Préparation des données pour un organisme de santé publique

Mission et objectifs

Mission : Explorer la faisabilité d'un modèle d'autocomplétion des données manquantes dans un jeu de données nutritionnelles.

Objectifs :

- Nettoyer et explorer la base de données nutritionnelle.
- Identifier les erreurs de saisie et les valeurs manquantes.
- Imputer les données manquantes.
- Évaluer la faisabilité de l'autocomplétion pour améliorer la qualité des données.

Plan

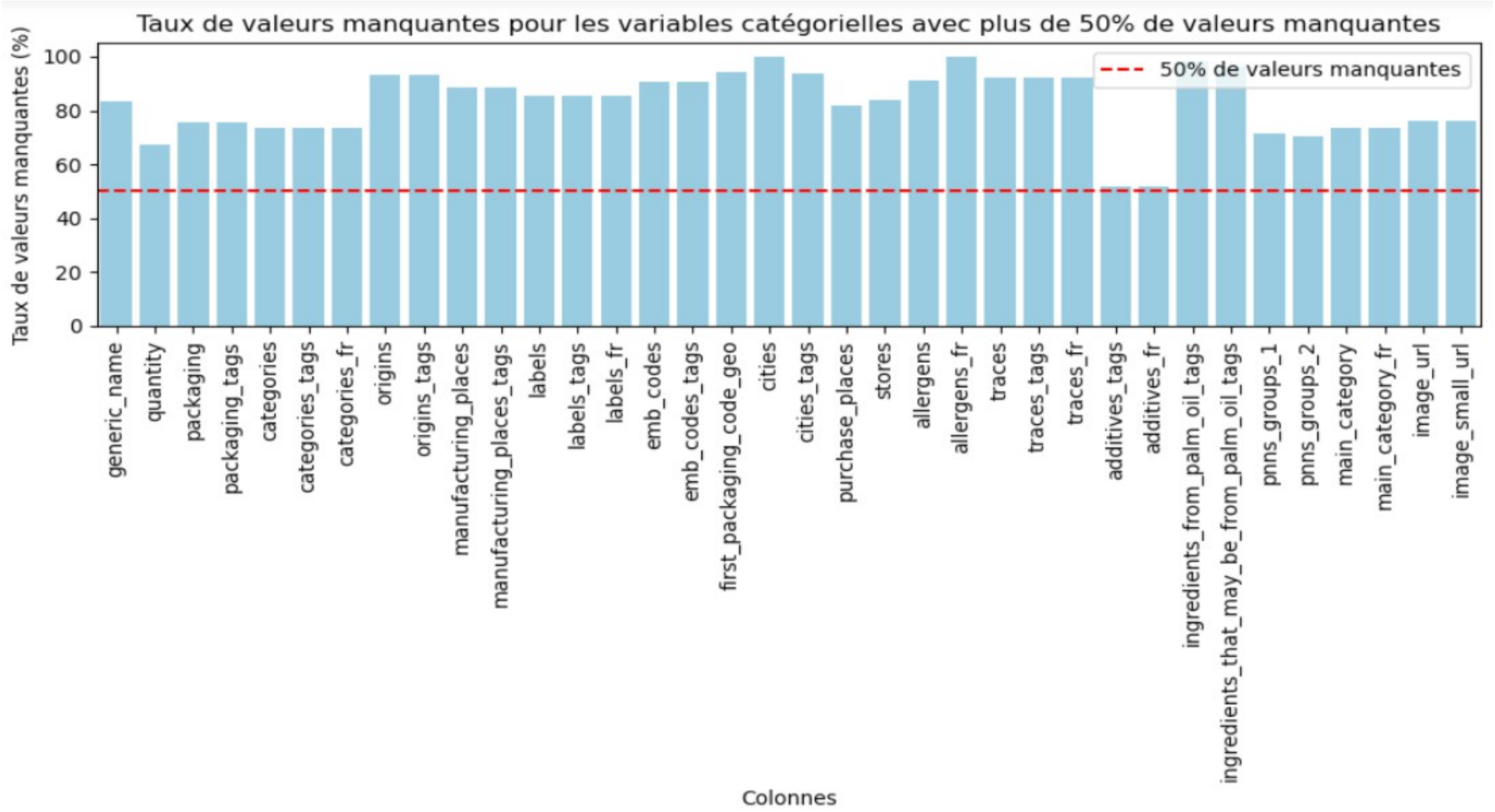
- Description du fichier
- Sélection de la variable cible
- Sélections des variables explicatives
- Règles RGPD
- Identification et traitement des valeurs incohérentes
- Traitement des données manquantes
- Analyse univariée des données
- Analyse bivariée des données
- Analyse multivariée des données

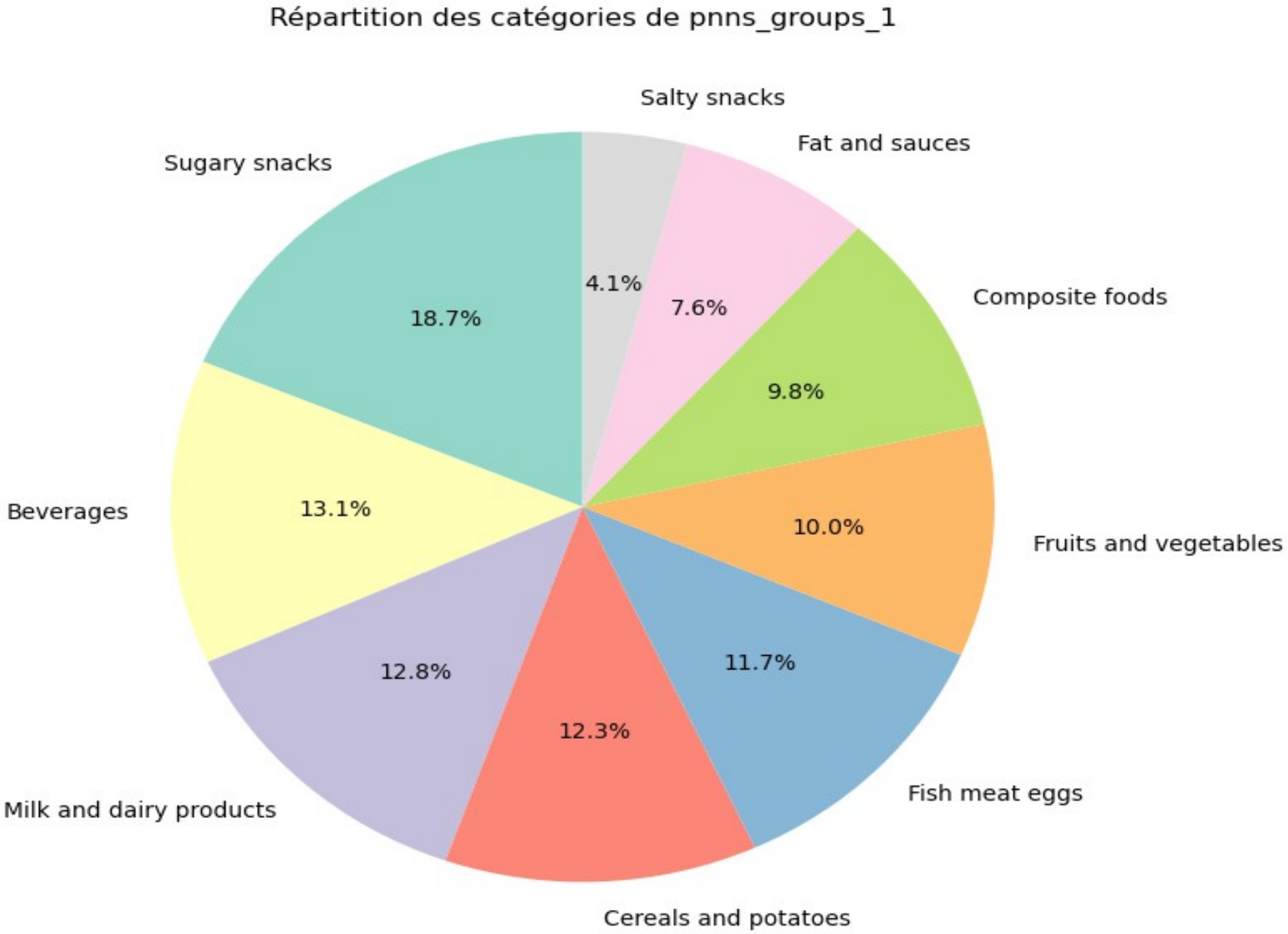
Description du fichier

320 772 lignes et 162 colonnes

Colonnes	Descriptions
Identifiants & Métadonnées	Informations sur le produit (code, créateur, dates de création, liens d'images).
Produit & Nom	Nom du produit et nom générique.
Quantité & Emballage	Détails sur le poids, le volume, et l'emballage.
Marques & Catégories	Marque et catégories du produit.
Origine & Fabrication	Origine et lieux de fabrication.
Labels & Tags	Labels, codes d'emballage, et tags géographiques.
Ingrédients & Allergènes	Liste des ingrédients, allergènes, et traces.
Additifs & Nutrition	Additifs, huile de palme, et notation nutritionnelle.
Informations Nutritionnelles	Composants nutritionnels pour 100g (calories, graisses, protéines, etc.).
Vitamines et Minéraux	Liste des vitamines et minéraux pour 100g.
Impact Environnemental	Empreinte carbone et dureté de l'eau.
Indicateurs de Consommation	Scores nutritionnels et indice glycémique.

Sélection de la variable cible




Répartition des catégories de pnns_groups_1
Salty snacks
Fat and sauces
Sugary snacks
Composite foods
4.1%
7.6%
18.7%
9.8%
Beverages
13.1%
10.0%
Fruits and vegetables
12.8%
11.7%
12.3%
Fish meat eggs
Milk and dairy products
Cereals and potatoes

Sélection des variables explicatives

Variable	Description	Catégories associées +	Catégories associées -
energy_100g	Teneur en énergie pour 100g d'un produit.	+ Fat and sauces + Sugary snacks + Cereals and potatoes	- Fruits and vegetables - Milk and dairy products
fat_100g	Quantité de graisses pour 100g d'un produit.	+ Fat and sauces + Salty snacks + Fish, Meat & Eggs	- Fruits and vegetables - Sugary snacks
carbohydrates_100g	Quantité de glucides pour 100g d'un produit.	+ Sugary snacks + Cereals and potatoes + Fruits and vegetables	- Fat and sauces - Fish, Meat & Eggs
sugars_100g	Quantité de sucres pour 100g d'un produit.	+ Sugary snacks + Beverages + Fruits and vegetables	- Fat and sauces - Fish, Meat & Eggs
proteins_100g	Quantité de protéines pour 100g d'un produit.	+ Fish, Meat & Eggs + Milk and dairy products + Cereals and potatoes	- Sugary snacks - Fat and sauces
salt_100g	Quantité de sel pour 100g d'un produit.	+ Salty snacks + Fat and sauces + Fish, Meat & Eggs + Composite foods	- Fruits and vegetables - Sugary snacks

Optimisation des données : Suppression des doublons et automatisation des processus

Objectif : Assurer la qualité des données en éliminant les doublons et en automatisant les tâches de nettoyage.

