

# PROJET 3

Préparez des données pour un organisme de santé publique



**Ethan VUILLEMIN**

**OPENCLASSROOMS**

2

# CONTEXTE DU PROJET

Nettoyer et explorer les données pour améliorer la base de données d'Open Food Facts?

+  
•

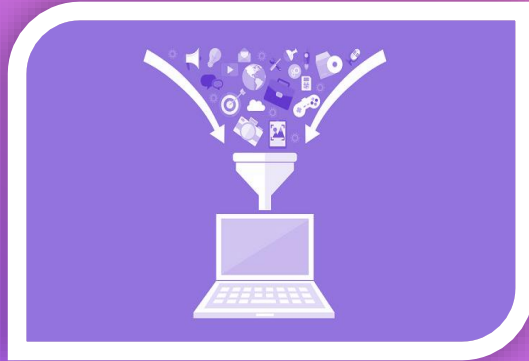
L'objectif :

- Etudier la faisabilité d'un autocomplète à partir de ces données

○



# SOMMAIRE



**Etape 1 – Nettoyer et  
filtrez les features**



**Etape 2 – Identifiez  
& Traitez les outliers**



**Etape 3 – Identifiez & Traitez  
les valeurs aberrantes**



**Etape 4 – Analyses  
uni/bi/multi variée**

# LECTURE ET PREMIÈRE CONSTATATION

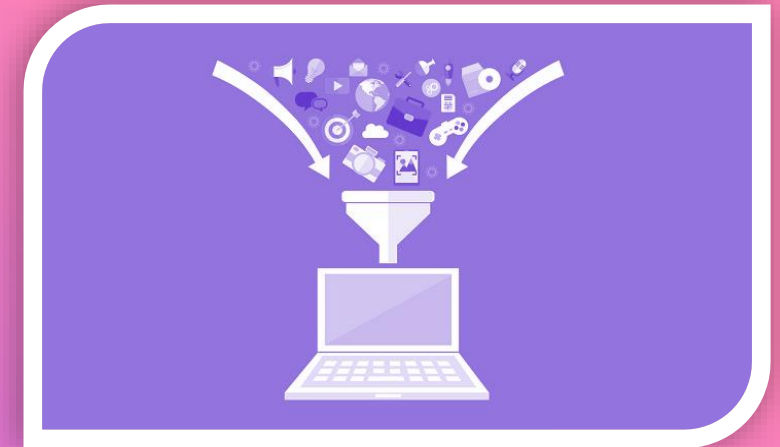
*Data Summary*

dataframe	Values
Number of rows	320772
Number of columns	159

*Data Types*

Column Type	Count
float64	106
string	53

○



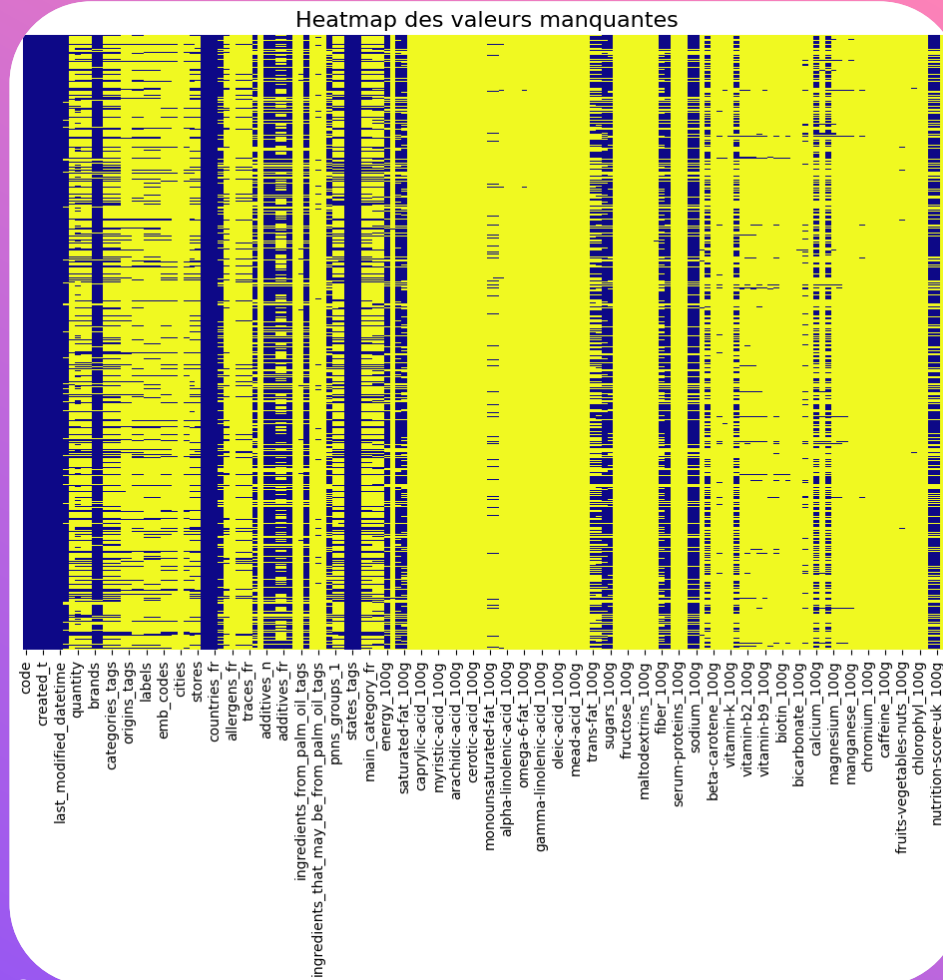
Feature Cible: nutrition\_grade\_fr

# ANALYSE DES FEATURES

Colonne	Moyenne Observée	Unité	Valeur Moyenne Tolérée	Valeur Minimale	Valeur Maximale
energy_100g	1141,92	kJ	800 à 2500	0	5000+
energy_100g	272,92	kcal	200 à 600	0	3500+
fat_100g	12,73	g	0 à 20	0	100
saturated-fat_100g	5,13	g	0 à 10	0	50
carbohydrates_100g	32,07	g	10 à 60	0	90
sugars_100g	16,00	g	0 à 30	0	80
fiber_100g	2,86	g	2 à 10	0	40
proteins_100g	7,08	g	5 à 25	0	60
salt_100g	2,03	g	0 à 5	0	15
sodium_100g	0,80	g	0 à 2	0	5
nutrition-score-fr_100g	9,17	score	-15 à 40	-15	100
nutrition-score-uk_100g	9,06	score	-15 à 40	-15	100



