**WATER OUALITY ANALYSIS**

**PHASE-5**

**TEAM MEMBER: KRITHIGA M**

**PROJECT OBJECTIVE:**

The project involves analysing 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.

The project's primary objective is to assess water quality suitability for drinking purposes by analyzing pertinent data and deviations from regulatory standards. This involves defining analysis goals, collecting a diverse range of water quality data, developing informative visualizations, and constructing a predictive model. The analysis will focus on parameters crucial to determining water portability, such as pH levels, microbial content, and various contaminants. Visual representations, including graphs and charts, will aid in identifying trends and communicating findings effectively. Additionally, a predictive model will be created to forecast water potability based on the gathered parameters, facilitating the anticipation of deviations from regulatory standards.   
  
The ultimate aim is to ensure the safety and fitness of water for consumption by addressing potential issues and deviations, allowing for informed decision-making regarding water safety. This project's outcomes will guide interventions, water treatment strategies, and policy implementations, contributing to public health safety through the assurance of access to clean, potable drinking water.

**DESIGN THINKING PROCESS:**

Design Thinking is a problem-solving approach that can be applied to various fields, including the analysis of water quality data. To effectively address the objectives you've outlined, a Design Thinking process can be applied. Here's how you can approach each objective using the five stages of Design Thinking:

1. Empathize:

* Understand the stakeholders involved in water quality analysis, such as government agencies, environmental organizations, and the public.
* Gather information about their needs and concerns regarding water quality data analysis.
* Collaborate with experts in water quality, environmental science, and data analysis to gain a comprehensive understanding of the domain.

2. Define:

* Clearly define the problem for each objective.
* Craft a problem statement that summarizes the challenges and goals for each objective.
* Identify key performance indicators (KPIs) to measure success, such as reducing contamination levels or meeting regulatory standards.

3. Ideate:

* Generate a wide range of ideas and potential solutions for each objective.
* Encourage interdisciplinary brainstorming sessions with experts from different fields.
* Consider innovative approaches, technologies, and methods for analyzing water quality data, identifying deviations, and understanding parameter relationships.

4. Prototype:

* Develop prototypes or mock-ups of the tools, systems, or processes needed to achieve each objective.
* Create data visualization dashboards that make it easy to understand and interpret water quality data.
* Test different analytical methods and technologies to detect deviations and understand parameter relationships.

5. Test and Iterate:

* Implement your prototypes and solutions on a small scale to test their effectiveness
* Collect feedback from stakeholders, experts, and end-users.
* Iterate and refine your solutions based on the feedback received.
* Continuously monitor and adjust your approach to adapt to changing water quality data and evolving needs.

**DEVELOPMENT PHASES:**

* Research and Planning: Conducting thorough research on existing water quality assessment methods and planning the approach for data collection and analysis.
* Data Collection: Gathering relevant datasets related to water quality parameters, sources of contamination, and potability criteria.
* Data Processing and Cleaning: Preprocessing the collected data to handle missing values, outliers, and inconsistencies.
* Analysis and Modeling: Utilizing statistical and machine learning techniques to analyze the data and build predictive models for water potability assessment.
* Testing and Validation: Evaluating the performance of the developed models through rigorous testing and validation procedures.
* Refinement and Optimization: Iteratively refining the models and methodologies to enhance accuracy and efficiency in assessing water quality and potability.
* Implementation and Deployment: Implementing the final solution and deploying it in real-world scenarios for continuous monitoring and assessment of water quality.

**ANALYSIS OBJECTIVE:**

* Determine the factors affecting water potability.
* Build a predictive model to classify water as potable or non-potable based on various features.

**DATA PREPROCESSING:**

* Handle missing values, outliers, and data normalization.
* Split the data into training and testing sets.

**EXPLORATORY DATA ANALYSIS (EDA):**

* Calculate summary statistics for each feature.
* Explore correlations between different features and the target variable.
* Identify any patterns or trends in the data.

**DATA VISUALIZATION:**

* Create visualizations such as histograms, box plots, and correlation matrices.
* Visualize the distribution of various features to understand their impact on water potability.

**PREDICTIVE MODELING:**

* Choose a suitable classification algorithm (e.g., logistic regression, random forest, or support vector machine) to build the predictive model.
* Train the model on the training data and evaluate its performance on the testing data.
* Fine-tune the model parameters to improve its predictive accuracy.

**INSIGHTS:**

Water quality analysis involves a detailed examination of various parameters that collectively determine the overall health and safety of water for human and environmental use. These insights aid in the comprehensive assessment of water quality through the following key components:

1. Chemical Analysis: This involves the measurement of various chemical constituents present in water, such as dissolved solids, nutrients, heavy metals, and organic compounds. By assessing these parameters, analysts can identify the presence of pollutants from industrial, agricultural, or domestic sources, enabling the determination of the water's contamination levels and the potential risks associated with consumption or ecosystem exposure.

2. Microbiological Analysis: This aspect of water quality assessment focuses on the detection of microorganisms, including bacteria, viruses, and parasites, that can pose significant health risks to humans. By conducting tests to identify fecal coliforms, E. coli, and other pathogenic organisms, analysts can evaluate the microbiological safety of water, particularly for drinking and recreational purposes.

3. Physical Analysis: This encompasses the evaluation of various physical characteristics of water, such as temperature, turbidity, and acidity (pH). These parameters provide insights into the water's clarity, temperature variations, and overall chemical balance. For example, high turbidity levels may indicate the presence of suspended particles or sediment, affecting water transparency and potentially harboring harmful contaminants.

4. Ecological Analysis: Understanding the ecological health of water systems involves examining parameters such as dissolved oxygen levels, aquatic biodiversity, and the presence of indicator species. By assessing these factors, analysts can gauge the overall well-being of aquatic ecosystems and identify potential threats to aquatic life, such as oxygen depletion or the presence of toxic substances that can disrupt the natural balance of the ecosystem.

5. Regulatory Standards: Insights from water quality analysis are often compared against established regulatory standards and guidelines set by government agencies or international organizations. This helps in determining compliance levels and the adherence to prescribed thresholds for various contaminants, ensuring that water resources meet the necessary safety requirements for specific uses, such as drinking, irrigation, or industrial processes.

**DETERMINING THE POTABILITY:**

1. Microbiological Quality: This involves testing for the presence of harmful bacteria, viruses, and parasites that can cause waterborne diseases. Water should be free from pathogenic microorganisms, ensuring it does not pose any health risks to consumers.

2. Chemical Composition: Assessing the concentration of various chemicals and substances, such as heavy metals, pesticides, and organic compounds, is crucial in determining water safety. Potable water should adhere to regulatory limits and be free from harmful chemical contaminants that can lead to acute or chronic health issues upon consumption.

3. Physical Characteristics: Parameters such as color, odor, and turbidity are important in determining water potability. Water should appear clear, free from unusual smells or tastes, and without visible contaminants, ensuring it is aesthetically acceptable and safe for consumption.

4. Compliance with Standards: Potability assessments often involve comparing water quality data with established national or international standards, such as those set by the World Health Organization (WHO) or the Environmental Protection Agency (EPA). Adherence to these standards ensures that water is safe for human consumption and meets the necessary health and safety requirements.

**PROVIDE INSTRUCTIONS ON HOW TO REPLICATE THE ANALYSIS, GENERATE VISUALIZATIONS, AND BUILD THE PREDICTIVE MODEL USING PYTHON. INCLUDE EXAMPLE OUTPUTS OF THE VISUALIZATIONS, CORRELATION MATRICES, AND MODEL EVALUATION.**

**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 = [&#39;x\_column&#39;]

y\_column = [&#39;y\_column&#39;]

# Create a scatterplot.

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

plt.scatter(x\_column, y\_column, c=&#39;blue&#39;, alpha=0.5, edgecolors=&#39;k&#39;)

plt.xlabel(&#39;X-Axis Label&#39;)

plt.ylabel(&#39;Y-Axis Label&#39;)

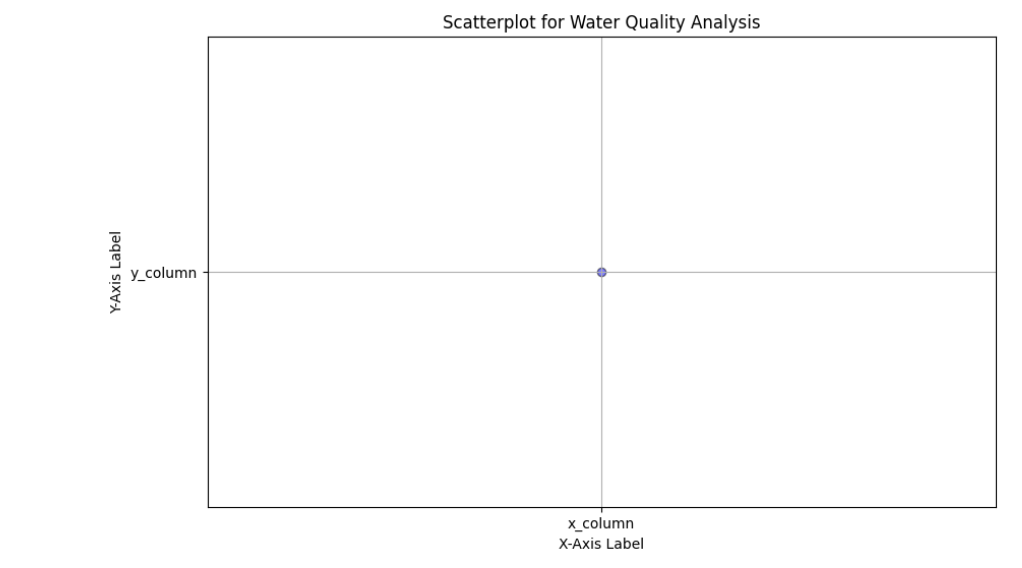
plt.title(&#39;Scatterplot for Water Quality Analysis&#39;)

