DATA ANALYSIS WITH COGNOS

Group 2

Project 11 – 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**

**Project overview:**

Access to clean and safe water is a fundamental human right and a critical component of public health and environmental sustainability. The project "Water Analysis Using Data Analysis" aims to utilize modern data analysis techniques to assess, monitor, and improve the quality of water resources. This project recognizes the importance of data-driven decision-making in addressing challenges related to potability, environmental impact, and resource management. Through the comprehensive analysis of water quality data, this initiative seeks to contribute to the protection of public health, the preservation of ecosystems, and the responsible use of this invaluable resource.

Objectives:

Regulatory standards:

This project seeks to identify and evaluate potential issues or deviations from regulatory standards governing water quality. It involves the collection and analysis of data related to various water quality parameters, ensuring that water sources meet the requirements set forth by relevant environmental agencies and regulatory bodies.

Water potability:

Another key objective is to determine the potability of water based on a comprehensive analysis of multiple parameters. The project will assess the suitability of water sources for human consumption by comparing measured data against established drinking water standards, such as those outlined by the World Health Organization (WHO) or local environmental agencies.

**Definition:**

Water is one of the basic resources for human survival. Most traditional water quality analysis systems, however, generally focus only on water quality data collection, data analysis and data mining, ignoring various properties of water such as pH, hardness. In addition, some dirty data and data loss may occur due to power failures or transmission failures, further affecting data analysis and its application.

The project involves analyzing water quality data to assess the suitability of water for specific purposes, such as drinking. The objective is to identify potential issues or deviations from regulatory standards and determine water potability based on various parameters. This project includes defining analysis objectives, collecting water quality data, designing relevant visualizations, and building a predictive model.

Data analytics techniques are applied to the collected data to identify trends, anomalies, and patterns, enabling informed decision-making and the implementation of measures whether the water is potable or not.

**Project approach:**

**Design Thinking:**

Design thinking for water analysis involves approach to ideate solutions and basic structures to solve and improving the analysis in a more efficient and user centric way.

Analyzing water quality data is a critical task with a range of specific objectives that help ensure the safety and sustainability of water resources. Design thinking in water analysis has four core components analysis objectives, data collections, visualization strategy, predictive modelling.

**Analysis Objectives:**

**1.Assessing Potability:**

The primary objective of water quality analysis is to determine whether the water meets the standards necessary for safe human consumption. This involves assessing parameters such as pH, turbidity, total dissolved solids (TDS), and concentrations of contaminants like bacteria, heavy metals, and organic compounds. The specific objective here is to compare the measured values to established drinking water standards, such as those set by the World Health Organization (WHO) or the Environmental Protection Agency (EPA) in the United States. If any parameter exceeds the maximum allowable concentration, it signifies a potential health risk, and corrective actions must be taken to ensure the water is safe to drink.

**2.Identifying Deviations from standards:**

Water quality data analysis aims to identify any deviations from established standards or regulatory limits. This involves detecting spikes or unusual trends in parameters that may indicate contamination or deterioration in water quality. For example, sudden increases in fecal coliform counts might suggest a sewage leak, while elevated levels of nitrates may indicate agricultural runoff. Identifying these deviations is essential for early warning and rapid response, allowing authorities to address potential threats to public health and the environment. The specific objective is to pinpoint the source and cause of these deviations and take appropriate corrective actions to restore water quality.

**3.Understanding Parameter Relationships:**

To manage and protect water resources effectively, it's crucial to understand the complex relationships between various water quality parameters. For instance, the pH level can impact the solubility of metals in water, while temperature fluctuations can affect the growth of aquatic organisms. The objective here is to use statistical and analytical methods to identify correlations and trends among different parameters. This helps in predicting how changes in one parameter may affect others and allows for more informed decision-making. For instance, understanding the relationship between temperature and dissolved oxygen levels can be vital in managing aquatic ecosystems.

**Data Collection:**

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.so we gather water quality data containing parameters.

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 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.

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.

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 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.

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.

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.

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.

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 US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

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.

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.

10. Potability:

Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

The real time water potability data is available in the csv file below:



**Dataset Link:**[**https://www.kaggle.com/datasets/adityakadiwal/water-potability**](https://www.kaggle.com/datasets/adityakadiwal/water-potability)

**Visualization strategies:**

Visualizing parameter distributions, correlations, and potability in water quality data is essential for gaining insights and making informed decisions. Here's a plan on how to visualize these aspects using suitable tools:

1.patrameter distibutions:

Visualize the distribution of individual water quality parameters like pH, hardness, solids, and contaminants to understand their variability and identify potential outliers.

Histogram**:** Use histograms to display the frequency distribution of each parameter. This helps in identifying common ranges and potential anomalies.

Box plots: Box plots provide a visual summary of the distribution's central tendency, spread, and outliers, making them useful for comparing parameters.

Density plots: Create density plots to visualize the probability density of continuous parameters, revealing patterns and variations.

2.Correlations:

Explore the relationships between different water quality parameters to identify dependencies and potential causations.

Scatter plots: Create scatter plots to visualize the pairwise relationships between parameters. Scatter plots are particularly useful for identifying linear or non-linear correlations

Correlation Matrices: Use heatmaps to display correlation matrices, showing the strength and direction of correlations between parameters. Tools like Python's Seaborn or R's ggplot2 are excellent for this purpose.

3.Potability Assessment:

Visualize the assessment of water potability based on the water quality parameters and regulatory standards.

Bar charts: Create bar charts to compare measured parameter values against regulatory standards. Parameters exceeding the standards can be highlighted to indicate non-potability.

Pie charts: Use pie charts to represent the percentage of potable and non-potable water samples in your dataset, providing a clear overview.

Map Visualization: If analyzing water sources across geographical locations, consider geographic information system (GIS) tools to create maps. Color-coded markers can indicate potable and non-potable water sources on the map.

**Predictive modelling:**

logistic regression is a machine learning algorithm primarily used for binary classification tasks, where the goal is to predict one of two possible outcomes based on input features. In your case, you want to predict water potability, which is typically a binary classification problem (potable or non-potable). Here's an explanation of logistic regression and features you can use for predicting water potability.

**Data set:**

For this piece of analysis, the Water Quality dataset has been taken from Kaggle¹.

Dataset link:

<https://www.kaggle.com/datasets/adityakadiwal/water-potability/>

About dataset

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

**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.

**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.

**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.

**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.

**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.

**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.

**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.

**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.

**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.

**10. Potability:** Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

**Methods and Algorithms**:

A jupyter notebook instance with Python code was used for processing.

Following are the list of algorithms that are used in this notebook.

* Logistic regression
* Decision tree
* Random tree
* SVM
* Adaboost

Understanding the data:

Firstly, we need to understand the data that we are working with. As the file format is a csv file, the standard pandas import statement using read\_csv will be used.

# Import the dataset for review as a DataFrame

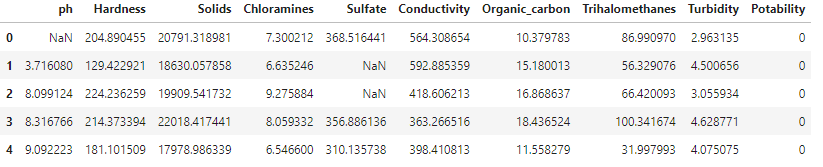
df = pd.read\_csv("../input/water-potability/water\_potability.csv")

# Review the first five observations

df.head()

Having imported the data, the code assigns the variable df with the DataFrame output results from the pandas method.

As with any dataset that you will process, reviewing a sample of records will help you to gain comfort. A DataFrame has a large number of methods associated with it, with the pandas API a great resource to use. Within the API a head method can be used. Output 1.1 shows the first 5 rows of the DataFrame by default. In order to produce a larger number of rows to be displayed a numeric value would be required inside the parenthesis. Two alternatives could be applied to sample the DataFrame with i) sample (df.sample()) selecting random rows from the index, or ii) tail (df.tail()) selecting the last n rows from the index.



Output 1.1 First five record details from the DataFrame

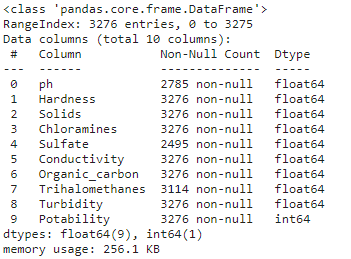
When running any method, the parenthesis is included after the method name allowing the Python interpreter to produce the result.

