**Water quality analysis**

**Phase 4: Development part 2**

Continue building the analysis by creating visualization libraries(e.g Matplotlib,seaborn)to create histogram,scatterplots, and correlation matrices.

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

**Feature Engineering**

Using pH readings, we may roughly determine the type of water.

0 7.1

1 3.7

2 8.1

3 8.3

4 9.1

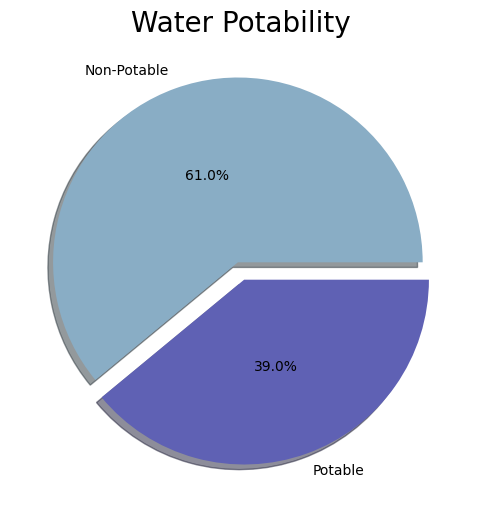
Name: ph, dtype: float64

**Data Visualization**

pH Level Chart



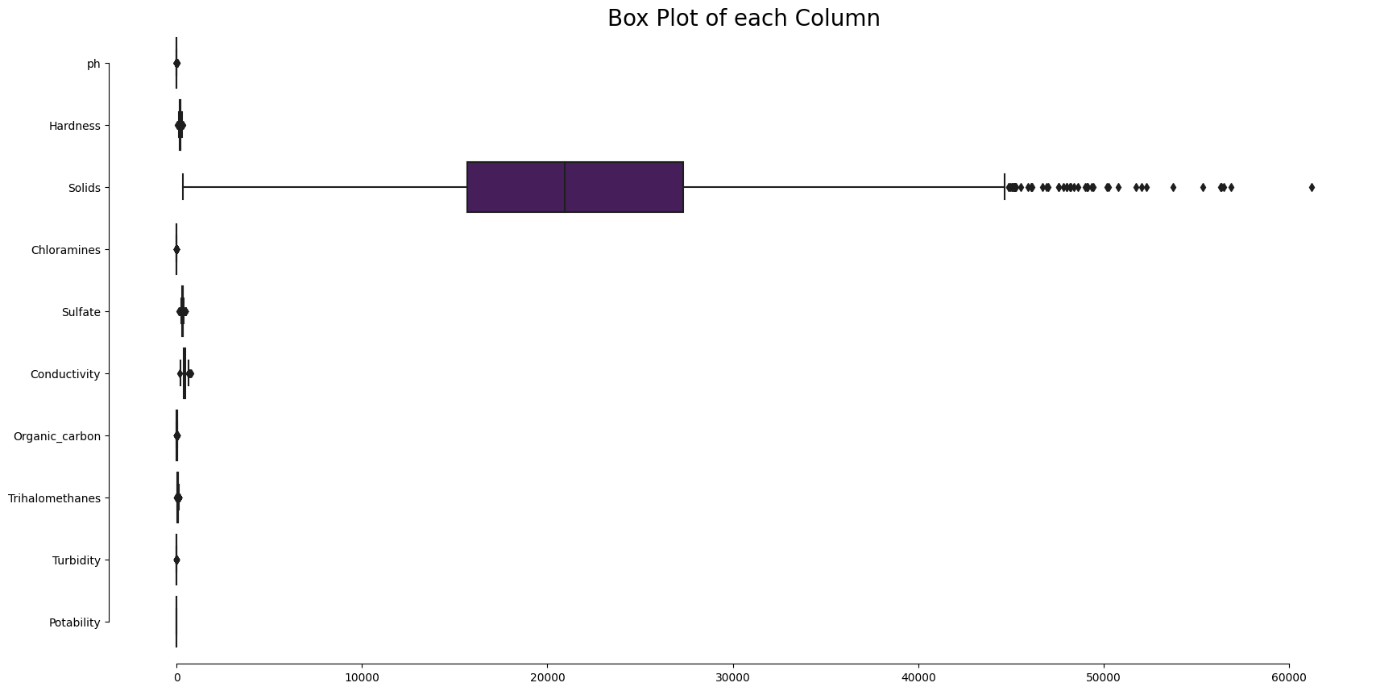
The different colors in this chart indicate the pH level of the given data and the correlation of another parameter.The pH readings are used to roughly determine the type of water, with values above 7 being alkaline and values below 7 being acidic.

