Santé publique France

Application sur l'alimentation au service de la santé publique

Data Science | Projet 3

Fırat Yasar

03/08/2021

Sommaire

Présentation de la problématique

- L'agence **Santé publique France** recherche des idées innovantes d'applications en lien avec l'alimentation.
- Pour cela, nous avons à notre disposition la base de données Open Food Facts :

  https://world.openfoodfacts.org

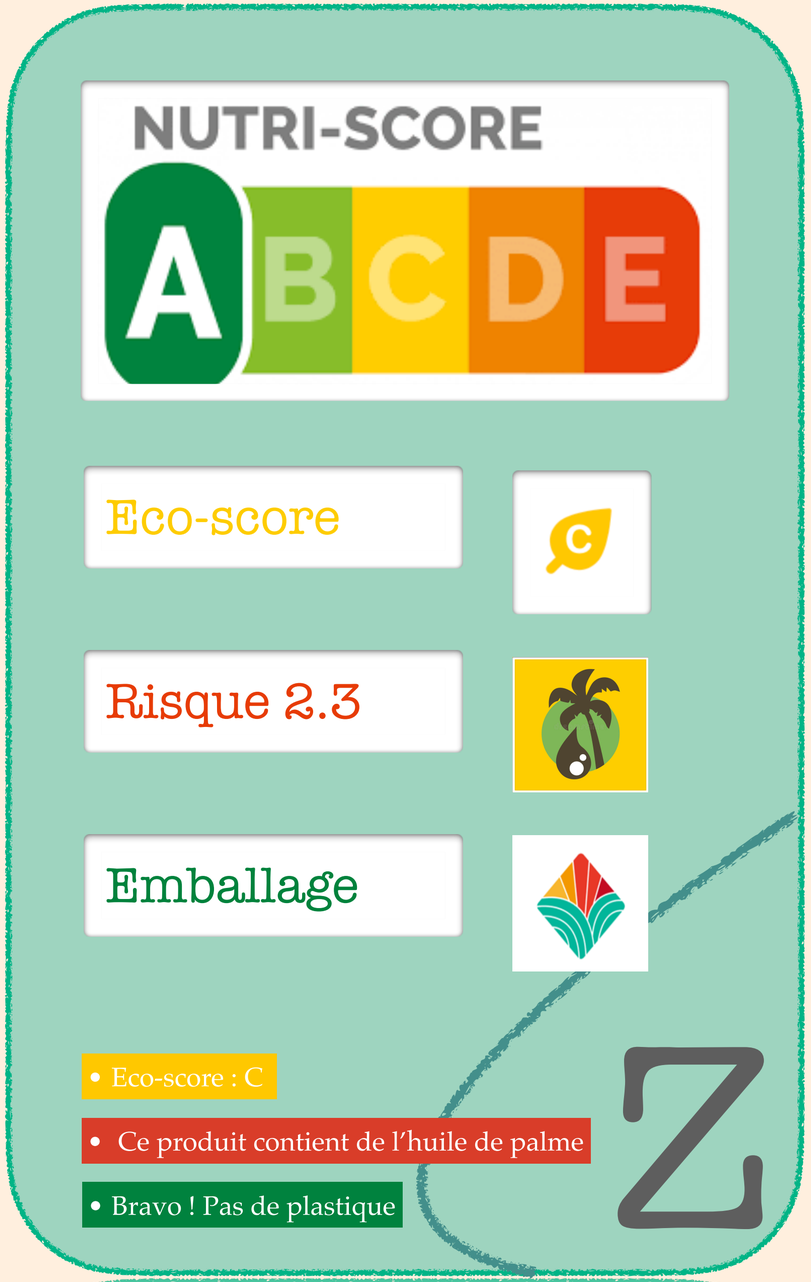
- Nous visons à proposer une idée d'application

L'idée de l'application : NUTRI +Z

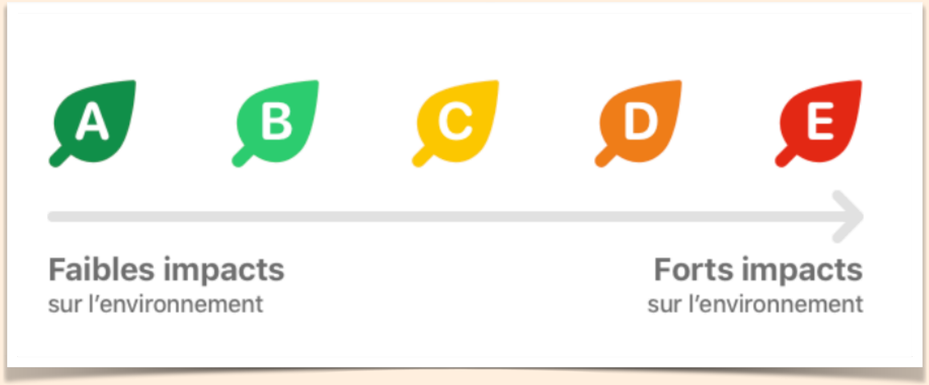
- Une application simple qui va à l'essentiel pour la plupart des consommateurs.

- Affichage
  - ➡ NUTRI : le nutriscore du produit
  - ➡ + Z : une alerte

- Les objectifs :
  - ➡ aider les consommateurs à choisir des aliments de meilleure qualité nutritionnelle
  - ➡ ralentir l'utilisation massive de matériaux plastiques



Eco-score
Une échelle représentant l'impact environnemental des produits alimentaires

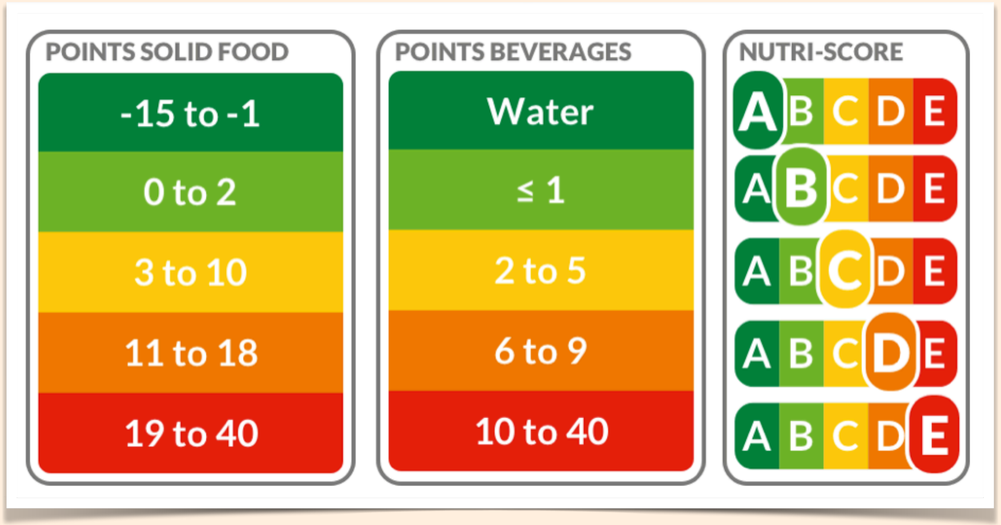


L'impact environnemental tient compte de plusieurs facteurs sur la pollution de l'air, des eaux, des océans, du sol, ainsi que les impacts sur la biosphère.

Source : https://docs.score-environnemental.com

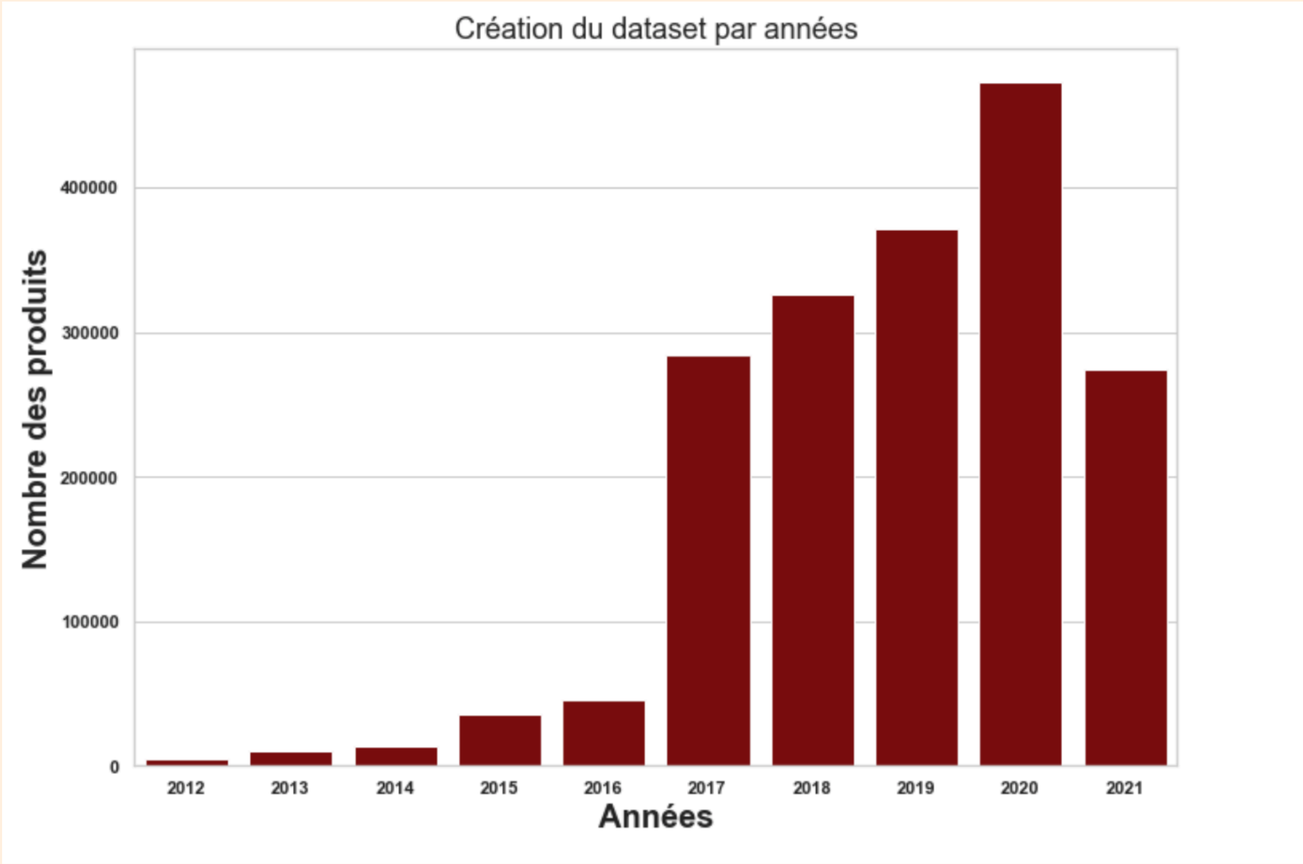
Nutri-grade
Une échelle alphabétique de A à E (A étant le meilleur score)

Nutri-score
Une échelle numérique de -15 à 40 (-15 étant le meilleur score)



Le calcul de nutri-score est basé sur les nutriments comme sucres, protéines, gras, gras-saturé, sel, etc… pour 100g de produit et la valeur énergétique du produit.

Découverte du jeu de données



Le jeu de données compte 186 colonnes et 1 837 365 lignes.

Chaque ligne représente un produit.

Il y a des produits alimentaires, des boissons, etc.

