

# TP1 - Exploration et transformation des données

Alain Nyeck - Folly Tata Ayeboua

```
In [1]: import pandas as pd
import numpy as np
import seaborn as sns
from scipy.stats import chi2_contingency
import matplotlib.pyplot as plt
import time
```

## Étape 1 : On considère le fichier train\_users\_2.csv

- Indiquer les points marquants l'exploration.
- Pour chaque observation, indiquer l'opération à effectuer qui serait la plus appropriée.

```
In [2]: df = pd.read_csv('train_users_2.csv', index_col=0)
```

```
In [3]: print('\nAffichage du dataset\n')
display(df.head(10))
```

Affichage du dataset

| id         | date_account_created | timestamp_first_active | date_first_booking | gender   | age  | signup_method | signup_flow | language |
|------------|----------------------|------------------------|--------------------|----------|------|---------------|-------------|----------|
| gxn3p5htnn | 2010-06-28           | 20090319043255         | NaN                | unknown- | NaN  | facebook      | 0           |          |
| 820tgsjxq7 | 2011-05-25           | 20090523174809         | NaN                | MALE     | 38.0 | facebook      | 0           |          |
| 4ft3gnwmtx | 2010-09-28           | 20090609231247         | 2010-08-02         | FEMALE   | 56.0 | basic         | 3           |          |
| bjlt8pjhuk | 2011-12-05           | 20091031060129         | 2012-09-08         | FEMALE   | 42.0 | facebook      | 0           |          |
| 87mebub9p4 | 2010-09-14           | 20091208061105         | 2010-02-18         | unknown- | 41.0 | basic         | 0           |          |
| osr2jwljor | 2010-01-01           | 20100101215619         | 2010-01-02         | unknown- | NaN  | basic         | 0           |          |
| lsw9q7uk0j | 2010-01-02           | 20100102012558         | 2010-01-05         | FEMALE   | 46.0 | basic         | 0           |          |
| 0d01nltbrs | 2010-01-03           | 20100103191905         | 2010-01-13         | FEMALE   | 47.0 | basic         | 0           |          |
| a1vcnhxeij | 2010-01-04           | 20100104004211         | 2010-07-29         | FEMALE   | 50.0 | basic         | 0           |          |
| 6uh8zyj2gn | 2010-01-04           | 20100104023758         | 2010-01-04         | unknown- | 46.0 | basic         | 0           |          |

## 1.1. Quels sont les descripteurs (colonnes) du dataset?

```
In [4]: print("Les descripteurs du dataset:")
print(df.columns.tolist())
```

Les descripteurs du dataset:

```
['date_account_created', 'timestamp_first_active', 'date_first_booking', 'gender', 'age', 'signup_method', 'signup_flow', 'language', 'affiliate_channel', 'affiliate_provider', 'first_affiliate_tracked', 'signup_app', 'first_device_type', 'first_browser', 'country_destination']
```

## 1.2. Combien d'enregistrements (lignes) ont été fournis ?

```
In [5]: nombre_enregistrements = df.shape[0]
print("Le nombre d'enregistrements:", nombre_enregistrements)
```

Le nombre d'enregistrements: 213451

### 1.3. Quel est le format des données. Par exemple, dans quel format les dates sont fournies, existe-t-il des valeurs numériques, à quoi ressemblent les différentes valeurs catégorielles ?

In [6]: `print(df.info())`

```
<class 'pandas.core.frame.DataFrame'>
Index: 213451 entries, gxn3p5htnn to nw9fwlyb5f
Data columns (total 15 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   date_account_created                 213451 non-null object
1   timestamp_first_active               213451 non-null int64
2   date_first_booking                  88908 non-null  object
3   gender                              213451 non-null object
4   age                                  125461 non-null float64
5   signup_method                       213451 non-null object
6   signup_flow                         213451 non-null int64
7   language                            213451 non-null object
8   affiliate_channel                   213451 non-null object
9   affiliate_provider                  213451 non-null object
10  first_affiliate_tracked              207386 non-null object
11  signup_app                           213451 non-null object
12  first_device_type                   213451 non-null object
13  first_browser                       213451 non-null object
14  country_destination                 213451 non-null object
dtypes: float64(1), int64(2), object(12)
memory usage: 26.1+ MB
None
```

Les données de types date: 'date\_account\_created', 'timestamp\_first\_active', 'date\_first\_booking'</br> Les données numériques: 'age', 'signup\_flow'</br> Les données catégorielles: 'gender', 'signup\_method', 'language', 'affiliate\_channel', 'affiliate\_provider', 'first\_affiliate\_tracked', 'signup\_app', 'first\_device\_type', 'first\_browser', 'country\_destination'</br></br>

Les dates sont de types objet et int. elles seront converties en type datetime (format ci-dessous) pour en extraire proprement les champs</br> 'date\_account\_created' utilise le format 'YYYY-MM-DD'</br> 'timestamp\_first\_active' utilise le format 'YYYYMMDDhhmmss'</br> 'date\_first\_booking' utilise le format 'YYYY-MM-DD'</br></br>

```
In [7]: cols = ['gender', 'signup_method', 'signup_flow', 'language', 'affiliate_channel', 'affiliate_provider', 'fir

print('Ci-dessous les valeurs catégorielles:\n')
for col in cols:
```

```
print(col,':', df[col].unique(), '\n')
```

Ci-dessous les valeurs catégorielles:

gender : ['-unknown-' 'MALE' 'FEMALE' 'OTHER']

signup\_method : ['facebook' 'basic' 'google']

signup\_flow : [ 0 3 2 1 24 8 6 5 10 25 12 4 16 15 20 21 23]

language : ['en' 'fr' 'de' 'es' 'it' 'pt' 'zh' 'ko' 'ja' 'ru' 'pl' 'el' 'sv' 'nl'  
'hu' 'da' 'id' 'fi' 'no' 'tr' 'th' 'cs' 'hr' 'ca' 'is']

affiliate\_channel : ['direct' 'seo' 'other' 'sem-non-brand' 'content' 'sem-brand'  
'remarketing' 'api']

affiliate\_provider : ['direct' 'google' 'other' 'craigslist' 'facebook' 'vast' 'bing' 'meetup'  
'facebook-open-graph' 'email-marketing' 'yahoo' 'padmapper' 'gsp' 'wayn'  
'naver' 'baidu' 'yandex' 'daum']

first\_affiliate\_tracked : ['untracked' 'omg' nan 'linked' 'tracked-other' 'product' 'marketing'  
'local ops']

signup\_app : ['Web' 'Moweb' 'iOS' 'Android']

first\_device\_type : ['Mac Desktop' 'Windows Desktop' 'iPhone' 'Other/Unknown'  
'Desktop (Other)' 'Android Tablet' 'iPad' 'Android Phone'  
'SmartPhone (Other)']

first\_browser : ['Chrome' 'IE' 'Firefox' 'Safari' '-unknown-' 'Mobile Safari'  
'Chrome Mobile' 'RockMelt' 'Chromium' 'Android Browser' 'AOL Explorer'  
'Palm Pre web browser' 'Mobile Firefox' 'Opera' 'TenFourFox' 'IE Mobile'  
'Apple Mail' 'Silk' 'Camino' 'Arora' 'BlackBerry Browser' 'SeaMonkey'  
'Iron' 'Sogou Explorer' 'IceWeasel' 'Opera Mini' 'SiteKiosk' 'Maxthon'  
'Kindle Browser' 'CoolNovo' 'Conqueror' 'wOSBrowser' 'Google Earth'  
'Crazy Browser' 'Mozilla' 'OmniWeb' 'PS Vita browser' 'NetNewsWire'  
'CometBird' 'Comodo Dragon' 'Flock' 'Pale Moon' 'Avant Browser'  
'Opera Mobile' 'Yandex.Browser' 'TheWorld Browser' 'SlimBrowser' 'Epic'  
'Stainless' 'Googlebot' 'Outlook 2007' 'IceDragon']

country\_destination : ['NDF' 'US' 'other' 'FR' 'CA' 'GB' 'ES' 'IT' 'PT' 'NL' 'DE' 'AU']

## 1.4. Y a-t-il des valeurs manquantes?

```
In [8]: print("Valeurs manquantes par colonne:\n")
print(df.isnull().sum())
```

Valeurs manquantes par colonne:

```
date_account_created      0
timestamp_first_active    0
date_first_booking      124543
gender                    0
age                      87990
signup_method             0
signup_flow              0
language                  0
affiliate_channel         0
affiliate_provider        0
first_affiliate_tracked   6065
signup_app               0
first_device_type         0
first_browser            0
country_destination       0
dtype: int64
```

## 1.5. Est-ce qu'il y'a des dépendances évidentes au niveau des descripteurs?

Oui, il peut exister des dépendances entre certains descripteurs. Par exemple, entre:

- ('language', 'country\_destination'): Un utilisateur qui s'inscrit avec la langue française a plus de chances de réserver en France
- ('date\_account\_created', 'timestamp\_first\_active', 'date\_first\_booking'): 'date\_account\_created' est toujours antérieure à 'timestamp\_first\_active' et 'date\_first\_booking' est toujours postérieure aux deux premières dates.
- ('first\_device\_type', 'first\_browser'): certains types d'appareils influencent fortement le navigateur utilisé
- ('affiliate\_provider', 'affiliate\_channel'): certains providers peuvent privilégier certains canaux.

Nous allons valider ces dépendances à l'aide de matrices de corrélations ci-dessous

### Corrélations entre variables qualitatives

```
In [9]: categorical_columns = [
        'gender', 'signup_method', 'language', 'affiliate_channel', 'affiliate_provider', 'first_affiliate_tracked'
        ]
```

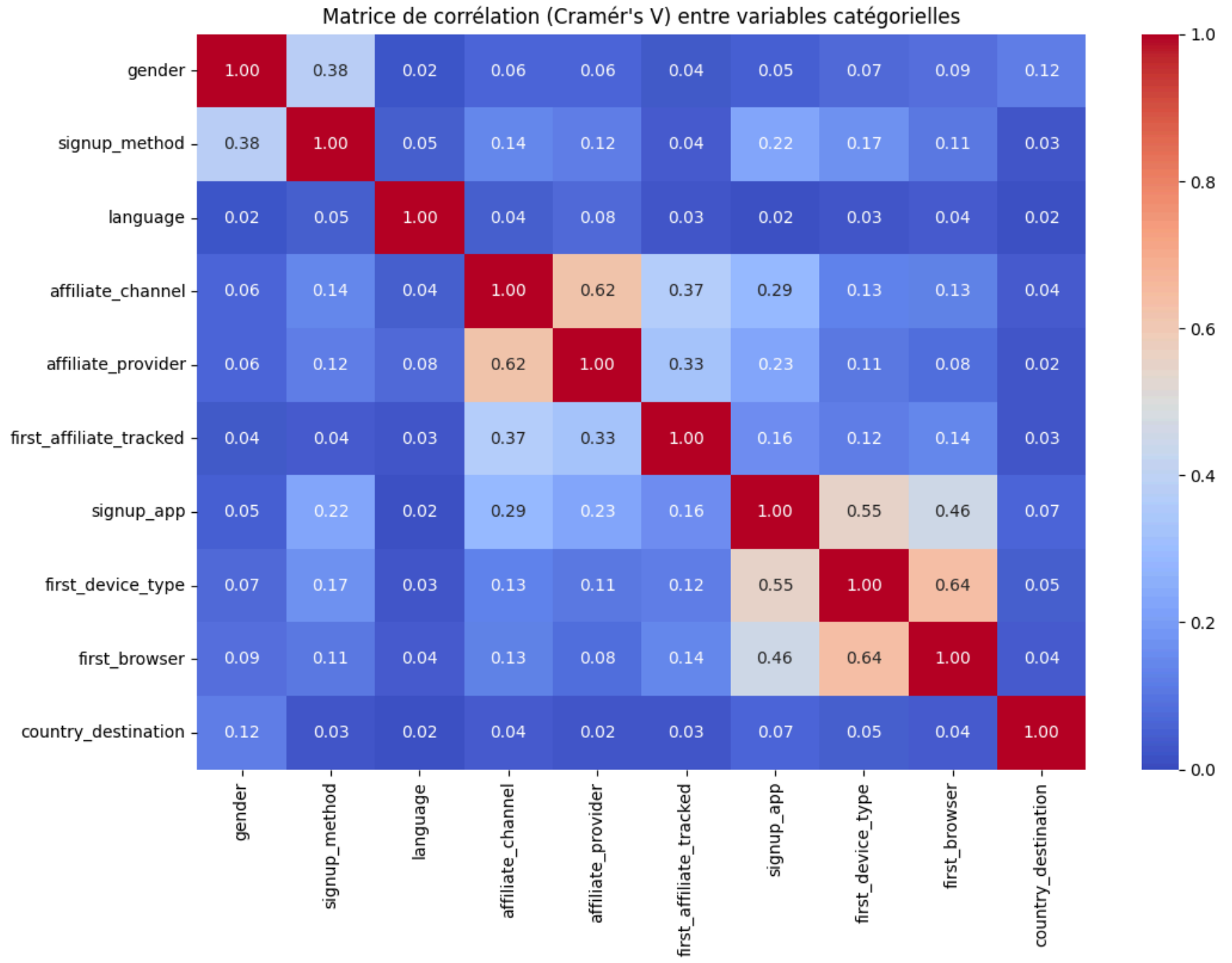
```
# Fonction pour calculer le coefficient de Cramér
def cramers_v(x, y):
    confusion_matrix = pd.crosstab(x, y)
    chi2 = chi2_contingency(confusion_matrix)[0]
    n = confusion_matrix.sum().sum()
    phi2 = chi2 / n
    r, k = confusion_matrix.shape
    phi2corr = max(0, phi2 - ((k-1)*(r-1))/(n-1))
    rcorr = r - ((r-1)**2)/(n-1)
    kcorr = k - ((k-1)**2)/(n-1)
    return np.sqrt(phi2corr / min((kcorr-1), (rcorr-1)))

# Créer une matrice de corrélation
correlation_matrix = pd.DataFrame(index=categorical_columns, columns=categorical_columns)

# Remplir la matrice avec les coefficients de Cramér
for col1 in categorical_columns:
    for col2 in categorical_columns:
        correlation_matrix.loc[col1, col2] = cramers_v(df[col1], df[col2])

# Convertir la matrice en valeurs numériques
correlation_matrix = correlation_matrix.astype(float)

# Afficher la matrice de corrélation avec une heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f", vmin=0, vmax=1)
plt.title("Matrice de corrélation (Cramér's V) entre variables catégorielles")
plt.show()
```



La matrice de corrélation met en lumière les dépendances mentionnées précédemment

## Corrélations entre variables numériques

```
In [10]: df_original = df.copy()
# Vérifier les dépendances temporelles
# Vérifier que les timestamps sont bien ordonnés
df["date_account_created"] = pd.to_datetime(df["date_account_created"]) #Convesion de la colonne "date_accoun

# Conversion de la colonne 'date_account_created' en datetime si ce n'est pas déjà fait
df["timestamp_first_active"] = pd.to_datetime(df["timestamp_first_active"], format='%Y%m%d%H%M%S')
df["date_first_booking"] = pd.to_datetime(df["date_first_booking"])

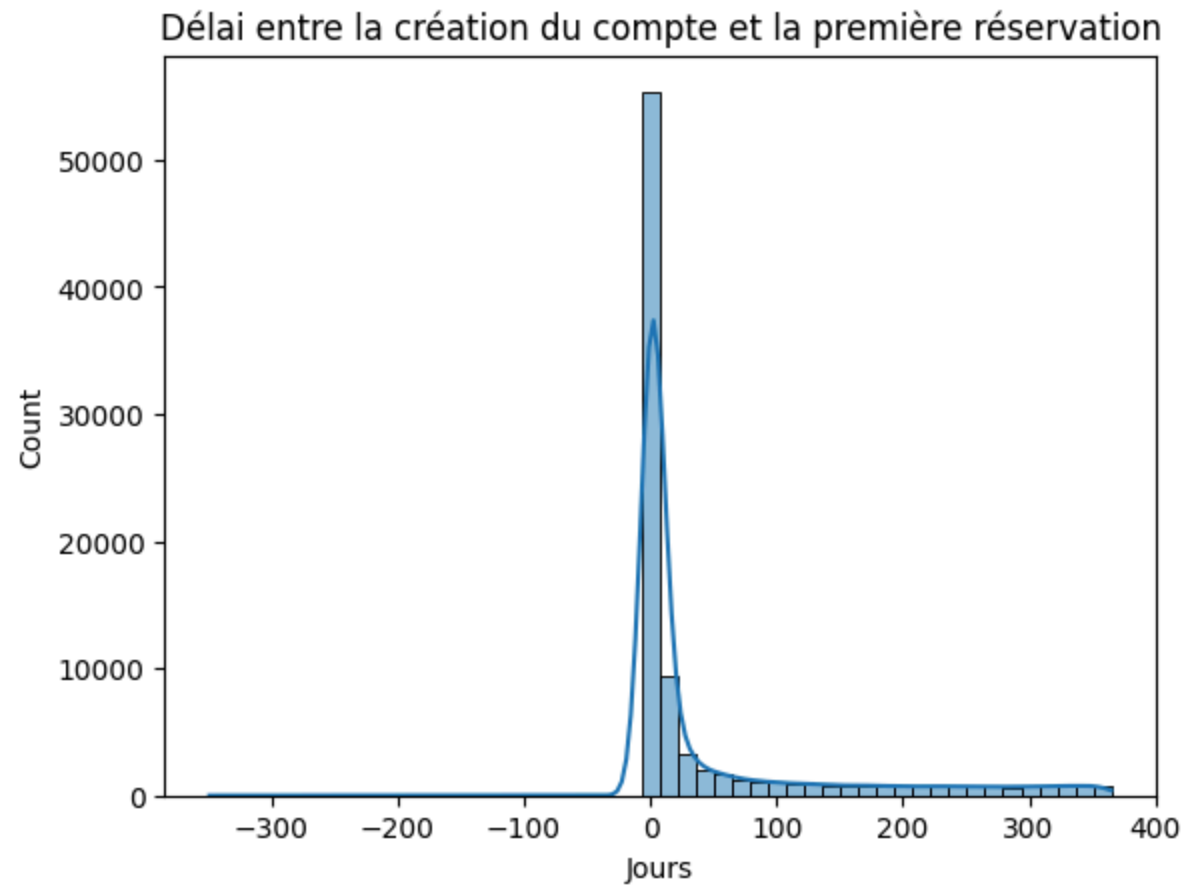
# Vérifier si `timestamp_first_active` est toujours avant ou égal à `date_account_created`
df["timestamp_issue"] = df["timestamp_first_active"] > df["date_account_created"]
print("Nombre de cas où `timestamp_first_active` est postérieur à `date_account_created` :", df["timestamp_is

# Visualiser l'écart entre `date_account_created` et `date_first_booking`
df["booking_delay"] = (df["date_first_booking"] - df["date_account_created"]).dt.days
sns.histplot(df["booking_delay"].dropna(), bins=50, kde=True)
plt.title("Délai entre la création du compte et la première réservation")
plt.xlabel("Jours")
plt.show()

print('Dataset avec dates converties:\n')
display(df)
```

Nombre de cas où `timestamp\_first\_active` est postérieur à `date\_account\_created` : 213273





Dataset avec dates converties:

|            | date_account_created | timestamp_first_active | date_first_booking | gender   | age  | signup_method | signup_flow | language |
|------------|----------------------|------------------------|--------------------|----------|------|---------------|-------------|----------|
| id         |                      |                        |                    |          |      |               |             |          |
| gxn3p5htnn | 2010-06-28           | 2009-03-19 04:32:55    | NaT                | unknown- | NaN  | facebook      | 0           |          |
| 820tgsjxq7 | 2011-05-25           | 2009-05-23 17:48:09    | NaT                | MALE     | 38.0 | facebook      | 0           |          |
| 4ft3gnwmtx | 2010-09-28           | 2009-06-09 23:12:47    | 2010-08-02         | FEMALE   | 56.0 | basic         | 3           |          |
| bjlt8pjhuk | 2011-12-05           | 2009-10-31 06:01:29    | 2012-09-08         | FEMALE   | 42.0 | facebook      | 0           |          |
| 87mebub9p4 | 2010-09-14           | 2009-12-08 06:11:05    | 2010-02-18         | unknown- | 41.0 | basic         | 0           |          |
| ...        | ...                  | ...                    | ...                | ...      | ...  | ...           | ...         | ...      |
| zxodksqep  | 2014-06-30           | 2014-06-30 23:56:36    | NaT                | MALE     | 32.0 | basic         | 0           |          |
| mhewnxesx9 | 2014-06-30           | 2014-06-30 23:57:19    | NaT                | unknown- | NaN  | basic         | 0           |          |
| 6o3arsjbb4 | 2014-06-30           | 2014-06-30 23:57:54    | NaT                | unknown- | 32.0 | basic         | 0           |          |
| jh95kwisub | 2014-06-30           | 2014-06-30 23:58:22    | NaT                | unknown- | NaN  | basic         | 25          |          |
| nw9fwlyb5f | 2014-06-30           | 2014-06-30 23:58:24    | NaT                | unknown- | NaN  | basic         | 25          |          |

213451 rows × 17 columns

Si timestamp\_first\_active est après date\_account\_created, il y a un problème dans les données.

La distribution des délais de réservation permet de voir combien de temps les utilisateurs attendent avant leur première réservation.

```
In [11]: # Convertir l'âge en numérique et traiter les valeurs aberrantes
df["age"] = pd.to_numeric(df["age"], errors="coerce")
df_FilteredAge = df[(df["age"] > 17) & (df["age"] <= 120)] # Filtrer des âges aberrants
display(df_FilteredAge['age'])

# Matrice de corrélation
num_vars = ["age", "signup_flow", "booking_delay"]
corr_matrix = df_FilteredAge[num_vars].corr()

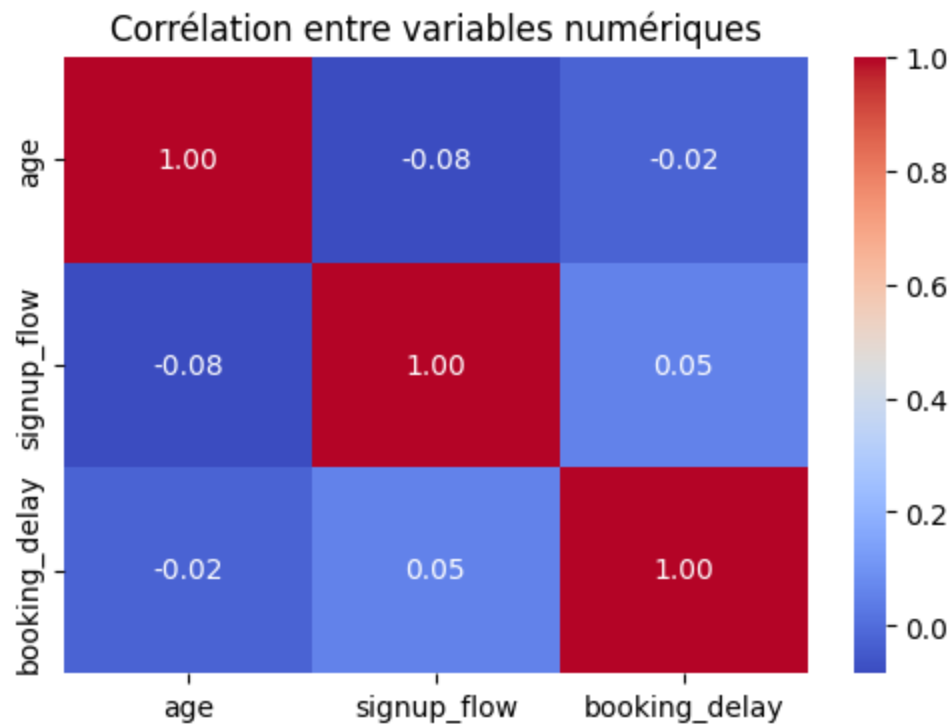
# Affichage
```

```
plt.figure(figsize=(6, 4))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Corrélation entre variables numériques")
plt.show()
```

```
id
820tgsjxq7    38.0
4ft3gnwmtx    56.0
bjjt8pjhuk    42.0
87mebub9p4    41.0
lsw9q7uk0j    46.0
```

```
...
omlc9iku7t    34.0
0k26r3mir0    36.0
qbxza0xojf    23.0
zxodksqpep    32.0
6o3arsjbb4    32.0
```

Name: age, Length: 124522, dtype: float64

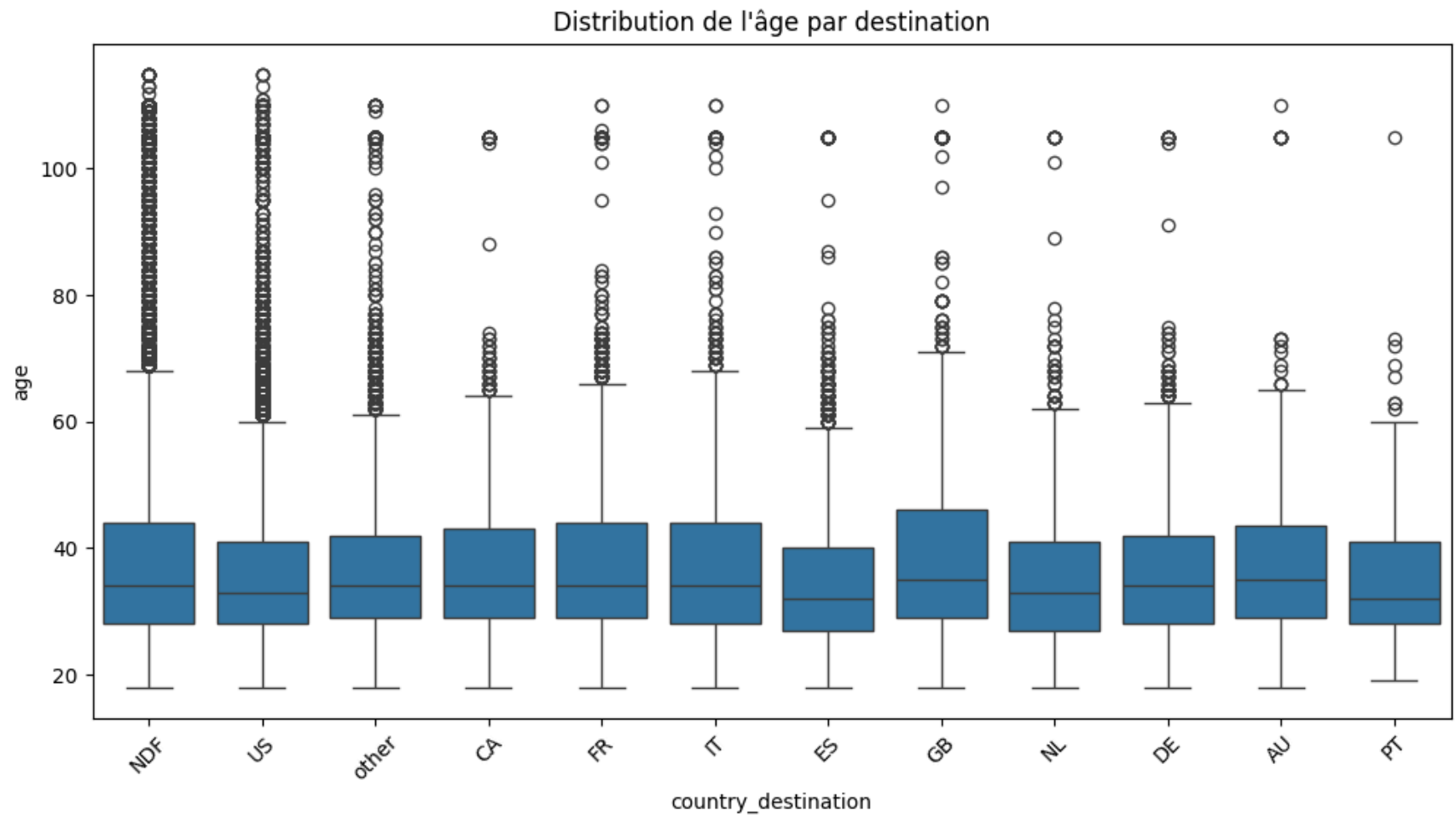


La matrice de corrélation montre qu'il n'y a pas de dépendances entre les variables quantitatives

## 1.6. D'autres observations sur le dataset qui pourraient être pertinentes ?

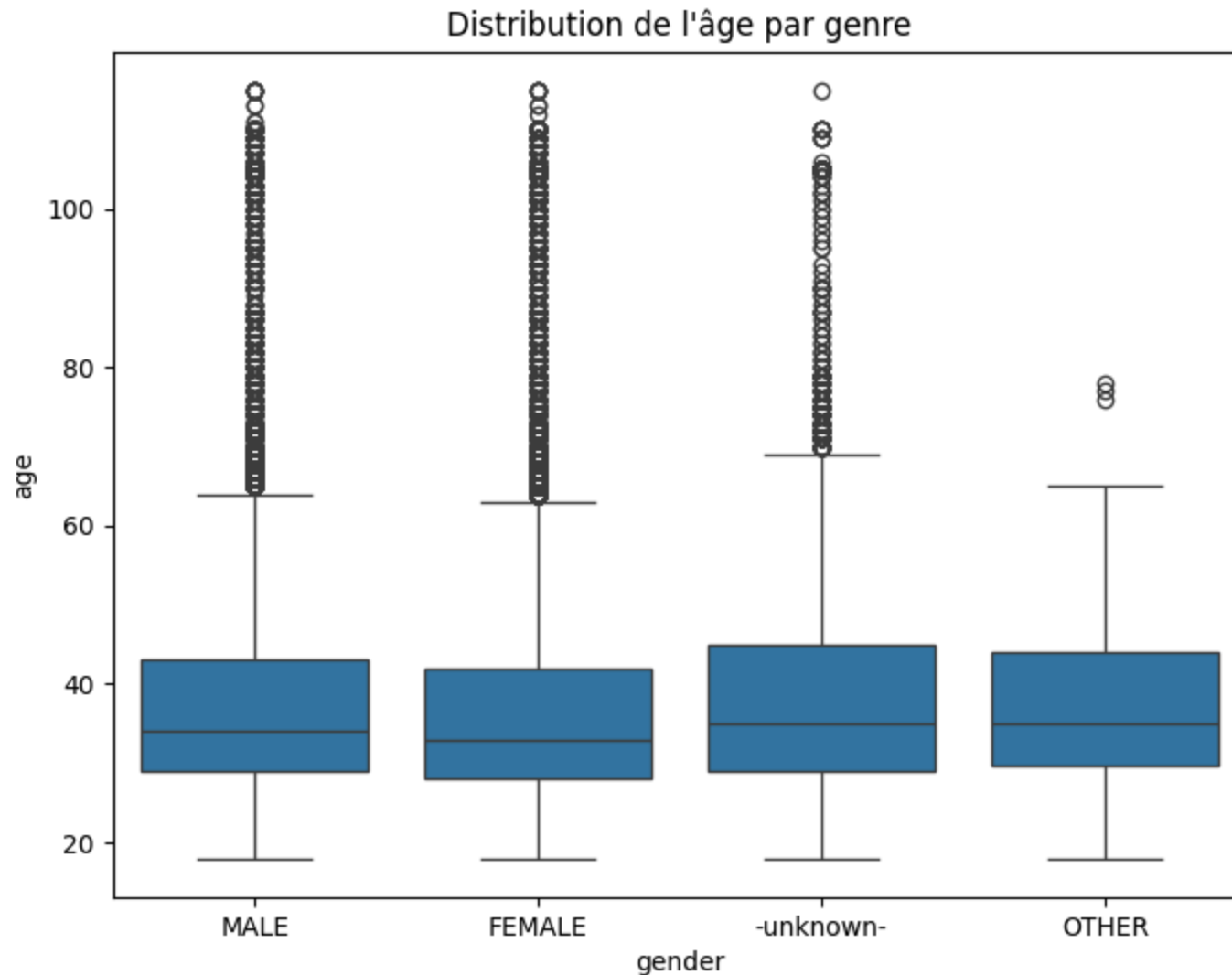
### Boxplot de l'âge en fonction du pays de destination

```
In [12]: #L'âge semble être la seule variable continue intéressante pour un boxplot.  
plt.figure(figsize=(12, 6))  
sns.boxplot(x="country_destination", y="age", data=df_FilteredAge)  
plt.xticks(rotation=45)  
plt.title("Distribution de l'âge par destination")  
plt.show()
```



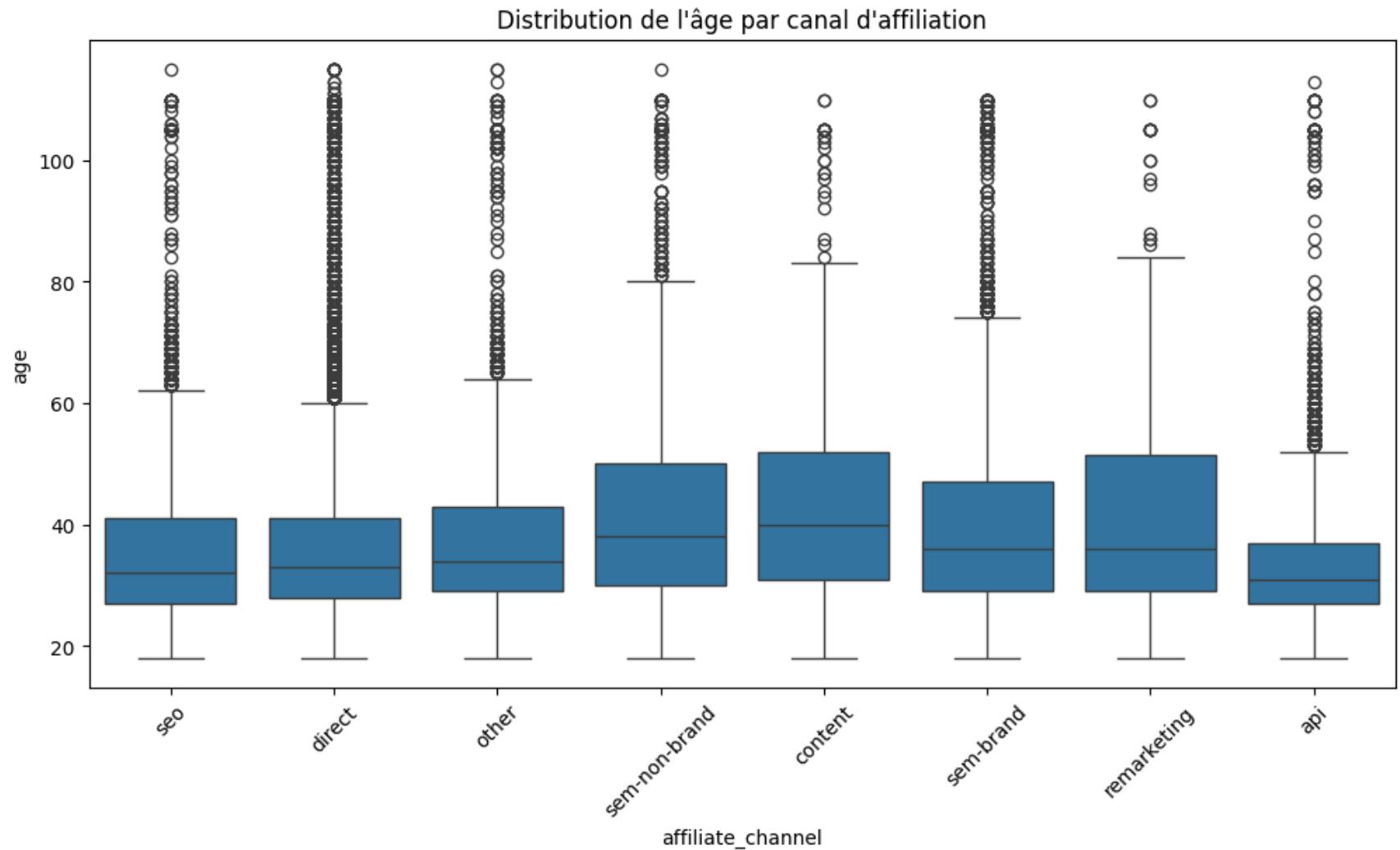
### Boxplot de l'âge selon le sexe :

```
In [13]: plt.figure(figsize=(8, 6))  
sns.boxplot(x="gender", y="age", data=df_FilteredAge)  
plt.title("Distribution de l'âge par genre")  
plt.show()
```



### Boxplot de l'âge selon le canal d'affiliation :

```
In [14]: plt.figure(figsize=(12, 6))
sns.boxplot(x="affiliate_channel", y="age", data=df_FilteredAge)
plt.xticks(rotation=45)
plt.title("Distribution de l'âge par canal d'affiliation")
plt.show()
```



Détection des valeurs aberrantes de l'âge (outliers)

```
In [15]: Q1 = df['age'].quantile(0.25)
Q3 = df['age'].quantile(0.75)
```

```
IQR = Q3 - Q1
```

```
outliers = df[(df['age'] < (Q1 - 1.5 * IQR)) | (df['age'] > (Q3 + 1.5 * IQR))]  
display(outliers[['age']]) # Liste des outliers
```

| age        |       |
|------------|-------|
| id         |       |
| dgatsm5ocq | 69.0  |
| 3qsa4lo7eg | 5.0   |
| 47wdhtdini | 72.0  |
| uhbkw5exeg | 70.0  |
| kw7qyvlhsq | 70.0  |
| ...        | ...   |
| pw9nfo1ulb | 95.0  |
| y37l7vzjpa | 66.0  |
| jl5f10hu4t | 69.0  |
| gfend4omwv | 105.0 |
| l8lttghomx | 69.0  |

5594 rows × 1 columns

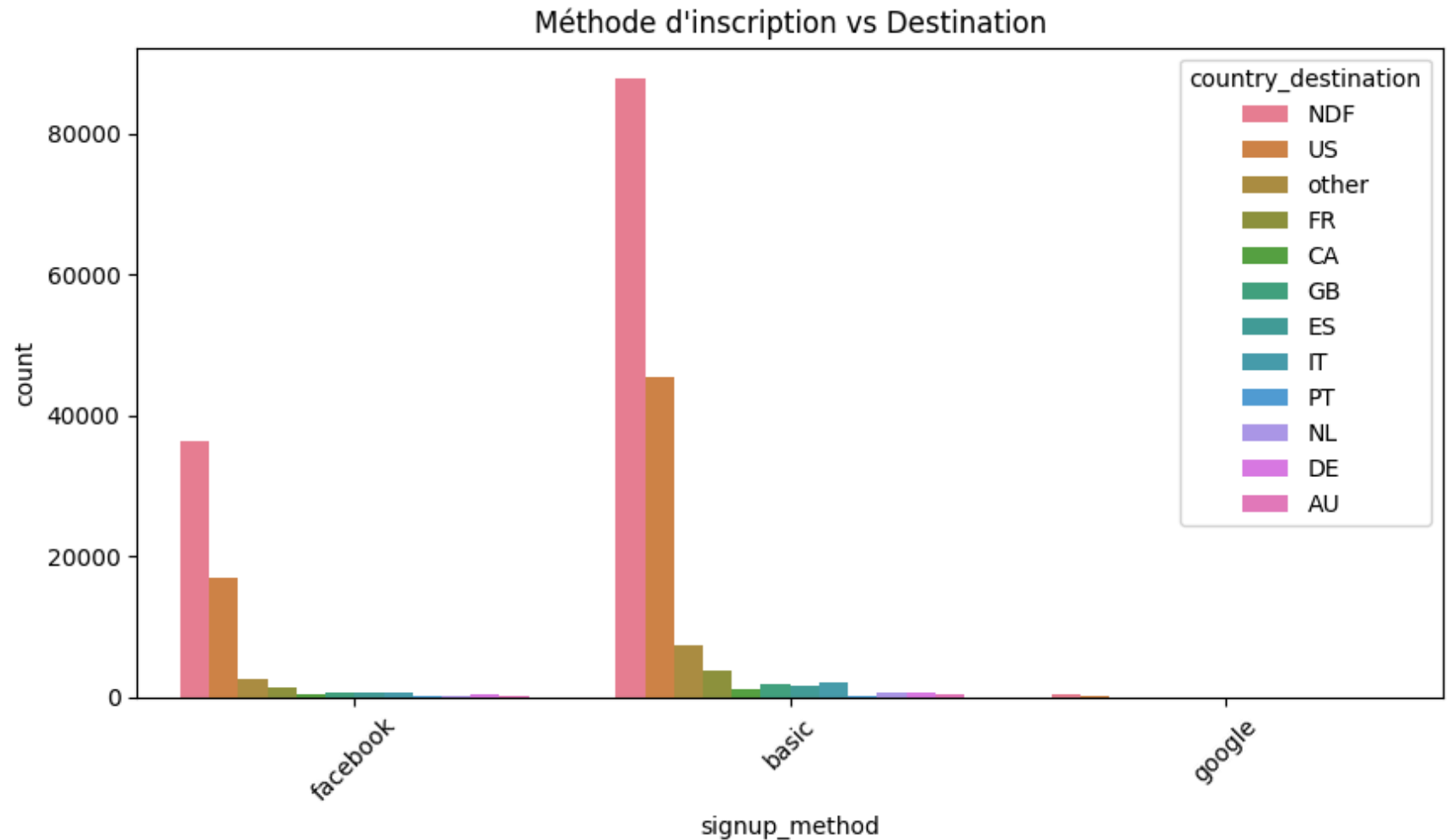
```
In [16]: print(df['age'].describe())
```

```
count    125461.000000  
mean      49.668335  
std      155.666612  
min        1.000000  
25%       28.000000  
50%       34.000000  
75%       43.000000  
max      2014.000000  
Name: age, dtype: float64
```

On constate que la colonne 'age' contient des données aberrantes. Nous ferons un filtrage à l'étape 2.

## Relations entre variables catégorielles

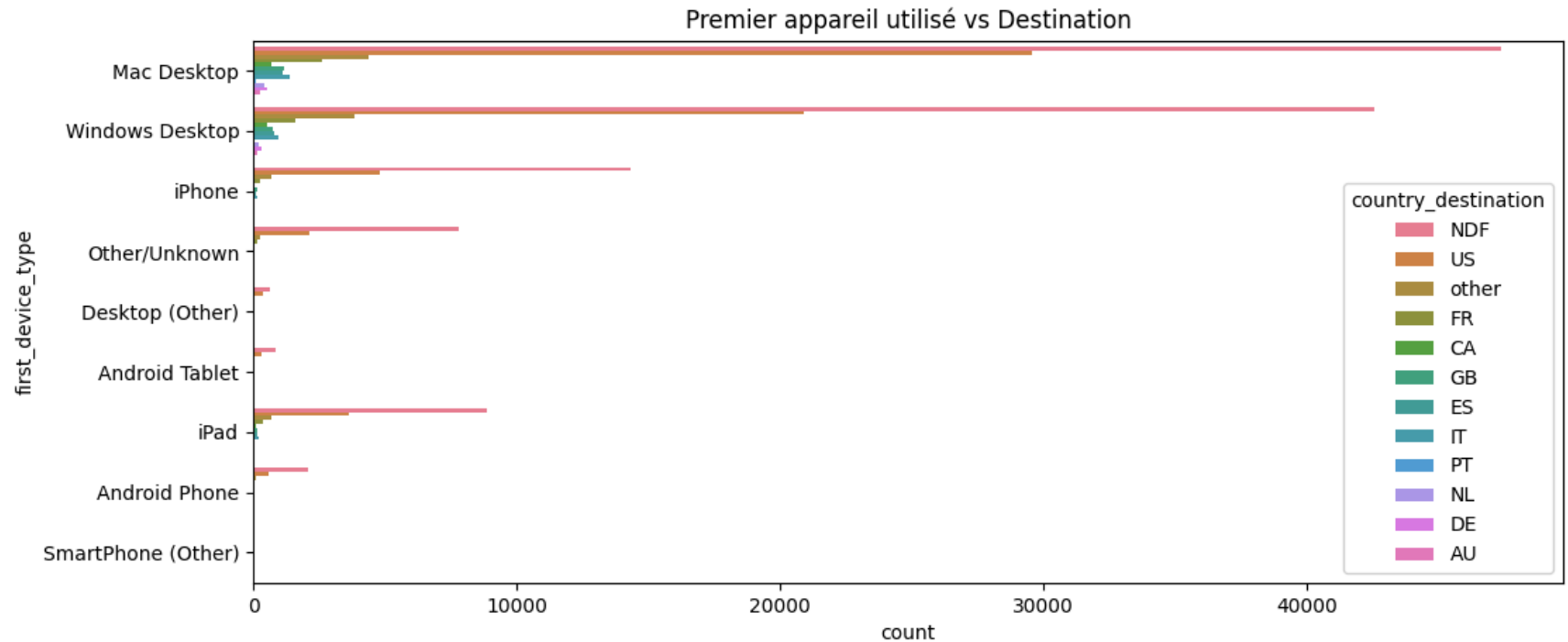
```
In [17]: #Impact de signup_method sur country_destination
plt.figure(figsize=(10, 5))
sns.countplot(data=df, x="signup_method", hue="country_destination")
plt.title("Méthode d'inscription vs Destination")
plt.xticks(rotation=45)
plt.show()
```



Certains modes d'inscription sont peut-être plus populaires pour certaines destinations. Par exemple, les utilisateurs inscrits via Google ou Facebook peuvent être différents de ceux inscrits par email.

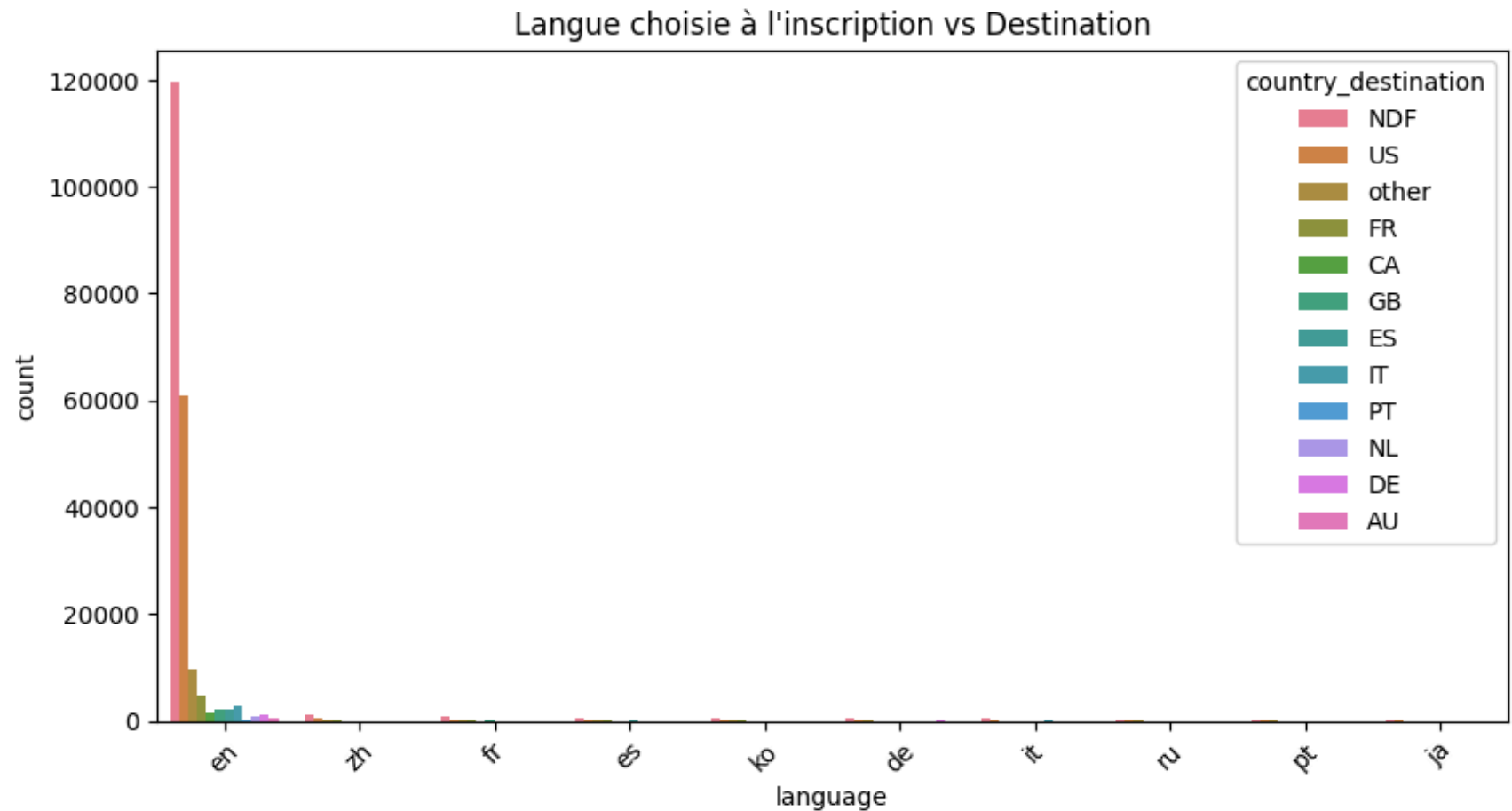


```
In [18]: #Influence de first_device_type sur country_destination
plt.figure(figsize=(12, 5))
sns.countplot(data=df, y="first_device_type", hue="country_destination")
plt.title("Premier appareil utilisé vs Destination")
plt.show()
```



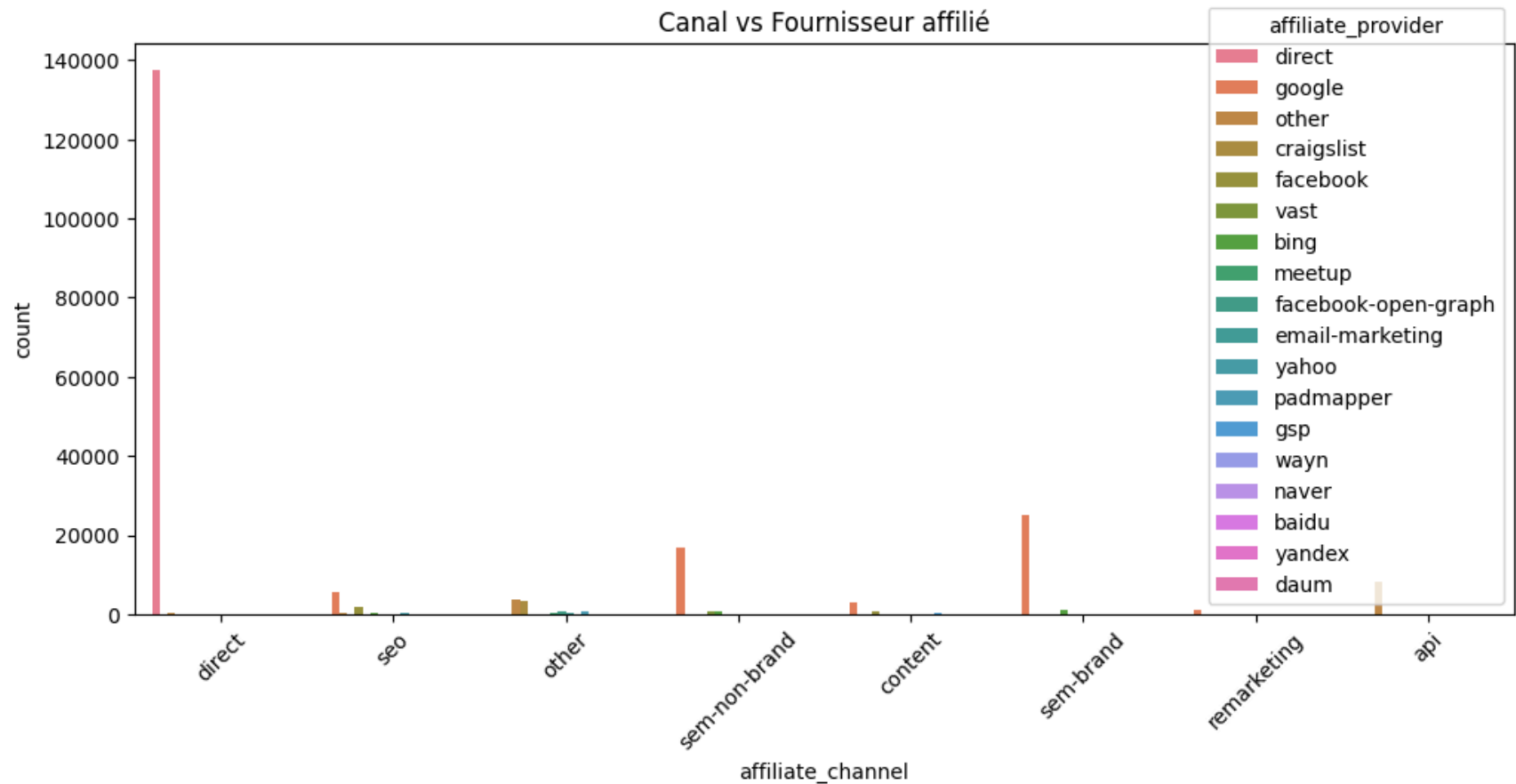
On constate que les utilisateurs mobiles (iPhone, Android) réservent plus rapidement que ceux sur ordinateur

```
In [19]: #Langue (language) et destination
plt.figure(figsize=(10, 5))
sns.countplot(data=df, x="language", hue="country_destination", order=df["language"].value_counts().index[:10])
plt.title("Langue choisie à l'inscription vs Destination")
plt.xticks(rotation=45)
plt.show()
```



On confirme que La langue d'inscription influence la destination finale

```
In [20]: # Vérifier les relations entre affiliés
plt.figure(figsize=(12, 5))
sns.countplot(data=df, x="affiliate_channel", hue="affiliate_provider")
plt.title("Canal vs Fournisseur affilié")
plt.xticks(rotation=45)
plt.show()
```



Le diagramme ci-dessus confirme que certains fournisseurs affiliés sont spécialisés dans certains canaux de conversion

## Étape 2 : On considère le fichier train\_users\_2.csv et test\_users.csv

- Implémenter les correctifs soulignés dans l'étape 1.

```
In [21]: import pandas as pd
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

df = df_original.copy()
```

```
df_test = pd.read_csv('test_users.csv', index_col=0)

print("Données de Train:")
display(df)

print("Données de Test:")
display(df_test)
```

Données de Train:

|            | date_account_created | timestamp_first_active | date_first_booking | gender   | age  | signup_method | signup_flow | language |
|------------|----------------------|------------------------|--------------------|----------|------|---------------|-------------|----------|
| id         |                      |                        |                    |          |      |               |             |          |
| gxn3p5htnn | 2010-06-28           | 20090319043255         | NaN                | unknown- | NaN  | facebook      | 0           |          |
| 820tgsjq7  | 2011-05-25           | 20090523174809         | NaN                | MALE     | 38.0 | facebook      | 0           |          |
| 4ft3gnwmtx | 2010-09-28           | 20090609231247         | 2010-08-02         | FEMALE   | 56.0 | basic         | 3           |          |
| bjit8pjhuk | 2011-12-05           | 20091031060129         | 2012-09-08         | FEMALE   | 42.0 | facebook      | 0           |          |
| 87mebub9p4 | 2010-09-14           | 20091208061105         | 2010-02-18         | unknown- | 41.0 | basic         | 0           |          |
| ...        | ...                  | ...                    | ...                | ...      | ...  | ...           | ...         | ...      |
| zxodksqep  | 2014-06-30           | 20140630235636         | NaN                | MALE     | 32.0 | basic         | 0           |          |
| mhewnxesx9 | 2014-06-30           | 20140630235719         | NaN                | unknown- | NaN  | basic         | 0           |          |
| 6o3arsjbb4 | 2014-06-30           | 20140630235754         | NaN                | unknown- | 32.0 | basic         | 0           |          |
| jh95kwisub | 2014-06-30           | 20140630235822         | NaN                | unknown- | NaN  | basic         | 25          |          |
| nw9fwlyb5f | 2014-06-30           | 20140630235824         | NaN                | unknown- | NaN  | basic         | 25          |          |

213451 rows x 15 columns

Données de Test:

| id          | date_account_created | timestamp_first_active | date_first_booking | gender | age      | signup_method | signup_flow | language |
|-------------|----------------------|------------------------|--------------------|--------|----------|---------------|-------------|----------|
| 5uwns89zht  | 2014-07-01           | 20140701000006         |                    | NaN    | FEMALE   | 35.0          | facebook    | 0        |
| jtl0dijy2j  | 2014-07-01           | 20140701000051         |                    | NaN    | unknown- | NaN           | basic       | 0        |
| xx0ulgorjt  | 2014-07-01           | 20140701000148         |                    | NaN    | unknown- | NaN           | basic       | 0        |
| 6c6puo6ix0  | 2014-07-01           | 20140701000215         |                    | NaN    | unknown- | NaN           | basic       | 0        |
| czqghjk3yfe | 2014-07-01           | 20140701000305         |                    | NaN    | unknown- | NaN           | basic       | 0        |
| ...         | ...                  | ...                    | ...                | ...    | ...      | ...           | ...         | ...      |
| cv0na2lf5a  | 2014-09-30           | 20140930235232         |                    | NaN    | unknown- | 31.0          | basic       | 0        |
| zp8xfonng8  | 2014-09-30           | 20140930235306         |                    | NaN    | unknown- | NaN           | basic       | 23       |
| fa6260ziny  | 2014-09-30           | 20140930235408         |                    | NaN    | unknown- | NaN           | basic       | 0        |
| 87k0fy4ugm  | 2014-09-30           | 20140930235430         |                    | NaN    | unknown- | NaN           | basic       | 0        |
| 9uqfg8txu3  | 2014-09-30           | 20140930235901         |                    | NaN    | FEMALE   | 49.0          | basic       | 0        |

62096 rows × 14 columns

\* Existence les doublons

```
In [22]: print("Doublons dans les données de train:", df.duplicated().unique())
print("Doublons dans les données de test:", df_test.duplicated().unique())
```

Doublons dans les données de train: [False]

Doublons dans les données de test: [False]

## 2.1. Conversion de type/format (les dates)

## \* Dataset de Training

```
In [23]: df['date_account_created'] = pd.to_datetime(df['date_account_created'])
df['timestamp_first_active'] = pd.to_datetime(df['timestamp_first_active'], format='%Y%m%d%H%M%S')

df.drop(['date_first_booking'], axis=1, inplace=True) # Supprimer la colonne 'date_first_booking' car elle es
display(df)
```

| id         | date_account_created | timestamp_first_active | gender   | age  | signup_method | signup_flow | language | affiliate_channel |
|------------|----------------------|------------------------|----------|------|---------------|-------------|----------|-------------------|
| gxn3p5htnn | 2010-06-28           | 2009-03-19 04:32:55    | unknown- | NaN  | facebook      | 0           | en       | direct            |
| 820tgsjq7  | 2011-05-25           | 2009-05-23 17:48:09    | MALE     | 38.0 | facebook      | 0           | en       | search            |
| 4ft3gnwmtx | 2010-09-28           | 2009-06-09 23:12:47    | FEMALE   | 56.0 | basic         | 3           | en       | direct            |
| bjit8pjhuk | 2011-12-05           | 2009-10-31 06:01:29    | FEMALE   | 42.0 | facebook      | 0           | en       | direct            |
| 87mebub9p4 | 2010-09-14           | 2009-12-08 06:11:05    | unknown- | 41.0 | basic         | 0           | en       | direct            |
| ...        | ...                  | ...                    | ...      | ...  | ...           | ...         | ...      | ...               |
| zxodksqep  | 2014-06-30           | 2014-06-30 23:56:36    | MALE     | 32.0 | basic         | 0           | en       | semi-bran         |
| mhewnxesx9 | 2014-06-30           | 2014-06-30 23:57:19    | unknown- | NaN  | basic         | 0           | en       | direct            |
| 6o3arsjbb4 | 2014-06-30           | 2014-06-30 23:57:54    | unknown- | 32.0 | basic         | 0           | en       | direct            |
| jh95kwisub | 2014-06-30           | 2014-06-30 23:58:22    | unknown- | NaN  | basic         | 25          | en       | other             |
| nw9fwlyb5f | 2014-06-30           | 2014-06-30 23:58:24    | unknown- | NaN  | basic         | 25          | en       | direct            |

213451 rows × 14 columns

## \* Dataset de Test

```
In [24]: df_test['date_account_created'] = pd.to_datetime(df_test['date_account_created'])
df_test.drop(['date_first_booking'], axis=1, inplace=True)
```

```
df_test['timestamp_first_active'] = pd.to_datetime(df_test['timestamp_first_active'], format='%Y%m%d%H%M%S')
display(df_test)
```

| id         | date_account_created | timestamp_first_active | gender   | age  | signup_method | signup_flow | language | affiliate_channel |
|------------|----------------------|------------------------|----------|------|---------------|-------------|----------|-------------------|
| 5uwns89zht | 2014-07-01           | 2014-07-01 00:00:06    | FEMALE   | 35.0 | facebook      | 0           | en       | direct            |
| jtl0dijy2j | 2014-07-01           | 2014-07-01 00:00:51    | unknown- | NaN  | basic         | 0           | en       | direct            |
| xx0ulgorjt | 2014-07-01           | 2014-07-01 00:01:48    | unknown- | NaN  | basic         | 0           | en       | direct            |
| 6c6puo6ix0 | 2014-07-01           | 2014-07-01 00:02:15    | unknown- | NaN  | basic         | 0           | en       | direct            |
| czqhjk3yfe | 2014-07-01           | 2014-07-01 00:03:05    | unknown- | NaN  | basic         | 0           | en       | direct            |
| ...        | ...                  | ...                    | ...      | ...  | ...           | ...         | ...      | ...               |
| cv0na2lf5a | 2014-09-30           | 2014-09-30 23:52:32    | unknown- | 31.0 | basic         | 0           | en       | direct            |
| zp8xfonng8 | 2014-09-30           | 2014-09-30 23:53:06    | unknown- | NaN  | basic         | 23          | ko       | direct            |
| fa6260ziny | 2014-09-30           | 2014-09-30 23:54:08    | unknown- | NaN  | basic         | 0           | de       | direct            |
| 87k0fy4ugm | 2014-09-30           | 2014-09-30 23:54:30    | unknown- | NaN  | basic         | 0           | en       | sem-branc         |
| 9uqfg8txu3 | 2014-09-30           | 2014-09-30 23:59:01    | FEMALE   | 49.0 | basic         | 0           | en       | other             |

62096 rows × 13 columns

## 2.2. Remplacement de valeurs manquantes

### \* Dataset de Train

La colonne 'age', présentant un nombre considérable (87990) de valeurs manquantes, nous utiliser une méthode de prédiction (RandomForest) au lieu d'un Imputer pour remplacer les données manquantes et évite d'aplatir la distribution des âges. C'est

une méthode plus réaliste, surtout si l'âge a un impact sur la destination

```
In [25]: age_data = df[df['age'].notnull()]
age_target = age_data['age']
age_features = age_data.drop(['age'], axis=1).select_dtypes(include=[np.number])

age_model = RandomForestRegressor()
age_model.fit(age_features, age_target)

# Prédire les valeurs manquantes dans 'age'
missing_age_data = df[df['age'].isnull()]
predicted_ages = age_model.predict(missing_age_data.drop(['age'], axis=1).select_dtypes(include=[np.number]))
df.loc[df['age'].isnull(), 'age'] = predicted_ages
```

Remplacement des valeurs manquantes de 'first\_affiliate\_tracked' par sa valeur médiane

```
In [26]: # c. Remplacement de first_affiliate_tracked avec la médiane
df['first_affiliate_tracked'].fillna(df['first_affiliate_tracked'].mode()[0], inplace=True)
```

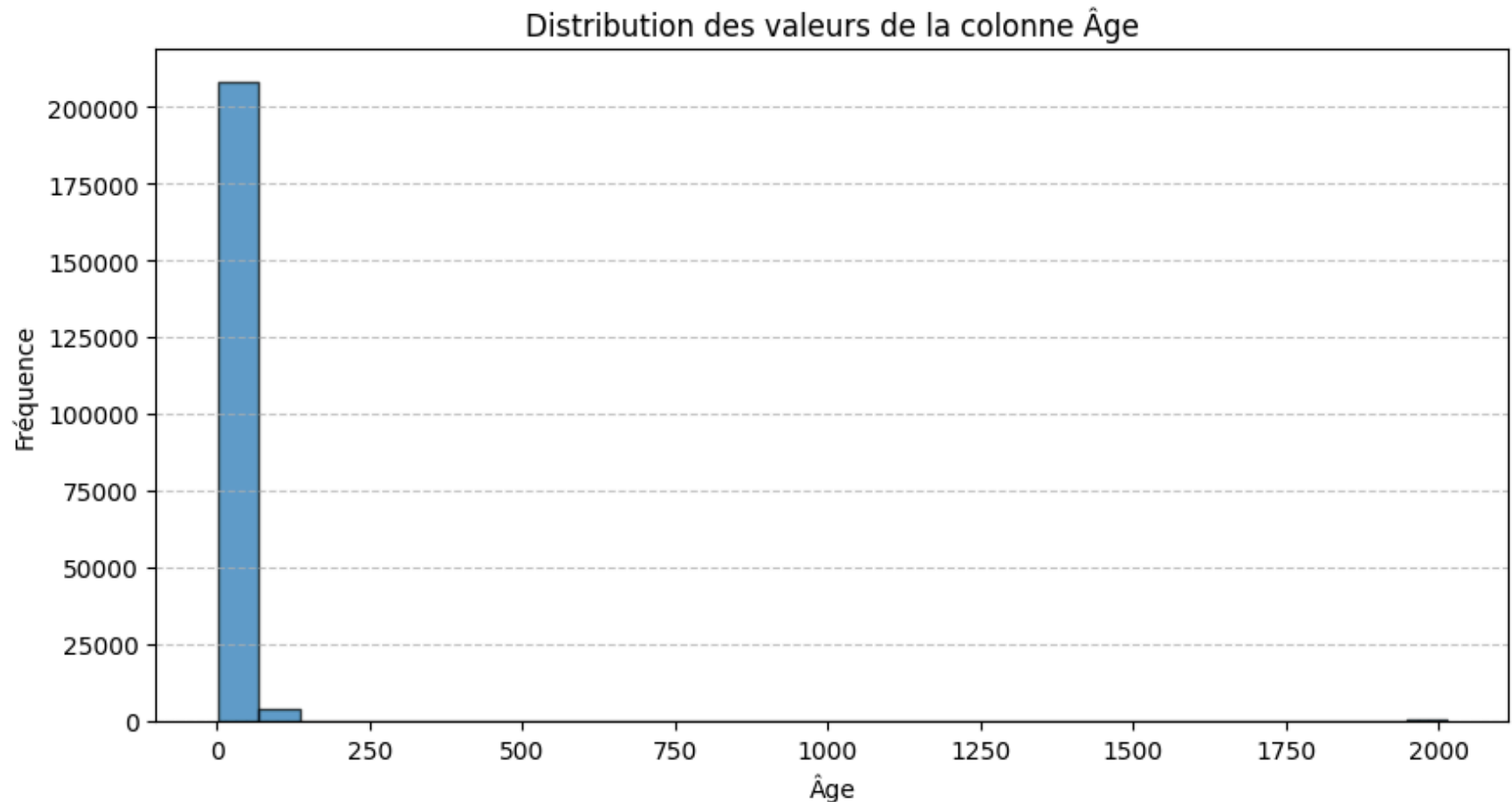
/var/folders/dz/dt7pkrls1kxg9y931v65tmz40000gn/T/ipykernel\_35413/4188757126.py:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method. The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

```
df['first_affiliate_tracked'].fillna(df['first_affiliate_tracked'].mode()[0], inplace=True)
```

```
In [27]: # Afficher la distribution des valeurs de la colonne 'age'
plt.figure(figsize=(10, 5))
plt.hist(df['age'].dropna(), bins=30, edgecolor='black', alpha=0.7)
plt.xlabel('Âge')
plt.ylabel('Fréquence')
plt.title('Distribution des valeurs de la colonne Âge')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



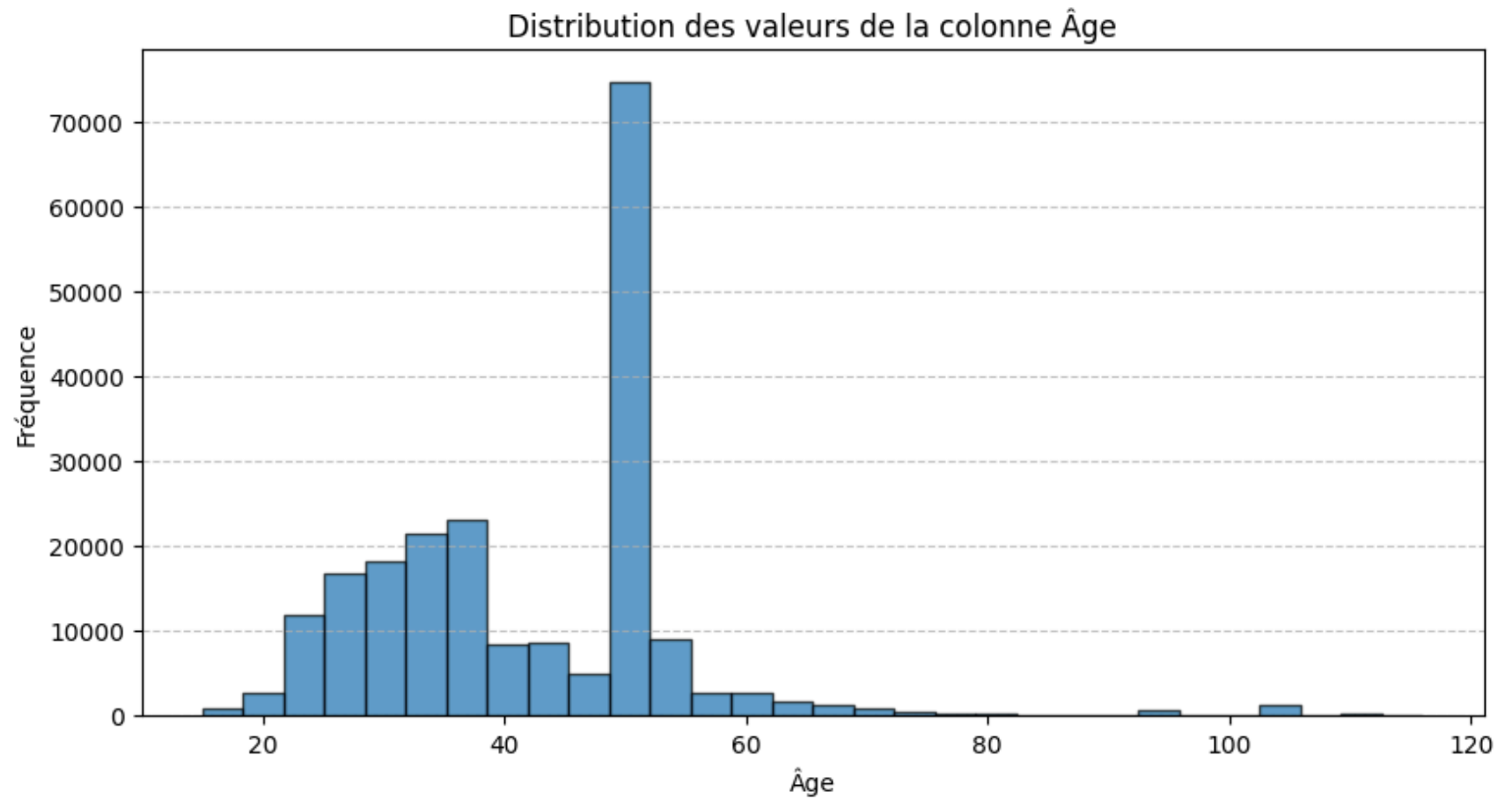


On constate la présence de valeurs aberrantes dans la colonne 'age'. Pour y remédier, nous supprimons les enregistrements où l'âge est inférieur à 15 ou supérieur à 120.

```
In [28]: # 3. Correction/Suppression de valeurs aberrantes/erronées
df = df[(df['age'] >= 15) & (df['age'] <= 120)] # Suppression des âges aberrants
df['age'] = df['age'].astype(int)
```

```
In [29]: # Afficher la distribution des valeurs de la colonne 'age'
plt.figure(figsize=(10, 5))
plt.hist(df['age'].dropna(), bins=30, edgecolor='black', alpha=0.7)
plt.xlabel('Âge')
plt.ylabel('Fréquence')
plt.title('Distribution des valeurs de la colonne Âge')
```

```
plt.grid(axis='y', linestyle='--', alpha=0.7)  
plt.show()
```



```
In [30]: display(df)
```

| id         | date_account_created | timestamp_first_active | gender   | age | signup_method | signup_flow | language | affiliate_channel |
|------------|----------------------|------------------------|----------|-----|---------------|-------------|----------|-------------------|
| gxn3p5htnn | 2010-06-28           | 2009-03-19 04:32:55    | unknown- | 50  | facebook      | 0           | en       | direct            |
| 820tgsjxq7 | 2011-05-25           | 2009-05-23 17:48:09    | MALE     | 38  | facebook      | 0           | en       | second            |
| 4ft3gnwmtx | 2010-09-28           | 2009-06-09 23:12:47    | FEMALE   | 56  | basic         | 3           | en       | direct            |
| bjlt8pjhuk | 2011-12-05           | 2009-10-31 06:01:29    | FEMALE   | 42  | facebook      | 0           | en       | direct            |
| 87mebub9p4 | 2010-09-14           | 2009-12-08 06:11:05    | unknown- | 41  | basic         | 0           | en       | direct            |
| ...        | ...                  | ...                    | ...      | ... | ...           | ...         | ...      | ...               |
| zxodksqep  | 2014-06-30           | 2014-06-30 23:56:36    | MALE     | 32  | basic         | 0           | en       | semi-bran         |
| mhewnxesx9 | 2014-06-30           | 2014-06-30 23:57:19    | unknown- | 50  | basic         | 0           | en       | direct            |
| 6o3arsjbb4 | 2014-06-30           | 2014-06-30 23:57:54    | unknown- | 32  | basic         | 0           | en       | direct            |
| jh95kwisub | 2014-06-30           | 2014-06-30 23:58:22    | unknown- | 38  | basic         | 25          | en       | other             |
| nw9fwlyb5f | 2014-06-30           | 2014-06-30 23:58:24    | unknown- | 38  | basic         | 25          | en       | direct            |

212613 rows × 14 columns

#### \* Dataset de Test

Les commentaires du dataset de train s'appliquent aussi pour le test.

```
In [31]: print(df_test.describe())
print("Valeurs manquantes par colonne:\n")
print(df_test.isnull().sum()) # Compte les valeurs manquantes par colonne
```

|       | date_account_created          | timestamp_first_active \      |
|-------|-------------------------------|-------------------------------|
| count | 62096                         | 62096                         |
| mean  | 2014-08-14 19:24:31.631022848 | 2014-08-15 07:57:04.415598080 |
| min   | 2014-07-01 00:00:00           | 2014-07-01 00:00:06           |
| 25%   | 2014-07-24 00:00:00           | 2014-07-24 00:04:38.500000    |
| 50%   | 2014-08-14 00:00:00           | 2014-08-14 02:36:11           |
| 75%   | 2014-09-05 00:00:00           | 2014-09-05 22:39:30           |
| max   | 2014-09-30 00:00:00           | 2014-09-30 23:59:01           |
| std   | NaN                           | NaN                           |

|       | age          | signup_flow  |
|-------|--------------|--------------|
| count | 33220.000000 | 62096.000000 |
| mean  | 37.616677    | 7.813885     |
| min   | 1.000000     | 0.000000     |
| 25%   | 26.000000    | 0.000000     |
| 50%   | 31.000000    | 0.000000     |
| 75%   | 40.000000    | 23.000000    |
| max   | 2002.000000  | 25.000000    |
| std   | 74.440647    | 11.254291    |

Valeurs manquantes par colonne:

|                         |       |
|-------------------------|-------|
| date_account_created    | 0     |
| timestamp_first_active  | 0     |
| gender                  | 0     |
| age                     | 28876 |
| signup_method           | 0     |
| signup_flow             | 0     |
| language                | 0     |
| affiliate_channel       | 0     |
| affiliate_provider      | 0     |
| first_affiliate_tracked | 20    |
| signup_app              | 0     |
| first_device_type       | 0     |
| first_browser           | 0     |

dtype: int64

```
In [32]: age_data = df_test[df_test['age'].notnull()]
age_target = age_data['age']
age_features = age_data.drop(['age'], axis=1).select_dtypes(include=[np.number])

age_model = RandomForestRegressor()
age_model.fit(age_features, age_target)

# Prédire les valeurs manquantes dans 'age'
missing_age_data = df_test[df_test['age'].isnull()]
```

```
predicted_ages = age_model.predict(missing_age_data.drop(['age'], axis=1).select_dtypes(include=[np.number]))  
df_test.loc[df_test['age'].isnull(), 'age'] = predicted_ages
```

In [33]: *# c. Remplacement de first\_affiliate\_tracked avec la médiane*

```
df_test['first_affiliate_tracked'].fillna(df_test['first_affiliate_tracked'].mode()[0], inplace=True)
```

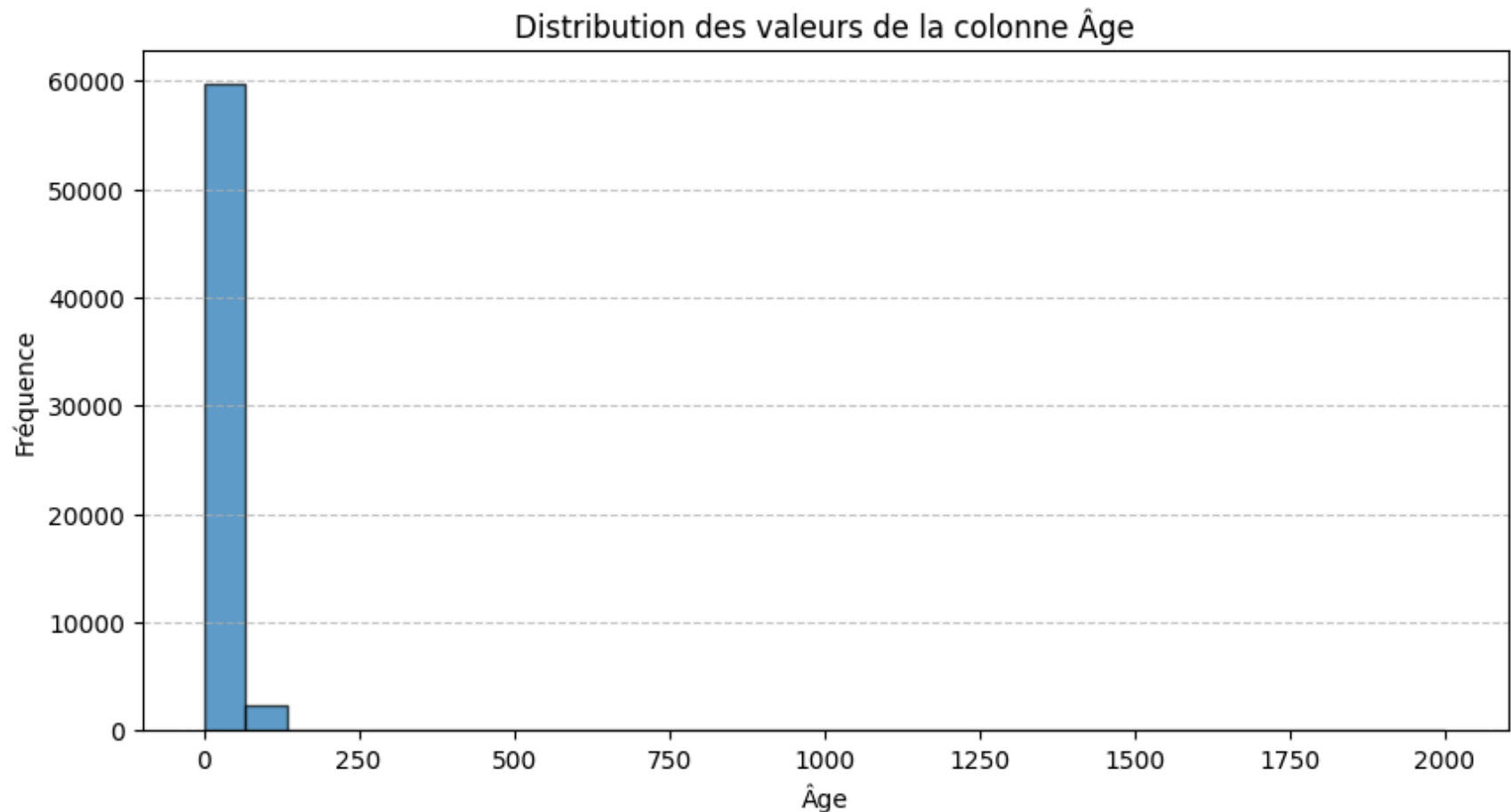
/var/folders/dz/dt7pkrls1kxg9y931v65tmz40000gn/T/ipykernel\_35413/961816202.py:3: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method. The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

```
df_test['first_affiliate_tracked'].fillna(df_test['first_affiliate_tracked'].mode()[0], inplace=True)
```

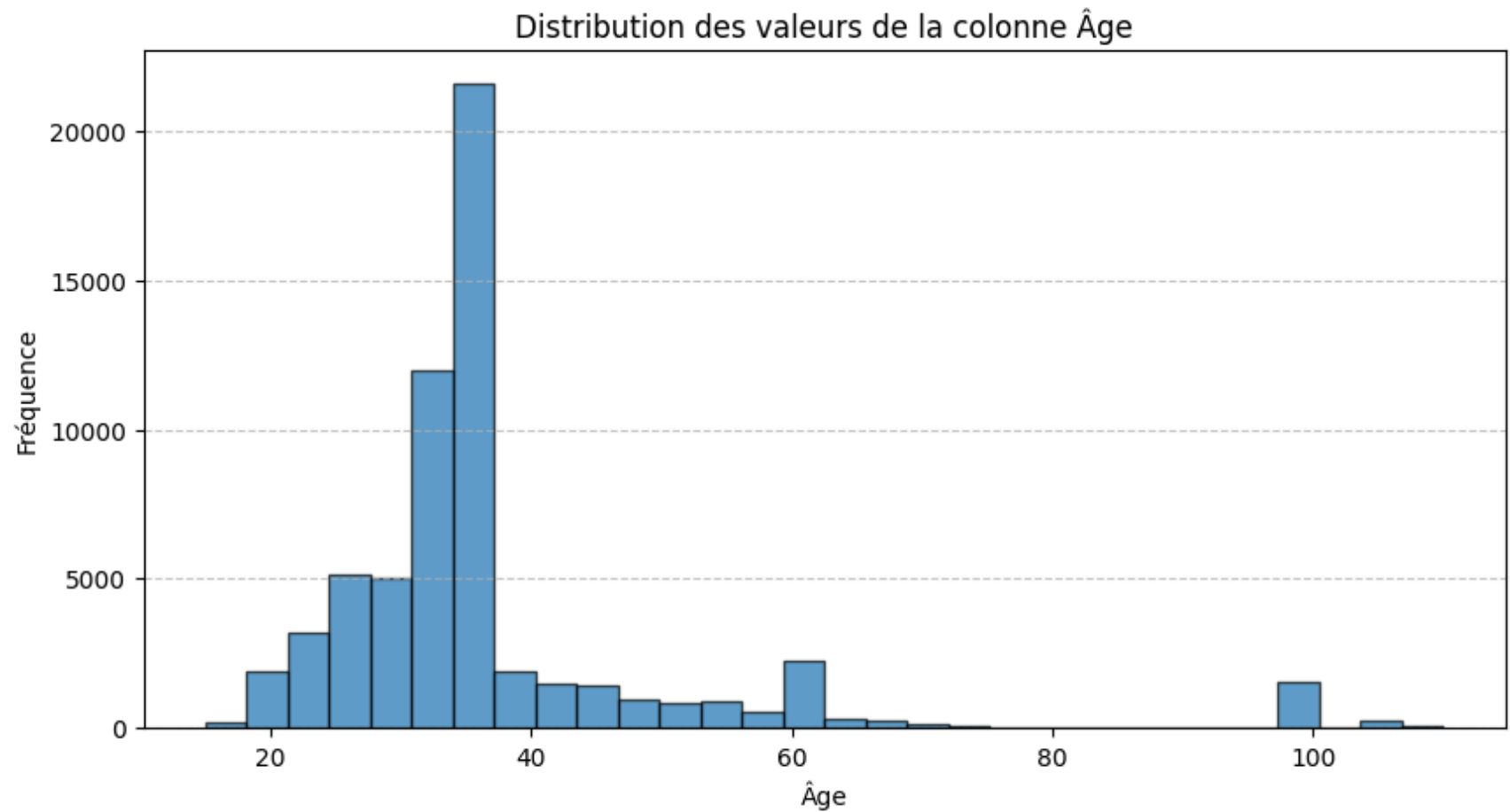
In [34]: *# Afficher la distribution des valeurs de la colonne 'age'*

