

Projet 3 : conception d'une application au service de la santé publique

- Projet s'inscrivant dans le cadre d'un appel à projet lancé par l'agence Santé publique France



- Objectif : Trouver une idée d'application en lien avec l'alimentation en partant du jeu de données Open Food Facts, disponible sur le site officiel <https://world.openfoodfacts.org>

Présentation du jeu de données

- Taille du dataset

- 320772 lignes correspondant aux individus que sont les produits alimentaires.

(Exemple: Filet de bœuf, Farine de blé noir, Lion Peanut x2 etc.)

- 62 colonnes correspondant aux variables que sont les caractéristiques qui définissent les produits.

(Exemple : code, product name, categories, salt_100g, nutrition-score-fr_100g, sugars_100g, vitamin-c_100g etc.)

- Taux de remplissage du tableau

- On a 76 % de valeurs manquantes

Plan d'étude

- I. Présentation de l'idée d'application
- II. Choix des variables utiles
- III. Nettoyage des données
- IV. Analyse univariée
- V. Analyse bivariée
- VI. Analyse multivariée : analyse en composantes principales(ACP)
- VII. Conclusions

Idée d'application

- Trouver le nutriscore correspondant à un produit à partir de la quantité de nutriments qu'il contient (les protéines, les sucres, les fibres, les lipides etc.)
- Objectif: aider le consommateur à identifier les produits de meilleure qualité.







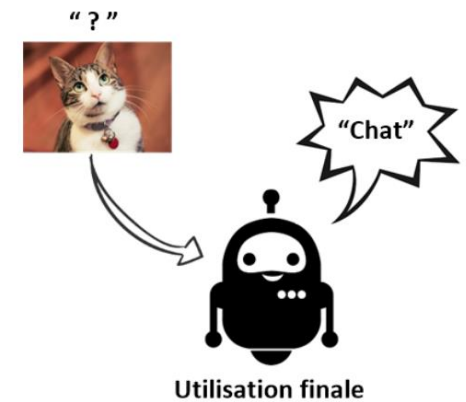
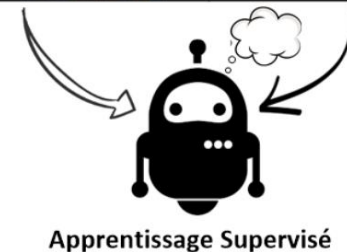
- Prédiction du nutriscore par un modèle de machine Learning (apprentissage automatique).

Idée d'application

- Utilisation de l'apprentissage supervisé: la machine peut apprendre à faire une tâche en étudiant un grand nombre d'exemples, avec 2 types de variables :

- **Une variable objectif (target) y**
(ici c'est le nutriscore)
- **des variables caractéristiques**
(features) x
(ici ce sont les nutriments)

x	y
	"Chien"
	"Chien"
	"Chat"
	"Chien"



Choix des variables

Variables caractéristiques (features)

x_1	x_2	x_3	...	x_n
$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$...	$x_n^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$...	$x_n^{(2)}$
$x_1^{(3)}$	$x_2^{(3)}$	$x_3^{(3)}$...	$x_n^{(3)}$
...
$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$...	$x_n^{(m)}$

energy_100g, fat_100g, saturated-fat_100g,
carbohydrates_100g, sugars_100g, salt_100g,
sodium_100g, proteins_100g



Objectif : Essayer de déterminer f

$$f(x) = y ??$$

variable cible

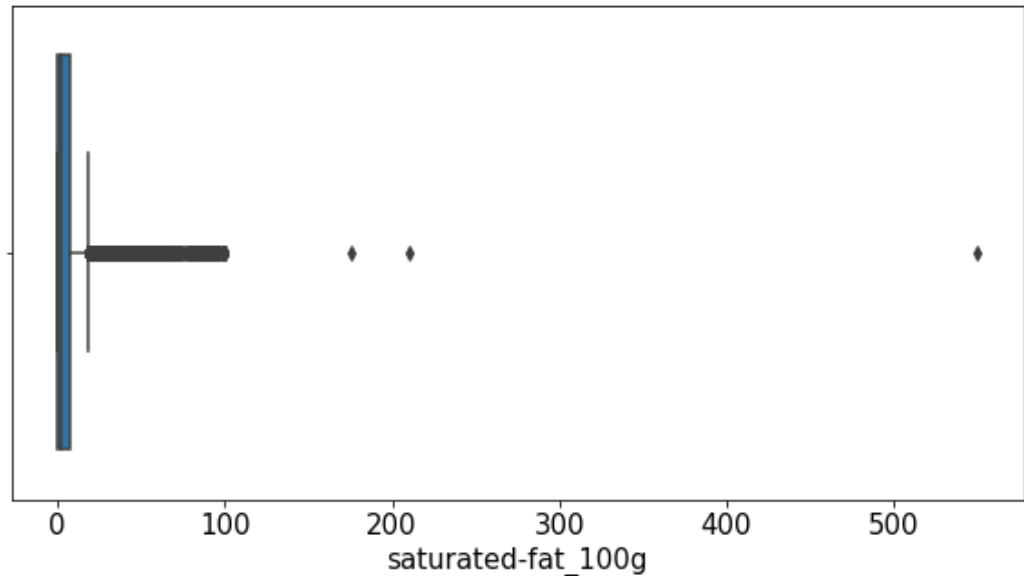
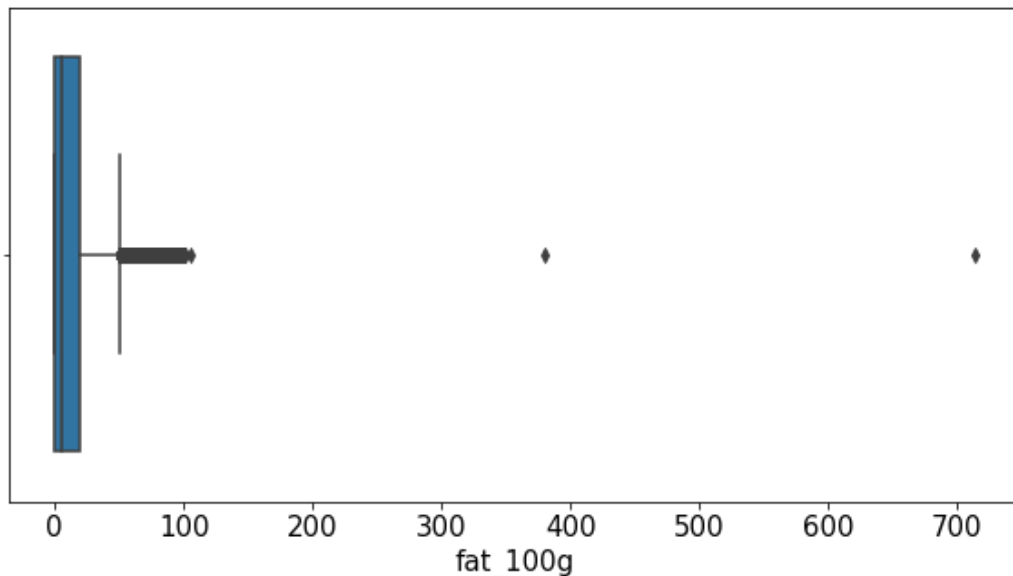
y
$y^{(1)}$
$y^{(2)}$
$y^{(3)}$
...
$y^{(m)}$



nutrition-score-fr_100g

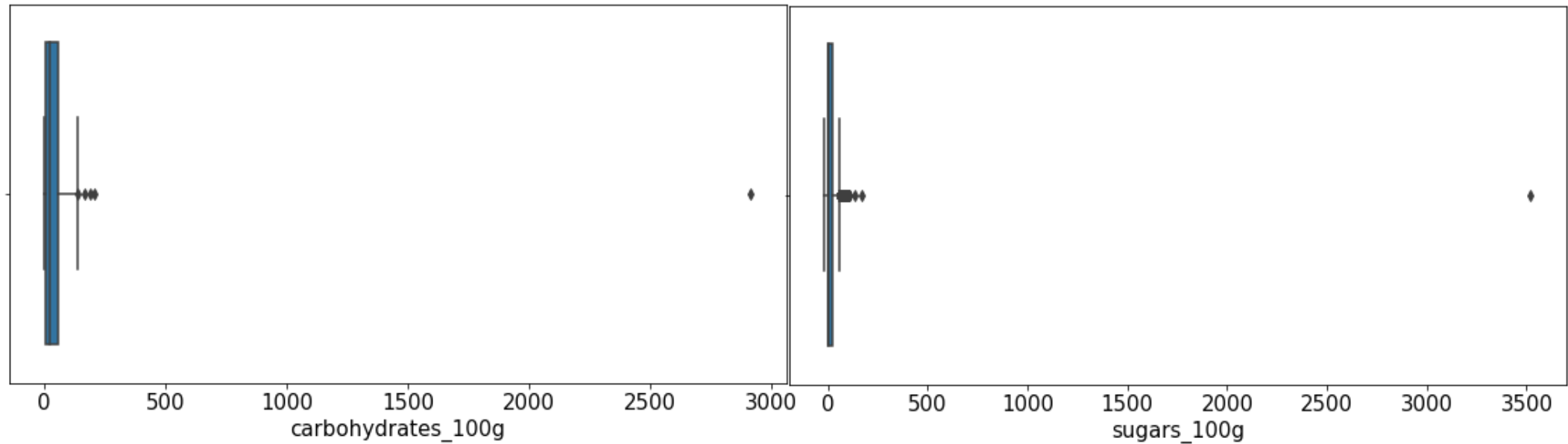
Nettoyage des données

- Pas de doublons dans le tableau de données
- Suppression des valeurs aberrantes pour une catégorie de variables :
 - **Outliers** =valeurs négatives et supérieures à 100 pour les variables : fat_100g, saturated-fat_100g, carbohydrates_100g, sugars_100g, salt_100g, sodium_100g, proteins_100g.



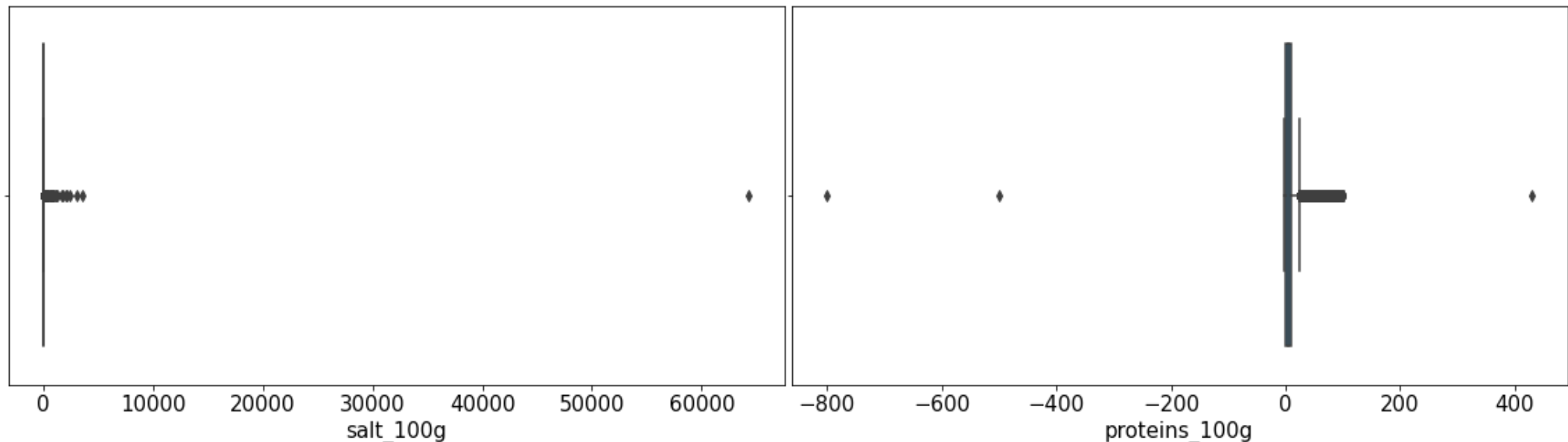
Nettoyage des données

- Suppression des valeurs aberrantes

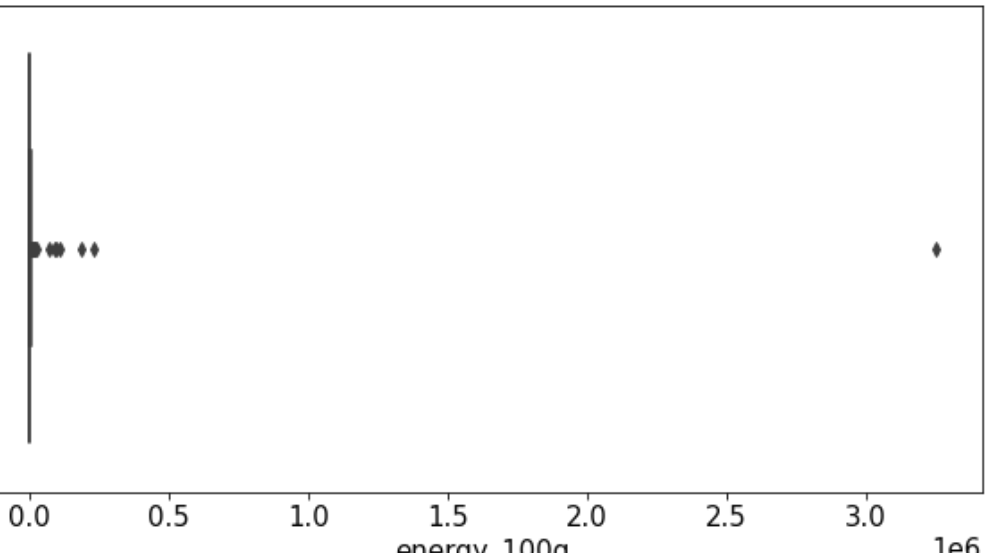


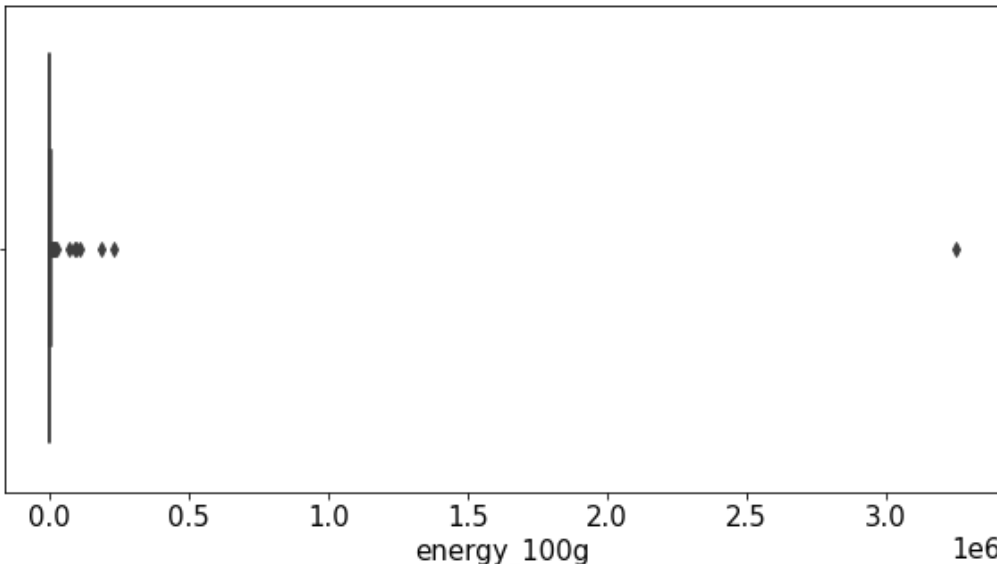
Nettoyage des données

- Suppression des valeurs aberrantes



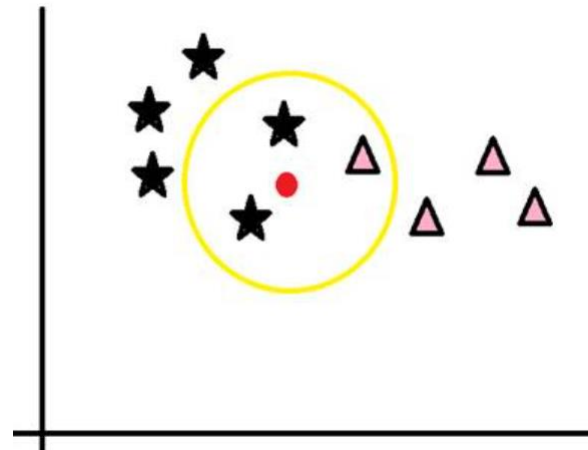
Nettoyage des données

- Suppression de la valeur aberrante pour l'énergie, celle des pois cassés.
 - L'énergie des pois cassés, c'est autour de 1.4 KJ pour les 100g et non 3.251 e6 J. C'est donc une valeur aberrante. (remplacement par la valeur maximale c'est à dire 1.431 Kj)
- 
- The scatter plot shows energy values for different masses. The x-axis is labeled 'energy_100g' and ranges from 0.0 to 3.0 with a multiplier of 1e6. The y-axis is unlabeled but has a vertical line at 0.0. Data points are clustered near the origin, with one outlier at approximately 3.251e6 J.
- | Mass (g) | Energy (J) |
|----------|------------|
| 100 | ~0.0 |
| 200 | ~0.0 |
| 300 | ~0.0 |
| 400 | ~0.0 |
| 500 | ~0.0 |
| 600 | ~0.0 |
| 700 | ~0.0 |
| 800 | ~0.0 |
| 900 | ~0.0 |
| 1000 | ~0.0 |
| 1100 | ~0.0 |
| 1200 | ~0.0 |
| 1300 | ~0.0 |
| 1400 | ~0.0 |
| 1500 | ~0.0 |
| 1600 | ~0.0 |
| 1700 | ~0.0 |
| 1800 | ~0.0 |
| 1900 | ~0.0 |
| 2000 | ~0.0 |
| 2100 | ~0.0 |
| 2200 | ~0.0 |
| 2300 | ~0.0 |
| 2400 | ~0.0 |
| 2500 | ~0.0 |
| 2600 | ~0.0 |
| 2700 | ~0.0 |
| 2800 | ~0.0 |
| 2900 | ~0.0 |
| 3000 | ~0.0 |
| 3100 | ~0.0 |
| 3200 | ~0.0 |
| 3300 | ~0.0 |
| 3400 | ~0.0 |
| 3500 | ~0.0 |
| 3600 | ~0.0 |
| 3700 | ~0.0 |
| 3800 | ~0.0 |
| 3900 | ~0.0 |
| 4000 | ~0.0 |
| 4100 | ~0.0 |
| 4200 | ~0.0 |
| 4300 | ~0.0 |
| 4400 | ~0.0 |
| 4500 | ~0.0 |
| 4600 | ~0.0 |
| 4700 | ~0.0 |
| 4800 | ~0.0 |
| 4900 | ~0.0 |
| 5000 | ~0.0 |
| 5100 | ~0.0 |
| 5200 | ~0.0 |
| 5300 | ~0.0 |
| 5400 | ~0.0 |
| 5500 | ~0.0 |
| 5600 | ~0.0 |
| 5700 | ~0.0 |
| 5800 | ~0.0 |
| 5900 | ~0.0 |
| 6000 | ~0.0 |
| 6100 | ~0.0 |
| 6200 | ~0.0 |
| 6300 | ~0.0 |
| 6400 | ~0.0 |
| 6500 | ~0.0 |
| 6600 | ~0.0 |
| 6700 | ~0.0 |
| 6800 | ~0.0 |
| 6900 | ~0.0 |
| 7000 | ~0.0 |
| 7100 | ~0.0 |
| 7200 | ~0.0 |
| 7300 | ~0.0 |
| 7400 | ~0.0 |
| 7500 | ~0.0 |
| 7600 | ~0.0 |
| 7700 | ~0.0 |
| 7800 | ~0.0 |
| 7900 | ~0.0 |
| 8000 | ~0.0 |
| 8100 | ~0.0 |
| 8200 | ~0.0 |
| 8300 | ~0.0 |
| 8400 | ~0.0 |
| 8500 | ~0.0 |
| 8600 | ~0.0 |
| 8700 | ~0.0 |
| 8800 | ~0.0 |
| 8900 | ~0.0 |
| 9000 | ~0.0 |
| 9100 | ~0.0 |
| 9200 | ~0.0 |
| 9300 | ~0.0 |
| 9400 | ~0.0 |
| 9500 | ~0.0 |
| 9600 | ~0.0 |
| 9700 | ~0.0 |
| 9800 | ~0.0 |
| 9900 | ~0.0 |
| 10000 | ~0.0 |
| 10100 | ~0.0 |
| 10200 | ~0.0 |
| 10300 | ~0.0 |
| 10400 | ~0.0 |
| 10500 | ~0.0 |
| 10600 | ~0.0 |
| 10700 | ~0.0 |
| 10800 | ~0.0 |
| 10900 | ~0.0 |
| 11000 | ~0.0 |
| 11100 | ~0.0 |
| 11200 | ~0.0 |
| 11300 | ~0.0 |
| 11400 | ~0.0 |
| 11500 | ~0.0 |
| 11600 | ~0.0 |
| 11700 | ~0.0 |
| 11800 | ~0.0 |
| 11900 | ~0.0 |
| 12000 | ~0.0 |
| 12100 | ~0.0 |
| 12200 | ~0.0 |
| 12300 | ~0.0 |
| 12400 | ~0.0 |
| 12500 | ~0.0 |
| 12600 | ~0.0 |
| 12700 | ~0.0 |
| 12800 | ~0.0 |
| 12900 | ~0.0 |
| 13000 | ~0.0 |
| 13100 | ~0.0 |
| 13200 | ~0.0 |
| 13300 | ~0.0 |
| 13400 | ~0.0 |
| 13500 | ~0.0 |
| 13600 | ~0.0 |
| 13700 | ~0.0 |
| 13800 | ~0.0 |
| 13900 | ~0.0 |
| 14000 | ~0.0 |
| 14100 | ~0.0 |
| 14200 | ~0.0 |
| 14300 | ~0.0 |
| 14400 | ~0.0 |
| 14500 | ~0.0 |
| 14600 | ~0.0 |
| 14700 | ~0.0 |
| 14800 | ~0.0 |
| 14900 | ~0.0 |
| 15000 | ~0.0 |
| 15100 | ~0.0 |
| 15200 | ~0.0 |
| 15300 | ~0.0 |
| 15400 | ~0.0 |
| 15500 | ~0.0 |
| 15600 | ~0.0 |
| 15700 | ~0.0 |
| 15800 | ~0.0 |
| 15900 | ~0.0 |
| 16000 | ~0.0 |
| 16100 | ~0.0 |
| 16200 | ~0.0 |
| 16300 | ~0.0 |
| 16400 | ~0.0 |
| 16500 | ~0.0 |
| 16600 | ~0.0 |
| 16700 | ~0.0 |
| 16800 | ~0.0 |
| 16900 | ~0.0 |
| 17000 | ~0.0 |
| 17100 | ~0.0 |
| 17200 | ~0.0 |



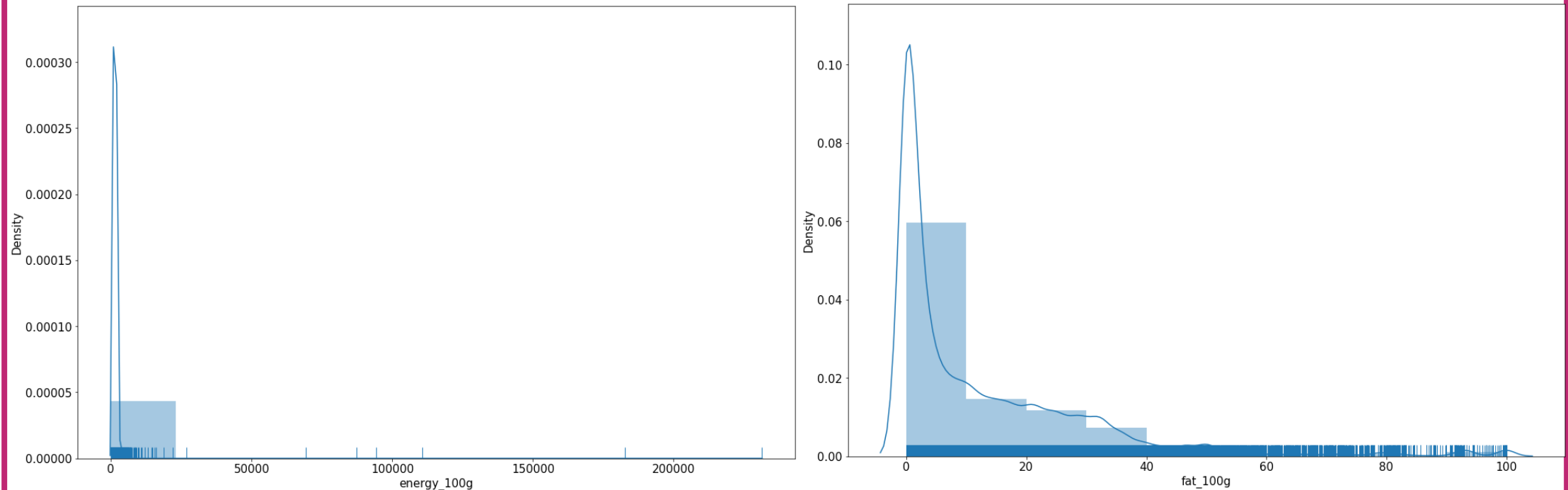
Traitement des valeurs manquantes

- Remplacement des valeurs manquantes à l'aide de la classe `KNNImputer` de `scikit-learn` (une bibliothèque libre de Python) :
 - Méthode utilisant l'algorithme KNN (k-Nearest Neighbours = k plus proches voisins).
- Principe: les valeurs manquantes de notre échantillon sont remplacées par celle de l'échantillon qui a les caractéristiques similaires à celui-ci.



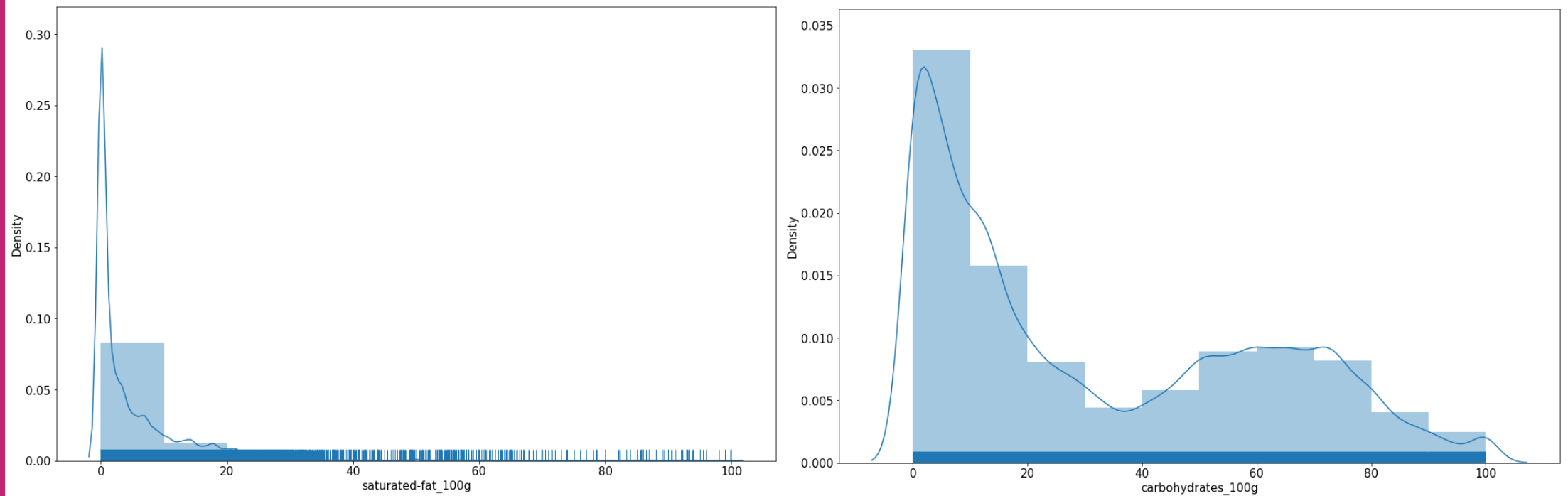
Analyse univariée

- Objectif : Description de la répartition des variables dans l'échantillon
 - Les distributions



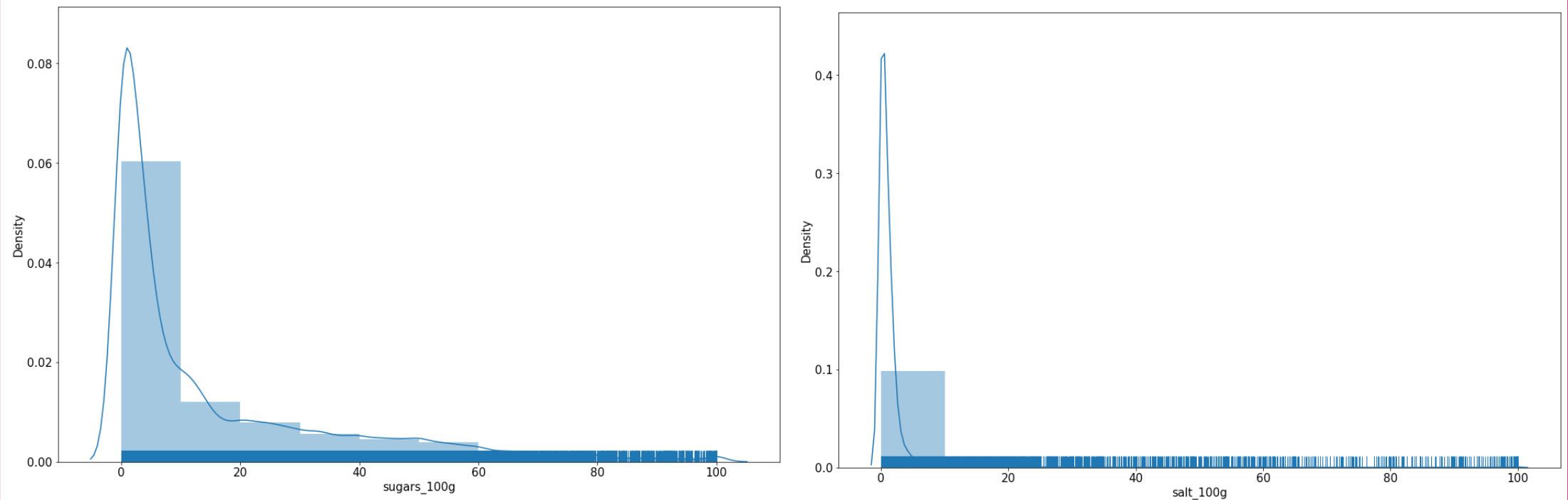
Analyse univariée

- Les distributions



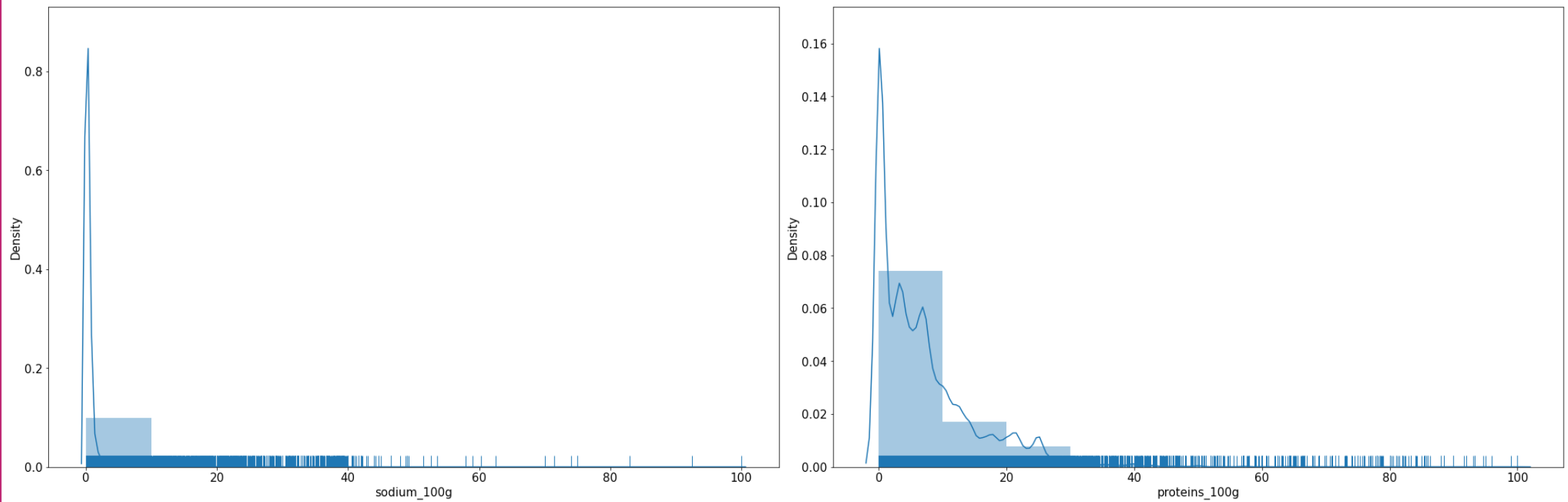
Analyse univariée

- Les distributions



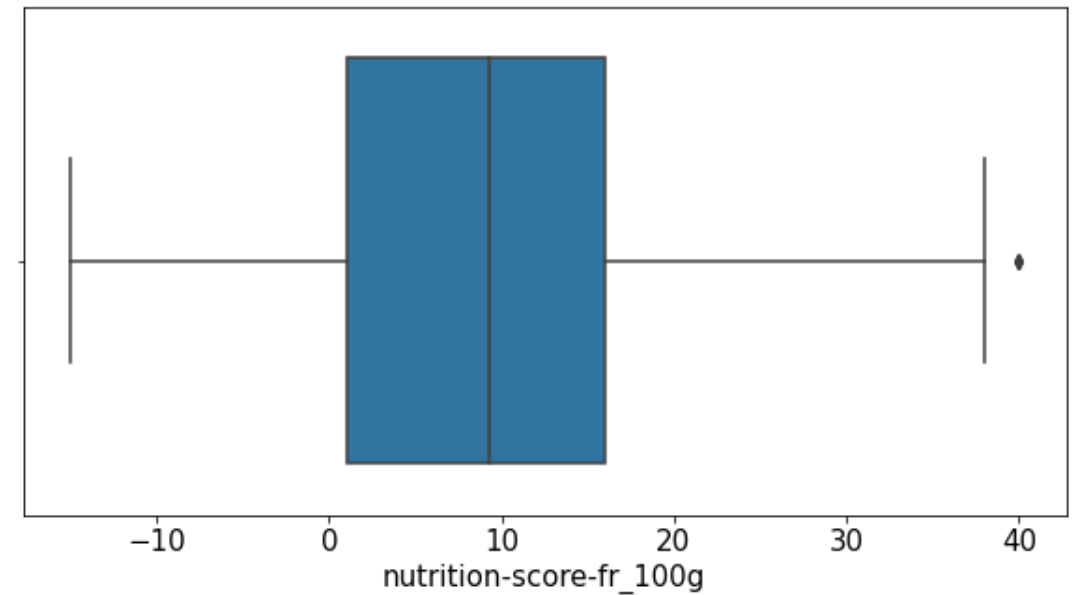
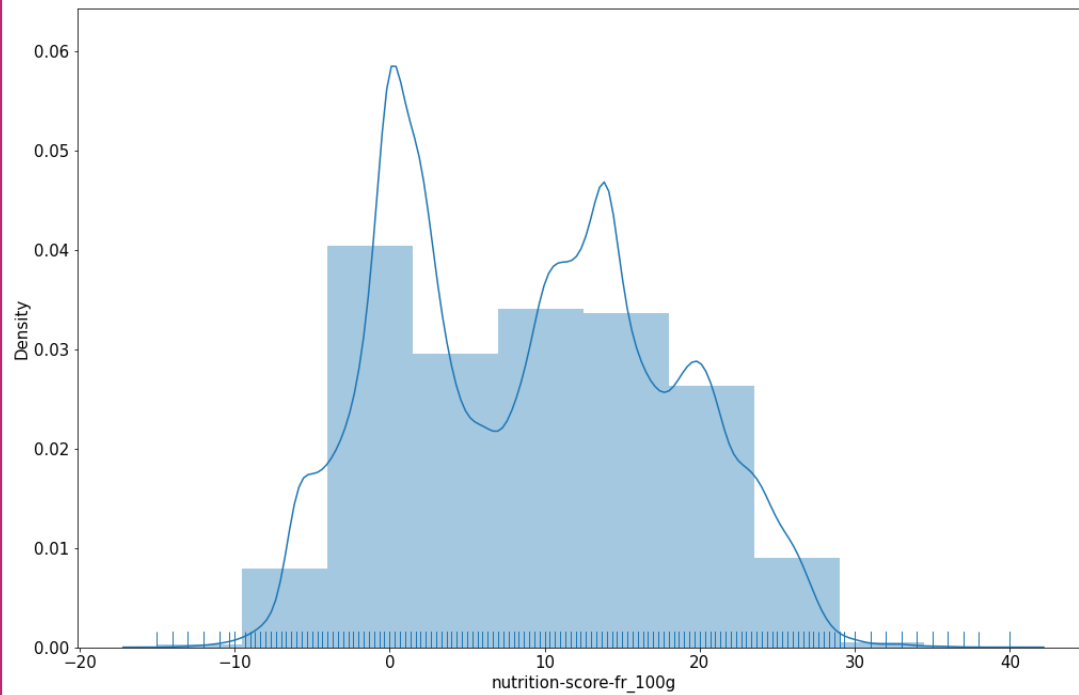
Analyse univariée

- Les distributions



Analyse univariée

- Les distributions

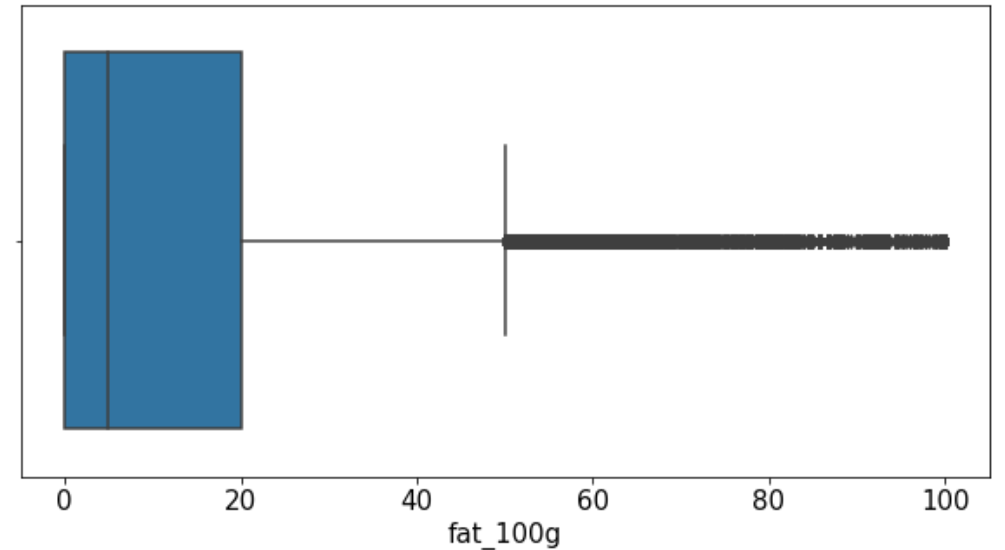
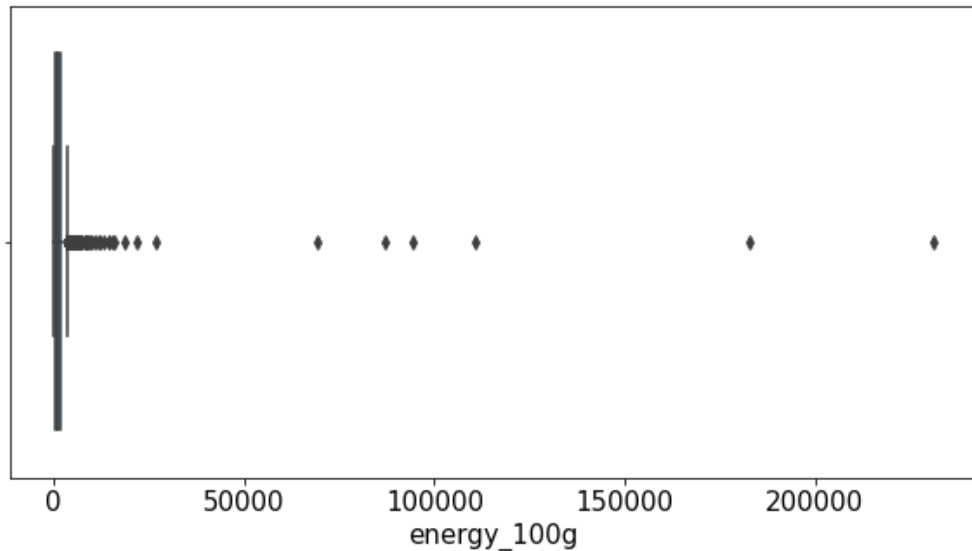


Analyse univariée

- Bilan:
 - Hormis le nutriscore, toutes les distributions des variables sont unimodales mais ne suivent pas une loi normale (une gaussienne).
 - Elles ne présentent pas de symétrie mais elles sont décalées vers la droite.
 - Ces observations peuvent être confirmées par des mesures de formes:
 - Le Kurtosis qui mesure l'aplatissement
 - Le Skewness qui mesure la symétrie
- Le nutriscore quant à lui présente une distribution bimodale.

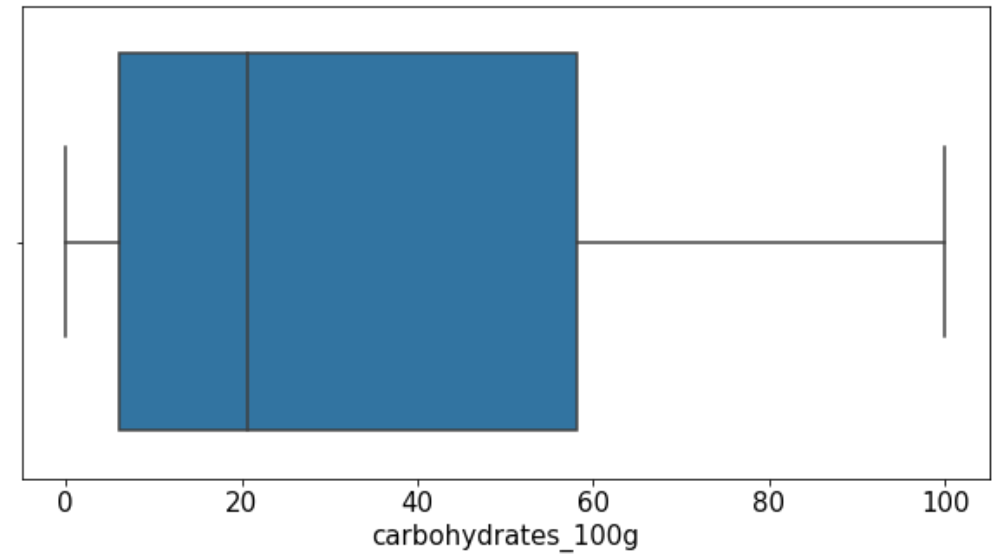
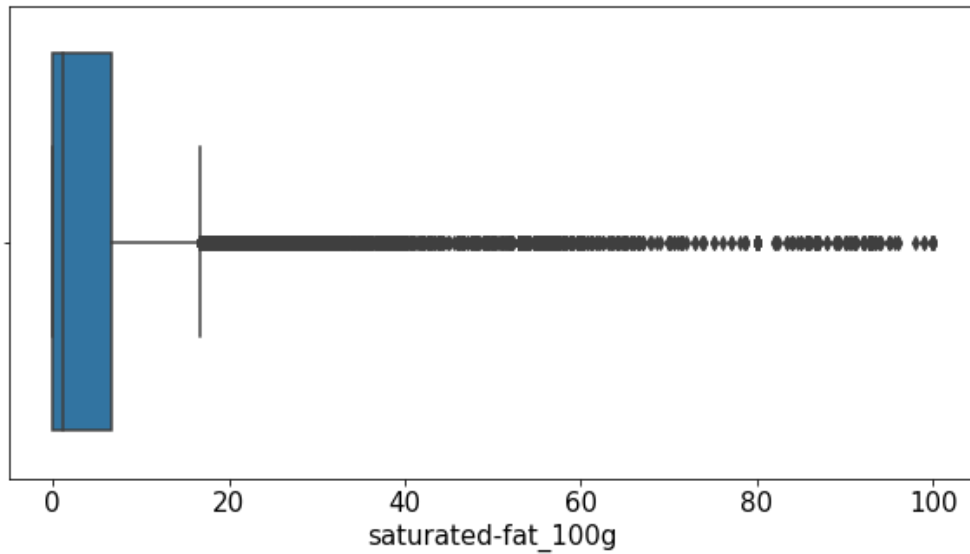
Analyse univariée

- Visualisation des dispersions par les boxplots



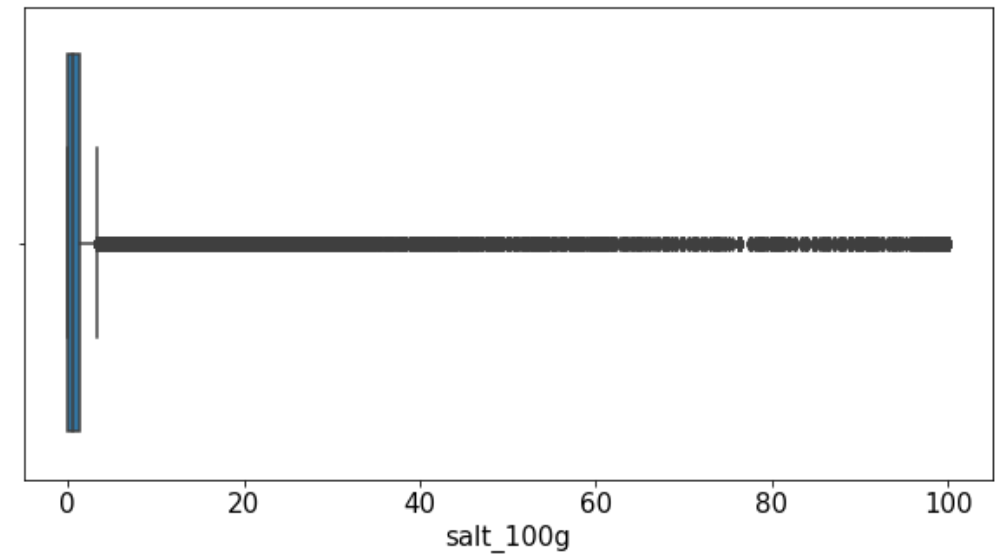
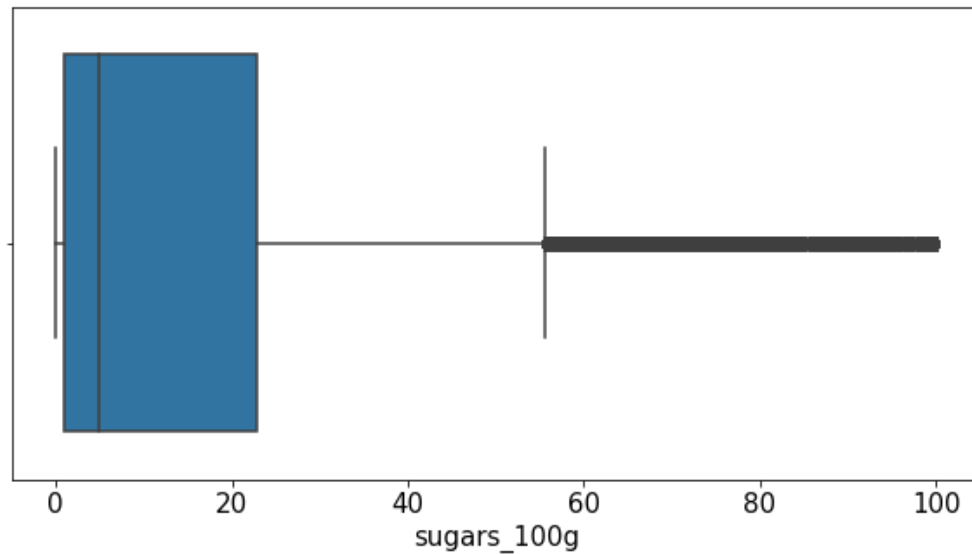
Analyse univariée

- Visualisation des dispersions par les boxplots



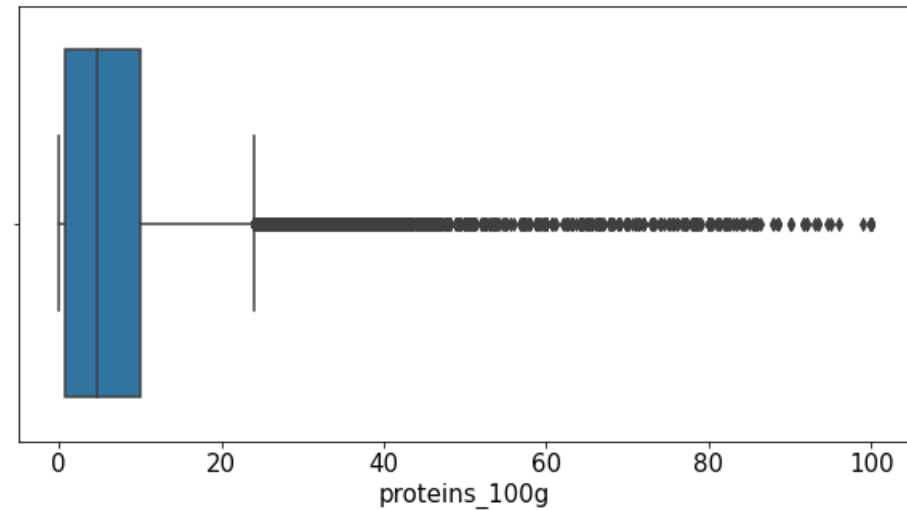
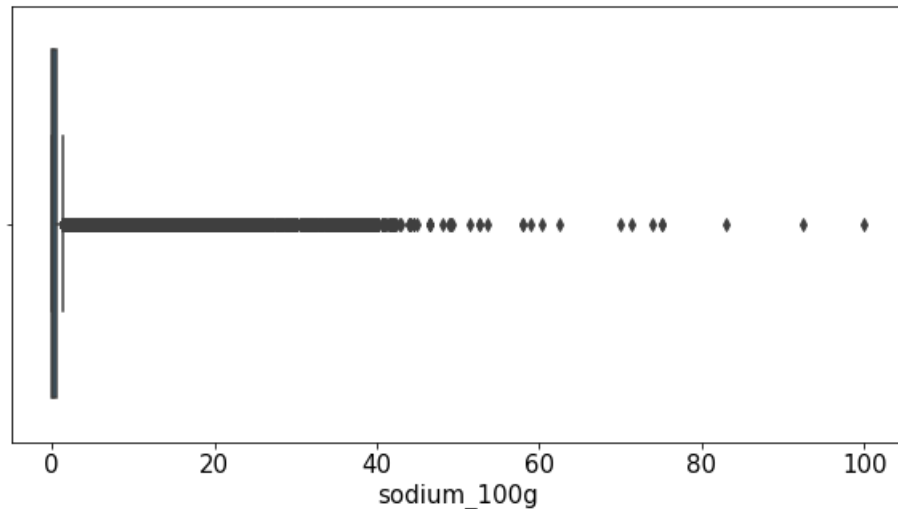
Analyse univariée

- Visualisation des dispersions par les boxplots



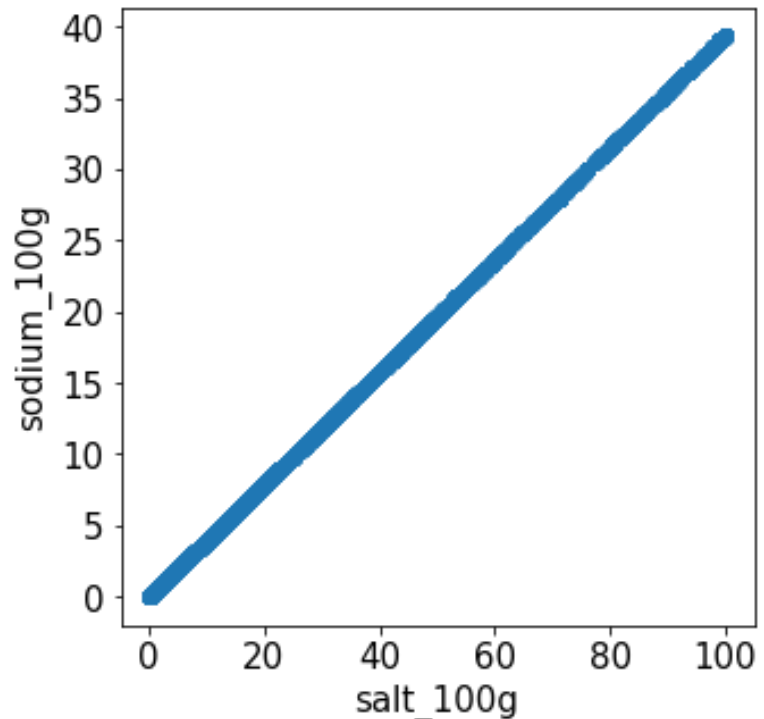
Analyse univariée

- Visualisation des dispersions par les boxplots



Analyse bivariée

- Objectif : déterminer les éventuelles corrélations entre nos variables pour pouvoir réduire la dimension de notre modèle.

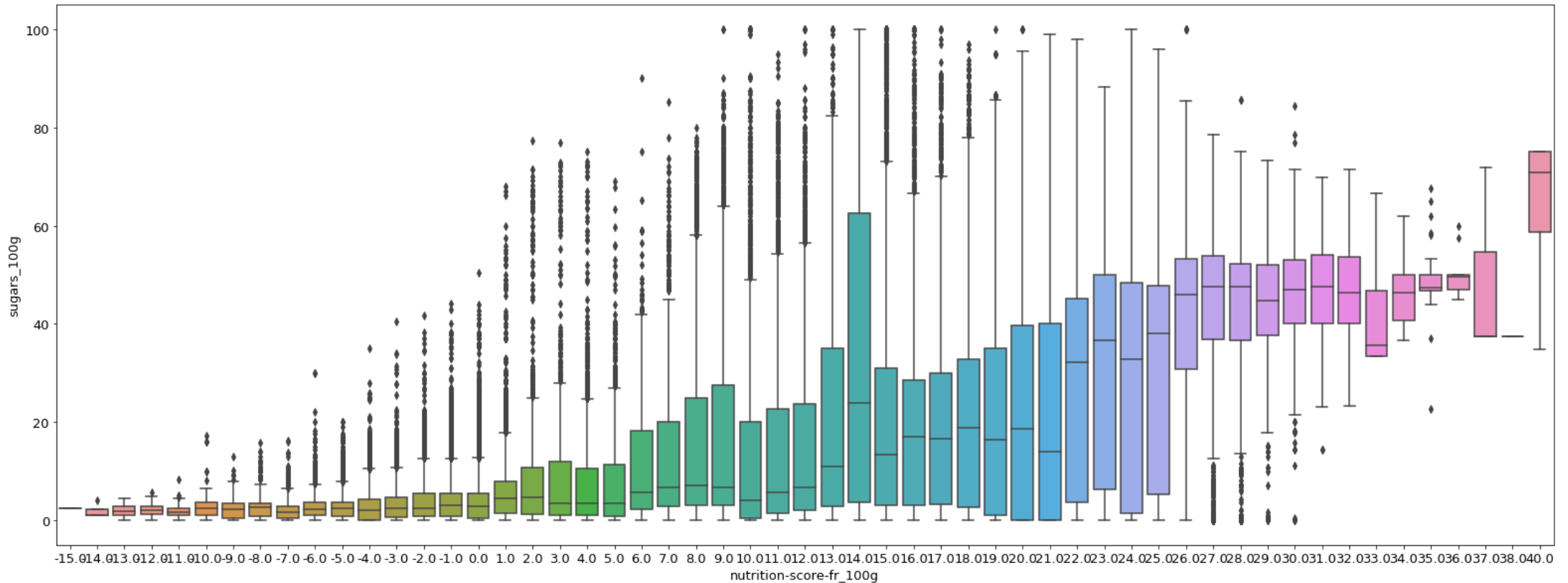


Corrélation entre le sodium et le sel (confirmation de notre hypothèse intuitive)
Retrait du sodium des variables d'entrée de notre modèle de prédiction.

Analyse bivariée

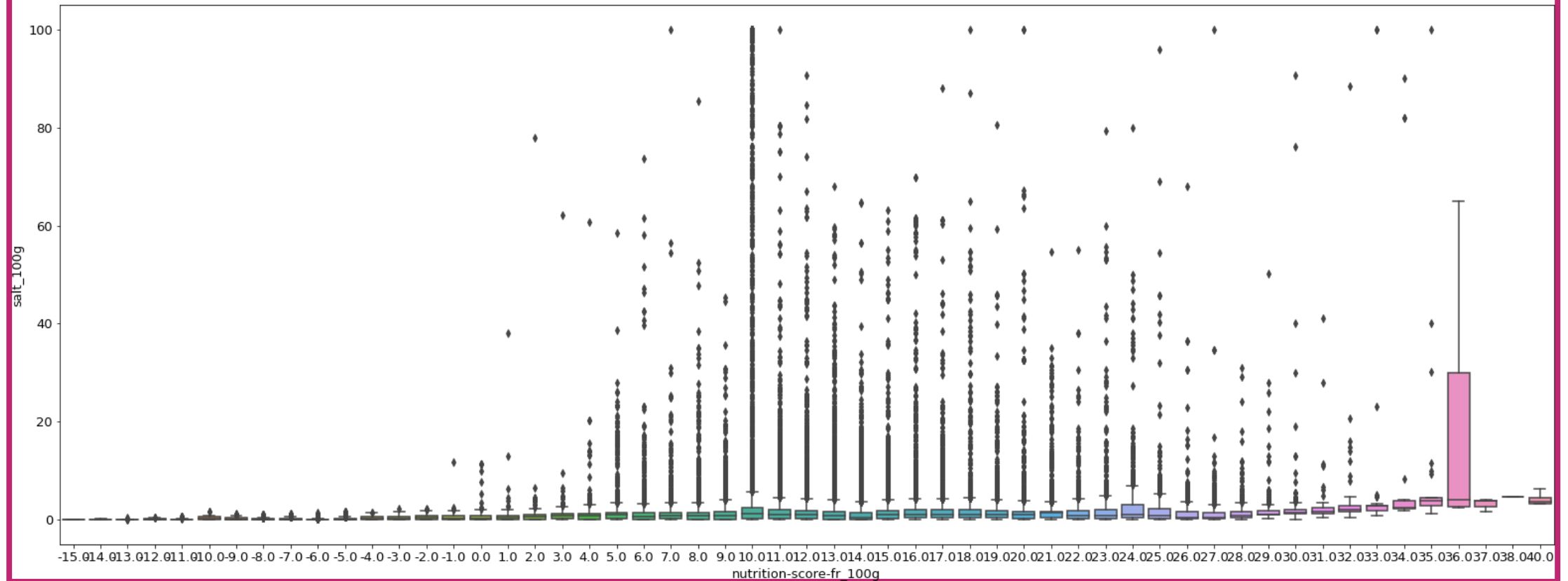
- Relation entre le sucre et le nutriscore

Lien entre le nutriscore et le taux de sucre



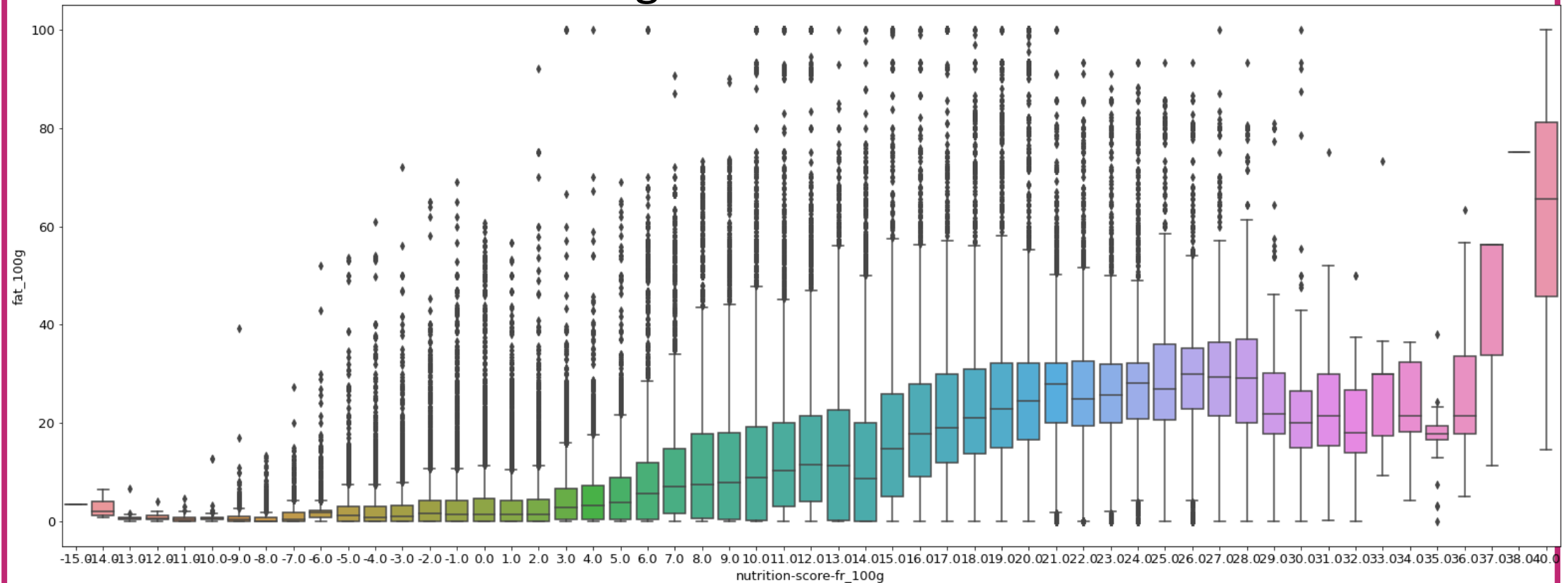
Analyse bivariable

- Relation entre le sel et le nutriscore



Analyse bivariée

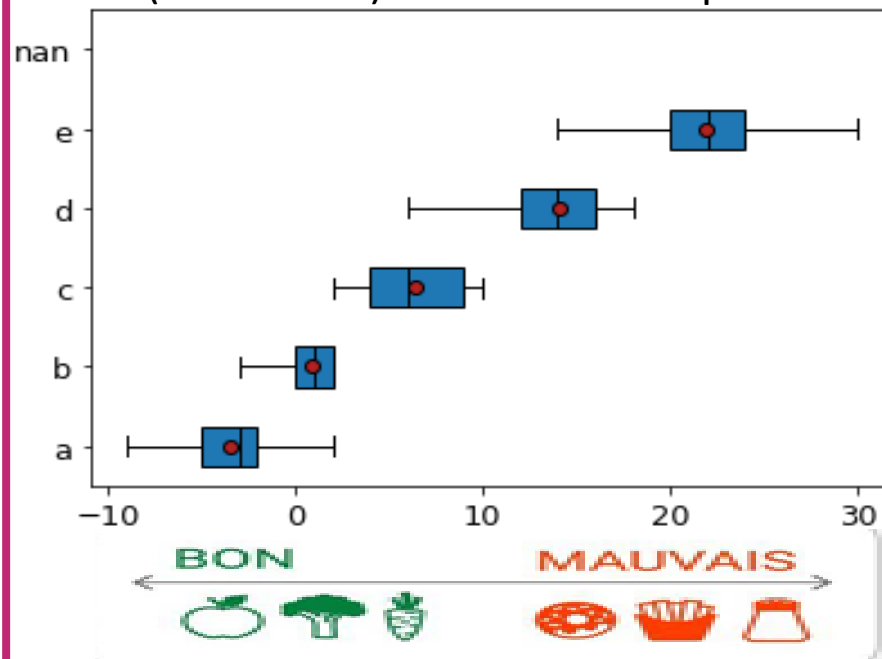
- Relation entre les matières grasses et le nutriscore



Analyse bivariée

- Bilan sur la manière d'attribution:

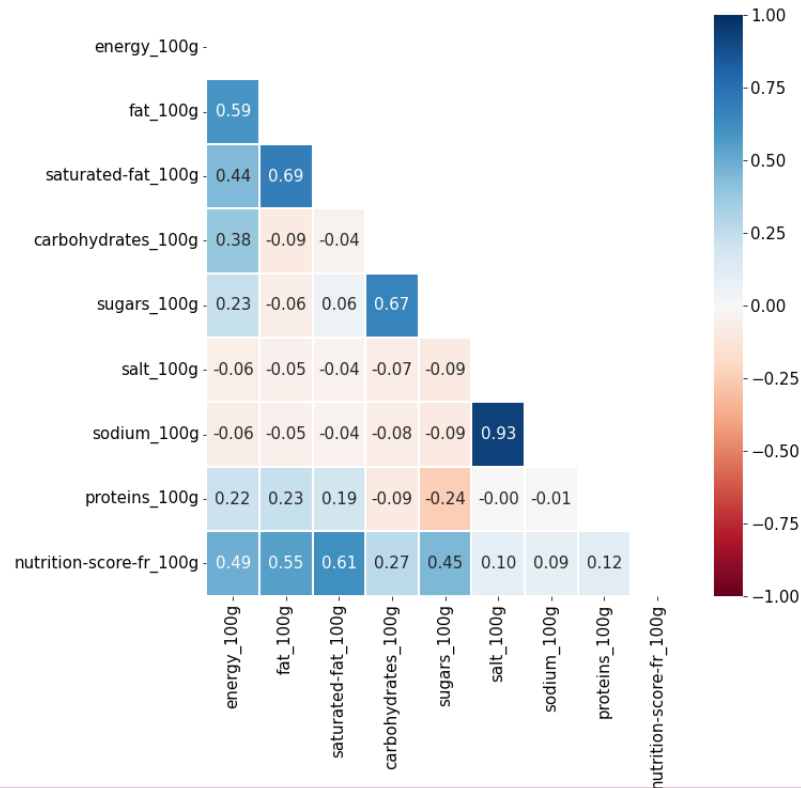
Relation entre le nutriscore lettre
(ou couleur) et le nutriscore points



Aliments solides (points)	Boissons (points)	Nutriscore (lettre)
-15 à -1	Eau	A B C D E
0 à 2	≤ 1	A B C D E
3 à 10	2 à 5	A B C D E
11 à 18	6 à 9	A B C D E
19 à 40	10 à 40	A B C D E

Analyse multivariée

• Etude de corrélations



Constat:

▪ corrélation positive du Nutriscore avec

- energy_100g
- fat_100g
- saturated_fat_100g
- Sugars_100g
- Carbohydrates_100g

corrélation positive → plus la composition du produit est riche en ces nutriments plus son nutriscore sera grand

Analyse multivariée : Analyse en composantes principales(ACP)


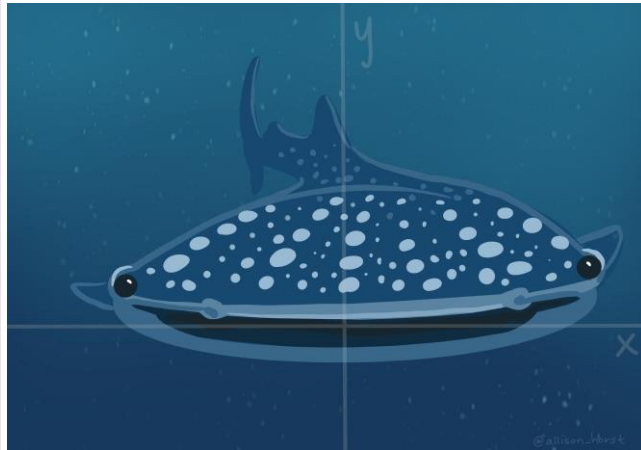
- Etude des corrélations entre les variables
- L'ACP est une méthode consistant à transformer des variables corrélées entre elles en nouvelles variables. Chacune de ces nouvelles variables est le résultat d'une combinaison linéaire des anciennes variables.
-  L'ACP projette nos données dans un nouvel espace. La première composante principale est construite de manière à capter la plus grande variance possible de nos données, la seconde la part la plus importante de la variance possible restant à expliquer, et ainsi de suite.