# 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:

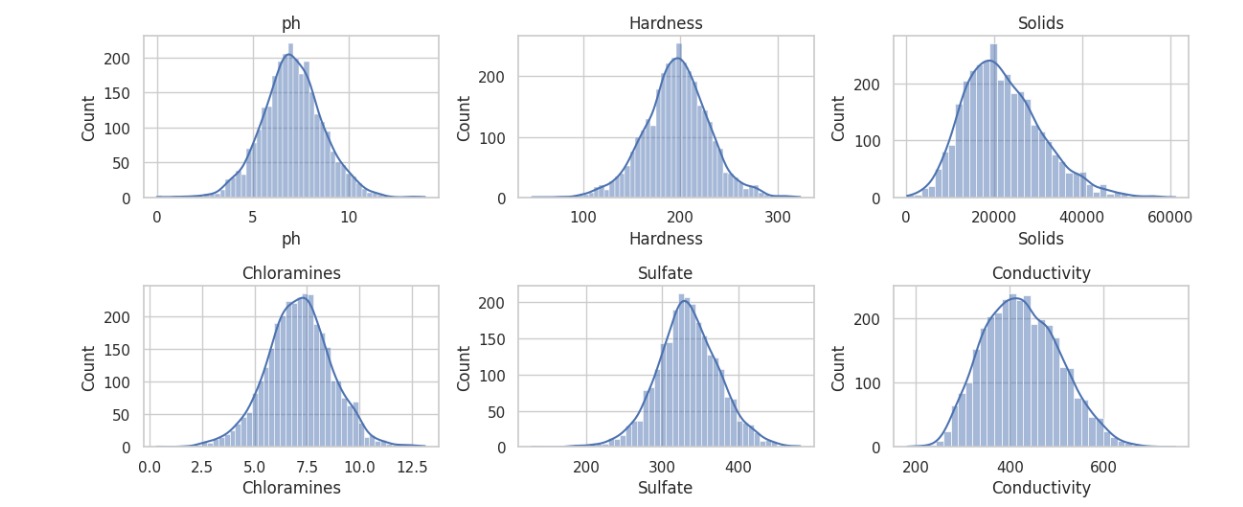
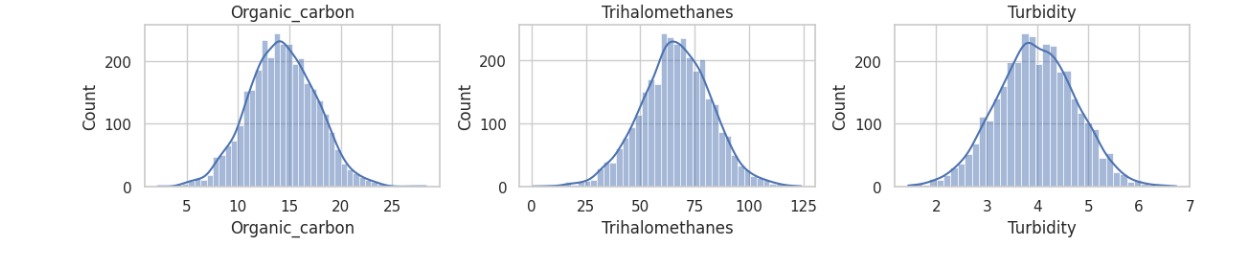
plt.subplot(3, 3, numerical\_columns.get\_loc(column) + 1)

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

plt.title(column)

plt.tight\_layout()

plt.show()



**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')

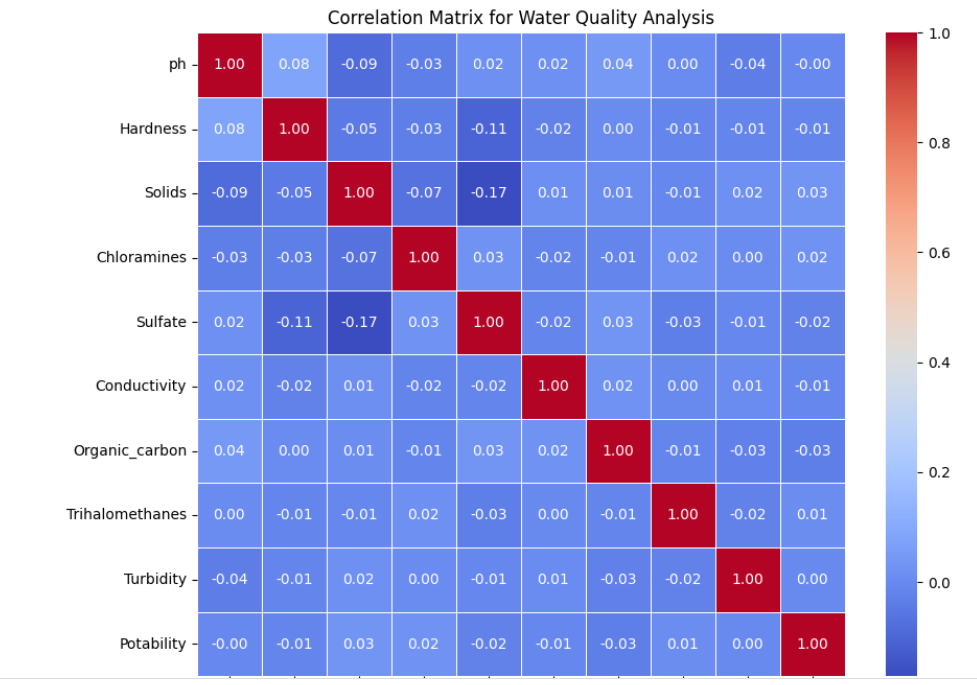
correlation\_matrix = data.corr()

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)

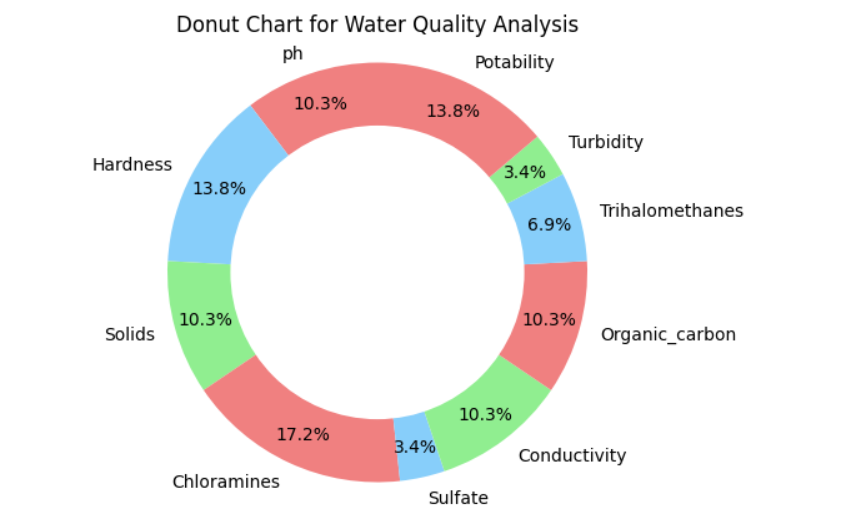
center\_circle = plt.Circle((0, 0), 0.70, fc='white')

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))

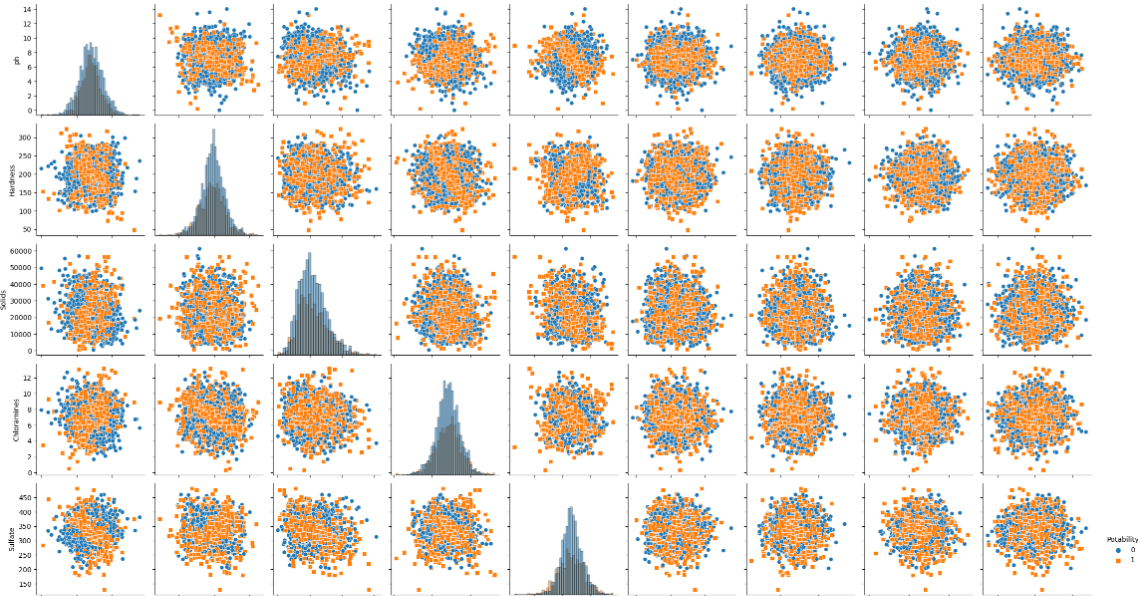
for i, col in enumerate(data.columns[:-1]):

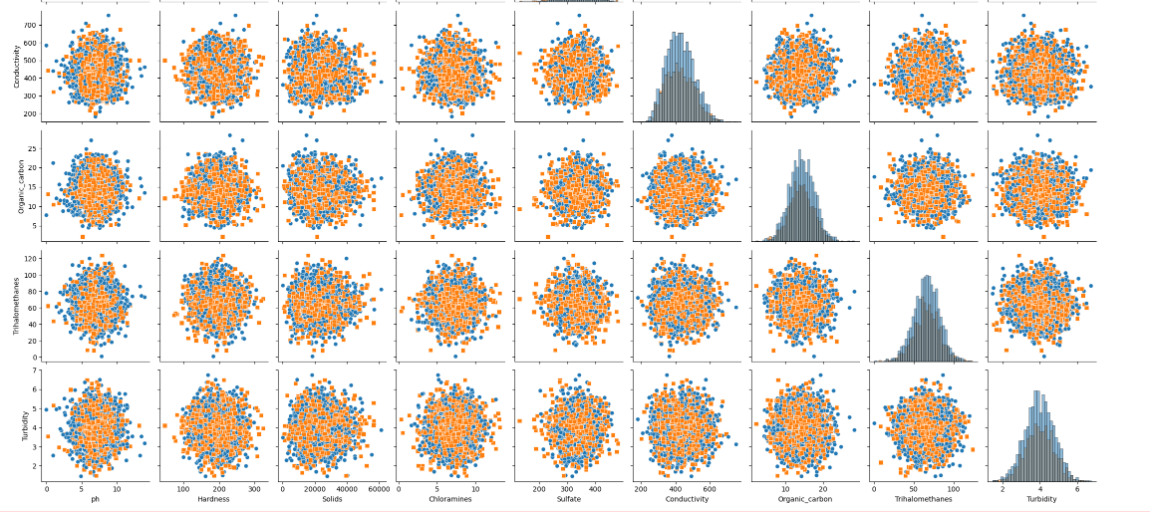
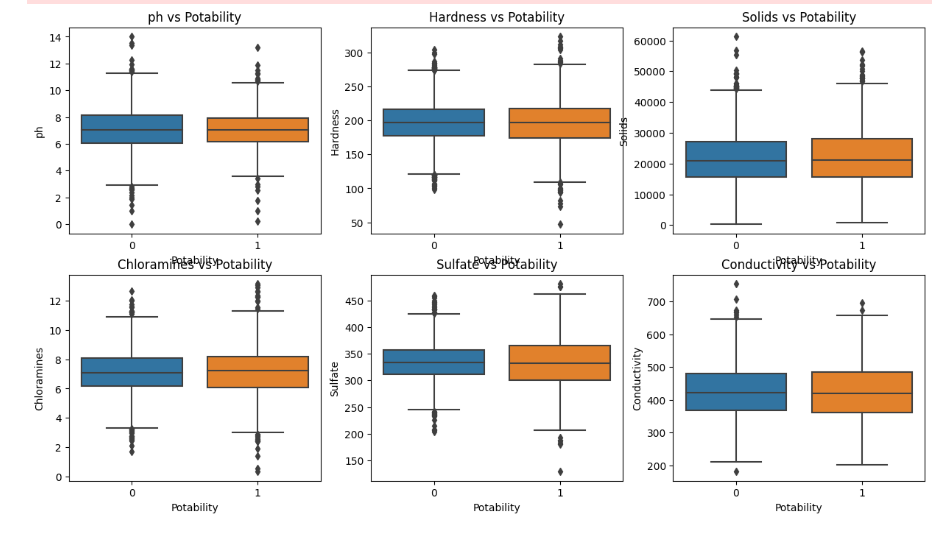
sns.boxplot(x='Potability', y=col, data=data, ax=axes[i//3, i%3])

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()

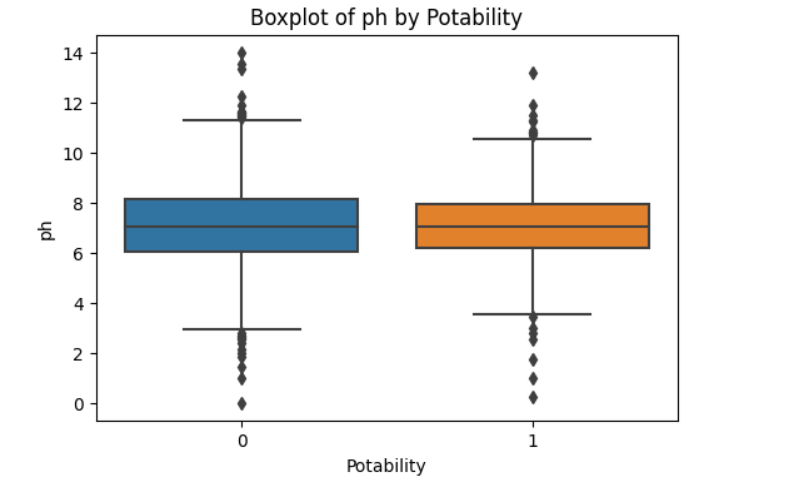
for column in data.columns[:-1]: # Excluding the target variable

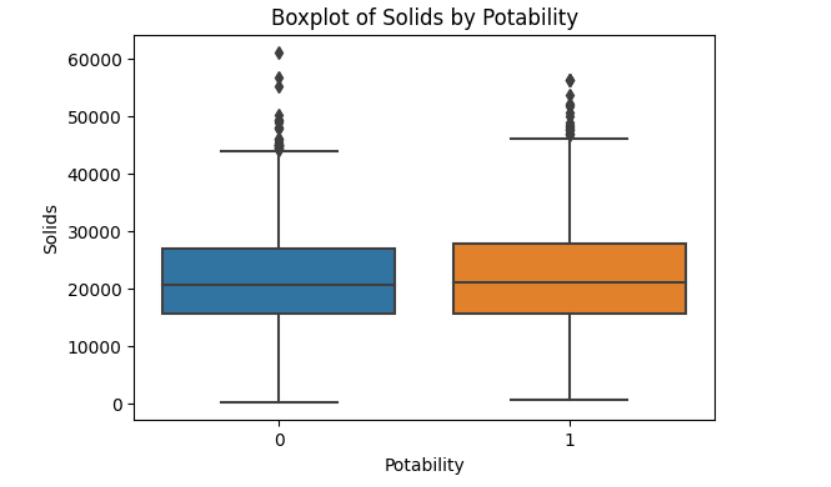
plt.figure(figsize=(6, 4))

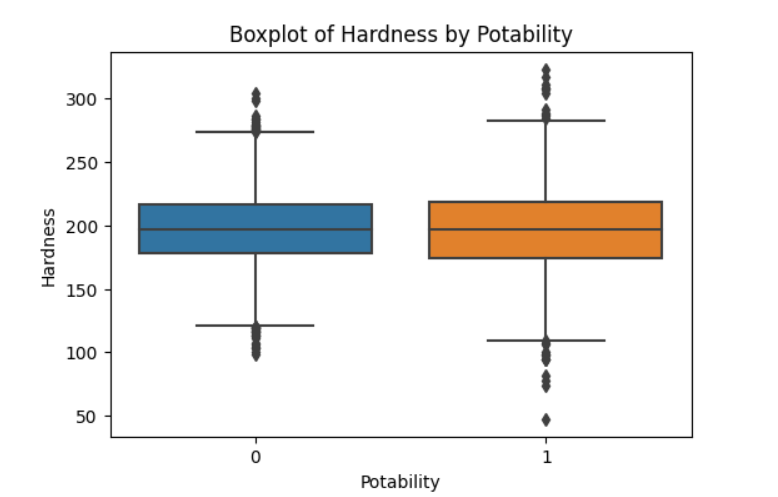
sns.boxplot(x='Potability', y=column, data=data)

