ETAPE 1 :

I) GRAPHIQUES SUR DONNEES D'ENTRAINEMENT

- X_train.csv

```
   Unnamed: 0                                        designation  ...   productid     imageid
0           0  Olivia: Personalisiertes Notizbuch / 150 Seite...  ...  3804725264  1263597046
1           1  Journal Des Arts (Le) N° 133 Du 28/09/2001 - L...  ...   436067568  1008141237
2           2  Grand Stylet Ergonomique Bleu Gamepad Nintendo...  ...   201115110   938777978
3           3  Peluche Donald - Europe - Disneyland 2000 (Mar...  ...    50418756   457047496
4           4                      La Guerre Des Tuques  ...   278535884  1077757786
5           5  Afrique Contemporaine N° 212 Hiver 2004 - Doss...  ...     5862738   393356830
6           6       Christof E: Bildungsprozessen Auf Der Spur  ...    91920807   907794536
7           7  Conquérant Sept Cahier Couverture Polypro 240 ...  ...   344240059   999581347
8           8        Puzzle Scooby-Doo Avec Poster 2x35 Pieces  ...  4239126071  1325918866
9           9  Tente Pliante V3s5-Pro Pvc Blanc - 3 X 4m50 - ...  ...  3793572222  1245644185

[10 rows x 5 columns]


<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84916 entries, 0 to 84915
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   84916 non-null  int64
