Data Reporting & Visualization

Homework

By

Nasiba Mammadli

Instructor: Camille DUQUESNE

École Pour l'Informatique et les Techniques Avancées- EPITA

Masters in Aritifical Intelligence Systems

Introduction

The dataset was **initially sourced from Numbeo**, a crowd-sourced database that collects real-time data on city living conditions. The data was compiled as an **aggregation of user voting**, meaning individuals from different locations contributed their assessments of air and water pollution levels. The dataset was collected in **2020 or 2021**, making it relatively recent and useful for studying modern urban pollution trends.The dataset is available under a **CC0 (Public Domain) License**, meaning it is free to use, modify, and share without restrictions.

The dataset includes several key environmental and geographic variables:

- **Air Quality Index (0-100):** Measures air pollution levels, where **0 represents very bad air quality** and **100 represents excellent air quality**.
- **Water Pollution Index (0-100):** Measures water contamination, where **0 represents clean water** and **100 indicates extreme pollution**.
- **City, Region, and Country:** Location-based details allow comparisons of air and water pollution levels across different geographical areas.

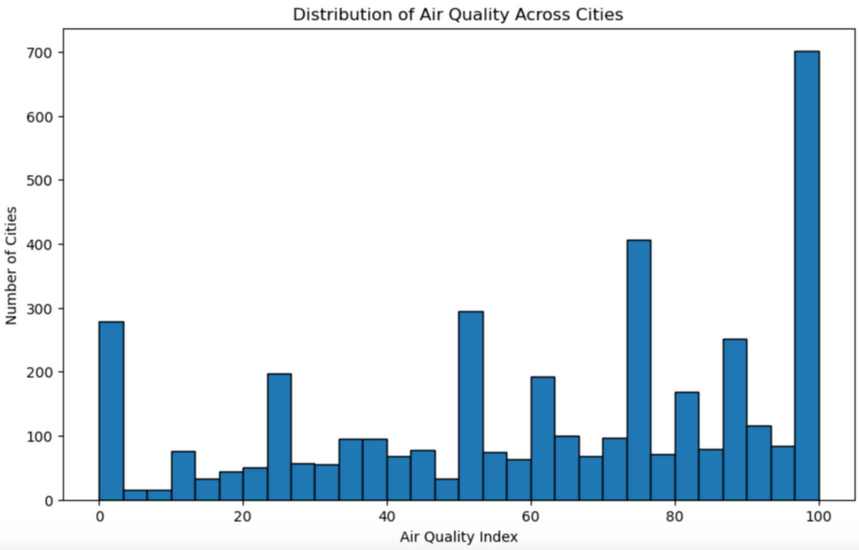
The dataset contains information on 3,796 cities worldwide, making it one of the **largest** open datasets for analyzing urban pollution levels.

The exact purpose of data collection is unclear, as the dataset does not specify if it was collected for academic research, policy development, or general public awareness. However, its potential applications include:

- **Identifying trends in air and water pollution** across cities.
- **Comparing pollution levels regionally or globally**.
- **Helping policymakers and researchers** better understand urban environmental challenges.

Additionally**,** Numbeo **is** a well-known crowd-sourced database that provides real-time information on cost of living, pollution, healthcare, and transportation worldwide. While it relies on user votes, making it susceptible to bias, it remains one of the largest publicly accessible environmental data sources.

This dataset provides valuable insights into global air and water pollution levels across 3,796 cities. While the purpose of collection remains **unclear**, it serves as an important resource for analyzing pollution trends, informing policy decisions, and studying the environmental challenges faced by urban areas worldwide.



Graph 1. Analysis of the Air Quality Distribution Graph

This graph represents the **distribution of Air Quality Index (AQI) across different cities**. It is a **histogram** that visualizes how air quality varies across the dataset.

X-Axis (Air Quality Index - AQI):

- The AQI values range from **0 to 100**.
- Lower values (closer to 0) indicate **better air quality**, while higher values (closer to 100) indicate **poor air quality**.

Y-Axis (Number of Cities):

- The vertical axis represents the **number of cities** that fall within each AQI range.
- Higher bars indicate that more cities fall within a specific AQI range.

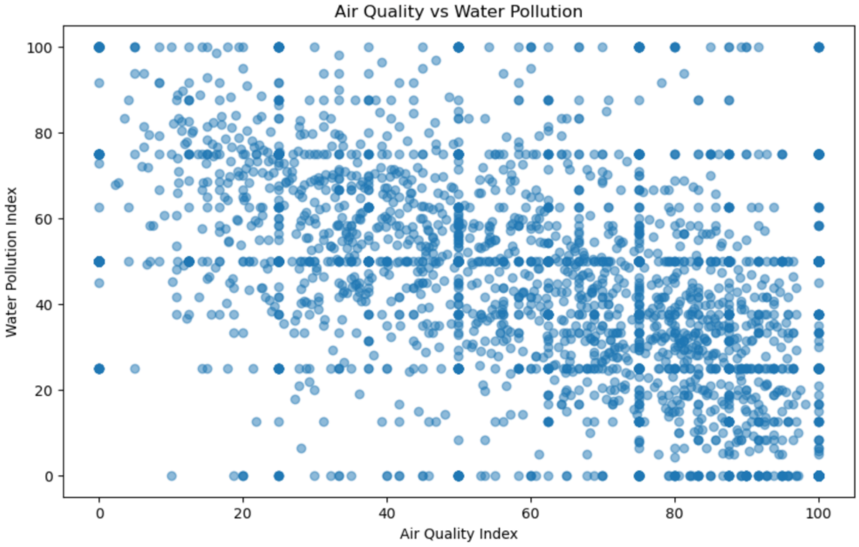
Distribution Insights:

- There are **several peaks**, indicating **concentrations of cities** at certain AQI values.
- The **largest peak is near 100**, meaning many cities experience **very poor air quality**.
- There are also noticeable peaks near **0-10 and around 30-40**, suggesting that some cities have **good air quality**, while others have moderate pollution levels.

Implications:

- A significant number of cities **suffer from very poor air quality (AQI ~100)**, which could indicate **high pollution levels, industrialization, or traffic congestion**.
- Some cities **maintain good air quality (AQI 0-10)**, possibly due to **strict environmental policies, lower industrial emissions, or better public transport systems**.

- The **uneven distribution** suggests a **wide variation in pollution levels across cities**, meaning air quality challenges are **not uniform globally**.



Graph 2. Analysis of the Scatter Plot: Air Quality vs. Water Pollution

This scatter plot visualizes the relationship between **Air Quality Index (AQI) and Water Pollution Index** across various cities.

X-Axis (Air Quality Index - AQI):

- Represents **air pollution levels**, ranging from **0 (clean air) to 100 (high pollution).**

Y-Axis (Water Pollution Index):

- Represents **water pollution levels**, ranging from **0 (clean water) to 100 (high pollution).**

Distribution of Data Points:

- The data points are **widely spread across the plot**, indicating **no immediate strong correlation** between air quality and water pollution.
- Some cities have **high air pollution but low water pollution**, while others show **the opposite trend**.
- There are **clusters of cities in the middle range (AQI ~40-80, Water Pollution ~40-80)**, suggesting that **many cities struggle with both air and water pollution at moderate levels.**

Possible Correlation:

- If there was a **strong correlation**, we would see a **clear pattern (e.g., upward or downward trend).**

- The scatter suggests a **weak or non-linear relationship** – meaning some cities have **both poor air and water quality**, while others experience only one form of pollution.

In the scatter plot visualizing the relationship between **Air Quality Index (AQI) and Water Pollution Index**, we can identify **potential outliers**—data points that **deviate significantly** from the overall trend.

What Are Outliers in This Context?

Outliers are **points that stand far from the majority of the data**, meaning cities where either:

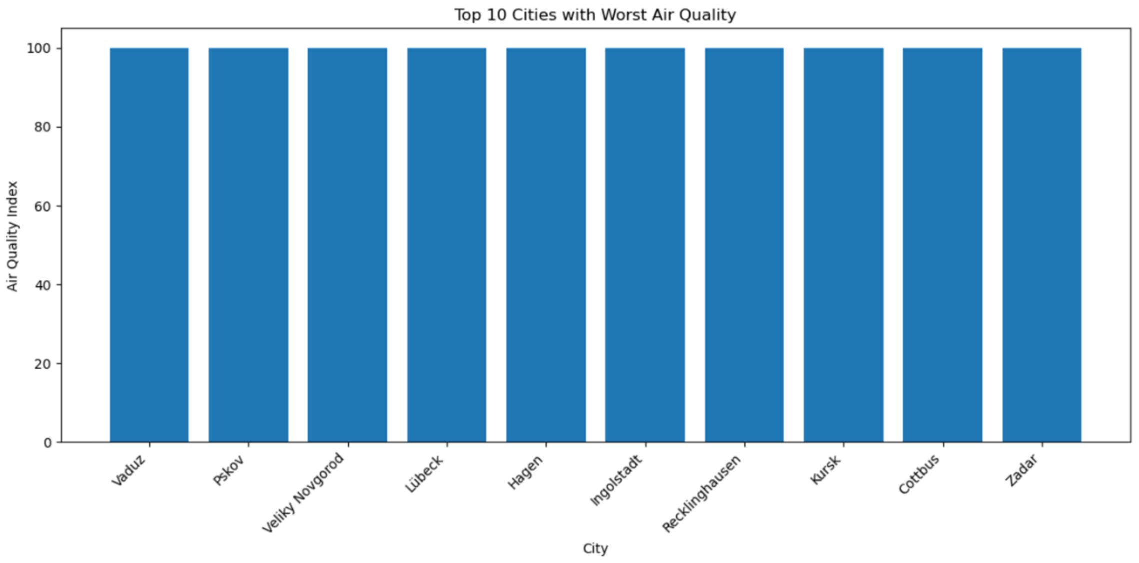
1. **Air Quality is extremely high (clean air) while Water Pollution is very high**, which is an unusual combination.
2. **Air Quality is extremely low (polluted air) while Water Pollution is very low**, which is also an uncommon scenario.
3. **Both Air Quality and Water Pollution are at extreme values (0 or 100)**, meaning a city experiences **either perfect environmental conditions or the worst pollution levels possible**.

Possible Causes of These Outliers

- **Geographical Factors**: Some cities may be located in regions where **air pollution is controlled (e.g., coastal cities with wind circulation)** but water pollution is high due to **industrial waste dumping into rivers**.
- **Data Collection Issues**: Since the dataset is based on **user-reported data (from Numbeo),** extreme values could be **biased perceptions rather than actual measured pollution levels**.
- **Unique Local Policies**: Some cities may have **strict air quality regulations** but lack enforcement on **water treatment** facilities, leading to contrasting pollution levels.

Implications:

- Some cities may have **effective policies for controlling one type of pollution but not the other.**
- Urbanization and industrial activities could be **impacting air and water pollution differently** across locations.
- Further analysis, such as calculating a **correlation coefficient (Pearson's r),** could help determine the exact relationship.



Graph 3. Analysis of the Bar Chart: Top 10 Cities with Worst Air Quality

This **bar chart** represents the **top 10 cities with the worst air quality**, ranked by their **Air Quality Index (AQI).** The higher the AQI value, the poorer the air quality in that city.

X-Axis (Cities)

- The cities listed on the x-axis are the **10 worst-ranked cities** in terms of air quality.
- Cities like **Vaduz, Pskov, Nizhny Novgorod, Lübeck, and Hagen** are included, suggesting a **geographical spread** of pollution.

Y-Axis (Air Quality Index - AQI)

- The **AQI scale** measures **air pollution levels**, where **higher values indicate worse air quality**.
- In this graph, all cities seem to have an **AQI near 100**, which is classified as **"Hazardous"** according to most air quality standards.

Distribution of Pollution Across Cities:

- **All 10 cities have nearly identical AQI values (~100),** indicating **severe pollution in these urban areas**.
- This suggests **high levels of industrial emissions, vehicle pollution, and poor air circulation**.

Potential Causes of High AQI in These Cities:

- **Industrial zones:** Cities with major factories or coal-based power plants often have higher pollution levels.
- **Heavy traffic congestion:** Urban areas with high vehicle density contribute significantly to poor air quality.
- **Geographical factors:** Some cities may have natural barriers (e.g., valleys) that trap pollutants.

References

1. **City API Project.** (2021). *Air and Water Pollution Dataset.* Kaggle. Retrieved from https://www.kaggle.com
2. **Montgomery, D. C., Peck, E. A., & Vining, G. G.** (2021). *Introduction to Linear Regression Analysis (6th ed.).*Wiley.
3. **Numbeo.** (2021). *Global Pollution Index: Air and Water Quality Data.* Retrieved from https://www.numbeo.com
4. **Tukey, J. W.** (1977). *Exploratory Data Analysis.* Addison-Wesley.