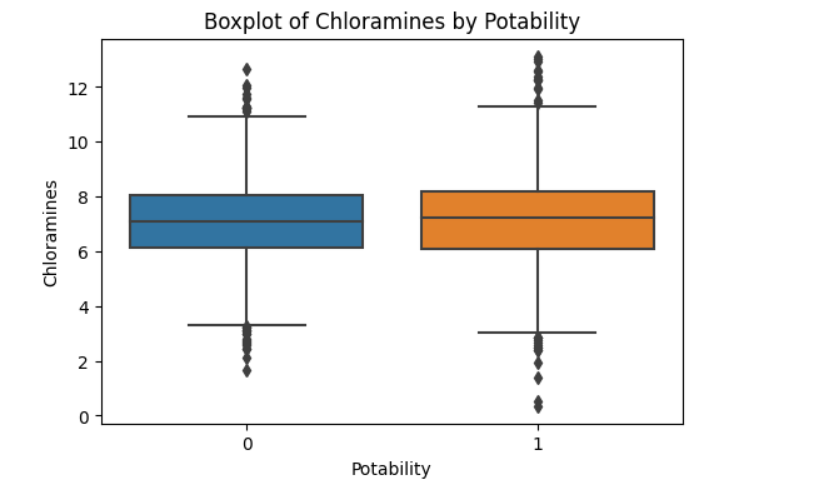
plt.title(f'Boxplot of {column} by Potability')

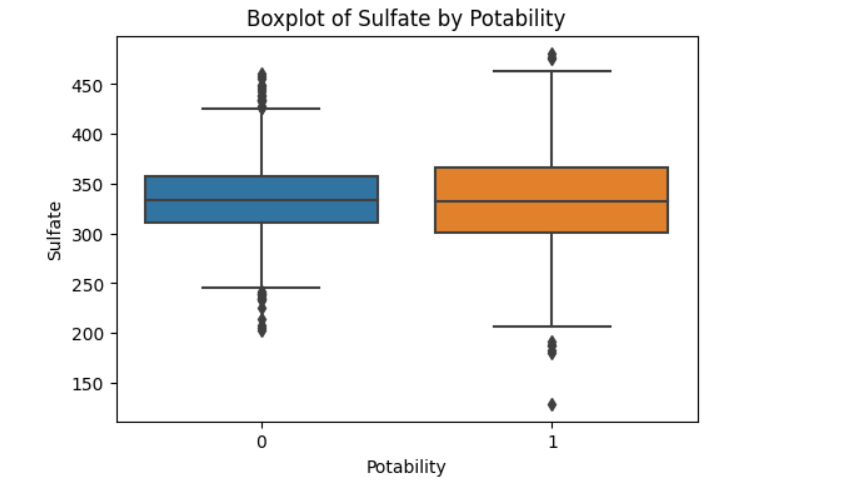
plt.show()

