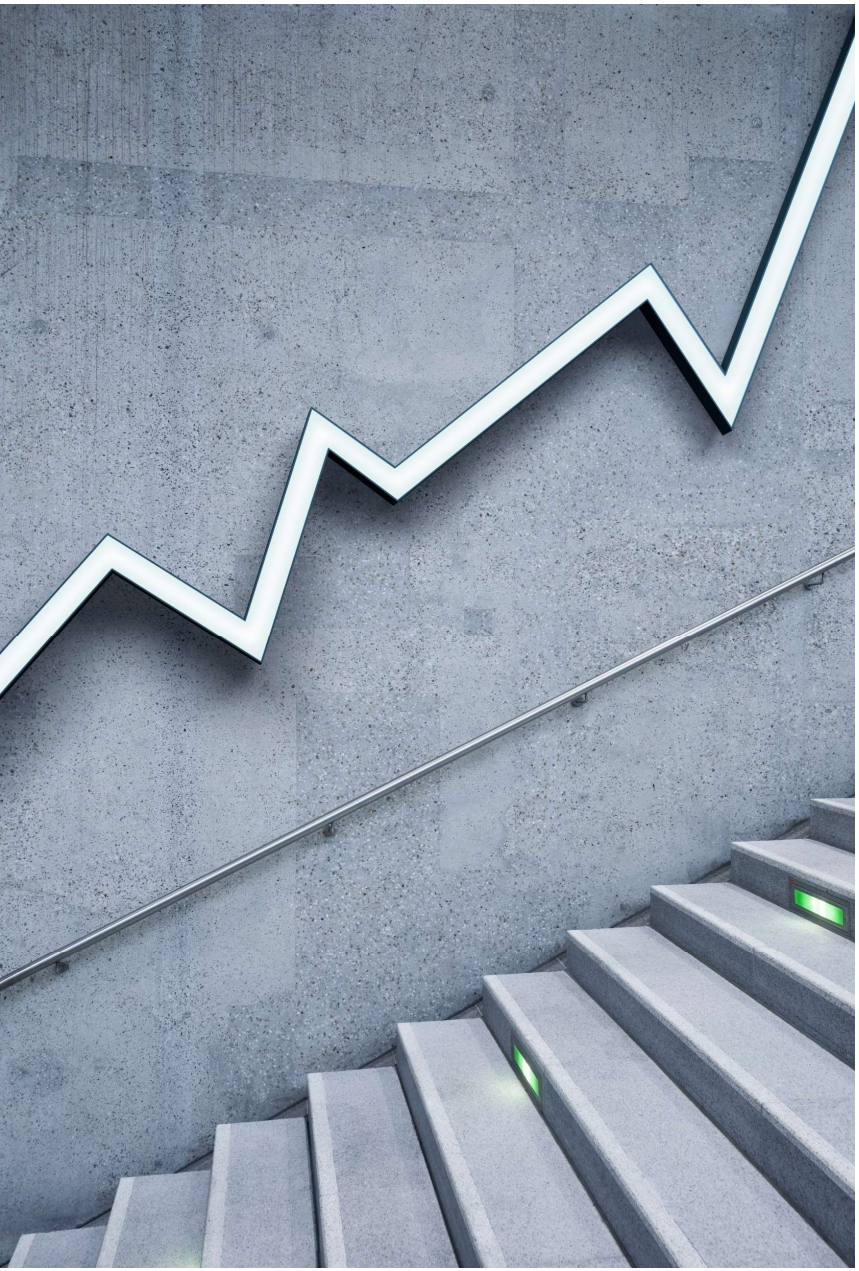


Santé publique France

Amélioration de la base de données Open Food Facts

PRÉPARATION DES DONNÉES

Cédric – Aout 2024

Objectifs de l’étude

Etablir la faisabilité de suggérer les valeurs manquantes pour une variable dont plus de 50% des valeurs sont manquantes..

Fichier `fr.openfoodfacts.org.products.csv`
`Toutes les données sont dans un seul fichier zippé.`

Structure du DataFrame d'origine : **162 colonnes pour 320 772 lignes**

Le fichier est composé de quatre blocs d'informations :

- Informations générales :
  le nom des produits, leur code barre, des informations sur leur création et mise à jour.
- Données nutritionnelles :
  des quantités en g (ou en Kj pour l'énergie) pour 100g ou 100ml de produits
- Détails en différentes langues
- Tags :
  permettant de catégoriser les produits

Qualité :

Doublons : 0.1%
Valeurs Nulles : nombreuses mais normal sachant que la complétion est à la liberté du contributeur

Les 5 principes fondamentaux du RGPD sont :

Principe de licéité, loyauté et transparence
Principe de limitation des finalités
Principe de minimisation des données
Principe d'exactitude
Principe de conservation

La Feature « Creator » qui comporte le pseudonyme du premier utilisateur a avoir ajouté le produit dans la base ou le site d'où provient l'information ne nous intéresse aucunement nous pouvons donc retirer cette information.

Aucune donnée personnelle n'est récupérée ni stockée,

notre étude n'a donc pas de lien avec les RGPD.

Méthodologie du nettoyage

- Mise en conformité caractères retour charriot et espaces vides [320 772 → 320 749 lignes]
- Filtrage des variables à partir de 75% de valeurs vides [161 → 49 features]
- Vérification des doublons (produits ayant le même code barre 0,1%) [320 398 lignes]
- Création de 5 jeux de données :

    - Informations nutritionnelles (14 features)
    - Informations énergétiques (5 features)
    - Marqueurs (5 features)
    - Labels langue française (6 features)
    - Informations générales (19 features)

	fat	saturated-fat	trans-fat	cholesterol	carbohydrates	sugars	fiber	proteins	salt	sodium	vitamin-a	vitamin-c	calcium	iron	energy	nutrition-score-fr
320733	31.0	NaN	NaN	NaN	12.2	9.60	1.1	2.10	1.1000	0.433071	NaN	NaN	NaN	NaN	569.0	NaN
320734	NaN	3.73	NaN	NaN	NaN	3.89	12.2	21.22	0.1000	0.039370	NaN	NaN	NaN	NaN	2406.0	0.0
320740	0.2	0.20	NaN	NaN	0.5	0.50	0.2	0.50	0.0254	0.010000	NaN	NaN	NaN	NaN	21.0	2.0

Gestion des données aberrantes

Après un balayage des aspects nutritionnels et énergétiques pour 100g il convient de supprimer les individus ayant :

1. Une valeur nutritionnelle négative ou supérieure à 100g (219 individus)
2. Un composé nutritif égal à 100g (1784 individus) sauf pour :
    1. Les protéines dont le nom de produit contient 'gelatin'
    2. Le sel dont le nom de produit contient 'sel' ou 'salt'
    3. Le sucre dont le nom de produit contient 'sucre' ou 'sugar'
    4. Les glucides dont le nom de produit contient 'sucre','sugar','sweetener' ou 'candy'
    5. Les graisses dont le nom de produit contient 'huiles', 'beurre', 'saindoux', 'graisse', …
3. Une valeur énergétique supérieure à 3900Kj pour 100g (151 individus)
4. Une quantité de fer supérieure à 70mg pour 100g (48 individus)

