Assignment_2_AirQuality

September 26, 2021

```
[1]: %pylab inline
     import warnings
     warnings.filterwarnings('ignore')
     import pandas as pd
     import numpy as np
     import seaborn as sns
```

```
Populating the interactive namespace from numpy and matplotlib
```

```
[2]: df = pd.read_excel("AirQualityUCI.xlsx")
     df.head()
```

```
[2]:         Date      Time  CO(GT)  PT08.S1(CO)  NMHC(GT)   C6H6(GT)  \
     0 2004-03-10  18:00:00     2.6      1360.00       150  11.881723
     1 2004-03-10  19:00:00     2.0      1292.25       112   9.397165
     2 2004-03-10  20:00:00     2.2      1402.00        88   8.997817
     3 2004-03-10  21:00:00     2.2      1375.50        80   9.228796
     4 2004-03-10  22:00:00     1.6      1272.25        51   6.518224

        PT08.S2(NMHC)  NOx(GT)  PT08.S3(NOx)  NO2(GT)  PT08.S4(NO2)  PT08.S5(O3)  \
     0        1045.50    166.0       1056.25    113.0       1692.00      1267.50
     1         954.75    103.0       1173.75     92.0       1558.75       972.25
     2         939.25    131.0       1140.00    114.0       1554.50      1074.00
     3         948.25    172.0       1092.00    122.0       1583.75      1203.25
     4         835.50    131.0       1205.00    116.0       1490.00      1110.00

            T         RH        AH  Unnamed: 15  Unnamed: 16
     0  13.60  48.875001  0.757754          NaN          NaN
     1  13.30  47.700000  0.725487          NaN          NaN
     2  11.90  53.975000  0.750239          NaN          NaN
     3  11.00  60.000000  0.786713          NaN          NaN
     4  11.15  59.575001  0.788794          NaN          NaN
```

```
[3]: x = df.drop(['Date', 'Time', 'Unnamed: 15', 'Unnamed: 16'], axis=1)
     print("RANGE for all the features:")
     print(x.max()-x.min())
     print("=====================================")
```

```python
print("VARIANCE for all the features:")
print(x.var())
print("=====================================")

x.describe()
```

```
RANGE for all the features:
CO(GT)            211.900000
PT08.S1(CO)      2239.750000
NMHC(GT)         1389.000000
C6H6(GT)          263.741476
PT08.S2(NMHC)    2414.000000
NOx(GT)          1679.000000
PT08.S3(NOx)     2882.750000
NO2(GT)           539.700000
PT08.S4(NO2)     2975.000000
PT08.S5(O3)      2722.750000
T                 244.600000
RH                288.725000
AH                202.231036
dtype: float64
=====================================
VARIANCE for all the features:
CO(GT)             6030.636106
PT08.S1(CO)      108779.263095
NMHC(GT)          19540.990493
C6H6(GT)           1712.317143
PT08.S2(NMHC)    117180.176653
NOx(GT)           66267.404793
PT08.S3(NOx)     103669.208719
NO2(GT)           16111.587462
PT08.S4(NO2)     218268.721729
PT08.S5(O3)      208778.379165
T                  1866.537024
RH                 2623.042273
AH                 1519.180817
dtype: float64
=====================================
```

```
[3]:             CO(GT)  PT08.S1(CO)     NMHC(GT)     C6H6(GT)  PT08.S2(NMHC)  \
     count  9357.000000  9357.000000  9357.000000  9357.000000    9357.000000
     mean    -34.207524  1048.869652  -159.090093     1.865576     894.475963
     std      77.657170   329.817015   139.789093    41.380154     342.315902
     min    -200.000000  -200.000000  -200.000000  -200.000000    -200.000000
     25%       0.600000   921.000000  -200.000000     4.004958     711.000000
     50%       1.500000  1052.500000  -200.000000     7.886653     894.500000
     75%       2.600000  1221.250000  -200.000000    13.636091    1104.750000
```

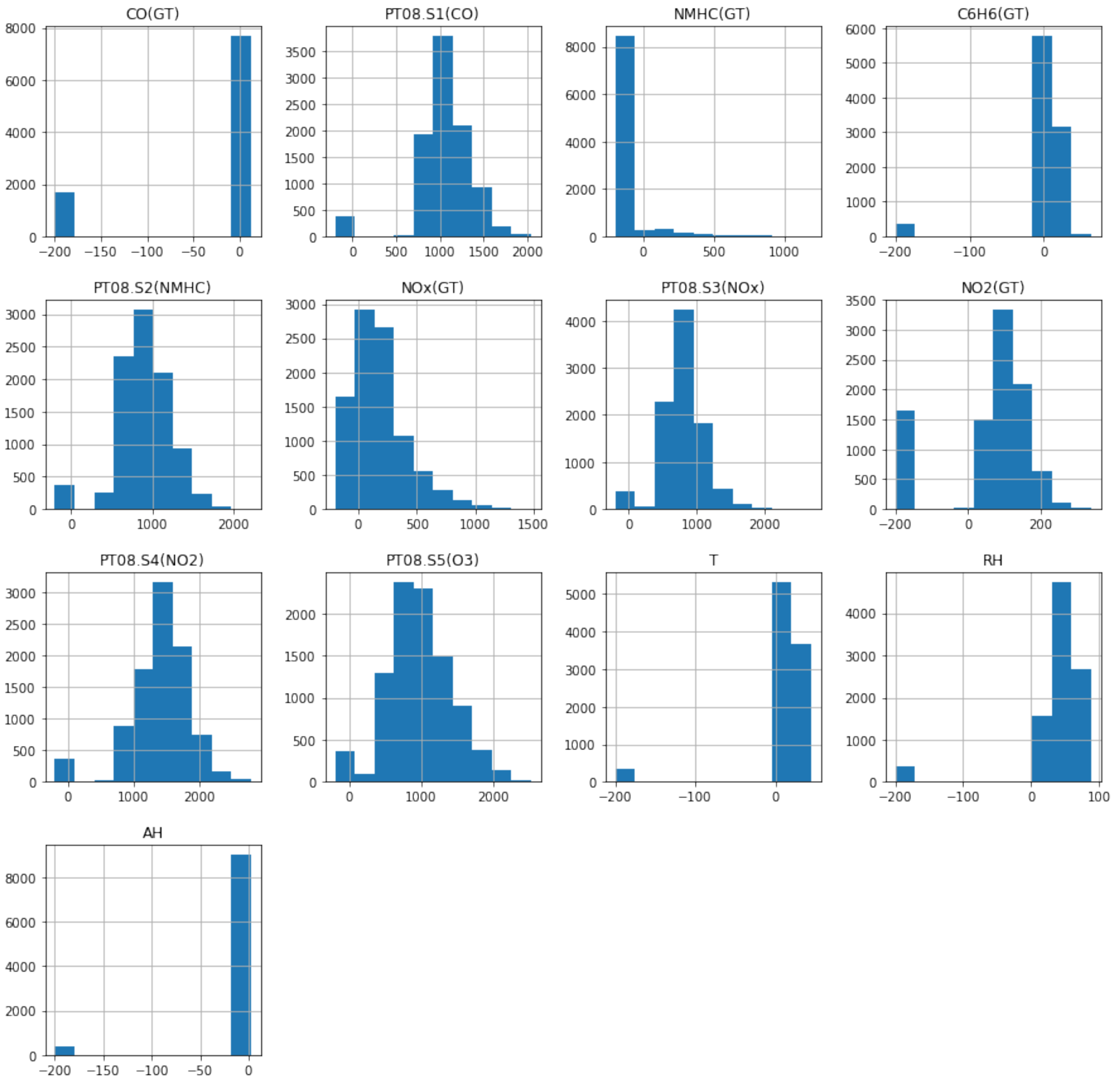
```
max      11.900000  2039.750000  1189.000000    63.741476    2214.000000

           NOx(GT)  PT08.S3(NOx)      NO2(GT)  PT08.S4(NO2)  PT08.S5(O3)  \
count  9357.000000   9357.000000  9357.000000   9357.000000  9357.000000
mean    168.604200    794.872333    58.135898   1391.363266   974.951534
std     257.424561    321.977031   126.931428    467.192382   456.922728
min    -200.000000   -200.000000  -200.000000   -200.000000  -200.000000
25%      50.000000    637.000000    53.000000   1184.750000   699.750000
50%     141.000000    794.250000    96.000000   1445.500000   942.000000
75%     284.200000    960.250000   133.000000   1662.000000  1255.250000
max    1479.000000   2682.750000   339.700000   2775.000000  2522.750000

                 T           RH           AH
count  9357.000000  9357.000000  9357.000000
mean      9.776600    39.483611    -6.837604
std      43.203438    51.215645    38.976670
min    -200.000000  -200.000000  -200.000000
25%      10.950000    34.050000     0.692275
50%      17.200000    48.550000     0.976823
75%      24.075000    61.875000     1.296223
max      44.600000    88.725000     2.231036
```

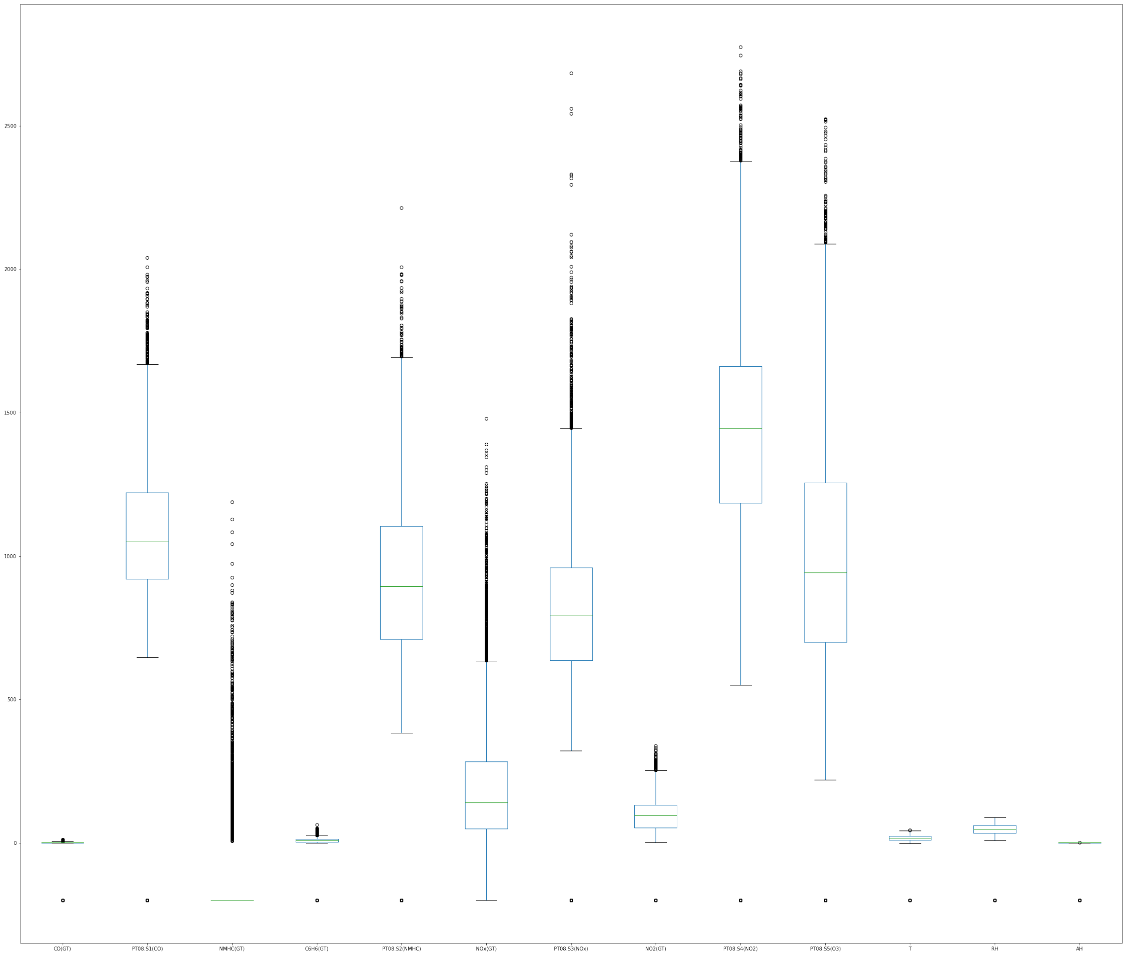
[22]:

```
# Histogram
fig = plt.figure(figsize = (15,15))
ax = fig.gca()
histogram = x.hist(ax = ax)
```



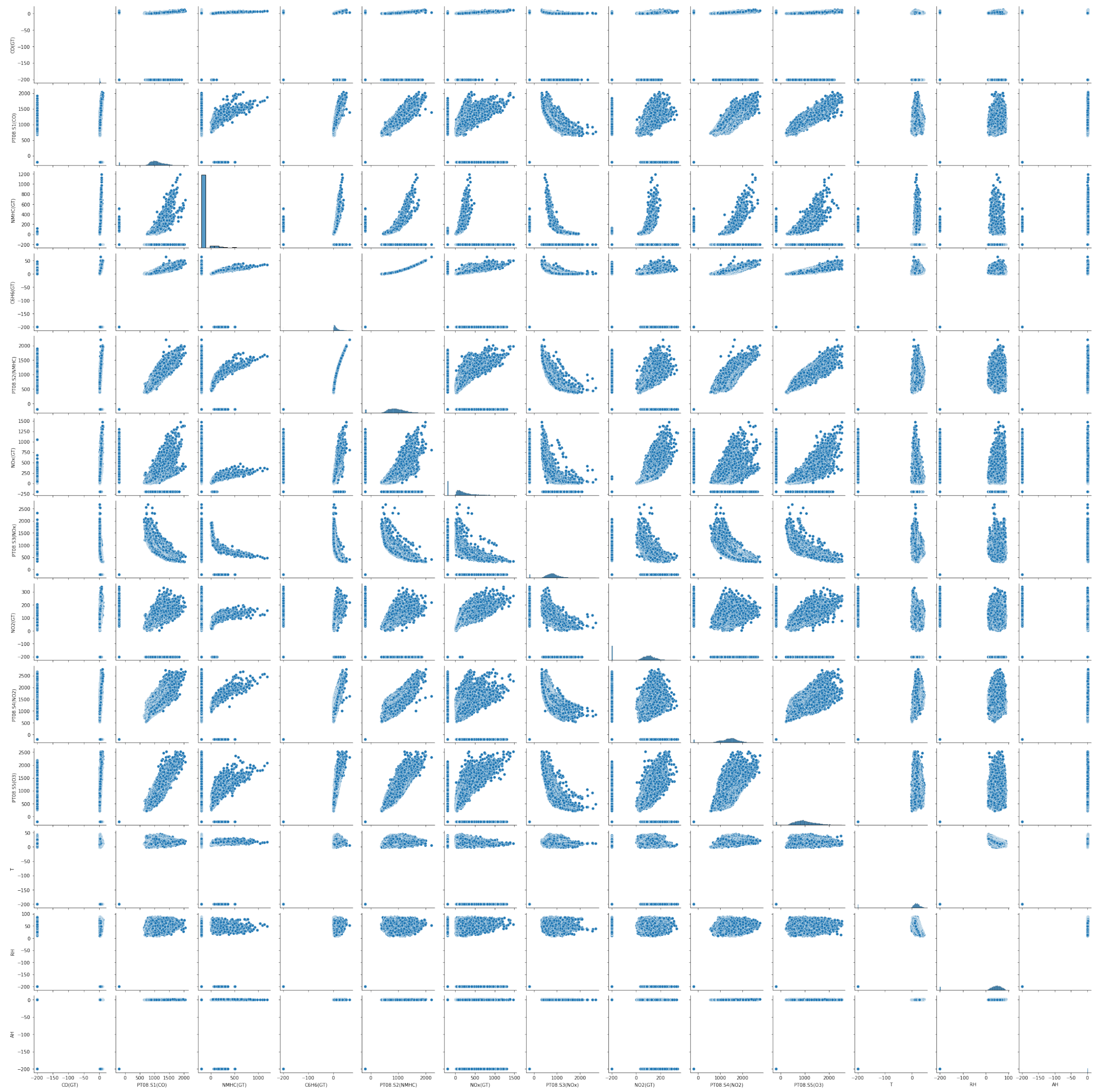
[23]:

```
# Box plots
fig = plt.figure(figsize = (40,35))
ax = fig.gca()
box_plot = x.boxplot(ax = ax, grid=False, return_type='axes')
```

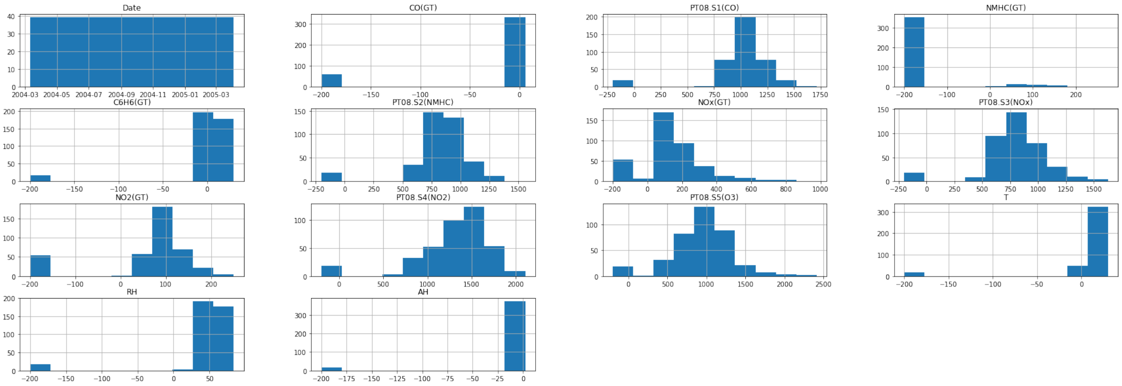


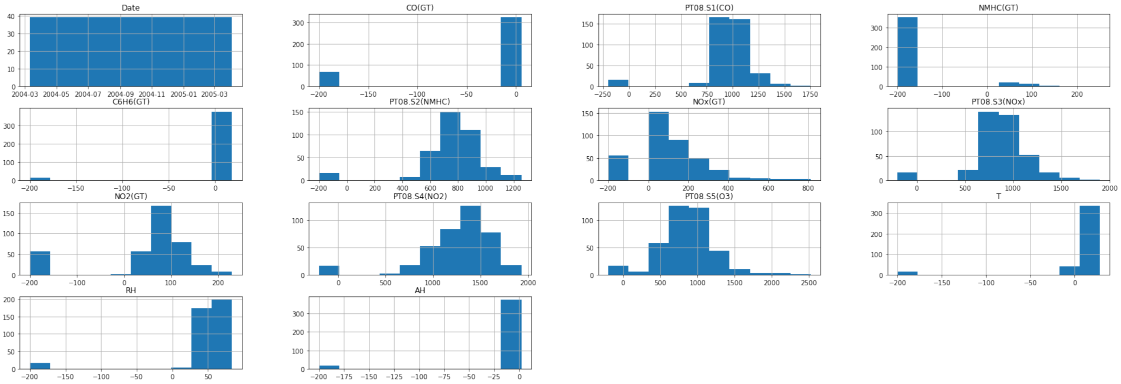
```
[24]: # Pairwise Plots
      sns.pairplot(x)
```

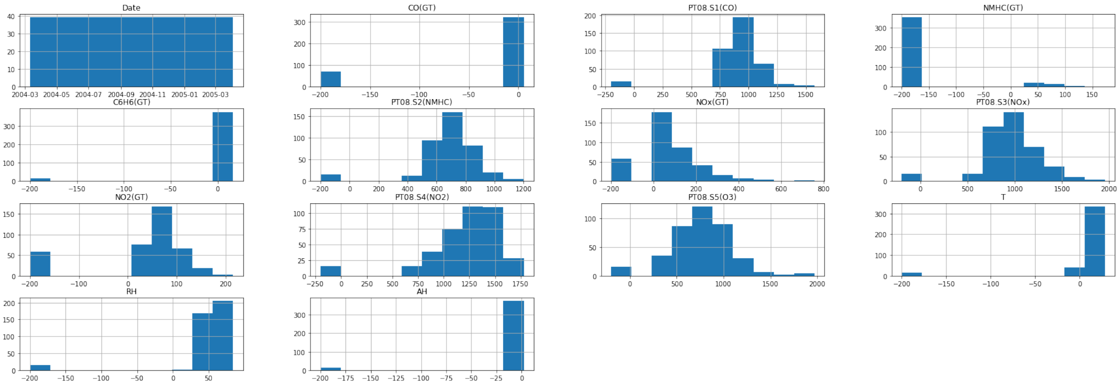
```
[24]: <seaborn.axisgrid.PairGrid at 0x2adfafa60>
```

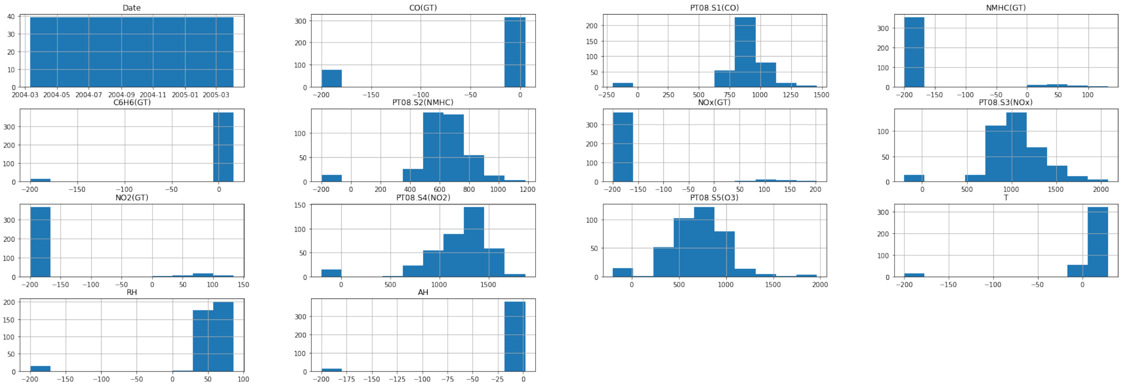


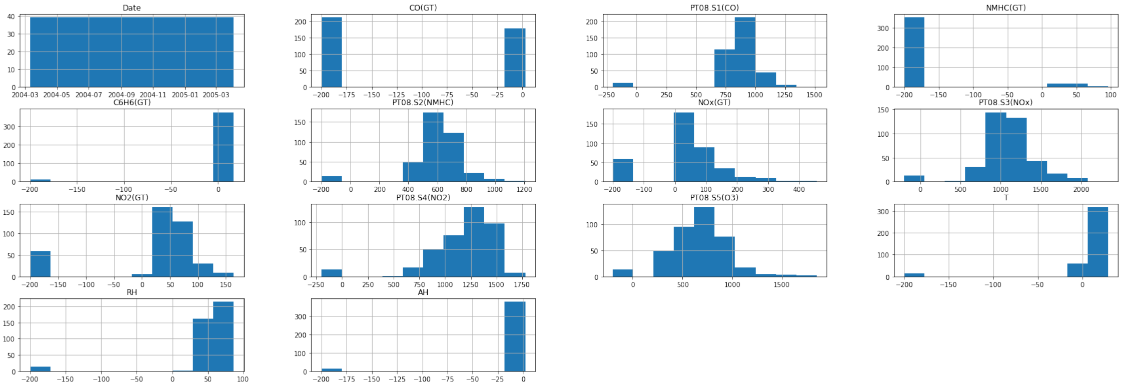
```
[26]: # Classwise Plot
      df = df.drop(['Unnamed: 15', 'Unnamed: 16'], axis=1)
      cp = df.groupby(['Time']).hist(figsize=(30,10))
```