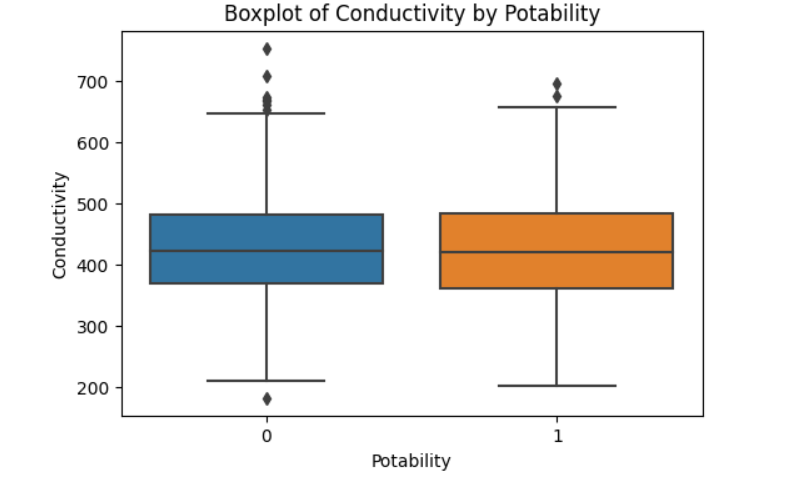


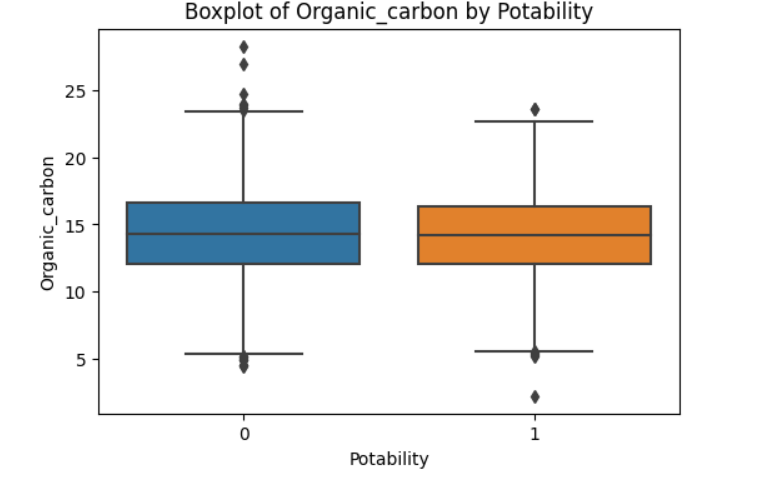


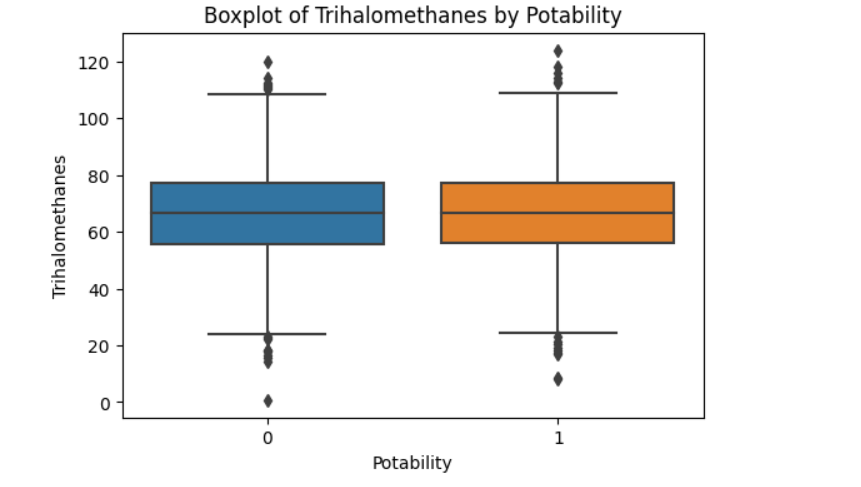


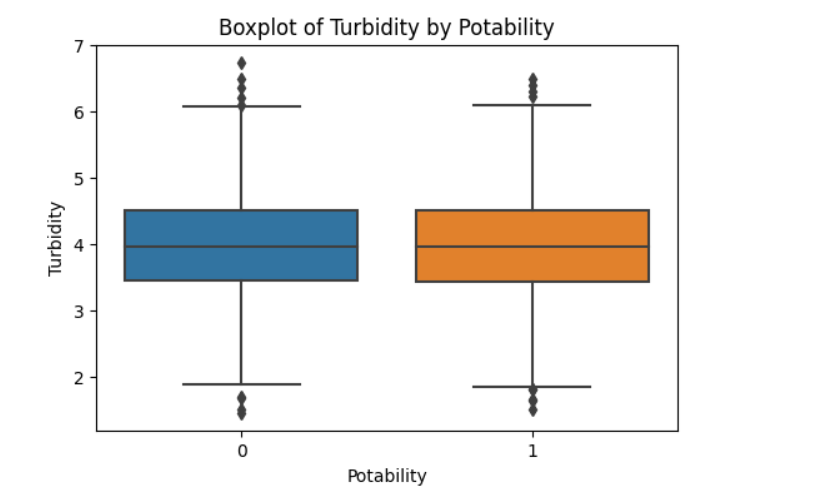












**CONCLUSION:**

Using Python, we can efficiently analyze water quality data, visualize key parameters, and develop a predictive model to assess water safety. These processes aid in making informed decisions about the potability of water, ensuring it meets regulatory standards and poses no health risks. By leveraging various Python libraries and tools, we can effectively handle data preprocessing, conduct in-depth analysis, and build robust models for sustainable water resource management. This approach facilitates the proactive protection of public health and the environment, supporting the continuous availability of clean and safe water for communities worldwide.