2024

FORMATION DATA SCIENTIST

OPENCLASSROOMS

PROJET 3 : PRÉPAREZ DES DONNÉES POUR UN ORGANISME DE SANTÉ PUBLIQUE

Soutenu par:
Kourouma Sekouba Aissatou

SOMMAIRE

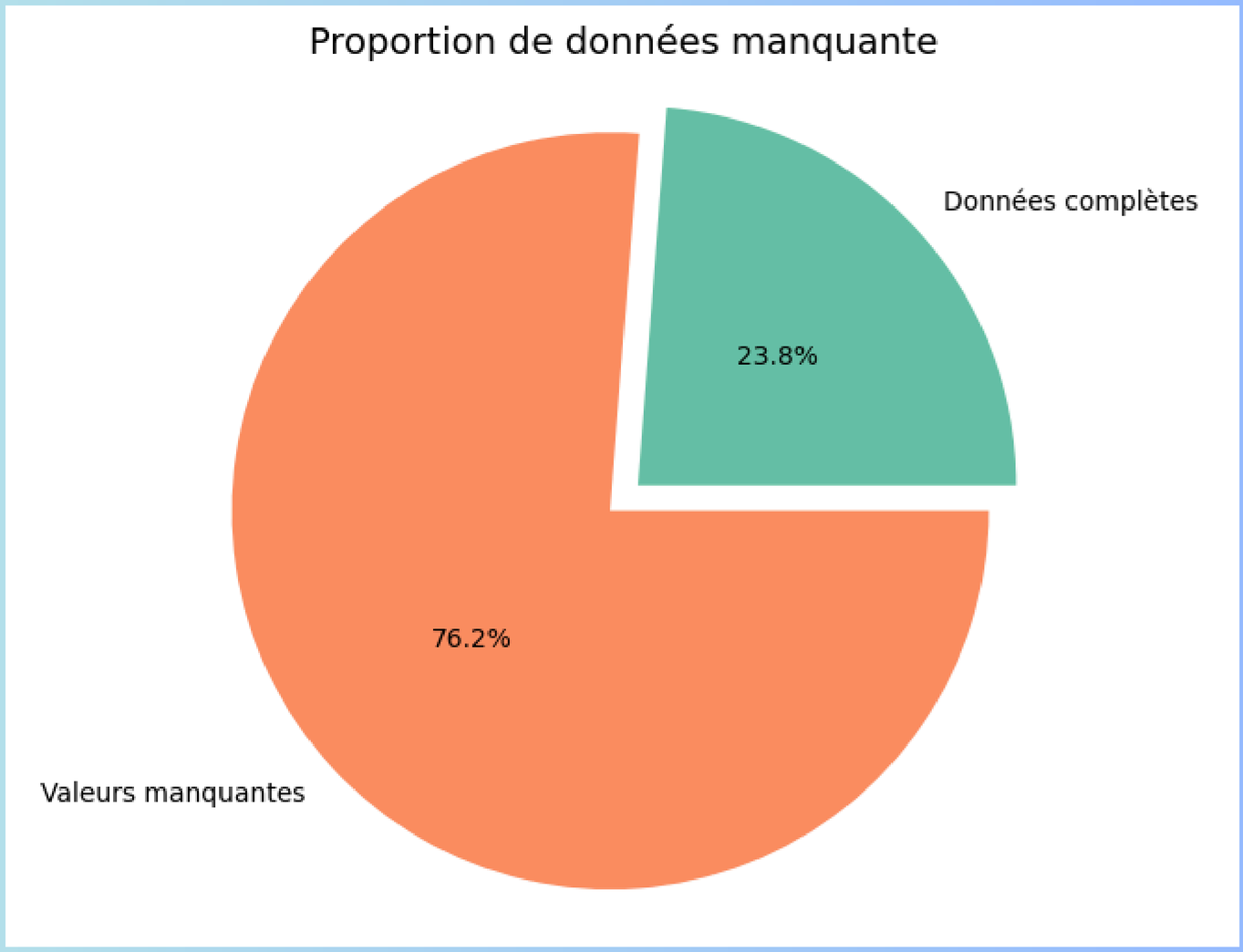
1. MISSION

L'agence **Santé publique France** souhaite améliorer sa base de données Open Food Facts et fait appel aux services de notre entreprise.

- Comment traiter les valeurs manquantes dans les données ?
- Quelles sont les méthodes les plus efficaces pour détecter et corriger les valeur aberrantes ?

2. Presentation du jeu de donnée

Le jeux de donnée contient une taille de (320772, 162), dont il pèse 808.17 Mo
le nombre de produit alimentaire est 320749
le nombre de variable est 162

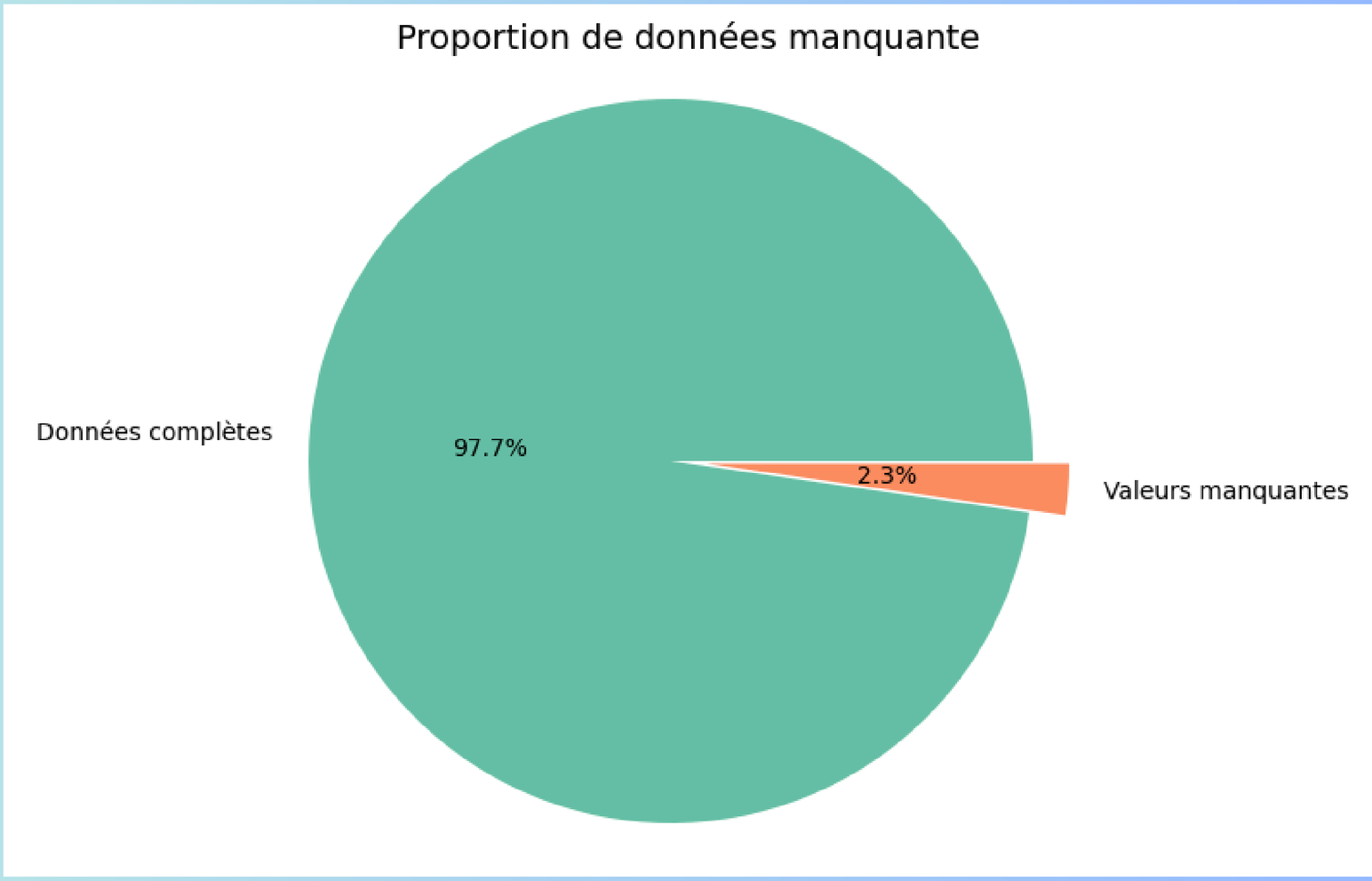


3. Nettoyage des Données

- Filtrer en fonction des variables pertinentes
    - Traitement des doublons
        - supression des lignes dont nutrition_grade_fr est NaN
        - Traitement des valeurs aberrantes et manquantes

3.1 Filtrer en fonction des variables pertinentes

- code
- countries_fr
- product_name
- energy_100g
- salt_100g
- nutrition_grade_fr_100g
- nutrition-score-fr_100g
- proteins_100g
- carbohydrates_100g
- sugars_100g
- saturated-fat_100g
- fat_100g
- pnns_groups_1
- pnns_groups_2



3.2 Traitement des doublons

	code	product_name	countries_fr	energy_100g	salt_100g	nutrition_grade_fr	score-fr_100g	proteins_100g	carbohydrates_100g	sugars_100g	saturated-fat_100g	fat_100g	pnns_groups_1	pn
189162	NaN	France	en:fruit-yogurts	NaN	NaN	NaN	NaN	0.158	NaN	NaN	NaN	NaN	NaN	
189168	NaN	France	en:stirred-yogurts	NaN	NaN	NaN	NaN	0.120	NaN	NaN	NaN	NaN	NaN	
189242	NaN	France	en:whole-milk-yogurts	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
189244	NaN	France	en:whole-milk-yogurts	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
189248	NaN	France	en:stirred-yogurts	NaN	NaN	NaN	NaN	0.120	NaN	NaN	NaN	NaN	NaN	
189345	NaN	France	en:stirred-yogurts	NaN	NaN	NaN	NaN	0.120	NaN	NaN	NaN	NaN	NaN	
189362	NaN	France	en:yogurts	NaN	NaN	NaN	NaN	0.180	NaN	NaN	NaN	NaN	NaN	
189364	NaN	France	en:yogurts	NaN	NaN	NaN	NaN	0.180	NaN	NaN	NaN	NaN	NaN	
189417	NaN	France	en:fruit-yogurts	NaN	NaN	NaN	NaN	0.158	NaN	NaN	NaN	NaN	NaN	

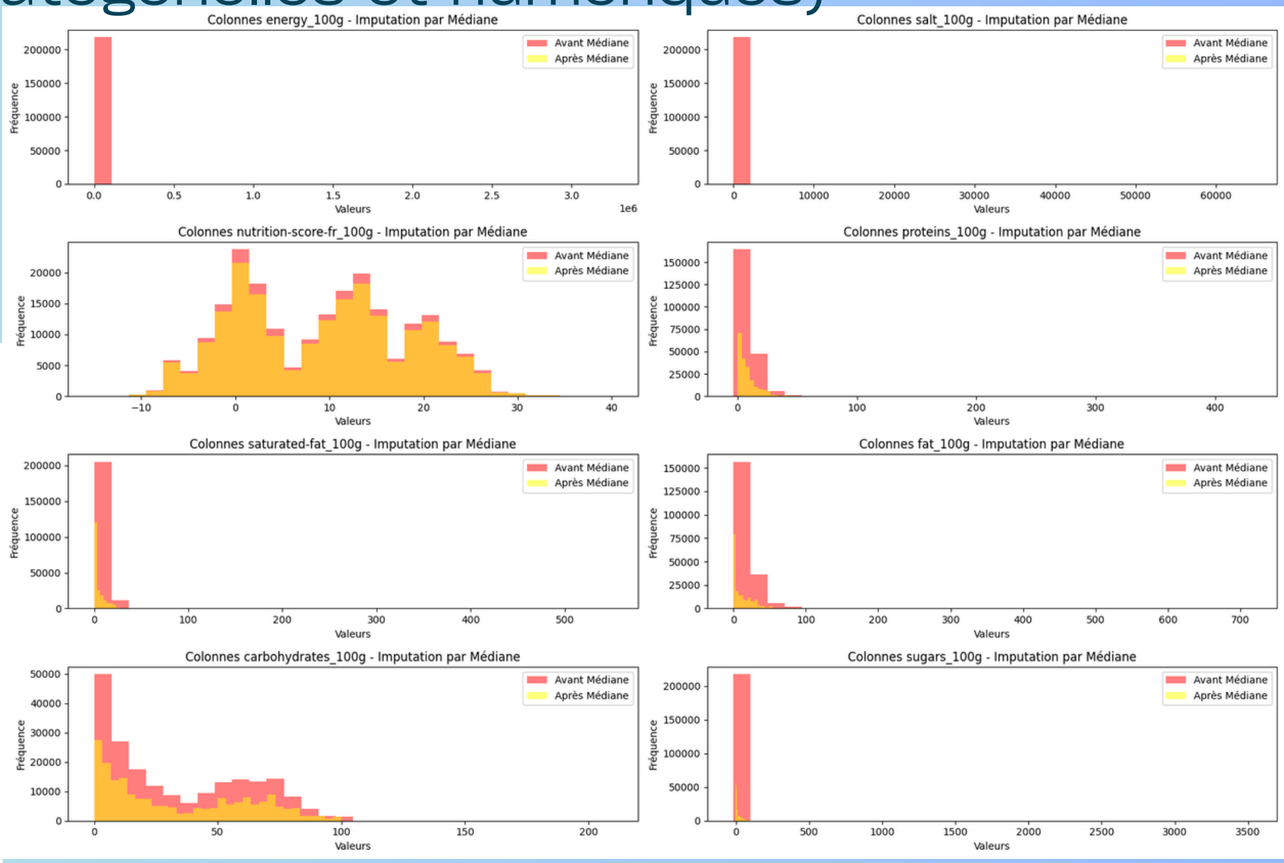
3.3 Traitement Des valeurs aberrantes et manquantes

