October 30, 2018

Analysez des données nutritionnelles

Maurice clère

Objectifs

1. Problématique & OpenFoodFacts
2. Présentation du nettoyage effectué
3. Analyse exploratoire
4. Pistes de modélisation
5. Q&A

Problématique

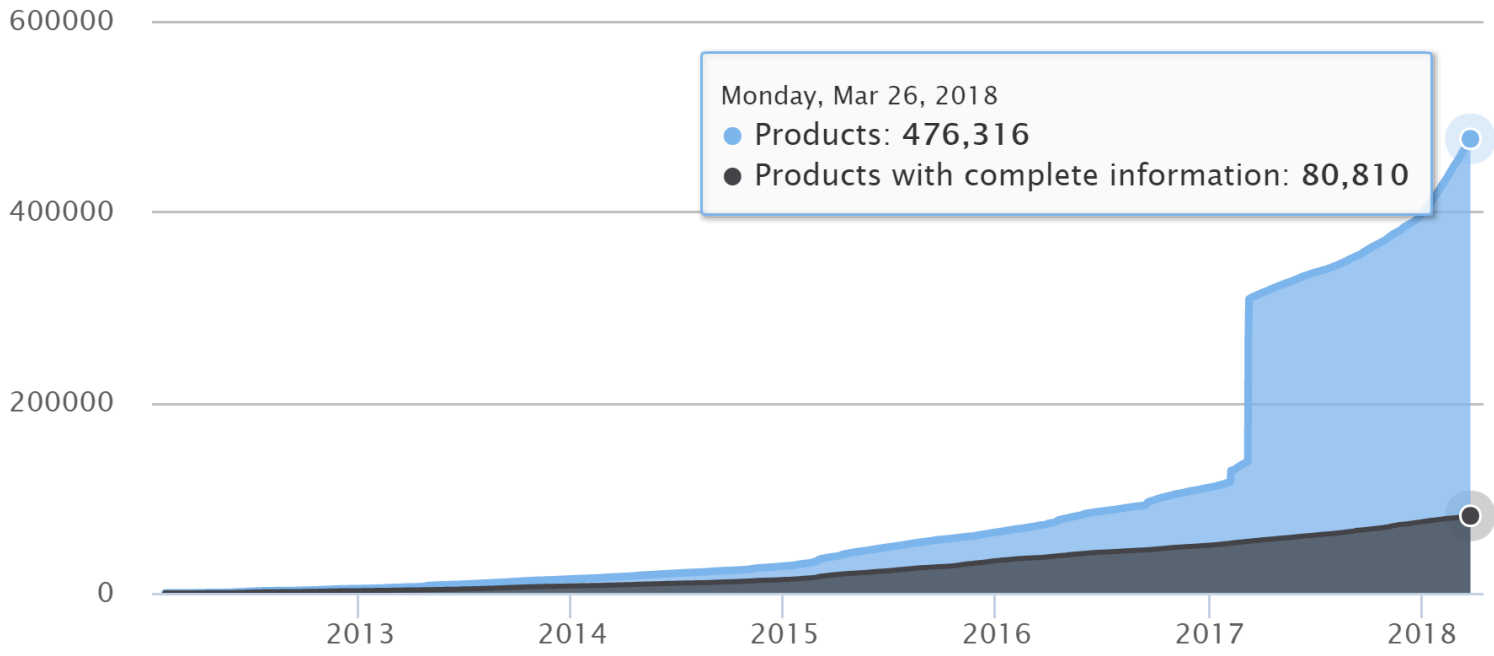
- Utilisé la base de données OpenFoodFact comme socle informatif pour générer des recettes saines.

