DATA ANALYSIS WITH COGNOS GROUP 2

PROJECT 10 – WATER ANALYSIS



**COLLEGE CODE:5113**

**TEAM 10**

**TEAM LEADER :**  INIYAN P – au511321104032

**TEAM MATES :**  KAMALESH L - au511321104039

DHANUSH M – au511321104021

MUKESH C - au511321104057

KUGAN K - au511321104044

WATER QUALITY ANALYSIS

**Phase 1 :**

**Project Definition and Design Thinking :**

**explanation of water quality analysis :**

## Introduction:

Water quality analysis involves testing and evaluating the characteristics of water to determine its safety and suitability for various purposes. It includes assessing parameters such as pH, dissolved oxygen, turbidity, and the presence of contaminants like bacteria, chemicals, and heavy metals. This analysis helps in ensuring that water meets the required standards for drinking, industrial use, and environmental protection.

Start by introducing the importance of water quality analysis. Explain the significance of monitoring and ensuring clean water for various purposes, such as drinking, agriculture, and aquatic ecosystems.

## *Goal:*

State the specific goal of your analysis. For example, your goal might be to assess the water quality of a local river and identify potential contaminants.

Methods:

Describe the methods and data sources you will use for the analysis. This may include collecting water samples, using sensors, or utilizing existing water quality datasets. Mention any relevant parameters you plan to measure, like pH, dissolved oxygen, turbidity, or specific contaminants.

**1. What is Water Quality Analysis?**

Water quality analysis is the process of evaluating the physical, chemical, and biological characteristics of water to determine its suitability for various purposes, such as drinking, recreational activities, or industrial use.

2. Importance of Water Quality Analysis

- Public Health: Ensuring safe drinking water is critical to prevent waterborne diseases.

- Environmental Protection: Maintaining aquatic ecosystems and biodiversity.

- Industrial and Agricultural Use: Assessing water quality for industrial and agricultural processes.

3. Parameters Analyzed

Water quality analysis examines several key parameters, including:

- Physical Parameters: Temperature, turbidity, color, odor, and taste.

- Chemical Parameters: pH, dissolved oxygen, nutrients (nitrates, phosphates), heavy metals, and organic contaminants.

- Biological Parameters: Presence of microorganisms (bacteria, algae), and indicators of fecal contamination (E. coli).

4. Sampling and Testing

- Sample Collection: Collect water samples from various sources, considering depth, location, and frequency.

- Laboratory Testing: Analyze samples using specialized equipment and chemical tests to measure parameters.

5. Regulatory Standards

- Water Quality Standards: Government agencies set regulatory standards to ensure water quality is safe for human consumption and ecosystem health.

- Legal Requirements: Facilities discharging pollutants must meet certain standards to protect water bodies.

6. Challenges in Water Quality Analysis

- Natural Variation: Water quality can fluctuate due to environmental factors, making consistent analysis challenging.

- Emerging Contaminants: Identifying and measuring new contaminants in water supplies.

- Infrastructure Monitoring: Aging infrastructure can lead to water quality issues.

7. Application of Results

The data obtained from water quality analysis is used to make informed decisions:

- Treatment Processes: Adjusting water treatment methods to remove specific contaminants.

- Environmental Management: Protecting and restoring aquatic ecosystems.

- Public Awareness: Informing the public about water safety and potential health risks.

8. Continuous Monitoring

Water quality analysis is an ongoing process, with regular monitoring to ensure that water remains safe and meets established standards. In summary, water quality analysis is a comprehensive process that evaluates the condition of water sources, making it possible to protect public health, the environment, and support various industries.

# Python :

Here's a simplified example of Python code to perform basic water quality analysis using some hypothetical data:

# code :

import matplotlib.pyplot as plt

# Load water quality data

water\_data = pd.read\_csv('water\_quality\_data.csv')

# Data preprocessing (cleaning and filtering)

# ...

# Calculate basic statistics

mean\_ph = water\_data['pH'].mean()

mean\_do = water\_data['Dissolved\_Oxygen'].mean()

# Create visualizations

plt.figure(figsize=(10, 6))

plt.scatter(water\_data['Date'], water\_data['pH'], label='pH')

plt.scatter(water\_data['Date'], water\_data['Dissolved\_Oxygen'], label='Dissolved Oxygen')

plt.xlabel('Date')

plt.ylabel('Value')

plt.legend()

plt.title('Water Quality Analysis')

plt.show()

# Perform contaminant analysis

# ...

# Generate a report or save the results

# ...

Explain what kind of output you expect to obtain from your analysis. This could be visualizations, statistical summaries, or a comprehensive report detailing water quality findings, potential issues, and recommendations.Remember to adapt the code and methods to your specific water quality analysis needs. Real water quality analysis projects can be much more complex and may involve more advanced statistical and machine learning techniques, depending on the scope of your analysis.

**PHASE-2**

**INNOVATION**

**INTRODUCTION:**

Water quality analysis is a critical aspect of environmental science and public health that seeks to assess and monitor the physical, chemical, and biological characteristics of water to ensure its suitability for various purposes. The quality of water has a profound impact on human health, ecosystems, and industrial processes, making it essential to understand and manage. This multifaceted field involves the collection, testing, and interpretation of data to determine the presence and concentration of various substances in water, such as pollutants, nutrients, minerals, and microorganisms.

Water quality analysis serves a variety of purposes, including the assessment of drinking water safety, the monitoring of aquatic ecosystems, the evaluation of wastewater treatment efficacy, and the identification of potential health hazards. It also plays a critical role in ensuring compliance with environmental regulations and standards.

This introduction sets the stage for a comprehensive exploration of the methods, significance, and applications of water quality analysis, highlighting its pivotal role in safeguarding our natural resources and protecting public health. In this context, it becomes evident that a thorough understanding of water quality is crucial for sustainable environmental management and responsible resource utilization.

PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–

6.83 which are in the range of WHO standards.

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.



Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine

levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

## COLUMNS THAT WE USED:

⦁ The water\_potability.csv file contains water quality metrics for 3276 different water bodies.

## pH value:

PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from

6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

Hardness:

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

Solids (Total dissolved solids - TDS):

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals

produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

## Chloramines:

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

## Sulfate:

Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

## Conductivity:

Pure water is not a good conductor of electric current rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to

transmit current. According to WHO standards, EC value should not

exceeded 400 μS/cm.

## Organic\_carbon:

Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

## Trihalomethanes:

THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

## Turbidity:

The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

## DETAILS OF LIBRARIES USED AND WAY TO DOWNLOAD:

In our water analysis forecasting project, we leverage a range of powerful libraries to facilitate data manipulation, model development, and evaluation. These libraries play a crucial role in transforming our design into an innovative solution. Here are the key libraries and a detailed explanation of how to download and install them:

**Pandas:**

⦁ Download: Pandas can be installed using Python's package manager, pip. Open a command prompt or terminal and run pip install pandas.

⦁ Installation: After downloading, Pandas can be imported in your Python script using import pandas as pd.

## 

## NumPy:

⦁ **Download**: NumPy can also be installed using pip. Run pip install numpy in your command prompt or terminal.

⦁ **Installation**: Import NumPy in your Python script using import numpy as np.

Matplotlib and Seaborn:

⦁ **Download**: These visualization libraries can be installed with pip. Use pip install matplotlib seaborn in your command prompt or terminal.

⦁ **Installation**: Import Matplotlib and Seaborn in your Python script using import matplotlib.pyplot as plt and import seaborn as sns, respectively.

Scikit-Learn:

⦁ **Download**: Scikit-Learn can be installed with pip. Run pip install scikit-learn in your command prompt or terminal.

⦁ **Installation**: Import Scikit-Learn in your Python script using import sklearn.

Keras with TensorFlow Backend:

⦁ **Download:** Install Keras with TensorFlow using pip install tensorflow keras in your command prompt or terminal.

⦁ **Installation**: Import Keras in your Python script using import keras.

Jupyter Notebooks:

⦁ Download: You can install Jupyter Notebooks by running pip install jupyter in your command prompt or terminal.

⦁ Installation: After installation, start a Jupyter Notebook server by running jupyter notebook in your command prompt or terminal. This will open a web-based interface for creating and running notebooks.

Once these libraries are downloaded and installed, you can import them into your Python scripts to access their functionality. These libraries provide a robust ecosystem for data analysis, machine learning, and data visualization, essential for transforming your design into an innovative solution for water analysis forecasting.

## HOW TO TRAIN AND TEST:

**Training and testing a water quality analysis dataset involves using machine learning or statistical methods to build a model that can predict or classify water quality parameters based on :**

### Data Collection:

Gather a comprehensive and representative dataset containing information on water quality parameters. This dataset should include features such as temperature, pH, turbidity, dissolved oxygen, chemical concentrations (e.g., nitrate, phosphate, heavy metals), and biological indicators (e.g., E. coli counts).

Ensure the dataset is labeled, meaning it includes the target variable you want to predict or classify (e.g., water quality categories like 'clean,' 'polluted,' or specific parameter concentrations).

**Data Preprocessing**:

Clean the dataset by handling missing values and outliers. This may involve imputation or removal of problematic data points.

Normalize or standardize numerical features to bring them to a common scale.

Encode categorical variables if needed (e.g., water source type) into numerical values.

## Data Splitting:

- Divide the dataset into two subsets: a training set and a testing set. The typical split is around 70-80% for training and 20-30% for testing.

Feature Selection/Engineering:

- Analyze the importance of features. You may choose to select the most relevant features or engineer new features if needed.

## Model Selection:

- Choose an appropriate machine learning algorithm based on your problem. Common algorithms for water quality analysis include decision trees, random forests, support vector machines, and neural networks.