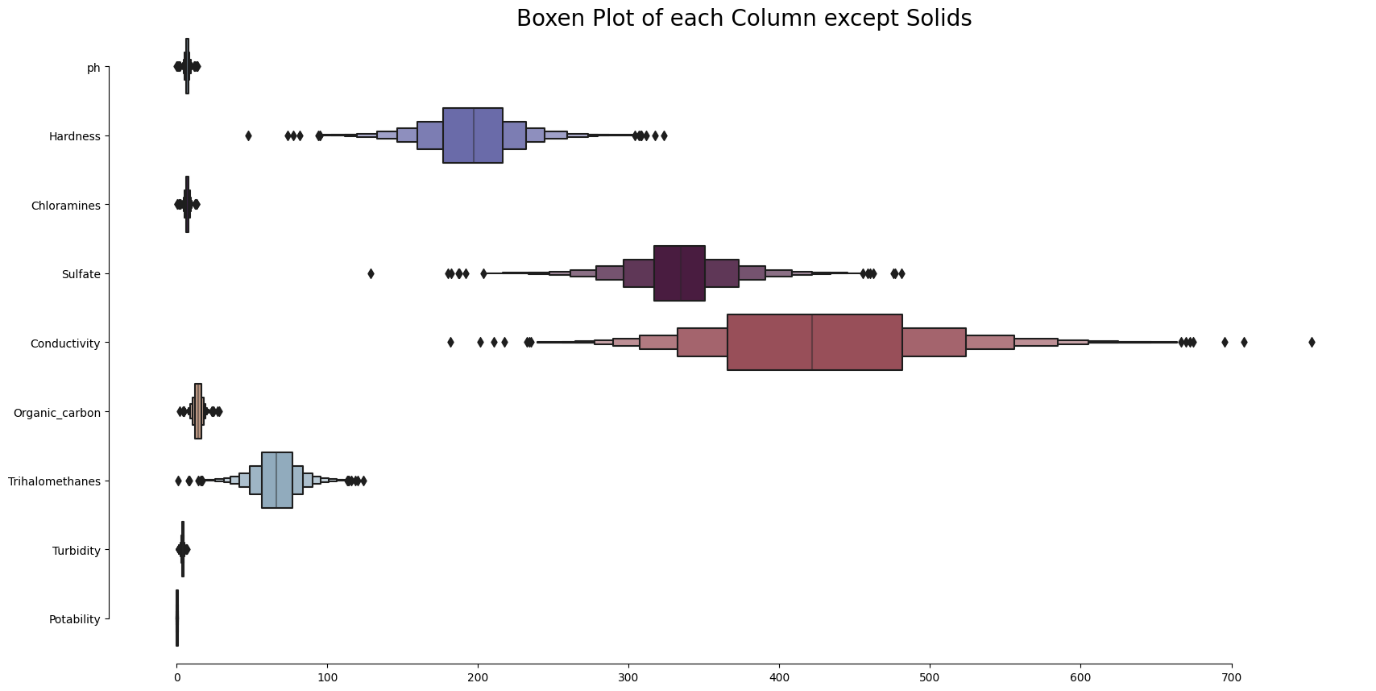


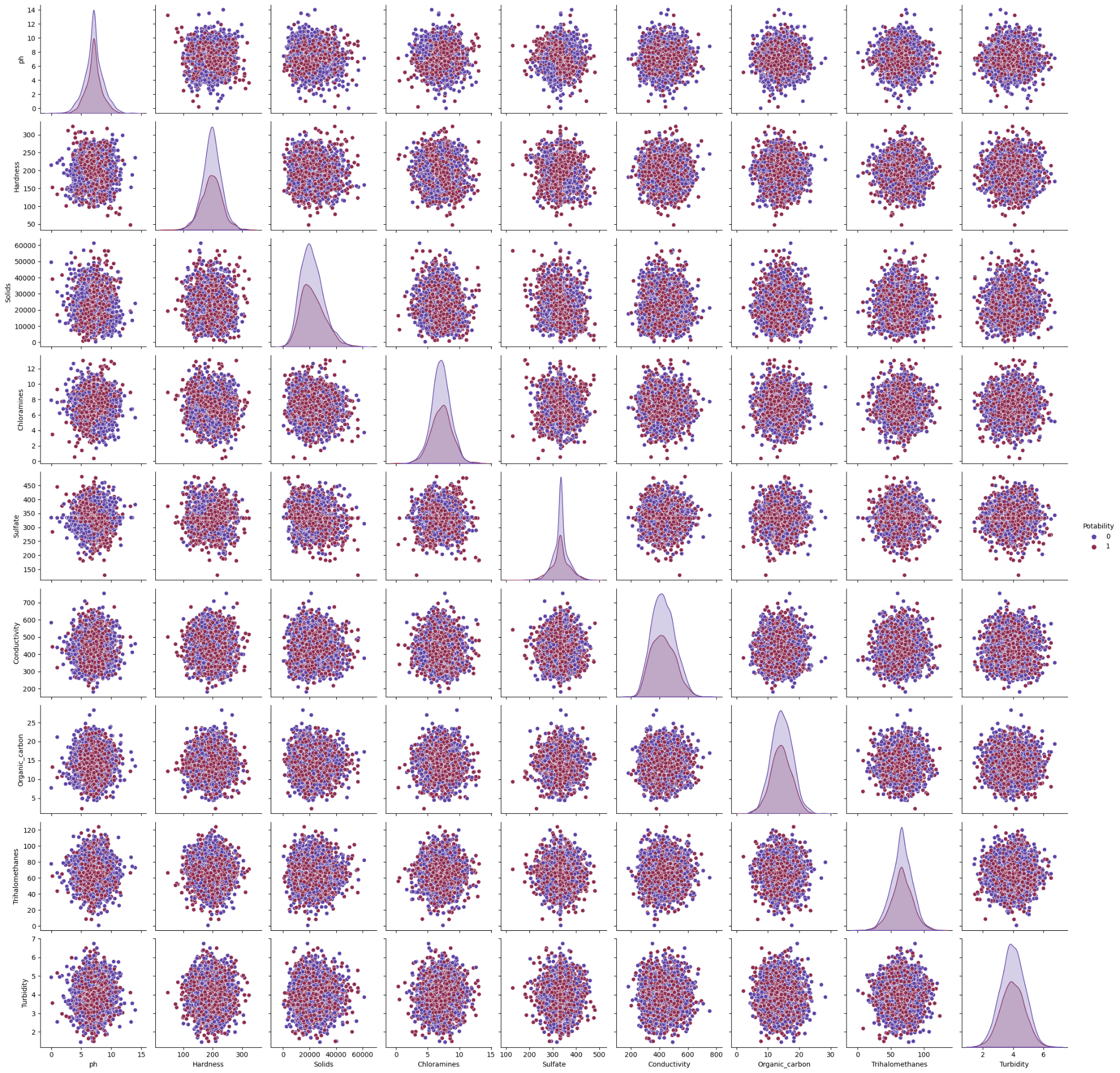
This chart provides insights into the total potability of the water samples. It shows that 61% of the water is non-potable, while 39% is potable

**Box plot for each column**

There is a box plot for each column in the dataset, which helps visualize the distribution, skewness, and potential outliers of the data



