Santé publique France

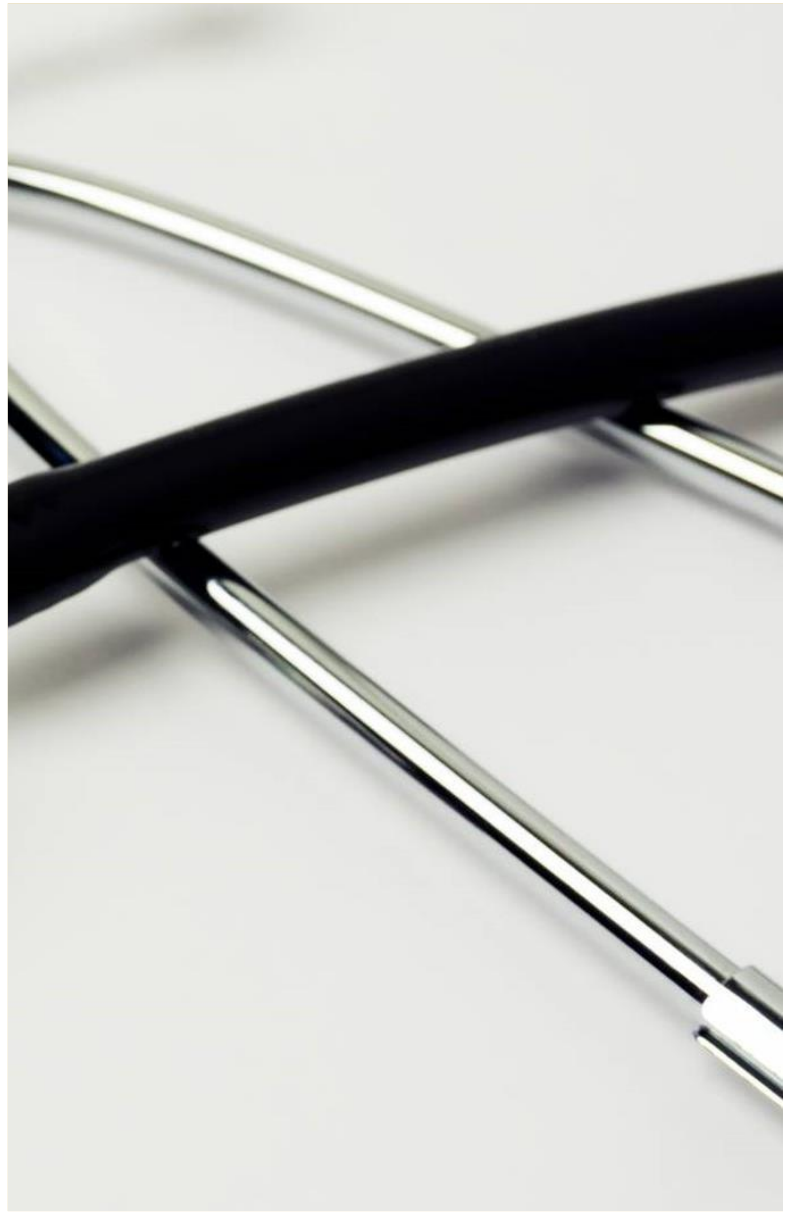
OPEN FOOD FACTS
l'information alimentaire ouverte

PROJET #3

CRÉATION DE L'APPLICATION "**HEALTH_AUTOFILL**" POUR L'AGENCE **SANTE PUBLIQUE FRANCE**

AGENDA

1. INTRODUCTION
2. PROBLEMATIQUE
3. EXPLORATION DES DONNEES
4. NETTOYAGE
5. ANALYSE DES DONNEES
6. SYNTHESE



I. INTRODUCTION

La base de données **Open Food Facts**

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

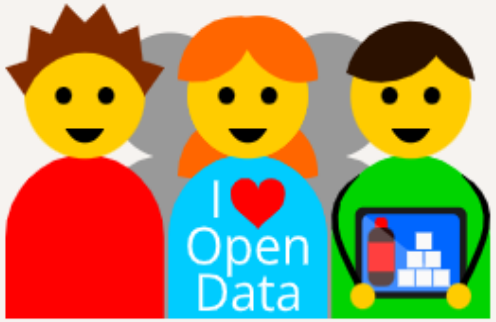
Base de données publiques

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

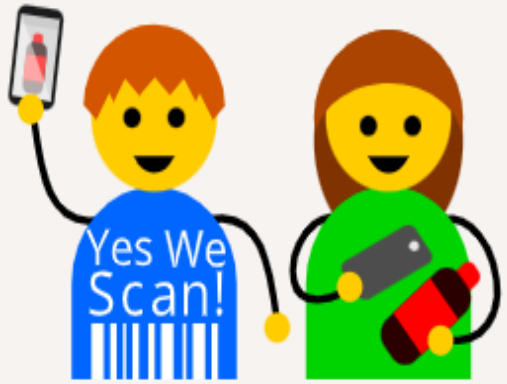
Création: 2012



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.

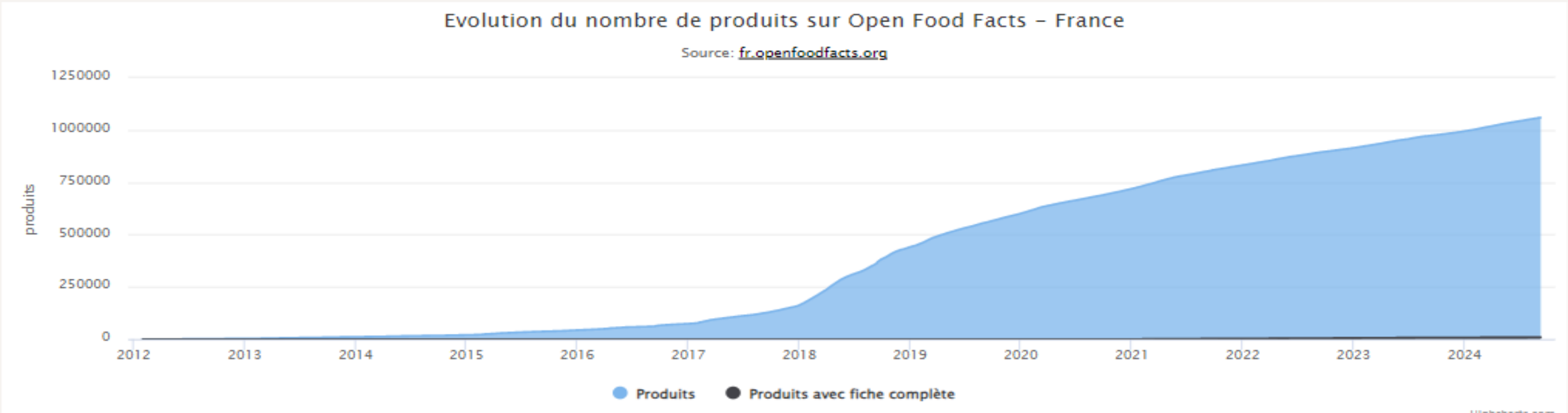
Plus de 9000 contributeurs comme vous ont ajouté 600 000 produits de 200 pays en utilisant notre app Android, iPhone ou Windows Phone ou leur appareil photo pour scanner les codes barres et envoyer des photos des produits et de leurs étiquettes.

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** grâce à leurs critères et d'effectuer des comparaisons.
- Consulter des informations détaillées sur les **ingredients** et **additifs** contenus pour chacun des produits de la base de données.
- Ci-dessous un graphique montrant l'évolution du nombre de produit disponible dans la base de données d'**Open Food Facts** depuis sa création en **2012** :



2. PROBLEMATIQUE

HEALTH_Autofill

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

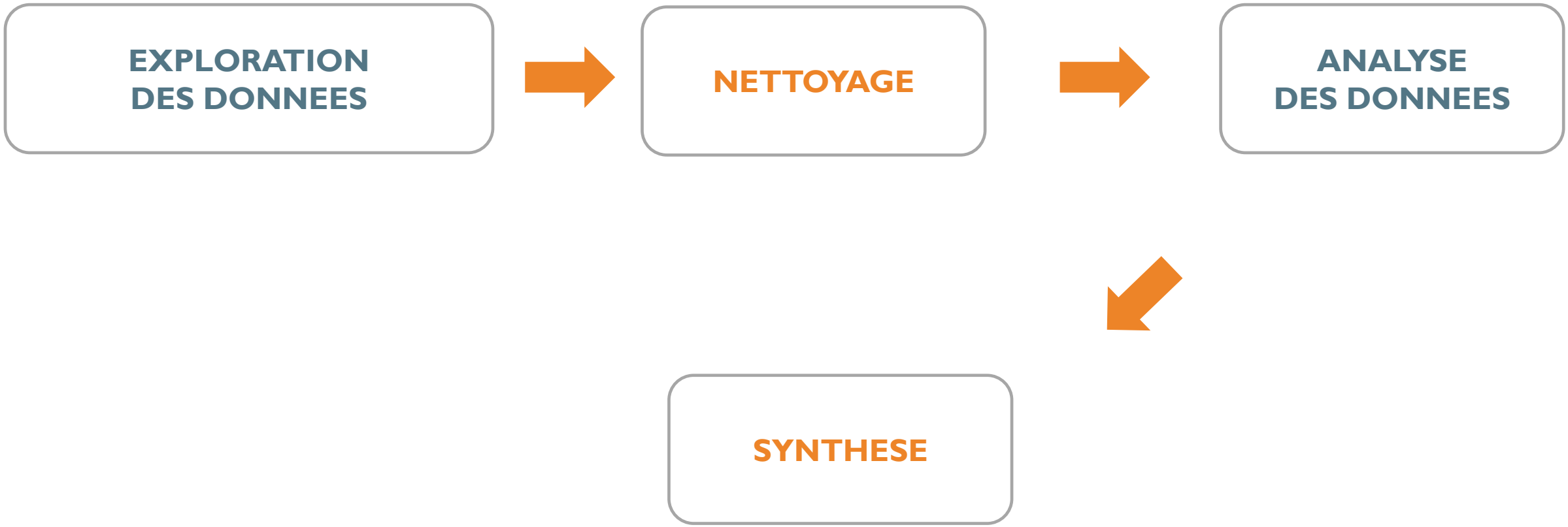
- **La saisie** de données :
  textuels et numériques

Santé publique France

Création d'une application dont l'objectif est de prédire des valeurs manquantes notamment pour l'information nutrionnelle d'un produit

/ Erreur de saisie potentielle
/ Valeurs manquantes potentielles

HEALTH_Autofill



EXPLORATION DES DONNÉES

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

est une base de données volumineuse: **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.

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

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

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

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

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

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

Des colonnes
NUMERIQUES:

Elles correspondent à la typologie des **ingredients** pour **100g**
(Exemple : **fructose_100g** / **lactose_100g**)

Des colonnes
CATEGORIELLES:

Elles correspondent aux informations **textuelles** des produits tells que les variables
(**categories_fr** / ou bien **nutrition_grade_fr**)

Le pourcentage de valeur **manquante pour l'ensemble des données de la base Open Food Facts est important**

Le choix de la variable à prédire dans
la base de données **Open Food Facts** :

La variable cible correspond à "**nutrition_grade_fr**" (la note nutritionnelle de chaque produit):

A PROPOS DE LA VARIABLE "nutrition_grade_fr":

- Variable **textuelle** (Object)

- **5 valeurs uniques** :

(**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, aidant les consommateurs à faire des choix alimentaires plus sains de manière rapide et simple.

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

Les variables numériques ci-dessous qui contiennent un taux de valeurs manquantes **inférieurs à 50%** et seront l'objet de la prediction 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 PREDIRE : nutrition_grade_fr

NETTOYAGE

LES ETAPES DU NETTOYAGE DES VARIABLES CIBLES:

La création du fichier plat avec les variables cibles :

LES VALEURS ABERRANTES

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

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

LES VALEURS MANQUANTES

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

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

LA CREATION DU DATAFRAME CONTENANT LES VARIABLES CIBLES

	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

D) VISUALISATION DU JEU DE DONNEE AVANT LE NETTOYAGE

Statistiques Descriptives:

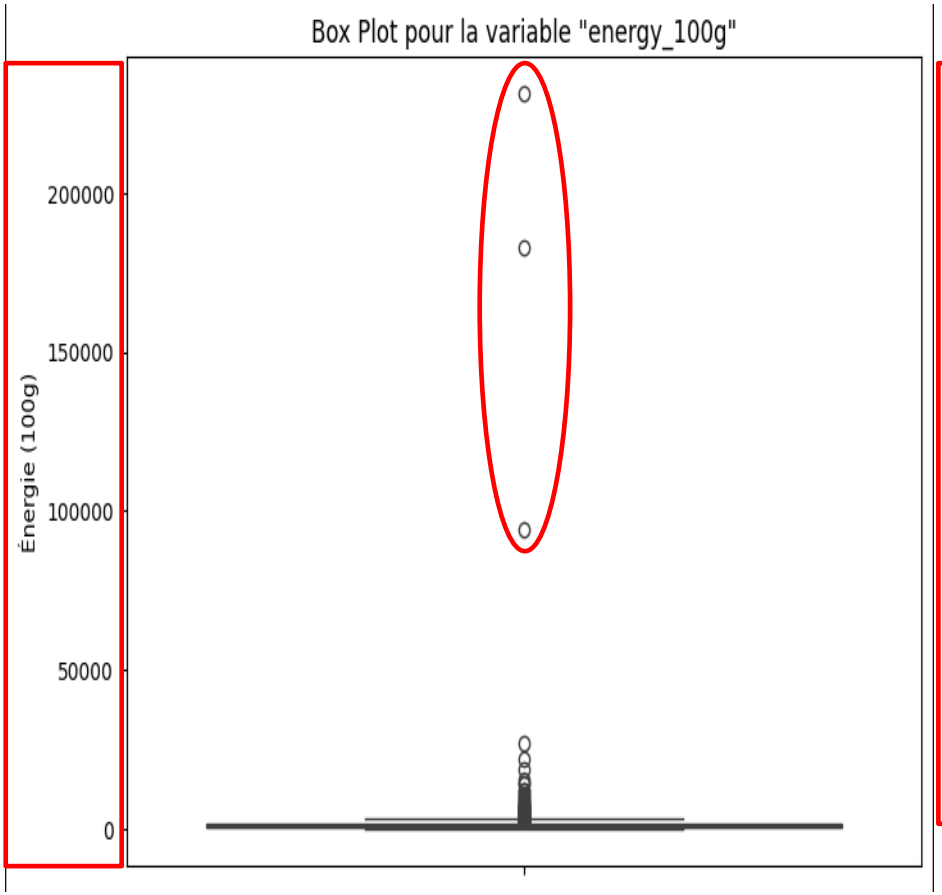
	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	\
count	184470.000000	169999.000000	162921.000000	169730.000000	
mean	1139.314255	12.497010	5.195458	33.116896	
std	1066.902948	16.476978	8.014654	29.981925	
min	0.000000	0.000000	0.000000	0.000000	
25%	418.000000	0.100000	0.000000	6.670000	
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	

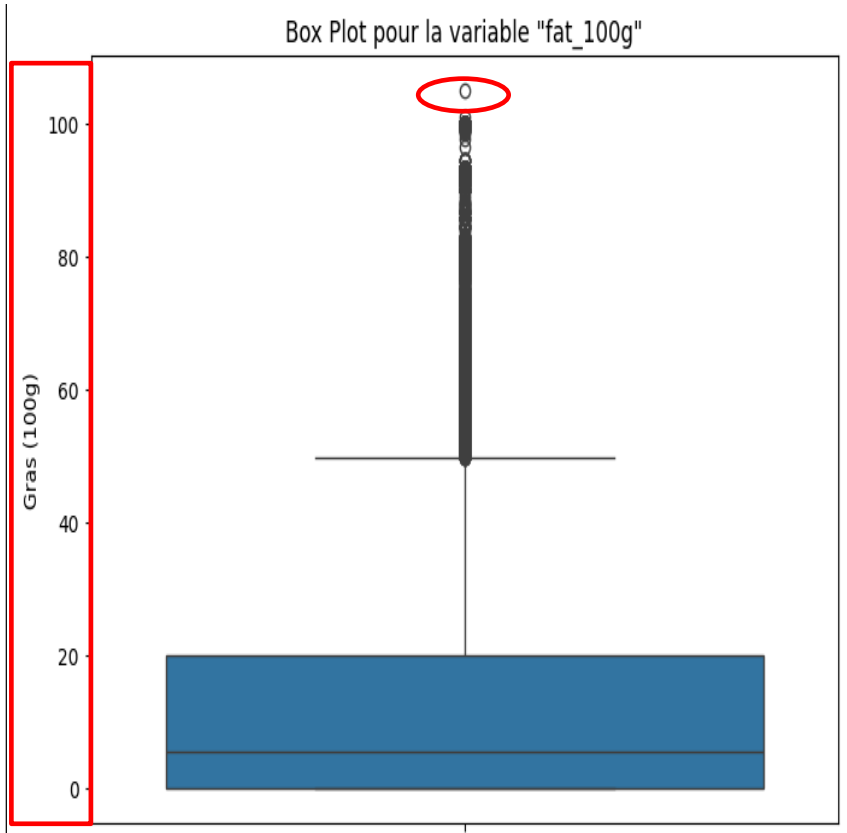
	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	173677.000000	141981.000000	183560.000000	180138.000000
mean	16.329856	2.898030	7.139842	2.080937
std	22.700836	15.016186	8.417052	152.389088
min	-6.250000	-6.700000	-800.000000	0.000000
25%	1.420000	0.000000	0.820000	0.071120
50%	6.190000	1.500000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

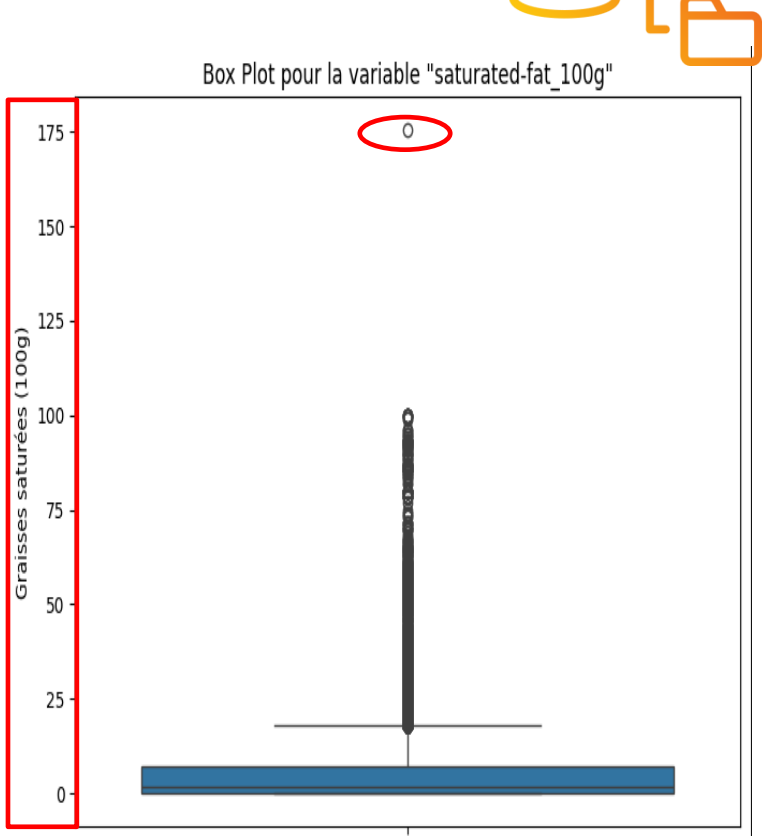
1. Présence de valeur **> 100g et < 0** => Mathématiquement **impossible**