## Model Training:

- Use the training dataset to train the chosen model. The model will learn patterns in the data to make predictions or classifications.

## Model Evaluation:

Evaluate the model's performance using the testing dataset.

Common evaluation metrics for classification tasks include accuracy, precision, recall, F1 score, and for regression tasks, metrics like mean squared error (MSE) or R-squared.

Perform cross-validation if you have a limited dataset to ensure the model's generalization.

## Tuning Hyperparameters:

- Optimize the model's hyperparameters through techniques like grid search or random search to improve its performance.

## 

## Interpret Results:

- Analyze the model's predictions to gain insights into water quality trends, identify influential factors, and understand the model's limitations.

## Deployment:

- Once you are satisfied with the model's performance, you can deploy it to make predictions or classifications on new, unseen data. This might involve integrating the model into a water quality monitoring system.

**Continuous Monitoring and Maintenance:**

Regularly update the model as more data becomes available or as water quality standards change.

Remember that the choice of modeling technique and the quality of your data are critical factors in the success of water quality analysis. Furthermore, interpretability and domain knowledge are crucial for understanding the implications of the model's predictions in the context of water quality management.

## 

## METRICS USED FOR ACCURACY CHECK:

In water quality analysis, the choice of evaluation metrics depends on the specific task you are trying to accomplish. Here are some common metrics used to check the accuracy of water quality analysis models:

## Accuracy:

## Accuracy is a general metric used for classification tasks. It measures the proportion of correctly classified samples out of the total samples in the dataset. While it provides a simple and easy-to- understand measure of overall model performance, it may not be the best choice if the dataset is imbalanced (e.g., if there are many more "clean" water samples than "polluted" ones), as it can be misleading in such cases.

## Precision: Precision is the ratio of true positive predictions to the total number of positive predictions (true positives + false positives). It is particularly important when you want to minimize false positives, such as in cases where classifying water as polluted when it's not could lead to unnecessary actions or costs.

## Recall (Sensitivity): Recall measures the ratio of true positive predictions to the total number of actual positive samples (true positives + false negatives). It is crucial when you want to avoid false negatives, such as when failing to detect actual pollution in water could have severe consequences.

## F1 Score: The F1 score is the harmonic mean of precision and recall, and it provides a balance between these two metrics. It's particularly useful when you want to strike a balance between minimizing false positives and false negatives.

## Specificity: Specificity is the ratio of true negative predictions to the total number of actual negative samples (true negatives + false positives). It is essential when you want to minimize false positives, similar to precision, but for the negative class.

## Mean Absolute Error (MAE): MAE is a common metric for regression tasks in water quality analysis. It measures the average absolute difference between the predicted values and the actual values of the water quality parameter. It is easy to interpret, with lower values indicating better accuracy.

## Root Mean Squared Error (RMSE):RMSE is another metric for regression tasks, and it measures the square root of the average of the squared differences between the predicted and actual values. RMSE gives more weight to larger errors and is sensitive to outliers.

## 

## Coefficient of Determination (R-squared or R²): R-squared quantifies the proportion of the variance in the water quality parameter that is explained by the model. An R-squared value close to 1 indicates a good fit, while a value close to 0 suggests the model does not explain much of the variance.

## Cohen's Kappa: Cohen's Kappa is a metric that takes into account the agreement between the model's predictions and what would be expected by chance. It is often used when dealing with imbalanced datasets.

The choice of metrics should align with the specific goals of your water quality analysis project. It's important to consider the context, potential consequences of false positives and false negatives, and the nature of the dataset when selecting the most appropriate evaluation metrics.

# PHASE 3 : INNOVATION

## Necessary step to follow:

## 1.Import Libraries:

Start by importing the necessary python libraries:

Import numpy as np # linear algebra

Import pandas as pd # data processing, CSV file I/O (e.g. pd.read\_csv)

Import seaborn as sns

Import matplotlib.pyplot as plt

Import plotly.express as px

Import missingno as msno

## 2.Load the Dataset:

Load your dataset into a Pandas DataFrame. Ensure that the data is in a format that Pandas can work with, such as CSV or Excel.

Data = pd.read\_csv(‘water\_quality\_data.csv’)

Preprocessing the dataset

Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate result

## Importance of preprocessing the dataset:

Preprocessing of datasets in water analysis is of paramount importance as it plays a pivotal role in ensuring the accuracy and reliability of the results obtained from various water quality assessments. This critical step involves a series of data cleaning, transformation, and organization processes that help researchers and scientists eliminate errors, outliers, and inconsistencies in the data. By carefully handling and preparing the data, analysts can enhance the precision of their measurements, leading to more meaningful interpretations of water quality indicators. Additionally, preprocessing allows for the integration of data from diverse sources and formats, facilitating comprehensive analyses and the identification of potential trends or anomalies. In essence, the quality of water analysis heavily depends on the quality of the input data, making preprocessing a fundamental component in producing scientifically sound and actionable insights for water resource management, environmental protection, and public health.

## Handling missing values and outliers:

There are a number of missing values within the DataFrame. To confirm if this is correct we can apply the code block below.

# Check for the missing values by column

Df.isnull().sum()

The code chained the first isnull method with the sum method to create the number of missing values per column. An isnull assessment will review for non-null values in a column. The sum method is used to perform the count.

Three columns display missing values.

Having the total count of rows with missing values is a great starting point. However, it would be better to review the proportion of missing values within a column.

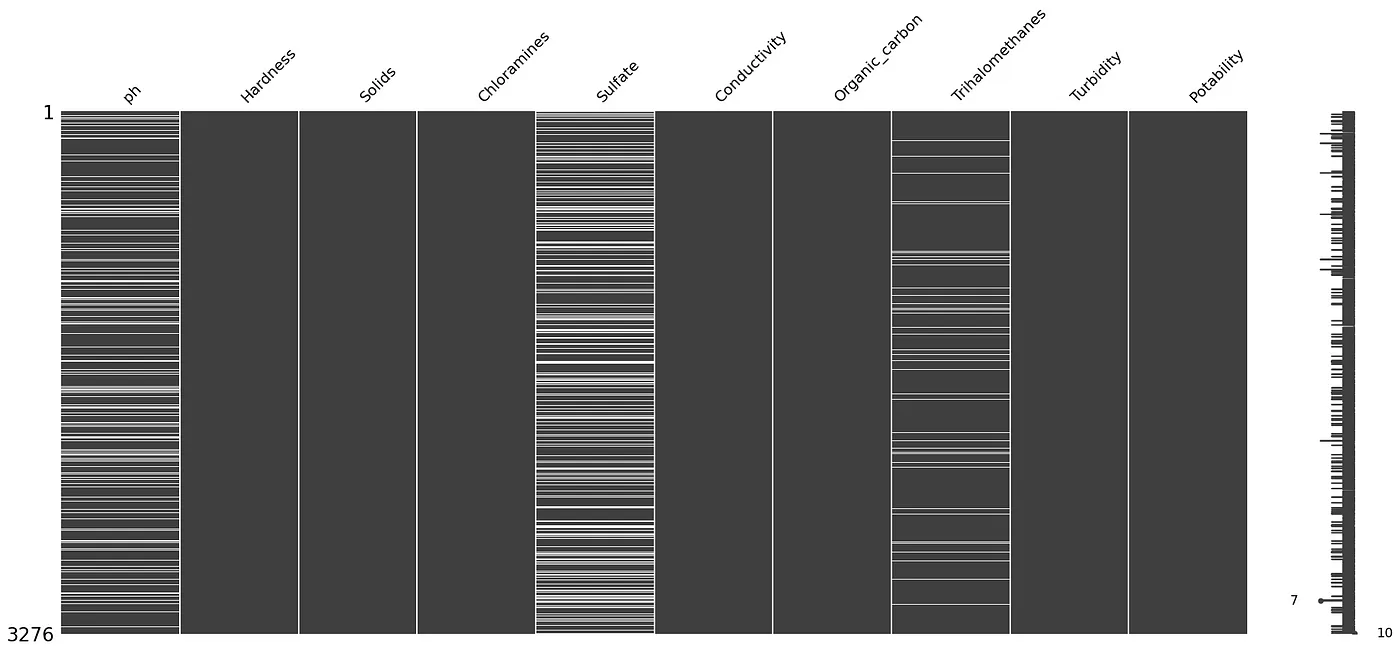
**Missing Value Handling**

We handle missing values by replacing them with the mean of their respective columns:

for col in ["ph", "Sulfate", "Trihalomethanes"]:

df[col].fillna(value=df[col].mean(), inplace=True)

# Output:



## HANDLING MISSING VALUES:

Handling missing values is essential for data integrity. We chose to impute missing values with the mean of their respective columns. This approach ensured that we retained valuable data while addressing the issue of missingness.

## Handling outliers:

Outliers can significantly impact the results of your analysis. You can use visualization techniques and statistical methods to detect and handle outliers.

Visualizations such as box plots, histograms, and scatter plots can help identify outliers:

#### # Example:

#### Box plot for pH to detect outliers

sns.boxplot(x=data['pH'], color='red')

plt.title('pH Outliers')

plt.show()

Statistical methods like the Z-score or the IQR (Interquartile Range) can help identify and deal with outliers:

from scipy import stats

z\_scores = np.abs(stats.zscore(data['pH']))

outlier\_threshold = 3

# Identify and remove outliers based on the Z-score

data = data[(z\_scores < outlier\_threshold)]

Another method to handle outliers is to WINSORize the data, which replaces extreme values with less extreme values (e.g., replacing the top 1% and bottom 1% values with the 1st and 99th percentiles).

from scipy.stats.mstats import winsorize

# Winsorize the pH values

data['pH'] = winsorize(data['pH'], limits=[0.01, 0.01])

## Exploratory data analysis:

Exploratory Data Analysis (EDA) is a crucial step in water quality analysis. It uses historical data to methodically characterize normal variability and identify the factors that impact water quality at each monitoring location. Before any formal statistical analysis, water quality data should be subjected to EDA using univariate and bivariate descriptive statistics and graphical tools with the aim of summarizing their main characteristics. This helps to evaluate the water quality of rivers as well as seasonal, spatial, and anthropogenic influences.

