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 2:INNOVATION

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.

**Introduction:**

Being able to provide enough fresh drinking water is a core requirement. Within the climate change debate, one of the largest challenges is ensuring enough freshwater to survive. Water quality is a big concern that impacts all the specifies. Only about three percent of Earth’s water is freshwater. Of that, only 1.2 percent can be used as drinking water, with the remainder locked up in glaciers, ice caps, and permafrost, or buried deep in the ground. Using a data-driven approach to assess the features that impact the water quality could greatly improve our understanding of what makes water drinkable.

We will seek to find hidden insights with data analysis techniques using pandas and numpy. For the data visualizations, the matplotlib and seaborn libraries will be used.

**Goal:**

The goal of water quality analysis using data analytics is to assess and monitor the quality of water in various environmental settings, such as natural bodies of water, industrial processes, drinking water supplies, and wastewater treatment systems. This analysis aims to ensure the safety of water for human consumption, protect aquatic ecosystems, and maintain overall environmental health. Data analytics plays a crucial role in achieving this goal by providing insights, patterns, and predictions from large datasets related to water quality.

**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/>

**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:**

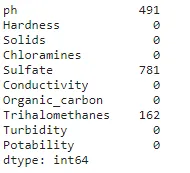
There are a number of missing values within the DataFrame. To confirm if this is correct we can apply the code block below.

# 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:

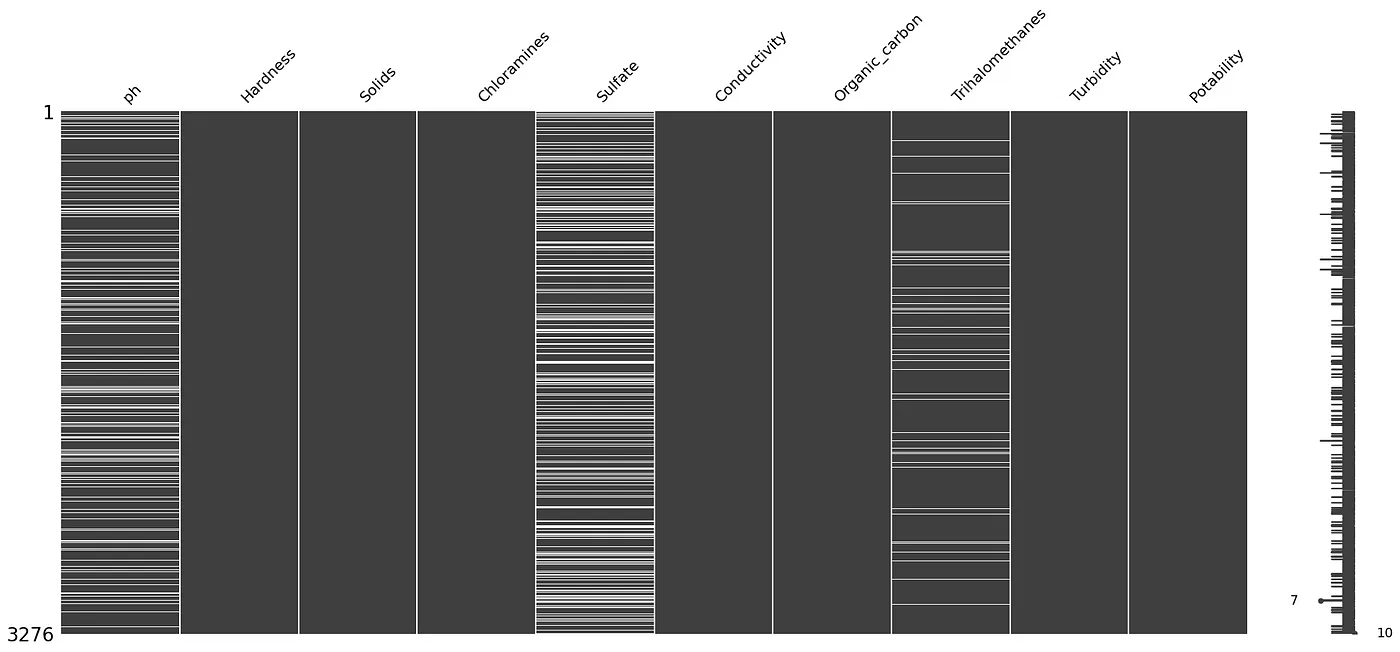
Having the total count of rows with missing values is a great starting point. However, it would be better to review the proportion of missing values within a column.

