Santé publique France

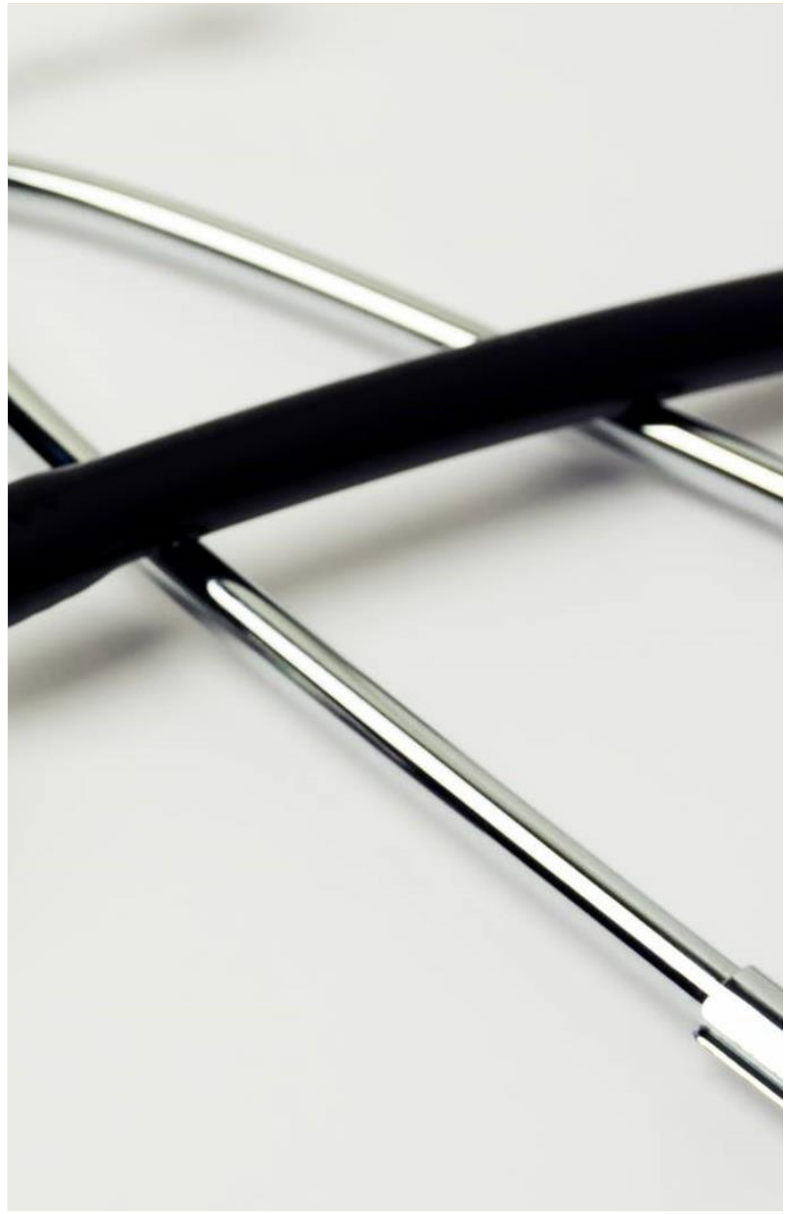
OPEN FOOD FACTS
l'information alimentaire ouverte

Préparation de données pour un organisme de santé publique
Projet n° 3

Formation : Data Scientist
Oumou Faye
Mentor : Medina Hadjem

AGENDA

1. INTRODUCTION
2. PROBLÉMATIQUE
3. ÉXPLORATION DES DONNÉES
4. NÉTTOYAGE
5. ANALYSE DES DONNÉES
6. SYNTHÈSE



I. INTRODUCTION

La base de données **Open Food Facts**

- La base de données **open-source Open Food Facts** est une base de données de **produits alimentaires.**
- Elle permet aux consommateurs **de connaître la qualité nutritionnelle** des produits grâce à **leurs fiches produit.**

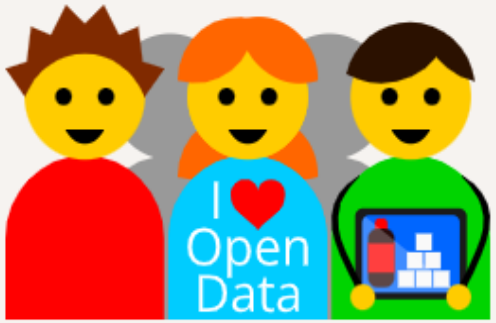
Base de données publiques

Une base de produits alimentaires

Open Food Facts est une base de données de produits alimentaires qui répertorie les ingrédients, les allergènes, la composition nutritionnelle et toutes les informations présentes sur les étiquettes des aliments.

OPEN FOOD FACTS
l'information alimentaire ouverte

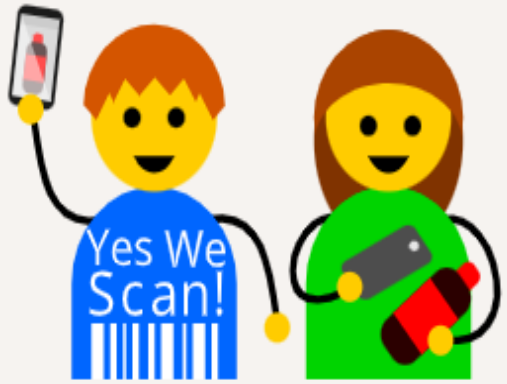
Création: 2012



Pour tout le monde

Les données sur la nourriture sont d'intérêt public et doivent être libres et ouvertes. Toute la base de données est publiée sous forme de données ouvertes (open data) qui peuvent être utilisées par tous et pour tous usages. Allez voir les réutilisations ou créez la vôtre !

Bénévoles et Contributeurs



Faite par tout le monde

Open Food Facts est une association à but non lucratif composée de volontaires.

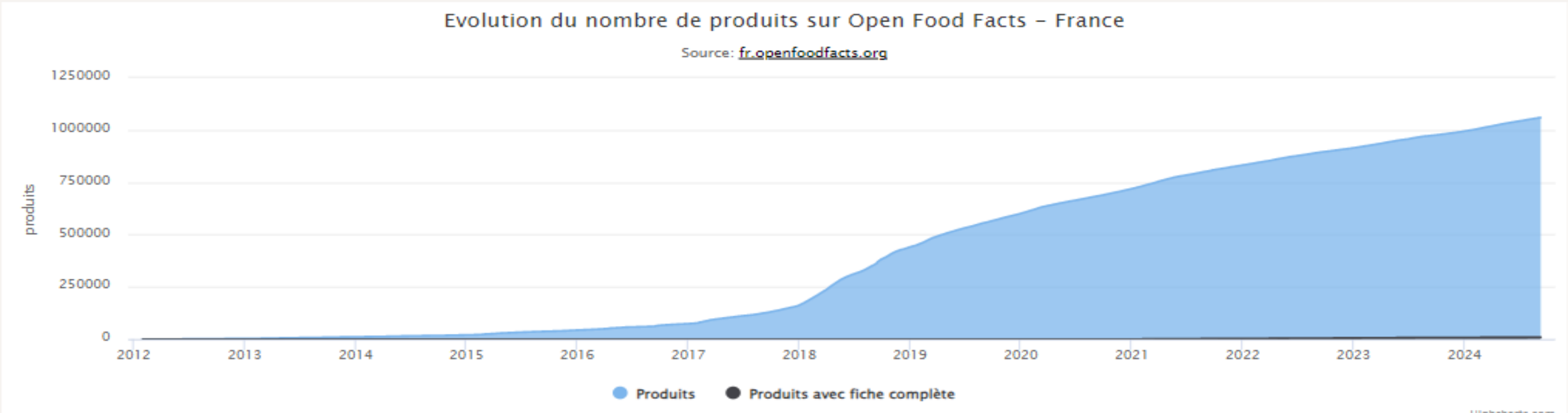
Plus de 9000 contributeurs comme vous ont ajouté 600 000 produits de 200 pays en utilisant notre app Android, iPhone ou Windows Phone ou leur appareil photo pour scanner les codes barres et envoyer des photos des produits et de leurs étiquettes.

Sur le site de la base de données **Open Food Facts,** Il y a la possibilité de :

lundi, Sep 9, 2024
Produits: 1 059 106
Produits avec fiche complète: 9 404

world.openfoodfacts.org

- Rechercher des **produits** en fonction de plusieurs critères et effectuer des comparaisons .
- Consulter des informations détaillées sur les **ingrédients** et **additifs** contenus dans les produits de la base de données.
- Ci-dessous un graphique montrant l'évolution du nombre de produits disponibles dans la base de données d'**Open Food Facts** depuis sa création en **2012** :



2. PROBLÉMATIQUE

HEALTH_Autofill

L' ajout d'un nouveau produit dans la base de données **Open Food Facts** requière :

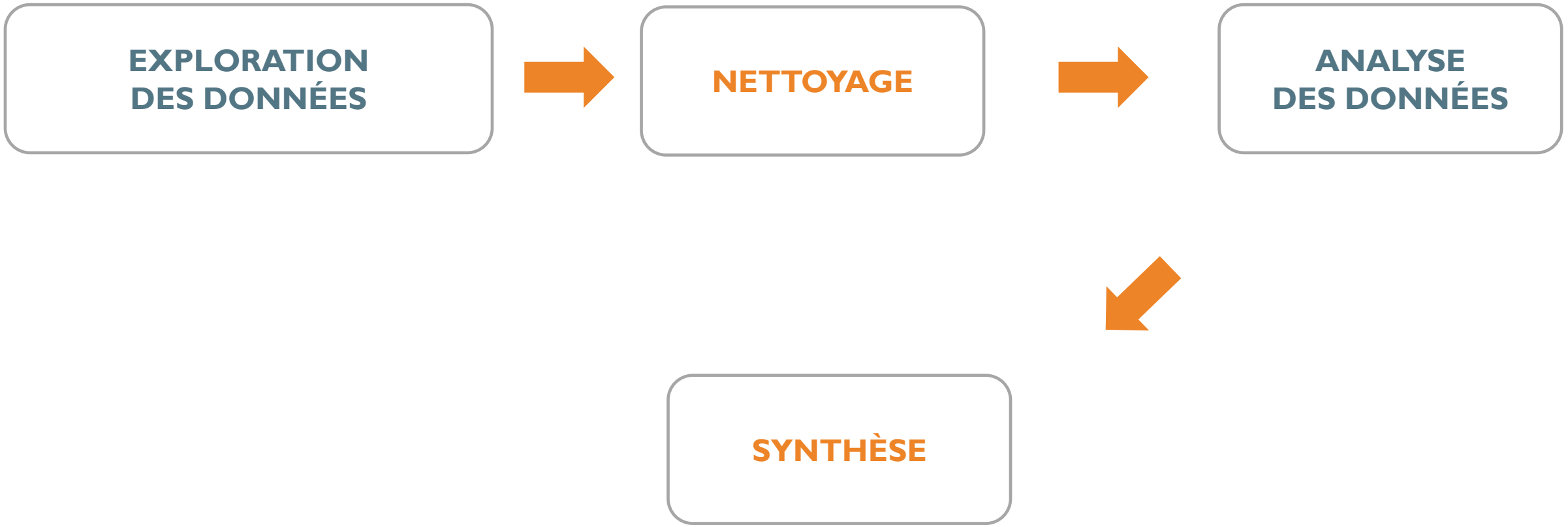
- **La saisie** de données :
  textuels et numériques

Santé publique France

Création d'une application dont l'objectif est afin de prédire les valeurs manquantes, notamment celles liées à l'information nutritionnelle d'un produit

/ Erreur de saisie potentielle
/ Valeurs manquantes potentielles

Structure de la gestion du projet



3. EXPLORATION DES DONNÉES

La base de données Open Food Facts contient : 320.772 lignes et 162 colonnes

**** Informations Générales ****
Contient des informations de base comme le code-barres, le nom du produit, les dates de création et de modification.

****Données Diverses****
Concernent des informations complémentaires comme la taille de la portion, les additifs et les ingrédients d'huile de palme.

Barcode:
3366321051983(EAN / EAN-13)

Common name: Matière grasse à tartiner et à cuire allégée (52% de MG), enrichie en vitamine B1

Quantity: 250 g

Packaging: Plastic, Tray

Brands: St Hubert, St hubert omega 3

Categories: Plant-based foods and beverages, Plant-based foods, Fats, Spreads, Plant-based spreads, Salted spreads, Spreadable fats, Vegetable fats, Margarines, Light margarines, Unsalted margarines, Light unsalted margarines, Plant-based pâtés, *50-63-unsalted-vegetable-fat-margarine-type-high-in-omega-3*

Labels, certifications, awards: Omega-3, Green Dot, Made in France, No palm oil, Nutriscore, Nutriscore Grade C, Triman

Origin of the product and/or its ingredients: Matière grasse à tartiner et à cuire allégée: France

Manufacturing or processing places: Ludres, 54710, Lorraine, France

Link to the product page on the official site of the producer: https://www.sthubert.fr/produit/st-huber...

Stores: Auchan, Leclerc, Magasins U, carrefour.fr

Countries where sold: France, Réunion, Switzerland

**** Ingrédients ****
Liste les ingrédients du produit et les traces d'allergènes possibles.

Informations Nutritionnelles	Pour 100 g	Portion (10 g)
Énergie	1887 kJ 459 kcal	189 kJ 46 kcal
Matières grasses, dont :	51 g	5,1 g
acides gras saturés	16 g	1,6 g
acides gras mono-insaturés	23 g	2,3 g
acides gras poly-insaturés	12 g	1,2 g
Sel	0,40 g	0,04 g
Vitamine E (% Apport de Référence)	11 mg (92%)	1,1 mg
Vitamine B1 (% Apport de Référence)	0,33 mg (30%)	0,03 mg

****Tags****
Regroupe les caractéristiques et classifications du produit, telles que les marques les catégories, et les lieux de fabrication

**** Informations Nutritionnelles ****
Détaille les éléments nutritifs incluant calories, protéines, graisses, vitamines, et minéraux

La base de données **Open Food Facts** contient :

Des variables
NUMÉRIQUES:
Elles correspondent à la typologie des **ingrédients** pour **100g**
(Exemple : **fructose_100g** / **lactose_100g**)

Des variables
CATÉGORIELLES:
Elles correspondent aux informations **textuelles** des produits, telles que les variables
(**categories_fr** / ou **nutrition_grade_fr**)

Le pourcentage de valeurs **manquantes dans l'ensemble des données de la base Open Food Facts est important**

Choix de la variable cible dans la base de données **Open Food Facts** :

La variable cible est ***nutrition_grade_fr,*** qui représente la note nutritionnelle attribuée à chaque produit.

A PROPOS DE LA VARIABLE “nutrition_grade_fr”:

- Variable **textuelle** (Object)

- **5 valeurs distinctes** :

(**A** : **Très bon sur le plan nutritionnel**. / **B** : **Bon**. / **C** : **Moyen**. / **D** : **Mauvais**. / **E** : **Très mauvais**.)

La variable `**nutrition_grade_fr**` est essentielle, car elle indique la qualité nutritionnelle d'un produit.
Elle aide les consommateurs à faire des choix **alimentaires plus sains, de manière simple et rapide**.

La base de données **Open Food Facts** :

Les variables numériques ci-dessous **dont le taux de valeurs manquantes** est **inférieur à 50%** feront l'objet de la prédiction de la variable cible : **nutrition_grade_fr**".

energy_100g :

Énergie en (**kilojoules**) pour 100 g. Utile pour identifier les produits énergétiques ou à faible teneur en calories. Valeurs manquantes : **18.60%**.

fat_100g :

Graisses totales pour 100 g. Indique les produits riches en graisses comme les snacks ou les produits frits. Valeurs manquantes : **23.97%**.

saturated-fat_100g :

Graisses saturées pour 100 g. Souvent présentes dans les viandes grasses, pâtisseries, ou produits laitiers entiers. Valeurs manquantes : **28.44%**.

carbohydrates_100g :

Glucides totaux pour 100 g. Clé pour identifier les produits sucrés, céréales, et produits de boulangerie. Valeurs manquantes : **24.06%**.

sugars_100g :

Sucres pour 100 g. Essentiel pour catégoriser les confiseries, boissons sucrées, ou desserts. Valeurs manquantes : **23.63%**.

fiber_100g :

Fibres pour 100 g. Indique les produits enrichis ou à base de grains entiers, souvent associés à une alimentation saine. Valeurs manquantes : **37.37%**.

proteins_100g :

Protéines pour 100 g. Présentes dans les viandes, produits laitiers, légumineuses, et substituts de viande. Valeurs manquantes : **18.97%**.

salt_100g :

Sel pour 100 g. Indicateur de la teneur en sodium, important pour évaluer les plats cuisinés et conserves. Valeurs manquantes : **20.35%**.

VARIABLE CIBLE A PRÉDIRE : nutrition_grade_fr

4. NETTOYAGE

LES ETAPES DU NETTOYAGE DES VARIABLES CIBLES:

La création du fichier tabulaire contenant uniquement les **variables explicatives** :

LES VALEURS ABERRANTES

1. **Détection** des valeurs aberrantes (outliers) grâce à la méthode **IQR** et à la visualisation d'un **box plot** par variable.

2. **Suppression** des valeurs aberrantes de la base de données **Open Food Facts.**

LES VALEURS MANQUANTES

1. **Détection** des valeurs manquantes pour chaque variable.

2. **Traitement** des valeurs manquantes pour chaque variable en privilégiant l'approche **métier.**

CREATION DU DATAFRAME CONTENANT LES VARIABLES CIBLES

Ces variables explicatives seront utilisées pour entraîner un modèle de prédiction du Nutri-score.

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2243.0	28.57	28.57	64.29	14.29	3.6	3.57	0.00000
2	1941.0	17.86	0.00	60.71	17.86	7.1	17.86	0.63500
3	2540.0	57.14	5.36	17.86	3.57	7.1	17.86	1.22428
4	1552.0	1.43	NaN	77.14	NaN	5.7	8.57	NaN

Les colonnes affichées représentent les caractéristiques nutritionnelles pour 100 g. Les valeurs **NaN** indiquent des données manquantes à traiter.

D) VISUALISATION DU JEU DE DONNÉES AVANT LE NETTOYAGE

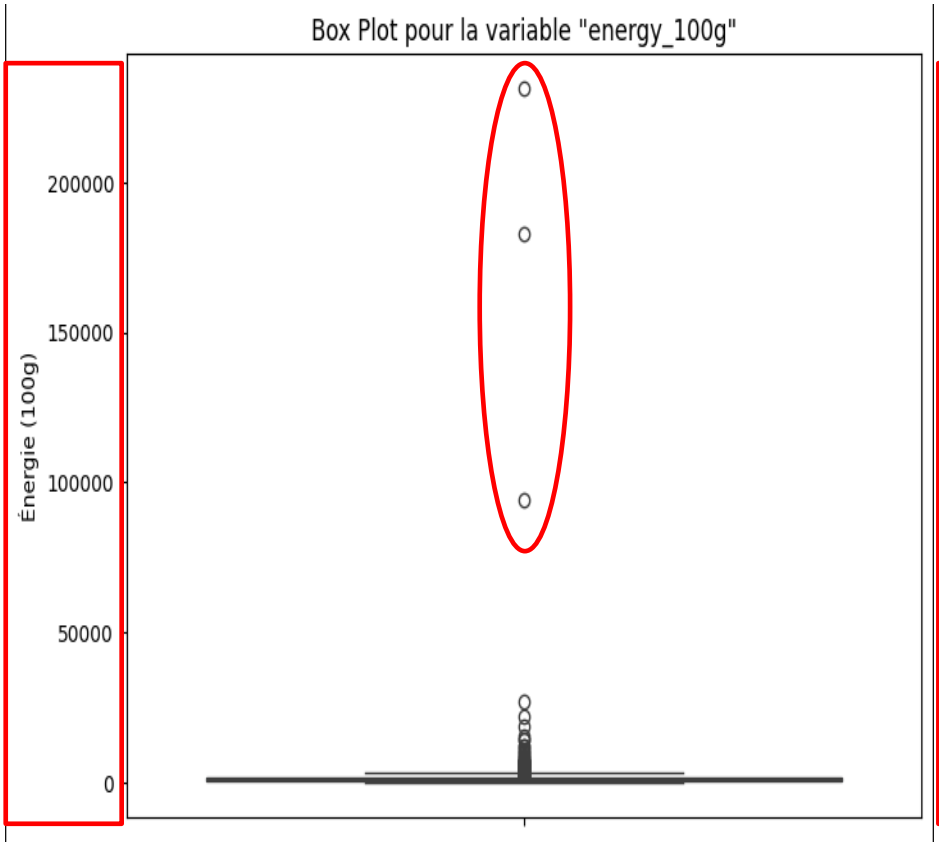
Statistiques Descriptives:

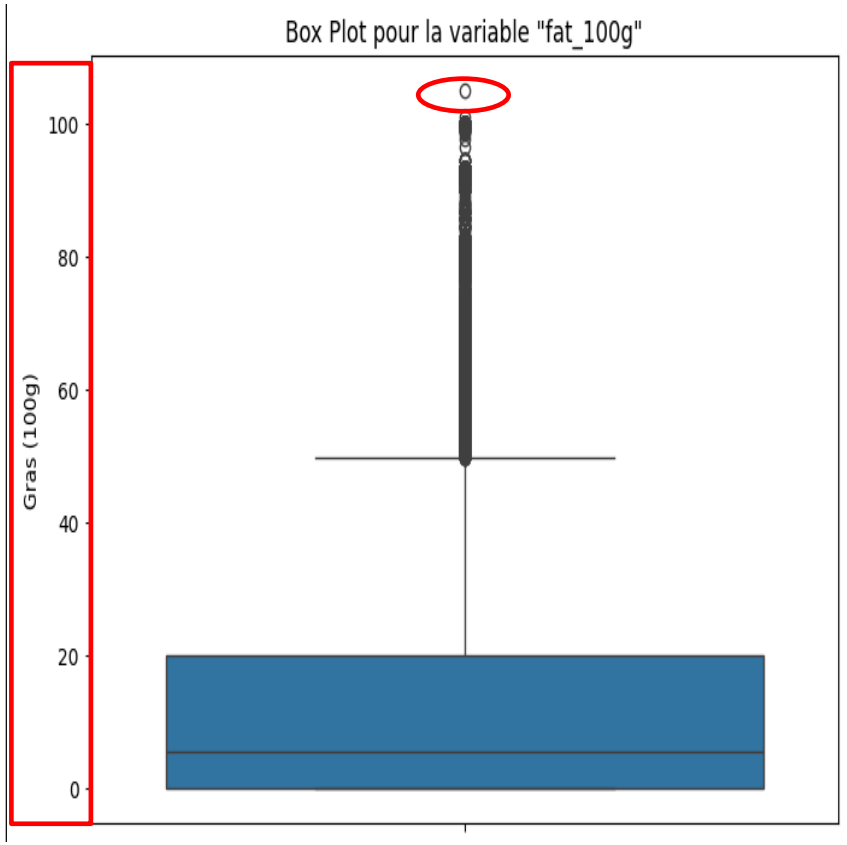
	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	\
count	184470.000000	169999.000000	162921.000000	169730.000000	
mean	1139.314255	12.497010	5.195458	33.116896	
std	1066.902948	16.476978	8.014654	29.981925	
min	0.000000	0.000000	0.000000	0.000000	
25%	418.000000	0.100000	0.000000	6.670000	
50%	1117.000000	5.630000	1.900000	23.080000	
75%	1674.000000	20.000000	7.140000	60.000000	
max	231199.000000	380.000000	550.000000	2916.670000	

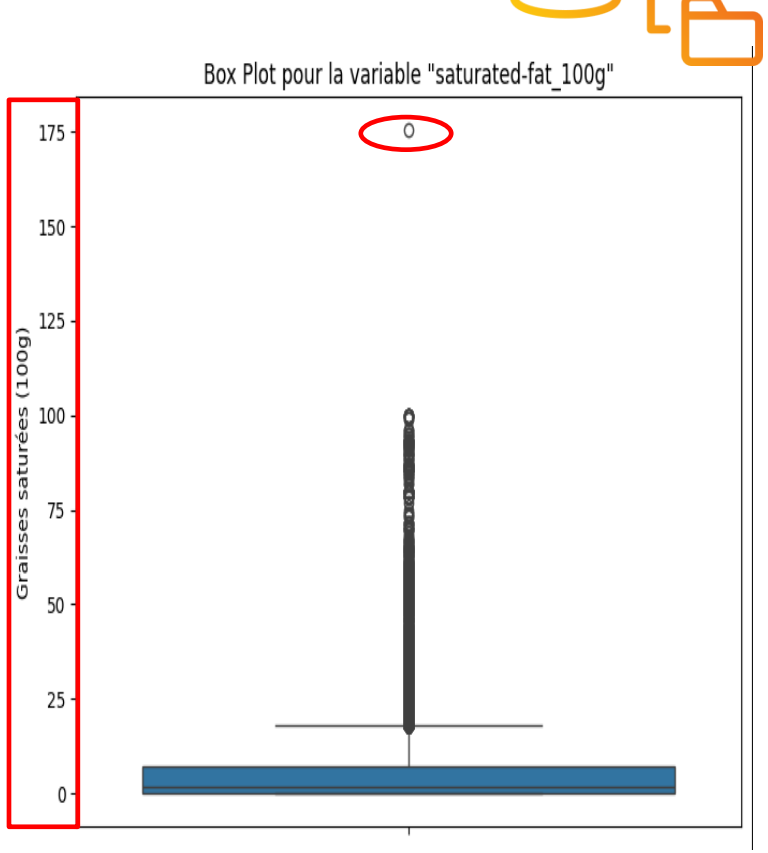
	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	173677.000000	141981.000000	183560.000000	180138.000000
mean	16.329856	2.898030	7.139842	2.080937
std	22.700836	15.016186	8.417052	152.389088
min	-6.250000	-6.700000	-800.000000	0.000000
25%	1.420000	0.000000	0.820000	0.071120
50%	6.190000	1.500000	5.000000	0.584200
75%	25.000000	3.600000	10.000000	1.361440
max	3520.000000	5380.000000	430.000000	64312.800000

1. Présence de valeurs **> 100g ou < 0** : incohérences nutritionnelles (mathématiquement impossibles)
2. Présence de valeurs **> 9000 kilojoules** dans la variable energy_100g : dépassement des seuils physiologiquement plausibles

LES VALEURS ABERRANTES DES VARIABLES CIBLES ON OBSERVE VISUELLEMENT LA PRESENCE D'OUTLIERS





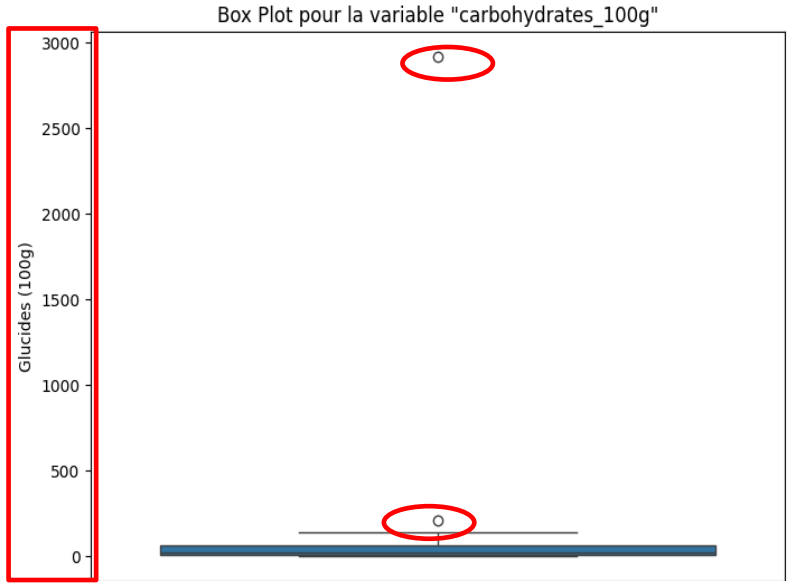


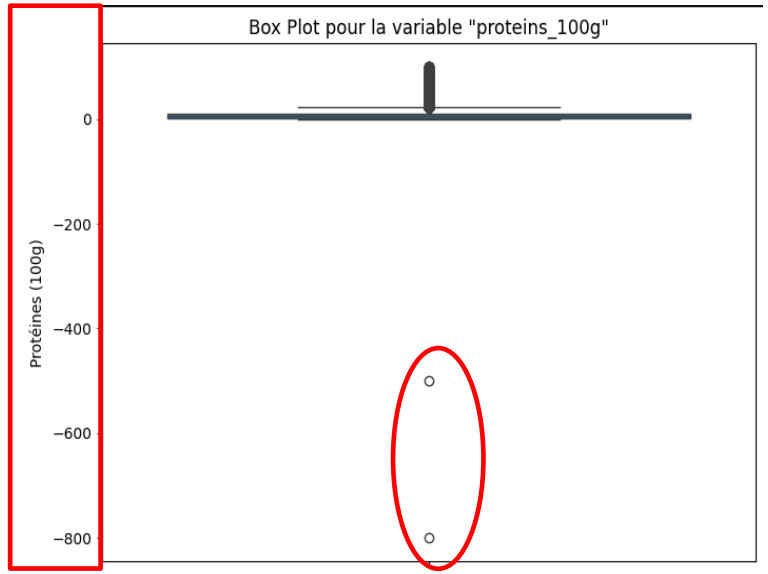
Valeurs énergétiques typiques pour les aliments – repères généraux :
Huiles et graisses : environ 3 500 à 4 000 kJ pour 100 g (≈ 800 à 950 kcal)
Viandes et poissons : environ 500 à 1 000 kJ pour 100 g (≈ 120 à 240 kcal)
Fruits et légumes : environ 100 à 500 kJ pour 100 g (≈ 24 à 120 kcal)
Produits céréaliers : environ 1 500 à 2 500 kJ pour 100 g (≈ 360 à 600 kcal)

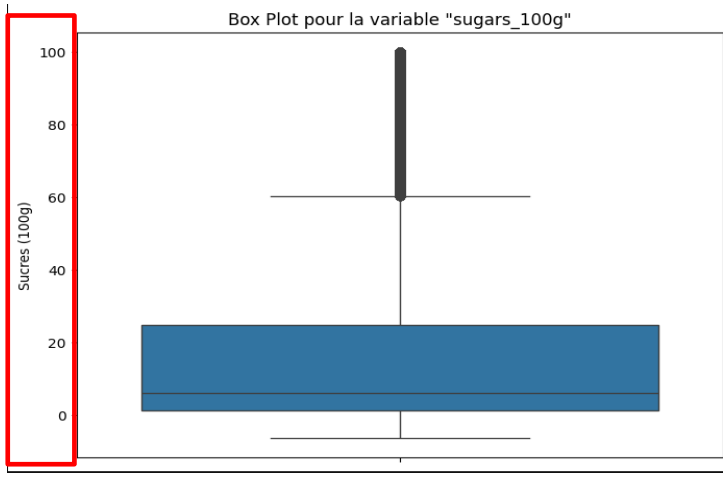
Après analyse, nous avons opté pour une approche métier afin de traiter les valeurs aberrantes.
Cela consiste à afficher les valeurs extrêmes dont l'énergie est supérieure à 9 000 kJ ou inférieure à 0 kJ,
et à supprimer ces valeurs car elles sont mathématiquement impossibles ou physiologiquement incohérentes.

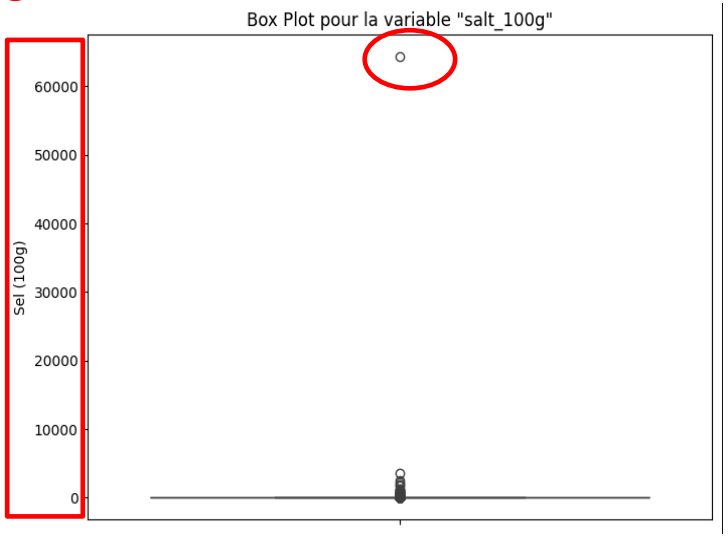
LES VALEURS ABERRANTES DES VARIABLES CIBLES

ON OBSERVE VISUELLEMENT LA PRESENCE D'OUTLIERS









LES VALEURS ABERRANTES DES VARIABLES CIBLES

NETTOYAGE DES VALEURS ABERRANTES PAR L'APPROCHE METIER

Objectif de nettoyage :

- Si la quantité pour 100g est **supérieure à 100g** → Suppression
- Si la quantité pour 100g est **inférieure à 0g** → Suppression

OUTLIERS QUI ONT ÉTÉ SUPPRIMÉES DE LA BASE DE DONNÉES Open Food Facts

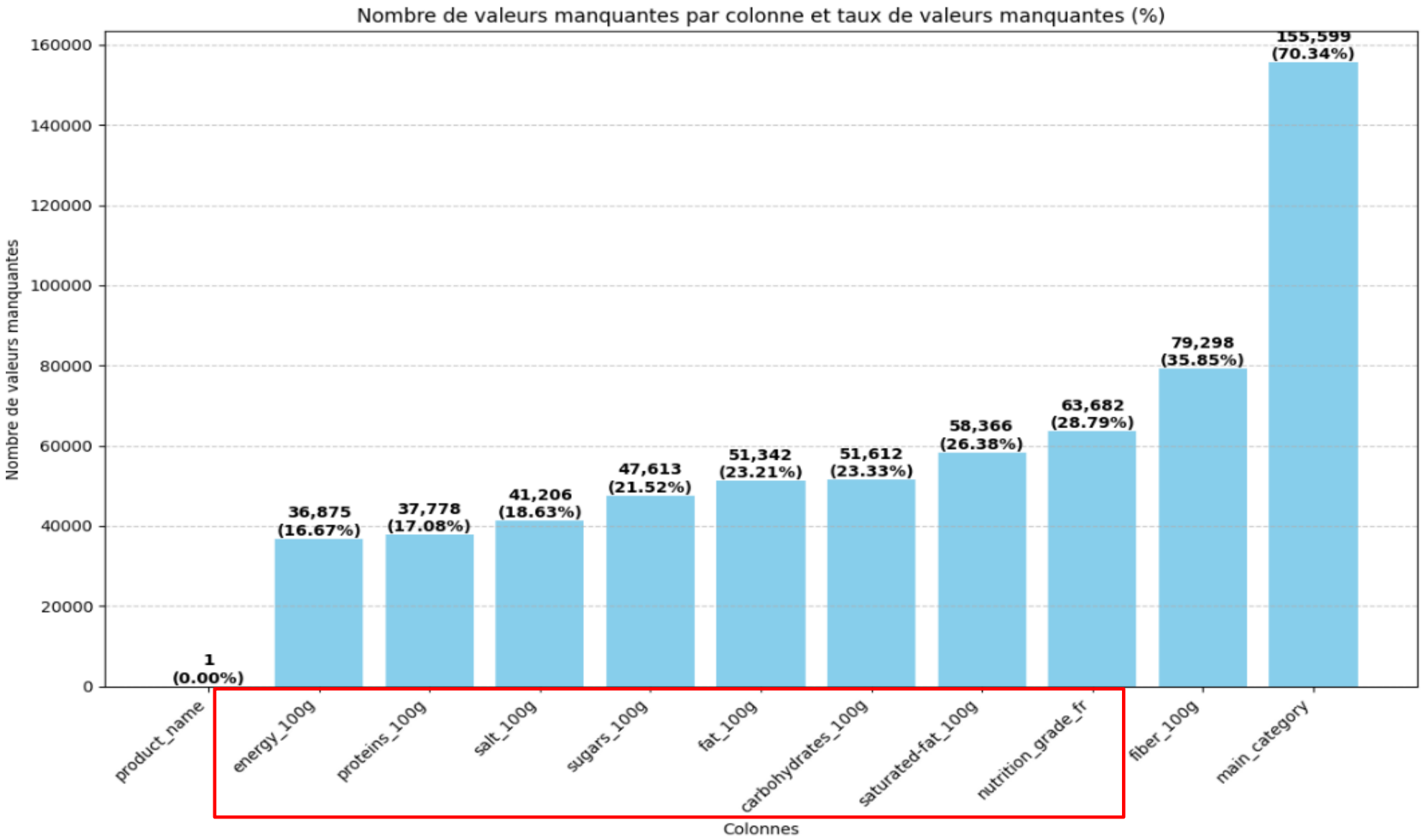
	product_name	fat_100g
209593	Ekstra Jomfru Olivenolie	101.0
210931	Graine de couscous moyen	105.0

	product_name	saturated-fat_100g
78048	Raw 100% Cacao, With Bits Of Delicate Dates	175.38

	product_name	sugars_100g
6553	Grade A Fancy Chopped Spinach	-1.20
13410	Select, Spicy Red Bell Pepper Pasta Sauce	-0.80
76779	Hummous, Black Truffle	-3.57
99419	Italianavera, Tomato Sauce With Gaeta Olives &...	-6.25
133294	Caprice des dieux	-0.10

	product_name	carbohydrates_100g
47640	Toaster Pastries, Strawberry	209.38
78334	Naturally Caffeinated Pure Empower Mint Dietar...	120.00
87603	Mango Jalapeno	125.00
102906	Tamarind Juice With Nata De Coco, Tamarind	2916.67
138720	Sirop d'Agave brun bio	104.00
181254	Agave Syrup dark	103.50
181255	Sirop d agave pur	103.50
181264	Agavendicksaft	103.50
184533	Agavendicksaft Dunkel	103.00
198473	Sauce Caramel	101.00
219387	Banane sèche	139.00

LES VALEURS MANQUANTES



TRAITEMENT DES VALEURS MANQUANTES (Not A Number) PAR L'APPROCHE MÉTIER

&

A) AFFICHAGE DE VALEURS MANQUANTES PAR VARIABLES

	product_name	energy_100g
0	Farine de blé noir	NaN
25	Real Salt Granular	NaN
46	Filet de bœuf	NaN
48	NaN	NaN
71	Fine Sea Salt	NaN

	product_name	proteins_100g
0	Farine de blé noir	NaN
9	Organic Sunflower Oil	NaN
25	Real Salt Granular	NaN
36	Sweeteners, Demerara Turbinado Sugar	NaN
46	Filet de bœuf	NaN
47	Marks % Spencer 2 Blueberry Muffins	NaN

	product_name	salt_100g
0	Farine de blé noir	NaN
4	Organic Polenta	NaN
5	Breadshop Honey Gone Nuts Granola	NaN
6	Organic Long Grain White Rice	NaN
8	Organic Dark Chocolate Minis	NaN
9	Organic Sunflower Oil	NaN

	product_name	sugars_100g
0	Farine de blé noir	NaN
4	Organic Polenta	NaN
6	Organic Long Grain White Rice	NaN
9	Organic Sunflower Oil	NaN
10	Organic Adzuki Beans	NaN
11	Organic Penne Pasta	NaN
13	Organic Golden Flax Seeds	NaN
14	Organic Spicy Punks	NaN

	product_name	fat_100g
0	Farine de blé noir	NaN
6	Organic Long Grain White Rice	NaN
25	Real Salt Granular	NaN
36	Sweeteners, Demerara Turbinado Sugar	NaN
39	Organic Black Beans	NaN
46	Filet de bœuf	NaN
47	Marks % Spencer 2 Blueberry Muffins	NaN

	product_name	carbohydrates_100g
0	Farine de blé noir	NaN
9	Organic Sunflower Oil	NaN
25	Real Salt Granular	NaN
46	Filet de bœuf	NaN
47	Marks % Spencer 2 Blueberry Muffins	NaN

B) Traitement des valeurs manquantes

REMPLACEMENT DES VALEURS MANQUANTES PAR LA LOGIQUE MÉTIER

SI le taux de sucre (sugars_100g) est égal à 100 g

ALORS IMPUTER

proteins_100g = 0 et fat_100g = 0

	sugars_100g
36	100.0
72	100.0
166	100.0
370	100.0
371	100.0

	product_name	proteins_100g
36	Sweeteners, Demerara Turbinado Sugar	0.0
72	Sweeteners, Organic Fair Trade Sugar	0.0
166	Organic Unrefined Mascobado Sugar	0.0
370	Tnt Exploding Candy	0.0
371	Exploding Candy	0.0

	sugars_100g
36	100.0
72	100.0
166	100.0
370	100.0
371	100.0
2425	100.0
2426	100.0

	product_name	fat_100g
36	Sweeteners, Demerara Turbinado Sugar	0.0
72	Sweeteners, Organic Fair Trade Sugar	0.0
166	Organic Unrefined Mascobado Sugar	0.0
370	Tnt Exploding Candy	0.0
371	Exploding Candy	0.0
2425	Dessert Topping, Red Sugar	0.0
2426	Green Sugar Dessert Toppings	0.0

Hypothèse métier : un aliment composé à 100 % de sucre contient logiquement 0 g de protéines et 0 g de matières grasses.

B) Traitement des valeurs manquantes

REMPLACEMENT DES VALEURS MANQUANTES PAR LOGIQUE

SI le taux de graisses (fat_100g) est égal à 100 g

ALORS IMPUTER :
sugars_100g = 0 et proteins_100g =

	product_name	fat_100g	\
9	Organic Sunflower Oil	100.0	
96	Organic Extra Virgin Olive Oil	100.0	
98	Organic Canola Oil Refined	100.0	
163	Organic Unrefined Extra Virgin Coconut Oil	100.0	
247	100% Pure Canola Oil	100.0	
475	Ventura, Soybean - Peanut Frying Oil Blend	100.0	

	sugars_100g
9	0.0
96	0.0
98	0.0
163	0.0
247	0.0
475	0.0

	product_name	fat_100g	\
9	Organic Sunflower Oil	100.0	
96	Organic Extra Virgin Olive Oil	100.0	
98	Organic Canola Oil Refined	100.0	
163	Organic Unrefined Extra Virgin Coconut Oil	100.0	
247	100% Pure Canola Oil	100.0	
475	Ventura, Soybean - Peanut Frying Oil Blend	100.0	

	proteins_100g
9	0.0
96	0.0
98	0.0
163	0.0
247	0.0
475	0.0

Un produit composé à 100 % de matières grasses (ex : huile végétale) contient logiquement 0 g de sucre et 0 g de protéines.

