Santé publique France

Application relative au score d'un produit alimentaire

Projet 3

Marc Sellam.

Sommaire:

Idée d'application

Un jeu de données Open Food Facts est fourni, celui-ci contenant des informations concernant des produits alimentaires dont le Nutri-score.

La démarche étant de trouver une idée d'application afin de mieux se nourrir.

L'idée serait de scanner un produit avec une application mobile afin de savoir s'il est bon pour la santé ou pas, et dans quelle mesure.

C'est la que le Nutri score intervient, car lorsque le consommateur scannera le produit, une voix lui dira en l'affichant sur l'écran: 'très bien !!' pour le nutriscore A, 'bien !!' pour le nutriscore B et 'pas bien !!' pour les nutriscore C D et E.

Les données fournies

La moitié des variables ont plus de 97% de données manquantes

```
nb de lignes x nb de colonnes :  (320772, 162) .

% de données manquantes :
 water-hardness_100g                  100.000000
no_nutriments                         100.000000
ingredients_that_may_be_from_palm_oil 100.000000
nutrition_grade_uk                    100.000000
nervonic-acid_100g                    100.000000
                                          ...
created_datetime                        0.002806
created_t                               0.000935
creator                                 0.000623
last_modified_datetime                  0.000000
last_modified_t                         0.000000
Length: 162, dtype: float64

nb de lignes dupliquées : 0 .
```

```
vitamin-d_100g                        0.978000
vitamin-b6_100g                       0.978851
magnesium_100g                        0.980506
phosphorus_100g                       0.981778
vitamin-b12_100g                      0.983477
                                          ...
ingredients_that_may_be_from_palm_oil 1.000000
ingredients_from_palm_oil             1.000000
no_nutriments                         1.000000
nervonic-acid_100g                    1.000000
water-hardness_100g                   1.000000
Length: 83, dtype: float64
```

Choix des variables

Le fichier fourni contient 162 variables, nous en sélectionnons 18.

```
nb de lignes x nb de colonnes :  (320772, 18) .

% de données manquantes :
 labels_fr                      85.451972
categories_fr                   73.685047
fiber_100g                      37.374210
nutrition_grade_fr              31.038245
nutrition-score-fr_100g         31.038245
saturated-fat_100g              28.437021
carbohydrates_100g              24.061951
fat_100g                        23.967491
sugars_100g                     23.630803
ingredients_from_palm_oil_n     22.393787
additives_n                     22.393787
salt_100g                       20.345292
proteins_100g                   18.969860
energy_100g                     18.598568
brands                           8.857382
product_name                     5.537266
countries_fr                     0.087289
code                             0.007170
dtype: float64

nb de lignes dupliquées : 18 .
```

Informations générales:

code, product_name.

Informations nutritionnelles:

energy_100g, fat_100g, saturated-fat_100g, carbohydrates_100g, sugars_100g, proteins_100g, salt_100g et fiber_100g.

Informations diverses:

ingredients_from_palm_oil_n, nutrition_grade_fr, nutrition-score-fr_100g et additives_n.

Certains tags: *brands, categories_fr, countries_fr, labels_fr.*

Labels_fr et categories_fr ont plus de 73% de valeurs manquantes.

Inspection des données

Les valeurs aberrantes.

Il apparait des valeurs extrêmes

	code	product_name	brands	categories_fr	countries_fr	nutrition_grade_fr	labels_fr
count	320749	303010	292360	84411	320492	221210	46666
unique	320579	221347	58784	21152	722	5	13655
top	70650800367	Ice Cream	Carrefour	Snacks sucrés,Biscuits et gâteaux,Biscuits	États-Unis	d	Bio,Bio européen,AB Agriculture Biologique
freq	3	410	2978	802	172998	62763	5311

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	nutrition-score-fr_100g	ingredients_from_palm_oil_n	additives_n
count	2.611130e+05	243891.000000	229554.000000	243588.000000	244971.000000	200886.000000	259922.000000	255510.000000	221210.000000	248939.000000	248939.000000
mean	1.141915e+03	12.730379	5.129932	32.073981	16.003484	2.862111	7.075940	2.028624	9.165535	0.019659	1.936024
std	6.447154e+03	17.578747	8.014238	29.731719	22.327284	12.867578	8.409054	128.269454	9.055903	0.140524	2.502019
min	0.000000e+00	0.000000	0.000000	0.000000	-17.860000	-6.700000	-800.000000	0.000000	-15.000000	0.000000	0.000000
25%	3.770000e+02	0.000000	0.000000	6.000000	1.300000	0.000000	0.700000	0.063500	1.000000	0.000000	0.000000
50%	1.100000e+03	5.000000	1.790000	20.600000	5.710000	1.500000	4.760000	0.581660	10.000000	0.000000	1.000000
75%	1.674000e+03	20.000000	7.140000	58.330000	24.000000	3.600000	10.000000	1.374140	16.000000	0.000000	3.000000
max	3.251373e+06	714.290000	550.000000	2916.670000	3520.000000	5380.000000	430.000000	64312.800000	40.000000	2.000000	31.000000

Nettoyage

Nombre de catégories.

categorie	
Aliments et boissons à base de végétaux	18294
Aliments d'origine végétale	15487
Snacks sucrés	9206
Produits laitiers	7077
Céréales et pommes de terre	6334
Plats préparés	6026
Boissons	5614
Aliments à base de fruits et de légumes	4960
Céréales et dérivés	4821
Desserts	4327
Biscuits et gâteaux	3939
Frais	3856
Viandes	3576
Boissons à base de végétaux	3440
Conserves	3342
Surgelés	3069
Produits à tartiner	2992
Fromages	2982
Petit-déjeuners	2884
Epicerie	2882
Chocolats	2689
Biscuits	2589
Boissons aux fruits	2510
Charcuteries	2468
Sauces	2434
Légumes et dérivés	2365
Fruits et produits dérivés	2183
Snacks salés	2144
Confiseries	2104
Boissons non sucrées	2036

Name: total, dtype: Int64

Apres le retrait des produits qui ne mentionnent pas le code –barre, le pays, le nom du produit, la marque ou son nutri grade et des produits dont aucune des mentions obligatoires n'est renseignée.

Combiens a ton de catégories et dans quelles proportions?

Retrait de la variable categorie_fr

Nettoyage

Filtrage des codes, pays et doublons.

Retrait des codes barres qui ne sont pas de 8 et 13 chiffres.

Retrait des produits qui ne sont pas vendus en France.

Retrait des code en doublons.

	code	product_name	brands	nutrition_grade_fr	labels_fr
count	59788	59788	59788	59788	27108
unique	59788	50001	12323	5	7744
top	3366321050849	Mayonnaise	Carrefour	d	Bio,Bio européen,AB Agriculture Biologique
freq	1	42	2311	16573	4634

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	nutrition-score-fr_100g	ingredients_from_palm_oil_n	additives_n
count	5.978800e+04	43144.000000	59788.000000	43093.000000	59788.000000	43532.000000	59788.000000	59788.000000	59788.000000	40360.000000	40360.000000
mean	1.180748e+03	13.339136	5.447547	27.666490	13.324111	2.523057	7.767034	1.081057	8.680504	0.078791	2.052131
std	1.333686e+04	16.771379	8.555680	27.244518	19.013593	4.544715	7.834633	3.816733	9.047314	0.273341	2.616622
min	0.000000e+00	0.000000	0.000000	0.000000	-0.100000	0.000000	0.000000	0.000000	-15.000000	0.000000	0.000000
25%	4.350000e+02	1.400000	0.300000	4.100000	1.000000	0.000000	1.900000	0.080000	1.000000	0.000000	0.000000
50%	1.040000e+03	7.000000	2.000000	14.500000	4.000000	1.300000	6.000000	0.558800	9.000000	0.000000	1.000000
75%	1.657000e+03	21.000000	7.500000	53.000000	17.400000	3.200000	11.000000	1.244600	15.000000	0.000000	3.000000
max	3.251373e+06	380.000000	210.000000	190.000000	104.000000	178.000000	100.000000	211.000000	40.000000	2.000000	31.000000

Nettoyage

Nombre de produits ayant des valeurs aberrantes.

