

Projet 3 : Concevez une application au service de la santé publique

Analyse Exploratoire

Sommaire

- Partie 1 : Introduction
- Partie 2 : Nettoyage du jeu de données
- Partie 3 : Exploration des données
- Partie 4 : Scoring et conclusion

1-1-Objectif et sélection de l'indicateur

- L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.
- 1) Traiter le jeu de données, en :
 - Réfléchissant à une idée d'application avec les variables pertinentes.
 - Nettoyant les données en :
 - mettant en évidence les éventuelles valeurs manquantes, avec au moins 3 méthodes de traitement adaptées.
 - identifiant et en quantifiant les éventuelles valeurs aberrantes.
 - Automatisant ces traitements.

1-1-Objectif et sélection de l'indicateur

- 2) Tout au long de l'analyse, produire des visualisations, effectuer une analyse univariée pour chaque variable intéressante.
- 3) Confirmer ou infirmer les hypothèses à l'aide d'une analyse multivariée.
- 4) Justifier votre idée d'application. Identifier des arguments justifiant la faisabilité (ou non) de l'application à partir des données Open Food Facts.
- 5) Rédiger un rapport d'exploration et pitcher votre idée durant la soutenance du projet.

1-1-Objectif et sélection de l'indicateur

- Idée d'application :
 - De nombreuses personnes atteintes de MICI (Maladie Inflammatoires Chroniques de l'Intestin) doivent faire attention à ne pas surcharger le travail de leur intestin lors de la digestion.
 - Elles doivent avoir une alimentation pauvre en fibres et en graisses pour éviter les selles, tout en augmentant les autres apports nutritionnels (protéines, sucres, énergie, carbohydrates).
 - Nous allons présenter une idée d'application pour mettre en œuvre ce type de régime.

1-2-Importation des librairies et des données

- Les librairies Pandas, Numpy, Matplotlib, Seaborn, Missingno et Sklearn ont été importées pour réaliser cette analyse.
- La base de données analysée est la suivante :
 - <https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-scientist/P2/fr.openfoodfacts.org.products.csv.zip>

1-3-Description de la base de données

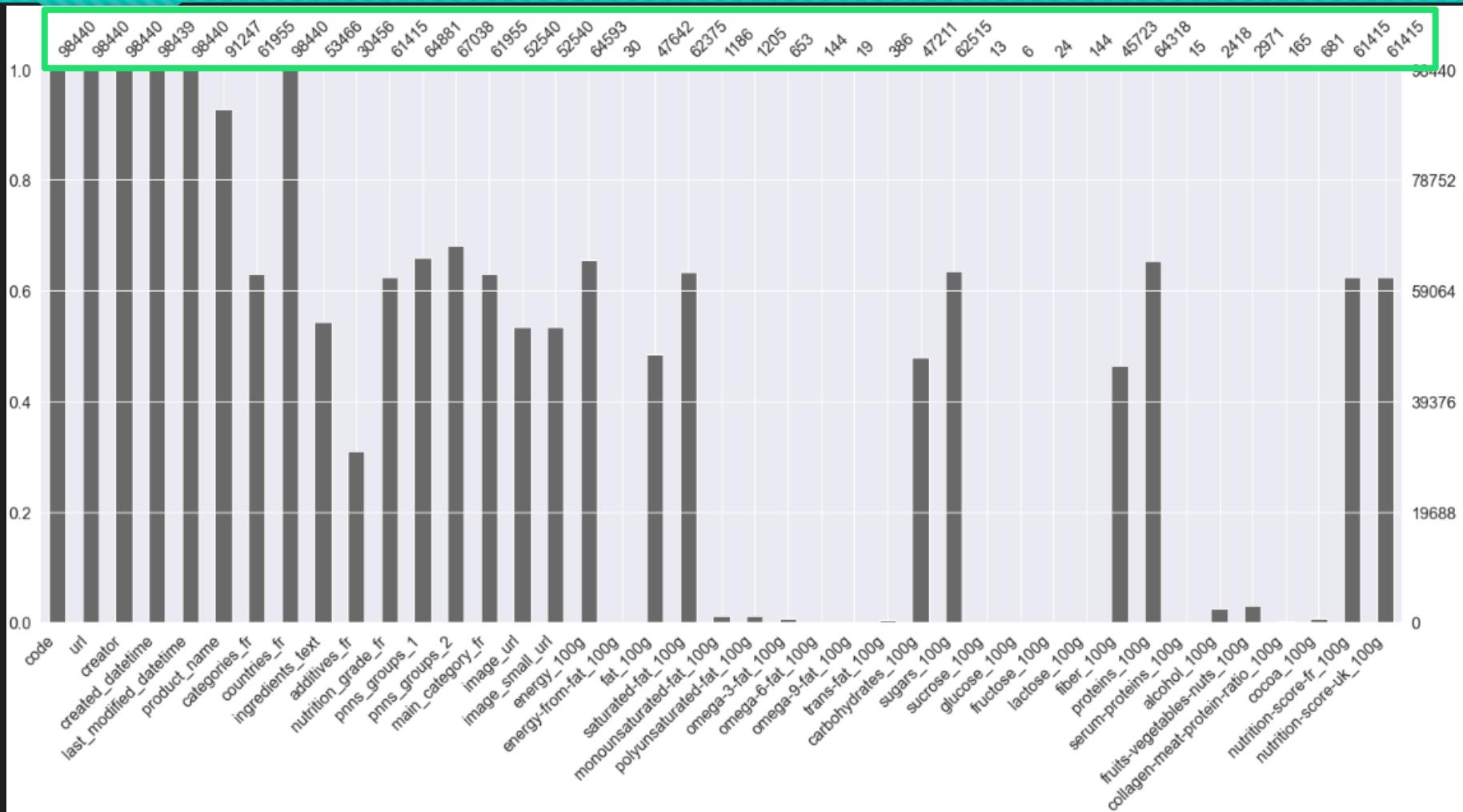
	count	mean	std	min	25%	50%	75%	max		count	mean	std	min	25%	50%	75%	max
no_nutriments	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	zinc_100g	3929.0	0.007950	0.080953	0.000000	0.001150	0.003700	0.007500	4.000000
additives_n	248939.0	1.936024	2.502019	0.00000	0.000000	1.00	3.000	31.00	copper_100g	2106.0	0.025794	0.914247	-6.896552	0.000177	0.000417	0.001000	37.600000
ingredients_from_palm_oil_n	248939.0	0.019659	0.140524	0.00000	0.000000	0.00	0.000	2.00	manganese_100g	1620.0	0.003014	0.028036	0.000000	0.000000	0.001000	0.002000	0.700000
ingredients_from_palm_oil	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	fluoride_100g	79.0	0.012161	0.067952	0.000000	0.000018	0.000060	0.000450	0.560000
ingredients_that_may_be_from_palm_oil_n	248939.0	0.055246	0.269207	0.00000	0.000000	0.00	0.000	6.00	selenium_100g	1168.0	0.003126	0.104503	-0.000002	0.000005	0.000022	0.000061	3.571429
ingredients_that_may_be_from_palm_oil	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	chromium_100g	20.0	0.001690	0.006697	0.000007	0.000011	0.000023	0.000068	0.030000
nutrition_grade_uk	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	molybdenum_100g	11.0	0.000401	0.001118	0.000005	0.000020	0.000039	0.000074	0.003760
energy_100g	261113.0	1141.914605	6447.154093	0.00000	377.000000	1100.00	1674.000	3251373.00	iodine_100g	259.0	0.000427	0.001285	0.000000	0.000015	0.000034	0.000103	0.014700
energy-from-fat_100g	857.0	585.501214	712.809943	0.00000	49.40000	300.00	898.000	3830.00	caffeine_100g	78.0	1.594563	6.475588	0.000000	0.015500	0.021000	0.043000	42.280000
fat_100g	243891.0	12.730379	17.578747	0.00000	0.000000	5.00	20.000	714.29	taurine_100g	29.0	0.145762	0.172312	0.001800	0.035000	0.039000	0.400000	0.423000
saturated-fat_100g	229554.0	5.129932	8.014238	0.00000	0.000000	1.79	7.140	550.00	ph_100g	49.0	6.425698	2.047841	0.000000	6.300000	7.200000	7.400000	8.400000
butyric-acid_100g	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	fruits-vegetables-nuts_100g	3036.0	31.458587	31.967918	0.000000	0.000000	23.000000	51.000000	100.000000
caproic-acid_100g	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	collagen-meat-protein-ratio_100g	165.0	15.412121	3.753028	8.000000	12.000000	15.000000	15.000000	25.000000
caprylic-acid_100g	1.0	7.400000	NaN	7.40000	7.400000	7.40	7.400	7.40	cocoa_100g	948.0	49.547785	18.757932	6.000000	32.000000	50.000000	64.250000	100.000000
capric-acid_100g	2.0	6.040000	0.226274	5.88000	5.960000	6.04	6.120	6.20	chlorophyl_100g	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lauric-acid_100g	4.0	36.136182	24.101433	0.04473	34.661183	47.60	49.075	49.30	carbon-footprint_100g	268.0	341.700764	425.211439	0.000000	98.750000	195.750000	383.200000	2842.000000
myristic-acid_100g	1.0	18.900000	NaN	18.90000	18.900000	18.90	18.900	18.90	nutrition-score-fr_100g	221210.0	9.165535	9.055903	-15.000000	1.000000	10.000000	16.000000	40.000000
palmitic-acid_100g	1.0	8.100000	NaN	8.10000	8.100000	8.10	8.100	8.10	nutrition-score-uk_100g	221210.0	9.058049	9.183589	-15.000000	1.000000	9.000000	16.000000	40.000000
stearic-acid_100g	1.0	3.000000	NaN	3.00000	3.000000	3.00	3.000	3.00	glycemic-index_100g	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
arachidic-acid_100g	24.0	10.752667	4.019993	0.06400	7.275000	12.85	13.375	15.40	water-hardness_100g	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

2-1-Suppression des variables inutiles et des produits étrangers pour l'analyse

- Les variables conservées :

code	monounsaturated-fat_100g
url	polyunsaturated-fat_100g
creator	omega-3-fat_100g
created_datetime	omega-6-fat_100g
last_modified_datetime	omega-9-fat_100g
product_name	trans-fat_100g
categories_fr	carbohydrates_100g
countries_fr	sugars_100g
ingredients_text	sucrose_100g
additives_fr	glucose_100g
nutrition_grade_uk	fructose_100g
nutrition_grade_fr	lactose_100g
pnns_groups_1	fiber_100g
pnns_groups_2	proteins_100g
main_category_fr	serum-proteins_100g
image_url	alcohol_100g
image_small_url	fruits-vegetables-nuts_100g
energy_100g	collagen-meat-protein-ratio_100g
energy-from-fat_100g	cocoa_100g
fat_100g	chlorophyl_100g
saturated-fat_100g	nutrition-score-fr_100g
	nutrition-score-uk_100g
	glycemic-index_100g

2-2-Suppression des colonnes vides



2-3-Suppression des valeurs dupliquées

- Il n'y a pas de ligne 100% identiques. Néanmoins un même produit peut avoir été renseignées sur plusieurs lignes différentes. Exemple :

	product_name	brands	energy_100g	nutrition-score-fr_100g
221137	100% Pur Boeuf 5% MG	Bigard	NaN	NaN
221136	100% Pur Boeuf 5% MG	Bigard	NaN	NaN
221135	100% Pur Boeuf 5% MG	Bigard	NaN	NaN
174934	12 crevettes apéritives	Picard	1120.0	17.0
219537	12 crevettes apéritives	Picard	1120.0	17.0

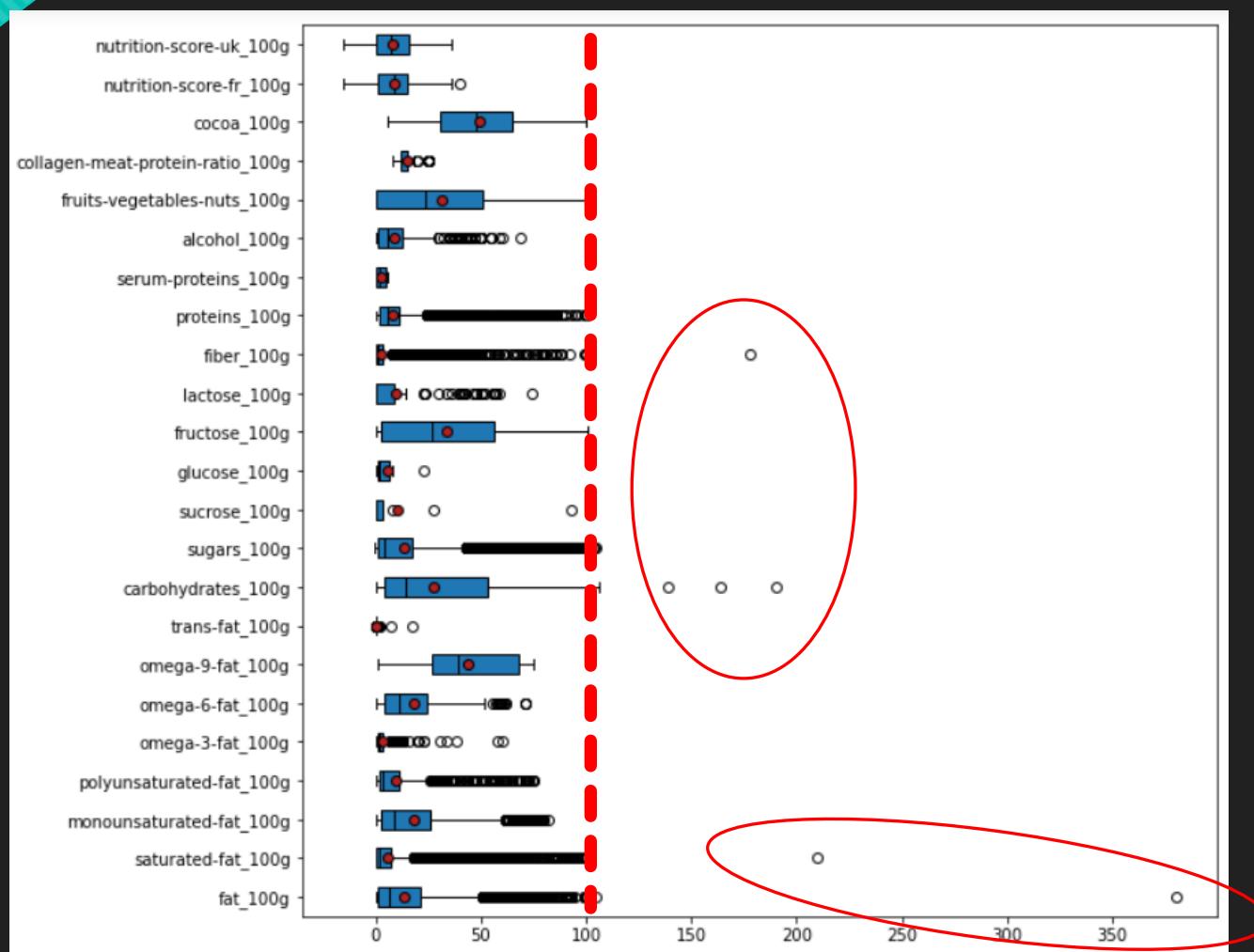
- On supprime tous les doublons. 8196 doublons sont supprimés.

2-4-Remise en conformité des types

code	object	
url	object	
creator	object	
created_datetime	object	
last_modified_datetime	object	
product_name	object	
categories_fr	object	
countries_fr	object	
ingredients_text	object	
additives_fr	object	
nutrition_grade_fr	object	
pnns_groups_1	object	
pnns_groups_2	object	
main_category_fr	object	
image_url	object	
image_small_url	object	
energy_100g	float64	
energy-from-fat_100g	float64	
fat_100g	float64	
		polyunsaturated-fat_100g float64
		omega-3-fat_100g float64
		omega-6-fat_100g float64
		omega-9-fat_100g float64
		trans-fat_100g float64
		carbohydrates_100g float64
		sugars_100g float64
		sucrose_100g float64
		glucose_100g float64
		fructose_100g float64
		lactose_100g float64
		fiber_100g float64
		proteins_100g float64
		serum-proteins_100g float64
		alcohol_100g float64
		fruits-vegetables-nuts_100g float64
		collagen-meat-protein-ratio_100g float64
		cocoa_100g float64
		nutrition-score-fr_100g float64

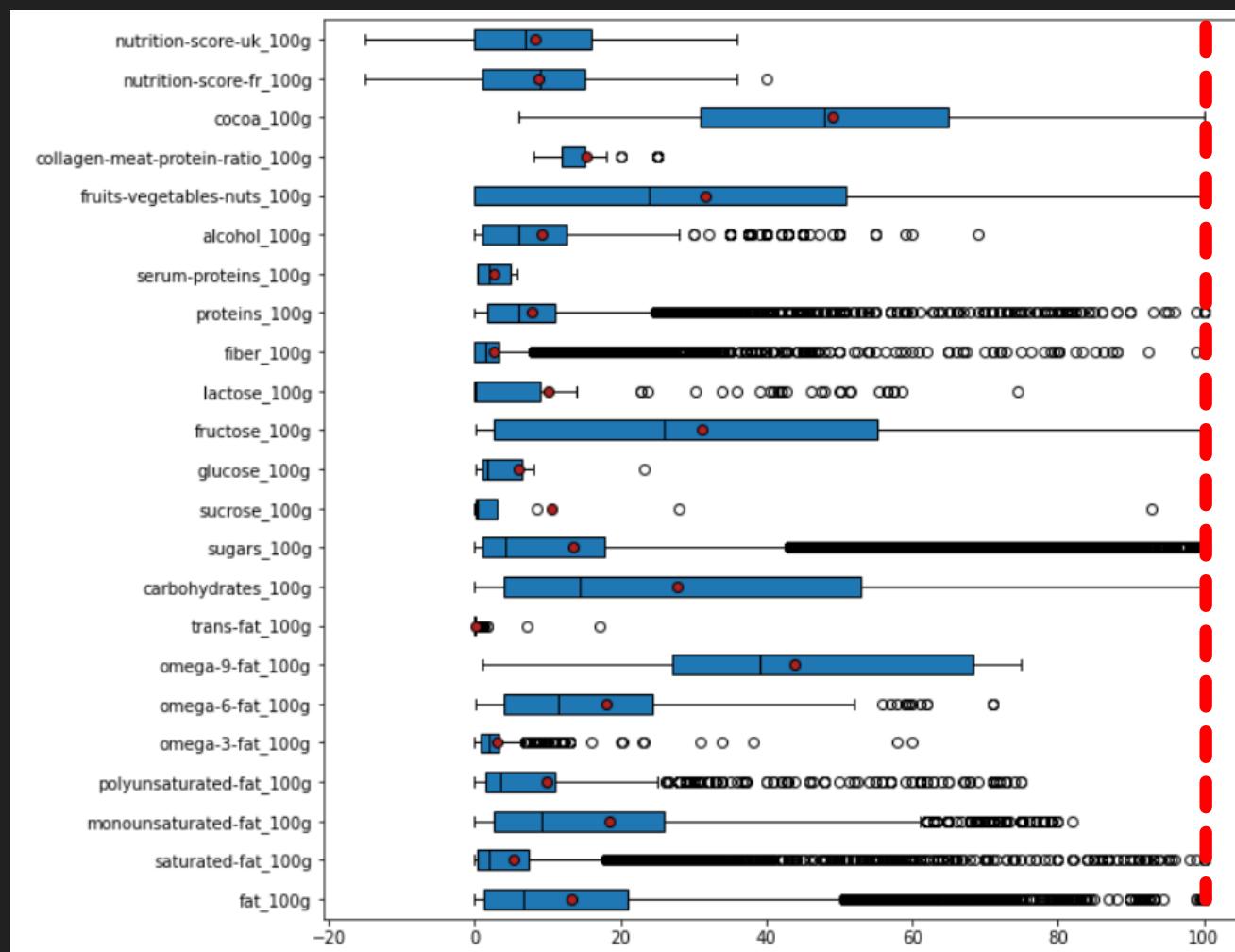
2-5-Suppression des valeurs abérrantes

○ Avant :

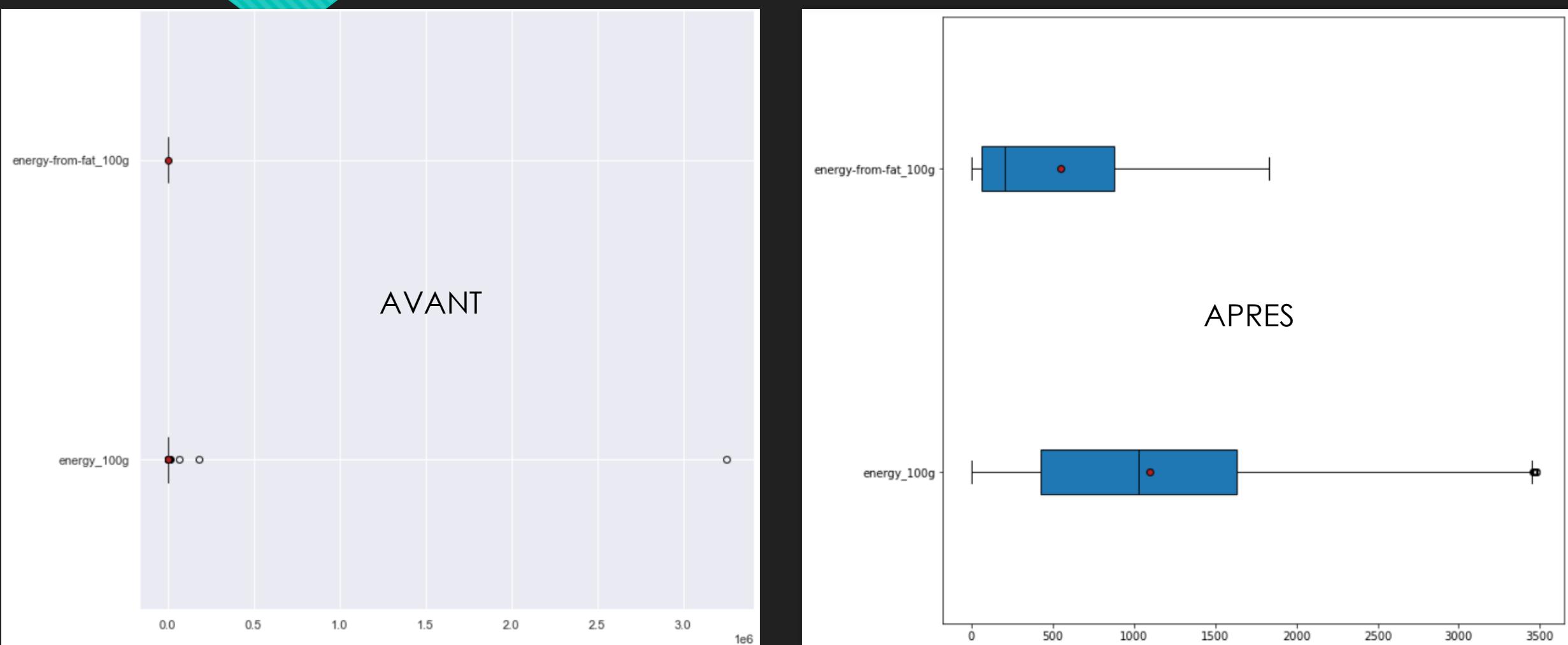


2-5-Suppression des valeurs abérrantes

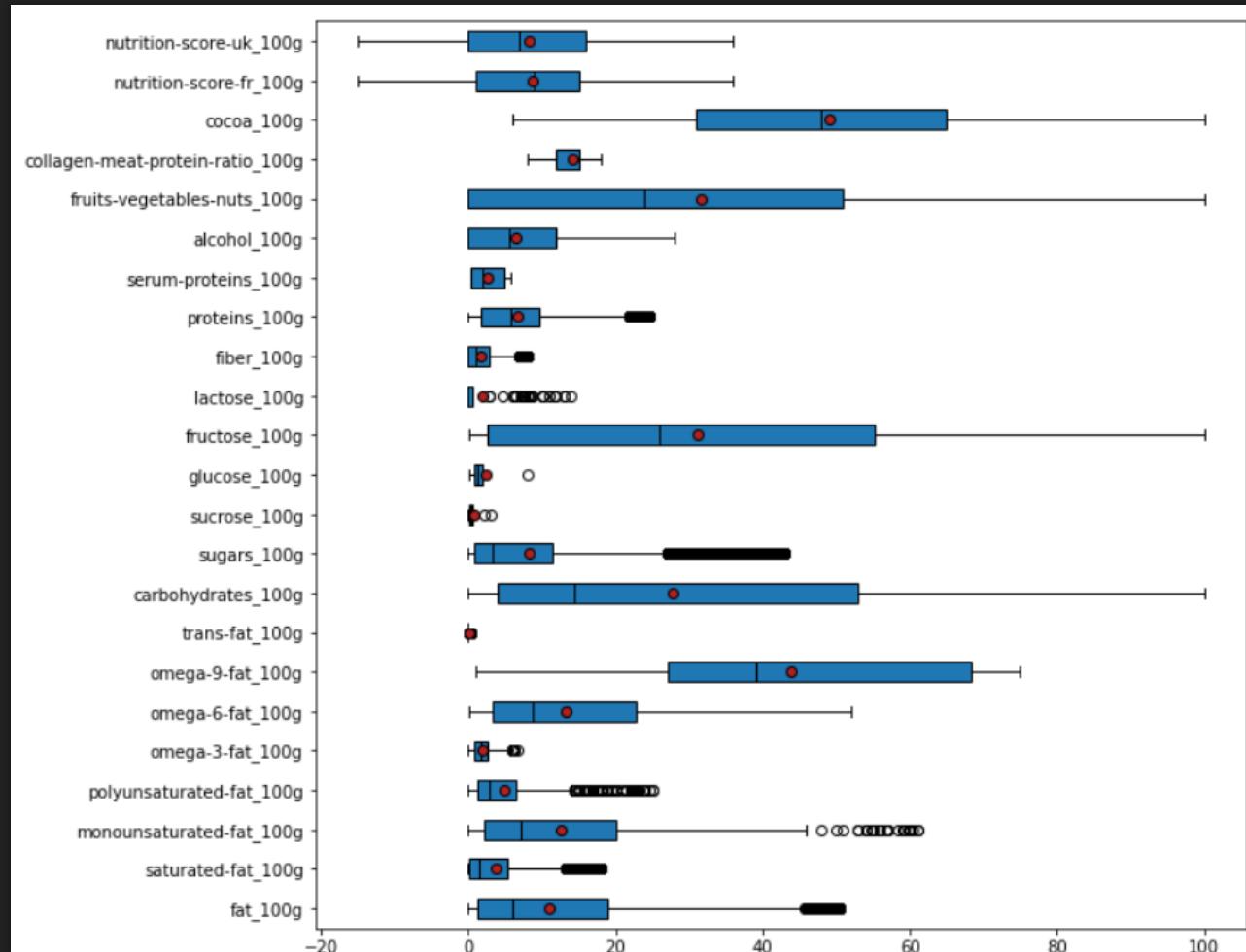
○ Après :



2-6-Suppression des outliers (méthode interquartile)



2-6-Suppression des outliers (méthode interquartile)

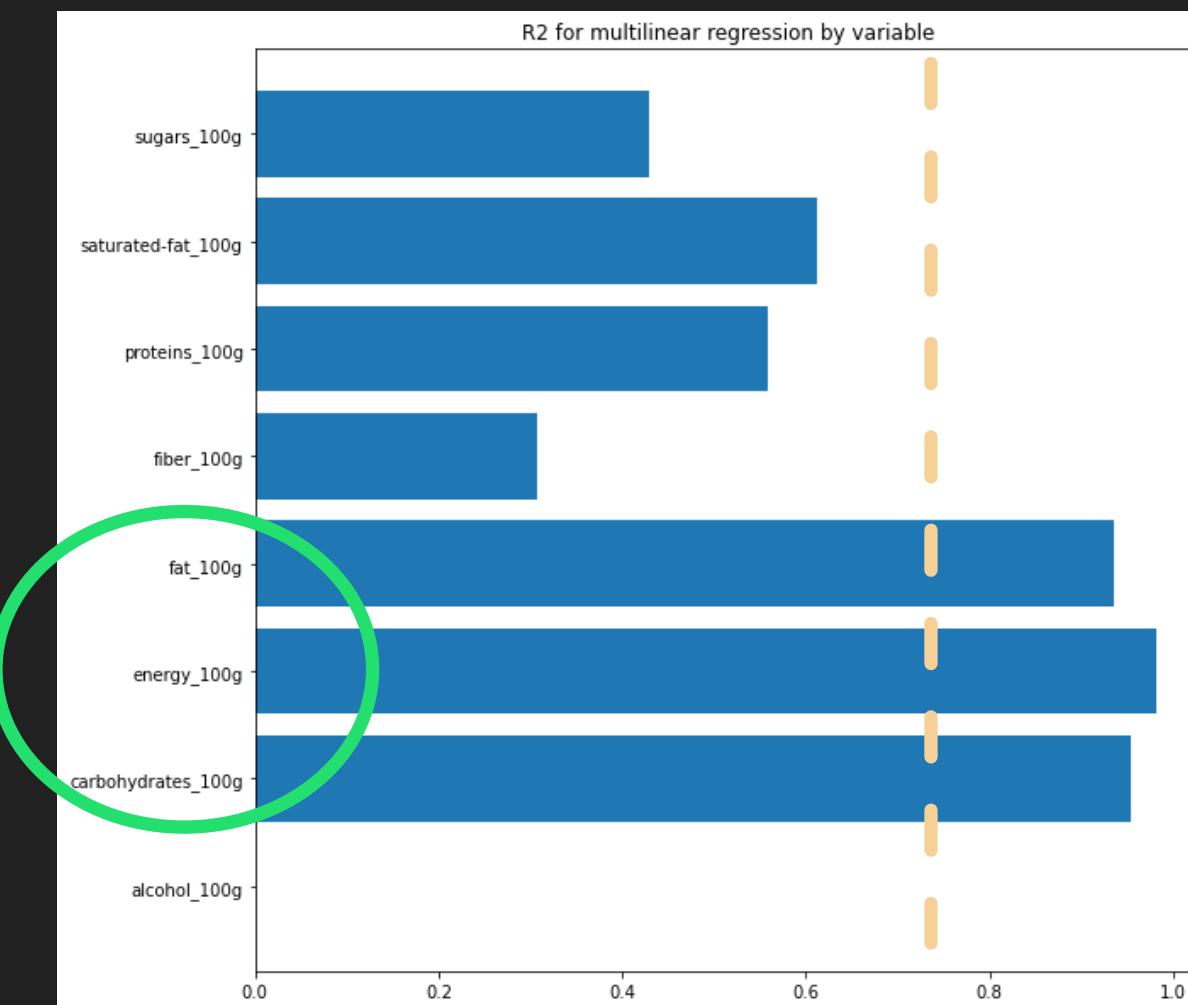


2-7-Régression multilinéaire sur la base

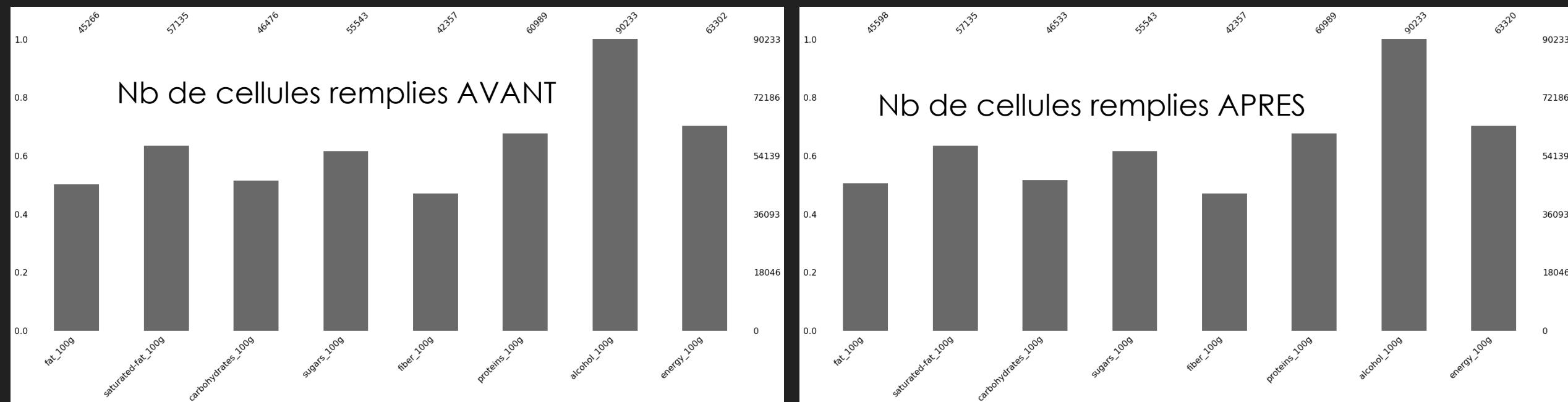
- Energy_100g étant une combinaison linéaire des autres composantes (source : <https://jmmanger.fr/JM-savoir/energie-kcal>) on choisit de faire une régression multilinéaire pour remplir les cellules.
- On fixe un seuil du coefficient de détermination à 0.75 pour effectuer le remplacement

2-7-Régression multilinéaire sur la base

Imputation effectuée



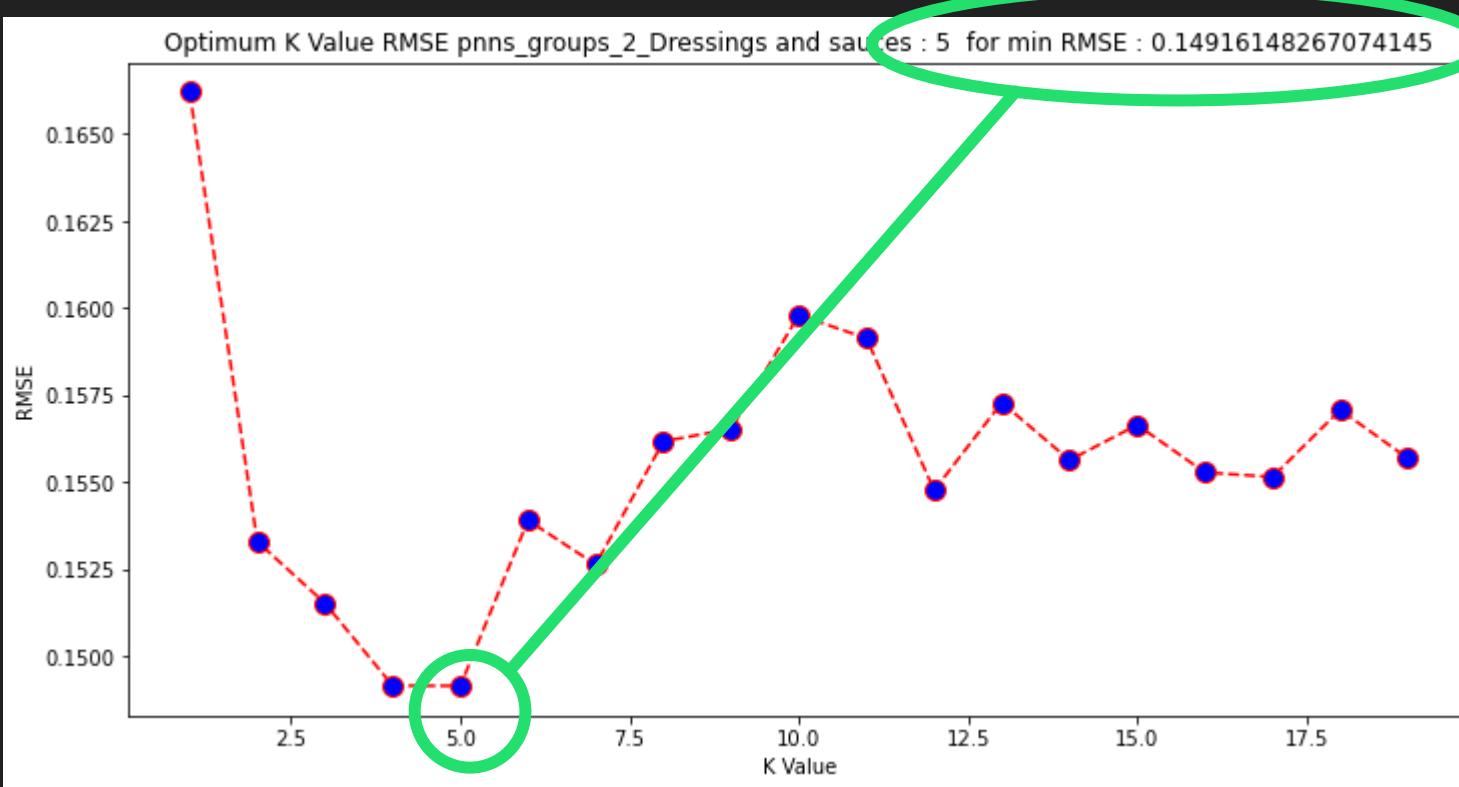
2-7-Régression multilinéaire sur la base



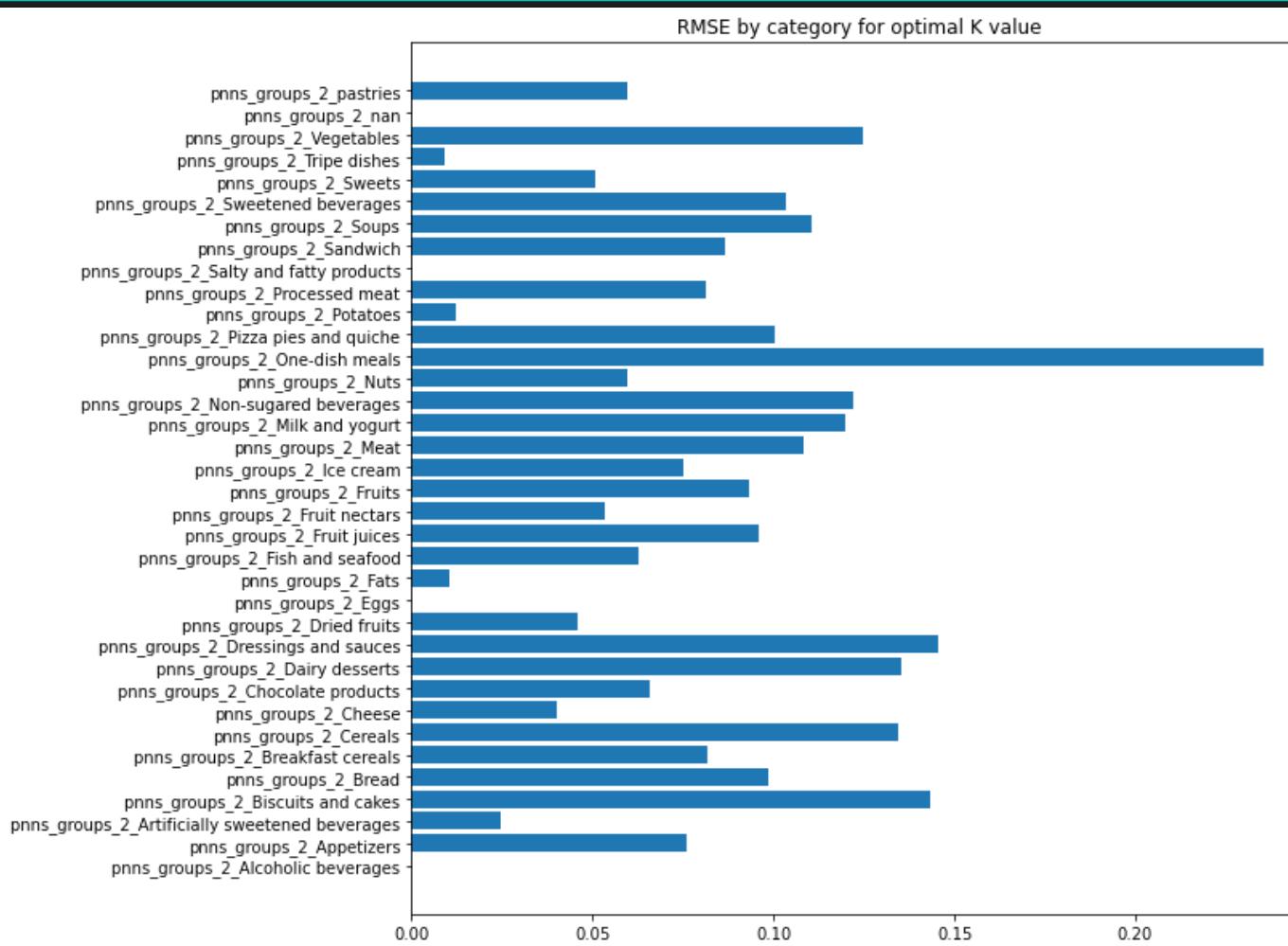
2-8-Remplissage des valeurs catégorielles manquantes (méthode OneHotEncoder+K-NN)

- On va remplir la colonne pnns-2 (35 valeurs possibles) pour catégoriser les lignes.
- En réalisant un HotEncoding sur la colonne catégorielle pnns_groups_2
- En effectuant ensuite un K-NN imput en prenant les variables _100g
- La mesure de l'efficacité de l'imputation est mesurée sur un échantillon de 2500 individus avec le RMSE

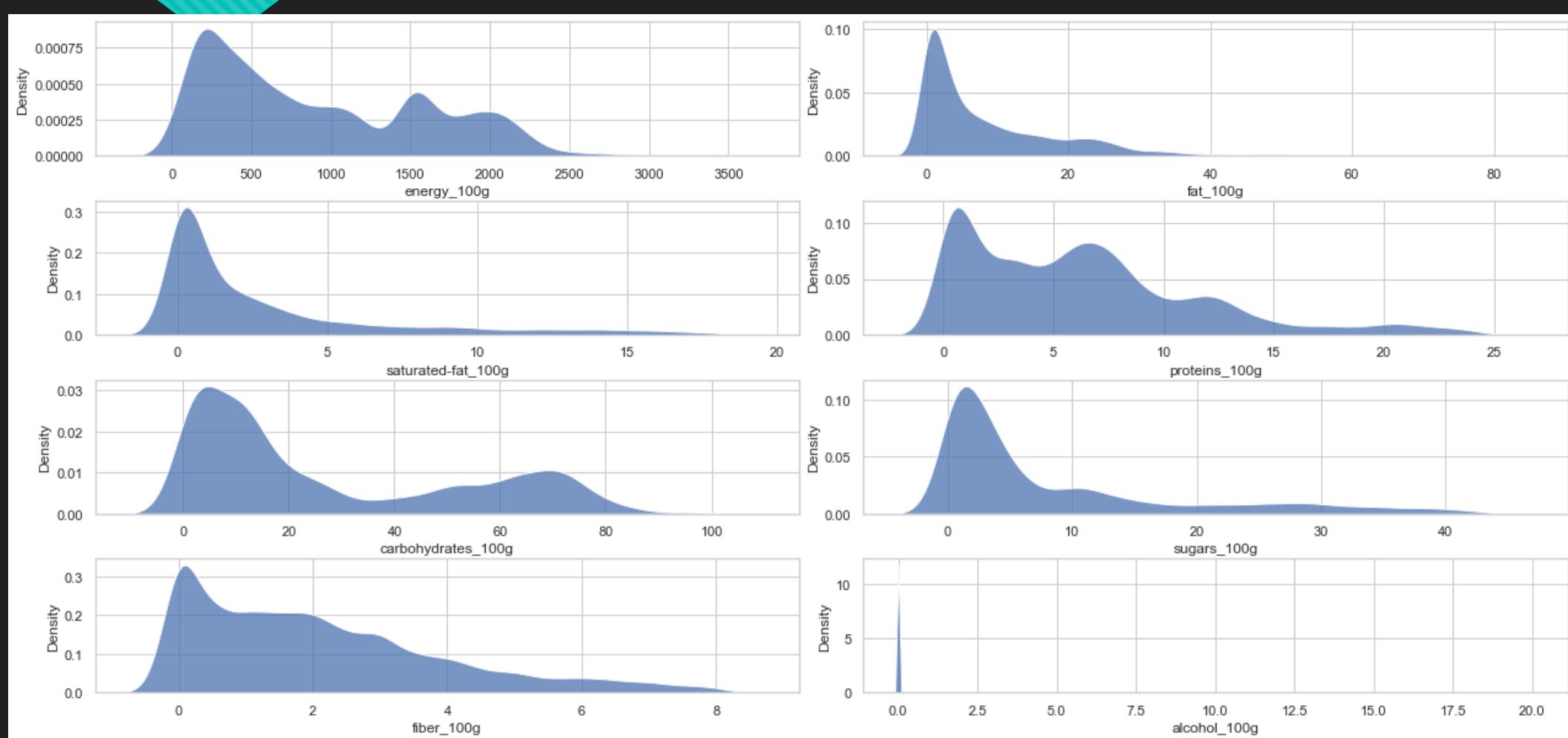
2-8-Remplissage des valeurs catégorielles manquantes (méthode HotEncoding+K-NN)



2-8-Remplissage des valeurs catégorielles manquantes (méthode HotEncoding+K-NN)



3-1-Analyse univariée



3-1-Analyse univariée

	count	mean	std	min	25%	50%	75%	max
energy_100g	22621.0	926.604502	698.657996	0.0	310.0	729.00	1523.0	3464.000000
fat_100g	22621.0	9.007942	11.096148	0.0	1.0	4.40	14.0	81.410768
saturated-fat_100g	22621.0	3.053632	4.185160	0.0	0.2	1.10	4.0	18.000000
carbohydrates_100g	22621.0	27.685410	26.302938	0.0	5.9	15.00	52.1	100.000000
sugars_100g	22621.0	8.566406	10.472359	0.0	1.3	3.70	12.0	43.000000
fiber_100g	22621.0	2.112936	1.901332	0.0	0.5	1.70	3.1	8.000000
proteins_100g	22621.0	6.362030	5.407720	0.0	1.7	5.67	9.0	24.700000
alcohol_100g	22621.0	0.006436	0.251146	0.0	0.0	0.00	0.0	20.000000

3-2-Analyse bivariée (ANOVA)

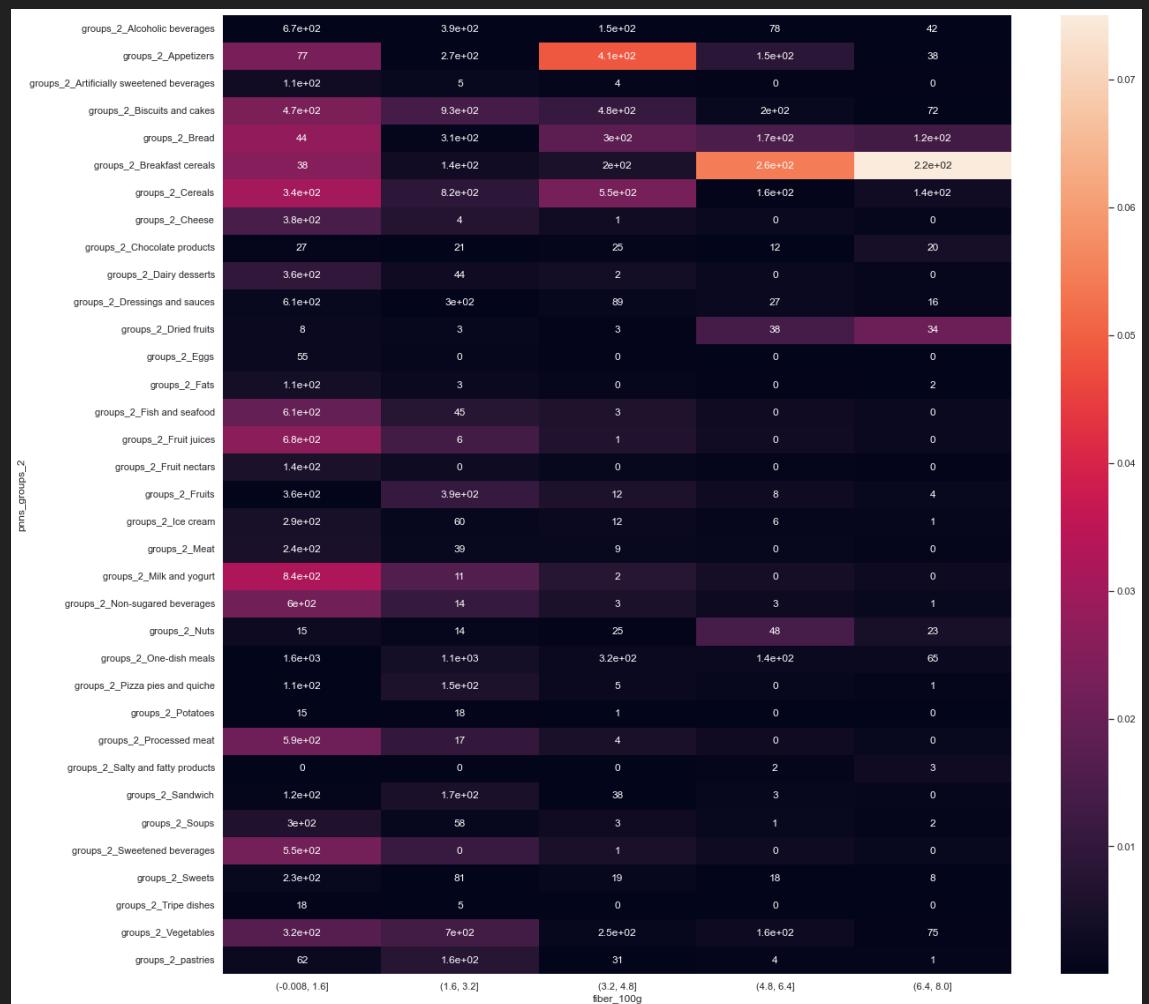
- Les deux variables pour l'ANOVA :
 - d'une part les catégories pnns_2
 - d'autres part la variable quantitative fiber_100g divisée en cinq classes suivant le taux de fibre par aliment

3-2-Analyse bivariée (ANOVA)

- valeur de Khi_2 : 13588.331054085862
- nb de degrés de liberté : 136

Loi de Khi-deux																		
p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001				
ddl	77,7551	83,8516	86,9233	90,3867	95,7046	100,6236	106,8056	132,8063	140,2326	146,5674	153,9182	158,9502	163,6482	173,6174				
120	77,7551	83,8516	86,9233	90,3867	95,7046	100,6236	106,8056	132,8063	140,2326	146,5674	153,9182	158,9502	163,6482	173,6174				
140	93,9256	100,6548	104,0344	107,8149	113,6593	119,0293	125,7581	153,8537	161,8270	168,6130	176,4709	181,8403	186,8468	197,4508				

- La valeur de Khi_2 : 13588.3310 est bien supérieure à 247.3705 (ddl 180). On est certain à plus de 99,9% que nos deux variables ne sont pas indépendantes.
- C'est-à-dire que selon le groupe le taux de fibres sera significativement différent.

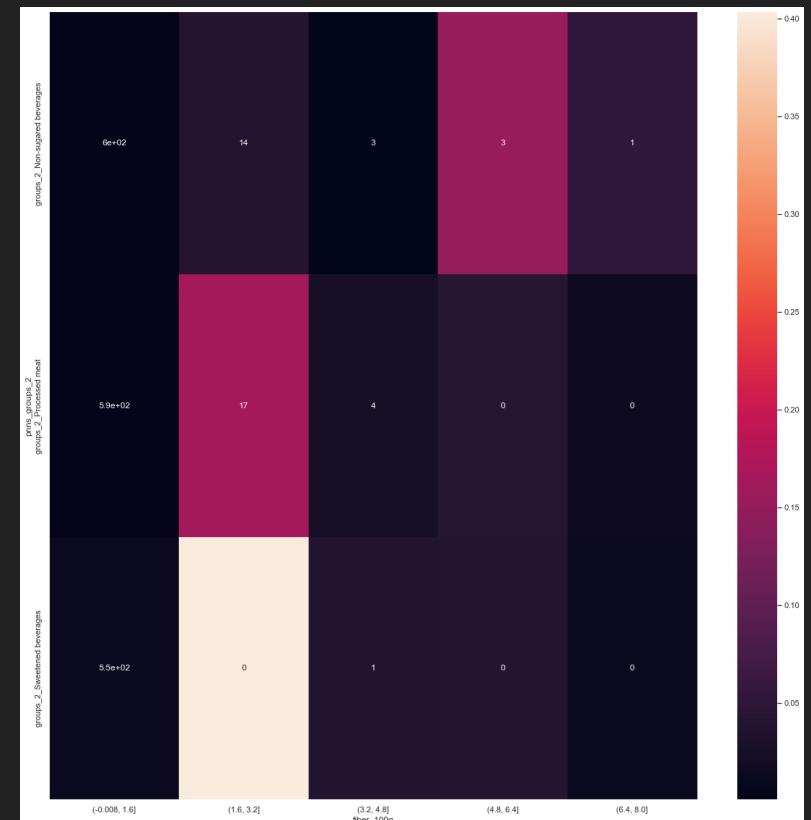


3-2-Analyse bivariée (ANOVA)

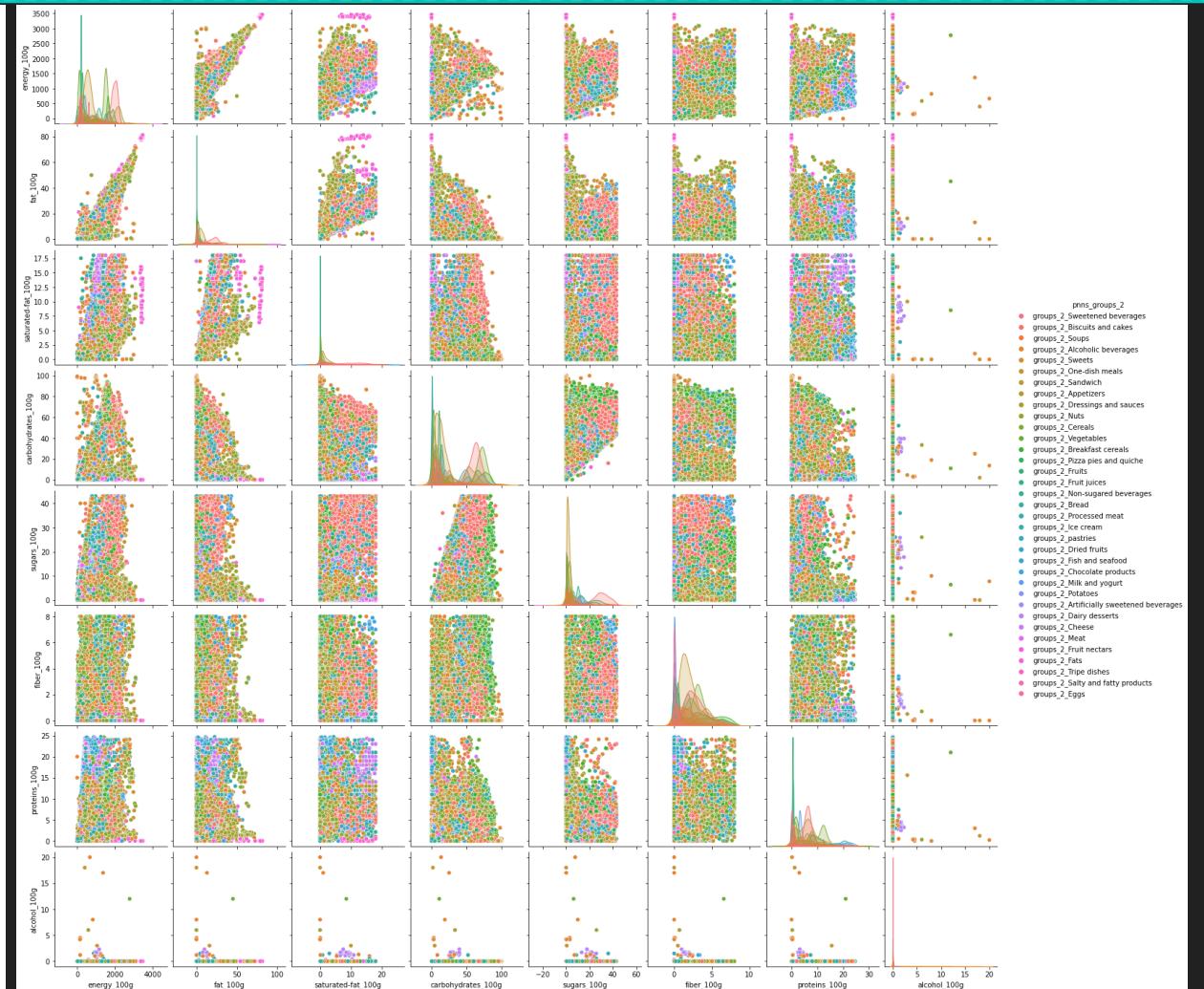
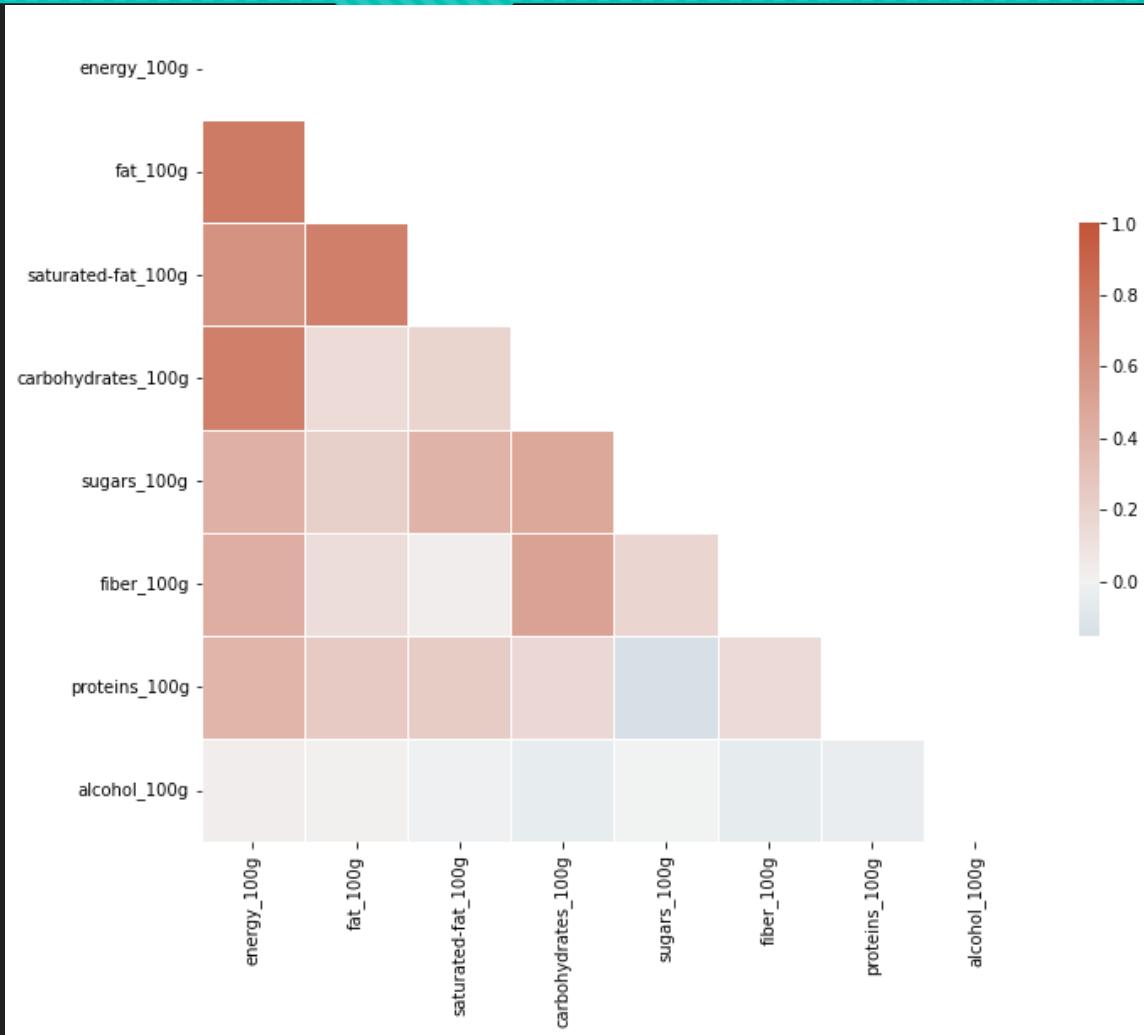
- Plus dans le détail, on regarde une ANOVA pour trois groupes à la répartition similaire pour les fibres : 'groups_2_Processed meat', 'groups_2_Sweetened beverages', 'groups_2_Non-sugared beverages'
- valeur de Khi₂ : 21.2666
- nb de degrés de liberté : 8

Loi de Khi-deux															
Le tableau donne x tel que P(K > x) = p															
p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001	
ddl	8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245

- 21.2666 > 15.5073 pour ddl=8 et risque alpha 5%
- Le taux de fibre reste significativement différent selon le groupe avec un taux de confiance de 95%

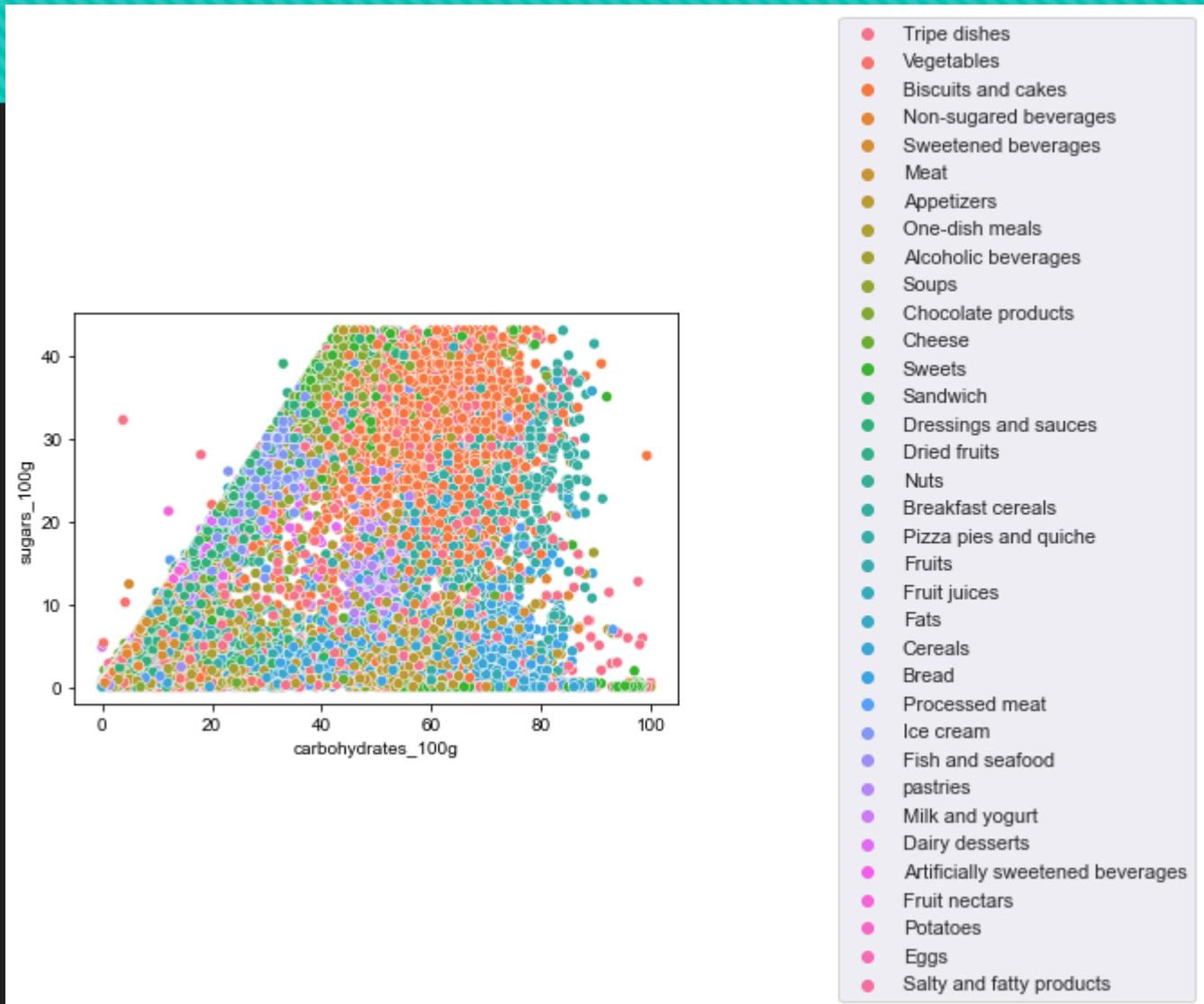


3-3-Analyse bivariée (matrice de corrélation)



3-3-Analyse bivariée (matrice de corrélation)

