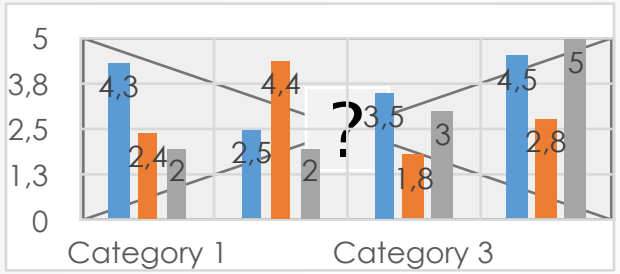
Concevez une Application au sevice de la Santé Publique

Moustapha Abdellahi- OPENCLASSROOMS

Problématique : Concevoir une idée d'application basée sur le traitement des données OpenFoodFacts en termes de nettoyage et d'exploration.

0	A	B	C	D
1				
2				
3				

Analyse du dataset
Forme, valeurs aberrantes, valeurs manquantes, …



Nettoyage des data
Variables pertinentes, outliers, imputation, …

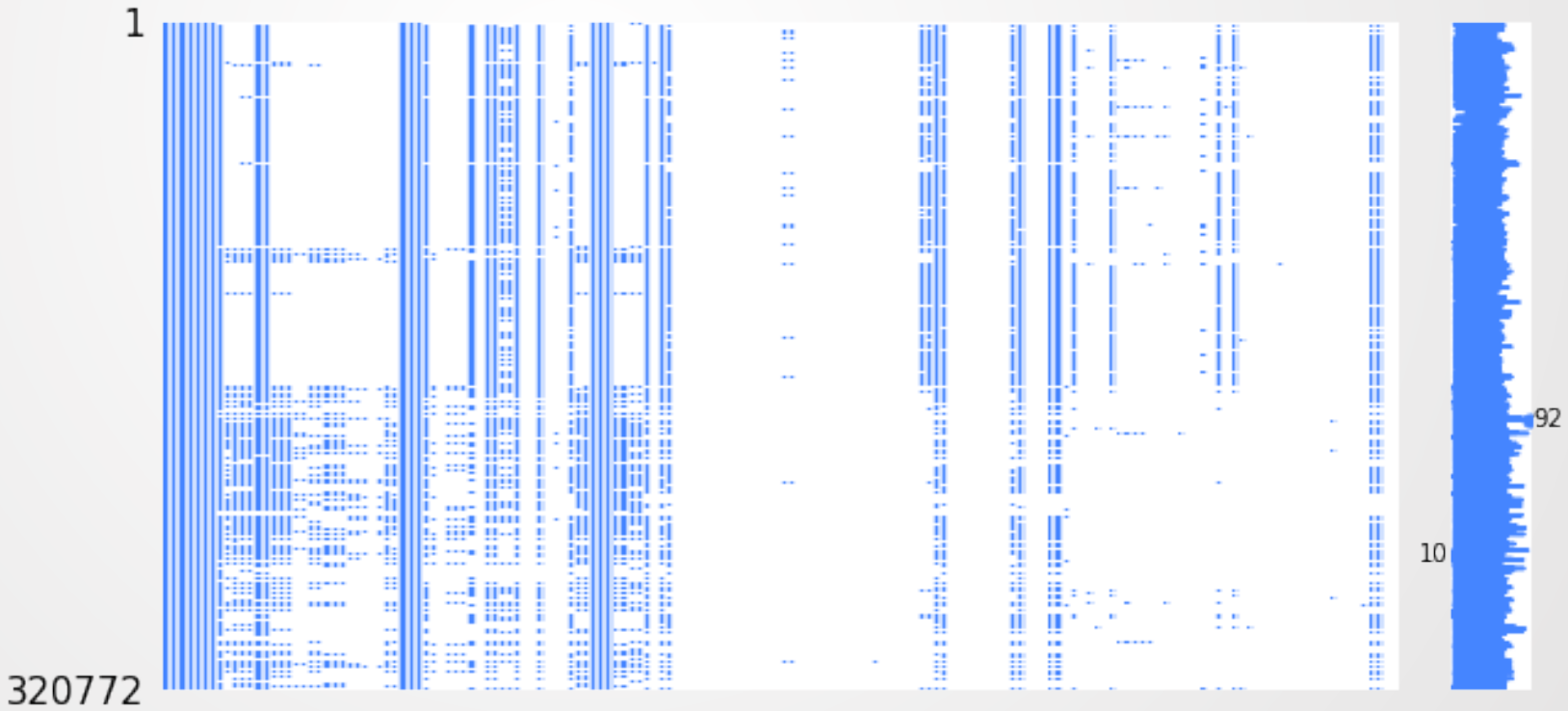
Analyse exploratoire
Analyse univariée, analyse, bivariée, ACP, ANOVA, …

Idée d'application :

- Application qui propose de calculer un nouveau score nommé Nutriclass, qui est basé sur les additifs et la présence d'Oméga 3.
- Prend en compte le Nutri-score (Nutrition-grade) pour l'améliorer
- Trois variables définies par cette étude: **Ingredients_n**, **Nutriratio**, **Nutriclass**

Analyse du dataset :

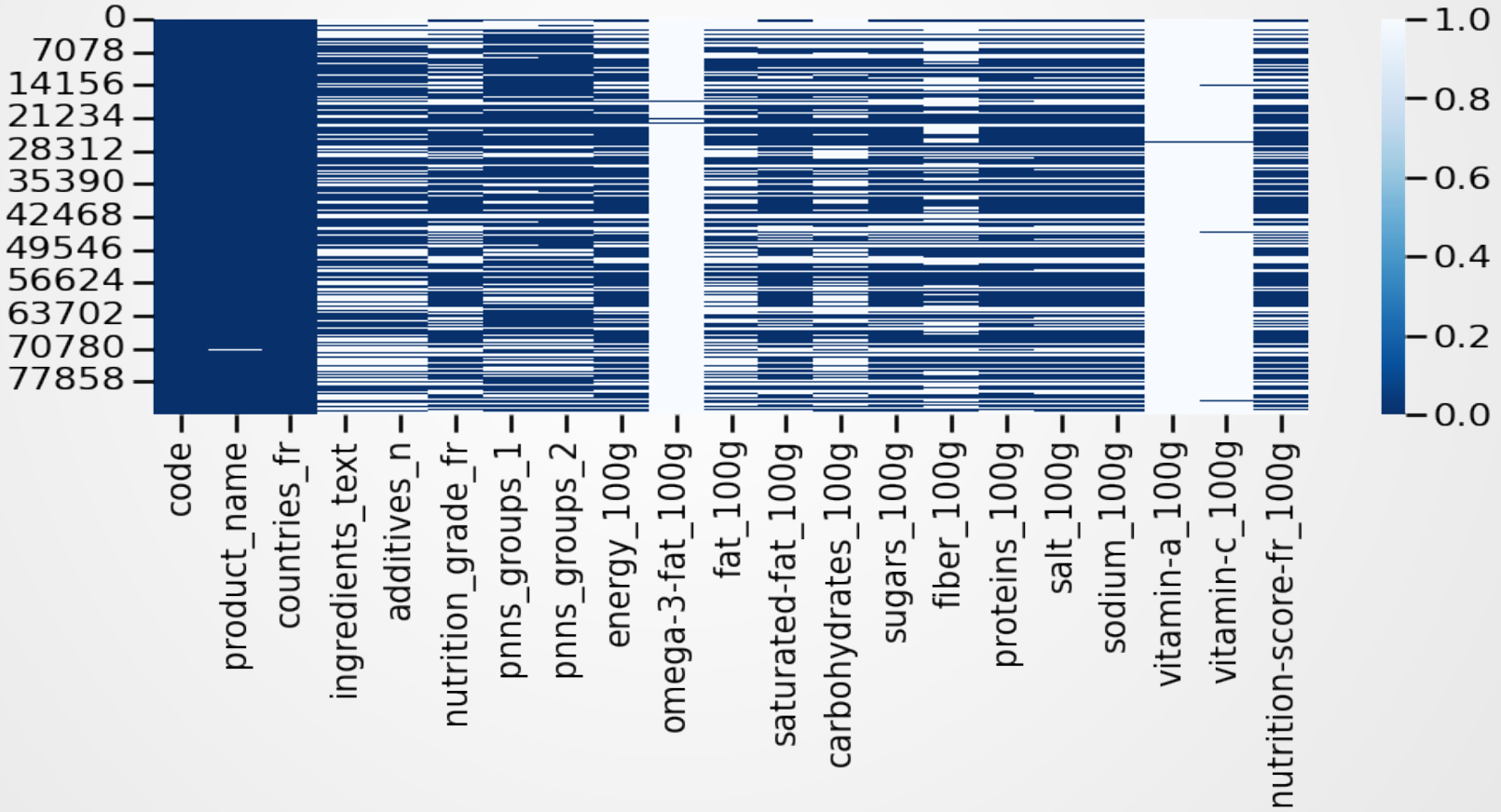
- Source : **OpenFoodFacts** (https://world.openfoodfacts.org/)
- Dimension : **~ 321k lignes × 162 colonnes**
- Lignes : produits
- Colonnes : variables (pays, noms des produits, ingrédients, valeur énergétique,…)
- Données uniquement pour la France**:~ 94k lignes × 20 colonnes**
- Data type



Dataset filtré pour produits français :

Valeurs manquantes:

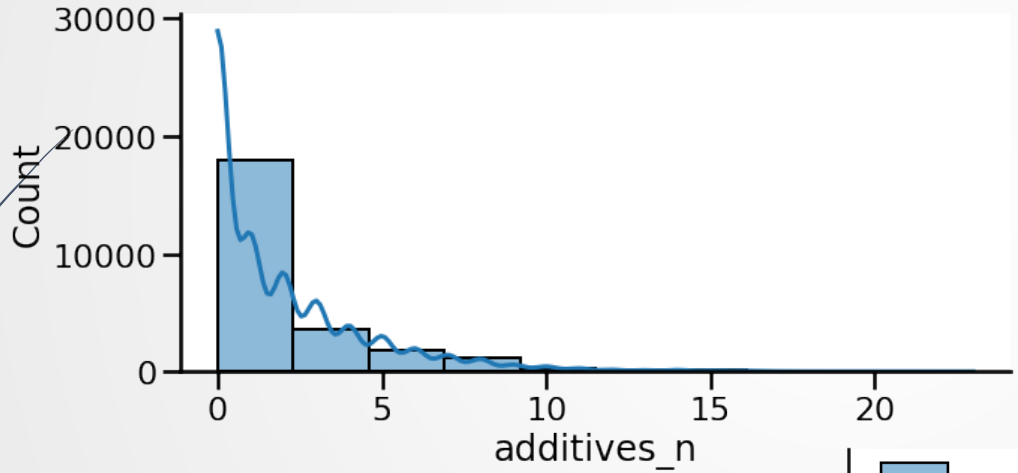
- **Taux de remplissage :~20%** car NaN sur 80% du dataset

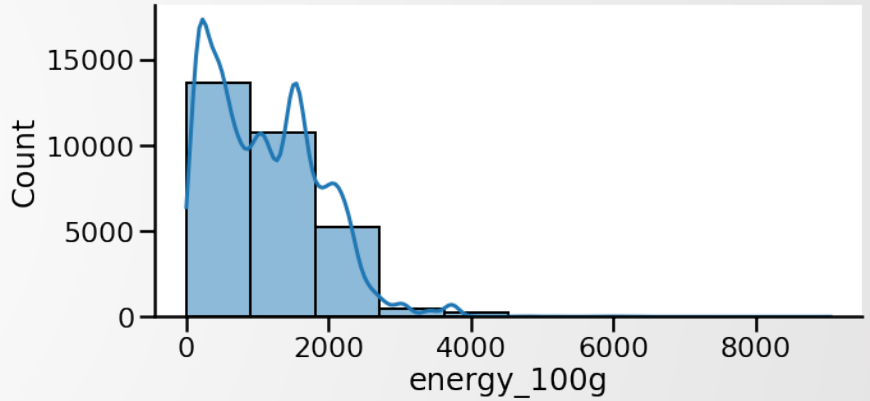


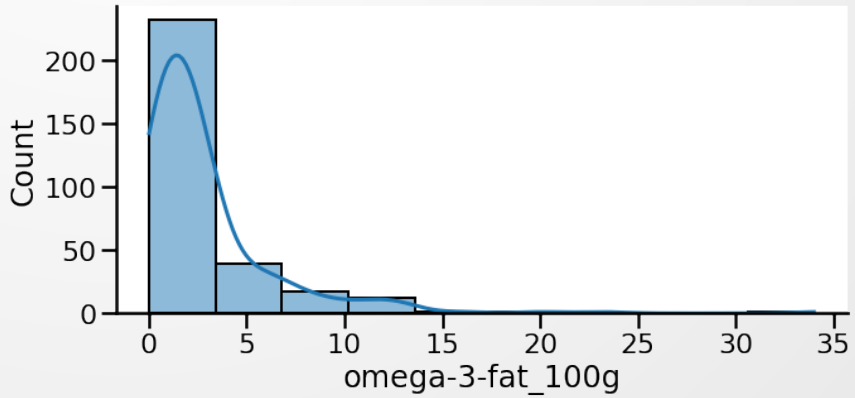
Nettoyage du dataset :

- Elimination des valeurs aberrantes : (si **valeur>100**) dans une **portion de100 g** d'un produit
- Distribution des variables pertinentes
- Remplissage des valeurs manquantes avec trois méthodes différentes

- Distribution des variables et remplissage des NaN par **0**.

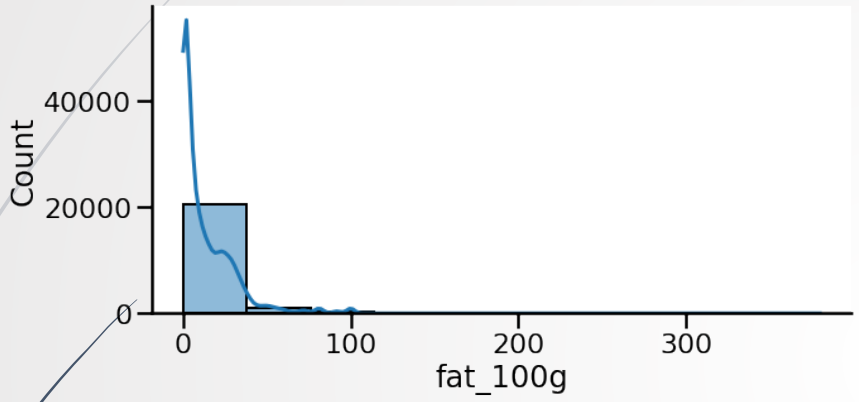


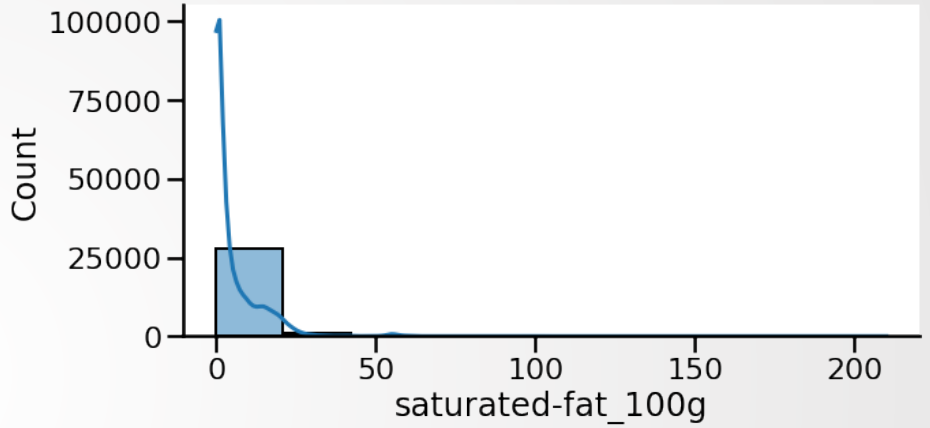


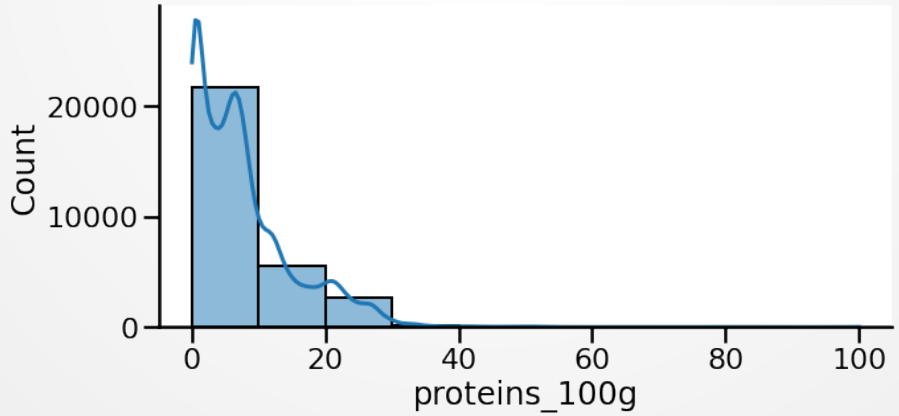


Nettoyage du dataset :

- Distribution des variables et remplissage des NaN par la **médiane** des valeurs.

