Project Report: Water Quality prediction using Logistic Regression

1. Problem Statement: The goal is to create a system that can analyze various water parameters and determine whether the water is suitable for human consumption (potability) or for other purposes (non-potability).
For potability, the model should consider factors such as the concentration of contaminants, pH levels, dissolved oxygen, and other relevant indicators that affect the safety and suitability of water for drinking. The prediction could classify water samples into categories like "safe for drinking" or "not safe for drinking."
For non-potability, the model may focus on different parameters depending on the specific use case. For example, water quality for industrial processes, agriculture, or aquatic ecosystems might require consideration of different pollutants and environmental factors.

2. Machine Learning Algorithm Used and Block Diagram:

2.1 Machine Learning Algorithm:

Logistic regression is a statistical method used for binary classification problems. It predicts the probability that an instance belongs to a particular category, making it suitable for problems where the dependent variable is binary, meaning it has two possible outcomes (e.g., yes/no, 1/0, true/false).

In logistic regression, the logistic function (also called the sigmoid function) is used to model the relationship between the independent variables and the probability of the binary outcome. The logistic function produces values between 0 and 1, which can be interpreted as probabilities.

In logistic regression, the glm function in R stands for Generalized Linear Model, and it's used to fit various types of generalized linear models, including logistic regression.

Logistic regression is used instead of linear regression in situations where the dependent variable is binary or categorical. Linear regression is suitable for predicting continuous outcomes, but when the outcome

variable is binary (e.g., yes/no, 1/0), logistic regression is a more appropriate choice
Binary Outcome:

Logistic regression is designed for binary outcomes, where the dependent variable has only two possible values (e.g., 1 or 0, yes or no).

Linear regression, which predicts a continuous outcome, may not be suitable for modeling binary responses as it assumes a linear relationship between the predictors and the response, leading to predictions outside the [0, 1] range.

Logistic regression is a supervised machine learning model that has three types: binary, multinomial, and ordinal. These types differ in theory and execution.

2.2 Block Diagram:
Include a visual representation of the system architecture or workflow using a block diagram. Highlight the key components and their interconnections.

3. Data Set Used:

3.1 Description of DataSet:

We have taken the dataset for Water Quality prediction from Kaggle website.
Link: Water Quality (kaggle.com)

The dataset contains above 2000 data and the columns in the dataset are :
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 un-wanted 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 the 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 used 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.

4. Results and Discussions:

4.1 Data Structures Used:

i. To fill the null values we used mean values: Imputing null values (missing data) in a dataset by using the mean values is a common strategy, known as mean imputation. The rationale behind this approach is to replace missing values with the mean of the available values in the respective variable

ii. Use of Matrix : Matrices are employed in feature extraction techniques such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). These methods transform high-dimensional data into a lower-dimensional space using matrix operations, facilitating data compression and noise reduction. We use Matrix for training the model as it contains multiple independent variables and a single dependent variable

iii. We used "Data frame" as the dataset is in the form of CSV (.csv). Using Data Frames when dealing with CSV files enhances the ease of data manipulation, analysis, and interpretation. The combination of tabular representation, labelled columns, and integration with data analysis libraries makes Data Frames a powerful tool for working with structured data stored in CSV format.

iv. We scaled all the independent variables for easier understanding of the data. Scaling independent variables is crucial for several reasons to ensure that they contribute to the model in a meaningful and comparable way. Calling independent variables is a common pre-processing step that promotes fairness, stability, and interpretability in statistical modelling and machine learning. It helps address issues associated with disparate scales, making the modelling process more reliable and aiding in the interpretation of results

4.2 Data Loading Mechanism:

We have used the in-built functions to load the data into the dataset.

read.csv:

These functions are used to read data from flat files (e.g., text files, CSV files) into a data frame. The read.csv function is a specialized version of read.table for reading CSV files.

Packages used for this data:

i. ggplot2 :
ggplot2 is used for statistical computing and data representation. It can improve the quality and aesthetics of graphics, and make them more efficient to create. ggplot2 is an open-source package for an open-source programming language. It comes with a number of built-in data sets.

ii.dplyr:
The dplyr package in R is a data manipulation structure that provides a set of verbs to help solve common data manipulation problems. It is a powerful package that can manipulate, clean, and summarize unstructured data.

iii. Scales:
The scales package in R implements scales in a way that is graphics system agnostic. The package's graphical scales map data to aesthetics and provide methods for automatically determining breaks and labels for axes and legends. The scales package in R is part of the tidyverse and provides a set of functions for transforming and formatting scales in data visualizations.

iv. DescTools: The DescTools package in R is a collection of functions for data description, summary, and exploration. It is used by data scientists, researchers, and data analysts to understand their data and identify their findings.

4.3 Statistics Used:
The Statistics used in this data are:
i. Mean: In statistics, the mean is a measure of central tendency that represents the average value of a set of numbers. It is calculated by summing up all the values in a dataset and then dividing the sum by the total number of observations.

ii. Median: In R, you can calculate the median of a dataset using the median() function. The median is another measure of central tendency, and it represents the middle value of a dataset when it is sorted in ascending or descending order.

iii. Mode: Unlike mean and median, R does not have a built-in function specifically named mode to directly compute the mode of a dataset. However, you can find the mode using different approaches, as the mode is the value that occurs most frequently in a dataset.

