Santé publique France

P 3
-
PRÉPAREZ DES DONNÉES POUR UN ORGANISME DE SANTÉ PUBLIQUE

Charlotte DUBUS

Soutenance du 22 décembre 2023 – Parcours Data Scientist

INTRODUCTION

Projet en collaboration avec Santé Publique France

Développement d'un système de suggestion ou d'auto-complétions

Présentation du nettoyage et de l'exploration des données internes :

- Utilisation des données de l'application 'Open Food Facts'
- 2 concepts clés omniprésents : Le Nutri-score et le Plan Nationale Nutrition Santé (PNNS)

OPEN FOOD FACTS

Santé publique France

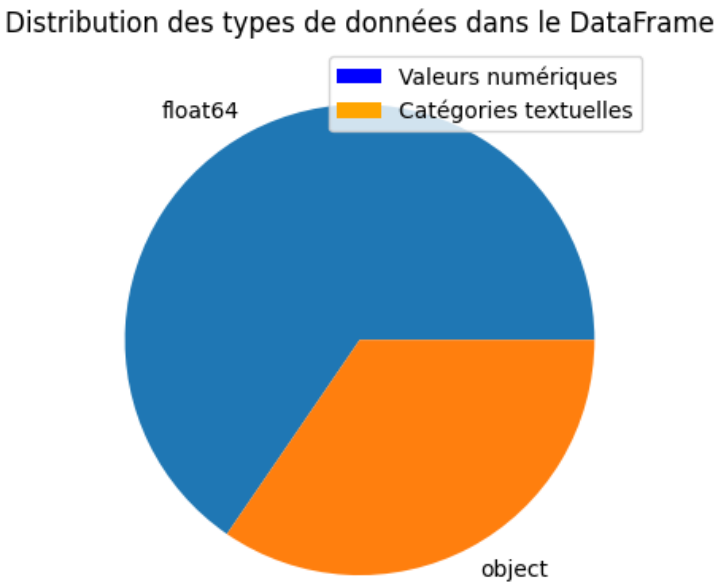
SOMMAIRE

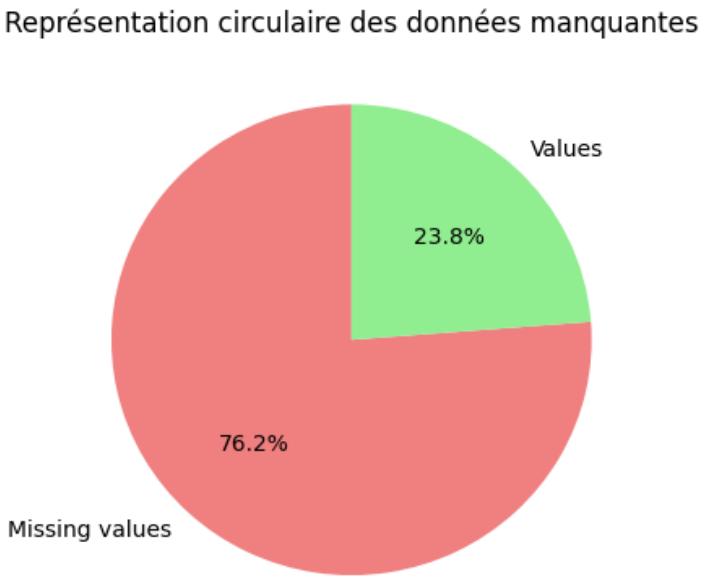
RESPECT DES PRINCIPES RGPD

1. Licéité, loyauté et transparence
2. Limitation des finalités
3. Minimisation des données
4. Exactitude
5. Limitation de la conservation

Analyse de forme

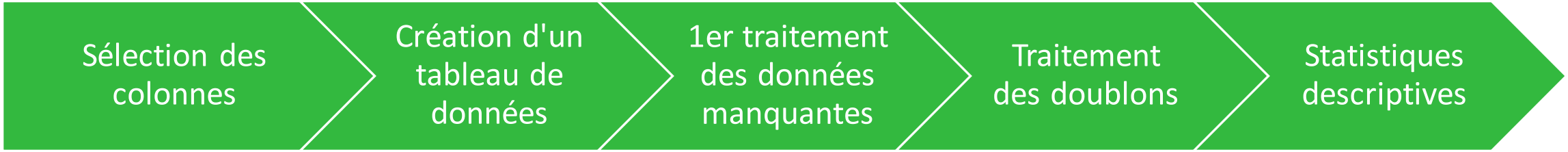
- 320 772 produits alimentaires
- 162 colonnes : 106 numériques & 56 catégorielles
- Beaucoup de données manquantes : Filtrage





1. ANALYSE ET PRÉ-EXPLORATION DES DONNÉES

Analyse de fond



1 .ANALYSE ET PRE-EXPLORATION DES DONNEES

Données catégorielles

Valeurs numériques

code	product_name	pnns_groups_1	nutrition_grade_fr	energy_100g	fiber_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	proteins_100g	salt_100g	sugars_100g	iron_100g	nutrition-score-fr_100g
0000000003087	Farine de blé noir	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
0000000004530	Banana Chips Sweetened (Whole)	NaN	d	2243.0	3.6	28.57	28.57	0.018	64.29	3.57	0.00000	14.29	0.00129	14.0
0000000004559	Peanuts	NaN	b	1941.0	7.1	17.86	0.00	0.000	60.71	17.86	0.63500	17.86	0.00129	0.0
0000000016087	Organic Salted Nut Mix	NaN	d	2540.0	7.1	57.14	5.36	NaN	17.86	17.86	1.22428	3.57	0.00514	12.0
0000000016094	Organic Polenta	NaN	NaN	1552.0	5.7	1.43	NaN	NaN	77.14	8.57	NaN	NaN	NaN	NaN

- **Après cette étape, l'ensemble de données comprend désormais 160 544 entrées et 15 colonnes**

2. NETTOYAGE DES DONNEES

Préparer le jeu de données pour les analyses

Conserver un maximum d'informations

Colonnes quantitatives

Aperçus statistiques et graphiques de la variable

- Recherches nutritionnelles
- Statistique descriptive
- Graphique de répartitions des données

Recherche des valeurs dites aberrantes

- Filtrage
- Modification
- Suppression
- Transformation des aberrants en manquants

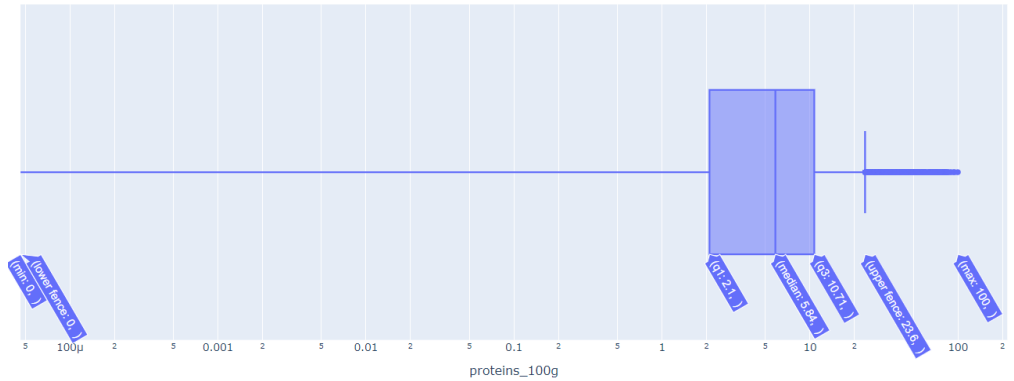
Traitement des données manquantes

- Completion
- Conversion des valeurs manquantes
- Vérification statistique & graphique

Statistiques descriptives

```
count    156415.000000
mean          7.739460
std           7.964189
min           0.000000
25%           2.100000
50%           5.840000
75%          10.710000
max         100.000000
Name: proteins_100g, dtype: float64
```

Graphique 'Boxplot' de la répartition



Données

code	product_name	pnns_groups_1	nutrition_grade_fr	energy_100g	fiber_100g	fat_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	proteins_100g	salt_100g	sugars_100g	iron_100g	nutrition-score-fr_100g
0016473902352	bari, pomegranate balsamic vinegar	NaN	c	724.0	0.0	0.0	0.0	0.000000	40.0	93.33	0.05080	38.33	NaN	5.0
0070552903203	unflavored gelatin	NaN	b	1795.0	0.0	0.0	0.0	0.017307	0.0	100.00	0.36322	0.00	NaN	1.0
0718122726011	bari, balsamic vinegar, fig	NaN	c	724.0	0.0	0.0	0.0	0.000000	40.0	93.33	0.05080	38.33	NaN	5.0
0812603011556	tcho-a-day dark chocolate	NaN	e	2301.0	0.0	35.0	22.5	0.000000	55.0	100.00	0.00000	40.00	0.00495	24.0
3183280016354	le saunier de camargue	unknown	e	243.0	0.0	NaN	36.0	0.017307	NaN	96.00	100.00000	0.50	NaN	20.0
3257983765946	edulcorant a l'extrait de stevia	unknown	b	1556.0	0.0	NaN	0.0	0.017307	NaN	93.10	0.08000	6.90	NaN	0.0
3286011051744	lingettes pocket pour visage et mains, biodégr...	unknown	d	1912.0	1.0	NaN	6.0	0.017307	NaN	99.00	2.00000	6.00	NaN	18.0
3350033331259	fromage blanc brebis	NaN	a	364.0	0.0	NaN	3.2	0.017307	NaN	94.70	0.10000	2.90	NaN	-1.0
8009958000218	mini gressins	NaN	d	1640.0	3.5	NaN	4.0	0.017307	NaN	95.00	2.00000	2.50	NaN	11.0

EXEMPLE AVEC PROTEINS_100G

EXEMPLE AVEC PROTEINS_100G

Recherche nutritionnelle

- 100g de produit = 100g de protéines >>> impossible
- Exemple d'aliments riche en protéines : Gélatine alimentaire (84.5g)

Suppression

- Produit non alimentaire : 'lingettes pocket pour visage'

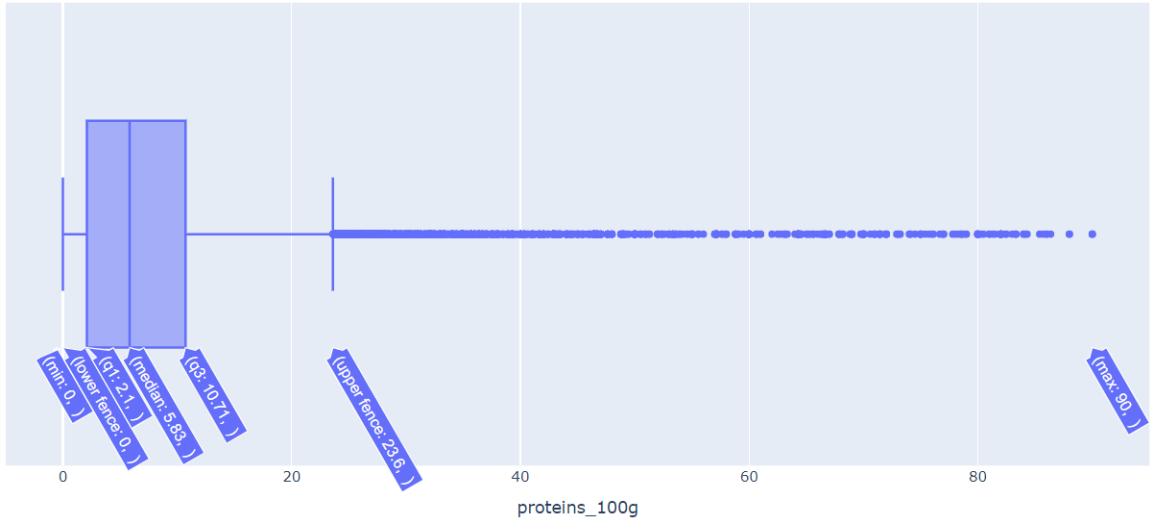
Modification

- Vérification des données sur 'OpenFoodFacts'
- Changement des données nutritionnelles incorrects

Traitement

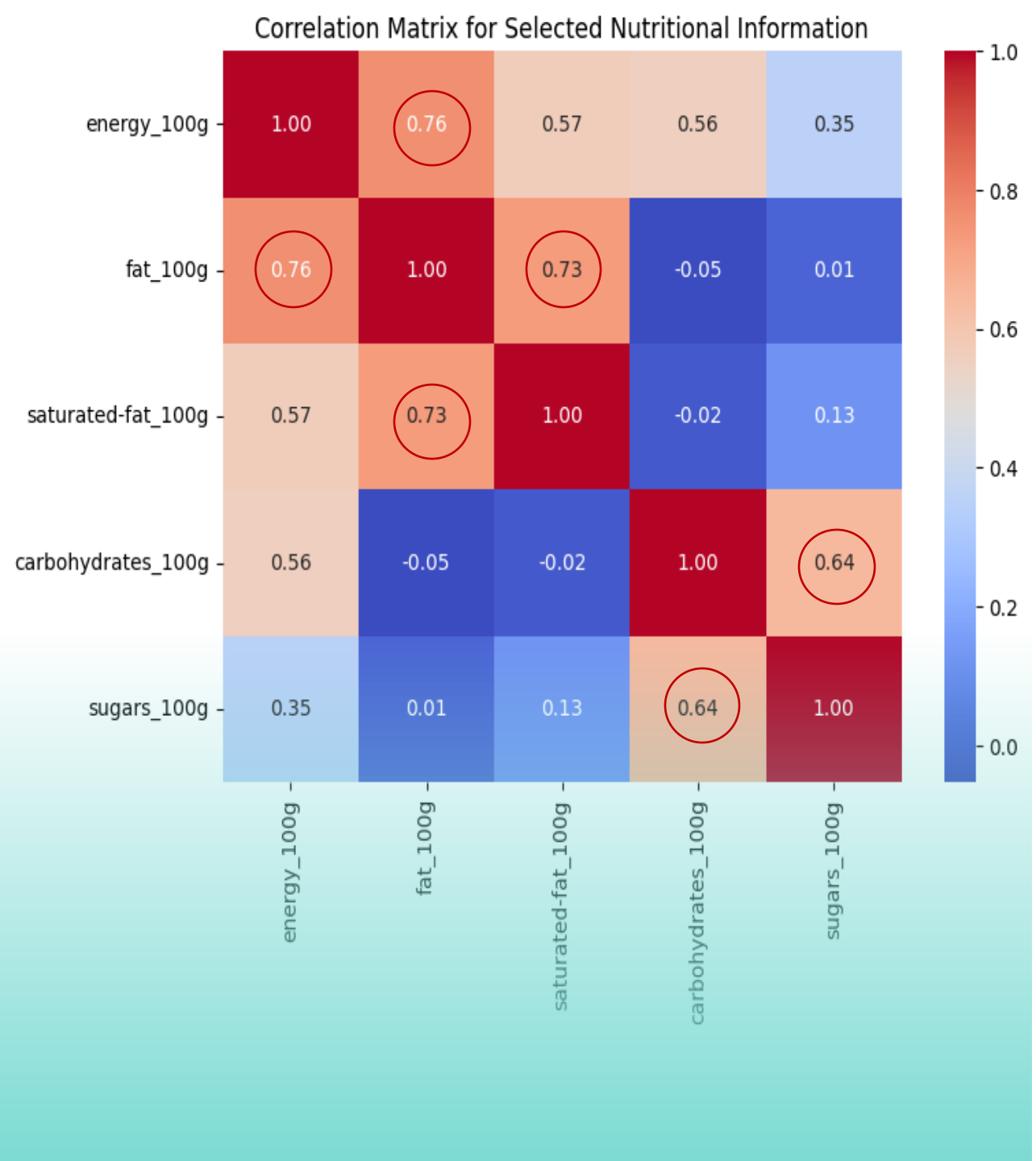
- Remplacement des informations dites aberrantes en manquantes
- Attribution de la valeur de la médiane aux manquants