# HEATMAP VALEURS MANQUANTES

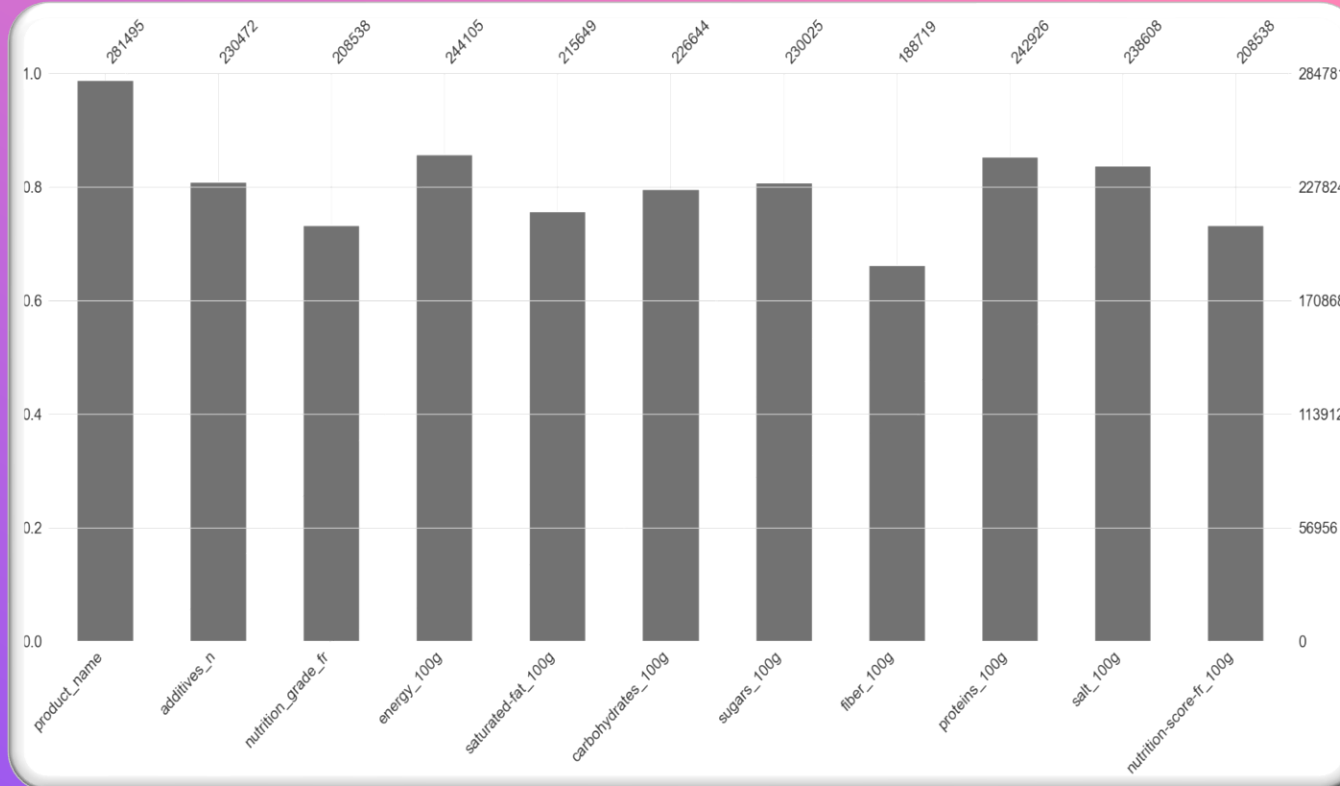


Nombres de colonnes: 168



Avant suppressions des colonnes avec plus de 50% de valeurs manquantes

# DATASET APRÈS TRAITEMENT



**Après suppressions des colonnes avec plus de 50% de valeurs manquantes**

**Nombres de colonnes: 11**



# FEATURES RESTANTES

## Valeurs nutritionnelles

Variable	Description
<b>energy_100g</b>	L'énergie (en kilojoules) contenue dans 100 grammes du produit.
<b>fat_100g</b>	La quantité de graisses (en grammes) contenue dans 100 grammes du produit.
<b>saturated-fat_100g</b>	La quantité de graisses saturées (en grammes) contenue dans 100 grammes du produit.
<b>carbohydrates_100g</b>	La quantité de glucides (en grammes) contenue dans 100 grammes du produit.
<b>sugars_100g</b>	La quantité de sucres (en grammes) contenue dans 100 grammes du produit.
<b>fiber_100g</b>	La quantité de fibres alimentaires (en grammes) contenue dans 100 grammes du produit.
<b>proteins_100g</b>	La quantité de protéines (en grammes) contenue dans 100 grammes du produit.
<b>salt_100g</b>	La quantité de sel (en grammes) contenue dans 100 grammes du produit.
<b>sodium_100g</b>	La quantité de sodium (en grammes) contenue dans 100 grammes du produit.
<b>energy_kcal</b>	L'énergie (en kilocalories) contenue dans 100 grammes du produit.

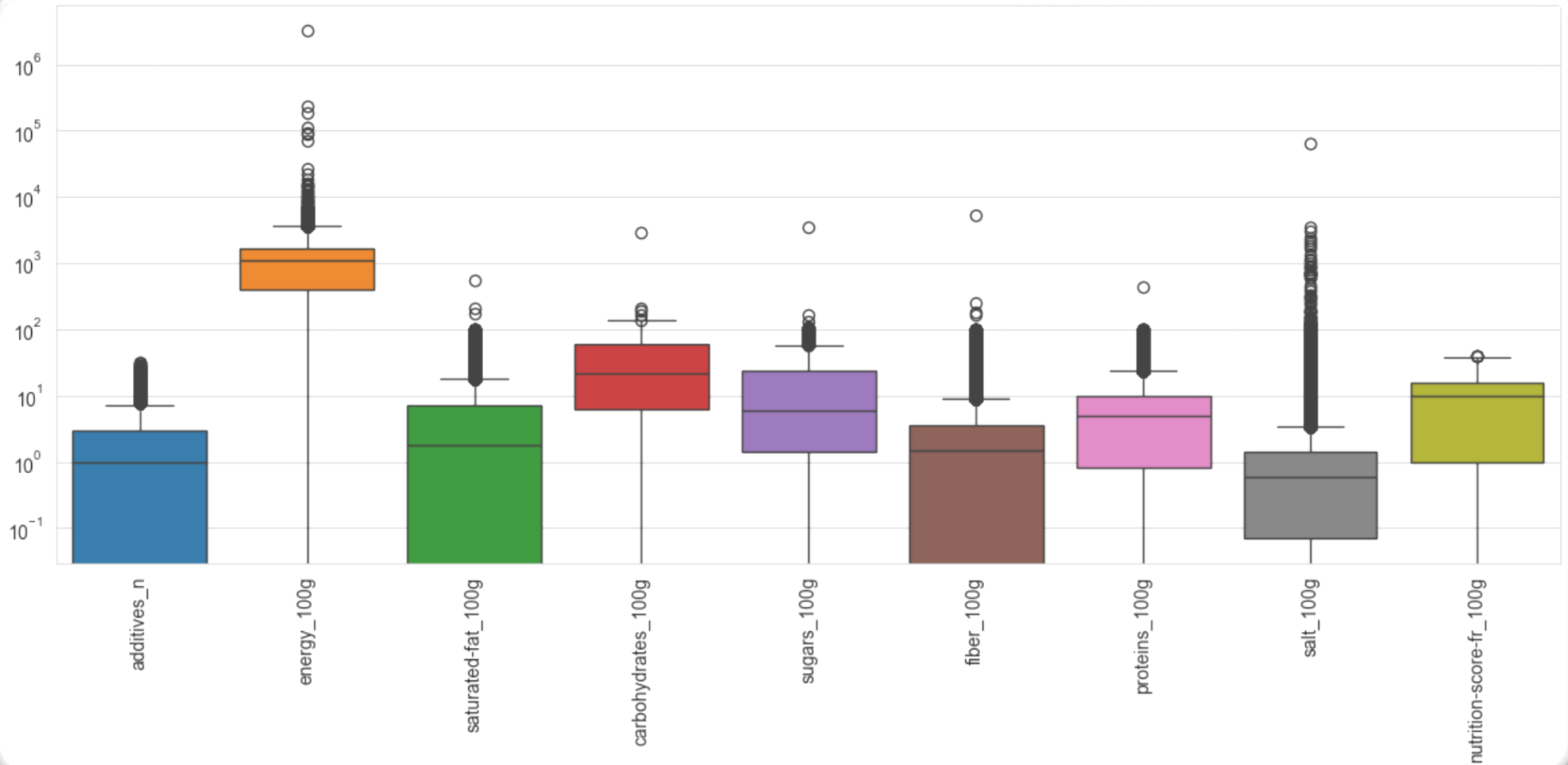
## Scores et Étiquettes

Variable	Description
<b>nutrition_grade_fr</b>	La note nutritionnelle du produit, souvent exprimée sous forme de lettre, en français.
<b>nutrition-score-fr_100g</b>	Le score nutritionnel du produit pour le marché français, basé sur des critères de santé.

