01_Dataset

December 2, 2021

[1]:
```
# ZOOM de la SFGP - GTIAP
# 02 / 12 / 2021
# Roda Bounaceur
# LRGP - Nancy
# roda.bounaceur@univ-lorraine.fr
#
#
# Importation des bibliothèques de bases - Pandas et Numpy - pour manipuler les␣
 ↪data
#
import pandas as pd
import numpy as np
```

[2]:
```
#
# Importation de matplotlib pour les graphs et sklearn pour les algorithmes de␣
 ↪ML
#
import matplotlib
import sklearn
#
# Versions utilisées
#
print('Version de matplotlib = ',matplotlib.__version__)
print('Version de sklearn = ',sklearn.__version__)
!python -V
```

```
Version de matplotlib =  3.3.4
Version de sklearn =  0.24.1
Python 3.8.8
```

[ ]:
```
#
# Analyse simple du dataset
#
# 0. Importer le dataset
# 1. Repérer les features et les targets
# 2. Repérer les nb lignes et colonnes
# 3. type de variables
# 4. analyse des valeurs manquantes
```

```
# 5. analyse des valeurs négatives, très petites, etc ....
```

[3]:
```
#
# Importation du dataset
#
df =  pd.read_csv('Dataset_Complet.csv',sep = ',')
```

[4]:
```
#
# Affichage du dataset
#
df.head()
```

[4]:

	Pressure_(bar)	Resident_Time_(s)	Temperature_(C)	Time_(sec)	\
0	0.159	1.985937	958.0	0.000000e+00	
1	0.159	1.985937	958.0	1.440000e-12	
2	0.159	1.985937	958.0	1.670000e-09	
3	0.159	1.985937	958.0	1.700000e-08	
4	0.159	1.985937	958.0	4.430000e-08	

	Mole_fraction_H2	Mole_fraction_CH4	Mole_fraction_Biomasse	\
0	0.000000e+00	0.000000e+00	1.0	
1	8.830000e-13	2.840000e-43	1.0	
2	1.020000e-09	2.270000e-34	1.0	
3	1.040000e-08	2.390000e-31	1.0	
4	2.710000e-08	4.240000e-30	1.0	

	Mole_fraction_C2H4Z	Mole_fraction_pC3H4	Mole_fraction_C4H4	…	\
0	0.000000e+00	0.000000e+00	0.000000e+00	…	
1	3.060000e-26	5.910000e-40	3.330000e-12	…	
2	3.310000e-20	3.590000e-28	3.850000e-09	…	
3	3.960000e-18	3.680000e-24	3.920000e-08	…	
4	3.350000e-17	1.600000e-22	1.020000e-07	…	

	Mole_fraction_fluorene	Mole_fraction_A3	Mole_fraction_A4	\
0	0.000000e+00	0.000000e+00	0.000000e+00	
1	1.040000e-103	5.730000e-70	4.400000e-96	
2	3.980000e-70	2.240000e-49	2.400000e-69	
3	2.990000e-58	1.930000e-42	1.100000e-60	
4	1.640000e-53	1.420000e-39	3.170000e-57	

	Mole_fraction_A5	Mole_fraction_A6	Mole_fraction_A7	Mole_fraction_A8	\
0	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
1	3.860000e-84	3.680000e-107	5.420000e-115	1.170000e-117	
2	1.550000e-60	4.130000e-77	6.960000e-82	5.290000e-85	
3	1.590000e-52	3.380000e-66	3.040000e-70	1.640000e-72	
4	3.450000e-49	1.290000e-61	1.600000e-65	7.090000e-67	

```
   Mole_fraction_A9  Mole_fraction_A10  Mole_fraction_C10H10
0      0.000000e+00       0.000000e+00          0.000000e+00
1      2.010000e-93      1.550000e-117          1.630000e-48
2      9.360000e-67       1.530000e-84          5.660000e-39
3      8.220000e-58       1.280000e-72          3.980000e-35
4      5.140000e-54       1.250000e-67          1.900000e-33

[5 rows x 35 columns]
```

[5]:
```
#
# nb lignes colonnes
#
df.shape
```

[5]:
```
(14756, 35)
```

[6]:
```
#
# Type de variables
#
df.dtypes
```

[6]:
```
Pressure_(bar)                float64
Resident_Time_(s)             float64
Temperature_(C)               float64
Time_(sec)                    float64
Mole_fraction_H2              float64
Mole_fraction_CH4             float64
Mole_fraction_Biomasse        float64
Mole_fraction_C2H4Z           float64
Mole_fraction_pC3H4           float64
Mole_fraction_C4H4            float64
Mole_fraction_C4H6Z2          float64
Mole_fraction_iC5H8           float64
Mole_fraction_C6H6#           float64
Mole_fraction_toluene         float64
Mole_fraction_etC6H5          float64
Mole_fraction_styrene         float64
Mole_fraction_naphtalene      float64
Mole_fraction_indene          float64
Mole_fraction_indane          float64
Mole_fraction_phenanthrene    float64
Mole_fraction_pyrene          float64
Mole_fraction_A1              float64
Mole_fraction_chrysene        float64
Mole_fraction_antra           float64
Mole_fraction_A2              float64
Mole_fraction_fluorene        float64
```

```
Mole_fraction_A3              float64
Mole_fraction_A4              float64
Mole_fraction_A5              float64
Mole_fraction_A6              float64
Mole_fraction_A7              float64
Mole_fraction_A8              float64
Mole_fraction_A9              float64
Mole_fraction_A10             float64
Mole_fraction_C10H10          float64
dtype: object
```

[7]:
```
#
# test de la présence de NaN
#
df.isna()
```

[7]:
```
       Pressure_(bar)  Resident_Time_(s)  Temperature_(C)  Time_(sec)  \
0               False              False            False       False
1               False              False            False       False
2               False              False            False       False
3               False              False            False       False
4               False              False            False       False
…                 …                  …                …           …
14751           False              False            False       False
14752           False              False            False       False
14753           False              False            False       False
14754           False              False            False       False
14755           False              False            False       False

       Mole_fraction_H2  Mole_fraction_CH4  Mole_fraction_Biomasse  \
0                 False              False                   False
1                 False              False                   False
2                 False              False                   False
3                 False              False                   False
4                 False              False                   False
…                   …                  …                       …
14751             False              False                   False
14752             False              False                   False
14753             False              False                   False
14754             False              False                   False
14755             False              False                   False

       Mole_fraction_C2H4Z  Mole_fraction_pC3H4  Mole_fraction_C4H4  …  \
0                    False                False               False  …
1                    False                False               False  …
2                    False                False               False  …
3                    False                False               False  …
```

```
4                   False                False                False  …
…                   …                  …                  …  …
14751               False                False                False  …
14752               False                False                False  …
14753               False                False                False  …
14754               False                False                False  …
14755               False                False                False  …

       Mole_fraction_fluorene  Mole_fraction_A3  Mole_fraction_A4  \
0                       False             False             False
1                       False             False             False
2                       False             False             False
3                       False             False             False
4                       False             False             False
…                       …               …                 …
14751                   False             False             False
14752                   False             False             False
14753                   False             False             False
14754                   False             False             False
14755                   False             False             False

       Mole_fraction_A5  Mole_fraction_A6  Mole_fraction_A7  Mole_fraction_A8  \
0                 False             False             False             False
1                 False             False             False             False
2                 False             False             False             False
3                 False             False             False             False
4                 False             False             False             False
…                 …               …                 …               …
14751             False             False             False             False
14752             False             False             False             False
14753             False             False             False             False
14754             False             False             False             False
14755             False             False             False             False

       Mole_fraction_A9  Mole_fraction_A10  Mole_fraction_C10H10
0                 False              False                 False
1                 False              False                 False
2                 False              False                 False
3                 False              False                 False
4                 False              False                 False
…                 …                …                   …
14751             False              False                 False
14752             False              False                 False
14753             False              False                 False
14754             False              False                 False
14755             False              False                 False
```

```
[14756 rows x 35 columns]
```

[8]:
```
#
# Comptage du nombre de valeurs NaN
#
df.isna().sum()
```

[8]:
```
Pressure_(bar)                0
Resident_Time_(s)             0
Temperature_(C)               0
Time_(sec)                    0
Mole_fraction_H2              0
Mole_fraction_CH4             0
Mole_fraction_Biomasse        0
Mole_fraction_C2H4Z           0
Mole_fraction_pC3H4           0
Mole_fraction_C4H4            0
Mole_fraction_C4H6Z2          0
Mole_fraction_iC5H8           0
Mole_fraction_C6H6#           0
Mole_fraction_toluene         0
Mole_fraction_etC6H5          0
Mole_fraction_styrene         0
Mole_fraction_naphtalene      0
Mole_fraction_indene          0
Mole_fraction_indane          0
Mole_fraction_phenanthrene    0
Mole_fraction_pyrene          0
Mole_fraction_A1              0
Mole_fraction_chrysene        0
Mole_fraction_antra           0
Mole_fraction_A2              0
Mole_fraction_fluorene        0
Mole_fraction_A3              0
Mole_fraction_A4              0
Mole_fraction_A5              0
Mole_fraction_A6              0
Mole_fraction_A7              0
Mole_fraction_A8              0
Mole_fraction_A9              0
Mole_fraction_A10             0
Mole_fraction_C10H10          0
dtype: int64
```

[9]:
```
#
# Comptage du nombre de valeurs négatives
#
```

```
(df<0).sum()
```

[9]:
```
Pressure_(bar)                0
Resident_Time_(s)             0
Temperature_(C)               0
Time_(sec)                    0
Mole_fraction_H2              0
Mole_fraction_CH4             0
Mole_fraction_Biomasse        0
Mole_fraction_C2H4Z           0
Mole_fraction_pC3H4           0
Mole_fraction_C4H4            0
Mole_fraction_C4H6Z2          0
Mole_fraction_iC5H8           0
Mole_fraction_C6H6#           0
Mole_fraction_toluene         0
Mole_fraction_etC6H5          0
Mole_fraction_styrene         0
Mole_fraction_naphtalene      0
Mole_fraction_indene          0
Mole_fraction_indane          0
Mole_fraction_phenanthrene    0
Mole_fraction_pyrene          0
Mole_fraction_A1              0
Mole_fraction_chrysene        0
Mole_fraction_antra           0
Mole_fraction_A2              0
Mole_fraction_fluorene        0
Mole_fraction_A3              0
Mole_fraction_A4              0
Mole_fraction_A5              0
Mole_fraction_A6              0
Mole_fraction_A7              0
Mole_fraction_A8              0
Mole_fraction_A9              0
Mole_fraction_A10             0
Mole_fraction_C10H10          0
dtype: int64
```

[10]:
```
#
# Comptage du nombre de valeurs très petites
#
(df<1.0e-25).sum()
```

[10]:
```
Pressure_(bar)                0
Resident_Time_(s)             0
Temperature_(C)               0
```

```
Time_(sec)                    100
Mole_fraction_H2              100
Mole_fraction_CH4             822
Mole_fraction_Biomasse          0
Mole_fraction_C2H4Z           200
Mole_fraction_pC3H4           302
Mole_fraction_C4H4            100
Mole_fraction_C4H6Z2          300
Mole_fraction_iC5H8          1046
Mole_fraction_C6H6#           100
Mole_fraction_toluene         741
Mole_fraction_etC6H5          903
Mole_fraction_styrene         218
Mole_fraction_naphtalene      300
Mole_fraction_indene          709
Mole_fraction_indane         1134
Mole_fraction_phenanthrene    967
Mole_fraction_pyrene         1017
Mole_fraction_A1              692
Mole_fraction_chrysene       1259
Mole_fraction_antra          1042
Mole_fraction_A2              993
Mole_fraction_fluorene       1285
Mole_fraction_A3             1031
Mole_fraction_A4             1480
Mole_fraction_A5             1259
Mole_fraction_A6             1643
Mole_fraction_A7             1971
Mole_fraction_A8             1758
Mole_fraction_A9             1335
Mole_fraction_A10            1883
Mole_fraction_C10H10          887
dtype: int64
```

[11]:
```
#
# Remplacement des valeurs très petites
# Fonction where :
# Where cond is True, keep the original value.
# Where False, replace with corresponding value
#
df_filtre = df.where( ((df>1.0e-15) | (df == 0.0)) ,1.0e-15)
```

[12]:
```
#
# Affichage du nouveau dataset
#
df_filtre.head()
```

[12]:

	Pressure_(bar)	Resident_Time_(s)	Temperature_(C)	Time_(sec)	\
0	0.159	1.985937	958.0	0.000000e+00	
1	0.159	1.985937	958.0	1.440000e-12	
2	0.159	1.985937	958.0	1.670000e-09	
3	0.159	1.985937	958.0	1.700000e-08	
4	0.159	1.985937	958.0	4.430000e-08	

	Mole_fraction_H2	Mole_fraction_CH4	Mole_fraction_Biomasse	\
0	0.000000e+00	0.000000e+00	1.0	
1	8.830000e-13	1.000000e-15	1.0	
2	1.020000e-09	1.000000e-15	1.0	
3	1.040000e-08	1.000000e-15	1.0	
4	2.710000e-08	1.000000e-15	1.0	

	Mole_fraction_C2H4Z	Mole_fraction_pC3H4	Mole_fraction_C4H4	…	\
0	0.000000e+00	0.000000e+00	0.000000e+00	…	
1	1.000000e-15	1.000000e-15	3.330000e-12	…	
2	1.000000e-15	1.000000e-15	3.850000e-09	…	
3	1.000000e-15	1.000000e-15	3.920000e-08	…	
4	1.000000e-15	1.000000e-15	1.020000e-07	…	

	Mole_fraction_fluorene	Mole_fraction_A3	Mole_fraction_A4	\
0	0.000000e+00	0.000000e+00	0.000000e+00	
1	1.000000e-15	1.000000e-15	1.000000e-15	
2	1.000000e-15	1.000000e-15	1.000000e-15	
3	1.000000e-15	1.000000e-15	1.000000e-15	
4	1.000000e-15	1.000000e-15	1.000000e-15	

	Mole_fraction_A5	Mole_fraction_A6	Mole_fraction_A7	Mole_fraction_A8	\
0	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
1	1.000000e-15	1.000000e-15	1.000000e-15	1.000000e-15	
2	1.000000e-15	1.000000e-15	1.000000e-15	1.000000e-15	
3	1.000000e-15	1.000000e-15	1.000000e-15	1.000000e-15	
4	1.000000e-15	1.000000e-15	1.000000e-15	1.000000e-15	

	Mole_fraction_A9	Mole_fraction_A10	Mole_fraction_C10H10
0	0.000000e+00	0.000000e+00	0.000000e+00
1	1.000000e-15	1.000000e-15	1.000000e-15
2	1.000000e-15	1.000000e-15	1.000000e-15
3	1.000000e-15	1.000000e-15	1.000000e-15
4	1.000000e-15	1.000000e-15	1.000000e-15

[5 rows x 35 columns]

[ ]:
```
#
# Relation entre les inputs et les outputs
#
```

[13]:
```
# Les inputs
X  = df[['Pressure_(bar)', 'Resident_Time_(s)', 'Temperature_(C)']]
# Les outputs
y = df.drop(['Pressure_(bar)', 'Resident_Time_(s)', 'Temperature_(C)',␣
↪'Time_(sec)'],axis=1)
```

[14]:
```
#
# Analyse statistique simple
#
X.describe().T
```

[14]:

	count	mean	std	min	25%	50%	\
Pressure_(bar)	14756.0	0.094702	0.037526	0.03	0.0625	0.095	
Resident_Time_(s)	14756.0	1.097209	0.519141	0.20	0.6500	1.100	
Temperature_(C)	14756.0	950.497493	57.708335	850.00	900.0000	952.000	

	75%	max
Pressure_(bar)	0.128	0.159000
Resident_Time_(s)	1.550	1.985937
Temperature_(C)	1000.000	1050.000000

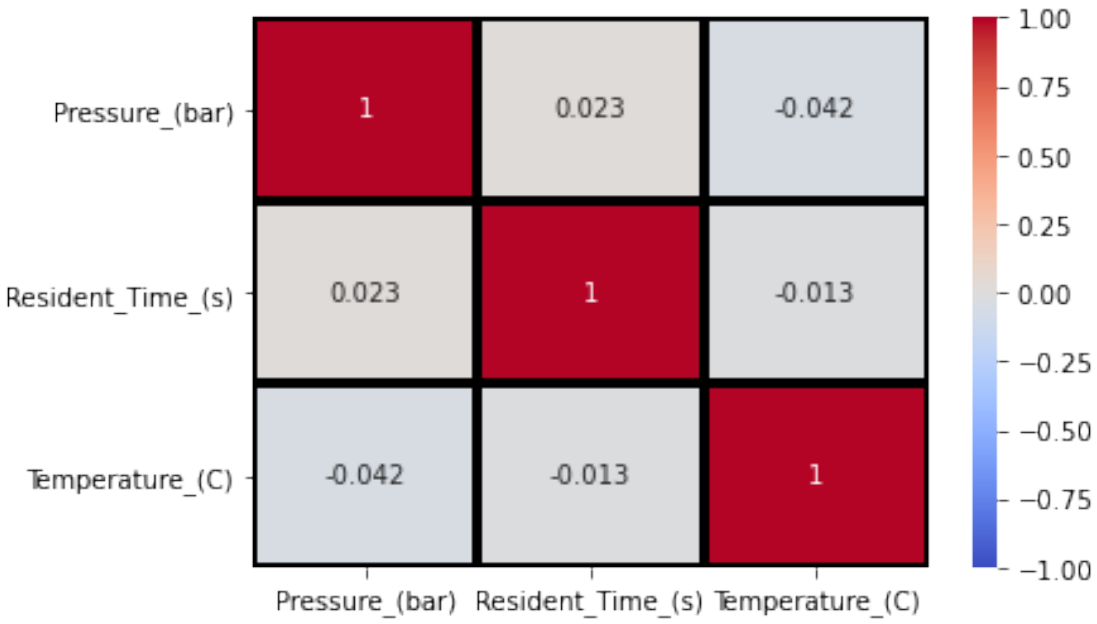
[15]:
```
#
# Matrice de corrélation
#
X.corr()
```

[15]:

	Pressure_(bar)	Resident_Time_(s)	Temperature_(C)
Pressure_(bar)	1.000000	0.022761	-0.041563
Resident_Time_(s)	0.022761	1.000000	-0.013416
Temperature_(C)	-0.041563	-0.013416	1.000000

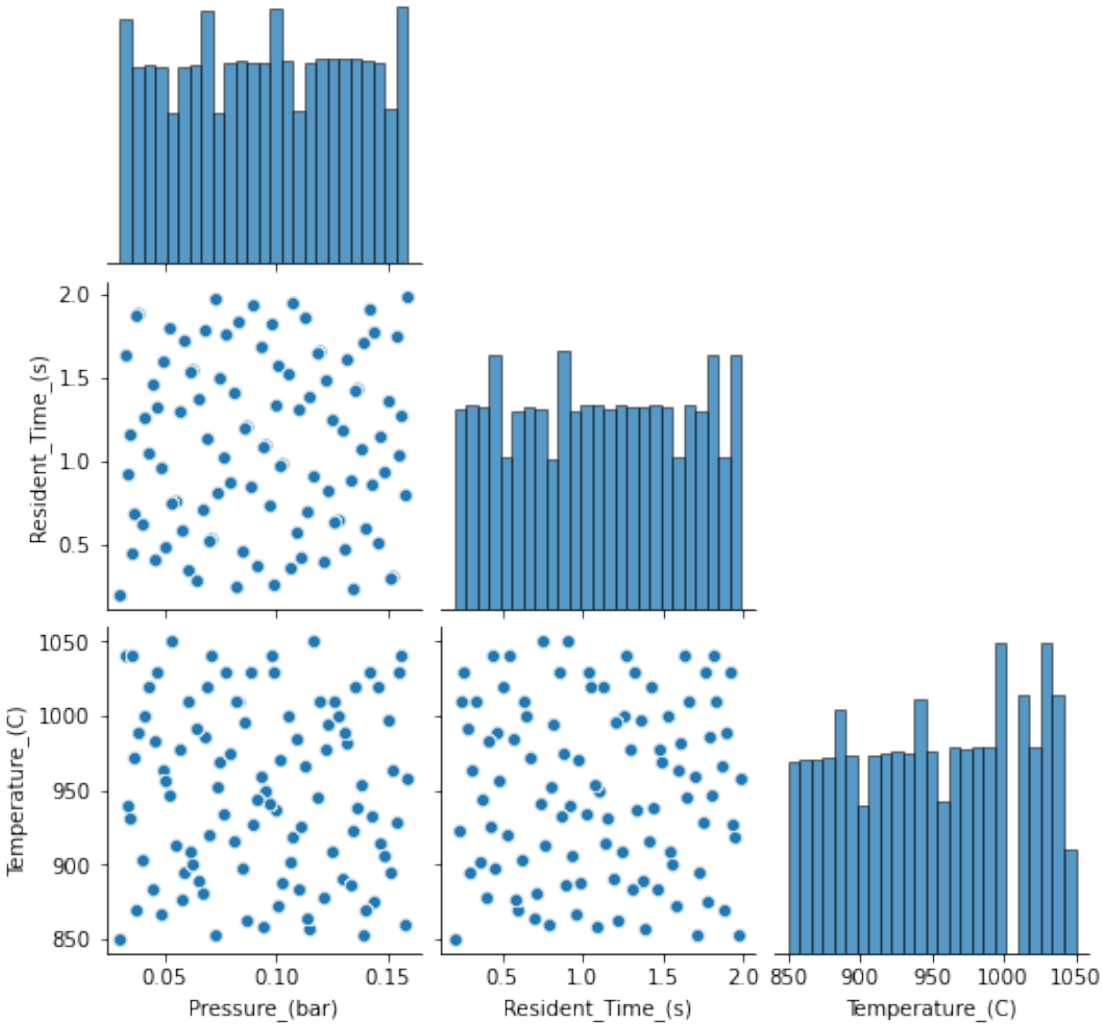
[19]:
```
#
# Matrice de corrélation - version graphique
#
sns.heatmap(X.
↪corr(),annot=True,vmin=-1,vmax=1,center=0,cmap='coolwarm',linewidth=3,linecolor='black')
```

[19]: <AxesSubplot:>



[18]:
```
#
# Courbes des plans projetés
#
import seaborn as sns
sns.pairplot(X, corner=True)
```

[18]:
```
<seaborn.axisgrid.PairGrid at 0x1bdb030c0d0>
```


Resident_Time_(s)
Temperature_(C)
Pressure_(bar)
Resident_Time_(s)
Temperature_(C)
2.0
1.5
1.0
0.5
1050
1000
950
900
850
0.05
0.10
0.15
0.5
1.0
1.5
2.0
850
900
950
1000
1050