Project Overview

This project analyzes the potability of water based on various physicochemical properties using **Exploratory Data Analysis (EDA)** techniques. The dataset consists of 3,276 water samples with attributes such as **pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity** to determine whether the water is potable (drinkable) or non-potable.

Objectives

- Understand the distribution and characteristics of water quality parameters.
- Handle missing data and ensure data consistency.
- Visualize relationships between different attributes.
- Identify potential insights that could help in determining potability.

Data Preprocessing

- **Missing Values:** Found missing values in pH, Sulfate, and Trihalomethanes columns.
- **Imputation:** Replaced missing values with the mean of respective columns.
- **Data Type Check:** Confirmed all columns have appropriate data types.

Key Findings

1. pH Analysis

- The pH values are **normally distributed** around 7.
- There are **outliers** on both the acidic and basic ends (below 4 and above 10).

2. Potability Distribution

- **Non-potable water samples (0):** 1,998
- **Potable water samples (1):** 1,278
- The dataset is **imbalanced**, with more non-potable samples.

3. Correlation Analysis

- Weak correlations between most variables, indicating that **no single parameter alone determines potability**.
- A **slight positive correlation** between Conductivity and Sulfate.

4. Visualization Insights

- **Box Plots & Violin Plots**: Show differences in pH distribution for potable vs. non-potable water.
- **Heatmap**: No strong relationships between features.
- **Scatter Plot (pH vs Conductivity)**: No clear separation between potable and non-potable water.

Conclusion

- **Most potable water samples have pH values in the range of 6-8.**
- **Non-potable water has more extreme values in pH and other attributes.**
- **Feature relationships are weak**, meaning water potability depends on multiple combined factors rather than a single parameter.
- Further analysis, such as machine learning classification models, could provide better predictions for potability.

Future Improvements

- **Feature Engineering:** Create additional attributes to improve potability classification.
- **Outlier Treatment:** Apply techniques like IQR-based filtering or transformation.
- **Machine Learning:** Implement classification models to predict water potability.
- **Data Balancing:** Handle the class imbalance in potable vs. non-potable samples.