- La frontière correspond à une ligne où $\% \text{sucres} = \% \text{carbohydrates}$. Elle montre que les sucres sont une partie des carbohydrates. Les valeurs à gauche de la ligne sont des valeurs aberrantes



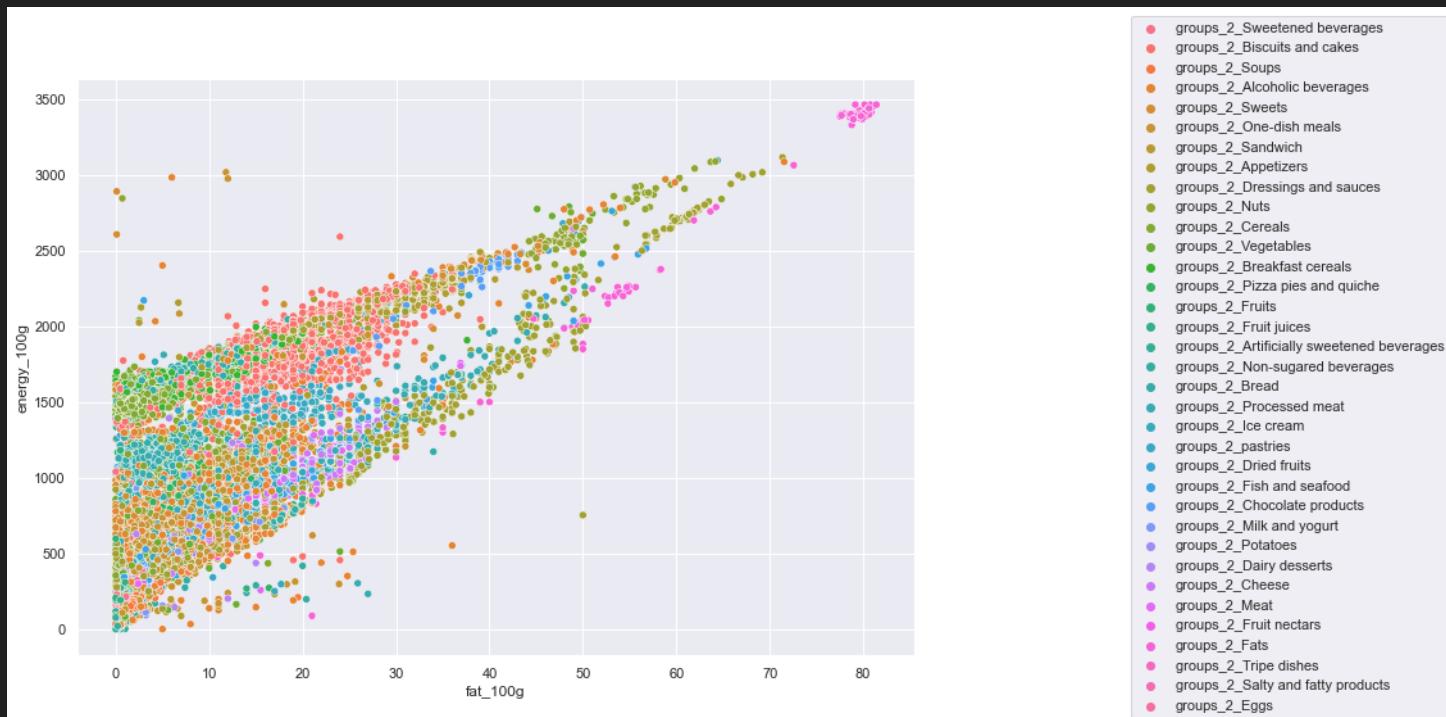
3-3-Analyse bivariée (matrice de corrélation)

- Sur internet on peut voir que 100g de carbohydrates correspond à 1700kJ.
- Les données convergent bien vers cette valeur.

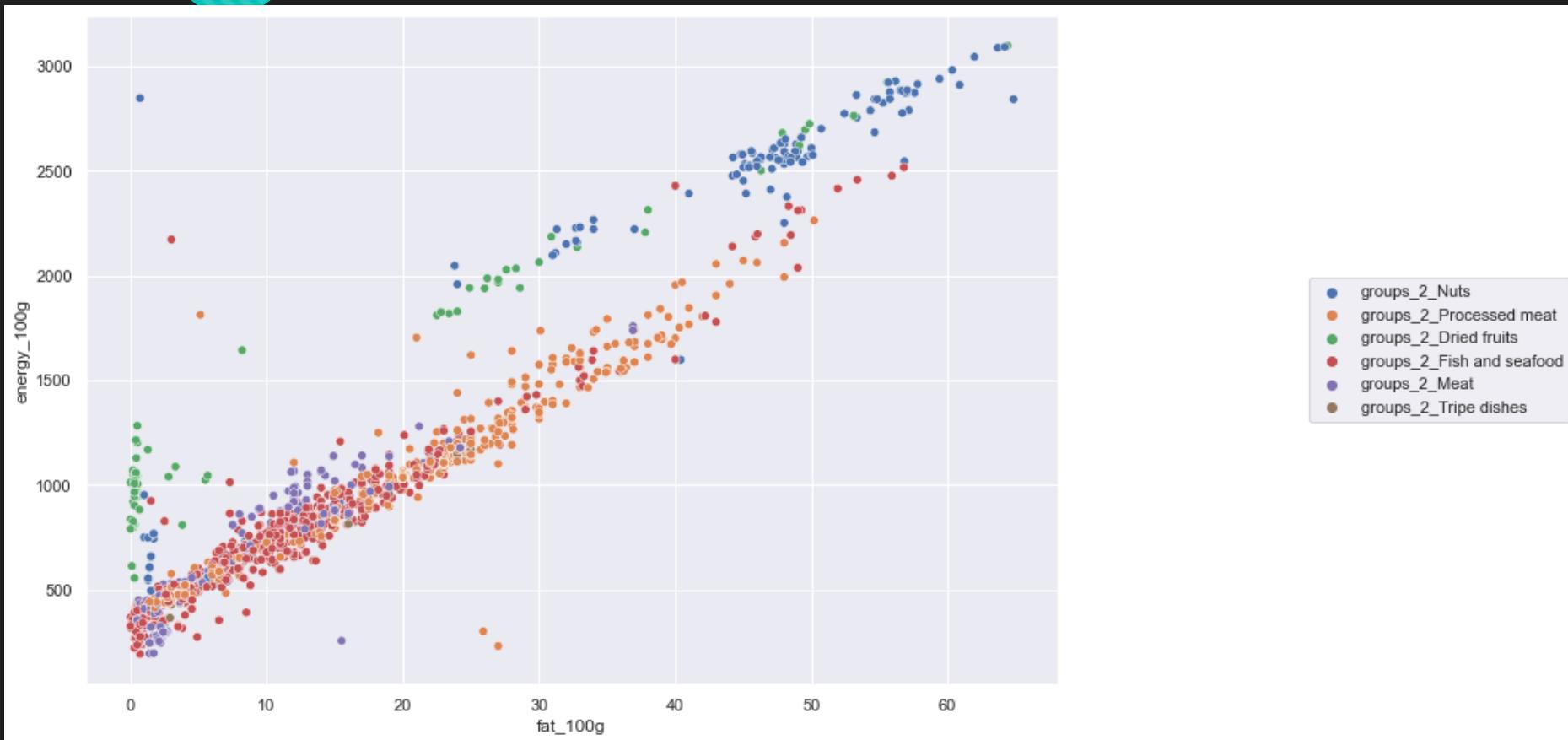


3-3-Analyse bivariée (matrice de corrélation)

- Sur internet on peut voir que 100g de Fat correspond à 3770kJ.
- Les données convergent bien vers cette valeur.

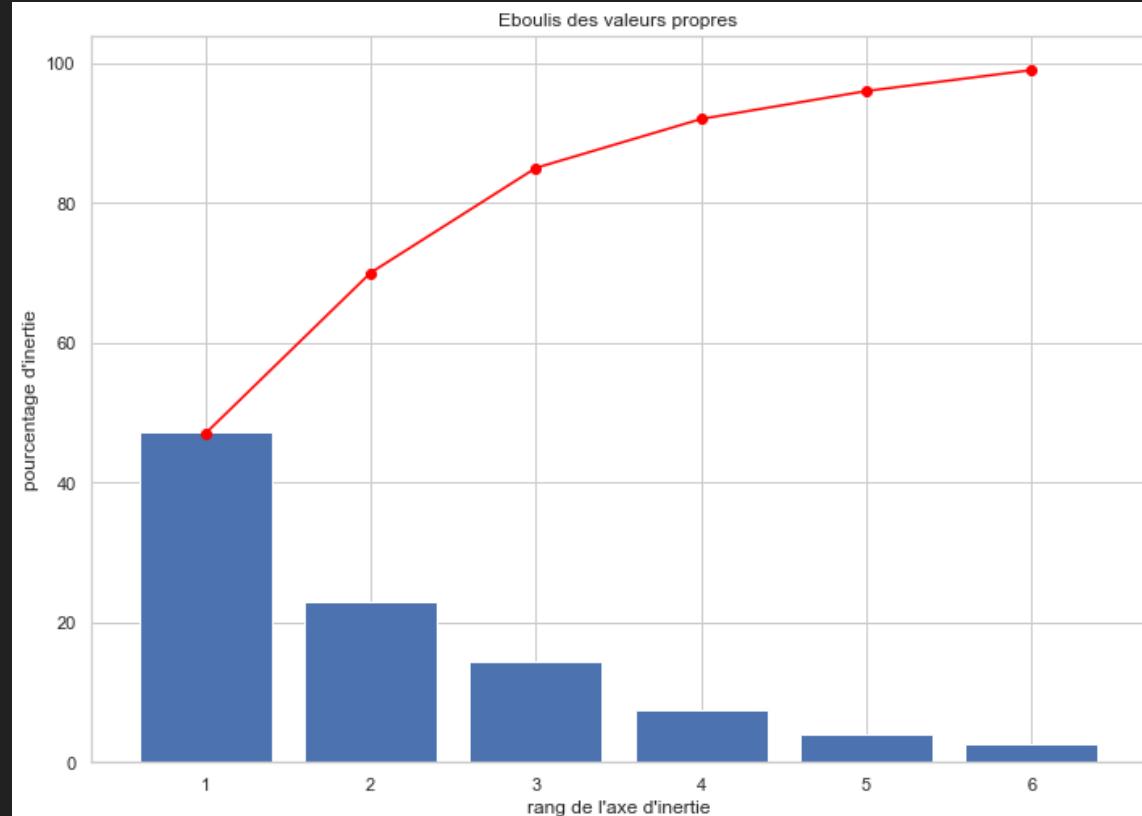


3-3-Analyse bivariée (matrice de corrélation)



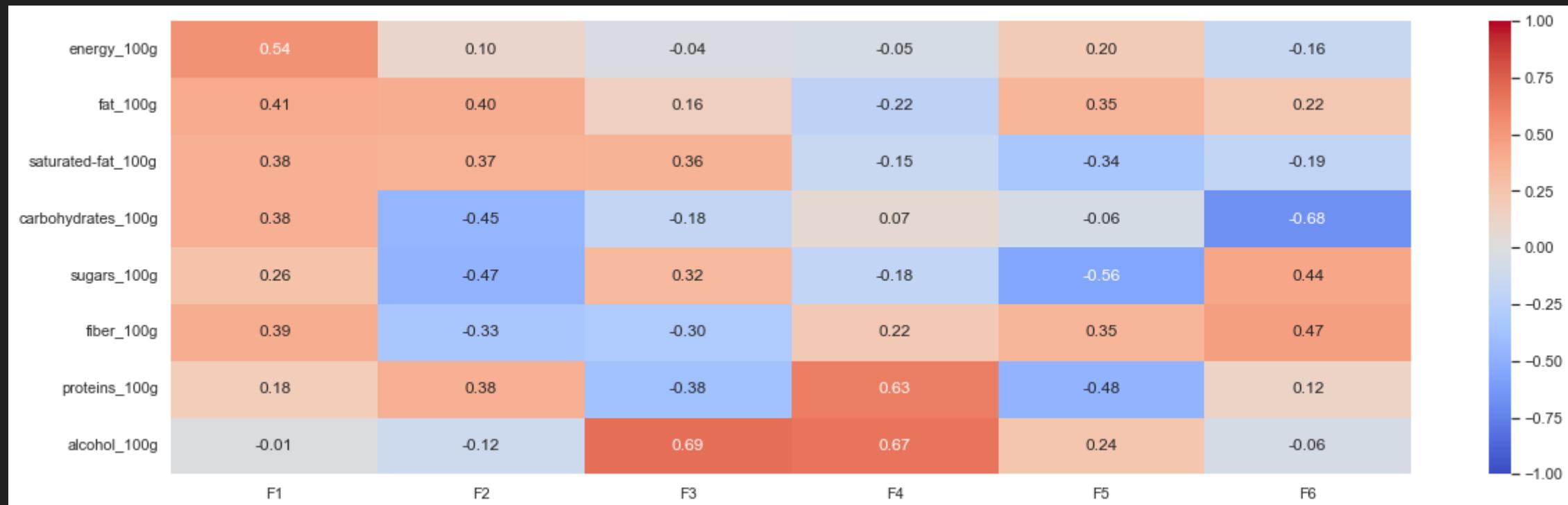
3-4-Réduction des dimensions (méthode ACP)

- Variance expliquée et diagramme des éboulis :
- Critère de Kaiser pour 8 dimensions => on élimine de l'analyse les composantes qui retiennent moins de $100/8=12,5\%$ de l'inertie.
C'est à dire que les composantes retenues sont F1, F2 et F3.



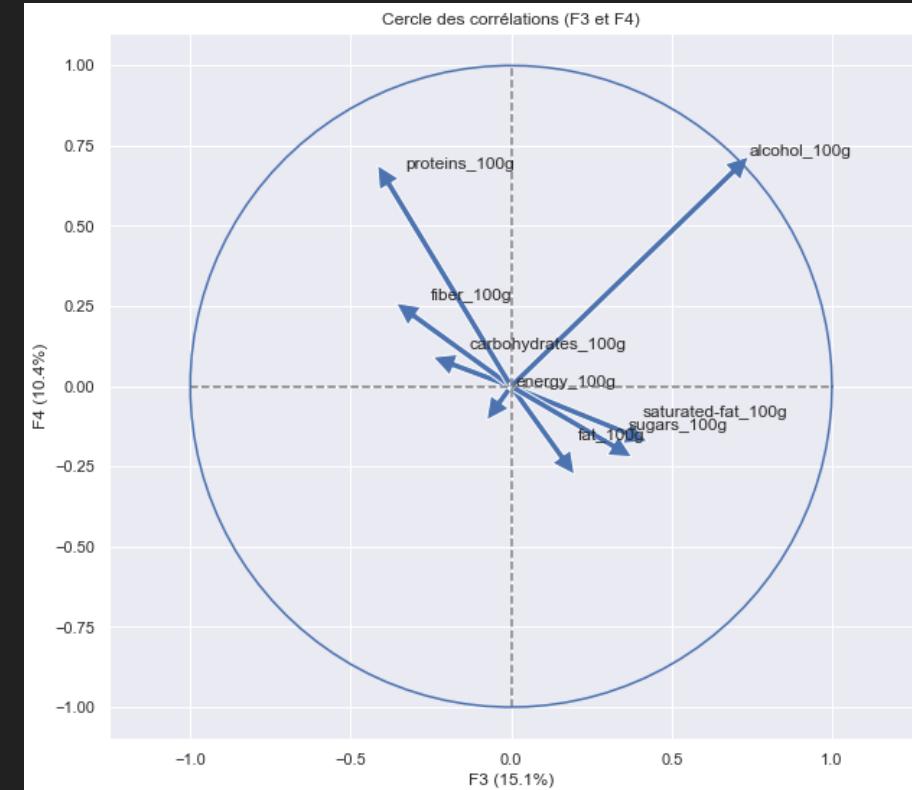
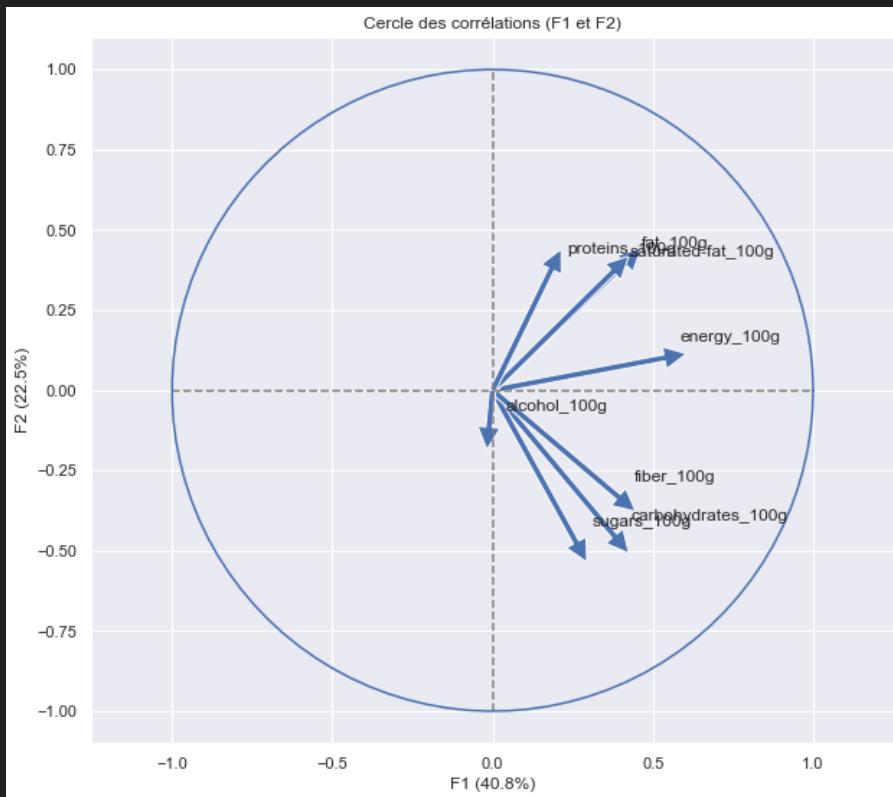
3-4-Réduction des dimensions (méthode ACP)

○ Description des composantes principales:



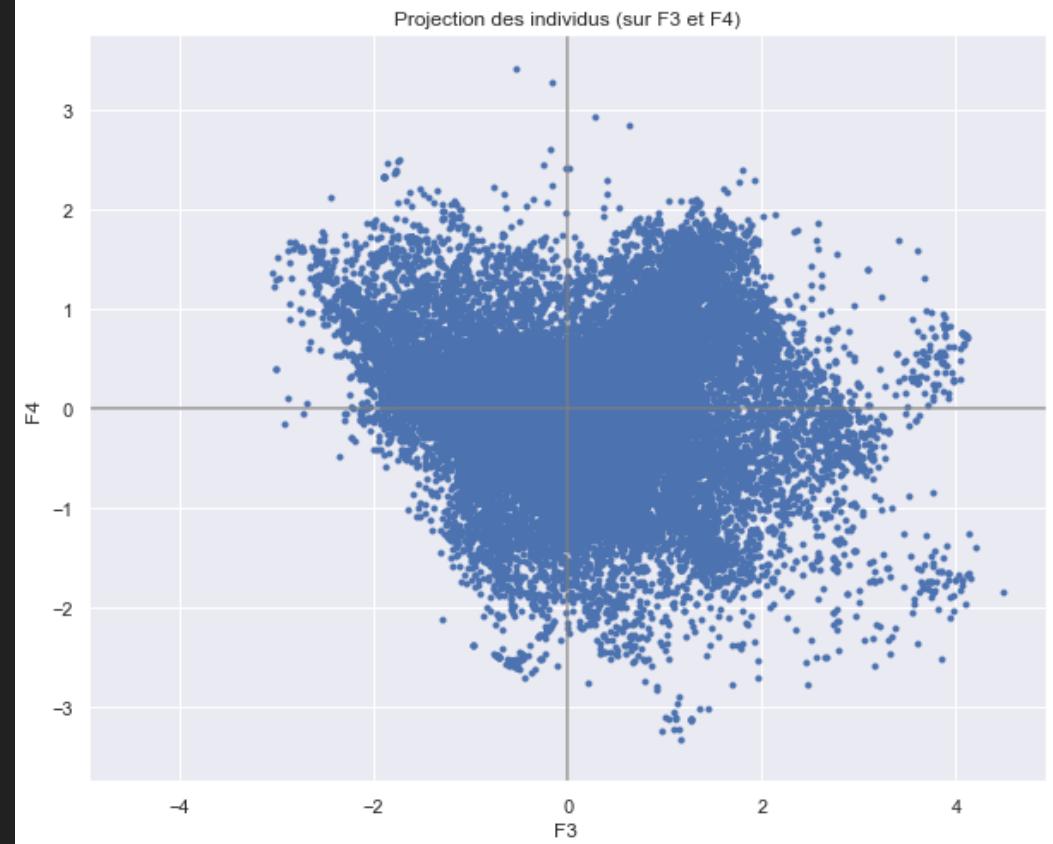
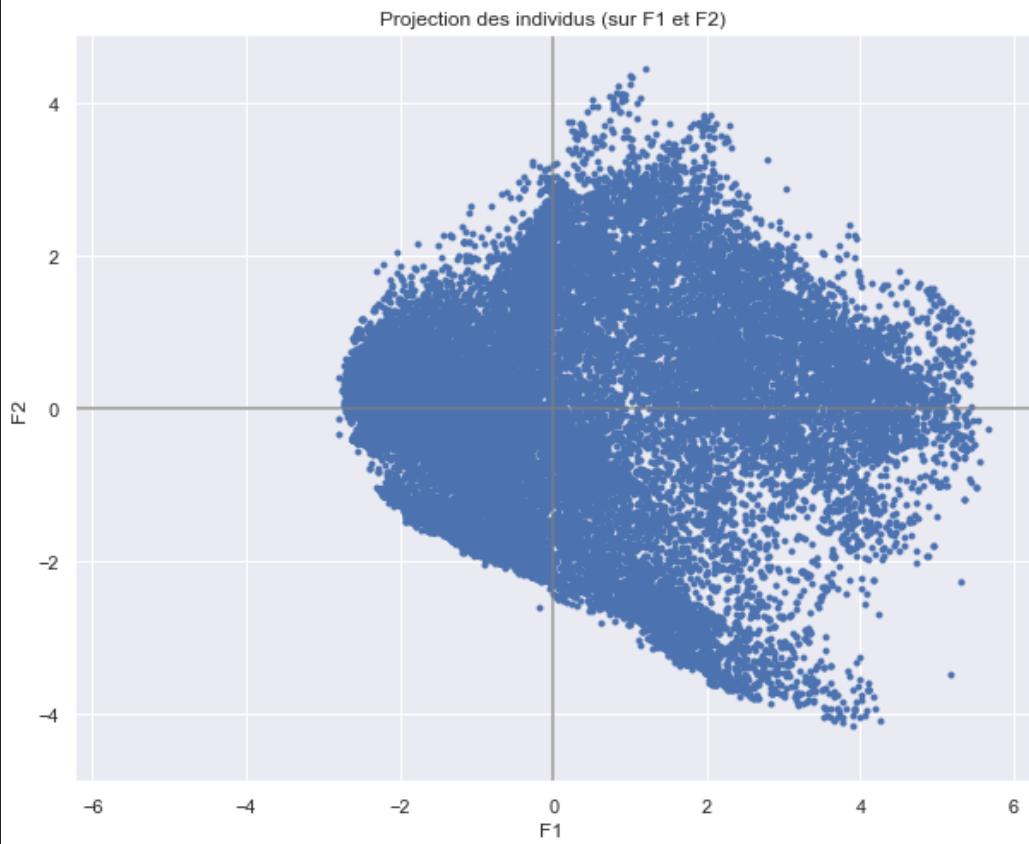
3-5-Réduction des dimensions (méthode ACP)

○ Graphes de corrélation:



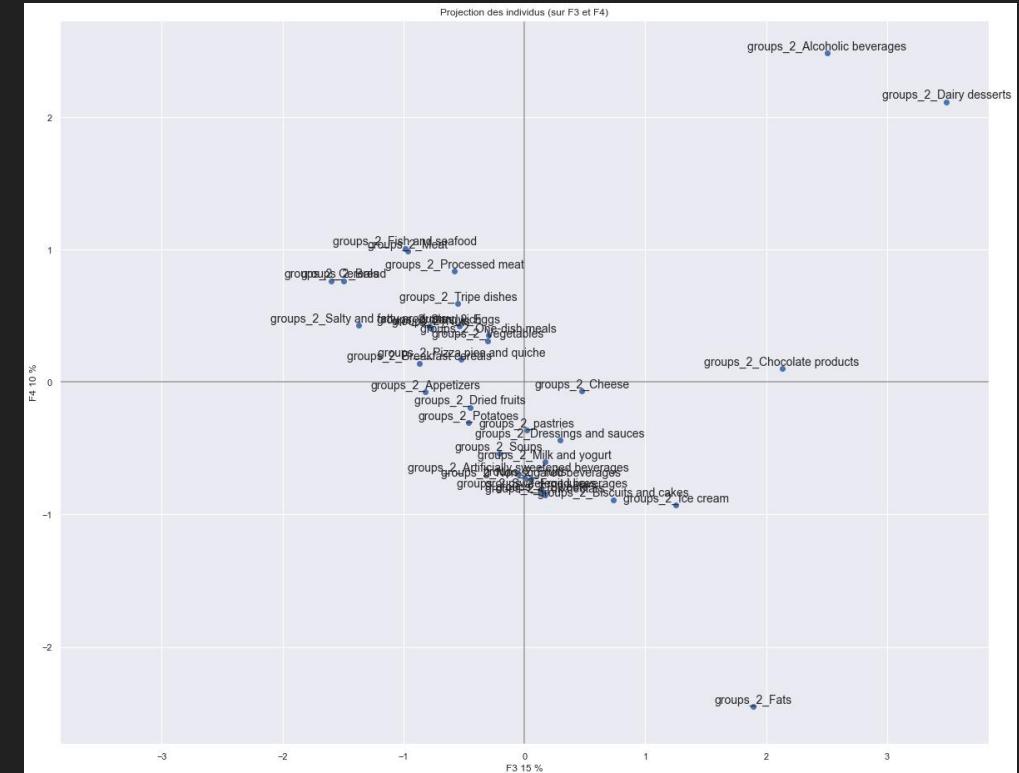
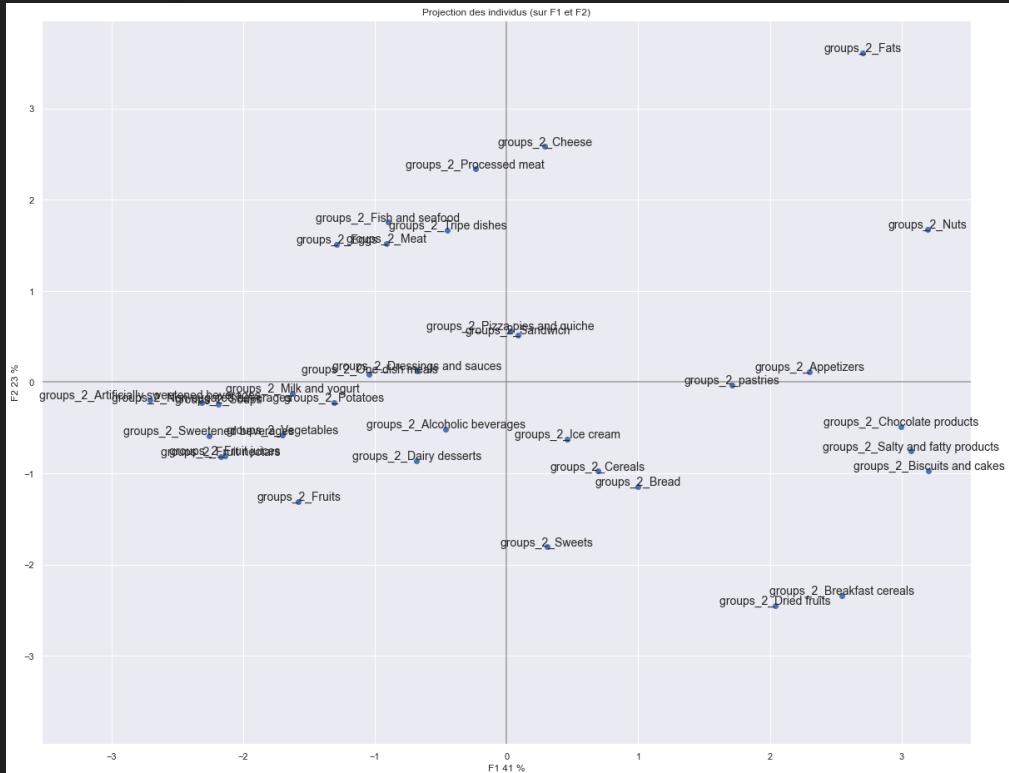
3-5-Réduction des dimensions (méthode ACP)

○ Projections:



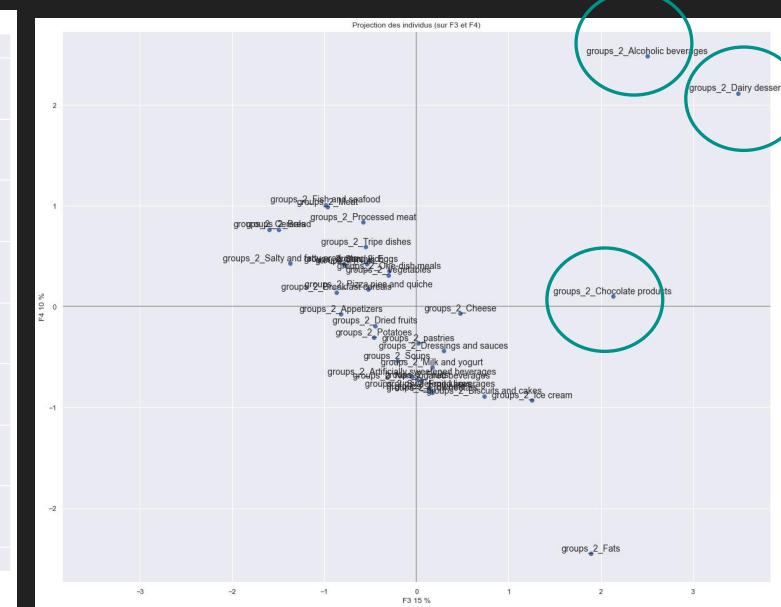
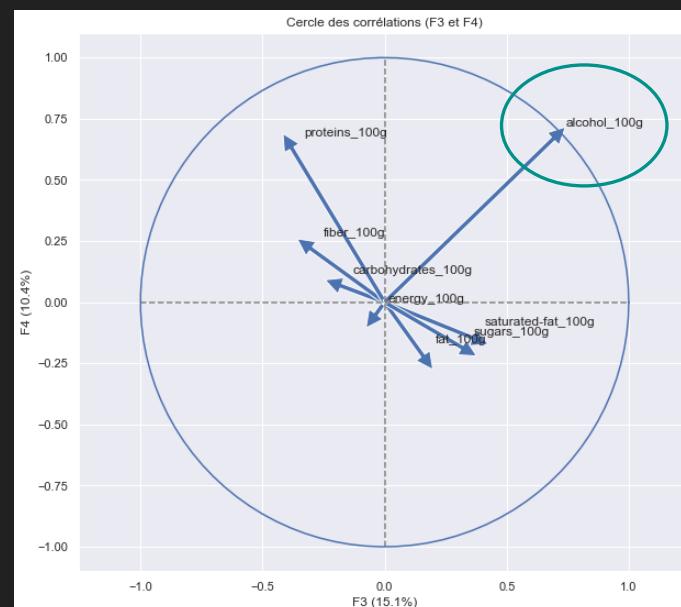
3-5-Réduction des dimensions (méthode ACP)

○ Projections:



3-5-Réduction des dimensions (méthode ACP)

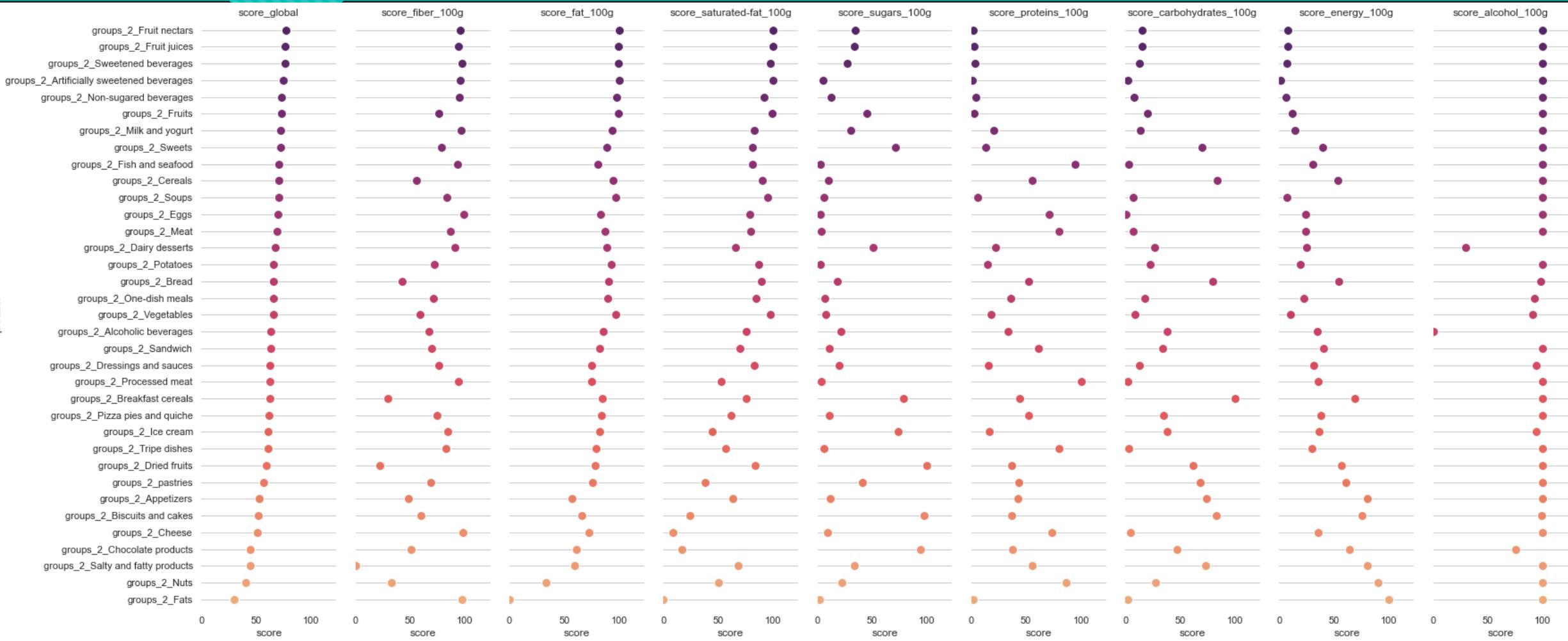
- Observation : les individus se trouvant dans le cadran nord-est du plan F2-F3 sont en partie expliqués par leur taux d'alcool (boissons alcoolisées mais aussi desserts et chocolats)



4-1-Scoring

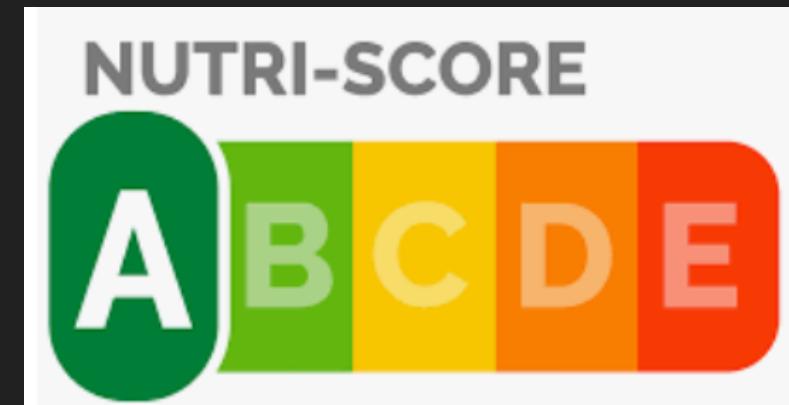
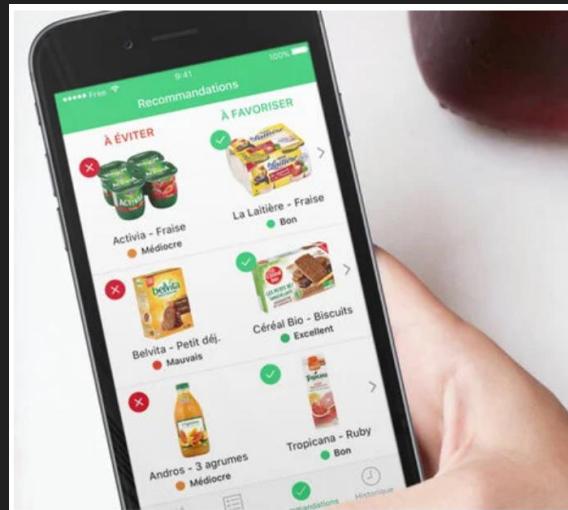
- Dans le but de créer une application qui permette de suivre un régime sans résidu on a créé un score par aliment.
- Les aliments sont évalués sur une échelle de 0 à 100 avec 0 pour l'aliment le moins bon pour le régime et 100 pour l'aliment le meilleur pour le régime.
- Ce score se base sur la teneur en %g des éléments suivants : fat_100g, saturated-fat_100g, fibers_100g, proteins_100g, carbohydrates_100g, sugars_100g, energy_100g.
- Un poids est donné à chaque élément suivant son importance dans le régime sans résidu :
 - 25% score_fat_100g
 - 25% score_saturated-fat_100g
 - 25% score_fibers_100g
 - 5% score_alcohols_100g
 - 5% score_proteins_100g
 - 5% score_carbohydrates_100g
 - 5% score_sugars_100g,
 - 5% score_energy_100g

4-1-Scoring



4-2-Conclusion

- Pour cette application de nutriscore pour le régime sans résidus, la base de données fournit les informations concernant les fibres, les matières grasses et les autres nutriments essentiels nécessaires comme les protéines, les glucides et l'énergie.
- En utilisant l'information que l'énergie est une combinaison linéaire des autres variables on a remplacé de façon fiable de nombreuses valeurs quantitatives manquantes par régression multilinéaire.
- Pour classifier les produits la colonne pnns_groups_2 fournit une bonne base qu'on a pu ensuite étoffer en effectuant un KNN imput.
- Au final on dresse le profil de 31 catégories d'aliments vis-à-vis du régime sans résidu. On constate que les jus de fruits sont particulièrement adapté avec peu de fibre et de matière grasse. En bas de tableau les chocolats, fromages et les noix, aliments ou très gras ou très riches en fibres, sont les moins recommandés.
- L'Algorithme utilisé pourra servir à créer un application de nutriscore.



4-2-Conclusion

A	
B	
C	
D	
E	