Le jeu de données contient les produits vendus dans le monde entier.

	pays
France	484479
United States	270432
Spain	119736
Germany	52906
Italy	48673
Switzerland	32925
Belgium	31708
United Kingdom	31220
Canada	18825

Le jeu de données

Information générale

- Nom du produit
- Code barre
- Date d'entrée dans le dataset
- etc.

Information categorielle

- Les catégories (snacks, légumes, boissons, etc.)
- Pays d'origine

Les nutriments (détail sur 100g de produit)

- Protéines
- Gras
- Gras saturé, etc.

Autres indicateurs sur la santé et l'écologie

- Energie
- Nutri-score
- Eco-score, etc.

Analyse exploratoire sur le jeu de données

DATA

186 colonnes
1 837 365 lignes

DATA_clean

24 colonnes
1 229 441 lignes

Etape 1 –

Il y a 1526 lignes où le nom du produit n'est pas précisé.

Il y a des lignes où la colonne 'product_name' est vide (315 lignes), mais on trouve les noms abrégés (des produits) dans la colonne 'abbreviated_product_name' (où les noms génériques (203 lignes)). Au total 518 lignes.

186 colonnes
1 759 154 lignes

	product_name	abbreviated_product_name
887676	NaN	Delice de poulet -25% sel 150g 4tr+1gt
887593	NaN	Id petit saucisdelice 230gr

Etape 2 –

401 582 lignes dupliquées (le nom du produit et sa marque)

186 colonnes
1 357 572 lignes

170911	2019-11-19T19:50:23Z	Aceite de oliva virgen extra	NaN	NaN	NaN	castela notti	castela-notti	Plant-based foods and beverages,Plant-based fo...
264335	2019-08-27T19:26:41Z	Aceite de oliva virgen extra	NaN	NaN	NaN	Kirkland	kirkland	Plant-based foods and beverages,Plant-based fo...
325591	2019-12-12T20:18:06Z	Aceite de oliva virgen extra	NaN	NaN	NaN	NaN	NaN	Plant-based foods and beverages,Plant-based fo...
325602	2019-12-23T08:32:01Z	Aceite de oliva virgen extra	NaN	NaN	NaN	Torres	torres	Plant-based foods and beverages,Plant-based fo...
326833	2019-10-28T23:02:39Z	Aceite de oliva virgen extra	NaN	NaN	NaN	NaN	NaN	Plant-based foods and beverages,Plant-based fo...
...	...	...	...	...	...	...	...	...
1742764	2019-10-28T19:12:42Z	Aceite de oliva virgen extra	NaN	NaN	NaN	Vivo	vivo	Plant-based foods and beverages,Plant-based fo...
1742765	2019-08-31T08:07:18Z	Aceite de oliva virgen extra	250 ml	NaN	NaN	NaN	NaN	Plant-based foods and beverages,Plant-based fo...

Etape 3 –

- 1386 outliers pour les nutriments (valeur > 100 ou < 0)
- 8887 outliers pour les energy_100g (valeur énergie > 3700)

186 colonnes
1 348 685 lignes

Etape 4 –

Il y avait aussi des valeurs atypiques (166 lignes)

186 colonnes
1 348 519 lignes

	created_datetime	product_name	quantity	packaging
17049	2020-10-19T07:03:56Z	Chargement…	NaN	plastique
55321	2017-02-19T23:09:29Z	Chargement…	NaN	NaN
134912	2016-10-08T15:51:04Z	Loading…	355 ml	Can,Aluminum
271130	2020-09-13T17:02:45Z	جاري التحميل…	500	groupe said
288924	2020-09-17T07:32:43Z	Chargement… emmental	0.168 kg	coupe pre emballée
...	...	...	...	...
1829718	2020-09-10T06:42:10Z	Loading…	NaN	plastic
1832440	2020-09-08T15:51:22Z	Loading…	NaN	NaN
1835977	2020-10-04T08:26:45Z	درحال بارگذاری…	NaN	NaN
1836530	2021-04-05T16:04:27Z	Cargando…	NaN	NaN

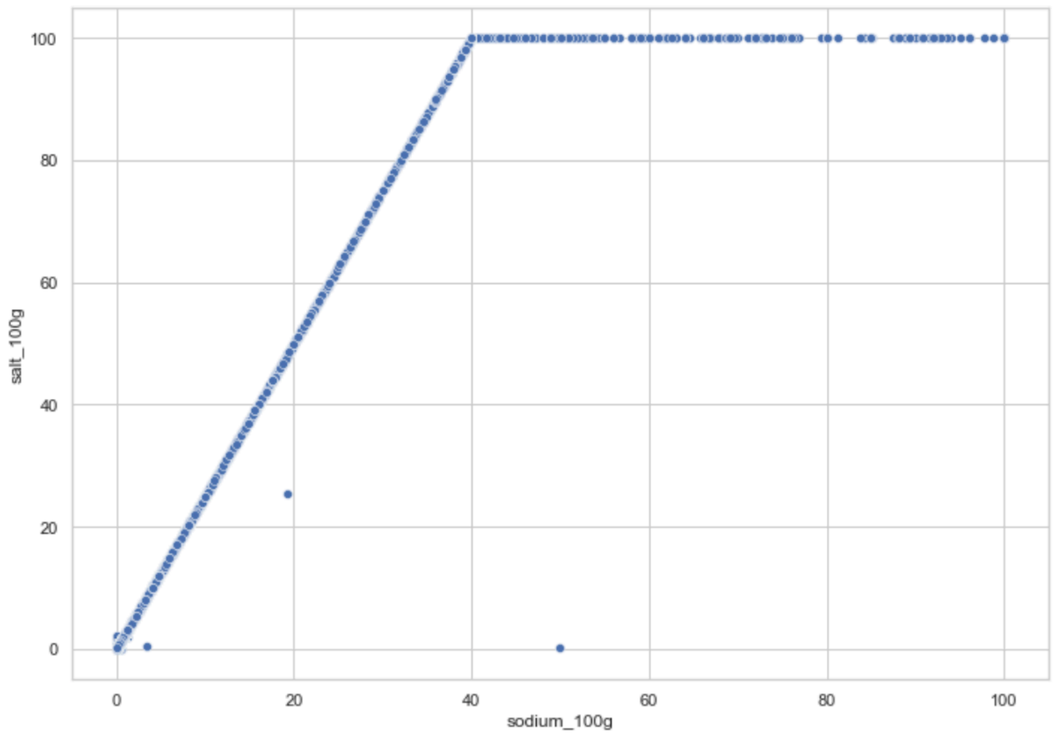
Etape 5 –

Elimination de quelques indicateurs fortement corrélés (considérés comme doublons)

- Sodium vs Sel
- energy_100g vs energy-kcal_100g

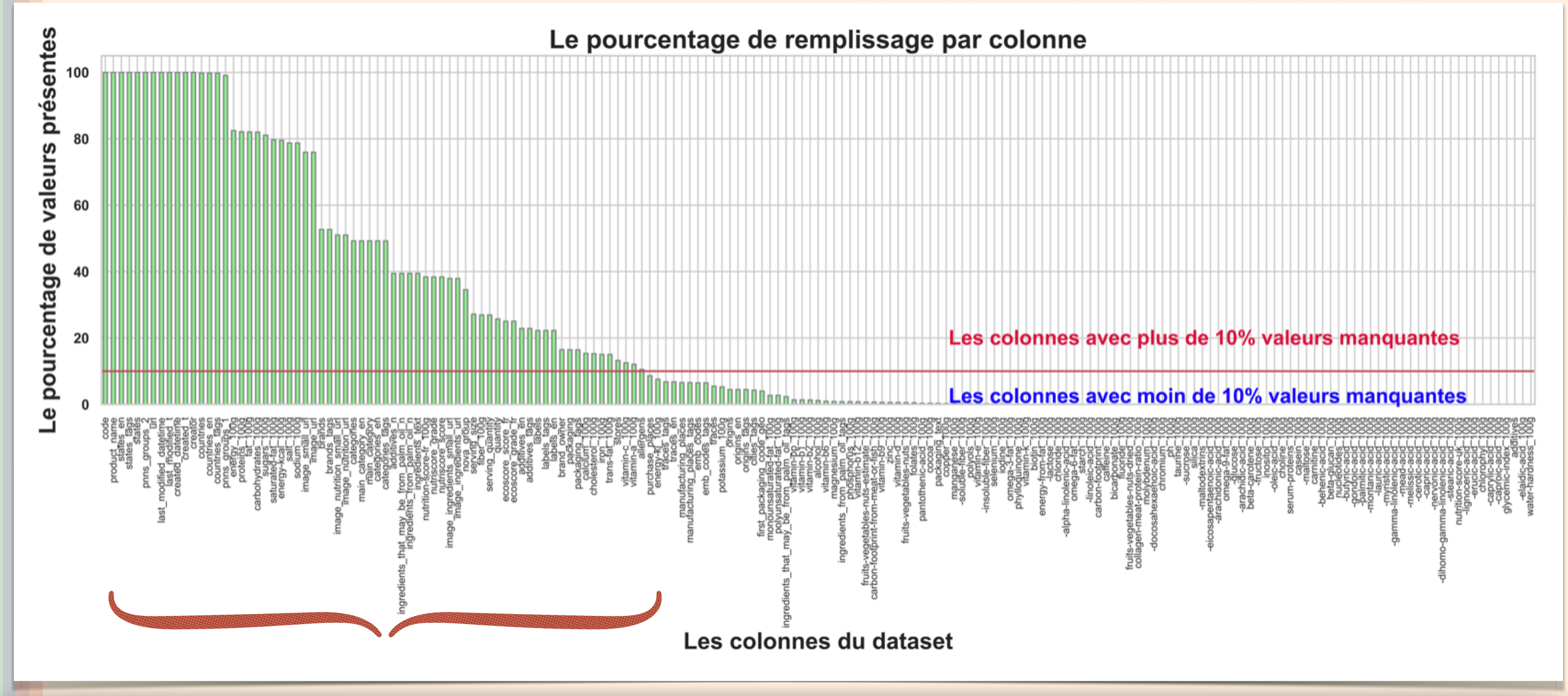
Et les valeurs redondantes (les colonnes avec un suffixe _tag et _en)

162 colonnes
1 348 519 lignes



38 758 mg sodium dans 100g du sel

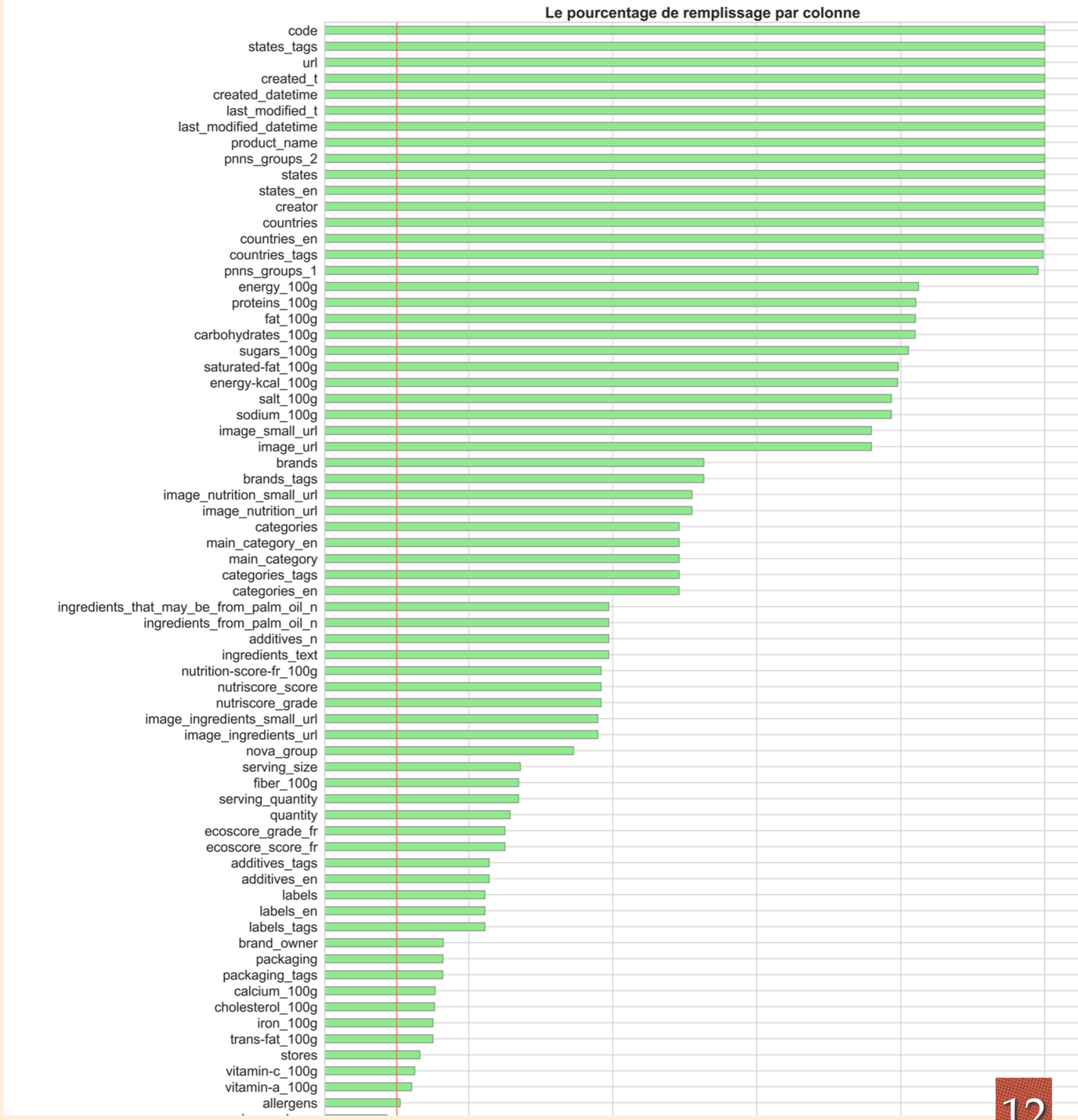
https://fdc.nal.usda.gov


Le pourcentage de remplissage par colonne
Le pourcentage de valeurs présentes
100
80
60
40
20
0
Les colonnes avec plus de 10% valeurs manquantes
Les colonnes avec moin de 10% valeurs manquantes
Les colonnes du dataset

Etape 6 –

Suppression de colonnes vides d'un taux de moins de 10 %

44 colonnes
1 348 519 lignes



Etape 7 –

Sélection des indicateurs

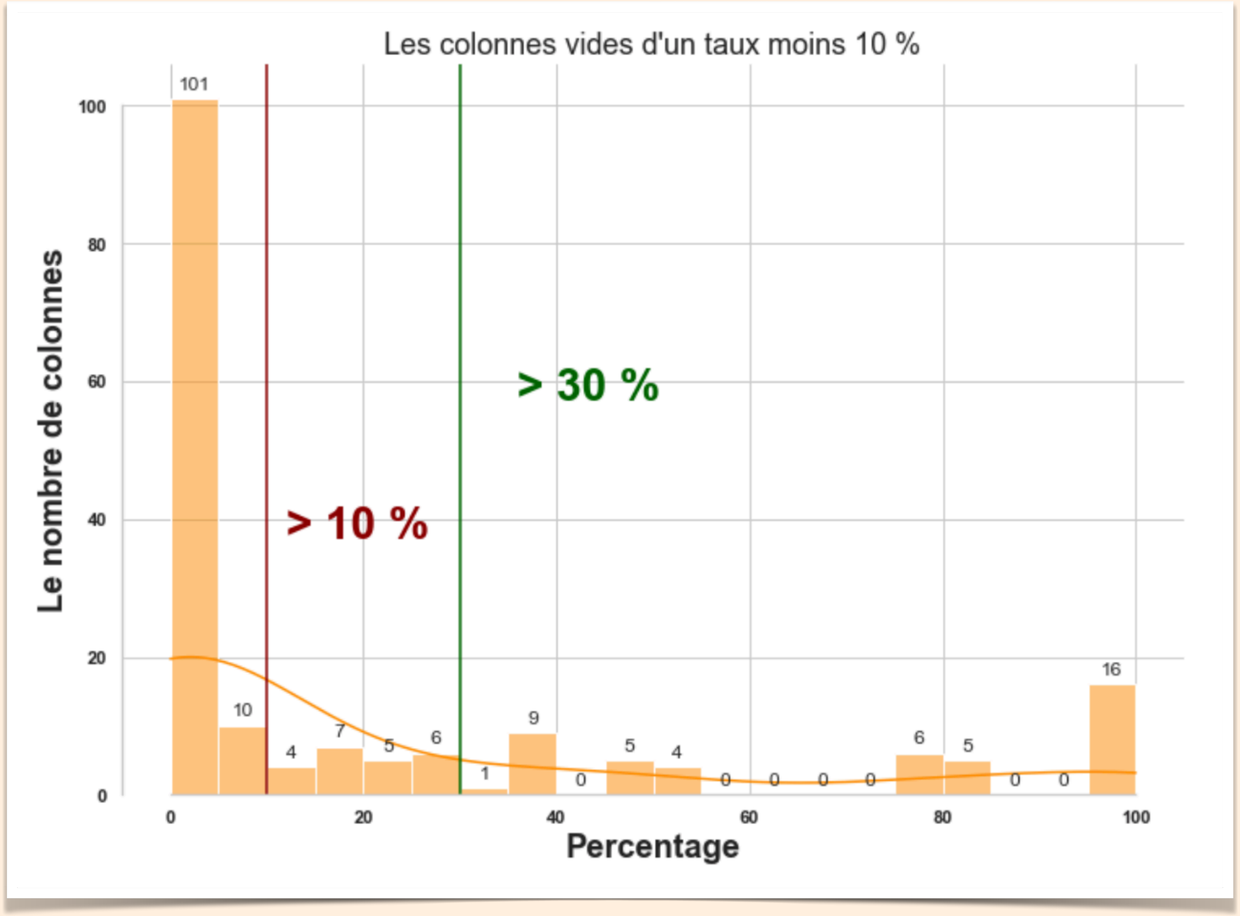
6 indicateurs pour "Ecologie" (taux de remplissage > 10 %)
- packaging
- ecoscore_grade_fr
- ecoscore_score_fr
- ingredients_from_palm_oil_n
- ingredients_that_may_be_from_palm_oil_n

Les nutriments (taux de remplissage > 30 %)
- protein_100g
- fat_100g
- saturated-fat_100g
- carbohydrates_100g
- sugars_100g
- salt_100g
- fiber_100g

Les indicateurs complémentaires (taux de remplissage > 30 %)
- energy_100g
- additives_n
- nutriscore_score
- nutriscore_grade
- nova_group

Les indicateurs catégoriels et les autres indicateurs (taux de remplissage > 30 %)
- categorie
- main_category_en
- pnns_groups_1
- pnns_groups_2
- pays
- created_datetime

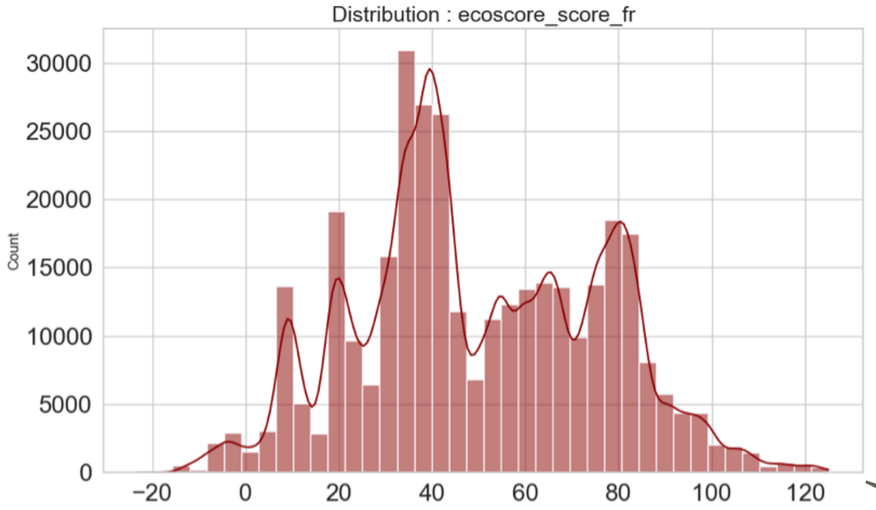
24 colonnes
1 229 441 lignes

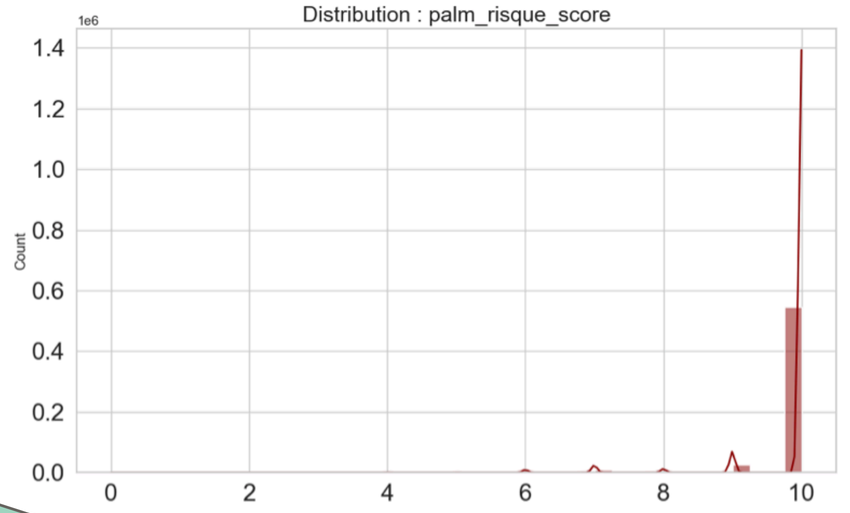


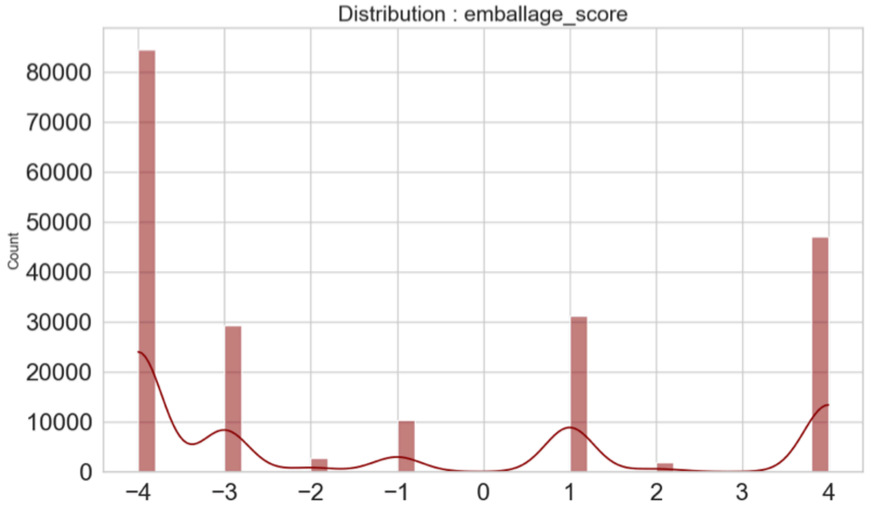
Première partie : Analyse sur l'écologie

