Predicting Water Quality using Machine Learning

Vincent Katunga-Phiri

2024-12-11

Introduction

Water is an essential component of life and the earth's ecosystems. Increased urbanisation and industrialisation have increased water demand while also reducing water quality (Ejigu, 2021). Oceans, lakes, dams, and rivers are the primary water sources for a variety of applications, including drinking, irrigation, and public use. According to Spellman (2008), water quality includes the physical, chemical, and biological properties of water. Potable water is defined as water that is safe to drink, tastes good, and is appropriate for domestic use (Spellman, 2008). This project aims to predict potable water using machine learning techniques based on available water parameters such as ph, hardness, solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, and turbidity.

The project will include several essential steps, including the methods section describing the machine learning methodologies utilised, the data cleaning processes, and the data exploration processes, along with an interpretation of the insights derived from each exploration. The model development methodology will also be addressed. The results section will summarise the model's outcomes and performance metrics using the accuracy, precision, recall and f1 score. The conclusion section will summarise the project, emphasising its limitations and prospective avenues for future research and lastly a reference section.

Methods

Dataset Description

The dataset has 3276 observations and 10 variables. the variables include; `ph`: An indicator of acidic or alkaline condition of water status. `Hardness`: The concentration of calcium and magnesium ions in water, which affects its ability to lather with soap. `Solids`: The total dissolved solids in water, indicating the amount of inorganic and organic substances dissolved in it. `Chloramines`: Compounds of chlorine and ammonia used as a disinfectant in water treatment to control bacteria and pathogens. `Sulfate`: The concentration of sulfate ions (SO^{2}) in water, which can affect taste and, at high levels, cause laxative effects. `Conductivity`: A measure of water's ability to conduct electricity, which reflects the concentration of dissolved salts or ions. `Organic_carbon`: The amount of organic compounds in water, often used as an indicator of water quality and pollution. `Trihalomethanes`: The cloudiness or haziness of water caused by suspended particles, which can affect aesthetic quality and microbial safety. `Potability`: An indicator of whether water is safe to drink, considering its chemical, physical, and biological quality.

```
## 'data.frame':    3276 obs. of  10 variables:
##  $ ph             : num  NA 3.72 8.1 8.32 9.09 ...
##  $ Hardness       : num  205 129 224 214 181 ...
##  $ Solids         : num  20791 18630 19910 22018 17979 ...
##  $ Chloramines    : num  7.3 6.64 9.28 8.06 6.55 ...
##  $ Sulfate        : num  369 NA NA 357 310 ...
##  $ Conductivity   : num  564 593 419 363 398 ...
##  $ Organic_carbon : num  10.4 15.2 16.9 18.4 11.6 ...
##  $ Trihalomethanes: num  87 56.3 66.4 100.3 32 ...
##  $ Turbidity      : num  2.96 4.5 3.06 4.63 4.08 ...
```

```
##  $ Potability     : int  0 0 0 0 0 0 0 0 0 0 ...
```

From the output above, all variables are numeric variables (continuous) except for the potability variable which is numeric. We can also see NA values in the dataset and these have to be handled before performing the machine learning models.

```
##        ph            Hardness          Solids         Chloramines
##  Min.   : 0.000   Min.   : 47.43   Min.   :  320.9   Min.   : 0.352
##  1st Qu.: 6.093   1st Qu.:176.85   1st Qu.:15666.7   1st Qu.: 6.127
##  Median : 7.037   Median :196.97   Median :20927.8   Median : 7.130
##  Mean   : 7.081   Mean   :196.37   Mean   :22014.1   Mean   : 7.122
##  3rd Qu.: 8.062   3rd Qu.:216.67   3rd Qu.:27332.8   3rd Qu.: 8.115
##  Max.   :14.000   Max.   :323.12   Max.   :61227.2   Max.   :13.127
##  NA's   :491
##     Sulfate       Conductivity   Organic_carbon  Trihalomethanes
##  Min.   :129.0   Min.   :181.5   Min.   : 2.20   Min.   :  0.738
##  1st Qu.:307.7   1st Qu.:365.7   1st Qu.:12.07   1st Qu.: 55.845
##  Median :333.1   Median :421.9   Median :14.22   Median : 66.622
##  Mean   :333.8   Mean   :426.2   Mean   :14.28   Mean   : 66.396
##  3rd Qu.:360.0   3rd Qu.:481.8   3rd Qu.:16.56   3rd Qu.: 77.337
##  Max.   :481.0   Max.   :753.3   Max.   :28.30   Max.   :124.000
##  NA's   :781                                     NA's   :162
##    Turbidity       Potability
##  Min.   :1.450   Min.   :0.0000
##  1st Qu.:3.440   1st Qu.:0.0000
##  Median :3.955   Median :0.0000
##  Mean   :3.967   Mean   :0.3901
##  3rd Qu.:4.500   3rd Qu.:1.0000
##  Max.   :6.739   Max.   :1.0000
##
```

From the output above, we can notice that some variables have a huge difference between the 3rd quantile and the maximum value in the dataset indicating some outliers.

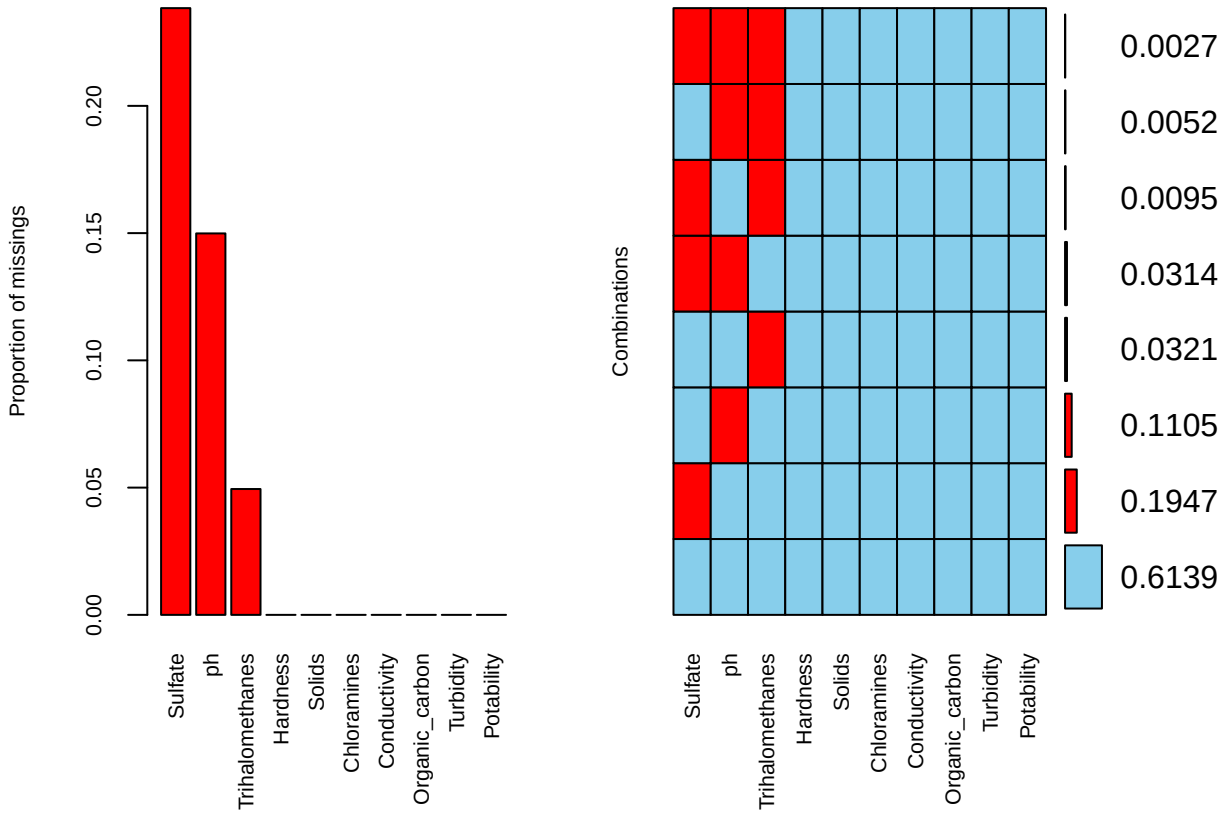
```
##         ph Hardness   Solids Chloramines  Sulfate Conductivity Organic_carbon
## 1       NA 204.8905 20791.32    7.300212 368.5164     564.3087      10.379783
## 2 3.716080 129.4229 18630.06    6.635246       NA     592.8854      15.180013
## 3 8.099124 224.2363 19909.54    9.275884       NA     418.6062      16.868637
## 4 8.316766 214.3734 22018.42    8.059332 356.8861     363.2665      18.436524
## 5 9.092223 181.1015 17978.99    6.546600 310.1357     398.4108      11.558279
## 6 5.584087 188.3133 28748.69    7.544869 326.6784     280.4679       8.399735
##   Trihalomethanes Turbidity Potability
## 1        86.99097  2.963135          0
## 2        56.32908  4.500656          0
## 3        66.42009  3.055934          0
## 4       100.34167  4.628771          0
## 5        31.99799  4.075075          0
## 6        54.91786  2.559708          0
```

The output above is just a glimpse of the 6 first observations in the dataset. NA values can be seen.

Missing Data

Since we have seen that our dataset comprises of missing values, the figure below is a visual representation of the extent of missingness in the dataset. There are 1434 observations with missing values in either of the variables in the dataset.

