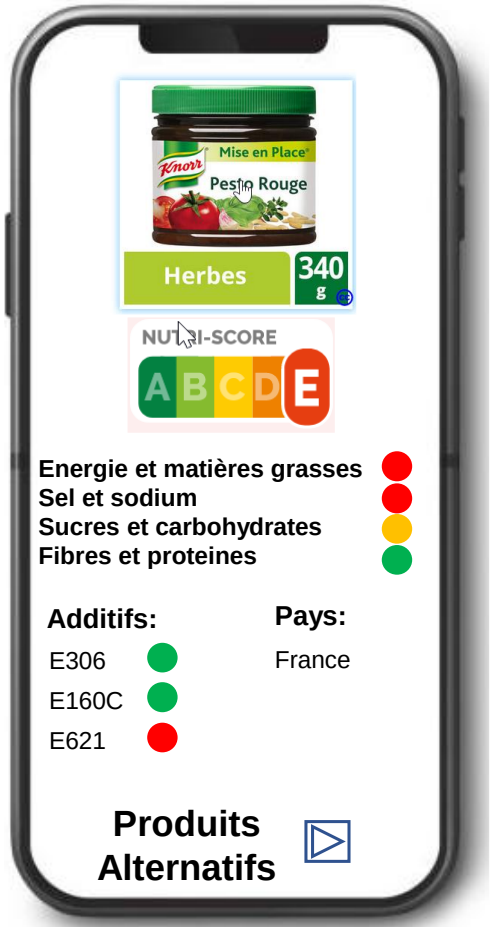
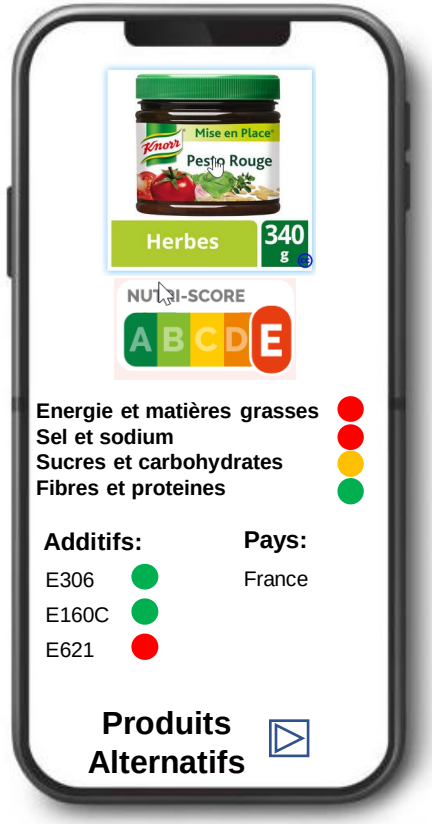
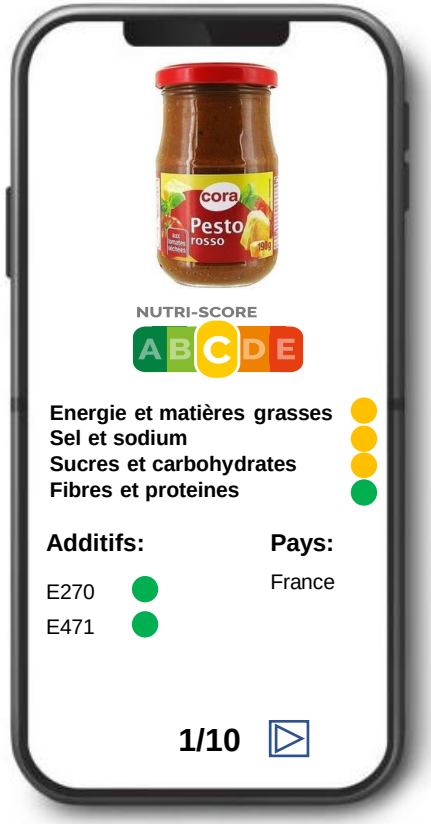


Projet P2 – Application au service de la santé publique

Description de l’application et schéma de fonctionnement

L’application permet de signaler aux consommateurs l’évaluation Nutriscore d’un produit alimentaire sélectionné et l’éventuelle présence et typologie d’additifs et conseiller des produits alternatifs sans additifs interdits, avec nombre total réduit d’additifs et meilleures valeurs nutritionnelles ordonnées par additifs et Nutrigrade. Si l’accès au database du supermarché est disponible l’application pourrait filtrer les propositions selon les produits disponibles dans les rayons.





ETAPE 1
Identification produit via barcode

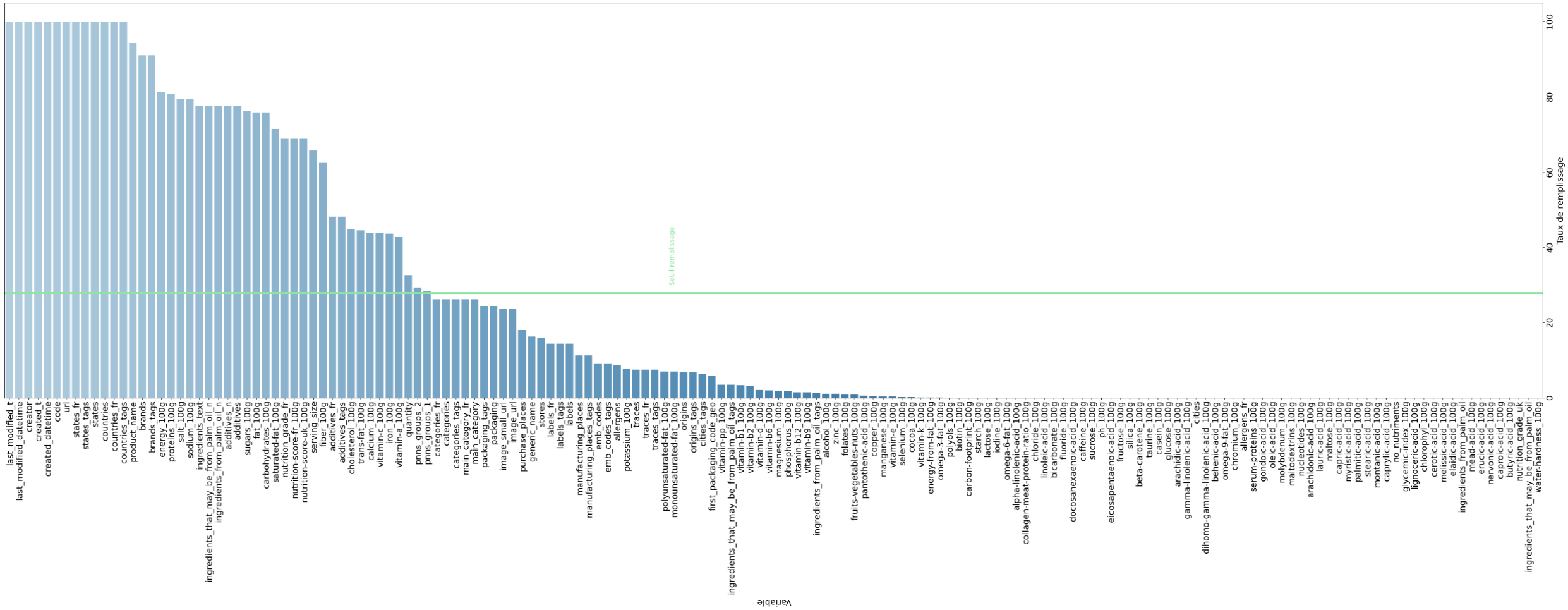
ETAPE 2
Vérification additifs et valeurs nutritionnelles

ETAPE 3
Proposition produits alternatifs

Description jeu de données initial

Open Food Facts	
Nr. individus	Nr. variables
320772	162

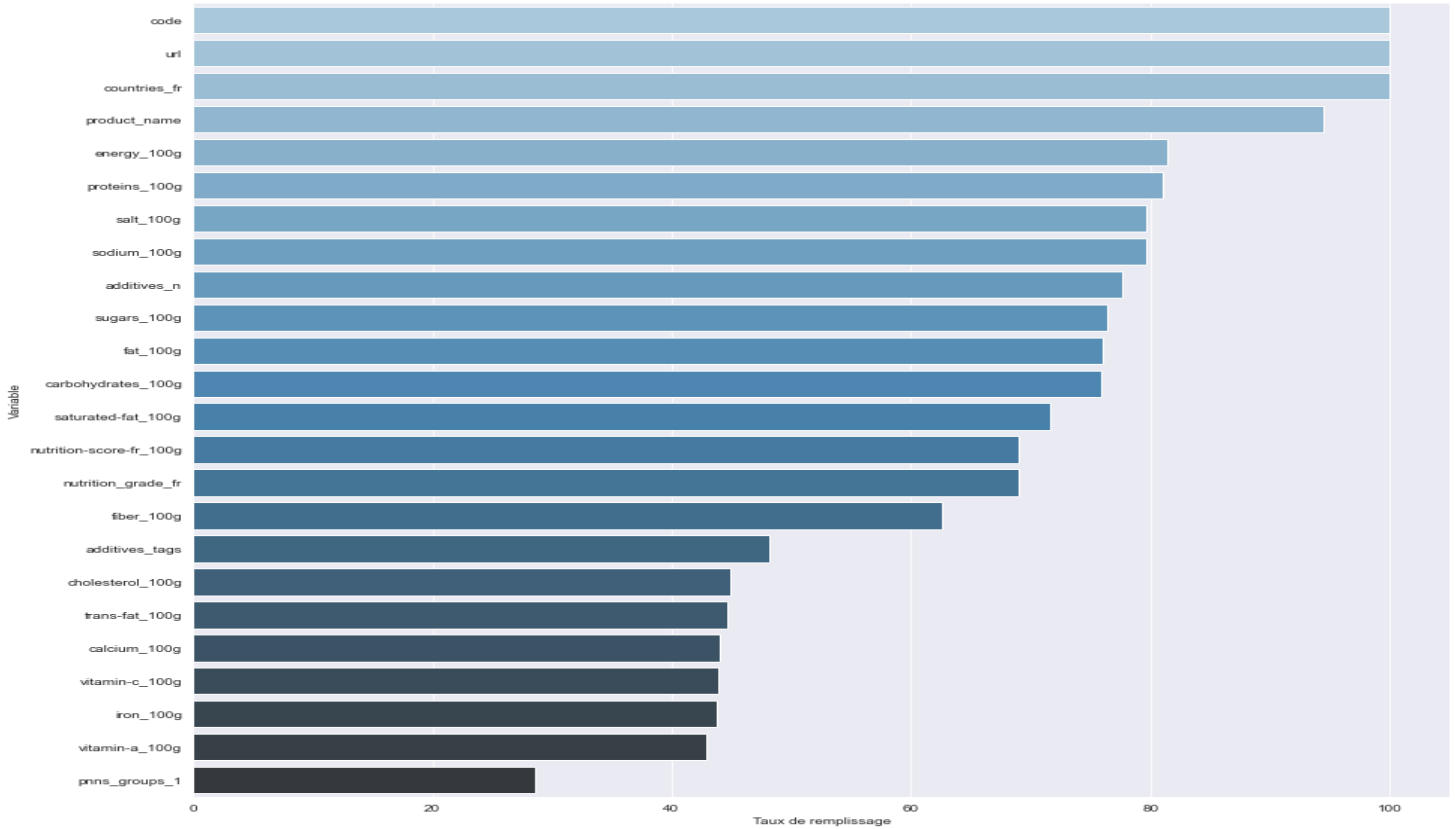
Taux de remplissage avant nettoyage (Seuil remplissage = 28%)



Nettoyage jeu de données – Réduction

Opération	Step	Critère	Nr. individus Apres Step	Nr. Variables Après Step
Suppression colonnes	1	Taux de remplissage	320772	45
	2	Utilité pour l'application	320772	32
	3	Redondance	320772	24
Suppression lignes	4	Doublons	320749	24
	5	Aucune valeur nutritionnelle	262292	24

Variables après réduction




```
data.dtypes
```

```
code                      object
url                       object
product_name              object
countries_fr              object
additives_n              float64
additives_tags            object
nutrition_grade_fr        object
pnns_groups_1             object
energy_100g              float64
fat_100g                 float64
saturated-fat_100g       float64
trans-fat_100g           float64
cholesterol_100g         float64
carbohydrates_100g       float64
sugars_100g              float64
fiber_100g               float64
proteins_100g            float64
salt_100g                float64
sodium_100g              float64
vitamin-a_100g           float64
vitamin-c_100g           float64
calcium_100g             float64
iron_100g                float64
nutrition-score-fr_100g  float64
```