16236 produits aux caractéristiques extrêmes

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
228574	NaN	54.0	NaN	NaN	NaN	NaN	NaN	NaN
245362	NaN	NaN	NaN	NaN	NaN	NaN	NaN	10.86
245386	NaN	NaN	NaN	NaN	NaN	9.7	NaN	NaN
245318	NaN	NaN	23.0	NaN	NaN	NaN	25.0	NaN
245398	NaN	NaN	NaN	NaN	NaN	26.4	NaN	NaN
...	...	...	...	...	...	...	...	...
234999	NaN	NaN	20.0	NaN	NaN	NaN	NaN	NaN
234998	NaN	NaN	20.0	NaN	NaN	NaN	NaN	NaN
234886	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.60
234866	NaN	NaN	NaN	NaN	73.0	NaN	NaN	NaN
234845	NaN	NaN	32.0	NaN	56.0	NaN	NaN	NaN

16236 rows × 8 columns

Nombre de valeurs aberrantes par variable

```
energy_100g              428
fat_100g                1516
saturated-fat_100g      4309
carbohydrates_100g         2
sugars_100g             6152
fiber_100g              2717
proteins_100g           2385
salt_100g               3094
dtype: int64
```

Nettoyage

Traitement des valeurs aberrantes ou manquantes.

Toutes les variables nutritionnelles dont la valeur dépasse 100 sera replacée par 100.

La variable 'energy_100g' ne dépasse théoriquement pas les 3768 Kj: retrait de 90 produits.

Les valeurs manquantes de la variable label_fr remplacé par no_label les labelisés par labeled.

Les valeurs manquantes de la variable ingredients_from_palm_oil_n'remplacé par 0.0.

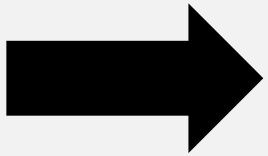
Les valeurs manquantes de la variable additives_n' remplacé par 1.0 pour les nutriscore 'e' et 0.0 pour les autres.

```
nb de lignes x nb de colonnes :  (59698, 16) .

% de données manquantes :
 labels_fr                      54.628296
ingredients_from_palm_oil_n    32.421522
additives_n                    32.421522
carbohydrates_100g             27.845154
fat_100g                       27.759724
fiber_100g                     27.223693
code                            0.000000
product_name                    0.000000
brands                          0.000000
nutrition_grade_fr              0.000000
energy_100g                     0.000000
saturated-fat_100g              0.000000
sugars_100g                     0.000000
proteins_100g                   0.000000
salt_100g                       0.000000
nutrition-score-fr_100g         0.000000
dtype: float64

nb de lignes dupliquées : 0 .
```

Choix de laisser les valeurs manquantes pour les variables 'carbohydrates_100g ' , 'fat_100g' et 'fiber_100g'.



```
nb de lignes x nb de colonnes :  (59698, 16) .

% de données manquantes :
 carbohydrates_100g             27.845154
fat_100g                       27.759724
fiber_100g                     27.223693
code                            0.000000
product_name                    0.000000
brands                          0.000000
nutrition_grade_fr              0.000000
energy_100g                     0.000000
saturated-fat_100g              0.000000
sugars_100g                     0.000000
proteins_100g                   0.000000
salt_100g                       0.000000
nutrition-score-fr_100g         0.000000
labels_fr                       0.000000
ingredients_from_palm_oil_n     0.000000
additives_n                     0.000000
dtype: float64

nb de lignes dupliquées : 0 .
```

Nettoyage

Données après traitements.

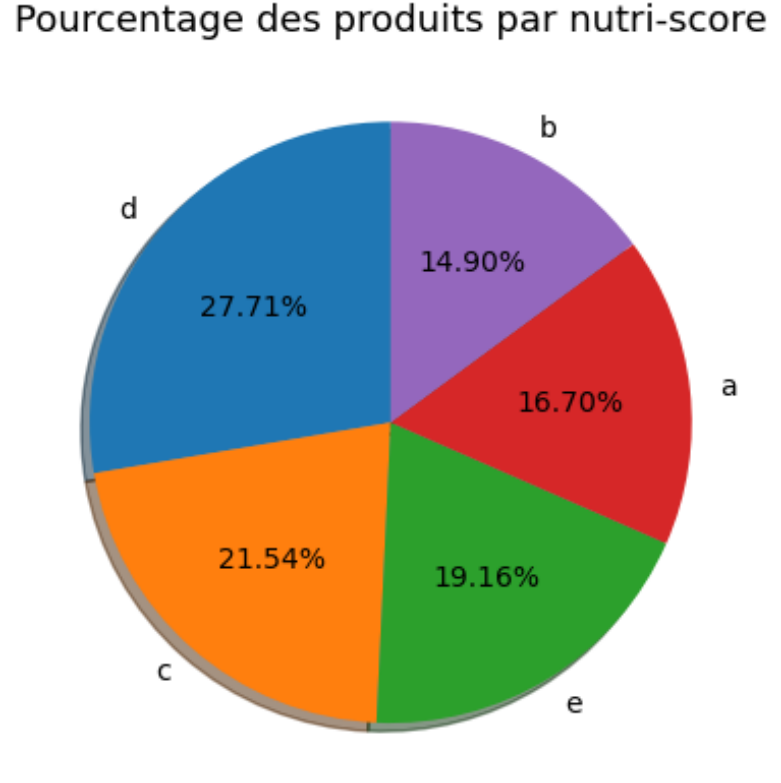
	code	product_name	brands	nutrition_grade_fr	labels_fr
count	59698	59698	59698	59698	59698
unique	59698	49926	12313	5	2
top	3366321050849	Mayonnaise	Carrefour	d	no_label
freq	1	41	2307	16541	32612

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	nutrition-score-fr_100g
count	59698.000000	43126.000000	59698.000000	43075.000000	59698.000000	43446.000000	59698.000000	59698.000000	59698.000000
mean	1117.074747	13.325059	5.442879	27.657283	13.312217	2.519176	7.765165	1.077638	8.665701
std	771.305599	16.667821	8.510855	27.227545	18.997108	4.470946	7.832324	3.727829	9.042082
min	0.000000	0.000000	0.000000	0.000000	-0.100000	0.000000	0.000000	0.000000	-15.000000
25%	434.000000	1.400000	0.300000	4.100000	1.000000	0.000000	1.900000	0.080000	1.000000
50%	1038.000000	7.000000	2.000000	14.500000	4.000000	1.300000	6.000000	0.558800	9.000000
75%	1653.000000	21.000000	7.500000	53.000000	17.300000	3.200000	10.900000	1.244600	15.000000
max	3768.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	40.000000

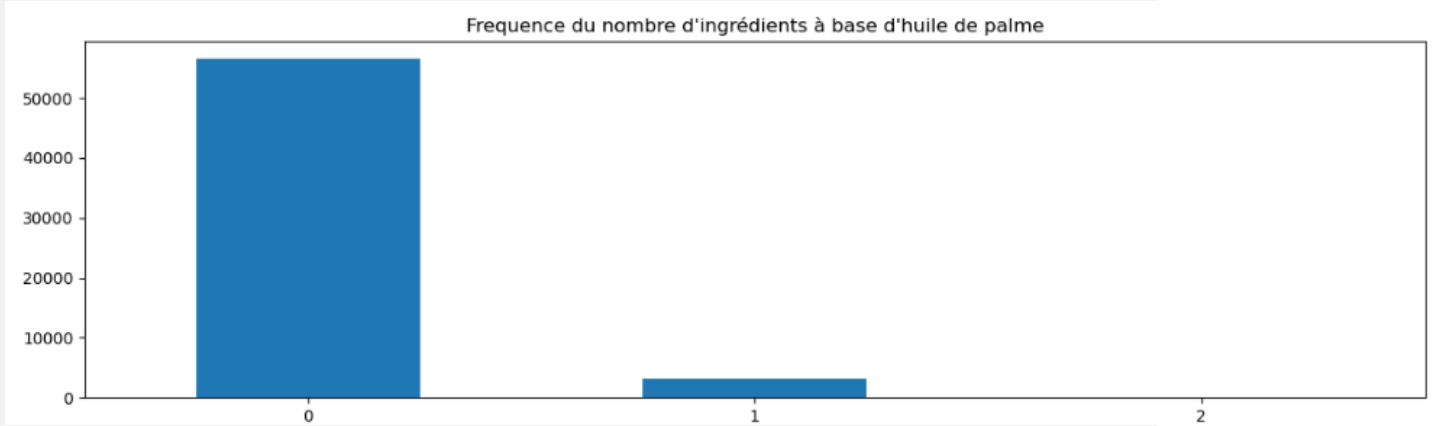
Analyses graphiques univariées

Les variables qualitatives

'nutrition_grade_fr'



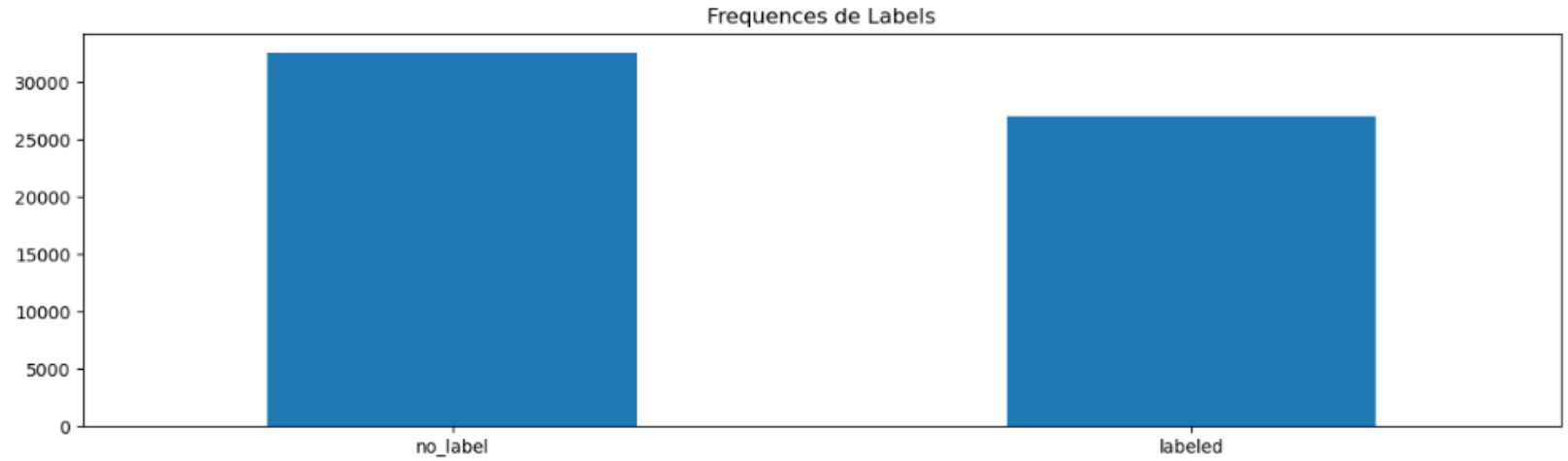
'ingredients_from_palm_oil_n'



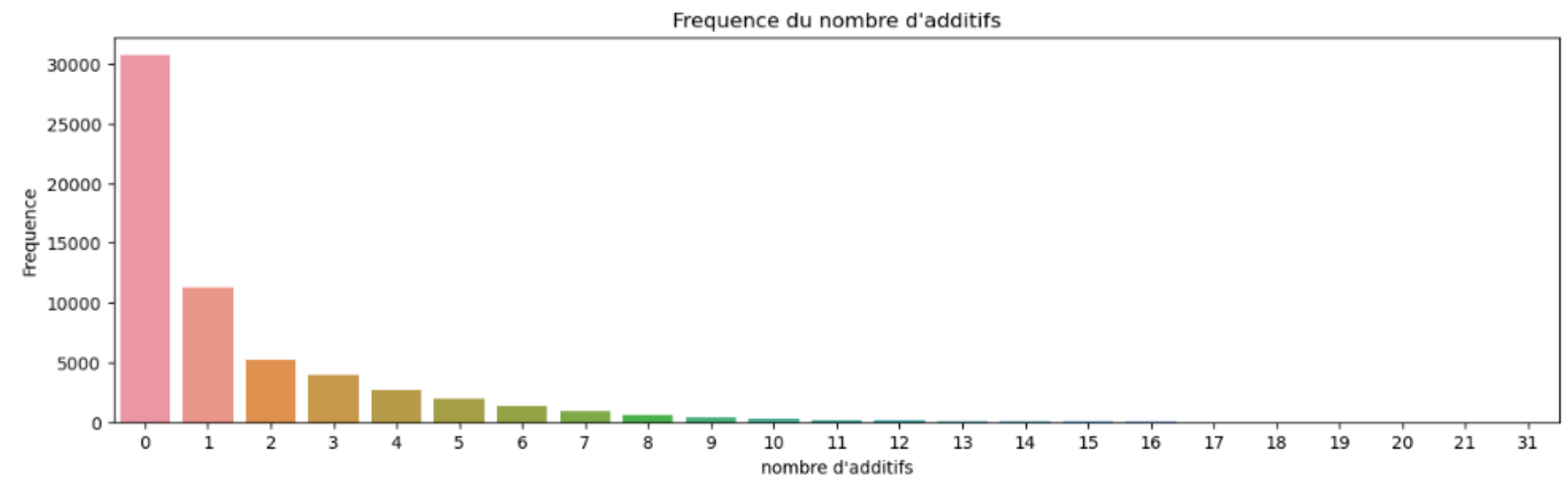
Analyses graphiques univariées

Les variables qualitatives

'labels_fr'

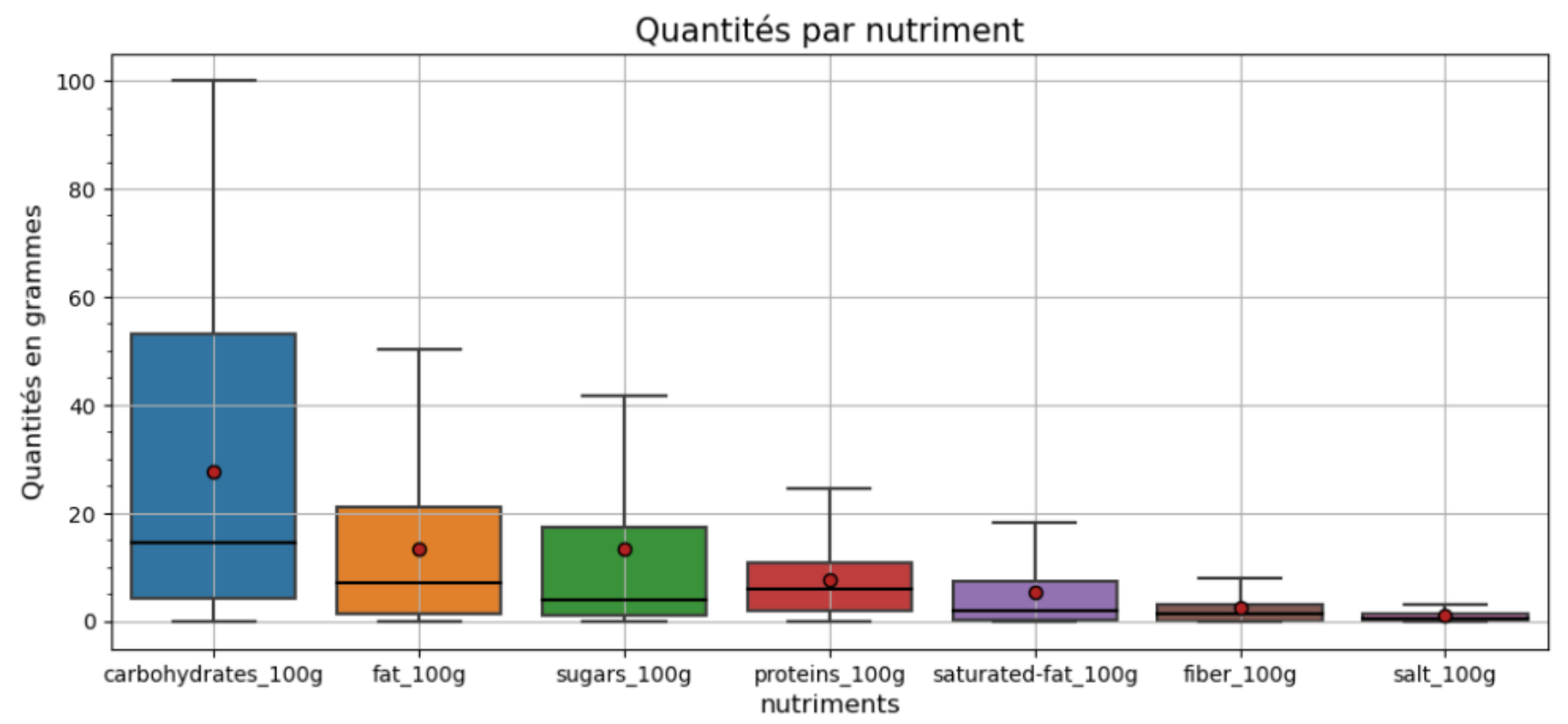


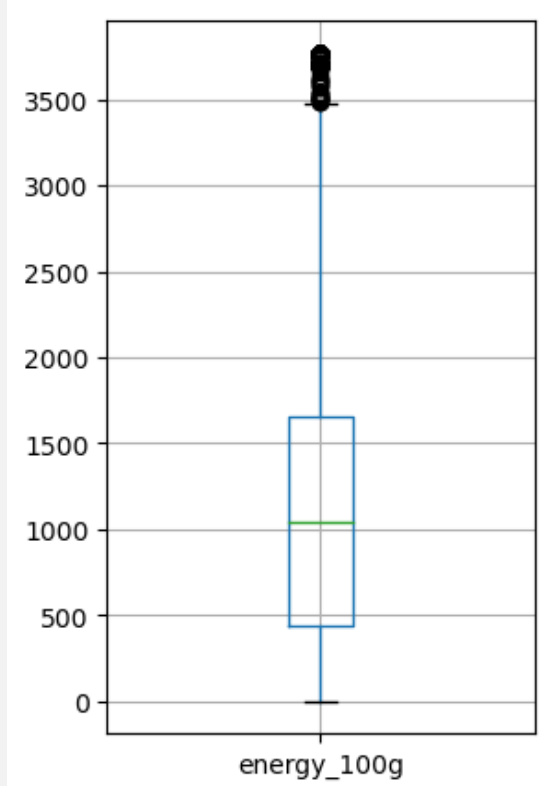
'nutrition_grade_fr'



Analyses graphiques univariées

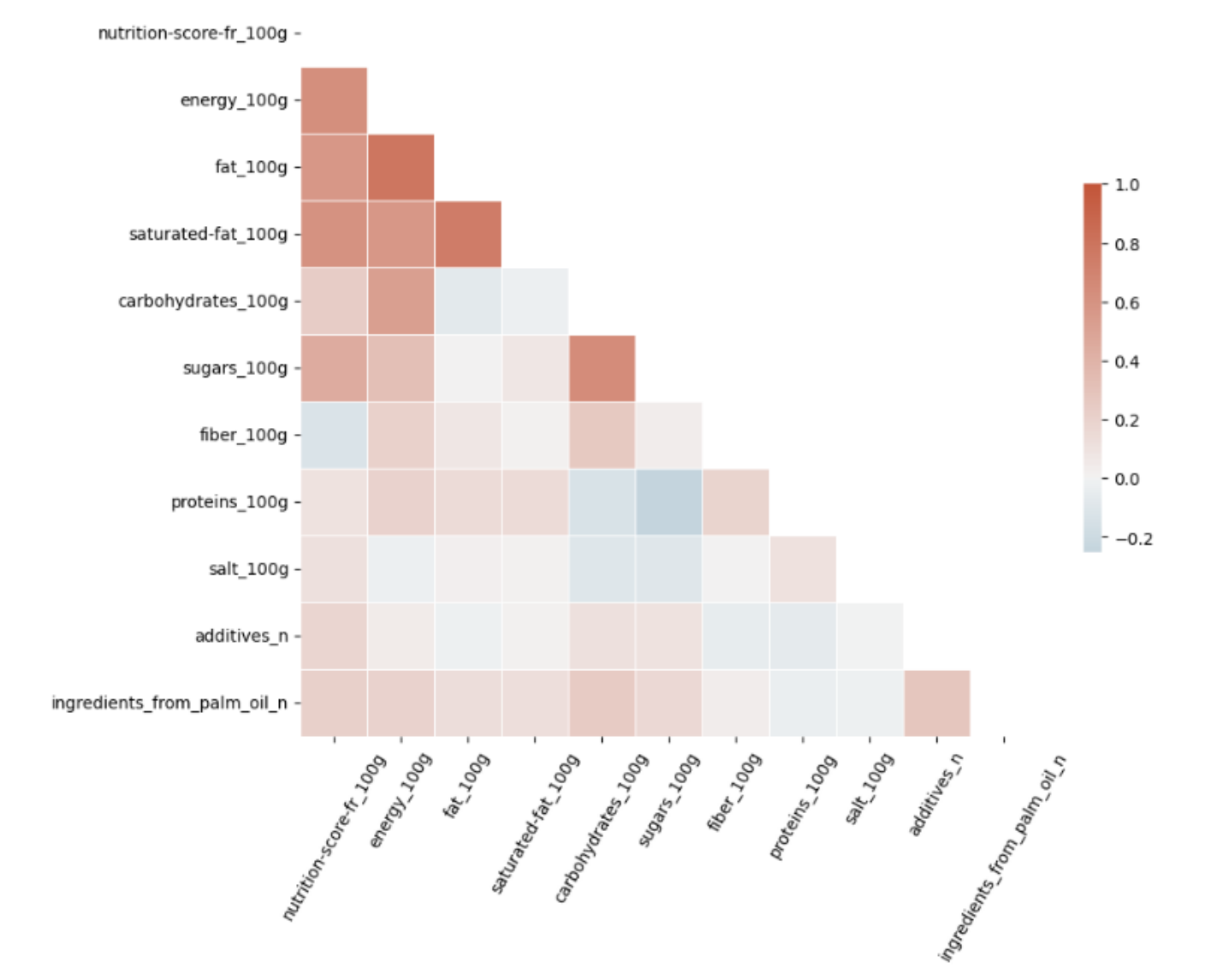
Les variables quantitatives





Analyses graphiques bivariées

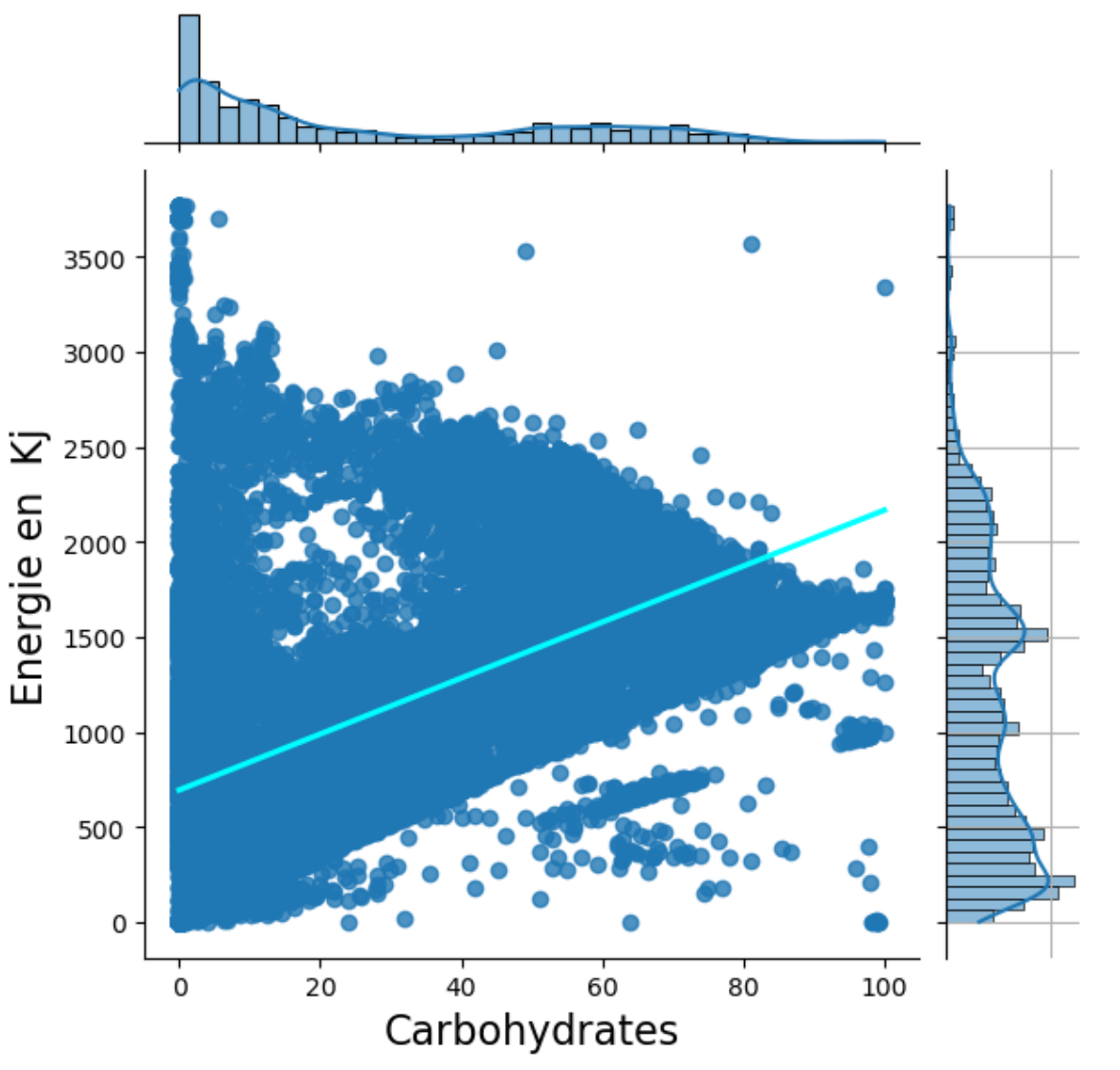
Corrélations entre les variables quantitatives.



Corrélations linéaires :
Energie / Carbohydrates
Energie / Matières grasses

Analyses graphiques bivariées

Corrélation entre l'énergie et les glucides.

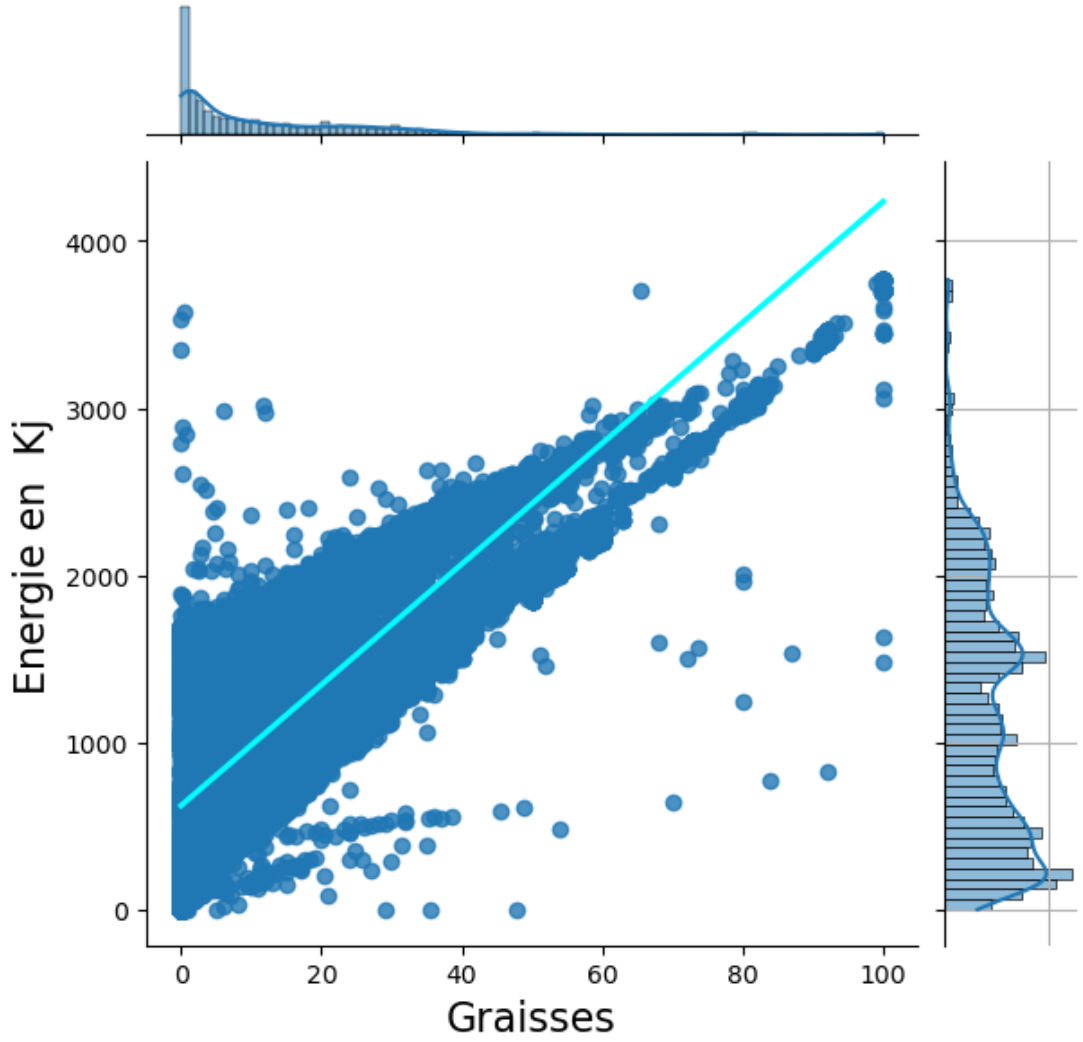


Coefficient de détermination (R^2): 0.5247817780841537

Il existe une corrélation entre l'énergie et les carbohydrates.

Analyse graphique bivariée

Corrélation entre l'énergie et les matières grasses.

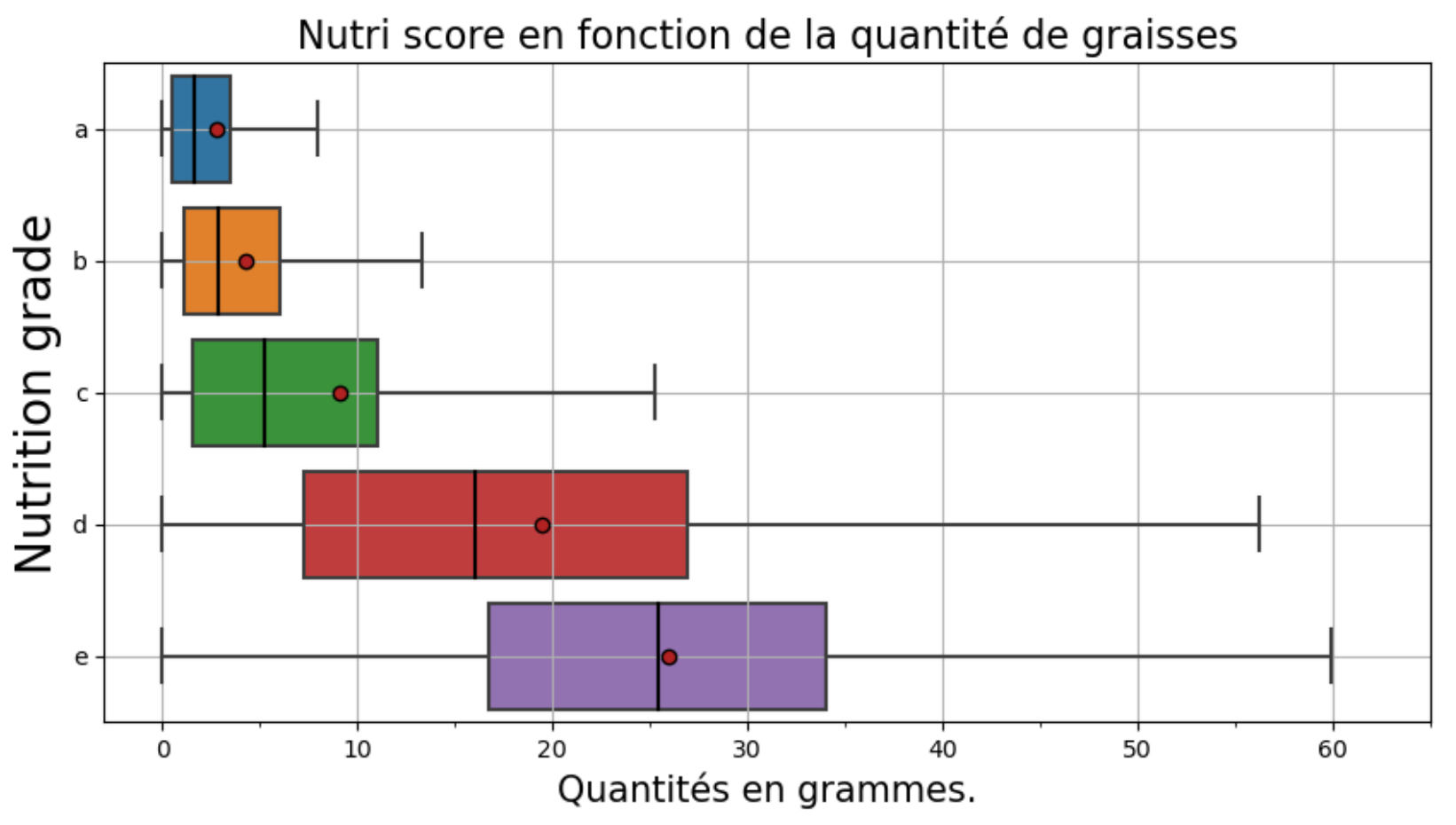


Coefficient de détermination (R^2): 0.7866063374321985

Il existe une corrélation forte entre l'énergie et la matière grasse.

Analyse graphique bivariée

Corrélation entre nutrigrade et les matières grasses.

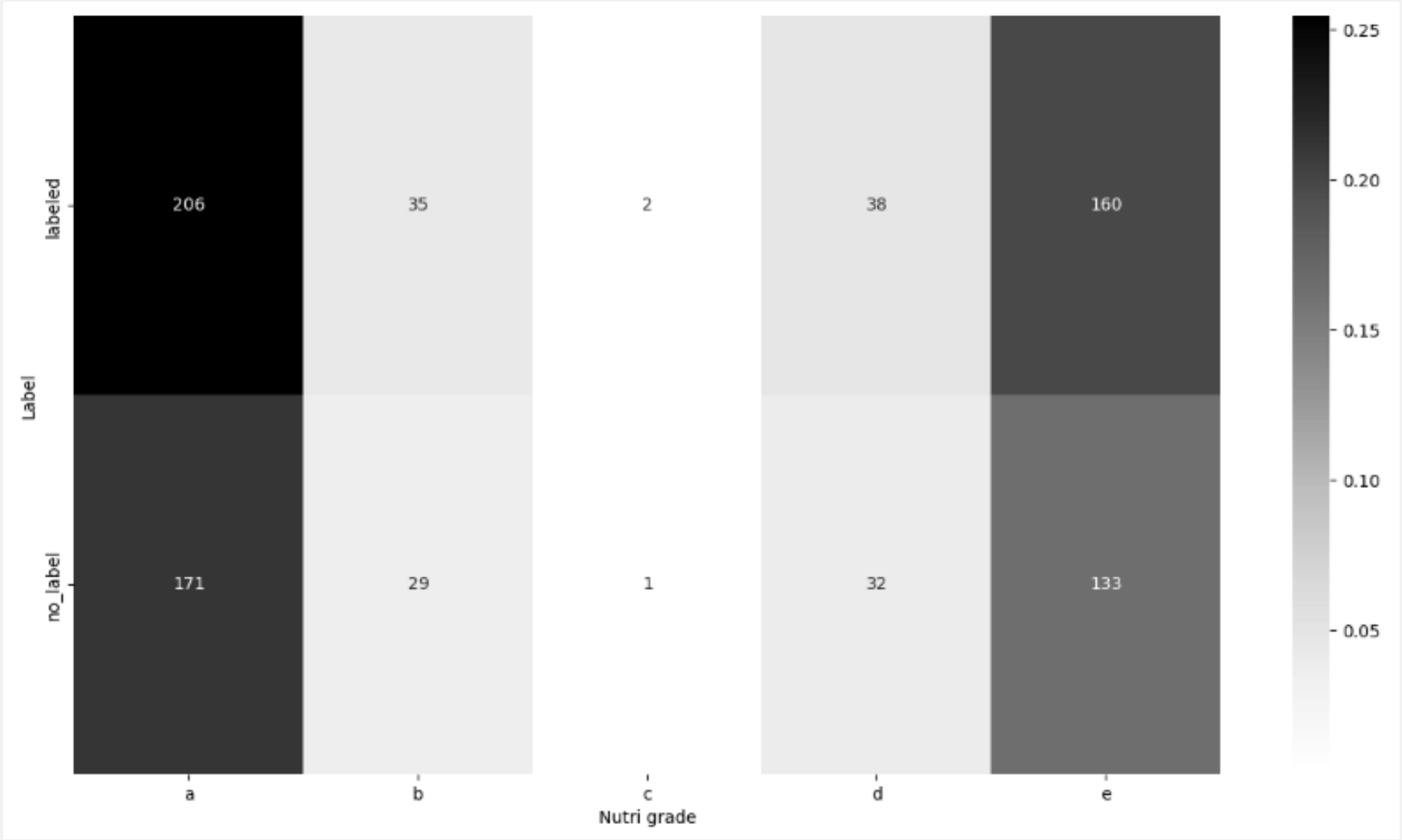


L'éta-carré (η2) est de 0.2711485781274317

```
Pour les Nutri-grade a.
Nombre de produits: 9967
Quantité moyenne: 2.79
Quantité médiane 1.6
Mode: 0.5
Ecart-type: 4.38
--------------------
Pour les Nutri-grade b.
Nombre de produits: 8894
Quantité moyenne: 4.31
Quantité médiane 2.9
Mode: 0.0
Ecart-type: 5.36
--------------------
Pour les Nutri-grade c.
Nombre de produits: 12858
Quantité moyenne: 9.08
Quantité médiane 5.2
Mode: 0.0
Ecart-type: 13.26
--------------------
Pour les Nutri-grade d.
Nombre de produits: 16541
Quantité moyenne: 19.45
Quantité médiane 16.05
Mode: 0.0
Ecart-type: 18.35
--------------------
Pour les Nutri-grade e.
Nombre de produits: 11438
Quantité moyenne: 25.92
Quantité médiane 25.4
Mode: 0.0
Ecart-type: 18.44
```

Analyse graphique bivariée

Corrélation entre le nutrigrade et la présence d'un label.

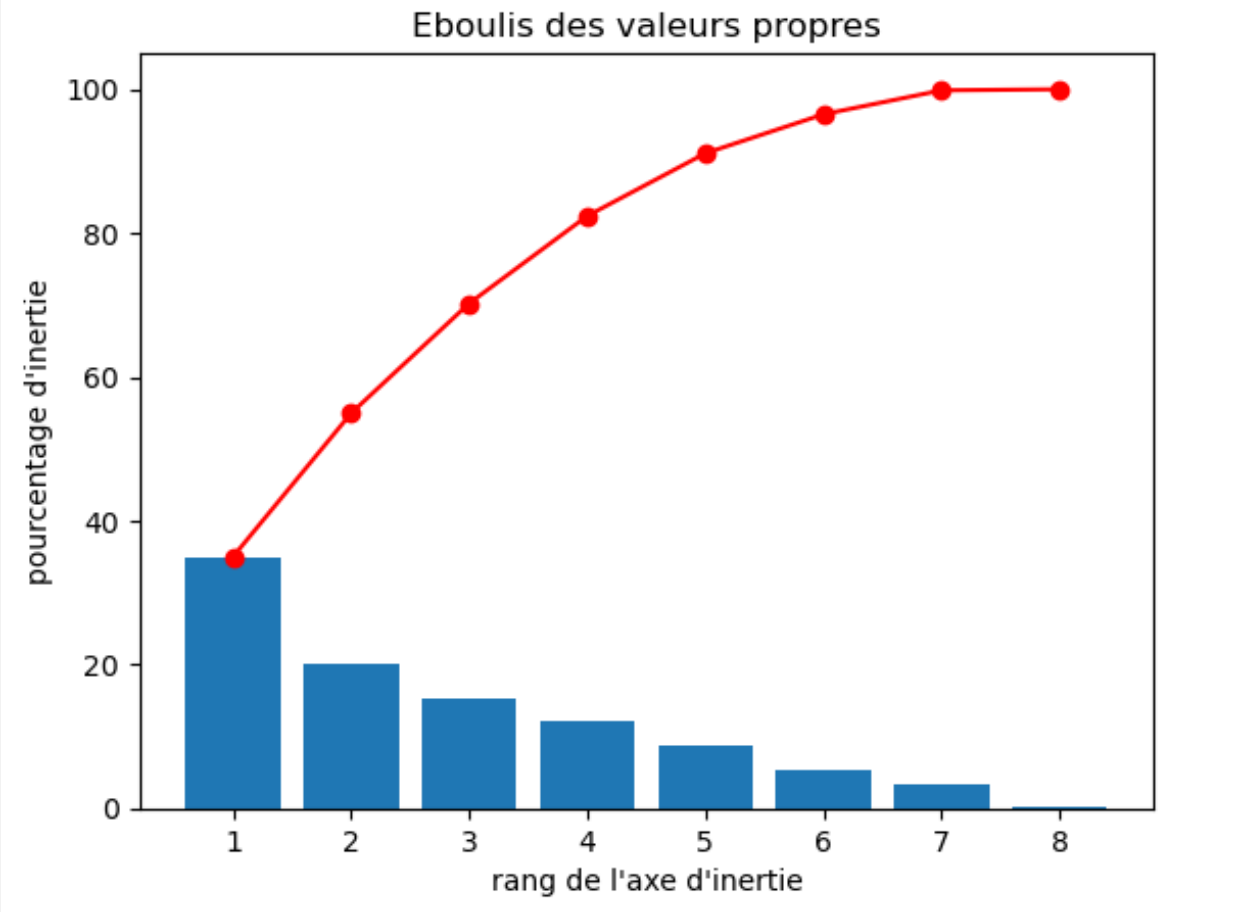


xi_n = 810

il existe un lien de dépendance entre le nutri grade et le label.