 1   designation  84916 non-null  object
 2   description  55116 non-null  object
 3   productid    84916 non-null  int64
 4   imageid      84916 non-null  int64
dtypes: int64(3), object(2)
memory usage: 3.2+ MB
None
```

- y_train.csv

```
   Unnamed: 0  prdtypecode
0           0           10
1           1         2280
2           2           50
3           3         1280
4           4         2705
5           5         2280
6           6           10
7           7         2522
8           8         1280
9           9         2582
```

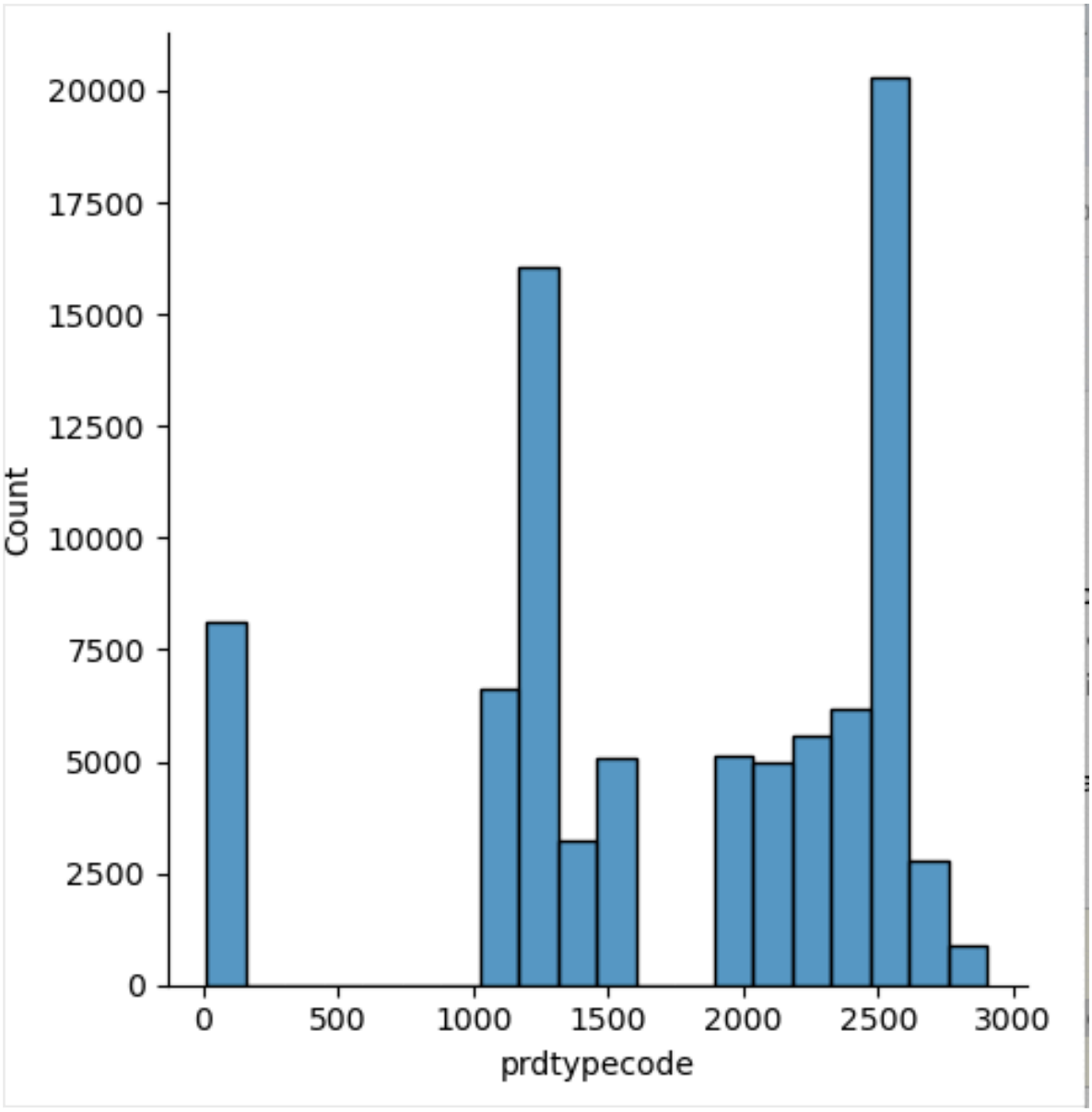
