PROJET 3 : CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

Damien Dous

SOMMAIRE

PRÉSENTATION DE L'IDÉE

LA BASE DE DONNÉES

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_nam
3	16087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salte Nut M
4	16094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organ Polen
5	16100	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055651	2017-03-09T10:34:11Z	1489055651	2017-03-09T10:34:11Z	Breadsho Honey Gor Nuts Grano
6	16117	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055730	2017-03-09T10:35:30Z	1489055730	2017-03-09T10:35:30Z	Organic Lor Grain Whi Ric
7	16124	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055711	2017-03-09T10:35:11Z	1489055712	2017-03-09T10:35:12Z	Organic Mue
...	...	...	...	...	...	...	...	
320741	9782401029101	http://world-fr.openfoodfacts.org/produit/9782...	kiliweb	1491508021	2017-04-06T19:47:01Z	1491508021	2017-04-06T19:47:01Z	Fiche Brev

- **320 772 lignes * 162 colonnes = 52 millions cases**
- **12 millions de cases non vide**
- **Colonnes :**
  - **56 colonnes données descriptives/qualitatives**
  - **106 colonnes données quantitatives dont 100 sont continues**

CRÉATION DE L'ECO-NUTRISCORE

- **Amélioration des indicateurs évaluant les qualités nutritives des articles**

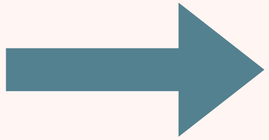
- **Nouveaux indicateurs discrets évaluant la notion d'écologie des articles :**
  - **Nombre d'ingrédients provenant de l'huile de palme**
  - **Nombre d'allergènes**
  - **Si l'article est labellisé BIO**
  - **Score d'emballage de l'article**

NETTOYAGE DES DONNÉES

NETTOYAGE DES DONNÉES

- Supprime les colonnes qualitatives ayant moins de 1/3 de données remplies
- Supprime les colonnes quantitatives ayant moins de 1/3 de données remplies

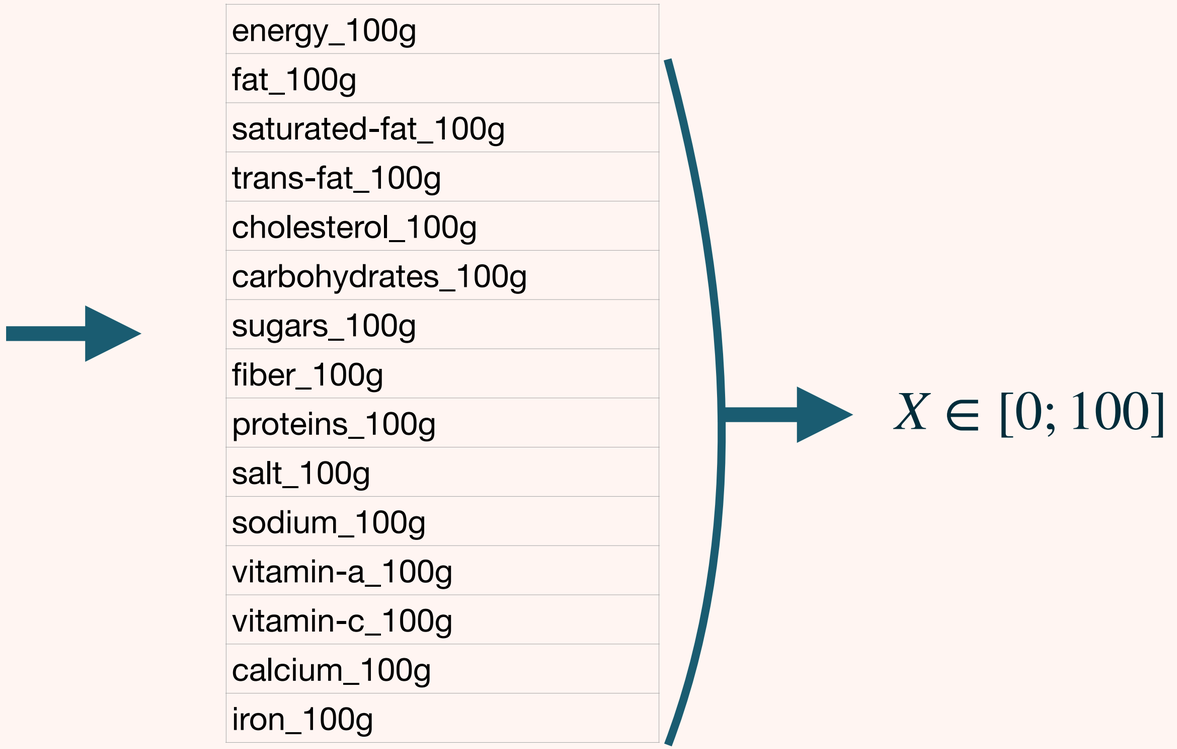
- 320 772 lignes * (106+56) colonnes :
  52 millions cases



- 320 772 lignes * (20 + 51) colonnes :
  23 millions cases

VALEURS ABERRANTES

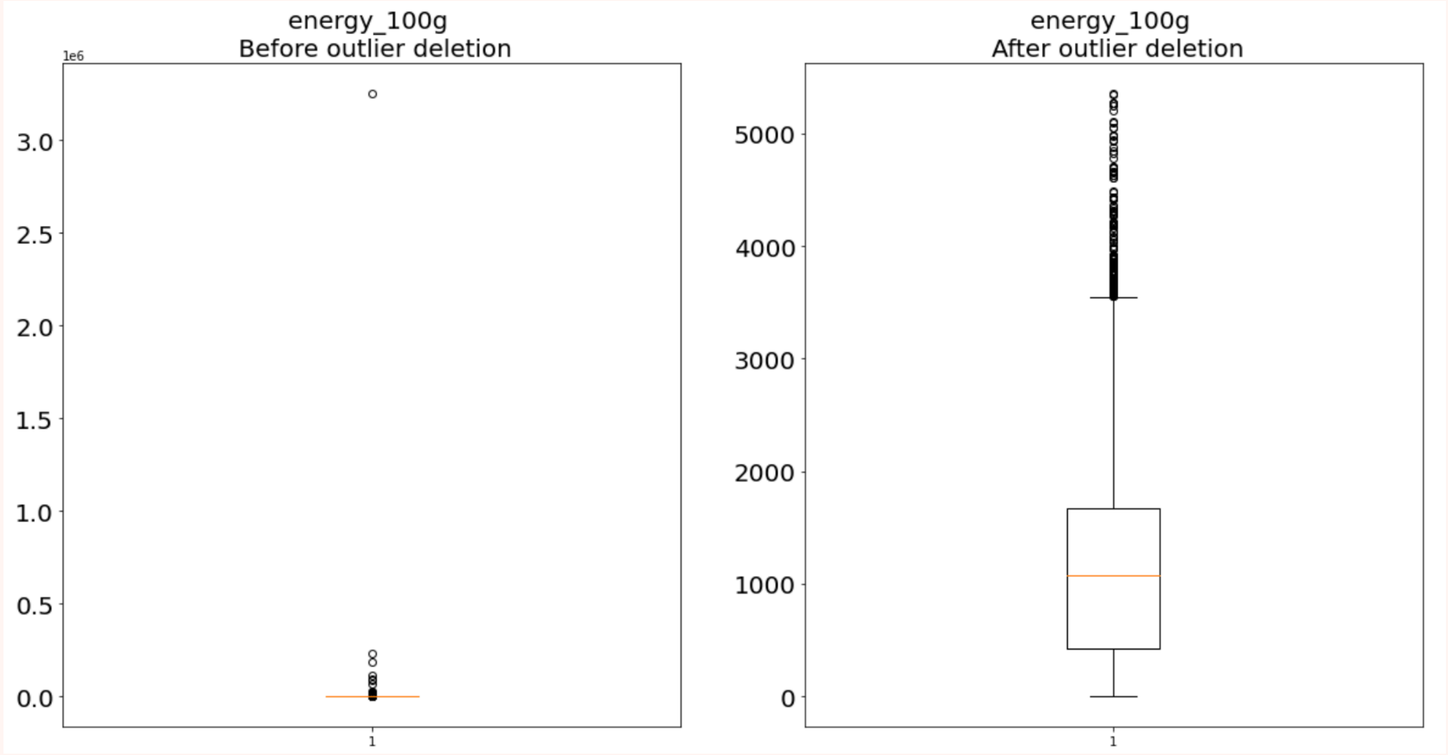
- **Données quantitatives :**
  - **5 discrètes**
  - **15 continues : dont 14 en g et 1 en KJ**



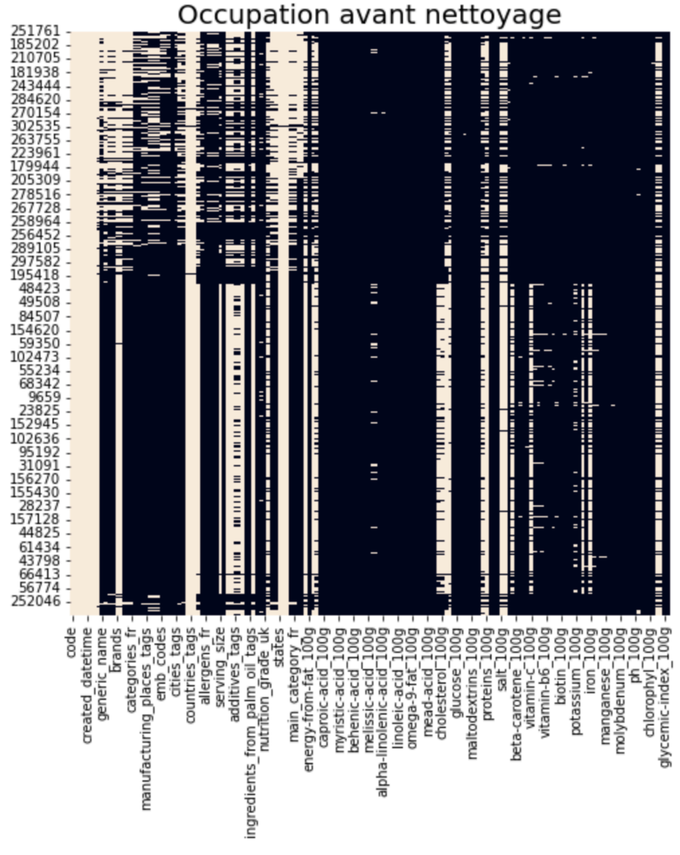
Suppression de 271 valeurs non comprise entre 0 et 100 g

