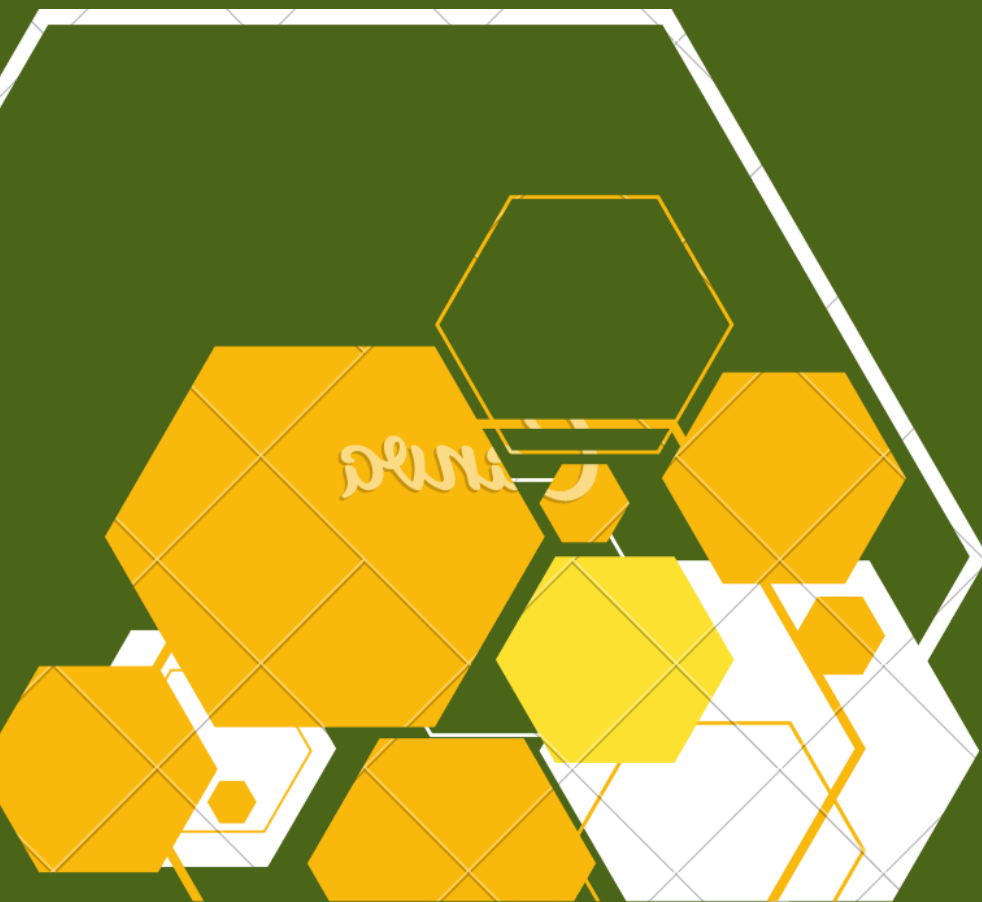


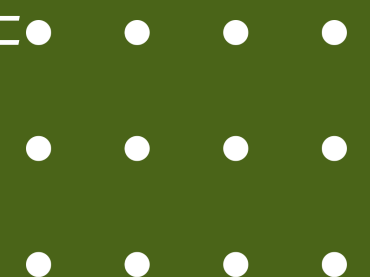
Santé publique France

Amélioration de la base de
données Open Food Facts



1/27

Présenté par Thierry KAPPE.



Le contexte

Santé publique France met à la disposition des particuliers et organisations sa base de données Open Food Facts pour connaître la qualité nutritionnelle des produits.

L'ajout de nouveaux produits dans cette base nécessite de remplir de nombreux champs créant ainsi plusieurs erreurs et de valeurs manquantes.

Notre mission est de proposer un système de suggestion qui permettra aux usagers de remplir efficacement la base de données.

Il est question dans un premier temps de **prendre en main les données existantes**, de les **nettoyer** et les **explorer** afin de déterminer la faisabilité du projet.

2/27



Choix de la thématique

Nous souhaitons dans notre approche, apporter une solution innovante à cet objectif de santé publique d'information nutritive, d'orientation vers de produits plus sains et d'efficacité de complétion pour l'ensemble des utilisateurs de cette base de données open source.

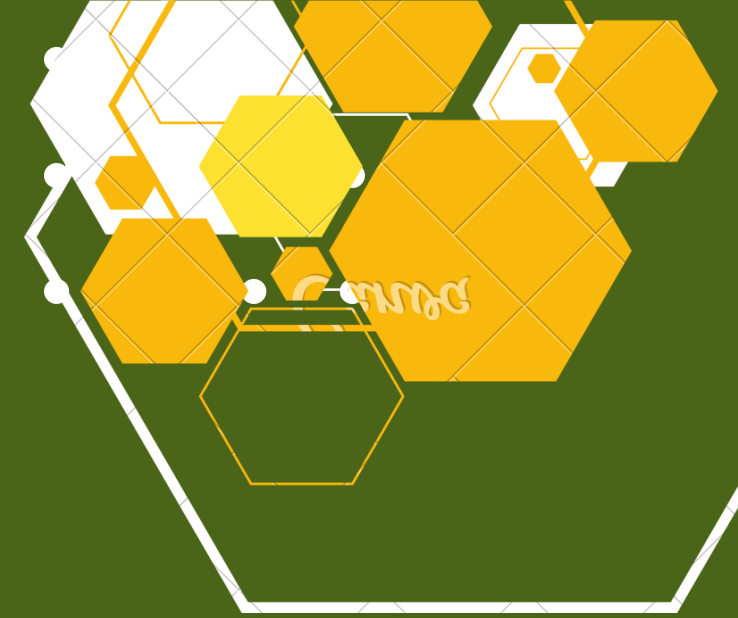
Notre solution s'articulera ainsi :

- Champ d'actions : ensemble des produits vendus sur le territoire français,
- Donner pour chaque produit la classe de son Nutri-score,
- Proposer aux usagers un produit alternatif mieux classé que celui recherché.

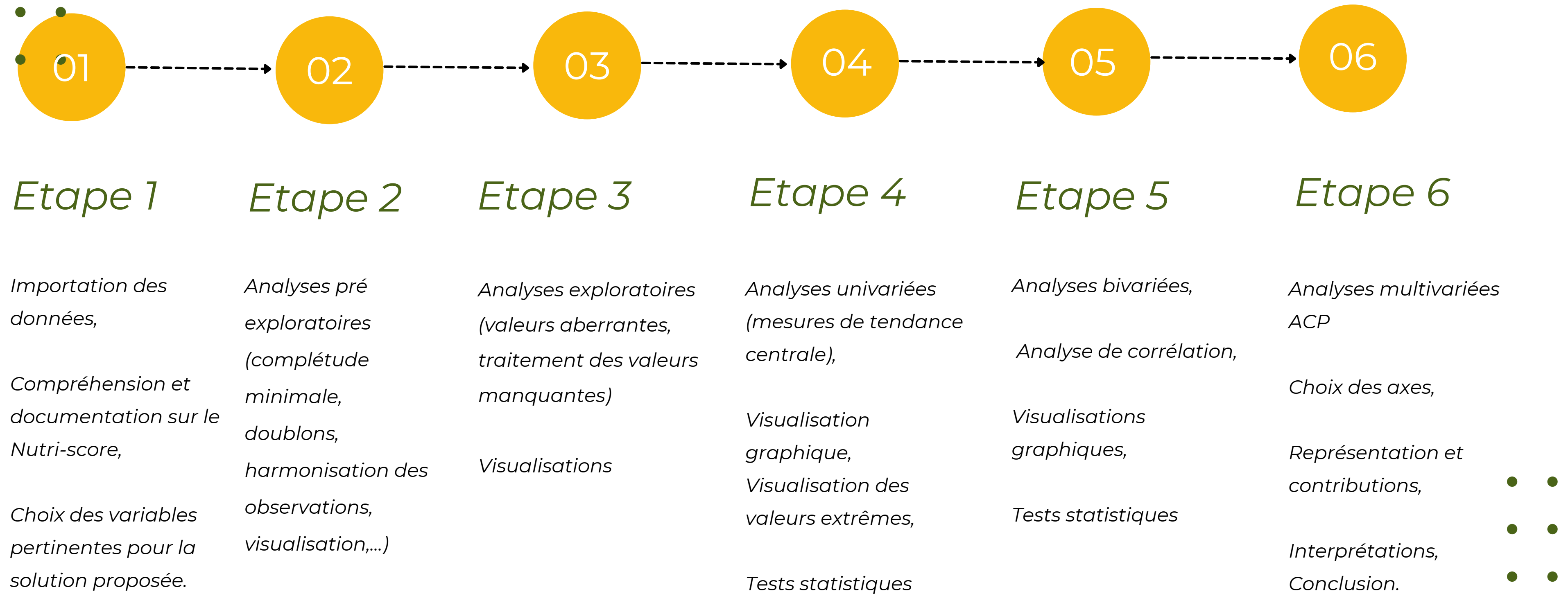
Les Objectifs de notre analyse

Nos travaux sont repartis ainsi :

- Description et compréhension du système de détermination du Nutri-score,
- Analyses pré exploratoires des données,
- Analyses exploratoires des données,
- Analyses univariées,
- Analyses bivariées,
- Analyses multivariées : ACP,
- Observations et conclusion.
- Point sur le RGPD

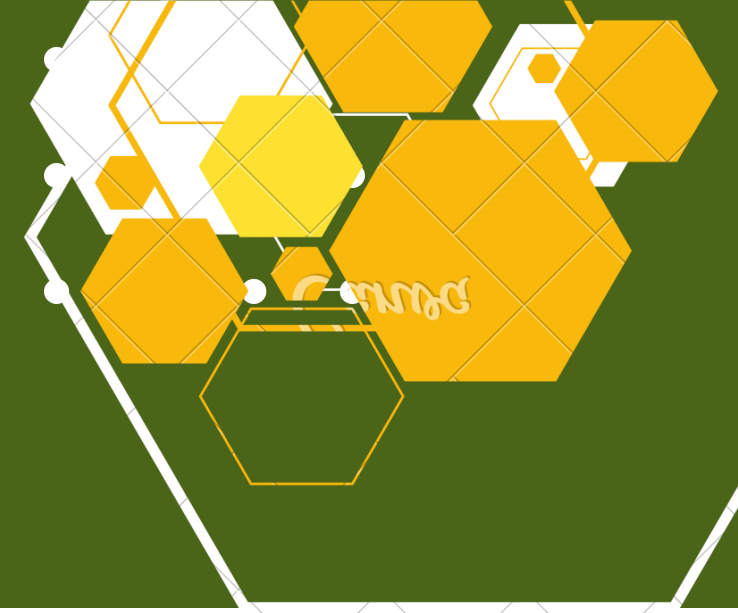


Vue globale des étapes clés



Compréhension et documentation du Nutri-score

Choix des variables pertinentes

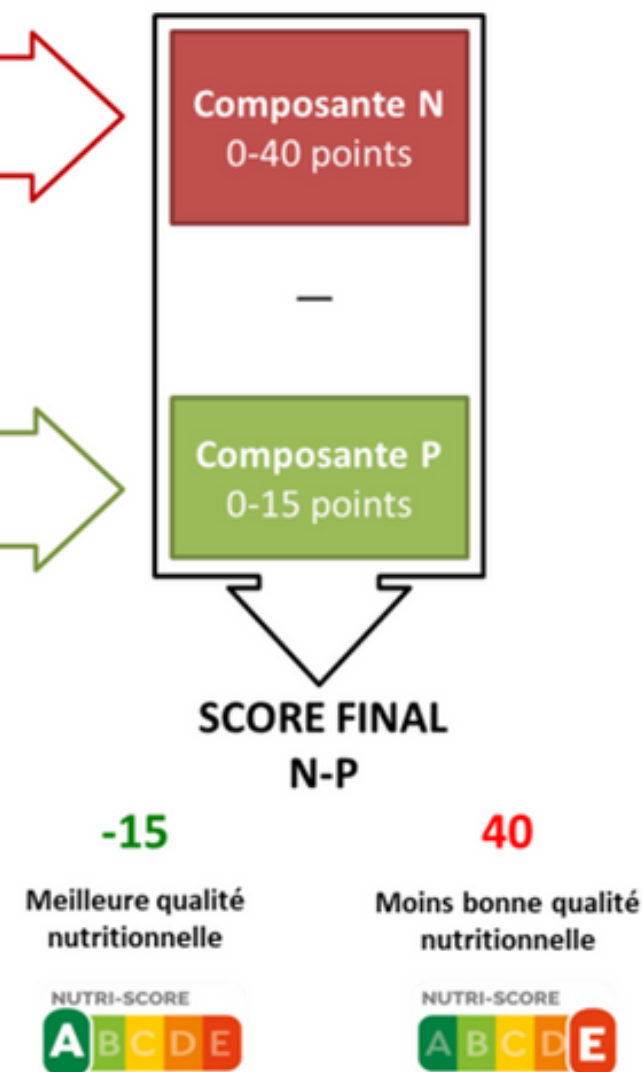


Système de calcul du Nutri-score

Elément /100g	Points
Energie (KJ)	0-10
Sucres (g)	0-10
Acides gras saturés (g)	0-10
Sodium (g)	0-10

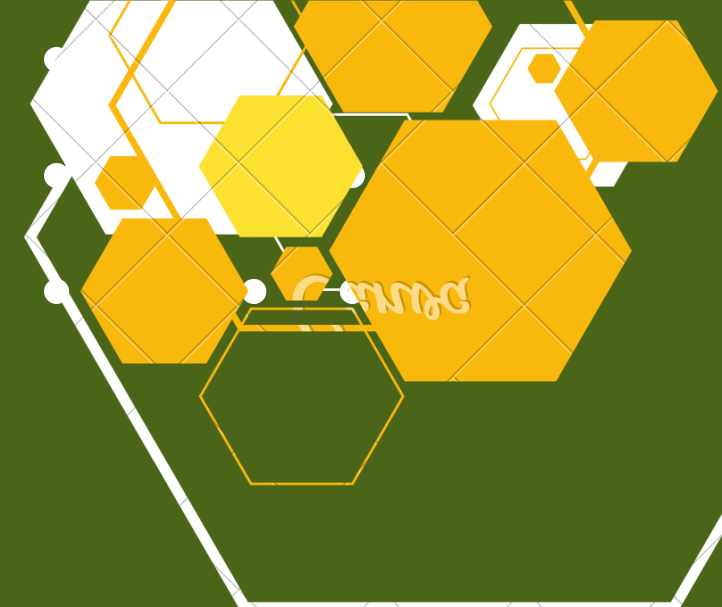
Elément /100g	Points
Fruits, légumes, légumineuses, fruits à coque, huiles de colza, de noix et d'olive (%)	0-5*
Fibres (g)	0-5
Protéines (g)**	0-5

*Dans le cas des boissons, un maximum de 10 points peut être attribué
**La prise en compte des protéines dépend du total de la composante N et des points pour les « Fruits, légumes, légumineuses, fruits à coque, huiles de colza, de noix et d'olive »



Liste des variables pertinentes retenues :

- Variables contenant les informations générales :
code, url, created_t, created_datetime, last_modified_t, last_modified_datetime, product_name, generic_name, quantity, categories.
- Variables nécessaires pour la détermination des composants N et P :
energy_100g, saturated-fat_100g, sugars_100g, fiber_100g, proteins_100g, sodium_100g, fruits-vegetables-nuts_100g.
- Variables retenues pour le traitement de certaines valeurs manquantes :
salt_100g, carbohydrates_100g, pnns_groups_1 et pnns_groups_2.
- Variables donnant les scores nutritionnels et le nutri-score :
nutrition-score-fr_100g, nutrition_grade_fr.



- **Fixation d'un taux de complétude minimal**

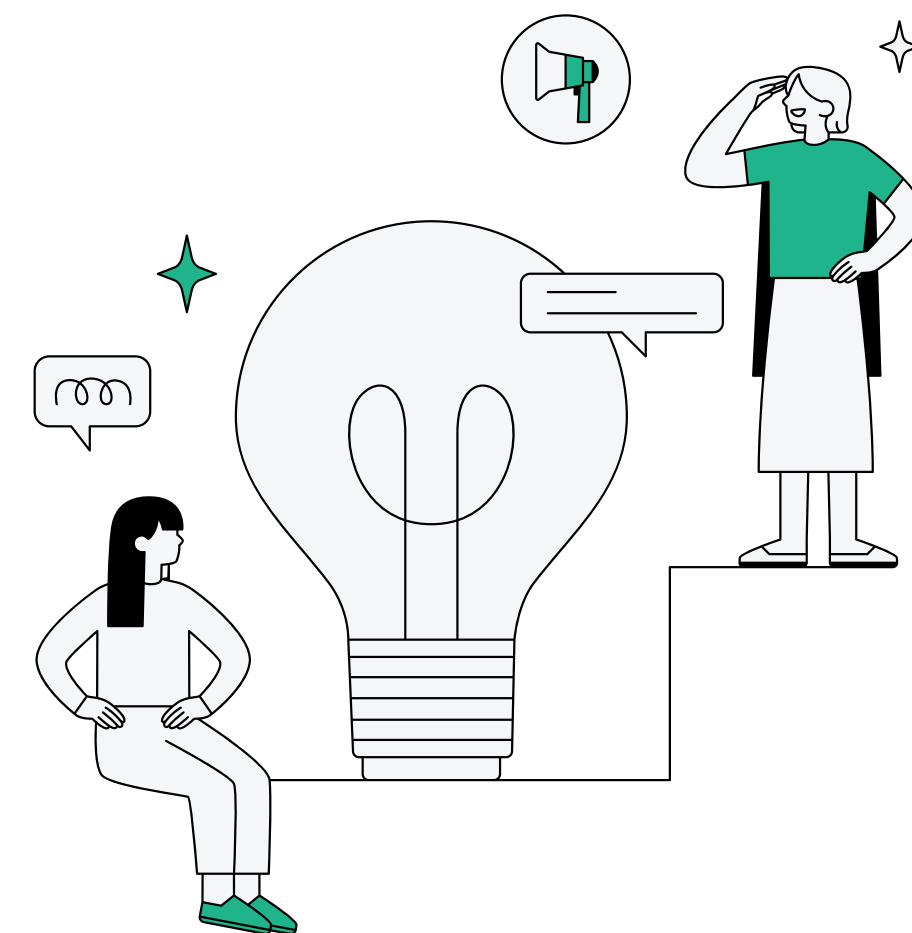
Pour les variables prises en compte pour la détermination du score nutritionnel(7 au total), nous avons pris comme postulat de ne retenir que les produits ayant au moins 3 indicateurs renseignés sur les 7 soit 43% de taux de remplissage.

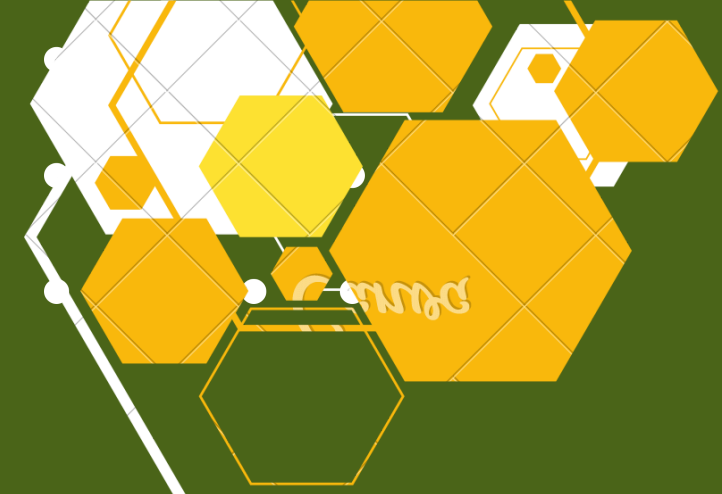
- **Extraction des produits vendus en France**

Notre champ d'action dans la solution proposée se limite aux produits vendus en France

- **Elimination des doublons**

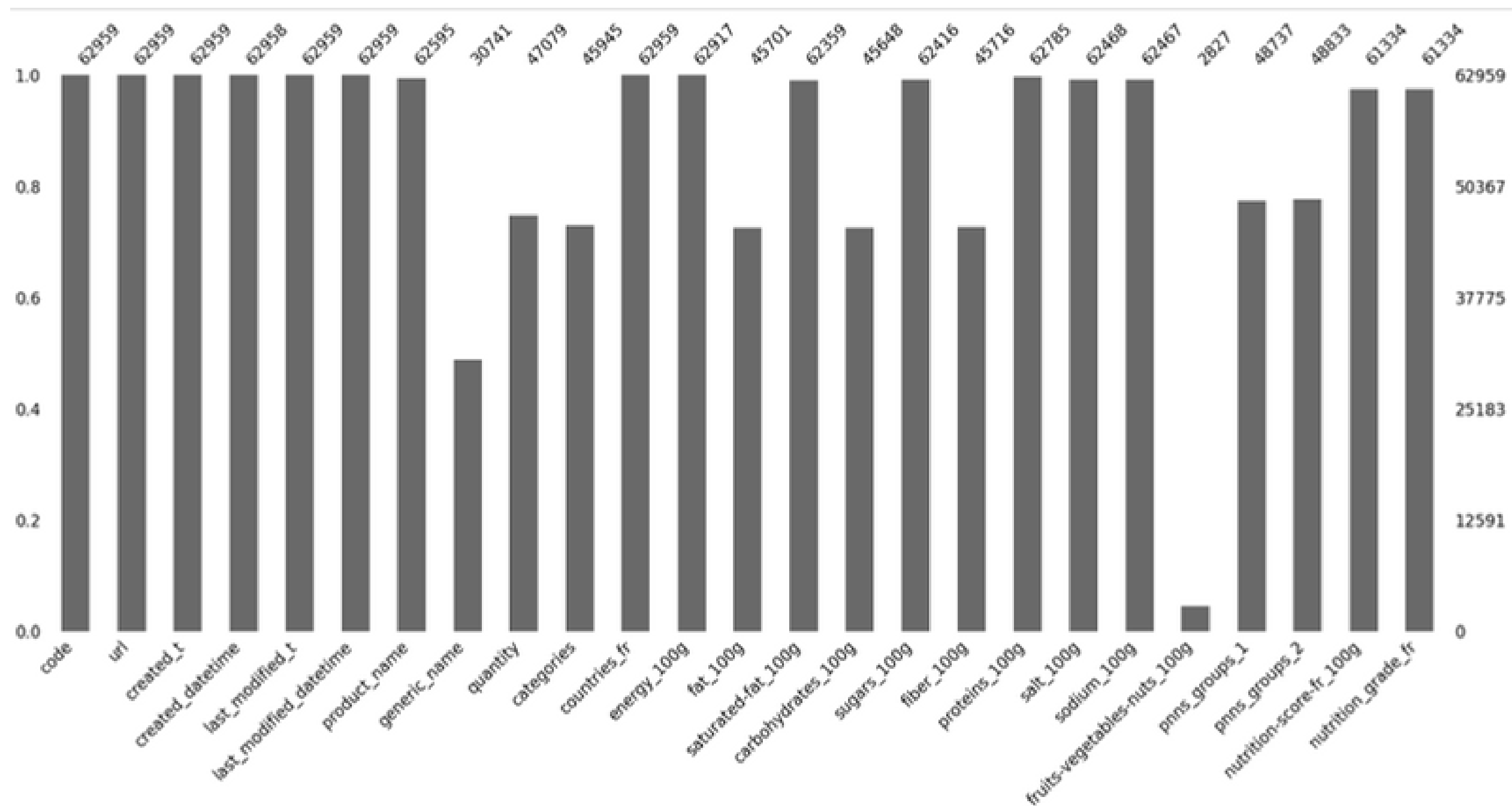
Suppression des produits (code) présents plusieurs fois dans le jeu des données



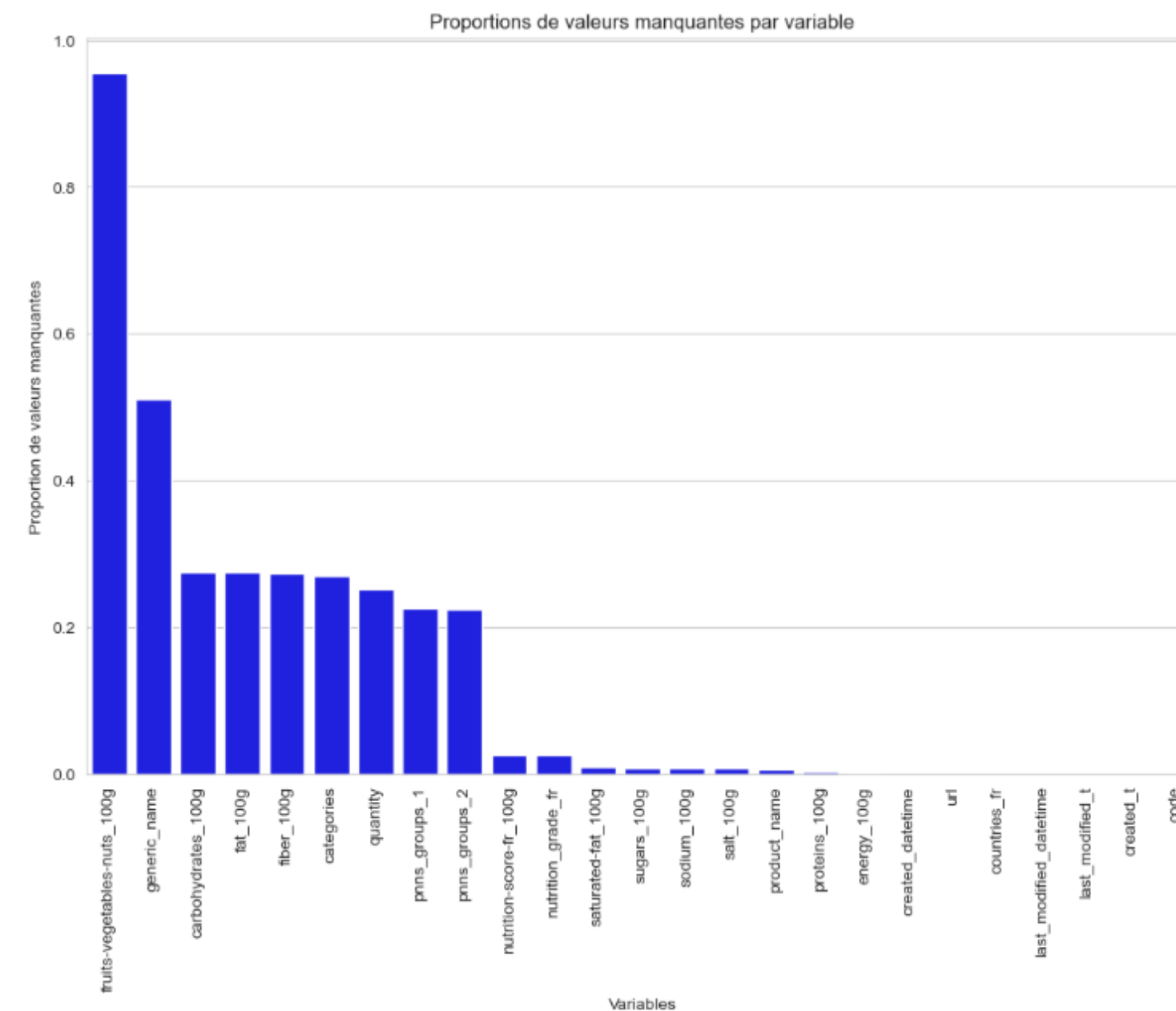
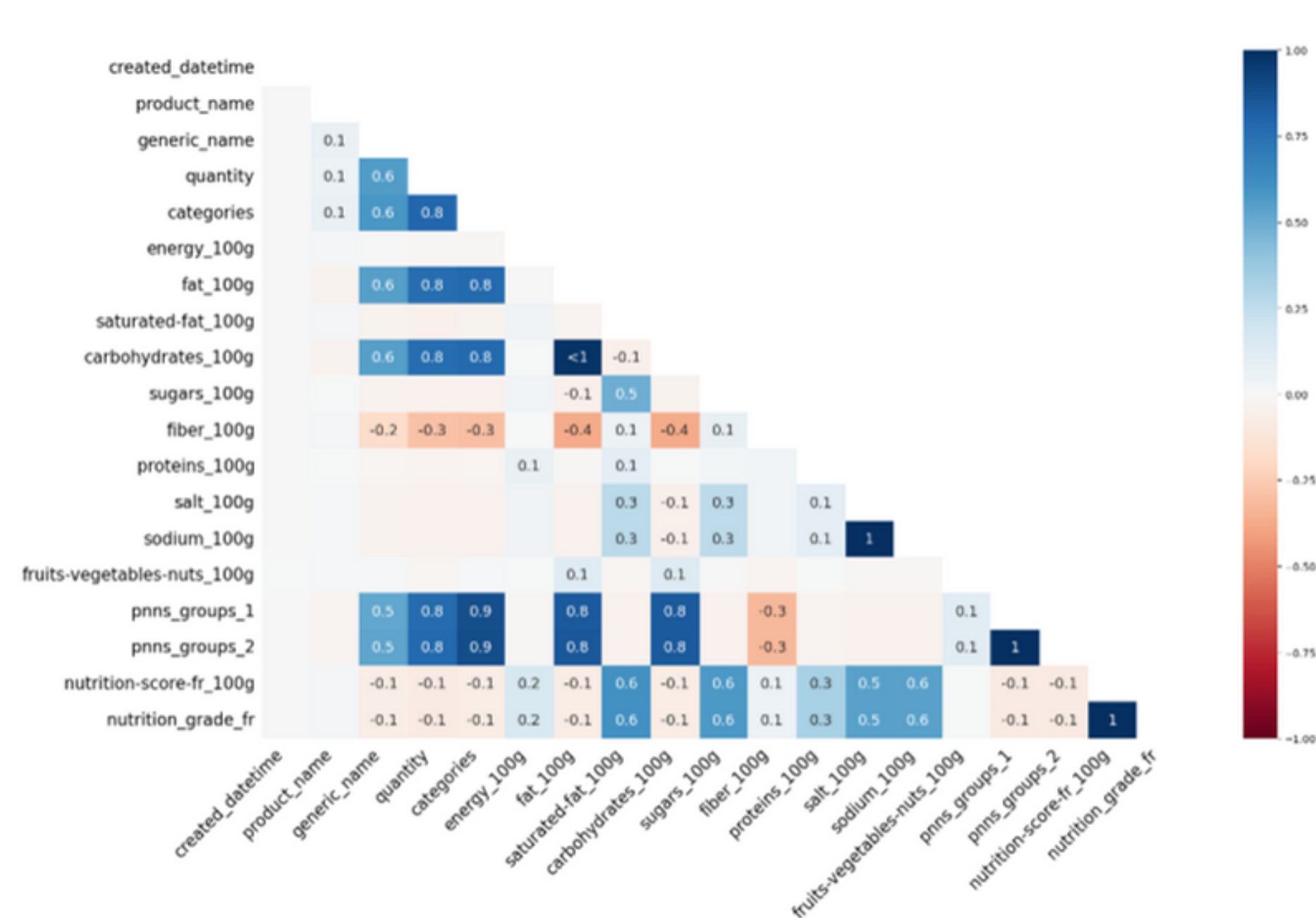


- Observation et visualisation globales des données (taille, type de données, valeurs manquantes)**

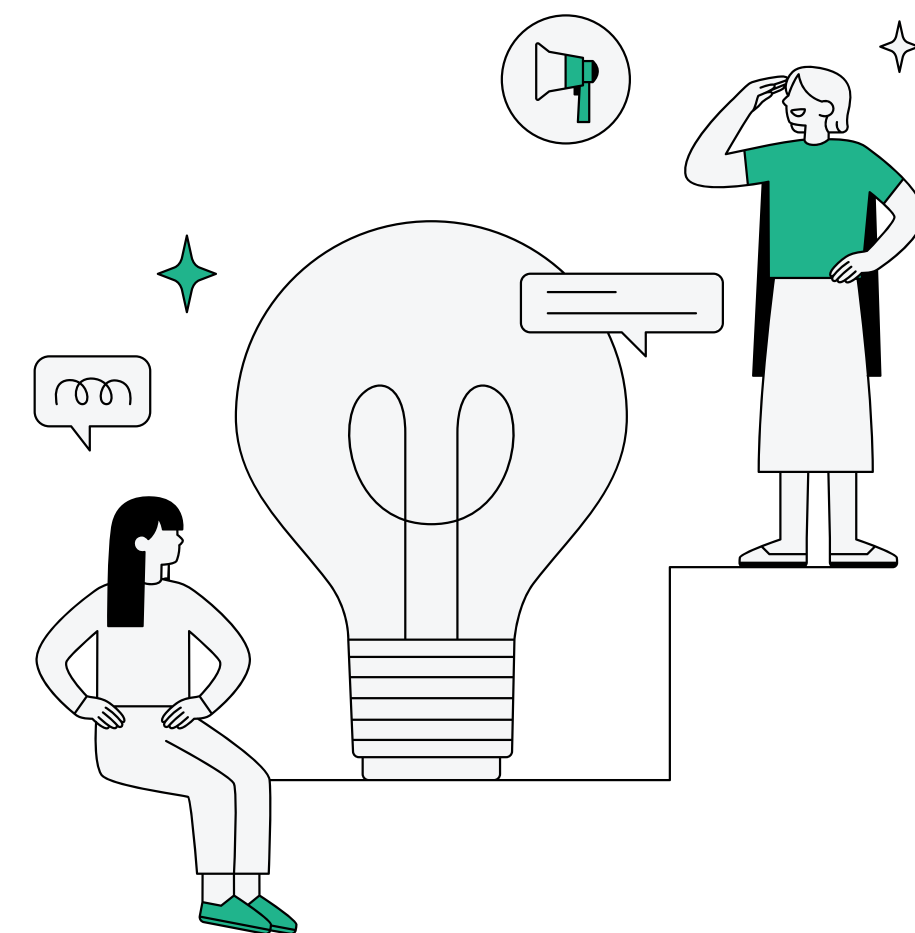
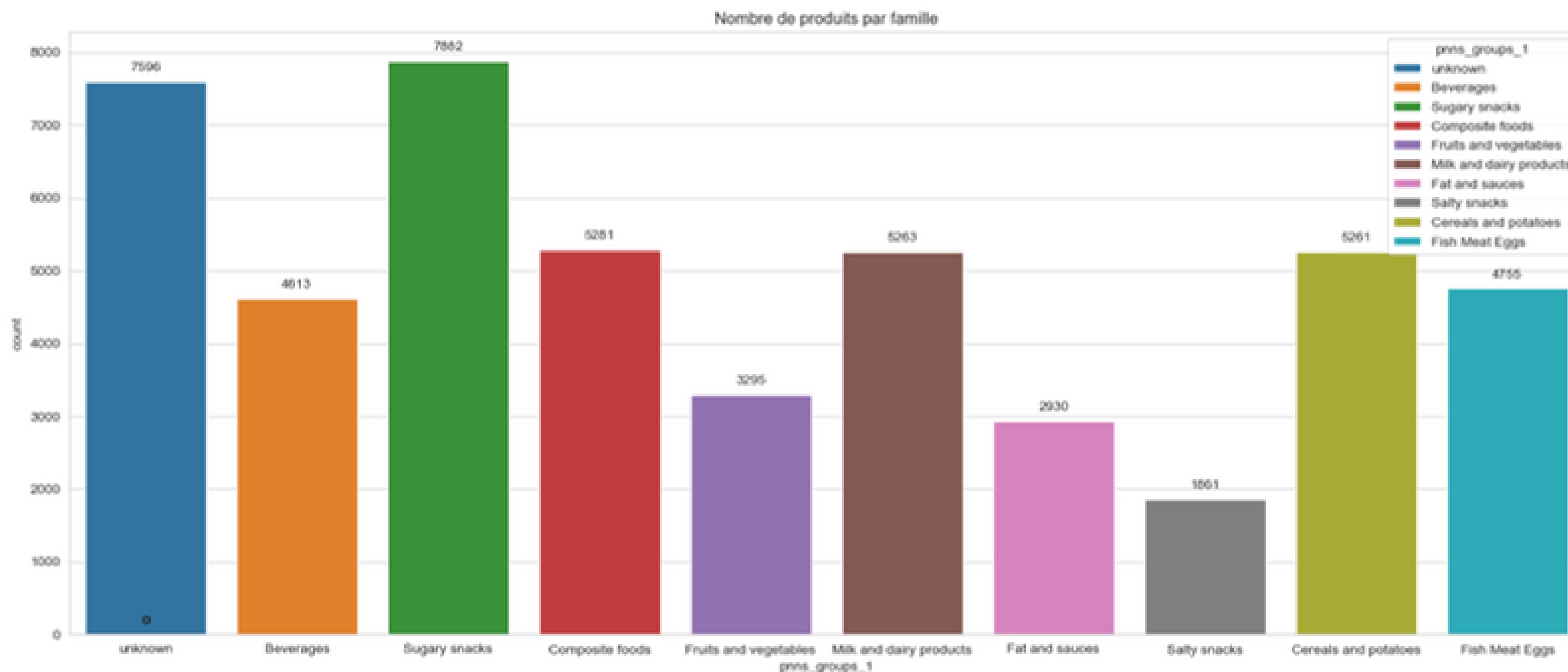
```
<class 'pandas.core.frame.DataFrame'>
Index: 62959 entries, 106 to 320763
Data columns (total 25 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   code                                  62959 non-null  object
1   url                                   62959 non-null  object
2   created_t                             62959 non-null  object
3   created_datetime                      62958 non-null  object
4   last_modified_t                       62959 non-null  object
5   last_modified_datetime                62959 non-null  object
6   product_name                          62595 non-null  object
7   generic_name                          30741 non-null  object
8   quantity                              47079 non-null  object
9   categories                            45945 non-null  object
10  countries_fr                          62959 non-null  object
11  energy_100g                           62917 non-null  float64
12  fat_100g                              45701 non-null  float64
13  saturated-fat_100g                   62359 non-null  float64
14  carbohydrates_100g                   45648 non-null  float64
15  sugars_100g                           62416 non-null  float64
16  fiber_100g                           45716 non-null  float64
17  proteins_100g                         62785 non-null  float64
18  salt_100g                            62468 non-null  float64
19  sodium_100g                          62467 non-null  float64
20  fruits-vegetables-nuts_100g          2827 non-null  float64
21  pnns_groups_1                         48737 non-null  object
22  pnns_groups_2                         48833 non-null  object
23  nutrition-score-fr_100g               61334 non-null  float64
24  nutrition_grade_fr                    61334 non-null  object
dtypes: float64(11), object(14)
memory usage: 12.5+ MB
None
```

