Santé publique France

Concevez une application au service de la santé publique

Recommandation Alimentaire pour Maladies Héréditaire du Métabolisme

Source : OPEN FOOD FACTS

https://world.openfoodfacts.org/

Sommaire

Sommaire

Idée d'application

Atteintes de Phénylcétonurie

Souhaite trouver le lait de soja recommandé pour leurs maladies

Recherche

Scoring
Tri

OPEN FOOD FACTS
Base de données

Moteur de Recommandation Alimentaire pour les personnes atteintes de Phénylcétonurie (Maladies Métaboliques Héréditaires)

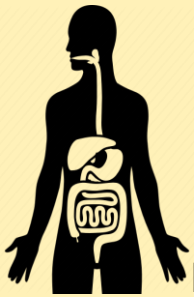
Idée d'application – La phénylcétonurie ?

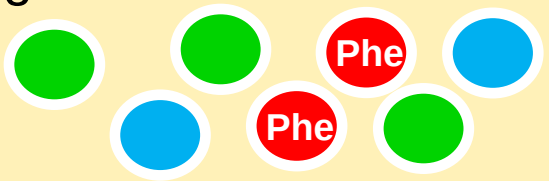
Diagnostic prénatal, test de Guthrie

Alimentation normale, riche en protéines

Molécules de protéine composées d'acides aminés

Digestion des aliments





Fragmentation en acides aminés

Transformation de la Phe en tyrosine à l'aide de PAH

Tyrosine

Phe

Phe

Enzyme PAH intact

Saine

Personne?

Atteinte PCU

Phe

Enzyme PAH **déficiente**

Phe

Phe **non transformée** en tyrosine, **ni assimilée**

Phe
Phe
Phe
Phe
Phe
Phe

Taux sanguin de Phe élevé : **intoxication**

Troubles du développement cérébral et dommages physiques et mentaux irréversibles

PAH = Phénylalanine hydroxylase

Idée d'application – Alimentation ?

Alimentation Saine, Équilibrée

POINTS ALIMENTS SOLIDES	POINTS BOISSONS
-15 à -1	Eau
0 à 2	≤ 1
3 à 10	2 à 5
11 à 18	6 à 9
19 à 40	10 à 40

Protéines : 2g/kg/Jour (> 4ans)

Produits Peu Transformés

Aspartame

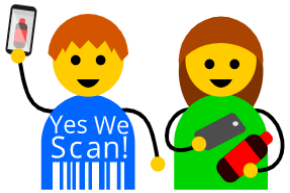
Idée d'application – Informations utiles

Régime hypoprotidique	Alimentation saine	Alimentation équilibrée	Peu transformés	Produits disponibles
• Peu de protéines	• Nutri-score faible • Peu de sel • Peu d'additifs • Pas huile de palme • Sucre limité • Peu calorique	• Macro-nutriments ➢ Lipides ➢ Glucides • Micro-nutriments ➢ Vitamines ➢ minéraux	• Groupe NOVA	• Nom du produit • Photo • Catégorie • Marque • Vendus en France

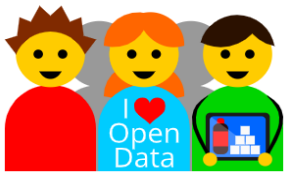
Sommaire

2 Jeu de données

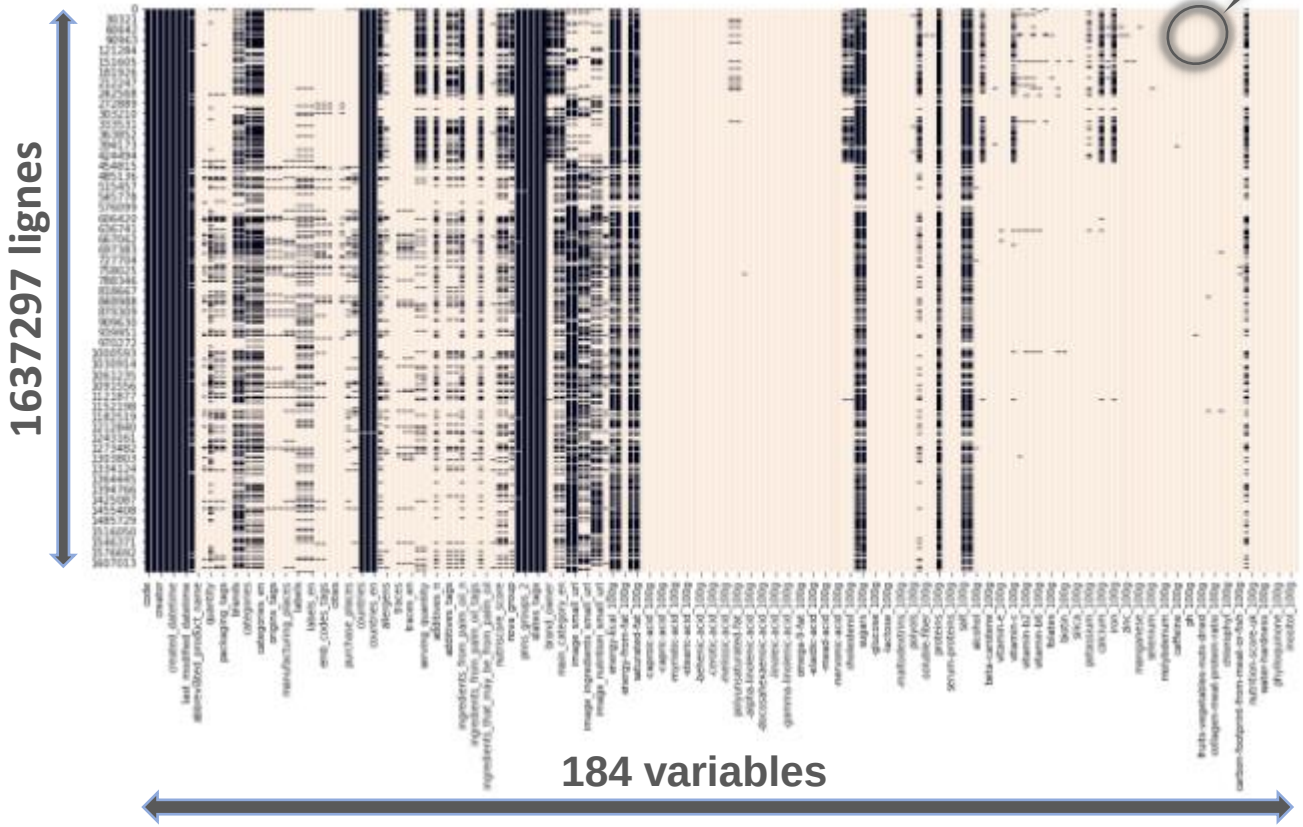
Une base de produits alimentaires



Faite par tout le monde



Pour tout le monde



> 79% de valeurs manquantes

118 variables > 90% de valeurs manquantes

5 sections

- Informations générales
- tags
- ingrédients
- Données diverses
- Infos nutritionnelles

2 Jeu de données - variables utiles

Régime hypoprotidique

proteins_100g

Alimentation saine

nutriscore_score
nutriscore_grade
sodium_100g
salt_100g
additives_n
additives_tags
additives_en
ingredients_from_palm_oil_n
ingredients_that_may_be_from_palm_oil_n
sugars_100g
energy-kcal_100g
energy_100g
nutrition-score-fr_100g
additives
ingredients_from_palm_oil_tags
ingredients_that_may_be_from_palm_oil
ingredients_that_may_be_from_palm_oil_tags
energy-kj_100g
energy-from-fat_100g
nutrition-score-uk_100g
allergens
ingredients_from_palm_oil
allergens_en

Alimentation équilibrée

fat_100g
saturated-fat_100g
carbohydrates_100g
fiber_100g
-butyric-acid_100g
-caproic-acid_100g
-caprylic-acid_100g
-capric-acid_100g
-lauric-acid_100g
-myristic-acid_100g
-palmitic-acid_100g
-stearic-acid_100g
-arachidic-acid_100g
-behenic-acid_100g
-lignoceric-acid_100g
-cerotic-acid_100g
-montanic-acid_100g
-melissic-acid_100g
monounsaturated-fat_100g
polyunsaturated-fat_100g
omega-3-fat_100g
-alpha-linolenic-acid_100g
-eicosapentaenoic-acid_100g
-docosahexaenoic-acid_100g
omega-6-fat_100g
-linoleic-acid_100g
-arachidonic-acid_100g
-gamma-linolenic-acid_100g
-dihomo-gamma-linolenic-acid_100g
-nervonic-acid_100g
-sucrose_100g
zinc_100g
copper_100g
manganese_100g
-lactose_100g
-maltose_100g
-maltodextrins_100g
starch_100g
polyols_100g
-soluble-fiber_100g
-insoluble-fiber_100g
casein_100g
serum-proteins_100g
nucleotides_100g
alcohol_100g
vitamin-a_100g
beta-carotene_100g
vitamin-d_100g
vitamin-e_100g
vitamin-k_100g
vitamin-c_100g
vitamin-b1_100g
vitamin-b2_100g
vitamin-pp_100g
vitamin-b6_100g
vitamin-b9_100g
folates_100g
vitamin-b12_100g
biotin_100g
pantothenic-acid_100g
silica_100g
bicarbonate_100g
potassium_100g
chloride_100g
calcium_100g
phosphorus_100g
magnesium_100g
fluoride_100g
selenium_100g
chromium_100g
molybdenum_100g
iodine_100g
caffeine_100g
taurine_100g
ph_100g
fruits-vegetables-nuts_100g
fruits-vegetables-nuts-dried_100g
fruits-vegetables-nuts-estimate_100g
collagen-meat-protein-ratio_100g
cocoa_100g
chlorophyl_100g
glycemic-index_100g
water-hardness_100g
choline_100g
phylloquinone_100g
beta-glucan_100g
inositol_100g
-glucose_100g
-fructose_100g
iron_100g
magnesium_100g
cholesterol_100g
trans-fat_100g
omega-9-fat_100g
-oleic-acid_100g
-elaidic-acid_100g
-gondoic-acid_100g
-mead-acid_100g
-erucic-acid_100g
carnitine_100g

