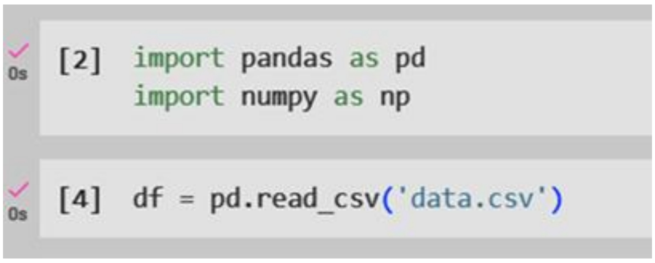
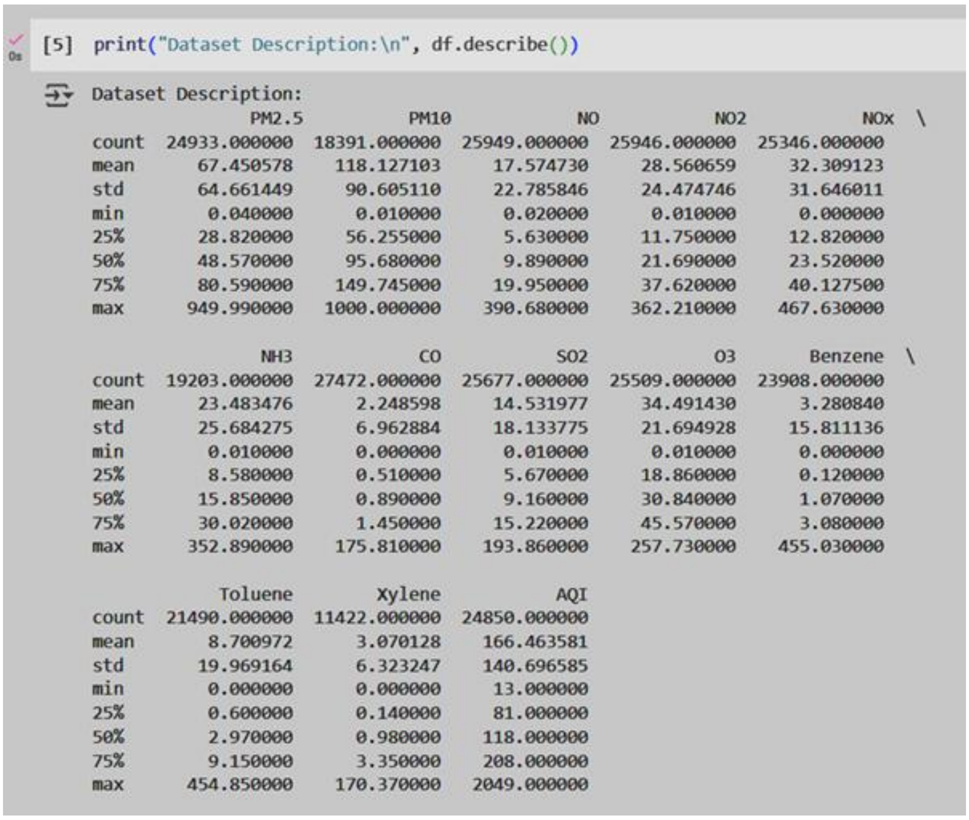
- Reading data with the help of pandas library.



```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv('data.csv')
```

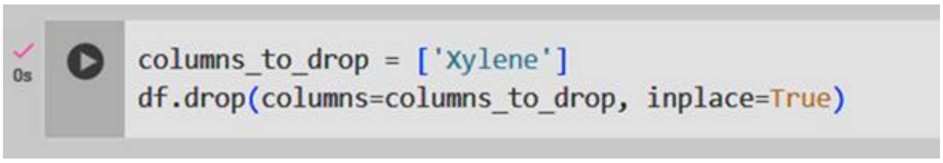
- Description of the data set using describe method.



```
print("Dataset Description:\n", df.describe())
```

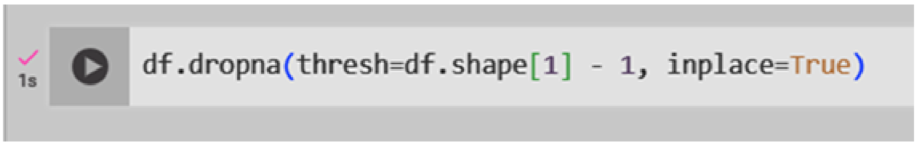
```
Dataset Description:
               PM2.5          PM10            NO           NO2           NOx  \
count  24933.000000  18391.000000  25949.000000  25946.000000  25346.000000
mean      67.450578    118.127103     17.574730     28.560659     32.309123
std       64.661449     90.605110     22.785846     24.474746     31.646011
min        0.040000      0.010000      0.020000      0.010000      0.000000
25%       28.820000     56.255000      5.630000     11.750000     12.820000
50%       48.570000     95.680000      9.890000     21.690000     23.520000
75%       80.590000    149.745000     19.950000     37.620000     40.127500
max      949.990000   1000.000000    390.680000    362.210000    467.630000

                NH3            CO           SO2            O3       Benzene  \
count  19203.000000  27472.000000  25677.000000  25509.000000  23908.000000
mean      23.483476      2.248598     14.531977     34.491430      3.280840
std       25.684275      6.962884     18.133775     21.694928     15.811136
min        0.010000      0.000000      0.010000      0.010000      0.000000
25%        8.580000      0.510000      5.670000     18.860000      0.120000
50%       15.850000      0.890000      9.160000     30.840000      1.070000
75%       30.020000      1.450000     15.220000     45.570000      3.080000
max      352.890000    175.810000    193.860000    257.730000    455.030000

            Toluene        Xylene           AQI
count  21490.000000  11422.000000  24850.000000
mean       8.700972      3.070128    166.463581
std       19.969164      6.323247    140.696585
min        0.000000      0.000000     13.000000
25%        0.600000      0.140000     81.000000
50%        2.970000      0.980000    118.000000
75%        9.150000      3.350000    208.000000
max      454.850000    170.370000   2049.000000
```

- Dropped unnecessary column



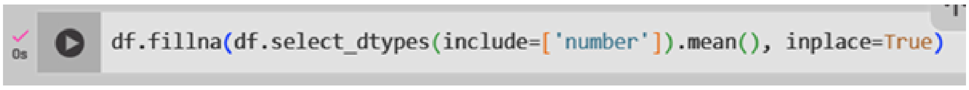
```
columns_to_drop = ['Xylene']
df.drop(columns=columns_to_drop, inplace=True)
```

- Dropped rows with maximum number of missing values.