2. Présence de valeur **> 9000 Kilojules** pour la variable **energy_100g** => Mathématiquement **impossible**

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



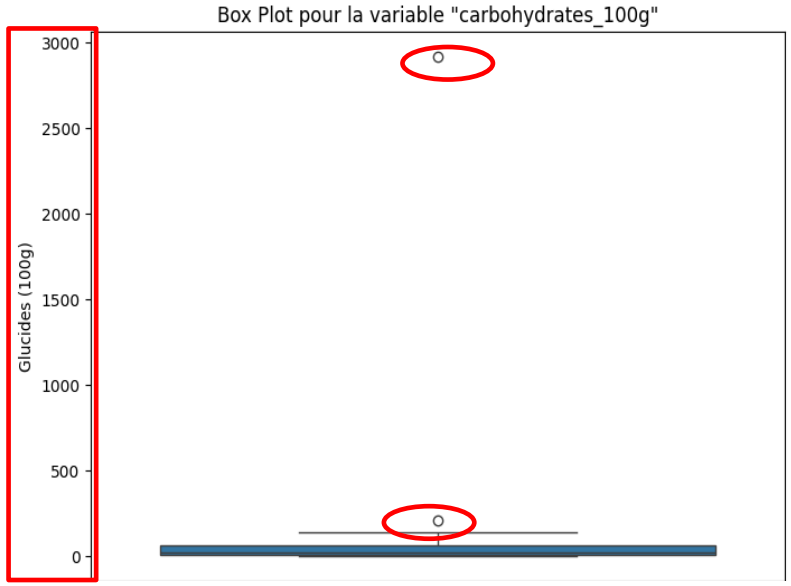


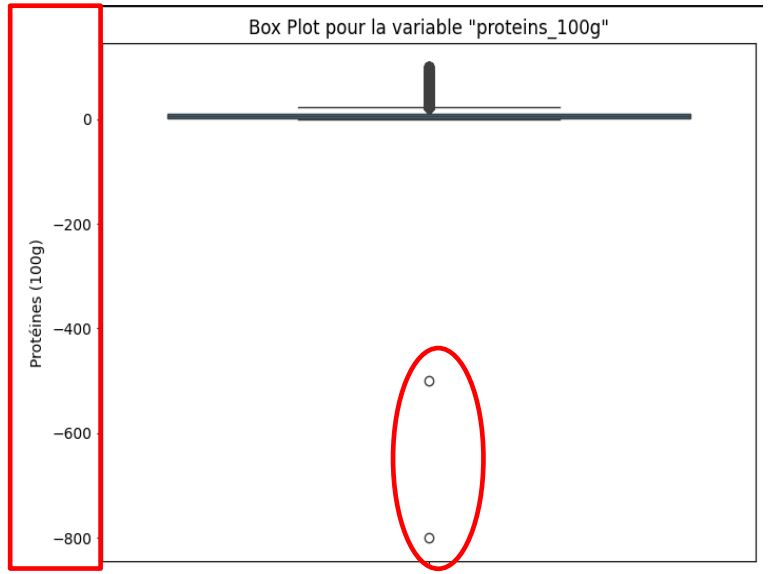


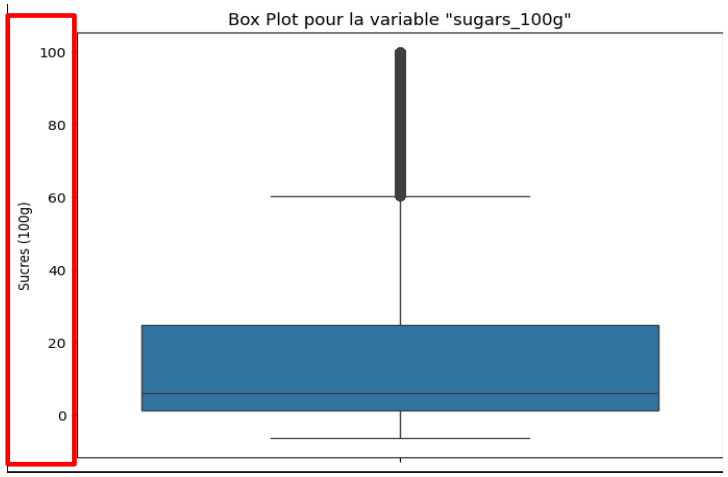
1.Valeurs Energétique Typiques pour les Aliments quelques repères généraux :
Huiles et graisse : Environ 3 500 à 4 000 kJ pour 100 g (environ 800 à 950 kcal). Viandes et poissons : Environ 500 à 1 000 kJ pour 100 g (environ 120 à 240 kcal). Fruits et légumes : Environ 100 à 500 kJ pour 100 g (environ 24 à 120 kcal). Produits céréaliers : Environ 1 500 à 2 500 kJ pour 100 g (environ 360 à 600 kcal).
APRES ANALYSE DE LA QUANTITE DE VALEUR ABERRANTE NOUS DECIDONS D'OPTER PAR UNE APPROCHE DE TRAITEMENT DES VALEURS ABERRANTES METIER, CELA SIGNIFIE D'AFFICHER LES VALEURS ABERRANTES (VALEUR POUR LESQUELLES LA QUANTITE DE KILOJULES EST SUPERIEUR ET 9.000 KILOJULES ET INFERIEUR A 0 KILOJULES) ET **DE SUPPRIMER LES VALEURS QUI SONT SUPERIEURS A 9000 KILOJULES ET INFERIEURS A 0 KILOJULES**

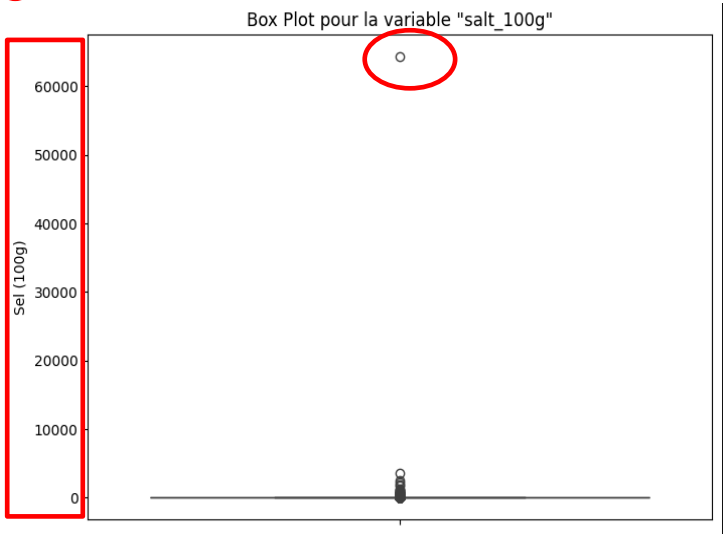
LES VALEURS ABERRANTES DES VARIABLES CIBLES

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 QTE produit pour 100 grammes
> à 100g = SUPPRESSION

Si la QTE produit pour 100 grammes
< à 0g = SUPPRESSION

OUTLIERS QUI ONT ÉTÉ SUPRIMES DE LA BASE DE DONNEES Open Food Facts

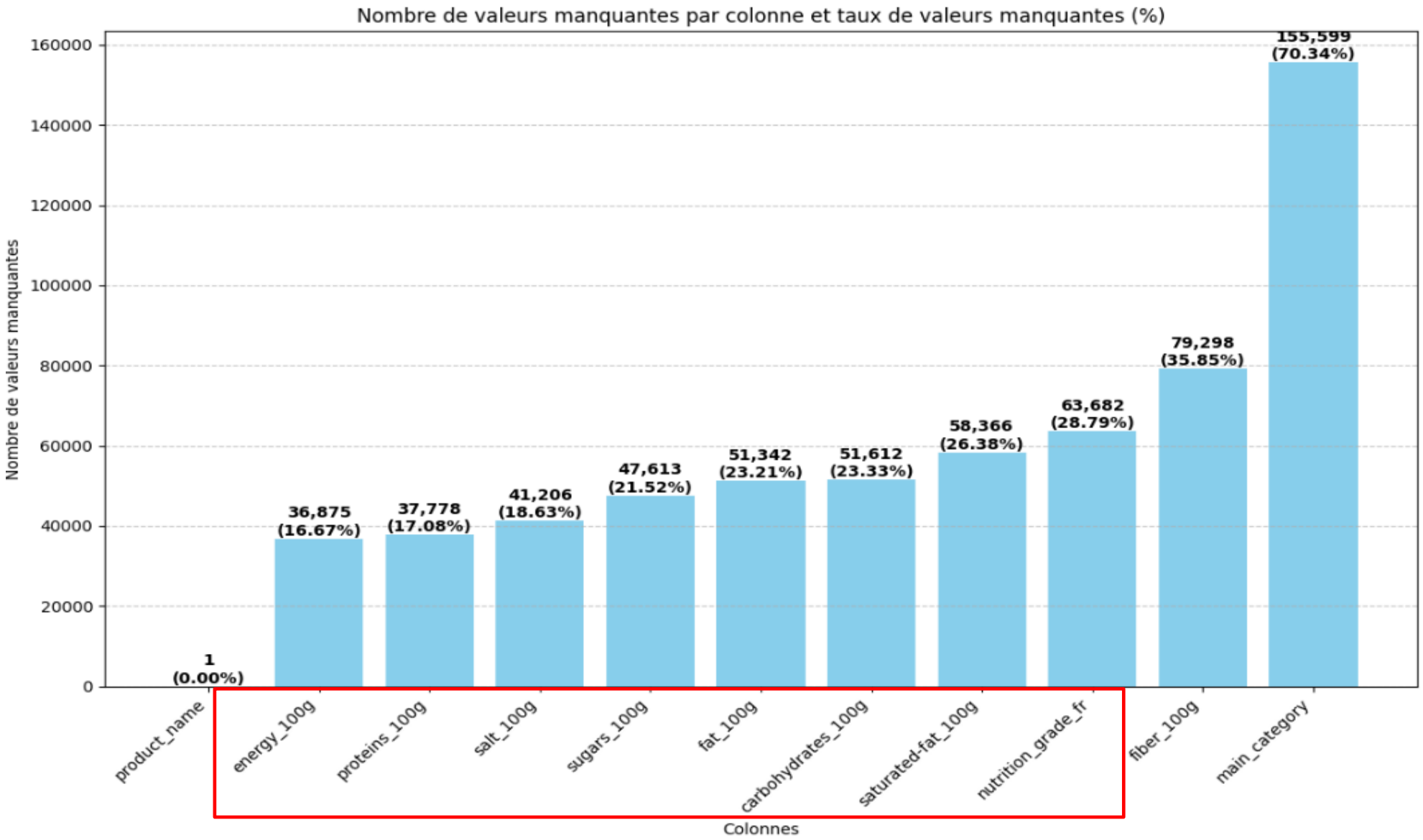
	product_name	fat_100g
209593	Ekstra Jomfru Olivenolie	101.0
210931	Graine de couscous moyen	105.0

	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 METIER
&
A) AFFICHAGE DE VALEURS MANQUANTES POUR CHACUNE DES 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

SI TAUX DE SUCRE
Sugars_100g = 100g

ALORS IMPUTER

Proteins_100g = 0 I 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

B) TRAITEMENT DES VALEURS MANQUANTES

REMPLACEMENT DES VALEURS MANQUANTES PAR LOGIQUE

SI TAUX DE GRAISSE
Fat_100g = 100g

ALORS IMPUTER

Sugars_100g = 0 I Proteins_100g = 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	

	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

C) TRAITEMENT DES VALEURS MANQUANTES

REMPLACER LE RESTANT DES VALEURS MANQUANTES PAR LA MOYENNE DE CHAQUE COLONNE

```
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
```

D) VISUALISATION DU JEU DE DONNEE NETTOYE

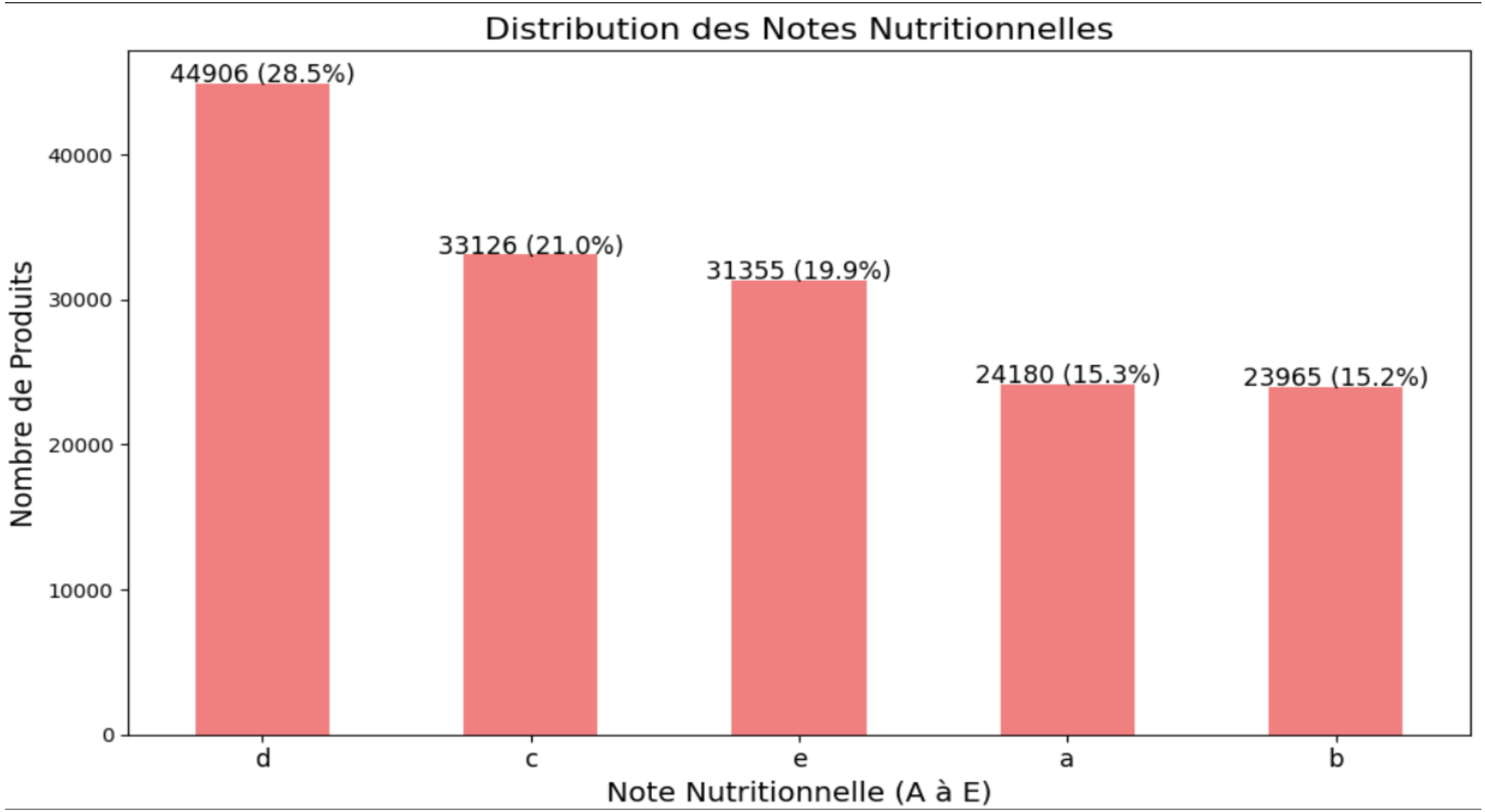
	index	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000	221214.000000
mean	162051.289430	1135.982292	12.496463	5.190437	33.104861	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. Pas de valeur **> à 100g**
2. Pas de **> à 0g** pour toutes les variables cibles sauf pour **energy_100g** qui est exprimé en kilojules
3. Pas de valeur **> à 9000 KILOJULES** pour la variable energy_100g

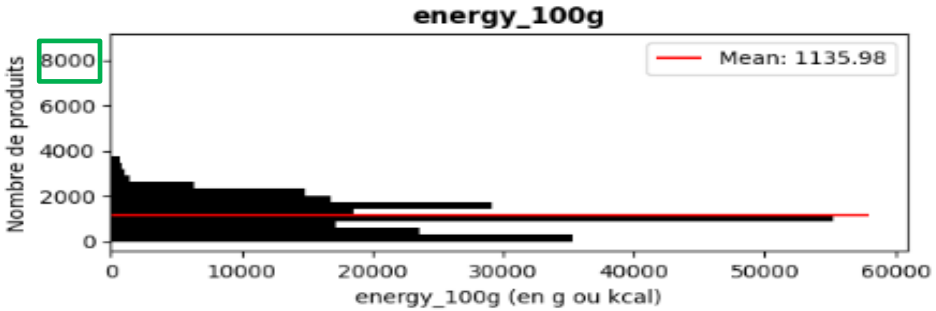
ANALYSE DES DONNEES

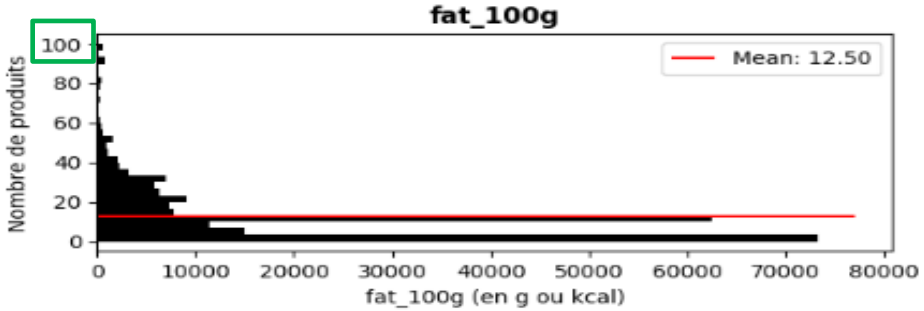
ANALYSE UNIVARIEE

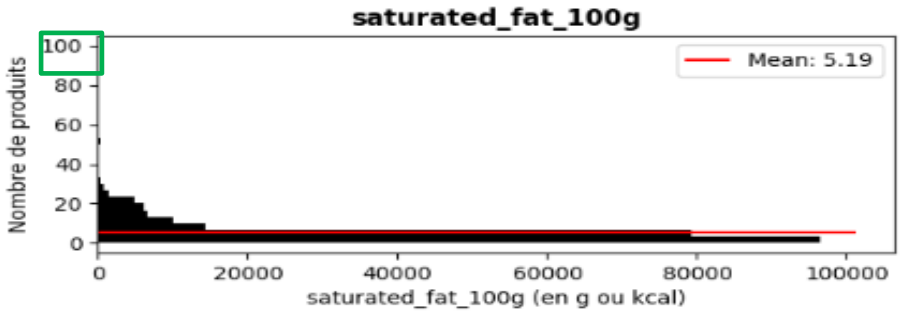
Variable : nutrition_grade_fr

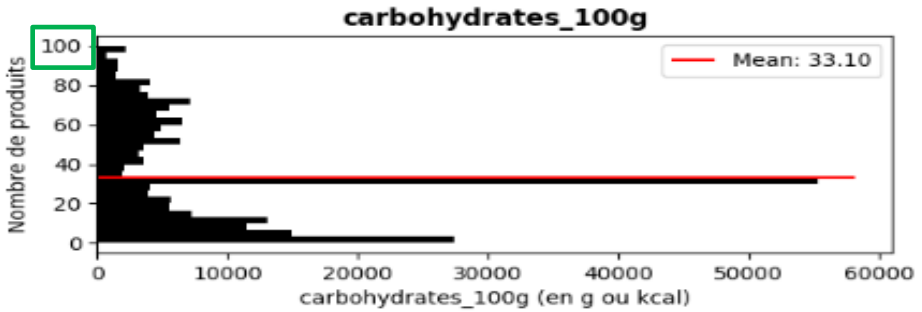


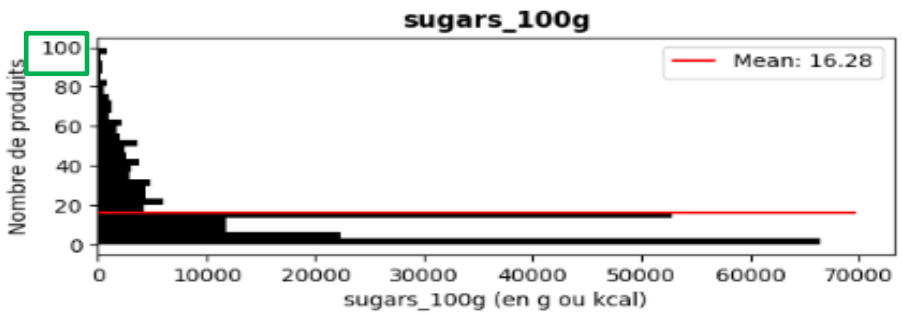
ANALYSE UNIVARIEE

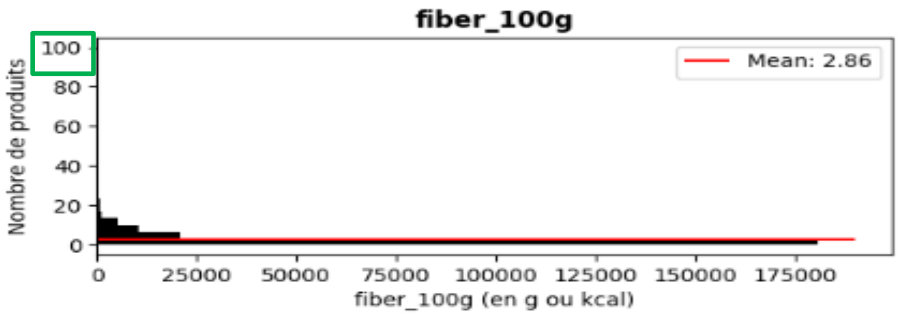


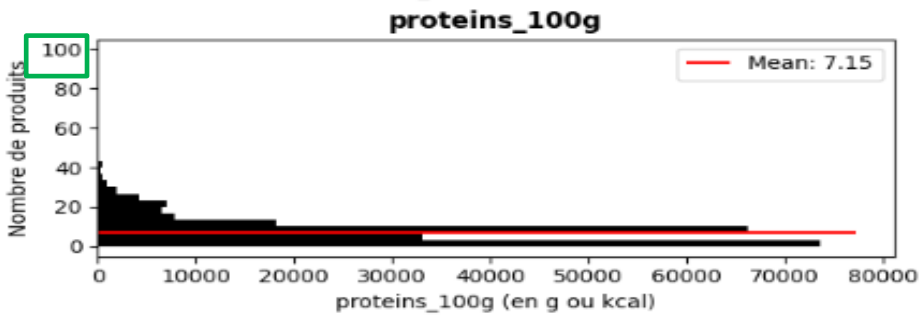


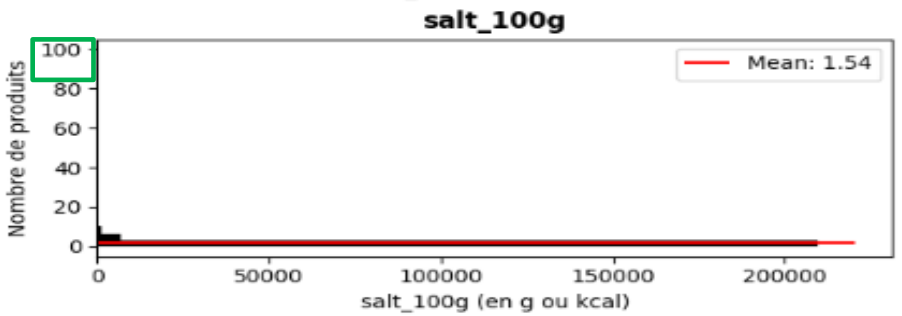






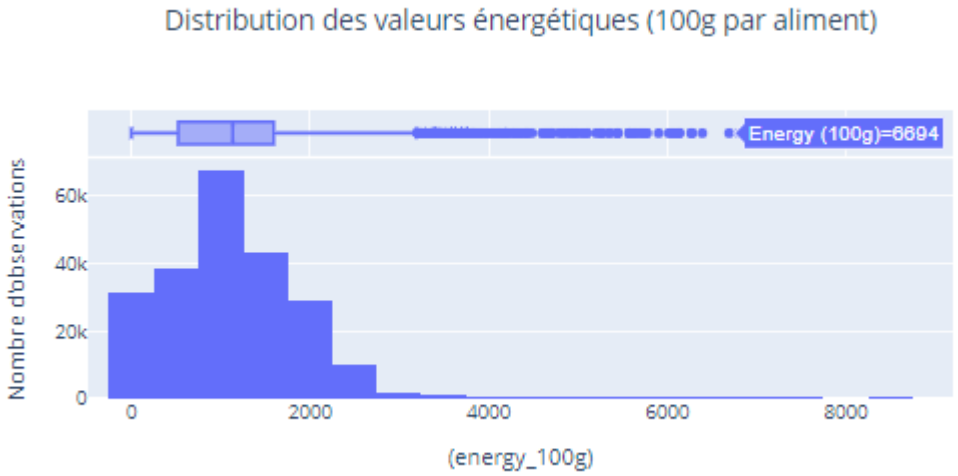


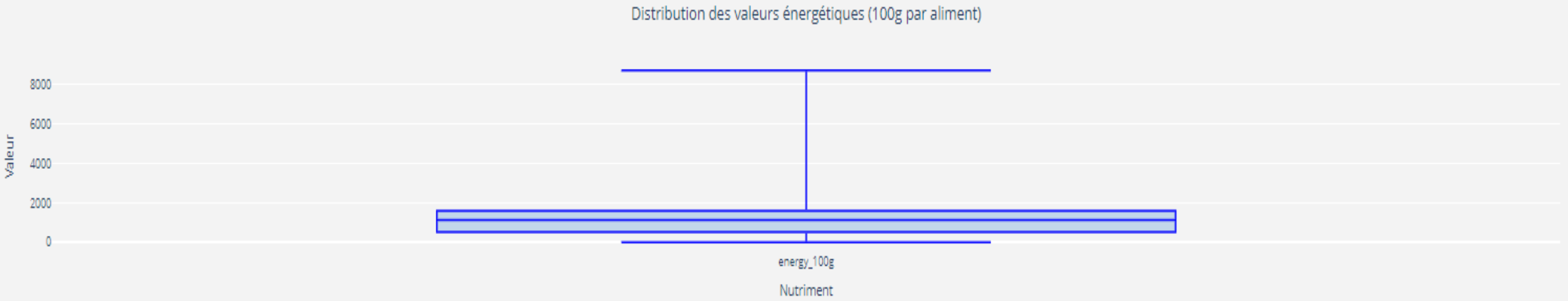




ANALYSE UNIVARIEE

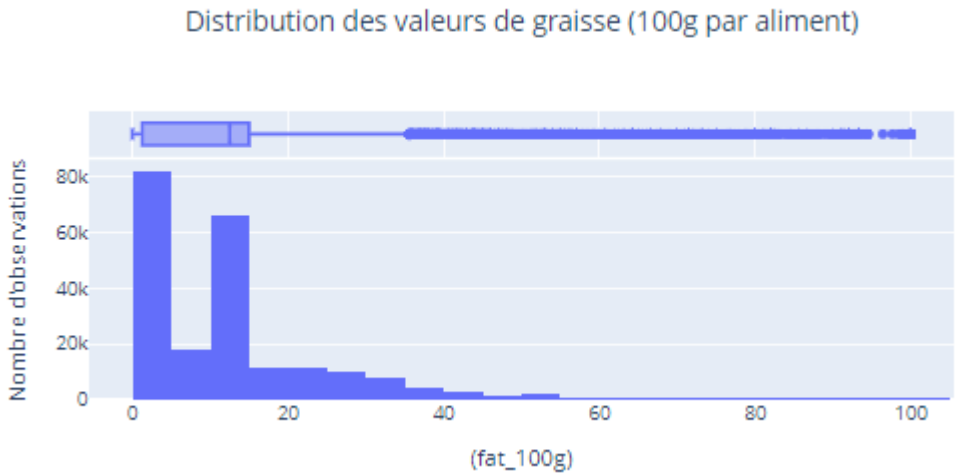
Distribution des valeurs par variable

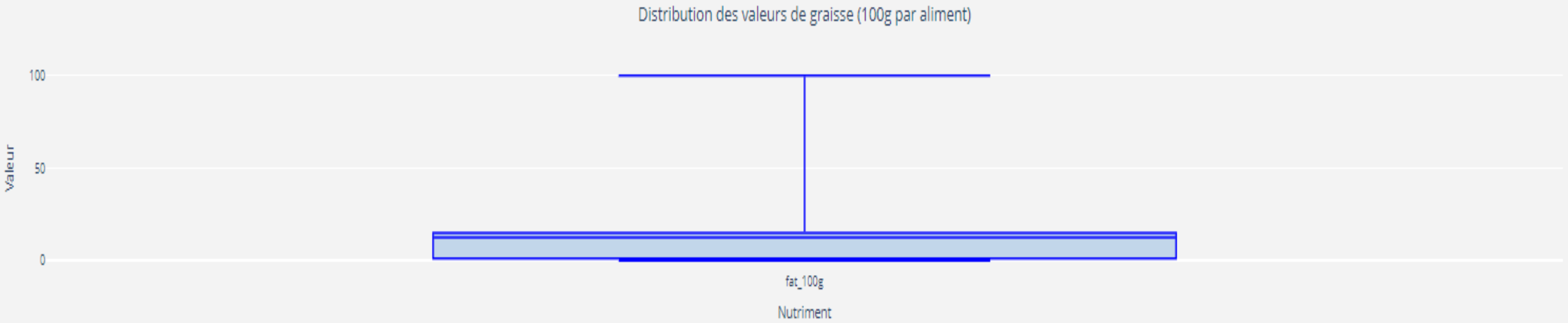




ANALYSE UNIVARIEE

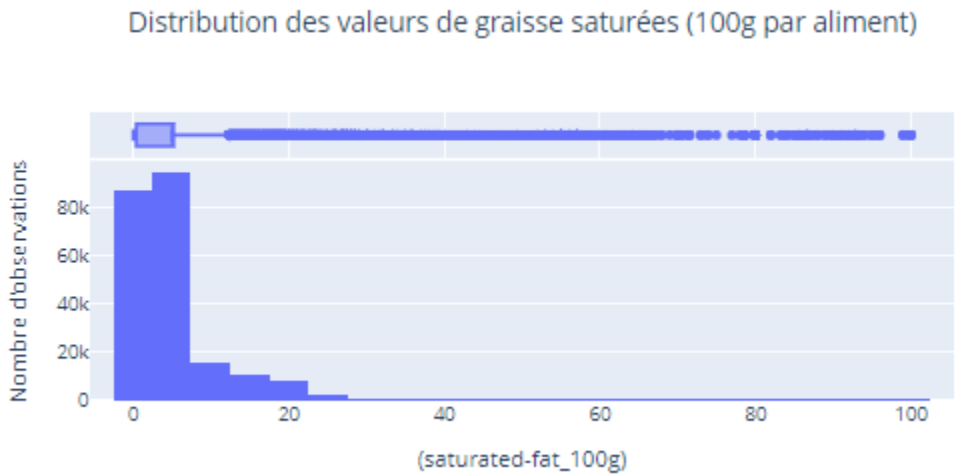
Distribution des valeurs par variable

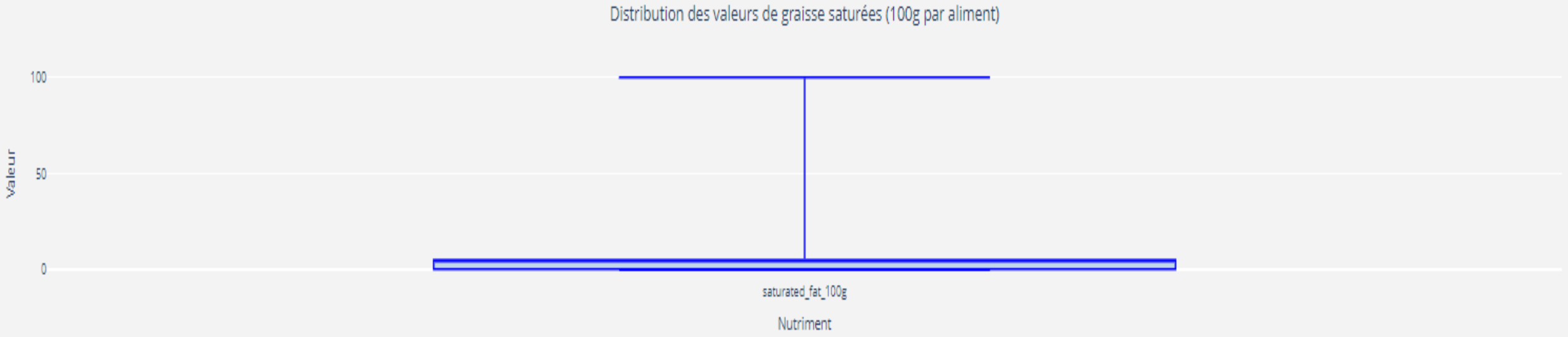




ANALYSE UNIVARIEE

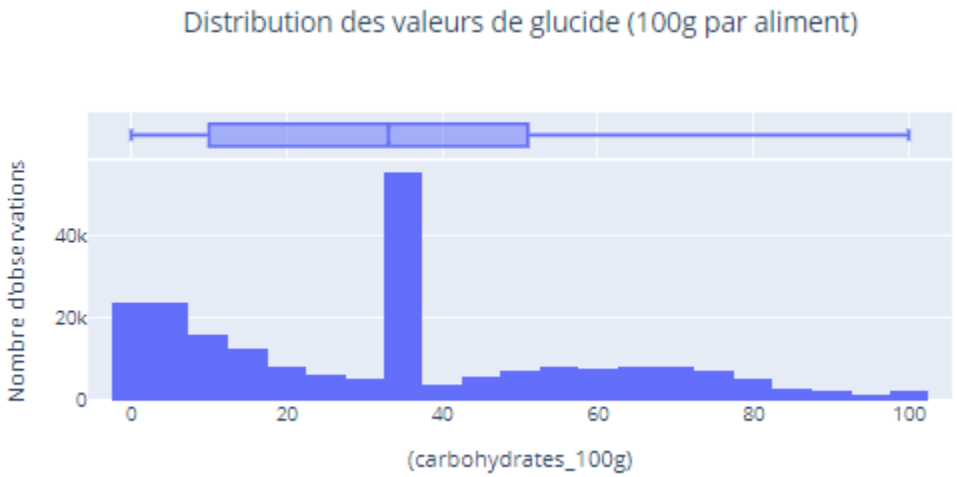
Distribution des valeurs par variable

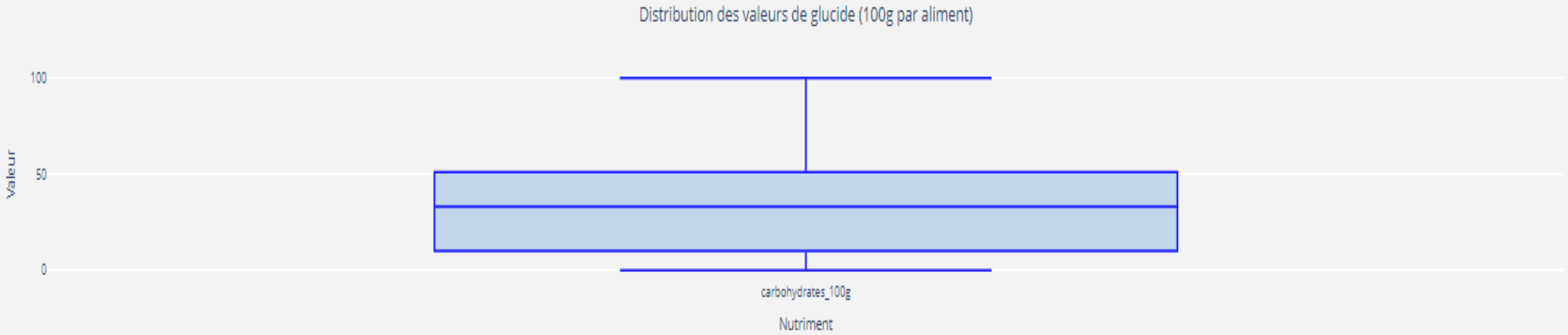




ANALYSE UNIVARIEE

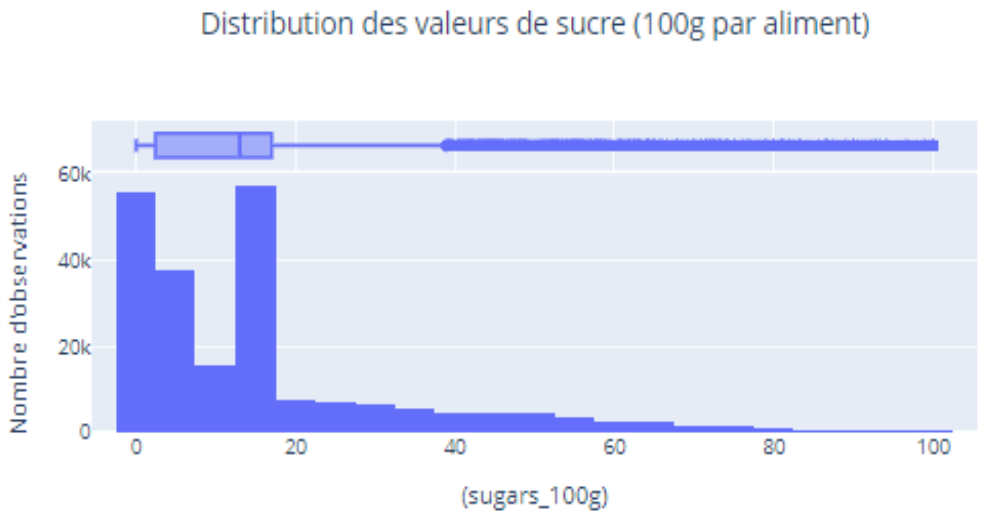
Distribution des valeurs par variable

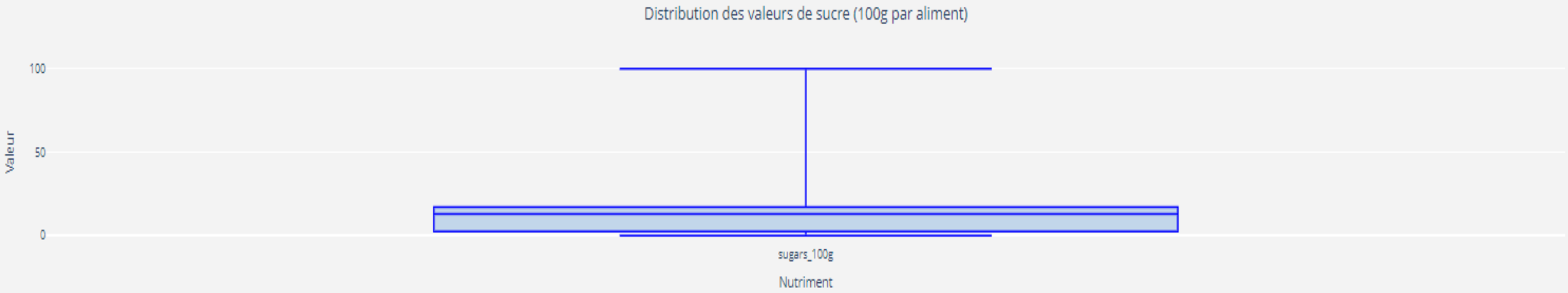




