

Projet FoodFlix

Application

Recommandation du meilleur produit à un utilisateur selon un mot-clé ou un ensemble de mot-clés.

MVP (Minimum Viable Product):

- Meilleur produit = meilleur Nutri-Score
- Sélection des données pour la France

Gestion de projet sur l'outil Trello

The screenshot displays a Trello project board for a workspace named 'FoodFlix'. The board is organized into six vertical columns representing different stages of a project workflow:

- Fait** (Done):
 - Création de base MySQL
 - importation des data
 - Définition du besoin client = besoin en termes de data / MVP / shouldbe
 - Création environnement de travail
 - + Ajouter une autre carte
- Sprint En Cours** (Sprint In Progress):
 - + Ajouter une carte
- En Cours** (In Progress):
 - recherche d'info sur les notions métier nutriscore /
 - Création de slides
 - + Ajouter une autre carte
- En Attente** (Waiting):
 - échantillonner 10000 lignes
 - + Ajouter une autre carte
- À Venir** (Upcoming):
 - analyse notebooks existants sur kaggle
 - livrable n°1: état des lieux des data not clean et analyse
 - livrable n°2: data cleaning (!! missiong values) Data viz pour avant / après (prouver des hypothèses + analyse des données aberrantes)
 - + Ajouter une autre carte
- Questions** (Questions):
 - combien de valeurs manquantes ?
 - tx de remplissage par ligne ?
 - tx de remplissage par colonne ?
 - redondance ?
 - ajouter des colonnes ?
 - duplicates ?
 - variables numériques ou catégorielles ?
 - nutriscore ? grade ?
 - que va-t-on nettoyer ?
 - yuka ?
 - + Ajouter une autre carte

The interface includes a top navigation bar with icons for home, boards, and search, and a sidebar on the right with a user profile for 'Butler'.

1. Analyse de la donnée

Source des données

Le dataset disponible sur le site Kaggle vient de Open Food Facts, base de données gratuite, libre et collaborative de produits alimentaires autour du monde, avec les ingrédients, allergènes, des informations nutritionnelles et toutes les informations que nous pouvons trouver sur les étiquettes des produits.

Cette base de données est constituée de 356027 lignes et 163 colonnes.

and



Nutri-Score, c'est quoi ?

- Un logo apposé en face avant des emballages qui informe sur la qualité nutritionnelle des produits sous une forme simplifiée et complémentaire à la déclaration nutritionnelle obligatoire (fixée par la réglementation européenne)
- Basé sur une échelle de 5 couleurs : du vert foncé au orange foncé
- Associé à des lettres allant de A à E pour optimiser son accessibilité et sa compréhension par le consommateur



Source: santepubliquefrance.fr

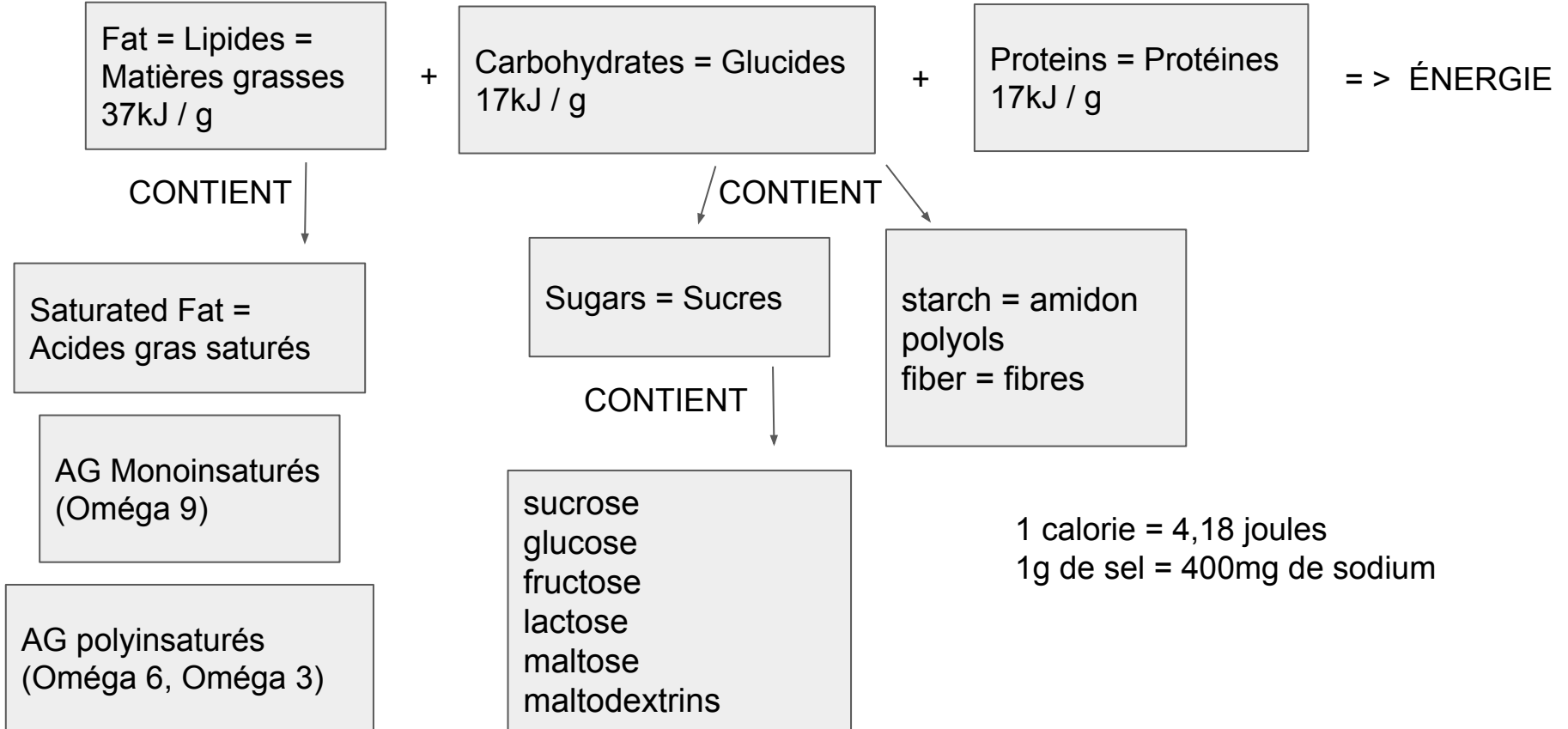
Comment est-il attribué ?

Le logo est attribué sur la base d'un score prenant en compte pour 100 gr ou 100 mL de produit, la teneur :

- en nutriments et aliments à favoriser (fibres, protéines, fruits, légumes, légumineuses, fruits à coques, huile de colza, de noix et d'olive),
- et en nutriments à limiter (énergie, acides gras saturés, sucres, sel).

Après calcul, le score obtenu par un produit permet de lui attribuer une lettre et une couleur.

Compréhension métier



Analyse de la donnée

L'analyse de la donnée brute se fait dans un Jupyter Notebook

<https://github.com/ATelders/foodflix/blob/main/notebook/2021-03-20-at-raw-data-analysis.ipynb>

et dans un rapport Pandas profiling, dans lequel on visualise pour chaque colonne les informations suivantes :

- **Type inference**: detect the types of columns in a dataframe.
- **Essentials**: type, unique values, missing values
- **Quantile statistics** like minimum value, Q1, median, Q3, maximum, range, interquartile range
- **Descriptive statistics** like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
- **Most frequent values**
- **Histograms**
- **Correlations** highlighting of highly correlated variables, Spearman, Pearson and Kendall matrices
- **Missing values** matrix, count, heatmap and dendrogram of missing values
- **Duplicate rows** Lists the most occurring duplicate rows
- **Text analysis** learn about categories (Uppercase, Space), scripts (Latin, Cyrillic) and blocks (ASCII) of text data

2. Sélection des données

Sélection des colonnes

1. Suppression de 75 colonnes vides à 99 %
2. Sélection de colonnes utiles pour notre application

- product_name + brands
- ingredients_text (on peut déduire les allergènes de la liste des ingrédients)
- nutrition_grade
- fiber_100g
- proteins_100g
- energy_100g
- fat_100g
- saturated_fat_100g
- carbohydrates_100g
- sugars_100g
- salt_100g

Le nutriscore prend en compte les données indiquées en rouge et en vert

Cependant il manque le taux de fruits /légumes /légumineuses / fruits à coque / huile de colza / noix / olive qui ont été éliminés à l'étape1 (ex : colonne 'fruits-vegetables-nuts_100g' : 3228 lignes non nulles sur 352799 -> moins de 1%)

Sélection des données

Observations:

Il y a beaucoup de colonnes vides ou quasi vides.

Je décide de supprimer les colonnes vides à 99% dans un premier temps.

Dans un second temps, je choisis les colonnes nécessaires au calcul du nutriscore. Je décide de garder la colonnes 'fruits et légumes' même si elle est très peu renseignée. Je garde aussi les allergènes, bien qu'on puisse les déduire de la colonne ingrédients, qui est elle même très incomplète.

Pandas profiling report

Données brutes : échantillon de 10000 lignes

https://htmlpreview.github.io/?https://github.com/ATelders/foodflix/blob/main/results/profiling_raw.html

Données nettoyées : échantillon de 10000 lignes

https://htmlpreview.github.io/?https://github.com/ATelders/foodflix/blob/main/results/profiling_intermediate.html

3. Nettoyage des données

Nettoyage des données

Sélection des produits français. (MVP)

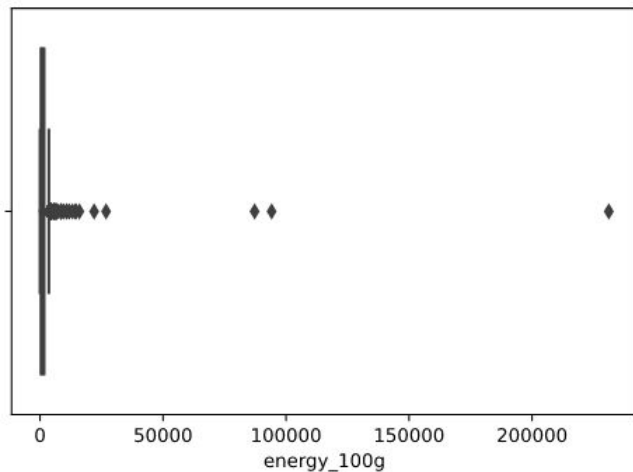
Il y a beaucoup d'erreurs lexicales dans la colonne 'countries'. Certains produits ont le nom du pays, tandis que d'autres ont des abréviations telles que 'fr', 'be' ... Certains produits ont plusieurs pays.

On décide de regrouper les valeurs contenant 'fr' dans la colonne 'countries', et de les renommer 'France'.

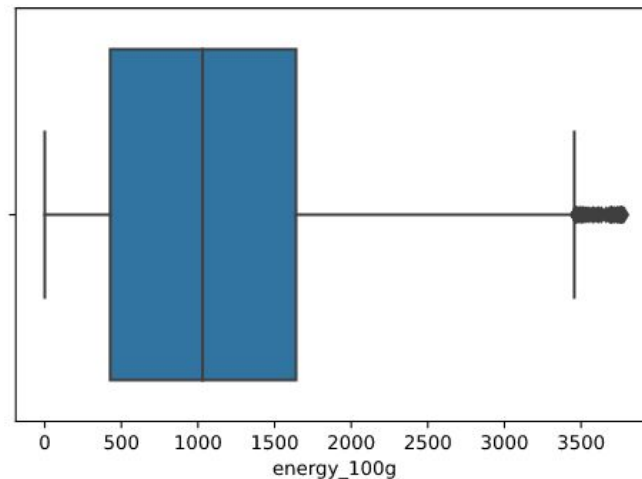
Ceci permet de sélectionner les produits fabriqués et/ou transformés en France, en incluant les produits multi-origines.

Nettoyage des données

Suppression des valeurs aberrantes pour l'énergie par 100g. Elle ne peut être supérieure à 3800 kJ. Les valeurs maximales (3700-3800) correspondent à des huiles.



Avant



Après

Nettoyage des données

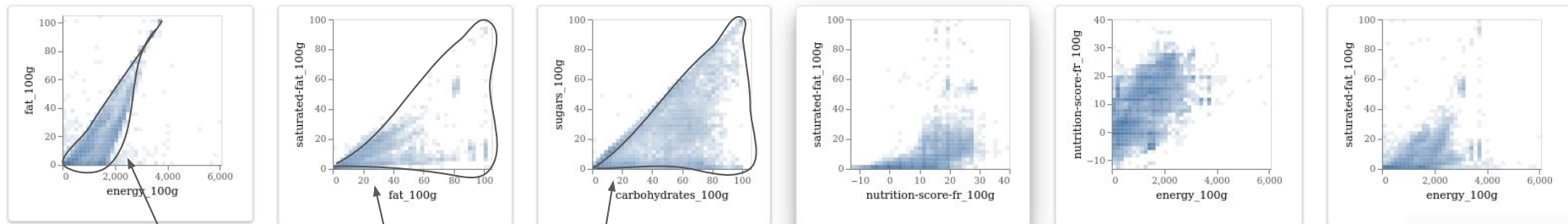
Suppression des valeurs aberrantes pour les données nutritionnelles par 100g.
Ces valeurs ne peuvent être inférieures à 0 ou supérieures à 100.

La somme des 'carbohydrates', 'fat' et 'proteins' ne peut être supérieure à 100g.

Nettoyage des données

Suppression des lignes où aucune donnée nutritionnelle n'existe, ces lignes ne permettent pas de calculer un nutriscore.

Lux report



Scroll for 12 more charts

Toutes les valeurs qui sont en dehors de ces triangles sont des valeurs impossibles.
En effet, les graisses saturées ne peuvent pas être supérieures aux lipides, les sucres aux glucides, etc.

Recalcul de l'énergie

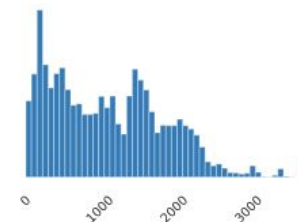
À partir des colonnes 'fat_100g', 'carbohydrates_100g' et 'proteins_100g'

energy_100g

Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION

Distinct	2245	Minimum	0
Distinct (%)	22.6%	Maximum	3766
Missing	57	Zeros	79
Missing (%)	0.6%	Zeros (%)	0.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1115.18276	Memory size	78.2 KiB



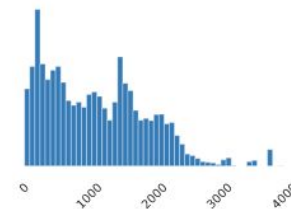
calculated_energy

Real number ($\mathbb{R}_{\geq 0}$)

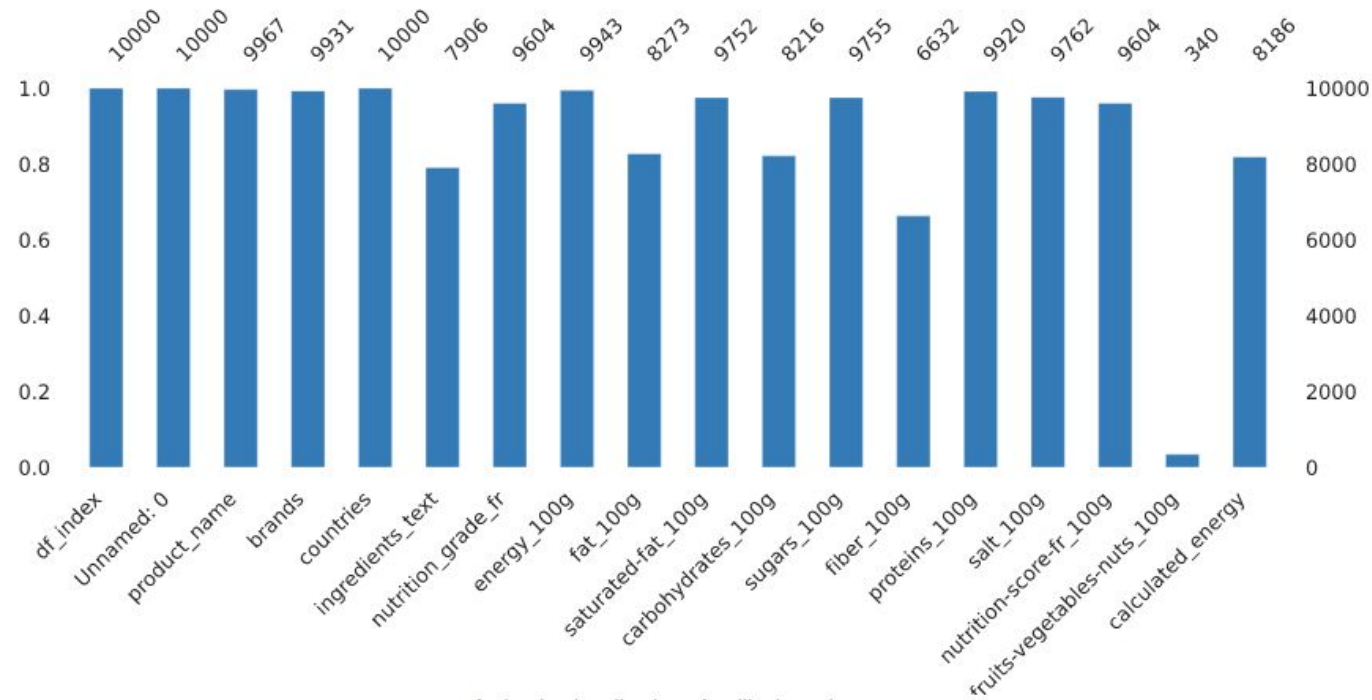
HIGH CORRELATION

MISSING

Distinct	6171	Minimum	0
Distinct (%)	75.4%	Maximum	3968.2
Missing	1814	Zeros	83
Missing (%)	18.1%	Zeros (%)	0.8%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	1103.202935	Memory size	78.2 KiB

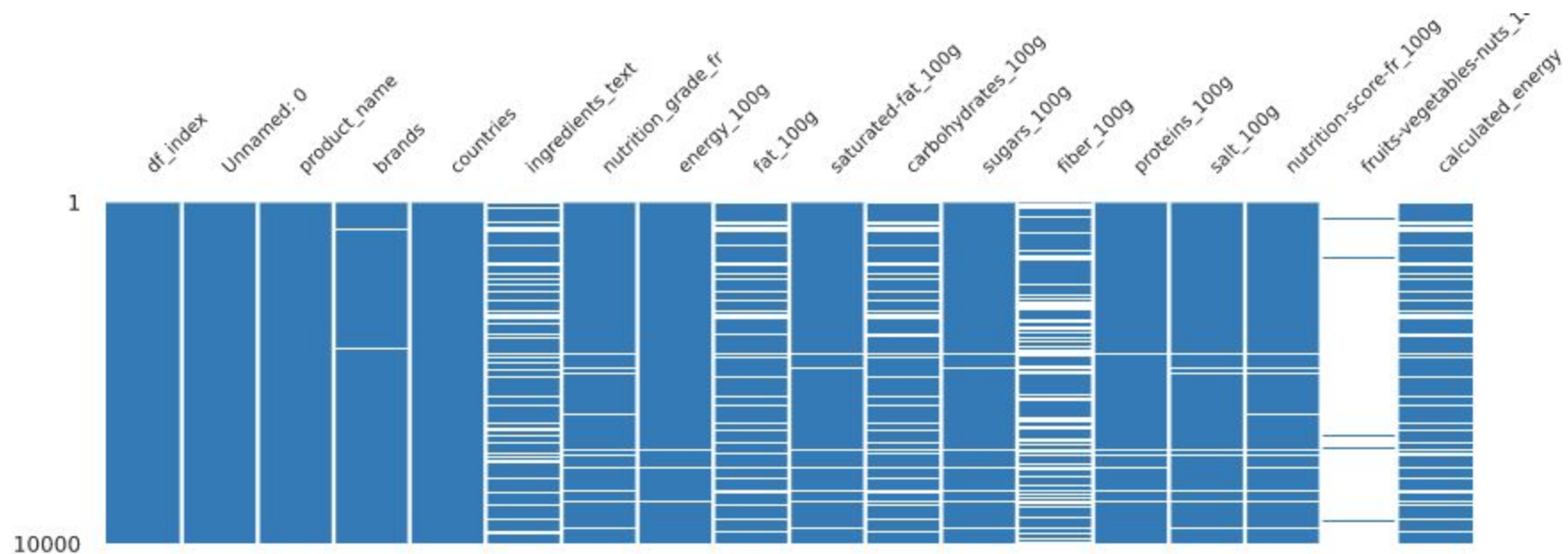


Données manquantes



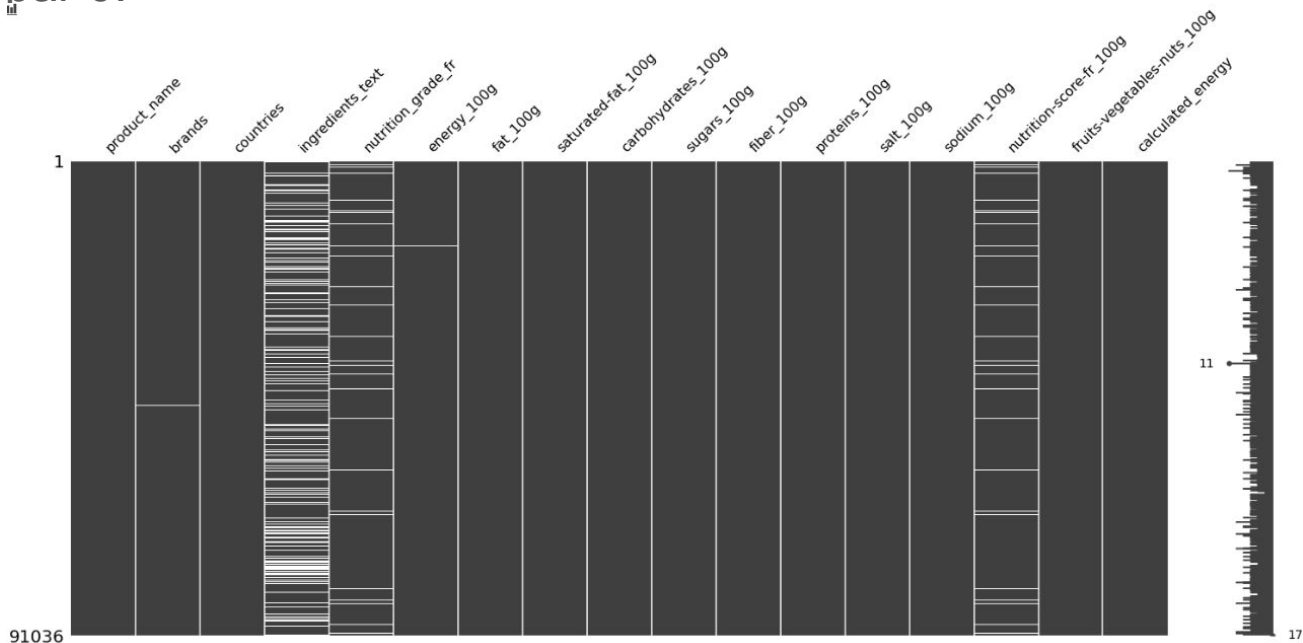
A simple visualization of nullity by column.

Données manquantes



Stratégie de remplissage des valeurs manquantes.

Pour les catégories utiles au calcul du nutriscore, nous remplissons les valeurs manquantes par 0.



Stratégie de remplissage des valeurs manquantes.

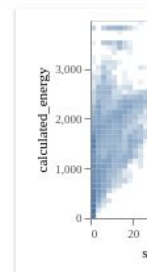
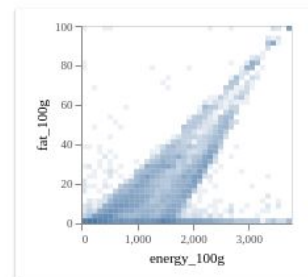
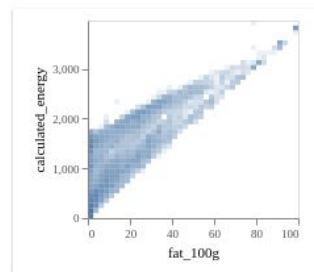
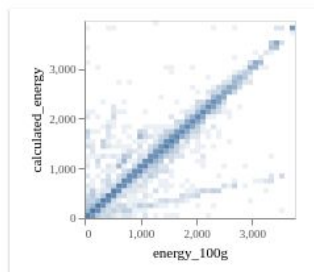
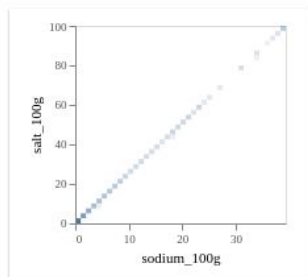
Correlation

Distribution

Occurrence



Show relationships between two quantitative attributes.



On remarque ici que le sel et le sodium sont directement corrélés. En effet, la valeur du sodium pour le calcul du nutriscore est calculée à partir du taux de sel. On pourrait choisir de ne garder qu'une seule des deux colonnes.

Stratégie de remplissage des valeurs manquantes.