Illustration de l' ACP



Jeux de données à 2D



Projection des données sur un
espace 1D la construction de
la 1^{ère} composante principale

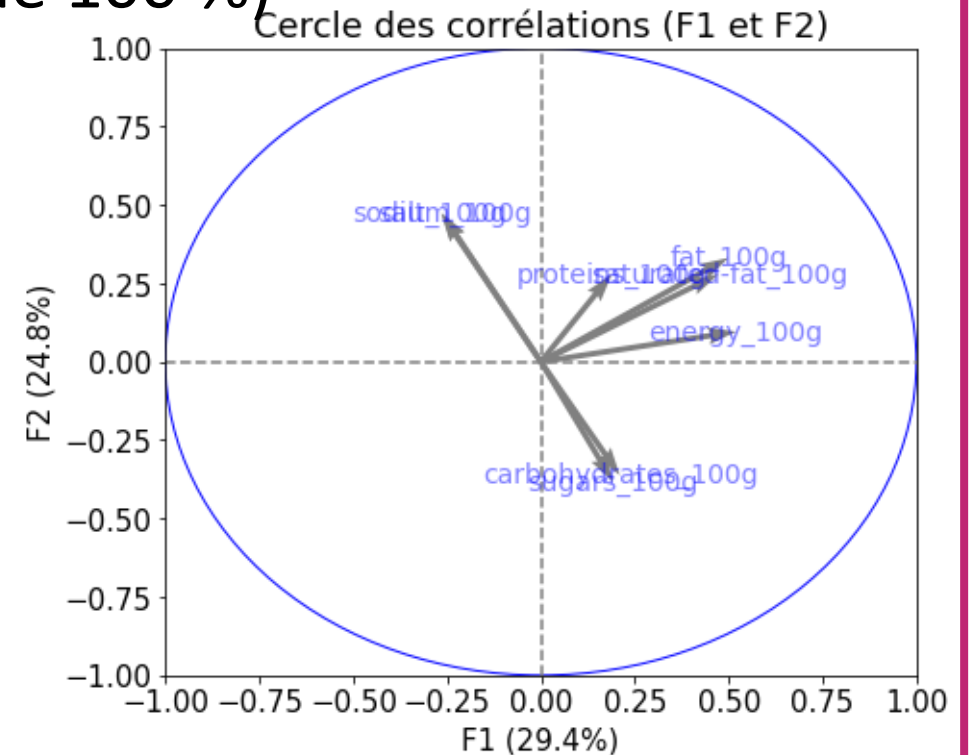
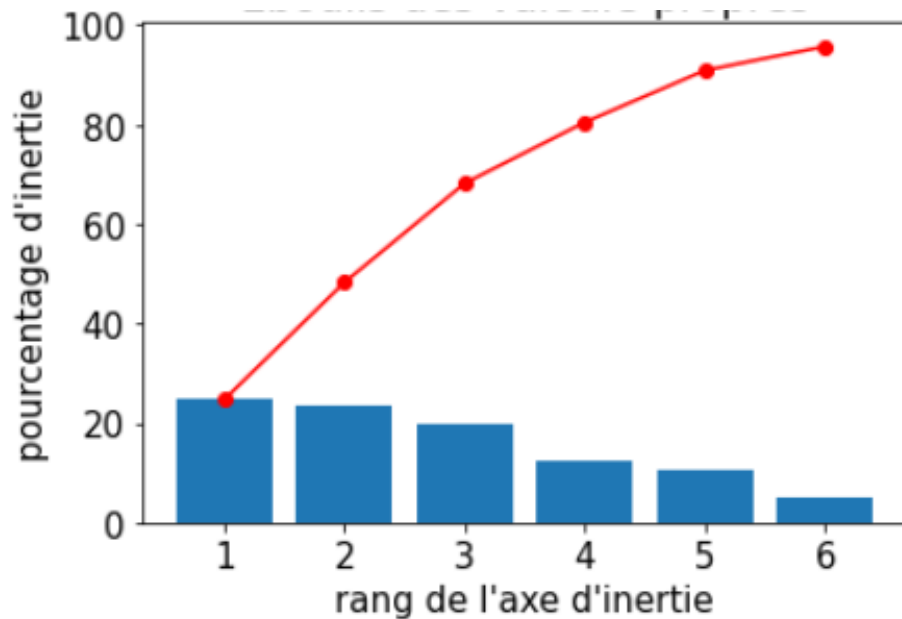
Nouvel espace à 1D



Objectif : réduire la dimension du jeux de données
avec un minimum de pertes d'information possible
(= variance maximum)

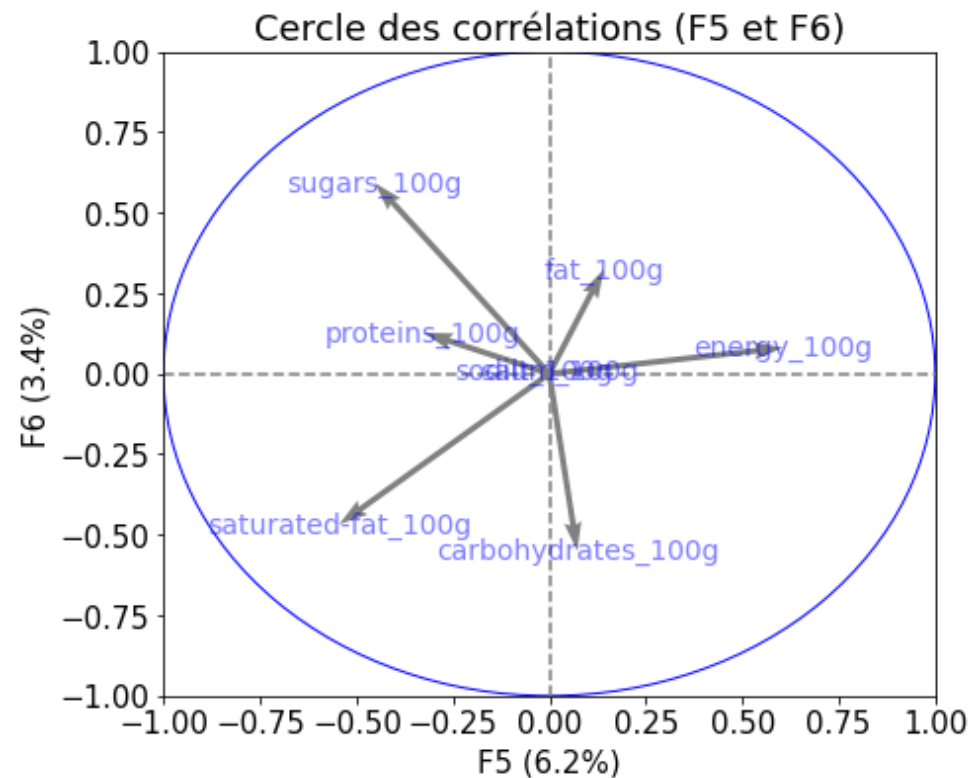
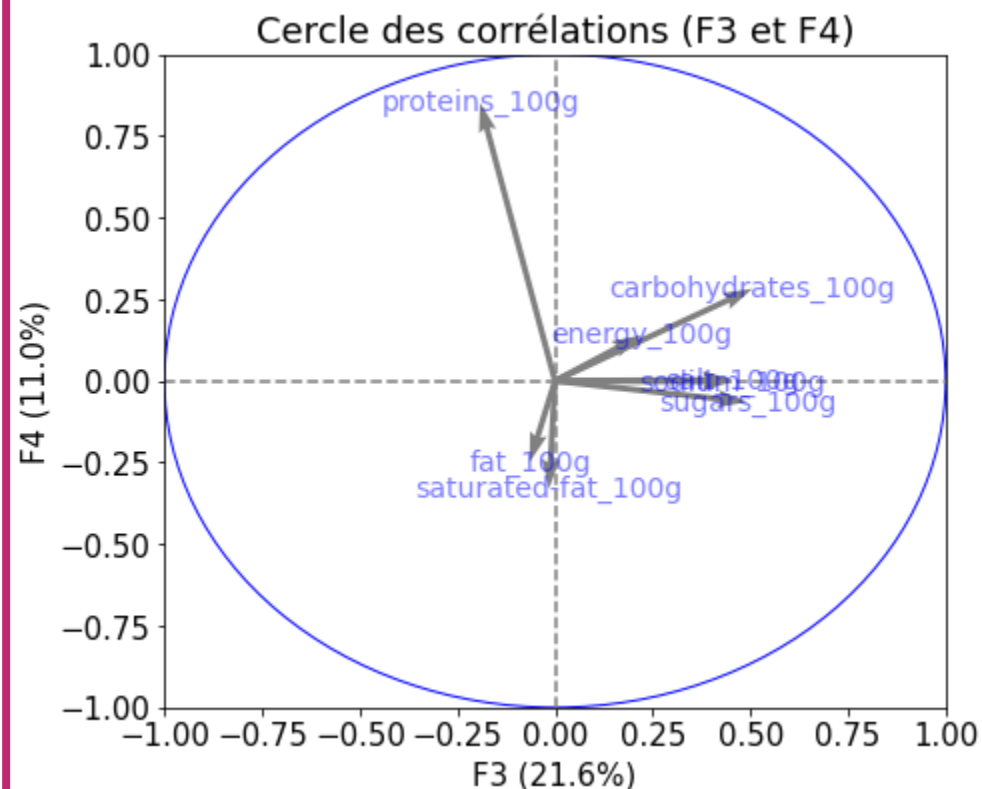
ACP: corrélations entre les variables et réduction de dimension

- Projections des 7 données sur 3 plans factoriels avec 6 composantes principales (variance cumulée proche de 100 %)



ACP: corrélations entre les variables et réduction de dimension

- Les autres composantes principales



ACP: corrélations entre les variables et réduction de dimension

- Certaines variables présentent une bonne corrélation avec les composantes principales ➡ possibilités de réduire la dimension de nos variables
 - fat_100g, saturated-fat_100g et energie_100g sont bien corrélées positivement à F1, la 1^{ère} composante principale.
 - sodium_100g et salt_100g sont bien corrélées positivement à F2, la 2^{ème} composante principale
 - carbohydrates_100g et sugars_100g sont bien corrélées négativement à F2 et positivement à F3.
 - proteins_100g est bien corrélées positivement à F4, la 4^{ème} composante principale
- ➡ Ces 4 composantes présentent à 80% d'inertie et sont corrélées à nos variables

Synthèse et conclusion

- La corrélation entre les variables avec le nutriscore nous permet d'étudier un modèle de prédiction de ce dernier à travers une application.
- Possibilité également de réduire la dimension de notre modèle.
- Plusieurs possibilités de modèles de machine learning s'offre à nous pour faire notre application (Régression linéaire, Kneighbors Classifier, Arbre de décision etc.)
- Il faudra choisir le mieux adapté, le paramétrer pendant la phase d'apprentissage. Il faudra ensuite étudier la fonction coût qui calcul nos erreurs. Les bons paramètres seront ceux pour lesquels, on a moins d'erreurs