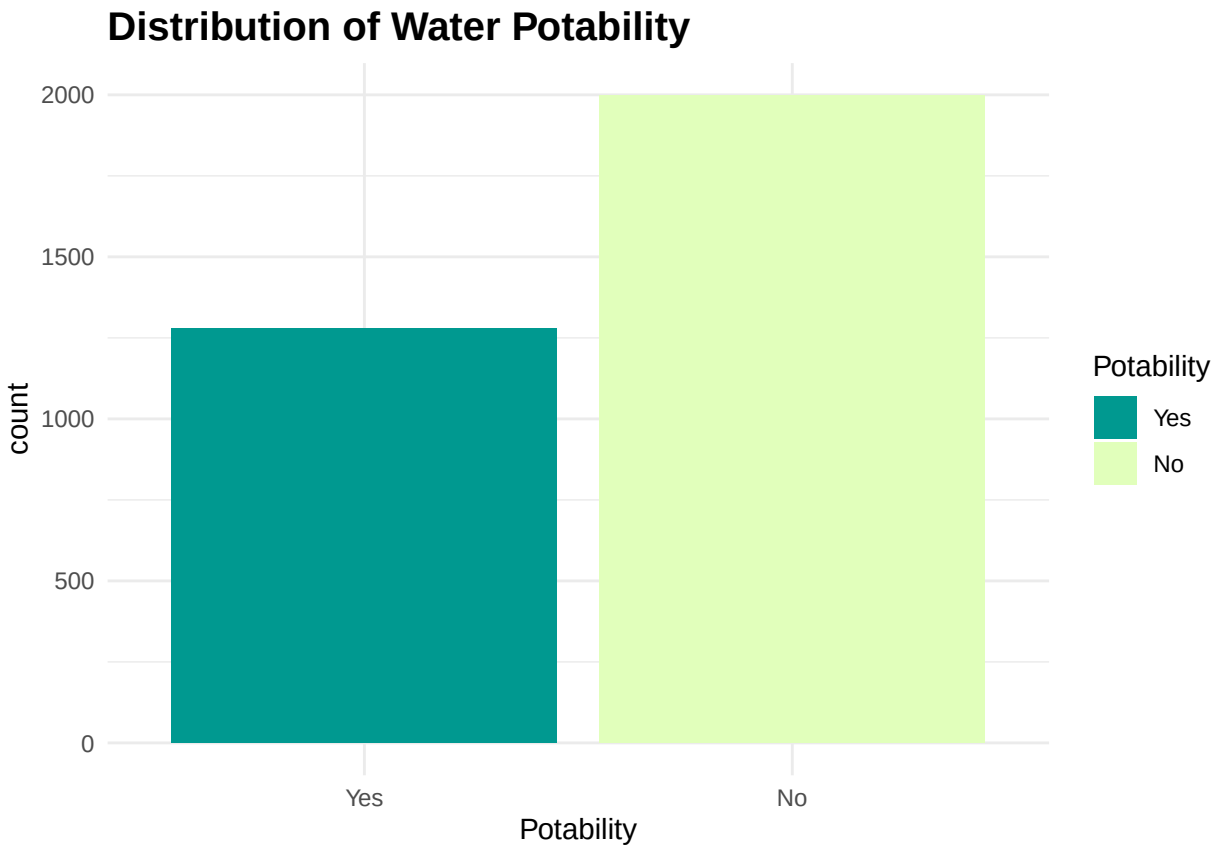
```
## [1] 1434
```




```
##
##  Variables sorted by number of missings:
##          Variable      Count
##           Sulfate 0.23840049
##                ph 0.14987790
##  Trihalomethanes 0.04945055
##          Hardness 0.00000000
##            Solids 0.00000000
##       Chloramines 0.00000000
##      Conductivity 0.00000000
##    Organic_carbon 0.00000000
##         Turbidity 0.00000000
##        Potability 0.00000000
```

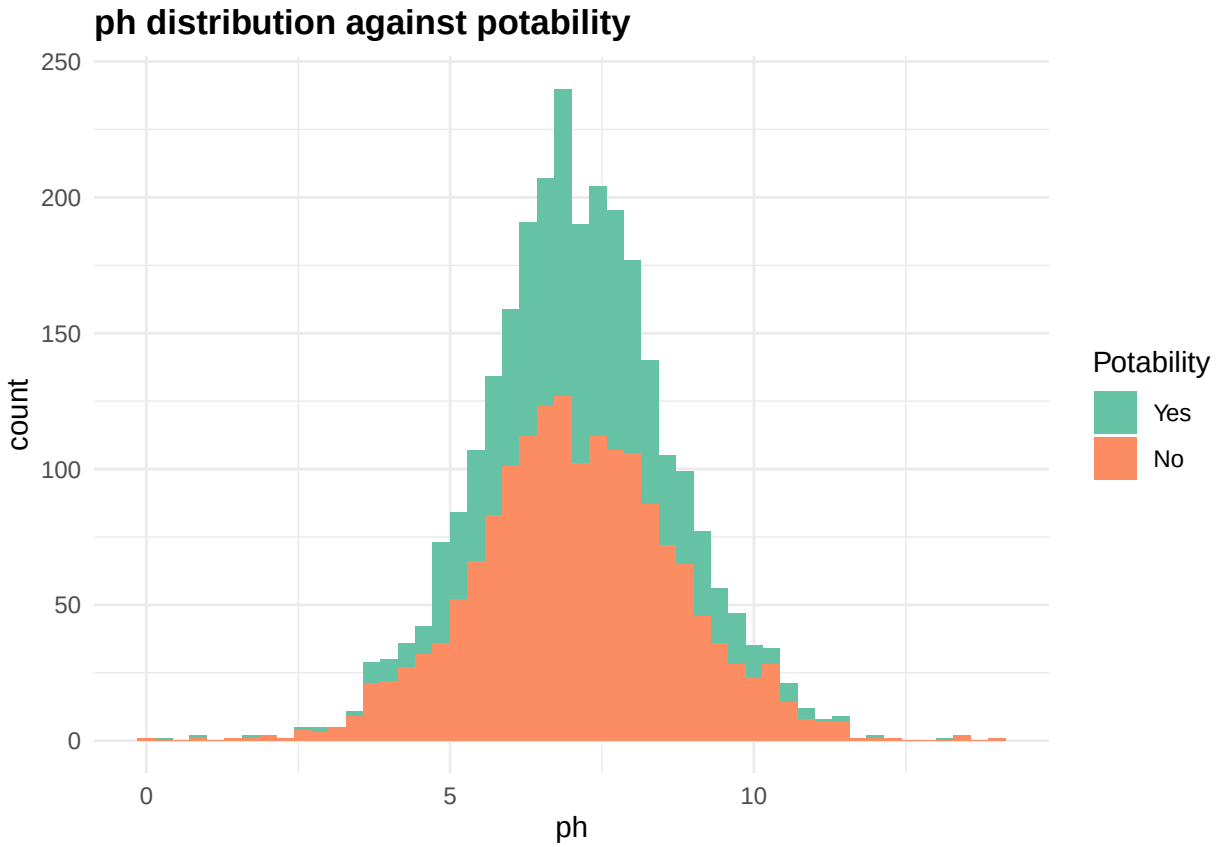
From the chart above, the variables `Sulfate` and `ph` have the highest number of missing values with over 20% missing values for `Sulfate` and 15% for `ph`. Looking at this, there is need for dealing with the missing data, either by imputing with the mean, mode, median or using a complex algorithm for imputation like k Nearest Neighbors (KNN) algorithm.