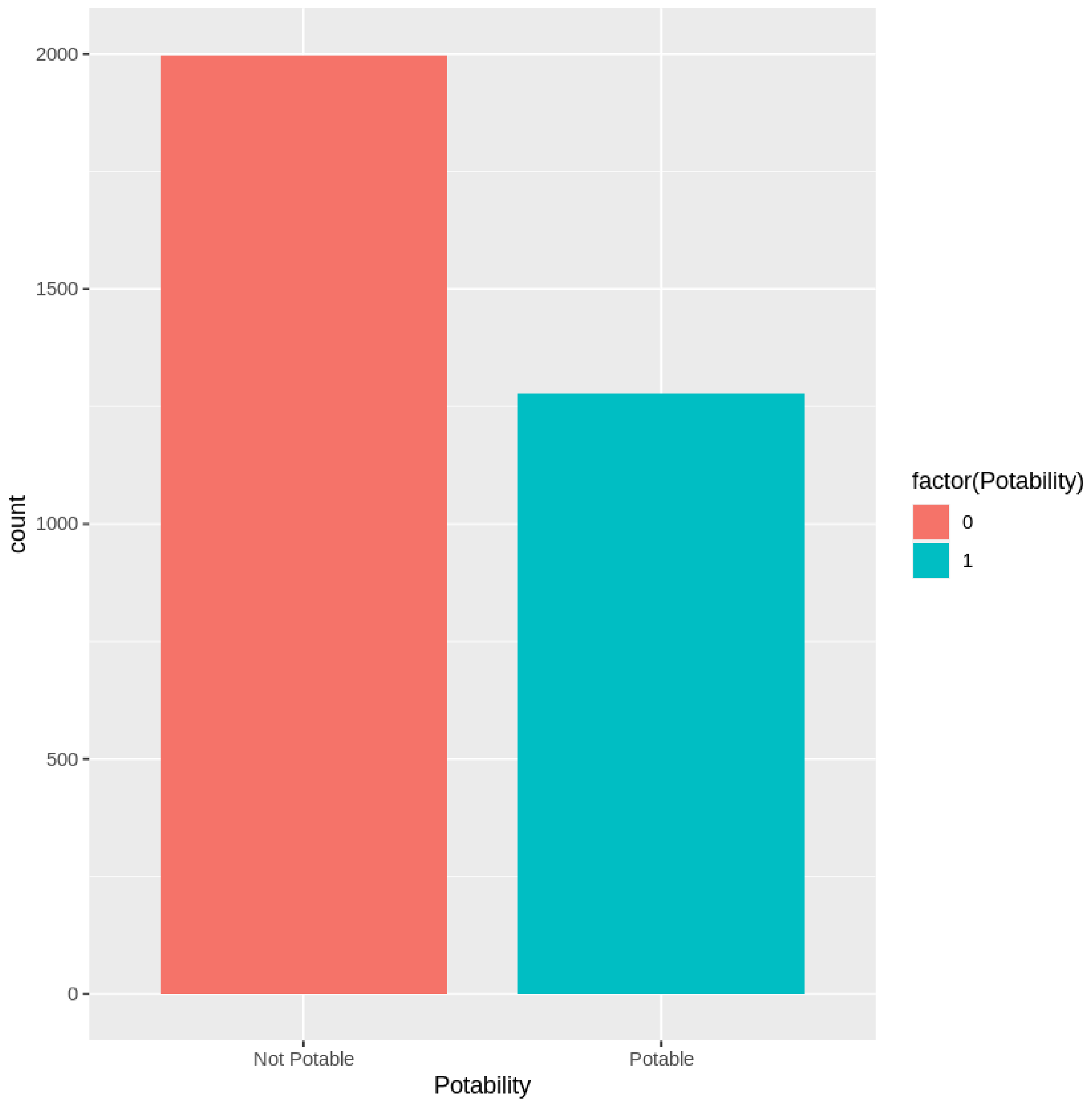
iv. Standard Deviation(sd): In R, you can calculate the standard deviation of a dataset using the sd() function. The standard deviation is a measure of the amount of variation or dispersion in a set of values. It indicates how much individual data points differ from the mean of the dataset.

v. Variance (var): In statistics, variance is a measure of the spread or dispersion of a set of values. In R, you can calculate the variance of a dataset using the var() function. The variance is computed as the average of the squared differences between each data point and the mean of the dataset.