Nettoyage jeu de données – Correction valeurs

Opération	Step	Critère	Stratégie	Correction	
Correction valeurs quantitatives	1	Valeurs énergétiques in [0,3700] kcal/100g	Clip variable energy_100g	Nr. correction: 631	
	2	Valeurs Nutritionnelles in [0,100] g	Clip variables nutritionnelles	Nr. correction: 271	
Correction valeurs qualitatives	3	Correction/traduction nom pays	Librairie FuzzyWuzzy	Nr. initial pays: 1614	Nr. final pays: 98
	4	Standardisation noms groupes nutritionnels	Librairie FuzzyWuzzy	Nr. corrections: 1185	
	5	Standardisation nom additif	Split et Replace		

Correction/traduction nom pays

```
 533 / 553
string_to_match: Portugal,en:Belgie,en:Duitsland,en:Frankrijk,en:Nederland,en:Spanje,en:Zwitserland
countries_to_match: ['Portugal', 'en:Belgie', 'en:Duitsland', 'en:Frankrijk', 'en:Nederland', 'en:Spanje', 'en:Zwitserland']
Portugal
Best match: Portugal Langue: da Max score: 100 Country: Portugal
en:Belgie
en:Duitsland
Best match: Duitsland Langue: nl Max score: 86 Country: Allemagne
en:Frankrijk
Best match: Frankrijk Langue: nl Max score: 86 Country: France
en:Nederland
Best match: Nederland Langue: nl Max score: 86 Country: Pays-Bas
en:Spanje
en:Zwitserland
Best match: Zwitserland Langue: nl Max score: 88 Country: Suisse
Version corrigée: Portugal,Allemagne,France,Pays-Bas,Suisse
```

```
 130 / 553
string_to_match: en:الإمارات-العربية-المتحدة
countries_to_match: ['en:الإمارات-العربية-المتحدة']
en:الإمارات-العربية-المتحدة
Best match: الإمارات العربية المتحدة Langue: ar Max score: 86 Country: Émirats arabes unis
Version corrigée: Émirats arabes unis
```

Standardisation noms groupes nutritionnels

```
**************************************************************************************************
Nombre valeurs uniques en  pnns_groups_1 variable:  13

Valeurs uniques en  pnns_groups_1 variable:
 ['Fruits and vegetables' 'Sugary snacks' 'Composite foods'
 'Fish Meat Eggs' 'Beverages' 'Fat and sauces' 'Cereals and potatoes'
 'Milk and dairy products' 'Salty snacks' 'fruits-and-vegetables'
 'sugary-snacks' 'cereals-and-potatoes' 'salty-snacks']
--------------------------------------------------------------------------------------------------
Fruits and vegetables
Step  1 / 13  :  fruits-and-vegetables (counted  [765] times in pnns_groups_1 ) replaced with Fruits and vegetables (counted
[3980] times in  pnns_groups_1  )
Score de similarité: 86 %
--------------------------------------------------------------------------------------------------
Sugary snacks
Step  2 / 13  :  sugary-snacks (counted  [404] times in pnns_groups_1 ) replaced with Sugary snacks (counted  [10232] times in
pnns_groups_1  )
Score de similarité: 85 %
--------------------------------------------------------------------------------------------------
Cereals and potatoes
Step  7 / 13  :  cereals-and-potatoes (counted  [15] times in pnns_groups_1 ) replaced with Cereals and potatoes (counted  [743
4] times in  pnns_groups_1  )
Score de similarité: 85 %
--------------------------------------------------------------------------------------------------
Salty snacks
Step  9 / 13  :  salty-snacks (counted  [1] times in pnns_groups_1 ) replaced with Salty snacks (counted  [2468] times in  pnns
_groups_1  )
Score de similarité: 83 %

Nombre totale valeurs corrigées dans pnns_groups_1: pnns_groups_1 :  [1185]
```

Nettoyage jeu de données – Remplissage valeurs manquantes

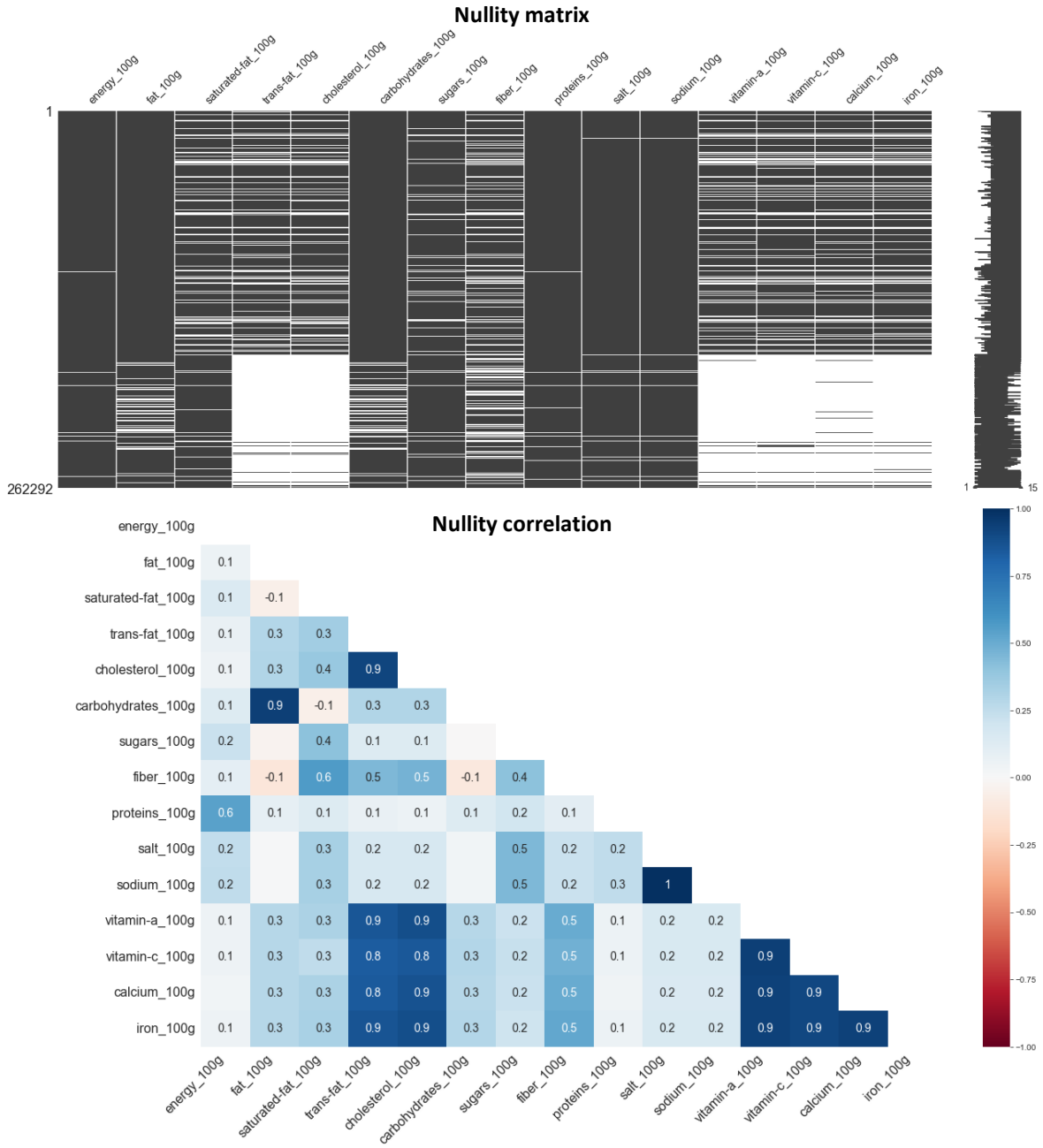
Techniques remplissage:

Variables nutritionnelles MNAR:
- Médiane des valeurs de la variables pour le même groupe nutritionnel 'pnns_groupe_1')
- Médiane des valeurs de la variables si 'pnns_groupe_1' = NaN

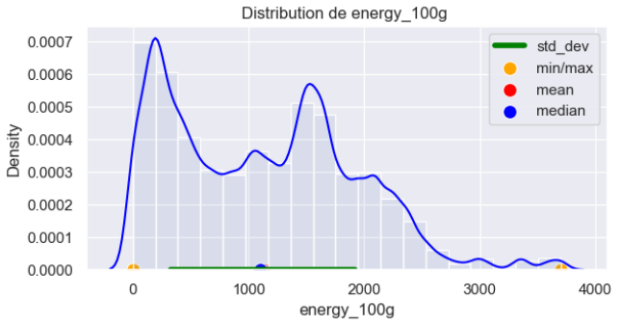
Variables nutritionneles MAR:
- Iterative imputation avec variables corrélées

Variable qualitative nutrition_grade-fr:
- KNN Imputer

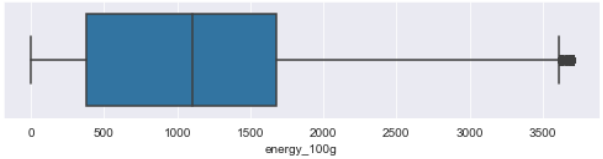
Variable	Nr. Valeurs manquantes	Type Valeurs manquantes	Technique remplissage
energy_100g	1513	MAR	Iterative Imputation
fat_100g	18401	MAR	Iterative Imputation
saturated_fat_100g	32738	MAR	Iterative Imputation
trans-fat_100g	118994	MNAR	Médiane
cholesterol_100g	118202	MNAR	Médiane
carbohydrates_100g	18704	MAR	Iterative Imputation
sugars_100g	17321	MAR	Iterative Imputation
fiber_100g	61406	MAR	Iterative Imputation
proteins_100g	2386	MAR	Iterative Imputation
salt_100g	6782	MAR	Iterative Imputation
sodium_100g	6829	MAR	Iterative Imputation
vitamin-a_100g	124738	MNAR	Médiane
vitamin-c_100g	121425	MNAR	Médiane
calcium_100g	121242	MNAR	Médiane
iron_100g	121830	MNAR	Médiane
nutrition_grade_fr	41273	-	KNN Imputer



Variable energy_100g

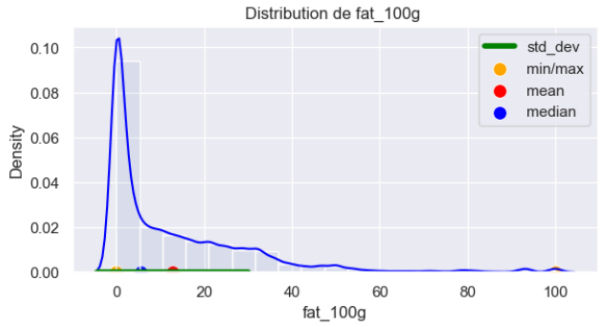


	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	3700.00	3700.00	1126.00	1100.00	[329.5, 1922.5]	0.46	-0.37

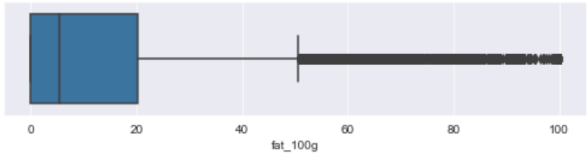


	whis_low	quant25	median	quant75	whis_high	IQR
0	-1558.5	381.0	1100.0	1674.0	3613.5	1293.0

Variable fat_100g

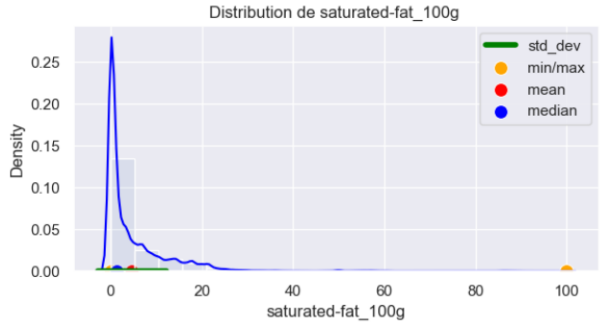


	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	100.00	100.00	12.80	5.45	[-4.5, 30.1]	2.17	6.11

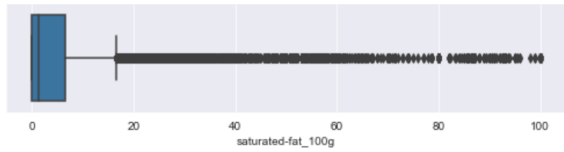


	whis_low	quant25	median	quant75	whis_high	IQR
0	-30.43	0.0	5.45	20.28	50.71	20.28

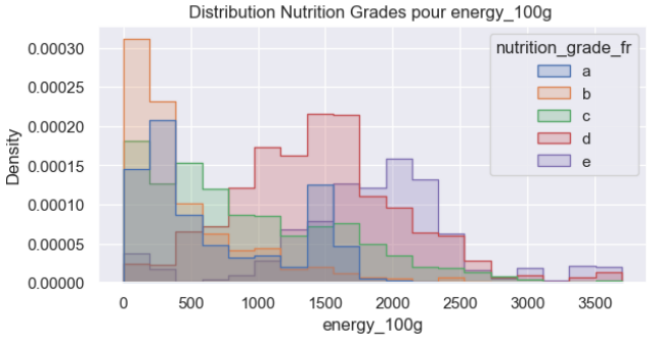
Variable saturated-fat_100g

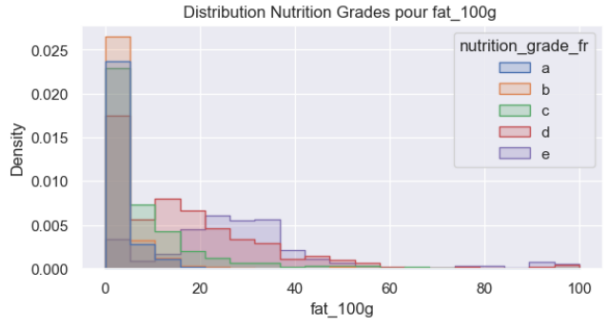


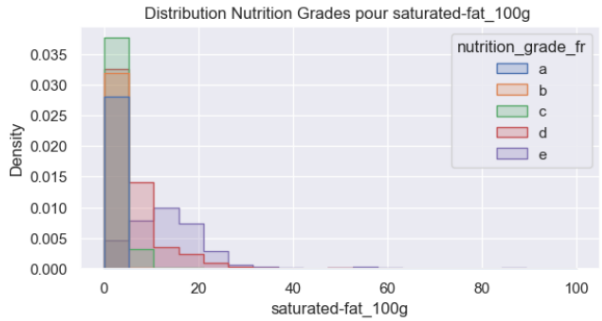
	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	100.00	100.00	4.70	1.25	[-2.9, 12.3]	3.66	24.35



	whis_low	quant25	median	quant75	whis_high	IQR
0	-10.0	0.0	1.25	6.67	16.67	6.67



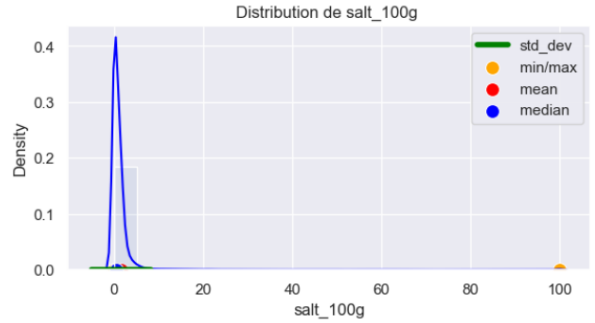




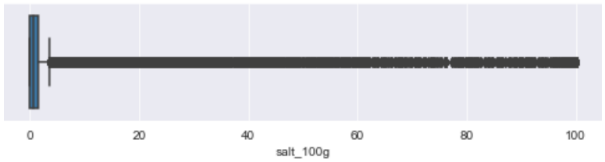
Exploration jeu de données – Analyse univariée variables quantitatives

Sel et sucres

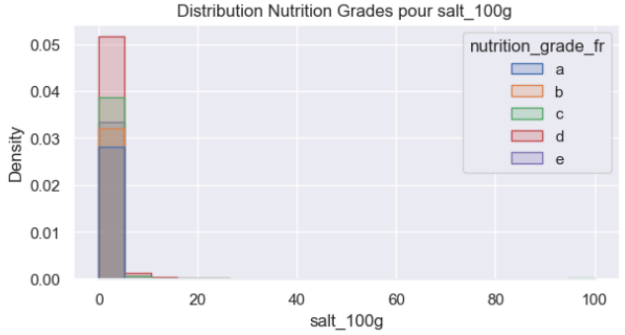
Variable salt_100g



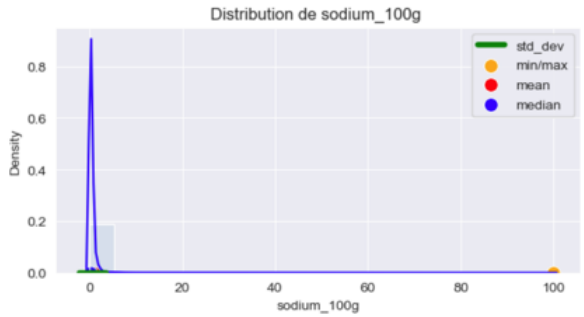
	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	100.00	100.00	1.60	0.61	[-5.0, 8.2]	10.99	137.19



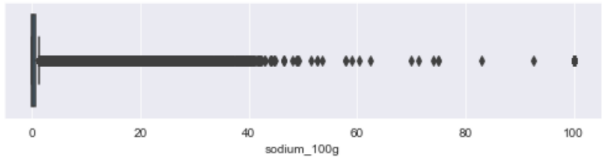
	whis_low	quant25	median	quant75	whis_high	IQR
0	-2.04	0.07	0.61	1.48	3.58	1.41



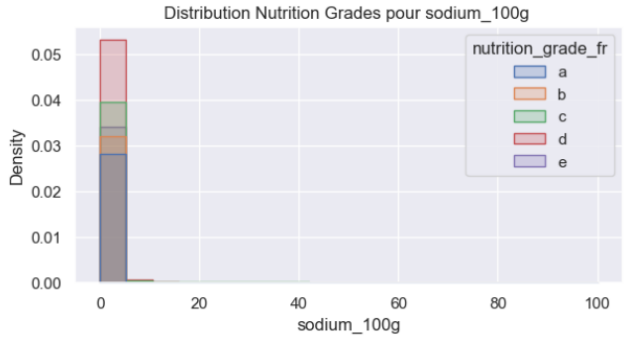
Variable sodium_100g



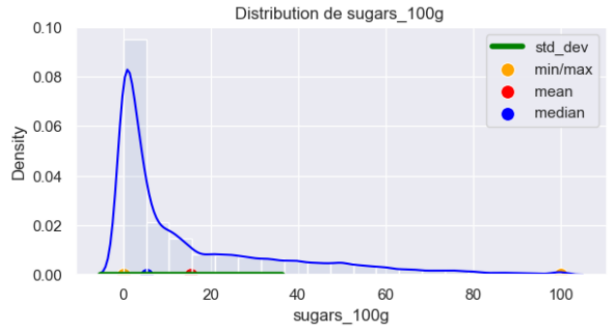
	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	100.00	100.00	0.70	0.24	[-2.2, 3.6]	14.47	306.68



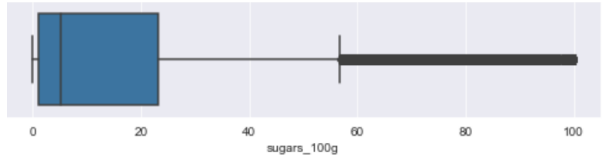
	whis_low	quant25	median	quant75	whis_high	IQR
0	-0.8	0.03	0.24	0.58	1.41	0.55



Variable sugars_100g



	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	100.00	100.00	15.50	5.26	[-5.3, 36.3]	1.71	2.42

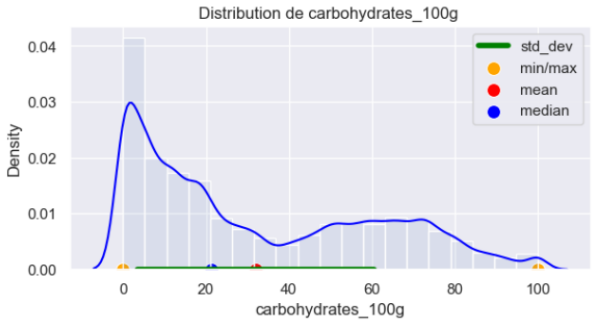


	whis_low	quant25	median	quant75	whis_high	IQR
0	-32.49	1.0	5.26	23.33	56.82	22.33

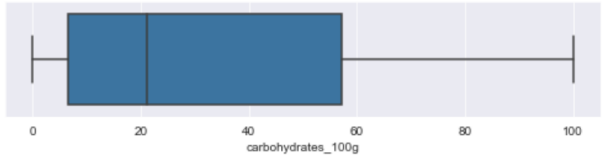
Exploration jeu de données – Analyse univariée variables quantitatives

Carbohydrates, fibres et proteines

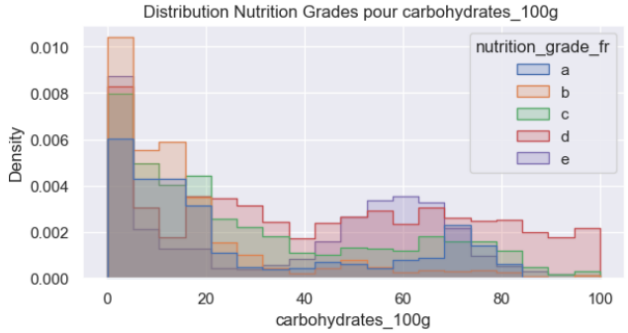
Variable carbohydrates_100g



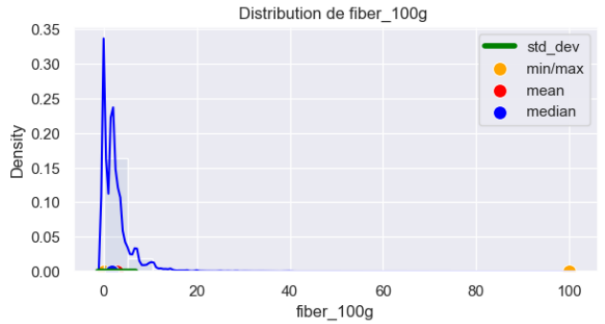
	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	100.00	100.00	32.00	21.21	[3.5, 60.5]	0.60	-0.97



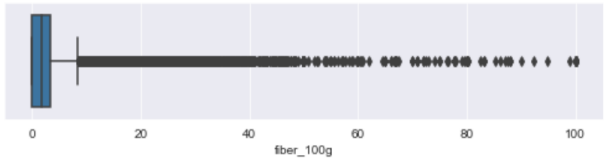
	whis_low	quant25	median	quant75	whis_high	IQR
0	-69.03	6.67	21.21	57.14	132.84	50.47



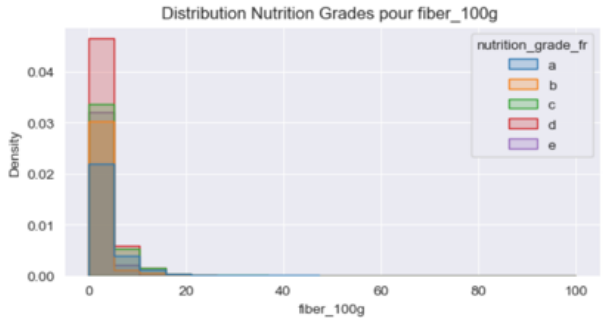
Variable fiber_100g



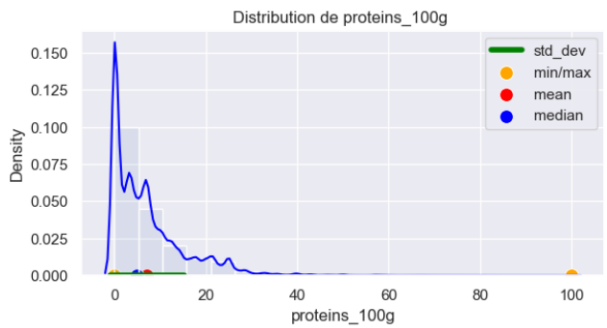
	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	100.00	100.00	2.80	1.79	[-1.3, 6.9]	5.80	68.31



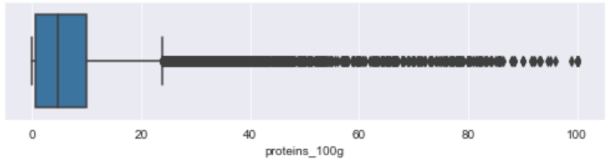
	whis_low	quant25	median	quant75	whis_high	IQR
0	-5.13	0.0	1.79	3.42	8.56	3.42



