Project 10: Water Quality Analysis (DAC_Phase5)

PHASE 5 : Project Documentation & Submission

Title: Water Quality Analysis and Potability Prediction Project

Abstract:

This project aims to assess and determine the potability of water using a given dataset through a comprehensive data analysis approach. We will follow the design thinking process and go through several development phases, including data preprocessing, exploratory data analysis (EDA), data visualization, and predictive modelling. The insights derived from this analysis will be instrumental in evaluating water quality and ensuring its safety for consumption.

1. Introduction

1.1 Background

- ❖ Clean and safe drinking water is a fundamental necessity for human survival.
- ❖ Access to potable water is critical for health and well-being, and the quality of water can vary significantly depending on various factors.
- ❖ Water quality analysis is essential to ensure that water is safe for consumption, free from contaminants, and meets established standards.
- ❖ In this project, we will use a dataset to perform a comprehensive analysis of water quality and predict its potability.

1.2 Objectives

- ❖ Analyze water quality data to determine its potability.
- ❖ Identify key factors affecting water quality.
- ❖ Develop a predictive model for potability assessment.
- ❖ Provide valuable insights to help ensure safe drinking water.

2. Design Thinking Process

2.1 Empathize

- ❖ Understand the problem and the stakeholders involved.
- ❖ Identify the concerns and expectations related to water quality and potability.

2.2 Define

- ❖ Clearly define the problem statement and project objectives.
- ❖ Set measurable goals for potability prediction and water quality analysis.

2.3 Ideate

- ❖ Brainstorm potential data sources, analytical techniques, and predictive models to achieve the defined objectives.

2.4 Prototype

- ❖ Develop a prototype data analysis and modelling framework.

2.5 Test

- ❖ Test the prototype on a small scale to identify any issues and refine the process.

3. Development Phases

3.1 Data Collection

- ❖ Acquire a comprehensive dataset containing water quality parameters, including physical, chemical, and biological properties.
- ❖ The dataset may include information on various water sources.

  DATASET LINK : https://www.kaggle.com/datasets/adityakadiwal/water-potability

3.2 Data Preprocessing

- ❖ Clean and prepare the data for analysis.
- ❖ Handle missing values, outliers, and inconsistencies.
- ❖ Normalize or scale the data as needed.

```
In [11]: data['Potability']=data['Potability'].astype('category')
```

```
In [12]: data.describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='PuBu')
```

Out[12]:

	count	mean	std	min	25%	50%	75%	max
ph	2785.000000	7.080795	1.594320	0.000000	6.093092	7.036752	8.062066	14.000000
Hardness	3276.000000	196.369496	32.879761	47.432000	176.850538	196.967627	216.667456	323.124000
Solids	3276.000000	22014.092526	8768.570828	320.942611	15666.690300	20927.833605	27332.762125	61227.196010
Chloramines	3276.000000	7.122277	1.583085	0.352000	6.127421	7.130299	8.114887	13.127000
Sulfate	2495.000000	333.775777	41.416840	129.000000	307.699498	333.073546	359.950170	481.030642
Conductivity	3276.000000	426.205111	80.824064	181.483754	365.734414	421.884968	481.792305	753.342620
Organic_carbon	3276.000000	14.284970	3.308162	2.200000	12.065801	14.218338	16.557652	28.300000
Trihalomethanes	3114.000000	66.396293	16.175008	0.738000	55.844536	66.622485	77.337473	124.000000
Turbidity	3276.000000	3.966786	0.780382	1.450000	3.439711	3.955028	4.500320	6.739000

```
In [13]: data[data['Potability']==1].describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='PuBu')
```