- Suppression des doublons

But : Éviter les redondances dans les données, garantir des résultats d’analyse plus précis.

- Automatisation des tâches

But : Gagner du temps, améliorer l'efficacité et réduire les erreurs humaines.

Respect des normes RGPD dans le projet d'analyse nutritionnelle

Principes clés du RGPD :

- Licéité, loyauté, transparence
- Limitation des finalités
- Minimisation des données
- Exactitude des données
- Limitation de la conservation

Absence de lien avec le RGPD dans ce projet :

- Utilisation exclusive de données sur les produits alimentaires (pas de données personnelles).
- Aucune donnée permettant d’identifier des individus spécifiques.
- Sources publiques et objectifs non liés à des informations personnelles.

Conclusion : Le projet respecte les normes éthiques puisqu'il ne traite que des données non personnelles.

Identification des valeurs incohérentes

Variable	Plage cohérente	Action	Valeurs supprimées
energy_100g	0 à 3900 kJ	Suppression des valeurs > 3900 kJ et <0	23
fat_100g	0 à 100 g	Suppression des valeurs > 100 kJ et <0	3
carbohydrates_100g	0 à 100 g	Suppression des valeurs > 100 kJ et <0	5
proteins_100g	0 à 80 g	Suppression des valeurs > 80 kJ et <0	1
salt_100g	0 à 100 g	Suppression des valeurs > 100 kJ et <0	3

Identification des valeurs incohérentes de sugars_100g

Problème détecté :

- 3 lignes avec sugars_100g < 0 ou > 100.
- Relation attendue : sugars_100g ≤ carbohydrates_100g.

Méthodologie :

- Acceptables : Écart < 20% → valeurs conservées.
- Inacceptables :
  - ✔ Écarts ≤ 0.5 g : 17 valeurs de sugars_100g supprimées.
  - ✔ Écarts > 0.5 g : Suppression de 12 produits (~0.025%).

Résultats :

- 12 produits supprimés, 20 incohérences supprimées.
- Les suppressions de produits affectent 0.025% des produits du dataset, garantissant ainsi la fiabilité des données.

Nettoyage des incohérences : Mission accomplie

- Toutes les valeurs incohérentes identifiées ont été supprimées.

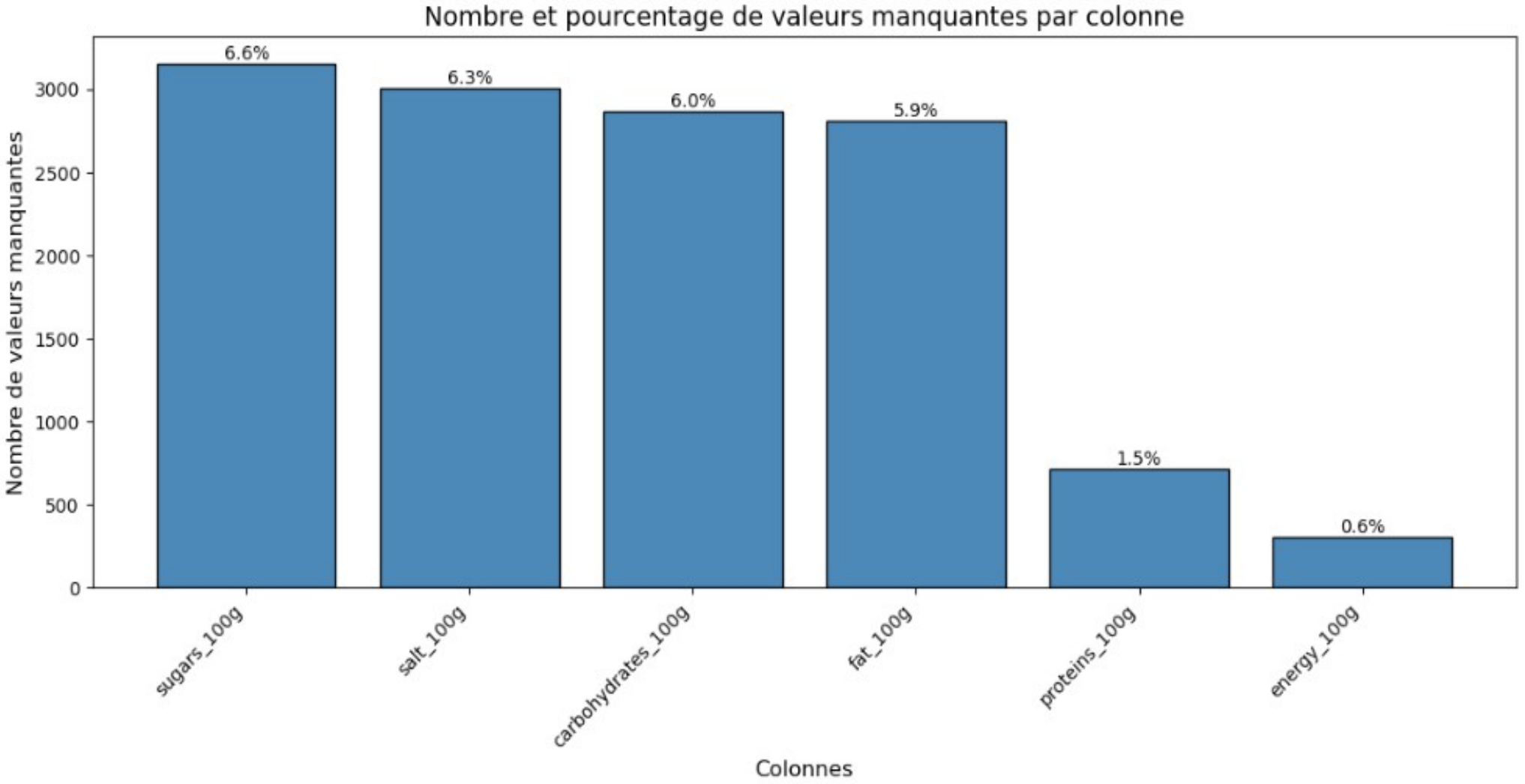
- Processus automatisé :
  - ✔ Analyse des incohérences basée sur des règles (marges d'erreur, seuils définis).
  - ✔ Suppression des valeurs aberrantes.

- Impact final :

Dataset entièrement nettoyé et prêt pour l'analyse.

Traitement des données manquantes

Visualisation des données manquantes



Identification du type des valeurs manquantes

Types de valeurs manquantes :

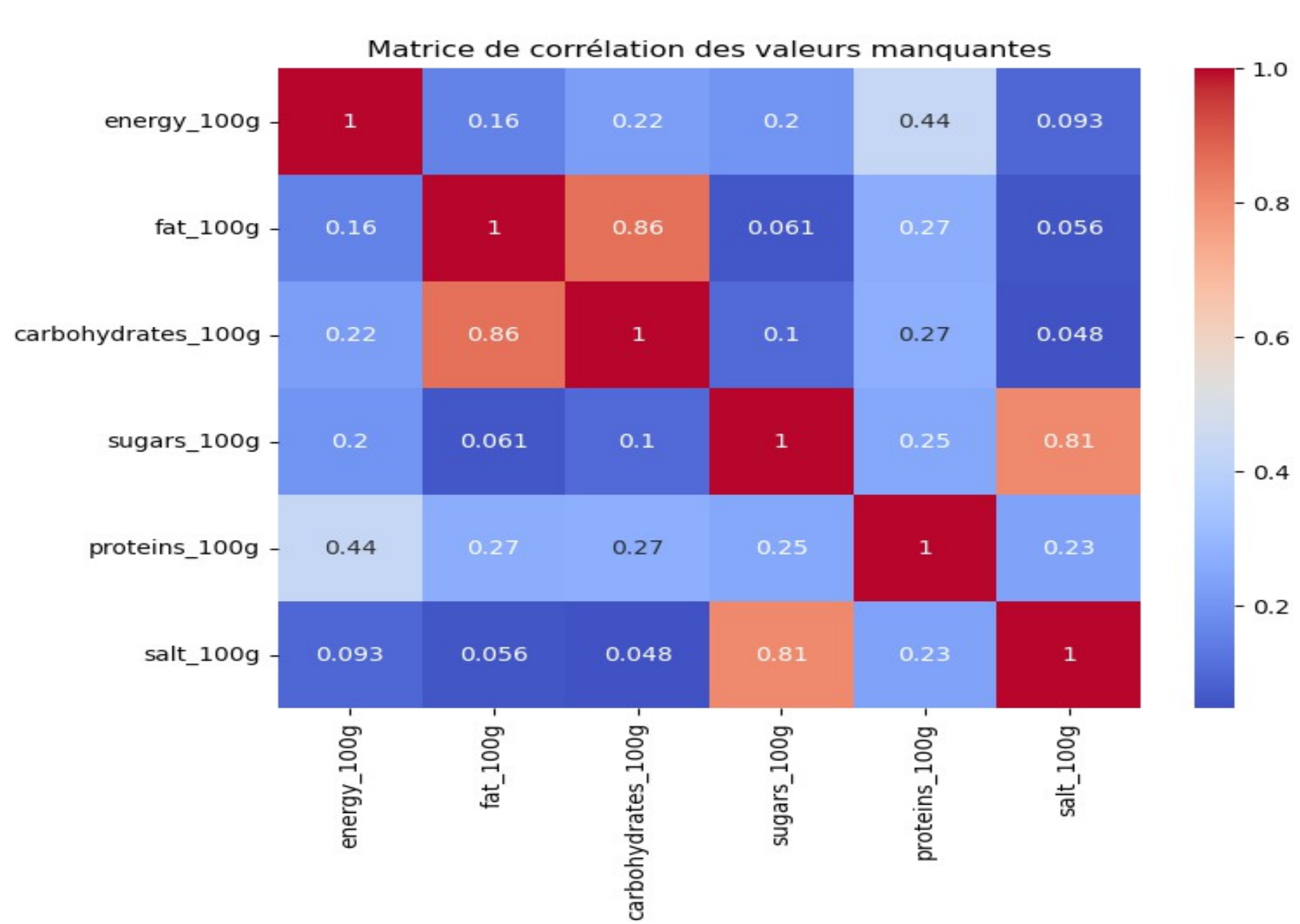
- MCAR : Complètement aléatoires
- MAR : Dépendent d'autres variables observées
- MNAR : Liées à la valeur manquante elle-même

Test de Little :

- Résultat : p-value = 0.0
- Interprétation : Les données manquantes ne sont pas MCAR

Implications :

- Les données sont probablement MAR ou MNAR
- Nécessité d'utiliser des méthodes d'imputation avancées
- Imputation simple (moyenne, médiane) non recommandée


Matrice de corrélation des valeurs manquantes
energy_100g
fat_100g
carbohydrates_100g
sugars_100g
proteins_100g
salt_100g
1 0.16 0.22 0.2 0.44 0.093
0.16 1 0.86 0.061 0.27 0.056
0.22 0.86 1 0.1 0.27 0.048
0.2 0.061 0.1 1 0.25 0.81
0.44 0.27 0.27 0.25 1 0.23
0.093 0.056 0.048 0.81 0.23 1
1.0
0.8
0.6
0.4
0.2

Stratégie d'imputation des valeurs manquantes

- Objectif : Restaurer des données cohérentes et exploitables.

- Méthodes principales utilisées :

  - ✔ Proportions pour certains nutriments.
  - ✔ Médiane
  - ✔ KNN pour les cas plus complexes.

- Vérifications systématiques après chaque imputation.

Imputations de sugars_100g

Contexte :

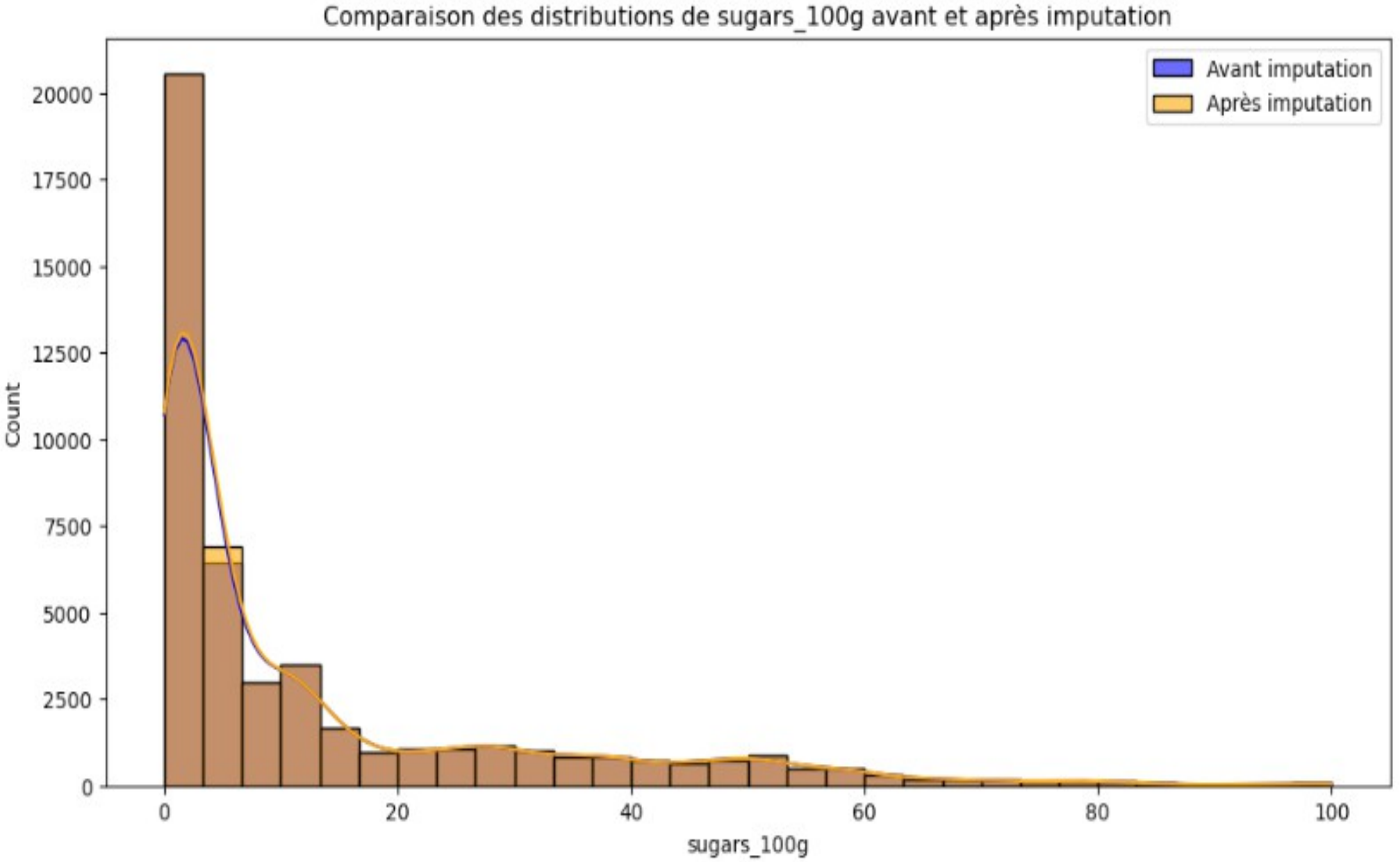
- Sucres ⊆ Glucides
- Relation utilisée pour estimer les valeurs manquantes

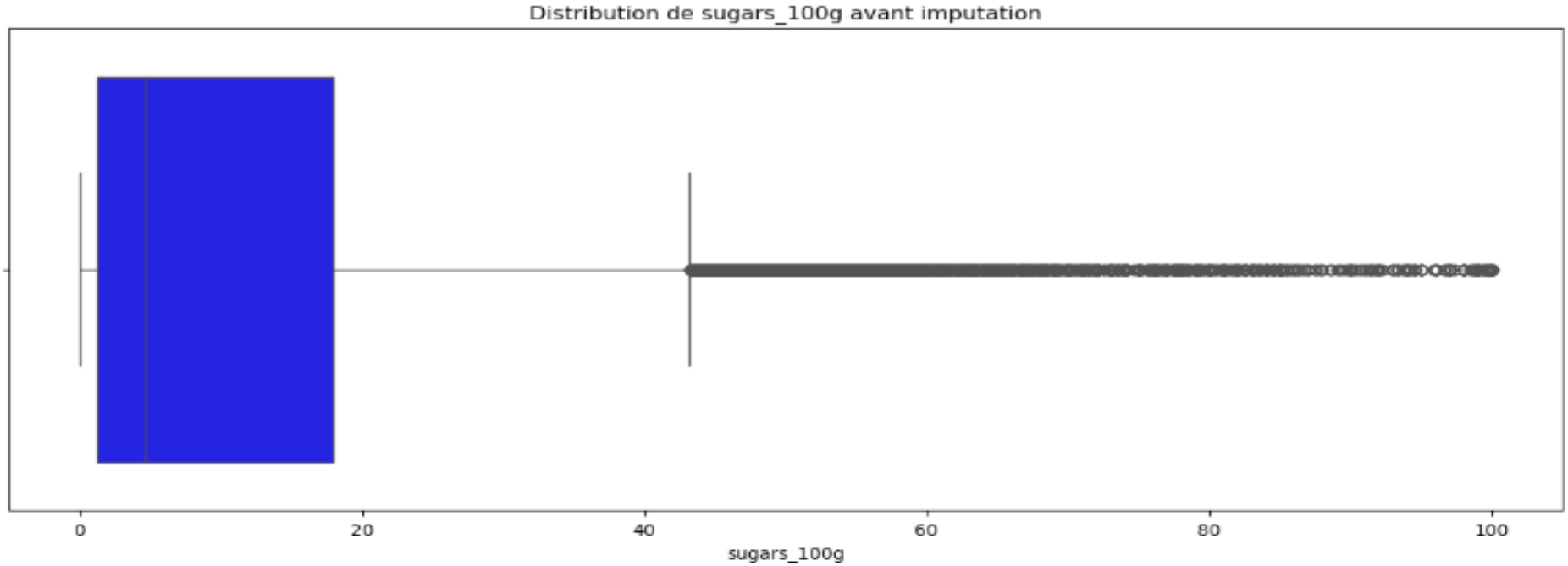
Méthode :

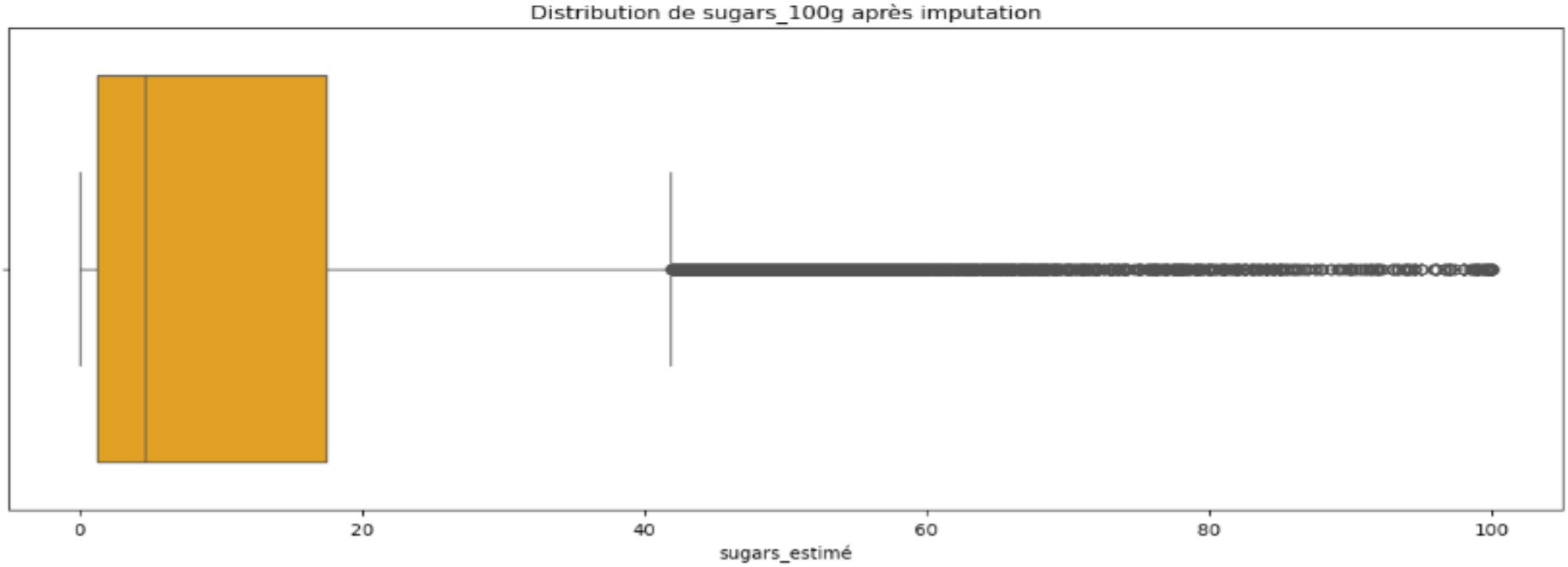
- Proportion (sugars_100g / carbohydrates_100g)
- Imputation :
  - ✔ Si glucides connus →

    Sucre estimé = Médiane des proportions × Glucides
  - ✔ Si glucides inconnus (473 valeurs)→

    Sucre estimé = Médiane globale des sucres


Comparaison des distributions de sugars_100g avant et après imputation
Avant imputation
Après imputation
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
0
20
40
60
80
100
sugars_100g


Distribution de sugars_100g avant imputation
0
20
40
60
80
100
sugars_100g


Distribution de sugars_100g après imputation
0
20
40
60
80
100
sugars_estimé

Impact de l'imputation sur les corrélations avec sugars_100g

Variables	Corrélation avant imputation	Corrélation après imputation	Différence
Salt_100g	-0,38	-0,37	0,01
energy_100g	0,26	0,27	0,01
proteins_100g	-0,33	-0,31	0,02
fat_100g	-0,01	-0,02	0,01
carbohydrates_100g	0,63	0,66	0,03

Imputation de carbohydrates_100g

Contexte :

- Sucres $\subseteq$ Glucides
- Relation utilisée pour estimer les valeurs manquantes

Méthode :

- Proportion (sugars_100g / carbohydrates_100g)
- Imputation :

Glucides estimés = Sucres /Médiane des proportions

Validation de l'imputation des sucres et glucides

Critères de validation :

- Distribution des données conservée ✓
- Relations entre variables maintenues ✓
- Proportions stables ✓
- Aucune incohérence n'a été introduite ✓

Implications :

- Méthode d'imputation fiable et robuste
- Données complétées sans biais significatif
- Prêtes pour les analyses statistiques suivantes

Imputation des autres variables par KNN

Méthode d'imputation :

- KNN (K-Nearest Neighbors) :
  - ✔ Utilise les valeurs des autres nutriments pour estimer les valeurs manquantes.
  - ✔ Se base sur la proximité entre les observations, permettant de conserver la structure des données.

Pourquoi KNN ?

- Prend en compte les relations entre plusieurs variables.
- Permet une imputation plus contextuelle et précise que d'autres méthodes simples.

Détails supplémentaires :

- K = racine carrée du nombre d'individus.
- Normalisation des données avant l'imputation.

Choix des variables impliquées dans les imputations

Variables imputée	Sucre	Glucides	Energie	Gras	Protéines	Sel
Energie		X		X	X	
Gras		X	X		X	
Protéines			X	X		X
Sel	X		X	X	X	

Relations utilisées pour les imputations :

- X = relation entres les variables
- X = relation entre les données manquantes

Validation des imputations

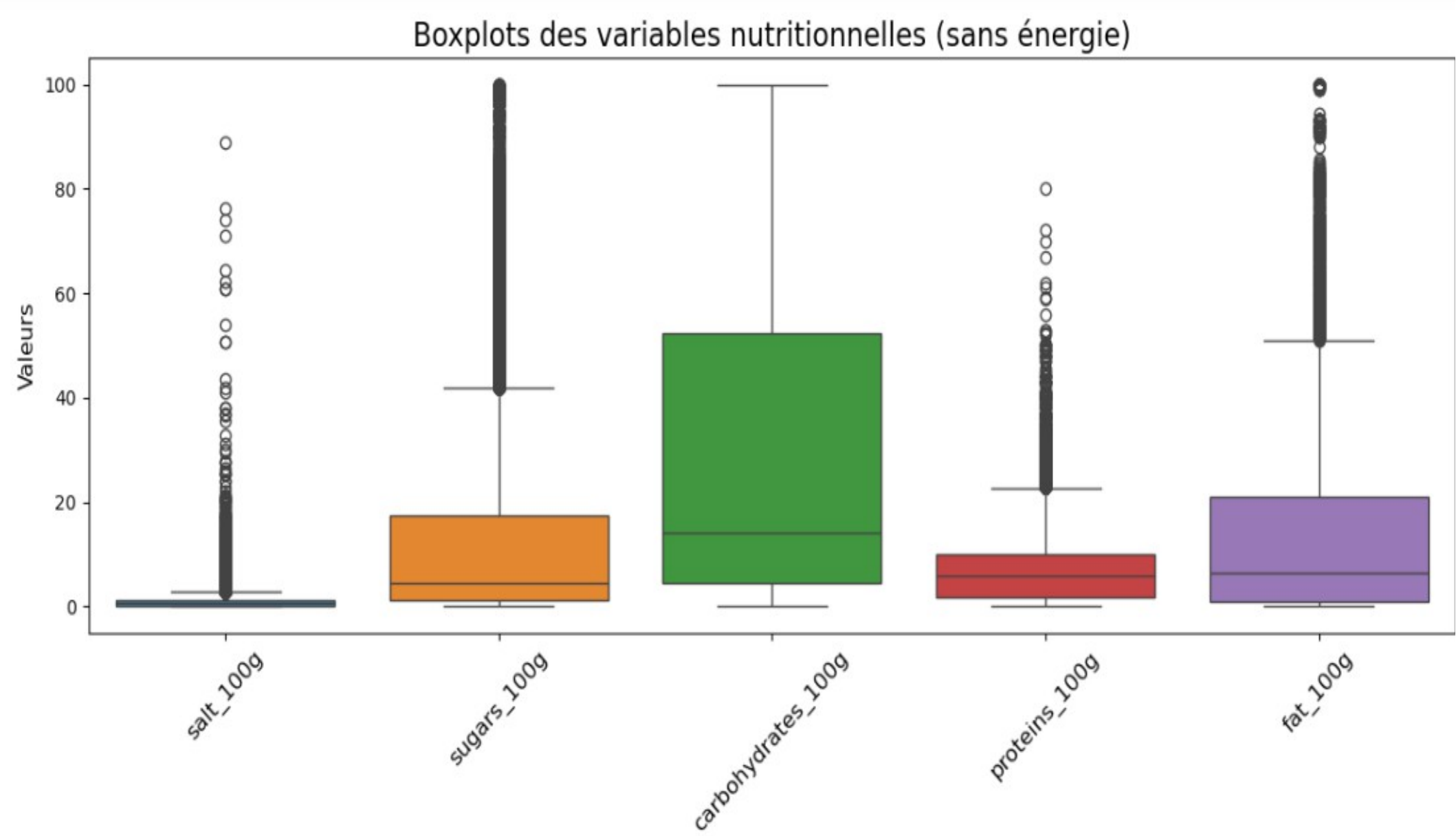
Critères de validation :

- Distribution des données conservée ✓
- Relations entre variables maintenues ✓
- Aucune incohérence n'a été introduite ✓

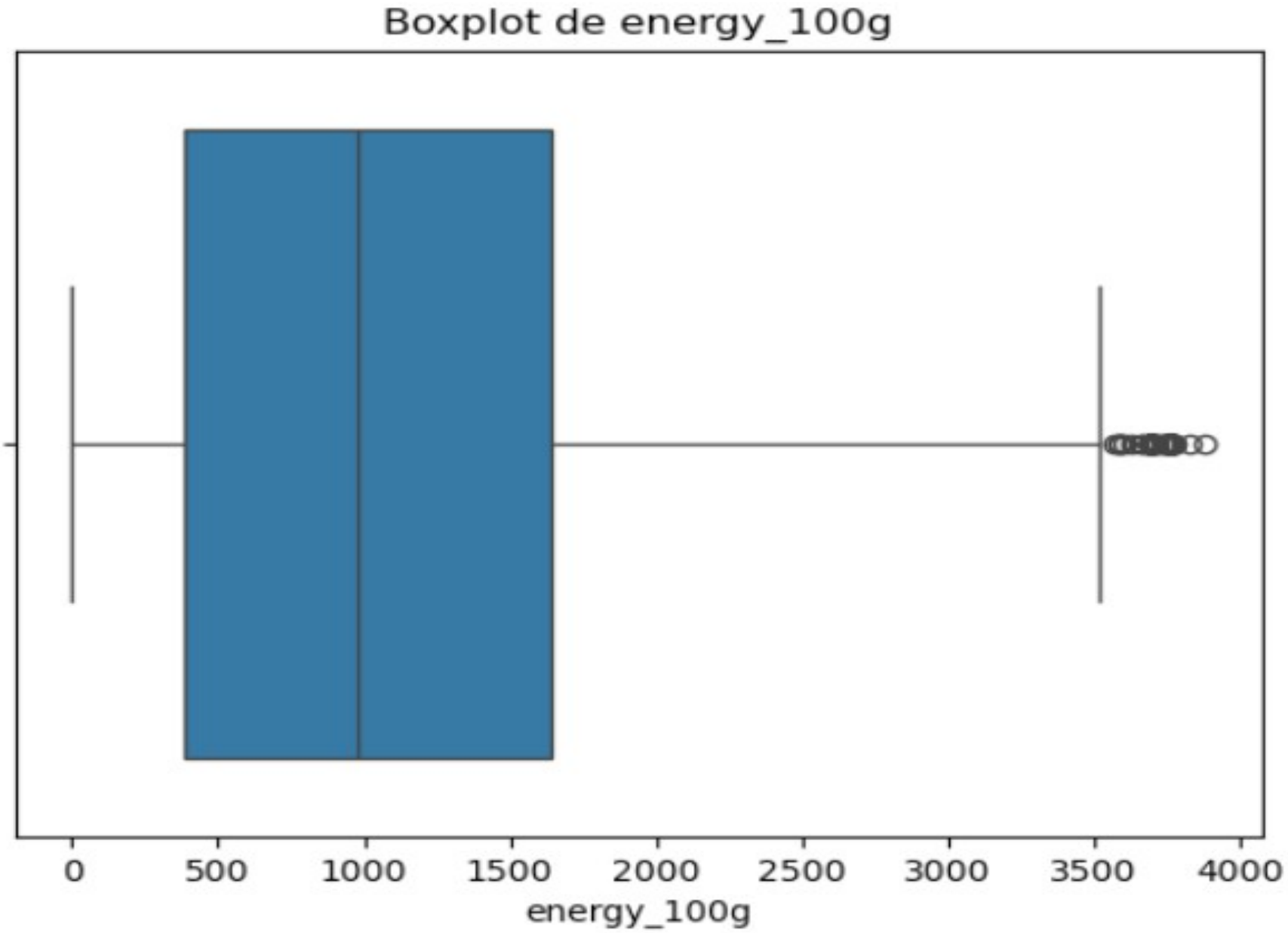
Implications :

- Méthode d'imputation fiable et robuste
- Données complétées sans biais significatif
- Prêtes pour les analyses statistiques suivantes

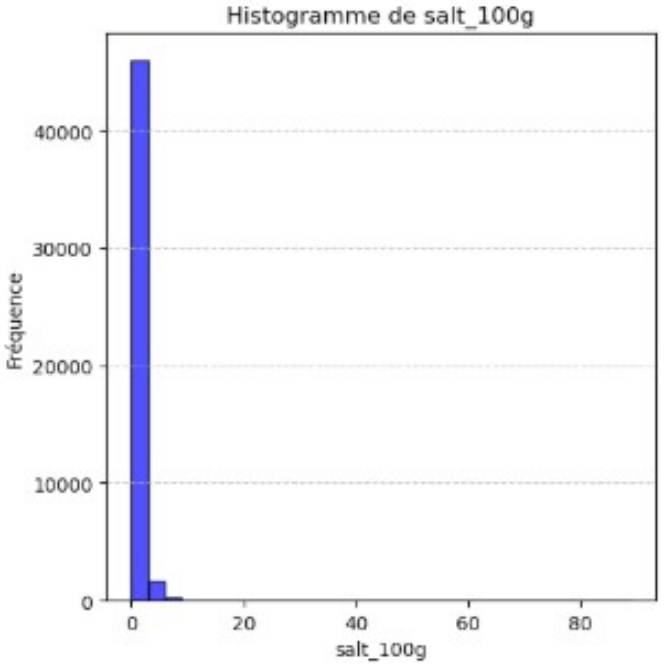
Analyse univariée des données

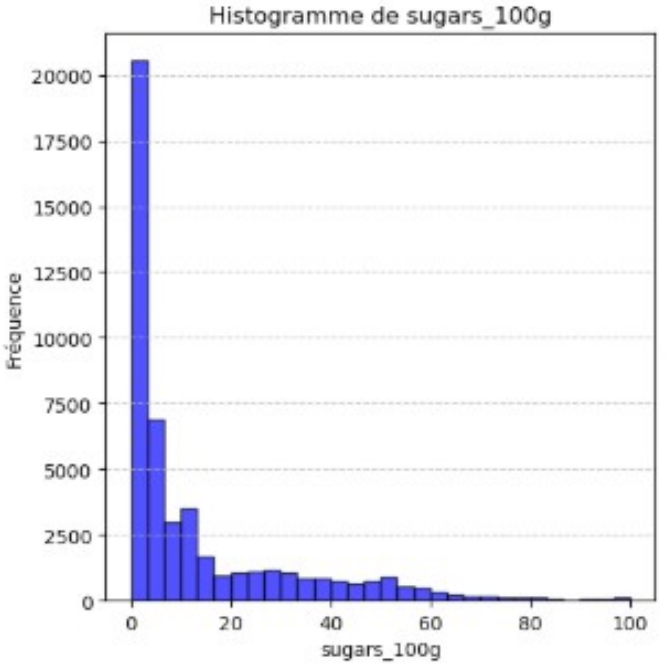


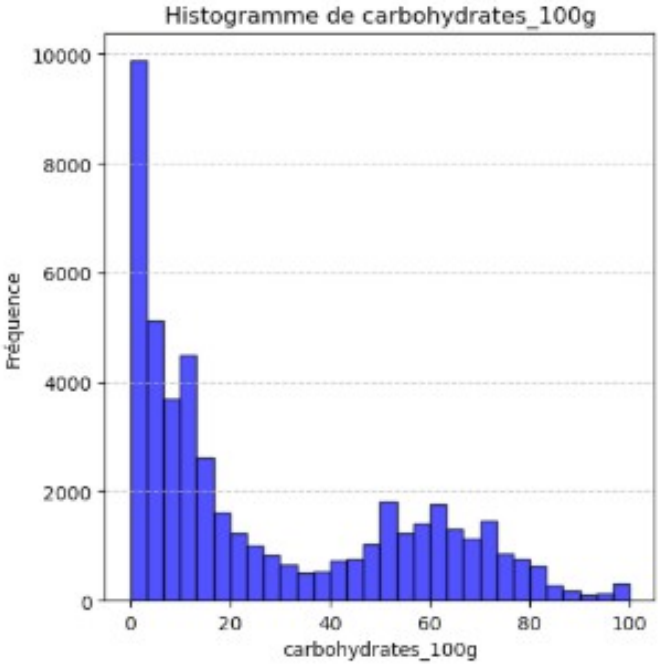
Matières grasses : Médiane 6.5g – Outliers (produits très gras)
Protéines : Médiane 5.9g – Variabilité modérée
Sel : Médiane 0.6g – Produits très salés identifiés
Sucre : Médiane 4.6g – Moyenne élevée (produits très sucrés)
Glucides : Médiane 14.1g – Variabilité importante

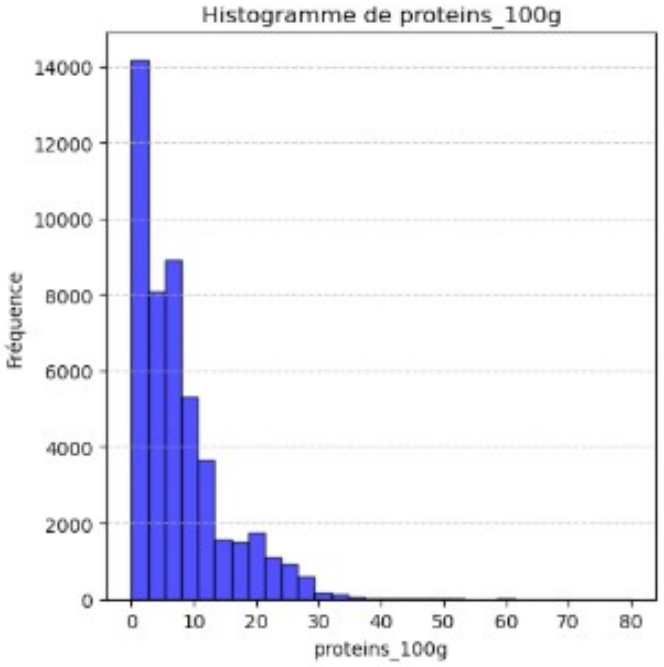


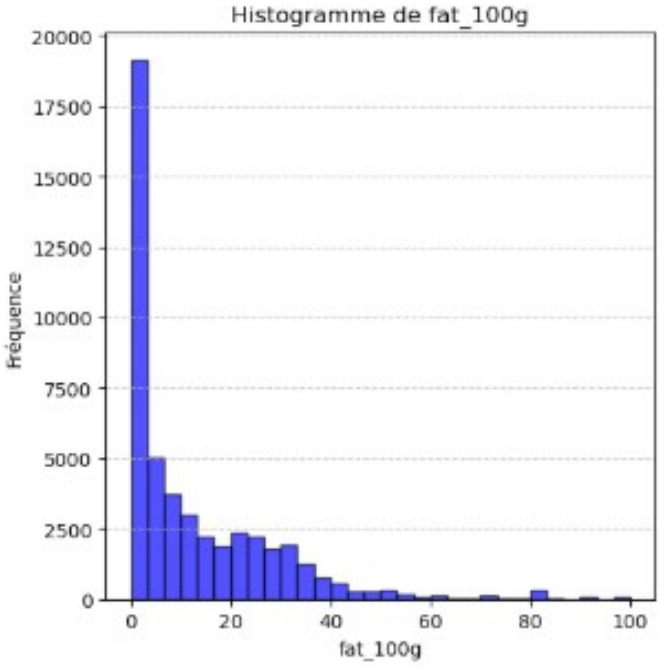
Médiane : 982 kJ
Outliers : produits extrêmement caloriques (>3500 kJ)

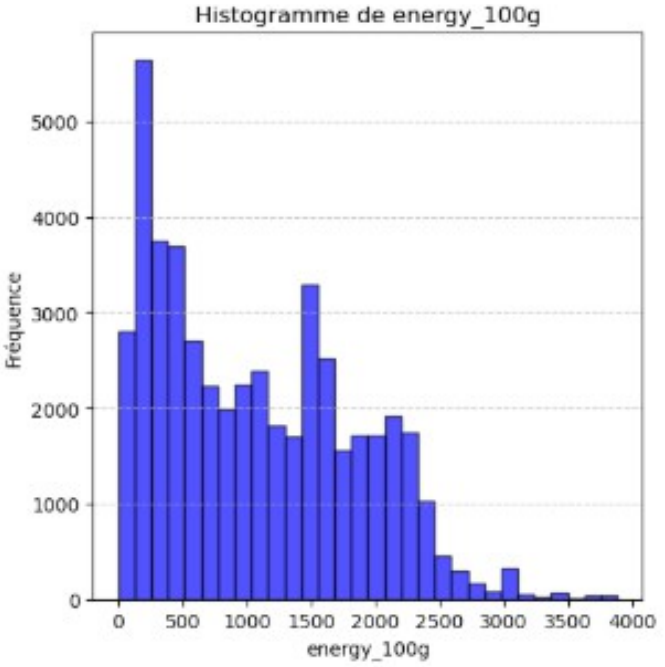






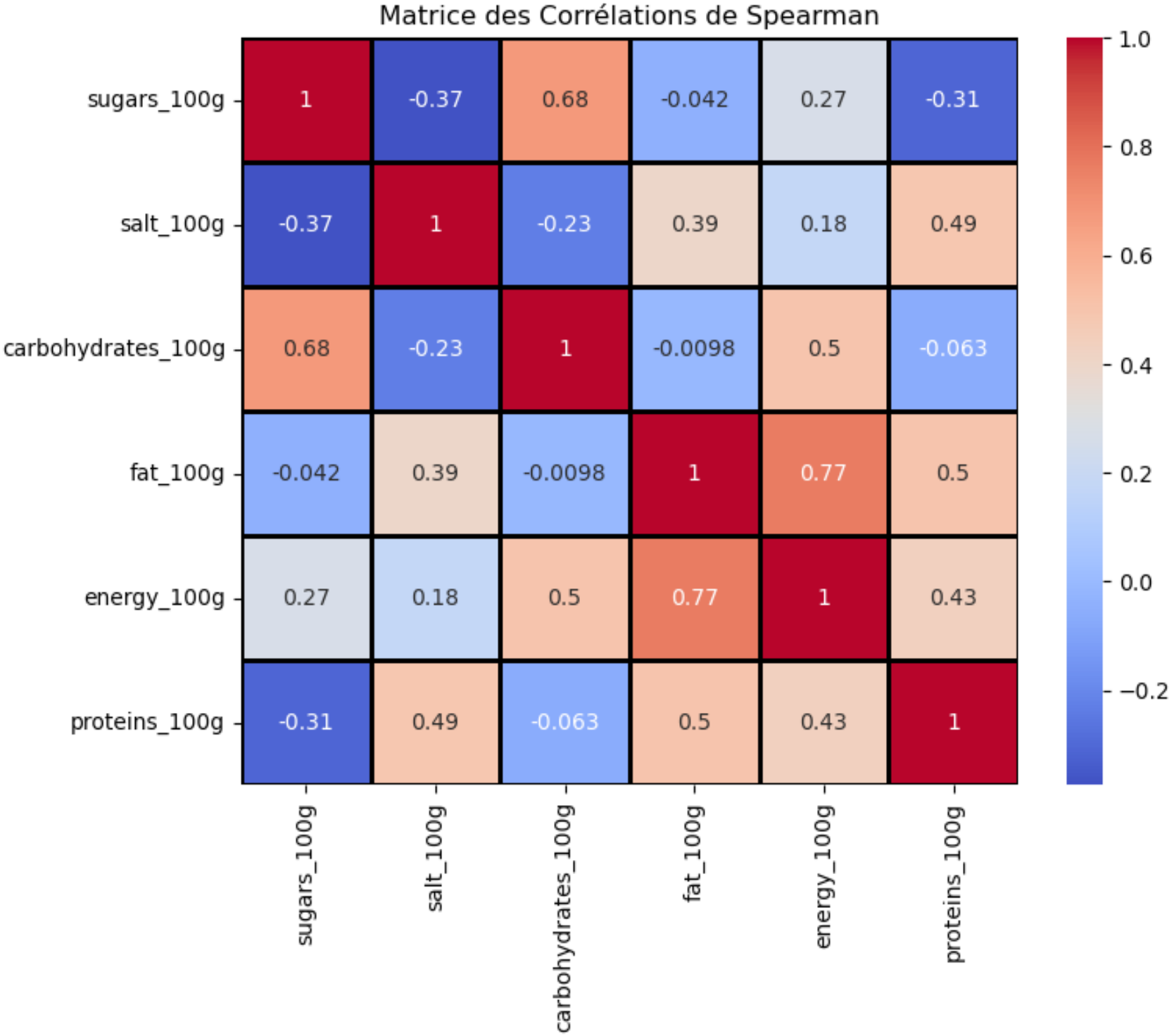






Sucre : Majorité faible en sucre – Longue traîne (produits très sucrés)
Sel : Produits majoritairement faibles en sel – Quelques valeurs élevées
Glucides : Large gamme – Pic autour de valeurs modérées

Analyse bivariée des données


Matrice des Corrélations de Spearman
sugars_100g
salt_100g
carbohydrates_100g
fat_100g
energy_100g
proteins_100g
1 -0.37 0.68 -0.042 0.27 -0.31
-0.37 1 -0.23 0.39 0.18 0.49
0.68 -0.23 1 -0.0098 0.5 -0.063
-0.042 0.39 -0.0098 1 0.77 0.5
0.27 0.18 0.5 0.77 1 0.43
-0.31 0.49 -0.063 0.5 0.43 1
sugars_100g
salt_100g
carbohydrates_100g
fat_100g
energy_100g
proteins_100g
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Résultats des Tests de Kruskal-Wallis

Utilité de ce test:

- Comparaison de groupes : Évalue les différences entre trois groupes ou plus.
- Données non paramétriques : Utilisé lorsque **les données ne suivent pas une distribution normale.**
- Résultats interprétables : Indique si au moins un groupe diffère des autres.

Variable	P-value du test de Kruskall-Wallis	Interprétation
Energie	0	Différence significative
Sucres	0	Différence significative
Glucides	0	Différence significative
Gras	0	Différence significative
Protéines	0	Différence significative
Sel	0	Différence significative

Analyse détaillée des différences d'énergie entre groupes alimentaires

Test de Dunn

Utilité :

- Compare les groupes par paires
- Identifie quels groupes diffèrent significativement

Résultat :

- Tous les groupes alimentaires diffèrent significativement en termes d'énergie

Analyse multivariée des données

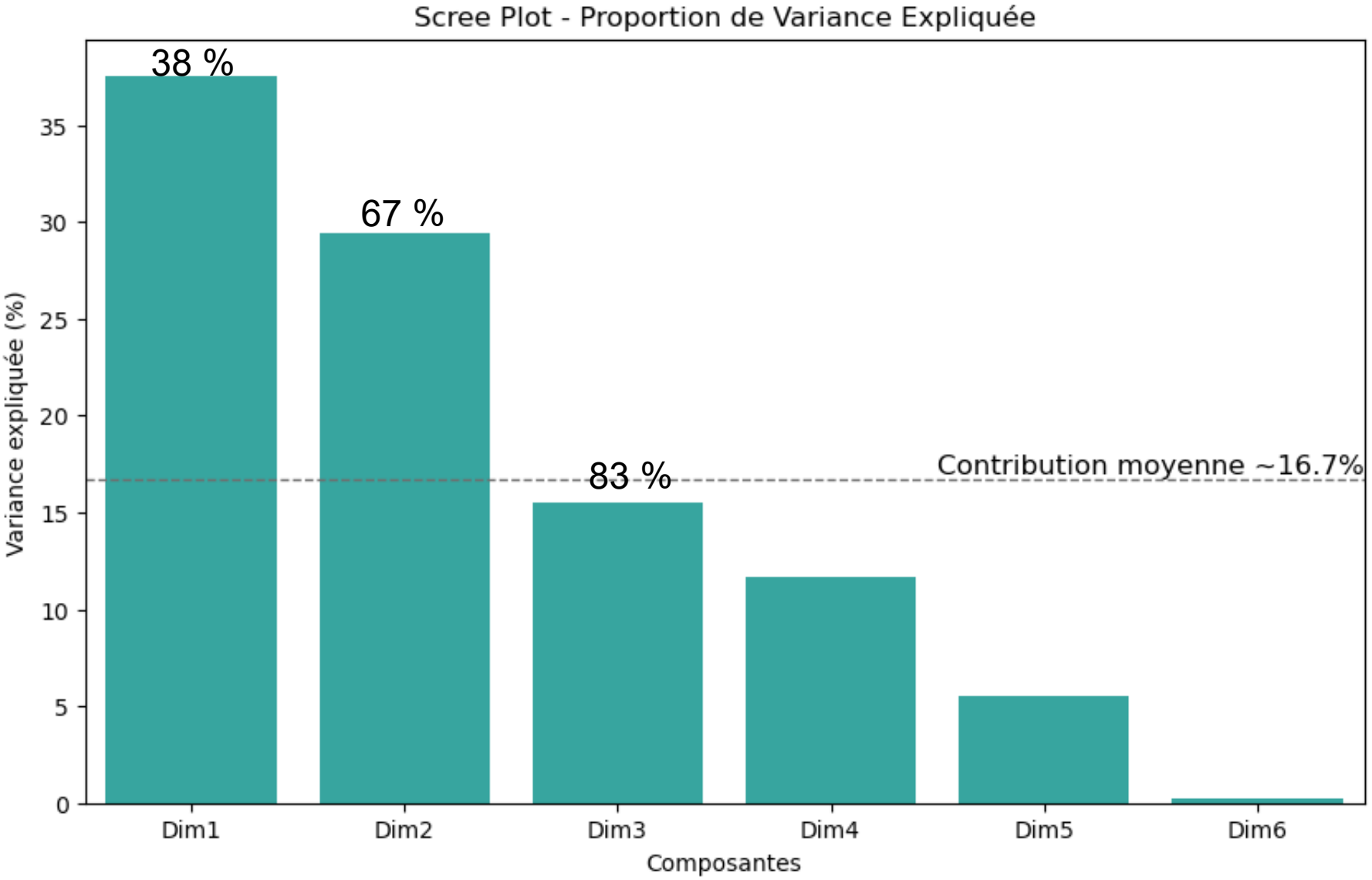
Qu'est-ce qu'une Analyse en Composantes Principales (ACP) ?

Objectif :
•Réduire la dimensionnalité des données tout en conservant un maximum d'information.
•Identifier des combinaisons linéaires des variables initiales (composantes principales) qui expliquent la variance des données.

