Analyse de données nutritionnelles

Thomas WEBER

Introduction

- Le site Lamarmite souhaite réaliser un générateur de recettes saines.
- Source des informations nutritionnelles: Open Food Facts (une base de données libre et ouverte).
- Objectif: nettoyer la base de données et réaliser une analyse exploratoire.

Nutrition – les bases

- Lipides:
  - Acides gras saturés
  - Acides gras insaturés (mono et poly)
- Glucides:
  - Sucres (ou glucides simples) – Fruits, lait, bonbons, biscuits
  - Glucides complexes – Riz, pâtes, céréales
- Protéines
- Fibres

1g de lipides = 9 kcal
1g de glucides = 4 kcal
1g de protéines = 4 kcal

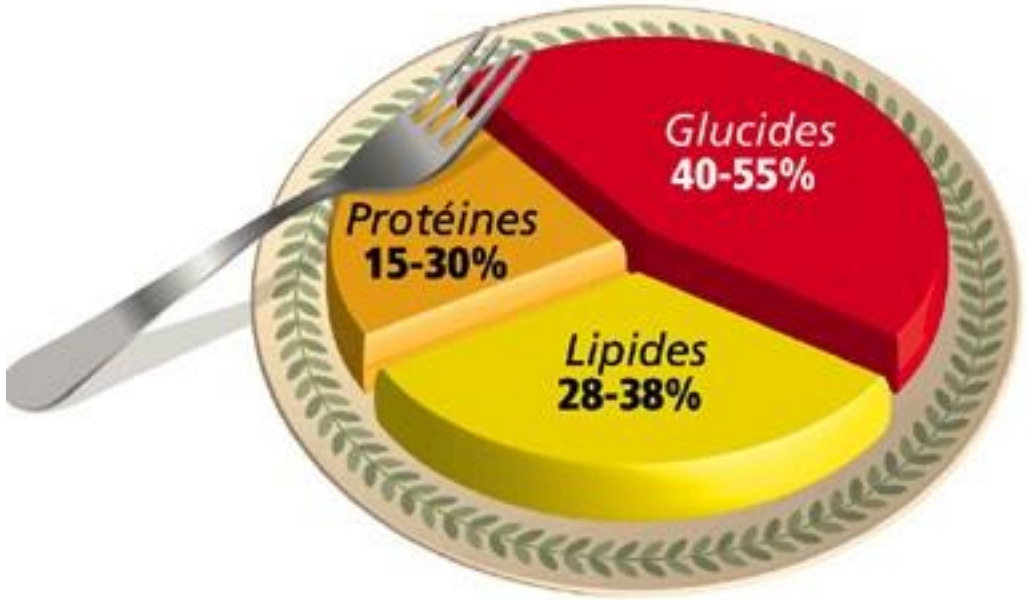
Qu’est qu’une alimentation saine ?

Limiter (sans éliminer):

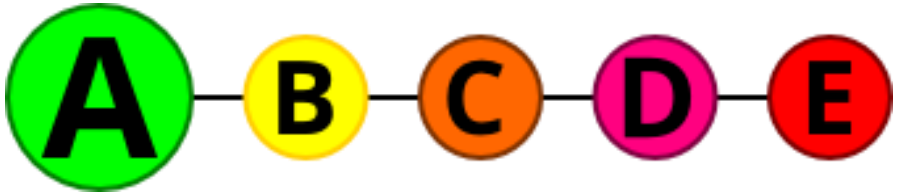
- Acides gras saturés
- Sucres (glucides simples)

Privilégier:

- Fibres
- Fruits, légumes



Score nutritionnel



- Un score calculé entre -15 (meilleure note) et 40 (plus mauvaise note). Un barème donne la notation entre A et E pour plus de lisibilité sur les emballages.
- Les facteurs qui améliorent la note: la teneur en fruits et légumes (en %), les fibres et éventuellement les protéines.
- Les facteurs qui détériorent la note: le nombre de calories, les acides gras saturés, les sucres, la teneur en sel.
- **Une valeur de sortie pour le générateur de recettes.**

Nettoyage – Vue d’ensemble

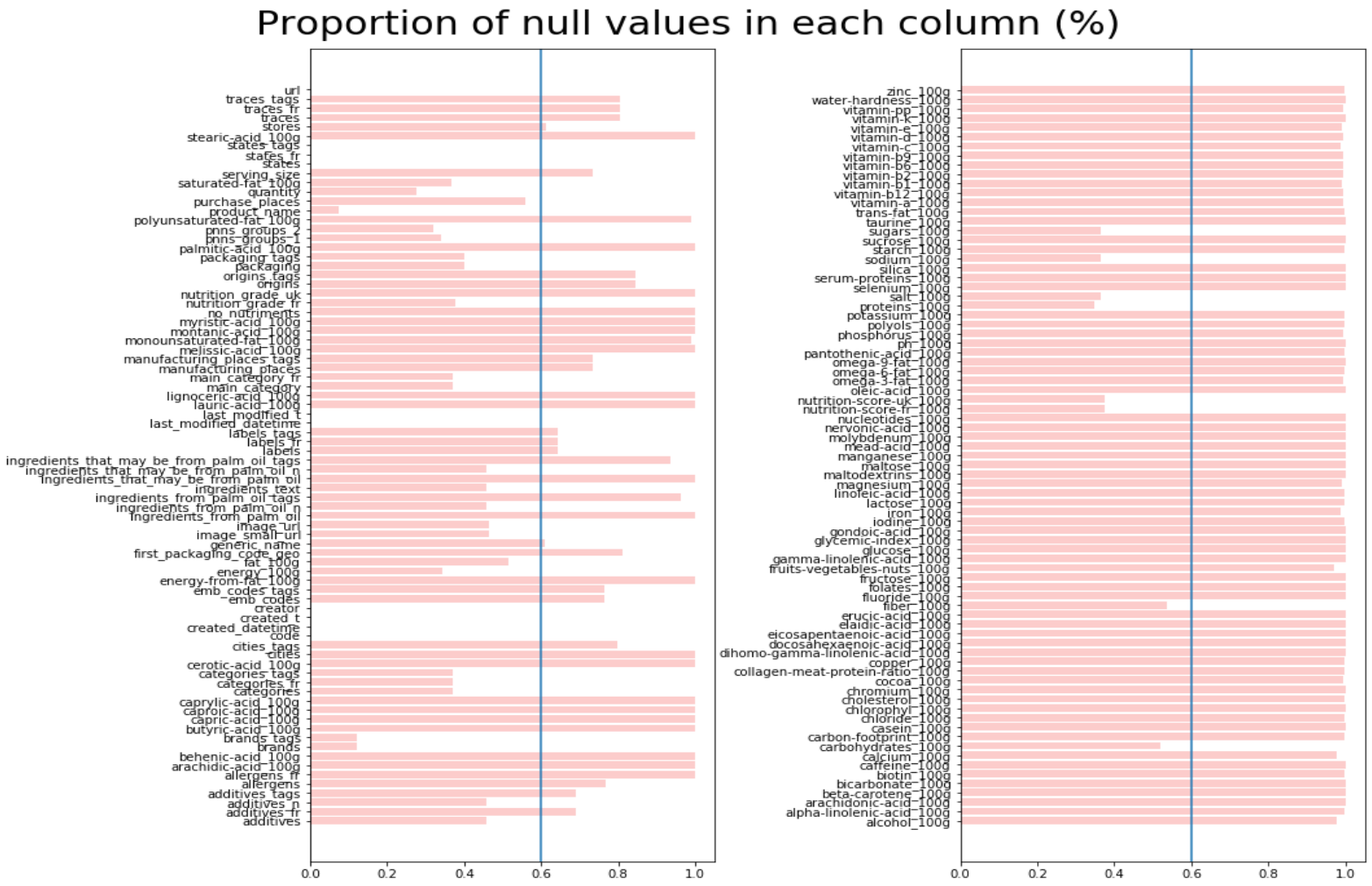
- Environ 320 000 lignes et 162 features pour la base de données.
- Base de données libre et ouverte à tout le monde => Peu de consistance.
- Parmi les features, on retrouve:
  - Des informations générales (nom du produit, date de creation, auteur)
  - Des tags (categories, marques, pays de vente, pays de fabrication)
  - Les informations sur le produit (ingredients, allergènes, additifs)
  - Les valeurs nutritionnelles détaillées (macro-nutriments mais aussi sels minéraux, vitamines)

Nettoyage – Adaptation au marché français

- Filtrage uniquement sur les produits vendus en France
  - On passe de 320 000 lignes à 98 000.
- Conversion de l'énergie de kilojoules en calories
  - 1 kcal = 4.184 kJ
- Deux scores nutritionnels (FR et UK)
  - Très proches mais avec un barème différent pour les boissons, les fromages ou les matières grasses.
  - On garde le score français.

Nettoyage – Filtrage des features

- On élimine les features avec plus de **60%** de valeurs manquantes.
- On passe de 162 à 43 features.
- Parmi celles restantes, un certain sont sans intérêts (timestamps, code barre, …)
- On passe de 43 à 24 features.



Nettoyage – Valeurs manquantes

- Suppression de la ligne s'il manque:
  - Le nom du produit
  - Le nombre de calories
  - Le score nutritionnel (entre -15 et 40)

- On passe de 98 000 à 61 000 lignes.

- Pour les lignes restantes: on complète avec np.NaN.

Nettoyage – Valeurs aberrantes

- On élimine certaines lignes si:
  - Nombre de calories > 900
  - Quantitié d'un nutriment pour 100g > 100g
  - Somme des nutriments pour 100g > 100g (à 10% près)
  - Nombre de calories différent à plus de 10% près de:

    9 * lipides + 4 * (glucides + protéines) + 1.9 * fibres

- On passe de 61 000 à 59 000 lignes.

Nettoyage – Features semblables

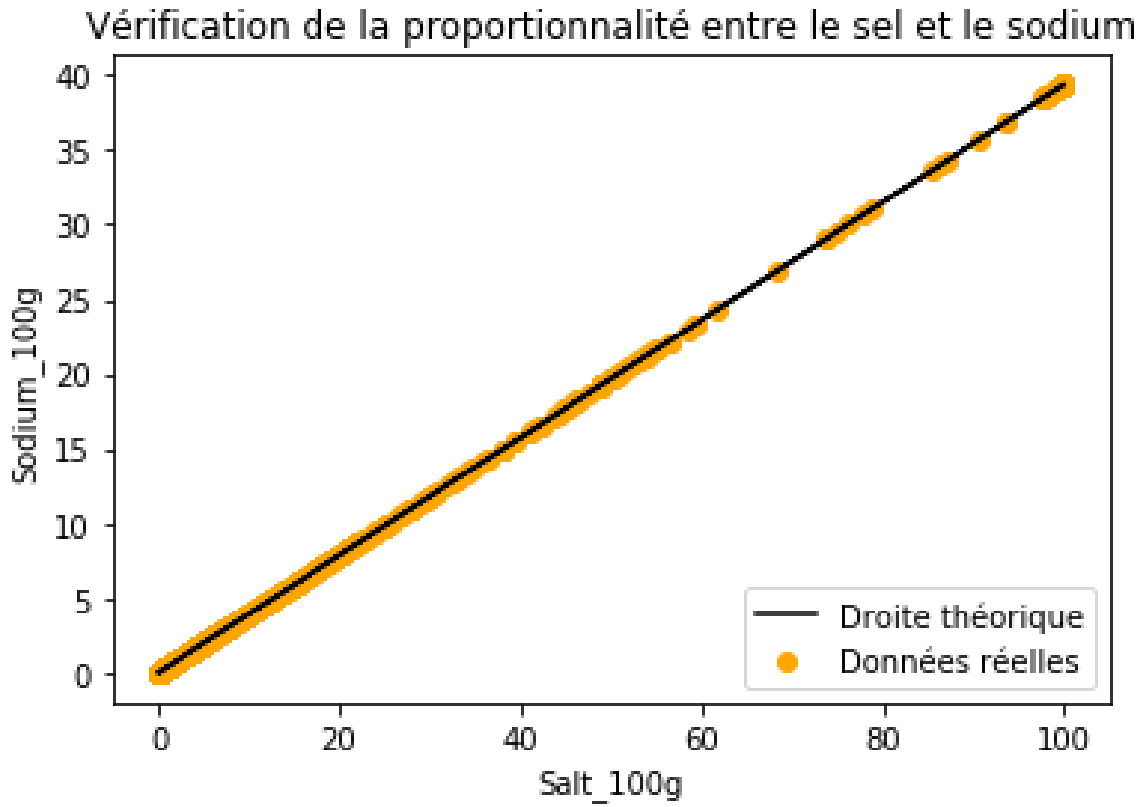
- Choix d'une variable « catégorie »
  - Pour les 7 choix possibles, on regarde le nombre d'occurrences différentes pour chacune.
  - 'pnns_groups_1' est la seule avec un nombre raisonnable d'occurrences pour permettre une analyse plus tard.

- Choix d'une variable « marque »
  - 2 choix possibles, brands et brands_tags mais identiques (sauf au niveau de la casse).
  - On garde « brands ».

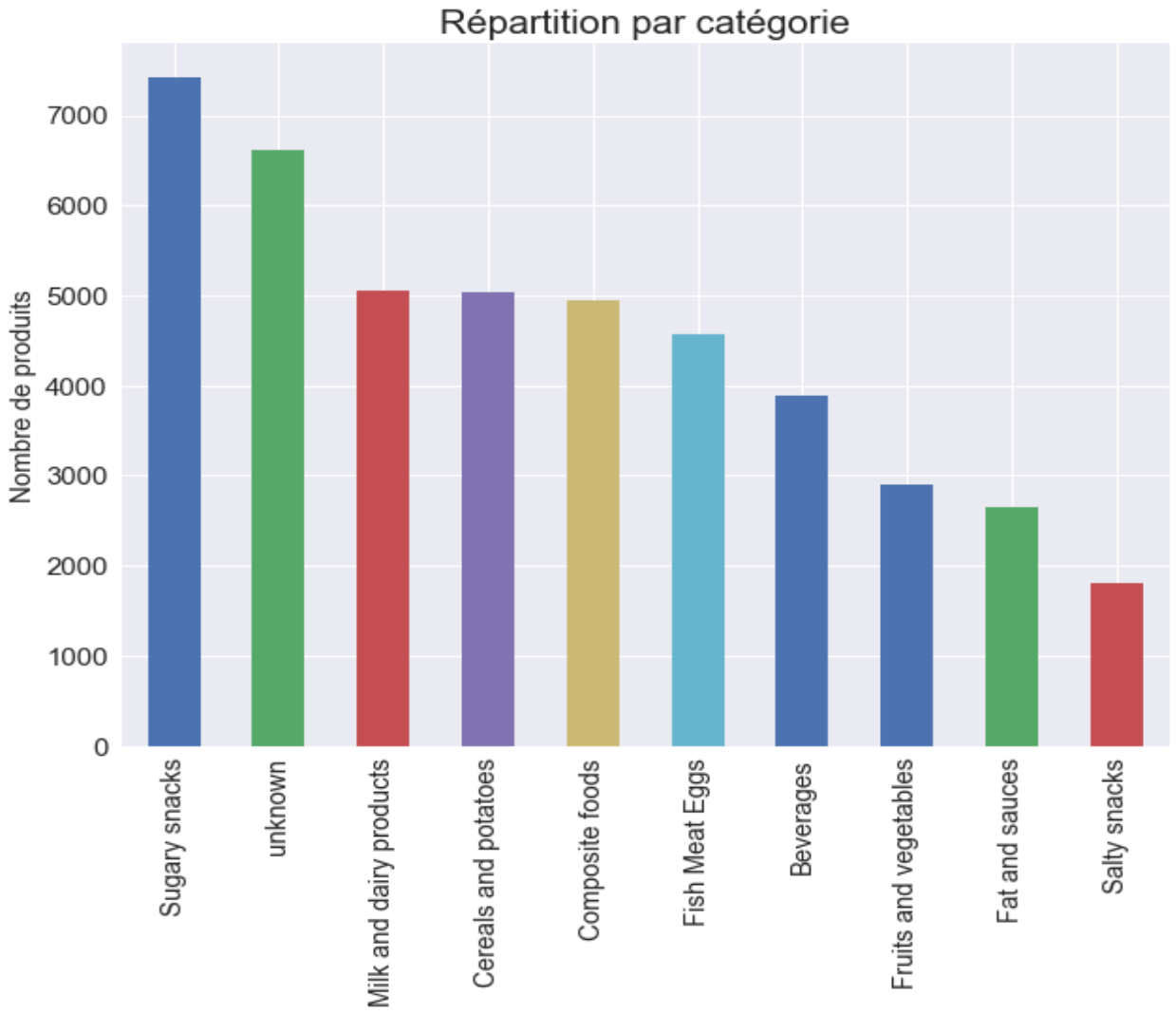
Nettoyage – Features semblables

- Entre sel et sodium:

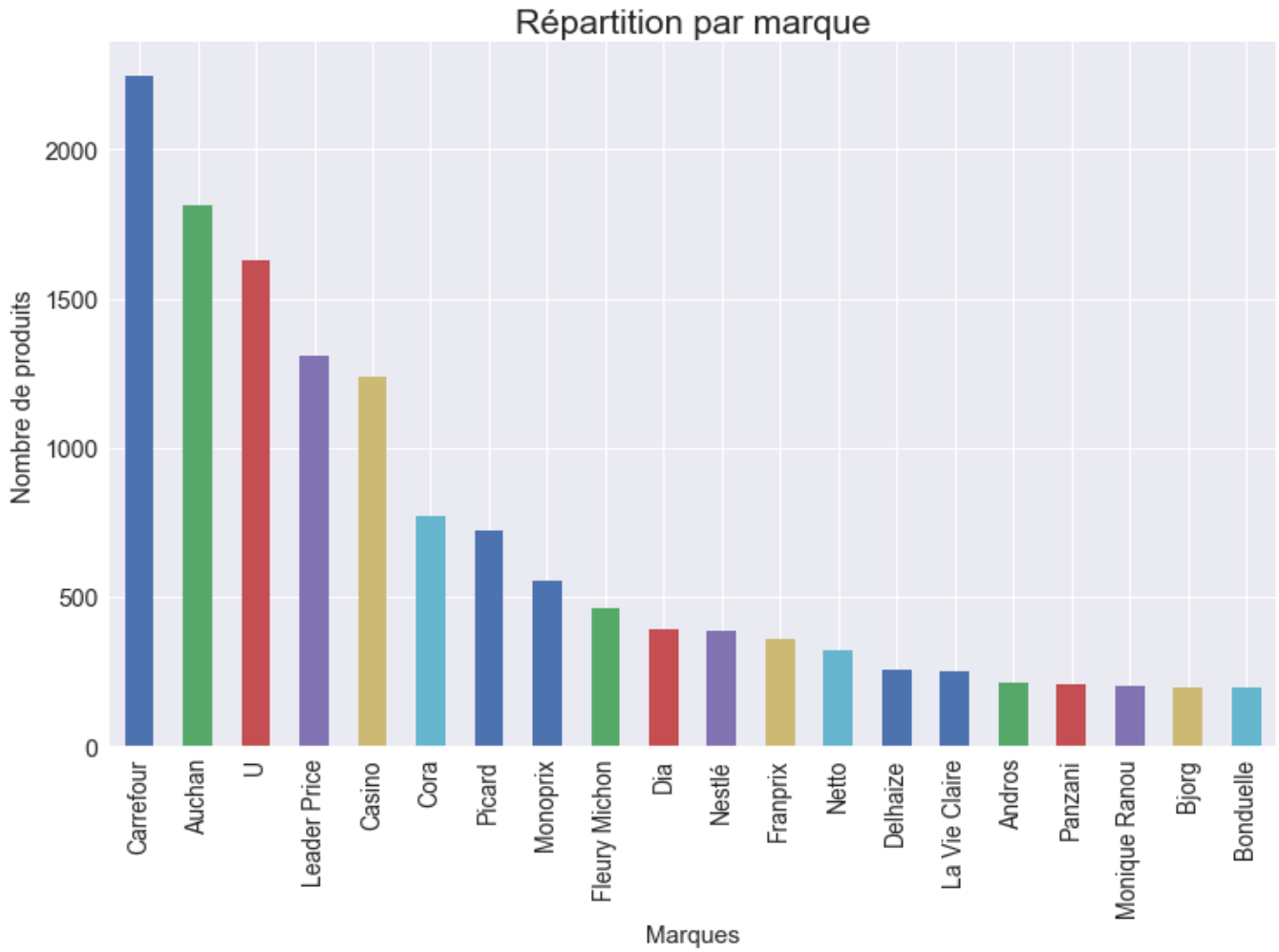
Il s’agit de la même chose:
1g de sodium = 2.54g de sel



