Parcours Data Scientist

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

Daniel CHASTANET

Sommaire

- ❑ Rappel de la problématique et idée d'application
- ❑ Nettoyage et analyse pré-exploratoire
- ❑ Analyse exploratoire
- ❑ Remplacement des données manquantes
- ❑ Mise en œuvre de l'application
- ❑ Conclusion

Partie 1 : Présentation de l'idée d'application

Rappel de la problématique

Projet de création d'application en lien avec la nutrition/santé

BDD : Open Food Fact

- Automatisation du traitements des données
- Visualisations simples pour présenter les données
- Hypothèses et tests statistiques appropriés
- Elaborer une idée d'application

Etude de ce qu'on trouve sur le « marché »

Nutriscore :

Système d'étiquetage nutritionnel pour faciliter la compréhension des informations nutritionnelles (état : ministère de la Santé / volontariat)

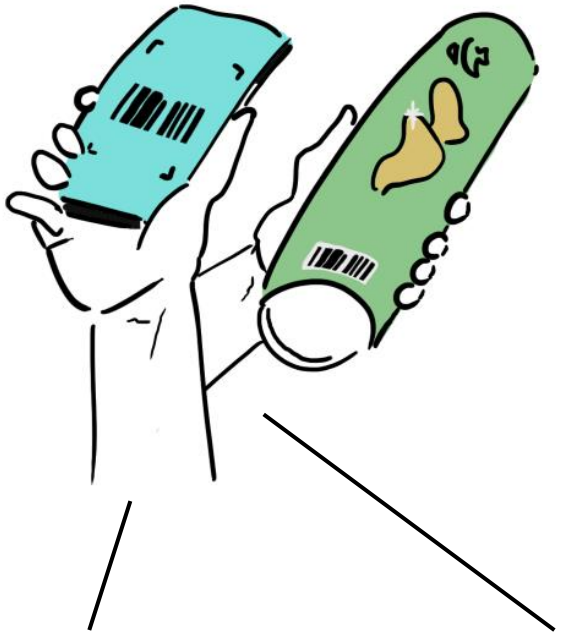
Eco score :

Indicateur représentant l'impact environnemental des produits alimentaires (état : ADEME / volontariat)

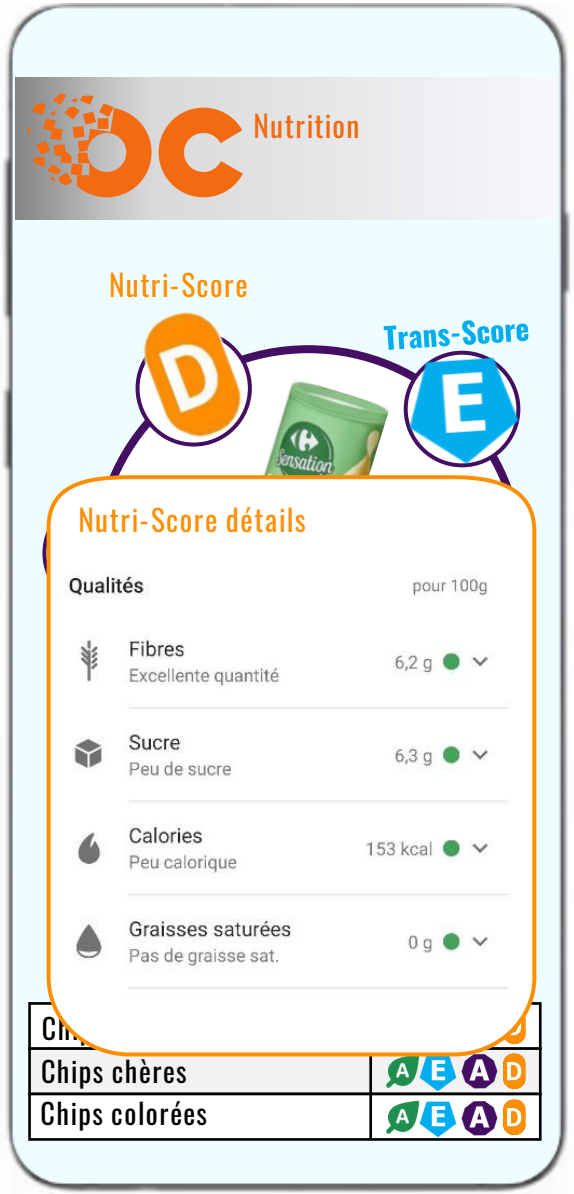
Yuka :

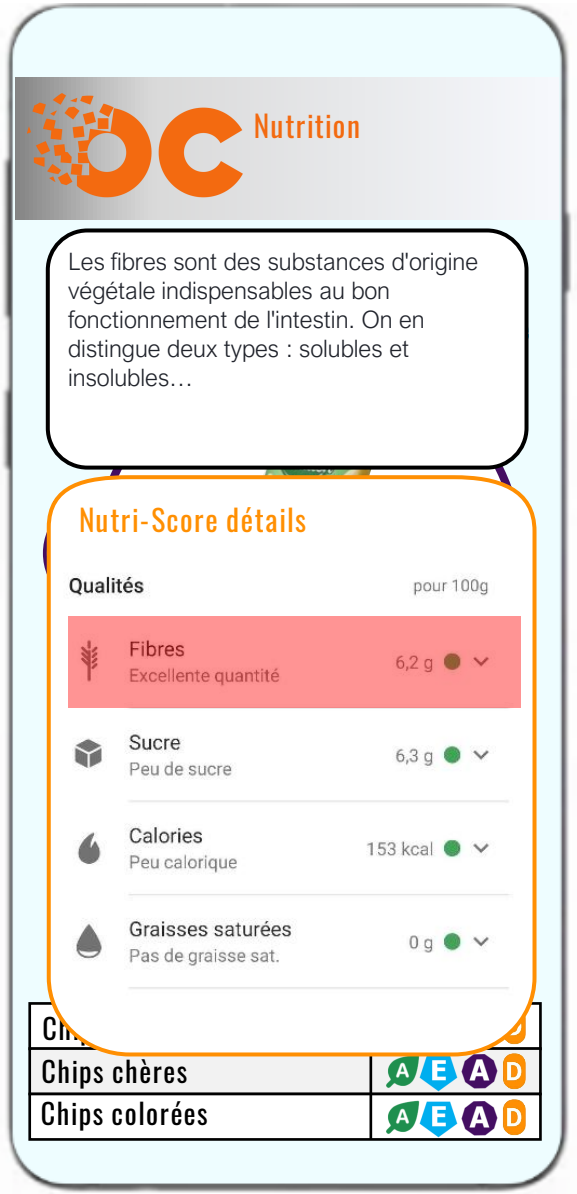
décrypte les étiquettes des produits alimentaires et cosmétiques et analyse leur impact sur la santé.
Une base de données complète : (1,5 million de produits alimentaires, 500 000 références cosmétiques, 800 nouveaux produits chaque jour)
Recommandation
Financement indépendant

L'application



Scan du code bar





L'application

Nom / Définition	Colonnes du dataframe
Nutri-Score : (Existant) : Barème calories, acides gras, sucres, fibres…	nutriscore_score / nutriscore_grade
k-score (non existant) : calories et mauvais gras !	energy_100g / fat_100g / saturated-fat_100g / carbohydrates_100g / sugars_100g / fiber_100g / proteins_100g / salt_100g
Addi-Score (mi - existant) : Nombre d'additifs (et dangerosité)	additives_n / Additives / additives_tags / additives_en
Nova-Score (existant) : Nombre de transformation et type	nova_group
Eco Score (Existant) : représente l'impact environnemental des produits alimentaires. Émissions de gaz à effet de serre (CO2) / Destruction de la couche d'ozone / Émissions de particules fines / Oxydation photochimique / Acidification / Radioactivité / etc…	ecoscore_score_fr / ecoscore_grade_fr
Σ Somme des autres score	

Nutri-Score

k-score

Addi-Score

Nova-Score

Eco-Score

Méga-score

Partie 2 : Nettoyage de la base de données

Prendre connaissances de la BDD

Légende des données :

```
dimension : (1918228, 186)
```

code barcode of the product (can be EAN-13 or internal codes for some food stores)
003. quantity : quantity and unit
004. url : url of the product page on Open Food Facts
005. creator : contributor who first added the product
006. created_t : date that the product was added (UNIX timestamp format)
007. created_datetime : date that the product was added (iso8601 format: yyyy-mm-ddThh:mn:ssZ)
008. last_modified_t : date that the product page was last modified
009. last_modified_datetime : date that the product was lastly modified (UNIX timestamp format)
010. product_name : name of the product
011. abbreviated_product_name
012. generic_name : generic name of the product
013. quantity : en g/l/capsules... ou no unit
014. packaging : type --> shape, material : plastique / bocal etc...
015. packaging_tags : same que packaging avec des tirets à la place des espaces
016. packaging_text : '1 Pot en verre à recycler', ..., '1 PET packet'
017. brands : 'courte paille', ..., 'miellerie de la natouze', 'Biocoop Bordeaux lac'
018. brands_tags : same stuff que 'brands' avec le tiret
019. categories : 'Epicerie, Condiments, Sauces, Moutardes
…
040. ingredients_text : description des ingrédients avec pourcent ou pas
041. allergens : 'en:mustard', 'en:eggs,en:mustard', ..., 'fr:Gs1:T4078:AI,fr:Gs1:T4078:BA
042. allergens_en : ...
043. traces : 'fr:Fruits à coques et/ou cacahuètes'
044. traces_tags : ...
…

050. additives : empty...
051. additives_tags : 'en:e150,en:e160a,en:e202
052. additives_en : ...
053. ingredients_from_palm_oil_n : in number from nan, 0 to 5
054. ingredients_from_palm_oil : vide
055. ingredients_from_palm_oil_tags : 'stearine-de-palme,oleine-de-palme,'e304-palmitate-d-ascorbyle'
056. ingredients_that_may_be_from_palm_oil_n : number
057. ingredients_that_may_be_from_palm_oil
058. ingredients_that_may_be_from_palm_oil_tags
059. nutriscore_score : nan, 18., 1., 14., -2.
060. nutriscore_grade : nan, 'd', 'b', 'a', 'c', 'e' https://fr.openfoodfacts.org/nutriscore
061. nova_group : Une classification en 4 groupes pour mettre en évidence le degré de transformation des aliments nan, 1 à 4 https://fr.openfoodfacts.org/nova
062. pnns_groups_1 : Programme national nutrition santé (PNNS) / 'Fat and sauces', 'Composite foods', 'Sugary snacks'
063. pnns_groups_2 : 'Dressings and sauces', 'One-dish meals'
…
080. energy-kj_100g : you know...
energy-kcal_100g
energy_100g
…
-linoleic-acid_100g
-arachidonic-acid_100g
-gamma-linolenic-acid_100g
-dihomo-gamma-linolenic-acid_100g
omega-9-fat_100g
-oleic-acid_100g
-elaidic-acid_100g

Nettoyage des données

Very large data frame (au moins pour mon pc) : `dimension : (1918228, 186)`

Filtrage colonnes

↳ Removing columns with 80% and more missing values :

(1918228, 57)

↳ Keeping only interesting columns (for the app) :

(1918228, 21)

Filtrage lignes

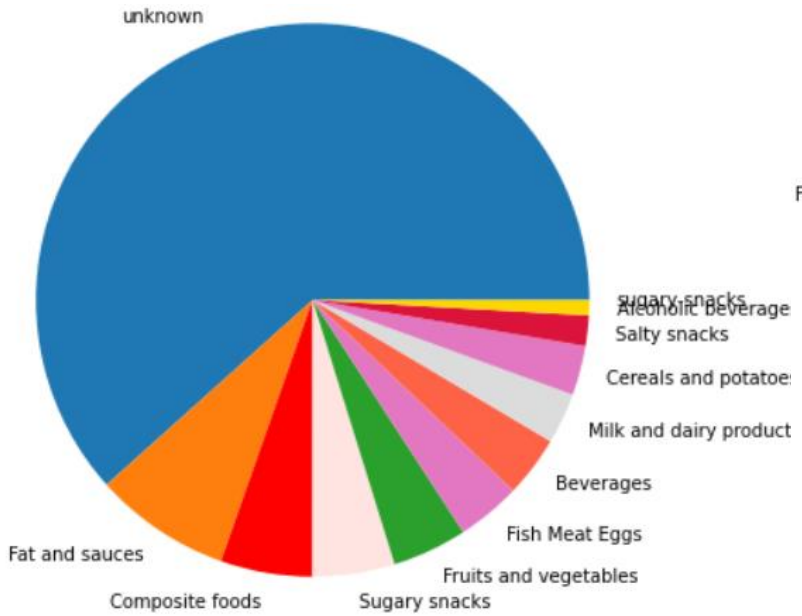
↳ Filtrage par « ecoscore_score » != isna() :

(462847, 21)

% de données manquantes	
code	0.000000
product_name	0.725942
ingredients_text	45.071698
additives_n	45.071482
additives_tags	71.674009
additives_en	71.674009
nutriscore_score	22.651978
nutriscore_grade	22.651978
nova_group	49.852327
pnns_groups_1	0.000000
pnns_groups_2	0.000000
ecoscore_score_fr	0.000000
ecoscore_grade_fr	0.000000
energy_100g	15.015113
fat_100g	15.299440
saturated-fat_100g	16.818301
carbohydrates_100g	15.357559
sugars_100g	16.347519
fiber_100g	65.960026
proteins_100g	15.245211
salt_100g	16.701199

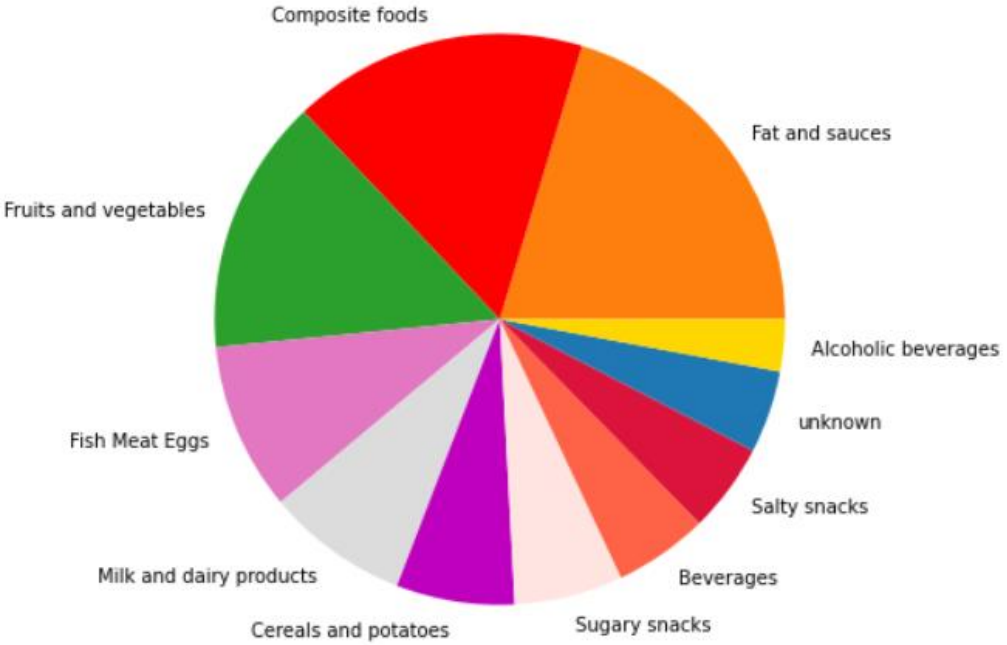
Nettoyage des données

Do I cut too much ? Let's check the pnns_groups (pnns_1)



pnns cat : 12 / 42

(1918228, 21)



pnns cat : 11 / 39

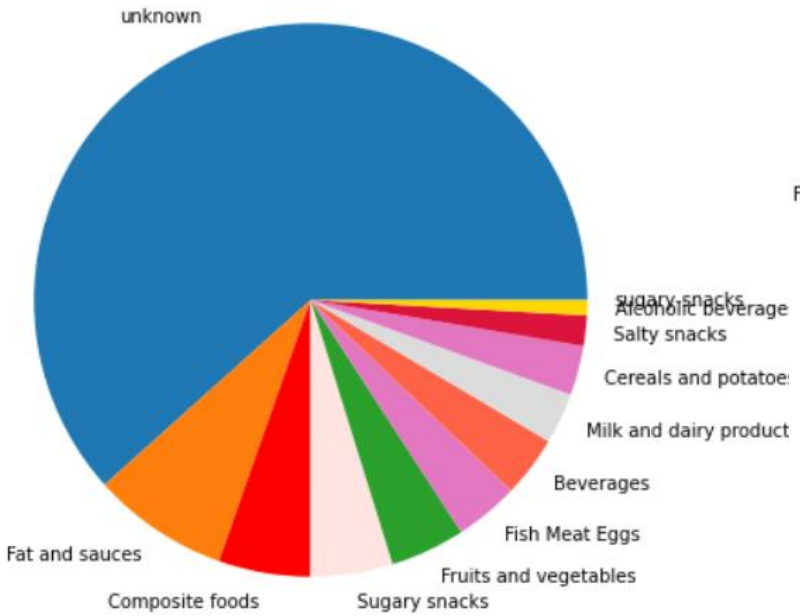
(462847, 21)

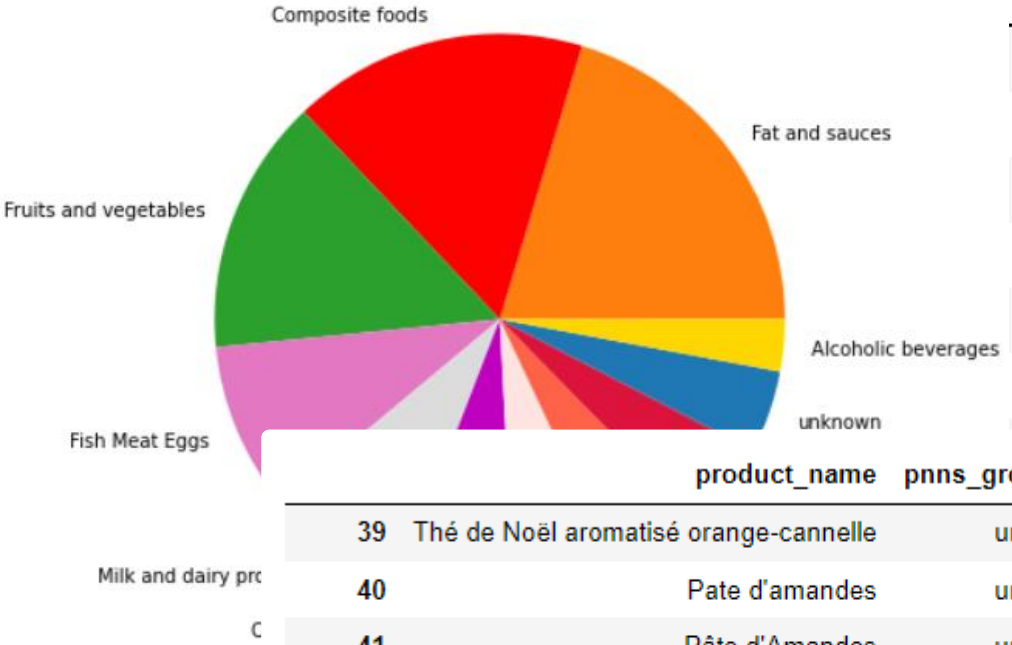
Waters and flavored waters 6863

	pnns_groups_1	n	n_filtrés
0	Alcoholic beverages	17040	13591.0
1	Beverages	68447	23376.0
2	Cereals and potatoes	83804	43985.0
3	Composite foods	55513	21630.0
4	Fat and sauces	71737	37372.0
5	Fish Meat Eggs	102955	67140.0
6	Fruits and vegetables	56971	31006.0
7	Milk and dairy products	92067	77362.0
8	Salty snacks	33976	28531.0
9	Sugary snacks	153243	94136.0
10	sugary-snacks	2	NaN
11	unknown	1181969	24718.0

Nettoyage des données

Do I cut too much ? Let's check the pnns_groups (pnns_1)





pnns cat : 12 / 42

(1918228, 21)

	pnns_groups_1	n	n_filtrés
0	Alcoholic beverages	17040	13591.0
1	Beverages	68447	23376.0
2	Cereals and potatoes	83804	43985.0
3	Composite foods	55513	21630.0
4	Fat and sauces	71737	37372.0
5	Fish Meat Eggs	102955	67140.0
		56971	31006.0
		92067	77362.0
		33976	28531.0
		153243	94136.0
		2	NaN
		1181969	24718.0

Waters and flavored waters 6863

	product_name	pnns_groups_1	ecoscore_score_fr
39	Thé de Noël aromatisé orange-cannelle	unknown	92.0
40	Pate d'amandes	unknown	5.0
41	Pâte d'Amandes	unknown	10.0
57	Tisane nerf - sommeil	unknown	94.0
73	Pomme dauphine	unknown	38.0

Traitement valeurs abhérantes

(462847, 21)

Filtrage valeurs aberrantes :

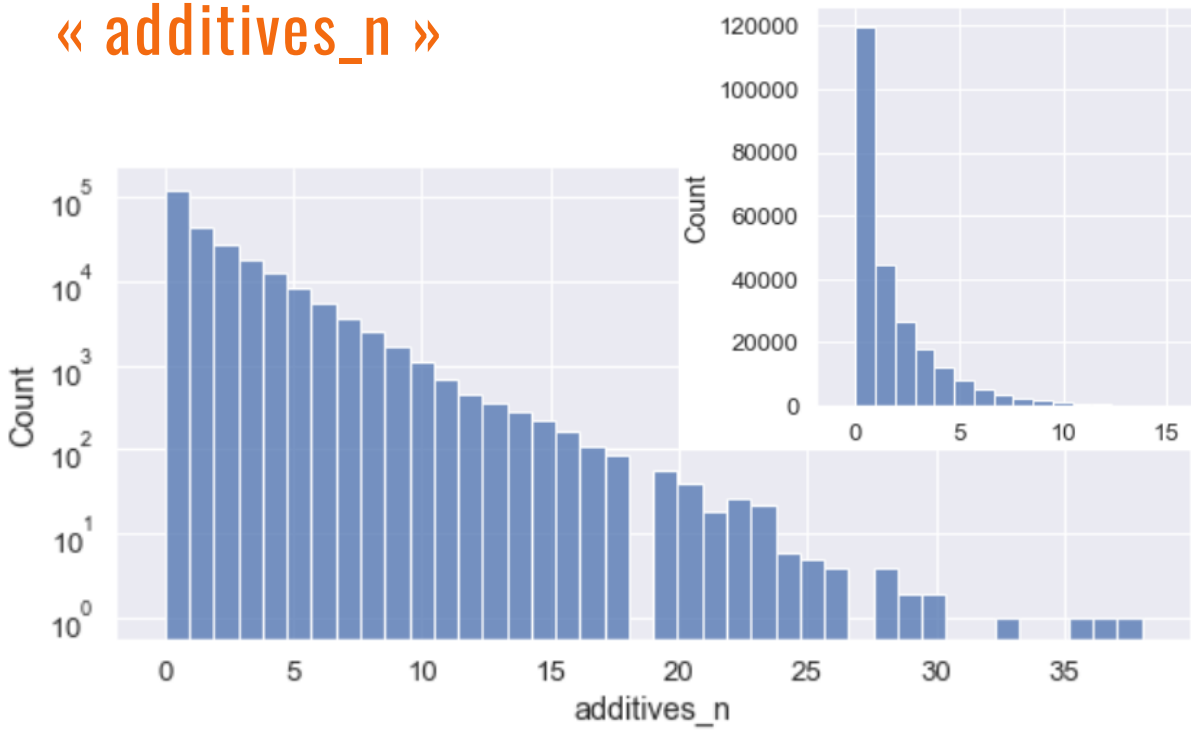
- ❑ Nombre de Kalories aberrantes (trop grande) 900kcal pour 100g au maximum
- ❑ 150 g de Ж dans 100 g de produit (Ж pouvant être du sel, sucre, fibre, etc)
- ❑ Valeurs négatives en g pour 100 g
- ❑ Produit dont la sommes des composants dépasse les 100g pour 100g
- ❑ Drop duplicate

(452724, 21)

Partie 3 : Analyse exploratoire

Analyses univariées

« additives_n »

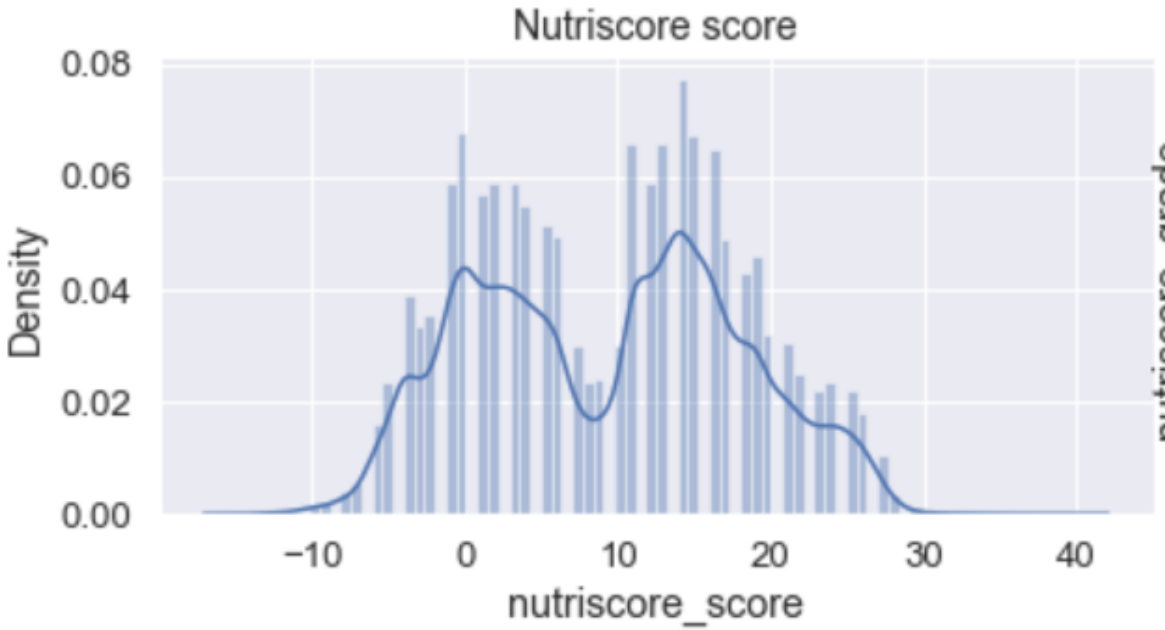


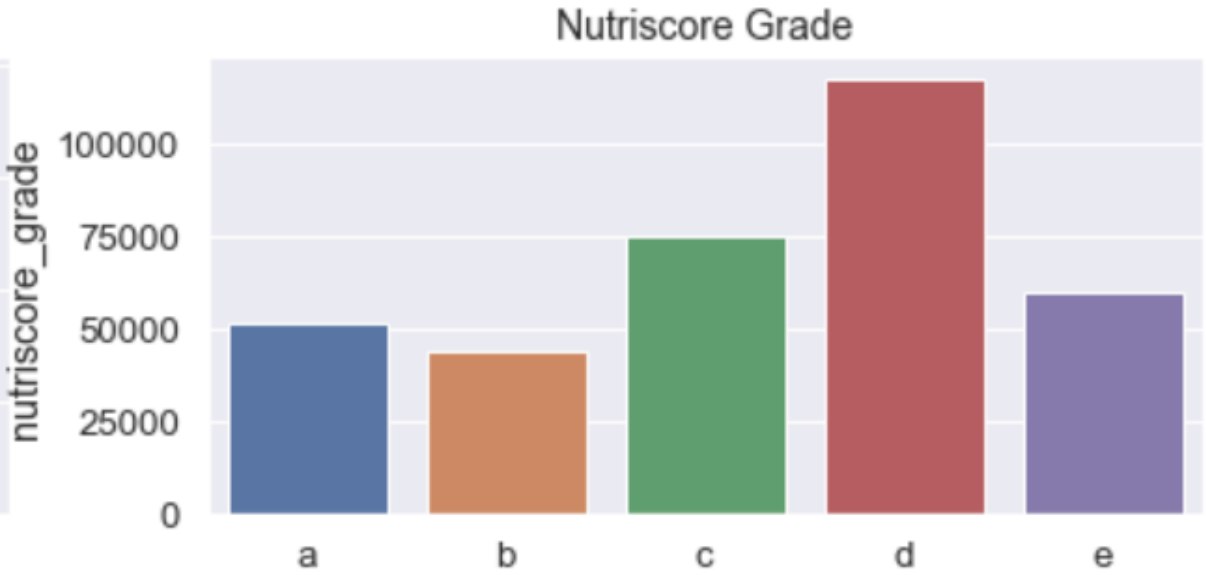
nombre d'additifs dans le produit

	additives_n	n	f	F
0	0.0	119387	0.263708	0.263708
1	1.0	44282	0.097812	0.361520
2	2.0	26805	0.059208	0.420729
3	3.0	17843	0.039413	0.460141
4	4.0	12395	0.027379	0.487520
5	5.0	8245	0.018212	0.505732
6	6.0	5507	0.012164	0.517896
7	7.0	3651	0.008065	0.525961
8	8.0	2466	0.005447	0.531408
9	9.0	1672	0.003693	0.535101
10	10.0	1103	0.002436	0.537537

Analyses univariées

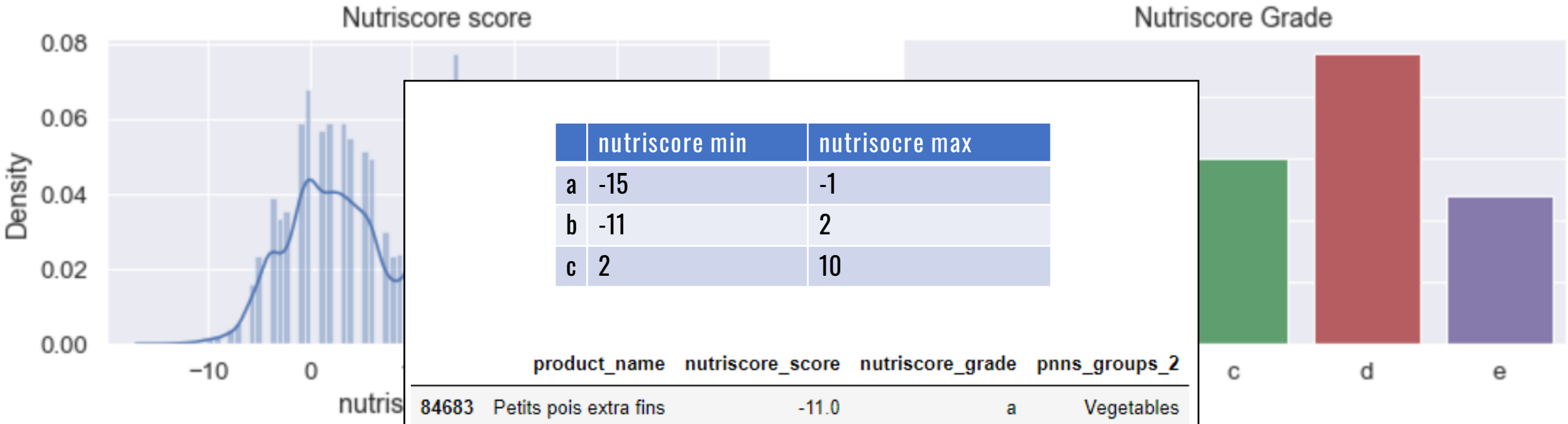
«nutriscore_score », « nutriscore_grade »





Analyses univariées

«nutriscore_score », « nutriscore_grade »