Ces divers traitements nous permettent de passer de 320 398 à 318 196 individus

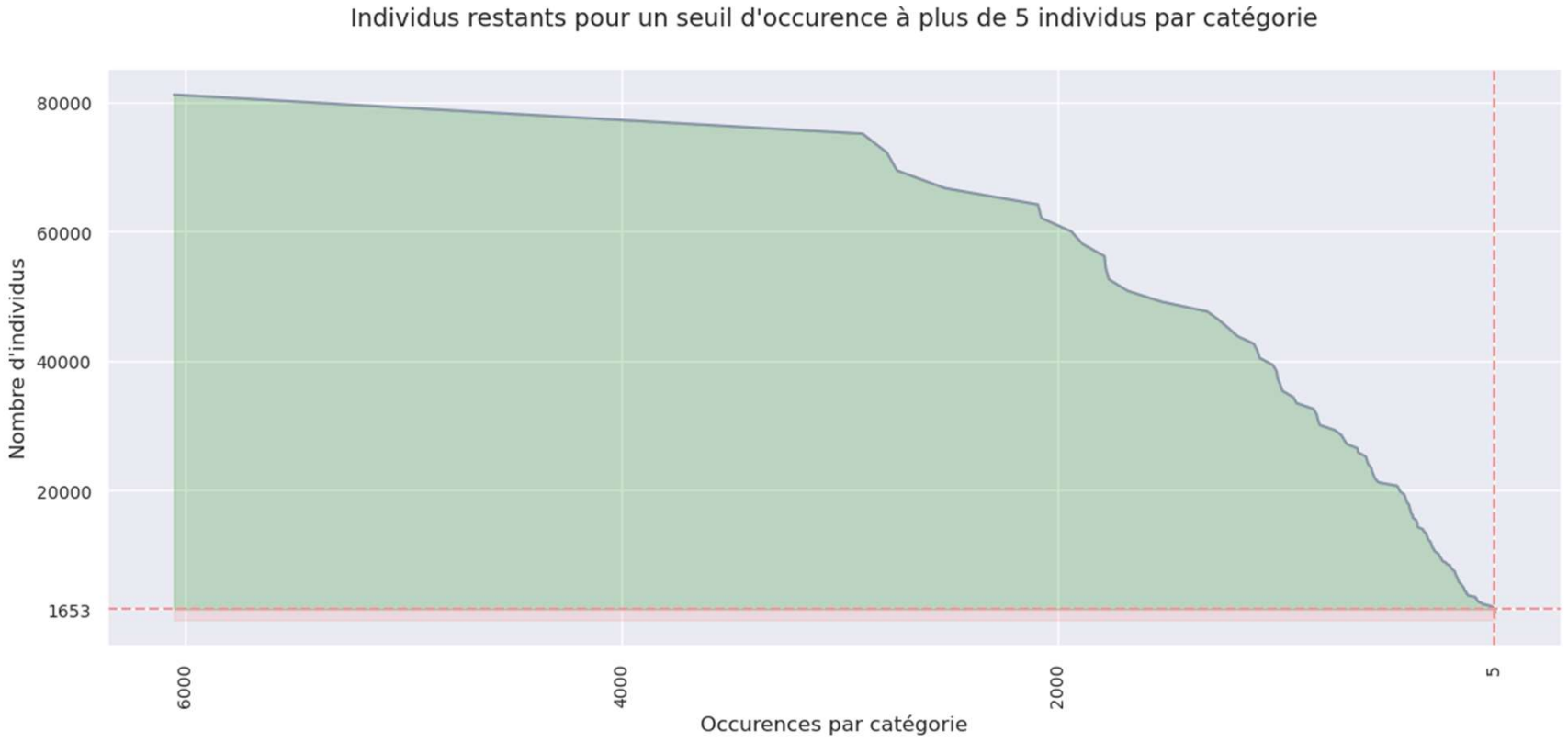
Choix de la cible

Possibilité de suggérer les valeurs de la variable catégorie principale en français en fonction des informations nutritionnelles des produits.

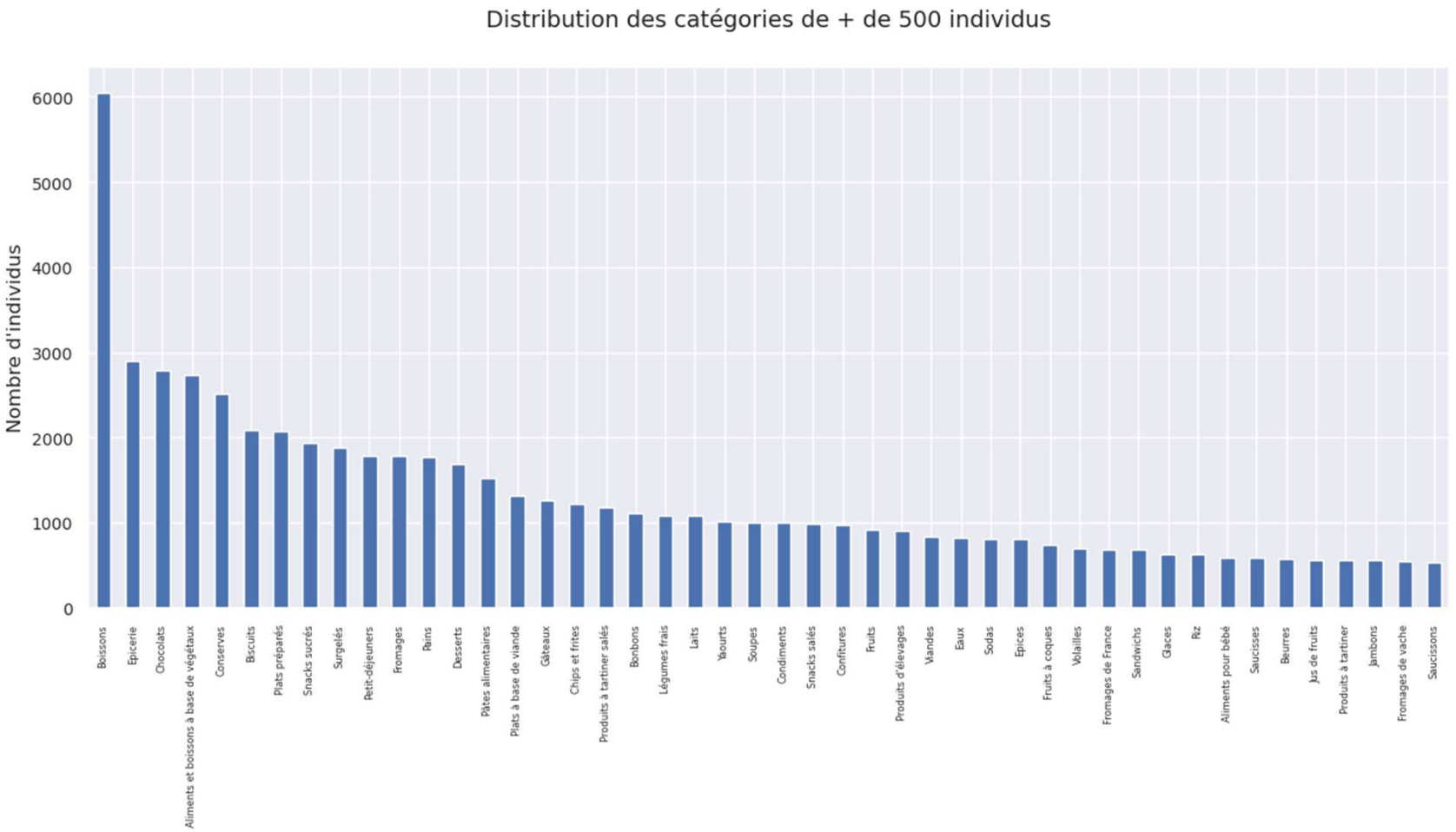
Suppression des individus dont la catégorie principale contient un marqueur d'une autre langue.
(2881 individus)

Nous obtenons ainsi 315 315 individus dont 81 185 ont une catégorie principale.
Ce qui représente 74% de valeurs manquantes.

Un filtre va être appliqué pour ne garder que les catégories qui disposent de plus de cinq produits.
(79 532 individus avec une catégorie principale)



Echantillon des catégories principales

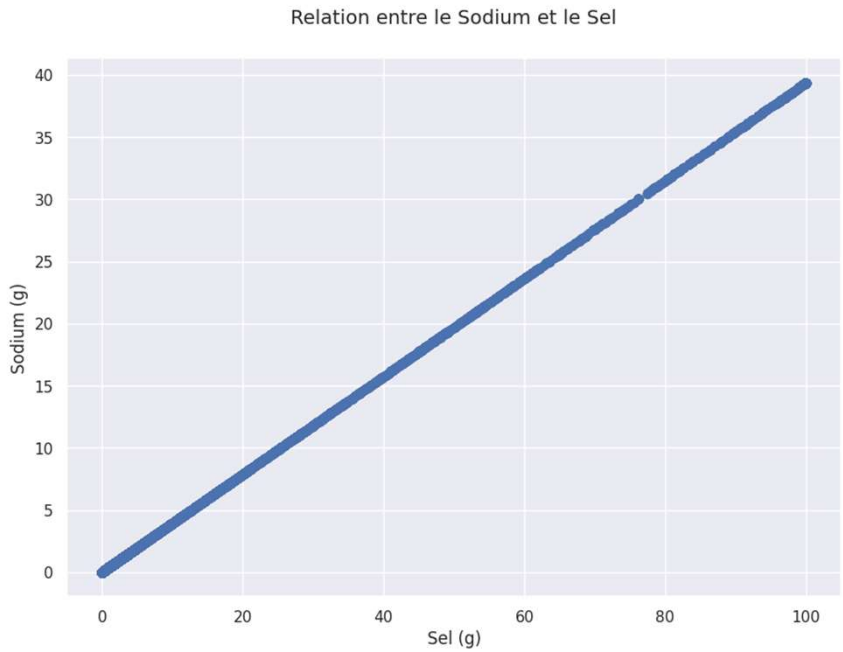


Gestion des valeurs manquantes

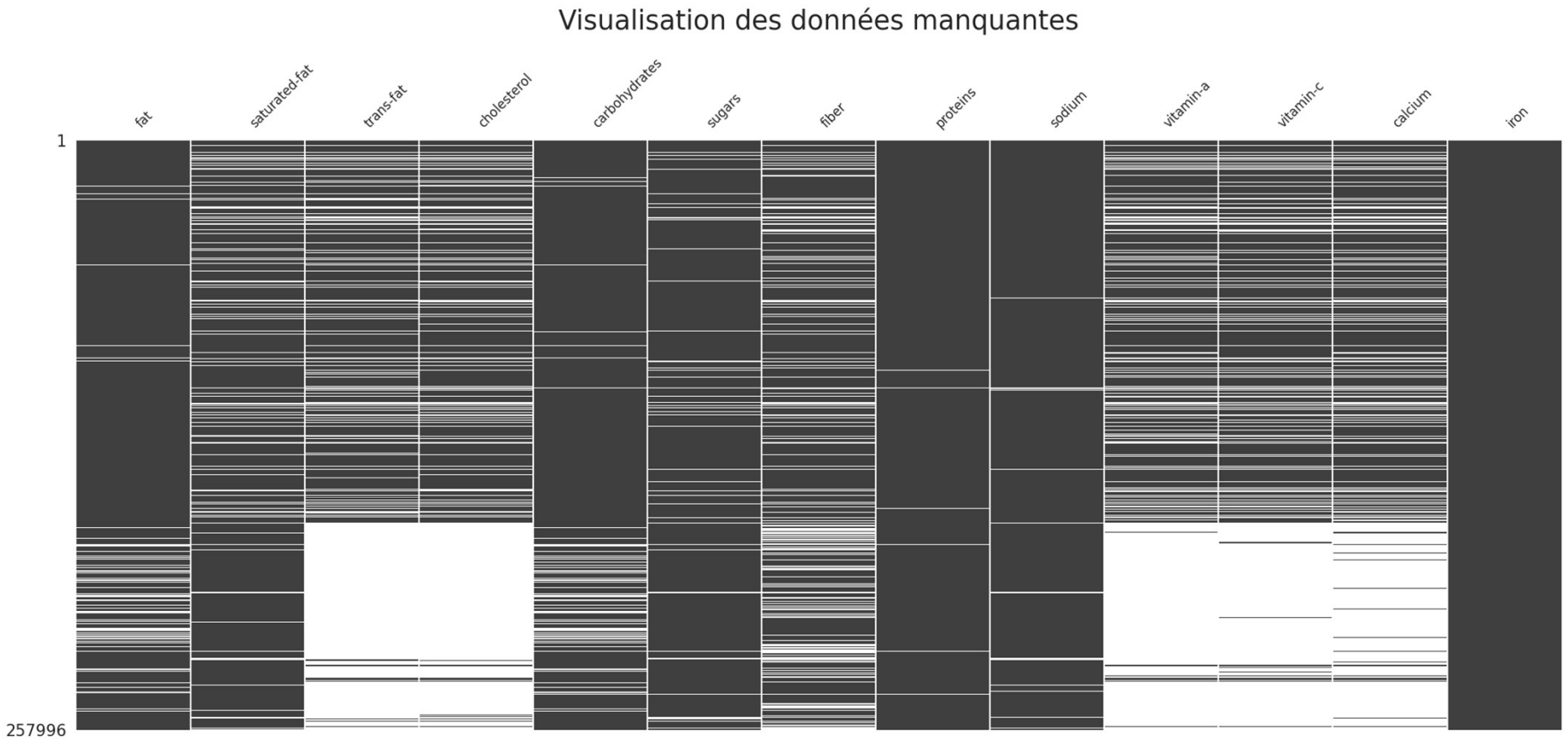
Pour commencer nous allons supprimer les individus n'ayant aucune feature nutritionnelle. (57 320 individus)

Il est flagrant que les informations du Fer n'ont pas été saisies quand celles-ci étaient nulles l'imputation par 0 est indiquée.

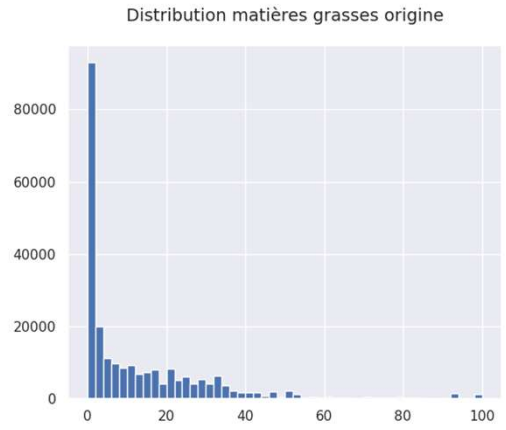
La relation entre le sel et le sodium étant linéaire nous pourrons nous affranchir de la feature Sel.

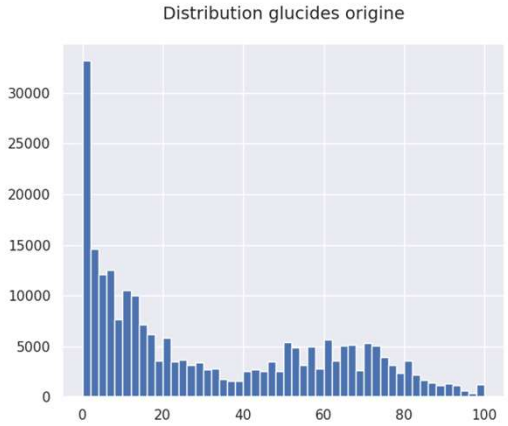


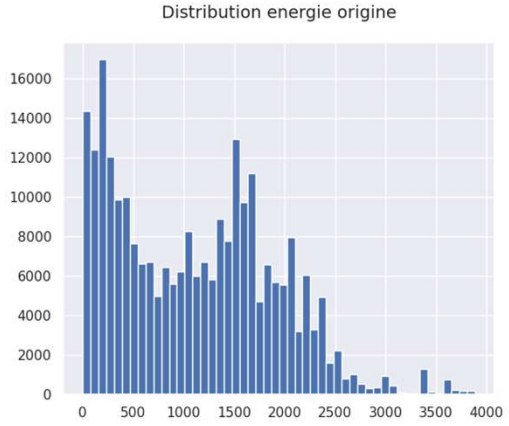
Les valeurs des acides gras, cholestérol, vitamines, calcium semblent s’arrêter toutes simultanément nous allons donc les ignorer vu le peu d’intérêt pour notre étude.

