Preprocessing: Water Treatment Plants

In this assignment, we worked on preprocessing data from the MERKUR project, which focuses on analyzing and optimizing water treatment plants in Denmark. The dataset, provided by VIA University College, contained various challenges such as missing values, outliers, and a mix of categorical and numerical data.

Key Tasks Performed

1. **Data Exploration**:
    - Loaded and explored the dataset to understand the features, distributions, and potential issues.
    - Identified columns with missing values and evaluated their significance based on the percentage of missing data.
2. **Data Cleaning**:
    - Removed irrelevant columns, such as IDs and other non-informative features.
    - Dropped or imputed missing values in columns depending on their importance and the extent of missing data.
    - Handled outliers by identifying and replacing or removing extreme values that could distort analysis.
3. **Feature Engineering**:
    - Applied one-hot encoding to categorical variables, ensuring they were usable in machine learning models.
    - For specific features like "Stages," evaluated alternatives to one-hot encoding to preserve meaningful relationships.
    - Scaled numeric data to standardize features and improve model performance.
4. **Correlation Analysis**:
    - Created a correlation matrix to identify relationships between features.
    - Based on the analysis, dropped highly correlated features to reduce redundancy and potential multicollinearity.
5. **Transformations and Observations**:
    - Applied transformations (e.g., logarithmic or square root) to skewed data to improve its distribution.
    - Discussed the implications of having only 80 rows in the dataset, highlighting challenges in building robust machine learning models.
6. **Finalization**:
    - Declared the dataset clean and ready for machine learning after completing the preprocessing steps.
    - Documented all preprocessing choices and their rationale to ensure reproducibility.

This assignment allowed us to work through real-world data preprocessing challenges, preparing a complex dataset for analysis and machine learning.