## **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:**

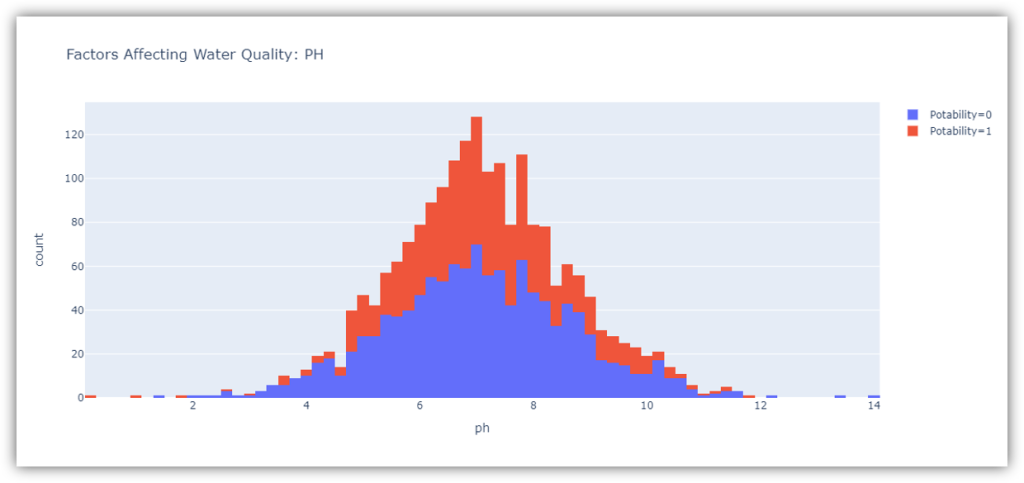
import plotly.express as px

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.

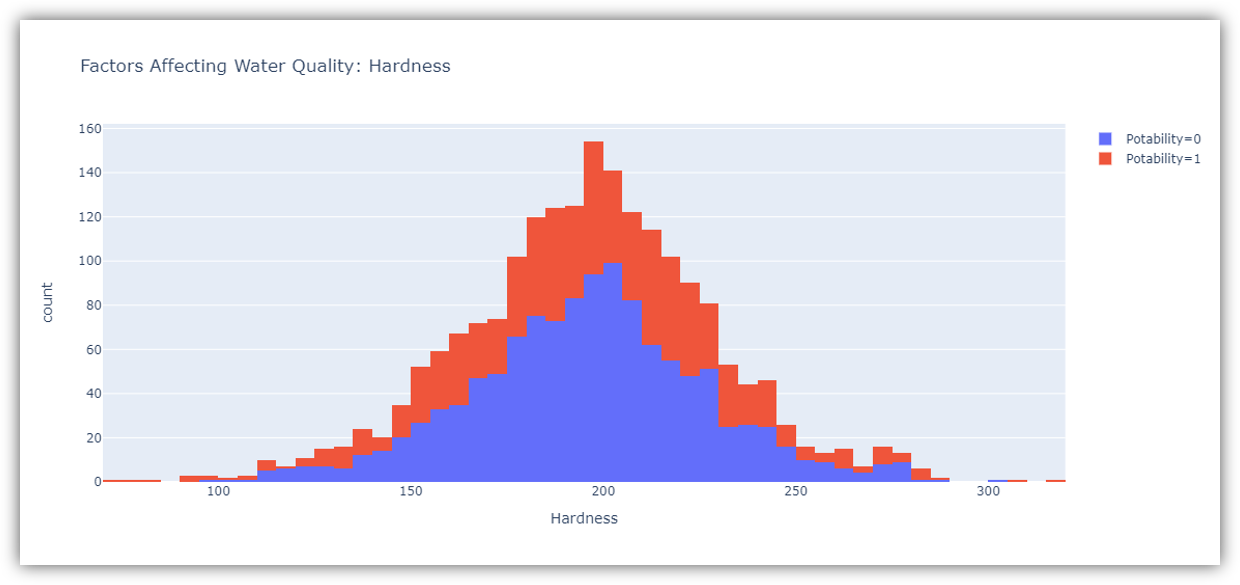
**Code:**

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.

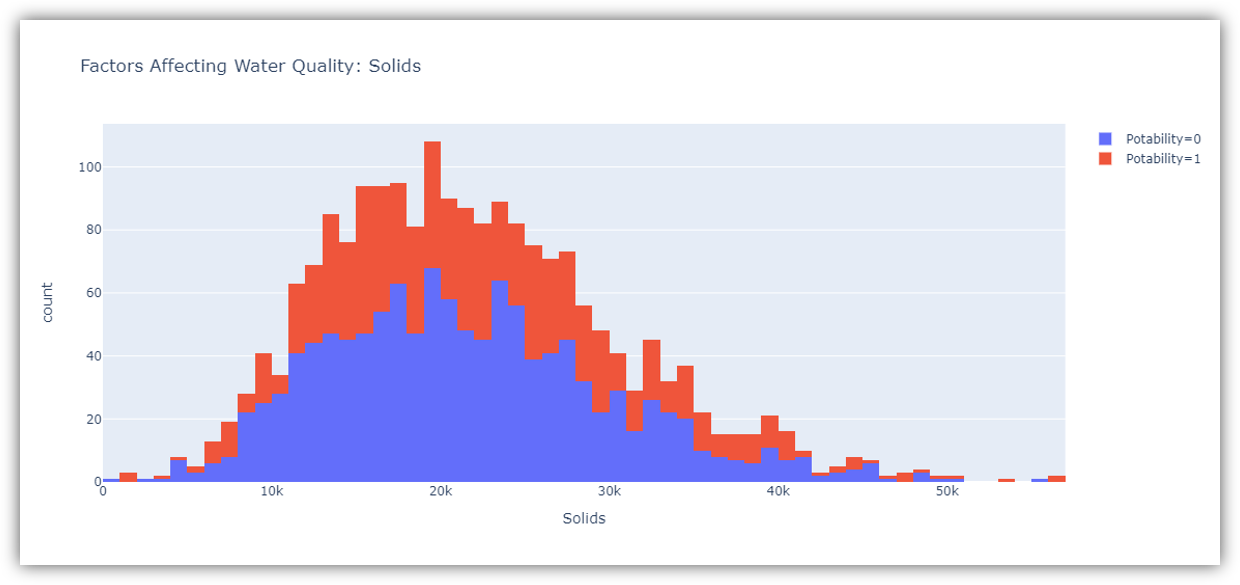
**Code:**

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.

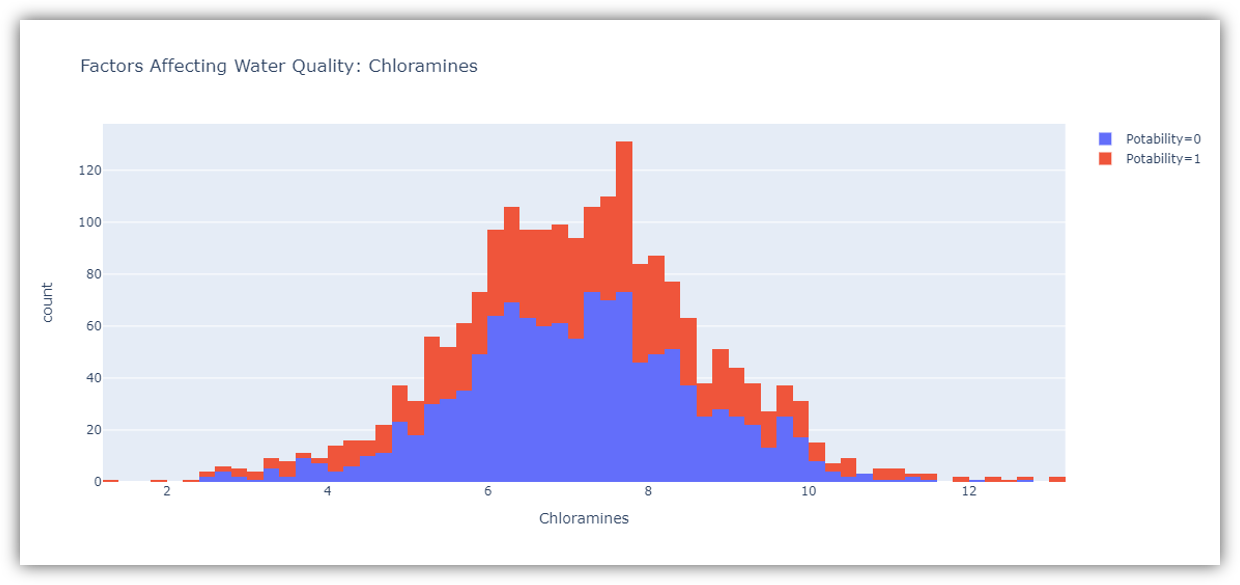
**Code:**

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.

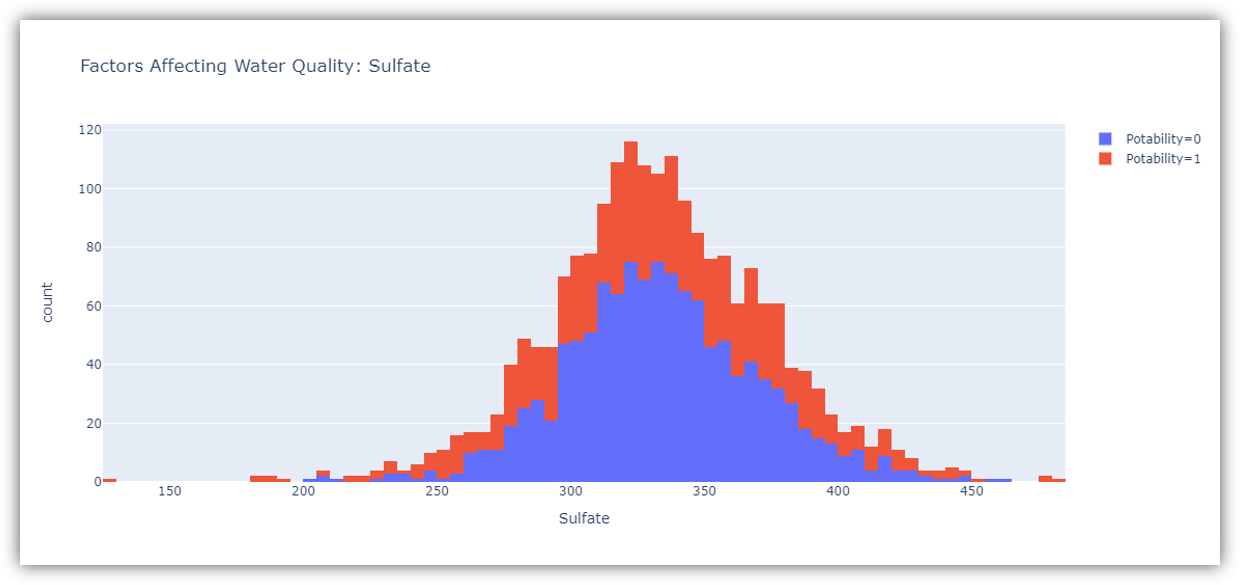
**Code:**

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.

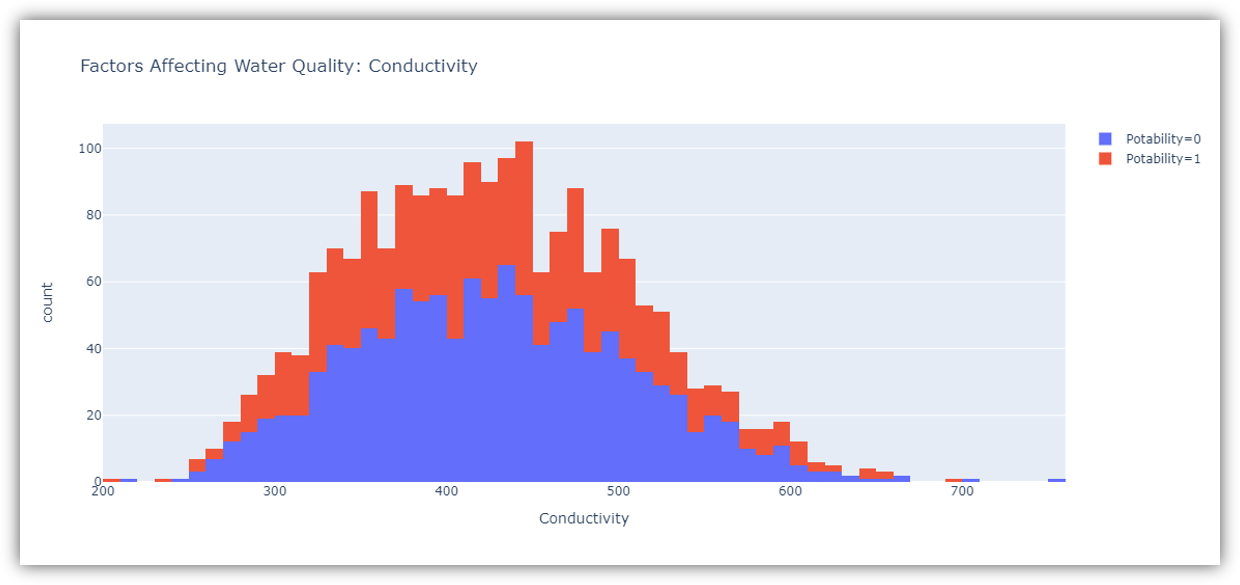
**Code:**

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.

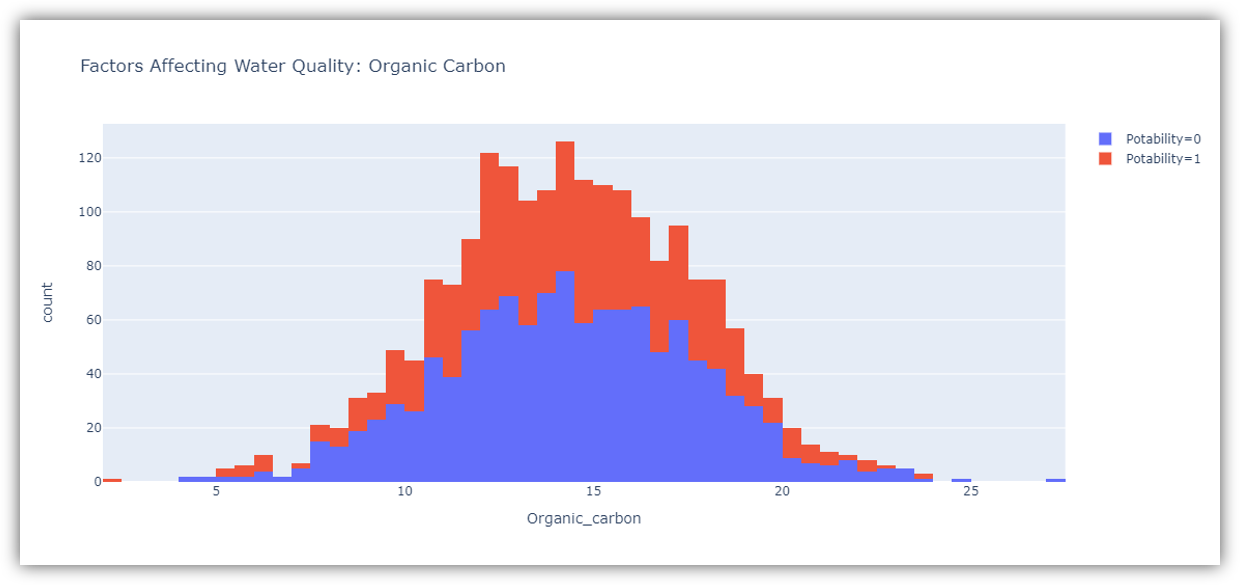
**Code:**

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

color = "Potability",

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

figure.show()



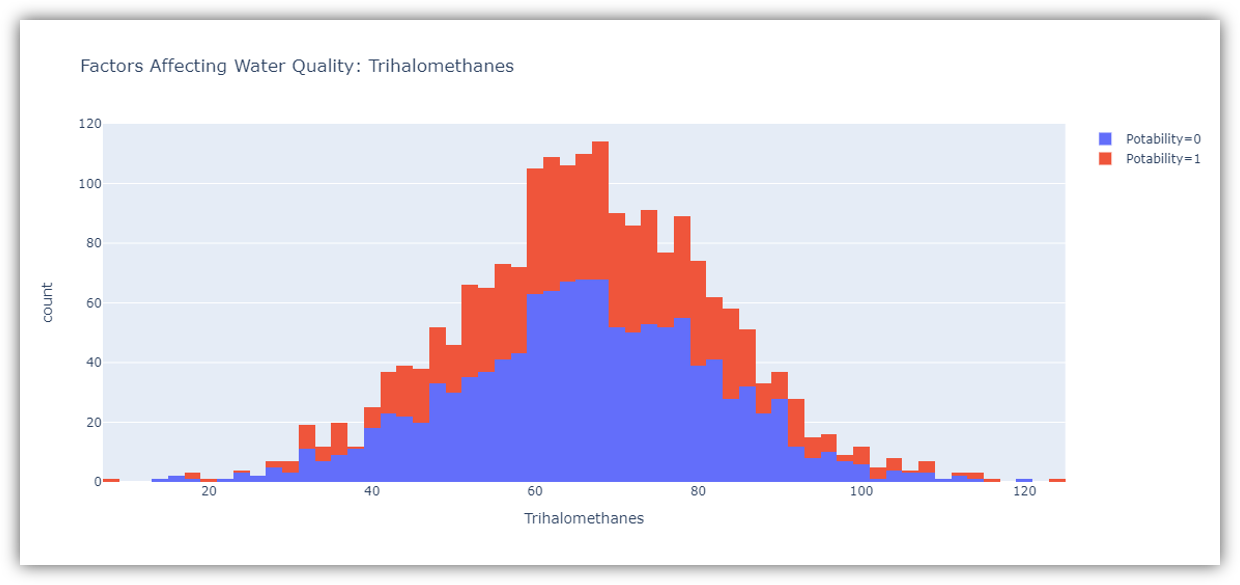
**8. Trihalomethanes:** THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being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

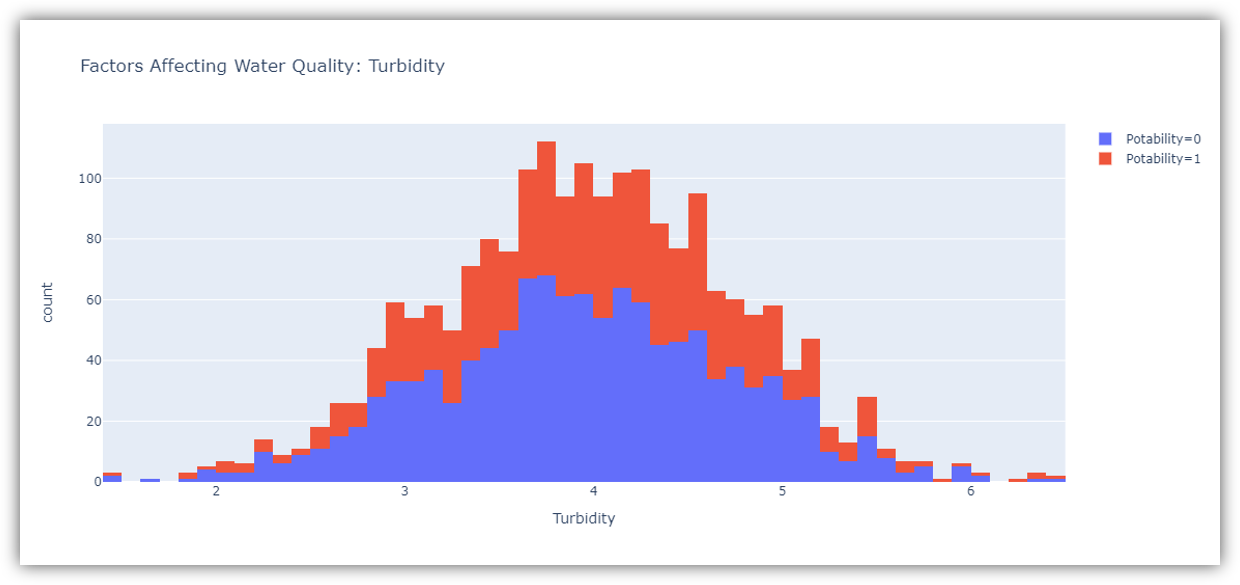
**9. Turbidity:** The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for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