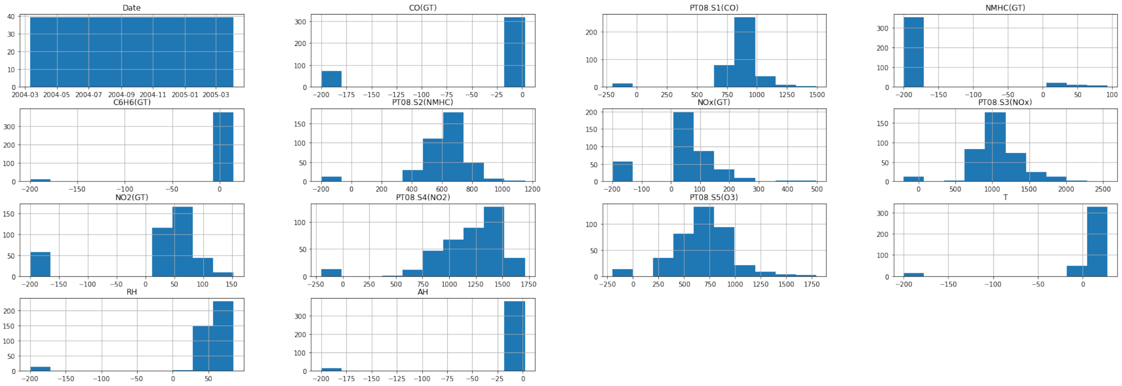

Date
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH

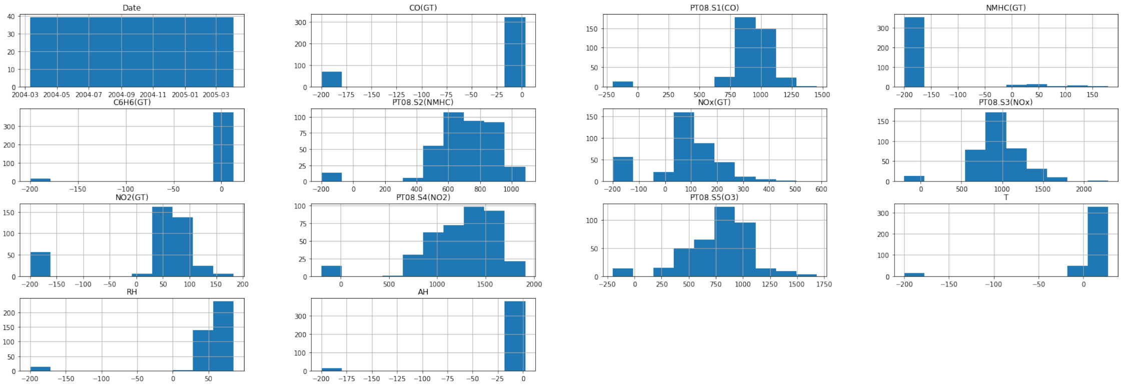

Date
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH

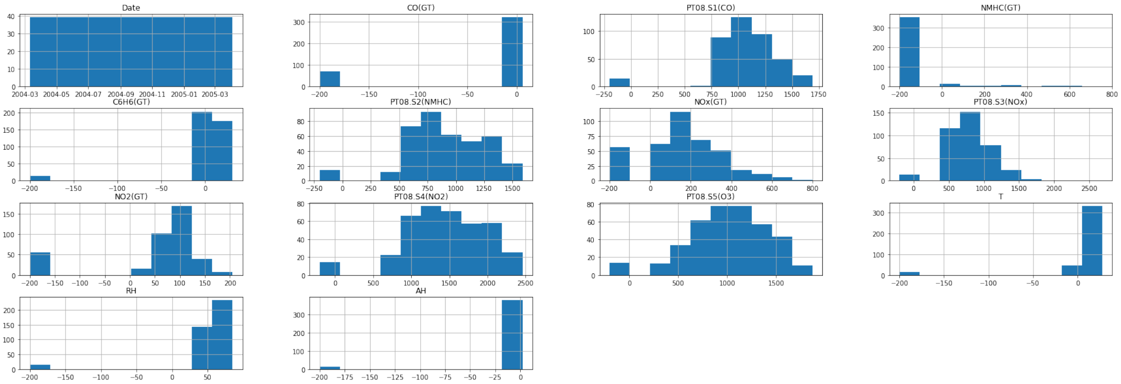

Date
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH

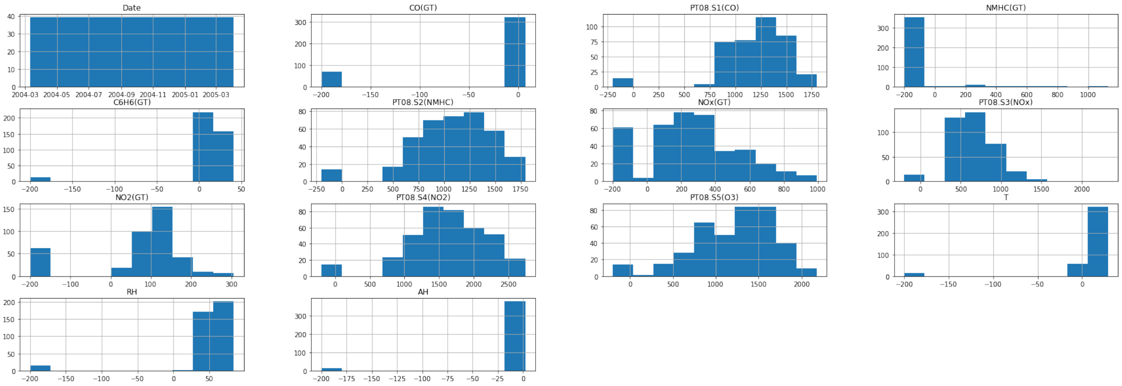

Date
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03

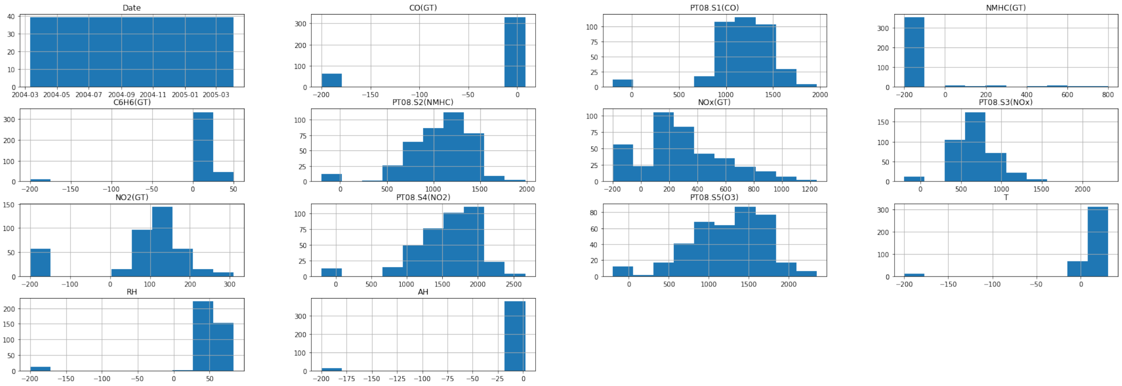

Date
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03


Date
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
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03

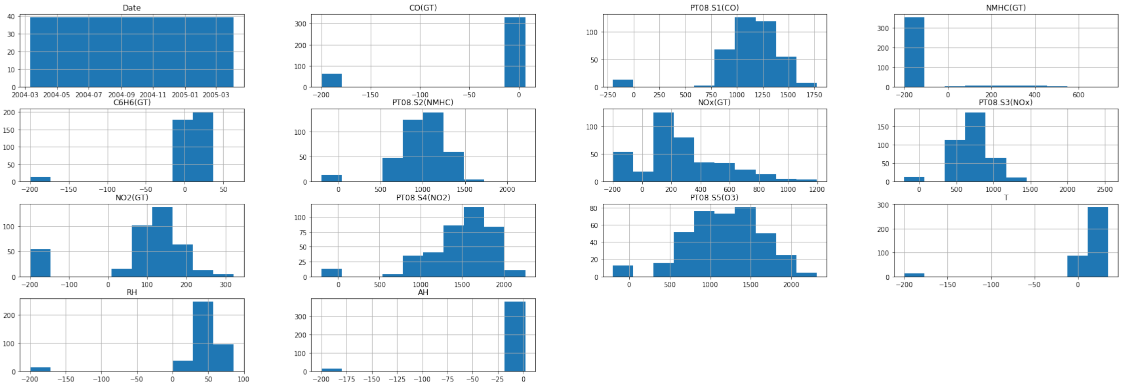

Date
0 10 20 30 40
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03
CO(GT)
0 100 200 300
−200 −175 −150 −125 −100 −75 −50 −25 0
PT08.S1(CO)
0 50 100 150
−250 0 250 500 750 1000 1250 1500
NMHC(GT)
0 100 200 300
−200 −150 −100 −50 0 50 100 150
C6H6(GT)
0 100 200 300
−200 −150 −100 −50 0
PT08.S2(NMHC)
0 25 50 75 100
−200 0 200 400 600 800 1000
NOx(GT)
0 50 100 150
−200 −100 0 100 200 300 400 500 600
PT08.S3(NOx)
0 50 100 150
0 500 1000 1500 2000
NO2(GT)
0 50 100 150
−200 −150 −100 −50 0 50 100 150 200
PT08.S4(NO2)
0 25 50 75 100
0 500 1000 1500 2000
PT08.S5(O3)
0 50 100
−250 0 250 500 750 1000 1250 1500 1750
T
0 100 200 300
−200 −150 −100 −50 0
RH
0 50 100 150 200
−200 −150 −100 −50 0 50
AH
0 100 200 300
−200 −175 −150 −125 −100 −75 −50 −25 0


