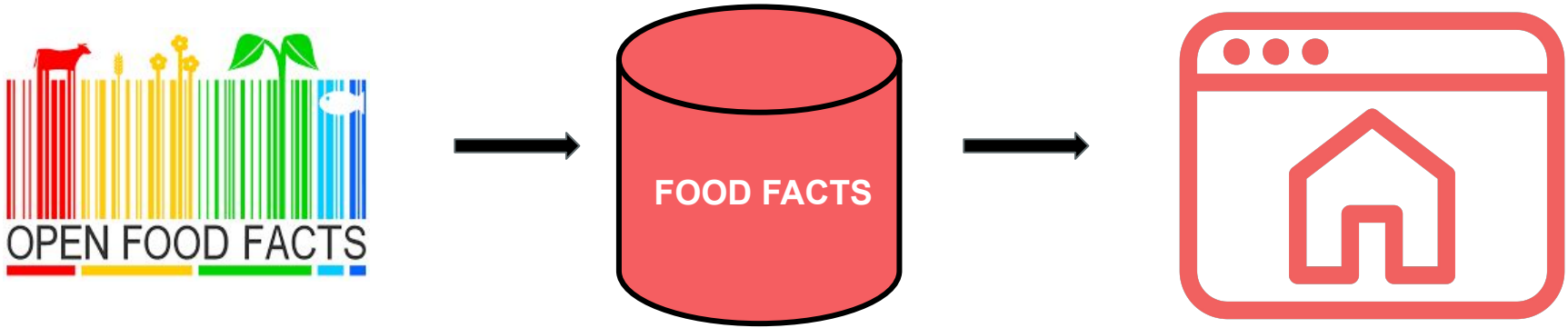
Analyse de données de santé publique

P3 - Sylwia Bankowska - Janvier 2022

Sommaire

1. Le projet

Analyser les données de santé publique



- Exploration
- Identification, quantification et traitement des failles

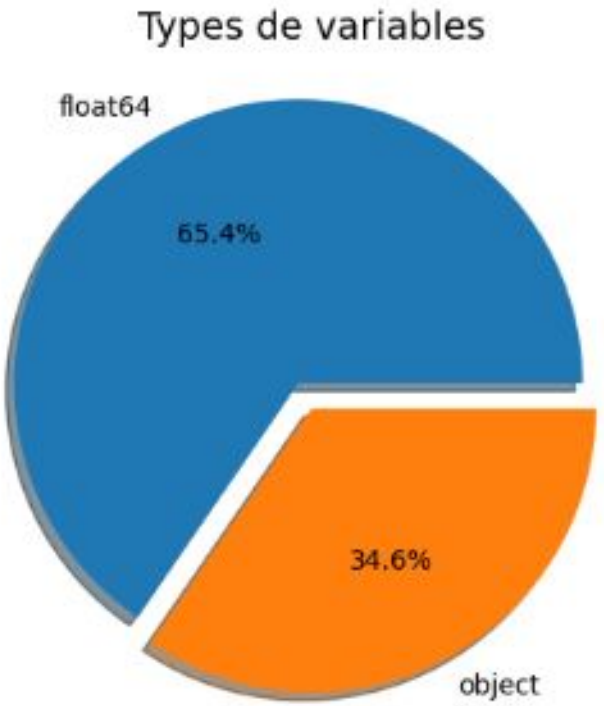
- Variables pertinentes
- Analyses statistiques

- Prototype

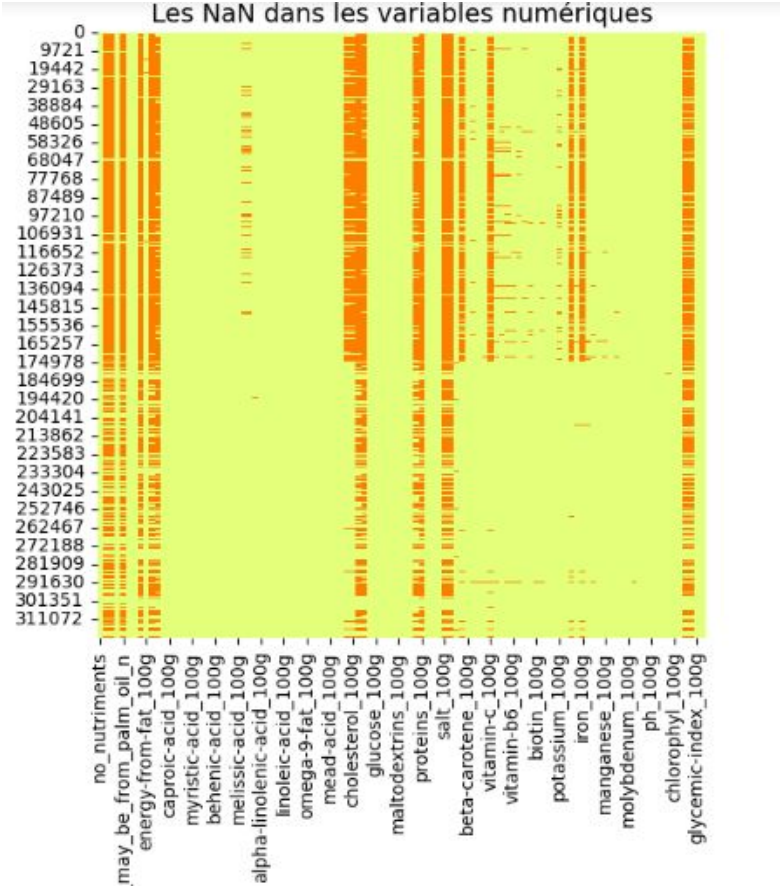
2. Les données

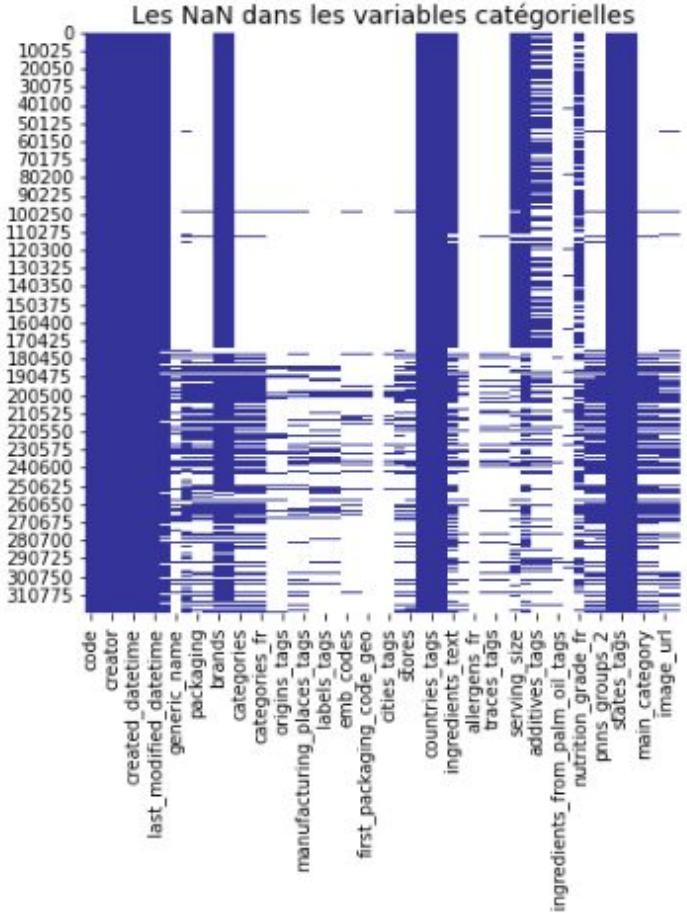
Présentation

- 320 772 lignes
- 162 colonnes
- 39 608 589 valeurs manquantes



Valeurs manquantes





Nettoyage

01	**Valeurs vides**	• 99 502 Nutri-Scores non renseignés • 2746 noms de produits non renseignés ○ Suppression de lignes
02	**Valeurs erronées**	• *Auchan , Picard, Géant, Leader Price* • *Sony, Blu Ray, Savarez* ○ Suppression de lignes
03	**Valeurs mal formatées**	• modalité “unknown” remplacées par np.nan • marque: P$$T - replace par Psst • ”Salty snacks', 'sugary-snacks', ‘Sugary snacks', 'salty-snacks’ ○ replace “ - “ par “ “ et capitalisation

04	**Valeurs dupliquées**	• ‘brands’, brands_tag’, ‘countries’, ‘countries_tag’, ‘countries_fr’
		  ○ Suppression de colonnes
		• Noms de produits
		  ○ Conservation des données les plus récentes.

05	**Valeurs non pertinentes**	• Colonnes à plus de 50% de valeurs de type NaN :
		  ○ 92 variables numériques
		  ○ 36 variables catégorielles
		  ■ Suppression

Valeurs aberrantes / atypiques

		Valeur aberrante	Valeur atypique	Caractéristiques
1	Eau	✗	✓	• Nutri Score : E • Teneur en fibres : 100g • il s'agit d'une eau révitalisante, un jus de détox
2	Huile , huile d'olive	✗	✓	• plus de 3700 kj d'énergie sur 100g d'aliment
3	Sel	✗	✓	• 100g de sel dans … du sel
4	Fruits	✗	✓	• Classés en Nutri Score E ! • Il s'agit de fruits confits ou bonbons fourrés cerise et liqueur
5	"Legumes"	✗	✓	• Classés en Nutri Score E ! • Il s'agit de pâtes à tartiner / beurres de cachuète

Outliers

Ne doivent pas être négatives.

Trop élevées ! A traiter au cas par cas.

Valeurs minimales

Valeurs maximales

	count	mean	std	min	25%	50%	75%	max
additives_n	248939.0	1.936024	2.502019	0.00	0.0000	1.00000	3.00000	31.00
energy_100g	261113.0	1141.914605	6447.154093	0.00	377.0000	1100.00000	1674.00000	3251373.00
fat_100g	243891.0	12.730379	17.578747	0.00	0.0000	5.00000	20.00000	714.29
saturated-fat_100g	229554.0	5.129932	8.014238	0.00	0.0000	1.79000	7.14000	550.00
carbohydrates_100g	243588.0	32.073981	29.731719	0.00	6.0000	20.60000	58.33000	2916.67
sugars_100g	244971.0	16.003484	22.327284	-17.86	1.3000	5.71000	24.00000	3520.00
fiber_100g	200886.0	2.862111	12.867578	-6.70	0.0000	1.50000	3.60000	5380.00
proteins_100g	259922.0	7.075940	8.409054	-800.00	0.7000	4.76000	10.00000	430.00
salt_100g	255510.0	2.028624	128.269454	0.00	0.0635	0.58166	1.37414	64312.80
sodium_100g	255463.0	0.798815	50.504428	0.00	0.0250	0.22900	0.54100	25320.00
nutrition-score-fr_100g	221210.0	9.165535	9.055903	-15.00	1.0000	10.00000	16.00000	40.00

Energie

La valeur maximale ne devrait pas dépasser les 3700 kj.

Score de nutrition

Valeurs entre -15 et +40

100g

Les features à 100g ne doivent pas dépasser … 100g.

Traitement des outliers

Méthode IQR

```
 - Les valeurs limites minimales :

 fat_100g              -30.00
saturated-fat_100g    -10.71
sugars_100g           -32.75
fiber_100g             -5.40
proteins_100g         -13.25
salt_100g              -1.90
sodium_100g            -0.75
carbohydrates_100g    -72.50
dtype: float64

- Les valeurs limites maximales :

 fat_100g               50.00
saturated-fat_100g     17.85
sugars_100g            58.05
fiber_100g              9.00
proteins_100g          23.95
salt_100g               3.34
sodium_100g             1.32
carbohydrates_100g    136.82
dtype: float64
```

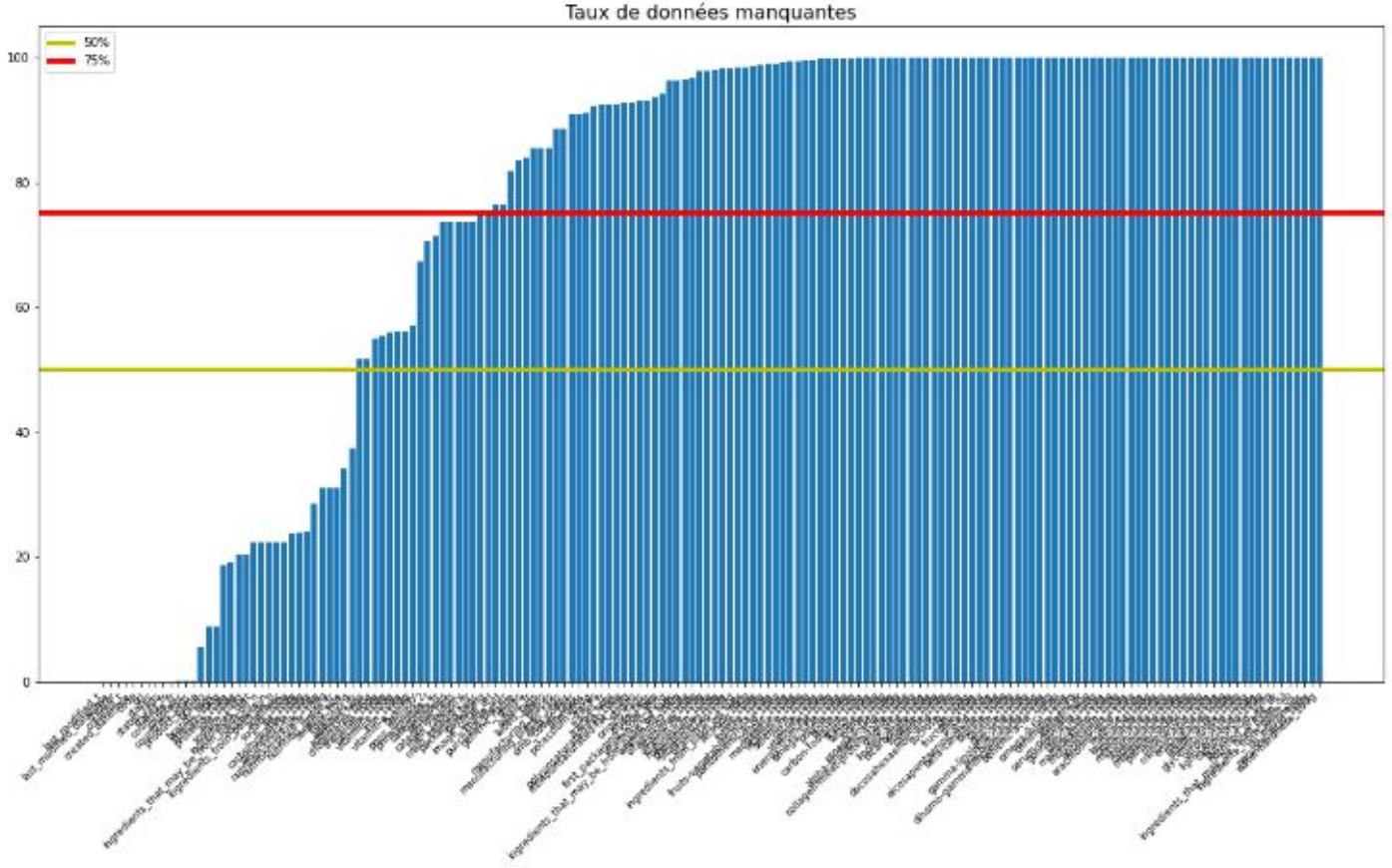
Fonctions Python

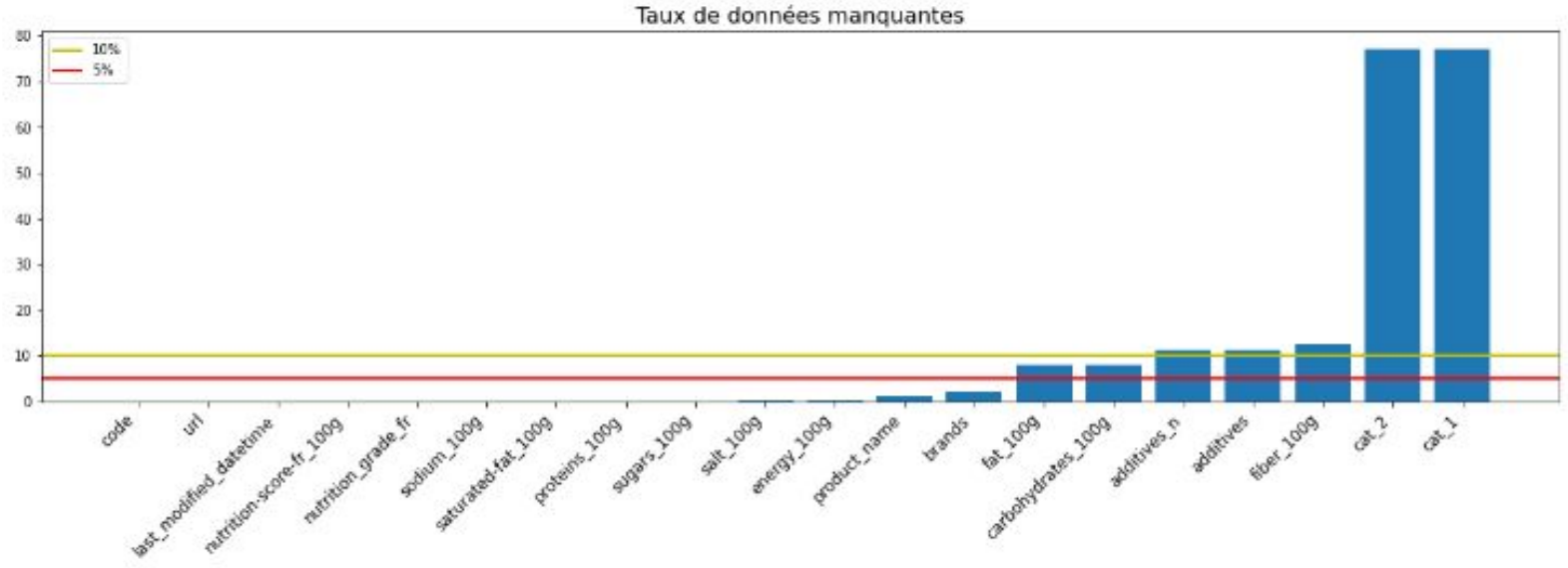
```
Mise à jour de la variable => " fat_100g " :
- "714.29" est une donnée invalide et sera remplacée par "NaN".
- "380.0" est une donnée invalide et sera remplacée par "NaN".
- "101.0" est une donnée invalide et sera remplacée par "NaN".
- "105.0" est une donnée invalide et sera remplacée par "NaN".
Mise à jour de la variable => " saturated-fat_100g " :
- "175.38" est une donnée invalide et sera remplacée par "NaN".
- "210.0" est une donnée invalide et sera remplacée par "NaN".
- "550.0" est une donnée invalide et sera remplacée par "NaN".
Mise à jour de la variable => " sugars_100g " :
- "-1.2" est une donnée invalide et sera remplacée par "NaN".
- "-0.8" est une donnée invalide et sera remplacée par "NaN".
- "134.0" est une donnée invalide et sera remplacée par "NaN".
- "-3.57" est une donnée invalide et sera remplacée par "NaN".
- "110.71" est une donnée invalide et sera remplacée par "NaN".
- "-6.67" est une donnée invalide et sera remplacée par "NaN".
- "-6.25" est une donnée invalide et sera remplacée par "NaN".
- "166.67" est une donnée invalide et sera remplacée par "NaN".
- "-17.86" est une donnée invalide et sera remplacée par "NaN".
```

Indicateurs statistiques

	count	mean	std	min	25%	50%	75%	max
additives_n	221116.0	1.91	2.40	0.0	0.00	1.00	3.00	31.0
energy_100g	221116.0	1175.79	760.66	0.0	450.90	1191.00	1715.00	3887.0
fat_100g	221116.0	13.15	15.69	0.0	1.17	7.14	21.43	100.0
saturated-fat_100g	221116.0	4.96	7.56	0.0	0.00	1.79	7.14	100.0
carbohydrates_100g	221116.0	32.93	27.50	0.0	7.96	23.80	57.14	100.0
sugars_100g	221116.0	15.00	19.81	0.0	1.30	5.00	23.06	100.0
fiber_100g	221116.0	2.58	4.22	0.0	0.00	1.20	3.30	100.0
proteins_100g	221116.0	7.77	8.06	0.0	1.90	5.70	10.71	100.0
salt_100g	221116.0	1.23	3.94	0.0	0.10	0.65	1.36	100.0
sodium_100g	221116.0	0.49	1.67	0.0	0.04	0.25	0.54	92.5
nutrition-score-fr_100g	221116.0	9.16	9.06	-15.0	1.00	10.00	16.00	40.0

Etat des lieux





320 772 — 160 487

Lignes avant — **Lignes après**

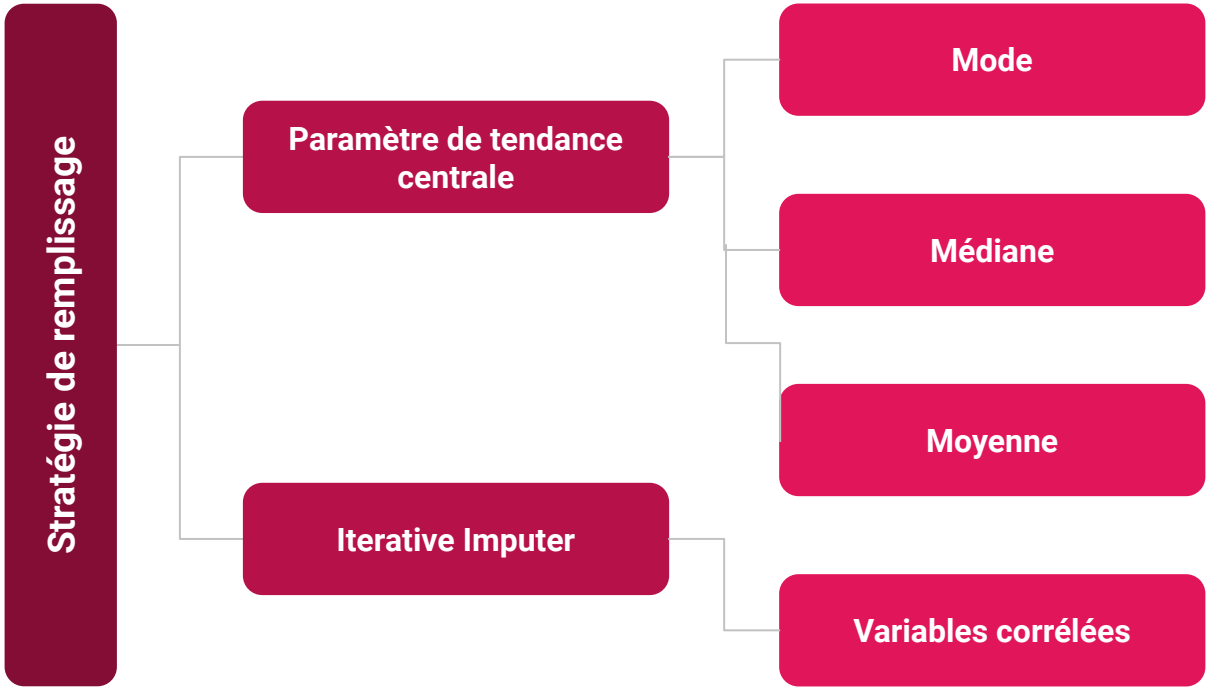
160 — 20

Colonnes avant — **Colonnes après**

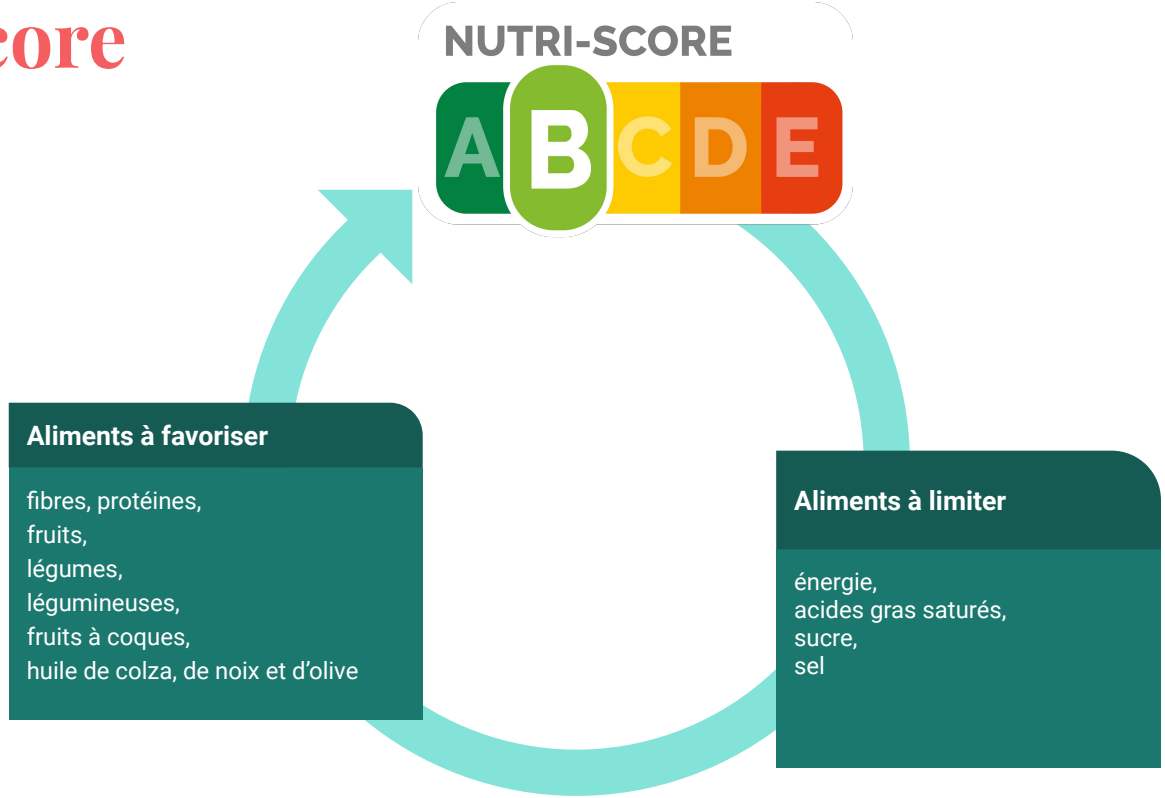
- Suppressions de doublons
- Suppressions de lignes vides

Suppression de variables => vides, peu renseignées, non pertinentes

Valeurs manquantes



Nutri-Score



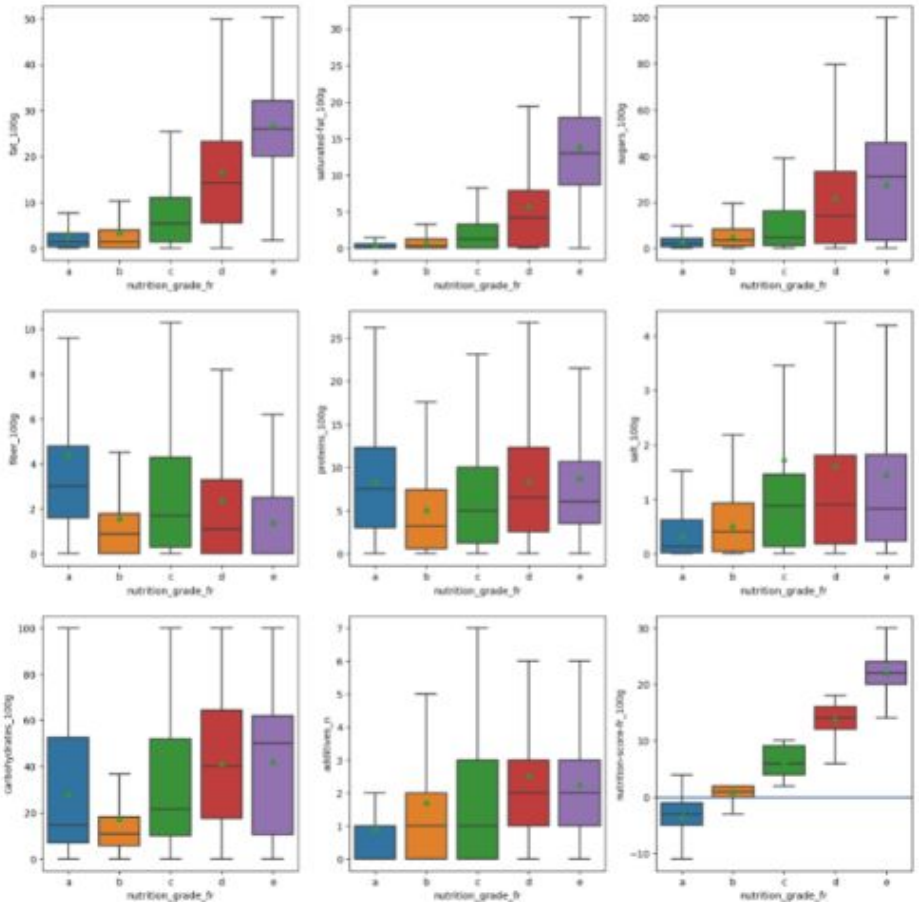
Objectif : remplir les NaN des substances nutritives
par la médiane des individus ayant le même score de nutrition.

Résultats

	Valeurs manquantes	Valeurs nulles
cat_1	119190	0.0
cat_2	119101	0.0
additives	19979	0.0
brands	2026	0.0
product_name	1	0.0
sugars_100g	0	0.0
nutrition-score-fr_100g	0	0.0
sodium_100g	0	0.0
salt_100g	0	0.0
proteins_100g	0	0.0
fiber_100g	0	0.0
code	0	0.0
carbohydrates_100g	0	0.0
url	0	0.0
fat_100g	0	0.0
energy_100g	0	0.0
nutrition_grade_fr	0	0.0
additives_n	0	0.0
last_modified_datetime	0	0.0
saturated-fat_100g	0	0.0

	count	mean	std	min	25%	50%	75%	max
additives_n	221116.0	1.91	2.40	0.0	0.00	1.00	3.00	31.0
energy_100g	221116.0	1175.79	760.66	0.0	450.90	1191.00	1715.00	3887.0
fat_100g	221116.0	13.15	15.69	0.0	1.17	7.14	21.43	100.0
saturated-fat_100g	221116.0	4.96	7.56	0.0	0.00	1.79	7.14	100.0
carbohydrates_100g	221116.0	32.93	27.50	0.0	7.96	23.80	57.14	100.0
sugars_100g	221116.0	15.00	19.81	0.0	1.30	5.00	23.06	100.0
fiber_100g	221116.0	2.58	4.22	0.0	0.00	1.20	3.30	100.0
proteins_100g	221116.0	7.77	8.06	0.0	1.90	5.70	10.71	100.0
salt_100g	221116.0	1.23	3.94	0.0	0.10	0.65	1.36	100.0
sodium_100g	221116.0	0.49	1.67	0.0	0.04	0.25	0.54	92.5
nutrition-score-fr_100g	221116.0	9.16	9.06	-15.0	1.00	10.00	16.00	40.0

Distribution de variables par Nutri-Score

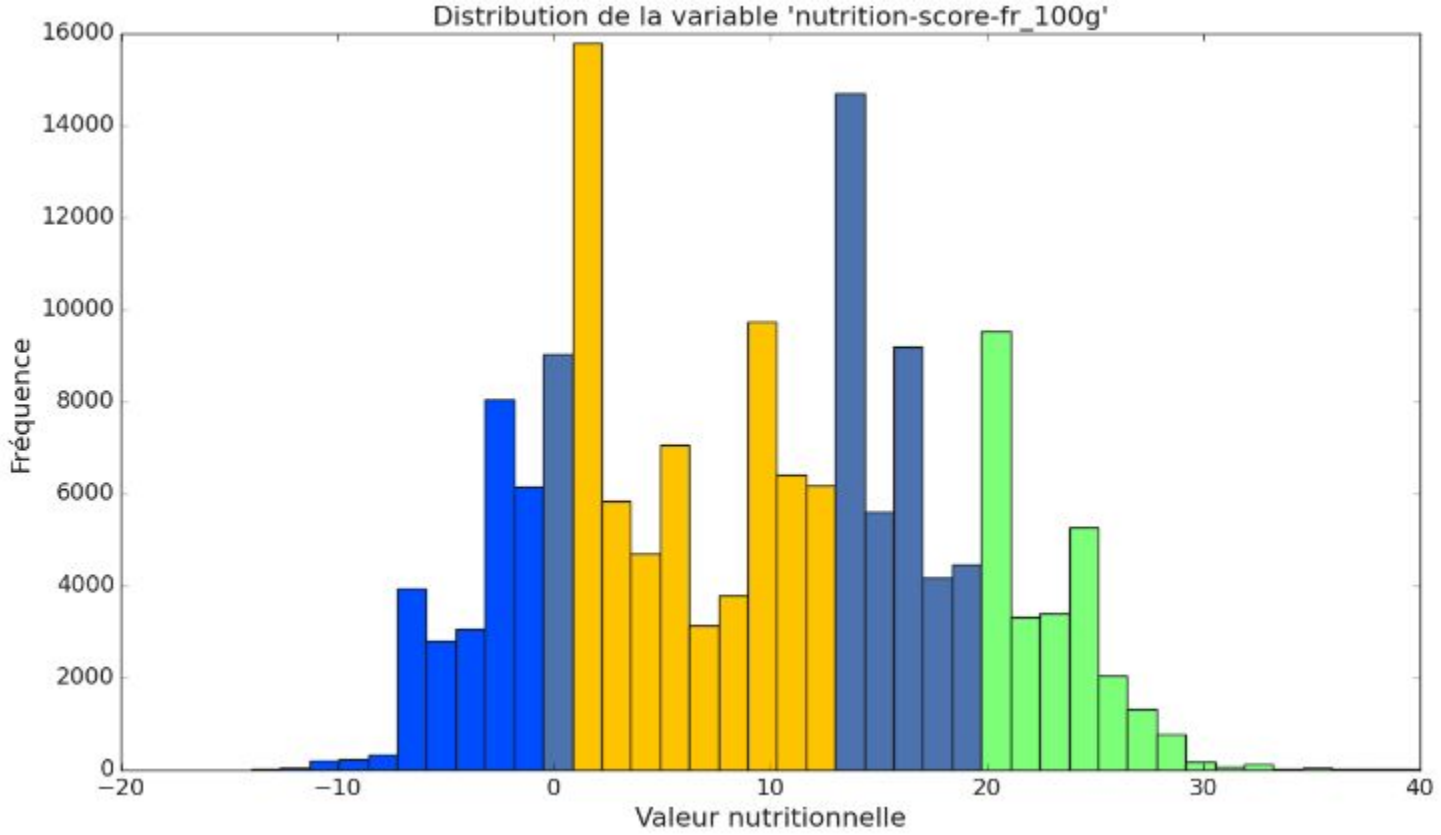


3. L’analyse statistique

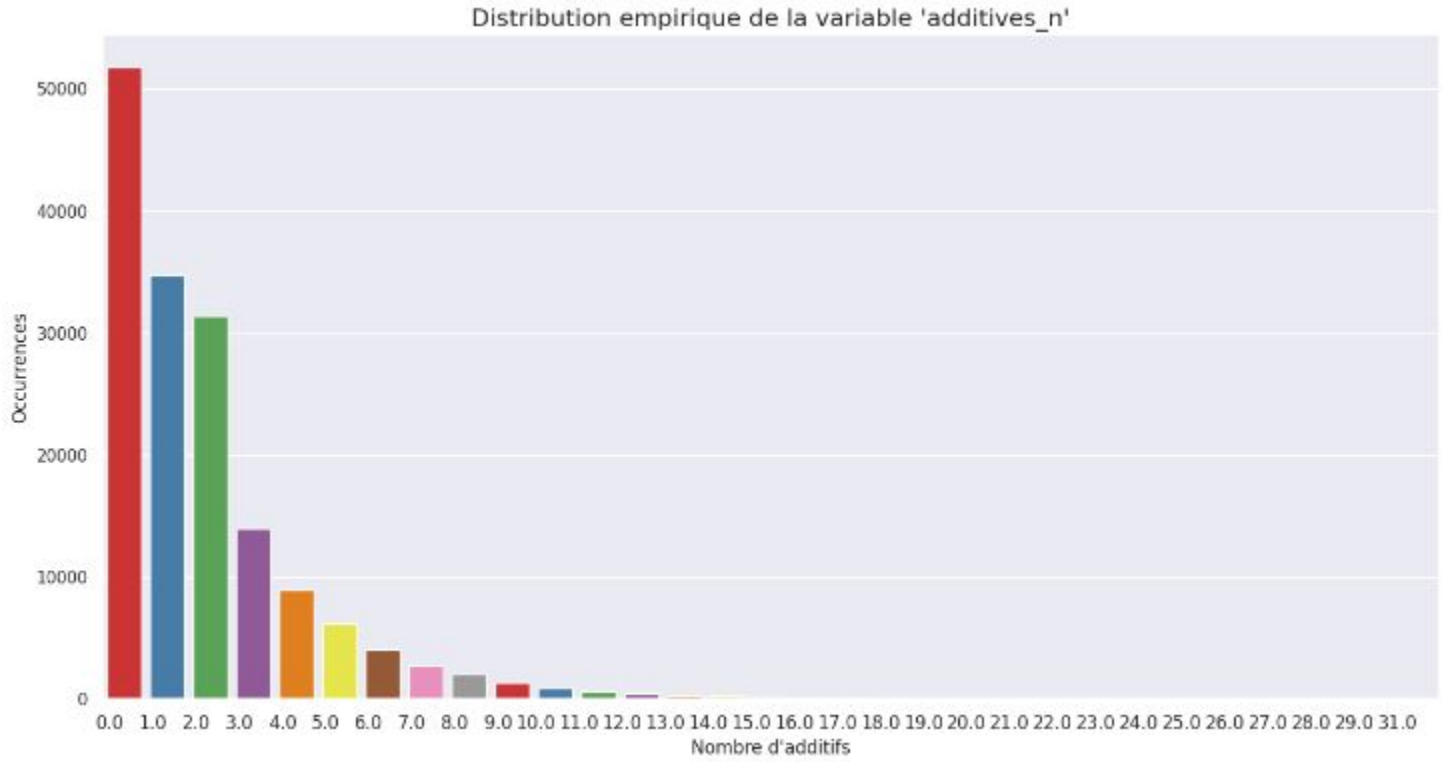
Analyse univariée

Variables numériques

Distribution des valeurs nutritionnelles



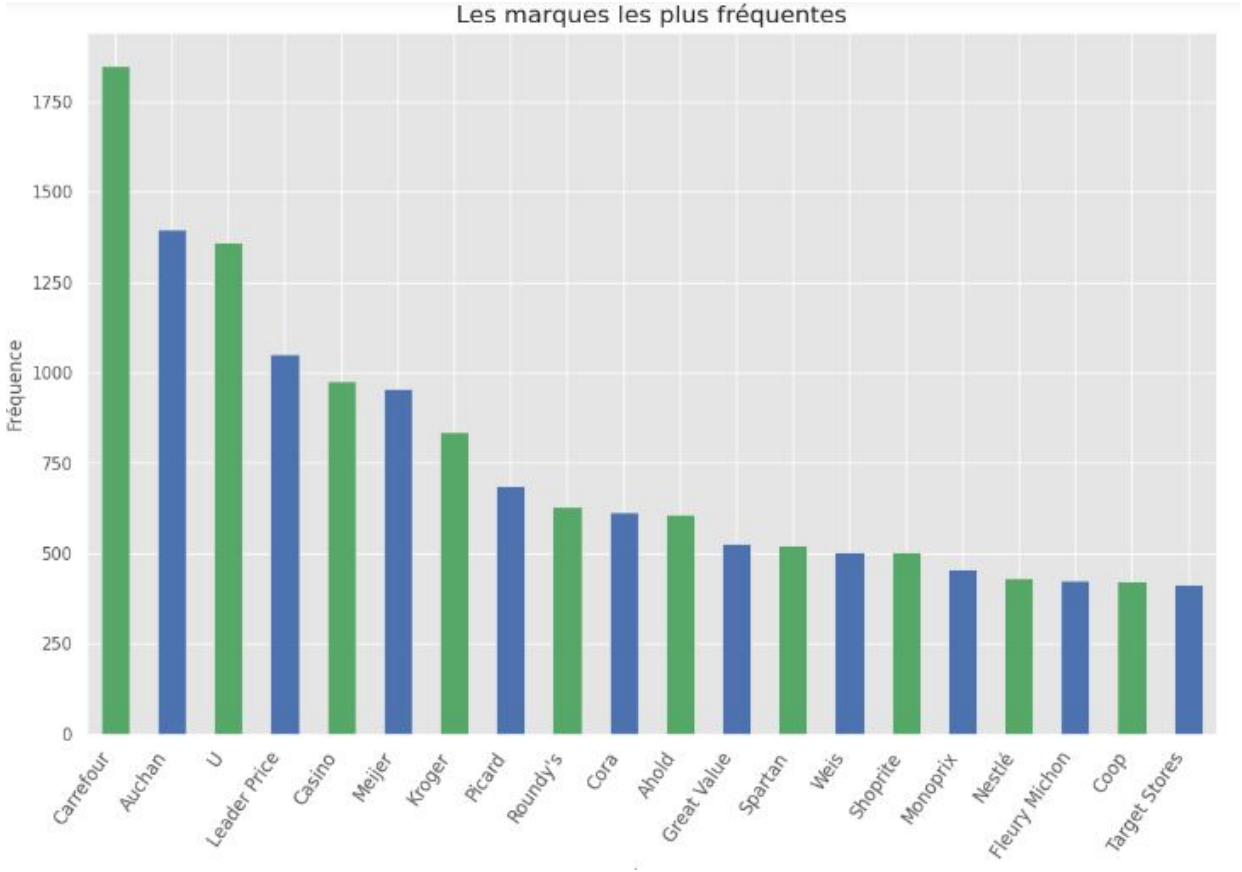
Distribution de la quantité des additifs