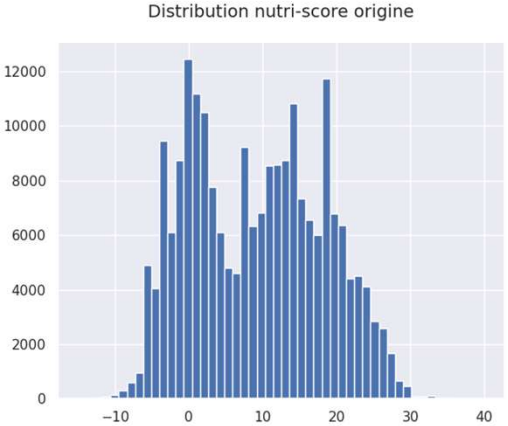


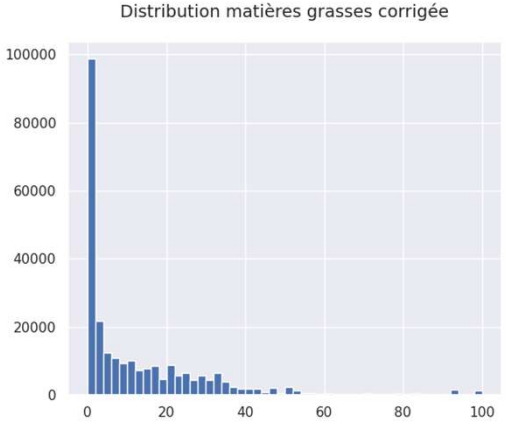
Une imputation par KNN est réalisée sur les autres valeurs manquantes.
Suite à une vérification de la complétude et des distributions avant et après imputation la méthodologie est validée.

