Pre-processing coding issues

# **Step 1: Data Loading and Temporal Feature Engineering**

The *Collect DateTime* column in the raw dataset used the American date format (MM/DD/YYYY hh:mm:ss AM/PM), which initially prevented automatic conversion to a datetime object. To address this, a custom format was specified in the pd.to\_datetime() function to correctly parse the 12-hour timestamps.  
Since the analysis focused on long-term and seasonal trends rather than hourly variation, timestamps were truncated to retain only the collection date.  
New temporal features — Year, Month, and Season — were then derived to support time-based aggregation and potential climate trend exploration.

# **Step 2: Parameter Selection and Dataset Restructuring**

The dataset was originally structured in a *long format*, where each row represented a single parameter measurement per sample.  
To enable multivariate analysis across physicochemical indicators, the data were reshaped into a *wide format* using pandas.pivot\_table(), producing one row per sample with columns representing individual water-quality parameters.

Parameters were selected based on their relevance to the Water Quality Index (WQI) and sensitivity to climate-related influences, such as nutrient enrichment and thermal variability.  
The final selection included:

* pH (Field)
* Temperature
* Dissolved Oxygen (Field)
* Conductivity (Field)
* Total Nitrogen
* Total Phosphorus
* Nitrate Nitrogen
* Orthophosphate Phosphorus

These variables are widely recognised as core indicators of water quality and are directly affected by climate-driven processes such as eutrophication, oxygen depletion, and seasonal thermal stratification.[1][2][3]

# **Step 3: Handling missing data and outlier removal**

Outliers across all parameters were filtered using the interquartile range (IQR) method (1.5×IQR criterion). This ensured that extreme values likely arising from sensor errors or rare sampling artefacts did not bias subsequent analyses.

KNN imputation was applied selectively to the *Temperature* parameter only (n\_neighbors=5). This decision was based on its low missingness (≈12%) and strong correlation with other environmental indicators such as dissolved oxygen and conductivity. The remaining parameters were excluded from imputation due to high structural missingness (>35%), which would make KNN unreliable and artificially inflate inter-feature similarity.[4]

# **Step 4: Investigation of Pre-1990 Data Gap and Temperature Spike**

Analysis of the dataset revealed a near absence of records prior to 1990. This discontinuity does not reflect an environmental phenomenon but rather a structural issue within the data itself. Earlier monitoring campaigns were less systematic and often lacked full parameter coverage, resulting in incomplete historical records.[5]

Several factors likely contributed to this pattern:

* Early monitoring programmes recorded fewer parameters and used inconsistent formats.
* Legacy datasets frequently included non-numeric entries such as “ND” or “<0.1,” which were interpreted as missing values during data import.
* The reshaping of data from long to wide format introduced listwise deletion, where samples missing any of the selected parameters were automatically excluded.[6].

As a result, only data from 1990 onwards can be considered reliable for temporal and multivariate analyses. The apparent absence of pre-1990 records therefore arise from structural missingness and digitisation bias rather than changes in water quality.

# **References**

[1]. Whitehead et al. (2009) – *A review of the potential impacts of climate change on surface water quality.*

[2]. Jane et al. (2021) – *“Widespread deoxygenation of temperate lakes*

[3]. EPA (2023) – *Climate Change Indicators: Water Quality.*

[4]. Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, *17*(5–6), 519–533. https://doi.org/10.1080/713827181

[5]. Helsel, D. R., & Hirsch, R. M. (2002). *Statistical methods in water resources.* U.S. Geological Survey.

[6]. Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley.