Projet n°3

Préparation de données pour un organisme de santé publique

SANTÉ PUBLIQUE FRANCE

OPENCLASSROOMS

Sommaire

INTRODUCTION

Mission et objectifs

Le projet de l'agence de Santé publique France est d'améliorer sa base de données **Open Food Facts**.
Notre objectif ici est la prise en main, le nettoyage et l'exploration des données en vue de la création d'un système d'auto-complétion.

PROCÉDURE :

- Exploration des données
- Sélection de la variable cible
- Nettoyage des données
- Analyse des données

État des lieux

Notre jeu de données est composé d'un fichier `.csv`, que l'on nommera `data` et dont les caractéristiques sont les suivantes :

Information	**Valeur**
Nombre de lignes	320 772
Nombre de colonnes	162
Nombre de colonnes float	106
Nombre de colonnes object	56

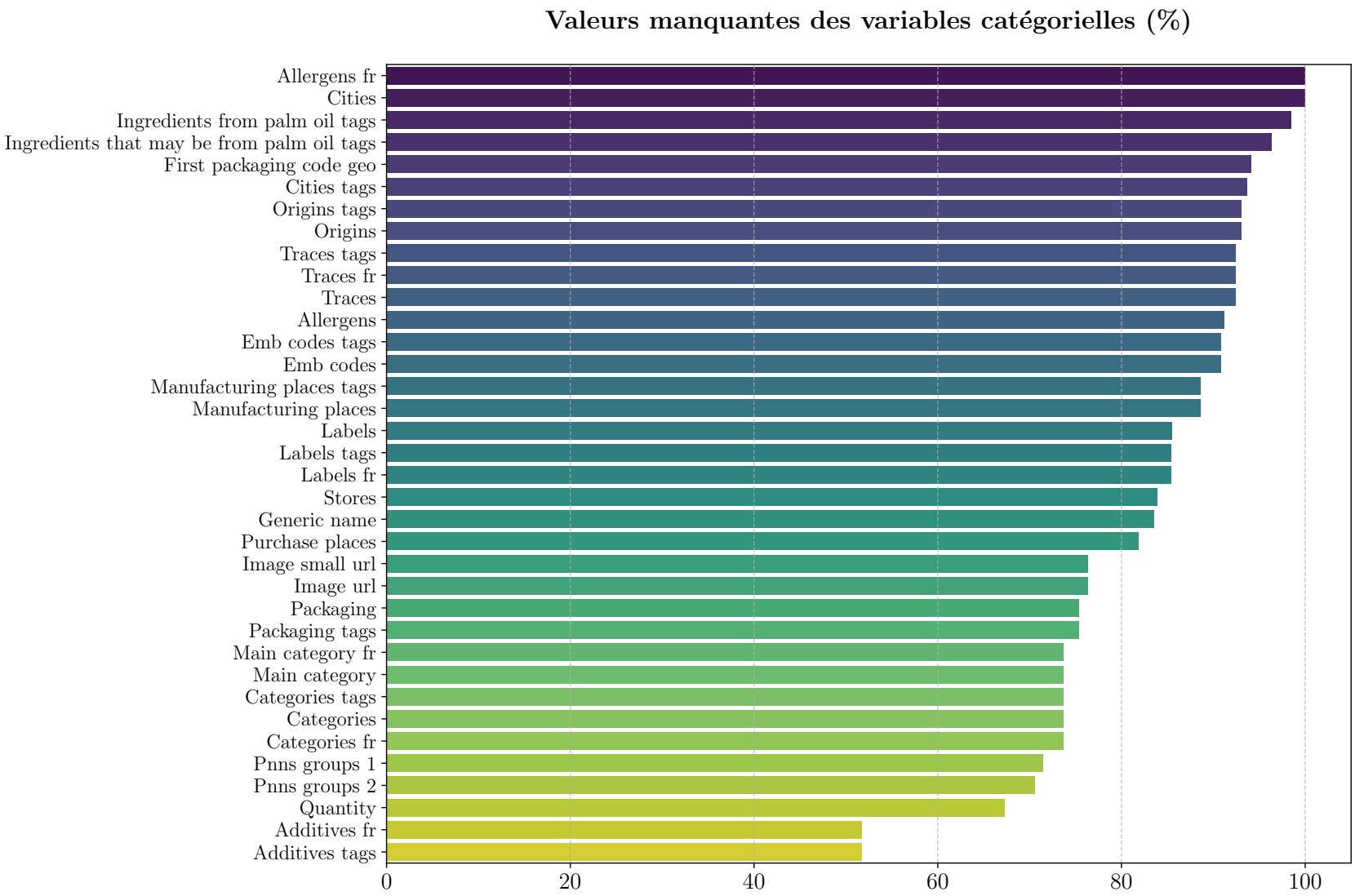
Table: Résumé descriptif de `data`

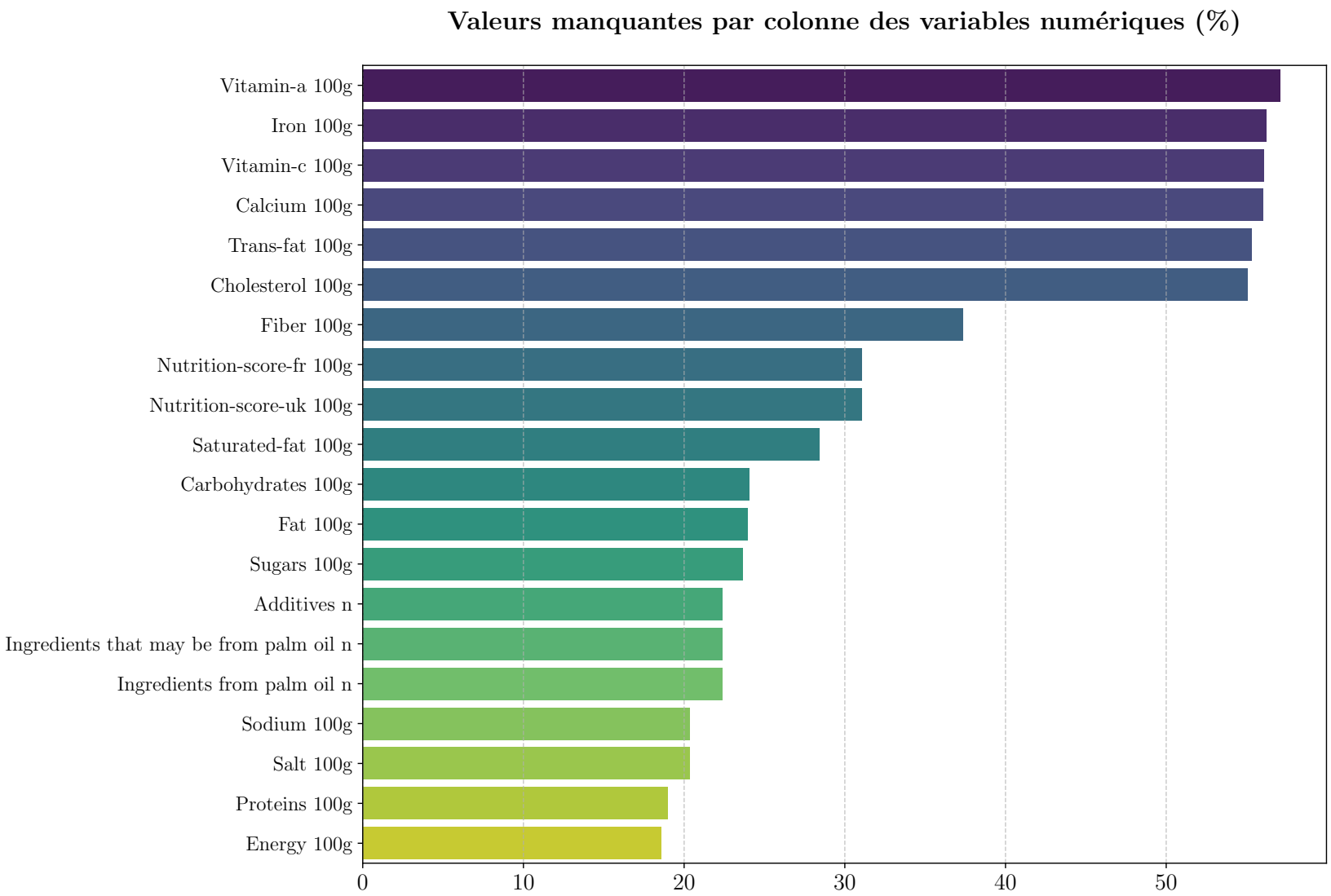
EXPLORATION DES DONNÉES

Description des données

- Chaque produit est identifié par son code barre.
- Variables catégorielles
  - Informations générales : Nom, Catégorie, Type, URL …
  - Informations complémentaires : Allergènes, Additifs, Origine, Label …
- Variables numériques
  - Données nutritionnelles : Energie, Glucides, Graisses, Sel …
  - Données complémentaires : Nutriscore, Nombre d'additifs …
- Aucun doublon détecté dans le jeu de données `data`

Valeurs manquantes




Valeurs manquantes par colonne des variables numériques (%)
Vitamin-a 100g
Iron 100g
Vitamin-c 100g
Calcium 100g
Trans-fat 100g
Cholesterol 100g
Fiber 100g
Nutrition-score-fr 100g
Nutrition-score-uk 100g
Saturated-fat 100g
Carbohydrates 100g
Fat 100g
Sugars 100g
Additives n
Ingredients that may be from palm oil n
Ingredients from palm oil n
Sodium 100g
Salt 100g
Proteins 100g
Energy 100g
0
10
20
30
40
50

SÉLECTION DES DONNÉES PERTINENTES

Choix de la variable cible

Critères de sélection de la variable cible :

- Plus de 50% de valeurs manquantes
- Moins de 80% de valeurs manquantes
- Un nombre de valeurs uniques relativement petit (< 100)
- Doit pouvoir être déterminé en fonction des variables numériques

On a donc 2 variables catégorielles qui respectent ces conditions

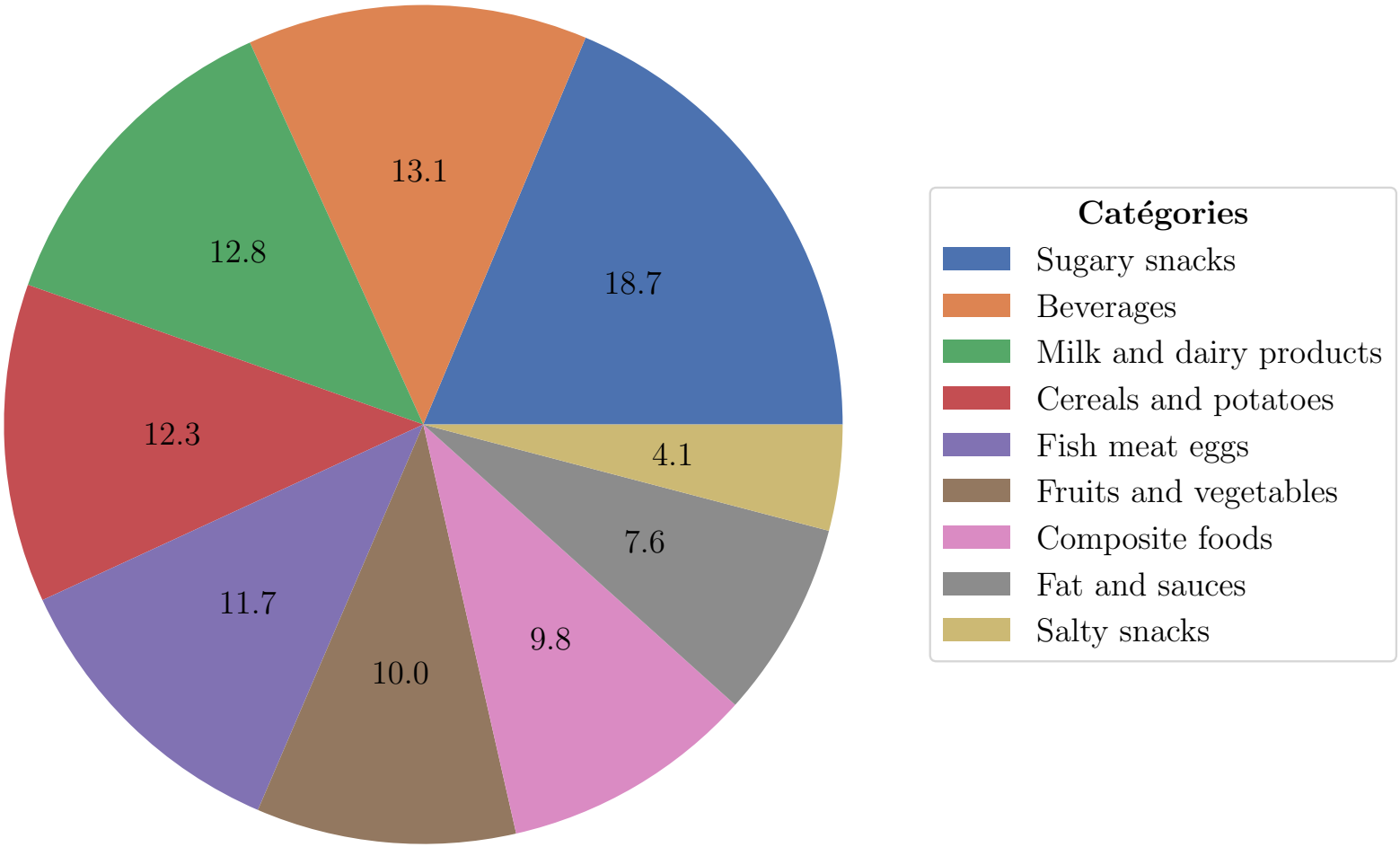
`pnns_groups_1` et `pnns_groups_2`

Variable	`pnns_groups_1`	`pnns_groups_2`
Lignes	68889	71867
Taux lignes (%)	54	52
Valeurs uniques	9	36

Finalement, on choisira `pnns_groups_1`.

On ne garde que les lignes de `data` pour lesquelles la colonne `pnns_groups_1` est renseignée.

Proportions des catégories PNNS GROUPS 1



Sélection des variables numériques

- Suppression de toutes les variables numériques ayant plus de 90% de valeurs manquantes
- Suppression des variables inutiles :
  - `additives_n`
  - `ingredients_from_palm_oil_n`
  - `ingredients_that_may_be_from_palm_oil_n`

- On a donc 11 variables numériques.
- On supprime tous les produits pour lesquels aucune variable n'est renseignée. (13022 lignes)

Variable	Traduction	VM (%)
`energy_100g`	Énergie	1, 50
`fat_100g`	Graisses	5, 70
`saturated-fat_100g`	Acides gras saturés	7, 55
`carbohydrates_100g`	Glucides	6, 44
`sugars_100g`	Sucres	7, 19
`fiber_100g`	Fibres	40, 62
`proteins_100g`	Protéines	2, 35
`salt_100g`	Sel	6, 91
`sodium_100g`	Sodium	6, 91
`ns-fr_100g`	Nutriscore français	9, 43
`ns-uk_100g`	Nutriscore anglais	9, 43

NETTOYAGE DES DONNÉES

Réglage des valeurs aberrantes

Avant tout, on sépare notre `data` en deux à l'aide des colonnes `quantity` et `product_name`

- `data_food` qui contiendra les produits de type `Nourriture`
- `data_drink` qui contiendra les produits de type `Boisson`

Dataframe	Nombre de lignes
`data_food`	48068
`data_drink`	7799

Transformation des valeurs aberrantes en valeurs manquantes

- Pour `energy_100g` :
  - Toutes les valeurs supérieures à 4000 pour `data_food`
  - Toutes les valeurs supérieures à 1200 pour `data_drink`
- Pour `ns-fr_100g` et `ns-uk_100g` :
  - Aucune valeur aberrante
- Pour toutes les autres variables :
  - Toutes les valeurs supérieures à 100
  - Toutes les valeurs négatives (inférieures à 0)

Cohérence des données

- Si `saturated-fat_100g` > `fat_100g` alors,

$$\texttt{saturated-fat\_100g} = \texttt{fat\_100g}$$

- Si `sugars_100g` > `carbohydrates_100g` alors,

$$\texttt{sugars\_100g} = \texttt{carbohydrates\_100g}$$

- Si `sodium_100g` > `salt_100g` alors,

$$\texttt{sodium\_100g} = \texttt{salt\_100g}$$

Remplissage des valeurs manquantes

- Pour la variable `fiber_100g`, les valeurs manquantes sont remplacées par 0
- Pour les autres, 3 méthodes utilisées en fonction des distributions
  - Imputation par la moyenne
  - Imputation par la médiane
  - Méthode des k plus proches voisins (k—NN)

Création d'un score

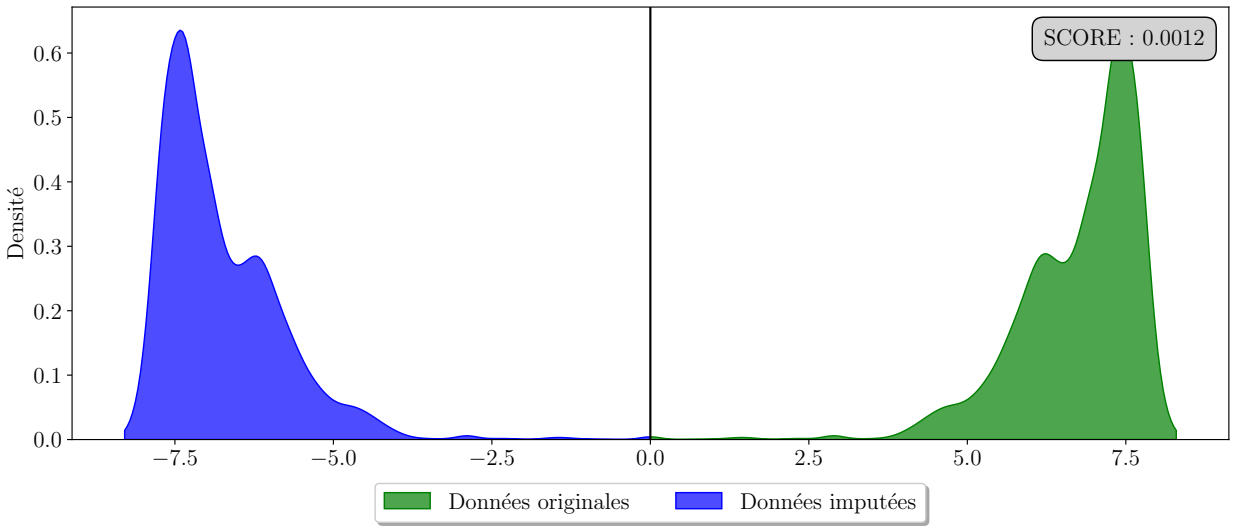
Les scores correspondent à la différence absolue moyenne entre la corrélation d'une variable avec les autres, avant et après imputation.

	data_food		data_drink	
Features	SCO MOY	SCO MED	SCO MOY	SCO MED
Energy	0, 0012	0, 0013	0, 1706	0, 1850
Fat	0, 0067	0, 0062	0, 0118	0, 0081
Saturated fat	0, 0065	0, 0072	0, 0069	0, 0047
Carbohydrates	0, 0074	0, 0090	0, 0079	0, 0081
Sugars	0, 0055	0, 0056	0, 0081	0, 0070
Proteins	0, 0009	0, 0011	0, 0034	0, 0035
Salt	0, 0177	0, 0178	0, 0010	0, 0016
Sodium	0, 0019	0, 0019	0, 0010	0, 0016

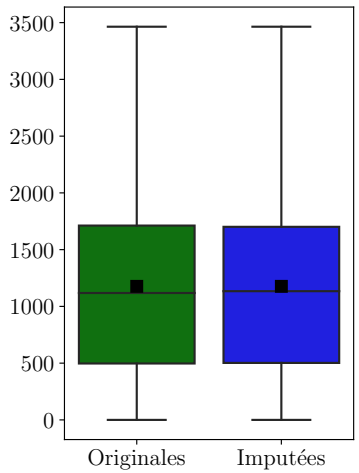
Nourritures

Energy – Imputations

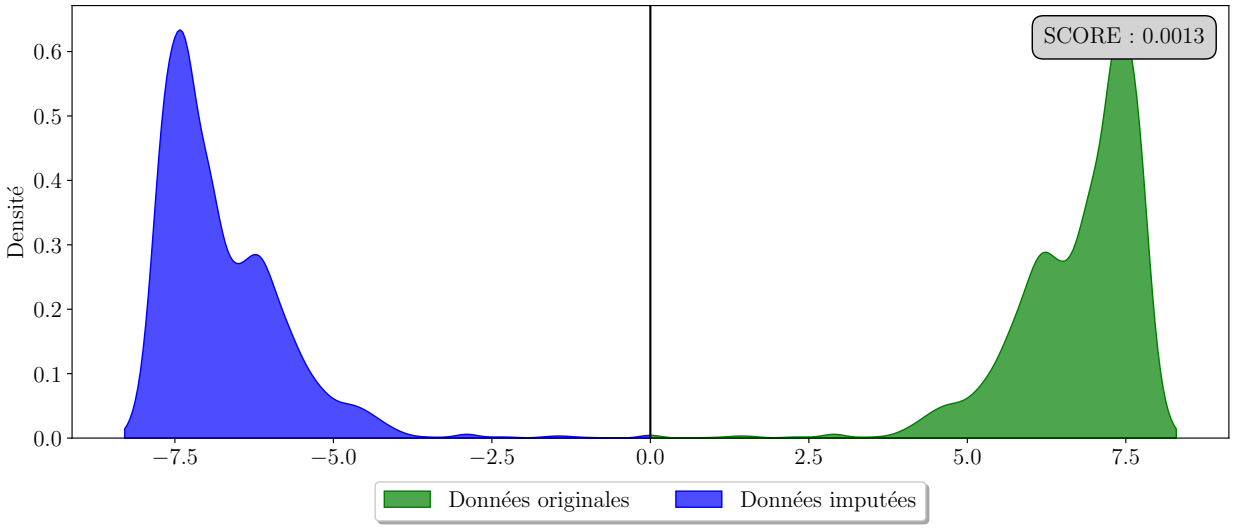
Distributions – Originales VS Moyenne



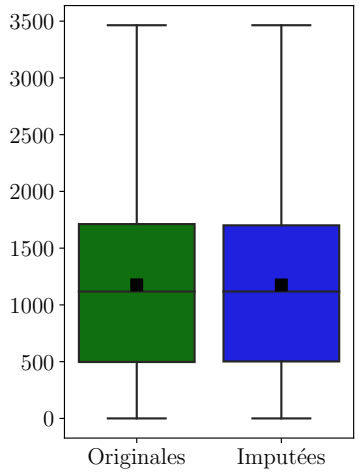
Boxplots



Distributions – Originales VS Médiane

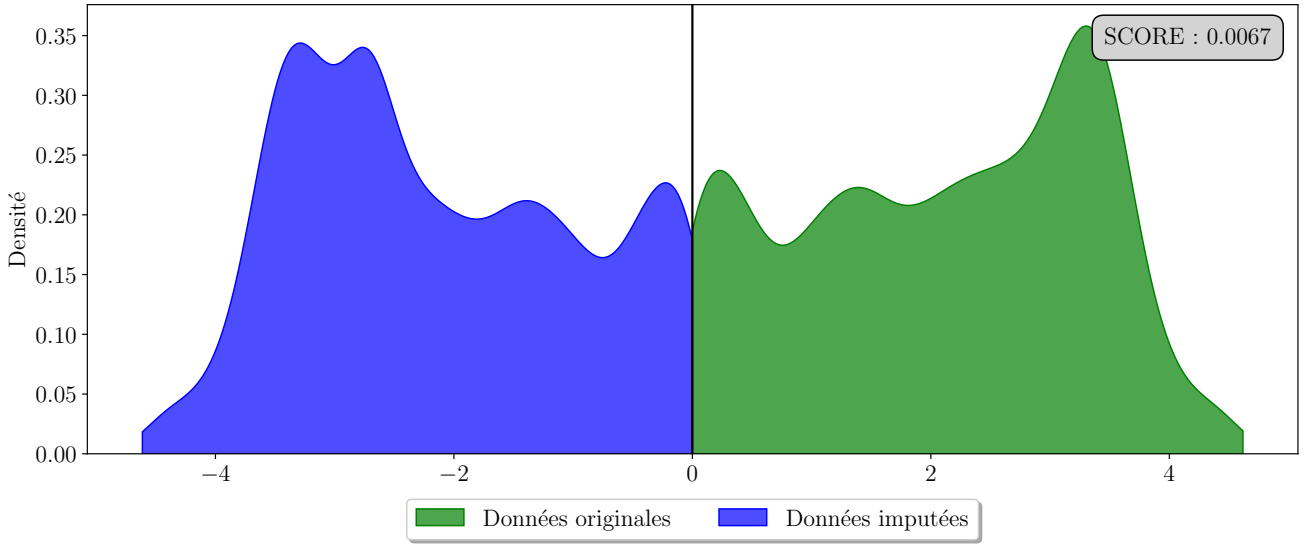


Boxplots

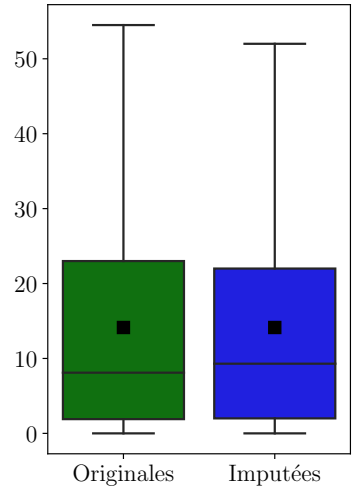


Fat – Imputations

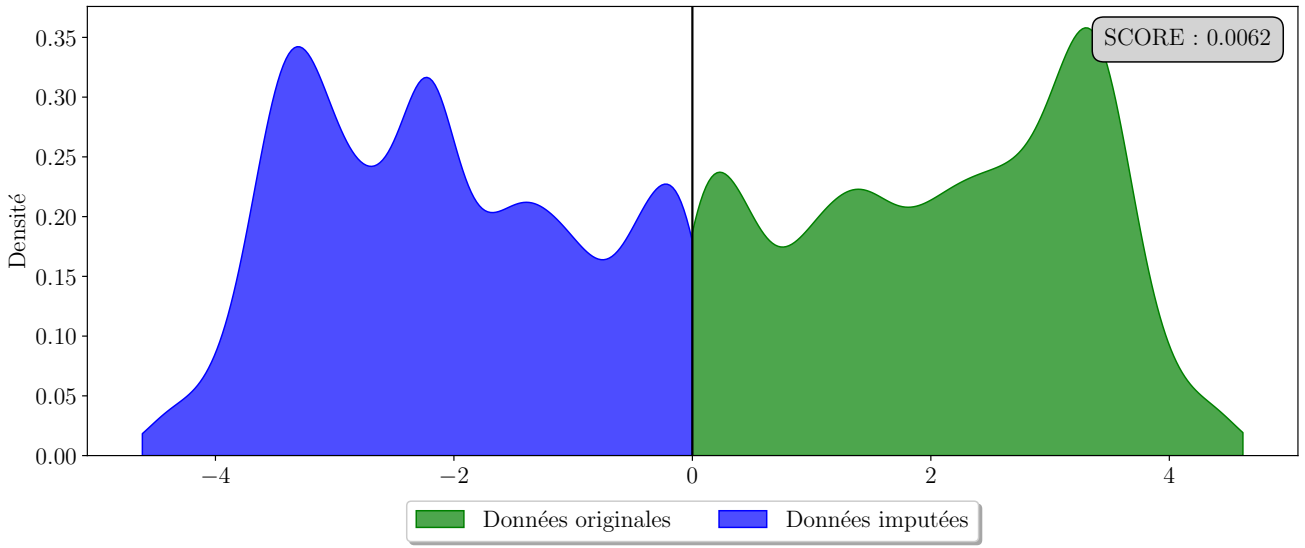
Distributions – Originales VS Moyenne



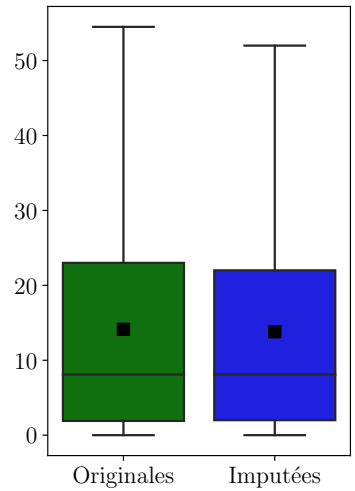
Boxplots



Distributions – Originales VS Médiane

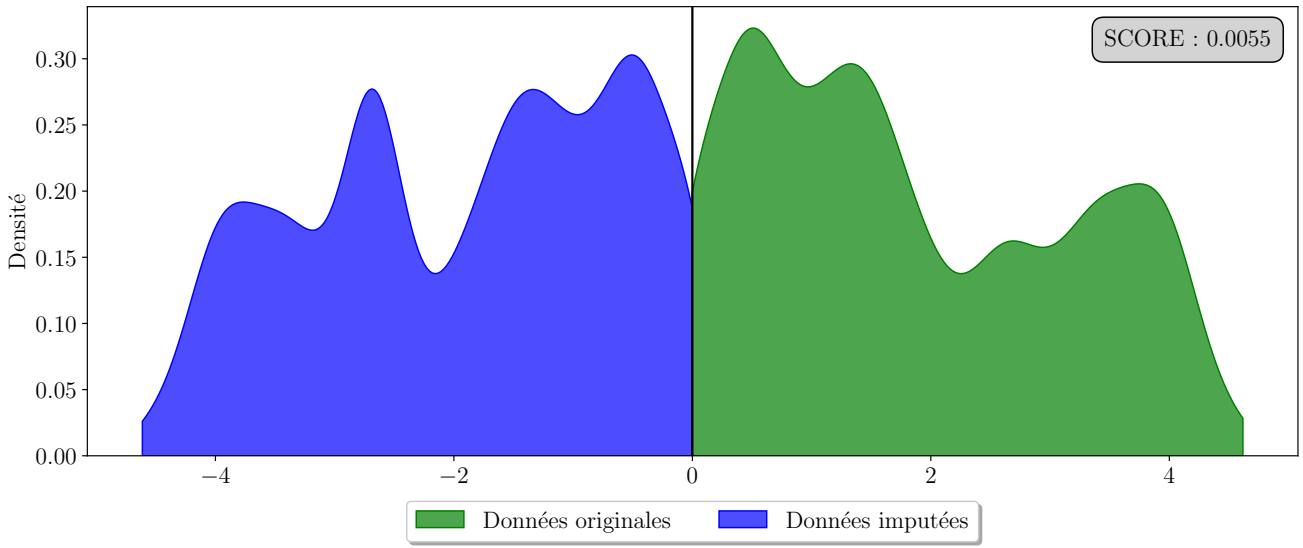


Boxplots

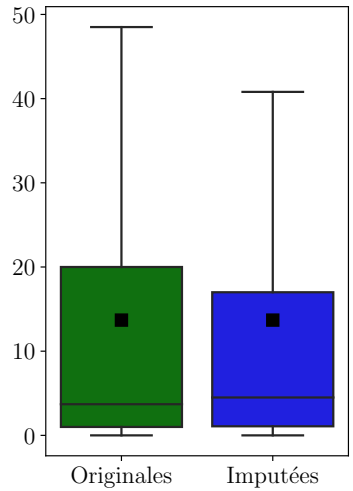


Sugars – Imputations

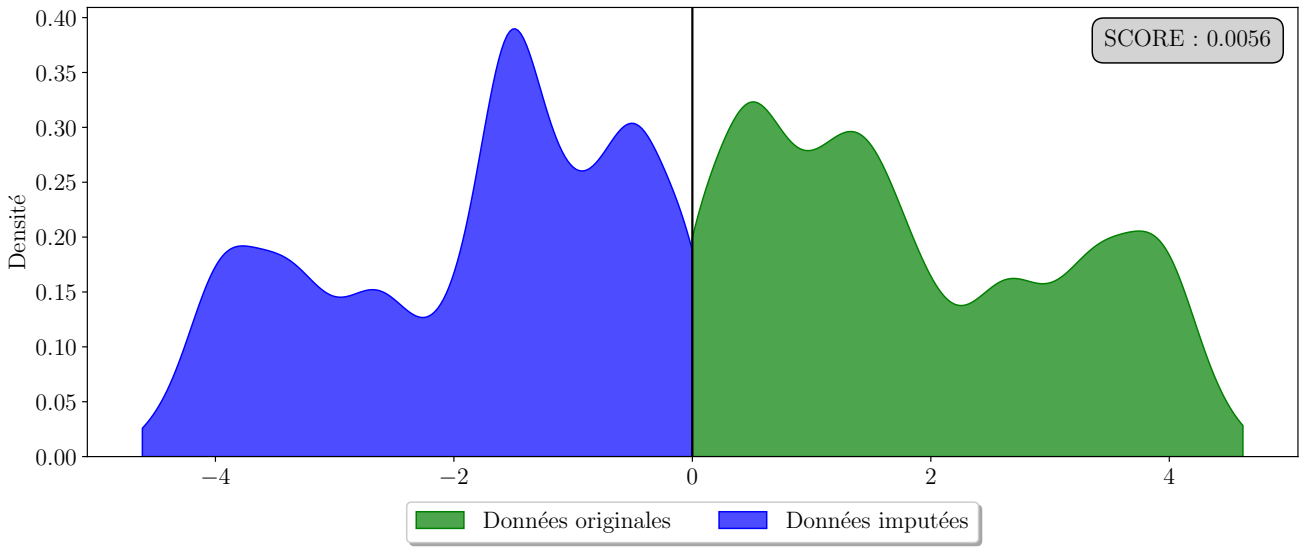
Distributions – Originales VS Moyenne



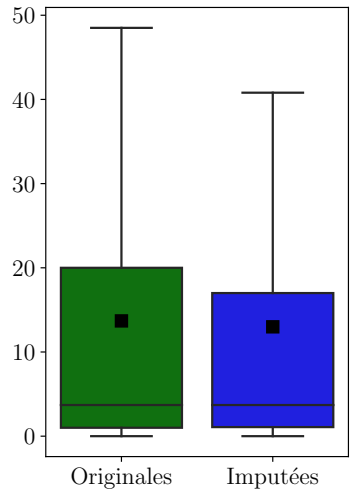
Boxplots



Distributions – Originales VS Médiane



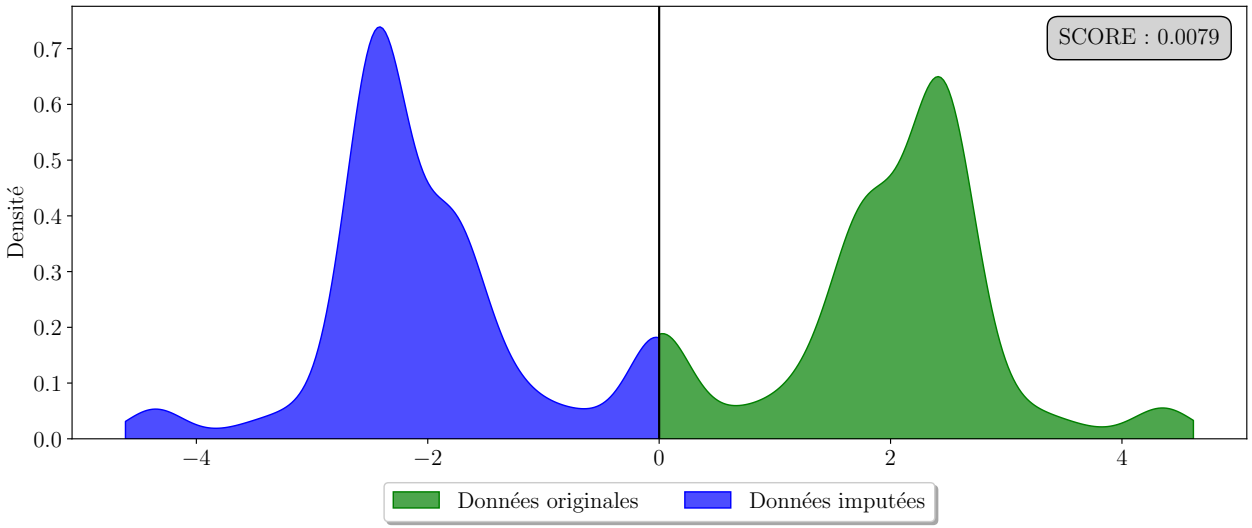
Boxplots



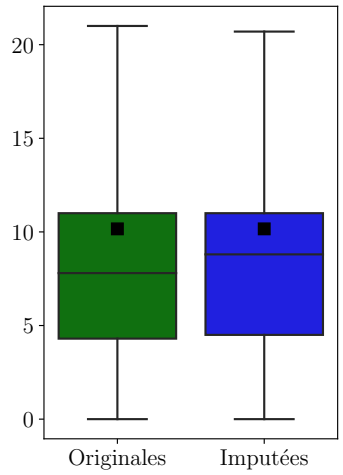
Boissons

Carbohydrates – Imputations

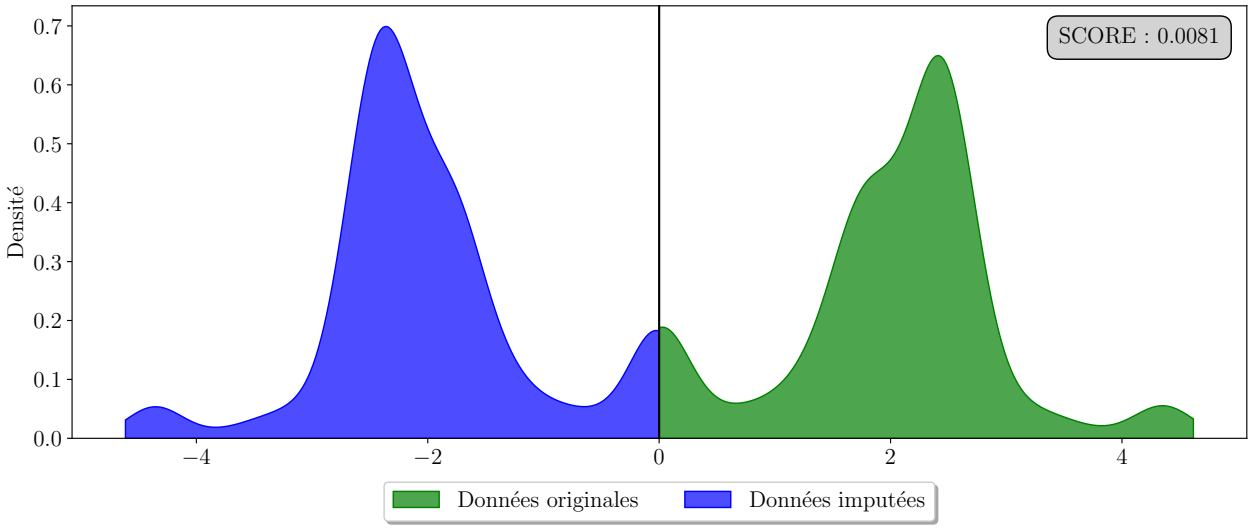
