data-preparation-and-cleaning

February 13, 2023

0.1 1- Import des librairies et de la data

[262]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import missingno as msno
```

[263]:
```
data = pd.read_excel("LIR-d.xlsx","Data")
```

[264]:
```
data.shape
```

[264]:
```
(10477, 12)
```

[265]:
```
data.head(5)
```

[265]:

	Lot	Civilité	Age	N° commande	Code Prescripteur	Date PSP	Date Point d'étape	Date démarrage	Date fin de prestation	Statut DE	Situation DE	Type 1er RDV
0	LIR AURA 14524	Madame	49	SSLIRL1757	1	2020-11-26 00:00:00	2021-04-14	2020-10-26	2021-08-12	Prestation aboutie	Sortie sans solution	Entretien individuel
1	LIR AURA 14524	Madame	29	SSLIRL0272	1	2020-08-17 00:00:00	2020-12-11	2020-08-03	2021-07-02	Prestation aboutie	Sortie sans solution	
2	LIR AURA 14524	Madame	63	SSLIRL2287	1	2020-12-04 00:00:00	2021-04-20	2020-11-19	2021-07-20	Prestation aboutie	Sortie sans solution	
3	LIR AURA 14524	Monsieur	58	SSLIRL1756	1	2020-11-16 00:00:00	2021-03-02	2020-10-26	2021-04-19	Sortie anticipée	Emploi durable	
4	LIR AURA 14524	Madame	57	SSLIRL1334	1	2020-10-22 00:00:00	2021-04-29	2020-10-05	2021-09-01	Sortie anticipée	Emploi durable	

```
1  Entretien individuel
2  Entretien individuel
3  Entretien individuel
4  Entretien individuel
```

[266]:
```
data.tail()
```

[266]:
```
                Lot  Civilité  Age N° commande  Code Prescripteur  \
10472  LIR12_14534    Madame   38  SKLIRN1243                 76
10473  LIR12_14534  Monsieur   30  SKLIRL0953                 76
10474  LIR12_14534  Monsieur   38  SKLIRP0486                 76
10475  LIR12_14534    Madame   60  SKLIRL0891                 76
10476  LIR12_14534  Monsieur   50  SKLIRL0123                 76

                  Date PSP Date Point d'étape Date démarrage  \
10472  2022-01-05 00:00:00         2022-04-06     2021-12-02
10473  2021-01-13 00:00:00         2021-04-07     2020-12-08
10474  2022-07-06 00:00:00         2022-10-26     2022-06-08
10475  2021-01-11 00:00:00         2021-04-01     2020-12-01
10476  2020-09-28 00:00:00         2020-12-10     2020-08-05

      Date fin de prestation           Statut DE          Situation DE  \
10472             2022-11-16   Prestation aboutie  Sortie sans solution
10473             2021-12-02   Prestation aboutie  Sortie sans solution
10474                    NaT  Prestation en cours                  NaN
10475             2021-11-10   Prestation aboutie  Sortie sans solution
10476             2021-02-15     Sortie anticipée       Emploi durable

               Type 1er RDV
10472  Entretien individuel
10473  Entretien individuel
10474  Entretien individuel
10475  Entretien individuel
10476  Entretien individuel
```

[267]:
```
data.describe()
```

[267]:
```
                Age  Code Prescripteur
count  10477.000000       10477.000000
mean      43.367853          73.249308
std       11.165074          24.423500
min       17.000000           1.000000
25%       34.000000          67.000000
50%       43.000000          78.000000
75%       53.000000          93.000000
max       71.000000          95.000000
```

[268]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10477 entries, 0 to 10476
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Lot                     10477 non-null  object
 1   Civilité                10477 non-null  object
 2   Age                     10477 non-null  int64
 3   N° commande             10477 non-null  object
 4   Code Prescripteur       10477 non-null  int64
 5   Date PSP                8005 non-null   object
 6   Date Point d'étape      6985 non-null   datetime64[ns]
 7   Date démarrage          10477 non-null  datetime64[ns]
 8   Date fin de prestation  8583 non-null   datetime64[ns]
 9   Statut DE               10477 non-null  object
 10  Situation DE            8587 non-null   object
 11  Type 1er RDV            10477 non-null  object
dtypes: datetime64[ns](3), int64(2), object(7)
memory usage: 982.3+ KB
```

[269]:
```
data.columns
```

[269]:
```
Index(['Lot', 'Civilité', 'Age', 'N° commande', 'Code Prescripteur',
       'Date PSP', 'Date Point d'étape', 'Date démarrage',
       'Date fin de prestation', 'Statut DE', 'Situation DE', 'Type 1er RDV'],
      dtype='object')
```

[270]:
```
data["Type 1er RDV"].unique()
```

[270]:
```
array(['Entretien individuel', 'Webcam'], dtype=object)
```

Nettoyage des données

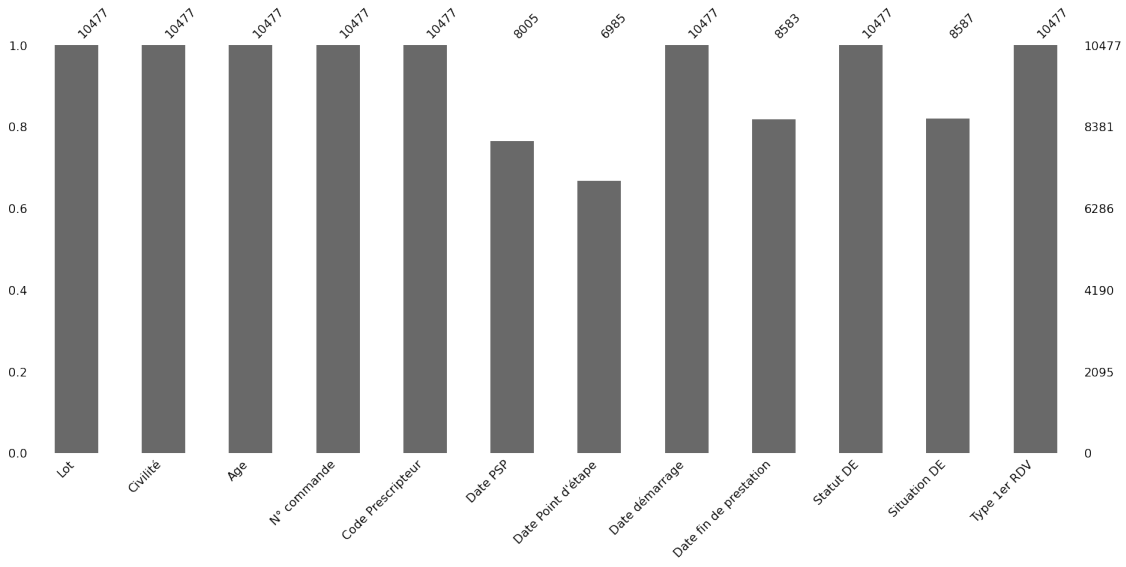
[271]:
```
data.isnull().sum()
```

[271]:
```
Lot                        0
Civilité                   0
Age                        0
N° commande                0
Code Prescripteur          0
Date PSP                2472
Date Point d'étape      3492
Date démarrage             0
Date fin de prestation  1894
Statut DE                  0
Situation DE            1890
```

```
Type 1er RDV                0
dtype: int64
```

[272]:
```
msno.bar(data)
```

[272]:
```
<AxesSubplot:>
```



Nous allons selectionner seulement les commandes qui ont finis.

[273]:
```
data = data[data["Statut DE"].isin(["Prestation aboutie","Sortie anticipée"])]
```

[274]:
```
data.shape
```

[274]:
```
(8472, 12)
```

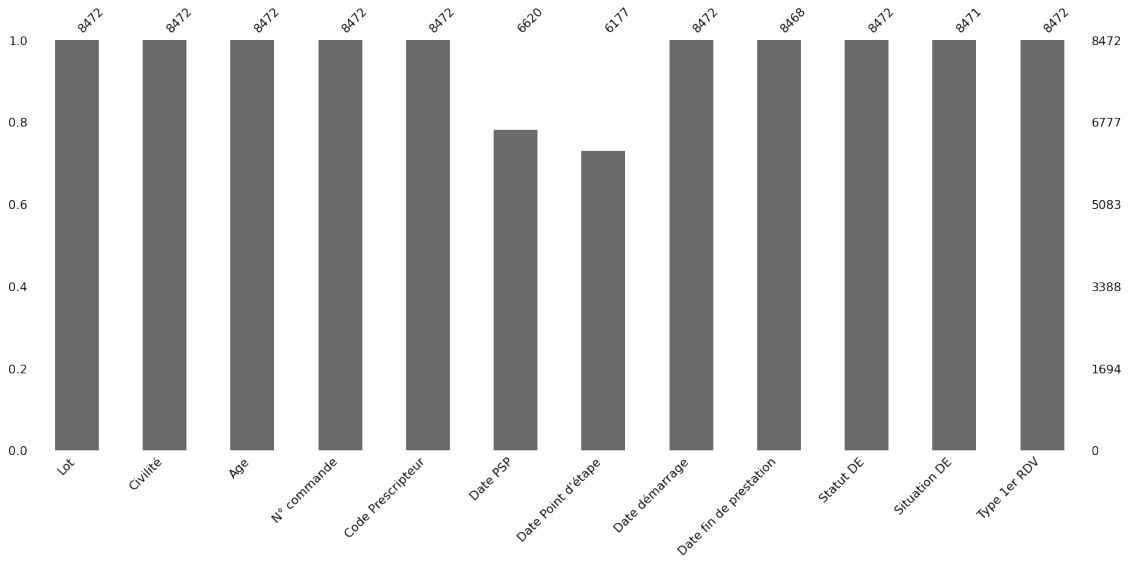
[275]:
```
data.isnull().sum()
```

[275]:
```
Lot                         0
Civilité                    0
Age                         0
N° commande                 0
Code Prescripteur           0
Date PSP                 1852
Date Point d'étape       2295
Date démarrage              0
Date fin de prestation      4
Statut DE                   0
Situation DE                1
Type 1er RDV                0
```

```
dtype: int64
```

[276]:
```
msno.bar(data)
```

[276]:
```
<AxesSubplot:>
```



Un des facteurs “Date PSP” n’as pas été transformer en datetime. Nous devons changer le type de données pour “Date PSP”.

