Santé publique France

OPEN FOOD FACTS
l'information alimentaire ouverte

AI Engineer – P03
OpenClassrooms

Préparez des données pour un organisme de santé publique

Anthony VIDRAGO – Aout 2024

Introduction

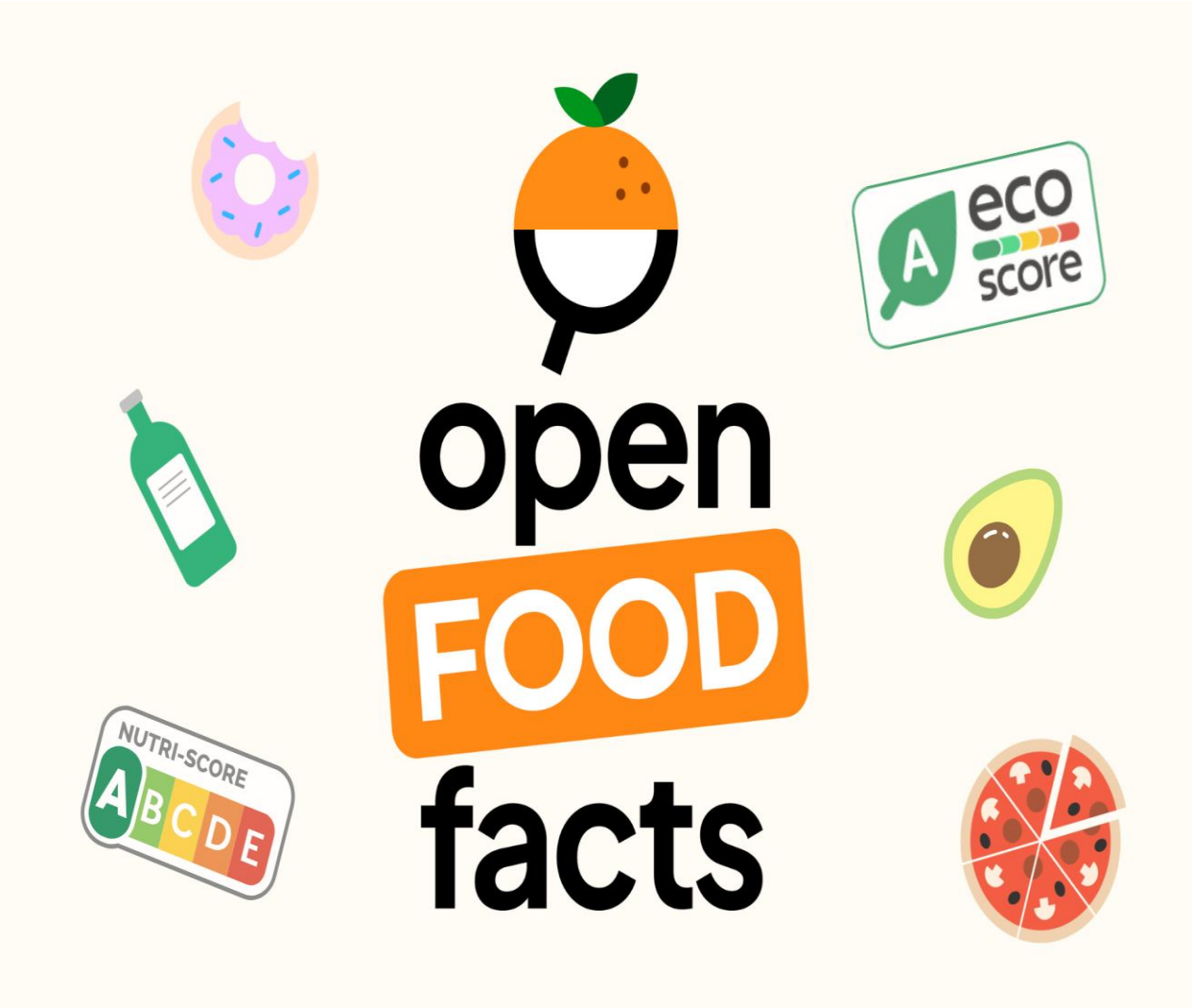
CONTEXTE DU PROJET

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 de produits.

Aujourd'hui, pour ajouter un produit à la base de données d'Open Food Facts, 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.

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

Nous sommes missionnés sur le projet de nettoyage et exploration des données, afin de déterminer la faisabilité de cette idée d'application.



Aujourd’hui

Demain

Sommaire

1er Notebook => Nettoyage

2ème Notebook => Analyse

1er Notebook

Le Nettoyage

```python
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
```

```python
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
```

```python
data.head()
```

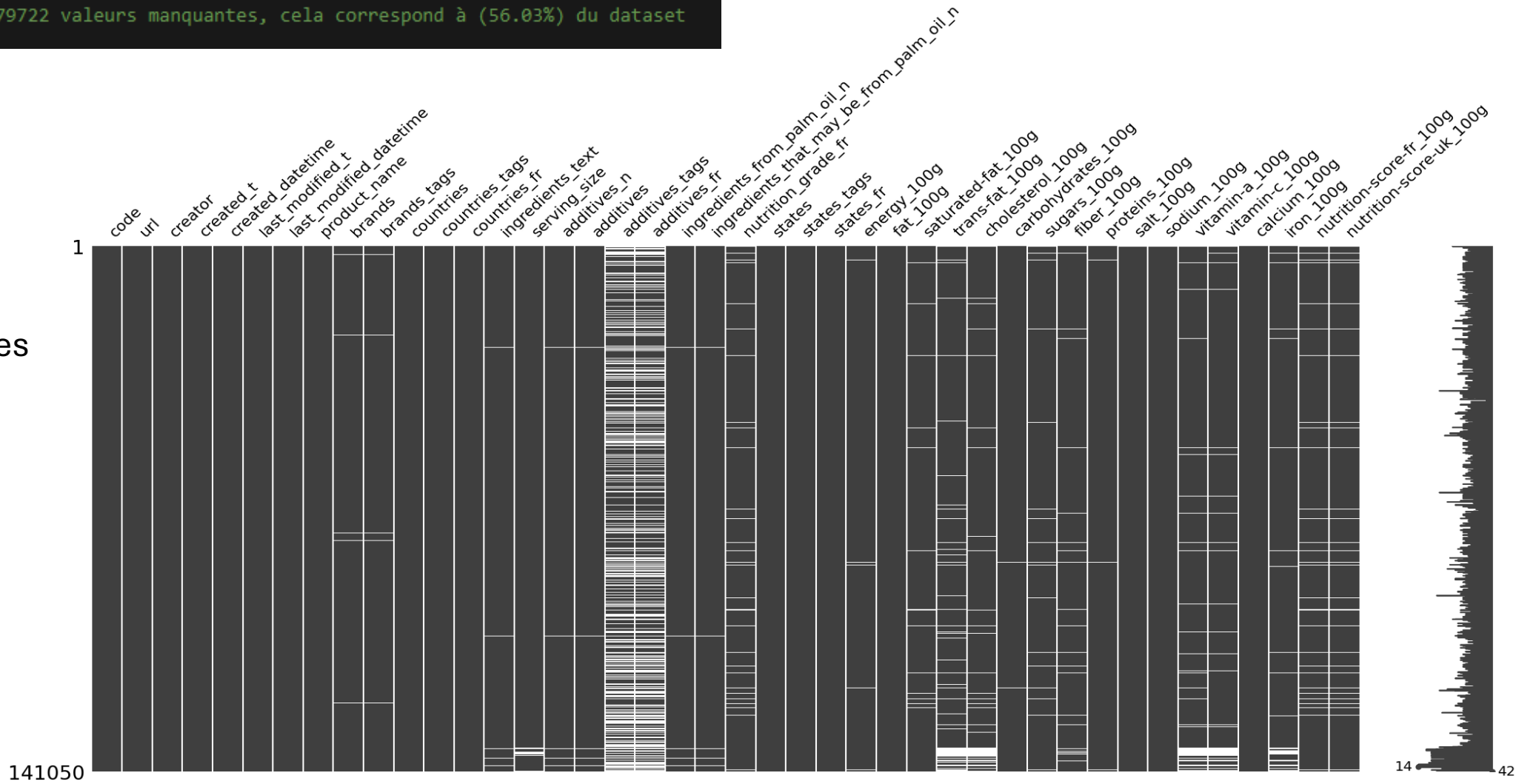
Python

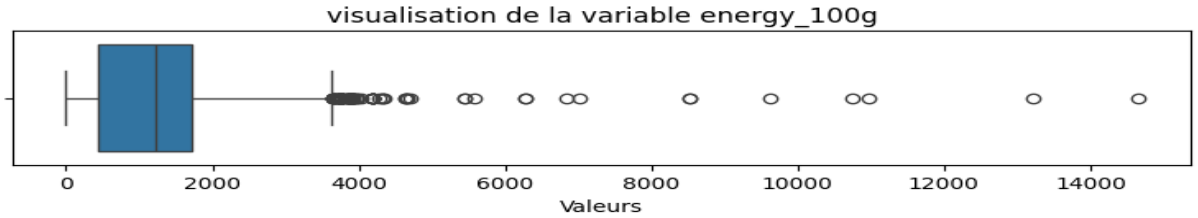
	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	quantity	...	ph_100g	fruits-vegetables-nuts_100g	collagen-meat-protein-ratio_100g	cocoa_100g	chlorophyl_100g	carbon-footprint_100g	nutrition-score-fr_100g	nutrition-score-uk_100g	glycemic-index_100g	hardne
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	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	14.0	14.0	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	0.0	0.0	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	12.0	12.0	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	NaN	NaN	NaN	

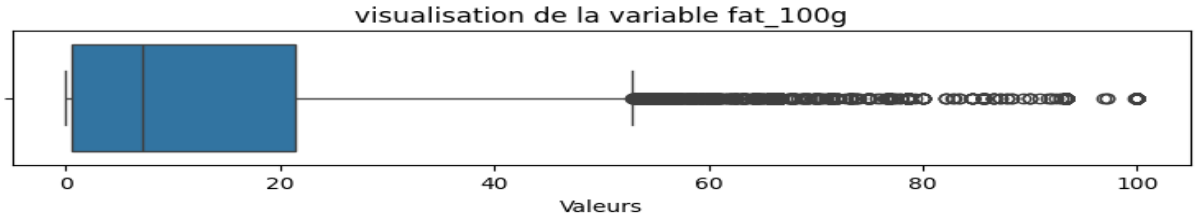
5 rows × 162 columns

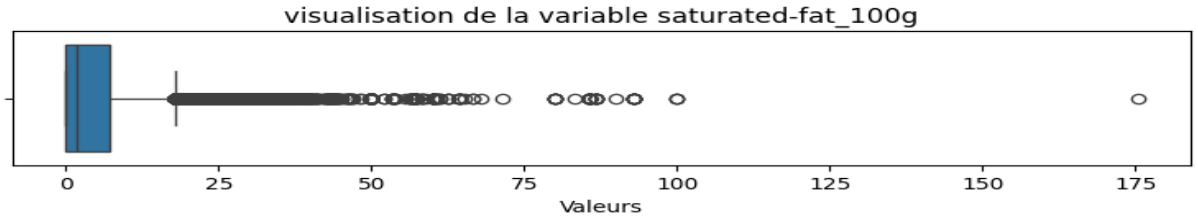
```
####################################################################
# 1.2 : Choix de la cible
####################################################################

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

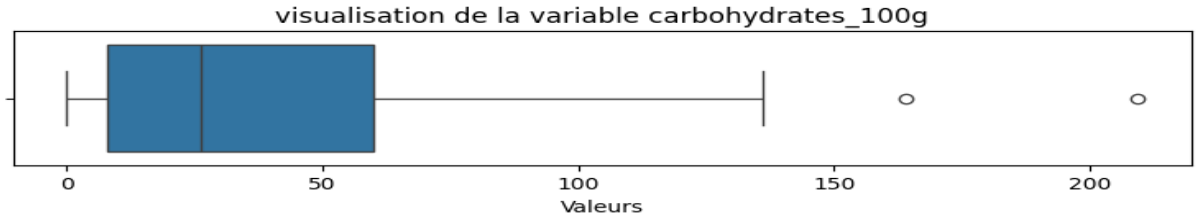
Les colonnes avec plus de 60% de valeurs manquantes ont été supprimé

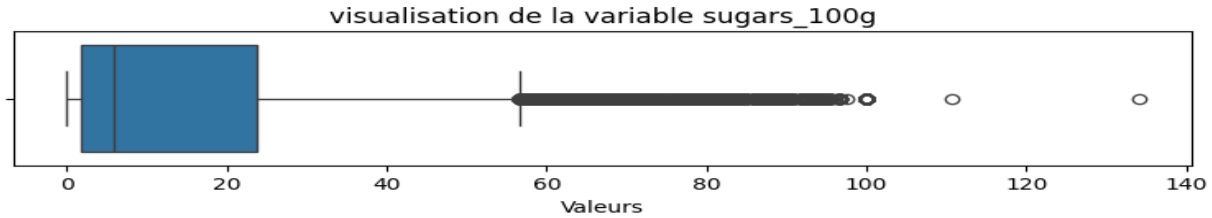


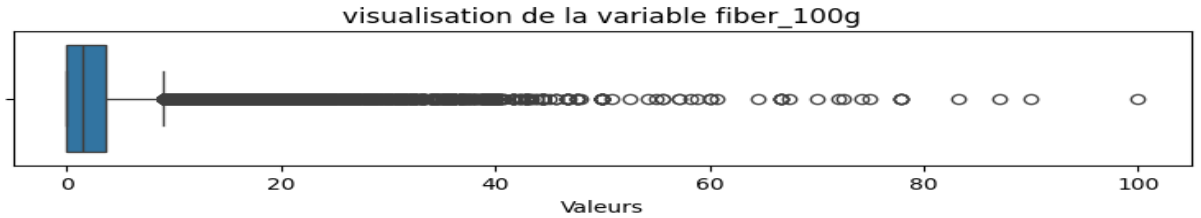

visualisation de la variable energy_100g
0
2000
4000
6000
8000
10000
12000
14000
Valeurs

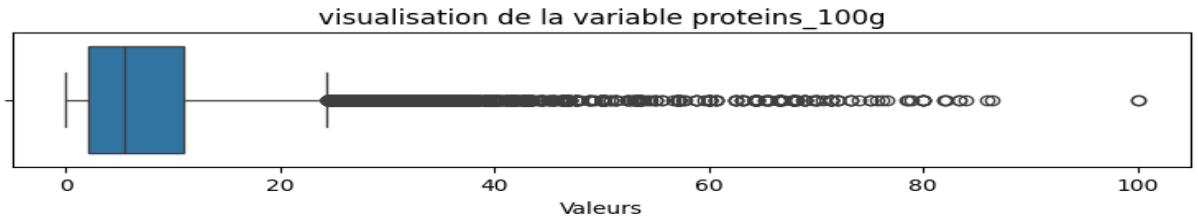

visualisation de la variable fat_100g
0
20
40
60
80
100
Valeurs

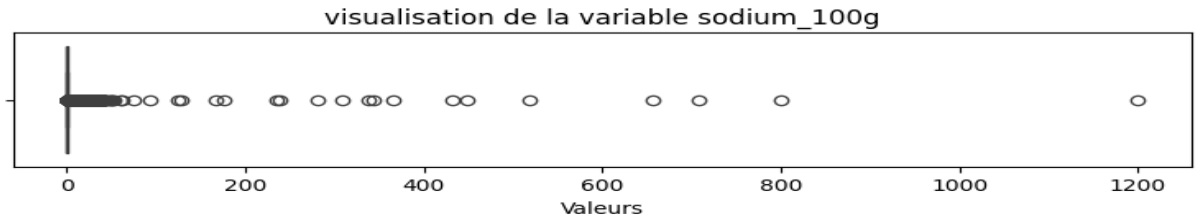

visualisation de la variable saturated-fat_100g
0
25
50
75
100
125
150
175
Valeurs

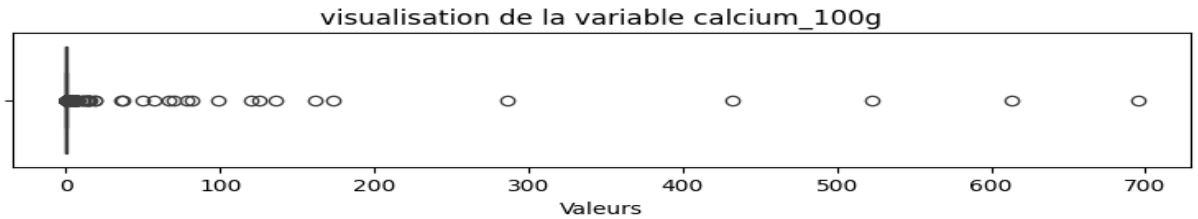

visualisation de la variable carbohydrates_100g
0
50
100
150
200
Valeurs

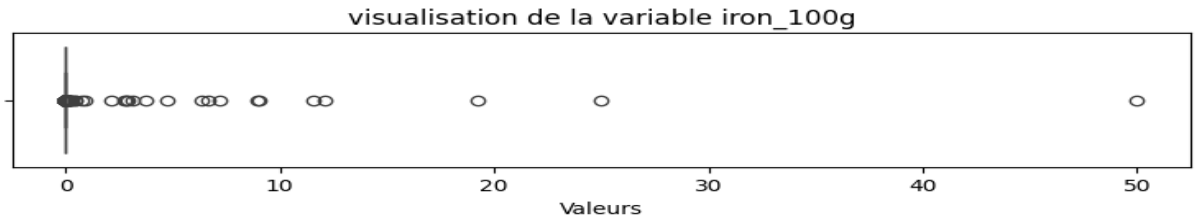

visualisation de la variable sugars_100g
0
20
40
60
80
100
120
140
Valeurs

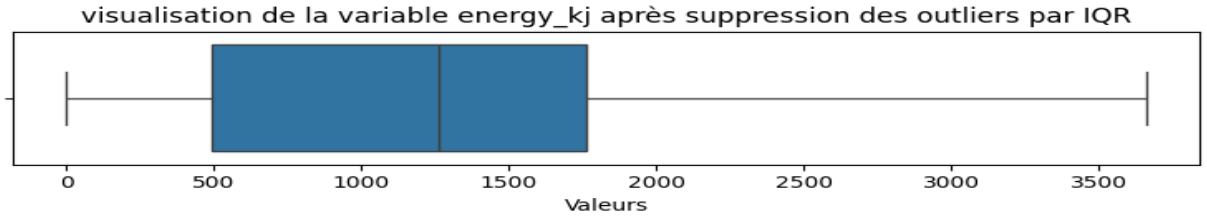

visualisation de la variable fiber_100g
0
20
40
60
80
100
Valeurs

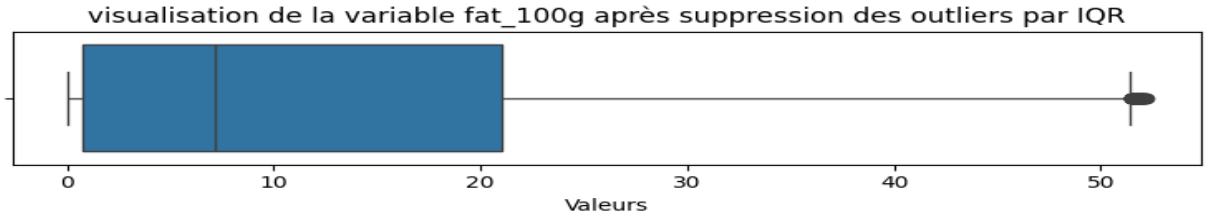

visualisation de la variable proteins_100g
0
20
40
60
80
100
Valeurs

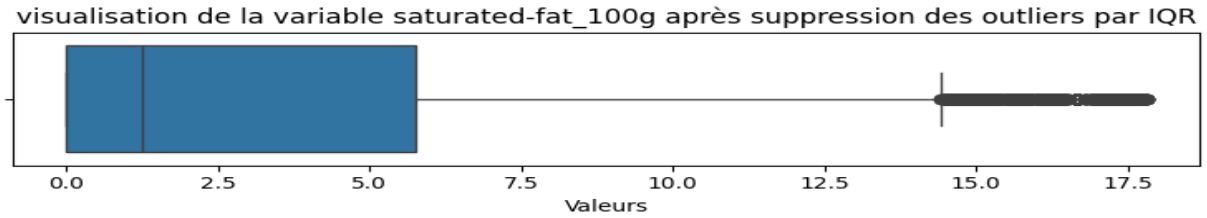

visualisation de la variable sodium_100g
0
200
400
600
800
1000
1200
Valeurs

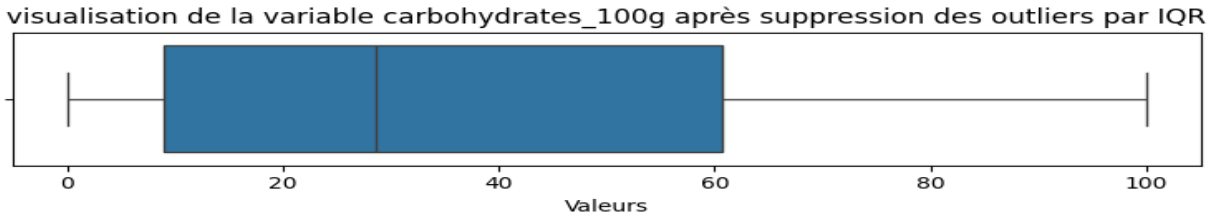

visualisation de la variable calcium_100g
0
100
200
300
400
500
600
700
Valeurs

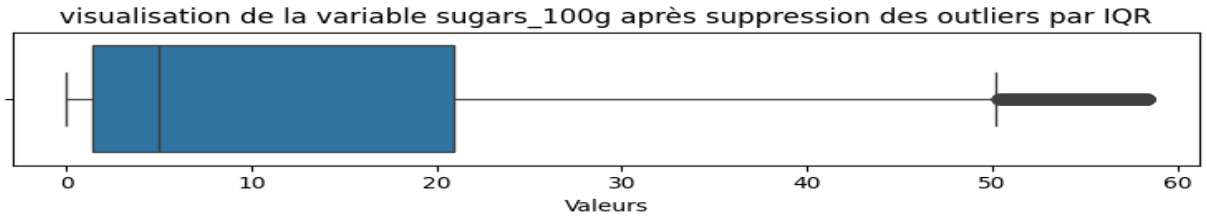

visualisation de la variable iron_100g
0
10
20
30
40
50
Valeurs

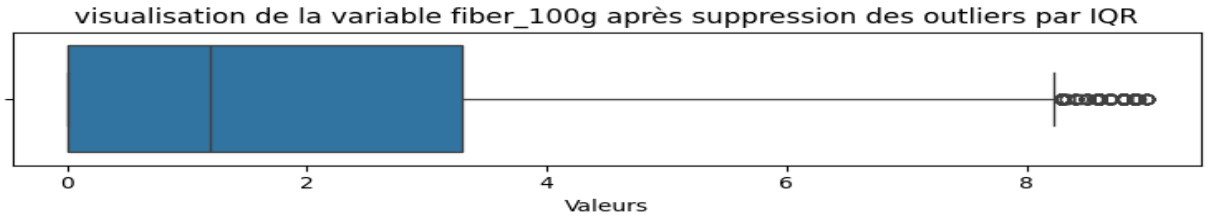

visualisation de la variable energy_kj après suppression des outliers par IQR
0
500
1000
1500
2000
2500
3000
3500
Valeurs

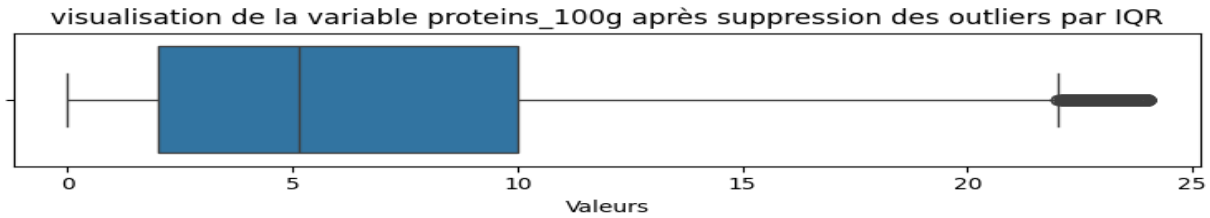

visualisation de la variable fat_100g après suppression des outliers par IQR
0
10
20
30
40
50
Valeurs

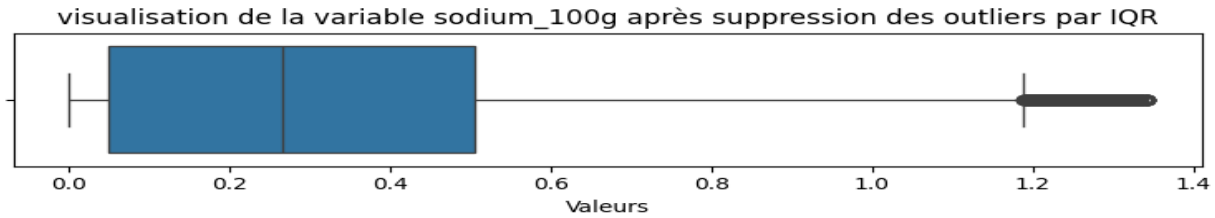

visualisation de la variable saturated-fat_100g après suppression des outliers par IQR
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Valeurs

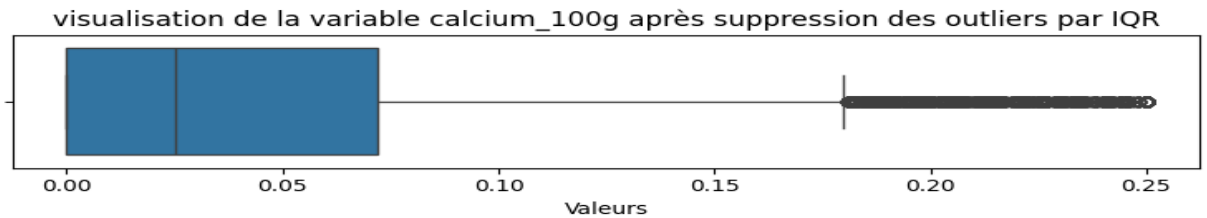

visualisation de la variable carbohydrates_100g après suppression des outliers par IQR
0
20
40
60
80
100
Valeurs


visualisation de la variable sugars_100g après suppression des outliers par IQR
0
10
20
30
40
50
60
Valeurs


visualisation de la variable fiber_100g après suppression des outliers par IQR
0
2
4
6
8
Valeurs


visualisation de la variable proteins_100g après suppression des outliers par IQR
0
5
10
15
20
25
Valeurs


visualisation de la variable sodium_100g après suppression des outliers par IQR
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Valeurs


visualisation de la variable calcium_100g après suppression des outliers par IQR
0.00
0.05
0.10
0.15
0.20
0.25
Valeurs

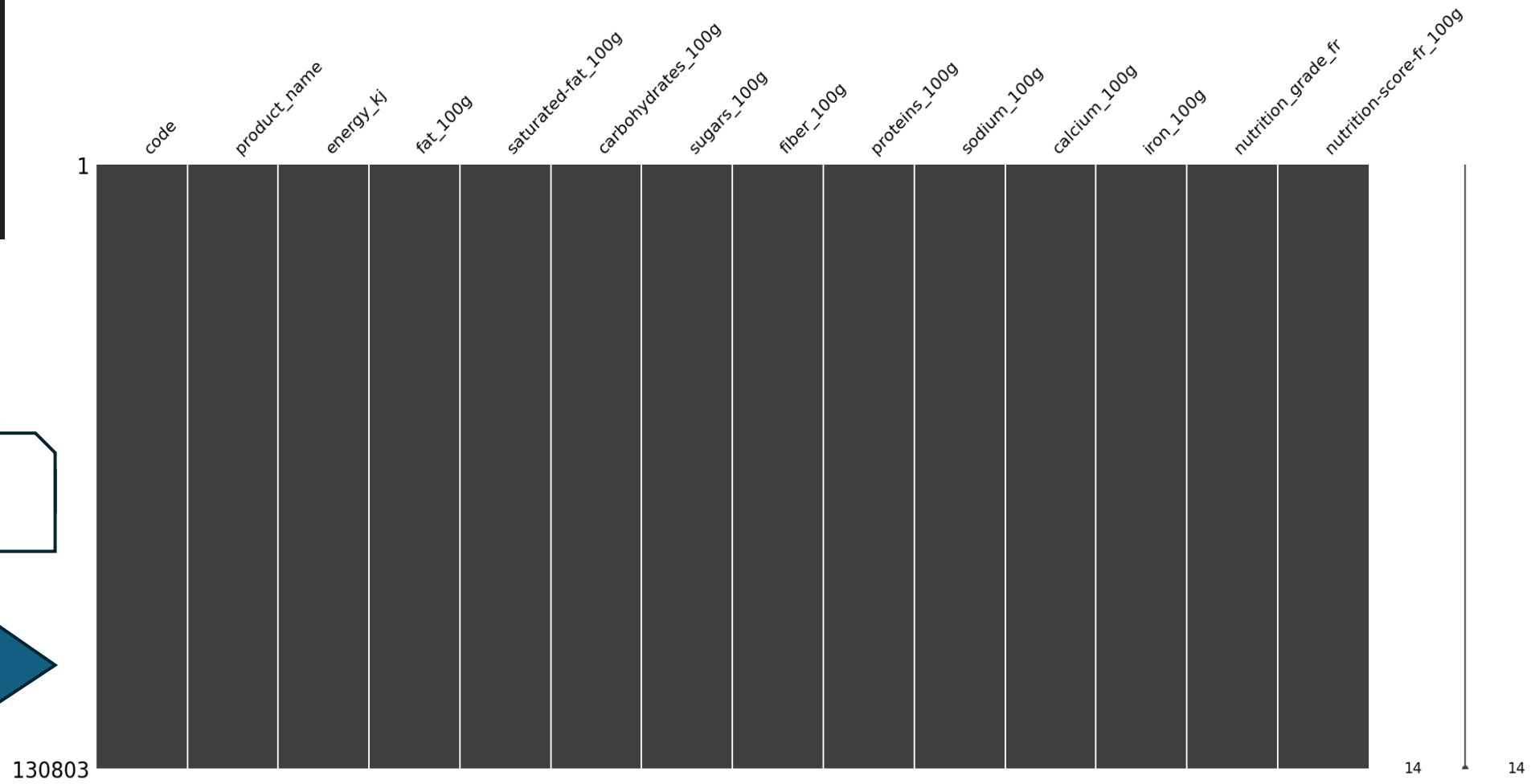

visualisation de la variable iron_100g après suppression des outliers par IQR
0.000
0.001
0.002
0.003
0.004
0.005
0.006
Valeurs

```python
num_features = data.select_dtypes(include = ["number"]).columns
num_fill_rate = data[num_features].notnull().mean().sort_values(ascending = False)

print("Taux de remplissage des variables numériques (en %) : ")
print(round(num_fill_rate * 100, 2))
```

✓ 0.0s

```
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
```

Imputation avec KNN



Analyse exploratoire des données

Réduction du dataset et sélection de la cible

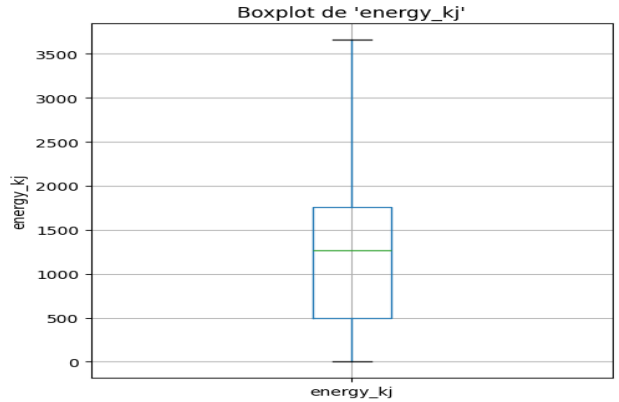
Gestion des valeurs : Les outliers (Avant)

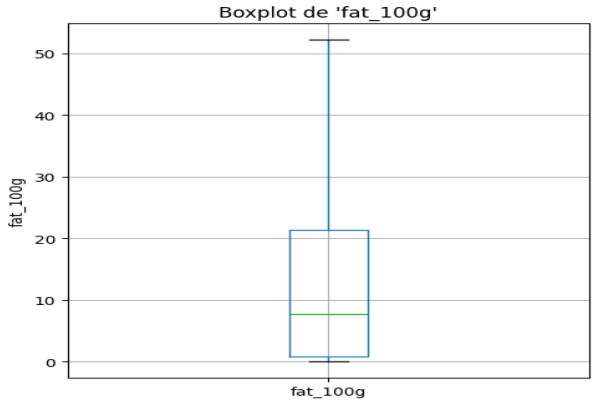
Gestion des valeurs : Les outliers (Après)

Gestion des valeurs : Imputation KNN

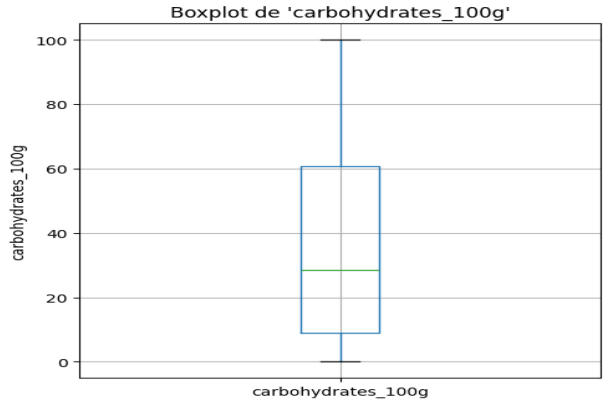
2ème Notebook

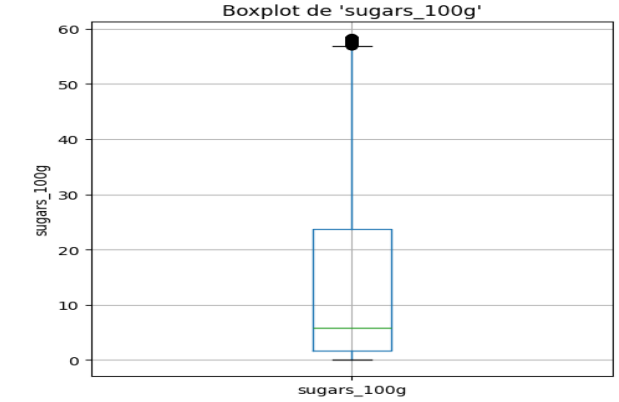
L’analyse

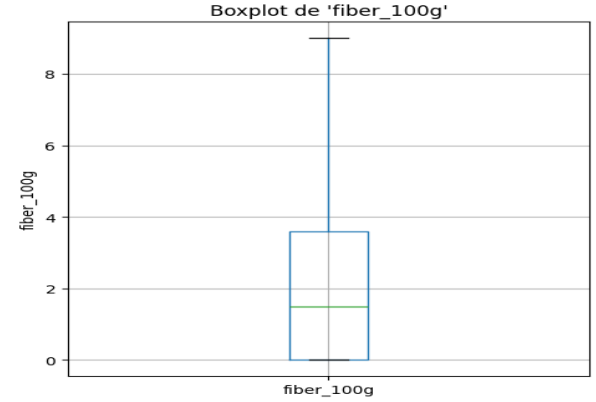

Boxplot de 'energy_kj'
energy_kj
0
500
1000
1500
2000
2500
3000
3500
energy_kj

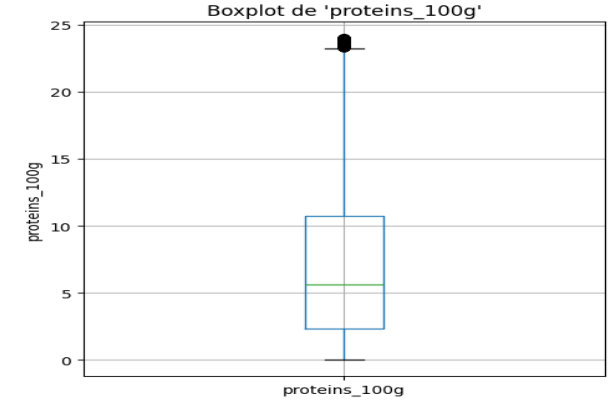

Boxplot de 'fat_100g'
fat_100g
0
10
20
30
40
50
fat_100g

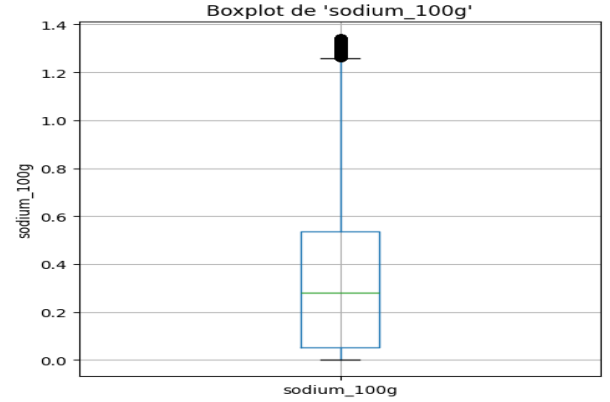

Boxplot de 'saturated-fat_100g'
saturated-fat_100g
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
saturated-fat_100g

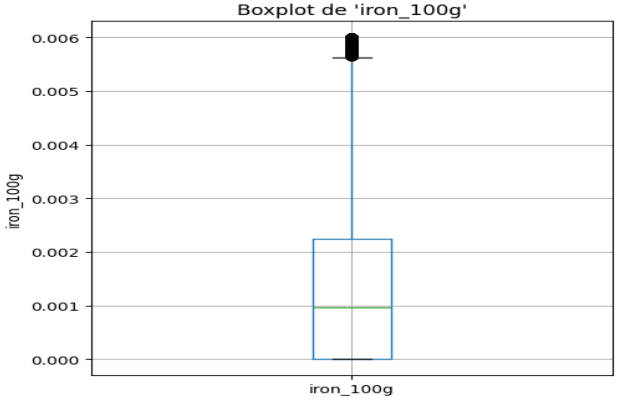

Boxplot de 'carbohydrates_100g'
carbohydrates_100g
0
20
40
60
80
100
carbohydrates_100g

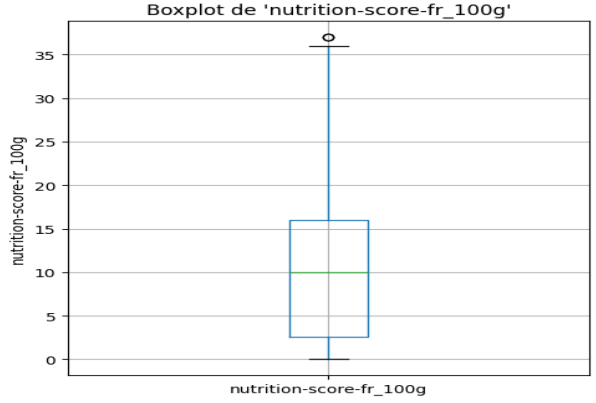

Boxplot de 'sugars_100g'
sugars_100g
0
10
20
30
40
50
60
sugars_100g