Analyse multivariée

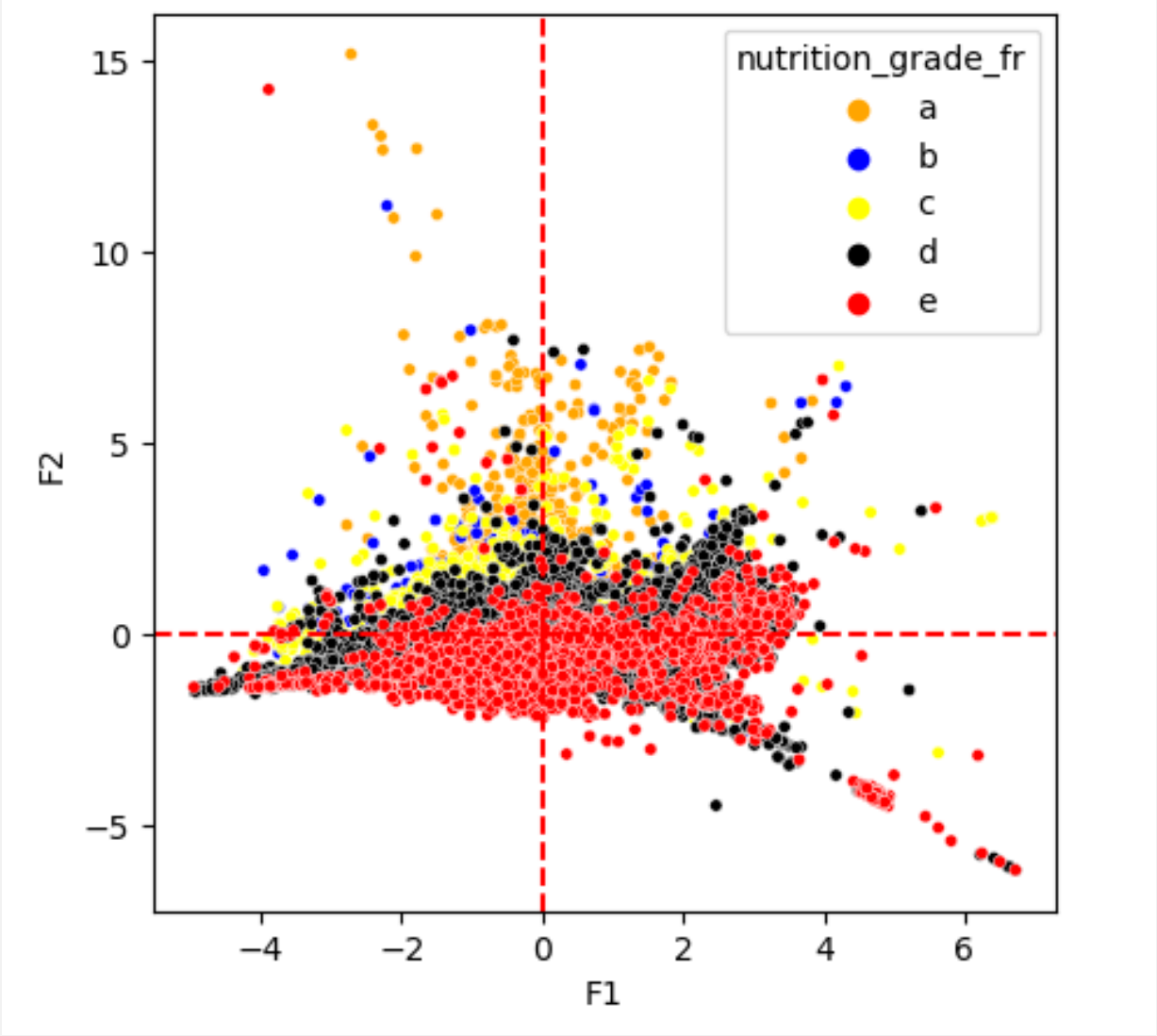
Analyse en composantes principales.

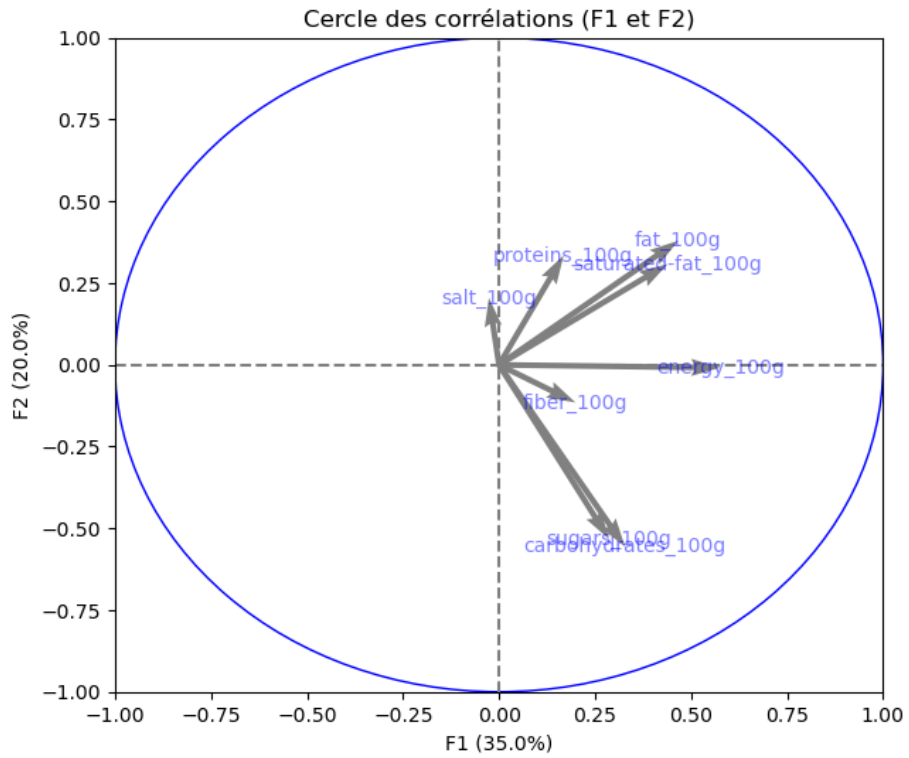


Le cumul de la variance expliqué nous indique que nous disposons de 35 % de l'information disponible sur la première composante et de 55 % avec les deux premiers facteurs.

Analyse multivariée

Cercle des corrélations et projections des individus.





Les variables les plus corrélées a F1 sont 'energy_100g','fat_100g','satured_fat_100g'(positivement) .Les variables 'sugar_100g' et 'carbohydrates_100g' sont corrélées négativement à F2.

Conclusion

Il existe une corrélation entre les glucides et l’énergie, ainsi qu’entre les matières grasses et l'énergie.

Autant il existe une influence de la présence d’un label sur le nutri grade , celle-ci n’est pas évidente entre le nutri grade et la présence de matières grasses.

Cependant, la faisabilité de l’application est bonne.