Graphique 'Boxplot' de la répartition



Statistiques descriptives

```
count    156563.000000
mean          7.732418
std           7.932460
min           0.000000
25%           2.100000
50%           5.830000
75%          10.710000
max          90.000000
Name: proteins_100g, dtype: float64
```



Colonnes quantitatives corrélées

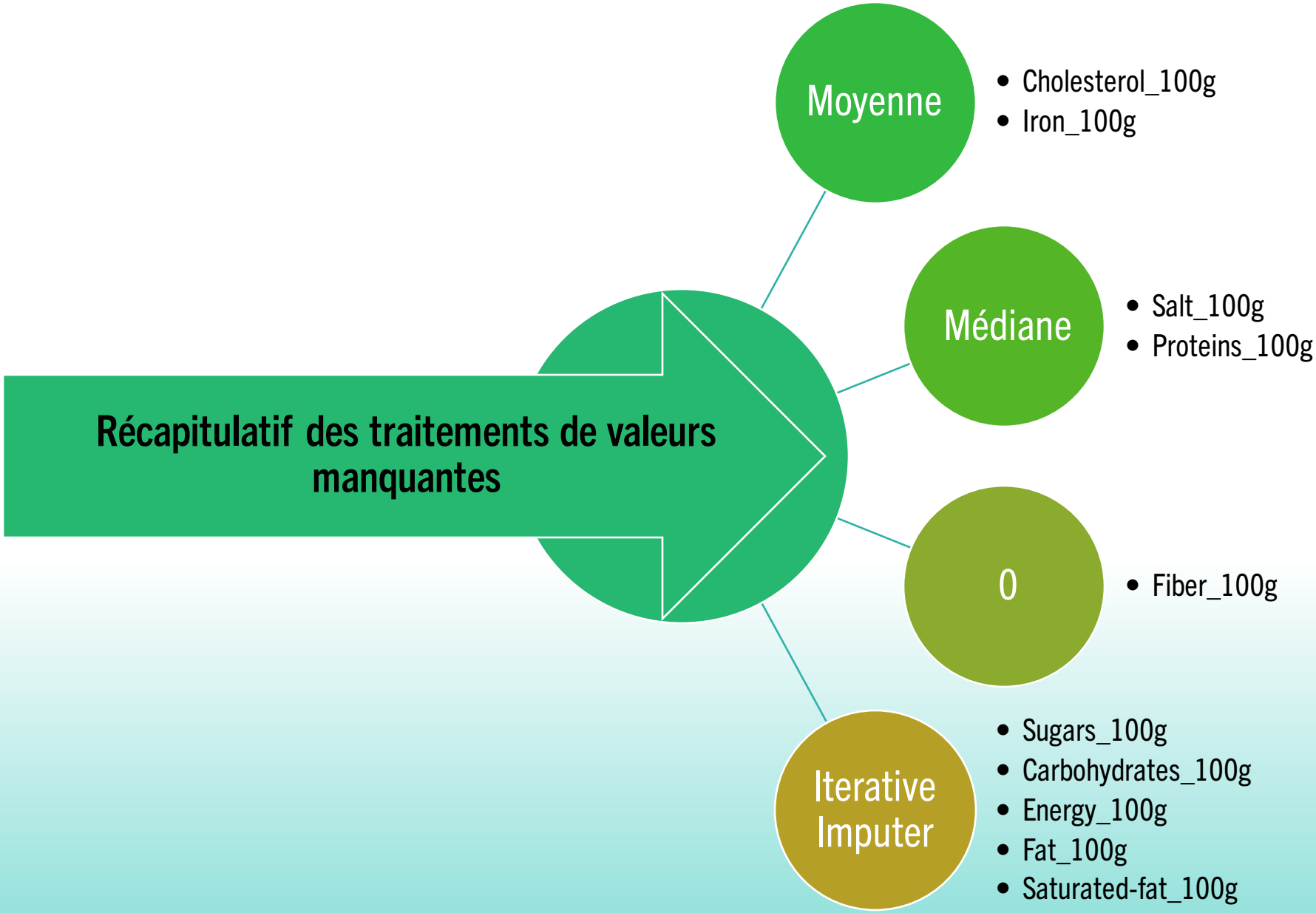
Traitement des manquants

Energy_100g >> Fat_100g

Saturated-fat_100g >> Fat_100g

Sugars_100g >> Carbohydrates_100g

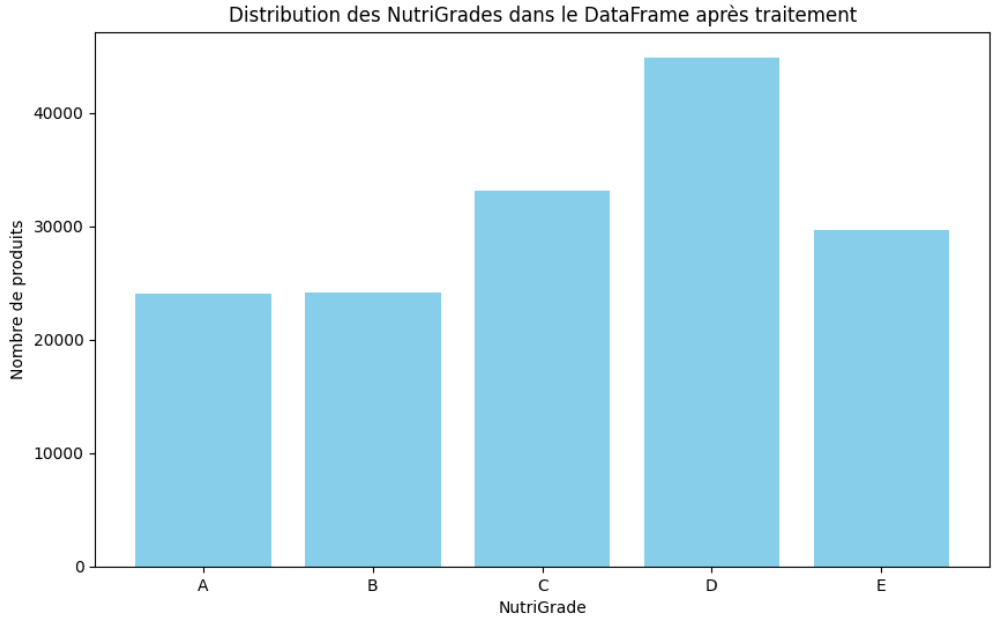
Iterative Imputer



2.NETTOYAGE DES DONNEES

Score final	NUTRI-SCORE
-15 à -1	
0 à 2	
3 à 10	
11 à 18	
19 à 40	

NUTRIGRADE