Eco-score | Risque de l'huile de palme | Emballage score

Analyse des indicateurs





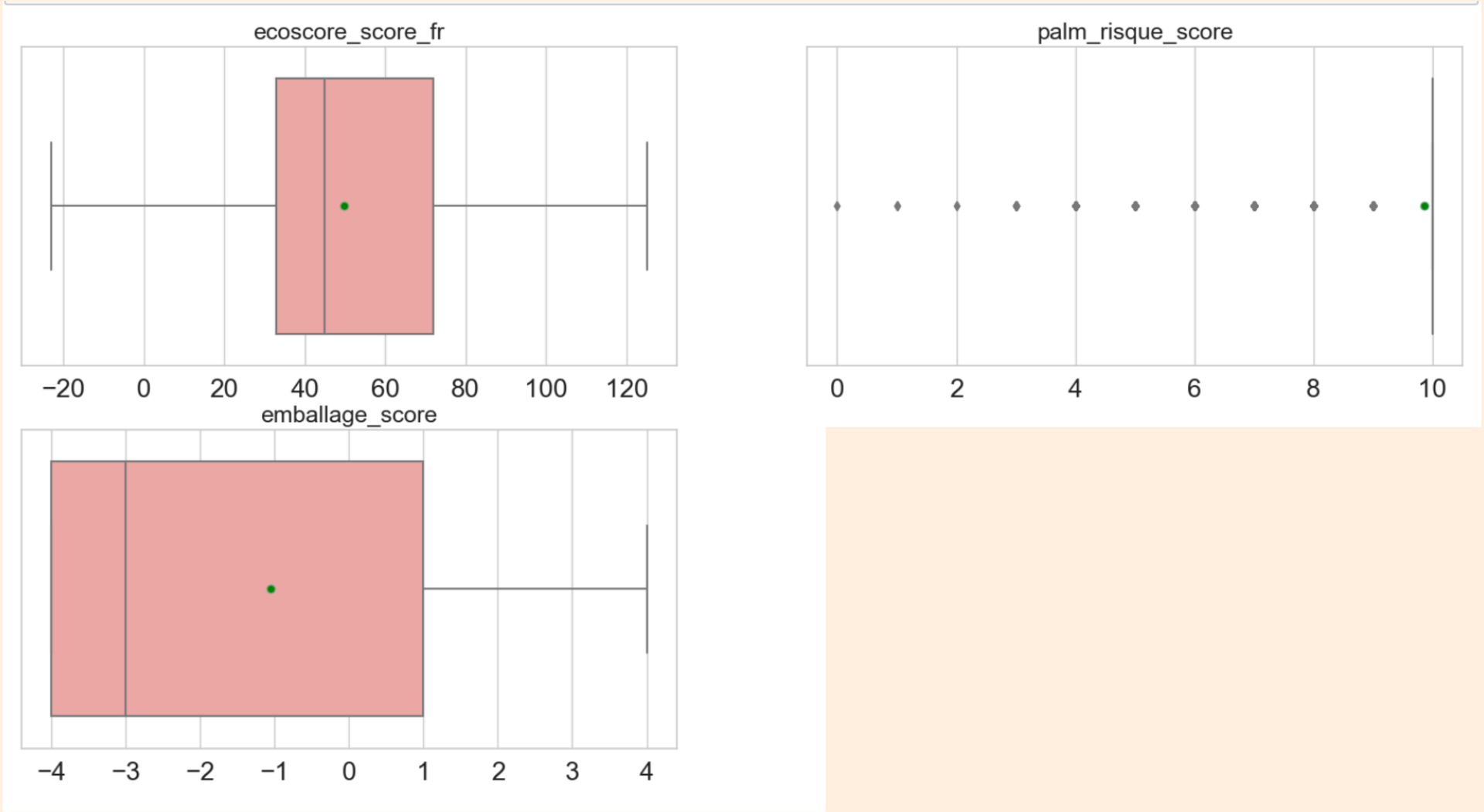


La distribution n'est pas Gaussienne

```python
# normalité test
stat, p = normaltest(df_eco_score['ecoscore_score_fr'])
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interprete
alpha = 0.05
if p > alpha:
    print('Sample looks Gaussian (fail to reject H0)')
else:
    print('Sample does not look Gaussian (reject H0)')
```

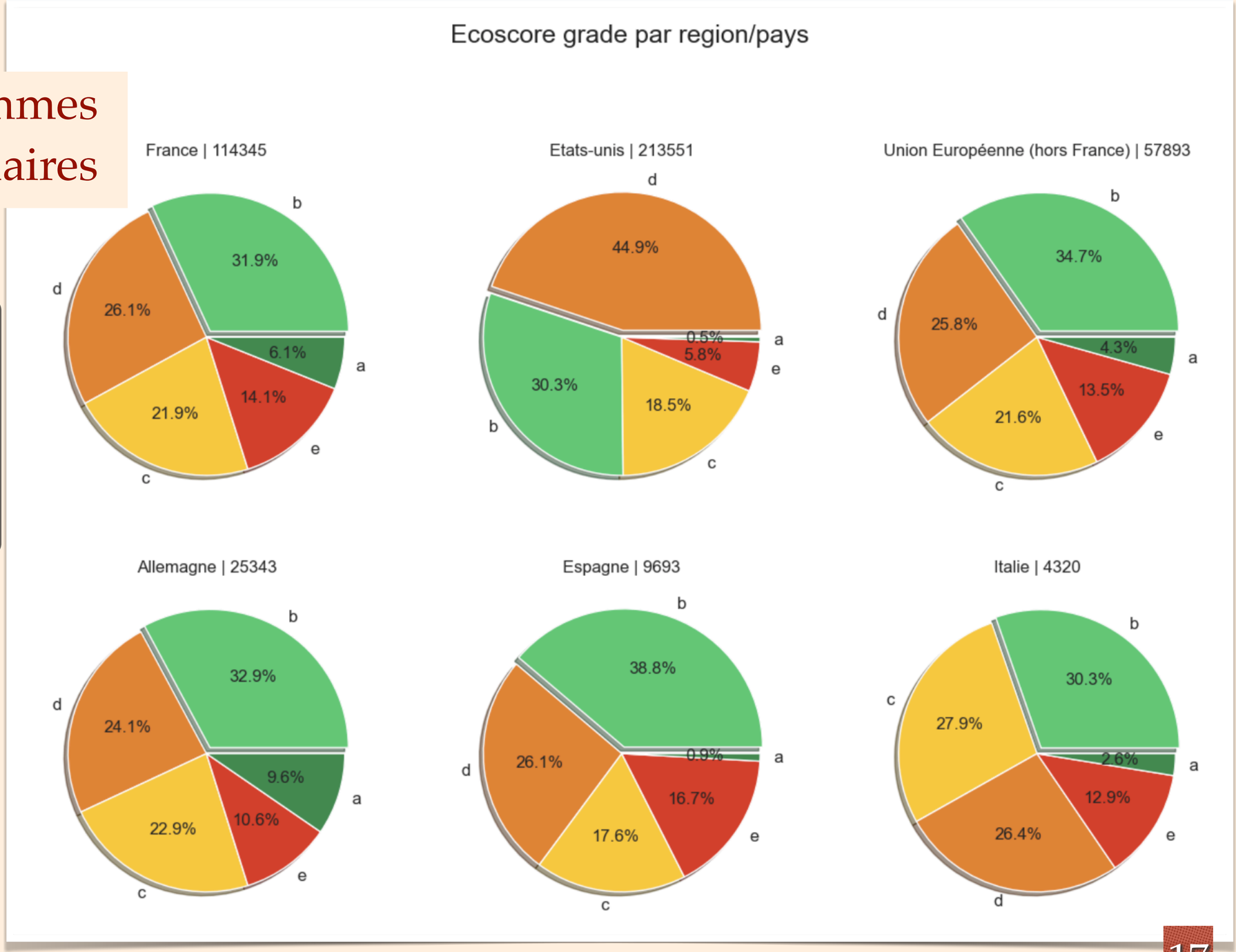
```
Statistics=nan, p=nan
Sample does not look Gaussian (reject H0)
```

Analyse du Box Plot



Visualisation des diagrammes circulaires

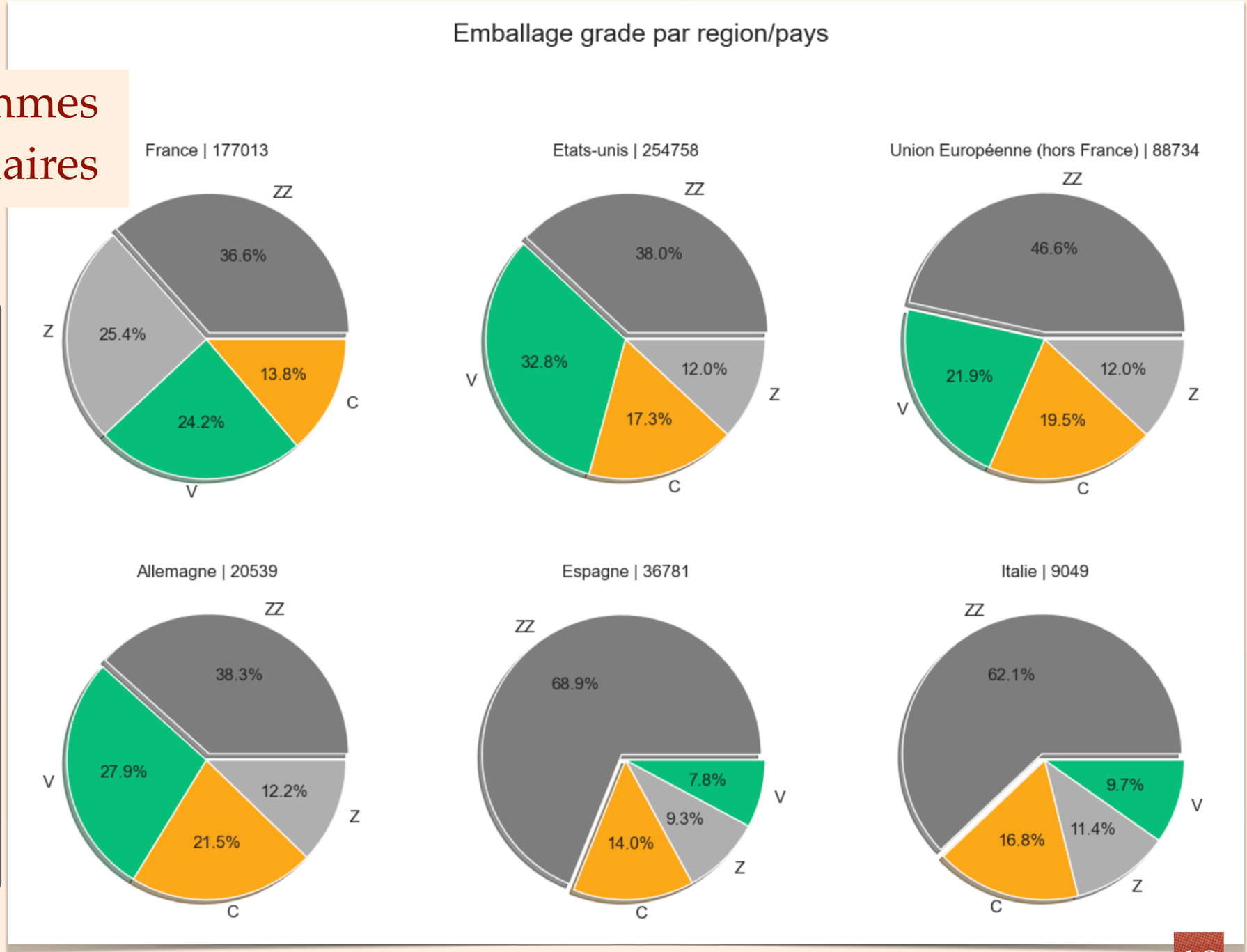
L'échelle représentant l'impact environnemental des produits alimentaires