Out[13]:

	count	mean	std	min	25%	50%	75%	max
ph	1101.000000	7.073783	1.448048	0.227499	6.179312	7.036752	7.933068	13.175402
Hardness	1278.000000	195.800744	35.547041	47.432000	174.330531	196.632907	218.003420	323.124000
Solids	1278.000000	22383.991018	9101.010208	728.750830	15668.985038	21199.386615	27973.236447	56488.672410
Chloramines	1278.000000	7.169338	1.702988	0.352000	6.094134	7.215163	8.199261	13.127000
Sulfate	985.000000	332.566990	47.692818	129.000000	300.763772	331.838167	365.941346	481.030642
Conductivity	1278.000000	425.383800	82.048446	201.619737	360.939023	420.712729	484.155911	695.369528
Organic_carbon	1278.000000	14.160893	3.263907	2.200000	12.033897	14.162809	16.356245	23.604298
Trihalomethanes	1223.000000	66.539684	16.327419	8.175876	56.014249	66.678214	77.380975	124.000000
Turbidity	1278.000000	3.968328	0.780842	1.492207	3.430909	3.958576	4.509569	6.494249

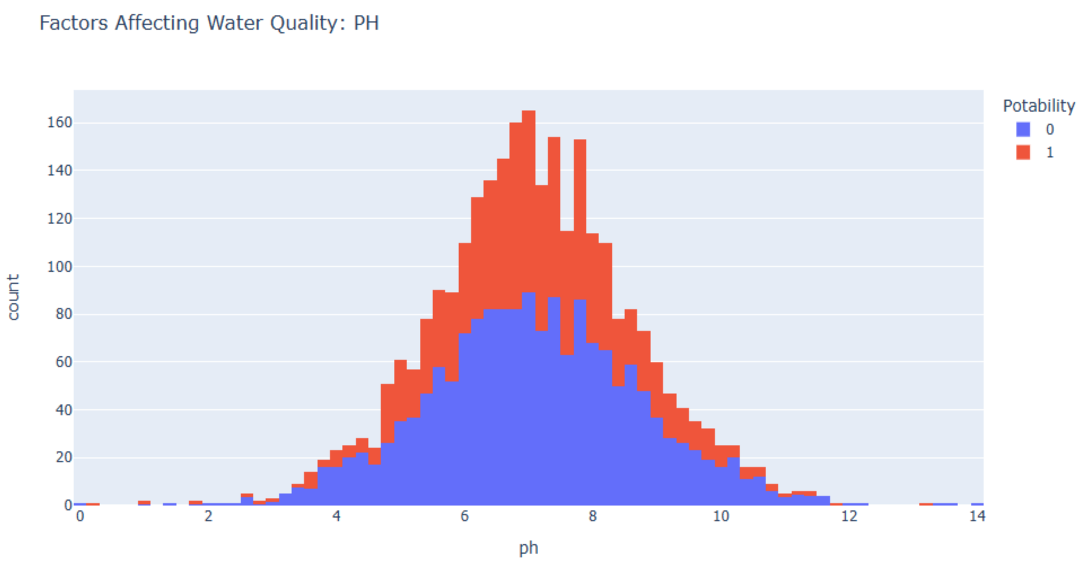
3.3 Exploratory Data Analysis (EDA)

- Conduct summary statistics and visualizations to understand the dataset's characteristics.
- Identify correlations between different water quality parameters.
- Explore potential factors influencing potability.

3.4 Data Visualization

- Utilize various data visualization techniques, such as scatter plots, histograms, box plots, and heatmaps, to visually represent the data.
- Visualizations will help in understanding the distribution of water quality parameters and identifying patterns.

```
In [2]: import plotly.express as px
        import pandas as pd
        data = pd.read_csv(r"C:\Users\divya\OneDrive\Documents\Untitled Folder\water_potability.csv")
        figure = px.histogram(data, x = "ph",
                             color = "Potability",
                             title= "Factors Affecting Water Quality: PH")
        figure.show()
```



3.5 Predictive Modelling for Potability

- ❖ Split the dataset into training and testing sets.
- ❖ Choose appropriate machine learning algorithms (e.g., logistic regression, decision trees, random forests, or neural networks) for potability prediction.
- ❖ Train and evaluate the models using appropriate performance metrics.
- ❖ Fine-tune the models to achieve the best predictive accuracy.

