## The report contains:

• Give your project a context- Story telling.

• A detailed description of the variables originally present in the dataset

o How many observations per variable?

o How many missing values? Express it in percentage of the total observations

o What type of variable?

o How many duplicates are present?

o What are the unique values present per variable?

o What are the frequencies of each unique observation? (when applicable)

o What are the statistical properties of each variable?

• Formulate a problem statement about the dataset.

o What do you intend to do with the dataset?

o Is it a regression type of analysis?

o Is it a classification type analysis?

o Both types can be applied to the same dataset

• A detailed strategy of selecting the variables.

o Why did you keep some variables?

o Why did you eliminate some variables?

o How and why did you decide to fill the missing values in some variables?

o How and why did you decide to handle the duplicates?

o How and why did you decide to apply a certain approach to handle extreme values?

• Choose at least five graphics to represent best your data and strategy.

o Each graphic should have a title and comments attached to it.

o Graphics can either show the relationship between variables or between many variables or the statistical properties of the target variable.

## Task 1

Give your project a context- Story telling.

## Solution

## Background

In the year 2023, a team of environmental scientists embarked on a mission to analyze water quality data collected from various monitoring stations across the States. The data spanned five years, from 2002 to 2016, and included measurements of key water quality parameters such as temperature, dissolved oxygen, pH, conductivity, biochemical oxygen demand, nitrate and nitrite nitrogen, fecal coliform bacteria, and total coliform bacteria.

The scientists were particularly interested in understanding the spatial and temporal variations in water quality across different regions and station types. They hoped to identify potential pollution sources, assess the overall health of aquatic ecosystems, and ultimately inform strategies for water quality management and protection.

## Task 2

A detailed description of the variables originally present in the dataset

## Solution

## Feature Explanation

1. **STATION CODE**:

- This feature represents a unique identifier assigned to each monitoring station where water quality measurements are taken. Each code corresponds to a specific sampling point.

2. **LOCATIONS**:

   - This feature contains the names of the locations where water samples are collected for quality assessment. It provides information about the specific places where the water quality measurements are recorded.

3. **STATE**:

   - This feature denotes the state where the monitoring station is situated. It indicates the geographical division or state within the country.

4. **Temp**:

   - This feature represents the water temperature measured at the sampling point. It indicates the degree of hotness or coldness of the water at the time of measurement.

5. **D.O. (mg/l)** (Dissolved Oxygen):

   - This feature represents the concentration of dissolved oxygen in the water, typically measured in milligrams per liter (mg/l). Dissolved oxygen is crucial for aquatic organisms and serves as an indicator of water quality.

6. **PH**:

   - pH is a measure of the acidity or alkalinity of the water. It indicates the hydrogen ion concentration and is measured on a scale of 0 to 14. A pH of 7 is considered neutral, while values below 7 are acidic and above 7 are alkaline.

7. **CONDUCTIVITY (µmhos/cm)**:

   - Conductivity is a measure of water's ability to conduct an electrical current. It is influenced by the presence of dissolved solids and ions. Conductivity is measured in microsiemens per centimeter (µS/cm) or micromhos per centimeter (µmhos/cm).

8. **B.O.D. (mg/l)** (Biochemical Oxygen Demand):

   - B.O.D. represents Biochemical Oxygen Demand, which measures the amount of dissolved oxygen required by microorganisms to break down organic matter in water. It is measured in milligrams per liter (mg/l).

9. **NITRATENAN N+ NITRITENANN (mg/l)**:

   - This feature may represent the combined concentration of nitrate and nitrite in water, measured in milligrams per liter (mg/l). Nitrate and nitrite are forms of nitrogen compounds and can be indicators of water pollution.

10. **FECAL COLIFORM (MPN/100ml)**:

    - Fecal coliform is a type of bacteria found in the feces of warm-blooded animals. Its presence in water indicates

contamination by fecal matter. The concentration is typically measured in Most Probable Number per 100 milliliters (MPN/100ml).

11. **TOTAL COLIFORM (MPN/100ml)Mean**:
    - Total coliform represents a group of bacteria, including fecal coliform, found in the environment. It's an indicator of overall water quality. The concentration is measured in Most Probable Number per 100 milliliters (MPN/100ml).
12. **year**:
    - This feature denotes the year in which the water quality measurements were recorded. It indicates the temporal aspect or the year of the observation for the corresponding water quality data.

