**WATER QUALITY ANALYSIS**

**PHASE 4: DEVELOPMENT PART 4**

**FEATURE ENGINEERING, MODEL TRAINING & EVALUATION**

**OVERVIEW OF THE PROCESS**:

DATA COLLECTION:

* Collect relevant data on water quality. This can come from various sources, including government agencies, sensors, or even crowdsourced data.
* Ensure the data is well-documented, clean, and in a format suitable for analysis.

DATA EXPLORATION AND PREPROCESSING:

* Explore the dataset to understand its structure and identify any missing or erroneous data.
* Preprocess the data, which may involve handling missing values, data imputation, and outlier detection.
* Conduct feature selection and engineering to create relevant input variables for your models. This can involve transforming, scaling, or creating new features from the raw data.

DATA VISUALISATION:

* Create visualizations to gain insights into the data. Plots, charts, and graphs can help you identify patterns, correlations, and anomalies.

MODEL EVALUATION:

* Evaluate your model(s) using appropriate evaluation metrics. For regression tasks, this might include Mean Absolute Error (MAE) or Root Mean Square Error (RMSE). For classification, you may use metrics like accuracy, precision, recall, and F1 score.
* Perform cross-validation to ensure your model's robustness.

EEATURE ENGINEERING:

* Feature engineering involves creating new features or transforming existing ones to improve the performance of your model.
* Some common techniques include one-hot encoding for categorical variables, scaling or normalizing numerical features, and creating interaction terms or polynomial features.

MONITORING AND MAINTAINANCE:

* Continuously monitor the model's performance in real-world settings and update it as needed.

DOCUMENTATION:

* Maintain comprehensive documentation of your work, including data sources, preprocessing steps, model details, and results.

PROGRAMMING:

import numpy as np *# linear algebra*

import pandas as pd *# data processing, CSV file I/O (e.g. pd.read\_csv)*

import os

for dirname, \_, filenames **in** os.walk('/kaggle/input'):

for filename **in** filenames:

print(os.path.join(dirname, filename))

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

import warnings

warnings.filterwarnings('ignore')

data = pd.read\_csv('/kaggle/input/water-quality-testing/Water Quality Testing.csv')

data.head()

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample ID | pH | Temperature (°C) | Turbidity (NTU) | Dissolved Oxygen (mg/L) | Conductivity (µS/cm) |
| 1 | 7.25 | 23.1 | 4.5 | 7.8 | 342 |  |
| 2 | 7.11 | 22.3 | 5.1 | 6.2 | 335 |  |
| 3 | 7.03 | 21.5 | 3.9 | 8.3 | 356 |  |
| 4 | 7.38 | 22.9 | 3.2 | 9.5 | 327 |  |
| 5 | 7.45 | 20.7 | 3.8 | 8.1 | 352 |  |

data.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 500 entries, 0 to 499

Data columns (total 6 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Sample ID 500 non-null int64

1 pH 500 non-null float64

2 Temperature (°C) 500 non-null float64

3 Turbidity (NTU) 500 non-null float64

4 Dissolved Oxygen (mg/L) 500 non-null float64

5 Conductivity (µS/cm) 500 non-null int64

dtypes: float64(4), int64(2)

memory usage: 23.6 KB

data.shape

(500, 6)

data.isnull().sum()

Sample ID 0

pH 0

Temperature (°C) 0

Turbidity (NTU) 0

Dissolved Oxygen (mg/L) 0

Conductivity (µS/cm) 0

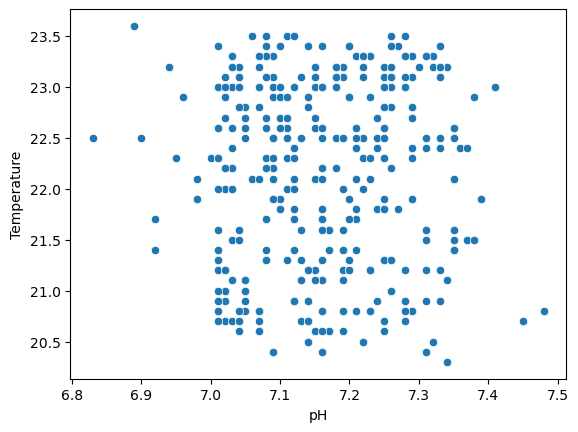
dtype: int64

data.head()

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample ID | pH | Temperature (°C) | Turbidity (NTU) | Dissolved Oxygen (mg/L) | Conductivity (µS/cm) | s |
| 1 | 7.25 | 23.1 | 4.5 | 7.8 | 342 |  |
| 2 | 7.11 | 22.3 | 5.1 | 6.2 | 335 |  |
| 3 | 7.03 | 21.5 | 3.9 | 8.3 | 356 |  |
| 4 | 7.38 | 22.9 | 3.2 | 9.5 | 327 |  |
| 5 | 7.45 | 20.7 | 3.8 | 8.1 | 352 |  |

sns.scatterplot(data=data,x='pH',y='Temperature')

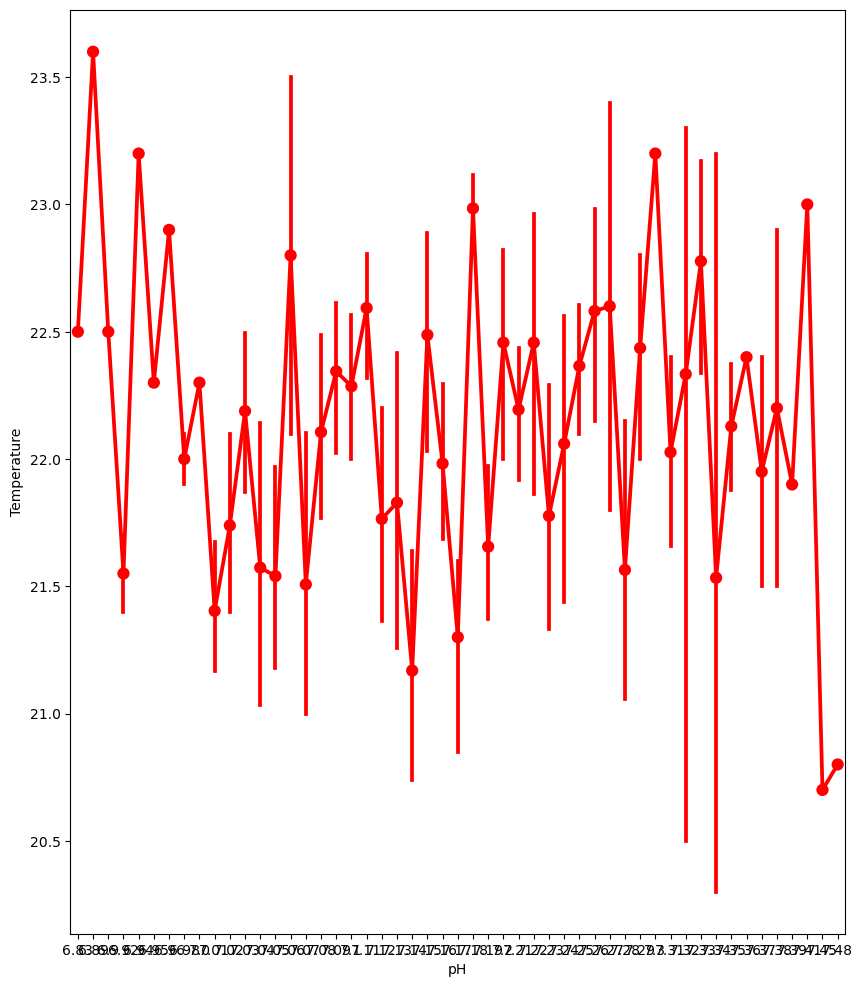
plt.show()



plt.figure(figsize=(10,12))

sns.pointplot(data=data,x='pH',y='Temperature',color='red')

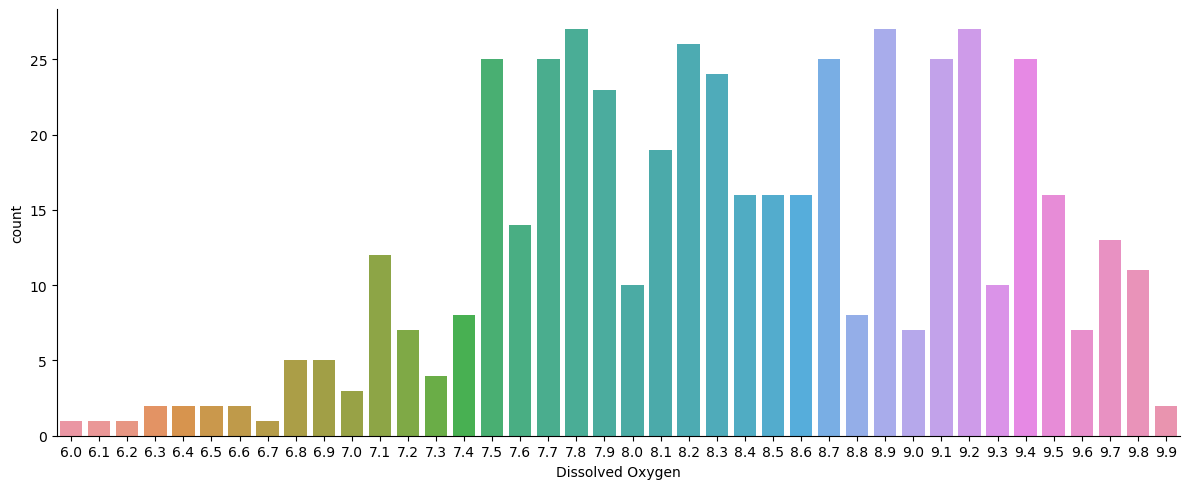
plt.show()



data['Dissolved Oxygen'].value\_counts()

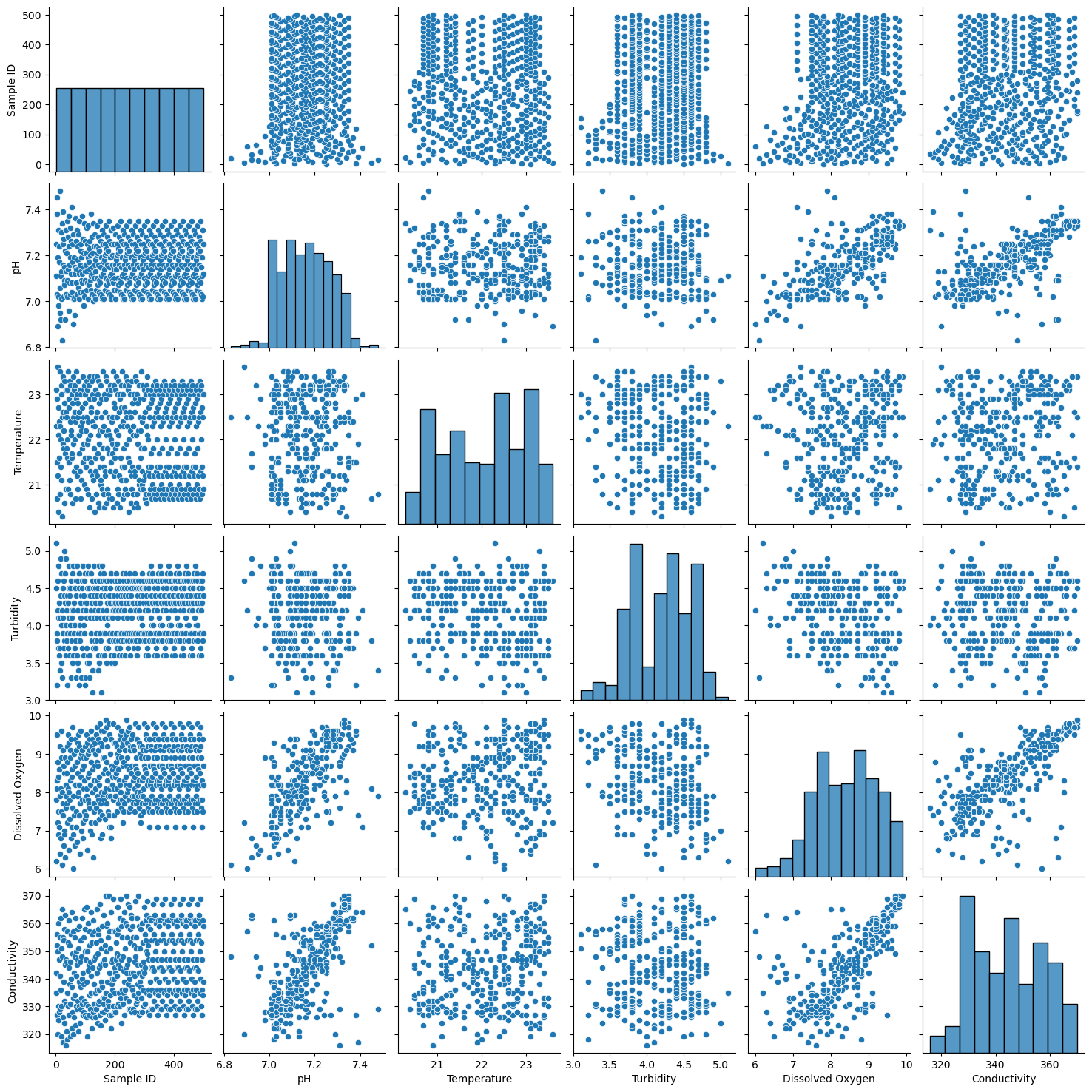
sns.catplot(data=data,x='Dissolved Oxygen',kind='count',aspect=2.4)

plt.show()



sns.pairplot(data)

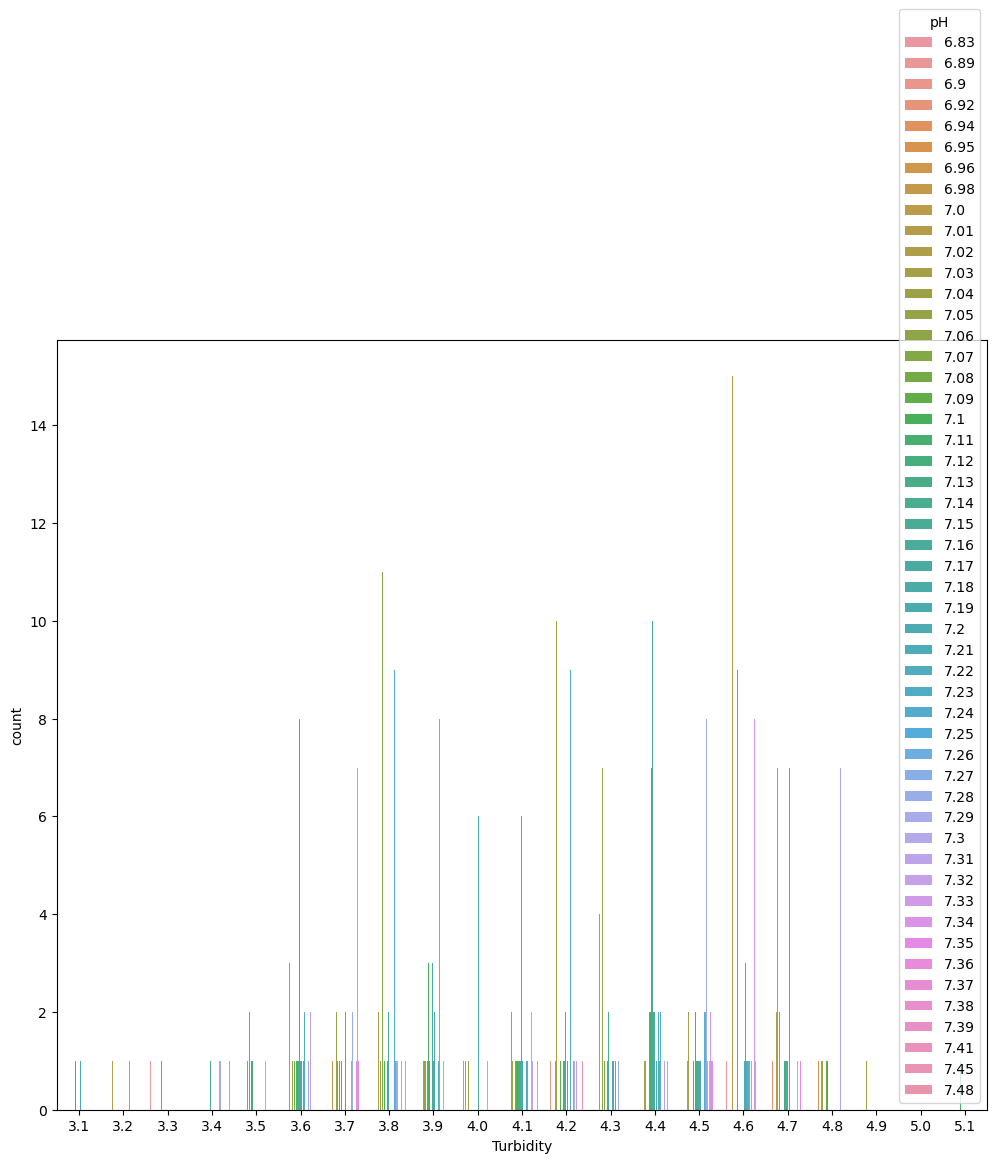
plt.show()



plt.figure(figsize=(12,10))

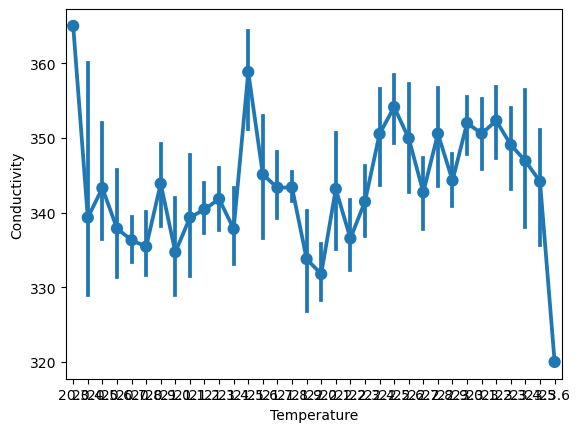
sns.countplot(data=data,x='Turbidity',hue='pH')

plt.show()



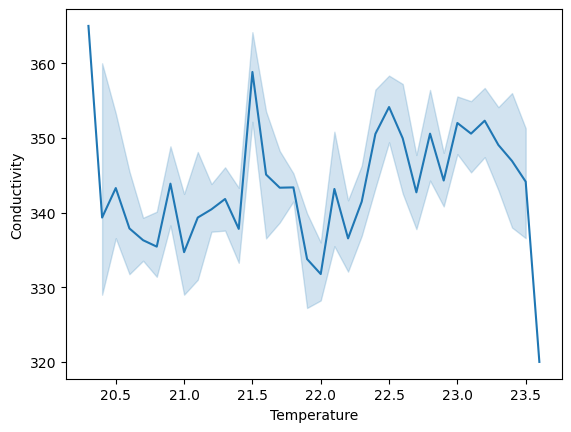
sns.pointplot(data=data,x='Temperature',y='Conductivity')

plt.show()



sns.lineplot(data=data,x='Temperature',y='Conductivity')

plt.show()



CONCLUSION:

* Evaluating water quality using the water quality parameters and esthetic characteristics has a profound significance for human health.
* Thus, Monitoring the quality of your water and testing it regularly is very important to maintain reliable and safe water sources and eliminate the potential health risks related to water contamination.