Open Food Facts

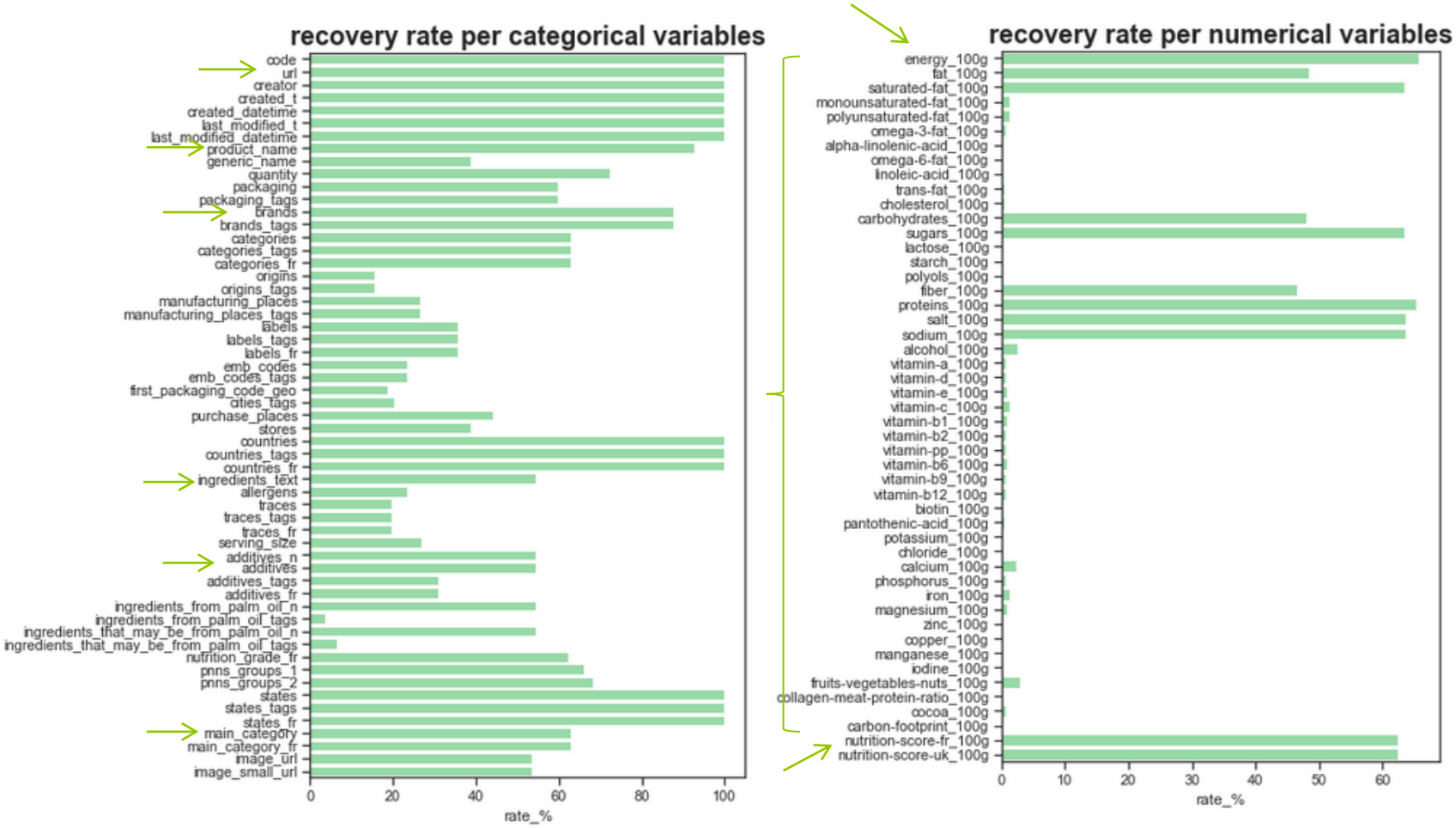
- le projet a débuté en mai 2012 sous l'impulsion de Stéphane Gigandet
- Une base de produits alimentaires
- Faite par tout le monde
- Pour tout le monde
- **Utilisation de licences libres**

La base de donnée

- Fichier: fr.openfoodfacts.org.products.csv
- 162 variables x 98440 produits (en France)



Les Variables



Sélection des variables

- Numériques:
  - toutes les variables avec un taux de recouvrement supérieur à 0.1%.

- Catégorical:
  1. **code:** indentification du produit (unique)
  2. **url:** lien sur la fiche produit du site officiel d'openFoodFacts
  3. **product_name:** nom du produit
  4. **Brands:** marque
  5. **categories_fr:** liste de mots clé (soupes, biscuit, etc…)
  6. **countries_fr:** pays dans lequel le produit est répertorié
  7. **ingredients_text**: liste d'ingrédients
  8. **ingredients_from_palm_oil_n:** huile de palme
  9. **additives_fr:** liste des additifs
  10. **main_category_fr:** catégorie qui représente le mieux le produit
  11. **nutrition_grade_fr:** notation (A, B, C, D, E)

Nutri-score

Sucres (g)

Energie (kJ)

Sodium (g)

Graisses saturées (g)

Fibres (g)

Protéines (g)

Fruits, légumes et noix (%)

Nettoyage

- Outliers
- Valeurs incohérentes
- Valeurs manquantes

	min	max
energy_100g	0.0	3251373.0
fat_100g	0.0	380.0
saturated-fat_100g	0.0	210.0
carbohydrates_100g	0.0	190.0
sugars_100g	-0.1	105.0
fiber_100g	0.0	178.0
salt_100g	0.0	211.0
vitamin-b1_100g	0.0	161.0
carbon-footprint_100g	0.0	2520.0
nutrition-score-fr_100g	-15.0	40.0
nutrition-score-uk_100g	-15.0	36.0

Nettoyage

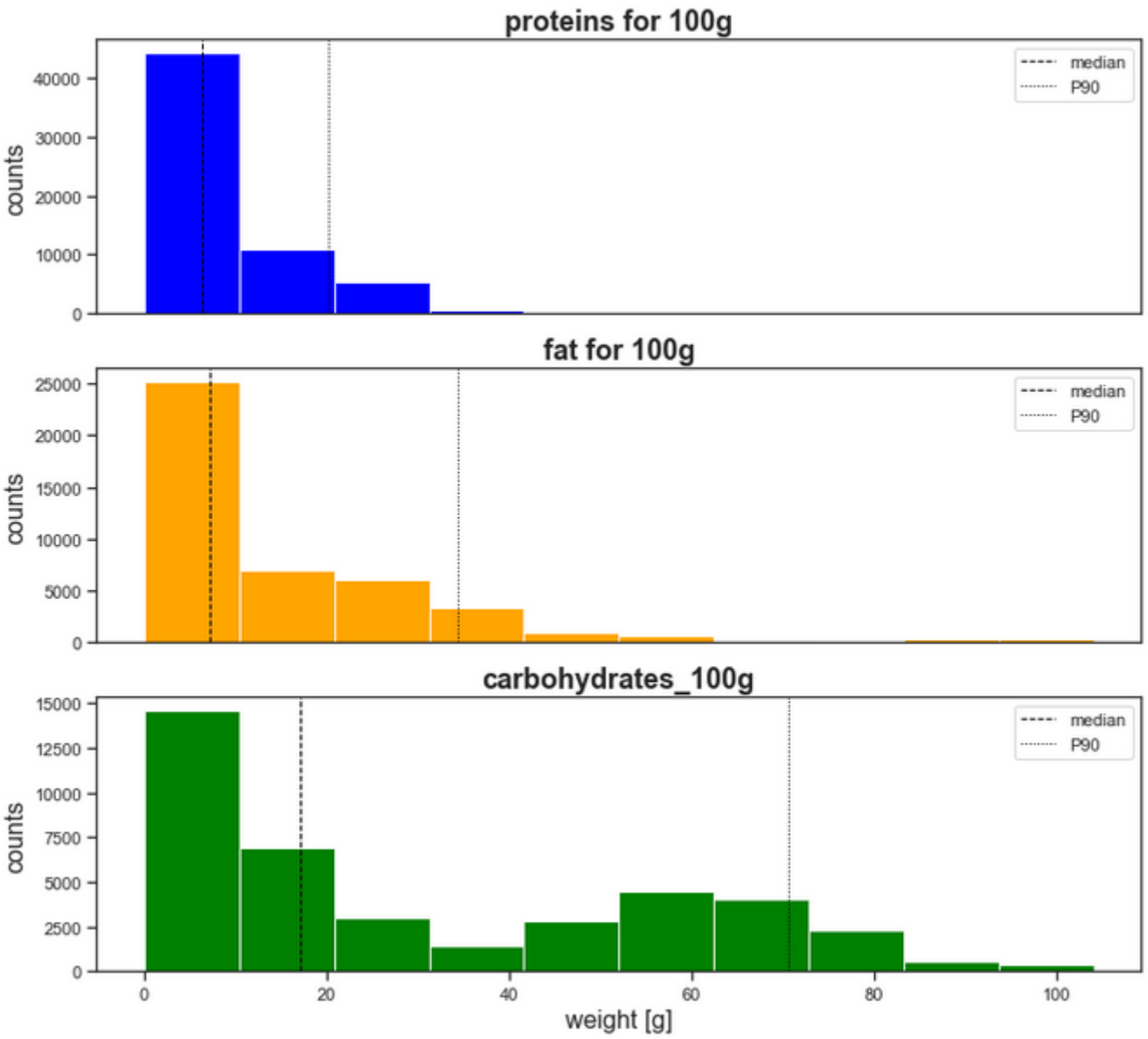
type	variables	Min	max	action	nombre d' occurrences
Outliers	toutes les valeurs pour 100g	0	110*	value set to None	10
Outliers	Energy_100g	0	3700**	value set to None	246
Incoherence	somme des macronutriments sur 100g	-	110*	produit retiré	69
Incoherence	somme des Graisses / 100g	-	110*total Graisse	value set to None	89
Incoherence	somme des Glucides / 100g	-	110*total Glucide	valeur set to None	3758
Valeur absentes	nutri-score	-	-	produit retiré	37020
Valeur absentes	product_name	-	-	produit retiré	343
Doublon	Code	-	-	un doublon retiré	0

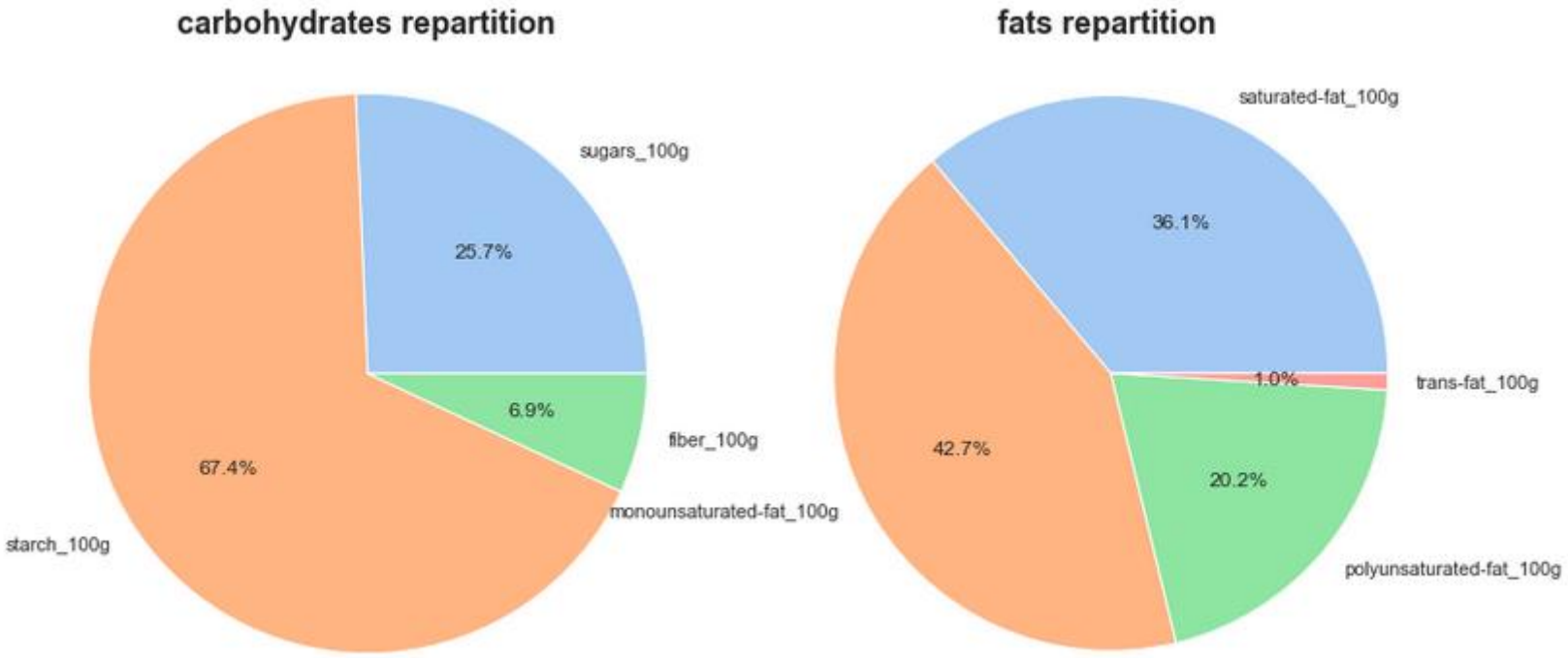
* 110 = limite haute + 10% car certains produits sont en ml
** 3700 = 37kj/g*100g (teneur maximum sur un produit composé à 100% de graisse)

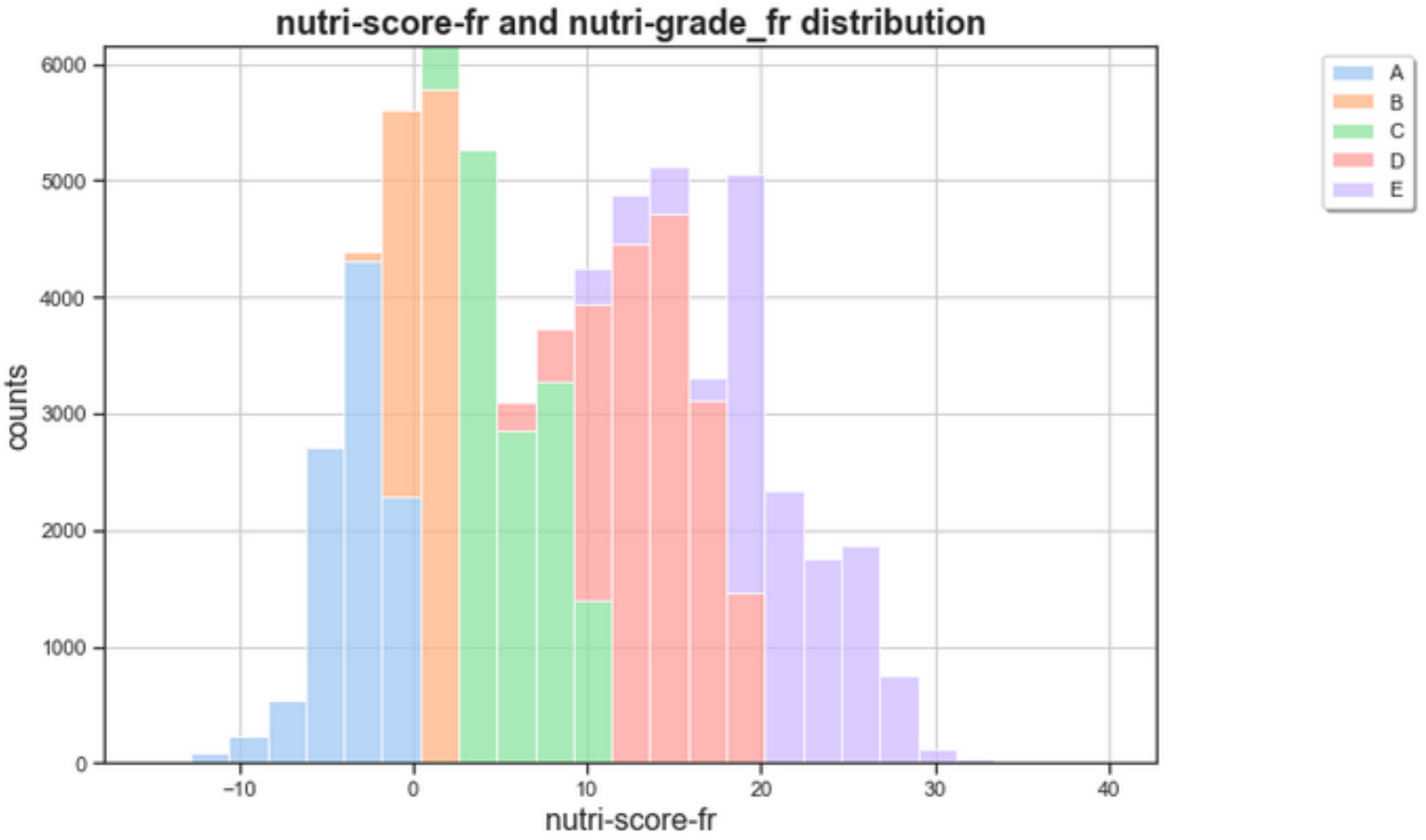
Autres

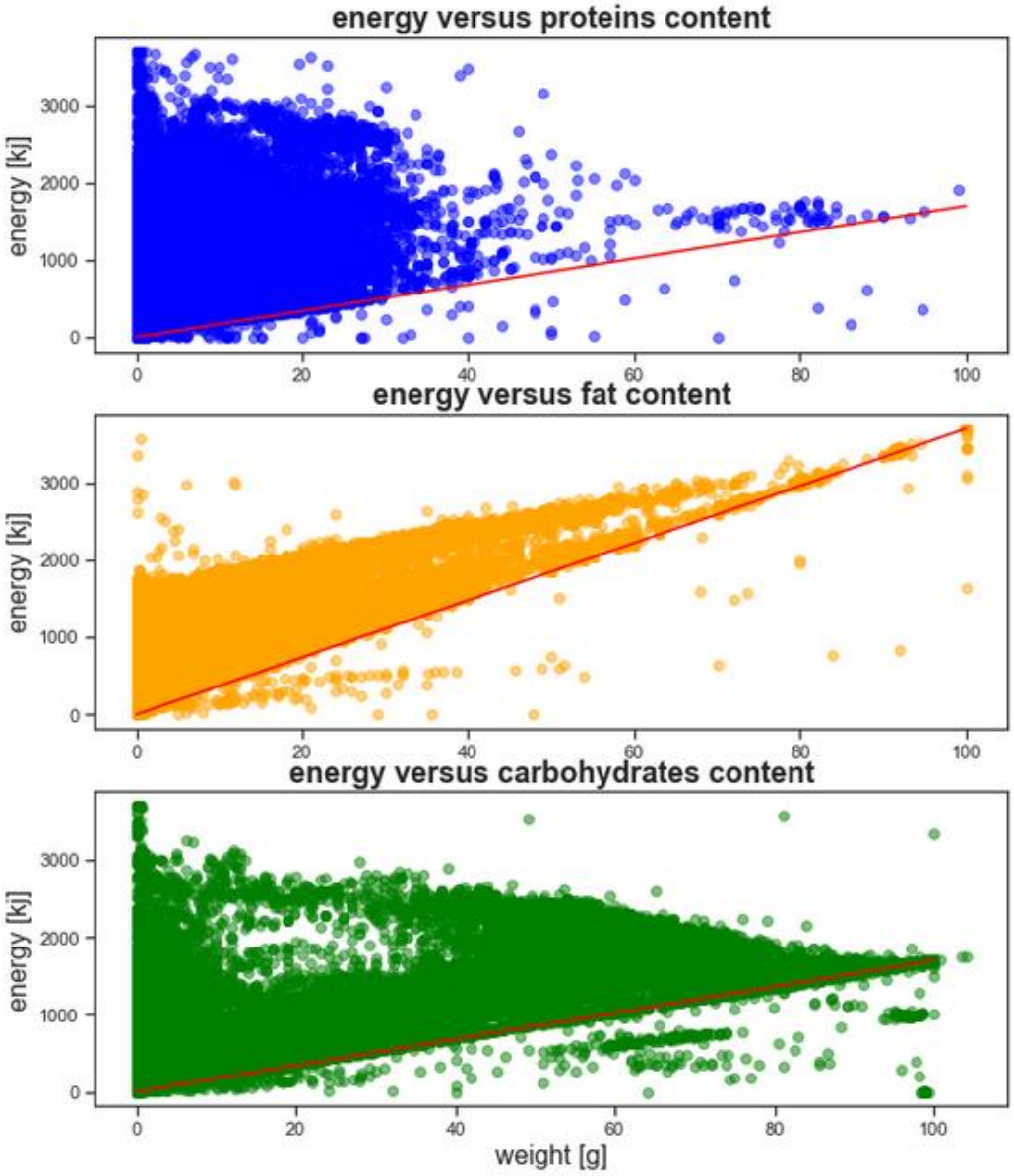
type	variables	action
formatage	Toutes les valeurs des variables catérogical	Changer en minuscule
formatage	Nom de la variable ‘energy_100g’	Renomme ‘energy_kj_100g’
feature engineering	’energy_Kcal_100g’	Energy convertie en Kcal
feature engineering	’check_energy_kj_100g’	Energy recalculer a partir des macronutriments

Analyse exploratoire


proteins for 100g
median
P90
counts
40000
30000
20000
10000
0
fat for 100g
median
P90
counts
25000
20000
15000
10000
5000
0
carbohydrates_100g
median
P90
counts
15000
12500
10000
7500
5000
2500
0
0
20
40
60
80
100
weight [g]


carbohydrates repartition
sugars_100g
25.7%
6.9%
fiber_100g
67.4%
starch_100g
monounsaturated-fat_100g
fats repartition
saturated-fat_100g
36.1%
1.0%
trans-fat_100g
42.7%
20.2%
polyunsaturated-fat_100g