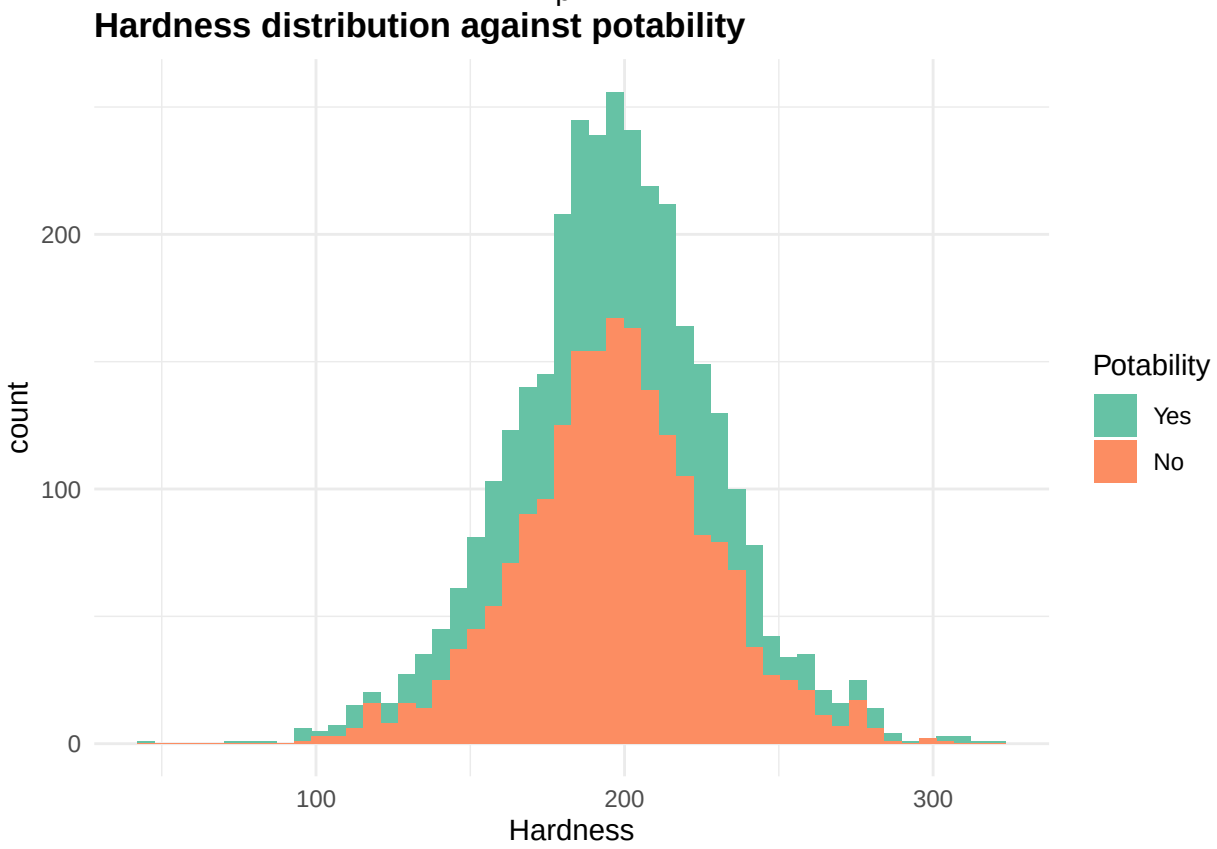
Explaratory Analysis

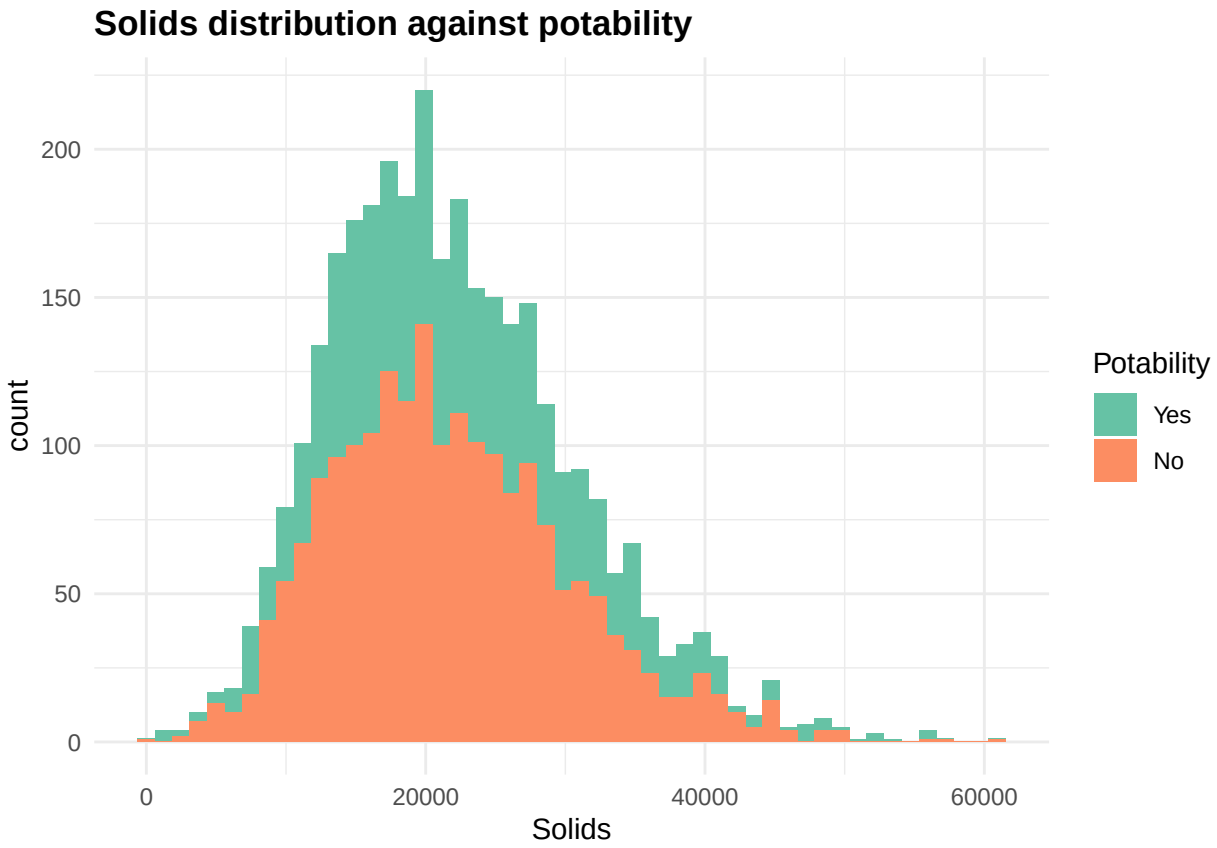


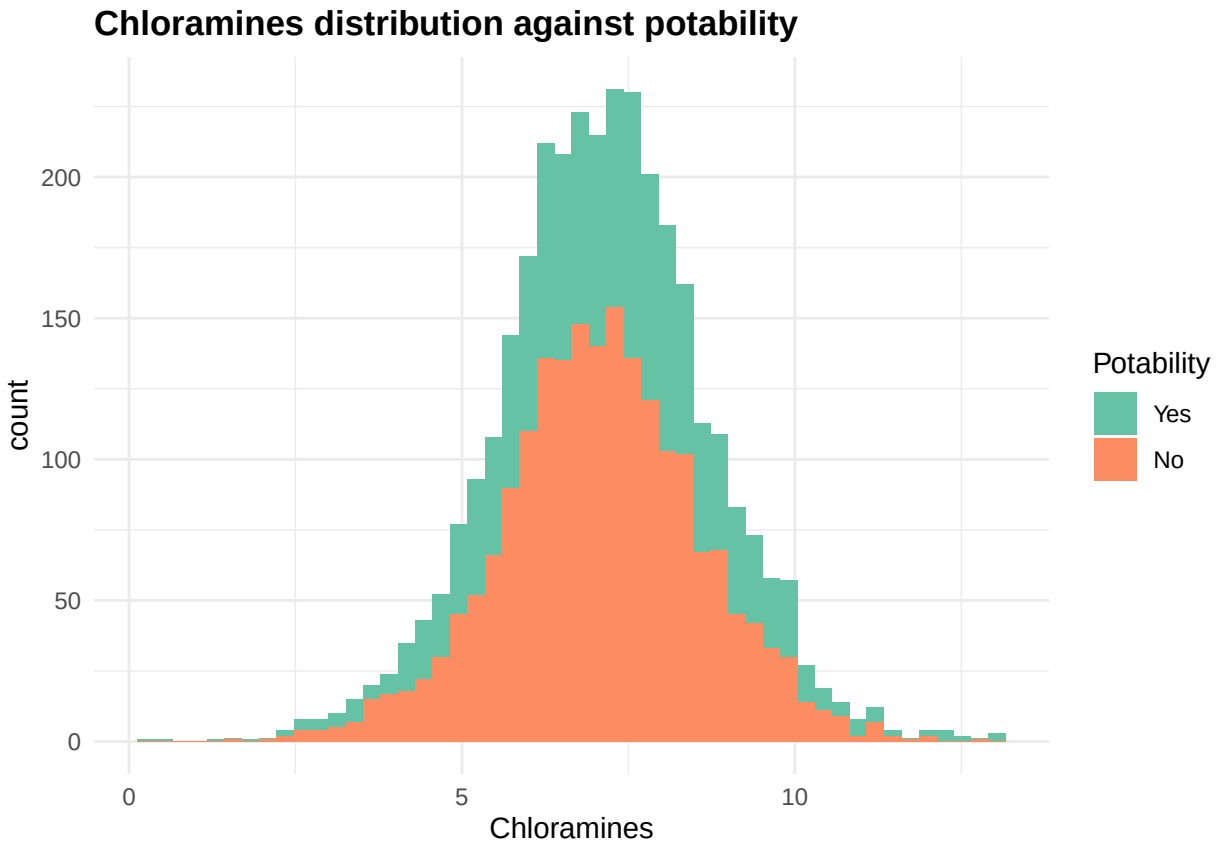
The plot above shows that the dataset contains a substantial number of observations with non-potable water, approximately 2000, representing approximately 59%, while observations of potable water make up approximately 40%.

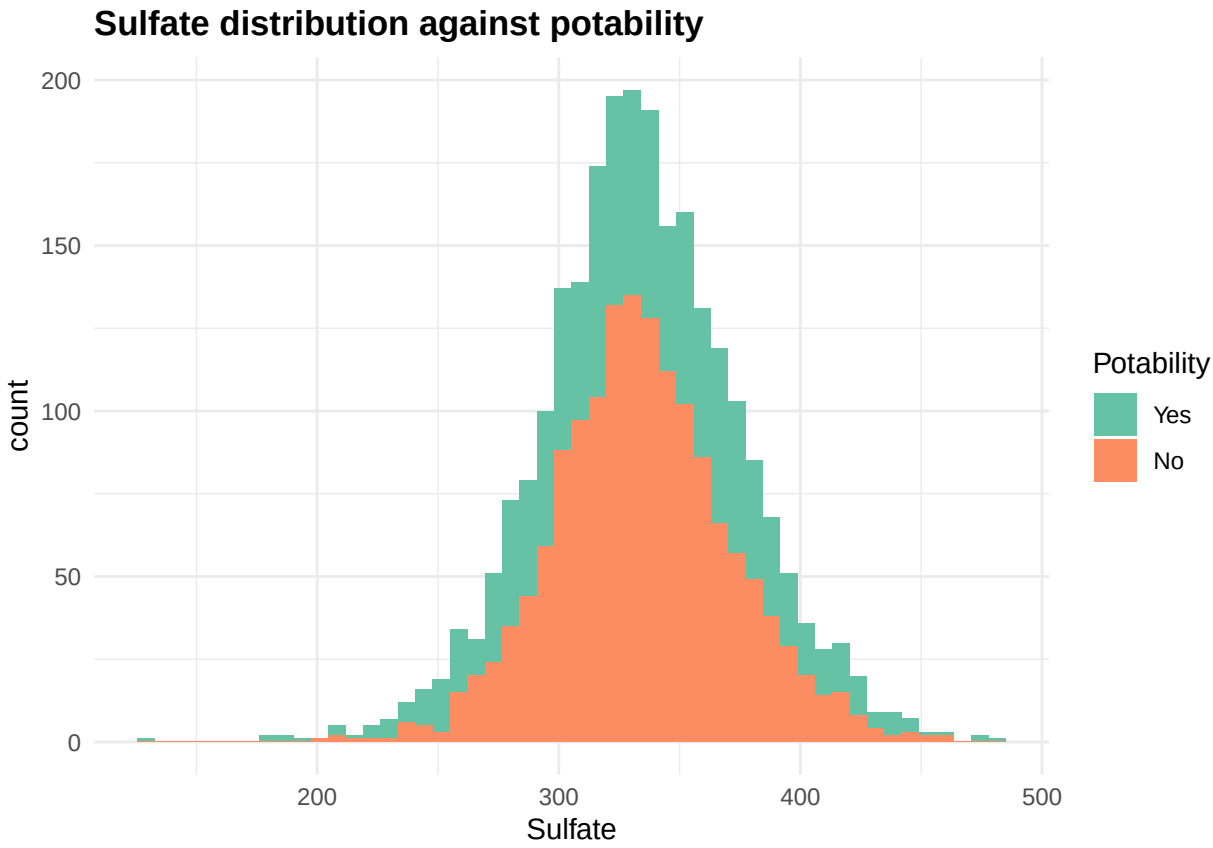
Distribution of the water parameters against Potability