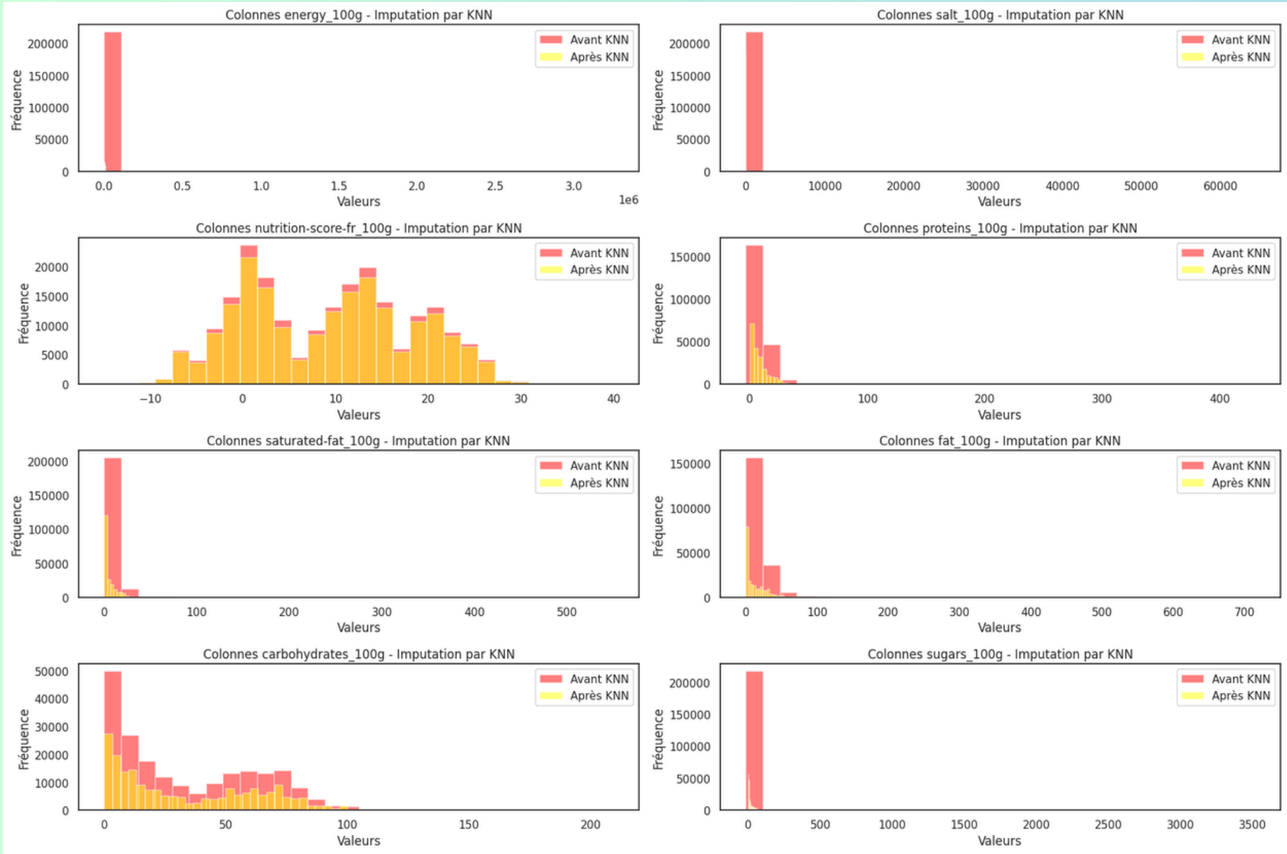
- Valeurs aberrantes

  - salt_100g <= 100
  - nutrition-score-fr_100g <= 100
  - carbohydrates_100g <= 100
  - sugars_100g <= 100
  - energy_100g <= 3800
  - proteins_100g <= 100
  - saturated-fat_100g <= 100
  - fat_100g <= 100

Imputation des valeurs manquantes

- Stratégie d'imputation (valeurs catégorielles et numériques)
- Impact d'imputation





4. Test de normalité

- Kolmogorov-Smirnov

```
Statistique de test Kolmogorov-Smirnov: 0.10516071177929126
p-value: 0.0
Les données ne suivent pas une distribution normale (rejetez H0)
```

- Anova

	Source	SS	DF	MS	F	p-unc	np2
0	nutrition_grade_fr	3.737632e+05	4	93440.801984	1509.740746	0.0	0.029239
1	Within	1.240946e+07	200502	61.891952	NaN	NaN	NaN

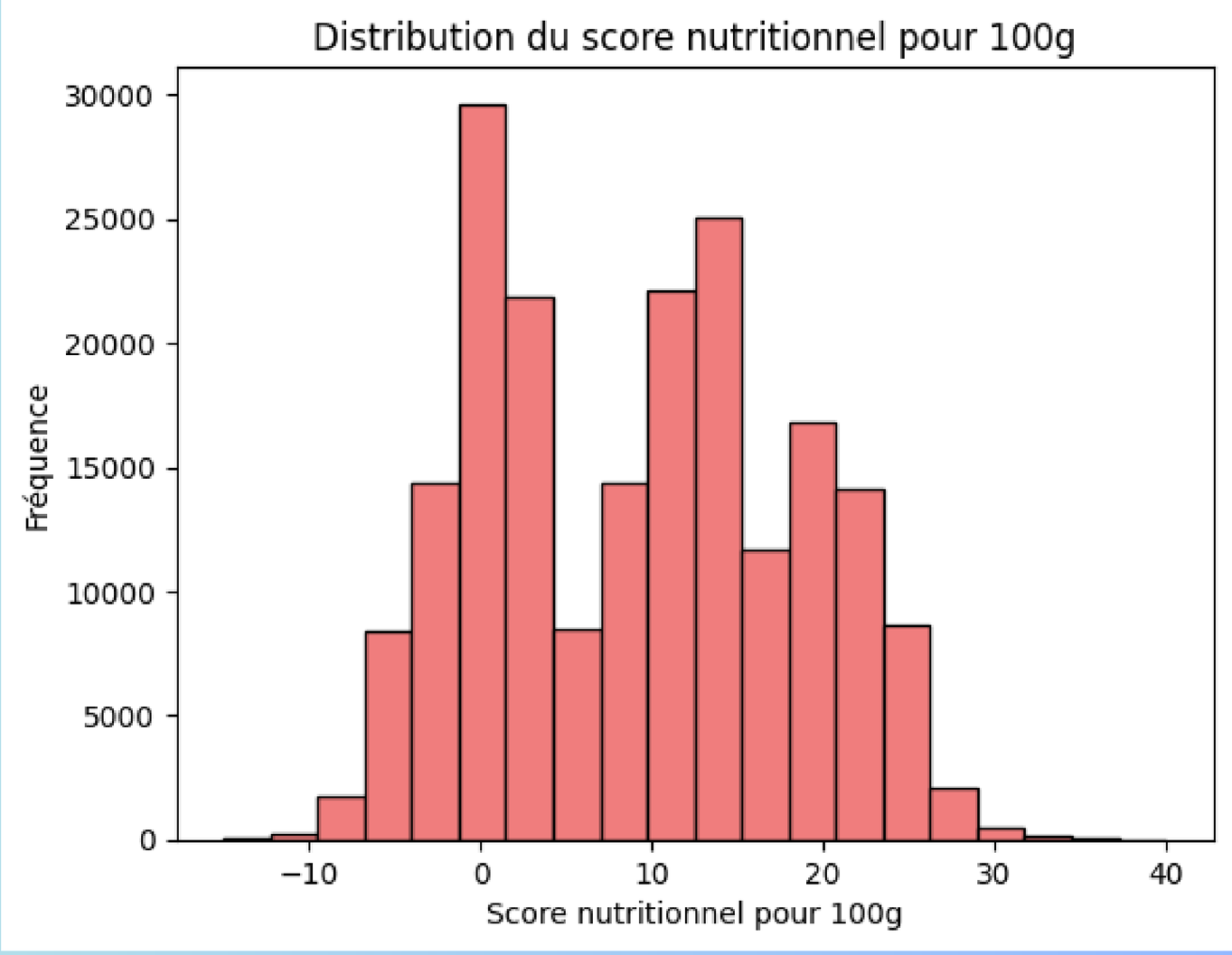
- Kruskal

	Source	ddof1	H	p-unc
Kruskal	nutrition_grade_fr	4	7726.501879	0.0

5. Analyse Univariée et multivariée

5.1 Le scores nutritionnel

- la plus part des scors se setue entre 0 et 15
- le scors max est 30 et min est -10