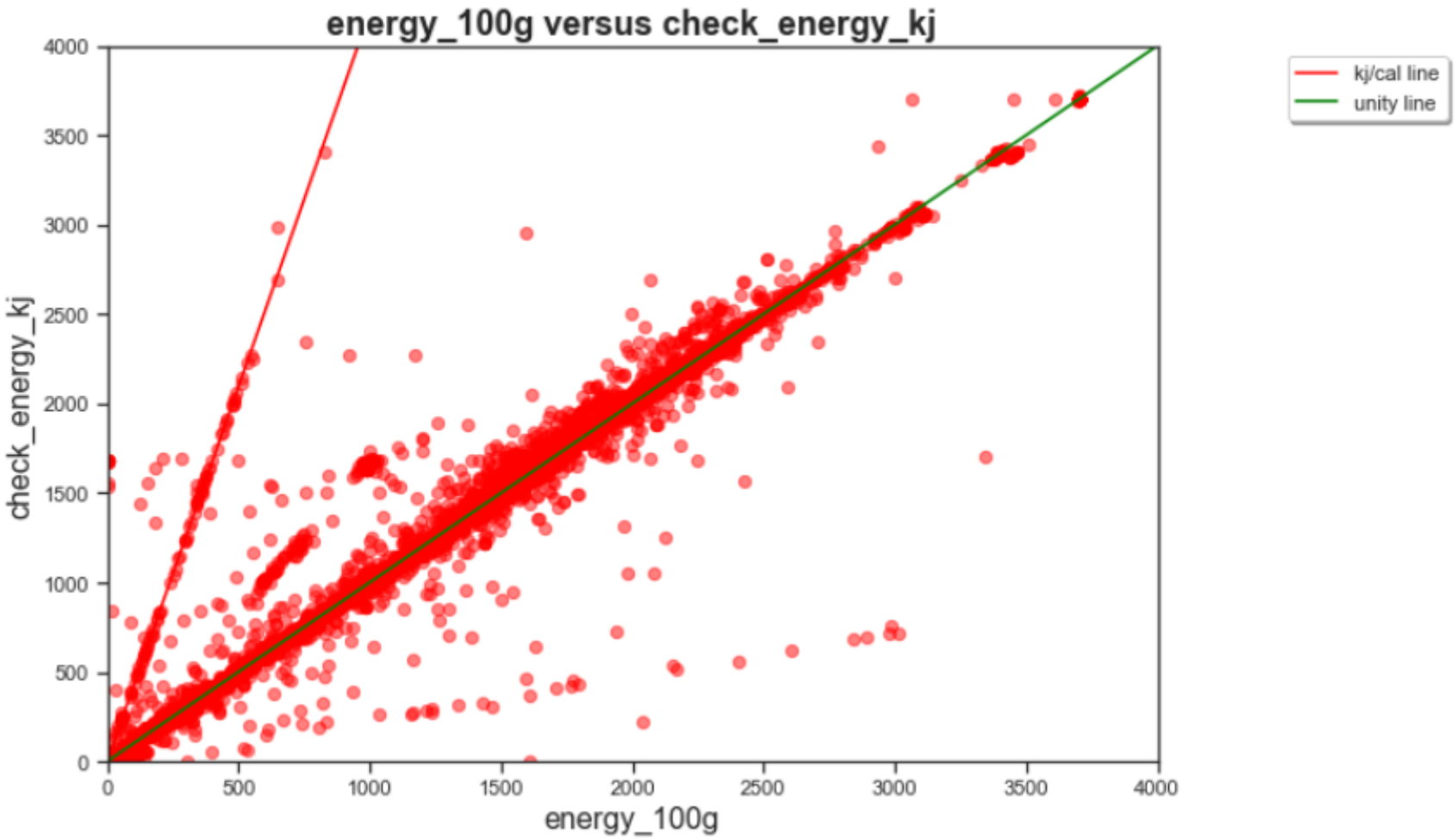

nutri-score-fr and nutri-grade_fr distribution
counts
nutri-score-fr
A
B
C
D
E
6000
5000
4000
3000
2000
1000
0
−10
0
10
20
30
40

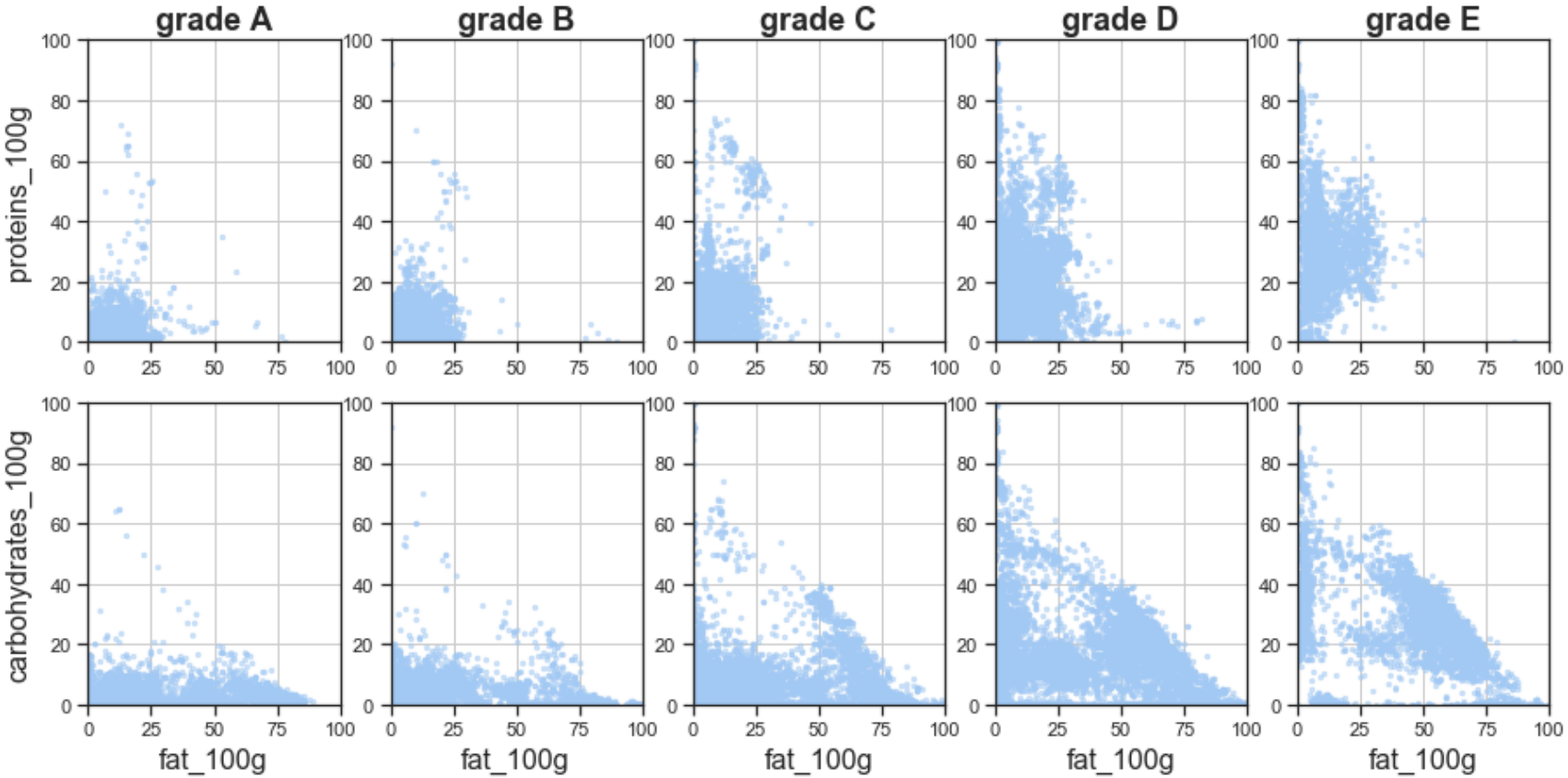


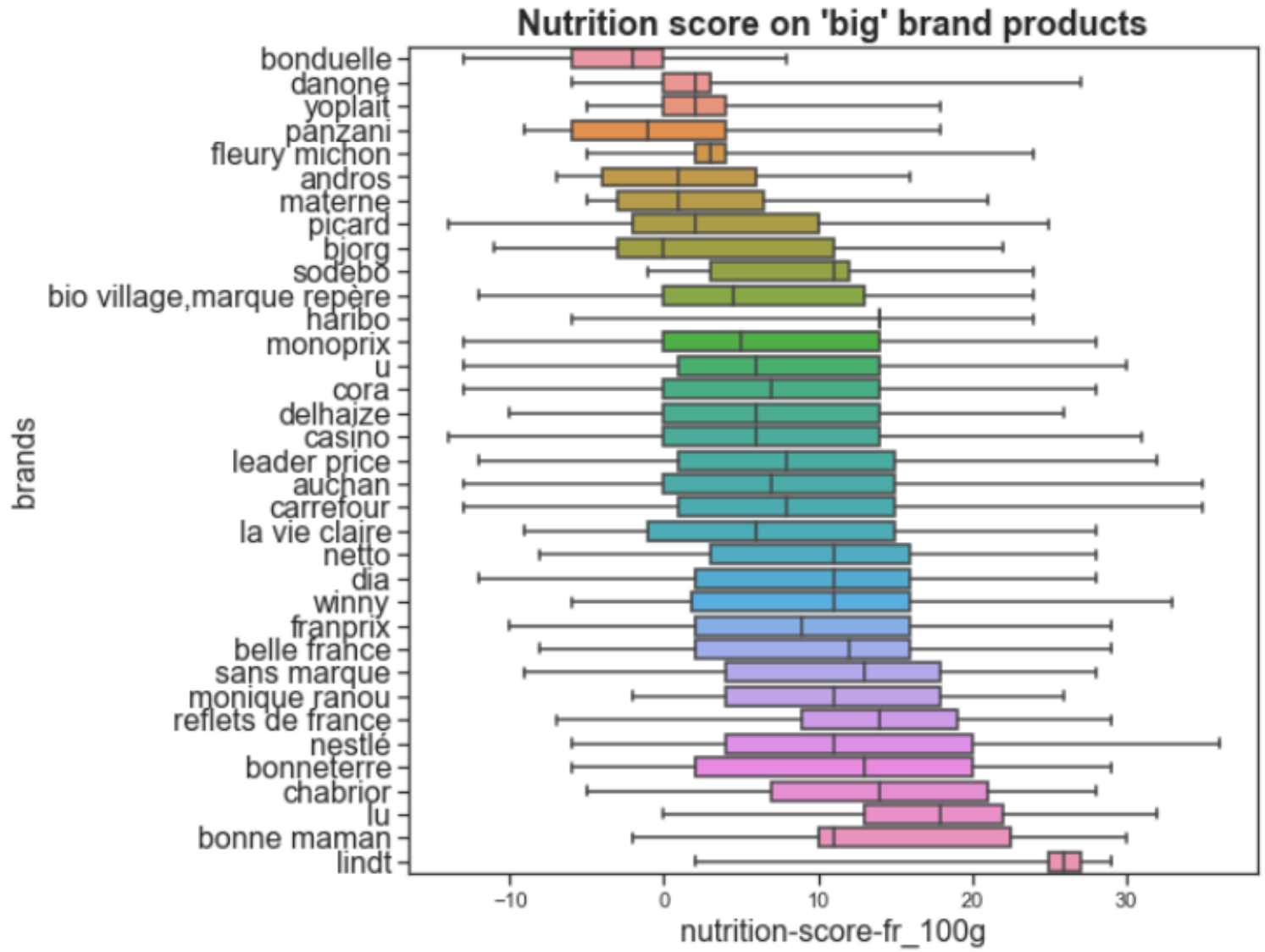
(1g protéines = 17kj)

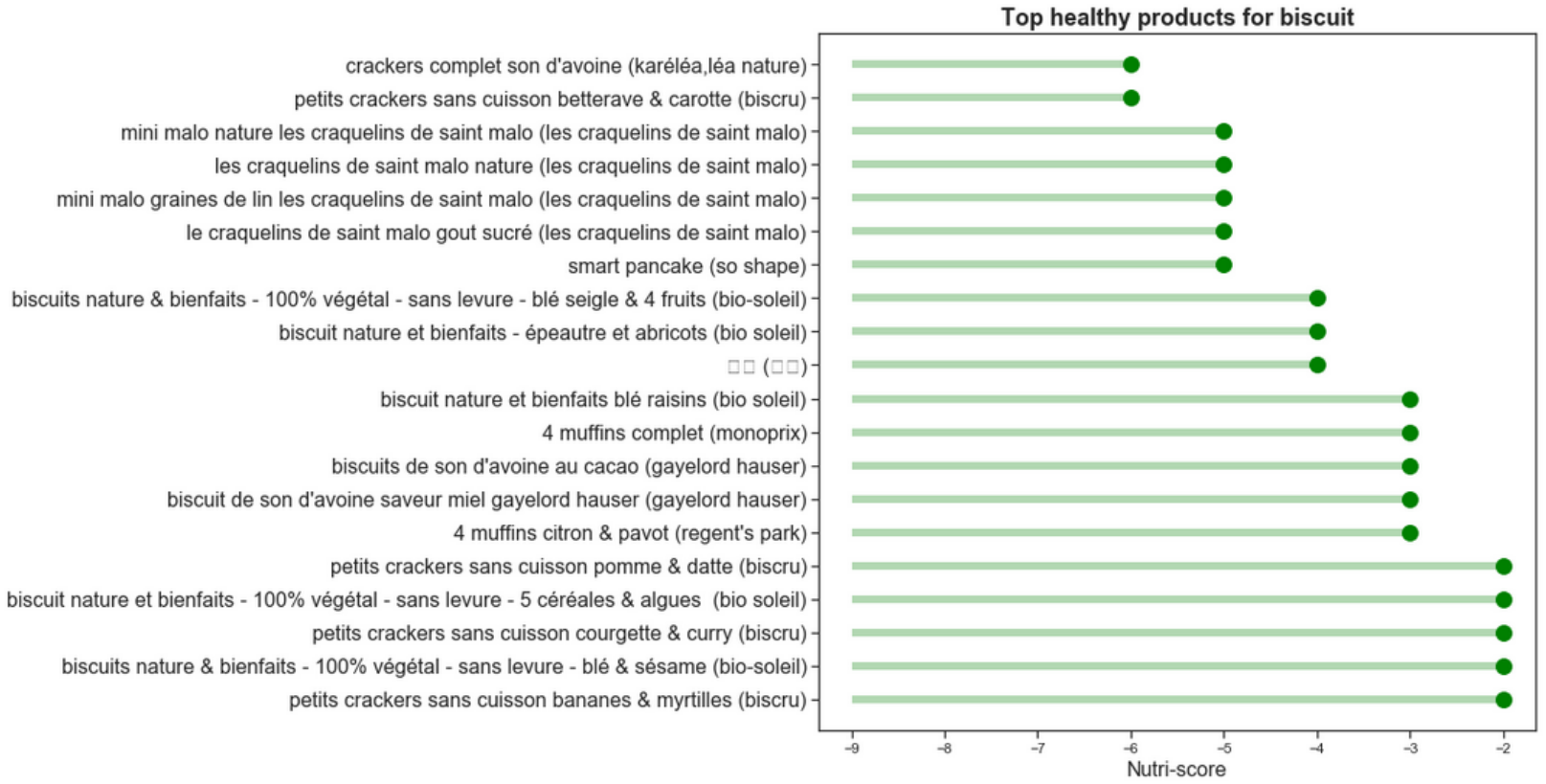
(1g graisse = 37kj)

(1g glucide = 17kj)


energy_100g versus check_energy_kj
kj/cal line
unity line
check_energy_kj
energy_100g
0
500
1000
1500
2000
2500
3000
3500
4000


grade A
grade B
grade C
grade D
grade E
proteins_100g
carbohydrates_100g
fat_100g
0
20
40
60
80
100
0
25
50
75
100


Nutrition score on 'big' brand products
brands
bonduelle
danone
yoplait
panzani
fleury michon
andros
materne
picard
bjorg
sodebo
bio village,marque repère
haribo
monoprix
u
cora
delhaize
casino
leader price
auchan
carrefour
la vie claire
netto
dia
winny
franprix
belle france
sans marque
monique ranou
reflets de france
nestlé
bonneterre
chabrior
lu
bonne maman
lindt
−10
0
10
20
30
nutrition-score-fr_100g


Top healthy products for biscuit
crackers complet son d'avoine (karéléa,léa nature)
petits crackers sans cuisson betterave & carotte (biscru)
mini malo nature les craquelins de saint malo (les craquelins de saint malo)
les craquelins de saint malo nature (les craquelins de saint malo)
mini malo graines de lin les craquelins de saint malo (les craquelins de saint malo)
le craquelins de saint malo gout sucré (les craquelins de saint malo)
smart pancake (so shape)
biscuits nature & bienfaits - 100% végétal - sans levure - blé seigle & 4 fruits (bio-soleil)
biscuit nature et bienfaits - épeautre et abricots (bio soleil)
□□ (□□)
biscuit nature et bienfaits blé raisins (bio soleil)
4 muffins complet (monoprix)
biscuits de son d'avoine au cacao (gayelord hauser)
biscuit de son d'avoine saveur miel gayelord hauser (gayelord hauser)
4 muffins citron & pavot (regent's park)
petits crackers sans cuisson pomme & datte (biscru)
biscuit nature et bienfaits - 100% végétal - sans levure - 5 céréales & algues (bio soleil)
petits crackers sans cuisson courgette & curry (biscru)
biscuits nature & bienfaits - 100% végétal - sans levure - blé & sésame (bio-soleil)
petits crackers sans cuisson bananes & myrtilles (biscru)
−9
−8
−7
−6
−5
−4
−3
−2
Nutri-score

Pistes de modélisation

- Recherche automatique des meilleurs produits par mots clés
- Sauvegarde des produits (ajoutés à la recette)
- Bilan nutritionnel par recette
- Ajout d'une colonne favori (suivant les disponibilités des produits)
- Update automatique de la base de donnée
- Redéfinition du nutri-score / nutri-grade
  - Intégration des additifs et de l'huile de palme

Q&A