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84916 entries, 0 to 84915
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Unnamed: 0   84916 non-null  int64
 1   prdtypecode  84916 non-null  int64
dtypes: int64(2)
memory usage: 1.3 MB
None
```

PRE-TRAITEMENT :

Merger les données dans un dataframe ‘fusion’.

#1 Histogramme_prdtypecode

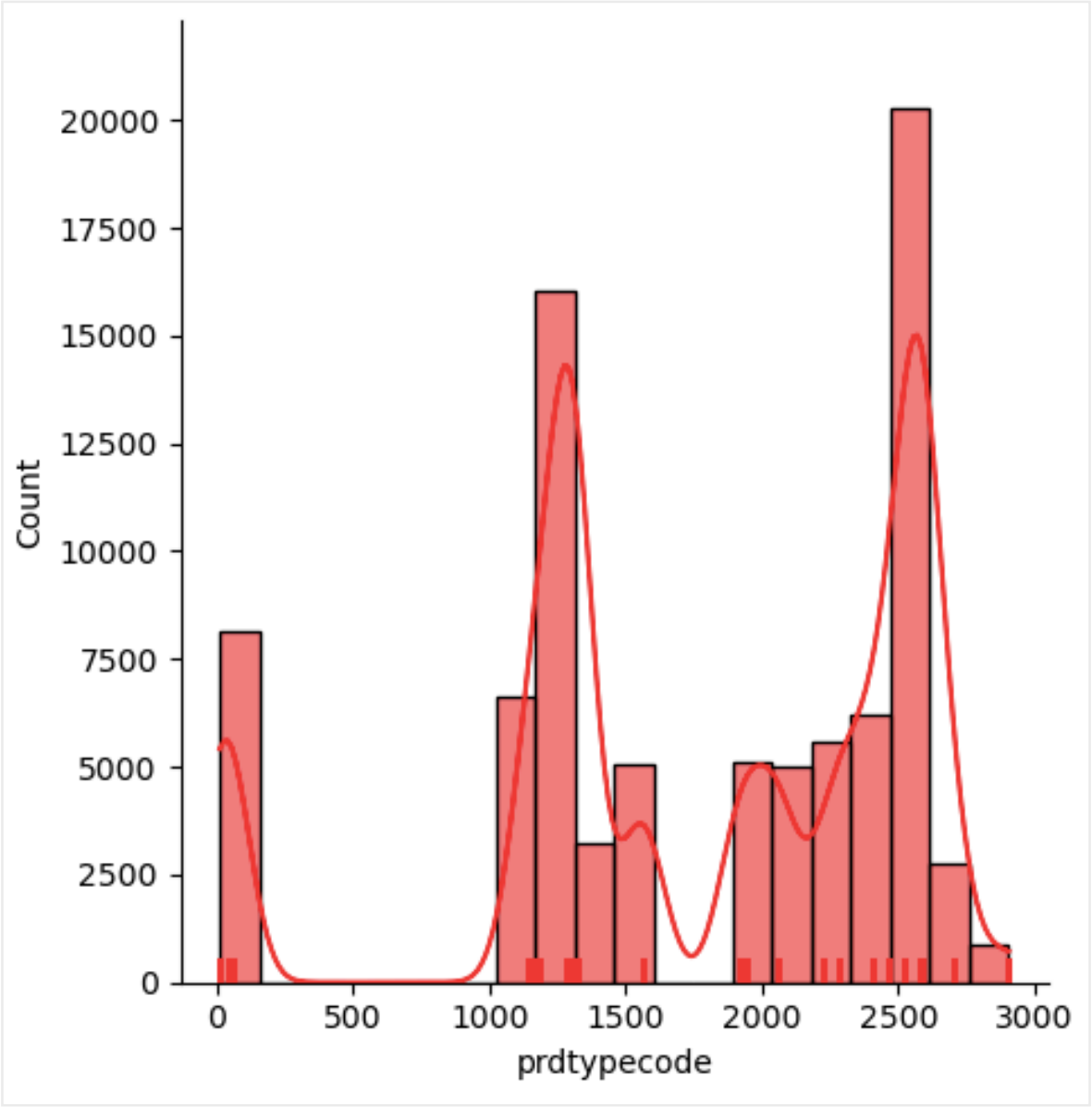
```
sns.displot(fusion.prdtypecode, bins=20)
plt.show()
```



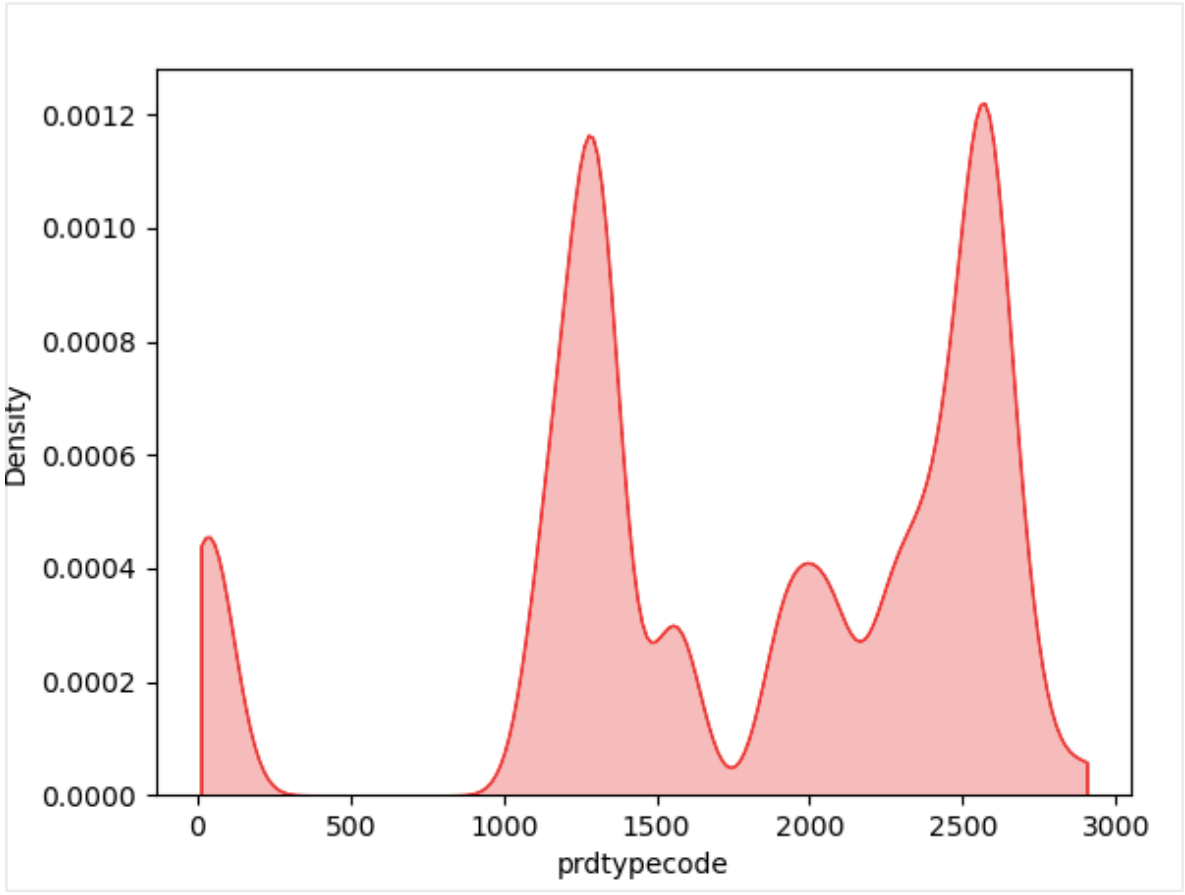
#2 histogramme prdtypecode

sur 20 intervalles et un *rug plot*, de couleur rouge, avec l'estimation de la densité.

```
sns.displot(fusion.prdtypecode, bins=20, kde = True, rug=True, color="red")