Variable proteins_100g



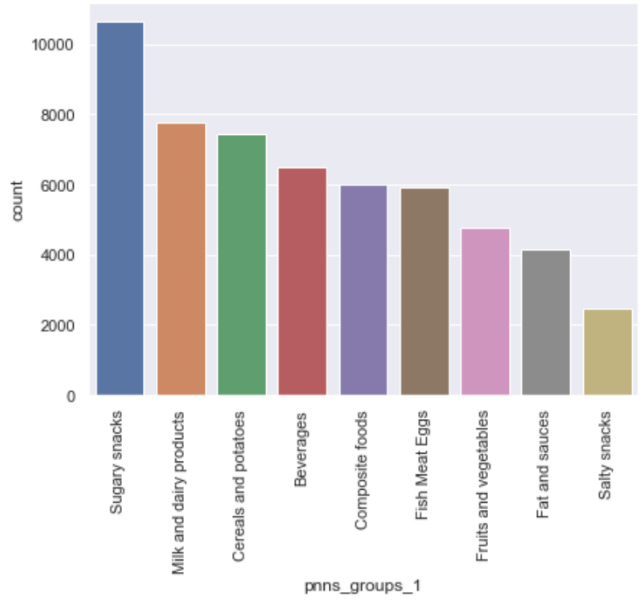
	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	262292	0	0.00	100.00	100.00	7.10	4.84	[-1.0, 15.2]	2.14	8.54

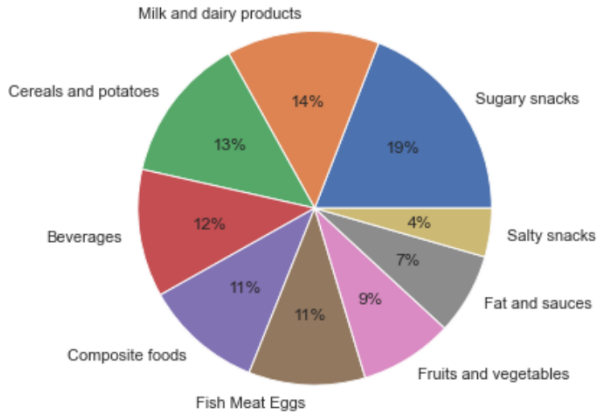


	whis_low	quant25	median	quant75	whis_high	IQR
0	-13.25	0.7	4.84	10.0	23.95	9.3

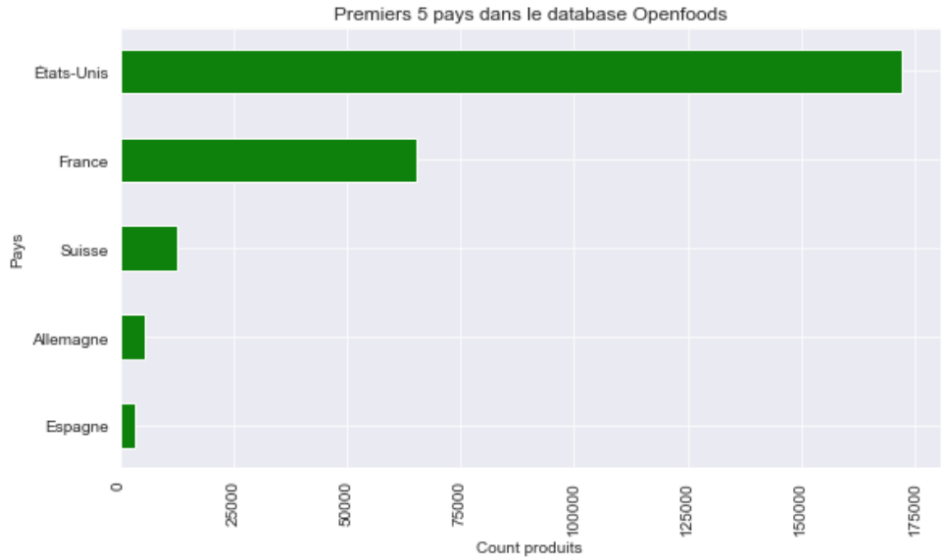
Exploration jeu de données – Analyse univariée variables qualitatives