ANALYSE UNIVARIEE

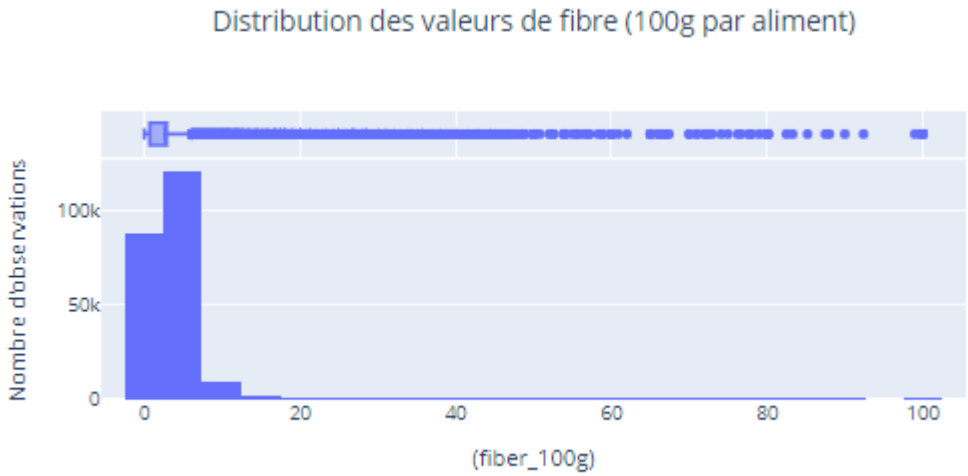
Distribution des valeurs par variable

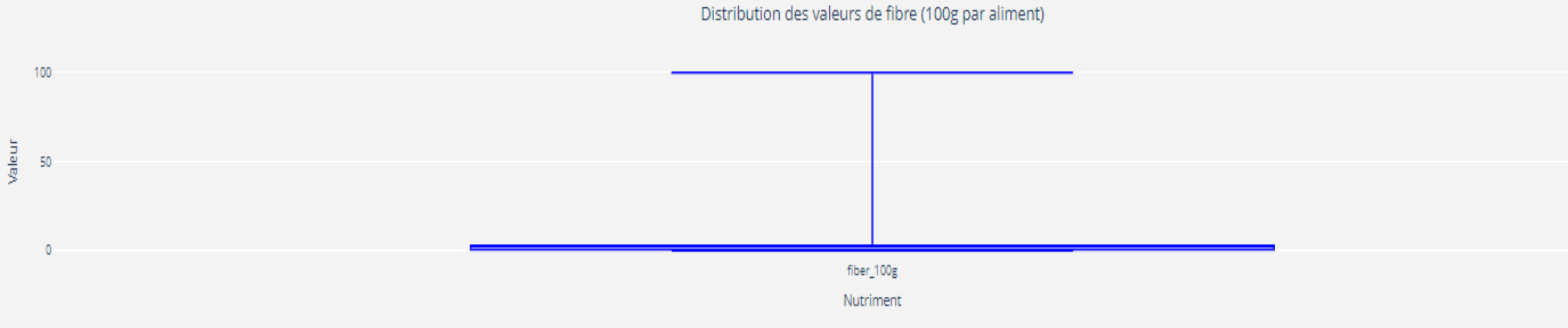




ANALYSE UNIVARIEE

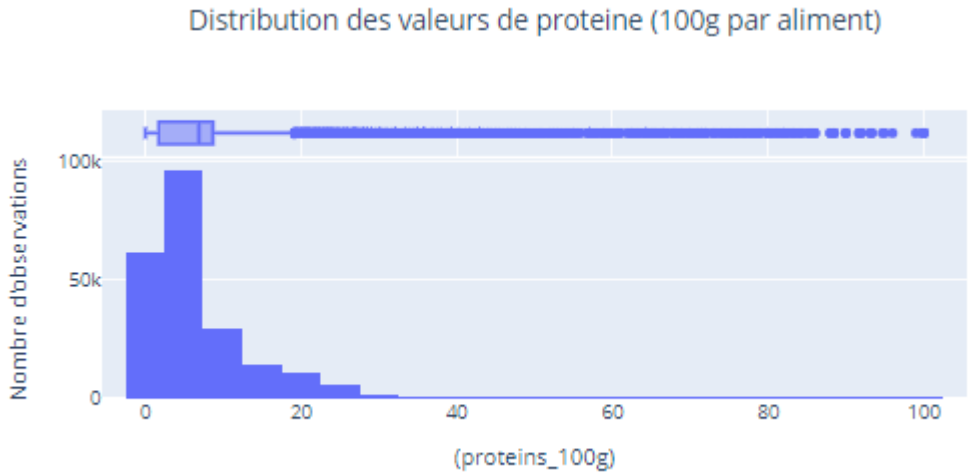
Distribution des valeurs par variable

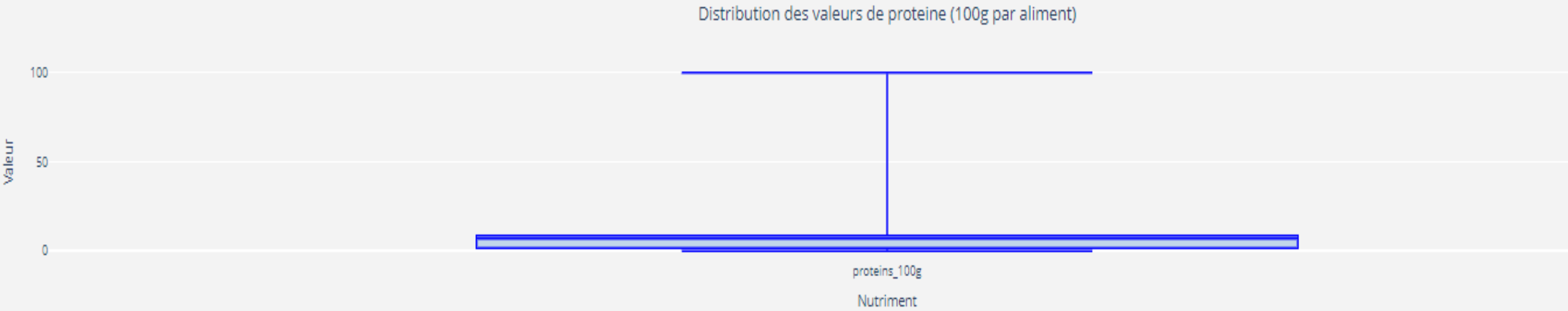




ANALYSE UNIVARIEE

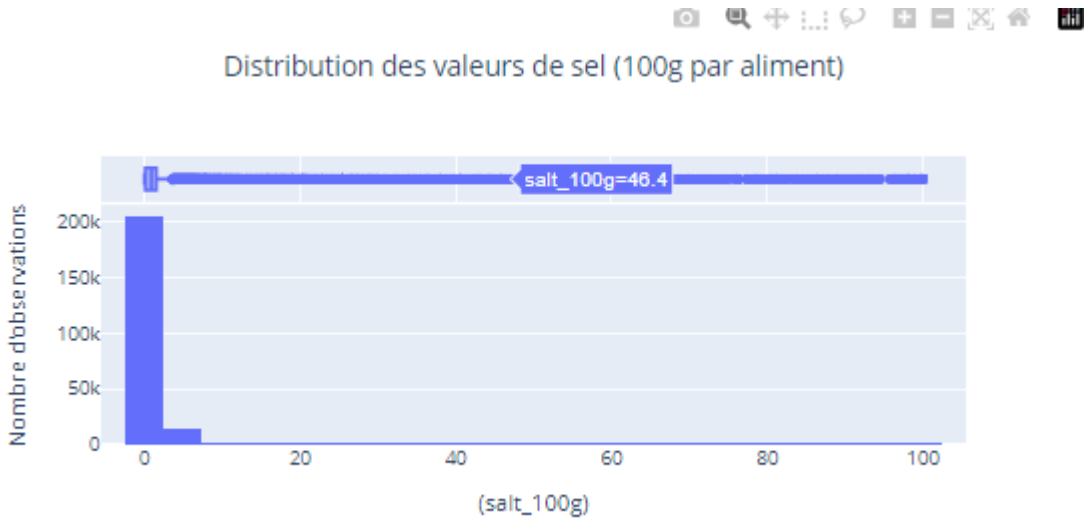
Distribution des valeurs par variable

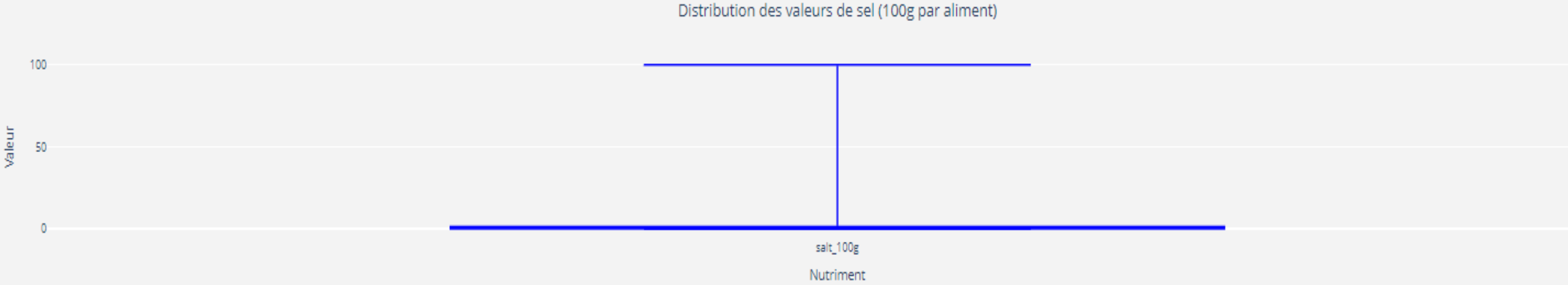




ANALYSE UNIVARIEE

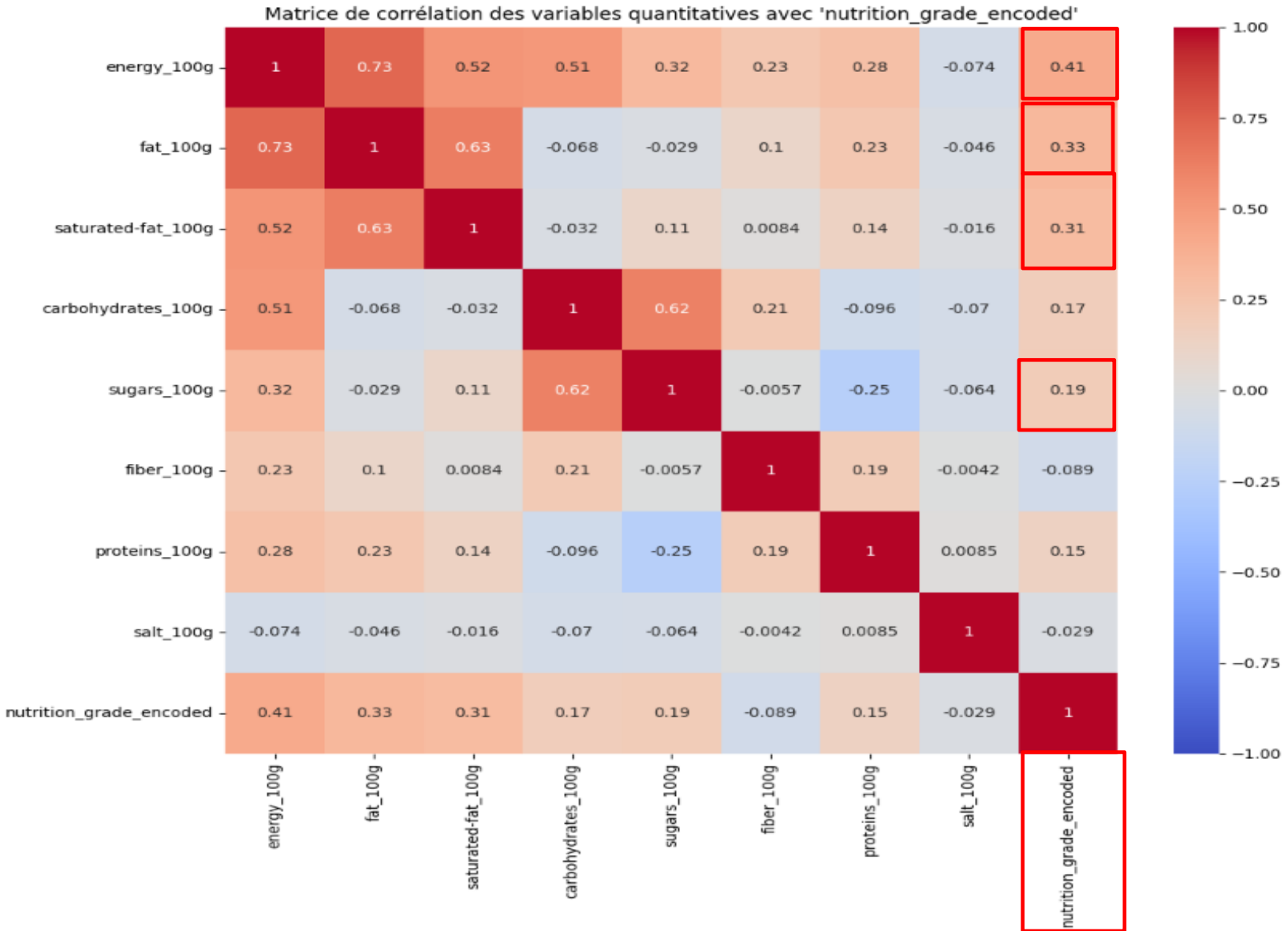
Distribution des valeurs par variable





ANALYSE BIVARIEE

Matrice de correlation



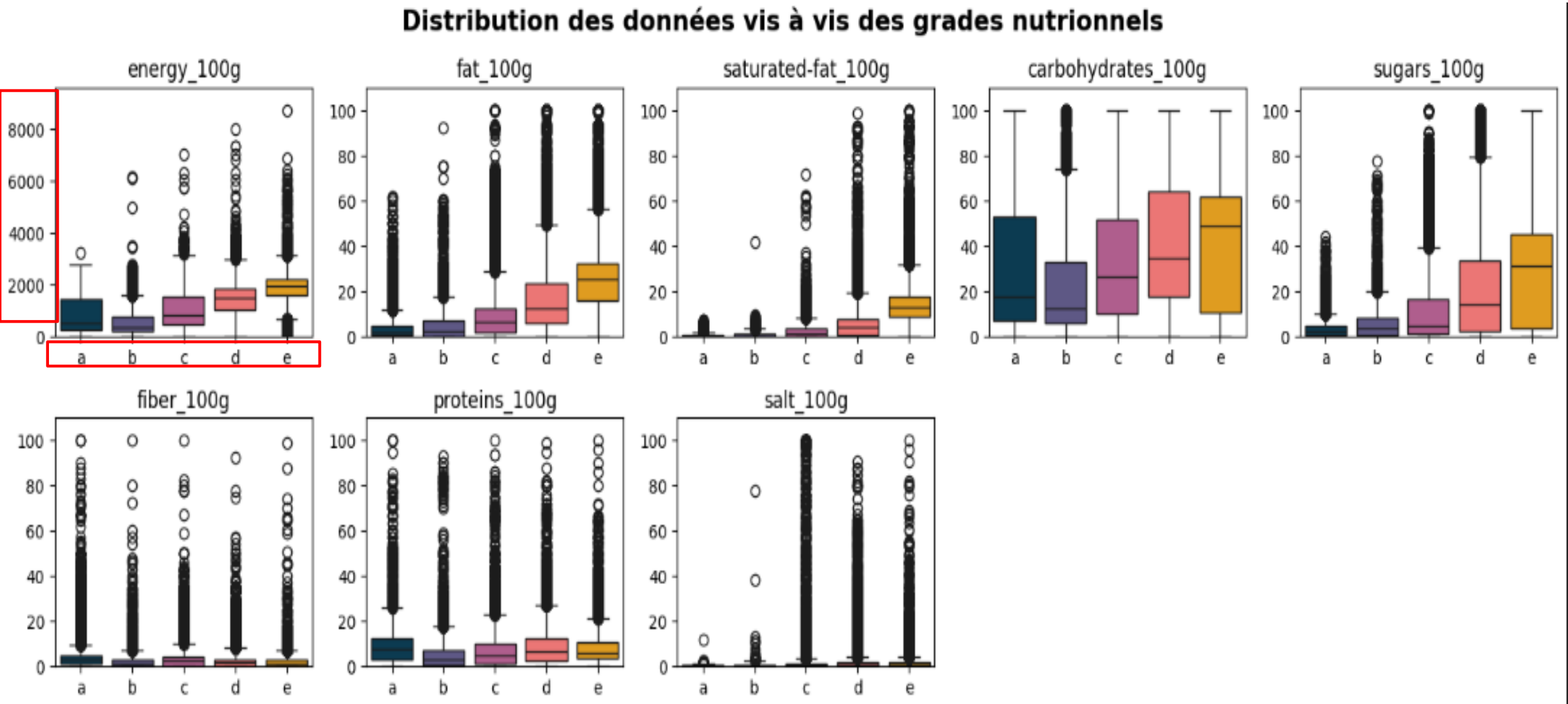
La variable catégorielle **nutrition_grade_fr** a été transformée en une variable numérique ordinale.

Les grades nutritionnels de **A** à **E** ont été convertis en code numériques pour analyser leur corrélation avec les autres variables **quantitatives**.

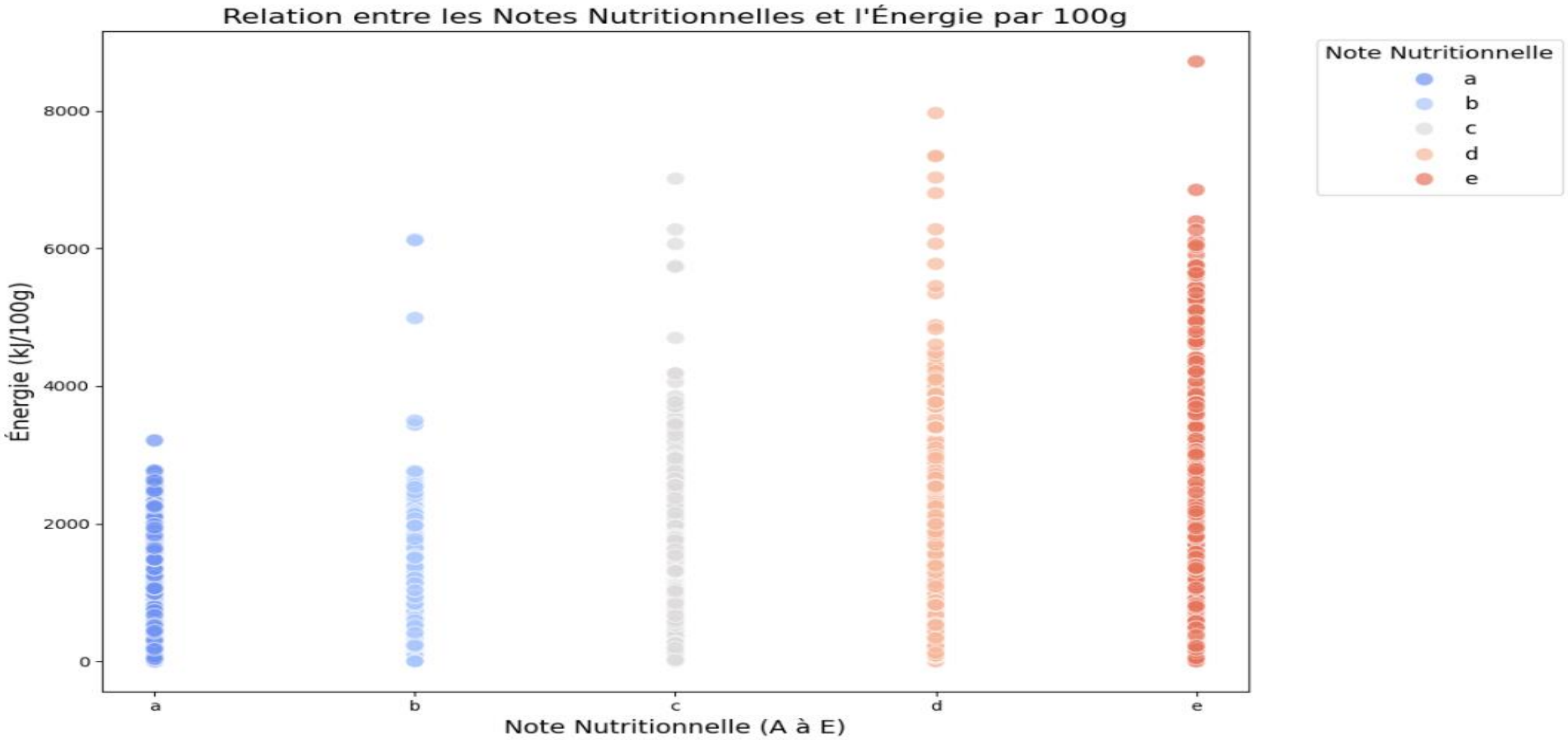
Cela permet d'obtenir un score de correspondance entre les variables quantitatives et le grade nutritionnel (de **A** à **E**) qui est initialement qualitatif.

