DATA ANALYSIS

# **Insights from analysis**

1. Most important attribute:

* After performing a correlation analysis, the attribute **PT08.S1(CO)** was found to be the most influential attribute, with a total correlation sum of 6.52. This suggests that **PT08.S1(CO)** has strong relationships with other attributes in the dataset and may play a central role in air quality measurements or equipment readings related to CO.

1. **Central Tendency and Dispersion for PT08.S1(CO)**:

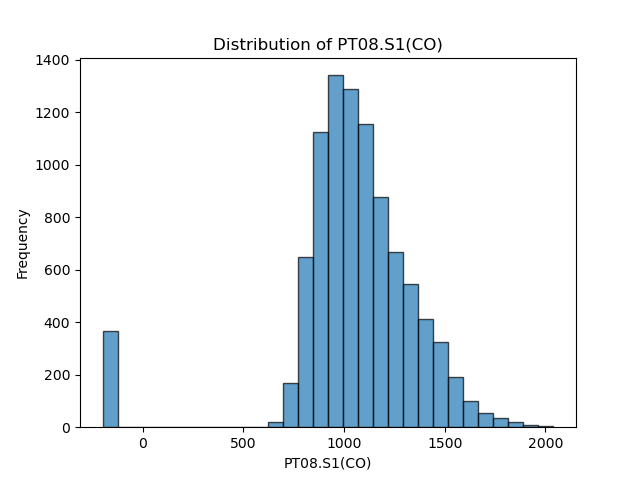
* **Mean**: The mean value of **PT08.S1(CO)** is **1048.87**, indicating that, on average, the sensor records readings around this value.
* **Median**: The median value is **1052.50**, which is close to the mean, suggesting a somewhat symmetric distribution with respect to its central tendency.
* **Mode**: The mode value is **-200.00**, which represents missing or invalid data in this dataset.
* **Standard Deviation**: The standard deviation is **329.82**, showing a moderate spread of the sensor readings around the mean. The high standard deviation indicates that the data points are spread out over a wider range.
* **Variance**: The variance is **108,779.26**, reinforcing that the data has significant variability.
* **Min and Max Values**: The minimum value is **-200.00** (which indicates missing data), and the maximum is **2039.75**. This large range (**2239.75**) demonstrates that the sensor captures a wide range of values.

1. **Missing Values**:

 The dataset contains **366 missing or invalid values** for the **PT08.S1(CO)** attribute, where a value of **-200** is used to represent missing data. These missing values may need to be handled before using the data for further analysis or modeling.

1. **Distribution of PT08.S1(CO)**:

* The histogram of the **PT08.S1(CO)** attribute shows how the values are distributed across the dataset. Given the presence of a mode of **-200**, the distribution might be skewed, and careful consideration should be given to whether these missing values impact the overall pattern in the dataset.



## After Handling missing values:

1. **Mean and Median Comparison**:

* **Mean (1048.87) vs. Median (1052.5)**: The mean is slightly lower than the median, indicating that the data distribution may be right-skewed. This means there could be some lower-value outliers pulling the mean down.

1. **Mode**:

* **Mode (1063.0)**: After cleaning, the mode has shifted from -200.0 to 1063.0. This reflects that 1063.0 is the most frequent value in the cleaned dataset, indicating that the previous mode was likely an anomaly or data entry error.

1. **Standard Deviation and Variance**:

* **Standard Deviation (212.92) and Variance (45332.84)**: Both the standard deviation and variance are high, reflecting a significant amount of dispersion around the mean. This suggests that the data values are spread out over a wide range.

1. **Range**:

* **Range (1392.5)**: The large range indicates that the dataset covers a broad spectrum of values. This supports the observation of high variability and suggests the presence of extreme values or outliers.

1. **Interquartile Range (IQR)**:

* **IQR (280.0)**: The IQR, which measures the spread of the central 50% of the data, is relatively smaller compared to the overall range. This implies that while the central portion of the data is concentrated, there are significant extremes outside this range.

**Summary**:

* **Data Distribution**: The data shows significant variability with a right-skewed distribution. The mean being higher than the median suggests some higher-value outliers.
* **Central Tendency**: The mode and median being the same reflects a central value that is common in the dataset.
* **Dispersion**: High standard deviation and variance, coupled with a large range, indicate substantial spread and variability in the data. The smaller IQR compared to the range highlights that while the middle 50% of the data is more concentrated, there are still significant extremes.

## Python codes and output:

// TO be Filled