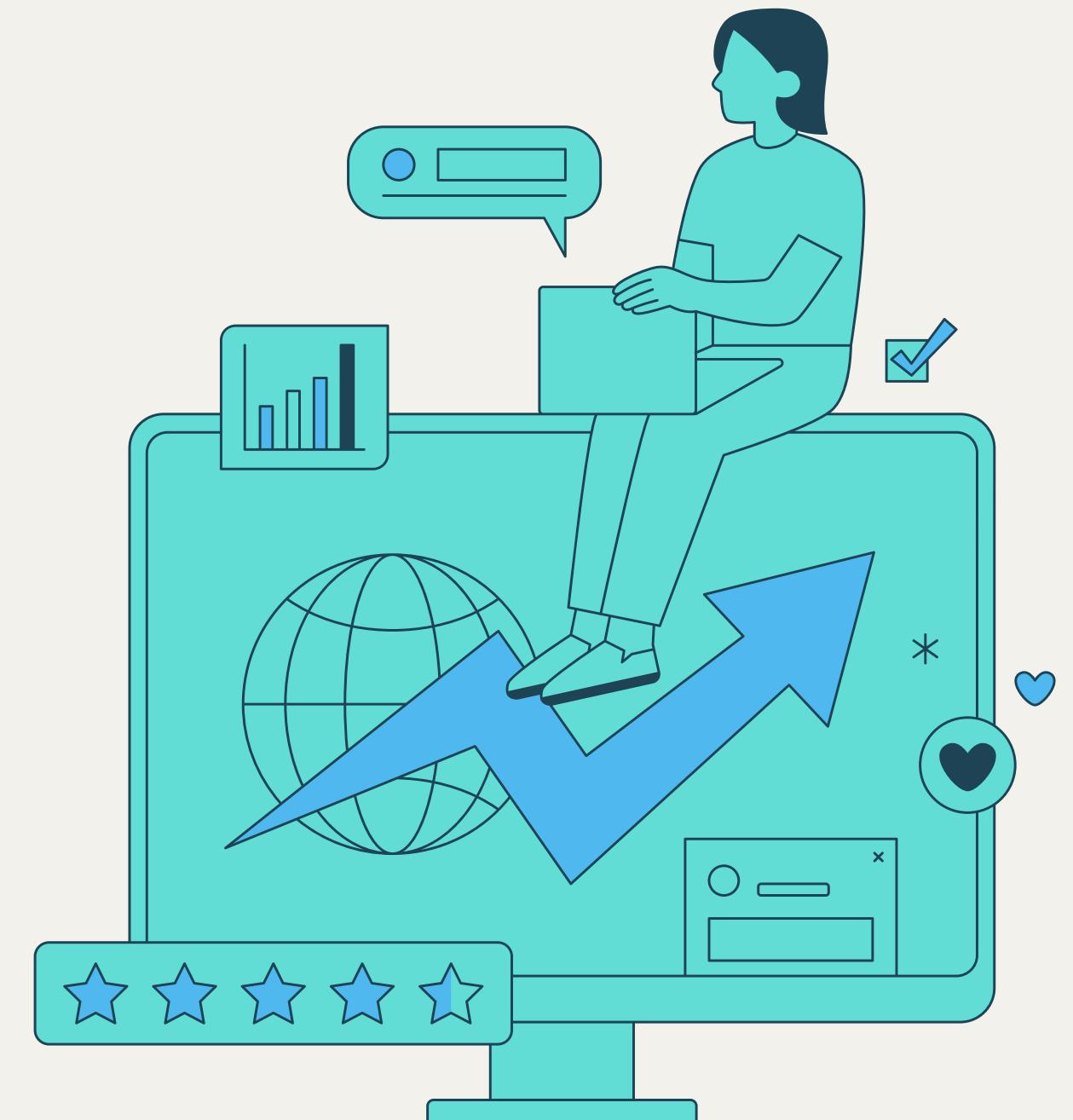


Améliorer la base de données Open Food Facts

Santé Publique France

Février 2025

Natascha Minnitt



La mission

1

Nettoyage et exploration
des données

2

Évaluation de la
prédictibilité des valeurs
manquantes

3

Conclusion sur la
faisabilité de l'outil de
suggestion automatisé

Nettoyage et exploration des données



Filtrage des données

Traitement des données



1 Approche métier :

- Suppression des produits non consommés en France
- Sélection des variables qui renseignent sur la qualité nutritionnelle des produits

2 Approche technique :

- Sélection d'une variable de référence (pnns_groups_1)
- Suppression des colonnes vides contenant plus de 90 % de valeurs manquantes

Structure avant :

- Lignes : 320 772
- Colonnes : 162 (106 catégorielles et 56 numériques)

Structure après :

- Lignes : 62 614
- Colonnes : 14 (4 catégorielles et 8 numériques)

L'identification

- [1] Nom du produit
- [2] Marques

Catégorisation

- [3] Groupes PNNS 1
- [4] Note nutritionnelle

Informations nutritionnelles

- [5] Matières grasses saturées (100g)
- [6] Sucres (100g)
- [7] Sodium (100g)
- [8] Fibres (100g)
- [9] Protéines (100g)

Calcul de l'énergie

- [11] Matières grasses (100g)
- [12] Glucides (100g)
- [13] Énergie (100g)

Traitement des données

Valeurs aberrantes

1 Seuils

- Matières grasses saturées (100g) : 35
- Sucres (100g) : 90
- Sodium (100g) : 5
- Fibres (100g) : 90
- Protéines (100g) : 90
- Matières grasses (100g) : 100
- Glucides (100g) : 100
- Énergie (100g) : 3700

	saturated-fat_100g	sugars_100g	sodium_100g	fiber_100g	proteins_100g	fat_100g	carbohydrates_100g	energy_100g
count	47667.0	47782.0	47843.0	31301.0	49520.0	46669.0	46239.0	49778.0
mean	5.3	13.0	0.4	2.7	7.6	13.3	27.7	1165.8
std	8.2	18.6	1.6	4.1	7.5	16.9	27.4	14611.9
min	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0
25%	0.3	1.0	0.0	0.3	1.7	1.3	4.0	410.0
50%	1.9	4.0	0.2	1.7	6.0	6.8	14.5	1011.0
75%	7.3	16.5	0.5	3.5	10.8	21.0	53.0	1636.0
max	210.0	105.0	83.0	178.0	100.0	380.0	190.0	3251373.0

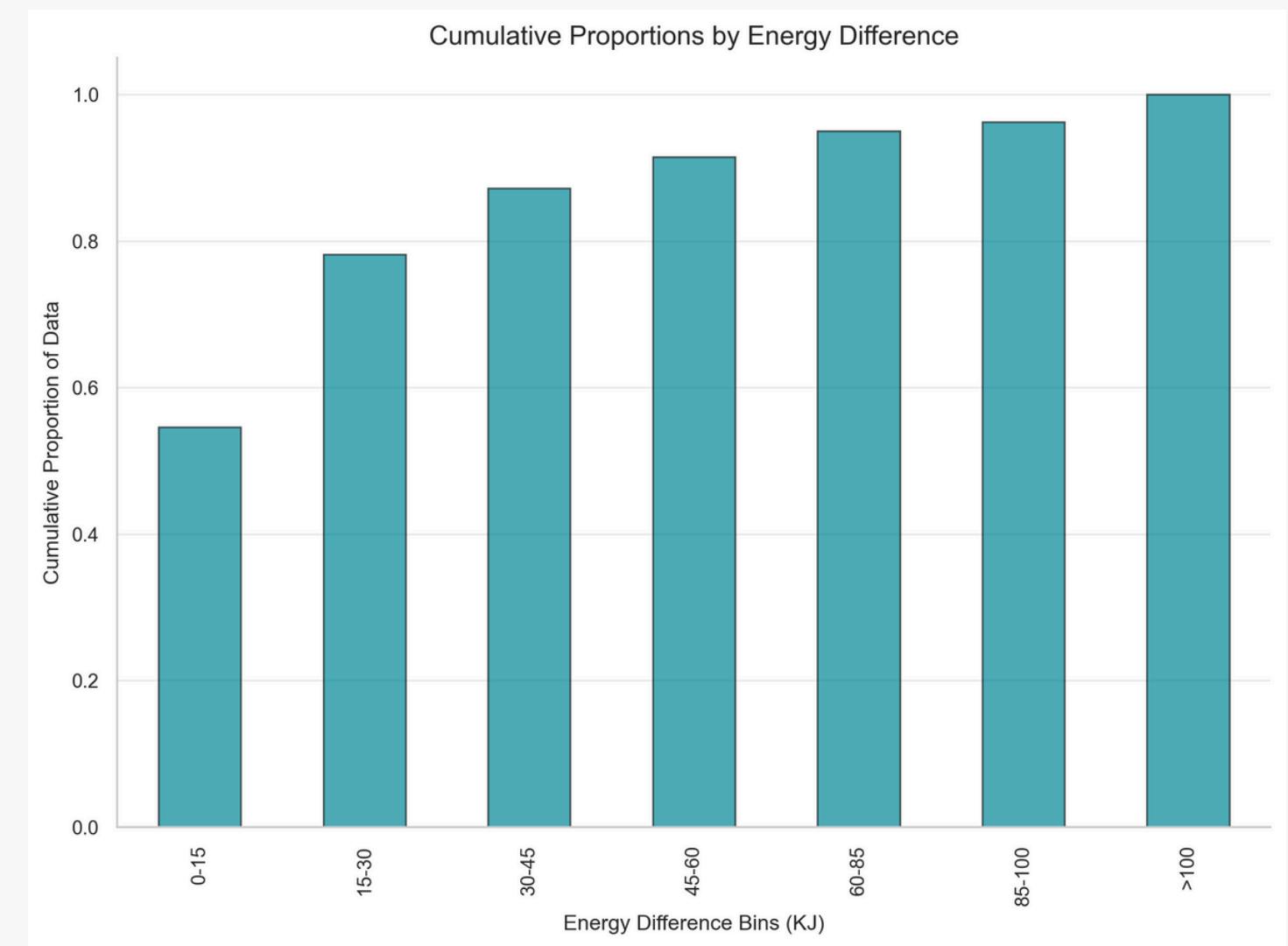
2 Différences et somme

- Matières grasses saturées \leq Matières grasses
- Sucres \leq Glucides
- Macronutriments $\leq 100\text{g}$