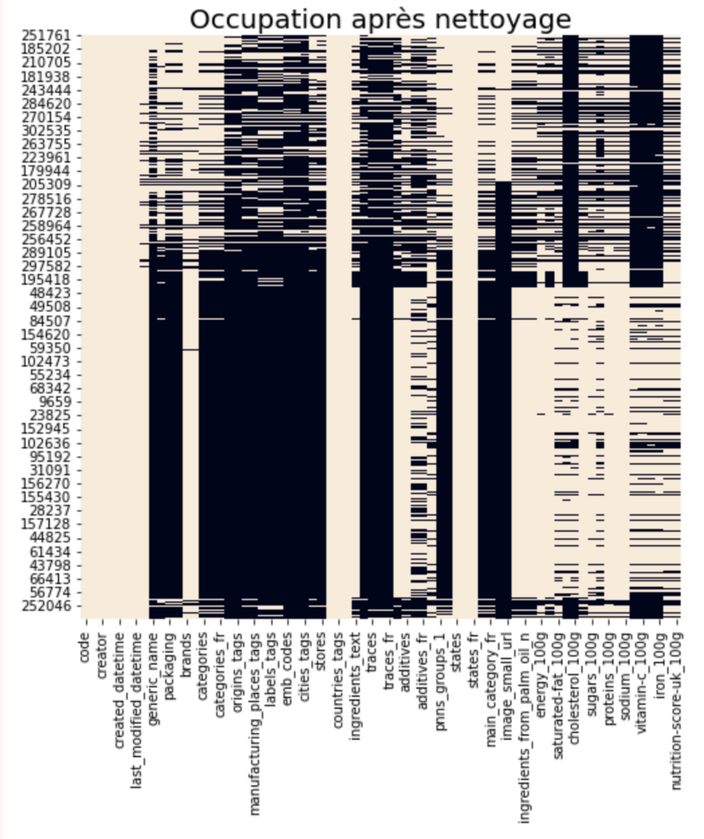
DETECTION D'OUTLIERS

- **Valeurs considérées comme outliers :**
- $X < Q1 - 1.5 * IQR$
- $Q3 + 1.5 * IQR < X$



OCCUPATION





- **320 772 lignes * 162 colonnes**
- **Taux de remplissage : 23%**

- **320 772 lignes * 76 colonnes**
- **Taux de remplissage : 62%**

SÉLECTION DES DONNÉES

CALCUL DU NUTRISCORE

- **Calculé à partir de :**
  - **8 données nutritionnelles**

    fruits-vegetables-nuts_100g
    energy_100g
    fat_100g
    saturated-fat_100g
    sugars_100g
    fiber_100g
    proteins_100g
    sodium_100g

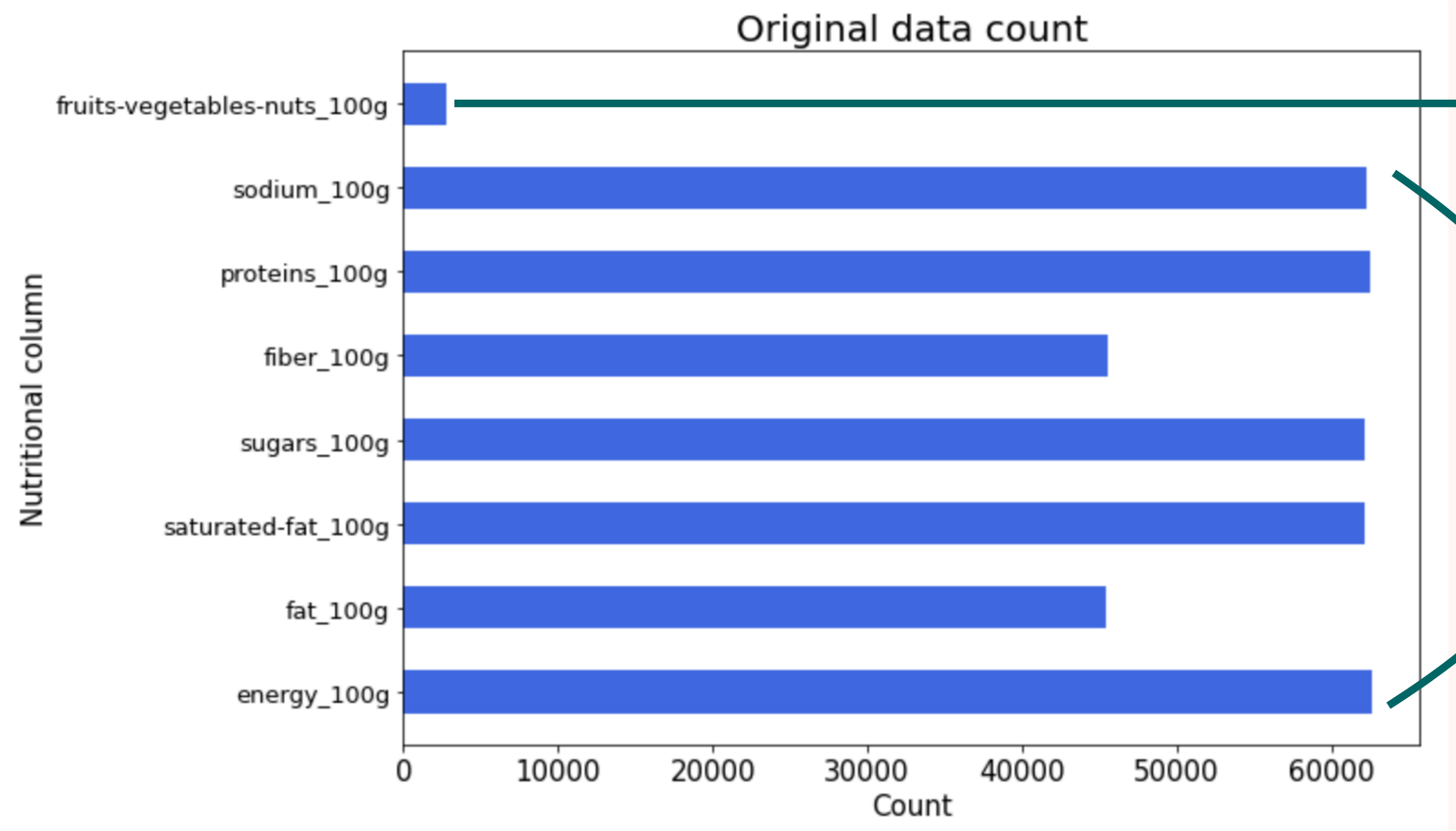
  - **Savoir si le produit est une boisson**
  - **Si c’est une boisson, savoir si le produit est de l’eau minéral**

ECO-NUTRISCORE

- **Colonnes utiles pour le calcul de l'éco-nutriscore (14 colonnes) :**
  - **8 données nutritionnelles**
  - **Boisson / Eau minérale**
  - **Nombre d'ingrédients provenant de l'huile de palme**
  - **Nombre d'allergènes**
  - **Si l'article est labellisé BIO**
  - **Score d'emballage de l'article**
- **Restriction BDD aux articles français : 98 103 articles**

DONNÉES NUTRITIONNELLES

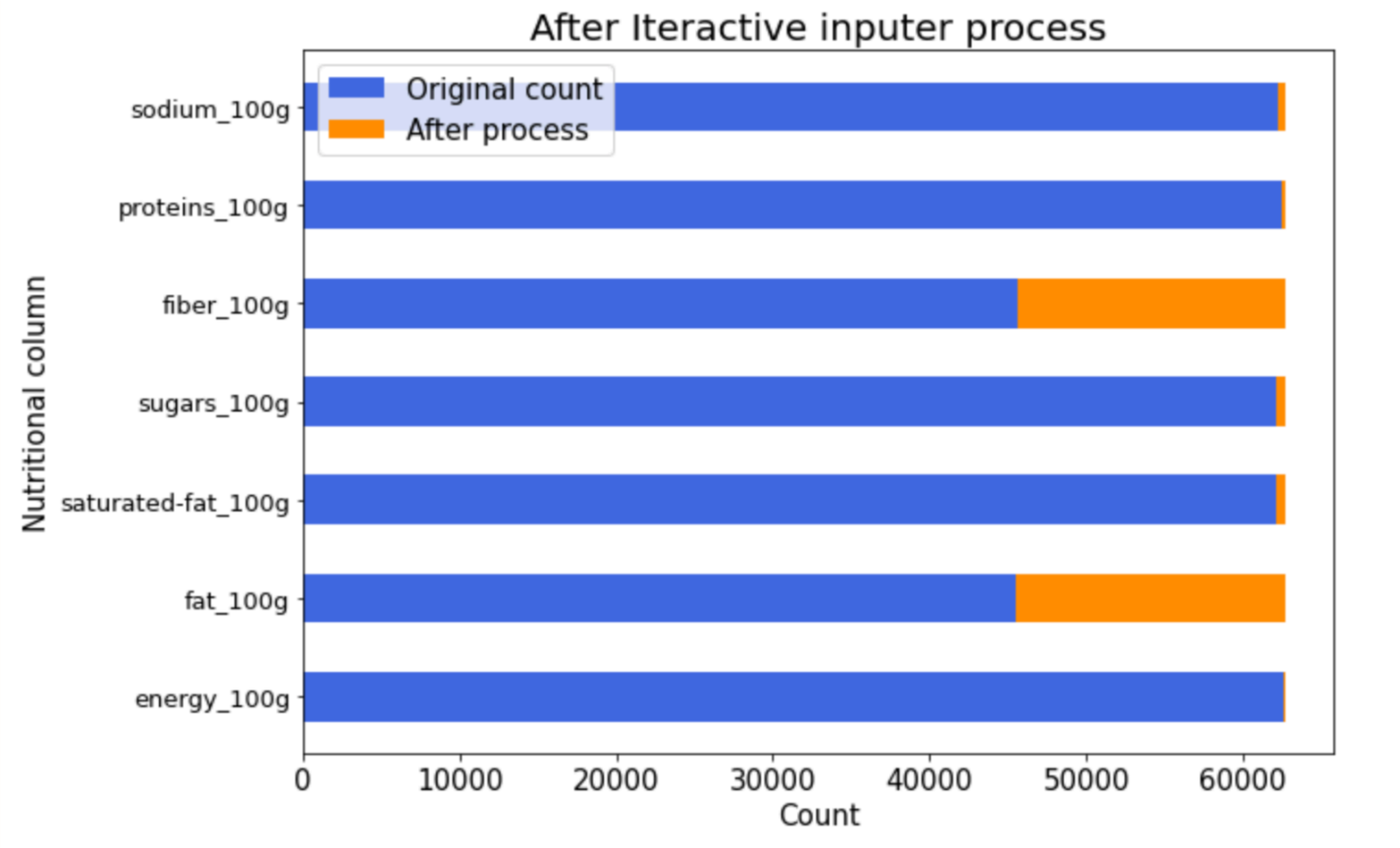
Sélection des articles remplis au moins de moitié



Amélioration/remplissage

Iterative Imputer de Scikit-Learn: Estimation of Missing Values in Multivariate Data

ITERATIVE IMPUTER



FRUITS LÉGUMES NOIX

- **Colonne « fruits-vegetables-nuts_100g »**

- **Création d'une liste non-exhaustive de catégories contenant des fruits, légumes et noix**

```
fln_list = ["Légumes frais", "Fruits", "Fruits à coques", "Jus de fruits",
            "Fruits secs", "Nectars de fruits", "Huiles", "Huiles d'olive",
            "Jus de pomme", "Tomates", "Olives vertes", "Jus d'orange",
            "Jus de pamplemousse"]
```

- **Recherche des mots clé dans les colonnes catégories**

- **Colonne « fruits-vegetables-nuts_100g » à 100g**

Parameters	Nb avant	Nb après
fruits-vegetables-nuts_100g	1839	3340

REMPLISSAGE PAR ALGORITHME KNN

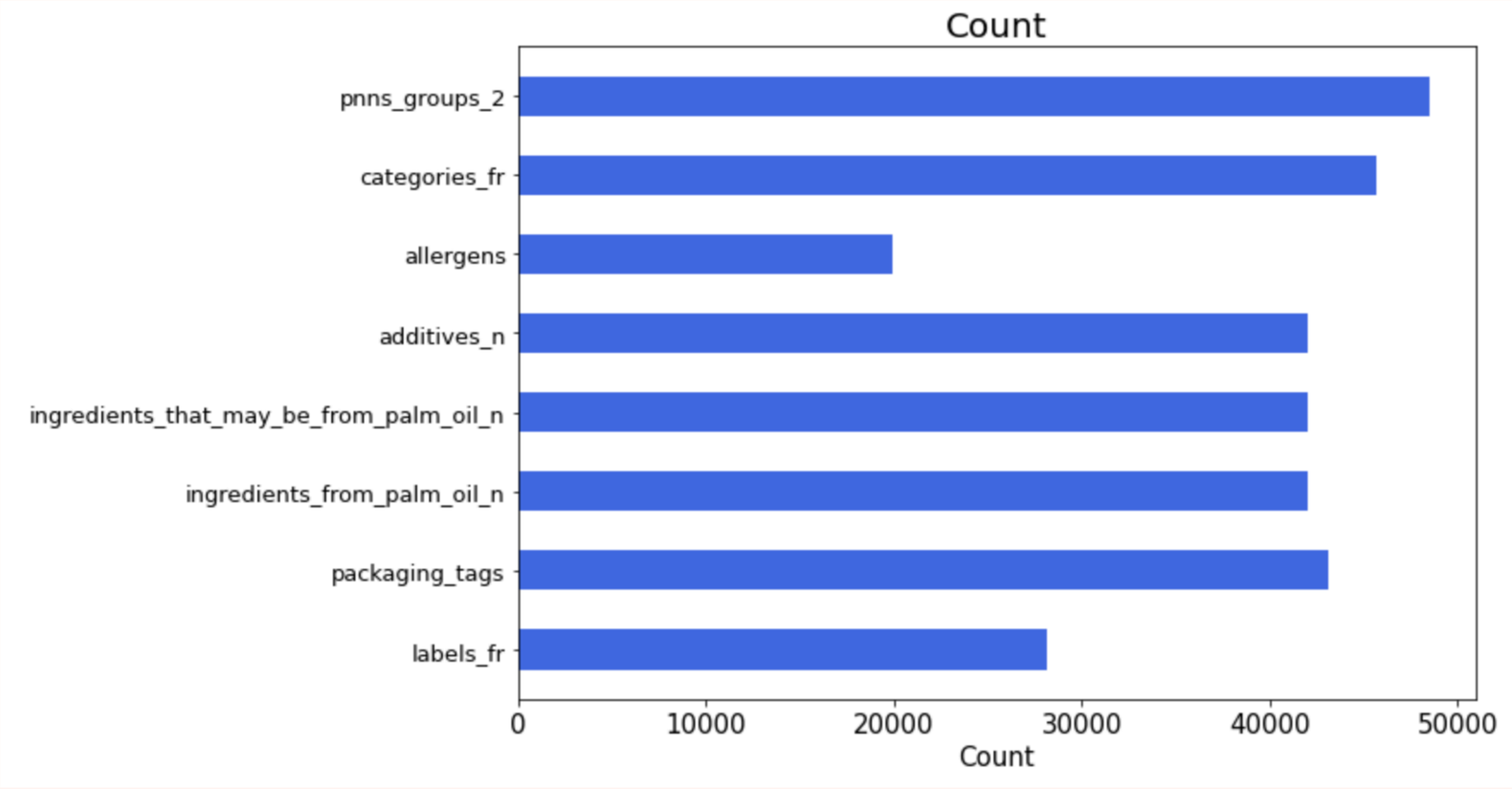
- **Entrainement avec les 7 données nutritionnelles**
- **3340 articles considérés comme « fruits, légumes, noix »**
- **Prédiction sur l'ensemble des données**
- **Résultat : 482 nouveaux articles considérés comme « fruits, légumes, noix »**

Gazpacho Original
Fanta orange
San Pellegrino limonata
Pur jus de raisin Sélection Muscaté
Pur jus de pomme Sélection Bretagne
Pur jus de pomme trouble
Pur jus de pamplemousse rose sélection Floride
Orange sans pulpe, 100% pur jus
Abricots moelleux dénoyautés
100 % Pur Jus Pamplemousse Rose avec pulpe

- **Non concluant**

AMÉLIORATION REMPLISSAGE

- **Boisson / Eau minérale**
- **Nombre d’allergènes**
- **Nombre d’additives**
- **Nombre d’ingrédients provenant de l’huile de palme**
- **Si l’article est labellisé BIO**
- **Score d’emballage de l’article**



INDICATEUR « SCORE EMBALLAGE »

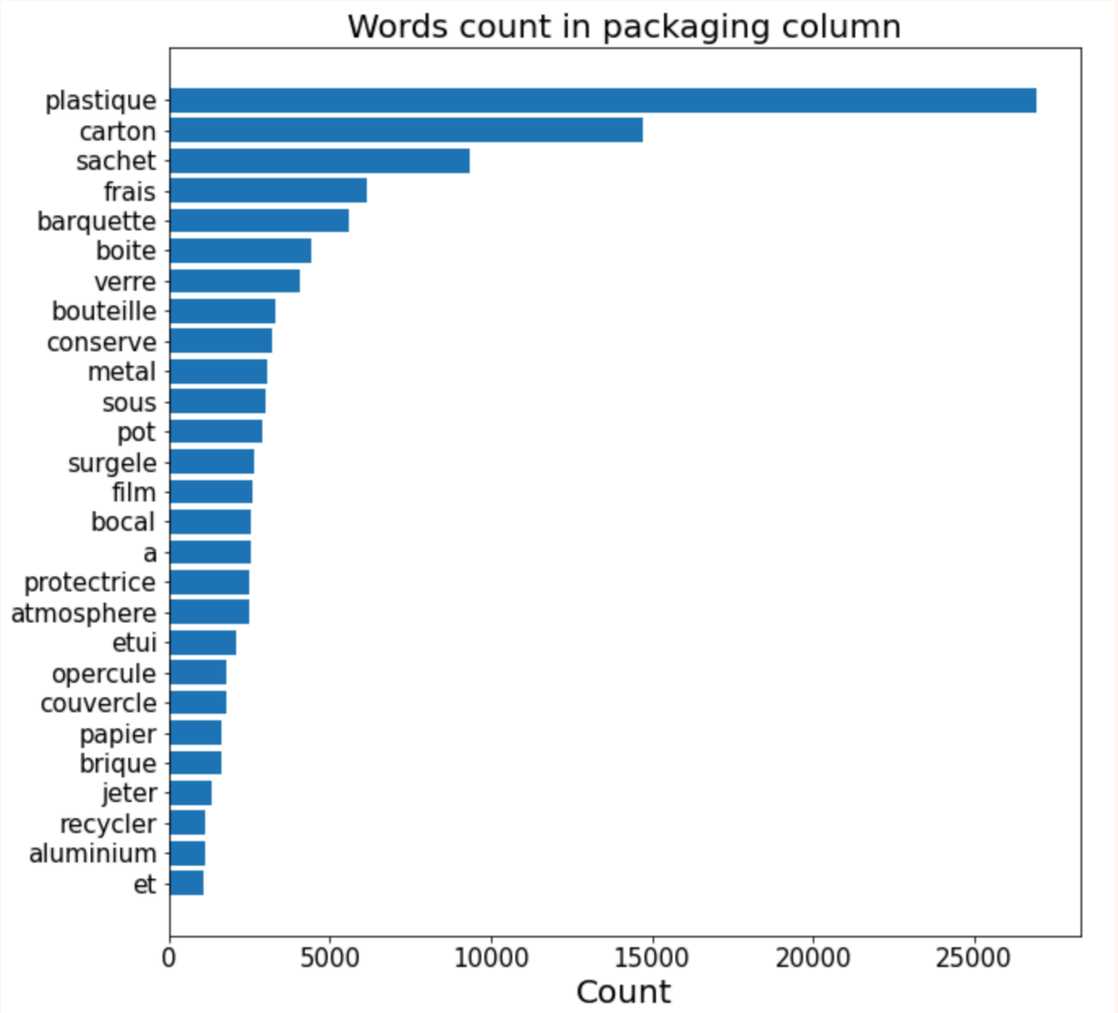
- **Indicateur « packaging_tags » : liste des matériaux des emballages**