Date
0 10 20 30 40
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03
CO(GT)
0 100 200 300
−200 −150 −100 −50 0
PT08.S1(CO)
0 50 100
−250 0 250 500 750 1000 1250 1500 1750
NMHC(GT)
0 100 200 300
−200 0 200 400 600 800
C6H6(GT)
0 50 100 150 200
−200 −150 −100 −50 0
PT08.S2(NMHC)
0 20 40 60 80
−250 0 250 500 750 1000 1250 1500
NOx(GT)
0 25 50 75 100
−200 0 200 400 600 800
PT08.S3(NOx)
0 50 100 150
0 500 1000 1500 2000 2500
NO2(GT)
0 50 100 150
−200 −150 −100 −50 0 50 100 150 200
PT08.S4(NO2)
0 20 40 60 80
0 500 1000 1500 2000 2500
PT08.S5(O3)
0 20 40 60 80
0 500 1000 1500
T
0 100 200 300
−200 −150 −100 −50 0
RH
0 50 100 150 200
−200 −150 −100 −50 0 50
AH
0 100 200 300
−200 −175 −150 −125 −100 −75 −50 −25 0

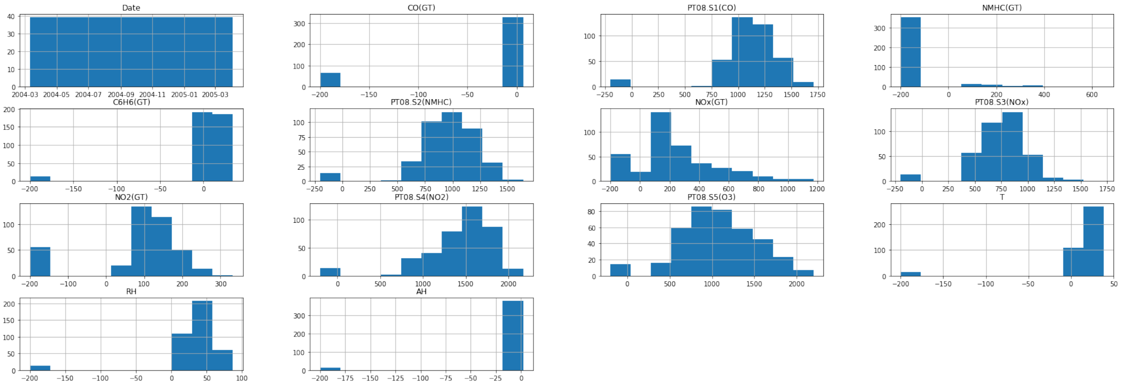

Date
0 10 20 30 40
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03
CO(GT)
0 100 200 300
−200 −150 −100 −50 0
PT08.S1(CO)
0 25 50 75 100
−250 0 250 500 750 1000 1250 1500 1750
NMHC(GT)
0 100 200 300
−200 0 200 400 600 800 1000
C6H6(GT)
0 50 100 150 200
−200 −150 −100 −50 0 50
PT08.S2(NMHC)
0 20 40 60 80
−250 0 250 500 750 1000 1250 1500 1750
NOx(GT)
0 20 40 60 80
−200 0 200 400 600 800 1000
PT08.S3(NOx)
0 50 100
0 500 1000 1500 2000
NO2(GT)
0 50 100 150
−200 −100 0 100 200 300
PT08.S4(NO2)
0 20 40 60 80
0 500 1000 1500 2000 2500
PT08.S5(O3)
0 20 40 60 80
0 500 1000 1500 2000
T
0 100 200 300
−200 −150 −100 −50 0
RH
0 50 100 150 200
−200 −150 −100 −50 0 50
AH
0 100 200 300
−200 −175 −150 −125 −100 −75 −50 −25 0



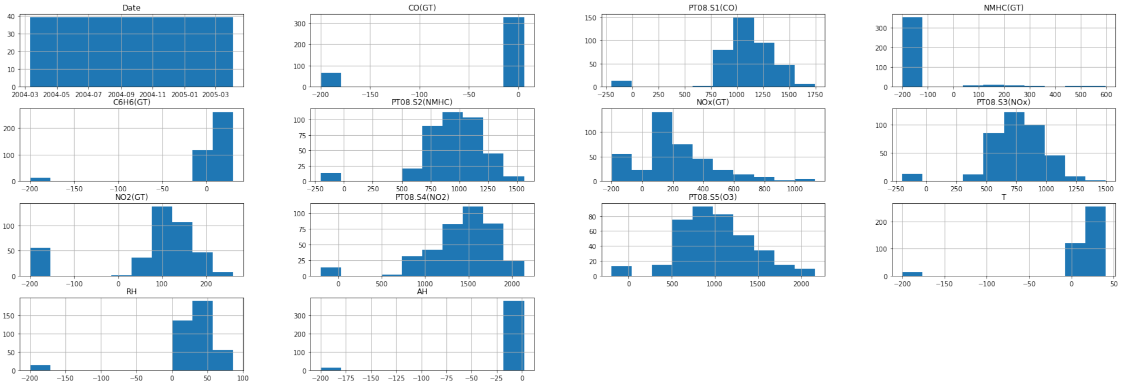
Date
0
10
20
30
40
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
0
100
200
300
−200
−150
−100
−50
0
PT08.S1(CO)
0
25
50
75
100
0
500
1000
1500
2000
NMHC(GT)
0
100
200
300
−200
0
200
400
600
800
C6H6(GT)
0
100
200
300
−200
−150
−100
−50
0
50
PT08.S2(NMHC)
0
25
50
75
100
0
500
1000
1500
2000
NOx(GT)
0
25
50
75
100
−200
0
200
400
600
800
1000
1200
PT08.S3(NOx)
0
50
100
150
0
500
1000
1500
2000
NO2(GT)
0
50
100
150
−200
−100
0
100
200
300
PT08.S4(NO2)
0
25
50
75
100
0
500
1000
1500
2000
2500
PT08.S5(O3)
0
20
40
60
80
0
500
1000
1500
2000
T
0
100
200
300
−200
−150
−100
−50
0
RH
0
50
100
150
200
−200
−150
−100
−50
0
50
AH
0
100
200
300
−200
−175
−150
−125
−100
−75
−50
−25
0

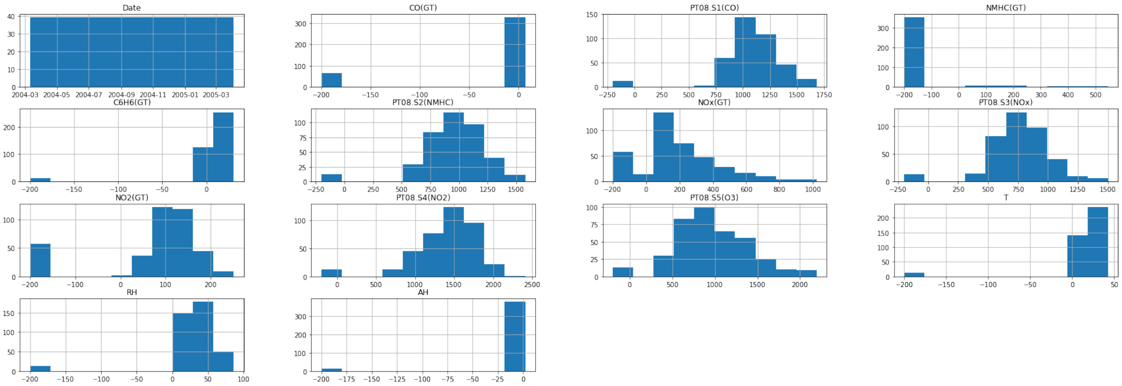


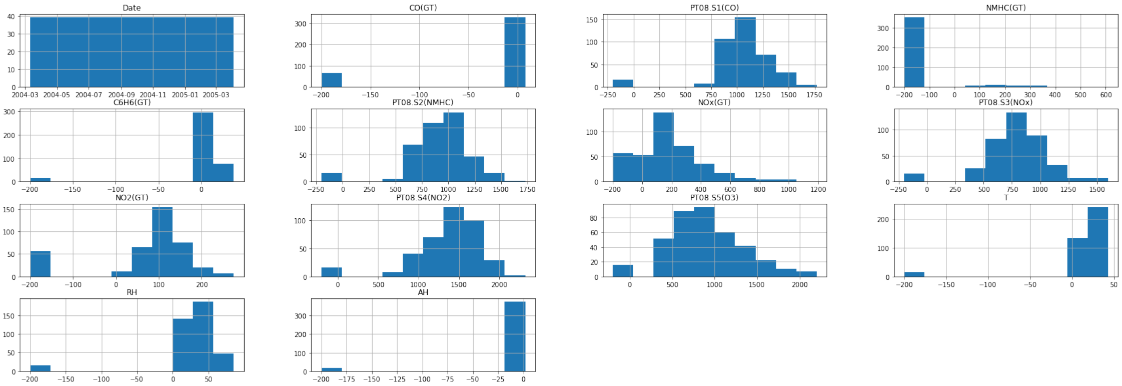
Date
0
10
20
30
40
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
0
100
200
300
−200
−150
−100
−50
0
PT08.S1(CO)
0
50
100
−250
0
250
500
750
1000
1250
1500
1750
NMHC(GT)
0
100
200
300
−200
0
200
400
600
C6H6(GT)
0
50
100
150
200
−200
−150
−100
−50
0
50
PT08.S2(NMHC)
0
50
100
0
500
1000
1500
2000
NOx(GT)
0
50
100
−200
0
200
400
600
800
1000
1200
PT08.S3(NOx)
0
50
100
150
0
500
1000
1500
2000
2500
NO2(GT)
0
50
100
−200
−100
0
100
200
300
PT08.S4(NO2)
0
25
50
75
100
0
500
1000
1500
2000
PT08.S5(O3)
0
20
40
60
80
0
500
1000
1500
2000
T
0
100
200
300
−200
−150
−100
−50
0
RH
0
100
200
−200
−150
−100
−50
0
50
100
AH
0
100
200
300
−200
−175
−150
−125
−100
−75
−50
−25
0

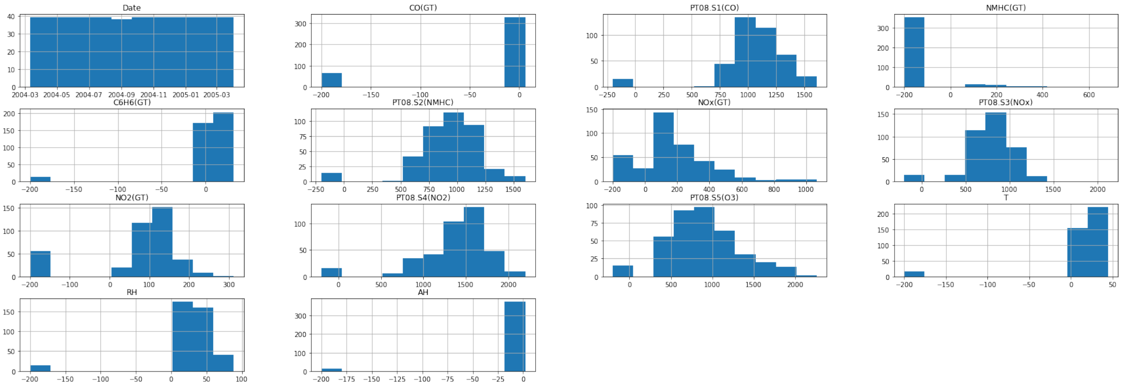


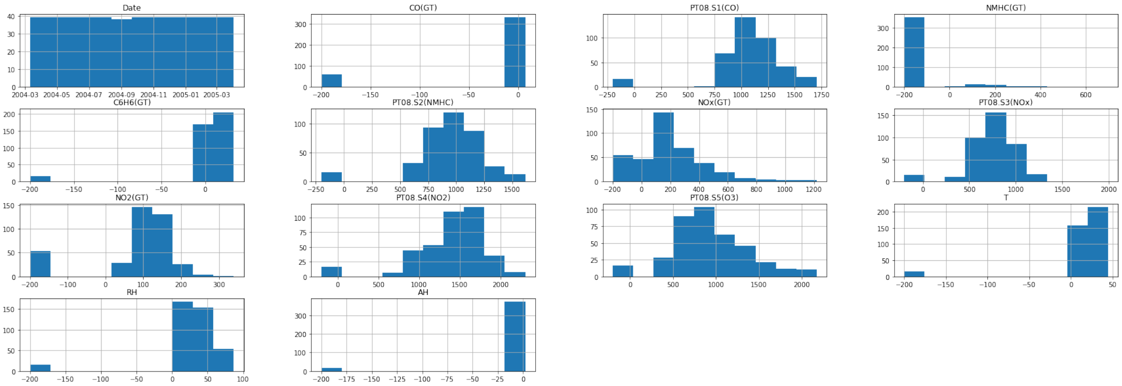
Date
0
10
20
30
40
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
0
100
200
300
−200
−150
−100
−50
0
PT08.S1(CO)
0
50
100
−250
0
250
500
750
1000
1250
1500
1750
NMHC(GT)
0
100
200
300
−200
0
200
400
600
C6H6(GT)
0
50
100
150
200
−200
−150
−100
−50
0
PT08.S2(NMHC)
0
25
50
75
100
−250
0
250
500
750
1000
1250
1500
NOx(GT)
0
50
100
−200
0
200
400
600
800
1000
1200
PT08.S3(NOx)
0
50
100
−250
0
250
500
750
1000
1250
1500
1750
NO2(GT)
0
50
100
−200
−100
0
100
200
300
PT08.S4(NO2)
0
50
100
0
500
1000
1500
2000
PT08.S5(O3)
0
20
40
60
80
0
500
1000
1500
2000
T
0
100
200
−200
−150
−100
−50
0
50
RH
0
50
100
150
200
−200
−150
−100
−50
0
50
100
AH
0
100
200
300
−200
−175
−150
−125
−100
−75
−50
−25
0

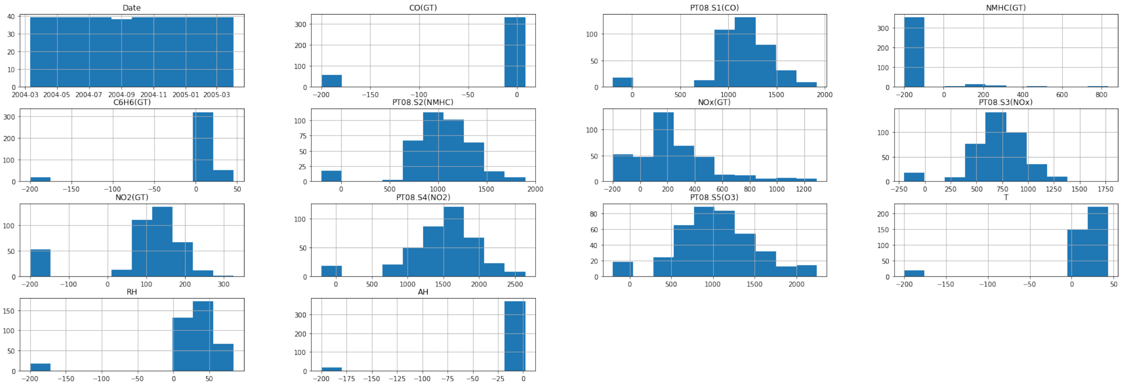

Date
0 10 20 30 40
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03
CO(GT)
0 100 200 300
−200 −150 −100 −50 0
PT08.S1(CO)
0 50 100 150
−250 0 250 500 750 1000 1250 1500 1750
NMHC(GT)
0 100 200 300
−200 −100 0 100 200 300 400 500 600
C6H6(GT)
0 100 200
−200 −150 −100 −50 0
PT08.S2(NMHC)
0 25 50 75 100
−250 0 250 500 750 1000 1250 1500
NOx(GT)
0 50 100
−200 0 200 400 600 800 1000
PT08.S3(NOx)
0 50 100
−250 0 250 500 750 1000 1250 1500
NO2(GT)
0 50 100
−200 −100 0 100 200
PT08.S4(NO2)
0 25 50 75 100
0 500 1000 1500 2000
PT08.S5(O3)
0 20 40 60 80
0 500 1000 1500 2000
T
0 100 200
−200 −150 −100 −50 0 50
RH
0 50 100 150
−200 −150 −100 −50 0 50 100
AH
0 100 200 300
−200 −175 −150 −125 −100 −75 −50 −25 0