3 Score calculé

Énergie (kJ/100g) :

- Matières grasses : 37 kJ par gramme
- Protéines : 17 kJ par gramme
- Glucides : 17 kJ par gramme

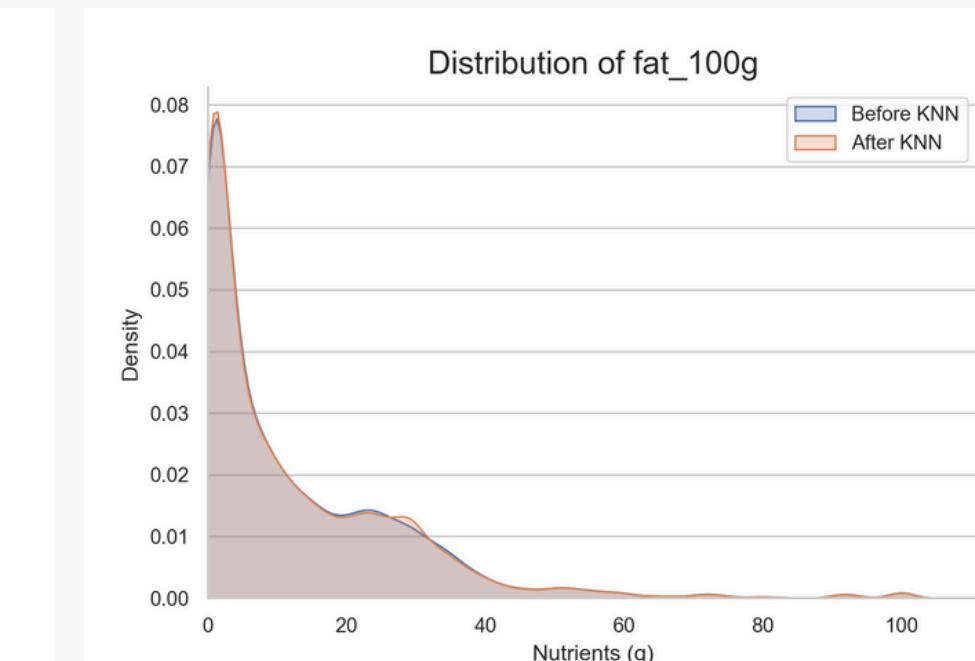
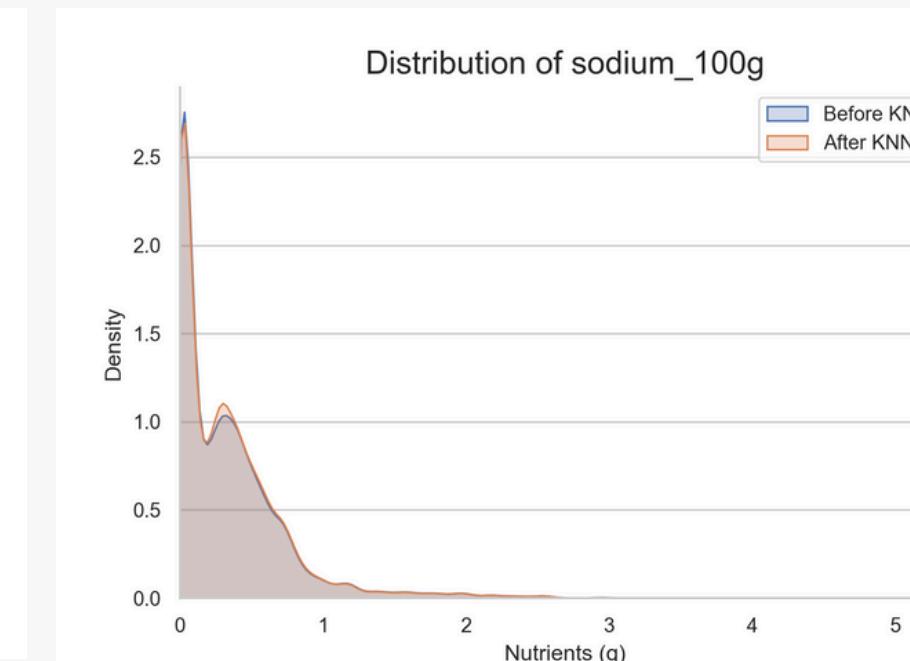
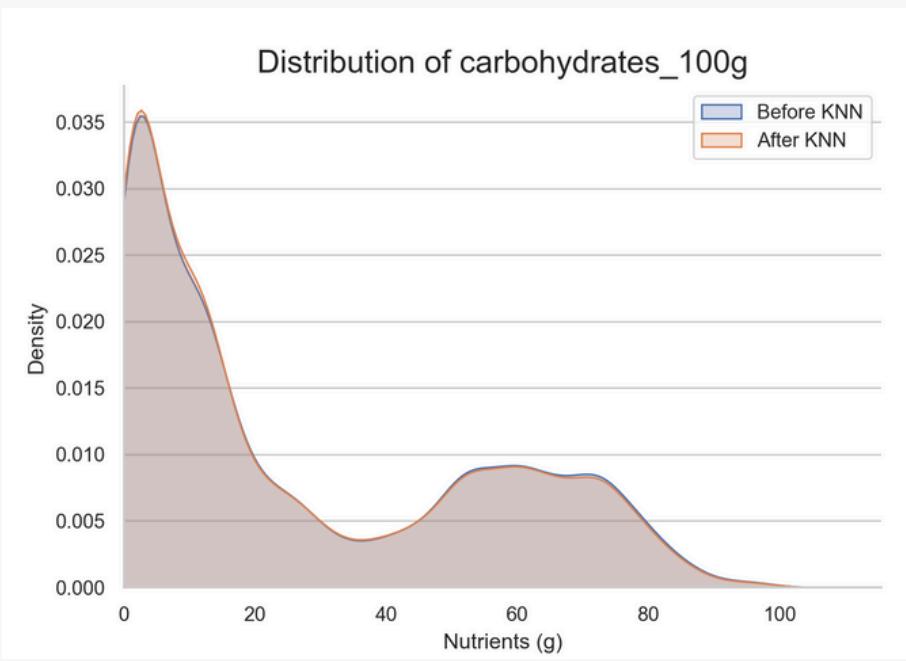
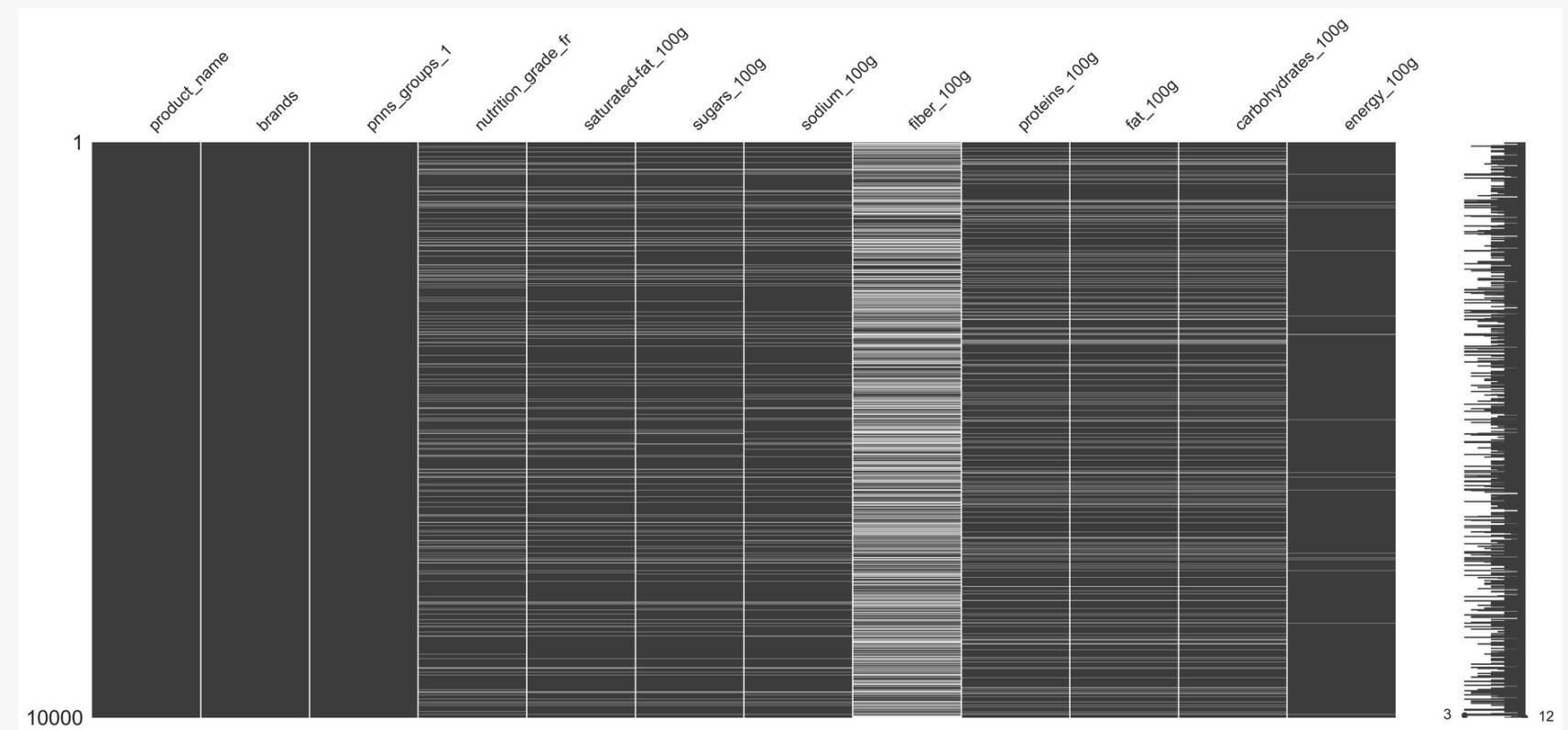


