In [1]:
```
#Data cleaning
import pandas as pd
```

In [2]:
```
import numpy as np
```

In [3]:
```
import matplotlib.pyplot as plt
```

In [4]:
```
import random as rd
```

In [5]:
```
ds = pd.read_csv("AirQuality1.csv")
```

In [6]:
```
ds_heart = pd.read_csv("heart.csv")
```

In [7]:
```
ds.head
```

Out[7]:
```
<bound method NDFrame.head of             Date      Time  CO(GT)  P
T08.S1(CO)  NMHC(GT)  C6H6_(GT)
0     10/03/2004  18.00.00       2            6      1360        15
0  \
1     10/03/2004  19.00.00       2         1292       112
9
2     10/03/2004  20.00.00       2            2      1402         8
8
3     10/03/2004  21.00.00       2            2      1376         8
0
4     10/03/2004  22.00.00       1            6      1272         5
1
...          ...       ...     ...          ...       ...
...
9352  04/04/2005  10.00.00       3            1      1314       -20
0
9353  04/04/2005  11.00.00       2            4      1163       -20
0
9354  04/04/2005  12.00.00       2            4      1142       -20
0
9355  04/04/2005  13.00.00       2            1      1003       -20
0
9356  04/04/2005  14.00.00       2            2      1071       -20
0

      PT08.S2(NMHC)  NOx(GT)  PT08.S3(NOx)  NO2(GT)  PT08.S4(NO2)
0                11        9          1046      166          1056
\
1                 4      955           103     1174            92
2                 9        0           939      131          1140
3                 9        2           948      172          1092
4                 6        5           836      131          1205
...             ...      ...           ...      ...           ...
9352             13        5          1101      472           539
9353             11        4          1027      353           604
9354             12        4          1063      293           603
9355              9        5           961      235           702
9356             11        9          1047      265           654

      PT08.S5(O3)     T    RH  AH  Unnamed: 15  Unnamed: 16  Unname
d: 17
0             113  1692  1268  13            6         48.0
9.0  \
1            1559   972    13   3           47          7.0
0.0
2             114  1555  1074  11            9         54.0
0.0
3             122  1584  1203  11            0         60.0
0.0
4             116  1490  1110  11            2         59.0
6.0
...           ...   ...   ...  ..          ...          ...
...
9352          190  1374  1729  21            9         29.0
3.0
9353          179  1264  1269  24            3         23.0
7.0
```

```
9354          175  1241  1092  26            9         18.0
3.0
9355          156  1041   770  28            3         13.0
5.0
9356          168  1129   816  28            5         13.0
1.0

      Unnamed: 18  Unnamed: 19
0             0.0       7578.0
1          7255.0          NaN
2             0.0       7502.0
3             0.0       7867.0
4             0.0       7888.0
...           ...          ...
9352          0.0       7568.0
9353          0.0       7119.0
9354          0.0       6406.0
9355          0.0       5139.0
9356          0.0       5028.0

[9357 rows x 20 columns]>
```

In [8]:
```
ds.info
```

Out[8]:
```
<bound method DataFrame.info of            Date      Time  CO(GT)
PT08.S1(CO)  NMHC(GT)  C6H6_(GT)
0     10/03/2004  18.00.00       2            6      1360        15
0  \
1     10/03/2004  19.00.00       2         1292       112
9
2     10/03/2004  20.00.00       2            2      1402         8
8
3     10/03/2004  21.00.00       2            2      1376         8
0
4     10/03/2004  22.00.00       1            6      1272         5
1
...          ...       ...     ...          ...       ...
...
9352  04/04/2005  10.00.00       3            1      1314       -20
0
9353  04/04/2005  11.00.00       2            4      1163       -20
0
9354  04/04/2005  12.00.00       2            4      1142       -20
0
9355  04/04/2005  13.00.00       2            1      1003       -20
0
9356  04/04/2005  14.00.00       2            2      1071       -20
0

      PT08.S2(NMHC)  NOx(GT)  PT08.S3(NOx)  NO2(GT)  PT08.S4(NO2)
0                11        9          1046      166          1056
\
1                 4      955           103     1174            92
2                 9        0           939      131          1140
3                 9        2           948      172          1092
4                 6        5           836      131          1205
...             ...      ...           ...      ...           ...
9352             13        5          1101      472           539
9353             11        4          1027      353           604
9354             12        4          1063      293           603
9355              9        5           961      235           702
9356             11        9          1047      265           654

      PT08.S5(O3)     T    RH  AH  Unnamed: 15  Unnamed: 16  Unname
d: 17
0             113  1692  1268  13            6         48.0
9.0  \
1            1559   972    13   3           47          7.0
0.0
2             114  1555  1074  11            9         54.0
0.0
3             122  1584  1203  11            0         60.0
0.0
4             116  1490  1110  11            2         59.0
6.0
...           ...   ...   ...  ..          ...          ...
...
9352          190  1374  1729  21            9         29.0
3.0
9353          179  1264  1269  24            3         23.0
7.0
```

```
9354          175  1241  1092  26            9         18.0
3.0
9355          156  1041   770  28            3         13.0
5.0
9356          168  1129   816  28            5         13.0
1.0

      Unnamed: 18  Unnamed: 19
0             0.0       7578.0
1          7255.0          NaN
2             0.0       7502.0
3             0.0       7867.0
4             0.0       7888.0
...           ...          ...
9352          0.0       7568.0
9353          0.0       7119.0
9354          0.0       6406.0
9355          0.0       5139.0
9356          0.0       5028.0

[9357 rows x 20 columns]>
```

In [9]:
```
ds.isnull().sum()
```

Out[9]:
```
Date                0
Time                0
CO(GT)              0
PT08.S1(CO)         0
NMHC(GT)            0
C6H6_(GT)           0
PT08.S2(NMHC)       0
NOx(GT)             0
PT08.S3(NOx)        0
NO2(GT)             0
PT08.S4(NO2)        0
PT08.S5(O3)         0
T                   0
RH                  0
AH                  0
Unnamed: 15         0
Unnamed: 16        61
Unnamed: 17       366
Unnamed: 18       366
Unnamed: 19      2442
dtype: int64
```

In [10]:
```
ds.dropna()
```

Out[10]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6_(GT)	PT08.S2(NMHC)	NC
0	10/03/2004	18.00.00	2	6	1360	150	11	
2	10/03/2004	20.00.00	2	2	1402	88	9	
3	10/03/2004	21.00.00	2	2	1376	80	9	
4	10/03/2004	22.00.00	1	6	1272	51	6	
5	10/03/2004	23.00.00	1	2	1197	38	4	
...	...	...	...	...	...	...	...	
9352	04/04/2005	10.00.00	3	1	1314	-200	13	
9353	04/04/2005	11.00.00	2	4	1163	-200	11	
9354	04/04/2005	12.00.00	2	4	1142	-200	12	
9355	04/04/2005	13.00.00	2	1	1003	-200	9	
9356	04/04/2005	14.00.00	2	2	1071	-200	11	

6915 rows × 20 columns

In [11]:
```
#Data integration
ds1 = ds.loc[111:999, ['Date', 'Time', 'C6H6_(GT)', 'RH']]
```

In [12]:
```
ds2 = ds.iloc[[1,3,5,2,4,22,43,54,67,7,8,9,50,10,11]]
```

In [13]:
```
ds_integration = pd.concat([ds1,ds2])
```

In [14]:
```
ds_integration
```

Out[14]:

	Date	Time	C6H6_(GT)	RH	CO(GT)	PT08.S1(CO)	NMHC(GT)	PT08.S2(NMHC
111	15/03/2004	09.00.00	618	2184	NaN	NaN	NaN	Nal
112	15/03/2004	10.00.00	438	1973	NaN	NaN	NaN	Nal
113	15/03/2004	11.00.00	334	1798	NaN	NaN	NaN	Nal
114	15/03/2004	12.00.00	221	1522	NaN	NaN	NaN	Nal
115	15/03/2004	13.00.00	207	1404	NaN	NaN	NaN	Nal
...	...	...	...	...	...	...	...	.
8	11/03/2004	02.00.00	24	620	0.0	9.0	1094.0	2.
9	11/03/2004	03.00.00	19	501	0.0	6.0	1010.0	1.
50	12/03/2004	20.00.00	488	1887	6.0	6.0	1843.0	32.
10	11/03/2004	04.00.00	1	10	-200.0	1011.0	14.0	3.
11	11/03/2004	05.00.00	8	422	0.0	7.0	1066.0	1.

904 rows × 20 columns

In [15]:
```
#Data transformation
ds_integration.transpose()
```

Out[15]:

	111	112	113	114	115	116	
Date	15/03/2004	15/03/2004	15/03/2004	15/03/2004	15/03/2004	15/03/2004	15/03/2
Time	09.00.00	10.00.00	11.00.00	12.00.00	13.00.00	14.00.00	15.0
C6H6_(GT)	618	438	334	221	207	191	
RH	2184	1973	1798	1522	1404	1263	
CO(GT)	NaN	NaN	NaN	NaN	NaN	NaN	
PT08.S1(CO)	NaN	NaN	NaN	NaN	NaN	NaN	
NMHC(GT)	NaN	NaN	NaN	NaN	NaN	NaN	
PT08.S2(NMHC)	NaN	NaN	NaN	NaN	NaN	NaN	
NOx(GT)	NaN	NaN	NaN	NaN	NaN	NaN	
PT08.S3(NOx)	NaN	NaN	NaN	NaN	NaN	NaN	
NO2(GT)	NaN	NaN	NaN	NaN	NaN	NaN	
PT08.S4(NO2)	NaN	NaN	NaN	NaN	NaN	NaN	
PT08.S5(O3)	NaN	NaN	NaN	NaN	NaN	NaN	
T	NaN	NaN	NaN	NaN	NaN	NaN	
AH	NaN	NaN	NaN	NaN	NaN	NaN	
Unnamed: 15	NaN	NaN	NaN	NaN	NaN	NaN	
Unnamed: 16	NaN	NaN	NaN	NaN	NaN	NaN	
Unnamed: 17	NaN	NaN	NaN	NaN	NaN	NaN	
Unnamed: 18	NaN	NaN	NaN	NaN	NaN	NaN	
Unnamed: 19	NaN	NaN	NaN	NaN	NaN	NaN	

20 rows × 904 columns

In [16]:
```
ds.drop(columns = "NOx(GT)")
```

Out[16]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6_(GT)	PT08.S2(NMHC)	PT
0	10/03/2004	18.00.00	2	6	1360	150	11	
1	10/03/2004	19.00.00	2	1292	112	9	4	
2	10/03/2004	20.00.00	2	2	1402	88	9	
3	10/03/2004	21.00.00	2	2	1376	80	9	
4	10/03/2004	22.00.00	1	6	1272	51	6	
...	...	...	...	...	...	...	...	
9352	04/04/2005	10.00.00	3	1	1314	-200	13	
9353	04/04/2005	11.00.00	2	4	1163	-200	11	
9354	04/04/2005	12.00.00	2	4	1142	-200	12	
9355	04/04/2005	13.00.00	2	1	1003	-200	9	
9356	04/04/2005	14.00.00	2	2	1071	-200	11	

9357 rows × 19 columns

In [17]:
```
ds2.drop(1)
```

Out[17]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6_(GT)	PT08.S2(NMHC)	NOx(
3	10/03/2004	21.00.00	2	2	1376	80	9	
5	10/03/2004	23.00.00	1	2	1197	38	4	
2	10/03/2004	20.00.00	2	2	1402	88	9	
4	10/03/2004	22.00.00	1	6	1272	51	6	
22	11/03/2004	16.00.00	2	2	1292	95	8	
43	12/03/2004	13.00.00	2	5	1252	140	11	
54	13/03/2004	00.00.00	2	7	1280	122	9	
67	13/03/2004	13.00.00	2	8	1328	154	12	
7	11/03/2004	01.00.00	1	1136	31	3	3	
8	11/03/2004	02.00.00	0	9	1094	24	2	
9	11/03/2004	03.00.00	0	6	1010	19	1	
50	12/03/2004	20.00.00	6	6	1843	488	32	
10	11/03/2004	04.00.00	-200	1011	14	1	3	
11	11/03/2004	05.00.00	0	7	1066	8	1	

In [18]:
```
ds.melt()
```

Out[18]:

	variable	value
0	Date	10/03/2004
1	Date	10/03/2004
2	Date	10/03/2004
3	Date	10/03/2004
4	Date	10/03/2004
...	...	...
187135	Unnamed: 19	7568.0
187136	Unnamed: 19	7119.0
187137	Unnamed: 19	6406.0
187138	Unnamed: 19	5139.0
187139	Unnamed: 19	5028.0

187140 rows × 2 columns

In [19]:
```
ds_merged = pd.concat([ds,ds_heart])
```

In [20]:
```
ds_merged
```

Out[20]:

	Date	Time	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6_(GT)	PT08.S2(NMHC)	NO
0	10/03/2004	18.00.00	2.0	6.0	1360.0	150.0	11.0	
1	10/03/2004	19.00.00	2.0	1292.0	112.0	9.0	4.0	
2	10/03/2004	20.00.00	2.0	2.0	1402.0	88.0	9.0	
3	10/03/2004	21.00.00	2.0	2.0	1376.0	80.0	9.0	
4	10/03/2004	22.00.00	1.0	6.0	1272.0	51.0	6.0	
...	...	...	...	...	...	...	...	
1020	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1021	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1023	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1024	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

10382 rows × 34 columns

In [21]:
```
#Error correcting:
ds_heart['ca'].unique()
```

Out[21]:
```
array([2, 0, 1, 3, 4])
```

In [22]:
```
ds_heart.ca.value_counts()
```

Out[22]:
```
ca
0    578
1    226
2    134
3     69
4     18
Name: count, dtype: int64
```

In [ ]:
```

```