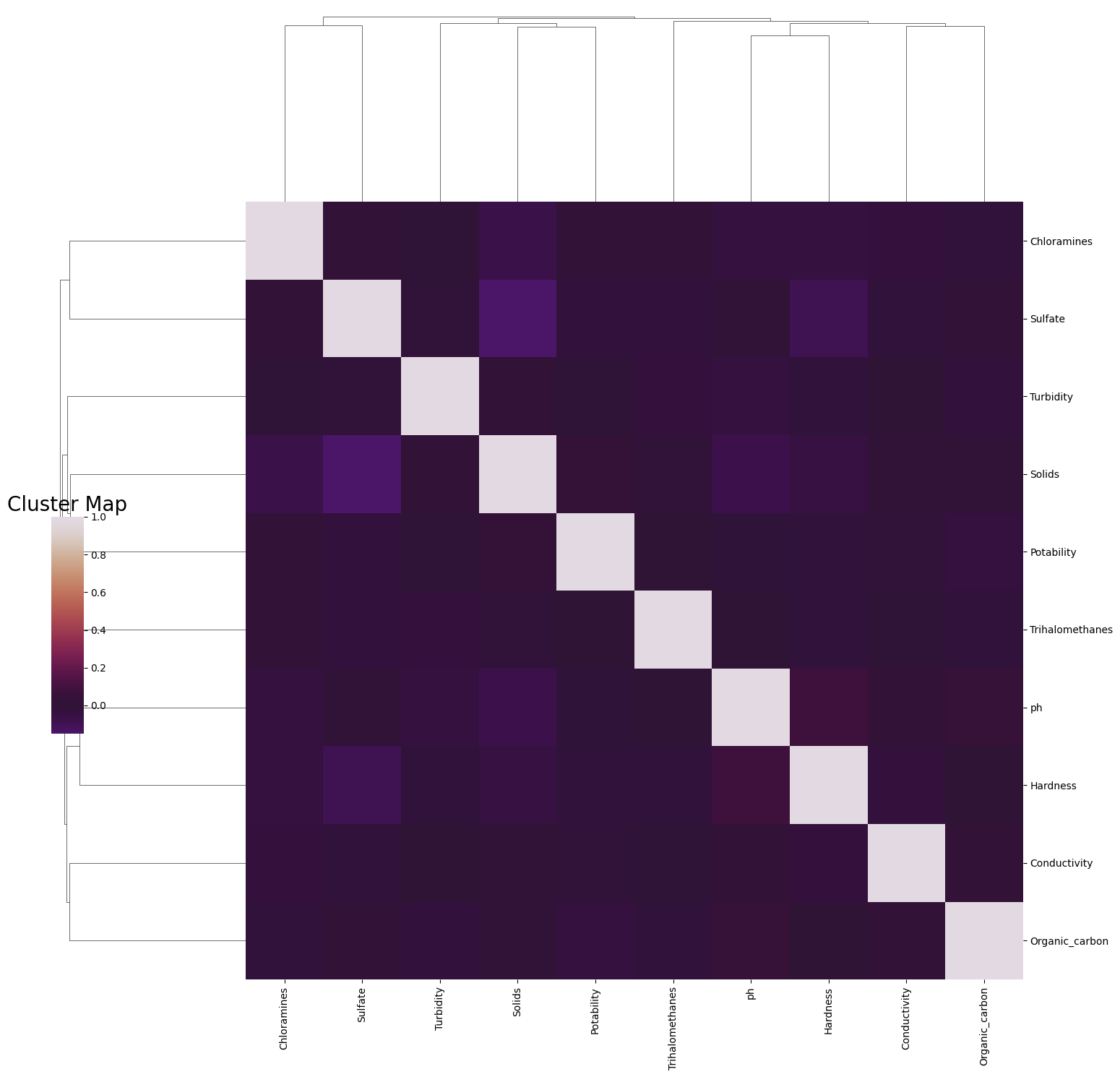