Peu transformés

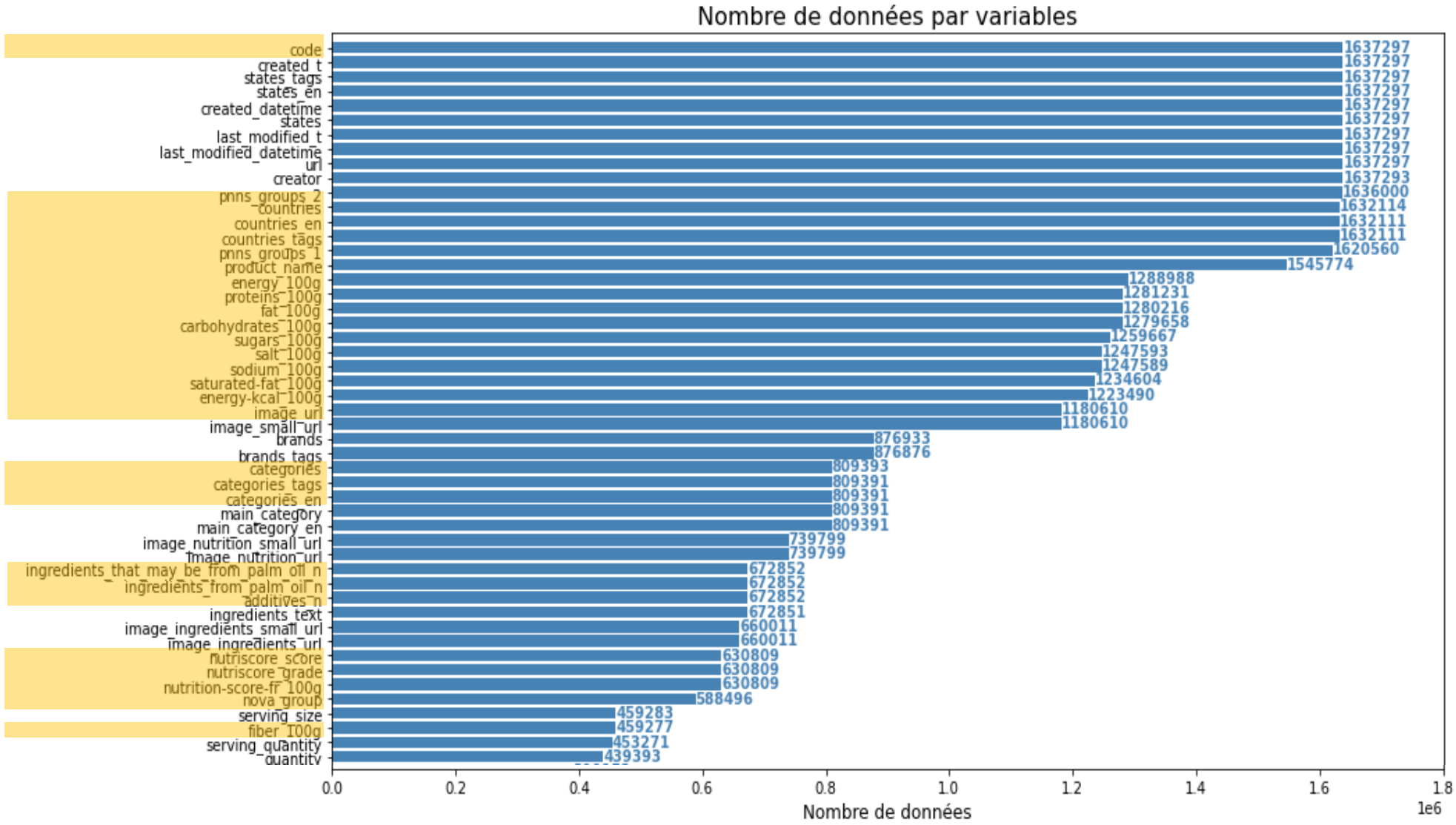
nova_group

Produits disponibles

code
product_name
brands
brands_tags
categories
categories_tags
categories_en
countries
countries_tags
countries_en
pnns_groups_1
pnns_groups_2
image_url
generic_name