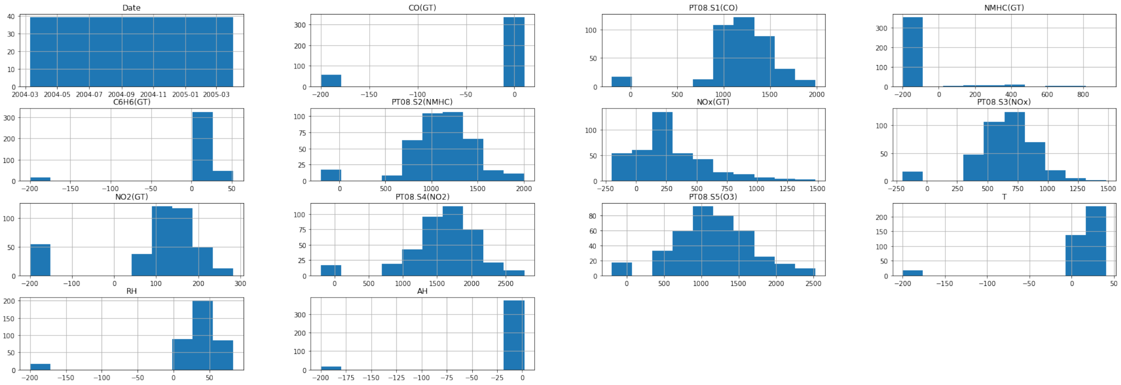

Date
0 10 20 30 40
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03
CO(GT)
0 100 200 300
−200 −150 −100 −50 0
PT08.S1(CO)
0 50 100 150
−250 0 250 500 750 1000 1250 1500 1750
NMHC(GT)
0 100 200 300
−200 −100 0 100 200 300 400 500
C6H6(GT)
0 100 200
−200 −150 −100 −50 0
PT08.S2(NMHC)
0 25 50 75 100
−250 0 250 500 750 1000 1250 1500
NOx(GT)
0 50 100
−200 0 200 400 600 800 1000
PT08.S3(NOx)
0 50 100
−250 0 250 500 750 1000 1250 1500
NO2(GT)
0 50 100
−200 −100 0 100 200
PT08.S4(NO2)
0 50 100
0 500 1000 1500 2000 2500
PT08.S5(O3)
0 25 50 75 100
0 500 1000 1500 2000
T
0 50 100 150 200
−200 −150 −100 −50 0 50
RH
0 50 100 150
−200 −150 −100 −50 0 50 100
AH
0 100 200 300
−200 −175 −150 −125 −100 −75 −50 −25 0

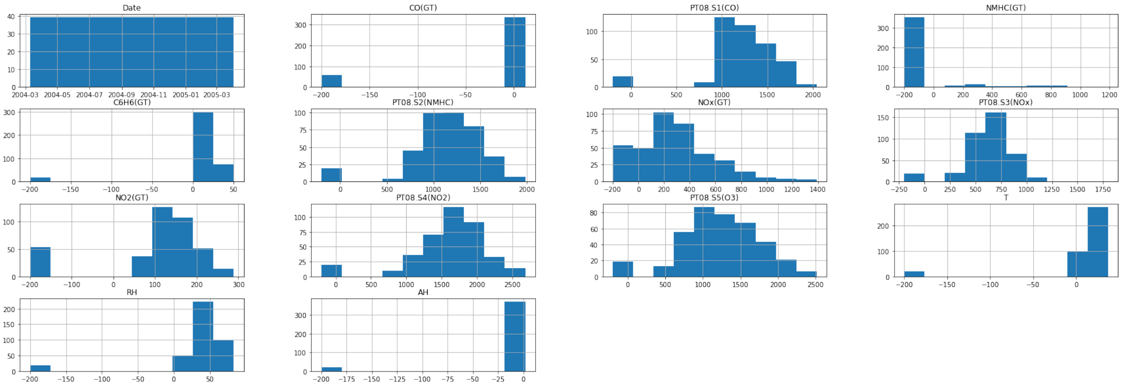

Date
0 10 20 30 40
2004-03 2004-05 2004-07 2004-09 2004-11 2005-01 2005-03
CO(GT)
0 100 200 300
−200 −150 −100 −50 0
PT08.S1(CO)
0 50 100 150
−250 0 250 500 750 1000 1250 1500 1750
NMHC(GT)
0 100 200 300
−200 −100 0 100 200 300 400 500 600
C6H6(GT)
0 100 200 300
−200 −150 −100 −50 0
PT08.S2(NMHC)
0 50 100
−250 0 250 500 750 1000 1250 1500 1750
NOx(GT)
0 50 100
−200 0 200 400 600 800 1000 1200
PT08.S3(NOx)
0 50 100
−250 0 250 500 750 1000 1250 1500
NO2(GT)
0 50 100 150
−200 −100 0 100 200
PT08.S4(NO2)
0 50 100
0 500 1000 1500 2000
PT08.S5(O3)
0 20 40 60 80
0 500 1000 1500 2000
T
0 100 200
−200 −150 −100 −50 0 50
RH
0 50 100 150
−200 −150 −100 −50 0 50
AH
0 100 200 300
−200 −175 −150 −125 −100 −75 −50 −25 0

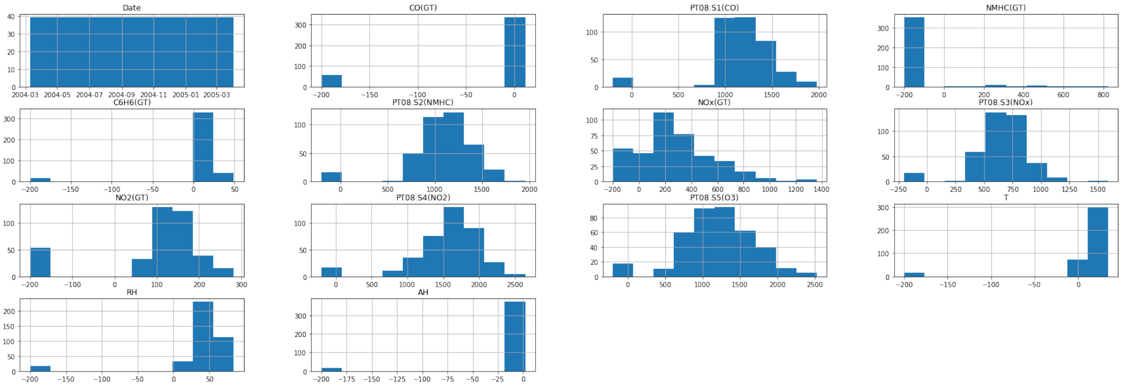

Date
0
10
20
30
40
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
−200
−150
−100
−50
0
PT08.S1(CO)
−250
0
250
500
750
1000
1250
1500
NMHC(GT)
−200
0
200
400
600
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
−175
−125
−75
−25
2000
800
1000
300
100
50

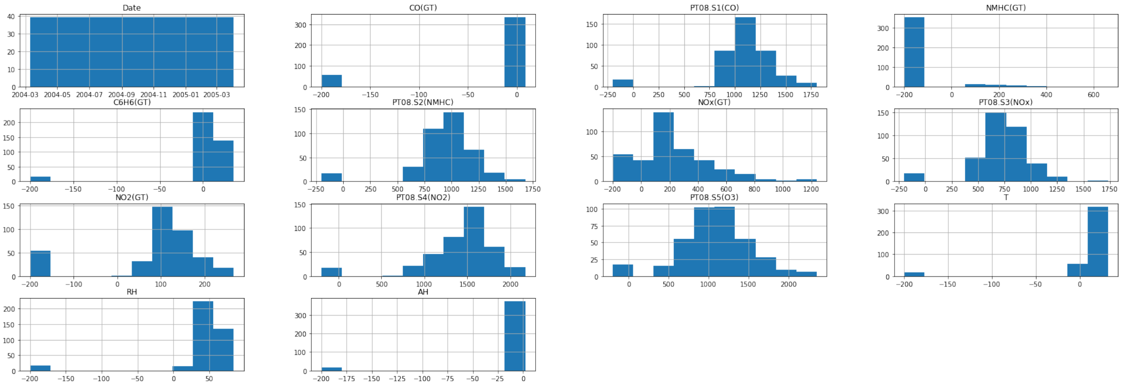

Date
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
−250
−200
−175
−150
−125
−100
−75
−50
−25
0
250
500
750
1000
1250
1500
1750
200
400
600
800
1200
2000