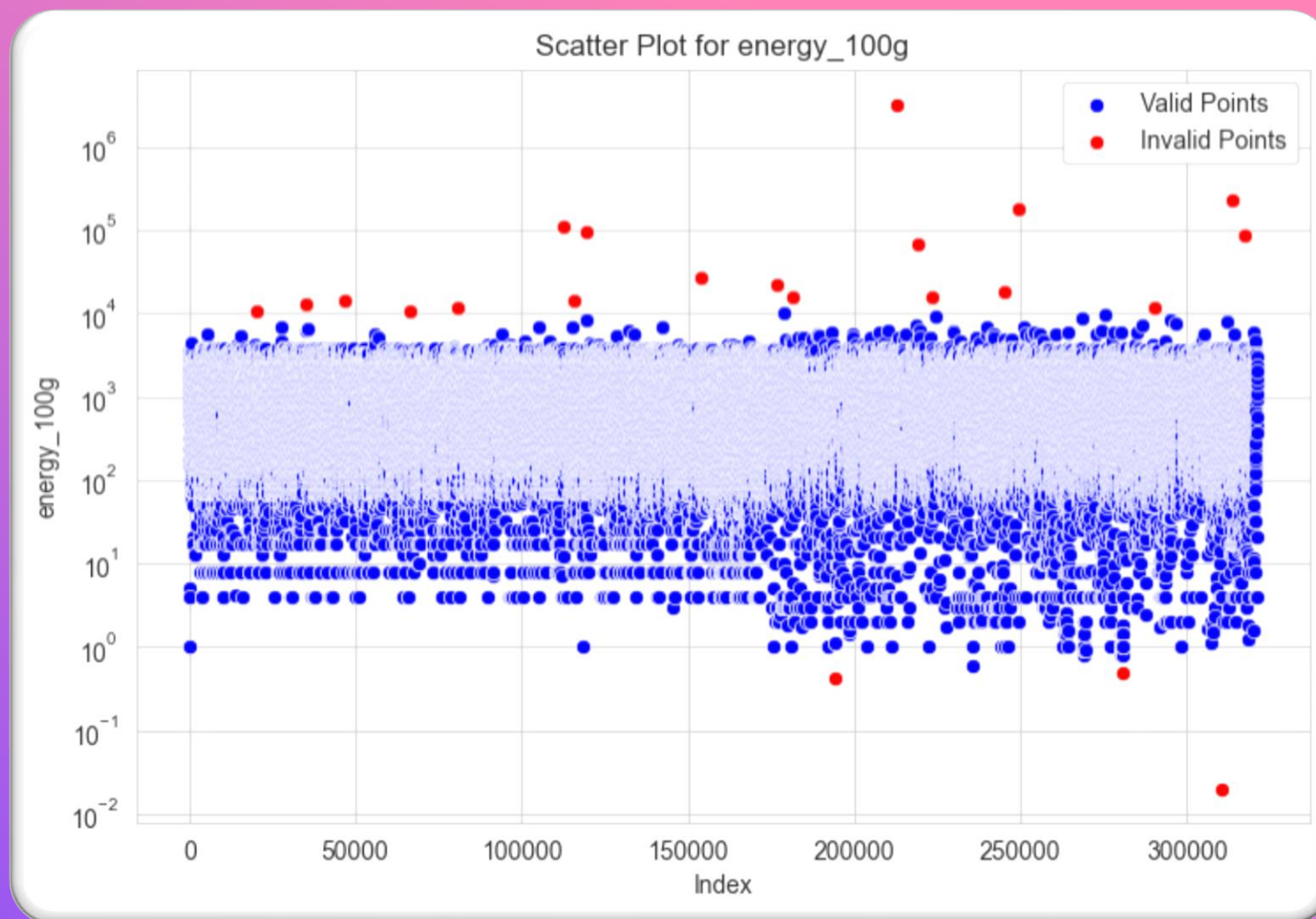


# IDENTIFIEZ LES VALEURS ABERRANTES (BOXPLOT)

9



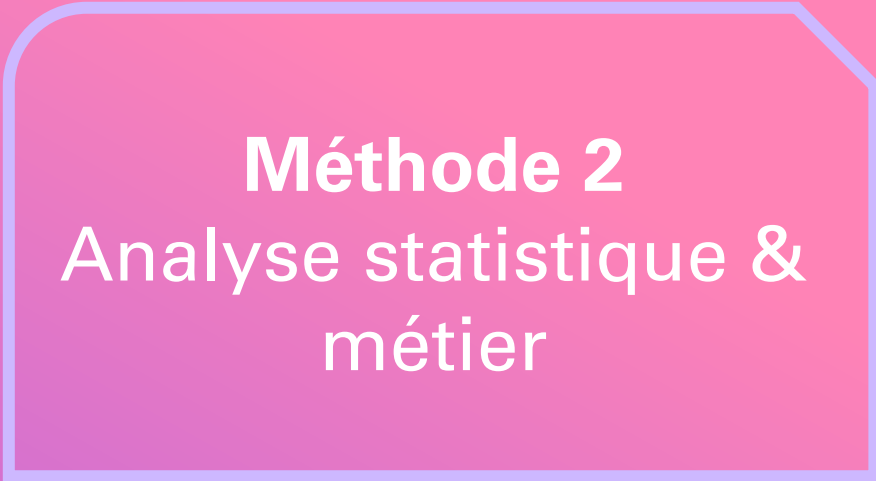
# IDENTIFIEZ LES VALEURS ABERRANTES (SCATTERPLOT)



# TRAITER LES VALEURS ABERRANTES



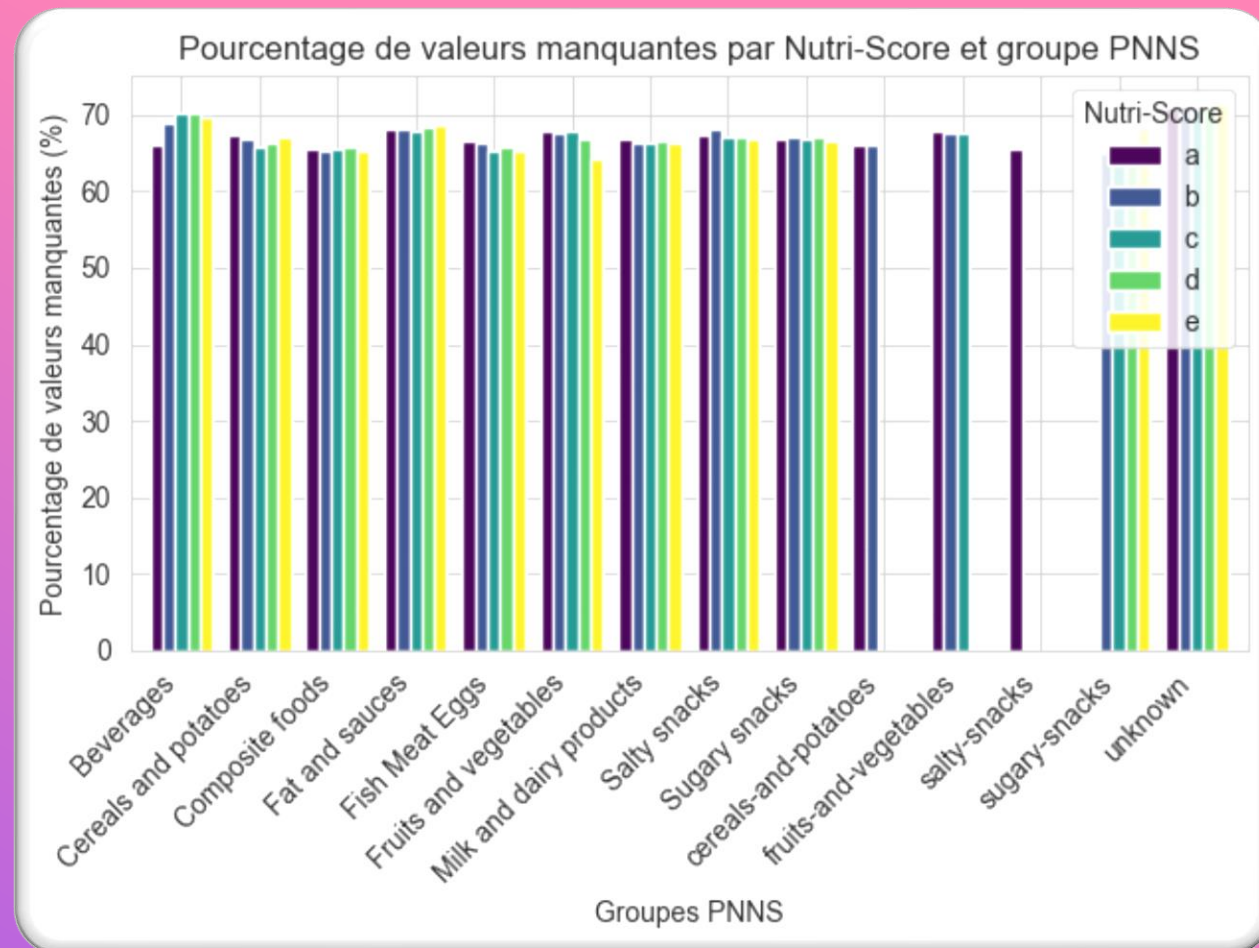
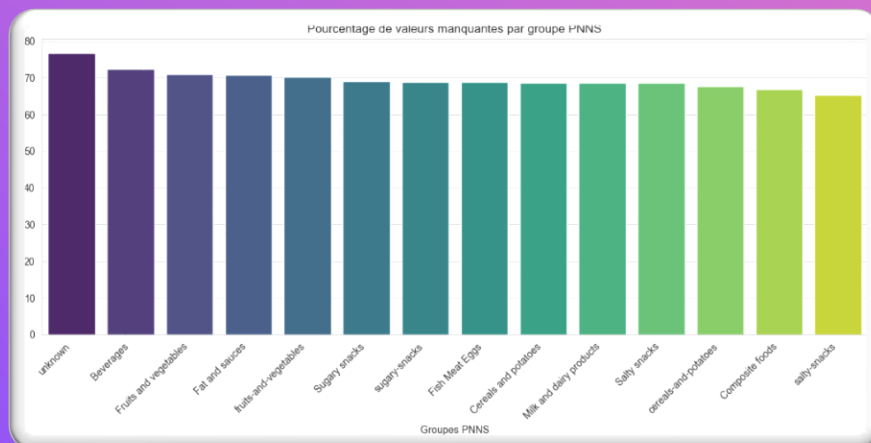
**Méthode 1**  
Analyse métier



**Méthode 2**  
Analyse statistique &  
métier

# PROGRAMME NATIONAL NUTRITION SANTÉ (PNNS)

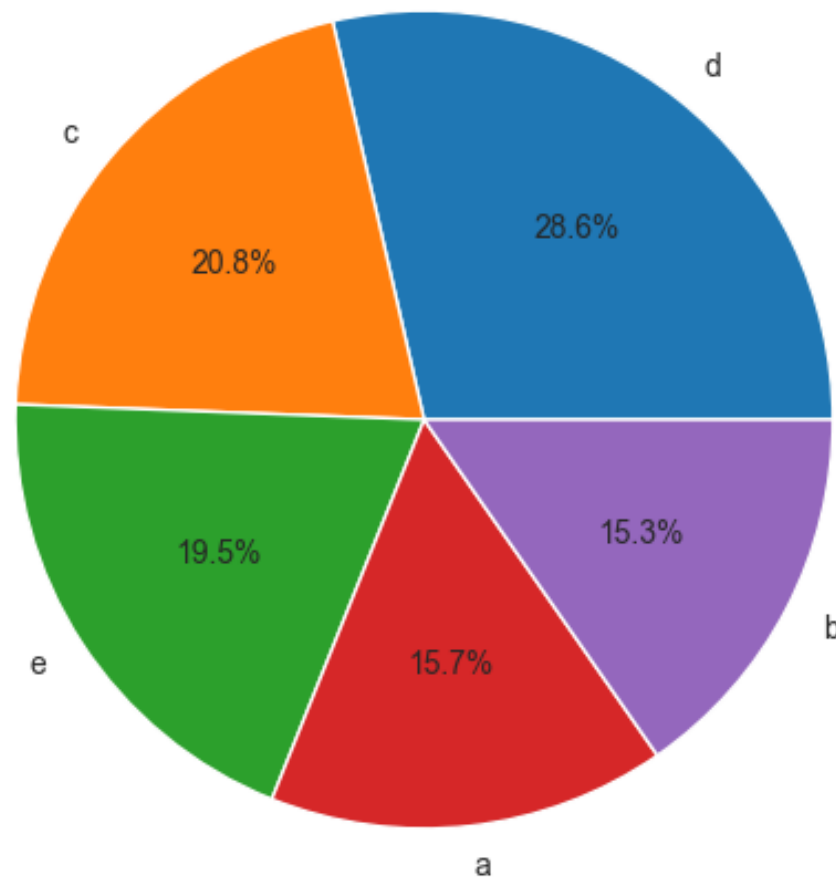
## Méthode 1<sup>+</sup> Analyse métier<sup>•</sup>