ANALYSE BIVARIEE

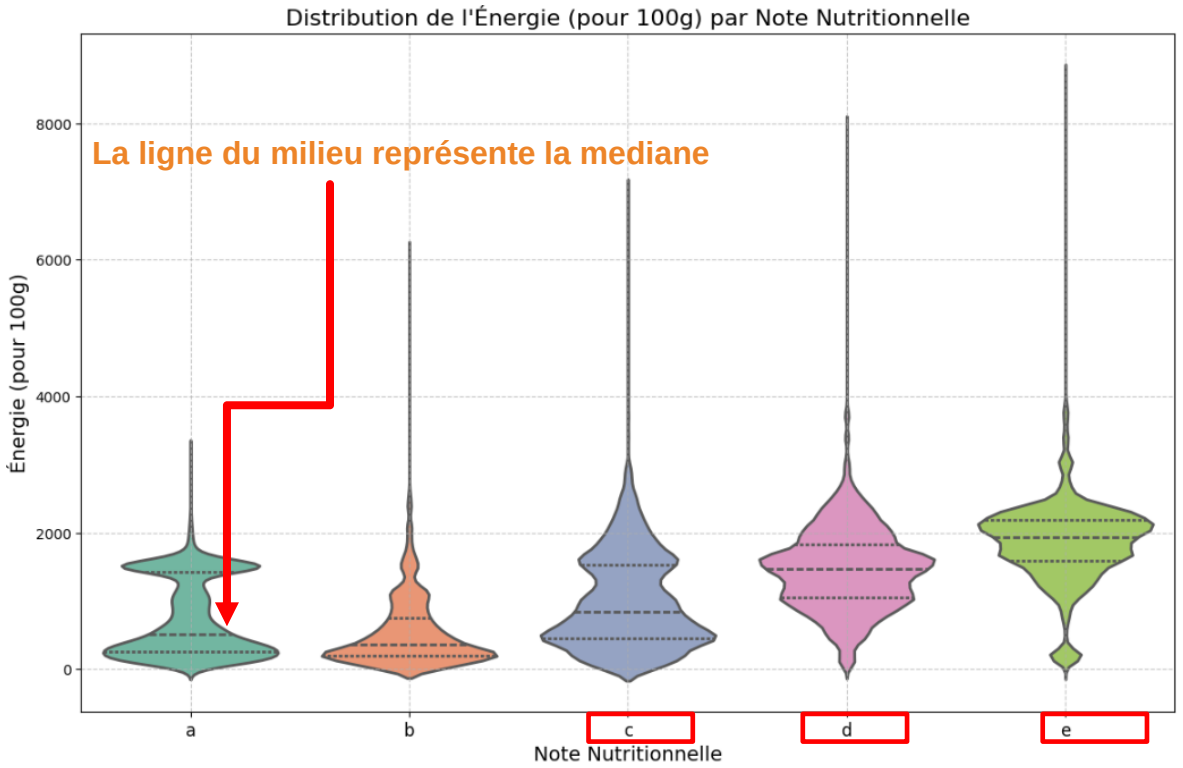
La variable cible nutrition_grade_fr avec les variables quantitatives précédemment néttoyées

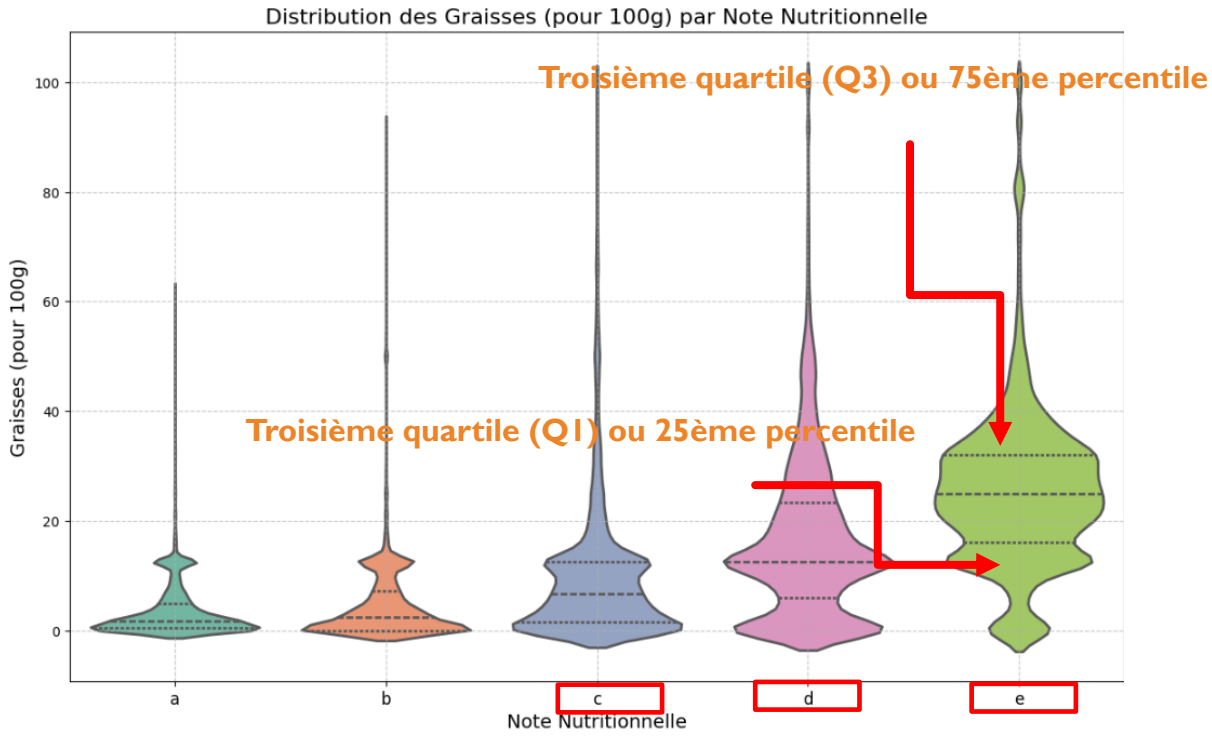


ANALYSE BIVARIEE



ANALYSE BIVARIEE

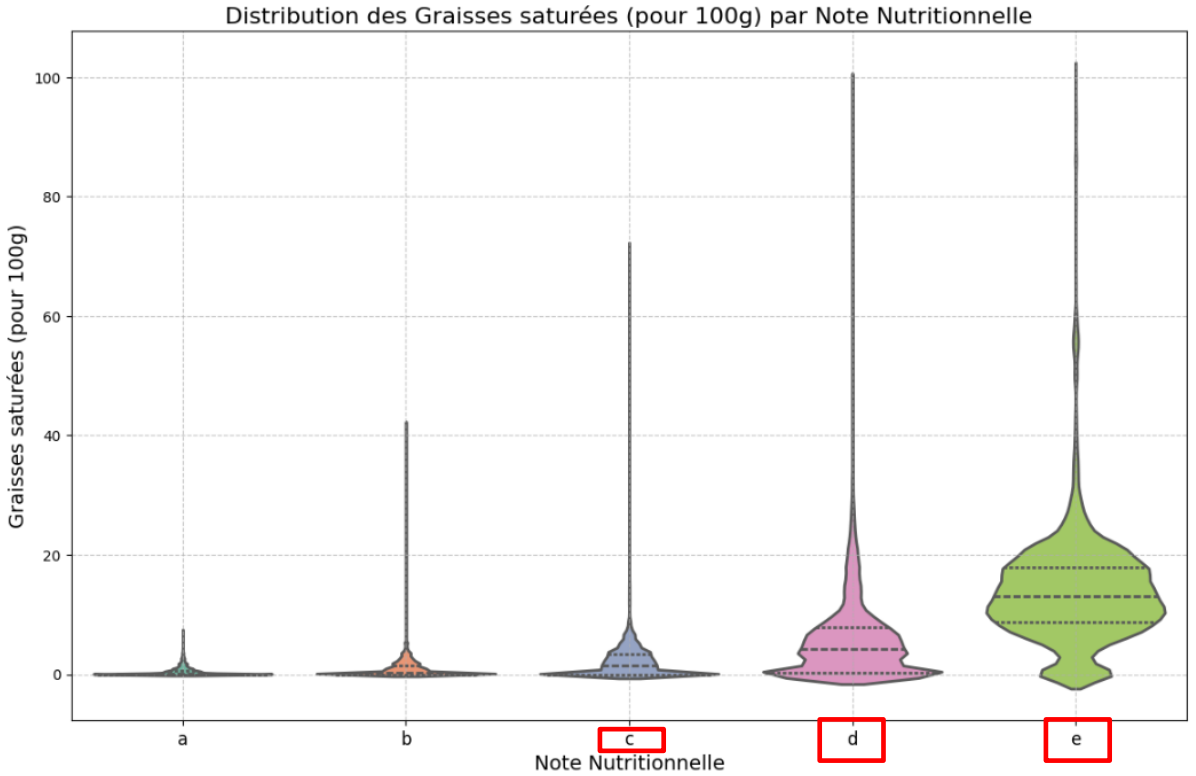


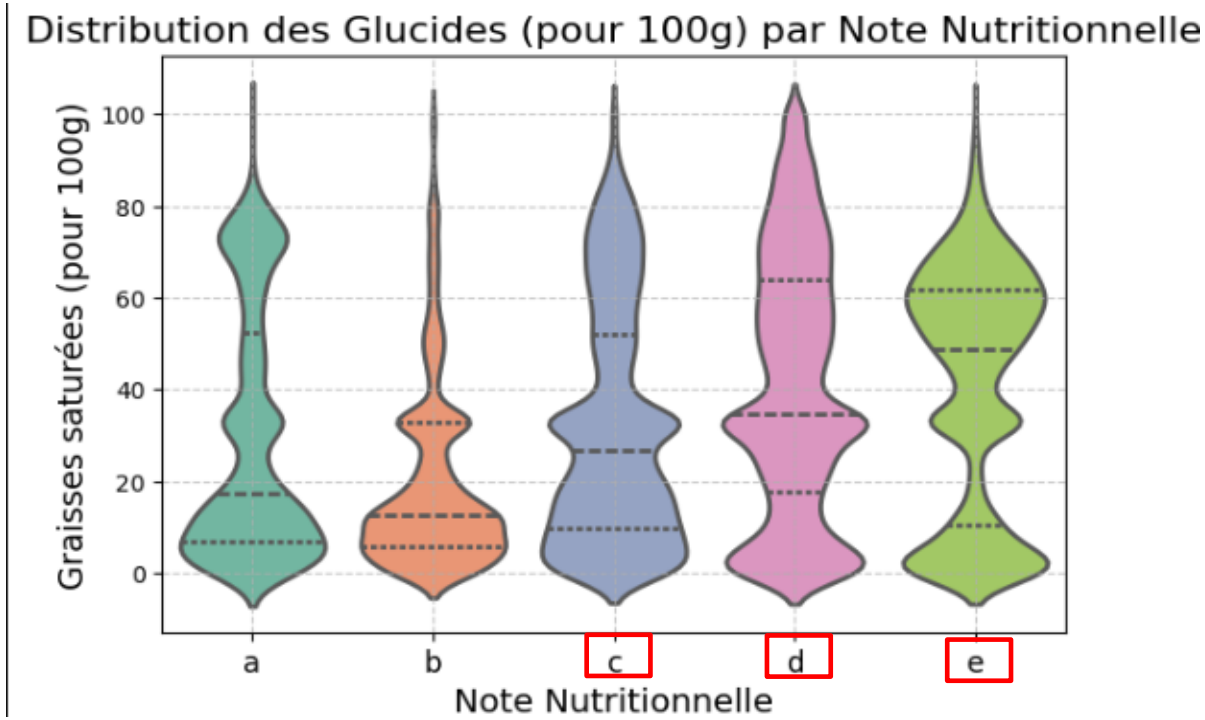


- La note nutritionnelle d'un **produit** à **Valeur energetique élevé** est => **Moyenne, Mauvaise voir très Mauvaise**
- La note nutritionnelle d'un **produit** avec **un taux de graisse élevé** est => **Moyenne, Mauvaise voir très Mauvaise**

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

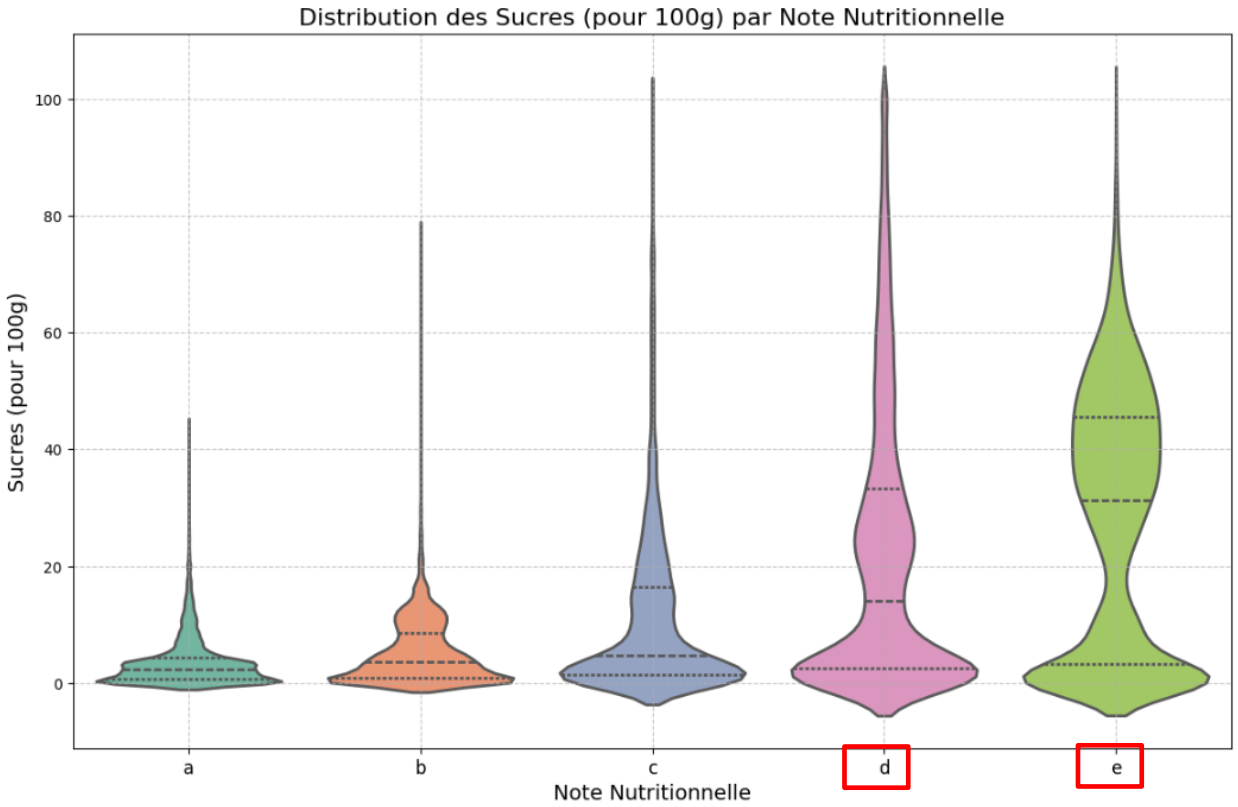
ANALYSE BIVARIEE

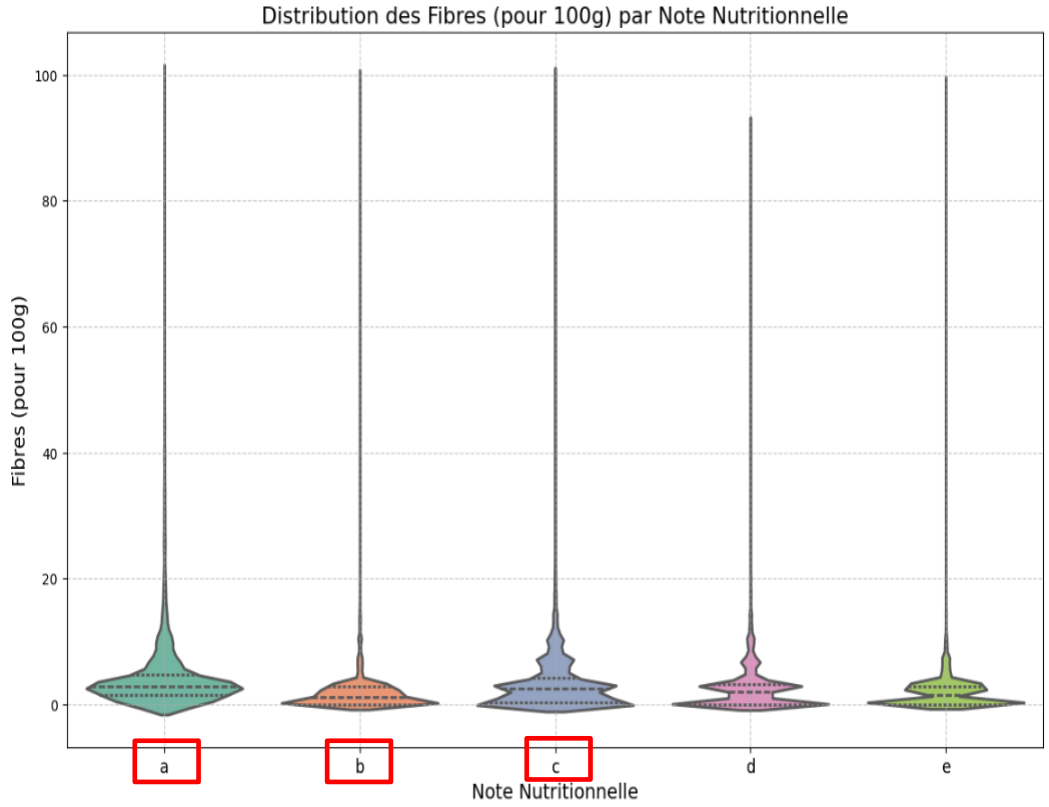




(**A** : Très bon sur le plan nutritionnel. / **B** : Bon. / **C** : Moyen. / **D** : Mauvais. / **E** : Très mauvais.)

