{

"cells": [

{

"cell\_type": "code",

"execution\_count": "null",

"id": "b360abbe-e87d-4c36-af38-b35bd48d2517",

"metadata": {},

"outputs": [],

"source": [

"# To install basic/necessary libraries\n",

"pip install pandas numpy matplotlib seaborn scikit-learn"

]

},

{

"cell\_type": "code",

"execution\_count": 1,

"id": "b024a389-e6d4-4e96-a75f-1d5ded1c4482",

"metadata": {},

"outputs": [],

"source": [

"# Import necessary libraries\n",

"import pandas as pd # data manipulation\n",

"import numpy as np # numerical python - linear algebra\n",

"\n",

"from sklearn.multioutput import MultiOutputRegressor\n",

"from sklearn.ensemble import RandomForestRegressor\n",

"from sklearn.model\_selection import train\_test\_split\n",

"from sklearn.metrics import mean\_squared\_error, r2\_score"

]

},

{

"cell\_type": "code",

"execution\_count": 2,

"id": "cbae76f5-2560-439f-b42a-7ee340766c00",

"metadata": {},

"outputs": [

{

"data": {

"text/html": [

"<div>\n",

"<style scoped>\n",

" .dataframe tbody tr th:only-of-type {\n",

" vertical-align: middle;\n",

" }\n",

"\n",

" .dataframe tbody tr th {\n",

" vertical-align: top;\n",

" }\n",

"\n",

" .dataframe thead th {\n",

" text-align: right;\n",

" }\n",

"</style>\n",

"<table border=\"1\" class=\"dataframe\">\n",

" <thead>\n",

" <tr style=\"text-align: right;\">\n",

" <th></th>\n",

" <th>id</th>\n",

" <th>date</th>\n",

" <th>NH4</th>\n",

" <th>BSK5</th>\n",

" <th>Suspended</th>\n",

" <th>O2</th>\n",

" <th>NO3</th>\n",

" <th>NO2</th>\n",

" <th>SO4</th>\n",

" <th>PO4</th>\n",

" <th>CL</th>\n",

" </tr>\n",

" </thead>\n",

" <tbody>\n",

" <tr>\n",

" <th>0</th>\n",

" <td>1</td>\n",

" <td>17.02.2000</td>\n",

" <td>0.330</td>\n",

" <td>2.77</td>\n",

" <td>12.0</td>\n",

" <td>12.30</td>\n",

" <td>9.50</td>\n",

" <td>0.057</td>\n",

" <td>154.00</td>\n",

" <td>0.454</td>\n",

" <td>289.50</td>\n",

" </tr>\n",

" <tr>\n",

" <th>1</th>\n",

" <td>1</td>\n",

" <td>11.05.2000</td>\n",

" <td>0.044</td>\n",

" <td>3.00</td>\n",

" <td>51.6</td>\n",

" <td>14.61</td>\n",

" <td>17.75</td>\n",

" <td>0.034</td>\n",

" <td>352.00</td>\n",

" <td>0.090</td>\n",

" <td>1792.00</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2</th>\n",

" <td>1</td>\n",

" <td>11.09.2000</td>\n",

" <td>0.032</td>\n",

" <td>2.10</td>\n",

" <td>24.5</td>\n",

" <td>9.87</td>\n",

" <td>13.80</td>\n",

" <td>0.173</td>\n",

" <td>416.00</td>\n",

" <td>0.200</td>\n",

" <td>2509.00</td>\n",

" </tr>\n",

" <tr>\n",

" <th>3</th>\n",

" <td>1</td>\n",

" <td>13.12.2000</td>\n",

" <td>0.170</td>\n",

" <td>2.23</td>\n",

" <td>35.6</td>\n",

" <td>12.40</td>\n",

" <td>17.13</td>\n",

" <td>0.099</td>\n",

" <td>275.20</td>\n",

" <td>0.377</td>\n",

" <td>1264.00</td>\n",

" </tr>\n",

" <tr>\n",

" <th>4</th>\n",

" <td>1</td>\n",

" <td>02.03.2001</td>\n",

" <td>0.000</td>\n",

" <td>3.03</td>\n",

" <td>48.8</td>\n",

" <td>14.69</td>\n",

" <td>10.00</td>\n",

" <td>0.065</td>\n",

" <td>281.60</td>\n",

" <td>0.134</td>\n",

" <td>1462.00</td>\n",

" </tr>\n",

" <tr>\n",

" <th>...</th>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2856</th>\n",

" <td>22</td>\n",

" <td>06.10.2020</td>\n",

" <td>0.046</td>\n",

" <td>2.69</td>\n",

" <td>3.6</td>\n",

" <td>8.28</td>\n",

" <td>3.80</td>\n",

" <td>0.038</td>\n",

" <td>160.00</td>\n",

" <td>0.726</td>\n",

" <td>77.85</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2857</th>\n",

" <td>22</td>\n",

" <td>27.10.2020</td>\n",

" <td>0.000</td>\n",

" <td>1.52</td>\n",

" <td>0.5</td>\n",

" <td>11.26</td>\n",

" <td>0.56</td>\n",

" <td>0.031</td>\n",

" <td>147.20</td>\n",

" <td>0.634</td>\n",

" <td>71.95</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2858</th>\n",

" <td>22</td>\n",

" <td>03.12.2020</td>\n",

" <td>0.034</td>\n",

" <td>0.29</td>\n",

" <td>0.8</td>\n",

" <td>11.09</td>\n",

" <td>2.58</td>\n",

" <td>0.042</td>\n",

" <td>209.92</td>\n",

" <td>0.484</td>\n",

" <td>61.17</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2859</th>\n",

" <td>22</td>\n",

" <td>12.01.2021</td>\n",

" <td>0.000</td>\n",

" <td>2.10</td>\n",

" <td>0.0</td>\n",

" <td>14.31</td>\n",

" <td>3.94</td>\n",

" <td>0.034</td>\n",

" <td>121.60</td>\n",

" <td>0.424</td>\n",

" <td>63.49</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2860</th>\n",

" <td>22</td>\n",

" <td>10.02.2021</td>\n",

" <td>0.000</td>\n",

" <td>1.78</td>\n",

" <td>0.0</td>\n",

" <td>14.30</td>\n",

" <td>6.30</td>\n",

" <td>0.033</td>\n",

" <td>134.40</td>\n",

" <td>0.582</td>\n",

" <td>66.31</td>\n",

" </tr>\n",

" </tbody>\n",

"</table>\n",

"<p>2861 rows × 11 columns</p>\n",

"</div>"

],

"text/plain": [

" id date NH4 BSK5 ... NO2 SO4 PO4 CL\n",

"0 1 17.02.2000 0.330 2.77 ... 0.057 154.00 0.454 289.50\n",

"1 1 11.05.2000 0.044 3.00 ... 0.034 352.00 0.090 1792.00\n",

"2 1 11.09.2000 0.032 2.10 ... 0.173 416.00 0.200 2509.00\n",

"3 1 13.12.2000 0.170 2.23 ... 0.099 275.20 0.377 1264.00\n",

"4 1 02.03.2001 0.000 3.03 ... 0.065 281.60 0.134 1462.00\n",

"... .. ... ... ... ... ... ... ... ...\n",

"2856 22 06.10.2020 0.046 2.69 ... 0.038 160.00 0.726 77.85\n",

"2857 22 27.10.2020 0.000 1.52 ... 0.031 147.20 0.634 71.95\n",

"2858 22 03.12.2020 0.034 0.29 ... 0.042 209.92 0.484 61.17\n",

"2859 22 12.01.2021 0.000 2.10 ... 0.034 121.60 0.424 63.49\n",

"2860 22 10.02.2021 0.000 1.78 ... 0.033 134.40 0.582 66.31\n",

"\n",

"[2861 rows x 11 columns]"

]

},

"execution\_count": 2,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"# load the dataset\n",

"df = pd.read\_csv('PB\_All\_2000\_2021.csv', sep=';')\n",

"df"

]

},

{

"cell\_type": "code",

"execution\_count": 3,

"id": "f0695fd7-6929-4c69-8898-e72896900ce7",

"metadata": {},

"outputs": [

{

"name": "stdout",

"output\_type": "stream",

"text": [

"<class 'pandas.core.frame.DataFrame'>\n",

"RangeIndex: 2861 entries, 0 to 2860\n",

"Data columns (total 11 columns):\n",

" # Column Non-Null Count Dtype \n",

"--- ------ -------------- ----- \n",

" 0 id 2861 non-null int64 \n",

" 1 date 2861 non-null object \n",

" 2 NH4 2858 non-null float64\n",

" 3 BSK5 2860 non-null float64\n",

" 4 Suspended 2845 non-null float64\n",

" 5 O2 2858 non-null float64\n",

" 6 NO3 2860 non-null float64\n",

" 7 NO2 2858 non-null float64\n",

" 8 SO4 2812 non-null float64\n",

" 9 PO4 2833 non-null float64\n",

" 10 CL 2812 non-null float64\n",

"dtypes: float64(9), int64(1), object(1)\n",

"memory usage: 246.0+ KB\n"

]

}

],

"source": [

"df.info() # dataset info"

]

},

{

"cell\_type": "code",

"execution\_count": 4,

"id": "21feb067-09f5-416b-a9a5-43b43cd8d831",

"metadata": {},

"outputs": [

{

"data": {

"text/plain": [

"(2861, 11)"

]

},

"execution\_count": 4,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"# rows and cols\n",

"df.shape"

]

},

{

"cell\_type": "code",

"execution\_count": 6,

"id": "b0e2f909-3b28-4ffd-87ff-689a8ab4d14d",

"metadata": {},

"outputs": [

{

"data": {

"text/html": [

"<div>\n",

"<style scoped>\n",

" .dataframe tbody tr th:only-of-type {\n",

" vertical-align: middle;\n",

" }\n",

"\n",

" .dataframe tbody tr th {\n",

" vertical-align: top;\n",

" }\n",

"\n",

" .dataframe thead th {\n",

" text-align: right;\n",

" }\n",

"</style>\n",

"<table border=\"1\" class=\"dataframe\">\n",

" <thead>\n",

" <tr style=\"text-align: right;\">\n",

" <th></th>\n",

" <th>count</th>\n",

" <th>mean</th>\n",

" <th>std</th>\n",

" <th>min</th>\n",

" <th>25%</th>\n",

" <th>50%</th>\n",

" <th>75%</th>\n",

" <th>max</th>\n",

" </tr>\n",

" </thead>\n",

" <tbody>\n",

" <tr>\n",

" <th>id</th>\n",

" <td>2861.0</td>\n",

" <td>12.397064</td>\n",

" <td>6.084226</td>\n",

" <td>1.00</td>\n",

" <td>8.0000</td>\n",

" <td>14.000</td>\n",

" <td>16.00000</td>\n",

" <td>22.000</td>\n",

" </tr>\n",

" <tr>\n",

" <th>NH4</th>\n",

" <td>2858.0</td>\n",

" <td>0.758734</td>\n",

" <td>2.486247</td>\n",

" <td>0.00</td>\n",

" <td>0.0800</td>\n",

" <td>0.220</td>\n",

" <td>0.50000</td>\n",

" <td>39.427</td>\n",

" </tr>\n",

" <tr>\n",

" <th>BSK5</th>\n",

" <td>2860.0</td>\n",

" <td>4.316182</td>\n",

" <td>2.973997</td>\n",

" <td>0.00</td>\n",

" <td>2.1600</td>\n",

" <td>3.800</td>\n",

" <td>5.80000</td>\n",

" <td>50.900</td>\n",

" </tr>\n",

" <tr>\n",

" <th>Suspended</th>\n",

" <td>2845.0</td>\n",

" <td>12.931905</td>\n",

" <td>16.543097</td>\n",

" <td>0.00</td>\n",

" <td>6.0000</td>\n",

" <td>10.000</td>\n",

" <td>15.00000</td>\n",

" <td>595.000</td>\n",

" </tr>\n",

" <tr>\n",

" <th>O2</th>\n",

" <td>2858.0</td>\n",

" <td>9.508902</td>\n",

" <td>4.428260</td>\n",

" <td>0.00</td>\n",

" <td>7.0925</td>\n",

" <td>8.995</td>\n",

" <td>11.52000</td>\n",

" <td>90.000</td>\n",

" </tr>\n",

" <tr>\n",

" <th>NO3</th>\n",

" <td>2860.0</td>\n",

" <td>4.316846</td>\n",

" <td>6.881188</td>\n",

" <td>0.00</td>\n",

" <td>1.3900</td>\n",

" <td>2.800</td>\n",

" <td>5.58250</td>\n",

" <td>133.400</td>\n",

" </tr>\n",

" <tr>\n",

" <th>NO2</th>\n",

" <td>2858.0</td>\n",

" <td>0.246128</td>\n",

" <td>2.182777</td>\n",

" <td>0.00</td>\n",

" <td>0.0300</td>\n",

" <td>0.059</td>\n",

" <td>0.12575</td>\n",

" <td>109.000</td>\n",

" </tr>\n",

" <tr>\n",

" <th>SO4</th>\n",

" <td>2812.0</td>\n",

" <td>59.362313</td>\n",

" <td>96.582641</td>\n",

" <td>0.00</td>\n",

" <td>27.0525</td>\n",

" <td>37.800</td>\n",

" <td>64.64000</td>\n",

" <td>3573.400</td>\n",

" </tr>\n",

" <tr>\n",

" <th>PO4</th>\n",

" <td>2833.0</td>\n",

" <td>0.418626</td>\n",

" <td>0.771326</td>\n",

" <td>0.00</td>\n",

" <td>0.1300</td>\n",

" <td>0.270</td>\n",

" <td>0.47000</td>\n",

" <td>13.879</td>\n",

" </tr>\n",

" <tr>\n",

" <th>CL</th>\n",

" <td>2812.0</td>\n",

" <td>93.731991</td>\n",

" <td>394.512184</td>\n",

" <td>0.02</td>\n",

" <td>26.8000</td>\n",

" <td>33.900</td>\n",

" <td>45.60750</td>\n",

" <td>5615.280</td>\n",

" </tr>\n",

" </tbody>\n",

"</table>\n",

"</div>"

],

"text/plain": [

" count mean std ... 50% 75% max\n",

"id 2861.0 12.397064 6.084226 ... 14.000 16.00000 22.000\n",

"NH4 2858.0 0.758734 2.486247 ... 0.220 0.50000 39.427\n",

"BSK5 2860.0 4.316182 2.973997 ... 3.800 5.80000 50.900\n",

"Suspended 2845.0 12.931905 16.543097 ... 10.000 15.00000 595.000\n",

"O2 2858.0 9.508902 4.428260 ... 8.995 11.52000 90.000\n",

"NO3 2860.0 4.316846 6.881188 ... 2.800 5.58250 133.400\n",

"NO2 2858.0 0.246128 2.182777 ... 0.059 0.12575 109.000\n",

"SO4 2812.0 59.362313 96.582641 ... 37.800 64.64000 3573.400\n",

"PO4 2833.0 0.418626 0.771326 ... 0.270 0.47000 13.879\n",

"CL 2812.0 93.731991 394.512184 ... 33.900 45.60750 5615.280\n",

"\n",

"[10 rows x 8 columns]"

]

},

"execution\_count": 6,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"# Statistics of the data\n",

"df.describe().T"

]

},

{

"cell\_type": "code",

"execution\_count": 7,

"id": "5f39d8e1-2f68-478f-8d64-838bcdbd2bb3",

"metadata": {},

"outputs": [

{

"data": {

"text/plain": [

"id 0\n",

"date 0\n",

"NH4 3\n",

"BSK5 1\n",

"Suspended 16\n",

"O2 3\n",

"NO3 1\n",

"NO2 3\n",

"SO4 49\n",

"PO4 28\n",

"CL 49\n",

"dtype: int64"

]

},

"execution\_count": 7,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"# Missing values\n",

"df.isnull().sum()"

]

},

{

"cell\_type": "code",

"execution\_count": 8,

"id": "d146fdb3-998c-461a-84d5-b45af90dffe0",

"metadata": {},

"outputs": [

{

"data": {

"text/html": [

"<div>\n",

"<style scoped>\n",

" .dataframe tbody tr th:only-of-type {\n",

" vertical-align: middle;\n",

" }\n",

"\n",

" .dataframe tbody tr th {\n",

" vertical-align: top;\n",

" }\n",

"\n",

" .dataframe thead th {\n",

" text-align: right;\n",

" }\n",

"</style>\n",

"<table border=\"1\" class=\"dataframe\">\n",

" <thead>\n",

" <tr style=\"text-align: right;\">\n",

" <th></th>\n",

" <th>id</th>\n",

" <th>date</th>\n",

" <th>NH4</th>\n",

" <th>BSK5</th>\n",

" <th>Suspended</th>\n",

" <th>O2</th>\n",

" <th>NO3</th>\n",

" <th>NO2</th>\n",

" <th>SO4</th>\n",

" <th>PO4</th>\n",

" <th>CL</th>\n",

" </tr>\n",

" </thead>\n",

" <tbody>\n",

" <tr>\n",

" <th>0</th>\n",

" <td>1</td>\n",

" <td>2000-02-17</td>\n",

" <td>0.330</td>\n",

" <td>2.77</td>\n",

" <td>12.0</td>\n",

" <td>12.30</td>\n",

" <td>9.50</td>\n",

" <td>0.057</td>\n",

" <td>154.00</td>\n",

" <td>0.454</td>\n",

" <td>289.50</td>\n",

" </tr>\n",

" <tr>\n",

" <th>1</th>\n",

" <td>1</td>\n",

" <td>2000-05-11</td>\n",

" <td>0.044</td>\n",

" <td>3.00</td>\n",

" <td>51.6</td>\n",

" <td>14.61</td>\n",

" <td>17.75</td>\n",

" <td>0.034</td>\n",

" <td>352.00</td>\n",

" <td>0.090</td>\n",

" <td>1792.00</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2</th>\n",

" <td>1</td>\n",

" <td>2000-09-11</td>\n",

" <td>0.032</td>\n",

" <td>2.10</td>\n",

" <td>24.5</td>\n",

" <td>9.87</td>\n",

" <td>13.80</td>\n",

" <td>0.173</td>\n",

" <td>416.00</td>\n",

" <td>0.200</td>\n",

" <td>2509.00</td>\n",

" </tr>\n",

" <tr>\n",

" <th>3</th>\n",

" <td>1</td>\n",

" <td>2000-12-13</td>\n",

" <td>0.170</td>\n",

" <td>2.23</td>\n",

" <td>35.6</td>\n",

" <td>12.40</td>\n",

" <td>17.13</td>\n",

" <td>0.099</td>\n",

" <td>275.20</td>\n",

" <td>0.377</td>\n",

" <td>1264.00</td>\n",

" </tr>\n",

" <tr>\n",

" <th>4</th>\n",

" <td>1</td>\n",

" <td>2001-03-02</td>\n",

" <td>0.000</td>\n",

" <td>3.03</td>\n",

" <td>48.8</td>\n",

" <td>14.69</td>\n",

" <td>10.00</td>\n",

" <td>0.065</td>\n",

" <td>281.60</td>\n",

" <td>0.134</td>\n",

" <td>1462.00</td>\n",

" </tr>\n",

" <tr>\n",

" <th>...</th>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" <td>...</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2856</th>\n",

" <td>22</td>\n",

" <td>2020-10-06</td>\n",

" <td>0.046</td>\n",

" <td>2.69</td>\n",

" <td>3.6</td>\n",

" <td>8.28</td>\n",

" <td>3.80</td>\n",

" <td>0.038</td>\n",

" <td>160.00</td>\n",

" <td>0.726</td>\n",

" <td>77.85</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2857</th>\n",

" <td>22</td>\n",

" <td>2020-10-27</td>\n",

" <td>0.000</td>\n",

" <td>1.52</td>\n",

" <td>0.5</td>\n",

" <td>11.26</td>\n",

" <td>0.56</td>\n",

" <td>0.031</td>\n",

" <td>147.20</td>\n",

" <td>0.634</td>\n",

" <td>71.95</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2858</th>\n",

" <td>22</td>\n",

" <td>2020-12-03</td>\n",

" <td>0.034</td>\n",

" <td>0.29</td>\n",

" <td>0.8</td>\n",

" <td>11.09</td>\n",

" <td>2.58</td>\n",

" <td>0.042</td>\n",

" <td>209.92</td>\n",

" <td>0.484</td>\n",

" <td>61.17</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2859</th>\n",

" <td>22</td>\n",

" <td>2021-01-12</td>\n",

" <td>0.000</td>\n",

" <td>2.10</td>\n",

" <td>0.0</td>\n",

" <td>14.31</td>\n",

" <td>3.94</td>\n",

" <td>0.034</td>\n",

" <td>121.60</td>\n",

" <td>0.424</td>\n",

" <td>63.49</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2860</th>\n",

" <td>22</td>\n",

" <td>2021-02-10</td>\n",

" <td>0.000</td>\n",

" <td>1.78</td>\n",

" <td>0.0</td>\n",

" <td>14.30</td>\n",

" <td>6.30</td>\n",

" <td>0.033</td>\n",

" <td>134.40</td>\n",

" <td>0.582</td>\n",

" <td>66.31</td>\n",

" </tr>\n",

" </tbody>\n",

"</table>\n",

"<p>2861 rows × 11 columns</p>\n",

"</div>"

],

"text/plain": [

" id date NH4 BSK5 ... NO2 SO4 PO4 CL\n",

"0 1 2000-02-17 0.330 2.77 ... 0.057 154.00 0.454 289.50\n",

"1 1 2000-05-11 0.044 3.00 ... 0.034 352.00 0.090 1792.00\n",

"2 1 2000-09-11 0.032 2.10 ... 0.173 416.00 0.200 2509.00\n",

"3 1 2000-12-13 0.170 2.23 ... 0.099 275.20 0.377 1264.00\n",

"4 1 2001-03-02 0.000 3.03 ... 0.065 281.60 0.134 1462.00\n",

"... .. ... ... ... ... ... ... ... ...\n",

"2856 22 2020-10-06 0.046 2.69 ... 0.038 160.00 0.726 77.85\n",

"2857 22 2020-10-27 0.000 1.52 ... 0.031 147.20 0.634 71.95\n",

"2858 22 2020-12-03 0.034 0.29 ... 0.042 209.92 0.484 61.17\n",

"2859 22 2021-01-12 0.000 2.10 ... 0.034 121.60 0.424 63.49\n",

"2860 22 2021-02-10 0.000 1.78 ... 0.033 134.40 0.582 66.31\n",

"\n",

"[2861 rows x 11 columns]"

]

},

"execution\_count": 8,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"# date is in object - date format\n",

"df['date'] = pd.to\_datetime(df['date'], format='%d.%m.%Y')\n",

"df"

]

},

{

"cell\_type": "code",

"execution\_count": 9,

"id": "fa9fc0e4-e6cd-4810-a68f-3acdc76c15e4",

"metadata": {},

"outputs": [

{

"name": "stdout",

"output\_type": "stream",

"text": [

"<class 'pandas.core.frame.DataFrame'>\n",

"RangeIndex: 2861 entries, 0 to 2860\n",

"Data columns (total 11 columns):\n",

" # Column Non-Null Count Dtype \n",

"--- ------ -------------- ----- \n",

" 0 id 2861 non-null int64 \n",

" 1 date 2861 non-null datetime64[ns]\n",

" 2 NH4 2858 non-null float64 \n",

" 3 BSK5 2860 non-null float64 \n",

" 4 Suspended 2845 non-null float64 \n",

" 5 O2 2858 non-null float64 \n",

" 6 NO3 2860 non-null float64 \n",

" 7 NO2 2858 non-null float64 \n",

" 8 SO4 2812 non-null float64 \n",

" 9 PO4 2833 non-null float64 \n",

" 10 CL 2812 non-null float64 \n",

"dtypes: datetime64[ns](1), float64(9), int64(1)\n",

"memory usage: 246.0 KB\n"

]

}

],

"source": [

"df.info()"

]

},

{

"cell\_type": "code",

"execution\_count": 10,

"id": "18b44f69-e8b2-4549-814a-a336db31fa94",

"metadata": {},

"outputs": [

{

"data": {

"text/html": [

"<div>\n",

"<style scoped>\n",

" .dataframe tbody tr th:only-of-type {\n",

" vertical-align: middle;\n",

" }\n",

"\n",

" .dataframe tbody tr th {\n",

" vertical-align: top;\n",

" }\n",

"\n",

" .dataframe thead th {\n",

" text-align: right;\n",

" }\n",

"</style>\n",

"<table border=\"1\" class=\"dataframe\">\n",

" <thead>\n",

" <tr style=\"text-align: right;\">\n",

" <th></th>\n",

" <th>id</th>\n",

" <th>date</th>\n",

" <th>NH4</th>\n",

" <th>BSK5</th>\n",

" <th>Suspended</th>\n",

" <th>O2</th>\n",

" <th>NO3</th>\n",

" <th>NO2</th>\n",

" <th>SO4</th>\n",

" <th>PO4</th>\n",

" <th>CL</th>\n",

" </tr>\n",

" </thead>\n",

" <tbody>\n",

" <tr>\n",

" <th>0</th>\n",

" <td>1</td>\n",

" <td>2000-02-17</td>\n",

" <td>0.330</td>\n",

" <td>2.77</td>\n",

" <td>12.0</td>\n",

" <td>12.30</td>\n",

" <td>9.50</td>\n",

" <td>0.057</td>\n",

" <td>154.0</td>\n",

" <td>0.454</td>\n",

" <td>289.5</td>\n",

" </tr>\n",

" <tr>\n",

" <th>1</th>\n",

" <td>1</td>\n",

" <td>2000-05-11</td>\n",

" <td>0.044</td>\n",

" <td>3.00</td>\n",

" <td>51.6</td>\n",

" <td>14.61</td>\n",

" <td>17.75</td>\n",

" <td>0.034</td>\n",

" <td>352.0</td>\n",

" <td>0.090</td>\n",

" <td>1792.0</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2</th>\n",

" <td>1</td>\n",

" <td>2000-09-11</td>\n",

" <td>0.032</td>\n",

" <td>2.10</td>\n",

" <td>24.5</td>\n",

" <td>9.87</td>\n",

" <td>13.80</td>\n",

" <td>0.173</td>\n",

" <td>416.0</td>\n",

" <td>0.200</td>\n",

" <td>2509.0</td>\n",

" </tr>\n",

" <tr>\n",

" <th>3</th>\n",

" <td>1</td>\n",

" <td>2000-12-13</td>\n",

" <td>0.170</td>\n",

" <td>2.23</td>\n",

" <td>35.6</td>\n",

" <td>12.40</td>\n",

" <td>17.13</td>\n",

" <td>0.099</td>\n",

" <td>275.2</td>\n",

" <td>0.377</td>\n",

" <td>1264.0</td>\n",

" </tr>\n",

" <tr>\n",

" <th>4</th>\n",

" <td>1</td>\n",

" <td>2001-03-02</td>\n",

" <td>0.000</td>\n",

" <td>3.03</td>\n",

" <td>48.8</td>\n",

" <td>14.69</td>\n",

" <td>10.00</td>\n",

" <td>0.065</td>\n",

" <td>281.6</td>\n",

" <td>0.134</td>\n",

" <td>1462.0</td>\n",

" </tr>\n",

" </tbody>\n",

"</table>\n",

"</div>"

],

"text/plain": [

" id date NH4 BSK5 Suspended ... NO3 NO2 SO4 PO4 CL\n",

"0 1 2000-02-17 0.330 2.77 12.0 ... 9.50 0.057 154.0 0.454 289.5\n",

"1 1 2000-05-11 0.044 3.00 51.6 ... 17.75 0.034 352.0 0.090 1792.0\n",

"2 1 2000-09-11 0.032 2.10 24.5 ... 13.80 0.173 416.0 0.200 2509.0\n",

"3 1 2000-12-13 0.170 2.23 35.6 ... 17.13 0.099 275.2 0.377 1264.0\n",

"4 1 2001-03-02 0.000 3.03 48.8 ... 10.00 0.065 281.6 0.134 1462.0\n",

"\n",

"[5 rows x 11 columns]"

]

},

"execution\_count": 10,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"df = df.sort\_values(by=['id', 'date'])\n",

"df.head()"

]

},

{

"cell\_type": "code",

"execution\_count": 11,

"id": "16c26fe3-9d39-47a9-80ee-b1332ecdb1aa",

"metadata": {},

"outputs": [],

"source": [

"df['year'] = df['date'].dt.year\n",

"df['month'] = df['date'].dt.month"

]

},

{

"cell\_type": "code",

"execution\_count": 12,

"id": "2b1a6f1f-16ff-42df-b0f1-d80a55ffc38d",

"metadata": {},

"outputs": [

{

"data": {

"text/html": [

"<div>\n",

"<style scoped>\n",

" .dataframe tbody tr th:only-of-type {\n",

" vertical-align: middle;\n",

" }\n",

"\n",

" .dataframe tbody tr th {\n",

" vertical-align: top;\n",

" }\n",

"\n",

" .dataframe thead th {\n",

" text-align: right;\n",

" }\n",

"</style>\n",

"<table border=\"1\" class=\"dataframe\">\n",

" <thead>\n",

" <tr style=\"text-align: right;\">\n",

" <th></th>\n",

" <th>id</th>\n",

" <th>date</th>\n",

" <th>NH4</th>\n",

" <th>BSK5</th>\n",

" <th>Suspended</th>\n",

" <th>O2</th>\n",

" <th>NO3</th>\n",

" <th>NO2</th>\n",

" <th>SO4</th>\n",

" <th>PO4</th>\n",

" <th>CL</th>\n",

" <th>year</th>\n",

" <th>month</th>\n",

" </tr>\n",

" </thead>\n",

" <tbody>\n",

" <tr>\n",

" <th>0</th>\n",

" <td>1</td>\n",

" <td>2000-02-17</td>\n",

" <td>0.330</td>\n",

" <td>2.77</td>\n",

" <td>12.0</td>\n",

" <td>12.30</td>\n",

" <td>9.50</td>\n",

" <td>0.057</td>\n",

" <td>154.0</td>\n",

" <td>0.454</td>\n",

" <td>289.5</td>\n",

" <td>2000</td>\n",

" <td>2</td>\n",

" </tr>\n",

" <tr>\n",

" <th>1</th>\n",

" <td>1</td>\n",

" <td>2000-05-11</td>\n",

" <td>0.044</td>\n",

" <td>3.00</td>\n",

" <td>51.6</td>\n",

" <td>14.61</td>\n",

" <td>17.75</td>\n",

" <td>0.034</td>\n",

" <td>352.0</td>\n",

" <td>0.090</td>\n",

" <td>1792.0</td>\n",

" <td>2000</td>\n",

" <td>5</td>\n",

" </tr>\n",

" <tr>\n",

" <th>2</th>\n",

" <td>1</td>\n",

" <td>2000-09-11</td>\n",

" <td>0.032</td>\n",

" <td>2.10</td>\n",

" <td>24.5</td>\n",

" <td>9.87</td>\n",

" <td>13.80</td>\n",

" <td>0.173</td>\n",

" <td>416.0</td>\n",

" <td>0.200</td>\n",

" <td>2509.0</td>\n",

" <td>2000</td>\n",

" <td>9</td>\n",

" </tr>\n",

" <tr>\n",

" <th>3</th>\n",

" <td>1</td>\n",

" <td>2000-12-13</td>\n",

" <td>0.170</td>\n",

" <td>2.23</td>\n",

" <td>35.6</td>\n",

" <td>12.40</td>\n",

" <td>17.13</td>\n",

" <td>0.099</td>\n",

" <td>275.2</td>\n",

" <td>0.377</td>\n",

" <td>1264.0</td>\n",

" <td>2000</td>\n",

" <td>12</td>\n",

" </tr>\n",

" <tr>\n",

" <th>4</th>\n",

" <td>1</td>\n",

" <td>2001-03-02</td>\n",

" <td>0.000</td>\n",

" <td>3.03</td>\n",

" <td>48.8</td>\n",

" <td>14.69</td>\n",

" <td>10.00</td>\n",

" <td>0.065</td>\n",

" <td>281.6</td>\n",

" <td>0.134</td>\n",

" <td>1462.0</td>\n",

" <td>2001</td>\n",

" <td>3</td>\n",

" </tr>\n",

" </tbody>\n",

"</table>\n",

"</div>"

],

"text/plain": [

" id date NH4 BSK5 Suspended ... SO4 PO4 CL year month\n",

"0 1 2000-02-17 0.330 2.77 12.0 ... 154.0 0.454 289.5 2000 2\n",

"1 1 2000-05-11 0.044 3.00 51.6 ... 352.0 0.090 1792.0 2000 5\n",

"2 1 2000-09-11 0.032 2.10 24.5 ... 416.0 0.200 2509.0 2000 9\n",

"3 1 2000-12-13 0.170 2.23 35.6 ... 275.2 0.377 1264.0 2000 12\n",

"4 1 2001-03-02 0.000 3.03 48.8 ... 281.6 0.134 1462.0 2001 3\n",

"\n",

"[5 rows x 13 columns]"

]

},

"execution\_count": 12,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"df.head()"

]

},

{

"cell\_type": "code",

"execution\_count": 13,

"id": "26bbbce1-8326-4166-be6d-2ce00e616630",

"metadata": {},

"outputs": [

{

"data": {

"text/plain": [

"Index(['id', 'date', 'NH4', 'BSK5', 'Suspended', 'O2', 'NO3', 'NO2', 'SO4',\n",

" 'PO4', 'CL', 'year', 'month'],\n",

" dtype='object')"

]

},

"execution\_count": 13,

"metadata": {},

"output\_type": "execute\_result"

}

],

"source": [

"df.columns"

]

},

{

"cell\_type": "code",

"id": "297f6475-d446-4147-8d7d-63652cf84205",

"metadata": {},

"outputs": [],

"source": [

"pollutants = ['O2', 'NO3', 'NO2', 'SO4',\n",

" 'PO4', 'CL']"

]

}

],

"metadata": {

"kernelspec": {

"display\_name": "Python 3 (ipykernel)",

"language": "python",

"name": "python3"

},

"language\_info": {

"codemirror\_mode": {

"name": "ipython",

"version": 3

},

"file\_extension": ".py",

"mimetype": "text/x-python",

"name": "python",

"nbconvert\_exporter": "python",

"pygments\_lexer": "ipython3",

"version": "3.12.8"

}

},

"nbformat": 4,

"nbformat\_minor": 5

}