Santé
Publique
France

OPEN FOOD FACTS

AI ENGINEER _ Projet 3

OpenClassrooms

Préparez des données pour un organisme de santé publique

Introduction

Santé publique France souhaite améliorer sa base de données Open Food Facts.
Cette base de données open source est mise à la disposition de particuliers et d'organisations afin de leur permettre de connaître la qualité nutritionnelle des produits.

Pour ajouter un produit à cette base de données, il est nécessaire de remplir de nombreux champs textuels et numériques, ce qui peut conduire à des erreurs de saisie et à des valeurs manquantes dans la base de données.

L'agence Santé publique France nous confie la création d'un système de suggestion ou d'auto-complétion afin d'aider les usagers à remplir plus efficacement la base de données.

En tant que ingénieur en intelligence artificiel, j'ai été missionné sur le projet de nettoyage et d'exploration des données, afin de déterminer la faisabilité de cette nouvelle application.

Sommaire

Nettoyage des données

. Analyse exploratoire des données
. Réduction du dataset et sélection de la cible
. Gestion des valeurs:
    Traitement des Outliers
    Imputation avec KNN

Partie Analyse

. Analyses univariées
. Analyses bivariées
. Analyses multivariées
. Test statistiques
. Analyses des composantes principales
. Cerles de corrélation
. Conclusion

Première partie

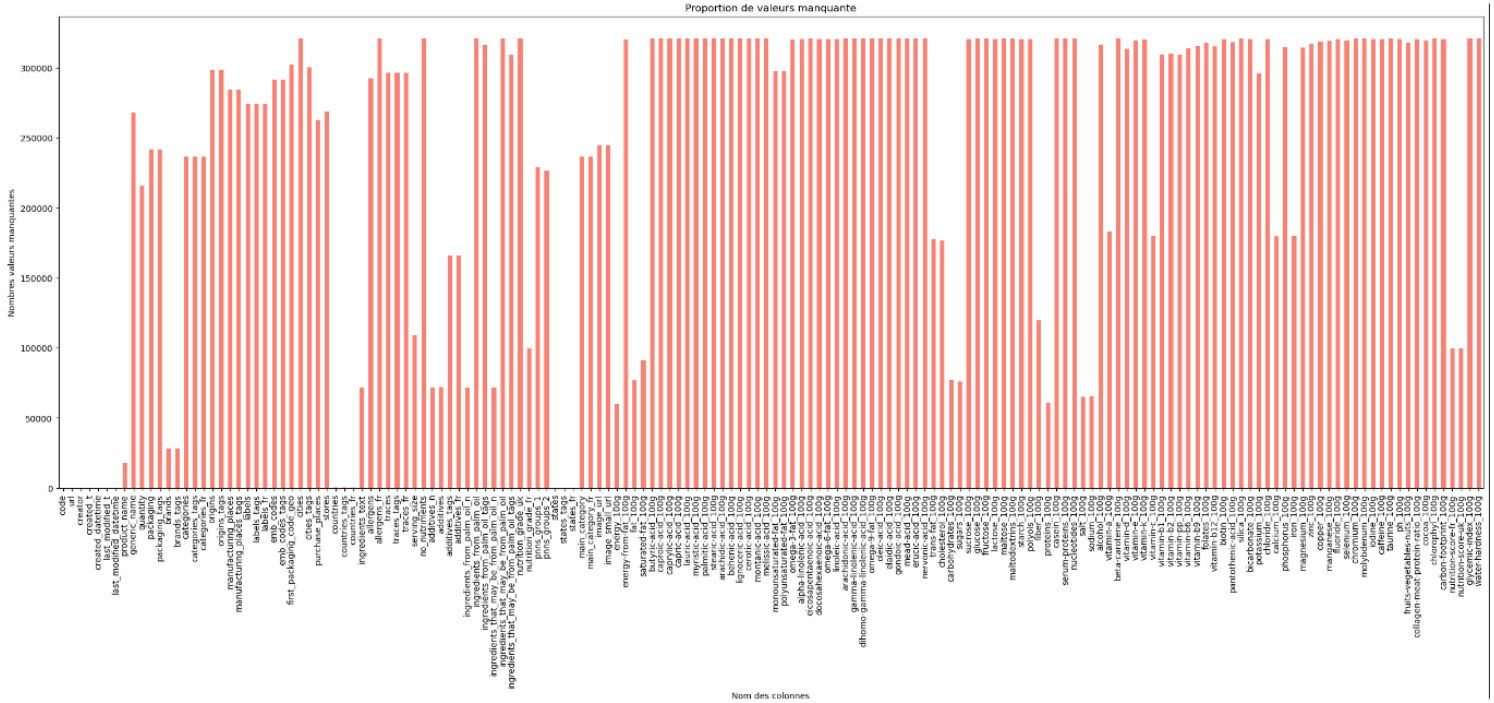
ANALYSE EXPLORATOIRE

```
# Dimensions du dataset
nb_lignes, nb_colonnes = data.shape
print(f"Nombre de lignes : {nb_lignes}")
print(f"Nombre de colonnes : {nb_colonnes}")
```

```
Nombre de lignes : 320772
Nombre de colonnes : 162
```

```
# Aperçu général du contenu
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 320772 entries, 0 to 320771
Columns: 162 entries, code to water-hardness_100g
dtypes: float64(106), object(56)
memory usage: 396.5+ MB
```




```
data.head()
```

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	quantity	...	ph_100g	fruits-vegetables-nuts_100g	collagen-meat-protein-ratio_100g	cocoa_100g	chlorophyl_100g	carbon-footprint
0	0000000003087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	NaN	1kg	...	NaN	NaN	NaN	NaN	NaN	NaN
1	0000000004530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
2	0000000004559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
3	0000000016087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salted Nut Mix	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN
4	0000000016094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN

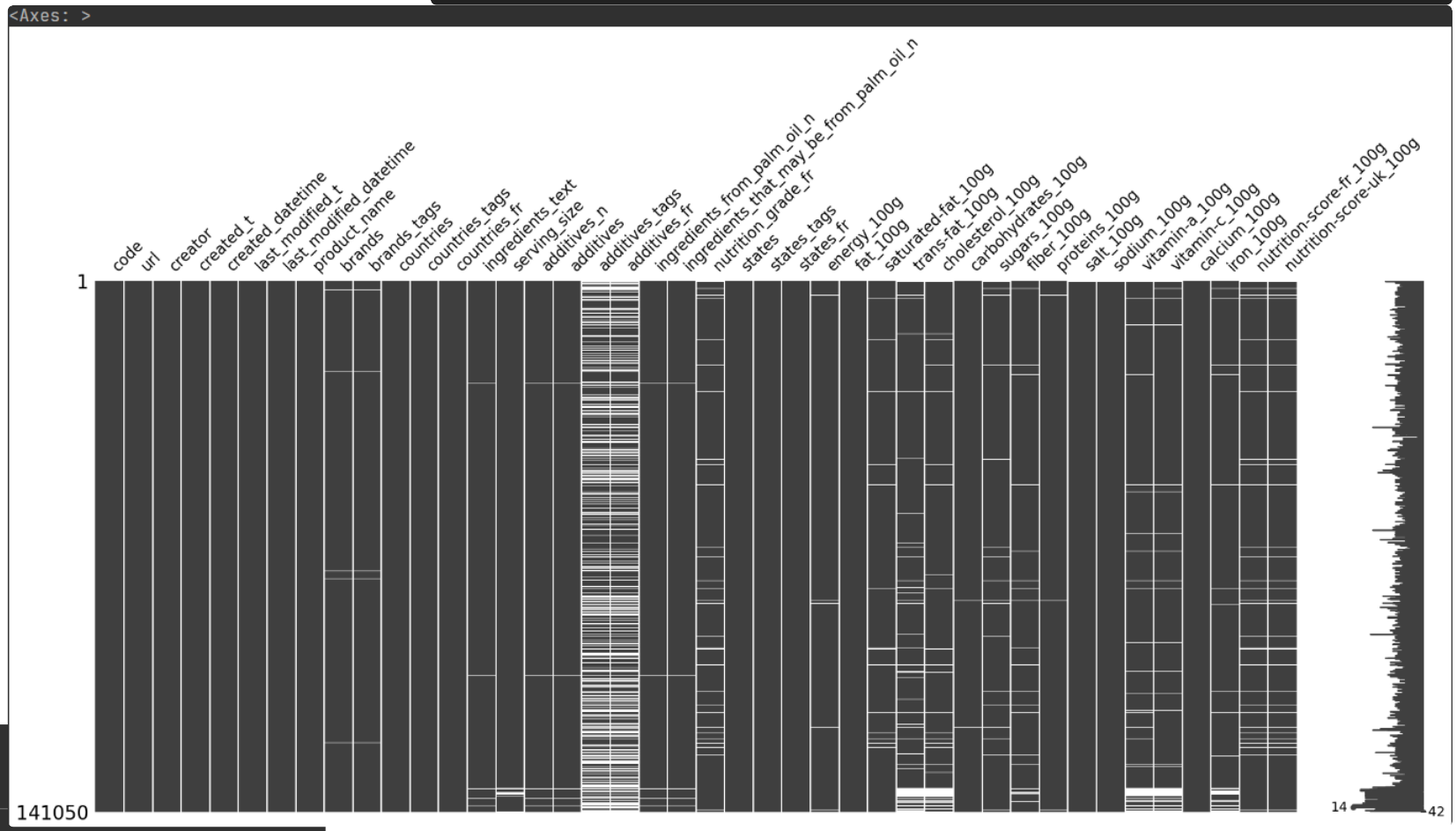
5 rows × 162 columns

REDUCTION DATA SET ET SELECTION DE LA CIBLE

Choix de la cible:

(quant) calcium_100g - 179722 valeurs manquantes, cela correspond à (56.03%) du dataset

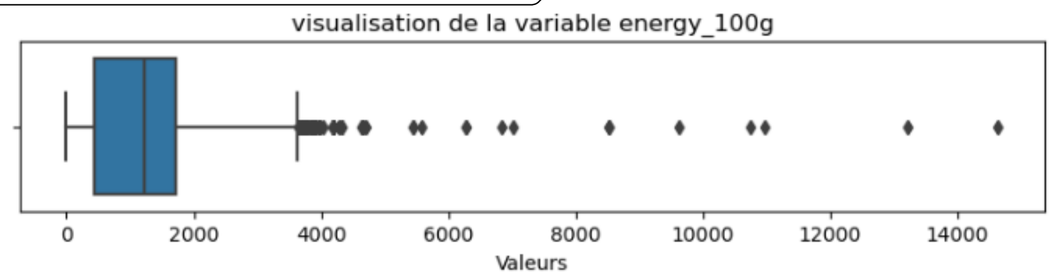
Apport du calcium	Impact sur la qualité
Information souvent manquante	Enrichit la base de données
Non pris en compte dans le Nutri-Score	Permet une évaluation complémentaire
Essentiel à la santé	Aide à mieux informer les consommateurs
Peut différencier des produits similaires	Encourage des choix plus équilibrés

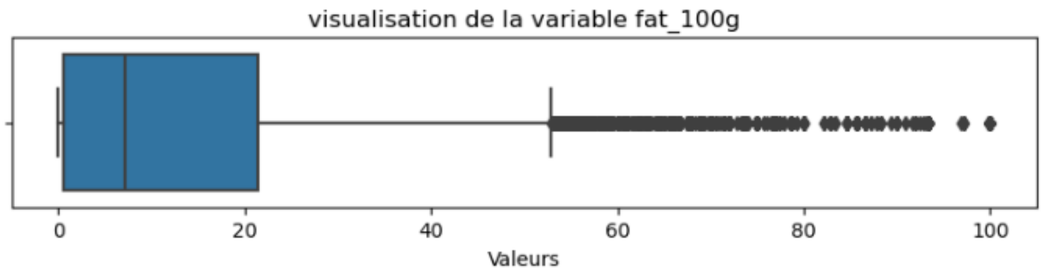


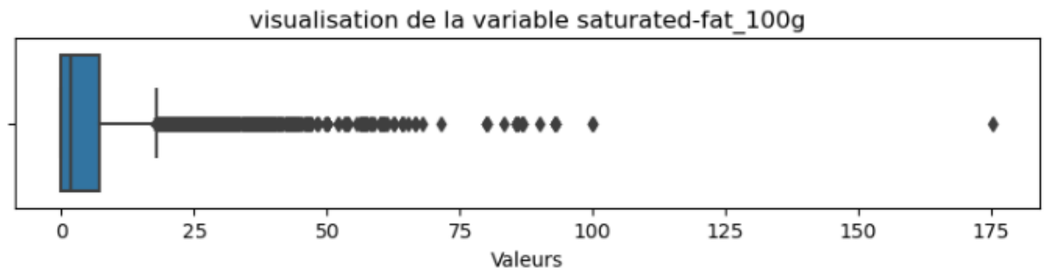

```
# Suppression des colonnes
data = data.drop(columns=colonne_supp_60)
```

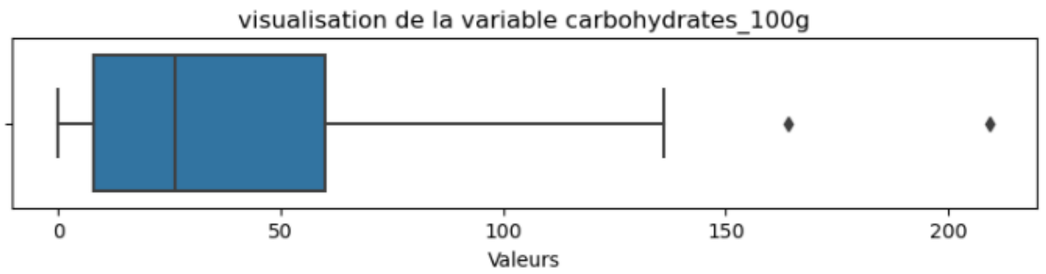
Les colonnes avec > 60 % de valeurs manquantes contiennent trop peu d'informations utiles.

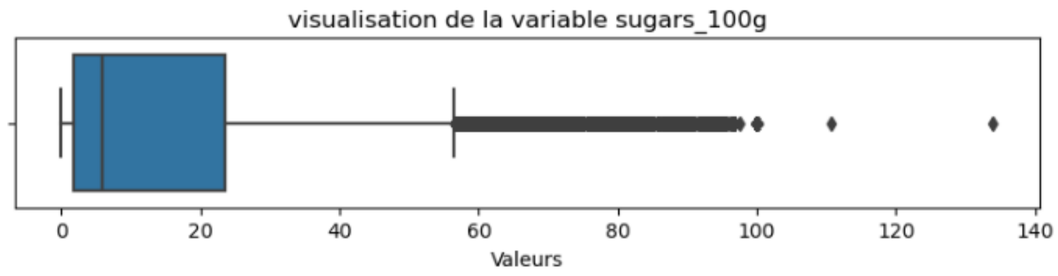
Gestion des valeurs: outliers (avant)

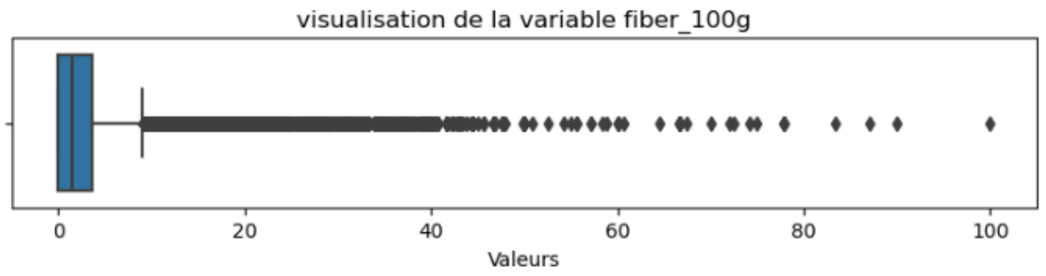


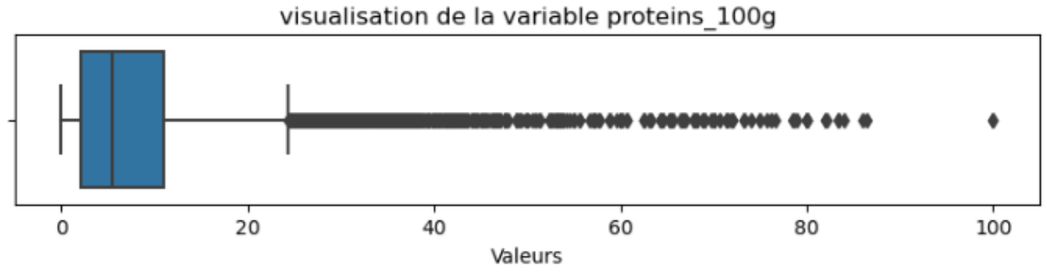


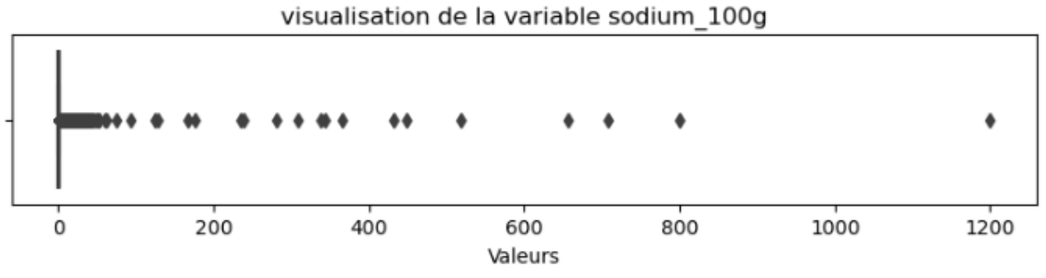


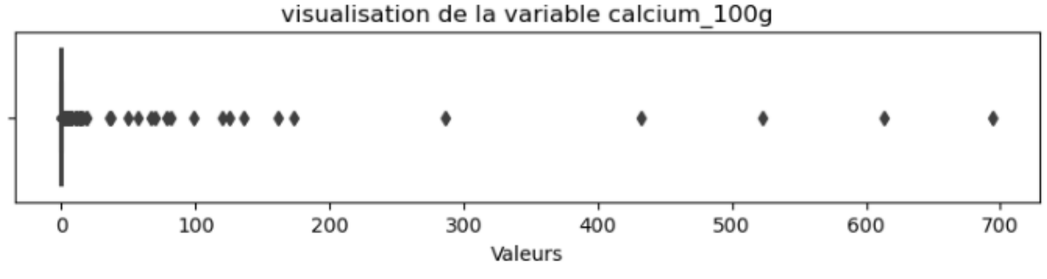


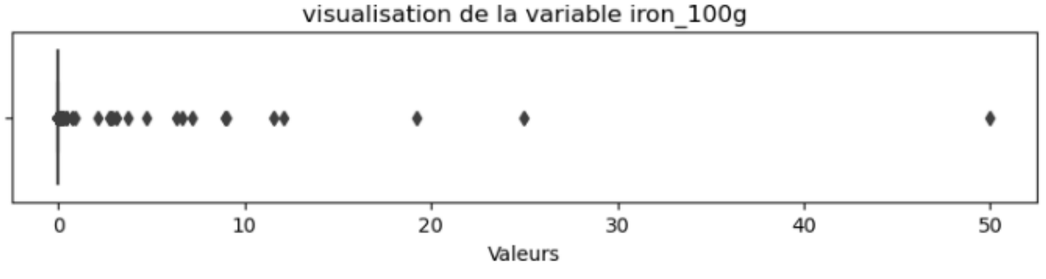




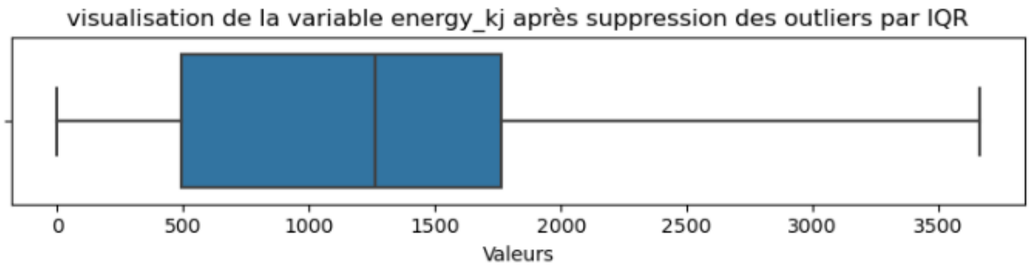


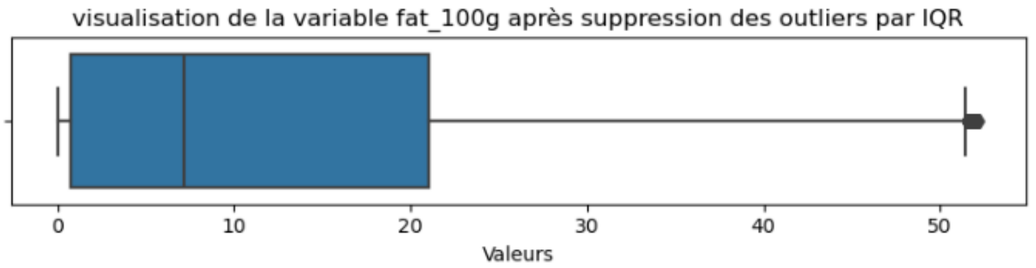


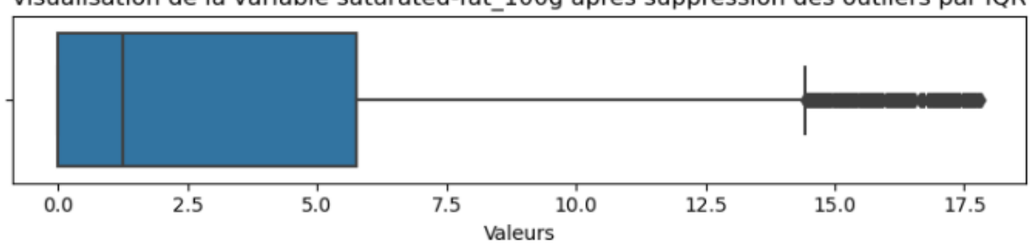


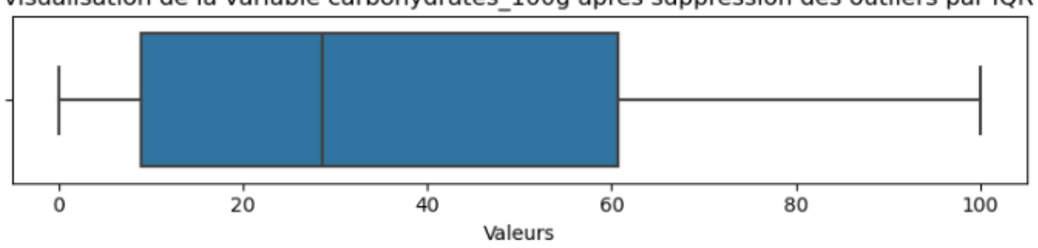


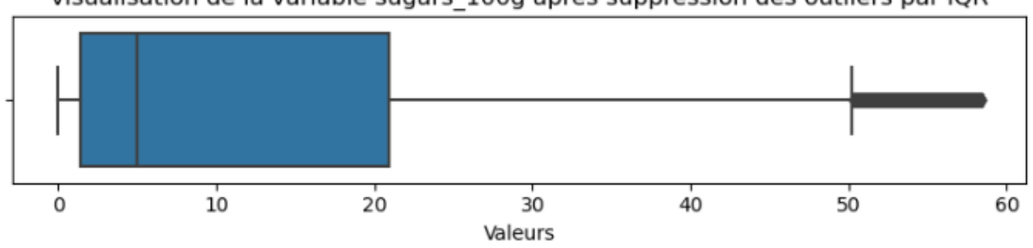
Gestion des valeurs: outliers (apres)

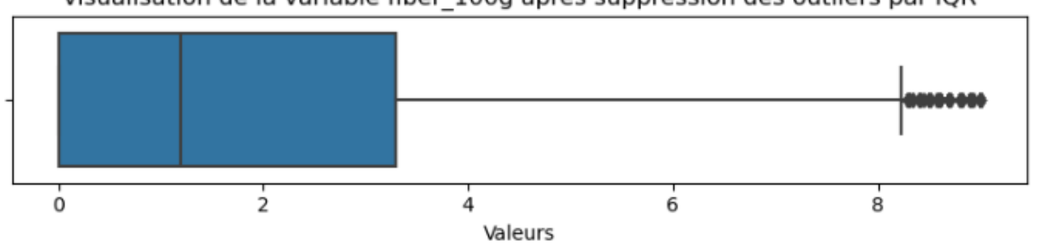


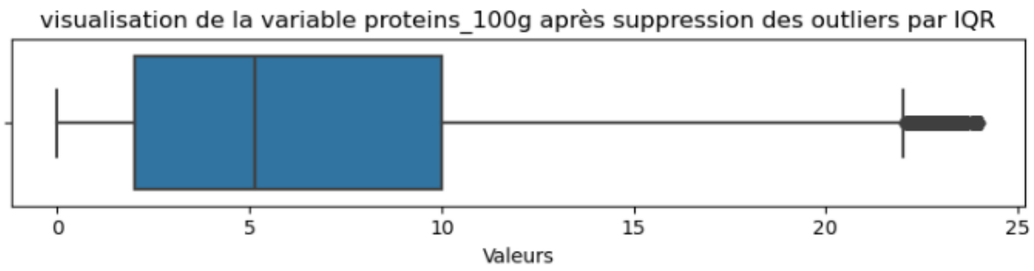


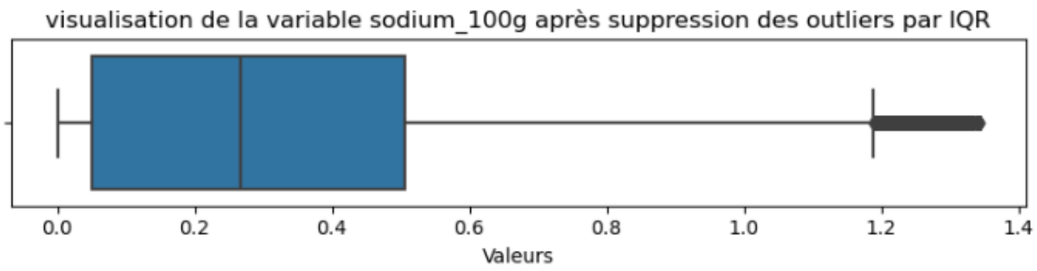


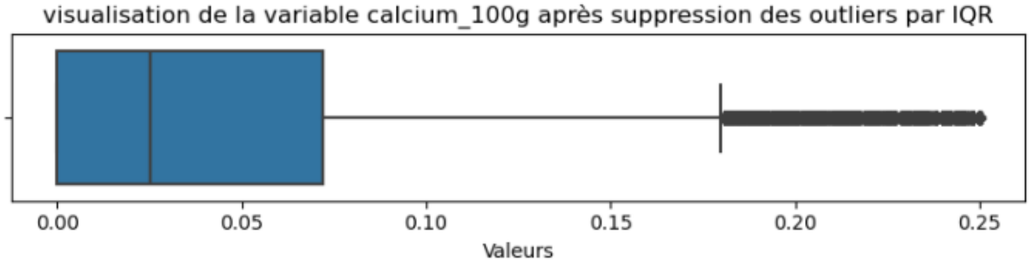


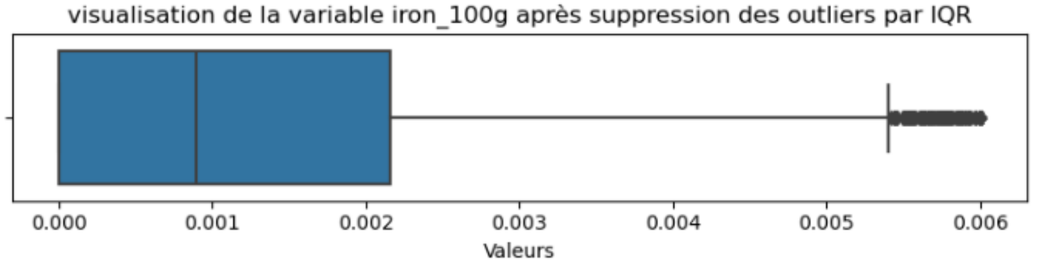








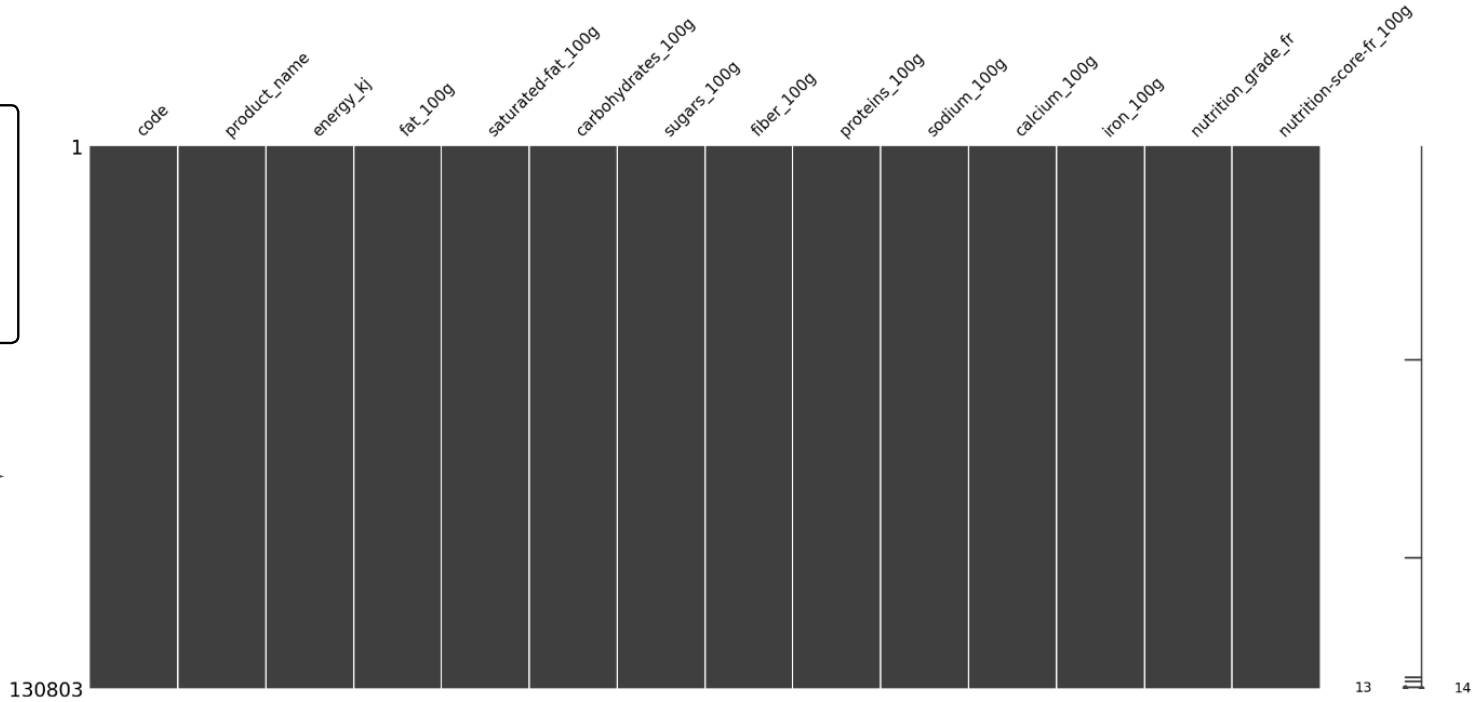




```
Taux de remplissage des variables numériques (en %) :
carbohydrates_100g         100.00
energy_kj                  100.00
fat_100g                    97.75
iron_100g                   95.73
sugars_100g                 95.33
sodium_100g                 94.88
proteins_100g               94.32
saturated-fat_100g          93.36
fiber_100g                  92.36
calcium_100g                91.68
nutrition-score-fr_100g     83.78
dtype: float64
```

Gestion des valeurs: imputation KNN

Imputation avec KNN

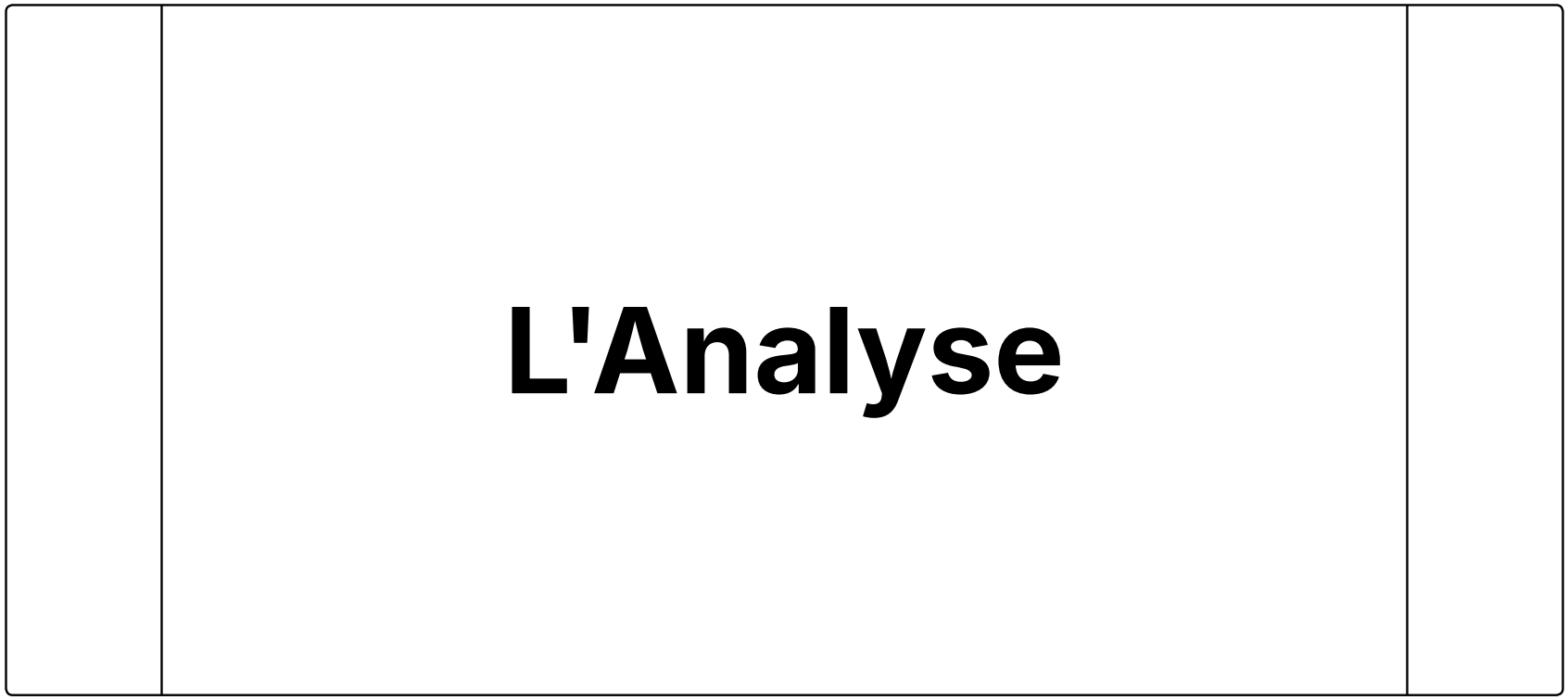


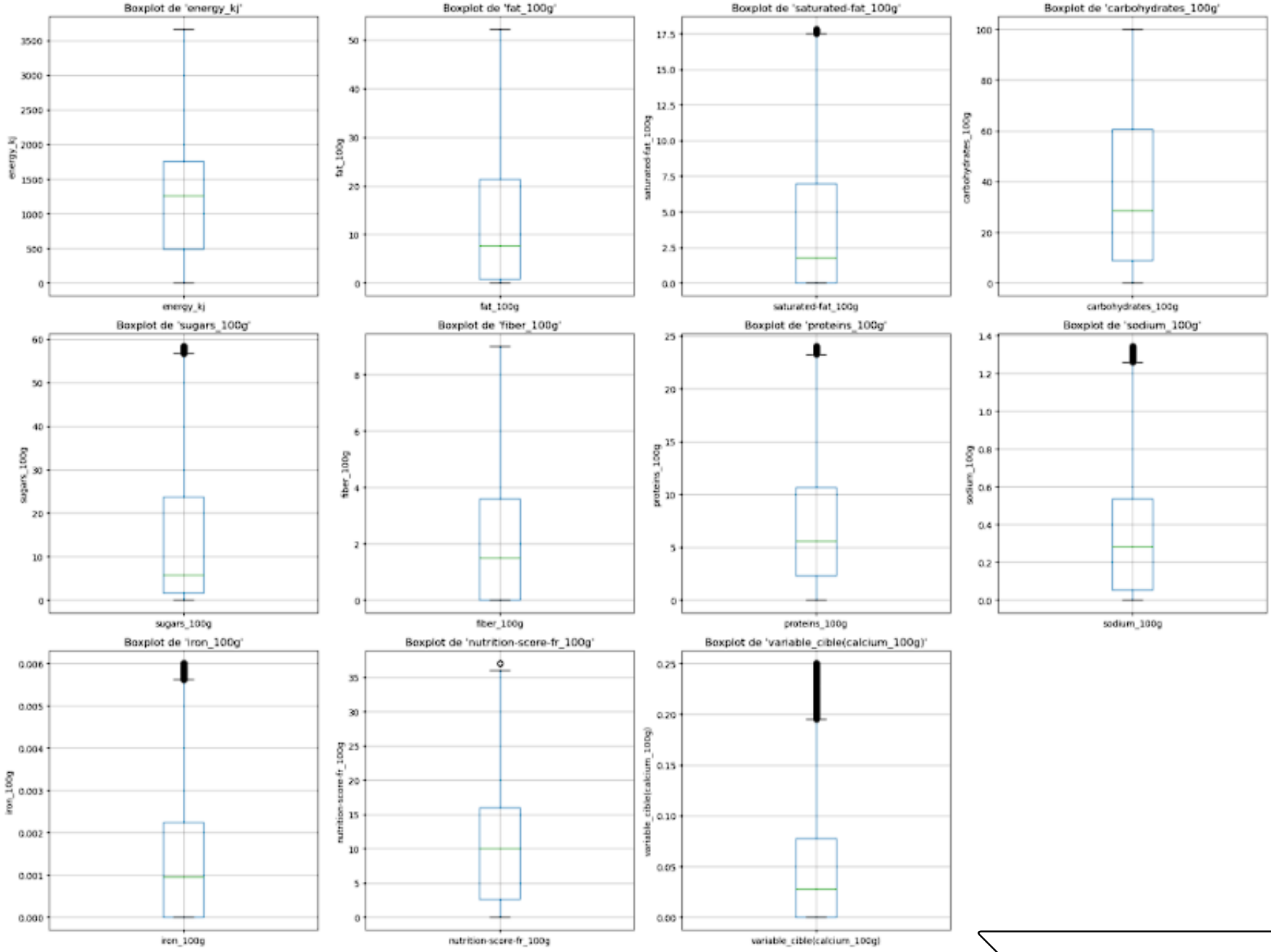
Data set final:

tailles du data set:	130 803 lignes, 14 colonnes
valeurs manquantes:	0 (numérique et cible)
outliers:	supprimés via IQR
doublons:	supprimés
variables inutiles:	supprimées (>60% manquants)
variables finales:	Conformes au nutri-scores

Mon objectif était de nettoyer les données tout en maximisant la conservation d'information.
J'ai combiné des méthodes statistiques (IQR, KNN) et supervisées (regression logistique) pour traiter les données manquantes et aberrantes.
Le jeu de données est maintenant propre, complet et prêt à être exploité dans une analyse exploratoire ou un modèle predictif.

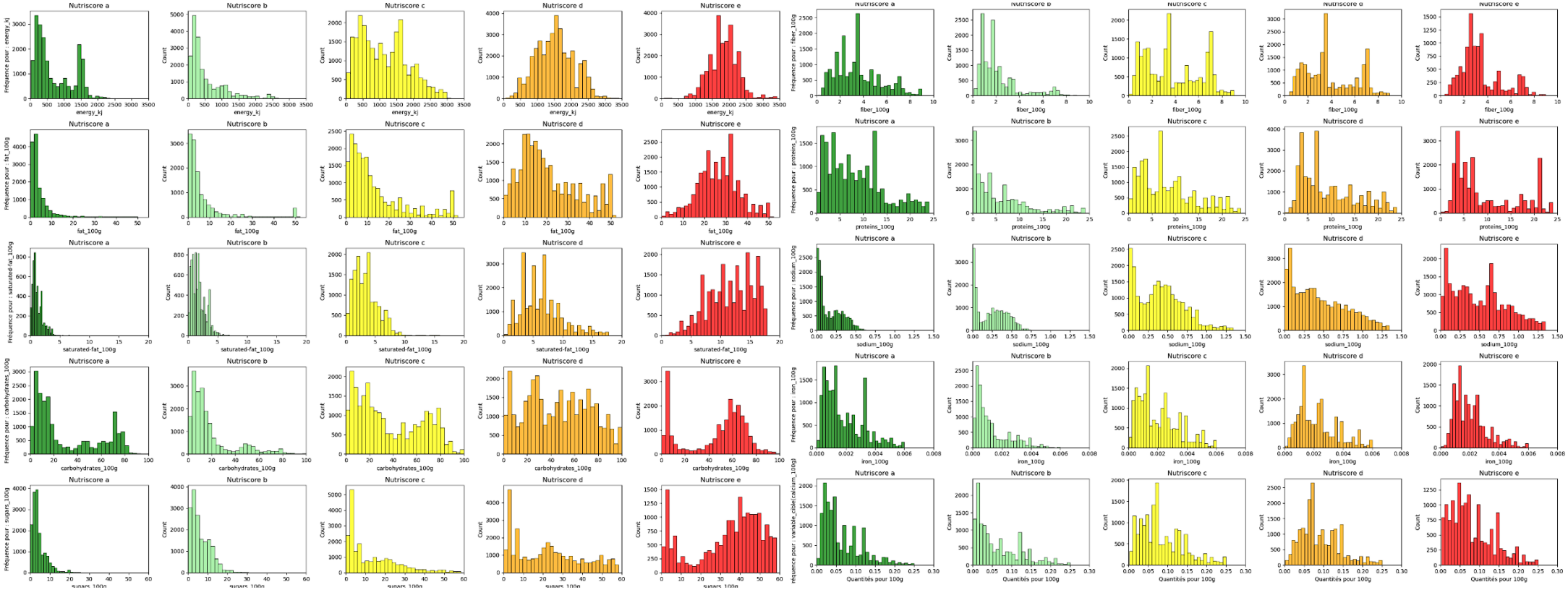
Seconde partie



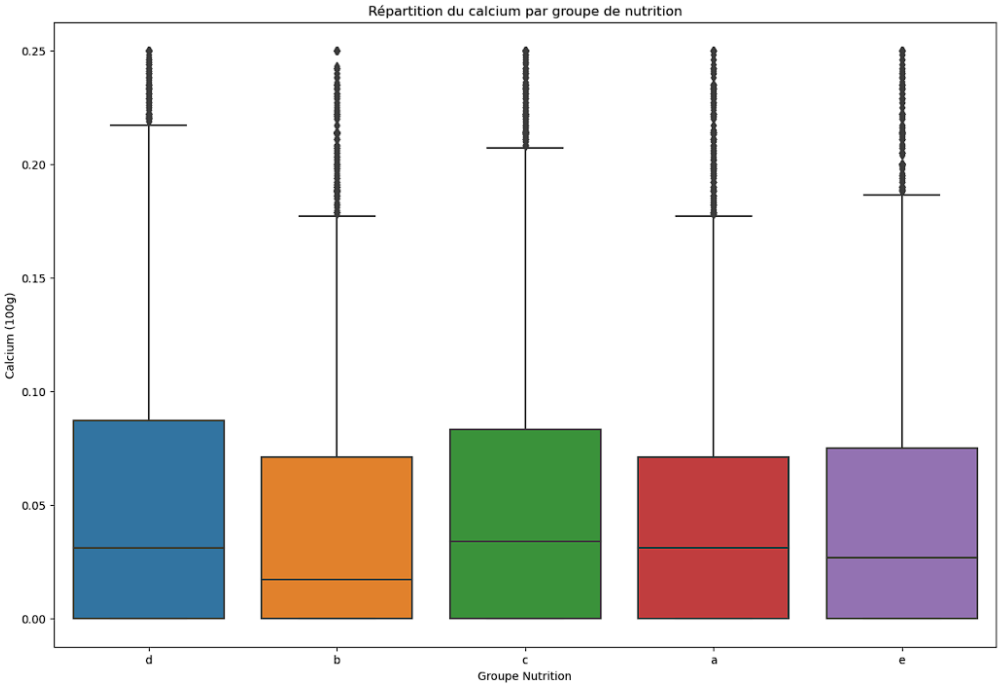


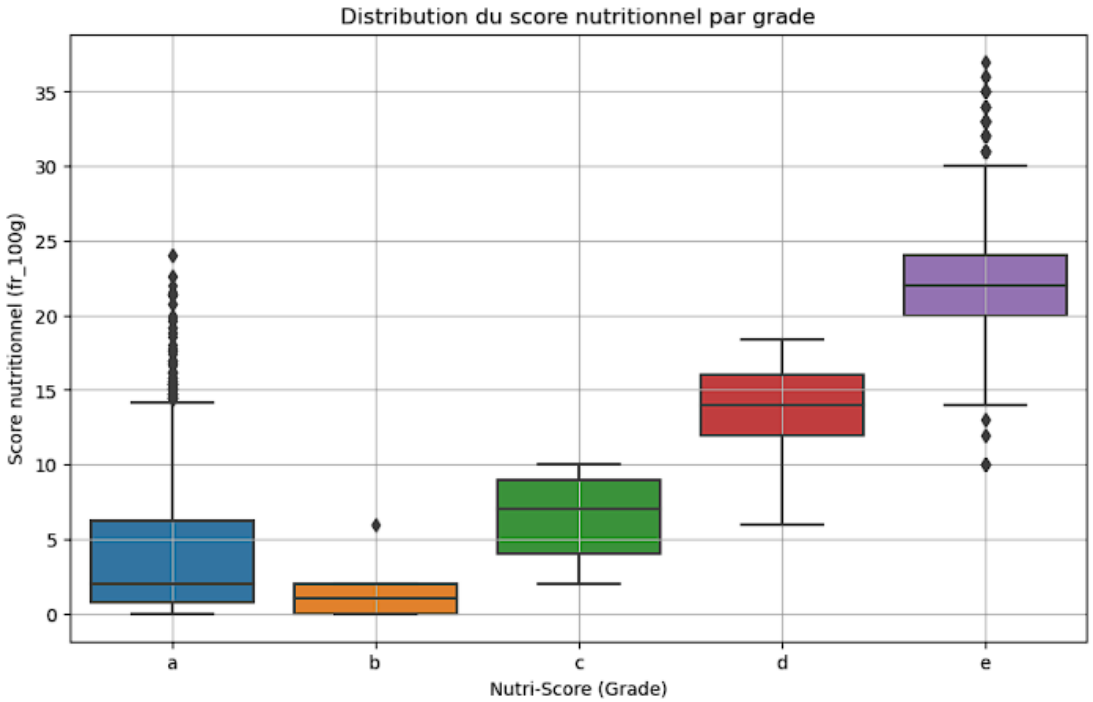
Analyse univariées

Distribution des nutriments par nutriscore

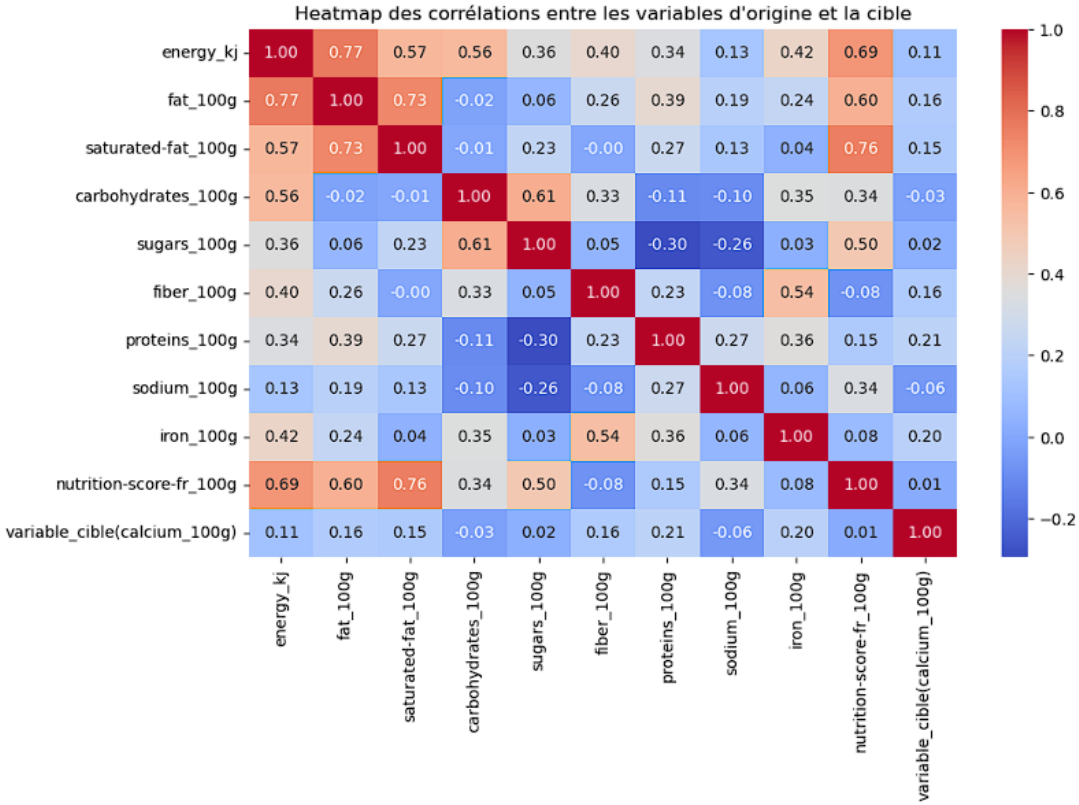


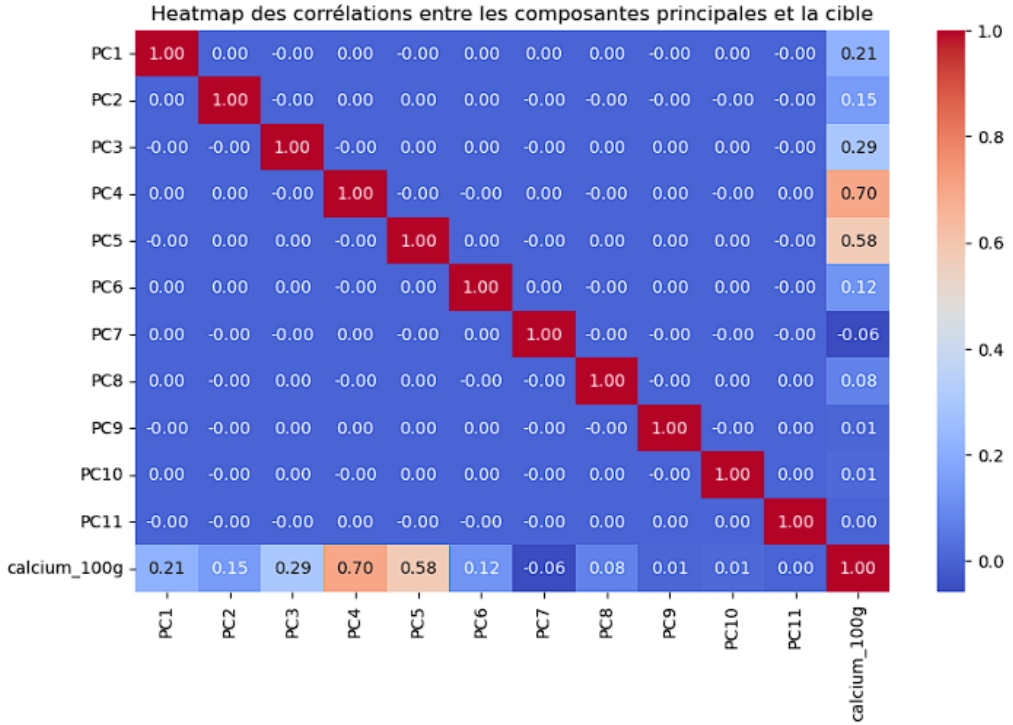
Analyse bivariée





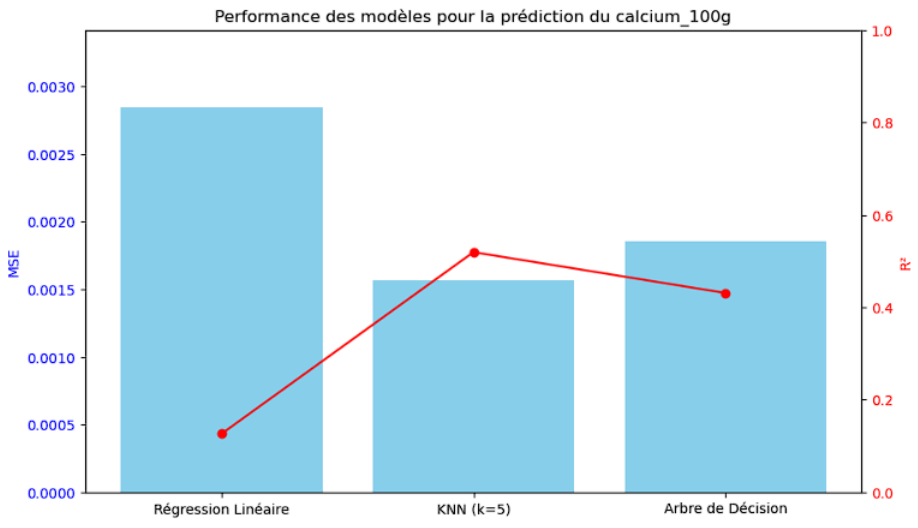
Analyse bivariées

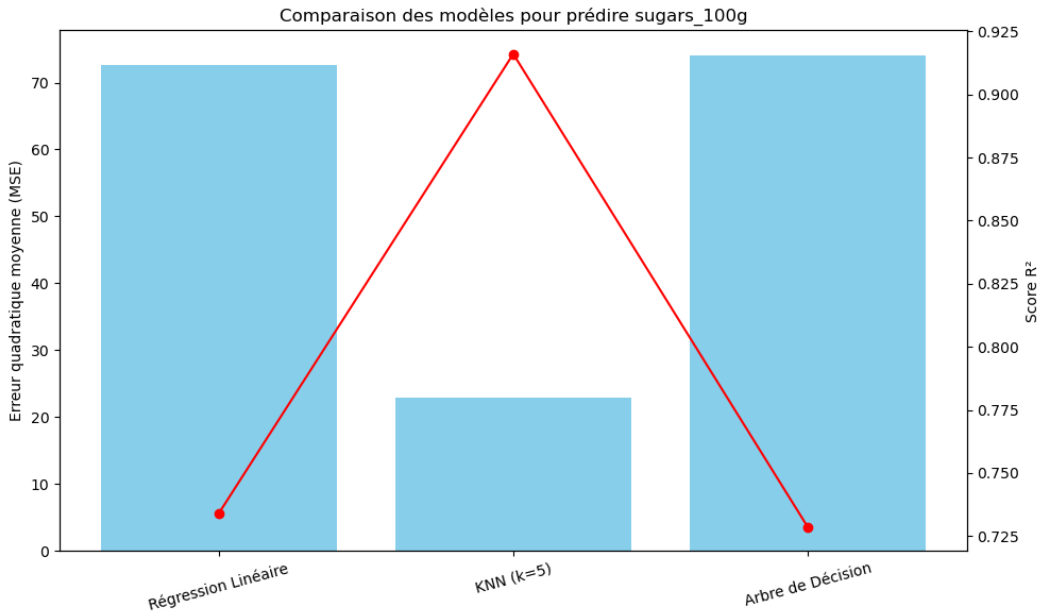


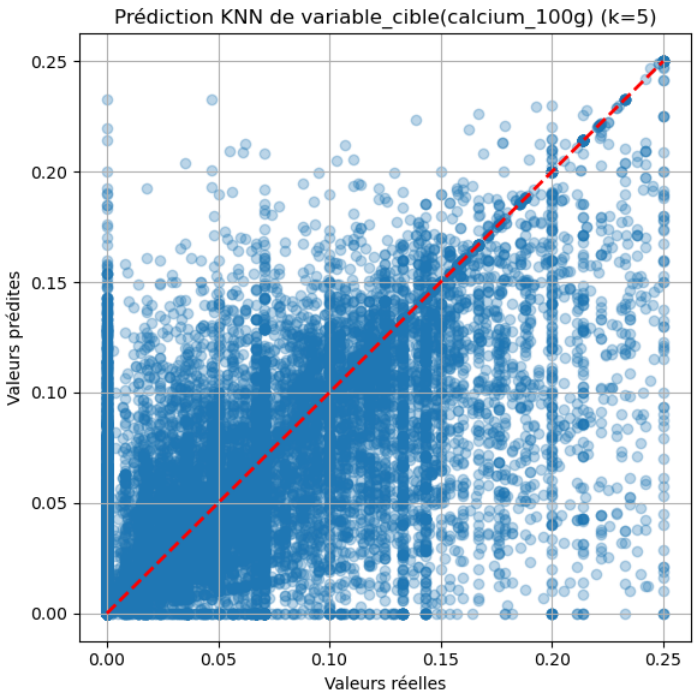


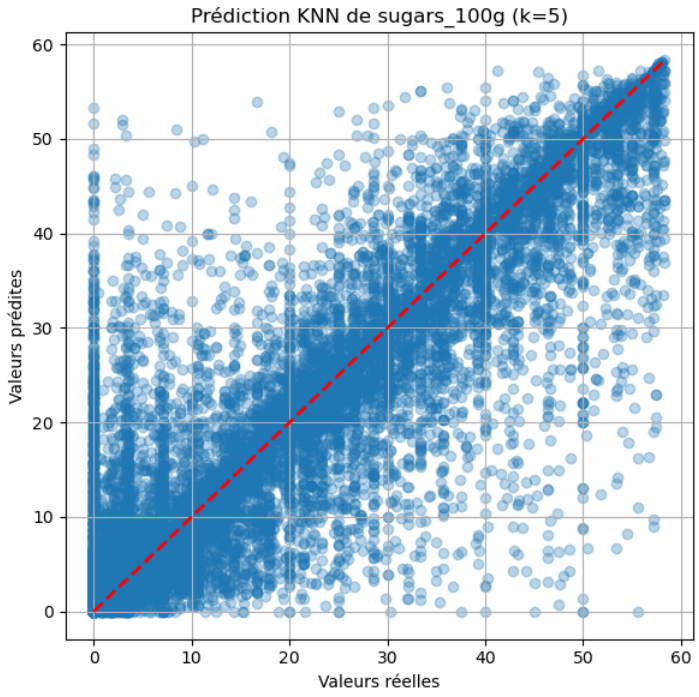
Analyses multivariées

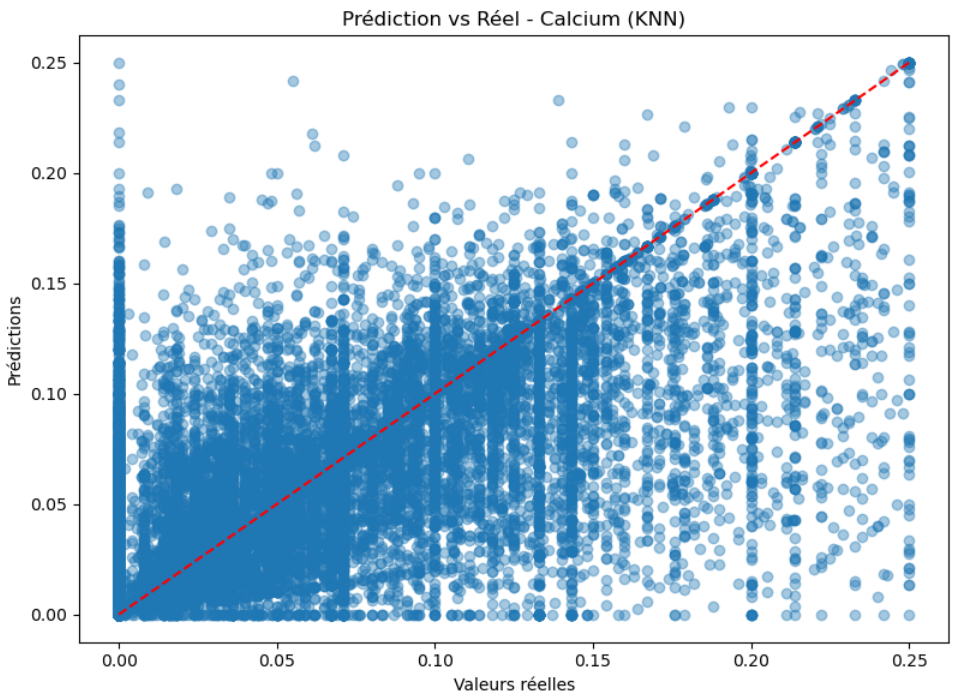
```
Régression Linéaire
MSE : 0.0028
R²  : 0.1263

KNN (k=5)
MSE : 0.0016
R²  : 0.5193

Arbre de Décision
MSE : 0.0019
R²  : 0.4309
```

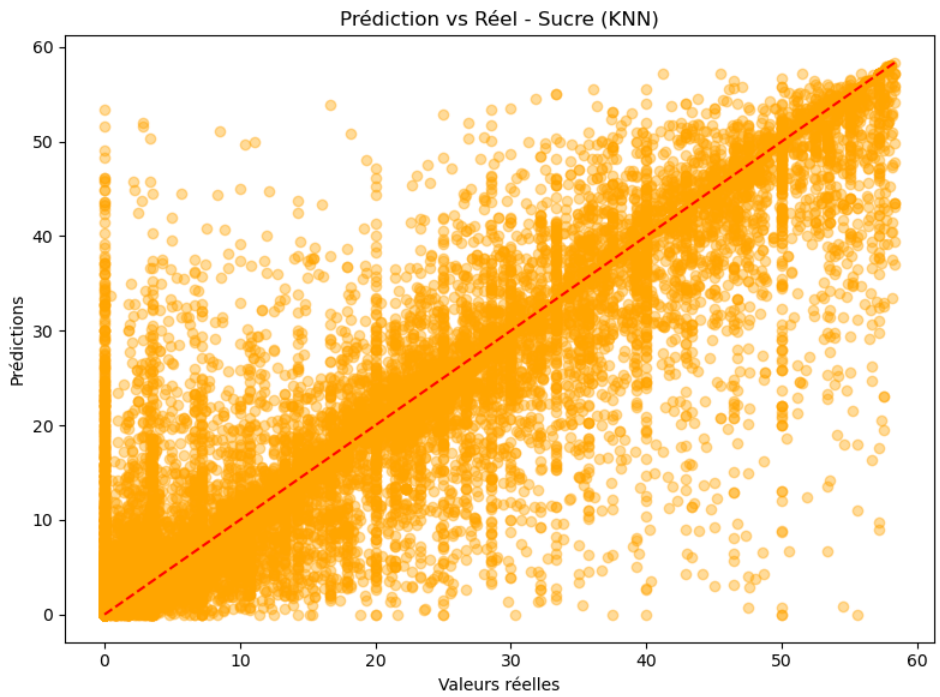




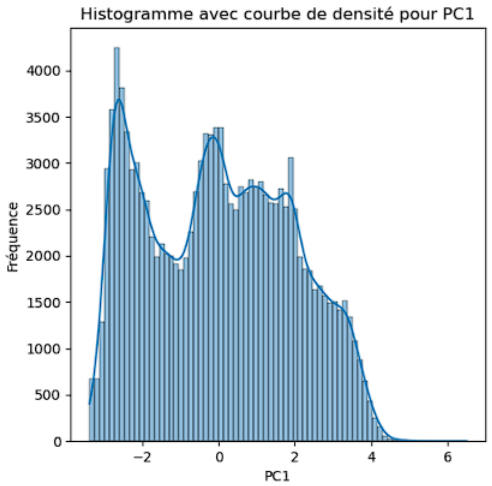


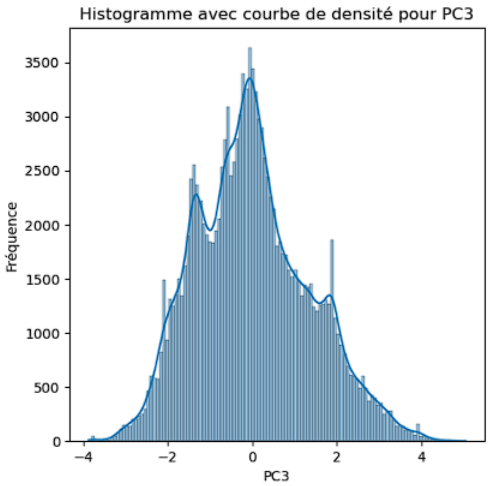


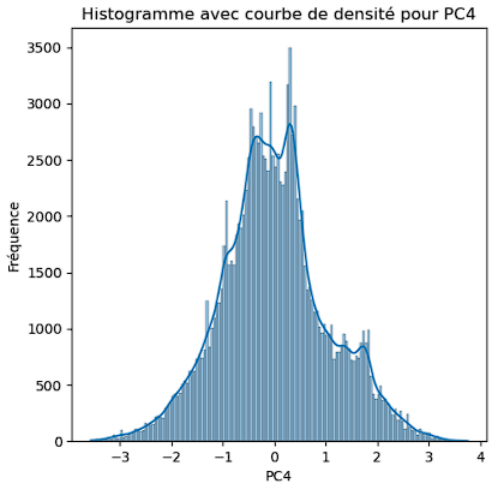


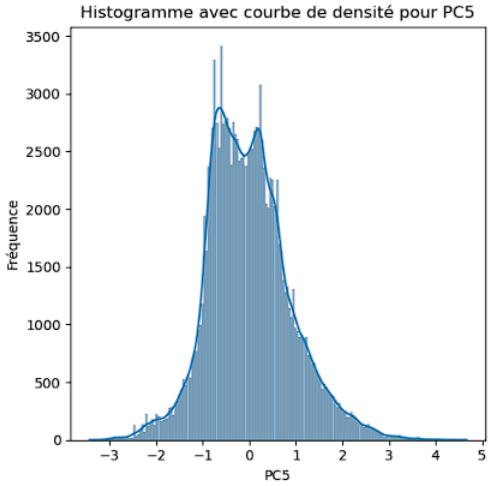


extra pour axe d'amelioration





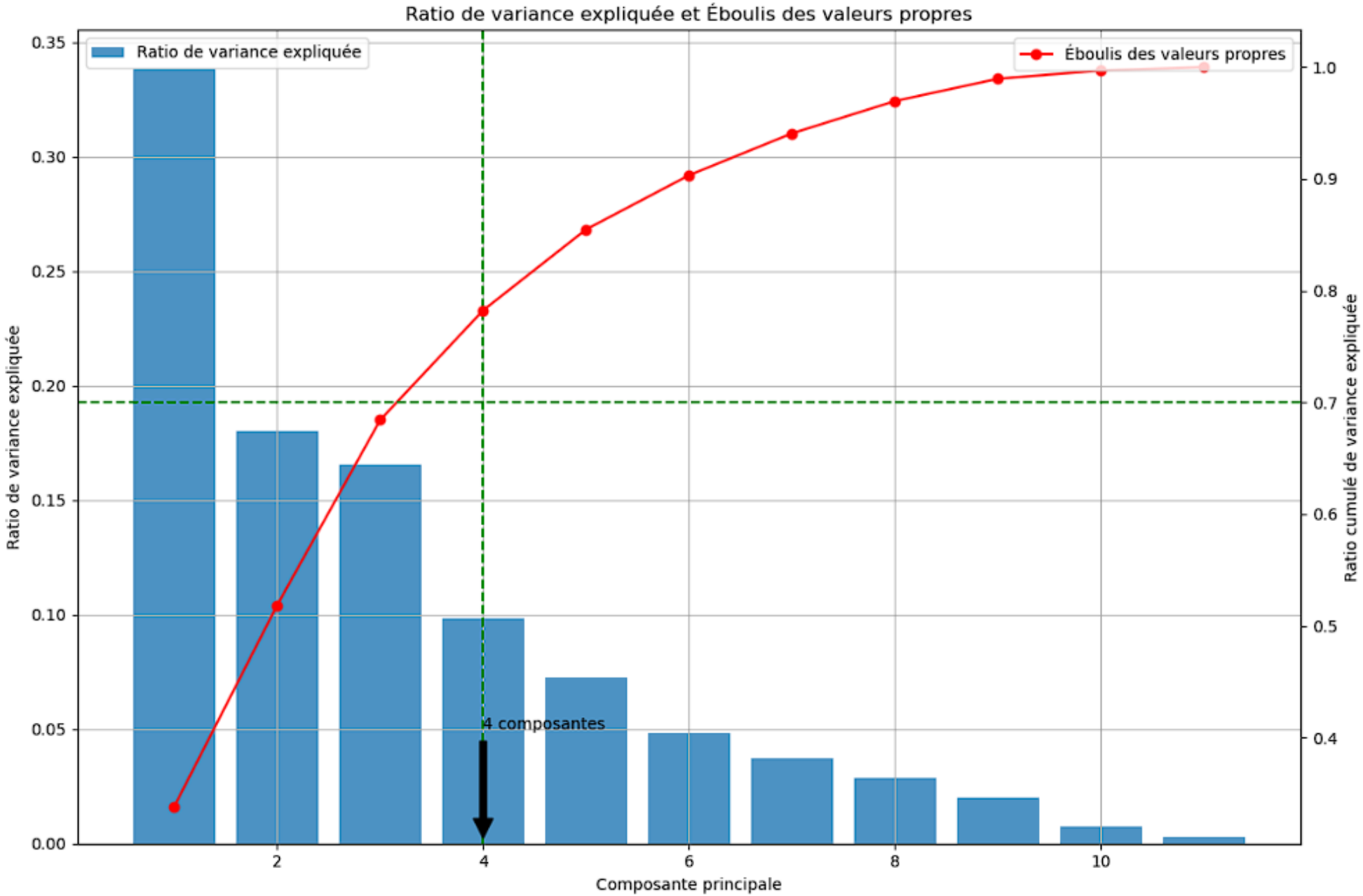




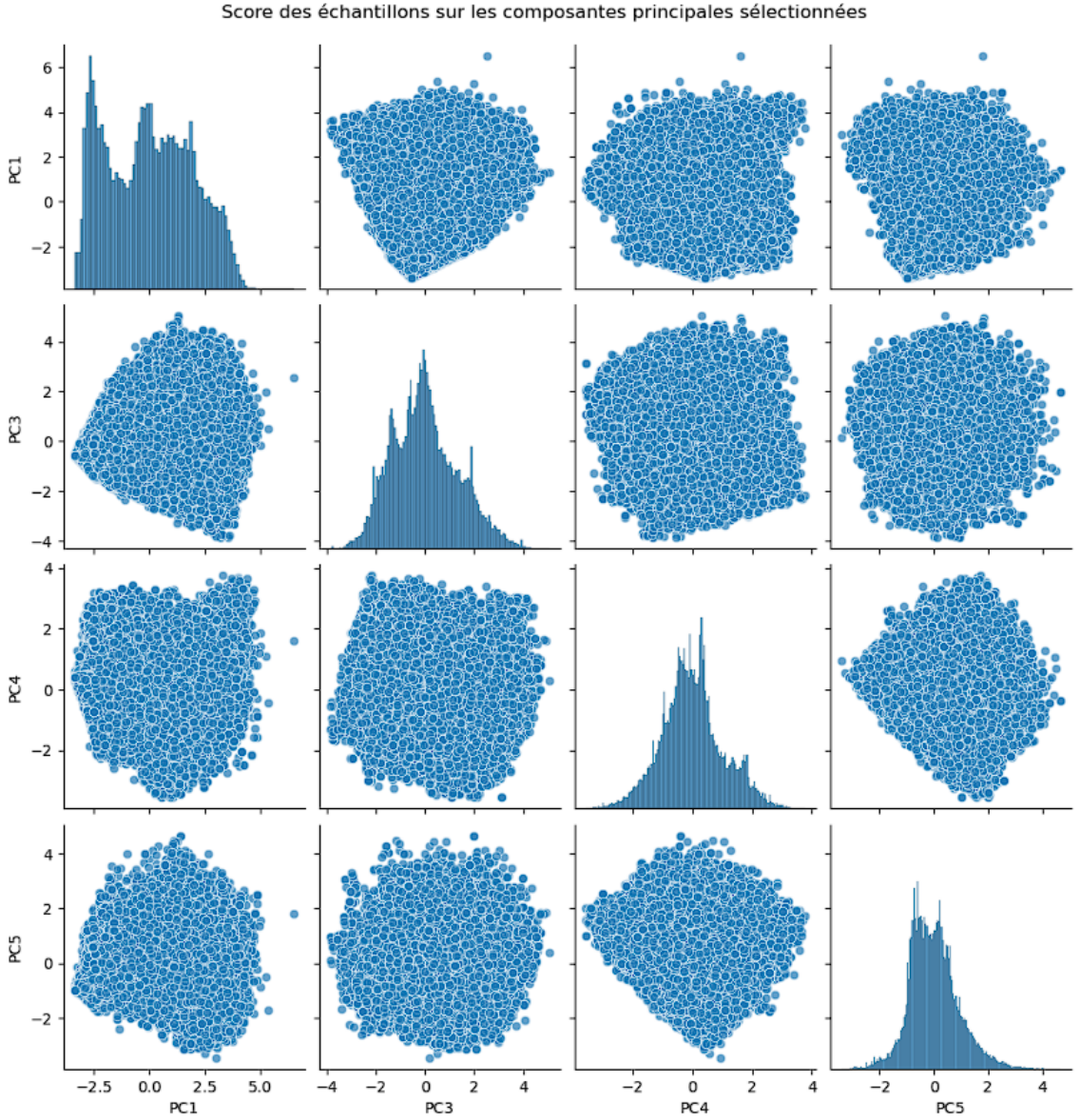
PC1 - Kolgomorov-Smirnov :
Statistique = 0.2165017522364619,
p-value = 0.0
PC1 n'est pas normalement distribué
PC3 - Kolgomorov-Smirnov :
Statistique = 0.09270263482065691,
p-value = 0.0
PC3 n'est pas normalement distribué
PC4 - Kolgomorov-Smirnov :
Statistique = 0.04044300673126888,
p-value = 2.4587742746593026e-186
PC4 n'est pas normalement distribué
PC5 - Kolgomorov-Smirnov :
Statistique = 0.06890482518182,
p-value = 0.0
PC5 n'est pas normalement distribué

Corrélation de Spearman pour PC1 et calcium 100g : 0.1715222410156125,
p-value :0.0
La corrélation entre PC1 et la cible est significative
Corrélation de Spearman pour PC3 et calcium 100g : 0.2885338063343441,
p-value :0.0
La corrélation entre PC3 et la cible est significative
Corrélation de Spearman pour PC4 et calcium 100g : 0.671148661601234,
p-value :0.0
La corrélation entre PC4 et la cible est significative
Corrélation de Spearman pour PC5 et calcium 100g : 0.4991215426662886,
p-value :0.0
La corrélation entre PC5 et la cible est significative

Tests statistiques



Analyses des composantes principales

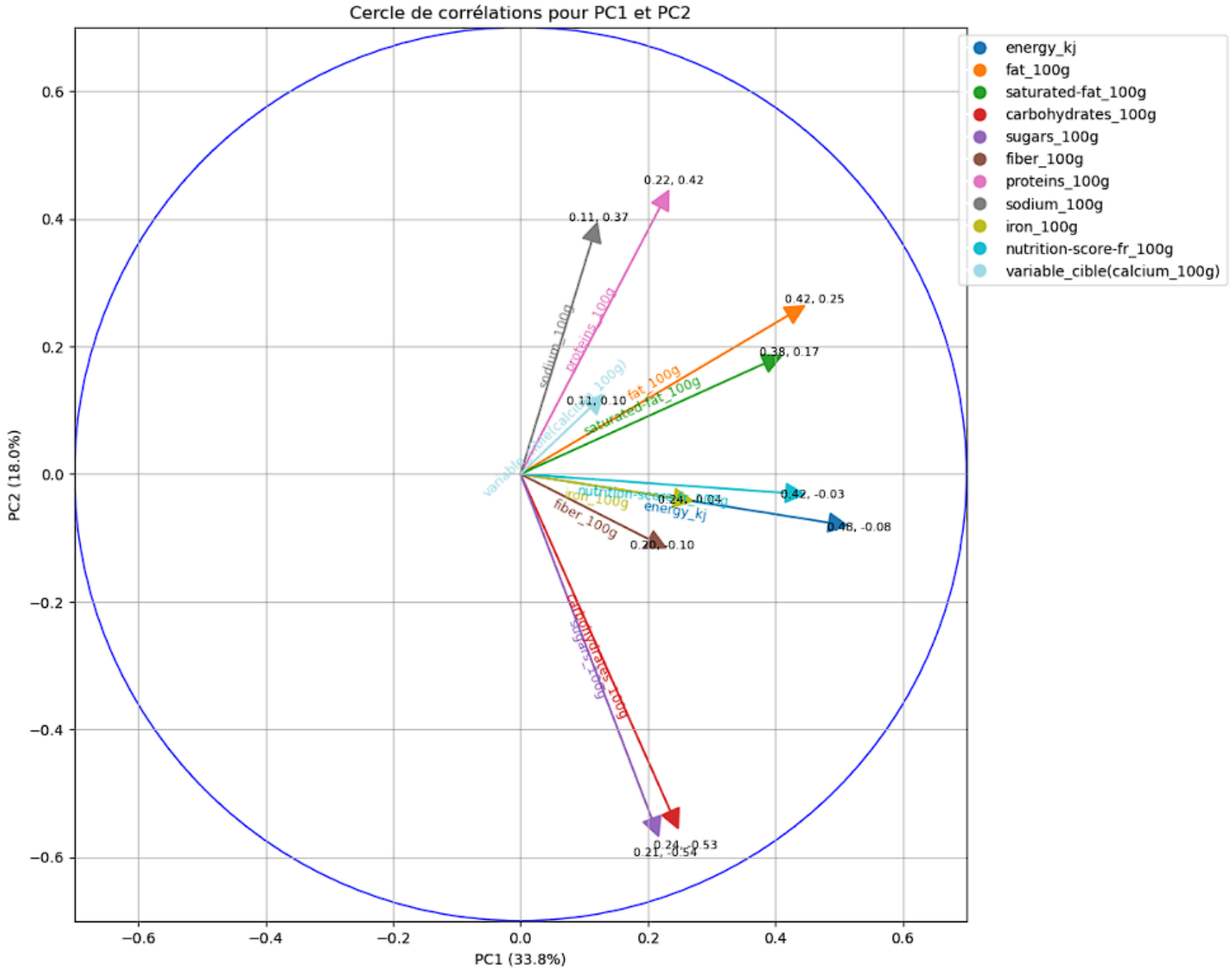

Score des échantillons sur les composantes principales sélectionnées
PC1
PC3
PC4
PC5

PC1 (33.8% de variance expliquée) semble être un axe d'opposition entre densité nutritionnelle (fat, saturated fat, energy) et éléments bénéfiques (fiber).

PC2 (18.0%) capte des différences liées à sugars_100g, iron_100g, calcium avec des effets modérés.

Les variables qui doivent être rempli pour la futur applications:

. **Energy_Kj**
. **Carbohydrates**
. **Sugar**
. **Saturated-fat**
. **Fat**
. **Proteins**
. **Sodium**



Cercle de corrélation

Résumé du projet:

. Améliorer la base de données d'OFF pour santé publique France
. Mise en place d'un système de suggestion et d'auto-complétion pour reduire les erreurs de saisie

Démarche:

Impact:

. Réduction des champs à saisir ce qui réduie les possibles erreurs de saisie et une amélioration de la qualité des données
. Meilleure compréhension des facteurs influençant la qualité nutritionnelle des produits

suggestion:

. Mettre en place dans l'application une validation automatique des valeurs saisies, avec des seuils maximaux réalistes, pour garantir la cohérence nutritionnelle des produits et éviter les erreurs de saisie.
pendant le nettoyage, j'ai vu des valeurs absurdes (ex : 1000 g de sucre).
Imposer des bornes logiques ou nutritionnelles réalistes (ex : max 100g pour les nutriments pour 100g) permettrait de prévenir les erreurs dès la saisie.

. Sensibiliser les utilisateurs sur l'importance de la qualité des données
la qualité de l'IA dépend de la qualité des données d'entrée.
La base Open Food Facts est participative, donc former ou informer les contributeurs est essentiel pour la fiabilité globale.

. Ajouter de nouvelles variables commes les vitamines, minéraux ou la liste d'ingrédients.
Cela permettrait d'enrichir le modèle et de mieux prédire certains nutriments difficiles à estimer à partir des seuls macronutriments, comme le calcium

Conclusion