

**Concevez une application au  
service de la santé publique**

**Projet 3 du parcour  
Data Scientist**

Dernière MàJ  
15 Janvier 2022

Raquel Sanchez Pellicer



## Présentation du CHEF CLOUD



Données disponibles  
Analyse et nettoyage



Analyse exploratoire  
Étude de faisabilité



Conclusions



# CHEF CLOUD



Comment transformer ce qu'il reste dans  
votre réfrigérateur en la recette qui convient  
le mieux à votre humeur diététique





# CHEF CLOUD



open **FOOD** facts





# CHEF CLOUD



## Quelles sont les données que nous utilisons ?

CHEF CLOUD combine notre base de recettes avec les données nécessaires au calcul du nutriscore des préparations.

Les informations nutritionnelles que nous utilisons sont, pour 100gr de produit :

- Énergie
- Graisses
  - Graisses saturées
- Glucides
  - Sucres
- Protéines
- Fibres
- Sel
  - Sodium
- Pourcentage de légumes, fruits et fruits secs
- Alcool

Nutri-score





## Présentation du CHEF CLOUD



Données disponibles  
Analyse et nettoyage



Analyse exploratoire  
Étude de faisabilité



Conclusions



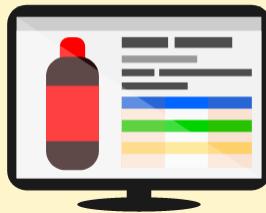
# Données Disponibles



1 ligne = 1 produit  
scanné par un  
collaborateur

- Colones
- Information collaborateur
- Information générales
- Informations ingrédients, additifs
- Informations nutritionnelles
- Informations diverses

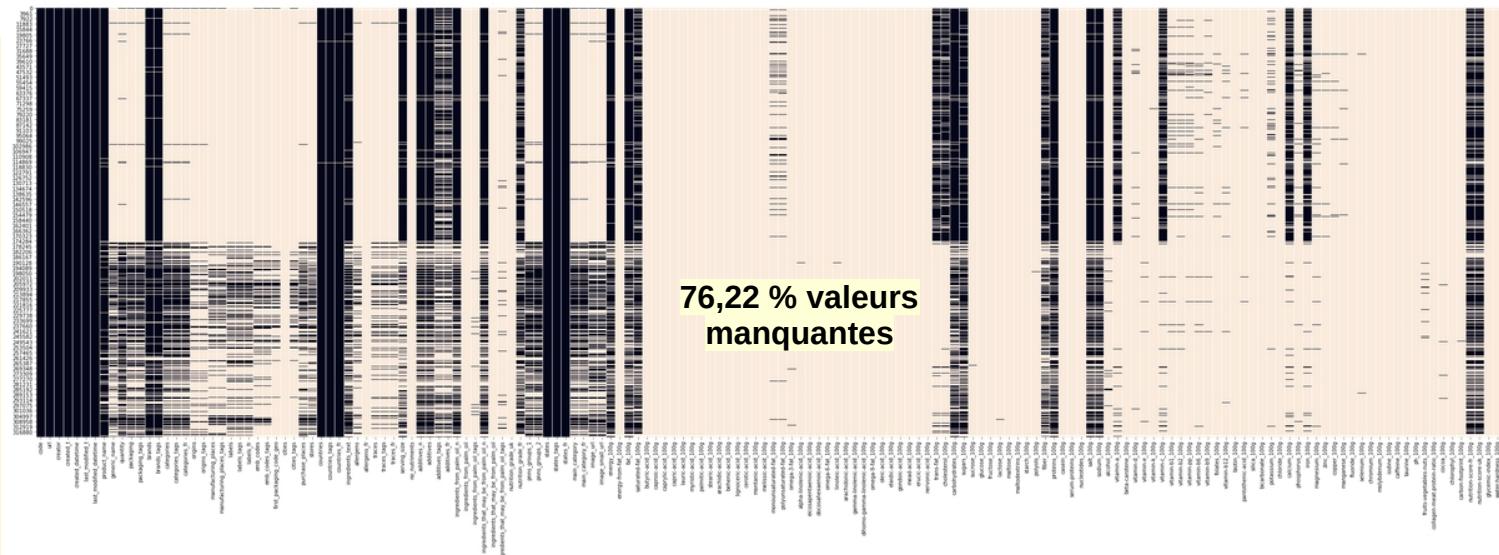
Une base de produits  
alimentaires



Faite par tout le monde



Pour tout le monde

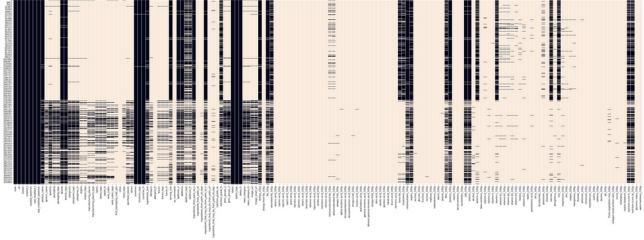


100 colonnes remplies moins de 10 %

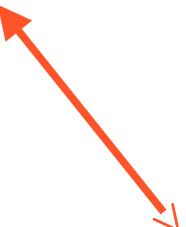
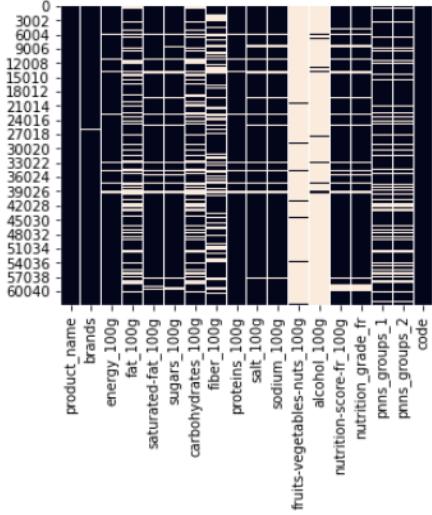
# Nettoyage



De....



à



Actions	Dimensions dataset
<b>Données disponibles</b> , 76.22 % NaN	320722 lignes 162 colonnes
<b>Suppression colonnes vides</b>	320722 lignes 146 colonnes
<b>Reduction : produits commercialisés en France</b>	98464 lignes 146 colonnes
<b>Suppression des lignes</b> sans information pour le <b>nom du produit</b>	91270 lignes 146 colonnes
Sélection <b>variables nutritionnelles</b> . <b>Réduction nombre de colonnes</b>	91270 lignes 19 colonnes
<b>Suppression des lignes</b> sans information pour les <b>variables sélectionnées</b>	66734 lignes 19 colonnes
<b>Suppression des lignes</b> avec <b>problèmes</b> dans les valeurs des <b>variables sélectionnées</b>	66469 lignes 19 colonnes
<b>Suppression doublons</b> (nom produit / marque)	63030 lignes 18 colonnes
<b>Gestion des valeurs manquantes</b> , 0 % NaN	63030 lignes 17 colonnes



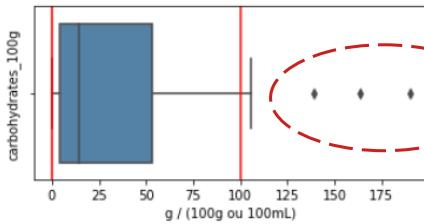
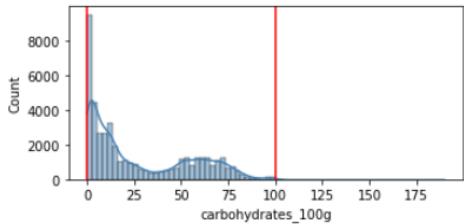
# Valeurs aberrantes