Displaying the memory of a DataFrame can be a common task, particularly when memory constraints are involved. An example is where the dataset to import is potentially larger than the memory available within the Python session. By using the pandas library a DataFrame is created in-memory so users should understand what memory can be used when performing these processing steps.

# Display information about the DataFrame - contains memory details

df.info(memory\_usage="deep")

The code above can be used as a method to display output 1.2. With the inclusion of the keyword memory\_usage, the Python interpreter is forced to do a deeper search to understand the memory usage that is displayed below. A default option would perform a general search to understand, so if accuracy in your assessment is required then ensure that the keyword phrase from above is applied.



Output 1.2 Provides an overview of the features and details of memory usage

From the results shown in output 1.2, it can show a range of details, from the column names and data types, to also confirming the class of the variable and number of non-null values. We can see that 3,276 rows are shown within the entire table. However, for the column Sulfate, there are only 2,495 non-null values present. Therefore, a number of missing values can be reviewed to understand if there is a pattern for these missing entries with other columns.

**Code generation for prediction:**

Start the water quality analysis task by importing the necessary

Python libraries and the dataset:

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

import plotly.express as px

import warnings

warnings.filterwarnings('ignore')

df.info

gives the various information properties for potability.

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 3276 entries, 0 to 3275

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 ph 2785 non-null float64

1 Hardness 3276 non-null float64

2 Solids 3276 non-null float64

3 Chloramines 3276 non-null float64

4 Sulfate 2495 non-null float64

5 Conductivity 3276 non-null float64

6 Organic\_carbon 3276 non-null float64

7 Trihalomethanes 3114 non-null float64

8 Turbidity 3276 non-null float64

9 Potability 3276 non-null int64

dtypes: float64(9), int64(1)

memory usage: 256.1 KB

|  | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic\_carbon | Trihalomethanes | Turbidity | Potability |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 2785.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 2495.000000 | 3276.000000 | 3276.000000 | 3114.000000 | 3276.000000 | 3276.000000 |
| mean | 7.080795 | 196.369496 | 22014.092526 | 7.122277 | 333.775777 | 426.205111 | 14.284970 | 66.396293 | 3.966786 | 0.390110 |
| std | 1.594320 | 32.879761 | 8768.570828 | 1.583085 | 41.416840 | 80.824064 | 3.308162 | 16.175008 | 0.780382 | 0.487849 |
| min | 0.000000 | 47.432000 | 320.942611 | 0.352000 | 129.000000 | 181.483754 | 2.200000 | 0.738000 | 1.450000 | 0.000000 |
| 25% | 6.093092 | 176.850538 | 15666.690297 | 6.127421 | 307.699498 | 365.734414 | 12.065801 | 55.844536 | 3.439711 | 0.000000 |
| 50% | 7.036752 | 196.967627 | 20927.833607 | 7.130299 | 333.073546 | 421.884968 | 14.218338 | 66.622485 | 3.955028 | 0.000000 |
| 75% | 8.062066 | 216.667456 | 27332.762127 | 8.114887 | 359.950170 | 481.792304 | 16.557652 | 77.337473 | 4.500320 | 1.000000 |
| max | 14.000000 | 323.124000 | 61227.196008 | 13.127000 | 481.030642 | 753.342620 | 28.300000 | 124.000000 | 6.739000 | 1.000000 |

*#unstacking to see correaltion*

corr = df.corr()

c1 = corr.abs().unstack()

c1.sort\_values(ascending = False)[12:24:2]

Hardness Sulfate 0.106923

ph Solids 0.089288

Hardness ph 0.082096

Solids Chloramines 0.070148

Hardness Solids 0.046899

ph Organic\_carbon 0.043503

dtype: float64

ax = sns.countplot(x = "Potability",data= df, saturation=0.8)

plt.xticks(ticks=[0, 1], labels = ["Not Potable", "Potable"])

plt.show()



:

x = df.Potability.value\_counts()

labels = [0,1]

print(x)

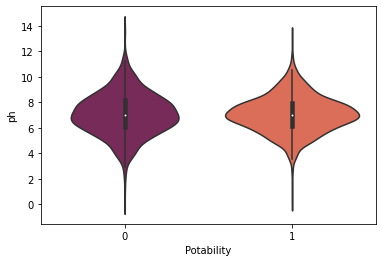
0 1998

1 1278

Name: Potability, dtype: int64

sns.violinplot(x='Potability', y='ph', data=df, palette='rocket')

<AxesSubplot:xlabel='Potability', ylabel='ph'>



Using Logistic Regression:

Logistic Regression is particularly useful in estimating the vulnerability of aquifers,which are underground layers of water bearing permeable rock from which groundwater can be extracted.

The logistic regression models relate the probability of a contaminant concentration exceeding a threshold concentration to a set of possible influencing variables.

from sklearn.linear\_model import LogisticRegression

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

*# Creating model object*

model\_lg = LogisticRegression(max\_iter=120,random\_state=0, n\_jobs=20)

*# Training Model*

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

LogisticRegression(max\_iter=120, n\_jobs=20, random\_state=0)

*# Making Prediction*

pred\_lg = model\_lg.predict(X\_test)

*# Calculating Accuracy Score*

lg = accuracy\_score(y\_test, pred\_lg)

print(lg)

0.6284658040665434

print(classification\_report(y\_test,pred\_lg))

precision recall f1-score support

0 0.63 1.00 0.77 680

1 0.00 0.00 0.00 402

accuracy 0.63 1082

macro avg 0.31 0.50 0.39 1082

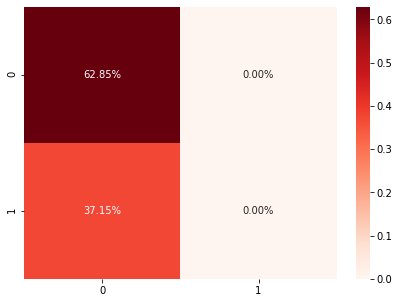
weighted avg 0.39 0.63 0.49 1082

*# confusion Maxtrix*

cm1 = confusion\_matrix(y\_test, pred\_lg)

sns.heatmap(cm1/np.sum(cm1), annot = True, fmt= '0.2%', cmap = 'Reds')

<AxesSubplot:>



**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.

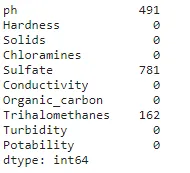
# Check for the missing values by column

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.

Output:



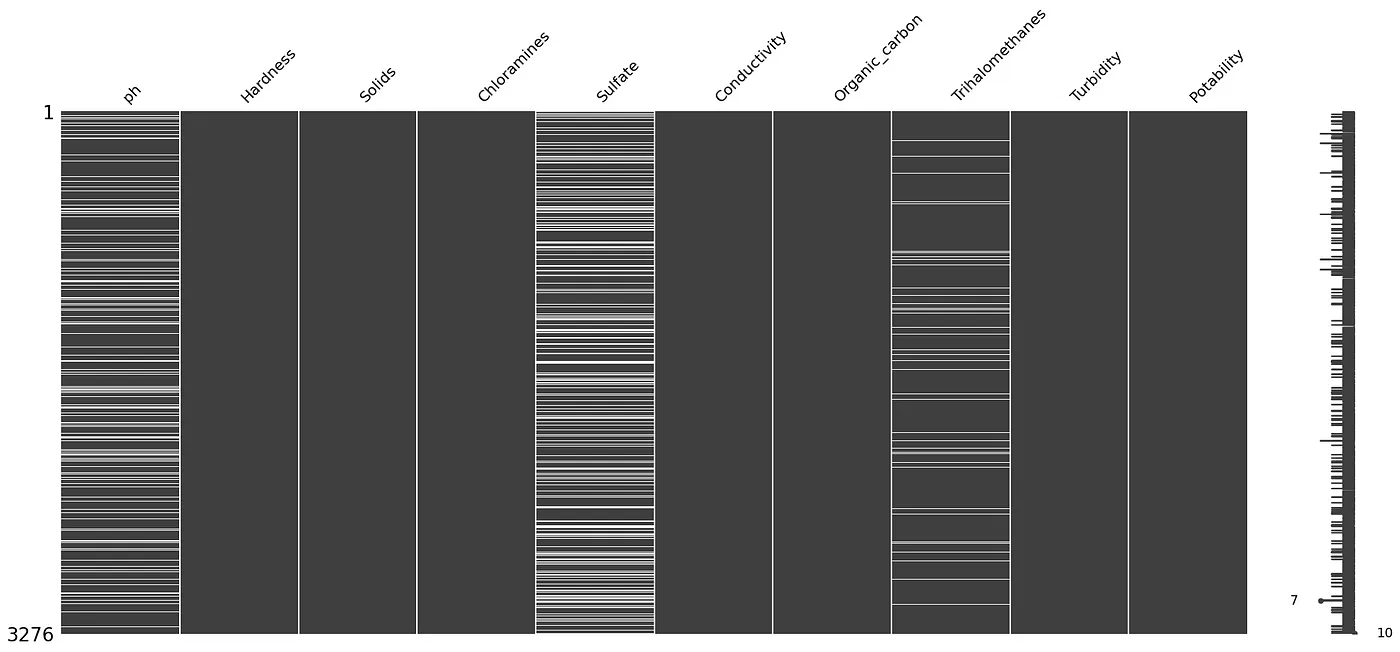
Having the total count of rows with missing values is a great starting point. However, it would be better to review the proportion of missing values within a column.

**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 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:**

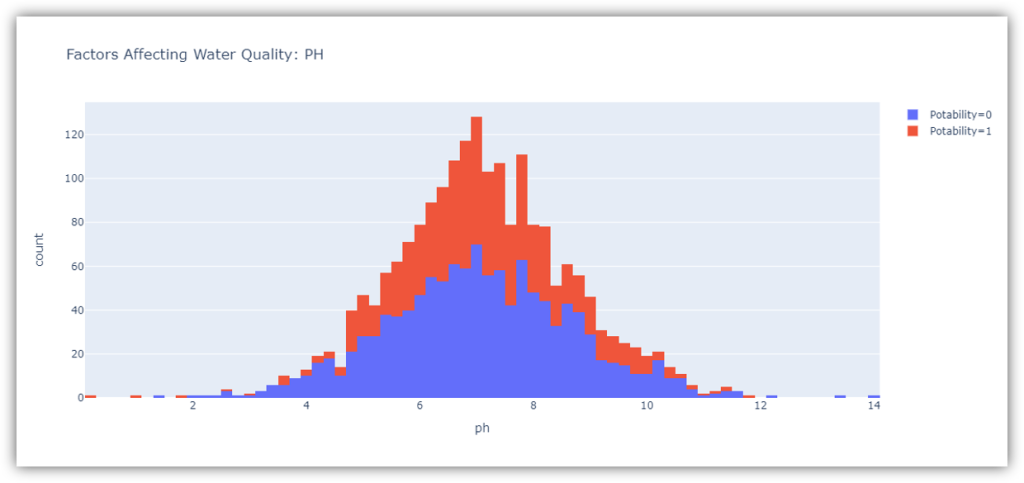
import plotly.express as px

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.

**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.

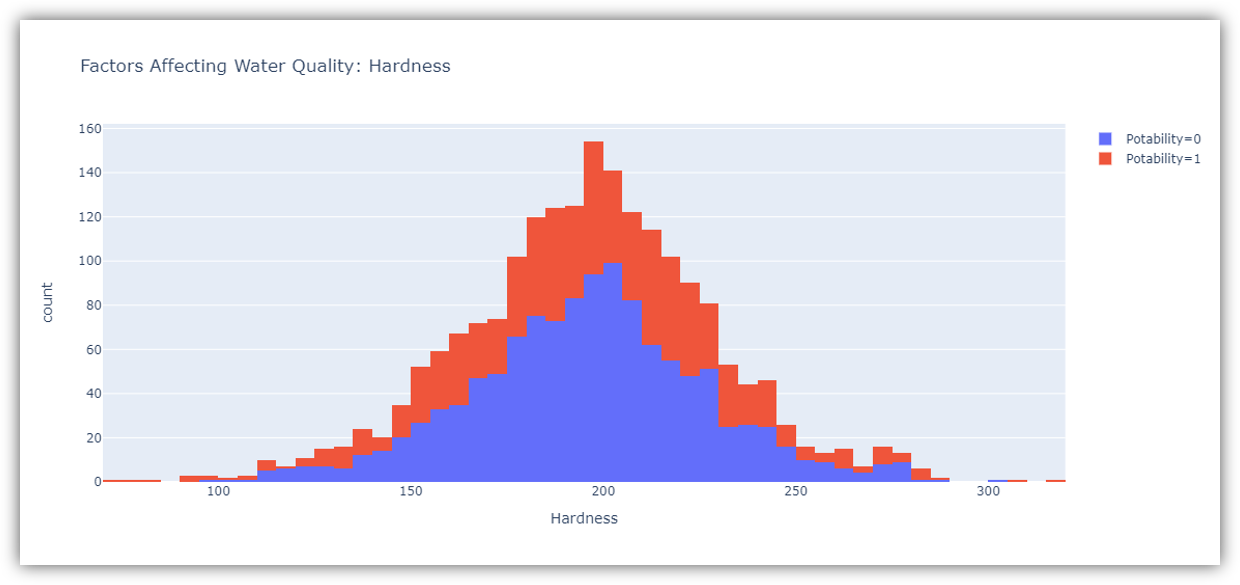
**Code:**

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

color = "Potability",

title= "Factors Affecting Water Quality: Hardness")

figure.show()



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.

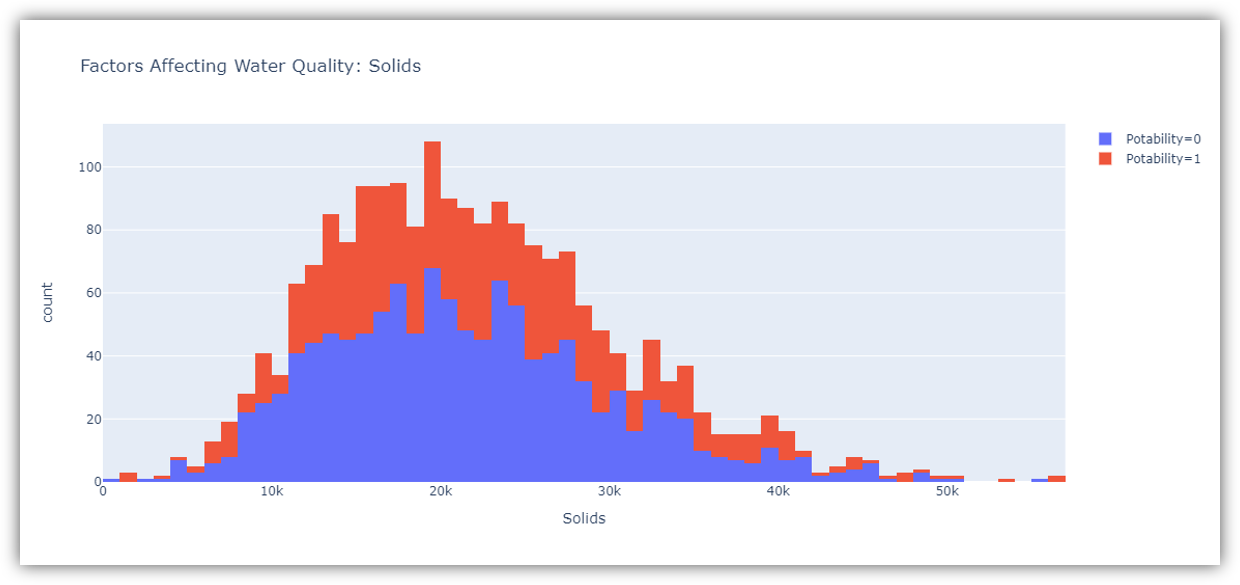
**Code:**

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

color = "Potability",

title= "Factors Affecting Water Quality: Solids")

figure.show()



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.

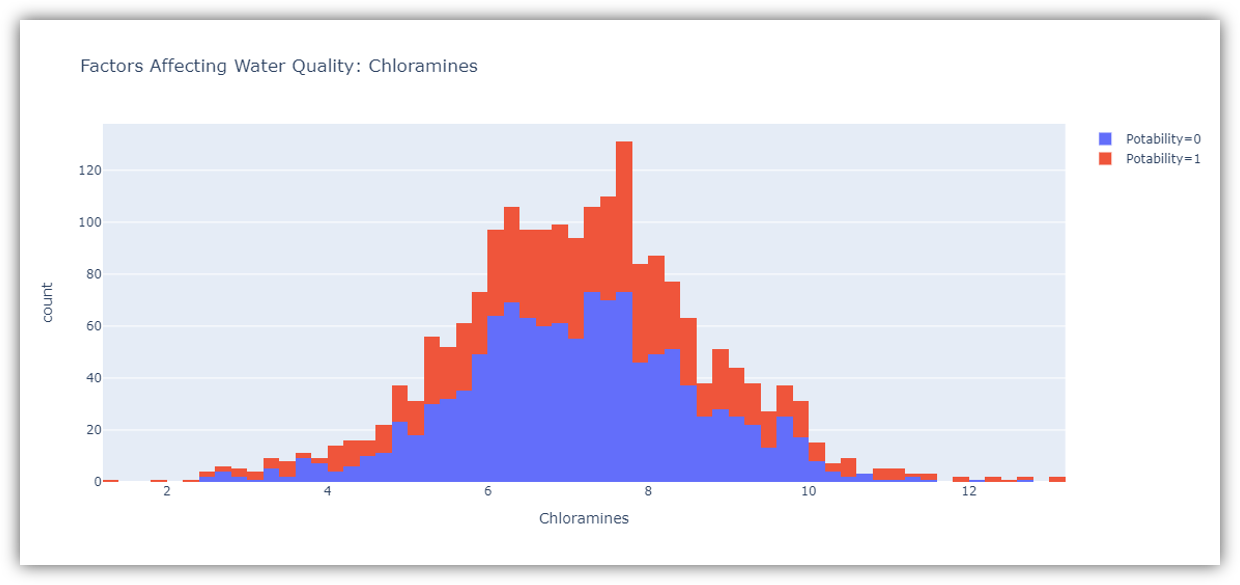
**Code:**

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

color = "Potability",

title= "Factors Affecting Water Quality: Chloramines")

figure.show()



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.

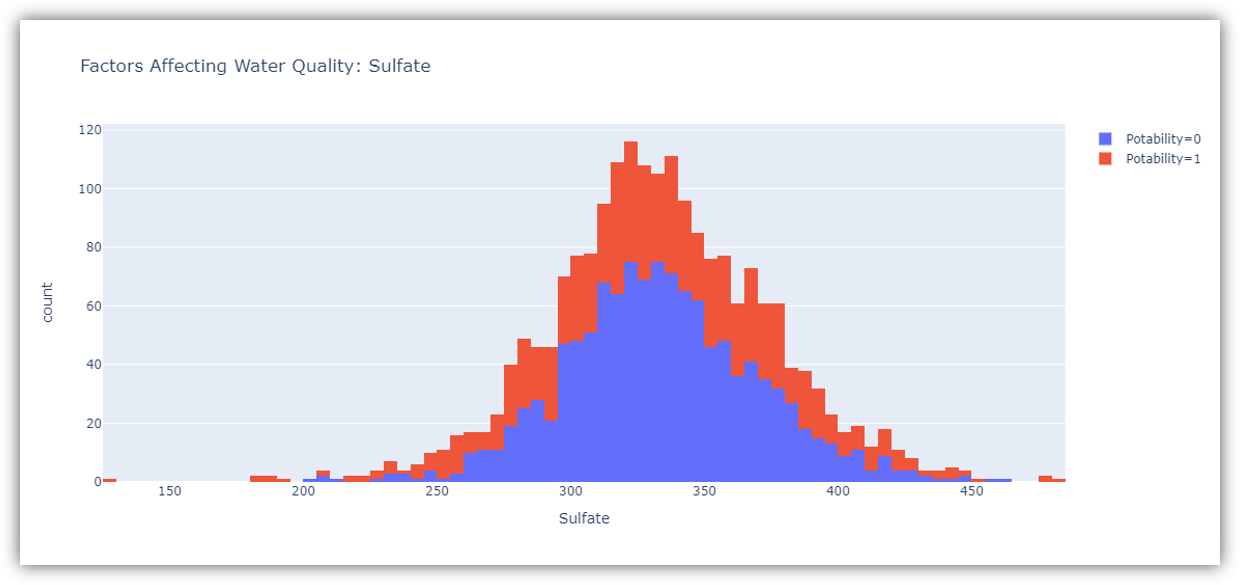
**Code:**

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

color = "Potability",

title= "Factors Affecting Water Quality: Sulfate")

figure.show()



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.

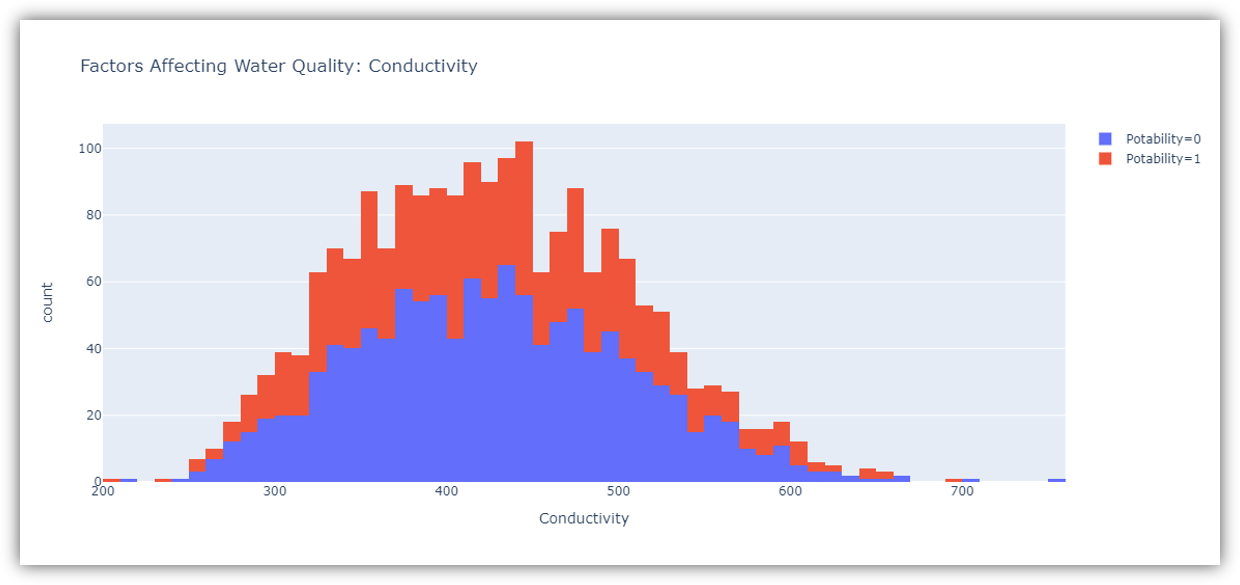
**Code:**

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

color = "Potability",

title= "Factors Affecting Water Quality: Conductivity")

figure.show()



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.

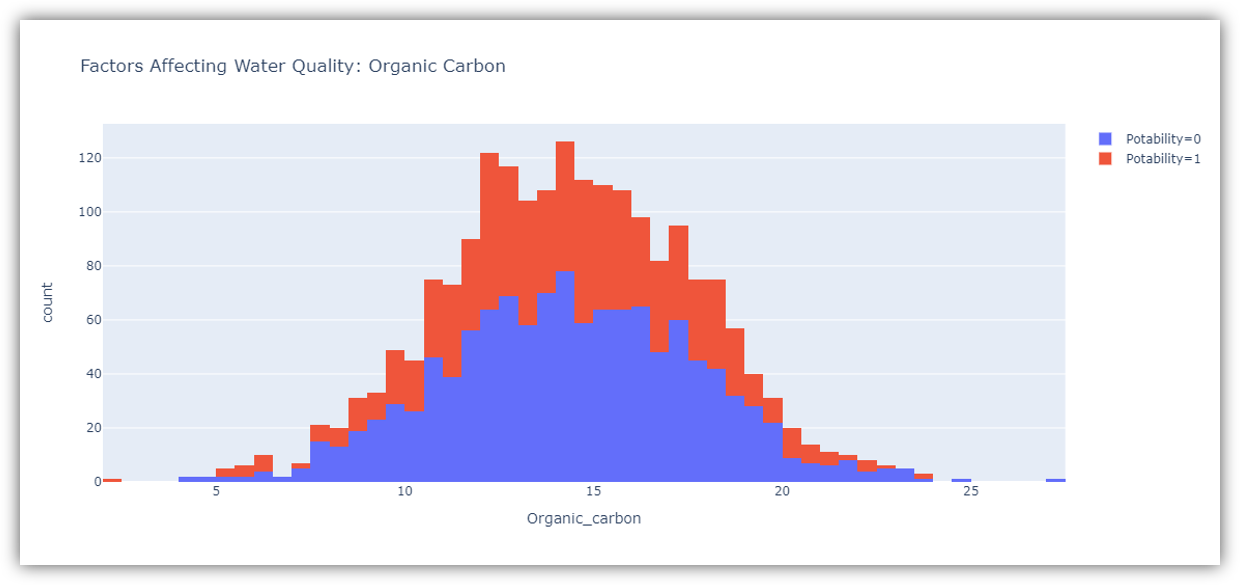
**Code:**

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

color = "Potability",

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

figure.show()



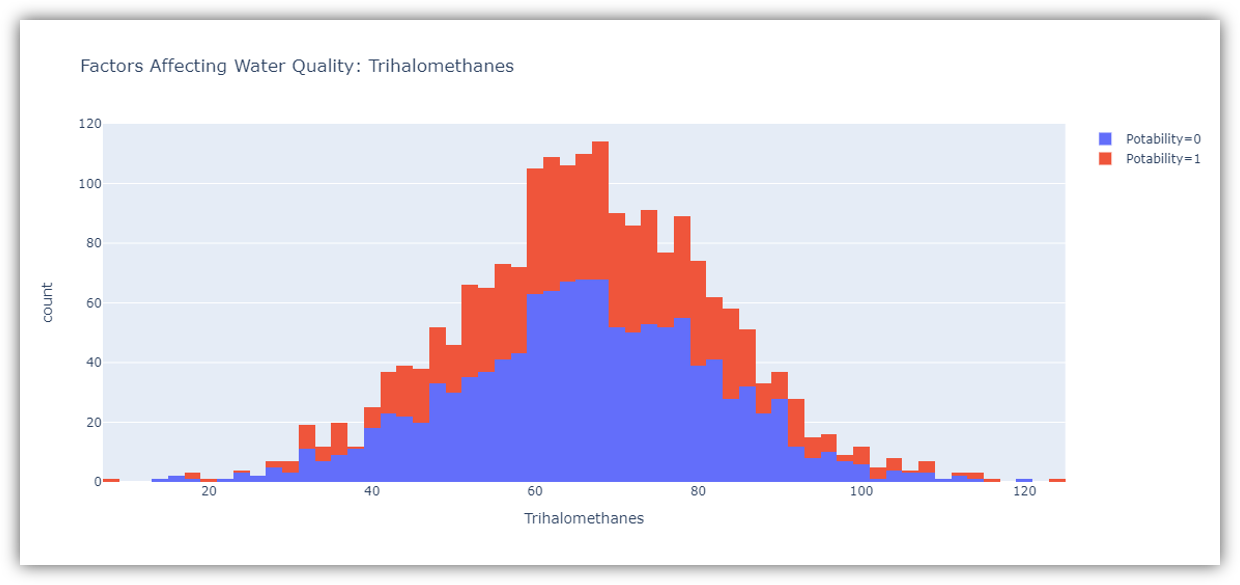
**8. Trihalomethanes:** THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being 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()

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

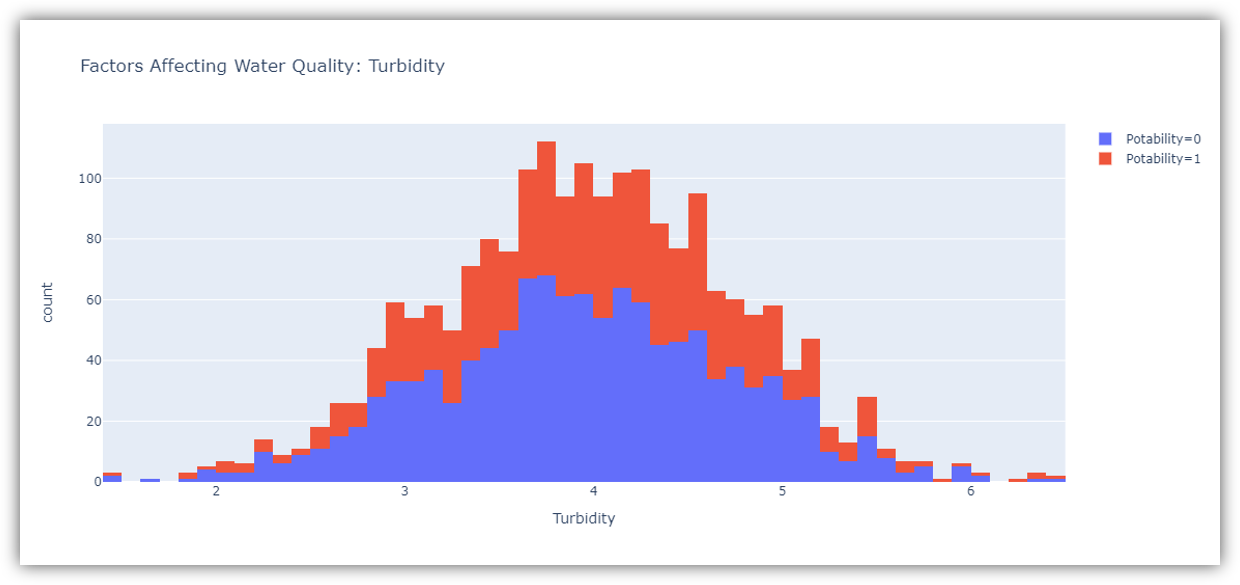
**9. Turbidity:** The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for 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")

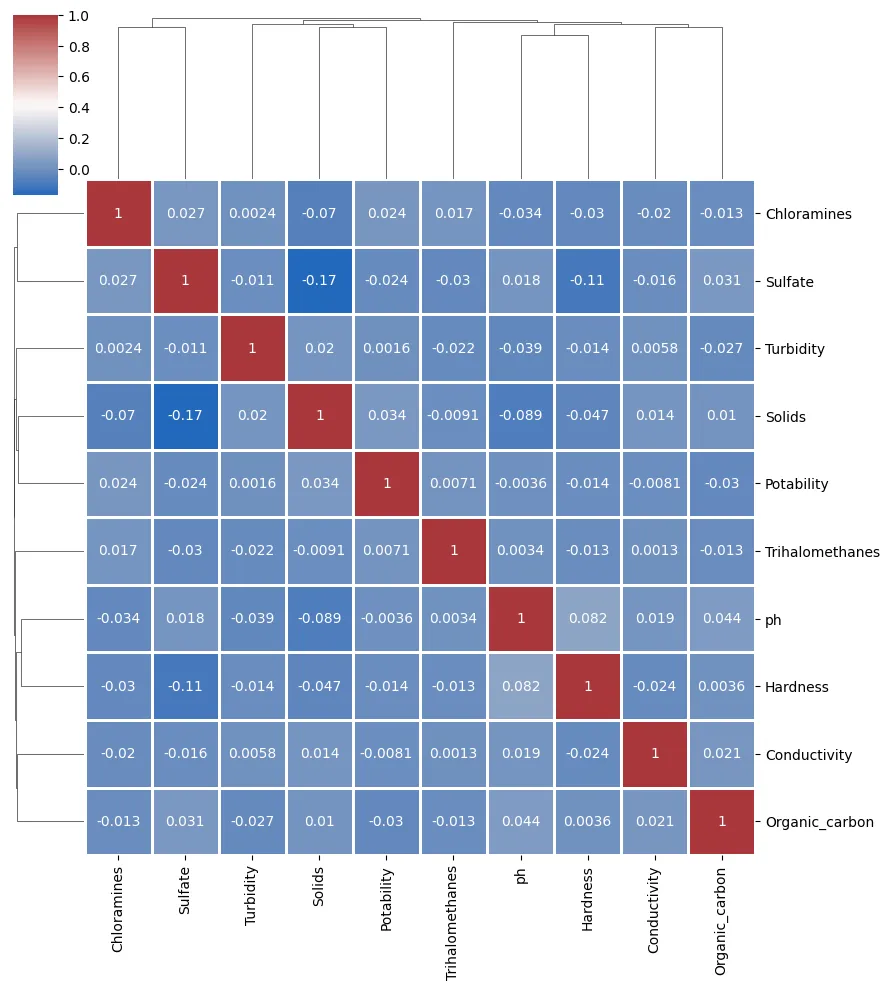
figure.show()

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

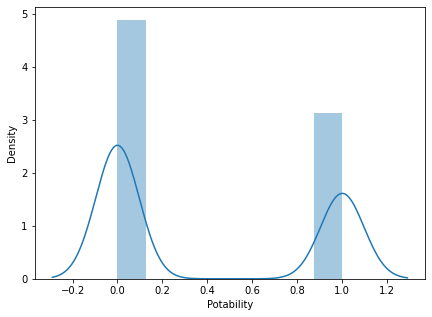
standard\_pH = 7.0 # Example standard pH value

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

plt.title('pH vs. Standard')

plt.ylabel('pH Value')

plt.show()



|  | **visualization:** |
| --- | --- |

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.

**potly:**

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:**

importpandasaspd importnumpyasnp

importmatplotlib.pyplot asplt importseabornassns

# Generatingsample waterqualitydata 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)

# Histogramsforeachwater quality parameter plt.figure(figsize=(12,10))

plt.subplot(2, 3, 1) sns.histplot(water\_quality["pH"],kde=True) plt.title("pHDistribution")

⁸

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()

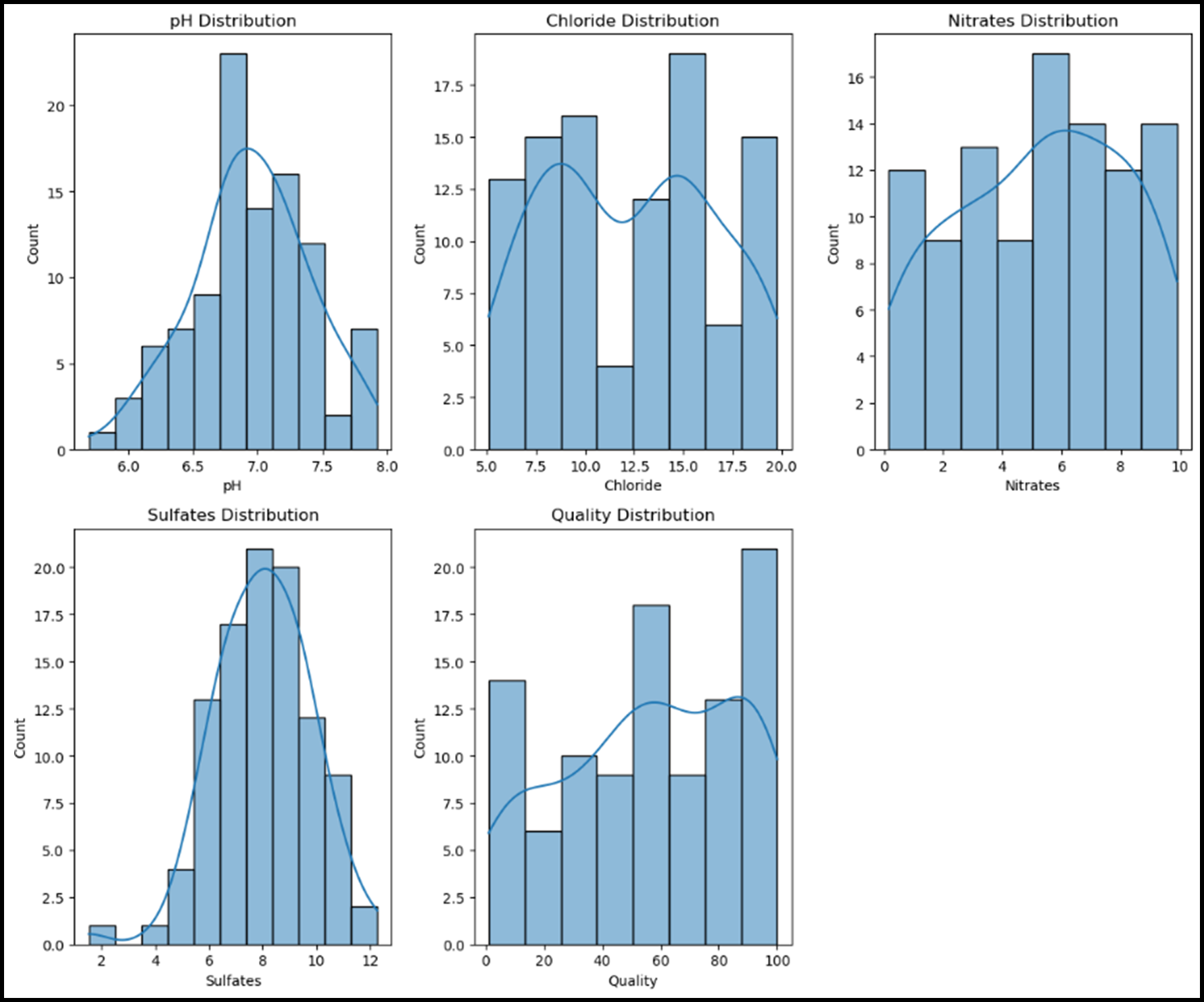
# Scatterplotmatrixforcorrelationvisualization sns.pairplot(water\_quality)

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()



**Correlation**

**Code:**

importpandasaspd importnumpyasnp importseabornassns

importmatplotlib.pyplot asplt

# Generatingsyntheticwaterqualitydata 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)

}

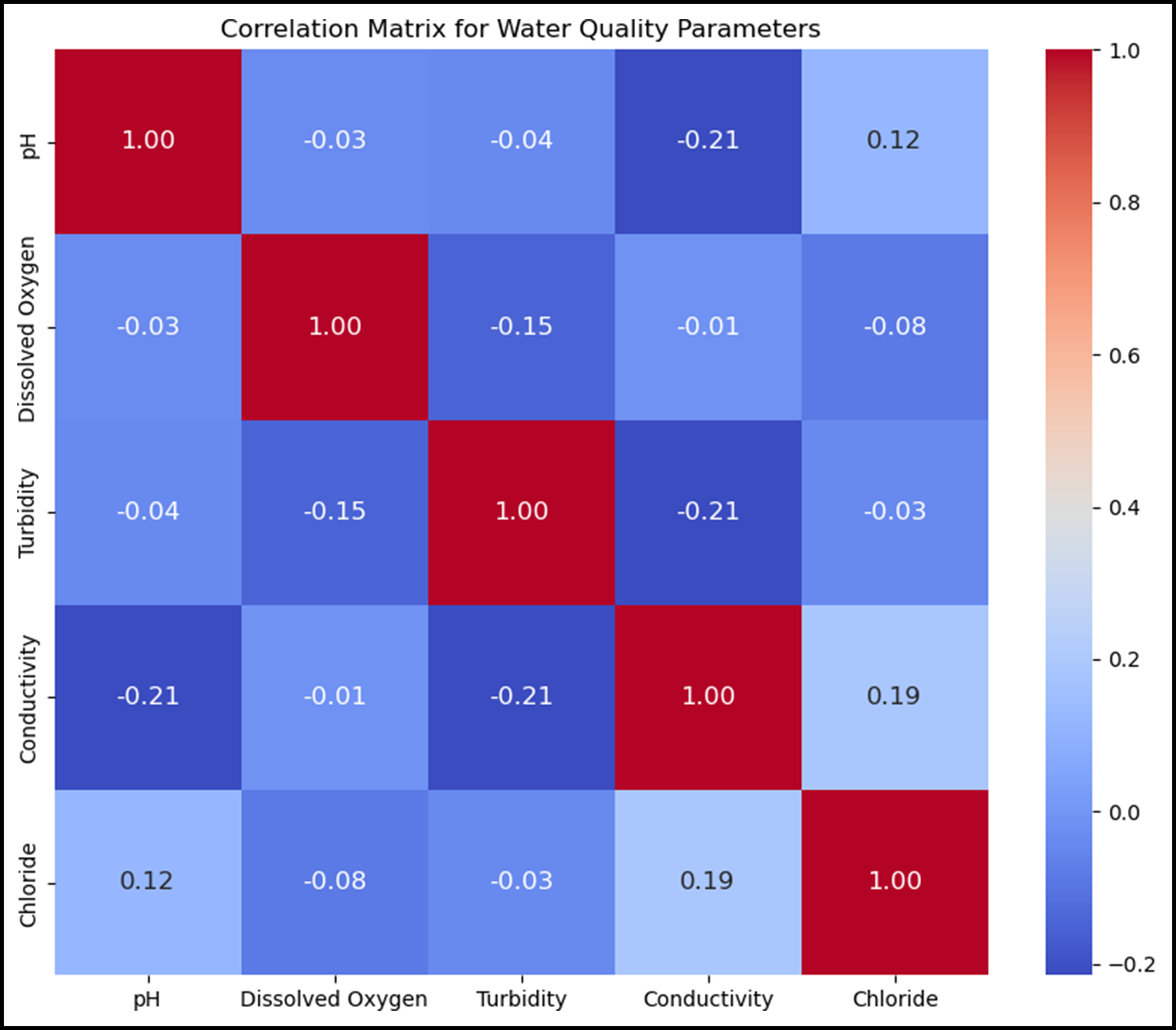
df=pd.DataFrame(data)

# Calculatingthecorrelationmatrix

corr=df.corr()

# Plotting the correlationmatrixusingSeaborn plt.figure(figsize=(10,8))

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

plt.title('Correlation Matrixfor Water Quality Parameters') plt.show() 

**Scatterplot**

**Code:**

importpandasaspd

importmatplotlib.pyplot asplt importseabornassns

# Samplewaterqualitydata (replacethiswithyourowndataset) 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)

# Createa scatterplotusingMatplotlib

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

# Using Matplotlib

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

plt.ylabel('Chlorine (ppm)') plt.title('ScatterplotofpHvs Chlorine') plt.legend()

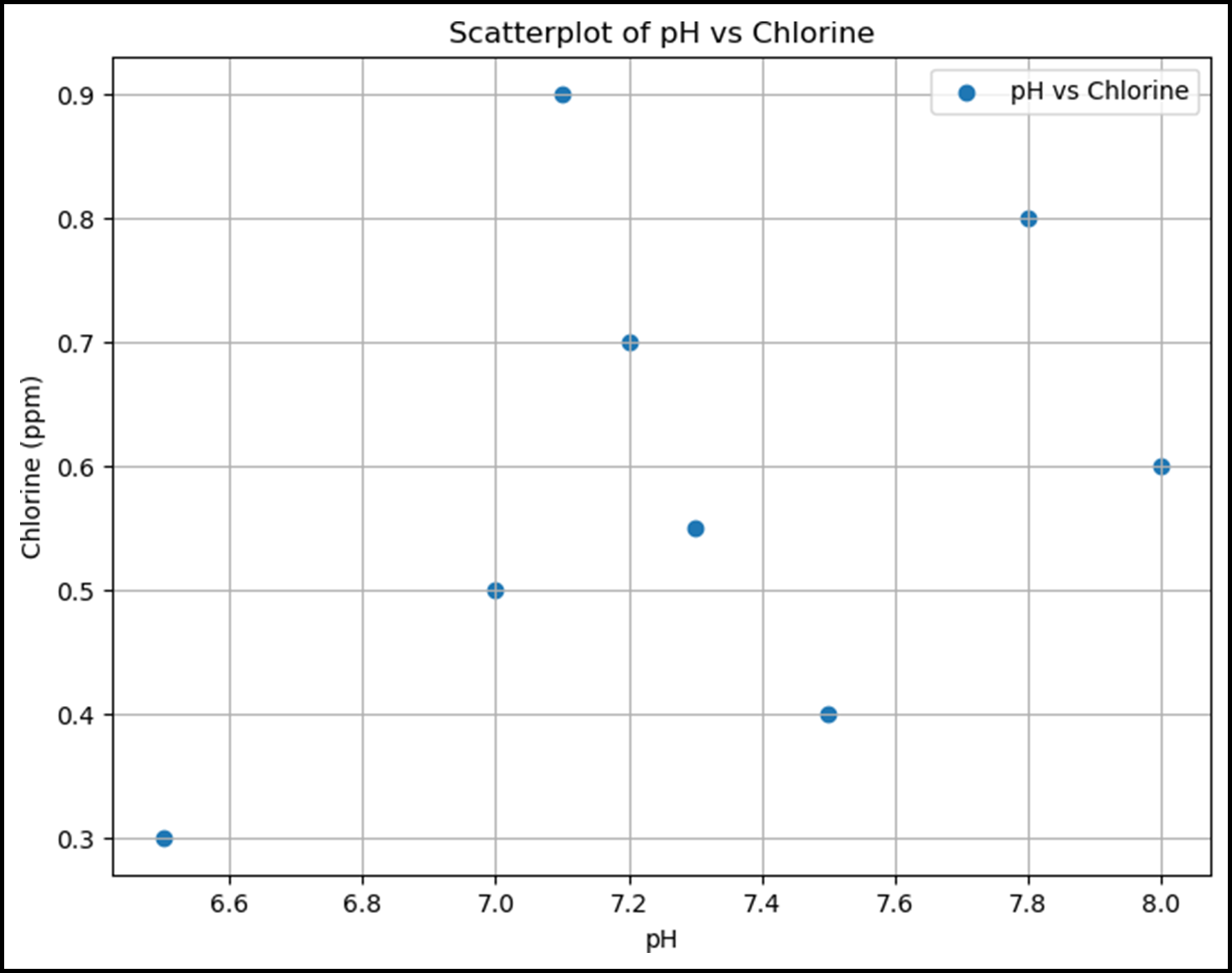
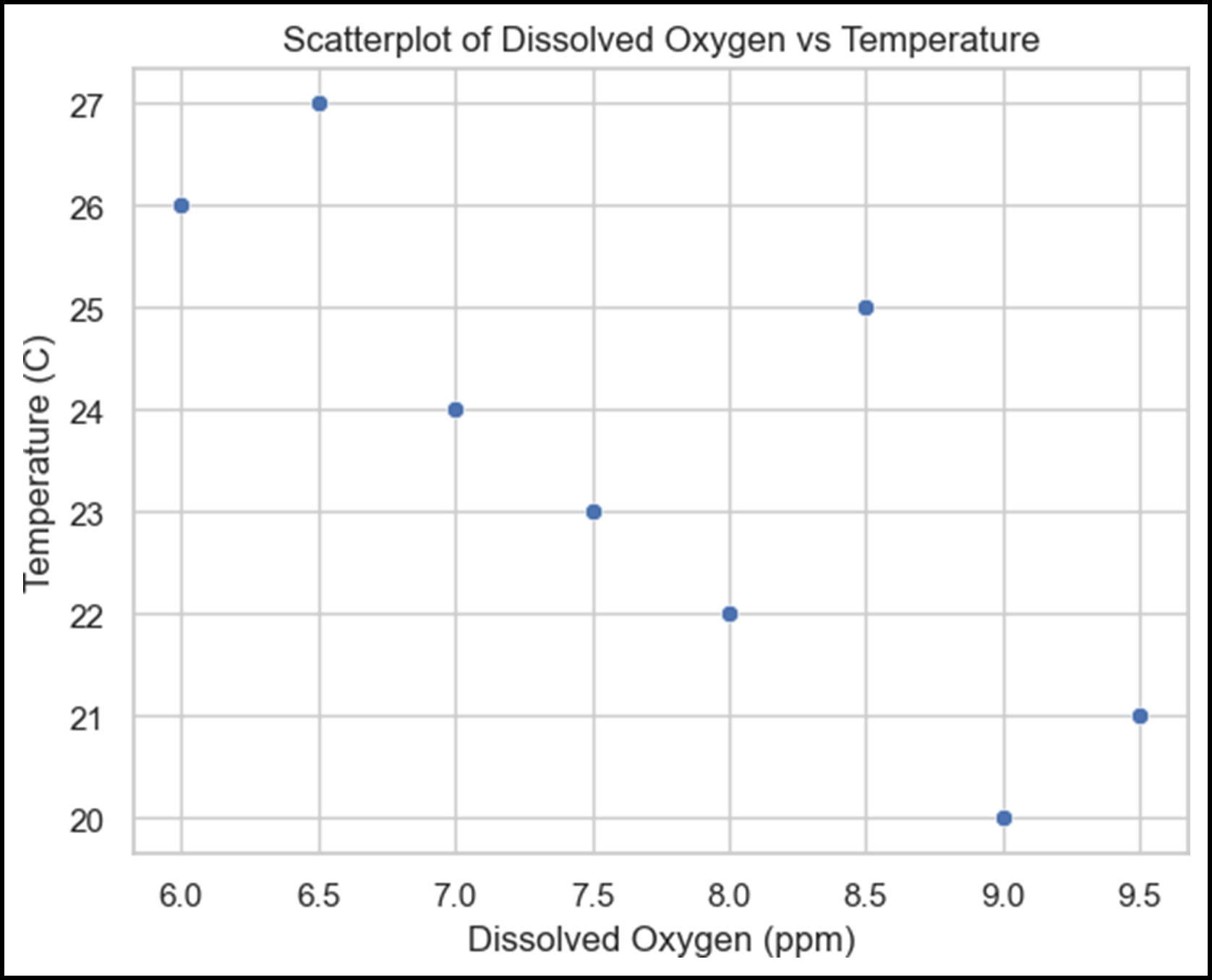
plt.grid(True) plt.show()

#Using Seaborn foranotherscatterplot sns.set(style='whitegrid')

sns.scatterplot(x='Dissolved Oxygen (ppm)',

y='Temperature (C)', data=df)

plt.title('Scatterplot of Dissolved OxygenvsTemperature') plt.show()



**Predictivemodel**

**Logisticregression**

**code:**

importpandasaspd

fromsklearn.model\_selection importtrain\_test\_split fromsklearn.linear\_model import LogisticRegression

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

importmatplotlib.pyplotasplt importseaborn assns

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

#Check the firstfewrowsof the dataset print(data.head())

#Separate features(waterquality parameters) andthetargetvariable X=data.drop('Potability',axis=1)

y= data['Potability']

#Splitthe dataintotrainingandtestingsets(80%train, 20%test)

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

#Createalogisticregressionmodel

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}")

#Createaconfusion 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("ClassificationReport:")

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

**Randomforest**

**code:**

importpandasaspd

fromsklearn.model\_selectionimport train\_test\_split fromsklearn.ensembleimport RandomForestClassifier

fromsklearn.metricsimportaccuracy\_score, classification\_report, confusion\_matrix

importseabornassns

importmatplotlib.pyplot asplt

# Loadthe water quality dataset(replacewithyourdataset)

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())

# Checkformissingvaluesandhandleifneeded print(data.isnull().sum()) # Checkfor missingvalues

# Handlemissingvalues(ifany)

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

# Definefeaturesandtargetvariable

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

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

# Splitthe data intotrainingandtestingsets

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

# Initialize the Random Forestclassifier

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

# Trainthe model rf\_classifier.fit(X\_train,y\_train)

# Makepredictions

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

# Evaluate the model

accuracy=accuracy\_score(y\_test,predictions) print(f"Accuracy: {accuracy:.2f}")

# Generate aclassificationreportandconfusion 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

ph 491

Hardness

0 31.997993 4.075075 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

0.65 0.84 0.73 231

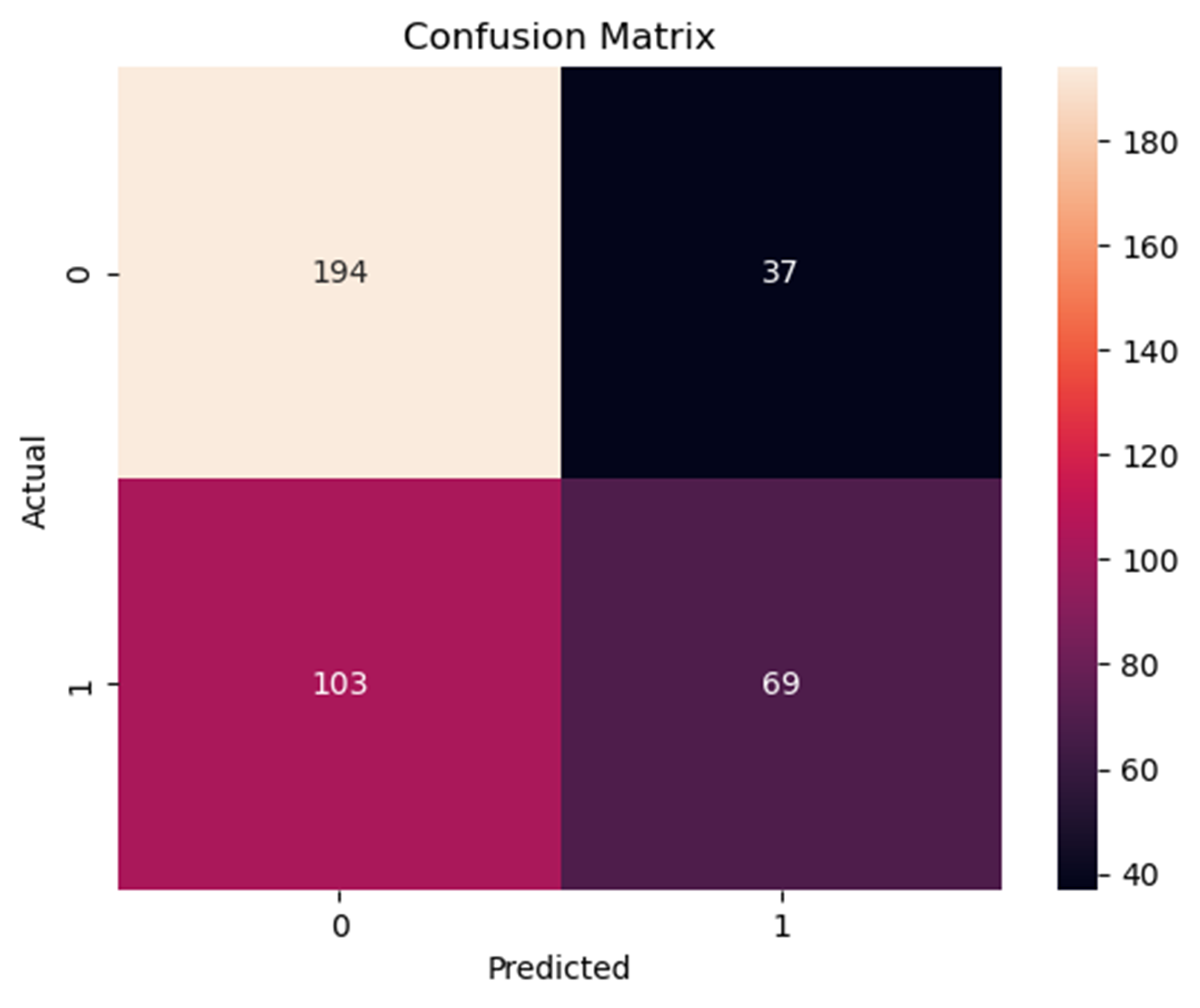
0.65 0.40 0.50 172

accuracy 0.65 403

macroavg 0.65 0.62 0.62 403

weightedavg 0.65 0.65 0.63 403

**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 that water quality parameters such as pH, turbidity, dissolved oxygen, and pollutant levelsplay a crucial role in assessing the overall health of water bodies. 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 waterqualityinthe region.