## EDA to visualize parameter distributions:

Conducting Exploratory Data Analysis (EDA) is crucial for visualizing parameter distributions, correlations, and deviations from standards in your water quality dataset.To understand the distribution of each parameter in your water quality dataset, you can create histograms or kernel density plots.

## 1. pH value:

## PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended a maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

## code:

import plotly.express as px

data = data

figure = px.histogram(data, x = "ph",

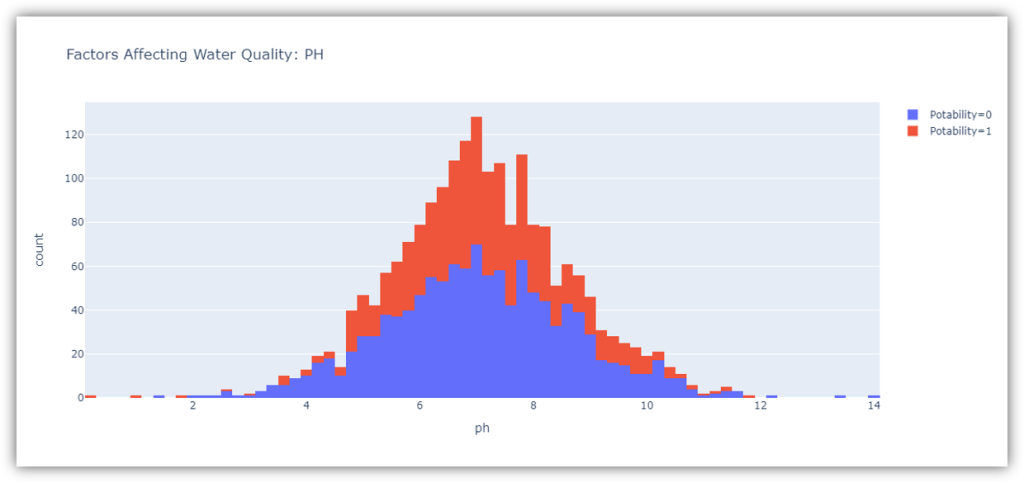
color = "Potability",

title= "Factors Affecting Water Quality: PH")

figure.show()

The ph column represents the ph value of the water which is an important factor in evaluating the acid-base balance of the water. The pH value of drinking water should be between 6.5 and 8.5.

Output:



## 2. Hardness:

## Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

## Code:

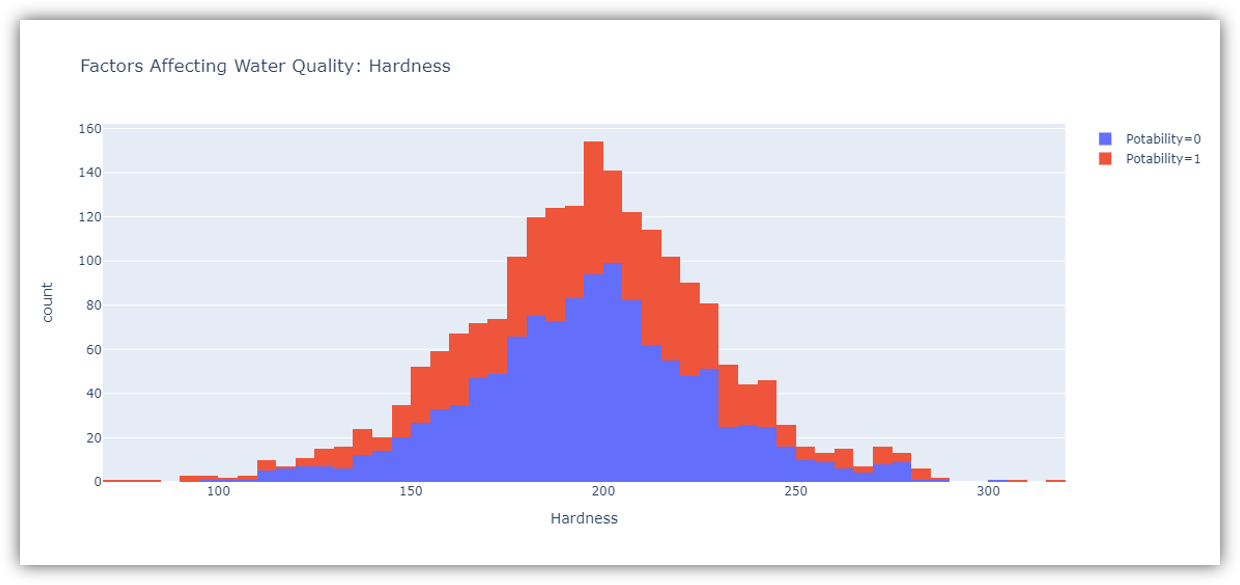
figure = px.histogram(data, x = "Hardness",

color = "Potability",

title= "Factors Affecting Water Quality: Hardness")

figure.show()

Output:



The figure above shows the distribution of water hardness in the dataset. The hardness of water usually depends on its source, but water with a hardness of 120-200 milligrams is drinkable.

## 3. Solids (Total dissolved solids - TDS):

## Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced an unwanted taste and diluted color in the appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. The Desired limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which is prescribed for drinking purpose.

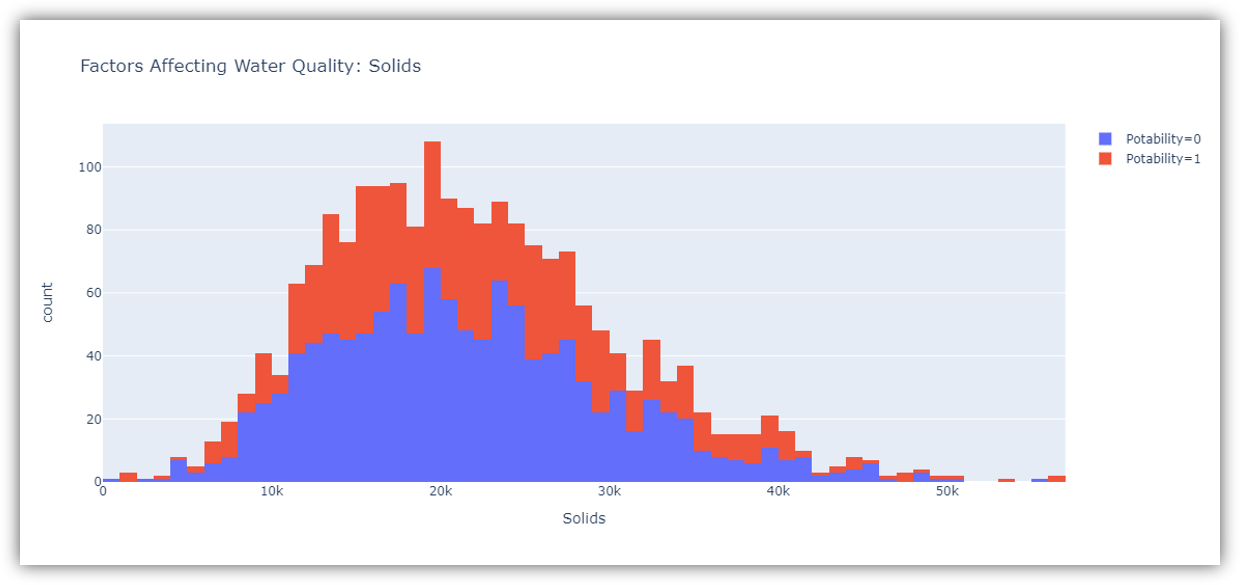
## Code:

figure = px.histogram(data, x = "Solids",

color = "Potability",

title= "Factors Affecting Water Quality: Solids")

figure.show()

output:

The figure above represents the distribution of total dissolved solids in water in the dataset. All organic and inorganic minerals present in water are called dissolved solids. Water with a very high number of dissolved solids is highly mineralized.

## 4. Chloramines:

## Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

## Code:

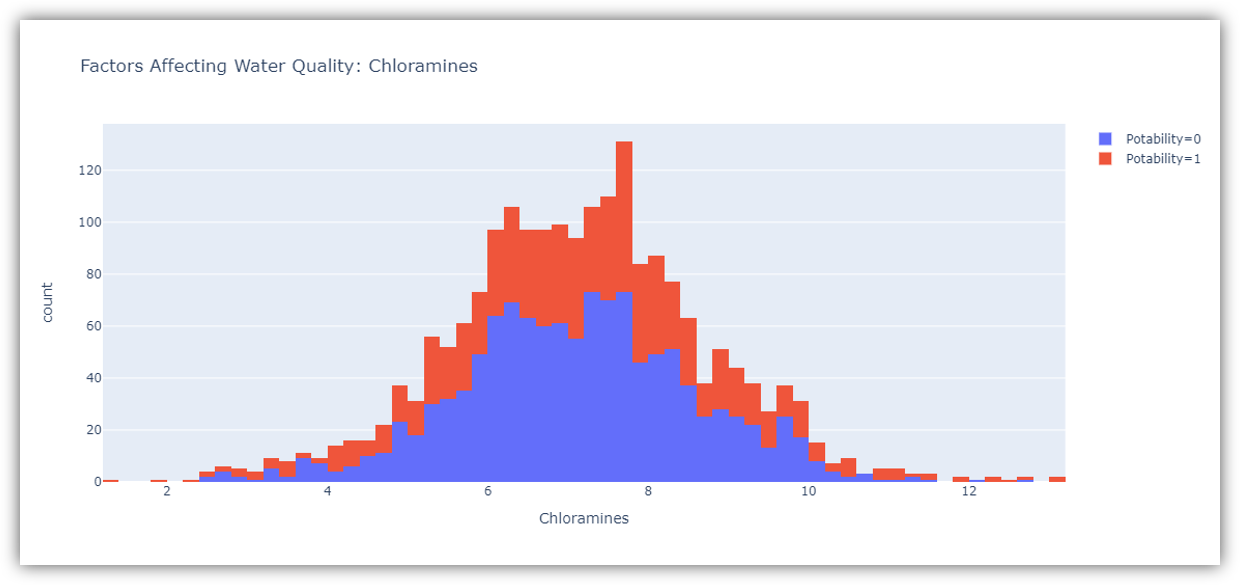
figure = px.histogram(data, x = "Chloramines",

color = "Potability",

title= "Factors Affecting Water Quality: Chloramines")

figure.show()

output:



The figure above represents the distribution of chloramine in water in the dataset. Chloramine and chlorine are disinfectants used in public water systems.