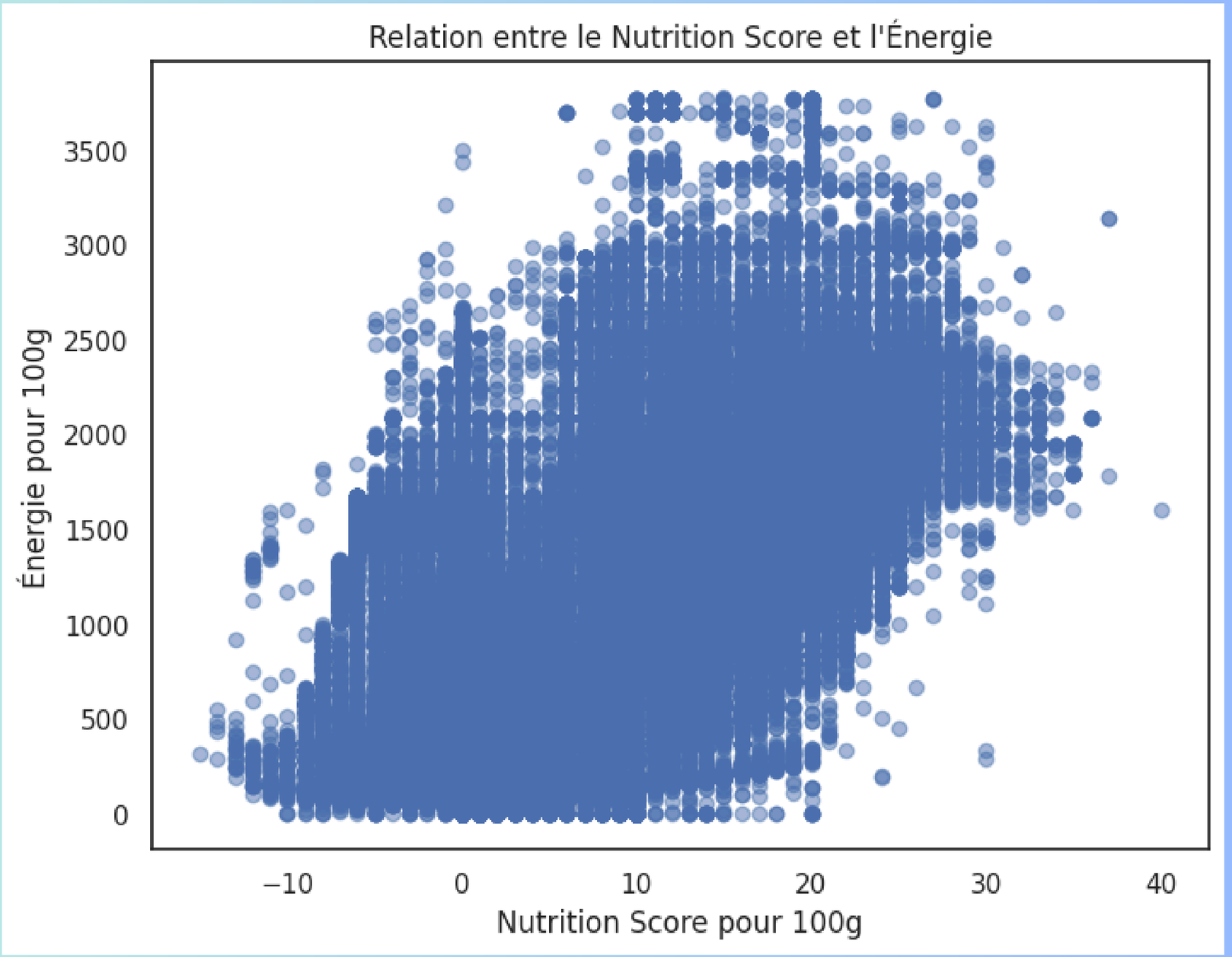


5.2 Le score nutritionnel et l'Energie

- le score nutritionnel augmente, l'énergie pour 100g tend également à augmenter.

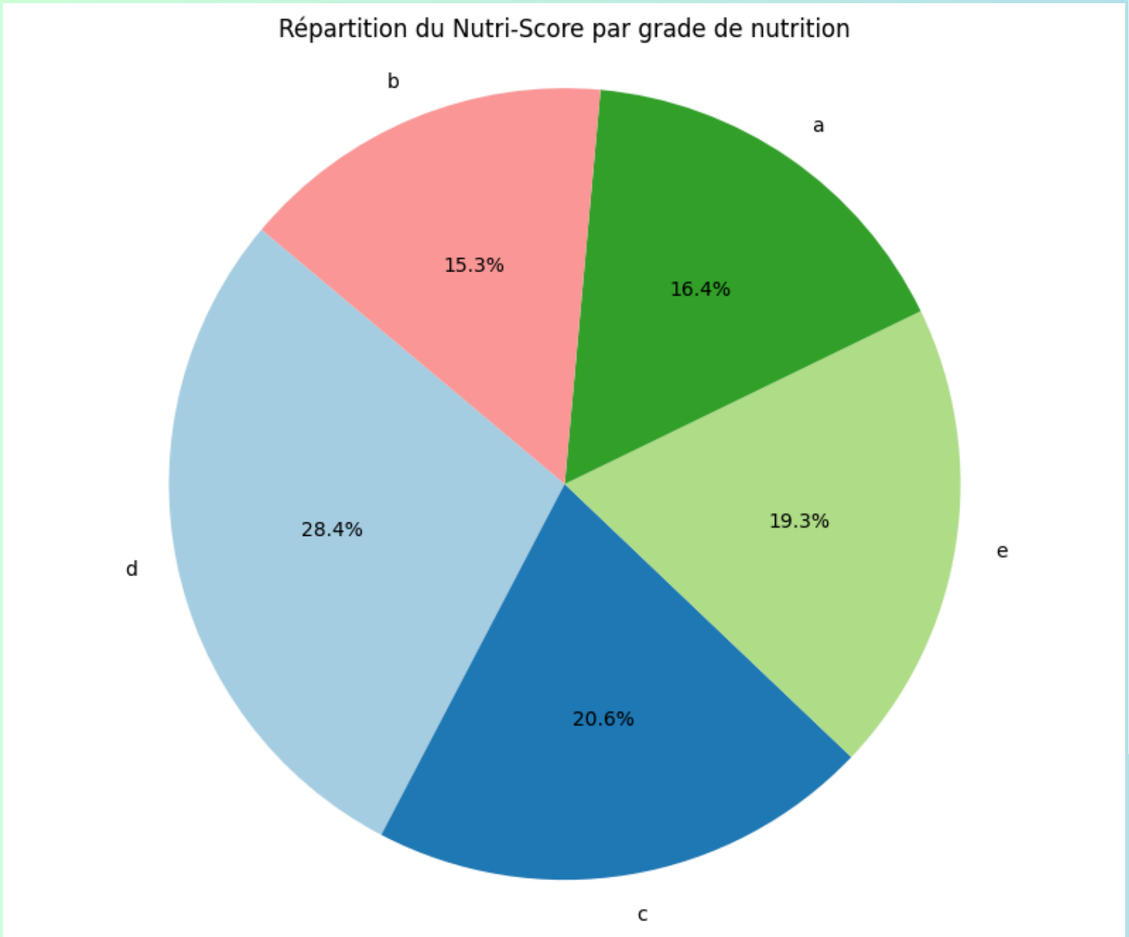
-Cela peut indiquer une corrélation positive entre le score nutritionnel et l'énergie

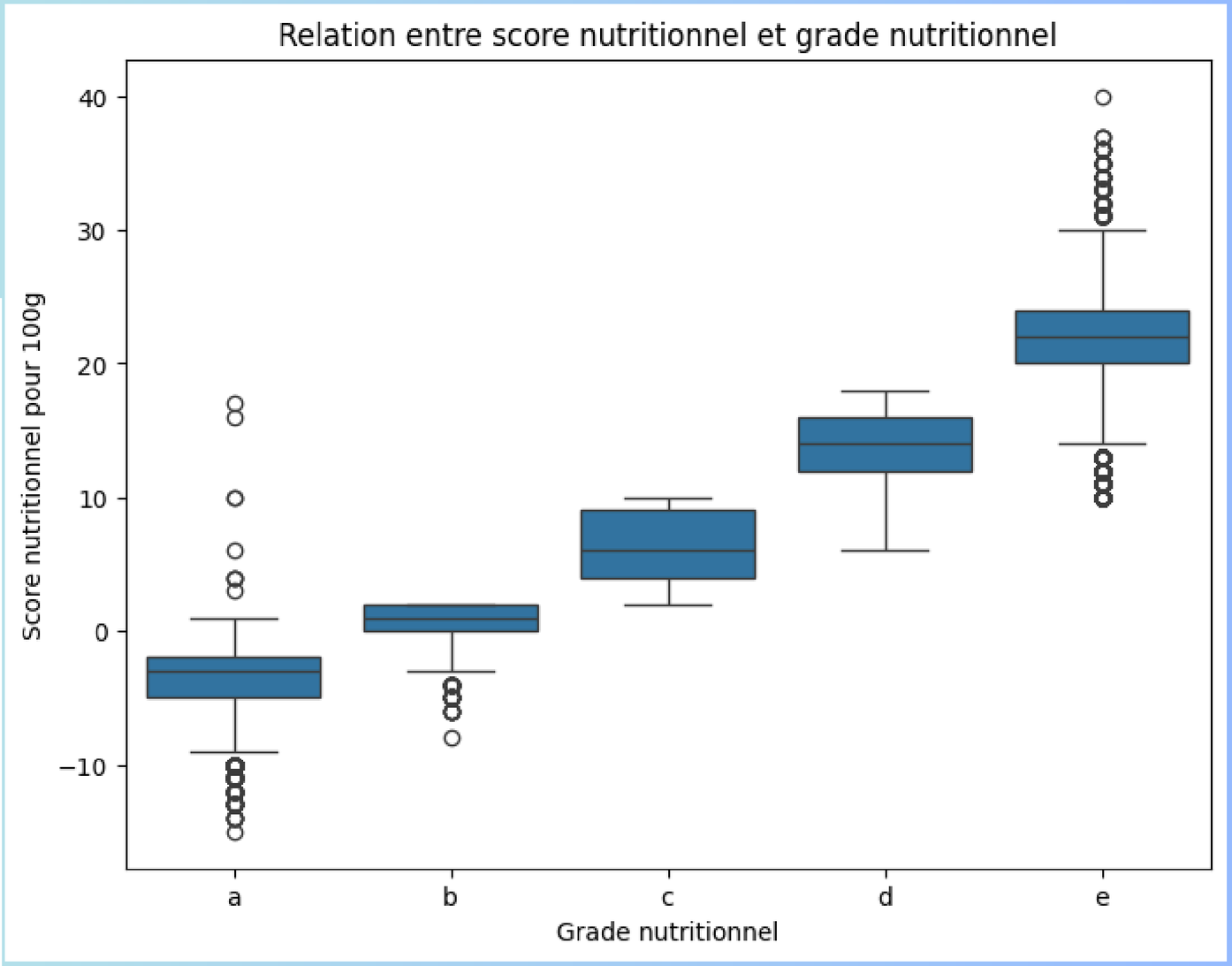
- Les données sont très dispersées, la dispersion est particulièrement notable pour les scores nutritionnels situés entre 0 et 20



5.3 Relation entre score nutritionnel et grade

- Tendance Générale
- Dispersion des Données

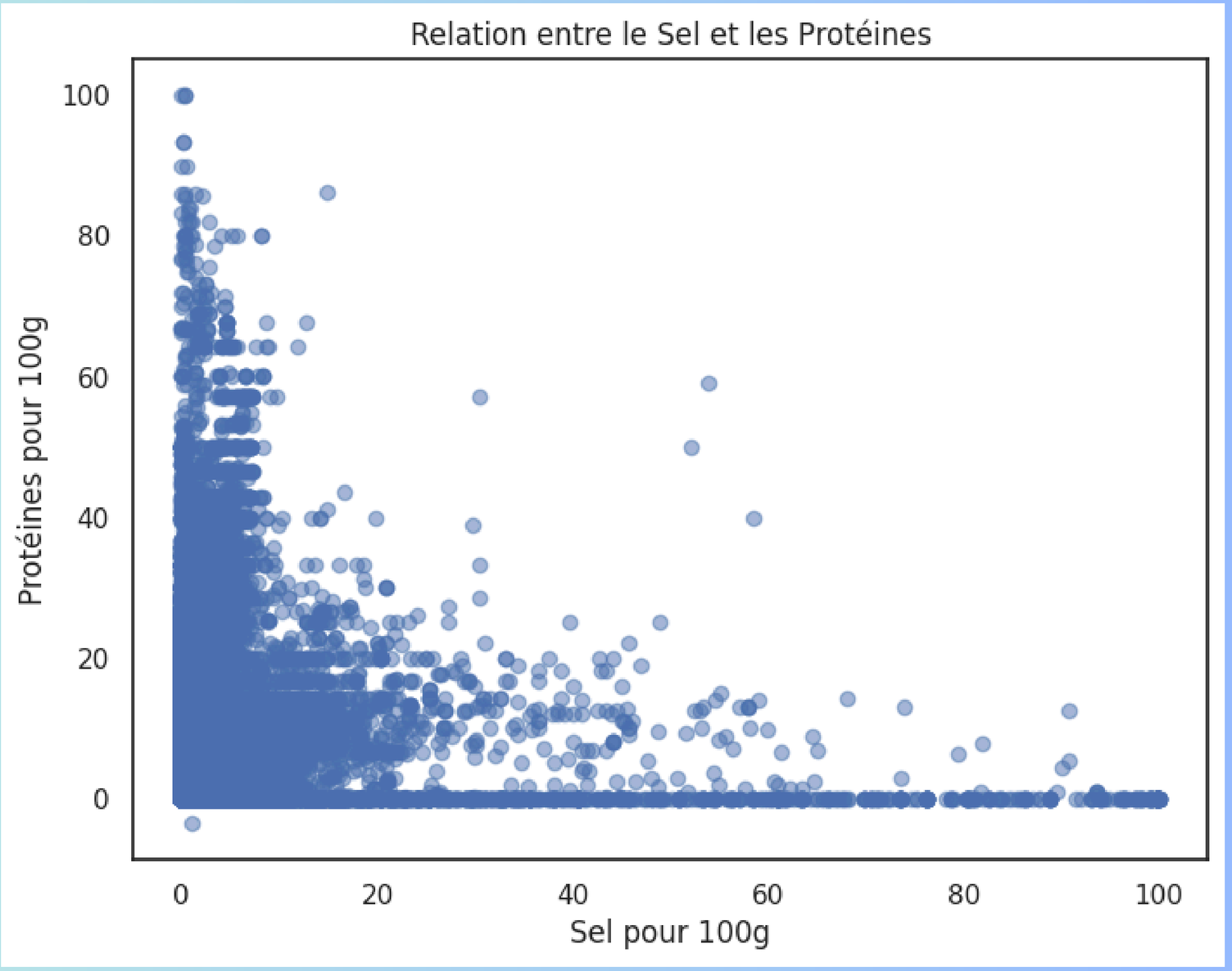




5.4 Relation protéine et le Sel

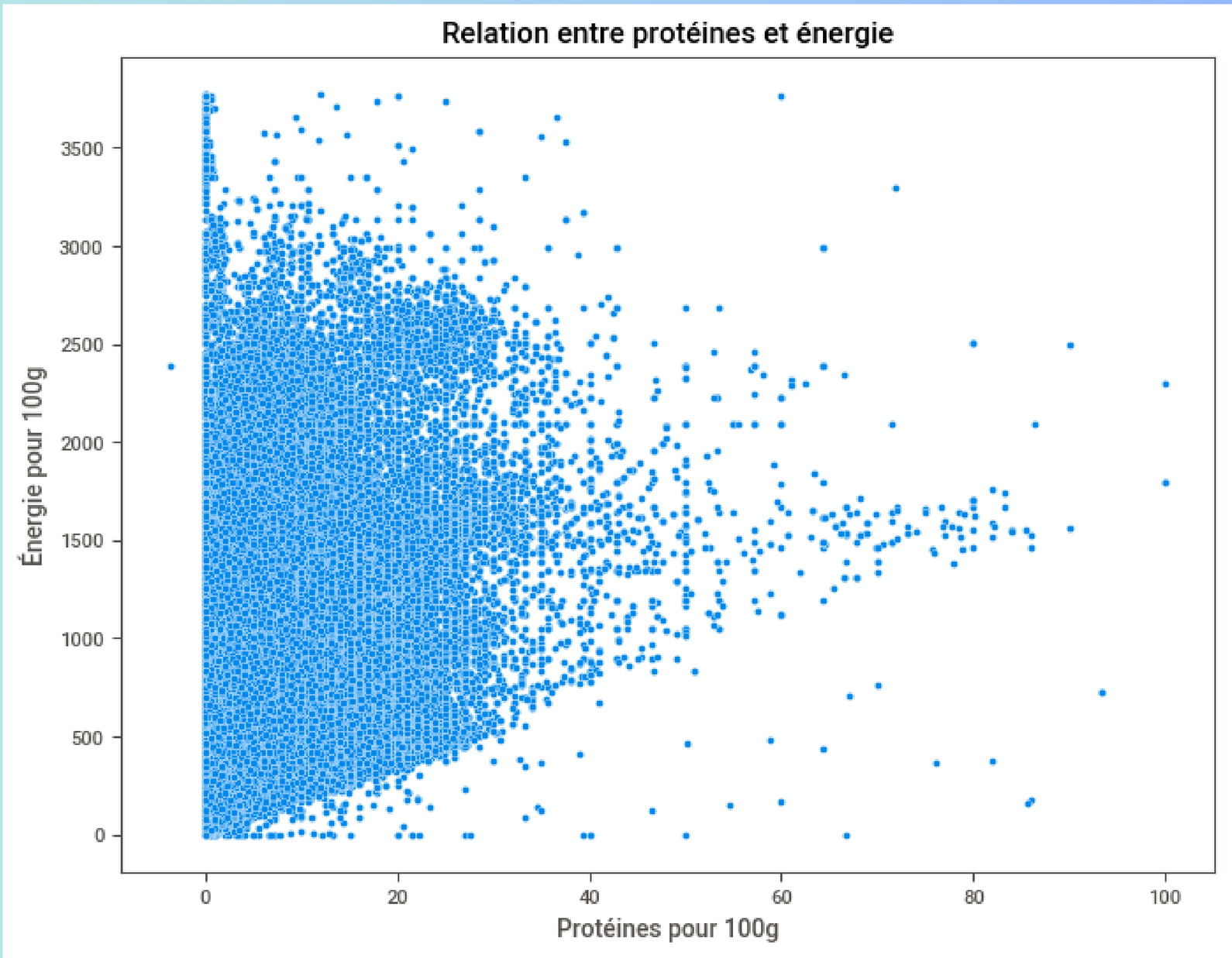
- La majorité des points se concentre dans la région où la teneur en sel est faible (inférieure à 10g pour 100g) et la teneur en protéines varie plus largement (jusqu'à environ 40g pour 100g).