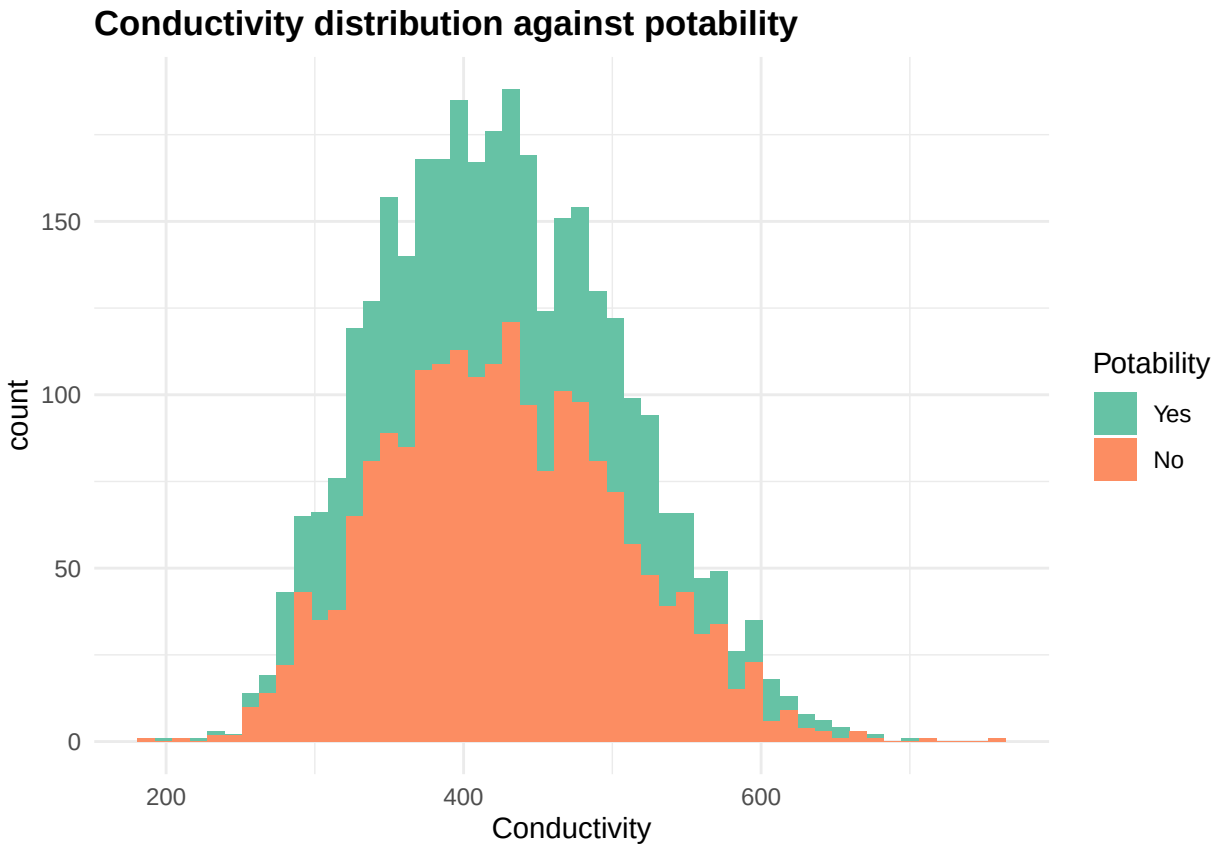


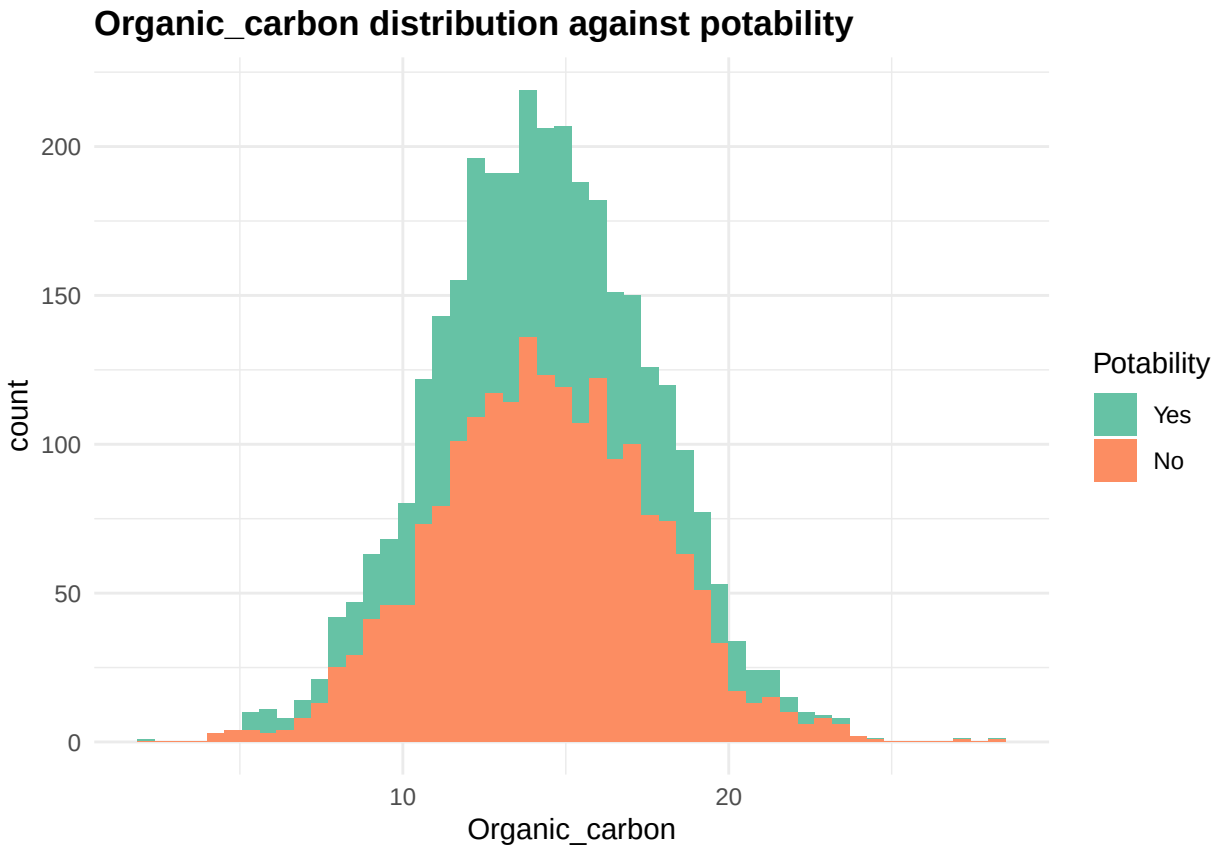


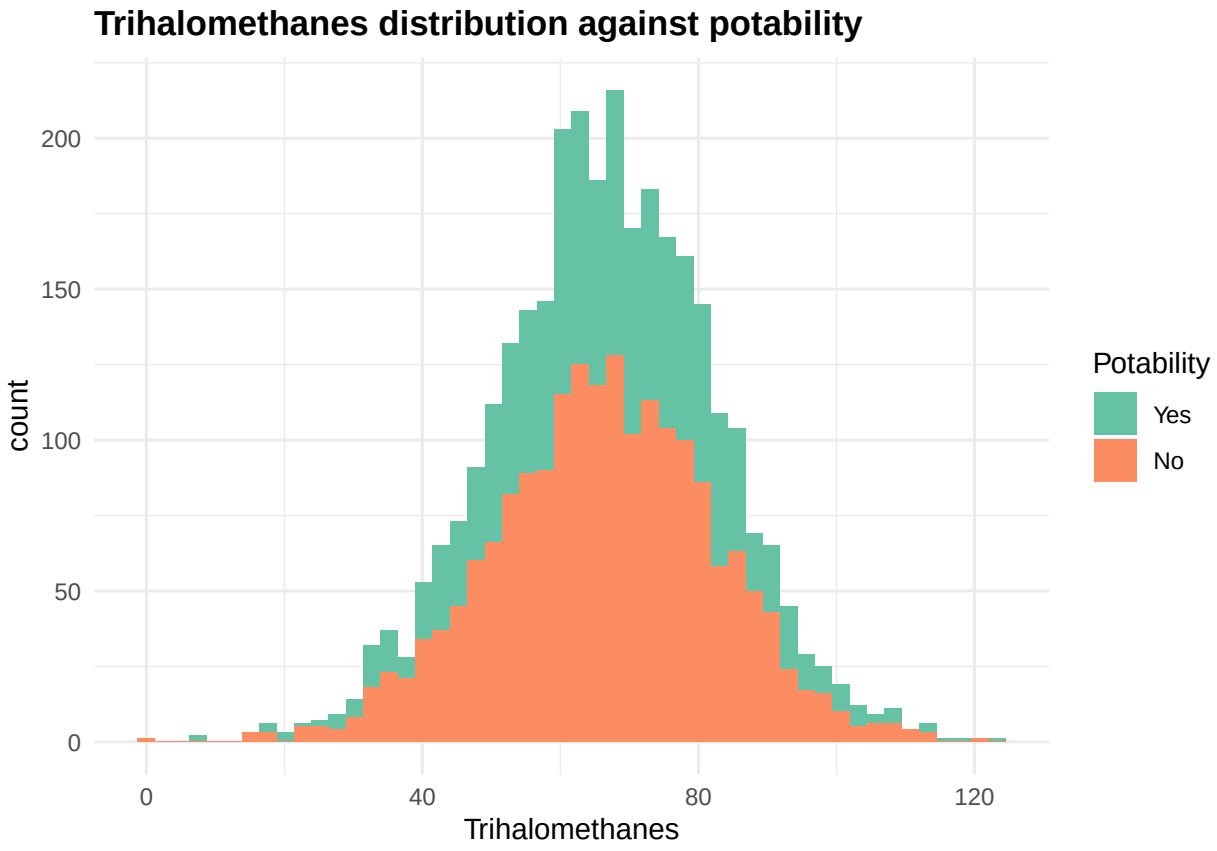

Solids distribution against potability
count
200
150
100
50
0
0
20000
40000
60000
Solids
Potability
Yes
No

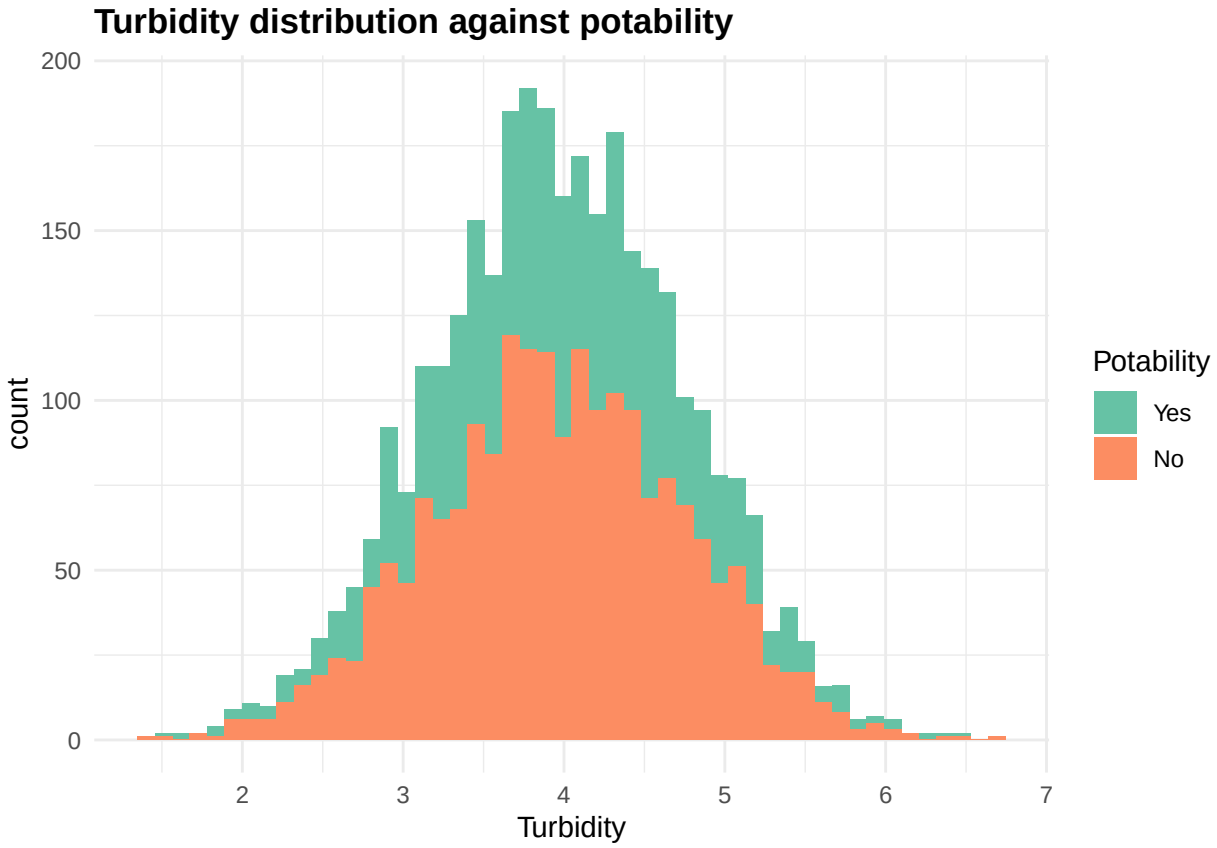

Chloramines distribution against potability
count
200
150
100
50
0
0
5
10
Chloramines
Potability
Yes
No


Sulfate distribution against potability
count
200
150
100
50
0
200
300
400
500
Sulfate
Potability
Yes
No


Conductivity distribution against potability
count
150
100
50
0
200
400
600
Conductivity
Potability
Yes
No


Organic_carbon distribution against potability
count
200
150
100
50
0
10
20
Organic_carbon
Potability
Yes
No

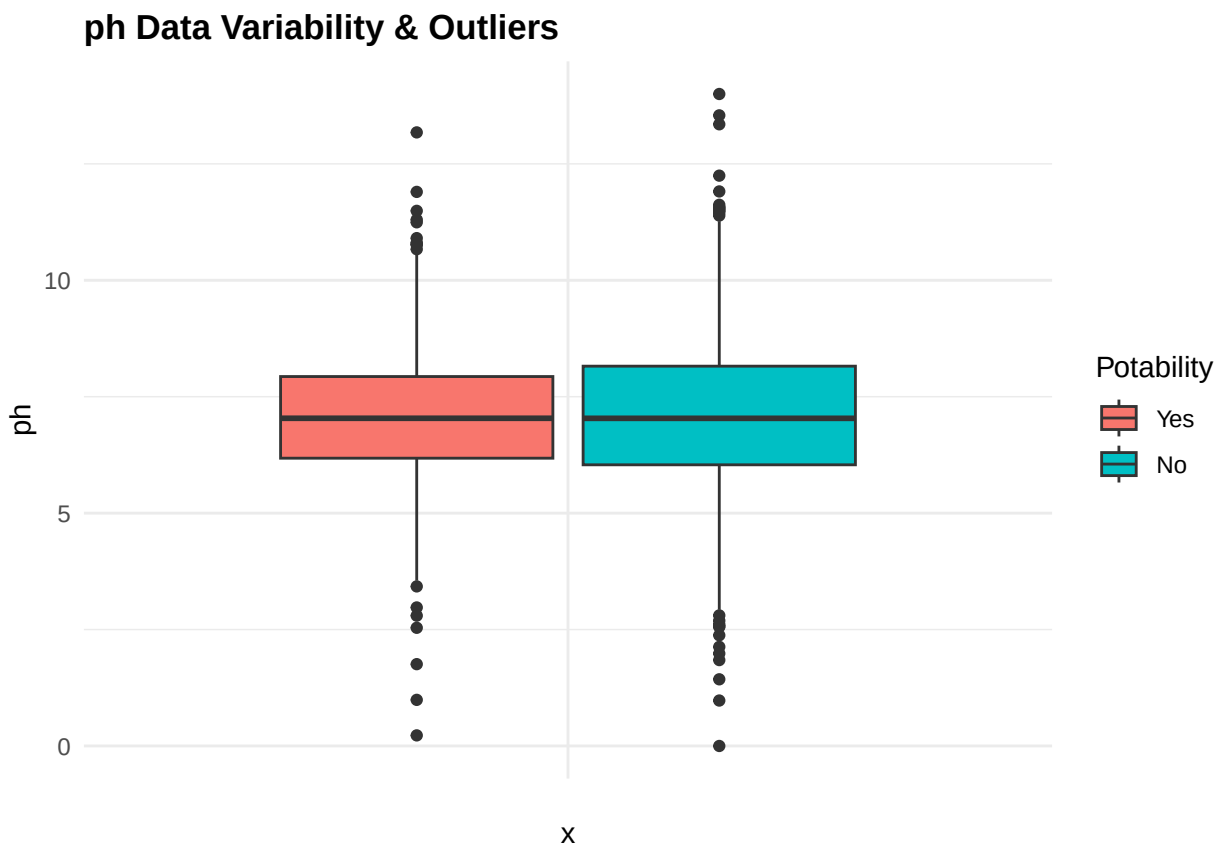

Trihalomethanes distribution against potability
count
200
150
100
50
0
0
40
80
120
Trihalomethanes
Potability
Yes
No

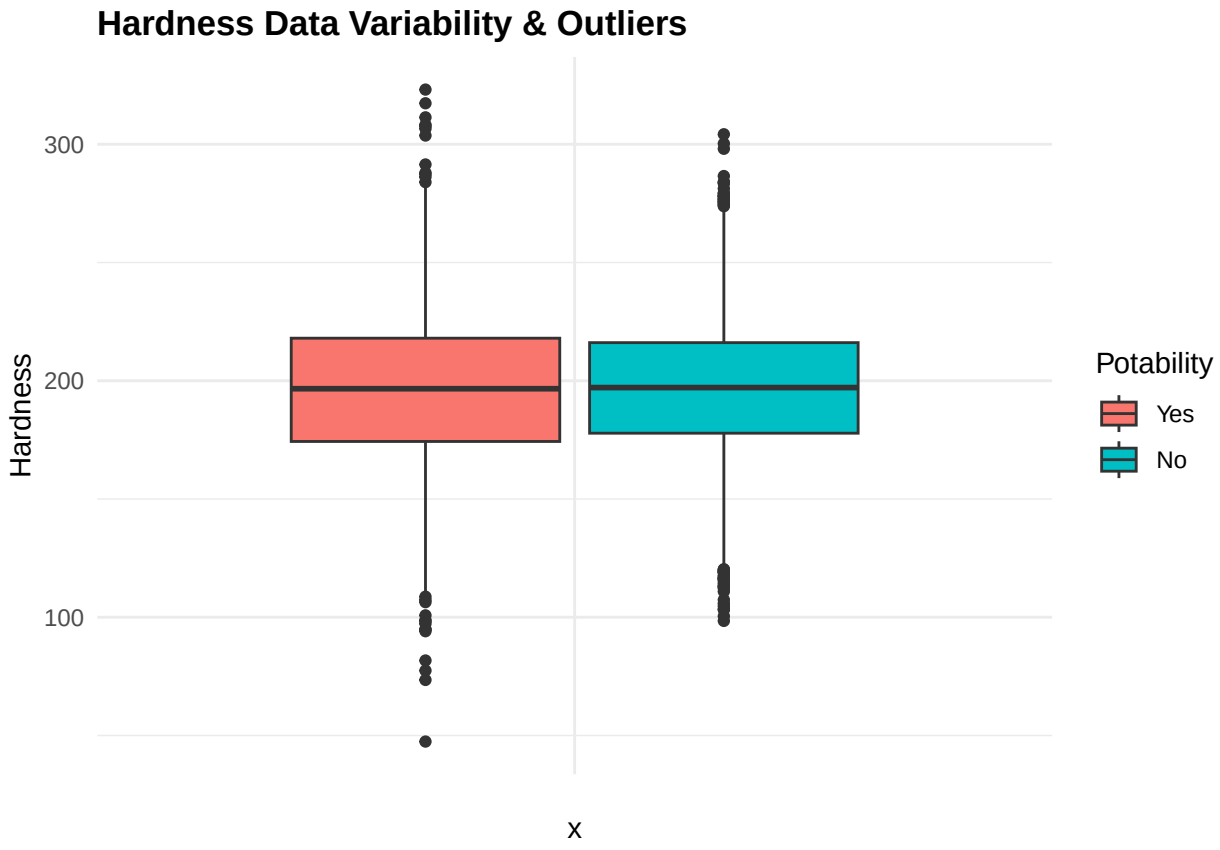


The output indicates that nearly all variables in the dataset show a normal distribution. Despite the presence of outliers, imputation utilising mean, mode, or median will be easily performed due to the normal distribution of the data. It also improves model efficacy. In nearly all distributions, the spread is similar, indicating comparable variability of parameter values for potable and non-potable water. Furthermore, all distributions indicate that the dataset contains a greater quantity of non-potable water compared to potable water. The distributions for both potable and non-potable water appear to be centred at comparable levels across all distributions.

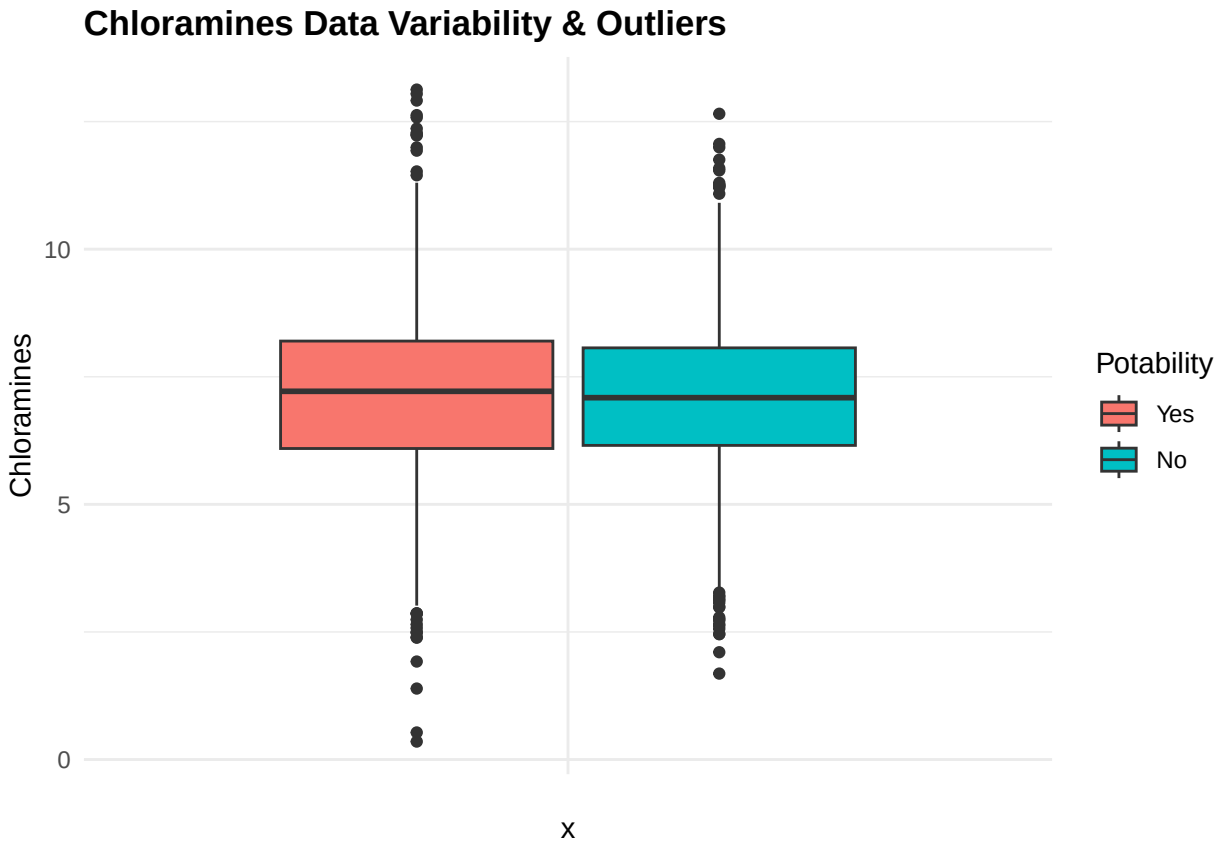
Variability of water parameters against potability