4. Analysis Objectives

4.1 Water Quality Assessment

- ❖ Determine the overall quality of water by analysing various parameters, including pH, hardness, turbidity, etc.

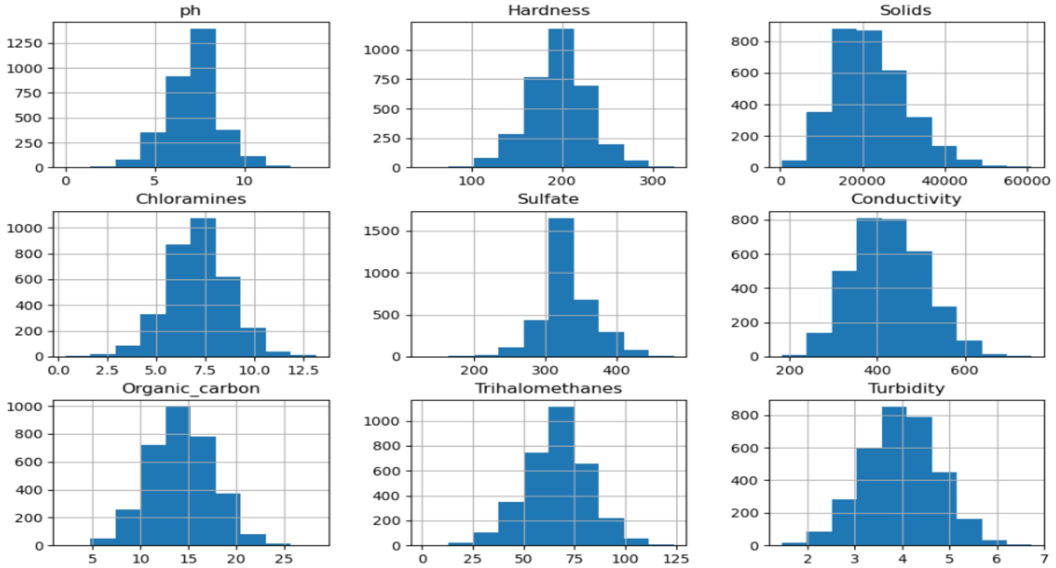
4.2 Potability Prediction

- ❖ Develop a predictive model to assess the potability of water based on the analysed parameters.
- ❖ This model will classify water as potable or non-potable.

4.3 Factors Influencing Potability

- ❖ Identify the most influential factors affecting water potability, providing actionable insights for water quality improvement.

```
In [27]: data.drop('Potability', axis=1).hist(figsize=(12,8));
```



5. Data Preprocessing

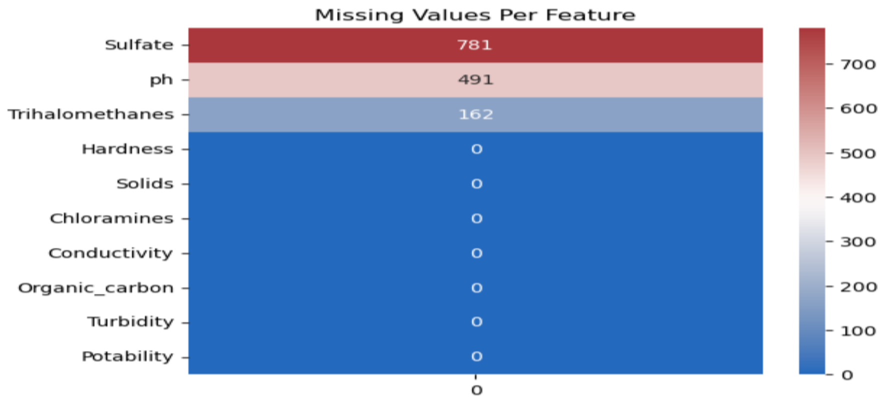
Data preprocessing is a crucial phase to ensure the reliability and accuracy of the analysis. This phase includes the following steps:

5.1 Handling Missing Values

- ❖ Identify and address missing data points by either imputing values or removing incomplete records.

```
In [23]: import matplotlib.pyplot as plt
         import seaborn as sns
         plt.title('Missing Values Per Feature')
         nans = data.isna().sum().sort_values(ascending=False).to_frame()
         sns.heatmap(nans,annot=True,fmt='d',cmap='vlag')

Out[23]: <Axes: title={'center': 'Missing Values Per Feature'}>
```



5.2 Outlier Detection and Treatment

- Detect and handle outliers that may skew the analysis.
- Determine whether to remove or transform outlier data points.

5.3 Data Normalization

- Normalize the data if necessary to bring all variables to the same scale for accurate modelling.

6.Exploratory Data Analysis (EDA)

EDA is a critical step to understand the dataset and uncover insights. The following EDA tasks will be performed:

6.1 Summary Statistics

- Calculate descriptive statistics for all water quality parameters, including mean, median, standard deviation, and percentiles.

6.2 Data Distribution

- Create histograms, density plots, and box plots to visualize the distribution of each parameter.

6.3 Correlation Analysis

- Explore the relationships between different parameters by calculating correlation coefficients and creating correlation matrices.

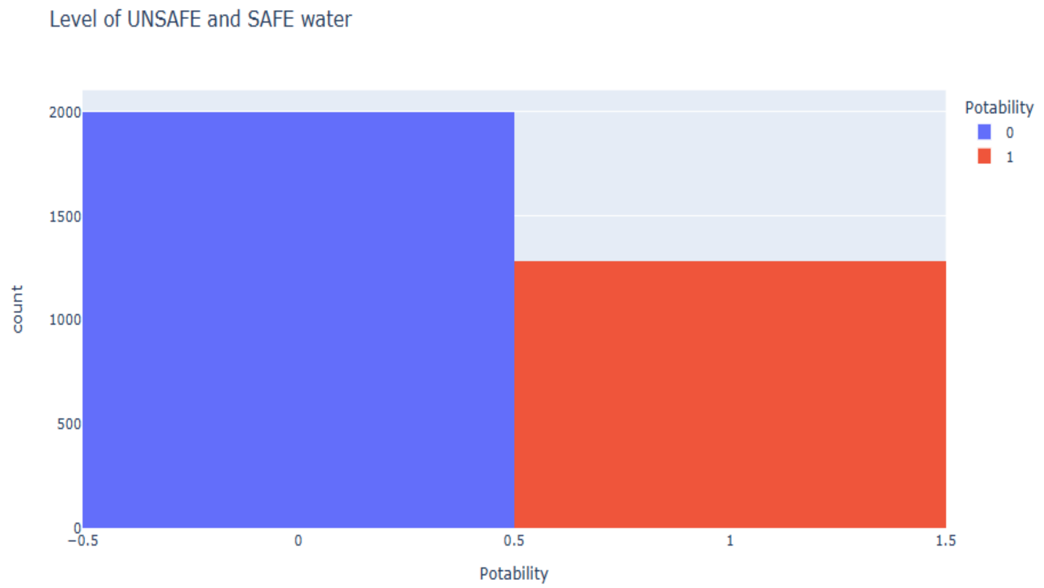
```
In [2]: correlation = data.corr()
        correlation["ph"].sort_values(ascending=False)

Out[2]: ph                 1.000000
        Hardness           0.082096
        Organic_carbon     0.043503
        Conductivity       0.018614
        Sulfate            0.018203
        Trihalomethanes    0.003354
        Potability        -0.003556
        Chloramines       -0.034350
        Turbidity         -0.039057
        Solids            -0.089288
        Name: ph, dtype: float64
```

6.4 Potability Analysis

- Analyze the distribution of potable and non-potable water samples and identify differences in key parameters.

```
In [4]: import plotly.express as px
        import pandas as pd
        data = pd.read_csv(r"C:\Users\divya\OneDrive\Documents\Untitled Folder\water_potability.csv")
        figure = px.histogram(data, x = "Potability",
                              color = "Potability",
                              title= "Level of UNSAFE and SAFE water")
        figure.show()
```



7. Data Visualization

Data visualization is essential for conveying information and patterns in the data. The following visualizations will be used:

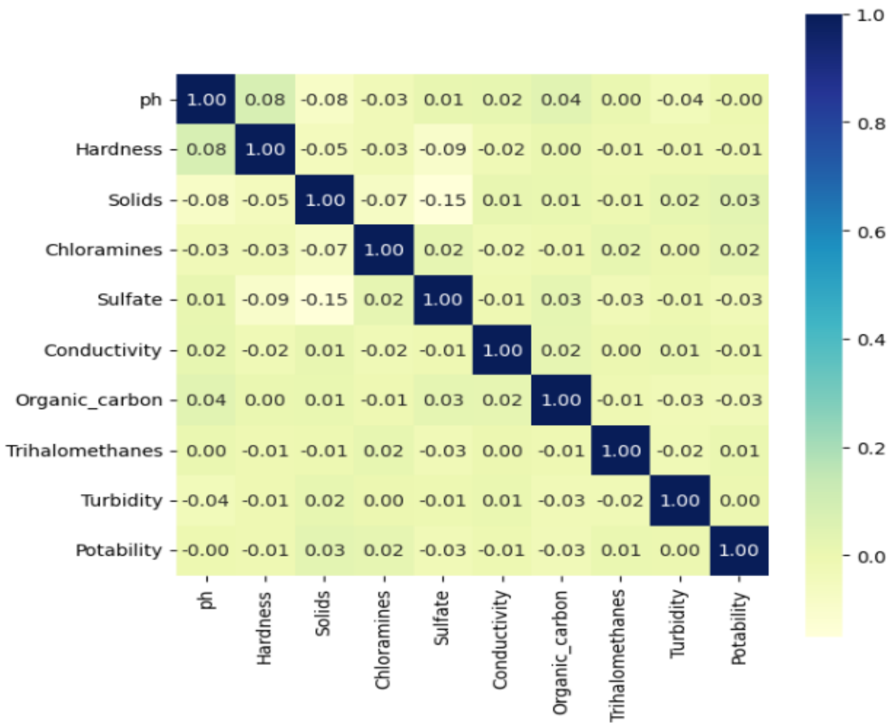
7.1 Scatter Plots

- Visualize the relationships between two continuous variables to identify patterns and trends.

7.2 Heatmaps

❖ Create heatmaps to visualize the correlation between water quality parameters and potability.

```
In [26]: import seaborn as sns
         Corrmat = data.corr()
         plt.subplots(figsize=(7,7))
         sns.heatmap(Corrmat, cmap="YlGnBu", square = True, annot=True, fmt='.2f')
         plt.show()
```



7.3 Box Plots

❖ Use box plots to compare the distribution of various parameters for potable and non-potable water samples.

7.4 Time Series Plots (if applicable)

❖ If the dataset includes time-related data, create time series plots to visualize trends over time.