[277]:
```
data['Date PSP'] = pd.to_datetime(data['Date PSP'], errors='coerce')
```

[278]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8472 entries, 0 to 10476
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Lot                     8472 non-null   object
 1   Civilité                8472 non-null   object
 2   Age                     8472 non-null   int64
 3   N° commande             8472 non-null   object
 4   Code Prescripteur       8472 non-null   int64
 5   Date PSP                6617 non-null   datetime64[ns]
 6   Date Point d'étape      6177 non-null   datetime64[ns]
 7   Date démarrage          8472 non-null   datetime64[ns]
 8   Date fin de prestation  8468 non-null   datetime64[ns]
 9   Statut DE               8472 non-null   object
 10  Situation DE            8471 non-null   object
```

```
 11  Type 1er RDV            8472 non-null   object
dtypes: datetime64[ns](4), int64(2), object(6)
memory usage: 860.4+ KB
```

0.2 Création des indicateurs des sorties positifs

[279]:
```
conditions = [
    data['Statut DE'].eq('Sortie anticipée') & data['Situation DE'].
 ↪eq('Création / Reprise d´entreprise')|data['Statut DE'].eq('Sortie␣
 ↪anticipée') & data['Situation DE'].eq('Création / Reprise␣
 ↪d'entreprise')|data['Statut DE'].eq('Sortie anticipée') & data['Situation␣
 ↪DE'].eq('Emploi durable')
]
choice = [1]
data["Sortie Positif"] = np.select(conditions, choice, default=0)
data.head(5)
```

[279]:

	Lot	Civilité	Age	N° commande	Code Prescripteur	Date PSP	Date Point d'étape	Date démarrage	Date fin de prestation	Statut DE	Situation DE	Type 1er RDV	Sortie Positif
0	LIR AURA 14524	Madame	49	SSLIRL1757	1	2020-11-26	2021-04-14	2020-10-26	2021-08-12	Prestation aboutie	Sortie sans solution	Entretien individuel	0
1	LIR AURA 14524	Madame	29	SSLIRL0272	1	2020-08-17	2020-12-11	2020-08-03	2021-07-02	Prestation aboutie	Sortie sans solution	Entretien individuel	0
2	LIR AURA 14524	Madame	63	SSLIRL2287	1	2020-12-04	2021-04-20	2020-11-19	2021-07-20	Prestation aboutie	Sortie sans solution	Entretien individuel	0
3	LIR AURA 14524	Monsieur	58	SSLIRL1756	1	2020-11-16	2021-03-02	2020-10-26	2021-04-19	Sortie anticipée	Emploi durable	Entretien individuel	1
4	LIR AURA 14524	Madame	57	SSLIRL1334	1	2020-10-22	2021-04-29	2020-10-05	2021-09-01	Sortie anticipée	Emploi durable	Entretien individuel	1

Création des intervalles entre les différentes étapes

[280]:
```
data['temps_psp'] = data['Date PSP'] - data['Date démarrage']
data['temps_pe'] = data["Date Point d'étape"] - data['Date démarrage']
data['temps_fin'] = data["Date fin de prestation"] - data['Date démarrage']
```

[281]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8472 entries, 0 to 10476
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Lot                     8472 non-null   object
 1   Civilité                8472 non-null   object
 2   Age                     8472 non-null   int64
 3   N° commande             8472 non-null   object
 4   Code Prescripteur       8472 non-null   int64
 5   Date PSP                6617 non-null   datetime64[ns]
 6   Date Point d'étape      6177 non-null   datetime64[ns]
 7   Date démarrage          8472 non-null   datetime64[ns]
 8   Date fin de prestation  8468 non-null   datetime64[ns]
 9   Statut DE               8472 non-null   object
 10  Situation DE            8471 non-null   object
 11  Type 1er RDV            8472 non-null   object
 12  Sortie Positif          8472 non-null   int64
 13  temps_psp               6617 non-null   timedelta64[ns]
 14  temps_pe                6177 non-null   timedelta64[ns]
 15  temps_fin               8468 non-null   timedelta64[ns]
dtypes: datetime64[ns](4), int64(3), object(6), timedelta64[ns](3)
memory usage: 1.1+ MB
```

Convertion de secondes à jours:

[282]:
```
data['temps_psp'] = data['temps_psp'].dt.total_seconds()
data['temps_pe'] = data['temps_pe'].dt.total_seconds()
data['temps_fin'] = data['temps_fin'].dt.total_seconds()
```