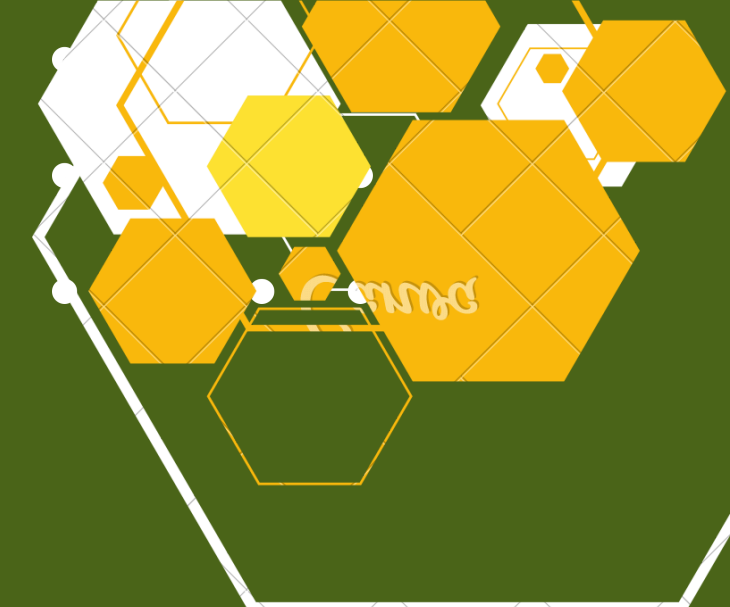


- **Observation et visualisation globales des données (taille, type de données, valeur manquantes)**



- **Harmonisation des observations du `pnnns_groups_1` et `pnnns_groups_2`**
Pour la définir les familles de produits et être plus pertinent au niveau de l'analyse des nutriments.
- **Visualisation des produits par famille**



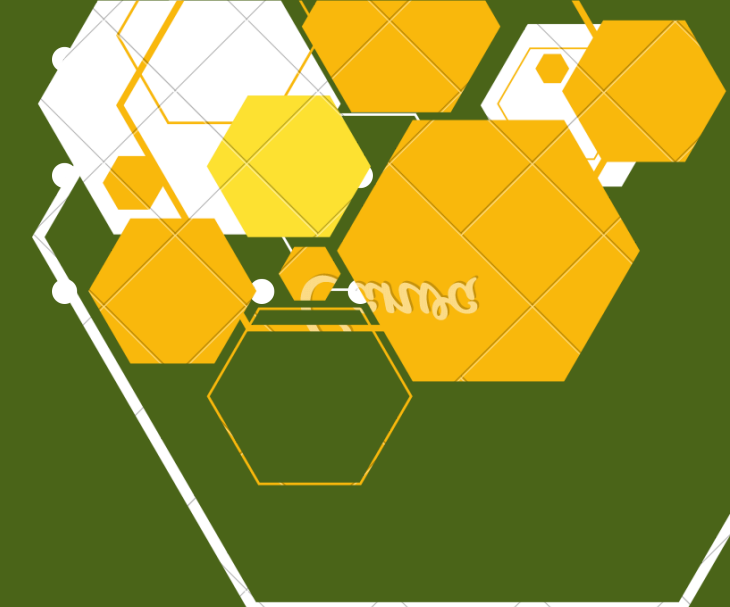


• Analyse des variables numériques

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	fruits-vegetables-nuts_100g	nutrition-score-fr_100g
count	6.291700e+04	45701.000000	62359.000000	45648.000000	62416.000000	45716.000000	62785.000000	62468.000000	62467.000000	2827.000000	61334.000000
mean	1.176158e+03	13.276054	5.423381	27.778450	13.378901	2.559466	7.737782	1.156485	0.455315	30.524609	8.694851
std	1.300515e+04	16.888191	8.531668	27.319808	19.026129	4.635102	7.871803	4.268056	1.680293	32.178404	9.046375
min	0.000000e+00	0.000000	0.000000	0.000000	-0.100000	0.000000	0.000000	0.000000	0.000000	0.000000	-15.000000
25%	4.300000e+02	1.300000	0.300000	4.100000	1.000000	0.000000	1.800000	0.080000	0.031496	0.000000	1.000000
50%	1.038000e+03	6.800000	2.000000	14.700000	4.000000	1.380000	6.000000	0.560000	0.220472	19.900000	9.000000
75%	1.653000e+03	21.000000	7.400000	53.000000	17.600000	3.200000	10.800000	1.244600	0.490000	50.000000	15.000000
max	3.251373e+06	380.000000	210.000000	190.000000	104.000000	178.000000	100.000000	211.000000	83.000000	100.000000	40.000000

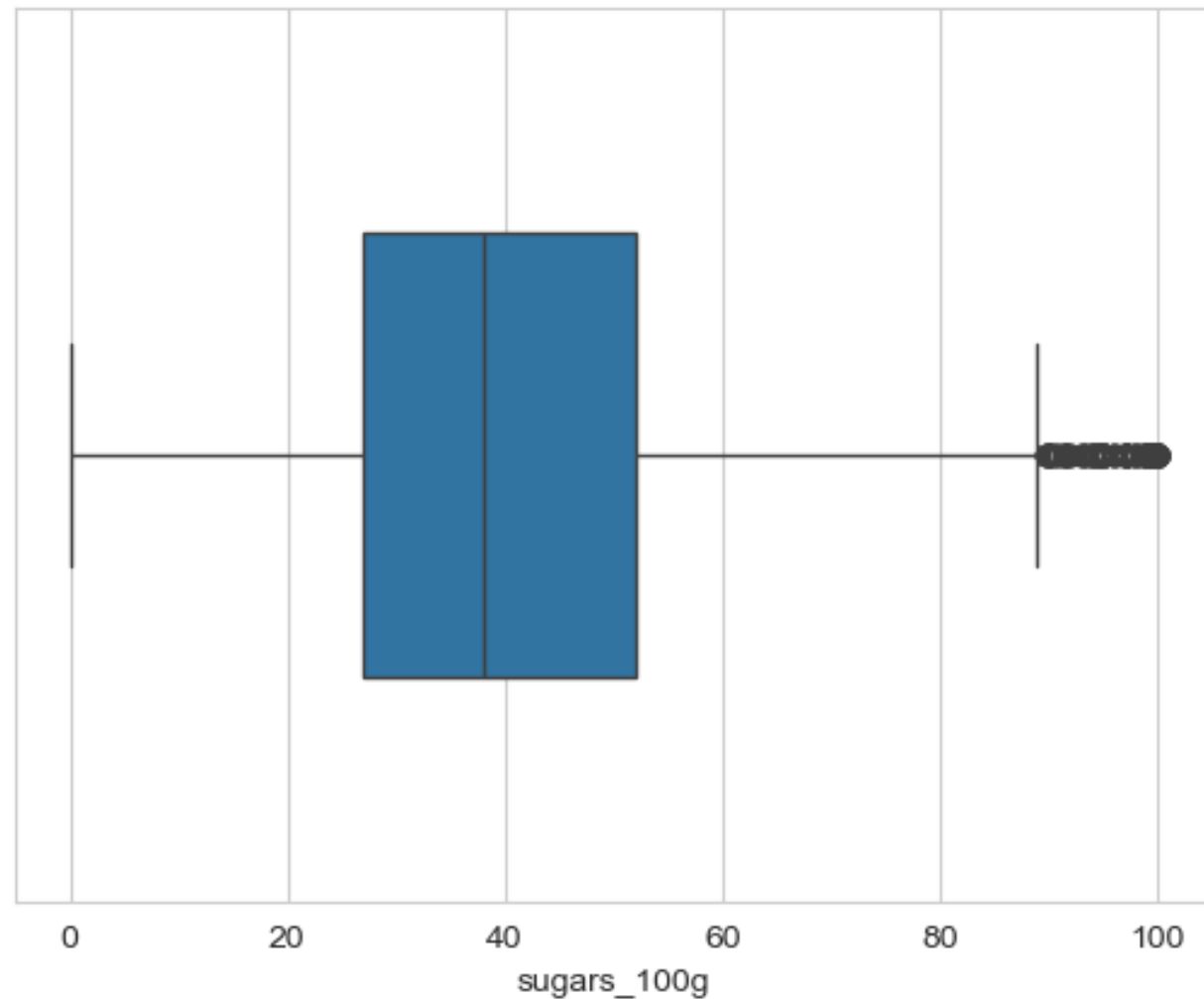
Quelques observations métier :

- Les nutriments dont les valeurs nutritionnelles sont définies pour 100g et ayant unité de mesure le g, ces valeurs ne peuvent qu'être comprises entre 0g et 100g.
- Concernant la variable energy_100g la valeur ici est donnée en KJ/100g. Dans 100g d'aliment, nous pouvons avoir plus de 100KJ mais pas plus de 480 Kcal soit 2008.32 KJ.
- La teneur en saturated-fat_100g ne peut être supérieure à celle de fat_100g car ce dernier regroupe l'ensemble des fat y compris saturated-fat.
- La teneur en sugars_100g ne peut être supérieure à celle de carbohydrates car ce dernier regroupe l'ensemble des glucides y compris sugars.



- **Visualisation des boxplot par famille**

Sugary snacks



- **Analyse via la méthode IQR**

$q1$ = 1er quartile, $q3$ = 3ème quartile

$iqr = q3 - q1$

limite inférieure = $q1 - 1.5 * iqr$

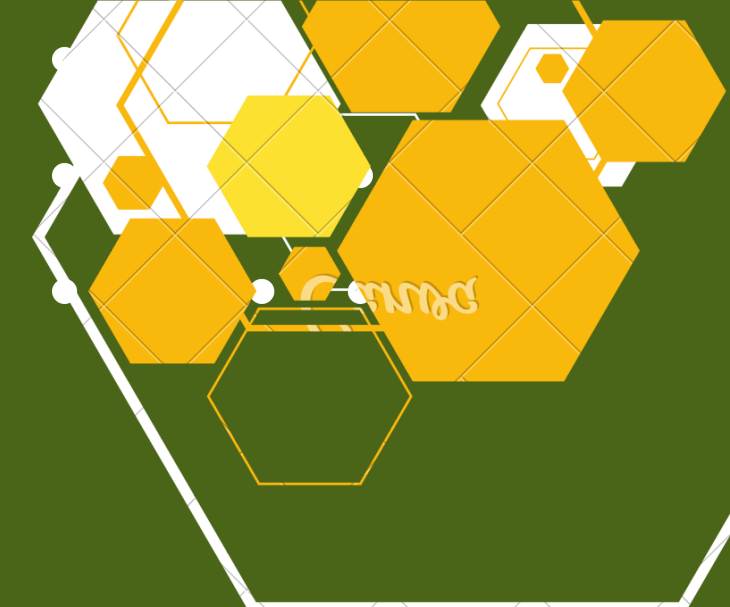
limite supérieure = $q3 + 1.5 * iqr$

- **Exemple : variable `sugars_100`, famille `Sugary snacks`**

Sugary snacks

`q1: 27.0 q3: 52.0 iqr: 25.0 lower_whisker: -10.5 upper_whisker: 89.5`

	product_name	generic_name	categories	sugars_100g
99449	Rainbow Nerds	NaN	Snacks sucrés,Confiseries,Bonbons	93.3
99452	Wonka Nerds Starwberry And Grape Theatre	NaN	Bonbons	98.0
178777	Bonbons fruités cœur liquide	NaN	Bonbons	92.9
179003	sucré cassonade	pure canne	Sucres roux	99.0
179053	Sucres en morceaux	NaN	Édulcorants,Sucres,Sucres en morceaux	100.0



- **Valeurs aberrantes statistiques**

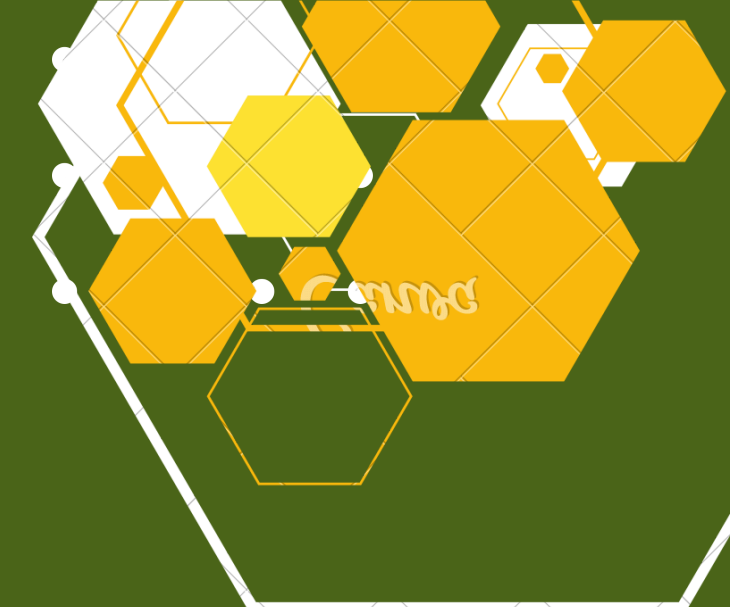
Les outliers identifiés statistiquement ne sont pas des valeurs aberrantes compte tenu des spécificités métier mais plutôt des valeurs atypiques. Nous n'allons par conséquent pas appliqué la méthode statistique pour traiter ces valeurs identifiées comme aberrantes.