C) Traitement des valeurs manquantes (fin du nettoyage)

Imputation des valeurs manquantes restantes par la moyenne de chaque variable

```
Nombre de valeurs manquantes après l'imputation :
index                    0
energy_100g              0
fat_100g                 0
saturated-fat_100g       0
carbohydrates_100g       0
sugars_100g              0
fiber_100g               0
proteins_100g            0
salt_100g                0
```

Résultat :

Toutes les valeurs manquantes ont été imputées

Le jeu de données est désormais complet et prêt à être analysé

D) VISUALISATION DU JEU DE DONNÉES NETTOYÉ

	index	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000
mean	162051.289430	1135.982292	12.496463	5.190437	33.104861	16.277209	2.856398	7.145728	1.536851
std	91795.600365	713.757393	14.416562	6.755245	25.516641	18.681516	3.670840	7.338418	5.231870
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	82924.250000	523.000000	1.270000	0.400000	10.000000	2.400000	0.700000	1.700000	0.120000
50%	166039.500000	1135.982292	12.496463	5.190437	33.104861	13.000000	2.856398	6.900000	0.906780
75%	239206.500000	1594.000000	15.000000	5.190437	51.000000	17.000000	2.856398	8.700000	1.536851
max	320770.000000	8715.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000

1. Aucune valeur **> à 100g**

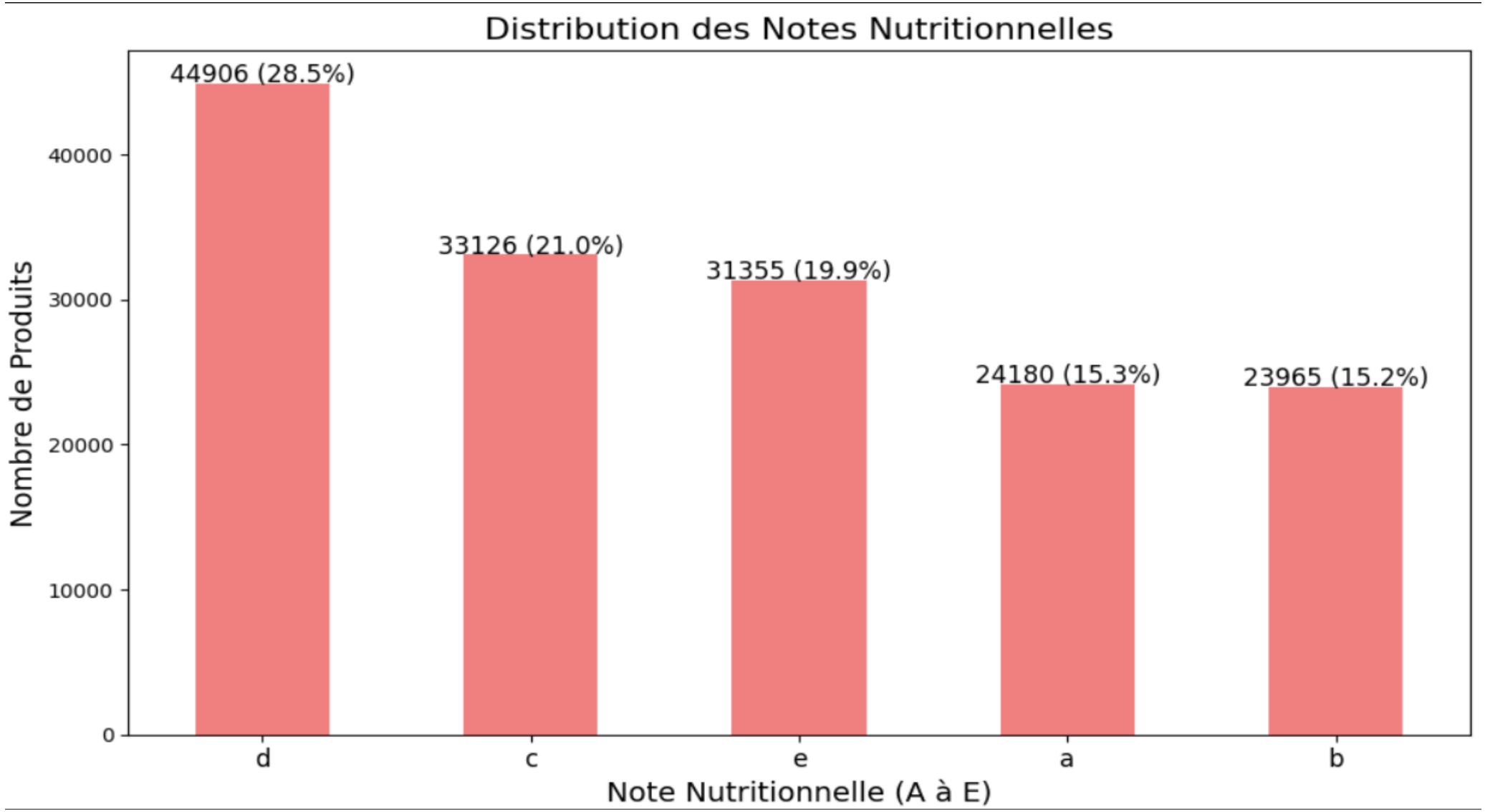
2. Aucune valeur inférieure à 0 g pour les variables cibles, sauf pour *energy_100g* (exprimée en kilojoules)