```
plt.figure(figsize=(10, 5))  
plt.hist(df_test['age'].dropna(), bins=30, edgecolor='black', alpha=0.7)  
plt.xlabel('Âge')  
plt.ylabel('Fréquence')  
plt.title('Distribution des valeurs de la colonne Âge')  
plt.grid(axis='y', linestyle='--', alpha=0.7)  
plt.show()
```



```
In [35]: # 3. Correction/Suppression de valeurs aberrantes/erronées
df_test = df_test[(df_test['age'] >= 15) & (df_test['age'] <= 120)] # Suppression des âges aberrants
df_test['age'] = df_test['age'].astype(int)
```

```
In [36]: # Afficher la distribution des valeurs de la colonne 'age'
plt.figure(figsize=(10, 5))
plt.hist(df_test['age'].dropna(), bins=30, edgecolor='black', alpha=0.7)
plt.xlabel('Âge')
plt.ylabel('Fréquence')
plt.title('Distribution des valeurs de la colonne Âge')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



```
In [37]: display(df_test)
```

| id          | date_account_created | timestamp_first_active | gender   | age | signup_method | signup_flow | language | affiliate_channel |
|-------------|----------------------|------------------------|----------|-----|---------------|-------------|----------|-------------------|
| 5uwns89zht  | 2014-07-01           | 2014-07-01 00:00:06    | FEMALE   | 35  | facebook      | 0           | en       | direct            |
| jtl0dijy2j  | 2014-07-01           | 2014-07-01 00:00:51    | unknown- | 36  | basic         | 0           | en       | direct            |
| xx0ulgorjt  | 2014-07-01           | 2014-07-01 00:01:48    | unknown- | 36  | basic         | 0           | en       | direct            |
| 6c6puo6ix0  | 2014-07-01           | 2014-07-01 00:02:15    | unknown- | 36  | basic         | 0           | en       | direct            |
| czqghjk3yfe | 2014-07-01           | 2014-07-01 00:03:05    | unknown- | 36  | basic         | 0           | en       | direct            |
| ...         | ...                  | ...                    | ...      | ... | ...           | ...         | ...      | ...               |
| cv0na2lf5a  | 2014-09-30           | 2014-09-30 23:52:32    | unknown- | 31  | basic         | 0           | en       | direct            |
| zp8xfonng8  | 2014-09-30           | 2014-09-30 23:53:06    | unknown- | 61  | basic         | 23          | ko       | direct            |
| fa6260ziny  | 2014-09-30           | 2014-09-30 23:54:08    | unknown- | 36  | basic         | 0           | de       | direct            |
| 87k0fy4ugm  | 2014-09-30           | 2014-09-30 23:54:30    | unknown- | 36  | basic         | 0           | en       | sem-brand         |
| 9uqfg8txu3  | 2014-09-30           | 2014-09-30 23:59:01    | FEMALE   | 49  | basic         | 0           | en       | other             |

62045 rows × 13 columns

## 2.3. Standardisation de la dataset train

```
In [38]: import pandas as pd

df_qualitatives = df.select_dtypes(include=['object'])
df_quantitatives = df.select_dtypes(include=[np.number])
df_dates = df.select_dtypes(include=['datetime64'])
df_target = df['country_destination']

display(df_qualitatives)
display(df_quantitatives)
```



```
display(df_dates)
display(df_target)
```

|                   | gender        | signup_method | language | affiliate_channel | affiliate_provider | first_affiliate_tracked | signup_app | first_device |
|-------------------|---------------|---------------|----------|-------------------|--------------------|-------------------------|------------|--------------|
| id                |               |               |          |                   |                    |                         |            |              |
| <b>gxn3p5htnn</b> | -<br>unknown- | facebook      | en       | direct            | direct             | untracked               | Web        | Mac I        |
| <b>820tgsjxq7</b> | MALE          | facebook      | en       | seo               | google             | untracked               | Web        | Mac I        |
| <b>4ft3gnwmtx</b> | FEMALE        | basic         | en       | direct            | direct             | untracked               | Web        | Windows I    |
| <b>bjlt8pjhuk</b> | FEMALE        | facebook      | en       | direct            | direct             | untracked               | Web        | Mac I        |
| <b>87mebub9p4</b> | -<br>unknown- | basic         | en       | direct            | direct             | untracked               | Web        | Mac I        |
| ...               | ...           | ...           | ...      | ...               | ...                | ...                     | ...        |              |
| <b>zxodksqep</b>  | MALE          | basic         | en       | sem-brand         | google             | omg                     | Web        | Mac I        |
| <b>mhewnxesx9</b> | -<br>unknown- | basic         | en       | direct            | direct             | linked                  | Web        | Windows I    |
| <b>6o3arsjbb4</b> | -<br>unknown- | basic         | en       | direct            | direct             | untracked               | Web        | Mac I        |
| <b>jh95kwisub</b> | -<br>unknown- | basic         | en       | other             | other              | tracked-other           | iOS        |              |
| <b>nw9fwlyb5f</b> | -<br>unknown- | basic         | en       | direct            | direct             | untracked               | iOS        |              |

212613 rows × 10 columns

|            | age | signup_flow |
|------------|-----|-------------|
| id         |     |             |
| gxn3p5htnn | 50  | 0           |
| 820tgsjxq7 | 38  | 0           |
| 4ft3gnwmtx | 56  | 3           |
| bjlt8pjhuk | 42  | 0           |
| 87mebub9p4 | 41  | 0           |
| ...        | ... | ...         |
| zxodksqpep | 32  | 0           |
| mhewnxesx9 | 50  | 0           |
| 6o3arsjbb4 | 32  | 0           |
| jh95kwisub | 38  | 25          |
| nw9fwlyb5f | 38  | 25          |

212613 rows × 2 columns

|            | date_account_created | timestamp_first_active |
|------------|----------------------|------------------------|
| id         |                      |                        |
| gxn3p5htnn | 2010-06-28           | 2009-03-19 04:32:55    |
| 820tgsjq7  | 2011-05-25           | 2009-05-23 17:48:09    |
| 4ft3gnwmtx | 2010-09-28           | 2009-06-09 23:12:47    |
| bjjt8pjhuk | 2011-12-05           | 2009-10-31 06:01:29    |
| 87mebub9p4 | 2010-09-14           | 2009-12-08 06:11:05    |
| ...        | ...                  | ...                    |
| zxodksqep  | 2014-06-30           | 2014-06-30 23:56:36    |
| mhewnxesx9 | 2014-06-30           | 2014-06-30 23:57:19    |
| 6o3arsjbb4 | 2014-06-30           | 2014-06-30 23:57:54    |
| jh95kwisub | 2014-06-30           | 2014-06-30 23:58:22    |
| nw9fwlyb5f | 2014-06-30           | 2014-06-30 23:58:24    |

212613 rows × 2 columns

```

id
gxn3p5htnn      NDF
820tgsjq7       NDF
4ft3gnwmtx      US
bjjt8pjhuk      other
87mebub9p4      US
...
zxodksqep       NDF
mhewnxesx9      NDF
6o3arsjbb4      NDF
jh95kwisub      NDF
nw9fwlyb5f      NDF
Name: country_destination, Length: 212613, dtype: object

```

## Encodage des variables catégorielles de la dataset train

Dans cette étape, nous appliquons le OneHotEncoder afin de convertir les données catégorielles en format numérique. Cette transformation crée des indicateurs binaires pour chaque modalité.

```
In [39]: from sklearn.preprocessing import OneHotEncoder

df_qualitatives = df_qualitatives.drop(['country_destination'], axis=1)
colonnes_qualitatives = df_qualitatives.select_dtypes(include=['object']).columns

encoder = OneHotEncoder()
encoded_data = encoder.fit_transform(df_qualitatives[colonnes_qualitatives]).toarray()

df_encoded = pd.DataFrame(encoded_data, columns=encoder.get_feature_names_out(colonnes_qualitatives))
print(df_encoded)

p_k = np.mean(df_encoded.values, axis=0)
print(p_k)

ZD = df_encoded.values/np.sqrt(p_k)
display(ZD.round(3))

df_encoded = pd.DataFrame(ZD, columns=encoder.get_feature_names_out(colonnes_qualitatives), index=df.index)
print(df_encoded)
```

|        | gender_—unknown— | gender_FEMALE | gender_MALE | gender_OTHER | \ |
|--------|------------------|---------------|-------------|--------------|---|
| 0      | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 1      | 0.0              | 0.0           | 1.0         | 0.0          |   |
| 2      | 0.0              | 1.0           | 0.0         | 0.0          |   |
| 3      | 0.0              | 1.0           | 0.0         | 0.0          |   |
| 4      | 1.0              | 0.0           | 0.0         | 0.0          |   |
| ...    | ...              | ...           | ...         | ...          |   |
| 212608 | 0.0              | 0.0           | 1.0         | 0.0          |   |
| 212609 | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 212610 | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 212611 | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 212612 | 1.0              | 0.0           | 0.0         | 0.0          |   |

|        | signup_method_basic | signup_method_facebook | signup_method_google | \ |
|--------|---------------------|------------------------|----------------------|---|
| 0      | 0.0                 | 1.0                    | 0.0                  |   |
| 1      | 0.0                 | 1.0                    | 0.0                  |   |
| 2      | 1.0                 | 0.0                    | 0.0                  |   |
| 3      | 0.0                 | 1.0                    | 0.0                  |   |
| 4      | 1.0                 | 0.0                    | 0.0                  |   |
| ...    | ...                 | ...                    | ...                  |   |
| 212608 | 1.0                 | 0.0                    | 0.0                  |   |
| 212609 | 1.0                 | 0.0                    | 0.0                  |   |
| 212610 | 1.0                 | 0.0                    | 0.0                  |   |
| 212611 | 1.0                 | 0.0                    | 0.0                  |   |
| 212612 | 1.0                 | 0.0                    | 0.0                  |   |

|        | language_ca | language_cs | language_da | ... | first_browser_SeaMonkey | \ |
|--------|-------------|-------------|-------------|-----|-------------------------|---|
| 0      | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| 1      | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| 2      | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| 3      | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| 4      | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| ...    | ...         | ...         | ...         | ... | ...                     |   |
| 212608 | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| 212609 | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| 212610 | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| 212611 | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |
| 212612 | 0.0         | 0.0         | 0.0         | ... | 0.0                     |   |

|   | first_browser_Silk | first_browser_SiteKiosk | \ |
|---|--------------------|-------------------------|---|
| 0 | 0.0                | 0.0                     |   |
| 1 | 0.0                | 0.0                     |   |
| 2 | 0.0                | 0.0                     |   |
| 3 | 0.0                | 0.0                     |   |
| 4 | 0.0                | 0.0                     |   |

|        |     |     |
|--------|-----|-----|
| ...    | ... | ... |
| 212608 | 0.0 | 0.0 |
| 212609 | 0.0 | 0.0 |
| 212610 | 0.0 | 0.0 |
| 212611 | 0.0 | 0.0 |
| 212612 | 0.0 | 0.0 |

|        | first_browser_SlimBrowser | first_browser_Sogou Explorer \ |
|--------|---------------------------|--------------------------------|
| 0      | 0.0                       | 0.0                            |
| 1      | 0.0                       | 0.0                            |
| 2      | 0.0                       | 0.0                            |
| 3      | 0.0                       | 0.0                            |
| 4      | 0.0                       | 0.0                            |
| ...    | ...                       | ...                            |
| 212608 | 0.0                       | 0.0                            |
| 212609 | 0.0                       | 0.0                            |
| 212610 | 0.0                       | 0.0                            |
| 212611 | 0.0                       | 0.0                            |
| 212612 | 0.0                       | 0.0                            |

|        | first_browser_Stainless | first_browser_TenFourFox \ |
|--------|-------------------------|----------------------------|
| 0      | 0.0                     | 0.0                        |
| 1      | 0.0                     | 0.0                        |
| 2      | 0.0                     | 0.0                        |
| 3      | 0.0                     | 0.0                        |
| 4      | 0.0                     | 0.0                        |
| ...    | ...                     | ...                        |
| 212608 | 0.0                     | 0.0                        |
| 212609 | 0.0                     | 0.0                        |
| 212610 | 0.0                     | 0.0                        |
| 212611 | 0.0                     | 0.0                        |
| 212612 | 0.0                     | 0.0                        |

|        | first_browser_TheWorld Browser | first_browser_Yandex.Browser \ |
|--------|--------------------------------|--------------------------------|
| 0      | 0.0                            | 0.0                            |
| 1      | 0.0                            | 0.0                            |
| 2      | 0.0                            | 0.0                            |
| 3      | 0.0                            | 0.0                            |
| 4      | 0.0                            | 0.0                            |
| ...    | ...                            | ...                            |
| 212608 | 0.0                            | 0.0                            |
| 212609 | 0.0                            | 0.0                            |
| 212610 | 0.0                            | 0.0                            |
| 212611 | 0.0                            | 0.0                            |
| 212612 | 0.0                            | 0.0                            |

```

first_browser_w0SBrowser
0      0.0
1      0.0
2      0.0
3      0.0
4      0.0
...
212608 0.0
212609 0.0
212610 0.0
212611 0.0
212612 0.0

```

[212613 rows x 130 columns]

```

[4.49694986e-01 2.94196498e-01 2.54796273e-01 1.31224337e-03
 7.15878145e-01 2.81553809e-01 2.56804617e-03 2.35169063e-05
 1.45804819e-04 2.72796113e-04 3.44287508e-03 1.12881150e-04
 9.66493112e-01 4.28948371e-03 6.58473377e-05 5.50295608e-03
 9.40676252e-06 7.99574814e-05 1.03474388e-04 2.35169063e-05
 2.41283459e-03 1.05826078e-03 3.50401904e-03 4.51524601e-04
 1.41101438e-04 2.53982588e-04 1.12410812e-03 1.82491193e-03
 5.73812514e-04 1.12881150e-04 3.01016401e-04 7.67591822e-03
 3.83513708e-02 1.85266188e-02 6.45350943e-01 4.18742034e-02
 5.12668557e-03 1.21986896e-01 8.82260257e-02 4.05572566e-02
 1.36398057e-04 1.09118445e-02 1.61655214e-02 4.70338126e-06
 6.43954039e-01 7.80761289e-04 1.06672687e-02 2.56334279e-03
 2.42078330e-01 2.12592833e-03 1.63207330e-03 2.44575826e-04
 5.88110793e-02 3.60279005e-03 3.87558616e-03 3.76270501e-05
 2.32817372e-03 7.99574814e-05 2.16755325e-01 1.59914963e-04
 6.53769995e-04 2.06163311e-01 7.26672405e-03 2.87988035e-02
 5.40202151e-01 2.56099110e-02 2.93443957e-02 8.55676746e-01
 8.93689473e-02 1.31130270e-02 6.04384492e-03 5.62524399e-03
 4.19809701e-01 4.98699515e-02 3.52753595e-04 3.40595354e-01
 6.71689878e-02 9.74211361e-02 1.27833199e-01 1.12881150e-03
 3.96965378e-03 1.69321725e-04 4.70338126e-06 1.88135250e-05
 2.44575826e-04 4.23304313e-05 2.99233819e-01 5.91685363e-03
 3.43346832e-04 5.17371939e-05 9.40676252e-06 4.70338126e-06
 2.82202876e-05 9.40676252e-06 4.70338126e-06 1.57629120e-01
 9.40676252e-06 4.70338126e-06 4.70338126e-06 9.86487186e-02
 1.69321725e-04 4.70338126e-06 6.11439564e-05 7.99574814e-05
 4.70338126e-06 2.16355538e-04 1.41101438e-04 9.02814033e-02
 1.41101438e-05 4.70338126e-06 9.40676252e-06 8.79532296e-04
 1.88135250e-05 9.40676252e-06 4.70338126e-06 4.70338126e-06
 5.64405751e-05 4.70338126e-06 1.12881150e-04 2.11567496e-01

```

```

5.17371939e-05 5.83219276e-04 1.12881150e-04 9.40676252e-06
1.55211582e-04 4.70338126e-06 3.76270501e-05 9.40676252e-06
5.17371939e-05 2.82202876e-05]
array([[1.491, 0.    , 0.    , ..., 0.    , 0.    , 0.    ],
       [0.    , 0.    , 1.981, ..., 0.    , 0.    , 0.    ],
       [0.    , 1.844, 0.    , ..., 0.    , 0.    , 0.    ],
       ...,
       [1.491, 0.    , 0.    , ..., 0.    , 0.    , 0.    ],
       [1.491, 0.    , 0.    , ..., 0.    , 0.    , 0.    ],
       [1.491, 0.    , 0.    , ..., 0.    , 0.    , 0.    ]])

