In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [4]:

```
df=pd.read_csv('../DSBDAFINAL/AirQuality.csv',sep=';')
```

In [5]:

```
df
```

Out[5]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.0	1056.0	113.0	1692.0	
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.0	1174.0	92.0	1559.0	
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.0	1140.0	114.0	1555.0	
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.0	1092.0	122.0	1584.0	
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.0	1205.0	116.0	1490.0	
...	...	...	...	...	...	...	...	...	...	...	...	
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

9471 rows × 17 columns

In [6]:

```
df.head()
```

Out[6]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.0	1056.0	113.0	1692.0	
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.0	1174.0	92.0	1559.0	
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.0	1140.0	114.0	1555.0	
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.0	1092.0	122.0	1584.0	
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.0	1205.0	116.0	1490.0	

In [7]:

```
df.shape
```

Out[7]:

```
(9471, 17)
```

In [8]:

```
df.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  Unnamed: 16    0 non-null      float64
dtypes: float64(10), object(7)
memory usage: 1.2+ MB
```

In [9]:

```
#------------------Data cleaning------------------------------#
```

In [15]:

```
df.dropna(inplace=True)
```

In [16]:

```
df.drop_duplicates(inplace=True)
```

Out[16]:

Date	Time	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

In [17]:
```
df
```

Out[17]:

Date	Time	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

In [18]:
```
df['NMHC(GT)']=df['NMHC(GT)'].astype(int)
df['NOx(GT)']=df['NOx(GT)'].astype(int)
```

In [19]:
```
df
```

Out[19]:

Date	Time	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

In [20]:
```
df.head()
```

Out[20]:

Date	Time	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

In [21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 0 entries
Data columns (total 17 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Date           0 non-null      object 
 1   Time           0 non-null      object 
 2   CO(GT)         0 non-null      object 
 3   PT08.S1(CO)    0 non-null      float64
 4   NMHC(GT)       0 non-null      int32  
 5   C6H6(GT)       0 non-null      object 
 6   PT08.S2(NMHC)  0 non-null      float64
 7   NOx(GT)        0 non-null      int32  
 8   PT08.S3(NOx)   0 non-null      float64
 9   NO2(GT)        0 non-null      float64
 10  PT08.S4(NO2)   0 non-null      float64
 11  PT08.S5(O3)    0 non-null      float64
 12  T              0 non-null      object 
 13  RH             0 non-null      object 
 14  AH             0 non-null      object 
 15  Unnamed: 15    0 non-null      float64
 16  Unnamed: 16    0 non-null      float64
dtypes: float64(8), int32(2), object(7)
memory usage: 0.0+ bytes
```

In [23]:
```
df.drop(['Unnamed: 15','Unnamed: 16'],axis=1,inplace=True)
```

In [25]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 0 entries
Data columns (total 15 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Date           0 non-null      object 
 1   Time           0 non-null      object 
 2   CO(GT)         0 non-null      object 
 3   PT08.S1(CO)    0 non-null      float64
 4   NMHC(GT)       0 non-null      int32  
 5   C6H6(GT)       0 non-null      object 
 6   PT08.S2(NMHC)  0 non-null      float64
 7   NOx(GT)        0 non-null      int32  
 8   PT08.S3(NOx)   0 non-null      float64
 9   NO2(GT)        0 non-null      float64
 10  PT08.S4(NO2)   0 non-null      float64
 11  PT08.S5(O3)    0 non-null      float64
 12  T              0 non-null      object 
 13  RH             0 non-null      object 
 14  AH             0 non-null      object 
dtypes: float64(6), int32(2), object(7)
memory usage: 0.0+ bytes
```

In [26]:
```
df
```

Out[26]:

Date	Time	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

In [27]:
```
#++++++++++++++++++++++++data integration++++++++++++++++++++++++++#
```

In [30]:
```
df=pd.read_csv('../DSBDAFINAL/AirQuality.csv',sep=';')
```

In [31]:
```
df
```

Out[31]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.0	1056.0	113.0	1692.0	
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.0	1174.0	92.0	1559.0	
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.0	1140.0	114.0	1555.0	
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.0	1092.0	122.0	1584.0	
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.0	1205.0	116.0	1490.0	
...	...	...	...	...	...	...	...	...	...	...	...	
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

9471 rows × 17 columns

In [40]:
```
subset1=df[df['NMHC(GT)'] <15]
subset2=df[df['NMHC(GT)'] >=22]
```

In [41]:
```
subset1
```

Out[41]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT
10	11/03/2004	04.00.00	-200	1011.0	14.0	1,3	527.0	21.0	1818.0	34.0	1197.0	
11	11/03/2004	05.00.00	0,7	1066.0	8.0	1,1	512.0	16.0	1918.0	28.0	1182.0	
34	12/03/2004	04.00.00	-200	831.0	10.0	1,1	506.0	21.0	1893.0	32.0	1134.0	
35	12/03/2004	05.00.00	0,6	847.0	7.0	1,0	501.0	30.0	1895.0	44.0	1155.0	
39	12/03/2004	09.00.00	-200	1545.0	-200.0	22,1	1353.0	-200.0	767.0	-200.0	2058.0	
...	...	...	...	...	...	...	...	...	...	...	...	
9352	04/04/2005	10.00.00	3,1	1314.0	-200.0	13,5	1101.0	472.0	539.0	190.0	1374.0	
9353	04/04/2005	11.00.00	2,4	1163.0	-200.0	11,4	1027.0	353.0	604.0	179.0	1264.0	
9354	04/04/2005	12.00.00	2,4	1142.0	-200.0	12,4	1063.0	293.0	603.0	175.0	1241.0	
9355	04/04/2005	13.00.00	2,1	1003.0	-200.0	9,5	961.0	235.0	702.0	156.0	1041.0	
9356	04/04/2005	14.00.00	2,2	1071.0	-200.0	11,9	1047.0	265.0	654.0	168.0	1129.0	

8450 rows × 17 columns

In [42]:
```
subset2
```

Out[42]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.0	1056.0	113.0	1692.0	
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.0	1174.0	92.0	1559.0	
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.0	1140.0	114.0	1555.0	
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.0	1092.0	122.0	1584.0	
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.0	1205.0	116.0	1490.0	
...	...	...	...	...	...	...	...	...	...	...	...	
1226	30/04/2004	20.00.00	4,4	1449.0	501.0	19,5	1282.0	254.0	625.0	133.0	2100.0	
1227	30/04/2004	21.00.00	3,1	1363.0	234.0	15,1	1152.0	189.0	684.0	110.0	1951.0	
1228	30/04/2004	22.00.00	3	1371.0	212.0	14,6	1136.0	174.0	689.0	102.0	1927.0	
1229	30/04/2004	23.00.00	3,1	1406.0	275.0	13,7	1107.0	167.0	718.0	108.0	1872.0	
1230	01/05/2004	00.00.00	3,5	1425.0	275.0	15,2	1155.0	185.0	709.0	110.0	1936.0	