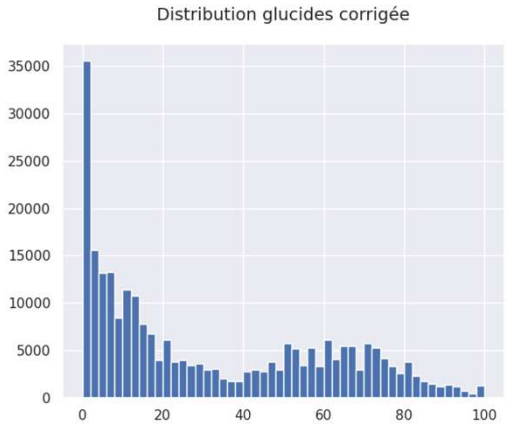


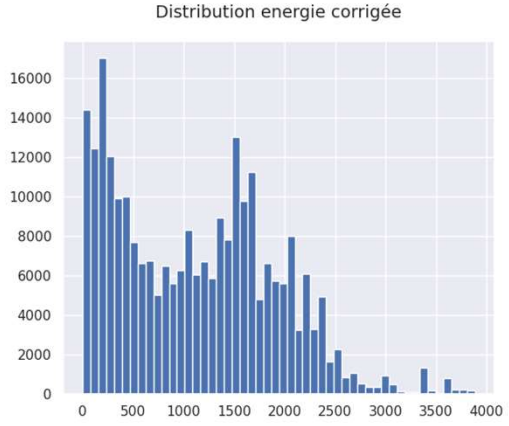


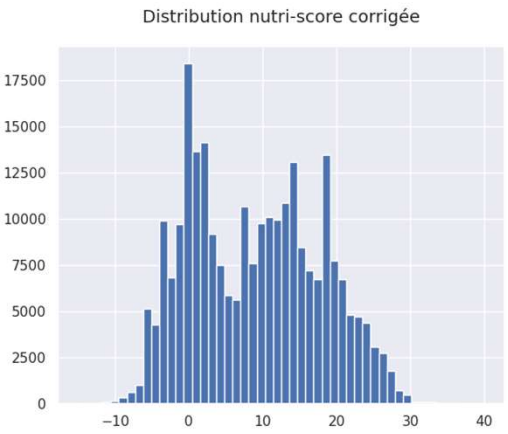






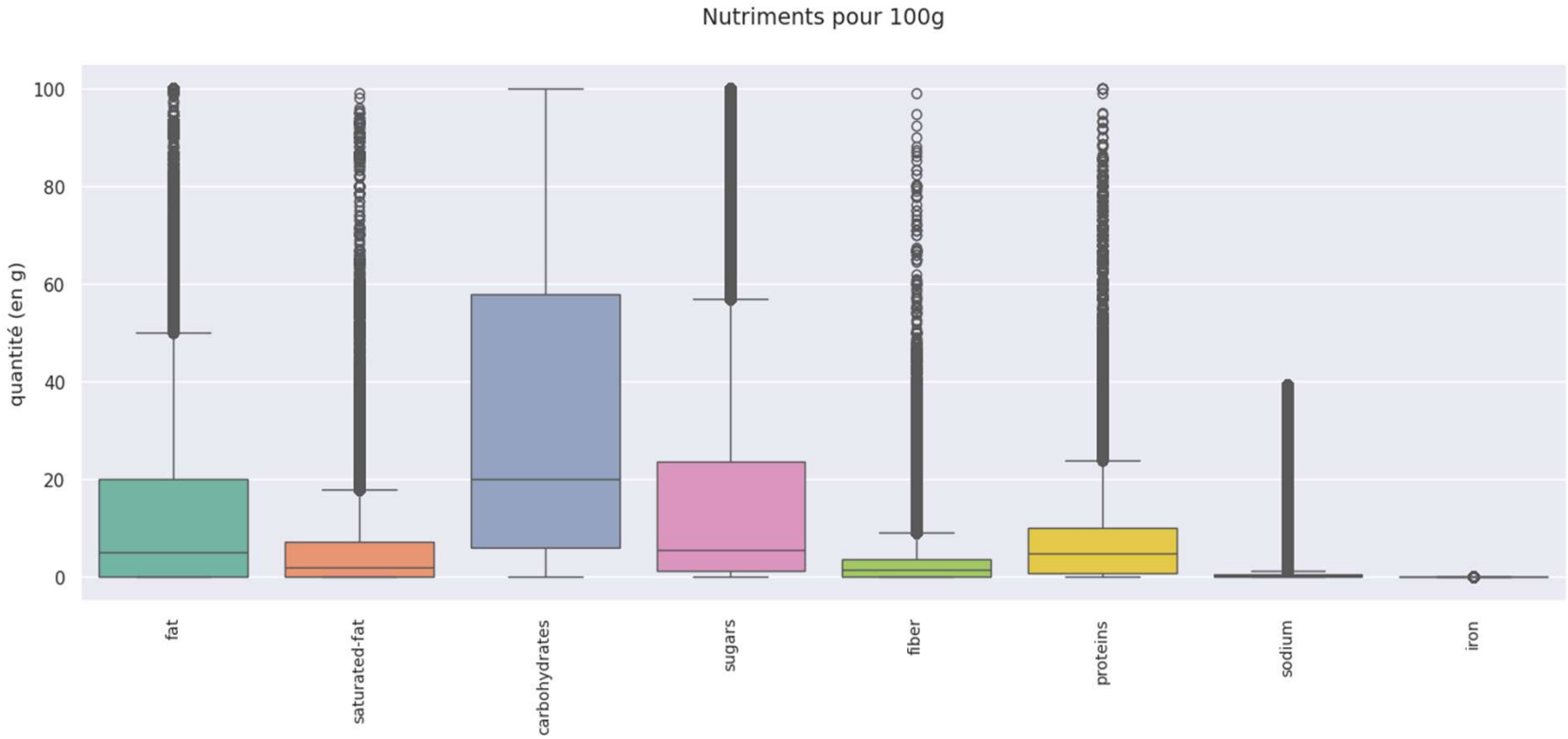




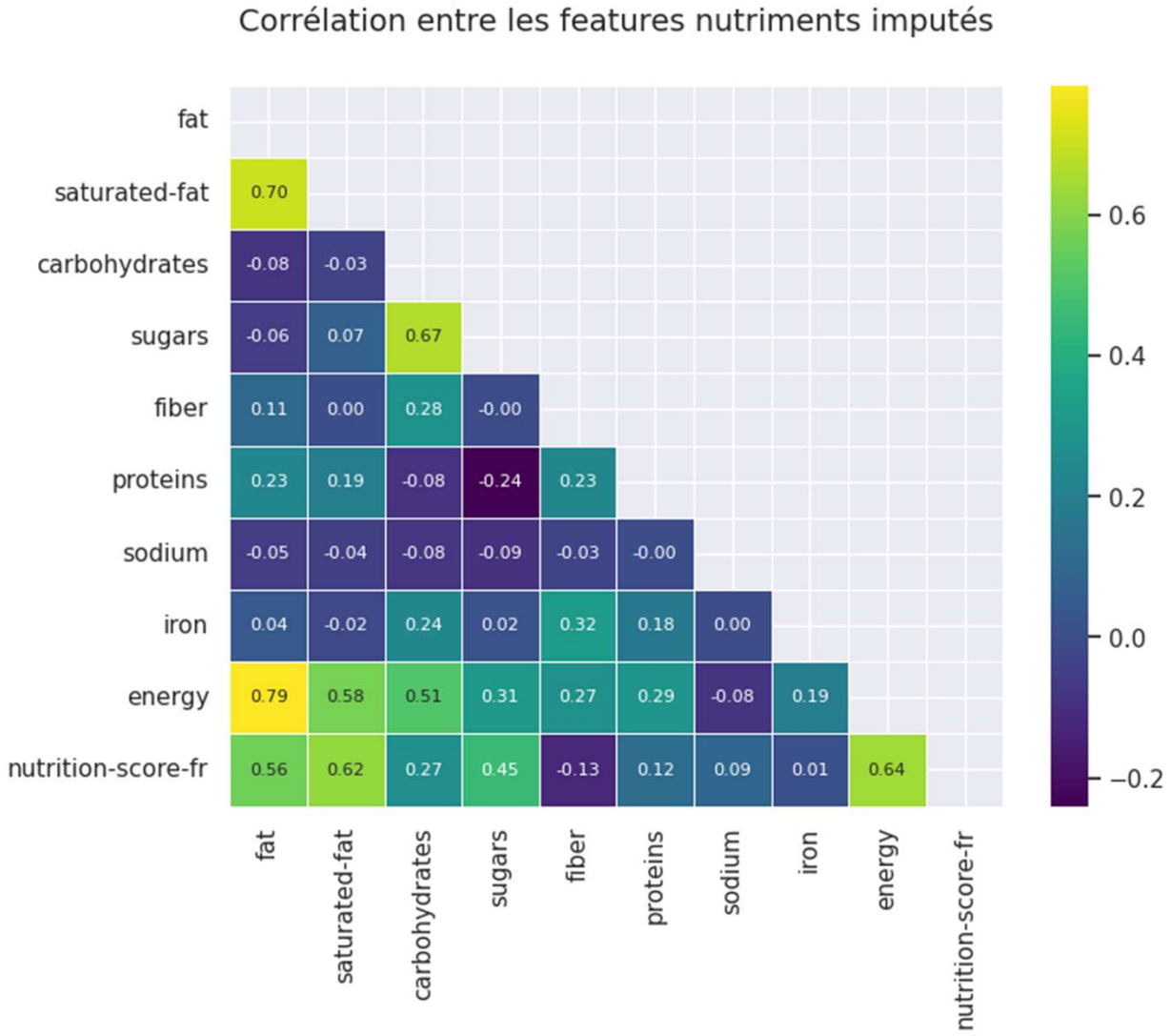


Distribution des nutriments pour 100g de produit

grande diversité des quantités de nutriments présents dans les produits alimentaires



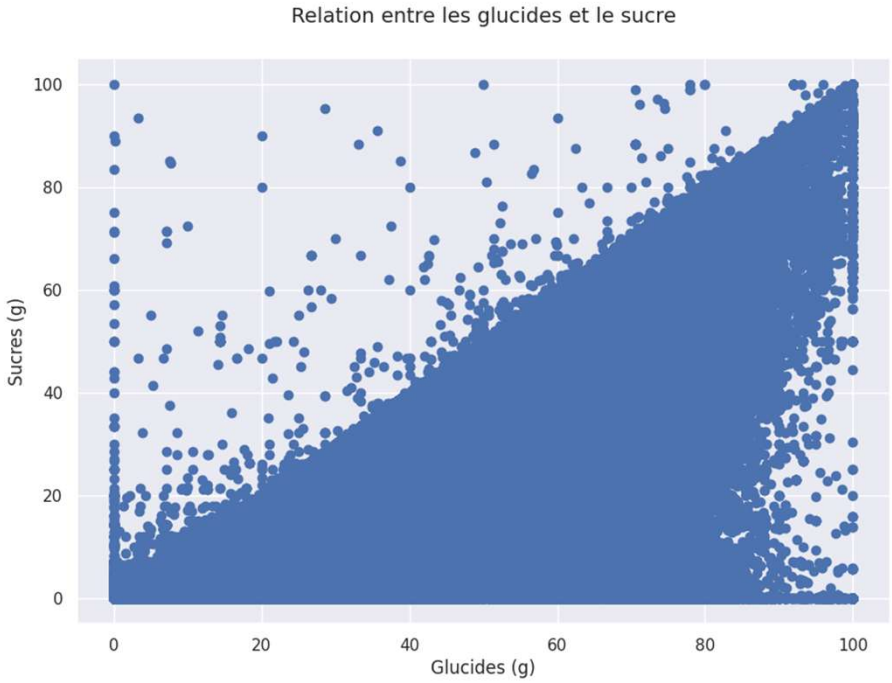
* Détail du sodium et du fer en annexe

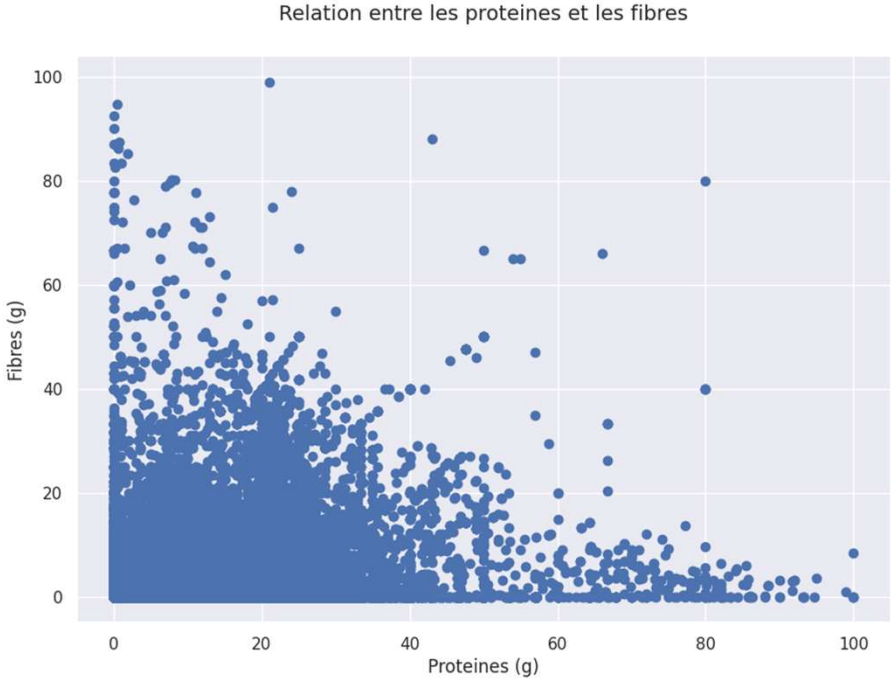


Ce heatmap montre de fortes corrélations entre divers nutriments :

Sucres et glucides
Energie graisses et glucides
Nutri-Score avec Energie

Le graphique sucres glucides met en évidence la corrélation, ainsi que le bruit généré par la famille des édulcorants qui n'a pas de relation entre ces deux nutriments.