[283]:
```
data['temps_psp'] = data['temps_psp'] / 86400
data['temps_pe'] = data['temps_pe'] / 86400
data['temps_fin'] = data['temps_fin'] / 86400
```

[284]:
```
data.head()
```

[284]:

	Lot	Civilité	Age	N° commande	Code Prescripteur	Date PSP	\
0	LIR AURA 14524	Madame	49	SSLIRL1757	1	2020-11-26	
1	LIR AURA 14524	Madame	29	SSLIRL0272	1	2020-08-17	
2	LIR AURA 14524	Madame	63	SSLIRL2287	1	2020-12-04	
3	LIR AURA 14524	Monsieur	58	SSLIRL1756	1	2020-11-16	
4	LIR AURA 14524	Madame	57	SSLIRL1334	1	2020-10-22	

```
  Date Point d'étape Date démarrage Date fin de prestation  \
0         2021-04-14     2020-10-26             2021-08-12
1         2020-12-11     2020-08-03             2021-07-02
2         2021-04-20     2020-11-19             2021-07-20
3         2021-03-02     2020-10-26             2021-04-19
4         2021-04-29     2020-10-05             2021-09-01

            Statut DE          Situation DE          Type 1er RDV  \
0  Prestation aboutie  Sortie sans solution  Entretien individuel
1  Prestation aboutie  Sortie sans solution  Entretien individuel
2  Prestation aboutie  Sortie sans solution  Entretien individuel
3    Sortie anticipée        Emploi durable  Entretien individuel
4    Sortie anticipée        Emploi durable  Entretien individuel

   Sortie Positif  temps_psp  temps_pe  temps_fin
0               0       31.0     170.0      290.0
1               0       14.0     130.0      333.0
2               0       15.0     152.0      243.0
3               1       21.0     127.0      175.0
4               1       17.0     206.0      331.0
```

Ajouter des colonnes: CODE ROME

[285]:
```
rome = pd.read_excel("Code ROME - Sortie Positive.xlsx","Data")
```

[286]:
```
rome = rome[["Numéro commande", "Taille dernière entreprise :", "Secteur"]]
```

[287]:
```
rome.shape
```

[287]:
```
(8829, 3)
```

[288]:
```
rome.head()
```

[288]:
```
  Numéro commande Taille dernière entreprise :  \
0      SVLIRL0070          Moins de 10 salariés
1      SVLIRL0056          De 50 à 499 salariés
2      SVLIRL0039           De 10 à 49 salariés
3      SVLIRL0830           De 10 à 49 salariés
4      SVLIRL0831          500 salariés et plus

                                          Secteur
0  HÔTELLERIE- RESTAURATION TOURISME LOISIRS ET A…
1          CONSTRUCTION, BÂTIMENT ET TRAVAUX PUBLICS
2             COMMERCE, VENTE ET GRANDE DISTRIBUTION
3                 COMMUNICATION, MEDIA ET MULTIMEDIA
4                           SUPPORT A L'ENTREPRISE
```

[289]:
```
rome["Taille dernière entreprise :"].unique()
```

[289]:
```
array(['Moins de 10 salariés', 'De 50 à 499 salariés',
       'De 10 à 49 salariés', '500 salariés et plus', nan], dtype=object)
```

[290]:
```
rome["Secteur"].unique()
```

[290]:
```
array(['HÔTELLERIE- RESTAURATION TOURISME LOISIRS ET ANIMATION',
       'CONSTRUCTION, BÂTIMENT ET TRAVAUX PUBLICS',
       'COMMERCE, VENTE ET GRANDE DISTRIBUTION',
       'COMMUNICATION, MEDIA ET MULTIMEDIA', "SUPPORT A L'ENTREPRISE",
       'BANQUE, ASSURANCE, IMMOBILIER', 'TRANSPORT ET LOGISTIQUE',
       'INDUSTRIE', 'SANTE',
       'SERVICES A LA PERSONNE ET A LA COLLECTIVITE',
       'INSTALLATION ET MAINTENANCE', 'SPECTACLE',
       "ARTS ET FACONNAGE D'OUVRAGES D'ART",
       'AGRICULTURE ET PÊCHE, ESPACES NATURELS ET ESPACES VERTS, SOINS AUX
ANIMAUX',
       nan], dtype=object)
```

[291]:
```
rome.columns
```

[291]:
```
Index(['Numéro commande', 'Taille dernière entreprise :', 'Secteur'],
dtype='object')
```

[292]:
```
data_prep = data.merge(rome, how='left', left_on="N° commande",␣
↪right_on='Numéro commande')
```

[293]:
```
data_prep.head()
```

[293]:

	Lot	Civilité	Age	N° commande	Code Prescripteur	Date PSP	Date Point d'étape	Date démarrage	Date fin de prestation	Statut DE	Situation DE	Type 1er RDV
0	LIR AURA 14524	Madame	49	SSLIRL1757	1	2020-11-26	2021-04-14	2020-10-26	2021-08-12	Prestation aboutie	Sortie sans solution	Entretien individuel
1	LIR AURA 14524	Madame	29	SSLIRL0272	1	2020-08-17	2020-12-11	2020-08-03	2021-07-02	Prestation aboutie	Sortie sans solution	Entretien individuel
2	LIR AURA 14524	Madame	63	SSLIRL2287	1	2020-12-04	2021-04-20	2020-11-19	2021-07-20			
3	LIR AURA 14524	Monsieur	58	SSLIRL1756	1	2020-11-16	2021-03-02	2020-10-26	2021-04-19			
4	LIR AURA 14524	Madame	57	SSLIRL1334	1	2020-10-22	2021-04-29	2020-10-05	2021-09-01			

```
2  Prestation aboutie  Sortie sans solution  Entretien individuel
3    Sortie anticipée        Emploi durable  Entretien individuel
4    Sortie anticipée        Emploi durable  Entretien individuel

   Sortie Positif  temps_psp  temps_pe  temps_fin Numéro commande  \
0               0       31.0     170.0      290.0      SSLIRL1757
1               0       14.0     130.0      333.0      SSLIRL0272
2               0       15.0     152.0      243.0      SSLIRL2287
3               1       21.0     127.0      175.0      SSLIRL1756
4               1       17.0     206.0      331.0      SSLIRL1334

  Taille dernière entreprise :  \
0          De 10 à 49 salariés
1          De 10 à 49 salariés
2          De 10 à 49 salariés
3          De 10 à 49 salariés
4         Moins de 10 salariés

                                          Secteur
0                                       INDUSTRIE
1                             SUPPORT A L'ENTREPRISE
2       SERVICES A LA PERSONNE ET A LA COLLECTIVITE
3                                       INDUSTRIE
4  HÔTELLERIE- RESTAURATION TOURISME LOISIRS ET A…
```

[294]:
```
data_prep = data_prep.drop(columns=["N° commande", "Numéro commande", "Lot",␣
↪"Date PSP", "Date démarrage", "Date Point d'étape", "Date fin de␣
↪prestation", "Statut DE", "Situation DE"])
```

[295]:
```
data_prep.shape
```

[295]:
```
(8472, 10)
```

[296]:
```
data_prep.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 8472 entries, 0 to 8471
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Civilité                    8472 non-null   object
 1   Age                         8472 non-null   int64
 2   Code Prescripteur           8472 non-null   int64
 3   Type 1er RDV                8472 non-null   object
 4   Sortie Positif              8472 non-null   int64
 5   temps_psp                   6617 non-null   float64
 6   temps_pe                    6177 non-null   float64
```

```
 7   temps_fin                       8468 non-null   float64
 8   Taille dernière entreprise :    7743 non-null   object
 9   Secteur                         7883 non-null   object
dtypes: float64(3), int64(3), object(4)
memory usage: 728.1+ KB
```

[297]:
```
data_prep.describe()
```

[297]:
```
                Age  Code Prescripteur  Sortie Positif    temps_psp  \
count  8472.000000        8472.000000     8472.000000  6617.000000
mean     43.539070          74.164188        0.344075    33.736285
std      11.078034          23.970397        0.475094    20.887738
min      17.000000           1.000000        0.000000     0.000000
25%      34.000000          68.000000        0.000000    21.000000
50%      43.000000          78.000000        0.000000    30.000000
75%      53.000000          93.000000        1.000000    43.000000
max      68.000000          95.000000        1.000000   325.000000

          temps_pe    temps_fin
count  6177.000000  8468.000000
mean    141.523393   246.076996
std      27.991674   102.409867
min      99.000000   -22.000000
25%     124.000000   175.000000
50%     133.000000   288.000000
75%     148.000000   318.000000
max     388.000000   627.000000
```

[298]:
```
data_prep.head()
```

[298]:
```
   Civilité  Age  Code Prescripteur          Type 1er RDV  Sortie Positif  \
0    Madame   49                 1  Entretien individuel               0
1    Madame   29                 1  Entretien individuel               0
2    Madame   63                 1  Entretien individuel               0
3  Monsieur   58                 1  Entretien individuel               1
4    Madame   57                 1  Entretien individuel               1

   temps_psp  temps_pe  temps_fin Taille dernière entreprise :  \
0       31.0     170.0      290.0          De 10 à 49 salariés
1       14.0     130.0      333.0          De 10 à 49 salariés
2       15.0     152.0      243.0          De 10 à 49 salariés
3       21.0     127.0      175.0          De 10 à 49 salariés
4       17.0     206.0      331.0         Moins de 10 salariés

                                    Secteur
0                                 INDUSTRIE
1                    SUPPORT A L'ENTREPRISE
```

```
2        SERVICES A LA PERSONNE ET A LA COLLECTIVITE
3                                              INDUSTRIE
4  HÔTELLERIE- RESTAURATION TOURISME LOISIRS ET A…
```

Transformation des valeurs catégoriques.