2.NETTOYAGE DES DONNEES

PNNS 1

Analyse visuelle des catégories

- Aperçu de la répartitions dans chaque groupe
- Détection d'anomalies

Modification des erreurs typographiques

- *Exemple: Fruits and vegetables = fruits-and-vegetables*

Intégration des 'Unknown' dans les manquants

Réduction des manquantes

- Utilisation de 'mots clés' avec 'product name' pour attribuer les catégories

Imputation des données manquantes

- Utilisation d'un classificateur appelé 'RandomForestClassifier'

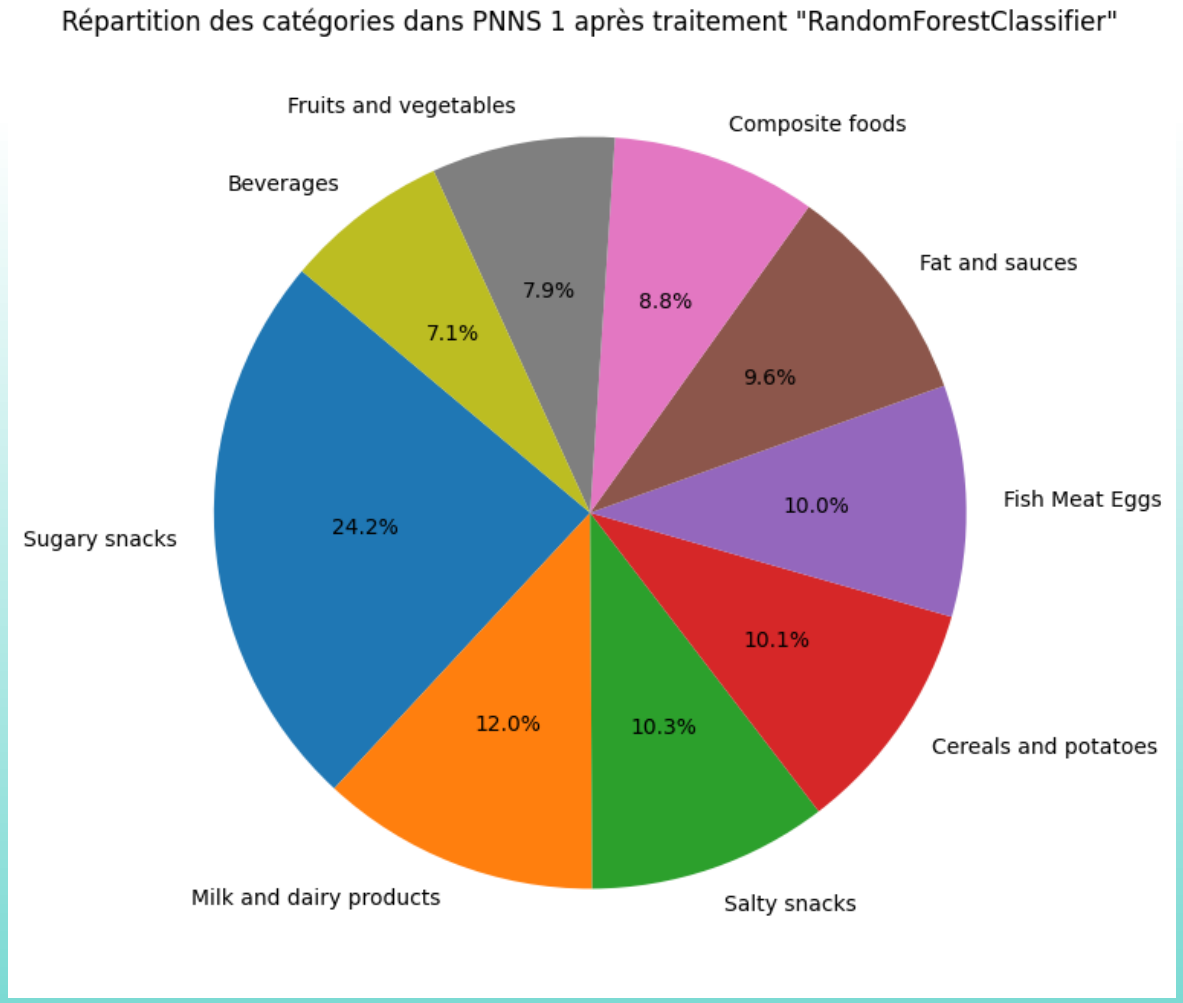
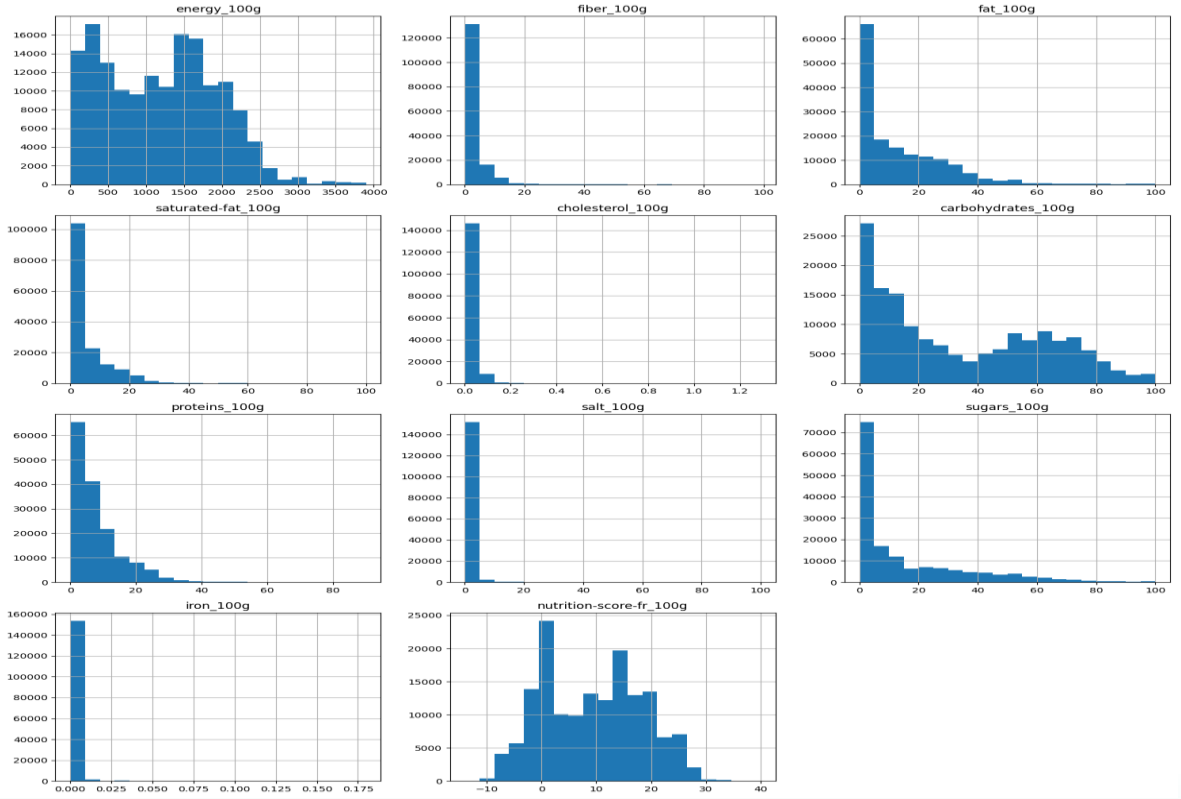
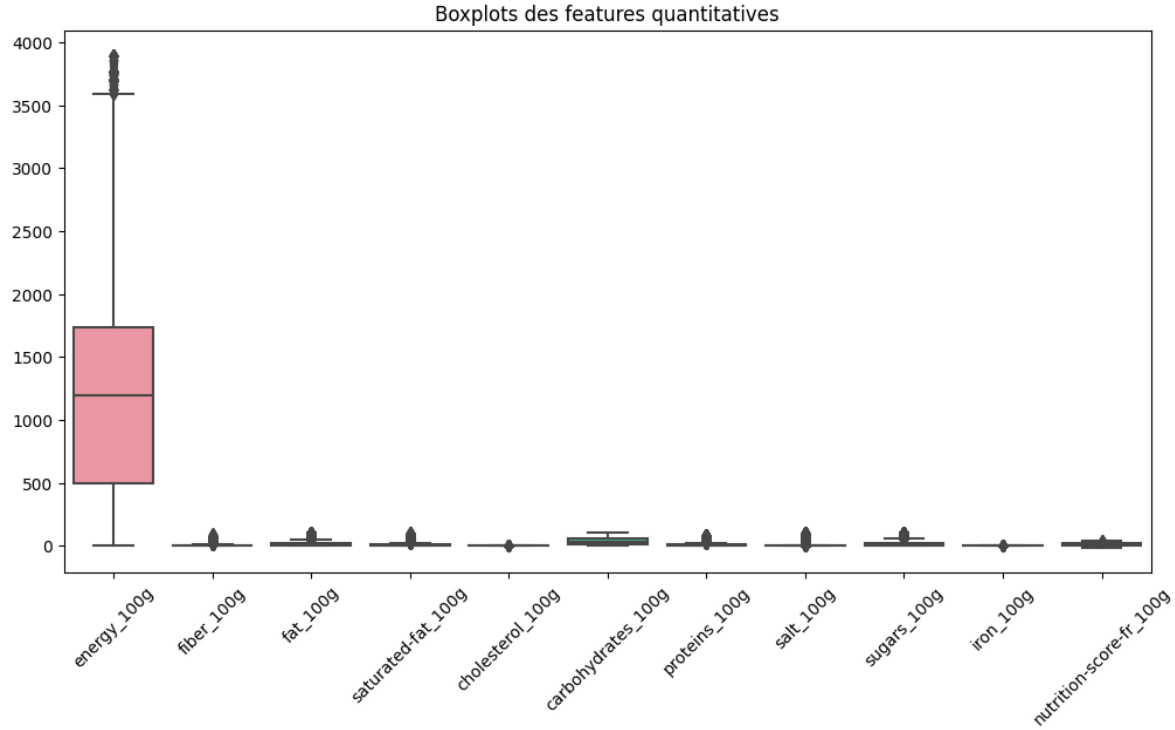
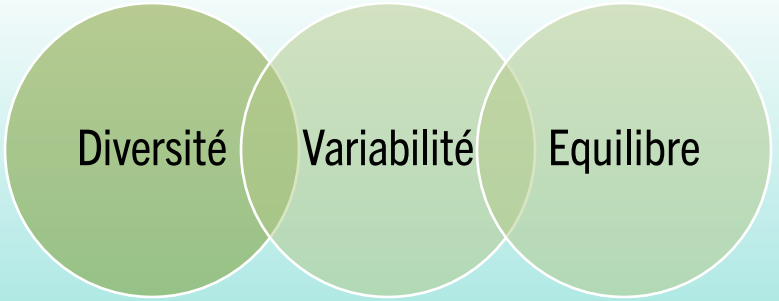


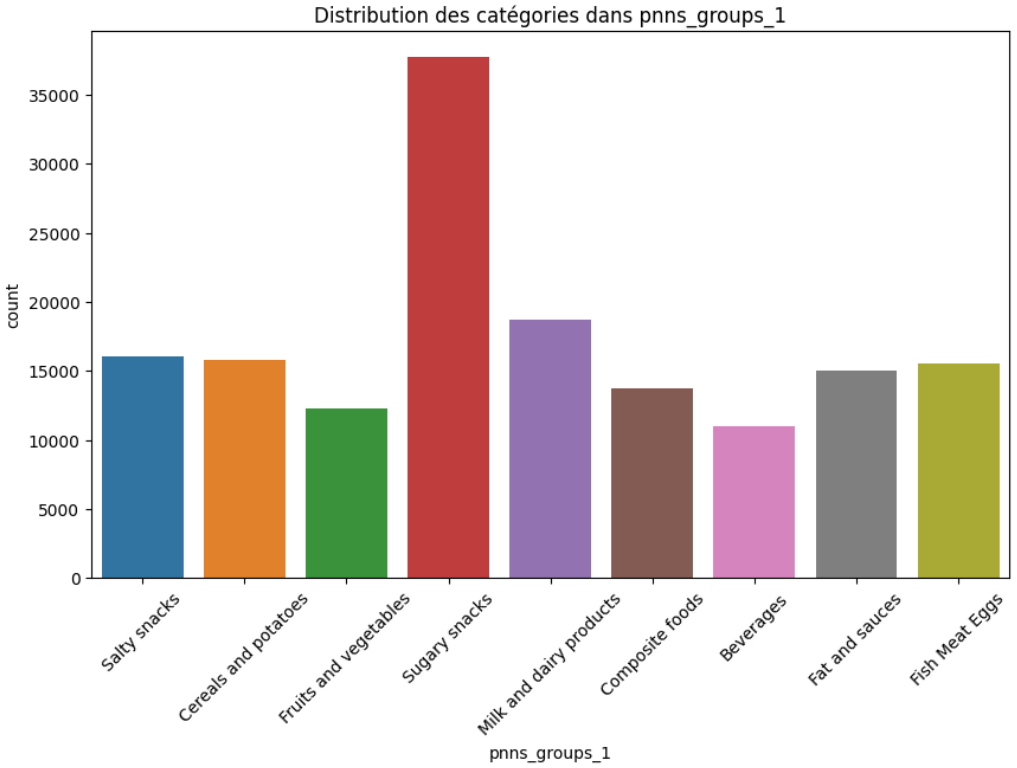
Diagramme de répartition des variables numériques

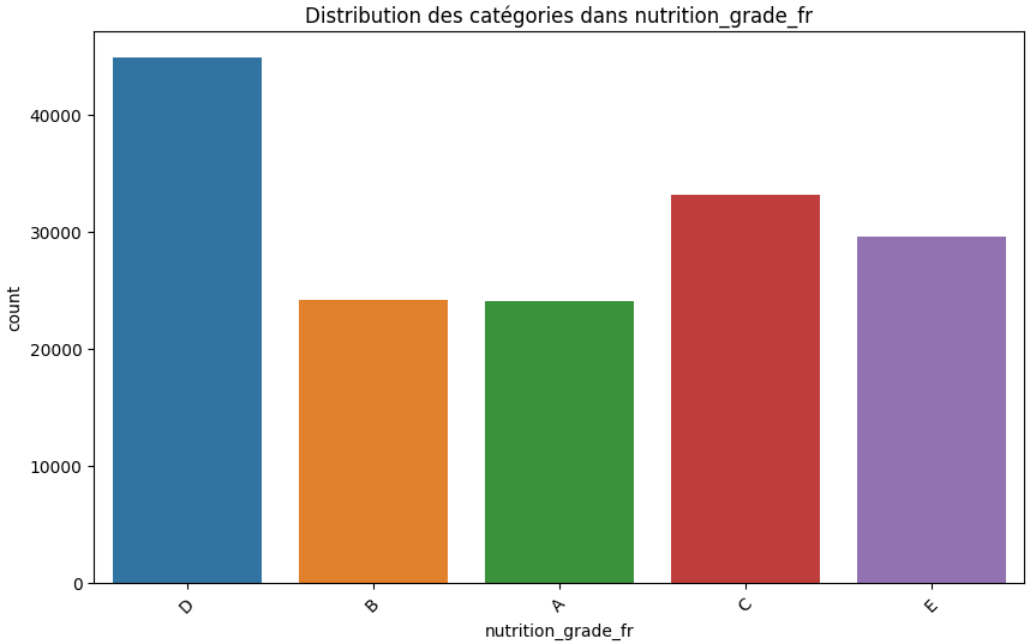






3. ANALYSE EXPLORATOIRE UNIVARIEE





Unicité des données

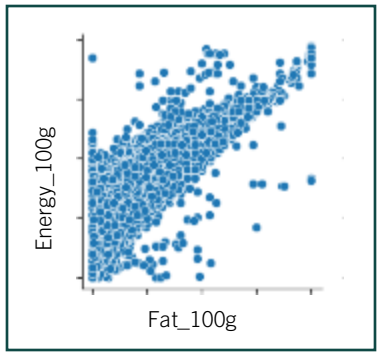
PNNS1 & Notes nutritionnelles

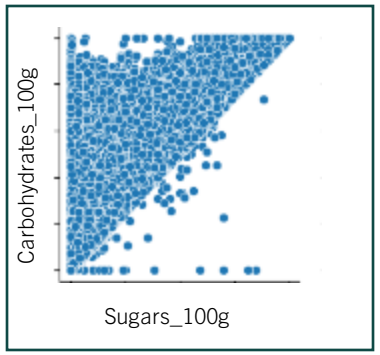
Prédominance : 'Sugary snacks' et D

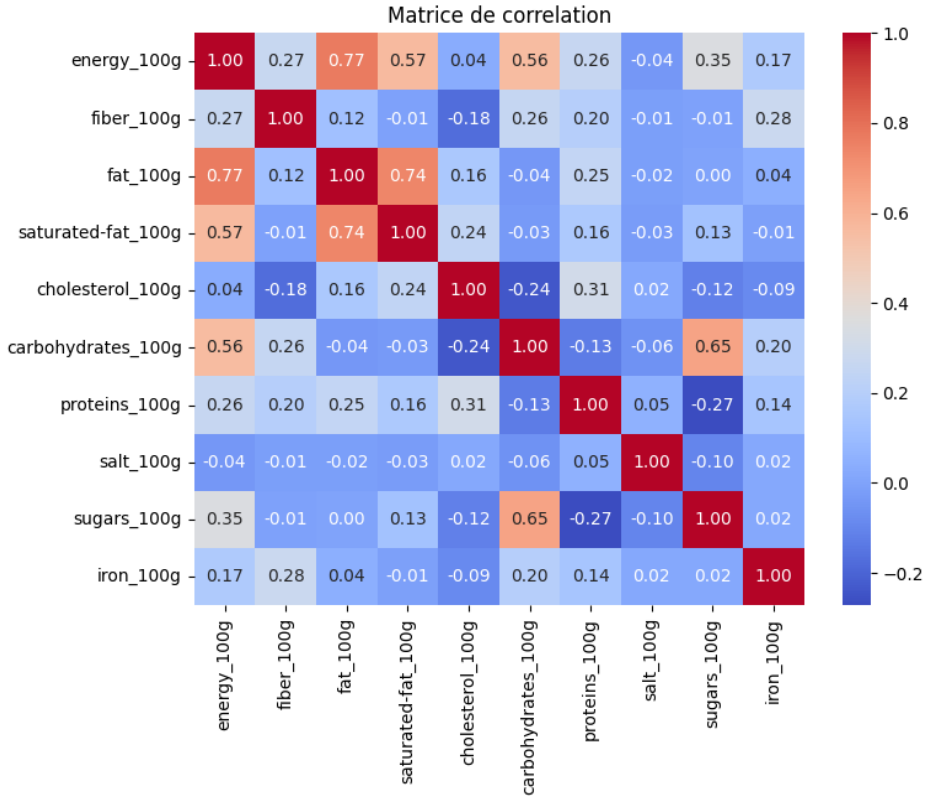
Milieu de gamme

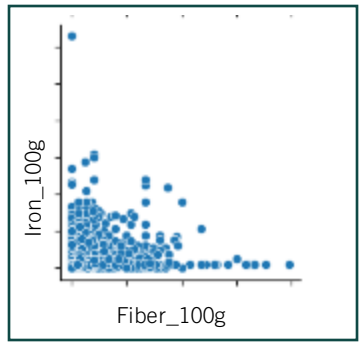
Moins représentées : 'Beverages' , 'Fruits and vegetables' et A,B

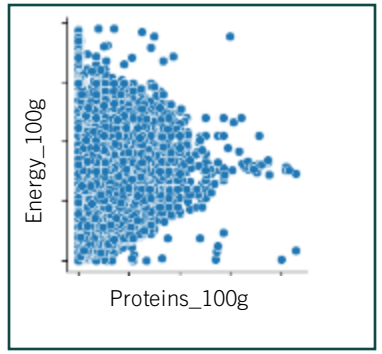
3. ANALYSE EXPLORATOIRE UNIVARIEE







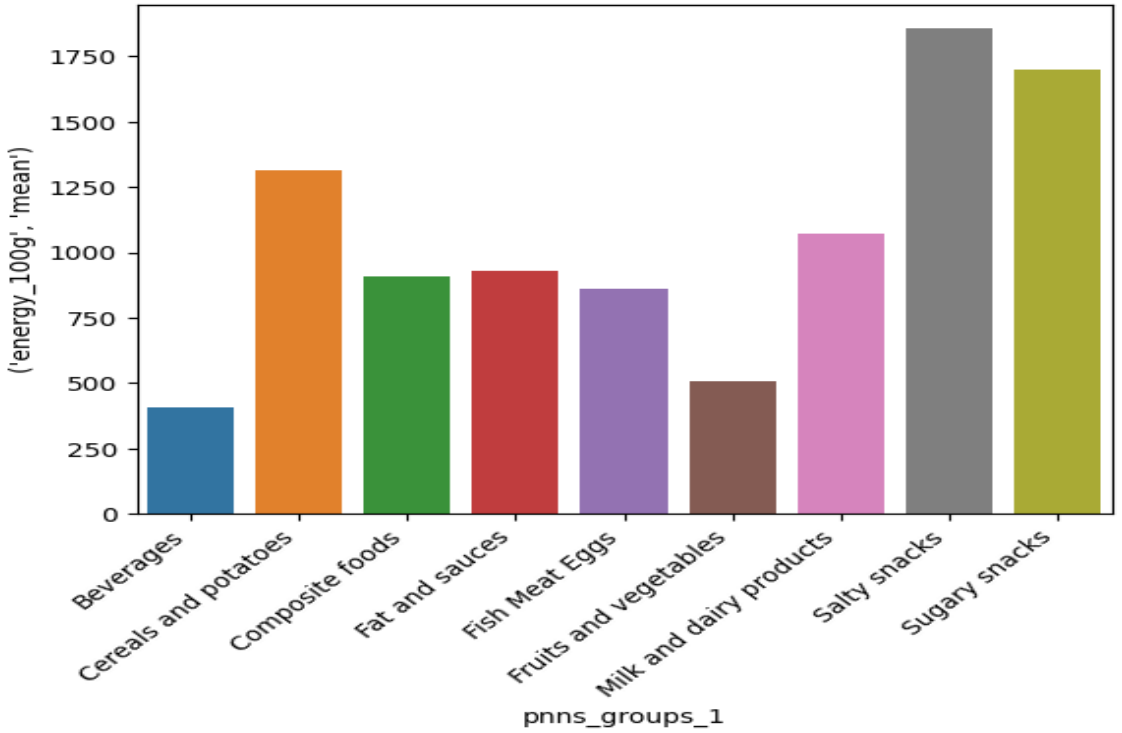


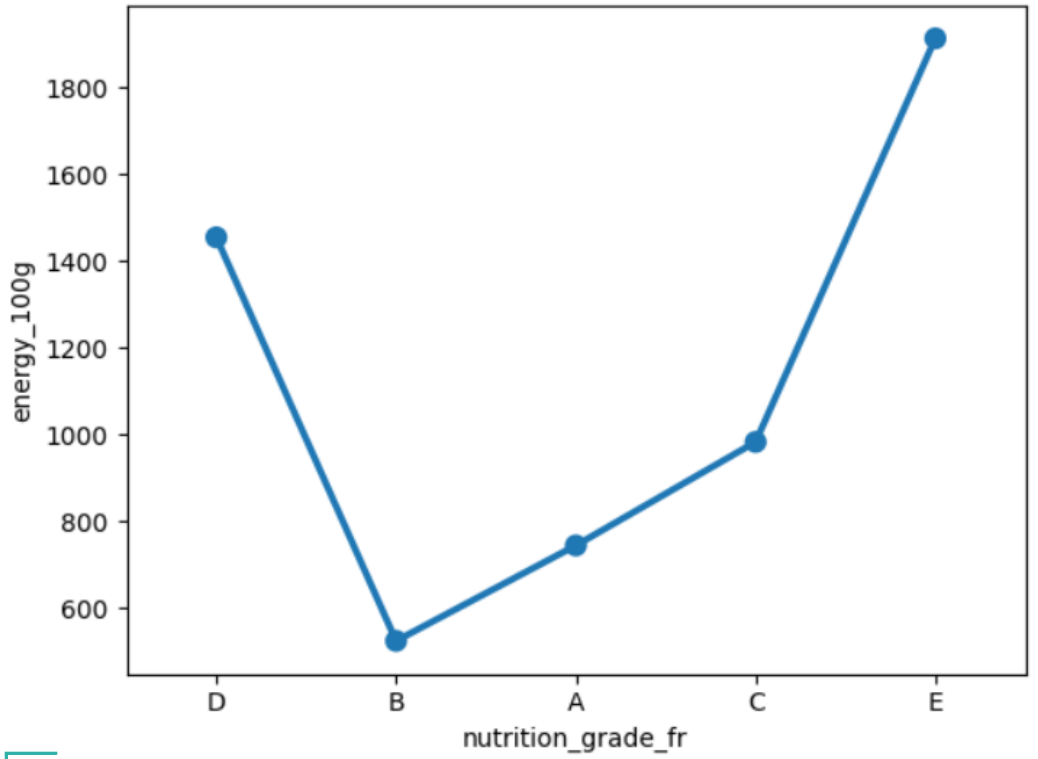


- **Énergie et Nutriments :** Corrélation positive entre l'énergie et les graisses, les acides gras saturés, et les sucres
- **Graisses et Acides Gras Saturés :** Forte corrélation positive
- **Sucres et Glucides :** Relation positive

- **Protéines** : Moins fortement corrélées à l'énergie par rapport aux graisses
- **Sel :** Pas de corrélation claire avec d'autres nutriments, indépendance relative.
- **Cholestérol, Fibres et Fer :** Pas de relations significatives ou fortes

3. ANALYSE EXPLORATOIRE BIVARIEE





"Fat and sauces" et "Sugary snacks" ont les valeurs énergétiques moyennes les plus élevées.

"Beverages" a la valeur énergétique moyenne la plus basse.

Teneur Énergétique par Catégories PNNS

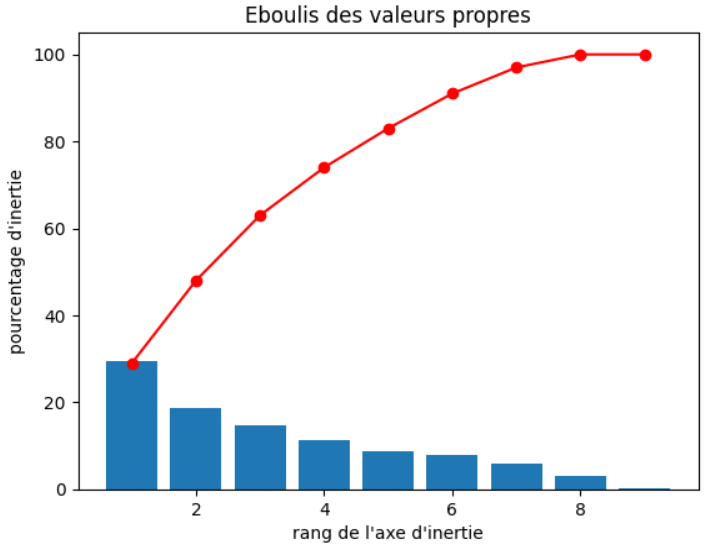
Les notes C et E ont des teneurs en énergie plus élevées

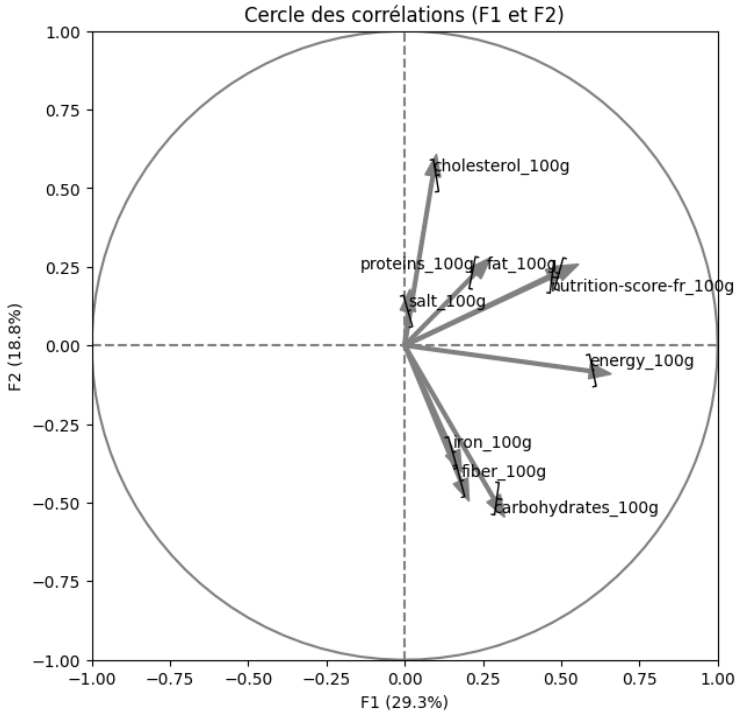
Les produits notés A présentent la teneur en énergie la plus basse

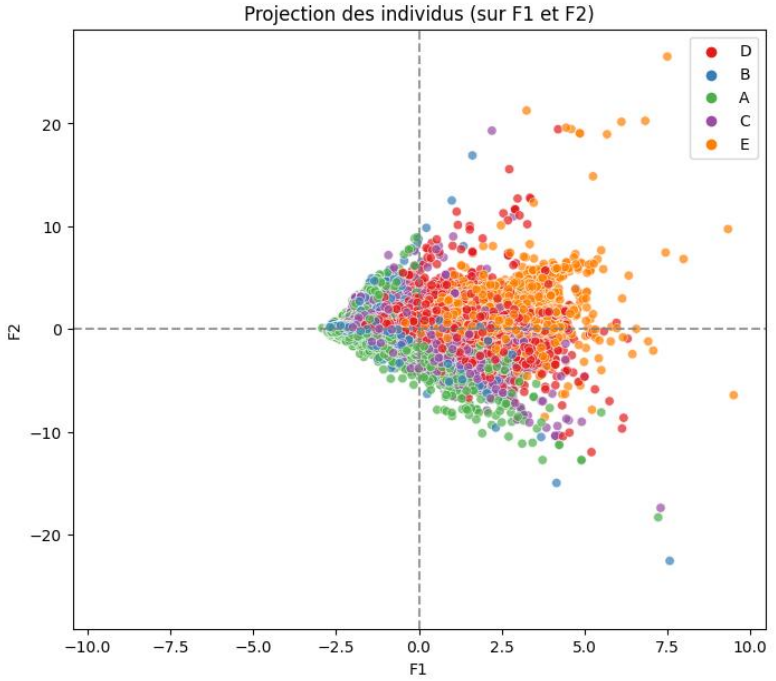
Teneur Énergétique par Nutri-Score

3. ANALYSE EXPLORATOIRE BIVARIEE

ANALYSE EN COMPOSANTES PRINCIPALES







Variables indicateurs clés

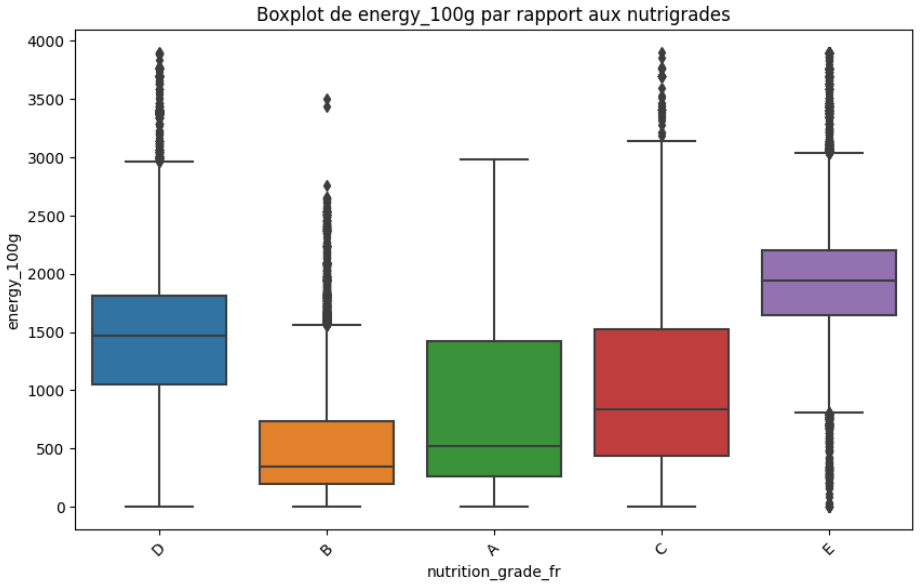
- Corrélations identifiées entre les graisses, l'énergie et le score nutritionnel

Dispersion sur les plans F1 et F2

- Variation de la qualité nutritionnelle avec séparation claire

3. ANALYSE EXPLORATOIRE MULTIVARIEE

TEST ANOVA



	Variable	F-statistic	p-value
0	energy_100g	387.992919	0.000000e+00
1	fiber_100g	101.176436	0.000000e+00
2	fat_100g	368.898356	0.000000e+00
3	saturated_fat_100g	263.342730	0.000000e+00
4	cholesterol_100g	222.909401	0.000000e+00
5	carbohydrates_100g	222.303407	0.000000e+00
6	proteins_100g	355.020250	0.000000e+00
7	salt_100g	50.968507	1.482197e-321
8	sugars_100g	482.967092	0.000000e+00
9	iron_100g	26.888951	1.095228e-159
10	nutrition_score_fr_100g	248.520798	0.000000e+00

Test de normalité → Test d'homogénéité → ANOVA → Test de Kruskal-Wallis

3. ANALYSE EXPLORATOIRE MULTIVARIEE

CONCLUSION

- Des **associations significatives** ont été identifiées entre les catégories alimentaires, les notes nutritionnelles, et les valeurs nutritionnelles comme les graisses, les sucres, l'énergie et les graisses saturées.

- **Limitation de l'étude** : Absence de données sur la teneur en fruits et légumes, un élément clé du Nutri-Score.

- **Potentiel pour l'auto-complétions** : Les modèles ACP et ANOVA suggèrent la viabilité d'un système d'auto-complétions pour la base de données nutritionnelle.

- **Développement d'algorithmes de suggestion** : Possibilité de prédire des valeurs nutritionnelles et de suggérer des corrections en temps réel.

- **Évolution réglementaire européenne** : Les améliorations prévues du Nutri-Score pourraient affecter les critères et seuils d'évaluation nutritionnelle.

MERCI POUR VOTRE ATTENTION