Distributions – Originales VS Moyenne



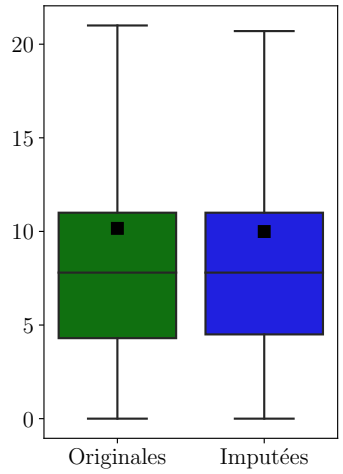
Boxplots



Distributions – Originales VS Médiane

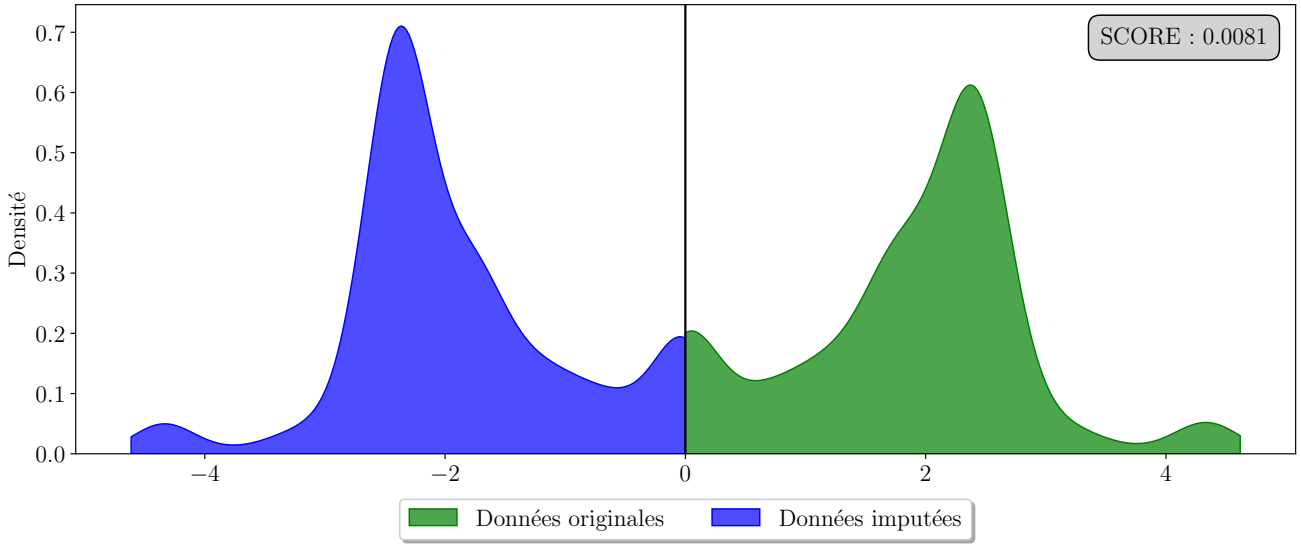


Boxplots

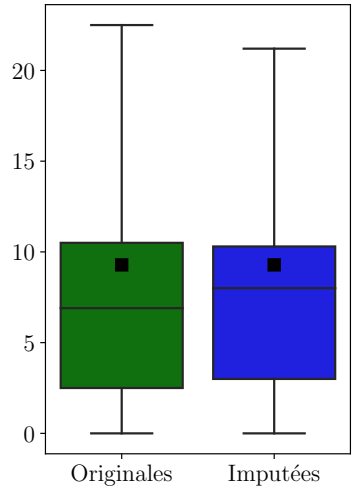


Sugars – Imputations

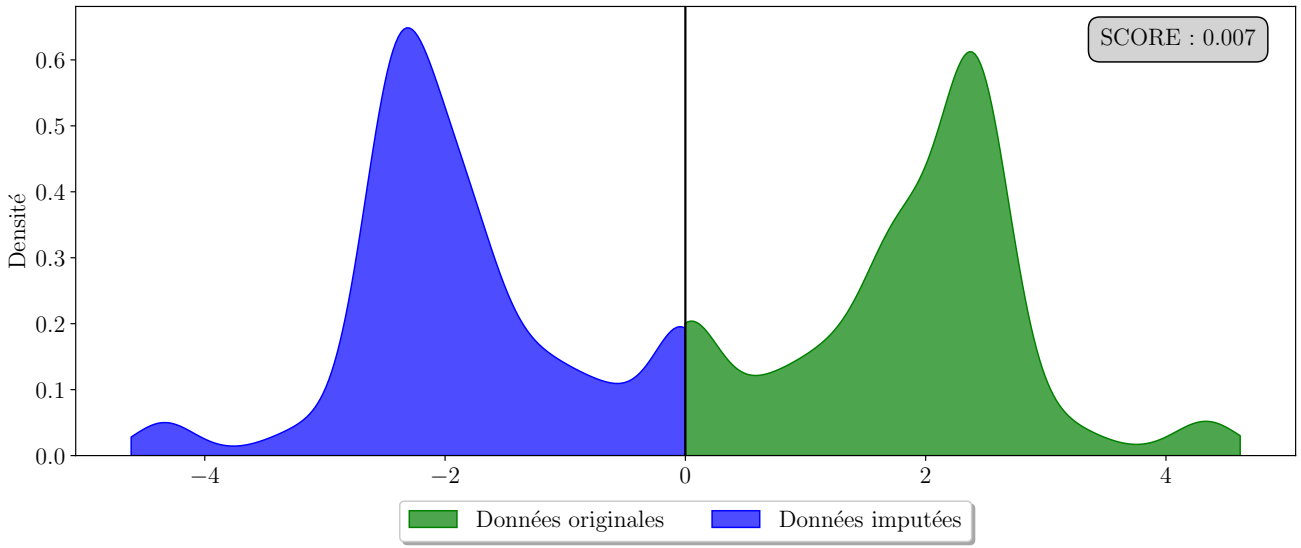
Distributions – Originales VS Moyenne



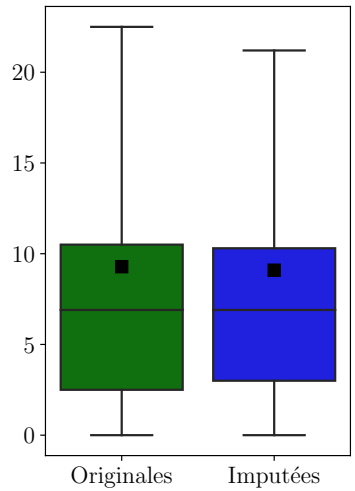
Boxplots



Distributions – Originales VS Médiane

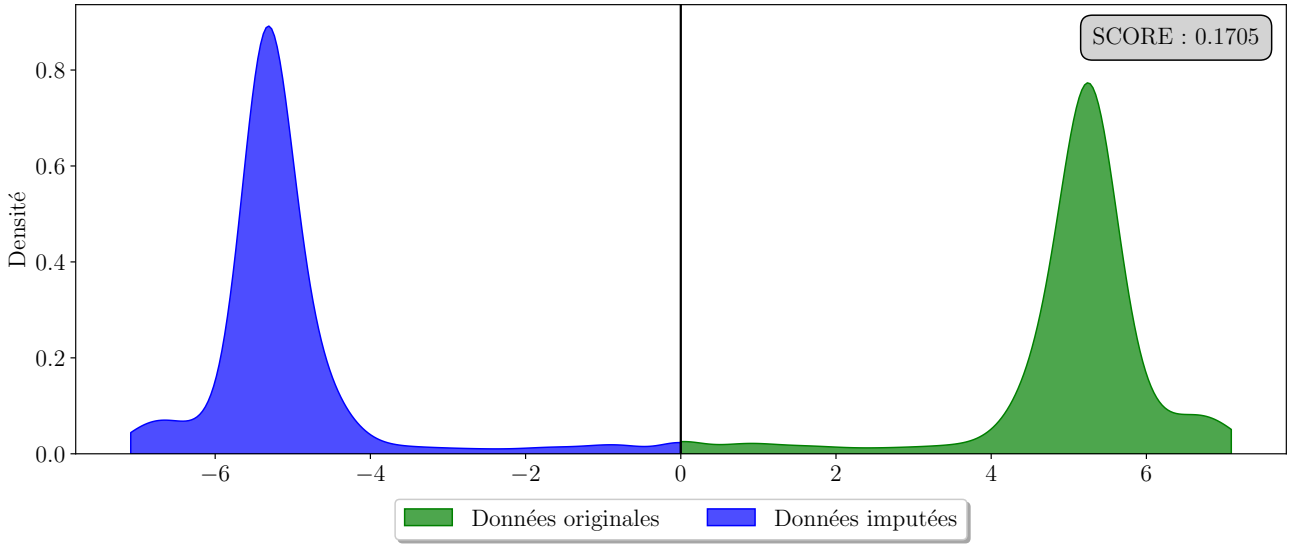


Boxplots

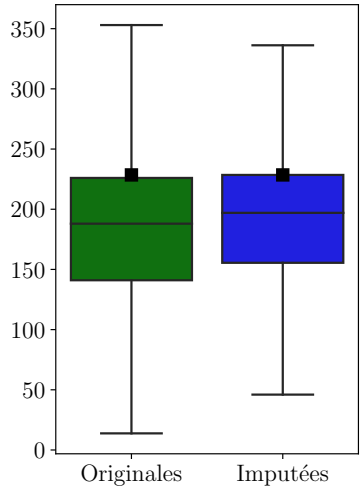


Energy – Imputations

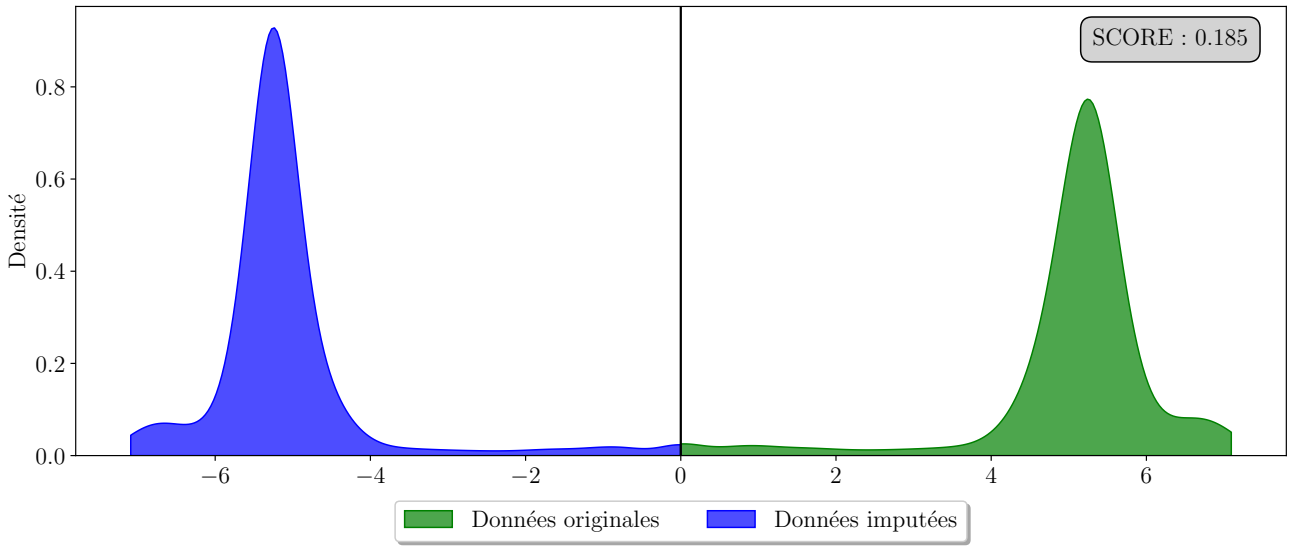
Distributions – Originales VS Moyenne



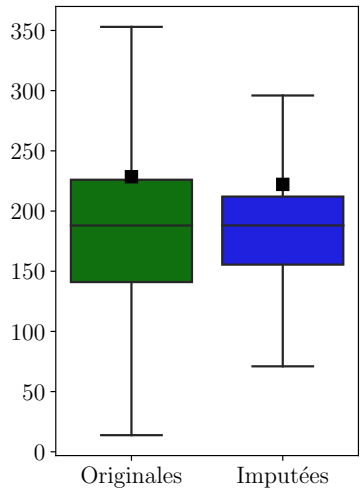
Boxplots



Distributions – Originales VS Médiane



Boxplots



Méthode des plus proches voisins

Avant d'imputer, on recherche le nombre de voisins optimal à l'aide de la technique du coude.

Exemple avec la variable `sugars_100g` de `data_food` :

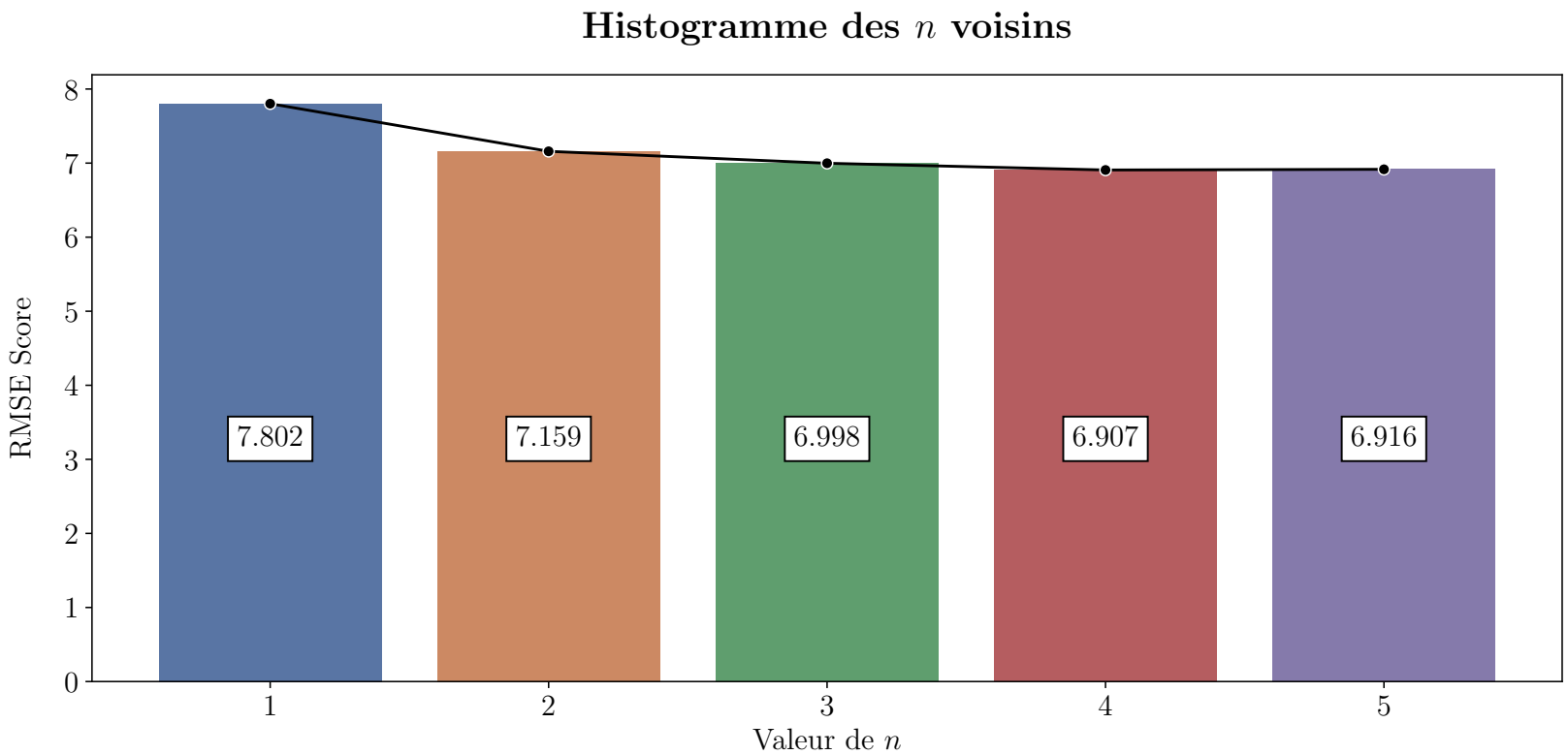
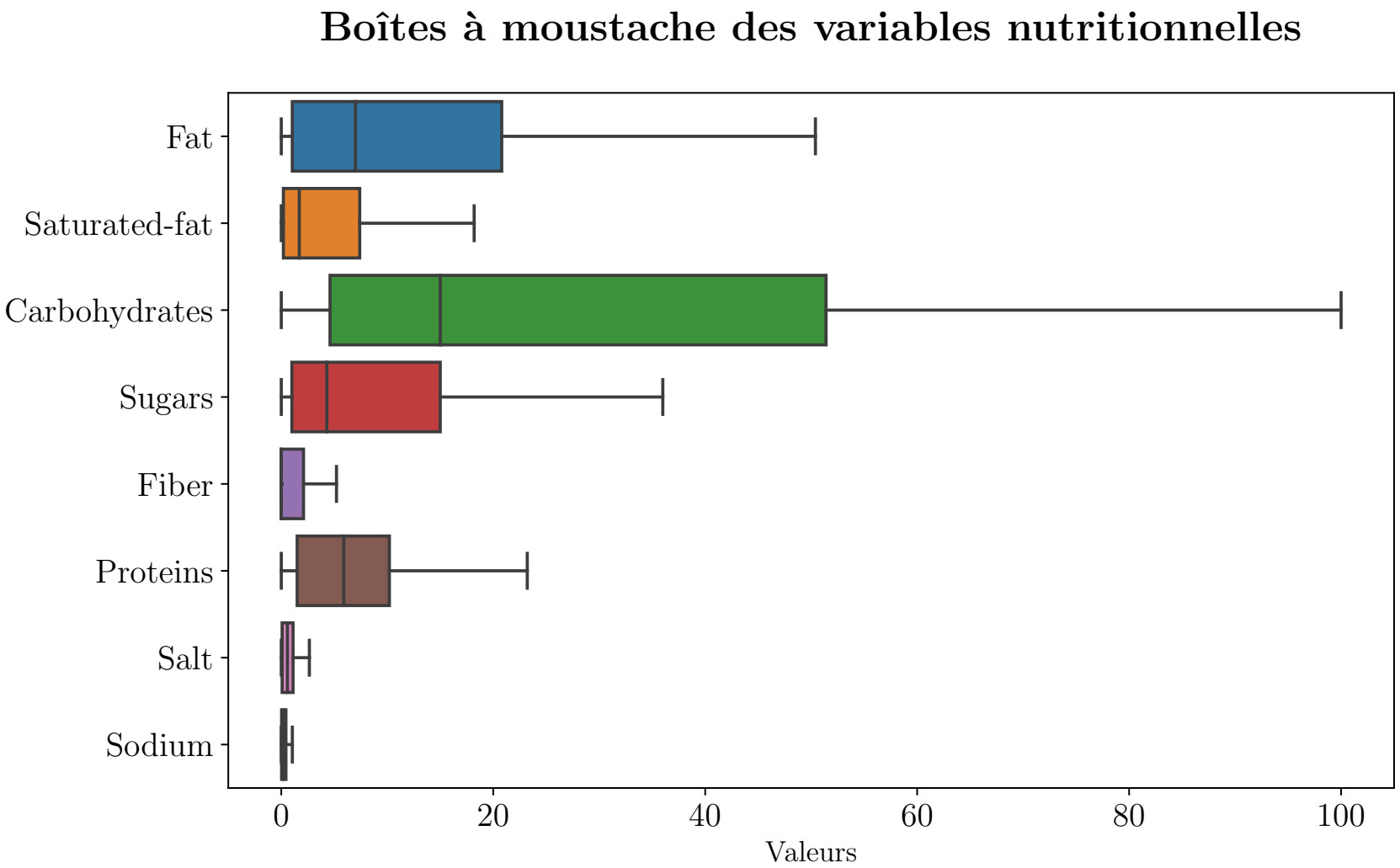


Tableau des imputations k—NN

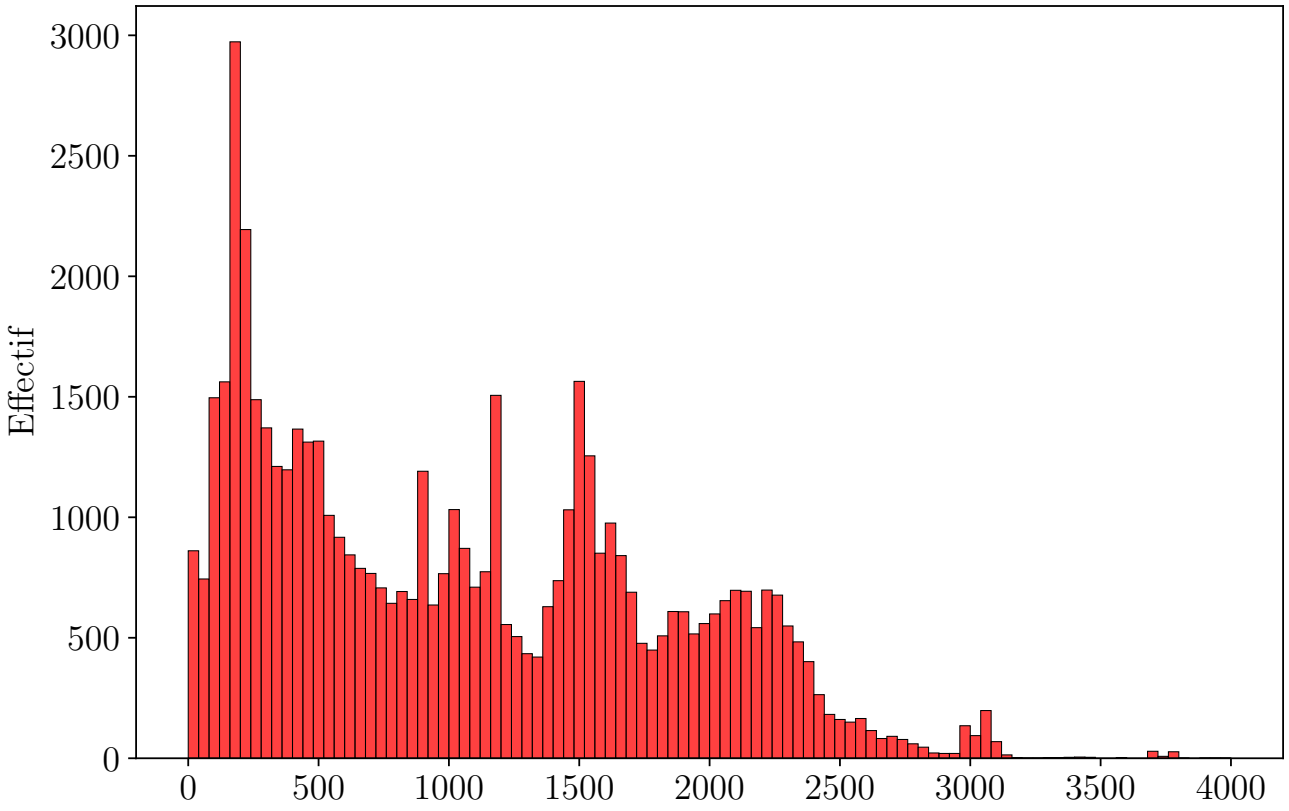
Dataframe	Variable	Valeur de k
data_food	saturated-fat_100g	3
	sugars_100g	2
	ns-fr_100g	2
	ns-uk_100g	2
data_drink	saturated-fat_100g	3
	energy_100g	4
	ns-fr_100g	4
	ns-uk_100g	4

ANALYSE DE DONNÉES

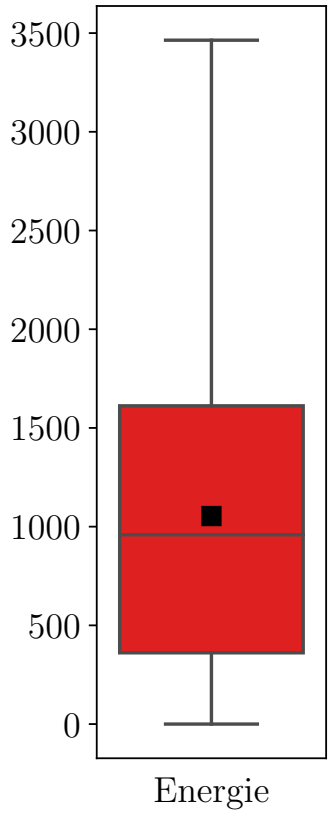
Analyse univariée

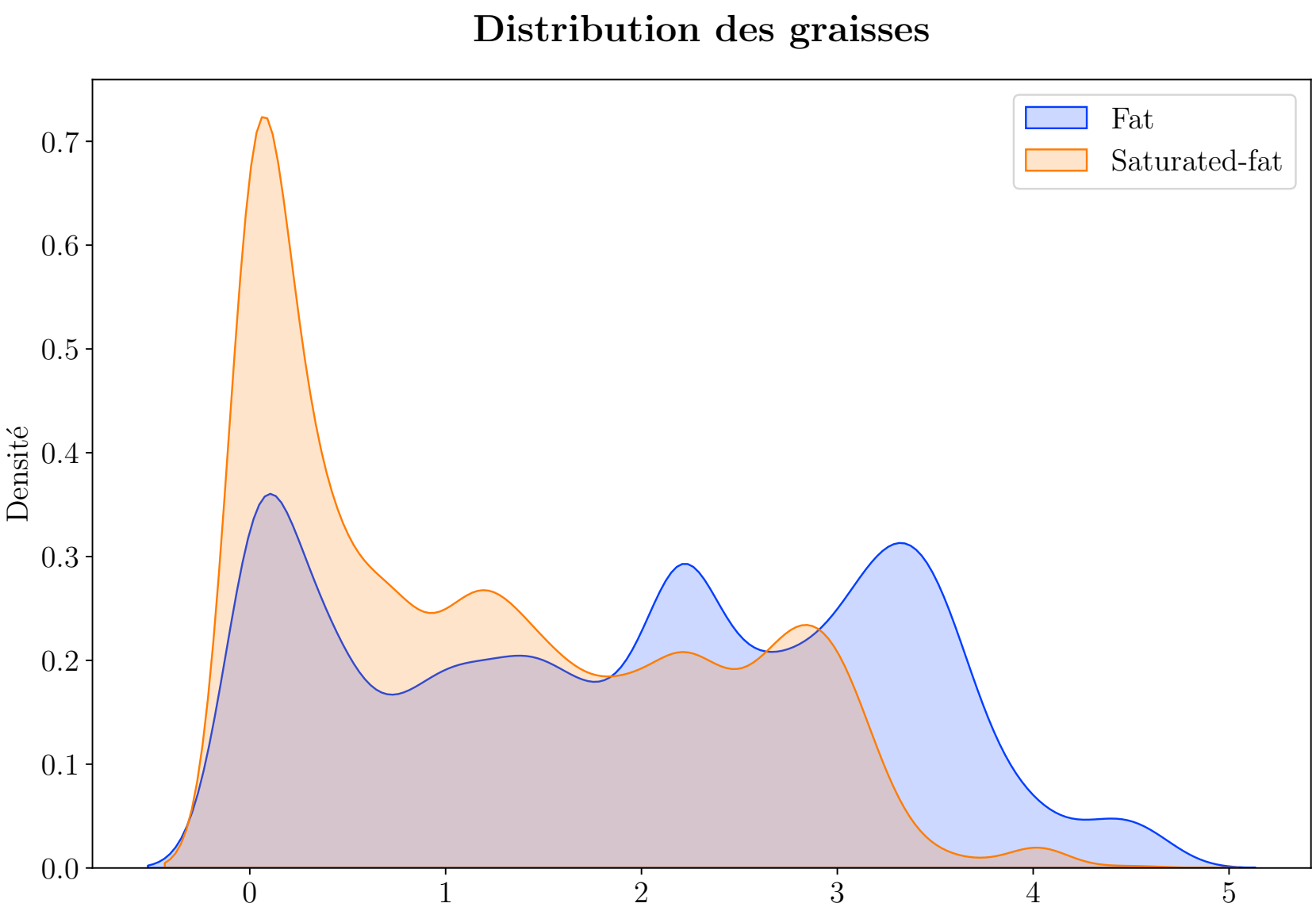


Distribution de l'énergie

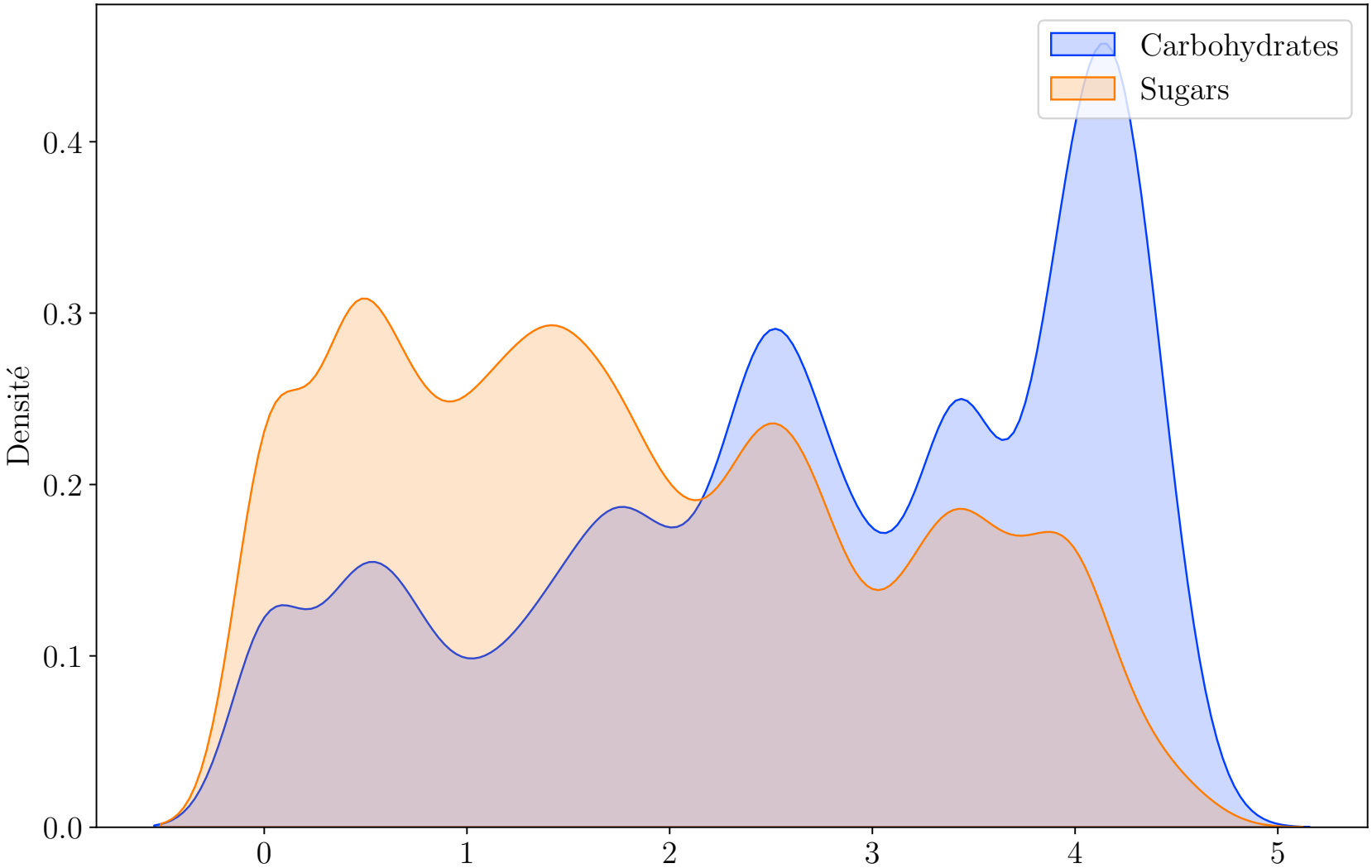


Boxplot

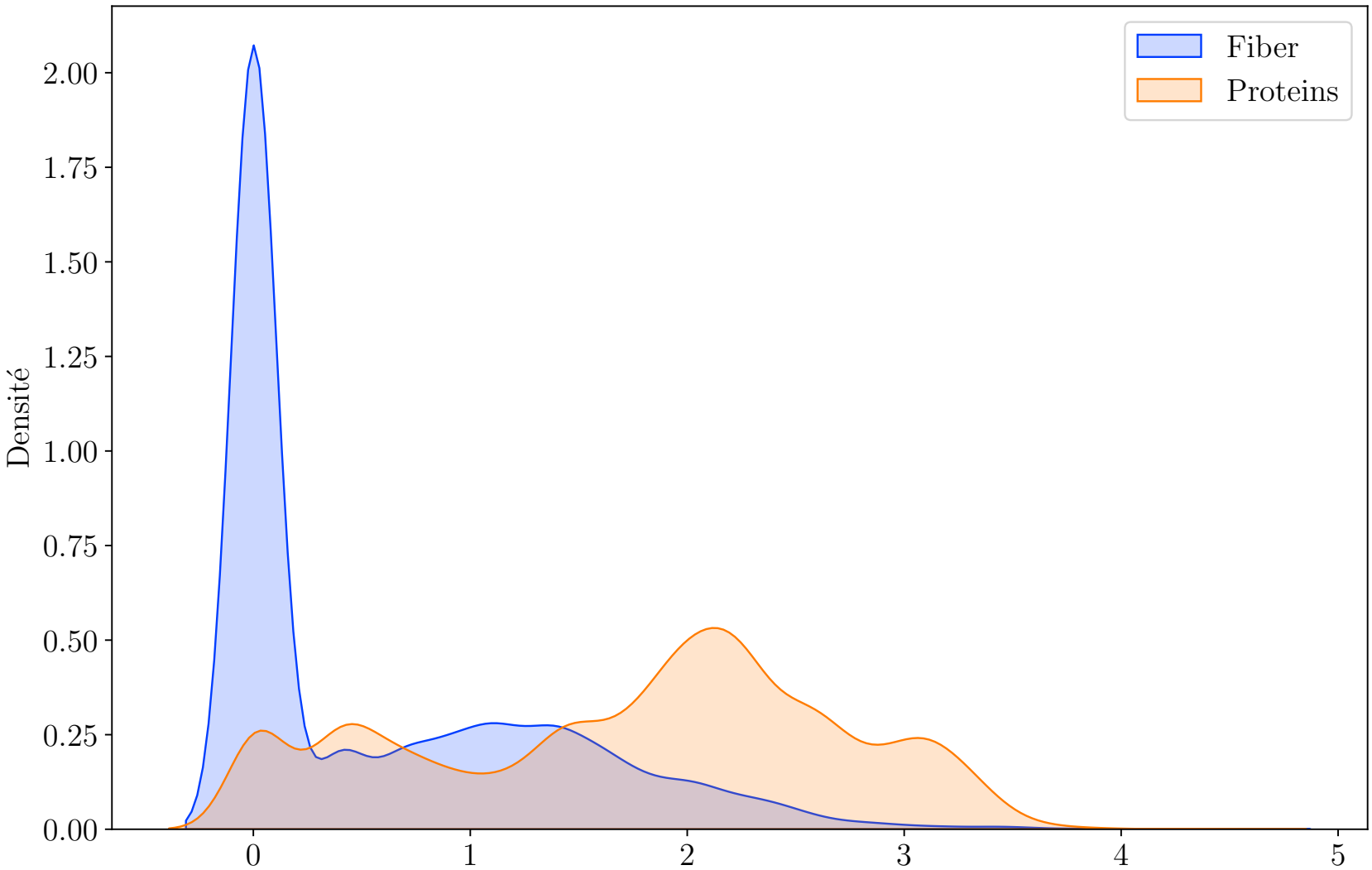



Distribution des graisses
Fat
Saturated-fat
Densité
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
4
5

