```
pip install pandas numpy matplotlib seaborn scikit-learn
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (2.2.2)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (2.0.2)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.11/dist-packages (3.10.0)
Requirement already satisfied: seaborn in /usr/local/lib/python3.11/dist-packages (0.13.2)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (1.6.1)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas) (2025.2)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (1.3.2)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (4.58.2)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (24.2)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (11.2.1)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (3.2.3)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.15.3)
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (1.5.1)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn) (3.6.0)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas) (1.17.0)
```

```
import pandas as pd
import numpy as np

from sklearn.multioutput import MultiOutputRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
```

```
df=pd.read_csv('/content/afa2e701598d20110228.csv', sep=';')
df
```

	id	date	NH4	BSK5	Suspended	O2	NO3	NO2	SO4	PO4	CL
0	1	17.02.2000	0.330	2.77	12.0	12.30	9.50	0.057	154.00	0.454	289.50
1	1	11.05.2000	0.044	3.00	51.6	14.61	17.75	0.034	352.00	0.090	1792.00
2	1	11.09.2000	0.032	2.10	24.5	9.87	13.80	0.173	416.00	0.200	2509.00
3	1	13.12.2000	0.170	2.23	35.6	12.40	17.13	0.099	275.20	0.377	1264.00
4	1	02.03.2001	0.000	3.03	48.8	14.69	10.00	0.065	281.60	0.134	1462.00
...	...	...	...	...	...	...	...	...	...	...	...
2856	22	06.10.2020	0.046	2.69	3.6	8.28	3.80	0.038	160.00	0.726	77.85
2857	22	27.10.2020	0.000	1.52	0.5	11.26	0.56	0.031	147.20	0.634	71.95
2858	22	03.12.2020	0.034	0.29	0.8	11.09	2.58	0.042	209.92	0.484	61.17
2859	22	12.01.2021	0.000	2.10	0.0	14.31	3.94	0.034	121.60	0.424	63.49
2860	22	10.02.2021	0.000	1.78	0.0	14.30	6.30	0.033	134.40	0.582	66.31

2861 rows × 11 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2861 entries, 0 to 2860
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         2861 non-null   int64
 1   date       2861 non-null   object
 2   NH4        2858 non-null   float64
 3   BSK5       2860 non-null   float64
 4   Suspended  2845 non-null   float64
 5   O2         2858 non-null   float64
 6   NO3        2860 non-null   float64
 7   NO2        2858 non-null   float64
 8   SO4        2812 non-null   float64
 9   PO4        2833 non-null   float64
 10  CL         2812 non-null   float64
dtypes: float64(9), int64(1), object(1)
memory usage: 246.0+ KB
```

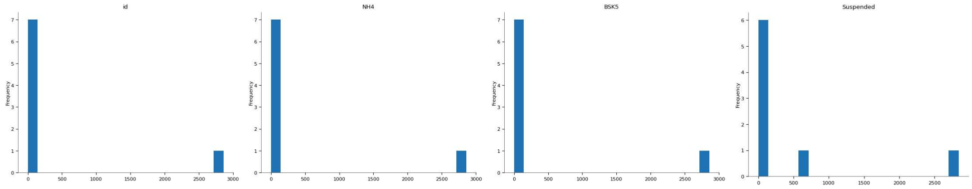
```
df.shape
```

```
(2861, 11)
```

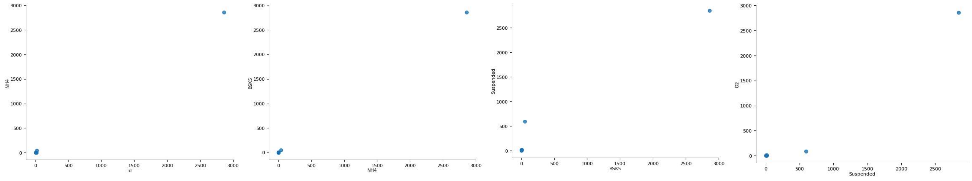
```
df.describe()
```

	id	NH4	BSK5	Suspended	O2	NO3	NO2	SO4	PO4	CL
count	2861.000000	2858.000000	2860.000000	2845.000000	2858.000000	2860.000000	2858.000000	2812.000000	2833.000000	2812.000000
mean	12.397064	0.758734	4.316182	12.931905	9.508902	4.316846	0.246128	59.362313	0.418626	93.731991
std	6.084226	2.486247	2.973997	16.543097	4.428260	6.881188	2.182777	96.582641	0.771326	394.512184
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.020000
25%	8.000000	0.080000	2.160000	6.000000	7.092500	1.390000	0.030000	27.052500	0.130000	26.800000
50%	14.000000	0.220000	3.800000	10.000000	8.995000	2.800000	0.059000	37.800000	0.270000	33.900000
75%	16.000000	0.500000	5.800000	15.000000	11.520000	5.582500	0.125750	64.640000	0.470000	45.607500
max	22.000000	39.427000	50.900000	595.000000	90.000000	133.400000	109.000000	3573.400000	13.879000	5615.280000

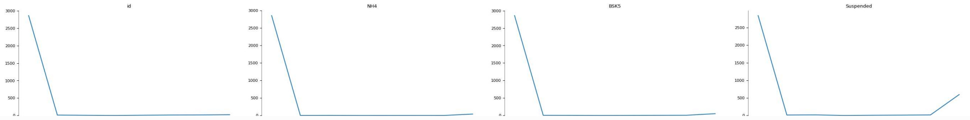
Distributions



2-d distributions



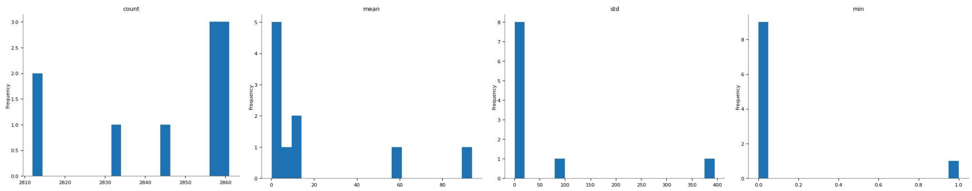
Values



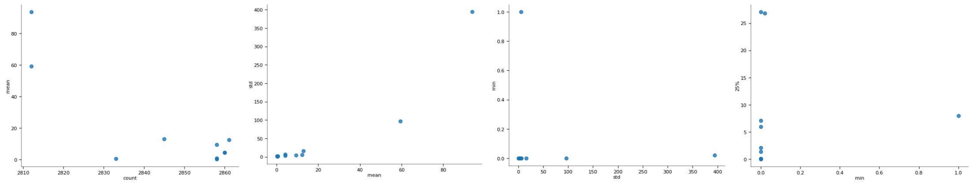
```
df.describe().T
```

```
           count       mean         std   min      25%     50%       75%       max
id        2861.0  12.397064    6.084226  1.00   8.0000  14.000  16.00000    22.000
NH4       2858.0   0.758734    2.486247  0.00   0.0800   0.220   0.50000    39.427
BSK5      2860.0   4.316182    2.973997  0.00   2.1600   3.800   5.80000    50.900
Suspended 2845.0  12.931905   16.543097  0.00   6.0000  10.000  15.00000   595.000
O2        2858.0   9.508902    4.428260  0.00   7.0925   8.995  11.52000    90.000
NO3       2860.0   4.316846    6.881188  0.00   1.3900   2.800   5.58250   133.400
NO2       2858.0   0.246128    2.182777  0.00   0.0300   0.059   0.12575   109.000
SO4       2812.0  59.362313   96.582641  0.00  27.0525  37.800  64.64000  3573.400
PO4       2833.0   0.418626    0.771326  0.00   0.1300   0.270   0.47000    13.879
CL        2812.0  93.731991  394.512184  0.02  26.8000  33.900  45.60750  5615.280
```

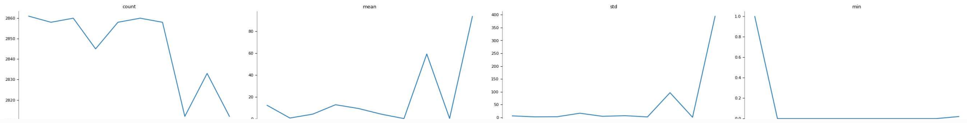
Distributions



2-d distributions



Values



```
df.isnull().sum()
```

```
           0
id         0
date       0
NH4        3
BSK5       1
Suspended 16
O2         3
NO3        1
NO2        3
SO4       49
PO4       28
CL        49
```

```
# Convert the 'date' column to datetime objects, using the correct format
df['date'] = pd.to_datetime(df['date'], format='%d.%m.%Y')
df
```

	id	date	NH4	BSK5	Suspended	O2	NO3	NO2	SO4	PO4	CL
0	1	2000-02-17	0.330	2.77	12.0	12.30	9.50	0.057	154.00	0.454	289.50
1	1	2000-05-11	0.044	3.00	51.6	14.61	17.75	0.034	352.00	0.090	1792.00
2	1	2000-09-11	0.032	2.10	24.5	9.87	13.80	0.173	416.00	0.200	2509.00
3	1	2000-12-13	0.170	2.23	35.6	12.40	17.13	0.099	275.20	0.377	1264.00
4	1	2001-03-02	0.000	3.03	48.8	14.69	10.00	0.065	281.60	0.134	1462.00
...	...	...	...	...	...	...	...	...	...	...	...
2856	22	2020-10-06	0.046	2.69	3.6	8.28	3.80	0.038	160.00	0.726	77.85
2857	22	2020-10-27	0.000	1.52	0.5	11.26	0.56	0.031	147.20	0.634	71.95
2858	22	2020-12-03	0.034	0.29	0.8	11.09	2.58	0.042	209.92	0.484	61.17
2859	22	2021-01-12	0.000	2.10	0.0	14.31	3.94	0.034	121.60	0.424	63.49
2860	22	2021-02-10	0.000	1.78	0.0	14.30	6.30	0.033	134.40	0.582	66.31

2861 rows × 11 columns

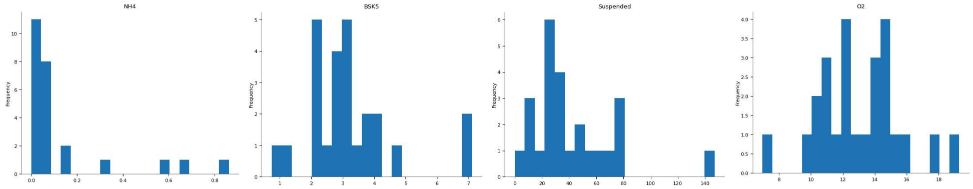
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2861 entries, 0 to 2860
Data columns (total 11 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         2861 non-null   int64
 1   date       2861 non-null   datetime64[ns]
 2   NH4        2858 non-null   float64
 3   BSK5       2860 non-null   float64
 4   Suspended  2845 non-null   float64
 5   O2         2858 non-null   float64
 6   NO3        2860 non-null   float64
 7   NO2        2858 non-null   float64
 8   SO4        2812 non-null   float64
 9   PO4        2833 non-null   float64
 10  CL         2812 non-null   float64
dtypes: datetime64[ns](1), float64(9), int64(1)
memory usage: 246.0 KB
```

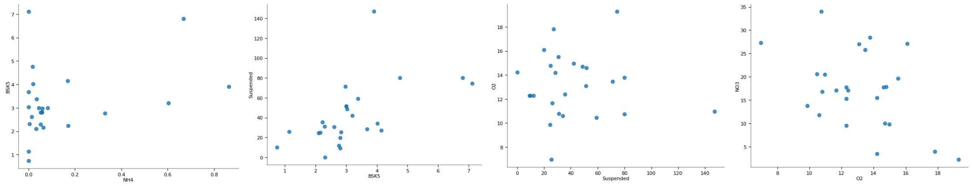
```
df = df.sort_values(by=['id','date'])
df.head(25)
```

```
0   1  2000-02-17  0.330  2.77   12.0  12.30   9.500  0.057  154.00  0.454   289.50
1   1  2000-05-11  0.044  3.00   51.6  14.61  17.750  0.034  352.00  0.090  1792.00
2   1  2000-09-11  0.032  2.10   24.5   9.87  13.800  0.173  416.00  0.200  2509.00
3   1  2000-12-13  0.170  2.23   35.6  12.40  17.130  0.099  275.20  0.377  1264.00
4   1  2001-03-02  0.000  3.03   48.8  14.69  10.000  0.065  281.60  0.134  1462.00
5   1  2001-06-07  0.020  4.02   34.0  10.61  11.800  0.016  287.00  0.208  1183.00
6   1  2001-09-10  0.863  3.91  147.0  10.96  20.500  0.284  595.20  0.674  4023.00
7   1  2001-11-06  0.060  2.97   71.2  13.47  25.800  0.095  314.00  0.390  1907.00
8   1  2002-03-12  0.168  4.15   27.0  17.82   3.945  0.058  153.60  0.110   473.00
9   1  2002-06-06  0.001  7.11   74.4  19.28   2.260  0.017  409.60  0.181  1782.00
10  1  2002-07-15  0.668  6.81   80.0  10.74  34.000  0.020  505.60  0.222  3160.00
11  1  2002-11-07  0.082  3.00   51.4  13.08  27.050  0.092  314.00  0.179  1907.00
12  1  2003-02-24  0.603  3.20   42.0  14.97   9.870  0.169  160.00  0.130   403.70
13  1  2003-06-13  0.058  2.80    9.3  12.30  15.300  0.060  500.00  0.300  2325.00
14  1  2003-09-16  0.053  2.30   31.2  10.80  16.800  0.078  498.50  0.380  2375.00
15  1  2003-10-14  0.017  4.76   80.0  13.77  28.400  0.048  518.40  0.371  3304.50
16  1  2004-03-15  0.014  2.61   30.8  15.51  19.610  0.052  557.80  0.150  1075.00
17  1  2004-06-10  0.035  3.38   59.2  10.47  20.610  0.081  708.40  0.200  2122.83
18  1  2004-09-06  0.057  2.83   25.6   6.97  27.300  0.120  416.16  0.780  2669.60
19  1  2004-12-14  0.053  2.80   19.9  16.10  27.080  0.130  540.60  0.820  2425.00
20  1  2005-01-17  0.000  1.14   26.0  11.66  17.100  0.043  262.40  0.530   836.60
21  1  2005-05-14  0.004  2.31    0.0  14.21   3.480  0.044  183.40  0.000   463.30
22  1  2005-09-02  0.000  3.68   28.5  14.20  15.510  0.033  422.40  0.900  2063.80
23  1  2005-12-22  0.063  2.16   24.8  14.78  17.900  0.052  288.00  0.380  1291.60
24  1  2006-03-20  0.000  0.74   10.0  12.30  17.730  0.087  224.00  0.270   489.00
```

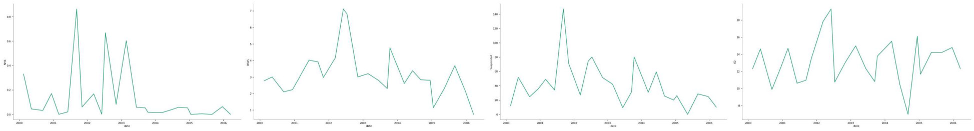
Distributions



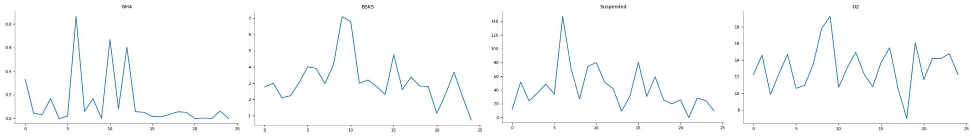
2-d distributions



Time series



Values



Next steps: Generate code with df View recommended plots New interactive sheet

```
df['year']=df['date'].dt.year
df['month']=df['date'].dt.month
```

```
df.head()
```

	id	date	NH4	BSK5	Suspended	O2	NO3	NO2	SO4	PO4	CL	year	month	day
0	1	2000-02-17	0.330	2.77	12.0	12.30	9.50	0.057	154.0	0.454	289.5	2000	2	17
1	1	2000-05-11	0.044	3.00	51.6	14.61	17.75	0.034	352.0	0.090	1792.0	2000	5	11
2	1	2000-09-11	0.032	2.10	24.5	9.87	13.80	0.173	416.0	0.200	2509.0	2000	9	11
3	1	2000-12-13	0.170	2.23	35.6	12.40	17.13	0.099	275.2	0.377	1264.0	2000	12	13
4	1	2001-03-02	0.000	3.03	48.8	14.69	10.00	0.065	281.6	0.134	1462.0	2001	3	2

Distributions