[299]:
```
from sklearn.preprocessing import OneHotEncoder
```

[300]:
```
categorical_val = ['Civilité', "Type 1er RDV", "Taille dernière entreprise :",␣
 ↪"Secteur", "Code Prescripteur"]
categorical_data = data_prep[categorical_val]
```

[301]:
```
encoder = OneHotEncoder(handle_unknown='ignore')
encoded_data = encoder.fit_transform(categorical_data).toarray()
```

[302]:
```
encoded_data = pd.DataFrame(encoded_data, columns=encoder.
 ↪get_feature_names(categorical_val))
```

```
/home/n_fuentes/.local/lib/python3.8/site-
packages/sklearn/utils/deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[303]:
```
data_prep = pd.concat([data_prep, encoded_data], axis=1)
```

[304]:
```
data_prep = data_prep.drop(categorical_val, axis=1)
```

[305]:
```
data_prep.shape
```

[305]:
```
(8472, 48)
```

[318]:
```
data_prep.columns
```

[318]:
```
Index(['Age', 'Sortie Positif', 'temps_psp', 'temps_pe', 'temps_fin',
       'Civilité_Madame', 'Civilité_Monsieur',
       'Type 1er RDV_Entretien individuel', 'Type 1er RDV_Webcam',
       'Taille dernière entreprise :_500 salariés et plus',
       'Taille dernière entreprise :_De 10 à 49 salariés',
       'Taille dernière entreprise :_De 50 à 499 salariés',
       'Taille dernière entreprise :_Moins de 10 salariés',
       'Taille dernière entreprise :_nan',
       'Secteur_AGRICULTURE ET PÊCHE, ESPACES NATURELS ET ESPACES VERTS, SOINS
AUX ANIMAUX',
       'Secteur_ARTS ET FACONNAGE D'OUVRAGES D'ART',
       'Secteur_BANQUE, ASSURANCE, IMMOBILIER',
       'Secteur_COMMERCE, VENTE ET GRANDE DISTRIBUTION',
```

```
       'Secteur_COMMUNICATION, MEDIA ET MULTIMEDIA',
       'Secteur_CONSTRUCTION, BÂTIMENT ET TRAVAUX PUBLICS',
       'Secteur_HÔTELLERIE- RESTAURATION TOURISME LOISIRS ET ANIMATION',
       'Secteur_INDUSTRIE', 'Secteur_INSTALLATION ET MAINTENANCE',
       'Secteur_SANTE', 'Secteur_SERVICES A LA PERSONNE ET A LA COLLECTIVITE',
       'Secteur_SPECTACLE', 'Secteur_SUPPORT A L'ENTREPRISE',
       'Secteur_TRANSPORT ET LOGISTIQUE', 'Secteur_nan', 'Code Prescripteur_1',
       'Code Prescripteur_26', 'Code Prescripteur_27', 'Code Prescripteur_38',
       'Code Prescripteur_42', 'Code Prescripteur_54', 'Code Prescripteur_61',
       'Code Prescripteur_63', 'Code Prescripteur_67', 'Code Prescripteur_68',
       'Code Prescripteur_69', 'Code Prescripteur_73', 'Code Prescripteur_75',
       'Code Prescripteur_76', 'Code Prescripteur_78', 'Code Prescripteur_92',
       'Code Prescripteur_93', 'Code Prescripteur_94', 'Code Prescripteur_95'],
      dtype='object')
```

[306]:
```
data_prep.head()
```

[306]:

	Age	Sortie Positif	temps_psp	temps_pe	temps_fin	Civilité_Madame	\
0	49	0	31.0	170.0	290.0	1.0	
1	29	0	14.0	130.0	333.0	1.0	
2	63	0	15.0	152.0	243.0	1.0	
3	58	1	21.0	127.0	175.0	0.0	
4	57	1	17.0	206.0	331.0	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]
```

Nous pouvons observer les min et max pour voir s'il y a des valeurs abérrantes.