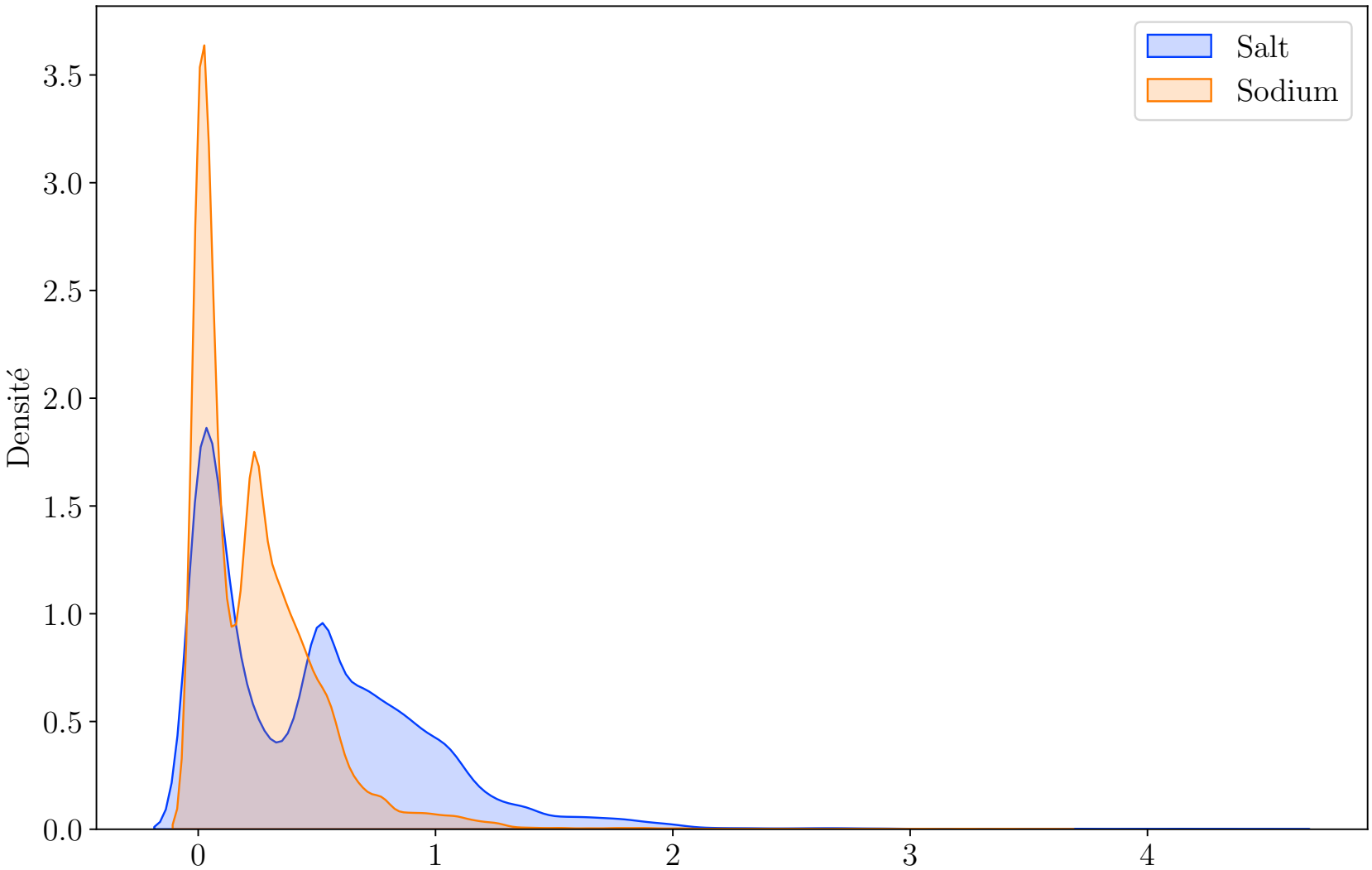
Distribution des glucides et sucres



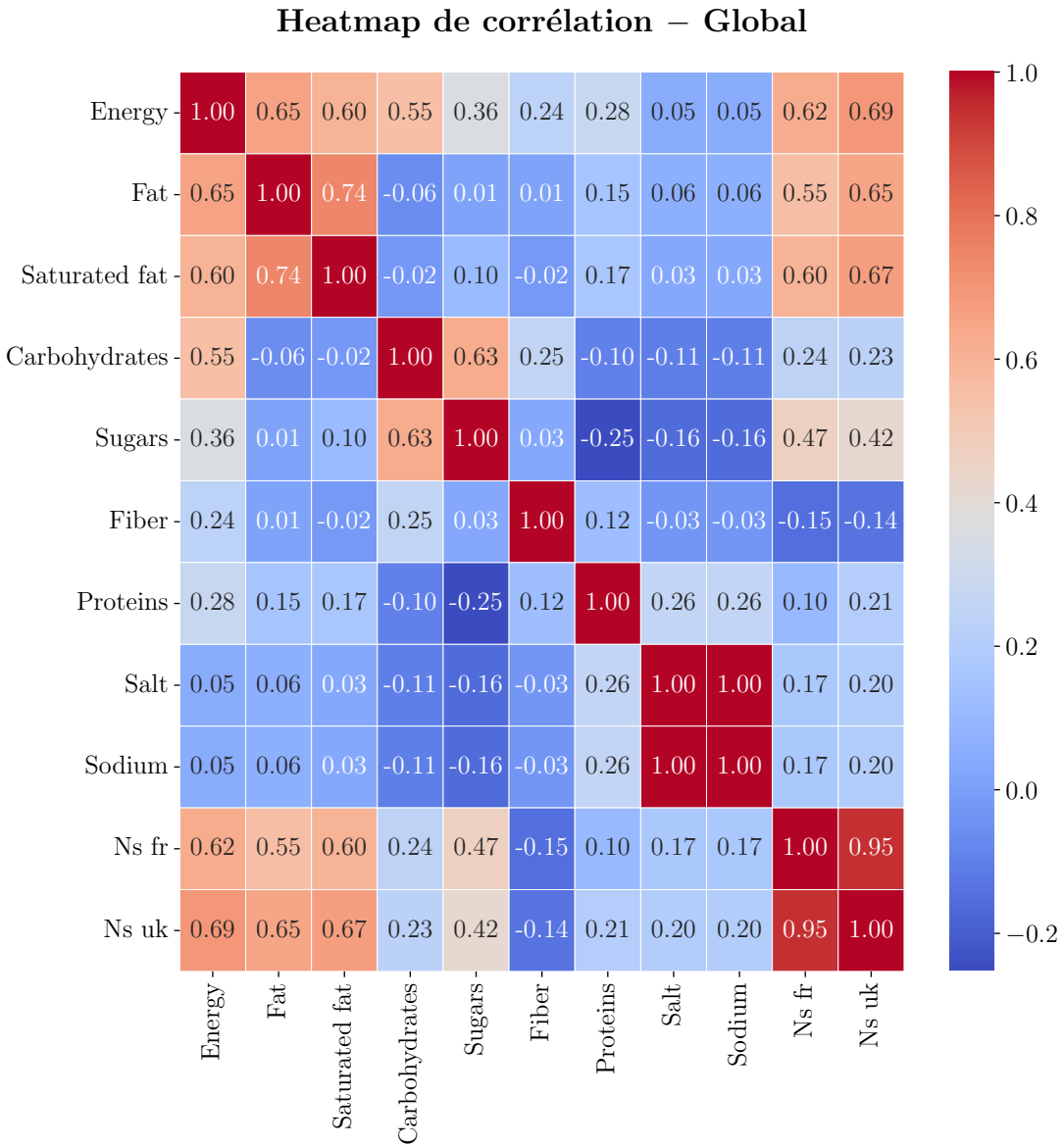
Distribution des fibres et protéines



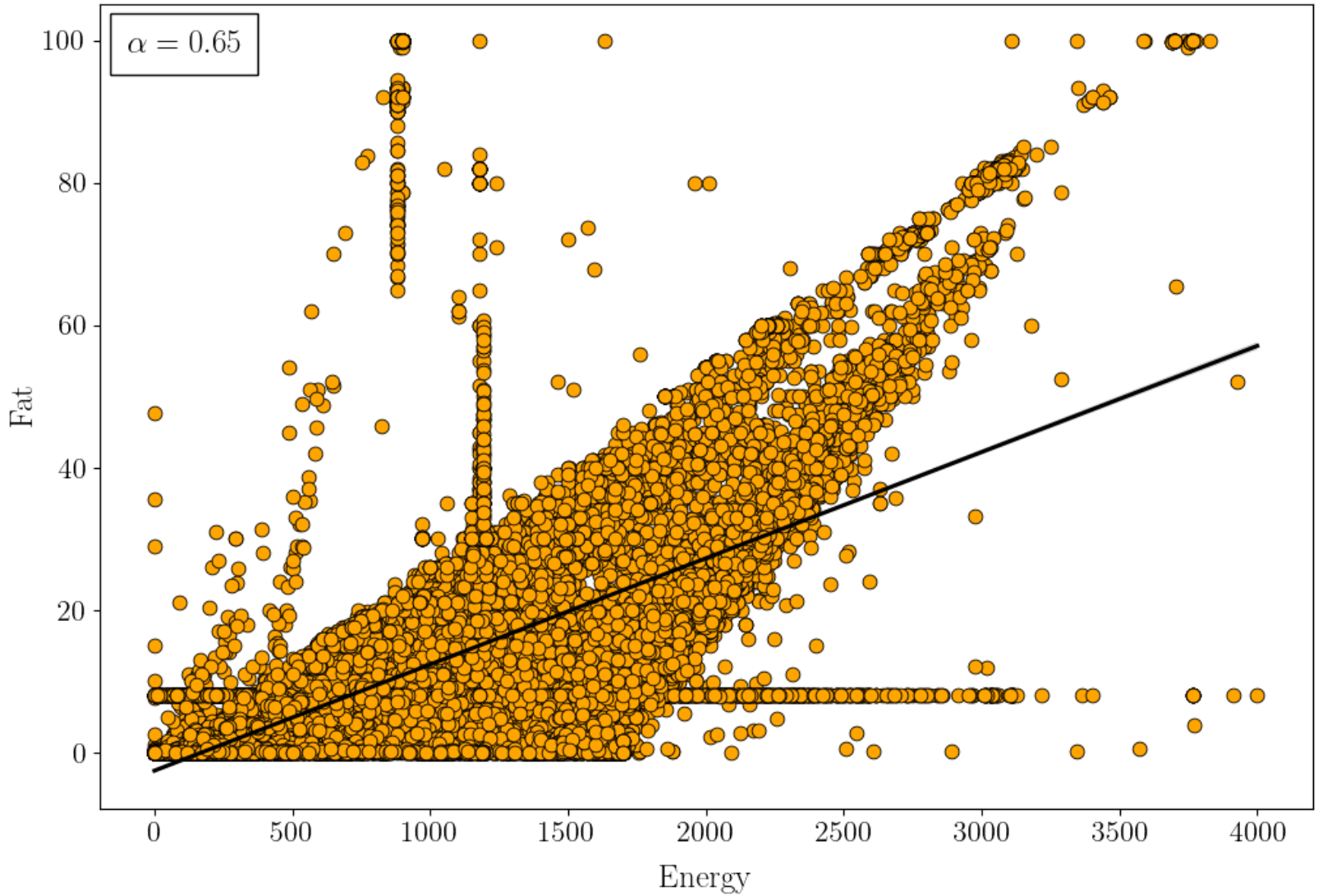
Distribution des sels et sodiums



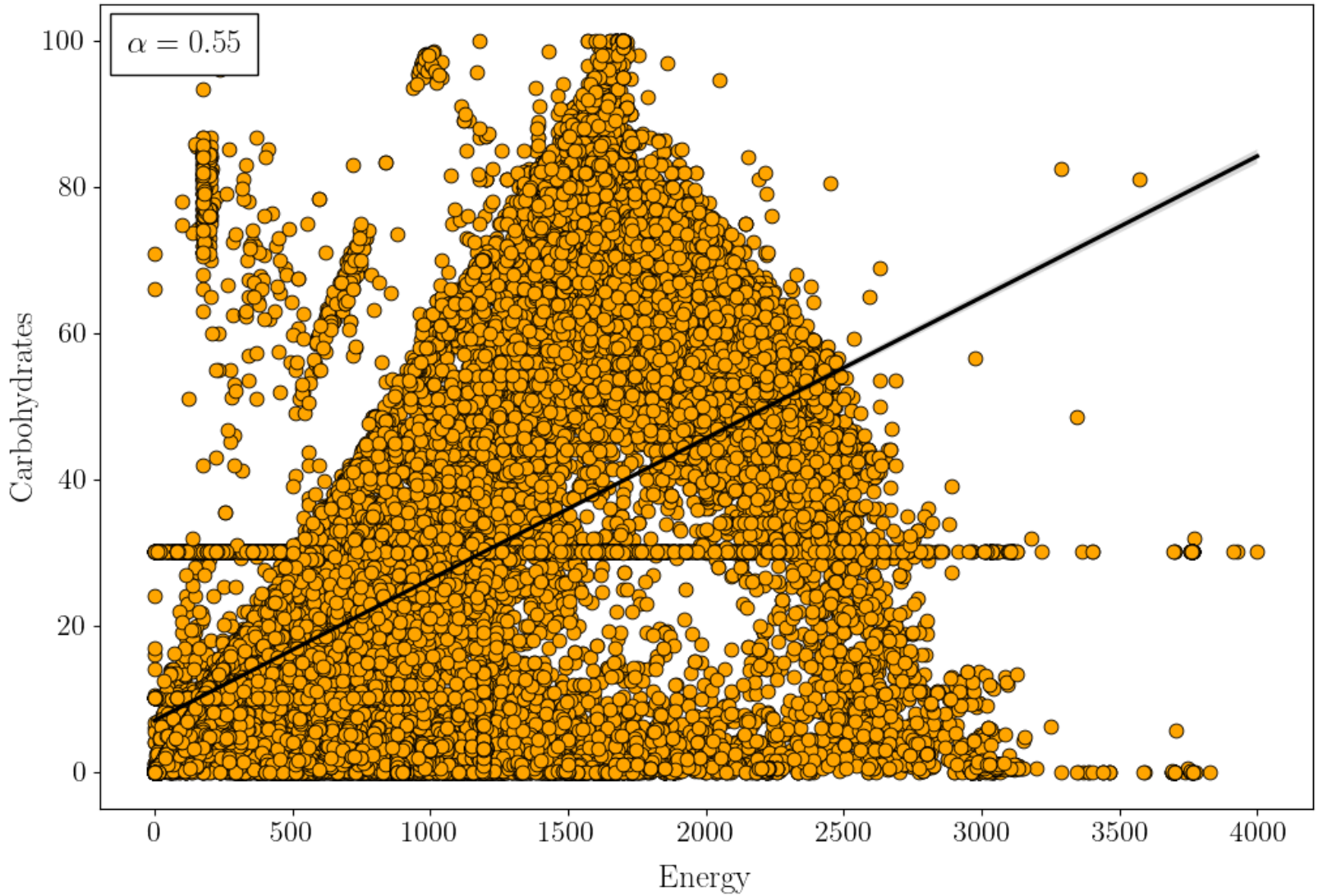
Analyse multivariée

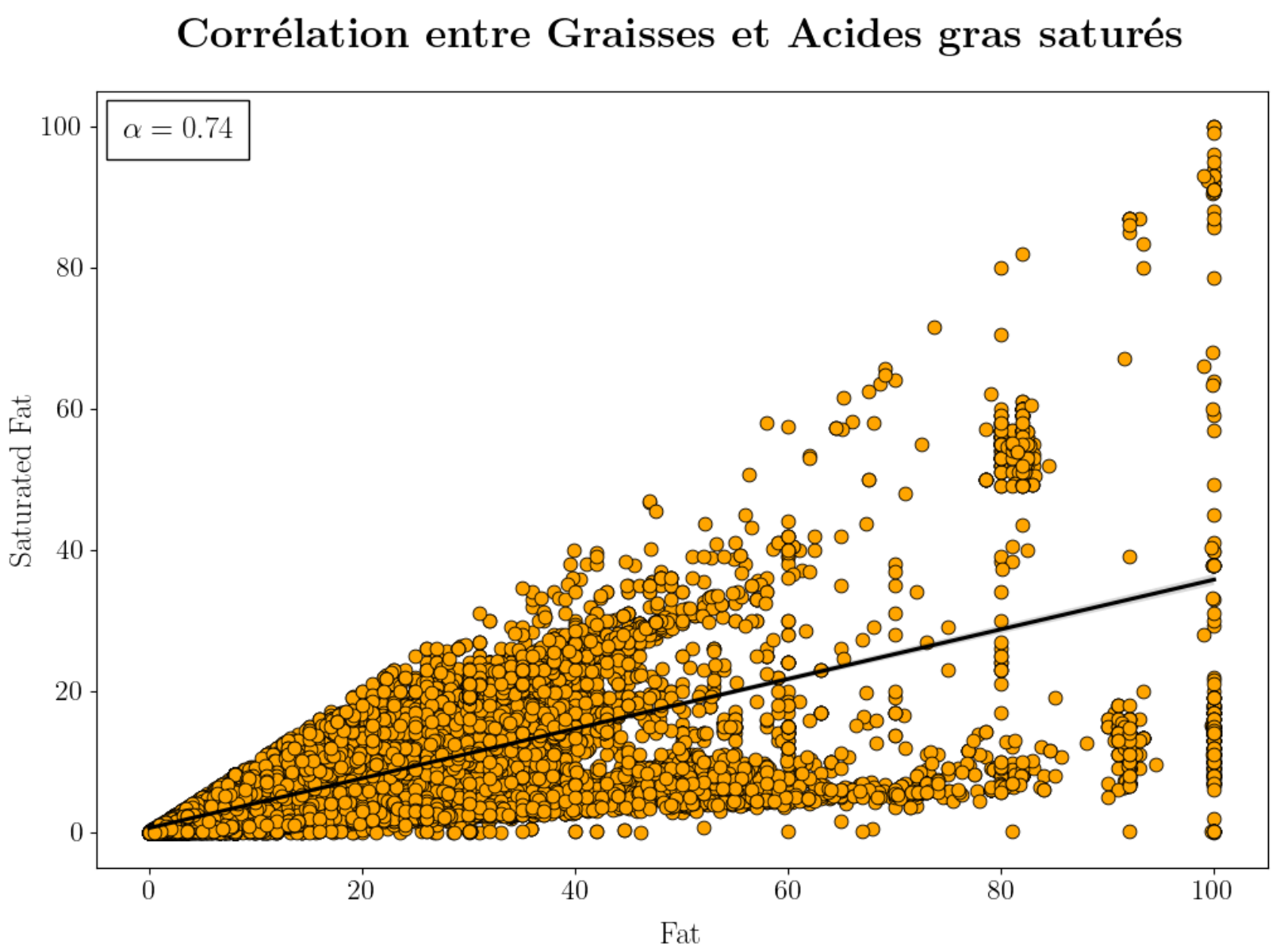


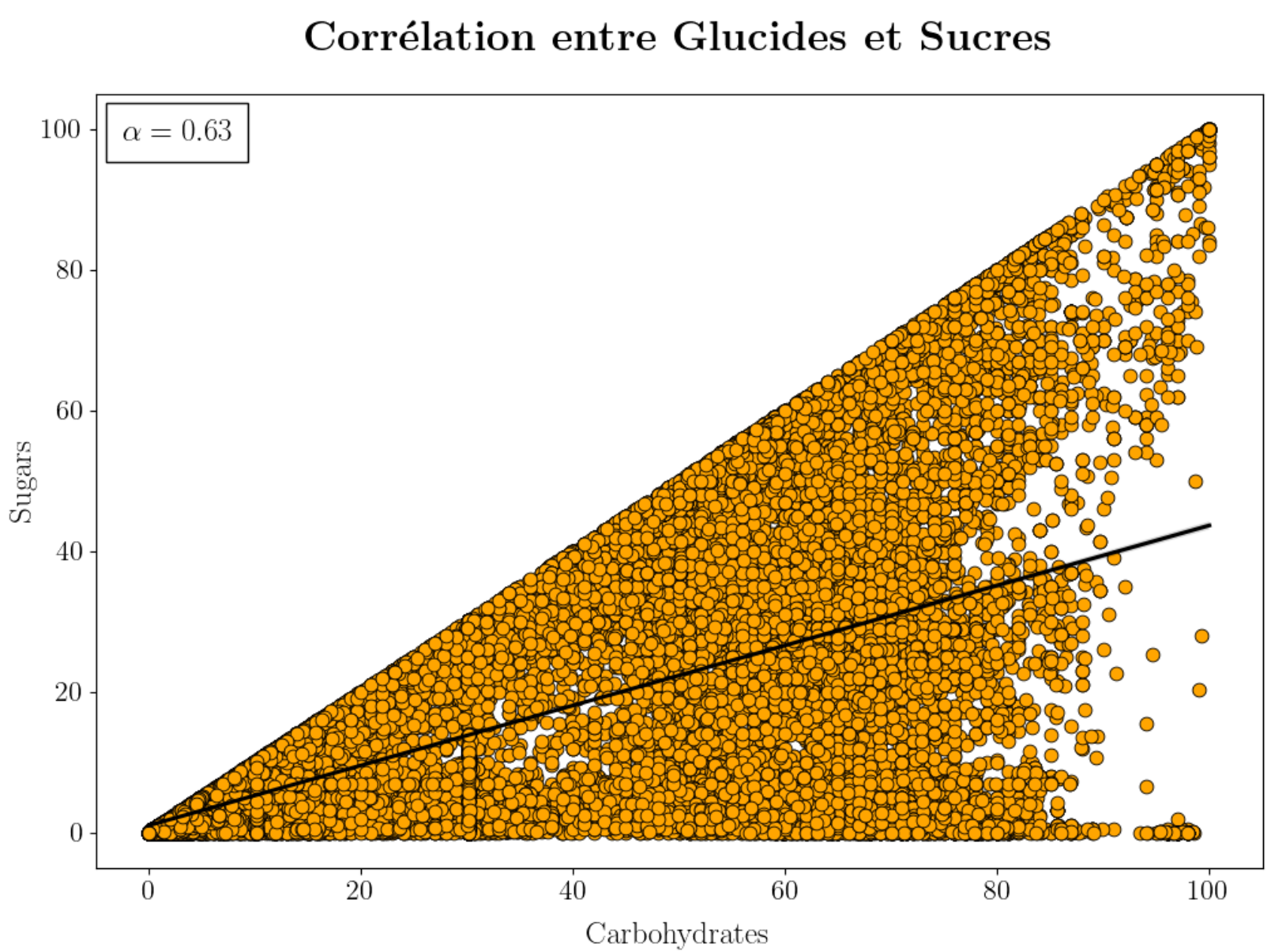
Corrélation entre Energie et Graisses