## Valeurs limite

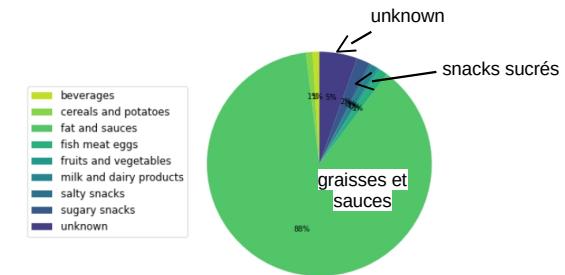
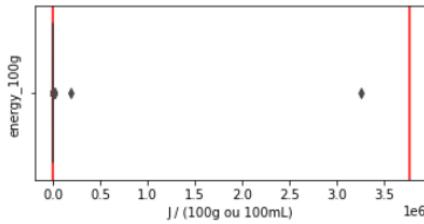
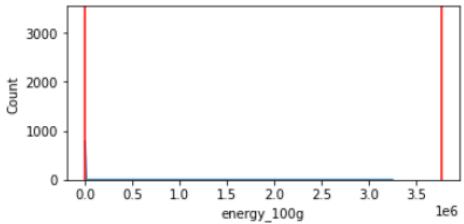
Vari_nutri	lim_min	lim_max	unit
energy_100g	0	3765690	J
fat_100g	0	100	g
saturated-fat_100g	0	100	g
sugars_100g	0	100	g
carbohydrates_100g	0	100	g
fiber_100g	0	100	g
proteins_100g	0	100	g
salt_100g	0	100	g
sodium_100g	0	100	g
fruits-vegetables-nuts_100g	0	100	%
alcohol_100g	0	100	g
nutrition-score-fr_100g	-15	40	

## Glucides

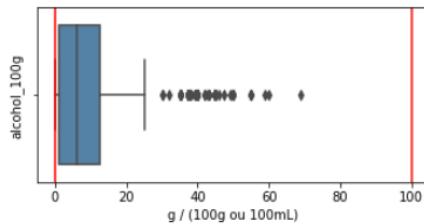
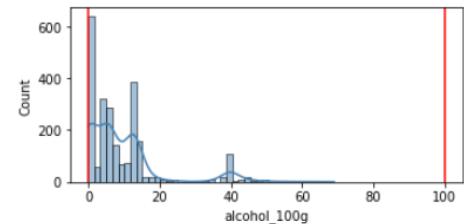


Suppression des valeurs hors limites

## Energie (J)



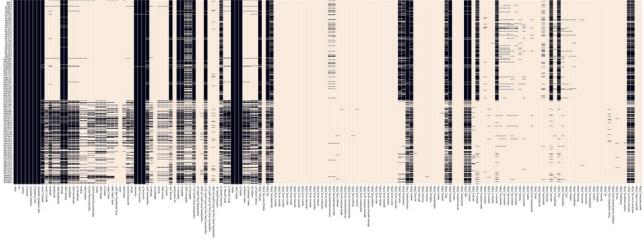
## Alcool



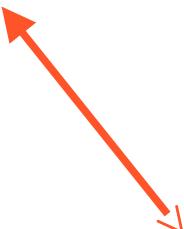
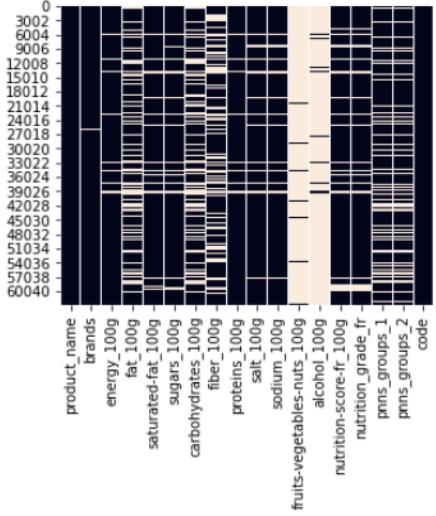
# Nettoyage



De....



... à



Actions	Dimensions dataset
<b>Données disponibles</b> , 76.22 % NaN	320722 lignes 162 colonnes
<b>Suppression colonnes vides</b>	320722 lignes 146 colonnes
<b>Reduction : produits commercialisés en France</b>	98464 lignes 146 colonnes
<b>Suppression des lignes</b> sans information pour le <b>nom du produit</b>	91270 lignes 146 colonnes
Sélection <b>variables nutritionnelles</b> . Réduction <b>nombre de colonnes</b>	91270 lignes 19 colonnes
<b>Suppression des lignes</b> sans information pour les <b>variables sélectionnées</b>	66734 lignes 19 colonnes
<b>Suppression des lignes</b> avec <b>problèmes</b> dans les valeurs des <b>variables sélectionnées</b>	66469 lignes 19 colonnes
<b>Suppression doublons</b> (nom produit / marque)	63030 lignes 18 colonnes
<b>Gestion des valeurs manquantes</b> , 0 % NaN	63030 lignes 17 colonnes



# Valeurs manquantes



## Informations non essentielles pour le calcul du Nutri-score de la recette

- ✓ Marque
- ✓ Catégorie et sous-catégorie d'aliments (groupes PNNS, Programme National Nutrition et Santé )
- ✓ Nutri-score

« non renseigné »

## Informations essentielles pour le calcul du Nutri-score de la recette

- ✓ 'fruit-vegetables-nuts\_100g'
- ✓ alcohol\_100g

Catégories PNNS dans lesquelles on ne s'attend pas à trouver ces éléments

« 0 »

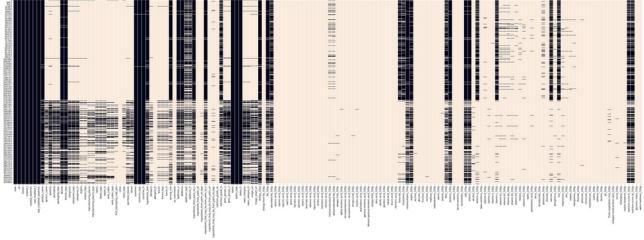
Catégories dans lesquelles on pourrait trouver ces éléments

*knn-imputer*

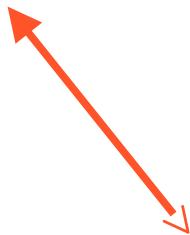
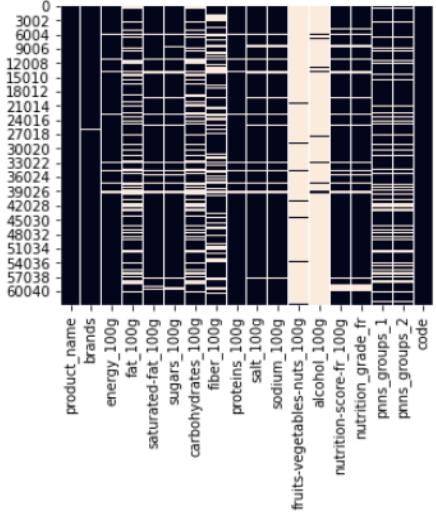
# Nettoyage



De....



... à



Actions	Dimensions dataset
<b>Données disponibles, 76.22 % NaN</b>	320722 lignes 162 colonnes
<b>Suppression colonnes vides</b>	320722 lignes 146 colonnes
<b>Reduction : produits commercialisés en France</b>	98464 lignes 146 colonnes
<b>Suppression des lignes</b> sans information pour le <b>nom du produit</b>	91270 lignes 146 colonnes
<b>Sélection variables nutritionnelles. Réduction nombre de colonnes</b>	91270 lignes 19 colonnes
<b>Suppression des lignes</b> sans information pour les <b>variables sélectionnées</b>	66734 lignes 19 colonnes
<b>Suppression des lignes</b> avec <b>problèmes</b> dans les valeurs des <b>variables sélectionnées</b>	66469 lignes 19 colonnes
<b>Suppression doublons</b> (nom produit / marque)	63030 lignes 18 colonnes
<b>Gestion des valeurs manquantes</b> , 0 % NaN	63030 lignes 17 colonnes