891 rows × 17 columns

In [43]:
```
merged_data=pd.concat([subset1,subset2],ignore_index=True)
```

In [44]:
```
merged_data
```

Out[44]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT
0	11/03/2004	04.00.00	-200	1011.0	14.0	1,3	527.0	21.0	1818.0	34.0	1197.0	
1	11/03/2004	05.00.00	0,7	1066.0	8.0	1,1	512.0	16.0	1918.0	28.0	1182.0	
2	12/03/2004	04.00.00	-200	831.0	10.0	1,1	506.0	21.0	1893.0	32.0	1134.0	
3	12/03/2004	05.00.00	0,6	847.0	7.0	1,0	501.0	30.0	1895.0	44.0	1155.0	
4	12/03/2004	09.00.00	-200	1545.0	-200.0	22,1	1353.0	-200.0	767.0	-200.0	2058.0	
...	...	...	...	...	...	...	...	...	...	...	...	
9336	30/04/2004	20.00.00	4,4	1449.0	501.0	19,5	1282.0	254.0	625.0	133.0	2100.0	
9337	30/04/2004	21.00.00	3,1	1363.0	234.0	15,1	1152.0	189.0	684.0	110.0	1951.0	
9338	30/04/2004	22.00.00	3	1371.0	212.0	14,6	1136.0	174.0	689.0	102.0	1927.0	
9339	30/04/2004	23.00.00	3,1	1406.0	275.0	13,7	1107.0	167.0	718.0	108.0	1872.0	
9340	01/05/2004	00.00.00	3,5	1425.0	275.0	15,2	1155.0	185.0	709.0	110.0	1936.0	

9341 rows × 17 columns

In [45]:
```
#---------------------------data transformation---------------------#
```

In [46]:
```
df=df.rename(columns={'Date':'DATE','Time':'TIME'})
```

In [47]:
```
df
```

Out[47]:

	DATE	TIME	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT
0	10/03/2004	18.00.00	2,6	1360.0	150.0	11,9	1046.0	166.0	1056.0	113.0	1692.0	
1	10/03/2004	19.00.00	2	1292.0	112.0	9,4	955.0	103.0	1174.0	92.0	1559.0	
2	10/03/2004	20.00.00	2,2	1402.0	88.0	9,0	939.0	131.0	1140.0	114.0	1555.0	
3	10/03/2004	21.00.00	2,2	1376.0	80.0	9,2	948.0	172.0	1092.0	122.0	1584.0	
4	10/03/2004	22.00.00	1,6	1272.0	51.0	6,5	836.0	131.0	1205.0	116.0	1490.0	
...	...	...	...	...	...	...	...	...	...	...	...	
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

9471 rows × 17 columns

In [49]:
```
df['NMHC(GT)']=df['NMHC(GT)'].map({150:'nan'})
```

In [50]:
```
df
```

Out[50]:

	DATE	TIME	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT
0	10/03/2004	18.00.00	2,6	1360.0	nan	11,9	1046.0	166.0	1056.0	113.0	1692.0	
1	10/03/2004	19.00.00	2	1292.0	NaN	9,4	955.0	103.0	1174.0	92.0	1559.0	
2	10/03/2004	20.00.00	2,2	1402.0	NaN	9,0	939.0	131.0	1140.0	114.0	1555.0	
3	10/03/2004	21.00.00	2,2	1376.0	NaN	9,2	948.0	172.0	1092.0	122.0	1584.0	
4	10/03/2004	22.00.00	1,6	1272.0	NaN	6,5	836.0	131.0	1205.0	116.0	1490.0	
...	...	...	...	...	...	...	...	...	...	...	...	
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

9471 rows × 17 columns

In [51]:

```
df=pd.get_dummies(df,columns=['T'])
```

In [52]:

```
df
```

Out[52]:

	DATE	TIME	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	...	T_9,0	T_9,1
0	10/03/2004	18.00.00	2,6	1360.0	nan	11,9	1046.0	166.0	1056.0	113.0	...	0	0
1	10/03/2004	19.00.00	2	1292.0	NaN	9,4	955.0	103.0	1174.0	92.0	...	0	0
2	10/03/2004	20.00.00	2,2	1402.0	NaN	9,0	939.0	131.0	1140.0	114.0	...	0	0
3	10/03/2004	21.00.00	2,2	1376.0	NaN	9,2	948.0	172.0	1092.0	122.0	...	0	0
4	10/03/2004	22.00.00	1,6	1272.0	NaN	6,5	836.0	131.0	1205.0	116.0	...	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9466	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0	0
9467	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0	0
9468	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0	0
9469	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0	0
9470	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0	0

9471 rows × 453 columns

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js