# NUTRISCORE GRADE MEDIAN

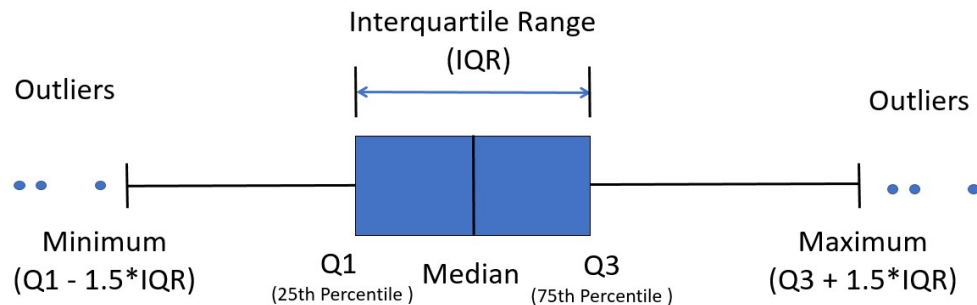
## Méthode 1 Analyse métier

Distribution des nutriscore grades

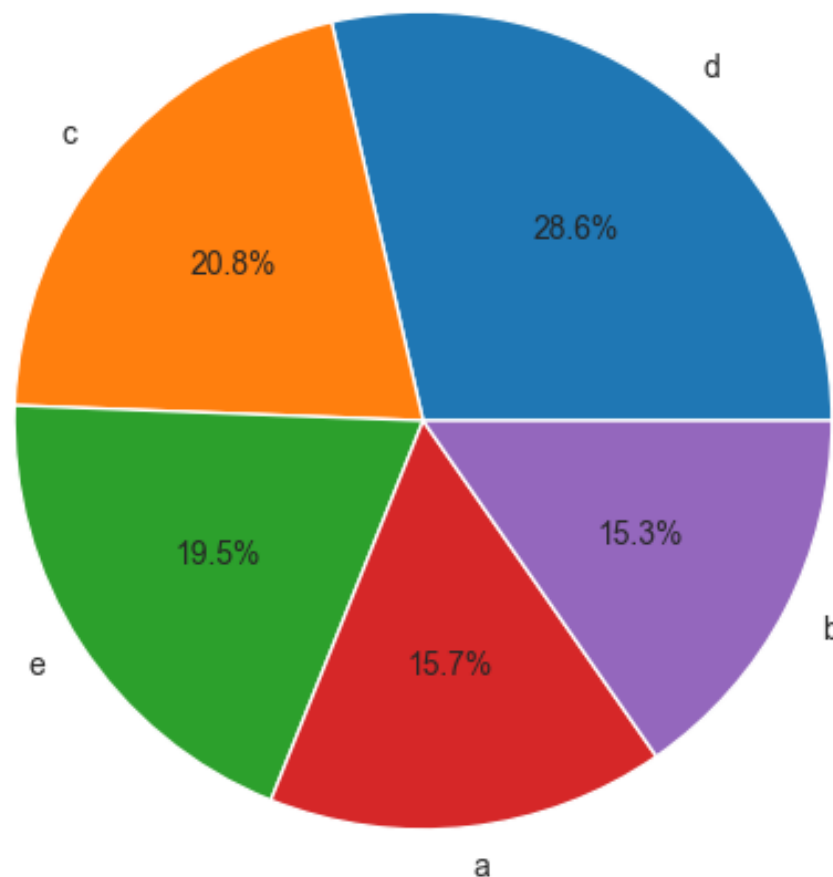


# TRAITER LES VALEURS ABERRANTES

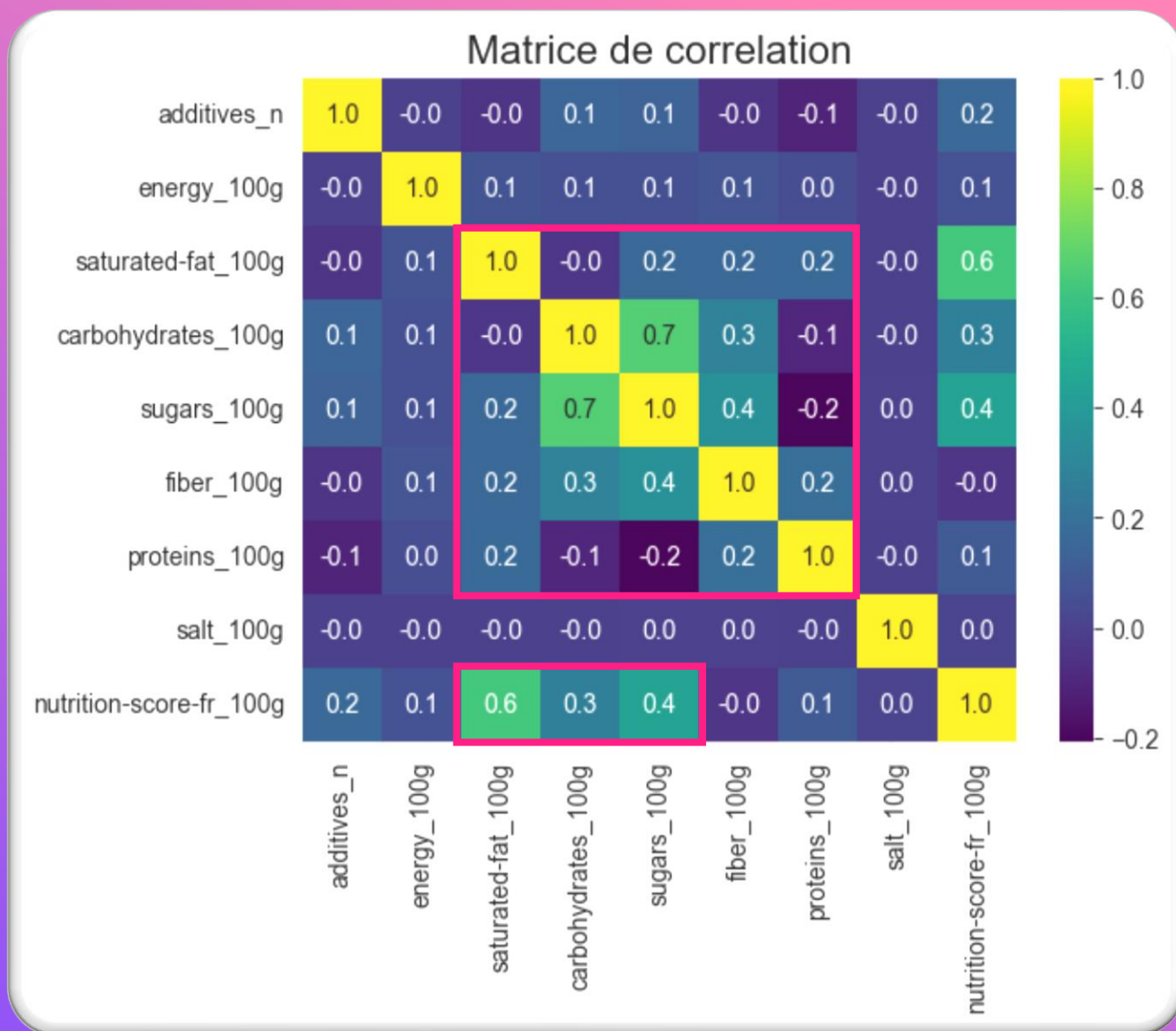
## Méthode 2 Analyse statistique & métier



Distribution des nutriments grades



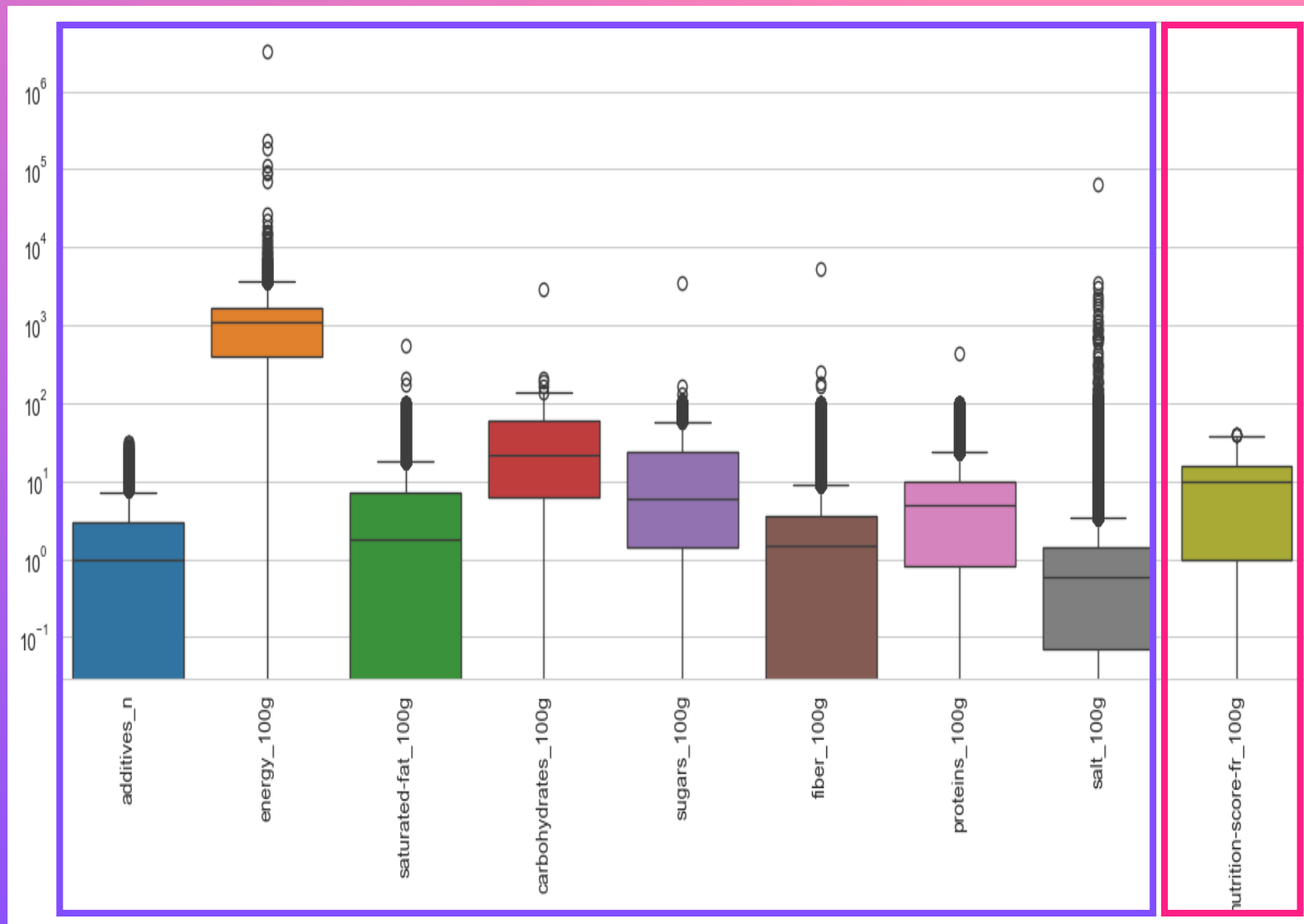
# MATRICE DE CORRELATION



- Corrélations majeures entre nutrition score et les valeurs nutritionnelles
- Corrélations entre toutes les valeurs nutritionnelles