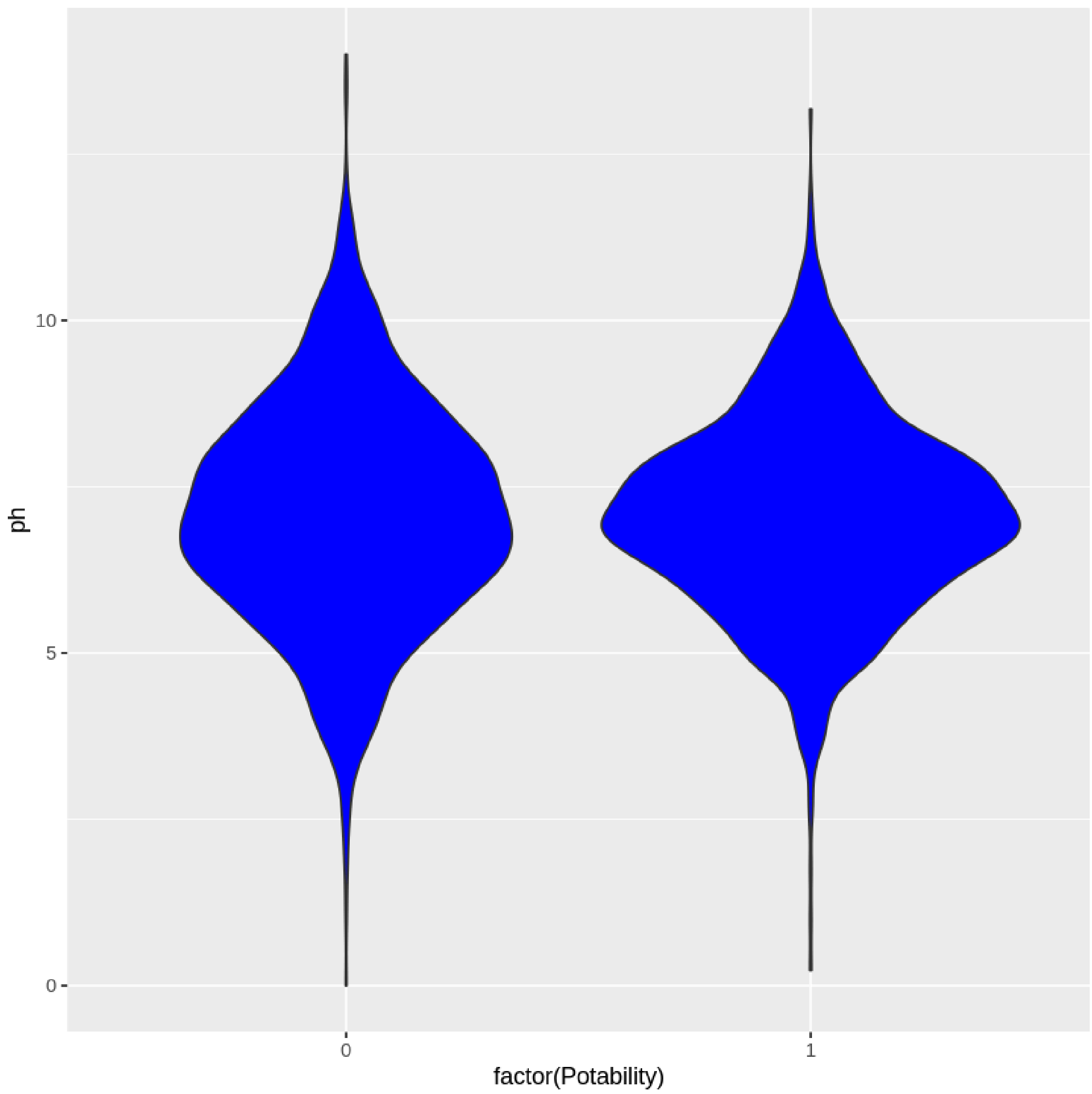
vi. Outliers: In statistics, outliers are data points that significantly differ from the rest of the observations in a dataset. Outliers can skew statistical analyses and models, so it's important to identify and handle them appropriately. For example : Boxplot and Z-score