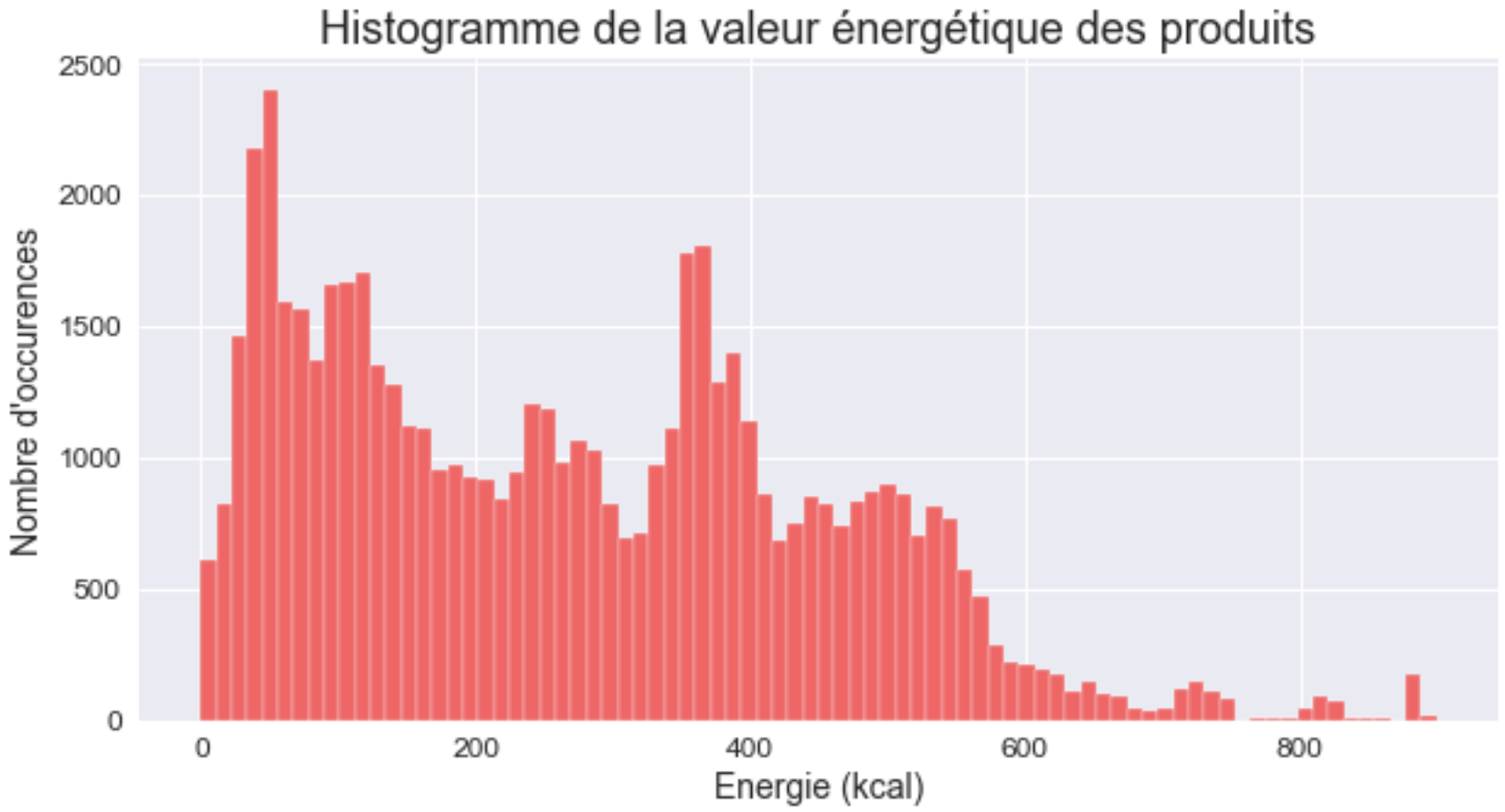
Analyse univariée - Catégories



Analyse univariée - Marques



Analyse univariée – Apport énergétique

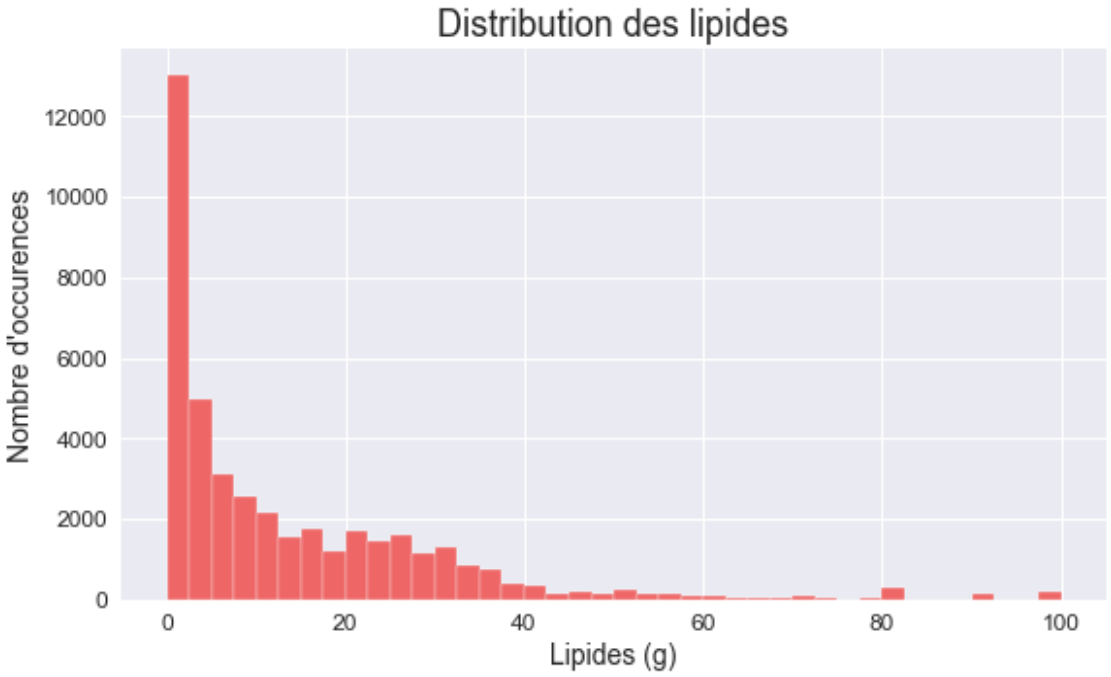


Distribution pluri-modale

Excentrée vers la gauche

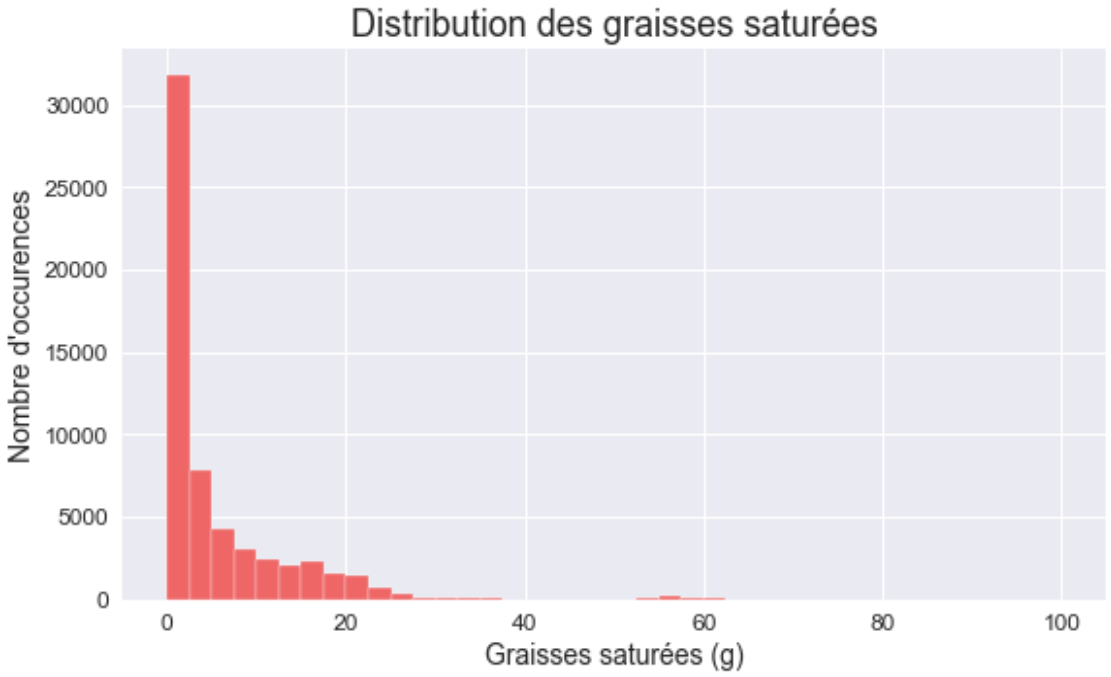
Majorité des produits < 550 kcal

Analyse univariée – Lipides



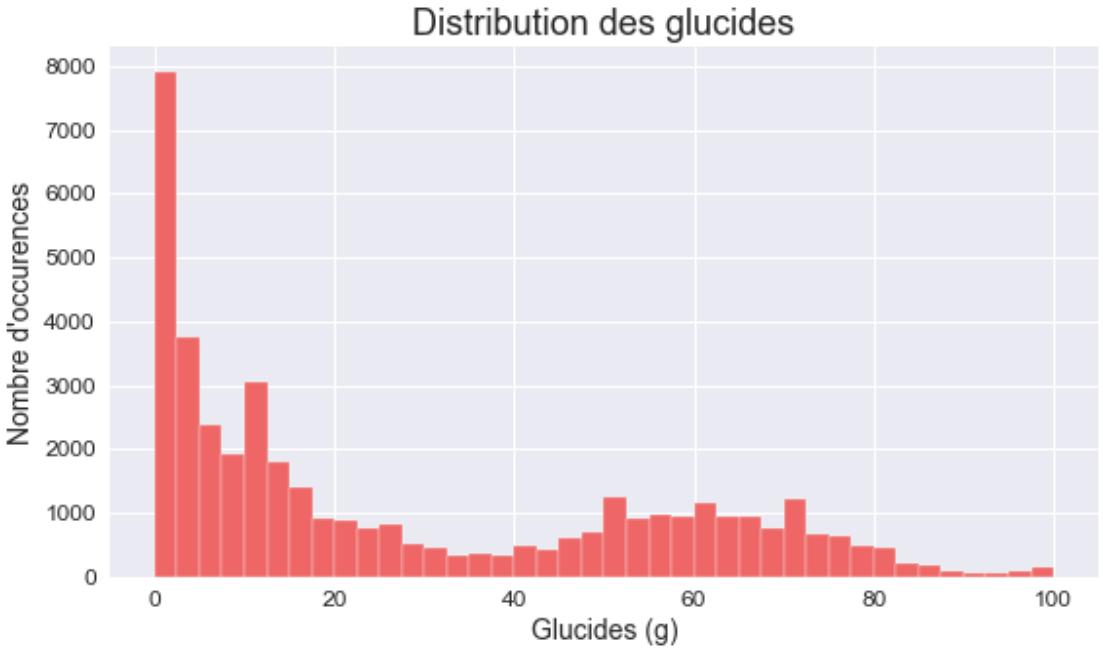
Excentrée vers la gauche

Moyenne: 13.56g
Ecart-type: 16.52
3e quartile: 21.7g



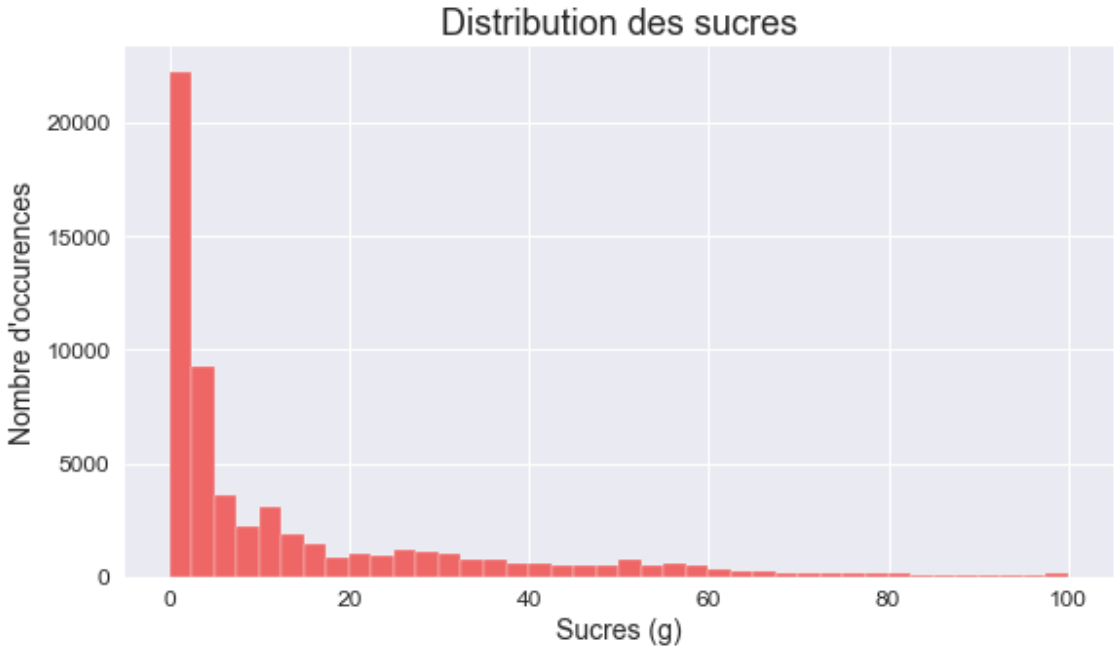
Excentrée vers la gauche
Moyennes: 5.46g
Ecart-type: 8.32
3e quartile: 7.5g
Les produits les plus gras: l'huile de coco

Analyse univariée - Glucides



Distribution bi-modale

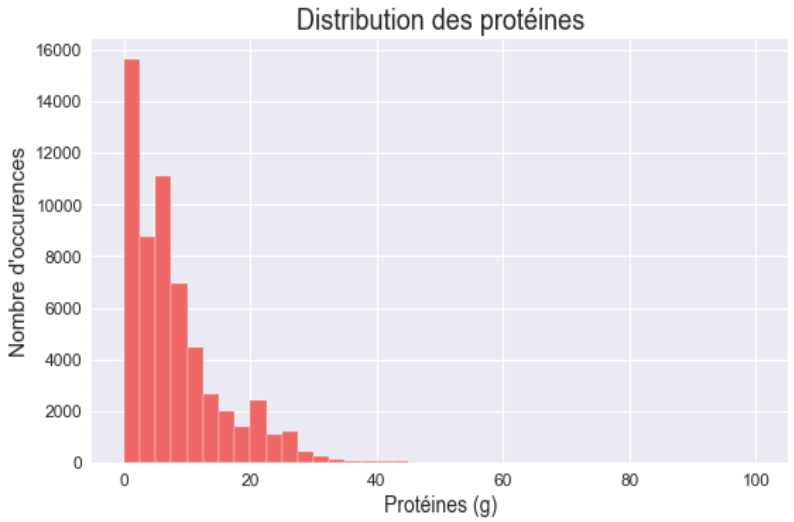
Moyenne: 27.93g
Ecart-type: 27.15
3e quartile: 53.1g



Distribution excentrée à gauche

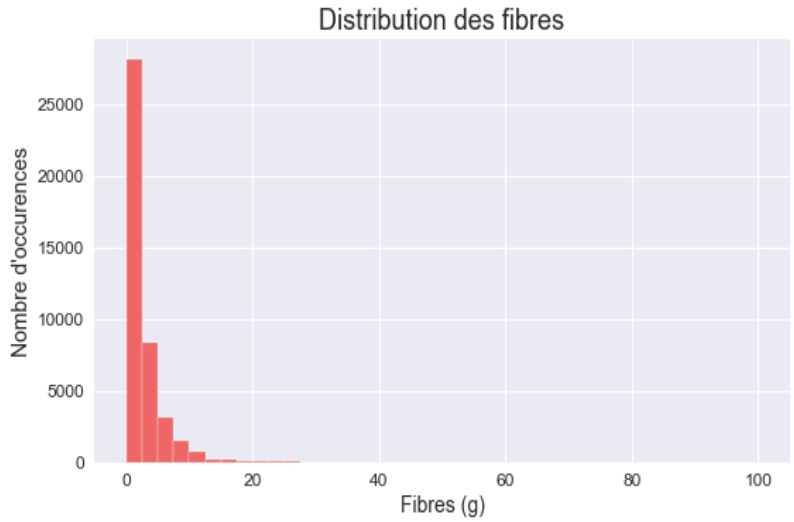
Moyenne: 13.59g
Ecart-type: 19.12
3e quartile: 18.4g

Analyse univariée – Protéines, fibres et sel



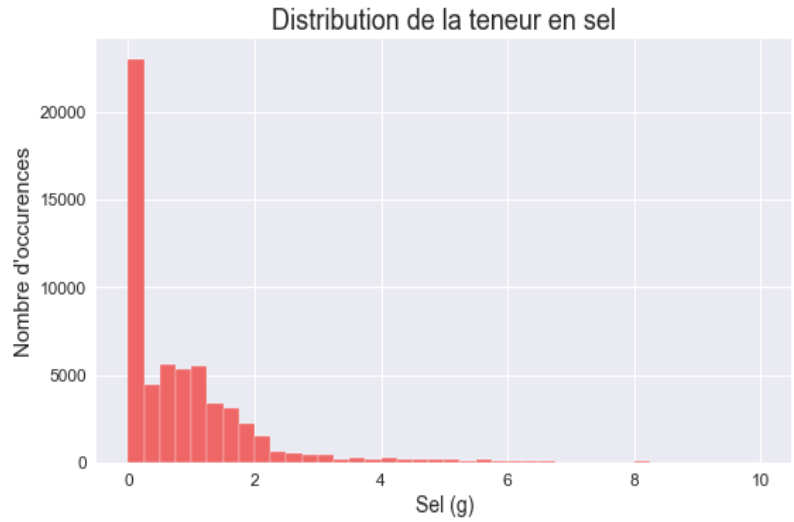
Excentrée vers la gauche

Moyenne: 7.84g
Ecart-type: 7.68
3e quartile: 11g



Excentrée vers la gauche

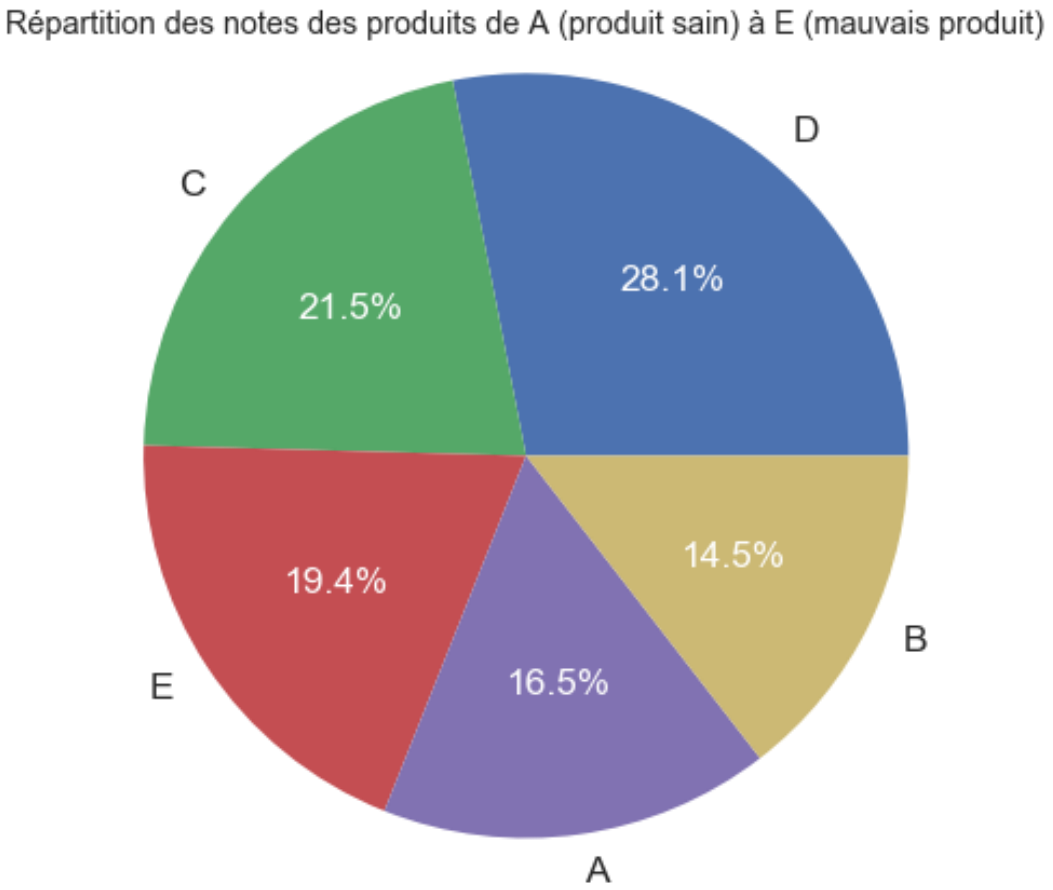
Moyenne: 2.47g
Ecart-type: 3.97
3e quartile: 3.2g



Excentrée vers la gauche

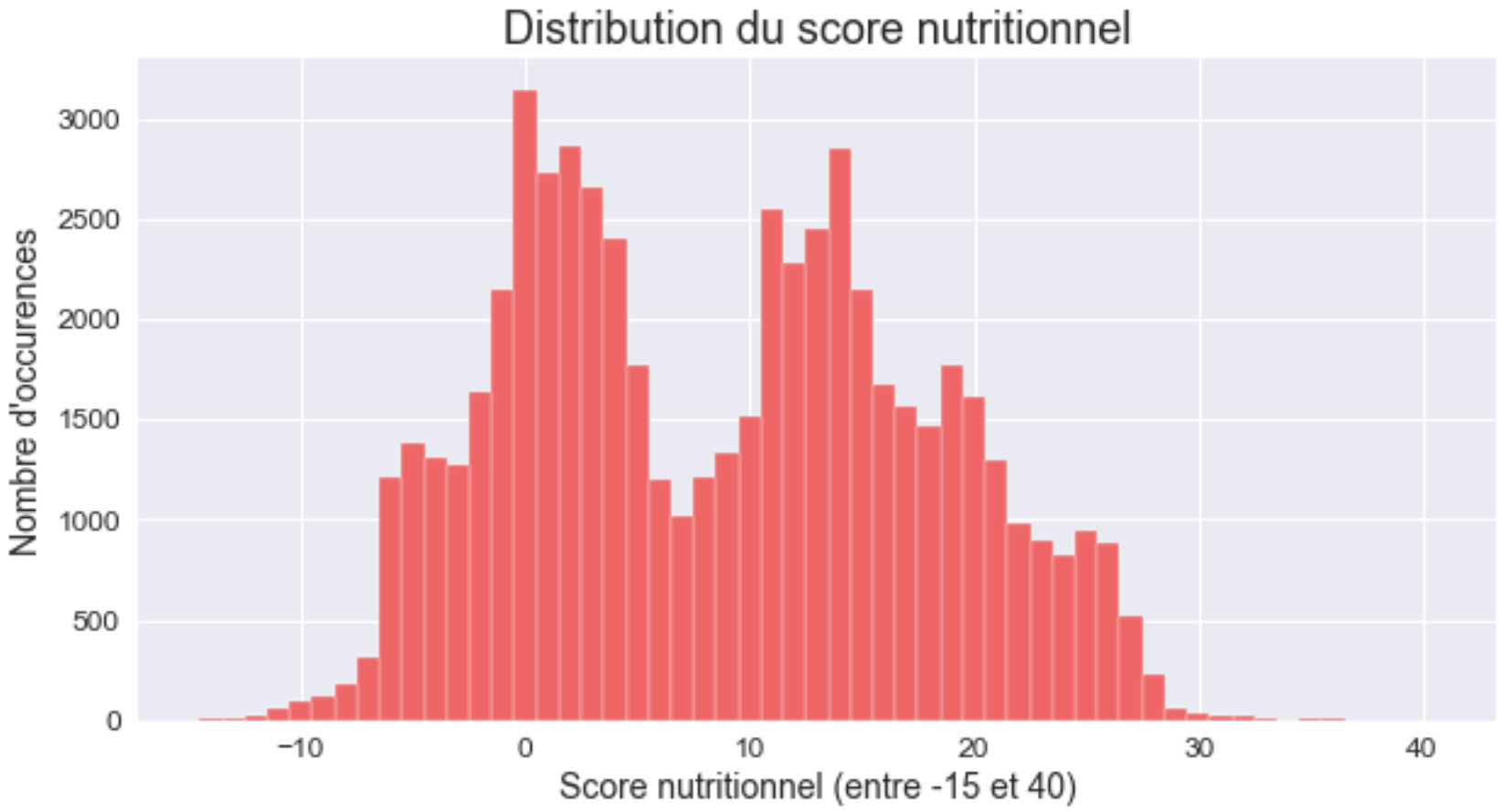
Moyenne: 1.06g
Ecart-type: 3.53
3e quartile: 1.25g

Analyse univariée – Note nutritionnelle (entre A et E)



Répartition homogène

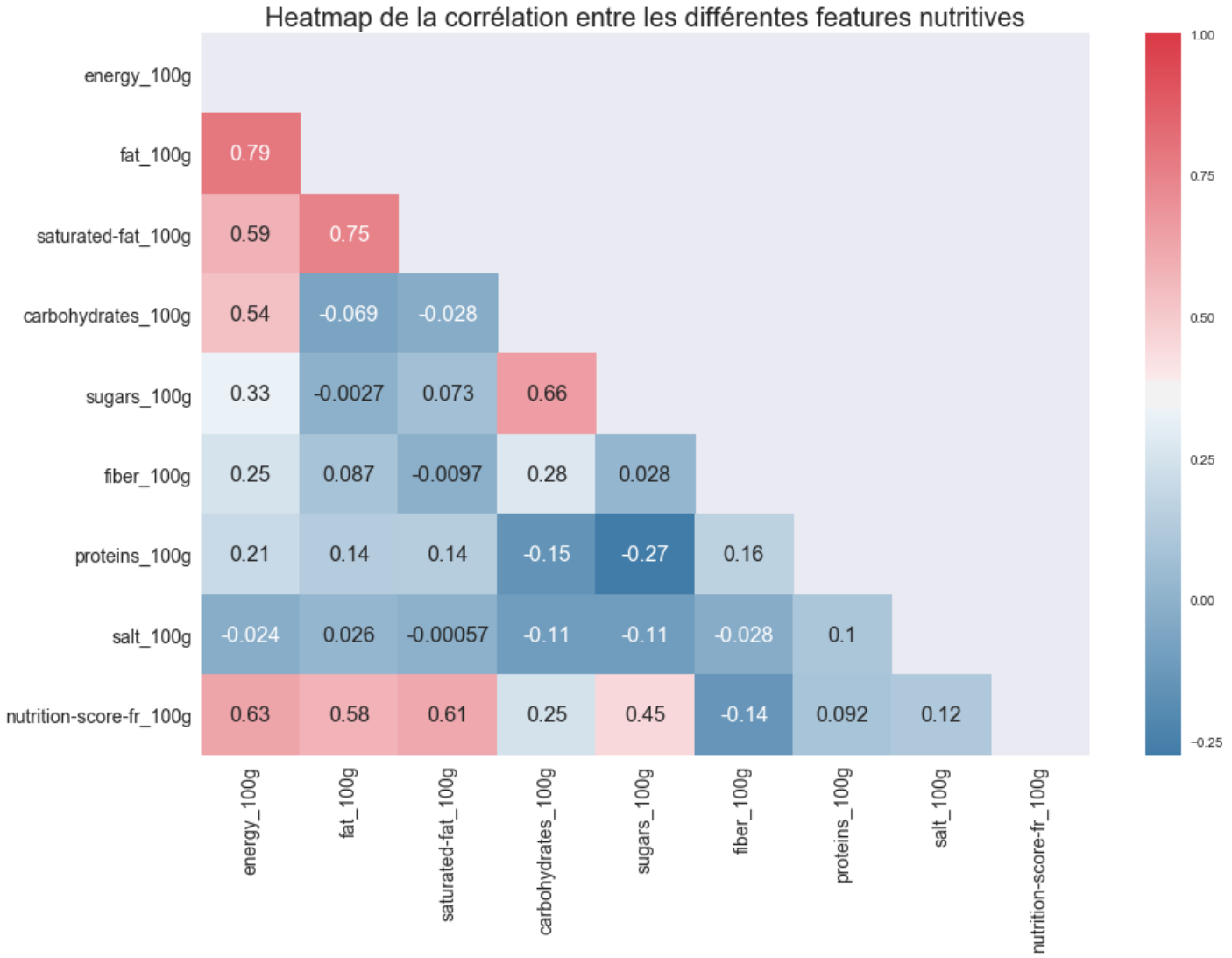
Analyse univariée – Score nutritionnel



Distribution bi-modale
(autour de 0 et 13)

Moyenne: 8.80
Ecart-type: 9.04

Analyse multivariée – Heatmap corrélation

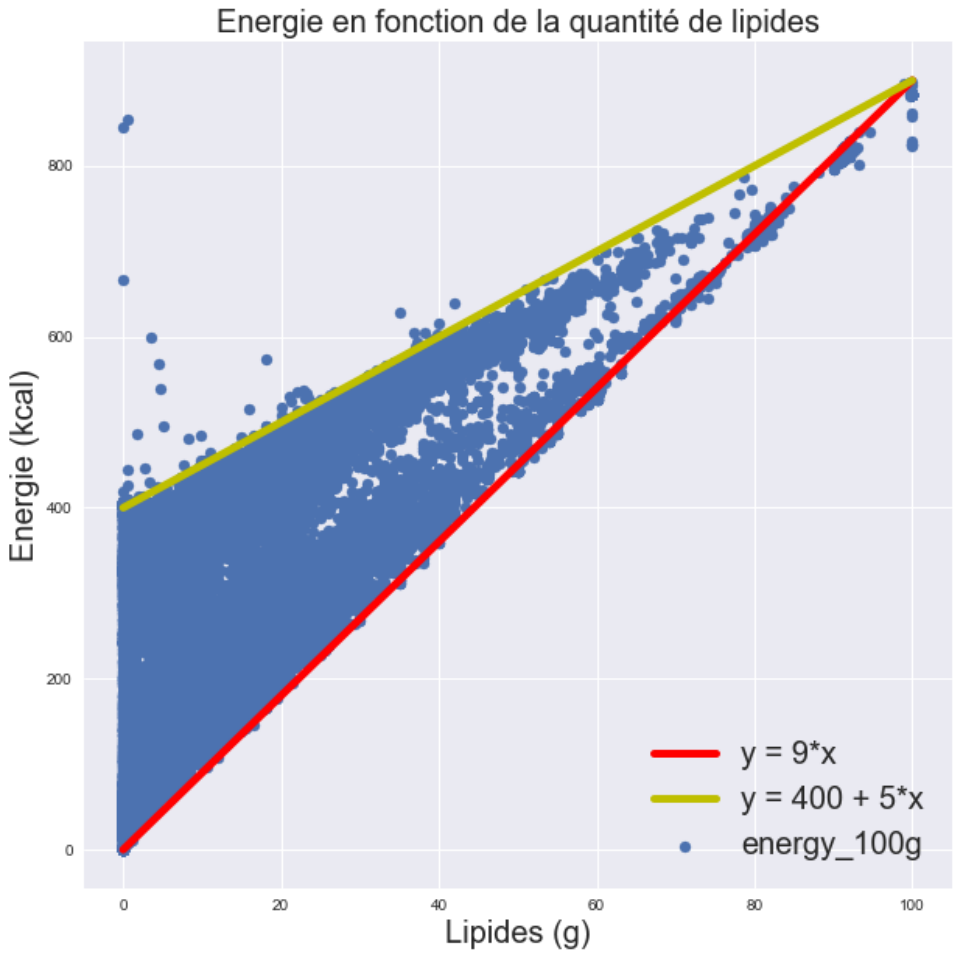


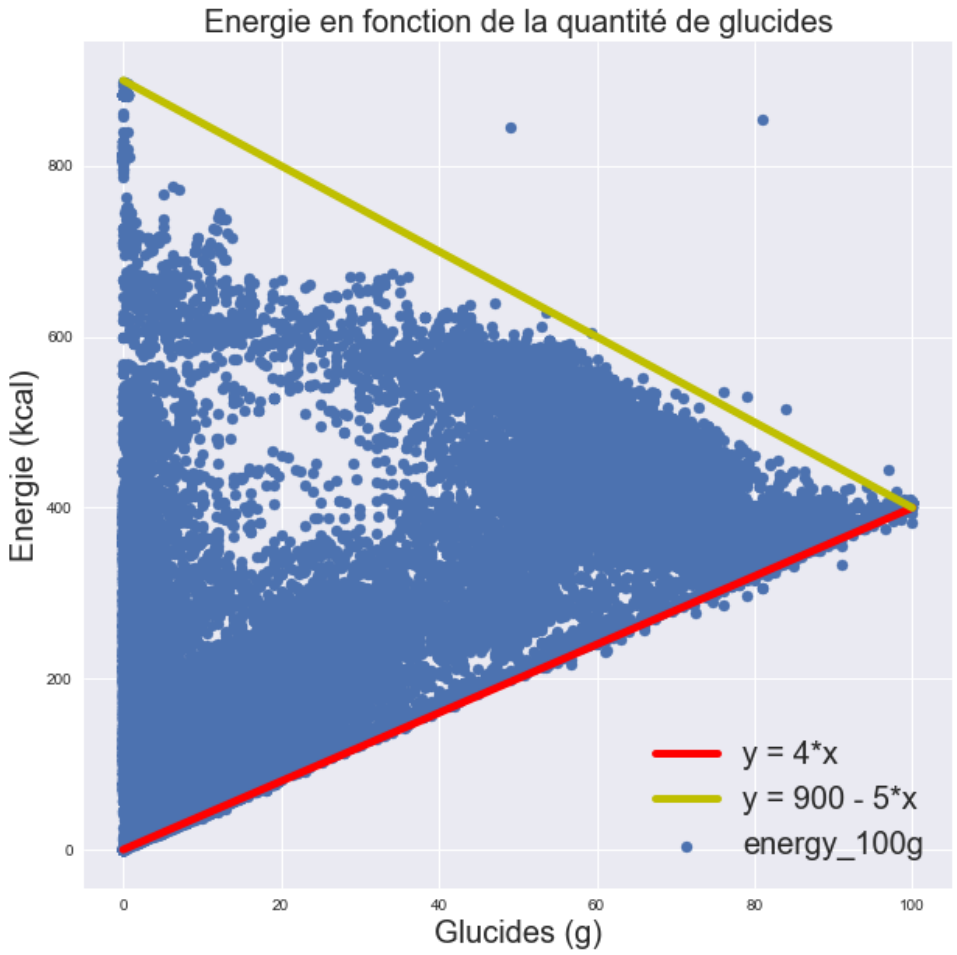
- Corrélations à examiner:
  - Energie / Lipides
  - Energie / Glucides
  - Lipides / Graisses saturées
  - Glucides / Sucres

- Variables qui semblent augmenter le score nutritionnel:
  - Energie
  - Lipides/Graisses saturées
  - Sucres

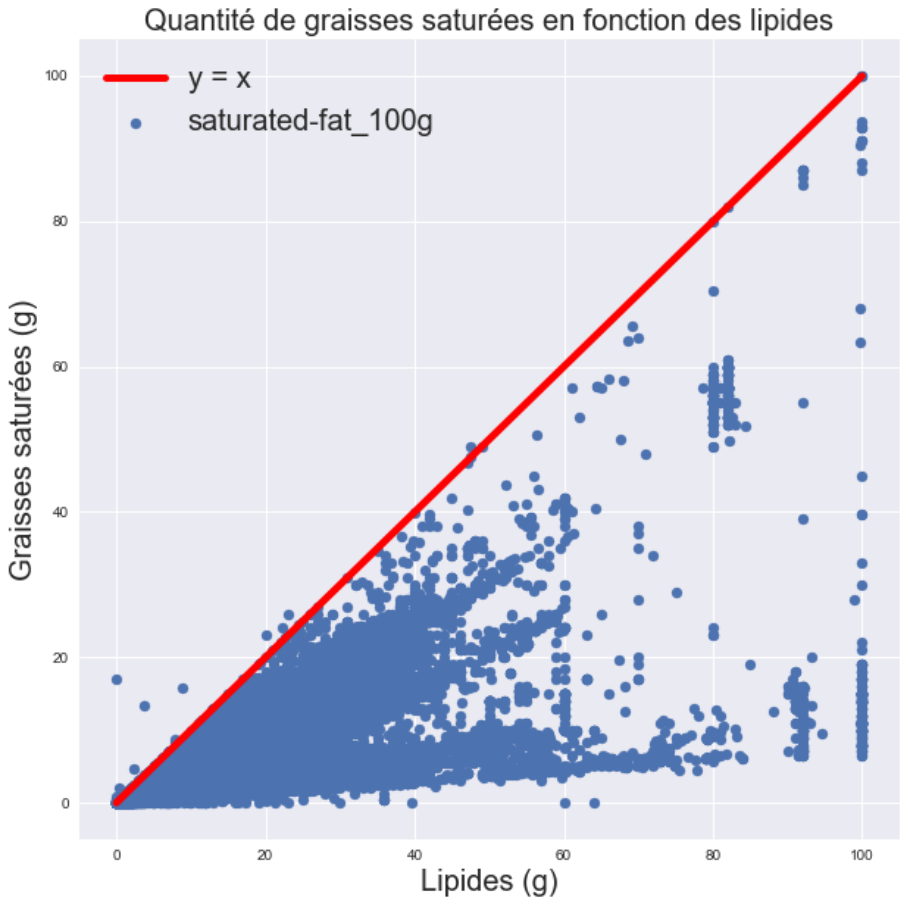
- Les autres ont l'air d'avoir peu d'effet significatif sur le score.

Analyse multivariée – Energie/Lipides/Glucides





Analyse multivariée – Lipides/Graisses saturées



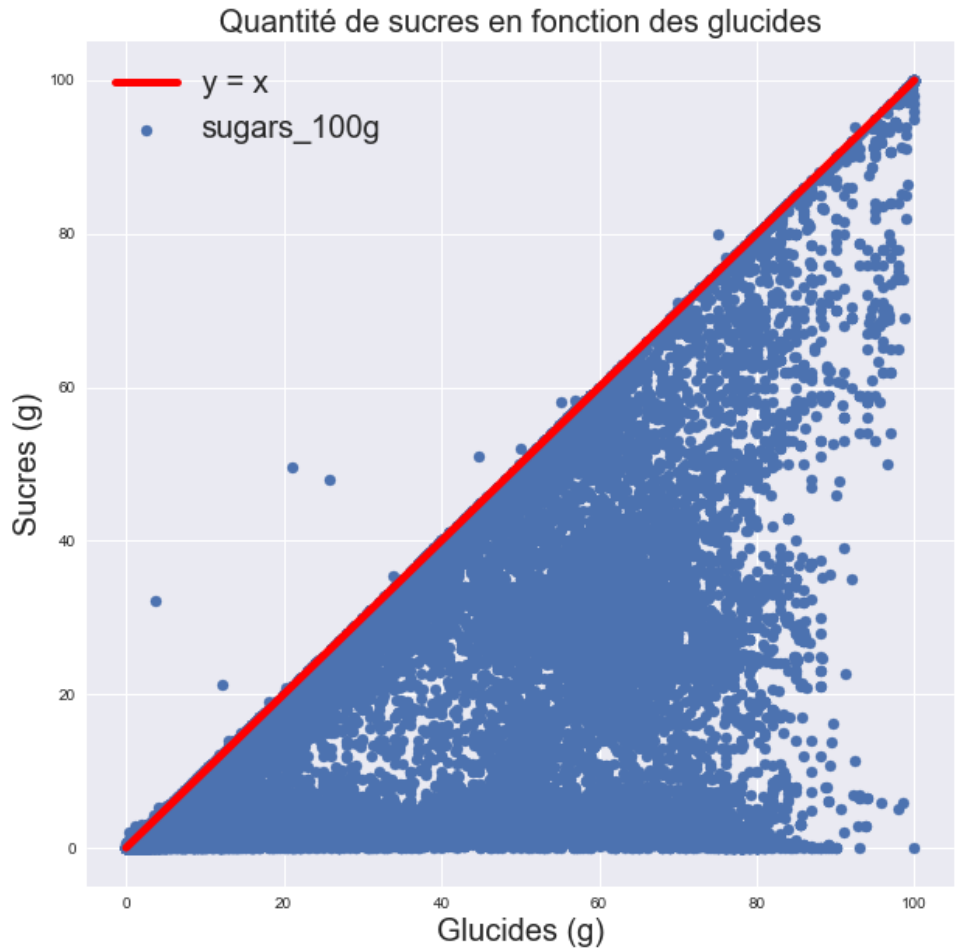
- Les graisses saturées sont un type de lipides: d'où le coefficient de corrélation élevé (0.75).

- L'inverse est fausse. C'est pourquoi on n'a pas plus l'allure d'une droite.

- Dans le cadre d'une alimentation saine, il faudra trouver des produits le plus loin possible de la ligne rouge.

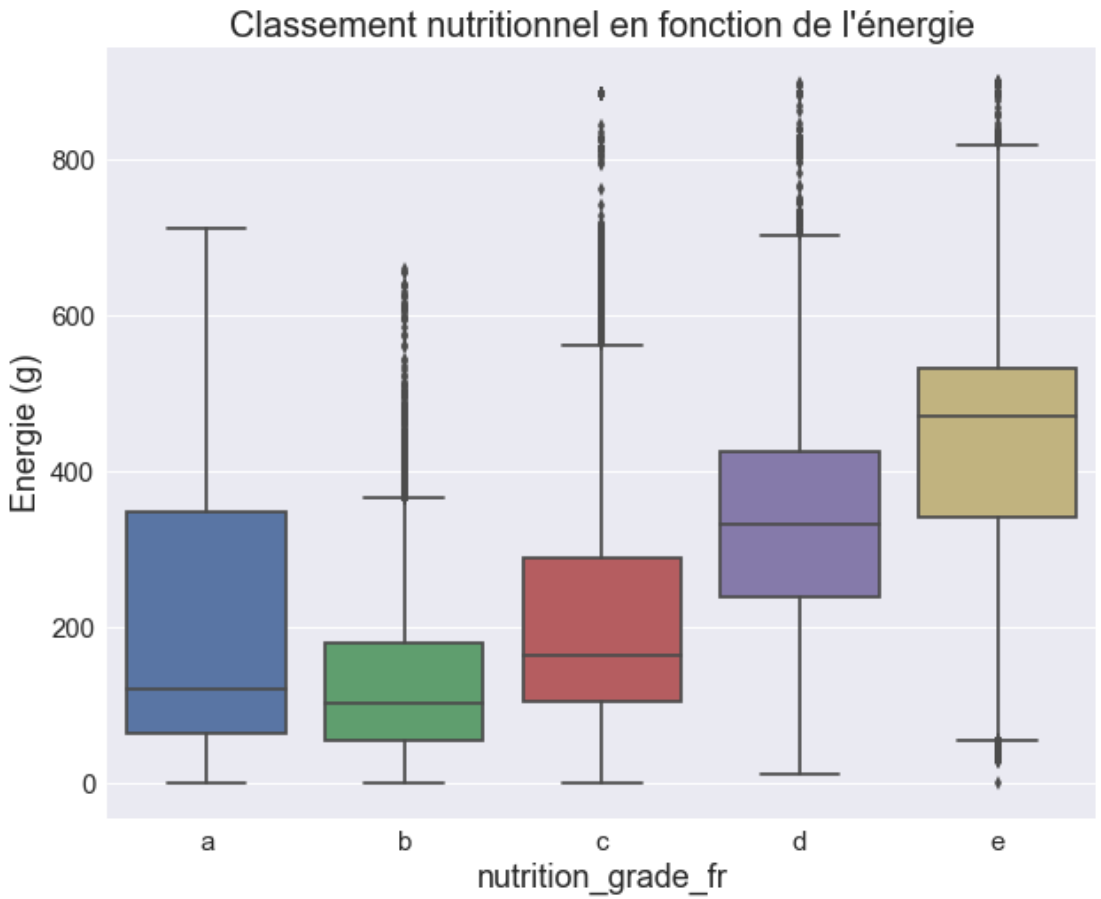
- Par exemple, à l'extrême en bas à droite on retrouve l'huile d'olive.