- **Traitement des valeurs aberrantes uniquement selon les spécificités métier**

Variables	Traitement apporté
energy_100g	Si la valeur > 2008.32, remplacer par None pour traitement comme valeur manquante
saturated-fat_100g	Si la valeur > 0, remplacer par None pour traitement comme valeur manquante Si > fat_100g, remplacer par None pour traitement comme valeur manquante
sugars_100g	Si la valeur < 0 ou > 100, remplacer par None pour traitement comme valeur manquante Si > carbohydrates_100g, remplacer par None pour traitement comme valeur manquante
fiber_100g	Si la valeur > 0, remplacer par None pour traitement comme valeur manquante

Analyses exploratoires

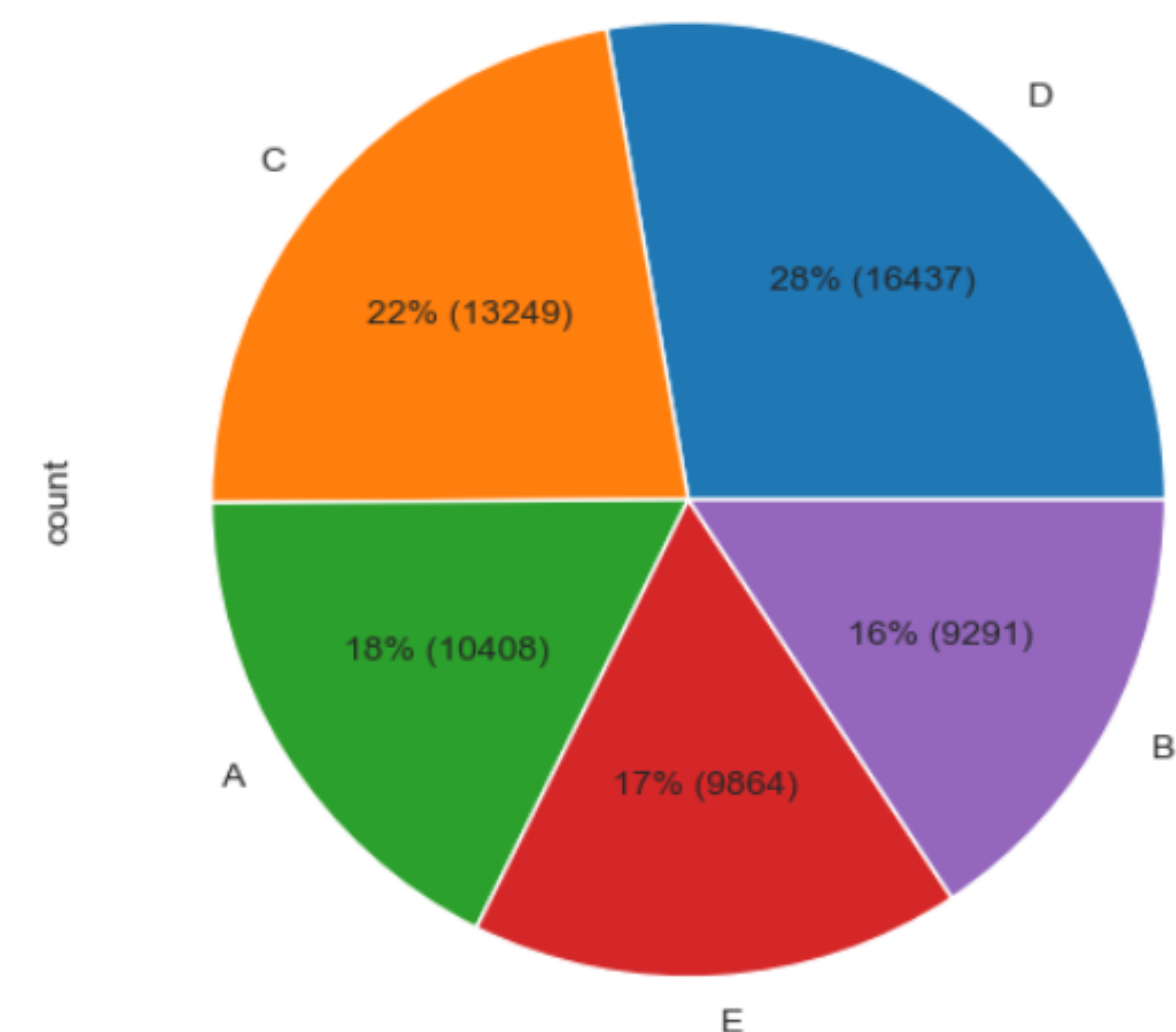
Traitement des valeurs manquantes

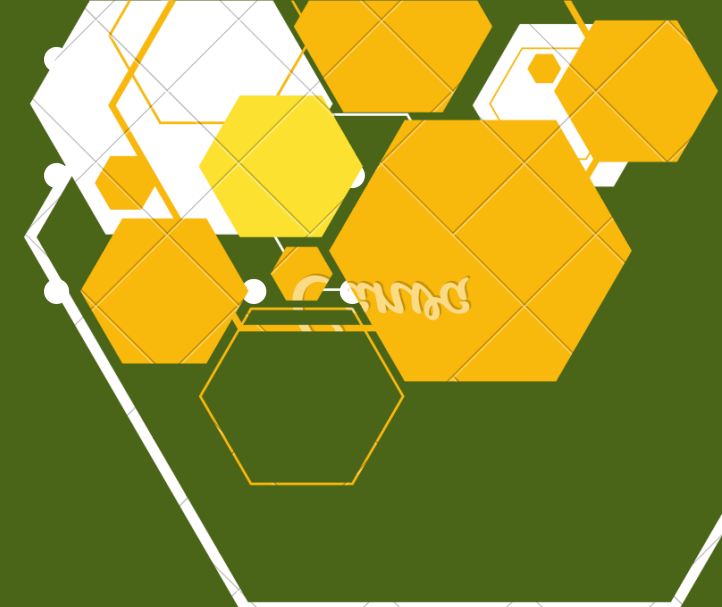


Variables	Traitement apporté à la valeur manquante identifiée dans la variable
created_datetime	Attribution du last_modified_datetime
product_name	Suppression car impossibilité de fournir une information à l'utilisateur
generic_name	Attribution du product_name pour éviter une perte importante de données
quantity	Marquer comme inconnu "unknown", information pas indispensable dans notre solution
fiber_100g	Affecter à 0, donnée prisée par le consommateur et très mise en avant par les distributeurs. En cas d'absence, nous considérons que le produit n'en contient pas.
fruits-vegetables-nuts_100g	Nutriment recommandé pour les consommateurs. Pour les produits de la famille Fruits and vegetables, nous imputons la moyenne de la famille, pour les autres familles nous imputons 0.
sodium_100g	Si la teneur en salt_100g = 0, nous imputons 0 Si la teneur en salt_100g est différente de 0, nous imputons salt_100g / 2.5. Ce sont des spécificités métier
pnns_groups_1	C'est la variable qui nous permet d'identifier les familles. Une analyse des correspondances avec le pnns_groups_2 nous permet de compléter les valeurs manquantes du pnns_groups_1 quand le pnns_groups_2 est connu. Lorsque le pnns_groups_2 n'est pas connu nous imputons "unknown".
energy_100, sugars_100g proteins_100, sodium_100g saturated-fat_100g	Imputation de la valeur manquante par la moyenne de la famille(pnns_groups_1) du produit concerné. Suppression des valeurs manquantes résiduelles (produits dont l'affectation à une famille est impossible)
nutrition-score-fr_100g	Imputation via KNN, moyenne des 20 plus proches voisins
nutrition_grade_fr	Le nutrition-score étant désormais connu, application du système de détermination du nutrition_grade pour imputer sachant leur nutrition-score.

Visualisation nombre et proportion des produits par grade

Nombre et proportion des produits par grade





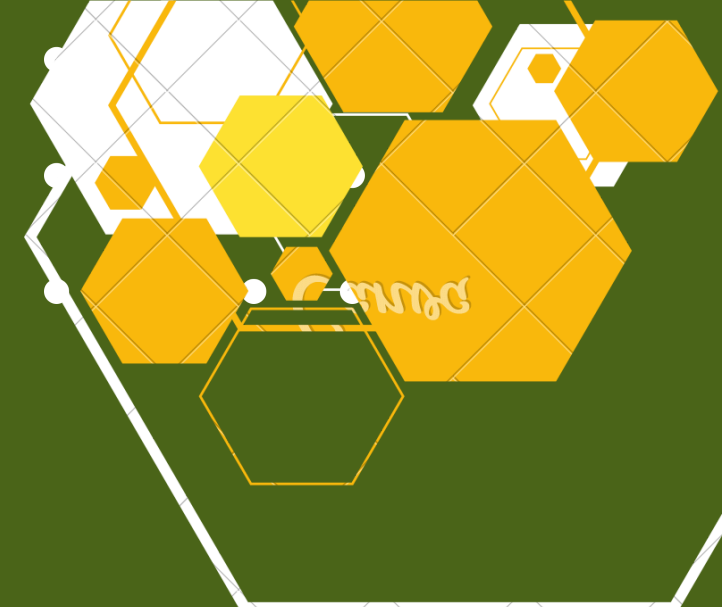
- **Le dataframe final : 9 variables et 59249 produits**

	energy_100g	saturated-fat_100g	sugars_100g	fiber_100g	proteins_100g	sodium_100g	fruits-vegetables-nuts_100g	nutrition-score-fr_100g	nutrition_grade_fr
0	1883.0	12.5	57.5	2.5	2.5	0.038000	0.0	22.0	E
1	1753.0	0.8	87.7	0.9	0.6	0.003937	0.0	14.0	D
2	177.0	0.0	10.4	0.0	0.0	0.010000	0.0	13.0	E
3	1079.0	11.0	1.0	1.4	7.5	0.314961	0.0	15.0	D
4	177.0	0.0	10.4	0.0	0.0	0.039370	0.0	13.0	E

- **Description des variables**

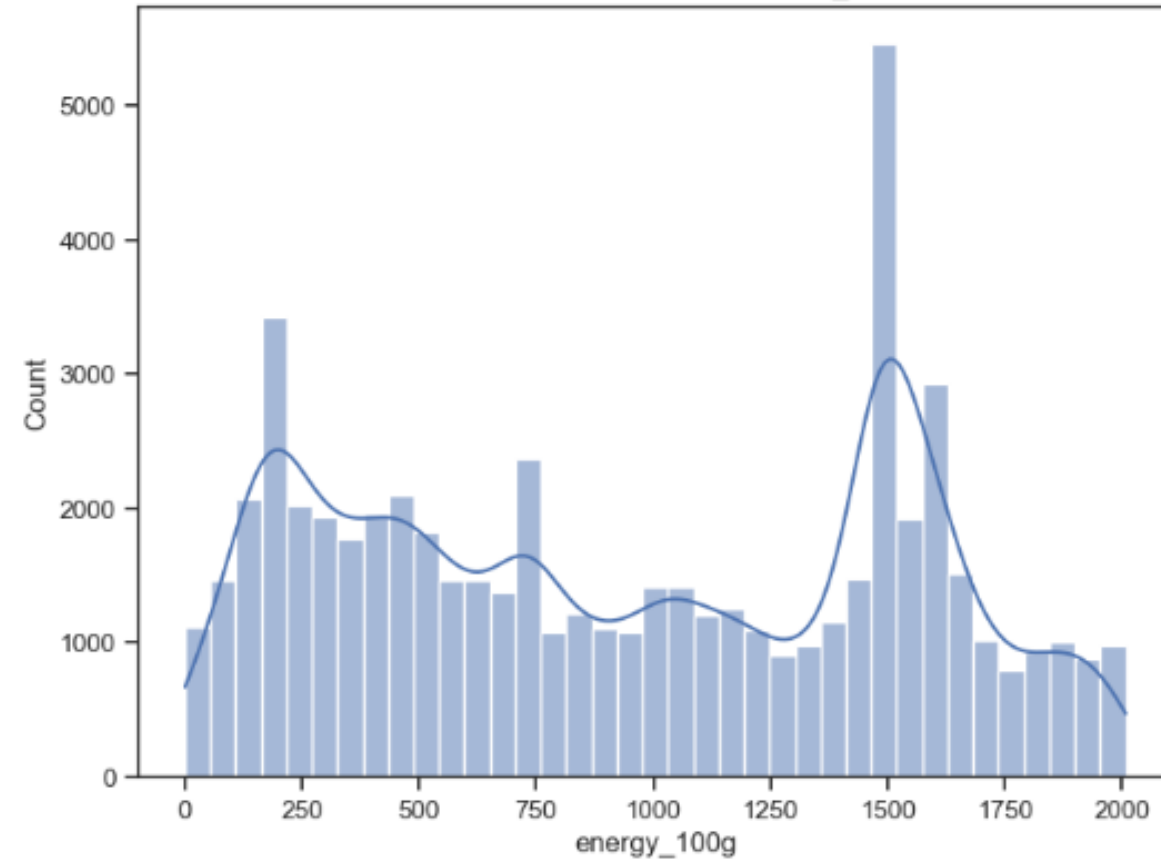
	energy_100g	saturated-fat_100g	sugars_100g	fiber_100g	proteins_100g	sodium_100g	fruits-vegetables-nuts_100g	nutrition-score-fr_100g
count	59249.000000	59249.000000	59249.000000	59249.000000	59249.000000	59249.000000	59249.000000	59249.000000
mean	940.232966	4.880625	12.864351	1.805627	7.681487	0.452837	4.747199	8.084643
std	577.607484	7.701774	18.756447	3.963776	7.857849	1.645478	16.879329	8.791658
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-15.000000
25%	406.000000	0.300000	1.000000	0.000000	1.700000	0.031496	0.000000	1.000000
50%	914.000000	1.700000	4.000000	0.100000	5.900000	0.228346	0.000000	8.000000
75%	1486.271803	6.500000	15.800000	2.300000	10.900000	0.492126	0.000000	15.000000
max	2008.000000	100.000000	100.000000	100.000000	100.000000	83.000000	100.000000	40.000000

nutrition_grade_fr	
count	59249
unique	5
top	D
freq	16437

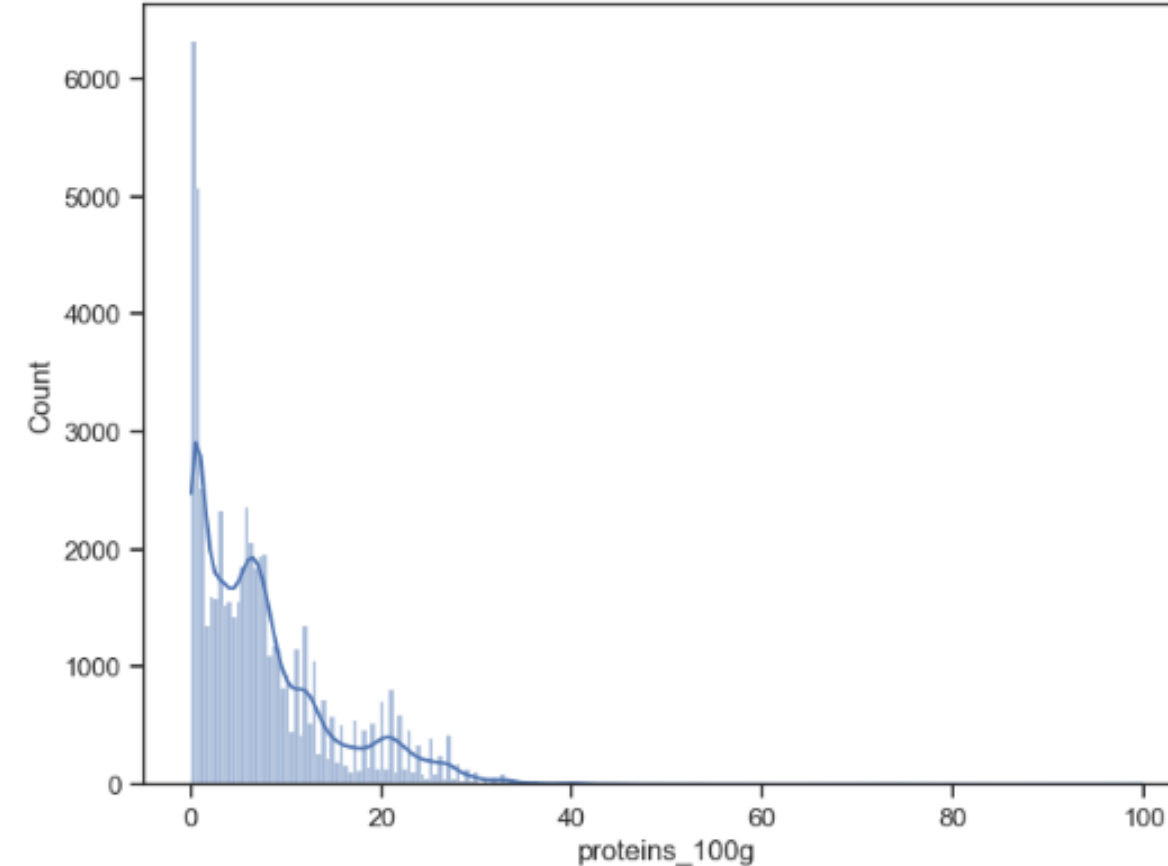


- Visualisation de quelques histogrammes

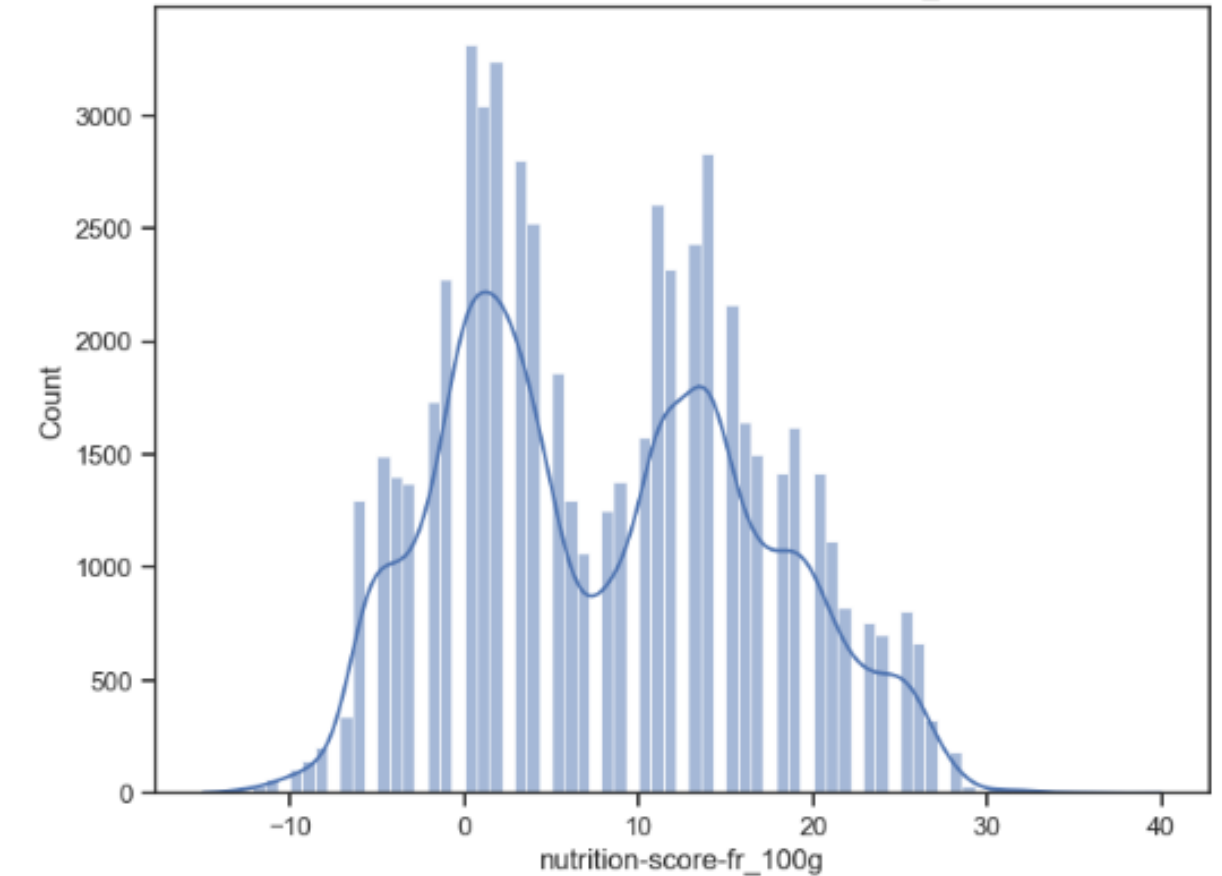
Histogramme de la variable energy_100g



Histogramme de la variable proteins_100g

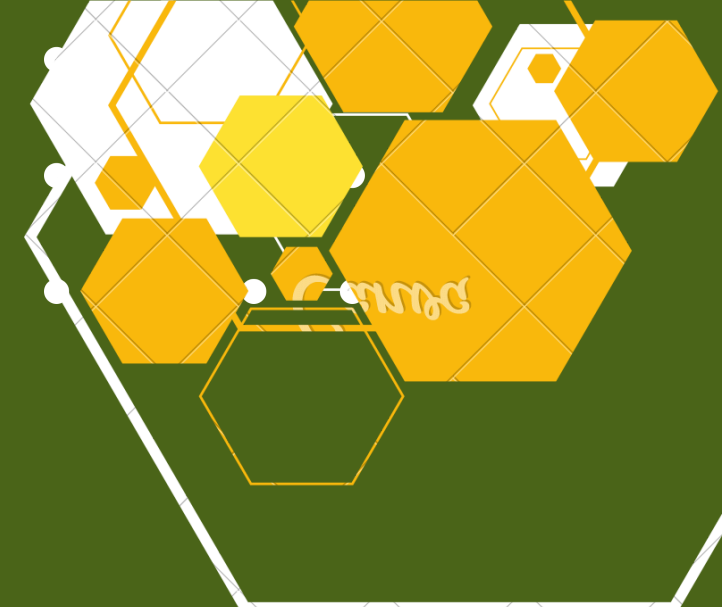


Histogramme de la variable nutrition-score-fr_100g

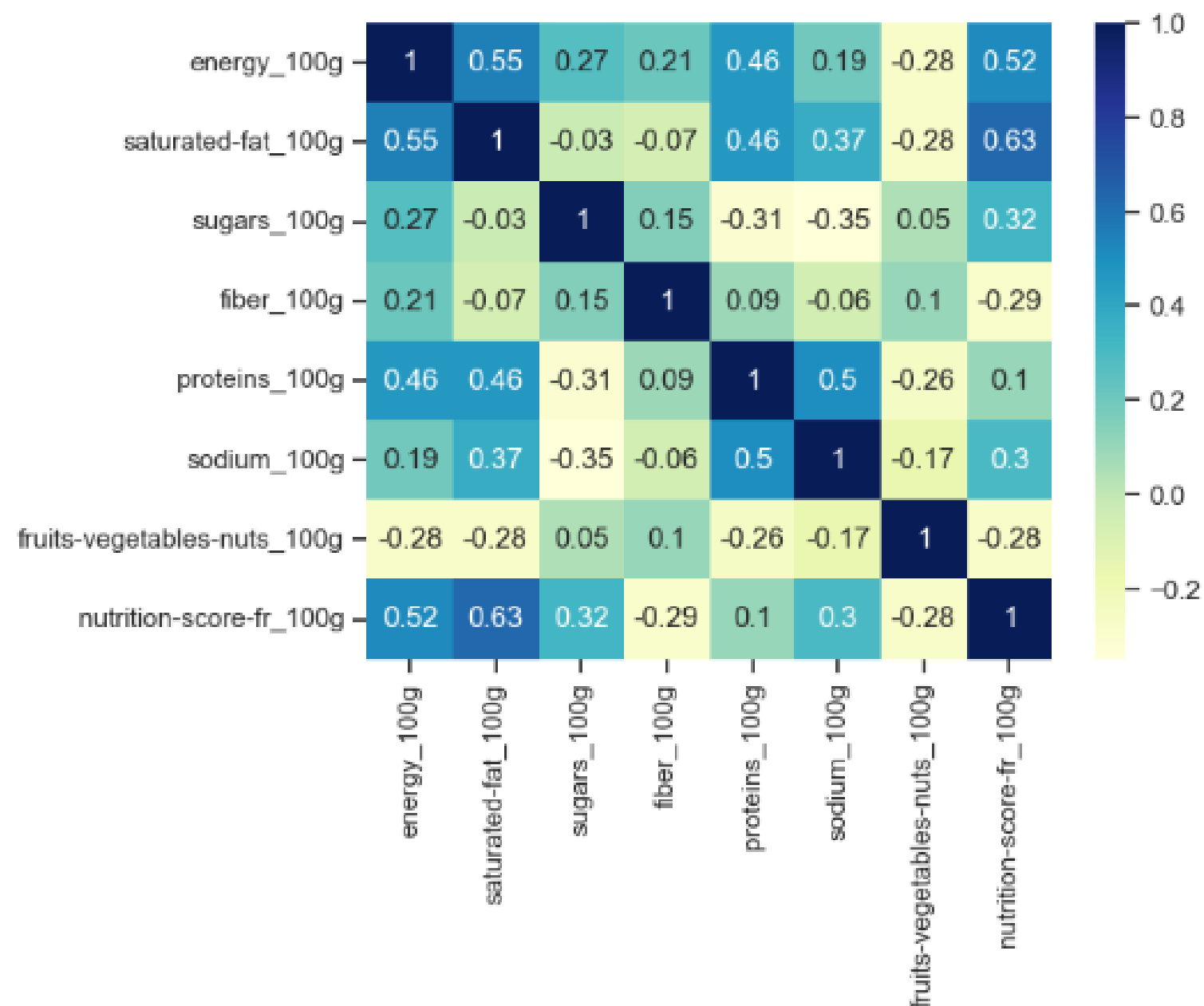


	Variable	Nom du test	Statistique de test	p-valeur	Normalité
0	energy_100g	Kolmogorov-Smirnov	9.922214e-01	0.0	Non
1	energy_100g	Anderson-Darling	1.273437e+03	NaN	Non
2	energy_100g	D'Agostino-Pearson	1.154580e+06	0.0	Non

	Variable	Nom du test	Statistique de test	p-valeur	Normalité
0	nutrition-score-fr_100g	Kolmogorov-Smirnov	0.693178	0.0	Non
1	nutrition-score-fr_100g	Anderson-Darling	599.226412	NaN	Non
2	nutrition-score-fr_100g	D'Agostino-Pearson	8273.143263	0.0	Non



- Matrix de corrélation de Spearman**



- Quelques observations**

==> Nous observons une bonne corrélation entre :

(energy et saturated-fat)

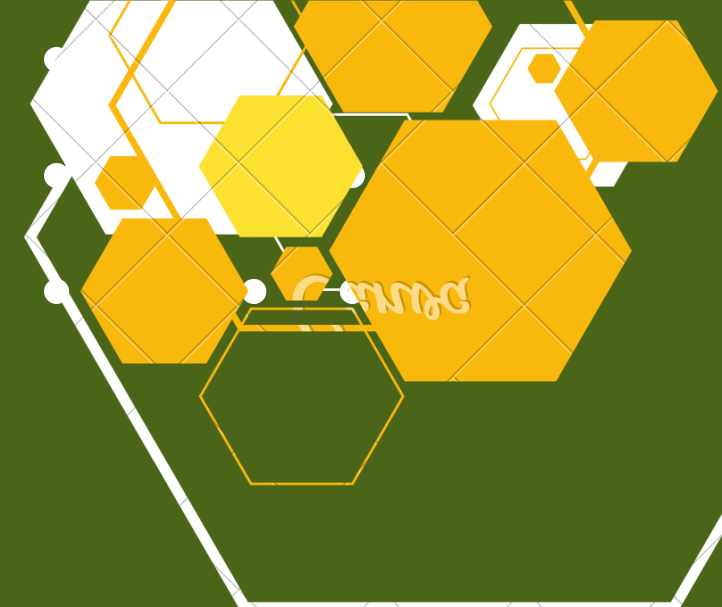
(energy et nutrition-score)

(saturated-fat et nutrition-score)

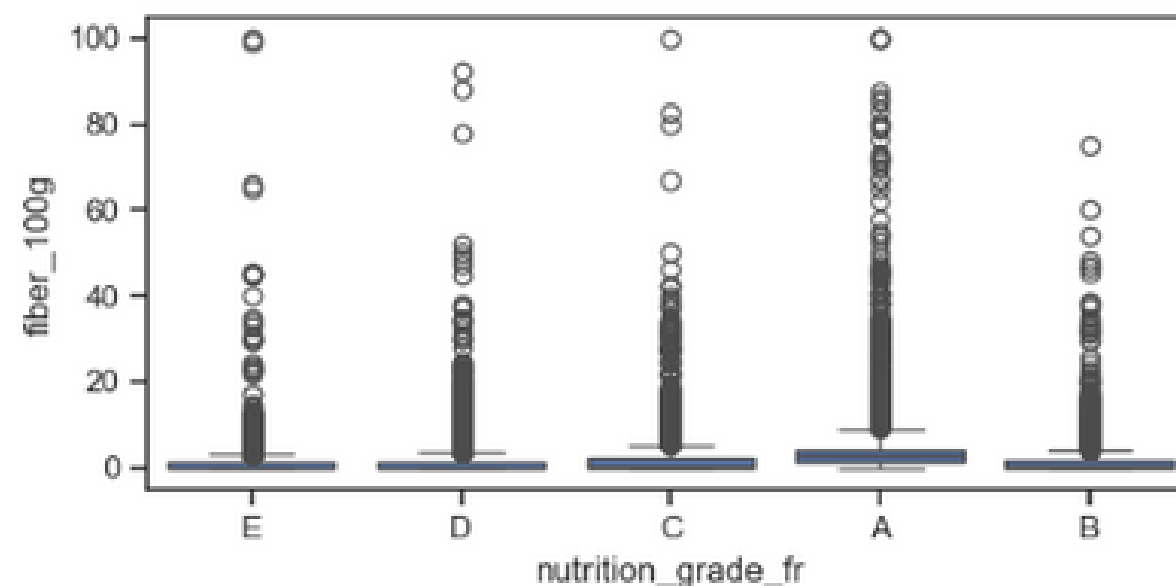
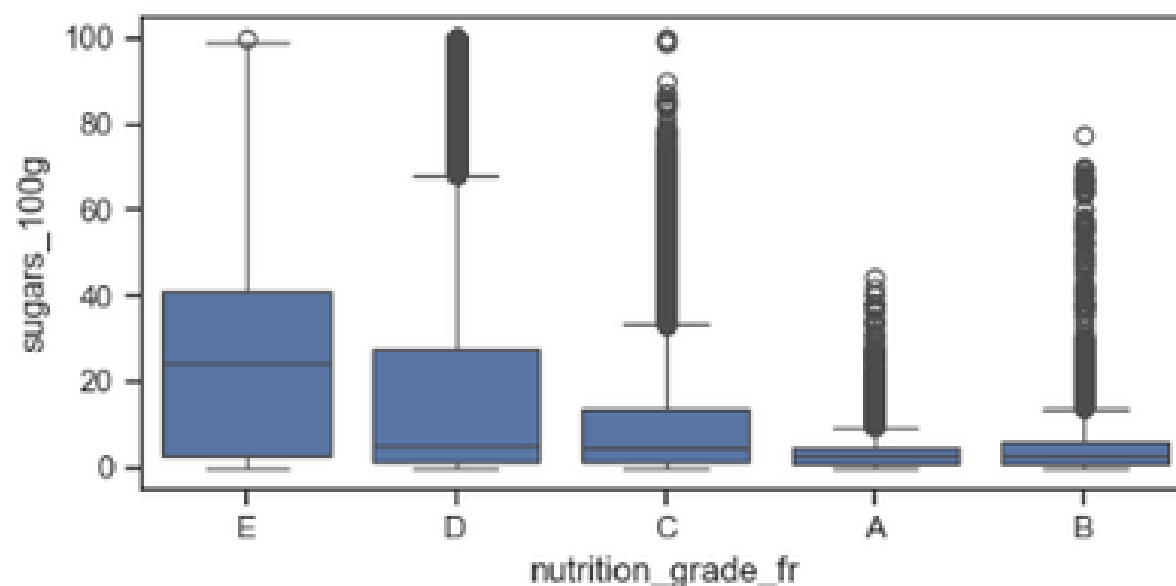
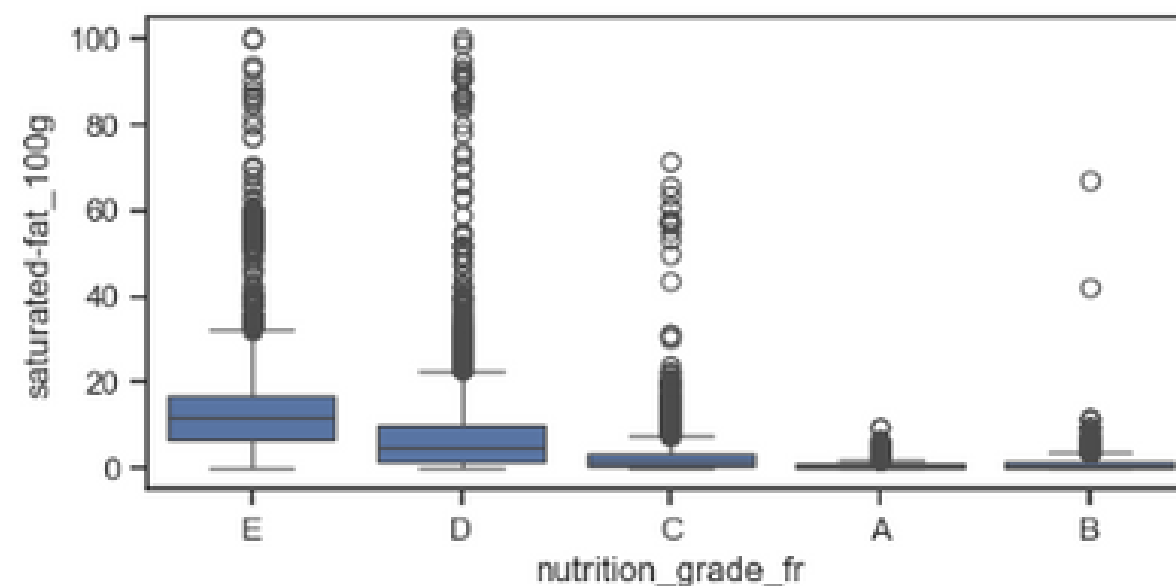
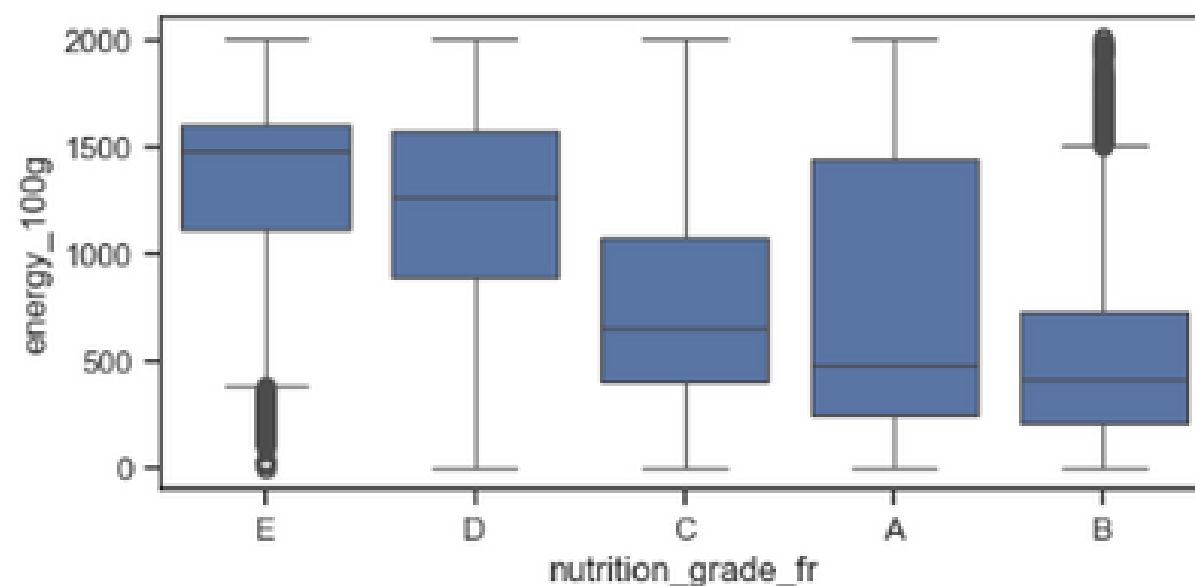
==> Nous observons une assez bonne corrélation entre :

(saturated-fat et proteins)

(sodium et proteins)

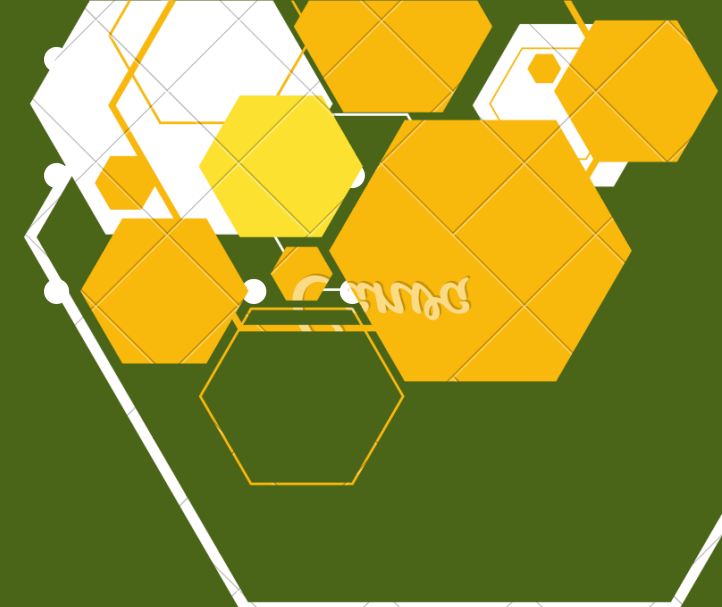


- **Analyse des relations entre les variables et le nutrition_grade des produits**

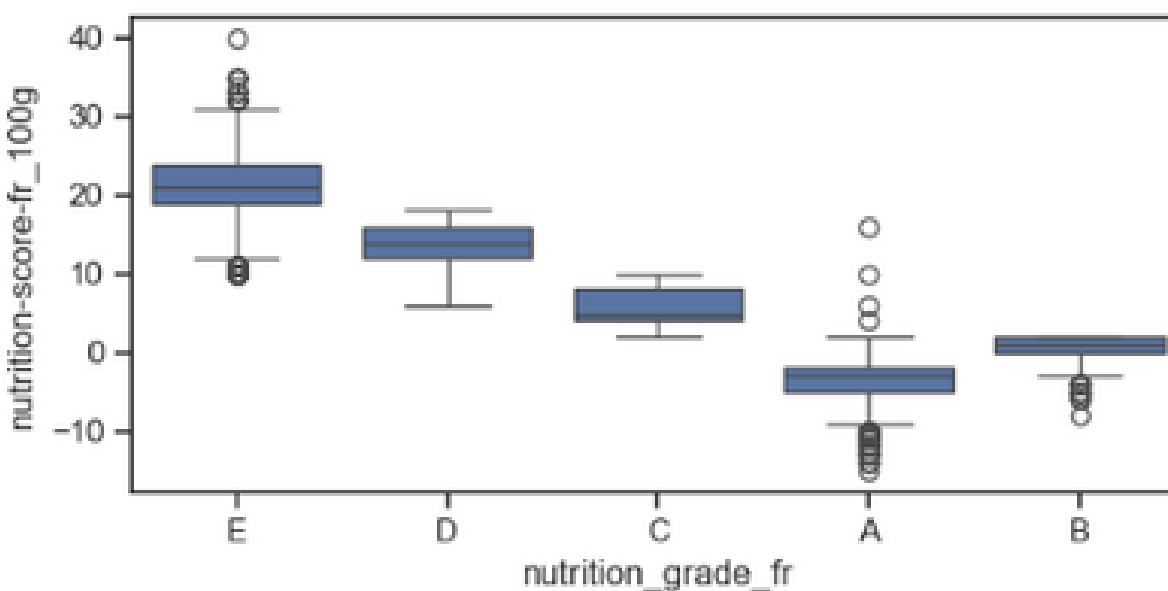
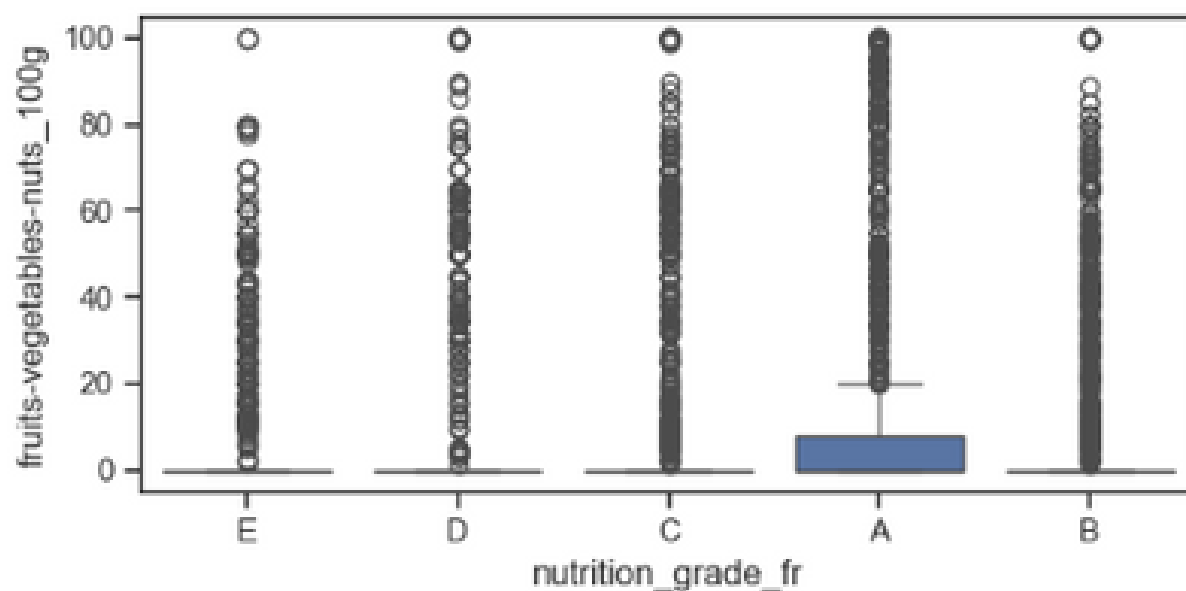
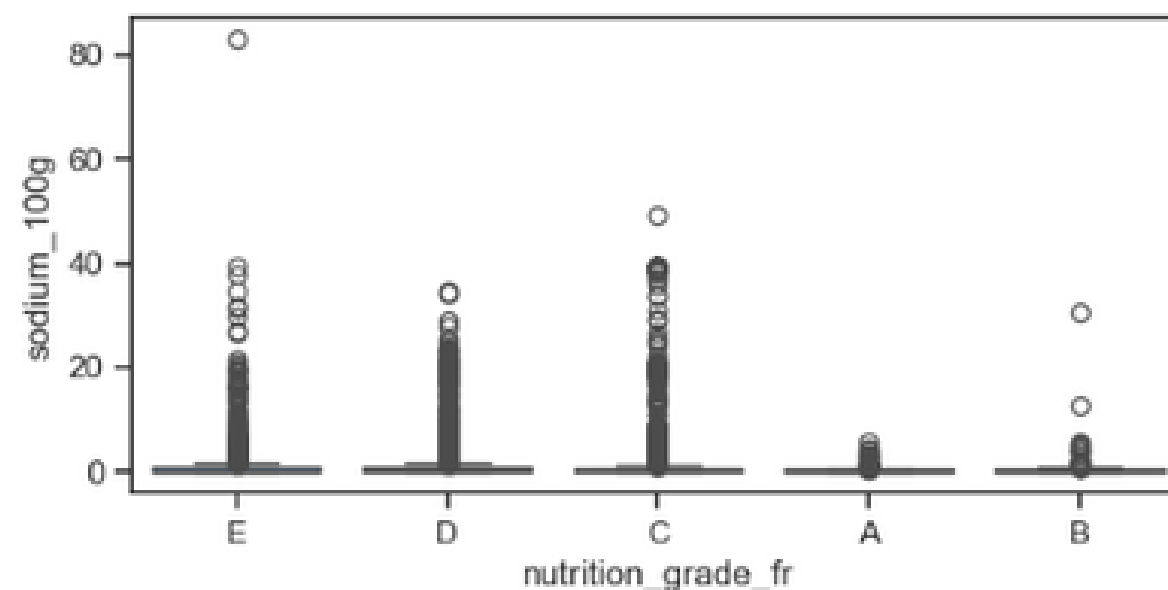
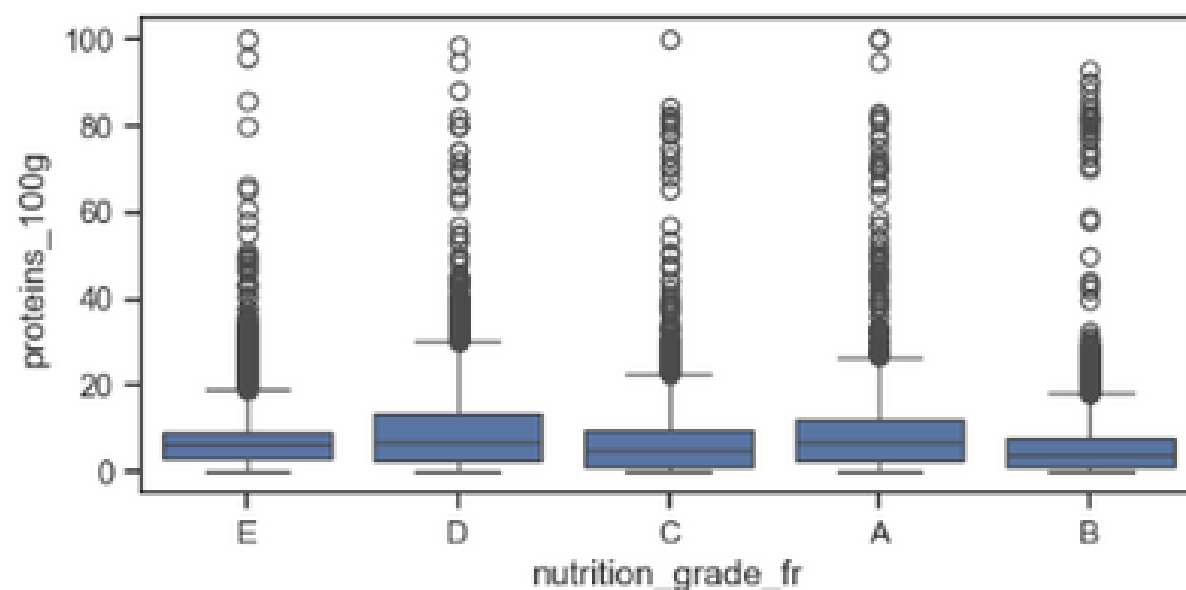


Quelques observations

- *energy / nutrition_grade* :
Les produits à forte teneur énergétique sont les moins bien classés en grade.
==> Plus la teneur en énergie est grande, moins le produit est sain pour la santé.
- *saturated-fat / nutrition_grade* :
==> Plus la teneur en matière grasse est élevée, moins le produit est sain pour la santé.
- *sugars / nutrition_grade* :
==> Plus la teneur en sucre est élevée, moins le produit est sain pour la santé.



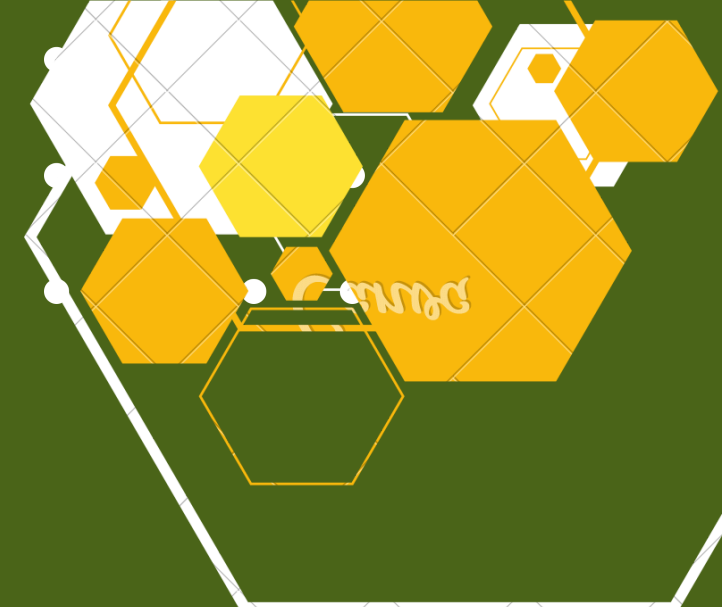
- **Analyse des relations entre les variables et le nutrition_grade des produits**



Quelsques observations

- *fruits-vegetables-nuts / nutrition_grade* :
==> Les fruits et légumes sont bénéfiques pour la santé.

- *nutrition-score / nutrition_grade* :
==> Plus le nutrition-score est élevé, moins bien le produit est classé. Ce qui confirme ce que nous avons appris dans le système de détermination du nutrition_grade.



Test de Kruskal-Wallis

Hypothèses:

- H_0 : Les distributions de tous les groupes sont égales.
- H_1 : Au moins une des distributions des groupes est différente des autres.

Conditions d'utilisation:

- Les observations sont indépendantes.
- Les données peuvent être ordonnées (au moins ordinale).

Interprétations:

- Si la p -valeur est inférieure au seuil de significativité (généralement 0,05), on rejette H_0 et on conclut qu'au moins une des distributions des groupes est significativement différente des autres

Conclusions:

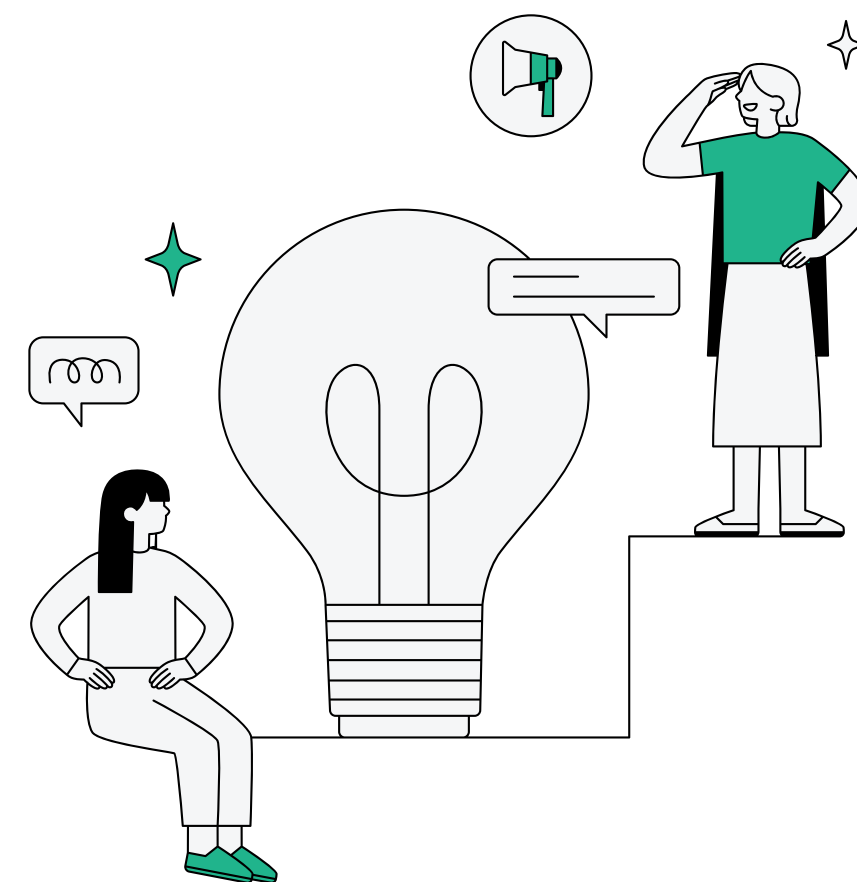
La variable energy_100g est significativement différente entre les grades (p-value=0.0)
 La variable saturated-fat_100g est significativement différente entre les grades (p-value=0.0)
 La variable sugars_100g est significativement différente entre les grades (p-value=0.0)
 La variable fiber_100g est significativement différente entre les grades (p-value=0.0)
 La variable proteins_100g est significativement différente entre les grades (p-value=5.931728993061146e-288)
 La variable sodium_100g est significativement différente entre les grades (p-value=0.0)
 La variable fruits-vegetables-nuts_100g est significativement différente entre les grades (p-value=0.0)
 La variable nutrition-score-fr_100g est significativement différente entre les grades (p-value=0.0)

Initialisation du test

La variable quantitative : nutrition-score-fr

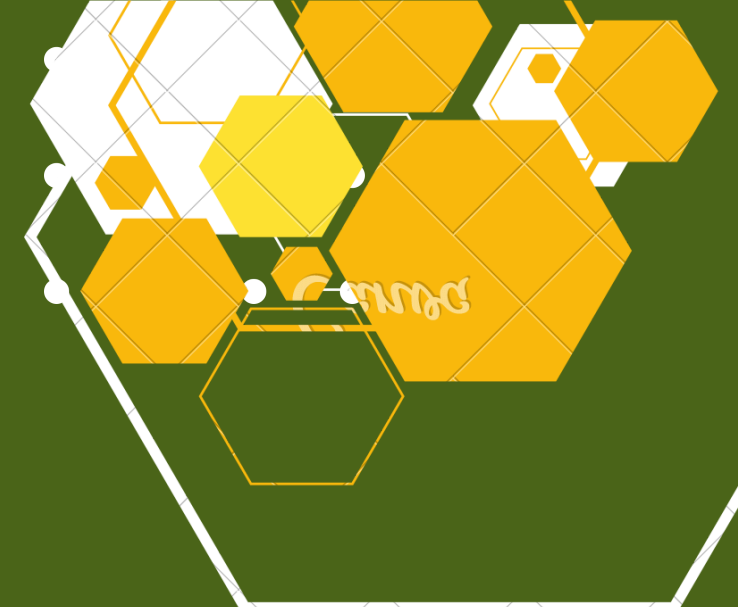
La variable qualitative : nutrition_grade

Les groupes formés sont faits à partir des différents grades (A à E)



Analyses bivariées

ANOVA



Test ANOVA (Analysis Of Variance)

Hypothèses :

- H_0 : Les moyennes de tous les groupes sont égales.
- H_1 : Au moins une des moyennes des groupes est différente des autres.

Conditions d'utilisation :

- La variable quantitative est normalement distribuée dans chaque groupe.
- Les variances des groupes sont égales.
- Les observations sont indépendantes.

Interprétations :

- Si la p-valeur est inférieure au seuil de significativité (généralement 0,05), on rejette H_0 et on conclut qu'au moins une des moyennes des groupes est significativement différente des autres.
- Si la p-valeur est inférieure au seuil de significativité (généralement 0,05), on rejette H_0 et on conclut que les distributions des deux groupes sont significativement différentes.

Initialisation du test

La variable quantitative : nutrition-score-fr

La variable qualitative : nutrition_grade

Les groupes formés sont faits à partir des différents grades (A à E)

Conclusions :

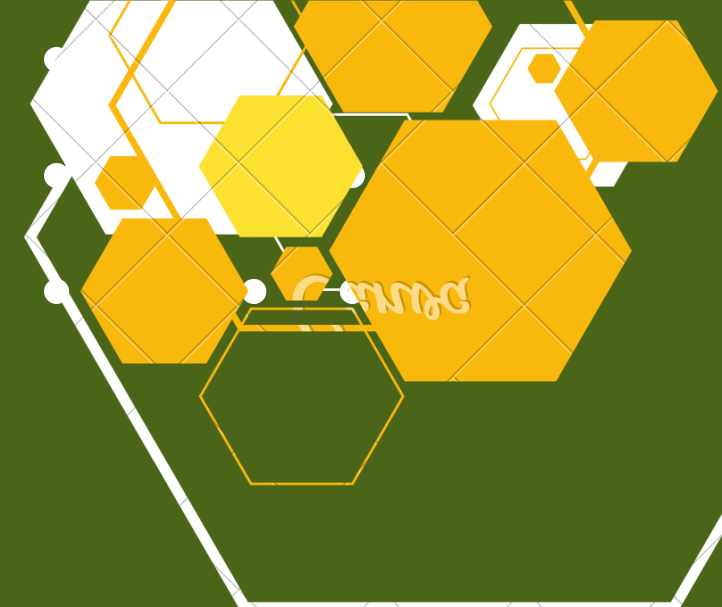
Test de normalité (Kolmogorov-Smirnov)

P-valeur de la normalité (Kolmogorov-Smirnov) - grade A : 0.0
P-valeur de la normalité (Kolmogorov-Smirnov) - grade B : 0.0
P-valeur de la normalité (Kolmogorov-Smirnov) - grade C : 0.0
P-valeur de la normalité (Kolmogorov-Smirnov) - grade D : 0.0
P-valeur de la normalité (Kolmogorov-Smirnov) - grade E : 0.0

ANOVA

Statistique F de l'ANOVA : 149574.59

P-valeur (ANOVA) : 0.0



Construction de l'ACP

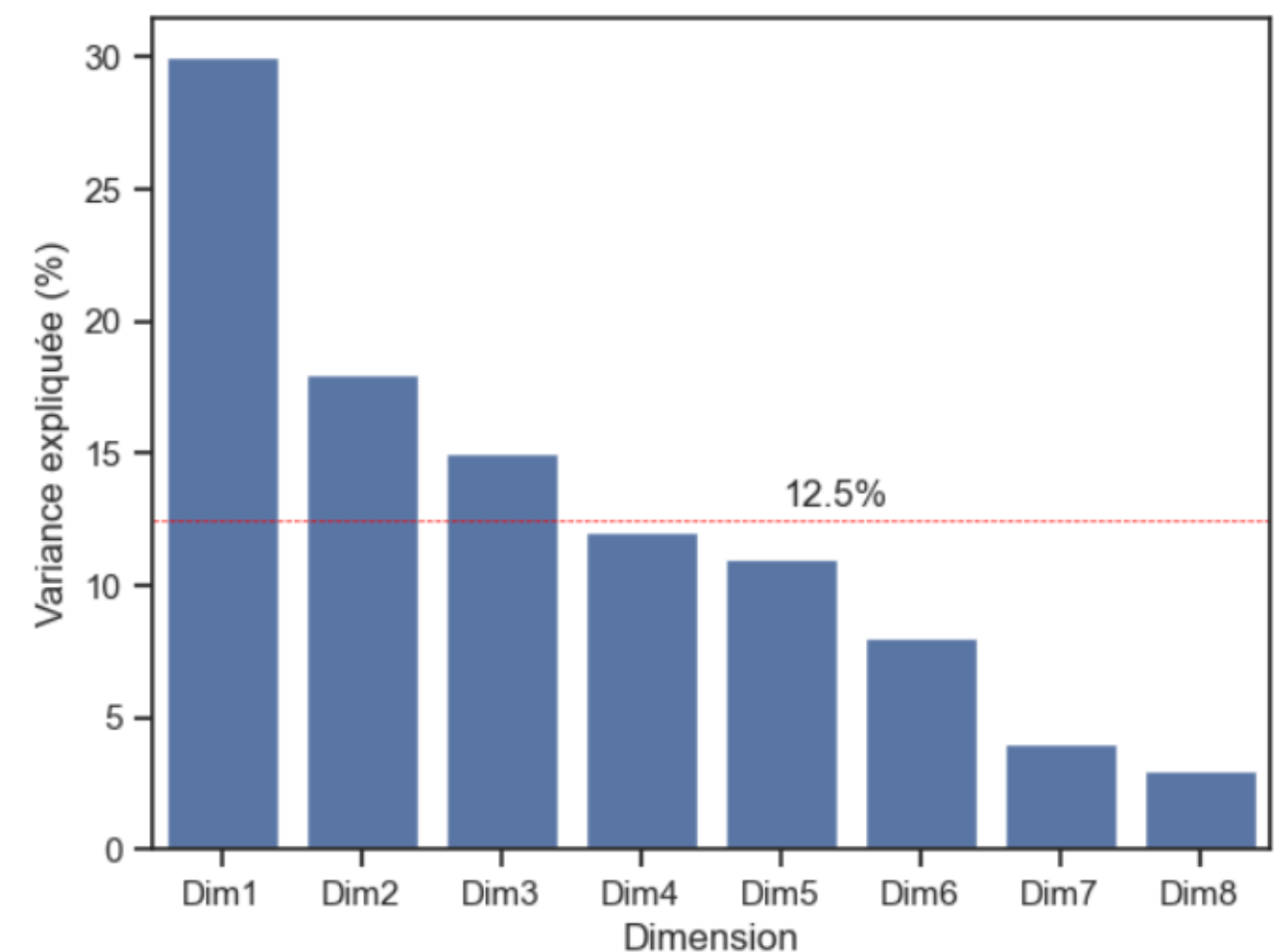
- Les variables explicatives sont constituées des 8 variables numériques : X
- La variable qualitative ici est le `nutrition_grade`, la dernière variable : y

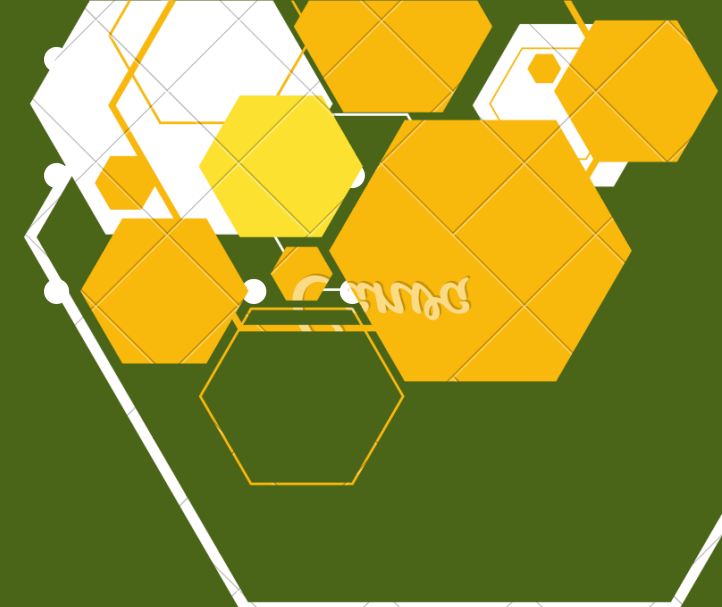
Choix du nombre d'axes

	Dimension	Valeur propre	% variance expliquée	% cum. var. expliquée
0	Dim1	2.399635	30.0	30.0
1	Dim2	1.404957	18.0	48.0
2	Dim3	1.173590	15.0	62.0
3	Dim4	0.960576	12.0	74.0
4	Dim5	0.855542	11.0	85.0
5	Dim6	0.656404	8.0	93.0
6	Dim7	0.333339	4.0	97.0
7	Dim8	0.216092	3.0	100.0

Scree plot : Eboulis des valeurs propres

Variance expliquée par dimension



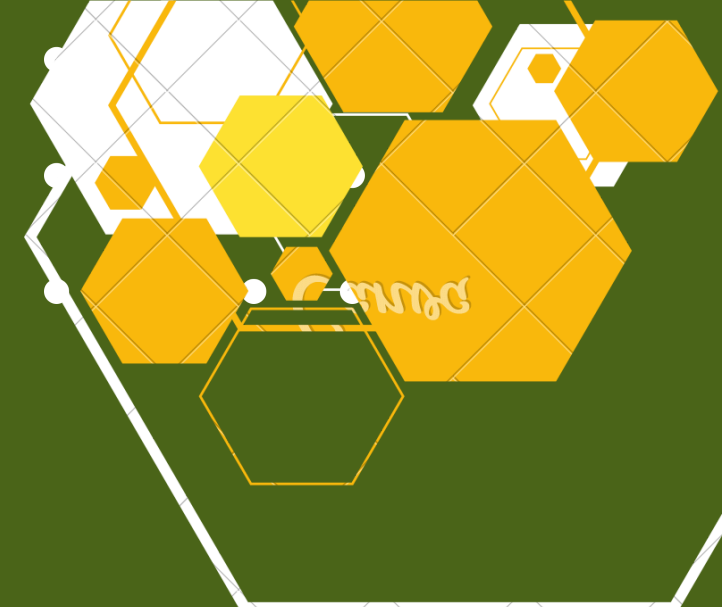


Représentation des variables sur chaque axe

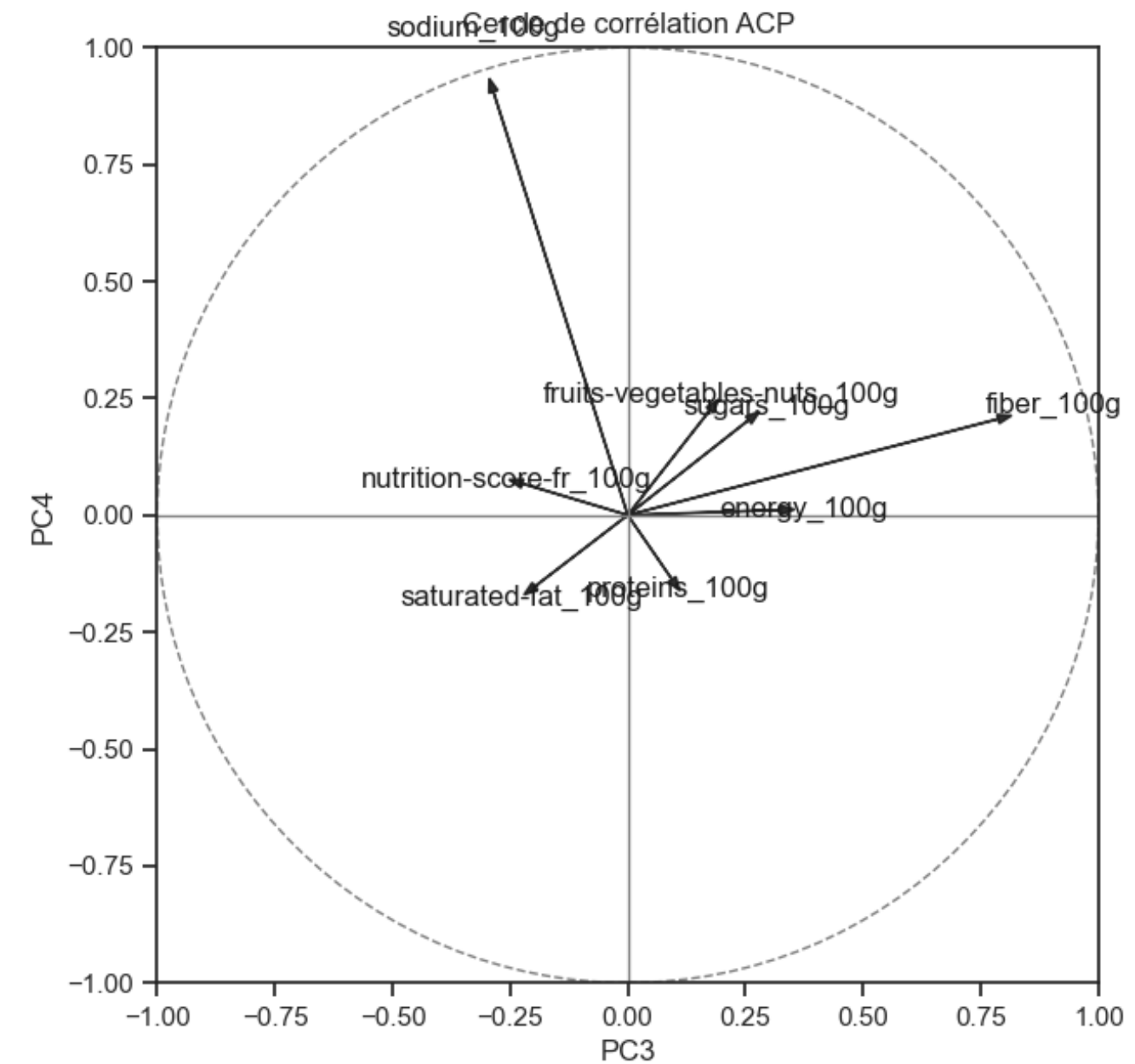
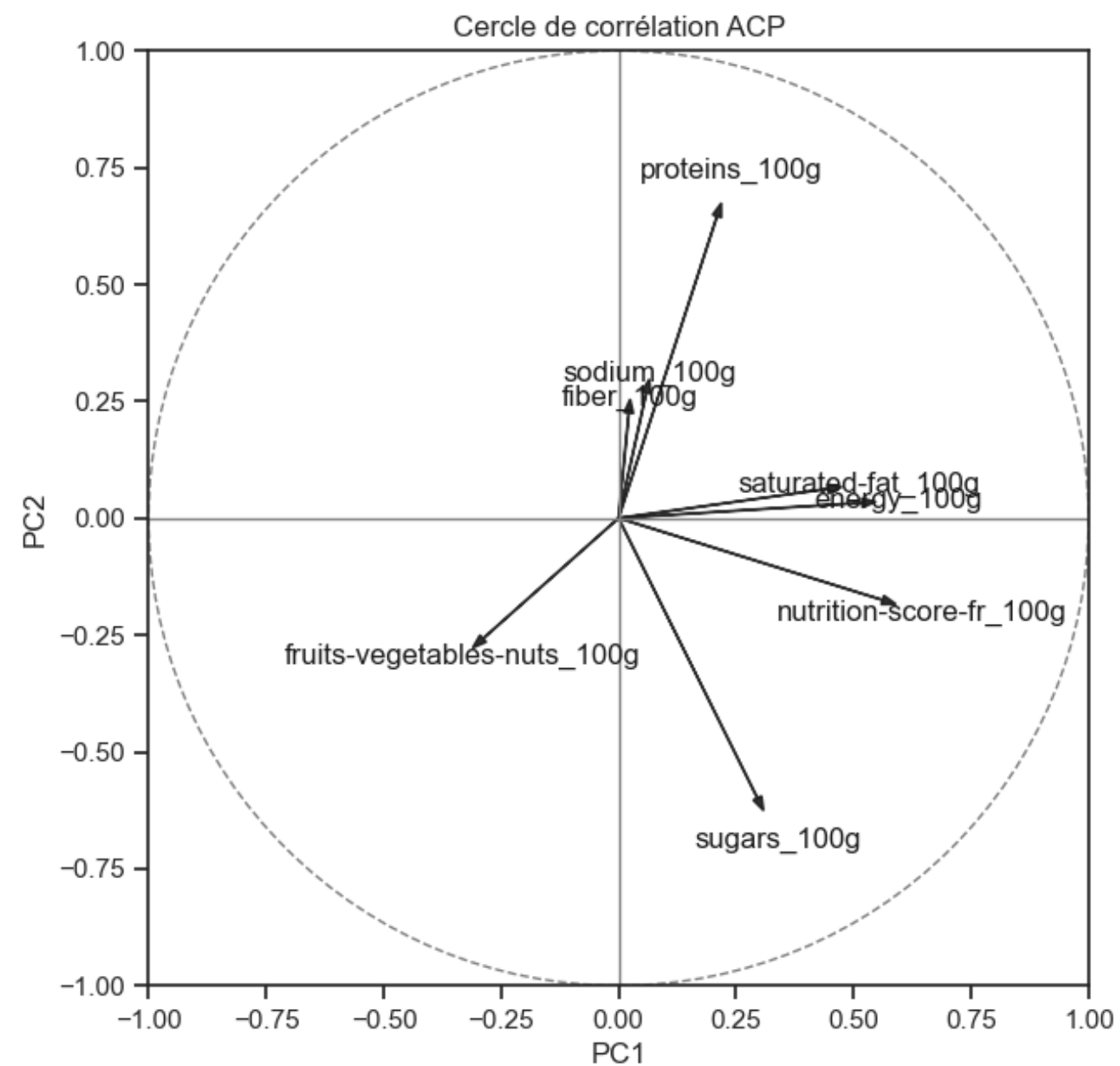
	PC1_contrib	PC2_contrib	PC3_contrib	PC4_contrib
energy_100g	0.643630	0.001518	0.123689	0.000115
saturated-fat_100g	0.475751	0.005521	0.045559	0.022280
sugars_100g	0.209274	0.503616	0.077216	0.039587
fiber_100g	0.001135	0.071066	0.726498	0.040541
proteins_100g	0.105286	0.585917	0.010332	0.018090
sodium_100g	0.008327	0.101001	0.095665	0.787262
fruits-vegetables-nuts_100g	0.199402	0.093237	0.035753	0.048347
nutrition-score-fr_100g	0.756829	0.043082	0.058878	0.004354

