Report on Air Quality Data Processing

1 Main Data Issues

- **Significant Gap**: During the period from 2014-2024, air quality data has significant gaps, particularly between **2018-2020**, greatly affecting the analysis.
- **Missing Data**: A large amount of missing data exists during this period, further highlighting the data deficiency.
- **Dividing the Data into Two Main Stages**:
    1. **Pre-COVID Stage (before 2018)**:
        - Data on **PM2.5 and PM10** is relatively complete, but indices such as **SO2, CO, O3, NO2** are significantly missing.
        - Numerous **outliers** are present in the data.
    2. **Post-COVID Stage (after 2020)**:
        - Data for the indices **SO2, CO, O3, NO2** is well recorded, but data for **PM2.5 and PM10** is missing.
        - Data for O3 is missing after 2024.

2 Proposed Solutions

2.1 Option 1: Sequentially Merge Data from Both Stages

- **This option was not selected** for the following reasons:
    - **Data Quality** is paramount compared to quantity. Having a smaller dataset that is accurate is far more valuable than having a larger dataset that is unreliable.
    - **Significant differences between the two stages before and after COVID**: Climate conditions, pollution levels, and social factors changed significantly after the pandemic.

2.2 Option 2: Choose One Stage for Analysis

- **Re-evaluation of the problem**: The analysis will focus on the **2020 to 2024** stage to ensure accuracy and relevance to the current situation.
- **Reasons for choosing the post-COVID stage**:
  - **More complete data**: Indices such as **SO2, CO, O3, NO2** are well recorded, despite the lack of **PM2.5 and PM10** data.
  - **PM data can be imputed**: Although missing, PM indices have clear cyclical patterns and can be **imputed** based on well-recorded years. This can be demonstrated through line charts for each year of PM2.5 and PM10.
  - **Closer to the current time**: This stage reflects the current situation more accurately and is easier to apply to contemporary analyses.

3 Sub-issues Identified After Re-evaluating the Problem

- **Filling 2020 and 2022 Data Logically**: Different methods can be applied to reasonably fill the missing data for 2020 and 2022:
  1. **2020 Imputation:** A local newspaper report indicated that the average PM index in 2020 dropped by 12.8% compared to 2019. After filling in the missing 2019 data using 2014's values increased by 72.5% (due to a high correlation shown in a heat map comparison and the difference after computing), we imputed it in 2020's missing data.
  2. **2022 Imputation:** Due to the cyclical nature of PM levels, the missing data for 2022 can be imputed by averaging the same day's values from corresponding days in years where data is available.
- **Solution for Missing O3 Data:** Utilize Selenium to scrape data from various websites and supplement the missing values accordingly.
- **Handling Small Amounts of Missing Values (null values)**: Apply **interpolation** methods to fill in missing values.
  1. **Linear Imputation:** Given the temporal continuity in PM data, linear imputation is employed. This method is straightforward and allows for a smooth transition between values, effectively capturing gradual changes over time while providing a reasonable estimate for missing data points.
  2. **Nearest-Neighbor Imputation:** For datasets exhibiting step-wise or abrupt changes, such as NO2, SO2, O3 and CO concentrations, nearest-neighbor imputation is employed. This method effectively preserves the discontinuous nature of the data, ensuring that sharp transitions remain intact.

- **Removing Outliers**: Remove outliers to improve the accuracy of the analysis.

4 Summary

- The **post-COVID** stage (2020-2023) was selected for analysis due to higher data availability and relevance to current environmental conditions.
- Missing data for 2020 and 2022 was compensated using reasonable methods, including adjusting 2020 values based on trends and averaging corresponding days from other years for 2022.
- Interpolation methods addressed small gaps, and outliers were removed to improve data reliability.
- **Selenium** was utilized to efficiently gather missing **O3** and **CO** data from websites, ensuring accuracy and saving processing time.
- The cleaned dataset is now ready for detailed statistical analyses and machine learning investigations on the correlation between weather conditions and air pollution in **Da Nang**.