Données manquantes

Traitement des données

1 Remplissage des colonnes numériques

- Remplacement des valeurs NaN de la colonne Fibres par 0
- Remplissage des colonnes numériques en utilisant l'imputation KNN



Données manquantes

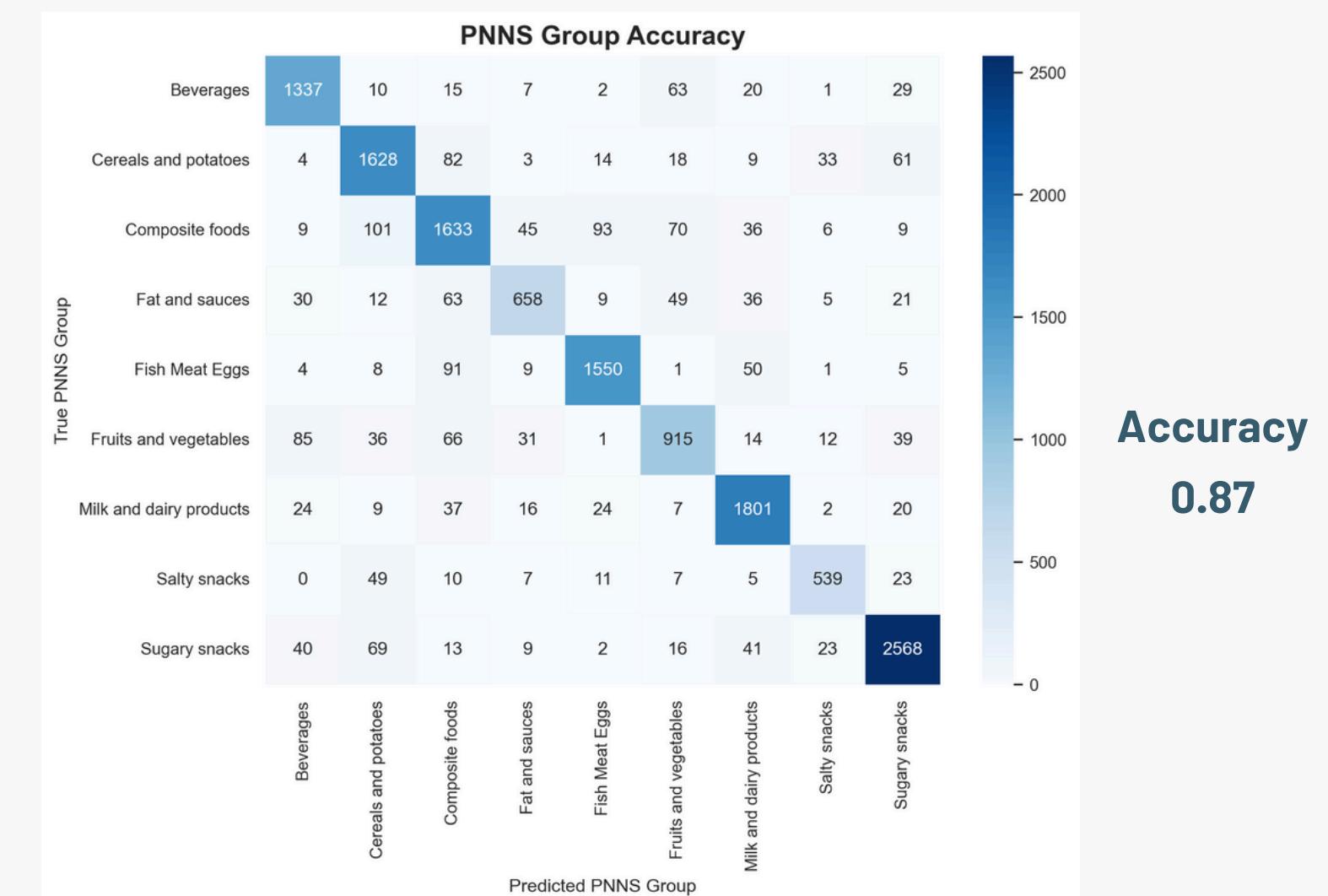
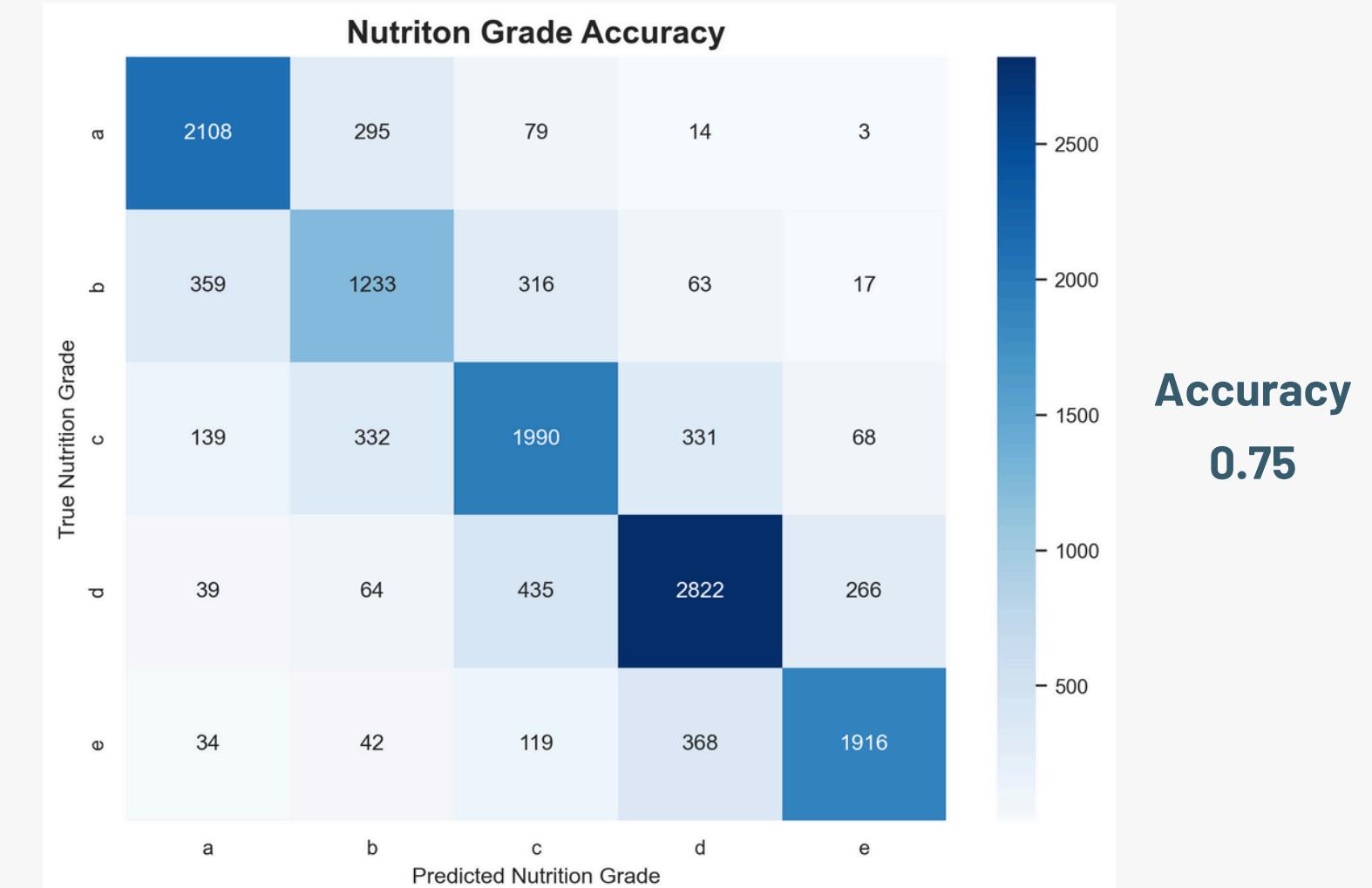
Traitement des données