## Task 3

How many observations per variable?

## Solution

We have a total of 1991 observations per variables with 12 features

## Task 4

How many missing values? Express it in percentage of the total observations

## Solution

We had about 11 features with missing values, The above shows the percentage representative of each features making state the highest with over 38.221999

## task 5

What type of variable ?

## Solution

The following are type of variables with missing values values

**STATE**
**Fec_col**
**NI**
**LOCATIONS**
**Tot_col**
**STATION CODE**
**Temp**
**BOD**
**DO**
**Conductivity**
**PH**

Excluding

## Task 6

How many duplicates are present?

## Solution

```
Duplicate Rows:
STATION CODE    0.0
LOCATIONS       0.0
STATE           0.0
Temp            0.0
DO              0.0
PH              0.0
Conductivity    0.0
BOD             0.0
NI              0.0
Fec_col         0.0
Tot_col         0.0
year            0.0
```

```
dtype: float64
Empty DataFrame
Columns: [STATION CODE, LOCATIONS, STATE, Temp, DO, PH, Conductivity, BOD, NI, Fec_col,
Tot_col, year]
Index: []

```

The output shows that there is absence of duplicate rows within the dataset containing water quality information across various parameters

## task 7

What are the unique values present per variable?

## Solution

```
STATION CODE     321
LOCATIONS        691
STATE             24
Temp             177
DO               165
PH               265
Conductivity    1004
BOD              407
NI               506
Fec_col          868
Tot_col         1093
year              12
dtype: int64

They above result display the unique values per variable in the datasets
```

## task 8

What are the frequencies of each unique observation? (when applicable)

## Solution

The frequencies of a unique observation can only be applicable when variable has a categorical data points

### task 9

What are the statistical properties of each variable?

### Solution

```
In [69]: 1 df_wqi.describe().T
```

Out[69]:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Temp** | 1862.0 | 26.243414 | 3.236175 | 10.000000 | 25.000000 | 27.00000 | 28.200000 | 35.000000 |
| **DO** | 1862.0 | 6.470459 | 1.246170 | 0.000000 | 6.000000 | 6.70000 | 7.200000 | 10.000000 |
| **PH** | 1862.0 | 21.391149 | 125.697450 | 0.000000 | 6.900000 | 7.30000 | 7.700000 | 3384.000000 |
| **Conductivity** | 1862.0 | 914.274560 | 2554.122951 | 3.700000 | 76.000000 | 169.15000 | 446.000000 | 18291.000000 |
| **BOD** | 1862.0 | 3.901351 | 6.993561 | 0.100000 | 1.100000 | 1.80000 | 3.400000 | 88.000000 |
| **NI** | 1862.0 | 1.091084 | 1.578783 | 0.000000 | 0.270000 | 0.51600 | 1.017500 | 12.150000 |
| **Fec_col** | 1862.0 | 2157.399943 | 8085.575387 | 0.000000 | 37.000000 | 221.00000 | 580.250000 | 150250.000000 |
| **Tot_col** | 1862.0 | 6974.150913 | 39355.379681 | 0.000000 | 112.250000 | 468.00000 | 1628.750000 | 967500.000000 |
| **year** | 1862.0 | 2010.067669 | 3.044138 | 2003.000000 | 2008.000000 | 2011.00000 | 2013.000000 | 2014.000000 |
| **WQI** | 1862.0 | 373.008553 | 2187.137257 | -65.334452 | 55.211945 | 71.74015 | 101.601333 | 58680.748613 |

There are lots of high variance in the dataset that needs to be normalise

### Task 10

• Formulate a problem statement about the dataset.

# Problem Statement:

Understanding the Temporal and Geographical Variations in Water Quality Parameters and Evaluating Water Pollution Contributing Factors.

### task 11

o What do you intend to do with the dataset?

## Solution

To carry out a prediction to know the water quality index for each observations when introduce to new data records

## task 12

o Is it a regression type of analysis?

## Solution

After i introduce feature engineering to calculate the water quality index, this features tens to be my target variable for each observation. also the feature is a continuous numerical datapoint. in conclusion its a regression type of analysis

## task 13

Is it a classification type analysis?

## Solution

No. its not a classification type analysis, as it doesn't require a binary or multiclass type of analysis or prediction

## task 14

Both types can be applied to the same dataset?