# MÉTHODE DE REMPLACEMENT DES OUTLIERS



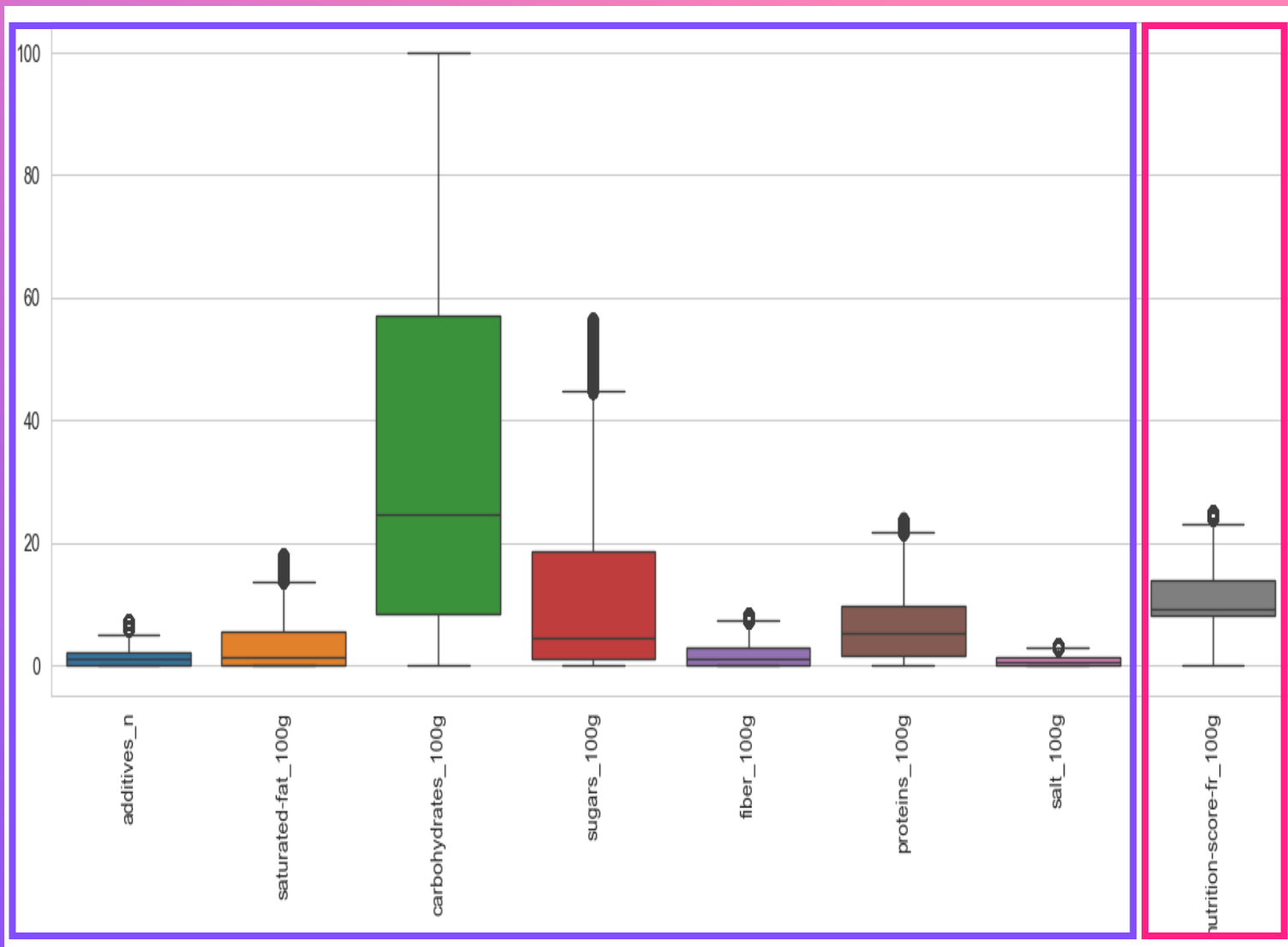
Méthodes de remplacement

Médiane par  
nutriscore group

Itérative imputer

Pour les données qualitative  
remplacement  
par« Unknow »

# APRÈS REMPLACEMENT DES OUTLIERS



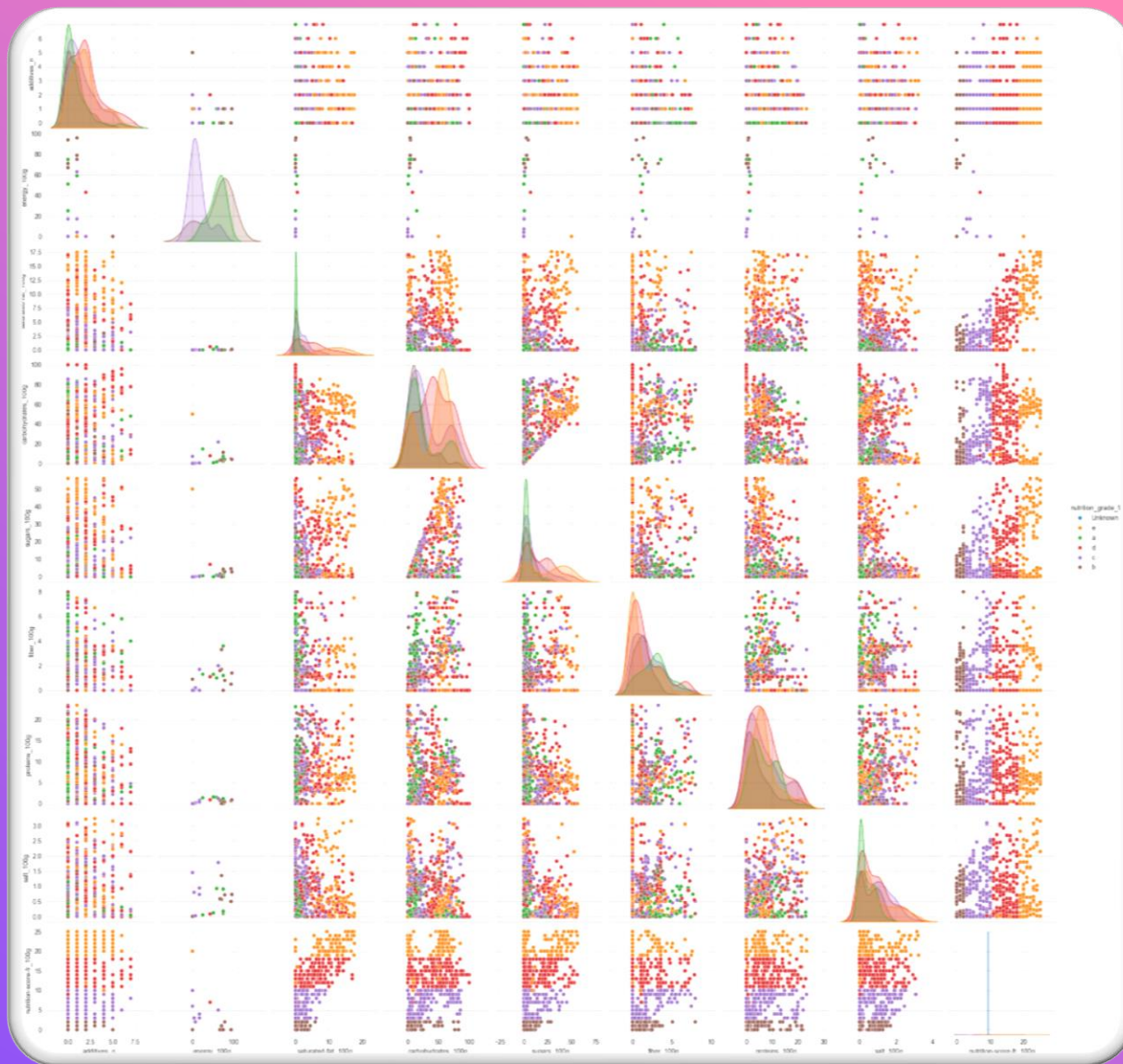
Méthodes de remplacement

■ Médiane par  
nutriscore group

■ Itérative imputer

Pour les données qualitative  
remplacement  
par« Unknow »

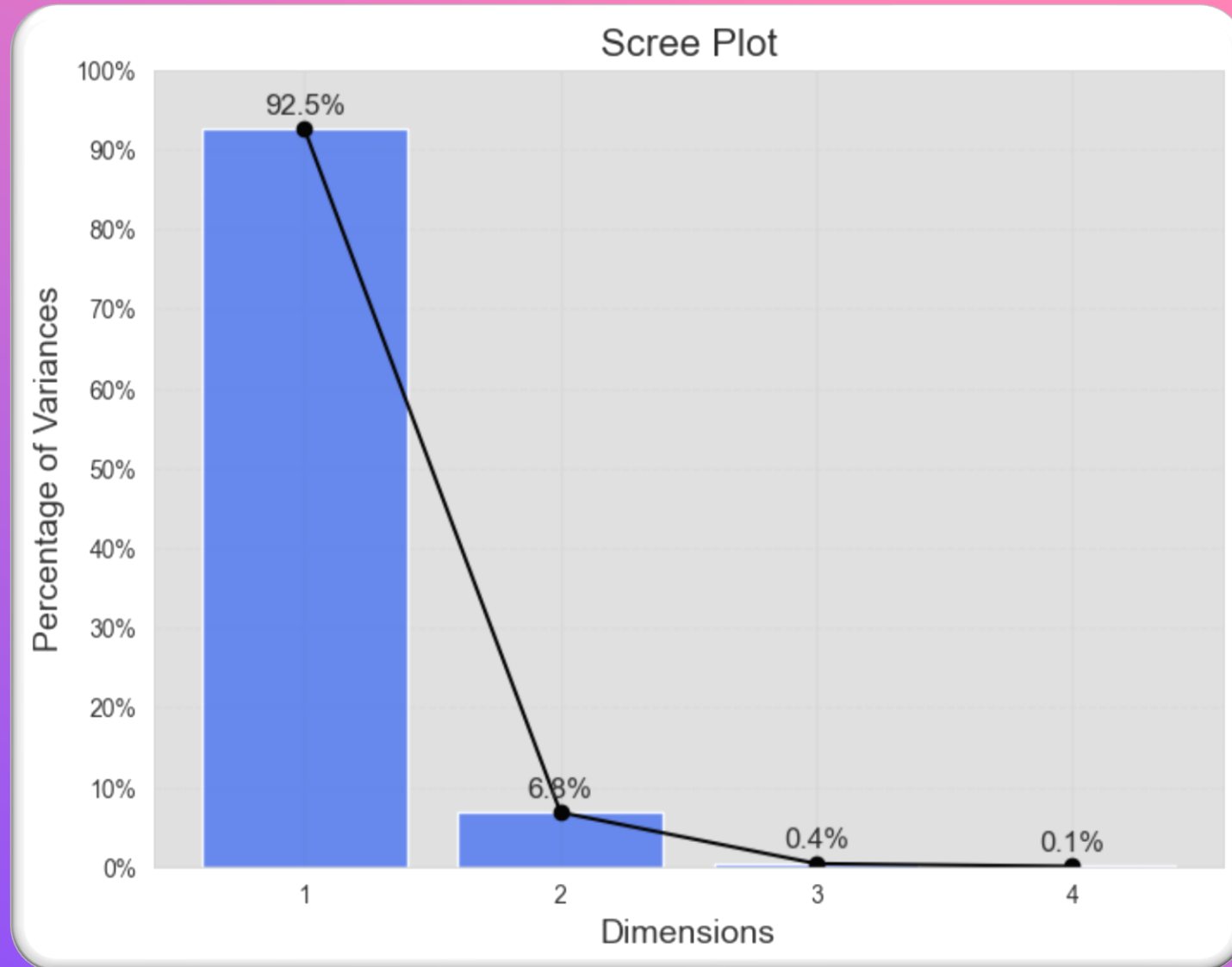
# ANALYSE POST TRAITEMENT



Explorer les corrélations entre les valeurs nutritionnelles et le nutriscore, en visualisant les liens potentiels entre chaque variable.