2

Remplissage des colonnes catégoriques

- Remplissage des notes nutritionnelles manquantes (3 527) en utilisant KNeighborsClassifier
 - Matières grasses saturées (100g)
 - Sucres (100g)
 - Sodium (100g)
 - Fibres (100g)
 - Protéines (100g)
- Remplissage des groupe pnns 'unknown' (7,055) en utilisant KNeighborsClassifier
- Suppression des doublons (3 151)
 - Numériques = médiane
 - Catégorielles = mode

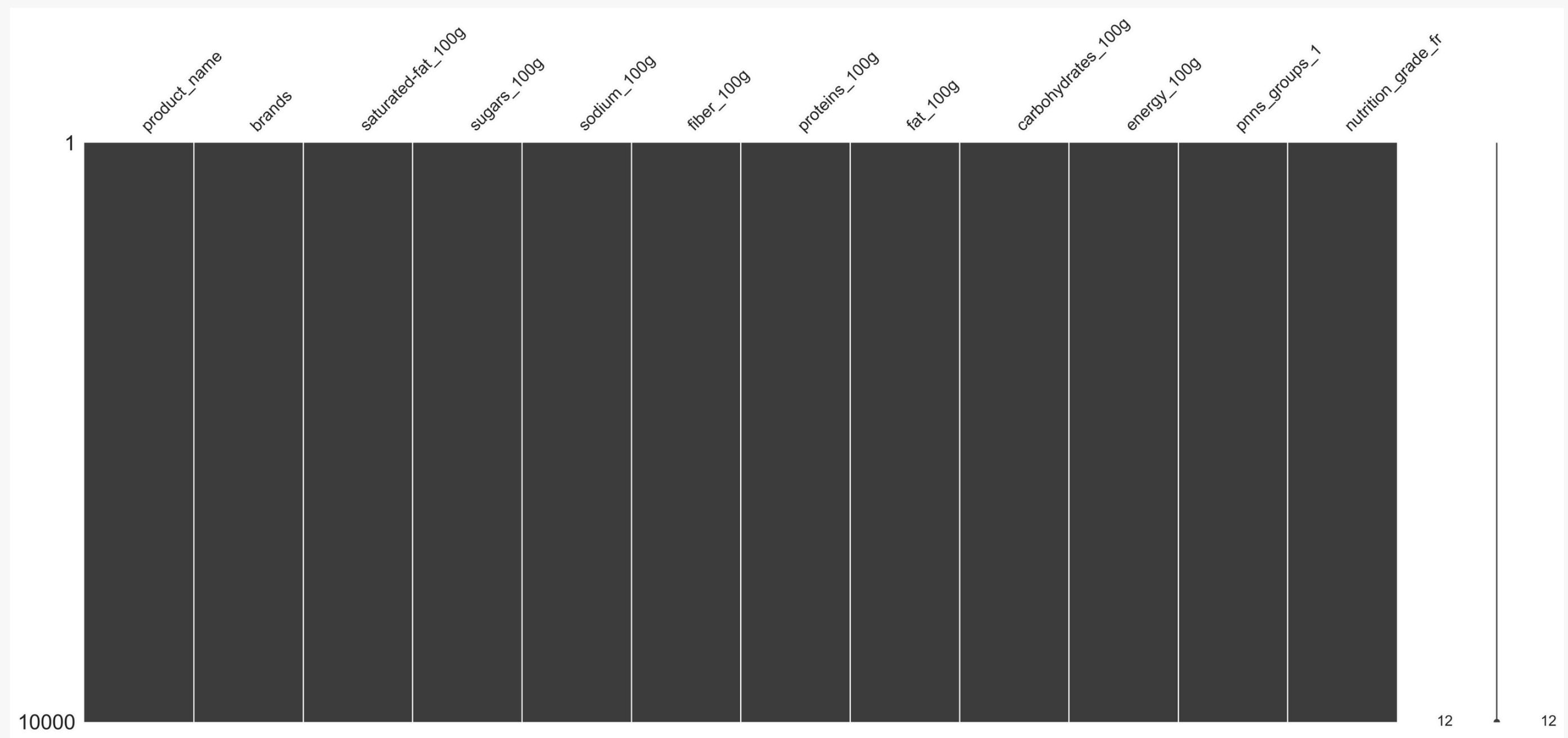


Base de données traitée

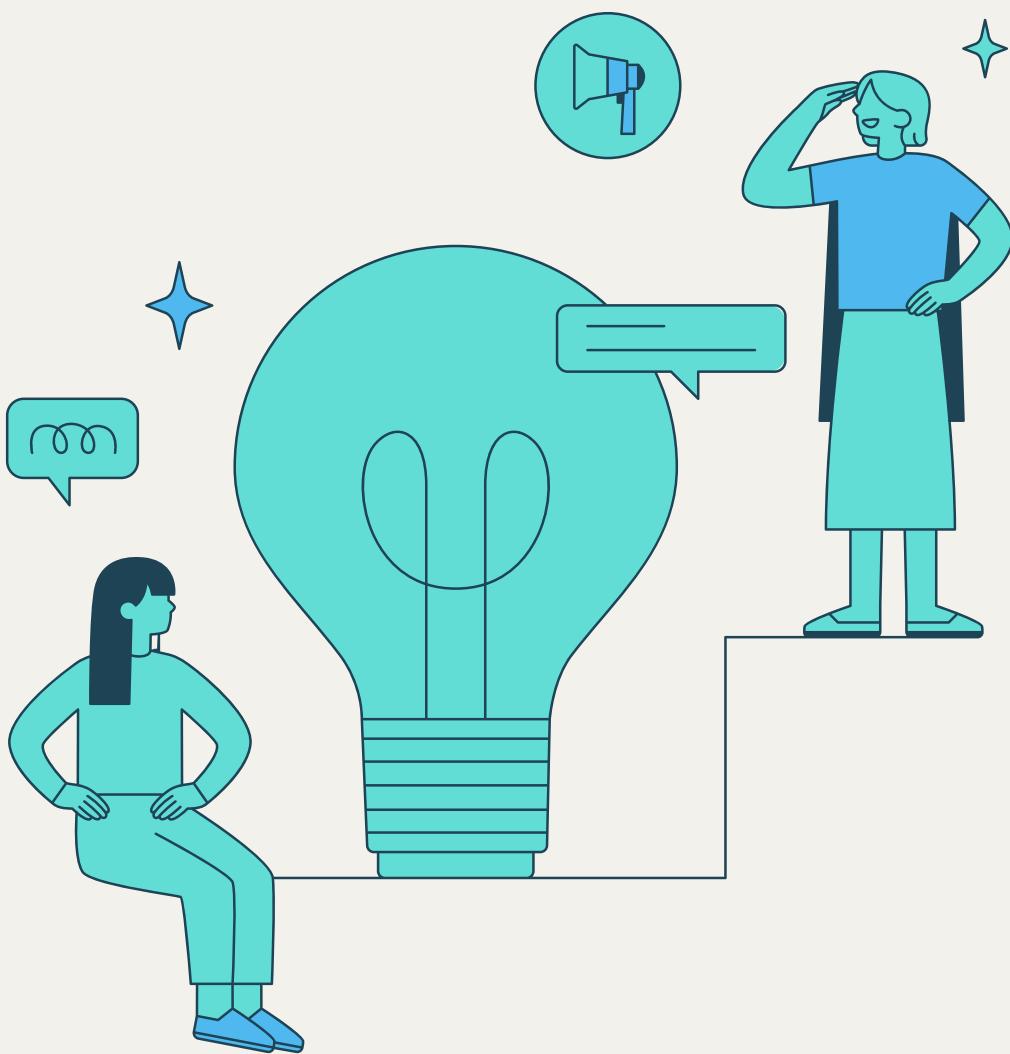
Traitement des données

Structure :

