parcours data scientist, projet 2

Le site Lamarmite souhaite construire un générateur de recettes saines. Lamarmite vous fournit le lien vers une base de données qui recense de nombreux produits de consommation. Vous allez explorer ces données afin de déterminer ce que vous pouvez en tirer d’utile pour ce projet. Vous devrez donc vous intéresser aux avantages et désavantages nutritionnels des aliments.

Analyse de données nutritionnelles

Table des matières

[Tables des figures 2](#_Toc503451192)

[1 Introduction 3](#_Toc503451193)

[2 Principes de base de la diététique 3](#_Toc503451194)

[2.1 Recherche de données. 3](#_Toc503451195)

[2.2 Sel et sodium 4](#_Toc503451196)

[2.3 Conclusion 5](#_Toc503451197)

[3 Traitement du jeu de données 5](#_Toc503451198)

[3.1 Travail sur la base de données. 5](#_Toc503451199)

[3.2 Conclusion chiffrée. 14](#_Toc503451200)

[4 Analyse 14](#_Toc503451201)

[4.1 Analyse univariée 14](#_Toc503451202)

[4.2 Visualisations de relations entre certaines variables 24](#_Toc503451203)

[4.3 Quelques graphiques et conclusions associées 24](#_Toc503451204)

[5 Analyse multivariée 30](#_Toc503451205)

[5.1 Matrice des corrélations 30](#_Toc503451206)

[5.2 Régression linéaire et coefficient de corrélation 32](#_Toc503451207)

[6 Feature engineering 33](#_Toc503451208)

[6.1 Définition des intervalles considérés corrects et non aberrants. 33](#_Toc503451209)

[6.2 Détails des variables proposées et crées 34](#_Toc503451210)

[7 Conclusion 36](#_Toc503451211)

# Tables des figures

[Figure 1 - Apports de référence en énergie et macronutriments du règlement 1169/2011 3](#_Toc503379732)

[Figure 2 - AJR de certains nutriments 4](#_Toc503379733)

[Figure 3 - Tableau Inserm 4](#_Toc503379734)

[Figure 4 – Tableau récapitulatif des données conservées ou supprimées 9](#_Toc503379735)

[Figure 5 – Tableau de complétude des données conservées 12](#_Toc503379736)

[Figure 6 – Tableau de complétude intermédiaire 13](#_Toc503379737)

[Figure 7 – Tableau de complétude finale 13](#_Toc503379738)

[Figure 8 – Tableau de description des données 15](#_Toc503379739)

[Figure 9 – Histogramme energy\_100g 16](#_Toc503379740)

[Figure 10 – Histogramme proteins\_100g 16](#_Toc503379741)

[Figure 11 – Histogramme salt\_100g 17](#_Toc503379742)

[Figure 12 – Histogramme sodium\_100g 17](#_Toc503379743)

[Figure 13 – Histogramme sugars\_100g 18](#_Toc503379744)

[Figure 14 – Histogramme fat\_100g 18](#_Toc503379745)

[Figure 15 – Histogramme carbohydrates\_100g 19](#_Toc503379746)

[Figure 16 – Histogramme saturated-fat\_100g 19](#_Toc503379747)

[Figure 17 – Histogramme nutrition-score-fr\_100g 20](#_Toc503379748)

[Figure 18 – Histogramme fiber\_100g 20](#_Toc503379749)

[Figure 19 – Histogramme cholesterol\_100g 21](#_Toc503379750)

[Figure 20 – Histogramme trans-fat\_100g 21](#_Toc503379751)

[Figure 21 – Histogramme calcium\_100g 22](#_Toc503379752)

[Figure 22 – Histogramme vitamin-c\_100g 22](#_Toc503379753)

[Figure 23 – Histogramme iron\_100g 23](#_Toc503379754)

[Figure 24 – Histogramme vitamin-a\_100g 23](#_Toc503379755)

[Figure 25 - Energie avant traitement 24](#_Toc503379756)

[Figure 26 - Energie après traitement 24](#_Toc503379757)

[Figure 27 - Sel avant traitement 25](#_Toc503379758)

[Figure 28 - Sel après traitement 25](#_Toc503379759)

[Figure 29 - Sodium avant traitement 26](#_Toc503379760)

[Figure 30 - Sodium après traitement 26](#_Toc503379761)

[Figure 31 - Fibres avant traitement 27](#_Toc503379762)

[Figure 32 - Fibres après traitement 27](#_Toc503379763)

[Figure 33 - Vitamine C avant traitement 28](#_Toc503379764)

[Figure 34 - Vitamine C après traitement 28](#_Toc503379765)

[Figure 35 - Sucres avant traitement 29](#_Toc503379766)

[Figure 36 - Sucres après traitement 29](#_Toc503379767)

[Figure 37 - Tableau récapitulatif de conclusion 30](#_Toc503379768)

[Figure 38 - Corrélogramme 31](#_Toc503379769)

[Figure 39 – Taux de similitude en fonction des paliers 34](#_Toc503379770)

[Figure 40 – Répartition des aliments avec 1 tranche 35](#_Toc503379771)

[Figure 41 – Répartition des aliments avec 2 tranches 35](#_Toc503379772)

[Figure 42 – Répartition des aliments avec 3 tranches 36](#_Toc503379773)

# Introduction

L’objectif de ce projet est d’explorer des données afin de déterminer ce qu’il est possible d’en tirer d’utile pour le société cliente.

Pour ce faire, la démarche suivante a été établie :

* Une recherche sur les principes de base de la diététique.
* Le traitement du jeu de données
* L’analyse en elle-même. Elle comprend la visualisation de relations entre certaines variables, l’analyse univariée et l’analyse multivariée sur certaines variables. Enfin, le feature engineering.
* Etablissement d’une conclusion avant une partie consacrée à des conseils qui pourront être appliqués plus tard par la société cliente.

Le cheminement de ces démarches est expliqué dans les chapitres correspondants.

# Principes de base de la diététique

## Recherche de données.

Une recherche en ligne a permis d’en savoir plus sur les principes de bases de la diététique humaine. La définition d’apports de référence est intéressante car elle nous indique en quantité, les ressources nécessaires que l’homme doit absorber quotidiennement pour être en bonne santé. Ci-dessous, un tableau qui récapitule les besoins journaliers en énergie, matières grasses, acides gras, glucides, protéines et sel.

|  |  |
| --- | --- |
| Énergie ou nutriment | Apport de Référence |
| Énergie | 8 400 kJ (2 000 kcal) |
| Matières Grasses Totales | 70 gr |
| Acides Gras Saturés | 20 gr |
| Glucides | 260 gr |
| Sucres | 90 gr |
| Protéines | 50 gr |
| Sel | 6 gr |

Figure - Apports de référence en énergie et macronutriments du règlement 1169/2011

Cependant, les « apports » ne s’arrêtent pas là. Il existe également les apports journaliers recommandés (AJR) ne s’arrêtent pas là. D’autres nutriments essentiels, tels que les vitamines, ont été recensés. Ici aussi, un tableau récapitulatif nous permet de connaitre en détail les AJR.

|  |  |
| --- | --- |
| Nutriment | Apport journalier recommandé |
| Vitamine A (rétinol) | 800 μg |
| Vitamine B1 (thiamine) | 1,1 mg |
| Vitamine B2 (riboflavine) | 1,4 mg |
| Vitamine B3 (ou PP, niacine) | 16 mg |
| Vitamine B5 (acide pantothénique) | 6 mg |
| Vitamine B6 (pyridoxine) | 1,4 mg |
| Vitamine B8 ou H (biotine) | 50 μg |
| Vitamine B9 (acide folique) | 200 μg |
| Vitamine B12 (cobalamine) | 2,5 μg |
| Vitamine C (acide ascorbique) | 80 mg |
| Vitamine D (cholécalciférol) | 5 μg |
| Vitamine E (tocophérol) | 12 mg |
| Vitamine K (anti-AVK) | 75 μg |
| Calcium | 800 mg |
| Fer | 14 mg |
| Iode | 150 μg |
| Magnésium | 375 mg |
| Phosphore | 700 mg |
| Sélénium | 55 μg |
| Zinc | 10 mg |
| Potassium | 2 000 mg |
| Chlorure | 800 mg |
| Cuivre | 1 mg |
| Manganèse | 2 mg |
| Fluorure | 3,5 mg |
| Chrome | 40 μg |
| Molybdène | 50 μg |

Figure - AJR de certains nutriments

Enfin, l’Inserm (Institut national de la santé et de la recherche médicale) a étudié 13.000 produits de marque différentes. Sur ces produits, elle a calculé un score de nutrition (ou note nutritionnelle). Ces scores varient de -5 pour les mieux notés à 25 pour les moins recommandables.

Ce calcul de score (ou points) attribué à chacun des nutriments de la composante du produit est dit « négatif » selon la méthodologie développée par *Rayner et al.*

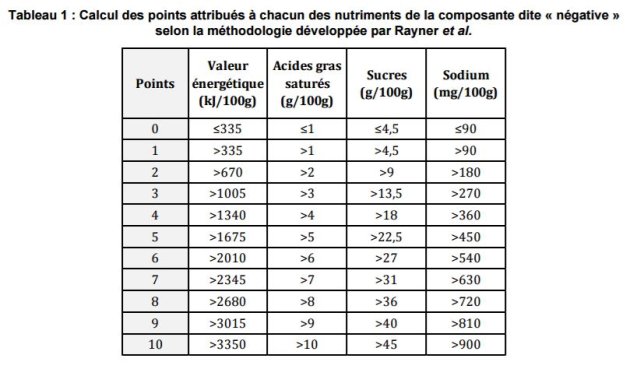


Figure - Tableau Inserm

## Sel et sodium

Le sel est considéré comme un aliment, dont le nom scientifique est le chlorure de sodium. Il est constitué de deux minéraux : le sodium (Na), pour 40 %, et le chlorure (Cl) pour 60 %. Pour 1 g de sel, nous retrouvons donc 400mg de Na et 600 mg de Cl. Le sodium est donc l’un des minéraux constituants du sel.

Il en résulte le calcul suivant. Pour retrouver une équivalence de la quantité de sel, il faut multiplier la quantité de sodium par 2.5.

Il se trouve sous forme de sel de table, mais aussi dans la plupart des aliments – surtout les coquillages et crustacés, les charcuteries, les fromages, les conserves, le pain – mais aussi dans les eaux de boissons, surtout pétillantes.

Les recommandations santé en matière de sel (5 g/jour) sont encore loin d’être atteintes aujourd’hui, malgré une baisse de la consommation de sel constatées depuis ces dernières années en France.

## Conclusion

Quelques données chiffrées retrouvées dans ces recherches sont importantes et vont être utilisées pour les analyses :

* Notes nutritionnelles.
* Apports Journaliers recommandés
* Apports de référence

# Traitement du jeu de données

Open Food Facts est une base de données sur les produits alimentaires faite par tout le monde, pour tout le monde. Elle permet de faire des choix plus informés, et comme les données sont ouvertes (open data), tout le monde peut les utiliser pour tout usage.

Cette base de données, bien que publique et ouverte, est la base de nombreux travaux français (et mondiaux). Je prends comme exemple l'équipe du Professeur Hercberg du Programme National Nutrition et Santé (PNNS) qui utilise les données d'Open Food Facts pour valider et affiner les formules des notes nutritionnelles proposées dans le cadre de la Loi Santé d'avril 2015.

C’est cette base de données qui va nous servir de fondement à notre travail.

## Travail sur la base de données.

Cette base de données est immense, et inutilisable telle qu’elle est récupérée sur le site. Il a donc fallu faire un certain nombre d’opérations pour la rendre propre, et cela a été divisée en trois parties. Nous nous intéressons aussi à la description du choix effectué sur les données manquantes et aberrantes.

### Premier nettoyage de gros

Un premier nettoyage de « gros » est fait afin de dégrossir très largement la base de données. Certaines colonnes ou lignes étant complètement vides, elles sont inutiles et ont donc été supprimées.

Un choix a été fait ensuite de conserver les données nutritionnelles uniquement. Ci-dessous, dans le détail, le choix qui a été fait pour chaque donnée avec sa justification.

|  |  |  |
| --- | --- | --- |
| Nom de la colonne | Choix | Justification |
| "code | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| url | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| creator | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| created\_t | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| created\_datetime | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| last\_modified\_t | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| last\_modified\_datetime | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| product\_name | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| generic\_name | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| quantity | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| packaging | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| packaging\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| brands | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| brands\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| categories | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| categories\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| categories\_fr | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| origins | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| origins\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| manufacturing\_places | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| manufacturing\_places\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| labels | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| labels\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| labels\_fr | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| emb\_codes | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| emb\_codes\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| first\_packaging\_code\_geo | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| cities | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| cities\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| purchase\_places | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| stores | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| countries | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| countries\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| countries\_fr | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| ingredients\_text | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| allergens | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| allergens\_fr | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| traces | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| traces\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| traces\_fr | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| serving\_size | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| no\_nutriments | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| additives\_n | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| additives | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| additives\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| additives\_fr | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| ingredients\_from\_palm\_oil\_n | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| ingredients\_from\_palm\_oil | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| ingredients\_from\_palm\_oil\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| ingredients\_that\_may\_be\_from\_palm\_oil\_n | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| ingredients\_that\_may\_be\_from\_palm\_oil | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| ingredients\_that\_may\_be\_from\_palm\_oil\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| nutrition\_grade\_uk | Supprimée | Peut être conservée si le calcul choisi est différent de la méthodologie donnée dans ce document. |
| nutrition\_grade\_fr | Conservée | Utile dans le feature engineering. |
| pnns\_groups\_1 | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| pnns\_groups\_2 | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| states | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| states\_tags | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| states\_fr | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| main\_category | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| main\_category\_fr | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| image\_url | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| image\_small\_url | Supprimée | Considérée sans lien avec la valeur nutritionnelle |
| energy\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| energy-from-fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| saturated-fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| butyric-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| caproic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| caprylic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| capric-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| lauric-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| myristic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| palmitic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| stearic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| arachidic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| behenic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| lignoceric-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| cerotic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| montanic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| melissic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| monounsaturated-fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| polyunsaturated-fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| omega-3-fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| alpha-linolenic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| eicosapentaenoic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| docosahexaenoic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| omega-6-fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| linoleic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| arachidonic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| gamma-linolenic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| dihomo-gamma-linolenic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| omega-9-fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| oleic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| elaidic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| gondoic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| mead-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| erucic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| nervonic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| trans-fat\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| cholesterol\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| carbohydrates\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| sugars\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| sucrose\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| glucose\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| fructose\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| lactose\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| maltose\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| maltodextrins\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| starch\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| polyols\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| fiber\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| proteins\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| casein\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| serum-proteins\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| nucleotides\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| salt\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| sodium\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| alcohol\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-a\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| beta-carotene\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-d\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-e\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-k\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-c\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-b1\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-b2\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-pp\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-b6\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-b9\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| folates\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| vitamin-b12\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| biotin\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| pantothenic-acid\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| silica\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| bicarbonate\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| potassium\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| chloride\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| calcium\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| phosphorus\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| iron\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| magnesium\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| zinc\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| copper\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| manganese\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| fluoride\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| selenium\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| chromium\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| molybdenum\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| iodine\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| caffeine\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| taurine\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| ph\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| fruits-vegetables-nuts\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| collagen-meat-protein-ratio\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| cocoa\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| chlorophyl\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| carbon-footprint\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| nutrition-score-fr\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| nutrition-score-uk\_100g | Supprimée | Le client étant français, nous avons conservé le calcul français et non pas britannique. |
| glycemic-index\_100g | Conservée | Rentre dans le calcul de la valeur nutritionnelle |
| water-hardness\_100g" | Conservée | Rentre dans le calcul de la valeur nutritionnelle |

Figure – Tableau récapitulatif des données conservées ou supprimées

En résumé, la démarche suivante a été appliquée :

* Récupération de la base de données.
* Sélection des colonnes qui nous intéressent. Par « colonne », il faut entendre « donnée nutritionnelle ». Les colonnes gardées sont toutes celles qui contiennent des valeurs nutritionnelles pour 100g de produit.
* Suppression des colonnes complètement vides.
* Suppression des lignes complètement vides.

### Deuxième nettoyage

Ensuite, il a fallu peaufiner le premier nettoyage effectué avec un deuxième, plus raffiné.

* Définition d’une limite de données manquantes pour considérer que la colonne peut être utilisée telle qu’elle.
* Application de cette limite et suppression des colonnes qui ne remplissent pas cette condition.

La limite de données a été fixée à 200 000 données manquantes pour les (environ) 320 000 données maximales disponibles ce qui donne un taux de complétude minimale acceptable de **40** % pour que la colonne soit gardée et considérée comme utilisable dans l’analyse.

|  |  |  |
| --- | --- | --- |
| column\_name | missing\_count | filling\_factor |
| energy\_100g | 59701 | 81.38944047682136 |
| proteins\_100g | 60888 | 81.01941762705313 |
| salt\_100g | 65289 | 79.64749634497227 |
| sodium\_100g | 65347 | 79.62941603723297 |
| sugars\_100g | 75838 | 76.35906244252489 |
| fat\_100g | 76919 | 76.02208291379746 |
| carbohydrates\_100g | 77226 | 75.92638197455665 |
| saturated-fat\_100g | 91255 | 71.55312960775085 |
| potassium\_100g | 296044 | 7.7143685452522055 |
| polyunsaturated-fat\_100g | 297932 | 7.125823355393387 |
| monounsaturated-fat\_100g | 297968 | 7.114601095417266 |
| nutrition-score-fr\_100g | 99600 | 68.9517473993971 |
| fiber\_100g | 119915 | 62.61896374898298 |
| cholesterol\_100g | 176701 | 44.9170955544264 |
| trans-fat\_100g | 177493 | 44.67020583495172 |
| calcium\_100g | 179742 | 43.96912631588791 |
| vitamin-c\_100g | 179925 | 43.91207982767597 |
| iron\_100g | 180329 | 43.78614113238838 |
| vitamin-a\_100g | 183237 | 42.879631909872785 |
| En dessous de cette ligne, les données en sont pas suffisantes pour être exploitables | | |
| vitamin-pp\_100g | 309062 | 3.656274646109149 |
| vitamin-b1\_100g | 309637 | 3.4770302159349855 |
| vitamin-b2\_100g | 309976 | 3.3713539344931744 |
| vitamin-d\_100g | 313734 | 2.1998746847636 |
| vitamin-b6\_100g | 314007 | 2.114772546611345 |
| magnesium\_100g | 314538 | 1.9492442119635527 |
| phosphorus\_100g | 314946 | 1.8220585989008422 |
| vitamin-b12\_100g | 315491 | 1.6521660520401134 |
| vitamin-b9\_100g | 315551 | 1.6334622854132441 |
| alcohol\_100g | 316646 | 1.292118544472881 |
| zinc\_100g | 316862 | 1.2247849846161518 |
| folates\_100g | 317749 | 0.9482809679822688 |
| fruits-vegetables-nuts\_100g | 317764 | 0.9436050263255515 |
| pantothenic-acid\_100g | 318308 | 0.774024208908604 |
| copper\_100g | 318685 | 0.6565022086031093 |
| manganese\_100g | 319171 | 0.5050016989254686 |
| vitamin-e\_100g | 319450 | 0.4180291841105268 |
| selenium\_100g | 319623 | 0.36409999033638724 |
| cocoa\_100g | 319843 | 0.29551951270453347 |
| vitamin-k\_100g | 319873 | 0.2861676293910989 |
| energy-from-fat\_100g | 319934 | 0.2671521333204485 |
| omega-3-fat\_100g | 319950 | 0.26216446221995005 |
| polyols\_100g | 320372 | 0.1306146369443033 |
| biotin\_100g | 320461 | 0.10287071644778063 |
| carbon-footprint\_100g | 320523 | 0.08354349093334912 |
| starch\_100g | 320525 | 0.08292003204578682 |
| lactose\_100g | 320529 | 0.0816731142706622 |
| iodine\_100g | 320532 | 0.08073792593931875 |
| omega-6-fat\_100g | 320603 | 0.058605135430856846 |
| alpha-linolenic-acid\_100g | 320605 | 0.057981676543294544 |
| collagen-meat-protein-ratio\_100g | 320626 | 0.051435358223890316 |
| chloride\_100g | 320633 | 0.04925325211742224 |
| linoleic-acid\_100g | 320642 | 0.04644768712339187 |
| bicarbonate\_100g | 320687 | 0.032419862153239956 |
| fluoride\_100g | 320712 | 0.02462662605871112 |
| docosahexaenoic-acid\_100g | 320713 | 0.02431489661492997 |
| caffeine\_100g | 320713 | 0.02431489661492997 |
| sucrose\_100g | 320719 | 0.022444519952243048 |
| ph\_100g | 320742 | 0.015274742745276518 |
| eicosapentaenoic-acid\_100g | 320753 | 0.011845718863683833 |
| fructose\_100g | 320753 | 0.011845718863683833 |
| silica\_100g | 320753 | 0.011845718863683833 |
| beta-carotene\_100g | 320758 | 0.010287071644778064 |
| taurine\_100g | 320762 | 0.00904015386965345 |
| casein\_100g | 320764 | 0.008416694982091143 |
| glucose\_100g | 320765 | 0.008104965538309989 |
| arachidic-acid\_100g | 320767 | 0.007481506650747684 |
| gamma-linolenic-acid\_100g | 320767 | 0.007481506650747684 |
| behenic-acid\_100g | 320768 | 0.007169777206966529 |
| dihomo-gamma-linolenic-acid\_100g | 320768 | 0.007169777206966529 |
| omega-9-fat\_100g | 320770 | 0.006546318319404223 |
| nucleotides\_100g | 320771 | 0.006234588875623069 |
| chromium\_100g | 320771 | 0.006234588875623069 |
| serum-proteins\_100g | 320775 | 0.004987671100498455 |
| gondoic-acid\_100g | 320777 | 0.004364212212936149 |
| oleic-acid\_100g | 320778 | 0.0040524827691549945 |
| caprylic-acid\_100g | 320779 | 0.003740753325373842 |
| palmitic-acid\_100g | 320779 | 0.003740753325373842 |
| stearic-acid\_100g | 320779 | 0.003740753325373842 |
| maltodextrins\_100g | 320780 | 0.003429023881592688 |
| molybdenum\_100g | 320780 | 0.003429023881592688 |
| arachidonic-acid\_100g | 320783 | 0.0024938355502492275 |
| caproic-acid\_100g | 320785 | 0.001870376662686921 |
| lauric-acid\_100g | 320787 | 0.0012469177751246137 |
| maltose\_100g | 320787 | 0.0012469177751246137 |
| capric-acid\_100g | 320789 | 0.0006234588875623069 |
| myristic-acid\_100g | 320790 | 0.00031172944378115344 |
| montanic-acid\_100g | 320790 | 0.00031172944378115344 |
| butyric-acid\_100g | 320791 | 0.0 |
| lignoceric-acid\_100g | 320791 | 0.0 |
| cerotic-acid\_100g | 320791 | 0.0 |
| melissic-acid\_100g | 320791 | 0.0 |
| elaidic-acid\_100g | 320791 | 0.0 |
| mead-acid\_100g | 320791 | 0.0 |
| erucic-acid\_100g | 320791 | 0.0 |
| nervonic-acid\_100g | 320791 | 0.0 |
| chlorophyl\_100g | 320791 | 0.0 |
| glycemic-index\_100g | 320791 | 0.0 |

Figure – Tableau de complétude des données conservées

La valeur est faible en pourcentage, mais on constate que la base de données est vraiment de mauvaise qualité. En effet, son taux initial de complétude est de **24,4 %** (cette valeur est calculée grâce au code suivant : missing\_data['filling\_factor'].mean(). Pour plus de détails, voir le fichier joint). Cela revient à dire que les ¾ des cases ne sont pas remplies !

### Troisième nettoyage

Enfin, un troisième et dernier nettoyage, qui a amené des choix, est fait :

* Parmi les colonnes sélectionnées, certaines valeurs (ou points) sont aberrantes :
* Les plus évidentes :
* On ne pouvait pas garder des poids (en gramme) négatif.
* On ne pouvait pas garder des poids (en gramme) qui étaient supérieur à 100g pour un nutriment. Il n’est clairement pas possible d’avoir (par exemple) 150g de graisses pour 100g de produit.
* Celles qui ont demandées un choix :
* Après avoir supprimer les valeurs négatives, seules les valeurs qui étaient inférieures au 98ème quantile ont été gardées[[1]](#footnote-1). Les autres valeurs ont été considérées comme aberrantes ou hors-normes.
  + Aberrantes quand, par exemple, on trouvait 90g de graisses pour 100g de produit.
  + Hors-normes, pour les nutriments particuliers comme les cacahuètes.

On retrouve les taux de complétudes suivant pour les données sélectionnées :

|  |  |  |
| --- | --- | --- |
| column\_name | missing\_count | filling\_factor |
| energy\_100g | 4351 | 98.36084101551758 |
| proteins\_100g | 5538 | 97.91366066282148 |
| salt\_100g | 9939 | 96.25566510071918 |
| sodium\_100g | 9997 | 96.23381467068012 |
| sugars\_100g | 20488 | 92.28152395447576 |
| fat\_100g | 21569 | 91.87427714633384 |
| carbohydrates\_100g | 21876 | 91.75862055974774 |
| saturated-fat\_100g | 35905 | 86.4734536111603 |
| nutrition-score-fr\_100g | 44250 | 83.32962880640142 |
| fiber\_100g | 64565 | 75.67632731944198 |
| cholesterol\_100g | 121351 | 54.283249384985744 |
| trans-fat\_100g | 122143 | 53.984877995486755 |
| calcium\_100g | 124392 | 53.137608734144315 |
| vitamin-c\_100g | 124575 | 53.06866686005554 |
| iron\_100g | 124979 | 52.91646731288686 |
| vitamin-a\_100g | 127887 | 51.82093195851433 |

Figure – Tableau de complétude intermédiaire

* Parmi les colonnes sélectionnées, certaines valeurs (ou points) sont manquantes :
* Pour les lignes, il a été envisagé de remplacer les valeurs manquantes par une valeur nulle. Il aurait été possible également de remplacer par la valeur moyenne (mean) du nutriment en question, cependant la valeur nulle est plus neutre que la valeur moyenne dans les calculs.
* Cette manœuvre est effectuée sauf pour le score nutritionnel. En effet, il était risqué de mettre une valeur de 0 qui a une signification importante (bon produit). Les valeurs manquantes ici n’ont pas été remplacées, elles ont été supprimées de la base de données.

Vu la quantité de données disponible, aucun remplacement n’a été effectué. Seules les lignes complètes avec ces éléments ont été conservées. Par ailleurs, pour aller plus loin, on pourrait étudier un moyen de calculer (même de manière approximative) un substitut pour remplir les données manquantes.

A la fin du troisième et dernier nettoyage, on retrouve les taux de complétudes suivant pour les données sélectionnées :

|  |  |  |
| --- | --- | --- |
| column\_name | missing\_count | filling\_factor |
| energy\_100g | 0 | 100 |
| proteins\_100g | 0 | 100 |
| salt\_100g | 0 | 100 |
| sodium\_100g | 0 | 100 |
| sugars\_100g | 0 | 100 |
| fat\_100g | 0 | 100 |
| carbohydrates\_100g | 0 | 100 |
| saturated-fat\_100g | 0 | 100 |
| nutrition-score-fr\_100g | 0 | 100 |
| fiber\_100g | 0 | 100 |
| cholesterol\_100g | 0 | 100 |
| trans-fat\_100g | 0 | 100 |
| calcium\_100g | 0 | 100 |
| vitamin-c\_100g | 0 | 100 |
| iron\_100g | 0 | 100 |
| vitamin-a\_100g | 0 | 100 |

Figure – Tableau de complétude finale

## Conclusion chiffrée.

La base de données contenait environ 320 000 lignes de départ et 162 colonnes de départ. Nous n’avions aucune idée de la pertinence des informations s’y trouvant. Après le « ménage » effectué, on comptabilise environ 97 000 lignes et 16 colonnes, cela correspond environ à **25** % de la base initiale.

Ce chiffre peut paraitre bas, mais il ne l’est pas car beaucoup de données ne nous intéressent pas, notamment les données qui ne concernent pas les valeurs nutritionnelles des aliments (tag, photo, heure de mise à jour etc.) Plus de la moitié des colonnes supprimées étaient inutiles.

En gardant cette base plus petite, elle sera aussi plus facile à mettre à jour suivant les besoins du client, si nécessaire.

Enfin, notons que les données « nutrition-score-fr\_100g » ne sont pas complètes. Pour aller plus loin, on pourrait essayer de définir une méthode pour retrouver ce score à partir des données déjà existantes de la base.

# Analyse

Dans ce chapitre, nous voyons la description et l'analyse univariée des différentes variables importantes avec leurs visualisations associées.

## Analyse univariée

Dans cette partie, nous allons analyser toutes les données que nous avons sélectionnées dans le chapitre précèdent de manière indépendante, c’est l’analyse univariée.

### Données

#### Description des données

Dans cette partie d’analyse, nous pouvons voir, pour chaque nutriment sélectionné :

* Le nombre de valeurs recensé,
* La moyenne,
* L’écart-type (déviation standard),
* La valeur minimale,
* La valeur maximale,
* Les 25ème, 50ème (médiane) et 75ème quantiles.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| index | count | mean | std | min | 25% | 50% | 75% | max |
| energy\_100g | 97884 | 1125,10 | 715,39 | 0 | 418 | 1117 | 1674 | 2870 |
| fat\_100g | 97884 | 12,04 | 12,88 | 0 | 1,02 | 7,395 | 20,51 | 50 |
| saturated-fat\_100g | 97884 | 4,20 | 5,50 | 0 | 0 | 1,79 | 6,67 | 22 |
| trans-fat\_100g | 97884 | 0,00 | 0,00 | 0 | 0 | 0 | 0 | 0,158 |
| cholesterol\_100g | 97884 | 0,02 | 0,03 | 0 | 0 | 0 | 0,02 | 0,107 |
| carbohydrates\_100g | 97884 | 33,74 | 26,60 | 0 | 8,8 | 26,67 | 58,06 | 90 |
| sugars\_100g | 97884 | 13,69 | 16,53 | 0 | 1,75 | 5,56 | 22,22 | 65 |
| fiber\_100g | 97884 | 2,35 | 2,88 | 0 | 0 | 1,5 | 3,6 | 14,7 |
| proteins\_100g | 97884 | 7,21 | 6,59 | 0 | 2,5 | 5,38 | 10,39 | 28,57 |
| salt\_100g | 97884 | 0,88 | 0,81 | 0 | 0,15494 | 0,73406 | 1,35382 | 3,71856 |
| sodium\_100g | 97884 | 0,35 | 0,32 | 0 | 0,061 | 0,289 | 0,533 | 1,464 |
| vitamin-a\_100g | 97884 | 0,00 | 0,00 | 0 | 0 | 0 | 9,51E-05 | 0,0010713 |
| vitamin-c\_100g | 97884 | 0,00 | 0,01 | 0 | 0 | 0 | 0,002 | 0,047 |
| calcium\_100g | 97884 | 0,08 | 0,14 | 0 | 0 | 0,033 | 0,1 | 0,714 |
| iron\_100g | 97884 | 0,00 | 0,00 | 0 | 0 | 0,0009 | 0,00212 | 0,00675 |
| nutrition-score-fr\_100g | 97884 | 8,58 | 9,01 | -10 | 1 | 8 | 16 | 36 |

Figure – Tableau de description des données

#### Conclusion

Toutes les données apparaissent le même nombre de fois (97884), ce qui prouve que le « ménage » effectué au précédent chapitre est respecté.

Certaines données ont des valeurs nulles très souvent, il s’agit notamment du cas de :

* trans-fat\_100g
* vitamin-a\_100g
* vitamin-c\_100g
* iron\_100g

Nous allons maintenant, dans le chapitre suivant, créer une vue directe de la répartition des valeurs grâce aux histogrammes.

### Histogrammes

Afin d’avoir une « vue » rapide de la distribution des données, des histogrammes ont été créés et reproduit dans ce document.

#### energy\_100g

La distribution est assez régulière sur l’ensemble des valeurs. C’est une distribution gaussienne multimodale.

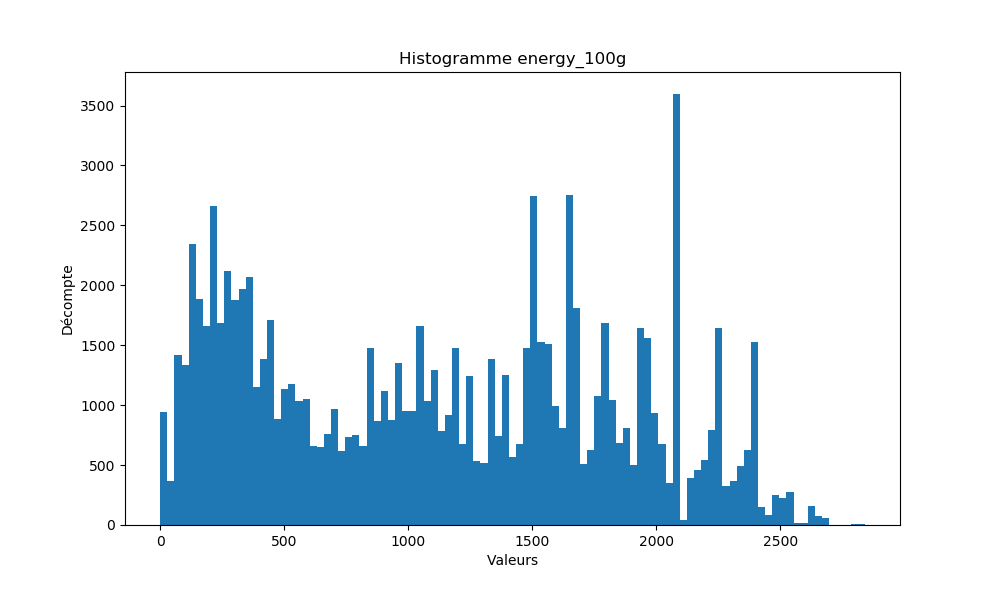


Figure – Histogramme energy\_100g

#### proteins\_100g

La distribution est plus forte sur les faibles valeurs, avec un pic pour la valeur 0.

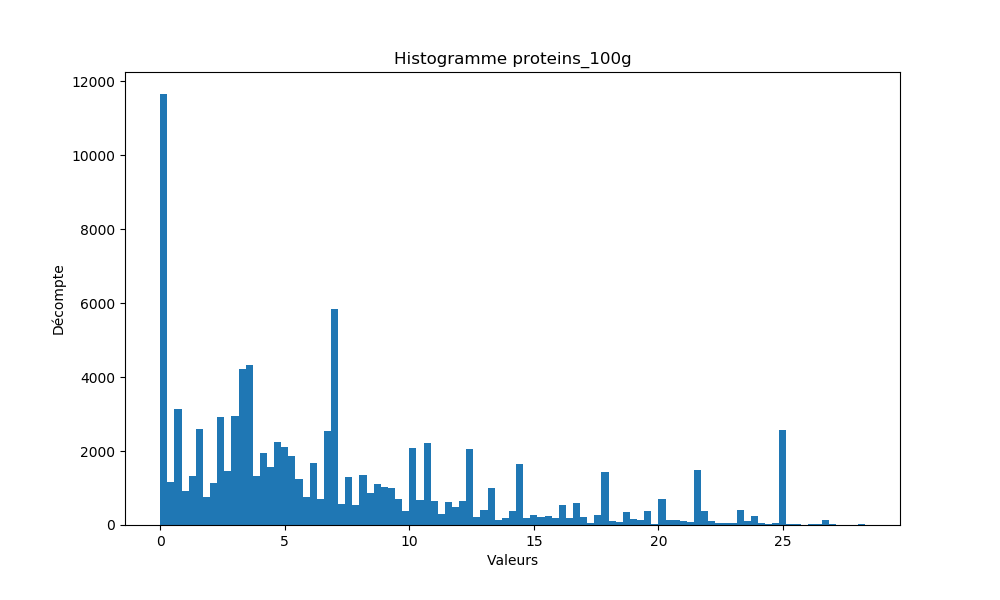


Figure – Histogramme proteins\_100g

#### salt\_100g

La distribution se fait sur les valeurs basses, avec un pic pour la valeur 0. Elle est assez uniforme jusqu’à une valeur de 1,2 g de sel par 100g.

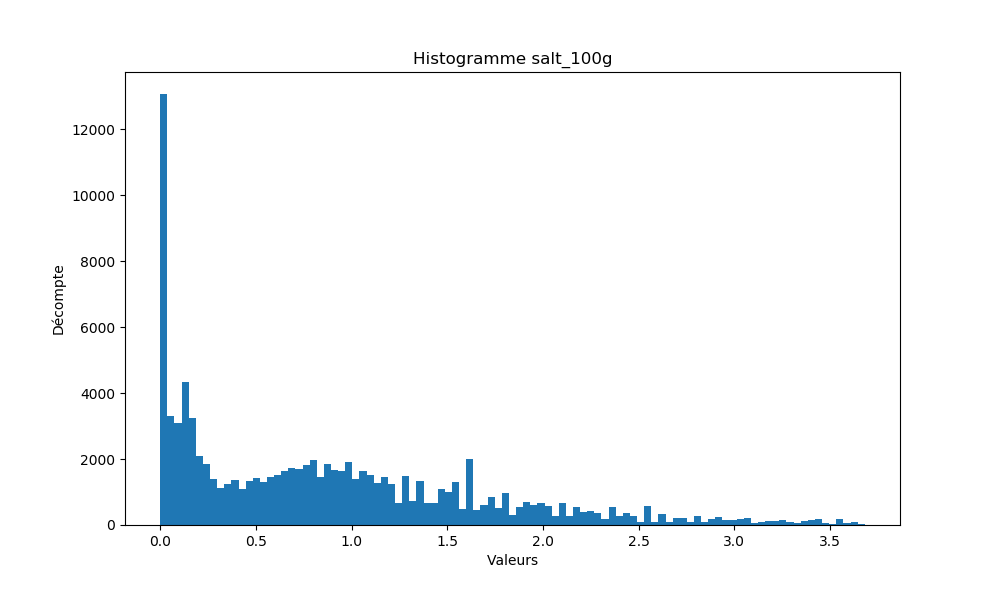


Figure – Histogramme salt\_100g

#### sodium\_100g

La distribution se fait sur les valeurs basses, avec un pic pour la valeur 0. Elle est assez uniforme jusqu’à une valeur de 0,5 g de sodium par 100g.

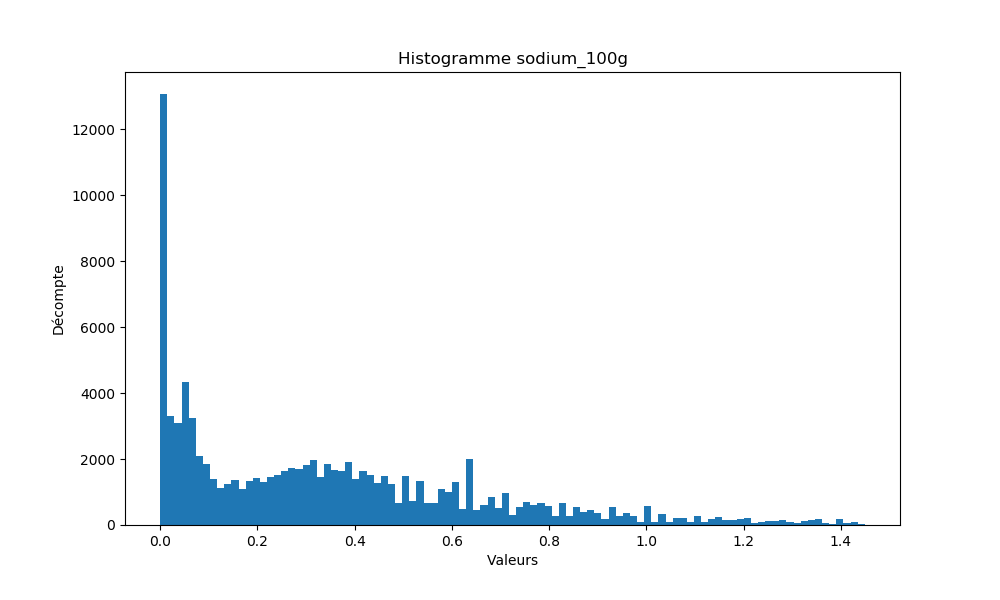


Figure – Histogramme sodium\_100g

#### sugars\_100g

Les valeurs basses prédominent, avec un grand pic à 0, et un pic plus petit pour la valeur de 4 environ.

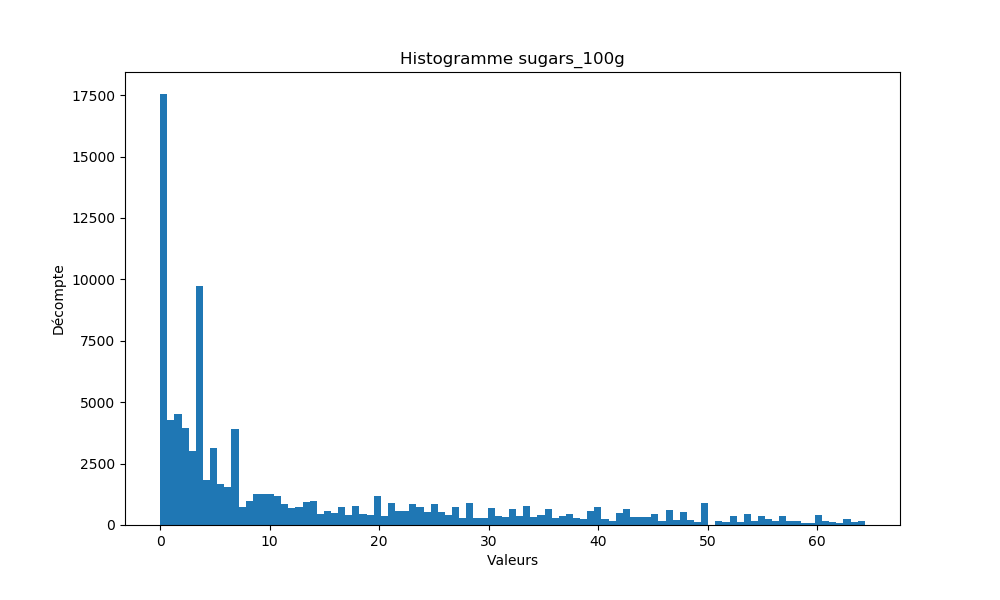


Figure – Histogramme sugars\_100g

#### fat\_100g

La distribution se fait sur les valeurs basses, avec un pic pour la valeur 0.

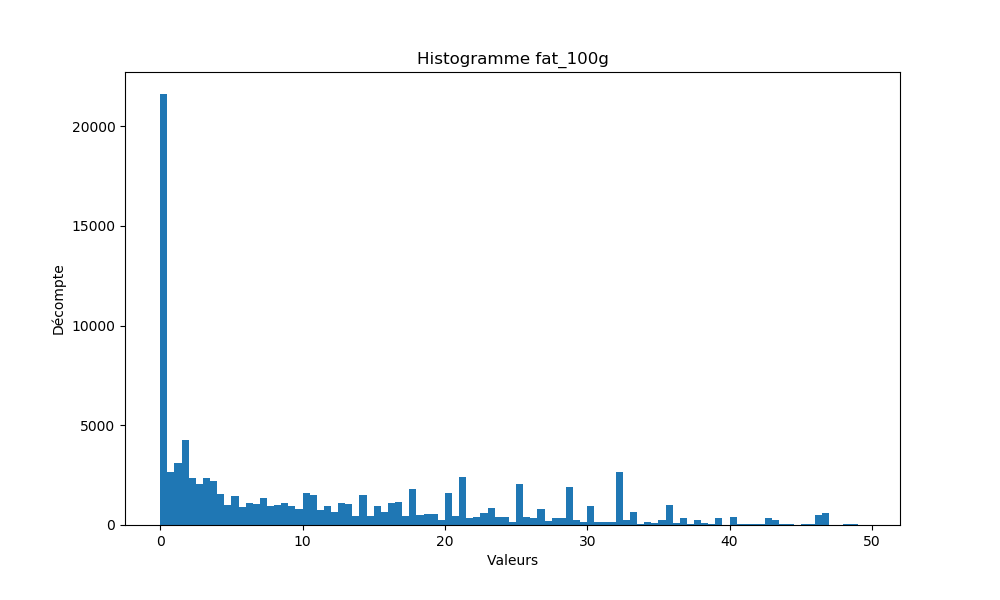


Figure – Histogramme fat\_100g

#### carbohydrates\_100g

Le pic n’est pas à la valeur 0 pour cette valeur. On constate deux parties distinctes séparés à la valeur de 40 (environ).

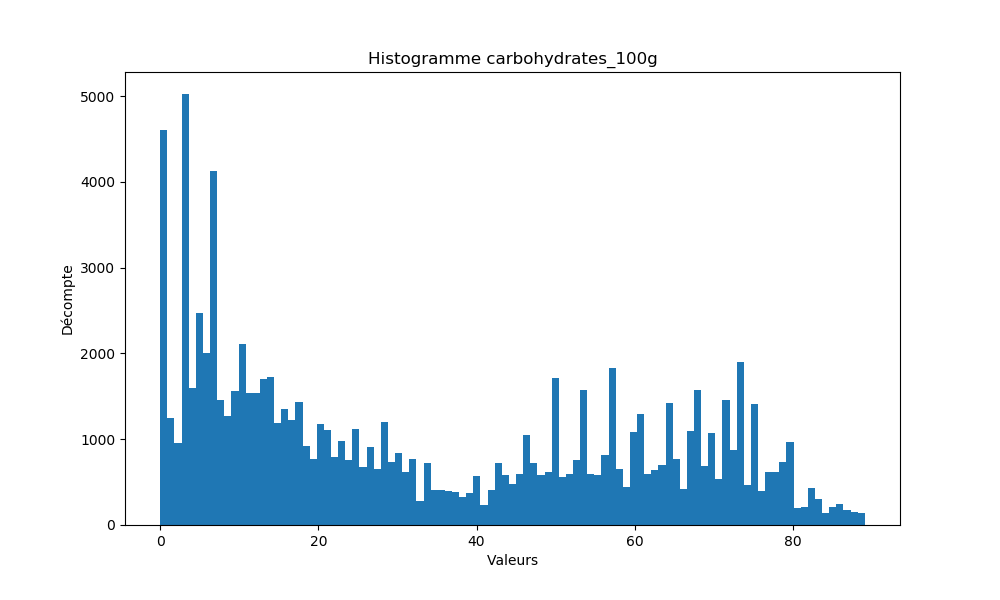


Figure – Histogramme carbohydrates\_100g

#### saturated-fat\_100g

Très peu d’aliments ont des graisses saturées d’après cette histogramme.

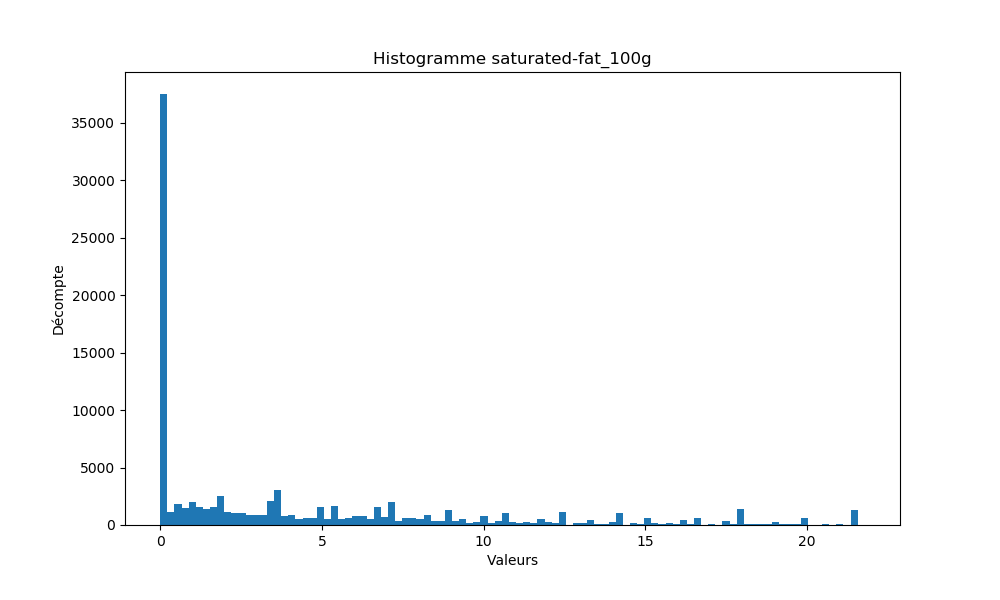


Figure – Histogramme saturated-fat\_100g

#### nutrition-score-fr\_100g

On note une prépondérance pour les valeurs basses (jusqu’à 2 environ). C’est une distribution gaussienne multimodale.

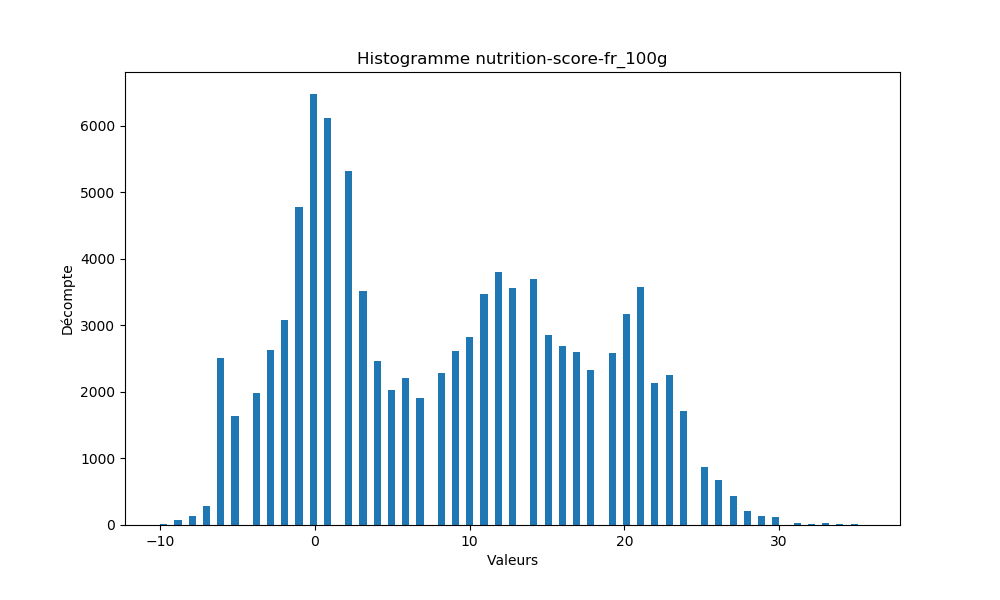


Figure – Histogramme nutrition-score-fr\_100g

#### fiber\_100g

La valeur 0 prédomine largement.

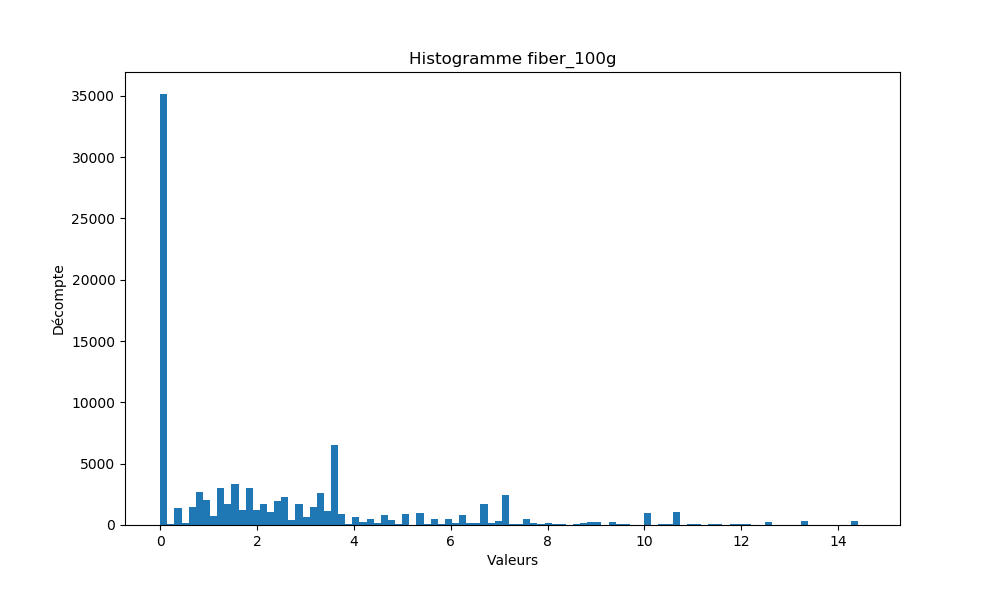


Figure – Histogramme fiber\_100g

#### cholesterol\_100g

La valeur 0 prédomine largement.

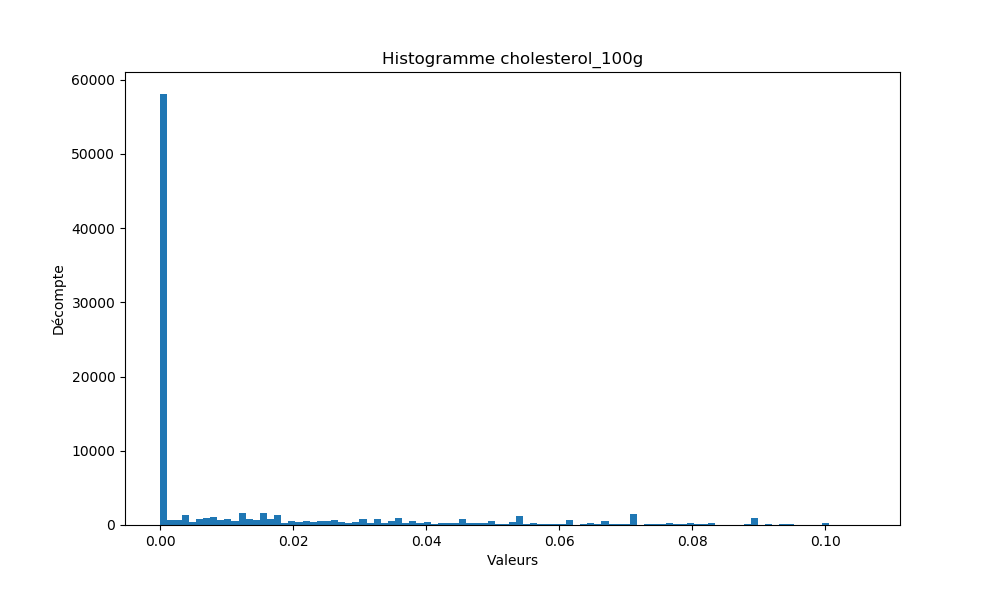


Figure – Histogramme cholesterol\_100g

#### trans-fat\_100g

La valeur 0 prédomine largement. Cette donnée ne semble pas vraiment pertinente.

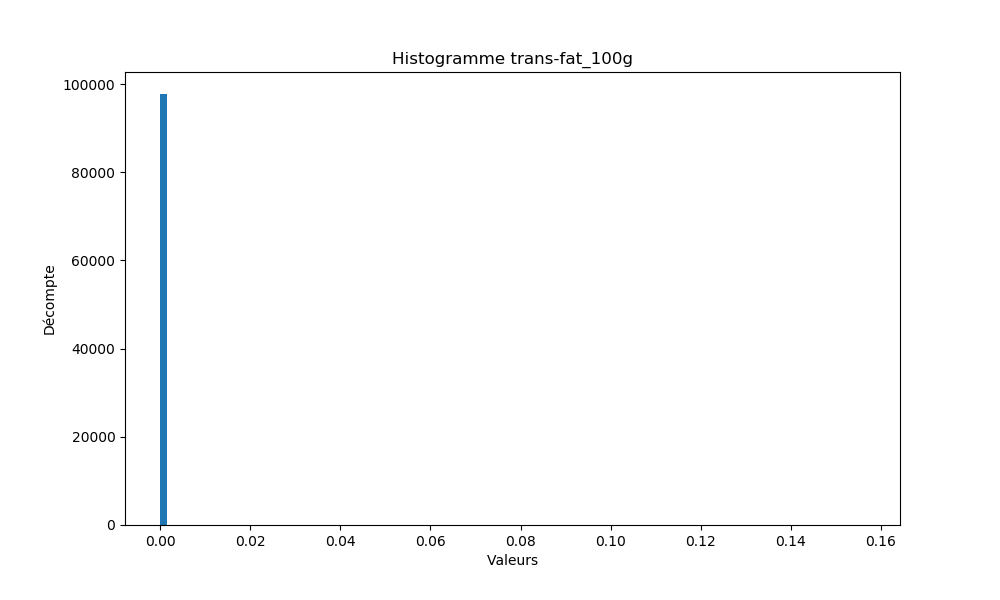


Figure – Histogramme trans-fat\_100g

#### calcium\_100g

La valeur 0 prédomine largement.

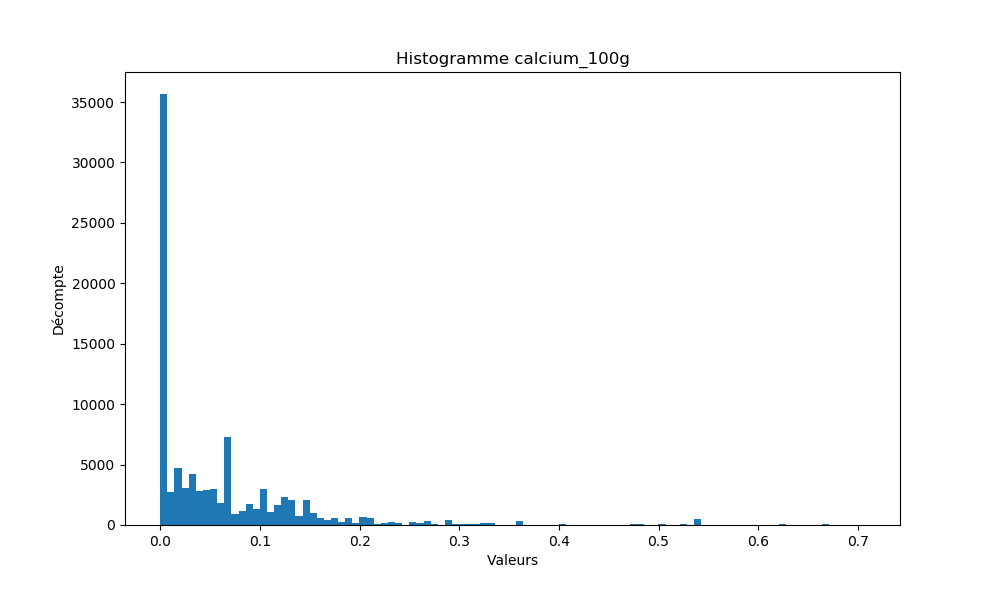


Figure – Histogramme calcium\_100g

#### vitamin-c\_100g

La valeur 0 prédomine largement.

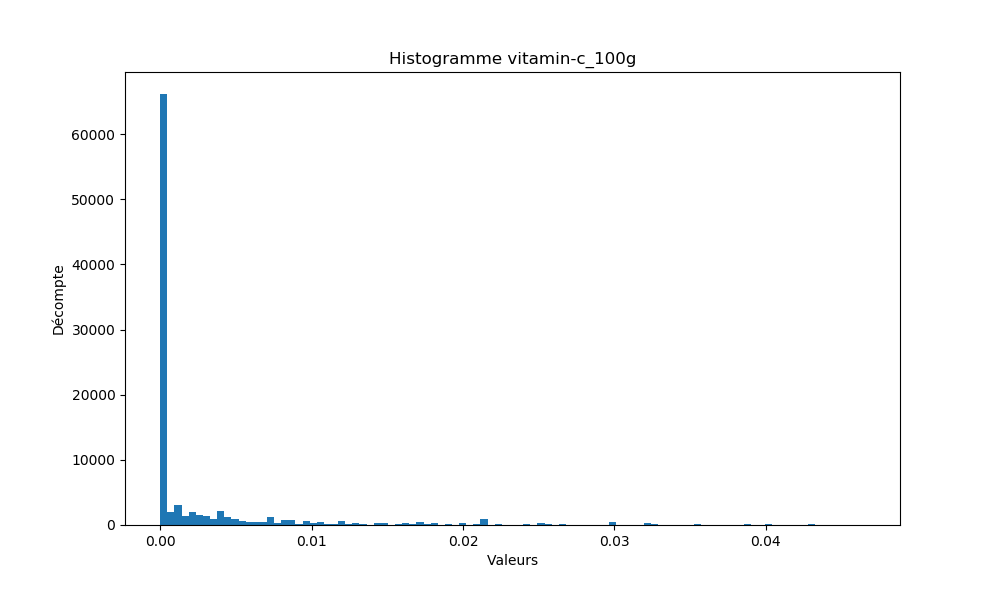


Figure – Histogramme vitamin-c\_100g

#### iron\_100g

La valeur 0 prédomine largement.

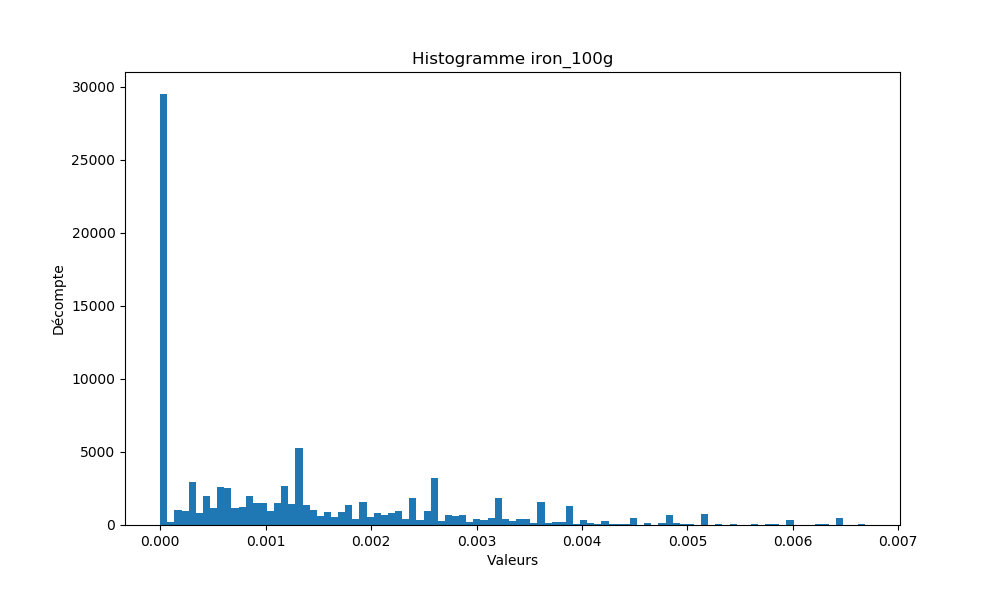


Figure – Histogramme iron\_100g

#### vitamin-a\_100g

La valeur 0 prédomine largement.

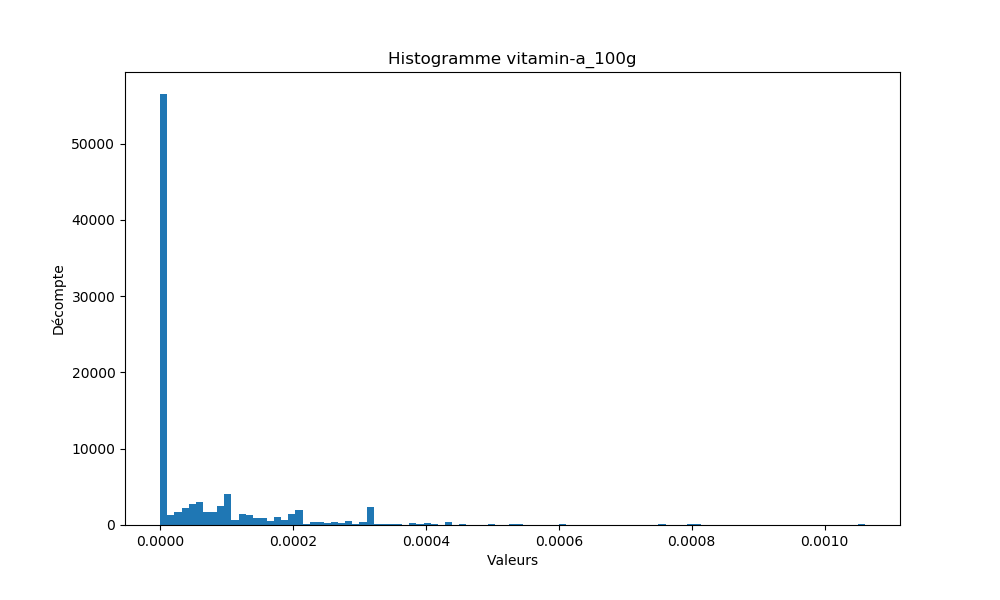


Figure – Histogramme vitamin-a\_100g

## Visualisations de relations entre certaines variables

Pour pouvoir observer des relations simples en certaines variables, la librairie *matplotlib.pyplot* de Python a été utilisée. Elle permet de créer simplement ces graphiques.

Afin d’avoir une base de comparaison saine, la référence qui a été choisie est la colonne « nutrition-score-fr\_100g ». La question s’est posée entre cette colonne ou la colonne « nutrition-score-uk\_100g ». Etant donné que le client est en France, il a été considéré plus logique de se prendre le référentiel français comme base. Cette donnée sera donc commune à tous nos graphiques.

## Quelques graphiques et conclusions associées

On retrouve en abscisse la caractéristique du nutriment, et en ordonnée sa note nutritionnelle. La graphique « Avant traitement » n’étant jamais interprétable, il ne sera pas commenté. Cependant il est montré afin de prouver l’efficacité discutée dans le § Traitement du jeu de données.

### Energie

#### Avant traitement

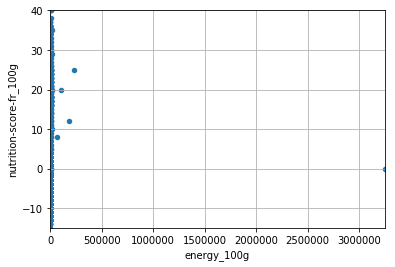


Figure - Energie avant traitement

#### Après traitement

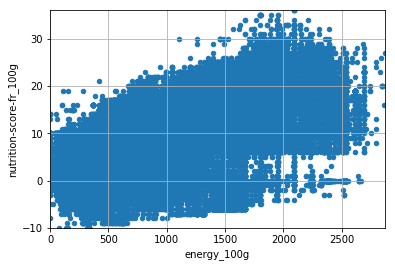


Figure - Energie après traitement

#### Conclusion

On voit une tendance négative[[2]](#footnote-2) dans ce graphique.

En effet, plus le nutriment est énergétique, plus le nombre de points dans son score nutritionnel total est **élevé**.

### Sel

#### Avant traitement

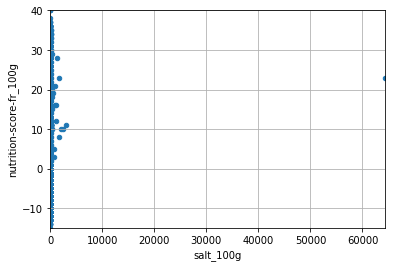


Figure - Sel avant traitement

#### Après traitement

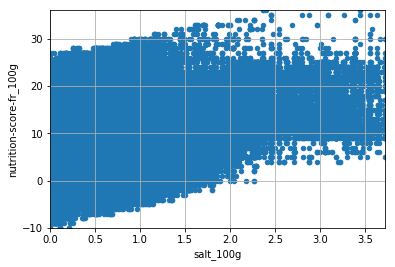


Figure - Sel après traitement

#### Conclusion

On voit une tendance négative dans ce graphique.

En effet, plus le nutriment contient de sel, plus le nombre de points dans son score nutritionnel total est **élevé**.

### Sodium

#### Avant traitement

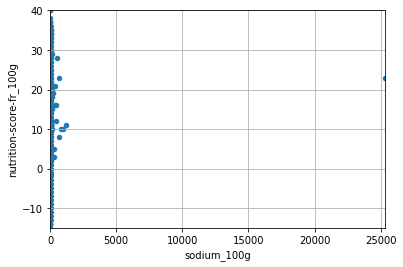


Figure - Sodium avant traitement

#### Après traitement

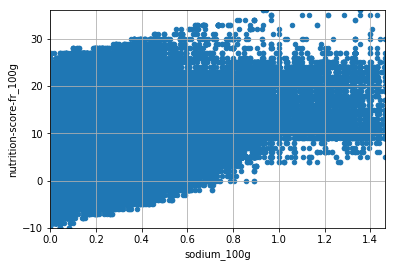


Figure - Sodium après traitement

#### Conclusion

On voit une tendance négative dans ce graphique.

En effet, plus le nutriment contient du sodium, plus le nombre de points dans son score nutritionnel total est **élevé**.

### Fibres

#### Avant traitement

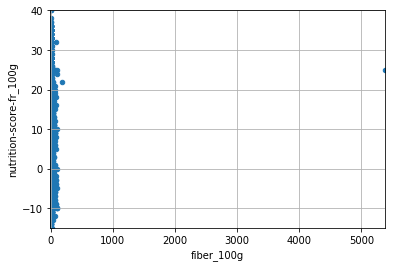


Figure - Fibres avant traitement

#### Après traitement

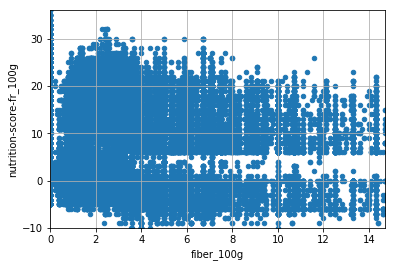


Figure - Fibres après traitement

#### Conclusion

On voit une tendance positive dans ce graphique.

En effet, plus le nutriment contient de fibre, plus le nombre de points dans son score nutritionnel total est **bas**.

### Vitamine C

#### Avant traitement

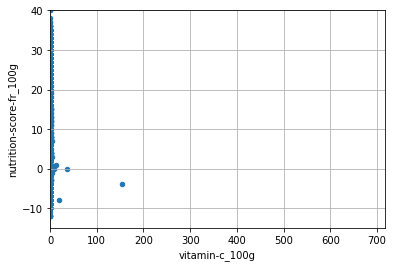


Figure - Vitamine C avant traitement

#### Après traitement

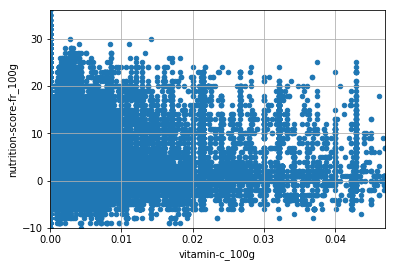


Figure - Vitamine C après traitement

#### Conclusion

On voit une tendance positive dans ce graphique.

En effet, plus le nutriment contient de vitamine C, plus le nombre de points dans son score nutritionnel total est **bas**.

### Sucres

#### Avant traitement

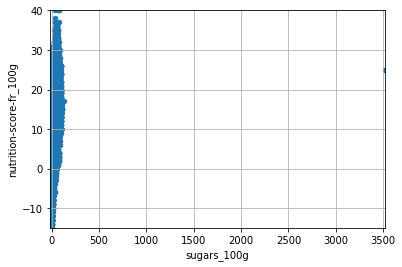


Figure - Sucres avant traitement

#### Après traitement

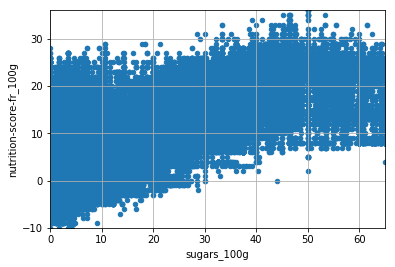


Figure - Sucres après traitement

#### Conclusion

On voit une tendance négative dans ce graphique.

En effet, plus le nutriment est sucré, plus le nombre de points dans son score nutritionnel total est **élevé**.

### Synthèse

Toutes les valeurs n’ont pas été représentés graphiquement dans ce rapport, mais voici une synthèse des conclusions observables :

| Caractéristique | Tendance |
| --- | --- |
| 'energy\_100g' | Négative |
| 'fat\_100g' | Négative |
| 'saturated-fat\_100g' | Négative |
| 'trans-fat\_100g' | Négative |
| 'cholesterol\_100g' | Négative |
| 'carbohydrates\_100g' | Négative |
| 'sugars\_100g' | Négative |
| 'fiber\_100g' | Positive |
| 'proteins\_100g' | Légèrement positive |
| 'salt\_100g' | Négative |
| 'sodium\_100g' | Négative |
| 'vitamin-a\_100g' | Positive |
| 'vitamin-c\_100g' | Positive |
| 'calcium\_100g' | Indéterminée. La courbe ne permet pas de tirer une conclusion |
| 'iron\_100g' | Indéterminée. La courbe ne permet pas de tirer une conclusion |

Figure - Tableau récapitulatif de conclusion

# Analyse multivariée

L’analyse multivariée effectuée ici a été basée sur deux méthodes principales :

* Matrice des corrélations.
* Régression linéaire entre deux variables.

## Matrice des corrélations

Une matrice de corrélation est utilisée pour évaluer la dépendance entre plusieurs variables en même temps. Le résultat est une table contenant les coefficients de corrélation entre chaque variable et les autres.

La matrice de corrélation peut être visualisée en utilisant un corrélogramme.

Un corrélogramme est une représentation graphique mettant en évidence une ou plusieurs corrélations entre des séries de données.

### Analyse

Pour effectuer cette analyse, toutes les colonnes sont utilisées. Le but est de voir si le lien entre elles est fort (valeur proche de 1), faible (valeur proche de 0) ou inverse (valeur négative).

### Corrélogramme

Le corrélogramme résultant de l’analyse du § 5.1.1 est montré ci-dessous.

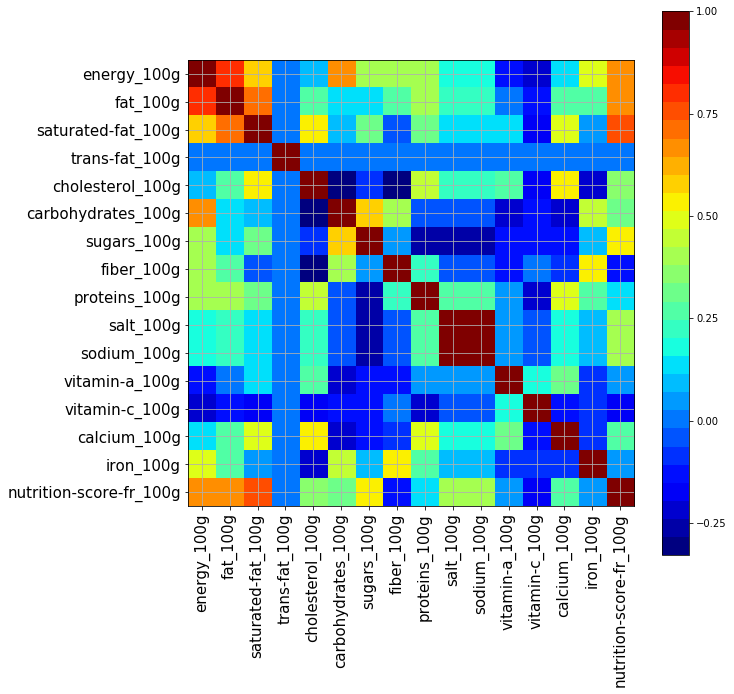


Figure - Corrélogramme

### Conclusion

Dans le corrélogramme, il ne faut pas prendre en compte la diagonale (supérieure gauche-inférieure droite) car elle « montre l’impact d’une variable sur elle-même ». Sa valeur sera toujours de 1.

Les conclusions essentielles à tirer de cette figure sont les suivantes :

* Sont très liées :
* Les quantités d’énergie et de graisse.
* Les quantités de graisse et de graisses saturées.
* Les quantités de carbohydrates et d’énergie.
* Vont faire progresser positivement le score nutritionnel :
* La quantité de vitamine A.
* La quantité de vitamine C.
* La quantité de fibres
* La quantité de protéines.
* Vont faire progresser négativement le score nutritionnel :
* La quantité d’énergie
* La quantité de graisse.
* La quantité de sucre.
* La quantité de graisse saturée.

## Régression linéaire et coefficient de corrélation

### Définitions

Un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.

Le type le plus simple de liaison est la relation affine. Dans le cas de deux variables numériques, elle se calcule à travers une régression linéaire, c’est ce que nous allons étudier ici et avons étudié (de manière qualitative dans le § 5.1).

La mesure de la corrélation linéaire entre les deux se fait alors par le calcul du coefficient de corrélation linéaire. Ce coefficient est égal au rapport de leur covariance et du produit non nul de leurs écarts types. Le coefficient de corrélation est compris entre -1 et 1.

Par contre, le fait que deux variables soient « fortement corrélées » ne démontre pas qu'il y ait une relation de causalité entre l'une et l'autre. Le contre-exemple le plus typique est celui où elles sont en fait liées par une causalité commune. Il faudra donc y être attentif.

Il est égal à 1 dans le cas où l'une des variables est une fonction affine croissante de l'autre variable, à -1 dans le cas où une variable est une fonction affine et décroissante. Les valeurs intermédiaires renseignent sur le degré de dépendance linéaire entre les deux variables. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation linéaire entre les variables est forte ; on emploie simplement l'expression « fortement corrélées » pour qualifier les deux variables. Une corrélation égale à 0 signifie que les variables ne sont pas corrélées linéairement.

Le coefficient de corrélation n’est pas sensible aux unités de chacune des variables. Ainsi, par exemple, le coefficient de corrélation linéaire entre l’âge et le poids d’un individu sera identique que l’âge soit mesuré en semaines, en mois ou en années.

L’échelle suivante est adoptée pour les futures conclusions :

|  |  |  |
| --- | --- | --- |
| Corrélation | Négative | Positive |
| Faible | de -0,5 à 0 | de 0 à 0,5 |
| Forte | de −1,0 à −0,5 | de 0,5 à 1,0 |

Figure 40 – Tableau des facteurs de corrélation

### Calculs de certains coefficients

Certains calculs entre deux données ont été effectués. Seules les valeurs supérieures à 0.5 (forte corrélation positive) sont présentées :

*Régression sur les deux données : nutrition-score-fr\_100g et energy\_100g*

*Score : 0.65*

*Régression sur les deux données : nutrition-score-fr\_100g et fat\_100g*

*Score : 0.66*

*Régression sur les deux données : nutrition-score-fr\_100g et saturated-fat\_100g*

*Score : 0.77*

*Régression sur les deux données : nutrition-score-fr\_100g et sugars\_100g*

*Score : 0.55*

*Régression sur les deux données : energy\_100g et fat\_100g*

*Score : 0.79*

*Régression sur les deux données : fat\_100g et saturated-fat\_100g*

*Score : 0.72*

*Régression sur les deux données : proteins\_100g et calcium\_100g*

*Score : 0.51*

Ces scores viennent confirmer, quantitativement cette fois-ci, les conclusions du § 5.1.3.

# Feature engineering

Le processus de « feature engineering » tente de créer des variables supplémentaires pertinentes à partir de données brutes existantes dans la base de données et peut permettre d’augmenter la performance de prédiction d’un futur algorithme d’apprentissage.

Dans ce chapitre, les différents choix de « feature engineering » sont listés et le renvoi de leur utilisation dans les études décrites est présenté.

## Définition des intervalles considérés corrects et non aberrants.

### Source

3.1 - Travail sur la base de données.

4.3 - Quelques graphiques et conclusions associées

### Conclusion

#### Valeurs supprimées

* Valeurs négatives supprimées pour les valeurs nutritionnelles des nutriments.
* Valeurs supérieures au 98ème quantile supprimées.

#### Données conservées

* energy\_100g
* fat\_100g
* saturated-fat\_100g
* trans-fat\_100g
* cholesterol\_100g
* carbohydrates\_100g
* sugars\_100g
* fiber\_100g
* proteins\_100g
* salt\_100g
* sodium\_100g
* vitamin-a\_100g
* vitamin-c\_100g
* calcium\_100g
* iron\_100g
* nutrition\_score\_fr\_100g

## Détails des variables proposées et crées

Des variables ont été créées à partir des données existantes afin de faciliter le travail de lecture immédiat de ces bases de données assez conséquentes. Dans le détail, nous allons nous intéresser aux deux indicateurs suivants :

* La recherche d’une correspondance avec le nutri score préexistant
* Boolean qui détermine si un aliment est sain ou ne l’est pas

### Correspondance avec le nutri score préexistant

Dans la partie ci-dessous, nous essayons de "valider" les similitudes entre l'échelle créé ici et l'échelle du nutriscore préexistant.

Nous recherchons une valeur "i" qui va maximiser le taux de similitude, sans toutefois chercher à l'atteindre.

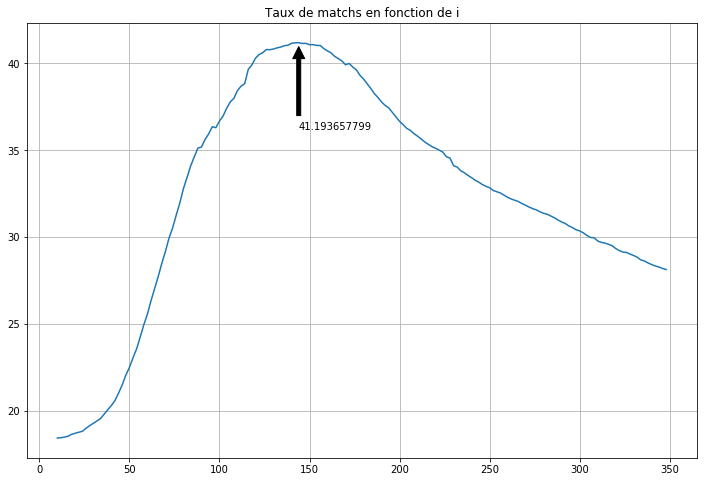


Figure – Taux de similitude en fonction des paliers

Il existe une valeur qui va modifier les valeurs des paliers qui ont été créés, et qui peut amener à un taux de similitude des échelles de 41,2 %. Attention néanmoins, le but ici est uniquement de donner un aperçu ce qui pourrait être fait dans un futur projet. Tel quel, ce résultat n’a pas grande valeur.

### Détermination de la sanité d’un aliment.

Nous avons également créé quelques diagrammes type camembert pour voir la réparation entre les bons et les mauvais aliments, suivant les échelles et également le nombre de tranche d’aliment que l’on désigne comme sain ou non.

#### Avec 1 tranche

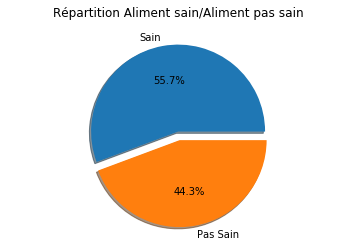


Figure – Répartition des aliments avec 1 tranche

Avec uniquement les aliments « a », on arrive à 56 %

#### Avec 2 tranches

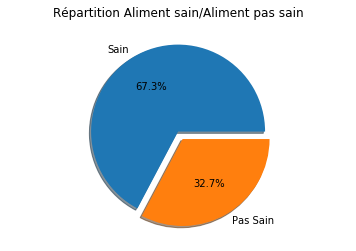


Figure – Répartition des aliments avec 2 tranches

Si on considère que les aliments « a » et « b » sont sains, on obtient les 2/3 des aliments sains.

#### Avec 3 tranches

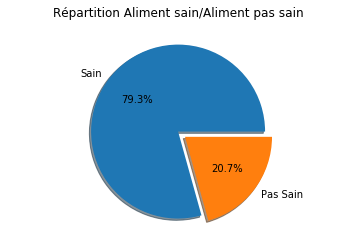


Figure – Répartition des aliments avec 3 tranches

Enfin, avec les aliments « a », « b » et « c » on arrive à 79,3 %

#### Conclusion

On peut facilement augmenter ou réduire la sévérité de l’échelle en jouant sur « l’équation » qui détermine les éléments positifs ou négatifs d’un aliment.

Les résultats montrés ici sont un seul et unique exemple, mais il serait très simple d’extrapoler avec d’autres paramètres d’entrées. Par exemple, une personne ayant un régime sans sel pourrait augmenter la négativité du sel (en multipliant sa valeur par 2 ou 3) et verrait alors tous les aliments contenant beaucoup de sel pencher vers le côté négatif du nutriscore.

# Conclusion

Pour conclure cette étude très intéressante, nous pouvons la résumer avec les points suivants :

* Comprendre les bases de la diététique est le premier point essentiel à aborder. Les nutriments contenus dans les aliments sont complexes et nombreux. Il faut donc être capable de cibler les éléments qui sont intéressantes pour les personnes concernées (Cf. §Principes de base de la diététique).
* Les bases de données sont souvent énormes, mal faites et incomplètes. Un travail important est nécessaire afin de pouvoir l’exploiter. Un avantage du fait de sa taille est que même après un nettoyage important, on conserve une grande quantité de données disponibles et utilisables (Cf. § Traitement du jeu de données).
* Trois analyses ont permis de tirer quelques conclusions :
  + Une première analyse nous permet de déterminer quels nutriments ont une influence positive et quels sont ceux qui ont une influence négative dans le calcul du score nutritionnel (Cf. § Visualisations de relations entre certaines variables).
  + Une deuxième analyse permet de comprendre le lien entre les nutriments 2 à 2 dans le calcul du score. Ainsi, il a été possible de voir des liens directs, de manière qualitative, entre « mauvais nutriment » et « mauvais aliment » (Cf. §. Matrice des corrélations).
  + Une troisième analyse permet de confirmer, de manière quantitative, les réponses obtenues dans la deuxième analyse.
* Une partie de feature engineering (Cf. § Détails des variables proposées et crées) permet d’aller un peu plus loin et de se projeter sur ce qu’on pourra faire avec les premières conclusions tirées des trois analyses décrites ci-dessus. Grace à elles, nous pouvons proposer des traitements potentiels des données qui pourront aider le client à développer son générateur de recettes. Parmi elles, on peut citer :
  + Déterminer la sévérité de l’échelle.
  + Déterminer la sévérité d’un nutriment par rapport à un autre.

1. En statistiques et en théorie des probabilités, les quantiles sont les valeurs qui divisent un jeu de données en intervalles contenant le même nombre de données. Il y a donc un quantile de moins que le nombre de groupes créés. Ainsi les quartiles sont les trois quantiles qui divisent un ensemble de données en quatre groupes de taille égale. [↑](#footnote-ref-1)
2. A partir de ce chapitre, et jusqu’à la fin du document, on parlera de tendance négative pour les aliments faisant augmenter le score nutritionnel. On rappelle que plus le score nutritionnel est élevé, plus l’aliment est mauvais pour la santé. Dans la même logique, une tendance positive va faire baisser le score nutritionnel. [↑](#footnote-ref-2)