GROUPES DE PRODUIT

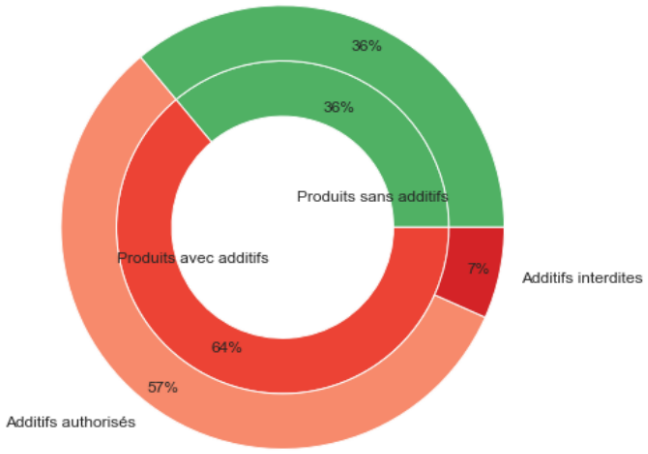


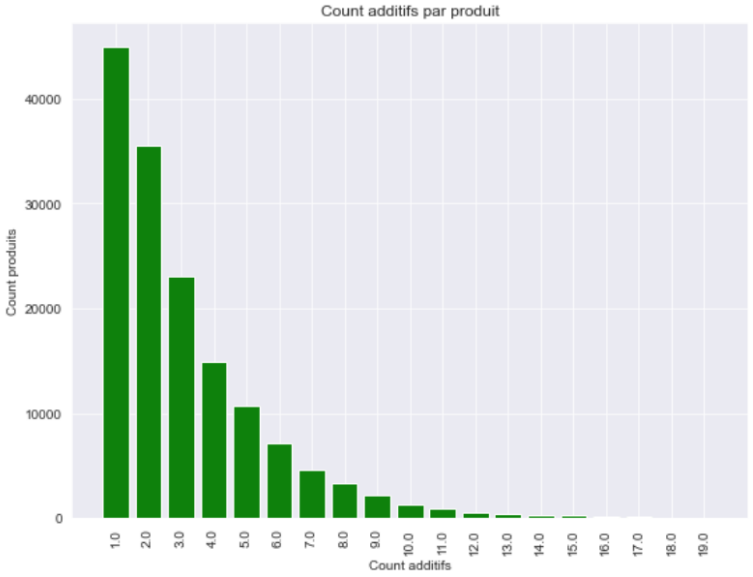


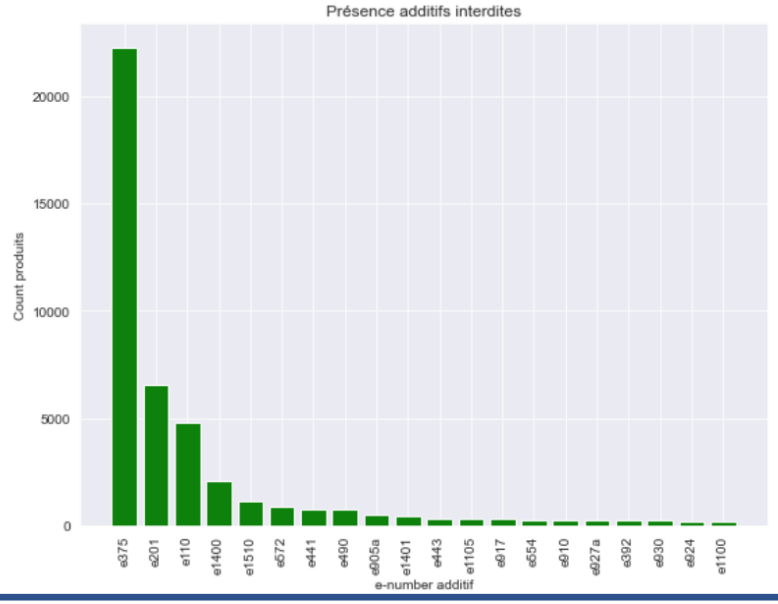
PAYS



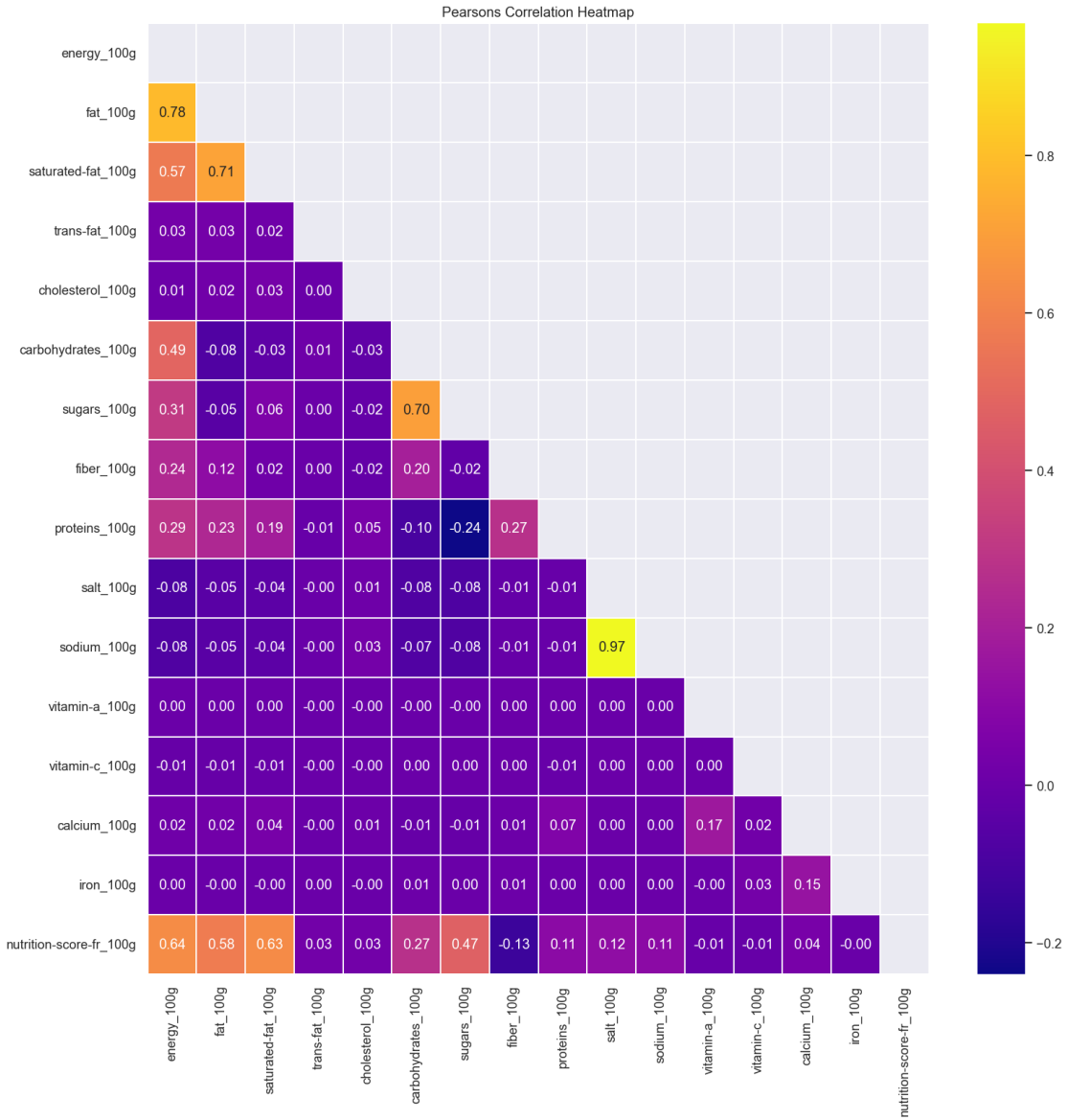
ADDITIFS

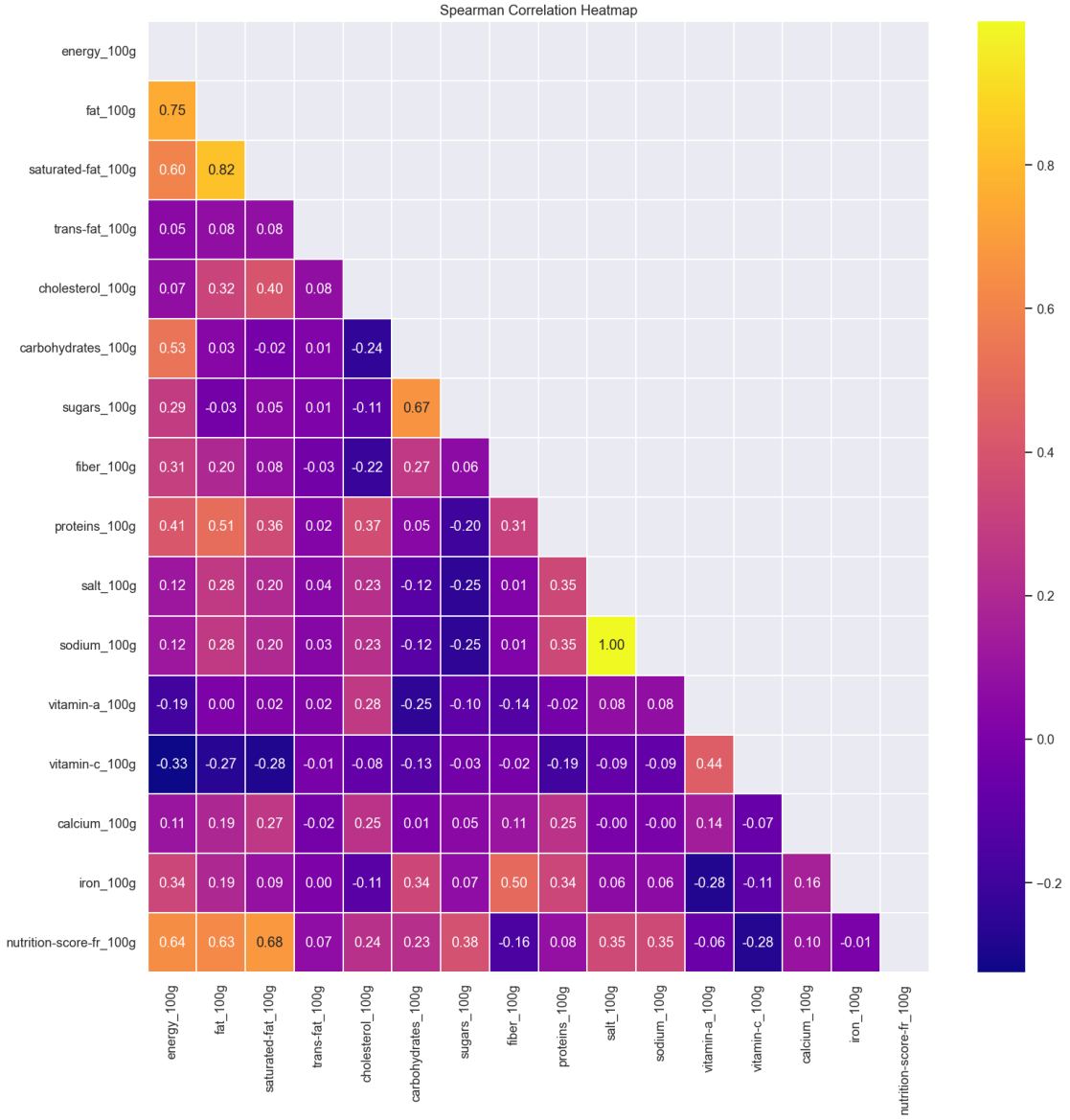






Exploration jeu de données – Analyse multivariée variables quantitatives

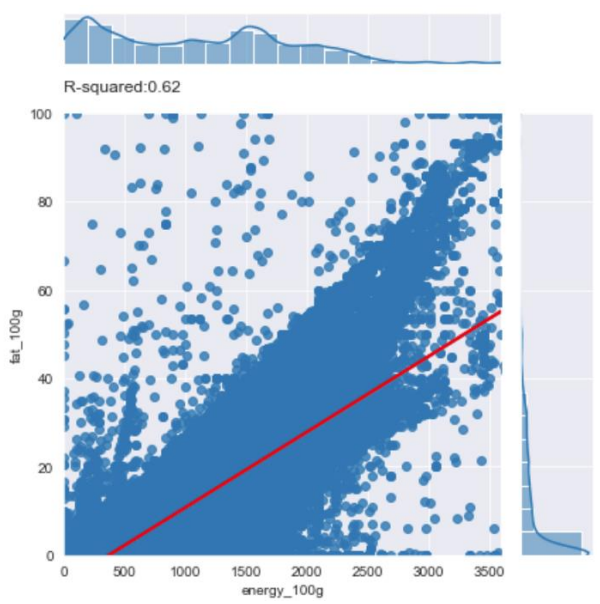


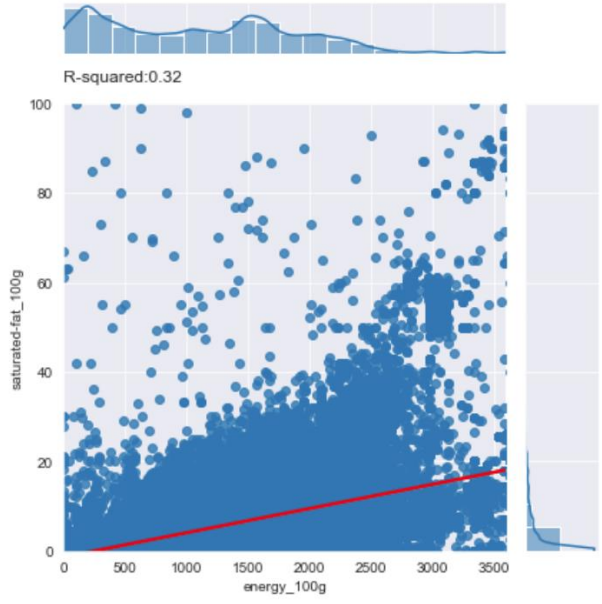


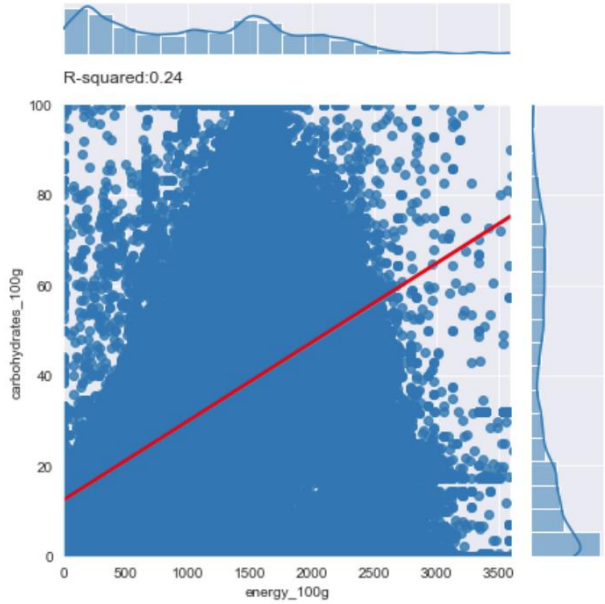
L'analyse des corrélations de Pearsons/Spearman confirme la présence de corrélation linéaires positives et liaison monotones entre 'energy_100g', 'fat_100g', 'saturated-fat_100g' et 'carbohydrates_100g', 'sodium_100g' et 'salt_100g', entre 'fat_100g' et 'saturated-fat_100g', entre 'carbohydrates_100g' et 'sugars_100g', entre 'nutrition-score-fr_100g', 'energy_100g', 'fat_100g', 'saturated-fat_100g' et 'sugars_100g'

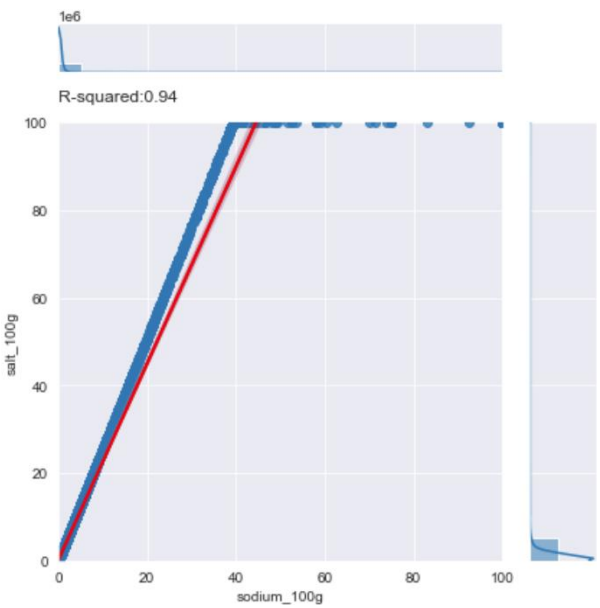
Exploration jeu de données – Analyse multivariée variables quantitatives

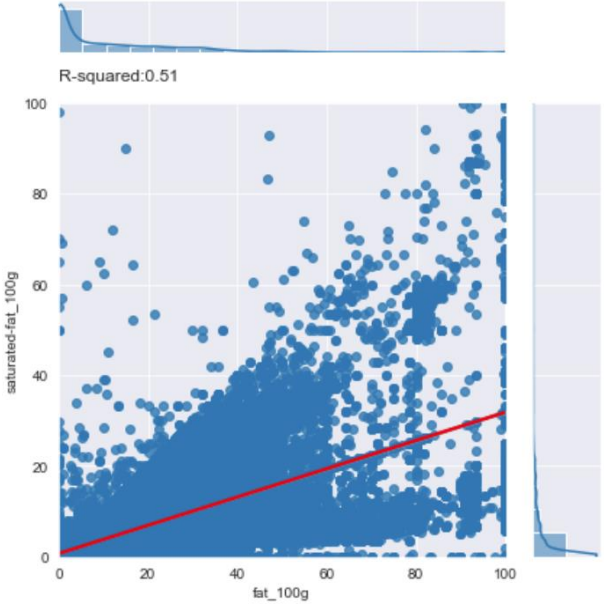
RELATIONS PAR PAIRES DE VARIABLES ET REGRESSIONS LINEAIRES

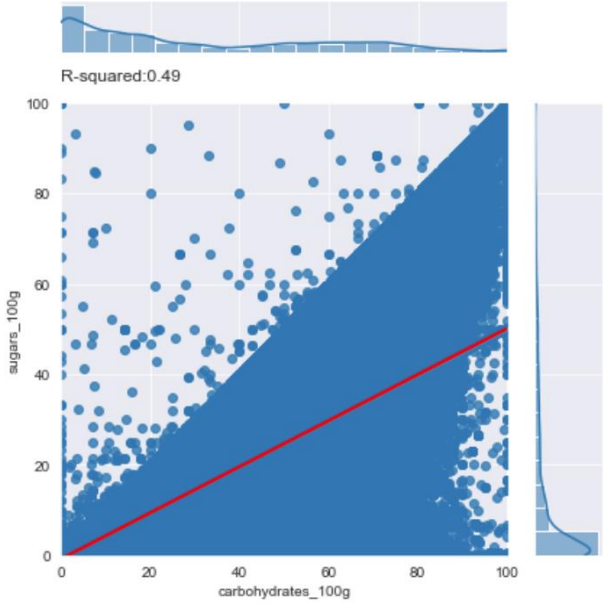






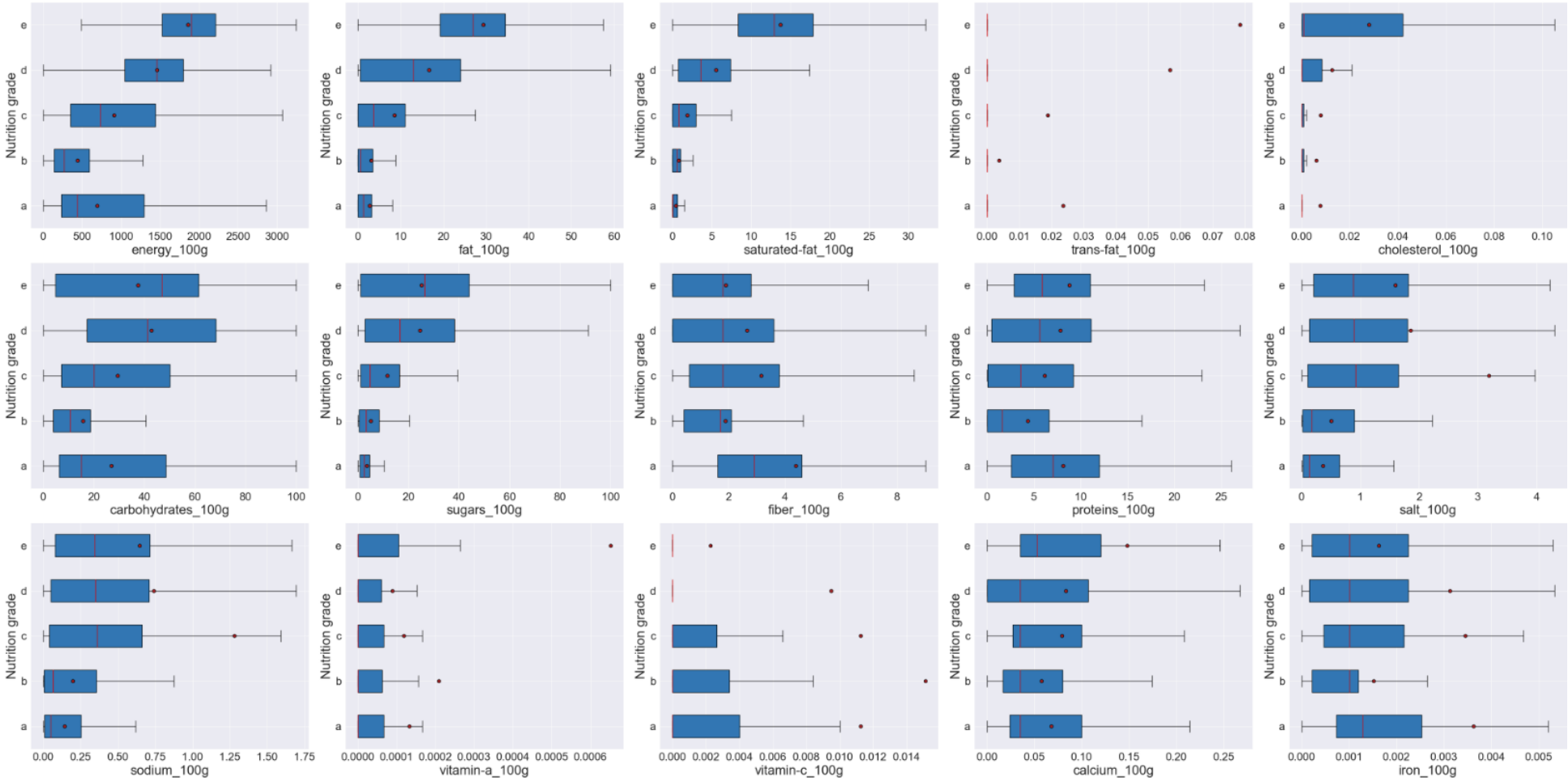






Exploration jeu de données – Analyse multivariée variables quantitative/qualitative

Objectif: vérifier si il y a une corrélation entre variables nutritionnelles et nutrigrade

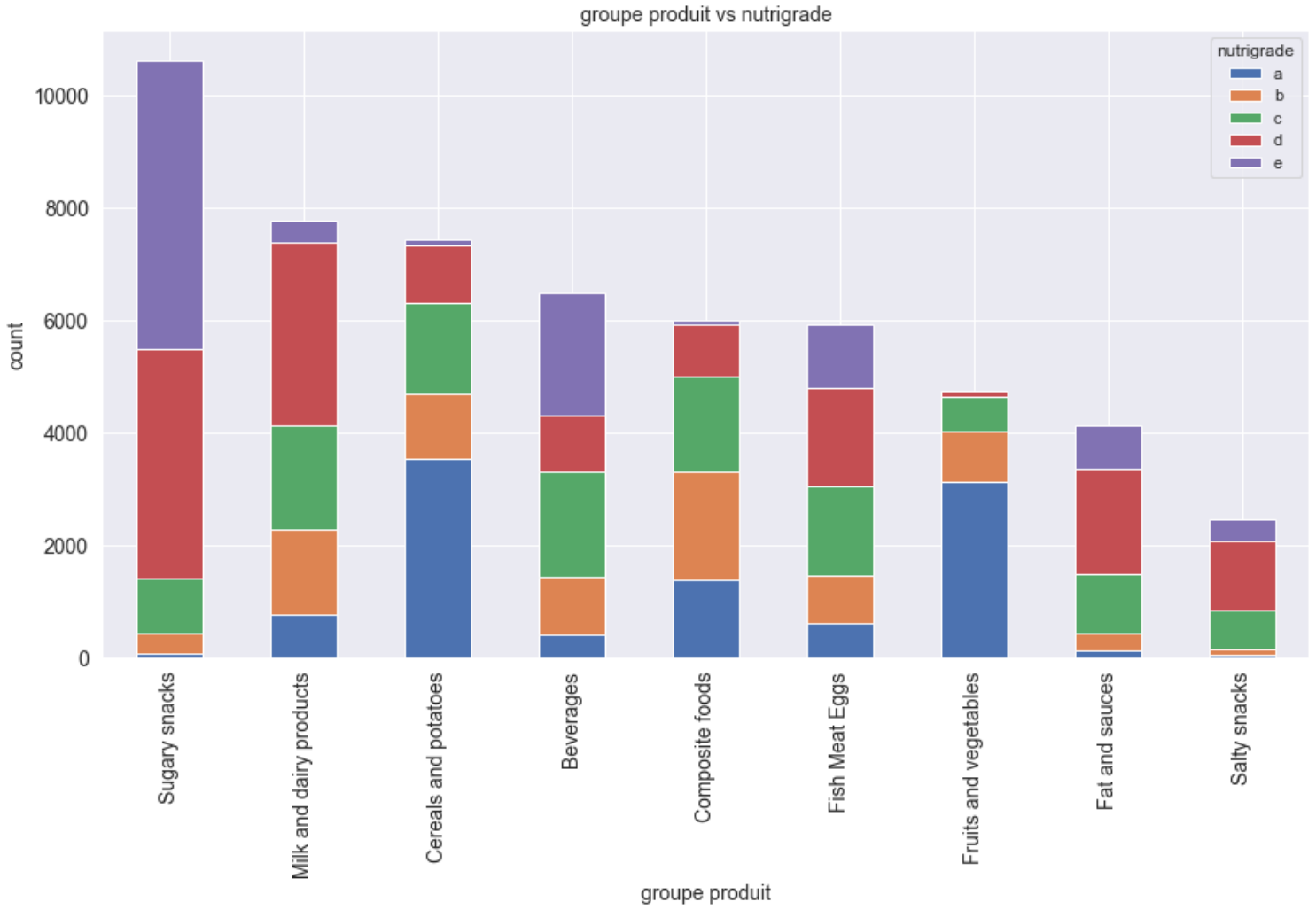


Test Kruskal Wallis

```
----------------------------
energy_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  105152.957016    0.0

----------------------------
fat_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  82485.660618    0.0

----------------------------
saturated-fat_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  107509.914911    0.0

----------------------------
trans-fat_100g
                  Source  ddof1              H          p-unc
Kruskal  nutrition_grade_fr      4  1208.671441  2.100525e-260

----------------------------
cholesterol_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  15192.683124    0.0

----------------------------
carbohydrates_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  24810.009237    0.0

----------------------------
sugars_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  35255.693175    0.0

----------------------------
fiber_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  14624.324609    0.0

----------------------------
proteins_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  13874.9745    0.0

----------------------------
salt_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  26804.055566    0.0

----------------------------
sodium_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  26784.232927    0.0

----------------------------
vitamin-a_100g
                  Source  ddof1              H          p-unc
Kruskal  nutrition_grade_fr      4  1002.845061  8.630402e-216

----------------------------
vitamin-c_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  15784.44658    0.0

----------------------------
calcium_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  3438.063931    0.0

----------------------------
iron_100g
                  Source  ddof1              H  p-unc
Kruskal  nutrition_grade_fr      4  4953.736692    0.0
```

Le test de Kruskal Wallis confirme une corrélation entre les variables nutritionnelles et le nutrition grade.

Objectif: vérifier si il y a une corrélation entre groupe de produit et nutrigrade



Test d'indépendnce du Chi2

```
(observed-expected).round(2)
```

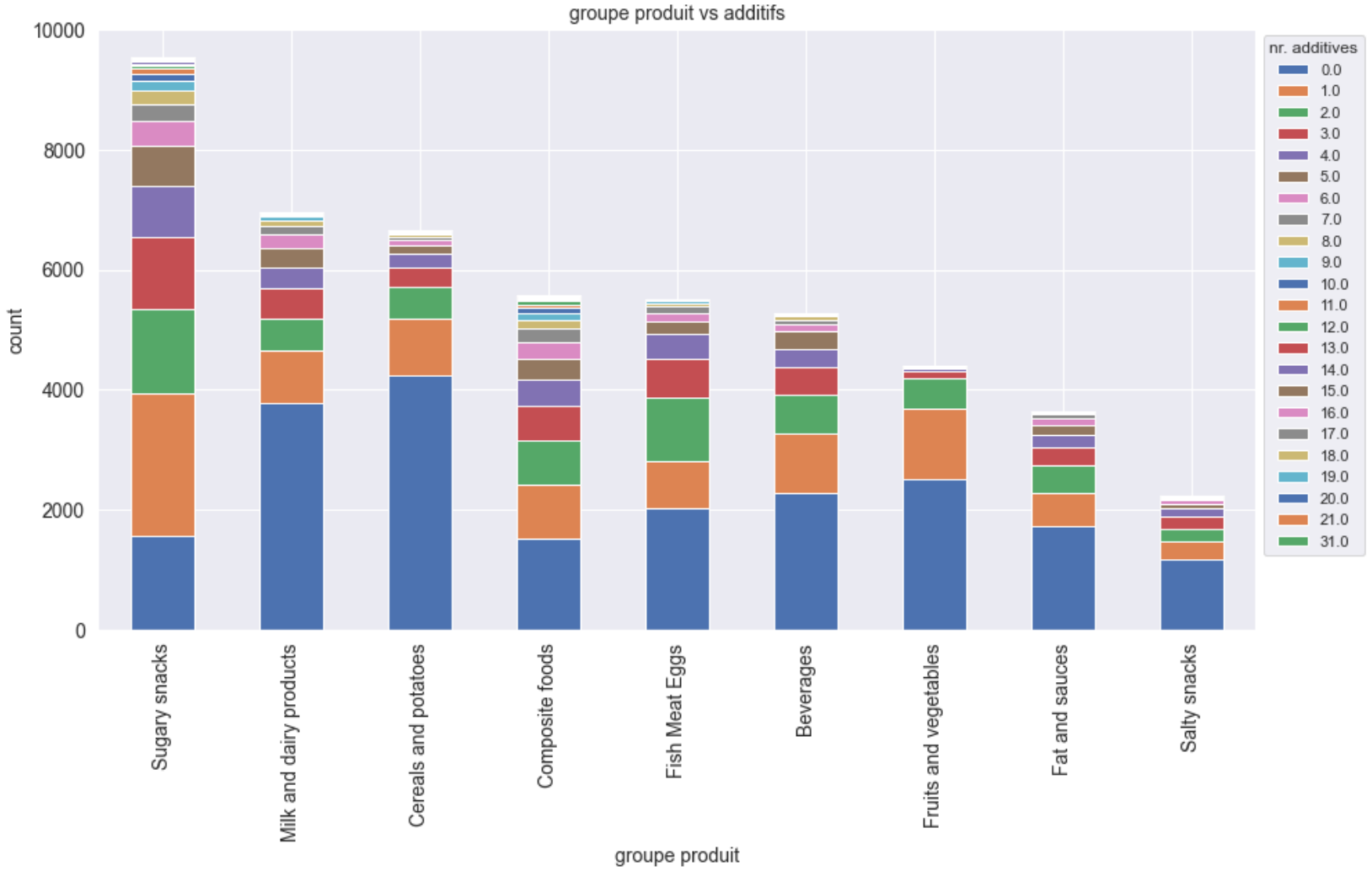
nutrition_grade_fr pnns_groups_1	a	b	c	d	e
Beverages	-775.35	92.69	472.83	-784.58	994.41
Cereals and potatoes	2196.30	51.50	18.71	-1019.48	-1247.04
Composite foods	301.69	1035.57	411.82	-728.58	-1020.50
Fat and sauces	-624.70	-306.14	171.06	738.52	21.25
Fish Meat Eggs	-467.18	-12.47	312.92	121.90	44.83
Fruits and vegetables	2266.61	207.18	-404.11	-1206.88	-862.80
Milk and dairy products	-655.09	405.62	160.62	1137.97	-1049.13
Salty snacks	-389.69	-275.46	166.24	564.02	-65.11
Sugary snacks	-1852.59	-1198.49	-1310.11	1177.10	3184.09

```
stats[stats['test']=='pearson'].round(2)
```

	test	lambda	chi2	dof	pval	cramer	power
0	pearson	1.0	31476.24	32.0	0.0	0.17	1.0

Le test d'indépendance du Chi2 confirme une corrélation entre groupe de produit et nutrigrade

