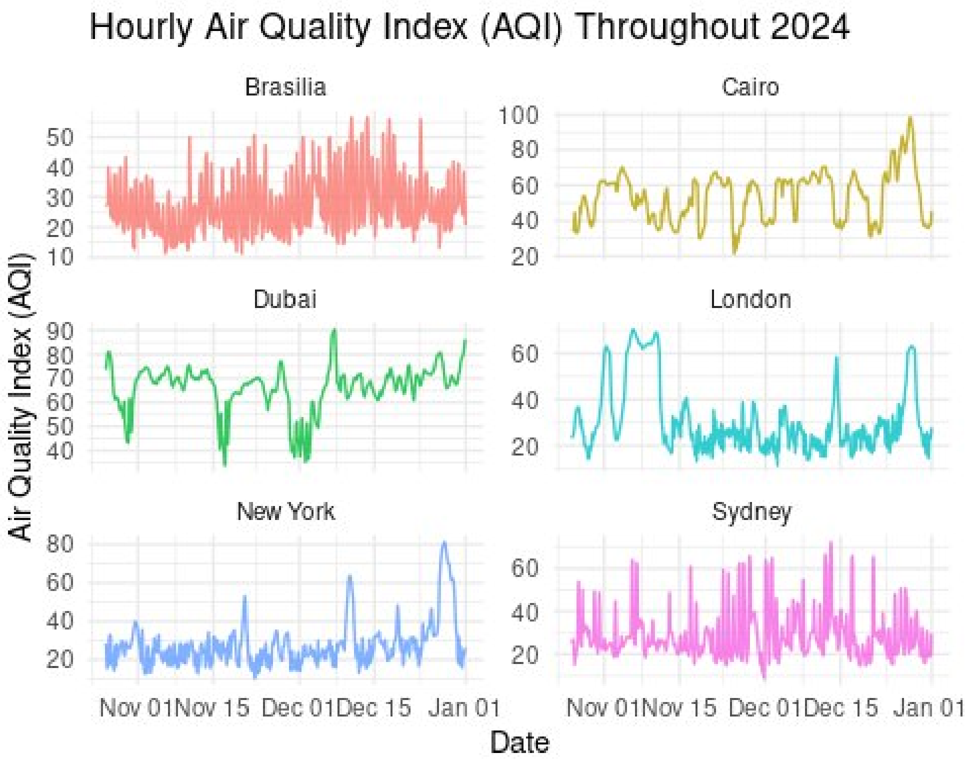
- **Strong Positive Correlations:** The AQI is strongly and positively correlated with **PM2.5**, **PM10**, **NO2**, and **SO2**. This indicates that these pollutants are major drivers of poor air quality.
- **Moderate Correlation: CO** also shows a moderate positive correlation with AQI.
- **Ozone (O3):** Interestingly, **O3** has a weak, slightly negative correlation with the overall AQI in this dataset. This can happen because ground-level ozone formation is a complex photochemical process that can be inversely related to other primary pollutants like NO2 under certain conditions.

.

5. Time Series Analysis of AQI

Finally, let's visualize how the AQI changes over time for each city. This can help us identify trends, seasonality, or specific pollution events.

```r
# Create a time series plot of AQI for all cities
ggplot(air_quality_clean, aes(x = Date, y = AQI, color = City)) +
  geom_line(alpha = 0.8, linewidth = 0.5) +
  facet_wrap(~City, ncol = 2, scales = "free_y") + # Create separate plots for each city
  labs(
    title = "Hourly Air Quality Index (AQI) Throughout 2024",
    x = "Date",
    y = "Air Quality Index (AQI)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



AQI Over Time for Each City

The time series plots reveal the dynamic nature of air pollution. We can observe daily and seasonal fluctuations in AQI for each city. Cities like Dubai and Cairo consistently show higher baseline AQI levels compared to London and Sydney.

.

6. Conclusion

This analysis provided a comprehensive overview of the air quality in six major global cities. The key findings are:

1. **Significant Disparities:** There are substantial differences in air quality, with **Dubai and Cairo experiencing significantly higher pollution levels** compared to **London and Sydney**.
2. **Key Pollutants:** Particulate matter (**PM2.5** and **PM10**) and **NO2** are the pollutants most strongly correlated with high AQI values, highlighting them as primary contributors to air pollution.
3. **Dynamic Nature:** Air quality is highly dynamic, with significant fluctuations observed over time in all cities.

This report serves as a foundational analysis. Further investigation could involve predictive modeling to forecast AQI, a more in-depth analysis of seasonal patterns, or correlating pollution data with public health records to study health impacts.