SANTÉ PUBLIQUE FRANCE

Projet d'amélioration de la base de données Open Food Facts

Marion Dedieu

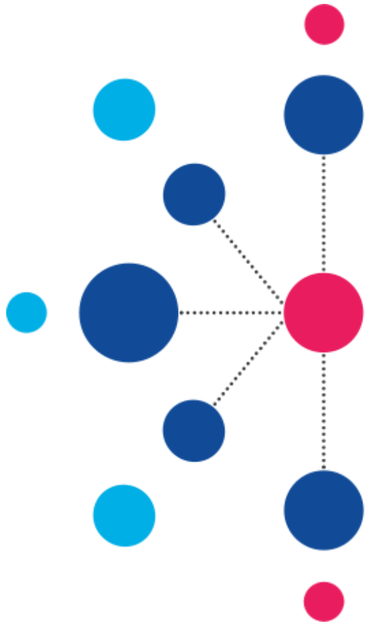
10/2023

SOMMAIRE

01.

CONTEXTE

Organisme
Problématique
Objectif



Santé publique France

PROBLEMATIQUE

Erreurs et données manquantes

OBJECTIF

Application d’autocomplétion

02.

ANALYSE DES DONNÉES

Jeu de données
Méthodes
Analyse univariée
Analyse multivariée

JEU DE DONNÉES

162 colonnes
320 772 lignes

SOURCE

Fichier Open Food Facts

VARIABLES

4 sections d’informations

TYPES

56 colonnes Texte
106 colonnes Numérique

COMPLÉTUDE

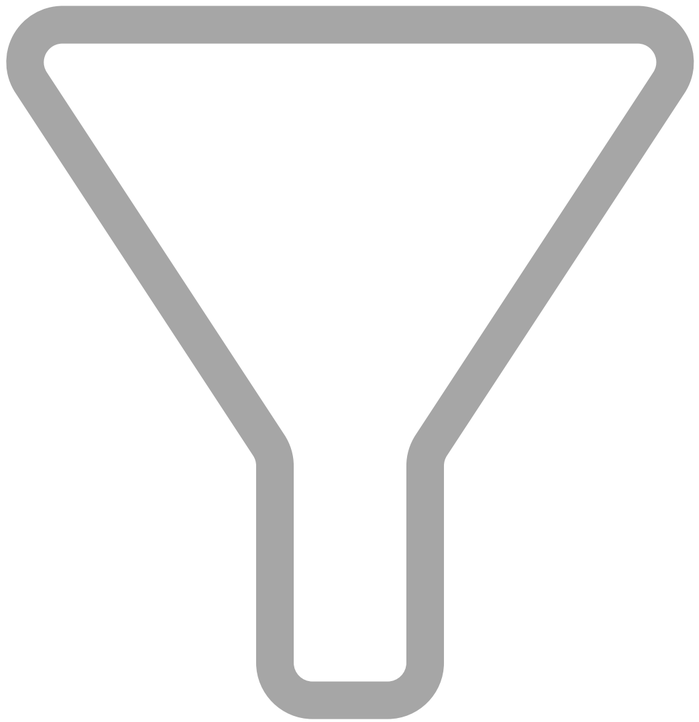
23,78 %

NUTRI-SCORE

= N – P

Class	Score ranges	Colour
A	Min - 1	Dark green
B	0 - 2	Light green
C	3 - 10	Light orange
D	11 - 18	Orange
E	19 - max	Dark orange

MÉTHODE DE NETTOYAGE



Suppression colonnes peu remplies

Taux de valeurs manquantes > 50%

Suppression colonnes inutiles

Données non pertinentes pour l'analyse
et l'application

Suppression des doublons

34 244 lignes

Traitement des valeurs aberrantes

Minimums négatifs
Maximums erronés ou dans une autre unité

Traitement des valeurs manquantes

Suppression ou techniques d'imputation

MÉTHODE D'ANALYSE

Univariée

Variables quantitatives :
Moyenne
Médiane
Écart-type
Variance
Min, Max
Q1, Q3
Skewness, Kurtosis
Histogramme

Qualitatives catégorielles :
Répartition

Bivariée

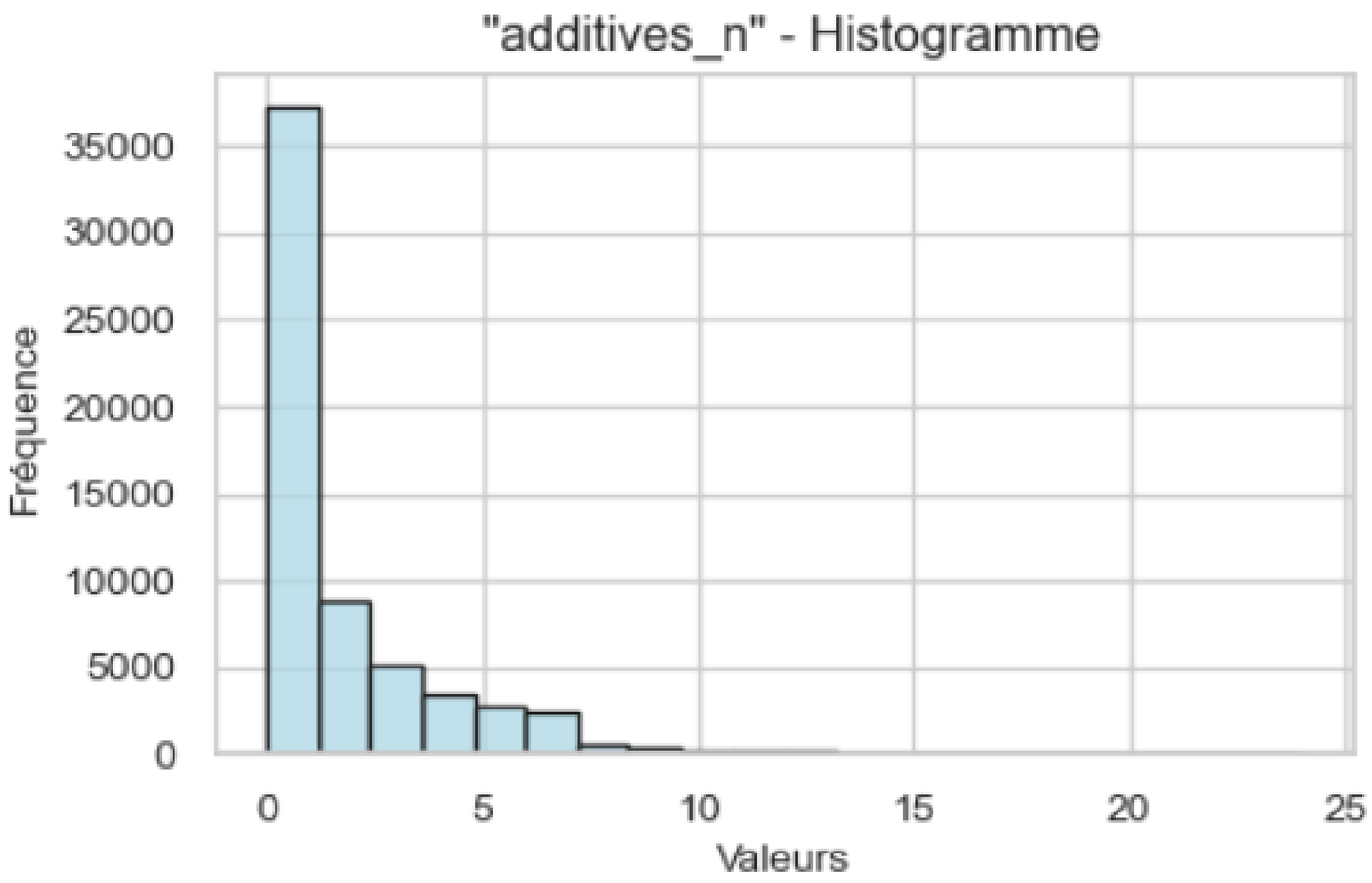
Heatmap des corrélations
Nuages de points
Diagrammes en bâtons
Diagrammes radar

Multivariée

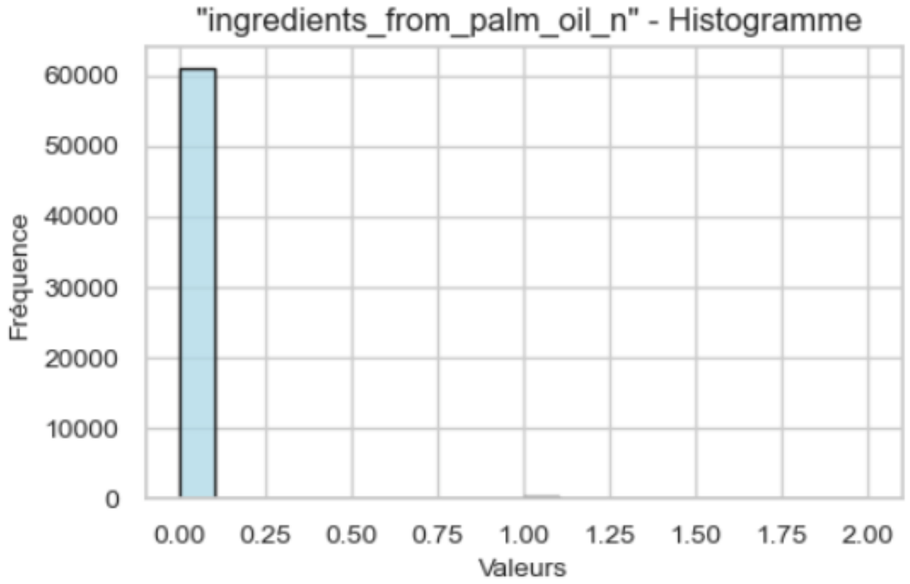
ACP
ANOVA

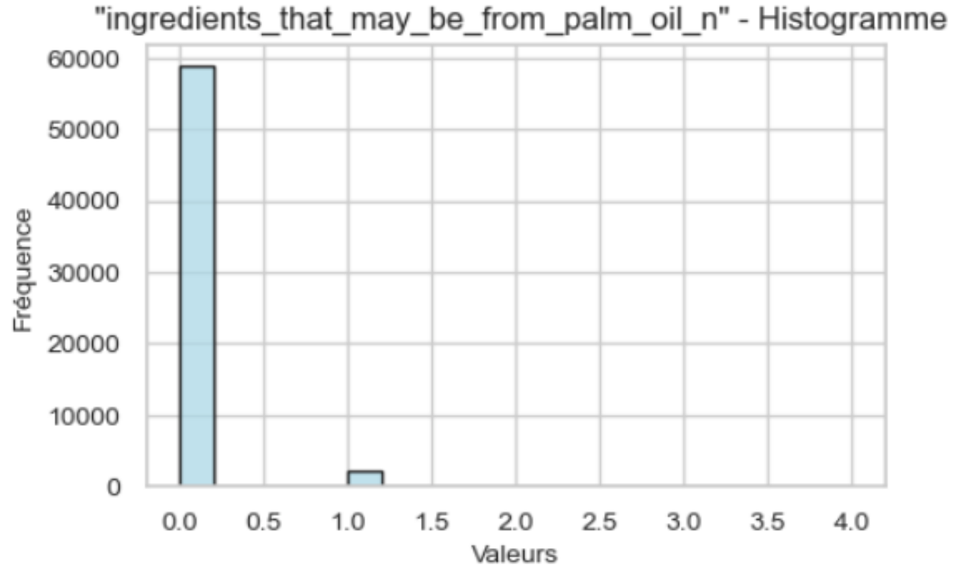
ANALYSE UNIVARIÉE

NOMBRE D'ADDITIFS

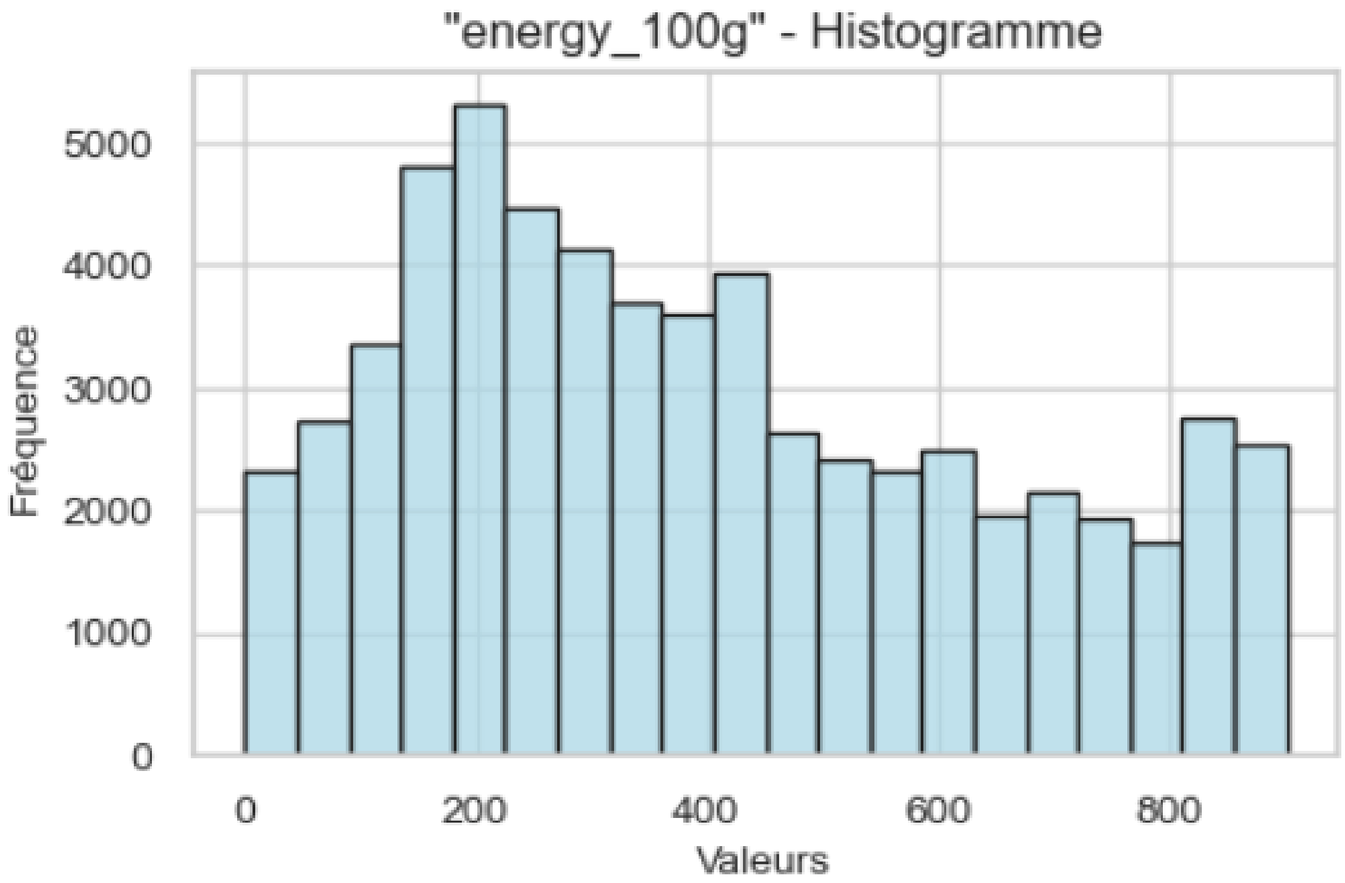


NOMBRE D'INGRÉDIENTS PROVENANT/ POUVANT PROVENIR DE L'HUILE DE PALME

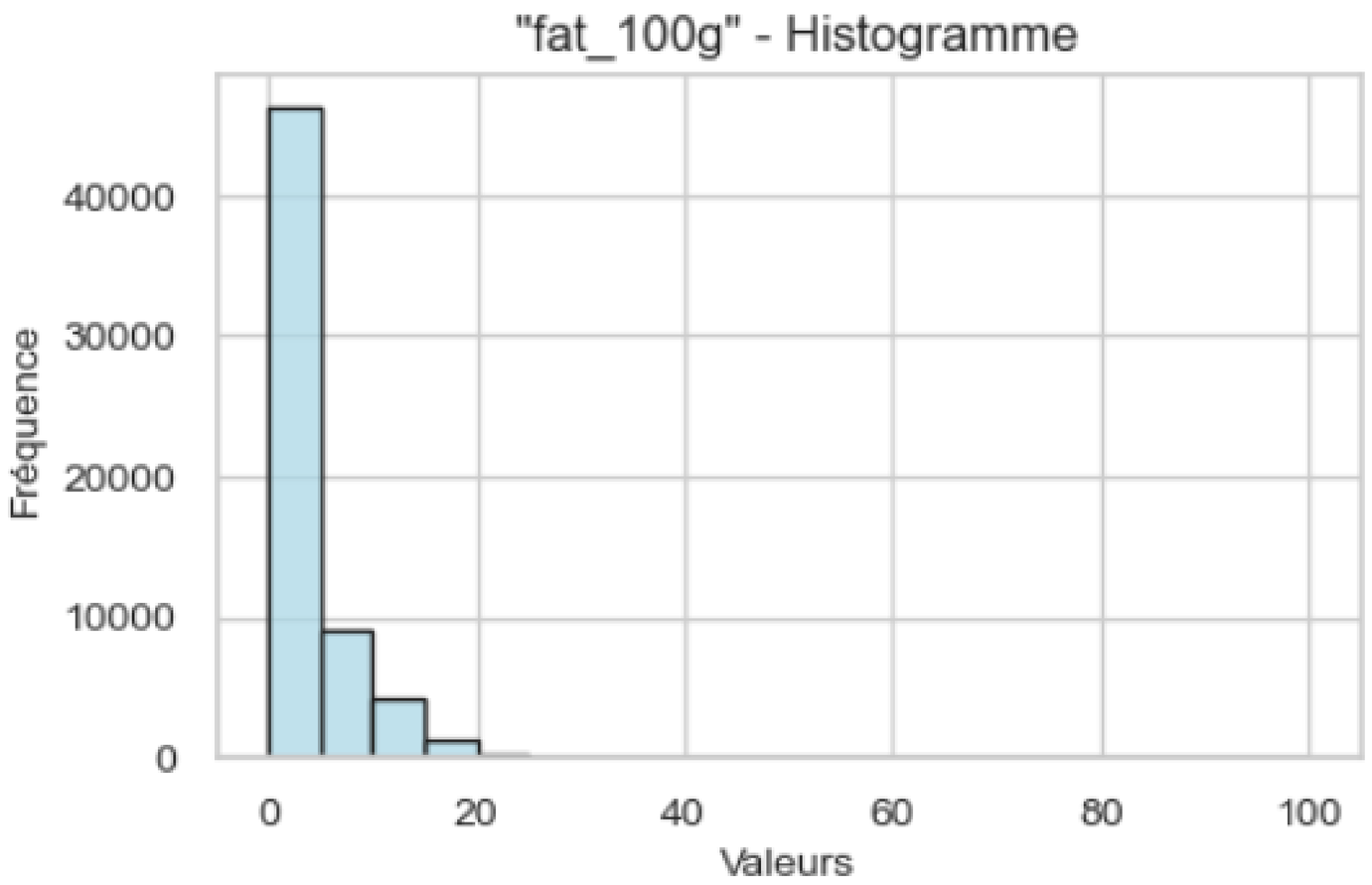




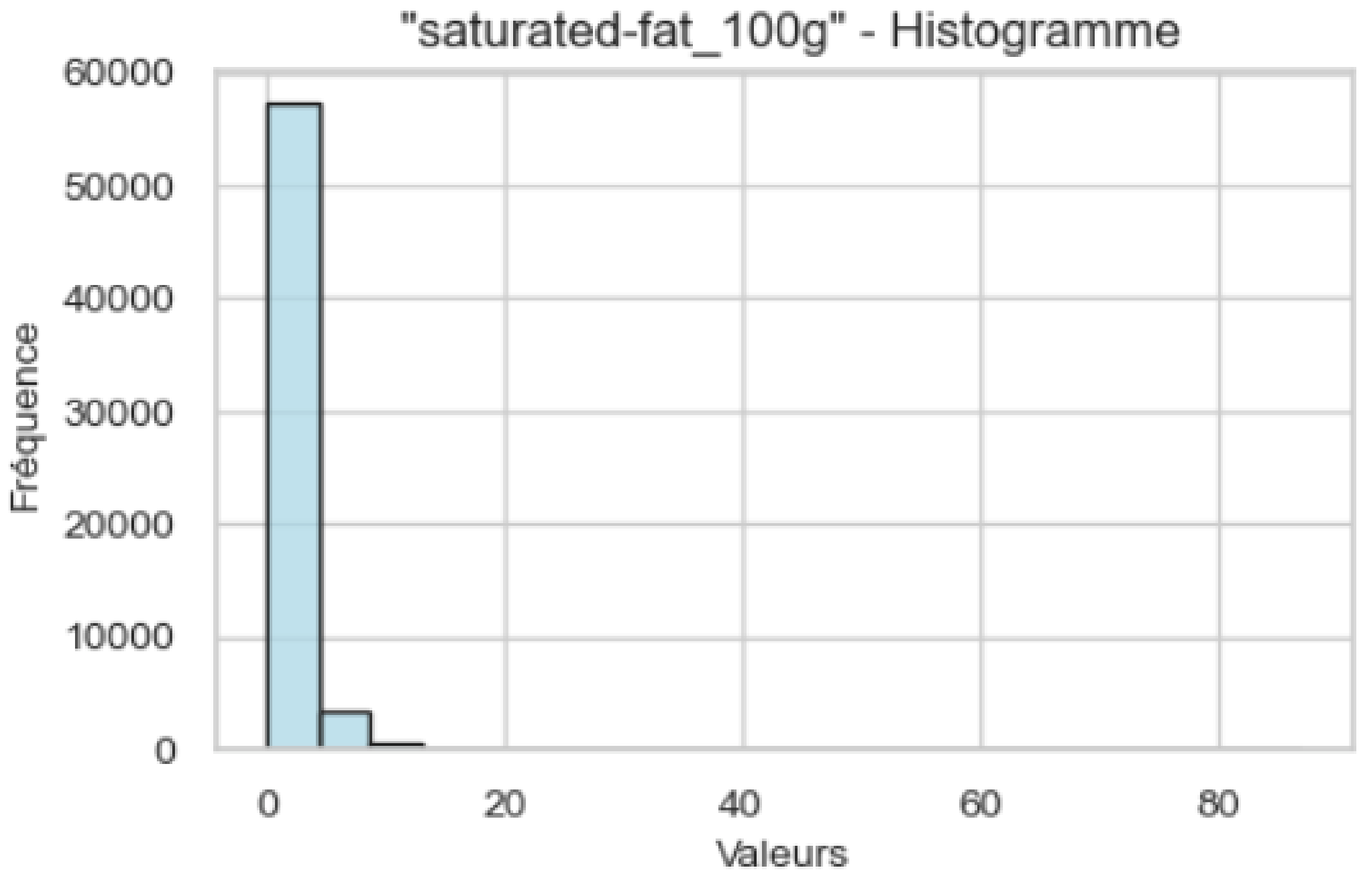
ÉNERGIE POUR 100G



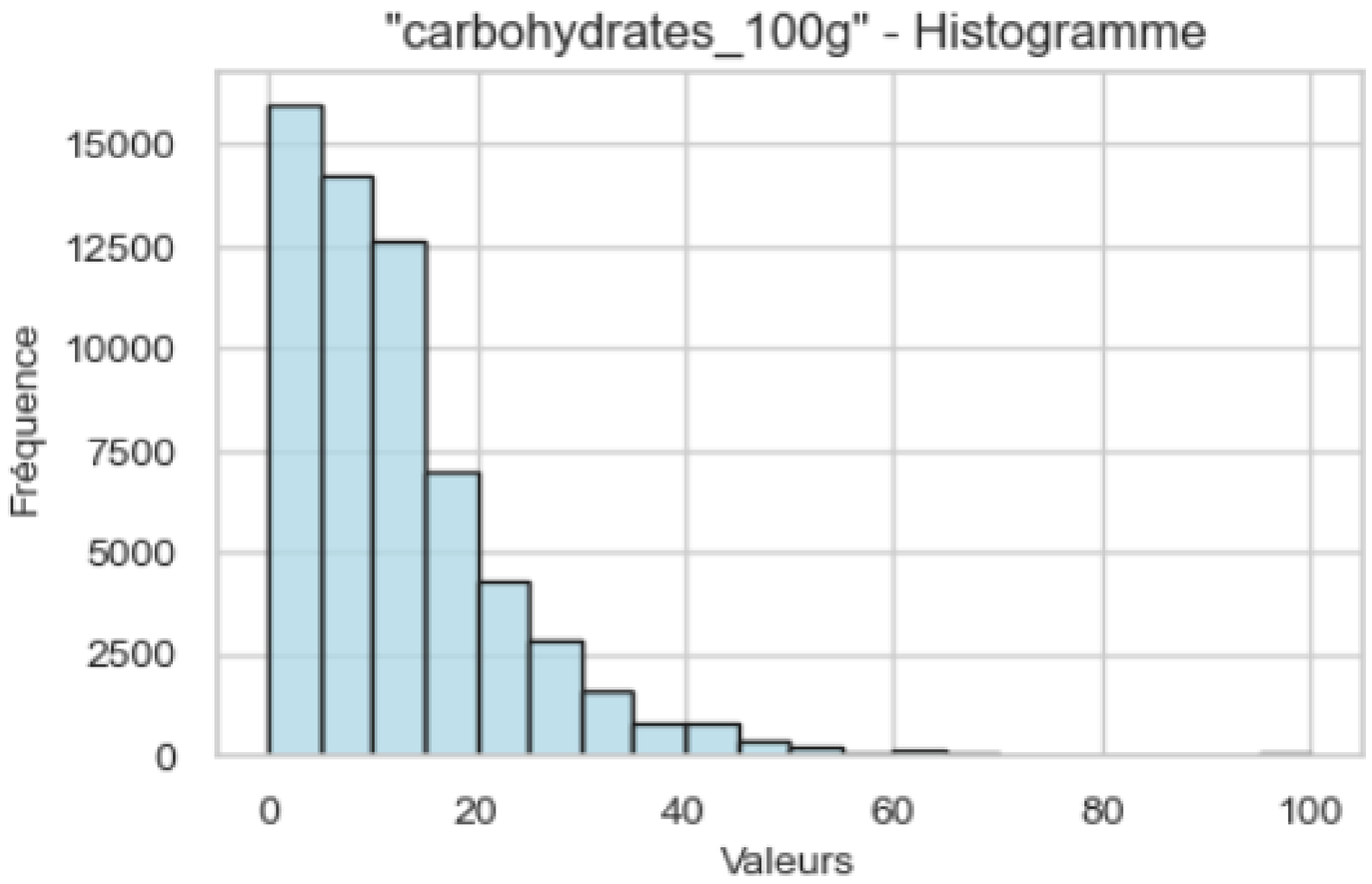
LIPIDES POUR 100G



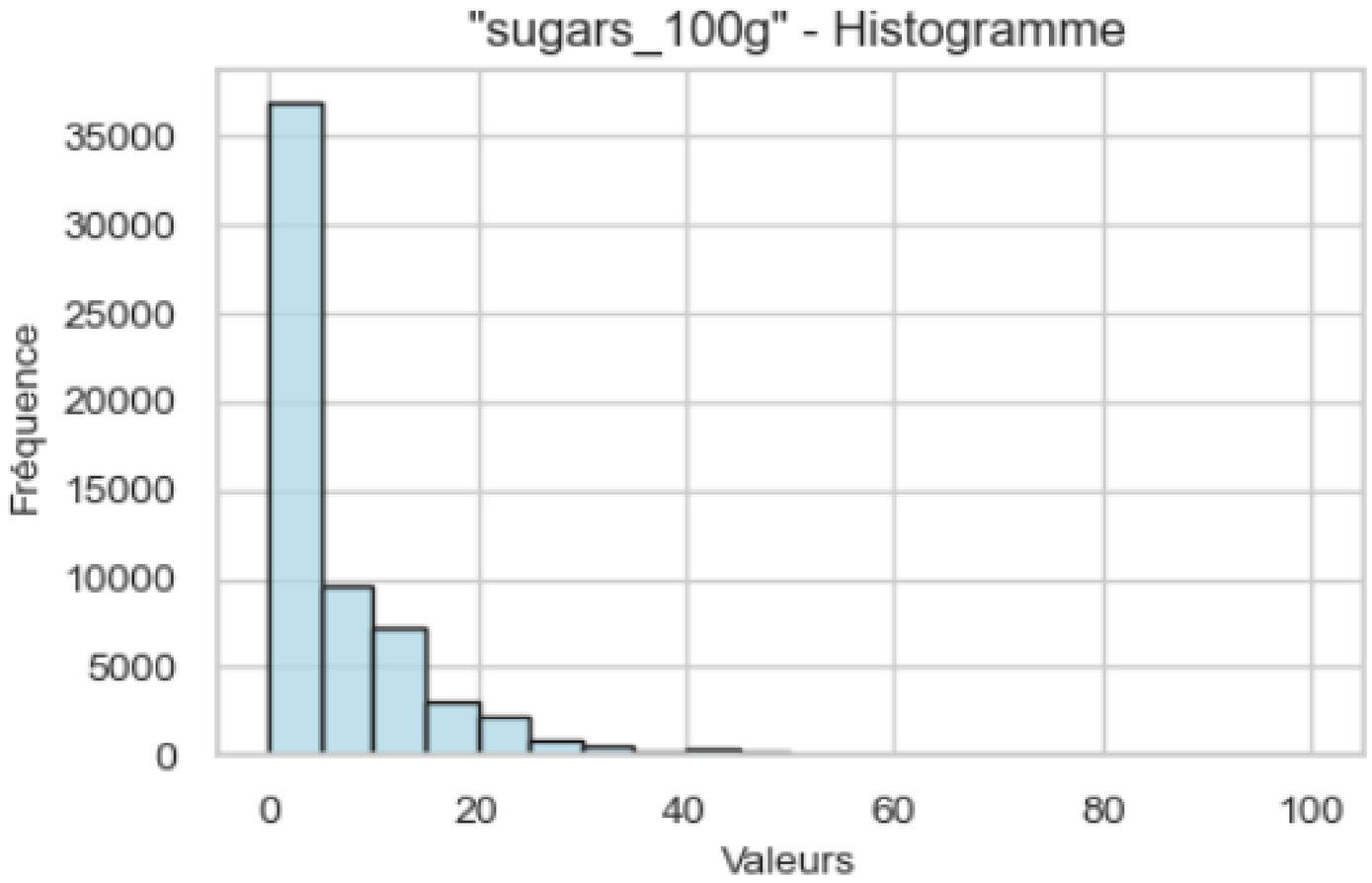
GRAISSES SATURÉES POUR 100G



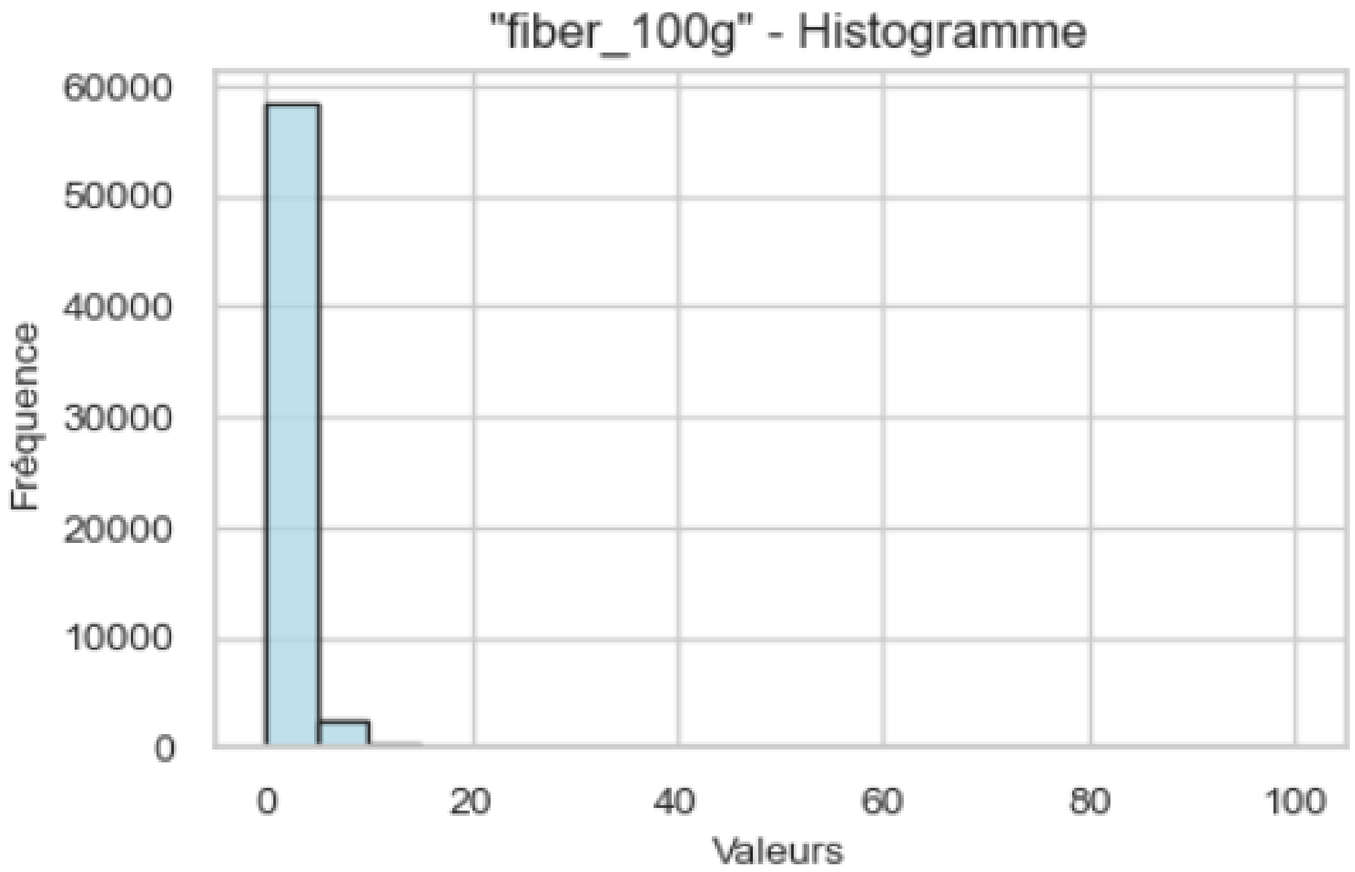
GLUCIDES POUR 100G



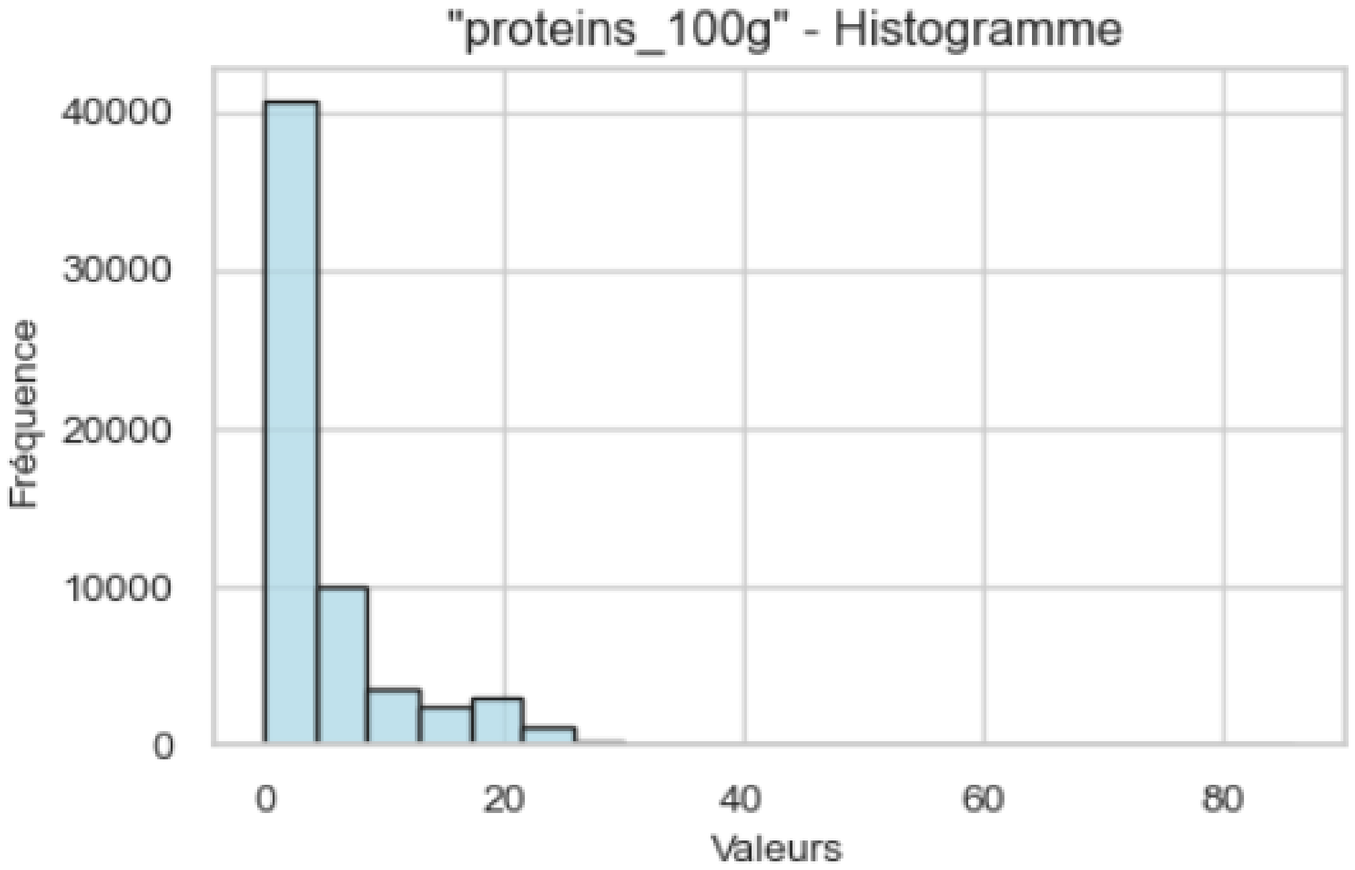
SUCRES POUR 100G



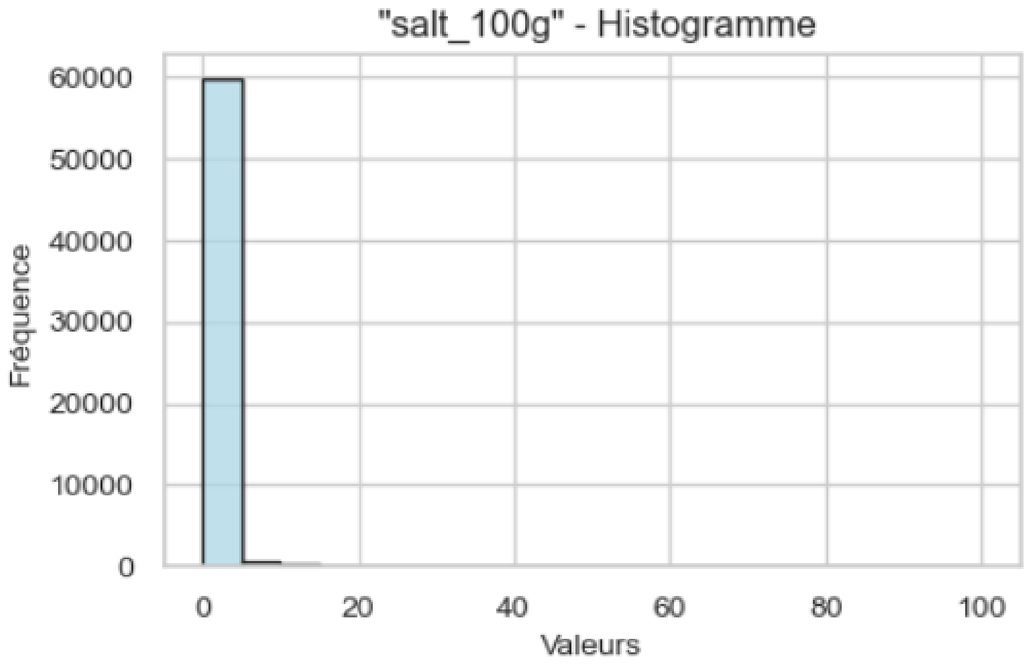
FIBRES POUR 100G

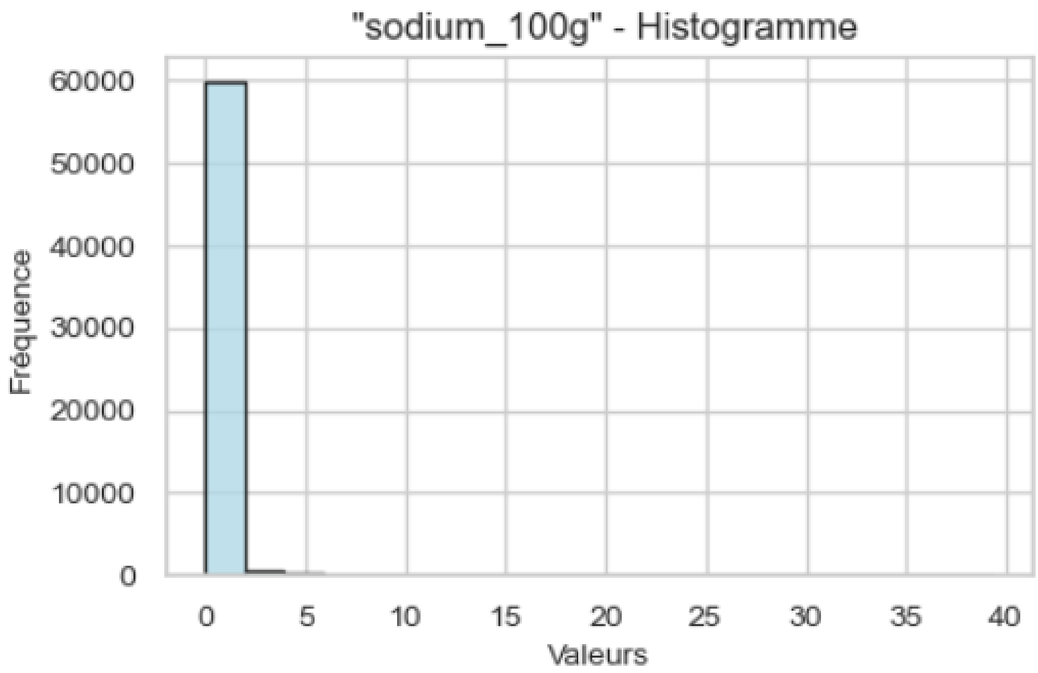


PROTÉINES POUR 100G

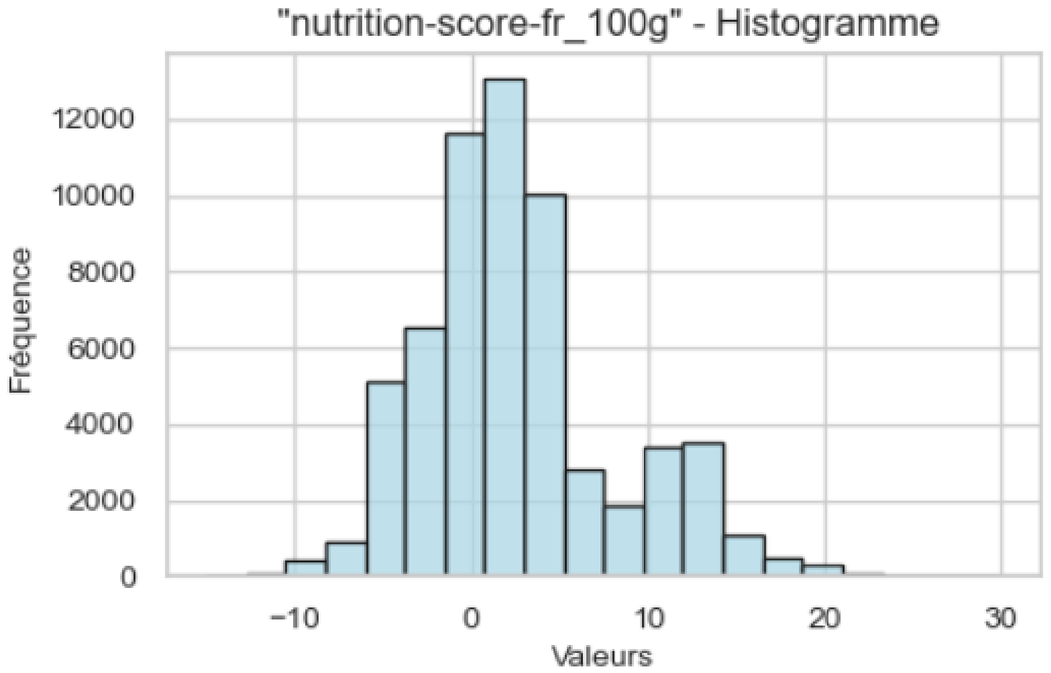


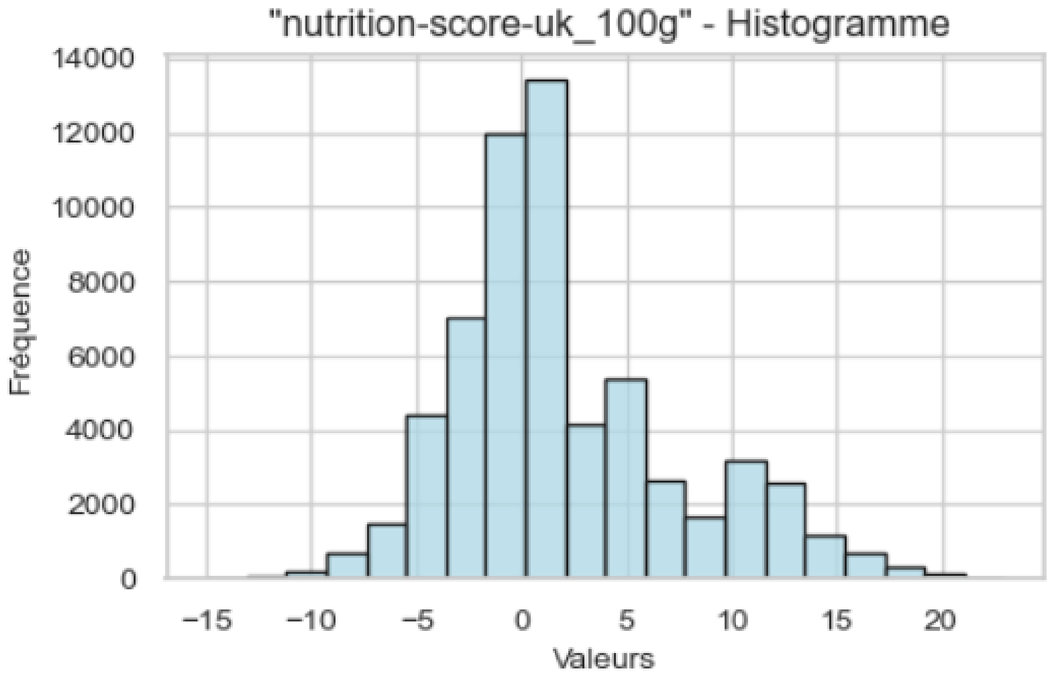
SEL & SODIUM POUR 100G



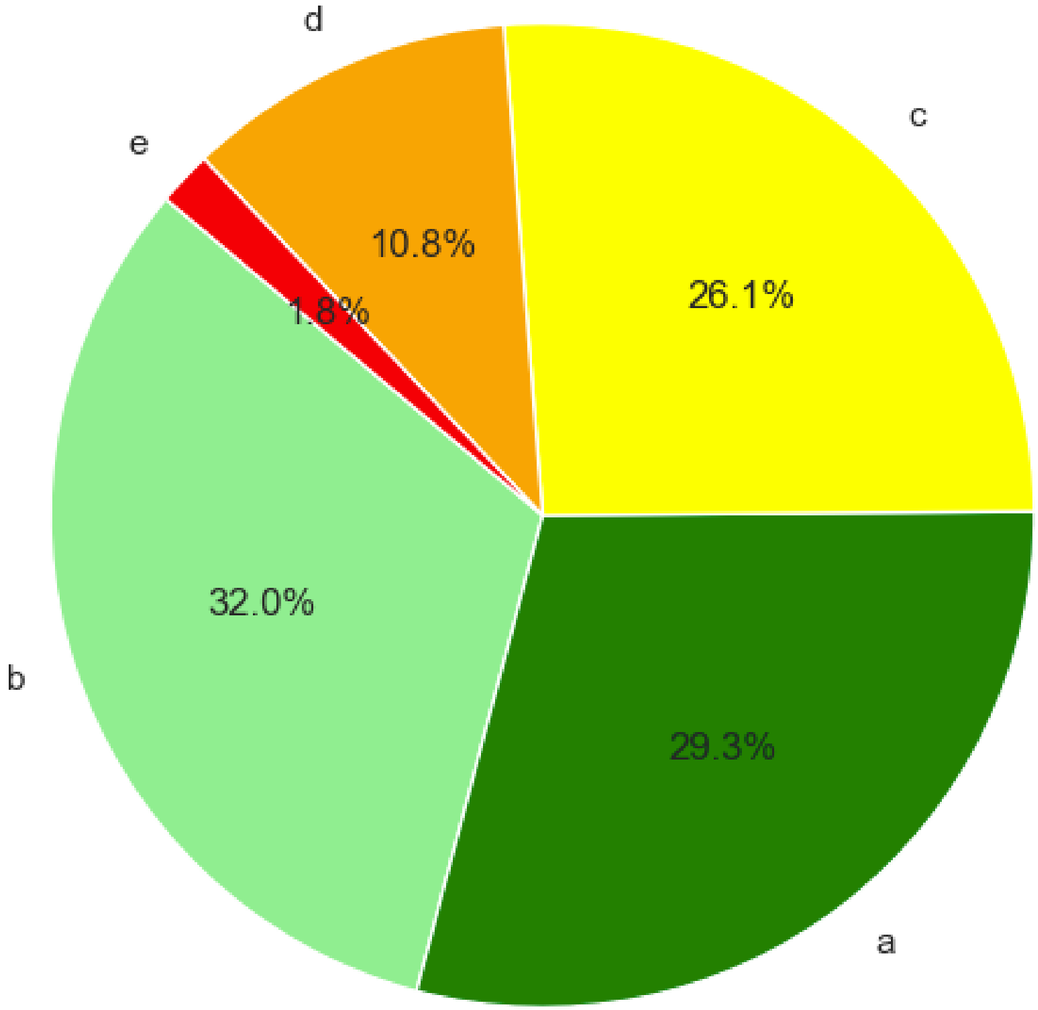


NUTRI-SCORE POUR 100G

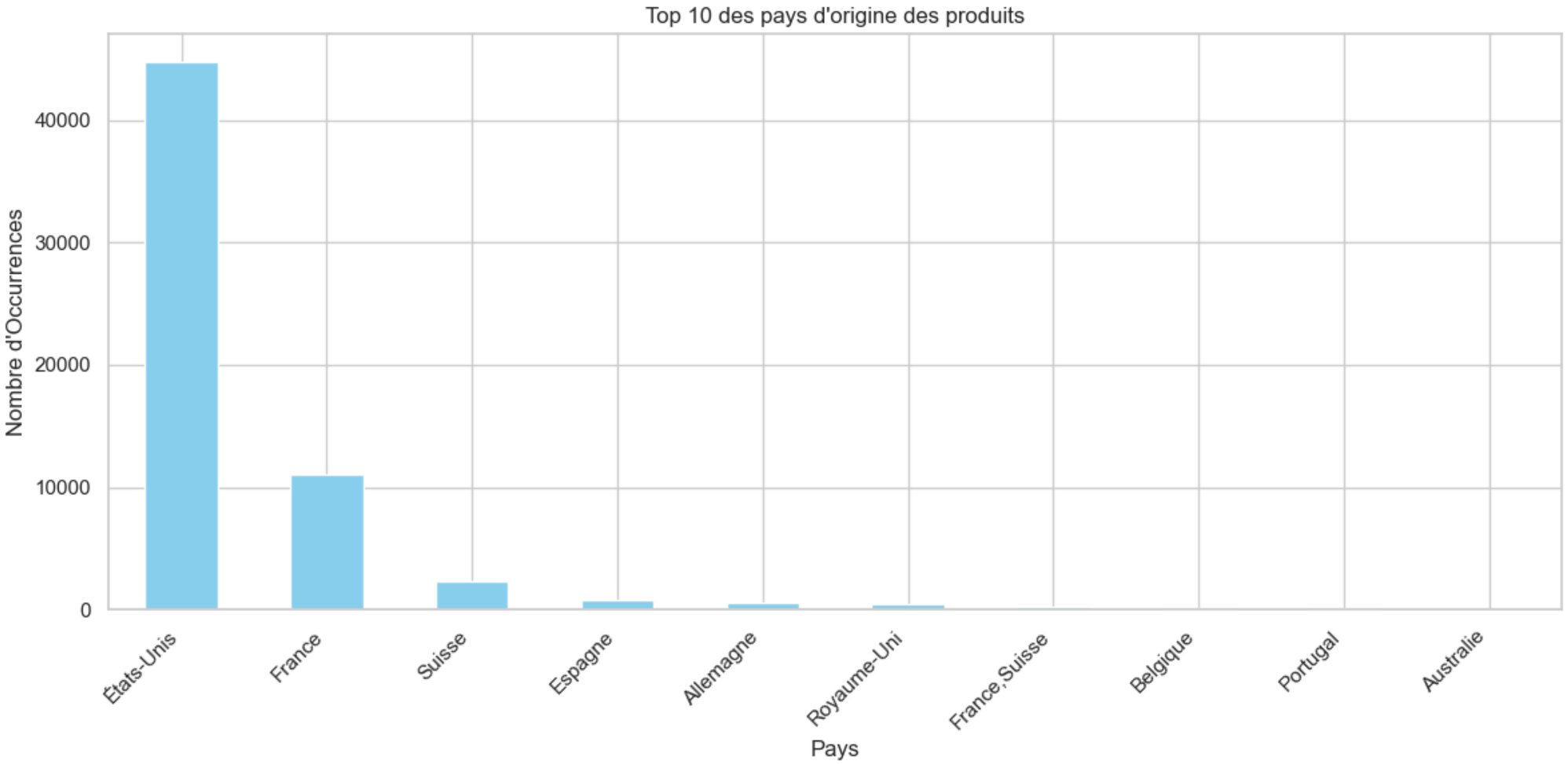




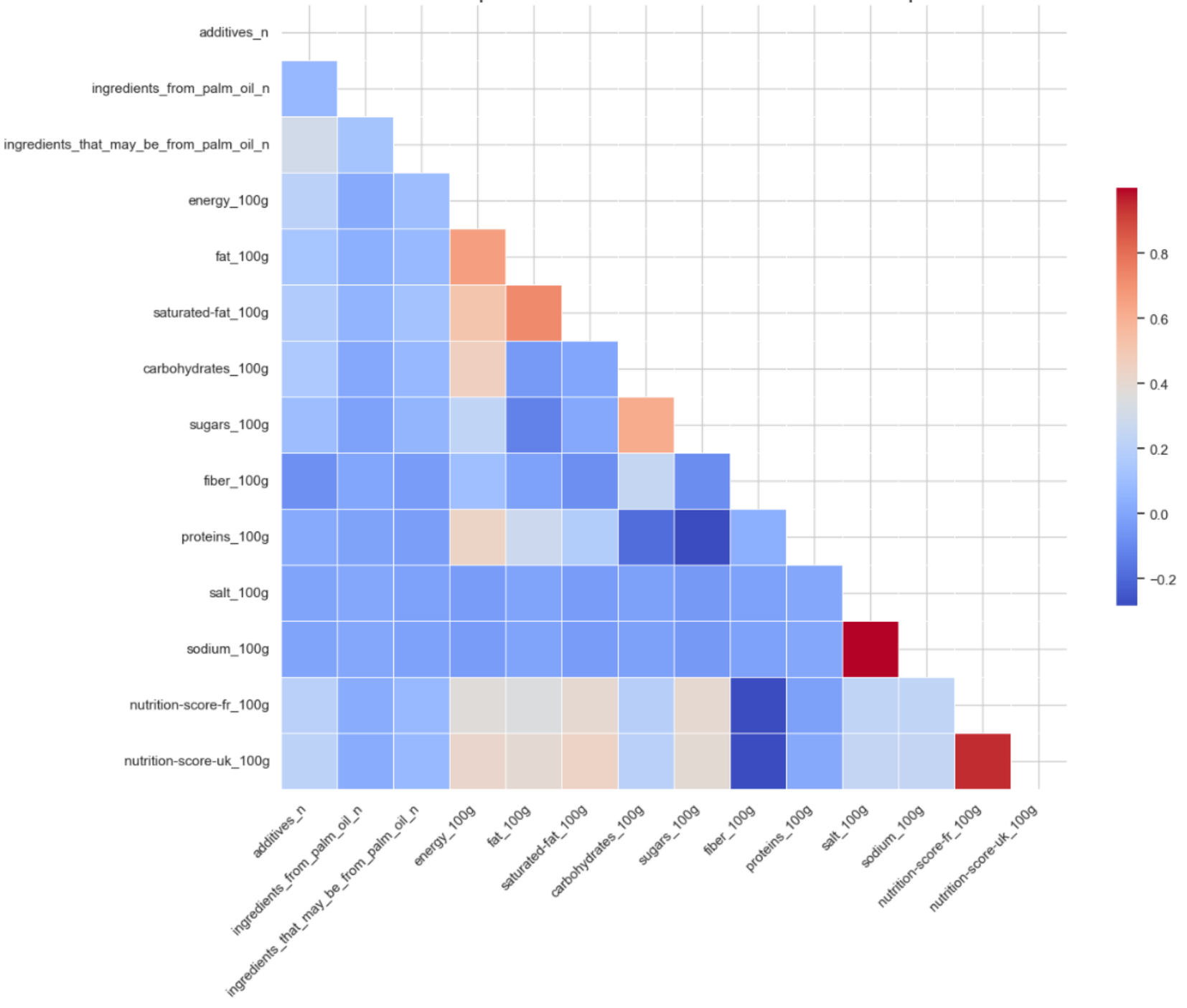
RÉPARTITION DU NUTRI-SCORE



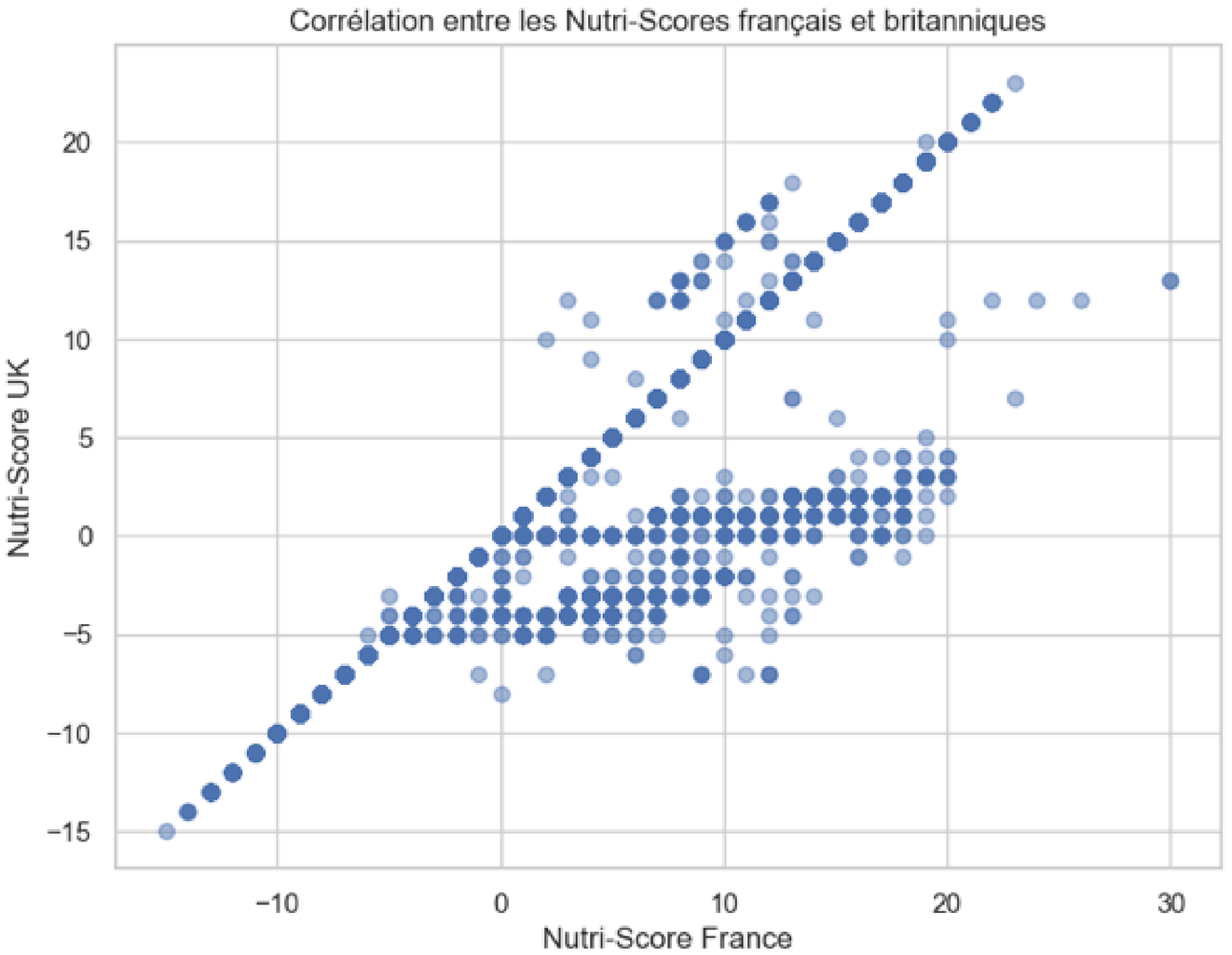
ORIGINE DES PRODUITS



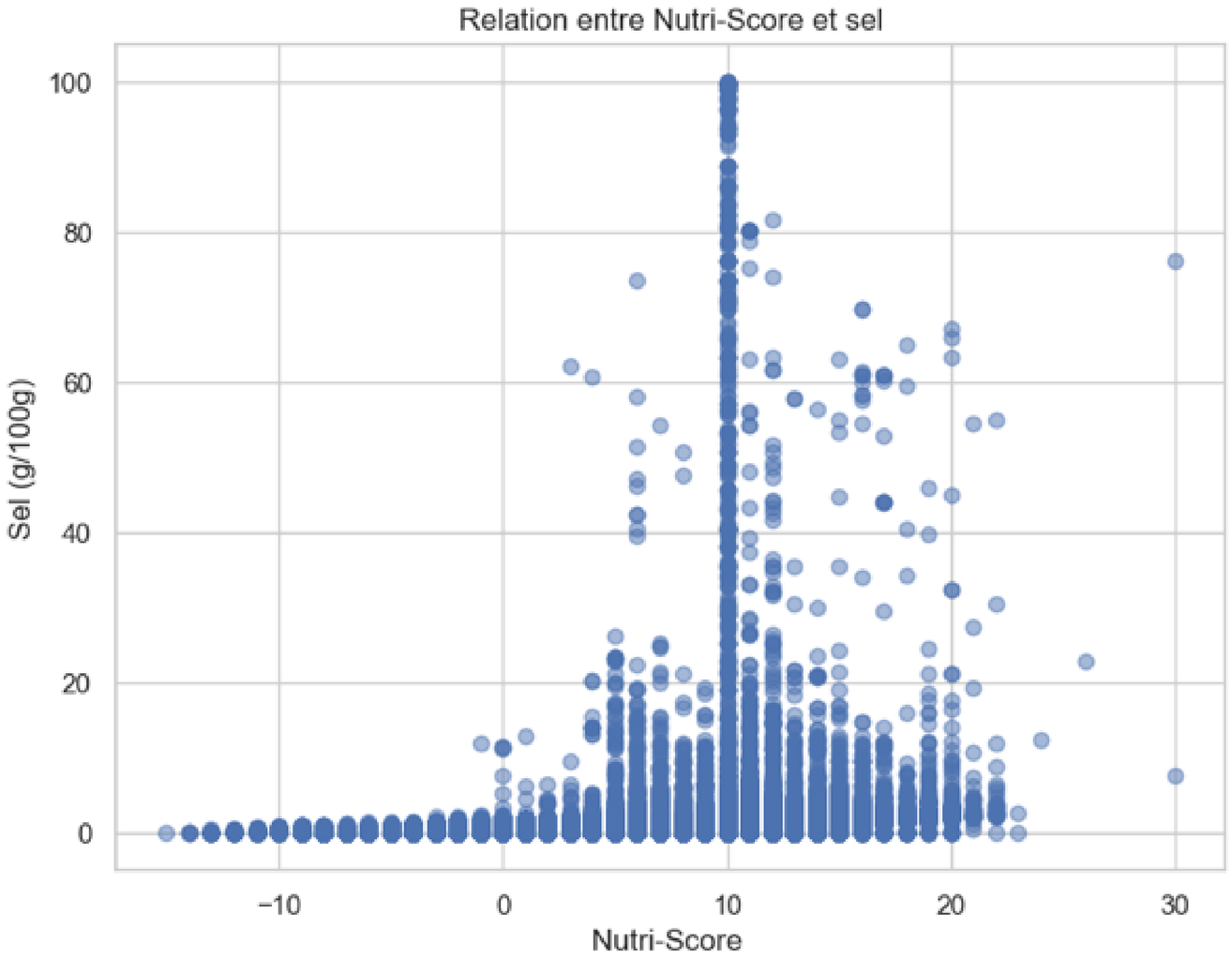
ANALYSE BIVARIÉE


Heatmap de corrélation entre les variables numériques
additives_n
ingredients_from_palm_oil_n
ingredients_that_may_be_from_palm_oil_n
energy_100g
fat_100g
saturated-fat_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
salt_100g
sodium_100g
nutrition-score-fr_100g
nutrition-score-uk_100g
0.8
0.6
0.4
0.2
0.0
−0.2

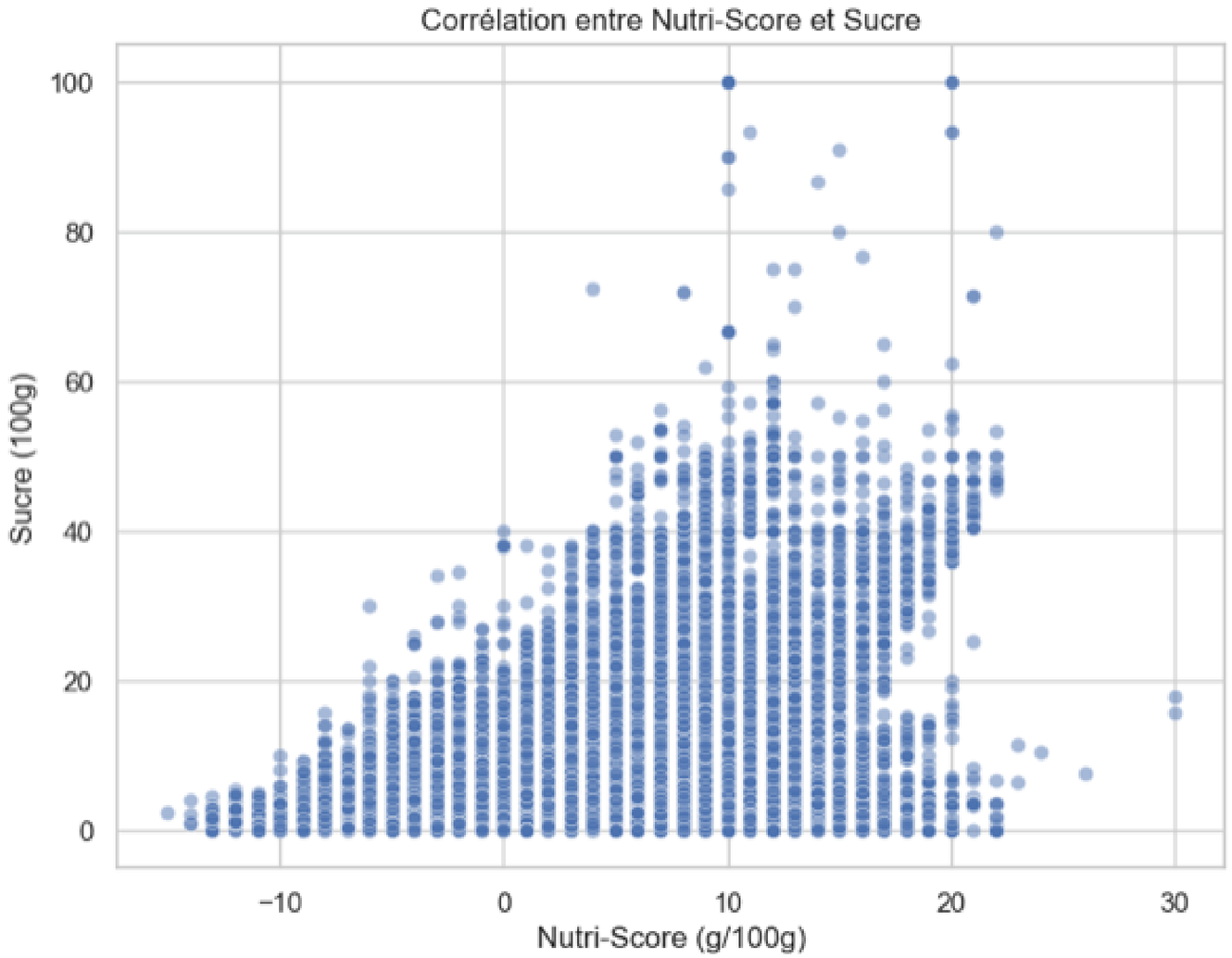
NUTRI-SCORE FR VS. UK



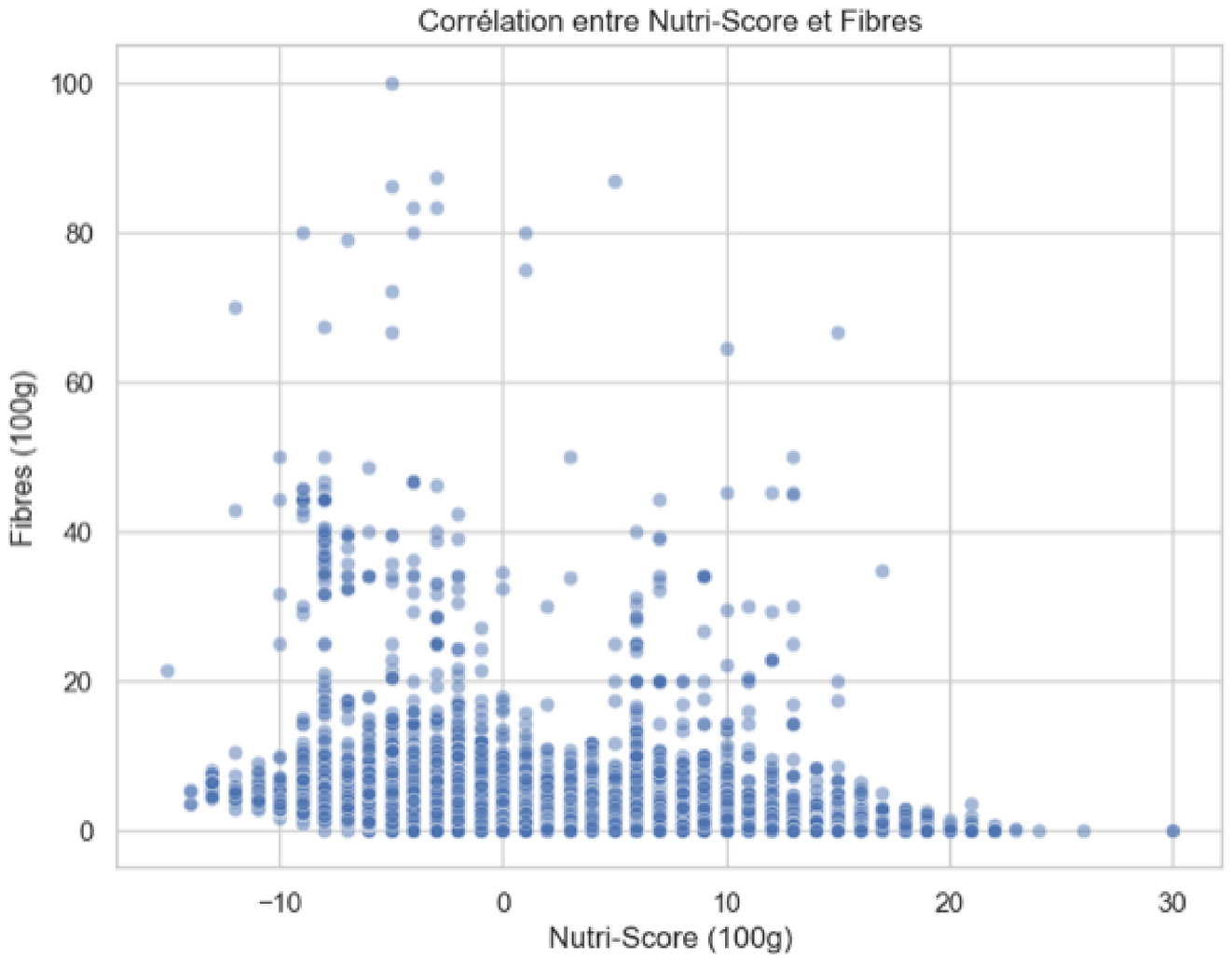
NUTRI-SCORE VS. SEL



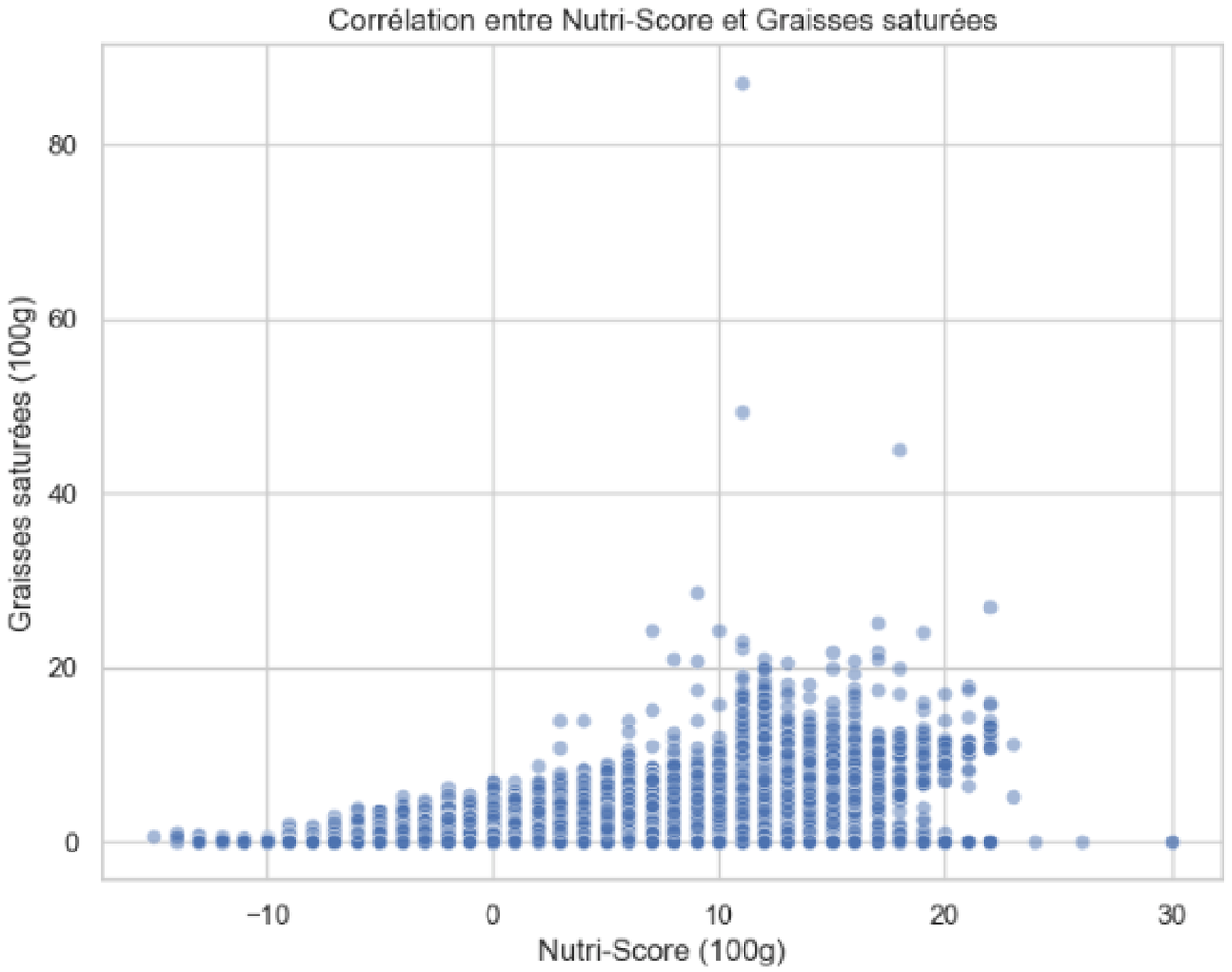
NUTRI-SCORE VS. SUCRE



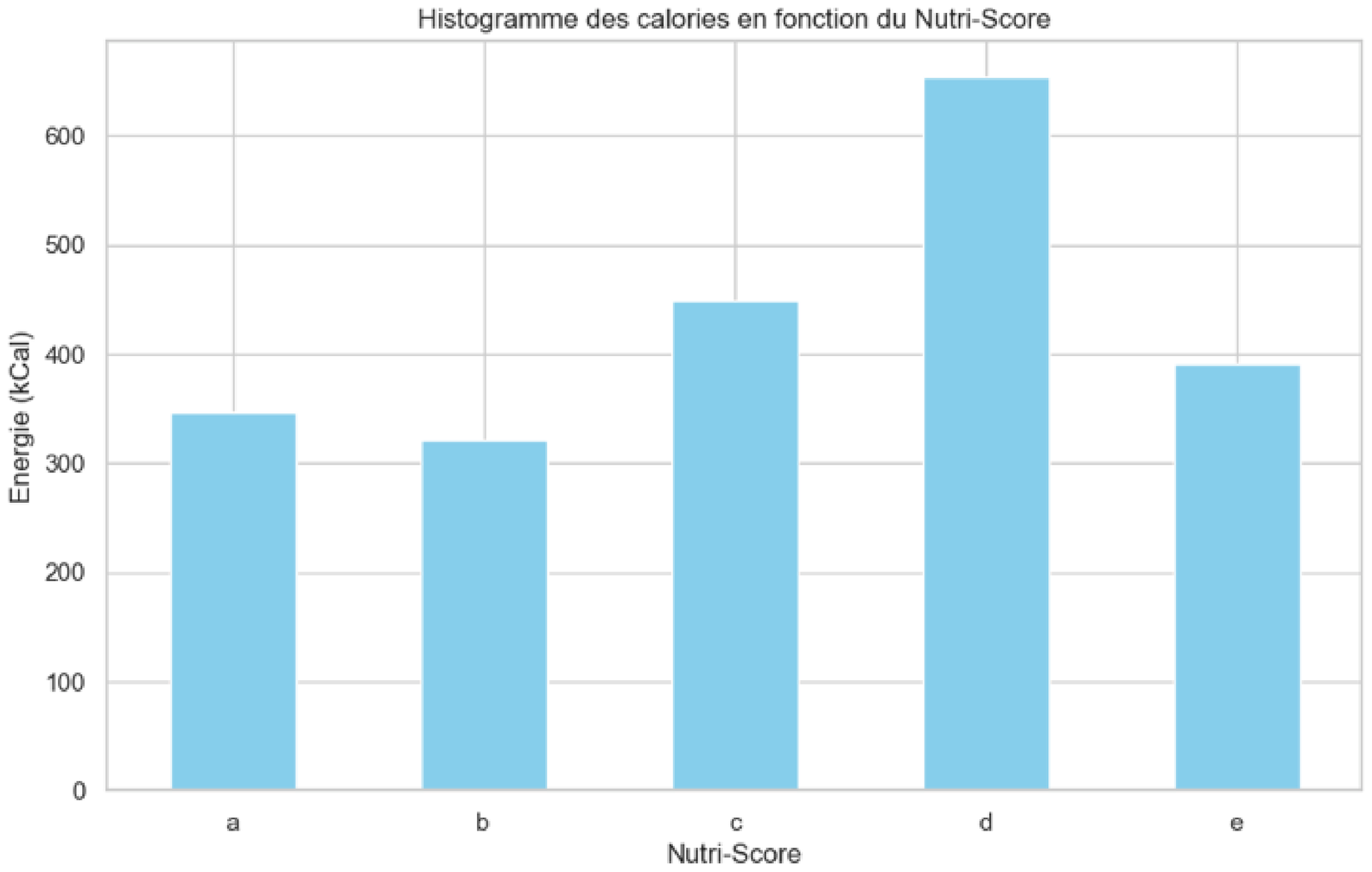
NUTRI-SCORE VS. FIBRES



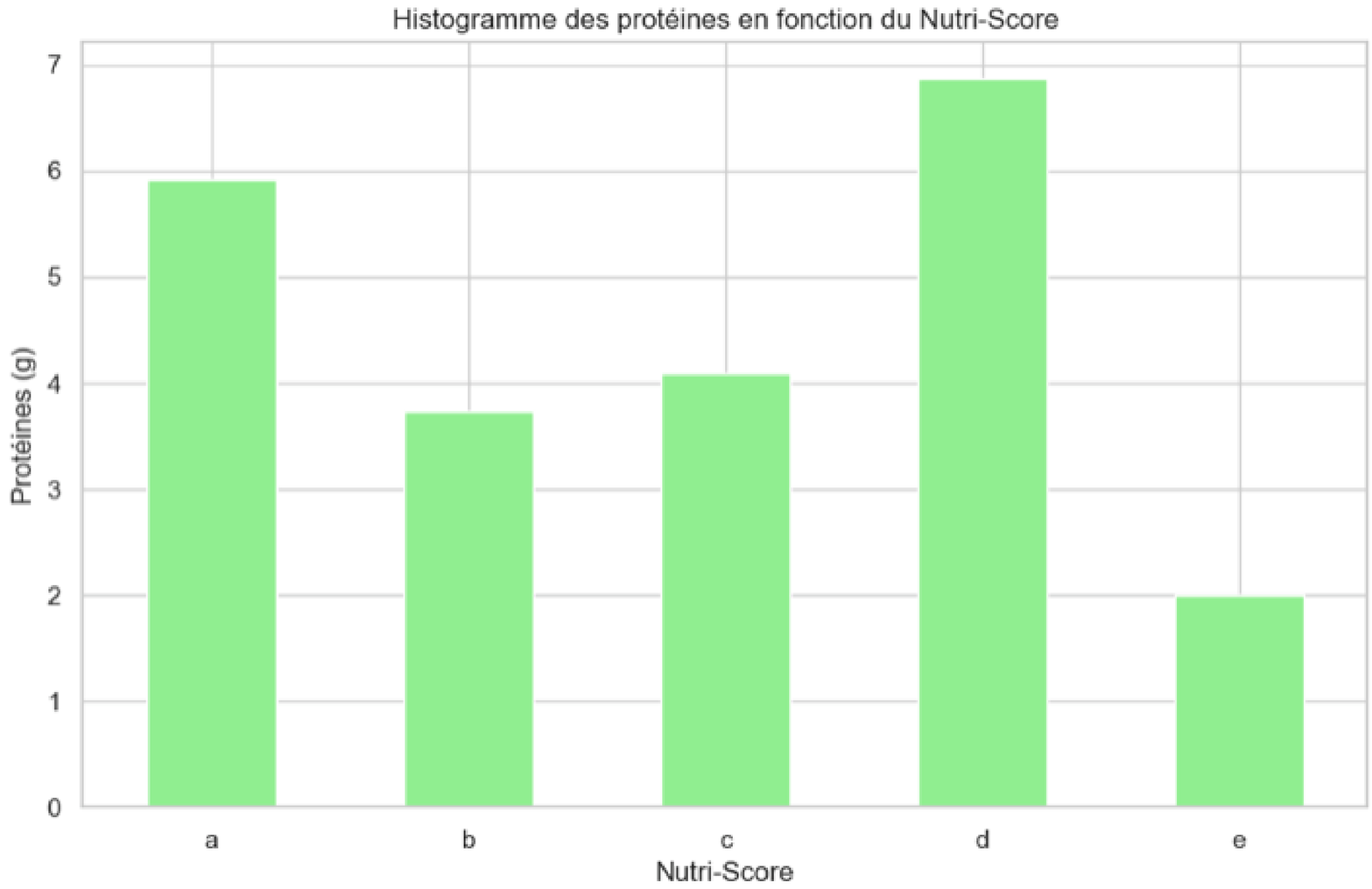
NUTRI-SCORE VS. GRAISSES SATURÉES



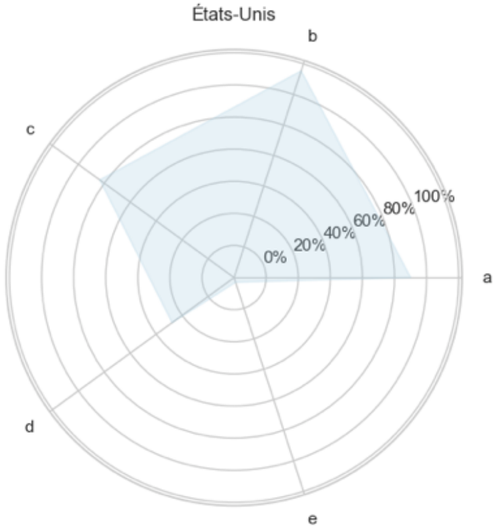
ÉNERGIE PAR NUTRI-SCORE

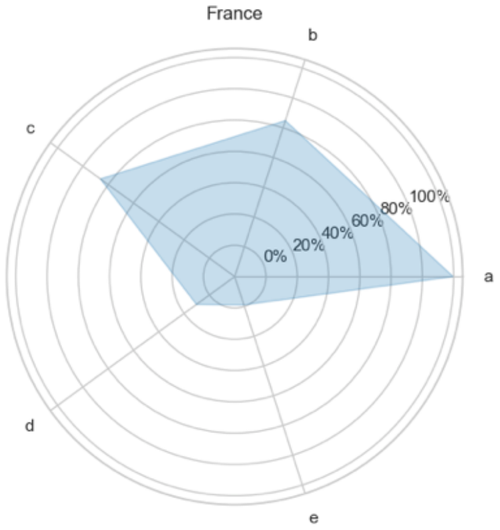


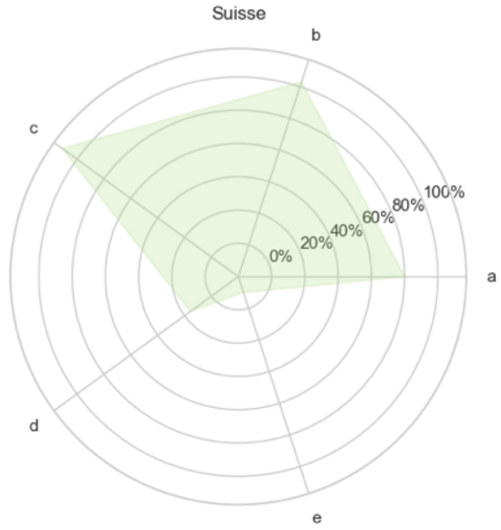
PROTÉINES PAR NUTRI-SCORE

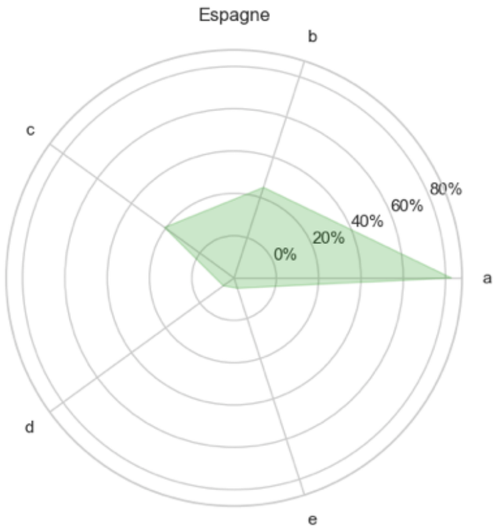


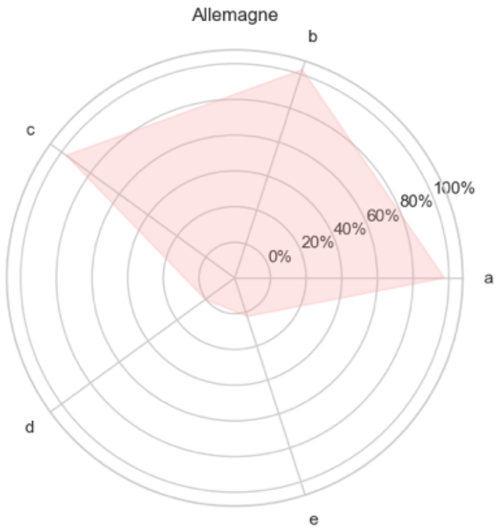
NUTRI-SCORE PAR PAYS

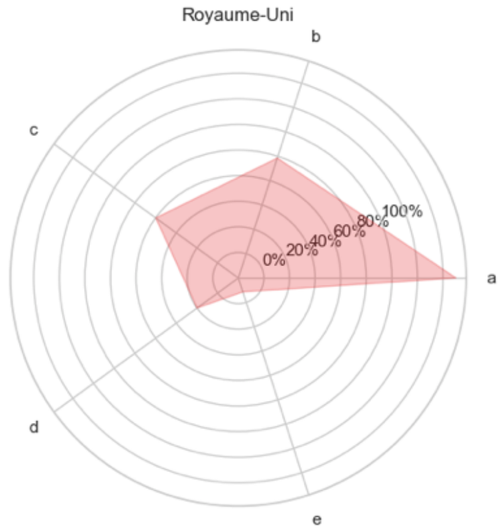






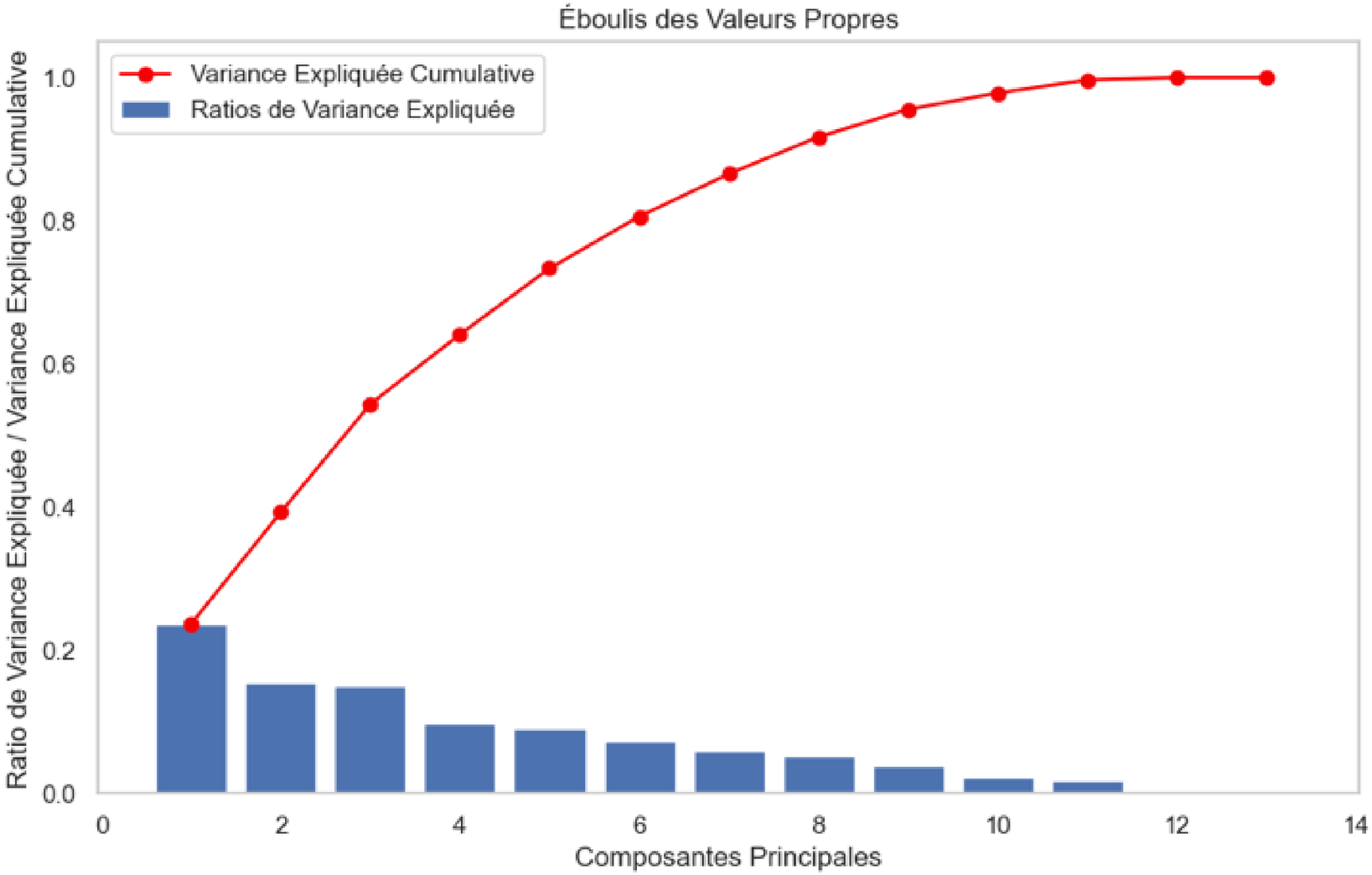




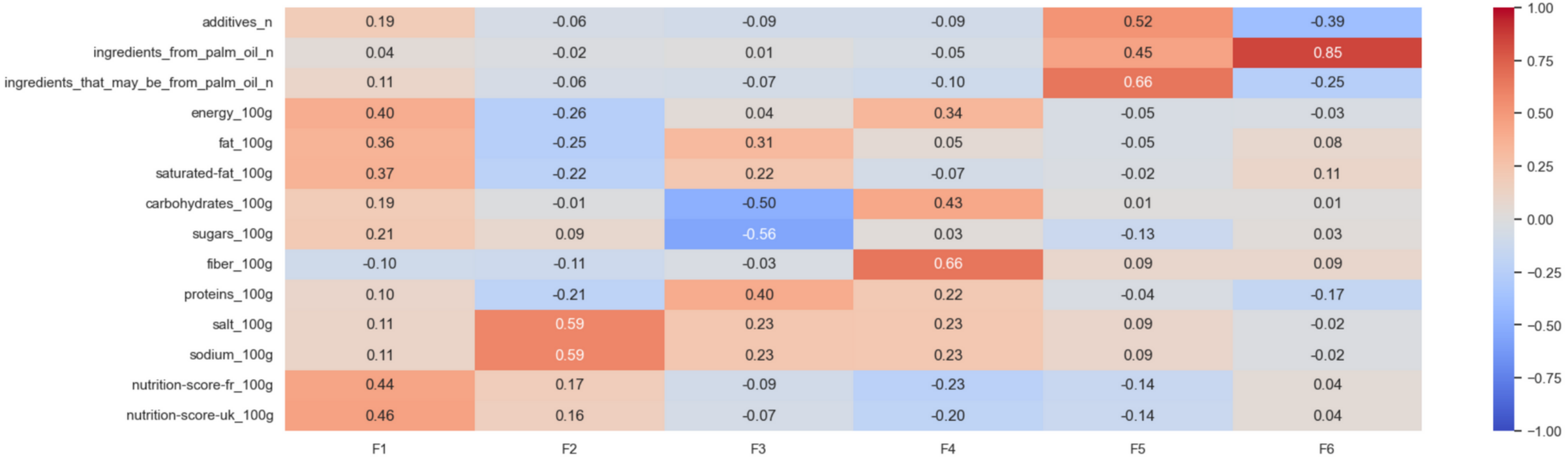


ANALYSE MULTIVARIÉE

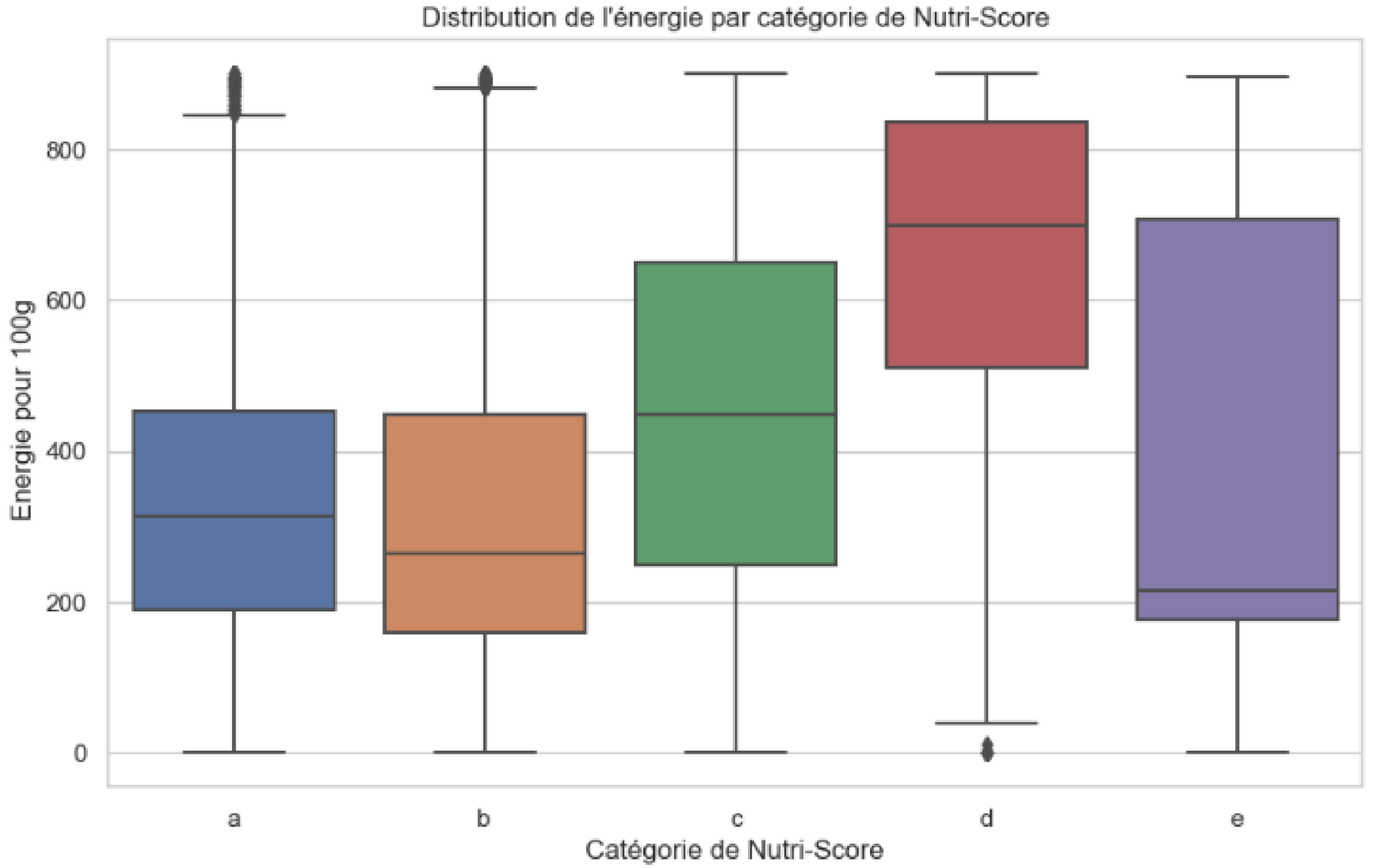
ACP


Éboulis des Valeurs Propres
Variance Expliquée Cumulative
Ratios de Variance Expliquée
Ratio de Variance Expliquée / Variance Expliquée Cumulative
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
Composantes Principales

6 PREMIÈRES COMPOSANTES



ANOVA


Distribution de l'énergie par catégorie de Nutri-Score
Energie pour 100g
800
600
400
200
0
a
b
c
d
e
Catégorie de Nutri-Score

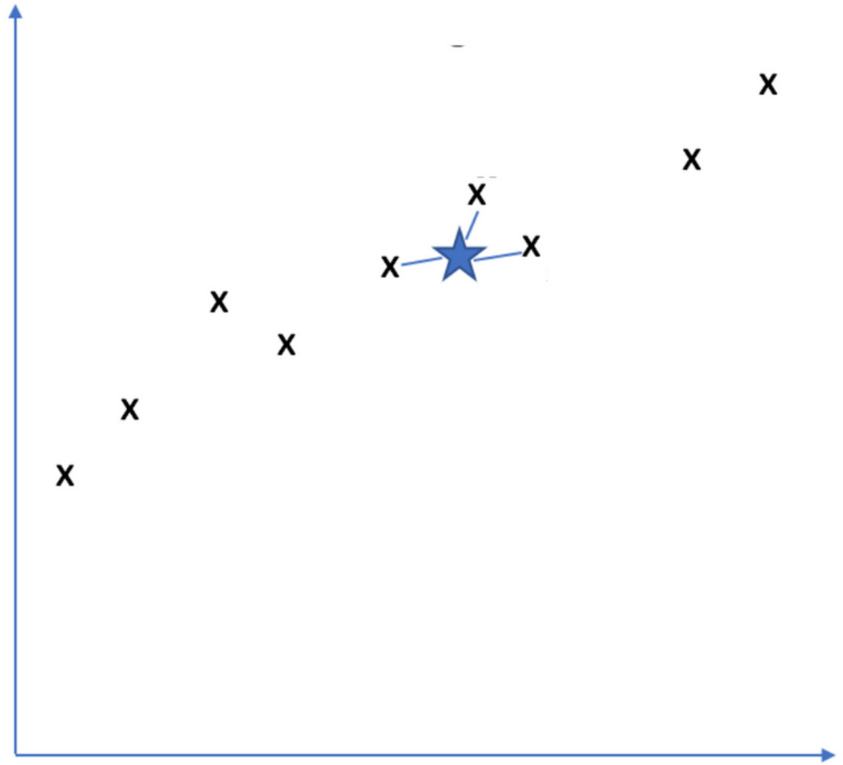
03.

APPLICATION D'AUTOCOMPLÉTION

Méthode
Résultats

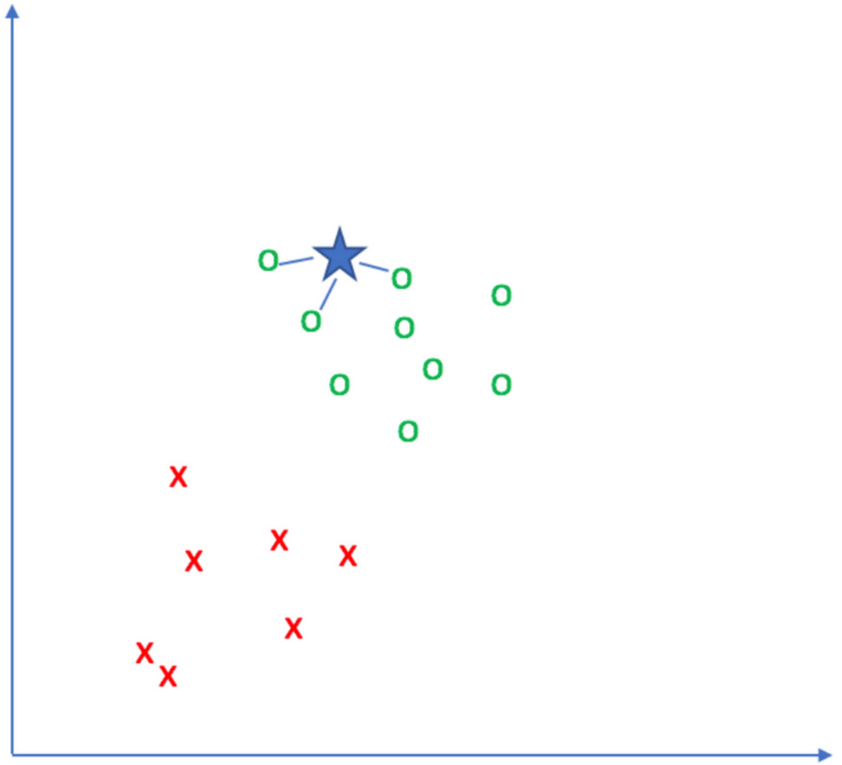
K-NN RÉGRESSION

Variables continues



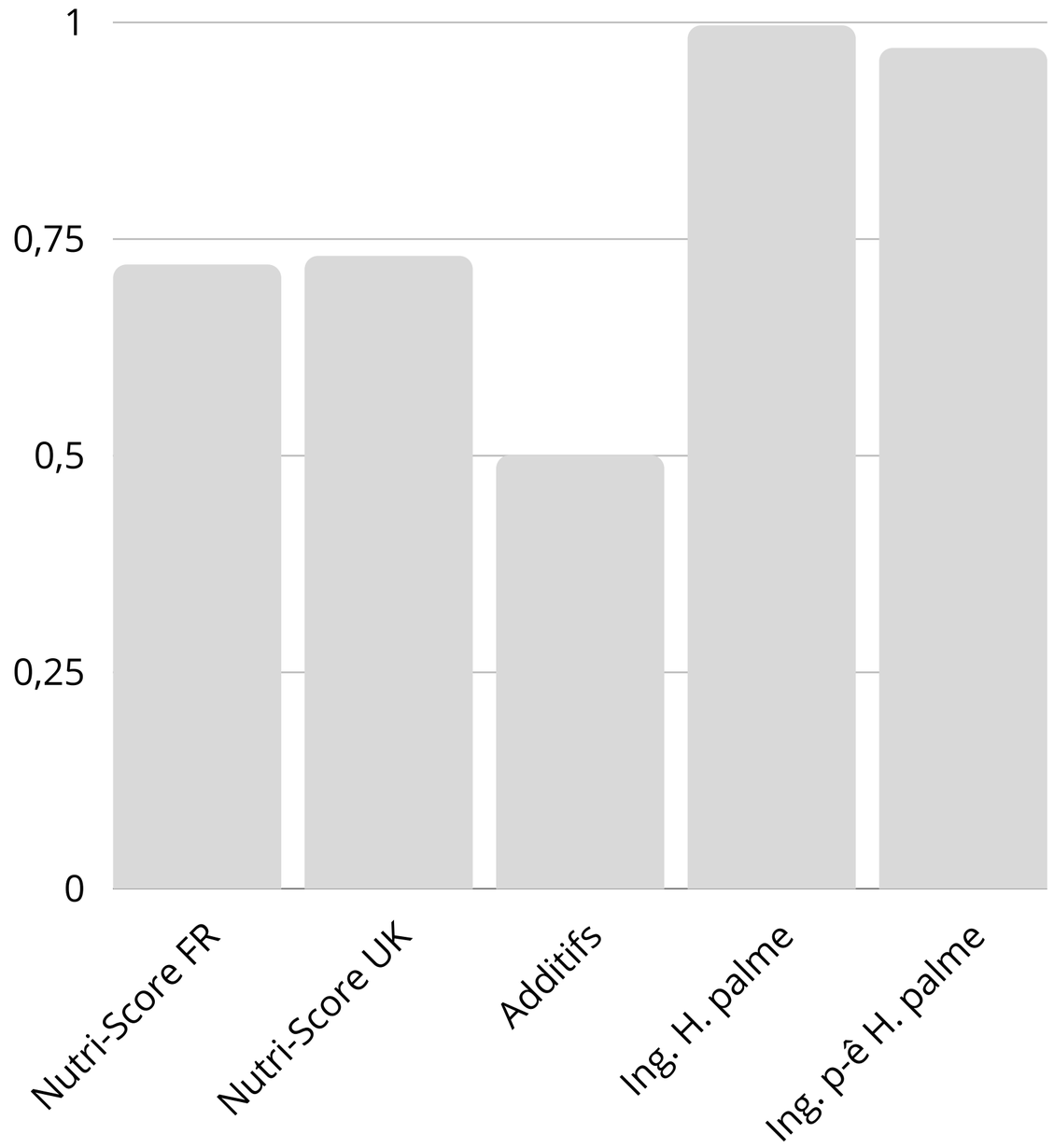
K-NN CLASSIFICATION

Variables discrètes



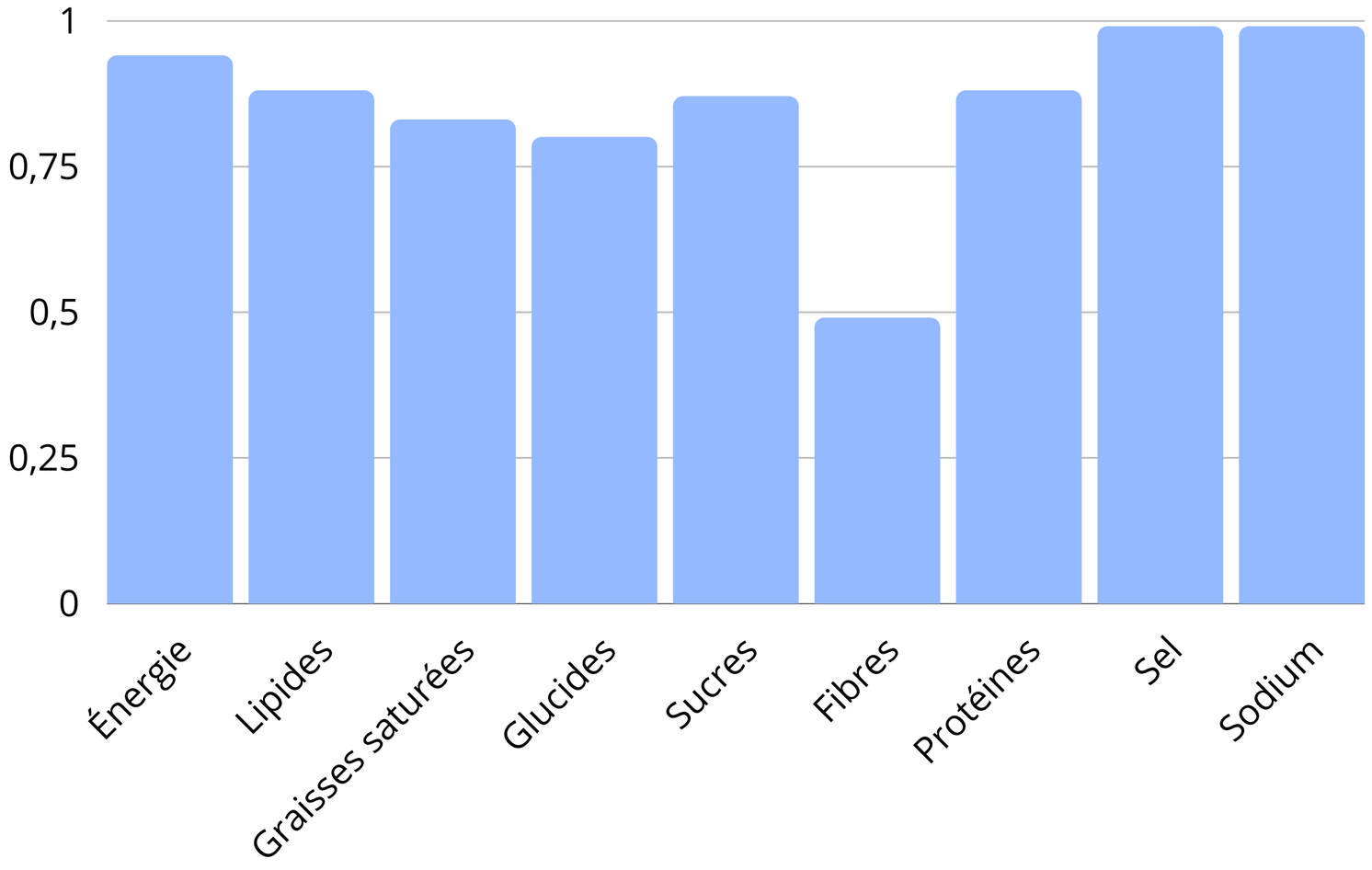
PRÉDICTIONS

Taux de précision



PRÉDICTIONS

Coefficient de détermination R²



04.

CONCLUSION

Faisabilité du projet
Nuances sur l’analyse

5 principes RGPD

- Licéité, Loyauté et Transparence
- Limitation des Finalités
- Minimisation des Données
- Exactitude des Données
- Conservation des Données

- FORTES CORRÉLATIONS
- PROJET COHÉRENT

NUANCES SUR L'ANALYSE

- SEUIL DE 50% DE VALEURS MANQUANTES
- TRAITEMENT DES VALEURS ABERRANTES
- TRAITEMENT DES VALEURS MANQUANTES
- ALGORITHMES DE MACHINE LEARNING

MERCI !

Des questions ?