[307]:
```
data_prep.describe()
```

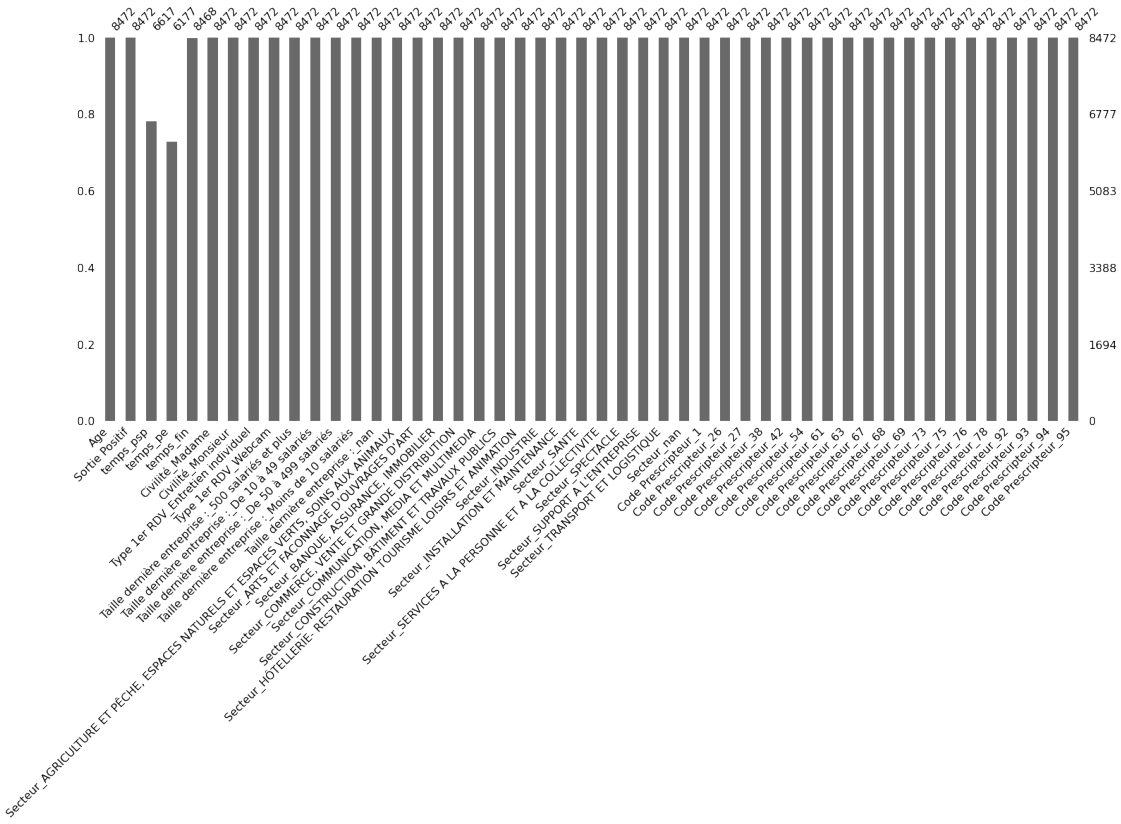
[307]:
```
                Age  Sortie Positif    temps_psp     temps_pe    temps_fin  \
count  8472.000000     8472.000000  6617.000000  6177.000000  8468.000000
mean     43.539070        0.344075    33.736285   141.523393   246.076996
std      11.078034        0.475094    20.887738    27.991674   102.409867
min      17.000000        0.000000     0.000000    99.000000   -22.000000
25%      34.000000        0.000000    21.000000   124.000000   175.000000
50%      43.000000        0.000000    30.000000   133.000000   288.000000
75%      53.000000        1.000000    43.000000   148.000000   318.000000
max      68.000000        1.000000   325.000000   388.000000   627.000000

       Civilité_Madame  Civilité_Monsieur  Type 1er RDV_Entretien individuel  \
count      8472.000000        8472.000000                         8472.000000
mean          0.485836           0.514164                            0.914660
std           0.499829           0.499829                            0.279403
min           0.000000           0.000000                            0.000000
25%           0.000000           0.000000                            1.000000
50%           0.000000           1.000000                            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          8472.000000                                        8472.000000
```

```
mean              0.085340                                              0.019358
std               0.279403                                              0.137788
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  …           8472.000000           8472.000000           8472.000000
mean   …              0.012040              0.035175              0.026086
std    …              0.109069              0.184232              0.159400
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           8472.00000          8472.000000          8472.000000
mean               0.10517             0.040486             0.095491
std                0.30679             0.197108             0.293909
min                0.00000             0.000000             0.000000
25%                0.00000             0.000000             0.000000
50%                0.00000             0.000000             0.000000
75%                0.00000             0.000000             0.000000
max                1.00000             1.000000             1.000000

       Code Prescripteur_92  Code Prescripteur_93  Code Prescripteur_94  \
count          8472.000000          8472.000000          8472.000000
mean              0.044381             0.190156             0.156870
std               0.205953             0.392447             0.363699
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          8472.000000
mean              0.051818
std               0.221672
min               0.000000
25%               0.000000
50%               0.000000
75%               0.000000
max               1.000000
```

[8 rows x 48 columns]

