**WATER QUALITY ANALYSIS**

**PHASE-4**

**TEAM MEMBER: KRITHIGA .M**

**OBJECTIVE:**

* To Continue building the analysis by creating visualizations and building a predictive model.
* Use visualization libraries (e.g., Matplotlib, Seaborn) to create histograms, scatter plots, and correlation matrices.
* Build a predictive model (e.g., Logistic Regression, Random Forest) to determine water potability based on water quality parameters.

**PROBLEM DEFINITION:**

Water quality monitoring is of paramount importance to ensure the safety and sustainability of our water sources. Detecting unusual patterns or anomalies in water quality parameters is a critical aspect of this monitoring process. Anomalies can indicate contamination, equipment malfunctions, or other issues that need immediate attention. To address this challenge, a range of advanced anomaly detection techniques have been developed, each with its own unique approach and capabilities. By exploring and implementing these anomaly detection techniques, water quality stakeholders can enhance their ability to monitor, detect, and respond to anomalies in real-time, ensuring the delivery of safe and clean water to communities while protecting the environment.

**DATA VISUALIZATION:**

Data visualization refers to the graphical representation of data to help people understand the significance of data by displaying it in a visual context. It involves using various charts, graphs, and other visual elements to present data in a way that is easily understandable and can reveal patterns, trends, and insights that might be less apparent in raw, numerical data.

The goal of data visualization is to simplify complex datasets, making it easier for individuals to grasp data, identify relationships, and draw conclusions. Effective data visualization can provide valuable insights, support decision-making, and communicate information more efficiently than raw data or text-based descriptions. Popular types of data visualizations include bar charts, line graphs, pie charts, scatterplots, heatmaps, and more, each of which is chosen based on the specific characteristics of the data and the message the visualization aims to convey.

**SCATTERPLOT:**

A scatterplot is a graphical representation of data that displays individual data points on a two-dimensional Cartesian plane. It is often used to visualize the relationship or correlation between two variables. Each data point in the scatterplot is represented by a dot or marker, with one variable plotted on the horizontal axis (x-axis) and the other on the vertical axis (y-axis). Scatterplots are useful for identifying patterns, trends, or outliers in data and for assessing the strength and direction of the relationship between the two variables.

**CODE:**

import matplotlib.pyplot as plt

import pandas as pd

x\_column = ['x\_column']

y\_column = ['y\_column']

# Create a scatterplot.

plt.figure(figsize=(10, 6)) # Adjust the figure size as needed.

plt.scatter(x\_column, y\_column, c='blue', alpha=0.5, edgecolors='k')

plt.xlabel('X-Axis Label')

plt.ylabel('Y-Axis Label')

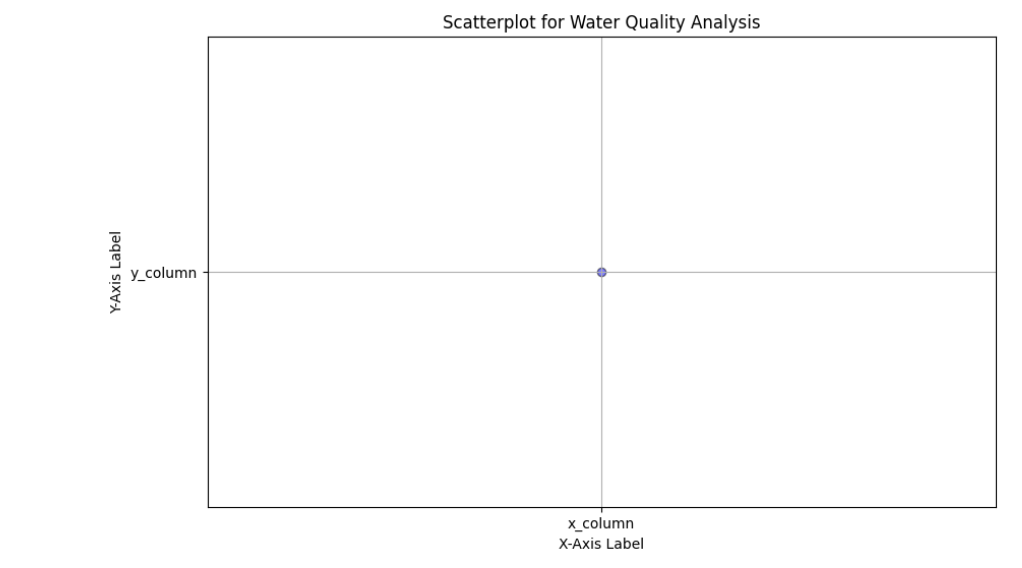
plt.title('Scatterplot for Water Quality Analysis')

# Display the scatterplot.

plt.grid(True) # Add a grid if needed.

plt.show()

**OUTPUT:**



**HISTOGRAM:**

A histogram is a graphical representation of data that displays the distribution of a dataset. It is a common tool in statistics and data analysis to visualize the frequency or count of data points falling within various ranges or "bins" of a continuous variable.

**CODE:**

sns.set(style="whitegrid")

plt.figure(figsize=(12, 8))

numerical\_columns = data.drop("Potability", axis=1).columns

for column in numerical\_columns:

plt.subplot(3, 3, numerical\_columns.get\_loc(column) + 1)

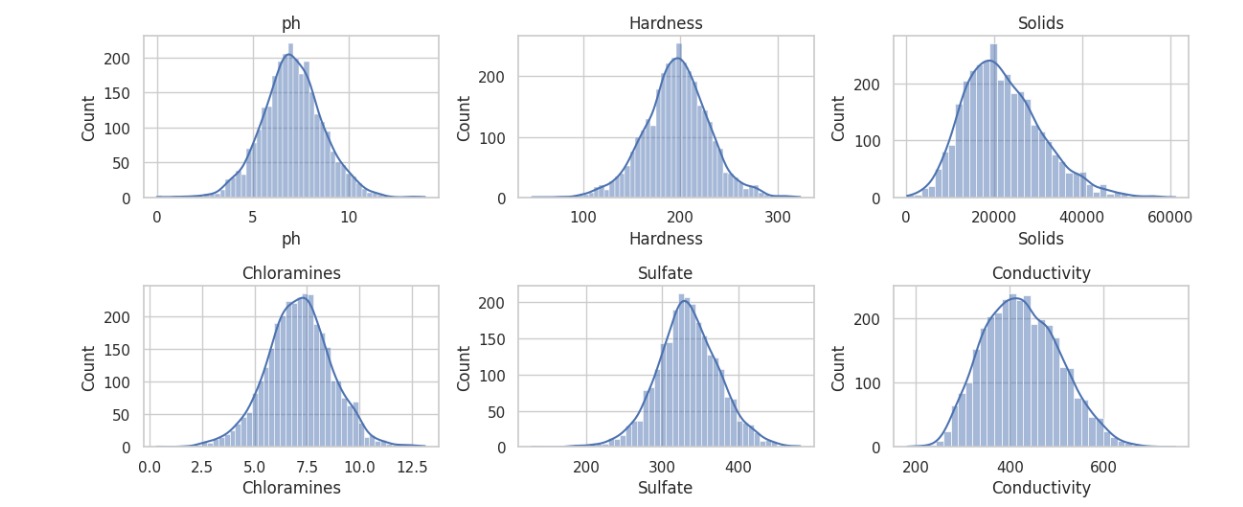
sns.histplot(data[column], kde=True)

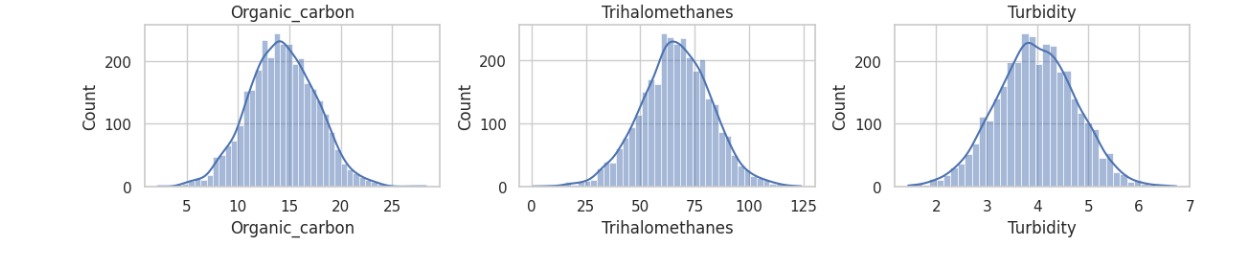
plt.title(column)

plt.tight\_layout()

plt.show()

**OUTPUT:**





**CORRELATION MATRICS:**

A correlation matrix is a table or matrix that displays the correlation coefficients between many variables. In statistics, a correlation coefficient measures the strength and direction of a linear relationship between two variables. Correlation matrices are often used in data analysis and research to understand how different variables are related to each other.

**CODE:**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

data = pd.read\_csv('/kaggle/input/water-quality-analysis/water\_potability.csv')

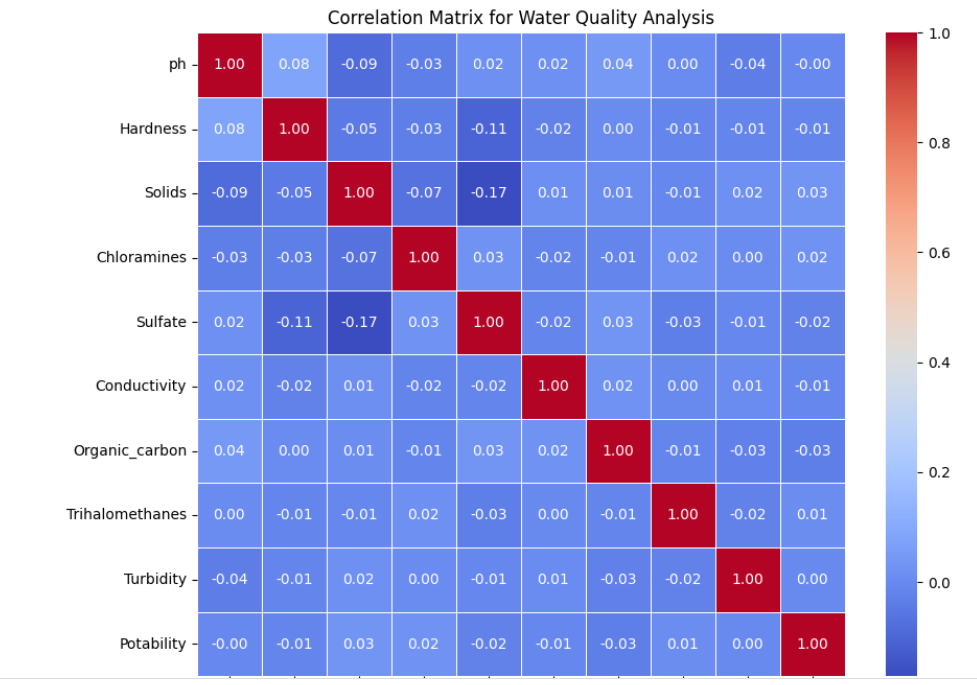
correlation\_matrix = data.corr()

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

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)

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

plt.show()



**DONUT CHART:**

A donut chart, also known as a doughnut chart, is a type of data visualization that is similar to a pie chart. It is essentially a variation of the pie chart with a hole in the center, creating a shape resembling a donut

**CODE:**

import matplotlib.pyplot as plt

labels=['ph','Hardness','Solids','Chloramines','Sulfate','Conductivity','Organic\_carbon','Trihalomethanes','Turbidity','Potability']

sizes = [30, 40, 30,50,10,30,30,20,10,40]

colors = ['lightcoral', 'lightskyblue', 'lightgreen']

fig, ax = plt.subplots()

ax.pie(sizes, labels=labels, autopct='%1.1f%%', startangle=90, pctdistance=0.85, colors=colors)

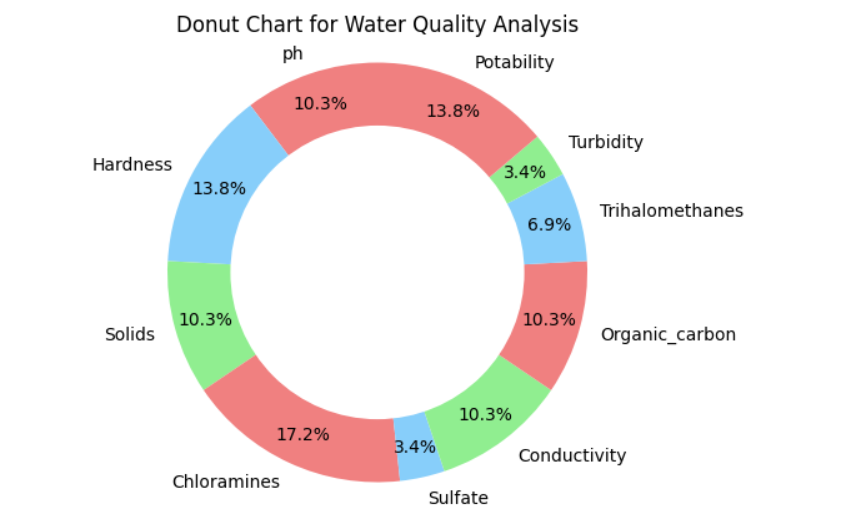
center\_circle = plt.Circle((0, 0), 0.70, fc='white')

fig.gca().add\_artist(center\_circle)

ax.axis('equal')

plt.title('Donut Chart for Water Quality Analysis')

plt.show()



**To Build a predictive model (e.g., Logistic Regression, Random Forest) to determine water potability based on water quality parameters.**

**LOGISTIC REGRESSION:**

# Import necessary libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read\_csv('/kaggle/input/water-quality-analysis/water\_potability.csv') # Replace 'water\_quality\_data.csv' with your dataset's file path

sns.pairplot(data, hue='Potability', markers=["o", "s"], diag\_kind='hist')

plt.show()

fig, axes = plt.subplots(2, 3, figsize=(15, 8))

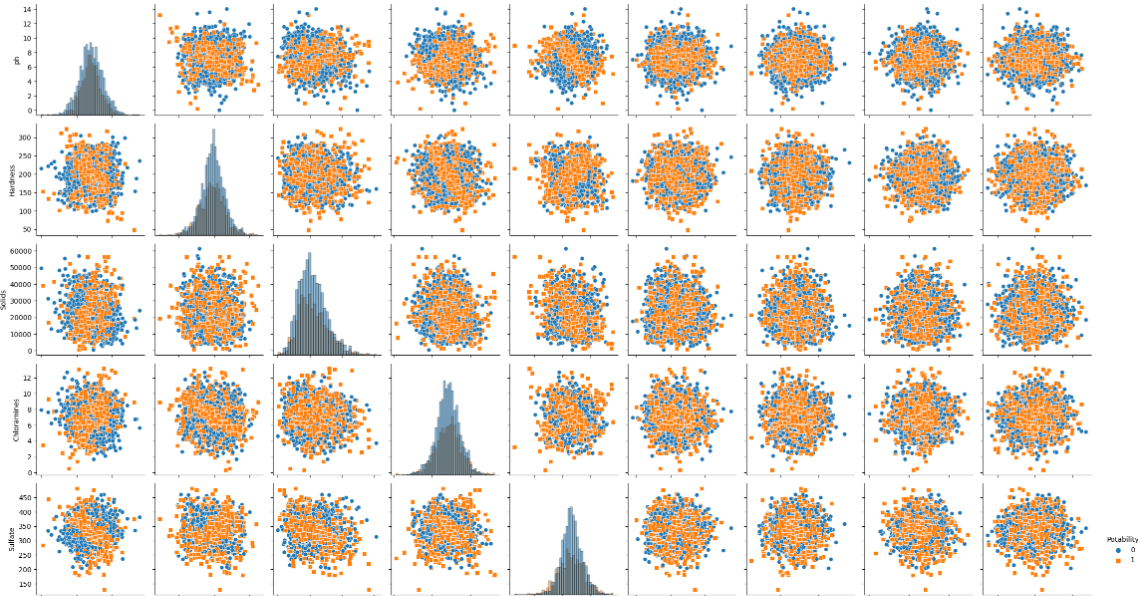
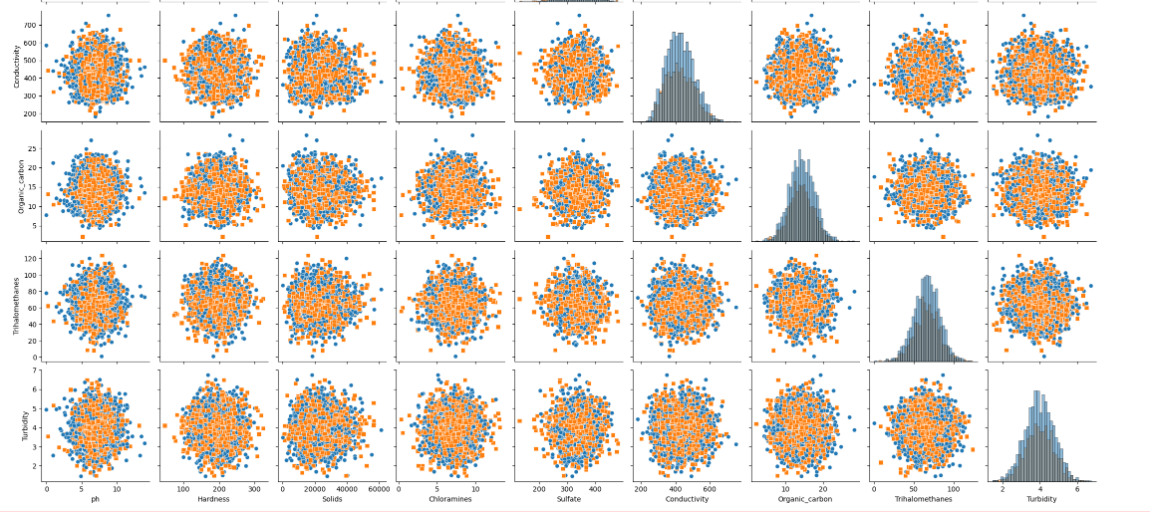
for i, col in enumerate(data.columns[:-1]):