3. Aucune valeur supérieure à 9000 kilojoules pour ***energy_100g***

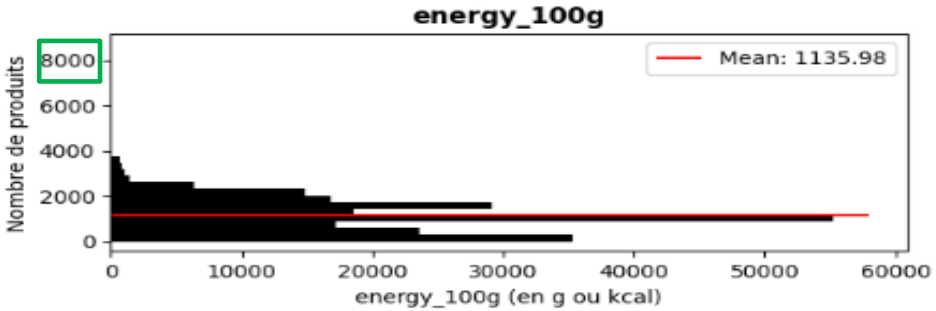
5. ANALYSE DES DONNÉES

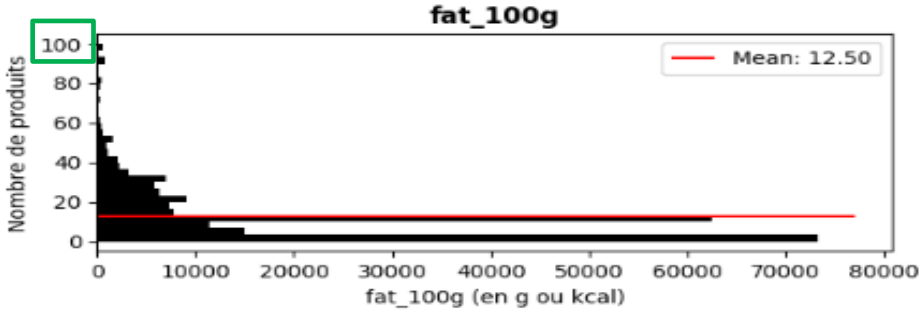
ANALYSE UNIVARIÉE

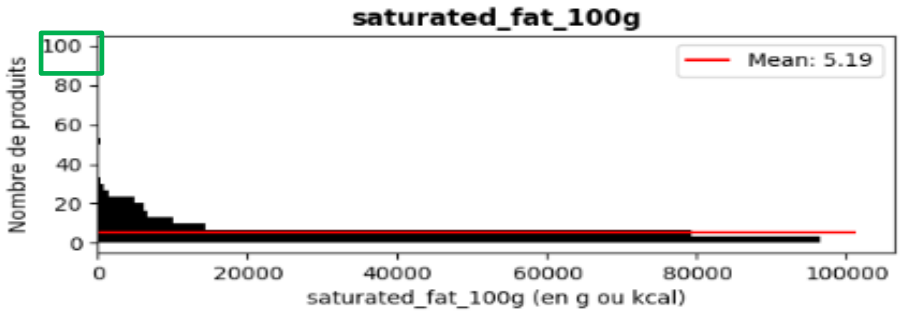
Variable : nutrition_grade_fr

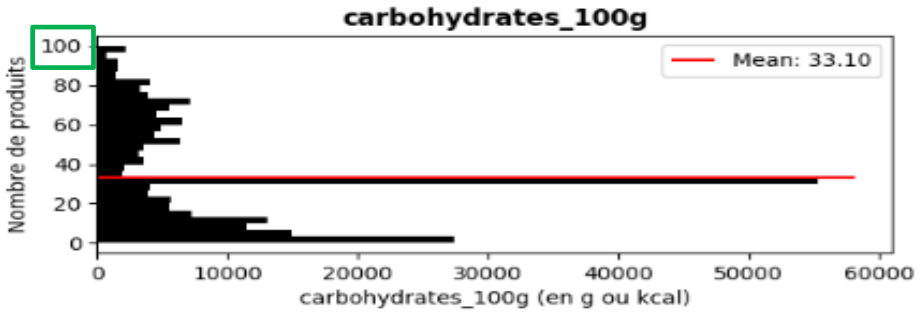


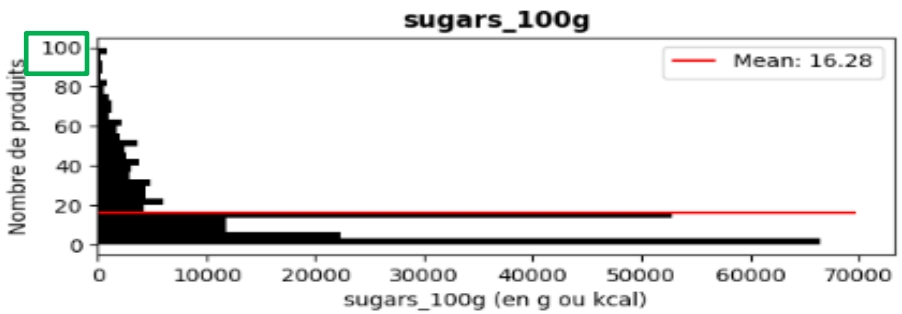
ANALYSE UNIVARIÉE

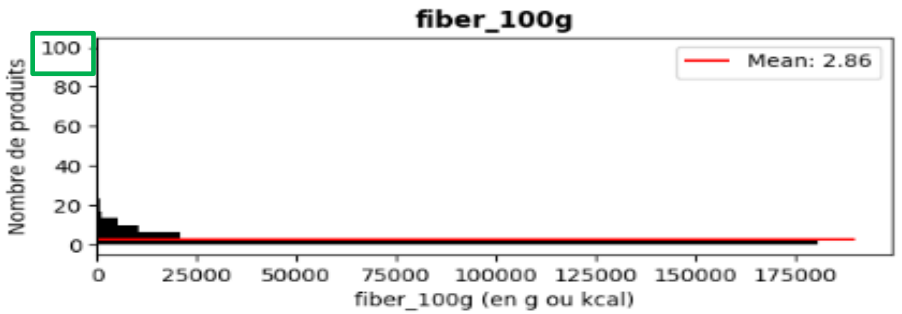


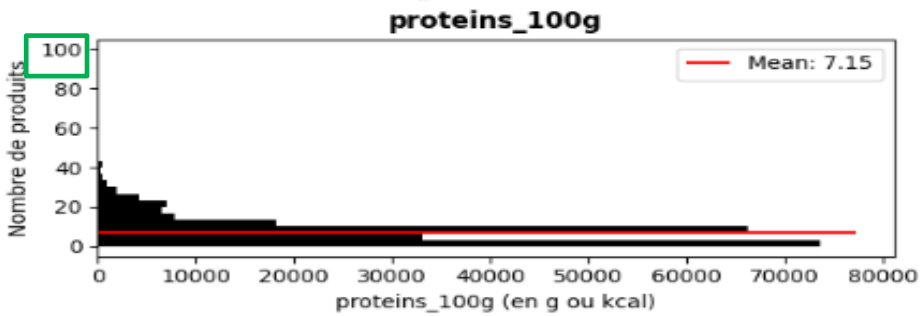


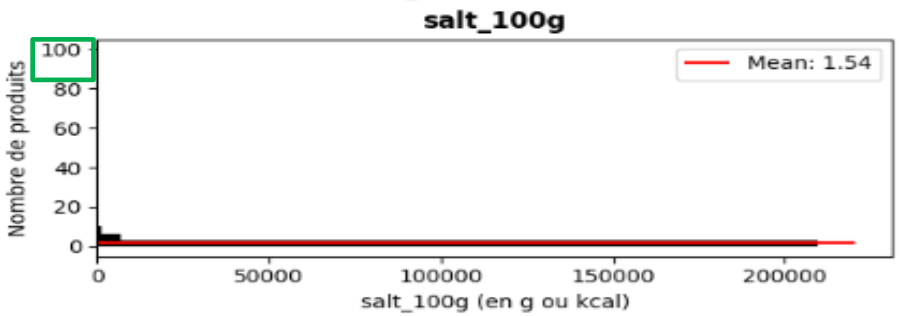






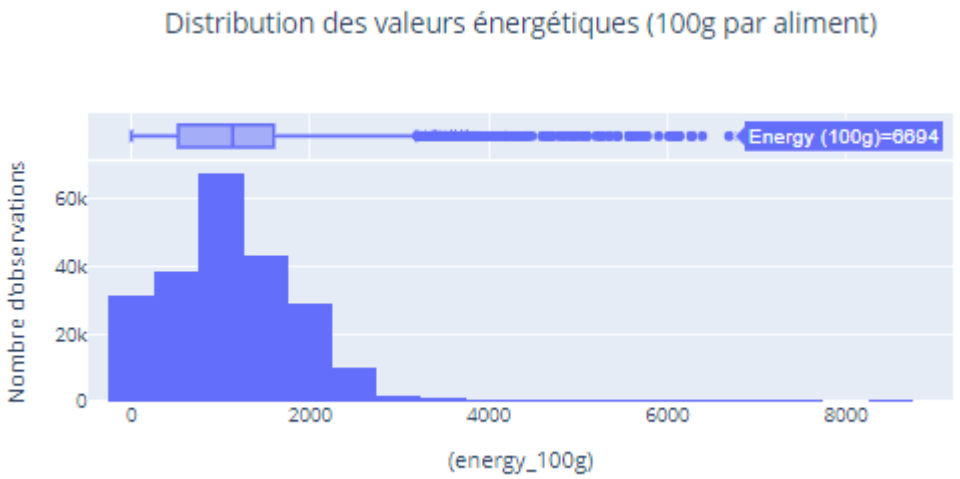


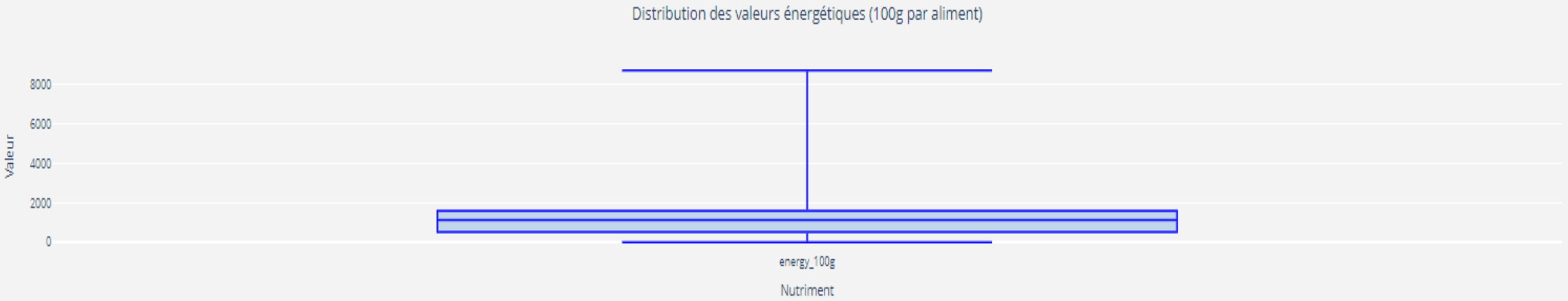




ANALYSE UNIVARIÉE

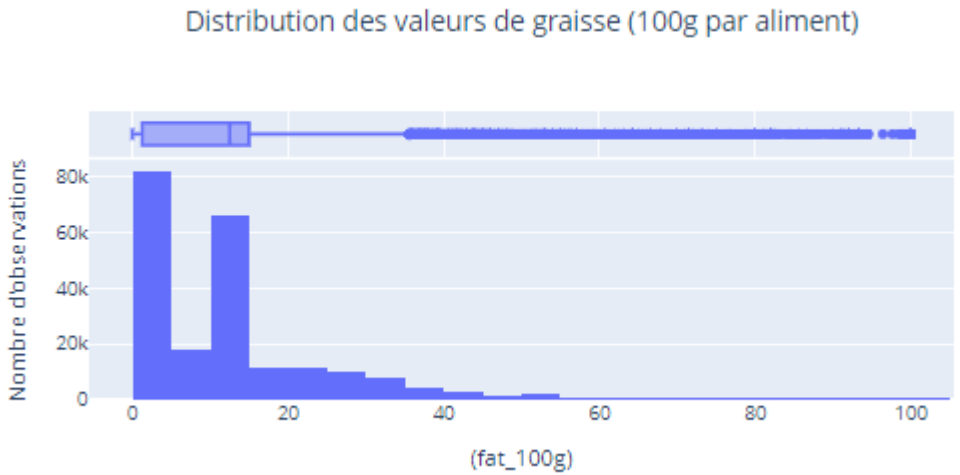
Distribution des valeurs par variable

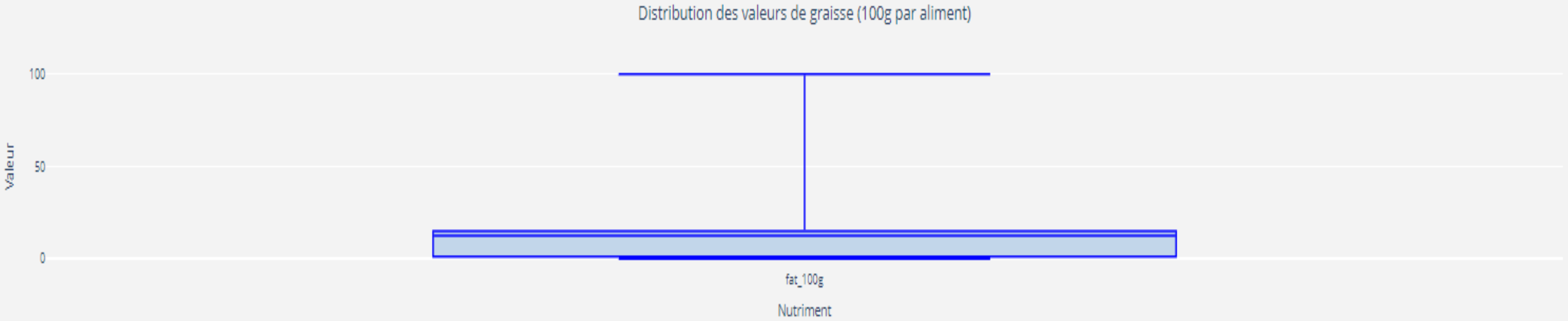




ANALYSE UNIVARIÉE

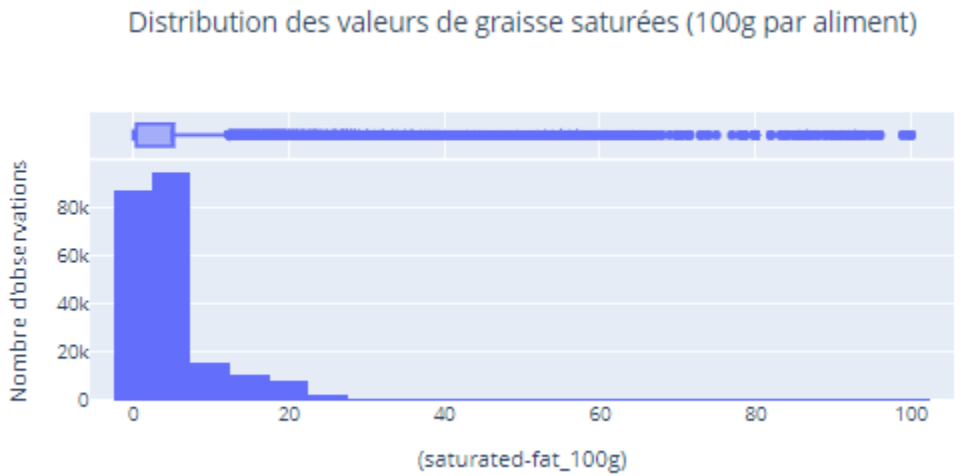
Distribution des valeurs par variable

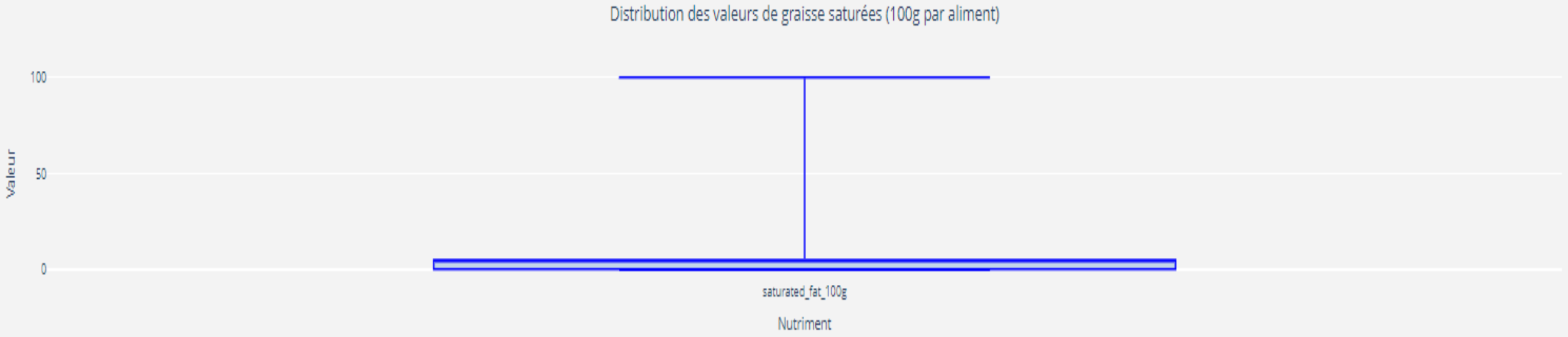




ANALYSE UNIVARIÉE

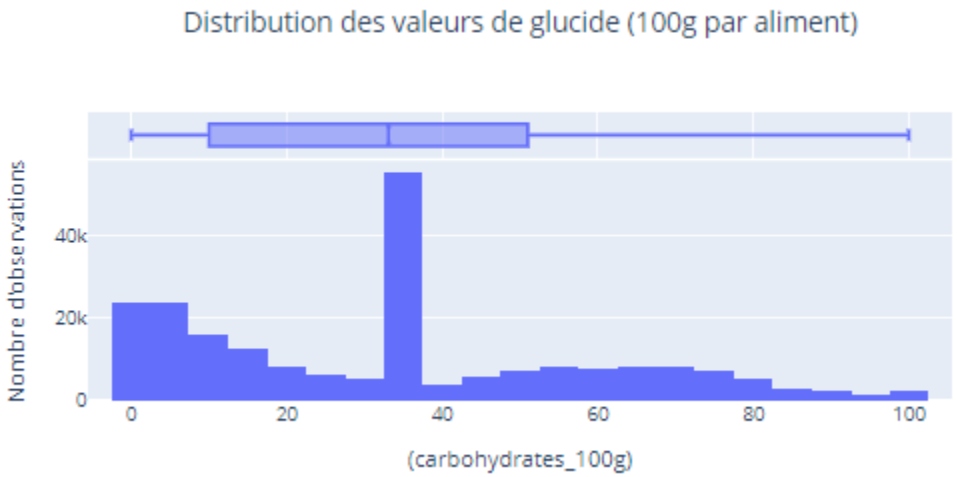
Distribution des valeurs par variable

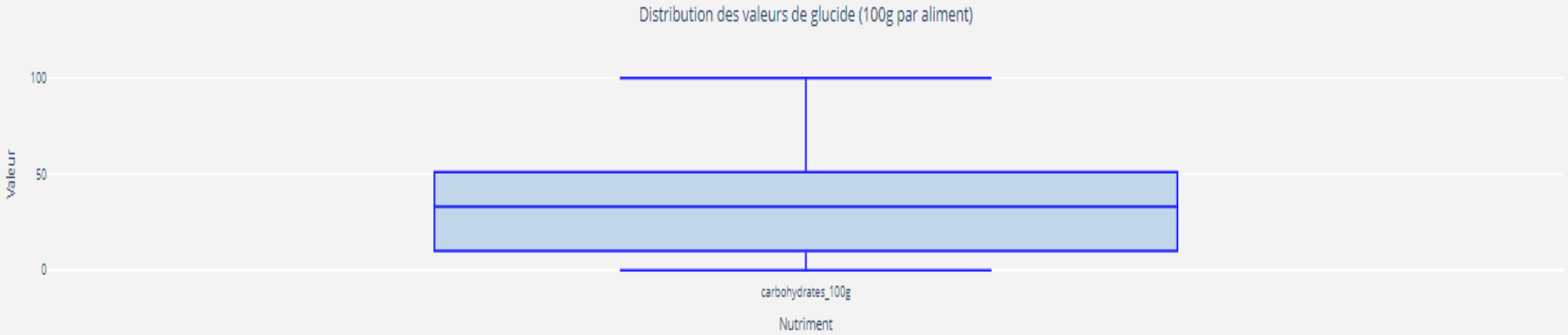




ANALYSE UNIVARIÉE

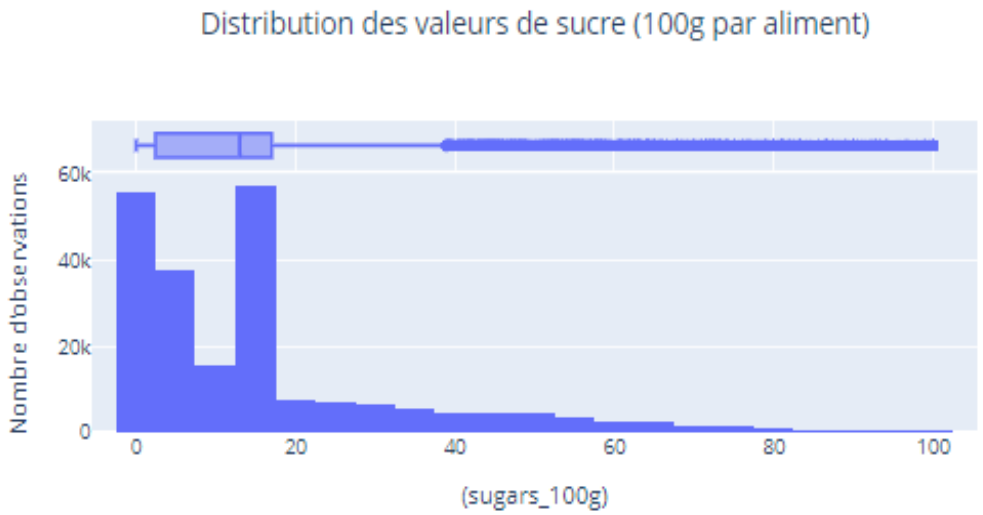
Distribution des valeurs par variable

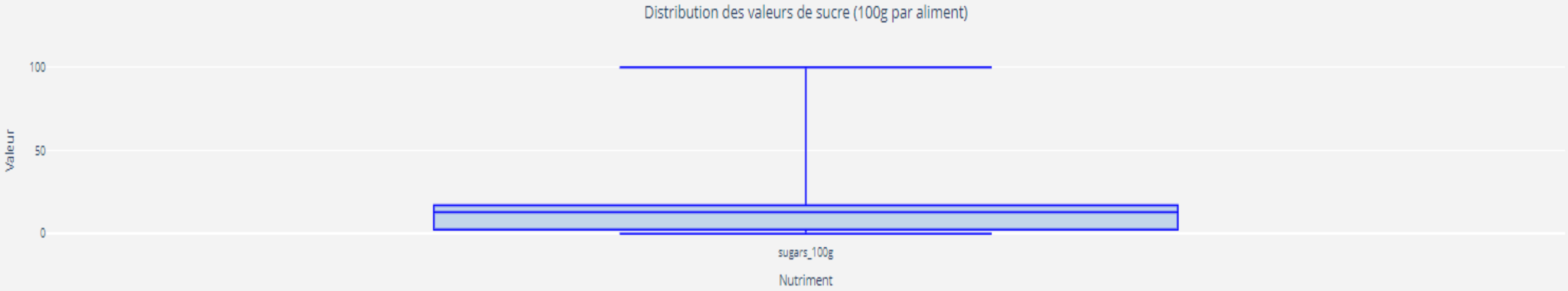




ANALYSE UNIVARIÉE

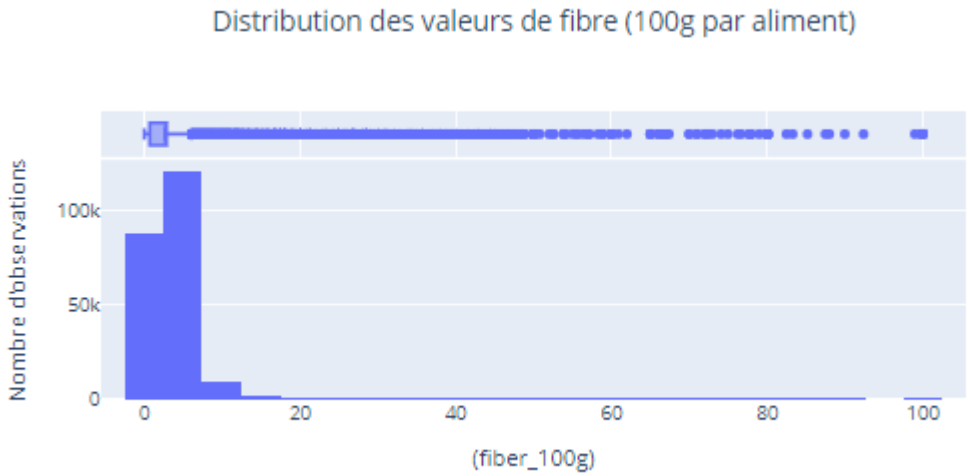
Distribution des valeurs par variable

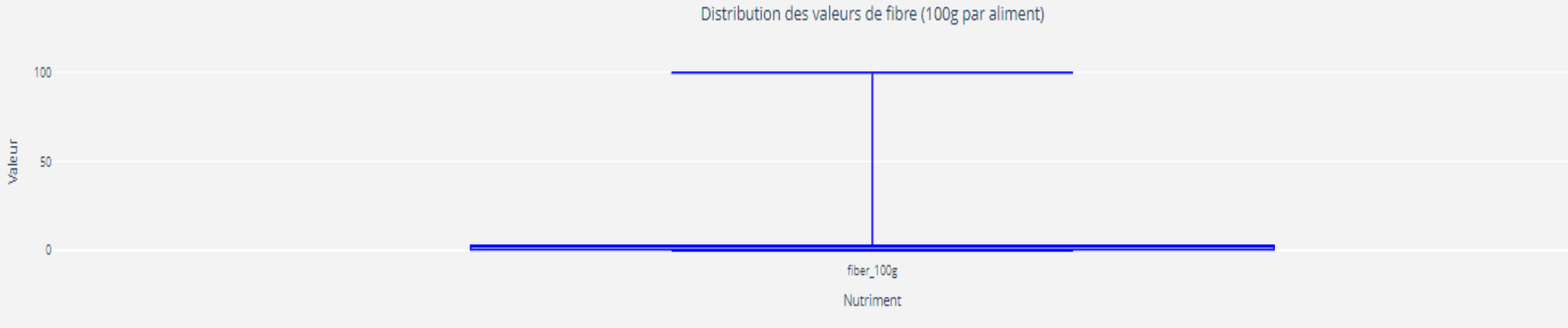




ANALYSE UNIVARIÉE

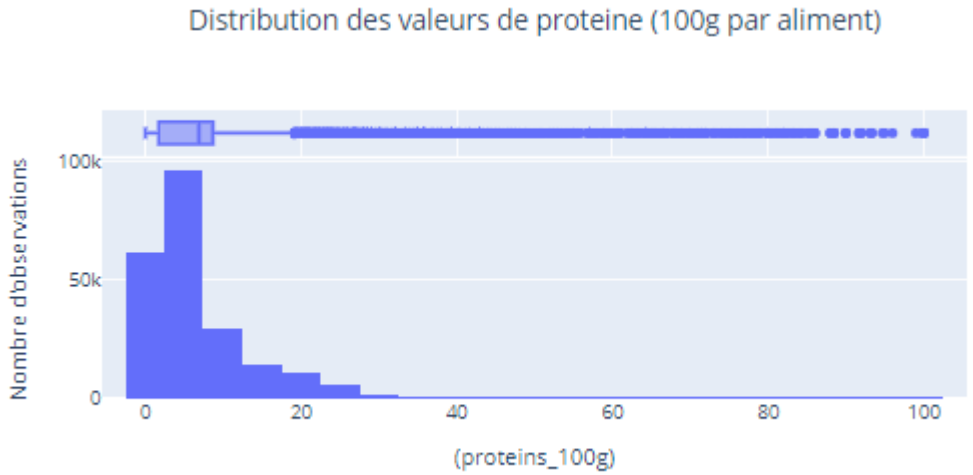
Distribution des valeurs par variable

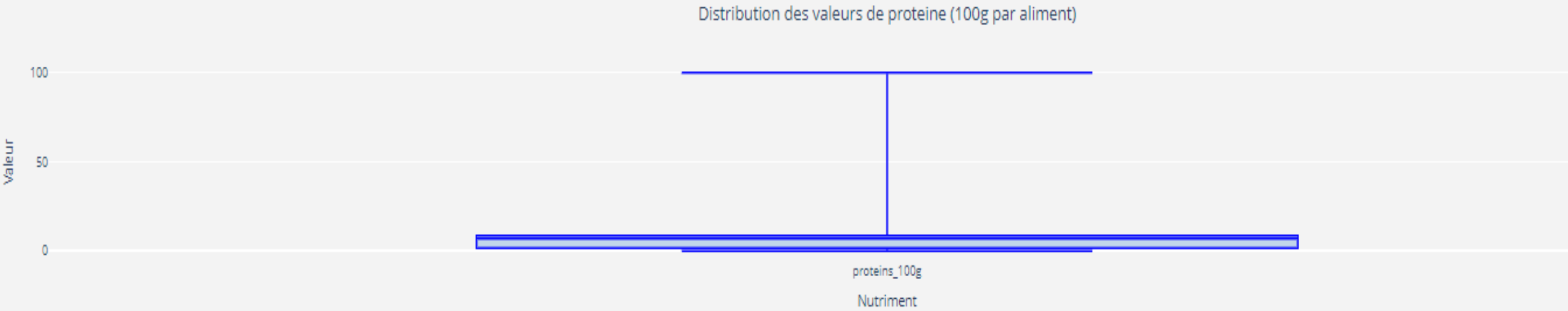




ANALYSE UNIVARIÉE

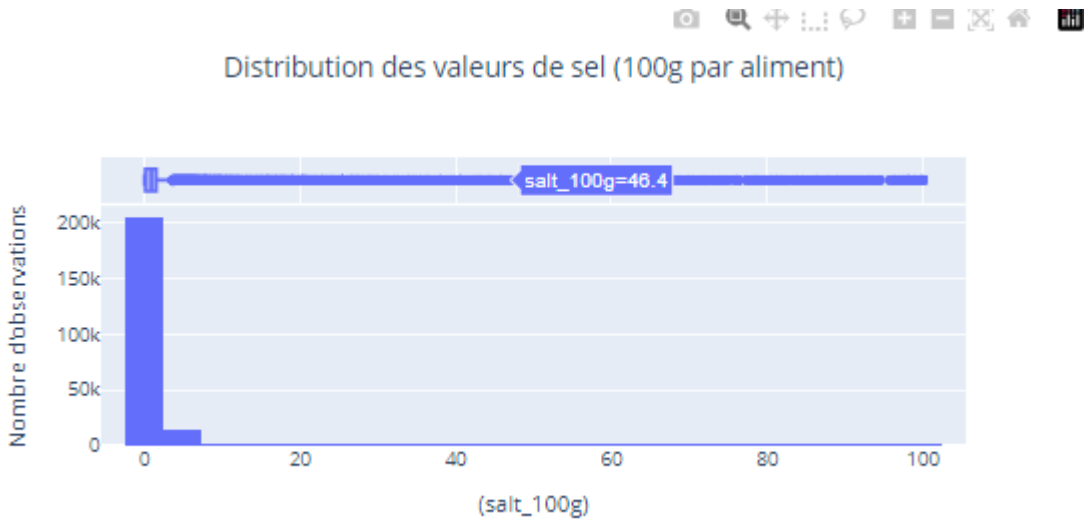
Distribution des valeurs par variable

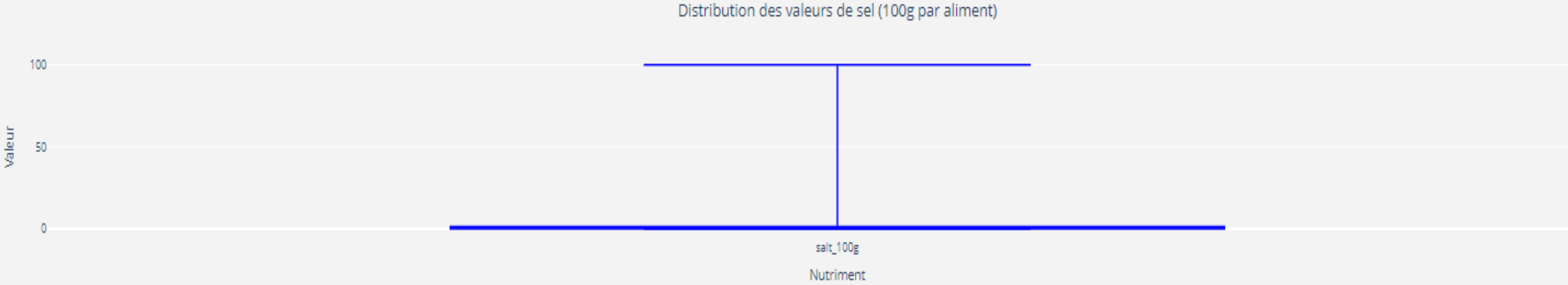




ANALYSE UNIVARIÉE

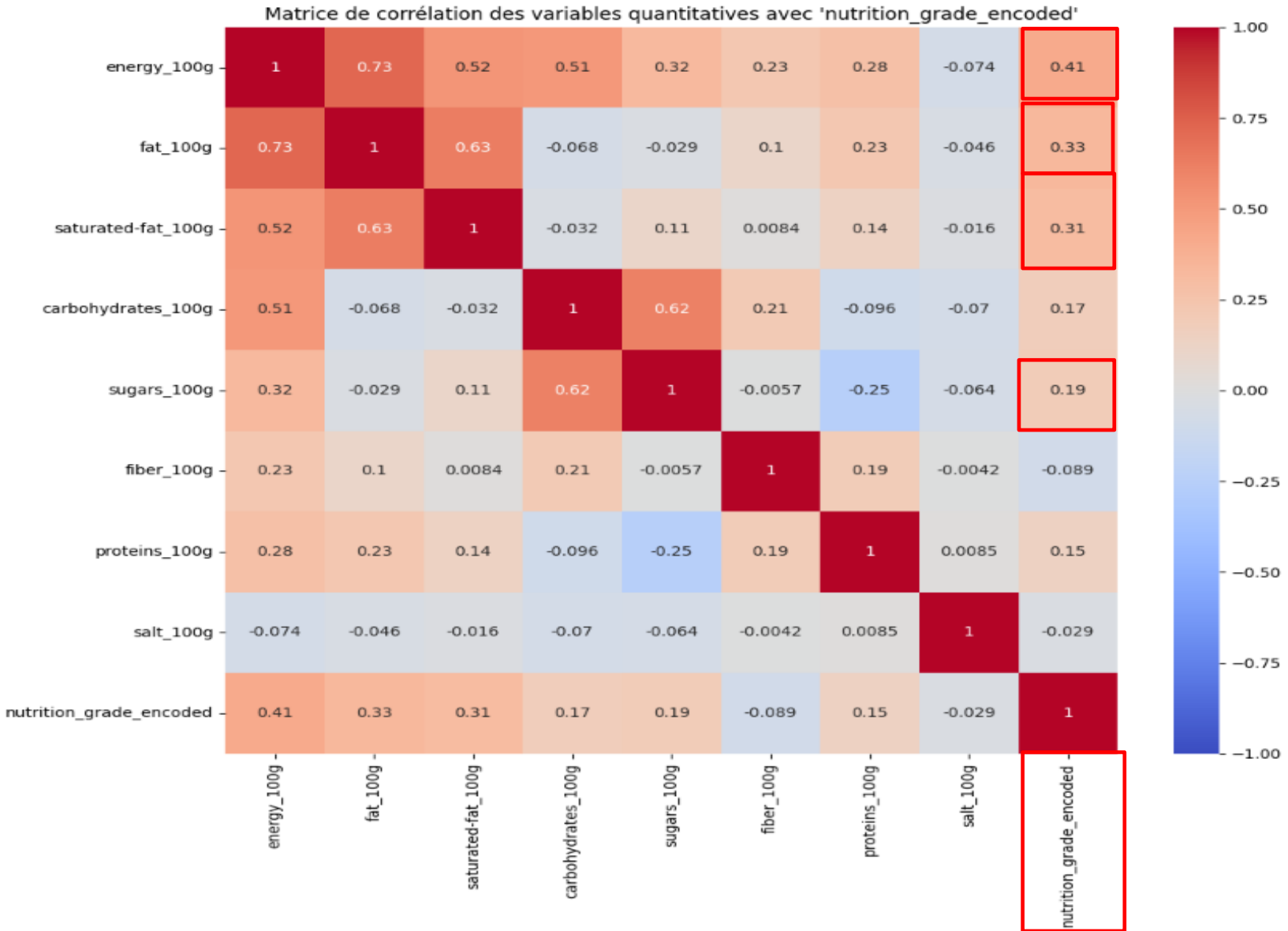
Distribution des valeurs par variable





ANALYSE BIVARIÉE

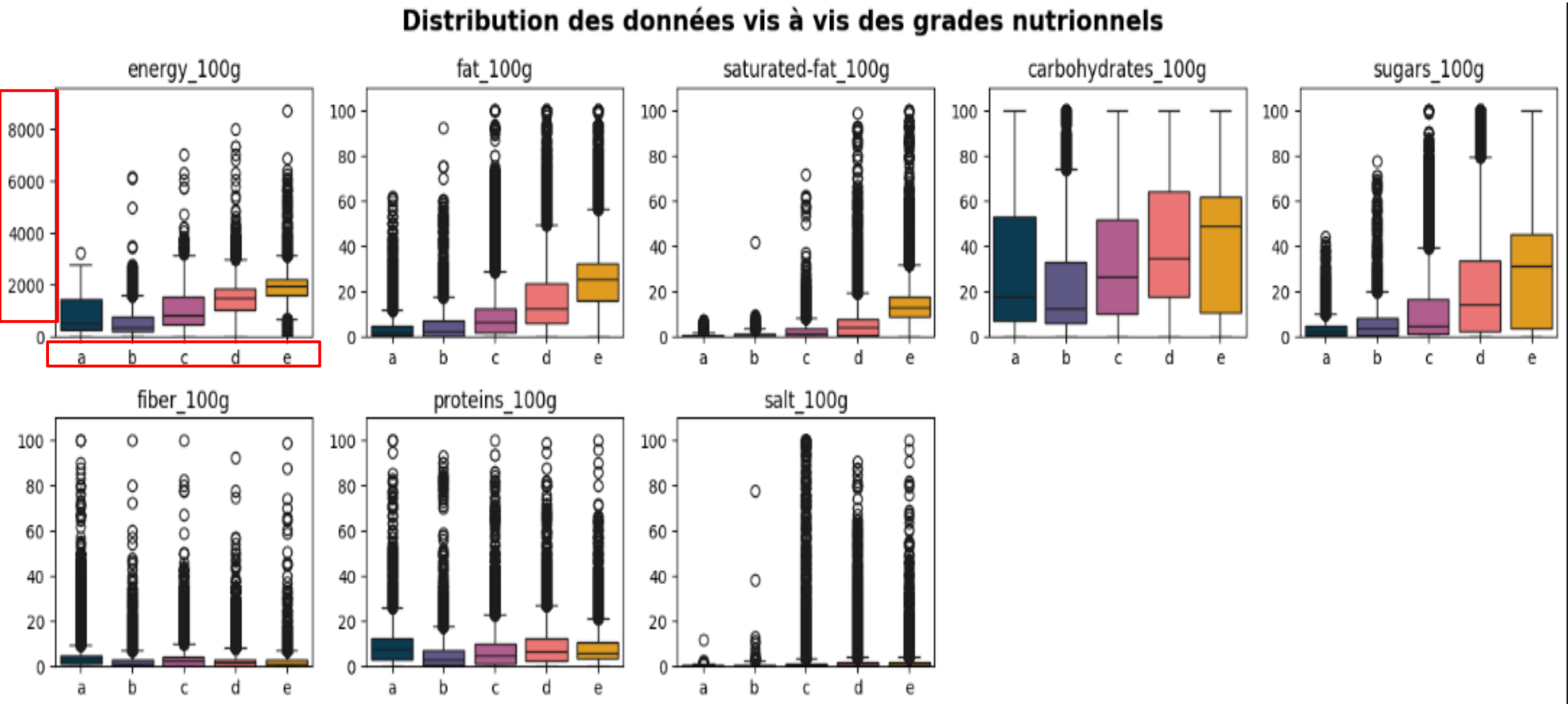
Matrice de correlation



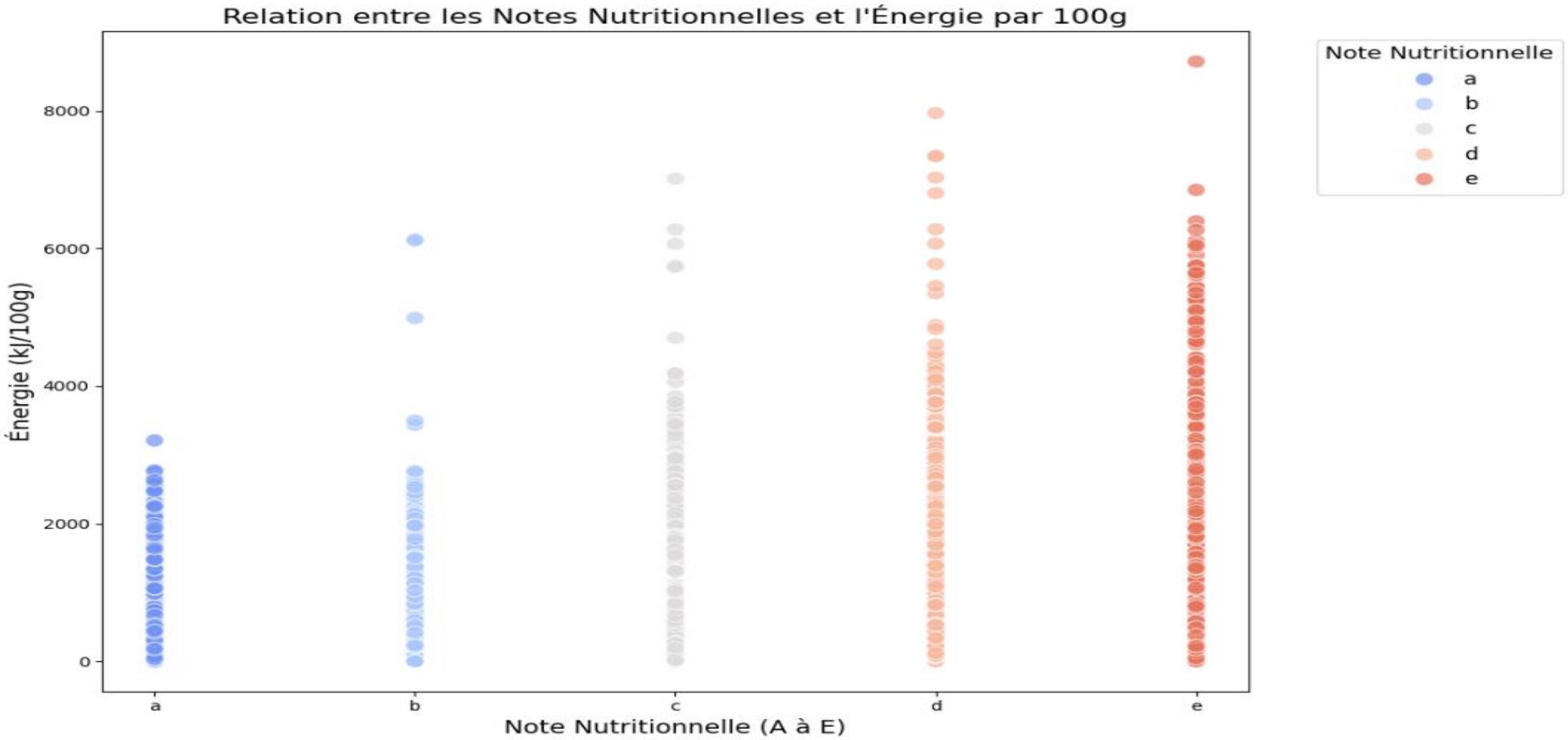
- La variable catégorielle **nutrition_grade_fr** a été convertie en une variable ordinale pour permettre une analyse de corrélation avec les variables quantitatives.
- La valeur énergétique, les graisses totales et les graisses saturées sont modérément corrélées positivement à un mauvais score nutritionnel
- Les protéines sont légèrement corrélées négativement, ce qui suggère une association avec de meilleures notes (A/B).
- Le **sel** et les **fibres** semblent avoir une influence plus faible sur la note globale

ANALYSE BIVARIÉE

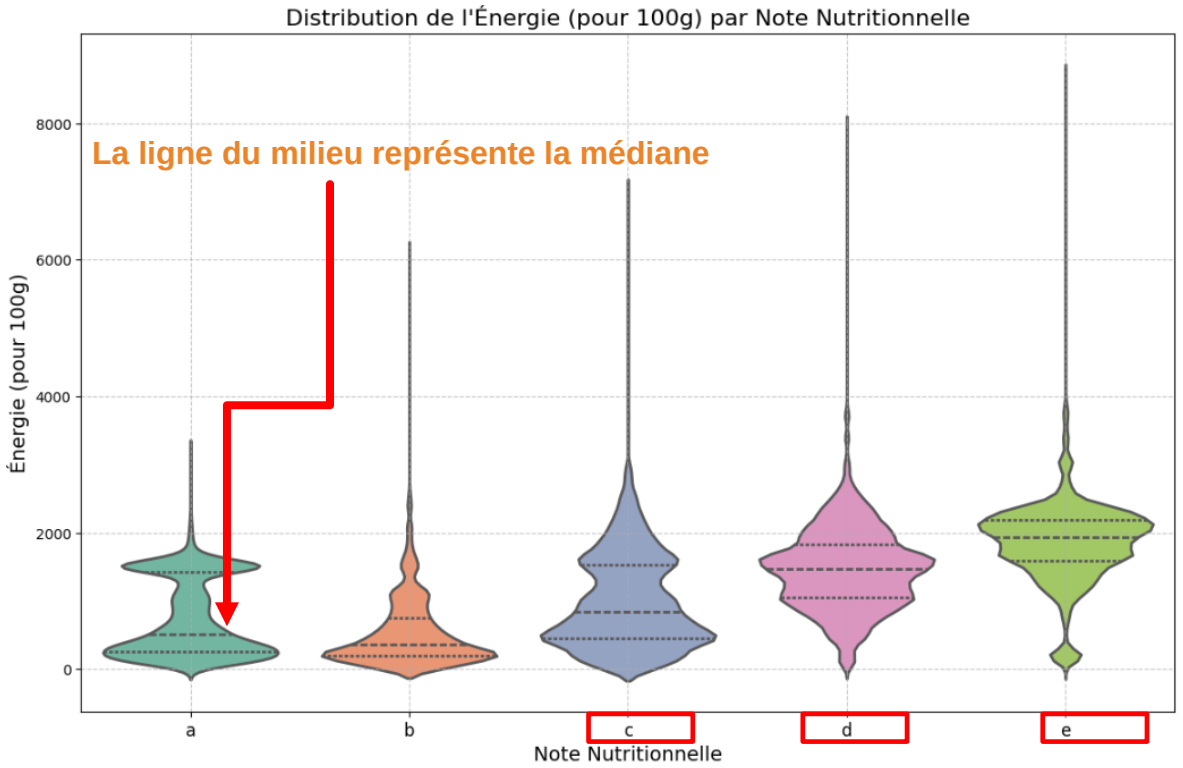
La variable cible nutrition_grade_fr avec les variables quantitatives précédemment nettoyées

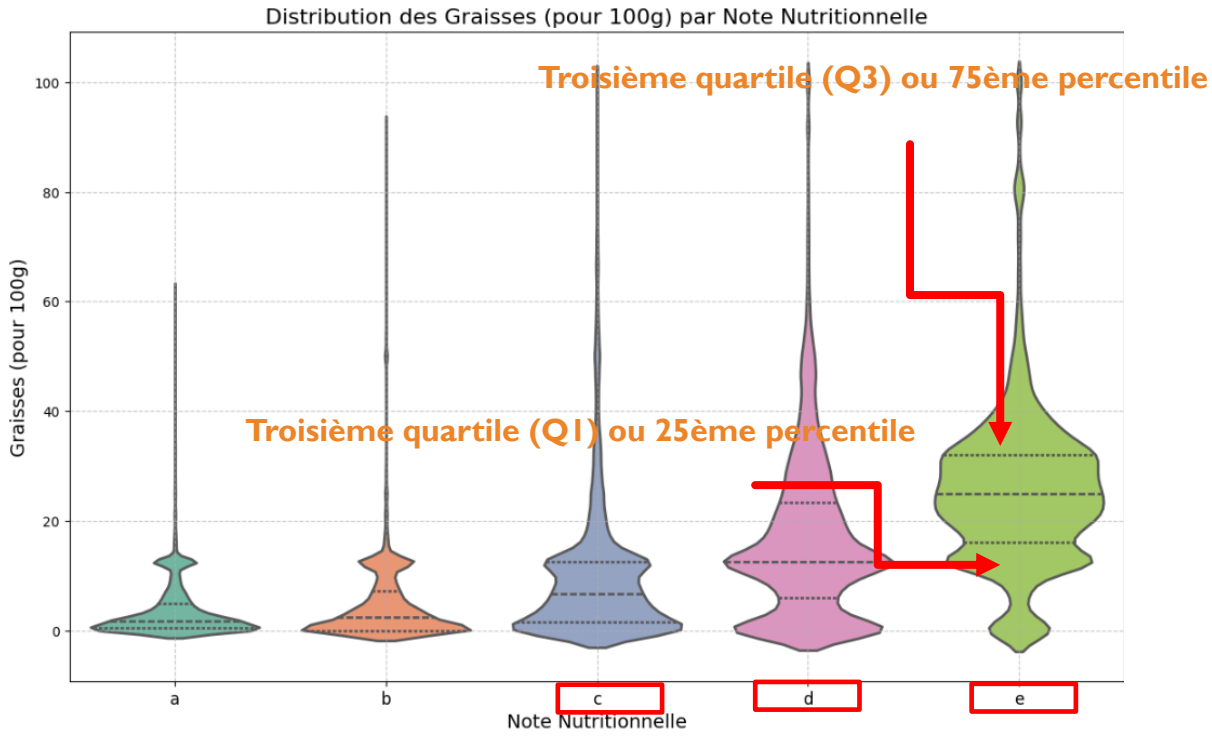


ANALYSE BIVARIEE



ANALYSE BIVARIÉE

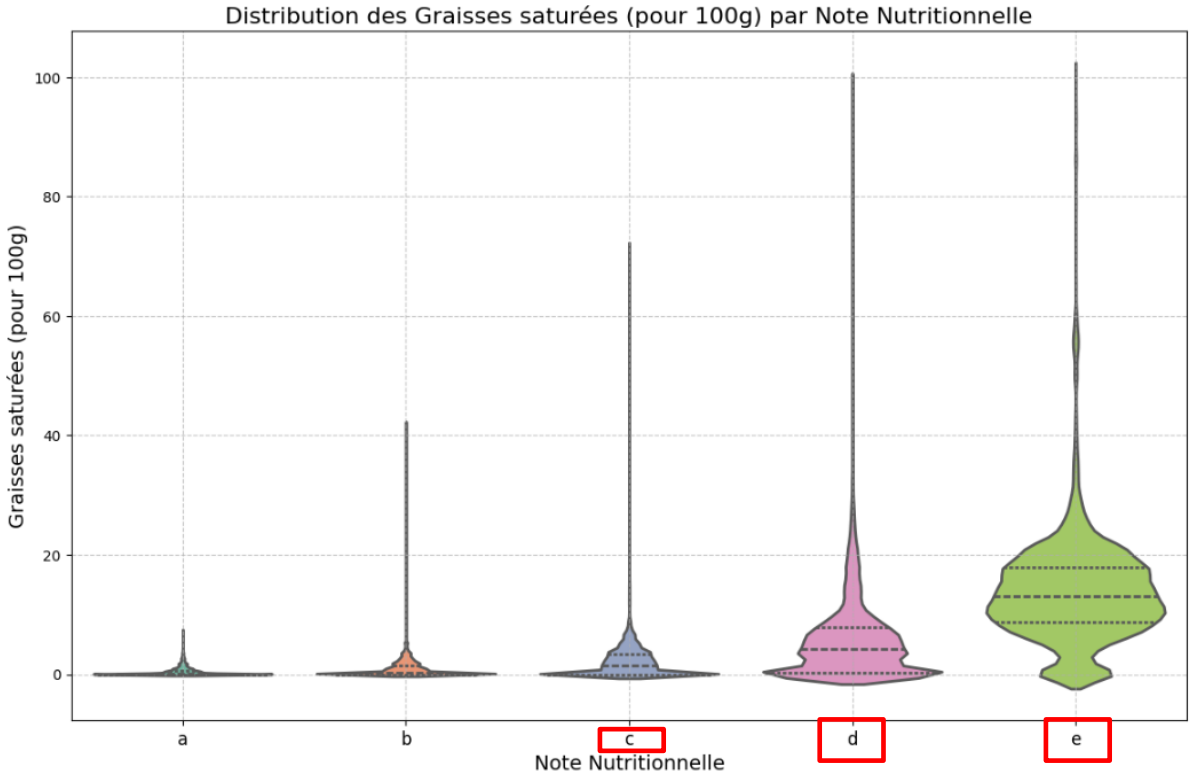


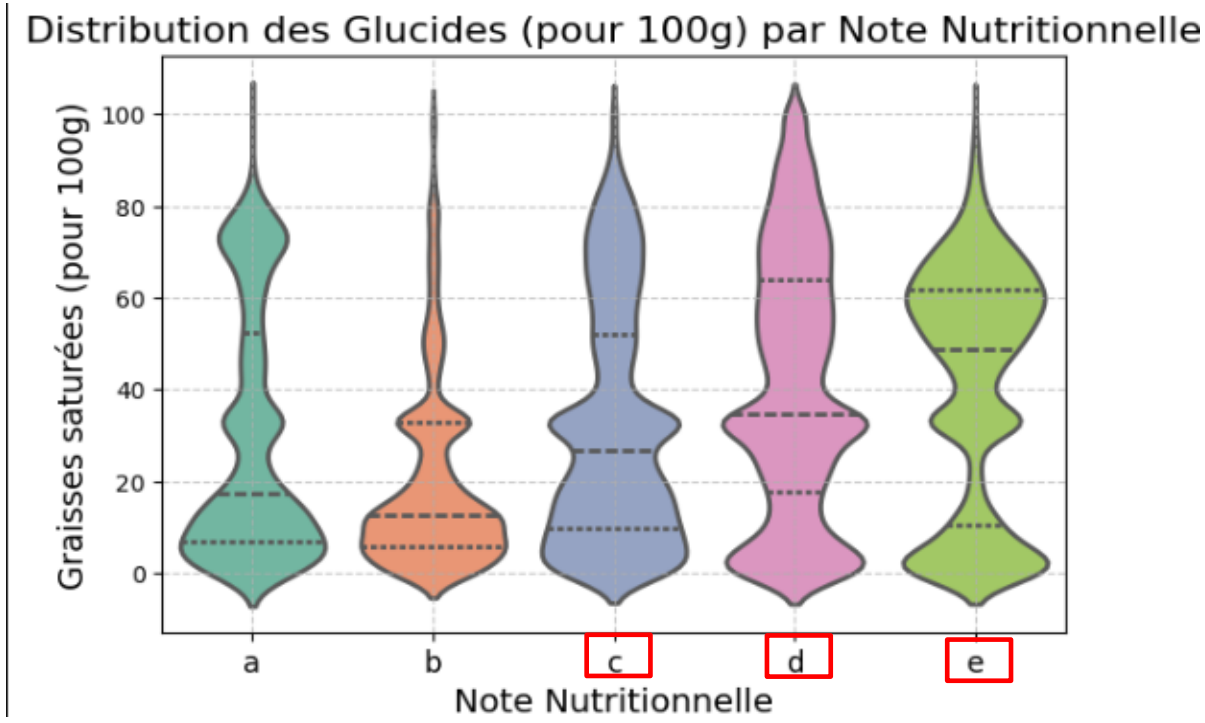


- La note nutritionnelle d'un **produit** à **valeur énergetique** **élevé** est => **Moyenne**, **mauvaise voire très mauvaise**
- La note nutritionnelle d'un **produit** avec **un taux de graisse** **élevé** est => **Moyenne, mauvaise voire très mauvaise**

(**A** : Très bon sur le plan nutritionnel. / **B** : Bon. / **C** : Moyen. / **D** : Mauvais. / **E** : Très mauvais.)

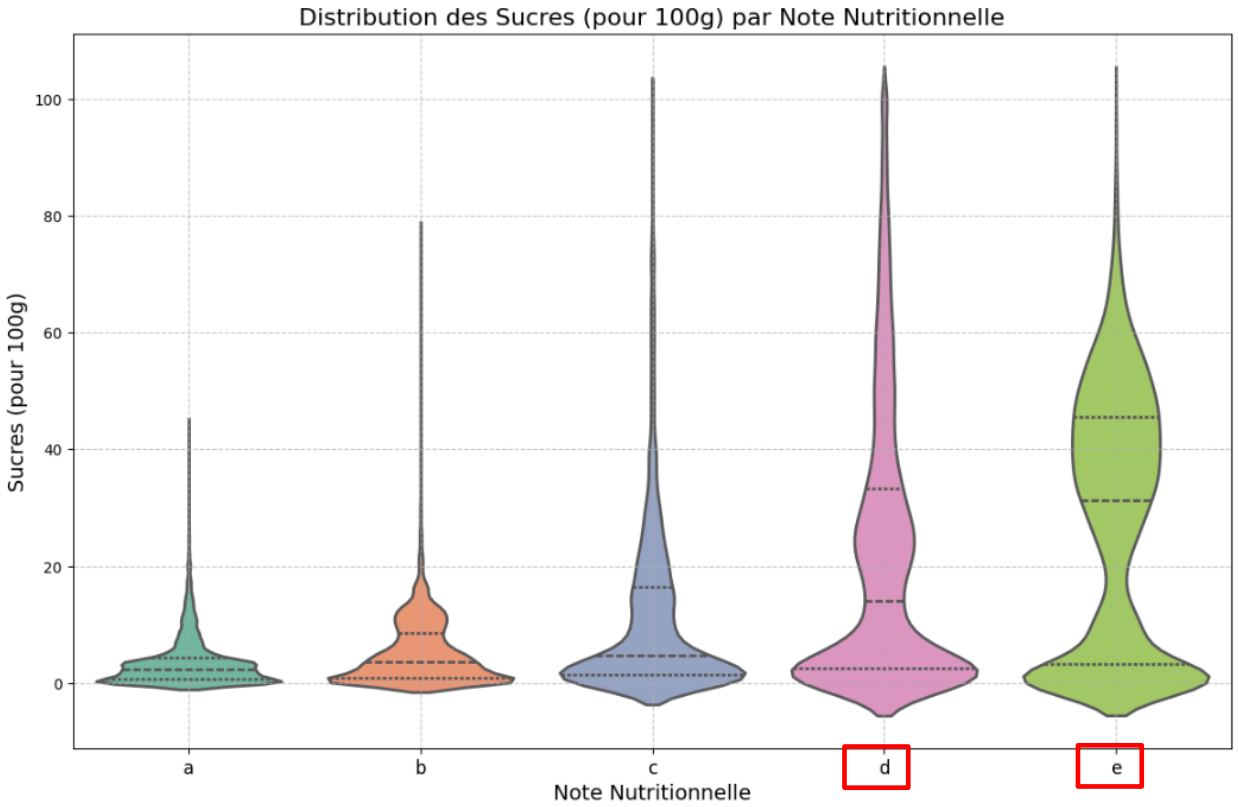
ANALYSE BIVARIÉE

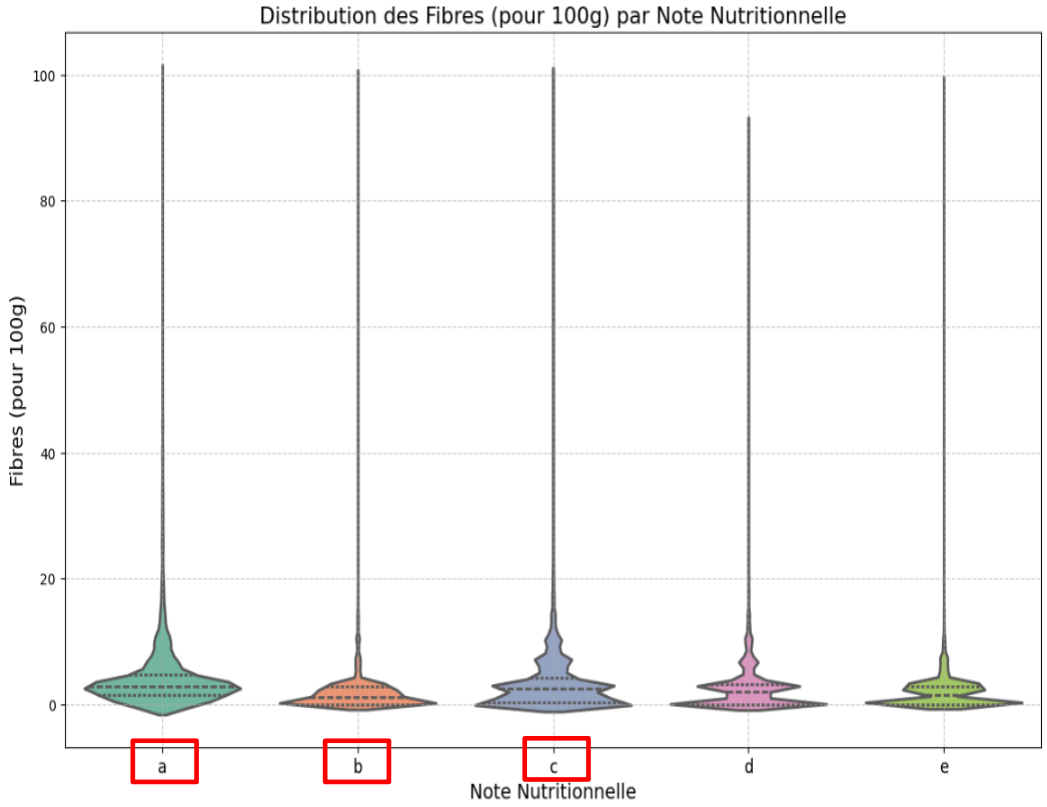




(**A** : Très bon sur le plan nutritionnel. / **B** : Bon. / **C** : Moyen. / **D** : Mauvais. / **E** : Très mauvais.)

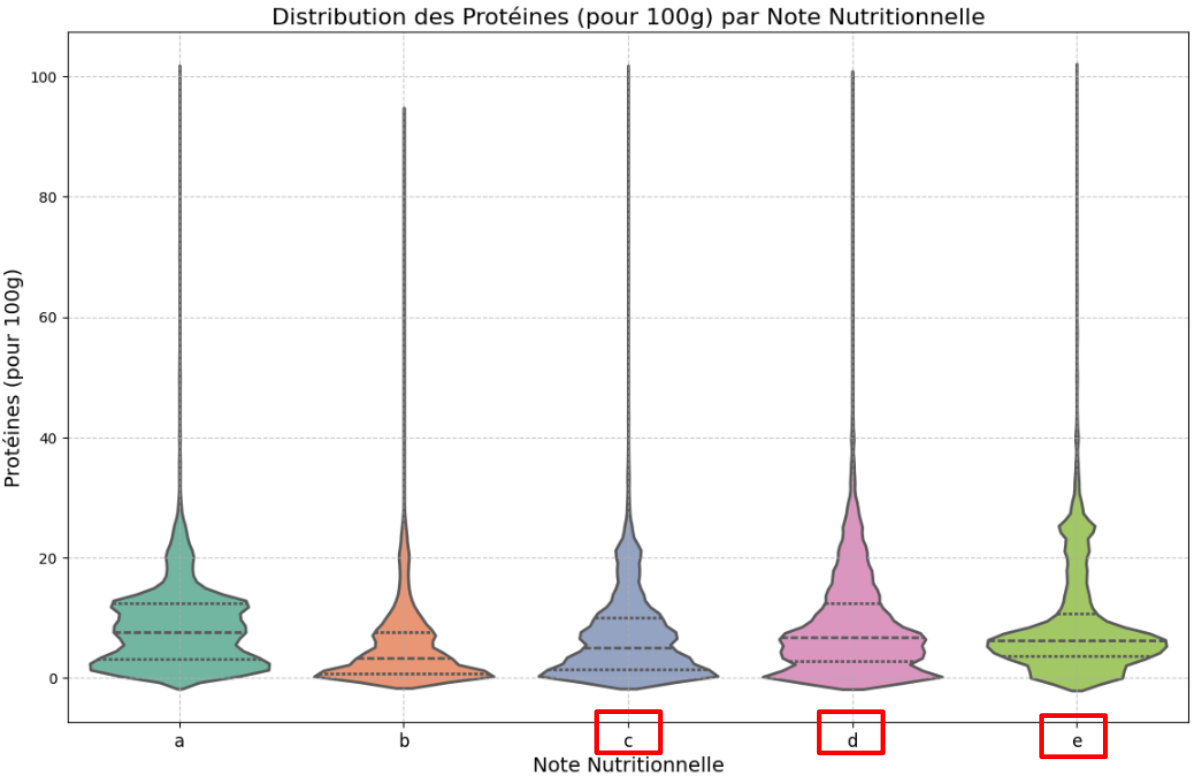
ANALYSE BIVARIÉE

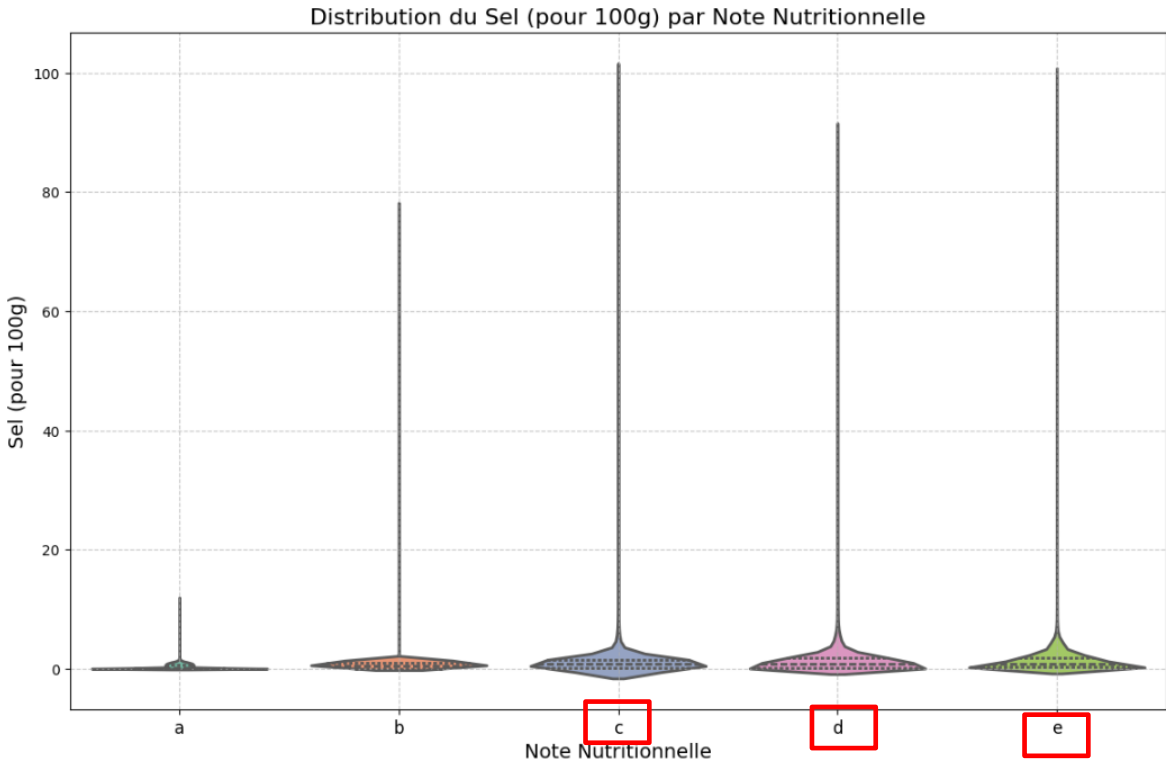




(**A** : Très bon sur le plan nutritionnel. / **B** : Bon. / **C** : Moyen. / **D** : Mauvais. / **E** : Très mauvais.)

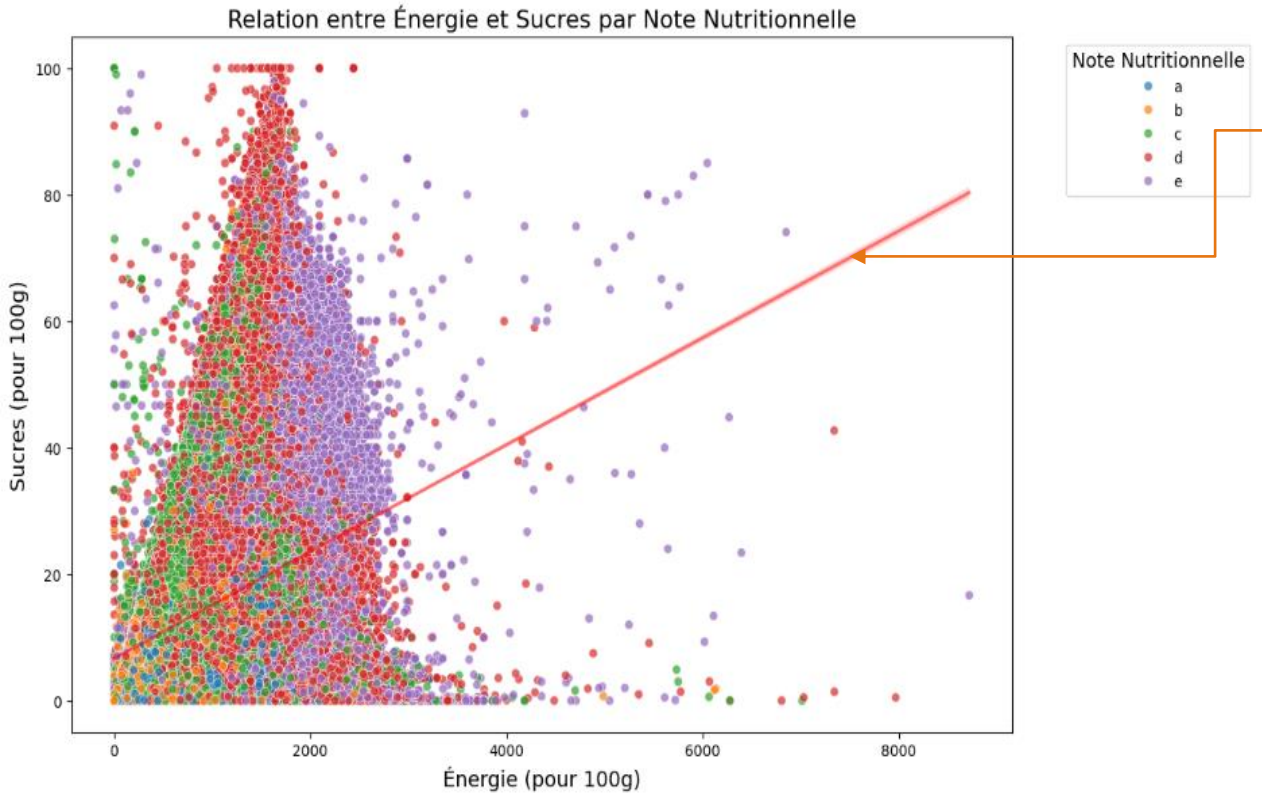
ANALYSE BIVARIÉE



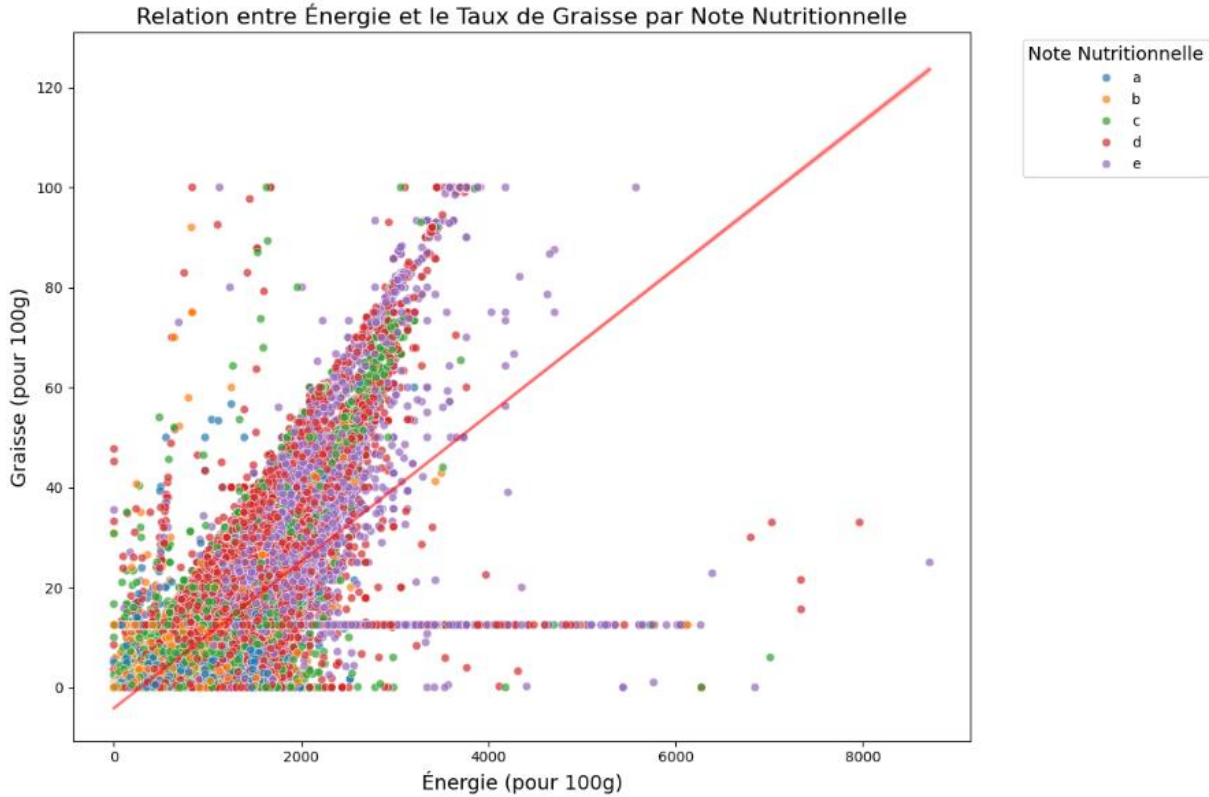


(**A** : Très bon sur le plan nutritionnel. / **B** : Bon. / **C** : Moyen. / **D** : Mauvais. / **E** : Très mauvais.)

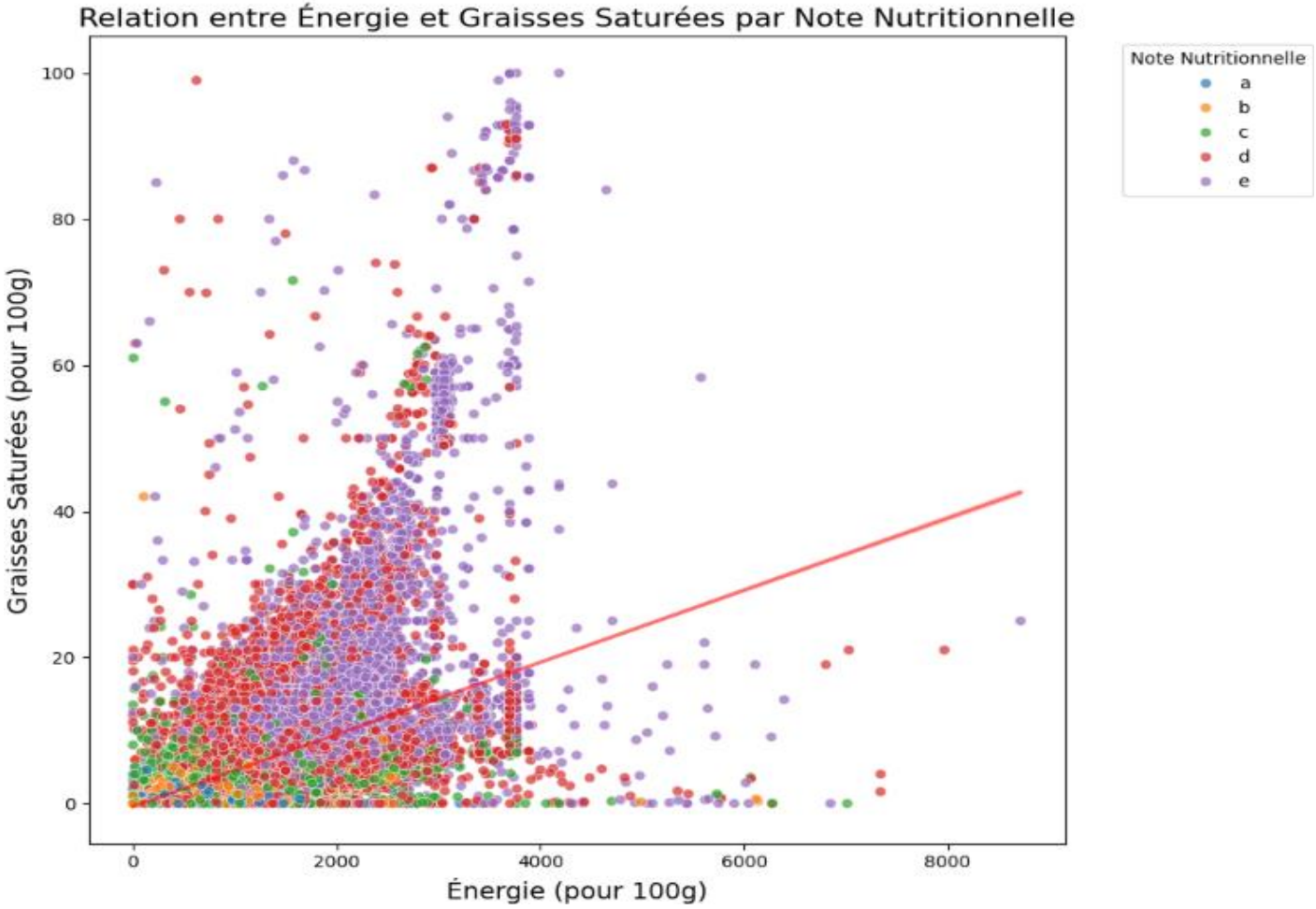
ANALYSE MULTIVARIÉE

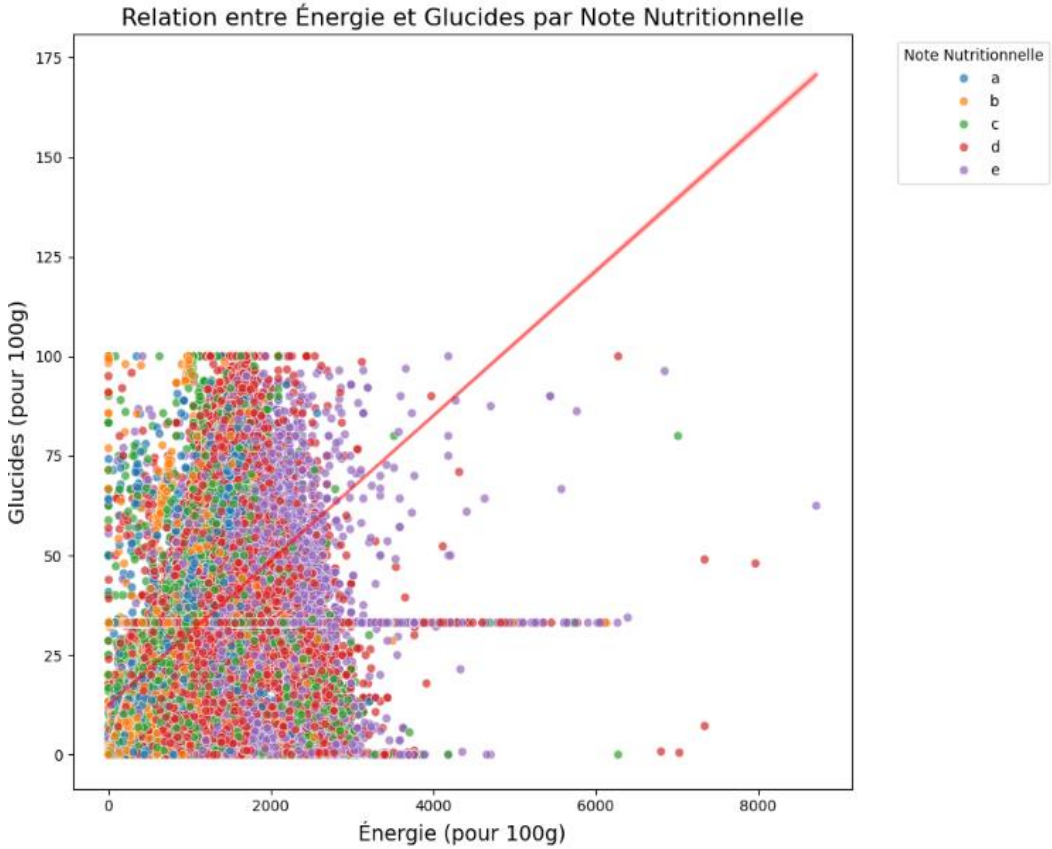


Ligne de regression permettant de visualiser la tendance entre l'énergie et le taux de sucre