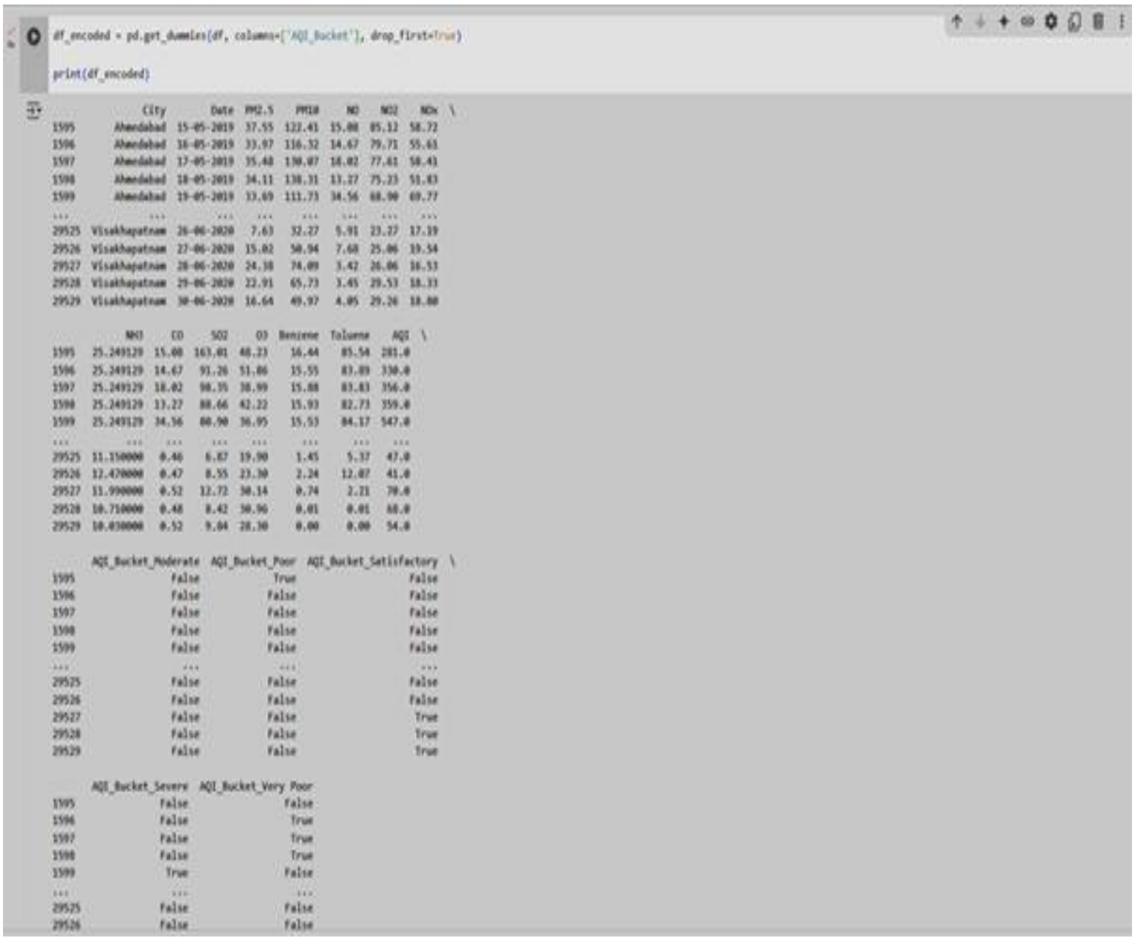
```
df.dropna(thresh=df.shape[1] - 1, inplace=True)
```

- Take care of missing data.



```
df.fillna(df.select_dtypes(include=['number']).mean(), inplace=True)
```

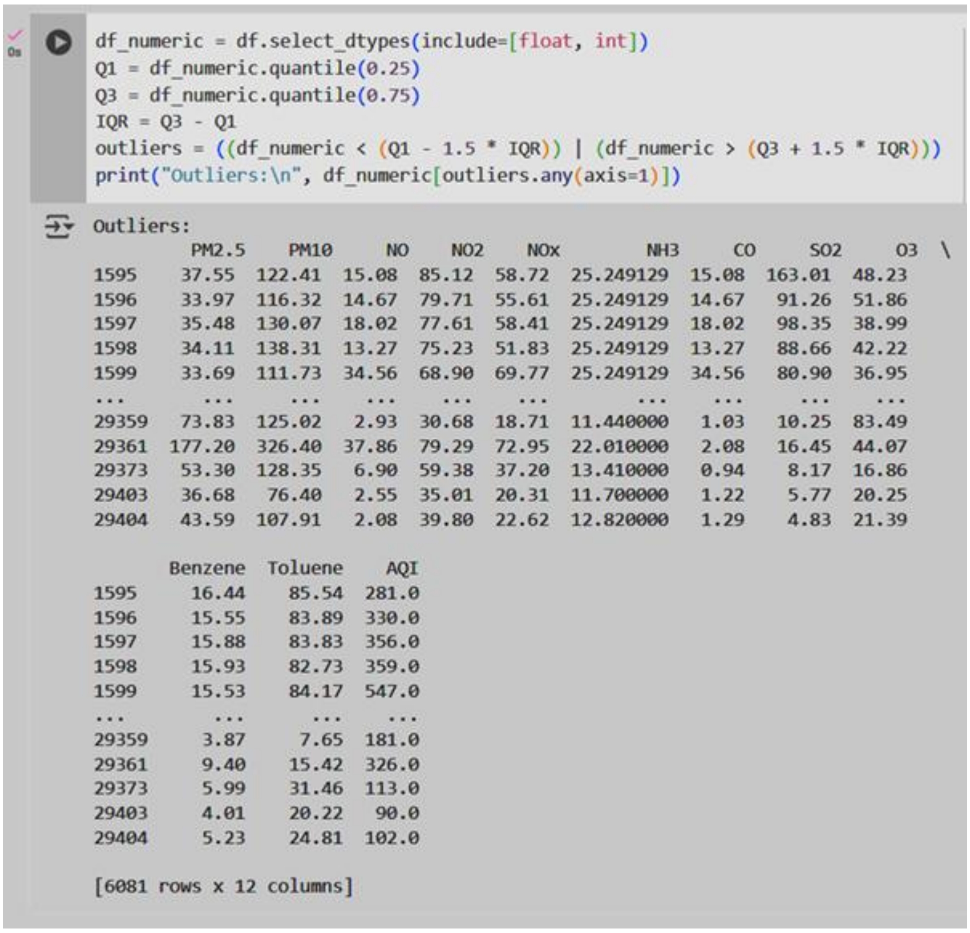
- Create dummy variables.



```
df_encoded = pd.get_dummies(df, columns=['AQI_Bucket'], drop_first=True)

print(df_encoded)
```

```
              City        Date  PM2.5    PM10     NO    NO2    NOx  \
1595     Ahmedabad  15-05-2019  37.55  122.41  15.00  85.12  58.72
1596     Ahmedabad  16-05-2019  33.97  116.32  14.67  79.71  55.61
1597     Ahmedabad  17-05-2019  35.48  130.07  18.02  77.61  58.41
1598     Ahmedabad  18-05-2019  34.11  138.31  13.27  75.23  51.83
1599     Ahmedabad  19-05-2019  33.69  111.73  34.56  68.90  69.77
...            ...         ...    ...     ...    ...    ...    ...
29525  Visakhapatnam  26-06-2020   7.63   32.27   5.91  23.27  17.19
29526  Visakhapatnam  27-06-2020  15.02   50.94   7.68  25.06  19.54
29527  Visakhapatnam  28-06-2020  24.38   74.09   3.42  26.06  16.53
29528  Visakhapatnam  29-06-2020  22.91   65.73   3.45  29.53  18.33
29529  Visakhapatnam  30-06-2020  16.64   49.97   4.05  29.26  18.80

             NH3     CO     SO2     O3  Benzene  Toluene    AQI  \
1595   25.249129  15.00  163.01  48.23    16.44    85.54  281.0
1596   25.249129  14.67   91.26  51.06    15.55    83.09  330.0
1597   25.249129  18.02   90.35  38.99    15.88    83.83  356.0
1598   25.249129  13.27   88.66  42.22    15.93    82.73  359.0
1599   25.249129  34.56   80.90  36.95    15.53    84.17  547.0
...          ...    ...     ...    ...      ...      ...    ...
29525  11.150000   0.46    6.87  19.90     1.45     5.37   47.0
29526  12.470000   0.47    8.55  23.30     2.24    12.07   41.0
29527  11.990000   0.52   12.72  30.14     0.74     2.21   70.0
29528  10.710000   0.48    8.42  30.96     0.01     0.01   68.0
29529  10.030000   0.52    9.04  28.30     0.00     0.00   54.0

       AQI_Bucket_Moderate  AQI_Bucket_Poor  AQI_Bucket_Satisfactory  \
1595                 False             True                    False
1596                 False            False                    False
1597                 False            False                    False
1598                 False            False                    False
1599                 False            False                    False
...                    ...              ...                      ...