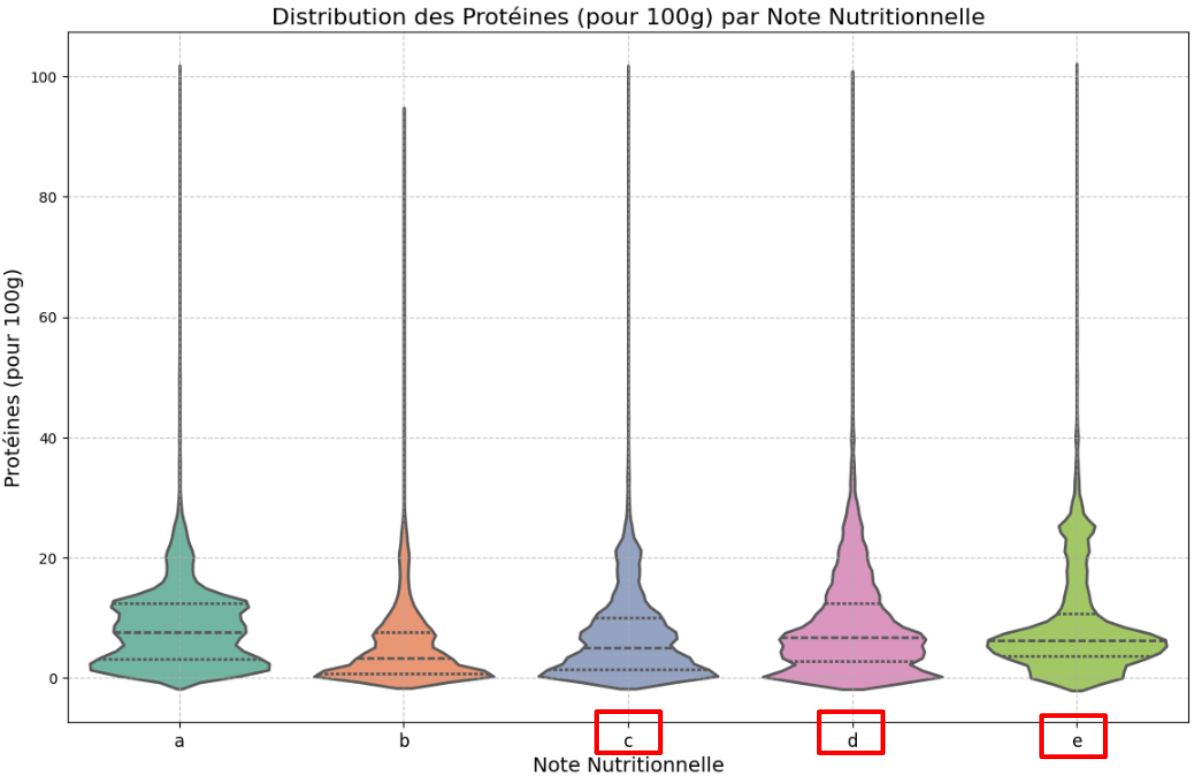
ANALYSE BIVARIEE

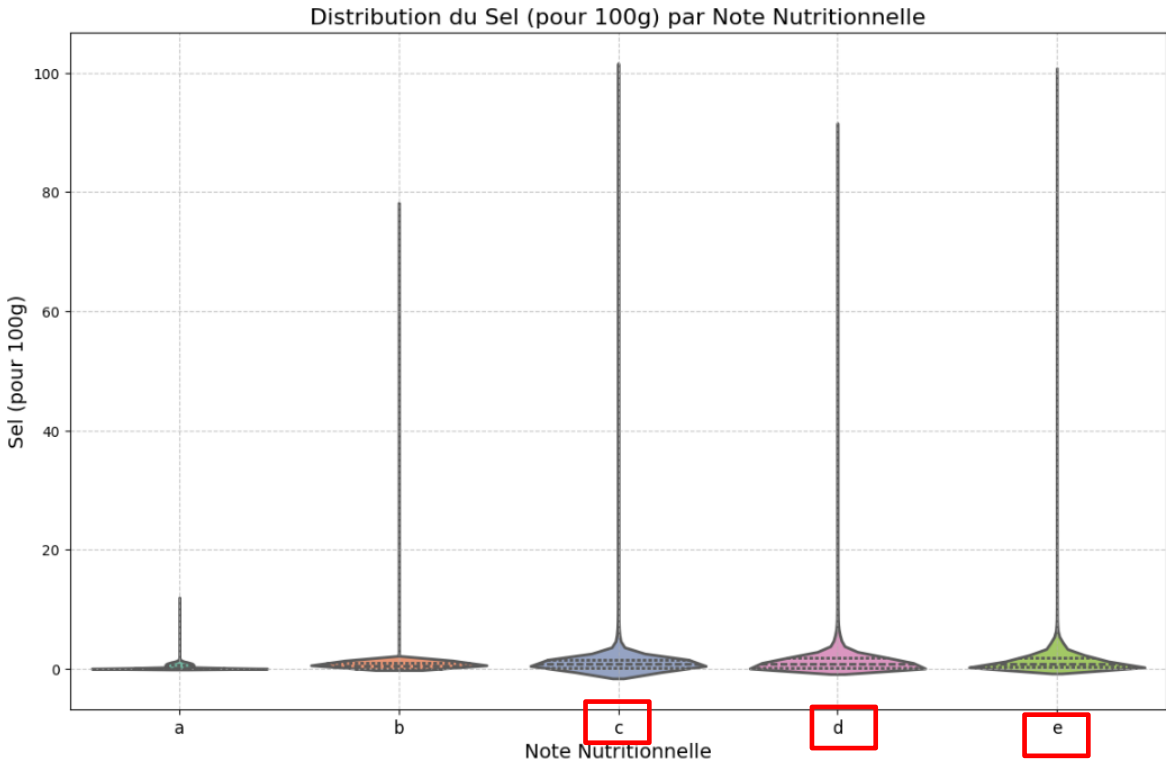




(**A** : Très bon sur le plan nutritionnel. / **B** : Bon. / **C** : Moyen. / **D** : Mauvais. / **E** : Très mauvais.)

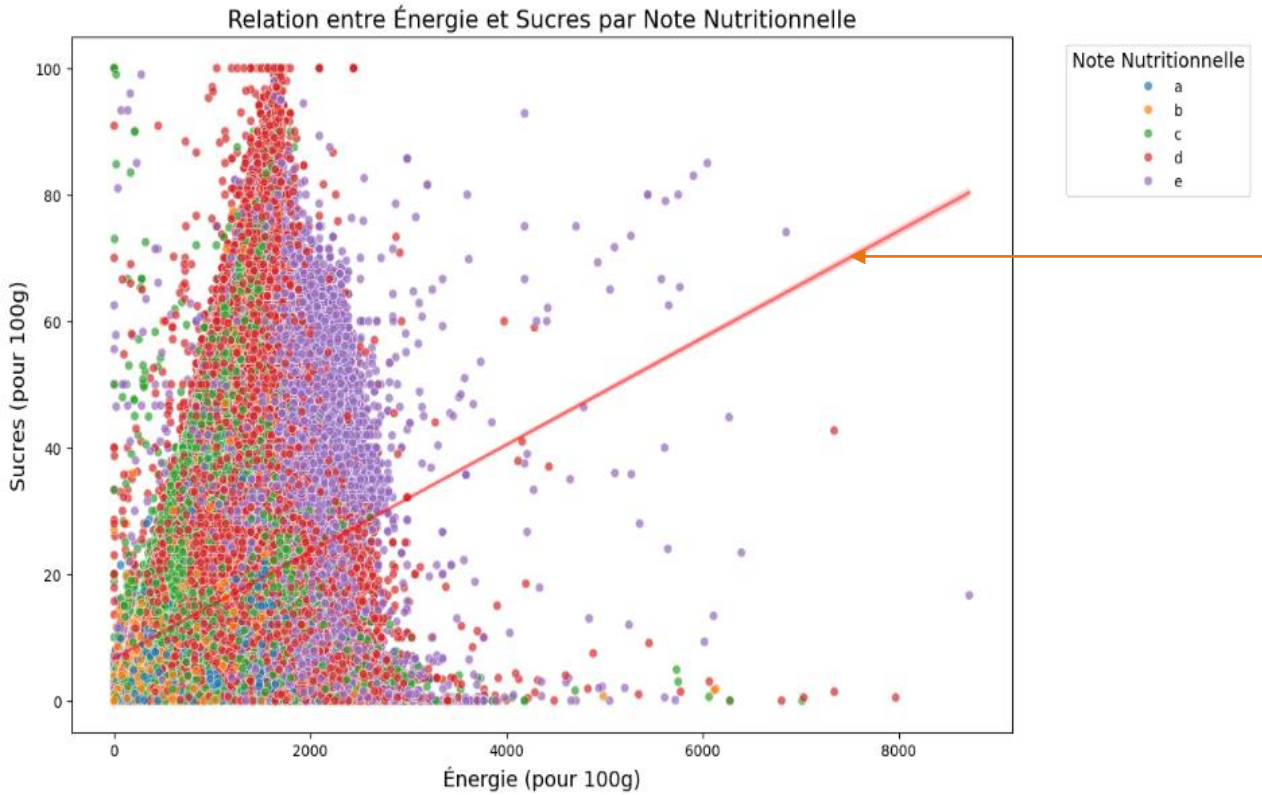
ANALYSE BIVARIEE





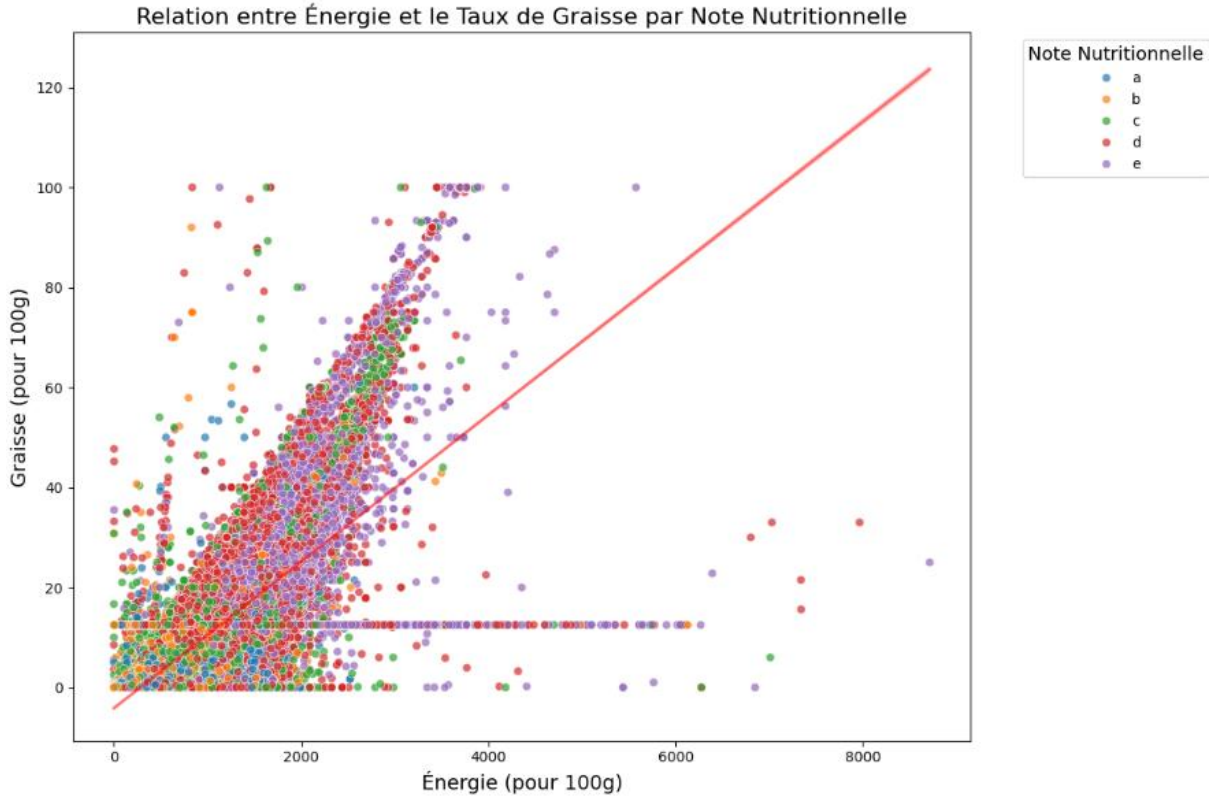
(**A** : Très bon sur le plan nutritionnel. / **B** : Bon. / **C** : Moyen. / **D** : Mauvais. / **E** : Très mauvais.)

ANALYSE MULTIVARIEE

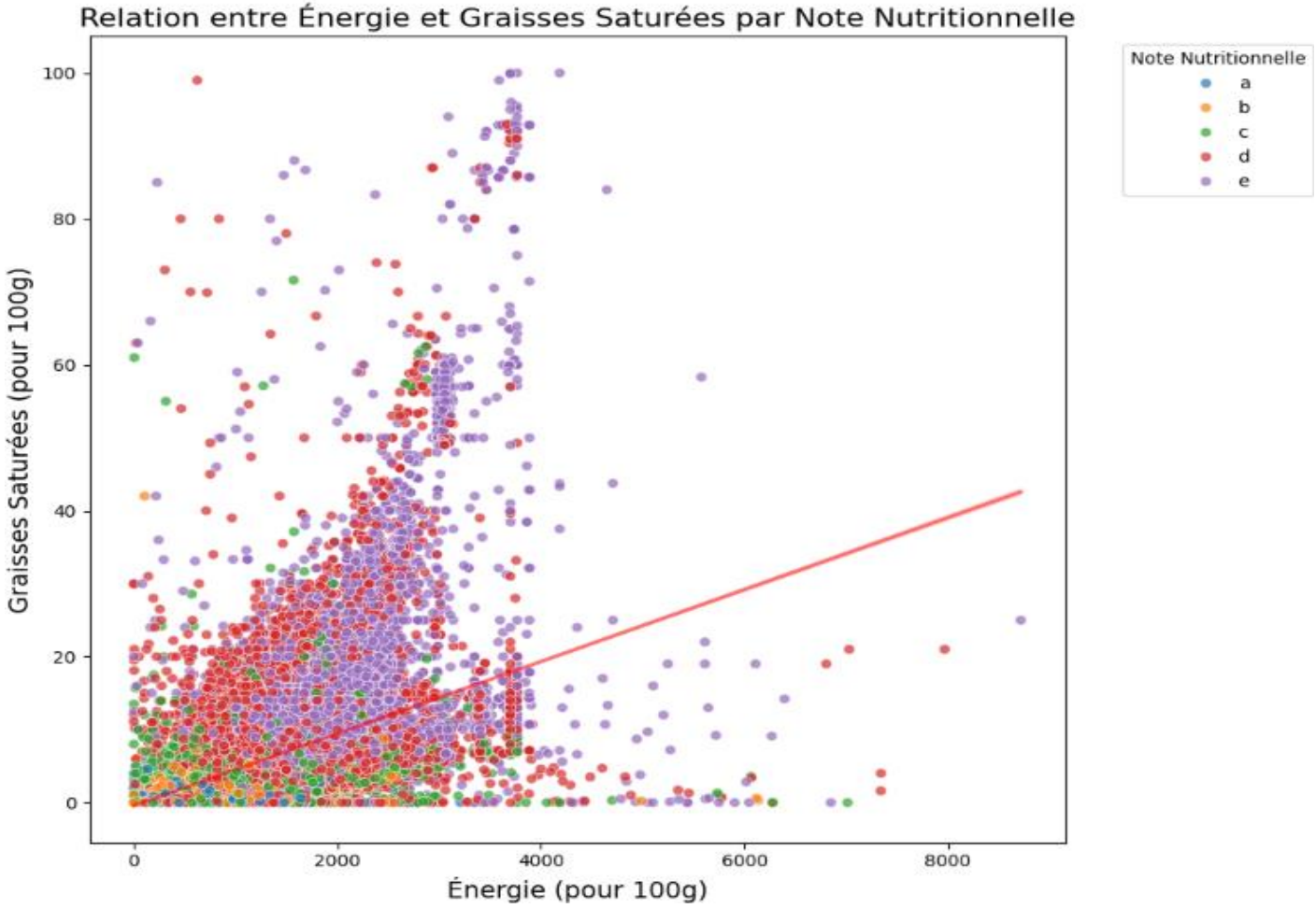


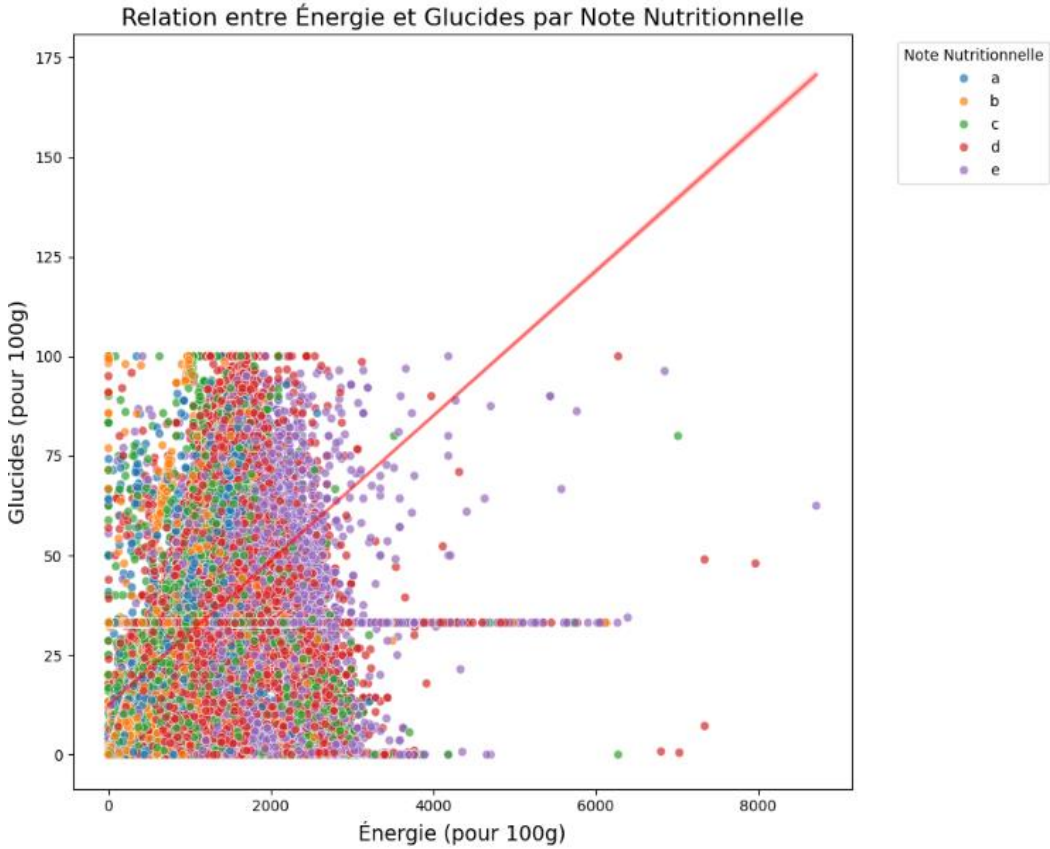
Ligne de régression pour visualiser la tendance entre les deux variables

sugars_100g et energy_100g.