plt.show()
```




```
#densité
sns.kdeplot(fusion.prdtypecode, shade=True, color="red", cut=0);
plt.show();
```



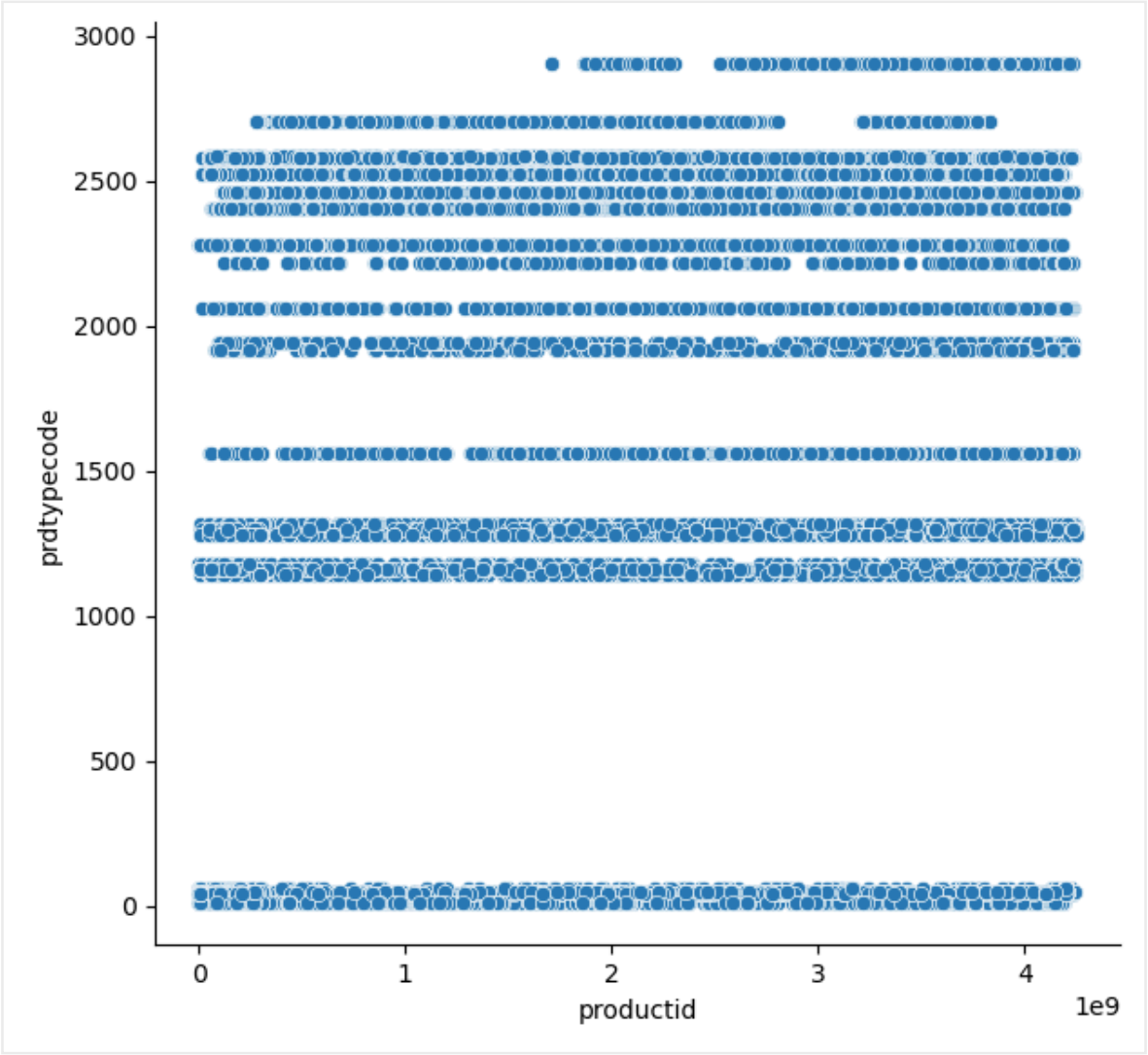
Conclusion : certaines catégories de produit sont plus représentées que d'autres.
Codes en 1300, en 2500.

#3 graphique prdtypecode = f(productid)

```
sns.relplot(x=fusion.productid, y=fusion.prdtypecode);
plt.show();
```

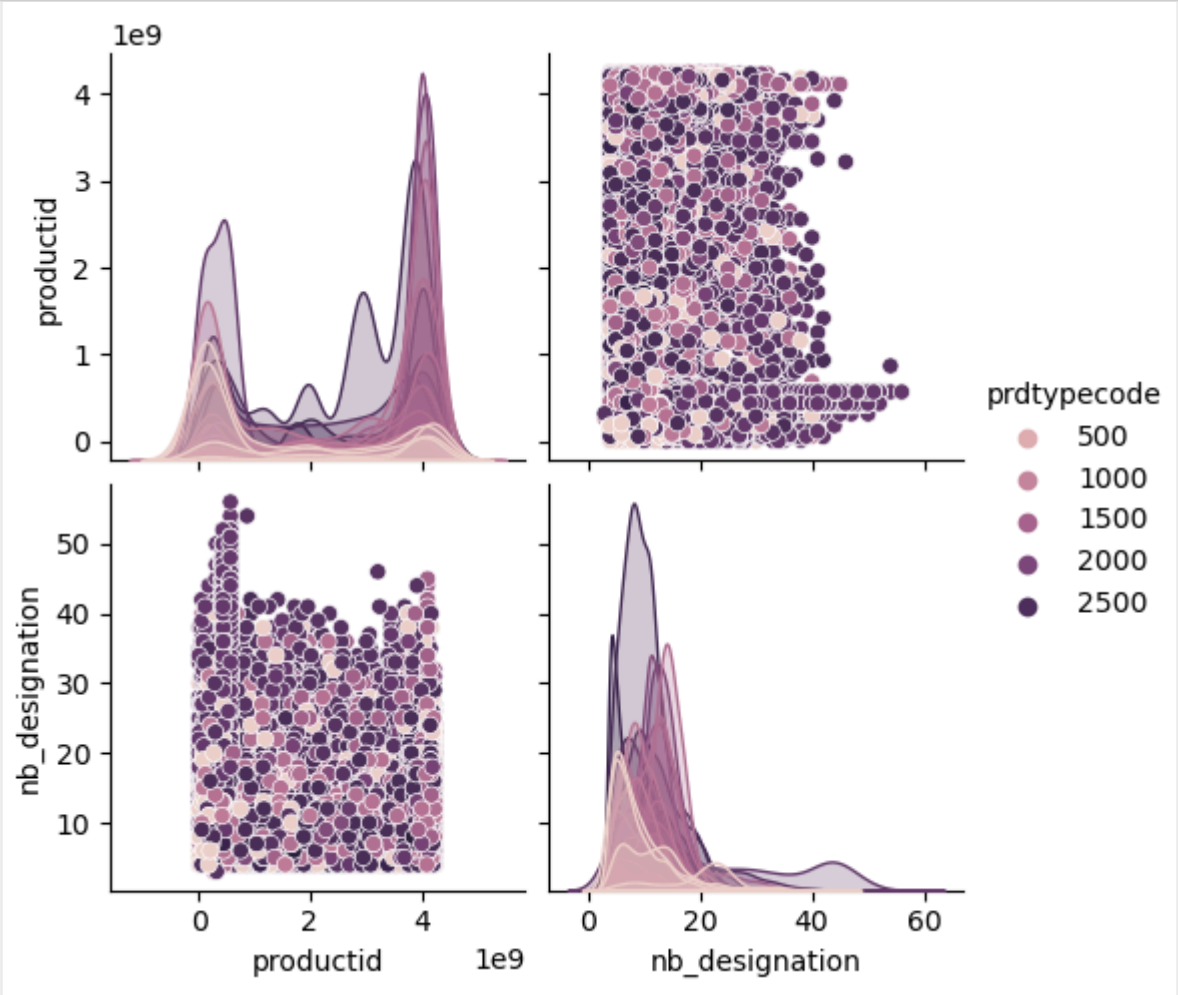
Conclusion :

1. Il y a des répartitions discrètes des produits selon les code type produit ==> Lignes horizontales sur la plage entière des productid
2. Dans la majorité des cas, pour chaque code type produit, les codes produits s'étendent sur les plages entière des productid



#4 graphiques croisés entre variables quantitatives

```
sns.pairplot(fusion[['productid', 'nb_designation', 'prdtypecode']],
hue="prdtypecode", diag_kind="kde");
plt.show();
```



1 2
3 4

Conclusion :
1er : (count(prdtypecode) = f(productid) ??) il y a des distributions similaires de productid selon le code type produit => accumulation en début et fin de numérotations
2ième, 3ième : pas de conclusions ...
4ième : Les désignations de produits se font majoritairement avec 5 à 15 mots toutes catégories confondues.

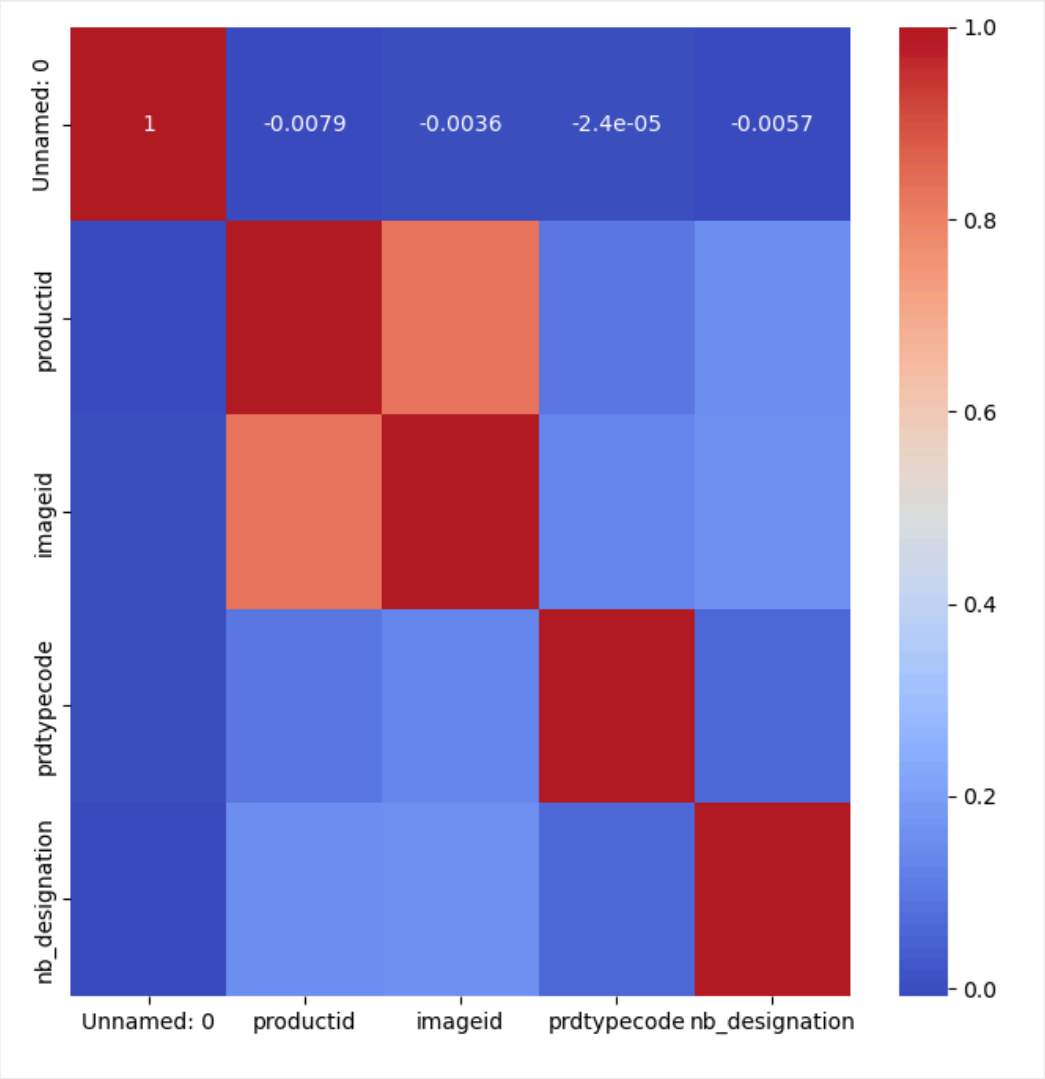
#5 matrice de corrélation heatmap

 # pb avec variables description et designation ??
ValueError: could not convert string to float: 'Olivia: Personalisiertes Notizbuch / 150 Seiten / Punktraster / Ca Din A5 / Rosen-Design'
ValueError: could not convert string to float: 'PILOT STYLE
==> retrait des 2 colonnes

Conclusion : corrélation entre imageid et productid ??

taille de cor 5

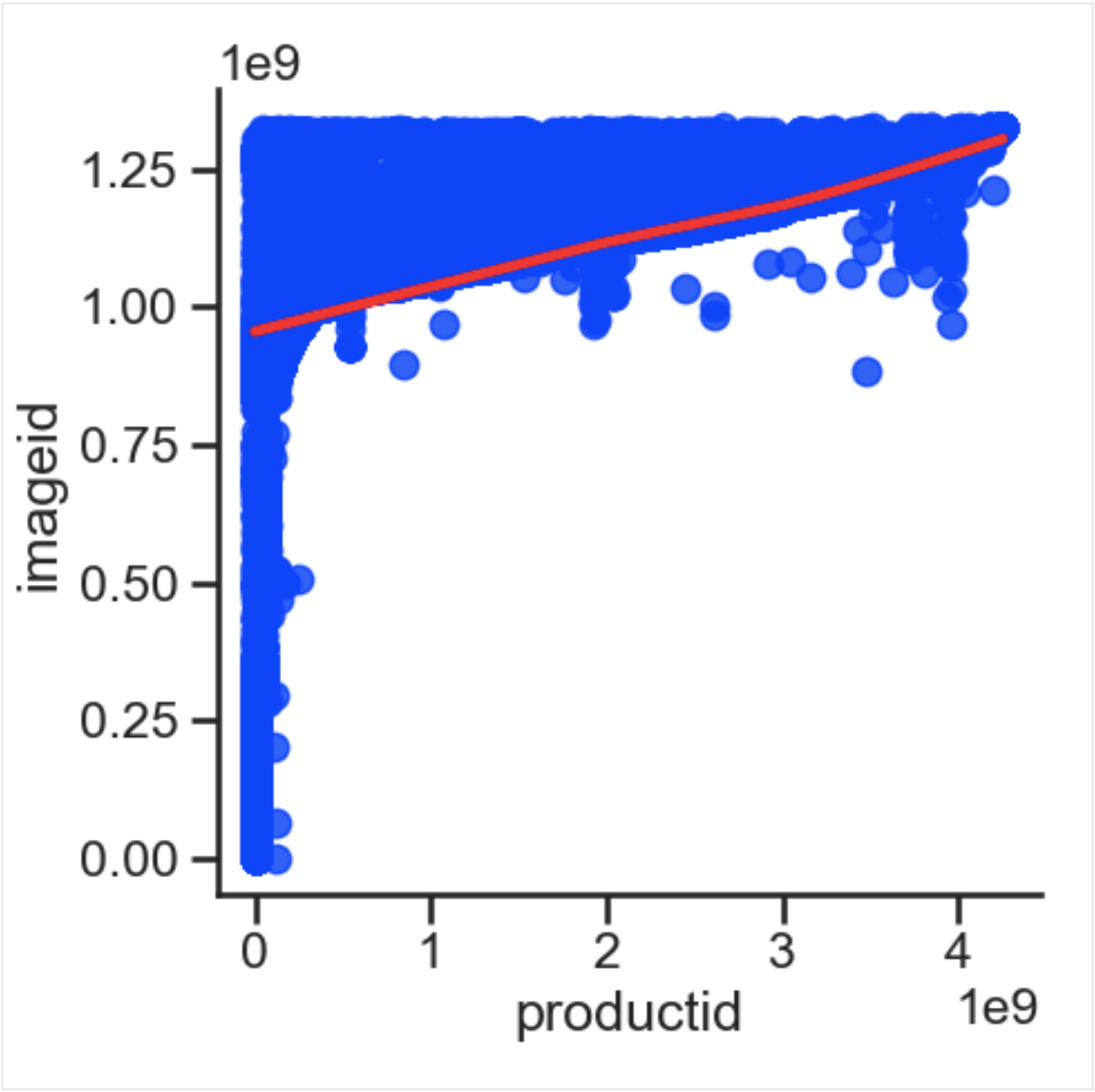
```
df = fusion.drop(['description'], axis=1)
df = df.drop(['designation'], axis=1)

df.info()
cor = df.corr()
print("taille de cor",len(cor))

fig, ax = plt.subplots(figsize = (8,8))
sns.heatmap(
cor,
ax = ax,
cmap = "coolwarm",
annot=True);

plt.show();
```



#6 Etude corrélation variables quantitatives x = "productid",y = "imageid"

ne fonctionne pas ....

```
sns.lmplot(x = "productid",y = "imageid", data = df,
lowess=True,line_kws={'color': 'red'});
plt.show();
```



AUTRES AXES

ETUDES DE LA DISTRIBUTION DU NOMBRE DE MOTS DES CHAMPS designation, description.