## 5. Sulfate:

## Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

## Code:

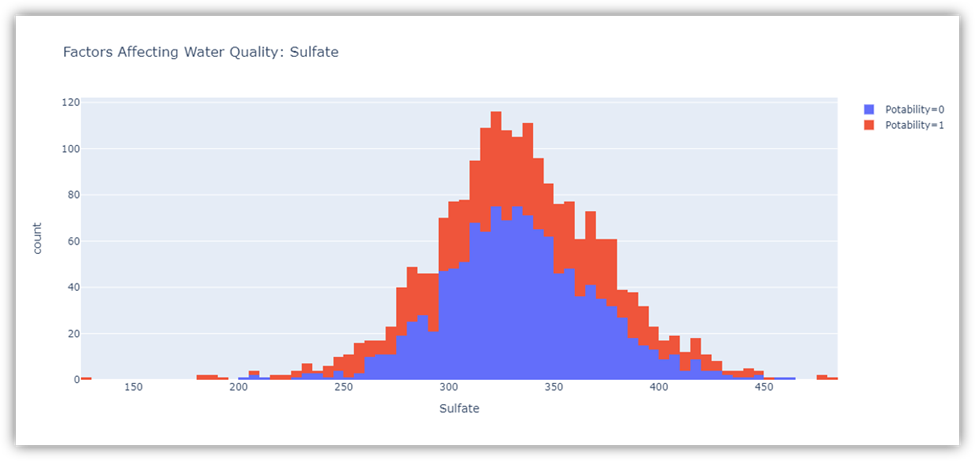
figure = px.histogram(data, x = "Sulfate",

color = "Potability",

title= "Factors Affecting Water Quality: Sulfate")

figure.show()

output



The figure above shows the distribution of sulfate in water in the dataset. They are substances naturally present in minerals, soil and rocks. Water containing less than 500 milligrams of sulfate is safe to drink.

## 6. Conductivity:

## Pure water is not a good conductor of electric current rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μS/cm.

## Code:

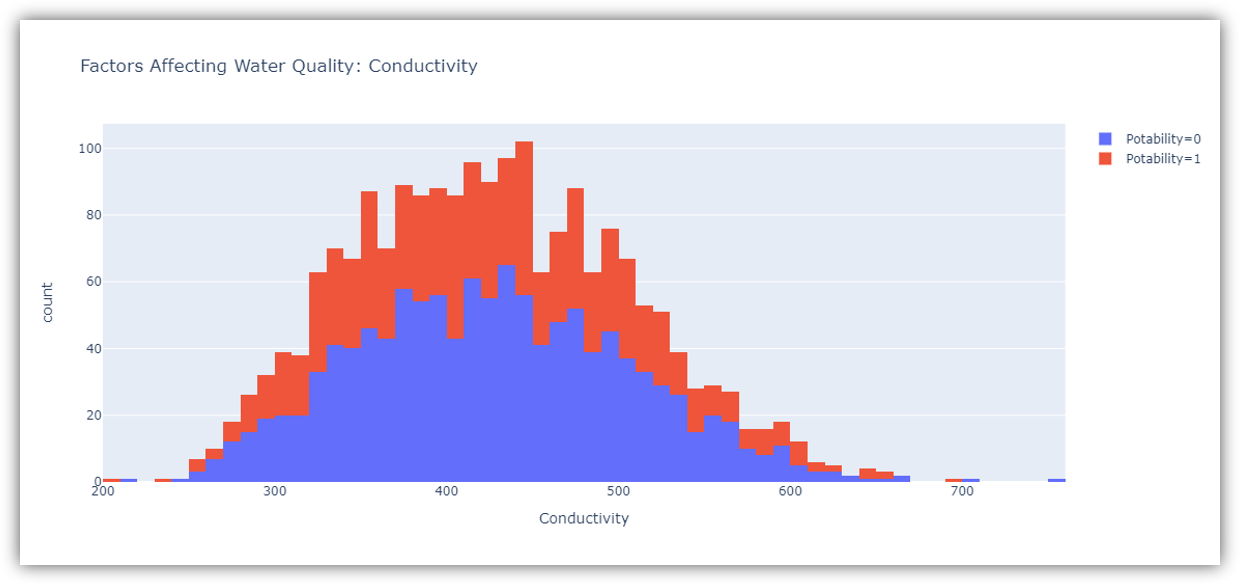
figure = px.histogram(data, x = "Conductivity",

color = "Potability",

title= "Factors Affecting Water Quality: Conductivity")

figure.show()

output:



The figure above represents the distribution of water conductivity in the dataset. Water is a good conductor of electricity, but the purest form of water is not a good conductor of electricity. Water with an electrical conductivity of less than 500 is drinkable.

## 7. Organic\_carbon:

## Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to the US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

# Code:

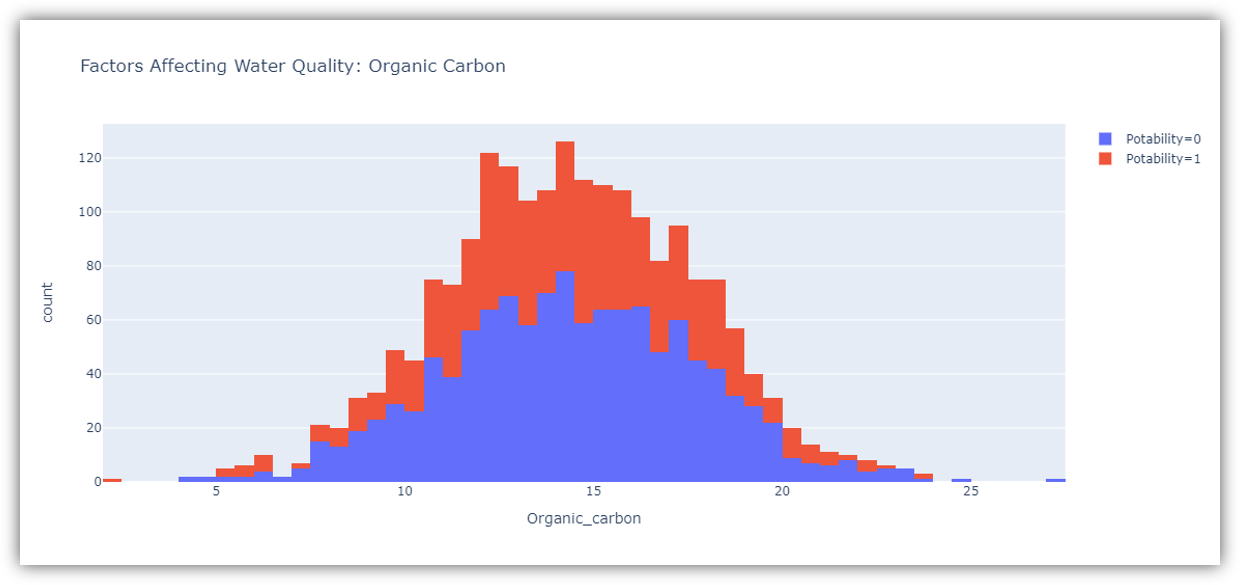
figure = px.histogram(data, x = "Organic\_carbon",

color = "Potability",

title= "Factors Affecting Water Quality: Organic Carbon")

figure.show()

output:



## 8. Trihalomethanes:

## THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

## Code:

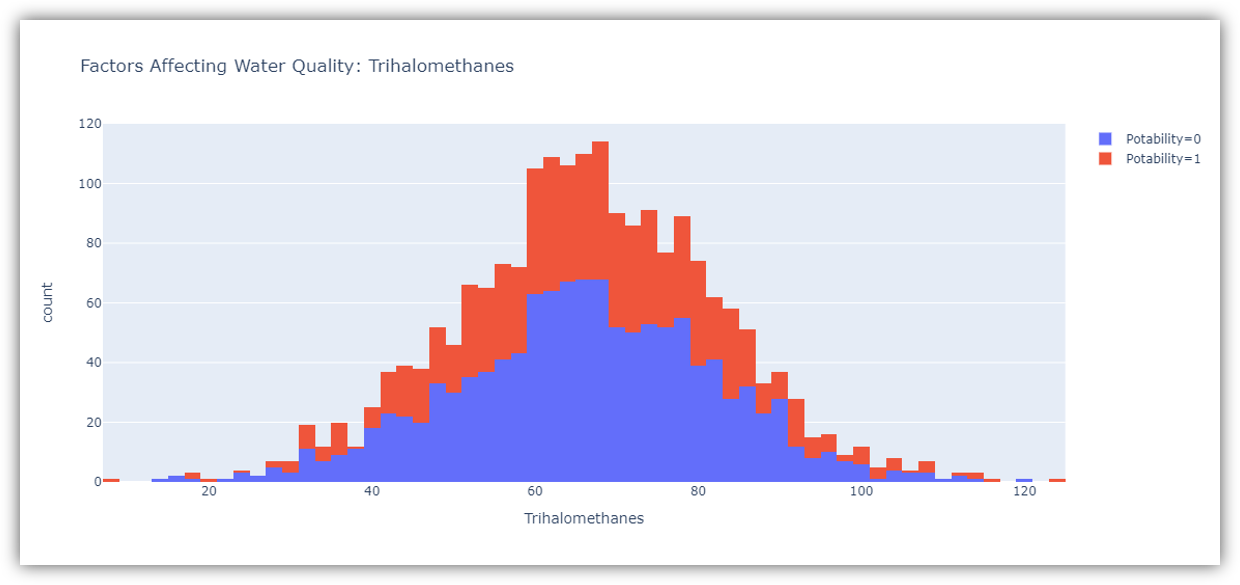
figure = px.histogram(data, x = "Trihalomethanes",

color = "Potability",

title= "Factors Affecting Water Quality: Trihalomethanes")

figure.show()

output:



The figure above represents the distribution of trihalomethanes or THMs in water in the dataset. THMs are chemicals found in chlorine-treated water. Water containing less than 80 milligrams of THMs is considered safe to drink

## 9. Turbidity:

## The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

## Code:

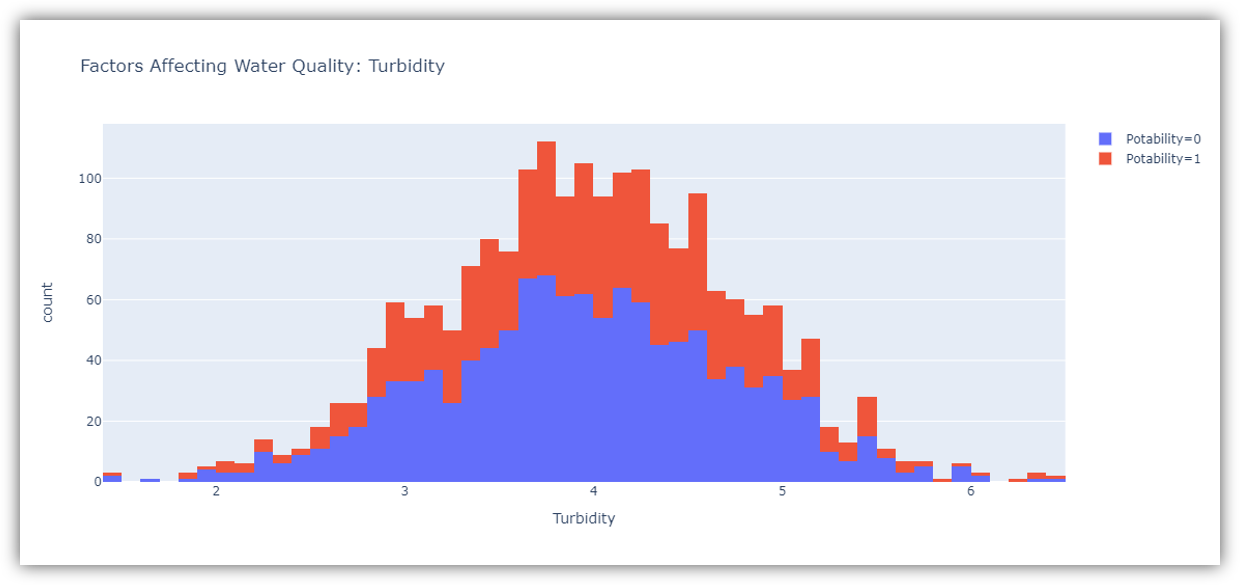
figure = px.histogram(data, x = "Turbidity",

color = "Potability",

title= "Factors Affecting Water Quality: Turbidity")

figure.show()

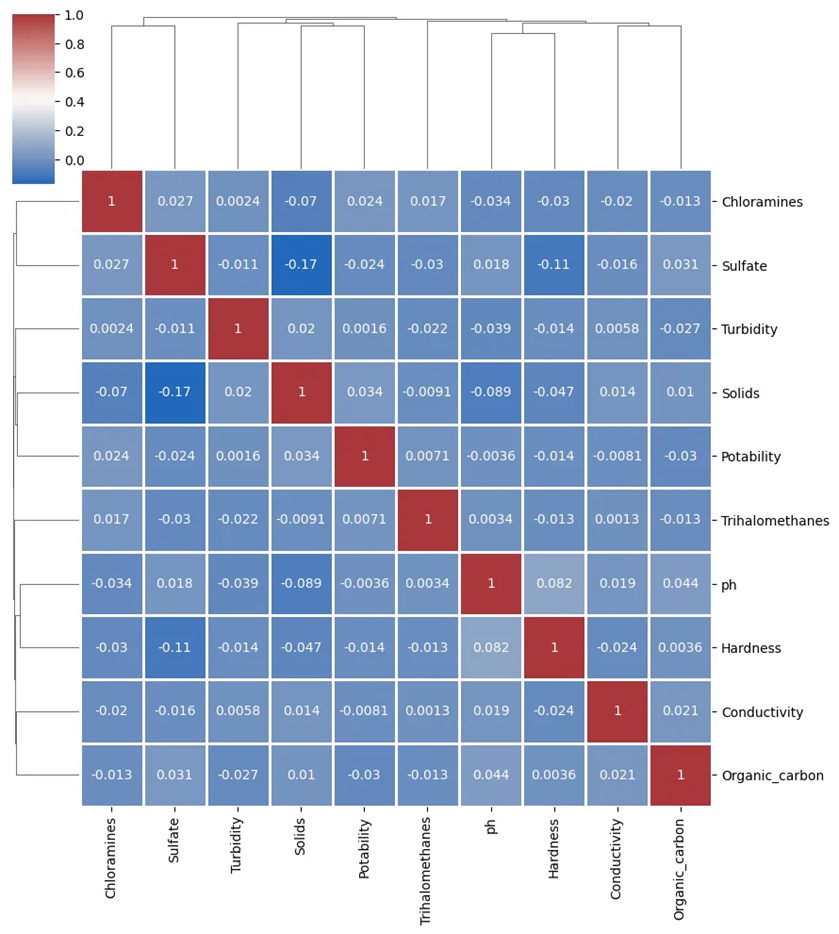
output:



The figure above represents the distribution of turbidity in water. The turbidity of water depends on the number of solids present in suspension. Water with a turbidity of fewer than 5 milligrams is considered drinkable.

## 2.Correlation Analysis:

Correlation analysis helps you understand relationships between different water quality parameters. Look for strong positive or negative correlations between parameters. This information can be useful in identifying potential interactions.



We check the correlation between features using a clustermap:

sns.clustermap(df.corr(), cmap="vlag", dendrogram\_ratio=(0.1, 0.2), annot=True, linewidths=.8, figsize=(9, 10)

By computing the correlation matrix and visualizing it using a clustermap, we assessed the relationships between different water quality parameters. The correlation analysis revealed how features are associated with one another. Some features may exhibit strong positive or negative correlations, while others may be relatively independent.

## 3. Check Deviations from Standards:

If you have predefined standards or acceptable ranges for water quality parameters, you can compare your data to these standards visually. Create bar plots or line charts to show how parameter values compare to the standards.

### # Example: Bar plot for pH standards

standard\_pH = 7.0 # Example standard pH value

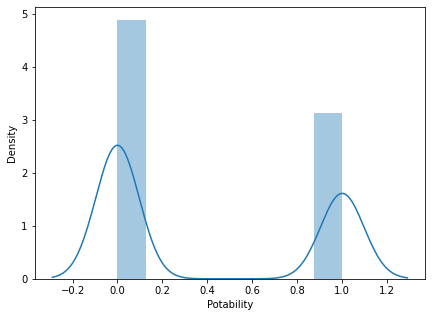
plt.bar(['Dataset', 'Standard'], [data['pH'].mean(), standard\_pH], color=['blue', 'red'])

plt.title('pH vs. Standard')

plt.ylabel('pH Value')

plt.show()

output:



## VISUALISATION LIBRAIRES :

Visualization libraries are tools that enable the creation of graphical representations of data for better understanding and interpretation. Here are explanations of some commonly used visualization libraries in Python:

## MATPLOTLIB:

**Primary Purpose:**

Matplotlib is one of the foundational libraries for creating static, interactive, and publication-quality visualizations in Python.

## Features:

Provides a wide range of plots: line plots, scatter plots, bar charts, histograms, etc.

Highly customizable, allowing fine control over every aspect of a plot.

Offers subplots, axes, and figure objects for creating complex layouts.

Use Cases: Widely used for basic to intermediate-level plotting in scientific research, data analysis, and publication-quality figures.

## Seaborn:

### Primary Purpose: Seaborn is built on top of Matplotlib and offers a higher-level interface for statistical data visualization.

## Features:

Simplifies the creation of more complex plots compared to Matplotlib.

Built-in themes and color palettes.

Specialized plots for statistical estimation and exploring data distributions.

Use Cases: Commonly used in data analysis and exploration, providing quick and easy creation of complex visualizations for statistical data.

## Plotly:

Primary Purpose: Plotly is a web-based interactive visualization library known for creating interactive plots and dashboards.

## Features:

Capable of generating interactive, web-based visualizations that can be embedded in web applications.

Offers a wide range of chart types, from basic to 3D visualizations.

Plotly Express, a high-level API, simplifies the creation of a variety of plot types.