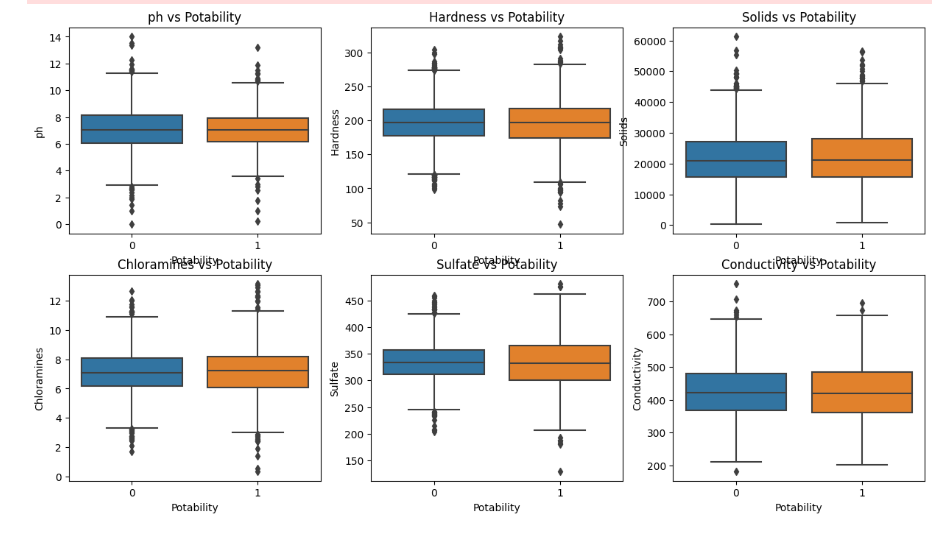
sns.boxplot(x='Potability', y=col, data=data, ax=axes[i//3, i%3])

axes[i//3, i%3].set\_title(f'{col} vs Potability')

plt.tight\_layout()

plt.show()





**RANDOM FOREST:**

**CODE:**

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read\_csv('/kaggle/input/water-quality-analysis/water\_potability.csv') # Replace 'water\_quality\_data.csv' with your dataset's file path

plt.figure(figsize=(6, 4))

sns.countplot(data['Potability'])

plt.title('Distribution of Potable and Non-Potable Water')

plt.show()