plastique	0
carton	1
metal	2
verre	3
bois	4

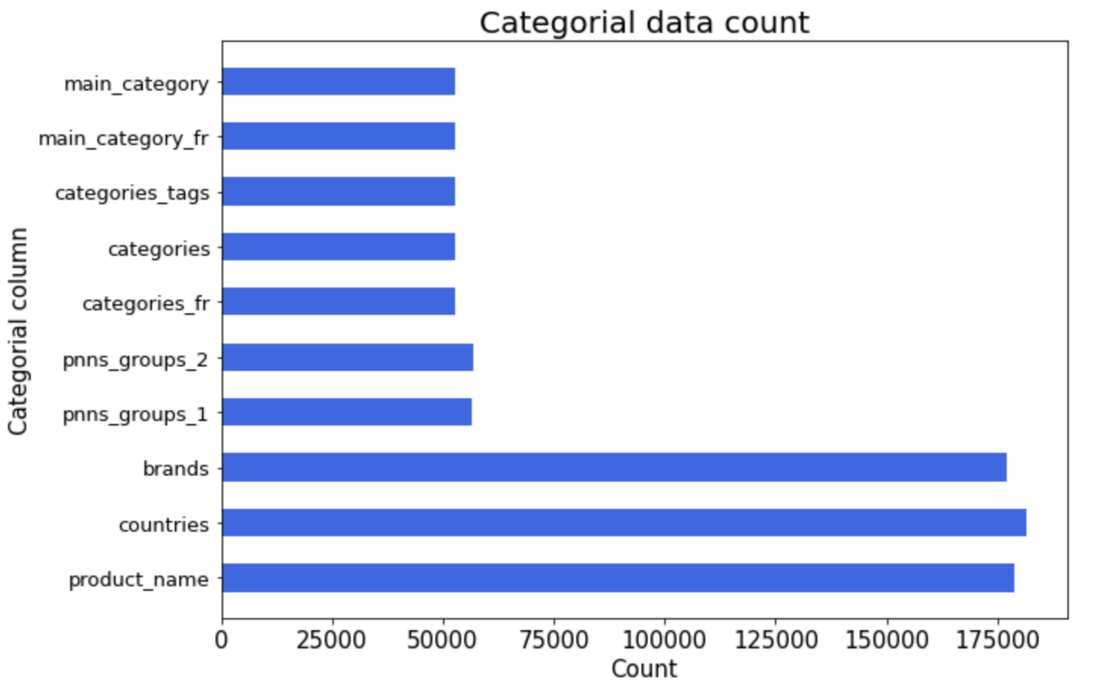
- **Attribue la note du matériel ayant le score le plus bas**

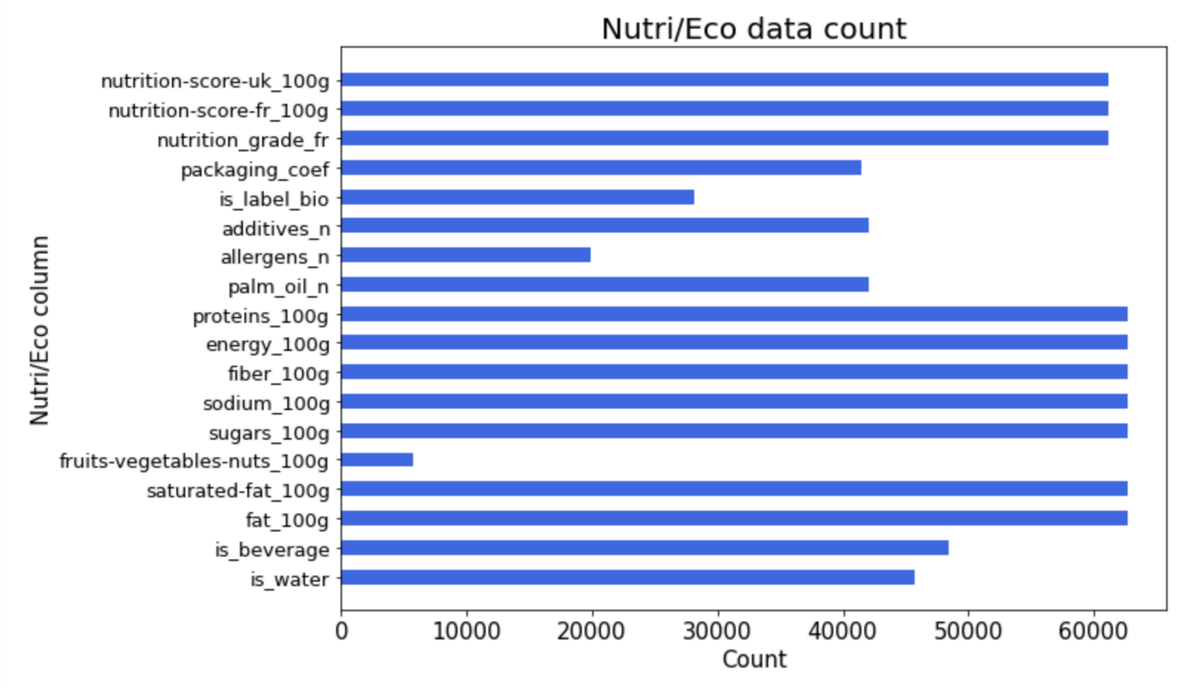
- **Remplissage du score**

Parameters	Nb
packaging_tags	43144
coef	41105



SELECTION DES DONNÉES

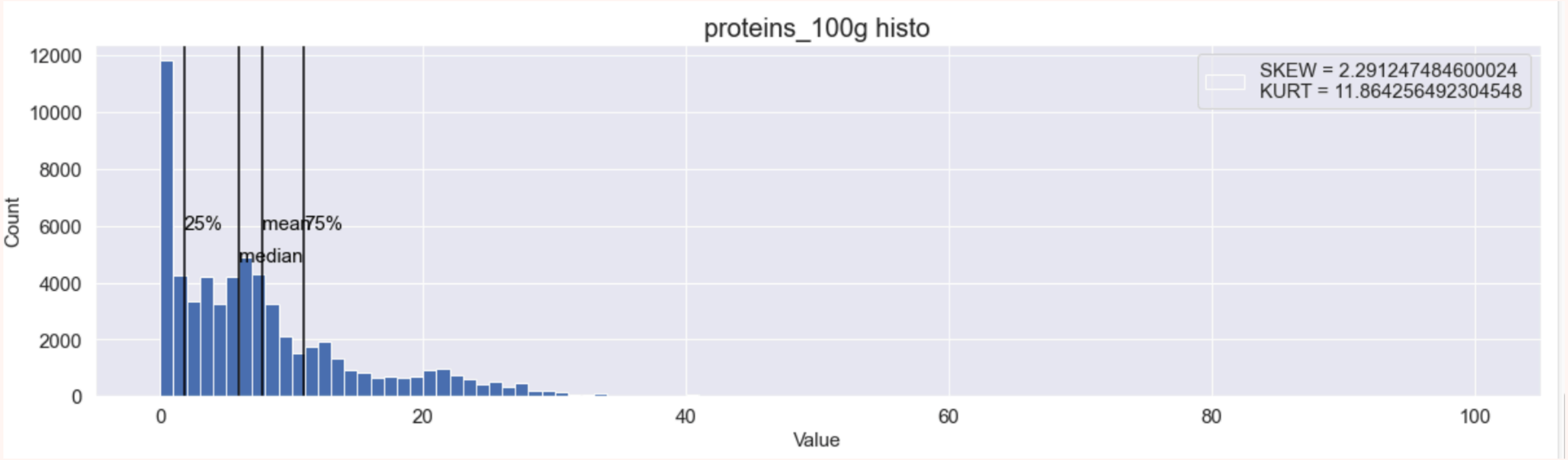




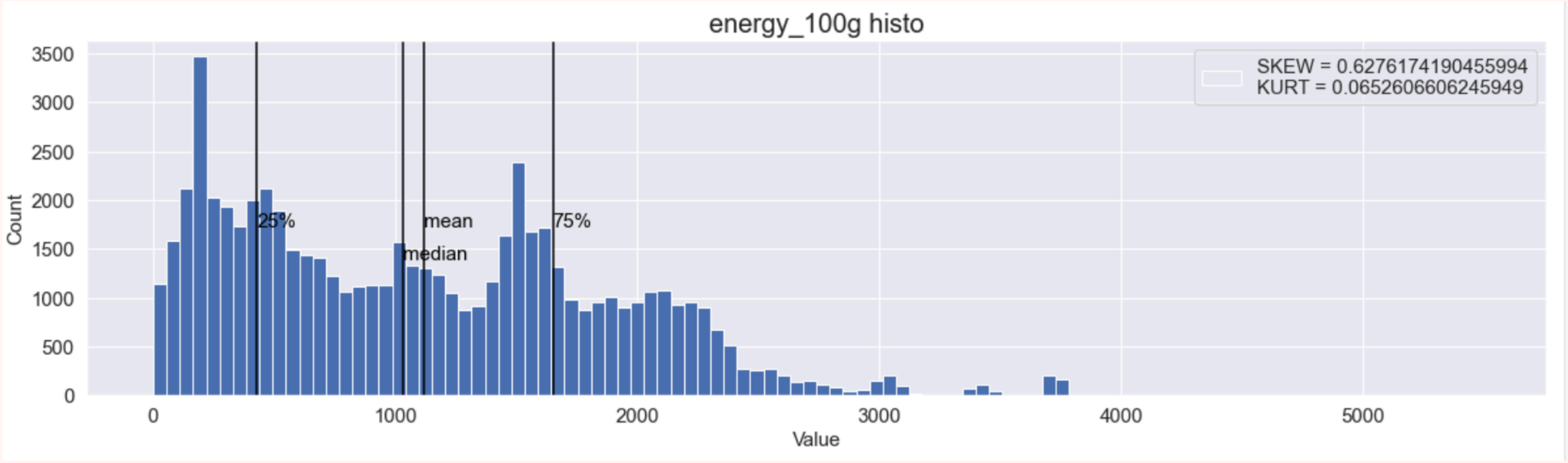
- **2 colonnes (article, marque)**
- **7 colonnes pour catégoriser**
- **10 colonnes pour calcul du nutriscore**
- **3 colonnes évaluant le nutriscore**
- **5 colonnes pour calcul de l'écoscore**