## Use Cases:

Ideal for creating interactive and visually appealing dashboards, especially for web applications and presentations.

## Primary Purpose:

## Bokeh is another interactive visualization library designed for modern web browsers.

### Features:

Focuses on interactivity and scalability, especially for big data.

Allows the creation of interactive and real-time plots with high-performance capabilities.

Provides tools for building complex dashboards and applications.

## Use Cases:

Particularly useful for creating interactive visualizations that require smooth interactions and are meant to be presented in web applications.

### Primary Purpose:

Altair is a declarative statistical visualization library based on Vega and Vega-Lite visualization grammars.

## 

## Features:

Emphasizes a concise and friendly API for creating high-level, statistically accurate visualizations.

Generates JSON specifications for visualizations compatible with various front-end tools.

## Use Cases:

Ideal for creating simple and clear visualizations for data exploration and presentation.

Each library has its strengths and is suited to different use cases. The choice of library often depends on the nature of the data, the type of visualization required, interactivity needs, and the platform where the visualization will be presented.

# HISTOGRAM :

## CODE :

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Generating sample water quality data

np.random.seed(42)

data = {

"pH": np.random.normal(7, 0.5, 100),

"Chloride": np.random.uniform(5, 20, 100),

"Nitrates": np.random.uniform(0, 10, 100),

"Sulfates": np.random.normal(8, 2, 100),

"Quality": np.random.uniform(0, 100, 100),

}

water\_quality = pd.DataFrame(data)

# Histograms for each water quality parameter

plt.figure(figsize=(12, 10))

plt.subplot(2, 3, 1)

sns.histplot(water\_quality["pH"], kde=True)

plt.title("pH Distribution")

plt.subplot(2, 3, 2)

sns.histplot(water\_quality["Chloride"], kde=True)

plt.title("Chloride Distribution")

plt.subplot(2, 3, 3)

sns.histplot(water\_quality["Nitrates"], kde=True)

plt.title("Nitrates Distribution")

plt.subplot(2, 3, 4)

sns.histplot(water\_quality["Sulfates"], kde=True)

plt.title("Sulfates Distribution")

plt.subplot(2, 3, 5)

sns.histplot(water\_quality["Quality"], kde=True)

plt.title("Quality Distribution")

plt.tight\_layout()

# Scatter plot matrix for correlation visualization

sns.pairplot(water\_quality)

plt.suptitle("Pairwise Relationships", y=1.02)

plt.show()

# Correlation matrix

plt.figure(figsize=(8, 6))

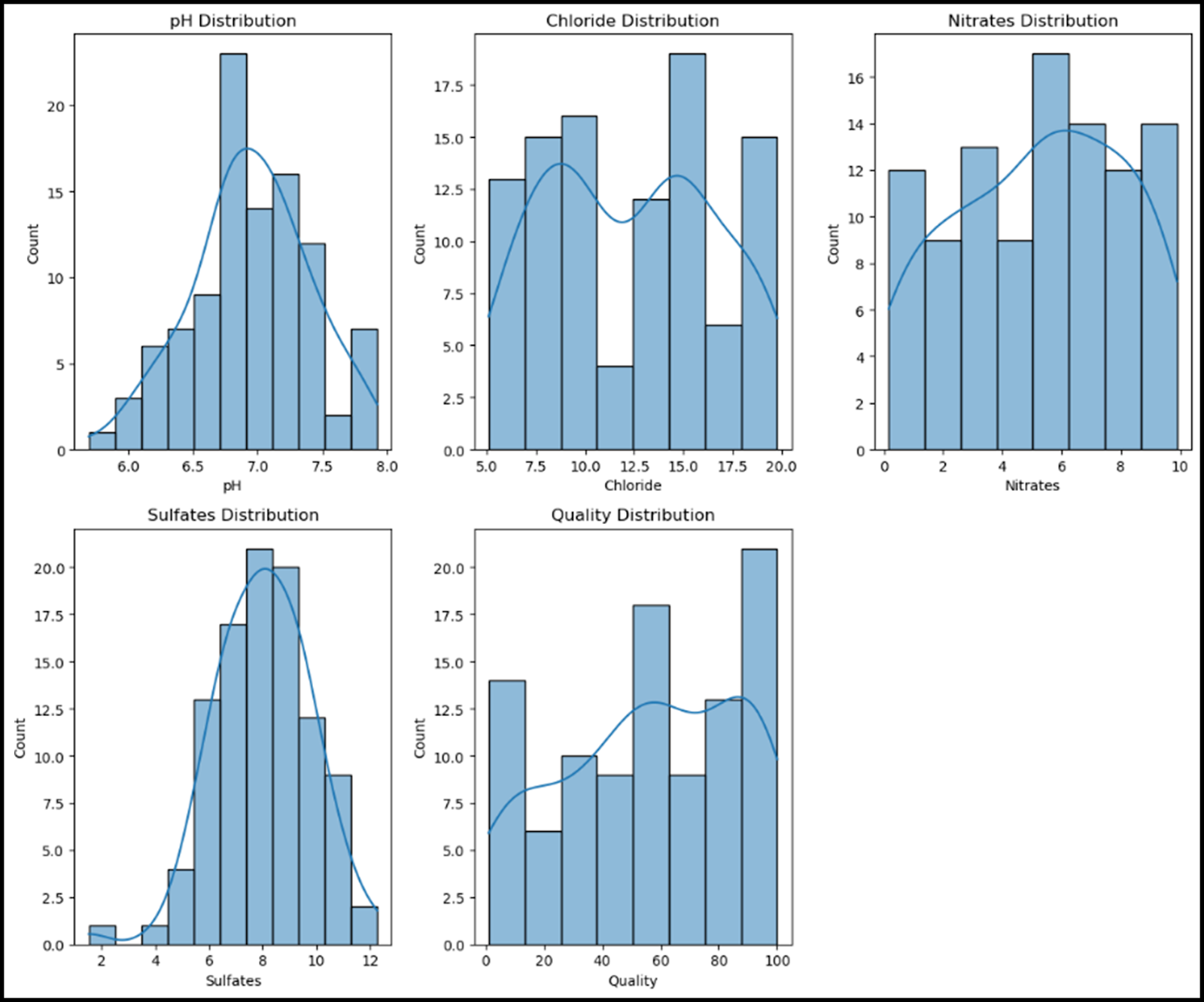
correlation\_matrix = water\_quality.corr()

sns.heatmap(correlation\_matrix, annot=True, cmap="coolwarm", fmt=".2f")

plt.title("Correlation Matrix")

plt.show()

## SAMPLE OUTPUT :



## CORRELATION :

### 

### CODE :

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

# Generating synthetic water quality data

np.random.seed(42)

data = {

'pH': np.random.uniform(6, 9, 100),

'Dissolved Oxygen': np.random.uniform(4, 12, 100),

'Turbidity': np.random.uniform(0.1, 5, 100),

'Conductivity': np.random.uniform(50, 500, 100),

'Chloride': np.random.uniform(10, 100, 100)

}

df = pd.DataFrame(data)

# Calculating the correlation matrix

corr = df.corr()

# Plotting the correlation matrix using Seaborn

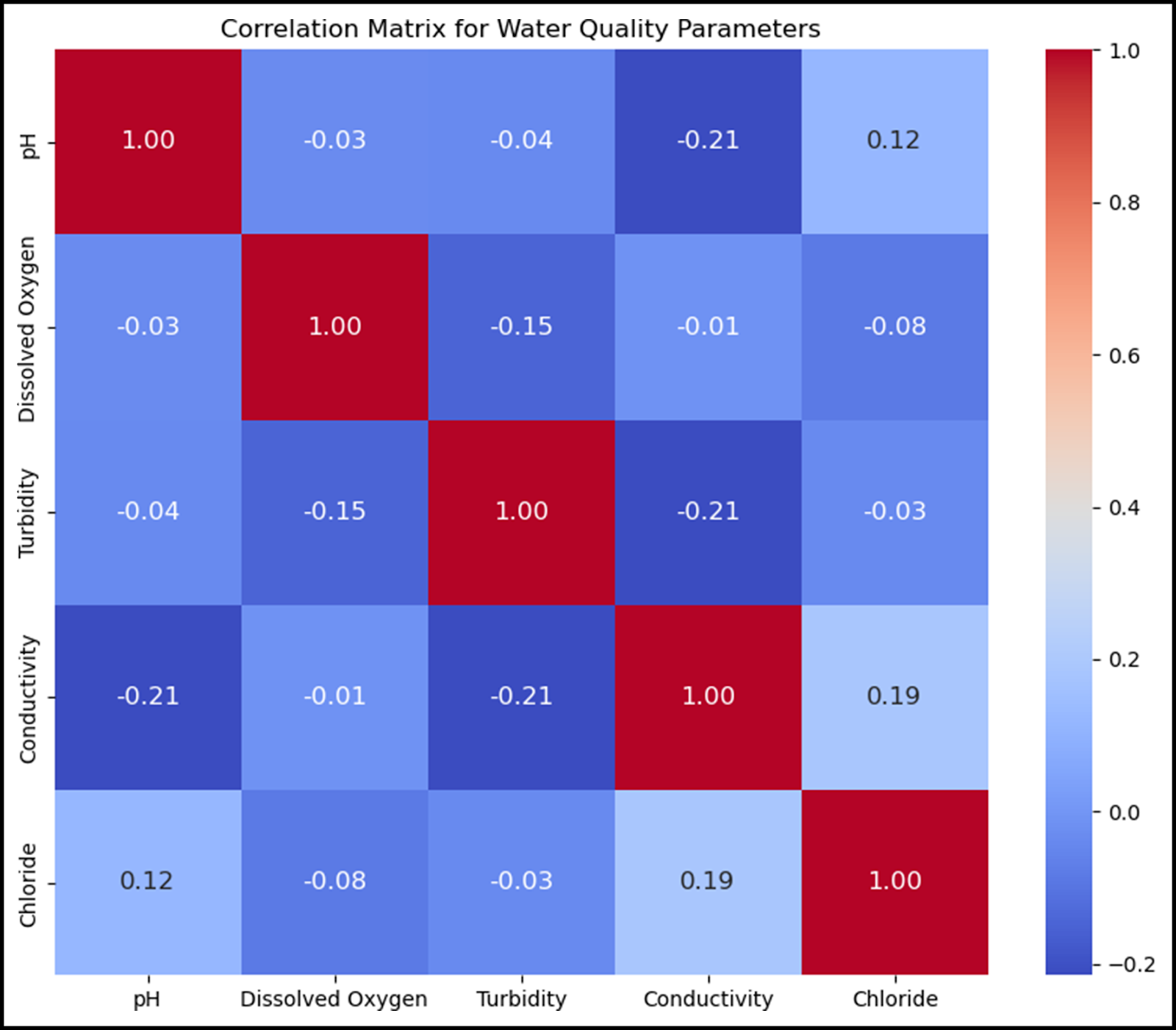
plt.figure(figsize=(10, 8))