Légende :
En_marron : >80% NaN
En_rouge : 100% NaN

2 Jeu de données - variables utiles



Régime hypoprotidique ✓

Alimentation saine ✓

Alimentation équilibrée ✗

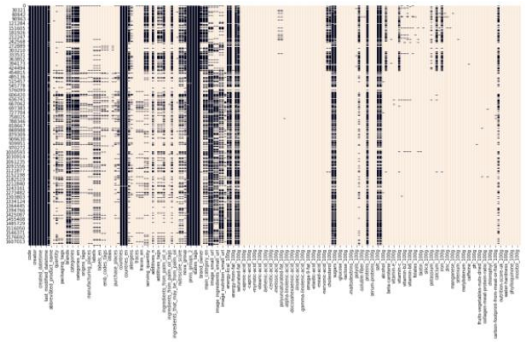
Peu transformés ✓

Produits disponibles ✓

Nettoyage – réduction du jeu de données

De…

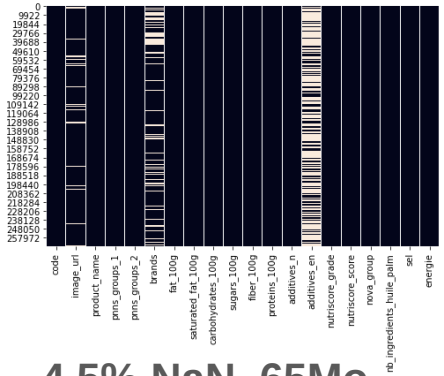
1637297 lignes, 184 variables



79,5% NaN, 3,5Go

… À

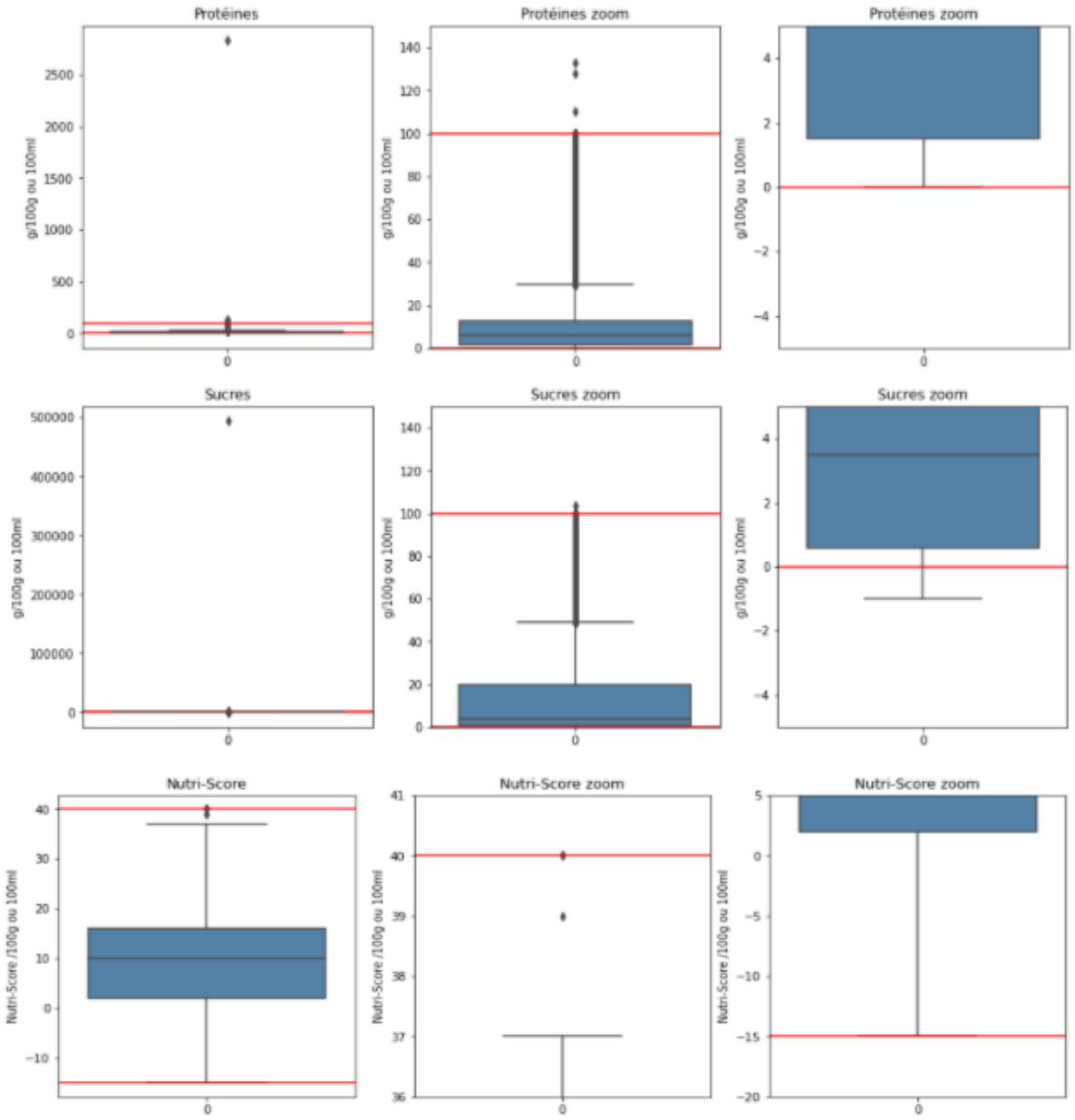
267881 lignes, 20 variables



4,5% NaN, 65Mo

	Nb lignes	Nb var.
Chargement des données	**1637297**	**184**
Suppression des 9 variables vides	1637297	175
Remplacer '-' par '_' dans le nom des variables Suppression des 2 doublons sur le code produit	1637295	175
Sélection des variables pertinentes pour l'application	1637295	124
Suppression des variables avec %NaN > 80%	1637295	24
Suppression des lignes sans nom produit	1545773	24
Conservation des produits vendus en France Traduction des groupes et sous-groupes de produit	748948	24
Gestion des valeurs aberrantes	748703	23
Feature engineering huile de palme	748703	22
Feature engineering sel	748703	21
Feature engineering énergie	748703	20
Gestion des valeurs manquantes	**267881**	**20**

2 Nettoyage – Valeurs aberrantes



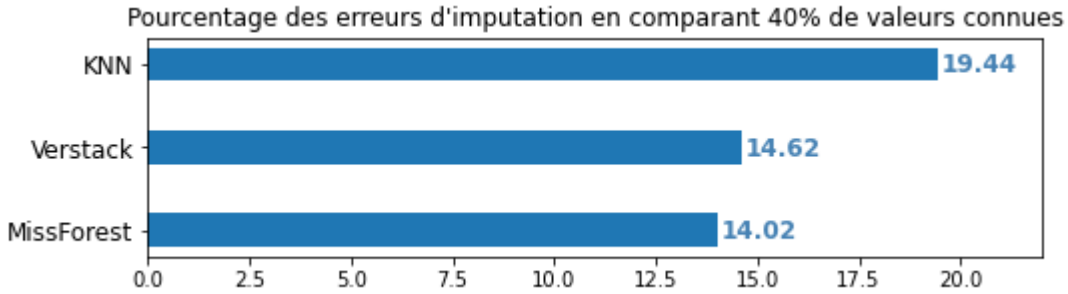
Outliers pour 100g ou ml de produit
énergie nutritionnelle > 900 kcal ou > 3766 kJ
nutriments _100g > 100g (ou 100ml)
nutriments _100g < 0g (ou 0ml)
le score du nutri-score au dessus de 40

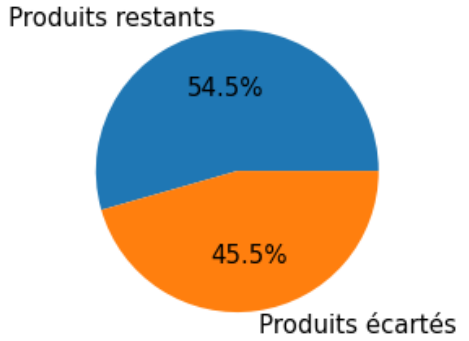
Macro-nutriment et sous-groupe de nutriment
si la masse totale des glucides inférieure à la masse de sucre
si la masse total des lipides est inférieure à la masse des acides gras saturés.
si la masse de sodium est supérieure à la masse de sel (en g).

Nettoyage – Valeurs manquantes

1 **Valeurs obligatoires** suppression des produits sans les valeurs obligatoires (application de santé)

2 **Imputations**





	Nombres de valeurs manquantes	% de valeurs manquantes
additives_en	176027	65.71000
fiber_100g	168919	63.06000
nova_group	119060	44.45000
additives_n	106190	39.64000
nb_ingredients_huile_palm	106190	39.64000
brands	51438	19.20000
image_url	14431	5.39000
saturated_fat_100g	144	0.05000

Pas d'imputation

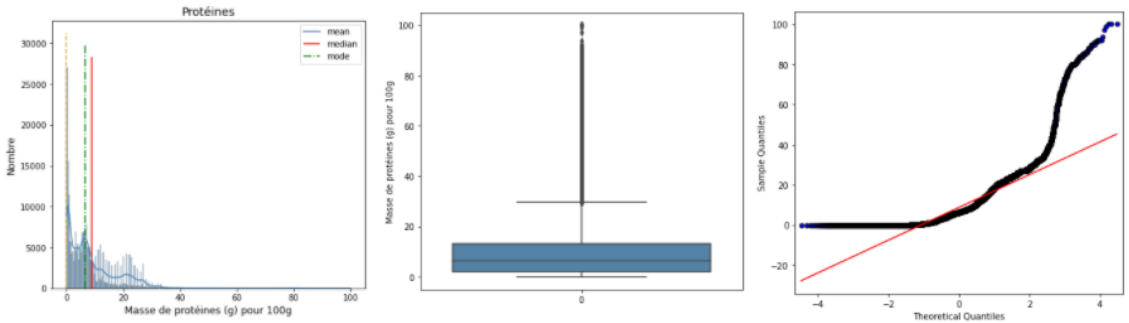
Imputation par 0

Imputation par MissForest

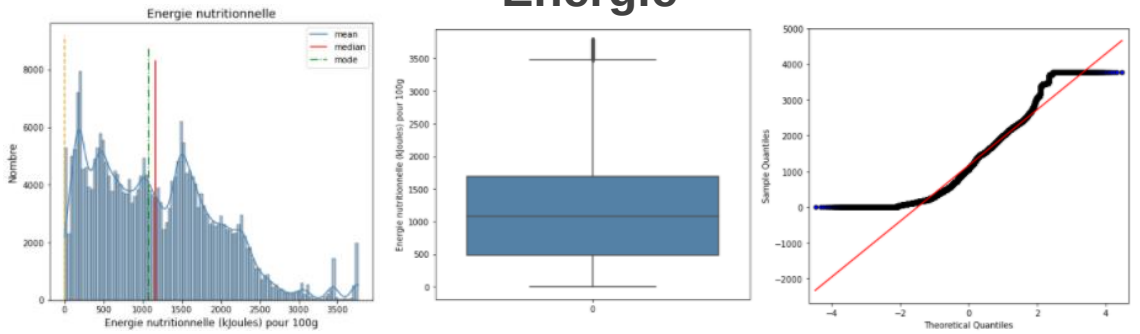
Sommaire

Analyse univariée – les nutriments

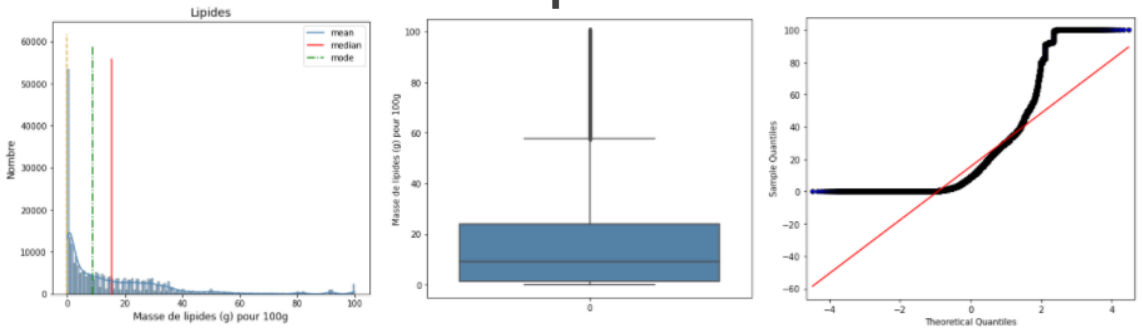
Protéines



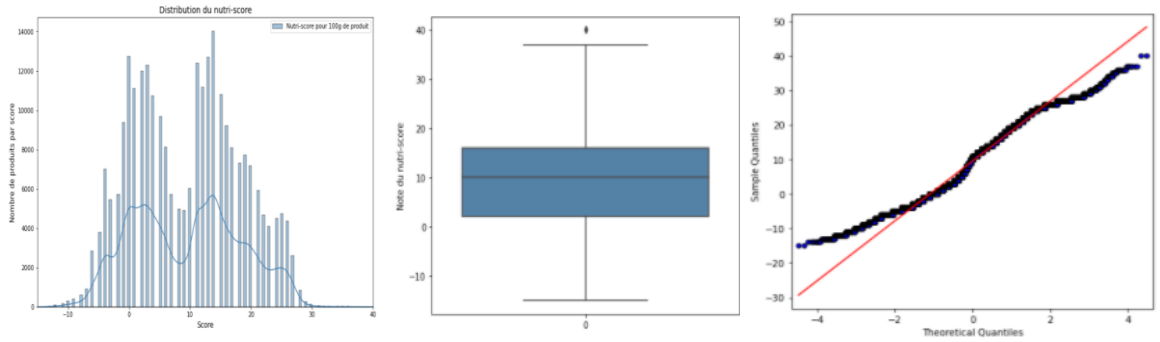
Energie



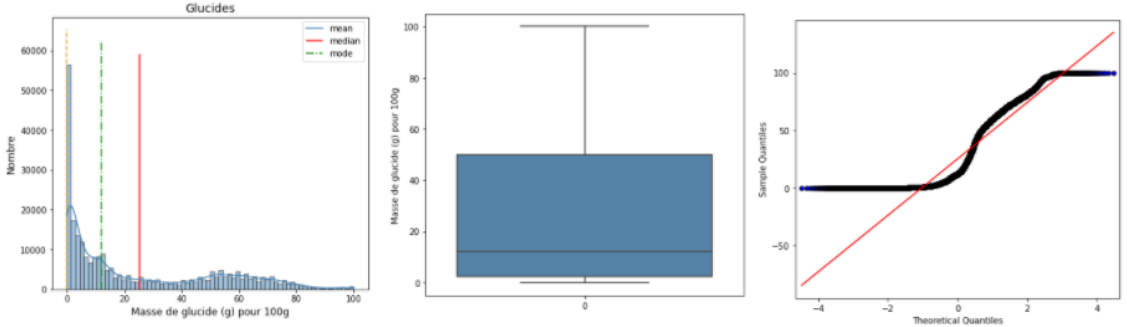
Lipides



Nutri-score



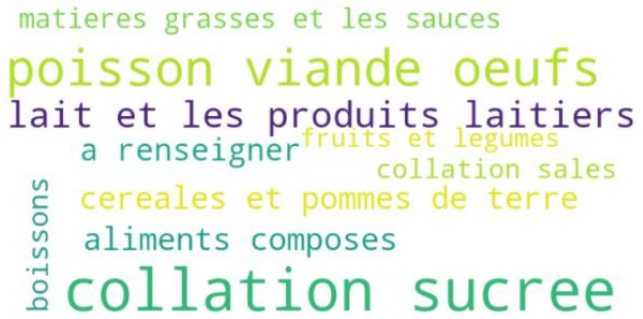
Glucides

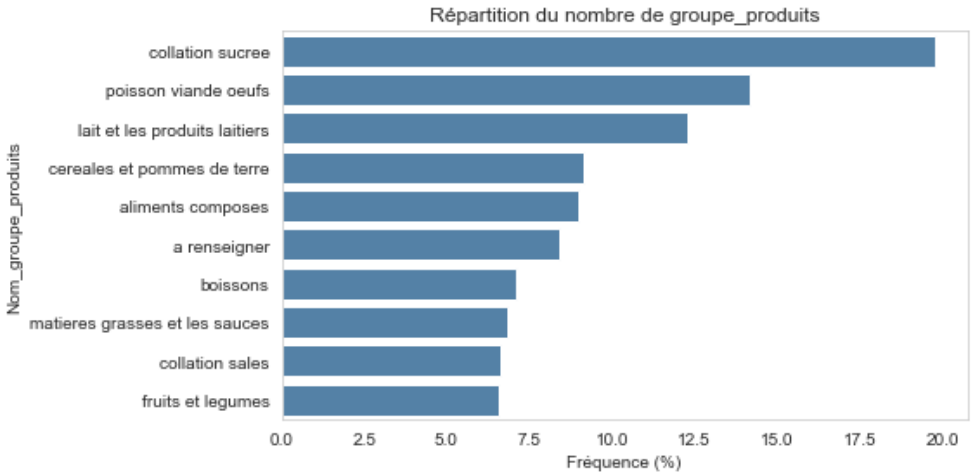


Desc	fat_100g	saturated_fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	sel	energie	nutriscore_score
mean	15.42544	5.89300	25.20521	12.67856	1.05942	8.81753	1.02137	1168.11084	9.50895
median	9.00000	2.30000	12.00000	3.30000	0.00000	6.40000	0.60000	1079.00000	10.00000
var	352.80405	73.00647	711.97937	347.02606	9.13648	79.57175	5.69719	635797.68750	76.91982
std	18.78308	8.54438	26.68294	18.62864	3.02266	8.92030	2.38688	797.36920	8.77039
skew	2.06164	3.22234	0.81542	1.86204	8.10557	1.92781	17.69135	0.65176	0.10425
kurtosis	5.30530	19.15100	-0.69722	3.04304	130.59885	7.77652	519.88153	0.11710	-0.93296
mode	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 0.0	0 14.0
Min	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	-15.00000
Max	100.00000	100.00000	100.00000	100.00000	100.00000	100.00000	100.00000	3766.00000	40.00000