DATA ANALYTICS WITH IBM COGNOS

I. IBM Cognos Introduction

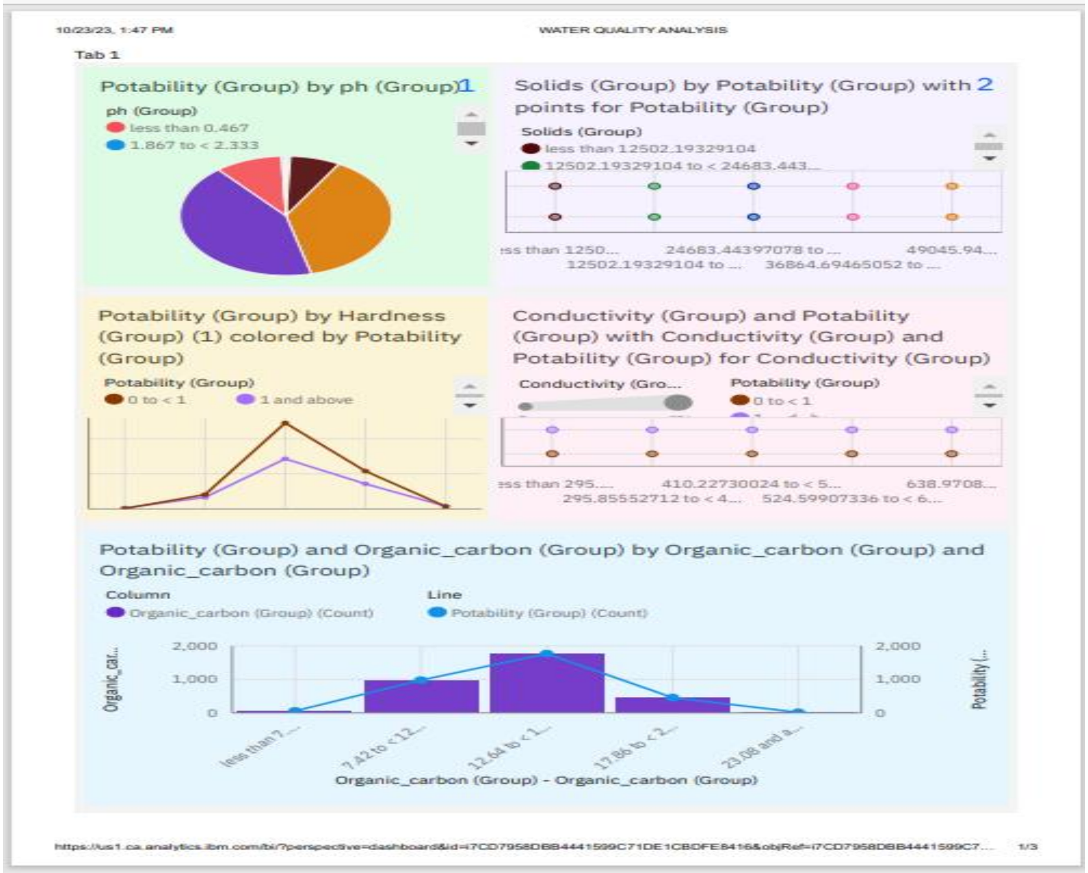
Introduce IBM Cognos as a tool for data analytics.

II. Data Exploration

Showcase how IBM Cognos aids in exploring and understanding the dataset.

III. Visualization

Demonstrate the creation of visualizations in IBM Cognos.



DATA VISUALIZATION WITH JUPYTER NOTEBOOK

i. Jupyter Notebook Introduction

Present Jupyter Notebook as a tool for data analysis and visualization.

ii. Visualizing Water Quality Parameters

Use Jupyter Notebook to create visualizations of water quality parameters.

iii. Geographic Mapping

Visualize water quality by location using Jupyter Notebook.

iv. Time Series Analysis

Analyze temporal changes in water quality using Jupyter Notebook.

8. Predictive Modelling for Potability

The predictive modelling phase aims to develop a model that can classify water samples as potable or non-potable based on the analysed parameters. This phase includes the following steps:

8.1 Data Splitting

- Divide the dataset into a training set and a testing set for model training and evaluation.

8.2 Model Selection

- Choose appropriate machine learning algorithms for binary classification.
- Evaluate multiple models to select the most suitable one.

8.3 Model Training

- Train the selected model on the training data using the water quality parameters as features and the potability as the target variable.

8.4 Model Evaluation

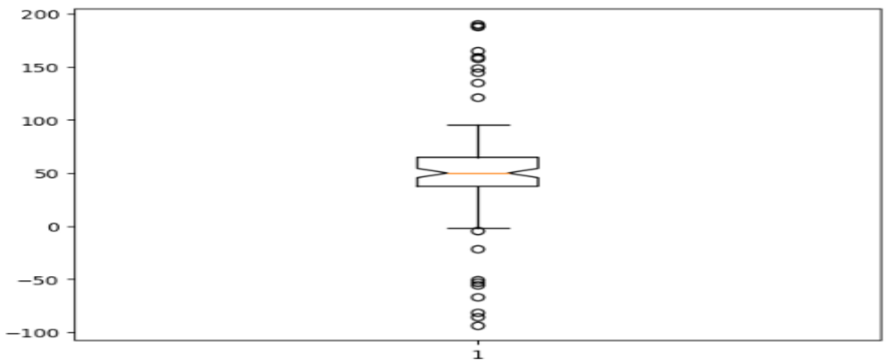
- Assess the model's performance on the testing set using relevant metrics such as accuracy, precision, recall, F1-score, and the ROC curve.

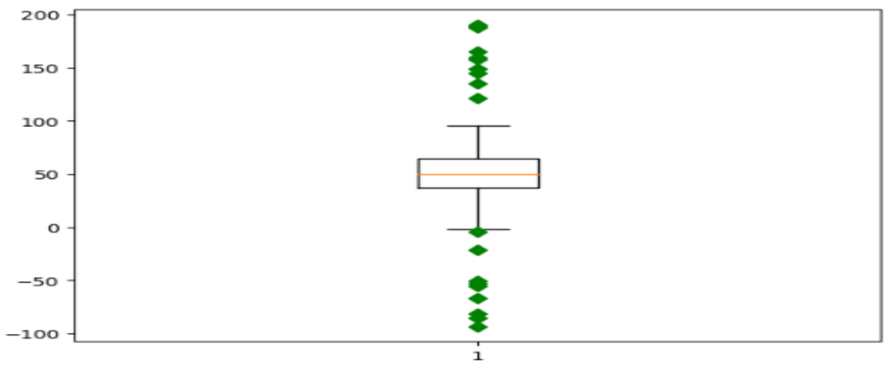
8.5 Model Optimization

- Fine-tune the model parameters, perform feature selection, and optimize hyperparameters to improve model accuracy.

```
In [21]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
spread = np.random.rand(50) * 100
center = np.ones(25) * 50
flier_high = np.random.rand(10) * 100 + 100
flier_low = np.random.rand(10) * -100
data = np.concatenate((spread, center, flier_high, flier_low), 0)
print (data)
plt.figure(figsize = (7, 5))
plt.boxplot(data, 1)
plt.show()
plt.figure(figsize = (7, 5))
plt.boxplot(data, 0, 'gD')
plt.show()
plt.figure(figsize = (7, 5))
plt.boxplot(data, 0, 'rs', 0, 0.75)
plt.show()
```

```
[ 76.00992876  65.91711212  68.18394798  50.13641895  91.8789882
  49.30853186  36.71959485  53.64321859  54.68554503  46.12751574
   1.98824902  94.19763425  70.15024029  39.48739256  40.27811781
  63.7901335    1.51570733  35.8664904   66.20022493  95.40368517
  33.23315008  26.50091155  58.3549899   55.29389813  83.63424892
  85.93968355   1.41949078  50.28720052  17.89504118  51.17823429
  38.13887785  31.33908749  63.80073629   4.93058491  15.25825846
  54.78652661  49.32384777  69.18585901  56.66409472  42.26304259
  57.08504526  75.30971113  25.95235052  51.40094291  41.79676876
  26.8334635    4.42275024  50.51908669  91.16687873  80.61094611
  50.          50.          50.          50.          50.
  50.          50.          50.          50.          50.
  50.          50.          50.          50.          50.
  50.          50.          50.          50.          50.
  50.          50.          50.          50.          50.
 148.63433008 190.45391435 188.31602969 187.62505984 121.7429514
 135.28761195 145.15853198 159.7340601  165.47596769 158.17283819
 -93.26456687 -80.91345169 -50.20532417 -66.38054113 -52.98736074
  -3.91547163 -21.35458017 -85.58189656 -55.23019443  -1.70216355]
```





9. Insights from Analysis

The insights derived from the analysis will provide valuable information for assessing water quality and ensuring potability. Here are some of the key insights that can be obtained:

9.1 Identification of Critical Parameters

- Determine which water quality parameters have the most significant impact on potability.

9.2 Potability Prediction

- Develop a predictive model that can accurately classify water samples as potable or non-potable.

9.3 Pattern Recognition

- Discover patterns and trends in the data that may indicate specific factors affecting water quality.

9.4 Data-Driven Recommendations

- Provide data-driven recommendations for improving water quality based on the identified factors.

9.5 Decision Support

- Offer decision support tools for stakeholders, such as water treatment plants or regulatory authorities, to make informed decisions about water quality management.

```
In [22]:
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.ensemble import RandomForestRegressor
data = pd.read_csv(r"C:\Users\divya\OneDrive\Documents\Untitled Folder\water_potability.csv")
print(data)
```

```
            ph    Hardness        Solids  Chloramines     Sulfate  \
0          NaN  204.890456  20791.31898     7.300212  368.516441
1     3.716080  129.422921  18630.05786     6.635246         NaN
2     8.099124  224.236259  19909.54173     9.275884         NaN
3     8.316766  214.373394  22018.41744     8.059332  356.886136
4     9.092223  181.101509  17978.98634     6.546600  310.135738
...        ...         ...          ...          ...         ...
3271  4.668102  193.681736  47580.99160     7.166639  359.948574
3272  7.808856  193.553212  17329.80216     8.061362         NaN
3273  9.419510  175.762646  33155.57822     7.350233         NaN
3274  5.126763  230.603758  11983.86938     6.303357         NaN
3275  7.874671  195.102299  17404.17706     7.509306         NaN

      Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
0       564.308654       10.379783        86.990970   2.963135           0
1       592.885359       15.180013        56.329076   4.500656           0
2       418.606213       16.868637        66.420093   3.055934           0
3       363.266516       18.436525       100.341674   4.628771           0
4       398.410813       11.558279        31.997993   4.075075           0
...            ...             ...              ...        ...         ...
3271    526.424171       13.894419        66.687695   4.435821           1
3272    392.449580       19.903225              NaN   2.798243           1
3273    432.044783       11.039070        69.845400   3.298875           1
3274    402.883113       11.168946        77.488213   4.708658           1
3275    327.459761       16.140368        78.698446   2.309149           1

[3276 rows x 10 columns]
```

Conclusion

This project's objective is to analyze water quality data and predict water potability using a dataset, following the design thinking process and various development phases. By conducting data preprocessing, exploratory data analysis, data visualization, and predictive modelling, we aim to provide valuable insights for assessing water quality and ensuring its potability. The insights obtained from this analysis can have a significant impact on water management, public health, and environmental conservation.

In summary, this comprehensive analysis will not only assess the quality of water but also empower decision-makers with the tools and knowledge needed to safeguard and enhance the safety of drinking water sources. Water quality analysis and potability prediction play a vital role in ensuring access to clean and safe drinking water, a fundamental human right.

LINK FOR JUPYTER NOTEBOOK (ipynb) :
https://github.com/cutieemagic/Divya_Venkatesan/blob/357d50f18c72d05efaeaedae7620a5685004af3c/DAC_Phase4.ipynb

LINK FOR JUPYTER NOTEBOOK (pdf) :
https://github.com/cutieemagic/Divya_Venkatesan/blob/357d50f18c72d05efaeaedae7620a5685004af3c/DAC_Phase4%20(part%201).pdf

LINK FOR IBM COGNOS VISUALIZATION (pdf) :
https://github.com/cutieemagic/Divya_Venkatesan/blob/357d50f18c72d05efaeaedae7620a5685004af3c/DAC_Phase4%20(part%202).pdf