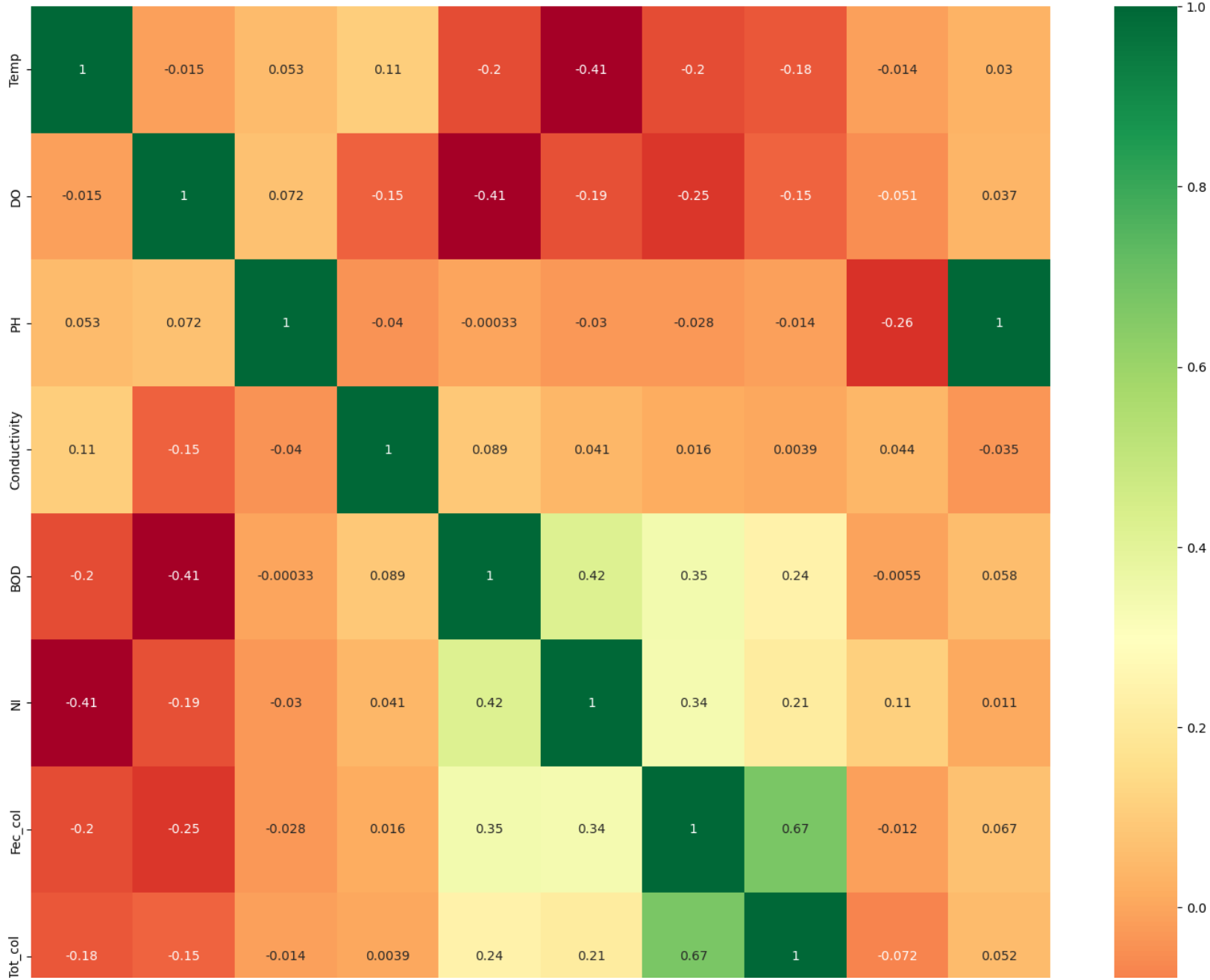
## Solution

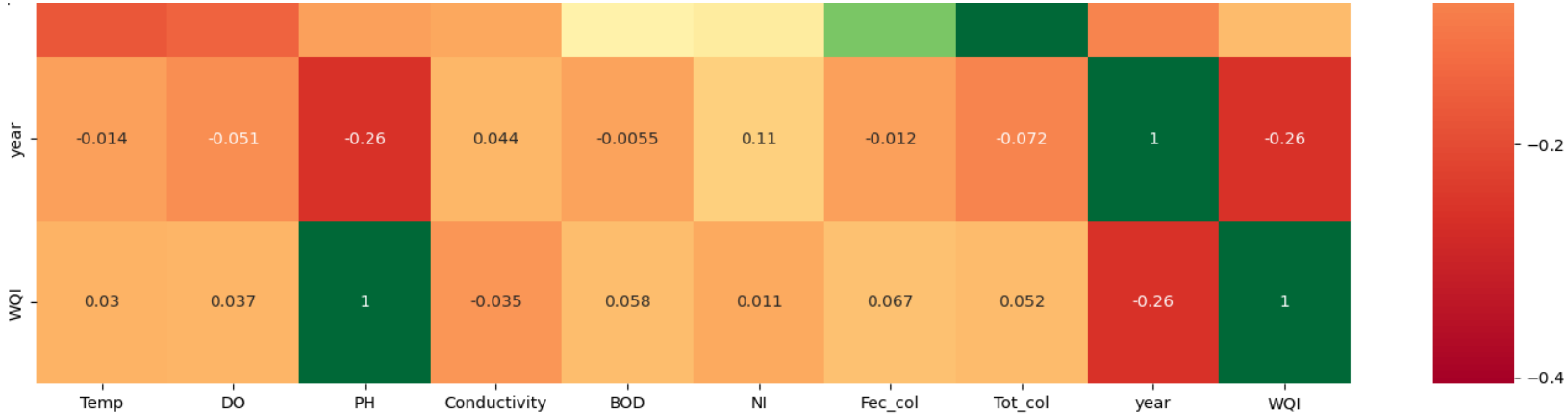
No. we can only apply regression type of analysis to the dataset

## task 15

A detailed strategy of selecting the variables.

## Solution


Temp
DO
PH
Conductivity
BOD
NI
Fec_col
Tot_col
1 -0.015 0.053 0.11 -0.2 -0.41 -0.2 -0.18 -0.014 0.03
-0.015 1 0.072 -0.15 -0.41 -0.19 -0.25 -0.15 -0.051 0.037
0.053 0.072 1 -0.04 -0.00033 -0.03 -0.028 -0.014 -0.26 1
0.11 -0.15 -0.04 1 0.089 0.041 0.016 0.0039 0.044 -0.035
-0.2 -0.41 -0.00033 0.089 1 0.42 0.35 0.24 -0.0055 0.058
-0.41 -0.19 -0.03 0.041 0.42 1 0.34 0.21 0.11 0.011
-0.2 -0.25 -0.028 0.016 0.35 0.34 1 0.67 -0.012 0.067
-0.18 -0.15 -0.014 0.0039 0.24 0.21 0.67 1 -0.072 0.052
1.0
0.8
0.6
0.4
0.2
0.0




I use Heat map to visually representing the relationships and correlations between features in a dataset to aid in the selection of relevant variables.

Using a heatmap in feature selection involves visually representing the relationships and correlations between features in a dataset to aid in the selection of relevant variables. Here's a summary of the process:

**Correlation Visualization:**

- **Heatmap Representation:** Utilize a heatmap, often a colored matrix, to display correlation coefficients between different features in a dataset.
- **Color Gradients:** Assign color gradients to represent the strength and direction of correlations (positive or negative) between features.

**Identifying Correlated Features:**

- **Visual Inspection:** Analyze the heatmap to identify clusters or patterns of high correlation between features.
- **Threshold Selection:** Set a correlation threshold to identify features that exhibit significant correlations with one another.

**Feature Selection Criteria:**

- **Highly Correlated Features:** Consider removing one from a pair of highly correlated features to mitigate multicollinearity and reduce redundancy.
- **Relevant Features:** Prioritize features that show strong correlations with the target variable or with other key variables of interest.

**Considerations and Interpretation:**

- **Domain Knowledge:** Complement heatmap analysis with domain-specific understanding to interpret and select features effectively.
- **Iterative Process:** Feature selection using a heatmap might involve multiple iterations to refine choices and account for

## task 16

Why did you keep some variables?