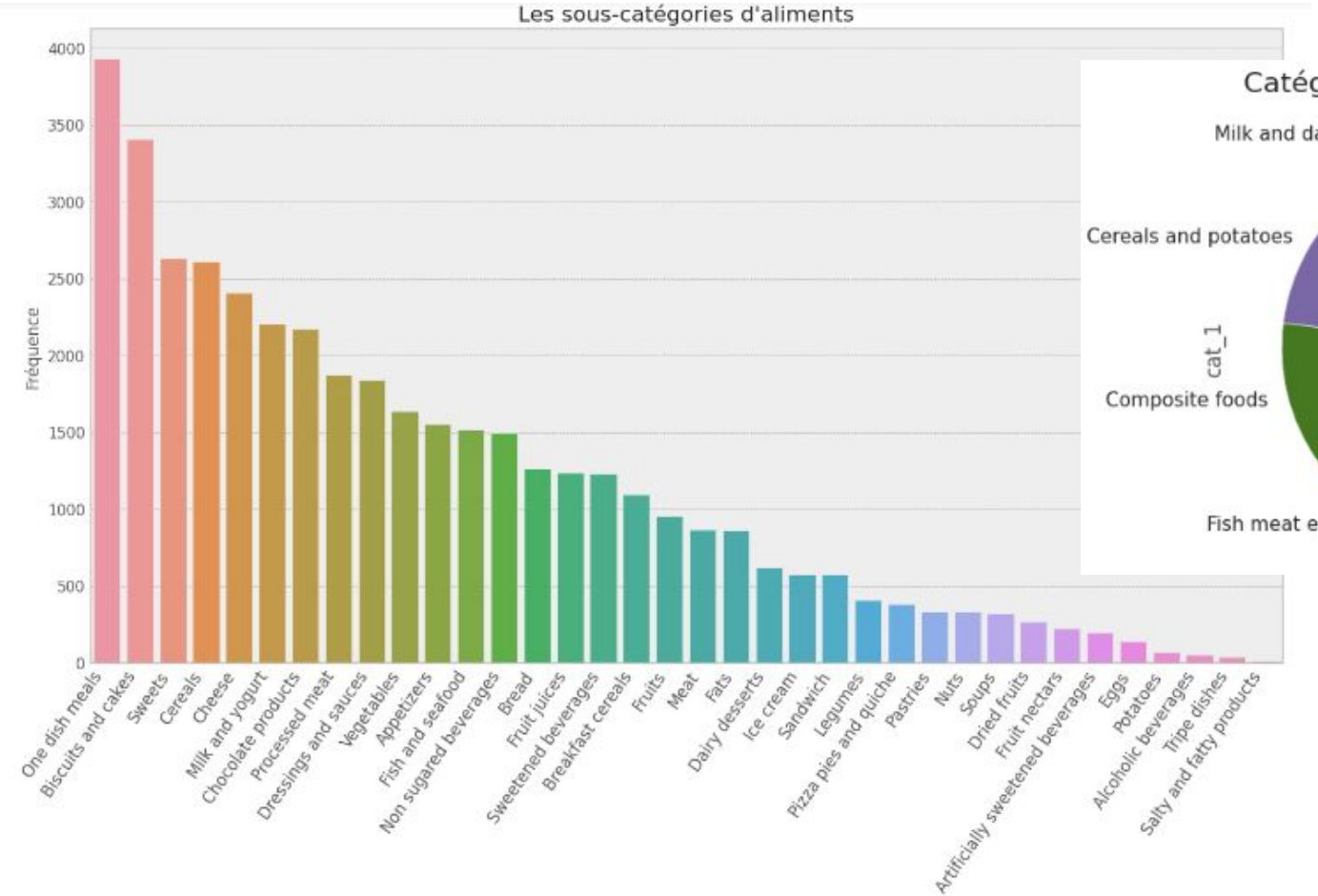
Analyse univariée

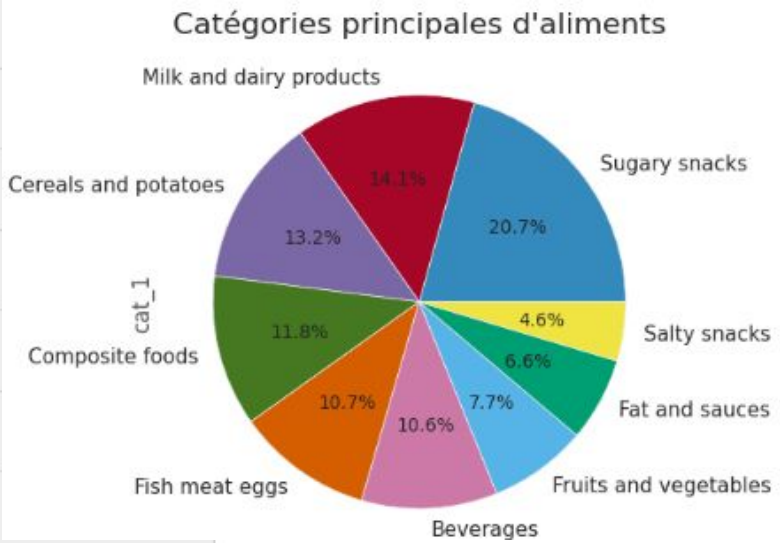
Variables catégorielles

Distribution des noms de marques



Distribution des catégories d'aliments

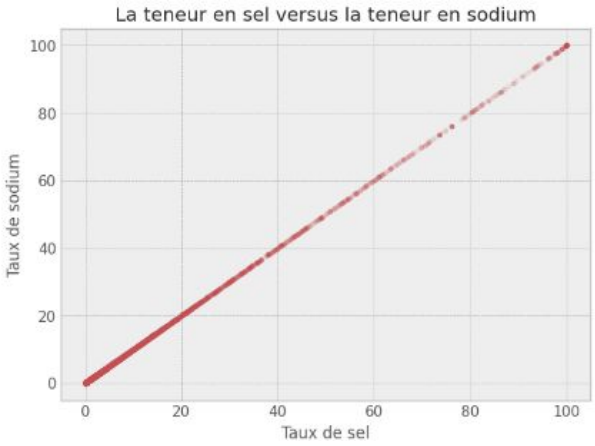




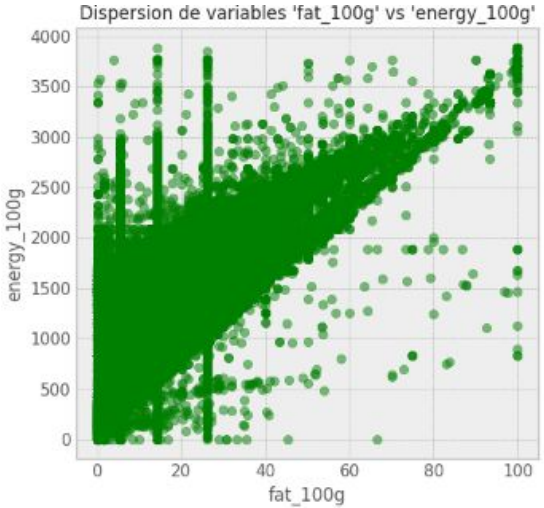
Analyse bivariée

Variables numériques

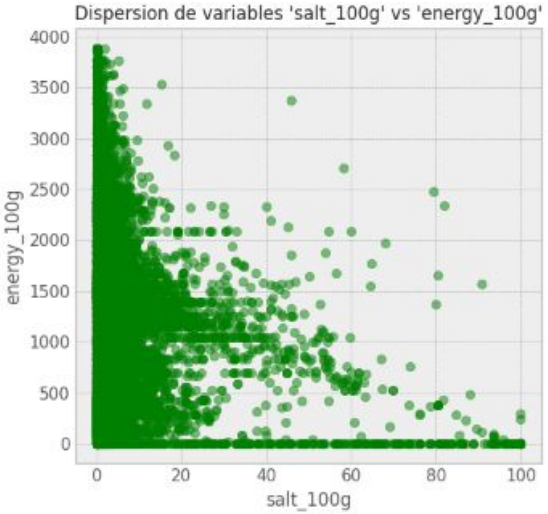
Covariance. Corrélation de Pearson. P-value.



r=1.0
p-value= 0.0



r=0.747
p-value= 0.0



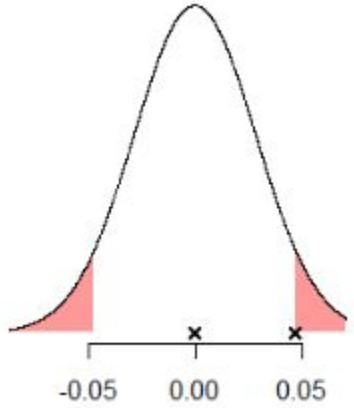
r=-0.269
p-value= 0.0

P-value

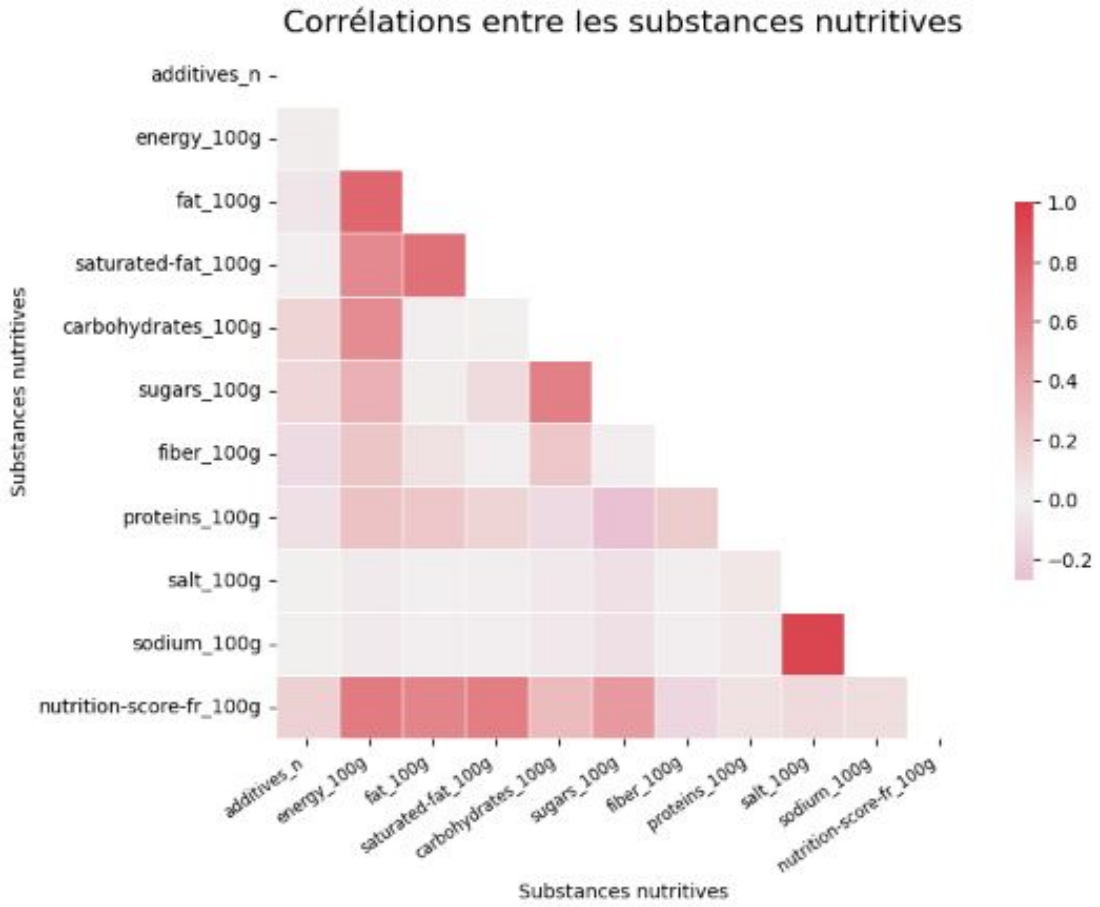
- un test statistique est significatif ou non ?
- quel seuil de certitude ?

	on accepte H_0	on rejette H_0
H_0 est vraie	bonne décision	*erreur (première espèce)*
H_0 est fausse	*erreur (seconde espèce)*	bonne décision

$p < 0.5$	Test significatif
$p = 0.0$	Test très significatif
$p > 0.5$	Test pas significatif



Analyse linéaire des corrélations.



Variables corrélées.

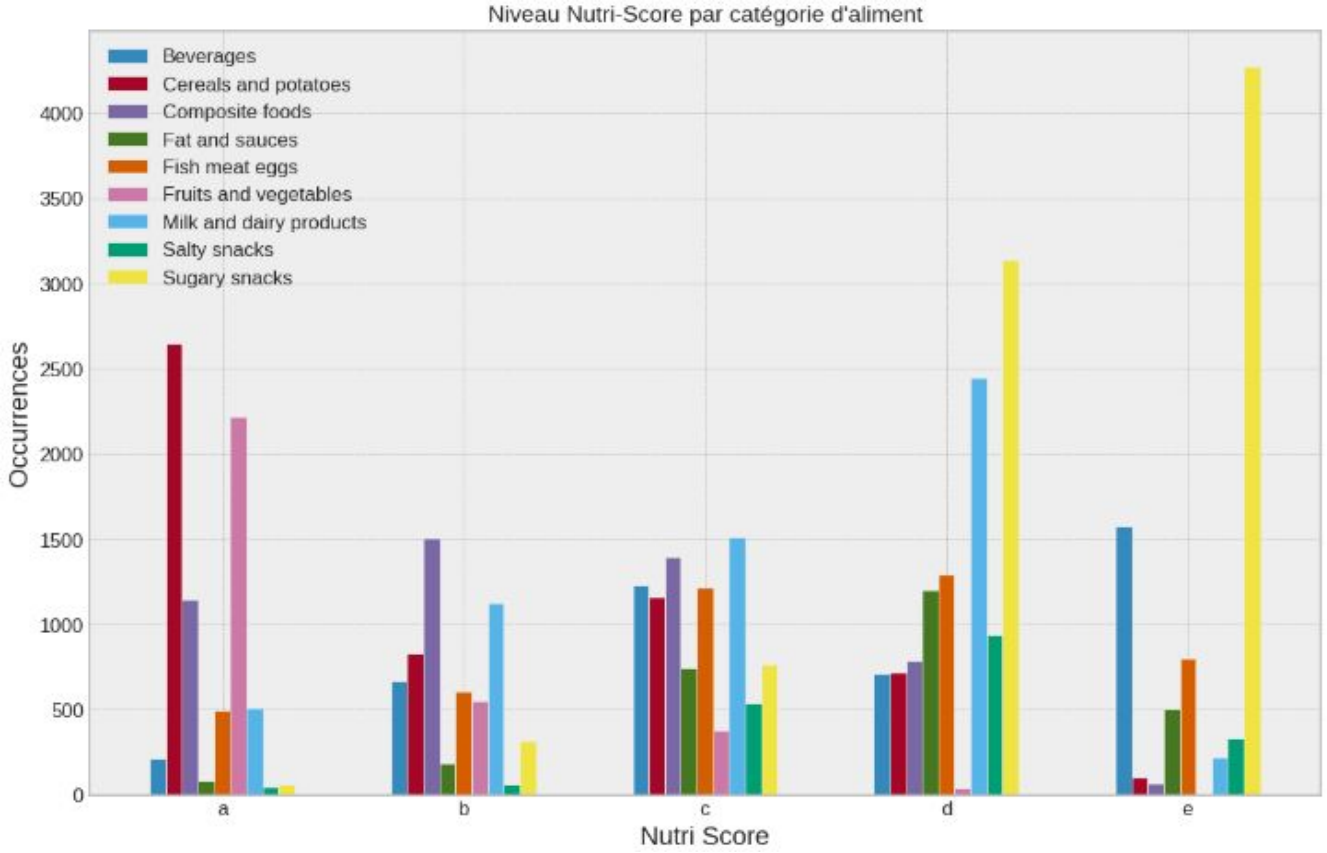
Variables anticorrélées.

Variables sans relation.

Analyse bivariée

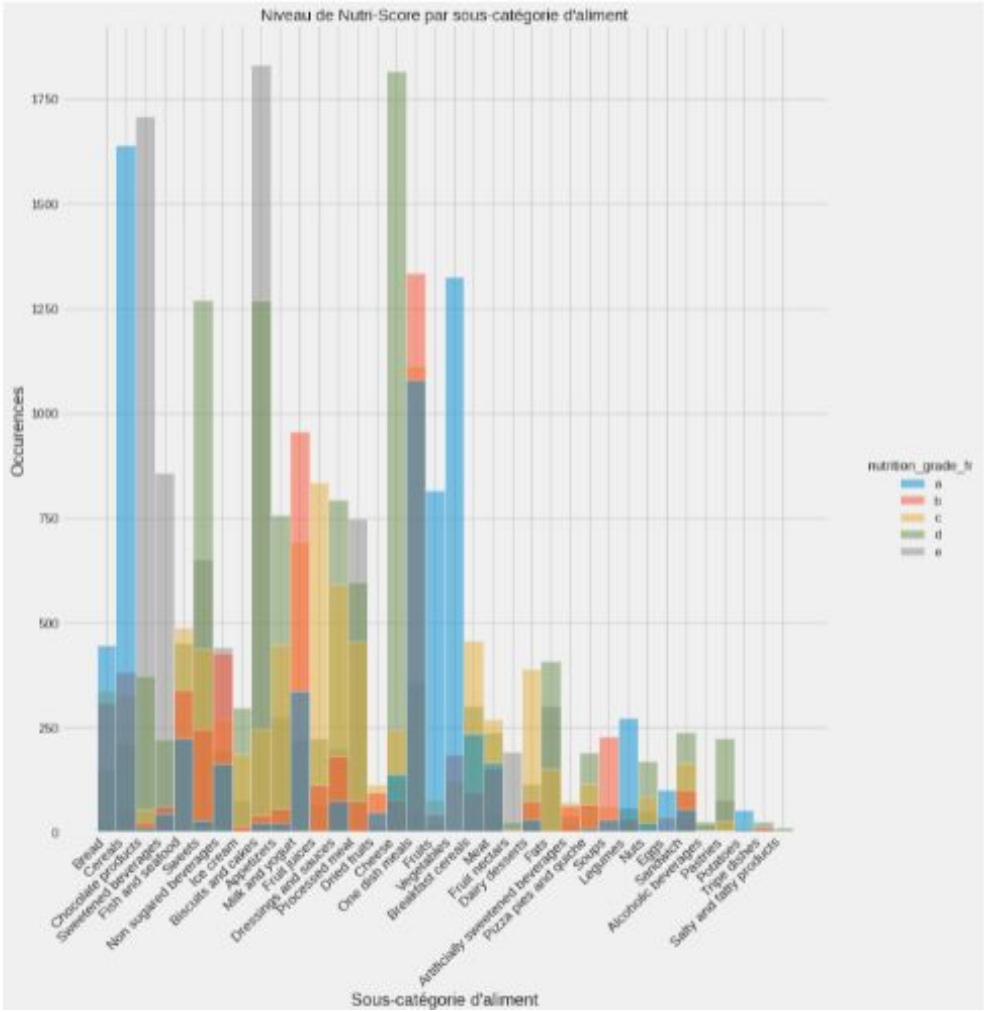
Variables catégorielles

Corrélation catégorie d'aliment vs note Nutri-Score

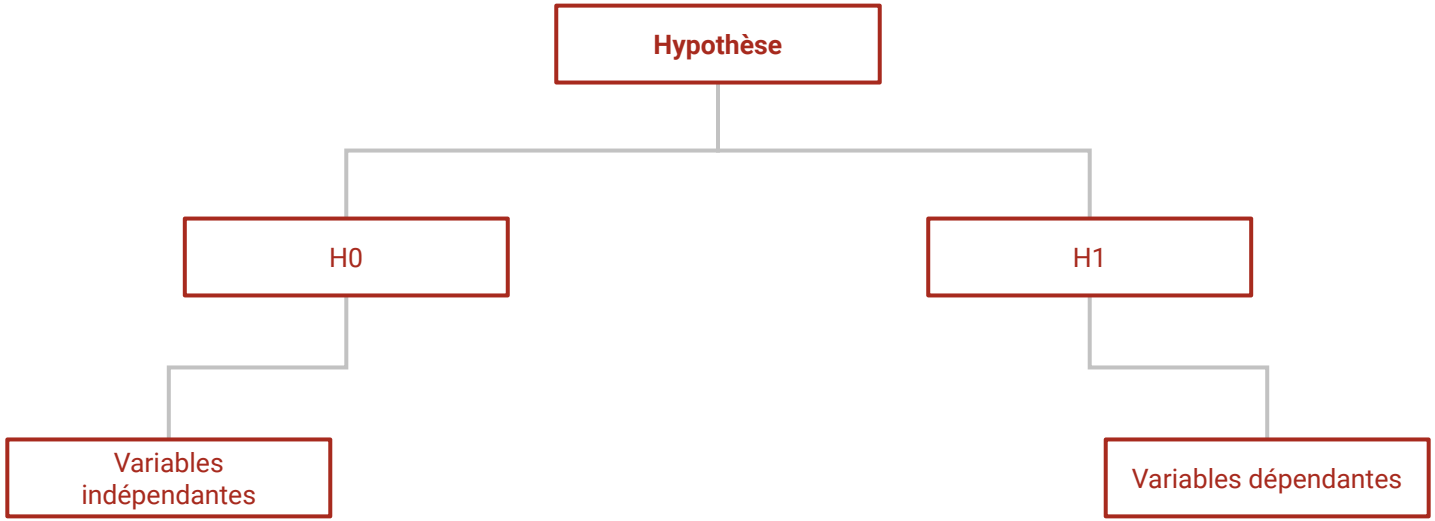


Sous-catégorie d'aliments vs note Nutri-Score

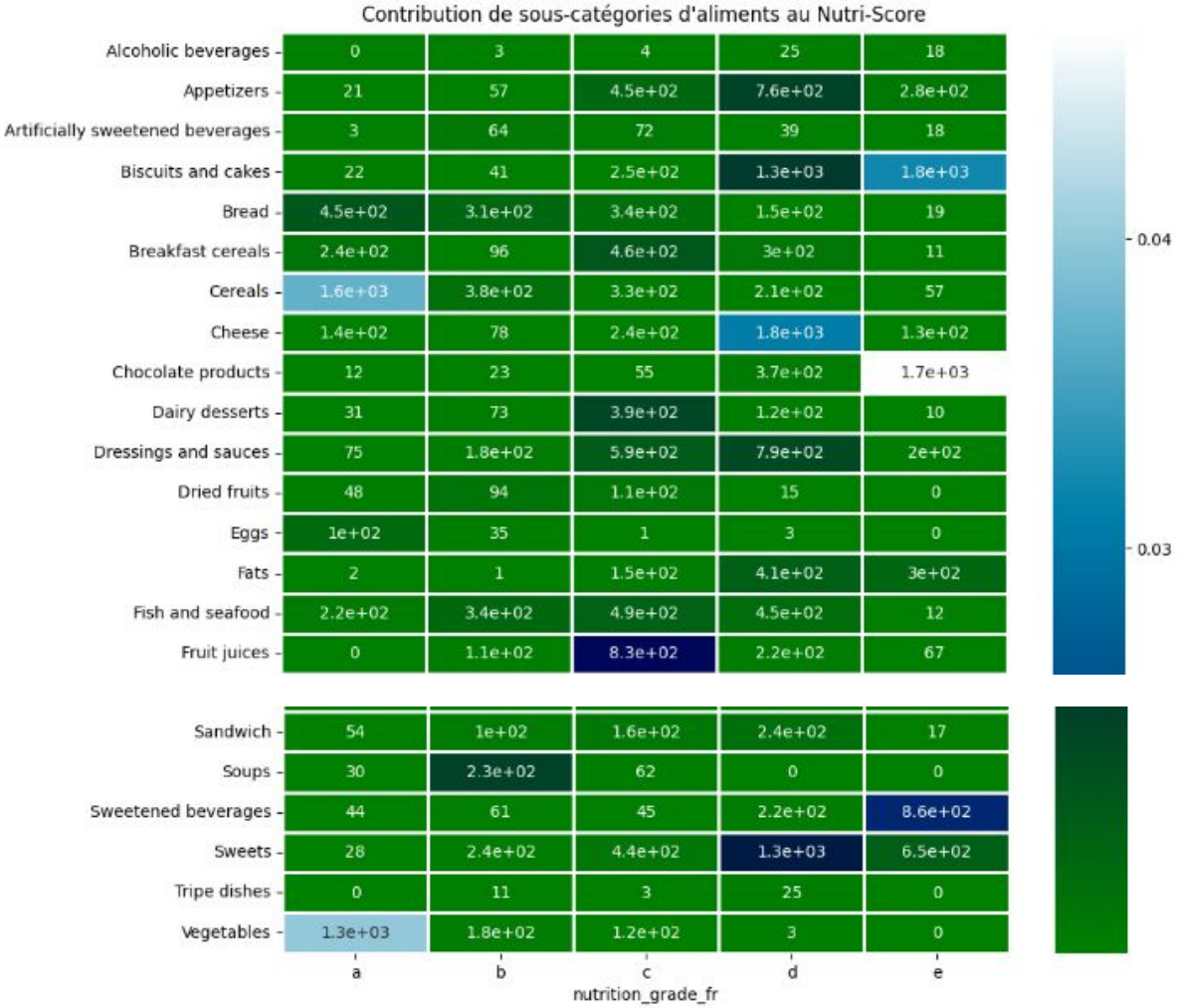
A	Cereals Fruits Vegetables
B	Milk and yogurt Soups
C	Fruits juices Dairy deserts
D	Sweets Cheese Biscuits and cakes
E	Chocolate products Sweetened beverages



Test d'indépendance CHI-2



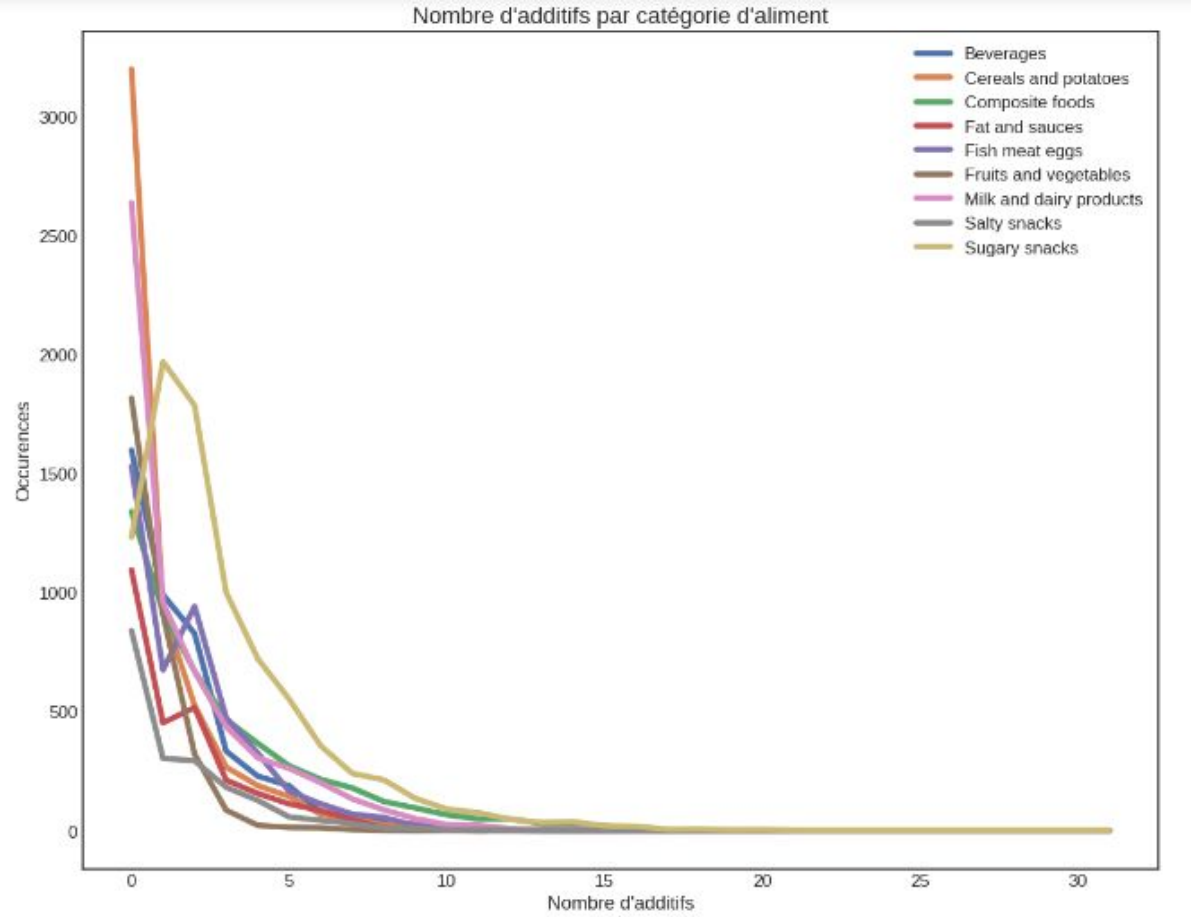
Un test $\chi 2$ d'indépendance nous permettra de définir la relation entre les variables catégorielles.


Contribution de sous-catégories d'aliments au Nutri-Score
Alcoholic beverages 0 3 4 25 18
Appetizers 21 57 4.5e+02 7.6e+02 2.8e+02
Artificially sweetened beverages 3 64 72 39 18
Biscuits and cakes 22 41 2.5e+02 1.3e+03 1.8e+03
Bread 4.5e+02 3.1e+02 3.4e+02 1.5e+02 19
Breakfast cereals 2.4e+02 96 4.6e+02 3e+02 11
Cereals 1.6e+03 3.8e+02 3.3e+02 2.1e+02 57
Cheese 1.4e+02 78 2.4e+02 1.8e+03 1.3e+02
Chocolate products 12 23 55 3.7e+02 1.7e+03
Dairy desserts 31 73 3.9e+02 1.2e+02 10
Dressings and sauces 75 1.8e+02 5.9e+02 7.9e+02 2e+02
Dried fruits 48 94 1.1e+02 15 0
Eggs 1e+02 35 1 3 0
Fats 2 1 1.5e+02 4.1e+02 3e+02
Fish and seafood 2.2e+02 3.4e+02 4.9e+02 4.5e+02 12
Fruit juices 0 1.1e+02 8.3e+02 2.2e+02 67
Sandwich 54 1e+02 1.6e+02 2.4e+02 17
Soups 30 2.3e+02 62 0 0
Sweetened beverages 44 61 45 2.2e+02 8.6e+02
Sweets 28 2.4e+02 4.4e+02 1.3e+03 6.5e+02
Tripe dishes 0 11 3 25 0
Vegetables 1.3e+03 1.8e+02 1.2e+02 3 0
a b c d e
nutrition_grade_fr
0.04
0.03

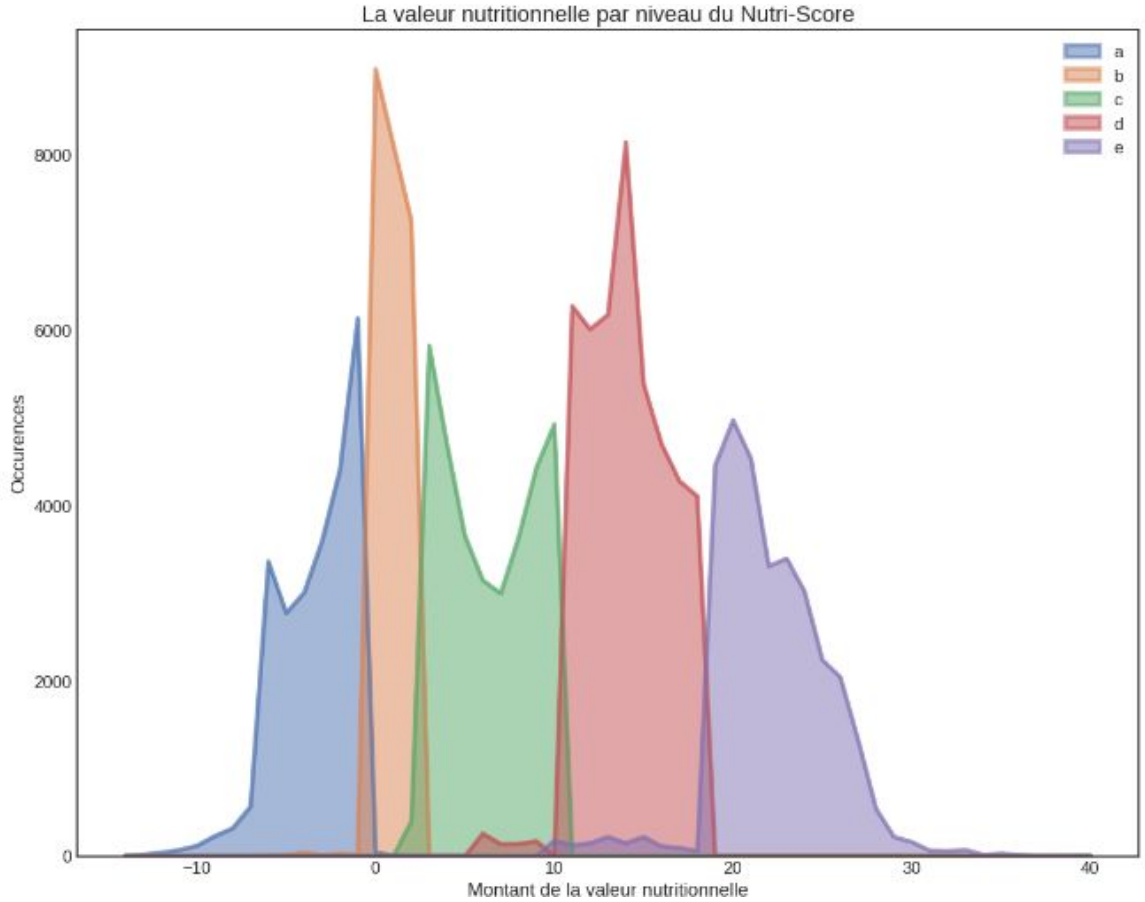
Analyse bivariée

Variables catégorielles
&
numériques

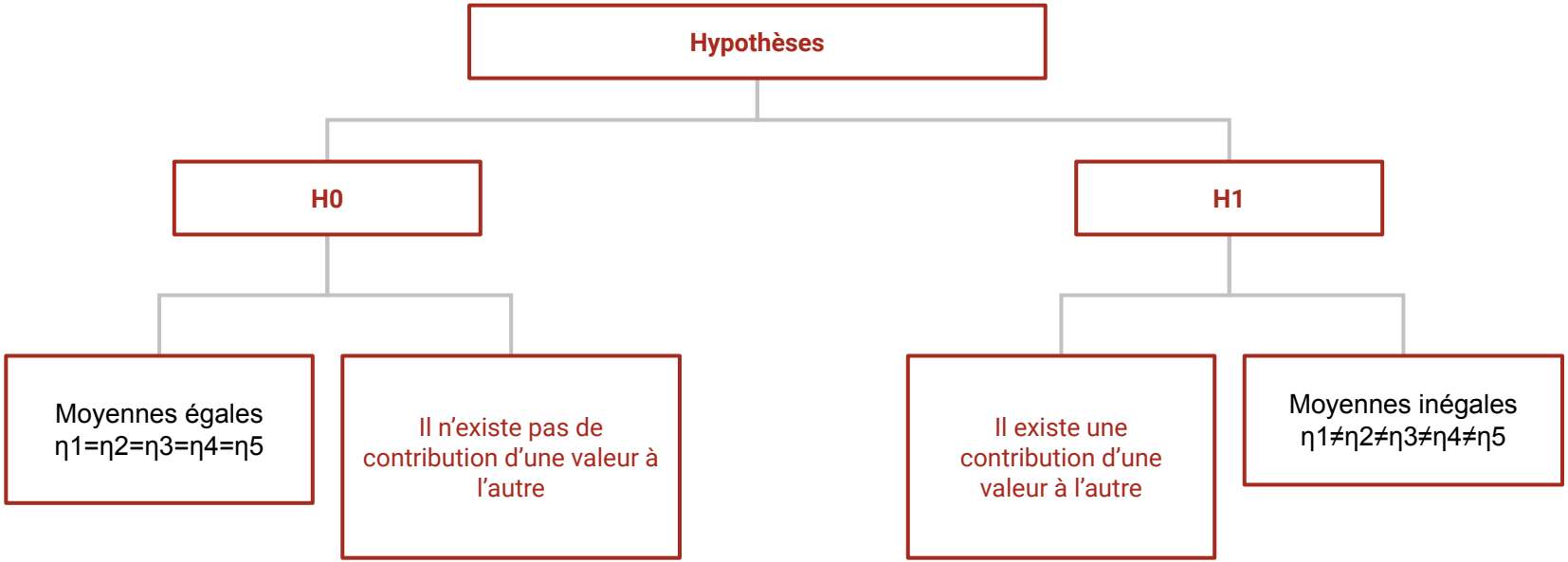
Corrélation additifs vs catégorie d'aliments



Corrélation valeur nutritionnelle vs note Nutri-Score



ANOVA - introduction



Un test de la variance permet de quantifier la relation entre les variables :

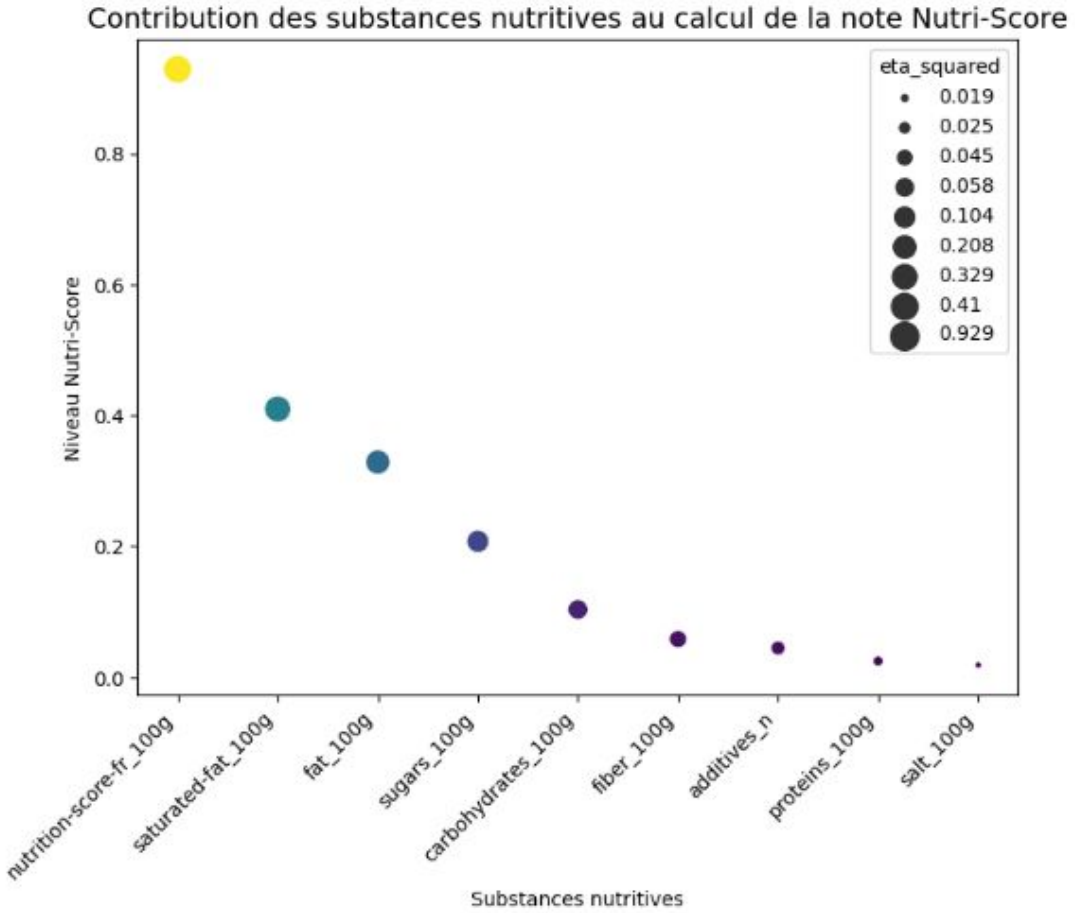
- **numériques (les substances nutritives)**
- **catégorielles (les niveaux de Nutri-Score)**

ANOVA - résultats

Résultats

Rejet H0
η2 > 0,5%

Validation H0
η2 < 0,5%



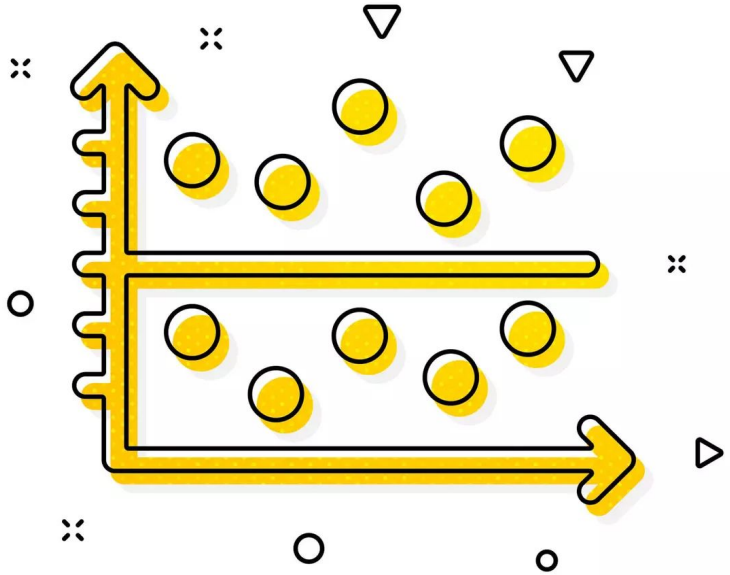
Analyse multi-variée

ACP

Analyse en Composantes Principales

Explorer de vastes jeux de données multidimensionnels

Repose sur des variables quantitatives

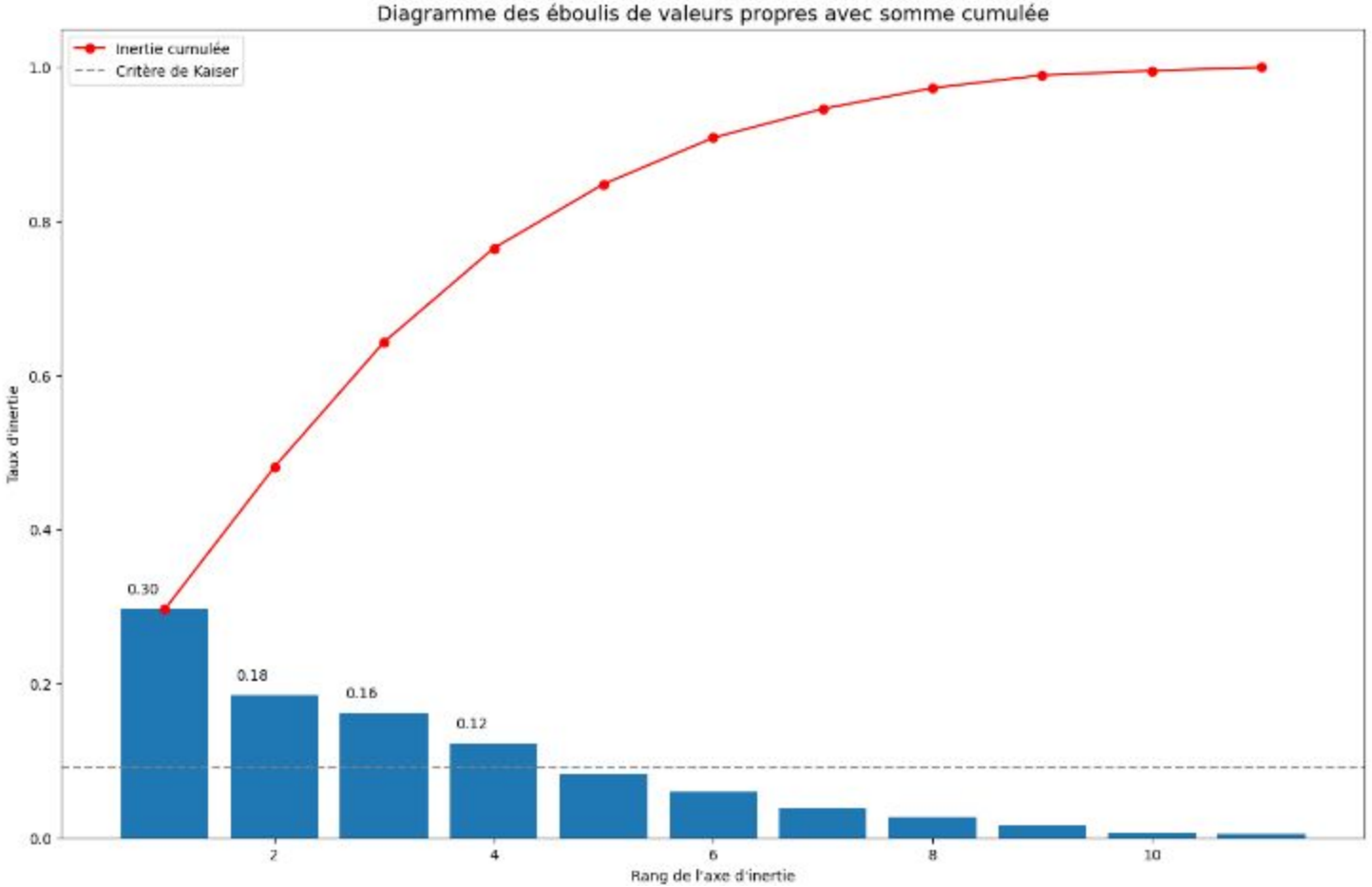


Transformer des variables corrélées en variables décorrélées appelées "composantes principales".