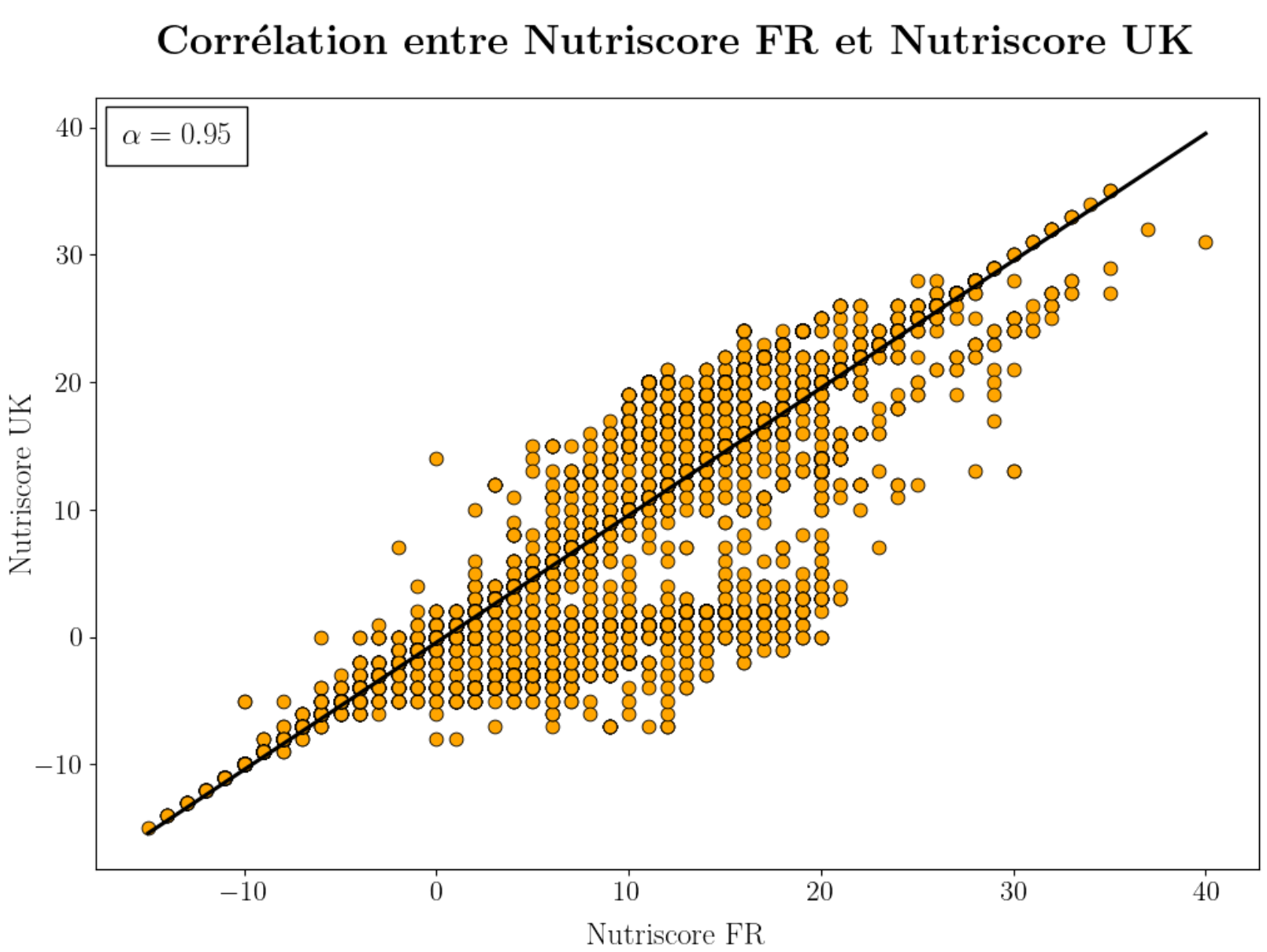


Corrélation entre Energie et Glucides




Corrélation entre Graisses et Acides gras saturés
$\alpha = 0.74$
Saturated Fat
Fat
0
20
40
60
80
100


Corrélation entre Glucides et Sucres
$\alpha = 0.63$
Sugars
Carbohydrates
0
20
40
60
80
100

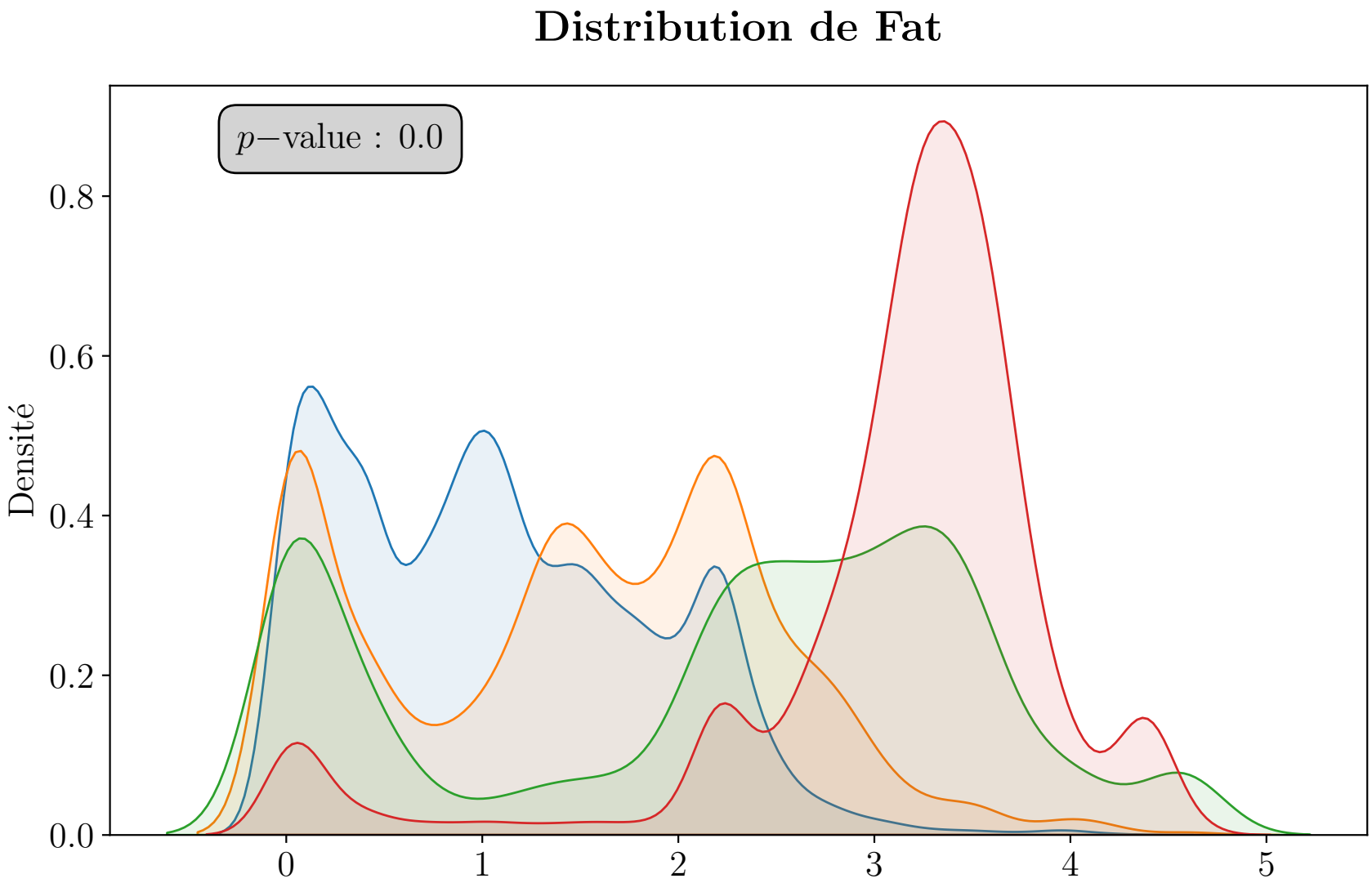

Corrélation entre Nutriscore FR et Nutriscore UK
$\alpha = 0.95$
Nutriscore UK
Nutriscore FR
40
30
20
10
0
−10
−10
0
10
20
30
40