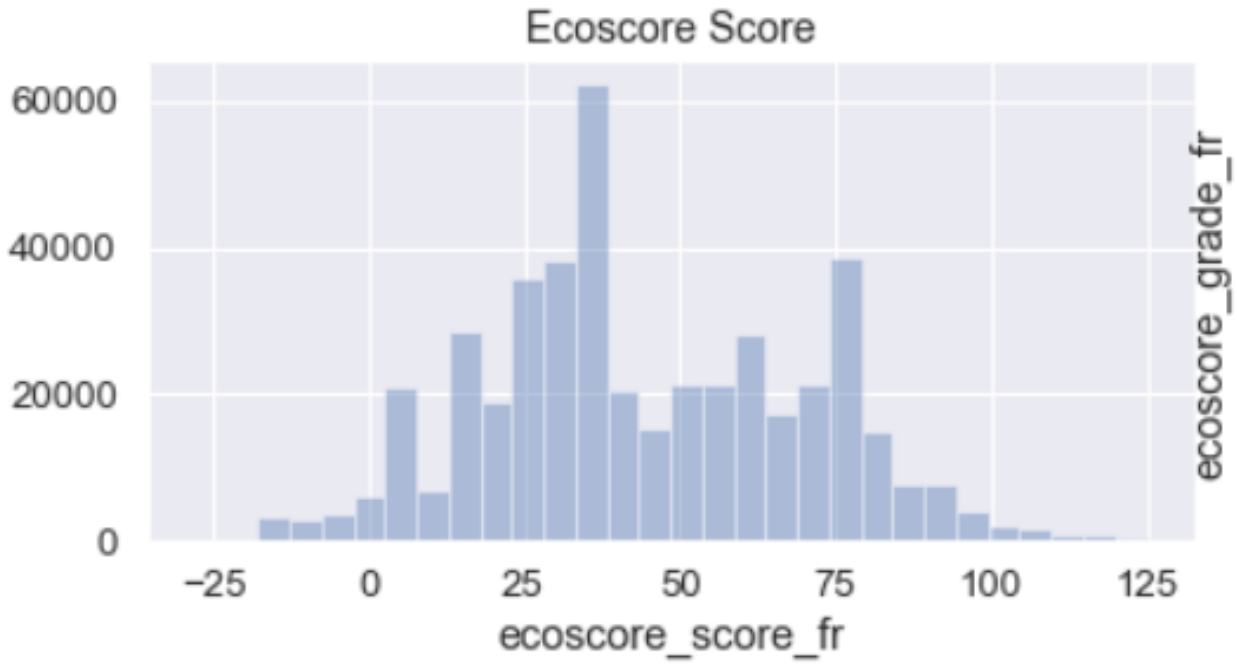


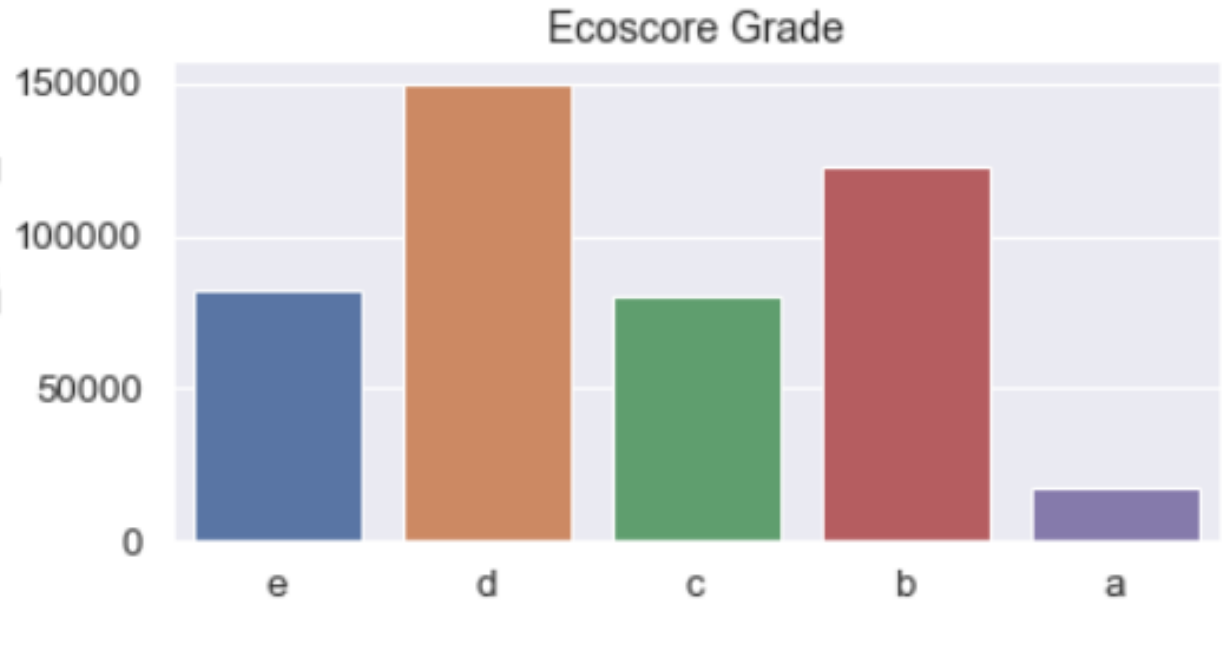
	nutriscore min	nutrisocre max
a	-15	-1
b	-11	2
c	2	10

	product_name	nutriscore_score	nutriscore_grade	pnns_groups_2
84683	Petits pois extra fins	-11.0	a	Vegetables
79542	Zumo	-11.0	b	Fruit juices

Analyses univariées

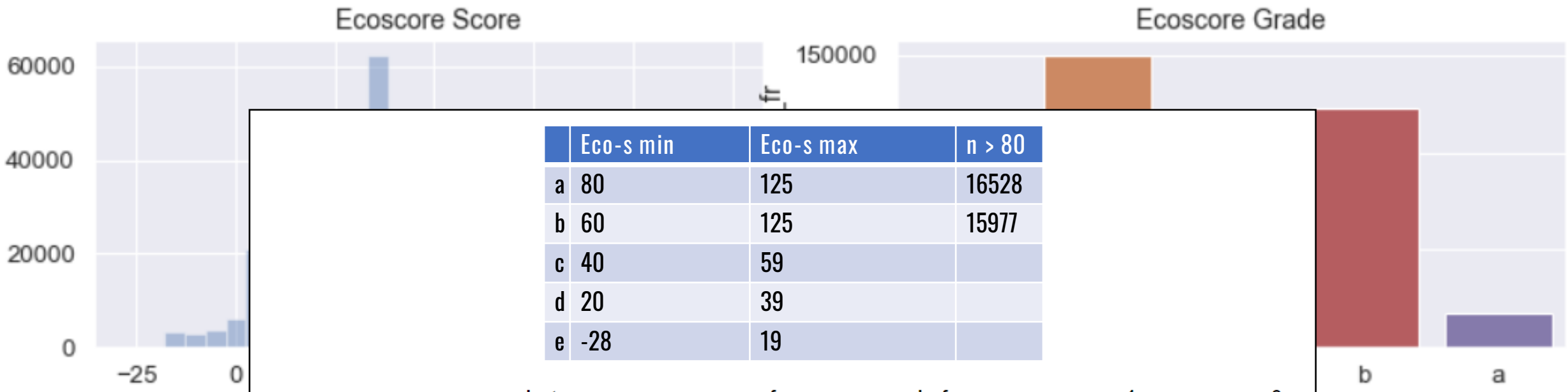
«ecoscore_score », « ecoscore_grade »





Analyses univariées

«ecoscore_score », « ecoscore_grade »

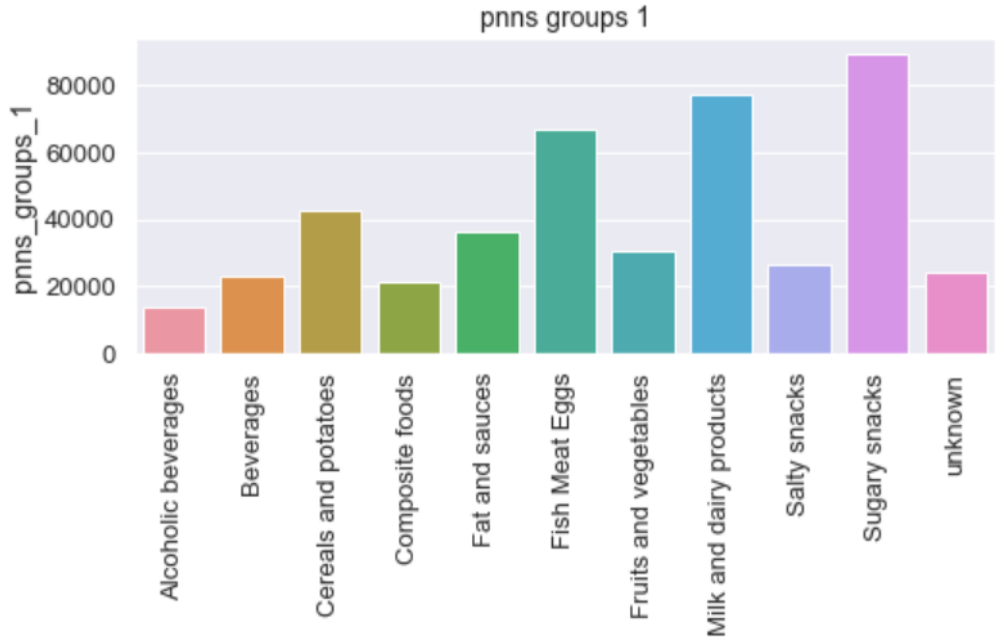


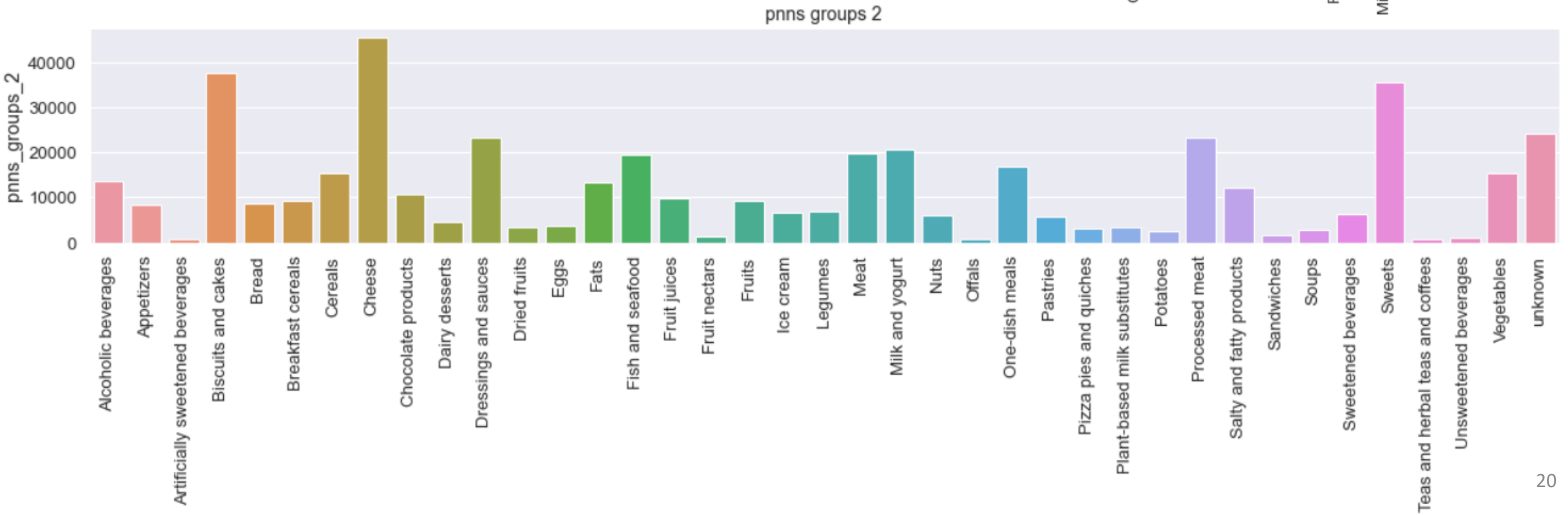
	Eco-s min	Eco-s max	n > 80
a	80	125	16528
b	60	125	15977
c	40	59	
d	20	39	
e	-28	19	

	product_name	ecoscore_score_fr	ecoscore_grade_fr	pnns_groups_1	pnns_groups_2
31	lentilles vertes	99.0	b	Cereals and potatoes	Legumes
39	Thé de Noël aromatisé orange-cannelle	92.0	a	unknown	unknown

Analyses univariées

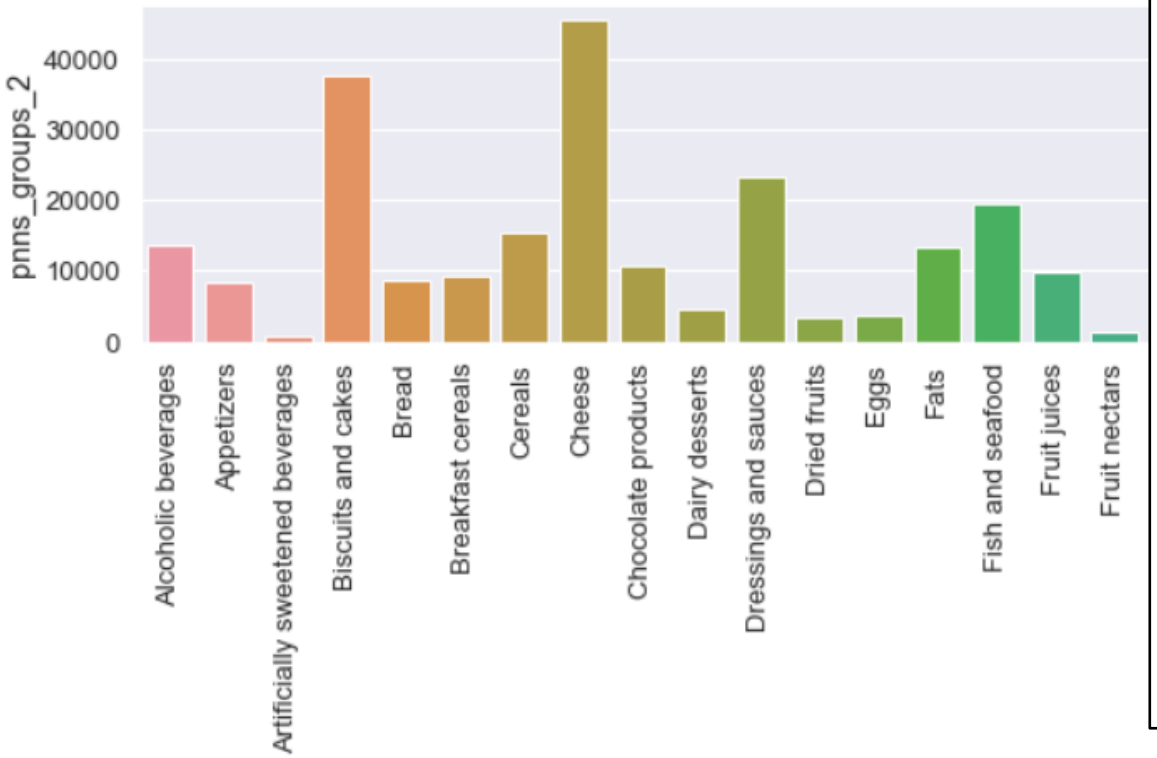
« pnns_groups_1 », « pnns_groups_2 »
Nombre d'individus