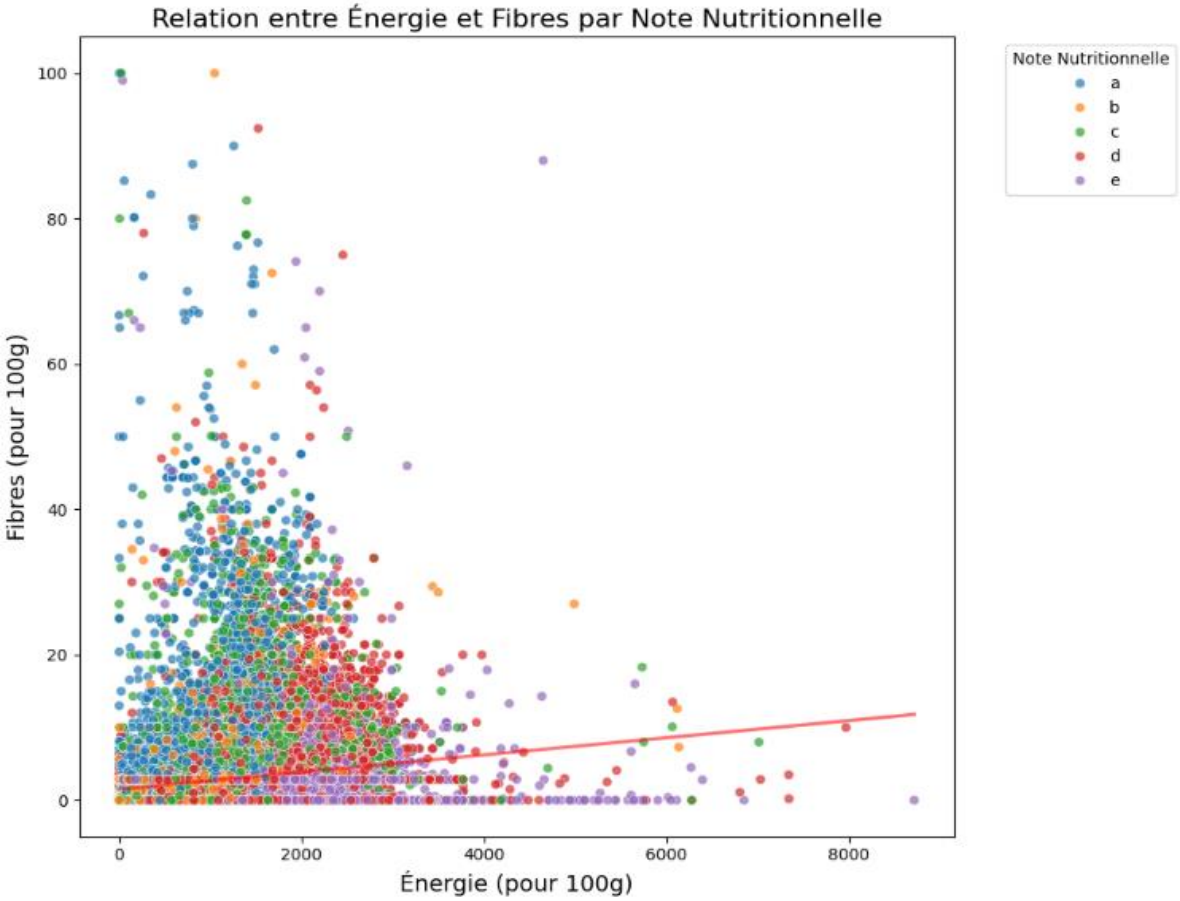


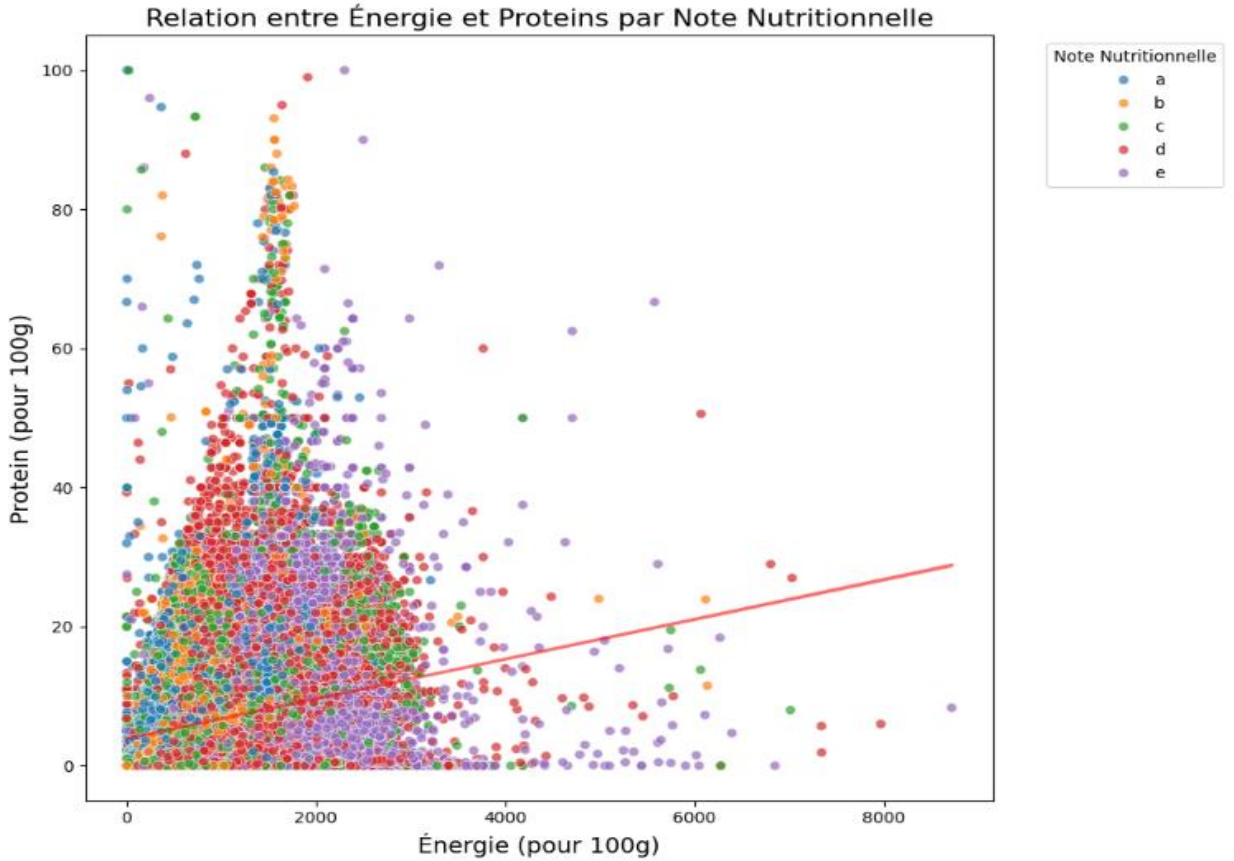
ANALYSE MULTIVARIÉE



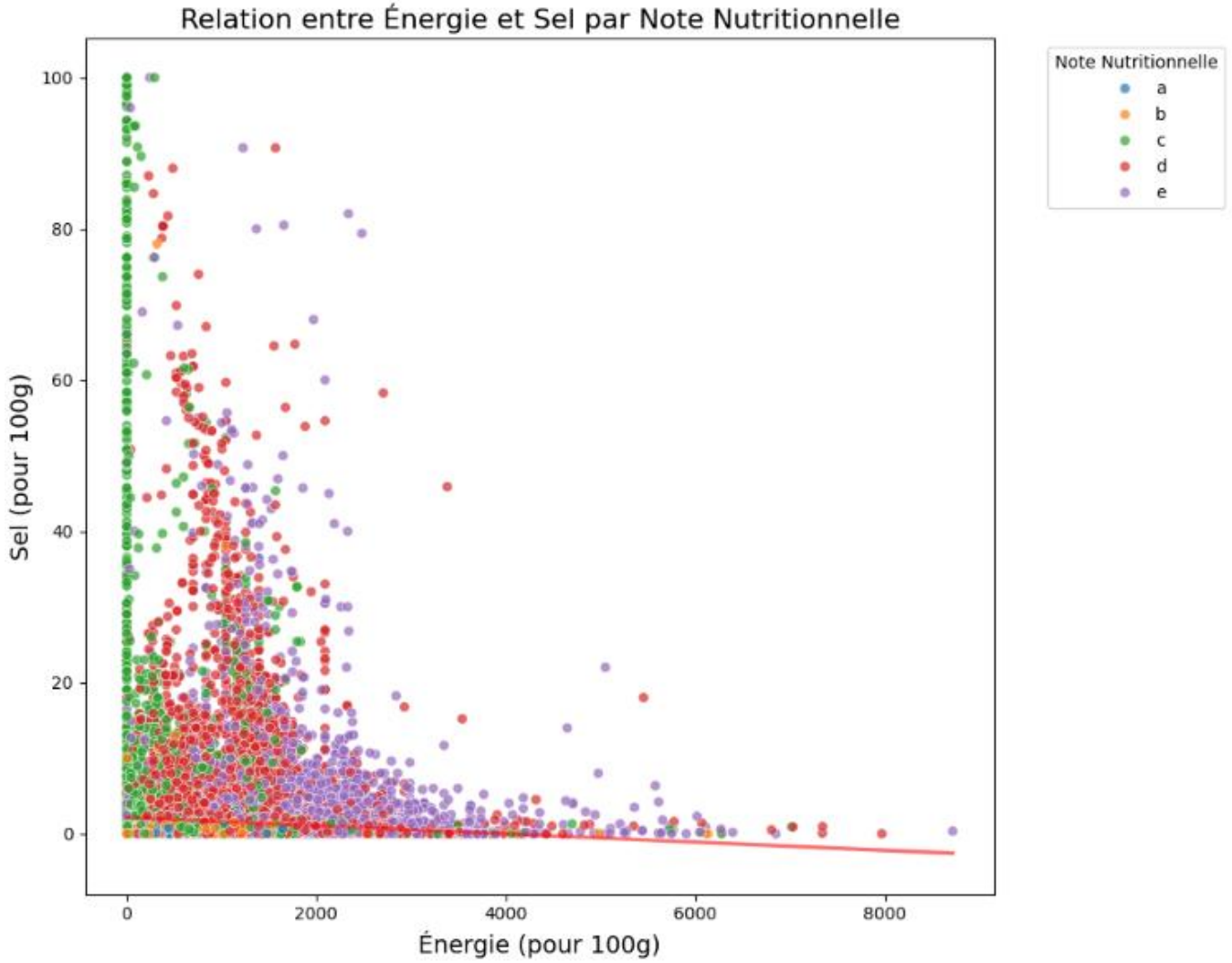


ANALYSE MULTIVARIÉE



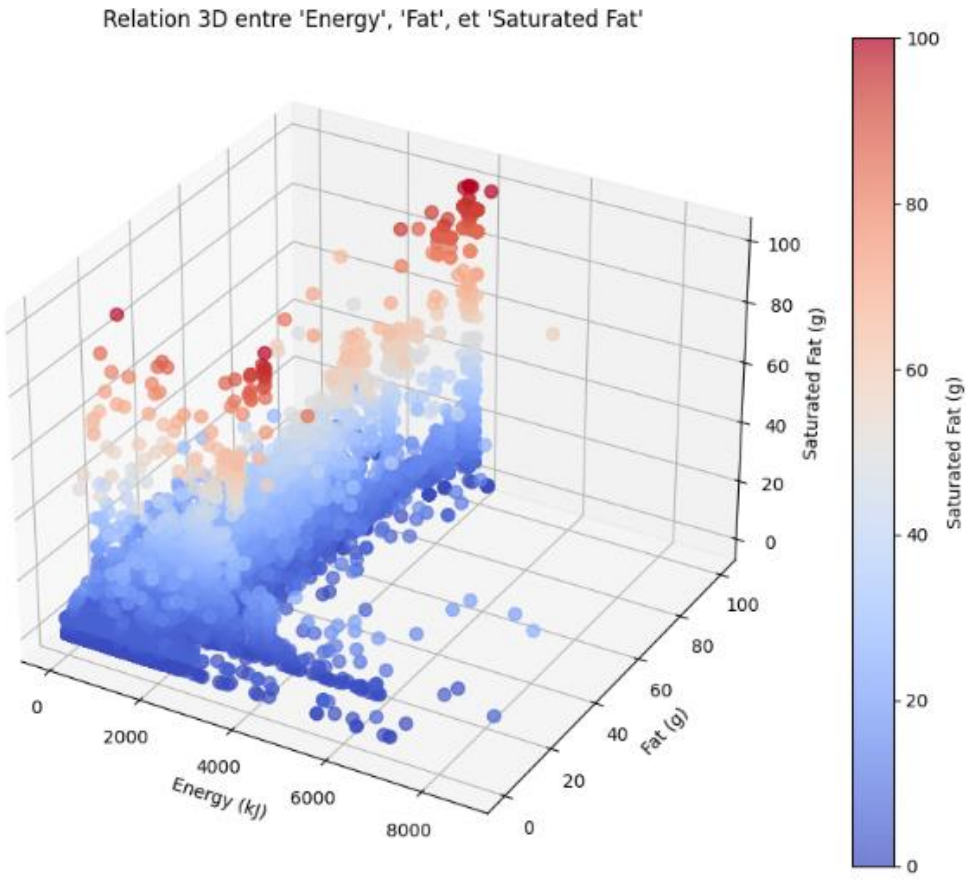


ANALYSE MULTIVARIÉE



ANALYSE MULTIVARIÉE

Relation entre 3 variables

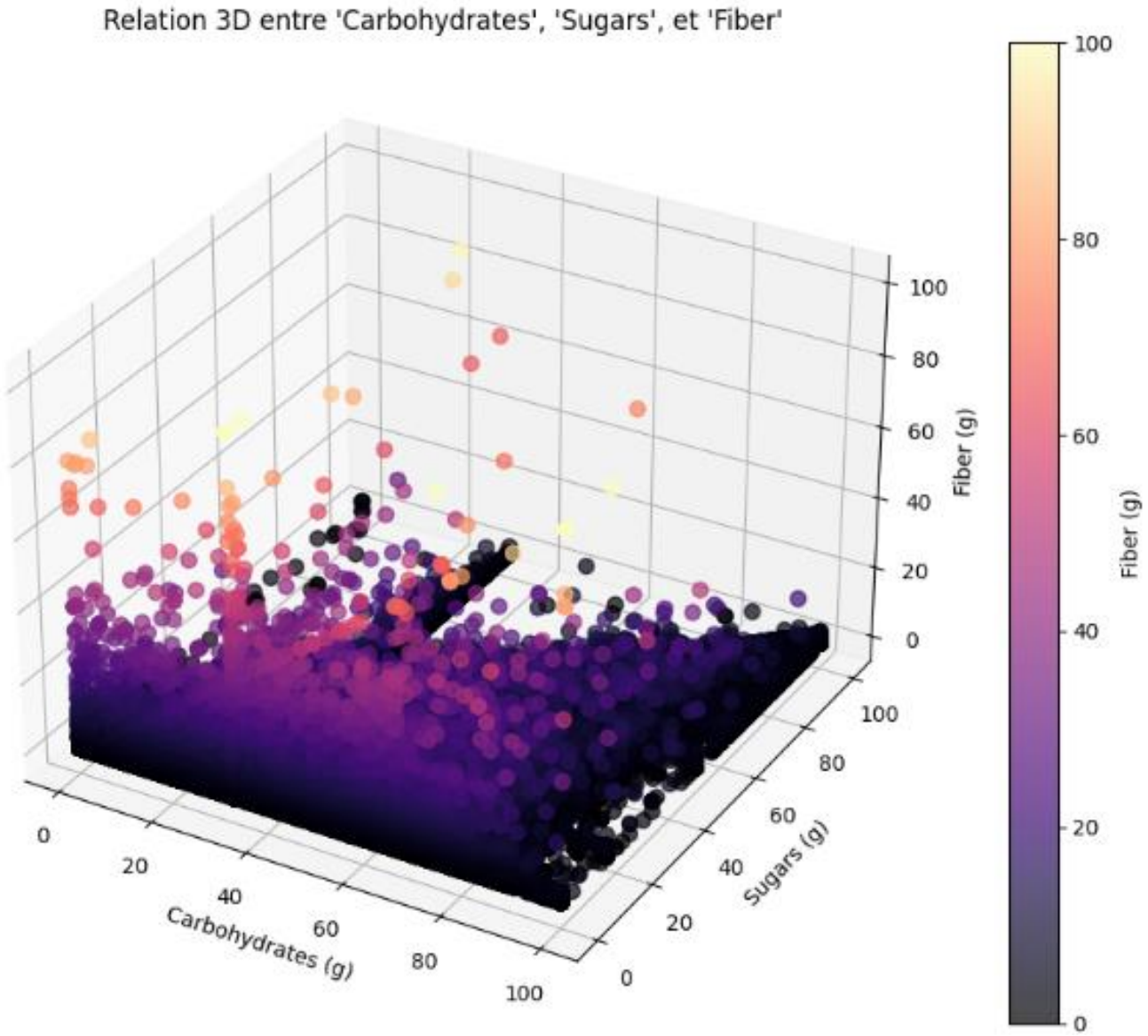


- Le graphique met en évidence la relation conjointe entre la teneur en énergie (KJ), en graisse totale (g), et en graisses saturée (g)
- Plus la teneur **en énergie et en graisses augmente**, plus la quantité de graisses saturées est élevée.
- Les points colorés illustrent cette corrélation.

- Visualisation basée sur un nuage de points coloré selon la teneur en graisses saturées (Saturated Fat).

ANALYSE MULTIVARIEE

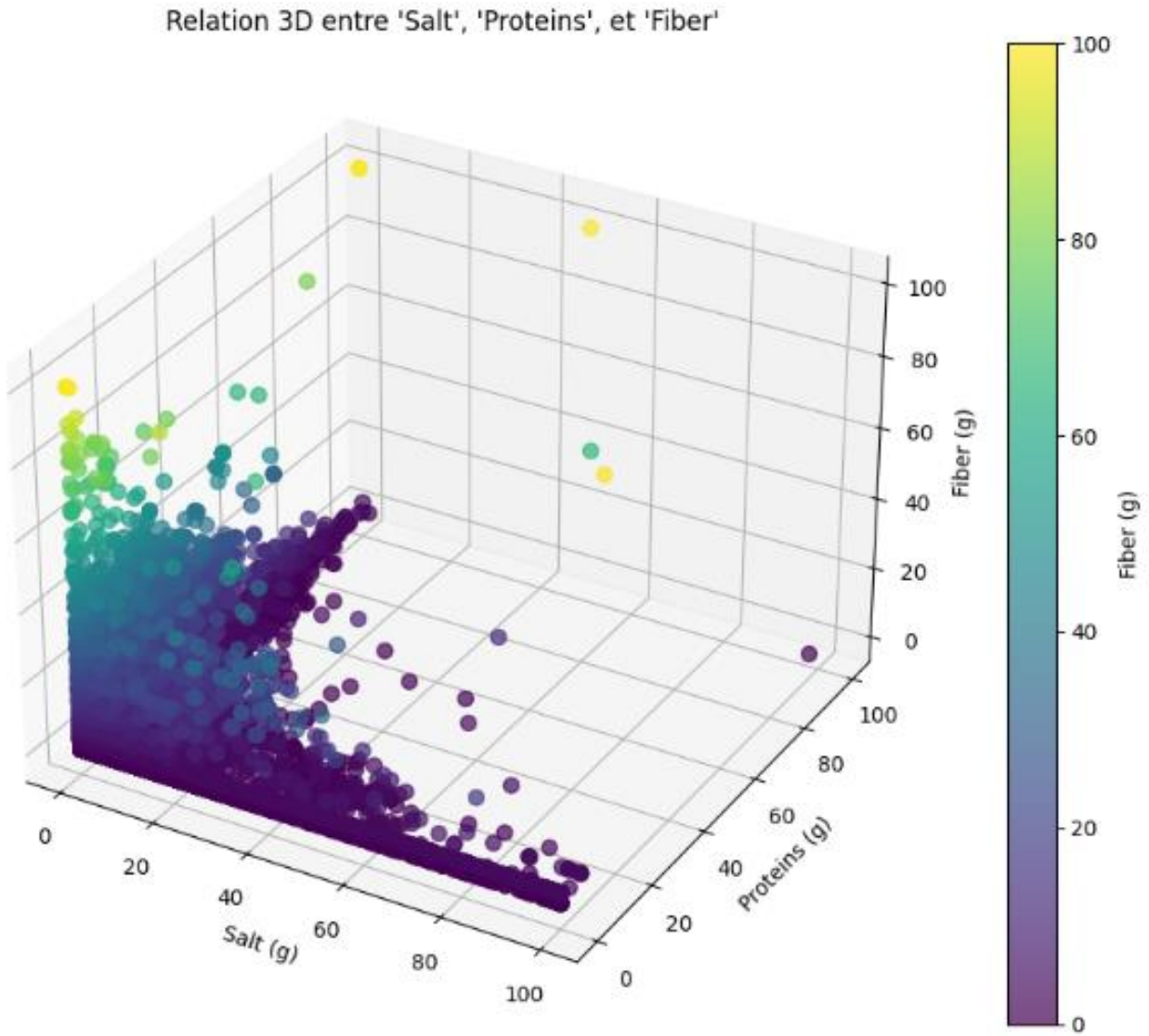
Relation entre 3 variables



- Ce graphique analyse la **relation** entre les **glucides**, les **sucres**, et les **fibres**.
- La palette de couleurs montre que les produits riches en **fibres** sont moins fréquents dans ces données.
- On observe une faible concentration de **fibres** pour des niveaux élevés de **glucides** et de **sucres.**

ANALYSE MULTIVARIEE

Relation entre 3 variables



- Ce graphique analyse la **relation** entre le **sel**, les **protéines**, et les **fibres**.
- La palette de couleurs montre que les produits riches en **fibres** sont moins fréquents dans ces données.
- On constate que les **aliments riches en sel** ont tendance à contenir **peu de fibres**.

