Date – 26.10.2023

Team ID -714

Project Title – Water quality analysis

**Import Dependencies**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

**Load Dataset**

```
dataset = pd.read_csv('/content/drive/MyDrive/water_potability.csv')
```

**Data Exploration**

```
dataset
```

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | NaN | 204.890456 | 20791.31898 | 7.300212 | 368.516441 | 564.308654 | 10.379783 | 86.990970 | 2.963135 | 0 |
| **1** | 3.716080 | 129.422921 | 18630.05786 | 6.635246 | NaN | 592.885359 | 15.180013 | 56.329076 | 4.500656 | 0 |
| **2** | 8.099124 | 224.236259 | 19909.54173 | 9.275884 | NaN | 418.606213 | 16.868637 | 66.420093 | 3.055934 | 0 |
| **3** | 8.316766 | 214.373394 | 22018.41744 | 8.059332 | 356.886136 | 363.266516 | 18.436525 | 100.341674 | 4.628771 | 0 |
| **4** | 9.092223 | 181.101509 | 17978.98634 | 6.546600 | 310.135738 | 398.410813 | 11.558279 | 31.997993 | 4.075075 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | .. |
| **3271** | 4.668102 | 193.681736 | 47580.99160 | 7.166639 | 359.948574 | 526.424171 | 13.894419 | 66.687695 | 4.435821 | 1 |
| **3272** | 7.808856 | 193.553212 | 17329.80216 | 8.061362 | NaN | 392.449580 | 19.903225 | NaN | 2.798243 | 1 |
| **3273** | 9.419510 | 175.762646 | 33155.57822 | 7.350233 | NaN | 432.044783 | 11.039070 | 69.845400 | 3.298875 | 1 |
| **3274** | 5.126763 | 230.603758 | 11983.86938 | 6.303357 | NaN | 402.883113 | 11.168946 | 77.488213 | 4.708658 | 1 |
| **3275** | 7.874671 | 195.102299 | 17404.17706 | 7.509306 | NaN | 327.459761 | 16.140368 | 78.698446 | 2.309149 | 1 |

3276 rows × 10 columns

```
dataset.shape
```

```
(3276, 10)
```

```
dataset.columns
```

```
Index(['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
       'Organic_carbon', 'Trihalomethanes', 'Turbidity', 'Potability'],
      dtype='object')
```

**Data Preprocessing**

```
dataset.isnull()
```

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | True | False | False | False | False | False | False | False | False | False |
| **1** | False | False | False | False | True | False | False | False | False | False |
| **2** | False | False | False | False | True | False | False | False | False | False |
| **3** | False | False | False | False | False | False | False | False | False | False |
| **4** | False | False | False | False | False | False | False | False | False | False |

```
dataset['ph'] = dataset['ph'].fillna(dataset['ph'].mean())
```

```
dataset['Sulfate'] = dataset['Sulfate'].fillna(dataset['Sulfate'].mode)
```

```
dataset['Trihalomethanes'] = dataset['Trihalomethanes'].fillna(dataset['Trihalomethanes'].mean())
```

```
dataset.isnull().sum()
```

```
ph                0
Hardness          0
Solids            0
Chloramines       0
Sulfate           0
Conductivity      0
Organic_carbon    0
Trihalomethanes   0
Turbidity         0
Potability        0
dtype: int64
```

```
dataset.isnull().sum().sum()
```

```
0
```

```
dataset.describe()
```

| | ph | Hardness | Solids | Chloramines | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 3276.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 3276.000000 |
| **mean** | 7.080795 | 196.369496 | 22014.092526 | 7.122277 | 426.205111 | 14.284970 | 66.396293 | 3.966786 | 0.390110 |
| **std** | 1.469956 | 32.879761 | 8768.570828 | 1.583085 | 80.824064 | 3.308162 | 15.769881 | 0.780382 | 0.487849 |
| **min** | 0.000000 | 47.432000 | 320.942611 | 0.352000 | 181.483754 | 2.200000 | 0.738000 | 1.450000 | 0.000000 |
| **25%** | 6.277673 | 176.850538 | 15666.690300 | 6.127421 | 365.734414 | 12.065801 | 56.647656 | 3.439711 | 0.000000 |
| **50%** | 7.080795 | 196.967627 | 20927.833605 | 7.130299 | 421.884968 | 14.218338 | 66.396293 | 3.955028 | 0.000000 |
| **75%** | 7.870050 | 216.667456 | 27332.762125 | 8.114887 | 481.792305 | 16.557652 | 76.666609 | 4.500320 | 1.000000 |
| **max** | 14.000000 | 323.124000 | 61227.196010 | 13.127000 | 753.342620 | 28.300000 | 124.000000 | 6.739000 | 1.000000 |

```
dataset.describe(include='all')
```

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Tri |
|---|---|---|---|---|---|---|---|---|
| **count** | 3276.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 3276 | 3276.000000 | 3276.000000 | |
| **unique** | NaN | NaN | NaN | NaN | 2496 | NaN | NaN | |
| **top** | NaN | NaN | NaN | NaN | <bound method Series.mode of 0 368.51644... | NaN | NaN | |
| **freq** | NaN | NaN | NaN | NaN | 198 | NaN | NaN | |

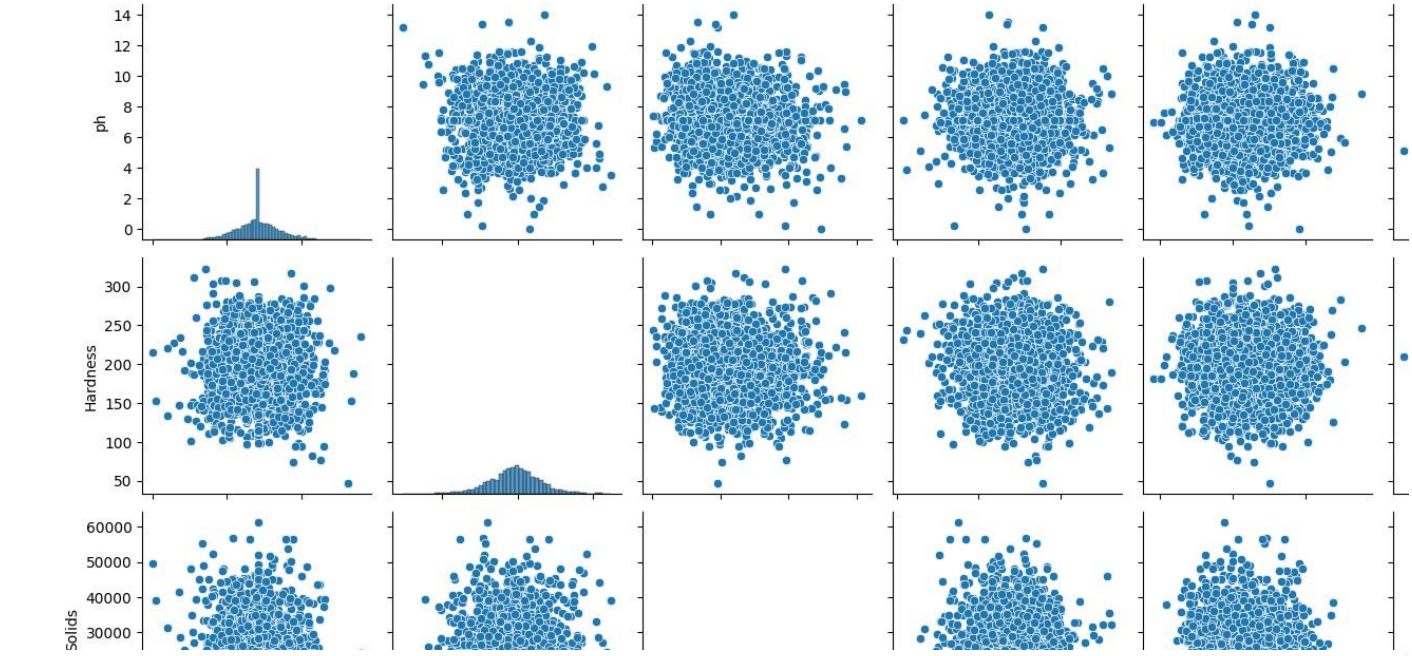
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   ph               3276 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          3276 non-null   object 
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3276 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64  
dtypes: float64(8), int64(1), object(1)
memory usage: 256.1+ KB
```

**Data Visualization**

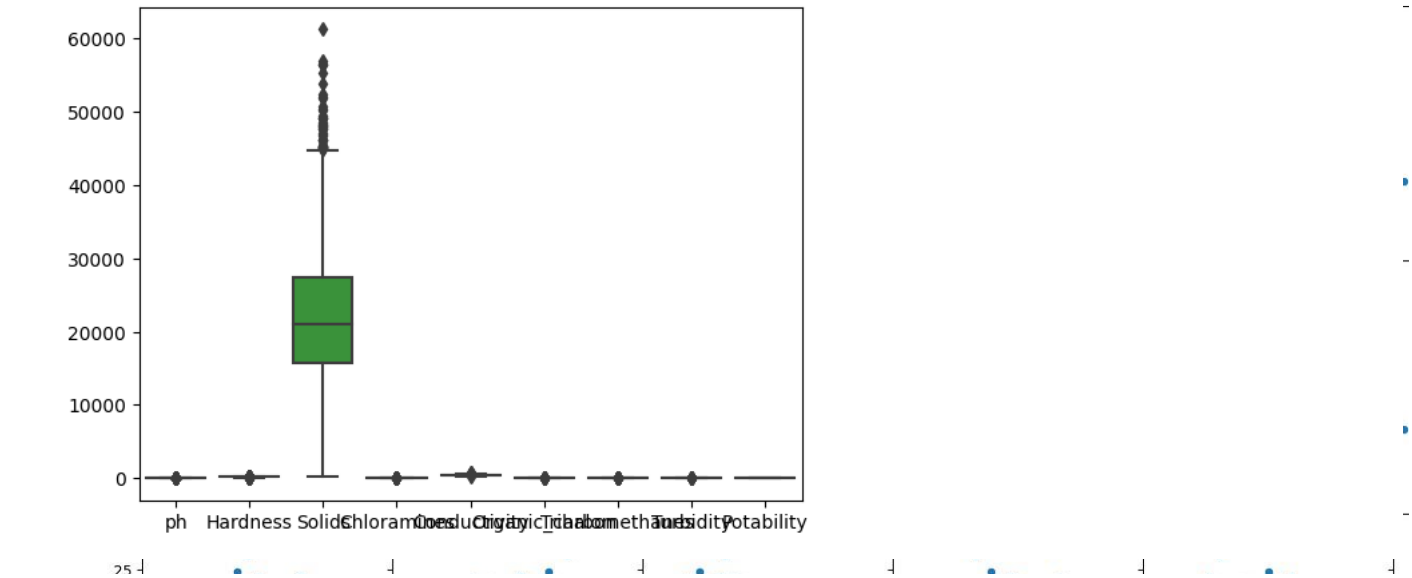
```
plt.figure(figsize=(10,10))
sns.pairplot(dataset)
```

```
<seaborn.axisgrid.PairGrid at 0x7f413b5390f0>
<Figure size 1000x1000 with 0 Axes>
```



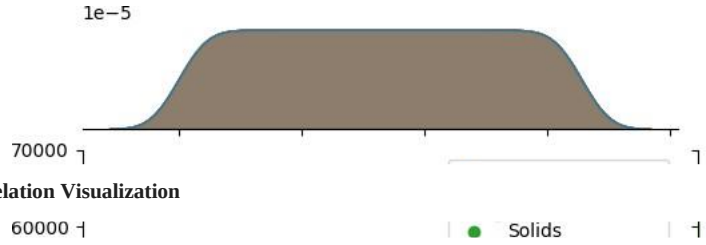

```
sns.boxplot(dataset)
```

```
<Axes: >
```




```
sns.jointplot(dataset)
```

```
<seaborn.axisgrid.JointGrid at 0x7f4132951ae0>
```




**Correlation Visualization**

```
dataset.corr()
```

```
<ipython-input-38-c187c74d1e71>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future ver
  dataset.corr()
```

| | ph | Hardness | Solids | Chloramines | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potability |
|---|---|---|---|---|---|---|---|---|---|
| **ph** | 1.000000 | 0.075833 | -0.081884 | -0.031811 | 0.017192 | 0.040061 | 0.002994 | -0.036222 | -0.003287 |
| **Hardness** | 0.075833 | 1.000000 | -0.046899 | -0.030054 | -0.023915 | 0.003610 | -0.012690 | -0.014449 | -0.013837 |
| **Solids** | -0.081884 | -0.046899 | 1.000000 | -0.070148 | 0.013831 | 0.010242 | -0.008875 | 0.019546 | 0.033743 |
| **Chloramines** | -0.031811 | -0.030054 | -0.070148 | 1.000000 | -0.020486 | -0.012653 | 0.016627 | 0.002363 | 0.023779 |
| **Conductivity** | 0.017192 | -0.023915 | 0.013831 | -0.020486 | 1.000000 | 0.020966 | 0.001255 | 0.005798 | -0.008128 |
| **Organic_carbon** | 0.040061 | 0.003610 | 0.010242 | -0.012653 | 0.020966 | 1.000000 | -0.012976 | -0.027308 | -0.030001 |
| **Trihalomethanes** | 0.002994 | -0.012690 | -0.008875 | 0.016627 | 0.001255 | -0.012976 | 1.000000 | -0.021502 | 0.006960 |
| **Turbidity** | -0.036222 | -0.014449 | 0.019546 | 0.002363 | 0.005798 | -0.027308 | -0.021502 | 1.000000 | 0.001581 |
| **Potability** | -0.003287 | -0.013837 | 0.033743 | 0.023779 | -0.008128 | -0.030001 | 0.006960 | 0.001581 | 1.000000 |

```
sns.heatmap(dataset.corr(),annot=True)
```

```
<ipython-input-40-9d3fd451b567>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future ver
  sns.heatmap(dataset.corr(),annot=True)
<Axes: >
```