MTH208 – Final Project Report

Air Quality Index (AQI) Analysis and Visualization using R Shiny

Team 7

Anusha Gupta • Shreya Tarafdar
Sruthi Subramanian • Saroj Karwal

1. Introduction

The AQI or the Air Quality Index is a measure of the Air Quality of a City. It is given on a scale of 0-500, with 0-50 being the best air quality implying that the city has clean air, and hence minimal air pollution, while a rating of 300-500 implies that the air quality is dangerous and could have health implications. The Air Quality Index is measured based on the amount of 6 pollutants in the air - PM2. 5, PM10, carbon monoxide, sulfur dioxide, nitrogen dioxide, and ground-level ozone.

Aim: In this project we aim to look at the AQI over major cities in the World, visualizing it using a heat map. We also want to see if there is a correlation between the AQI and any other parameters like rainfall, humidity, and population.

2. Main questions addressed

1. How does air quality vary across different regions of the world?
2. Is there a relationship between development (measured using GDP per capita) and AQI levels?
3. Is there a relationship between population and AQI levels?
4. How do climatic factors like temperature, humidity, precipitation, and sunshine hours relate to AQI?
5. Which cities show the highest and lowest AQIs? Where are they located?

3. Data Scraping and Cleaning

3.1 City-wise AQI Data:

Visit the IQAir Global AQI Data.

We scraped city-wise PM2.5 data from this website, navigating through 180 pages with 50 cities on each, totally collecting data for around 9000 cities all over the world.

As these webpages were in JavaScript, we had to use the Chromote library to convert it into an HTML format, from which we have extracted the data. We used functions to split strings to clean the data, and obtain the city and country separately, finally saving a table with "cities", "countries", and "AQI" as the columns.

We faced one issue in this extraction. As the number of entries was high, and the total data we needed to collect was of a large magnitude, the R session used to crash before complete extraction happened. So this data has been extracted in batches, over 3 separate R sessions to account for this. All the data has been finally saved in the csv file : "AQI_by_city.csv", from where it is extracted when we run the app, to get AQI data.

AQI of a city is actually obtained by scaling PM2.5 by a factor. We have done this in the App code, to convert the PM2.5 data into AQI levels. (reference article).

3.2 Country-wise Weather and Climate Data:

Visit the Weather and Climate data.

The weather parameters - Daily mean temperature, Average precipitation, Average relative humidity, and Mean monthly sunshine, have been extracted from this site.

This site has links to country specific webpages which have tables with overall annual values for these parameters, which is actually the average over many of the previous years. We use this as a measure for these weather parameters, and scraped it country wise by going to each webpage. Here as the webpages were already in HTML, we directly scraped it using the rvest package, but had to extract the values for these parameters using methods used for strings. Finally we stored all of these values country wise as a data frame in the file: "Countrywise_Climate_Data.csv".

3.3 Country-wise Population Data and Geological Data:

Sources:

1. Population data.
2. Geological data.

We extracted both the country wise population data and geological data (longitudes and latitudes) from the above sites by using the `html_table()` function of the `rvest` library.

3.4 Country-wise GDP per capita Data:

Visit the Country wise GDP data.

We extracted the data of GDP per capita from the above site using the `html_table()` function of the `rvest` library.

3.5 City-wise Geological Data:

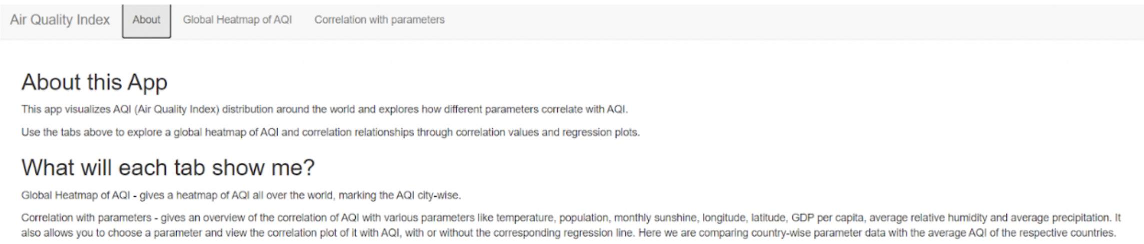
Visit the City-wise Geological data.

"worldcities.csv" was taken from this online database where the geological data (latitude and longitude) of cities is maintained. That was not scraped by us. We directly downloaded that file and used the coordinates for plotting the AQI in the heatmap.

4. Our App and Visualizations

There are three tabs in our app:

1. About - which gives an overview of what the App does, and explains what you can find in each of the tabs.



About this App

This app visualizes AQI (Air Quality Index) distribution around the world and explores how different parameters correlate with AQI.

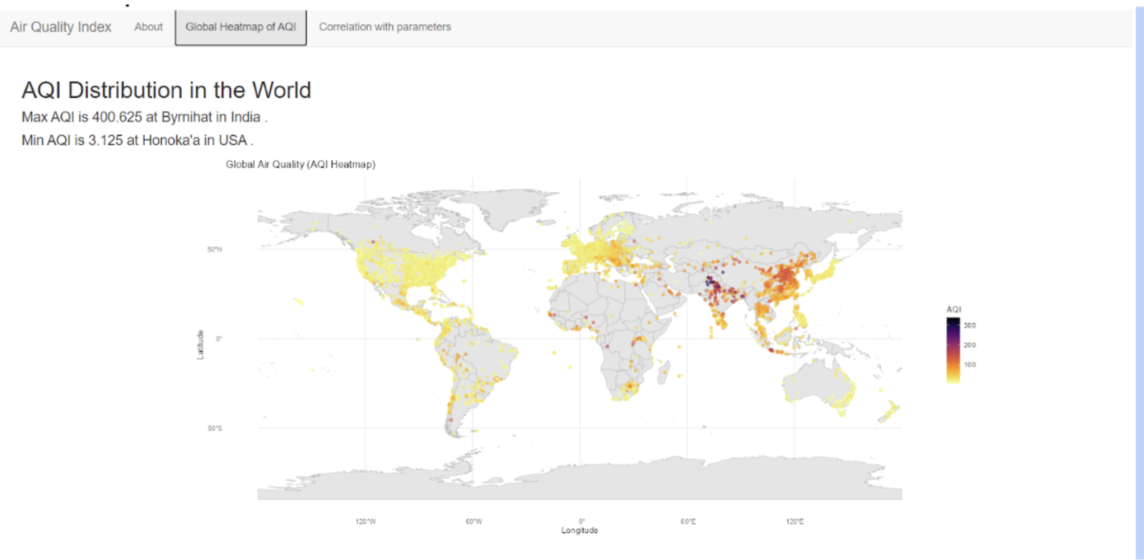
Use the tabs above to explore a global heatmap of AQI and correlation relationships through correlation values and regression plots.

What will each tab show me?

Global Heatmap of AQI - gives a heatmap of AQI all over the world, marking the AQI city-wise.

Correlation with parameters - gives an overview of the correlation of AQI with various parameters like temperature, population, monthly sunshine, longitude, latitude, GDP per capita, average relative humidity and average precipitation. It also allows you to choose a parameter and view the correlation plot of it with AQI, with or without the corresponding regression line. Here we are comparing country-wise parameter data with the average AQI of the respective countries.

2. Global Heatmap of AQI - which gives a heatmap of city-wise AQI on a world map.

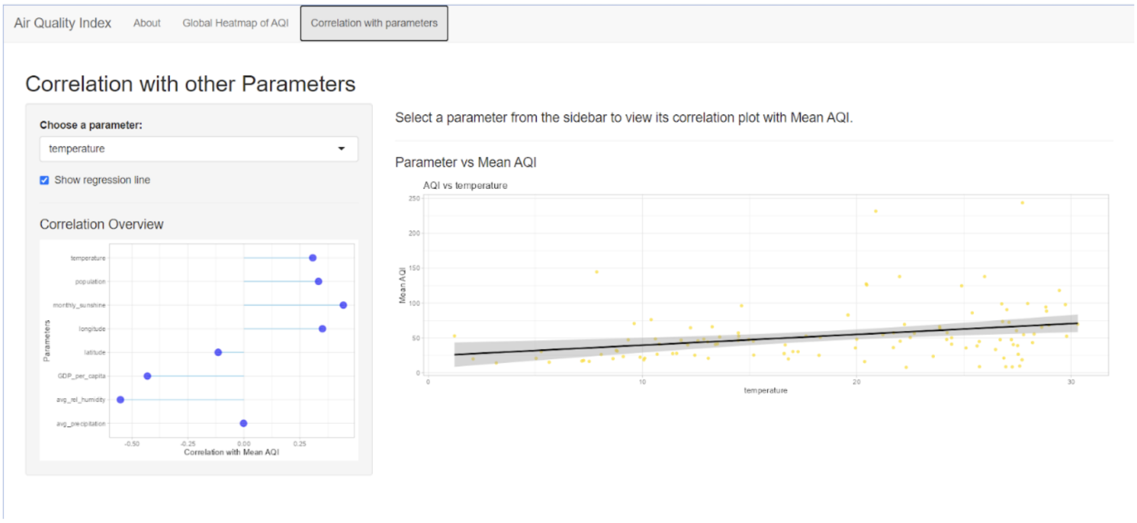


AQI Distribution in the World

Max AQI is 400.625 at Byrnihat in India .

Min AQI is 3.125 at Honoka'a in USA .

3. Correlation with parameters - which shows an overview of how the various parameters we have correlate (or don't) with AQI. This analysis is done country-wise. Here you can see a lollipop chart with an overview of correlation country-wise. There is also an option for you to pick one parameter, and see the correlation plot of that parameter with respect to AQI of a country. You also have an option to add a regression line to this plot for better analysis.



Correlation with other Parameters

Choose a parameter: temperature

☑ Show regression line

Select a parameter from the sidebar to view its correlation plot with Mean AQI.

Parameter vs Mean AQI

5. Libraries We Used

The following R packages were used during this project:

- `ggplot2`, `viridis` – for creating graphs and plots with visually appealing colors
- `dplyr`, `tidyverse` – for data manipulation and cleaning
- `rvest` – for web scraping of HTML pages
- `chromote` – for web scraping of sites using JavaScript (to convert them to HTML)
- `stringr` – for string manipulation (data cleaning)
- `rnaturalearth`, `rnaturalearthdata`, `sf` – for world map data needed for creating the heatmap
- `shiny` – for creation of the App

6. Reproducibility Notes

Download the zip file “MTH208 - Team 7 - Project” and extract the contents.

Install the packages listed below if not already installed, by executing the lines:

```
install.packages(c("ggplot2", "rnaturalearth", "rnaturalearthdata", "sf", "viridis"))
install.packages("shiny")
```

To run the app: Run the following commands in the console:

```
setwd("path to the extracted file")
shiny::runApp("app.R")
```

You can now see the app. You can go through the various tabs and view the visualizations. In the tab "Correlation with parameters", you can also use the controls on the left top of the screen to pick which parameter you would like to see the correlation of AQI with, and you can choose to have a regression line or not in that correlation plot.

To see Exploratory Data Analysis:

Within the R file "MTH208 - Team 7 - Exploratory Data Analysis.R", you can also see how we built up the plot codes. We have included all the graph codes for reference. Before executing this, make sure that you change the working directory at the top of the code to the extracted file path.

For Data:

The files "Country_wise_parameters.csv" and "AQI_by_city.csv", have been scraped. The code for this can be seen in "MTH208 - Team 7 - Data Scraping Code.R". Before executing this, make sure that you change the working directory at the top of the code to the extracted file path. On executing this file, you will be able to get the same csv files that we have saved in the folder. The scraping process took a lot of time, and the session had to be renewed to prevent timeouts a few times. Hence, we are attaching the files that are already scraped so that the app can be run directly using these. This will save computational power and is hence more efficient.

"worldcities.csv" was taken from an online database where the geological data (latitude and longitude) of cities is maintained. That was not scraped by us. We directly downloaded that file and used the coordinates for plotting the AQI in the heatmap.

7. Key Findings from Correlation Analysis

The correlation dashboard in the app reveals the following insights between Mean AQI and various weather parameters across countries:

- **Temperature:** A mild positive correlation of magnitude 0.31 was observed with AQI. This implies that the regions with higher average temperatures tend to have slightly higher AQI levels. Which makes sense as when temperature increases, the air becomes still and traps pollutants near the ground. Heat also speeds up chemical reactions that create smog. Because of this, hotter places usually have slightly higher AQI levels.

- **Population:** A strong positive correlation of magnitude 0.34 exists. Densely populated countries generally experience higher AQI values due to urbanization, vehicular emissions, and industrial concentration.

- **GDP per Capita:** Shows a negative of magnitude 0.43 correlation. Richer countries usually have better technology and stricter pollution laws. They can afford clean energy and waste control systems, so their air is cleaner. That's why higher GDP per capita is linked to lower AQI.

- **Average Relative Humidity:** Displays a negative correlation of magnitude 0.55. When humidity is high, water vapor in the air helps tiny dust and smoke particles stick together and settle down. It also helps scatter pollutants away. This reduces the number of particles in the air, which is why higher humidity is linked to lower AQI.
- **Average Precipitation:** Has a slight negative or near-zero correlation, as rainfall aids in pollutant washout but the effect is not consistent globally.
- **Monthly Sunshine Hours:** Shows a positive correlation of magnitude 0.44. This depicts correctly as sunlight increases the rate of chemical reactions in the air that create ozone and smog. This means areas with longer sunshine hours often have higher pollution levels, leading to a higher AQI.