ANALYSE DES DONNÉES

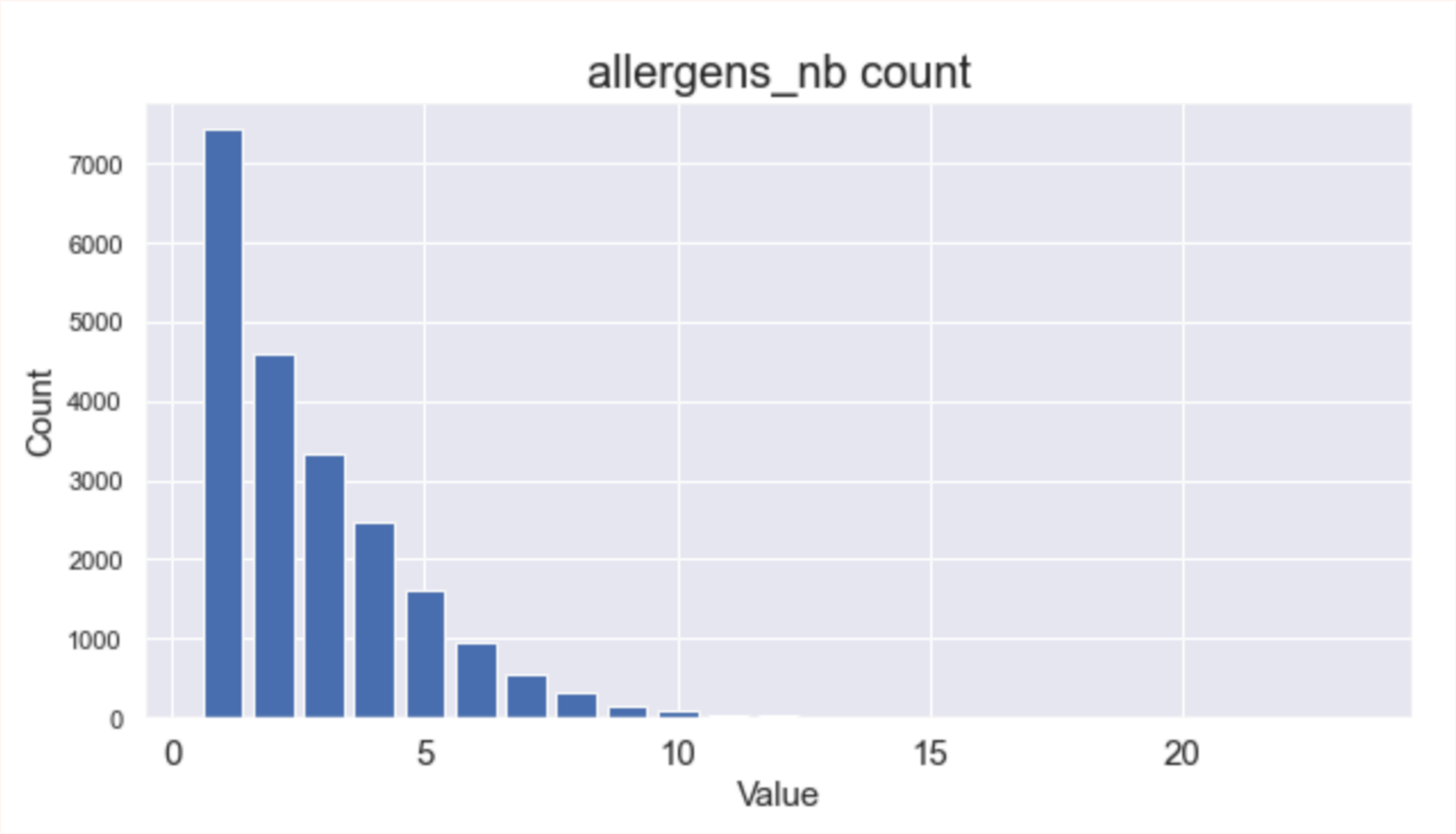
ANALYSE VARIABLE CONTINUE



ANALYSE VARIABLE CONTINUE



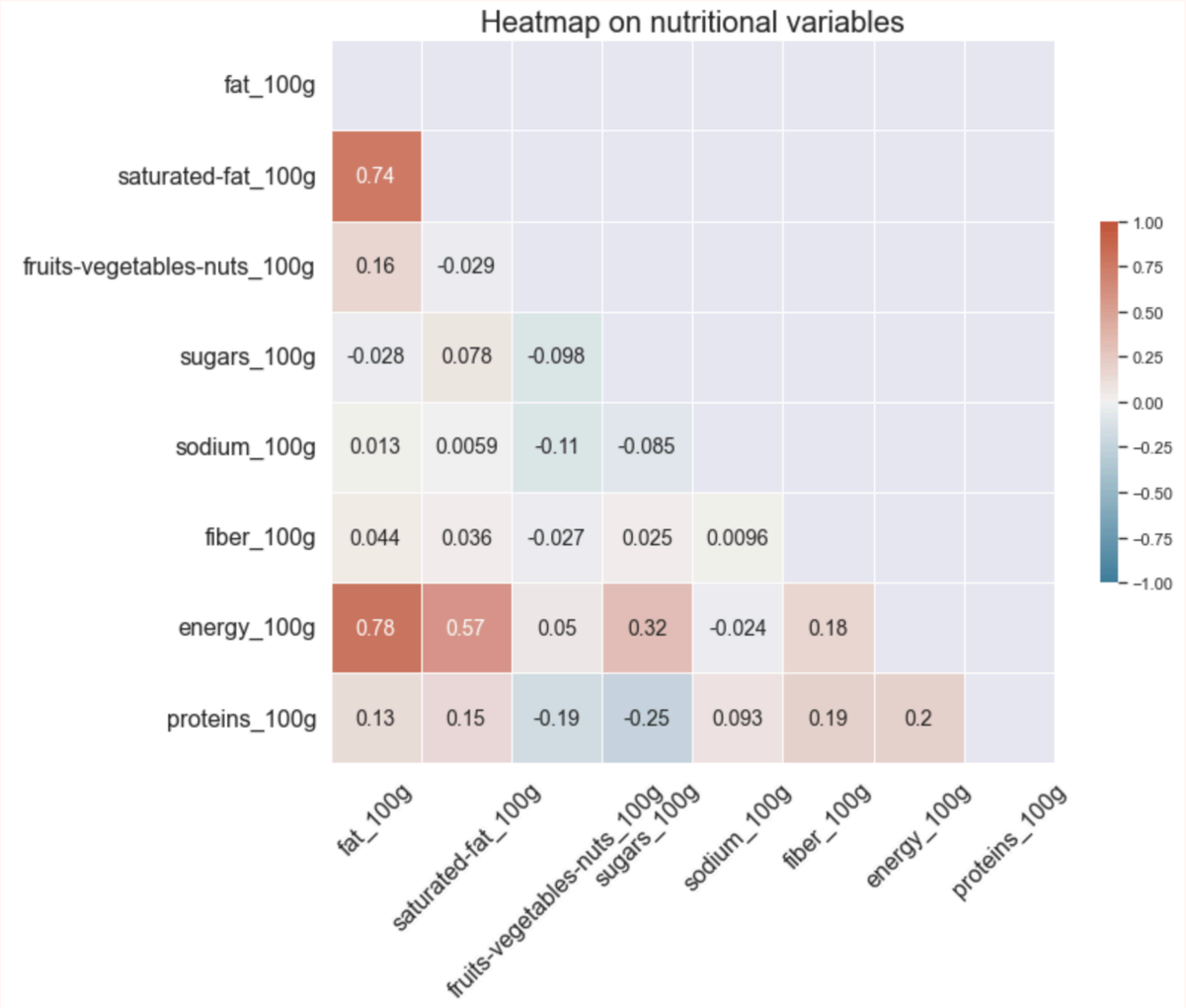
ANALYSE VARIABLE DISCRETE



ANALYSE MULTIVARIÉE

- **Variable continue VS**

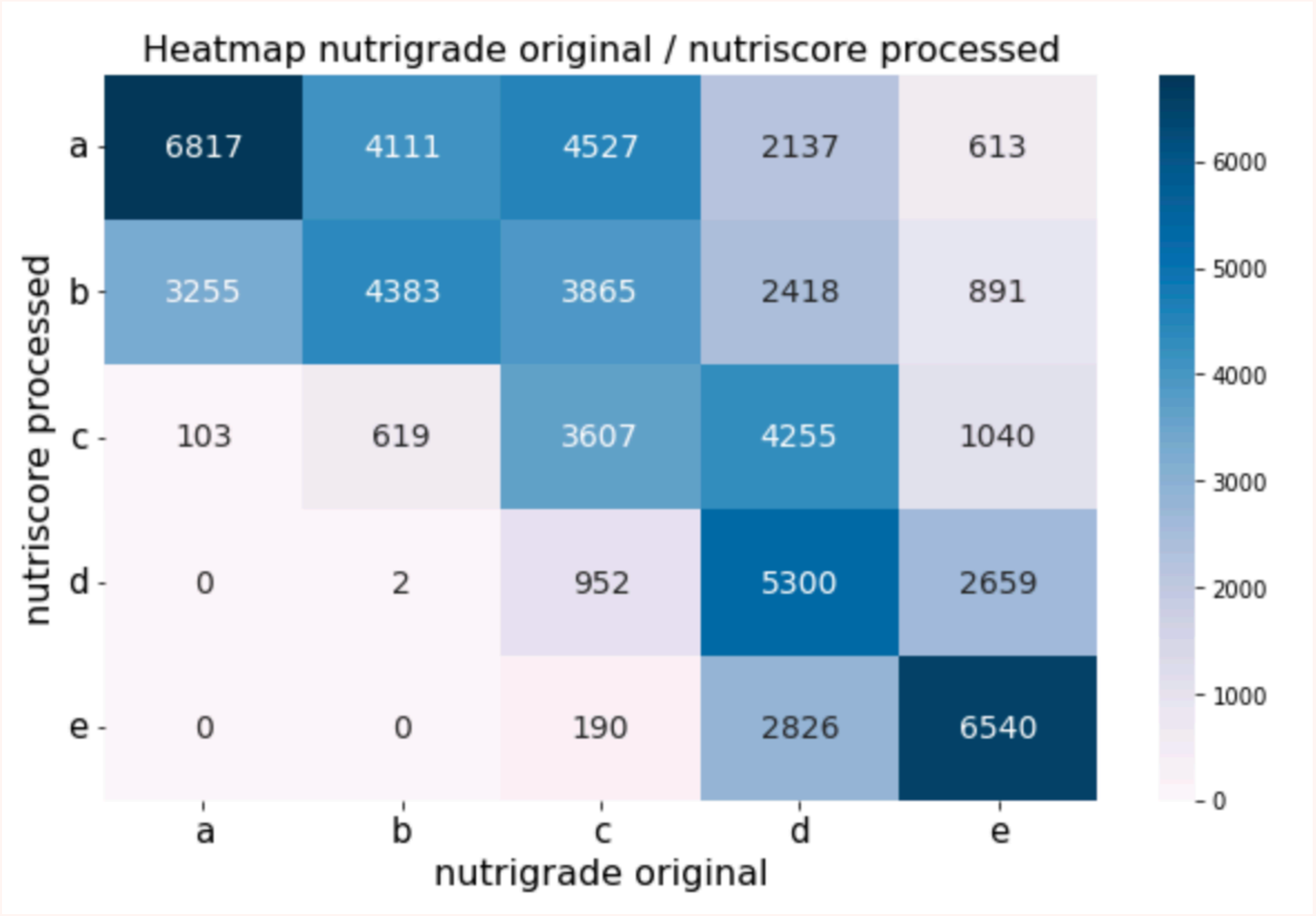
  Variable continue



ANALYSE MULTIVARIÉE

- **Variable discret VS**

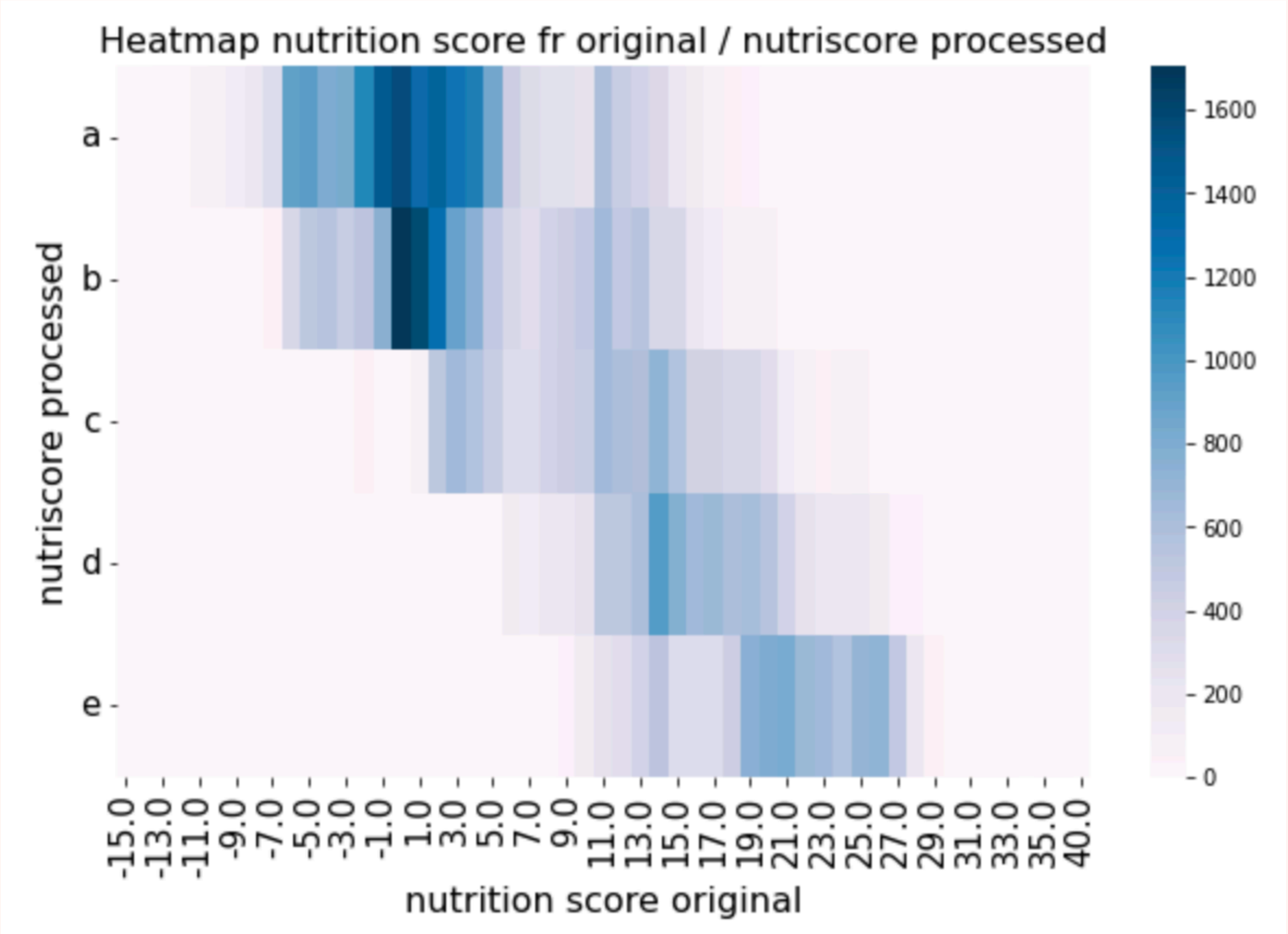
  Variable discret



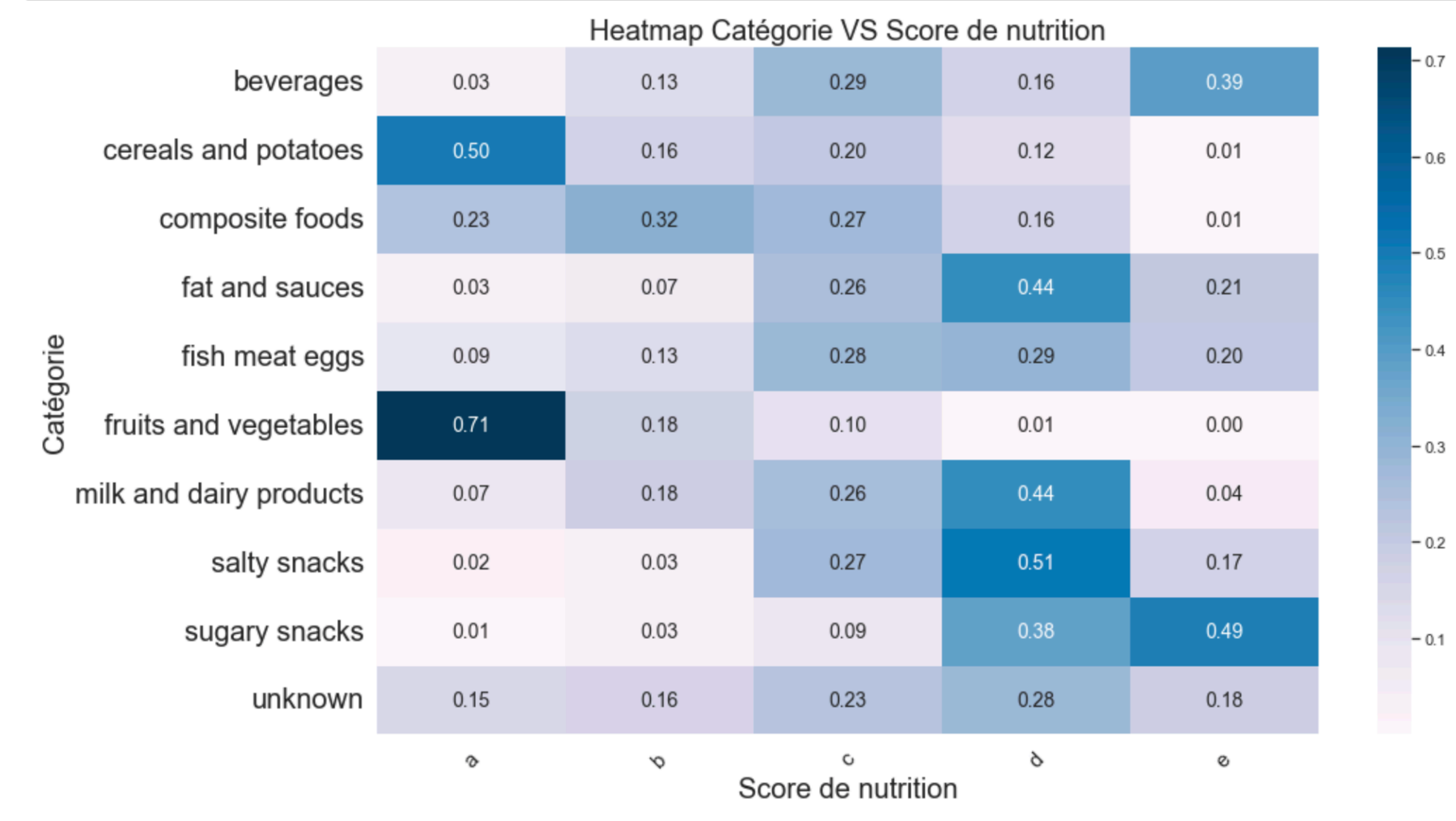
ANALYSE MULTIVARIÉE

- **Variable discret VS**

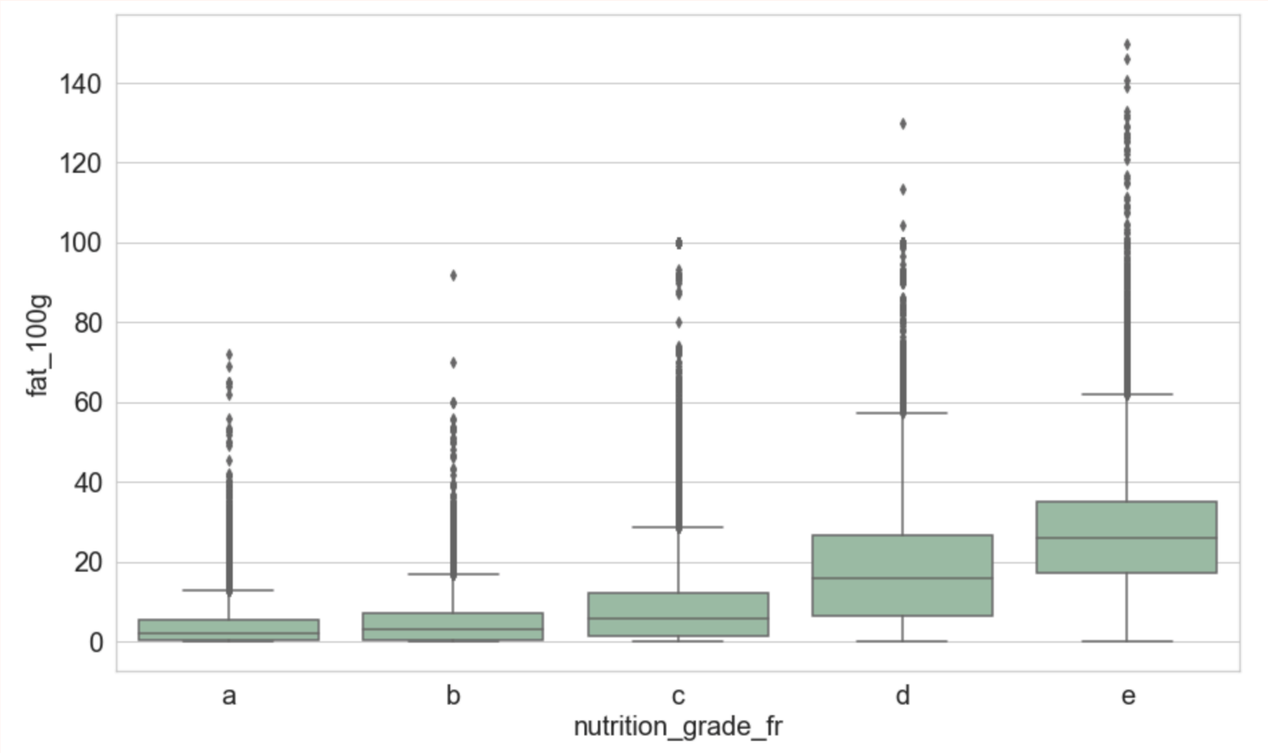
  Variable discret



HEATMAP CATEGORIE

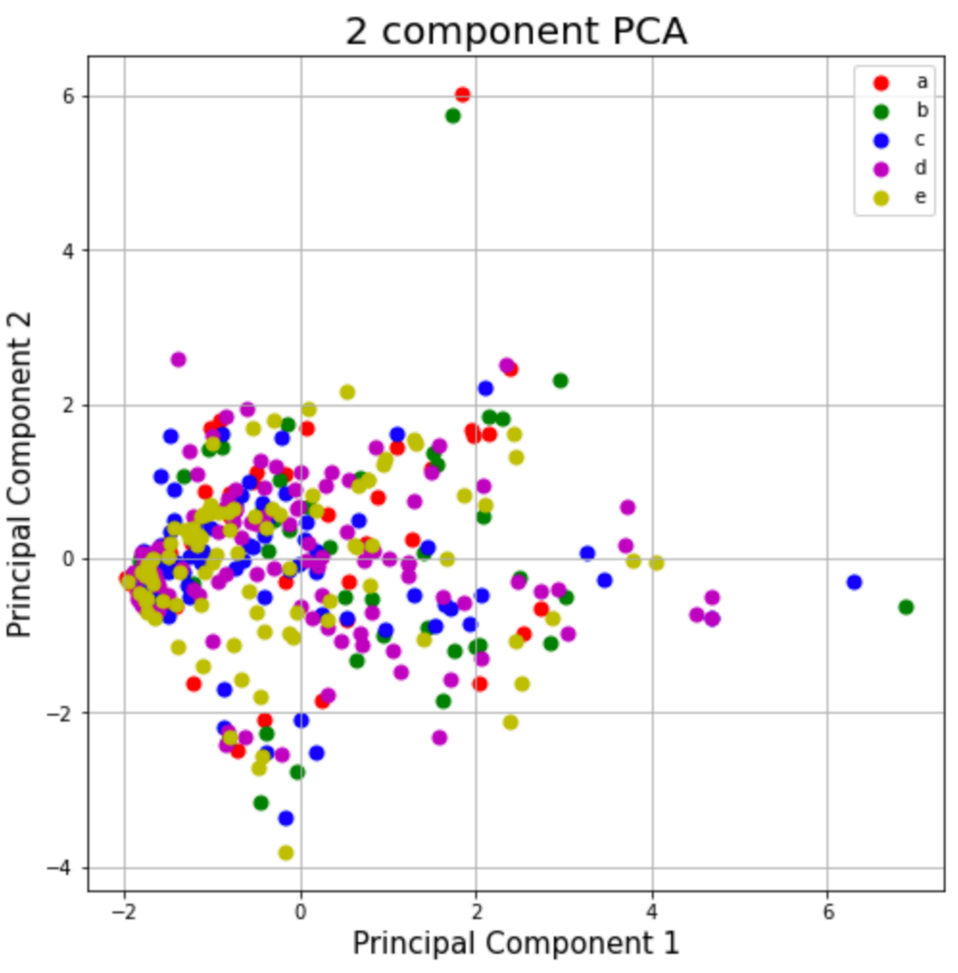


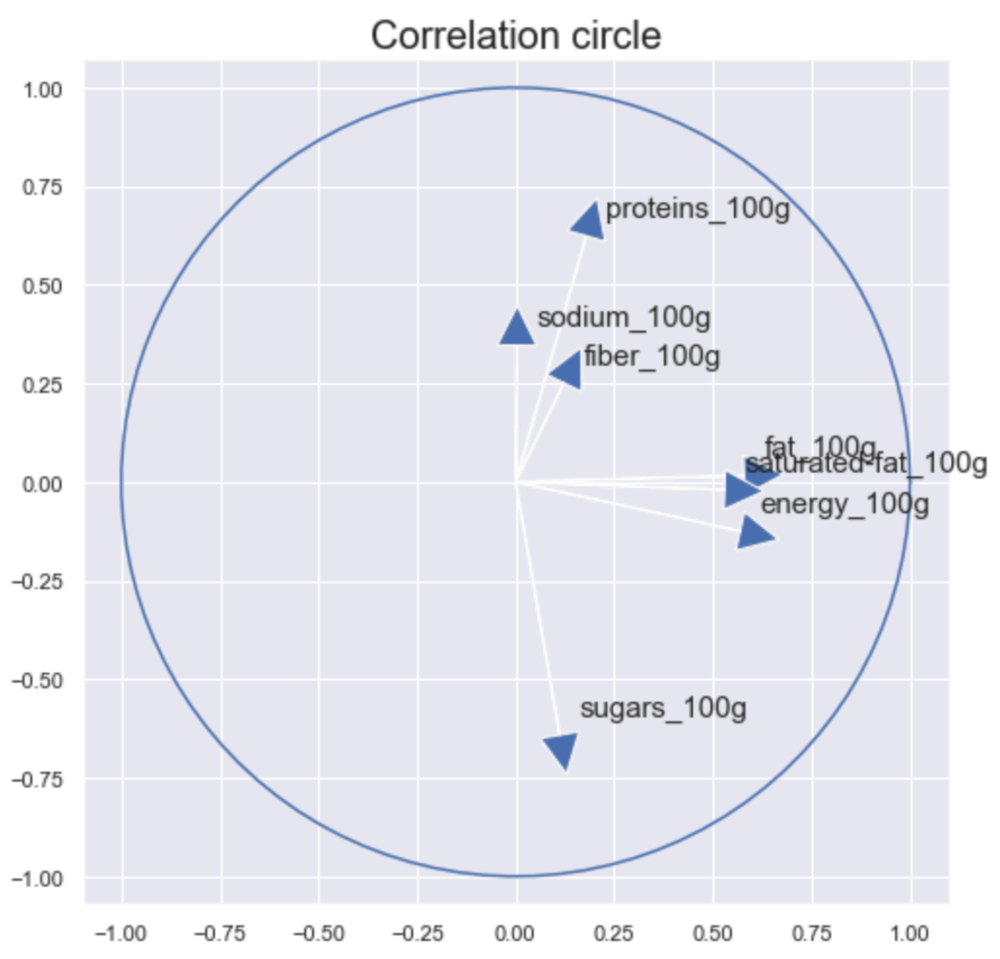
ANOVA



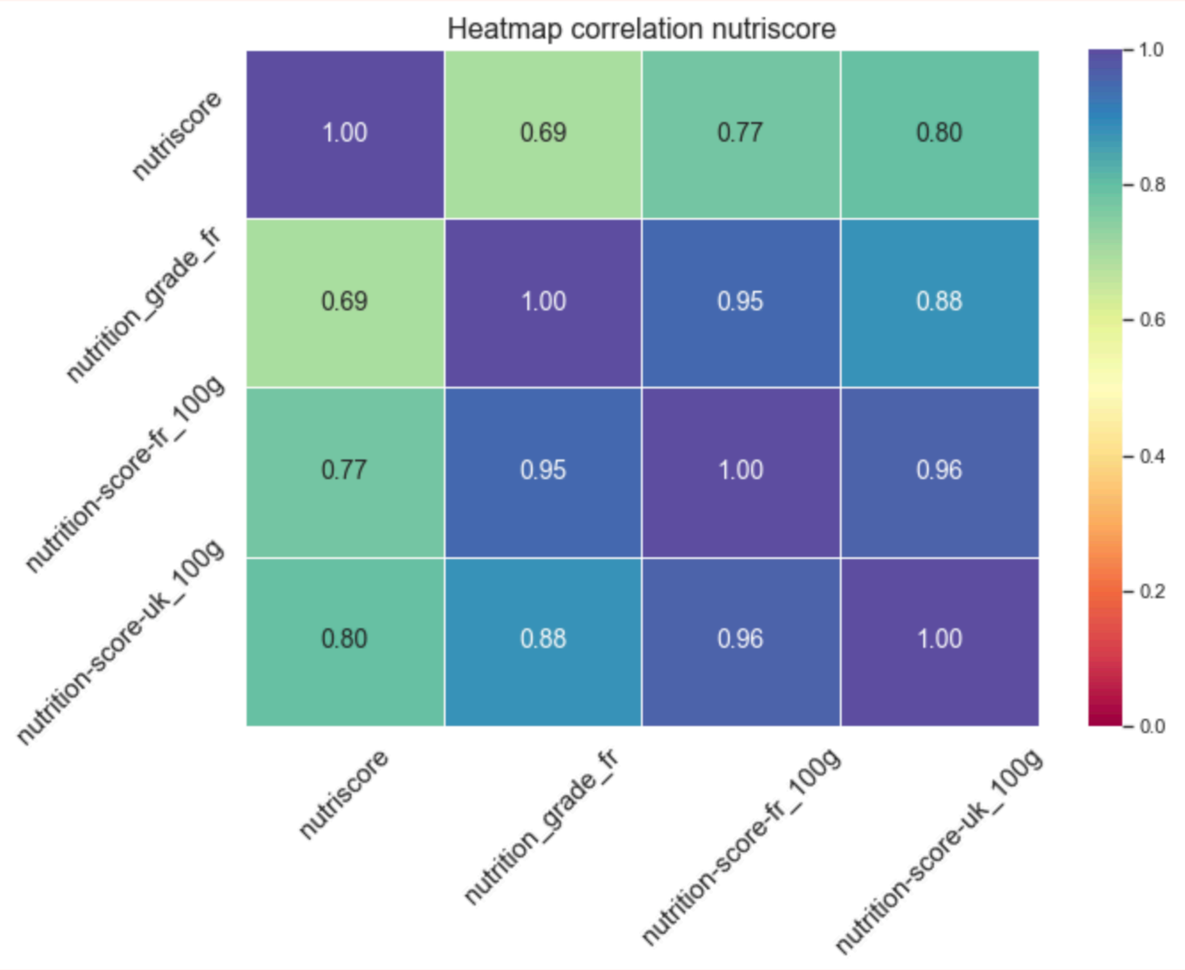
	df	sum_sq	mean_sq	F	PR(>F)
C(gradeN)	4.0	3.519006e+06	879751.534486	4697.162435	0.0
Residual	45570.0	8.534999e+06	187.294254	NaN	NaN

PCA

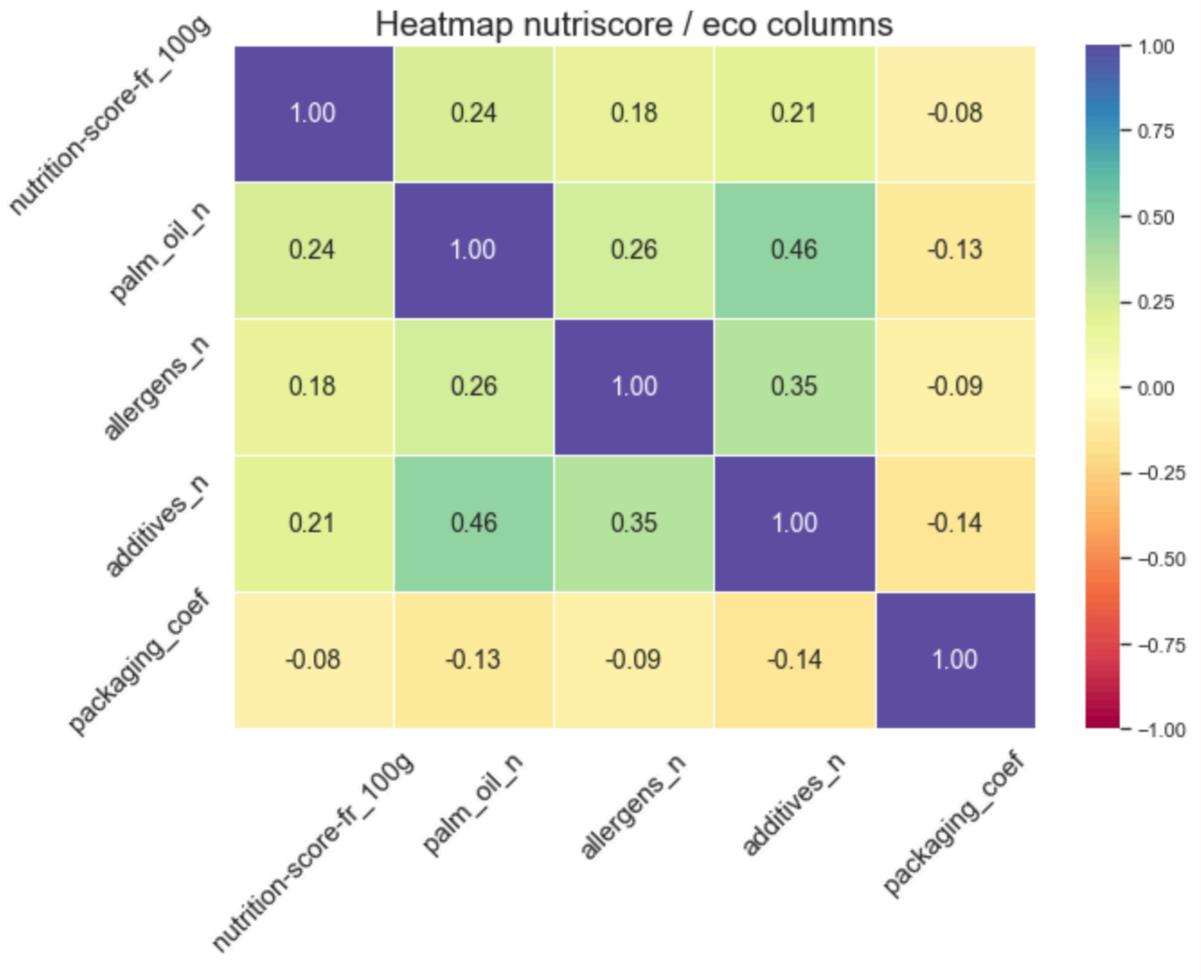

2 component PCA
Principal Component 2
Principal Component 1
a
b
c
d
e
6
4
2
0
−2
−4
−2
0
2
4
6


Correlation circle
proteins_100g
sodium_100g
fiber_100g
fat_100g
saturated-fat_100g
energy_100g
sugars_100g
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

CORRELATION NUTRISCORES



CORRELATION ECO-NUTRISCORE



CONCLUSION

- Corrélations existantes entre nutriscore et colonnes écologiques donc à approfondir
- Analyse non complète par manque de données (colonnes boisson, fuits-légumes-noix)
- Calcul du eco-nutriscore complexe : faire des analyses séparées en fonction de catégories de produits
- Meilleure base de données existe?