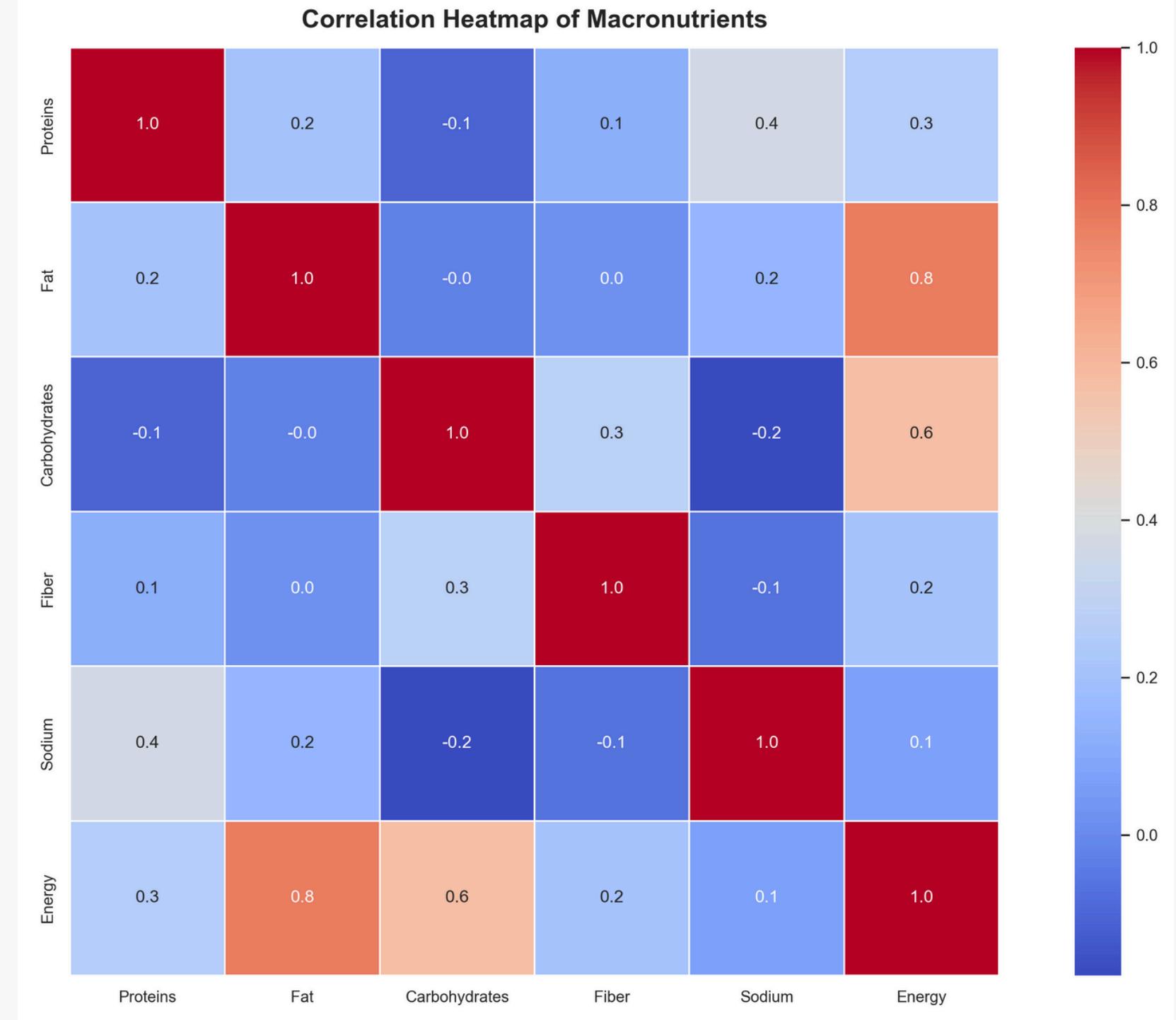
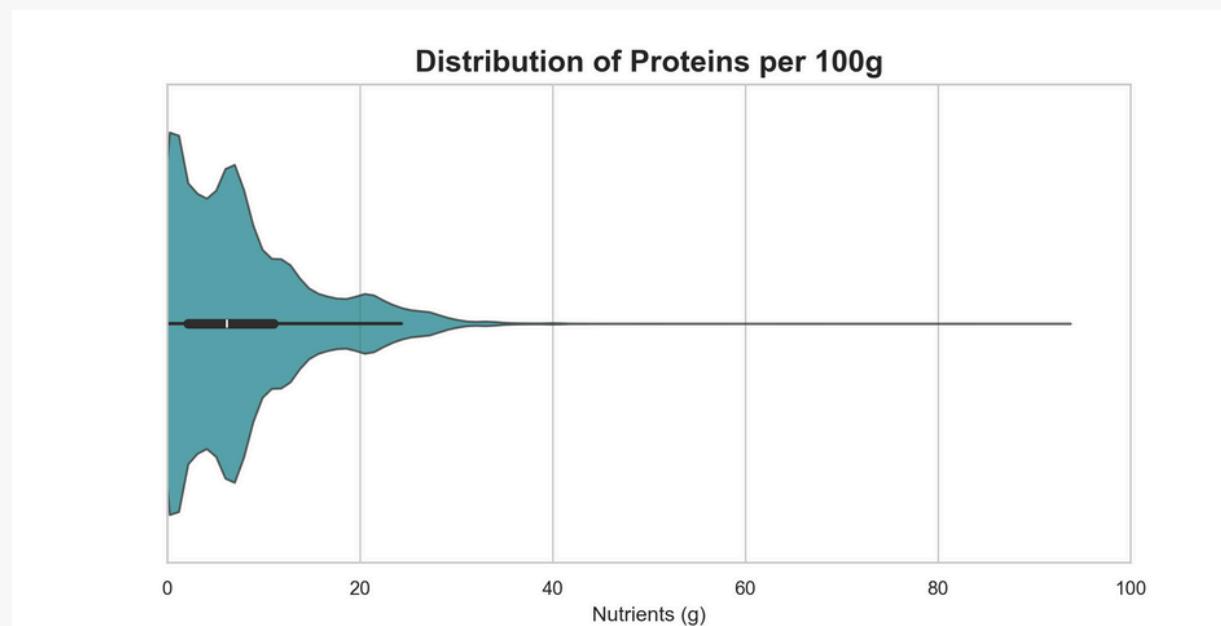
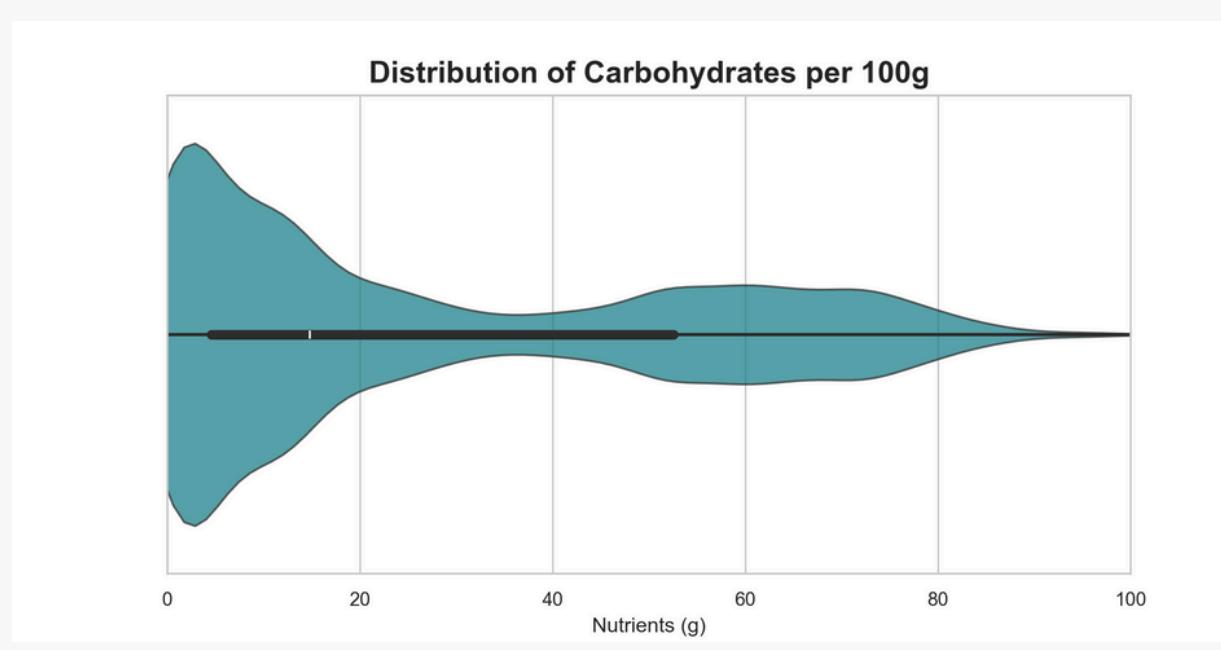
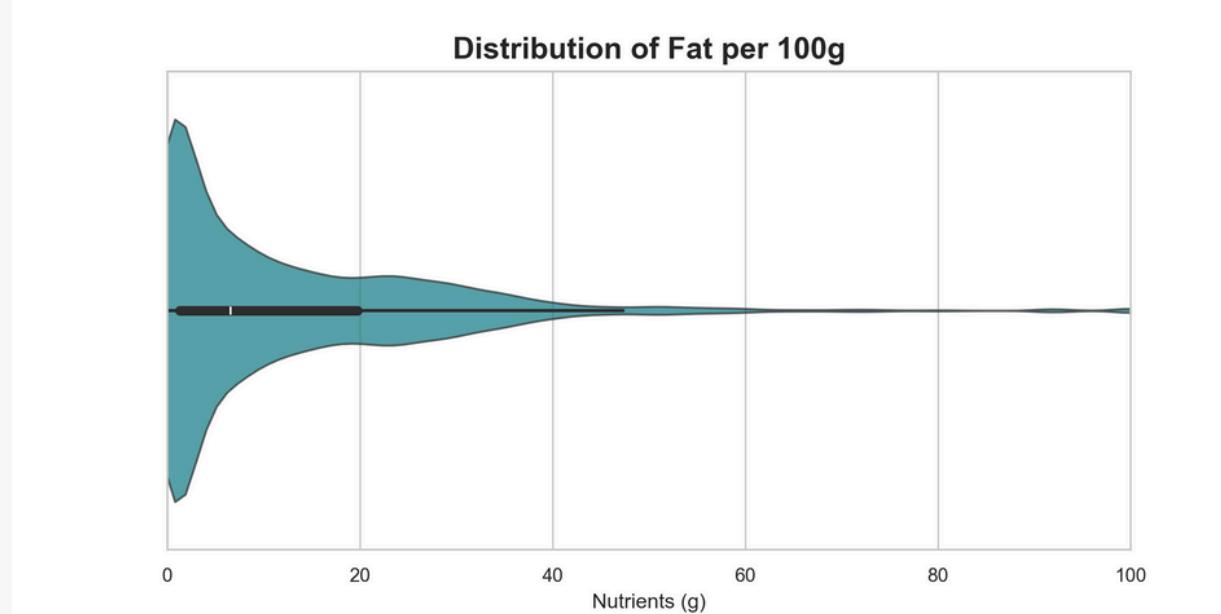
- Lignes : 46 594
- Colonnes : 12 (4 objets et 8 nombres)



Analyse exploratoire



Analyse univariée

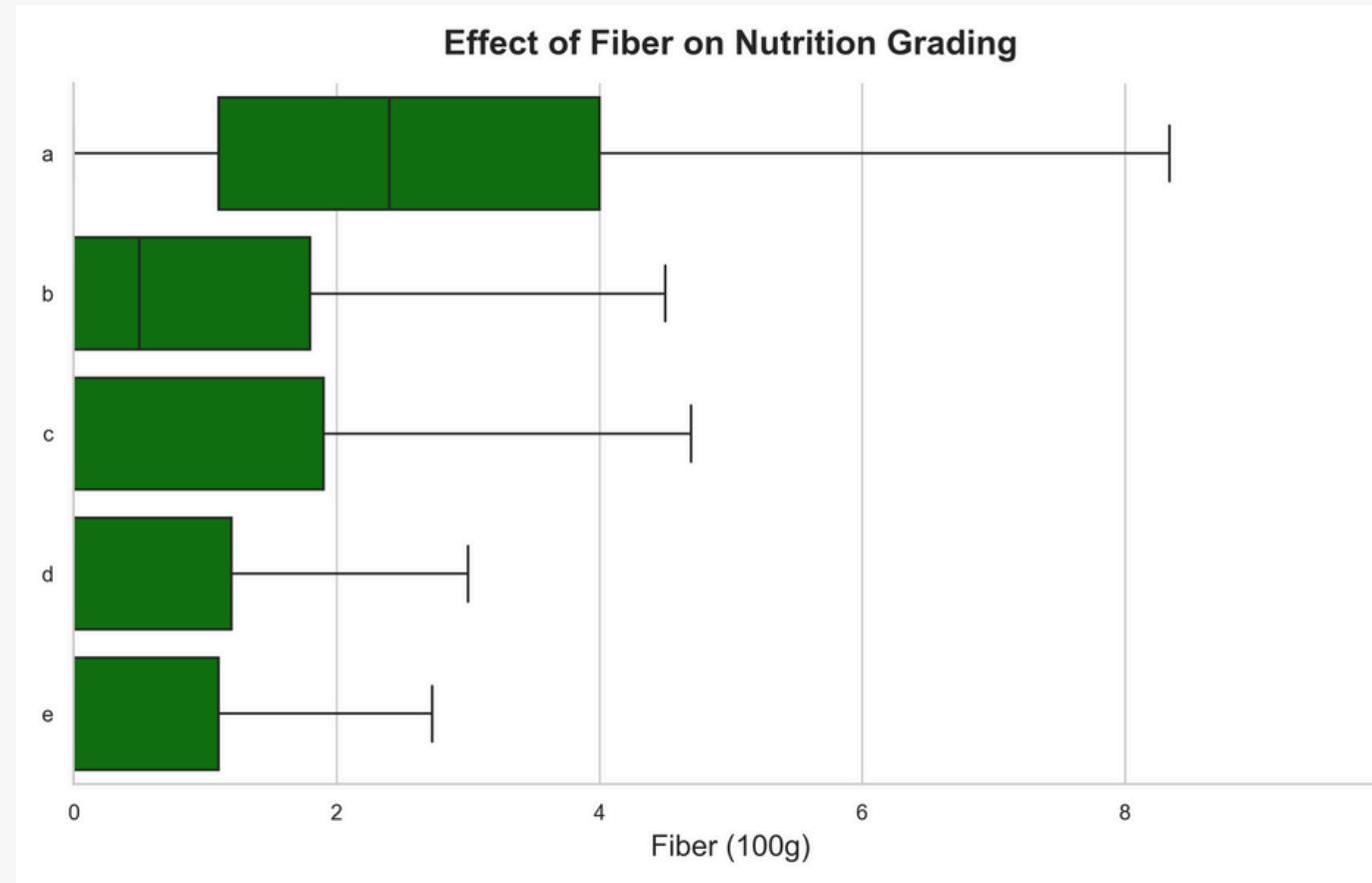


Analyse exploratoire

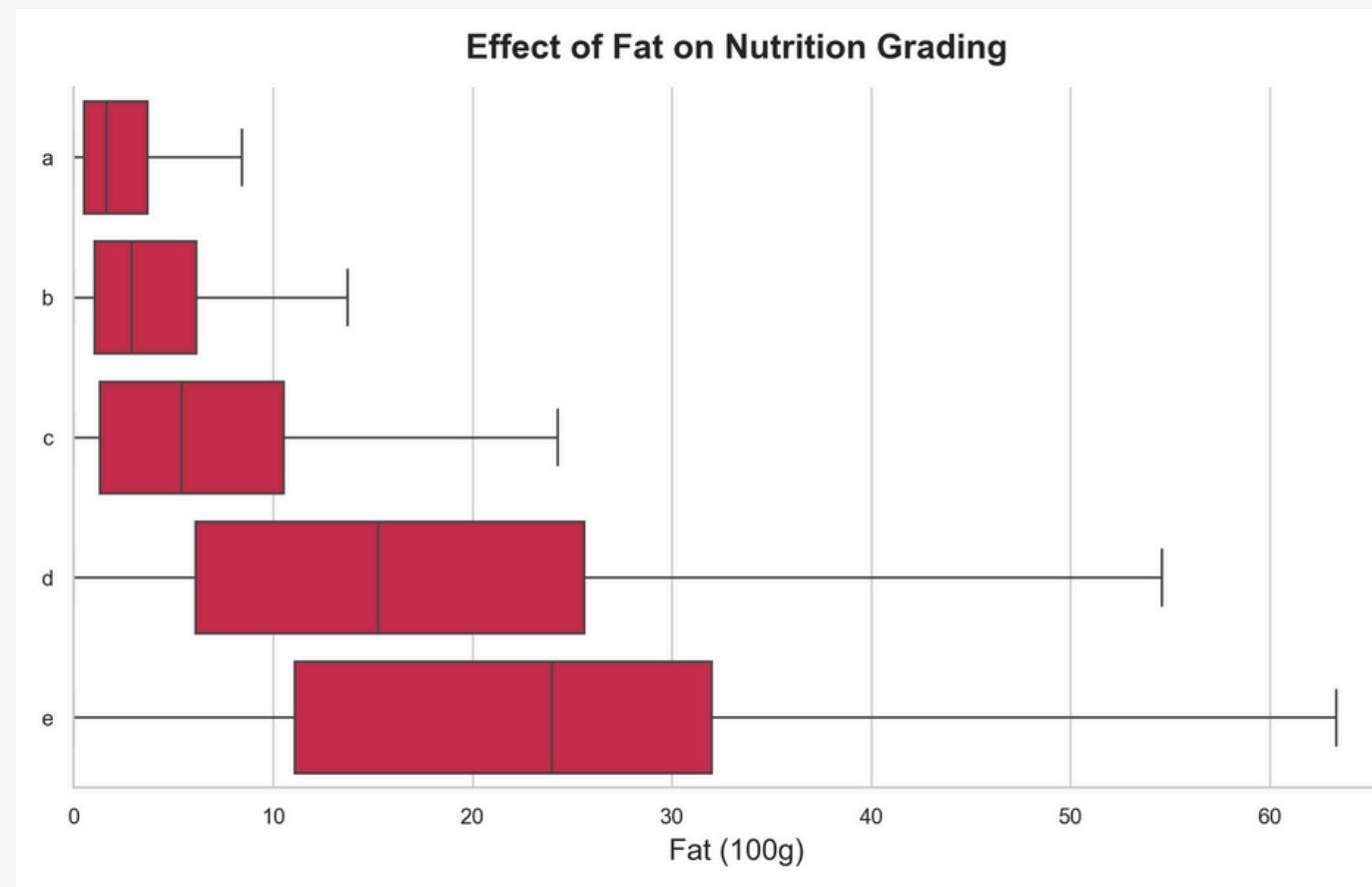


Analyse bi-variée

Analyse exploratoire



Plus de fibres => meilleur score nutritionnel



Plus de matières grasses
=> moins bon score nutritionnel

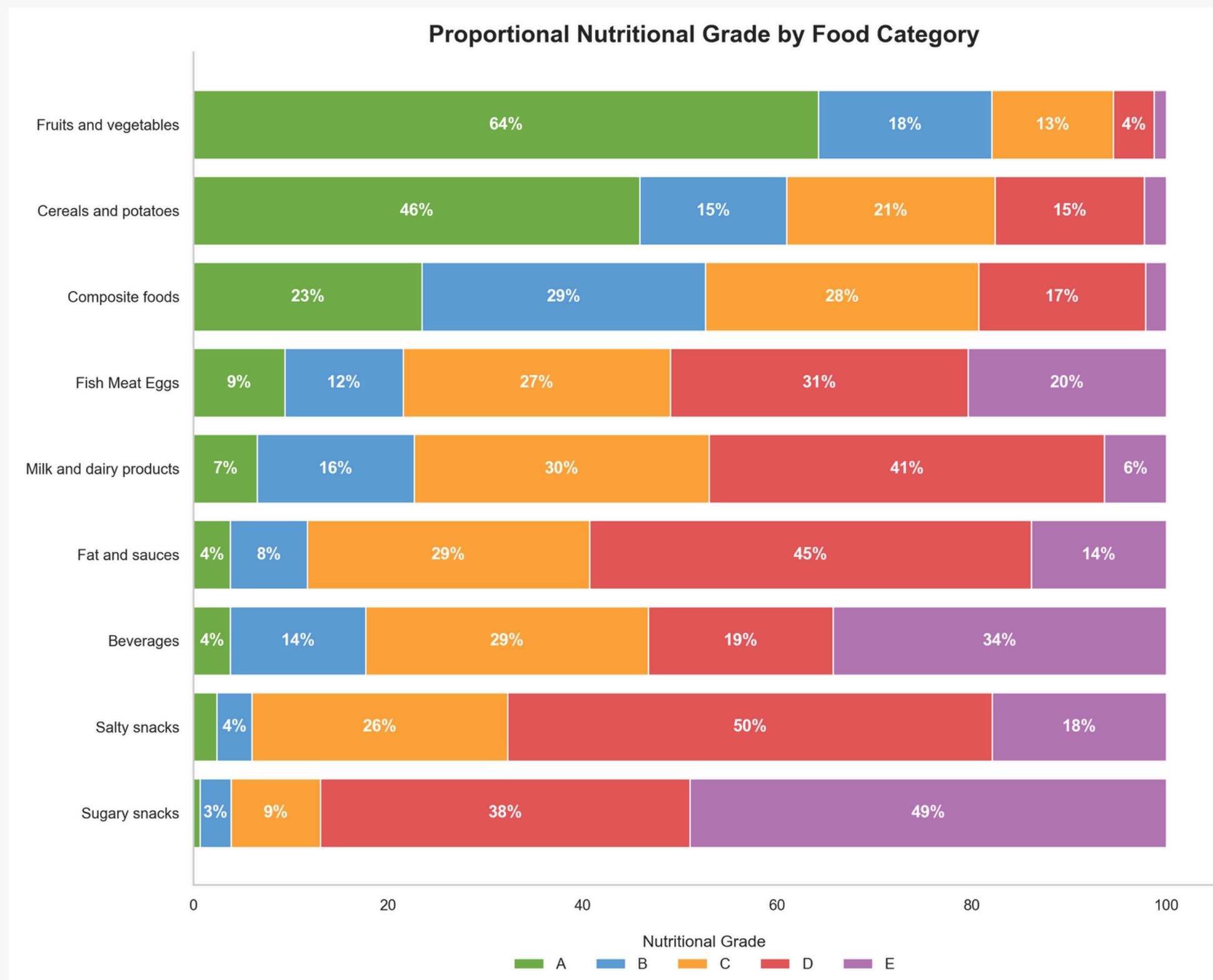


Analyse bi-variée

Analyse exploratoire



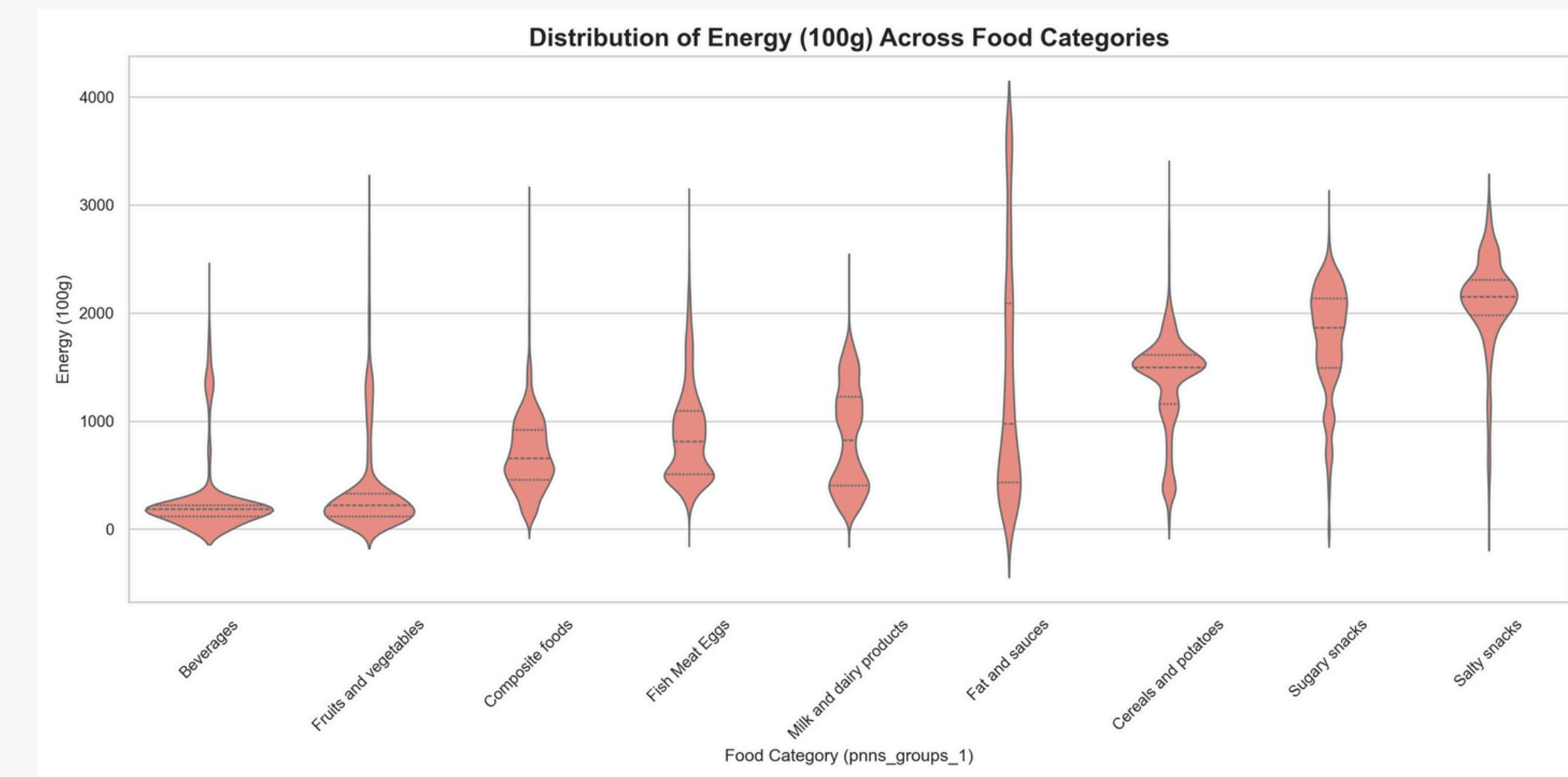
Quelles catégories d'aliments ont la plus grande proportion de produits sains (score A) ou malsains (score E) ?



Analyse multi-variée

Analyse exploratoire

Les différences d'énergie peuvent-elles être expliquées par la catégorie alimentaire ?



- L'hypothèse nulle (H_0) était que le score moyen d'énergie est identique dans toutes les catégories alimentaires.
 - Le test de Kruskal-Wallis a donné une statistique de 26090,35 ($p < 0,0$)
 - Nous rejetons l'hypothèse nulle.
- Le test post hoc de Dunn a montré que toutes les comparaisons par paires entre les catégories alimentaires sont statistiquement significatives (valeur de $p < 0,0$).



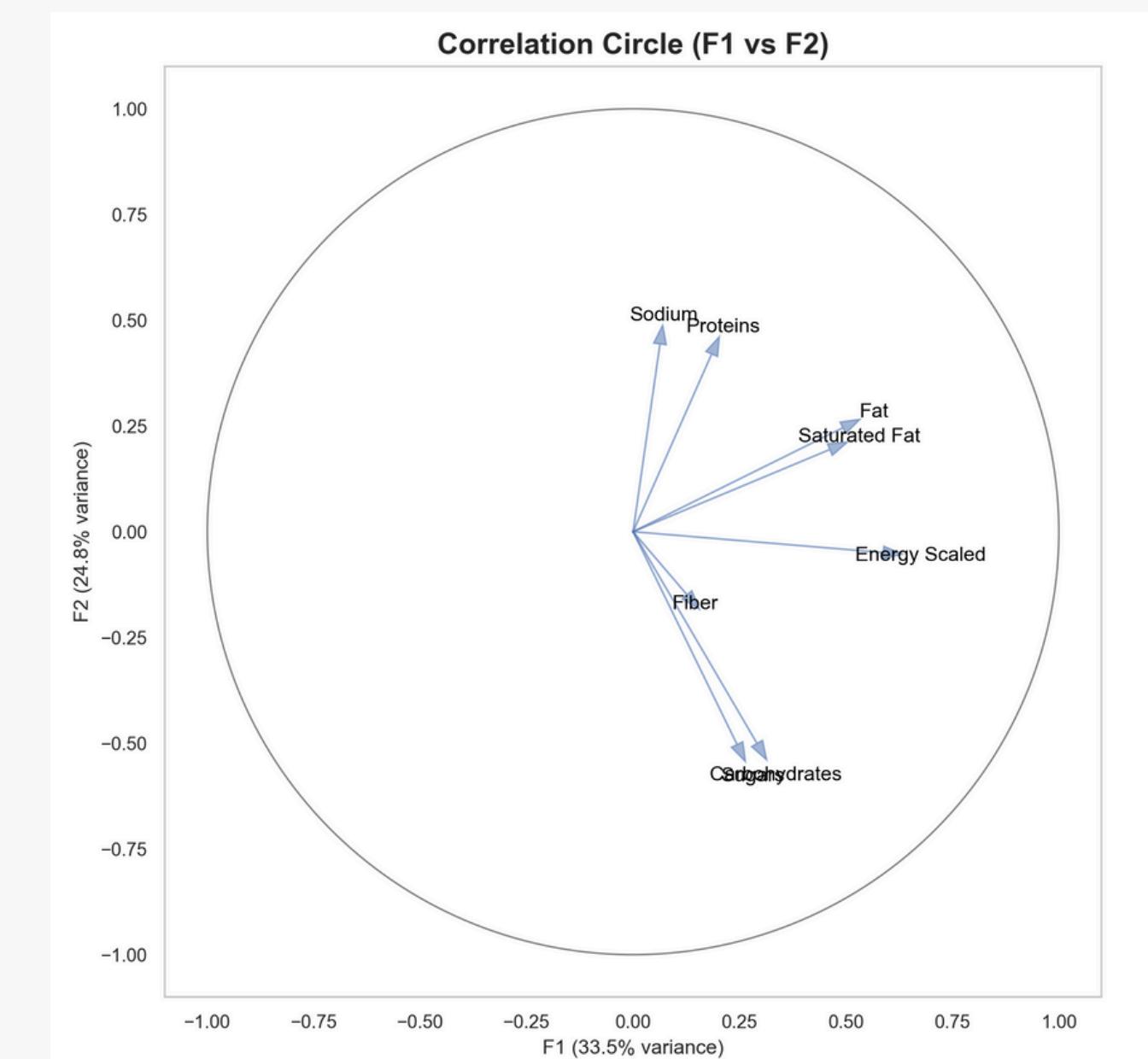
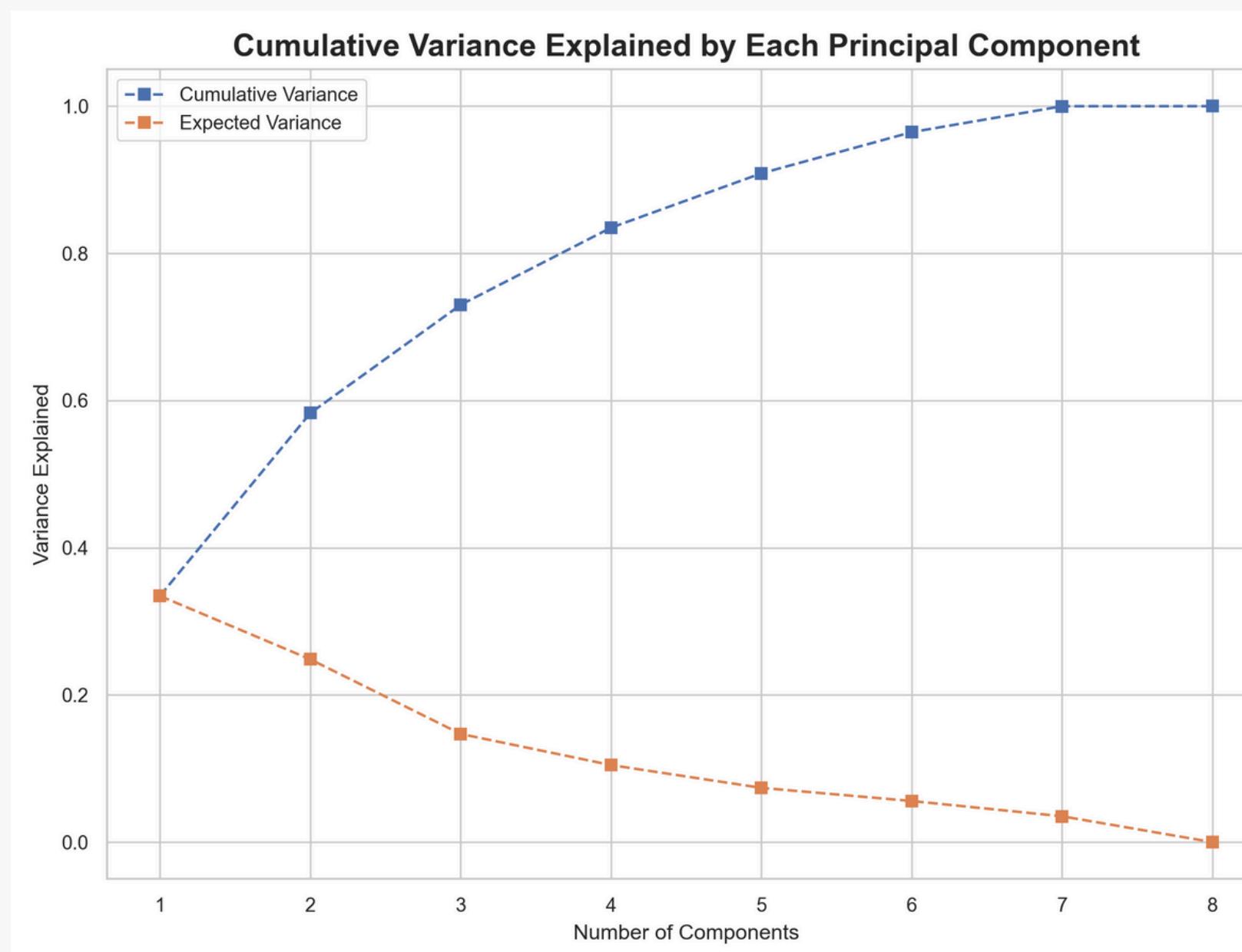
Analyse multi-variée

Analyse exploratoire



Combien de composantes principales (PCs) sont nécessaires pour expliquer la majeure partie de la variance des données ?

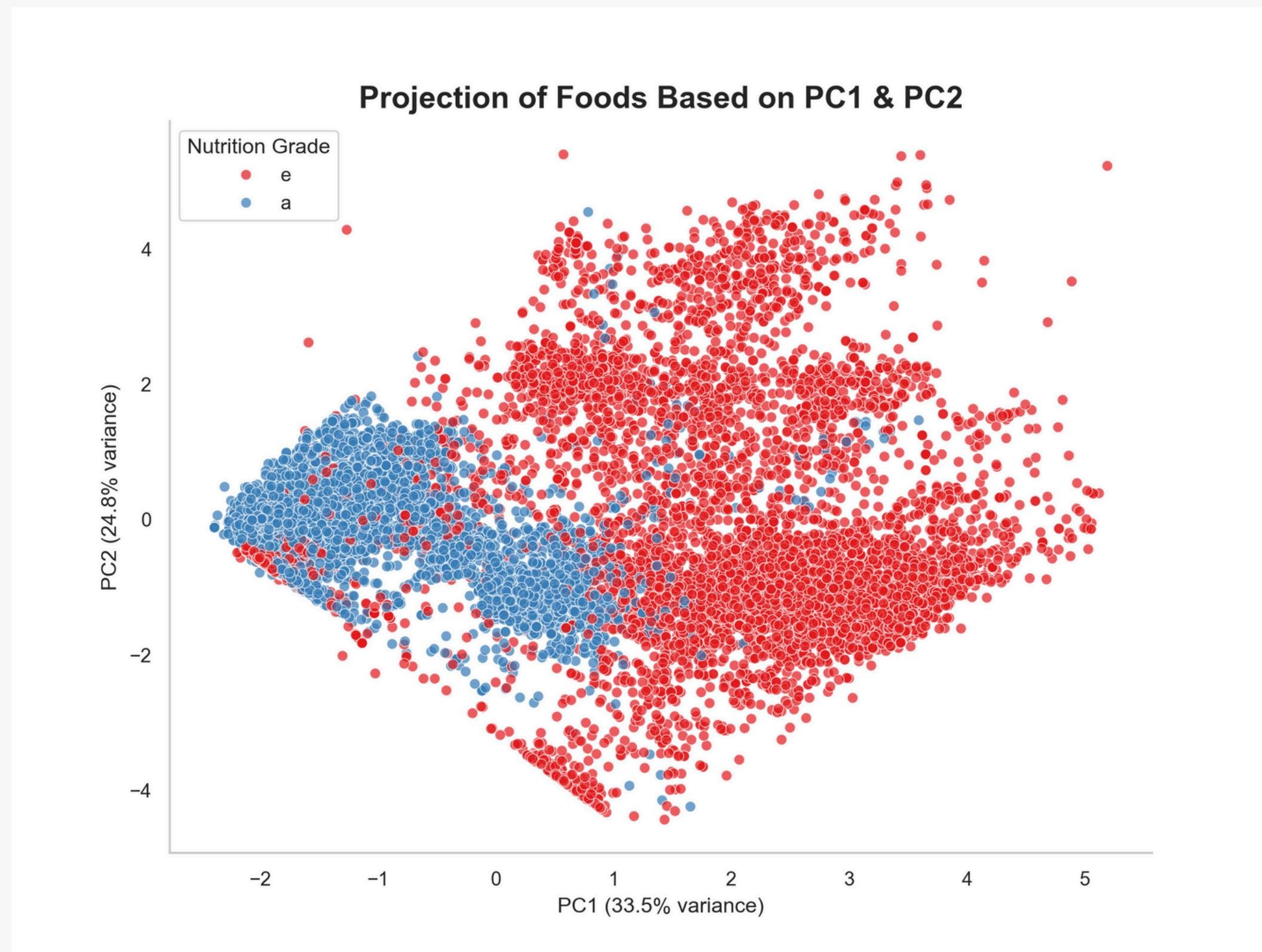
Comment chaque macronutriments contribue-t-il aux composantes principales ?



Analyse multi-variée

Analyse exploratoire

En quoi les aliments les plus sains et les moins sains (scores A et E) diffèrent-ils dans leur composition nutritionnelle ?



Conclusion