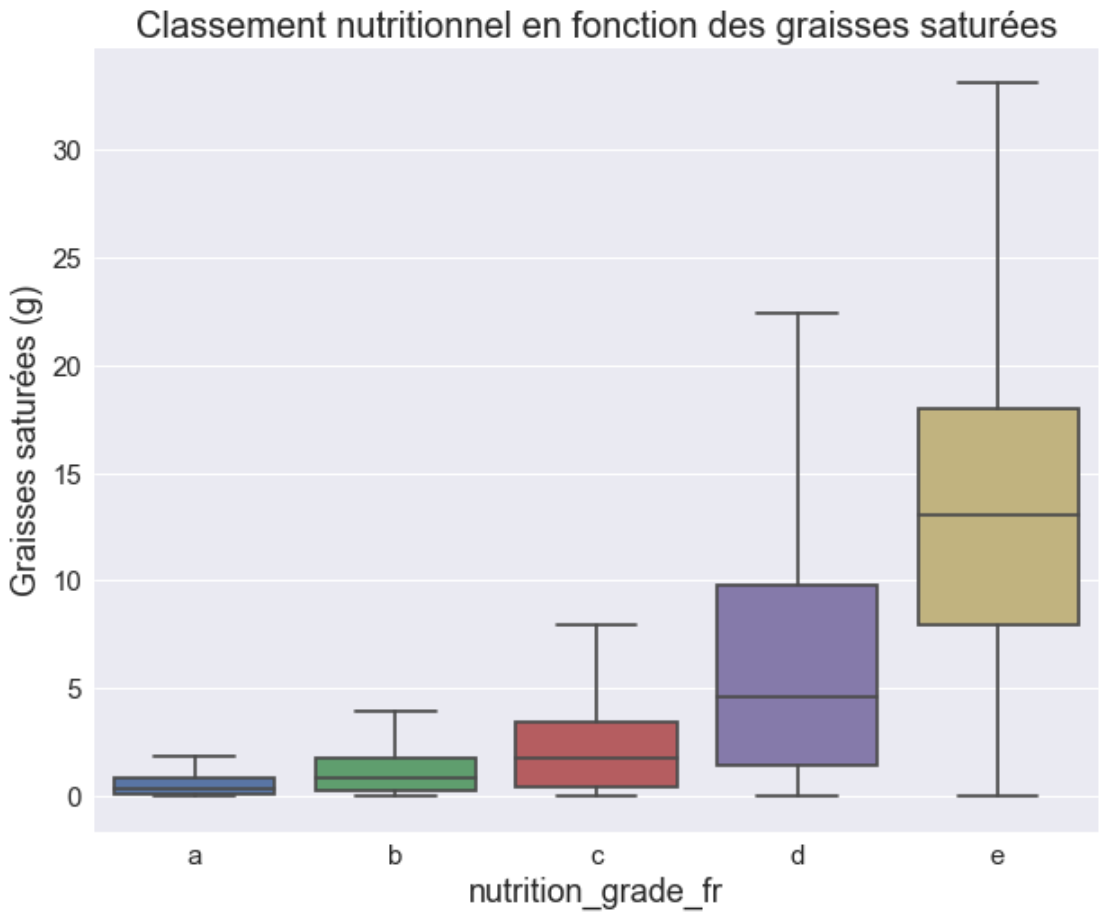
Analyse multivariée – Glucides/Sucres



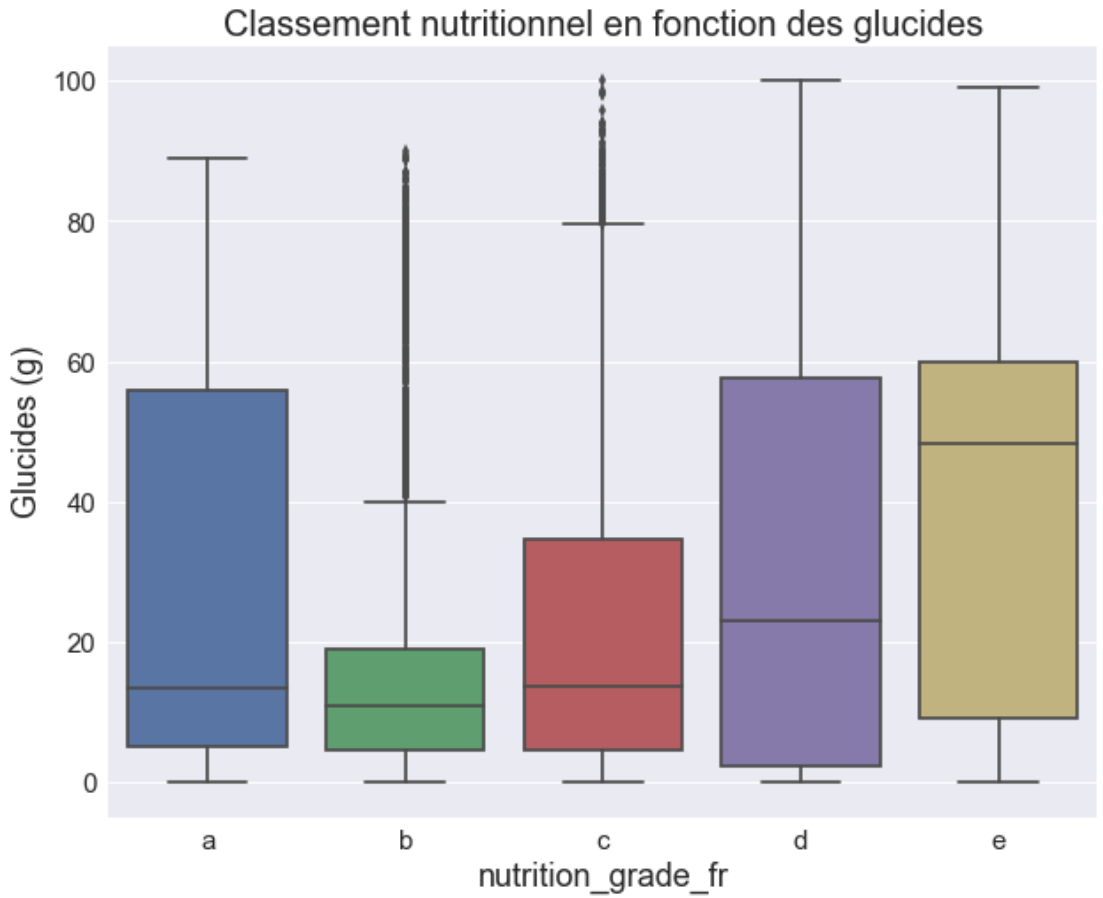
- Même constat que pour les lipides/graisses saturées, les sucres sont un type de glucides.
- L'inverse est fausse. C'est pourquoi on n'a pas plus l'allure d'une droite.
- Dans le cadre d'une alimentation saine, il faudra trouver des produits le plus loin possible de la ligne rouge.
- Par exemple, en bas à droite on retrouve les céréales comme le riz ou les pâtes.

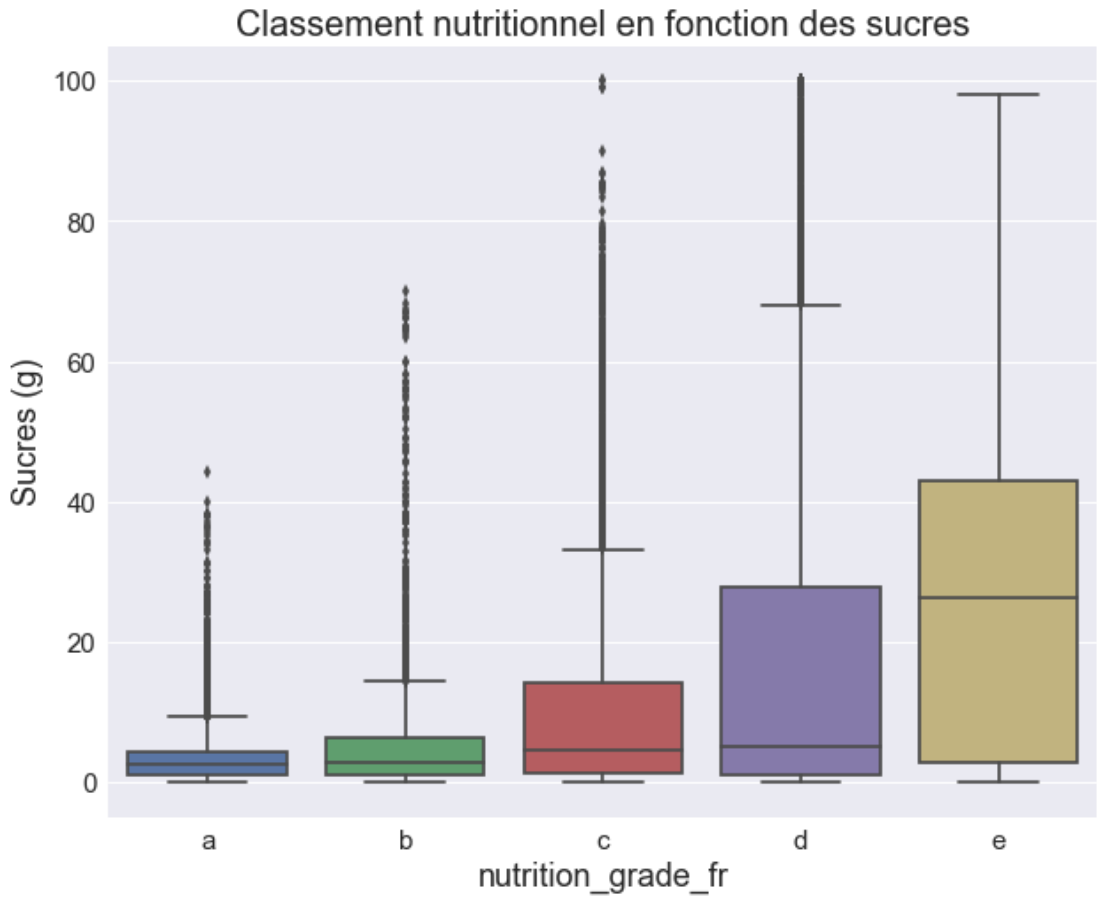
Analyse multivariée – Note nutritionnelle



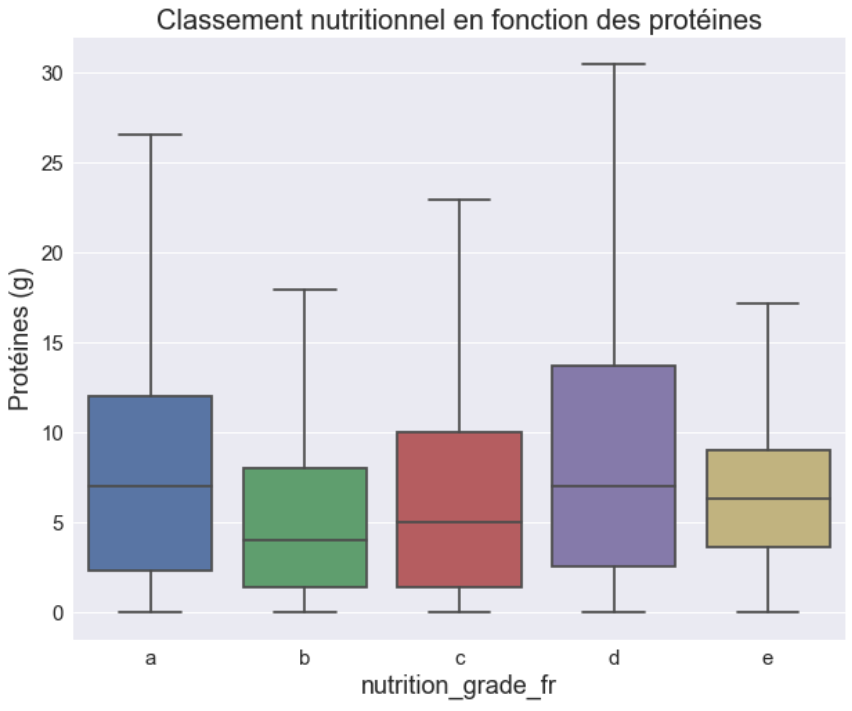


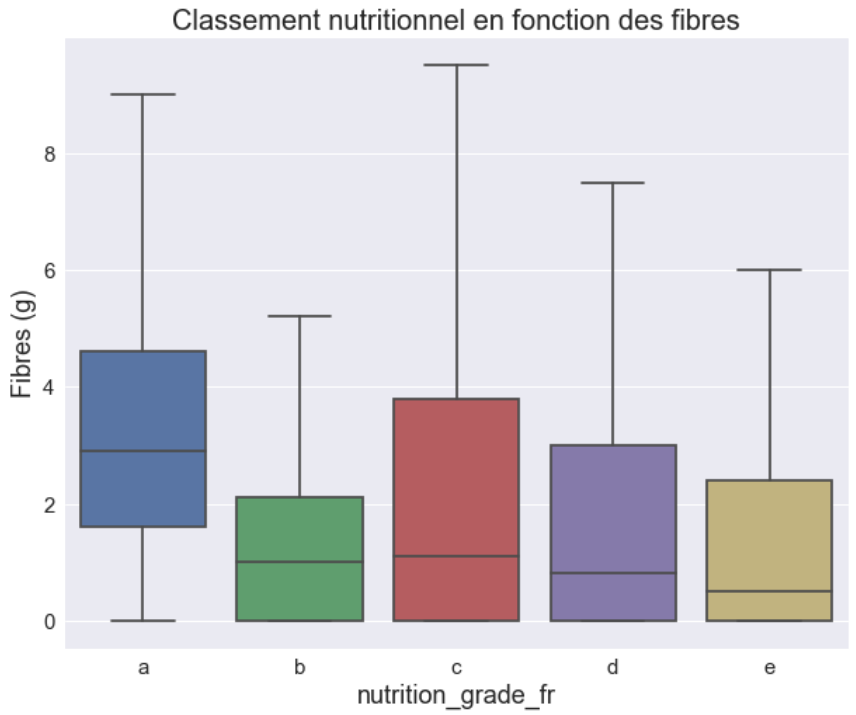
Analyse multivariée – Note nutritionnelle 2

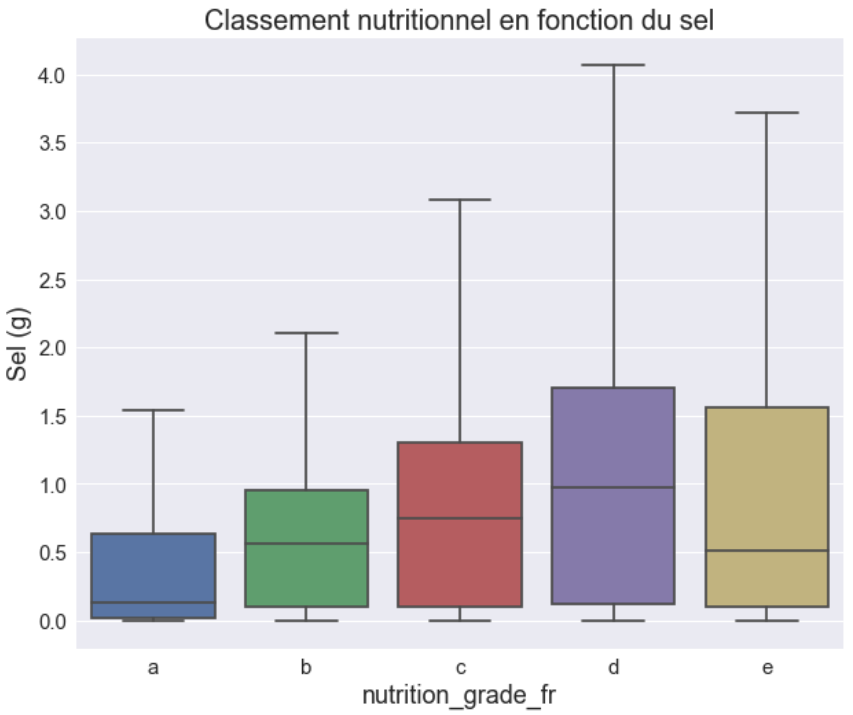




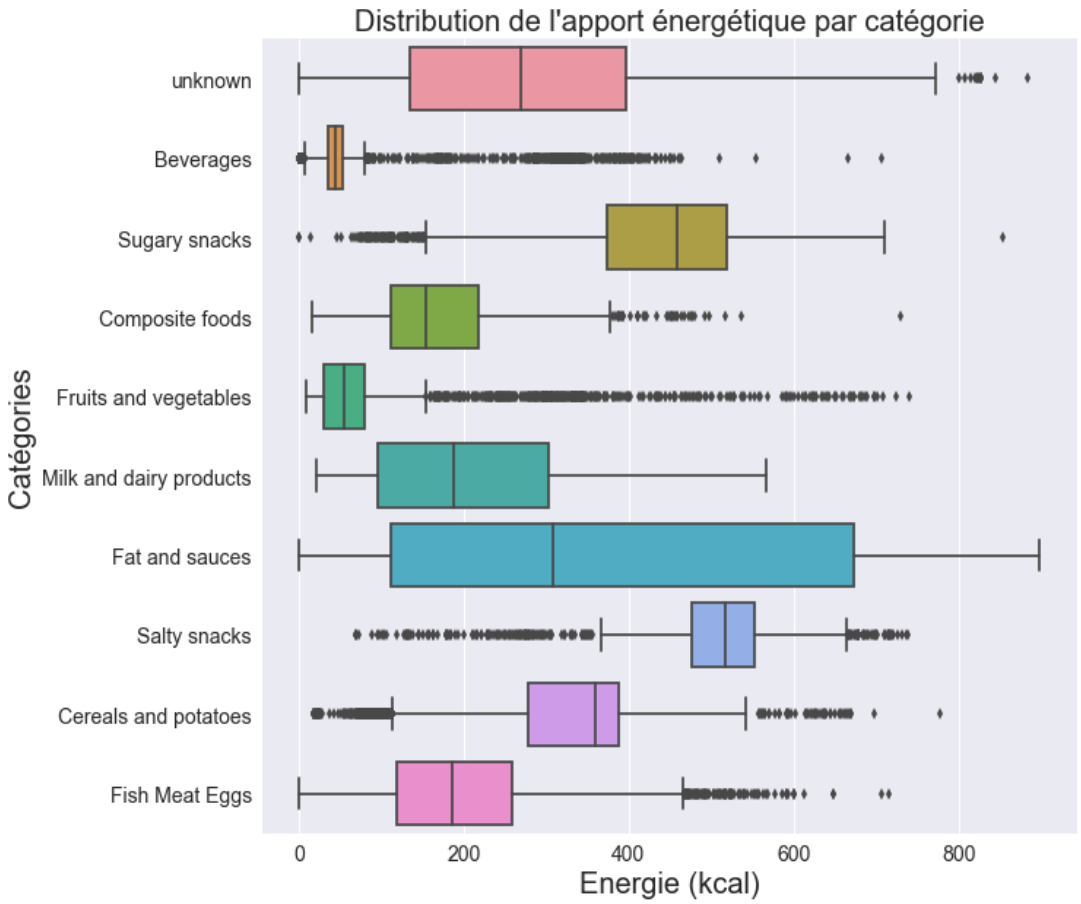
Analyse multivariée – Note nutritionnelle 3

