In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#Loading csv file on a dataframe
ar = pd.read_csv('AirQuality.csv',sep=';')
ar.head()
```

Out[1]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.

In [2]:
```
ar.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9471 entries, 0 to 9470
Data columns (total 17 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Date           9357 non-null   object 
 1   Time           9357 non-null   object 
 2   CO(GT)         9357 non-null   object 
 3   PT08.S1(CO)    9357 non-null   float64
 4   NMHC(GT)       9357 non-null   float64
 5   C6H6(GT)       9357 non-null   object 
 6   PT08.S2(NMHC)  9357 non-null   float64
 7   NOx(GT)        9357 non-null   float64
 8   PT08.S3(NOx)   9357 non-null   float64
 9   NO2(GT)        9357 non-null   float64
 10  PT08.S4(NO2)   9357 non-null   float64
 11  PT08.S5(O3)    9357 non-null   float64
 12  T              9357 non-null   object 
 13  RH             9357 non-null   object 
 14  AH             9357 non-null   object 
 15  Unnamed: 15    0 non-null      float64
 16  ,,,,,          2556 non-null   object 
dtypes: float64(9), object(8)
memory usage: 1.2+ MB
```

In [6]:
```
#Formatting some object columns from strings to floats

ar.replace(to_replace=',',value='.',regex=True,inplace=True)

for i in'C6H6(GT) T RH AH'.split():
    ar[i]=pd.to_numeric(ar[i],errors='coerce')
```

In [8]:
```
#Dropping CO(GT) and Unnamed columns
ar = ar.loc[:,~ar.columns.str.contains('^Unnamed')]
```

In [9]:
```
ar.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9471 entries, 0 to 9470
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Date           9357 non-null   object 
 1   Time           9357 non-null   object 
 2   CO(GT)         9357 non-null   object 
 3   PT08.S1(CO)    9357 non-null   float64
 4   NMHC(GT)       9357 non-null   float64
 5   C6H6(GT)       9357 non-null   float64
 6   PT08.S2(NMHC)  9357 non-null   float64
 7   NOx(GT)        9357 non-null   float64
 8   PT08.S3(NOx)   9357 non-null   float64
 9   NO2(GT)        9357 non-null   float64
 10  PT08.S4(NO2)   9357 non-null   float64
 11  PT08.S5(O3)    9357 non-null   float64
 12  T              9357 non-null   float64
 13  RH             9357 non-null   float64
 14  AH             9357 non-null   float64
 15  ,,,,,          2556 non-null   object 
dtypes: float64(12), object(4)
memory usage: 1.2+ MB
```

In [10]:
```
#Replacing null data from -200 to NaN for posterior treatment
ar.replace(to_replace=-200,value=np.nan,inplace=True)
ar.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9471 entries, 0 to 9470
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Date           9357 non-null   object 
 1   Time           9357 non-null   object 
 2   CO(GT)         9357 non-null   object 
 3   PT08.S1(CO)    8991 non-null   float64
 4   NMHC(GT)       914 non-null    float64
 5   C6H6(GT)       8991 non-null   float64
 6   PT08.S2(NMHC)  8991 non-null   float64
 7   NOx(GT)        7718 non-null   float64
 8   PT08.S3(NOx)   8991 non-null   float64
 9   NO2(GT)        7715 non-null   float64
 10  PT08.S4(NO2)   8991 non-null   float64
 11  PT08.S5(O3)    8991 non-null   float64
 12  T              8991 non-null   float64
 13  RH             8991 non-null   float64
 14  AH             8991 non-null   float64
 15  ,,,,,          2556 non-null   object 
dtypes: float64(12), object(4)
memory usage: 1.2+ MB
```

```
C:\Users\admin\AppData\Local\Temp\ipykernel_7628\1792958475.py:2: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy
  ar.replace(to_replace=-200,value=np.nan,inplace=True)
```

In [11]:
```
NMHC_ratio = ar['NMHC(GT)'].isna().sum()/len(ar['NMHC(GT)'])
print('The NMHC(GT) sensor has {:.2f}% of missing data.'.format(NMHC_ratio*100))
#Removing NMHC(GT) sensor due to amount of null values

ar.drop('NMHC(GT)',axis=1,inplace=True)