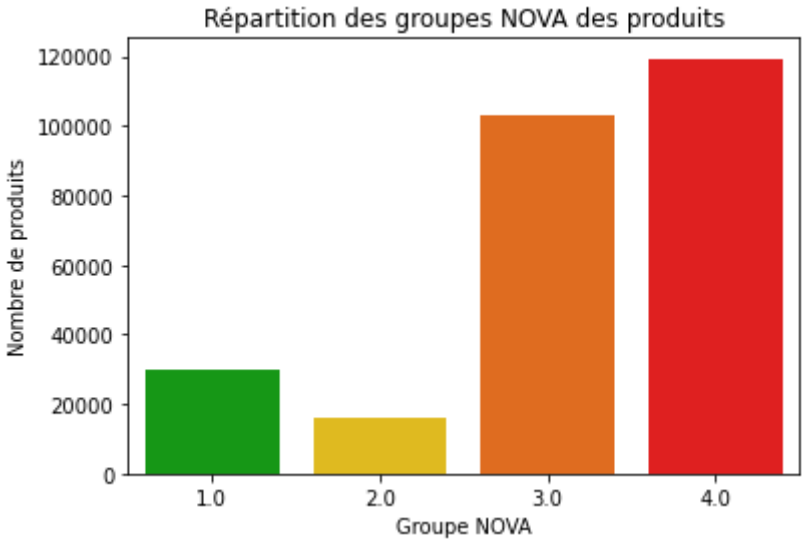
Analyse univariée – les produits

Groupe de produits

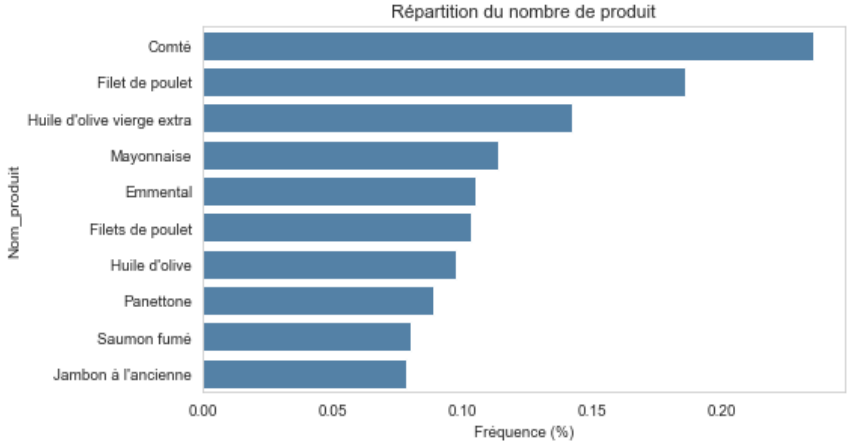




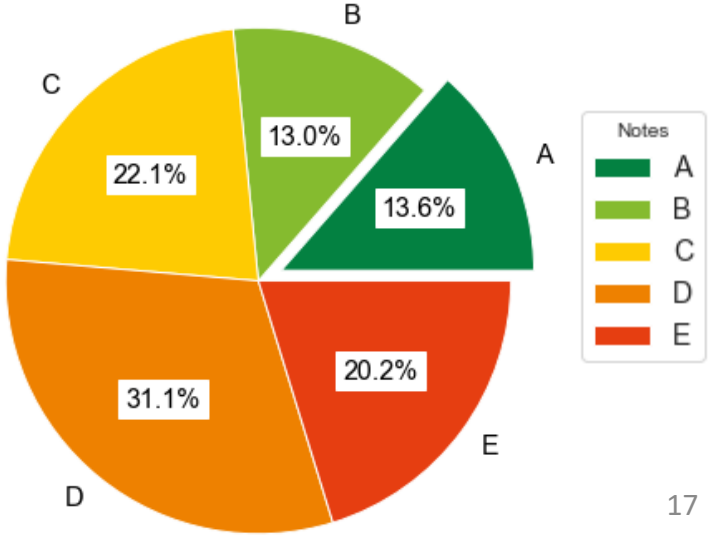
Groupe Nova



Produits

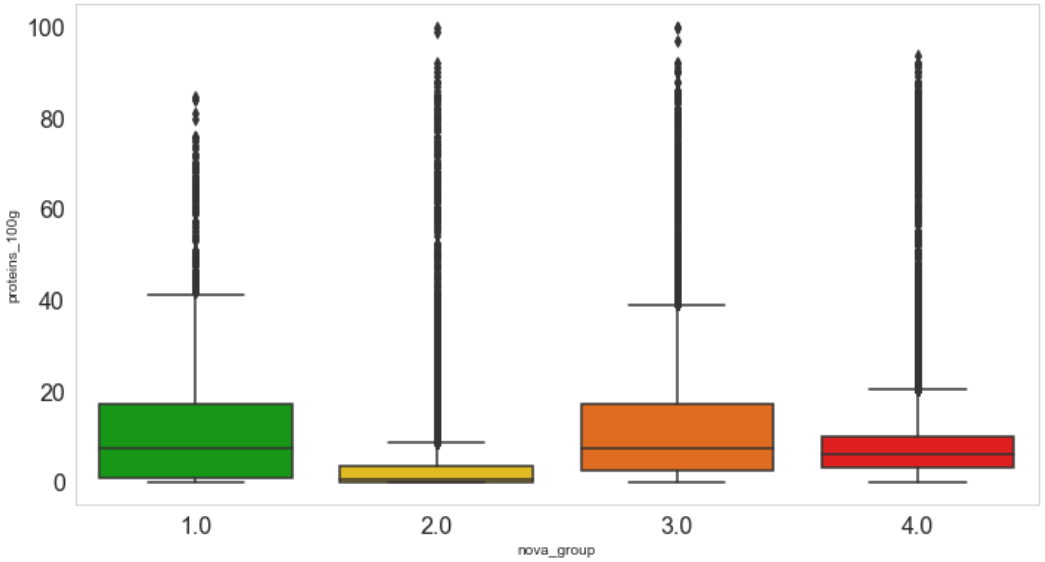


Note du nutri-score

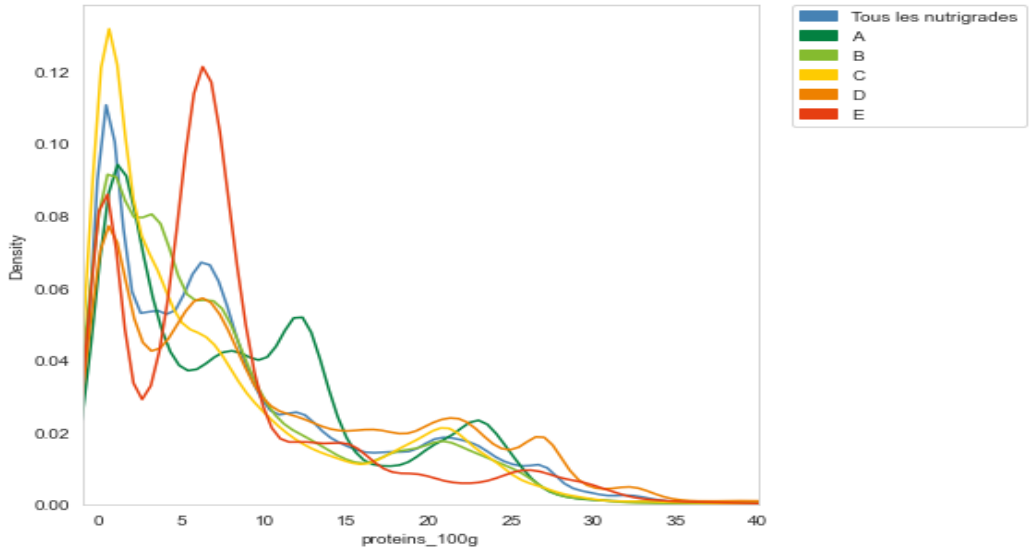


Analyse bivariée - protéines

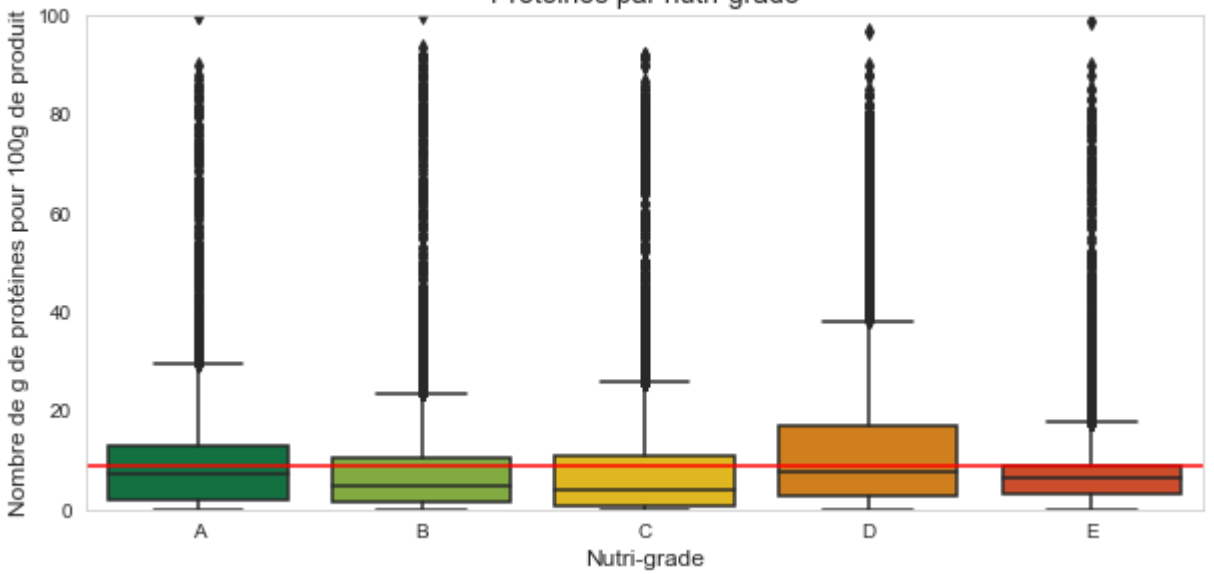
Groupe Nova

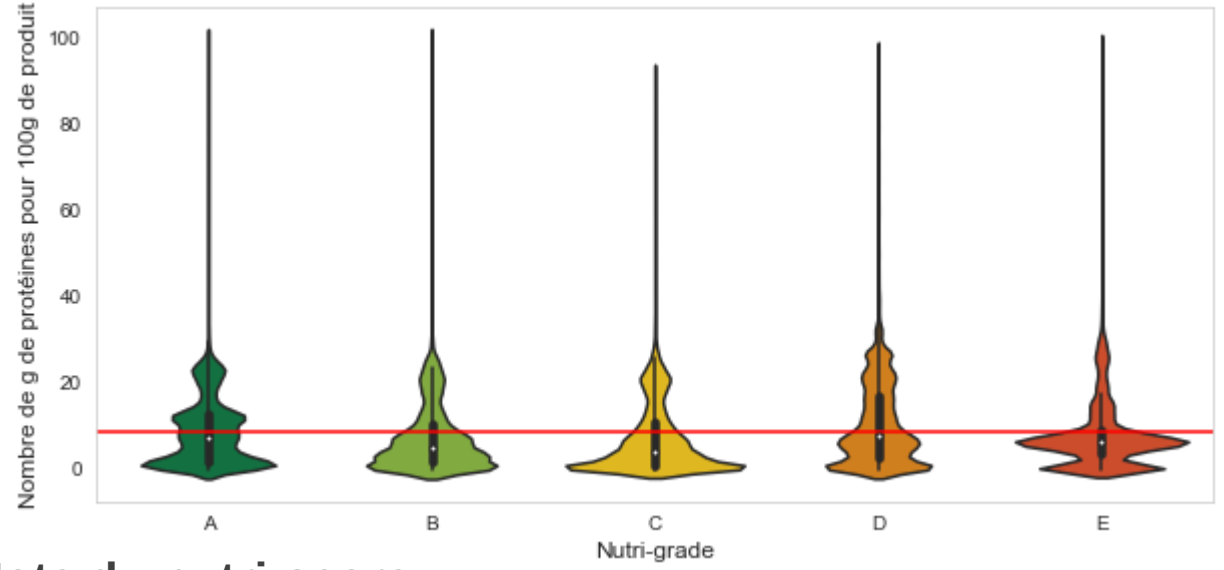


Distribution des protéines



Protéines par nutri-grade

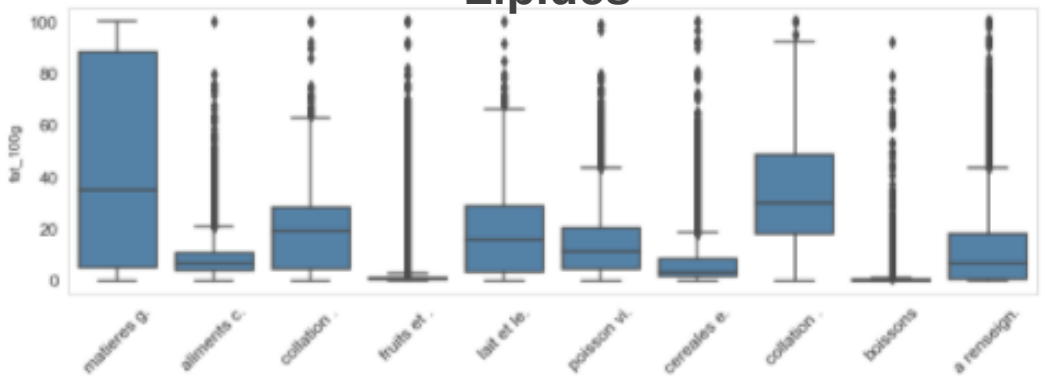




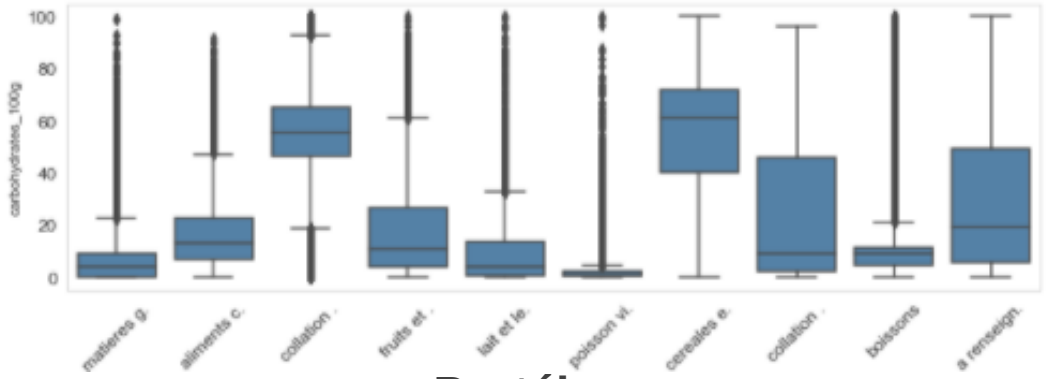
Note du nutri-score

Analyse bivariée - produits

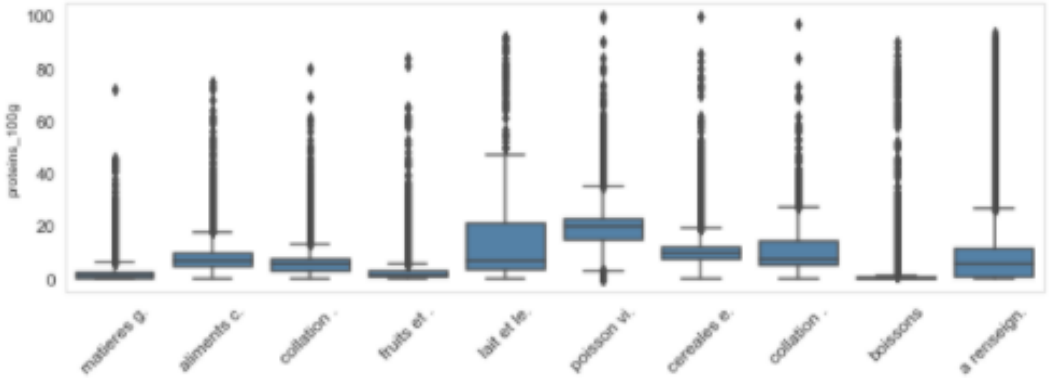
Lipides



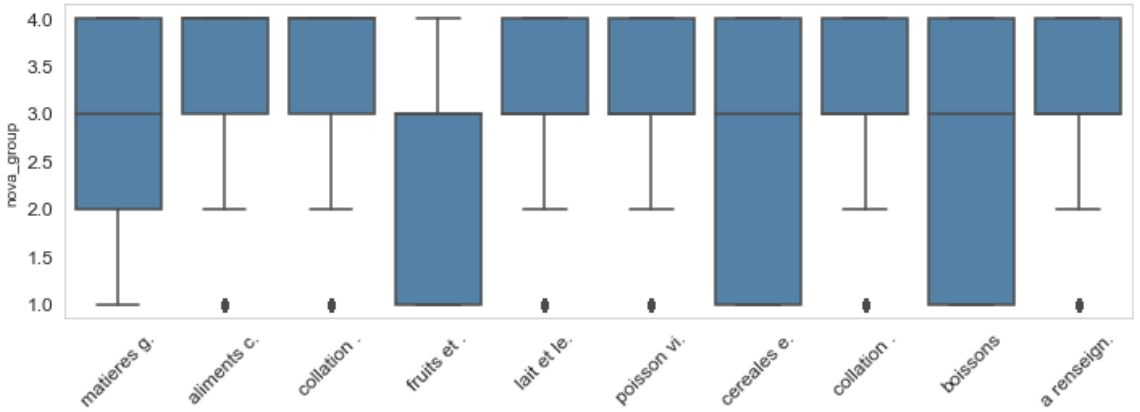
Glucides



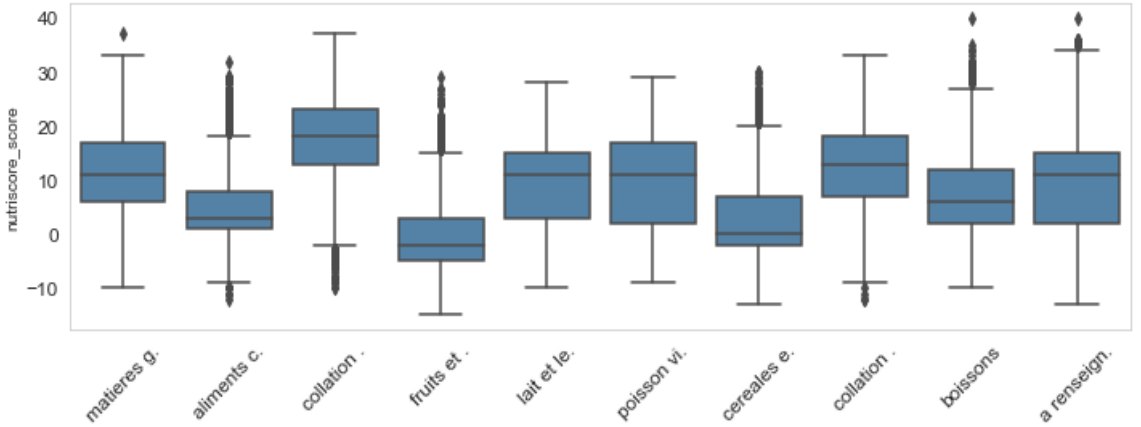
Protéines



Groupe Nova

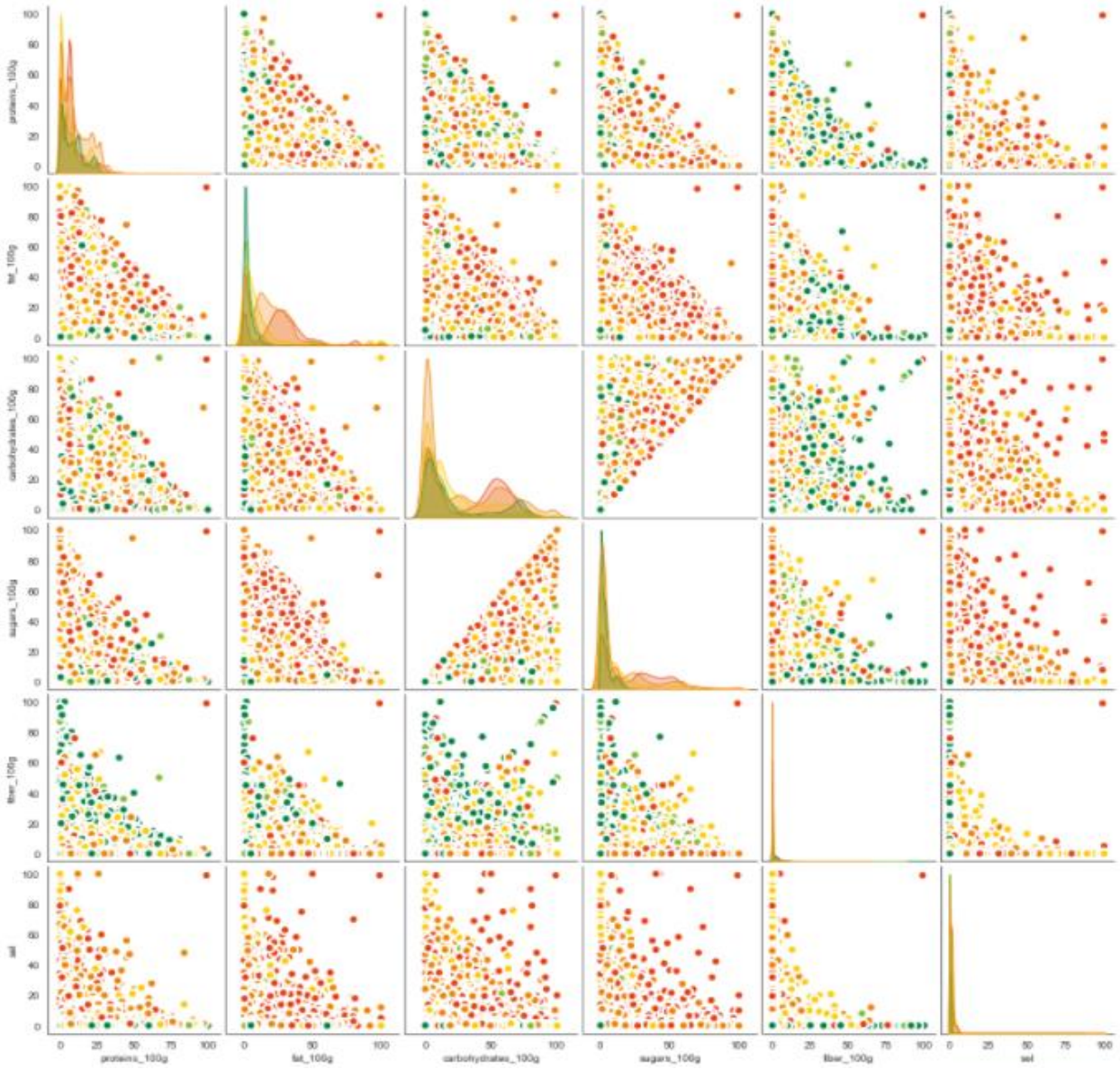


Nutri-score

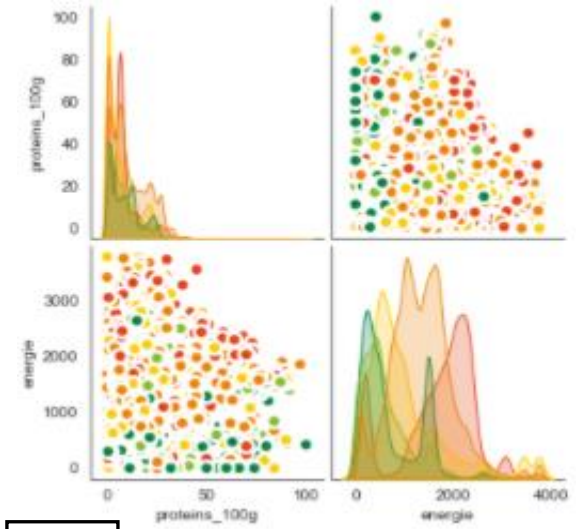


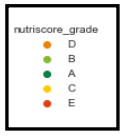
Analyse multivariée

Nutriments

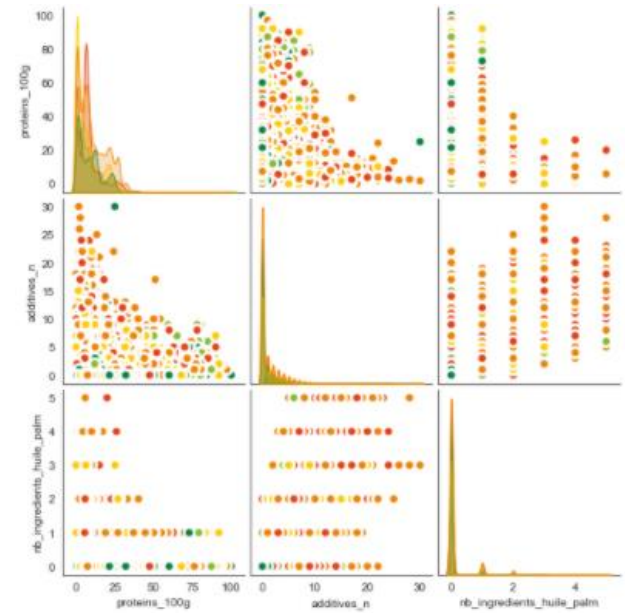


Énergie

