Data Preprocessing

The dataset had missing values, which were represented by -200 as a placeholder. Additionally, some columns had gaps that required filling to avoid errors.

Steps:

1. Replaced -200 with NaN across the dataset. This ensures that these values are recognised as missing data, which is essential to handling these missing values.

2. Handled missing data with interpolation
    - For Environmental Columns: [T, RH, AH]
      I applied linear interpolation as it would estimate missing values based on surrounding data points. I used this because these variables are in continuous trend, and this interpolation technique was the best fit, according to me.
    - For Sensor Columns
      I checked the proportion of missing values. If the missing values were less than 5%, I used linear interpolation again for the column. But if the missing data was more than 5%, I filled the missing values with the mean of the column. This is because larger gaps in the data are less predictable and filling withthe mean avoids introducing bias from interpolation.

3. Dropped the NMHC column as it was having 90% NULL Values
   NMHC(GT) 8443 90.231912

4. Combining Date and Time into a Single Column
   The dataset had separate date and time columns; thus, for the ease of working in time-based analysis, I changed it to a unified date and time column.

5. Ensuring Consistent Data Types
   The new date and time column was converted to datetime format and all the other columns were converted to int or float.