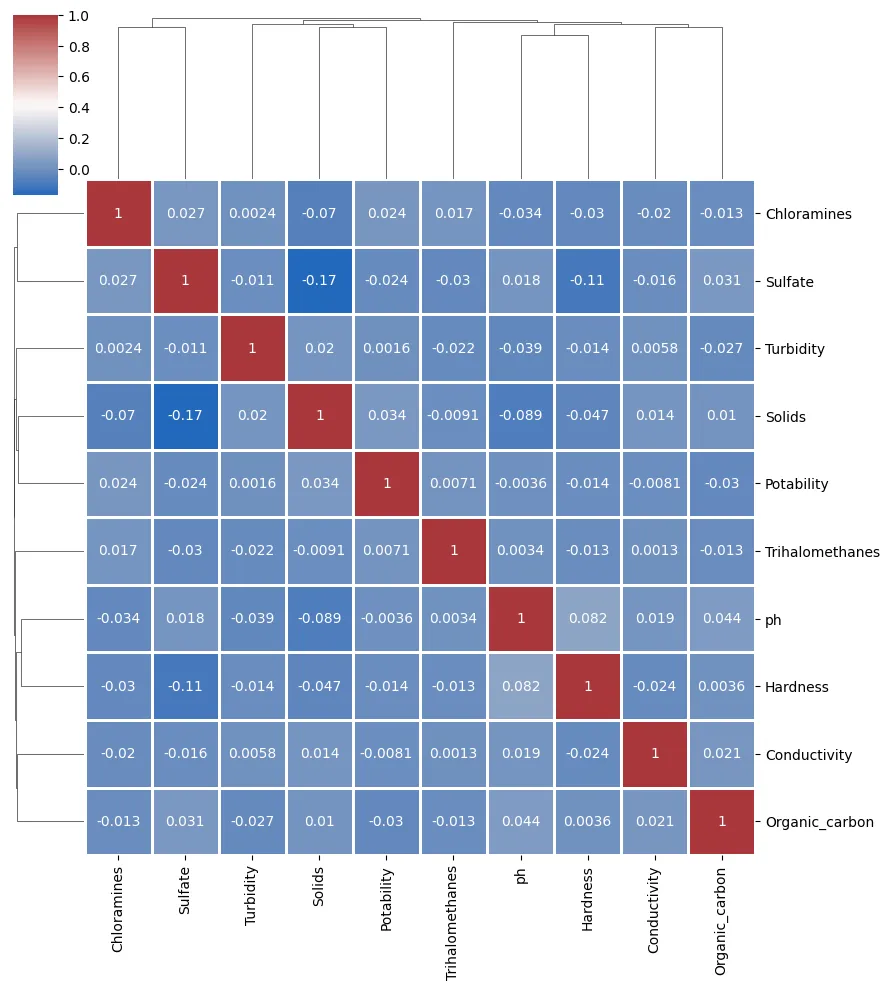
figure.show()

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

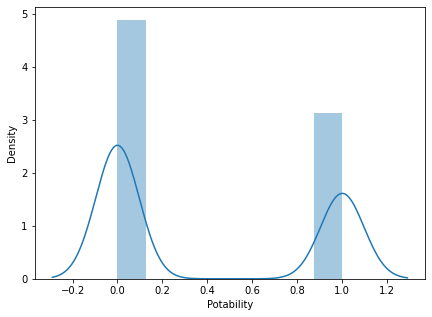
standard\_pH = 7.0 # Example standard pH value

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

plt.title('pH vs. Standard')

plt.ylabel('pH Value')

plt.show()



Conclusion:

In conclusion, the process of water quality analysis is greatly enhanced by effective data preprocessing and exploratory data analysis (EDA). Through careful preparation and examination of the data, we can draw valuable insights and make informed decisions regarding the quality of water sources.

Data preprocessing,This step is crucial in eliminating errors and inconsistencies that can affect the validity of any subsequent analysis. It also plays a significant role in standardizing data for meaningful comparisons.

Exploratory data analysis, on the other hand, empowers us to understand the inherent patterns and relationships within the water quality dataset. EDA allows us to visualize trends, identify outliers, and uncover potential correlations between various water quality parameters. These insights can inform the selection of appropriate analytical techniques and the development of predictive models.

In summary, the combination of data preprocessing and EDA provides a solid foundation for more advanced water quality analysis. By optimizing data quality and gaining a deeper understanding of the data, we can confidently make recommendations for improving water quality, mitigating environmental risks, and protecting public health. This approach ultimately leads to more effective decision-making and resource allocation in the realm of water quality management.