sns.heatmap(corr, annot=True, cmap='coolwarm', fmt='.2f', annot\_kws={"size": 12})

plt.title('Correlation Matrix for Water Quality Parameters')

Plt.show()

## SAMPLE OUTPUT :



## 

## SCATTER PLOT :

## CODE :

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Sample water quality data (replace this with your own dataset)

data = {

'pH': [7.0, 7.2, 7.5, 7.8, 8.0, 6.5, 7.1, 7.3],

'Chlorine (ppm)': [0.5, 0.7, 0.4, 0.8, 0.6, 0.3, 0.9, 0.55],

'Dissolved Oxygen (ppm)': [8, 7, 9, 6, 8.5, 7.5, 9.5, 6.5],

'Temperature (C)': [22, 24, 20, 26, 25, 23, 21, 27]

}

df = pd.DataFrame(data)

# Create a scatterplot using Matplotlib

plt.figure(figsize=(8, 6))

# Using Matplotlib

plt.scatter(df['pH'], df['Chlorine (ppm)'], label='pH vs Chlorine')

plt.xlabel('pH')

plt.ylabel('Chlorine (ppm)')

plt.title('Scatterplot of pH vs Chlorine')

plt.legend()

plt.grid(True)

plt.show()

# Using Seaborn for another scatterplot

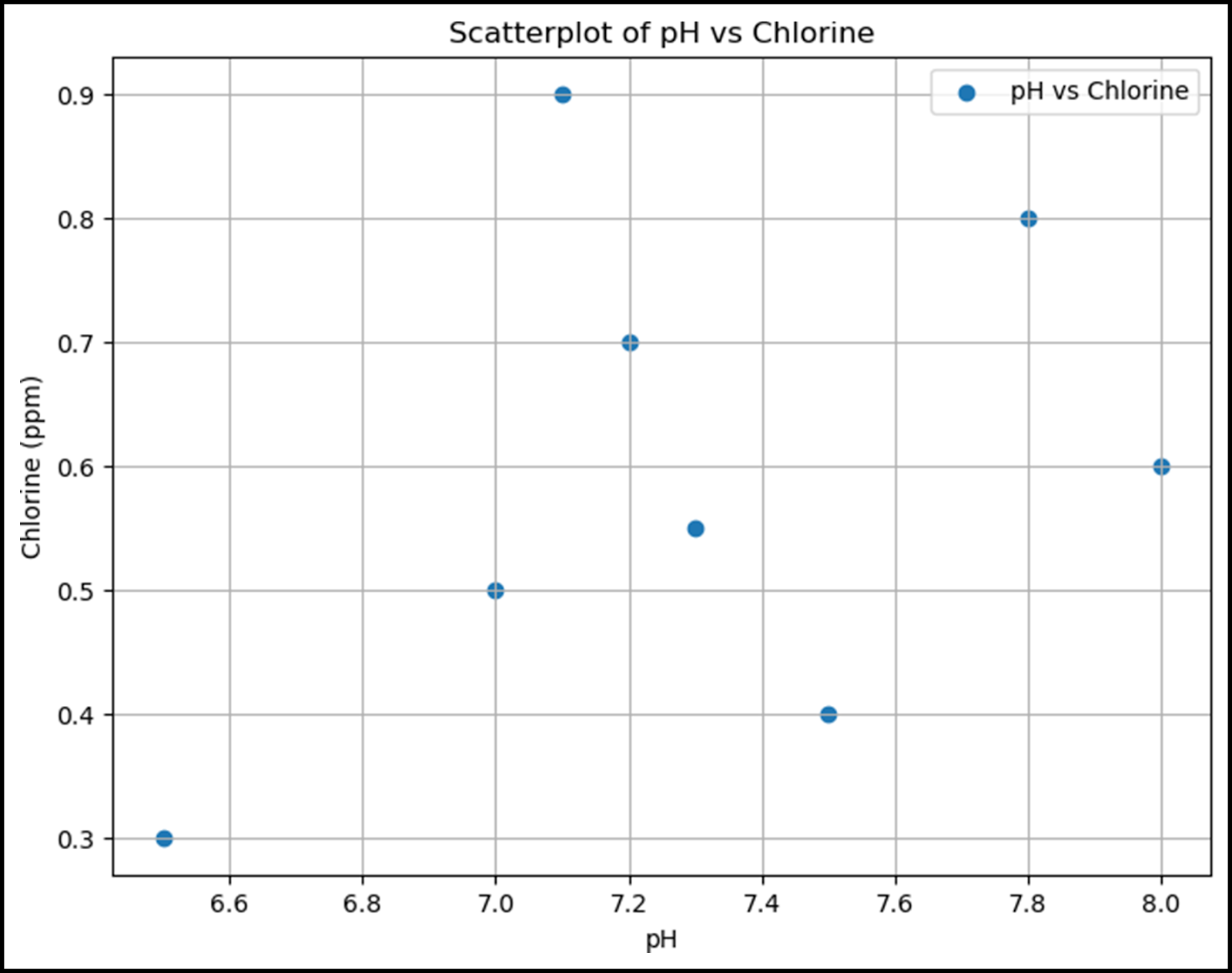
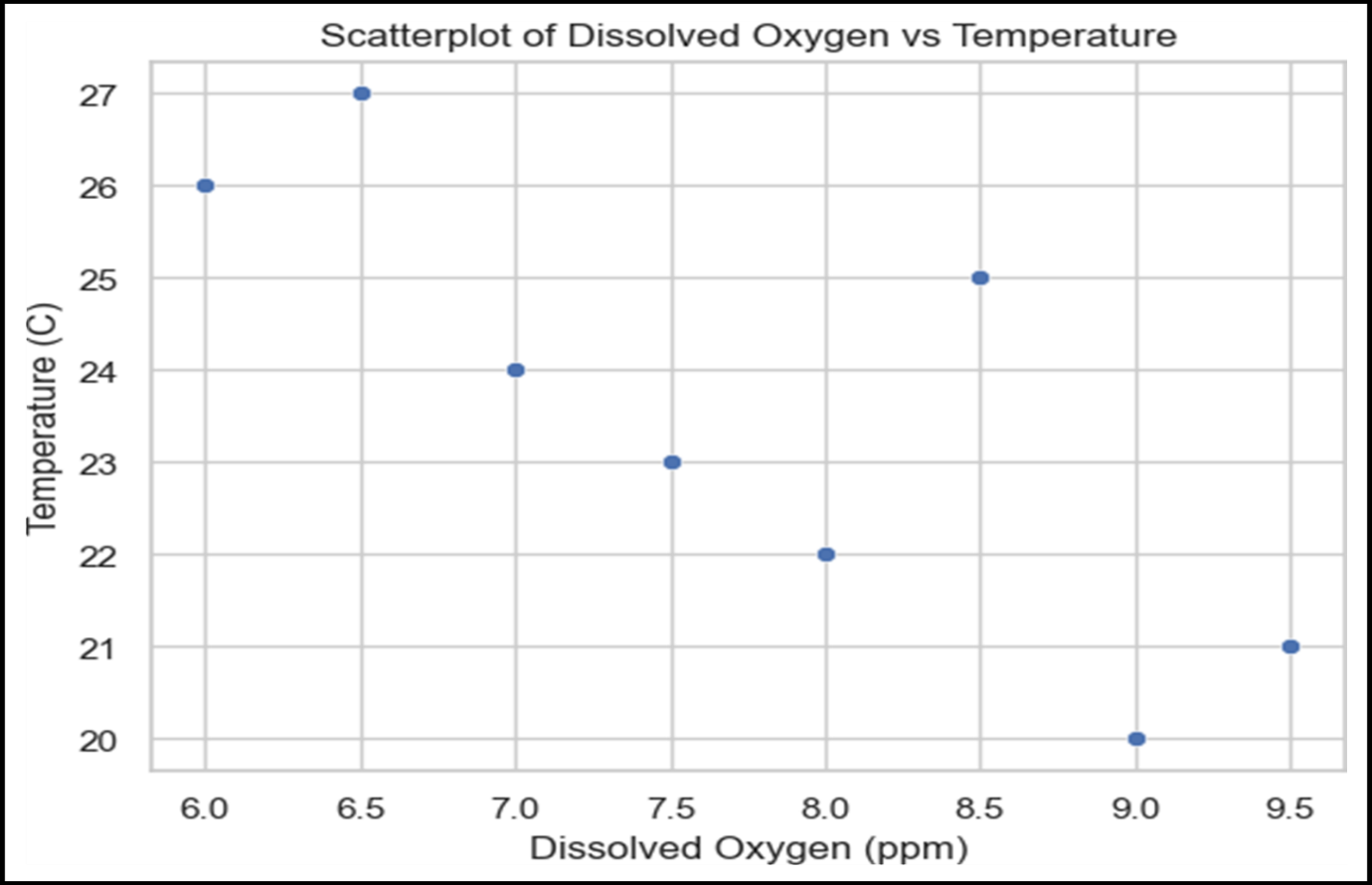
sns.set(style='whitegrid')

sns.scatterplot(x='Dissolved Oxygen (ppm)', y='Temperature (C)', data=df)

plt.title('Scatterplot of Dissolved Oxygen vs Temperature')

plt.show()