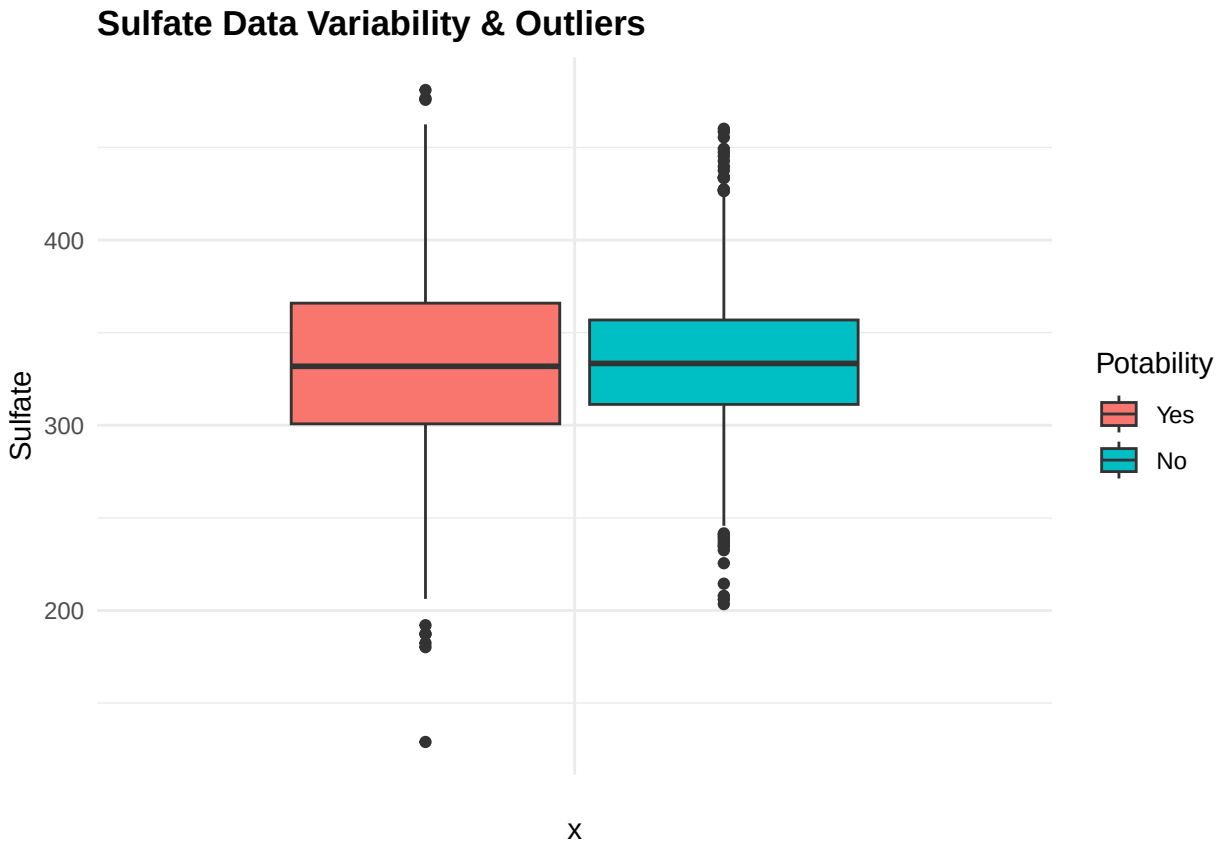
The box plots below confirms our observation that the variability of the data in both potable and non-potable is similar e.g., the mean looks the same in all distributions. We can also notice outliers in the dataset. The quantiles also look similar.

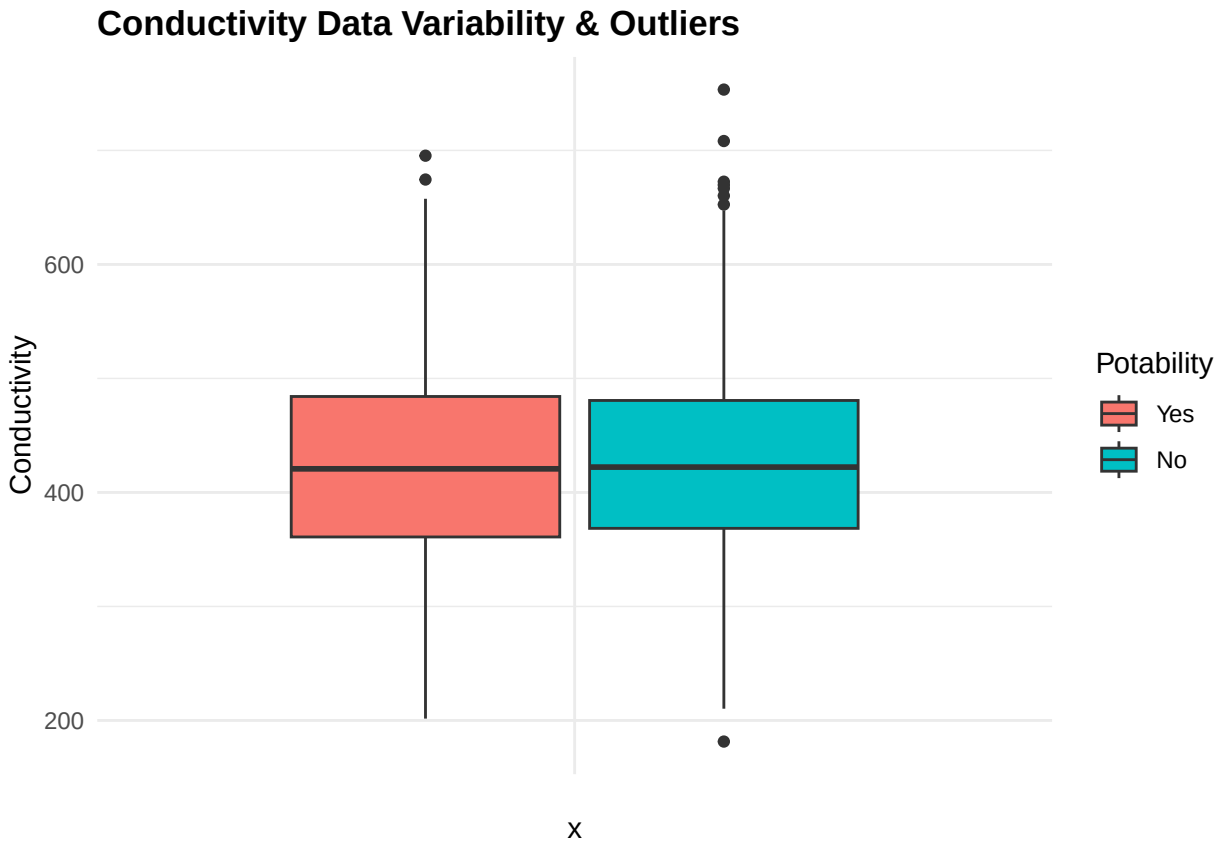

ph Data Variability & Outliers
10
5
0
ph
x
Potability
Yes
No

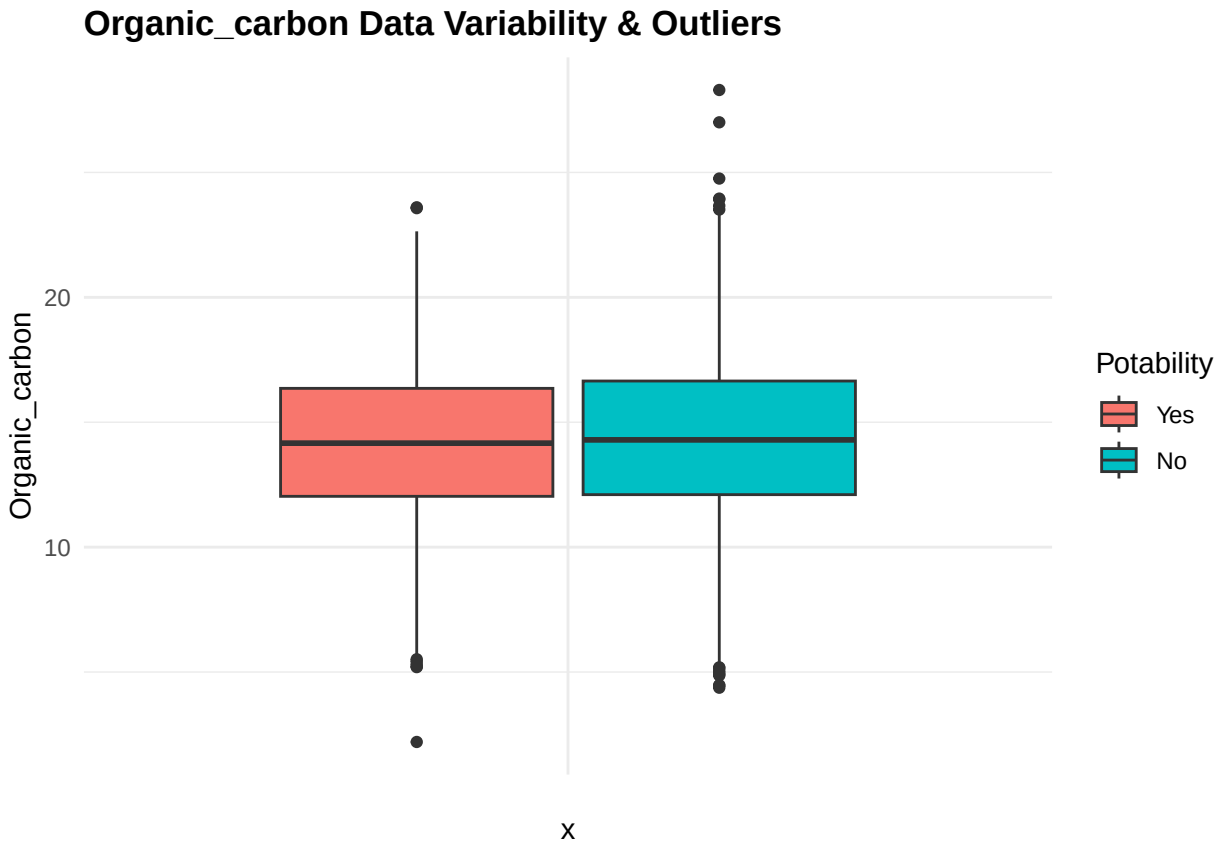

Hardness Data Variability & Outliers
300
200
100
Hardness
x
Potability
Yes
No

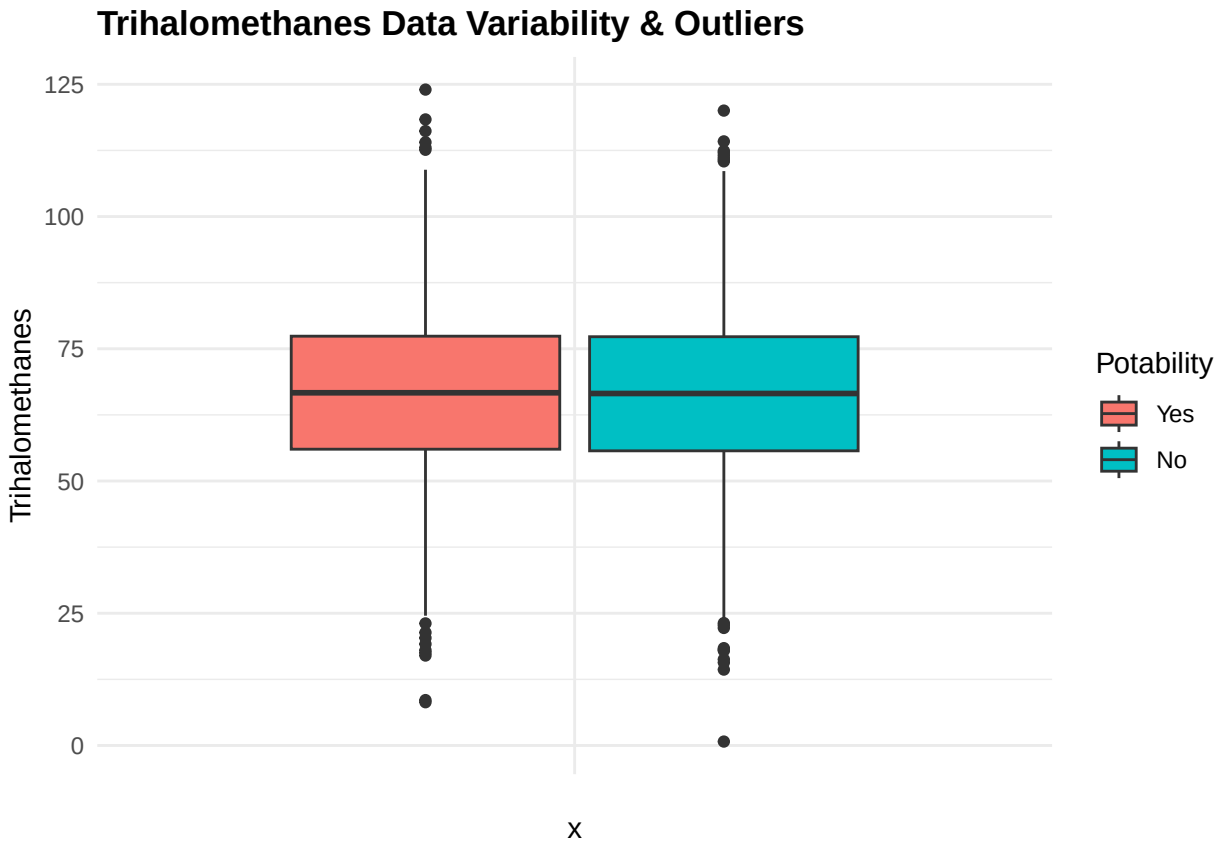

Solids Data Variability & Outliers
60000
40000
20000
0
Solids
x
Potability
Yes
No

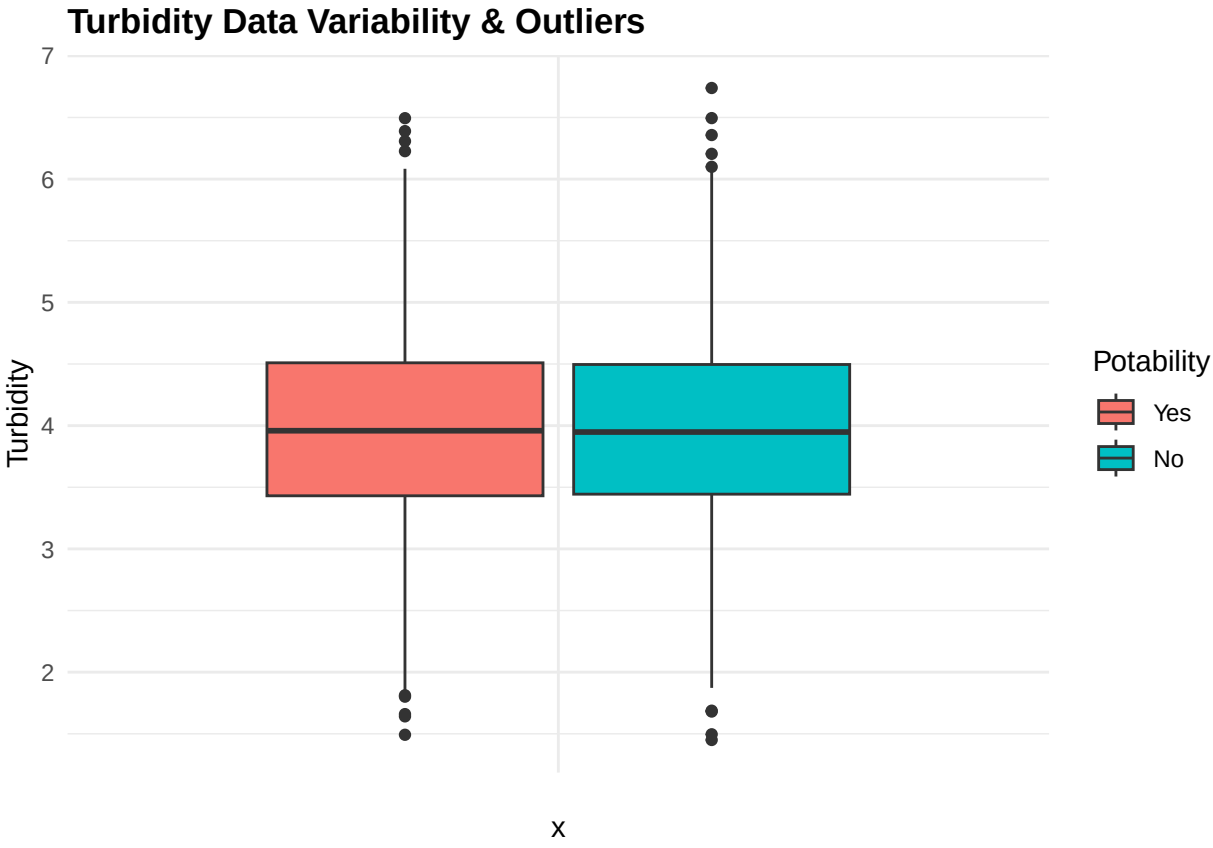

Chloramines Data Variability & Outliers
10
5
0
Chloramines
x
Potability
Yes
No


Sulfate Data Variability & Outliers
Sulfate
400
300
200
x
Potability
Yes
No


Conductivity Data Variability & Outliers
Conductivity
600
400
200
x
Potability
Yes
No


Organic_carbon Data Variability & Outliers
Organic_carbon
20
10
x
Potability
Yes
No


Trihalomethanes Data Variability & Outliers
Trihalomethanes
125
100
75
50
25
0
x
Potability
Yes
No