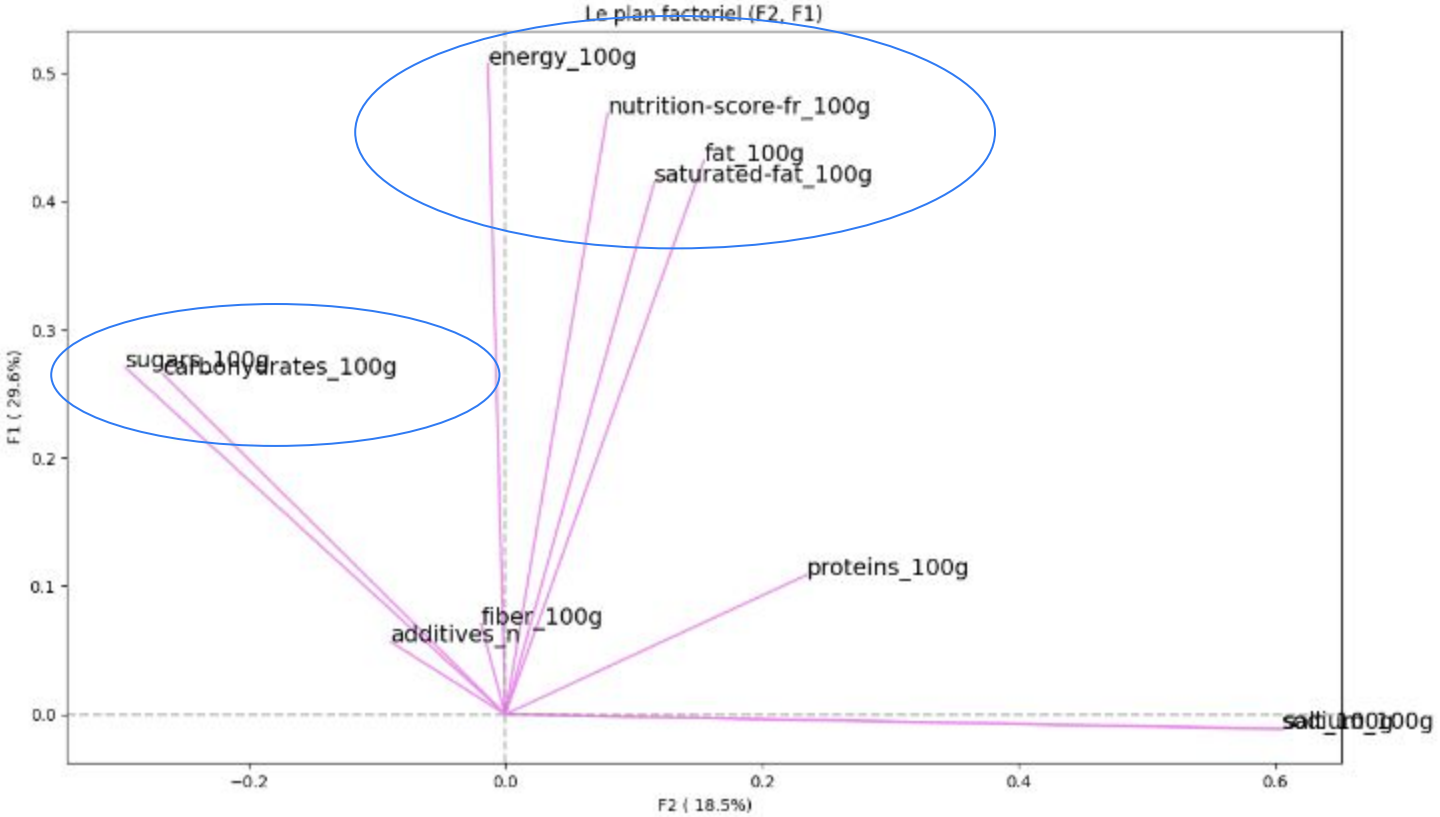
Réduire le nombre de variables appliquées à des individus.

Simplifier les observations tout en conservant un maximum d'informations.

Eboulis de valeurs propres



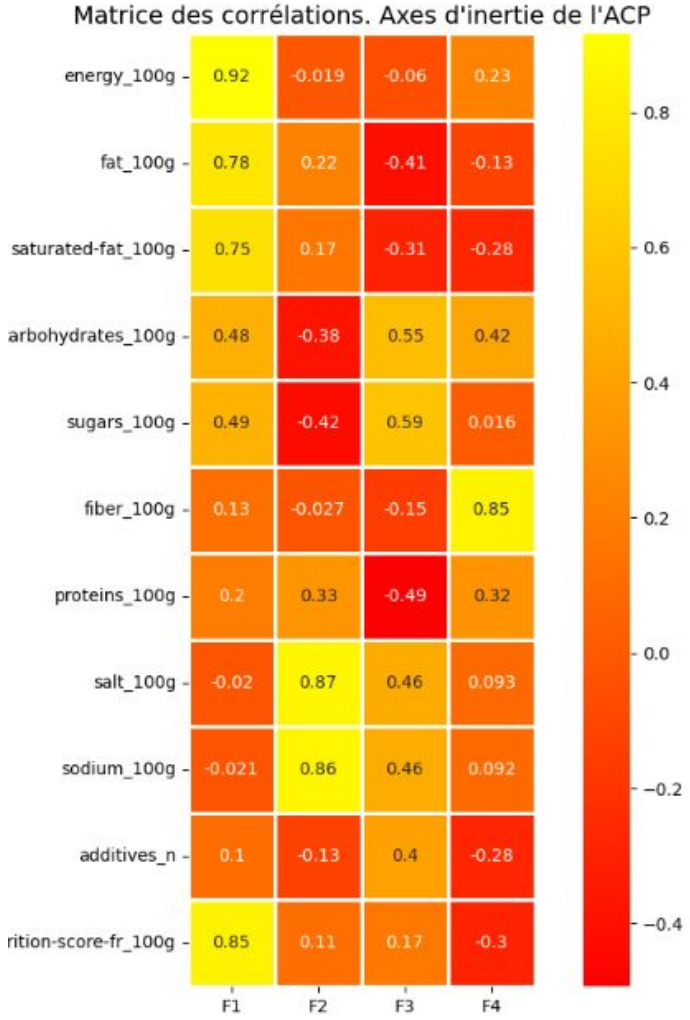
Plans factoriels. Cercle des corrélations.



Matrice des corrélations. Axes d'inertie.

Variables synthétiques :

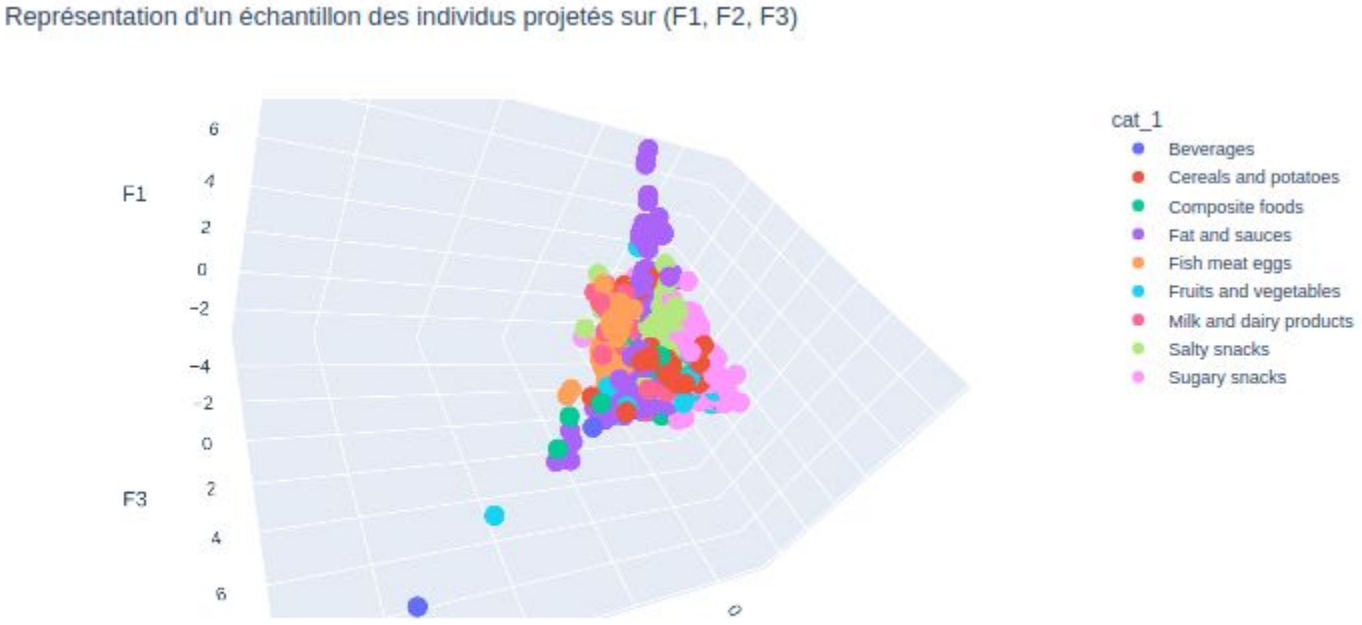
F1	peut correspondre à L'APPORT ÉNERGÉTIQUE
F2	peut correspondre à la SALINITÉ



Projection des individus

Groupes homogènes

Observations atypiques



Produits représentés par les nouvelles variables synthétiques.

4. La simulation


Surveiller sa dose de fructose

Conception

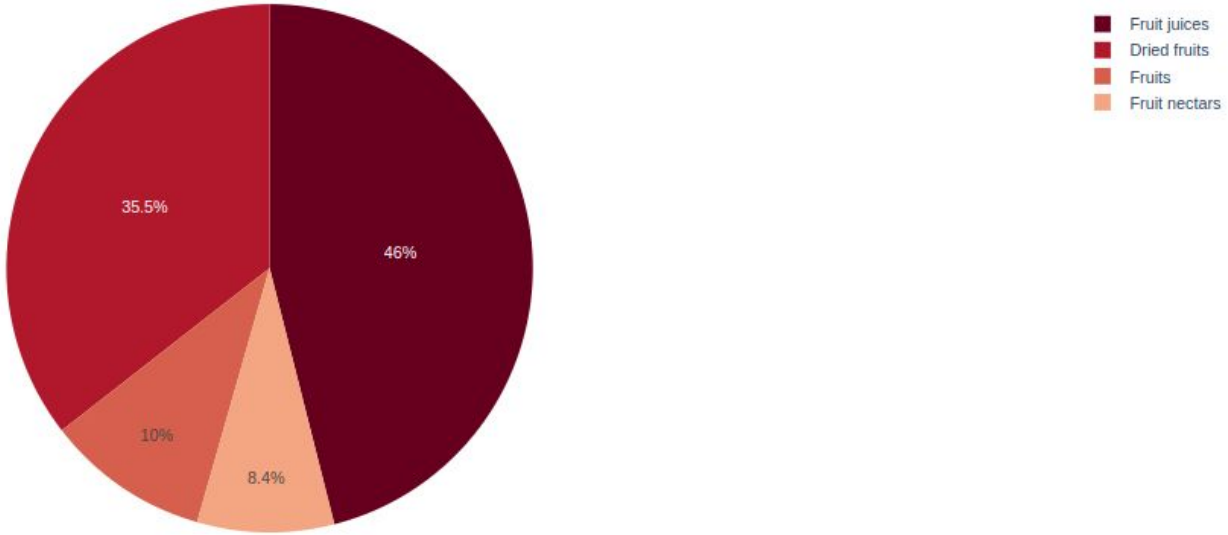
product_name	brands	additives_n	additives	nutrition_grade_fr	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	cat_1	cat_2
Frutas de Aragón con chocolate	Caro	4.0	[frutas-confitadas-en-proporcion-variable ->...	e	1668.48	13.09	8.09	64.52	Fruits and vegetables	Fruits
Bombones rellenos Cereza & Licor	Dia,//Propiedad de://,Dia - Distribuidora Inte...	2.0	[azucar -> es:azucar] [pasta-de-cacao ->...	e	1830.00	20.00	13.00	60.00	Fruits and vegetables	Fruits
Bombones Mon Chéri	Mon Chéri,Ferrero,//Propiedad de://,Ferrero S....	2.0	[chocolate-negro-49 -> es:chocolate-negro-49...	e	1902.00	20.30	13.20	52.80	Fruits and vegetables	Fruits

Base de données

Informations générales

Quels fruits contient notre base de données ?

Les catégories de fruits



Application

Détection du produit

Recherche d'informations sur un produit

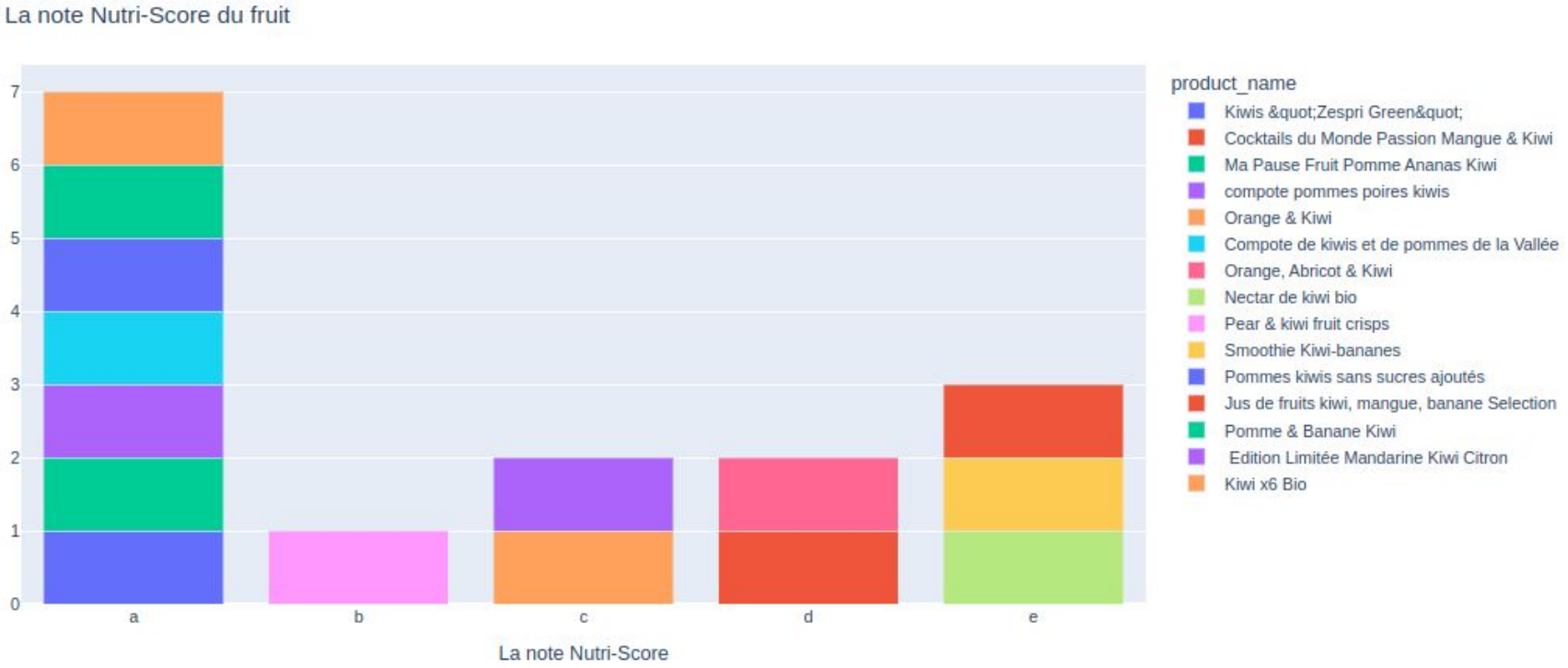
===> Par le code barre

	code	url	product_name	brands	additives_n	nutrition_grade_fr	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g
31271	3270160680924	http://world-fr.openfoodfacts.org/produit/3270...	Bol de fruits rouges	Picard	0.0	a	343.0	0.6	0.1	15.7	11.0	4.4	1.2