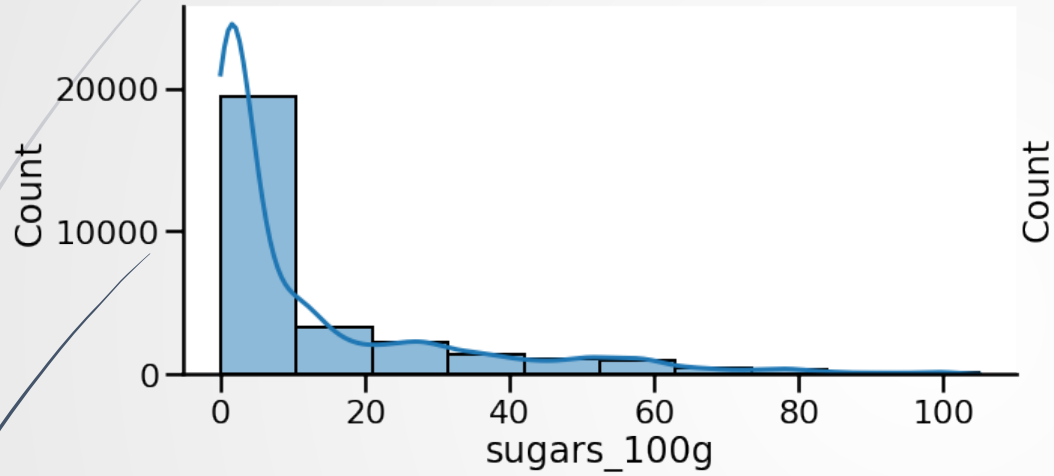


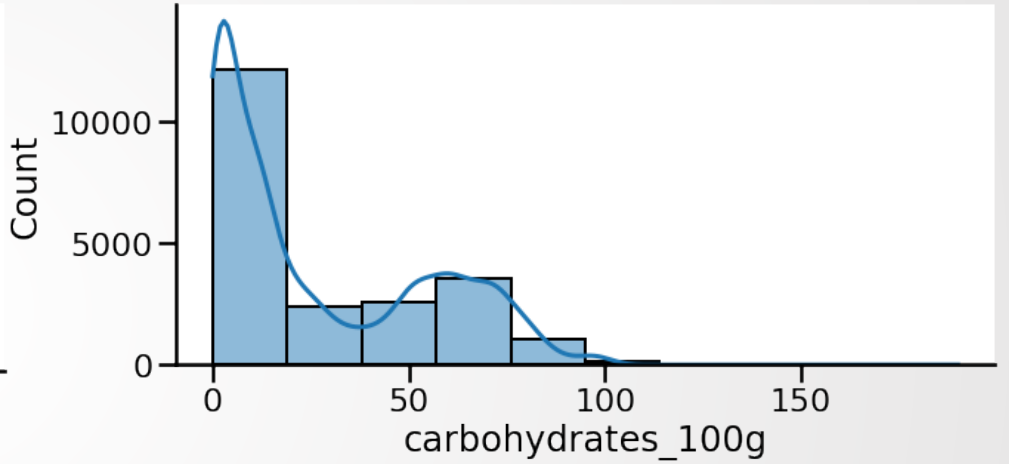


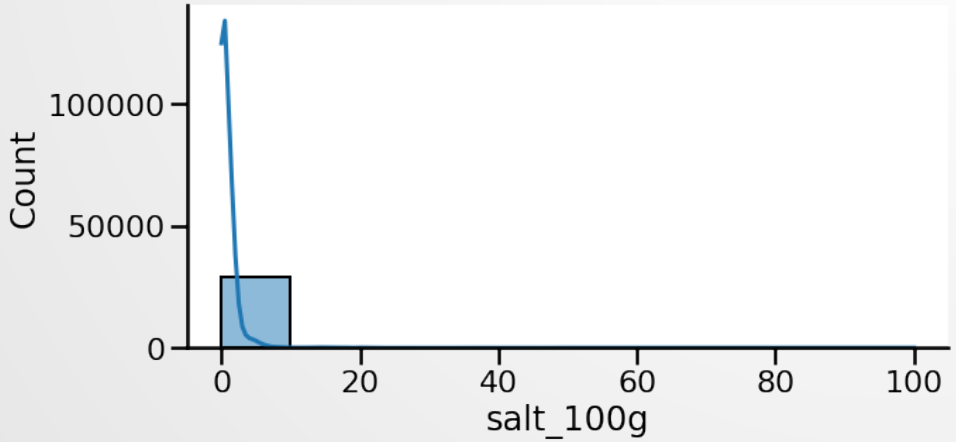


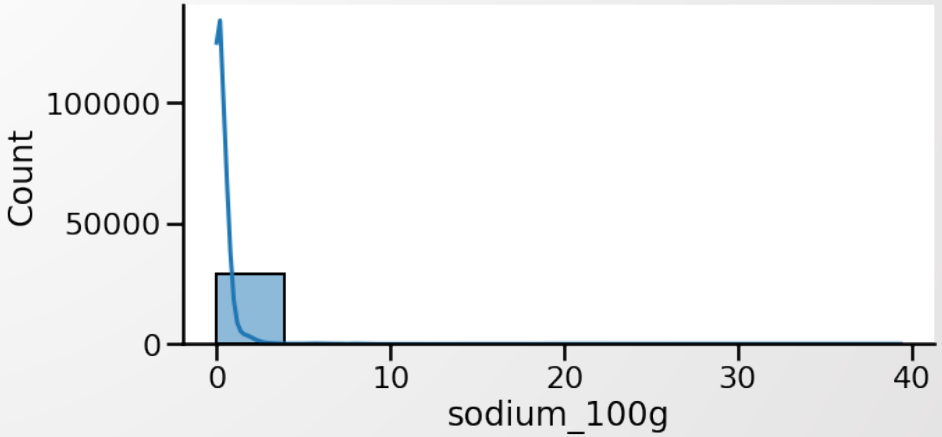
Nettoyage du dataset :

- Distribution des variables et remplissage des NaN par la méthode d'imputation **IterativeImputer** (Machine learning).
- Choix des variables corrélées



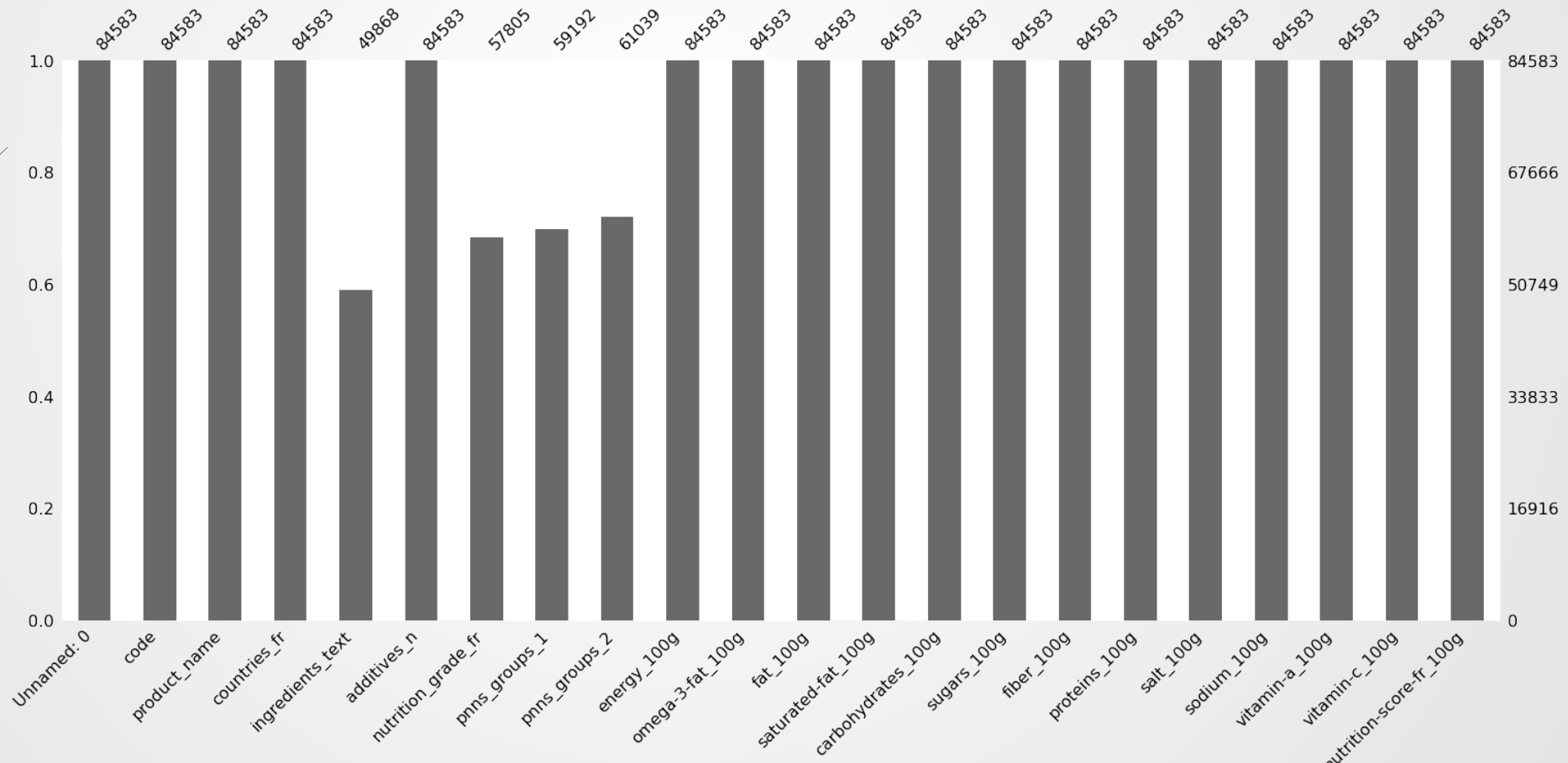






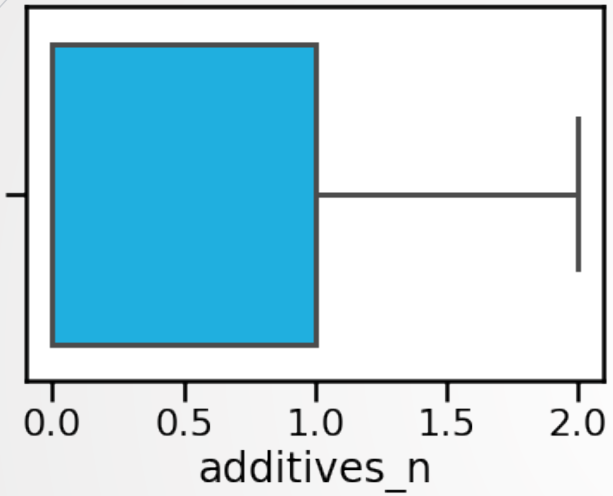
Dataset nettoyé :

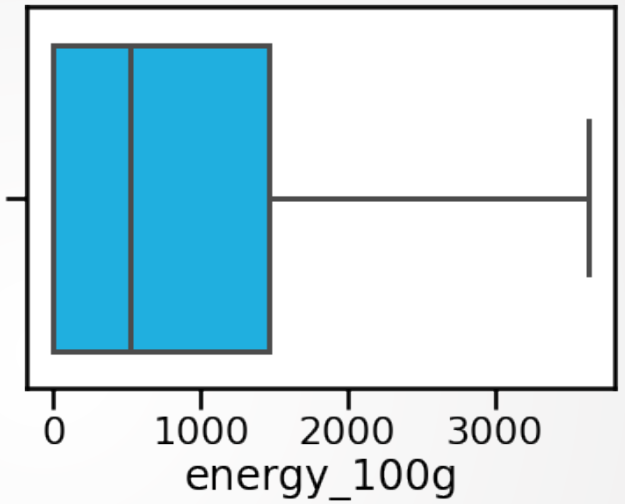
- Valeurs dupliquées supprimées
- Valeurs aberrantes supprimées
- Taux de remplissage des variables numériques = 100%
- **Nombre de variables = 20**

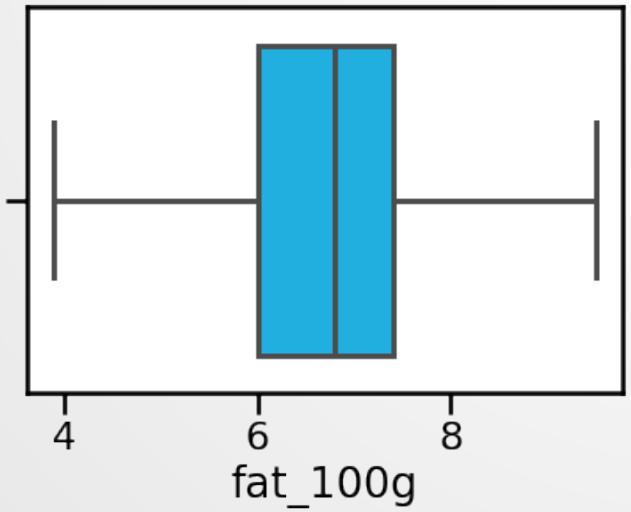


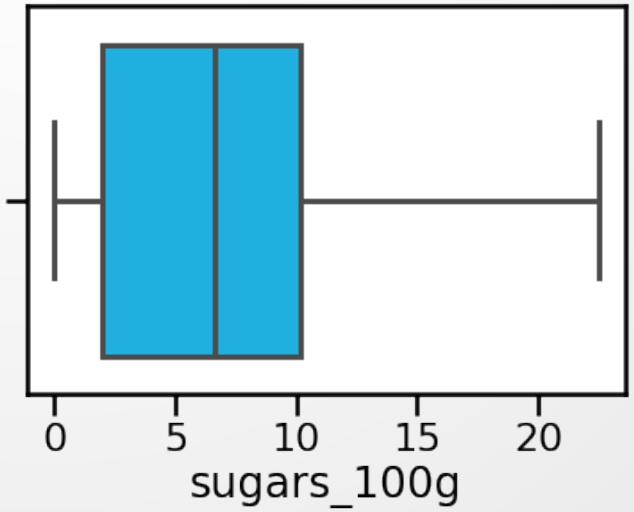
Analyse univariée:

- Features numériques.
- Targets
- Certains produits contiennent plusieurs additifs



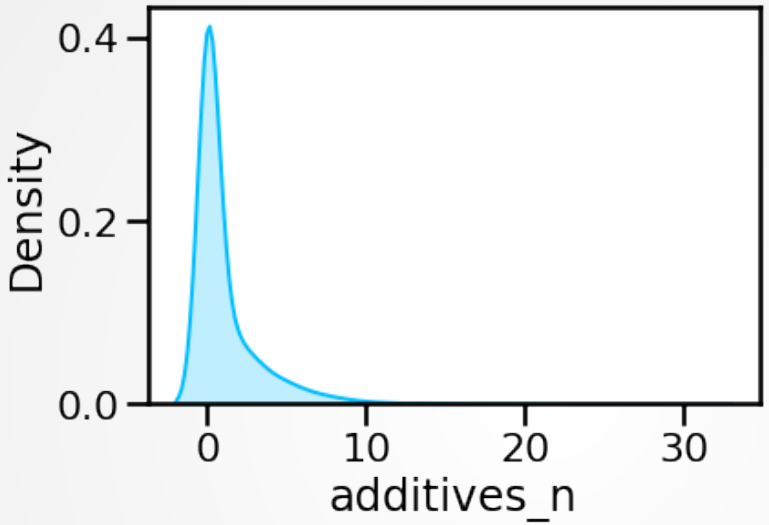


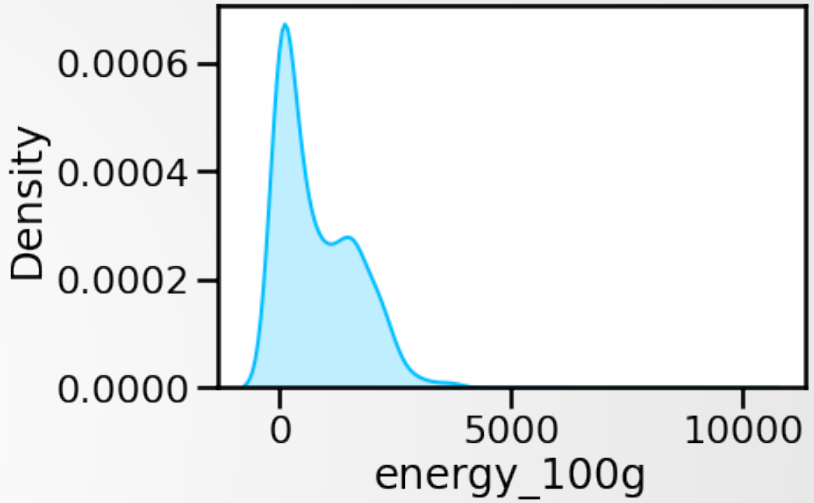


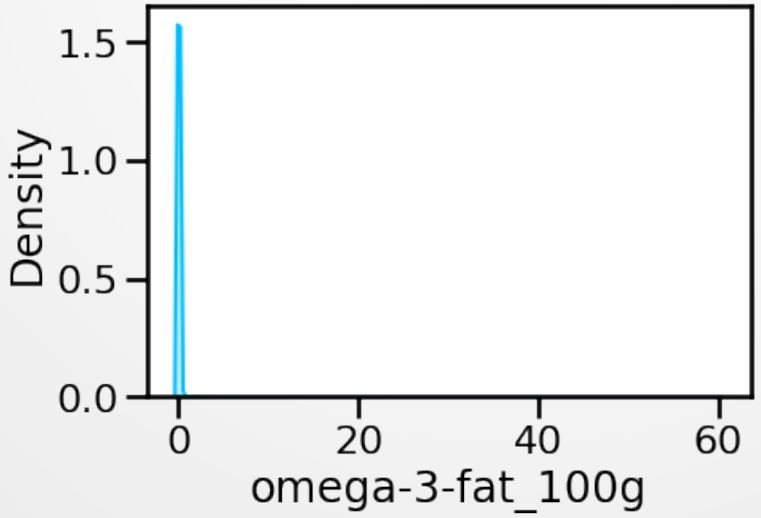


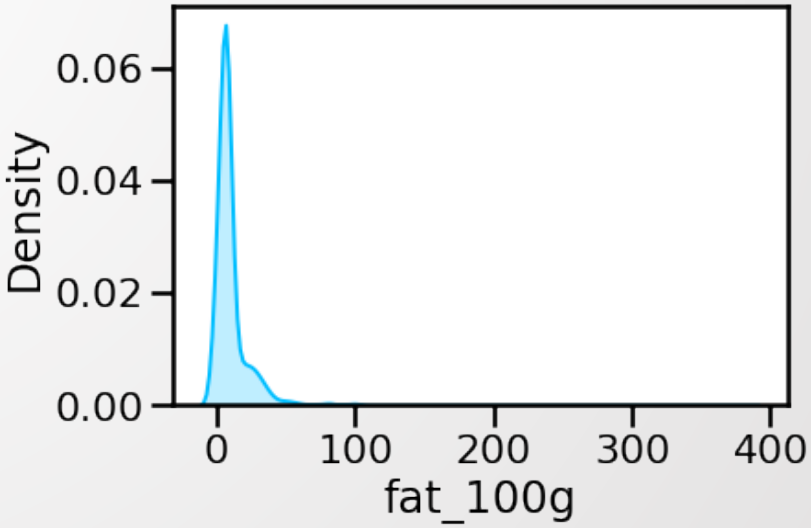
Analyse univariée (suite):

- Densités **non normale** (gaussienne)
- Features numériques comportant des outliers



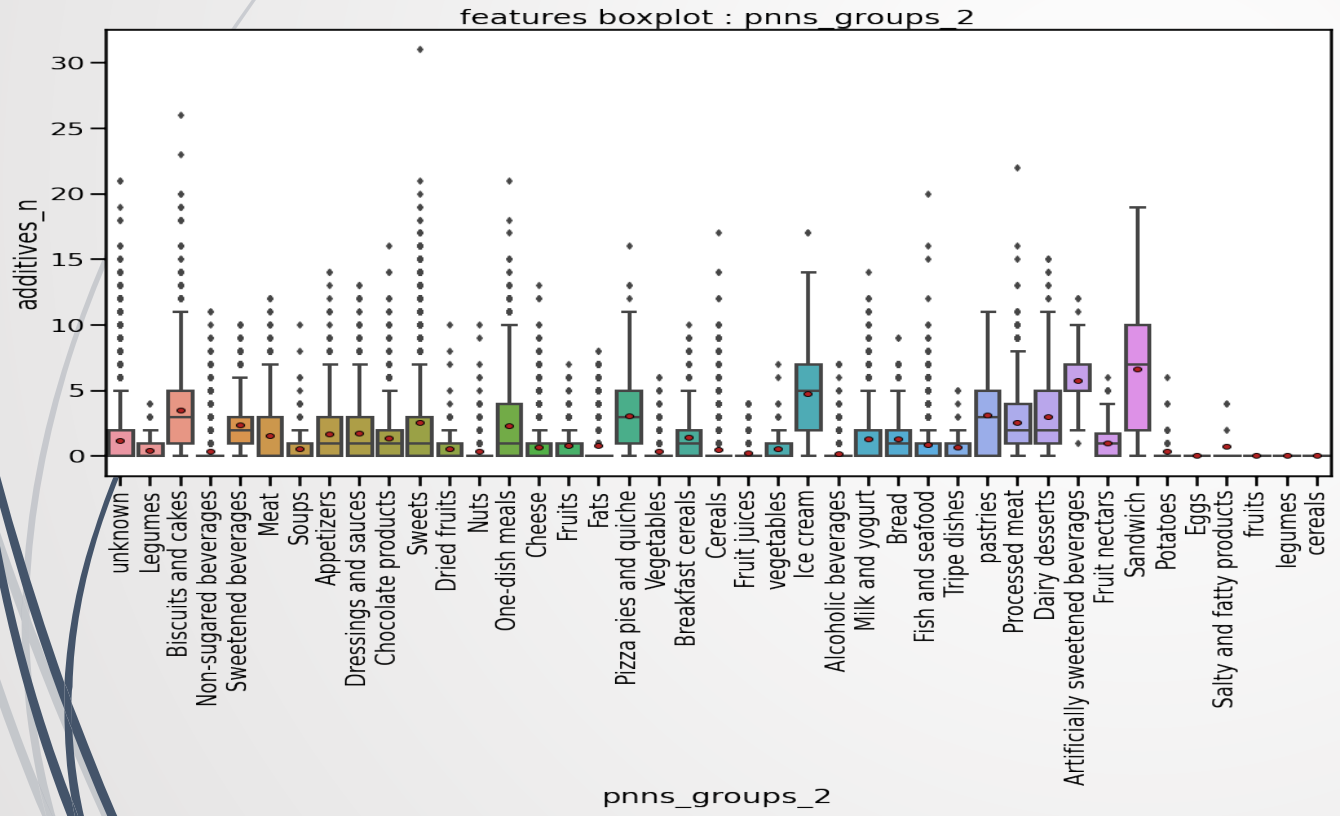


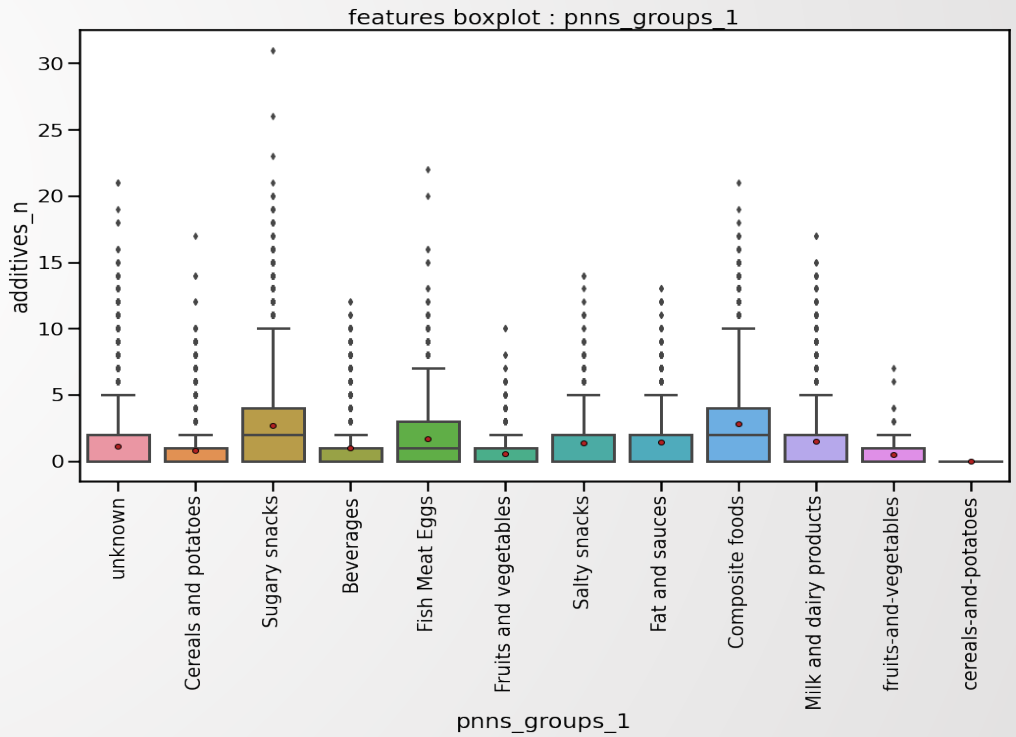




Analyse univariée:

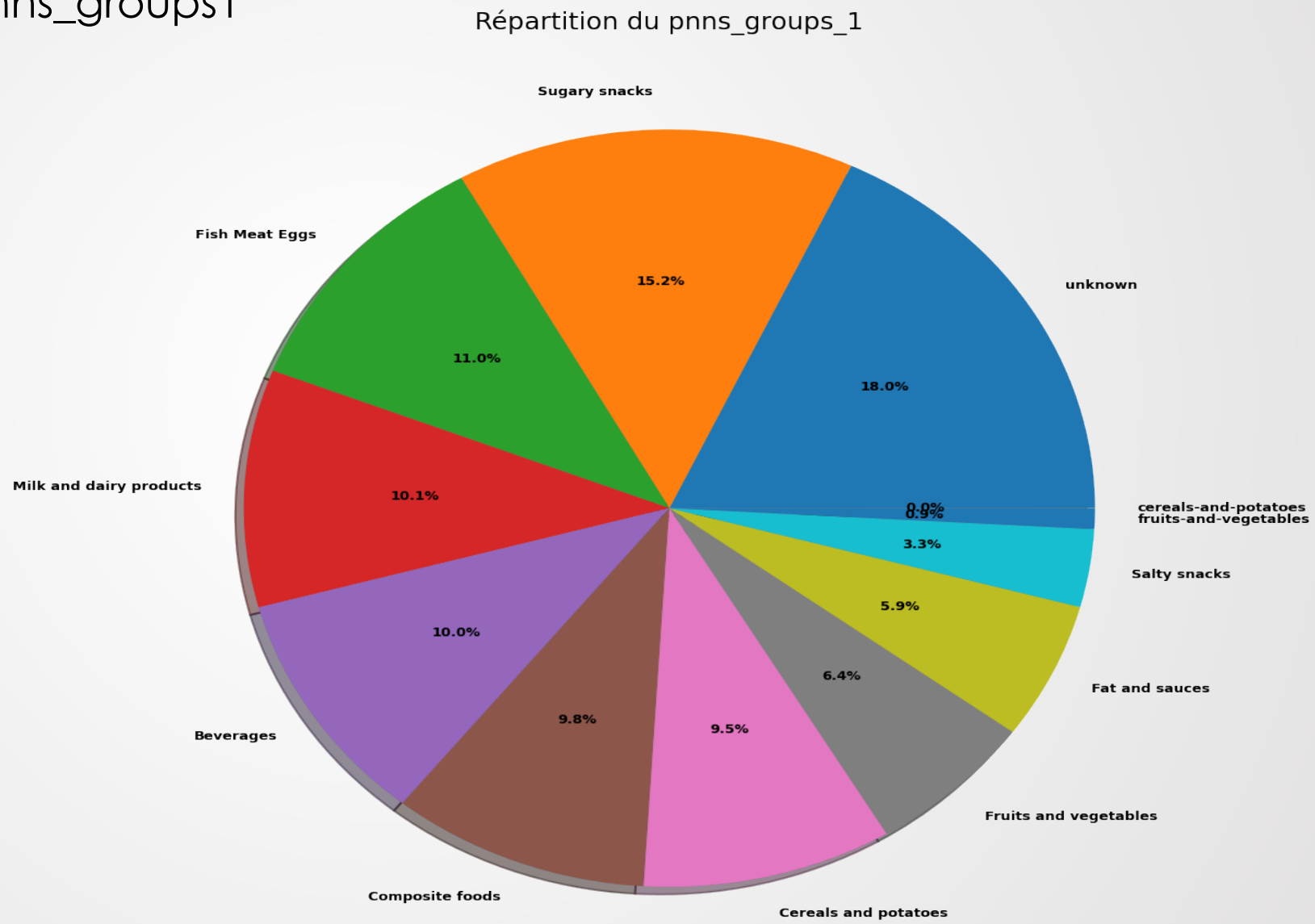
- Distributions des catégories non similaires
- Moyenne des additifs plus élevée pour les catégories **composite foods** (pnns_group_1) et **les sandwichs** (pnns_group_2)





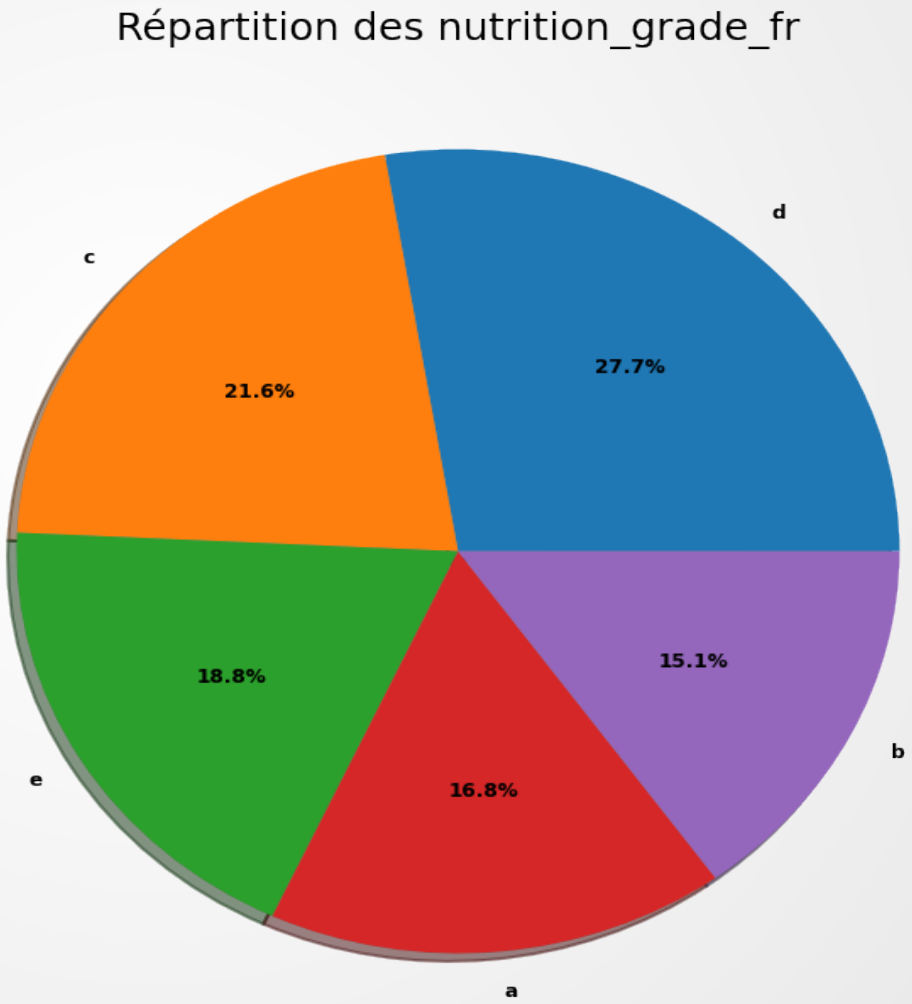
Analyse univariée (suite):

- Le groupe de test est **pnns_groups1**, car il contient les principales catégories
- Répartition de pnns_groups1



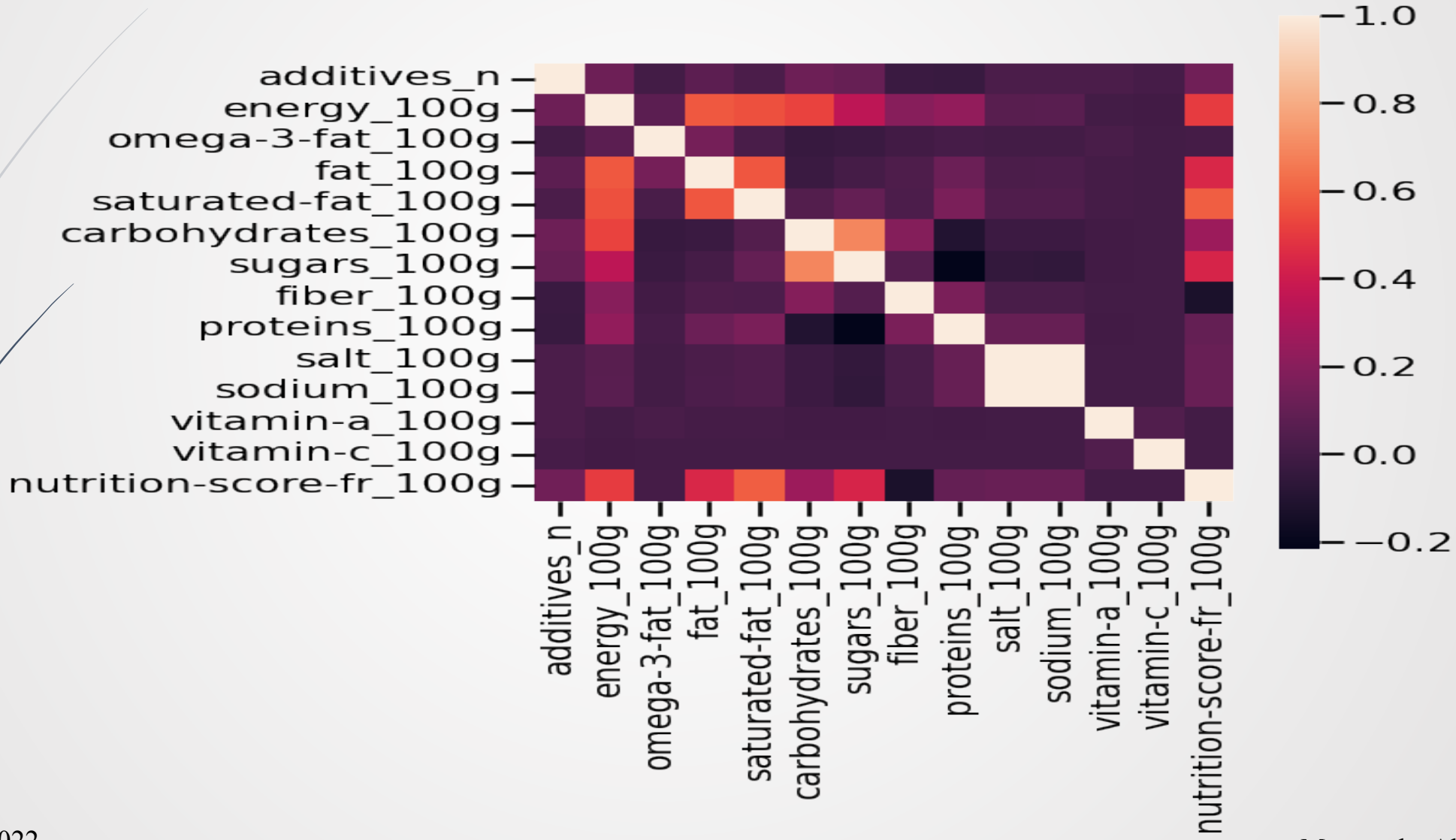
Analyse univariée:

- Les pourcentages de nutritions_gardes (**a**, **b**, **e**) sont proches
- Les pourcentages de nutritions_gardes (**c**, **d**) représentent près de la moitié des produits



Analyse bivariée:

- Forte corrélation entre les variables **salt et sodium**, ($r \simeq 1$)
- Corrélation entre les variables **fat, saturated fat, carbohydates, sugar et** la variable **energy** $(r < 0.9)$

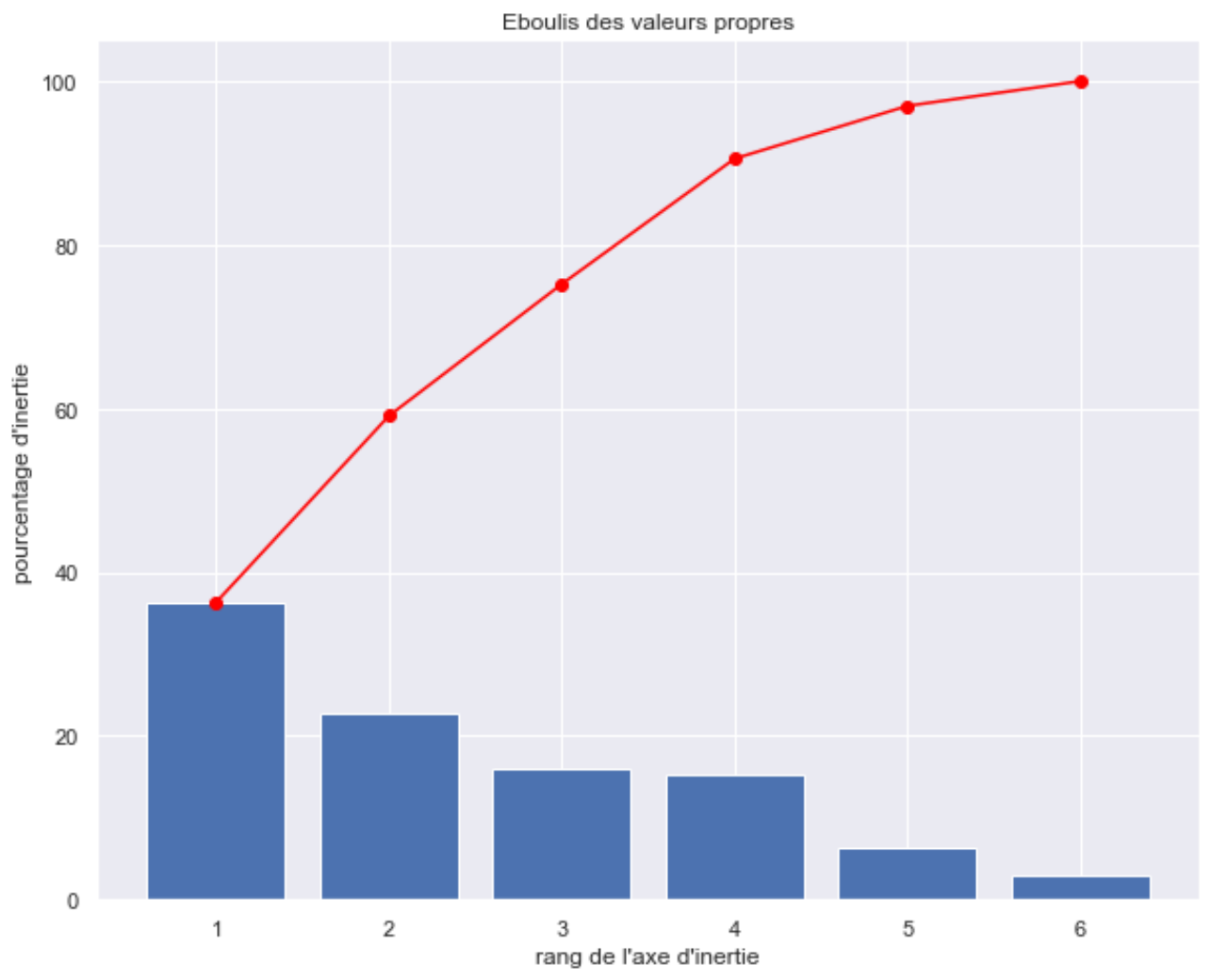


Analyse bivariée (suite):

- Faible corrélation entre les variables : **oméga 3 et energy**
- Pas de corrélation évidente entre : **additives et energy**

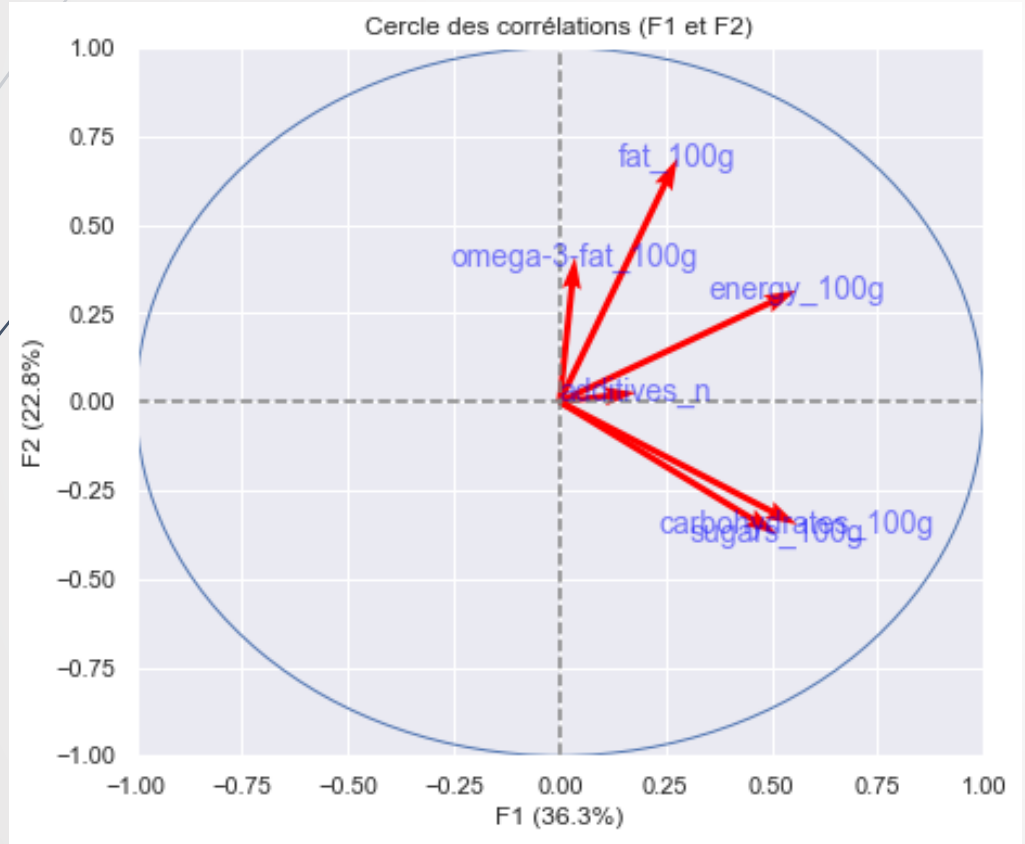
Réduction dimensionnelle: Analyse en Composantes Principales (ACP)

