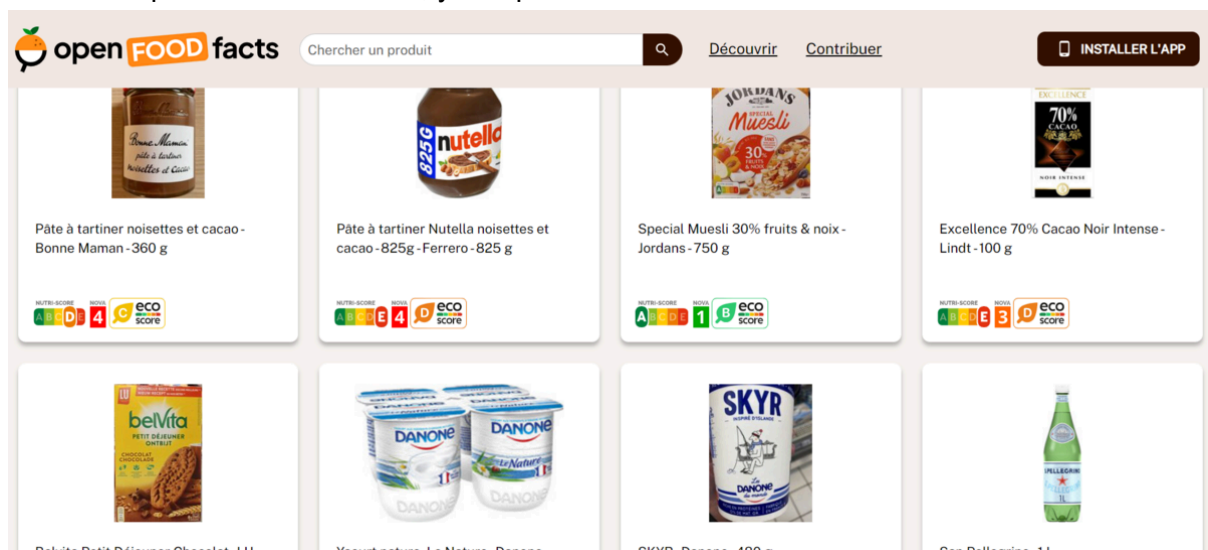


## Présentation d'Open Food Facts

Open Food Facts est une base de données collaborative, ouverte et mondiale, qui répertorie les caractéristiques nutritionnelles, les ingrédients, les additifs et diverses informations relatives à des milliers de produits alimentaires. Alimentée par une communauté internationale de bénévoles, elle permet aux consommateurs, aux chercheurs et aux développeurs d'accéder librement aux données pour mieux comprendre la composition des aliments, comparer leur qualité nutritionnelle et favoriser des choix alimentaires plus sains.

Open Food Facts possède un site web ainsi qu'une application mobile où ils répertorient des millions de produits alimentaires, y compris leur Nutri-Score








Depuis 2014, Open Food Facts calcule et affiche le Nutri-Score, un système d'étiquetage nutritionnel, avant même son adoption officielle par le gouvernement. C'est depuis 2017 que l'initiative est soutenue par Santé Publique France, et a été officiellement introduite en France .

## Présentation du Nutri-Score

Le nutri-score fournit aux consommateurs une note qui représente la qualité nutritionnelle du produit. Le score est calculé par un système de point, plus le score est élevé plus l'aliment est considéré comme mauvais.

Le résultat du calcul du nutri-score donne une valeur comprise entre -15 et +40 représentée par une étiquette appelée Nutri-grade :

Étiquette	Lettre	Couleur	Score
	A	Vert	-15 à -2
	B	Vert clair	-1 à +3
	C	Jaune	+4 à +11
	D	Orange	+12 à +16
	E	Rouge	+17 à +40

### Elements favorable / défavorable au score

Pour 100g on va évaluer la teneur en éléments favorable au score et élément défavorable au score.

Éléments défavorables au score :

- Apport calorique pour 100g
- Teneur en sucre
- Teneur en graisses saturées
- Teneur en sel

Element favorables au score :

- Teneur en fruits, légumes, légumineuses (dont les légumes secs), oléagineux, huiles de colza, de noix et d'olive.
- Teneur en fibres
- Teneur en protéines

### Transition vers l'analyse des traits distinctifs d'un aliment sain

Dans le cadre de notre mission en tant que Développeur IA pour une ONG dédiée à l'amélioration de la santé mondiale, nous visons à exploiter les données d'Open Food Facts pour créer un outil d'analyse et de visualisation des données nutritionnelles.

Mais pour ça nous devons comprendre quelles étaient les caractéristiques d'un aliment sain en nous basant sur le dataset d'Open Food Facts.

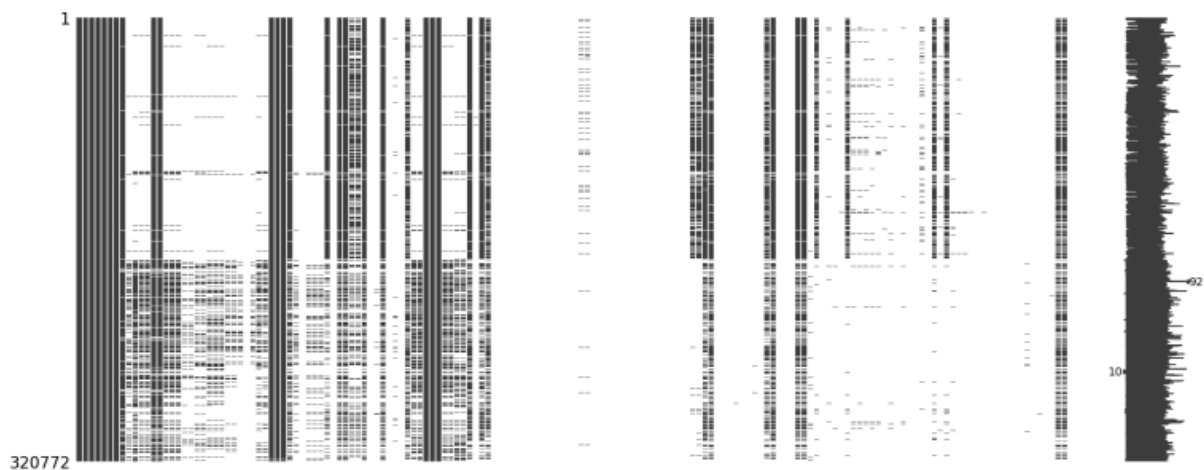
## Nettoyage des données

Le nettoyage et l'exploration des données sont des étapes fondamentales du processus d'analyse des données. Ils garantissent que les données sont exactes, cohérentes et prêtes à être analysées. La qualité des données impacte la qualité des résultats.

On dit souvent "Garbage In, Garbage Out". Cela signifie que si les données initiales sont erronées, incomplètes ou biaisées, toutes l'analyse basées sur ces données sera également erronée ou trompeuse. Par exemple :

- Une valeur manquante peut fausser des moyennes
- Une erreur dans les données (comme une valeur aberrante ou incohérente) peut déformer les résultats.

Voici une visualisation de notre dataset brut grâce à la librairie missingno :



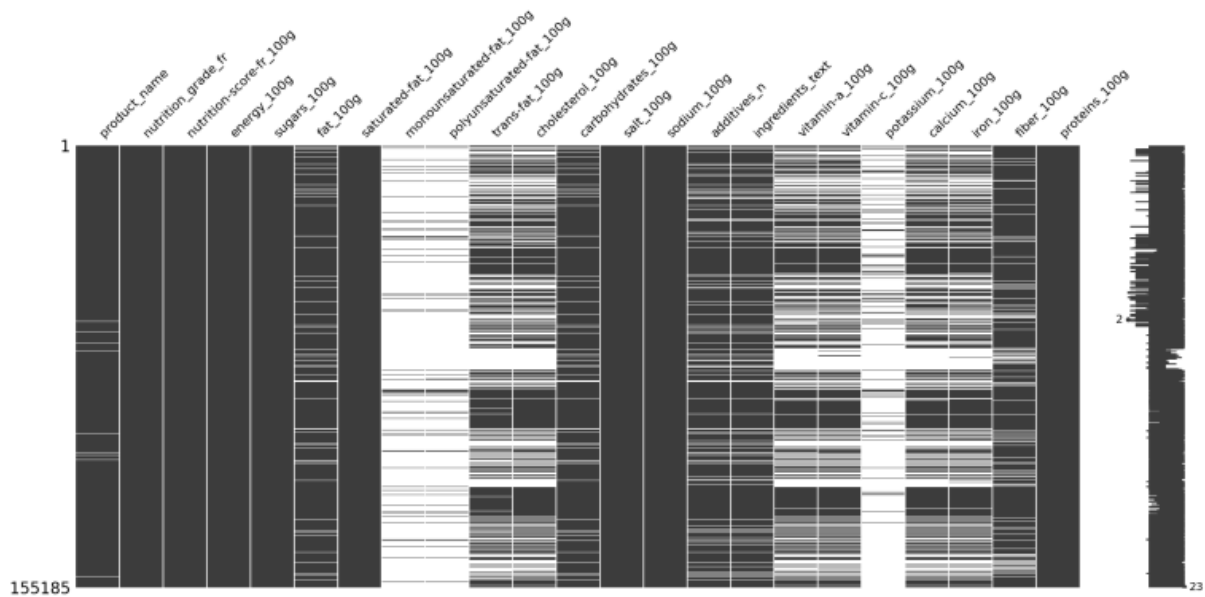
On peut voir qu'il y a beaucoup de colonnes avec des données manquantes. Elles n'apportent pas beaucoup d'informations pour l'analyse dans notre contexte. Et les garder rend le dataset plus complexe et difficile à traiter.

Pour notre analyse, nous avons supprimé les colonnes contenant trop de valeurs manquantes : celles qui sont entièrement remplies de NaN ou celles où plus de 95 % des valeurs sont manquantes.

Nous avons ensuite sélectionné les colonnes pertinentes en fonction des besoins spécifiques de chacune de nos analyses.

Nous obtenons ainsi un dataset épuré qui nous servira de base pour une analyse spécifique.

Voici un aperçu d'un de nos dataset nettoyé utilisé:



Pour chaque colonne on va ensuite essayer d'identifier les valeurs anormales:

	sugars_100g	fat_100g	saturated-fat_100g	monounsaturated-fat_100g	polyunsaturated-fat_100g	sodium_100g	fiber_100g	proteins_100g
count	221019.000000	203733.000000	221019.000000	20151.000000	20179.000000	221019.000000	193770.000000	221019.000000
mean	15.023803	13.367035	4.970118	7.235400	4.988618	0.645805	2.828862	7.775905
std	21.181012	16.208529	7.659902	11.930241	8.325063	54.100789	13.016507	8.123994
min	-17.860000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-3.570000
25%	1.300000	0.880000	0.000000	0.000000	0.000000	0.039370	0.000000	1.900000
50%	5.000000	7.140000	1.790000	3.330000	1.790000	0.255906	1.500000	5.700000
75%	23.080000	21.430000	7.140000	8.930000	6.060000	0.536000	3.600000	10.710000
max	3520.000000	714.290000	550.000000	557.140000	75.000000	25320.000000	5380.000000	430.000000

Les sucres ne peuvent pas être négatifs, donc cette valeur est clairement anormale. Une teneur en sucres de 3520 g/100 g semble irréaliste.

On a donc utilisé la méthode des bornes de l'écart interquartile. Elle est utilisée pour détecter et supprimer les valeurs aberrantes d'un dataset.

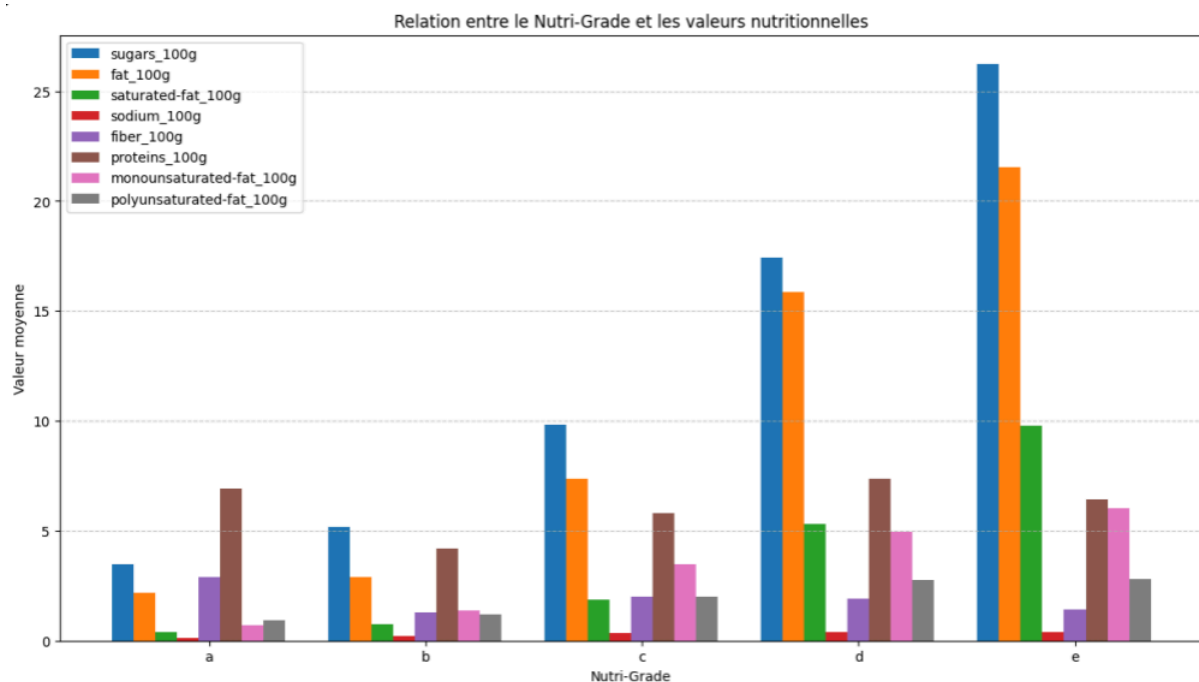
C'est une approche statistique simple et robuste, à laquelle nous avons ajouté une condition pour exclure les colonnes contenant des valeurs négatives.

Voici à quoi ressemble nos colonnes traitées :

	sugars_100g	fat_100g	saturated-fat_100g	monounsaturated-fat_100g	polyunsaturated-fat_100g	sodium_100g	fiber_100g	proteins_100g
count	154994.000000	142809.000000	154994.000000	12509.000000	12528.000000	154994.000000	136616.000000	154994.000000
mean	11.46179	9.195982	3.153834	3.209556	1.910874	0.292237	1.953648	6.142408
std	13.60918	10.654376	4.226416	4.012318	2.220527	0.275494	2.121229	5.392512
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.600000	0.500000	0.000000	0.000000	0.000000	0.040000	0.000000	1.670000
50%	4.900000	4.620000	1.160000	1.590000	1.160000	0.241000	1.400000	4.900000
75%	17.750000	15.220000	5.000000	5.290000	3.100000	0.465000	3.200000	9.090000
max	52.630000	52.220000	17.100000	17.800000	9.300000	1.121000	9.000000	22.220000

## Analyse du jeu de données

Sur ce graphique se basant sur nos colonnes pertinentes nettoyées, on observe des tendances claires qui semblent logiques et cohérentes avec ce que l'on peut attendre du Nutri-grade. Les produits considérés comme sains sont en moyenne plus riches en fibre et protéine, tout en affichant des valeurs faibles en sucre, graisses saturées et sodium. On peut remarquer que plus la quantité moyenne de sodium, gras et sucre augmente plus la note se dégrade et l'aliment est considéré comme mauvais.



Après avoir visualisé les moyennes des valeurs nutritionnelles par Nutri-Grade dans le graphique précédent, nous avons calculé les corrélations de Pearson pour quantifier la relation entre le Nutri-Score numérique et chaque variable nutritionnelle.

### Corrélations avec le Nutri-Score :

nutrition-score-fr_100g	1.000000
saturated-fat_100g	0.740575
sugars_100g	0.705451
fat_100g	0.672154
monounsaturated-fat_100g	0.484215
sodium_100g	0.401889
polyunsaturated-fat_100g	0.331790
fiber_100g	-0.131795
proteins_100g	-0.254014

Les résultats des corrélations de Pearson montrent des tendances claires et cohérentes. Les éléments défavorables, tels que la teneur en sucres (0.71), en graisses totales (0.67) et surtout en graisses saturées (0.74), présentent une forte corrélation positive avec le Nutri-Score. Cela signifie que plus ces valeurs augmentent, plus le Nutri-Score est mauvais (produit moins sain).

À l'inverse, les éléments favorables comme la teneur en fibres (-0.13) et en protéines (-0.25) montrent des corrélations négatives modérées avec le Nutri-Score. Cela indique que plus ces valeurs augmentent, plus le Nutri-Score s'améliore (produit plus sain) bien que leurs corrélations négatives soient plus faibles.

Pour approfondir notre analyse, on a identifié les différentes catégories d'aliments dans le dataset.

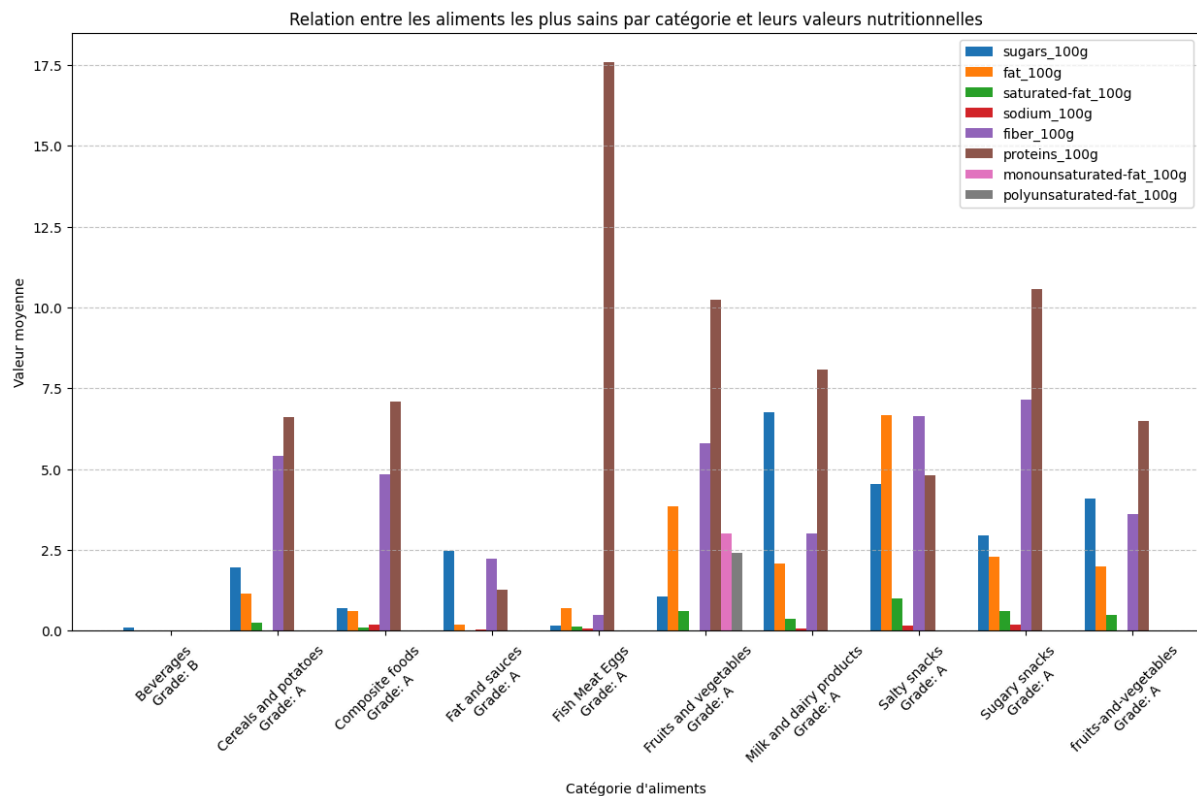
- Cereals and potatoes (Céréales et pommes de terre)
- Sugary snacks (Collations sucrées / Snacks sucrés)
- Composite foods (Aliments composés / Plats composés)
- Beverages (Boissons)
- Milk and dairy products (Lait et produits laitiers)
- Fruits and vegetables (Fruits et légumes)
- Fish Meat Eggs (Poisson, viande et œufs)
- Fat and sauces (Matières grasses et sauces)
- Salty snacks (Collations salées / Snacks salés)

Pour chaque catégorie d'aliments nous avons sélectionné deux sous-groupes :

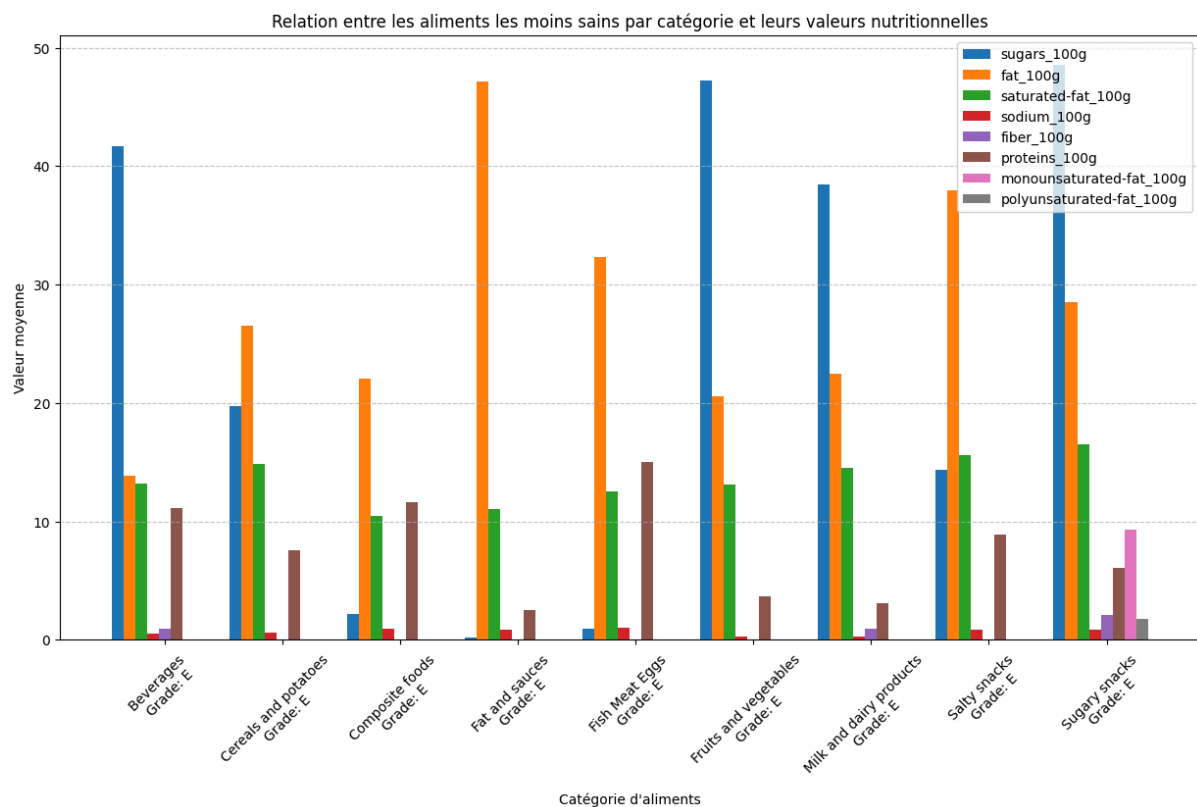
- Les aliments avec le Nutri-Score le plus élevé (catégorisés comme les plus sains).
- Les aliments avec le Nutri-Score le plus bas (catégorisés comme les moins sains).

Ce graphique ci-dessous confirme que les aliments considérés comme sains se distinguent par :

Des niveaux élevés de fibres et protéines (éléments favorables). De faibles teneurs en sucres, graisses saturées et sodium (éléments défavorables).



Ce graphique ci-dessous montre clairement que les aliments les moins sains se caractérisent par des niveaux très élevés de facteurs défavorables : sucres, graisses saturées, et sodium. Et des niveaux faibles ou très faibles de facteurs favorables : fibres et protéines.



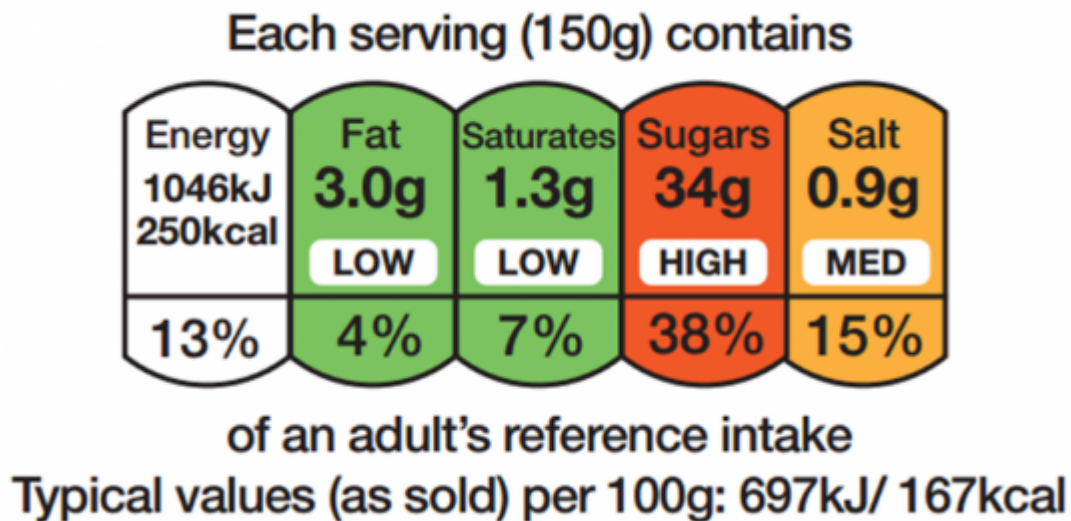
### Observations et hypothèses :

On a pu observer que les points défavorables pèsent plus lourd dans le calcul global. Cela signifie qu'un excès de sucres ou graisses domine et dégrade le score final. Les fibres et protéines améliorent le Nutri-Score, mais elles ne suffisent pas à contrebalancer une composition nutritionnelle médiocre.

Le système du nutri-score est bien pensé, il empêche les fabricants de tricher en ajoutant des ingrédients favorables pour "compenser" des aliments trop riches en sucres ou graisses saturées.

### Nutri-Score et informations nutritionnelles

Bien que le Nutri-Score soit largement adopté en France et dans d'autres pays européens pour simplifier l'interprétation de la qualité nutritionnelle, ce n'est pas le cas partout. Par exemple, le Royaume-Uni a choisi de conserver son propre système d'étiquetage nutritionnel, connu sous le nom de « Traffic Light System ».



Il ne s'agit pas d'un score unique, mais d'une étiquette indiquant séparément les teneurs en graisses, graisses saturées, sucres et sel, chacune associée à une couleur en fonction de seuils définis. Ce système segmenté fournit une vision détaillée nutriment par nutriment, plutôt qu'un score global comme le nutriscore.

En revenant au Nutri-Score, il est intéressant de souligner que la qualité nutritionnelle d'un produit, bien que simplifiée par ce système, n'est pas nécessairement liée à son prix. En effet, la corrélation entre le prix et le Nutri-Score n'est pas stricte, car de nombreux autres



facteurs entrent en jeu : stratégies marketing, saisonnalité, disponibilité, ou encore marque versus premier prix.

- Certains aliments ultra-transformés (snacks sucrés, sodas, confiseries) sont souvent peu chers, mais ont un Nutri-Score moins bon (C, D, E).
- Les fruits et légumes frais, de bonne qualité nutritionnelle (Nutri-Score A ou B en général), peuvent parfois être plus coûteux, surtout hors saison.
- Il est également possible de trouver des aliments peu coûteux et bien notés (comme des légumineuses sèches, des céréales complètes peu transformées, de l'eau, etc.).

Au-delà des considérations économiques, il est important de noter que la composition nutritionnelle des aliments varie selon le pays.

<b>U.S. Version</b>   <b>Ingredients:</b> Tomato Concentrate, Distilled Vinegar, High Fructose Corn Syrup, Corn Syrup, Salt, Spice, Onion Powder, Natural Flavoring.	<b>U.K. Version</b>   <b>Ingredients:</b> Tomatoes, Spirit Vinegar, Sugar, Salt, Spice and Herb Extracts, Spice.
<b>U.S. Version</b>   <b>Ingredients:</b> Carbonated Water, High Fructose Corn Syrup, Concentrated Orange Juice, Citric Acid, Natural Flavor, Sodium Benzoate, Caffeine, Sodium Citrate, Erythorbic Acid, Gum Arabic, Calcium Disodium EDTA, Brominated Vegetable Oil, Yellow 5.	<b>U.K. Version</b>   <b>Ingredients:</b> Carbonated Water, Sugar, Citric Acid, Ascorbic Acid, Caffeine, Flavourings, Potassium Sorbate, Gum Arabic, Colour (Beta Carotene).
<b>U.S. Version</b>   <b>Ingredients:</b> Whole Grain Rolled Oats, Sugar, Creaming Agent, Maltodextrin, Sunflower and Palm Oils, Whey, Sodium Caseinate, Flavored and Colored Fruit Pieces, Dehydrated Apples (Treated With Sodium Sulfite), Artificial Strawberry Flavor, Citric Acid, Red 40, Salt, Guar Gum, Artificial Flavor, Citric Acid, Nicotinamide, Vitamin A Palmitate, Reduced Iron, Pantothenic Hydrochloride, Riboflavin, Thiamin Mononitrate, Folic Acid.	<b>U.K. Version</b>   <b>Ingredients:</b> Quaker Wholegrain Rolled Oats, Sugar, Freeze Dried Raspberry Pieces, Freeze Dried Strawberry Pieces, Natural Flavouring.
<b>U.S. Version</b>   <b>Ingredients:</b> Corn, Vegetable Oil (Corn, Canola, And/or Sunflower Oil), Maltodextrin, Salt, Tomato Powder, Corn Starch, Lactose, Whey, Skim Milk, Corn Syrup Solids, Onion Powder, Sugar, Garlic Powder, Monosodium Glutamate, Cheddar Cheese (Milk, Cheese Cultures, Salt, Enzymes), Dextrose, Malic Acid, Buttermilk, Natural And Artificial Flavors, Sodium Acetate, Artificial Color (Red #40, Blue #1, Yellow #5), Sodium Caseinate, Spice, Citric Acid, Disodium Inosinate, And Disodium Guanylate.	<b>U.K. Version</b>   <b>Ingredients:</b> Corn, Vegetable Oils (Sunflower, Rapeseed), Cool Onion Flavour, Salt, Glucose Syrup, Sugar, Potassium Chloride, Cheese Powder, Flavour Enhancers (Monosodium Glutamate, Disodium 5'-Ribonucleotides), Acidity Regulators (Malic Acid, Sodium Acetate, Citric Acid), Colour (Annatto), Milk Proteins, Spice.
<b>U.S. Version</b>   <b>Ingredients:</b> Milk Chocolate [Sugar; Milk; Cocoa Butter; Chocolate; Milk Fat; Nonfat Milk; Lecithin (Soy); Natural and Artificial Flavor]; Sugar; Corn Syrup; High Fructose Corn Syrup; Artificial Color [Yellow 6]; Artificial Flavor; Calcium Chloride; Egg Whites.	<b>U.K. Version</b>   <b>Ingredients:</b> Sugar; Milk; Glucose Syrup; Cocoa Butter; Invert Sugar Syrup; Dried Whey; Cocoa Mass; Vegetable Fats (Palm, Shea); Ammonium Phosphatides; Dried Egg White; Flavourings, Colour (Paprika Extract).
<b>U.S. Version</b>   <b>Ingredients:</b> Water, High Fructose Corn Syrup, Concentrated Juice (Orange, Tangerine, Apple, Lime, Grapefruit, Pear), Citric Acid, Vitamins C and B1, Natural Flavors, Modified Cornstarch, Canola Oil, Sodium Citrate, Cellulose Gum, Sucralose, Sodium Hexametaphosphate, Potassium Sorbate, Yellow #5, Yellow #6, Calcium Disodium EDTA.	<b>U.K. Version</b>   <b>Ingredients:</b> Water, Fruit Juice from Concentrate (Orange, Mandarin, Red Grapefruit, Lime), Sugar, Citric Acid, Acacia Gum, Vitamin C, E, A and D, Guar Gum, Natural Orange Flavourings, Sucralose.

Ces différences, influencées par les préférences locales, la réglementation et l'accessibilité des ingrédients, peuvent modifier leur profil nutritionnel global. Par conséquent, ces variations peuvent parfois donner l'impression de doublons dans des bases de données collaboratives comme Open Food Facts, où un même produit peut apparaître avec des formulations différentes selon le marché local.

## Contribution et origine des données

Dans notre jeu de données, lorsque l'on examine le nom du contributeur on constate qu'il s'agit de « usda-ndb-import ». Cela indique que Open Food Facts compte sur les utilisateurs pour enrichir sa base de données, mais a également importé un grand nombre d'informations depuis le Département de l'Agriculture des États-Unis (USDA).



Le Département de l'Agriculture des États-Unis (USDA) joue un rôle clé dans la collecte et la diffusion d'informations nutritionnelles précises. Il gère une base de données complète, appelée USDA National Nutrient Database (NDB), qui recense la composition nutritionnelle d'un large éventail d'aliments. Cette base inclut des détails sur les nutriments tels que les protéines, les fibres, les sucres, les vitamines, les minéraux, et bien d'autres encore.

Les données de l'USDA sont issues de recherches rigoureuses et sont largement reconnues pour leur exactitude, ce qui renforce la crédibilité des informations nutritionnelles sur Open Food Facts.