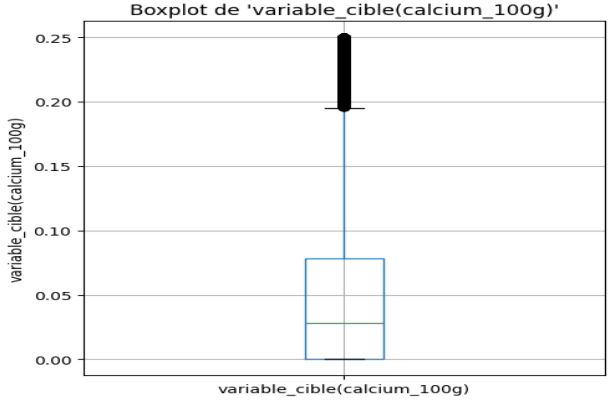

Boxplot de 'fiber_100g'
fiber_100g
0
2
4
6
8
fiber_100g


Boxplot de 'proteins_100g'
proteins_100g
0
5
10
15
20
25
proteins_100g

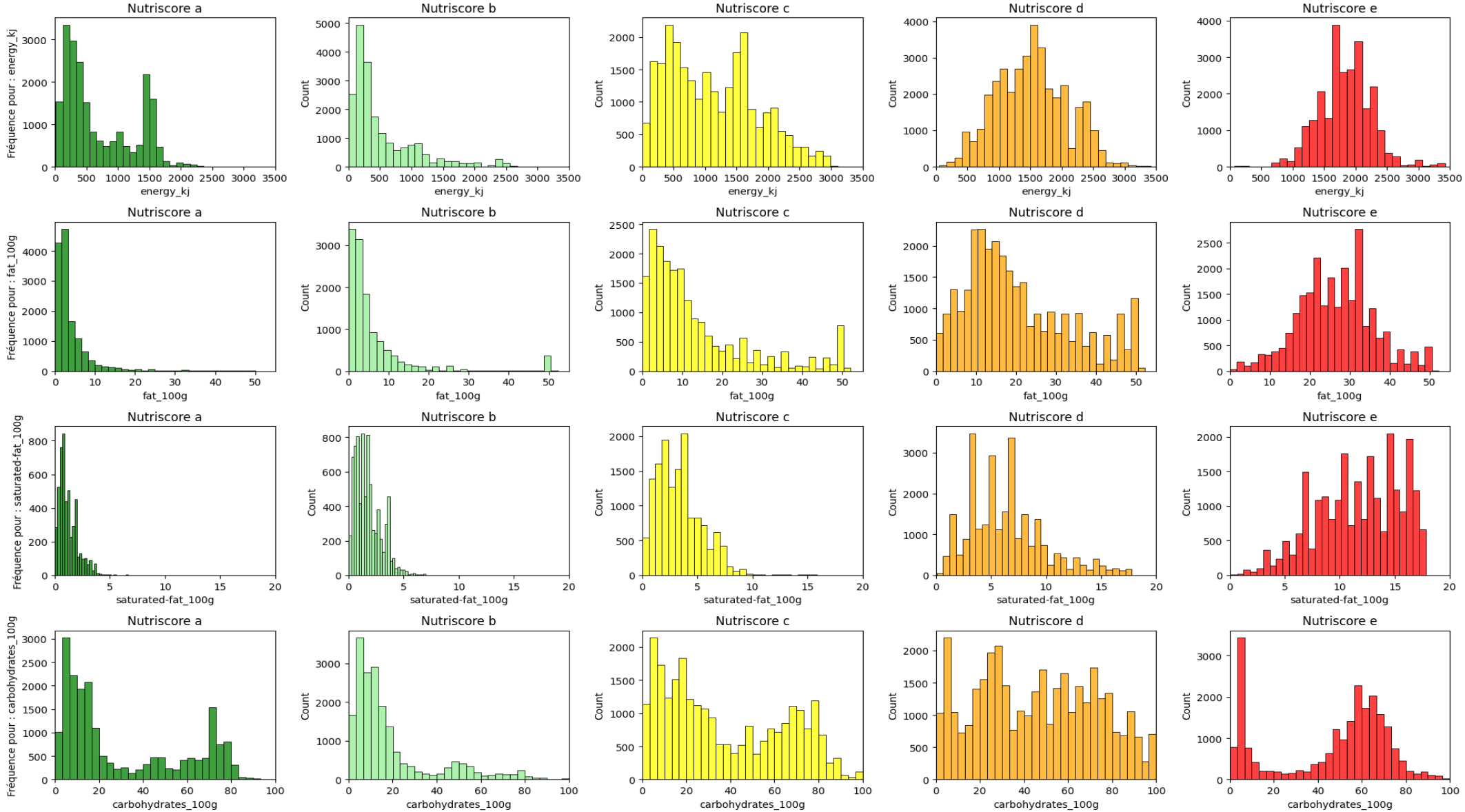

Boxplot de 'sodium_100g'
sodium_100g
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
sodium_100g


Boxplot de 'iron_100g'
iron_100g
0.000
0.001
0.002
0.003
0.004
0.005
0.006
iron_100g


Boxplot de 'nutrition-score-fr_100g'
nutrition-score-fr_100g
0
5
10
15
20
25
30
35
nutrition-score-fr_100g


Boxplot de 'variable_cible(calcium_100g)'
variable_cible(calcium_100g)
0.00
0.05
0.10
0.15
0.20
0.25
variable_cible(calcium_100g)

Distribution des nutriments par nutriscore

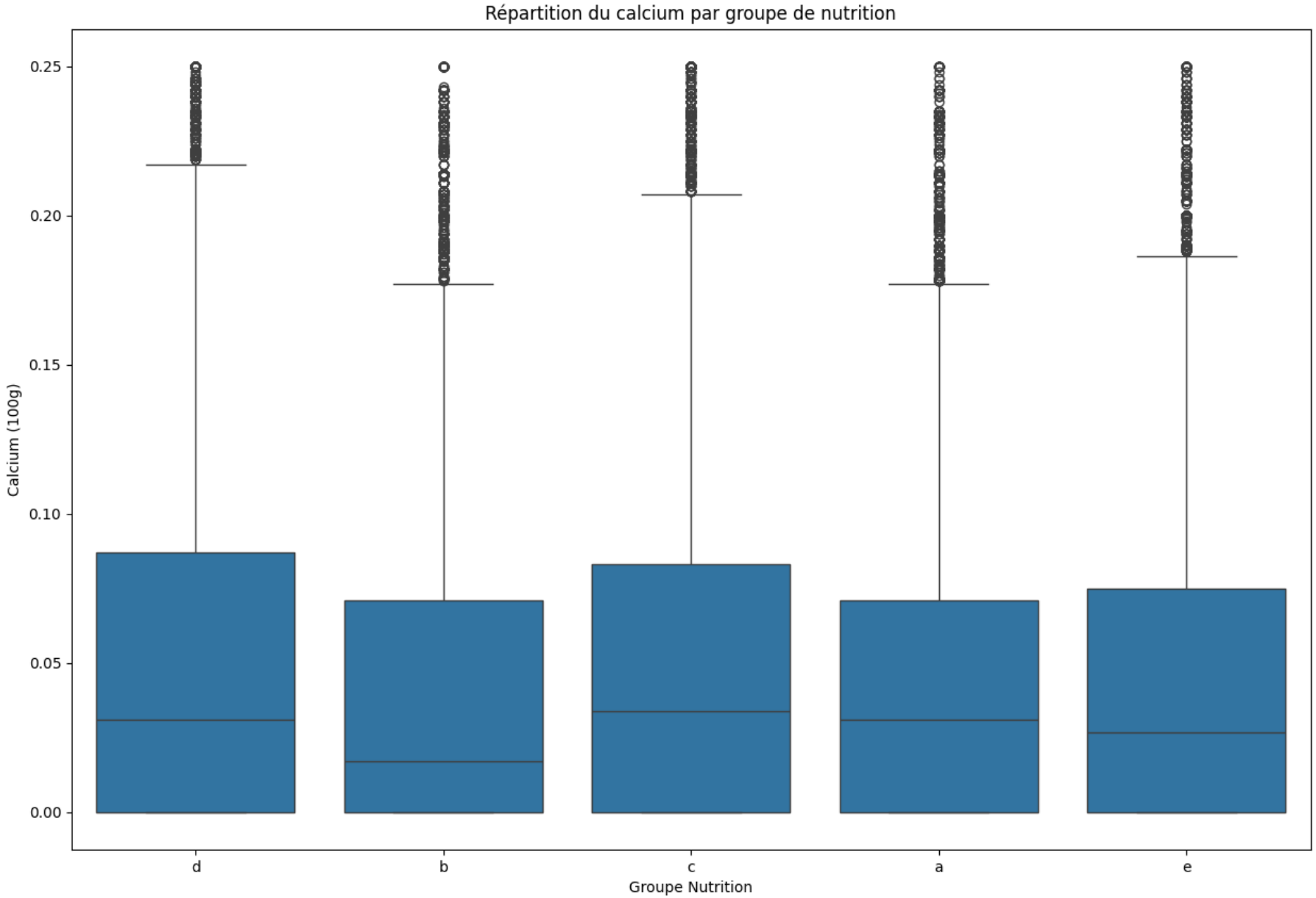


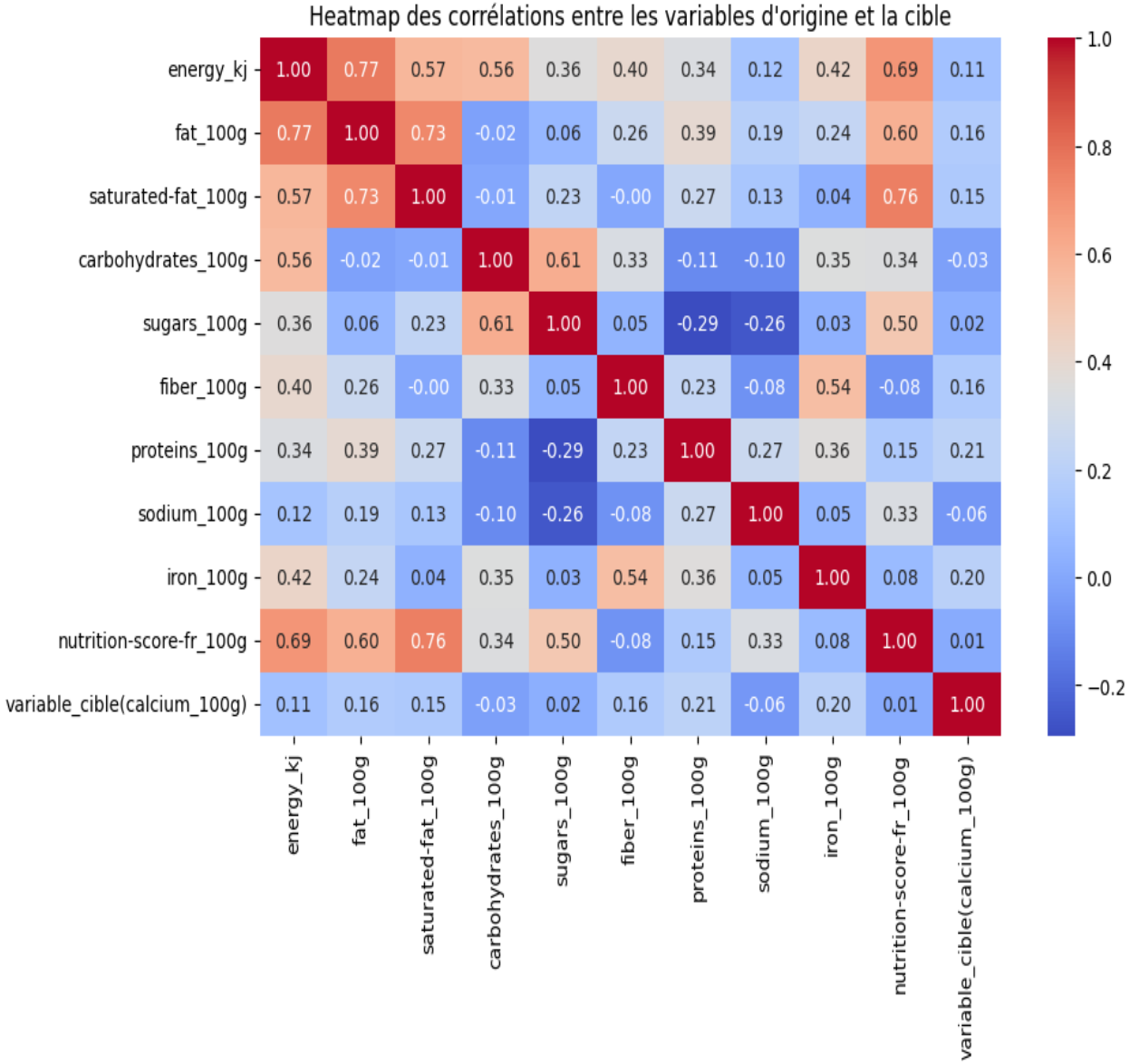
Analyses bivariées

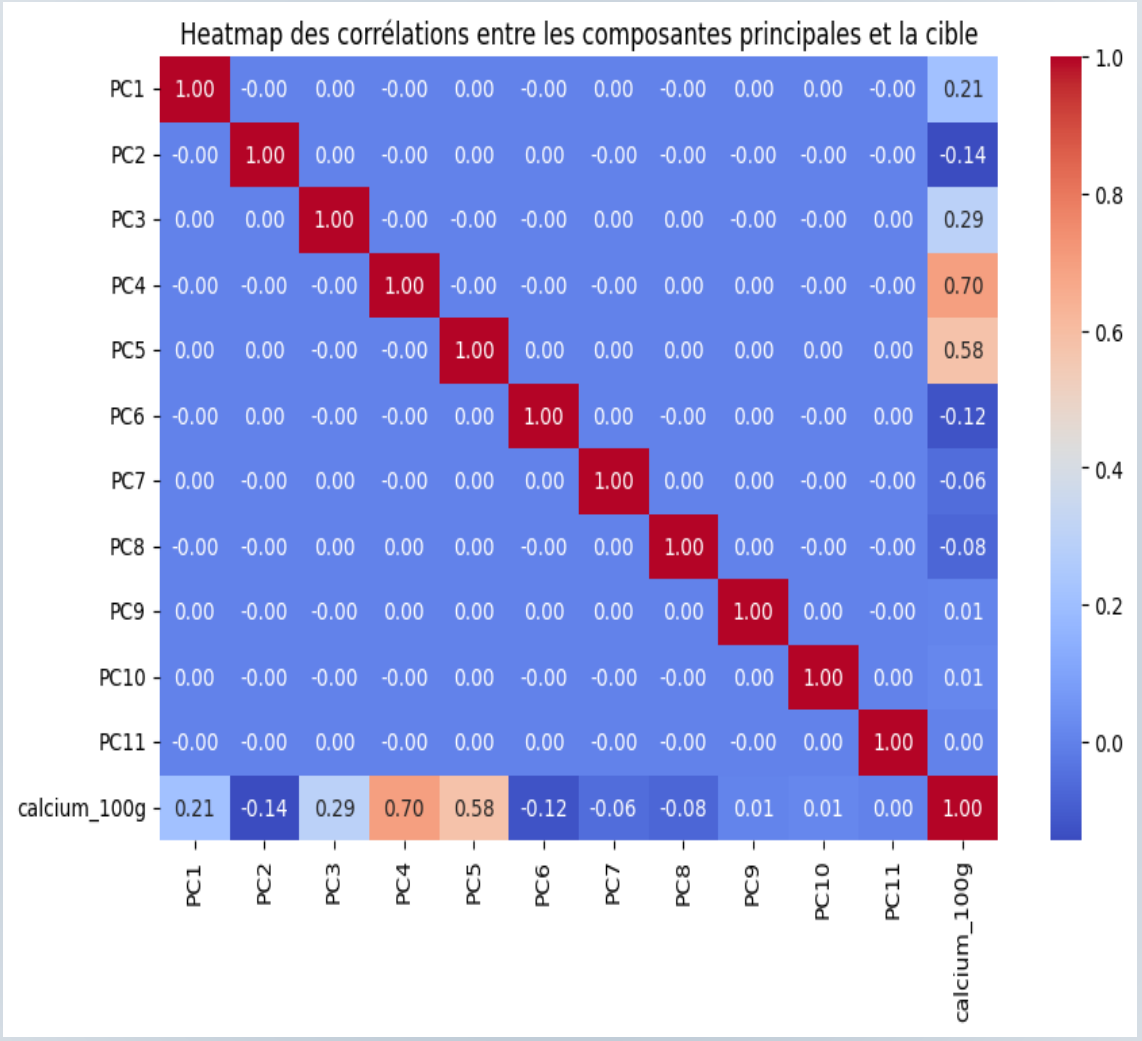
Pour la variable catégorielle

Test Shapiro-Wilk
W-statistic = 0.0003906209121684600 7
p-value = 9.436948635914407e-211

Test Kruskal-Wallis
Statistique = 175.8750825499985
p-value = 5.731868141181024e-37




Heatmap des corrélations entre les variables d'origine et la cible
energy_kj 1.00 0.77 0.57 0.56 0.36 0.40 0.34 0.12 0.42 0.69 0.11
fat_100g 0.77 1.00 0.73 -0.02 0.06 0.26 0.39 0.19 0.24 0.60 0.16
saturated-fat_100g 0.57 0.73 1.00 -0.01 0.23 -0.00 0.27 0.13 0.04 0.76 0.15
carbohydrates_100g 0.56 -0.02 -0.01 1.00 0.61 0.33 -0.11 -0.10 0.35 0.34 -0.03
sugars_100g 0.36 0.06 0.23 0.61 1.00 0.05 -0.29 -0.26 0.03 0.50 0.02
fiber_100g 0.40 0.26 -0.00 0.33 0.05 1.00 0.23 -0.08 0.54 -0.08 0.16
proteins_100g 0.34 0.39 0.27 -0.11 -0.29 0.23 1.00 0.27 0.36 0.15 0.21
sodium_100g 0.12 0.19 0.13 -0.10 -0.26 -0.08 0.27 1.00 0.05 0.33 -0.06
iron_100g 0.42 0.24 0.04 0.35 0.03 0.54 0.36 0.05 1.00 0.08 0.20
nutrition-score-fr_100g 0.69 0.60 0.76 0.34 0.50 -0.08 0.15 0.33 0.08 1.00 0.01
variable_cible(calcium_100g) 0.11 0.16 0.15 -0.03 0.02 0.16 0.21 -0.06 0.20 0.01 1.00
energy_kj
fat_100g
saturated-fat_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
sodium_100g
iron_100g
nutrition-score-fr_100g
variable_cible(calcium_100g)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Heatmap des corrélations entre les composantes principales et la cible
PC1 1.00 -0.00 0.00 -0.00 0.00 -0.00 0.00 -0.00 0.00 0.00 -0.00 0.21
PC2 -0.00 1.00 0.00 -0.00 0.00 0.00 -0.00 -0.00 -0.00 -0.00 -0.00 -0.14
PC3 0.00 0.00 1.00 -0.00 -0.00 -0.00 0.00 0.00 -0.00 -0.00 0.00 0.29
PC4 -0.00 -0.00 -0.00 1.00 -0.00 -0.00 -0.00 0.00 0.00 -0.00 -0.00 0.70
PC5 0.00 0.00 -0.00 -0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.58
PC6 -0.00 0.00 -0.00 -0.00 0.00 1.00 0.00 -0.00 0.00 -0.00 0.00 -0.12
PC7 0.00 -0.00 0.00 -0.00 0.00 0.00 1.00 0.00 0.00 -0.00 -0.00 -0.06
PC8 -0.00 -0.00 0.00 0.00 0.00 -0.00 0.00 1.00 0.00 -0.00 -0.00 -0.08
PC9 0.00 -0.00 -0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 -0.00 0.01
PC10 0.00 -0.00 -0.00 -0.00 0.00 -0.00 -0.00 -0.00 0.00 1.00 0.00 0.01
PC11 -0.00 -0.00 0.00 -0.00 0.00 0.00 -0.00 -0.00 -0.00 0.00 1.00 0.00
calcium_100g 0.21 -0.14 0.29 0.70 0.58 -0.12 -0.06 -0.08 0.01 0.01 0.00 1.00
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9
PC10
PC11
calcium_100g
1.0
0.8
0.6
0.4
0.2
0.0

Tests statistiques

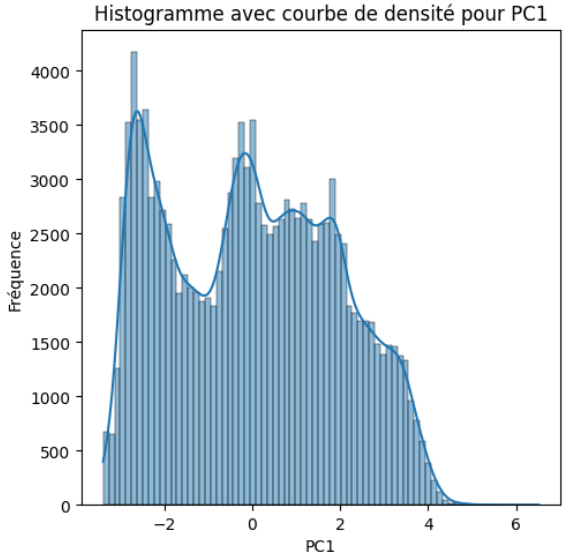
Pour la variable catégorielle

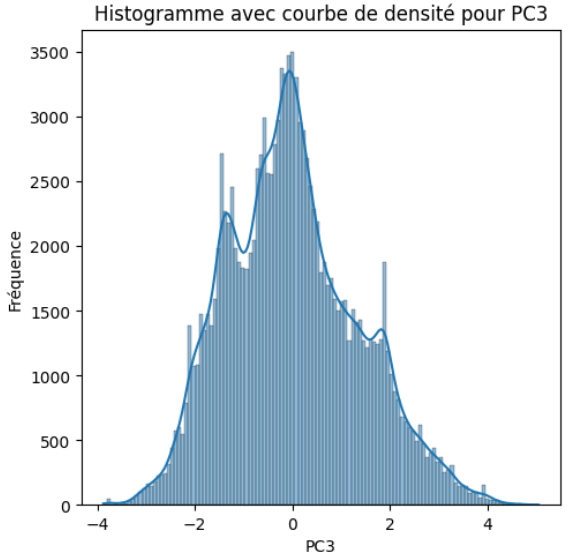
Test Kolmorov-Smirnov

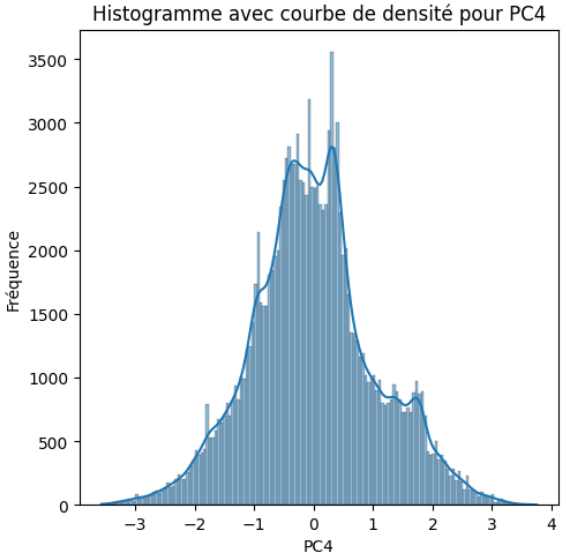
PC1 : Statistique = 0.2165941598036073,
p-value = 0.0
PC3 : Statistique = 0.09339074217058335,
p-value = 0.0
PC4 : Statistique = 0.039900171733098566,
p-value = 2.2360712623746O4e-181
PC5 : Statistique = 0.06908826726005846,
p-value = 0.0

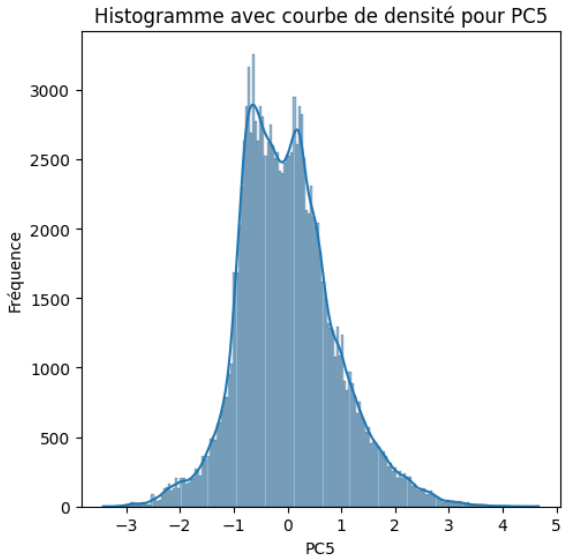
Test Spearman

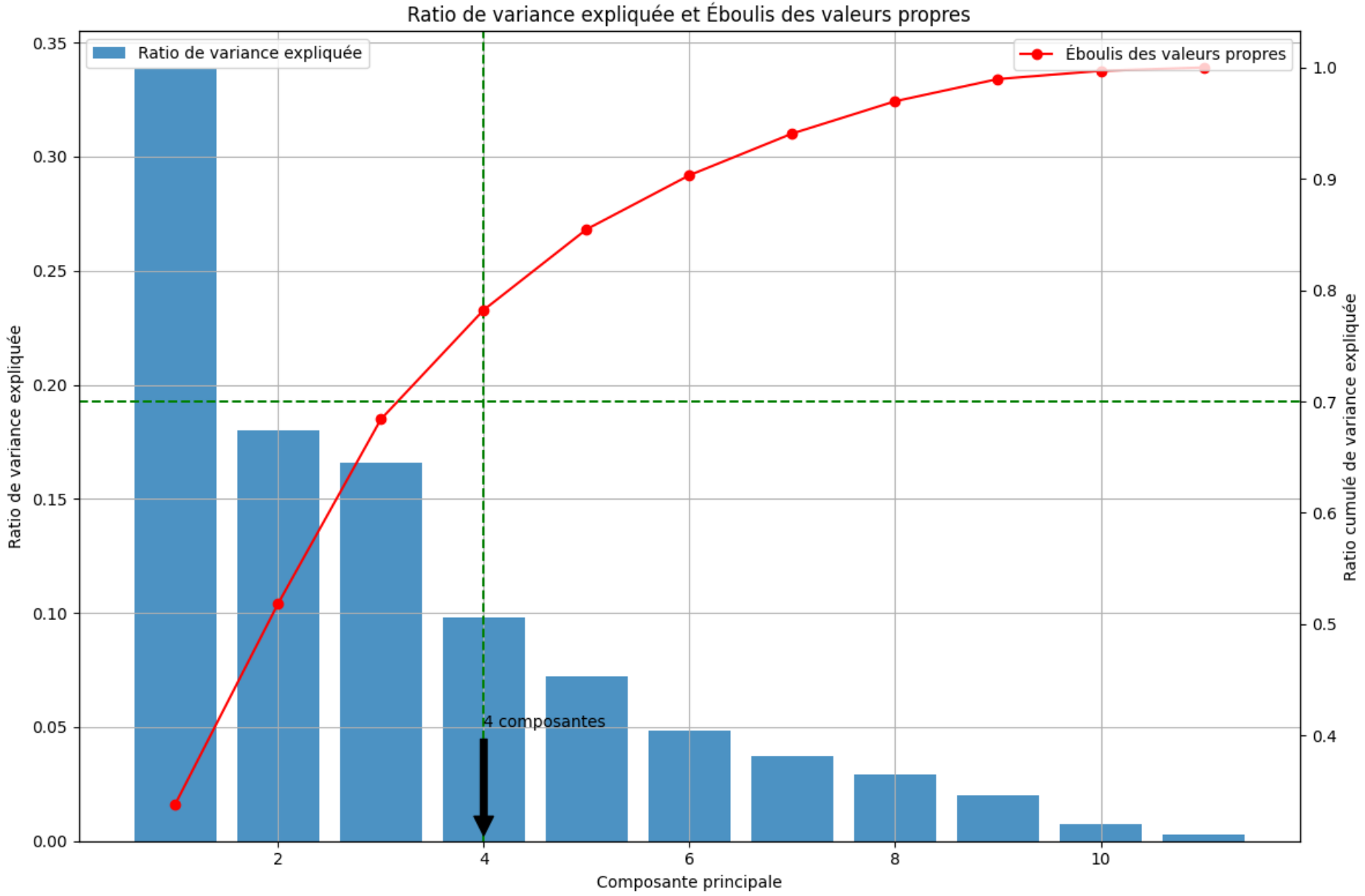
```
PC1 et calcium 100g : 0.17159376490111275,
          p-value :0.0
PC3 et calcium 100g : 0.29003305611411057,
          p-value :0.0
PC4 et calcium 100g : 0.6703412004890147,
          p-value :0.0
PC5 et calcium 100g : 0.5003067599770551,
          p-value :0.0
```










Ratio de variance expliquée et Éboulis des valeurs propres
Ratio de variance expliquée
Éboulis des valeurs propres
4 composantes
Ratio de variance expliquée
Ratio cumulé de variance expliquée
Composante principale
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.4
0.5
0.6
0.7
0.8
0.9
1.0
2
4
6
8
10

Cercle de corrélation

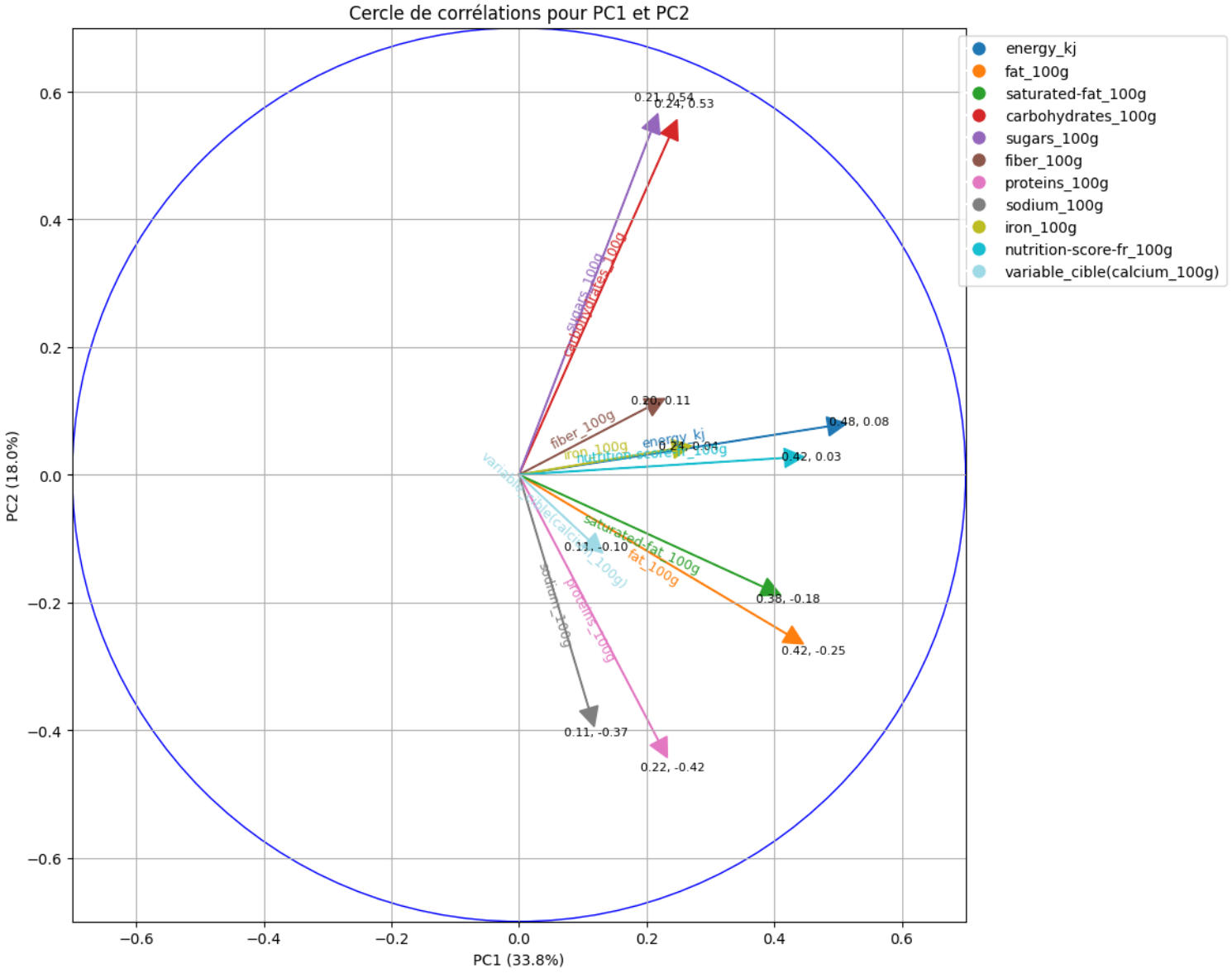
Pour la variable catégorielle

PC1
Explique 33,8% de la variance

PC2
Explique 18,0% de la variance

Les variables qui doivent impérativement être rempli dans la future application sont :
- Energy kj
- Carbohydrates
- sugars
- saturated-fat
- fat
- proteins
- sodium

Ces informations permettront de définir le nutriscore (A, B, C, D et E), l'ajout de nouvelles variables comme les fibres, le calcium permettrait de compléter l'information nutritionnelles.



Résumé du projet :

* Améliorer la base de données d'Open Food Facts pour santé publique France
* Mise en place d'un système de suggestion et d'auto-complétion pour réduire les erreurs de saisie

Réalisation :

Impact :

Réduction des champs à saisir, cela signifie moins de possibilité d'erreur de saisie donc une amélioration de la qualité des données

Nous avons maintenant une meilleure compréhension des facteurs influençant la qualité nutritionnelle des produits

Proposition d'amélioration :

* Dans l'application final limité la saisie à une valeurs maximal (si 100g impossibilité de saisir 101)
* Sensibiliser les utilisateurs sur l'importance de la qualité des données
* Ajouter d'autre informations nutritionnelles comme les différentes vitamines


MERCI