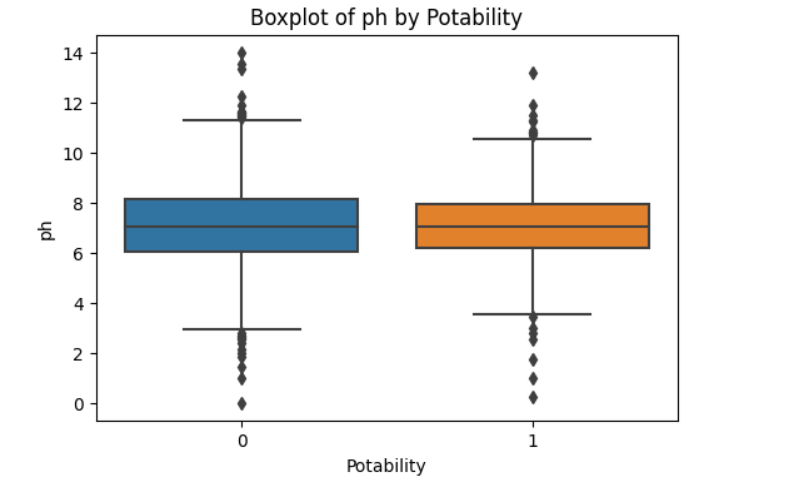
for column in data.columns[:-1]: # Excluding the target variable

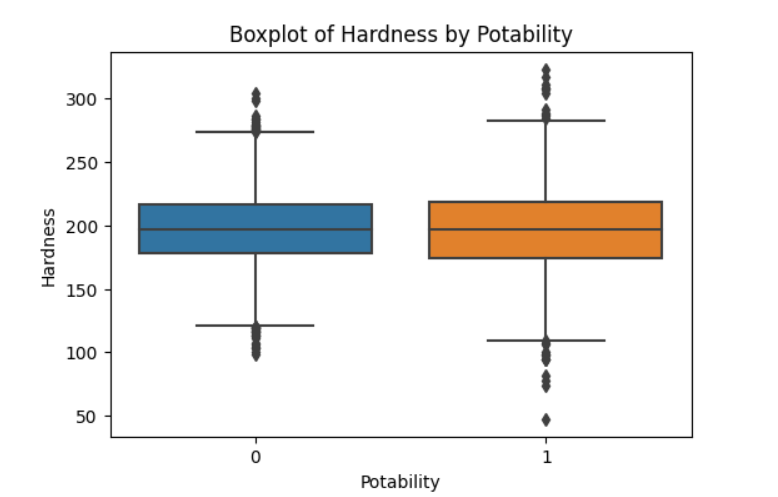
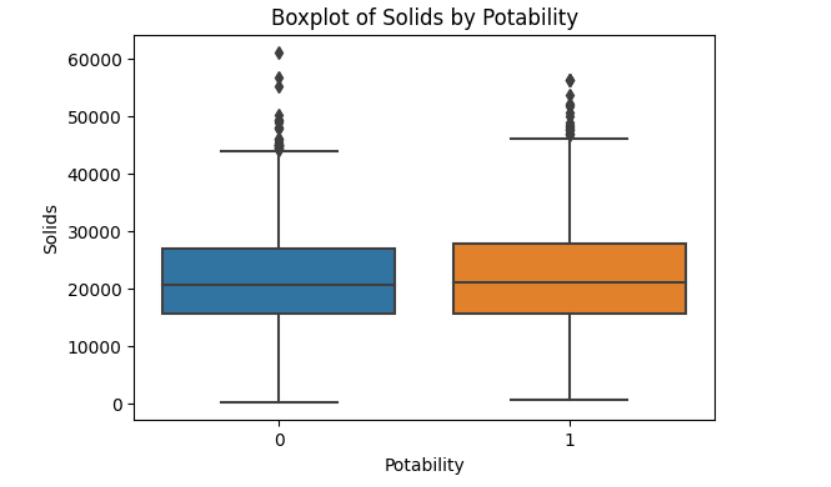
plt.figure(figsize=(6, 4))