- Cela suggère que la plupart des aliments ont une faible teneur en sel et une large gamme de teneur en protéines



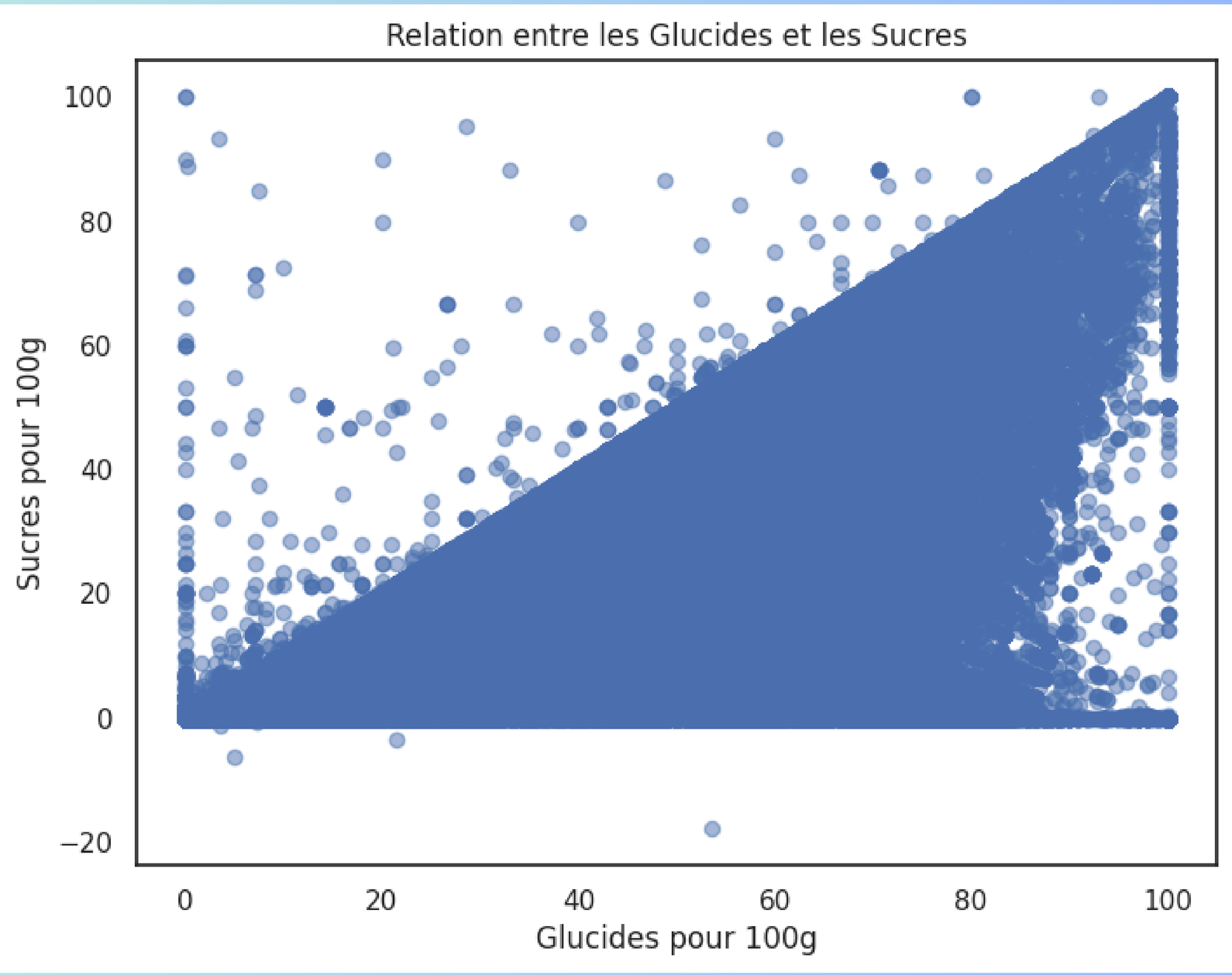
5.5 Relation entre protéine et énergie

- Tendance générale

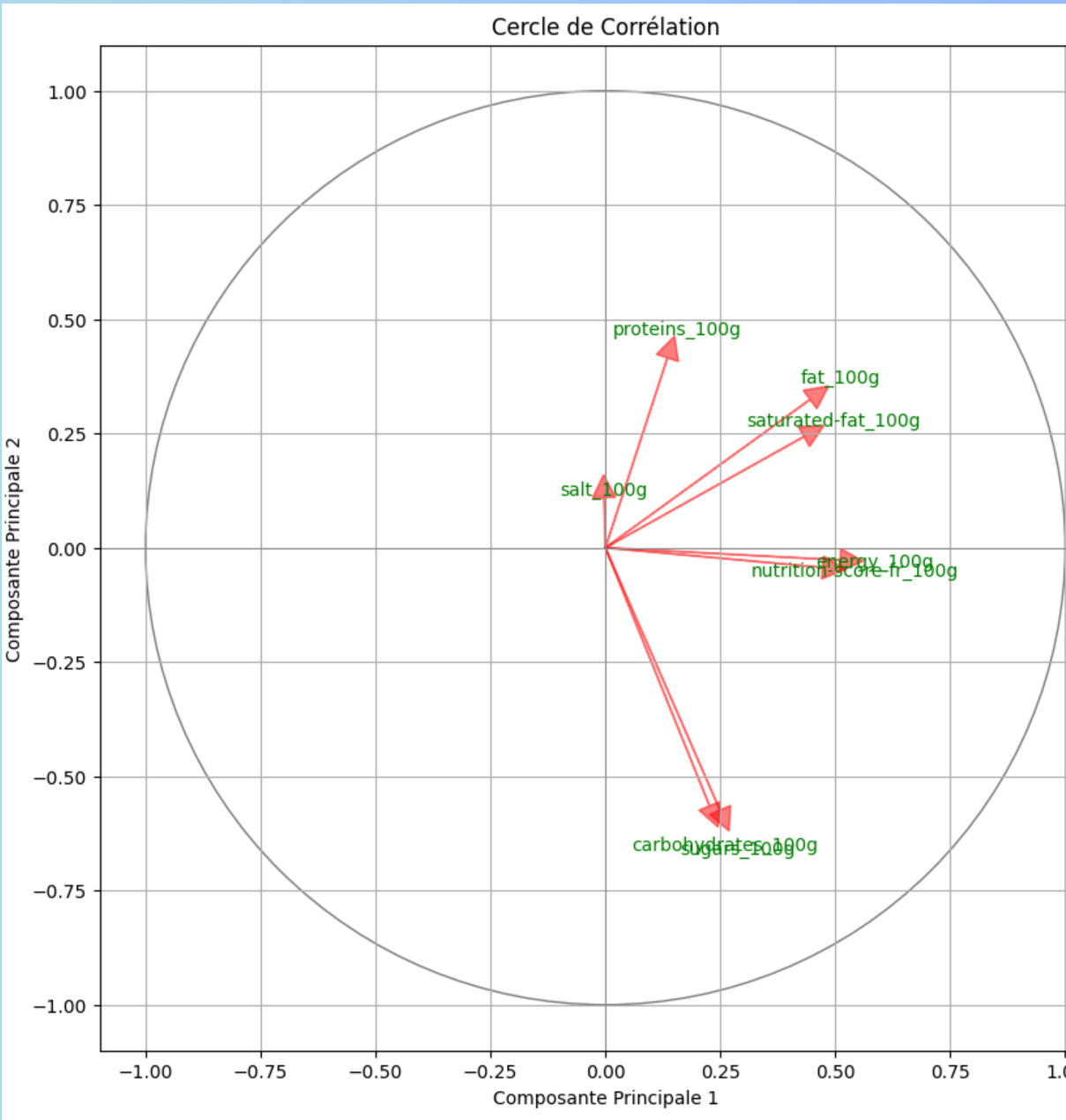


5.6 Relation entre les glucide et sucres

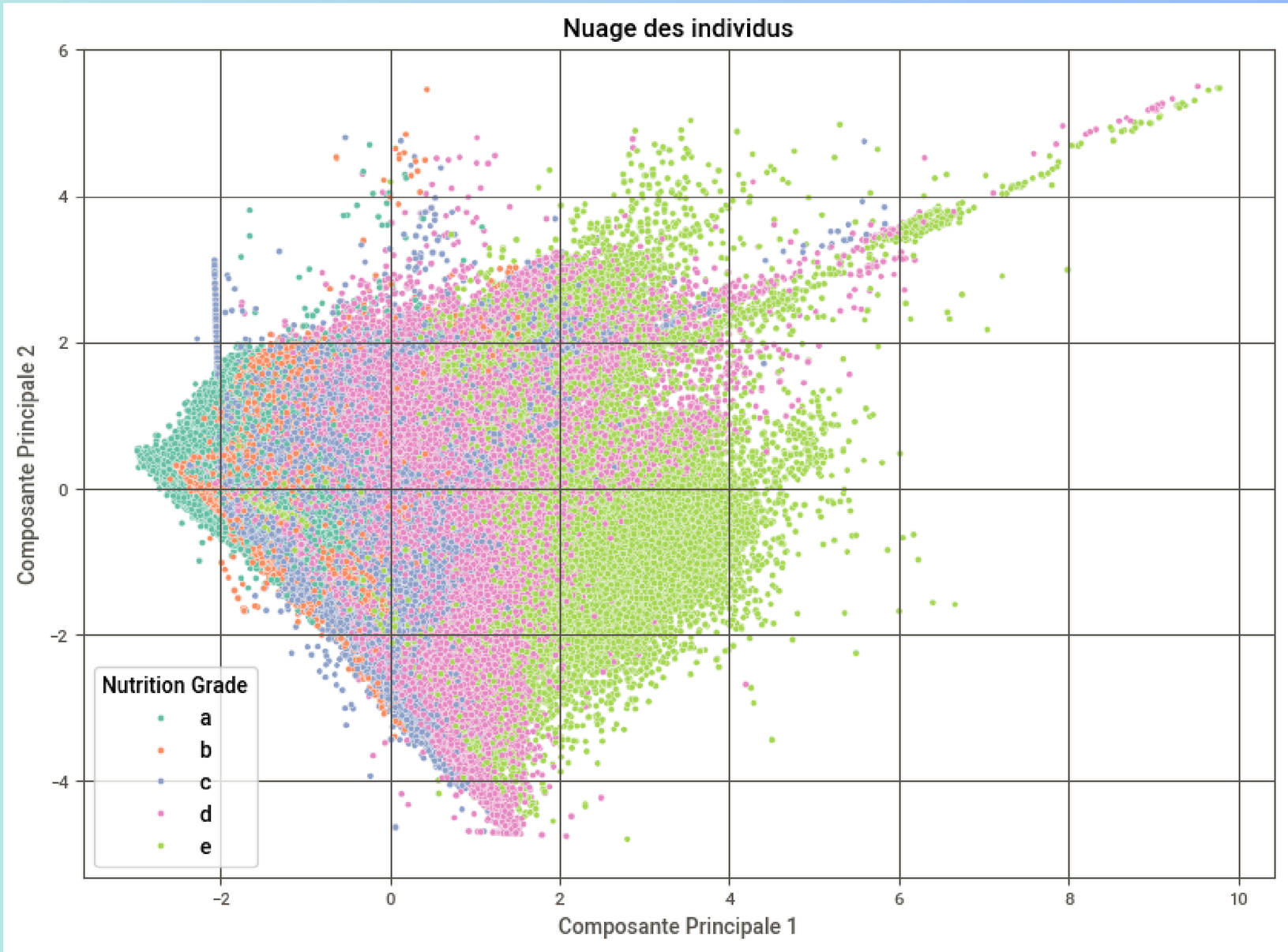
- Tendance générale
- Dispersion des données



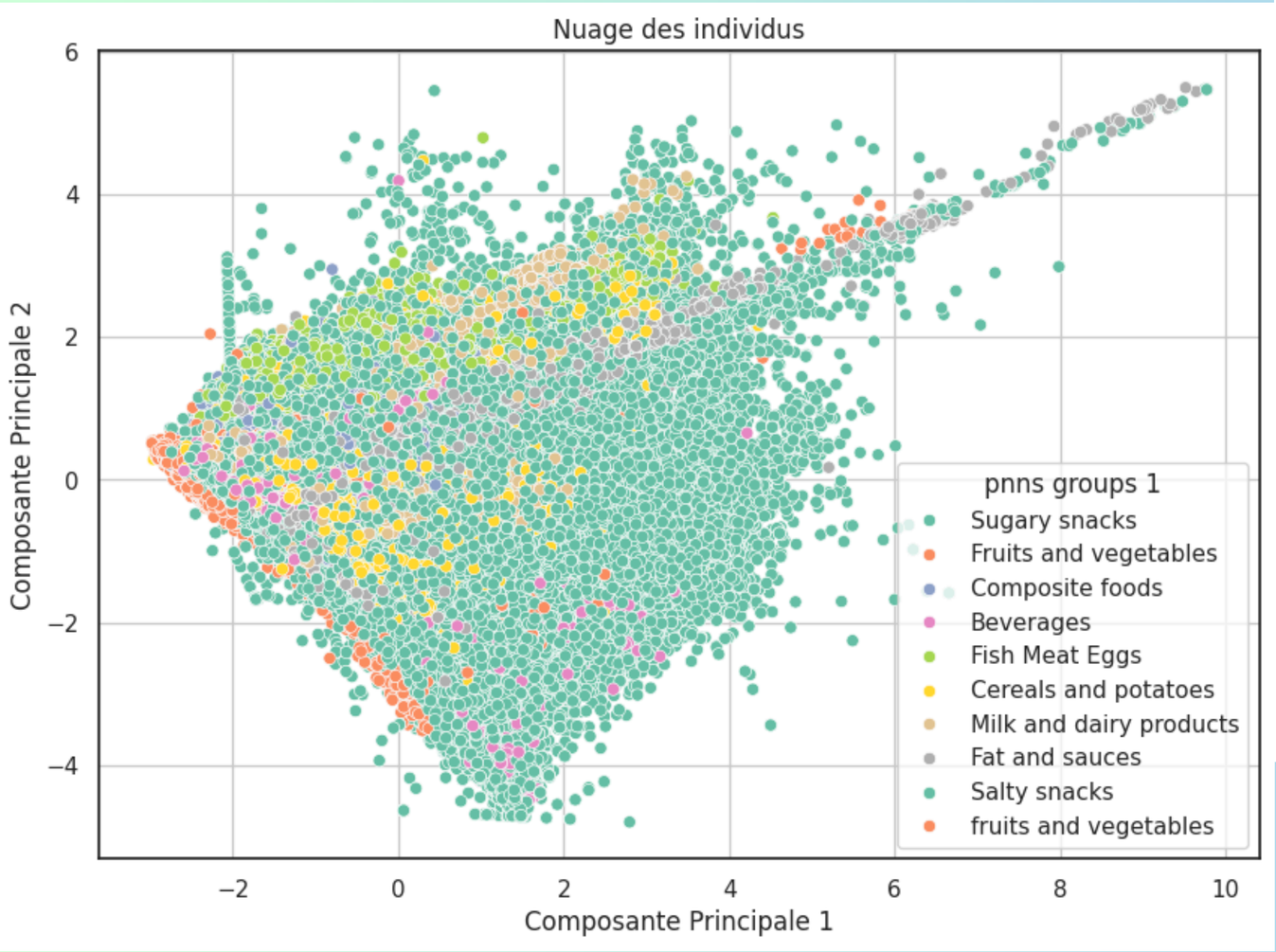
5.7 Cercle de corrélation

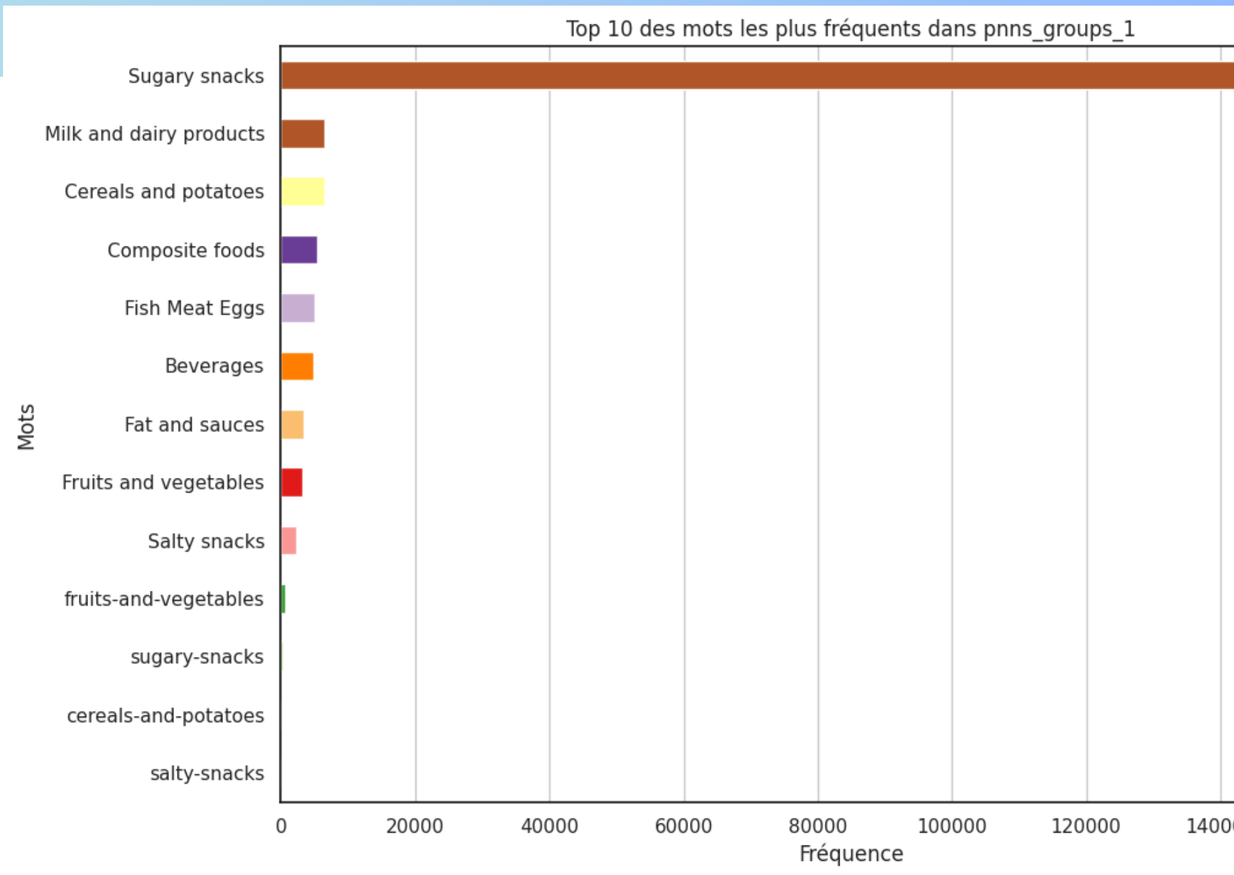


5.8 Nuage des indivudus du grade nutritionnel

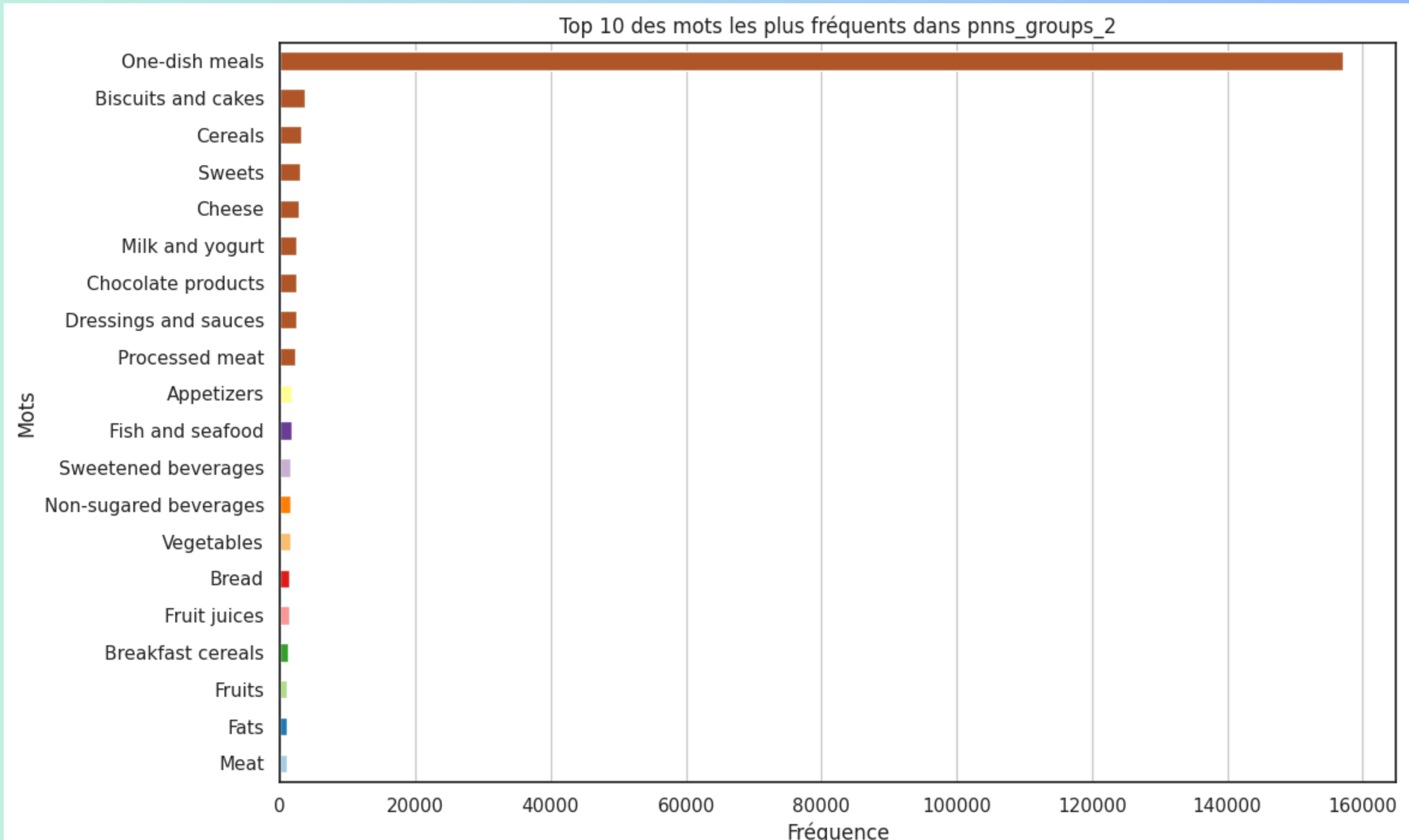


5.10 Top 10 des mots plus fréquents groupe PNNS-1





5.10 Top 10 des mots plus fréquents groupe PNNS-2



6. Respect du RGPD

Conclusion