ANALYSE MULTIVARIÉE

Analyse en Composantes Principales (ACP)

```
Valeurs manquantes par colonne avant suppression :
 energy_100g               0
fat_100g                   0
saturated-fat_100g         0
carbohydrates_100g         0
sugars_100g                0
fiber_100g                 0
proteins_100g              0
salt_100g                  0
nutrition_grade_fr     63682
dtype: int64
Taille des données avant nettoyage : 221214
Taille des données après nettoyage : 157532

Valeurs manquantes par colonne après suppression :
 energy_100g            0
fat_100g                0
saturated-fat_100g      0
carbohydrates_100g      0
sugars_100g             0
fiber_100g              0
proteins_100g           0
salt_100g               0
nutrition_grade_fr      0
dtype: int64
```

```
Variance expliquée par chaque composante :
 [0.32424523 0.22448912 0.14736178 0.12248594 0.08361582 0.05164922
 0.03745037 0.00870252]
```

I. Standardisation des données

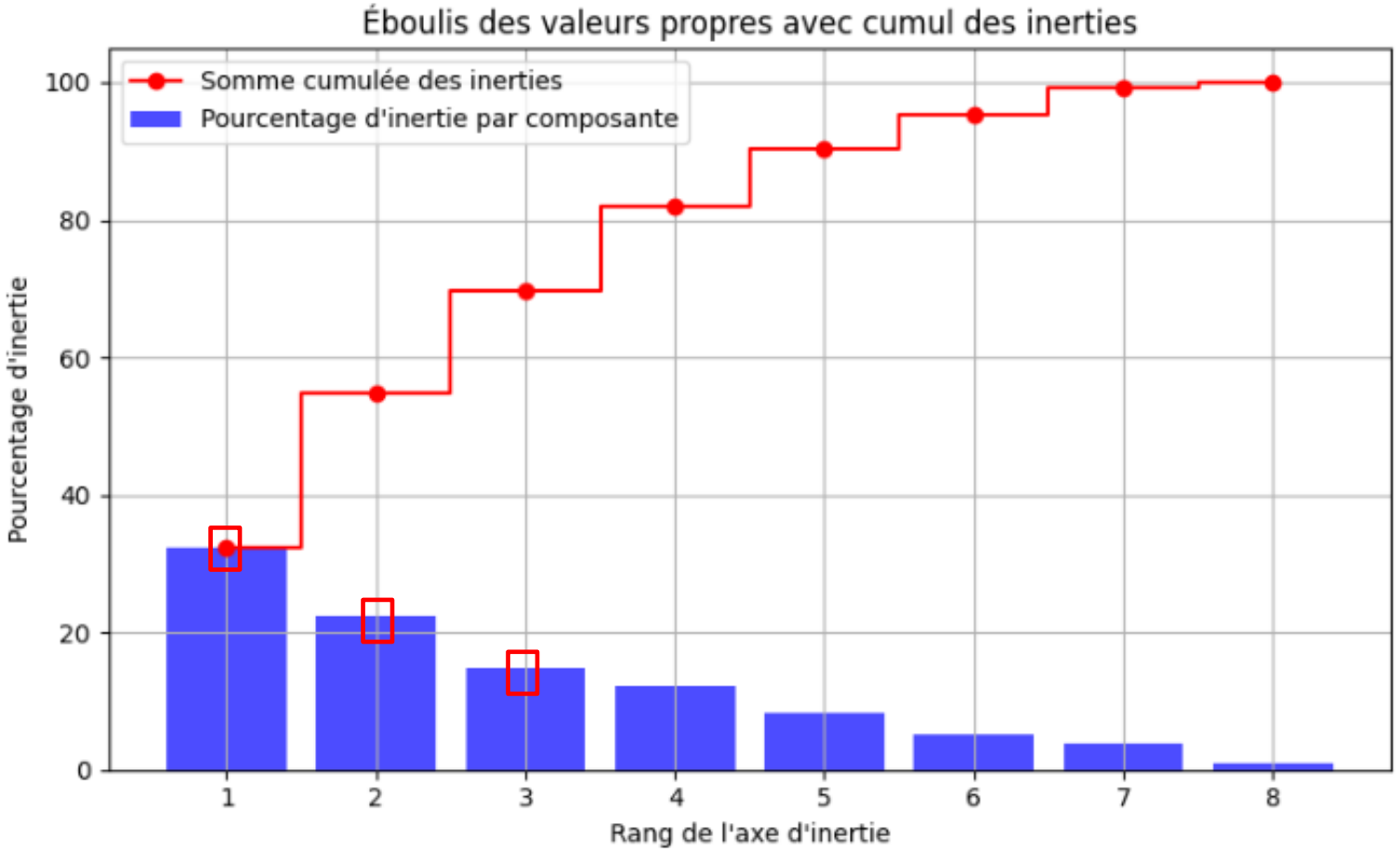
- Toutes les variables ont été centrées et réduites (moyenne = 0, écart-type = 1)
- Objectif : comparer des variables exprimées sur des échelles différentes

II. L'ACP projette les données dans un espace de composantes principales non corrélées

- La première composante explique la plus grande part de la variance, suivie de la deuxième.
- la première composante (F1) explique environ 32% de la variance, et la deuxième (F2) environ 22%.

ANALYSE MULTIVARIEE

Analyse en Composantes Principales – Eboulis des valeurs propres



Les barres (bleues) représentent le pourcentage d'inertie (variance) expliqué par chaque composante principale.

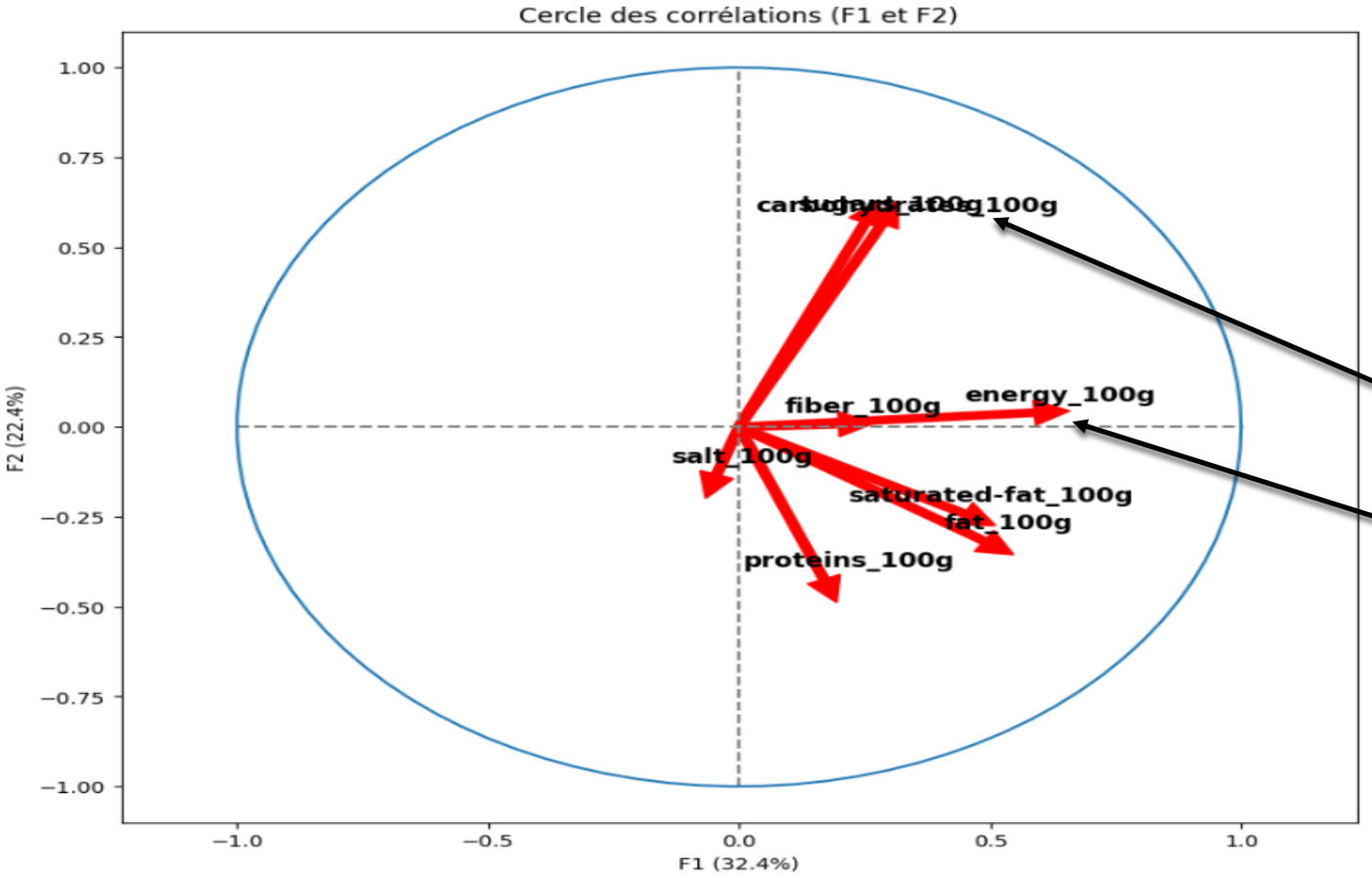
- La **première** composante explique environ **32%** de la variance

- La **seconde** en explique environ **22%**.

Ligne rouge : Affiche la somme cumulée de la variance expliquée.

ANALYSE MULTIVARIEE

Analyse en Composantes Principales – Cercle des correlations (F1, F2)



- Plus une flèche est longue et proche du cercle, plus la variable est bien représentée.
- Des flèches proches entre elles indiquent des variables corrélées.

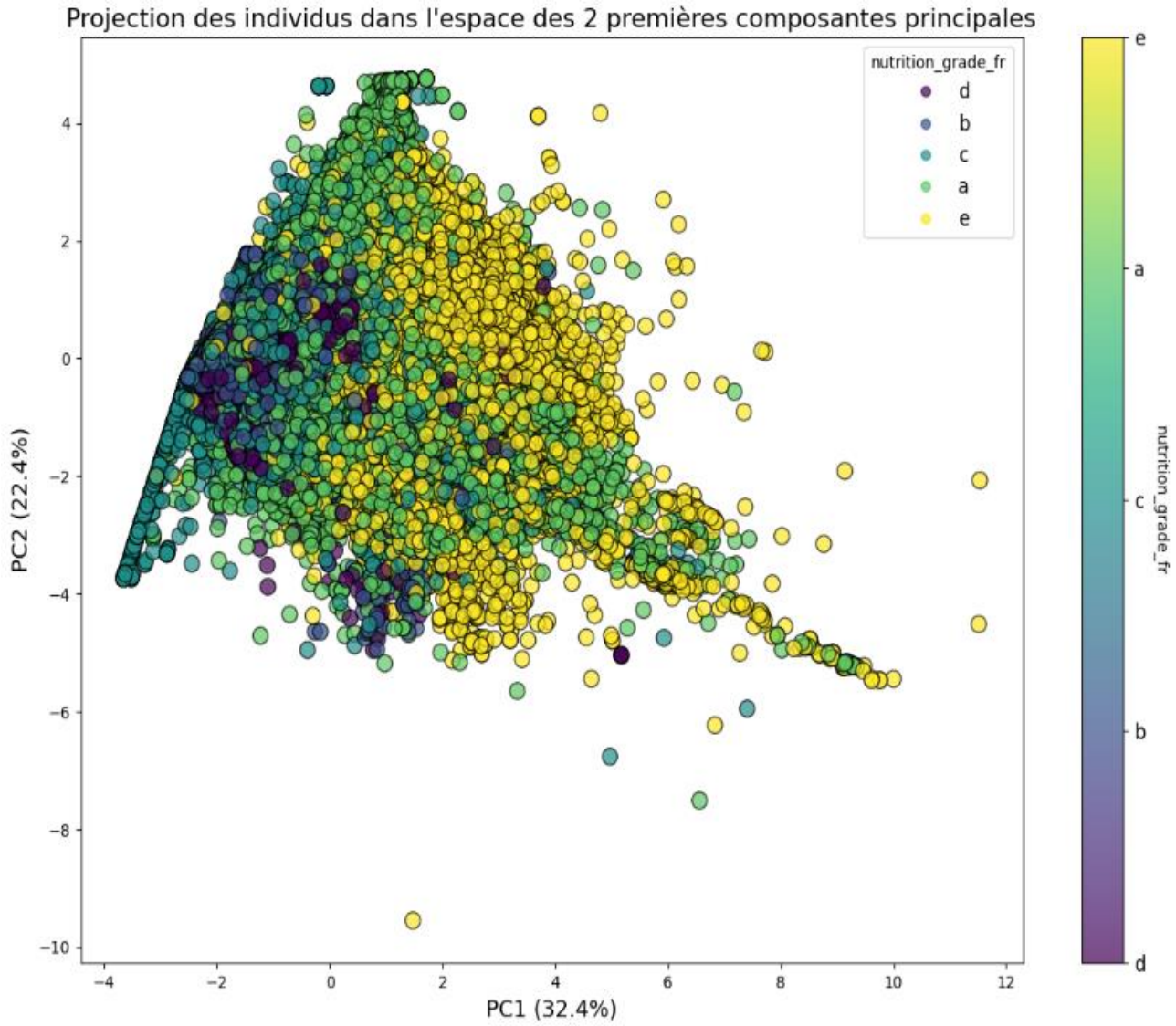
- Le cercle des corrélations visualise la manière dont les variables originales se projettent sur les composantes principales (F1 et F2).
- Il montre la relation entre les variables d'origine et les composantes.

Dans ce cas:

- Les variables energy_100g, sugars_100g sont bien corrélées à la composante F1.
- Ces variables sont donc parmi les plus déterminantes pour expliquer la variance des données et potentiellement prédire le score nutritionnel (nutrition_grade_fr).

ANALYSE MULTIVARIÉE

Analyse en Composantes Principales – **Projection des individus dans l'espace**

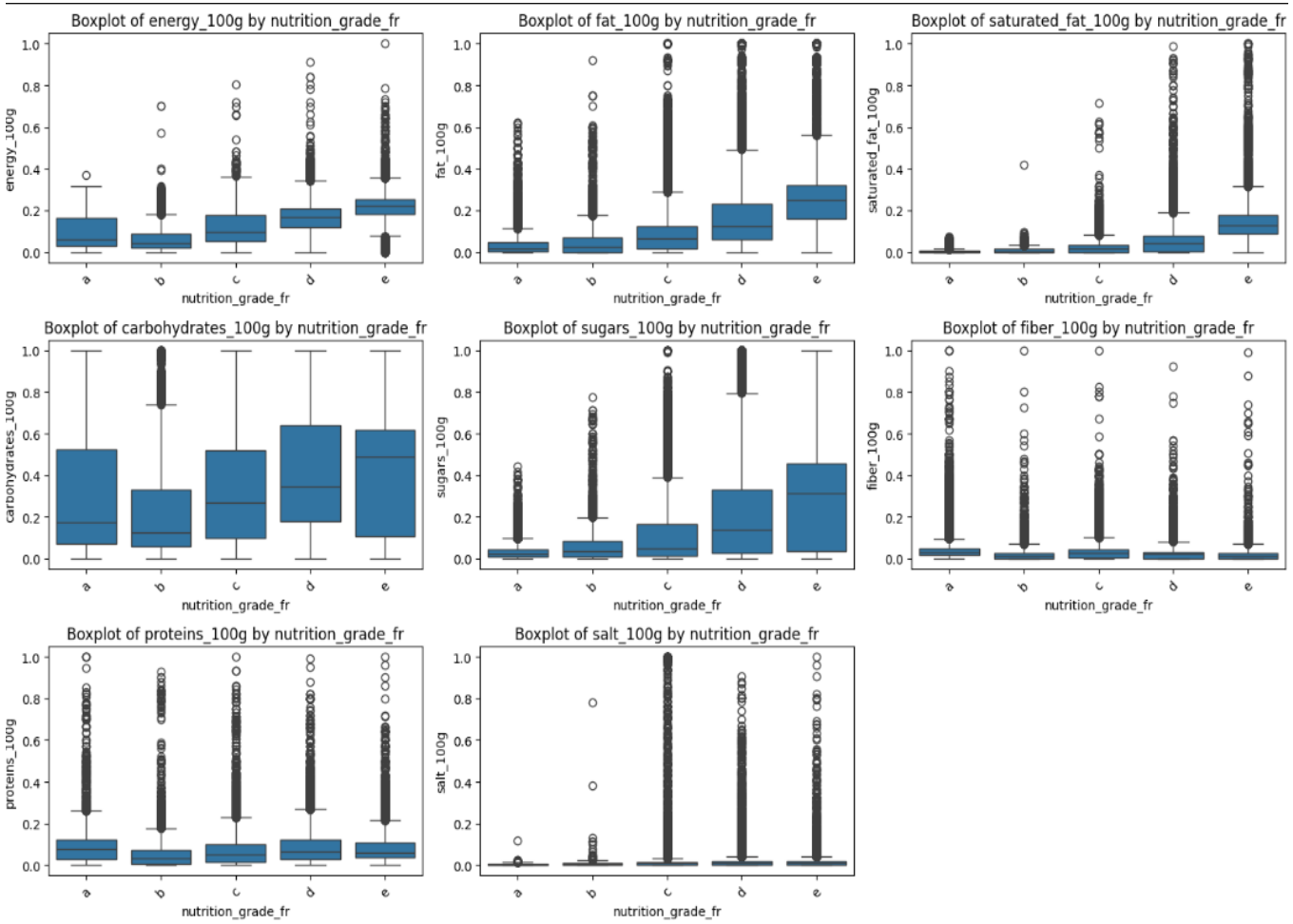


- Chaque point du nuage représente un produit, projeté dans l'espace défini par les deux premières composantes principales (PC1 et PC2).

- La couleur des points correspond à la valeur de la variable `nutrition_grade_fr`, c'est-à-dire le score nutritionnel associé à chaque produit.

ANALYSE MULTIVARIEE
(ANOVA) Analyse de la Variance

L'ANOVA permet de tester si les moyennes d'une variable numérique (par ex. énergie, sucres, graisses…) diffèrent significativement selon les groupes de la variable catégorielle nutrition_grade_fr (notes A à E).



	F-statistic	p-value
energy_100g	23822.012559	0.0
fat_100g	15414.645893	0.0
saturated_fat_100g	27952.187035	0.0
carbohydrates_100g	3304.766667	0.0
sugars_100g	10349.573270	0.0
fiber_100g	1853.191059	0.0
proteins_100g	984.906388	0.0
salt_100g	763.642079	0.0

Une p-value proche de 0 pour chaque variable signifie qu'il existe une différence significative entre les groupes nutrition_grade_fr (de A à E) pour toutes les variables analysées (énergie, graisses, sucres, etc.).

- comme les p-values sont toutes inférieures au seuil critique de 0,05, on rejette l'hypothèse nulle :
- Les moyennes sont différentes selon les groupes nutritionnels.
- Les valeurs élevées des F-statistics confirment que ces différences sont statistiquement fortes.
- Les variables étudiées (énergie, sucres, graisses…) sont pertinentes pour distinguer les niveaux de score nutritionnel.
  Elles joueront donc un rôle important dans la prédiction de nutrition_grade_fr

6. SYNTHÈSE

SYNTHESE

Le **RGPD** (**R**ègleement **G**énéral sur la **P**rotection des **D**onnées) est au cœur de la protection des données à caractère personnel. Il repose sur les principes listés ci-dessous :

- Licéité, loyauté, transparence
  Les données doivent être collectées légalement, avec un objectif clair et connu des utilisateurs.
- Limitation des finalités
  Elles sont utilisées uniquement pour les objectifs définis à l'avance.
- Minimisation des données
  Seules les données strictement nécessaires sont traitées.
- Exactitude
  Les données doivent être justes et mises et à jour.
- Limitation de la conservation
  Elles ne doivent pas être conservées plus longtemps que nécessaire.

RESPECT DU PRINCIPE RGPD LORS DU NETTOYAGE DE LA BASE DE DONNEES OPEN FOOD FACTS

- **La base ne contient aucune information permettant d'identifier une personne.**
- **Seules des données techniques anonymisées ont été utilisées. Aucun traitement de données sensibles ou personnelles n'a été réalisé.**

SYNTHESE

- La base de données **Open Food Facts** nous permet de conclure que les produits qui y sont intégrés ne sont pas majoritairement des produits sains et qu'elle a fait l'objet d'un important travail de nettoyage (valeurs manquantes, aberrantes).

- Une majorité de produits sont classés avec un score nutritionnel défavorable (D et E), ce qui confirme la nécessité de mieux informer les consommateurs.

- Une analyse exploratoire, combinée à une analyse en composantes principales (ACP), a permis d'identifier les variables les plus discriminantes pour expliquer les différences nutritionnelles :
  → notamment energy_100g, sugars_100g et carbohydrates_100g.

- L'ACP met en évidence **sugars_100g** et **energy_100g** comme les variables les plus liées à la prédiction de **nutrition_grade_fr**, avec une forte contribution à la composante F1.

Elles sont situées positivement dans le cercle des corrélations concernant la prédiction de la **nutrition_grade_fr**