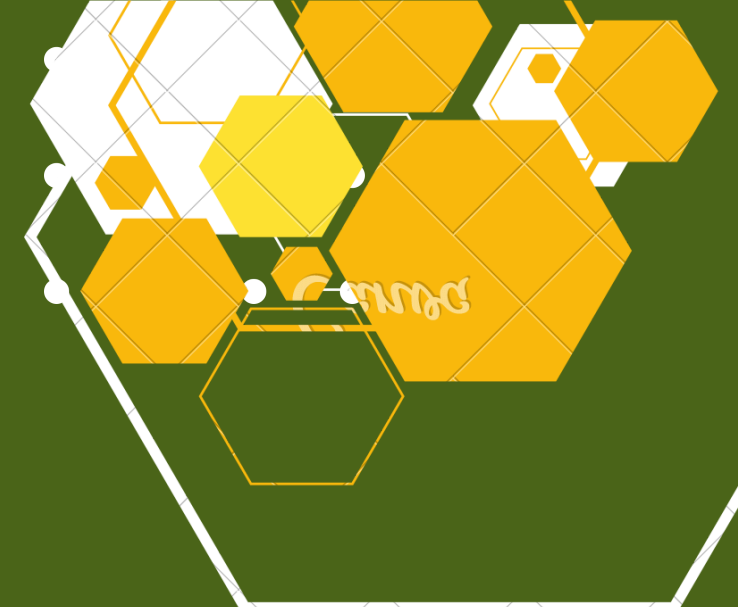
Contribution de chaque variable à la formation des axes

	PC1_contrib	PC2_contrib	PC3_contrib	PC4_contrib
energy_100g	26.822010	0.108028	10.539343	0.011985
saturated-fat_100g	19.825972	0.392982	3.881983	2.319454
sugars_100g	8.721073	35.845621	6.579479	4.121124
fiber_100g	0.047319	5.058234	61.903920	4.220505
proteins_100g	4.387603	41.703545	0.880376	1.883234
sodium_100g	0.346996	7.188878	8.151493	81.957252
fruits-vegetables-nuts_100g	8.309675	6.636304	3.046503	5.033132
nutrition-score-fr_100g	31.539353	3.066407	5.016903	0.453314



Cercles de corrélation sur les principaux plans





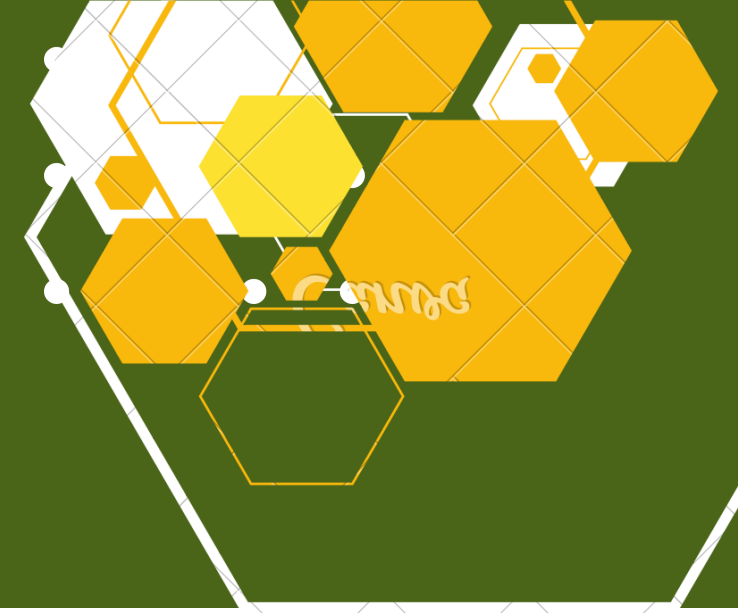
Sur le choix du nombre d'axes

- Nous constatons que pour expliquer la totalité de la variance il faut 8 dimensions
- En analysant les valeurs propres, nous constatons qu'avec 4 dimensions, nous parvenons à expliquer 74% de la variance
- Le scree plot de la variance expliquée par dimension (Eboulis des valeurs propres) montre un coude à partir de la 4ème dimension. Nous retenons donc 4 dimensions pour nos analyses

Sur la contribution des variables

- Energy_100g , nutrition-score-fr_100g contribuent bien à la formation de l'axe PC1 et saturated-fat_100g aussi mais dans une moindre mesure.
- Proteins_100g et sugars_100g contribuent bien à la formation de l'axe PC2
- Fiber_100g contribue fortement à la formation de l'axe PC3
- Sodium_100g contribue très fortement à la formation de l'axe PC4

Analyse multivariée : ACP Interprétations



Sur les axes principaux

- Sur l'axe PC1 : $r(\text{energy_100g}, \text{PC1}) = 0.64$ $r(\text{nutrition-score-fr_100g}, \text{PC1}) = 0.76$

L'axe PC1 pourrait représenter les aliments à forte teneur énergétique, ils ont un nutrition-score élevé ce qui confirme ce que nous avons vu sur leur moins bon classement en grade dans l'analyse bivariée.

- Sur l'axe PC2 : $r(\text{proteins_100g}, \text{PC2}) = 0.6$ $r(\text{sugars_100g}, \text{PC2}) = -0.5$

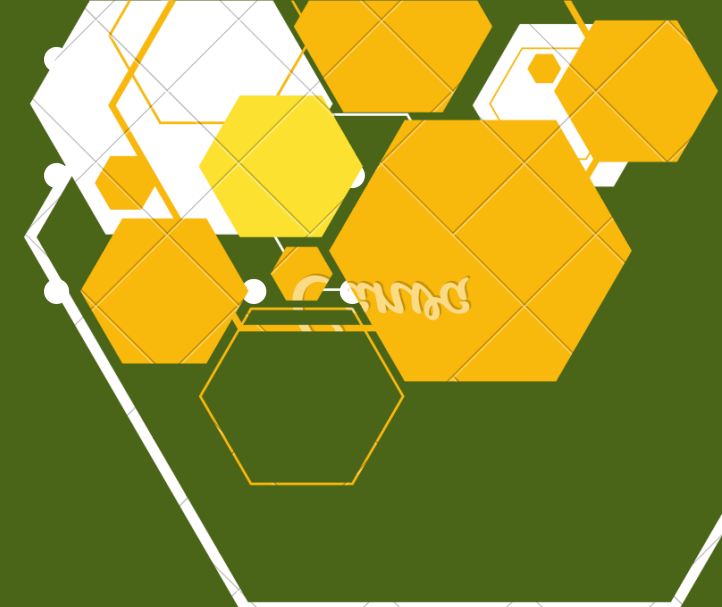
L'axe PC2 pourrait représenter les aliments riches en protéines et faibles en sucre

- Sur l'axe PC3 : $r(\text{fiber_100g}, \text{PC3}) = 0.73$

L'axe PC3 pourrait représenter les aliments riches en fibres

- Sur l'axe PC4 : $r(\text{sodium_100g}, \text{PC4}) = 0.78$

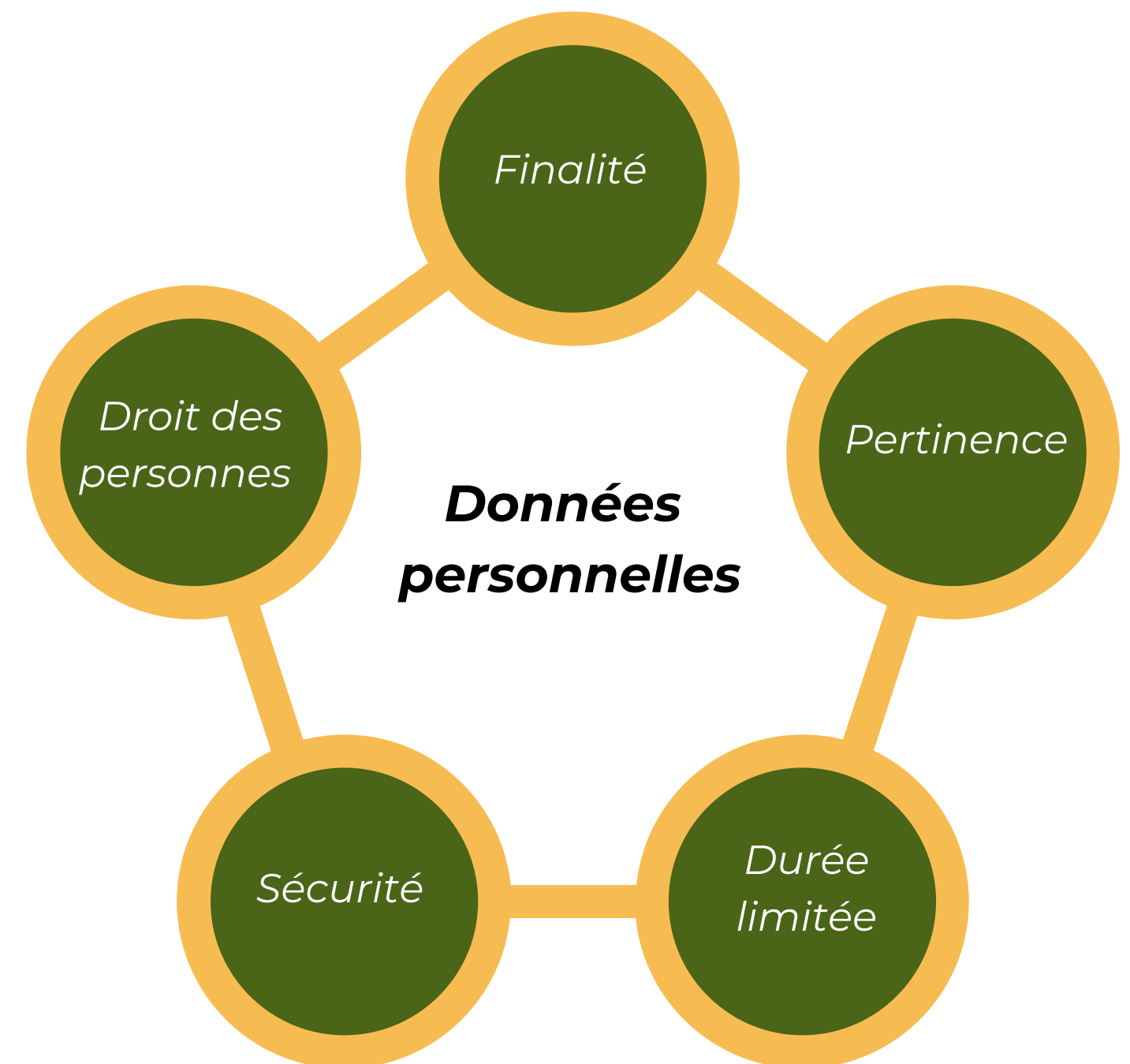
L'axe PC4 pourrait représenter les aliments à forte teneur en sel

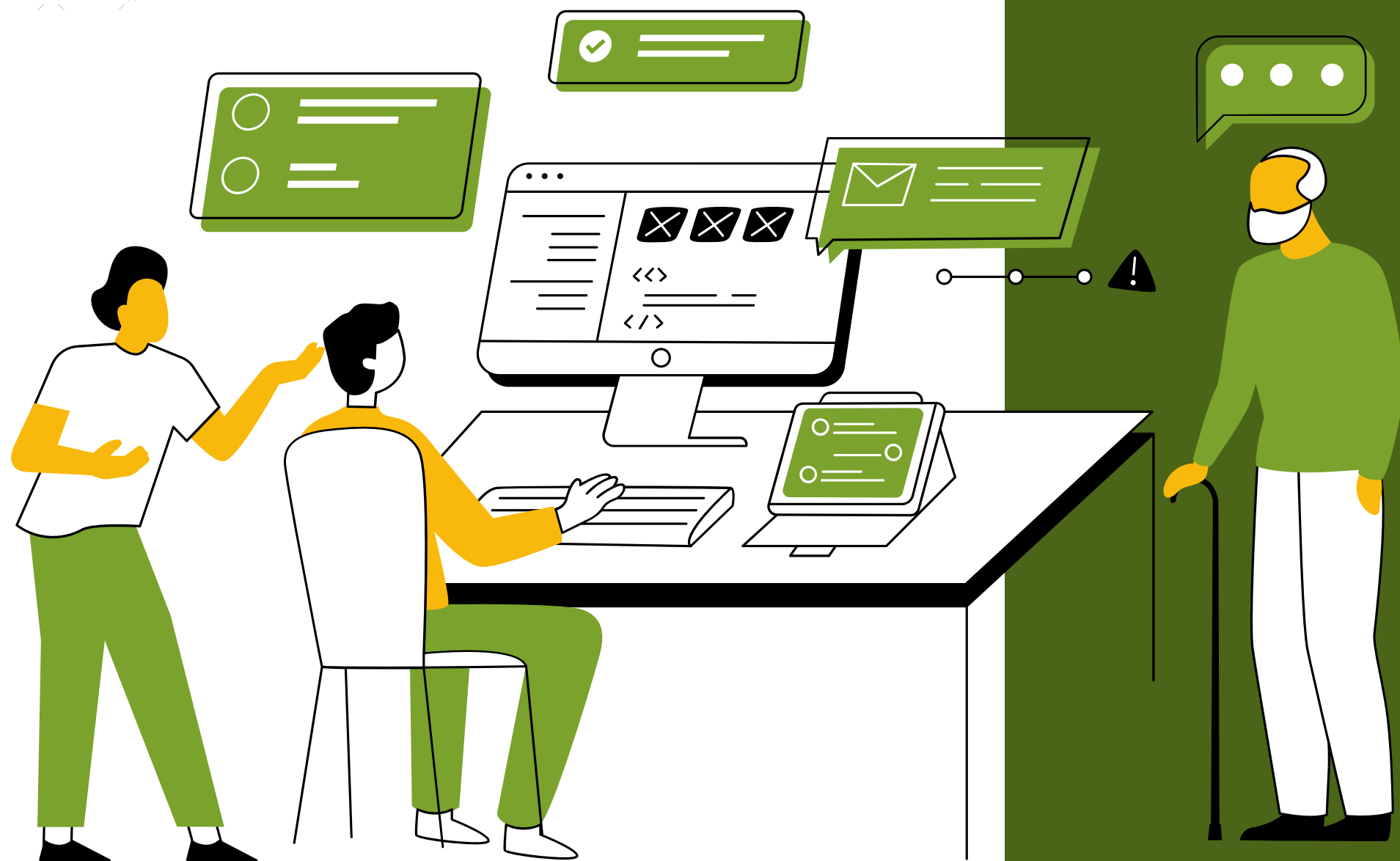
**Le RGPD ou Règlement Général sur la Protection des Données**

- Les données personnelles appartiennent à chaque citoyen. Chaque citoyen doit pouvoir exercer un contrôle.
- Les organisations ne sont que des dépositaires temporaires des données.
- Des responsables doivent être désignés pour l'utilisation des données, leur sécurité et leur confidentialité.

Le RGPD : En quoi le projet de Santé publique France respecte t-il ses principes fondamentaux?

La base de données Open Food Facts ne collecte pas de données personnelles telles que définies dans les principes fondamentaux du RGPD. Son projet de mettre une base de données open source aux usagers pour informer sur la qualité nutritionnelle des produits aliments est hors du champ d'application du RGPD.

Principes fondamentaux du RGPD



Merci

27/27