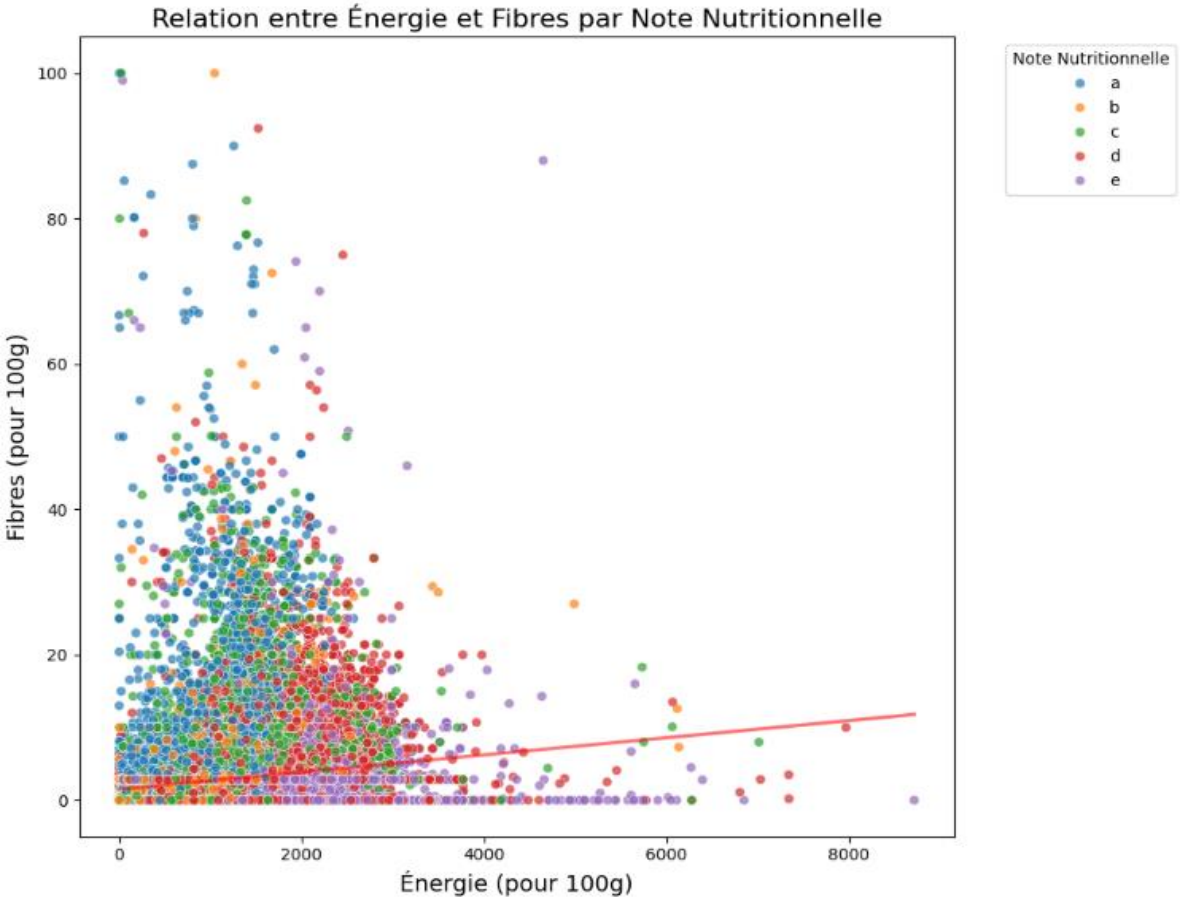


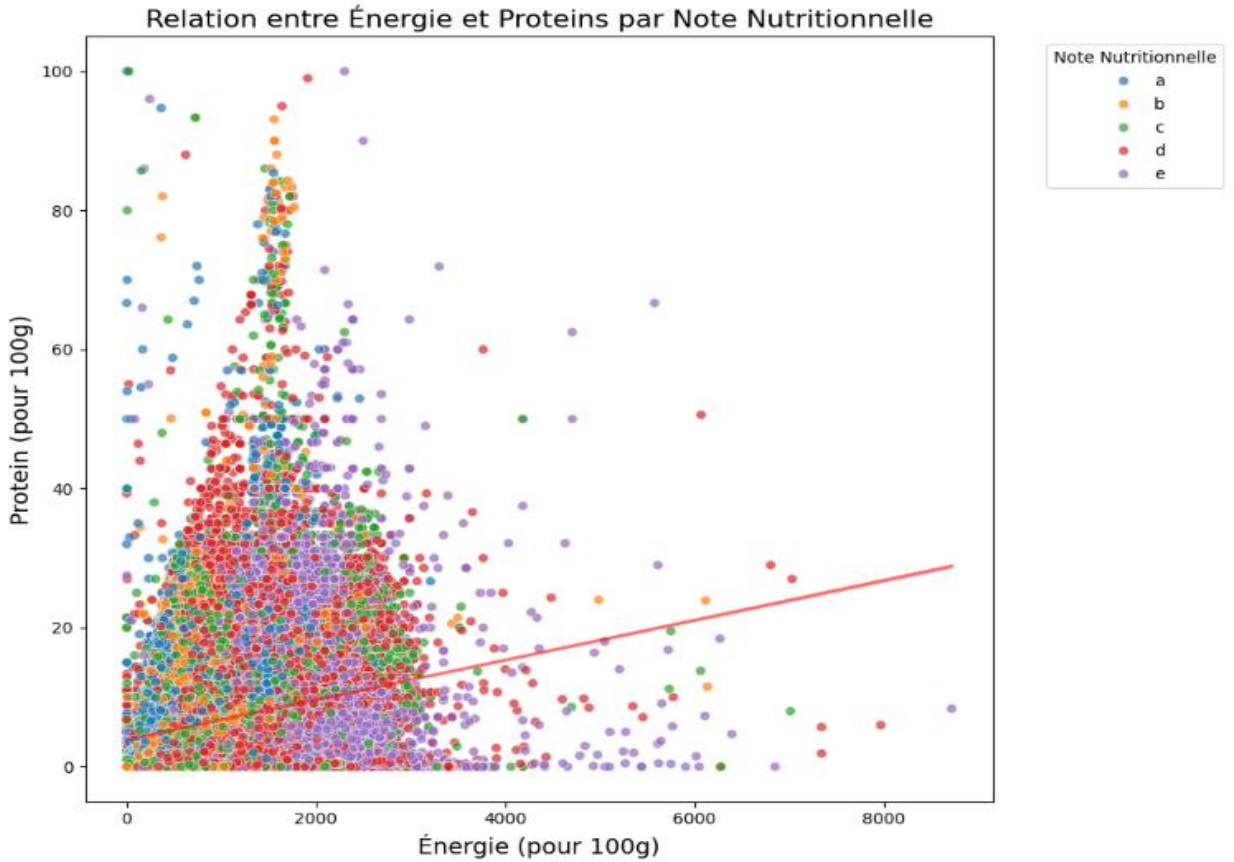
ANALYSE MULTIVARIEE



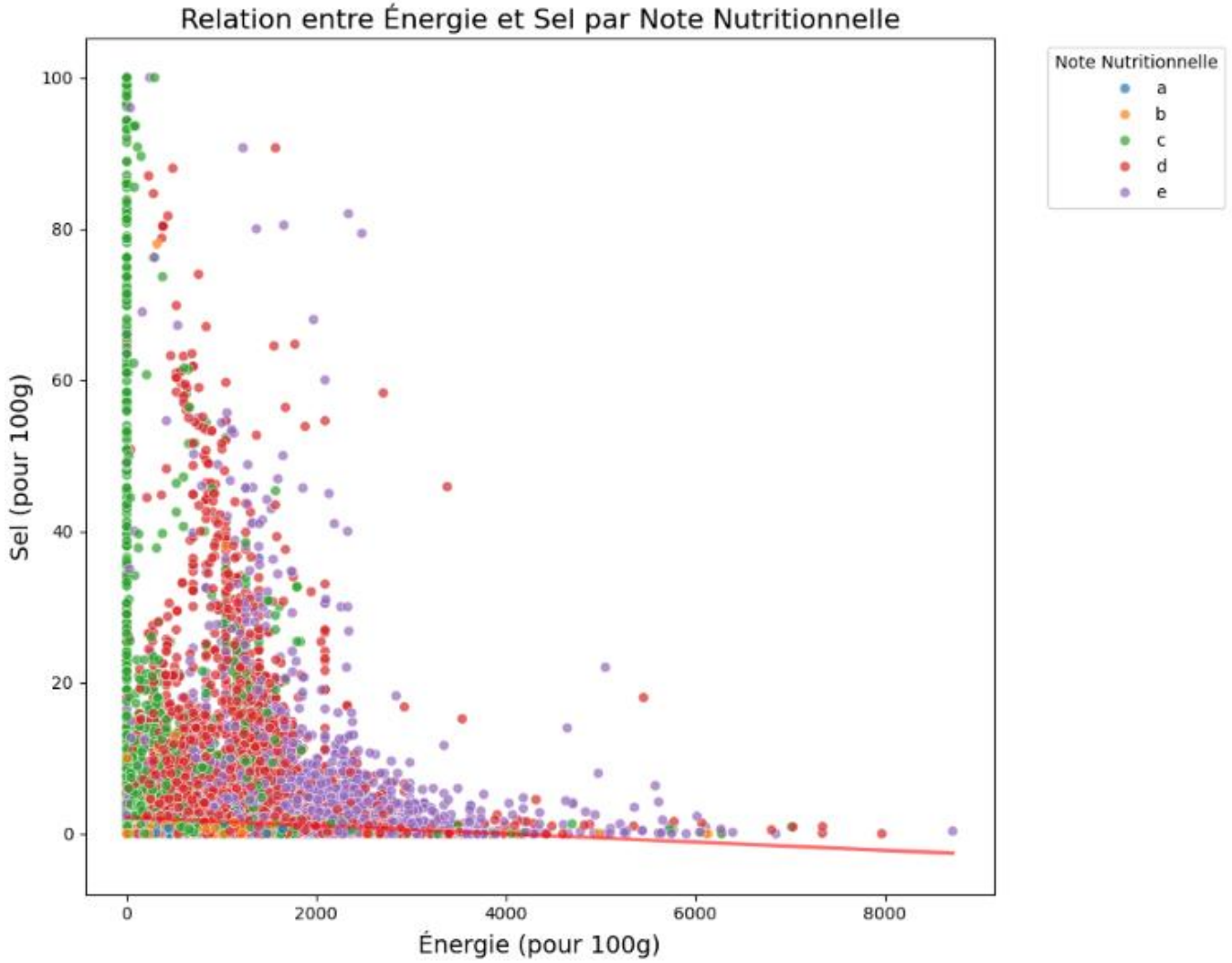


ANALYSE MULTIVARIEE



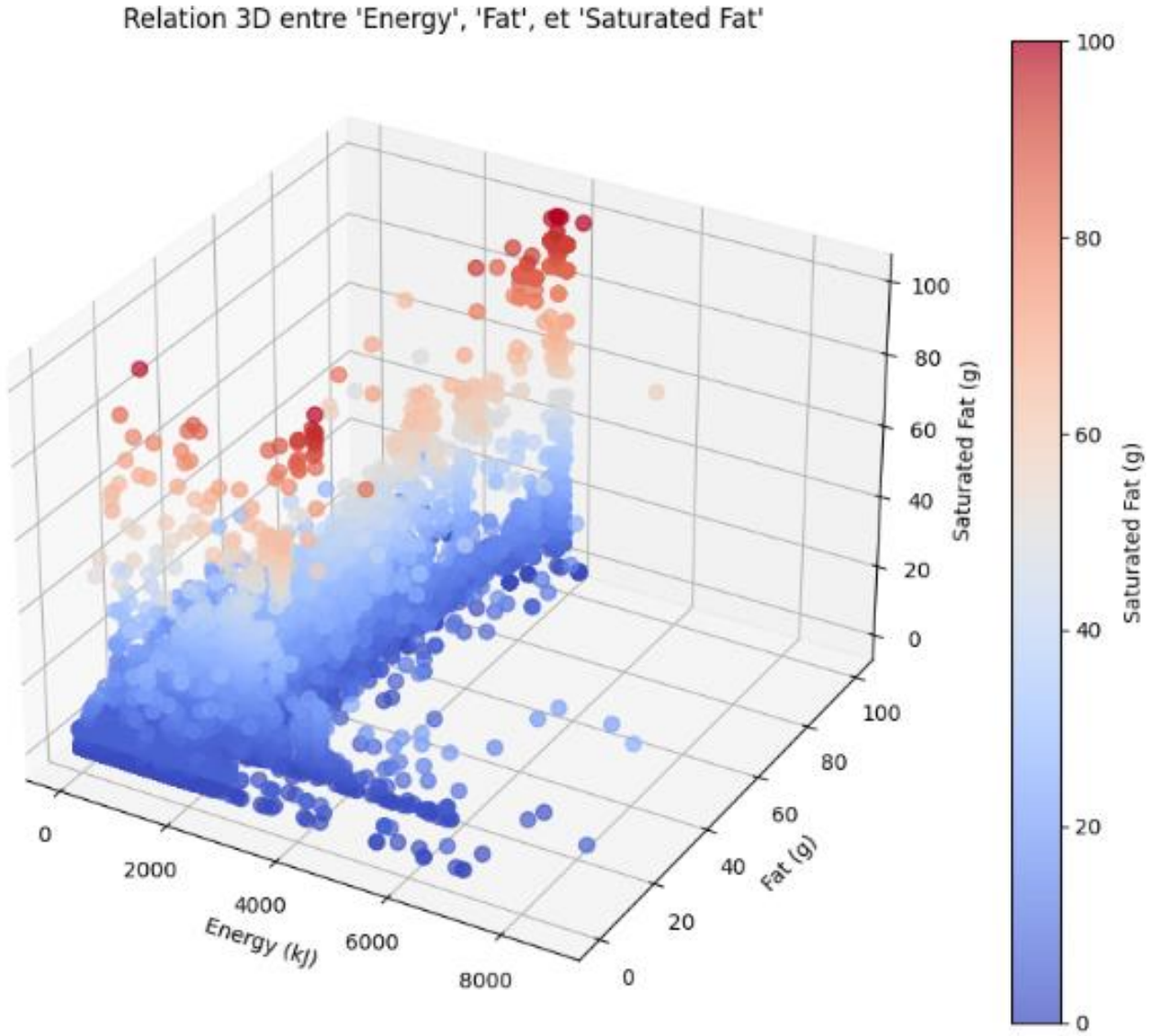


ANALYSE MULTIVARIEE



ANALYSE MULTIVARIEE

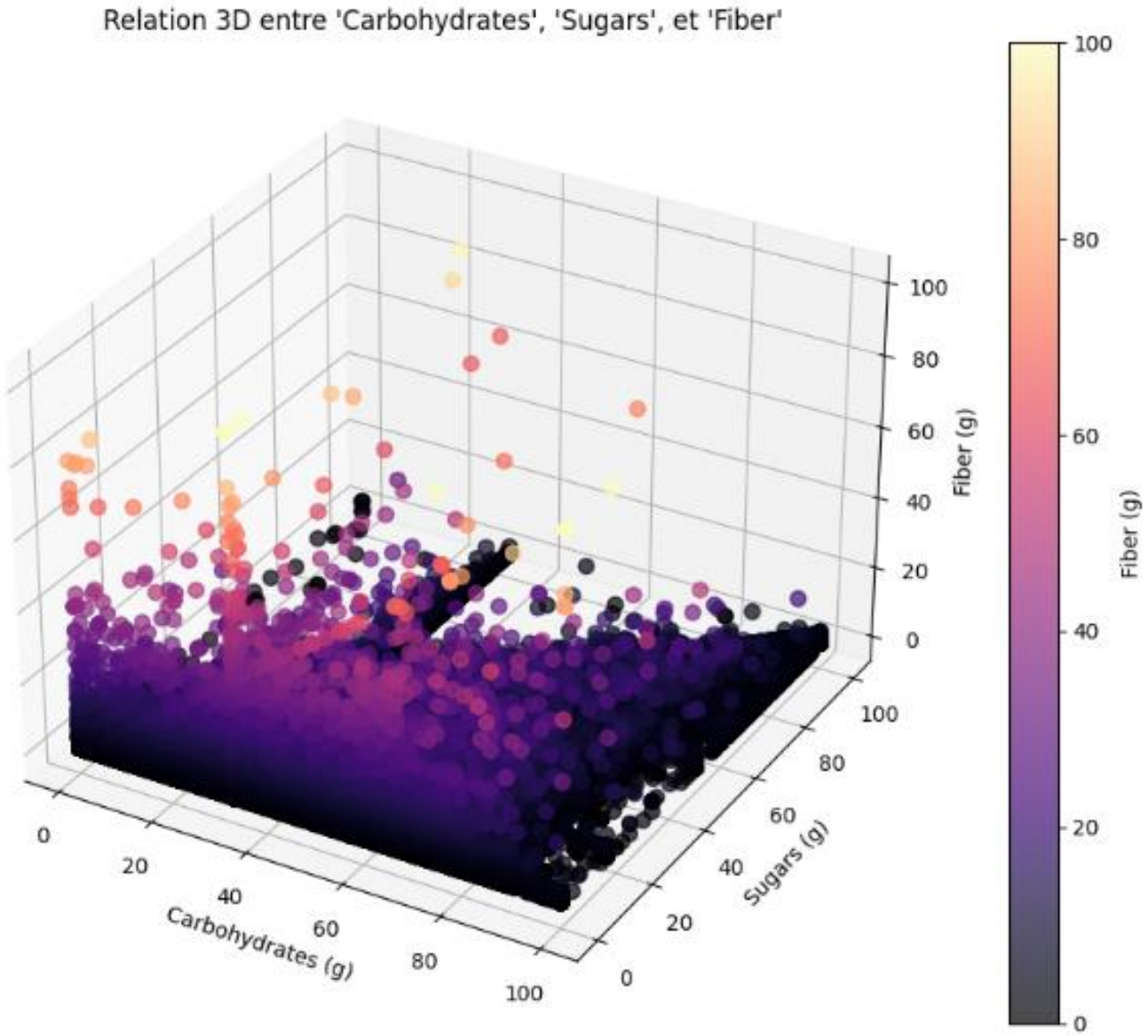
Relation entre 3 variables



- Le graphique montre la relation entre **l'énergie**, les **graisses**, et les **graisses saturées**.
- Plus la teneur **en énergie et en graisses augmente**, plus la quantité de graisses saturées est élevée.
- Les points colorés illustrent cette corrélation.

ANALYSE MULTIVARIEE

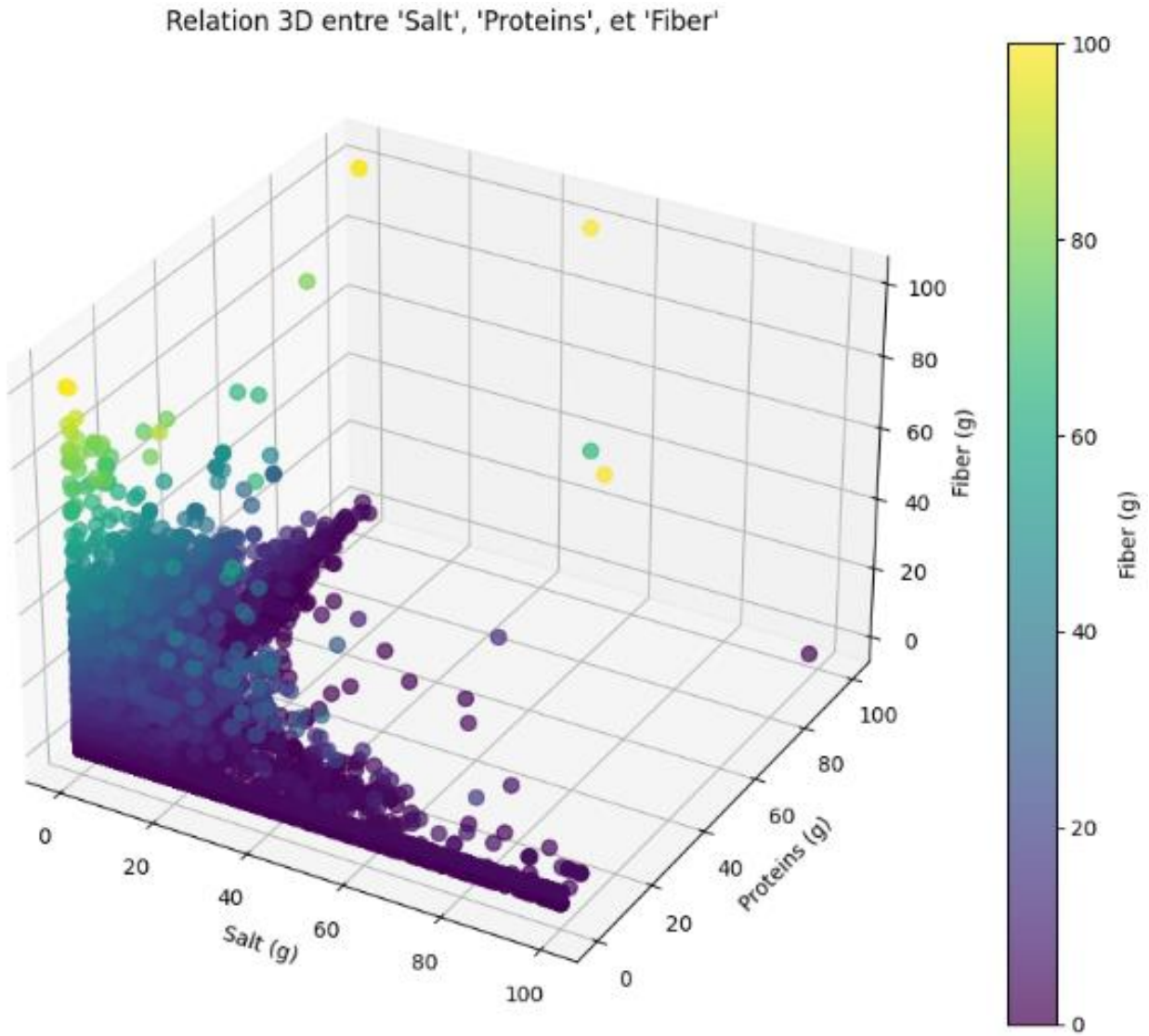
Relation entre 3 variables



- 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 observe une faible concentration de **fibres** pour des **niveaux élevés de sel**