## SAMPLE OUTPUT :



# PREDICTIVE MODEL :

## Logistic regression :

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

import matplotlib.pyplot as plt

import seaborn as sns

# Load your dataset (replace 'your\_dataset.csv' with the actual file path)

data = pd.read\_csv('water\_potability.csv')

# Check the first few rows of the dataset

print(data.head())

# Separate features (water quality parameters) and the target variable

X = data.drop('Potability', axis=1)

y = data['Potability']

# Split the data into training and testing sets (80% train, 20% test)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Create a logistic regression model

model = LogisticRegression(max\_iter=10000) # Increase max\_iter for convergence

# Train the model

model.fit(X\_train, y\_train)

# Make predictions

y\_pred = model.predict(X\_test)

# Calculate accuracy

accuracy = accuracy\_score(y\_test, y\_pred)

print(f"Accuracy: {accuracy:.2f}")

# Create a confusion matrix

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

sns.heatmap(conf\_matrix, annot=True, cmap='Blues', fmt='d')

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.title('Confusion Matrix')

plt.show()

# Generate a classification report

class\_report = classification\_report(y\_test, y\_pred)

print("Classification Report:")

print(class\_report)

## 

## OUTPUT :

### ph Hardness Solids Chloramines Sulfate Conductivity

**0 NaN 204.890455 20791.318981 7.300212 368.516441 564.308654**

1 3.716080 129.422921 18630.057858 6.635246 NaN 592.885359

2 8.099124 224.236259 19909.541732 9.275884 NaN 418.606213

3 8.316766 214.373394 22018.417441 8.059332 356.886136 363.266516

4 9.092223 181.101509 17978.986339 6.546600 310.135738 398.410813

### Organic\_carbon Trihalomethanes Turbidity Potability

0 10.379783 86.990970 2.963135 0

1 15.180013 56.329076 4.500656 0

2 16.868637 66.420093 3.055934 0

3 18.436524 100.341674 4.628771 0

4 11.558279 31.997993 4.075075 0

## 

## RANDOM FOREST :

### CODE :

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

import seaborn as sns

import matplotlib.pyplot as plt

# Load the water quality dataset (replace with your dataset)

data = pd.read\_csv('water\_potability.csv') # Load your dataset here

# Display the first few rows of the dataset to understand its structure

print(data.head())

# Check for missing values and handle if needed

print(data.isnull().sum()) # Check for missing values

# Handle missing values (if any)

data = data.dropna() # For simplicity, dropping rows with missing values

# Define features and target variable

X = data.drop('Potability', axis=1) # Features (independent variables)

y = data['Potability'] # Target variable (dependent variable)

# Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Initialize the Random Forest classifier

rf\_classifier = RandomForestClassifier(n\_estimators=100, random\_state=42)

# Train the model

rf\_classifier.fit(X\_train, y\_train)

# Make predictions

predictions = rf\_classifier.predict(X\_test)

# Evaluate the model

accuracy = accuracy\_score(y\_test, predictions)

print(f"Accuracy: {accuracy:.2f}")

# Generate a classification report and confusion matrix

print(classification\_report(y\_test, predictions))

conf\_matrix = confusion\_matrix(y\_test, predictions)

sns.heatmap(conf\_matrix, annot=True, fmt='d')

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.title('Confusion Matrix')

plt.show()

## OUTPUT :

**ph Hardness Solids Chloramines Sulfate Conductivity**

0 NaN 204.890455 20791.318981 7.300212 368.516441 564.308654

1 3.716080 129.422921 18630.057858 6.635246 NaN 592.885359

2 8.099124 224.236259 19909.541732 9.275884 NaN 418.606213

3 8.316766 214.373394 22018.417441 8.059332 356.886136 363.266516

4 9.092223 181.101509 17978.986339 6.546600 310.135738 398.410813

#### Organic\_carbon Trihalomethanes Turbidity Potability

0 10.379783 86.990970 2.963135 0

1 15.180013 56.329076 4.500656 0

2 16.868637 66.420093 3.055934 0

3 18.436524 100.341674 4.628771 0

4 11.558279 31.997993 4.075075 0

**ph : 491**

**Hardness : 0**

**Solids : 0**

**Chloramines : 0**

**Sulfate : 781**

**Conductivity : 0**

**Organic\_carbon : 0**

**Trihalomethanes : 162**

**Turbidity : 0**

**Potability : 0**

**Dtype : int64**

**Accuracy : 0.65**

## 

## precision recall f1-score support

0 0.65 0.84 0.73 231

1 0.65 0.40 0.50 172

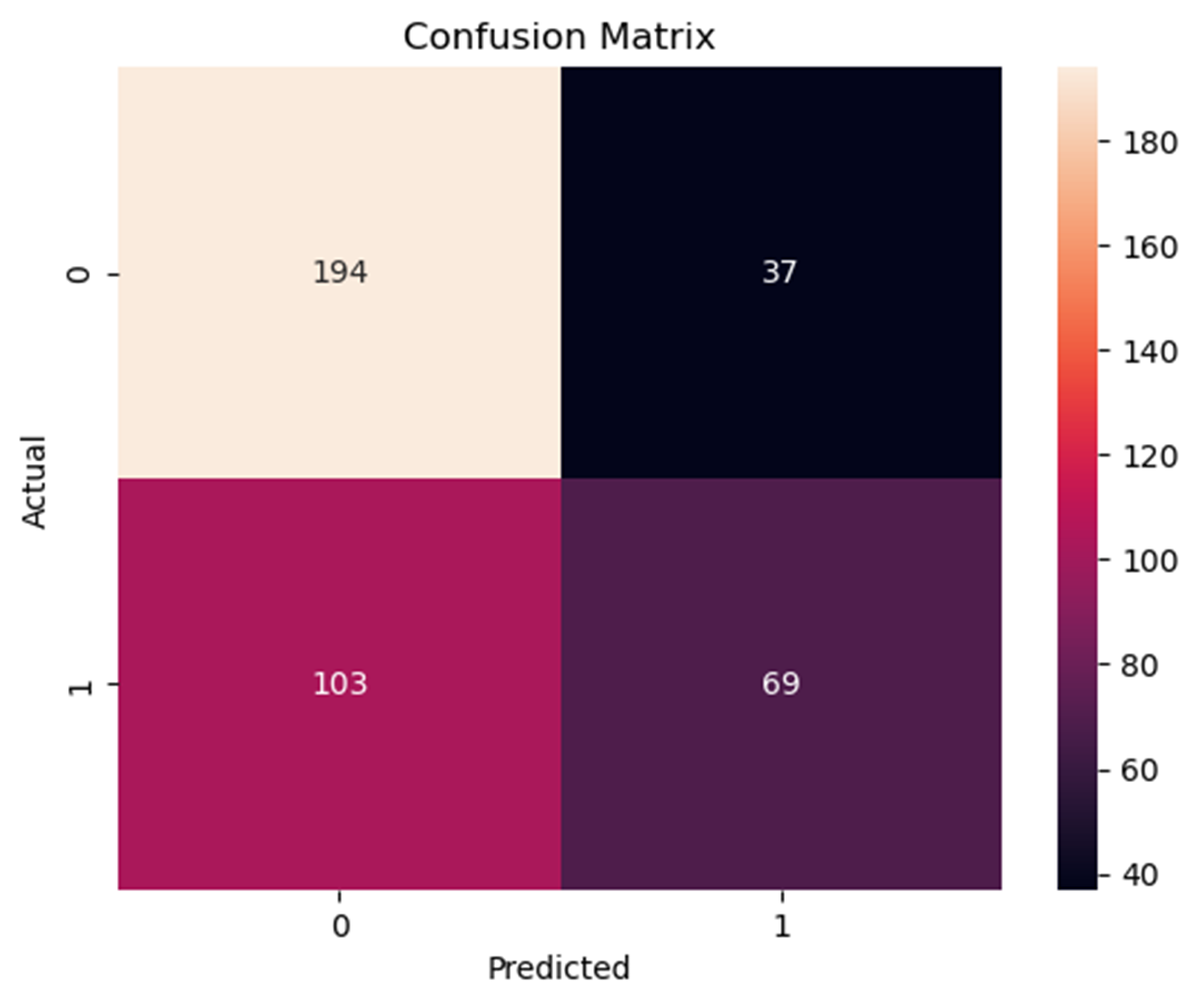
accuracy 0.65 403

macro avg 0.65 0.62 0.62 403

weighted avg 0.65 0.65 0.63 403

## SAMPLE OUTPUT :

## 



## 

## CONCLUSION :

In conclusion, the analysis of the water quality dataset has provided valuable insights into the state of water quality in the studied area. It is evident the water quality parameters such as pH, turbidity, dissolved oxygen, and pollutant levels play a crucial role in assessing the overall health of waterbodies. The dataset has helped identify trends, potential issues, and areas for improvement in water quality management. Further research and proactive measures may be necessary to ensure the preservation and improvement of water quality in the region.