ar.head()
```

```
The NMHC(GT) sensor has 90.35% of missing data.
```

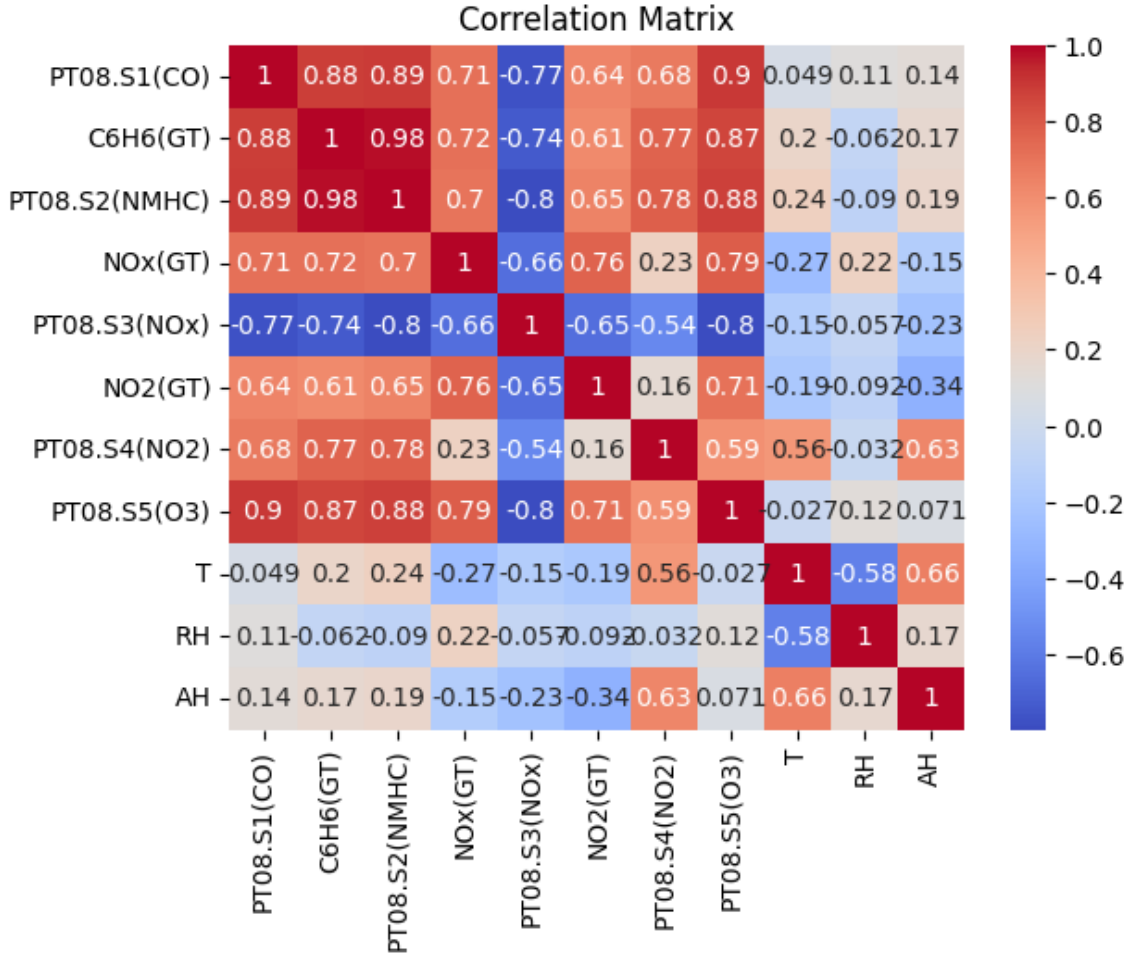
```
C:\Users\admin\AppData\Local\Temp\ipykernel_7628\1531951828.py:5: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/sta
ble/user_guide/indexing.html#returning-a-view-versus-a-copy
  ar.drop('NMHC(GT)',axis=1,inplace=True)
```

Out[11]:

	Date	Time	CO(GT)	PT08.S1(CO)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(N
0	10/03/2004	18.00.00	2.6	1360.0	11.9	1046.0	166.0	10
1	10/03/2004	19.00.00	2	1292.0	9.4	955.0	103.0	11
2	10/03/2004	20.00.00	2.2	1402.0	9.0	939.0	131.0	11
3	10/03/2004	21.00.00	2.2	1376.0	9.2	948.0	172.0	10
4	10/03/2004	22.00.00	1.6	1272.0	6.5	836.0	131.0	12

In [12]:
```
#Plotting correlation matrix
sns.heatmap(ar.corr(),annot=True,cmap = 'coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

```
C:\Users\admin\AppData\Local\Temp\ipykernel_7628\3740674684.py:2: FutureWarnin
g: The default value of numeric_only in DataFrame.corr is deprecated. In a futu
re version, it will default to False. Select only valid columns or specify the
value of numeric_only to silence this warning.
  sns.heatmap(ar.corr(),annot=True,cmap = 'coolwarm')
```


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

In [ ]: