eda

February 13, 2023

[19]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[20]:
```
df = pd.read_csv("data_prep.csv")
```

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5162 entries, 0 to 5161
Data columns (total 49 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Unnamed: 0
5162 non-null   int64
 1   Age
5162 non-null   float64
 2   Sortie Positif
5162 non-null   int64
 3   temps_psp
5162 non-null   float64
 4   temps_pe
5162 non-null   float64
 5   temps_fin
5162 non-null   float64
 6   Civilité_Madame
5162 non-null   float64
 7   Civilité_Monsieur
5162 non-null   float64
 8   Type 1er RDV_Entretien individuel
5162 non-null   float64
 9   Type 1er RDV_Webcam
5162 non-null   float64
 10  Taille dernière entreprise :_500 salariés et plus
5162 non-null   float64
```

```
 11  Taille dernière entreprise :_De 10 à 49 salariés
5162 non-null   float64
 12  Taille dernière entreprise :_De 50 à 499 salariés
5162 non-null   float64
 13  Taille dernière entreprise :_Moins de 10 salariés
5162 non-null   float64
 14  Taille dernière entreprise :_nan
5162 non-null   float64
 15  Secteur_AGRICULTURE ET PÊCHE, ESPACES NATURELS ET ESPACES VERTS, SOINS AUX
ANIMAUX  5162 non-null   float64
 16  Secteur_ARTS ET FACONNAGE D'OUVRAGES D'ART
5162 non-null   float64
 17  Secteur_BANQUE, ASSURANCE, IMMOBILIER
5162 non-null   float64
 18  Secteur_COMMERCE, VENTE ET GRANDE DISTRIBUTION
5162 non-null   float64
 19  Secteur_COMMUNICATION, MEDIA ET MULTIMEDIA
5162 non-null   float64
 20  Secteur_CONSTRUCTION, BÂTIMENT ET TRAVAUX PUBLICS
5162 non-null   float64
 21  Secteur_HÔTELLERIE- RESTAURATION TOURISME LOISIRS ET ANIMATION
5162 non-null   float64
 22  Secteur_INDUSTRIE
5162 non-null   float64
 23  Secteur_INSTALLATION ET MAINTENANCE
5162 non-null   float64
 24  Secteur_SANTE
5162 non-null   float64
 25  Secteur_SERVICES A LA PERSONNE ET A LA COLLECTIVITE
5162 non-null   float64
 26  Secteur_SPECTACLE
5162 non-null   float64
 27  Secteur_SUPPORT A L'ENTREPRISE
5162 non-null   float64
 28  Secteur_TRANSPORT ET LOGISTIQUE
5162 non-null   float64
 29  Secteur_nan
5162 non-null   float64
 30  Code Prescripteur_1
5162 non-null   float64
 31  Code Prescripteur_26
5162 non-null   float64
 32  Code Prescripteur_27
5162 non-null   float64
 33  Code Prescripteur_38
5162 non-null   float64
 34  Code Prescripteur_42
5162 non-null   float64
```

```
 35  Code Prescripteur_54
5162 non-null   float64
 36  Code Prescripteur_61
5162 non-null   float64
 37  Code Prescripteur_63
5162 non-null   float64
 38  Code Prescripteur_67
5162 non-null   float64
 39  Code Prescripteur_68
5162 non-null   float64
 40  Code Prescripteur_69
5162 non-null   float64
 41  Code Prescripteur_73
5162 non-null   float64
 42  Code Prescripteur_75
5162 non-null   float64
 43  Code Prescripteur_76
5162 non-null   float64
 44  Code Prescripteur_78
5162 non-null   float64
 45  Code Prescripteur_92
5162 non-null   float64
 46  Code Prescripteur_93
5162 non-null   float64
 47  Code Prescripteur_94
5162 non-null   float64
 48  Code Prescripteur_95
5162 non-null   float64
dtypes: float64(47), int64(2)
memory usage: 1.9 MB
```

[22]:

```
df = df.drop(columns="Unnamed: 0")
```

[23]:

```
df.head()
```

[23]:

```
        Age  Sortie Positif  temps_psp  temps_pe  temps_fin  Civilité_Madame  \
0  0.632653               0   0.095385  0.260073   0.480740              1.0
1  0.224490               0   0.043077  0.113553   0.546995              1.0
2  0.918367               0   0.046154  0.194139   0.408320              1.0
3  0.816327               1   0.064615  0.102564   0.303544              0.0
4  0.795918               1   0.052308  0.391941   0.543914              1.0

   Civilité_Monsieur  Type 1er RDV_Entretien individuel  Type 1er RDV_Webcam  \
0                0.0                                1.0                  0.0
1                0.0                                1.0                  0.0
2                0.0                                1.0                  0.0
3                1.0                                1.0                  0.0
```

```
4                0.0                                1.0                    0.0

   Taille dernière entreprise :_500 salariés et plus  …  \
0                                                0.0  …
1                                                0.0  …
2                                                0.0  …
3                                                0.0  …
4                                                0.0  …

   Code Prescripteur_68  Code Prescripteur_69  Code Prescripteur_73  \
0                   0.0                   0.0                   0.0
1                   0.0                   0.0                   0.0
2                   0.0                   0.0                   0.0
3                   0.0                   0.0                   0.0
4                   0.0                   0.0                   0.0

   Code Prescripteur_75  Code Prescripteur_76  Code Prescripteur_78  \
0                   0.0                   0.0                   0.0
1                   0.0                   0.0                   0.0
2                   0.0                   0.0                   0.0
3                   0.0                   0.0                   0.0
4                   0.0                   0.0                   0.0

   Code Prescripteur_92  Code Prescripteur_93  Code Prescripteur_94  \
0                   0.0                   0.0                   0.0
1                   0.0                   0.0                   0.0
2                   0.0                   0.0                   0.0
3                   0.0                   0.0                   0.0
4                   0.0                   0.0                   0.0

   Code Prescripteur_95
0                   0.0
1                   0.0
2                   0.0
3                   0.0
4                   0.0

[5 rows x 48 columns]
```

```
[24]: df.describe()
```

```
[24]:
               Age  Sortie Positif    temps_psp     temps_pe    temps_fin  \
count  5162.000000     5162.000000  5162.000000  5162.000000  5162.000000
mean      0.533336        0.245641     0.106407     0.156459     0.481013
std       0.227237        0.430508     0.063542     0.099806     0.095463
min       0.000000        0.000000     0.000000     0.000000     0.000000
25%       0.346939        0.000000     0.064615     0.095238     0.445300
```

```
50%      0.530612        0.000000     0.095385     0.124542     0.500770
75%      0.734694        0.000000     0.132308     0.179487     0.534669
max      1.000000        1.000000     1.000000     1.000000     1.000000

       Civilité_Madame  Civilité_Monsieur  Type 1er RDV_Entretien individuel  \
count      5162.000000        5162.000000                        5162.000000
mean          0.510074           0.489926                           0.901976
std           0.499947           0.499947                           0.297376
min           0.000000           0.000000                           0.000000
25%           0.000000           0.000000                           1.000000
50%           1.000000           0.000000                           1.000000
75%           1.000000           1.000000                           1.000000
max           1.000000           1.000000                           1.000000

       Type 1er RDV_Webcam  Taille dernière entreprise :_500 salariés et plus  \
count          5162.000000                                        5162.000000
mean              0.098024                                           0.019372
std               0.297376                                           0.137843
min               0.000000                                           0.000000
25%               0.000000                                           0.000000
50%               0.000000                                           0.000000
75%               0.000000                                           0.000000
max               1.000000                                           1.000000

       …  Code Prescripteur_68  Code Prescripteur_69  Code Prescripteur_73  \
count  …           5162.000000           5162.000000           5162.000000
mean   …              0.014723              0.043975              0.030608
std    …              0.120453              0.205060              0.172271
min    …              0.000000              0.000000              0.000000
25%    …              0.000000              0.000000              0.000000
50%    …              0.000000              0.000000              0.000000
75%    …              0.000000              0.000000              0.000000
max    …              1.000000              1.000000              1.000000

       Code Prescripteur_75  Code Prescripteur_76  Code Prescripteur_78  \
count           5162.000000           5162.000000           5162.000000
mean               0.129601              0.049981              0.100155
std                0.335896              0.217926              0.300236
min                0.000000              0.000000              0.000000
25%                0.000000              0.000000              0.000000
50%                0.000000              0.000000              0.000000
75%                0.000000              0.000000              0.000000
max                1.000000              1.000000              1.000000

       Code Prescripteur_92  Code Prescripteur_93  Code Prescripteur_94  \
count           5162.000000           5162.000000           5162.000000
mean               0.039907              0.114103              0.117978
```

```
std                   0.195760              0.317967              0.322613
min                   0.000000              0.000000              0.000000
25%                   0.000000              0.000000              0.000000
50%                   0.000000              0.000000              0.000000
75%                   0.000000              0.000000              0.000000
max                   1.000000              1.000000              1.000000

       Code Prescripteur_95
count           5162.000000
mean               0.059667
std                0.236891
min                0.000000
25%                0.000000
50%                0.000000
75%                0.000000
max                1.000000

[8 rows x 48 columns]
```

[25]:

```
df.groupby("Sortie Positif").count()
```

[25]:

```
                 Age  temps_psp  temps_pe  temps_fin  Civilité_Madame  \
Sortie Positif
0               3894       3894      3894       3894             3894
1               1268       1268      1268       1268             1268

                Civilité_Monsieur  Type 1er RDV_Entretien individuel  \
Sortie Positif
0                            3894                               3894
1                            1268                               1268

                Type 1er RDV_Webcam  \
Sortie Positif
0                              3894
1                              1268

                Taille dernière entreprise :_500 salariés et plus  \
Sortie Positif
0                                                            3894
1                                                            1268

                Taille dernière entreprise :_De 10 à 49 salariés  …  \
Sortie Positif                                                    …
0                                                           3894  …
1                                                           1268  …

                Code Prescripteur_68  Code Prescripteur_69  \
Sortie Positif
0                               3894                  3894
1                               1268                  1268

                Code Prescripteur_73  Code Prescripteur_75  \
Sortie Positif
0                               3894                  3894
1                               1268                  1268

                Code Prescripteur_76  Code Prescripteur_78  \
Sortie Positif
0                               3894                  3894
1                               1268                  1268

                Code Prescripteur_92  Code Prescripteur_93  \
Sortie Positif
0                               3894                  3894
1                               1268                  1268

                Code Prescripteur_94  Code Prescripteur_95
Sortie Positif
0                               3894                  3894
1                               1268                  1268

[2 rows x 47 columns]
```

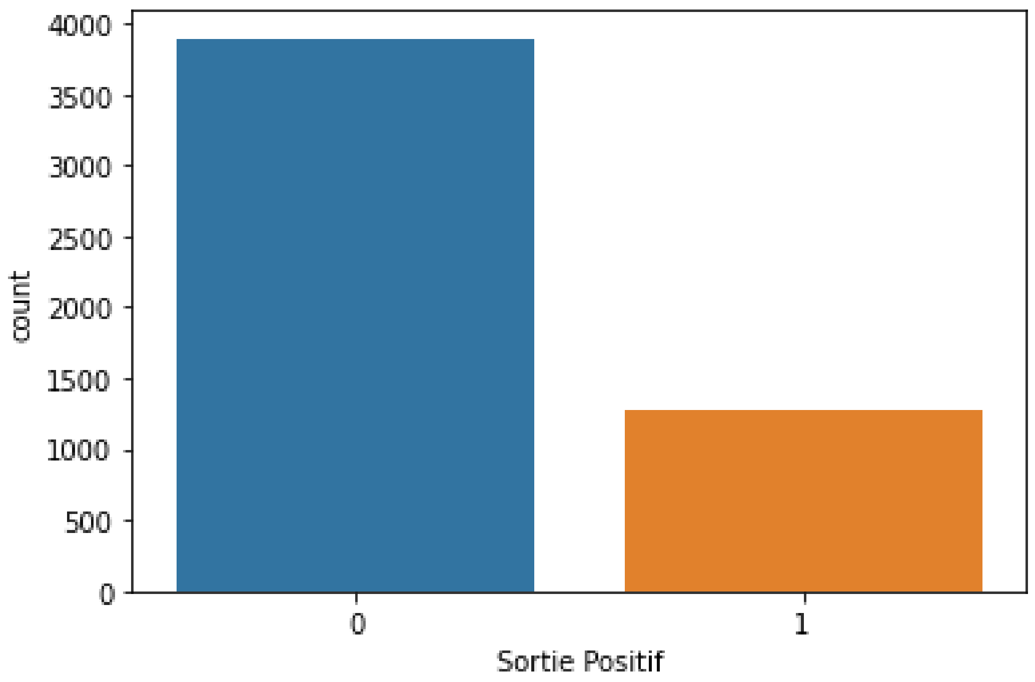
[26]:
```
1268/3894
```

[26]:
```
0.3256291730868002
```

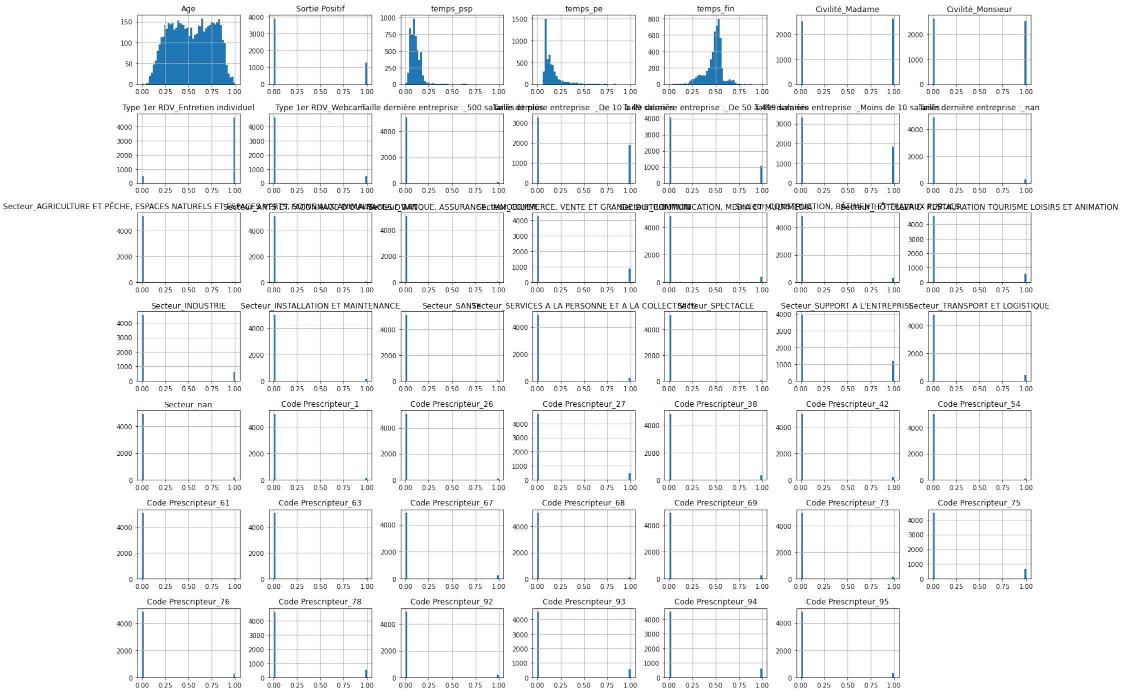
[27]:
```
sns.countplot(df['Sortie Positif'])
```

```
/home/n_fuentes/.local/lib/python3.8/site-packages/seaborn/_decorators.py:36:
FutureWarning: Pass the following variable as a keyword arg: x. From version
0.12, the only valid positional argument will be `data`, and passing other
arguments without an explicit keyword will result in an error or
misinterpretation.
  warnings.warn(
```

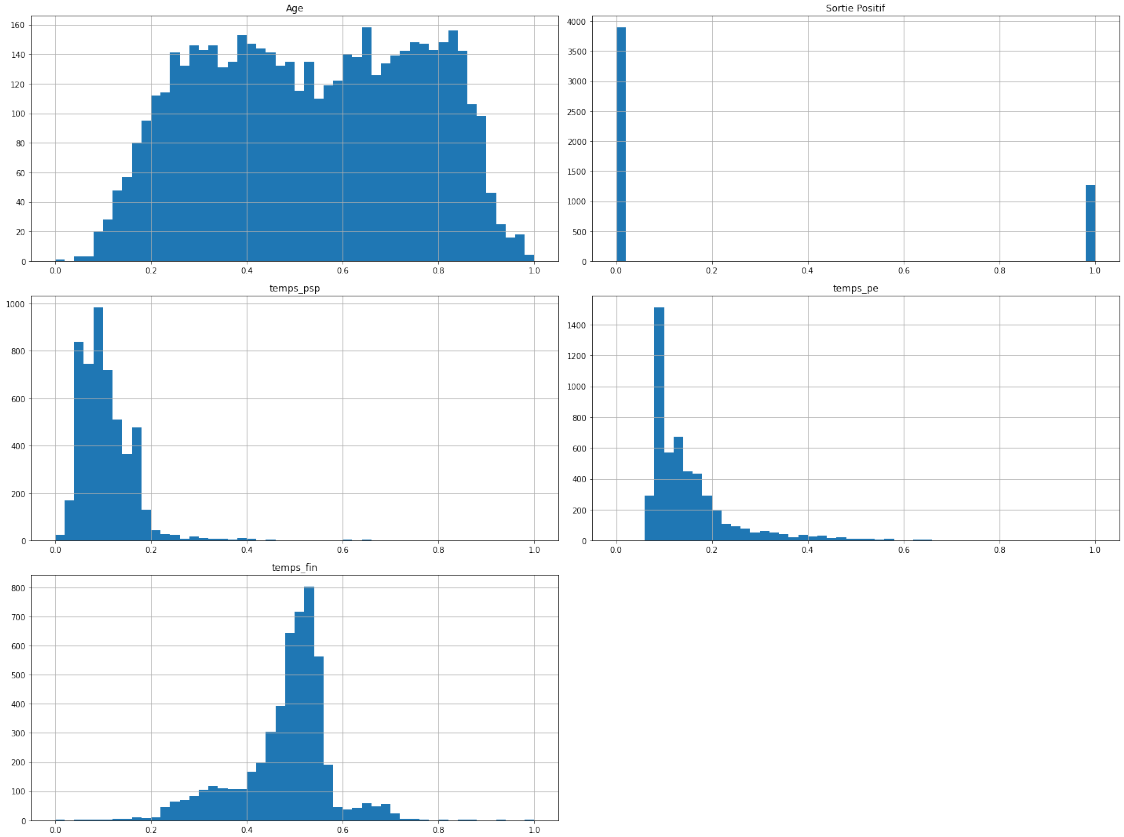
[27]:
```
<AxesSubplot:xlabel='Sortie Positif', ylabel='count'>
```



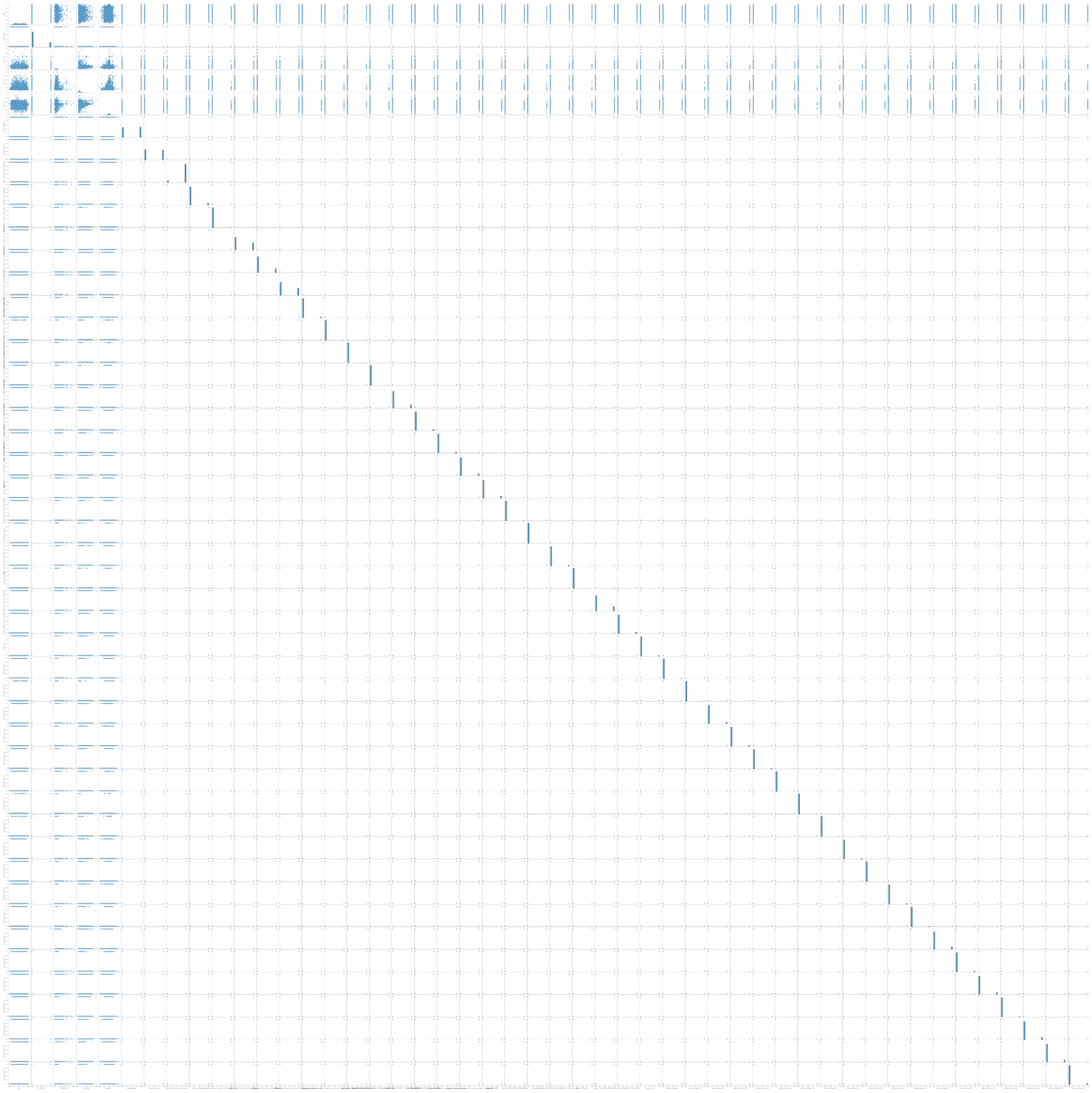
```
[28]: df.hist(bins=50, figsize=(20,15))
      plt.tight_layout()
      plt.show()
```



```
[35]: df[["Age",        "Sortie␣
       ↪Positif",        "temps_psp",        "temps_pe",        "temps_fin"]].
       ↪hist(bins=50, figsize=(20,15))
      plt.tight_layout()
      plt.show()
```

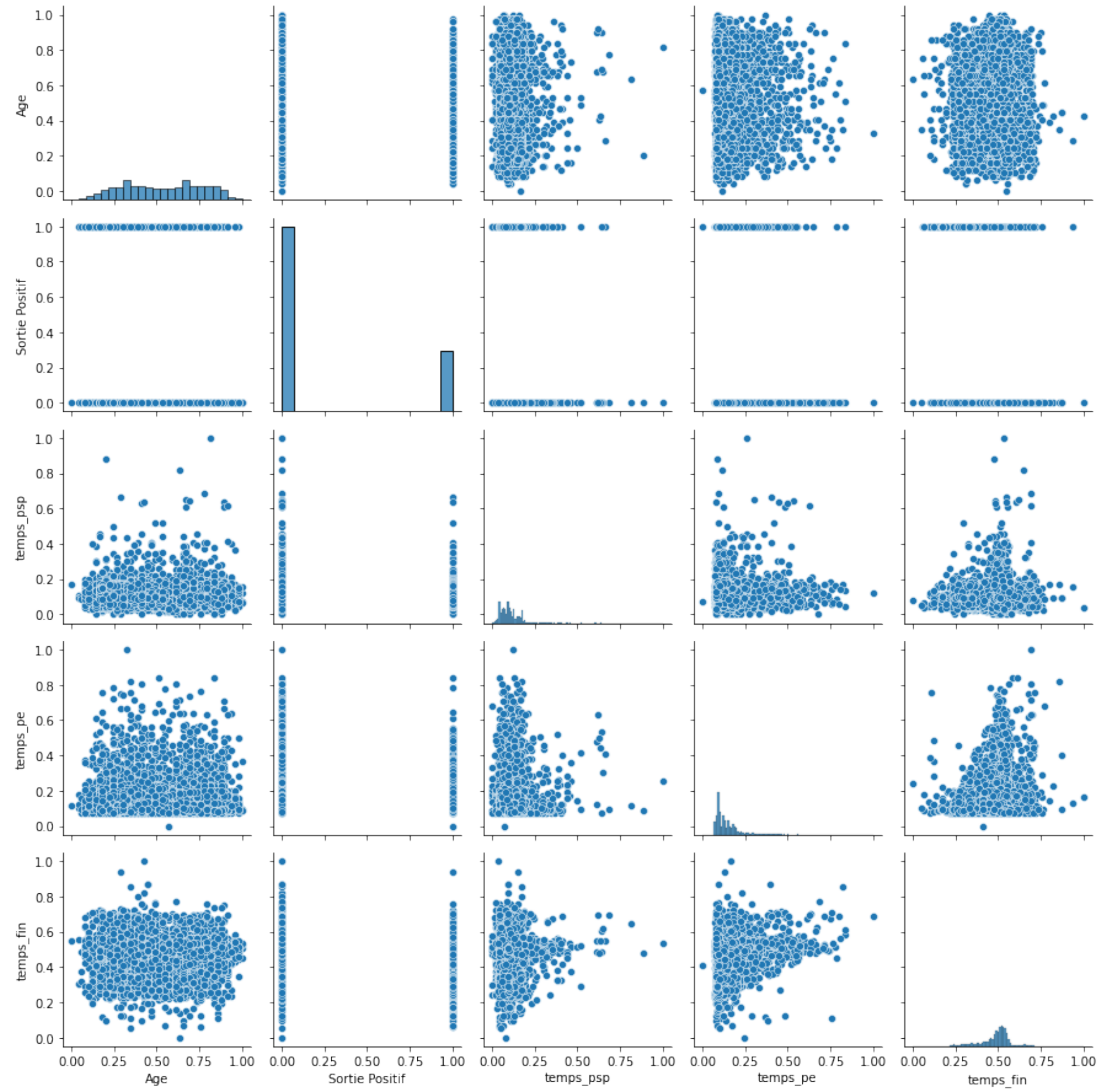


```
[29]: sns.pairplot(df)
      plt.show()
```

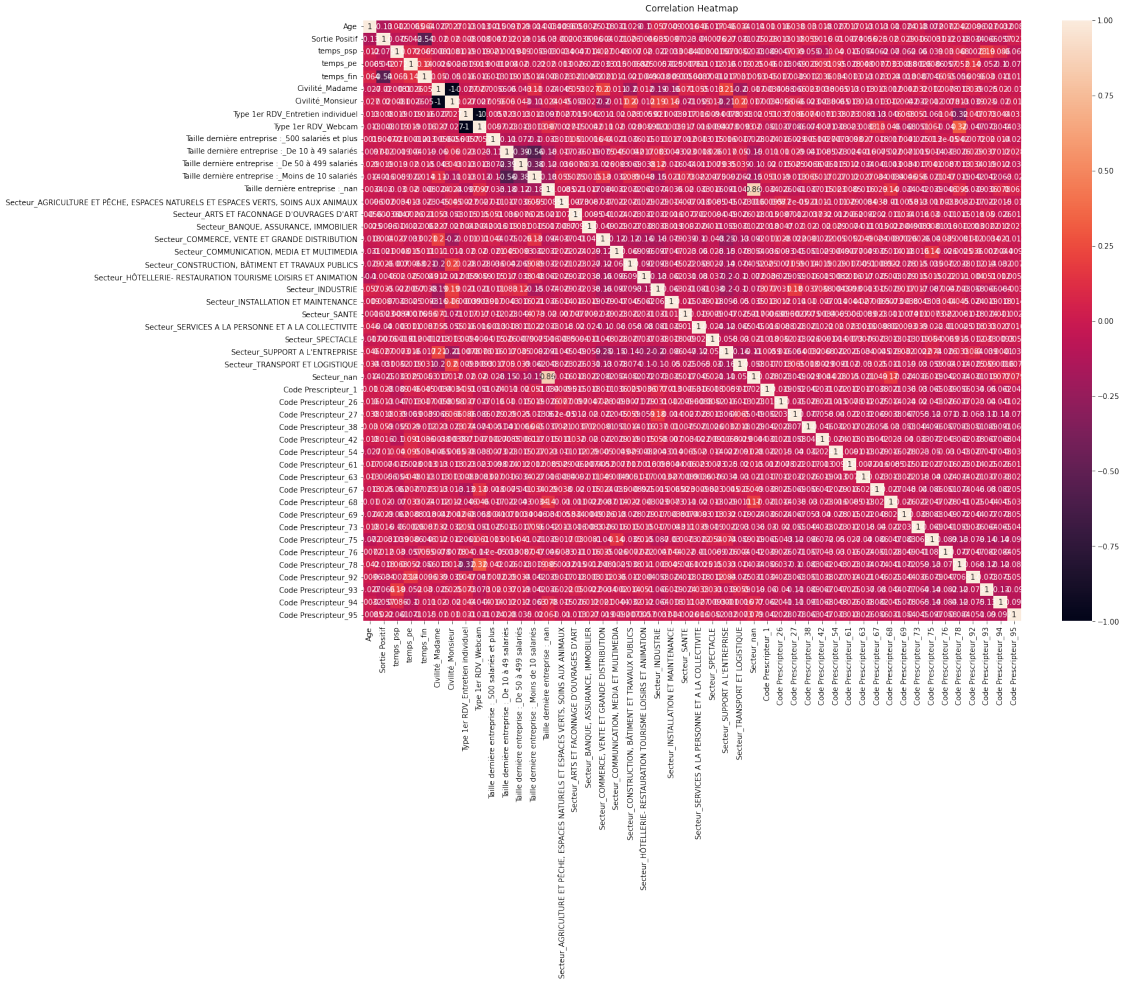


[30]:

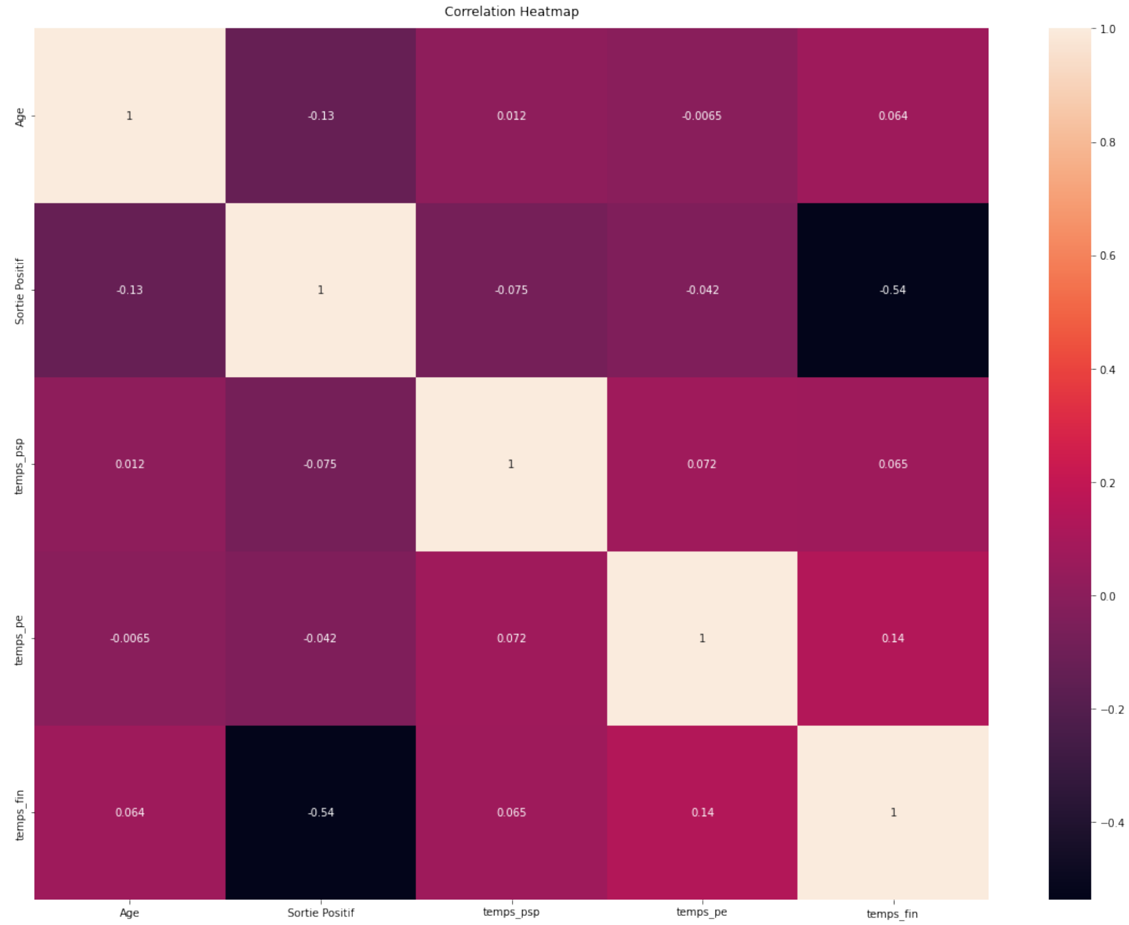
```
sns.pairplot(df[["Age",         "Sortie␣
 ↪Positif",         "temps_psp",         "temps_pe",         "temps_fin"]])
plt.tight_layout()
plt.show()
```



[31]:
```
corr = df.corr()
plt.figure(figsize=(20,15))
sns.heatmap(corr, annot=True).set_title('Correlation Heatmap',␣
 ↪fontdict={'fontsize':12}, pad=12)
plt.show()
```

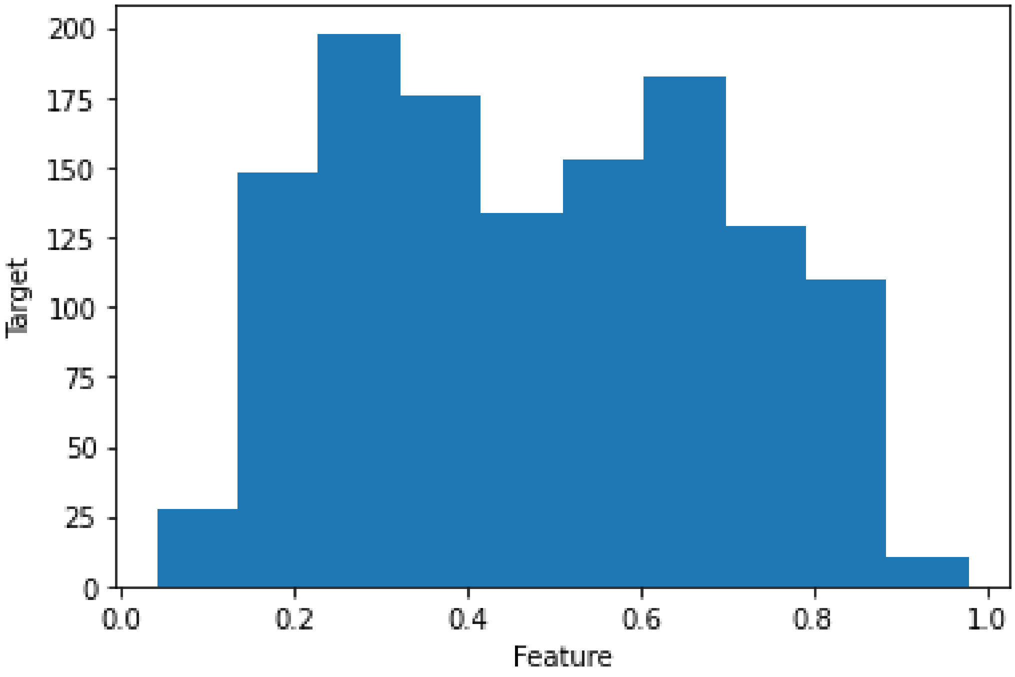


```
[32]: corr = df[["Age",          "Sortie␣
      ↪Positif",         "temps_psp",          "temps_pe",          "temps_fin"]].corr()
      plt.figure(figsize=(20,15))
      sns.heatmap(corr, annot=True).set_title('Correlation Heatmap',␣
      ↪fontdict={'fontsize':12}, pad=12)
      plt.show()
```

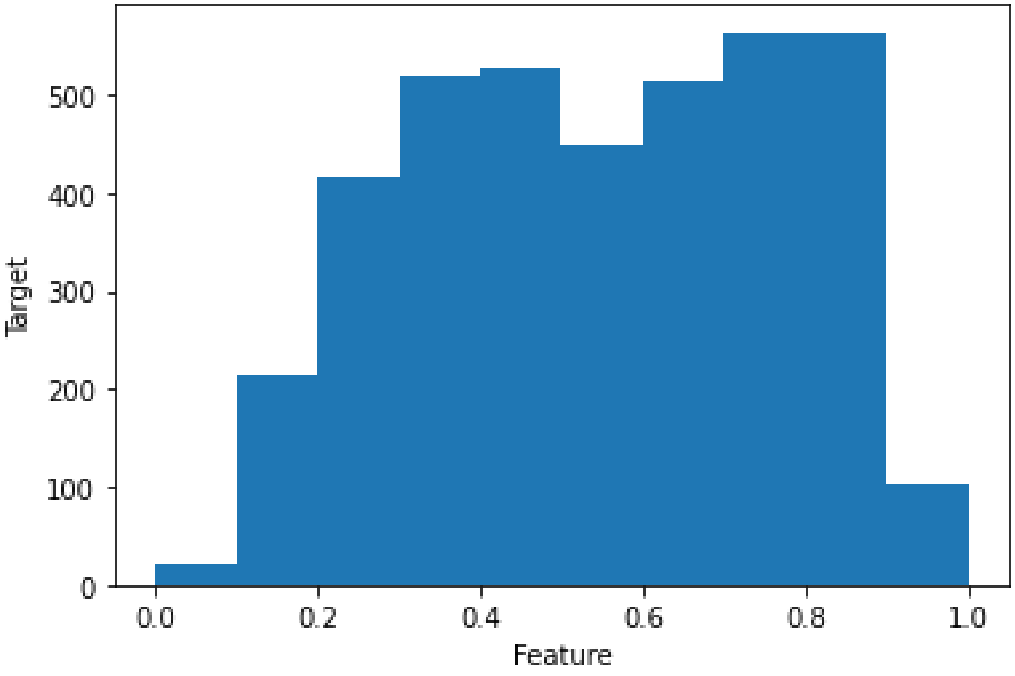


[33]:

```
plt.hist(df['Age'].where(df["Sortie Positif"]==1))
plt.xlabel("Feature")
plt.ylabel("Target")
plt.show()
```



[34]:
```
plt.hist(df['Age'].where(df["Sortie Positif"]==0))
plt.xlabel("Feature")
plt.ylabel("Target")
plt.show()
```


Target
Feature