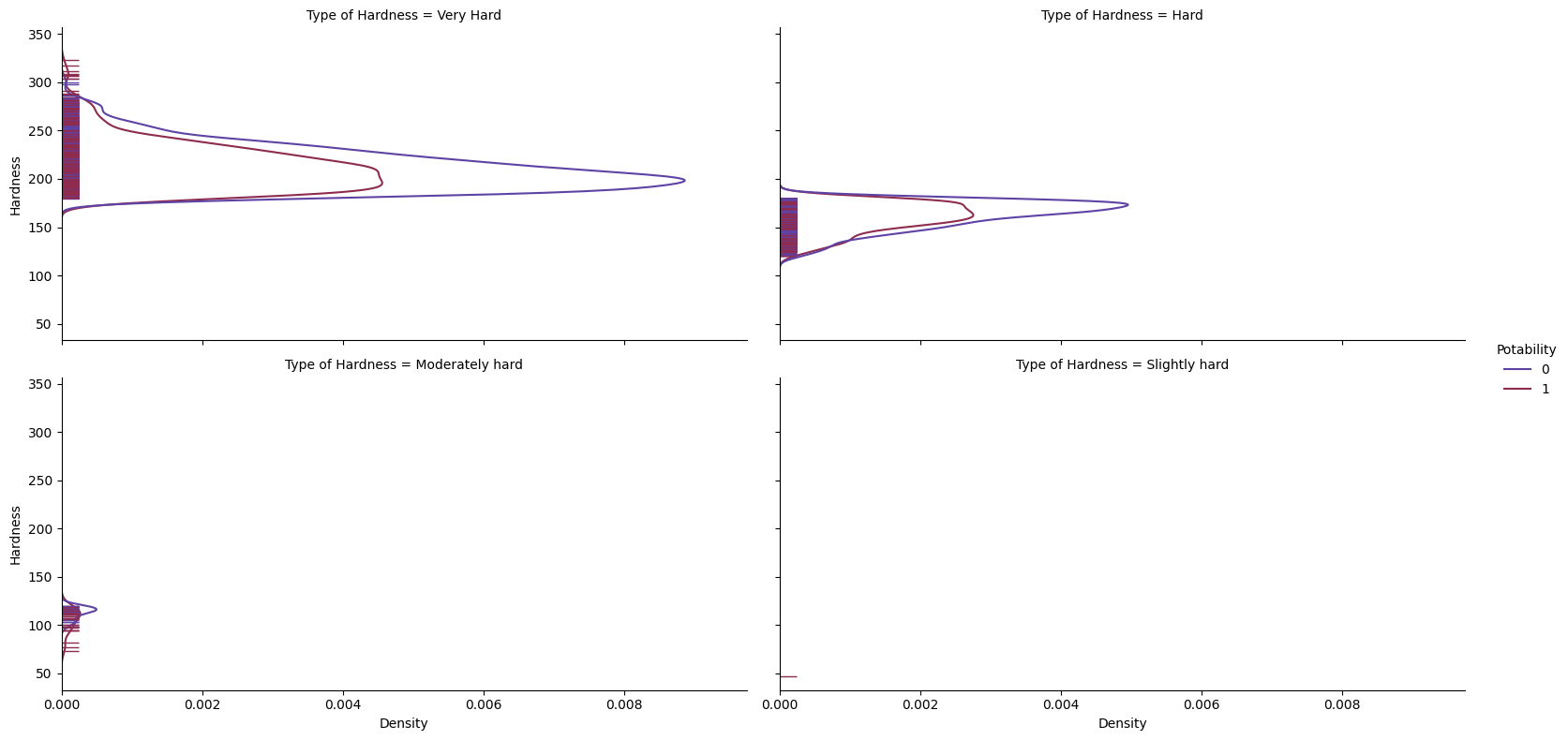