These relations are visualized through scatter plots with a regression line to depict the trend effectively.

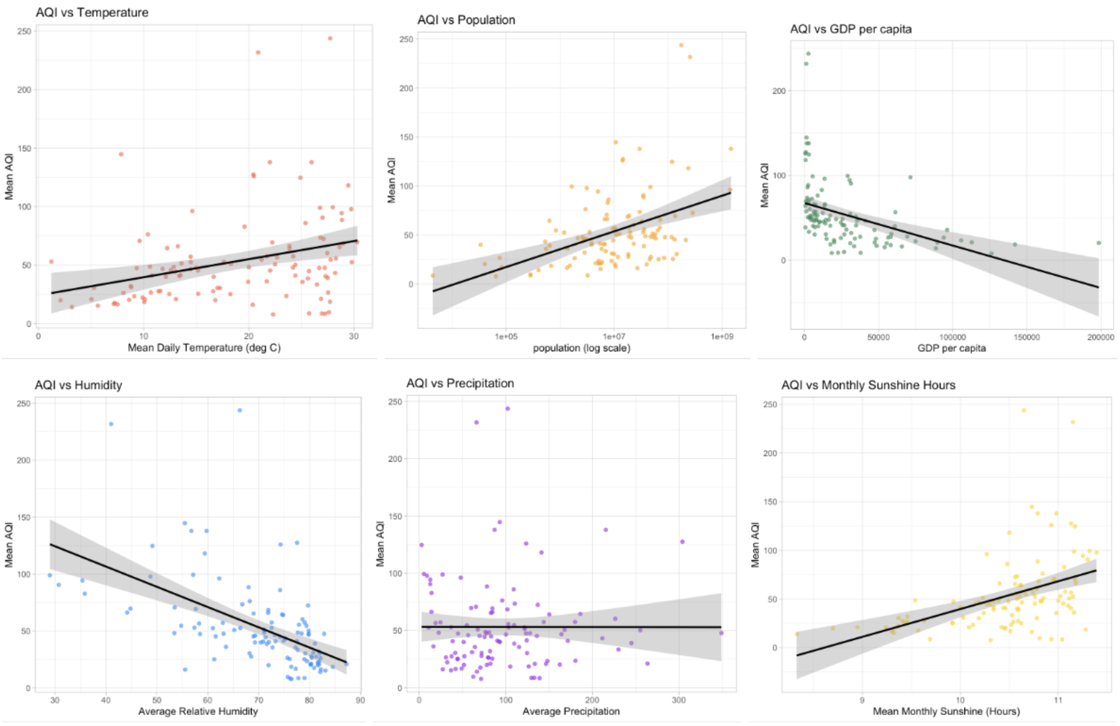


Figure 1: Corr. graphs for each parameter.

8. Data Sources Added Beyond the Proposal

We used a few extra data sources apart from those mentioned in the proposal. These helped make our results and visuals more complete and meaningful.

- Country wise GDP per capita data: Added to see how a country's income level affects its air quality. Richer countries usually have cleaner air because they use better technology and follow stricter pollution rules.

- City-wise Coordinates: Added so we could accurately show each city's location on the world map and create clear global heatmaps of AQI.

- Population data: Added to study how the number of people living in a country affects its AQI. Densely populated areas often have more vehicles and industries, which increase pollution levels.

9. Limitations and Ethics

- Incomplete Global Coverage: Data availability is higher for developed, densely populated regions, potentially overstating correlations in urban zones while rural pollution remains underrepresented.

- Temporal Simplification: Using averages overlooks daily and seasonal fluctuations.

- Standardization Differences: Various AQI data sources (IQAir, CPCB) apply different pollutant-weighting methods, affecting cross-country comparability.

- Ethical Scraping: All the data was taken from free and public websites and used only for learning and making visualizations. While collecting the data, we made sure not to overload any website by limiting how fast we downloaded the information.

- Technical Limitations: While collecting the data, R crashed several times because the files were very large. So, we collected the data in smaller parts and saved it as CSV files to make the process easier and repeatable.

10. Resources

- IQAir Global Air Quality Data

- CPCB India Air Quality Portal

- World Population Review

- Weather and Climate data

- City-wise Geological data

- Population data

- Country wise GDP data

- Geological data

- R Shiny Documentation

- ggplot2 Visualization Guide