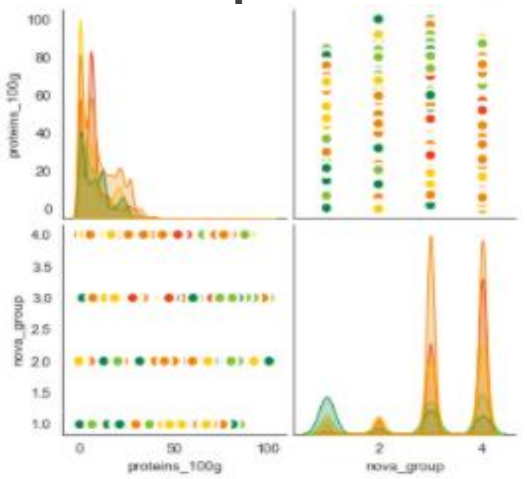




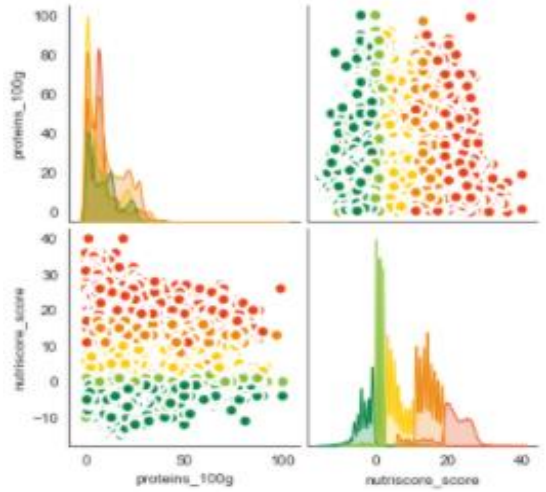
Additifs



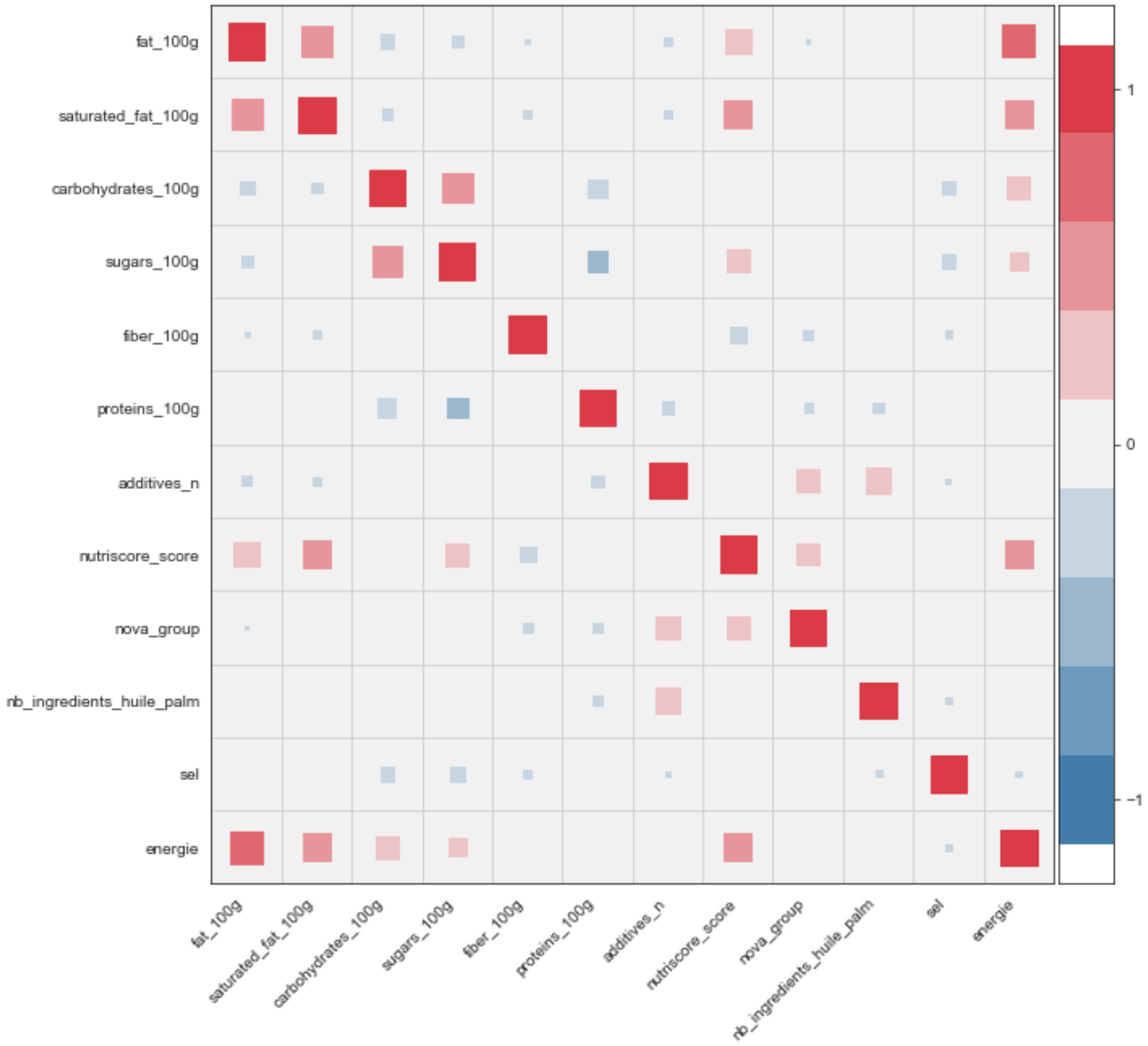
Groupe Nova



Nutri-score



Analyse multivariée – matrice de corrélation



Coef. corrélation

	proteins_100g
fat_100g	0.109967
saturated_fat_100g	0.144397
carbohydrates_100g	-0.238516
sugars_100g	-0.308174
fiber_100g	0.061575
proteins_100g	1.000000
additives_n	-0.100706
nutriscore_score	0.065837
nova_group	-0.046566
nb_ingredients_huile_palm	-0.071861
sel	0.177769
energie	0.148193