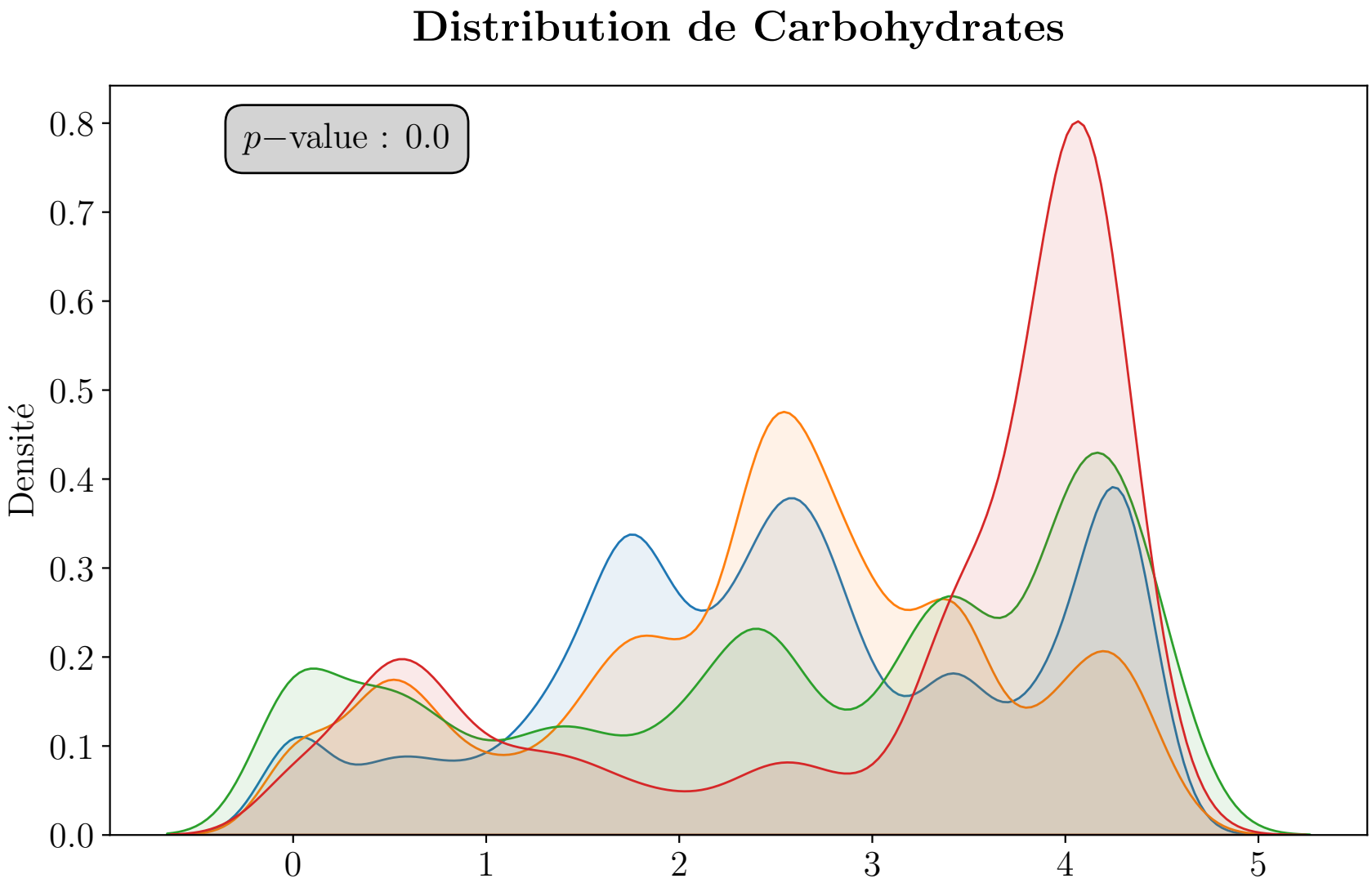
Test de Kruskal-Wallis

Vu que les distributions des variables ne sont pas normales, l'ANOVA n'est pas appropriée.
On effectue donc à la place un test de Kruskal-Wallis.

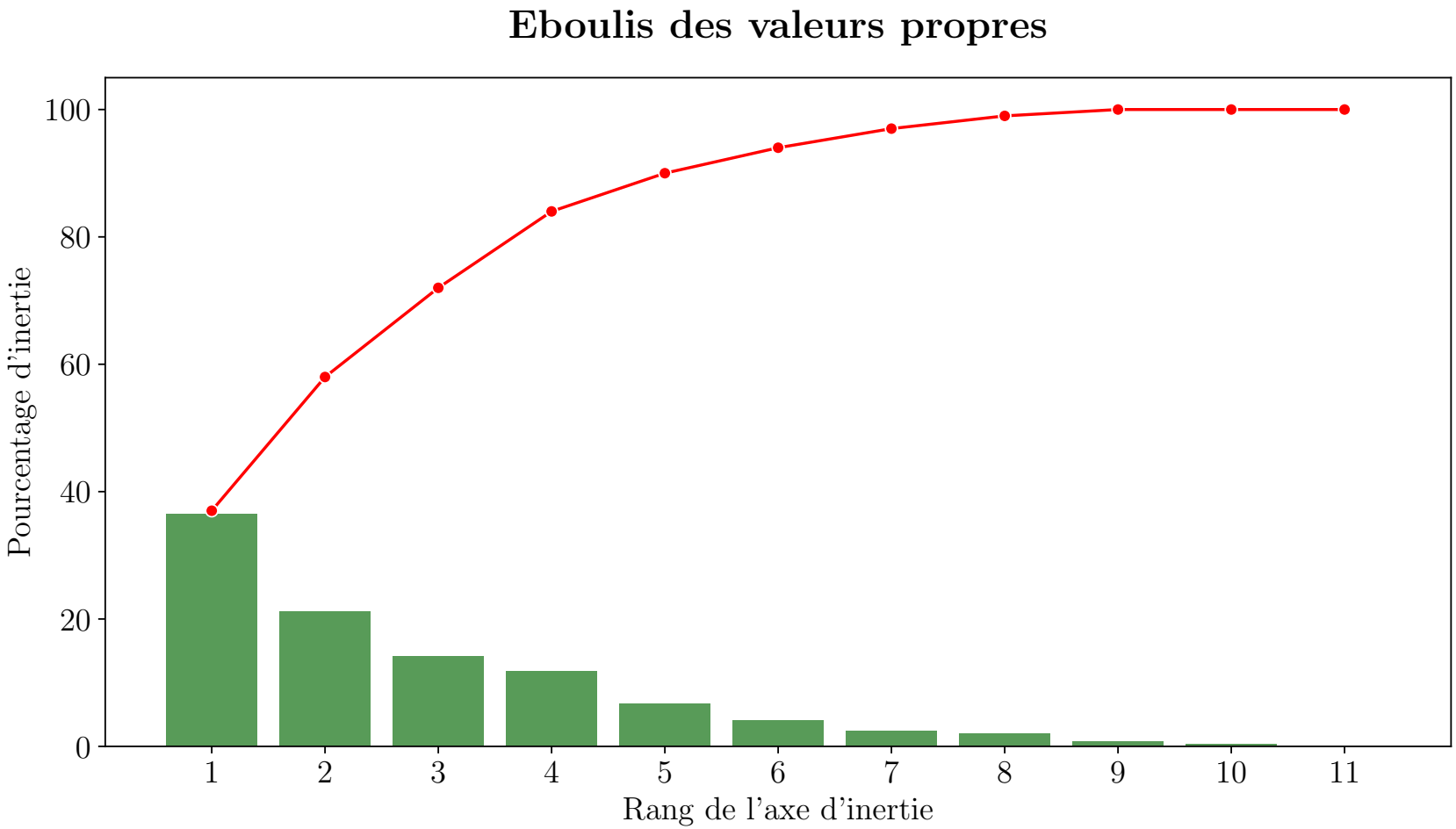
$\hookrightarrow$ On découpe la variable `ns-fr_100g` en 4 sous-groupes.

	Intervalle
Groupe 1	$(-15.001, 1.0]$
Groupe 2	$(1.0, 8.0]$
Groupe 3	$(8.0, 15.0]$
Groupe 4	$(15.0, 40.0]$


Distribution de Fat
$p-$value : 0.0
Densité
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5


Distribution de Carbohydrates
$p-$value : 0.0
Densité
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1
2
3
4
5

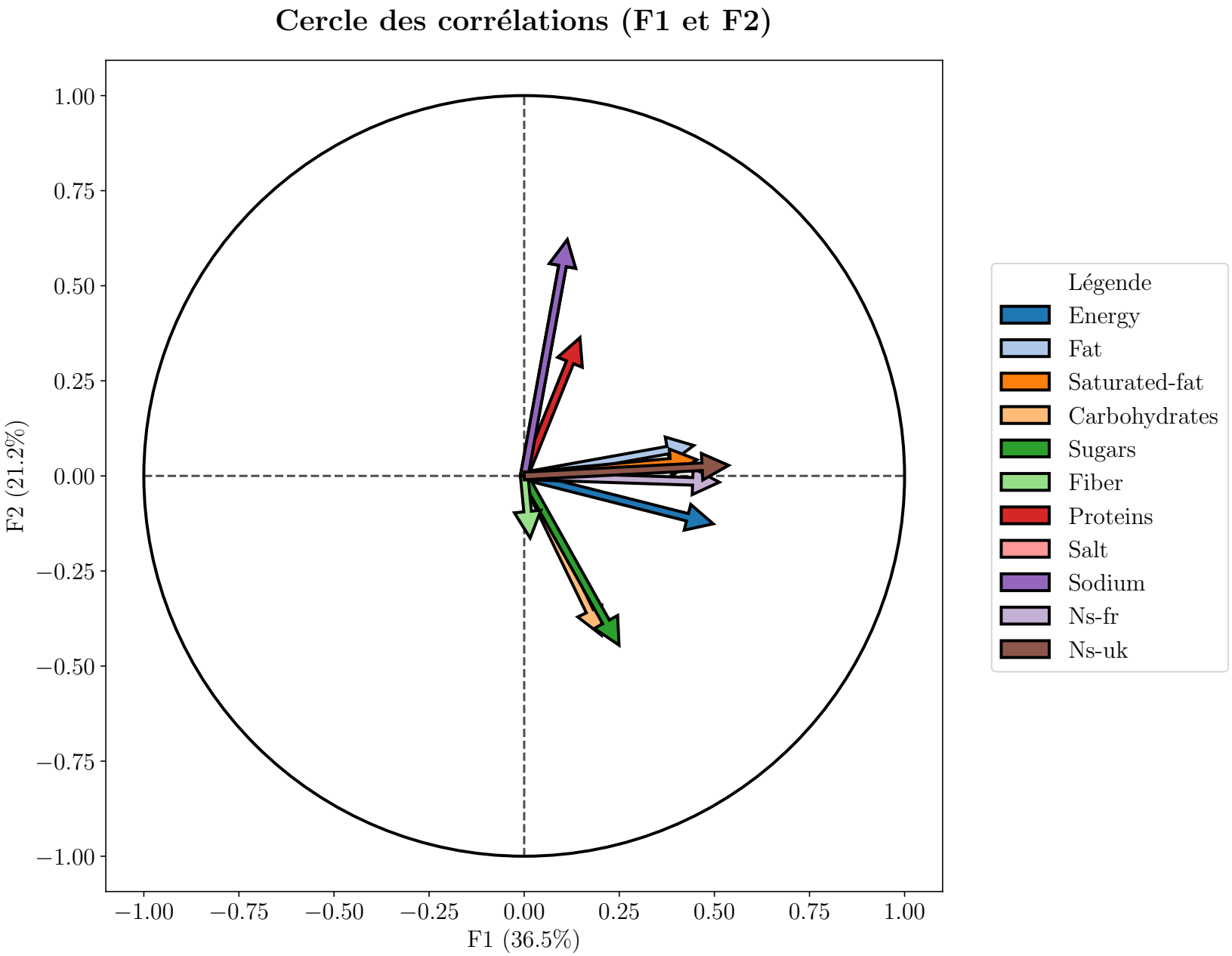
ANALYSE EN COMPOSANTES PRINCIPALES



Avec $F1$ et $F2$, on a 57, 7% de la variance expliquée.

Composants de l'ACP

	F1	F2	F3	F4
Energy	0, 42	−0, 11	0, 12	0, 28
Fat	0, 37	0, 07	−0, 35	0, 14
Saturated-fat	0, 39	0, 04	−0, 34	0, 09
Carbohydrates	0, 17	−0, 36	0, 53	0, 06
Sugars	0, 22	−0, 38	0, 35	−0, 29
Fiber	0, 01	−0, 09	0, 29	0, 68
Proteins	0, 12	0, 30	0, 00	0, 47
Salt	0, 10	0, 55	0, 36	−0, 13
Sodium	0, 10	0, 55	0, 36	−0, 13
Ns-fr	0, 45	−0, 01	−0, 02	−0, 25
Ns-uk	0, 47	0, 03	−0, 05	−0, 17


Cercle des corrélations (F1 et F2)
F2 (21.2%)
F1 (36.5%)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Légende
Energy
Fat
Saturated-fat
Carbohydrates
Sugars
Fiber
Proteins
Salt
Sodium
Ns-fr
Ns-uk

CONCLUSION

Conclusion

- Les valeurs aberrantes et manquantes ont été traitées.
- Les données ont été explorées et analysées de manière approfondie.

↪ Les données sont maintenant prêtes à être utilisées dans la création et l'entraînement d'un modèle de machine learning.

↪ Les traitements précédents permettront d'obtenir des résultats plus fiables et précis lors de la modélisation de la variable catégorielle cible `pnns_groups_1`

Règlement Général sur la Protection des Données

↪ Le RGPD est une réglementation européenne visant à renforcer la protection des données des citoyens de l'Union Européenne.

↪ Cette réglementation repose sur les 5 doctrines suivantes :

1 — Légalité, loyauté et transparence

2 — Limitation des finalités

3 — Minimisation des données collectées

4 — Exactitude des données collectées

5 — Limitation de la conservation

— FIN —