29525                False            False                    False
29526                False            False                    False
29527                False            False                     True
29528                False            False                     True
29529                False            False                     True

       AQI_Bucket_Severe  AQI_Bucket_Very Poor
1595               False                 False
1596               False                  True
1597               False                  True
1598               False                  True
1599                True                 False
...                  ...                   ...
29525              False                 False
29526              False                 False
```

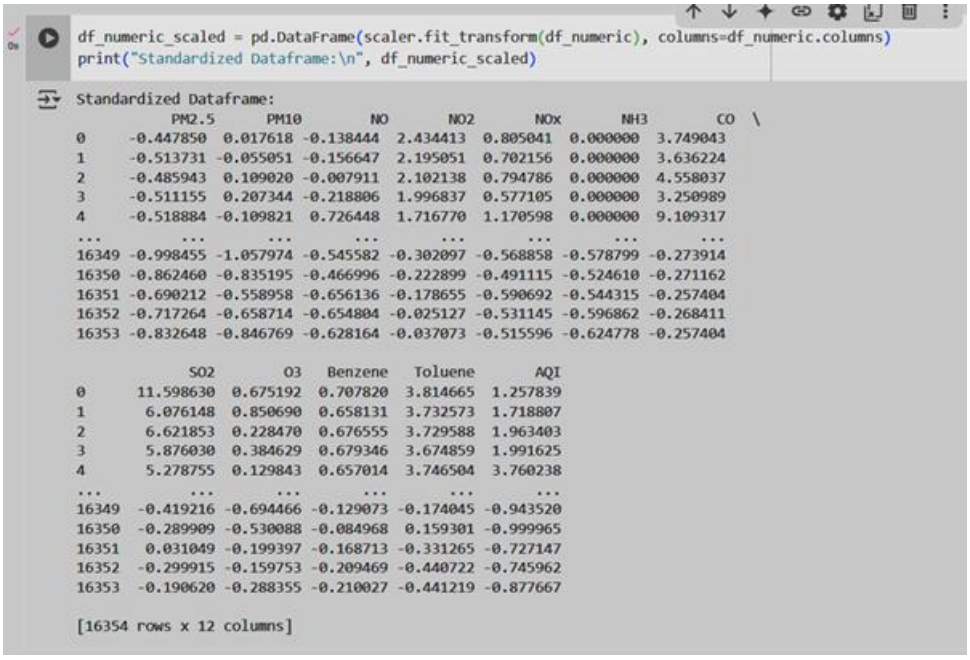
- Finding Outlier with the method of IQR :



```
df_numeric = df.select_dtypes(include=[float, int])
Q1 = df_numeric.quantile(0.25)
Q3 = df_numeric.quantile(0.75)
IQR = Q3 - Q1
outliers = ((df_numeric < (Q1 - 1.5 * IQR)) | (df_numeric > (Q3 + 1.5 * IQR)))
print("Outliers:\n", df_numeric[outliers.any(axis=1)])
```

```
Outliers:
         PM2.5    PM10     NO    NO2    NOx        NH3     CO     SO2     O3  \
1595    37.55  122.41  15.08  85.12  58.72  25.249129  15.08  163.01  48.23
1596    33.97  116.32  14.67  79.71  55.61  25.249129  14.67   91.26  51.86
1597    35.48  130.07  18.02  77.61  58.41  25.249129  18.02   98.35  38.99
1598    34.11  138.31  13.27  75.23  51.83  25.249129  13.27   88.66  42.22
1599    33.69  111.73  34.56  68.90  69.77  25.249129  34.56   80.90  36.95
...       ...     ...    ...    ...    ...        ...    ...     ...    ...
