

Egypt University of Informatics

Computer and Information Systems

Data Analysis Course

The Analysis of Air Quality Across countries and contributing factors

Submitted by:

Youssef Wael 23-101183

Youssef Ihab 23-101138

Youssef Ashoush 23-101049

Ahmed Yasser 23-101168

Under Supervision of:

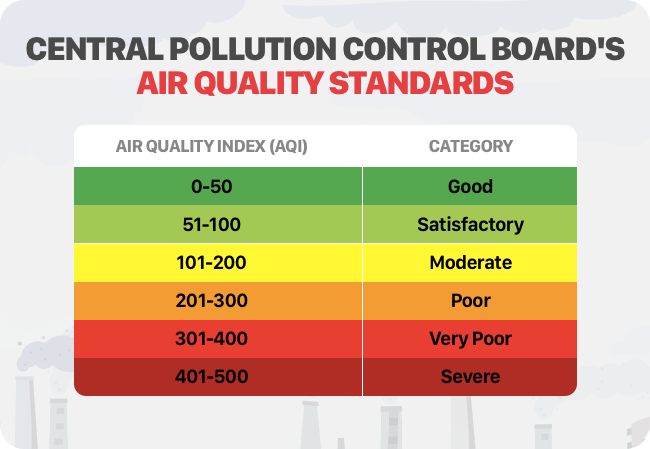
Prof. Mohamed Taher Alrefaie

Dr. Shereen el Fayoumy

23/5/2025

# Introduction:

T**he Air Quality Index (AQI) is used for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. The AQI focuses on health effects you may experience within a few hours or days after breathing polluted air. Think of the AQI as a yardstick that runs from 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern.**

**As students living in Cairo, we witness daily how the gray haze that blankets our city affects our health and quality of life, driving us to investigate the complex factors behind Air Pollution, not only in Egypt, but across all countries. Such knowledge would help us further face this pollution crisis.**

# Research Question:

**How do such factors ( Population density, GreenHouse Gasses Emissions, Forest percentage of lands ) affect the AQI across countries?**

# Hypothesis:

1. **Does increased GHG emission correlate with higher AQI values?**

**Null Hypothesis (H0):**

There is no correlation between AQI value and GHG emissions.

**Alternative Hypothesis (H1):**

There is a positive correlation between AQI value and GHG emissions.

1. **Does increased forest cover correlate with lower AQI values?**

**Null Hypothesis (H0):**

There is no correlation between AQI value and forest cover.

**Alternative Hypothesis (H1):**

There is a negative correlation between AQI value and forest cover.

1. **Does increased population density correlate with higher AQI values?**

**Null Hypothesis (H0):**

There is no correlation between AQI value and population density**.**

**Alternative Hypothesis (H1):**

There is a positive correlation between AQI value and population density.

1. **Does mean AQI values differ significantly across continents?**

**Null Hypothesis (H0):**

There is no difference in mean AQI values among different continents.

**Alternative Hypothesis (H1):**

There is a significant difference in mean AQI values among different continents.

# Population of Interest:

**Countries across the world**

# Sampling Method:

**We used a non-probability sampling method, specifically convenience sampling. Our data was collected by web scraping the Global AQI website, which provides real-time air quality data from various monitoring stations. This method was chosen for its accessibility and ease of data collection,**

# Bias Identification:

**One potential bias in our project is selection bias, as we relied on data scraped from the Global AQI website. This site may only report from certain monitoring stations, possibly missing areas with less coverage or lower socioeconomic visibility. We also have some missing countries in our data. Additionally, since we didn’t design survey questions but used existing environmental data, we avoided response bias. We stayed objective in our analysis by not letting personal opinions influence how we interpreted the AQI values or pollution sources.**

# Survey Questions/Collected Data/Dataset:

**Our Columns:**

(110, 8)  
Country - Country Name

Status - Air Quality status based on AQI value

Code - Country 3-letter Alpha code (to help us visualize the data on a map)

AQI Value - Value of Air Quality Index (<https://aqicn.org>) (as of 20/05/2025)

GHG\_Emissions - GreenHouse Gases Emissions in **kilotonnes of CO₂ equivalent (kt CO₂eq)**. This unit standardizes various greenhouse gases—such as carbon dioxide (CO₂), methane (CH₄), nitrous oxide (N₂O), and fluorinated gases—by converting their emissions into the equivalent amount of CO₂ based on their global warming potential (GWP).

(<https://edgar.jrc.ec.europa.eu/report_2024#data_download>)

Forest cover - Percentage of lands covered by forests (<https://ourworldindata.org/forest-area>)

Population Density - Number of People / km (<https://ourworldindata.org/grapher/population-density.csv?v=1&csvType=full&useColumnShortNames=true&utm_source=chatgpt.com>)

Continent - Which Continent does each country belong to

# Analysis:

**First, We scraped the AQI values for 110 countries off the aqicn website, then trying to find related data we started searching for population density data, forest land cover % data and GreenHouse Gas emissions data from reputable organizations listed above, then merged all these data into one dataframe to start performing our analysis on it.**

|  |
| --- |
|  |
| Descriptive Statistics regarding the numerical columns in our data |

|  |
| --- |
|  |
| AQI values by country plotted on the world map, We can see countries like India and China have experienced critical levels |

|  |
| --- |
|  |
| GHG Emissions by country plotted on the world map, We can see China leads the world by a huge margin in GreenHouse gas emissions possibly because of their huge economic growth and reliance on coal for Energy |

|  |
| --- |
|  |
| Population Density by country plotted on the world map, We can see India’s population density is very high which may be a contributing factor to why their AQI is that high |

|  |
| --- |
|  |
| Map of countries with the percentage of the country’s land is covered in Forests. We can see that Brazil is very much covered with Forests. |

|  |
| --- |
|  |
| Scatter plot of AQI values and GHG\_Emissions (removed an outlier in China as it skewed the graph very heavily and didn’t show the relation clearly) we can see that there is a positive relationship between AQI values and GHG\_Emissions |

|  |
| --- |
|  |
| Scatter plot of AQI values and Forest Cover , We can see that there’s hardly any correlation between both variables |

|  |
| --- |
|  |
| Scatter plot of AQI values and Population Density, Again we can see that there’s hardly any correlation between both variables |

|  |
| --- |
|  |
| Side-by-Side Box plots of AQI values by continent, showing how they differ from continent to continent |

# Hypothesis Testing Steps:

### Hypothesis 1

**Research Question:** Does increased GHG emission correlate with higher AQI values?

**Step 1: State the hypotheses**

* H₀: There is no correlation between AQI value and GHG emissions.
* H₁: There is a positive correlation between AQI value and GHG emissions.

**Step 2: Choose the significance level**

* α = 0.05

**Step 3: Conduct the statistical test**

* Pearson correlation test
* Correlation = 0.5060, p-value = 0.0000

**Step 4: Make a decision**

* Since p-value < 0.05, we **reject H₀**.
* **Conclusion:** There is a statistically significant positive correlation between GHG emissions and AQI values.

### Hypothesis 2

**Research Question:** Does increased forest cover correlate with lower AQI values?

**Step 1: State the hypotheses**

* H₀: There is no correlation between AQI value and forest cover.
* H₁: There is a negative correlation between AQI value and forest cover.

**Step 2: Choose the significance level**

* α = 0.05

**Step 3: Conduct the statistical test**

* Pearson correlation test
* Correlation = -0.0201, p-value = 0.8440

**Step 4: Make a decision**

* Since p-value > 0.05, we **fail to reject H₀**.
* **Conclusion:** There is no significant evidence of a correlation between forest cover and AQI values.

### Hypothesis 3

**Research Question:** Does increased population density correlate with higher AQI values?

**Step 1: State the hypotheses**

* H₀: There is no correlation between AQI value and population density.
* H₁: There is a positive correlation between AQI value and population density.

**Step 2: Choose the significance level**

* α = 0.05

**Step 3: Conduct the statistical test**

* Pearson correlation test
* Correlation = -0.1000, p-value = 0.3351

**Step 4: Make a decision**

* Since p-value > 0.05, we **fail to reject H₀**.
* **Conclusion:** No significant correlation was found between population density and AQI values.

### Hypothesis 4

**Research Question:** Do mean AQI values differ significantly across continents?

**Step 1: State the hypotheses**

* H₀: There is no difference in mean AQI values among different continents.
* H₁: There is a significant difference in mean AQI values among different continents.

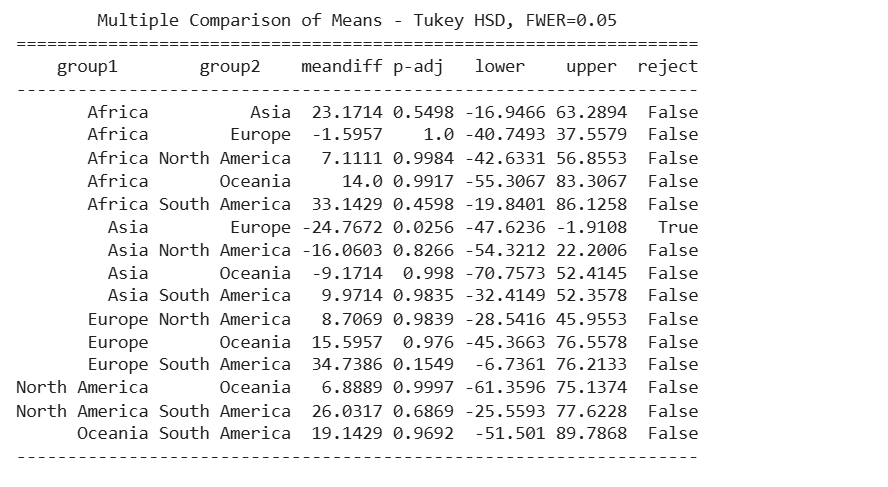
**Step 2: Choose the significance level**

* α = 0.05

**Step 3: Conduct the statistical test**

* ANOVA test
* F-statistic = 2.7460, p-value = 0.0227

**Step 4: Make a decision**

* Since p-value < 0.05, we **reject H₀**.
* **Conclusion:** There is a statistically significant difference in AQI values among continents.  
  Bonus points:  
  

We used the Tukey HSD test to find which differences between 2 means were the most significant and we found that the difference between Europe and Asia was the most significant.

# Conclusion:

1. **There is a statistically significant positive correlation between GHG emissions and higher AQI values, indicating that as emissions rise, air quality worsens possibly due to Industrial emissions and/or car emissions and other contributing factors.**
2. **The mean AQI values differ significantly between continents, highlighting regional disparities in air pollution levels specially between Europe and Asia.**

# Any potential issues:

1. **When collecting the datasets, unfortunately there were some countries that we couldn’t access the data for.**
2. **The AQI value changes everyday and the dataset we have was from 20/5/2025.**
3. **The newest dataset we found for the forest percentage land coverage in each country was from 2020.**
4. **There may be confounding variables that we didn’t consider that may have affected our analysis.**

.