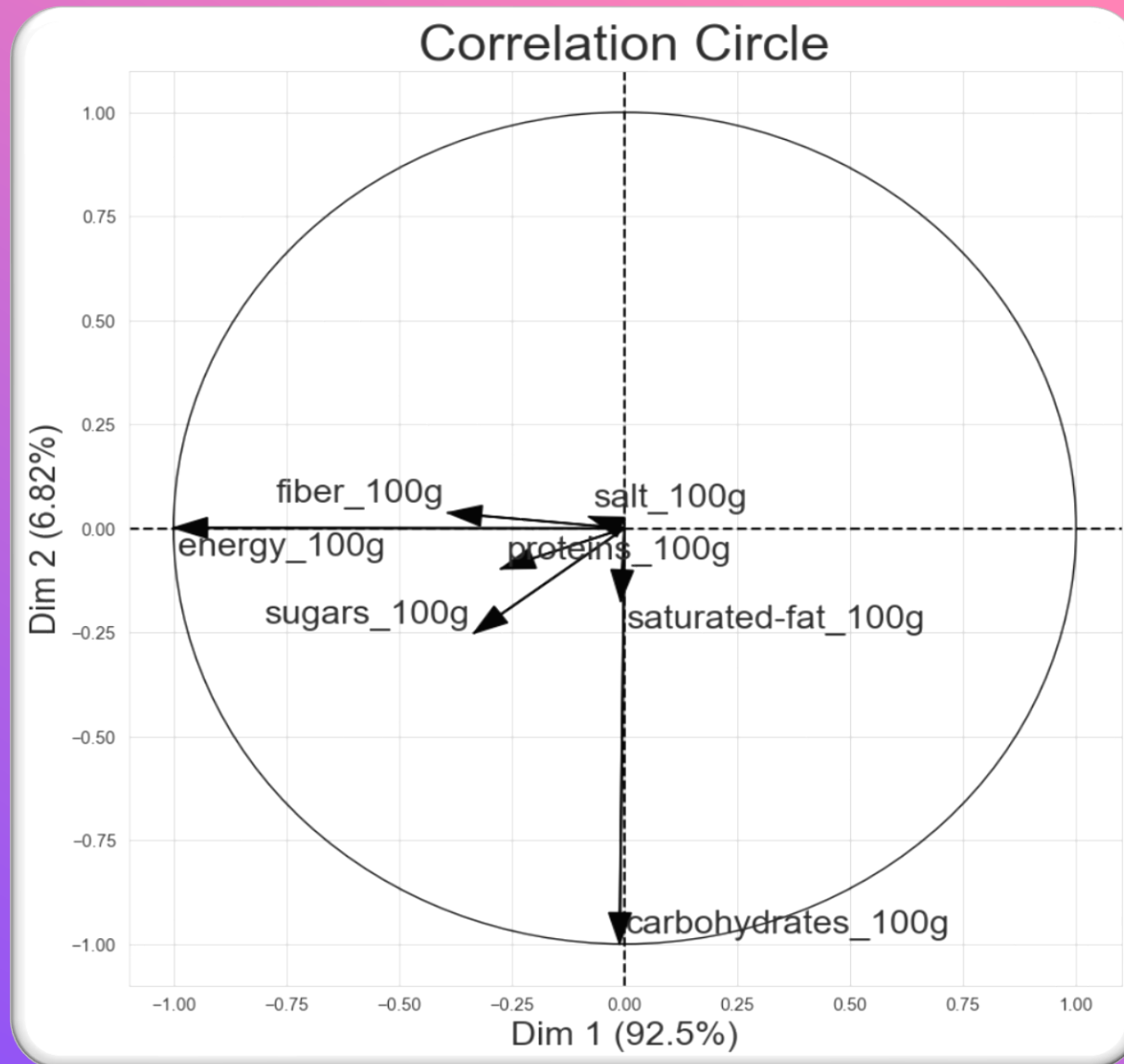
# SCREEPLOT

19



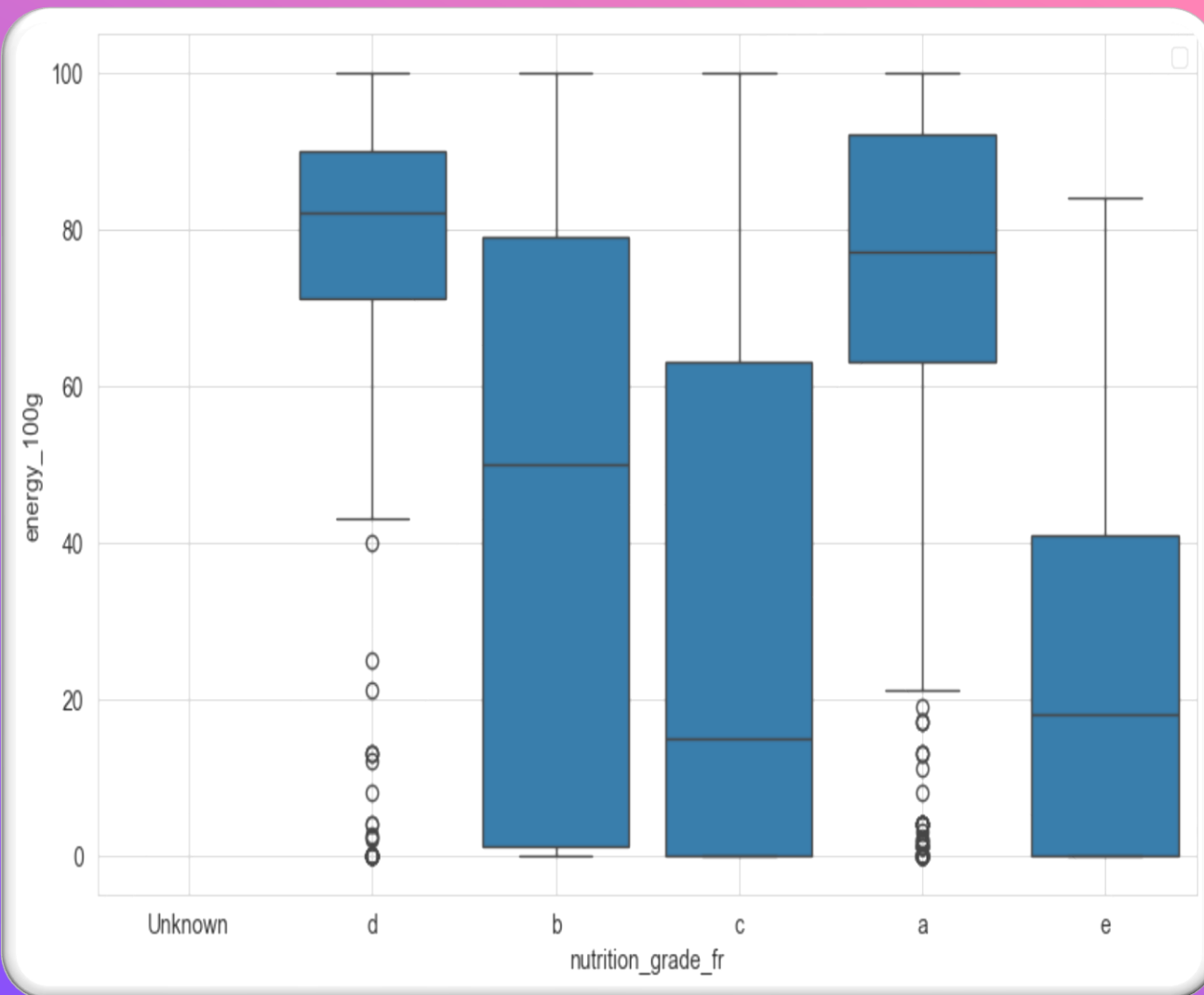
Nombre optimal de  
composante a conservé

# MATRICE DE CORRELATION



Complémentaire au pair plot, aide à la visualisation des relations inter variables

# ANOVA NUTRITION GRADE EN FONCTION DE L'ENERGIE



Variable	F-value	p-value
additives_n	2344.95	0.0
ingredients_from_palm_oil_n	791.87	0.0
ingredients_that_may_be_from_palm_oil_n	306.51	0.0
energy_kj	28023.35	0.0
fat_100g	18397.22	0.0
saturated-fat_100g	34942.42	0.0
carbohydrates_100g	4874.44	0.0
sugars_100g	11301.89	0.0
fiber_100g	3294.61	0.0
proteins_100g	2891.05	0.0
salt_100g	1011.53	0.0
sodium_100g	1011.53	0.0
nutrition-score-fr_100g	335269.58	0.0
nutrition-score-uk_100g	244228.88	0.0

# EST-T-IL POSSIBLE DE FAIRE UN SYSTÈME D'AUTO-COMPLÉTION ?



## Avantages de l'auto-complétion dans Open Food Facts

Gain de temps et d'efficacité

Précision et cohérence des données

Amélioration de la qualité des données

Facilitation de l'exploration des données

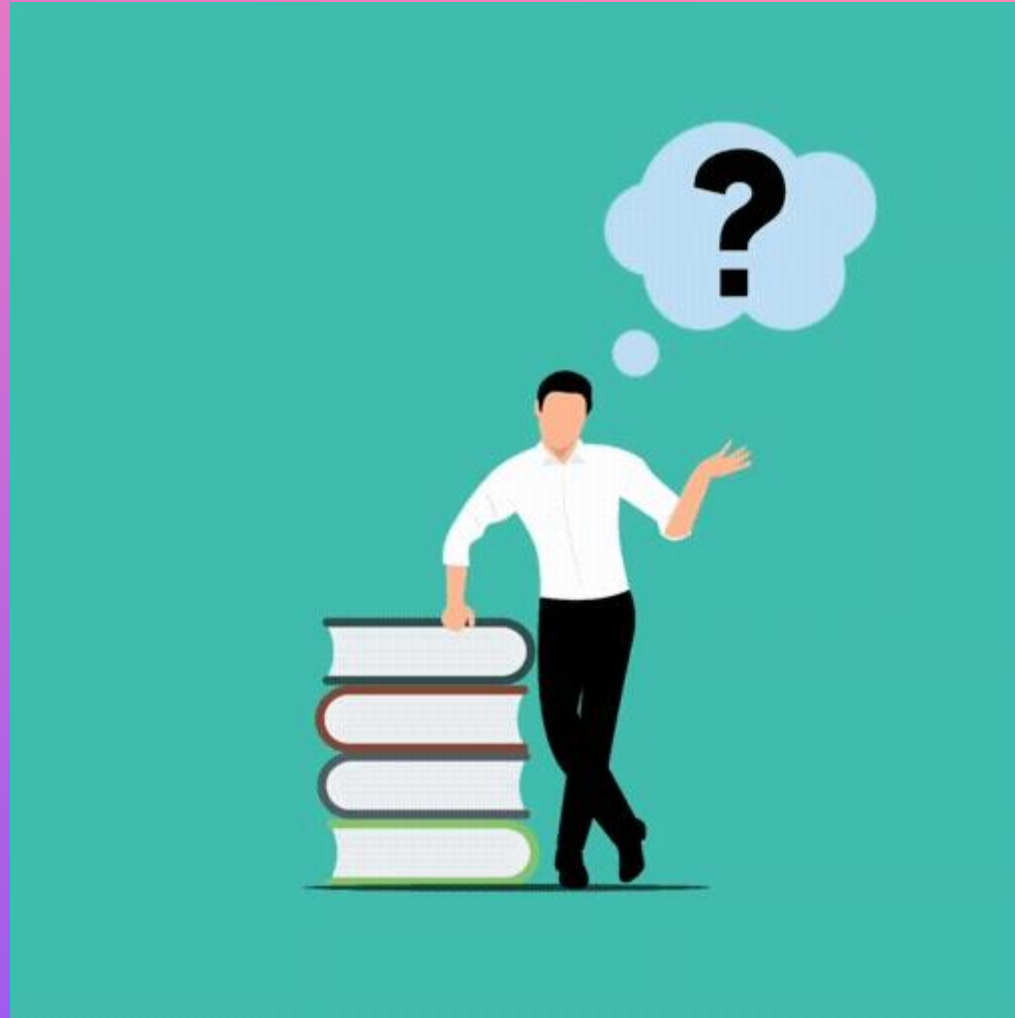
Personnalisation et adaptation aux besoins spécifiques



# RESPECT DES RGPD

Critères	Justifications
<b>Nature des Données Traitées</b>	Le dataset utilisé (openfoodfacts_dataset.csv) contient des informations sur des produits alimentaires (valeurs nutritionnelles, marques, pays, etc.) qui sont généralement anonymisées.
<b>Absence de Données Personnelles</b>	Les données manipulées ne contiennent pas d'informations personnelles (noms, adresses, numéros de téléphone, identifiants uniques) permettant d'identifier une personne physique.
<b>Finalité du Traitement</b>	L'objectif principal est de nettoyer et préparer des données pour des analyses statistiques et des problématiques métier liées à la santé publique, orienté vers l'amélioration des données au niveau agrégé.
<b>Conformité aux Principes du RGPD</b>	Bien que le RGPD ne soit pas directement applicable, des bonnes pratiques de gestion des données (minimisation des données, sécurisation des informations) sont respectées.
<b>Absence de Consentement et de Base Légale</b>	Les notions de consentement et de base légale ne s'appliquent pas, car les données ne concernent pas des personnes physiques.

# DES QUESTIONS ?



# MERCI

Ethan VUILLEMIN

07 63 76 58 31

[vuilleminethan@gmail.com](mailto:vuilleminethan@gmail.com)