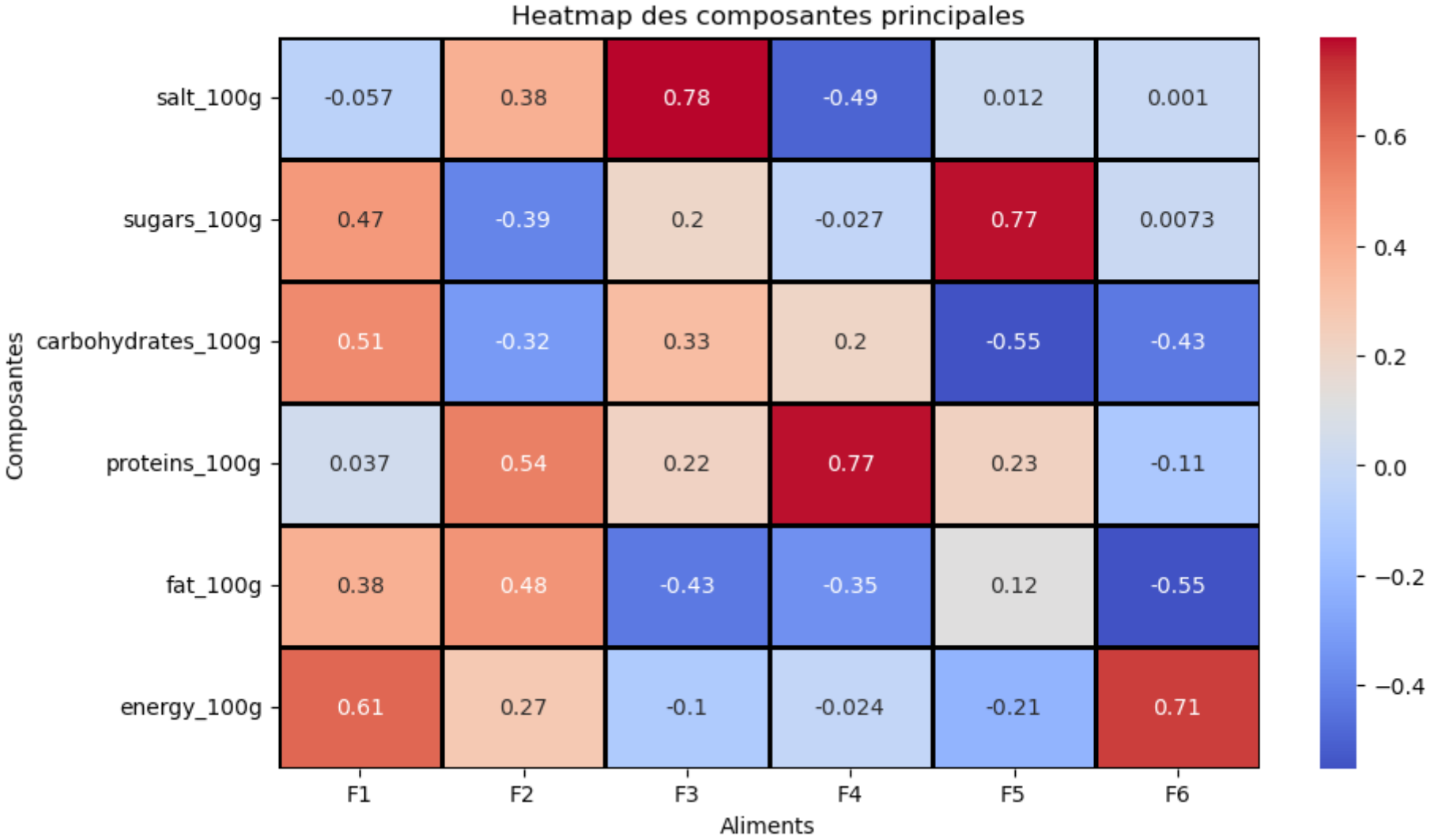
Pourquoi utiliser l'ACP ?
•Simplifier l'analyse des données multivariées complexes.
•Visualiser les relations entre les variables et les observations.
•Identifier les variables qui contribuent le plus à la structure globale des données.

Résultats clés :
•Chaque composante principale explique une part de la variance totale.
•Les premières composantes capturent généralement l'essentiel de l'information.

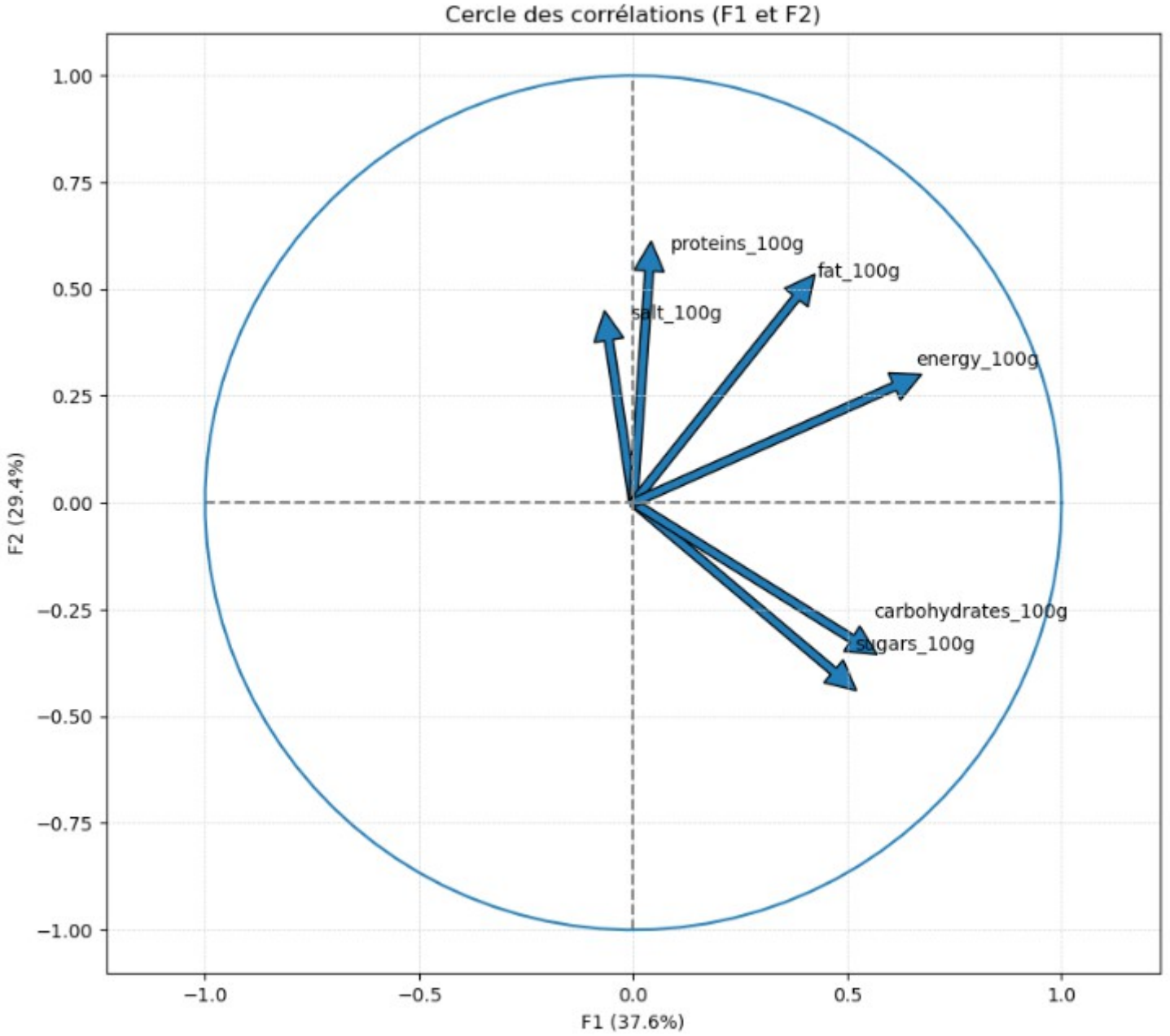
Proportion de variance expliquée par les composantes principales (ACP)



Coefficients des Variables dans les Composantes Principales (Heatmap)



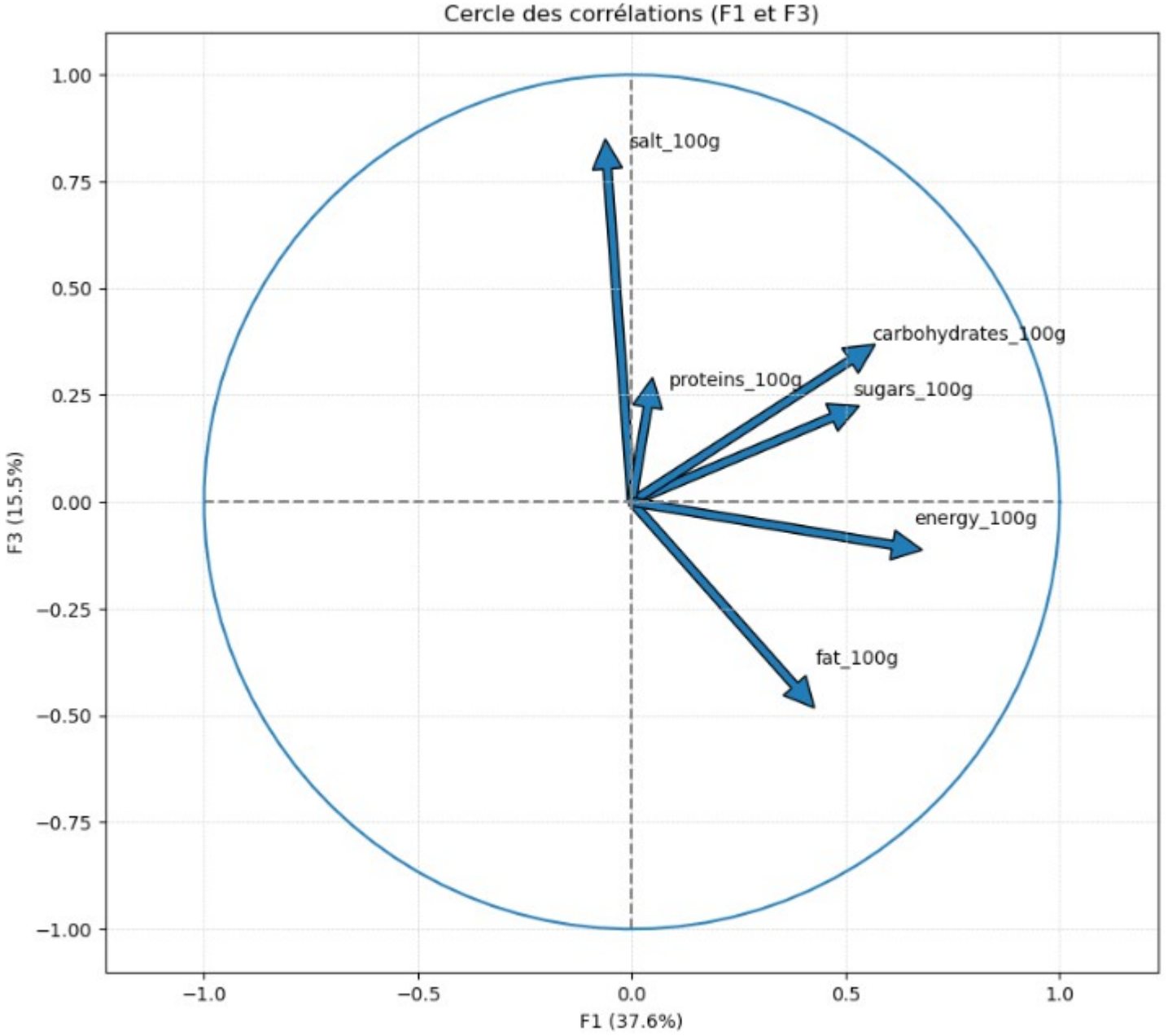
Cercle des Corrélations F1-F2 : Interprétation des Relations entre Variables



F1 : Densité énergétique (glucides, sucres, énergie).

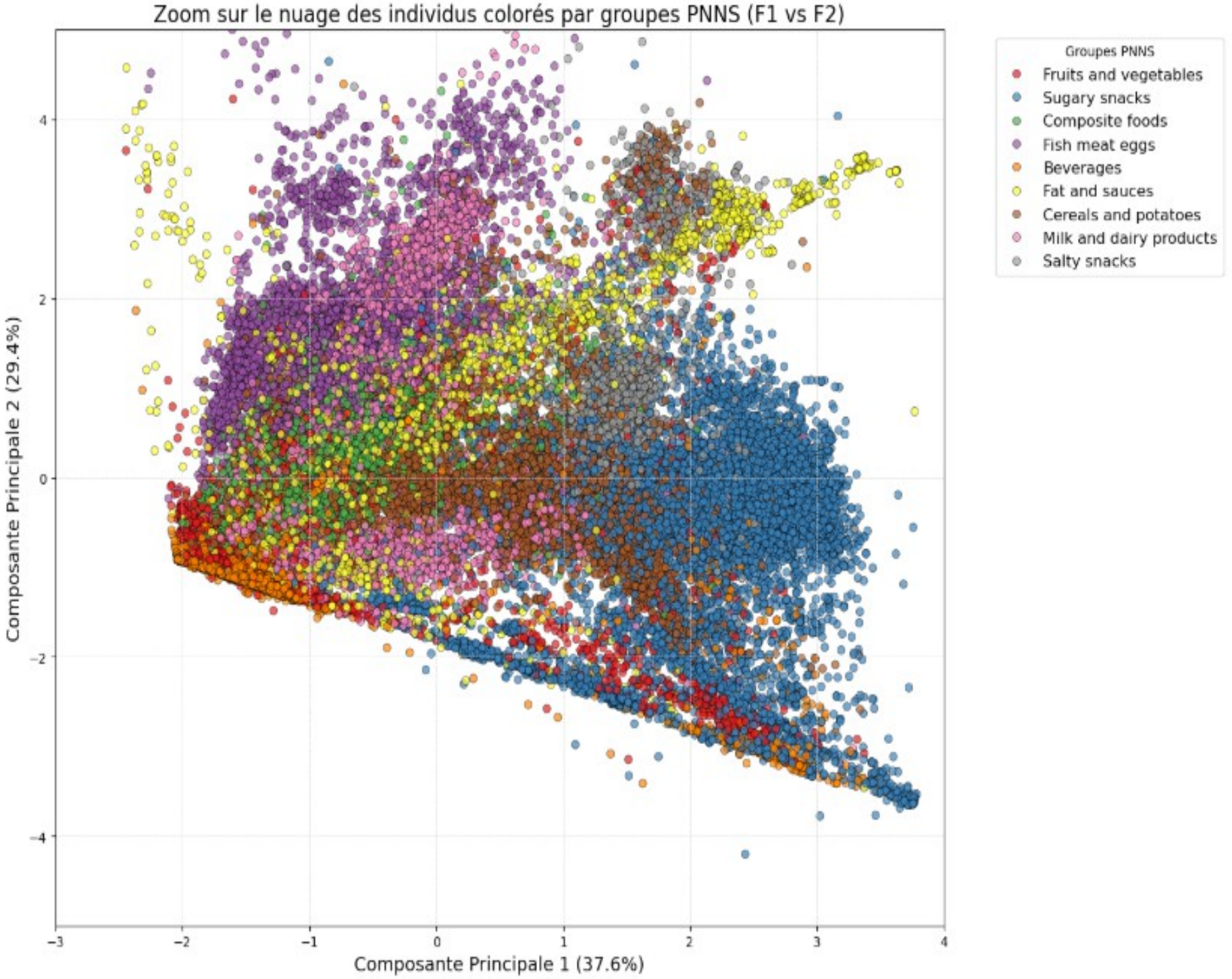
F2 : Opposition satiété (protéines/graisses) vs énergie rapide (glucides/sucres).

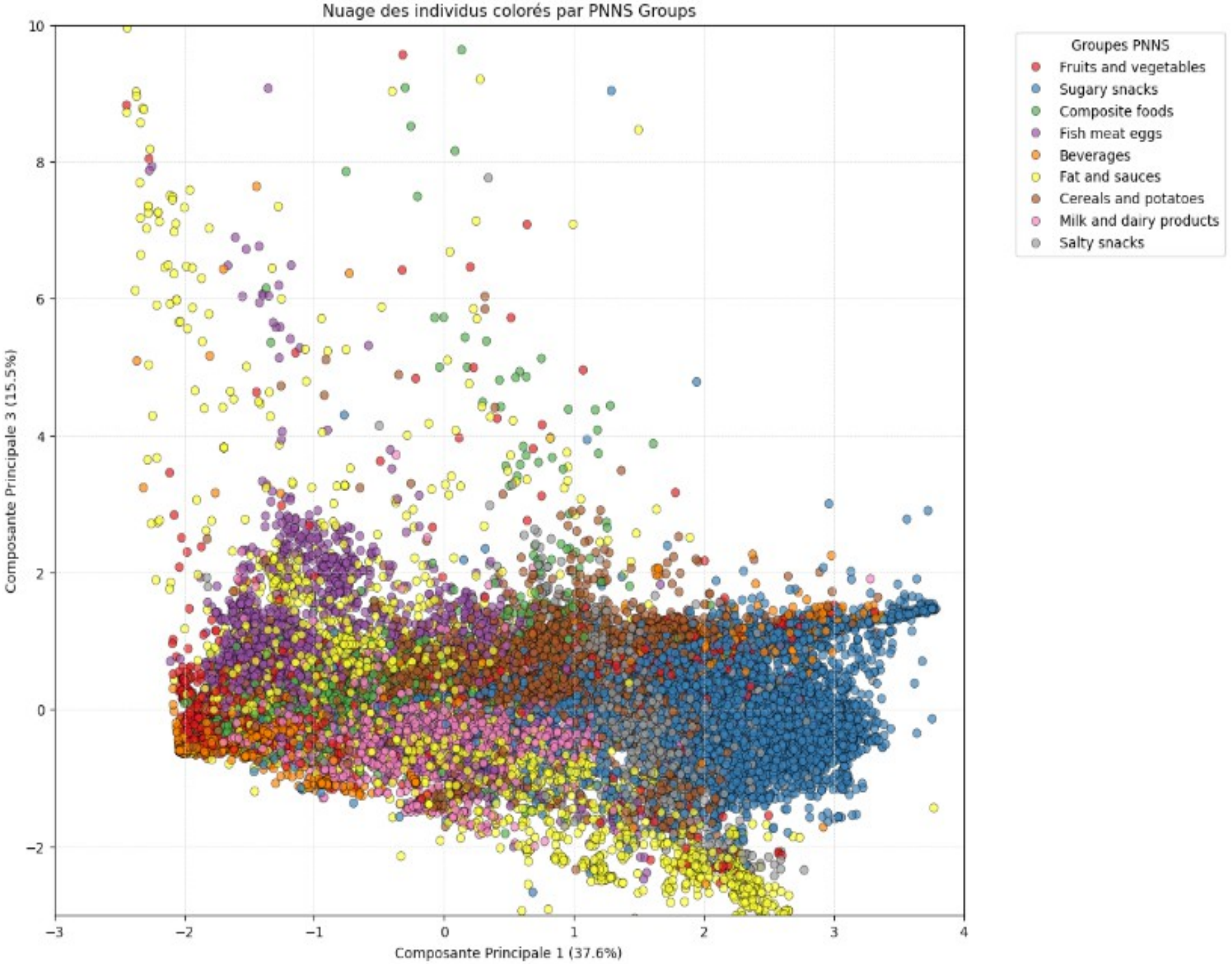
Cercle des Corrélations F1-F3 : Interprétation des Relations entre Variables



F1 : Densité énergétique (glucides, sucres).

F3 : Teneur en sel opposée à la teneur en graisses.


Zoom sur le nuage des individus colorés par groupes PNNS (F1 vs F2)
Groupes PNNS
Fruits and vegetables
Sugary snacks
Composite foods
Fish meat eggs
Beverages
Fat and sauces
Cereals and potatoes
Milk and dairy products
Salty snacks
Composante Principale 2 (29.4%)
Composante Principale 1 (37.6%)
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
4


Nuage des individus colorés par PNNS Groups
Groupes PNNS
Fruits and vegetables
Sugary snacks
Composite foods
Fish meat eggs
Beverages
Fat and sauces
Cereals and potatoes
Milk and dairy products
Salty snacks
Composante Principale 3 (15.5%)
Composante Principale 1 (37.6%)
10
8
6
4
2
0
−2
−3
−2
−1
0
1
2
3
4

Conclusion

Préparation des Données: Traitement des valeurs aberrantes et des valeurs manquantes pour garantir la qualité des données.

Résultats Statistiques
Les tests montrent des différences significatives entre les groupes PNNS pour nos variables

Analyse en Composantes Principales (ACP)
Réduction de la dimensionnalité avec 3 axes principaux :
•F1 : Densité énergétique
•F2 : Satiété
•F3 : Teneur en sel.

Modélisation Prédictive
•Construction d'un modèle prédictif basé sur les caractéristiques nutritionnelles.
•Précision satisfaisante dans la classification des produits.

Base solide pour classifier efficacement les produits alimentaires selon leurs groupes PNNS.