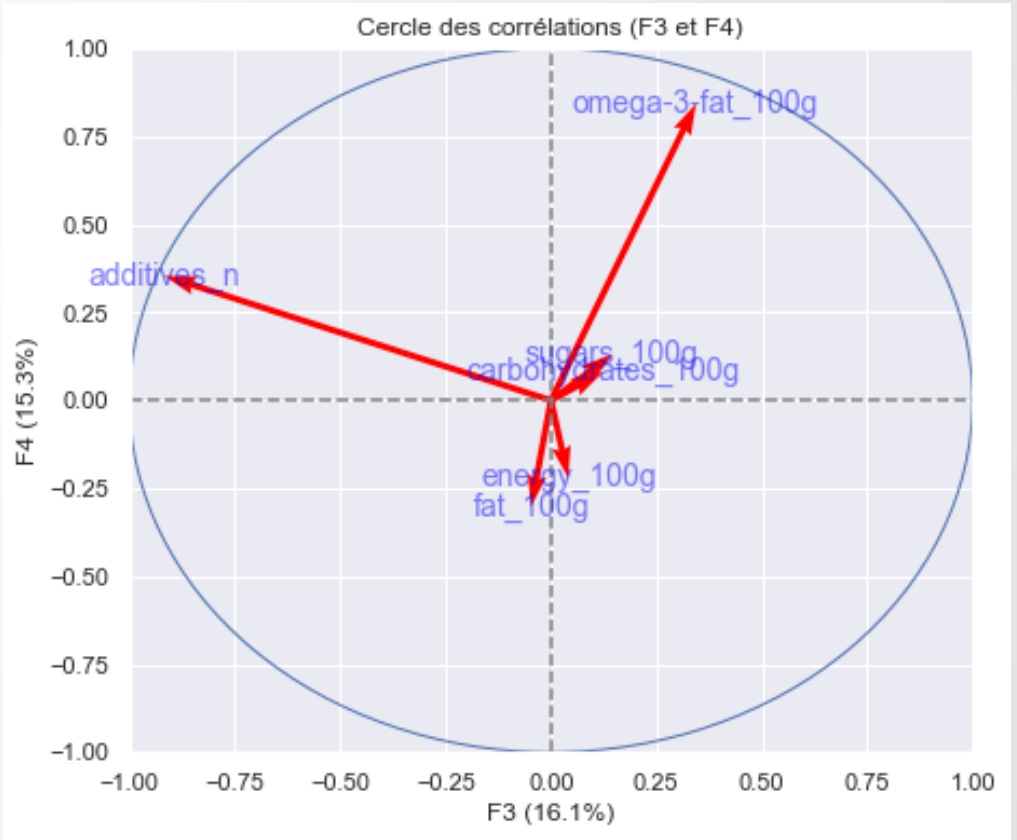
- Méthode pour explorer des jeux de données multidimensionnels composés de variables quantitatives.
- Les valeurs propres de la matrice de covariance



Réduction dimensionnelle (suite): Analyse en Composantes Principales (ACP)

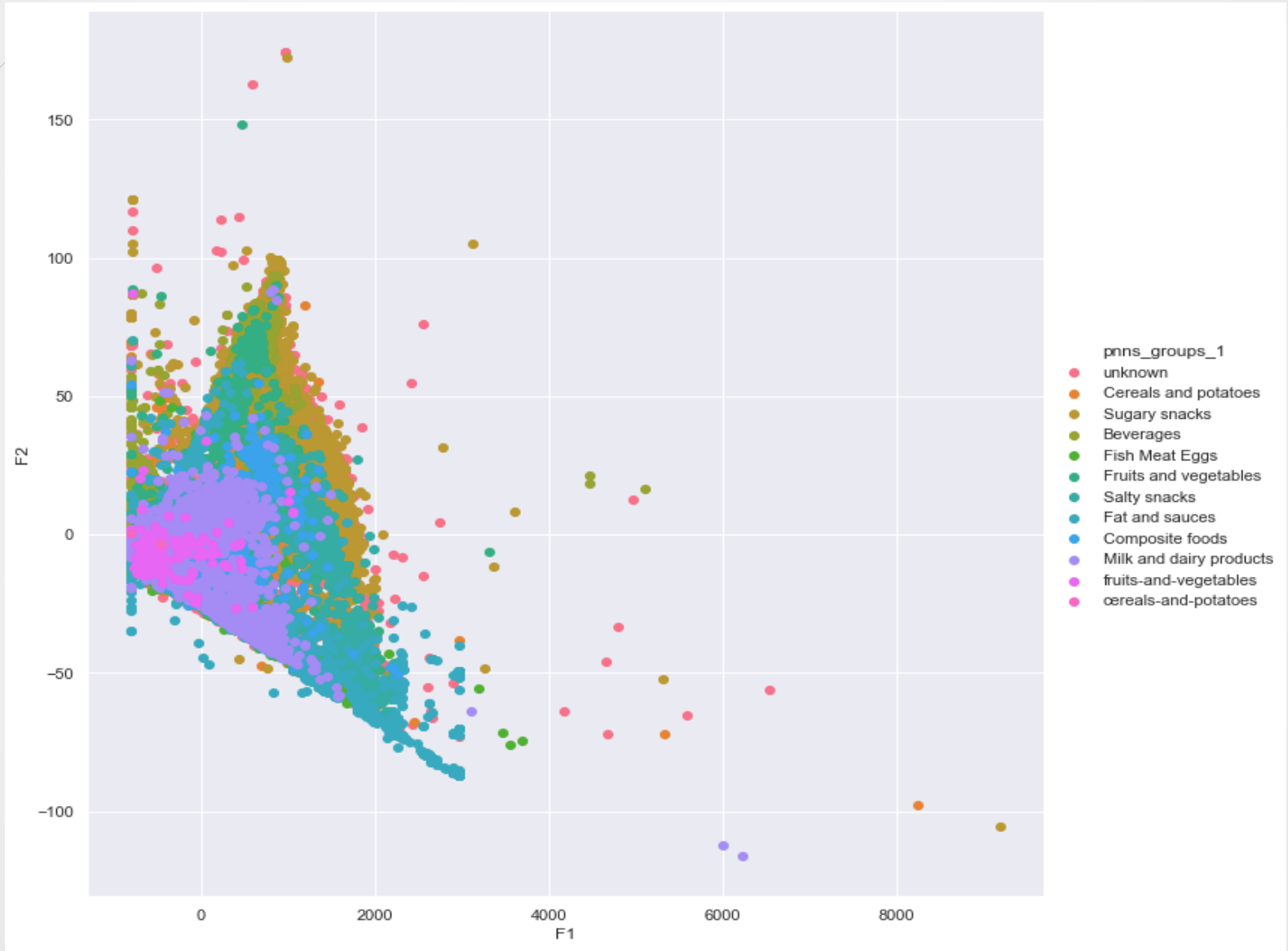
- Cercles de corrélation : $F1 \ et \ F2$ représentent environ **60%** des informations
- Cercles de corrélation : $F1, F2, F3 \ et \ F4$ représentent environ **90%** des informations





Réduction dimensionnelle: ACP

- Projection des individus de pnns_group_1 sur F1 et F2
- La catégorie **fruit-and-vegetables** concentrée près de l'origine.

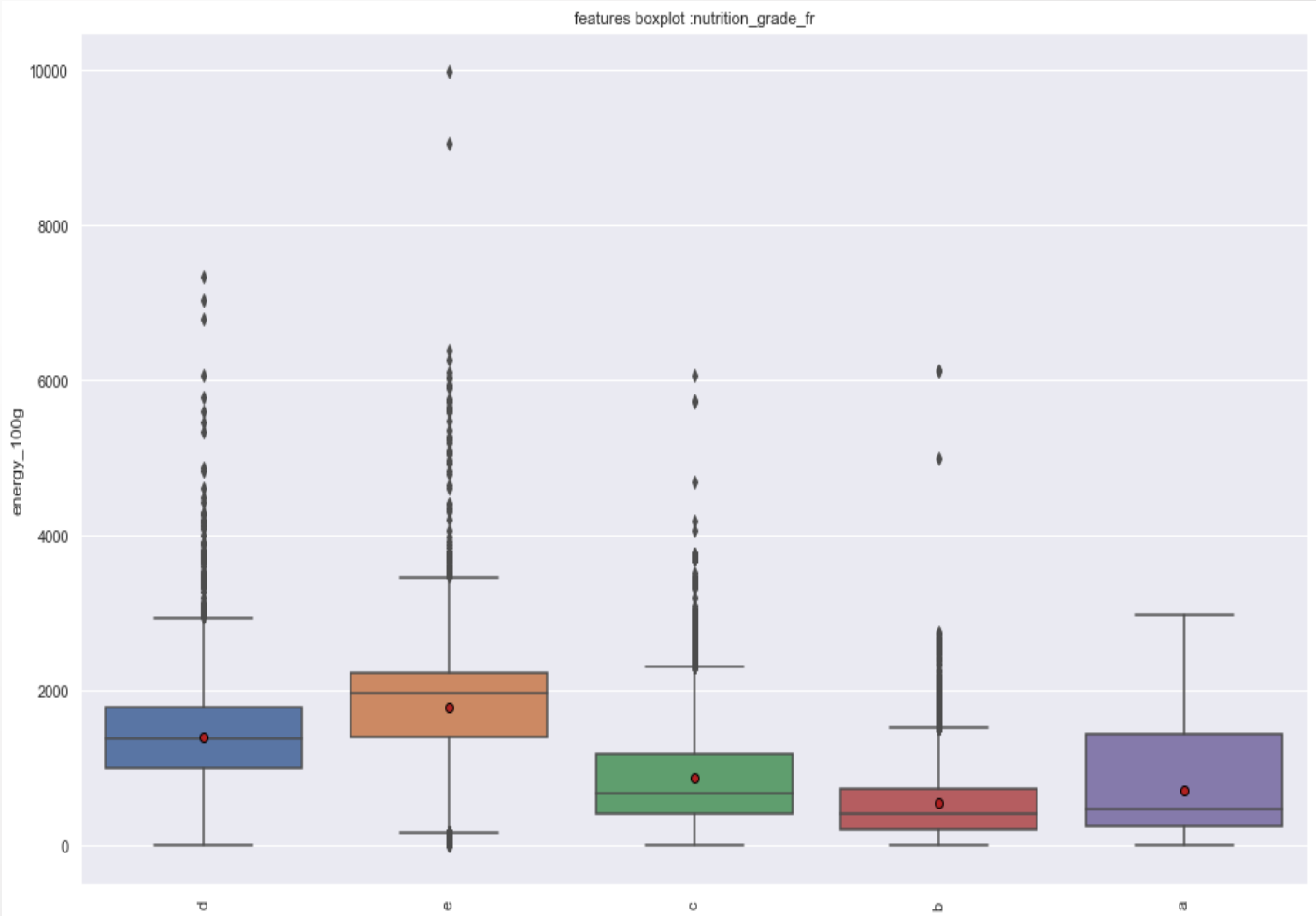


Analyse of Variance : ANOVA

- Influence d'une variable qualitative **nutrition_grade_fr** sur une variable quantitative **energy_100**
- Les hypothèses sont :

  H0 : La distribution des échantillons est identique (nutrition_grade_fr n'a aucune influence sur l'énergie).

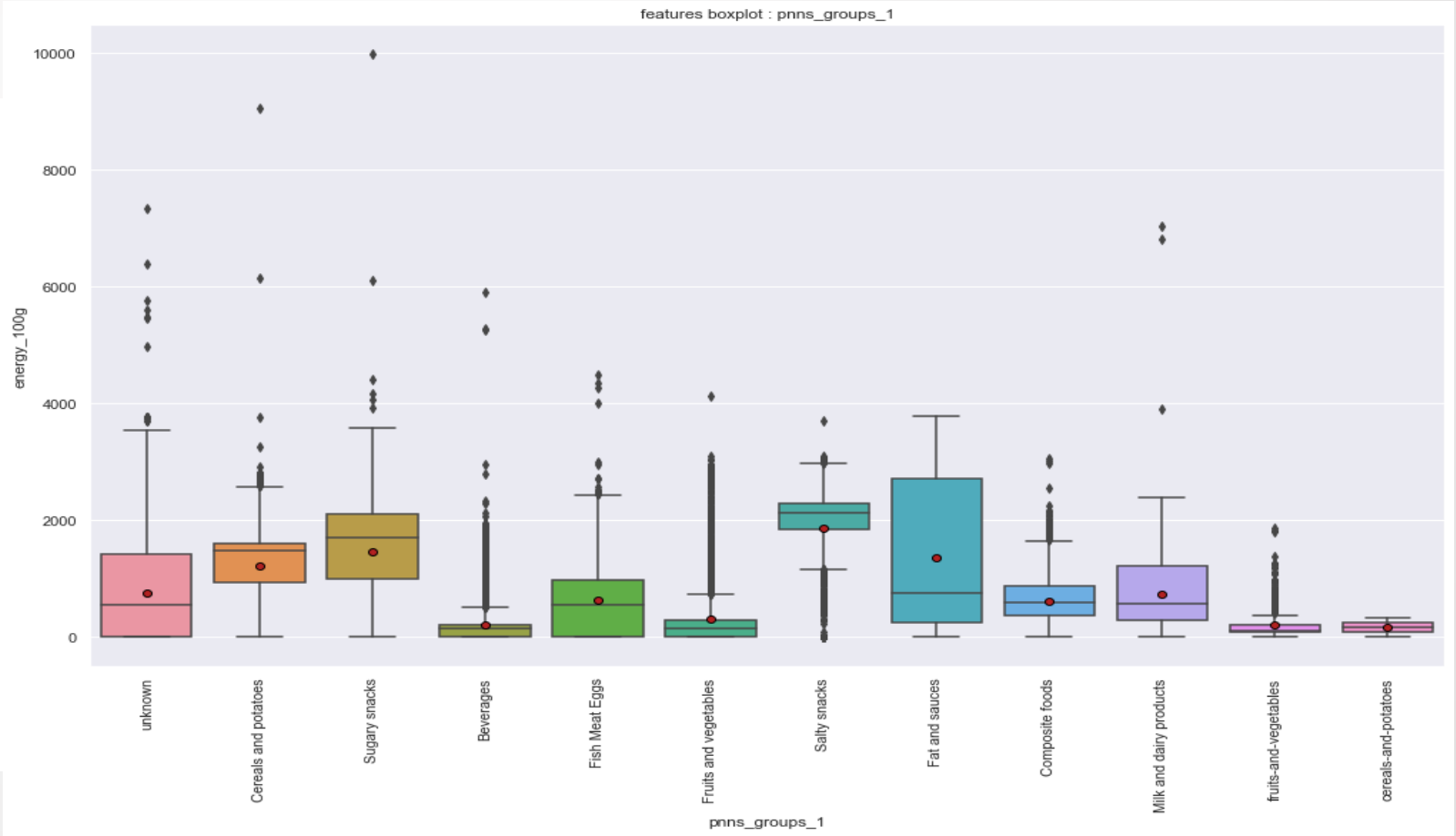
	df	sum_sq	mean_sq	F	PR(>F)
nutrition_grade_fr	4.0	1.140056e+10	2.850140e+09	6785.177746	0.0
Residual	57800.0	2.427911e+10	4.200539e+05	NaN	NaN



Analyse of Variance (suite) : ANOVA

- Influence d'une variable qualitative **pnns_group_1** sur une variable quantitative **energy_100**
- Les hypothèses sont :
  H0 : La distribution des échantillons est similaire (pnns_group_1 n'influence pas l'énergie).
  H1 : Une ou plusieurs distributions sont inégales
  Rejet de **H0** car **p-value =0**

	df	sum_sq	mean_sq	F	PR(>F)
pnns_groups_1	11.0	1.157223e+10	1.052021e+09	2204.363348	0.0
Residual	59180.0	2.824336e+10	4.772450e+05	NaN	NaN



Application et scoring :

- Donner le code d'un produit
- Calculer son nutriratio
- Vérifier les conditions de classement

Quelques exemples :

- Produit avec nutriclass =5

	product_name	nutriratio	nutriclass
2230	Pavé de Saumon	0.000000	5.0
4039	Pavè de Saumon	0.000000	5.0
11595	Farine de blé au lin	0.000000	5.0
19850	Filets de maquereaux grillés aux herbes	0.000000	5.0
22535	Les Croustillants Cabillaud Caviar d'Aubergines	0.029412	5.0
22585	Colin d'Alaska Bisque de crevettes, Surgelé	0.000000	5.0
25169	Risotto Champignons & Saumon	0.000000	5.0
26639	Oeufs moyens, frais, de poules élévées en plein air	0.000000	5.0
26842	Oeufs frais Bio (x 10) calibre Gros	0.000000	5.0
28296	Oeufs frais	0.000000	5.0
28297	6 Œufs frais de Poules Élevées en Plein Air (Calibre Moyen)	0.000000	5.0

Produit avec nutriclass =4

	product_name	nutriratio	nutriclass
174	Purée de marrons	0.0	4.0
183	Tranches d'Ananas au jus d'Ananas	0.0	4.0
265	Tropical Gold Premium Pineapple Chunks in Juice	0.0	4.0
325	Soupe rustique aux légumes à l'italienne	0.0	4.0
331	Spaghettis Italiano avec boulettes végétales	0.0	4.0
...	...	...	...
84003	Mangues en tranches au sirop léger	0.0	4.0
84207	Fusillini N° 37	0.0	4.0
84382	fromage blanc bio	0.0	4.0
84491	Ananas en morceaux	0.0	4.0
84493	Cœurs de palmier	0.0	4.0

Quelques exemples :

- Produit avec nutriclass =1

	product_name	nutriratio	nutriclass
557	Dragées Surprises de Bertie Crochue	0.538462	1.0
635	Wonka Nerds Starwberry And Grape Theatre	0.555556	1.0
696	Sriracha Stærk Chili Sauce	0.571429	1.0
1612	Carambar Mystery	0.642857	1.0
2142	Lipton Ice Tea Pêche	0.571429	1.0
2744	Confiture Abricot	0.600000	1.0
2832	Mini éclairs, chocolat/café - Coup de Coeur	0.625000	1.0
2851	Confiture de fraises	0.666667	1.0
2886	Gouda jeune	0.750000	1.0
2893	Levure chimique	0.666667	1.0
2966	Saucisse de Lyon en tranches	0.666667	1.0

Produit avec nutriclass =0

	product_name	nutriratio	nutriclass
236	Reese's Peanut Butter Cups Miniatures, 40 Oz	1.000000	0.0
3743	Allumettes de lardons nature	2.000000	0.0
4677	Freeway lemon	0.800000	0.0
6226	Sirop de Citron	2.000000	0.0
6855	6 Géants (Chocolat lait, Vanille & Amandes)	1.166667	0.0
7978	Monaco au fromage de Hollande	1.666667	0.0
8345	Pomme Poire Williams	1.000000	0.0
10793	Lait demi écrémé en poudre	1.000000	0.0
11443	BN Pocket - Céréales complètes - Vanille	0.800000	0.0
11453	BN vanille	1.000000	0.0
13111	Maxi Cocobat	1.666667	0.0

Conclusions:

- Dataset volumineux **(~ 321K entrées**) avec un faible taux de remplissage (**20%**)
- Analyse des **valeurs manquantes** permettant de choisir les **variables pertinentes et bien renseignées**
- **Imputation** des valeurs manquantes par différentes méthodes parmi lesquelles un **algorithme de Machine learning (IterativeImputer)**
- **Analyse bivariée** pour identifier les **corrélations entre variables**
- **Dévelopement d'une méthode de scoring** prenant en compte:

  - Nutrition_grade
  - Nutriratio
  - Oméga 3

- Avec ce scoring l'application peut aider le consommateur à choisir les meilleurs produits