CaseStudy-WaterQuality

May 9, 2023

The case study deals with Water quality data, there are actually six water samples involved. And in all the six samples the test to evaluate the parameters pH,EC,TDS,BOD and COD are repeated for three times and the readings are saved.Mean value and standard deviation value of the three test repetition has been calculated and recorded.

1. pH - Potential of Hydrogen. This is a measure of how acidic/basic the water is. The range goes from 0 to 14, with 7 being neutral. pH < 7 is acidic and pH > 7 is base.
2. EC - Electrical Conductivity. This is a measure of Water's ability to conduct electricity. The normal range for purified drinking water ranges from 0.5 to 3 µS/cm(millimhos/cm).
3. TDS - Total Dissolved Solids. This refers to the total concentration of dissolved substances in drinking water. Normal TDS value for purified drinking water is under 25 ppm (parts per milligram).
4. BOD - Biological Oxygen Demand. This is a measure of the amount of oxygen required to remove waste organic matter from water in the process of decomposition by aerobic bacteria. Aerobic bacteria lives only in an environment containing oxygen. Normal drinking water has a BOD level of 1 - 2 ppm (parts per milligram)
5. COD - Chemical Oxygen Demand. This is a water quality measure used not only to determine the amount of biologicaly active substances such as bacteria but also biologically inactive organic matter in water.It is an important amd rapidly measured variable for characterising water bodies, sewage, industrial wastes and treatment plant effluents.

Our main aim is to find out the correlation between pH values and other the numerical columns.

Import the Excel sheet containing the required data for water quality analysis and display the data present in the sheet.

```
  Parameter   Unit      S-1       S-2      S-3       S-4      S-5      S-6
0      pH-1   none     8.70      8.60     8.70      8.36     8.80     9.16
1      pH-2   none     8.92      8.61     8.48      8.42     8.41     8.52
2      pH-3   none     7.80      7.43     7.75      8.62     9.37     9.38
3      Mean    NaN     8.50      8.20     8.20      8.40     8.80     9.00
4        SD    NaN     0.48      0.56     0.17      0.12     0.36     0.38
5       NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
6       NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
7      EC-1  µS/cm  3130.00  12440.00  2340.00  14690.00  9130.00  7790.00
8      EC-2  µS/cm  3310.00  10340.00  4470.00  15210.00  4580.00  4760.00
9      EC-3  µS/cm  4580.00  13750.00  1880.00  18570.00  4480.00  4160.00
```

```
10       Mean    NaN  3673.30  12176.60  2896.70  16156.70  6063.30  5570.00
11         SD    NaN   645.30   1404.50  1128.30   1719.60  2168.80  1588.80
12        NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
13      TDS-1   mg/L  2000.00   8512.00  1490.00   9956.00  5850.00  5380.00
14      TDS-2   mg/L  2280.00   6750.00  2930.00   2750.00  3001.00  3110.00
15      TDS-3   mg/L  1230.00   8900.00  2270.00   2930.00  9960.00  4460.00
16       Mean    NaN  1836.70   8054.00  6690.00   5212.00  6270.30  4316.70
17         SD    NaN   443.90    935.60   588.60   3355.30  2856.50   932.20
18        NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
19      BOD-1   mg/L    16.00     15.00    31.00     12.00    24.00    38.00
20      BOD-2   mg/L    18.00     20.00    22.00     18.00    32.00    22.00
21      BOD-3   mg/L    80.00     80.00    72.00     58.00    60.00    48.00
22       Mean    NaN    38.00     38.30    41.70     29.30    38.70    36.00
23         SD    NaN    29.70     29.50    21.80     20.40    15.40    10.70
24        NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
25      COD-1   mg/L   323.00    282.00   229.00    217.00   175.00   161.00
26      COD-2   mg/L   312.00    201.00   105.00     97.00    63.20    46.40
27      COD-3   mg/L   548.00    391.00   229.00    201.00   180.00   135.00
28       Mean    NaN   394.30    291.30   187.70    171.70   139.40   342.40
29         SD    NaN   108.80     77.80    58.50     53.20    53.90    49.10
```

Below given is the number null values in the imported dataset.

```
Parameter     5
Unit         15
S-1           5
S-2           5
S-3           5
S-4           5
S-5           5
S-6           5
dtype: int64
```

```
   Parameter   Unit      S-1       S-2      S-3       S-4      S-5      S-6
0        pH-1   none     8.70      8.60     8.70      8.36     8.80     9.16
1        pH-2   none     8.92      8.61     8.48      8.42     8.41     8.52
2        pH-3   none     7.80      7.43     7.75      8.62     9.37     9.38
3        Mean    NaN     8.50      8.20     8.20      8.40     8.80     9.00
4          SD    NaN     0.48      0.56     0.17      0.12     0.36     0.38
5         NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
6         NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
7        EC-1  µS/cm  3130.00  12440.00  2340.00  14690.00  9130.00  7790.00
8        EC-2  µS/cm  3310.00  10340.00  4470.00  15210.00  4580.00  4760.00
9        EC-3  µS/cm  4580.00  13750.00  1880.00  18570.00  4480.00  4160.00
10       Mean    NaN  3673.30  12176.60  2896.70  16156.70  6063.30  5570.00
11         SD    NaN   645.30   1404.50  1128.30   1719.60  2168.80  1588.80
12        NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
```

```
13     TDS-1   mg/L  2000.00   8512.00  1490.00   9956.00  5850.00  5380.00
14     TDS-2   mg/L  2280.00   6750.00  2930.00   2750.00  3001.00  3110.00
15     TDS-3   mg/L  1230.00   8900.00  2270.00   2930.00  9960.00  4460.00
16      Mean    NaN  1836.70   8054.00  6690.00   5212.00  6270.30  4316.70
17        SD    NaN   443.90    935.60   588.60   3355.30  2856.50   932.20
18       NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
19     BOD-1   mg/L    16.00     15.00    31.00     12.00    24.00    38.00
20     BOD-2   mg/L    18.00     20.00    22.00     18.00    32.00    22.00
21     BOD-3   mg/L    80.00     80.00    72.00     58.00    60.00    48.00
22      Mean    NaN    38.00     38.30    41.70     29.30    38.70    36.00
23        SD    NaN    29.70     29.50    21.80     20.40    15.40    10.70
24       NaN    NaN      NaN       NaN      NaN       NaN      NaN      NaN
25     COD-1   mg/L   323.00    282.00   229.00    217.00   175.00   161.00
26     COD-2   mg/L   312.00    201.00   105.00     97.00    63.20    46.40
27     COD-3   mg/L   548.00    391.00   229.00    201.00   180.00   135.00
28      Mean    NaN   394.30    291.30   187.70    171.70   139.40   342.40
29        SD    NaN   108.80     77.80    58.50     53.20    53.90    49.10
```

Unit column needs to be removed, since it is no longer required for our analysis.

The dataset looks like below mentioned one after removing unit column

```
  Parameter   S-1   S-2   S-3   S-4   S-5   S-6
0      pH-1  8.70  8.60  8.70  8.36  8.80  9.16
1      pH-2  8.92  8.61  8.48  8.42  8.41  8.52
2      pH-3  7.80  7.43  7.75  8.62  9.37  9.38
3      Mean  8.50  8.20  8.20  8.40  8.80  9.00
4        SD  0.48  0.56  0.17  0.12  0.36  0.38
```

For the benefit of our analysis, the dataset WaterQuality is being transposed. And the data is displayed below for better understanding.

```
              0     1     2     3     4    5    6        7        8        9  \
Parameter  pH-1  pH-2  pH-3  Mean    SD  NaN  NaN     EC-1     EC-2     EC-3
S-1         8.7  8.92   7.8   8.5  0.48  NaN  NaN   3130.0   3310.0   4580.0
S-2         8.6  8.61  7.43   8.2  0.56  NaN  NaN  12440.0  10340.0  13750.0
S-3         8.7  8.48  7.75   8.2  0.17  NaN  NaN   2340.0   4470.0   1880.0
S-4        8.36  8.42  8.62   8.4  0.12  NaN  NaN  14690.0  15210.0  18570.0
S-5         8.8  8.41  9.37   8.8  0.36  NaN  NaN   9130.0   4580.0   4480.0
S-6        9.16  8.52  9.38   9.0  0.38  NaN  NaN   7790.0   4760.0   4160.0

           …     20     21    22    23   24     25     26     27     28  \
Parameter  …  BOD-2  BOD-3  Mean    SD  NaN  COD-1  COD-2  COD-3   Mean
S-1        …   18.0   80.0  38.0  29.7  NaN  323.0  312.0  548.0  394.3
S-2        …   20.0   80.0  38.3  29.5  NaN  282.0  201.0  391.0  291.3
S-3        …   22.0   72.0  41.7  21.8  NaN  229.0  105.0  229.0  187.7
S-4        …   18.0   58.0  29.3  20.4  NaN  217.0   97.0  201.0  171.7
S-5        …   32.0   60.0  38.7  15.4  NaN  175.0   63.2  180.0  139.4
```

```
S-6         …   22.0   48.0  36.0  10.7  NaN  161.0   46.4  135.0  342.4

               29
Parameter      SD
S-1         108.8
S-2          77.8
S-3          58.5
S-4          53.2
S-5          53.9
S-6          49.1

[7 rows x 30 columns]
```

Since we have transposed the dataset, Now we have need to convert the First row as column header and delete the first row.

```
Parameter   pH-1  pH-2  pH-3 Mean    SD  NaN  NaN     EC-1     EC-2     EC-3  \
S-1          8.7  8.92   7.8  8.5  0.48  NaN  NaN   3130.0   3310.0   4580.0
S-2          8.6  8.61  7.43  8.2  0.56  NaN  NaN  12440.0  10340.0  13750.0
S-3          8.7  8.48  7.75  8.2  0.17  NaN  NaN   2340.0   4470.0   1880.0
S-4         8.36  8.42  8.62  8.4  0.12  NaN  NaN  14690.0  15210.0  18570.0
S-5          8.8  8.41  9.37  8.8  0.36  NaN  NaN   9130.0   4580.0   4480.0
S-6         9.16  8.52  9.38  9.0  0.38  NaN  NaN   7790.0   4760.0   4160.0

Parameter   … BOD-2 BOD-3  Mean    SD  NaN  COD-1  COD-2  COD-3   Mean     SD
S-1         …  18.0  80.0  38.0  29.7  NaN  323.0  312.0  548.0  394.3  108.8
S-2         …  20.0  80.0  38.3  29.5  NaN  282.0  201.0  391.0  291.3   77.8
S-3         …  22.0  72.0  41.7  21.8  NaN  229.0  105.0  229.0  187.7   58.5
S-4         …  18.0  58.0  29.3  20.4  NaN  217.0   97.0  201.0  171.7   53.2
S-5         …  32.0  60.0  38.7  15.4  NaN  175.0   63.2  180.0  139.4   53.9
S-6         …  22.0  48.0  36.0  10.7  NaN  161.0   46.4  135.0  342.4   49.1

[6 rows x 30 columns]
```

The dataset after deleting the NULL values across rows and columns.

```
Parameter   pH-1  pH-2  pH-3 Mean    SD     EC-1     EC-2     EC-3     Mean  \
S-1          8.7  8.92   7.8  8.5  0.48   3130.0   3310.0   4580.0   3673.3
S-2          8.6  8.61  7.43  8.2  0.56  12440.0  10340.0  13750.0  12176.6
S-3          8.7  8.48  7.75  8.2  0.17   2340.0   4470.0   1880.0   2896.7
S-4         8.36  8.42  8.62  8.4  0.12  14690.0  15210.0  18570.0  16156.7
S-5          8.8  8.41  9.37  8.8  0.36   9130.0   4580.0   4480.0   6063.3
S-6         9.16  8.52  9.38  9.0  0.38   7790.0   4760.0   4160.0   5570.0

Parameter       SD  … BOD-1 BOD-2 BOD-3  Mean    SD  COD-1  COD-2  COD-3  \
S-1          645.3  …  16.0  18.0  80.0  38.0  29.7  323.0  312.0  548.0
S-2         1404.5  …  15.0  20.0  80.0  38.3  29.5  282.0  201.0  391.0
S-3         1128.3  …  31.0  22.0  72.0  41.7  21.8  229.0  105.0  229.0
```

```
 S-4        1719.6  …  12.0  18.0  58.0  29.3  20.4  217.0   97.0  201.0
 S-5        2168.8  …  24.0  32.0  60.0  38.7  15.4  175.0   63.2  180.0
 S-6        1588.8  …  38.0  22.0  48.0  36.0  10.7  161.0   46.4  135.0

 Parameter   Mean     SD
 S-1        394.3  108.8
 S-2        291.3   77.8
 S-3        187.7   58.5
 S-4        171.7   53.2
 S-5        139.4   53.9
 S-6        342.4   49.1

 [6 rows x 25 columns]
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 6 entries, S-1 to S-6
Data columns (total 25 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   pH-1    6 non-null      object
 1   pH-2    6 non-null      object
 2   pH-3    6 non-null      object
 3   Mean    6 non-null      object
 4   SD      6 non-null      object
 5   EC-1    6 non-null      object
 6   EC-2    6 non-null      object
 7   EC-3    6 non-null      object
 8   Mean    6 non-null      object
 9   SD      6 non-null      object
 10  TDS-1   6 non-null      object
 11  TDS-2   6 non-null      object
 12  TDS-3   6 non-null      object
 13  Mean    6 non-null      object
 14  SD      6 non-null      object
 15  BOD-1   6 non-null      object
 16  BOD-2   6 non-null      object
 17  BOD-3   6 non-null      object
 18  Mean    6 non-null      object
 19  SD      6 non-null      object
 20  COD-1   6 non-null      object
 21  COD-2   6 non-null      object
 22  COD-3   6 non-null      object
 23  Mean    6 non-null      object
 24  SD      6 non-null      object
dtypes: object(25)
memory usage: 1.2+ KB
```

Since all the columns are converted as object columns, we need to convert it to float columns to perform required actions on the numerical columns.

```
<class 'pandas.core.frame.DataFrame'>
Index: 6 entries, S-1 to S-6
Data columns (total 25 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   pH-1    6 non-null      float64
 1   pH-2    6 non-null      float64
 2   pH-3    6 non-null      float64
 3   Mean    6 non-null      float64
 4   SD      6 non-null      float64
 5   EC-1    6 non-null      float64
 6   EC-2    6 non-null      float64
 7   EC-3    6 non-null      float64
 8   Mean    6 non-null      float64
 9   SD      6 non-null      float64
 10  TDS-1   6 non-null      float64
 11  TDS-2   6 non-null      float64
 12  TDS-3   6 non-null      float64
 13  Mean    6 non-null      float64
 14  SD      6 non-null      float64
 15  BOD-1   6 non-null      float64
 16  BOD-2   6 non-null      float64
 17  BOD-3   6 non-null      float64
 18  Mean    6 non-null      float64
 19  SD      6 non-null      float64
 20  COD-1   6 non-null      float64
 21  COD-2   6 non-null      float64
 22  COD-3   6 non-null      float64
 23  Mean    6 non-null      float64
 24  SD      6 non-null      float64
dtypes: float64(25)
memory usage: 1.2+ KB
```

For better readability, the mean and Sd columns are renamed with relevant names. ‘pH-Mean’,‘pH-SD’,‘EC-Mean’,‘EC-SD’,‘TDS-Mean’,‘TDS-SD’,‘BOD-Mean’,‘BOD-SD’,‘COD-Mean’,COD-SD’

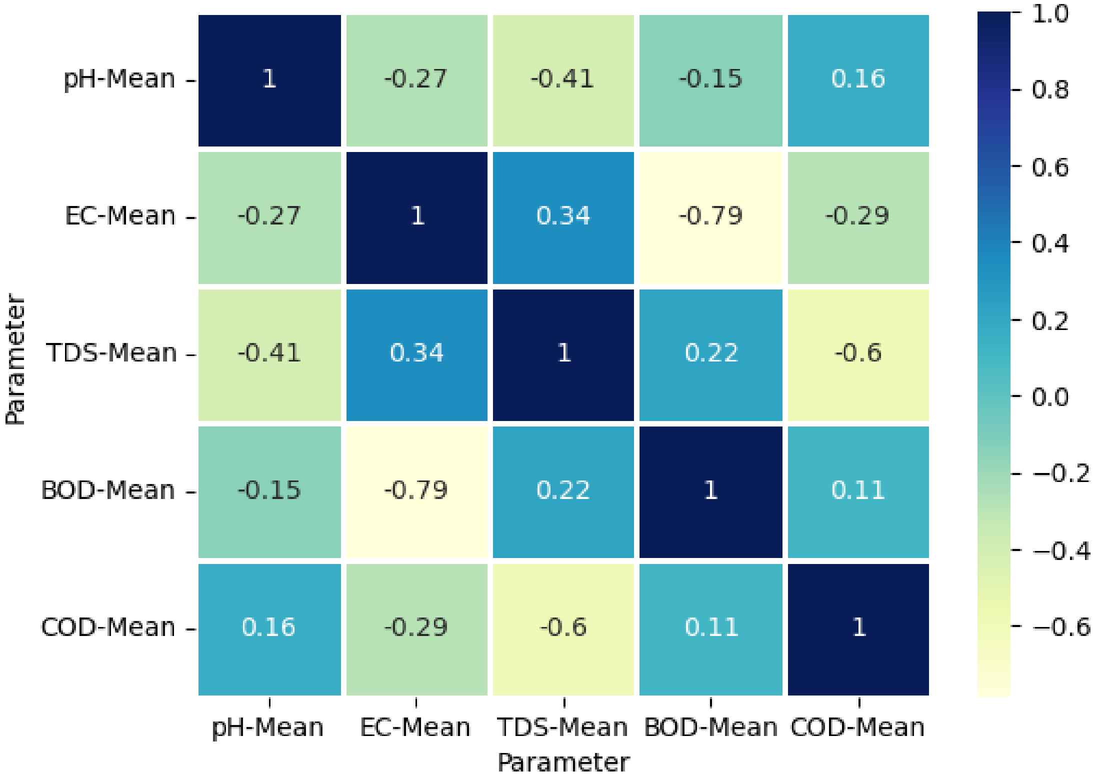
```
<class 'pandas.core.frame.DataFrame'>
Index: 6 entries, S-1 to S-6
Data columns (total 25 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   pH-1      6 non-null      float64
 1   pH-2      6 non-null      float64
 2   pH-3      6 non-null      float64
 3   pH-Mean   6 non-null      float64
 4   pH-SD     6 non-null      float64
 5   EC-1      6 non-null      float64
 6   EC-2      6 non-null      float64
 7   EC-3      6 non-null      float64
 8   EC-Mean   6 non-null      float64
 9   EC-SD     6 non-null      float64
 10  TDS-1     6 non-null      float64
 11  TDS-2     6 non-null      float64
 12  TDS-3     6 non-null      float64
 13  TDS-Mean  6 non-null      float64
 14  TDS-SD    6 non-null      float64
 15  BOD-1     6 non-null      float64
 16  BOD-2     6 non-null      float64
 17  BOD-3     6 non-null      float64
 18  BOD-Mean  6 non-null      float64
 19  BOD-SD    6 non-null      float64
 20  COD-1     6 non-null      float64
 21  COD-2     6 non-null      float64
 22  COD-3     6 non-null      float64
 23  COD-Mean  6 non-null      float64
 24  COD-SD    6 non-null      float64
dtypes: float64(25)
memory usage: 1.2+ KB
```

Correlation Between Mean columns

```
Parameter   pH-Mean   EC-Mean  TDS-Mean  BOD-Mean  COD-Mean
Parameter
pH-Mean    1.000000 -0.270410 -0.411112 -0.149625  0.159292
EC-Mean   -0.270410  1.000000  0.340275 -0.785475 -0.286521
TDS-Mean  -0.411112  0.340275  1.000000  0.219680 -0.597765
BOD-Mean  -0.149625 -0.785475  0.219680  1.000000  0.105256
COD-Mean   0.159292 -0.286521 -0.597765  0.105256  1.000000
```

Heat Map Generated for Mean Columns.

```
<Axes: xlabel='Parameter', ylabel='Parameter'>
```



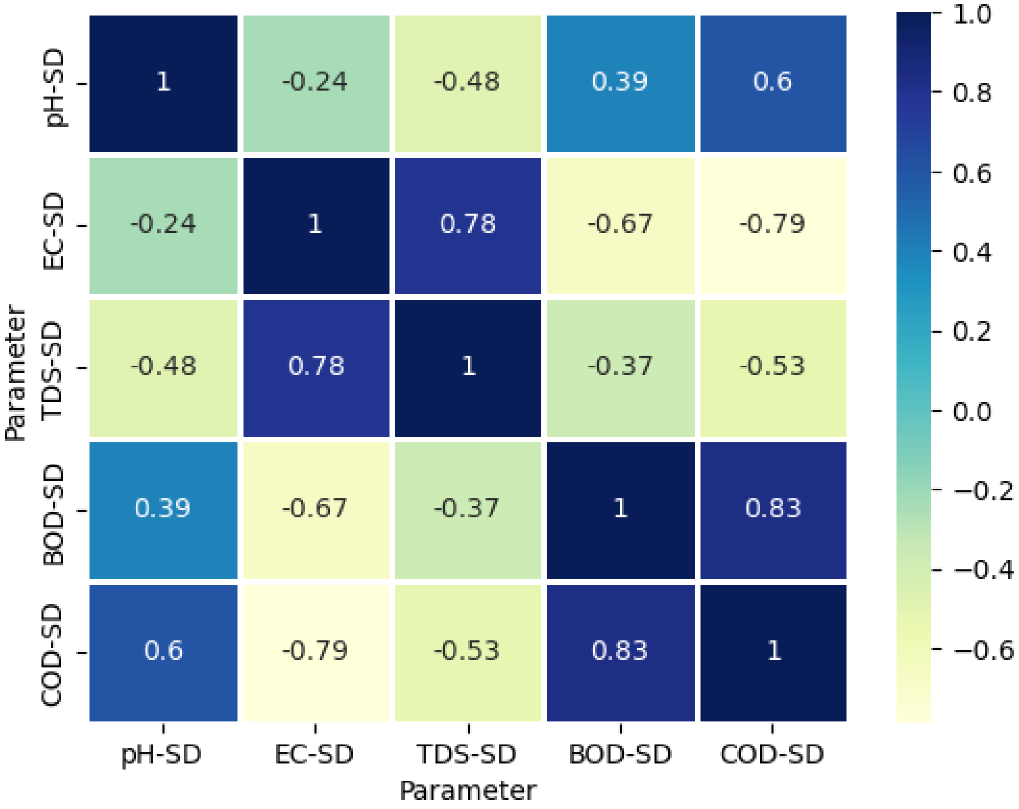
1. With respect to pH mean,there is Negative Weak Correlation with EC, BOD & COD. Negative Moderate correlation with TDS.
2. With respect to EC mean, there is Negative Weak Correlation with pH & COD. Positive weak correlation with TDS and Negative strong correlation with BOD
3. With respect to TDS mean, there is Negative moderate correlation with pH, COD. Positive Moderate correlation with EC and BOD
4. With respect to BOD mean, Negative weak correlation with pH, Negative strong correlation with EC, Positive wak correlation with TDS and COD
5. With respect to COD mean, there is positive wean correlation with pH & BOD, Negative weak correlation with EC and Negative moderate correlation with TDS.

Correlation Between Standard Deviation columns

```
Parameter      pH-SD     EC-SD    TDS-SD    BOD-SD    COD-SD
Parameter
pH-SD       1.000000 -0.239374 -0.484594  0.393452  0.597802
EC-SD      -0.239374  1.000000  0.784072 -0.671500 -0.787522
TDS-SD     -0.484594  0.784072  1.000000 -0.371113 -0.527358
BOD-SD      0.393452 -0.671500 -0.371113  1.000000  0.826020
COD-SD      0.597802 -0.787522 -0.527358  0.826020  1.000000
```

Heat Map Generated Standard Deviation Mean Columns.

```
<Axes: xlabel='Parameter', ylabel='Parameter'>
```

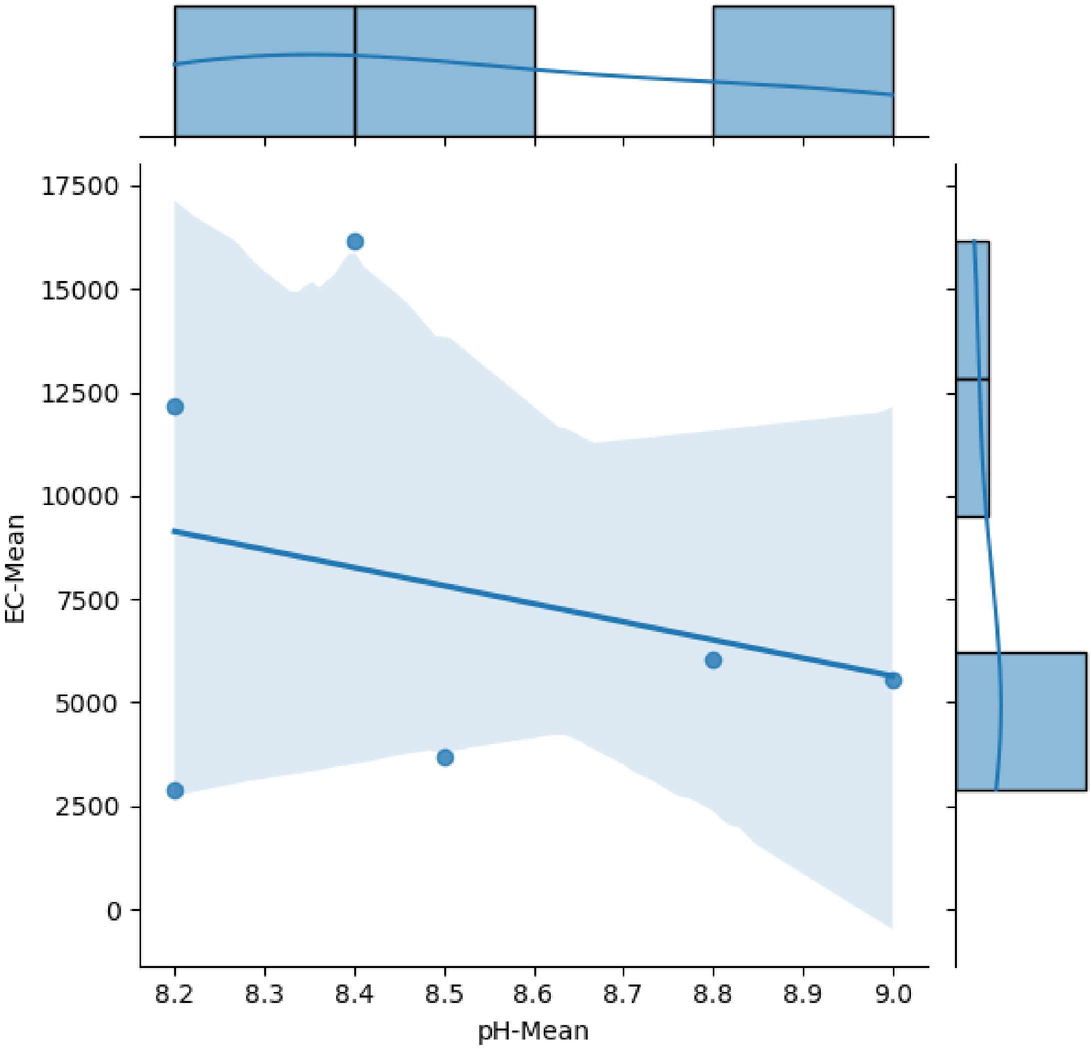


1. With respect to pH SD, there is a Negative Weak correlation with EC, Negative Moderate correlation with TDS, Positive Moderate correlation with BOD and Positive Strong correlation with COD.
2. With respect to EC SD, there is Negative Weak correlation with pH. Positive Strong correlation with TDS, Negatiive Strong correlation with BOD and COD
3. With respect to TDS SD, there is a Negative moderate correlation with pH & BOD, Positive Strong correlation with EC and Negative strong correlation wuth COD
4. with respect to BOD SD, there is positive Moderate correlation with pH, Negative Strong Correlation with EC, Negative Moderate correlation with TDS and Positive strong correlation with COD.
5. With respect to COD SD, there is positive strong correlation with pH & BOD, Negative strong correlation with EC & TDS

Mean values Comparison

Potential of Hydrogen VS Electrical Conductivity

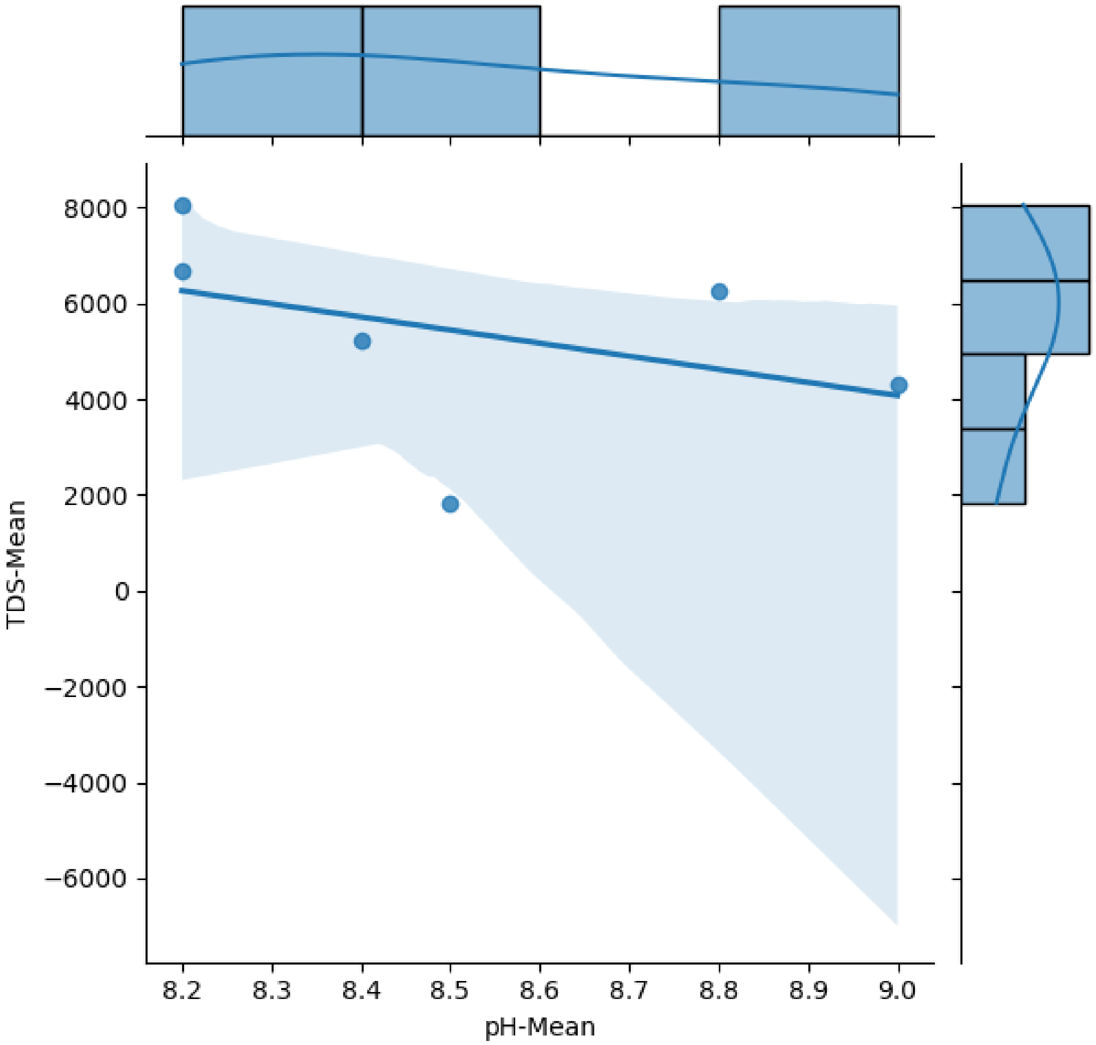
```
PearsonRResult(statistic=-0.2704097992258808, pvalue=0.604271680754778)
```



Potential of Hydrogen VS Total Dissolved Solids

```
PearsonRResult(statistic=-0.4111118465174884, pvalue=0.4180738432750332)
```

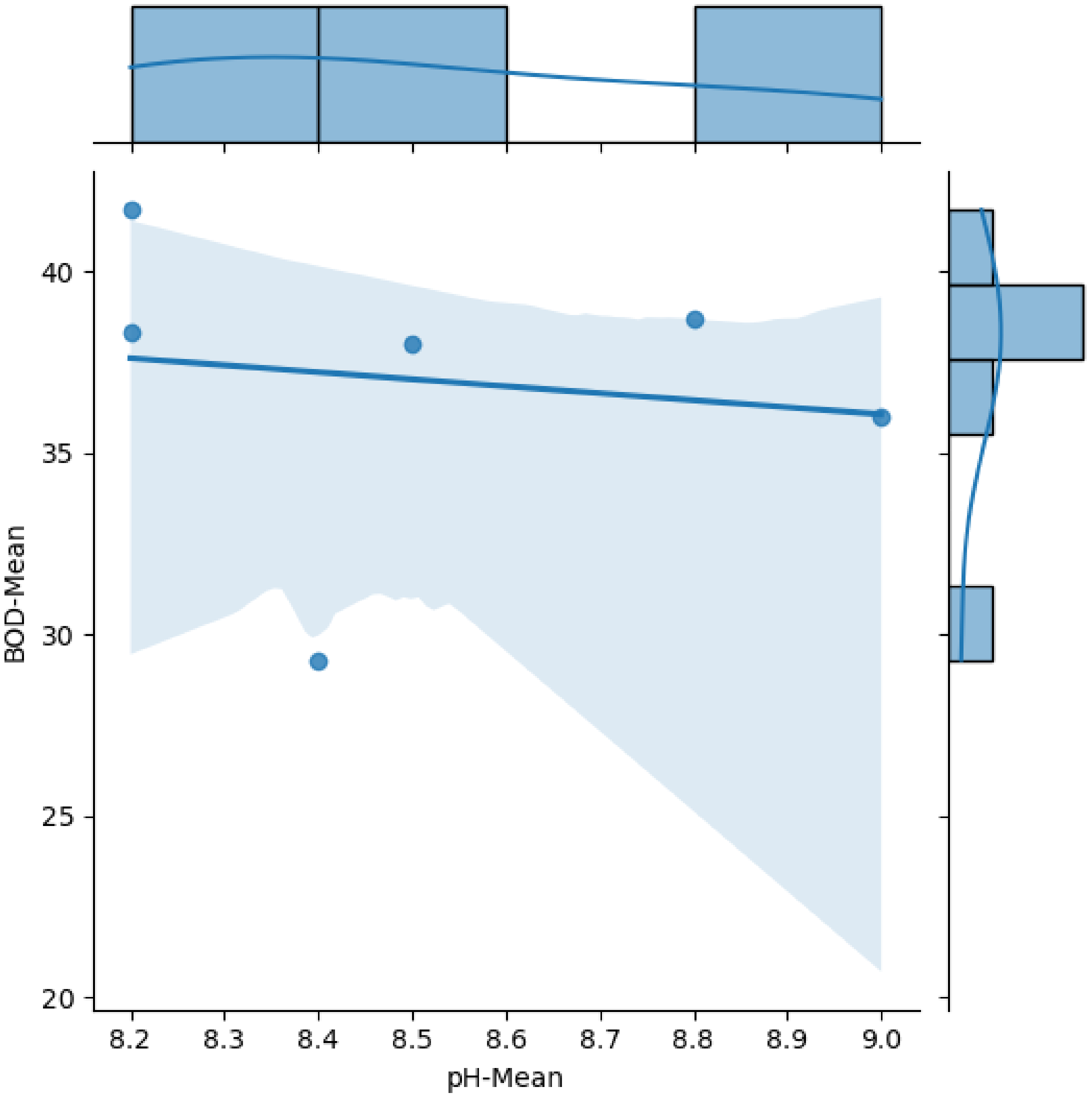
```
 <seaborn.axisgrid.JointGrid at 0x170d94b8c40>
```



Potential of Hydrogen VS Biological Oxygen Demand

```
PearsonRResult(statistic=-0.14962479055901207, pvalue=0.7772376824924154)
```

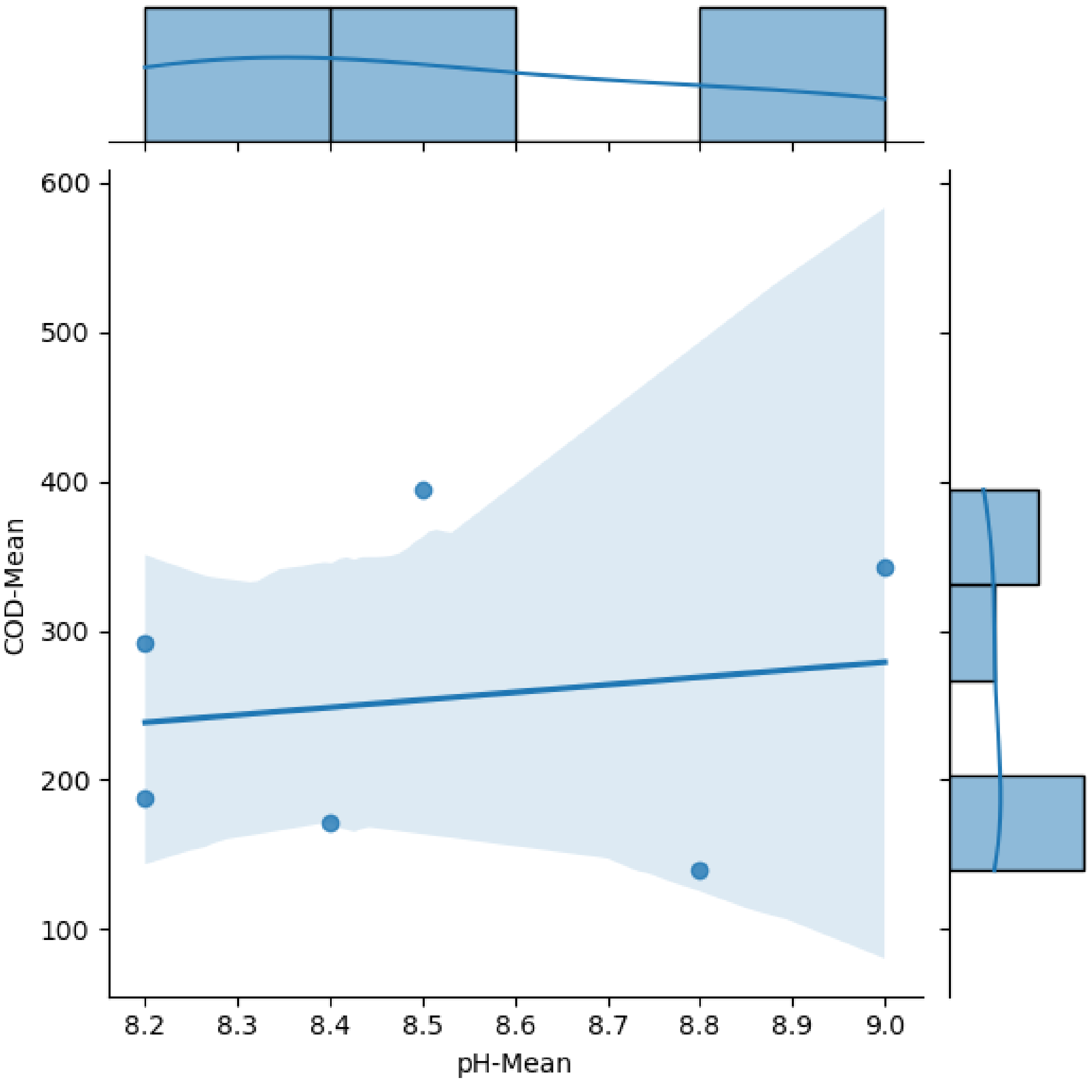
```
 <seaborn.axisgrid.JointGrid at 0x170d9f1b820>
```



Potential of Hydrogen VS Chemical Oxygen Demand

```
PearsonRResult(statistic=0.1592917629600689, pvalue=0.7630832794638668)
```

```
 <seaborn.axisgrid.JointGrid at 0x170dadd87f0>
```

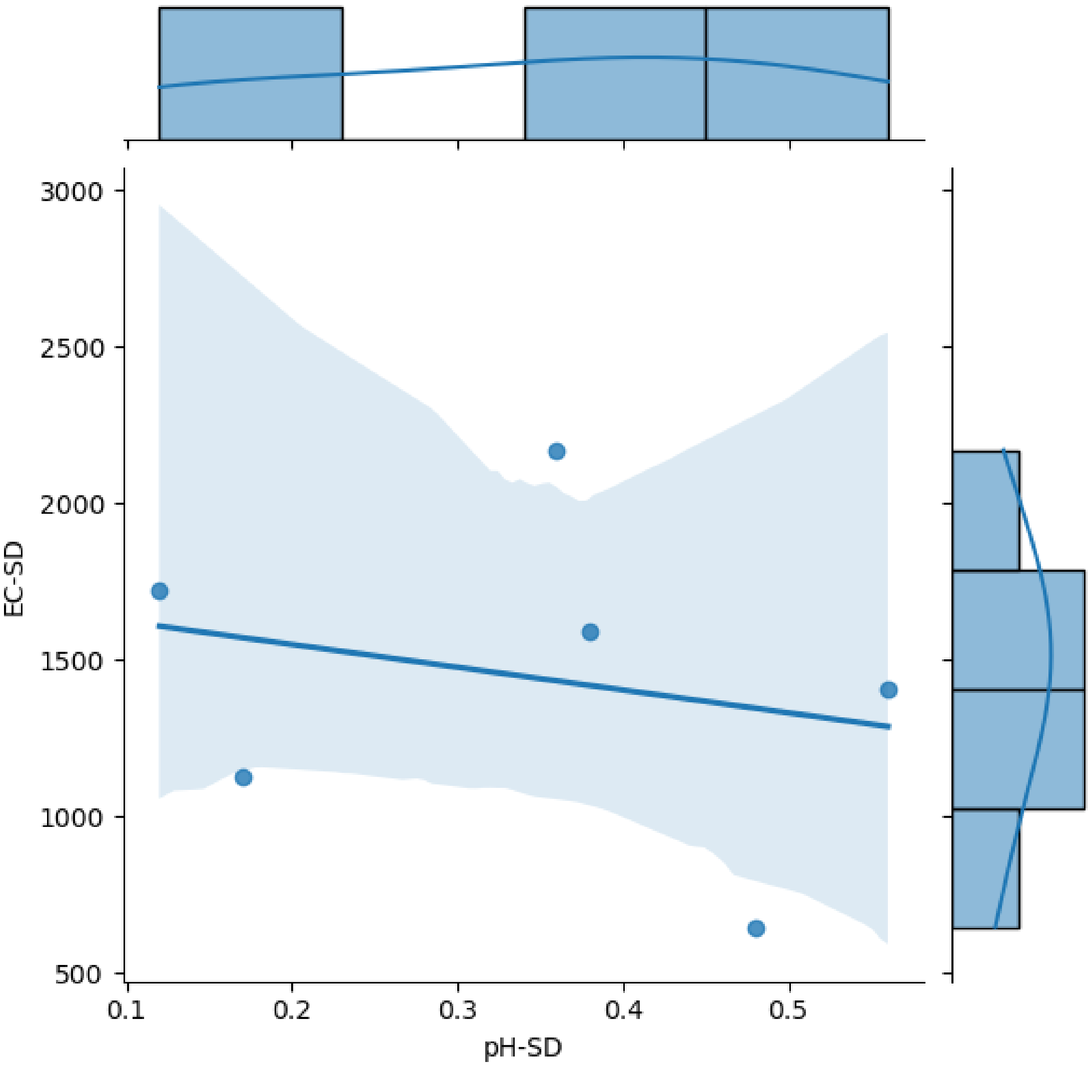


Standard Deviation values Comparison

Potential of Hydrogen VS Electrical Conductivity

```
PearsonRResult(statistic=-0.23937439009406186, pvalue=0.6477965029401975)
```

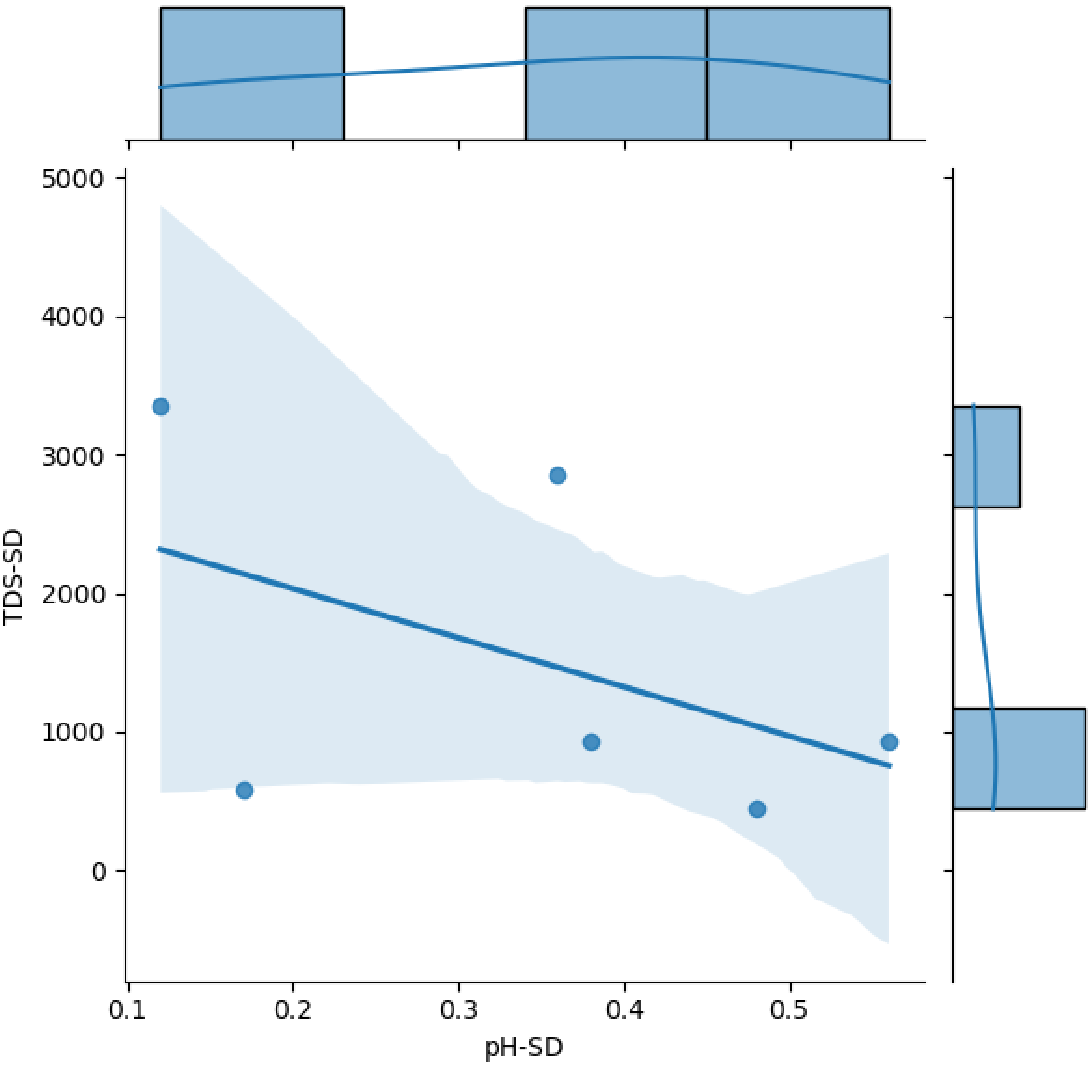
```
 <seaborn.axisgrid.JointGrid at 0x170db8d34c0>
```



Potential of Hydrogen VS Total Dissolved Solids

```
PearsonRResult(statistic=-0.48459389393723384, pvalue=0.3300080520935927)
```

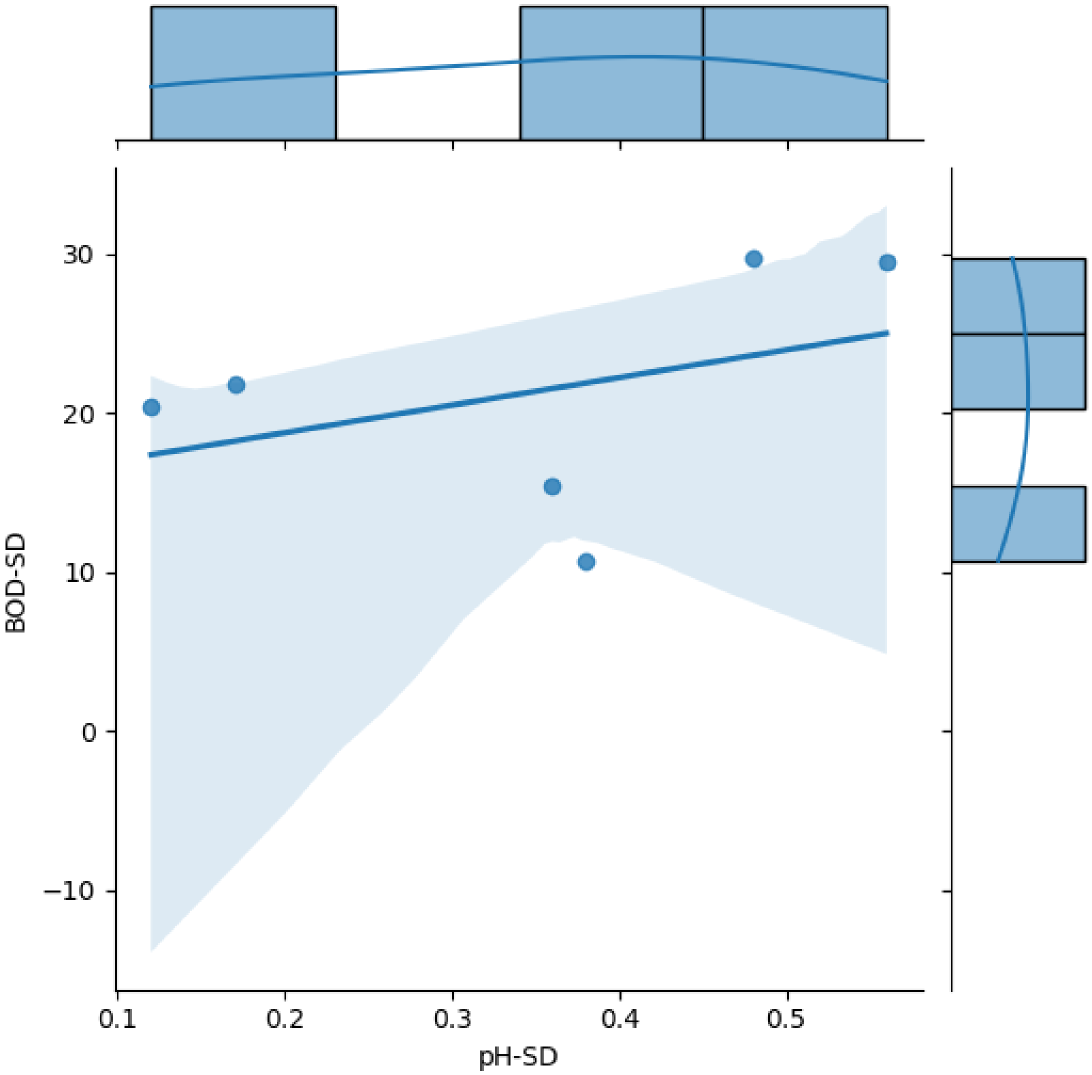
```
 <seaborn.axisgrid.JointGrid at 0x170db93d270>
```



Potential of Hydrogen VS Biological Oxygen Demand

```
PearsonRResult(statistic=0.39345150640909626, pvalue=0.44027669117675816)
```

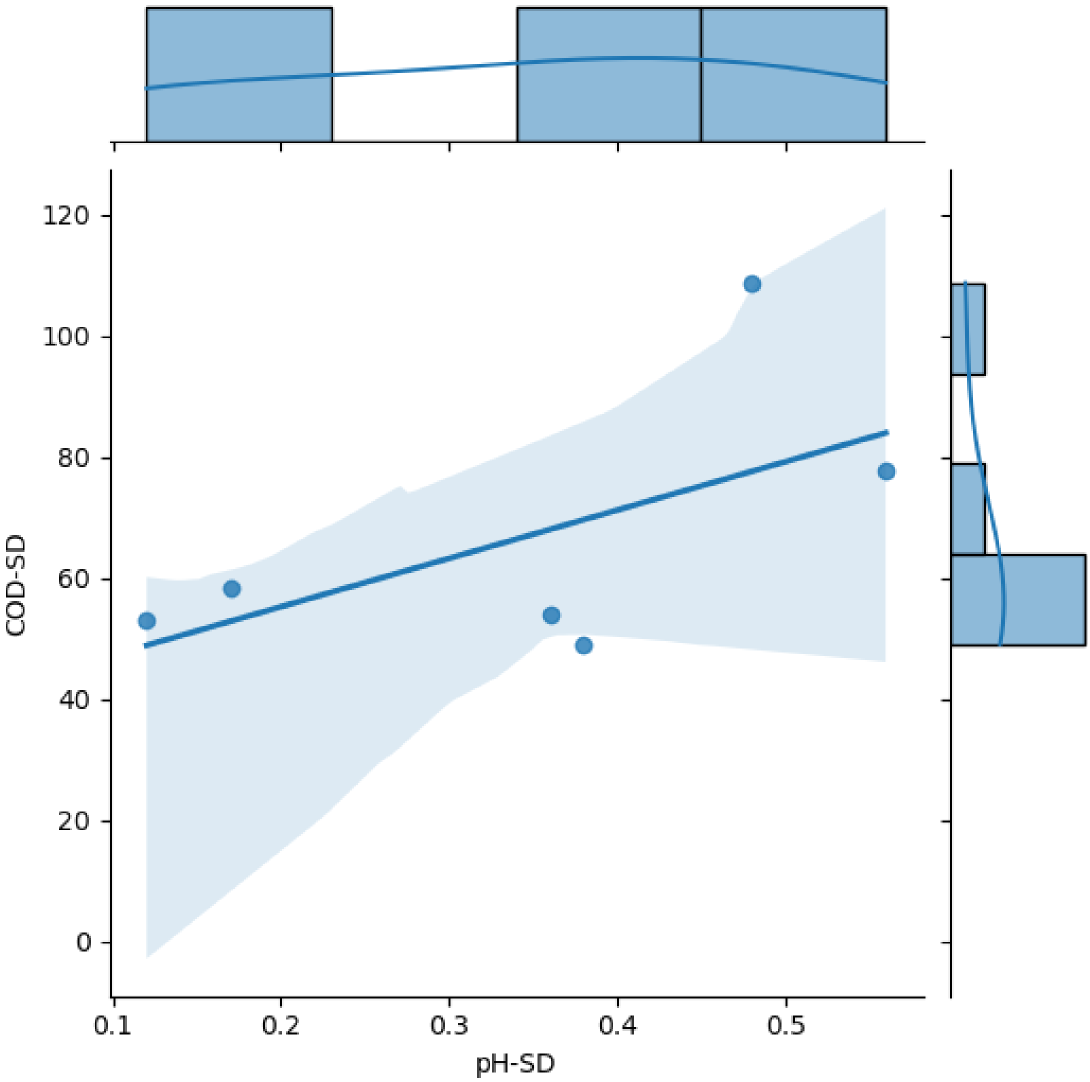
```
 <seaborn.axisgrid.JointGrid at 0x170dafcb8e0>
```



Potential of Hydrogen VS Chemical Oxygen Demand

```
PearsonRResult(statistic=0.5978021544725063, pvalue=0.2101142738704866)
```

```
 <seaborn.axisgrid.JointGrid at 0x170dc4cc5b0>
```


120
100
80
60
40
20
0
COD-SD
0.1
0.2
0.3
0.4
0.5
pH-SD