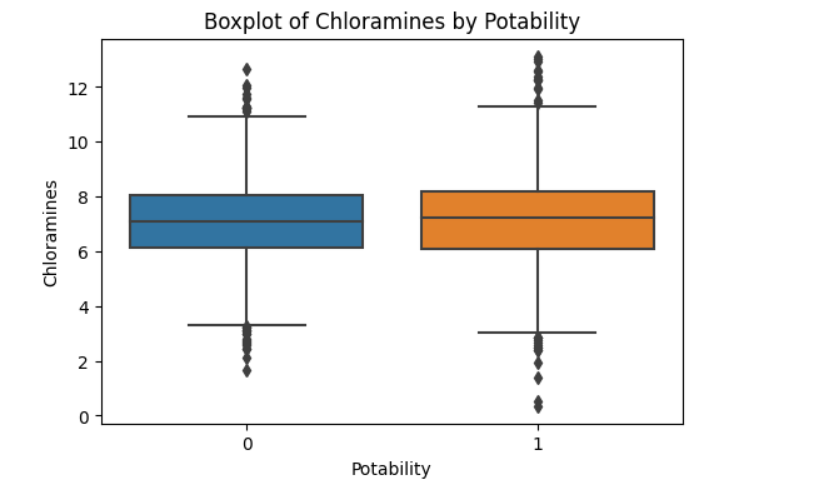
sns.boxplot(x='Potability', y=column, data=data)

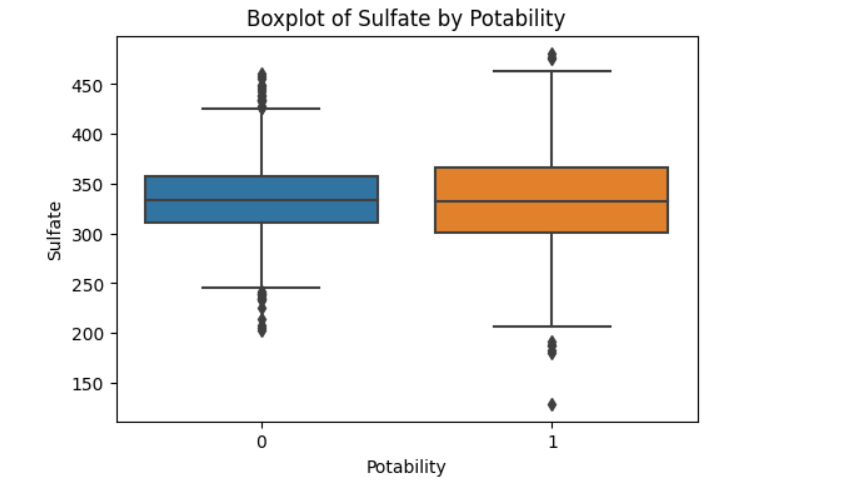
plt.title(f'Boxplot of {column} by Potability')

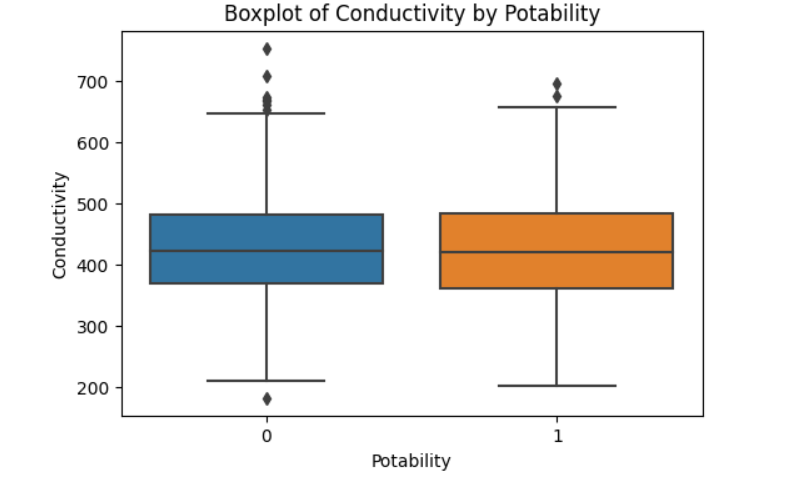
plt.show()

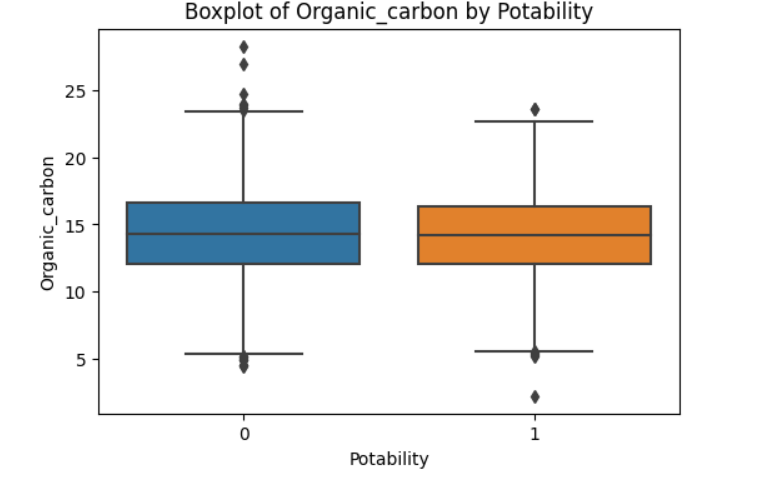


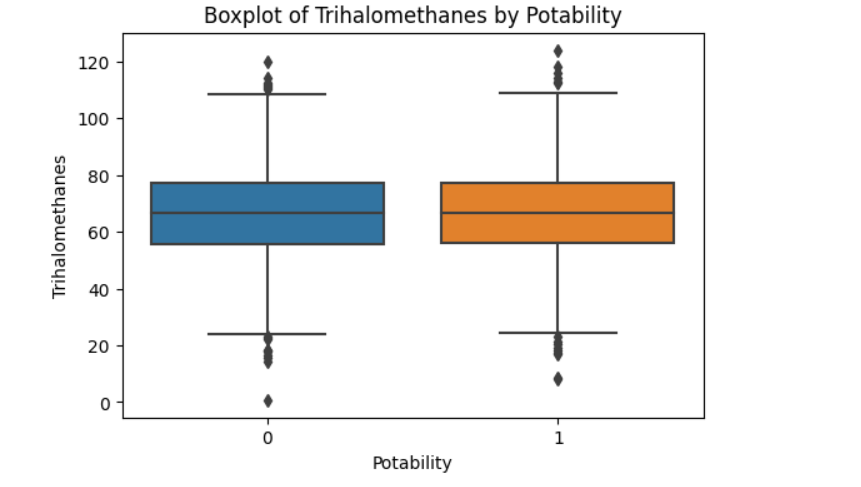


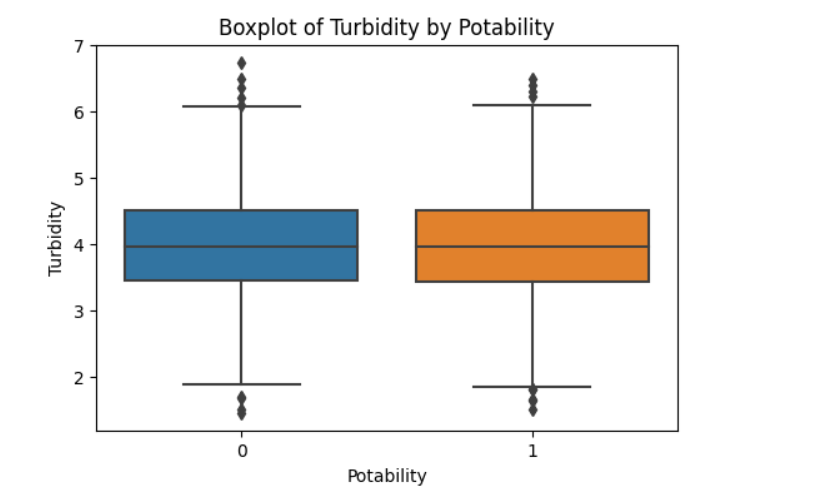












**CONCLUSION:**

We have implemented visualizations and building a predictive model and  used visualization libraries (e.g., Matplotlib, Seaborn) to create histograms, scatter plots, and correlation matrices and Build a predictive model (e.g., Logistic Regression, Random Forest) to determine water potability based on water quality parameters.