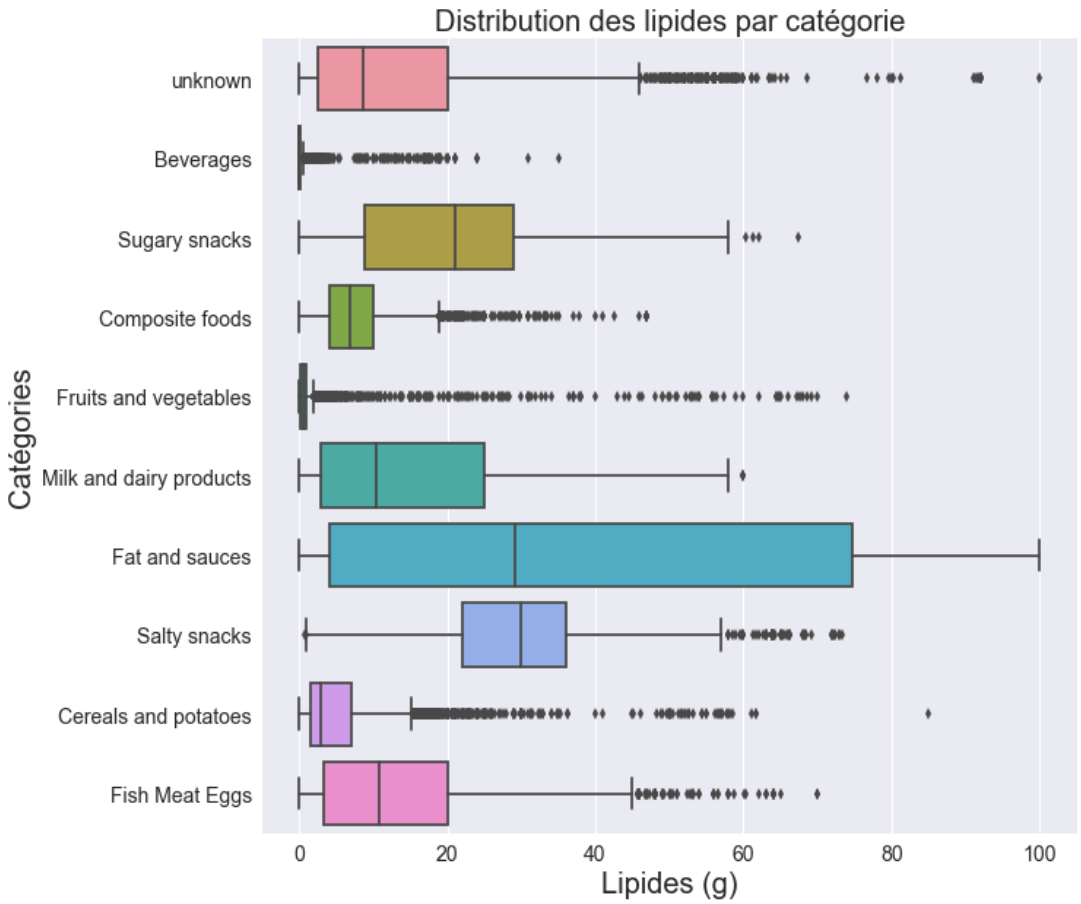




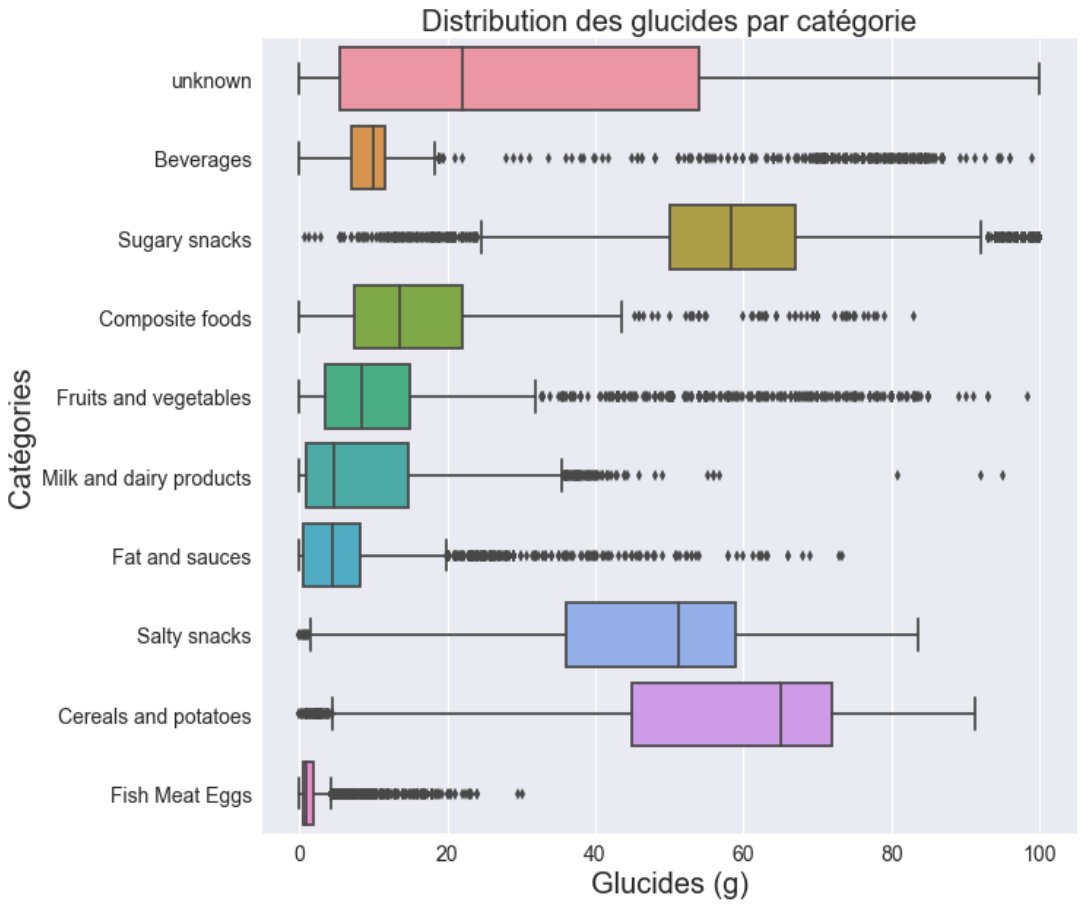


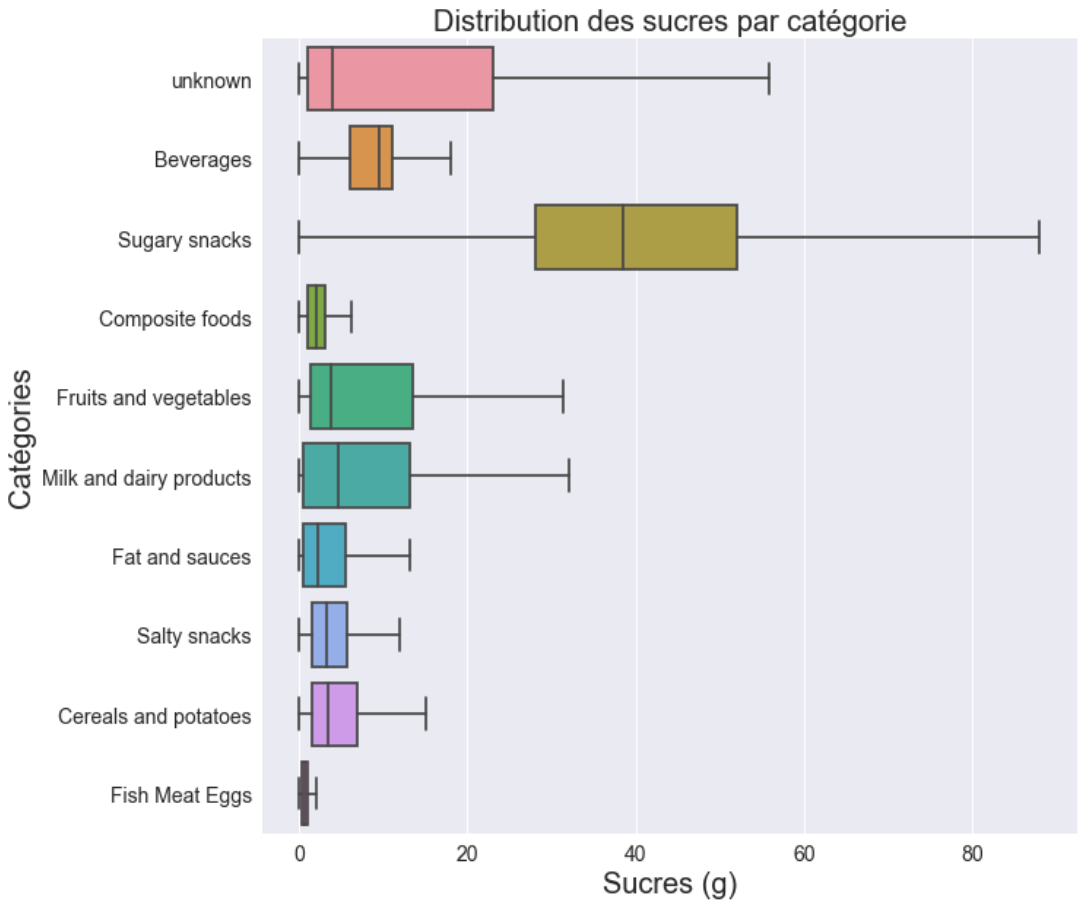
Analyse multivariée – Catégories 1



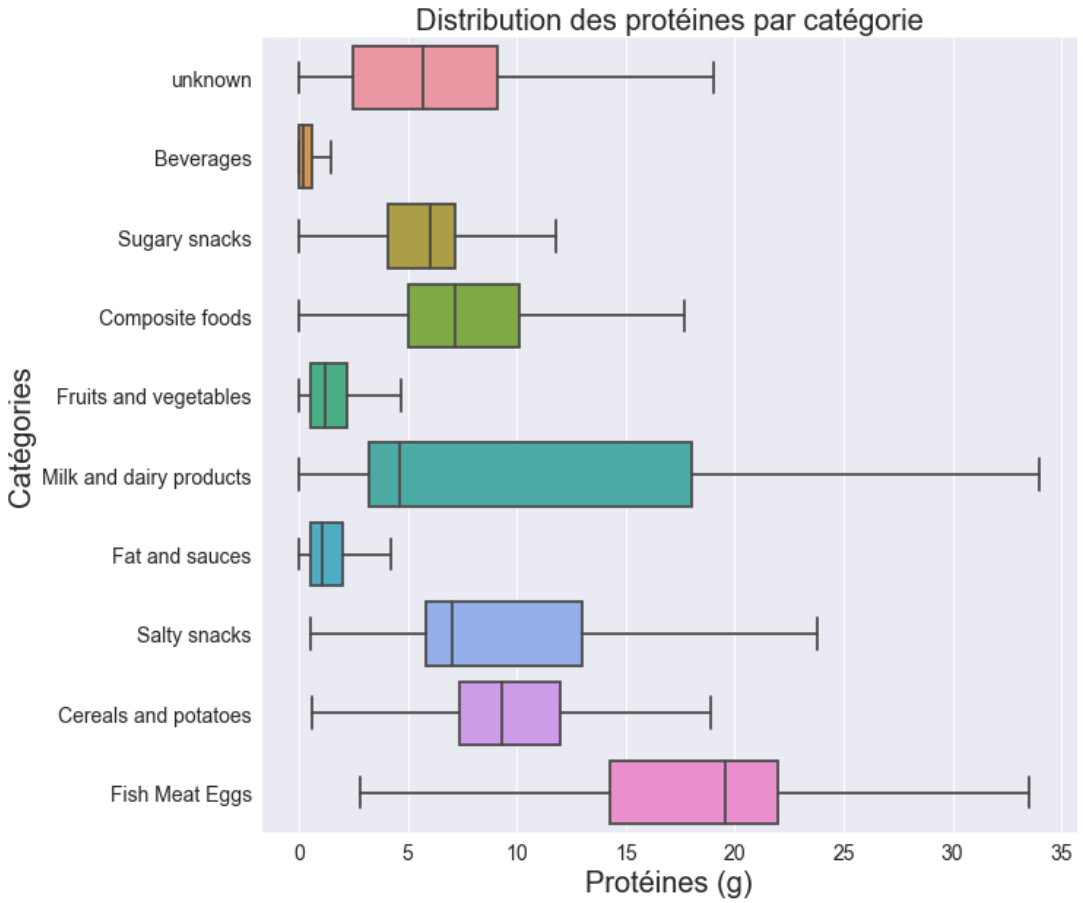


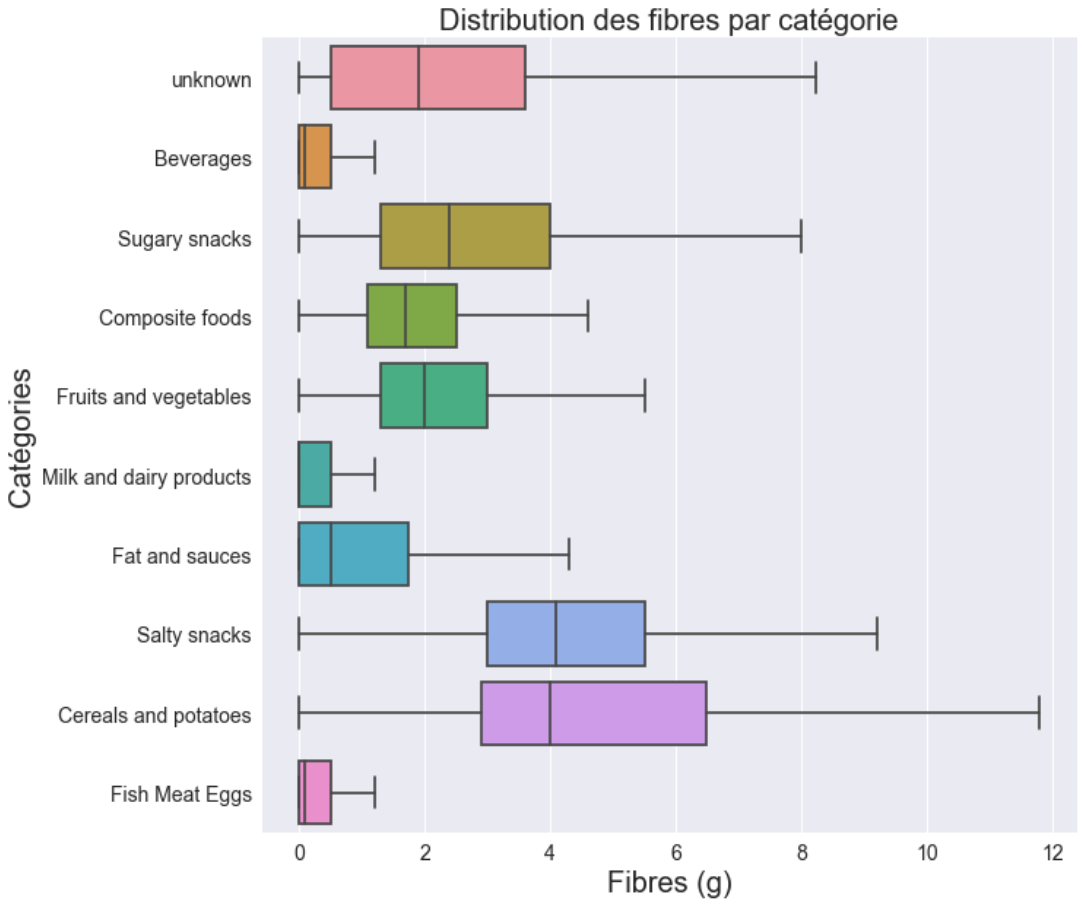
Analyse multivariée – Catégories 2



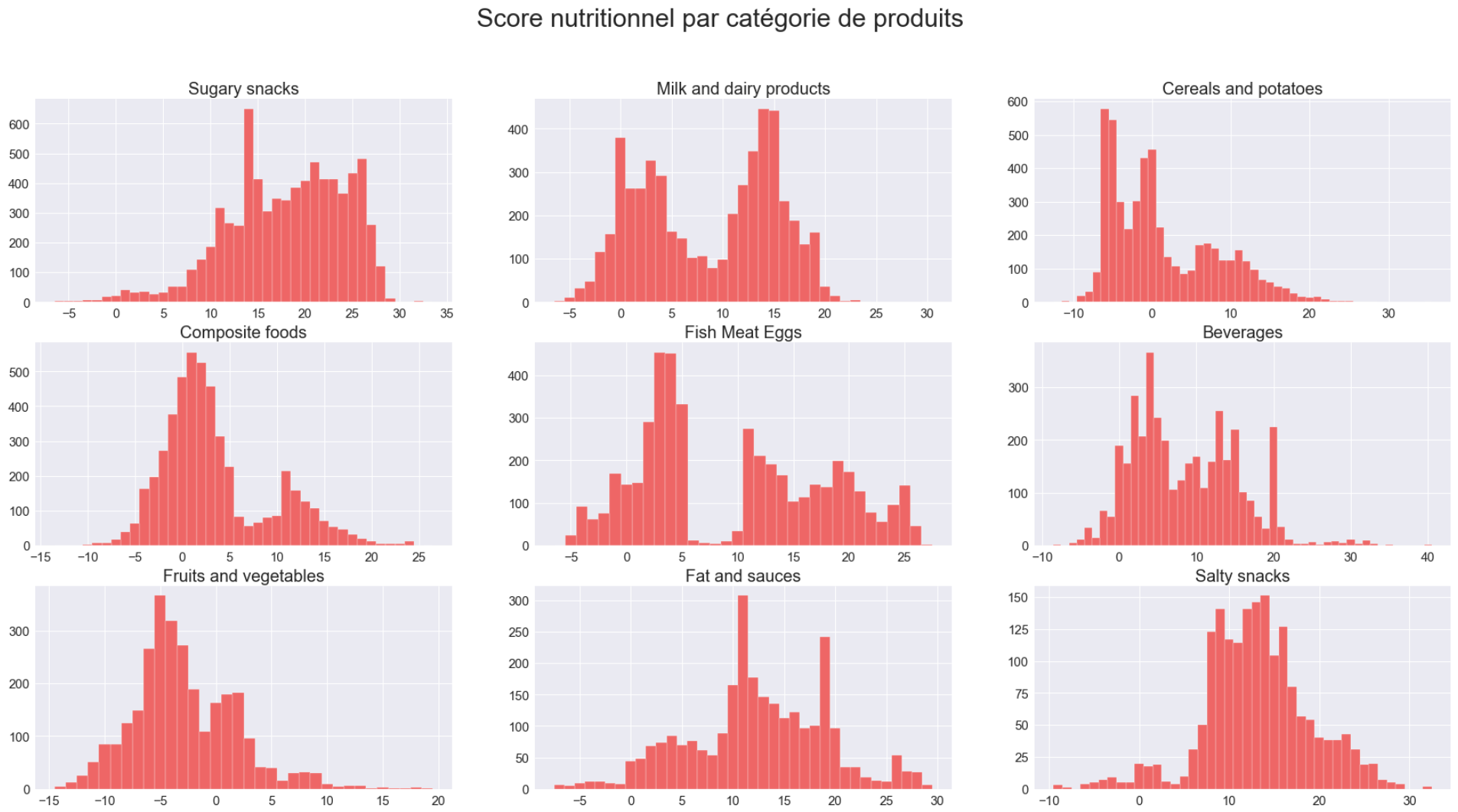


Analyse multivariée – Catégories 3

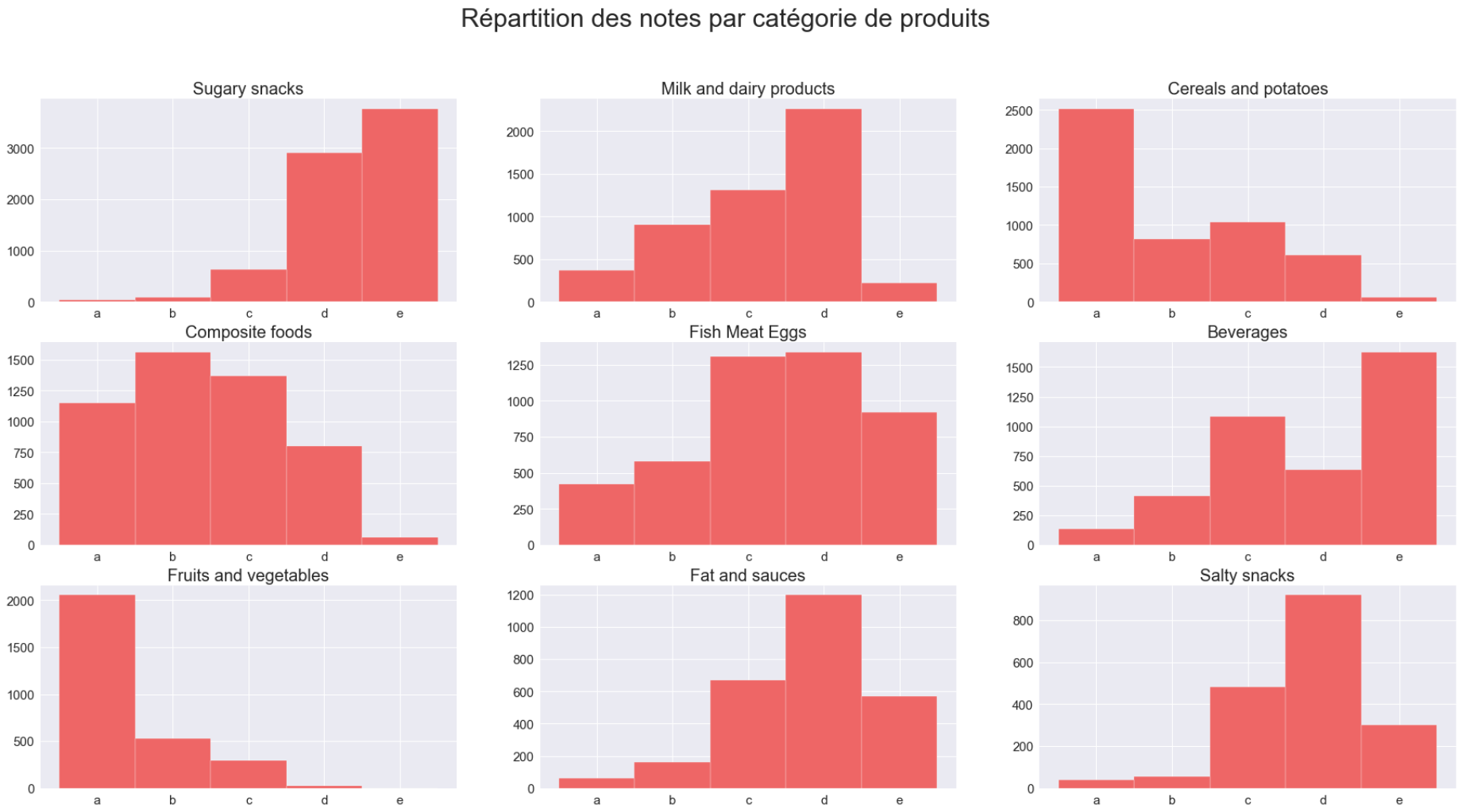




Analyse multivariée – Catégories 4