## Présentation du CHEF CLOUD



Données disponibles  
Analyse et nettoyage



Analyse exploratoire  
Étude de faisabilité



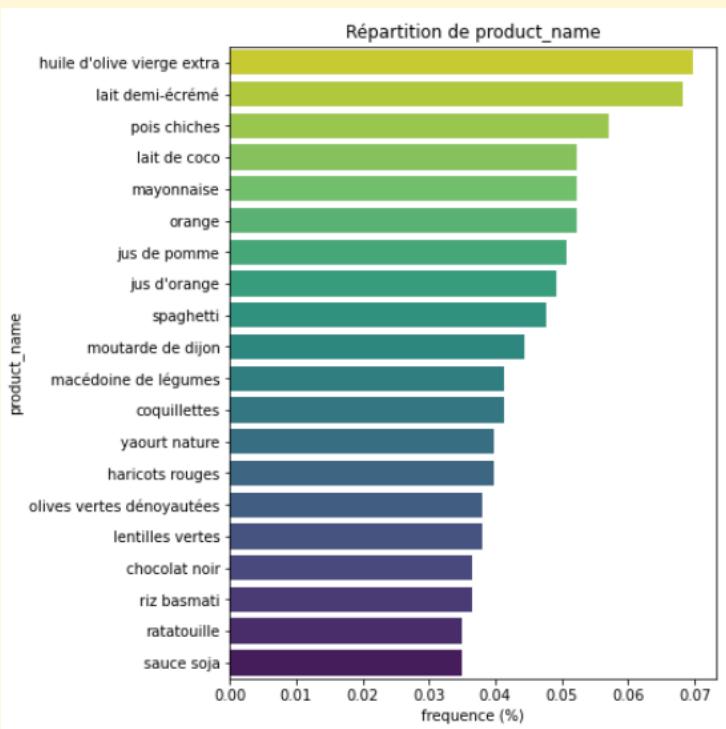
Conclusions



# Analyse exploratoire

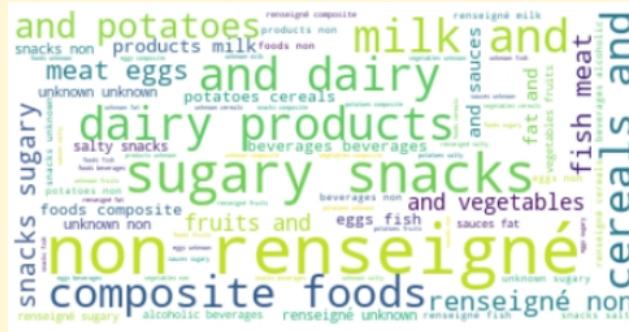


Les produits les plus fréquents dans la base de données sont d'usage quotidien

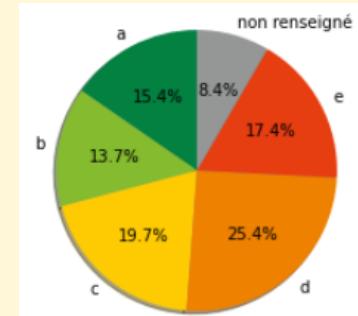


Bonne représentation des catégories de produits non transformés (céréales et pommes de terre, poisson, viande, œufs...)

30% des entrées : non renseigné ou unknown



Représentation homogène des catégories de Nutri-score

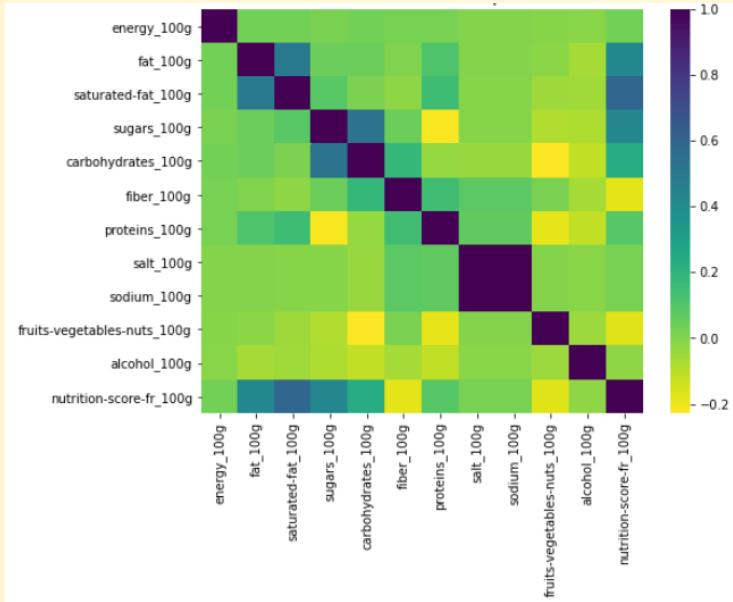




# Étude faisabilité



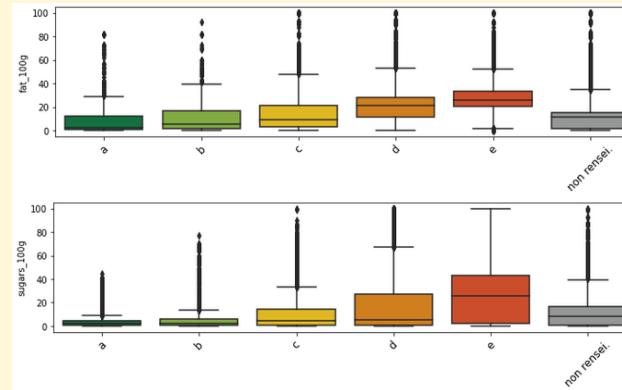
Étude de la dépendance/indépendance des variables nutritionnelles



Les variables nutritionnelles sont indépendantes elles.