[308]:
```
msno.bar(data_prep)
```

[308]: <AxesSubplot:>



[309]:
```
nan_mask = pd.isnull(data_prep)
nan_count = nan_mask.sum()
print(nan_count)
```

```
Age
0
Sortie Positif
0
temps_psp
1855
temps_pe
2295
temps_fin
4
```

```
Civilité_Madame
0
Civilité_Monsieur
0
Type 1er RDV_Entretien individuel
0
Type 1er RDV_Webcam
0
Taille dernière entreprise :_500 salariés et plus
0
Taille dernière entreprise :_De 10 à 49 salariés
0
Taille dernière entreprise :_De 50 à 499 salariés
0
Taille dernière entreprise :_Moins de 10 salariés
0
Taille dernière entreprise :_nan
0
Secteur_AGRICULTURE ET PÊCHE, ESPACES NATURELS ET ESPACES VERTS, SOINS AUX
ANIMAUX       0
Secteur_ARTS ET FACONNAGE D'OUVRAGES D'ART
0
Secteur_BANQUE, ASSURANCE, IMMOBILIER
0
Secteur_COMMERCE, VENTE ET GRANDE DISTRIBUTION
0
Secteur_COMMUNICATION, MEDIA ET MULTIMEDIA
0
Secteur_CONSTRUCTION, BÂTIMENT ET TRAVAUX PUBLICS
0
Secteur_HÔTELLERIE- RESTAURATION TOURISME LOISIRS ET ANIMATION
0
Secteur_INDUSTRIE
0
Secteur_INSTALLATION ET MAINTENANCE
0
Secteur_SANTE
0
Secteur_SERVICES A LA PERSONNE ET A LA COLLECTIVITE
0
Secteur_SPECTACLE
0
Secteur_SUPPORT A L'ENTREPRISE
0
Secteur_TRANSPORT ET LOGISTIQUE
0
Secteur_nan
0
```

```
Code Prescripteur_1
0
Code Prescripteur_26
0
Code Prescripteur_27
0
Code Prescripteur_38
0
Code Prescripteur_42
0
Code Prescripteur_54
0
Code Prescripteur_61
0
Code Prescripteur_63
0
Code Prescripteur_67
0
Code Prescripteur_68
0
Code Prescripteur_69
0
Code Prescripteur_73
0
Code Prescripteur_75
0
Code Prescripteur_76
0
Code Prescripteur_78
0
Code Prescripteur_92
0
Code Prescripteur_93
0
Code Prescripteur_94
0
Code Prescripteur_95
0
dtype: int64
```

[310]:
```
data_prep = data_prep.dropna()
```

[311]:
```
data_prep.shape
```

[311]: (5162, 48)

[312]:
```
nan_mask = pd.isnull(data_prep)
nan_count = nan_mask.sum()
print(nan_count)
```

```
Age
0
Sortie Positif
0
temps_psp
0
temps_pe
0
temps_fin
0
Civilité_Madame
0
Civilité_Monsieur
0
Type 1er RDV_Entretien individuel
0
Type 1er RDV_Webcam
0
Taille dernière entreprise :_500 salariés et plus
0
Taille dernière entreprise :_De 10 à 49 salariés
0
Taille dernière entreprise :_De 50 à 499 salariés
0
Taille dernière entreprise :_Moins de 10 salariés
0
Taille dernière entreprise :_nan
0
Secteur_AGRICULTURE ET PÊCHE, ESPACES NATURELS ET ESPACES VERTS, SOINS AUX
ANIMAUX    0
Secteur_ARTS ET FACONNAGE D'OUVRAGES D'ART
0
Secteur_BANQUE, ASSURANCE, IMMOBILIER
0
Secteur_COMMERCE, VENTE ET GRANDE DISTRIBUTION
0
Secteur_COMMUNICATION, MEDIA ET MULTIMEDIA
0
Secteur_CONSTRUCTION, BÂTIMENT ET TRAVAUX PUBLICS
0
Secteur_HÔTELLERIE- RESTAURATION TOURISME LOISIRS ET ANIMATION
0
Secteur_INDUSTRIE
```

0
Secteur_INSTALLATION ET MAINTENANCE
0
Secteur_SANTE
0
Secteur_SERVICES A LA PERSONNE ET A LA COLLECTIVITE
0
Secteur_SPECTACLE
0
Secteur_SUPPORT A L'ENTREPRISE
0
Secteur_TRANSPORT ET LOGISTIQUE
0
Secteur_nan
0
Code Prescripteur_1
0
Code Prescripteur_26
0
Code Prescripteur_27
0
Code Prescripteur_38
0
Code Prescripteur_42
0
Code Prescripteur_54
0
Code Prescripteur_61
0
Code Prescripteur_63
0
Code Prescripteur_67
0
Code Prescripteur_68
0
Code Prescripteur_69
0
Code Prescripteur_73
0
Code Prescripteur_75
0
Code Prescripteur_76
0
Code Prescripteur_78
0
Code Prescripteur_92
0
Code Prescripteur_93

```
0
Code Prescripteur_94
0
Code Prescripteur_95
0
dtype: int64
```

[313]:
```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
cols_to_norm = ["Age", "temps_psp","temps_pe", "temps_fin"]
data_prep[cols_to_norm] = scaler.fit_transform(data_prep[cols_to_norm])
```

[314]:
```
data_prep.head()
```

[314]:

	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]
```

[315]:
```
data_prep.describe()
```

[315]:
```
                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]
```

[260]:
```
data_prep.to_csv("data_prep.csv")
```