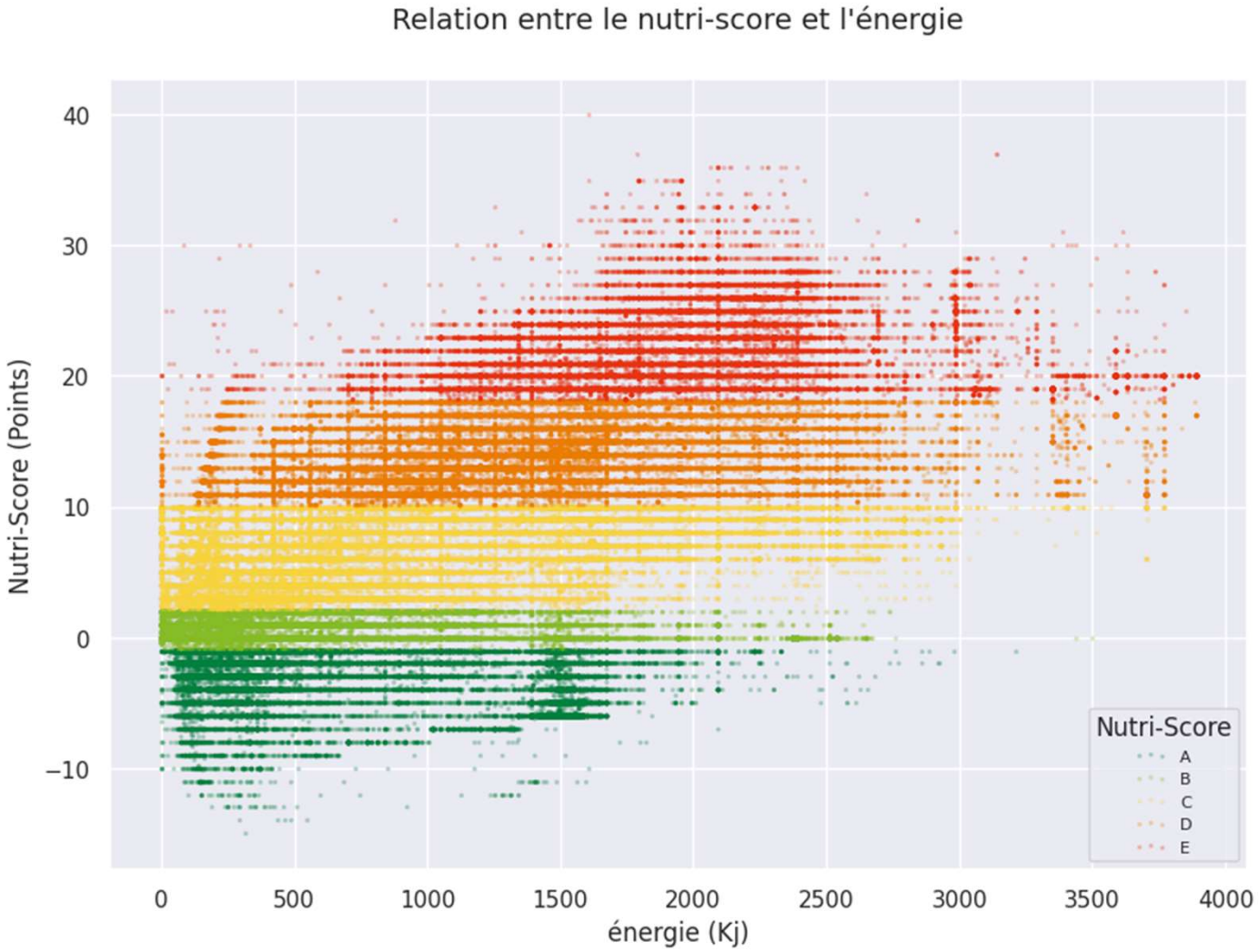


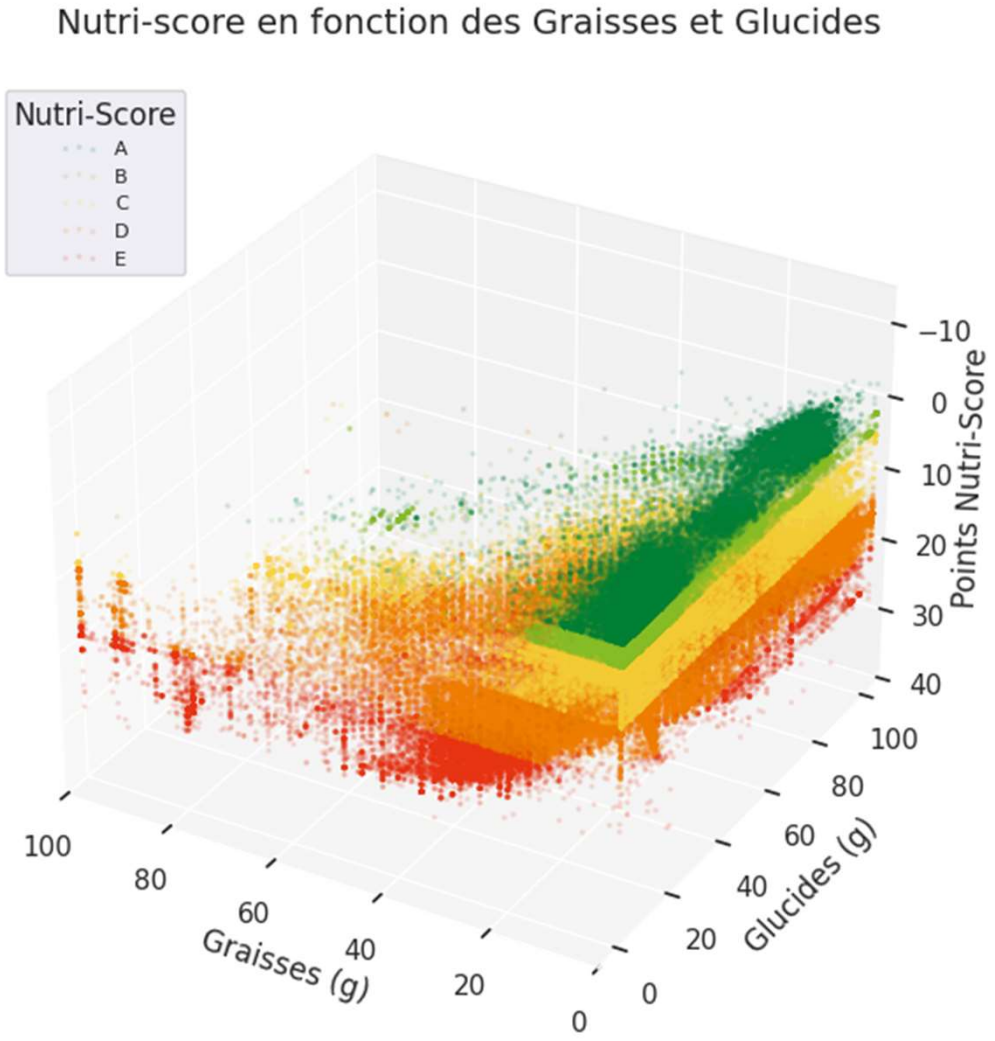
Nous voyons ici que les protéines ont tendance à augmenter quand la quantité de fibre augmente.

Ce qui nous indique un lien entre ceux-ci.

Bien que Le nutri-score soit lié à la valeur énergétique des produits, nous pouvons voir qu'il y a d'autres éléments qui agissent sur son score.

L'énergie est une combinaison entre les glucides et les graisses.





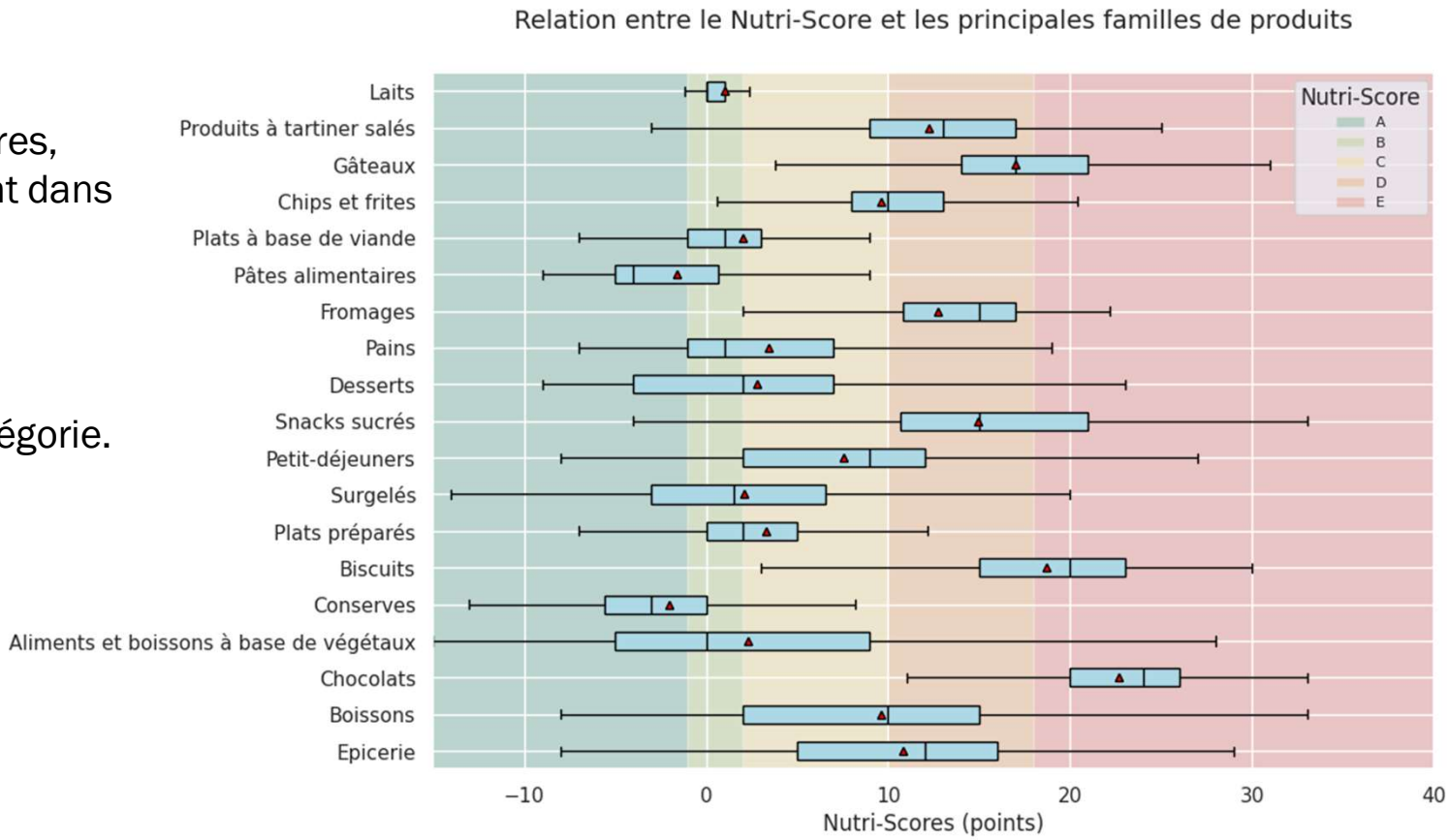
Nous pouvons maintenant observer que les glucides influencent beaucoup moins le Nutri-Score que les graisses.

Un produit ayant beaucoup de glucides peut avoir une catégorie A, ce qui n'est pas le cas pour les graisses.

Répartition des Nutri-Scores pour différentes familles de produits alimentaires.

Des catégories plus saines que d’autres, avec des Nutri-Scores majoritairement dans les catégories A et B.

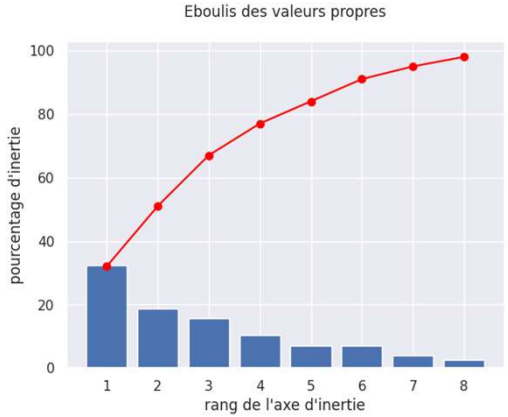
Un lien existe donc entre score et catégorie.



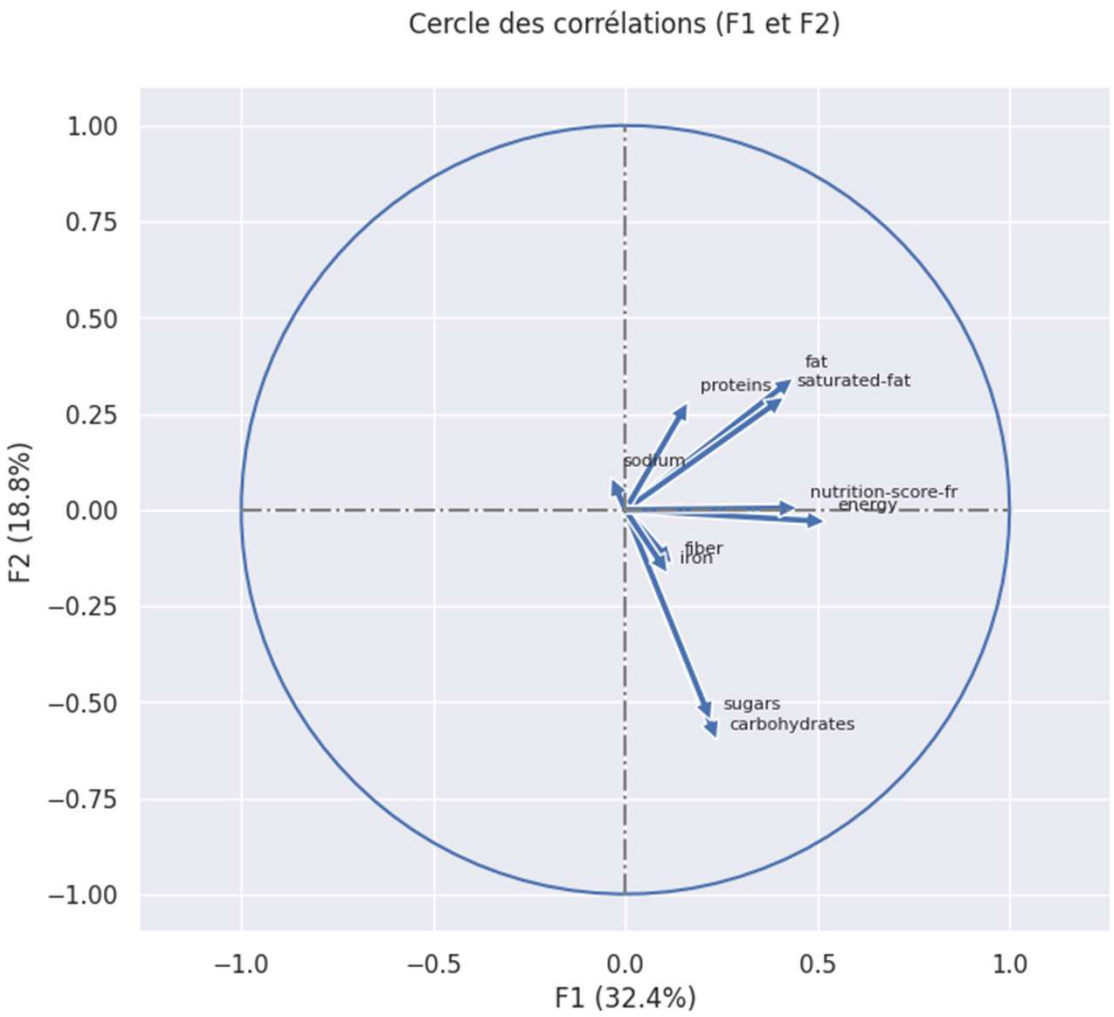
Analyse en composantes principales

Le heatmap de l'ACP met en évidence les influences des variables nutritionnelles sur les composantes principales.

Les composantes F1 et F2 captent 51% de la variance totale des données.



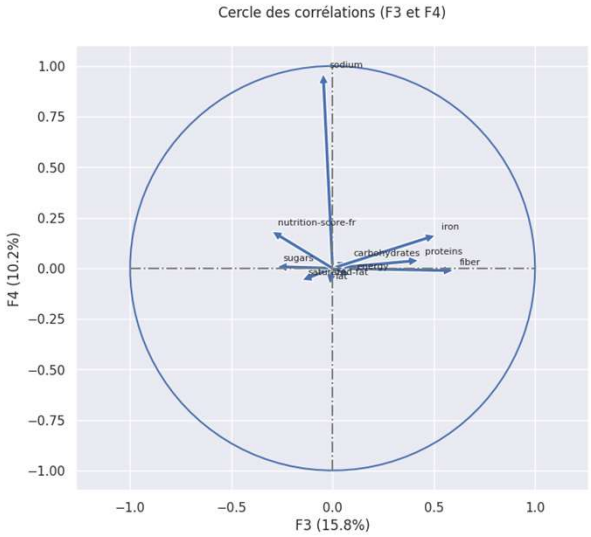
	F1	F2	F3	F4	F5	F6	F7	F8
fat	0.44	0.34	-0.01	-0.08	0.12	0.27	-0.33	-0.31
saturated-fat	0.41	0.29	-0.15	-0.06	0.12	0.11	0.53	0.62
carbohydrates	0.24	-0.60	0.07	0.03	-0.16	-0.03	-0.30	0.49
sugars	0.22	-0.55	-0.27	0.01	-0.09	-0.14	0.33	-0.32
fiber	0.12	-0.14	0.60	-0.01	-0.30	0.55	0.39	-0.21
proteins	0.16	0.28	0.43	0.04	-0.46	-0.68	0.06	0.00
sodium	-0.04	0.08	-0.05	0.96	-0.10	0.15	-0.05	0.07
iron	0.11	-0.17	0.51	0.16	0.78	-0.23	0.07	-0.04
energy	0.52	-0.03	0.09	-0.03	-0.08	0.08	-0.45	0.04
nutrition-score-fr	0.45	0.00	-0.30	0.18	0.04	-0.22	0.21	-0.36

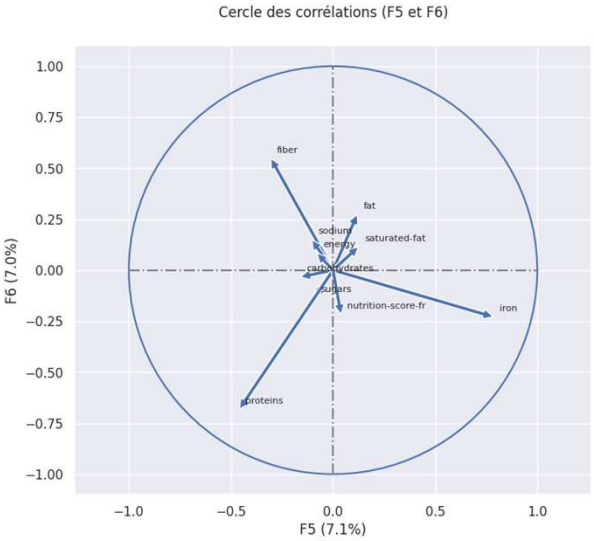


Représentation des variables en fonction des deux premières composantes
Corrélation entre les variables :

- Energie
- Nutri-Score
- Graisse

- Sucres
- Glucides





Forte corrélation entre F4 et le sodium.

Relations plus complexes avec le cercle des corrélations F5/F6.

Caractéristiques spécifiques des données pas immédiatement apparentes dans les premières composantes.

Observations utiles pour affiner la compréhension des données.

Rapports de corrélation

Catégories de produit et nutriments :
Graisses (0.7), Graisses saturées (0.68), Glucides (0.73), Sucres (0.63), Protéines (0.65) :
Forte corrélation entre catégories de produits et nutriments.
Catégories de produits ont des profils distincts en termes de contenu.

Catégories de produit et autres variables :
Energie (0.73) :
Forte corrélation, logique puisque l'énergie est liée au nutriments (graisses, glucides, protéines).

Nutri-score (0.59) :
Corrélation modérée, Score nutritionnel varie en fonction des catégories.

Résultats d’analyse

Forte influence des catégories de produits sur la répartition des nutriments et l'énergie. (graisses, graisses saturées, glucides, sucres, protéines)

Crucial pour la segmentation des produits et la suggestion de catégories de produits.

Moindre variation des fibres et du fer selon les catégories.

Suite

Choix de la technique de modélisation.

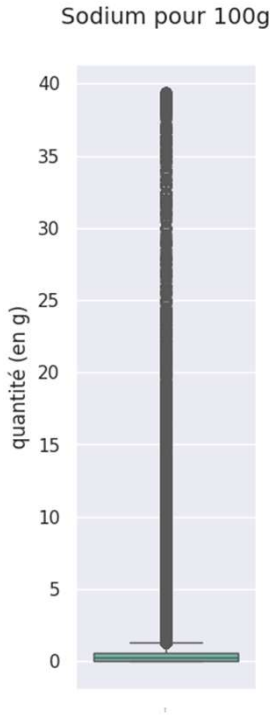
Processus d'évaluation rigoureux.

Comparaison des performances sur différents sous-ensembles de données.

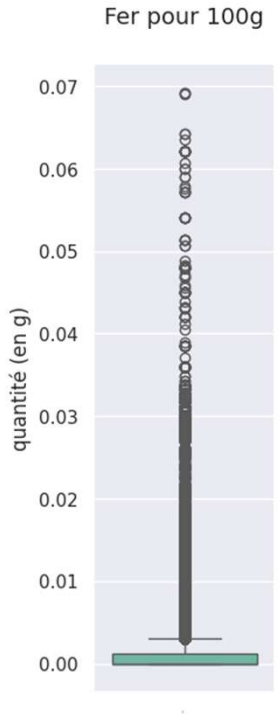
Validation de la robustesse des modèles avec validation croisée.

Annexe

Détail de la distribution du sodium et du fer pour 100g



Médiane 0.23g
Valeur max 39g



Médiane 0mg
Valeur max 70mg

Merci pour votre attention