```



|            | gender_—unknown— | gender_FEMALE | gender_MALE | gender_OTHER | \ |
|------------|------------------|---------------|-------------|--------------|---|
| id         |                  |               |             |              |   |
| gxn3p5htnn | 1.491217         | 0.000000      | 0.000000    | 0.0          |   |
| 820tgsjxq7 | 0.000000         | 0.000000      | 1.981087    | 0.0          |   |
| 4ft3gnwmtx | 0.000000         | 1.843662      | 0.000000    | 0.0          |   |
| bjjt8pjhuk | 0.000000         | 1.843662      | 0.000000    | 0.0          |   |
| 87mebub9p4 | 1.491217         | 0.000000      | 0.000000    | 0.0          |   |
| ...        | ...              | ...           | ...         | ...          |   |
| zxodksqpep | 0.000000         | 0.000000      | 1.981087    | 0.0          |   |
| mhewnxesx9 | 1.491217         | 0.000000      | 0.000000    | 0.0          |   |
| 6o3arsjbb4 | 1.491217         | 0.000000      | 0.000000    | 0.0          |   |
| jh95kwisub | 1.491217         | 0.000000      | 0.000000    | 0.0          |   |
| nw9fwlyb5f | 1.491217         | 0.000000      | 0.000000    | 0.0          |   |

|            | signup_method_basic | signup_method_facebook | signup_method_google | \ |
|------------|---------------------|------------------------|----------------------|---|
| id         |                     |                        |                      |   |
| gxn3p5htnn | 0.000000            | 1.8846                 | 0.0                  |   |
| 820tgsjxq7 | 0.000000            | 1.8846                 | 0.0                  |   |
| 4ft3gnwmtx | 1.181899            | 0.0000                 | 0.0                  |   |
| bjjt8pjhuk | 0.000000            | 1.8846                 | 0.0                  |   |
| 87mebub9p4 | 1.181899            | 0.0000                 | 0.0                  |   |
| ...        | ...                 | ...                    | ...                  |   |
| zxodksqpep | 1.181899            | 0.0000                 | 0.0                  |   |
| mhewnxesx9 | 1.181899            | 0.0000                 | 0.0                  |   |
| 6o3arsjbb4 | 1.181899            | 0.0000                 | 0.0                  |   |
| jh95kwisub | 1.181899            | 0.0000                 | 0.0                  |   |
| nw9fwlyb5f | 1.181899            | 0.0000                 | 0.0                  |   |

|            | language_ca | language_cs | language_da | ... | \ |
|------------|-------------|-------------|-------------|-----|---|
| id         |             |             |             | ... |   |
| gxn3p5htnn | 0.0         | 0.0         | 0.0         | ... |   |
| 820tgsjxq7 | 0.0         | 0.0         | 0.0         | ... |   |
| 4ft3gnwmtx | 0.0         | 0.0         | 0.0         | ... |   |
| bjjt8pjhuk | 0.0         | 0.0         | 0.0         | ... |   |
| 87mebub9p4 | 0.0         | 0.0         | 0.0         | ... |   |
| ...        | ...         | ...         | ...         | ... |   |
| zxodksqpep | 0.0         | 0.0         | 0.0         | ... |   |
| mhewnxesx9 | 0.0         | 0.0         | 0.0         | ... |   |
| 6o3arsjbb4 | 0.0         | 0.0         | 0.0         | ... |   |
| jh95kwisub | 0.0         | 0.0         | 0.0         | ... |   |
| nw9fwlyb5f | 0.0         | 0.0         | 0.0         | ... |   |

|            | first_browser_SeaMonkey | first_browser_Silk | \ |
|------------|-------------------------|--------------------|---|
| id         |                         |                    |   |
| gxn3p5htnn | 0.0                     | 0.0                |   |

|            |     |     |
|------------|-----|-----|
| 820tgsjq7  | 0.0 | 0.0 |
| 4ft3gnwmtx | 0.0 | 0.0 |
| bjjt8pjhuk | 0.0 | 0.0 |
| 87mebub9p4 | 0.0 | 0.0 |
| ...        | ... | ... |
| zxodksqpep | 0.0 | 0.0 |
| mhewnxesx9 | 0.0 | 0.0 |
| 6o3arsjbb4 | 0.0 | 0.0 |
| jh95kwisub | 0.0 | 0.0 |
| nw9fwlyb5f | 0.0 | 0.0 |

|            | first_browser_SiteKiosk | first_browser_SlimBrowser \ |
|------------|-------------------------|-----------------------------|
| id         |                         |                             |
| gxn3p5htnn | 0.0                     | 0.0                         |
| 820tgsjq7  | 0.0                     | 0.0                         |
| 4ft3gnwmtx | 0.0                     | 0.0                         |
| bjjt8pjhuk | 0.0                     | 0.0                         |
| 87mebub9p4 | 0.0                     | 0.0                         |
| ...        | ...                     | ...                         |
| zxodksqpep | 0.0                     | 0.0                         |
| mhewnxesx9 | 0.0                     | 0.0                         |
| 6o3arsjbb4 | 0.0                     | 0.0                         |
| jh95kwisub | 0.0                     | 0.0                         |
| nw9fwlyb5f | 0.0                     | 0.0                         |

|            | first_browser_Sogou Explorer | first_browser_Stainless \ |
|------------|------------------------------|---------------------------|
| id         |                              |                           |
| gxn3p5htnn | 0.0                          | 0.0                       |
| 820tgsjq7  | 0.0                          | 0.0                       |
| 4ft3gnwmtx | 0.0                          | 0.0                       |
| bjjt8pjhuk | 0.0                          | 0.0                       |
| 87mebub9p4 | 0.0                          | 0.0                       |
| ...        | ...                          | ...                       |
| zxodksqpep | 0.0                          | 0.0                       |
| mhewnxesx9 | 0.0                          | 0.0                       |
| 6o3arsjbb4 | 0.0                          | 0.0                       |
| jh95kwisub | 0.0                          | 0.0                       |
| nw9fwlyb5f | 0.0                          | 0.0                       |

|            | first_browser_TenFourFox | first_browser_TheWorld Browser \ |
|------------|--------------------------|----------------------------------|
| id         |                          |                                  |
| gxn3p5htnn | 0.0                      | 0.0                              |
| 820tgsjq7  | 0.0                      | 0.0                              |
| 4ft3gnwmtx | 0.0                      | 0.0                              |
| bjjt8pjhuk | 0.0                      | 0.0                              |

|            |     |     |
|------------|-----|-----|
| 87mebub9p4 | 0.0 | 0.0 |
| ...        | ... | ... |
| zxodksqpep | 0.0 | 0.0 |
| mhewnxesx9 | 0.0 | 0.0 |
| 6o3arsjbb4 | 0.0 | 0.0 |
| jh95kwisub | 0.0 | 0.0 |
| nw9fwlyb5f | 0.0 | 0.0 |

|            | first_browser_Yandex.Browser | first_browser_w0SBrowser |
|------------|------------------------------|--------------------------|
| id         |                              |                          |
| gxn3p5htnn | 0.0                          | 0.0                      |
| 820tgsjq7  | 0.0                          | 0.0                      |
| 4ft3gnwmtx | 0.0                          | 0.0                      |
| bjjt8pjhuk | 0.0                          | 0.0                      |
| 87mebub9p4 | 0.0                          | 0.0                      |
| ...        | ...                          | ...                      |
| zxodksqpep | 0.0                          | 0.0                      |
| mhewnxesx9 | 0.0                          | 0.0                      |
| 6o3arsjbb4 | 0.0                          | 0.0                      |
| jh95kwisub | 0.0                          | 0.0                      |
| nw9fwlyb5f | 0.0                          | 0.0                      |

[212613 rows x 130 columns]

## Normalisation des variables quantitatives : centrage et réduction

Ici, nous appliquons une standardisation afin de centrer (soustraire la moyenne) et réduire (diviser par l'écart-type) nos données numériques pour garantir une échelle uniforme lors de leur utilisation dans les modèles.

```
In [40]: df_scaled = (df_quantitatives.values - np.mean(df_quantitatives.values, axis=0)) / np.std(df_quantitatives.values,
df_scaled = pd.DataFrame(df_scaled, columns=df_quantitatives.columns, index=df.index)
display(df_scaled)
```

|            | age       | signup_flow |
|------------|-----------|-------------|
| id         |           |             |
| gxn3p5htnn | 0.616998  | -0.428185   |
| 820tgsjxq7 | -0.330476 | -0.428185   |
| 4ft3gnwmtx | 1.090735  | -0.035788   |
| bjlt8pjhuk | -0.014652 | -0.428185   |
| 87mebub9p4 | -0.093608 | -0.428185   |
| ...        | ...       | ...         |
| zxodksqpep | -0.804213 | -0.428185   |
| mhewnxesx9 | 0.616998  | -0.428185   |
| 6o3arsjbb4 | -0.804213 | -0.428185   |
| jh95kwisub | -0.330476 | 2.841790    |
| nw9fwlyb5f | -0.330476 | 2.841790    |

212613 rows × 2 columns

```
In [41]: train = pd.concat([df_dates, df_scaled, df_encoded], axis=1)
```

## Standardisation de la dataset Test

```
In [42]: df_qualitatives = df_test.select_dtypes(include=['object'])
df_quantitatives = df_test.select_dtypes(include=[np.number])
df_dates = df_test.select_dtypes(include=['datetime64'])

display(df_qualitatives)
display(df_quantitatives)
display(df_dates)
```

| id          | gender        | signup_method | language | affiliate_channel | affiliate_provider | first_affiliate_tracked | signup_app | first_device |
|-------------|---------------|---------------|----------|-------------------|--------------------|-------------------------|------------|--------------|
| 5uwns89zht  | FEMALE        | facebook      | en       | direct            | direct             | untracked               | Moweb      |              |
| jtl0dijy2j  | -<br>unknown- | basic         | en       | direct            | direct             | untracked               | Moweb      |              |
| xx0ulgorjt  | -<br>unknown- | basic         | en       | direct            | direct             | linked                  | Web        | Windows D    |
| 6c6puo6ix0  | -<br>unknown- | basic         | en       | direct            | direct             | linked                  | Web        | Windows D    |
| czqghjk3yfe | -<br>unknown- | basic         | en       | direct            | direct             | untracked               | Web        | Mac D        |
| ...         | ...           | ...           | ...      | ...               | ...                | ...                     | ...        |              |
| cv0na2lf5a  | -<br>unknown- | basic         | en       | direct            | direct             | untracked               | Web        | Windows D    |
| zp8xfonng8  | -<br>unknown- | basic         | ko       | direct            | direct             | untracked               | Android    | Android      |
| fa6260ziny  | -<br>unknown- | basic         | de       | direct            | direct             | linked                  | Web        | Windows D    |
| 87k0fy4ugm  | -<br>unknown- | basic         | en       | sem-brand         | google             | omg                     | Web        | Mac D        |
| 9uqfg8txu3  | FEMALE        | basic         | en       | other             | other              | tracked-other           | Web        | Windows D    |

62045 rows × 9 columns

|            | age | signup_flow |
|------------|-----|-------------|
| id         |     |             |
| 5uwns89zht | 35  | 0           |
| jtl0dijy2j | 36  | 0           |
| xx0ulgorjt | 36  | 0           |
| 6c6puo6ix0 | 36  | 0           |
| czqhjk3yfe | 36  | 0           |
| ...        | ... | ...         |
| cv0na2lf5a | 31  | 0           |
| zp8xfonng8 | 61  | 23          |
| fa6260ziny | 36  | 0           |
| 87k0fy4ugm | 36  | 0           |
| 9uqfg8txu3 | 49  | 0           |

62045 rows × 2 columns

| id         | date_account_created | timestamp_first_active |
|------------|----------------------|------------------------|
| 5uwns89zht | 2014-07-01           | 2014-07-01 00:00:06    |
| jtl0dijy2j | 2014-07-01           | 2014-07-01 00:00:51    |
| xx0ulgorjt | 2014-07-01           | 2014-07-01 00:01:48    |
| 6c6puo6ix0 | 2014-07-01           | 2014-07-01 00:02:15    |
| czqhjk3yfe | 2014-07-01           | 2014-07-01 00:03:05    |
| ...        | ...                  | ...                    |
| cv0na2lf5a | 2014-09-30           | 2014-09-30 23:52:32    |
| zp8xfonng8 | 2014-09-30           | 2014-09-30 23:53:06    |
| fa6260ziny | 2014-09-30           | 2014-09-30 23:54:08    |
| 87k0fy4ugm | 2014-09-30           | 2014-09-30 23:54:30    |
| 9uqfg8txu3 | 2014-09-30           | 2014-09-30 23:59:01    |

62045 rows × 2 columns

## Encodage des variables catégorielles de la dataset test

Dans cette étape, nous appliquons le OneHotEncoder afin de convertir les données catégorielles en format numérique. Cette transformation crée des indicateurs binaires pour chaque modalité, rendant ainsi les variables compatibles avec les algorithmes d'apprentissage automatique.

```
In [43]: colonnes_qualitatives = df_qualitatives.select_dtypes(include=['object']).columns

encoder = OneHotEncoder()
encoded_data = encoder.fit_transform(df_qualitatives[colonnes_qualitatives]).toarray()

df_encoded = pd.DataFrame(encoded_data, columns=encoder.get_feature_names_out(colonnes_qualitatives))
print(df_encoded)

p_k = np.mean(df_encoded.values, axis=0)
print(p_k)
```

```
ZD = df_encoded.values/np.sqrt(p_k)
display(ZD.round(3))

df_encoded = pd.DataFrame(ZD, columns=encoder.get_feature_names_out(colonnes_qualitatives), index=df_test.index)
print(df_encoded)
```



|       | gender_-unknown- | gender_FEMALE | gender_MALE | gender_OTHER | \ |
|-------|------------------|---------------|-------------|--------------|---|
| 0     | 0.0              | 1.0           | 0.0         | 0.0          |   |
| 1     | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 2     | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 3     | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 4     | 1.0              | 0.0           | 0.0         | 0.0          |   |
| ...   | ...              | ...           | ...         | ...          |   |
| 62040 | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 62041 | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 62042 | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 62043 | 1.0              | 0.0           | 0.0         | 0.0          |   |
| 62044 | 0.0              | 1.0           | 0.0         | 0.0          |   |

|       | signup_method_basic | signup_method_facebook | signup_method_google | \ |
|-------|---------------------|------------------------|----------------------|---|
| 0     | 0.0                 | 1.0                    | 0.0                  |   |
| 1     | 1.0                 | 0.0                    | 0.0                  |   |
| 2     | 1.0                 | 0.0                    | 0.0                  |   |
| 3     | 1.0                 | 0.0                    | 0.0                  |   |
| 4     | 1.0                 | 0.0                    | 0.0                  |   |
| ...   | ...                 | ...                    | ...                  |   |
| 62040 | 1.0                 | 0.0                    | 0.0                  |   |
| 62041 | 1.0                 | 0.0                    | 0.0                  |   |
| 62042 | 1.0                 | 0.0                    | 0.0                  |   |
| 62043 | 1.0                 | 0.0                    | 0.0                  |   |
| 62044 | 1.0                 | 0.0                    | 0.0                  |   |

|       | signup_method_weibo | language_-unknown- | language_ca | ... | \ |
|-------|---------------------|--------------------|-------------|-----|---|
| 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         | ... |   |
| ...   | ...                 | ...                | ...         | ... |   |
| 62040 | 0.0                 | 0.0                | 0.0         | ... |   |
| 62041 | 0.0                 | 0.0                | 0.0         | ... |   |
| 62042 | 0.0                 | 0.0                | 0.0         | ... |   |
| 62043 | 0.0                 | 0.0                | 0.0         | ... |   |
| 62044 | 0.0                 | 0.0                | 0.0         | ... |   |

|   | first_browser_Opera Mobile | first_browser_Pale Moon | \ |
|---|----------------------------|-------------------------|---|
| 0 | 0.0                        | 0.0                     |   |
| 1 | 0.0                        | 0.0                     |   |
| 2 | 0.0                        | 0.0                     |   |
| 3 | 0.0                        | 0.0                     |   |
| 4 | 0.0                        | 0.0                     |   |

```

...
62040      0.0      0.0
62041      0.0      0.0
62042      0.0      0.0
62043      0.0      0.0
62044      0.0      0.0

first_browser_Safari first_browser_SeaMonkey first_browser_Silk \
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      1.0      0.0      0.0
...
62040      0.0      0.0      0.0
62041      0.0      0.0      0.0
62042      0.0      0.0      0.0
62043      1.0      0.0      0.0
62044      0.0      0.0      0.0

first_browser_SiteKiosk first_browser_Sogou Explorer \
0      0.0      0.0
1      0.0      0.0
2      0.0      0.0
3      0.0      0.0
4      0.0      0.0
...
62040      0.0      0.0
62041      0.0      0.0
62042      0.0      0.0
62043      0.0      0.0
62044      0.0      0.0

first_browser_UC Browser first_browser_Yandex.Browser \
0      0.0      0.0
1      0.0      0.0
2      0.0      0.0
3      0.0      0.0
4      0.0      0.0
...
62040      0.0      0.0
62041      0.0      0.0
62042      0.0      0.0
62043      0.0      0.0
62044      0.0      0.0

```

```

first_browser_w0SBrowser
0          0.0
1          0.0
2          0.0
3          0.0
4          0.0
...
62040      0.0
62041      0.0
62042      0.0
62043      0.0
62044      0.0

```

[62045 rows x 107 columns]

```

[5.44588605e-01 2.33056652e-01 2.21516641e-01 8.38101378e-04
 7.29937948e-01 2.39422999e-01 3.02683536e-02 3.70698686e-04
 1.61173342e-05 1.61173342e-05 2.73994681e-04 2.73994681e-04
 3.94874688e-03 9.67040052e-05 9.53743251e-01 4.17438956e-03
 9.67040052e-05 5.41542429e-03 1.12821339e-04 1.61173342e-05
 1.91796277e-03 1.93408010e-03 5.91506165e-03 5.96341365e-04
 3.38464018e-04 3.38464018e-04 1.32162140e-03 1.91796277e-03
 8.70336046e-04 6.44693368e-05 4.51285357e-04 1.61495689e-02
 2.73994681e-03 7.05890886e-01 9.42864050e-03 2.77218148e-03
 1.67507454e-01 1.98243211e-02 9.18365702e-02 4.83520026e-05
 2.24192119e-02 6.44693368e-05 3.22346684e-05 7.05890886e-01
 1.67620276e-03 2.77701668e-02 3.38464018e-04 2.29833186e-01
 3.22346684e-05 1.77290676e-04 2.25642679e-04 7.84914175e-03
 1.07986139e-03 1.61173342e-05 2.53042147e-03 1.61173342e-05
 2.54234830e-01 5.64106697e-04 2.28866146e-03 1.75276009e-01
 1.28455154e-02 8.02643243e-03 5.46764445e-01 8.10057217e-02
 6.85953743e-02 5.99419776e-01 2.50979128e-01 1.06583931e-01
 1.29905714e-02 4.96413893e-03 2.69610766e-01 8.02643243e-03
 1.85349343e-03 2.29301314e-01 5.95857845e-02 3.07083568e-01
 2.75445241e-01 1.45056008e-04 1.16850673e-02 1.45056008e-04
 5.80224031e-04 2.38875010e-01 3.08002256e-02 1.61173342e-04
 1.61173342e-05 8.07478443e-02 1.61173342e-05 5.92312032e-02
 1.32162140e-03 1.61173342e-05 1.12821339e-04 2.25642679e-04
 5.47989363e-04 1.66991700e-01 1.61173342e-05 6.44693368e-04
 6.44693368e-05 3.22346684e-05 1.61173342e-05 1.31082279e-01
 1.61173342e-05 7.73632041e-04 4.83520026e-05 1.61173342e-04
 1.61173342e-05 4.83520026e-05 1.61173342e-05]

```

```
array([[0.      , 2.071, 0.      , ..., 0.      , 0.      , 0.      ],
       [1.355, 0.      , 0.      , ..., 0.      , 0.      , 0.      ],
       [1.355, 0.      , 0.      , ..., 0.      , 0.      , 0.      ],
       ...,
       [1.355, 0.      , 0.      , ..., 0.      , 0.      , 0.      ],
       [1.355, 0.      , 0.      , ..., 0.      , 0.      , 0.      ],
       [0.      , 2.071, 0.      , ..., 0.      , 0.      , 0.      ]])
```

|            | gender_—unknown— | gender_FEMALE | gender_MALE | gender_OTHER | \ |
|------------|------------------|---------------|-------------|--------------|---|
| id         |                  |               |             |              |   |
| 5uwns89zht | 0.000000         | 2.071425      | 0.0         | 0.0          |   |
| jtl0dijy2j | 1.355082         | 0.000000      | 0.0         | 0.0          |   |
| xx0ulgorjt | 1.355082         | 0.000000      | 0.0         | 0.0          |   |
| 6c6puo6ix0 | 1.355082         | 0.000000      | 0.0         | 0.0          |   |
| czqhjk3yfe | 1.355082         | 0.000000      | 0.0         | 0.0          |   |
| ...        | ...              | ...           | ...         | ...          |   |
| cv0na2lf5a | 1.355082         | 0.000000      | 0.0         | 0.0          |   |
| zp8xfonng8 | 1.355082         | 0.000000      | 0.0         | 0.0          |   |
| fa6260ziny | 1.355082         | 0.000000      | 0.0         | 0.0          |   |
| 87k0fy4ugm | 1.355082         | 0.000000      | 0.0         | 0.0          |   |
| 9uqfg8txu3 | 0.000000         | 2.071425      | 0.0         | 0.0          |   |

|            | signup_method_basic | signup_method_facebook | signup_method_google | \ |
|------------|---------------------|------------------------|----------------------|---|
| id         |                     |                        |                      |   |
| 5uwns89zht | 0.000000            | 2.0437                 | 0.0                  |   |
| jtl0dijy2j | 1.170461            | 0.0000                 | 0.0                  |   |
| xx0ulgorjt | 1.170461            | 0.0000                 | 0.0                  |   |
| 6c6puo6ix0 | 1.170461            | 0.0000                 | 0.0                  |   |
| czqhjk3yfe | 1.170461            | 0.0000                 | 0.0                  |   |
| ...        | ...                 | ...                    | ...                  |   |
| cv0na2lf5a | 1.170461            | 0.0000                 | 0.0                  |   |
| zp8xfonng8 | 1.170461            | 0.0000                 | 0.0                  |   |
| fa6260ziny | 1.170461            | 0.0000                 | 0.0                  |   |
| 87k0fy4ugm | 1.170461            | 0.0000                 | 0.0                  |   |
| 9uqfg8txu3 | 1.170461            | 0.0000                 | 0.0                  |   |

|            | signup_method_weibo | language_—unknown— | language_ca | ... | \ |
|------------|---------------------|--------------------|-------------|-----|---|
| id         |                     |                    |             | ... |   |
| 5uwns89zht | 0.0                 | 0.0                | 0.0         | ... |   |
| jtl0dijy2j | 0.0                 | 0.0                | 0.0         | ... |   |
| xx0ulgorjt | 0.0                 | 0.0                | 0.0         | ... |   |
| 6c6puo6ix0 | 0.0                 | 0.0                | 0.0         | ... |   |
| czqhjk3yfe | 0.0                 | 0.0                | 0.0         | ... |   |
| ...        | ...                 | ...                | ...         | ... |   |
| cv0na2lf5a | 0.0                 | 0.0                | 0.0         | ... |   |
| zp8xfonng8 | 0.0                 | 0.0                | 0.0         | ... |   |
| fa6260ziny | 0.0                 | 0.0                | 0.0         | ... |   |
| 87k0fy4ugm | 0.0                 | 0.0                | 0.0         | ... |   |
| 9uqfg8txu3 | 0.0                 | 0.0                | 0.0         | ... |   |

|            | first_browser_Opera Mobile | first_browser_Pale Moon | \ |
|------------|----------------------------|-------------------------|---|
| id         |                            |                         |   |
| 5uwns89zht | 0.0                        | 0.0                     |   |

|            |     |     |
|------------|-----|-----|
| jtl0dijy2j | 0.0 | 0.0 |
| xx0ulgorjt | 0.0 | 0.0 |
| 6c6puo6ix0 | 0.0 | 0.0 |
| czqhjk3yfe | 0.0 | 0.0 |
| ...        | ... | ... |
| cv0na2lf5a | 0.0 | 0.0 |
| zp8xfonng8 | 0.0 | 0.0 |
| fa6260ziny | 0.0 | 0.0 |
| 87k0fy4ugm | 0.0 | 0.0 |
| 9uqfg8txu3 | 0.0 | 0.0 |

|            | first_browser_Safari | first_browser_SeaMonkey | first_browser_Silk \ |
|------------|----------------------|-------------------------|----------------------|
| id         |                      |                         |                      |
| 5uwns89zht | 0.000000             | 0.0                     | 0.0                  |
| jtl0dijy2j | 0.000000             | 0.0                     | 0.0                  |
| xx0ulgorjt | 0.000000             | 0.0                     | 0.0                  |
| 6c6puo6ix0 | 0.000000             | 0.0                     | 0.0                  |
| czqhjk3yfe | 2.762028             | 0.0                     | 0.0                  |
| ...        | ...                  | ...                     | ...                  |
| cv0na2lf5a | 0.000000             | 0.0                     | 0.0                  |
| zp8xfonng8 | 0.000000             | 0.0                     | 0.0                  |
| fa6260ziny | 0.000000             | 0.0                     | 0.0                  |
| 87k0fy4ugm | 2.762028             | 0.0                     | 0.0                  |
| 9uqfg8txu3 | 0.000000             | 0.0                     | 0.0                  |

|            | first_browser_SiteKiosk | first_browser_Sogou Explorer \ |
|------------|-------------------------|--------------------------------|
| id         |                         |                                |
| 5uwns89zht | 0.0                     | 0.0                            |
| jtl0dijy2j | 0.0                     | 0.0                            |
| xx0ulgorjt | 0.0                     | 0.0                            |
| 6c6puo6ix0 | 0.0                     | 0.0                            |
| czqhjk3yfe | 0.0                     | 0.0                            |
| ...        | ...                     | ...                            |
| cv0na2lf5a | 0.0                     | 0.0                            |
| zp8xfonng8 | 0.0                     | 0.0                            |
| fa6260ziny | 0.0                     | 0.0                            |
| 87k0fy4ugm | 0.0                     | 0.0                            |
| 9uqfg8txu3 | 0.0                     | 0.0                            |

|            | first_browser_UC Browser | first_browser_Yandex.Browser \ |
|------------|--------------------------|--------------------------------|
| id         |                          |                                |
| 5uwns89zht | 0.0                      | 0.0                            |
| jtl0dijy2j | 0.0                      | 0.0                            |
| xx0ulgorjt | 0.0                      | 0.0                            |
| 6c6puo6ix0 | 0.0                      | 0.0                            |

|            |     |     |
|------------|-----|-----|
| czqhjk3yfe | 0.0 | 0.0 |
| ...        | ... | ... |
| cv0na2lf5a | 0.0 | 0.0 |
| zp8xfonng8 | 0.0 | 0.0 |
| fa6260ziny | 0.0 | 0.0 |
| 87k0fy4ugm | 0.0 | 0.0 |
| 9uqfg8txu3 | 0.0 | 0.0 |

|            | first_browser_w0SBrowser |
|------------|--------------------------|
| id         |                          |
| 5uwns89zht | 0.0                      |
| jtl0dijy2j | 0.0                      |
| xx0ulgorjt | 0.0                      |
| 6c6puo6ix0 | 0.0                      |
| czqhjk3yfe | 0.0                      |
| ...        | ...                      |
| cv0na2lf5a | 0.0                      |
| zp8xfonng8 | 0.0                      |
| fa6260ziny | 0.0                      |
| 87k0fy4ugm | 0.0                      |
| 9uqfg8txu3 | 0.0                      |

[62045 rows x 107 columns]

## Normalisation des variables quantitatives : centrage et réduction

Ici, nous appliquons une standardisation afin de centrer (soustraire la moyenne) et réduire (diviser par l'écart-type) nos données numériques pour garantir une échelle uniforme lors de leur utilisation dans les modèles.

```
In [44]: df_scaled = (df_quantitatives.values - np.mean(df_quantitatives.values, axis=0)) / np.std(df_quantitatives.values,
df_scaled = pd.DataFrame(df_scaled, columns=df_quantitatives.columns, index=df_test.index)
display(df_scaled)
```

|             | age       | signup_flow |
|-------------|-----------|-------------|
| id          |           |             |
| 5uwns89zht  | -0.149108 | -0.693777   |
| jtl0dijy2j  | -0.080702 | -0.693777   |
| xx0ulgorjt  | -0.080702 | -0.693777   |
| 6c6puo6ix0  | -0.080702 | -0.693777   |
| czqghjk3yfe | -0.080702 | -0.693777   |
| ...         | ...       | ...         |
| cv0na2lf5a  | -0.422728 | -0.693777   |
| zp8xfonng8  | 1.629426  | 1.350003    |
| fa6260ziny  | -0.080702 | -0.693777   |
| 87k0fy4ugm  | -0.080702 | -0.693777   |
| 9uqfg8txu3  | 0.808564  | -0.693777   |

62045 rows × 2 columns

```
In [45]: test = pd.concat([df_dates, df_scaled, df_encoded], axis=1)
```

```
In [46]: print("Data de training:")
display(train)
```

Data de training:



|            | date_account_created | timestamp_first_active | age       | signup_flow | gender_-<br>unknown- | gender_FEMALE | gender_MALE | gen |
|------------|----------------------|------------------------|-----------|-------------|----------------------|---------------|-------------|-----|
| id         |                      |                        |           |             |                      |               |             |     |
| gxn3p5htnn | 2010-06-28           | 2009-03-19 04:32:55    | 0.616998  | -0.428185   | 1.491217             | 0.000000      | 0.000000    |     |
| 820tgsjxq7 | 2011-05-25           | 2009-05-23 17:48:09    | -0.330476 | -0.428185   | 0.000000             | 0.000000      | 1.981087    |     |
| 4ft3gnwmtx | 2010-09-28           | 2009-06-09 23:12:47    | 1.090735  | -0.035788   | 0.000000             | 1.843662      | 0.000000    |     |
| bjjt8pjhuk | 2011-12-05           | 2009-10-31 06:01:29    | -0.014652 | -0.428185   | 0.000000             | 1.843662      | 0.000000    |     |
| 87mebub9p4 | 2010-09-14           | 2009-12-08 06:11:05    | -0.093608 | -0.428185   | 1.491217             | 0.000000      | 0.000000    |     |
| ...        | ...                  | ...                    | ...       | ...         | ...                  | ...           | ...         | ... |
| zxodksqpep | 2014-06-30           | 2014-06-30 23:56:36    | -0.804213 | -0.428185   | 0.000000             | 0.000000      | 1.981087    |     |
| mhewnxesx9 | 2014-06-30           | 2014-06-30 23:57:19    | 0.616998  | -0.428185   | 1.491217             | 0.000000      | 0.000000    |     |
| 6o3arsjbb4 | 2014-06-30           | 2014-06-30 23:57:54    | -0.804213 | -0.428185   | 1.491217             | 0.000000      | 0.000000    |     |
| jh95kwisub | 2014-06-30           | 2014-06-30 23:58:22    | -0.330476 | 2.841790    | 1.491217             | 0.000000      | 0.000000    |     |
| nw9fwlyb5f | 2014-06-30           | 2014-06-30 23:58:24    | -0.330476 | 2.841790    | 1.491217             | 0.000000      | 0.000000    |     |

212613 rows x 134 columns

```
In [47]: print("Dataset Target:")
display(df_target)
```

Dataset Target:

```
id
gxn3p5htnn      NDF
820tgsjxq7      NDF
4ft3gnwmtx      US
bjjt8pjhuk      other
87mebub9p4      US
...
zxodksqpep      NDF
mhewnxesx9      NDF
6o3arsjbb4      NDF
jh95kwisub      NDF
nw9fwlyb5f      NDF
```

Name: country\_destination, Length: 212613, dtype: object

```
In [48]: print("Dataset de test:")
display(test)
```

Dataset de test:

|             | date_account_created | timestamp_first_active | age       | signup_flow | gender_<br>unknown- | gender_FEMALE | gender_MALE | genc |
|-------------|----------------------|------------------------|-----------|-------------|---------------------|---------------|-------------|------|
| id          |                      |                        |           |             |                     |               |             |      |
| 5uwns89zht  | 2014-07-01           | 2014-07-01 00:00:06    | -0.149108 | -0.693777   | 0.000000            | 2.071425      | 0.0         |      |
| jtl0dijy2j  | 2014-07-01           | 2014-07-01 00:00:51    | -0.080702 | -0.693777   | 1.355082            | 0.000000      | 0.0         |      |
| xx0ulgorjt  | 2014-07-01           | 2014-07-01 00:01:48    | -0.080702 | -0.693777   | 1.355082            | 0.000000      | 0.0         |      |
| 6c6puo6ix0  | 2014-07-01           | 2014-07-01 00:02:15    | -0.080702 | -0.693777   | 1.355082            | 0.000000      | 0.0         |      |
| czqjhjk3yfe | 2014-07-01           | 2014-07-01 00:03:05    | -0.080702 | -0.693777   | 1.355082            | 0.000000      | 0.0         |      |
| ...         | ...                  | ...                    | ...       | ...         | ...                 | ...           | ...         | ...  |
| cv0na2lf5a  | 2014-09-30           | 2014-09-30 23:52:32    | -0.422728 | -0.693777   | 1.355082            | 0.000000      | 0.0         |      |
| zp8xfonng8  | 2014-09-30           | 2014-09-30 23:53:06    | 1.629426  | 1.350003    | 1.355082            | 0.000000      | 0.0         |      |
| fa6260ziny  | 2014-09-30           | 2014-09-30 23:54:08    | -0.080702 | -0.693777   | 1.355082            | 0.000000      | 0.0         |      |
| 87k0fy4ugm  | 2014-09-30           | 2014-09-30 23:54:30    | -0.080702 | -0.693777   | 1.355082            | 0.000000      | 0.0         |      |
| 9uqfg8txu3  | 2014-09-30           | 2014-09-30 23:59:01    | 0.808564  | -0.693777   | 0.000000            | 2.071425      | 0.0         |      |

62045 rows × 111 columns