Sommaire

4 Faisabilité de l'application

Le nettoyage et l'analyse exploratoire ont permis de vérifier que les variables sont disponibles en quantité avec qualité

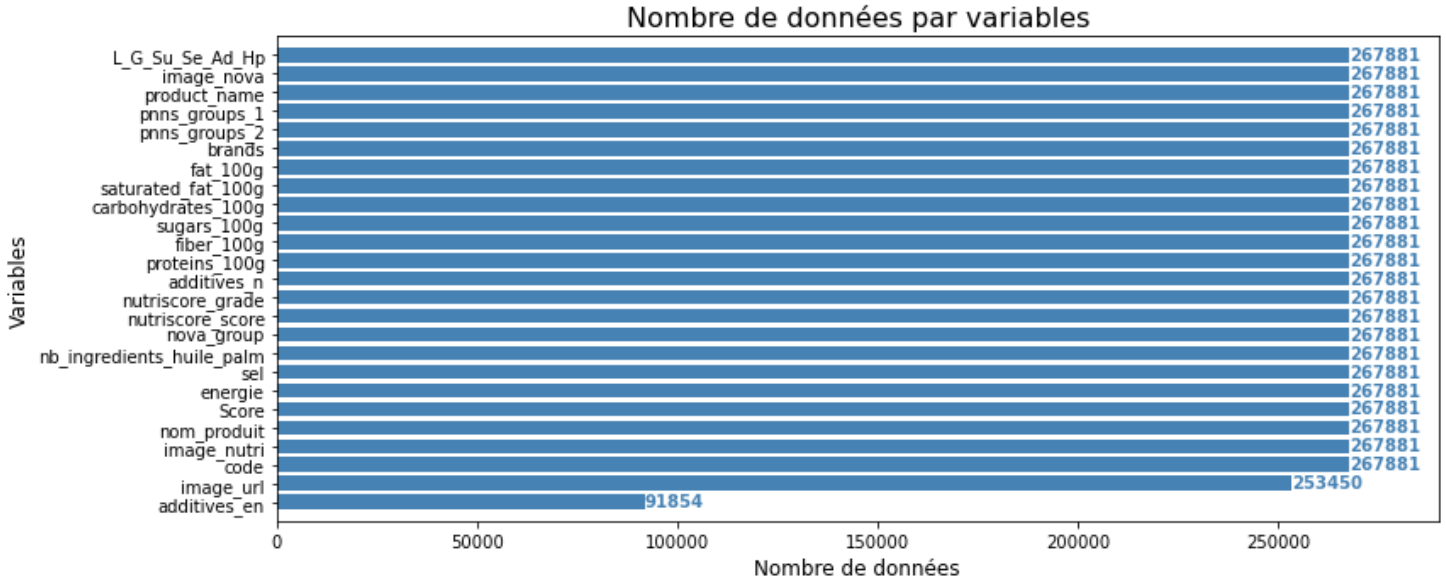
Régime hypoprotidique
- proteins_100g

Alimentation saine
- nutriscore_score
- nutriscore_grade
- sel
- sugars_100g
- additives_n
- ingredients_from_palm_oil_n
- energie

Alimentation équilibrée
- fat_100g
- carbohydrates_100g
- fiber_100g

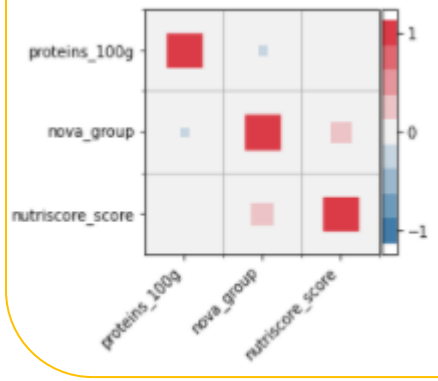
Peu transformés
- nova_group

Produits disponibles
- product_name
- brands
- image_url



Moteur de recommandation réalisable

Variables de scoring



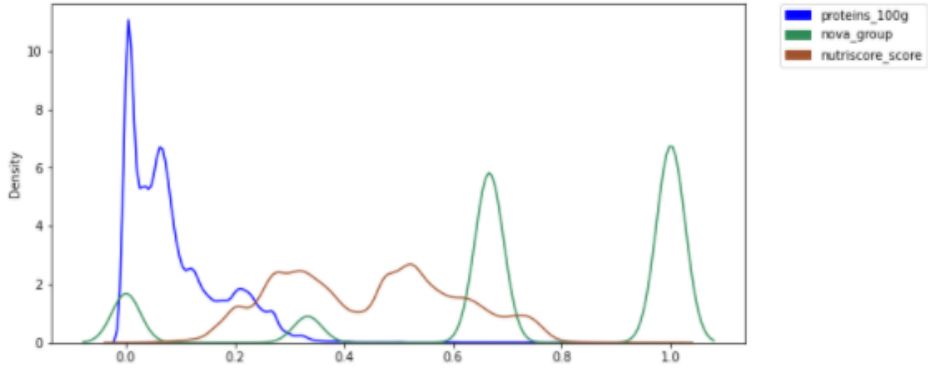
Nature	Pondération
protéine	50
nutri-score	2
groupe NOVA	2

4 Moteur de recommandation - Score

1. **Sélection des variables de scoring :**
   protéines, nutri-score et groupe nova

2. **Mise à l'échelle :**
   MinMaxScaler()

3. **Scoring avec pondération :**
   Protéines : 50 – Nutri-score et groupe Nova : 2



4 Moteur de recommandation

1

Pré-traitement des noms de produit:

Pré-processing : clean() de texthero

Suppression stopwords français : NLTK

2

Vectorisation :

TfidfVectorizer de scikit-learn

3

DataFrame de comparaison :

Features engineering (renommer variables, ajouter images, ajouter variables L_G_Su_Se_Ad_Hp)

4

Moteur de recommandation :

Vérification de la saisie, suppression des ponctuations, suppression des stopwords, similarité des cosinus scikit-learn, tri des produits similaires et récupération des n produits trier par score descendant

```
# test du pré-moteur
produit = input(str('Quel est le produit recherché ? : '))
rech_produits(produit)
```

```
Quel est le produit recherché ? : riz
Produits recommandés :
Remarque : L_G_Su_Se_Ad_Hp : lipides - Glucides - Sucre - Sel - Additif - huile palme
```

Produit	Photo	g de protéines/100g prod	Note_nutri_score	Nutri_score	Groupe_NOVA	Marque	L_G_Su_Se_Ad_Hp
Riz		2.8	B	0.0	1		0.4-27.5-0.2-0.0-0.0-0.0
riz		5.0	B	-2.0	2		1.0-4.0-2.0-0.0-0.0-0.0
Riz		6.6	B	0.0	1	Özde	1.0-78.0-0.4-0.01-0.0-0.0
Riz		7.0	B	-1.0	1	Carrefour	0.5-77.0-0.2-0.0-0.0-0.0
3 riz		7.6	B	-1.0	1	Lustucru	1.4-78.0-0.4-0.0-0.0-0.0

Sommaire

5 Conclusion

OPEN FOOD FACTS

données nettoyés, imputées et analysées

Santé

Basée sur 4 variables : protéines, nutri-score, groupe Nova et nom du produit

Limites

- **Pondération** protéines/nutri-score/groupe Nova
- **Moteur de recommandation** : fiabilité de la similarité, tests
- **Fiabilité** de la saisie

Amélioration

- Produits vendus hors de France (vacances, voyages scolaires)
- Conseils des **médecins/diététiciens**
- Ajout du **nombre de part** de phénylalanine pour chaque aliment
- Magasin où **trouver** le produit **proche de chez soi**
- Consulter les **associations** de personnes atteintes de phénylcétonurie

Prolongement

- Bilan protéinique journalier/mensuel/ calcul des parts
- Scanner pour les enfants avec logos interdits, autorisés

Autres MHM

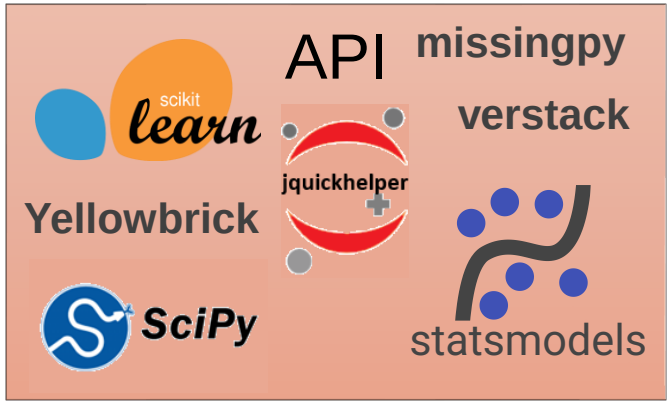
- **Protéines** : aminoacidopathies, troubles du cycle de l'urée, aciduries organiques), tyrosinémie, homocystinurie -> régime hypoprotidique.
- **Sucre** : troubles de la glycolyse et de la néogrucogénèse, glycogénoses) -> régime contrôlé en sucre.
- **Lipides** : déficit de l'oxydation des acides gras -> régime en lipides.
- **Glucides** : déficit en PDH, déficit en Glut1, cytopathies mitochondriales ->régime cétogène (glucides remplacés par lipides).

5 Boîte à Outils

Environnement

Librairies de base

API



Visualisation

Outil de bureau