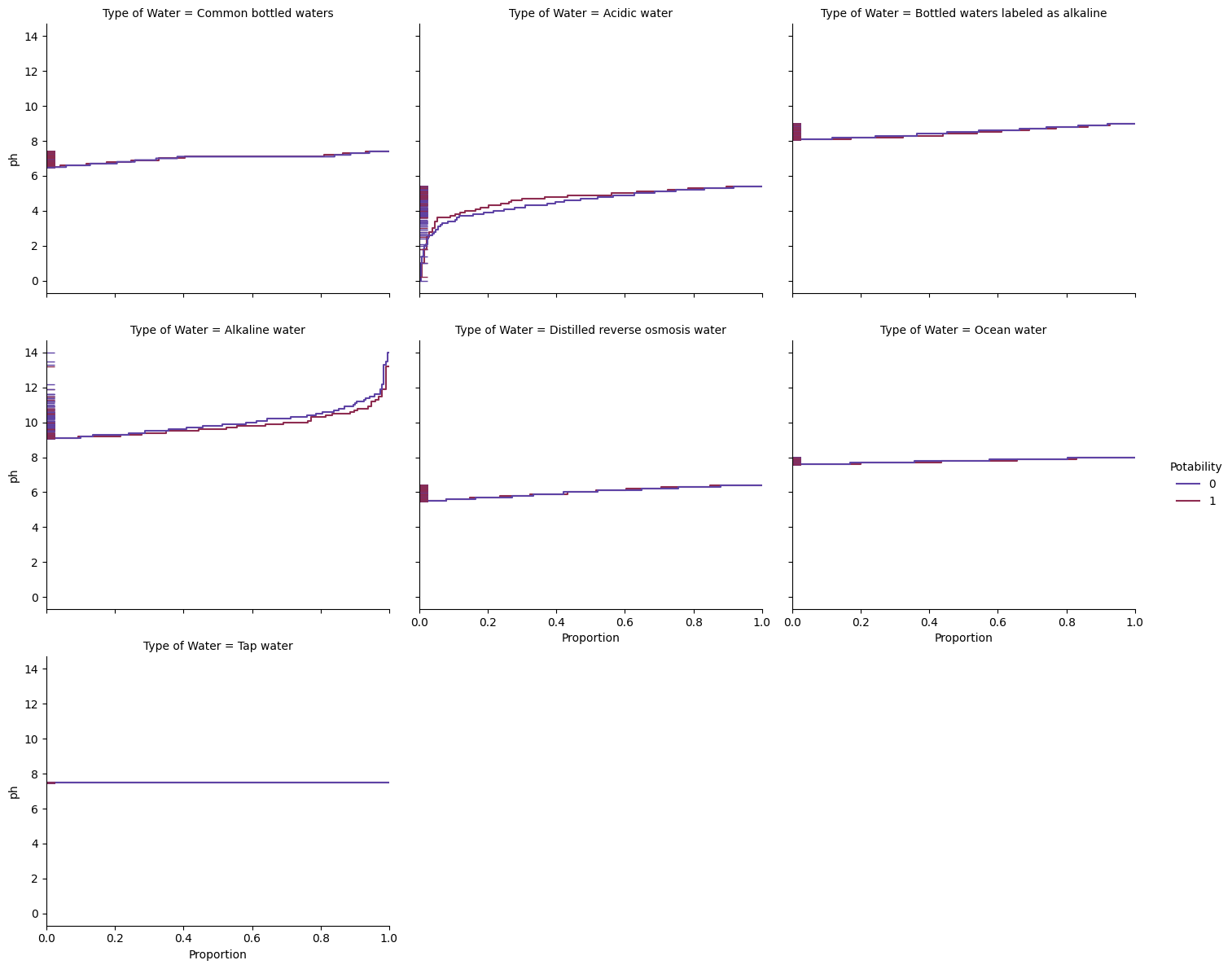
The pair plot shows the relationships between different variables in the dataset. It is a grid of scatterplots, where each variable is plotted against every other variable

**Cluster map**

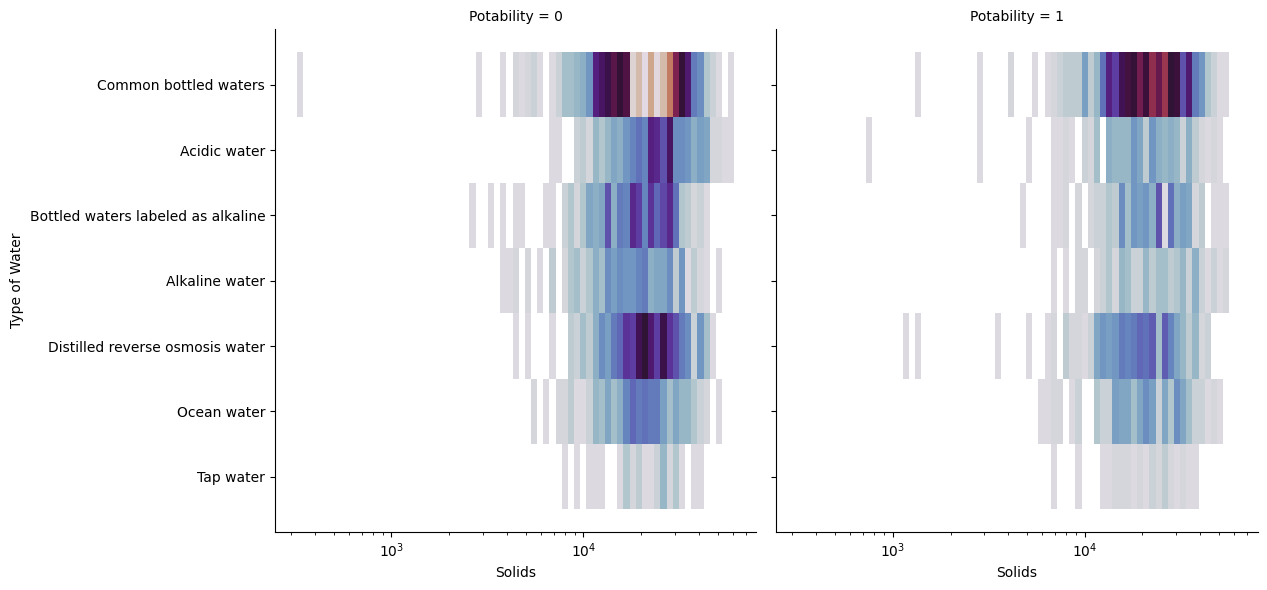
The cluster map visualizes the relationship between solids and potability. It uses a hierarchical clustering algorithm to group similar samples together based on their characteristics



There are several annotated heat maps that show the relationships between different parameters and potability. These heat maps use color to represent the strength and direction of the correlation between variables



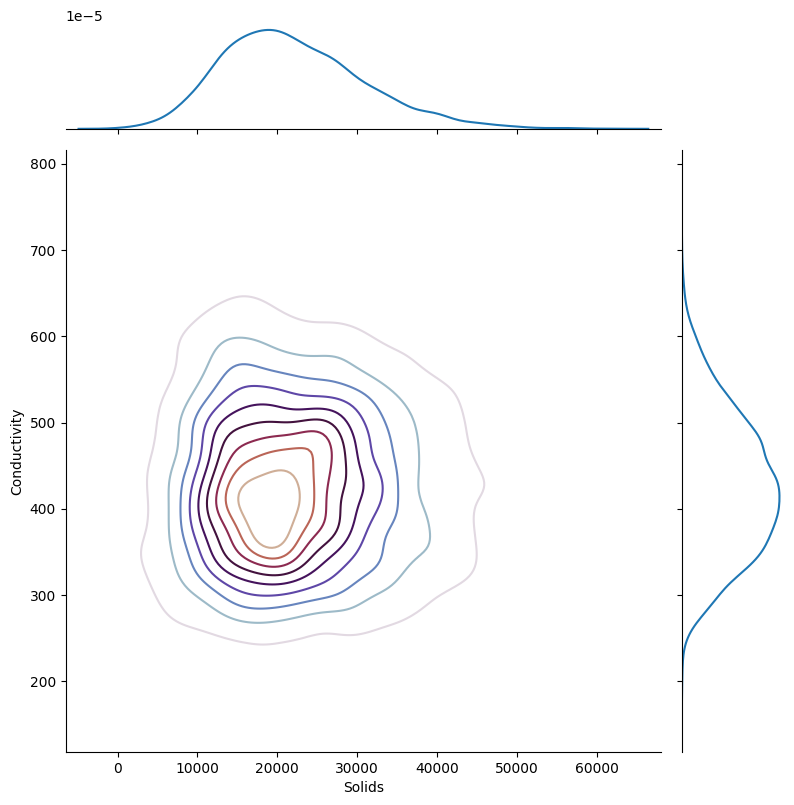
**Relation between solids and potability**



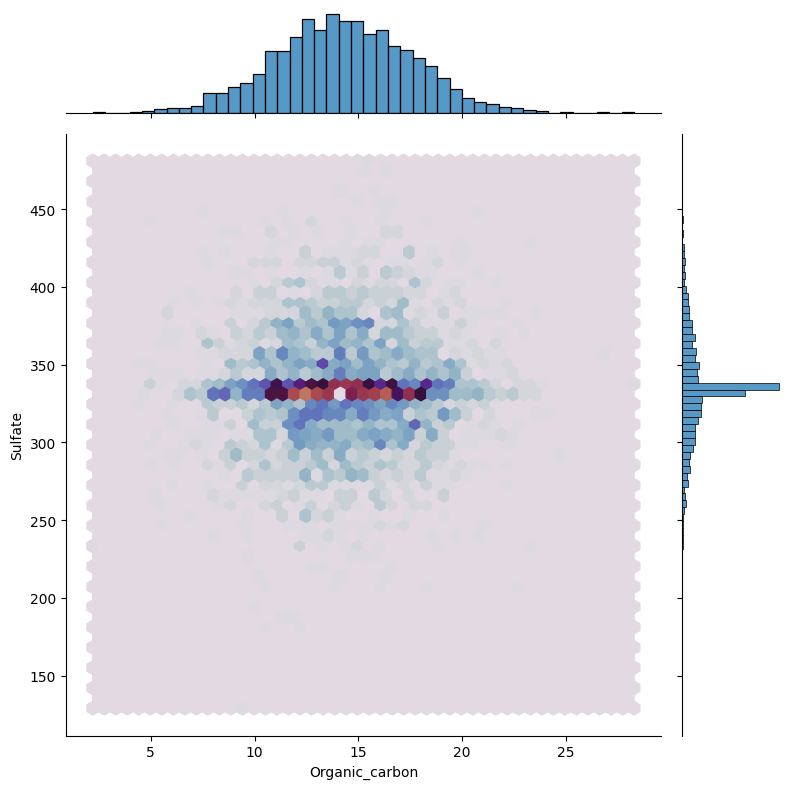
The cluster map visualizes the relationship between solids and potability. It uses a hierarchical clustering algorithm to group similar samples together based on their characteristics

**Relation between Solids and Conductivity**

One of the annotated heat maps shows the relationship between solids and conductivity. It uses color to represent the strength and direction of the correlation between these two variables

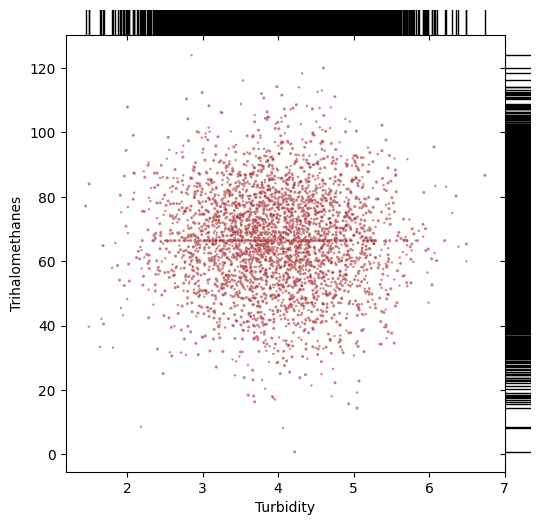


**Relation between Organic\_carbon and Sulfate**

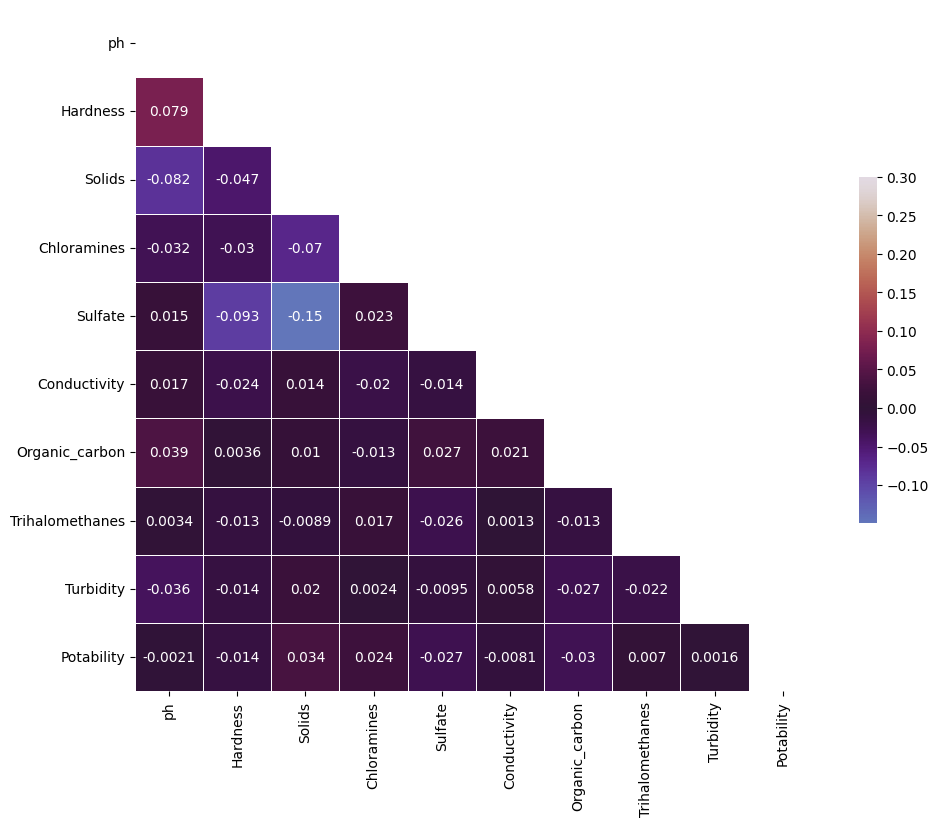


**Relation between Turbidity and Trihalomethanes**

A third annotated heat map shows the relationship between turbidity and trihalomethanes. It helps understand how these two parameters are related and their influence on water quality



**Annotated Heat Map**



**Models**

The analysis includes the use of various predictive models to determine water potability based on the water quality parameters. The model scores indicate the accuracy of each model in predicting water potability. The models used in this analysis include Random Forest, Decision Tree, KNN, Support Vector Machines, Logistic Regression, and Naive Bayes**1**. The Random Forest and Decision Tree models have a score of 100%, while the KNN model has a score of 78.27%. The Support Vector Machines model has a score of 73.50%, the Logistic Regression model has a score of 62.64%, and the Naive Bayes model has a score of 61.54%. These scores can be used to

|  |  |  |
| --- | --- | --- |
|  | **Model** | **Score** |
| **3** | Random Forest | 100.00 |
| **5** | Decision Tree | 100.00 |
| **1** | KNN | 78.27 |
| **0** | Support Vector Machines | 73.50 |
| **2** | Logistic Regression | 62.64 |
| **4** | Naive Bayes | 61.54 |