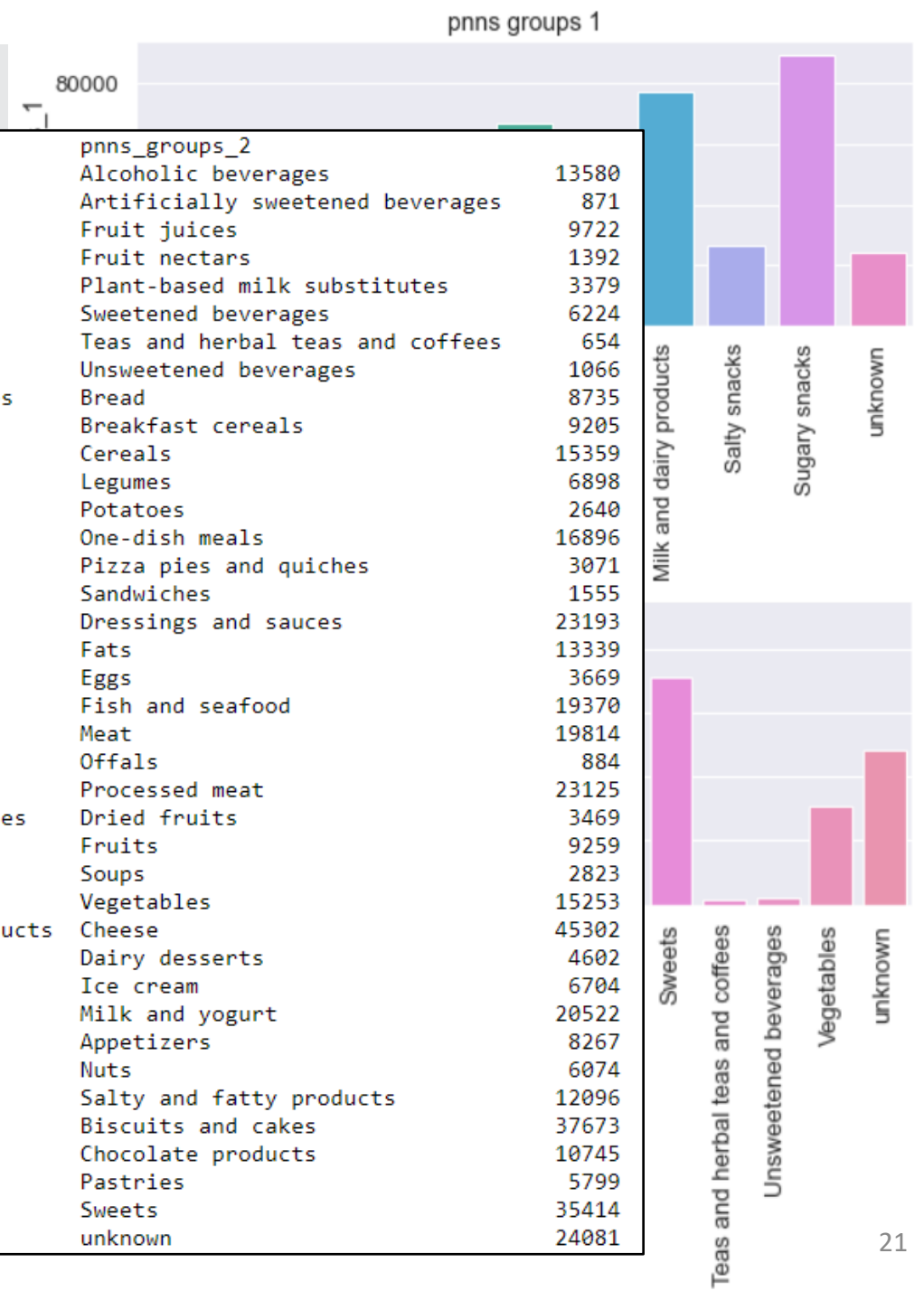


Analyses univariées

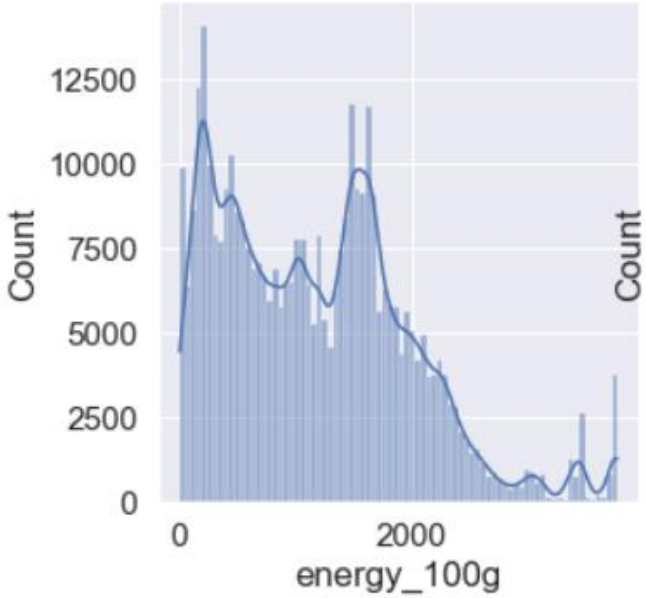
« pnns_groups_1 », « pnns_groups_2 »
Nombre d'individus

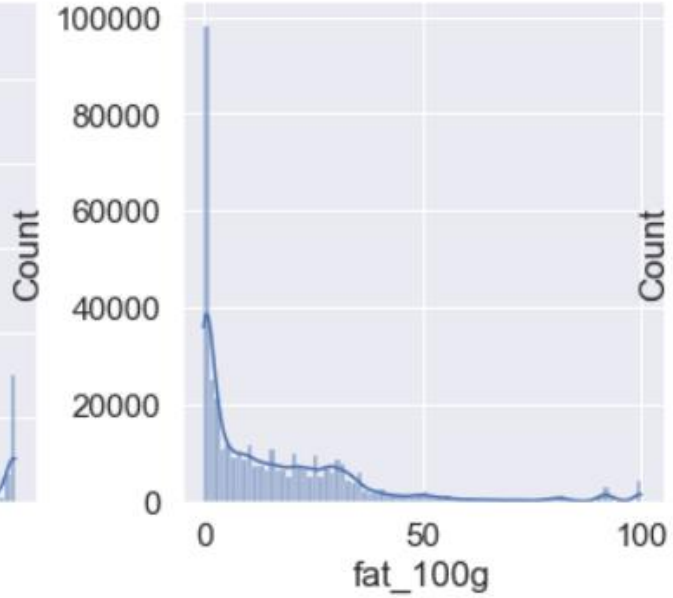


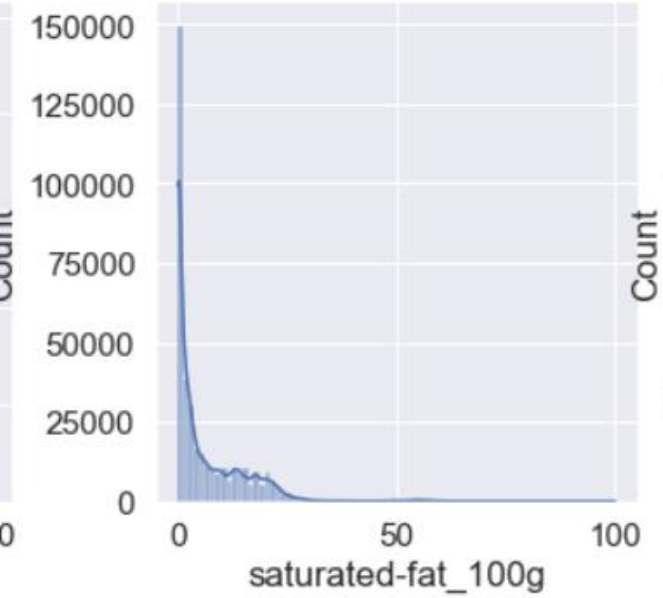
pnns_groups_1	pnns_groups_2	
Alcoholic beverages	Alcoholic beverages	13580
Beverages	Artificially sweetened beverages	871
	Fruit juices	9722
	Fruit nectars	1392
	Plant-based milk substitutes	3379
	Sweetened beverages	6224
	Teas and herbal teas and coffees	654
	Unsweetened beverages	1066
Cereals and potatoes	Bread	8735
	Breakfast cereals	9205
	Cereals	15359
	Legumes	6898
	Potatoes	2640
Composite foods	One-dish meals	16896
	Pizza pies and quiches	3071
	Sandwiches	1555
Fat and sauces	Dressings and sauces	23193
	Fats	13339
Fish Meat Eggs	Eggs	3669
	Fish and seafood	19370
	Meat	19814
	Offals	884
	Processed meat	23125
Fruits and vegetables	Dried fruits	3469
	Fruits	9259
	Soups	2823
	Vegetables	15253
Milk and dairy products	Cheese	45302
	Dairy desserts	4602
	Ice cream	6704
	Milk and yogurt	20522
Salty snacks	Appetizers	8267
	Nuts	6074
	Salty and fatty products	12096
Sugary snacks	Biscuits and cakes	37673
	Chocolate products	10745
	Pastries	5799
	Sweets	35414
unknown	unknown	24081

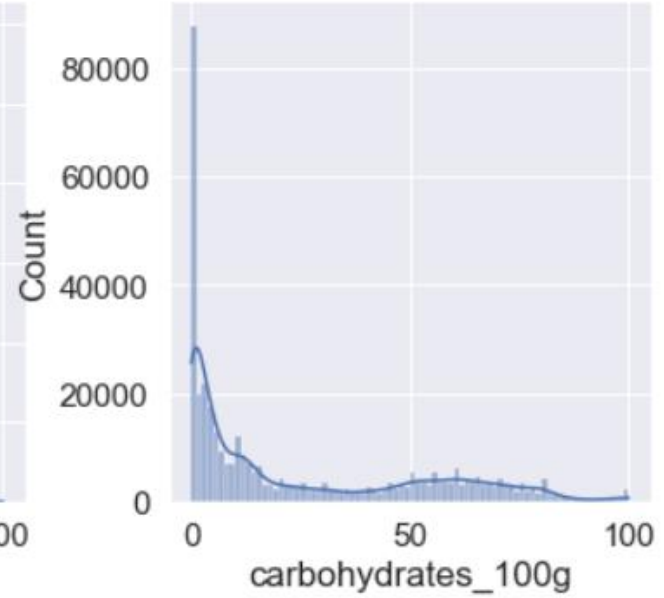


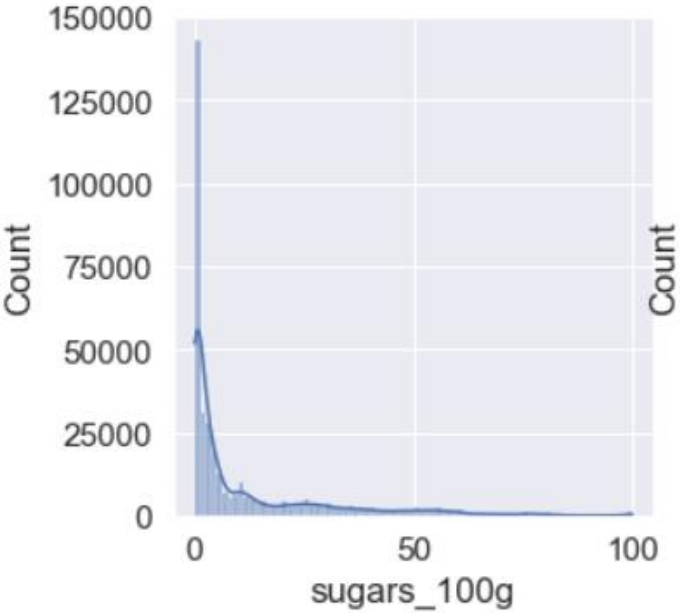
Analyses univariées « …100g »