29359   73.83  125.02   2.93  30.68  18.71  11.440000   1.03   10.25  83.49
29361  177.20  326.40  37.86  79.29  72.95  22.010000   2.08   16.45  44.07
29373   53.30  128.35   6.90  59.38  37.20  13.410000   0.94    8.17  16.86
29403   36.68   76.40   2.55  35.01  20.31  11.700000   1.22    5.77  20.25
29404   43.59  107.91   2.08  39.80  22.62  12.820000   1.29    4.83  21.39

       Benzene  Toluene    AQI
1595     16.44    85.54  281.0
1596     15.55    83.89  330.0
1597     15.88    83.83  356.0
1598     15.93    82.73  359.0
1599     15.53    84.17  547.0
...        ...      ...    ...
29359     3.87     7.65  181.0
29361     9.40    15.42  326.0
29373     5.99    31.46  113.0
29403     4.01    20.22   90.0
29404     5.23    24.81  102.0

[6081 rows x 12 columns]
```

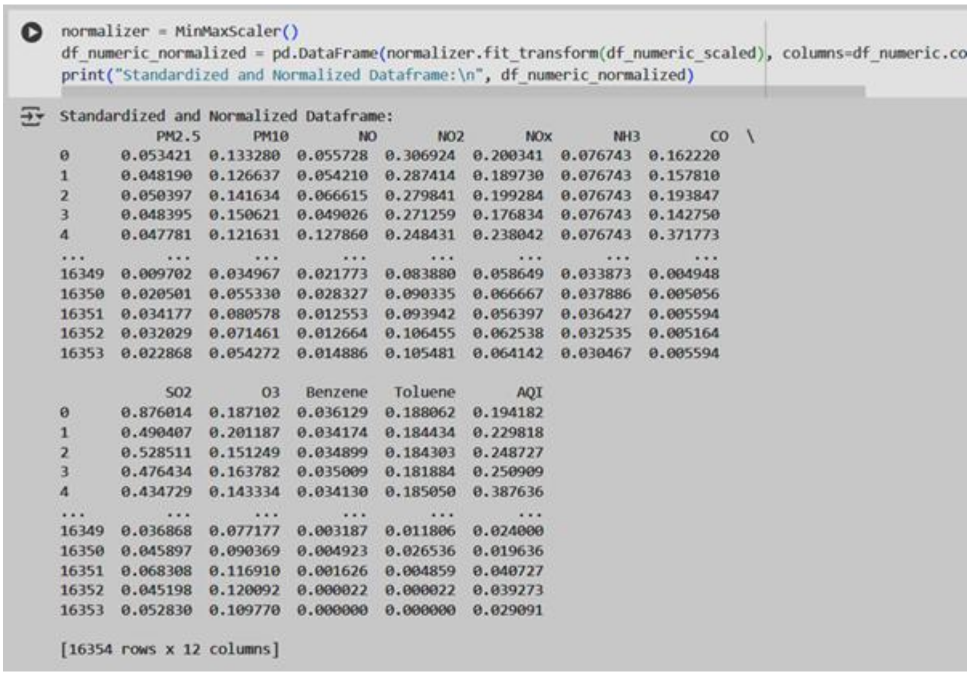
- Standardization using sk learn library



```
df_numeric_scaled = pd.DataFrame(scaler.fit_transform(df_numeric), columns=df_numeric.columns)
print("Standardized Dataframe:\n", df_numeric_scaled)
```

```
Standardized Dataframe:
           PM2.5      PM10        NO       NO2       NOx       NH3        CO  \
0     -0.447850  0.017618 -0.138444  2.434413  0.805041  0.000000  3.749043
1     -0.513731 -0.055051 -0.156647  2.195051  0.702156  0.000000  3.636224
2     -0.485943  0.109020 -0.007911  2.102138  0.794786  0.000000  4.558037
3     -0.511155  0.207344 -0.218806  1.996837  0.577105  0.000000  3.250989
4     -0.518884 -0.109821  0.726448  1.716770  1.170598  0.000000  9.109317
...         ...       ...       ...       ...       ...       ...       ...
16349 -0.998455 -1.057974 -0.545582 -0.302097 -0.568858 -0.578799 -0.273914
16350 -0.862460 -0.835195 -0.466996 -0.222899 -0.491115 -0.524610 -0.271162
16351 -0.690212 -0.558958 -0.656136 -0.178655 -0.590692 -0.544315 -0.257404
16352 -0.717264 -0.658714 -0.654804 -0.025127 -0.531145 -0.596862 -0.268411
16353 -0.832648 -0.846769 -0.628164 -0.037073 -0.515596 -0.624778 -0.257404

             SO2        O3   Benzene   Toluene       AQI
0      11.598630  0.675192  0.707820  3.814665  1.257839
1       6.076148  0.850690  0.658131  3.732573  1.718807
2       6.621853  0.228470  0.676555  3.729588  1.963403
3       5.876030  0.384629  0.679346  3.674859  1.991625
4       5.278755  0.129843  0.657014  3.746504  3.760238
...          ...       ...       ...       ...       ...
16349  -0.419216 -0.694466 -0.129073 -0.174045 -0.943520
16350  -0.289909 -0.530088 -0.084968  0.159301 -0.999965
16351   0.031049 -0.199397 -0.168713 -0.331265 -0.727147
16352  -0.299915 -0.159753 -0.209469 -0.440722 -0.745962
16353  -0.190620 -0.288355 -0.210027 -0.441219 -0.877667

[16354 rows x 12 columns]
```

- Normalization:



```
normalizer = MinMaxScaler()
df_numeric_normalized = pd.DataFrame(normalizer.fit_transform(df_numeric_scaled), columns=df_numeric.co
print("Standardized and Normalized Dataframe:\n", df_numeric_normalized)
```

```
Standardized and Normalized Dataframe:
           PM2.5      PM10        NO       NO2       NOx       NH3        CO  \
0      0.053421  0.133280  0.055728  0.306924  0.200341  0.076743  0.162220
1      0.048190  0.126637  0.054210  0.287414  0.189730  0.076743  0.157810
2      0.050397  0.141634  0.066615  0.279841  0.199284  0.076743  0.193847
3      0.048395  0.150621  0.049026  0.271259  0.176834  0.076743  0.142750
4      0.047781  0.121631  0.127860  0.248431  0.238042  0.076743  0.371773
...         ...       ...       ...       ...       ...       ...       ...
16349  0.009702  0.034967  0.021773  0.083880  0.058649  0.033873  0.004948
16350  0.020501  0.055330  0.028327  0.090335  0.066667  0.037886  0.005056
16351  0.034177  0.080578  0.012553  0.093942  0.056397  0.036427  0.005594
16352  0.032029  0.071461  0.012664  0.106455  0.062538  0.032535  0.005164
16353  0.022868  0.054272  0.014886  0.105481  0.064142  0.030467  0.005594

            SO2        O3   Benzene   Toluene       AQI
0      0.876014  0.187102  0.036129  0.188062  0.194182
1      0.490407  0.201187  0.034174  0.184434  0.229818
2      0.528511  0.151249  0.034899  0.184303  0.248727
3      0.476434  0.163782  0.035009  0.181884  0.250909
4      0.434729  0.143334  0.034130  0.185050  0.387636
...         ...       ...       ...       ...       ...
16349  0.036868  0.077177  0.003187  0.011806  0.024000
16350  0.045897  0.090369  0.004923  0.026536  0.019636
16351  0.068308  0.116910  0.001626  0.004859  0.040727
16352  0.045198  0.120092  0.000022  0.000022  0.039273
16353  0.052830  0.109770  0.000000  0.000000  0.029091

[16354 rows x 12 columns]
```