Objectif: vérifier si il y a une corrélation entre groupe de produit et nombre d'additifs



Test d'indépendance du Chi2

```
(observed-expected).round(2)
```

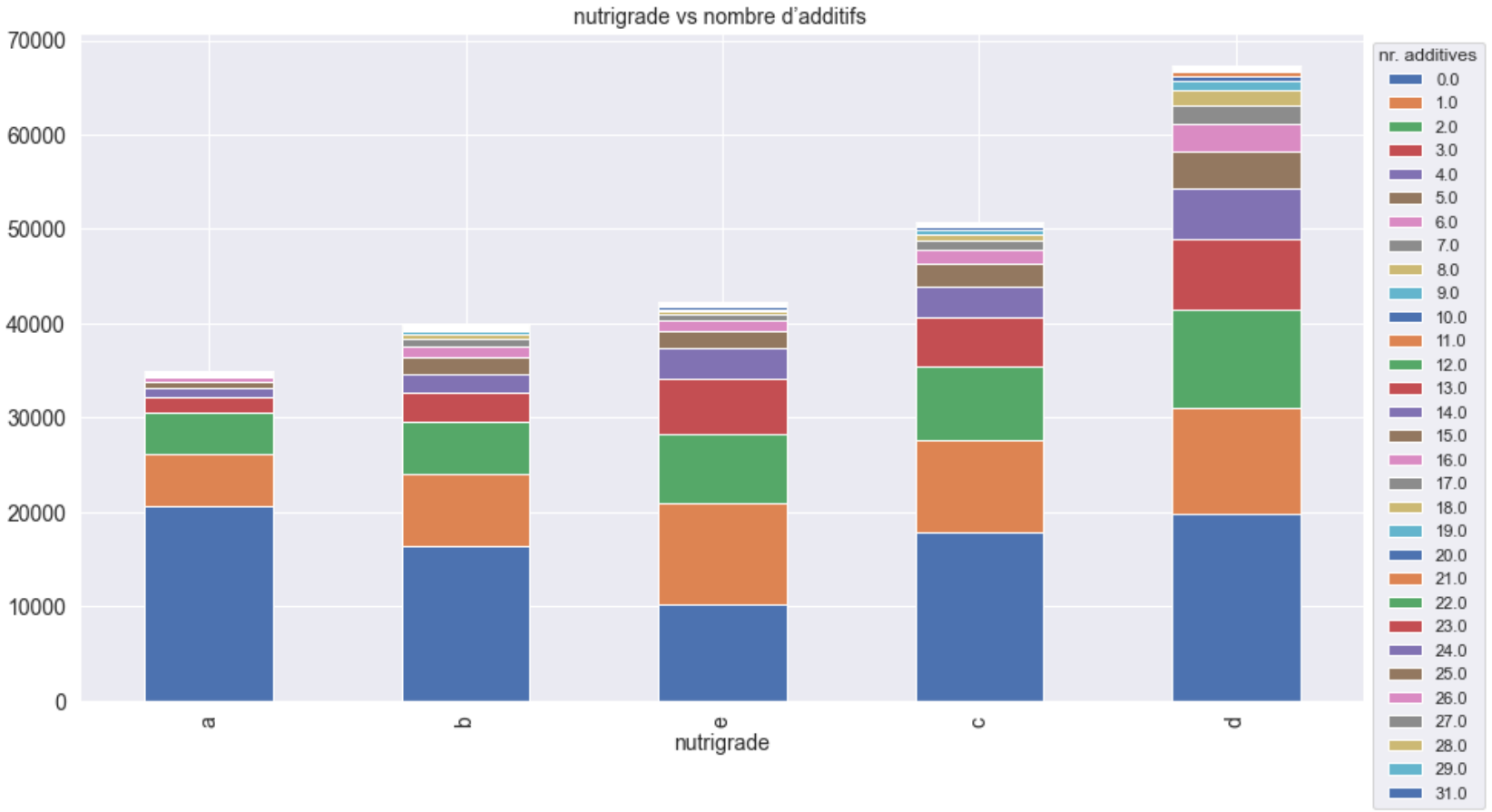
additives_n pnns_groups_1	0.0	1.0	2.0	3.0	4.0	5.0
Beverages	-1.24	9.57	-27.81	-0.09	-26.52	46.09
Cereals and potatoes	1292.44	-304.20	-331.92	-294.74	-206.80	-154.78
Composite foods	-561.11	19.30	136.25	133.76	144.61	127.19
Fat and sauces	167.58	-128.85	13.49	-32.80	-7.25	-12.18
Fish Meat Eggs	-321.28	-230.50	374.69	144.76	73.38	-41.05
Fruits and vegetables	508.77	308.43	-72.93	-302.01	-246.61	-195.64
Milk and dairy products	851.12	-358.18	-333.42	-99.36	-63.10	2.94
Salty snacks	196.53	-107.85	-70.62	7.78	9.05	-34.90
Sugary snacks	-2132.82	792.29	312.26	442.69	323.24	262.32

```
stats[stats['test']=='pearson'].round(2)
```

	test	lambda	chi2	dof	pval	cramer	power
0	pearson	1.0	5686.41	40.0	0.0	0.07	1.0

Le test d'indépendance du Chi2 confirme une corrélation entre groupe de produit et nombre d'additifs

Objectif: vérifier si il y a une corrélation entre nutrition grade et nombre d'additifs



Test d'indépendance du Chi2

```
(observed-expected).round(2)
```

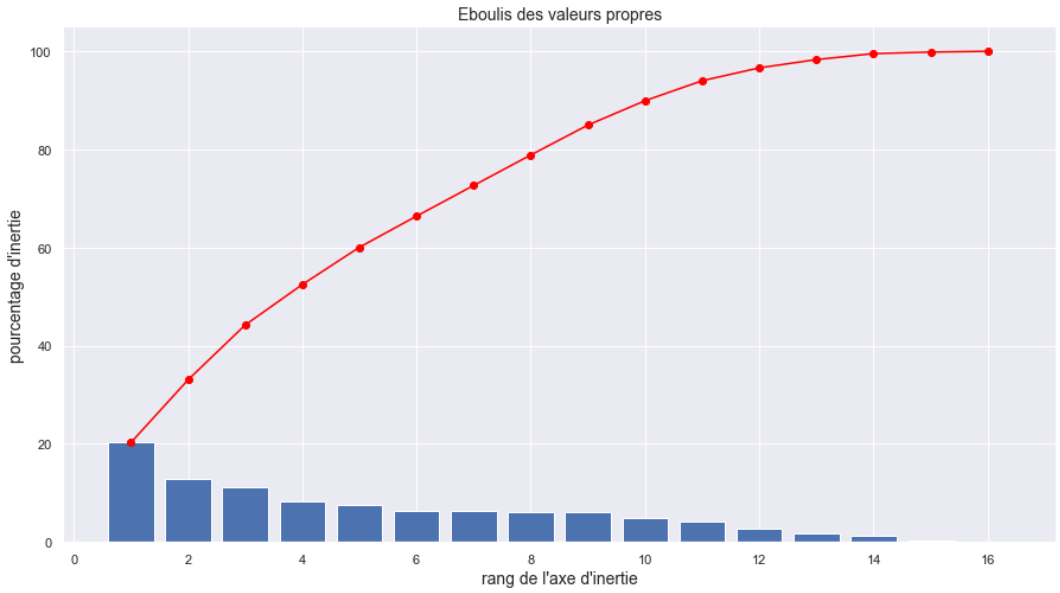
additives_n	0.0	1.0	2.0	3.0	4.0	5.0	6.0
nutrition_grade_fr							
a	7423.41	-1386.72	-1108.85	-1942.71	-1361.04	-941.91	-682.17
b	2025.42	-50.83	-486.30	-936.95	-462.27	-55.37	-33.70
c	-539.56	145.03	73.61	107.70	152.12	103.73	-42.64
d	-3618.94	-1240.92	571.64	1145.39	1182.89	972.41	987.53
e	-5290.33	2533.45	949.90	1626.57	488.31	-78.87	-229.02

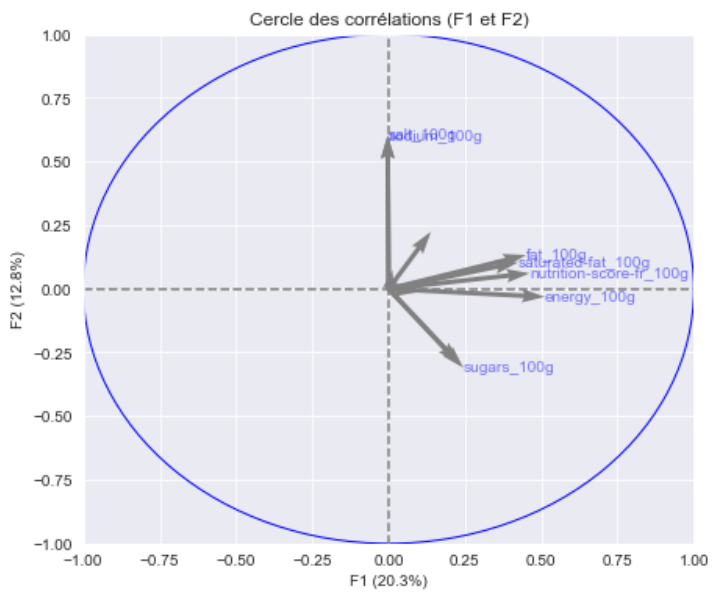
```
stats[stats['test']=='pearson'].round(2)
```

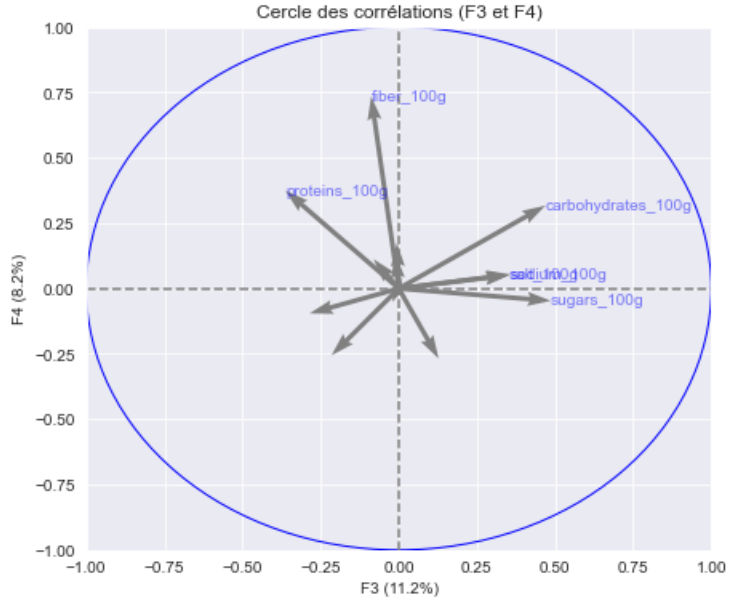
	test	lambda	chi2	dof	pval	cramer	power
0	pearson	1.0	13755.17	24.0	0.0	0.12	1.0

Le test d'indépendance du Chi2 confirme une corrélation entre nutrition grade et nombre d'additifs

Réduction dimensionnelle – ACP – Cercle des corrélations

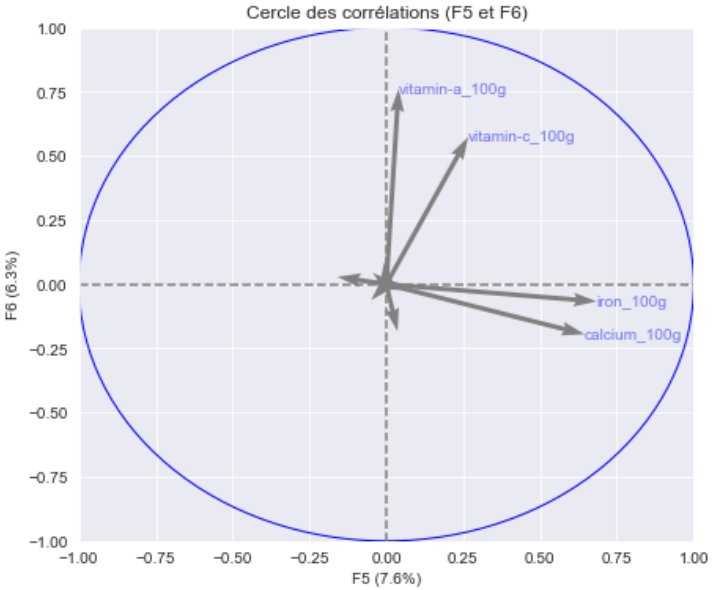






Composantes principales et corrélations entre variables et axes d'inertie

id		COR_1	COEFF_F1	COR_2	COEFF_F2	COR_3	COEFF_F3	COR_4	COEFF_F4	COR_5	COEFF_F5	COR_6	COEFF_F6
0	energy_100g	0.92	0.51	-0.04	-0.03	-0.01	-0.01	0.20	0.17	-0.06	-0.05	0.01	0.01
1	fat_100g	0.81	0.45	0.19	0.13	-0.38	-0.29	-0.11	-0.09	-0.05	-0.04	0.04	0.04
2	saturated-fat_100g	0.76	0.42	0.15	0.10	-0.29	-0.22	-0.29	-0.26	0.02	0.02	0.03	0.03
3	trans-fat_100g	0.04	0.02	-0.00	-0.00	0.02	0.01	-0.03	-0.03	-0.01	-0.01	0.09	0.09
4	cholesterol_100g	0.02	0.01	0.11	0.07	-0.03	-0.02	-0.06	-0.05	0.04	0.04	-0.18	-0.18
5	carbohydrates_100g	0.42	0.23	-0.42	-0.29	0.63	0.47	0.36	0.32	-0.02	-0.02	-0.02	-0.02
6	sugars_100g	0.44	0.24	-0.44	-0.31	0.65	0.49	-0.05	-0.05	0.07	0.06	-0.02	-0.02
7	fiber_100g	0.17	0.09	-0.02	-0.01	-0.12	-0.09	0.84	0.73	-0.17	-0.16	0.03	0.03
8	proteins_100g	0.25	0.14	0.32	0.22	-0.49	-0.36	0.43	0.38	-0.05	-0.05	-0.06	-0.06
9	salt_100g	-0.01	-0.00	0.86	0.60	0.47	0.35	0.06	0.05	-0.02	-0.01	-0.00	-0.00
10	sodium_100g	-0.01	-0.00	0.86	0.60	0.48	0.36	0.06	0.05	-0.01	-0.01	-0.00	-0.00
11	vitamin-a_100g	-0.01	-0.01	0.04	0.03	-0.00	-0.00	0.02	0.02	0.04	0.04	0.77	0.76
12	vitamin-c_100g	-0.02	-0.01	-0.00	-0.00	0.01	0.01	0.05	0.05	0.29	0.27	0.58	0.57
13	calcium_100g	0.06	0.03	0.07	0.05	-0.11	-0.08	0.13	0.11	0.71	0.65	-0.20	-0.20
14	iron_100g	0.01	0.01	0.01	0.00	-0.01	-0.01	0.15	0.13	0.75	0.68	-0.07	-0.07
15	nutrition-score-fr_100g	0.83	0.46	0.09	0.06	0.17	0.13	-0.31	-0.27	0.07	0.06	-0.01	-0.01



Les variables synthétiques F1, F2, F3, F4 vont exprimer:

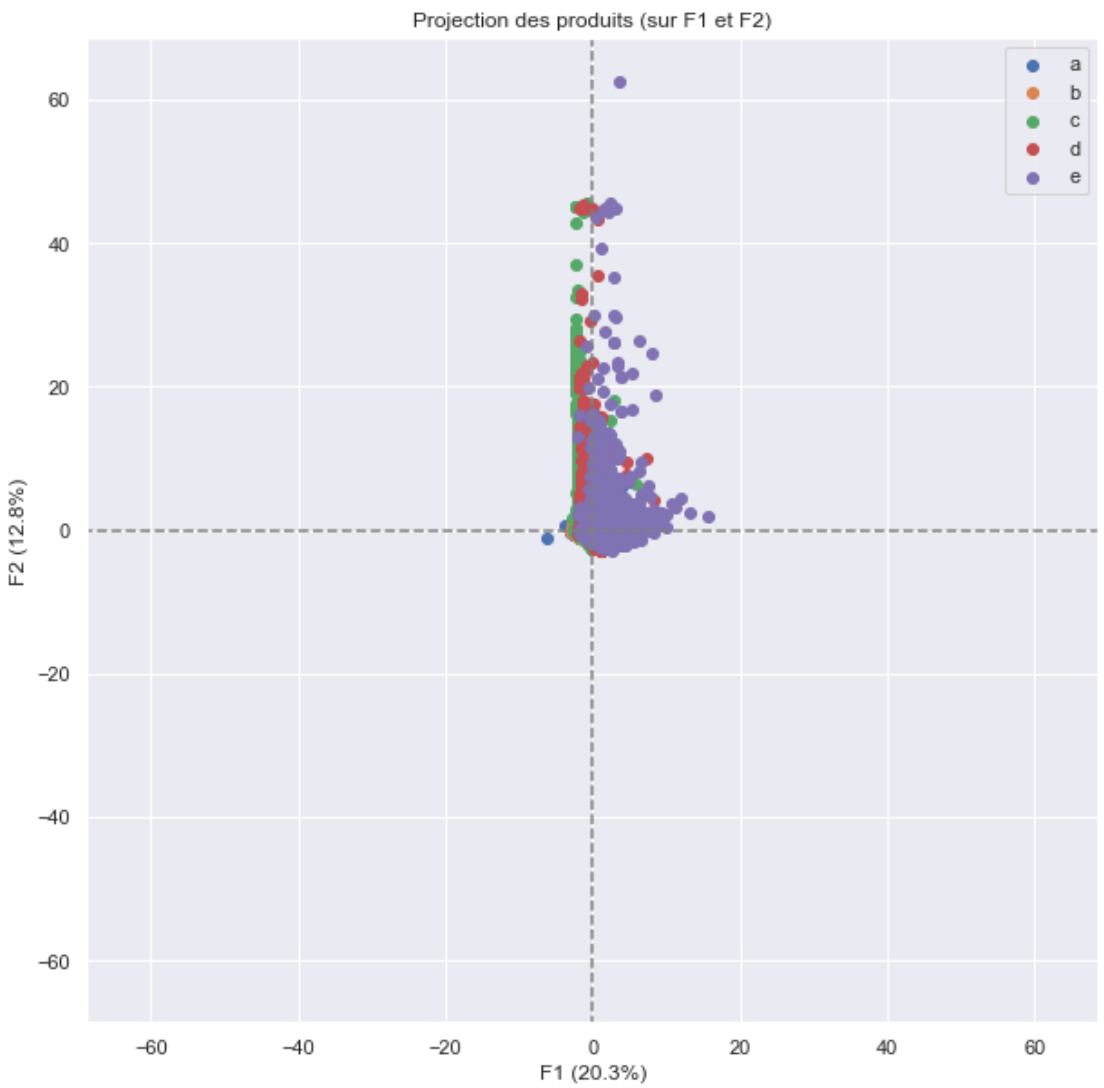
F1 = énergie et matières grasses

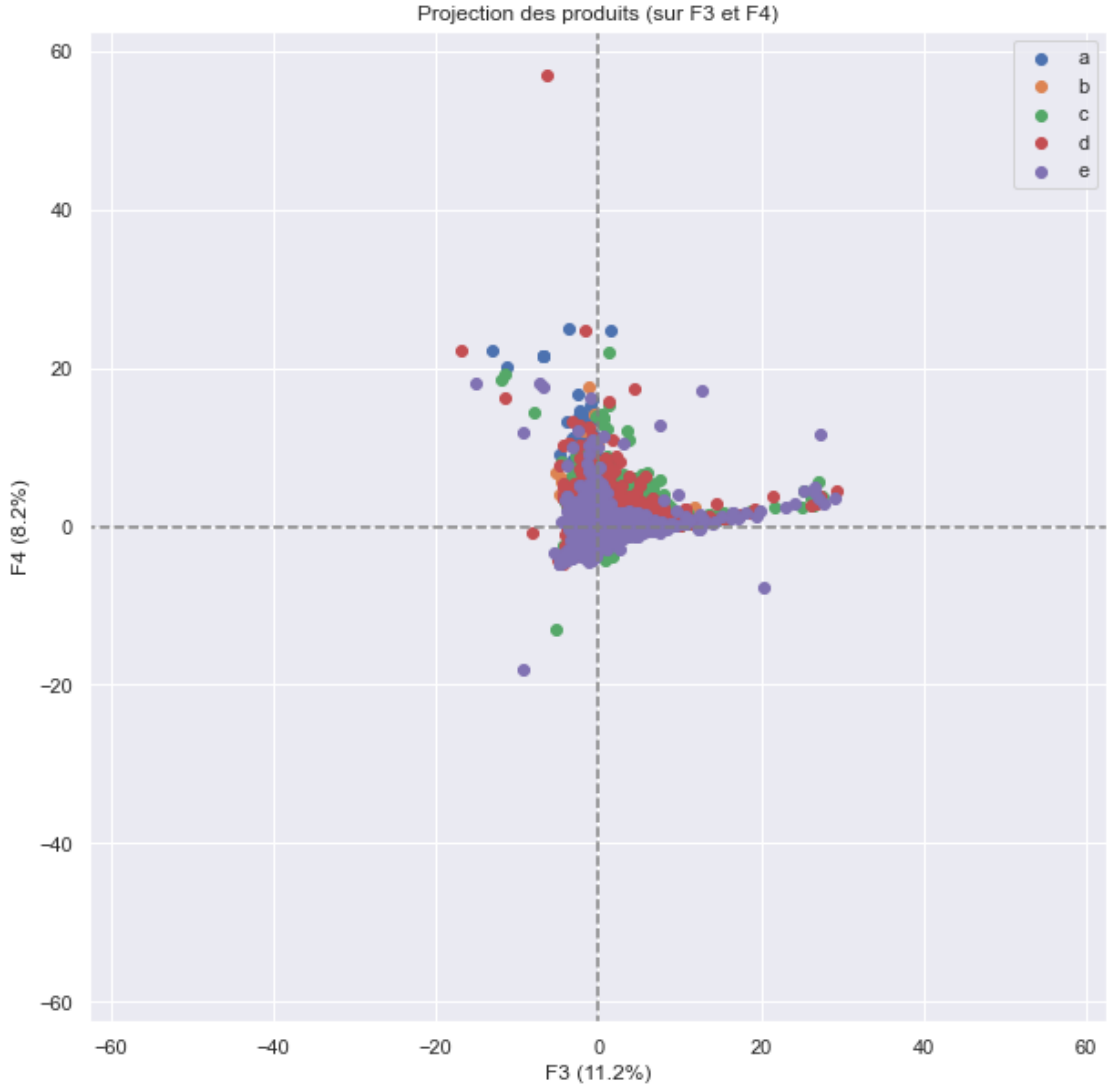
F2 = sel et sodium

F3 = sucres et carbohydrates

F4 = fibres et protéines

Réduction dimensionnelle – ACP – Projection des produits sur les plans factoriels





Conclusions

- Dans le même groupe de produits est toujours possible de trouver des alternatives avec différentes/meilleures évaluations

- L'évaluation de la typologie d'additifs est très importante car on retrouve dans les produits le 7% des additifs utilisés qui sont interdits en Europe. Il y a une relation entre nombre d'additifs et évaluation du produit, donc chercher un produit avec nombre réduit d'additifs signifie au même temps améliorer son Nutrigrade

- On a introduit quatre variables synthétiques avec l'ACP afin de regrouper les aspectes nutritionnelles importantes des produits en regroupent l'énergie avec les matières grasses, le sel avec le sodium, sucres et carbohydrates et fibres et protéines pour faciliter la compréhension de la part du consommateur. Les couleurs des quatre indicateurs de l'application pourraient se baser sur les valeurs des quatre variables synthétiques F1, F2, F3, F4.

- La projection sur les plans factoriels ne permet pas d'identifier des clusters séparés associés aux produits par rapport à leur Nutrigrade. Afin de permettre la mise en place d'un système de suggestion de produits alternatifs on pourrait utiliser une technique 'content based filtering' basée sur un algorithme KNN appliquée aux produits du même groupe alimentaire similaires au produit sélectionné par le consommateur.