Visualisation des diagrammes circulaires

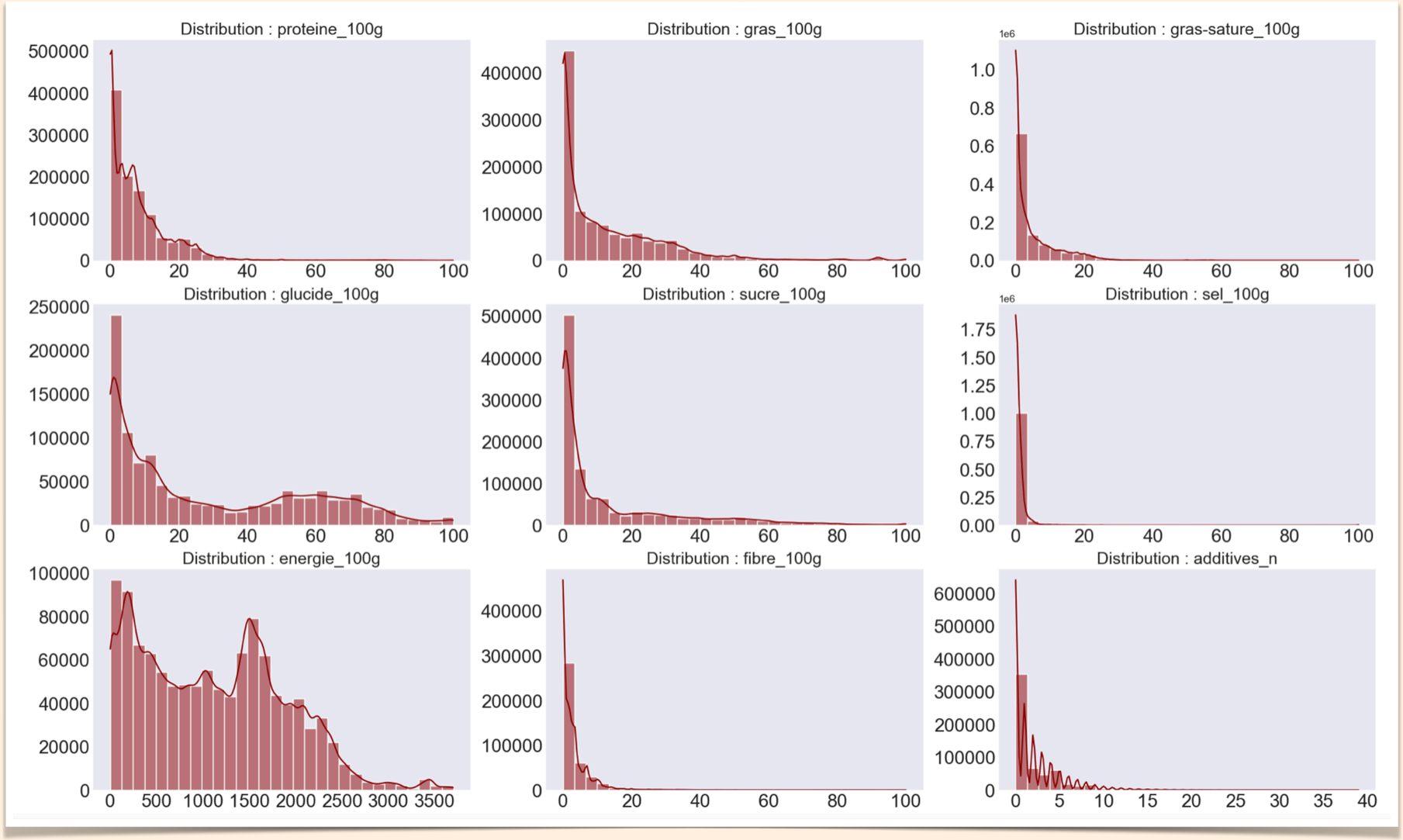
Explications

- **V** — verre
- **C** — carton
- **Z** — mixte (avec plastique)
- **ZZ** — plastique

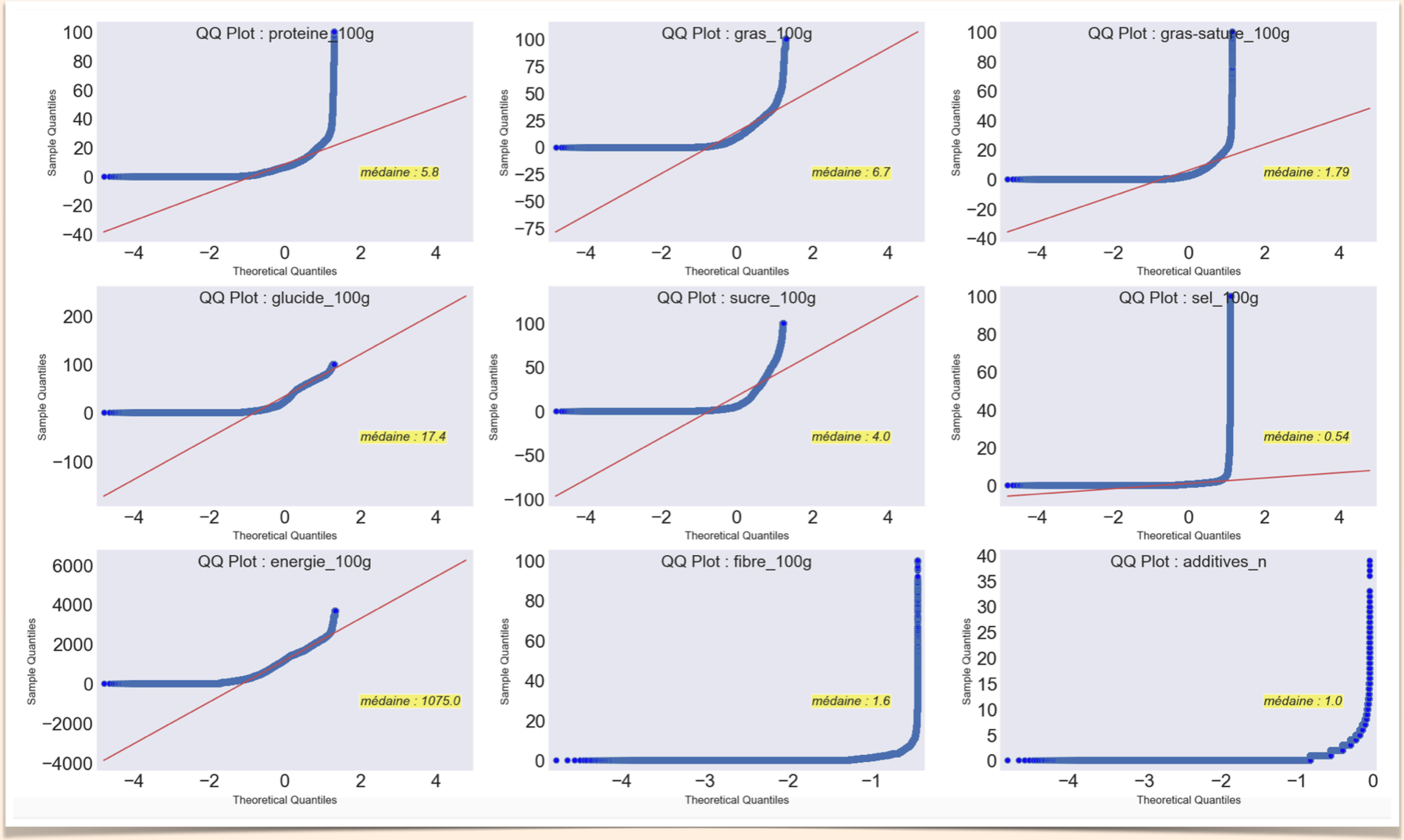


Deuxième partie : Analyse sur nutriscore

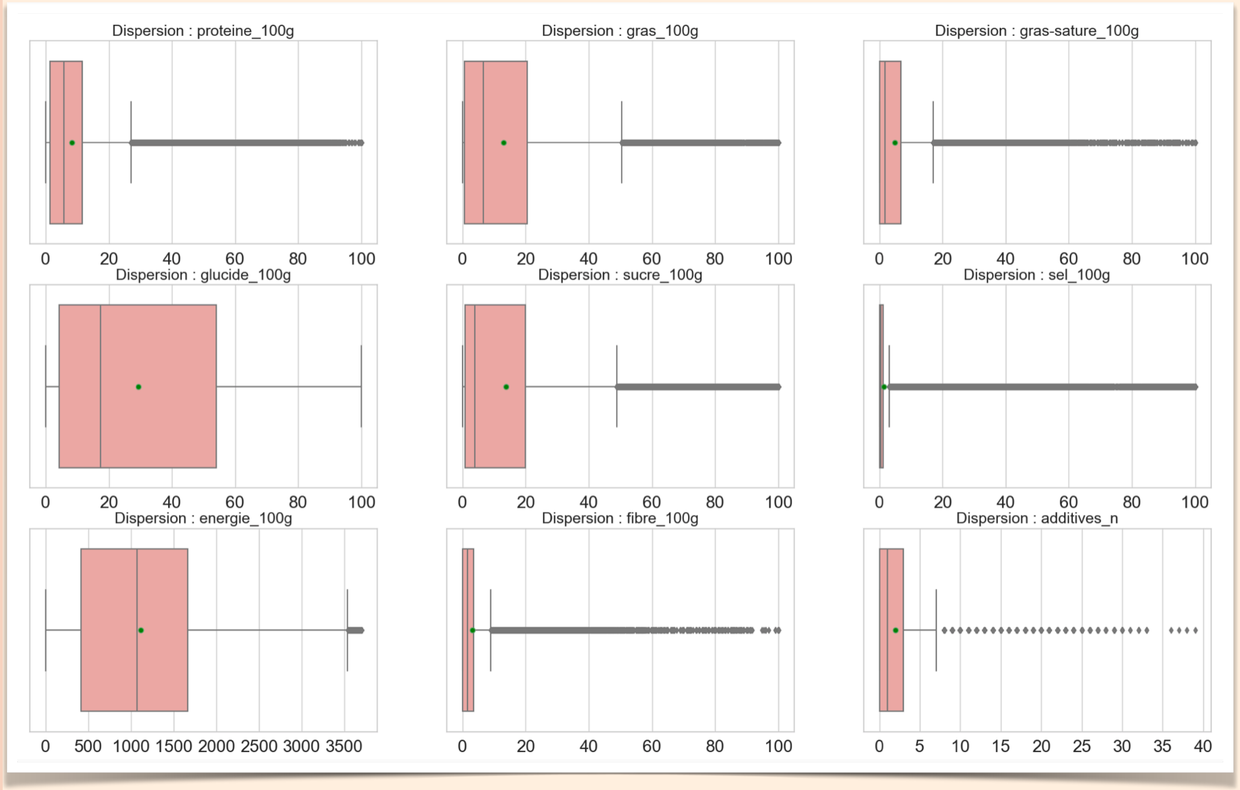
Analyse des indicateurs nutritionnels



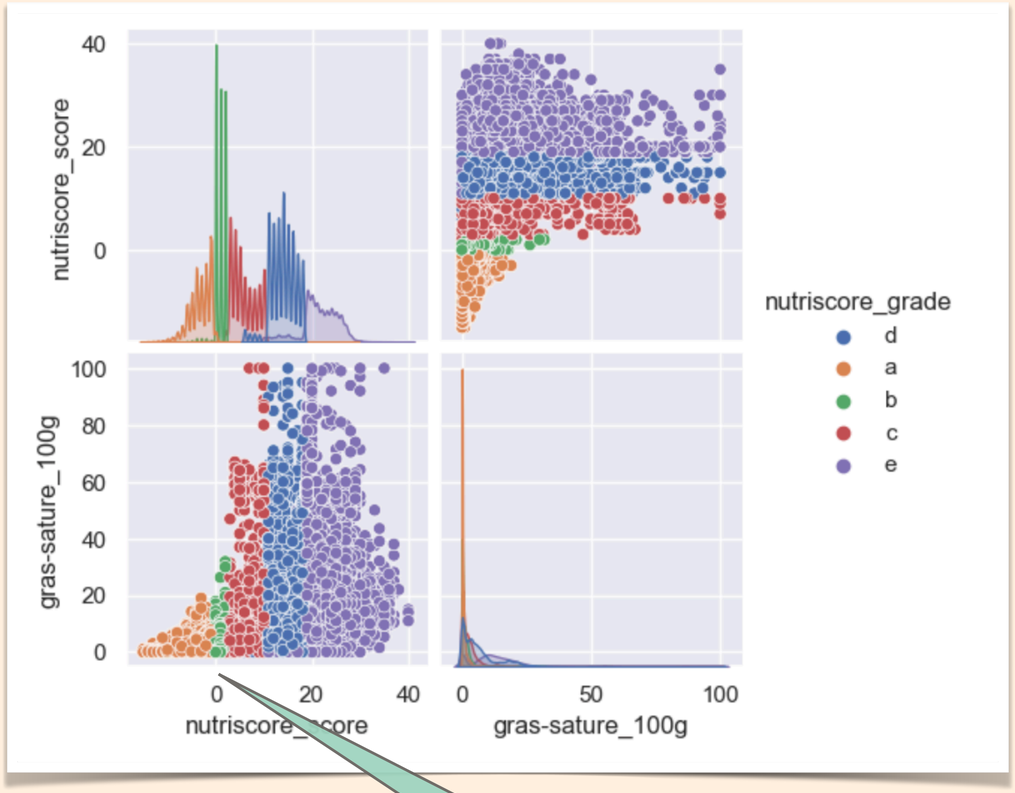
Analyse des indicateurs nutritionnels



Analyse du Box Plot



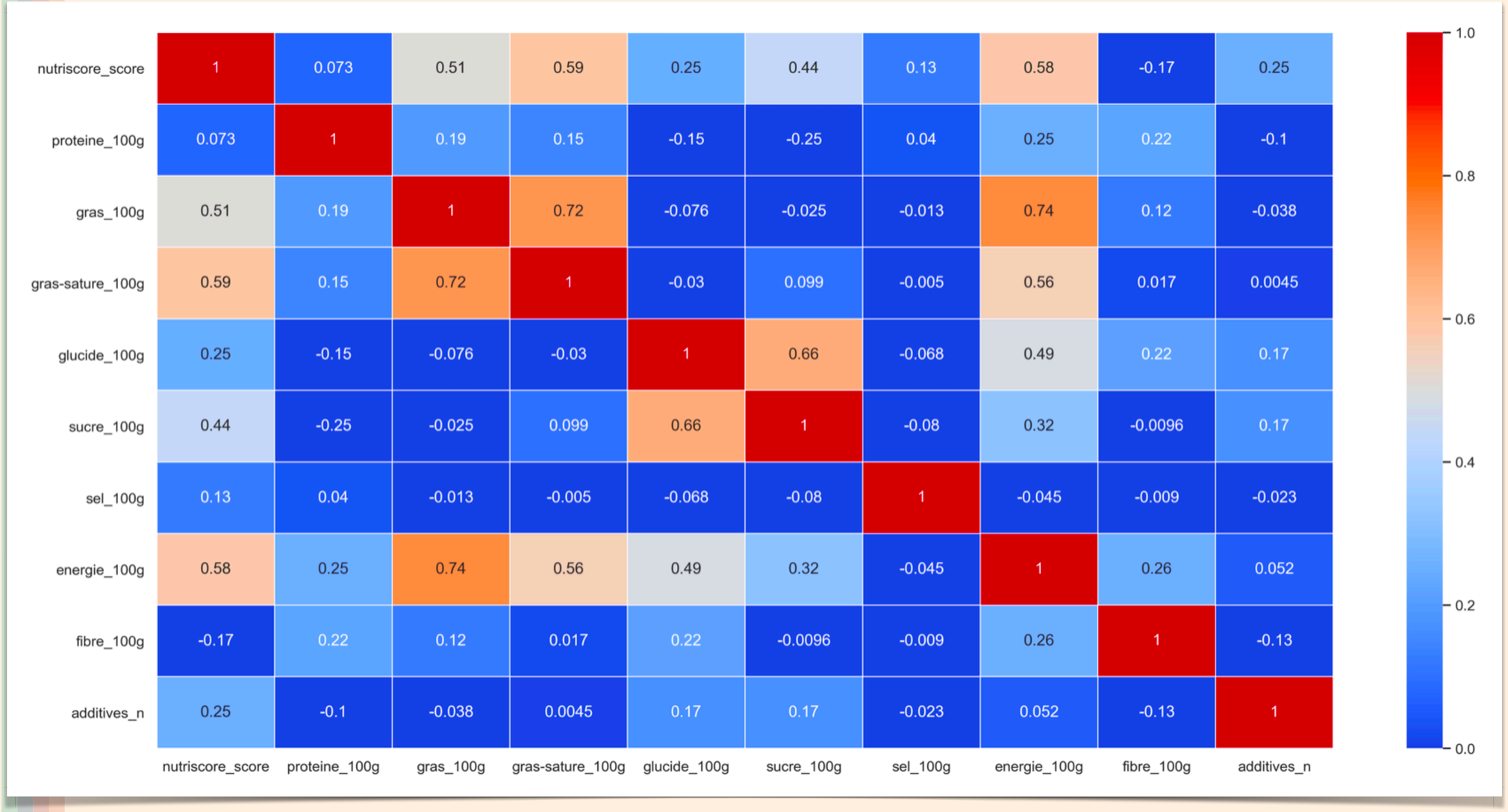
Analyse du Pair Plot



Moins il y a de gras saturé, plus le Nutriscore est faible

Recherche de corrélations

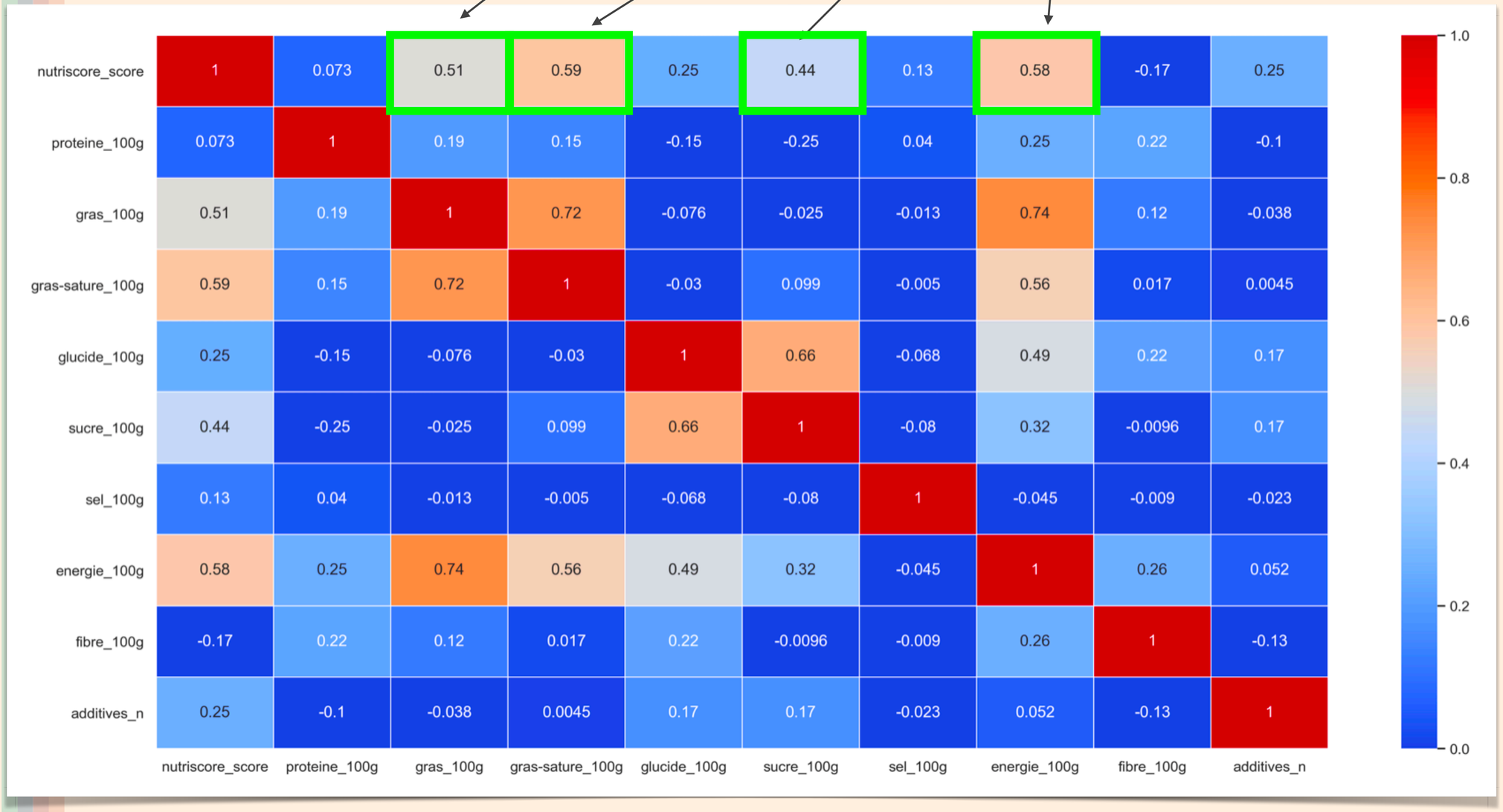
Nous cherchons à présent les corrélations possibles entre le NutriScore et les indicateurs :



La matrice de corrélation

- La **corrélation** entre deux (ou plusieurs) variables est une notion de liaison qui contredit leur indépendance.
- Utile pour voir les variables liées une à autre.
- Le coefficient de corrélation linéaire, toujours comprise entre 0 et 1, mesure le lien de corrélation entre 2 variables.
- Plus le nombre est grand (max : 1) et plus les corrélations sont représentatives. Par contre, si elle est proche de 0, cela veut dire qu'il n'y a pas de corrélation entre les deux variables

Nous constatons que Nutri-score est corrélé avec les gras, les gras saturés, le sucre et l'énergie

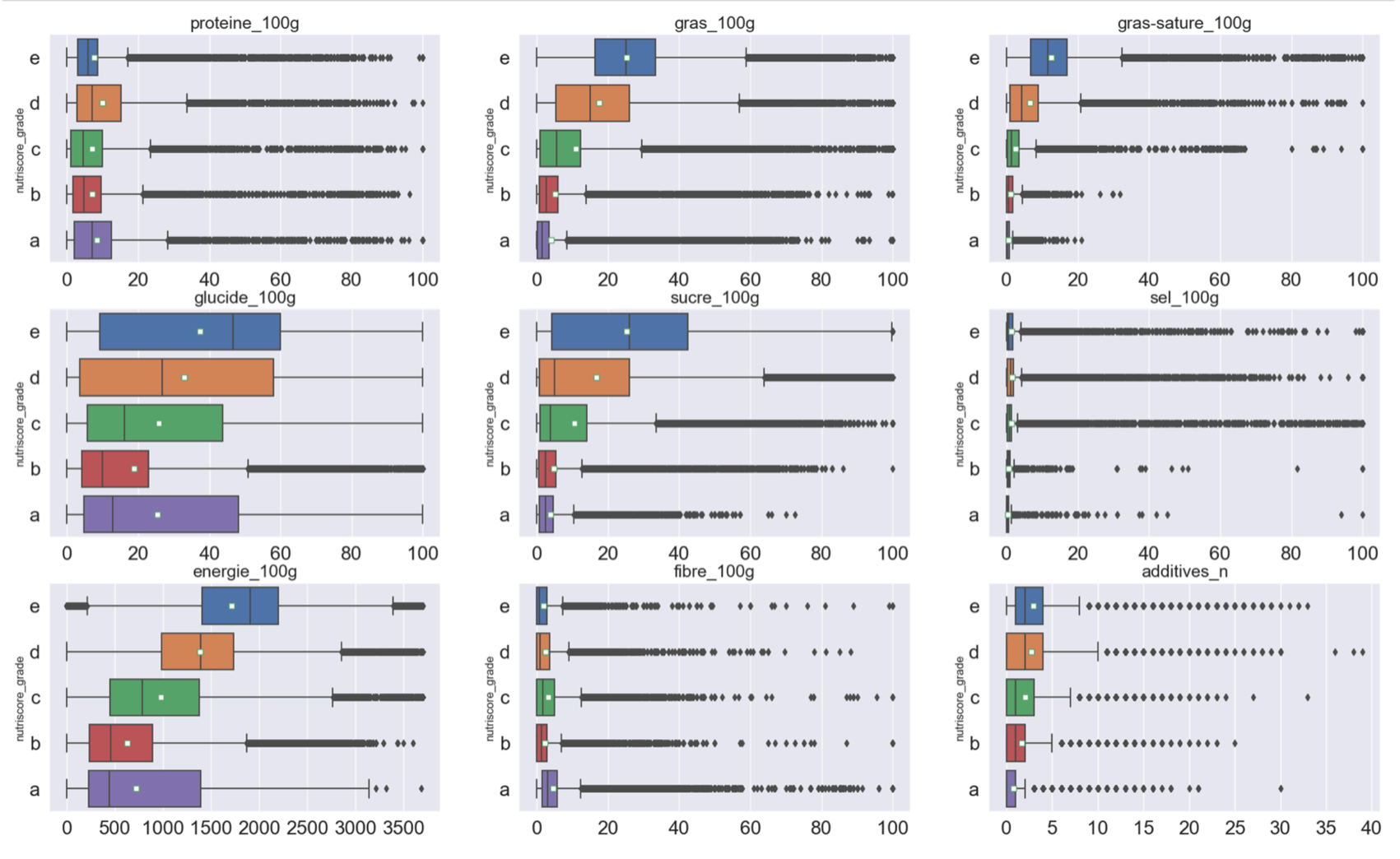


La matrice de corrélation

La méthode ANOVA

- Analyse de la variance *(anglais : **AN**alysis **O**f **VA**riance)*
- C’est une analyse de la covariance entre deux variables (quantitative et qualitative)
- Il nous permet de vérifier si les moyennes des groupes proviennent d'une même population

Analyse de la variance



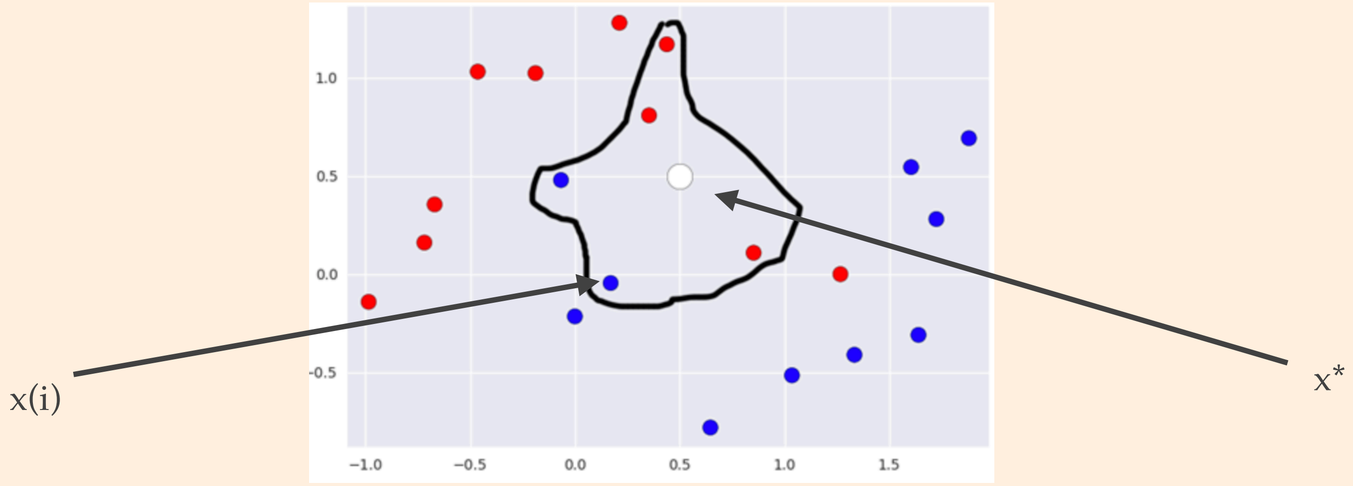
La méthode ANOVA

- Nous constatons que le NutriScore semble dépendre des gras, des gras saturés, des sucres, de l'énergie.
- Par contre, les glucides, les protéines, les fibres et le sel ne semblent pas avoir de lien avec le Nutriscore.
- Ces différentes analyses nous indique que pour construire un modèle de prédiction, nous pourrons nous appuyer sur les indicateurs suivants :
  - L'énergie
  - Les gras
  - Les gras saturés
  - Les sucres

La méthode KNNImputer

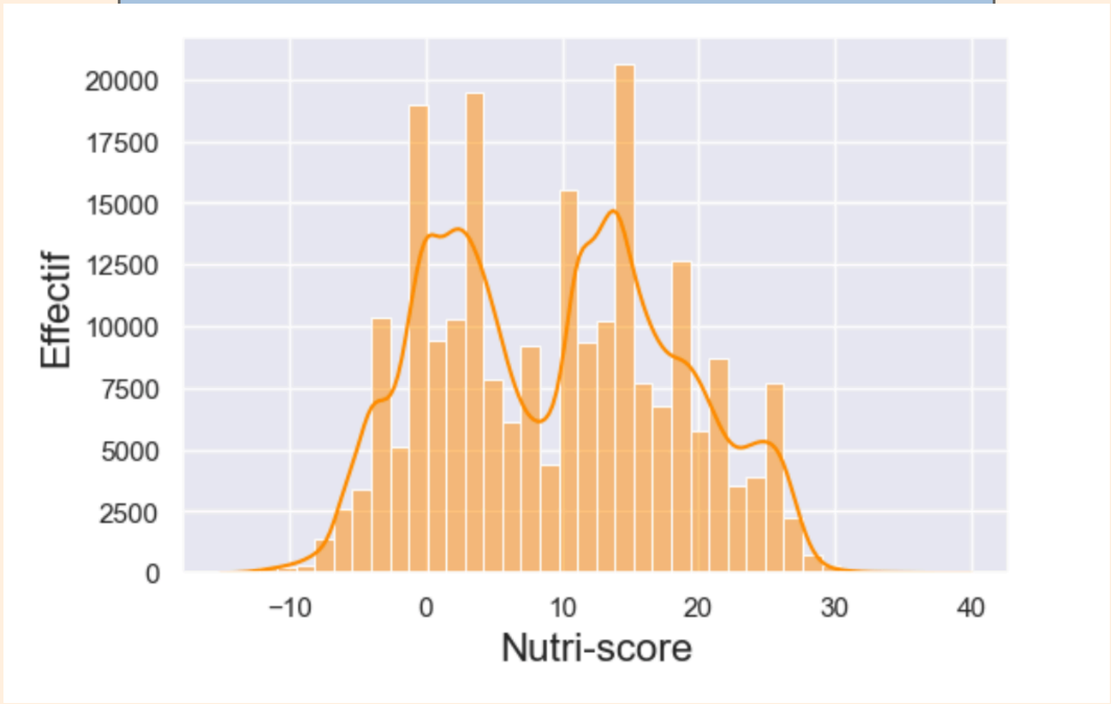
- La complétion par k plus proches voisins (*k-nearest neighbors* ou KNN) consiste à exécuter l'algorithme suivant qui modélise et prévoit les données manquantes.

- **<u>Algorithme des k plus proches voisins (k-nn)</u>**
- Choix d'un entier k
- Calculer les distances d(x*,x(i)), i=1,…,n
- Retenir les k observations x(i_1) , . . . , x(i_k) pour lesquelles ces distances sont les plus petites.
- Affecter aux valeurs manquantes la moyenne des valeurs des k voisins



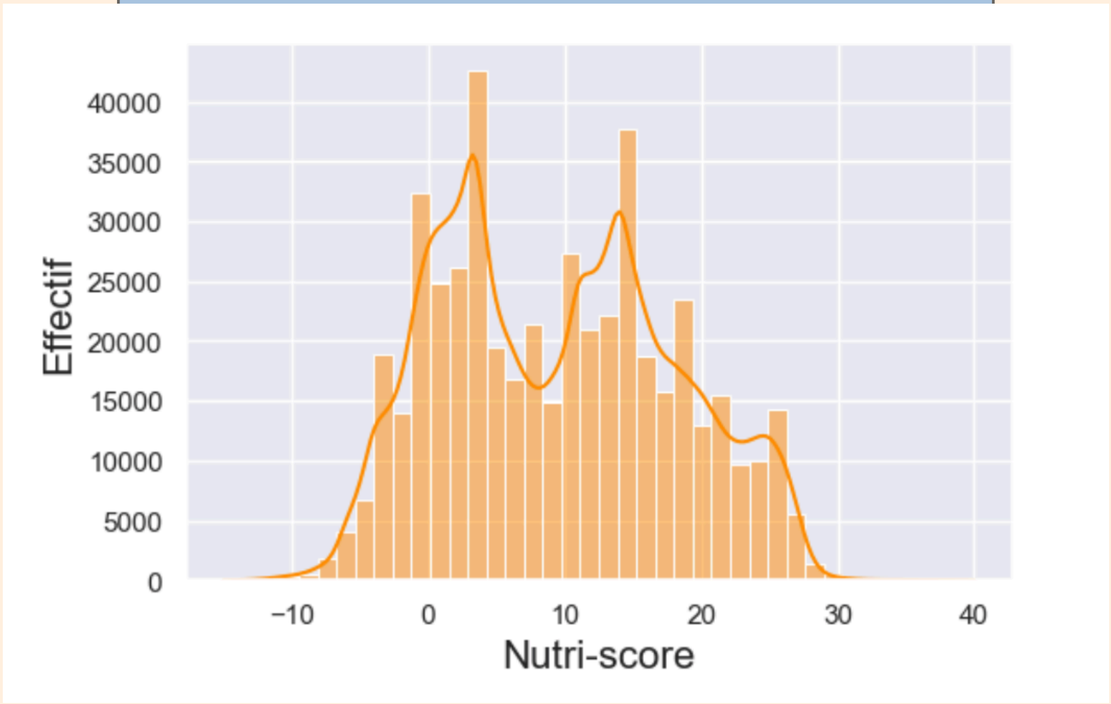
Prédiction du NutriScore

La distribution de NutriScore avant imputation



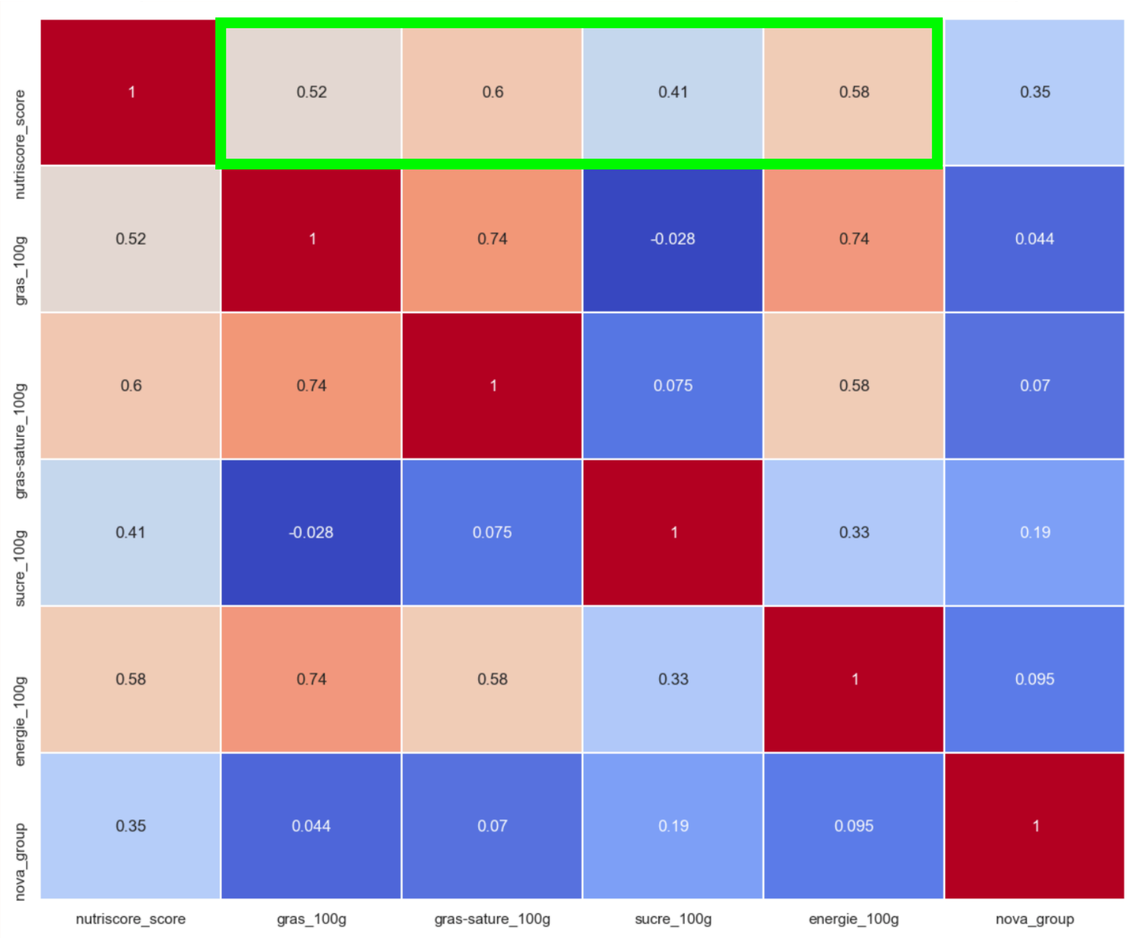

```
moyenne avant: 9.429067262427193
mediane avant: 10.0
std avant: 8.828784403178467
skw avant: 0.11560477812197563
kur avant: -0.935612659072417
```

La distribution de NutriScore après imputation




```
moy après: 9.427260159932828
med après: 9.333333333333334
std après: 8.605508394364007
skw après: 0.18722768368173523
kur après: -0.9207164670775105
```

La matrice de corrélation avant imputation

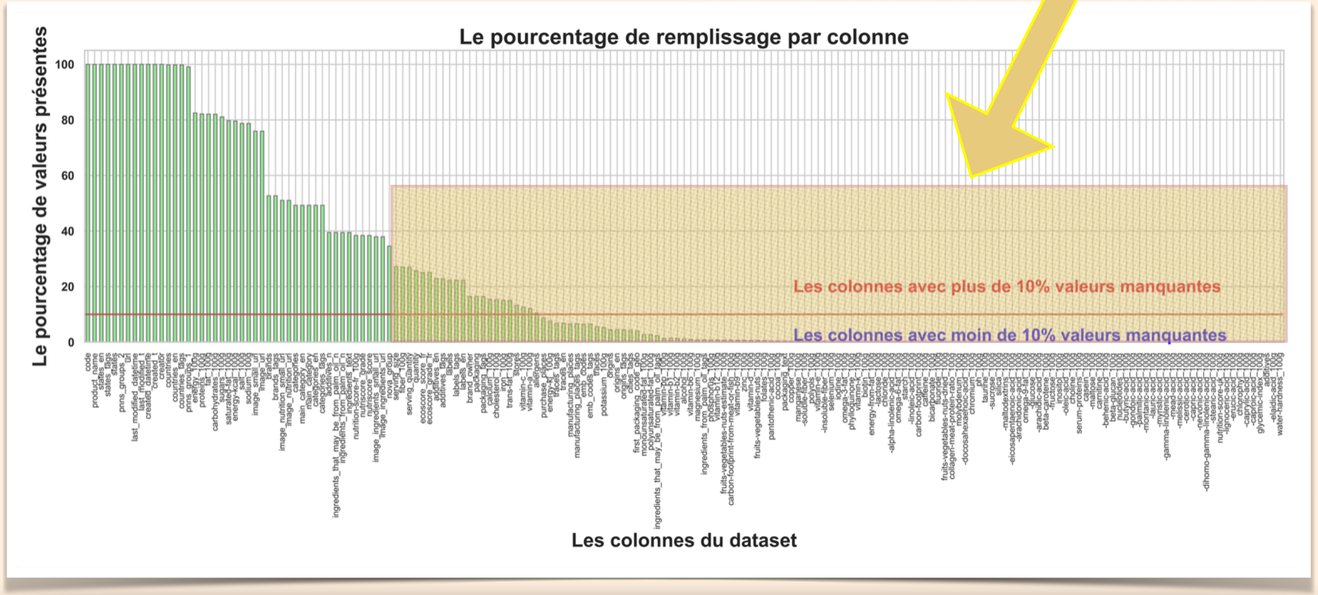


La matrice de corrélation après imputation

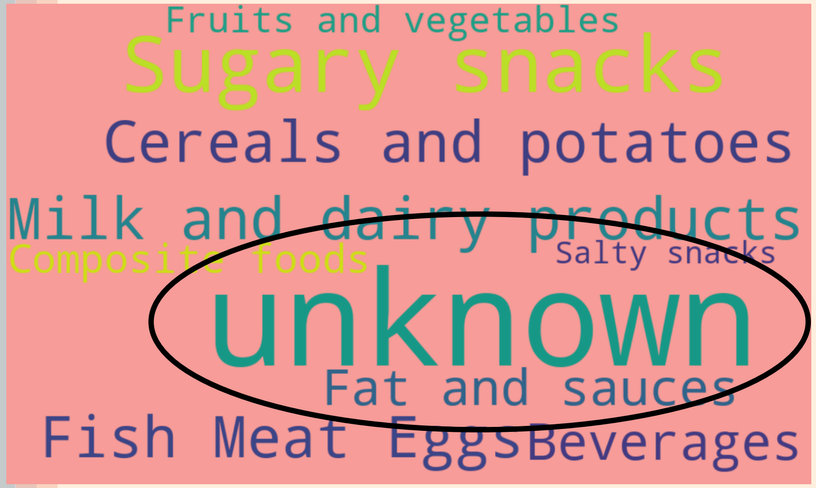
Conclusion

	column_name	nb_manquant	nb_present	Taux de remplissage
0	allergens	1191838	156681	11.618746
1	vitamin-a_100g	1172824	175695	13.028737
2	vitamin-c_100g	1166776	181743	13.477229
3	stores	1142242	206277	15.296559
4	trans-fat_100g	1137805	210714	15.625586
5	iron_100g	1137065	211454	15.680461
6	cholesterol_100g	1134952	213567	15.837152
7	calcium_100g	1132606	215913	16.011120
8	brand_owner	1122471	226048	16.762686
9	packaging_tags	1089537	258982	19.204920
10	packaging	1089508	259011	19.207071
11	labels_en	1019960	328559	24.364432
12	ecoscore_score_fr	1008716	339803	25.198236
13	ecoscore_grade_fr	1008716	339803	25.198236
14	additives_en	1007200	341	25.310656
15	serving_quantity	954541	393	29.215606
16	serving_size	950491	398028	15936
17	quantity	947073	401446	29.769
18	fiber_100g	946230	402289	29.831912

Rappel : la plupart des données est manquante et mal renseignée. Cela pose un problème pour obtenir une analyse de NutriScore cohérente.



Rappel : le pourcentage de remplissage pour les indicateurs retenus dans le calcul de l'emballage score pour l'écologie est inférieur à 30 %.



	pnns_groups_1
unknown	681060
sugary snacks	122061
milk and dairy products	67895
cereals and potatoes	66893
fish meat eggs	66083
beverages	51283

	pnns_groups_2
unknown	681060
sweets	53864
biscuits and cakes	51662
dressings and sauces	38231
one-dish meals	34890
cereals	32100

	categorie
Snacks	25401
Snacks,Sweet snacks,Biscuits and cakes,Biscuits	12471
Groceries,Sauces	12372
Snacks,Sweet snacks,Confectioneries	12160
Dairies,Fermented foods,Fermented milk products,Cheeses	11353

Le problème des catégories alimentaires

Faisabilité du projet

- Les données manquantes posent un réel problème d'analyse des alertes sur l'écologie.
- Les prédictions du Nutriscore sont encourageantes avec la méthode k-NN qui a encore un potentiel d'amélioration avec un paramétrage plus approfondi.
- Il existe d'autres modèles que nous n'avons pas encore testés ici.