ANALYSE MULTIVARIEE

(ACP) L' Analyse en Composantes Principales

```
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]
```

1. Standardisation des données

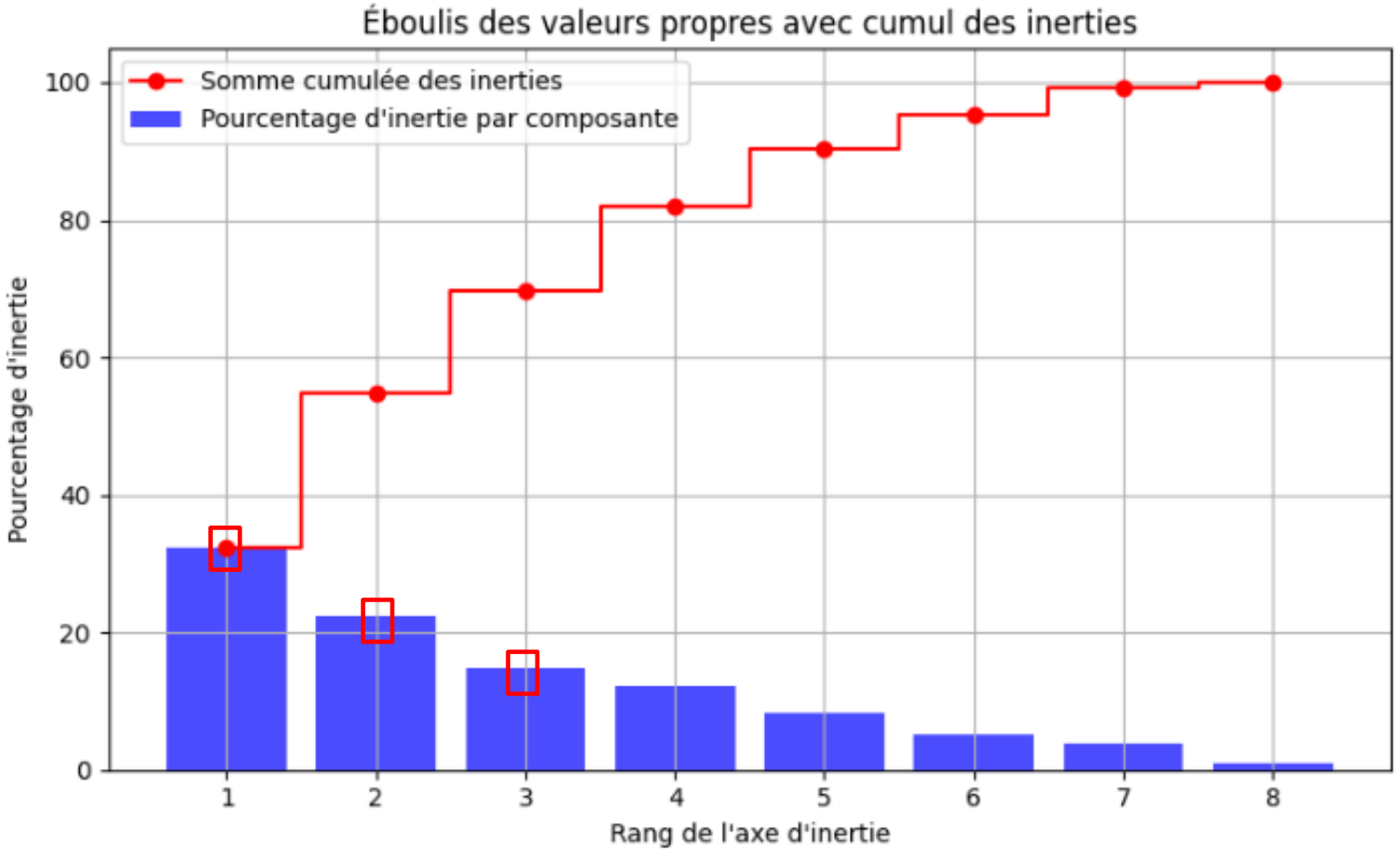
chaque variable est centrée (moyenne égale à 0) et réduite (écart-type égal à 1).
Mise à niveau et réduction de l'échelle des données.

2. L'ACP transforme les données en un nouvel ensemble de variables appelées « composantes principales » (8).

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.

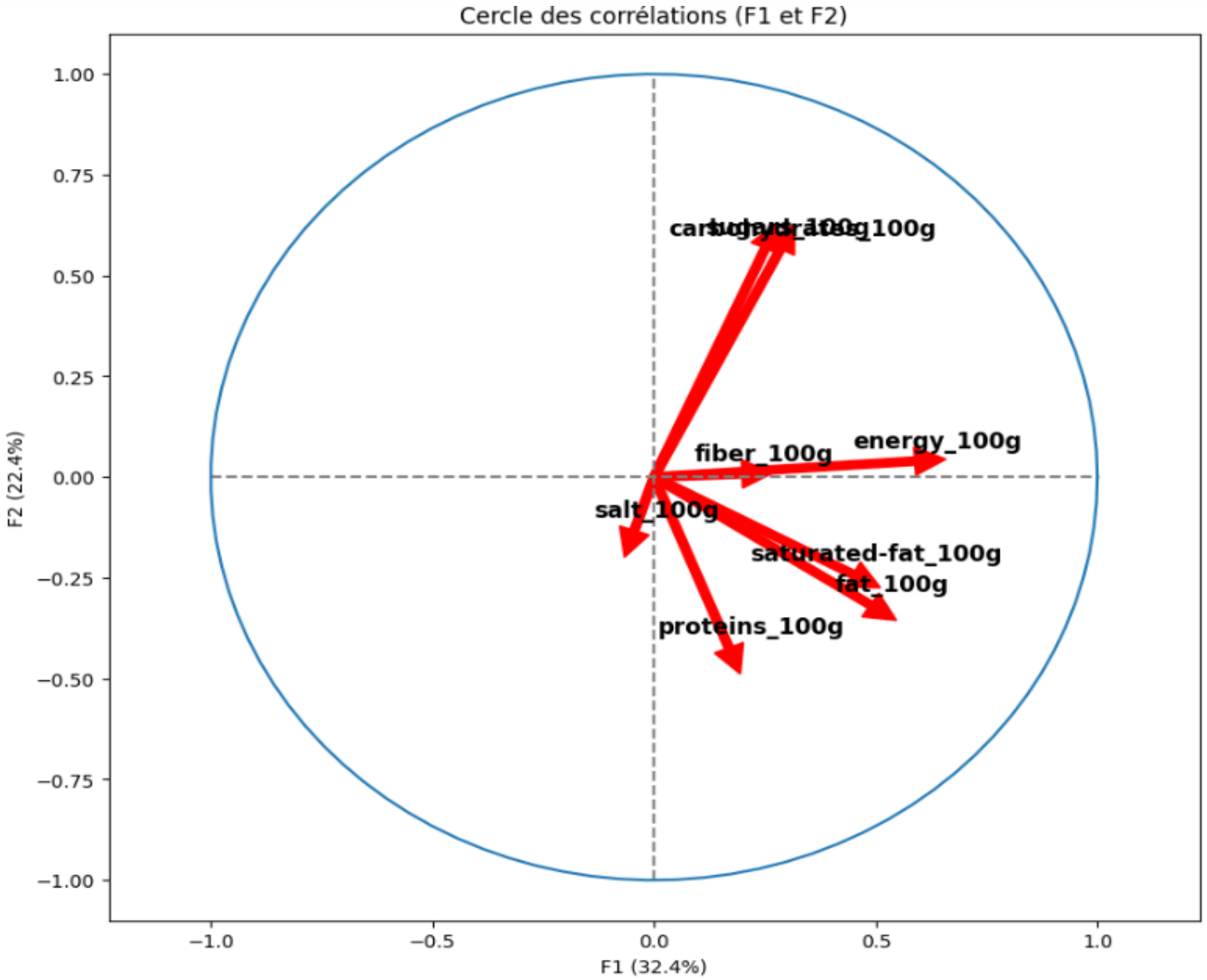
- La **première** composante explique environ **32%** de la variance

- 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)



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.

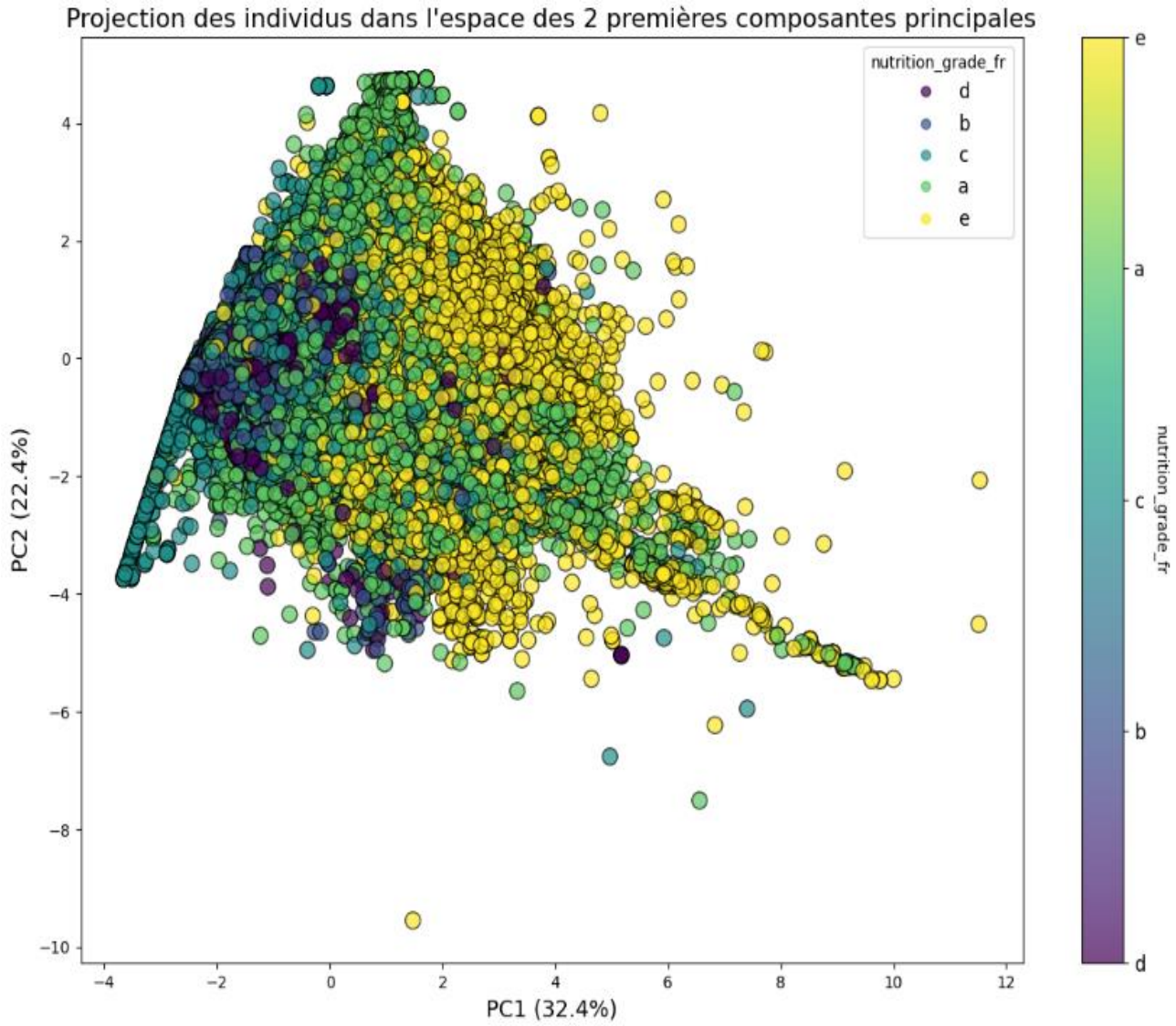
Dans ce cas:

/ Les **Gluicides** (**carbohydrates**_100g)
/ Les **Sucres** (**Sugars**_100g)

Elles figurent parmi les variables qui contribueront le plus à la prédiction du **score nutritionnelle** (**nutrition_grade_fr**)

ANALYSE MULTIVARIEE

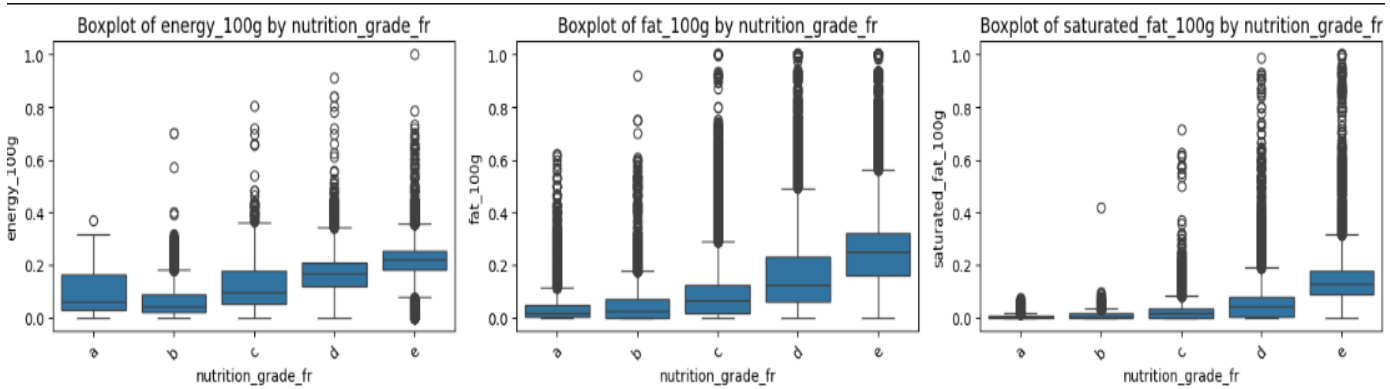
Analyse en Composantes Principales – **Projection des individus dans l'espace**

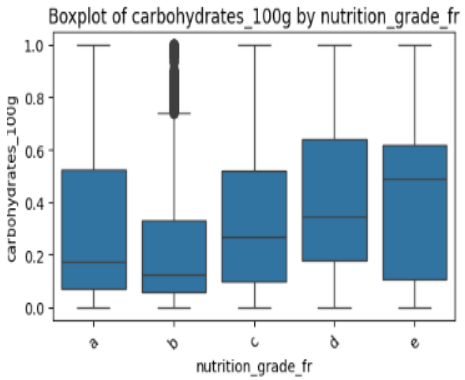


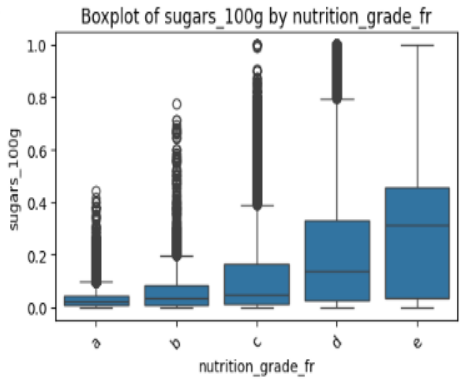
Un nuage de points représentant chaque produit projeté dans l'espace des deux premières composantes.

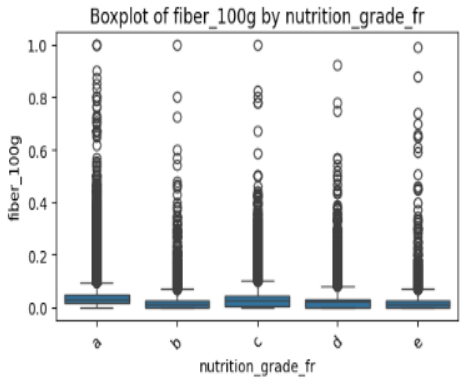
Les points sont colorés en fonction de la variable nutrition_grade_fr qui représente le score nutritionnel de chaque produit

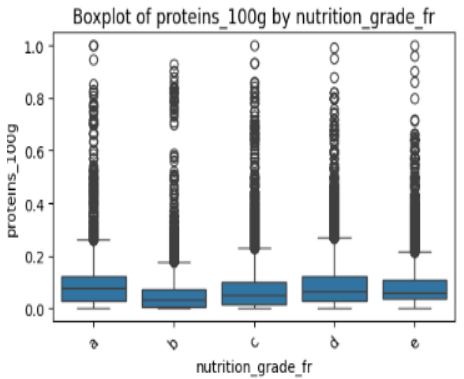
ANALYSE MULTIVARIEE
(ANOVA) Analyse de la Variance

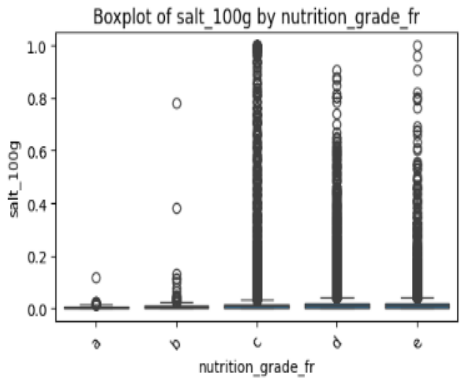












	F-statistic	p-value
energy_100g	23822.012559	0.0
fat_100g	15414.645893	0.0
saturated_fat_100g	27952.187035	0.0
carbohydrates_100g	3304.766667	0.0
sugars_100g	10349.573270	0.0
fiber_100g	1853.191059	0.0
proteins_100g	984.906388	0.0
salt_100g	763.642079	0.0

- La p-value égale à 0 pour toutes les features, ce qui signifie que les **différences entre les groupes de nutrition_grade_fr** sont statistiquement significatives pour toutes les variables analysées (énergie, graisses, sucres, etc.). Elle est inférieure à 0,05.
- Lorsque la p-value est inférieure au seuil critique (souvent 0.05), on rejette l'hypothèse nulle.
- Ici, avec des p-values égales à 0, nous pouvons conclure qu'il existe une différence statistiquement significative entre les groupes pour toutes les features étudiées, confirmée par les F-statistics élevées.

SYNTHESE

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

- **Principe de licéité, loyauté et transparence**

Les données doivent être traitées de façon légale, transparente, et avec un objectif légitime clairement expliqué aux utilisateurs.

- **Principe de limitation des finalités**

Les données doivent être collectées pour des objectifs spécifiques et ne pas être utilisées au-delà de ces objectifs.

- **Principe de minimisation des données**

Seules les données strictement nécessaires à la finalité doivent être collectées.

- **Principe d'exactitude**

Les données doivent être exactes et mises à jour, avec correction ou suppression des erreurs.

- **Principe de limitation de la conservation**

Les données doivent être conservées uniquement le temps nécessaire à la finalité du traitement.

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

- La base de données ne contient **aucune donnée** nous permettant **d'identifier une personne physique**

- Les opérations de **nettoyage** de la base de données **Open Food Facts** ne concernaient pas de données à **caractère personnel** mais des données **techniques** et **anosmisées** ajoutées par des utilisateurs.

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.

- On observe une majorité de produit classifiés sous un grade nutritionnel **D** et **E** qui respectivement signifie, **Mauvais** et **Très Mauvais**.

- Le traitement des valeurs **aberrantes** et des valeurs **manquantes** nous permet de mettre en place une technique **Machine Learning** (**ML**) consistant à la classification des produits par score nutritionnelle de **A** à **E**.

- L' **A**nalyse en **C**omposantes **P**rincipales (**ACP**) nous permet de réaliser l'importance des variables (**sugars_100g** et **carbohydrates_100g**), le taux de sucre et le taux de glucides.

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