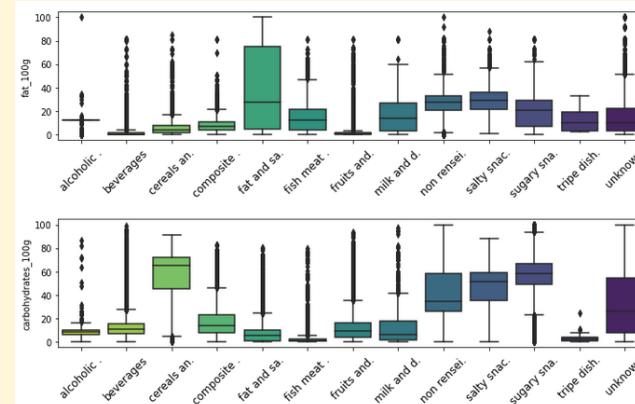
La corrélation entre celles-ci et le Nutri-score corresponds au protocole de calcul.

Distribution des valeurs des variables nutritionnelles / Nutri-score



la quantité d'une variable nutritionnelle ne détermine pas le Nutri-score

Distribution des valeurs des variables nutritionnelles / catégories PNNS



compositions nutritionnelles très différentes au sein d'une même catégorie

# Qualité nutritionnelle régimes perte poids



## Keto food pyramide

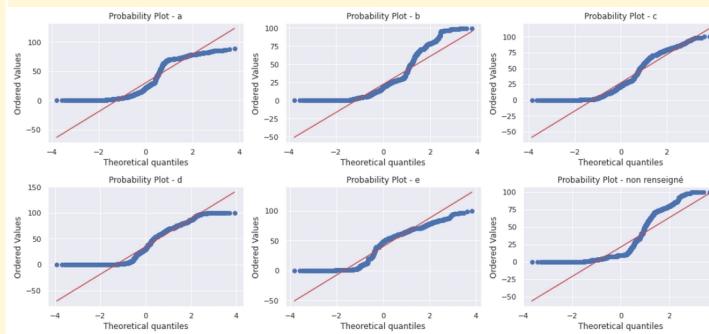
Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s,



# Relation entre la teneur moyenne en glucides et les catégories du Nutri-score

## Hypothèses pour le test ANOVA

- ✓ Les échantillons sont indépendants les uns des autres
    - ✗ Données sont normalement distribuées
    - ✗ Distribution variance homogène



## Êta carré, $n^2$

Mesure apparenté au coefficient de corrélation

$\eta^2$	Interpretation différence des moyennes
0.00 < 0.01	Négligeable
0.01 < 0.06	Petite
0.06 < 0.14	Moyenne
0.14 <= 1	Grande

Le résultat suggère que la différence des moyennes des glucides par Nutri-score est petite.

Pas de preuve de relation entre la teneur moyenne en glucides et les catégories du Nutri-score.

Pas de relation entre les restrictions du régime cétogène et la amélioration de la qualité nutritionnelle.





## Présentation du CHEF CLOUD



Données disponibles  
Analyse et nettoyage



Analyse exploratoire  
Étude de faisabilité



Conclusions



# Conclusions



- ☞ Toutes les informations nécessaires pour estimer le nutri-score d'une recette sont disponibles dans la base de données, Open Food Facts
- ☞ La qualité nutritionnelle, Nutri-score, est un concept complexe. Informations nutritionnelles 100g nécessaires .
- ☞ Les différentes catégories d'aliments ne sont pas liées à un Nutri-score donné.
- ☞ La proportion des ingrédients et leur combinaison permettent de composer des recettes avec Nutri-score différent.
- ☞ CHEF CLOUD peut être utile pour intégrer des notions de qualité nutritionnelle à certains régimes de perte de poids.



# Perspectives

analyse des données de la phase de révision

- il existent des boissons contenant 80% de fruits et légumes
- il existent des fruits et des légumes contenant plus de 50 g de graisses saturées

☛ Vérifier la relation entre l'énergie et la graisse

☛ Rechercher des filtres pour réduire le nombre de "non renseigné"

# BOITE À OUTILS