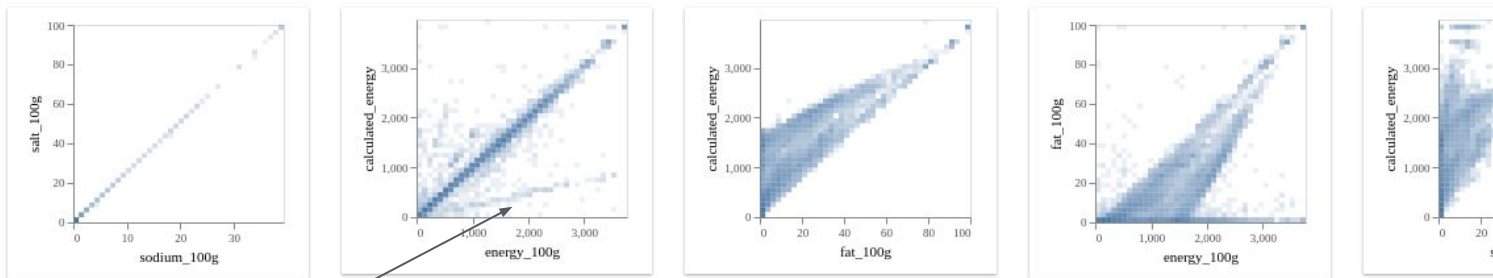
[Correlation](#)

[Distribution](#)

[Occurrence](#)



Show relationships between two quantitative attributes.



On remarque ici une ligne plus légère dans la relation entre l'énergie calculée en fonction du taux d'énergie pour 100g des données d'origine.

Une hypothèse est que c'est une erreur d'irrégularité, car la ligne fait penser à la valeur convertie en kilocalories.

Stratégie de remplissage des valeurs manquantes.

Nutrition-score

Il sert à calculer la note du nutriscore (ou nutrigrade)

Après nettoyage, nous avons 4% de données manquantes pour lesquelles il faudra essayer de recalculer à partir des colonnes

`nutrition-score-fr_100g`

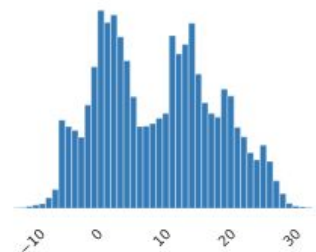
Real number (\mathbb{R})

MISSING

ZEROS

Distinct	48
Distinct (%)	0.5%
Missing	396
Missing (%)	4.0%
Infinite	0
Infinite (%)	0.0%
Mean	8.833402749

Minimum	-13
Maximum	35
Zeros	500
Zeros (%)	5.0%
Negative	1529
Negative (%)	15.3%
Memory size	78.2 KiB



Next steps

Calculer le 'nutrition-score-fr_100g' et le 'nutrition_grade_fr'

Détecter toutes les valeurs aberrantes.

CALCUL DU NUTRI-SCORE

01

Attribution des points à différents facteurs nutritionnels



Nutriments à limiter

Points N	Score pour les boissons			Score pour les produits gras		
Points	Energie (kJ)	Score (N)	Energie (kJ)	Servos (g)	Graines (g)	Sodium (mg)
0	≤ 335	≤ 4,5	≤ 0	≤ 0	≤ 1	≤ 10
1	> 335	> 4,5	≤ 90	≤ 1,5	> 1	≤ 16
2	> 470	> 9	≤ 180	≤ 3	> 2	≤ 22
3	> 1005	> 13,5	≤ 90	≤ 4,5	> 3	≤ 28
4	> 1340	> 18	≤ 120	≤ 6	> 4	≤ 34
5	> 1675	> 22,5	≤ 150	≤ 7,5	> 5	≤ 40
6	> 2010	> 27	≤ 180	≤ 9	> 6	≤ 46
7	> 2345	> 31	≤ 210	≤ 10,5	> 7	≤ 52
8	> 2680	> 36	≤ 240	≤ 12	> 8	≤ 58
9	> 3015	> 40	≤ 270	≤ 13,5	> 9	≤ 64
10	> 3350	> 45	≤ 300	≤ 15	> 10	≤ 70
Somme des points	0 à 30	0 à 130	0 à 300	0 à 150	0 à 30	0 à 30
Somme des points pour l'énergie, les sucres, les graisses saturées et le sodium						