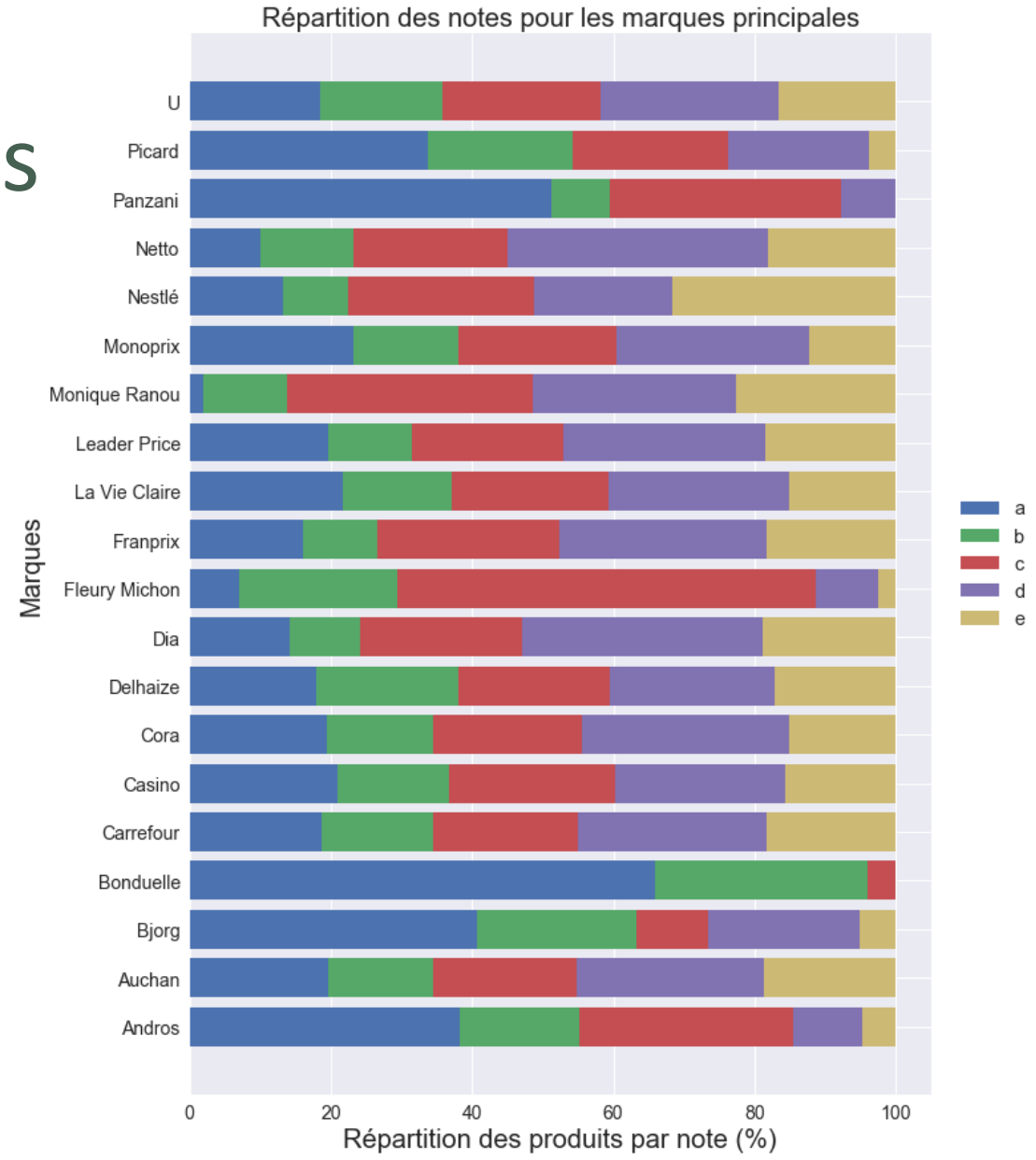


Analyse multivariée – Catégories 5



Analyse multivariée - Marques

- Mélange de marques et de distributeurs
- Filtrage possible pour garder les marques « saines »



Pistes de modélisation

- Filtrer davantage pour ne garder qu'un ensemble de produits sains:
  - Par score ou note nutritionnelle
  - Par catégorie de produits
  - Par marque

- Utiliser le score nutritionnel ou la note comme output des modèles