## Solution

When we use heatmap for feature selection, the decision to keep certain variables is based on several factors related to their relevance, significance, and relationships with other variables in the dataset. Here's why some variables was being kept in the dataset:

**High Correlation with Target Variable:**

- Variable like PH that exhibit strong correlations with the target variable or outcome of interest is being retained. These feature have a direct influence on predicting the target and is crucial for model performance.

**Independently Informative Features:**

- variables 'Temp', 'D.O. (mg/l)', 'PH', 'CONDUCTIVITY (µmhos/cm)', 'B.O.D. (mg/l)', 'NITRATENAN N+ NITRITENANN (mg/l)', 'FECAL COLIFORM (MPN/100ml)', 'TOTAL COLIFORM (MPN/100ml)Mean', 'year' does not exhibit high correlations with other variables but are independently informative for the analysis or predictive modeling. They provide unique information not captured by other features.

**Multicollinearity Mitigation:**

- In cases of multicollinearity (high correlations between multiple features), keeping a subset of highly correlated features while removing others helps mitigate redundancy and multicollinearity issues. These retained variables might represent essential aspects of the data.

**Strategic Model Complexity:**

- Keeping a subset of variables helps in controlling model complexity. Removing redundant or less informative features streamlines the model and prevents overfitting, leading to a more interpretable and robust model.

In essence, the decision to keep certain variables while using a heatmap for feature selection involves a balance between their individual importance, relationships with other variables, domain relevance, and their contribution to achieving the desired analysis

## task 17

Why did you eliminate some variables?

## Solution

In the context of using a heatmap for feature selection, elimination of certain variables is necessary to enhance model performance, reduce complexity, and improve the interpretability of the analysis.

Though, in the given dataset, there is no feature that shows less importance to the target variables.

This are some reasons why some variables might be eliminated in other datasets.:

1**High Multicollinearity:**

- When variables are highly correlated with each other, retaining all of them can introduce multicollinearity issues. Removing one of the highly correlated variables helps mitigate redundancy and stabilizes model estimation.

**Low Correlation with Target Variable:**

- Variables that exhibit weak correlations with the target variable or outcome might not contribute significantly to predicting the target. Eliminating such variables can streamline the model and improve predictive accuracy.

**Redundancy and Overfitting Prevention:**

- Redundant or unnecessary variables add complexity to the model without providing substantial information gain. Removing such variables helps prevent overfitting and enhances model generalization to new data.

**Noise Reduction:**

- Variables with random or insignificant relationships with other features or the target variable add noise to the model. Eliminating noisy variables improves the signal-to-noise ratio and model robustness.

**Dimensionality Reduction:**

- High-dimensional datasets with excessive variables can lead to the curse of dimensionality, making analysis and model

## task 18

How and why did you decide to fill the missing values in some variables?

## Solution

**Why Fill Missing Values:**

```
Filling missing values helps maintain the integrity of the dataset, preventing the loss of
information that might be valuable for analysis.

Many machine learning algorithms cannot handle missing values. Imputation allows for the u
se of these algorithms without discarding valuable data.

Complete datasets contribute to a larger sample size, providing more statistical power and
potentially improving the robustness of analysis or models.

Filling missing values can enhance interpretability by providing a more comprehensive data
set for analysis and decision-making.
```

**How did i implement Filling Missing Values:**

I introduce Data Imputation Techniques:

By using Median to Impute For numerical variables, then mode for categorical variables

## task 19

How and why did you decide to handle the duplicates?

## Solution

**Why I Handle Duplicates in the datasets is because:**

Eliminating duplicates helps ensure data accuracy and integrity, preventing overrepresentation of certain observations.

Duplicate entries can skew statistical analysis or machine learning models, leading to biased estimates or predictions.

Large numbers of duplicates increase computational burden and processing time, especially in complex analyses.

Many algorithms may give undue importance to duplicated instances, affecting model training and performance negatively.

**How I Handle Duplicates in the dataset:**

Use decided to use functions base method to detect duplicate rows or observations within the dataset.

Identify duplicate entries based on specific columns across the entire dataset.

When implementing this method, i found out that there was no duplicate in the datasets

## task 20

How and why did you decide to apply a certain approach to handle extreme values?

