eda

July 20, 2025

1 Air Quality EDA

Exploratory Data Analysis (EDA) on the cleaned UCI Air Quality dataset.

[1]:
```
# Imports
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

# Configs
sns.set(style="whitegrid")
```

[2]:
```
# Load cleaned dataset
df = pd.read_csv('E:/air_quality_forecasting/data/clean_air_quality.csv',␣
 ↪parse_dates=['datetime'])
df.set_index('datetime', inplace=True)
df.head()
```

[2]:
```
                     CO(GT)  PT08.S1(CO)  NMHC(GT)  C6H6(GT)  PT08.S2(NMHC)  \
datetime
2004-03-10 18:00:00     2.6       1360.0     150.0      11.9         1046.0
2004-03-10 19:00:00     2.0       1292.0     112.0       9.4          955.0
2004-03-10 20:00:00     2.2       1402.0      88.0       9.0          939.0
2004-03-10 21:00:00     2.2       1376.0      80.0       9.2          948.0
2004-03-10 22:00:00     1.6       1272.0      51.0       6.5          836.0

                     NOx(GT)  PT08.S3(NOx)  NO2(GT)  PT08.S4(NO2)  \
datetime
2004-03-10 18:00:00    166.0        1056.0    113.0        1692.0
2004-03-10 19:00:00    103.0        1174.0     92.0        1559.0
2004-03-10 20:00:00    131.0        1140.0    114.0        1555.0
2004-03-10 21:00:00    172.0        1092.0    122.0        1584.0
2004-03-10 22:00:00    131.0        1205.0    116.0        1490.0

                     PT08.S5(O3)     T    RH      AH
datetime
```

```
2004-03-10 18:00:00       1268.0  13.6  48.9  0.7578
2004-03-10 19:00:00        972.0  13.3  47.7  0.7255
2004-03-10 20:00:00       1074.0  11.9  54.0  0.7502
2004-03-10 21:00:00       1203.0  11.0  60.0  0.7867
2004-03-10 22:00:00       1110.0  11.2  59.6  0.7888
```

1.1 Missing Values Analysis

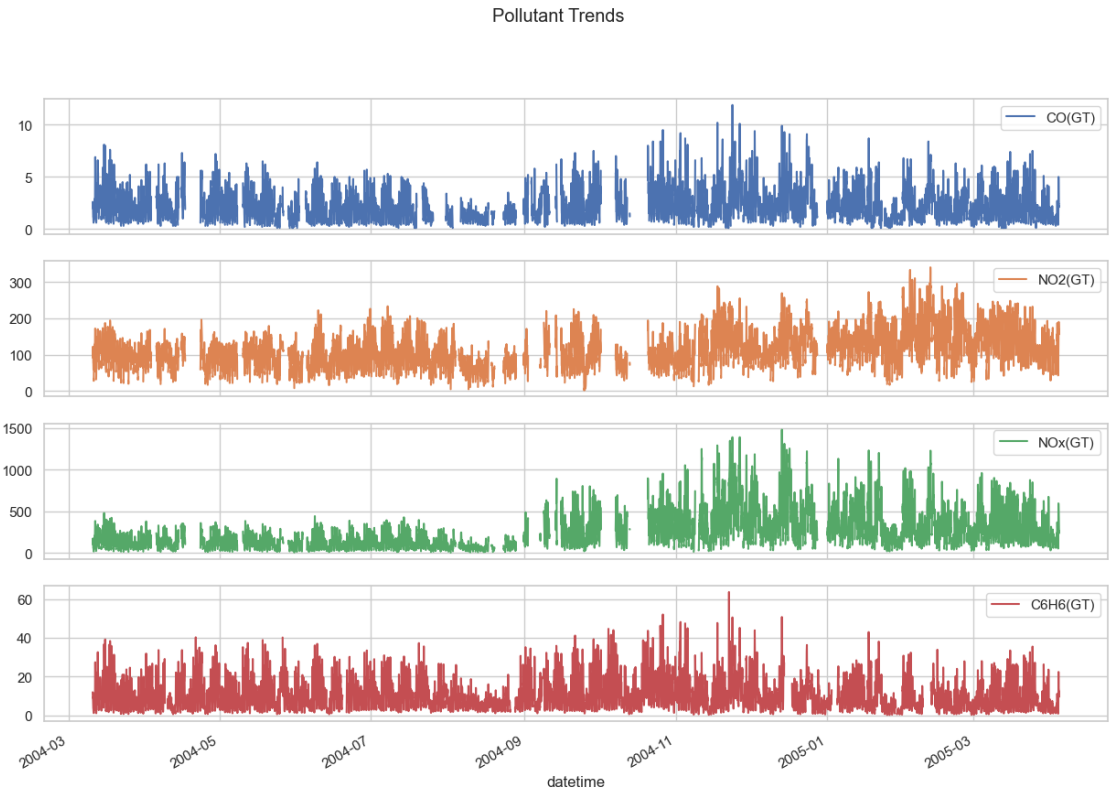
[3]:
```
missing = df.isna().sum()
missing[missing > 0].sort_values(ascending=False)
```

[3]:
```
NMHC(GT)         8557
CO(GT)           1797
NO2(GT)          1756
NOx(GT)          1753
PT08.S1(CO)       480
PT08.S2(NMHC)     480
C6H6(GT)          480
PT08.S3(NOx)      480
PT08.S4(NO2)      480
PT08.S5(O3)       480
T                 480
RH                480
AH                480
dtype: int64
```

1.2 Time Series Plot of Pollutants

[4]:
```
df[['CO(GT)', 'NO2(GT)', 'NOx(GT)', 'C6H6(GT)']].plot(subplots=True,␣
↪figsize=(15, 10), title='Pollutant Trends')
```

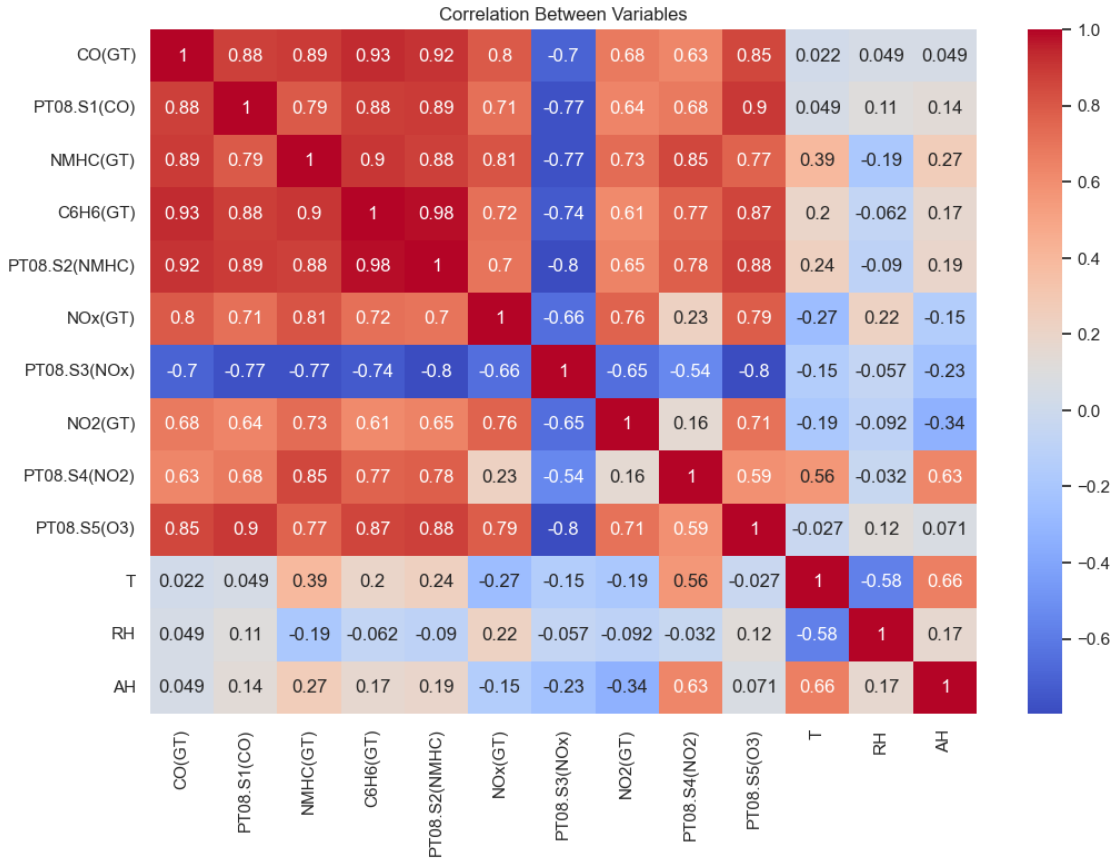
[4]:
```
array([<Axes: xlabel='datetime'>, <Axes: xlabel='datetime'>,
       <Axes: xlabel='datetime'>, <Axes: xlabel='datetime'>], dtype=object)
```



1.3 Correlation Heatmap

[5]:
```
plt.figure(figsize=(12, 8))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title('Correlation Between Variables')
```

[5]:
```
Text(0.5, 1.0, 'Correlation Between Variables')
```


Correlation Between Variables
CO(GT) 1 0.88 0.89 0.93 0.92 0.8 -0.7 0.68 0.63 0.85 0.022 0.049 0.049
PT08.S1(CO) 0.88 1 0.79 0.88 0.89 0.71 -0.77 0.64 0.68 0.9 0.049 0.11 0.14
NMHC(GT) 0.89 0.79 1 0.9 0.88 0.81 -0.77 0.73 0.85 0.77 0.39 -0.19 0.27
C6H6(GT) 0.93 0.88 0.9 1 0.98 0.72 -0.74 0.61 0.77 0.87 0.2 -0.062 0.17
PT08.S2(NMHC) 0.92 0.89 0.88 0.98 1 0.7 -0.8 0.65 0.78 0.88 0.24 -0.09 0.19
NOx(GT) 0.8 0.71 0.81 0.72 0.7 1 -0.66 0.76 0.23 0.79 -0.27 0.22 -0.15
PT08.S3(NOx) -0.7 -0.77 -0.77 -0.74 -0.8 -0.66 1 -0.65 -0.54 -0.8 -0.15 -0.057 -0.23
NO2(GT) 0.68 0.64 0.73 0.61 0.65 0.76 -0.65 1 0.16 0.71 -0.19 -0.092 -0.34
PT08.S4(NO2) 0.63 0.68 0.85 0.77 0.78 0.23 -0.54 0.16 1 0.59 0.56 -0.032 0.63
PT08.S5(O3) 0.85 0.9 0.77 0.87 0.88 0.79 -0.8 0.71 0.59 1 -0.027 0.12 0.071
T 0.022 0.049 0.39 0.2 0.24 -0.27 -0.15 -0.19 0.56 -0.027 1 -0.58 0.66
RH 0.049 0.11 -0.19 -0.062 -0.09 0.22 -0.057 -0.092 -0.032 0.12 -0.58 1 0.17
AH 0.049 0.14 0.27 0.17 0.19 -0.15 -0.23 -0.34 0.63 0.071 0.66 0.17 1
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6