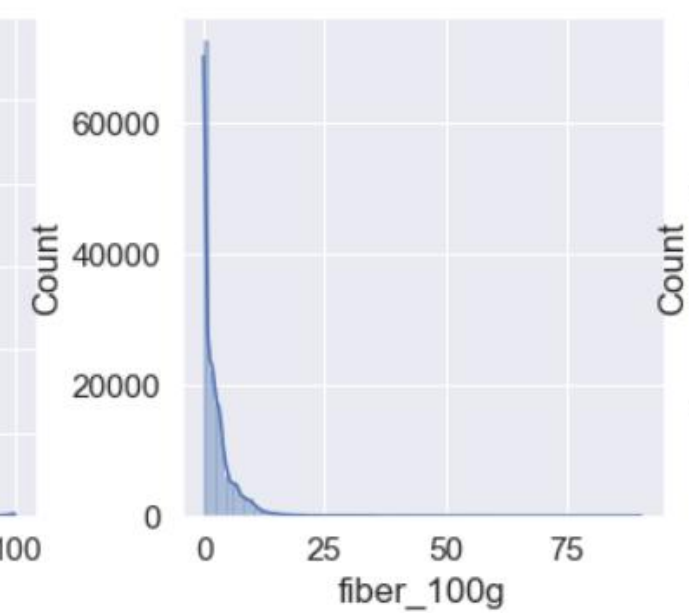


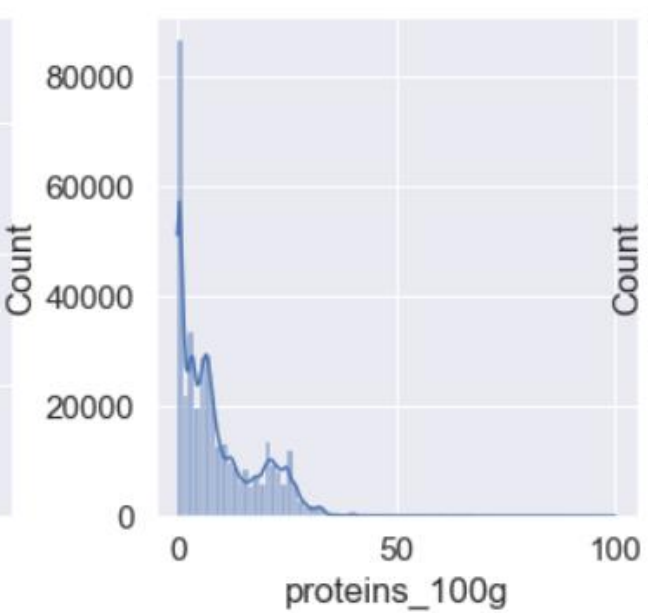


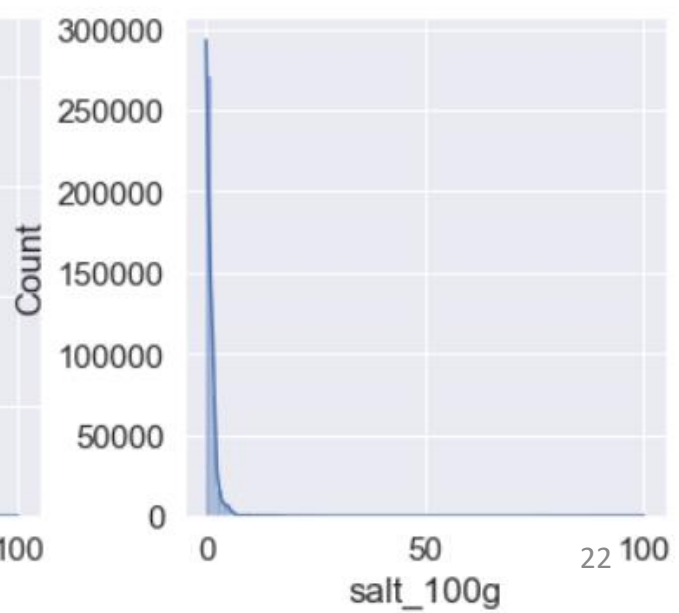






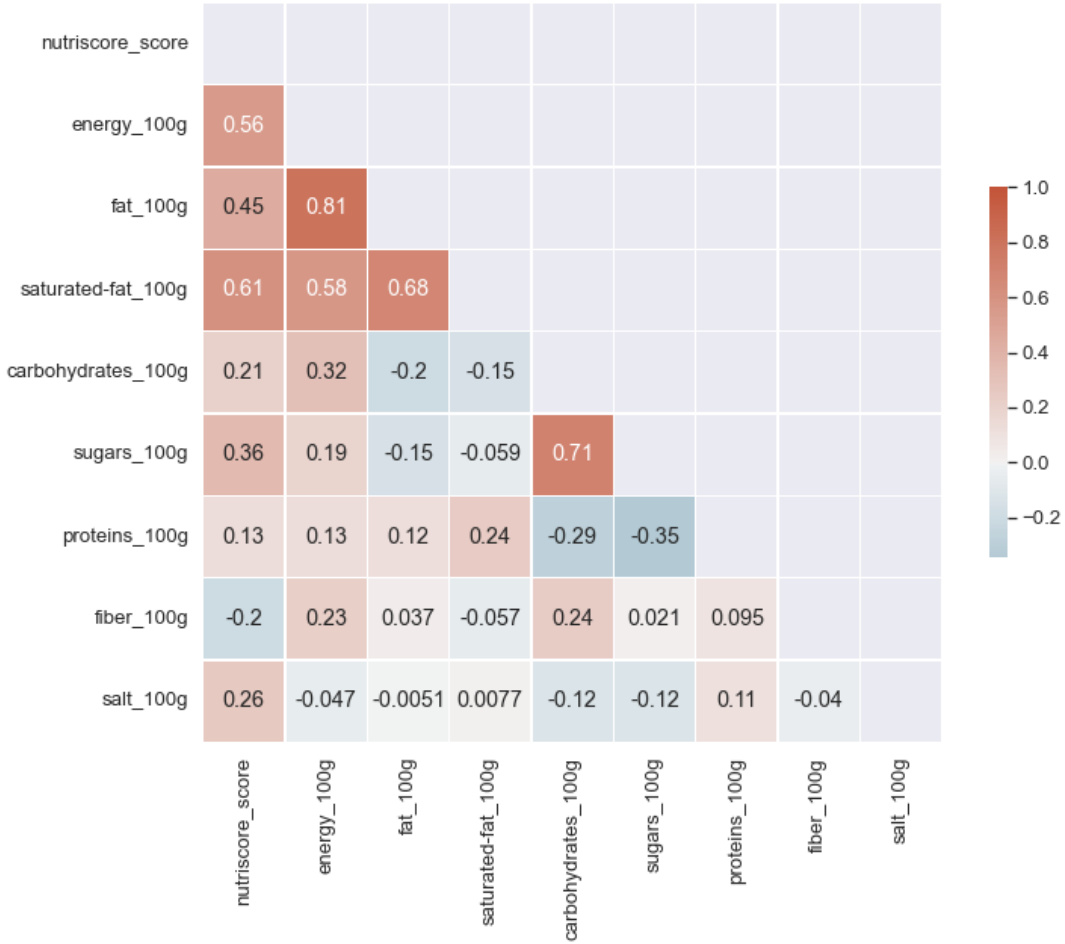


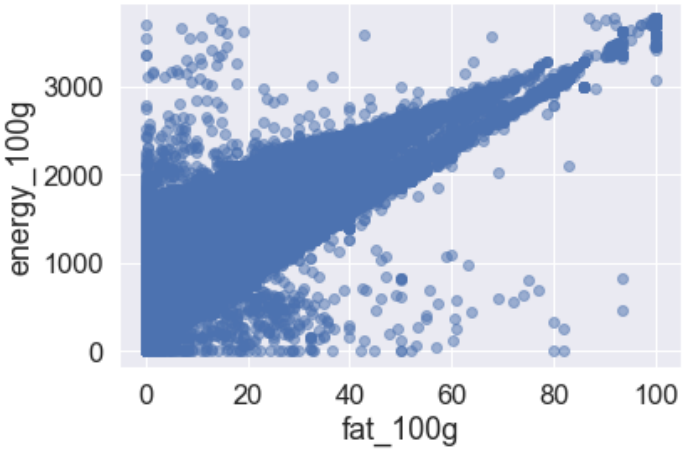




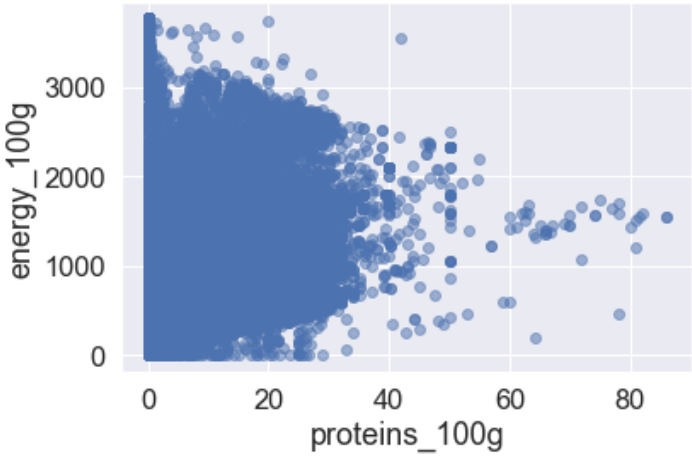
Analyses multivariées

Corrélation des catégorie « nutriscore et …_100g »





coefficient de Pearson : 0.7727228924095313
covariance : 9014.94379179825

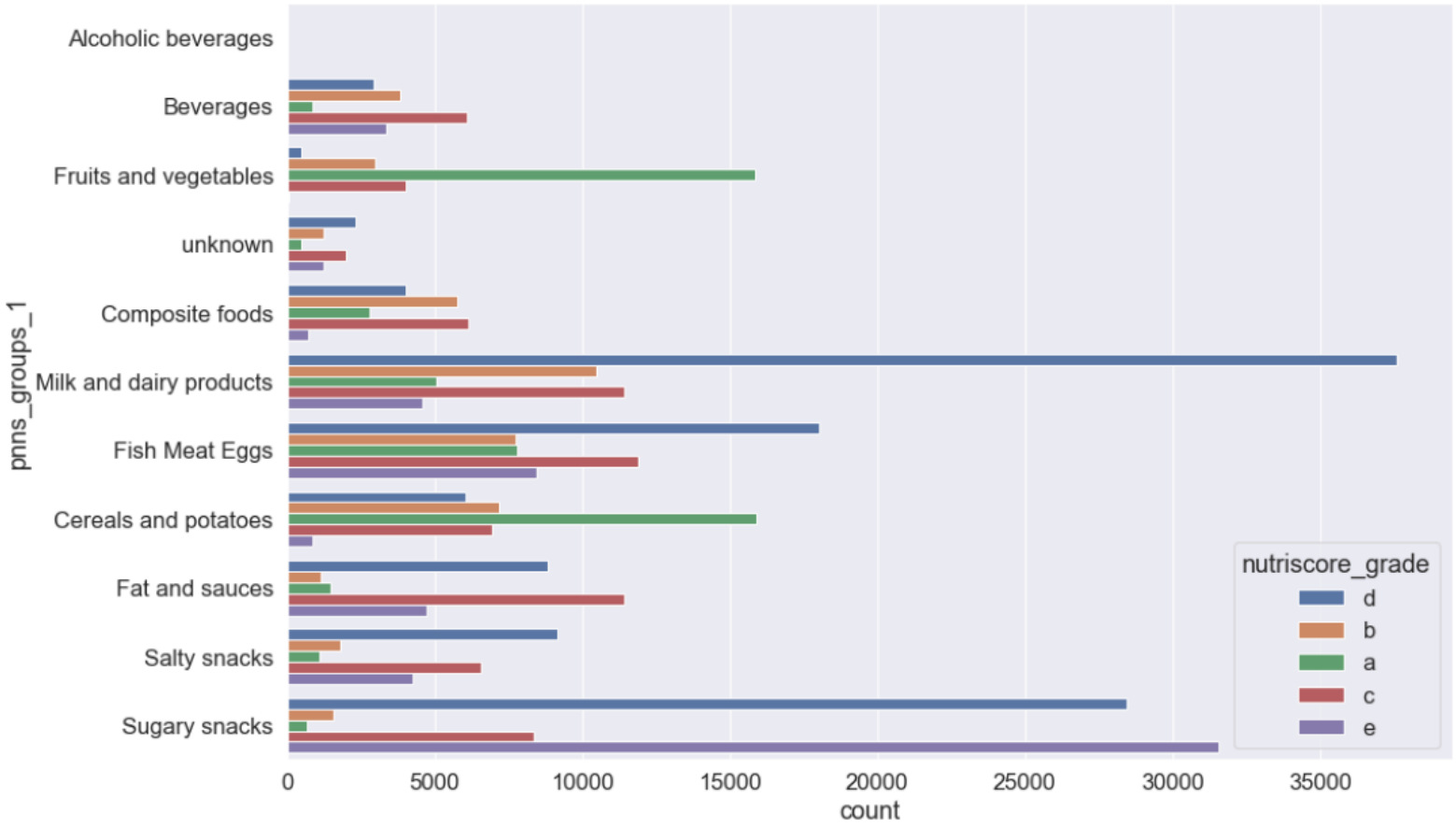


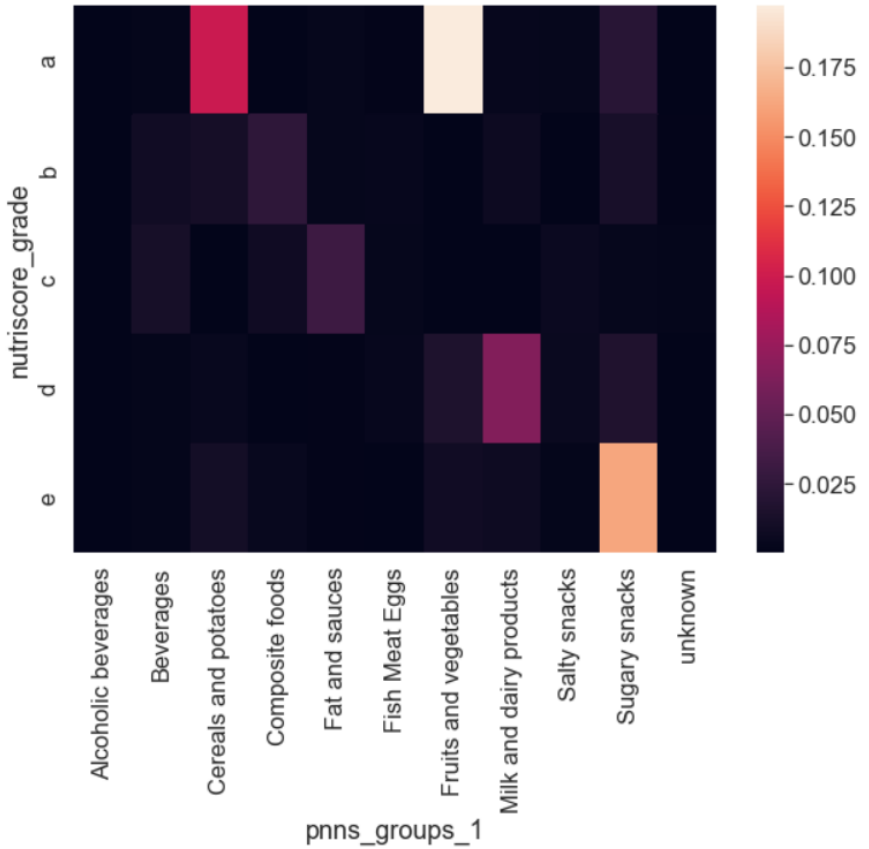
coefficient de Pearson : 0.25738134613637953
covariance : 1516.176868031524

	product_name	energy_100g	fat_100g
3290	Pure Sesame Oil	7.14	100.000000
5645	Organic Extra Virgin Olive Oil	558.00	93.333333
19086	California Extra Virgin Olive Oil	335.00	93.333333
28228	Premium Extra Virgin Olive Oil	0.00	93.333333
37080	Extra Virgin Olive Oil	463.00	93.333333

Analyses multivariées

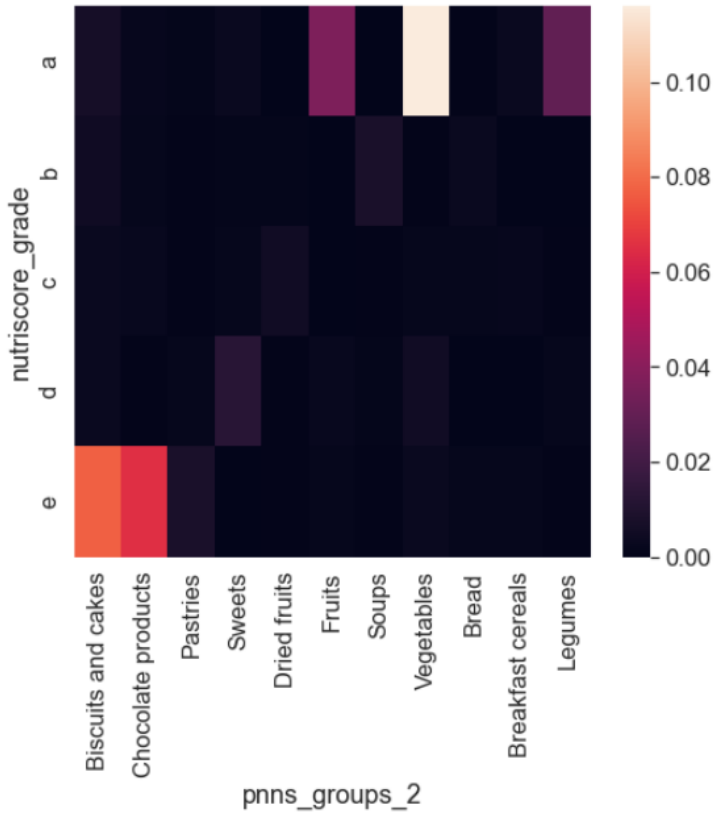
«nutriscore_grade » vs « pnns_1 »



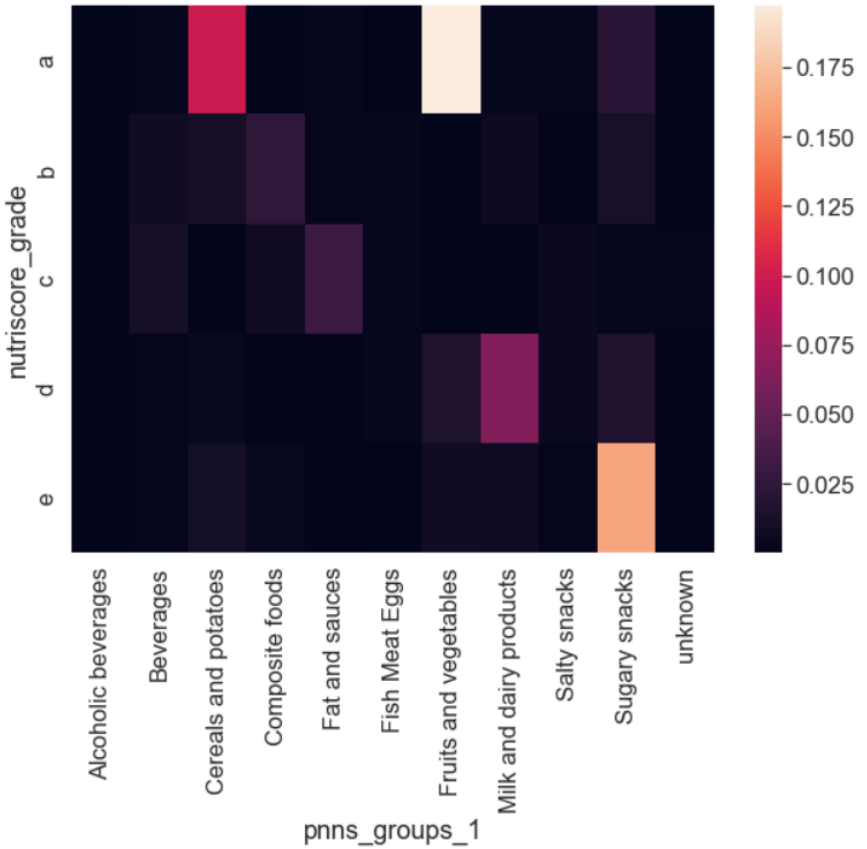


Analyses multivariées

«nutriscore_grade » vs « pnns_1 »

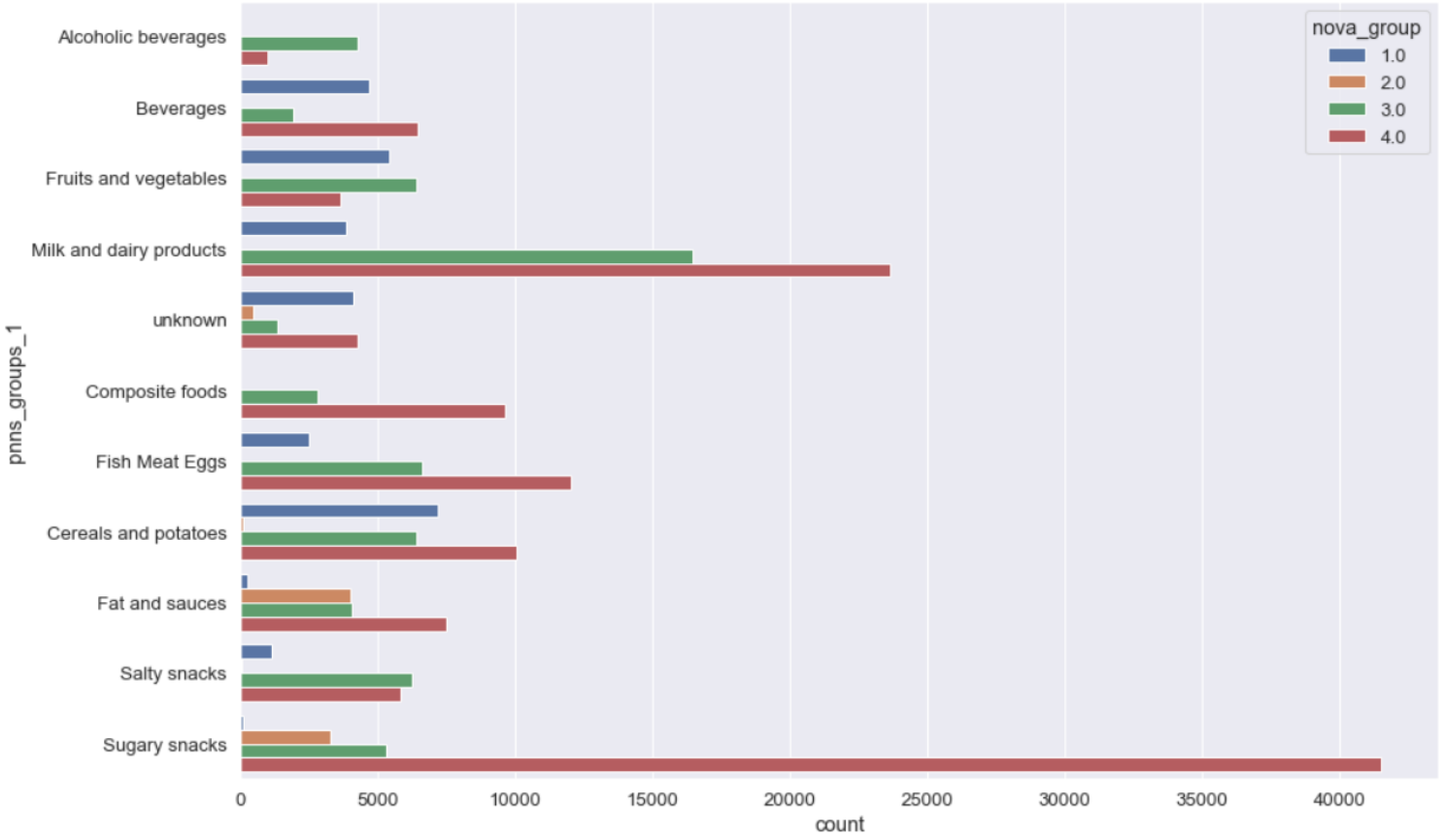


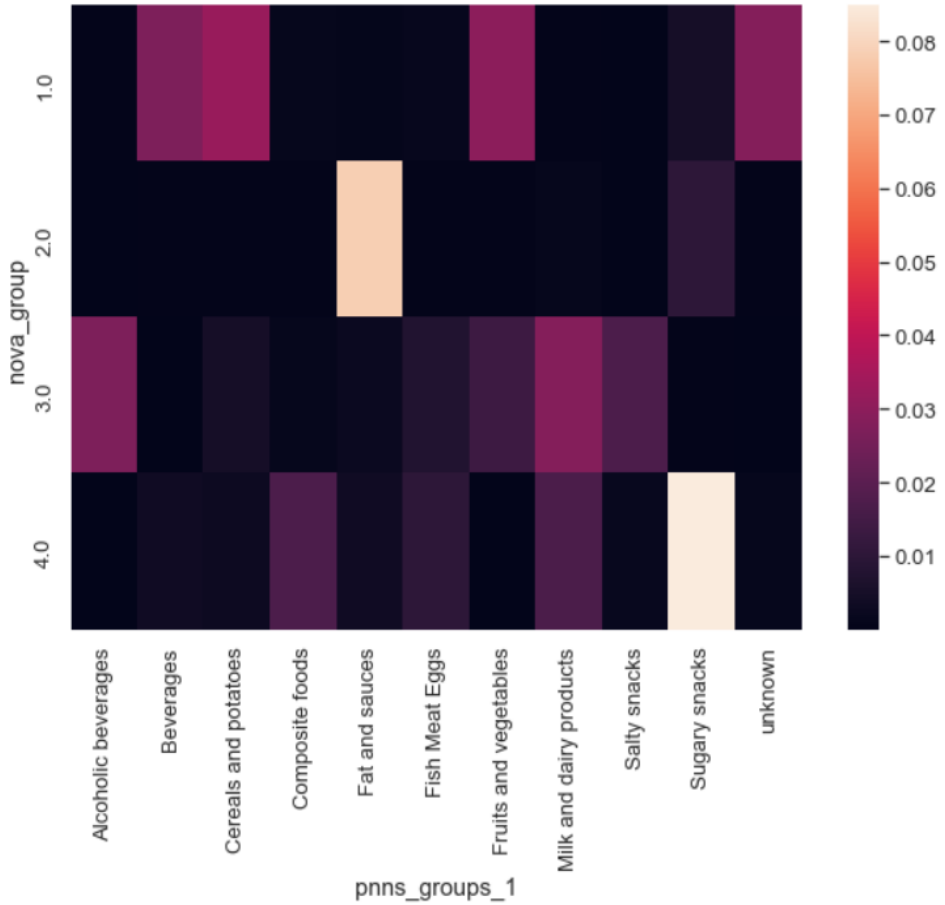
Sugary snacks	Biscuits and cakes	37673
	Chocolate products	10745
	Pastries	5799
	Sweets	35414
Fruits and vegetables	Dried fruits	3469
	Fruits	9259
	Soups	2823
	Vegetables	15253
Cereals and potatoes	Bread	8735
	Breakfast cereals	9205
	Cereals	15359
	Legumes	6898
	Potatoes	2640



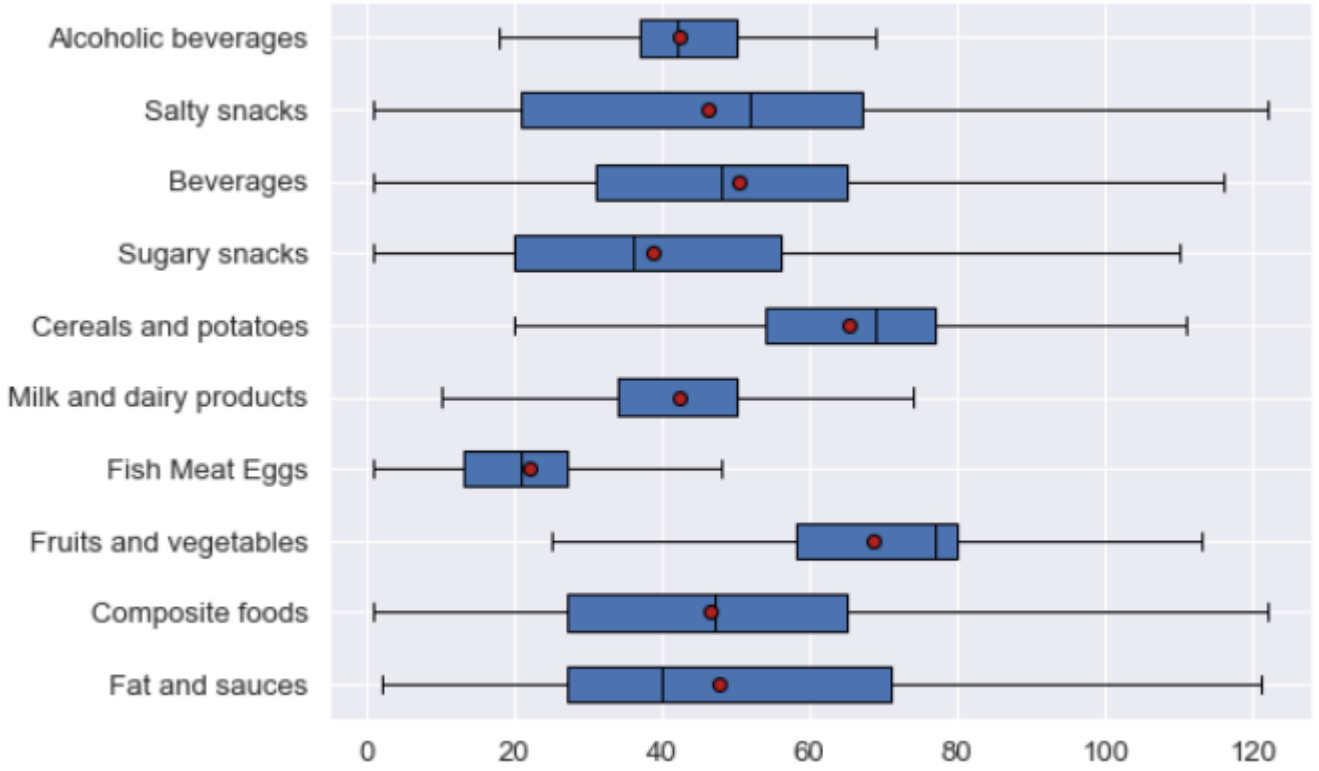
Analyses multivariées

«nova_score » vs « pnns_1 »





Analyses multivariées

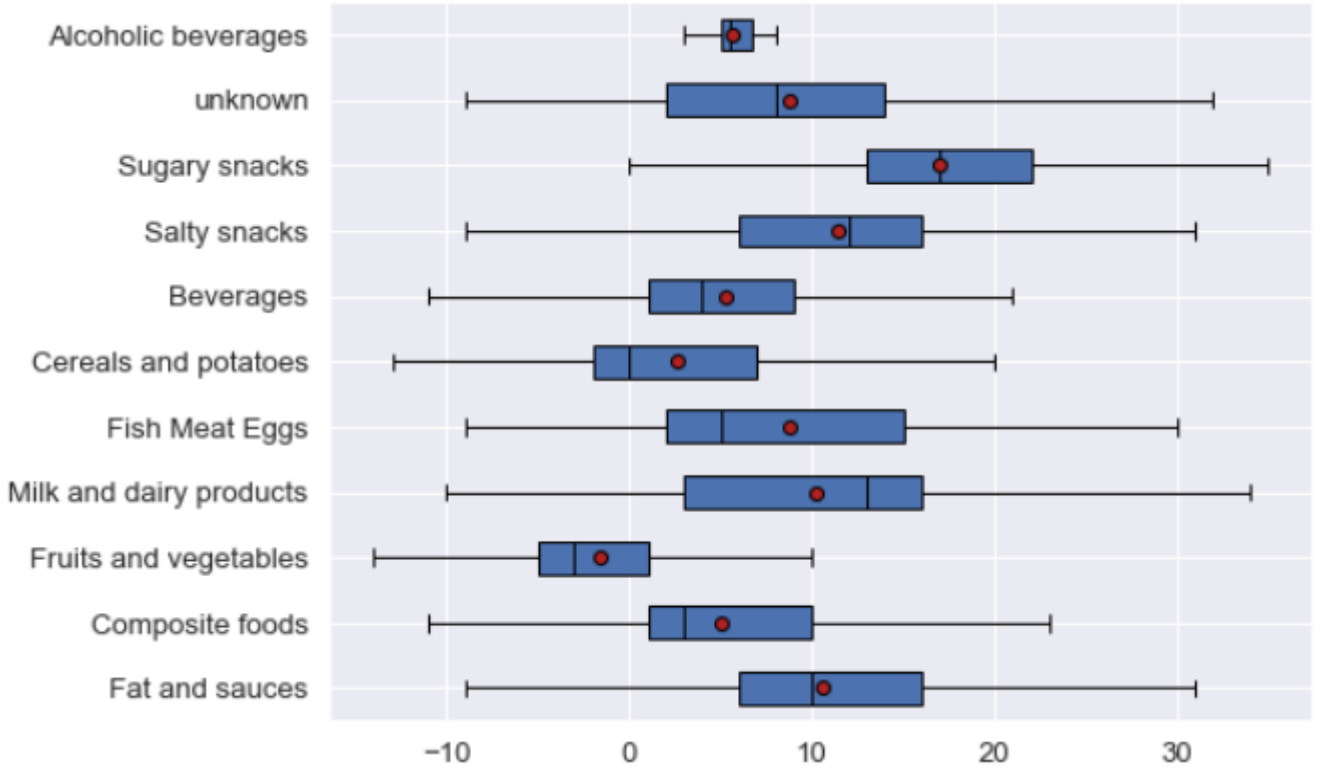


« ecoscore_score », « pnns_groups_1 »

	sum_sq	df	F	PR(>F)
C(treatments)	9.029856e+07	10.0	18992.520639	0.0
Residual	2.152391e+08	452713.0	NaN	NaN

	group1	group2	Diff	Lower	Upper	q-value	p-value
39	Milk and dairy products	Alcoholic beverages	0.246681	-0.40645	0.899813	1.719239	0.9

Analyses multivariées



«nutriscore_score », « pnns_groups_1 »

	sum_sq	df	F	PR(>F)
C(treatments)	9.482769e+06	10.0	19509.406494	0.0
Residual	1.694167e+07	348550.0	NaN	NaN

	group1	group2	p-value
9	Fat and sauces	Alcoholic beverages	0.788261
14	Composite foods	Beverages	0.264148
18	Composite foods	Alcoholic beverages	0.900000
26	Fruits and vegetables	Alcoholic beverages	0.281017
33	Milk and dairy products	Alcoholic beverages	0.868963
38	Fish Meat Eggs	unknown	0.900000
39	Fish Meat Eggs	Alcoholic beverages	0.900000
44	Cereals and potatoes	Alcoholic beverages	0.900000
48	Beverages	Alcoholic beverages	0.900000
51	Salty snacks	Alcoholic beverages	0.617893
54	unknown	Alcoholic beverages	0.900000

Partie 4 : Données manquantes

k-score

Méga-score

Application

Remplissage des données

product_name	0.733560
ingredients_text	45.679708
additives_n	45.679487
nutriscore_score	23.008058
nutriscore_grade	23.008058
nova_group	50.429842
pnns_groups_1	0.000000
pnns_groups_2	0.000000
ecoscore_score_fr	0.000000
ecoscore_grade_fr	0.000000
energy_100g	15.339368
fat_100g	15.635354
saturated-fat_100g	17.146208
carbohydrates_100g	15.692342
sugars_100g	16.669759
fiber_100g	67.034661
proteins_100g	15.576378
salt_100g	17.040183

Méthode de remplacement des données

Meanvalue

sklearn.impute.IterativeImputer

KNeighborsClassifier

Estimation via règle de calcul

sklearn.impute.IterativeImputer

pnns_groups_2	mean	std
Alcoholic beverages	0.085493	0.467025
Appetizers	1.907947	1.864349
Artificially sweetened beverages	4.153846	1.965620
Biscuits and cakes	3.876277	2.886765
Bread	1.975615	2.003255
Breakfast cereals	0.867898	1.064692
Cereals	0.154698	0.783192
Cheese	1.060064	1.350985
Chocolate products	0.994323	1.052275
Dairy desserts	1.817036	1.935492
Dressings and sauces	1.725133	1.429991
Dried fruits	0.156529	0.543390
Eggs	0.145271	0.793507

Remplissage des données

product_name	0.733560
ingredients_text	45.679708
additives_n	45.679487
nutriscore_score	23.008058
nutriscore_grade	23.008058
nova_group	50.429842
pnns_groups_1	0.000000
pnns_groups_2	0.000000
ecoscore_score_fr	0.000000
ecoscore_grade_fr	0.000000
energy_100g	15.339368
fat_100g	15.635354
saturated-fat_100g	17.146208
carbohydrates_100g	15.692342
sugars_100g	16.669759
fiber_100g	67.034661
proteins_100g	15.576378
salt_100g	17.040183

Méthode de remplacement des données

Meanvalue

sklearn.impute.IterativeImputer

KNeighborsClassifier

Estimation via règle de calcul

sklearn.impute.IterativeImputer

Multivariate imputer that estimates each feature from all the others.

A strategy for imputing missing values by modeling each feature with missing values as a function of other features in a round-robin fashion.

Remplissage des données

product_name	0.733560
ingredients_text	45.679708
additives_n	45.679487
nutriscore_score	23.008058
nutriscore_grade	23.008058
nova_group	50.429842
pnns_groups_1	0.000000
pnns_groups_2	0.000000
ecoscore_score_fr	0.000000
ecoscore_grade_fr	0.000000
energy_100g	15.339368
fat_100g	15.635354
saturated-fat_100g	17.146208
carbohydrates_100g	15.692342
sugars_100g	16.669759
fiber_100g	67.034661
proteins_100g	15.576378
salt_100g	17.040183

Méthode de remplacement des données

Meanvalue

sklearn.impute.IterativeImputer

KNeighborsClassifier

Algorithme des K plus proche voisins

Estimation via règle de calcul

sklearn.impute.IterativeImputer

Remplissage des données

product_name	0.733560
ingredients_text	45.679708
additives_n	45.679487
nutriscore_score	23.008058
nutriscore_grade	23.008058
nova_group	50.429842
pnns_groups_1	0.000000
pnns_groups_2	0.000000
ecoscore_score_fr	0.000000
ecoscore_grade_fr	0.000000
energy_100g	15.339368
fat_100g	15.635354
saturated-fat_100g	17.146208
carbohydrates_100g	15.692342
sugars_100g	16.669759
fiber_100g	67.034661
proteins_100g	15.576378
salt_100g	17.040183

Méthode de remplacement des données

Meanvalue

sklearn.impute.IterativeImputer

KNeighborsClassifier

Estimation via règle de calcul

sklearn.impute.IterativeImputer

Estimation basée sur les valeurs nutitrionnelles des aliments

Remplissage des données

Méthode de remplacement des données

product_name	0.733560
ingredients_text	45.679708
additives_n	45.679487
nutriscore_score	23.008058
nutriscore_grade	23.008058
nova_group	50.429842
pnns_groups_1	0.000000
pnns_groups_2	0.000000
ecoscore_score_fr	0.000000
ecoscore_grade_fr	0.000000
energy_100g	15.339368
fat_100g	15.635354
saturated-fat_100g	17.146208
carbohydrates_100g	15.692342
sugars_100g	16.669759
fiber_100g	67.034661
proteins_100g	15.576378
salt_100g	17.040183

	energy_100g	fat_100g	carbohydrates_100g
count	452724.000000	452724.000000	452724.000000
mean	1178.044098	16.107308	24.705077
std	755.527756	18.428249	25.436517
min	-440.726660	-1.792836	-70.347807
25%	565.000000	1.900000	2.900000
50%	1169.654285	14.120000	17.500000
75%	1610.000000	21.700000	41.000000
max	4012.088556	100.000000	100.000000

product_name	0.733560
ingredients_text	45.679708
additives_n	0.000000
nutriscore_score	0.000000
nutriscore_grade	23.008058
nova_group	0.000000
pnns_groups_1	0.000000
pnns_groups_2	0.000000
ecoscore_score_fr	0.000000
ecoscore_grade_fr	0.000000
energy_100g	0.000000
fat_100g	0.000000
saturated-fat_100g	0.000000
carbohydrates_100g	0.000000
sugars_100g	0.000000
fiber_100g	0.000000
proteins_100g	0.000000
salt_100g	0.000000

k-score et Méga-score

k_score = Σ sur les différentes catégories '...100g' (energy, fat, etc...)
par rapport aux apports journaliers recommandés (fat_100g - 35)/10
bonus / malus (+ saturated_fat / - fiber)

Méga_score = Σ sur les différents autres score (nutriscore_score / nova_group etc)

	k_score	mega_score
count	452724.000000	452724.000000
mean	-6.531733	5.936741
std	5.747572	32.093599
min	-24.400212	-118.960048
25%	-11.515675	-16.507641
50%	-6.585110	9.528842
75%	-2.793273	29.192871
max	14.847720	99.467778

L'application

Scan du code bare

```
Broccoli- Röshen
----
mega_score          -64.675226
nutriscore_score     -7.000000
nova_group            1.000000
ecoscore_score_fr    83.000000
additives_n           0.000000
k_score             -15.324774
Name: 4421, dtype: float64
----
ingrédients :
Brokkoliröschen
------------------------------
------------------------------
suggestion d'autres produit avec un meilleur score :
```

```
produit N° 1  :
----
Ajo morado
----
mega_score          -70.757388
nutriscore_score      7.678495
nova_group            1.000000
ecoscore_score_fr    95.000000
additives_n           0.000000
k_score              -6.564118
Name: 411503, dtype: float64
------------------------------
----
produit N° 2  :
----
Ajo negro
----
mega_score          -84.181571
nutriscore_score     -8.000000
nova_group            1.000000
ecoscore_score_fr    95.000000
additives_n           0.000000
k_score              -8.818429
Name: 429194, dtype: float64
------------------------------
----
produit N° 3  :
----
Annalisa Canned Tomatoes
----
mega_score          -76.611394
nutriscore_score     -5.000000
nova_group            1.000000
ecoscore_score_fr    97.000000
additives_n           0.000000
k_score             -15.388606
Name: 378609, dtype: float64
```

```
produit N° 1  :
----
Betteraves Rouges
----
mega_score          -83.441455
nutriscore_score     -7.000000
nova_group            1.000000
ecoscore_score_fr   101.000000
additives_n           1.000000
k_score             -13.558545
Name: 94016, dtype: float64
------------------------------
----
produit N° 2  :
----
Chopped Tomatoes
----
mega_score          -67.670297
nutriscore_score     -5.000000
nova_group            1.000000
ecoscore_score_fr    89.000000
additives_n           1.000000
k_score             -15.329703
Name: 327748, dtype: float64
------------------------------
----
produit N° 3  :
----
Bio Zwiebeln
----
mega_score          -92.757388
nutriscore_score      7.678495
nova_group            1.000000
ecoscore_score_fr   117.000000
additives_n           0.000000
k_score              -6.564118
Name: 301555, dtype: float64
```

L'application

2 autres examples

Lentilles cuisinées aux carottes et oignons

```
----
mega_score          -101.058968
nutriscore_score      -4.000000
nova_group             1.000000
ecoscore_score_fr    123.000000
additives_n            2.000000
k_score              -13.941032
Name: 125260, dtype: float64
----
ingrédients :
ingrédients à renseigner
------------------------------
------------------------------
suggestion d'autres produit avec un meilleur score :
produit N° 1  :
----
```

Couscous biologique de blé khorazan

```
----
mega_score          -114.430635
nutriscore_score      -5.000000
nova_group             1.000000
ecoscore_score_fr    123.000000
additives_n            0.000000
k_score               -3.569365
Name: 183210, dtype: float64
------------------------------
----
pas d'autre meilleur prodruit
pas d'autre meilleur prodruit
```

Vinagre balsámico de Módena

```
----
mega_score          -14.367607
nutriscore_score      8.229934
nova_group            4.000000
ecoscore_score_fr    71.000000
additives_n           2.000000
k_score              -6.402460
Name: 425260, dtype: float64
----
ingrédients :
ingrédients à renseigner
------------------------------
------------------------------
suggestion d'autres produit avec un meilleur score :
produit N° 1  :
----
Peppery Moroccan Ketchup
----
mega_score          -24.90817
nutriscore_score      7.00000
nova_group            3.00000
ecoscore_score_fr    74.00000
additives_n           0.00000
k_score             -12.09183
Name: 43227, dtype: float64
```

```
produit N° 2  :
----
Knoblauch Sauce
----
mega_score          -27.831084
nutriscore_score      7.000000
nova_group            4.000000
ecoscore_score_fr    83.000000
additives_n           2.000000
k_score              -6.168916
Name: 314984, dtype: float64
------------------------------
----
produit N° 3  :
----
Puré de tomates homogeneizado
----
mega_score          -53.423089
nutriscore_score      8.103689
nova_group            1.000000
ecoscore_score_fr    78.000000
additives_n           0.000000
k_score              -6.473222
Name: 371904, dtype: float64
```

Conclusion

- Analyses univariées
- Analyses multivariées
- Prémisses d'une application

Limitation de l'étude

- Beaucoup de valeurs manquantes
- Encore des valeurs aberrantes
- Limitations métier sur le k et Méga Score
- défauts techniques

Perspectives

- Une applications avec choix pour l'utilisateur
- De nouveaux scores

Fin de la présentation

Merci de m’avoir écouté !

Remplissage des données

Méthode de remplacement des données

sklearn.impute.IterativeImputer

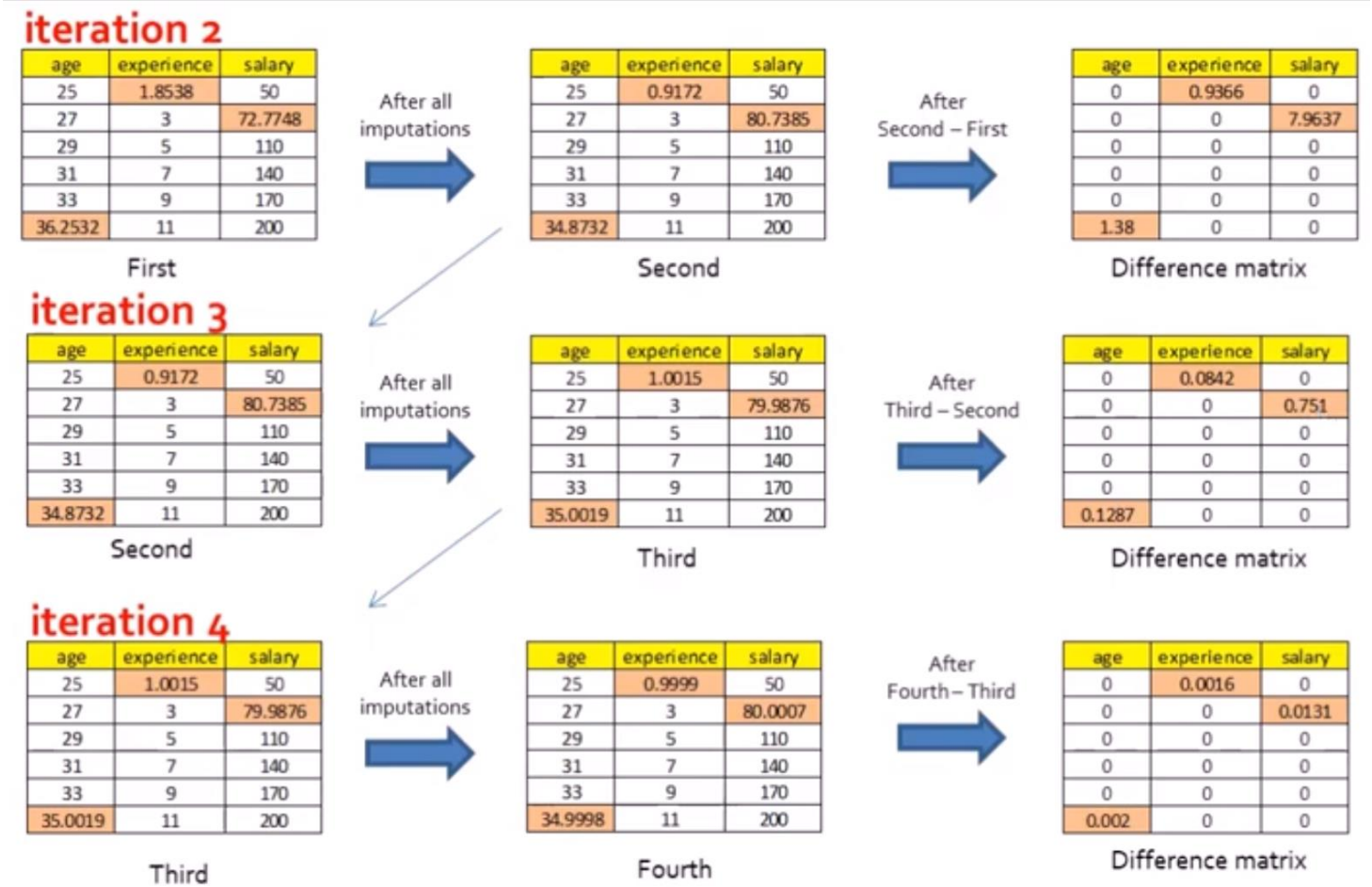
La supposition de base et que les données manquantes peuvent être expliqué par les autres colonnes et qu'on peut faire une estimation éclairée

Étape 1 : on fill tout avec mean de la colonne.

Étape 2 : remove les valeurs de la colonnes de gauche et ainsi de suite jusqu'à la colonne de droite (3 étatpes pour 3 colonnes)

Étape 3 : On soustrait les 2 « derniers » datasets

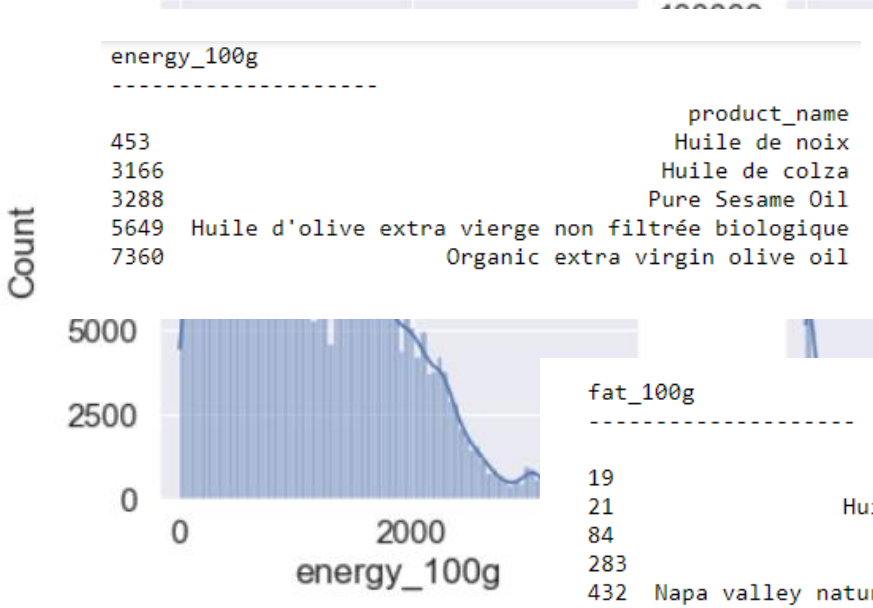
Etc etc… On s'arrête après 10 itérations



Analyses univariées « …100g »

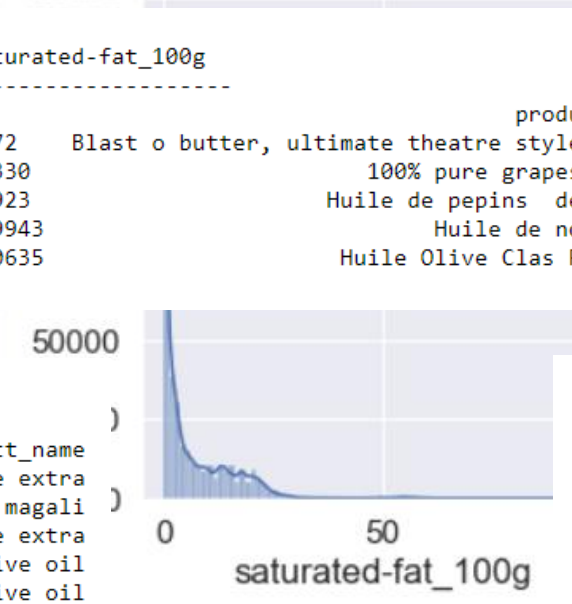
energy_100g

	product_name
453	Huile de noix
3166	Huile de colza
3288	Pure Sesame Oil
5649	Huile d'olive extra vierge non filtrée biologique
7360	Organic extra virgin olive oil



saturated-fat_100g

	product_name
9172	Blast o butter, ultimate theatre style butter
68330	100% pure grapeseed oil
78923	Huile de pepins de raisin
139943	Huile de noisettes
180635	Huile Olive Clas Puget 1L

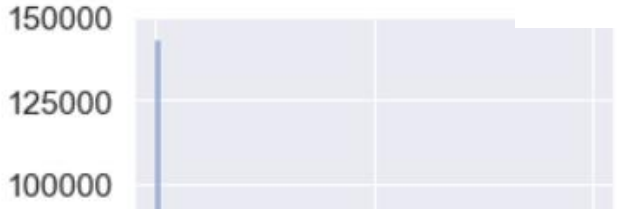


fat_100g

	product_name
19	Huile d'olive de Nîmes vierge extra
21	Huile d'olive vierge extra cuve magali
84	Huile d'olive vierge extra
283	Trader giotto's, olive oil
432	Napa valley naturals, organic extra virgin olive oil

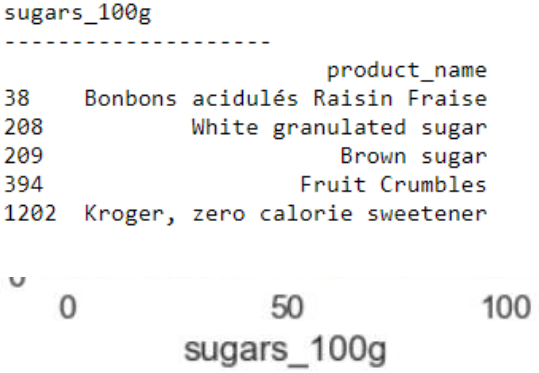
carbohydrates_100g

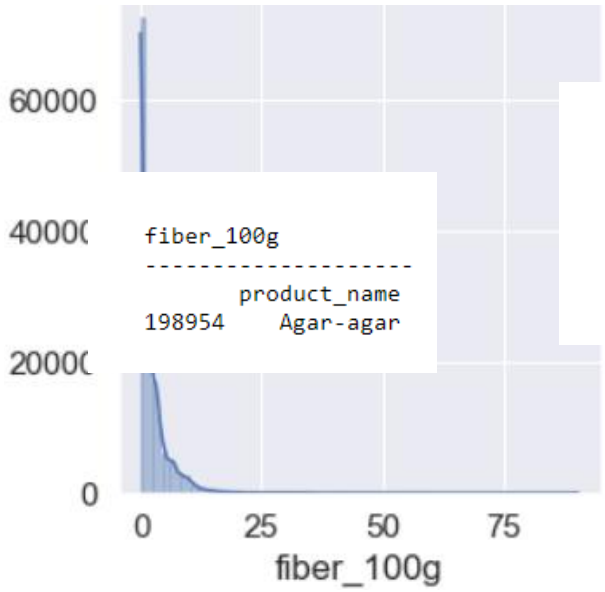
	product_name
37	Compote de Pomme
38	Bonbons acidulés Raisin Fraise
56	Pastille Vichy
102	Lagg's, bronchtea
165	Dextrose & Maltodextrine mixed berry



sugars_100g

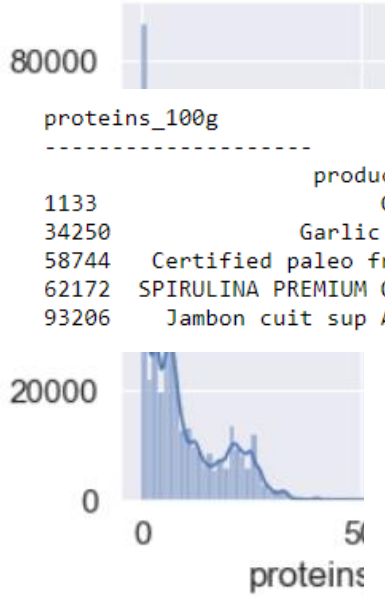
	product_name
38	Bonbons acidulés Raisin Fraise
208	White granulated sugar
209	Brown sugar
394	Fruit Crumbles
1202	Kroger, zero calorie sweetener





fiber_100g

	product_name
198954	Agar-agar



proteins_100g

	product_name
1133	Gelatin
34250	Garlic Powder
58744	Certified paleo friendly
62172	SPIRULINA PREMIUM QUALITY
93206	Jambon cuit sup AC fumé

salt_100g

	product_name
3511	Salt sense iodized
8196	Regular salt
8197	Iodized salt
8199	Iodized Salt
14877	Roland, fleur de sel, hand-harvested french sea salt

k-score et Méga-score

k_score = (energy_100 - 2403.5)/378 + ('fat_100g' - 35)/10 + ('saturated-fat_100g)/10 + (carbohydrates_100g - 53)/10 + (sugars_100g)/10 + salt_100g/10 + ('proteins_100g' - 12)/10 - fiber_100g/10

Méga_score = nutriscore_score + nova_group*10 + 'additives_n' - écoscore_score_fr - k_score

k-score et Méga-score

		k_score		mega_score	
		mean	std	mean	std
pnns_groups_1	pnns_groups_2				
Alcoholic beverages	Alcoholic beverages	-7.471919	2.473715	6.695231	13.553697
Beverages	Artificially sweetened beverages	-14.224208	3.300956	11.595462	19.434894
	Fruit juices	-13.034468	2.379500	-12.588925	14.911204
	Fruit nectars	-13.148338	2.351942	15.309572	17.265314
	Plant-based milk substitutes	-13.331134	2.797635	-0.713888	31.139878
	Sweetened beverages	-9.016513	7.004630	-1.668555	25.950041
	Teas and herbal teas and coffees	-13.930828	3.656838	-23.710676	21.383530
	Unsweetened beverages	-12.262877	4.024396	-27.722421	28.418722
Cereals and potatoes	Bread	-5.460284	2.736713	-20.101231	20.543182
	Breakfast cereals	-2.694466	2.551121	-13.707322	25.512619
	Cereals	-4.434512	2.972469	-40.348471	27.453369
	Legumes	-9.633180	5.043828	-30.541920	26.920877
	Potatoes	-10.293894	3.146938	-30.952647	19.316018
Composite foods	One-dish meals	-10.664973	3.378885	10.472520	29.435423
	Pizza pies and quiches	-7.326599	3.126628	13.026982	18.760789
	Sandwiches	-8.303774	1.597796	26.665906	22.498428
Fat and sauces	Dressings and sauces	-8.965931	4.700407	-2.467008	26.640280
	Fats	2.537131	4.217577	0.489337	17.223873
Fish Meat Eggs	Eggs	-10.375892	3.036371	-15.792344	17.155765
	Fish and seafood	-10.267857	2.267553	28.319594	21.621744
	Meat	-9.709818	2.674299	28.852349	19.030149
	Offals	-9.197778	1.815217	39.810246	17.898896
	Processed meat	-8.770102	2.970108	52.239887	15.285763
Fruits and vegetables	Dried fruits	-2.338639	3.323005	3.927407	23.729642
	Fruits	-9.221205	5.048805	-34.851242	27.641035
	Soups	-13.980892	2.791174	-31.092377	13.157553
	Vegetables	-12.978325	3.791422	-34.019588	27.392035
Milk and dairy products	Cheese	-5.963484	2.454805	20.795047	12.172390
	Dairy desserts	-10.470221	3.069992	-0.716787	24.612471
	Ice cream	-7.041939	2.651016	2.760947	11.875592
	Milk and yogurt	-12.056870	3.023035	-1.919599	22.294138
Salty snacks	Appetizers	-2.878896	3.280413	-5.929949	18.431676
	Nuts	-1.097395	3.082015	13.828993	20.384899
	Salty and fatty products	-7.289247	4.102381	8.693991	30.565605
Sugary snacks	Biscuits and cakes	-0.070130	3.395985	20.943153	17.956378
	Chocolate products	3.549001	3.178363	46.828448	13.245075
	Pastries	-3.283363	2.439530	20.143344	12.532068
	Sweets	-1.779095	5.277827	7.743143	30.659600
unknown	unknown	-8.368471	5.427680	-15.748180	32.993252

Backup slides

Légende des données (bdd complète):

% de données manquantes	
code	0.000000
url	0.000000
creator	0.000209
created_t	0.000000
created_datetime	0.000000
last_modified_t	0.000000
last_modified_datetime	0.000000
product_name	4.169265
abbreviated_product_name	99.633360
generic_name	93.772638
quantity	75.140807
packaging	84.390594
packaging_tags	84.392262
packaging_text	99.704884
brands	48.403787
stores	87.459259
countries	0.307002
countries_tags	0.307211
countries_en	0.307211
ingredients_text	63.014146
allergens	90.224103
allergens_en	100.000000
traces	94.810210
traces_tags	93.564425
traces_en	93.564425
serving_size	74.368271
serving_quantity	74.603436
no_nutriments	100.000000
additives_n	63.014094
additives	100.000000
additives_tags	78.625534
additives_en	78.625534
ingredients_from_palm_oil_n	63.014094
ingredients_from_palm_oil	100.000000
ingredients_from_palm_oil_tags	99.252748
ingredients_that_may_be_from_palm_oil_n	63.014094
ingredients_that_may_be_from_palm_oil	100.000000
ingredients_that_may_be_from_palm_oil_tags	97.796821
nutriscore_score	63.431667
nutriscore_grade	63.431667
nova_group	67.690389
pnns_groups_1	0.026274
pnns_groups_2	0.026170
states	0.000000
states_tags	0.000000
states_en	0.000000
brand_owner	84.890117
ecoscore_score_fr	75.871116
ecoscore_grade_fr	75.871116
main_category	52.830998
main_category_en	52.830998
image_url	23.825948
image_small_url	23.825948
image_ingredients_url	63.753944
image_ingredients_small_url	63.753944
image_nutrition_url	48.846644
image_nutrition_small_url	48.846644
energy-kj_100g	92.815609
energy-kcal_100g	23.275492
energy_100g	20.563353
energy-from-fat_100g	99.949328
fat_100g	20.976703
saturated-fat_100g	23.210275
-butyric-acid_100g	99.998332
-caproic-acid_100g	99.999687
-caprylic-acid_100g	99.999635
-capric-acid_100g	99.999427
-lauric-acid_100g	99.998853
-myristic-acid_100g	99.999166
-palmitic-acid_100g	99.998280
-stearic-acid_100g	99.999166
-elaidic-acid_100g	99.999896
-gondoic-acid_100g	99.997654
-mead-acid_100g	99.999114
-erucic-acid_100g	99.999531
-nervonic-acid_100g	99.999218
trans-fat_100g	86.194133
cholesterol_100g	85.987849
carbohydrates_100g	20.998286
sugars_100g	21.824621
-sucrose_100g	99.991398
-glucose_100g	99.994891
-fructose_100g	99.995829
-lactose_100g	99.954437
-maltose_100g	99.996403
-maltodextrins_100g	99.991033
starch_100g	99.972683
polyols_100g	99.808678
fiber_100g	74.775887
-soluble-fiber_100g	99.811805
-insoluble-fiber_100g	99.825933
proteins_100g	20.913729
casein_100g	99.997081
serum-proteins_100g	99.996559
nucleotides_100g	99.997706
salt_100g	24.773228
sodium_100g	24.773385
alcohol_100g	98.858686
vitamin-a_100g	88.919044
beta-carotene_100g	99.995725
vitamin-d_100g	99.483325
vitamin-e_100g	99.813057
vitamin-k_100g	99.935826
vitamin-c_100g	88.516537
vitamin-b1_100g	98.762869
vitamin-b2_100g	98.811924
vitamin-pp_100g	98.756665
vitamin-b6_100g	99.160736
vitamin-b9_100g	99.460700
folates_100g	99.558134
caffeine_100g	99.975029
taurine_100g	99.989574
ph_100g	99.989522
fruits-vegetables-nuts_100g	99.541608
fruits-vegetables-nuts-dried_100g	99.980503
fruits-vegetables-nuts-estimate_100g	99.383285
collagen-meat-protein-ratio_100g	99.983631
cocoa_100g	99.680121
chlorophyl_100g	99.999739
carbon-footprint_100g	99.975185
carbon-footprint-from-meat-or-fish_100g	99.387925
nutrition-score-fr_100g	63.431459
nutrition-score-uk_100g	99.999583
glycemic-index_100g	99.999791
water-hardness_100g	99.999948
choline_100g	99.996768
phylloquinone_100g	99.910751
beta-glucan_100g	99.998071
inositol_100g	99.996455
carnitine_100g	99.997967
brands_tags	48.407124
categories	52.830685
categories_tags	52.830998
categories_en	52.830998
origins	95.726942
origins_tags	95.733719
origins_en	95.733719
manufacturing_places	93.698664
manufacturing_places_tags	93.701531
labels	78.623344
labels_tags	78.623761
labels_en	78.623761
emb_codes	93.806784
emb_codes_tags	93.808974
first_packaging_code_geo	96.243356
cities	100.000000
cities_tags	95.952097
purchase_places	91.864106
-arachidic-acid_100g	99.994370
-behenic-acid_100g	99.997863
-lignoceric-acid_100g	99.999270
-cerotic-acid_100g	99.999166
-montanic-acid_100g	99.998853
-melissic-acid_100g	99.999114
monounsaturated-fat_100g	97.503842
polyunsaturated-fat_100g	97.505771
omega-3-fat_100g	99.893235
-alpha-linolenic-acid_100g	99.960432
-eicosapentaenoic-acid_100g	99.993588
-docosahexaenoic-acid_100g	99.988479
omega-6-fat_100g	99.971745
-linoleic-acid_100g	99.974925
-arachidonic-acid_100g	99.993379
-gamma-linolenic-acid_100g	99.999114
-dihomo-gamma-linolenic-acid_100g	99.999374
omega-9-fat_100g	99.994318
-oleic-acid_100g	99.996507
vitamin-b12_100g	99.327661
biotin_100g	99.937703
pantothenic-acid_100g	99.677828
silica_100g	99.991920
bicarbonate_100g	99.977792
potassium_100g	95.195097
chloride_100g	99.958347
calcium_100g	85.893908
phosphorus_100g	99.274070
iron_100g	86.184072
magnesium_100g	99.227829
zinc_100g	99.464245
copper_100g	99.774219
manganese_100g	99.786574
fluoride_100g	99.979617
selenium_100g	99.867065
chromium_100g	99.989052
molybdenum_100g	99.983735
iodine_100g	99.880515

Backup slides

Quelles sont les modalités de calcul du nutri score ?
Le Nutri Score est calculé en fonction de plusieurs composantes :
•le nombre de calories contenues dans le produit fini,
•la quantité d'acides gras,
•la dose de sucres,
•la proportion de fibres…
Pour l'attribution des points d'un aliment, il faut évaluer le rapport entre ses composants nocifs (les calories, le sel, les sucres…) et ses véritables apports nutritionnels (fruits, légumes, fibres, protéines).