## Solution

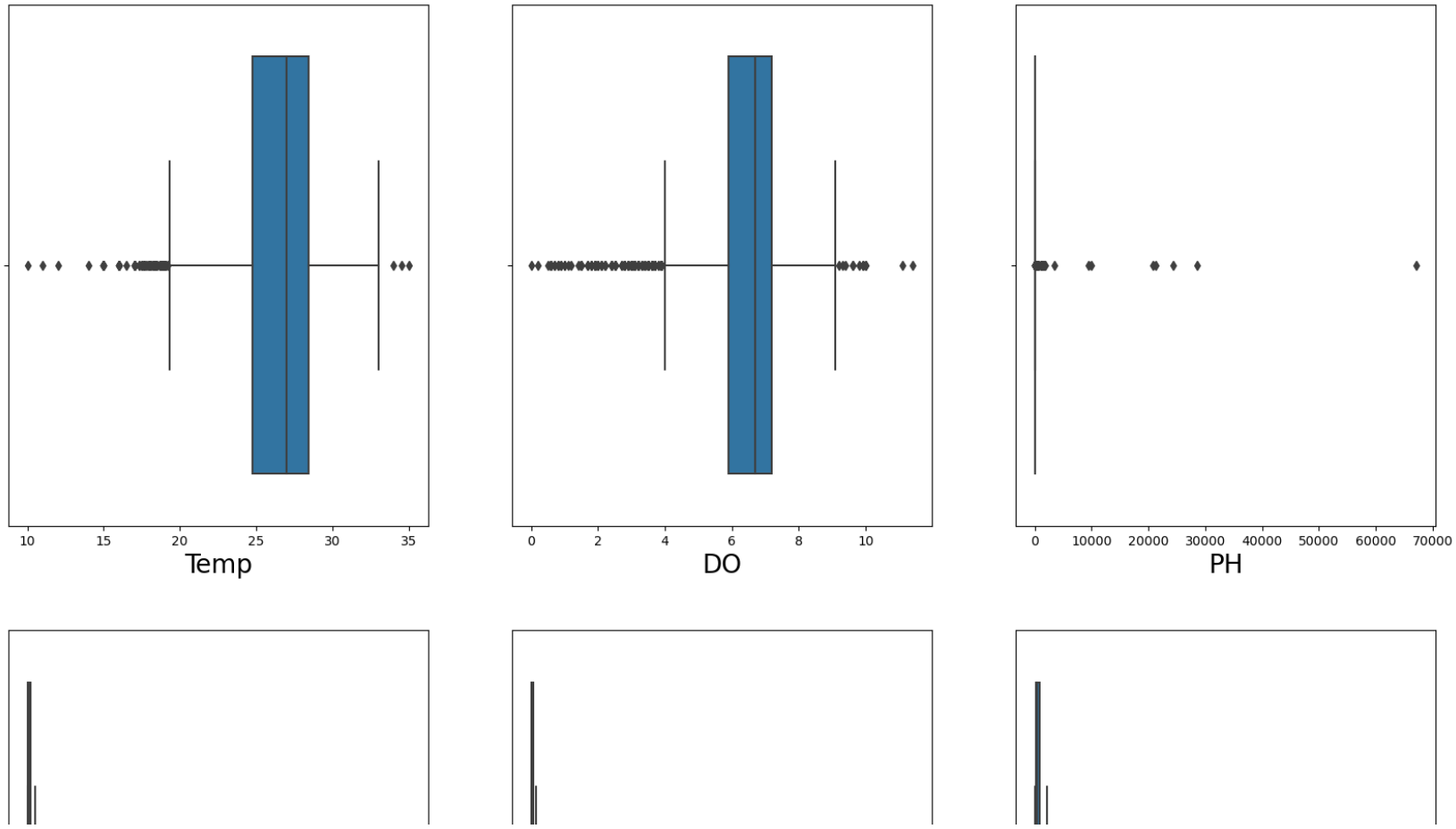
**Why Handle Extreme Values:**

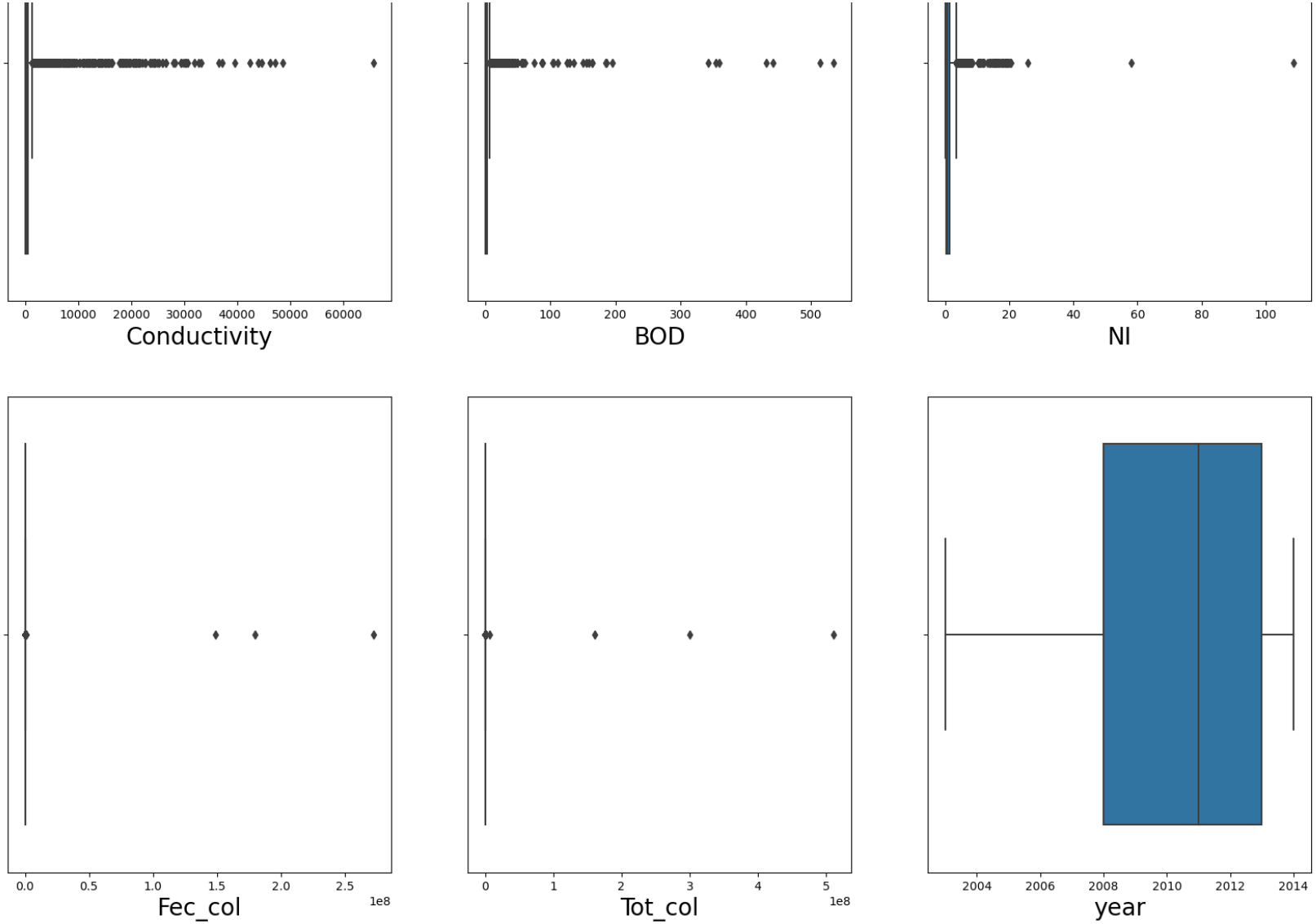
The main i decied to handle Extreme Values are:

```
Extreme values can adversely affect data integrity, leading to biased estimations or misle
ading analysis results.

Outliers might obscure patterns or trends in the data. Addressing extreme values enhances
interpretability by focusing on the underlying data distribution.

Outliers can bias statistical estimates, such as means and variances, affecting the validi
ty of conclusions drawn from the data.
```

**How to Decide Handling Extreme Values:**

I Ploted boxplots to visually identify observations that significantly deviate from the majority of the data.






I use the Z-score to measures how many standard deviations a particular data point is from the mean of the dataset.

Here's how the Z-score method works for outlier detection:

For each data point in a numerical variable, compute its Z-score using the formula: Z=(X−μ)σZ=σ(X−μ)

Where XX is the data point, μμ is the mean of the variable, and σσ is the standard deviation.

I then Determine a threshold value for Z-scores. Common thresholds are often around +2.5+2.5 or +3+3, indicating data points that

In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```