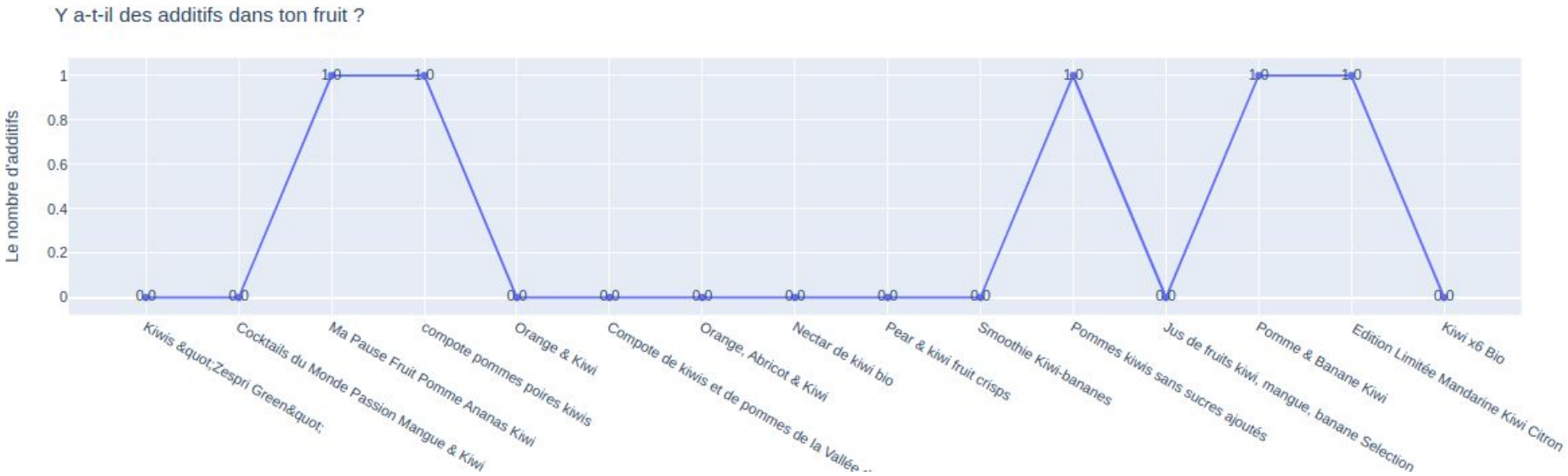
===> Par le nom du produit

```
Le fruit que vous avez choisi est un(e) " kiwi ".
```

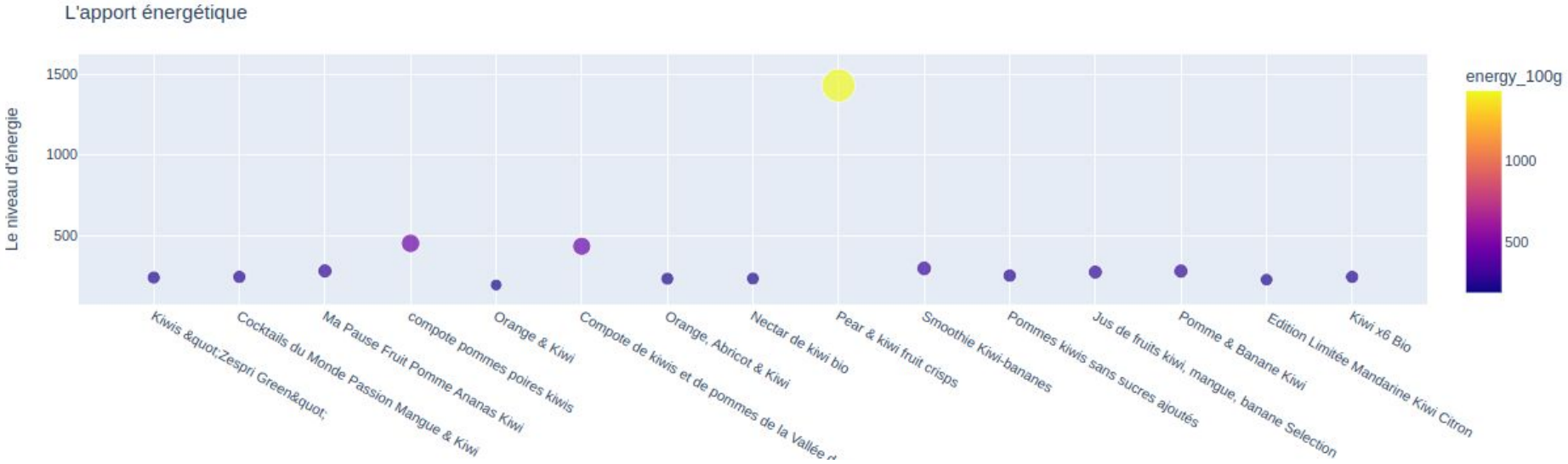
Note Nutri-Score



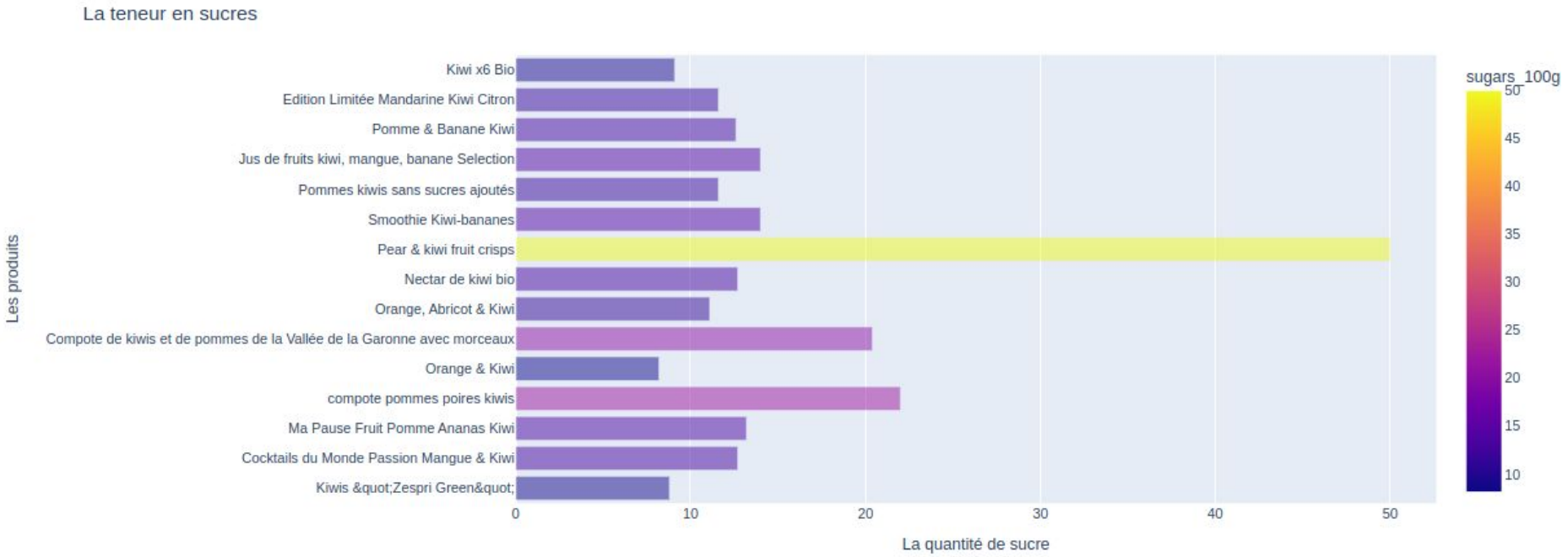
Additifs


Y a-t-il des additifs dans ton fruit ?
Le nombre d'additifs
1
0.8
0.6
0.4
0.2
0
Kiwis "Zespri Green"
Cocktails du Monde Passion Mangue & Kiwi
Ma Pause Fruit Pomme Ananas Kiwi
compote pommes poires kiwis
Orange & Kiwi
Compote de kiwis et de pommes de la Vallée ...
Orange, Abricot & Kiwi
Nectar de kiwi bio
Pear & kiwi fruit crisps
Smoothie Kiwi-bananes
Pommes kiwis sans sucres ajoutés
Jus de fruits kiwi, mangue, banane Selection
Pomme & Banane Kiwi
Edition Limitée Mandarine Kiwi Citron
Kiwi x6 Bio

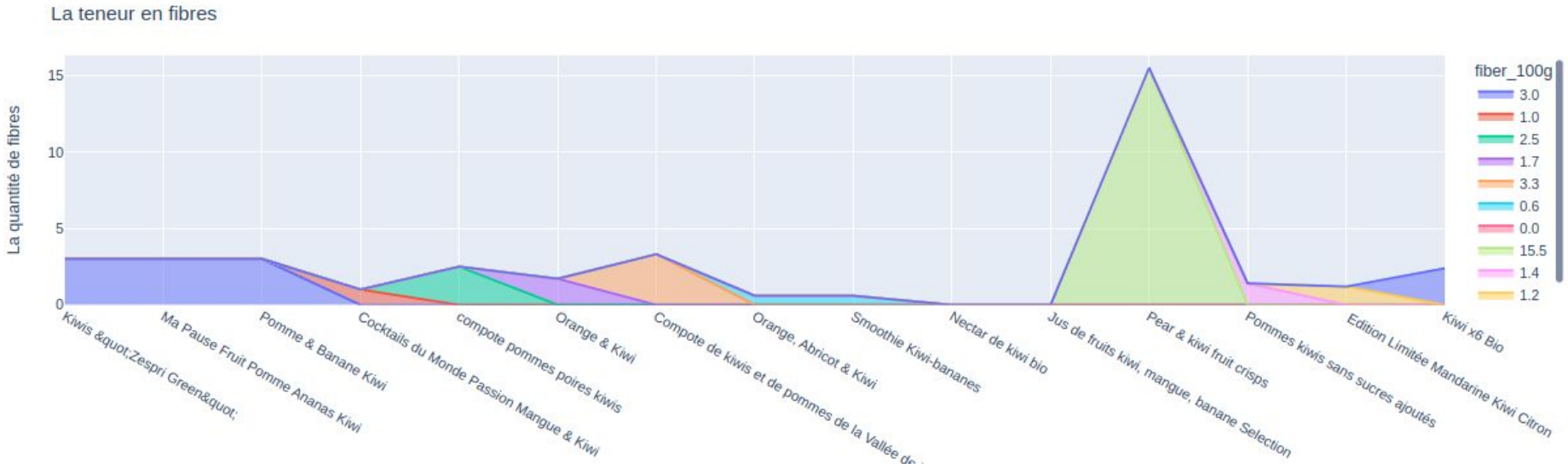
Apport énergétique



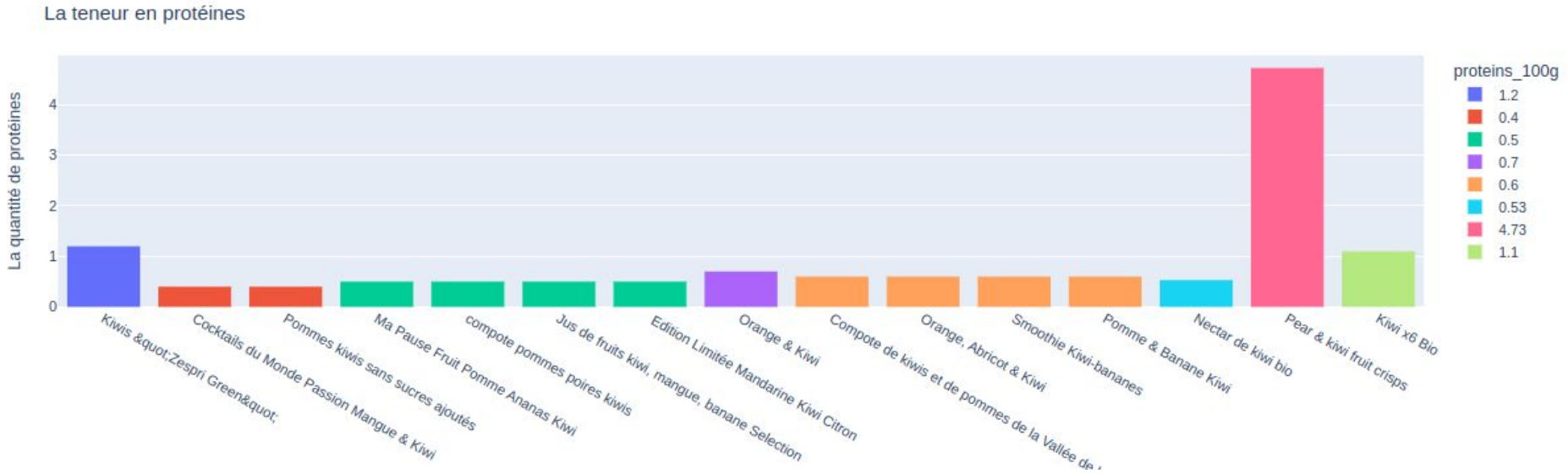
Sucres


La teneur en sucres
Les produits
Kiwi x6 Bio
Edition Limitée Mandarine Kiwi Citron
Pomme & Banane Kiwi
Jus de fruits kiwi, mangue, banane Selection
Pommes kiwis sans sucres ajoutés
Smoothie Kiwi-bananes
Pear & kiwi fruit crisps
Nectar de kiwi bio
Orange, Abricot & Kiwi
Compote de kiwis et de pommes de la Vallée de la Garonne avec morceaux
Orange & Kiwi
compote pommes poires kiwis
Ma Pause Fruit Pomme Ananas Kiwi
Cocktails du Monde Passion Mangue & Kiwi
Kiwis "Zespri Green"
0
10
20
30
40
50
La quantité de sucre
sugars_100g
50
45
40
35
30
25
20
15
10

Fibres


La teneur en fibres
La quantité de fibres
15
10
5
0
fiber_100g
3.0
1.0
2.5
1.7
3.3
0.6
0.0
15.5
1.4
1.2
Kiwis "Zespri Green"
Ma Pause Fruit Pomme Ananas Kiwi
Pomme & Banane Kiwi
Cocktails du Monde Passion Mangue & Kiwi
compote pommes poires kiwis
Orange & Kiwi
Compote de kiwis et de pommes de la Vallée de
Orange, Abricot & Kiwi
Smoothie Kiwi-bananes
Nectar de kiwi bio
Jus de fruits kiwi, mangue, banane Selection
Pear & kiwi fruit crisps
Pommes kiwis sans sucres ajoutés
Edition Limitée Mandarine Kiwi Citron
Kiwi x6 Bio

Protéines


La teneur en protéines
La quantité de protéines
0
1
2
3
4
proteins_100g
1.2
0.4
0.5
0.7
0.6
0.53
4.73
1.1
Kiwis "Zespri Green"
Cocktails du Monde Passion Mangue & Kiwi
Pommes kiwis sans sucres ajoutés
Ma Pause Fruit Pomme Ananas Kiwi
compote pommes poires kiwis
Jus de fruits kiwi, mangue, banane Selection
Edition Limitée Mandarine Kiwi Citron
Orange & Kiwi
Compote de kiwis et de pommes de la Vallée de
Orange, Abricot & Kiwi
Smoothie Kiwi-bananes
Pomme & Banane Kiwi
Nectar de kiwi bio
Pear & kiwi fruit crisps
Kiwi x6 Bio

Recherche avec itable

Show 10 entries

Search: goji

	product_name	nutrition_grade_fr	brands
280	Gayelord Hauser Goji	d	Gayelord Hauser
22966	Goji-Beeren	c	Kluth
26331	Goji	c	Gayelord Hauser
45517	Bayas de Goji secas	a	Sin marca
146764	Jus d'Orange Goji - Acérola - Passion	c	Sans marque,Agidra
149998	Baies de Goji Biologiques	a	La Vie Claire

Showing 1 to 6 of 6 entries (filtered from 2,691 total entries)

Previous 1 Next

Avez-vous des questions ?