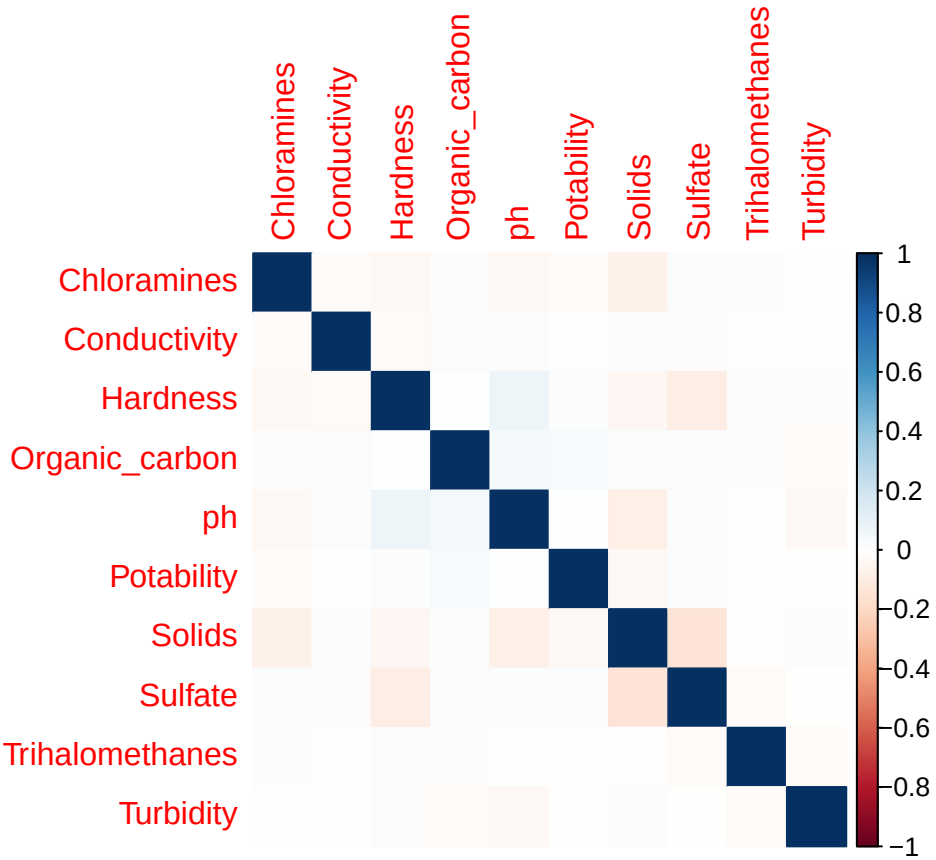


Data Cleaning and Pre-processing

Initially, as indicated in the dataset description, our dependent variable `potability` is an integer. To perform classification modelling, it is necessary to convert this variable into a factor. Secondly, the dataset's missing values must be addressed. Given the normal distribution of the dataset, the mean for each variable was employed to impute values for the missing data, with the exception of the `Solids` variable, which shows a significant deviation from the mean. The median was employed for the `solids` variable.

```
##        ph            Hardness          Solids         Chloramines
##  Min.   : 0.000   Min.   : 47.43   Min.   :  320.9   Min.   : 0.352
##  1st Qu.: 6.278   1st Qu.:176.85   1st Qu.:15666.7   1st Qu.: 6.127
##  Median : 7.000   Median :196.97   Median :20927.8   Median : 7.130
##  Mean   : 7.069   Mean   :196.37   Mean   :22014.1   Mean   : 7.122
##  3rd Qu.: 7.870   3rd Qu.:216.67   3rd Qu.:27332.8   3rd Qu.: 8.115
##  Max.   :14.000   Max.   :323.12   Max.   :61227.2   Max.   :13.127
##     Sulfate       Conductivity   Organic_carbon  Trihalomethanes
##  Min.   :129.0   Min.   :181.5   Min.   : 2.20   Min.   :  0.738
##  1st Qu.:317.1   1st Qu.:365.7   1st Qu.:12.07   1st Qu.: 56.648
##  Median :334.0   Median :421.9   Median :14.22   Median : 66.000
##  Mean   :333.8   Mean   :426.2   Mean   :14.28   Mean   : 66.377
##  3rd Qu.:350.4   3rd Qu.:481.8   3rd Qu.:16.56   3rd Qu.: 76.667
##  Max.   :481.0   Max.   :753.3   Max.   :28.30   Max.   :124.000
##    Turbidity     Potability
##  Min.   :1.450   Yes:1278
##  1st Qu.:3.440   No :1998
##  Median :3.955
##  Mean   :3.967
```

```
##  3rd Qu.:4.500
##  Max.   :6.739
```



The correlation matrix above clearly indicates the absence of correlation among the variables.

Modelling Approaches

For this project, Random Forest and K Nearest Neighbors were used to predict water potability. These models take into account that the dependent variable be a factor with more than one level (Han et al., 2022). According to (Breiman, 2001) is a classification and regression model that uses the decision tree model approach to create a forest consisting of multiple decision trees. K Nearest Neighbors is also a classification machine learning model that classifies a data point by determining the predominant class among its k nearest neighbours in the feature space (Cover & Hart, 1967).

Data Split

The caret package was utilised to partition the data according to the 80/20 rule, allocating 80% for training and 20% for testing. Set.seed is used to ensure the reproducibility of the model.

```
set.seed(123)
sample <- createDataPartition(water_quality$Potability, p = 0.8, list = FALSE)
train_data <- water_quality[sample, ]
test_data <- water_quality[-sample, ]
```

Random Forest Model

```
train_control <- trainControl(method = "cv", number = 8)
rf_model <- train(Potability ~ ., data = train_data, method = "rf", trControl = train_control) #The mod
```

```
rf_predict <- predict(rf_model, test_data) #Predicting the test data
test_data$Potability_pred <- rf_predict
```

KNN Model

```
train_control2 <- trainControl(method = "cv", number = 10)
knn_model <- train(Potability ~ ., data = train_data, method = "knn", trControl = train_control2) #The
knn_predict <- predict(knn_model, test_data) #Predicting the test data
test_data$Potability_pred2 <- knn_predict
```

Results

According to Han et al. (2022), when dealing with nominal dependent variables, evaluation metrics such as accuracy, precision, recall, and F1-Score should be utilised instead of the Root Mean Squared Error (RMSE). Given that the dependent variable `potability` is nominal, the following evaluation metrics will be employed.

Random Forest Model Confusion matrix The rows in the confusion matrix below represent actual labels and the column represent predicted label. So confusion matrix output below indicates that; **True Positives (TP):** Correctly predicted 92 observation as potable water. **True Negatives (TN):** Correctly predicted 348 observation as non-potable water. **False Positives (FP):** Incorrectly predicted 163 observations as potable water while they are non-potable. **False Negatives (FN):** Incorrectly predicted 51 observations as non-potable water while they are potable.

```
rf_confusionMatrix <- table(test_data$Potability, test_data$Potability_pred)
rf_confusionMatrix
```

```
##
##       Yes  No
##   Yes  92 163
##   No   51 348
```

Random Forest Model Accuracy

According to (Witten et al., 2016), accuracy is the proportion of correctly predicted instances to the total number of instances.For this random forest model, the model is correct by 67%.

```
rf_classification_accuracy <- sum(diag(rf_confusionMatrix)/sum(rf_confusionMatrix))
rf_classification_accuracy
```

```
## [1] 0.6727829
```

The code below extracts the elements of the confusion matrix

```
#Elements of confusion matrix
rf_TN <- rf_confusionMatrix [1,1]
rf_FP <- rf_confusionMatrix [1,2]
rf_FN <- rf_confusionMatrix [2,1]
rf_TP <- rf_confusionMatrix [2,2]
```

Random Forest Model Precision Evaluation

Precision denotes the proportion of accurately predicted positive cases (Witten et al., 2016). From the output of the precision calculation, it shows that 68% of the potable predicted observations are indeed potable water.

```
rf_precision <- rf_TP/(rf_TP+rf_FP)
rf_precision
```

```
## [1] 0.6810176
```

Random Forest Model Recall Evaluation

Recall, or sensitivity, measures the model's ability to accurately identify all positive instances present in the dataset (Witten et al., 2016). The model captures 87% of actual potable observations.

```
rf_recall <- rf_TP/(rf_TP+rf_FN)
rf_recall
```

```
## [1] 0.8721805
```

Random Forest Model F1 Score Evaluation

The F1 score combines metrics that incorporate precision and recall, providing a comprehensive evaluation of model performance (Witten et al., 2016). The balance between Precision and Recall for `potablity` is good at 76%.

```
rf_f1_score <- 2*(rf_precision * rf_recall)/(rf_precision + rf_recall)
rf_f1_score
```

```
## [1] 0.7648352
```

KNN Confusion Matrix Model Confusion matrix

The rows in the confusion matrix below represent actual labels and the column represent predicted label. So confusion matrix output below indicates that; **True Positives (TP):** Correctly predicted 60 observation as potable water. **True Negatives (TN):** Correctly predicted 323 observation as non-potable water. **False Positives (FP):** Incorrectly predicted 195 observations as potable water while they are non-potable. **False Negatives (FN):** Incorrectly predicted 76 observations as non-potable water while they are potable.

```
knn_confusionMatrix <- table(test_data$Potability, test_data$Potability_pred2)
knn_confusionMatrix
```

```
##
##       Yes  No
##   Yes  60 195
##   No   76 323
```

KNN Model Accuracy

For this random forest model, the model is correct by 58%.

```
knn_classification_accuracy <- sum(diag(knn_confusionMatrix)/sum(knn_confusionMatrix))
knn_classification_accuracy
```

```
## [1] 0.5856269
```

The code below extracts the elements of the confusion matrix

```
knn_TN <- knn_confusionMatrix [1,1]
knn_FP <- knn_confusionMatrix [1,2]
knn_FN <- knn_confusionMatrix [2,1]
knn_TP <- knn_confusionMatrix [2,2]
```

Random Forest Model Precision Evaluation

From the output of the precision calculation, it shows that 62% of the potable predicted observations are indeed potable water.

```
knn_precision <- knn_TP/(knn_TP+knn_FP)
knn_precision
```

```
## [1] 0.6235521
```

KNN Model Recall Evaluation

The model captures 81% of actual potable observations.

```
knn_recall <- knn_TP/(knn_TP+knn_FN)
knn_recall
```

```
## [1] 0.8095238
```

KNN Model F1 Score Evaluation

The balance between Precision and Recall for `potablity` is excellent at 70%.

```
knn_f1_score <- 2*(knn_precision * knn_recall)/(knn_precision + knn_recall)
knn_f1_score
```

```
## [1] 0.7044711
```

Model Performance Summary

The evaluation metrics indicate that the Random Forest model performed better than the K-Nearest Neighbours (KNN) model in several evaluation parameters. The accuracy of the random forest was 67%, while the KNN model was 58%, which signified a reduction in overall classification errors. Furthermore, the Random Forest model precision of 68% compared to 62% for KNN, resulting in a reduced incidence of false positives, which is vital in contexts where erroneous positive predictions incur significant costs like the water potability which is crucial in the health of human beings. Regarding recall, Random Forest outperforms KNN with a rate of 87% versus 81%, indicating its superior capacity to accurately identify positive instances. This renders Random Forest appropriate where the reduction of false negatives is essential, such as in water potability. Lastly, Random Forest shows a superior F1 Score of 76% compared to 70%, indicating its enhanced balance between precision and recall.

Conclusion

The report outlines and discusses the process of predicting water quality using 9 water quality parameters. The dataset required data preprocessing and cleaning steps for it to be usable for our models. Values were imputed to resolve missing data issues using the mean and median which were calculated for each variable. From the evaluation results, Random Forest is a more resilient and dependable model, rendering it the optimal selection for this dataset. The results from this project can be used to identify huge datasets where water quality parameters were collected but experiments to find out if the water is potable or not were not done whereby cutting costs that would require laboratory experiments.

The major limitation for this project was the missing data. Though imputation was done to account for this, the values do not carry the true value.

This project also opens up future experiment/research ideas like developing a mobile or computer application which a user can use to input water quality parameters and the application provides water potability status in real time. Lastly, other machine learning models and Artificial Intelligence models such as deep learning can be experimented on this data to see if the performance can be improved.

References

1. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

2. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964
3. Han, J., Pei, J., & Tong, H. (2022). Data mining: Concepts and techniques. Morgan kaufmann.
4. Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. Acm Transactions on Interactive Intelligent Systems (Tiis), 5(4), 1–19.
5. Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. Information Sciences, 585, 609–629. https://doi.org/10.1016/j.ins.2021.11.036
6. Witten, I., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques.