Water Quality Analysis

Phase 2: Innovation

Consider exploring anomaly detection techniques to identify unusual patterns in water quality parameters

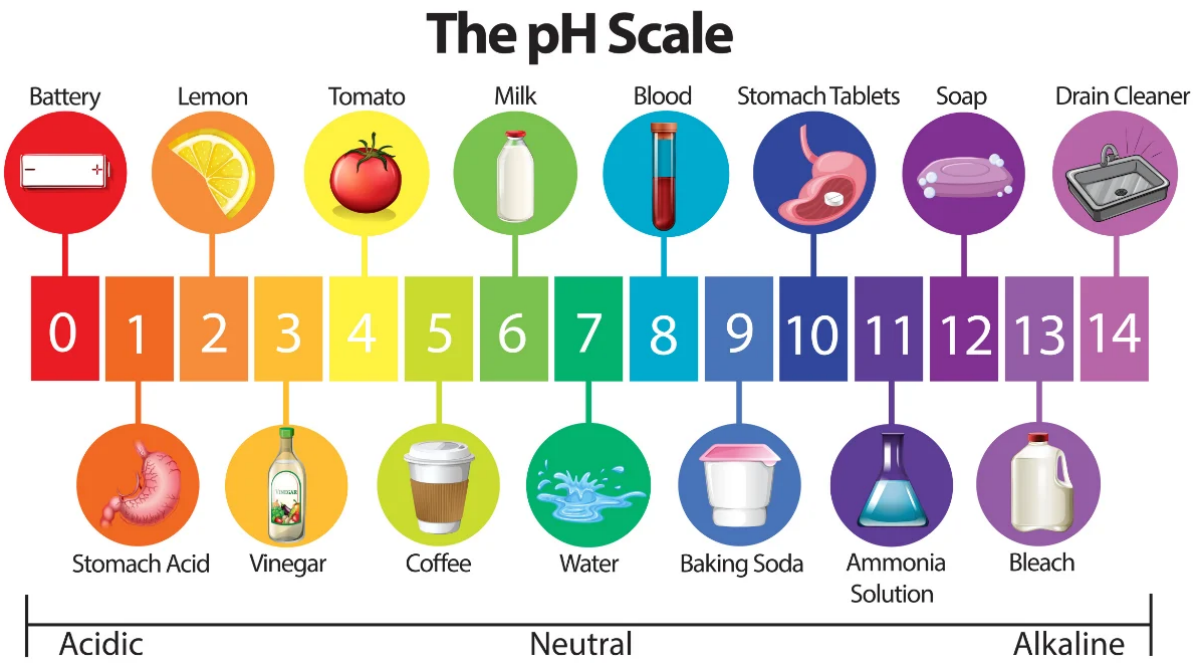
Anomaly Detection

dataset information

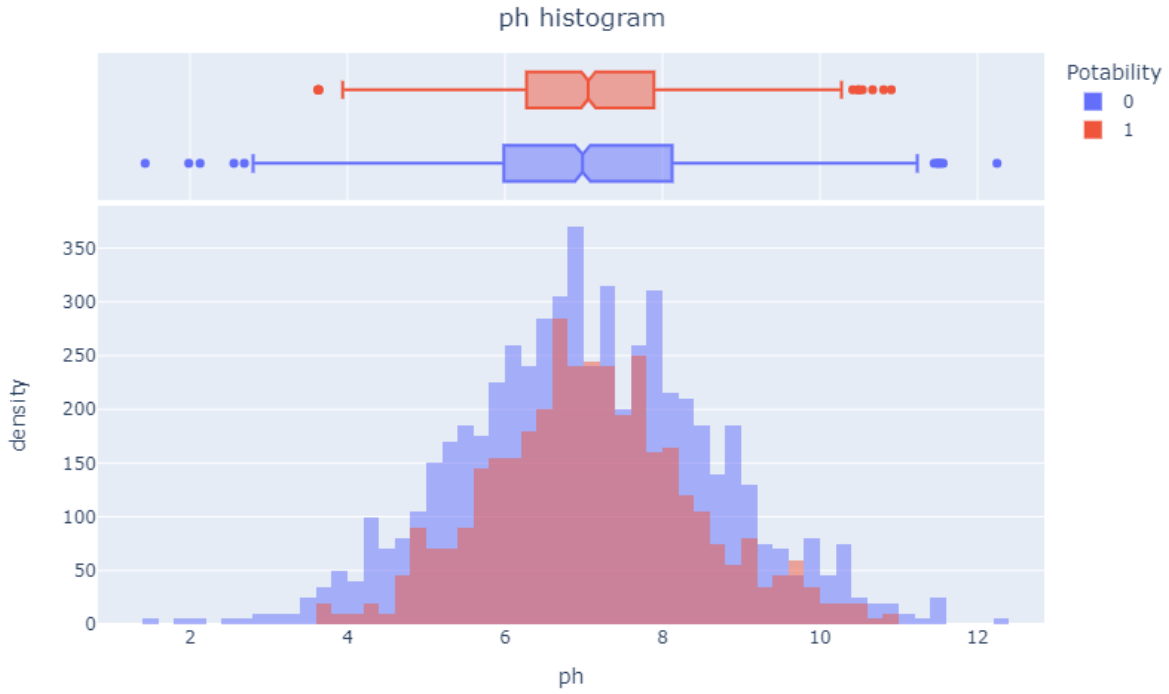
```
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
```

Histogram Based

The model seen above is a parametric one in the sense the we only estimate the parameters of the multivariate distribution, namely the mean and covariance matrix. Another option is to use a non-parametric statistical technique. Here, we consider using histograms to model the normal data profile. Bellow, we see 5 histograms for different features in our data, both for normal and anomalous instances.(each parameter applied this anomaly detection technique)



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 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.



Observation

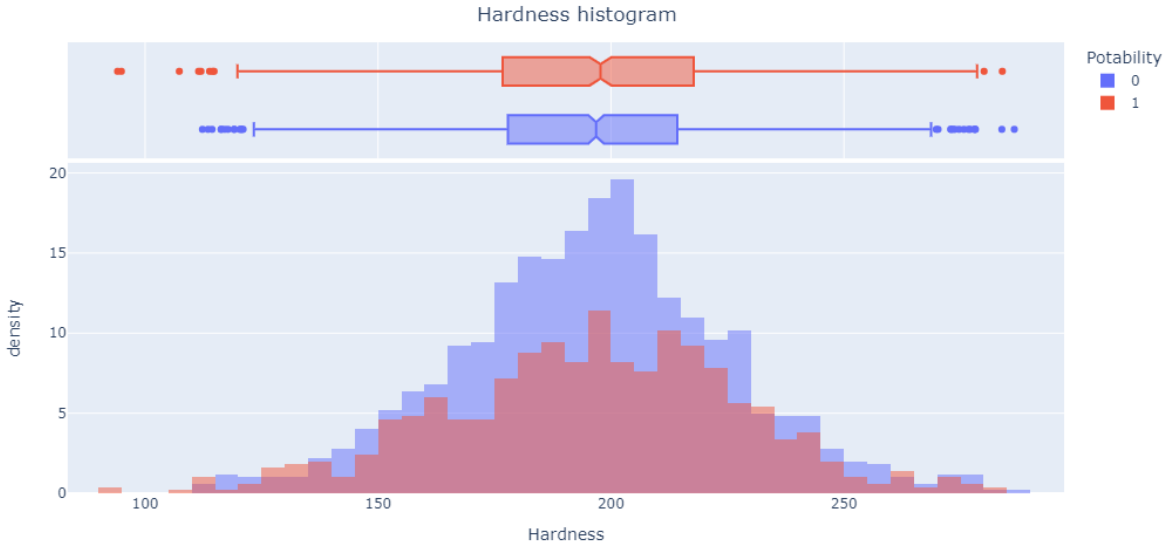
- If you look at the picture above, there are cases where it is judged that you can drink even if the pH is less than 6 or greater than 8.5.
- There is no significant difference in the distribution of water potability according to pH.

Derived Feature (is_ph_ok)

- We want to record the data included in the IQR of ph that is judged to be drinkable by creating a derived feature called is_ph_ok.
- Among the data, the pH values between 6.28 and 7.89 are included in the IQR.
-
- `0  108`
- `1  824`

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.



Observation:

- There is no significant difference in the distribution of water potability according to Hardness.

Derived Feature (is_Hardness_ok)

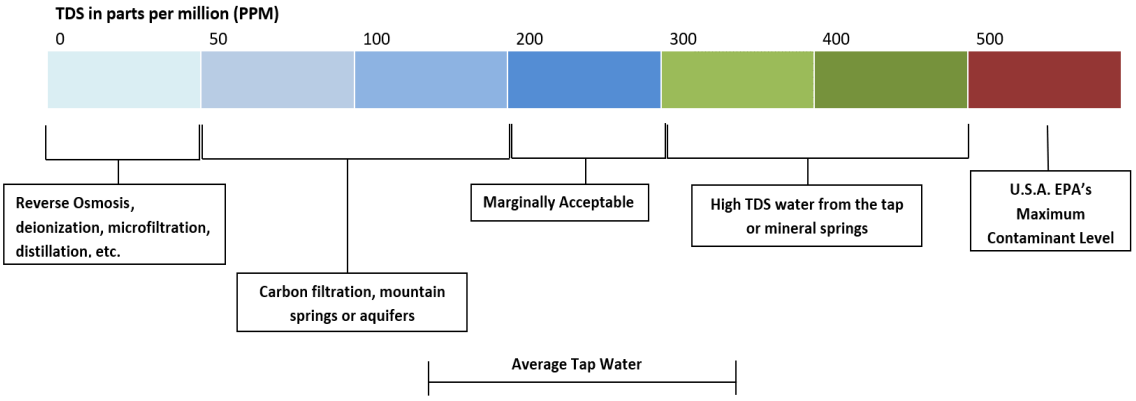
We want to record the data included in the IQR of Hardness that is judged to exist by creating a derived feature called is_Hardness_ok.

- Among the data, the Hardness values between 176.74 and 217.73 are included in the IQR.

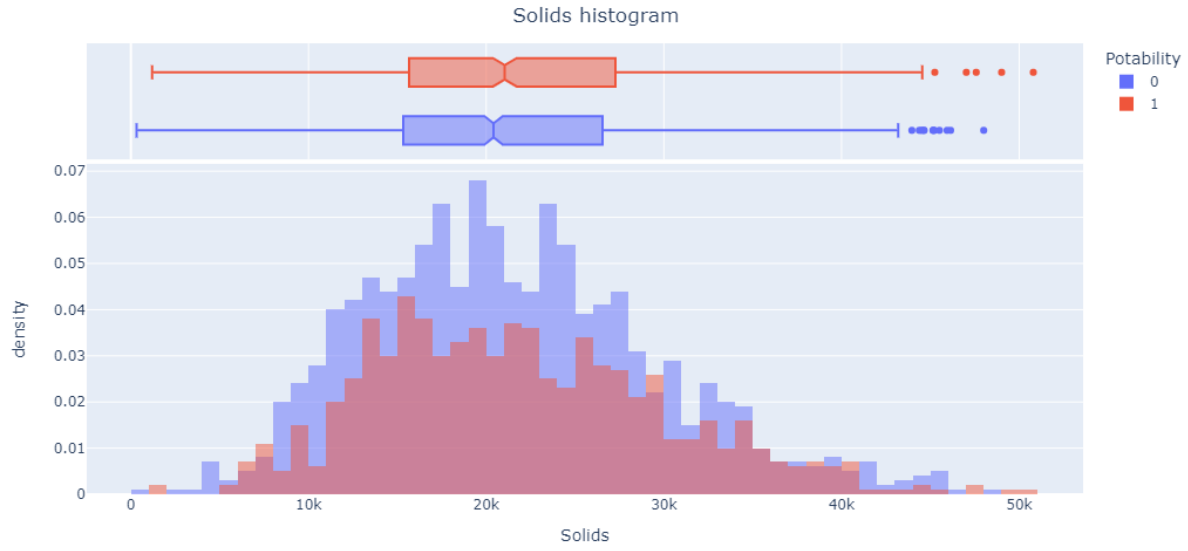
```
1    1013

0     897
```

Solids



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 un-wanted taste and diluted color in 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. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.



Observation:

- There is no significant difference in the distribution of water potability according to Solids.

Derived Feature (is_Solids_ok)

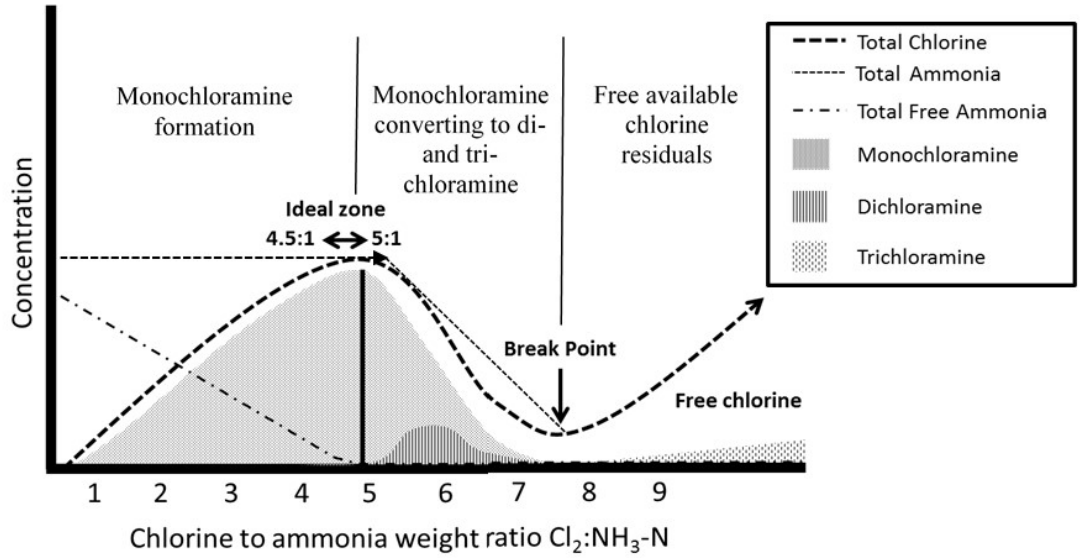
We want to record the data included in the IQR of Hardness that is judged to be drinkable by creating a derived feature called is_Solids_ok.

Among the data, the Solids values between 15665.10 and 27249.84 are included in the IQR

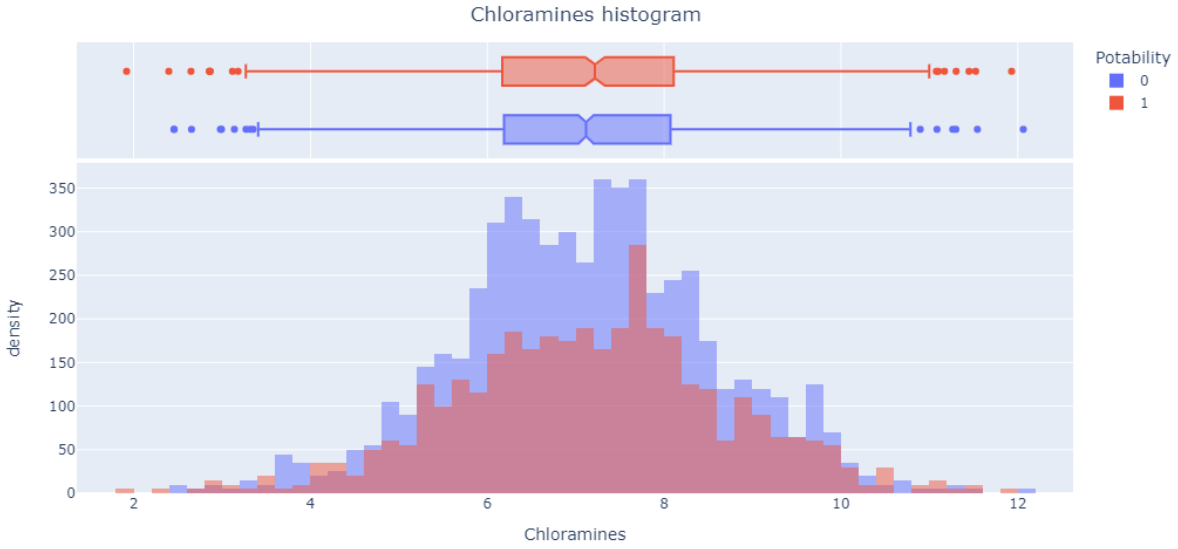
```
1    978

0    932
```

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.



Observation:

- There is no significant difference in the distribution of water potability according to Chloramines.

Derived Feature (is_Chloramines_ok)

We intend to record the data included in the IQR of Chloramines that are judged to be drinkable by creating a derived feature called is_Chloramines_ok.

- Among the data, the Chloramines values between 6.17 and 8.10 are included in the IQR.

```
1    970

0    940
```

Sulfate

TRACKING SULFATES IN WATER

This guide can help cattle producers keep on top of their water quality. Sulfate levels are measured in milligrams per litre.

GOOD

- **Diarrhea or refusal of water** by animals not accustomed to it
- 500 to 800 mg/L may **affect calves, including a trade mineral deficiency**
- Trace mineral deficiencies can cause **depressed growth rate, fertility and immune response**
- **Decreased performance** in feedlot cattle
- **1,000 mg/L recommended maximum** if feed level is high in sulfates or if ambient temperature is high

ACCEPTABLE

- **Laxative**
- **Performance reduced**
- High level of sulfates can also **contribute to copper and other trace mineral deficiencies**

POOR

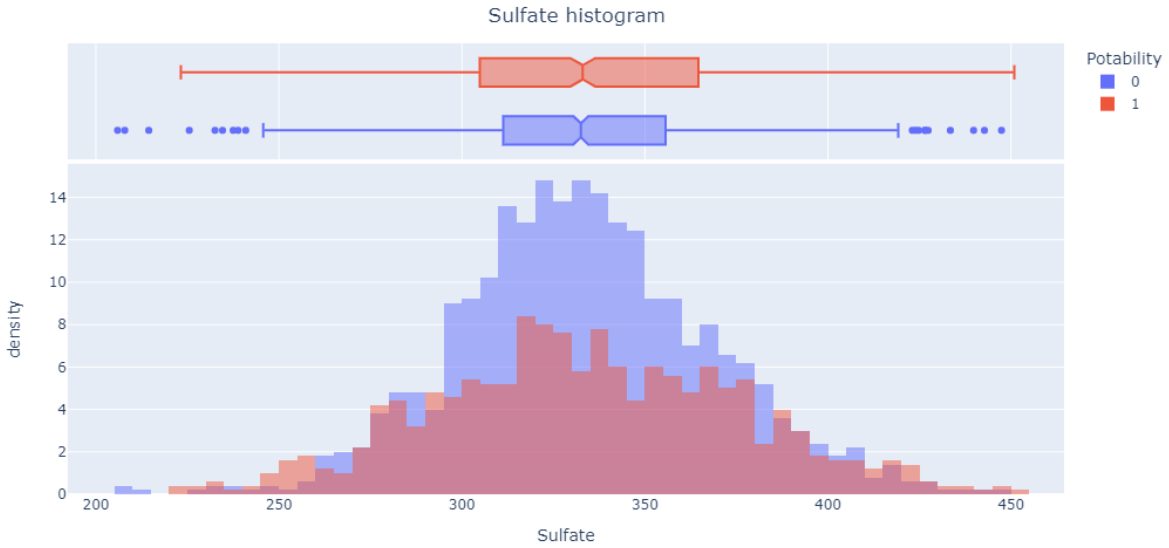
- **Chelated or Hydroxy minerals** may be required
- High chance of trace mineral deficiency
- Symptoms include: **decreased gains, depressed immunity and reduced conception, etc.**
- **Sporadic cases of polio** possible
- Can cause **diarrhea and reduced milk production** in dairy cows

UNSUITABLE

- **Sporadic cases of polio** are highly probable
- **Performance reduced**
- **Scours** likely
- Greater than 4,000 mg/L **dangerous health problems** expected

Source: beefresearch.ca, adapted from Saskatchewan Agriculture | RAELENE HOLTH GRAPHIC

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.



observation

Derived Feature (is_Sulfate_ok)

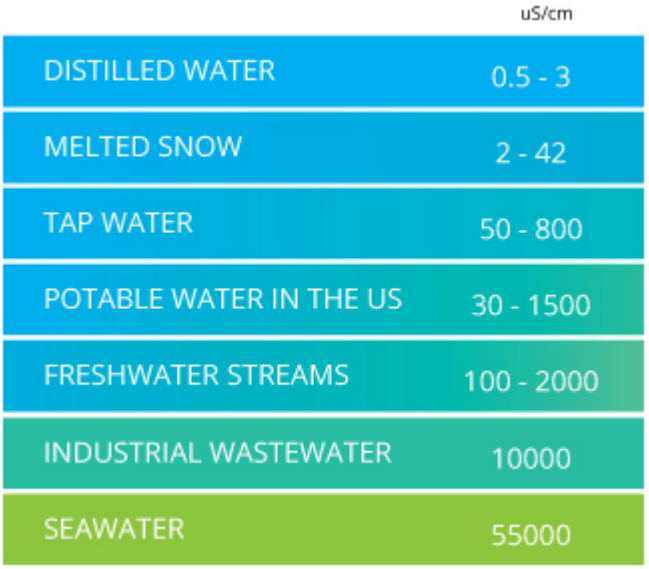
We want to record the data included in Sulfate's IQR, which is judged to be drinkable, by creating a derived feature called is_Sulfate_ok.

- Among the data, the Sulfate values between 304.97 and 364.57 are included in the IQR.

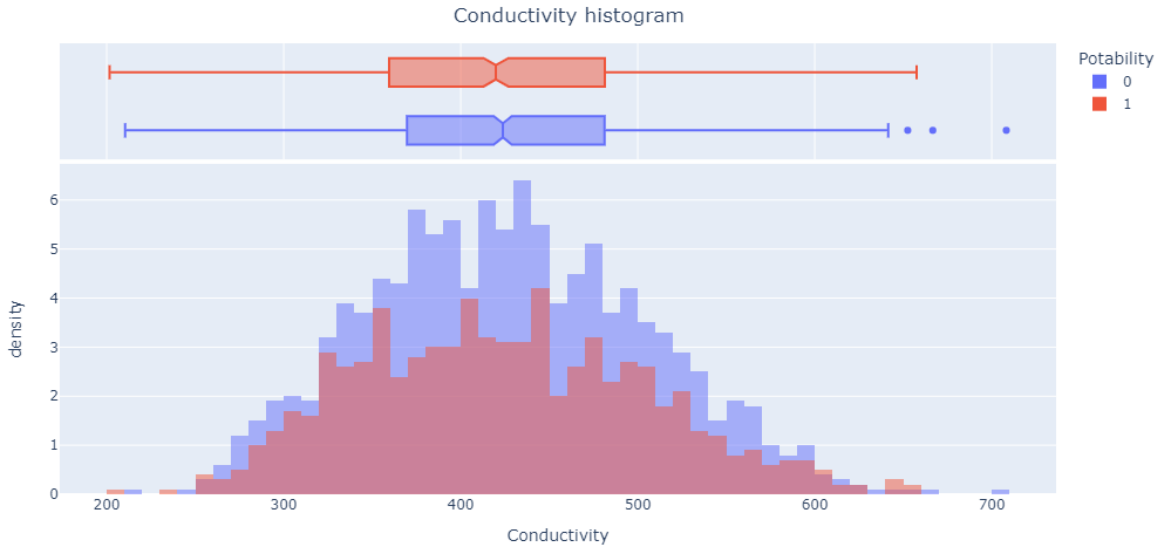
```
1     1094

0      816
```

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.



Observation

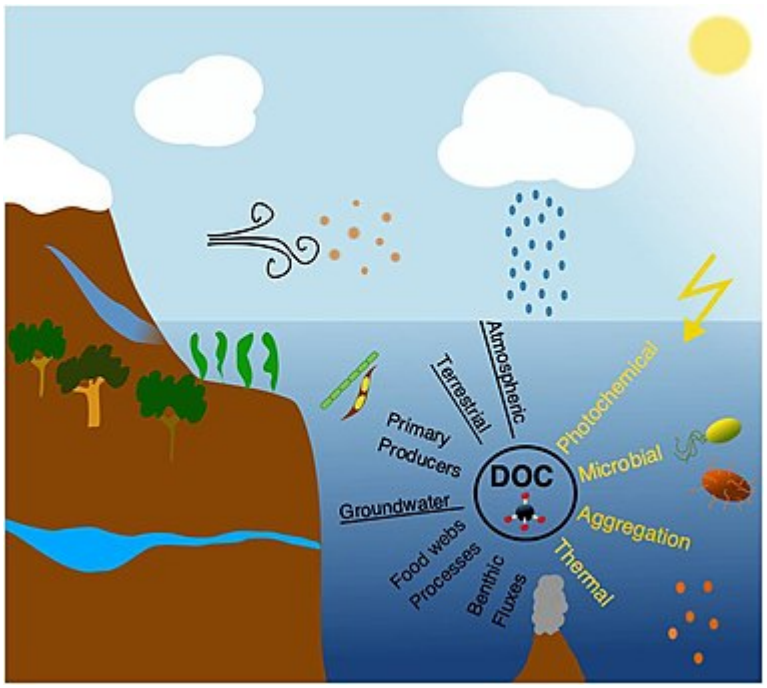
Derived Feature (is_Conductivity_ok)

We want to record the data included in the IQR of Conductivity that is judged to be drinkable by creating an is_Conductivity_ok derived feature.

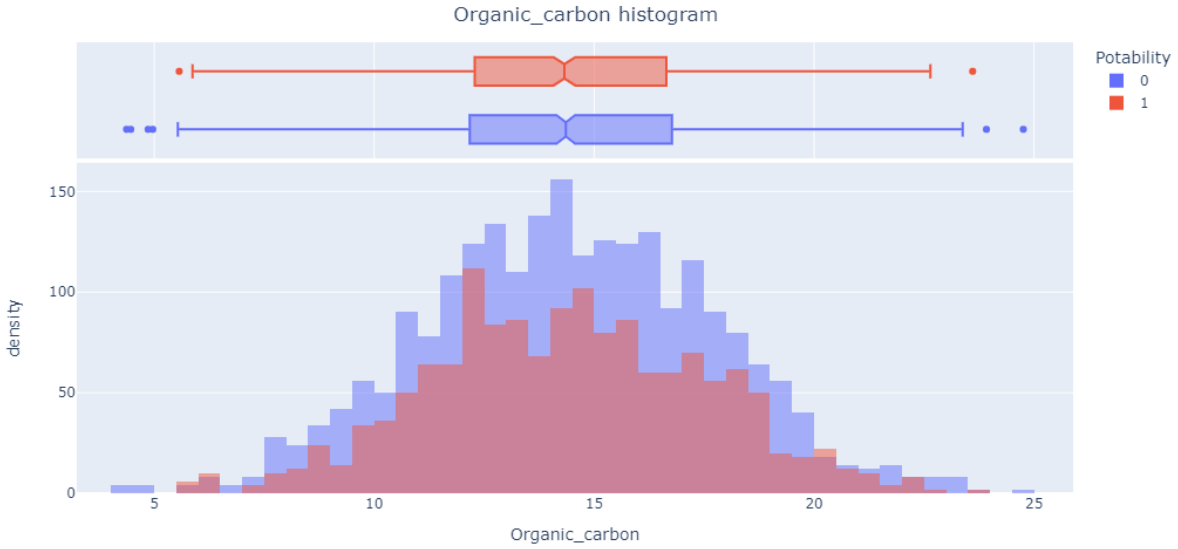
- Among the data, the Conductivity values between 359.51 and 481.41 are included in the IQR.

```
1    999
0    911
```

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 US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.



Observation

Derived Feature (is_Organic_carbon_ok)

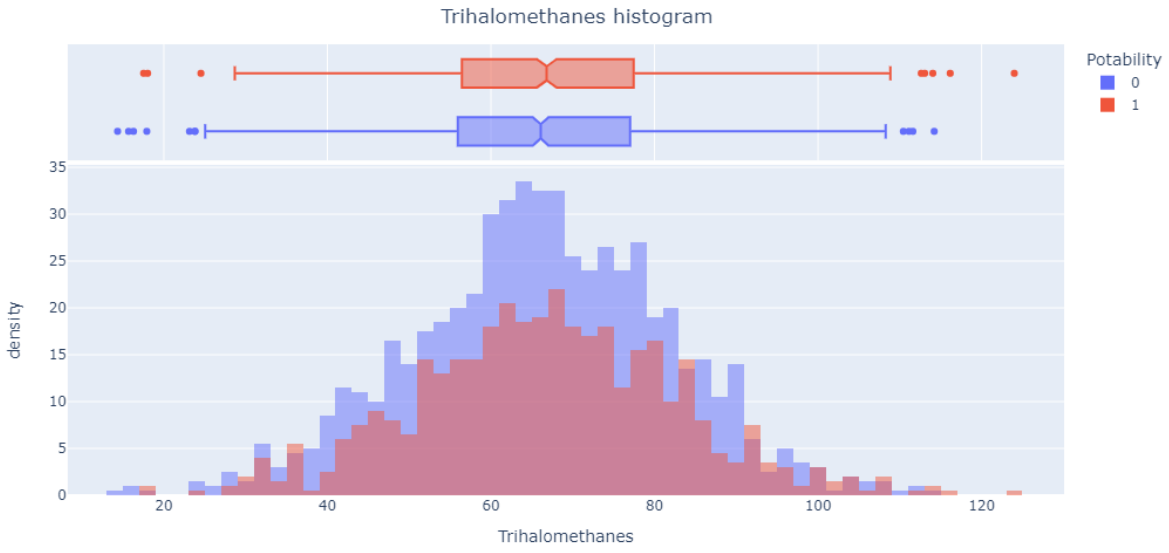
We want to record the data included in the IQR of Organic_carbon, which is judged to be drinkable, by creating a derived feature called is_Organic_carbon_ok.

- Among the data, the Organic_carbon values between 12.28 and 16.63 are included in the IQR.

```
0    981
1    929
```

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.



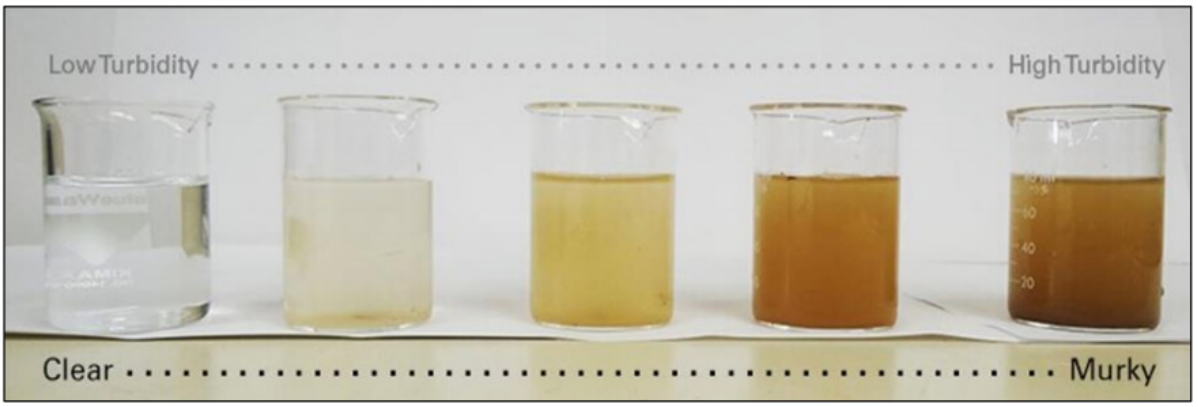
Observation

Derived Feature (is_Trihalomethanes_ok)

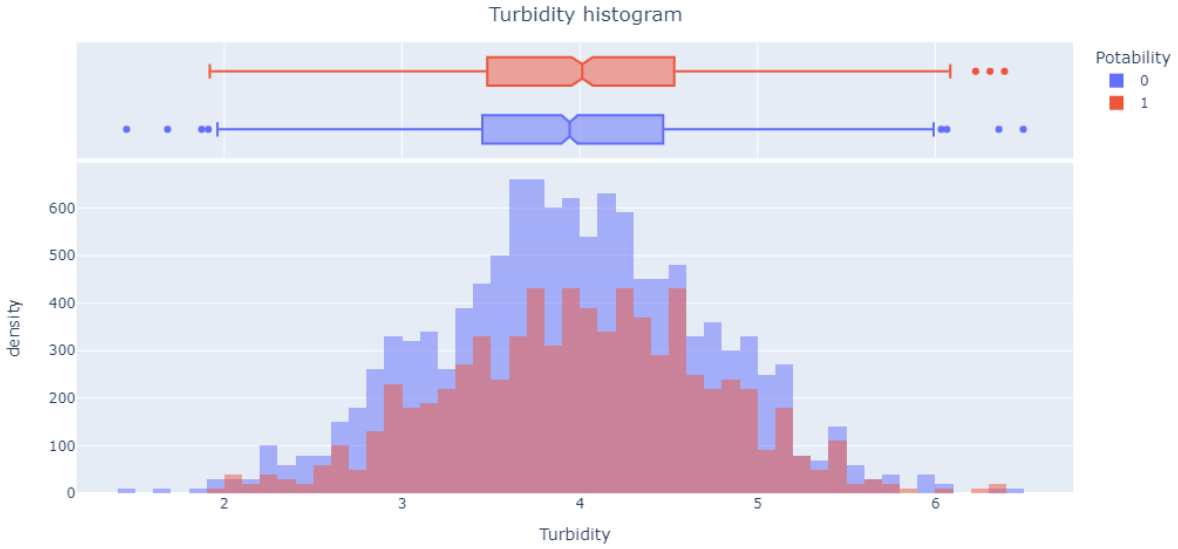
We want to record the data included in the IQR of Trihalomethanes that are judged to be drinkable by creating a derived feature called is_Trihalomethanes_ok.

- **Among the data, the Sulfate values between 56.45 and 77.41 are included in the IQR.**

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



Observation

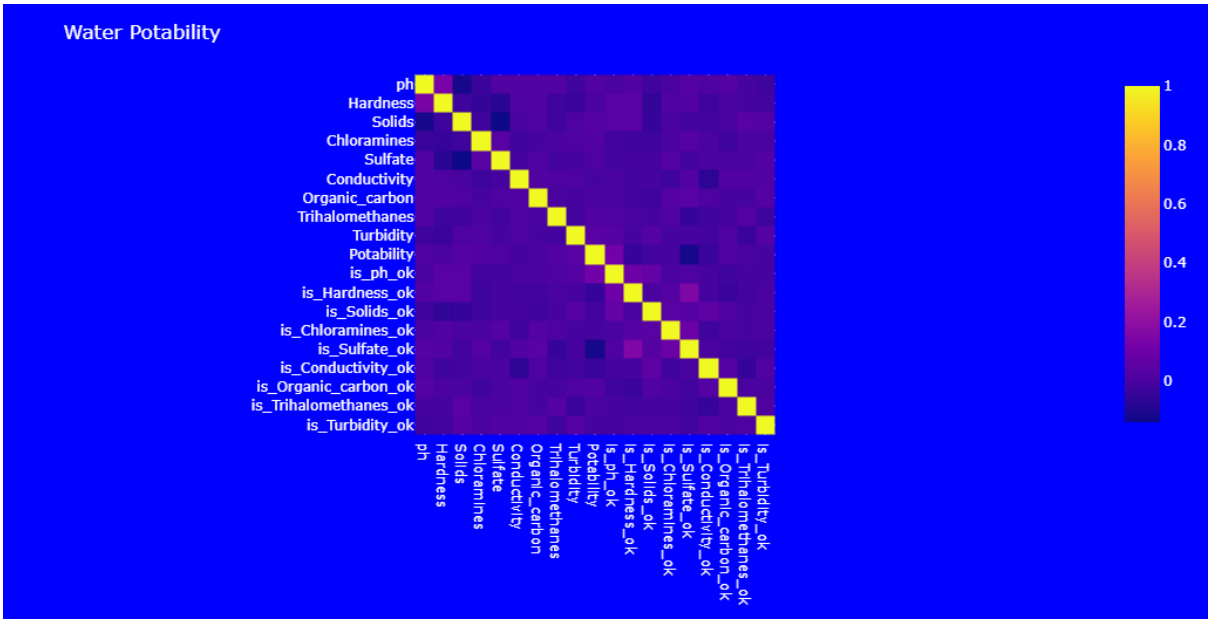
Derived Feature (is_Turbidity_ok)

We want to record the data included in the IQR of Turbidity, which is judged to be drinkable, by creating a derived feature called is_Turbidity_ok.

- Among the data, the Sulfate values between 3.48 and 4.53 are included in the IQR.

Potability

Potability: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.



Observation:

- The correlation coefficient of all features is low.
- Correlation coefficients of newly created derived features are relatively high.

Doing Anomaly Detection

We want to find outliers in the dataset through anomaly detection. When solving the water portability problem, it seems right to remove outliers.

Pycaret

PyCaret is a versatile library that can be used not only for traditional supervised machine learning tasks but also for anomaly detection. Anomaly detection involves identifying rare and unusual data points that deviate significantly from the norm. PyCaret makes it relatively easy to apply anomaly detection techniques to your datasets

PCA

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique and, when applied to anomaly detection, it becomes a powerful tool for identifying unusual patterns or outliers in datasets. In anomaly detection using PCA, the primary goal is to transform the data into a lower-dimensional space while retaining as much variance as possible. Anomalies can then be detected by identifying data points that deviate significantly from the expected distribution in this reduced space.

```
the size of anomaly = 101 # detection anomaly in dataset
```

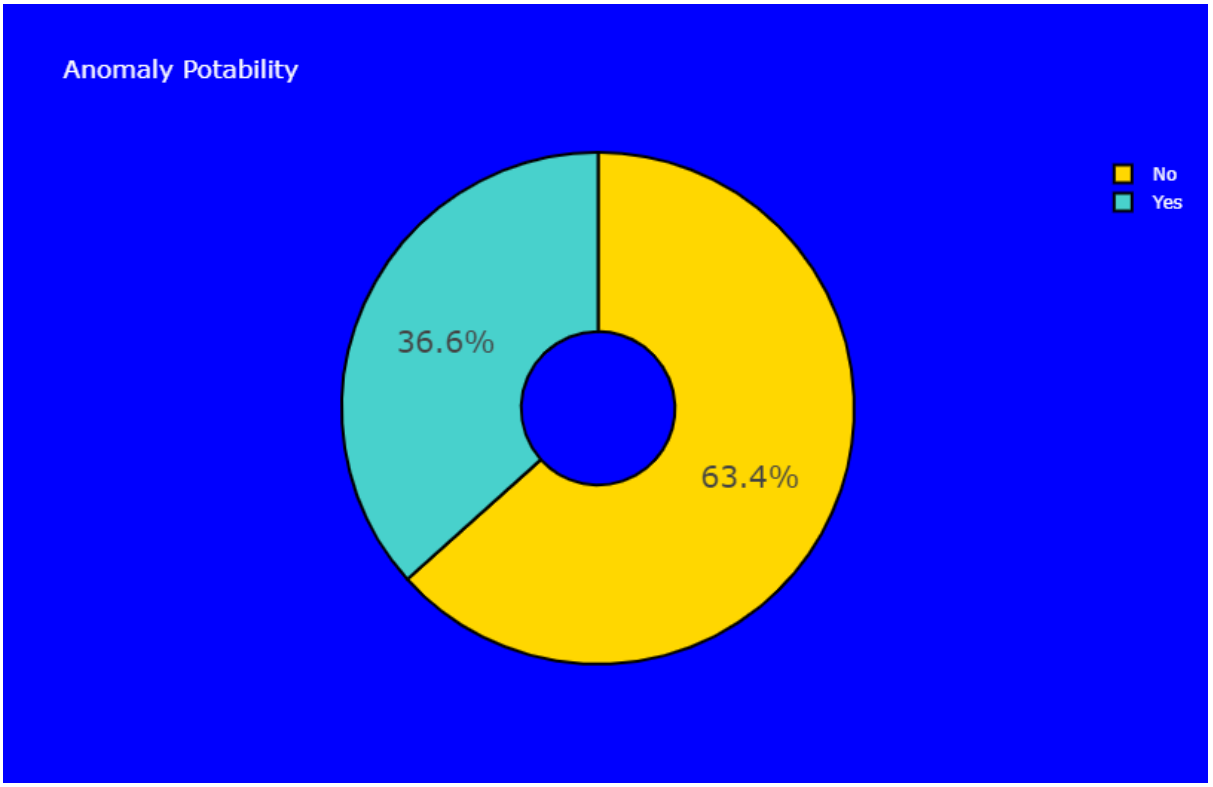
Observation:

There are 110 anomaly data.

If you look at the Top 10 anomaly dates, there are many data judged to have potability.

Looking at the target of data judged as anomaly, there are more cases judged as potability.

What can be predicted from this is that there are many cases in this dataset where undrinkable water is judged to be drinkable.



Conclusion

The analysis of dataset used histogram based technique to apply on each parameter and PCA algorithm is applied overall dataset determined anomaly of potability

Finally 36.6% yes and 63.4% No show the potability of water. 63.4% is large anomaly Of dataset so we need used hyperparameter and standardisation to enhance model Predict the potability accurately .