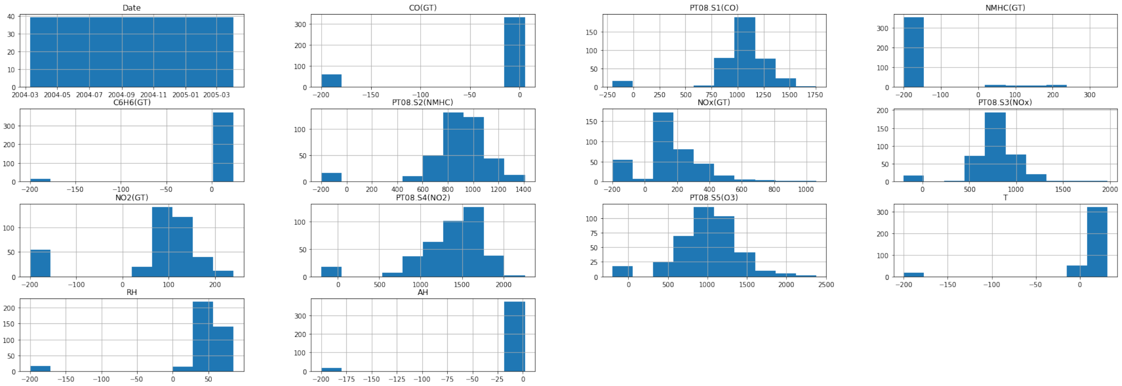

Date
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH
−250
−200
−175
−150
−125
−100
−75
−50
−25
0
500
1000
1500
2000
2500
200
400
600
800
1200
1750

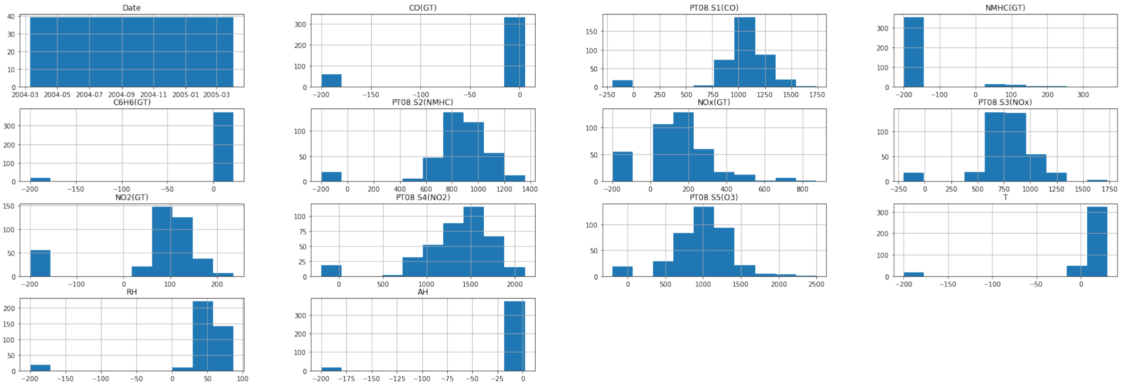

Date
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
CO(GT)
PT08.S1(CO)
NMHC(GT)
C6H6(GT)
PT08.S2(NMHC)
NOx(GT)
PT08.S3(NOx)
NO2(GT)
PT08.S4(NO2)
PT08.S5(O3)
T
RH
AH


Date
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
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


Date
2004-03
2004-05
2004-07
2004-09
2004-11
2005-01
2005-03
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 2.3 - CONCEPTUAL QUESTIONS

1.0.1 1.

The ranges for features like Relative Humidity(RH), Absolute Humidity(AH), CO(GT), and C6H6(GT) start with negative values. This is a big inconsistency in data as RH and AH should be greater than 0 and if CO or C6H6 is not there in atmosphere, they should also be 0 but not negative.

1.0.2 2.

The abnormalities and inconsistency in data can be found in the summary statistics also. They are minimum values for all the readings are -200, and mean for AH, NMHC(GT), and CO(GT) are negative values.

1.0.3 3.

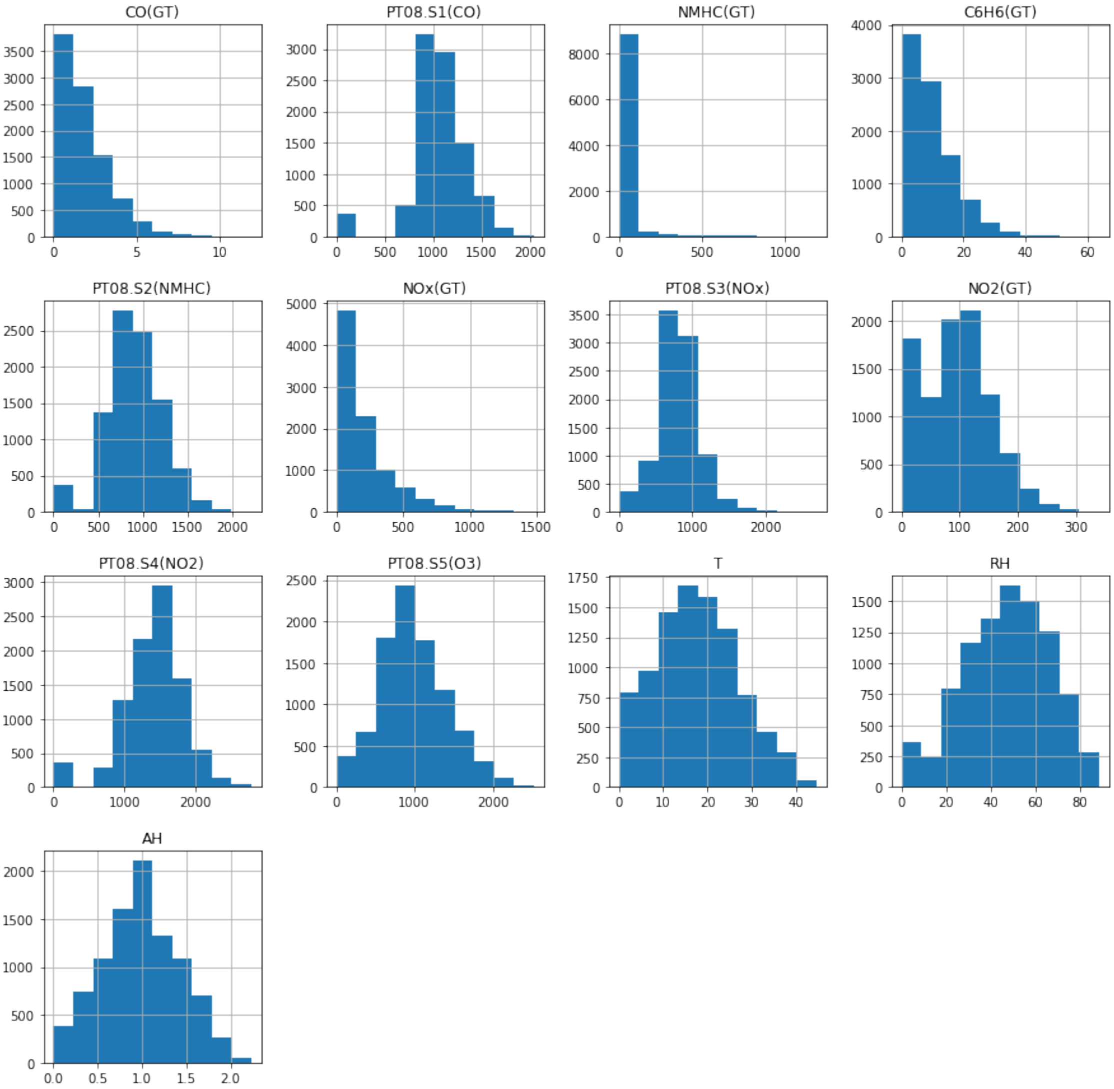
The abnormalities discussed above can be removed from the data by masking the negative values and irregular ranges.

[4]:

```
x.mask(x < 0, 0, inplace=True)
```

[5]:

```
#4. Histograms after masking negative values.
fig = plt.figure(figsize = (15,15))
ax = fig.gca()
histogram = x.hist(ax = ax)
```



1.0.4 4.

As we can see from the above plots, masking the negative values gave us the new ranges for RH, AH, CO(GT), C6H6(GT).