5. Visualization:

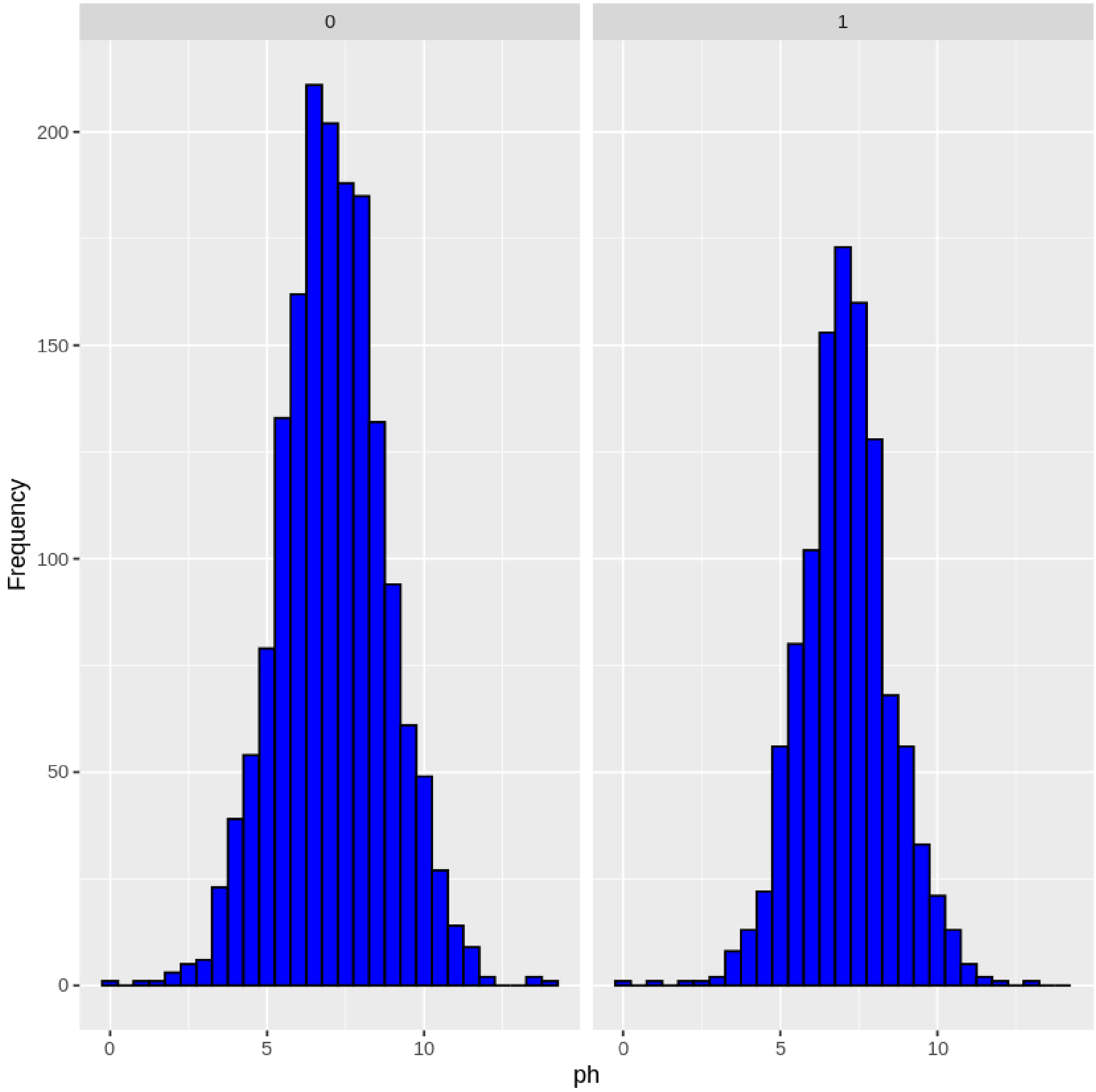
i. Bar chart of Dependent variable



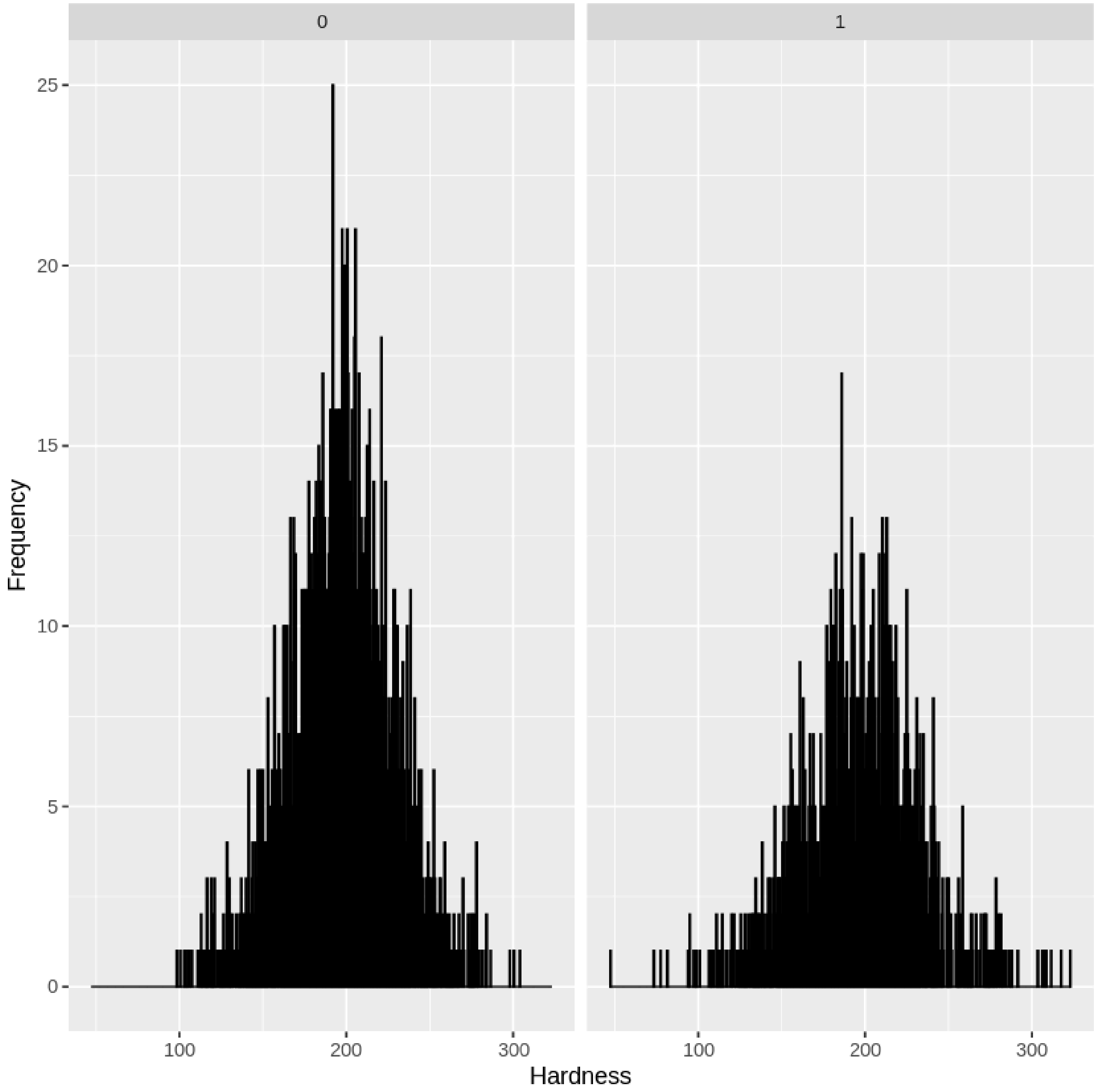
ii. Violin plot of Dependent variable



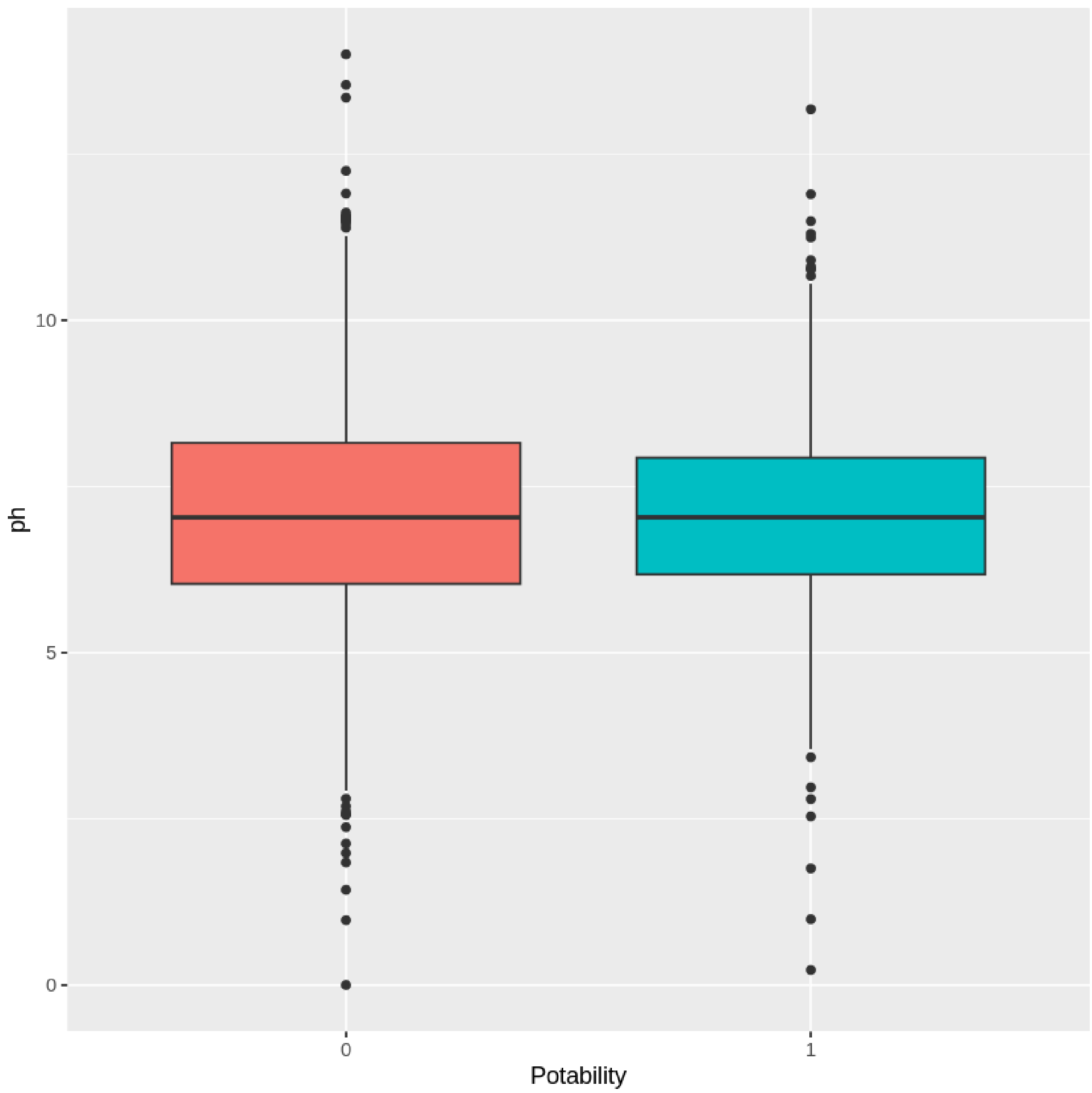
iii. Frequency of ph-value



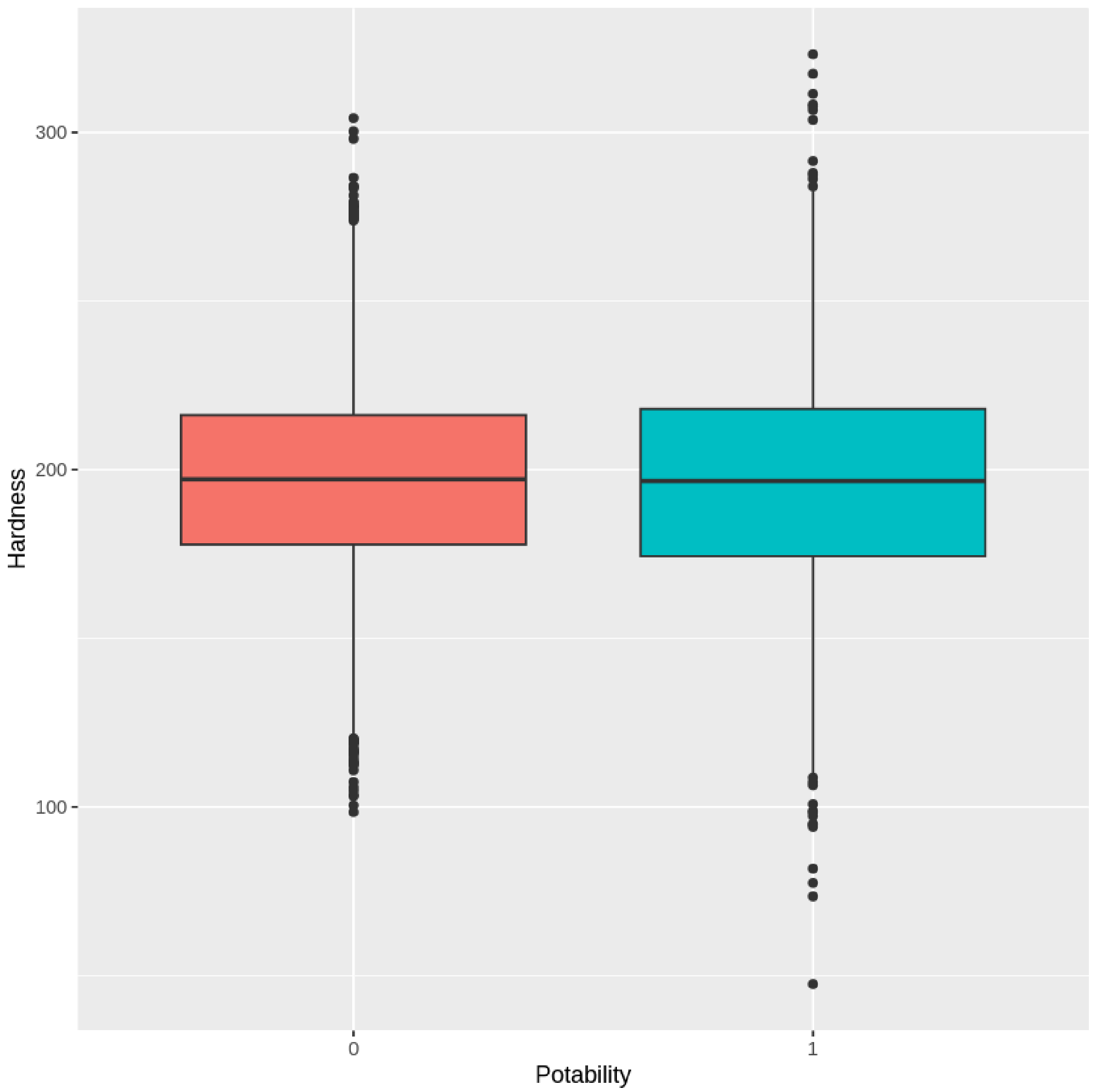
iv. Frequency of Hardness



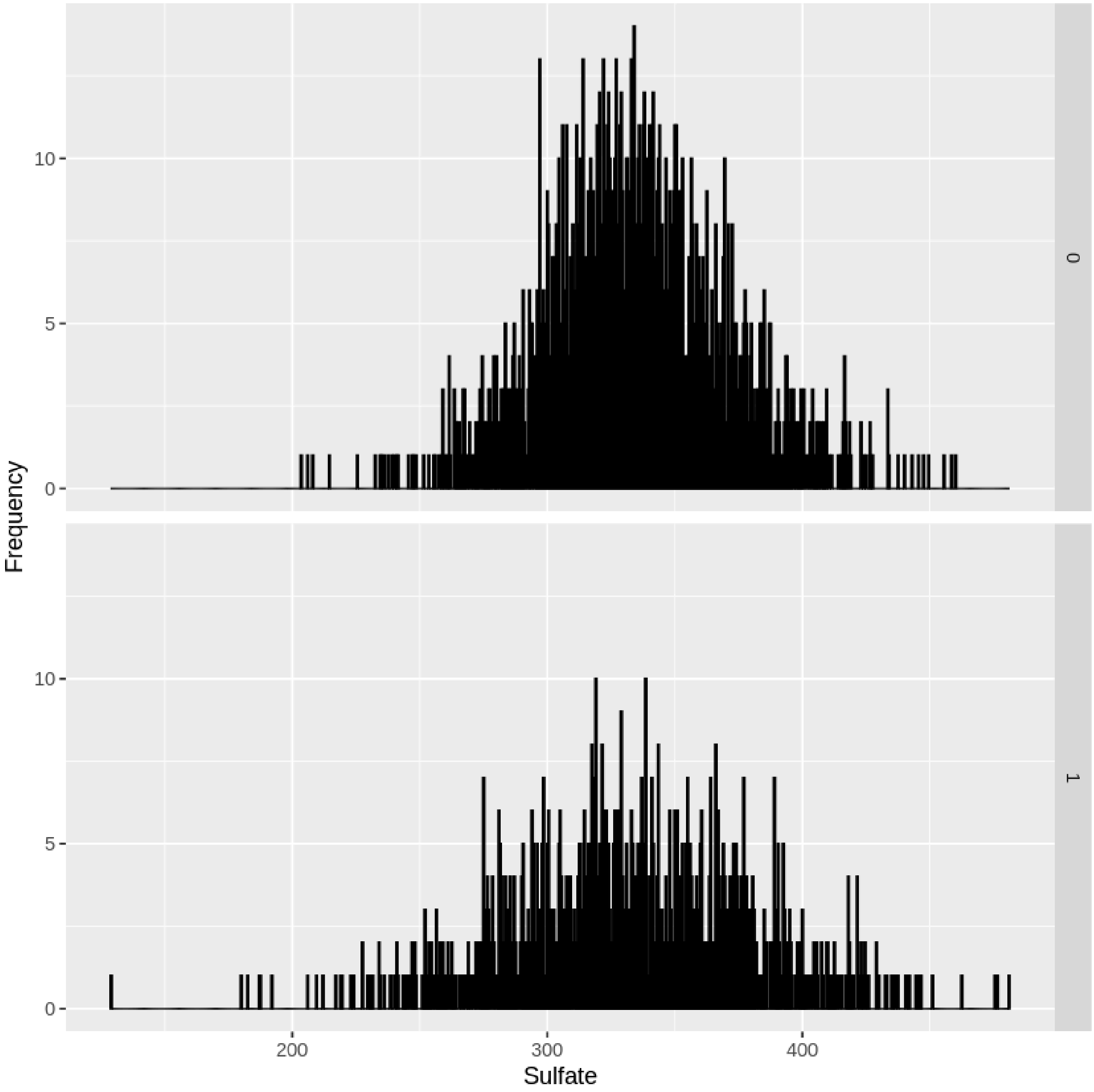
v. Boxplot of PH



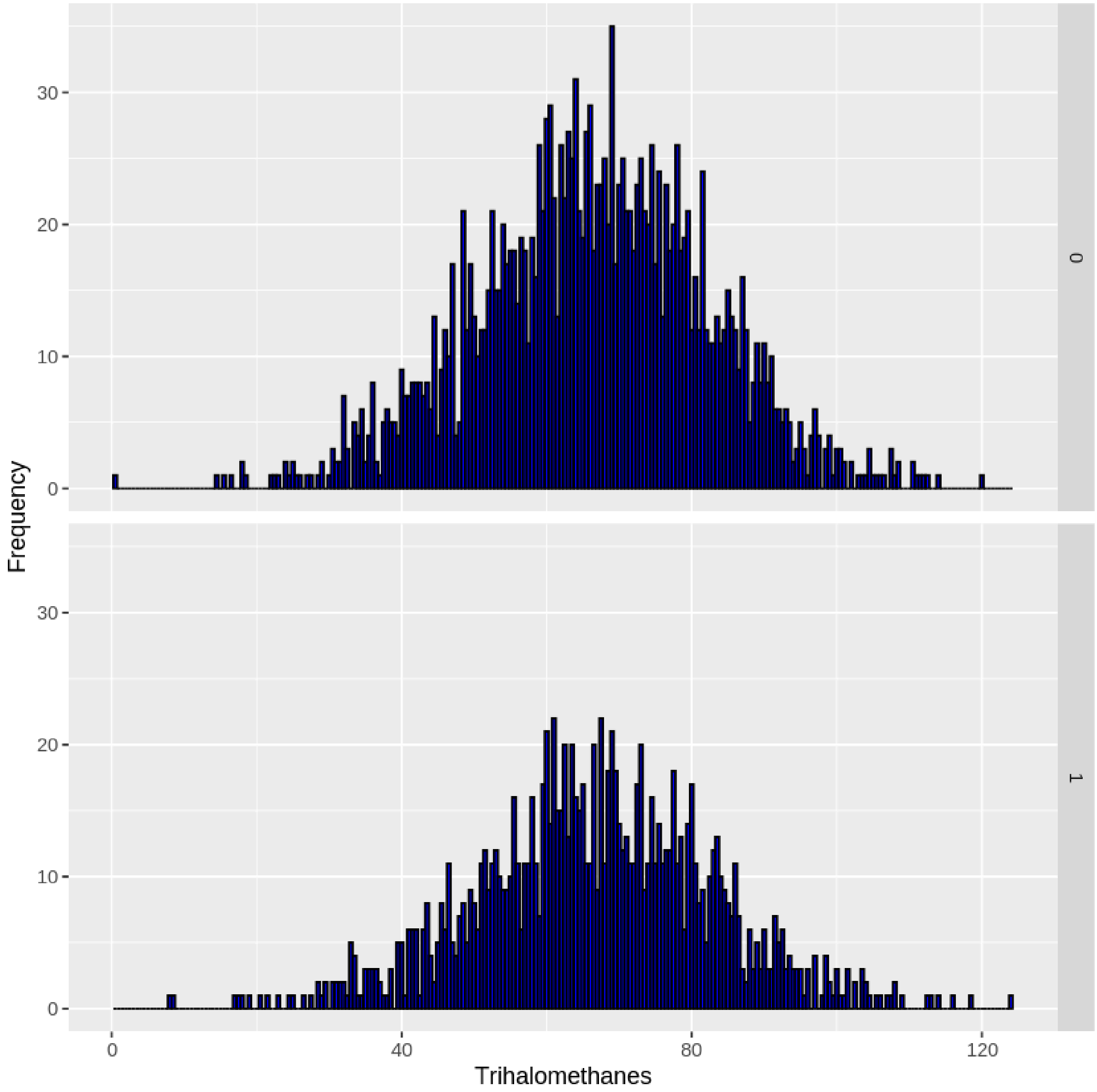
vi. Boxplot of Hardness



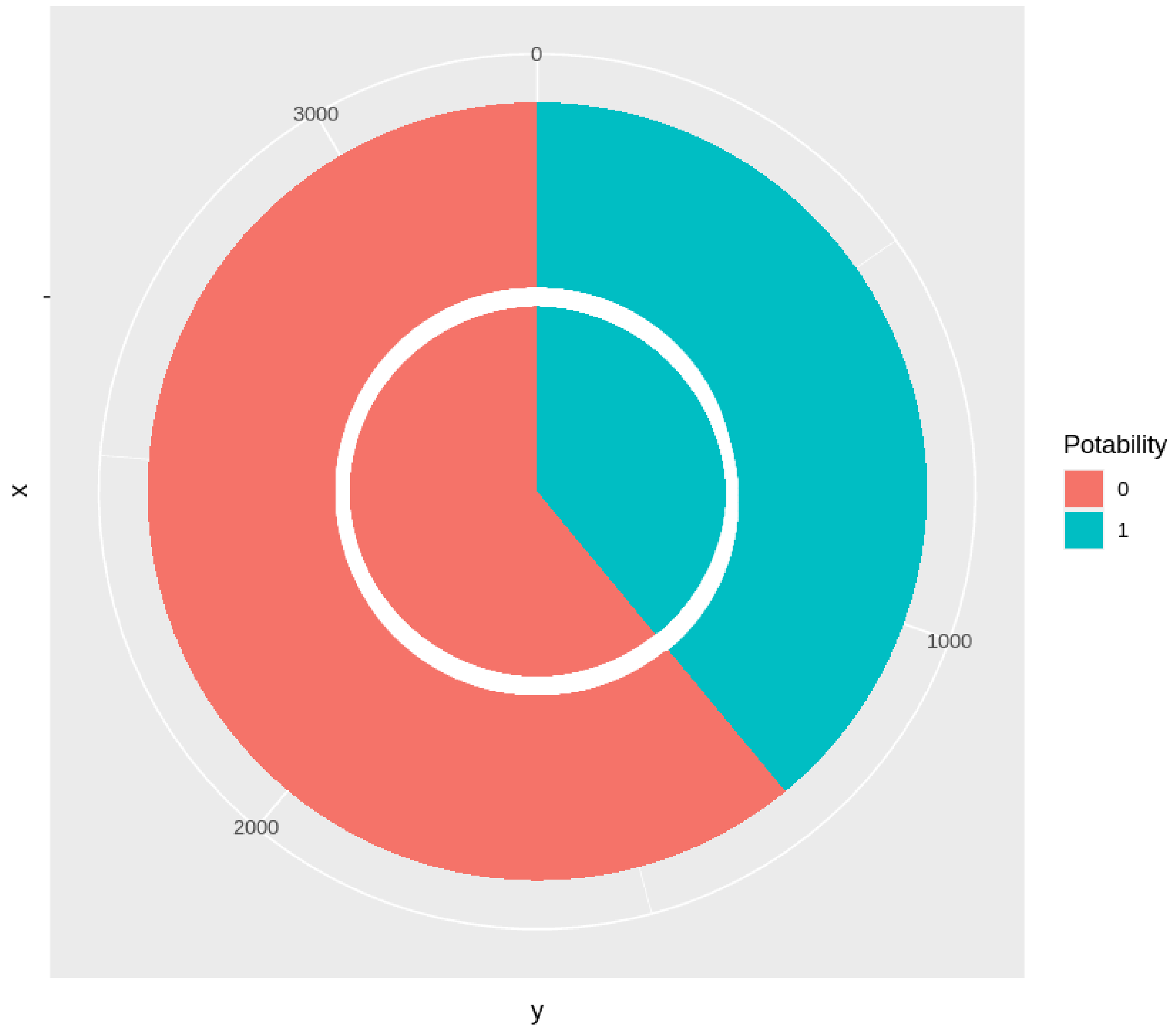
vii. Frequency of Sulfates



viii. Frequency of Tri-halo-methanes



ix. Pie chart of Potablity:



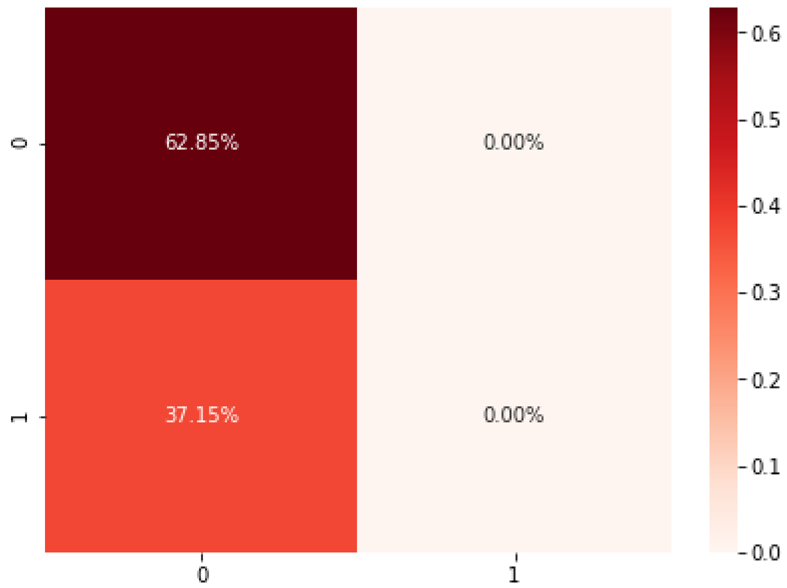
6. Conclusion:

The accuracy score of the model is `0.6284658040665434`

```
 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
```



7. Future Work:

Ensemble Methods: The accuracy of the logistic regression model can potentially be improved by using ensemble methods. For example, an ensemble model using the adaptive boosting technique was able to improve the forecast accuracy of the logistic regression model.

Comparative Analysis with Other Models: The performance of logistic regression can be compared with other machine learning models like Support Vector Machine (SVM), Decision Tree (DT), Random Forest, Gradient Boost, and Ada Boost. This comparative analysis can provide insights into the strengths and weaknesses of logistic regression in predicting water potability.

8. References:

https://www.geeksforgeeks.org/
https://stackoverflow.com/
ChatGPT (openai.com)
Water Quality (kaggle.com)