

Formation de Data Science - Openclassrooms

Formation Ouverte et à Distance – FOAD par Pôle Emploi
Solutions 100% à distance

Projet 3 : Concevez une application au service de la santé publique

Étudiant : Maria Daniela Barrios
Mentor : Dan Slama

03 janvier, 2022

Idée d'une application qui classe et montre les produits low sodium selon leur utilisation ou leur catégorie (ex. entrées, sauces, desserts, boissons, etc.) et donne le classement nutri-score.

Les régimes à low sodium sont couramment prescrits aux personnes souffrant d'une maladie rénale, d'une maladie cardiaque ou d'une hypertension artérielle, afin de gérer les symptômes et de prévenir les complications.

Les données sont extraites de **Open Food Facts** ([see https://world.openfoodfacts.org/data](https://world.openfoodfacts.org/data))

Stratégie pour accomplir notre mission

→ **Analyse exploratoire**

- Décrire les informations sur le nombre de lignes et de colonnes des données

→ **Sélection de variables ou d'indicateurs importants**

- Filtrage des données en fonction des variables importantes

→ **Nettoyage des données**

- Révision des valeurs manquantes et des doublons

→ **Validation des données**

- Traitement des valeurs aberrantes

→ **Analyse exploratoire des données : analyse univariée, multivariée et bivariée**

- Détermination des ordres de grandeur, des distributions, des relations entre les variables

Analyse exploratoire

Le fichier de données contient 2054909 lignes et 187 colonnes

Informations générales :

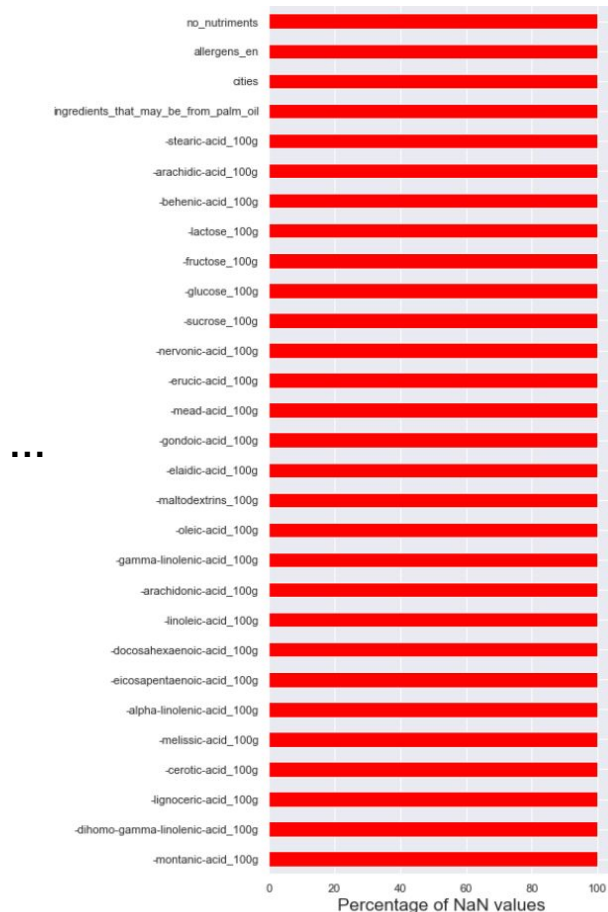
- Les champs qui se terminent ou commencent par un code de deux lettres de la langue (par exemple, fr pour le français) correspondent à l'ensemble des balises de cette langue
- Les champs qui finissent par “_100g” correspondent à la quantité d'un nutriment (en g, ou en kJ pour l'énergie) pour 100g ou 100ml de produit
- Les champs terminés par “_serving” correspondent à la quantité d'un nutriment (en g, ou en kJ pour l'énergie) pour 1 portion du produit
- Nutrition_grade_fr : note nutritionnelle (“a” à “e”) : <https://fr.openfoodfacts.org/nutriscore>
- Nutrition-score-fr_100g : Nutri-Score - Score nutritionnel dérivé du score UK FSA et adapté au marché français (formule définie par l'équipe du Professeur Hercberg). Les champs sont numériques allant de -15 à 40. Plus le score est bas, meilleur est le produit
- Code : code-barres du produit (peut être EAN-13 ou des codes internes pour certains magasins d'alimentation)

Analyse exploratoire

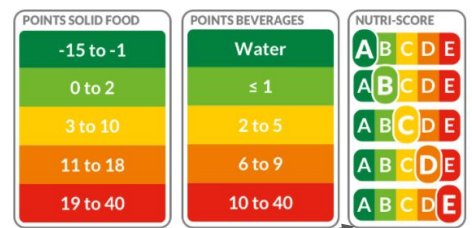
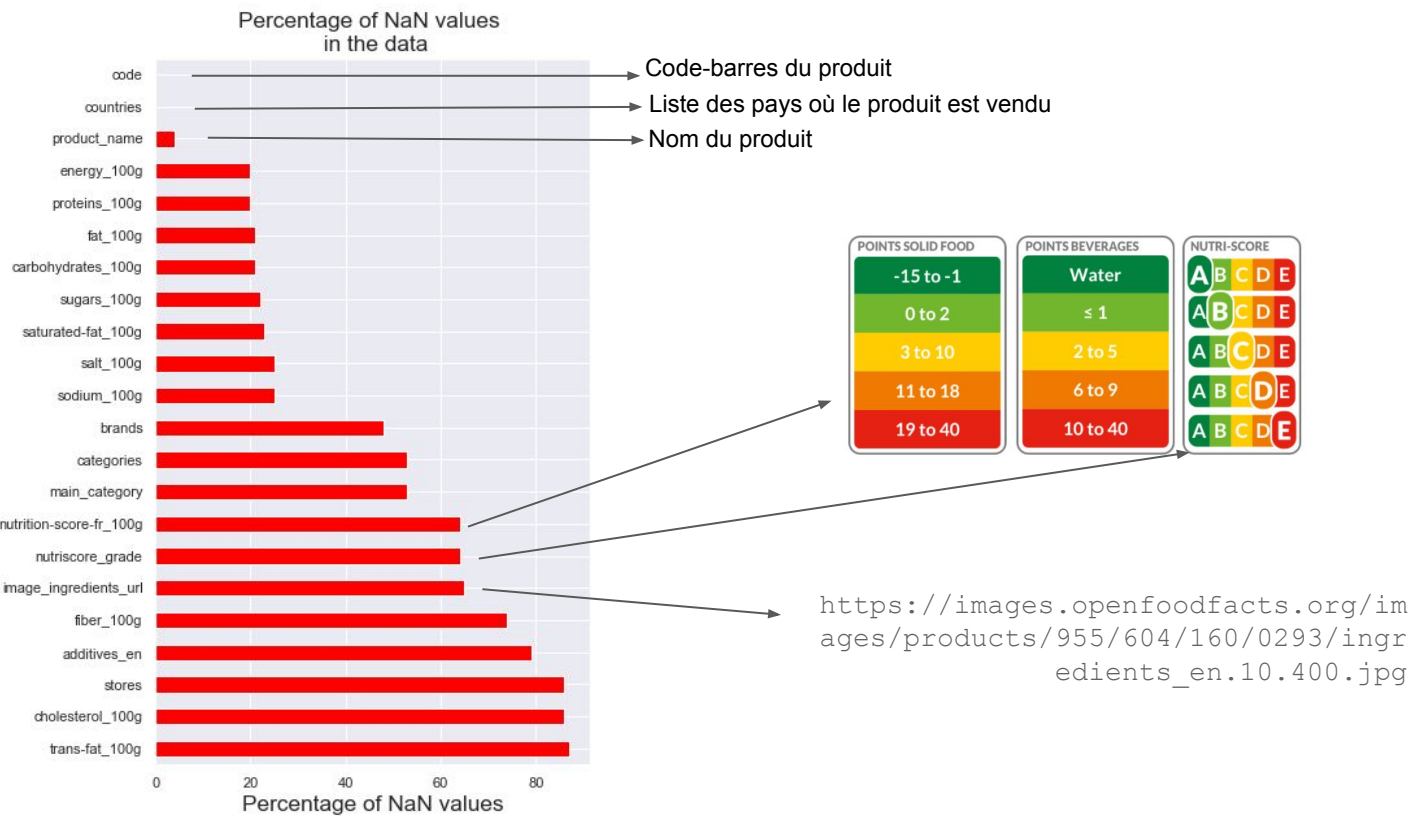
Percentage of NaN values
in the data



160 colonnes sur 187 ont plus de
40% de valeurs manquantes



Sélection de variables ou d'indicateurs importants



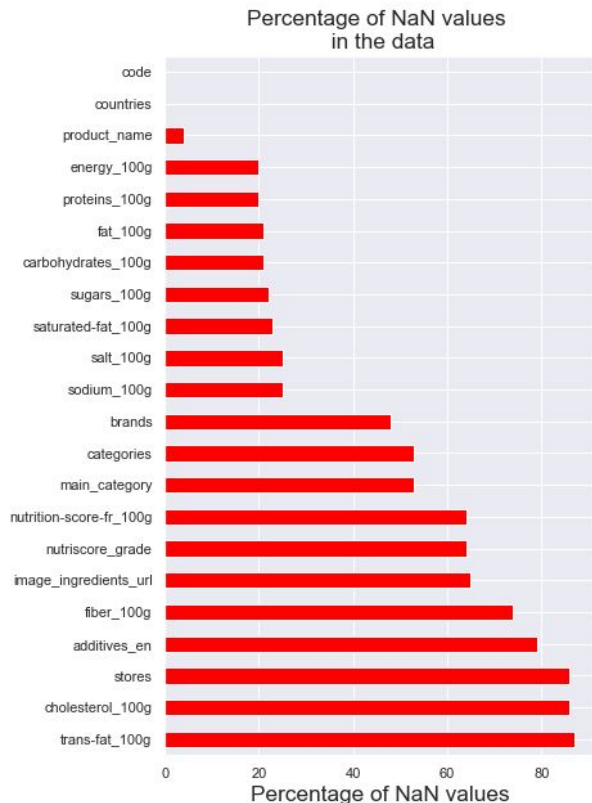
Ingredients: Wild Caught Sardines Fish, Extra Virgin Olive Oil, Salt

Net 120g

| NUTRITION INFORMATION | | |
|--|-------------|----------|
| Servings per package: 2 | | |
| Serving size: 39g (1/2 can) Drained Weight | | |
| | Per serving | Per 100g |
| Energy (1kcal = 4.2kJ) | 73kcal | 187kcal |
| Protein | 9.1g | 23.4g |
| Total fat | 4.0g | 10.3g |
| - saturated fat | 1.1g | 2.9g |
| - Omega 3 | 0.7g | 1.9g |
| - trans fat | 0.0g | 0.1g |
| Cholesterol | 23mg | 58mg |
| Carbohydrate | 0.0g | 0.1g |
| Dietary fibre | 0.5g | 1.2g |
| Sodium | 100mg | 257mg |
| Calcium | 127mg | 325mg |

Nettoyage des données

Valeurs manquantes et doublons



Il y avait 319 valeurs doublées pour la variable 'Code' sur 2054909 lignes

- Nous avons éliminé ces valeurs doublées
- Nous gardons 2054590 lignes et 22 colonnes

- Compte tenu de la nature de nos données, aucun remplacement des valeurs NaN par 0, la moyenne ou la médiane n'a été effectué
- Nous ne pouvons pas ajouter de valeurs aux produits qui pourraient modifier les statistiques et les distributions
- D'autres filtres ont été appliqués pour la validation des données

Nettoyage des données

Validation des données

| | salt_100g | sodium_100g | cholesterol_100g | saturated-fat_100g | fat_100g | trans-fat_100g | sugars_100g | carbohydrates_100g | proteins_100g | fiber_100g | energy_100g | nutrition-score-fr_100g |
|-------|--------------|--------------|------------------|--------------------|--------------|----------------|---------------|--------------------|---------------|---------------|--------------|-------------------------|
| count | 1.523258e+06 | 1.523255e+06 | 268984.000000 | 1.573472e+06 | 1.621962e+06 | 264954.000000 | 1.600814e+06 | 1.621548e+06 | 1.623261e+06 | 5.185520e+05 | 1.630186e+06 | 730993.000000 |
| mean | 6.560690e+04 | 2.624281e+04 | 0.055349 | 6.351608e+04 | 7.330625e+07 | 0.150308 | 6.253066e+07 | 6.165679e+04 | 6.157176e+04 | 1.892438e+41 | 4.088833e+36 | 9.113163 |
| std | 8.096942e+07 | 3.238780e+07 | 1.759469 | 7.966696e+07 | 8.664786e+10 | 29.553317 | 7.903688e+10 | 7.847708e+07 | 7.843566e+07 | 1.362755e+44 | 5.220571e+39 | 8.850129 |
| min | 0.000000e+00 | 0.000000e+00 | 0.000000 | 0.000000e+00 | 0.000000e+00 | 0.000000 | -1.000000e+00 | -1.000000e+00 | -5.000000e+02 | -2.000000e+01 | 0.000000e+00 | -15.000000 |
| 25% | 8.000000e-02 | 3.200000e-02 | 0.000000 | 1.000000e-01 | 8.000000e-01 | 0.000000 | 6.000000e-01 | 3.500000e+00 | 1.300000e+00 | 0.000000e+00 | 4.180000e+02 | 1.000000 |
| 50% | 5.800000e-01 | 2.320000e-01 | 0.000000 | 1.800000e+00 | 7.000000e+00 | 0.000000 | 3.570000e+00 | 1.510000e+01 | 6.000000e+00 | 1.600000e+00 | 1.079000e+03 | 10.000000 |
| 75% | 1.400000e+00 | 5.600000e-01 | 0.022000 | 7.100000e+00 | 2.120000e+01 | 0.000000 | 1.750000e+01 | 5.300000e+01 | 1.250000e+01 | 3.600000e+00 | 1.674000e+03 | 16.000000 |
| max | 9.993273e+10 | 3.997309e+10 | 362.000000 | 9.993273e+10 | 1.100000e+14 | 14800.000000 | 1.000000e+14 | 9.993273e+10 | 9.993273e+10 | 9.813275e+46 | 6.665559e+42 | 40.000000 |

Pour valider les données afin qu'elles correspondent à la réalité, quelques règles simples ont été appliquées :

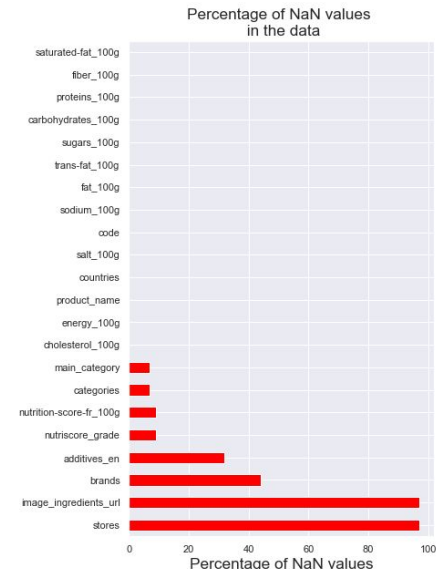
- La quantité en grammes de glucides doit être supérieure à celle du sucre.
- La quantité en grammes de graisses doit être supérieure à celle des graisses saturées et trans.
- La somme des lipides, glucides, protéines, fibres et sel doit être inférieure ou égale à 100 grammes.
- Nous ne garderons que les lignes où par exemple sucre_100g et les autres sont positives.

Nettoyage des données

Validation des données

| | salt_100g | sodium_100g | cholesterol_100g | saturated-fat_100g | fat_100g | trans-fat_100g | sugars_100g | carbohydrates_100g | proteins_100g | fiber_100g | energy_100g | nutrition-score-fr_100g |
|-------|---------------|---------------|------------------|--------------------|---------------|----------------|---------------|--------------------|---------------|---------------|---------------|-------------------------|
| count | 213996.000000 | 213996.000000 | 212254.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213601.000000 | 193907.000000 |
| mean | 1.238011 | 0.495207 | 0.039663 | 4.075873 | 10.715881 | 0.033794 | 14.329648 | 32.021267 | 7.242566 | 1.851611 | 1054.881900 | 8.556679 |
| std | 3.725091 | 1.490036 | 1.113710 | 6.299592 | 13.521522 | 0.410913 | 19.087392 | 27.762173 | 8.060134 | 2.662877 | 687.390436 | 9.080842 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | -14.000000 |
| 25% | 0.137160 | 0.054864 | 0.000000 | 0.000000 | 0.220000 | 0.000000 | 1.300000 | 7.140000 | 1.639344 | 0.000000 | 403.000000 | 1.000000 |
| 50% | 0.720000 | 0.288000 | 0.000000 | 1.250000 | 5.360000 | 0.000000 | 5.000000 | 22.350000 | 4.710000 | 1.000000 | 1046.000000 | 10.000000 |
| 75% | 1.422400 | 0.568960 | 0.026000 | 6.160000 | 17.020000 | 0.000000 | 21.840000 | 56.250000 | 10.000000 | 2.900000 | 1586.000000 | 16.000000 |
| max | 100.000000 | 40.000000 | 116.915000 | 100.000000 | 100.000000 | 35.710000 | 100.000000 | 100.000000 | 90.910000 | 72.500000 | 13213.000000 | 38.000000 |

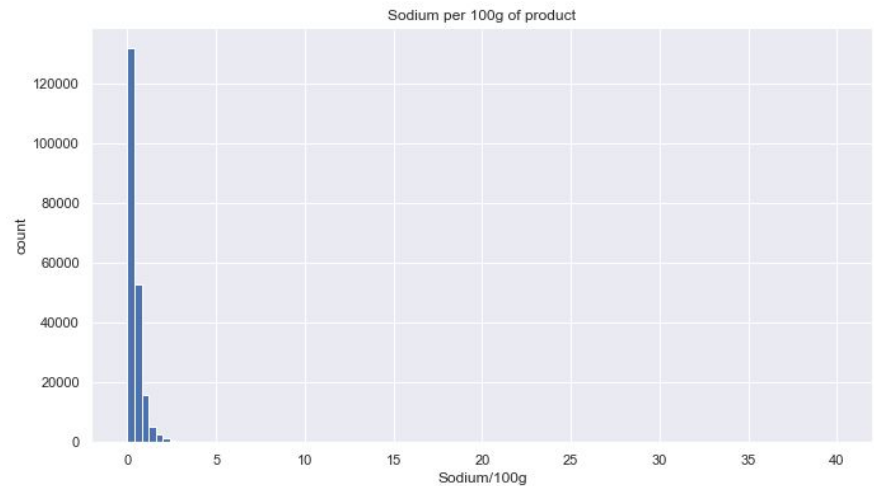
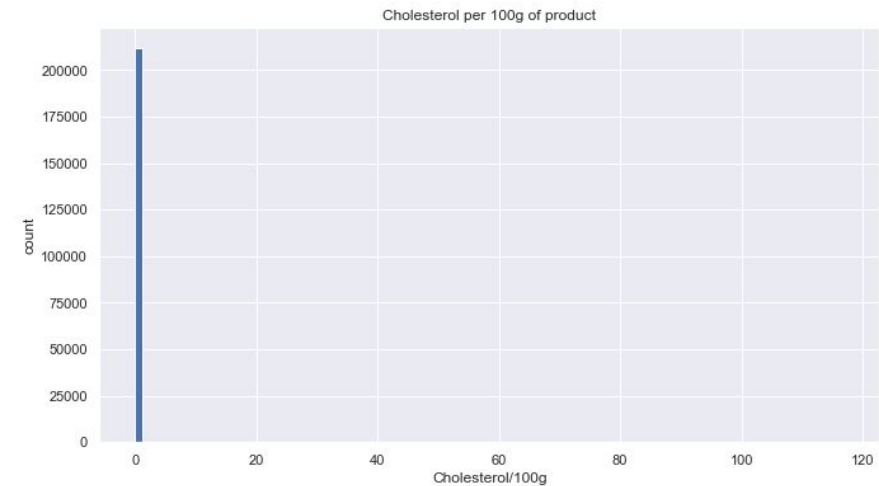
- Le cadre de données avec les variables pertinentes contient maintenant 213996 lignes et 22 colonnes
- Une observation intéressante est que la valeur maximale de cholesterol_100g est supérieure à 100
- Il serait pratique de vérifier les valeurs de cholestérol et de sodium en fonction de leurs unités - Les étiquettes des produits pourraient indiquer les valeurs de cholestérol et de sodium en mg par 100g de produit
- Si une valeur de cholestérol ou de sodium est supérieure à 100, cela reste logique si ses unités sont des mg - Cependant, il n'est pas précisé quelles sont les unités de ces valeurs



Nettoyage des données

Validation des données

| | salt_100g | sodium_100g | cholesterol_100g | saturated-fat_100g | fat_100g | trans-fat_100g | sugars_100g | carbohydrates_100g | proteins_100g | fiber_100g | energy_100g | nutrition-score-fr_100g |
|-------|---------------|---------------|------------------|--------------------|---------------|----------------|---------------|--------------------|---------------|---------------|---------------|-------------------------|
| count | 213996.000000 | 213996.000000 | 212254.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213601.000000 | 193907.000000 |
| mean | 1.238011 | 0.495207 | 0.039663 | 4.075873 | 10.715881 | 0.033794 | 14.329648 | 32.021267 | 7.242566 | 1.851611 | 1054.881900 | 8.556679 |
| std | 3.725091 | 1.490036 | 1.113710 | 6.299592 | 13.521522 | 0.410913 | 19.087392 | 27.762173 | 8.060134 | 2.662877 | 687.390436 | 9.080842 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | -14.000000 |
| 25% | 0.137160 | 0.054864 | 0.000000 | 0.000000 | 0.220000 | 0.000000 | 1.300000 | 7.140000 | 1.639344 | 0.000000 | 403.000000 | 1.000000 |
| 50% | 0.720000 | 0.288000 | 0.000000 | 1.250000 | 5.360000 | 0.000000 | 5.000000 | 22.350000 | 4.710000 | 1.000000 | 1046.000000 | 10.000000 |
| 75% | 1.422400 | 0.568960 | 0.026000 | 6.160000 | 17.020000 | 0.000000 | 21.840000 | 56.250000 | 10.000000 | 2.900000 | 1586.000000 | 16.000000 |
| max | 100.000000 | 40.000000 | 116.915000 | 100.000000 | 100.000000 | 35.710000 | 100.000000 | 100.000000 | 90.910000 | 72.500000 | 13213.000000 | 38.000000 |



Nettoyage des données

Validation des données

Cholesterol_100g et sodium_100g

| | product_name | brands | categories | stores | countries | additives_en | main_category | salt_100g | sodium_100g | cholesterol_100g |
|---------|--|------------|---|--------|-----------------------|--------------|---------------|-----------|-------------|------------------|
| 2045201 | Ayam brand, sardines in extra virgin olive oil | Ayam Brand | Canned foods, Seafood, Fishes, Sardines | NaN | France, United States | NaN | ensardines | 0.8925 | 0.357 | 0.036 |

| | product_name | brands | categories | stores | countries | additives_en | main_category | salt_100g | sodium_100g | cholesterol_100g |
|--------|---|---------------|---|-------------|---------------|--------------|---------------|-----------|-------------|------------------|
| 240519 | Red onion and thyme fresh spreadable cheese | Norseland Inc | Dairies, Fermented foods, Fermented milk produ... | Whole Foods | United States | NaN | encheeses | 1.34 | 0.536 | 53.571 |

<https://world.openfoodfacts.org/product/9556041600293/ayam-brand-sardines-in-extra-virgin-olive-oil>

- La valeur du cholestérol pour 100g de produit est de 36 mg. Dans les données nous avons 0,036, ce qui signifie que les valeurs originales ont été divisées par 1000.
- Le sodium pour 100 g de produit selon nos données, a une valeur de 0,357. La valeur indiquée sur l'étiquette est de 257 mg pour 100 g de produit et n'est pas exactement la même que la précédente.

Ingredients: Wild Caught Sardines Fish, Extra Virgin Olive Oil, Salt

Net 120g

| NUTRITION INFORMATION | | |
|--|-------------|----------|
| Servings per package: 2 | | |
| Serving size: 39g (1/2 can) Drained Weight | | |
| | Per serving | Per 100g |
| Energy (1kcal = 4.2kJ) | 73kcal | 187kcal |
| Protein | 9.1g | 23.4g |
| Total fat | 4.0g | 10.3g |
| - saturated fat | 1.1g | 2.9g |
| - Omega 3 | 0.7g | 1.9g |
| - trans fat | 0.0g | 0.1g |
| Cholesterol | 23mg | 58mg |
| Carbohydrate | 0.0g | 0.1g |
| Dietary fibre | 0.5g | 1.2g |
| Sodium | 100mg | 257mg |
| Calcium | 127mg | 325mg |

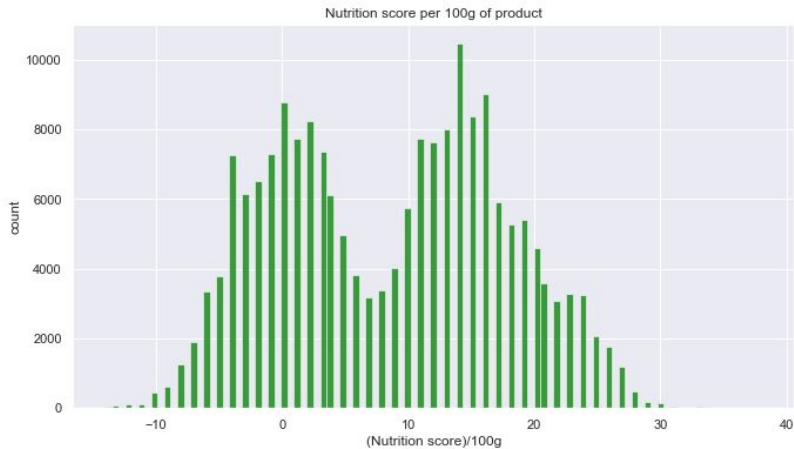
<https://world.openfoodfacts.org/product/0075501246202/red-onion-and-thyme-fresh-spreadable-cheese-norseland-inc>

- La valeur du cholestérol pour 100g de produit est de 53.6mg et dans nos données nous avons la même valeur.
- Le sodium pour 100g de produit dans nos données est de 0,536. Selon une source de données externe : <https://www.instacart.com/products/3198219-snofrisk-red-onion-and-thyme-fresh-spreadable-cheese-4-4-oz> - nous avons constaté que le sodium par portion (28g) est de 150mg. Si l'on fait le calcul pour 100g de produit, on obtient exactement 536 mg.

Analyse exploratoire des données

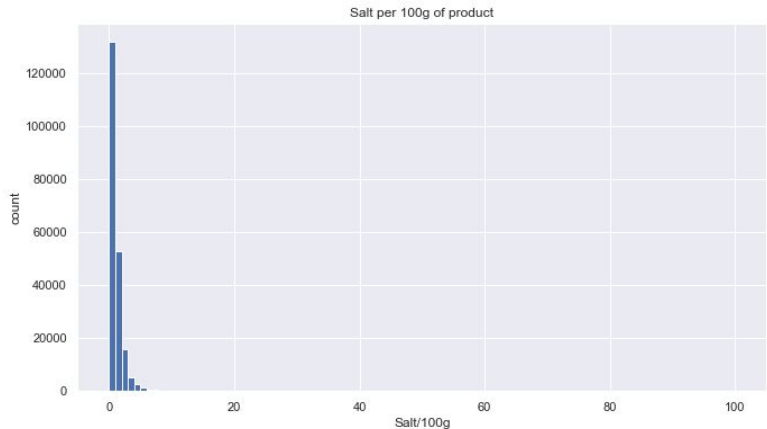
Analyse univariée

Distribution : nutrition-score-fr_100g



| nutrition-score-fr_100g | |
|-------------------------|---------------|
| count | 193907.000000 |
| mean | 8.556679 |
| std | 9.080842 |
| min | -14.000000 |
| 25% | 1.000000 |
| 50% | 10.000000 |
| 75% | 16.000000 |
| max | 38.000000 |

Distribution : salt_100g

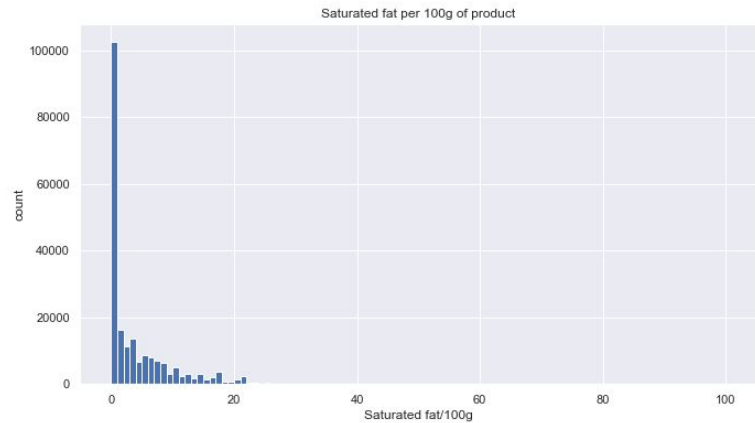


| salt_100g | |
|-----------|---------------|
| count | 213996.000000 |
| mean | 1.238011 |
| std | 3.725091 |
| min | 0.000000 |
| 25% | 0.137160 |
| 50% | 0.720000 |
| 75% | 1.422400 |
| max | 100.000000 |

Analyse exploratoire des données

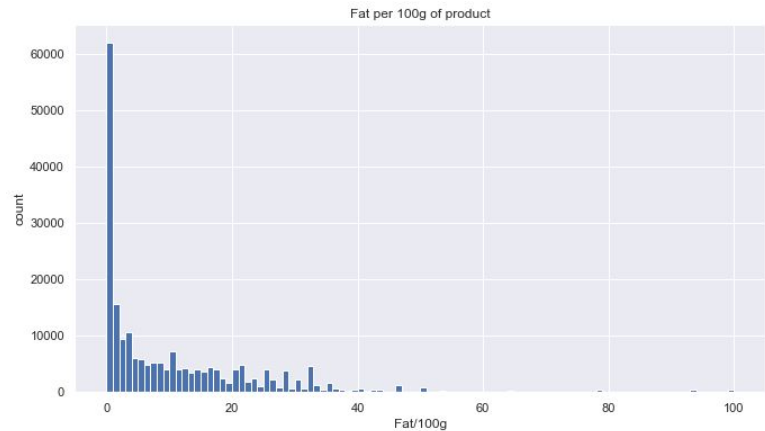
Analyse univariée

Distribution : saturated-fat_100g



| saturated-fat_100g | |
|--------------------|---------------|
| count | 213996.000000 |
| mean | 4.075873 |
| std | 6.299592 |
| min | 0.000000 |
| 25% | 0.000000 |
| 50% | 1.250000 |
| 75% | 6.160000 |
| max | 100.000000 |

Distribution : fat_100g

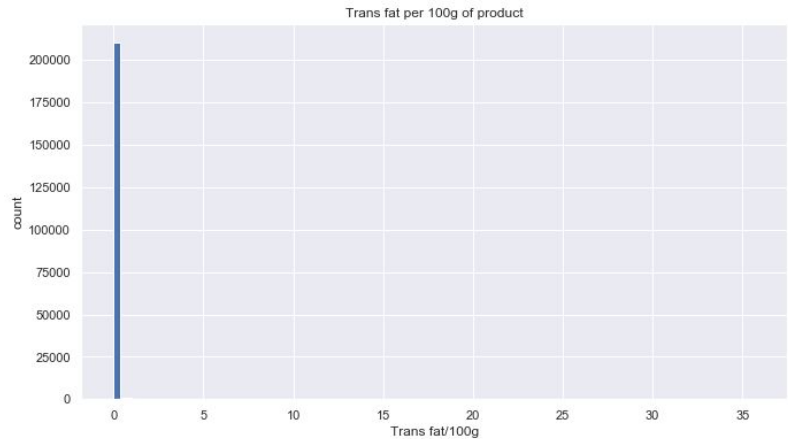


| fat_100g | |
|----------|---------------|
| count | 213996.000000 |
| mean | 10.715881 |
| std | 13.521522 |
| min | 0.000000 |
| 25% | 0.220000 |
| 50% | 5.360000 |
| 75% | 17.020000 |
| max | 100.000000 |

Analyse exploratoire des données

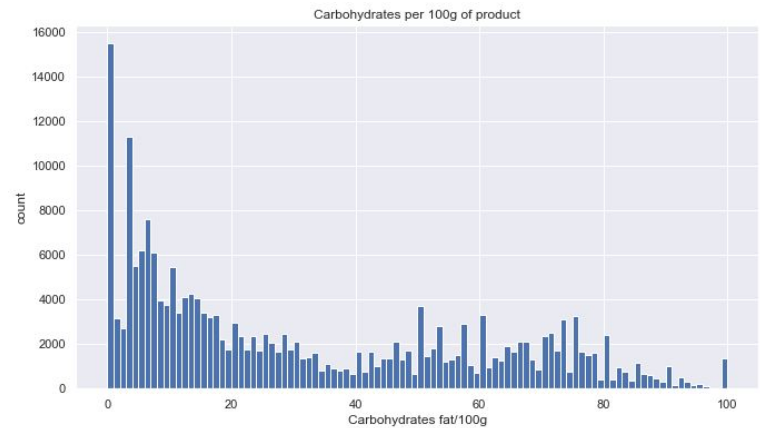
Analyse univariée

Distribution : trans-fat_100g



| trans-fat_100g | |
|----------------|---------------|
| count | 213996.000000 |
| mean | 0.033794 |
| std | 0.410913 |
| min | 0.000000 |
| 25% | 0.000000 |
| 50% | 0.000000 |
| 75% | 0.000000 |
| max | 35.710000 |

Distribution : carbohydrates_100g

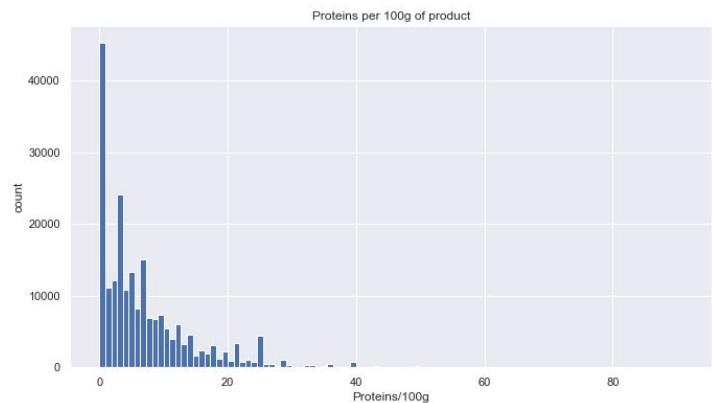


| carbohydrates_100g | |
|--------------------|---------------|
| count | 213996.000000 |
| mean | 32.021267 |
| std | 27.762173 |
| min | 0.000000 |
| 25% | 7.140000 |
| 50% | 22.350000 |
| 75% | 56.250000 |
| max | 100.000000 |

Analyse exploratoire des données

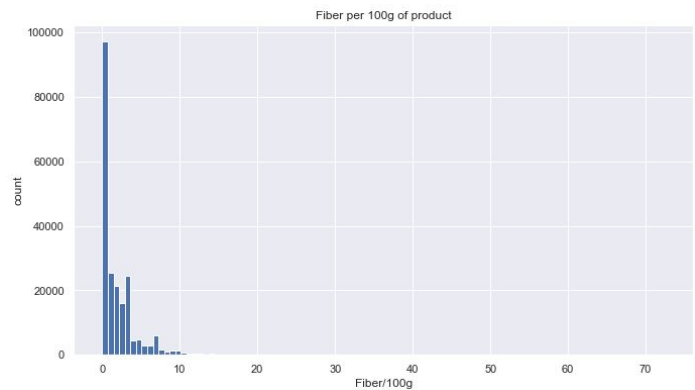
Analyse univariée

Distribution : proteins_100g



| | proteins_100g |
|-------|---------------|
| count | 213996.000000 |
| mean | 7.242566 |
| std | 8.060134 |
| min | 0.000000 |
| 25% | 1.639344 |
| 50% | 4.710000 |
| 75% | 10.000000 |
| max | 90.910000 |

Distribution : fiber_100g

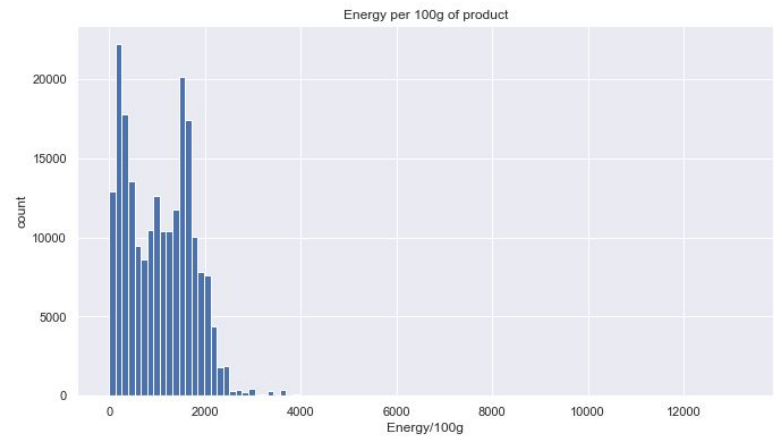


| | fiber_100g |
|-------|---------------|
| count | 213996.000000 |
| mean | 1.851611 |
| std | 2.662877 |
| min | 0.000000 |
| 25% | 0.000000 |
| 50% | 1.000000 |
| 75% | 2.900000 |
| max | 72.500000 |

Analyse exploratoire des données

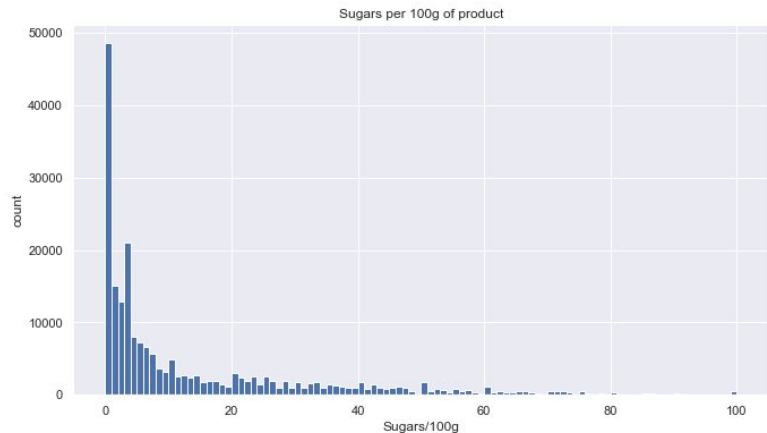
Analyse univariée

Distribution : energy_100g



| energy_100g | |
|-------------|---------------|
| count | 213601.000000 |
| mean | 1054.881900 |
| std | 687.390436 |
| min | 0.000000 |
| 25% | 403.000000 |
| 50% | 1046.000000 |
| 75% | 1586.000000 |
| max | 13213.000000 |

Distribution : sugars_100g

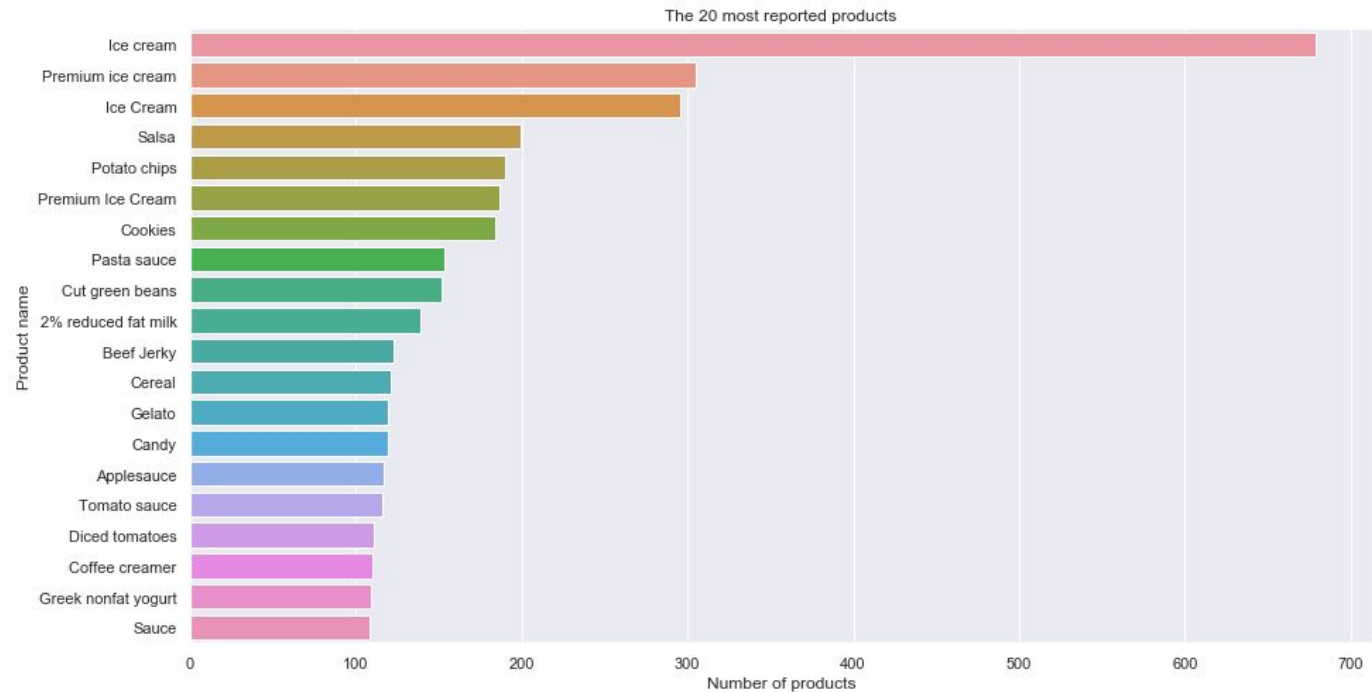


| sugars_100g | |
|-------------|---------------|
| count | 213996.000000 |
| mean | 14.329648 |
| std | 19.087392 |
| min | 0.000000 |
| 25% | 1.300000 |
| 50% | 5.000000 |
| 75% | 21.840000 |
| max | 100.000000 |

Analyse exploratoire des données

Analyse univariée

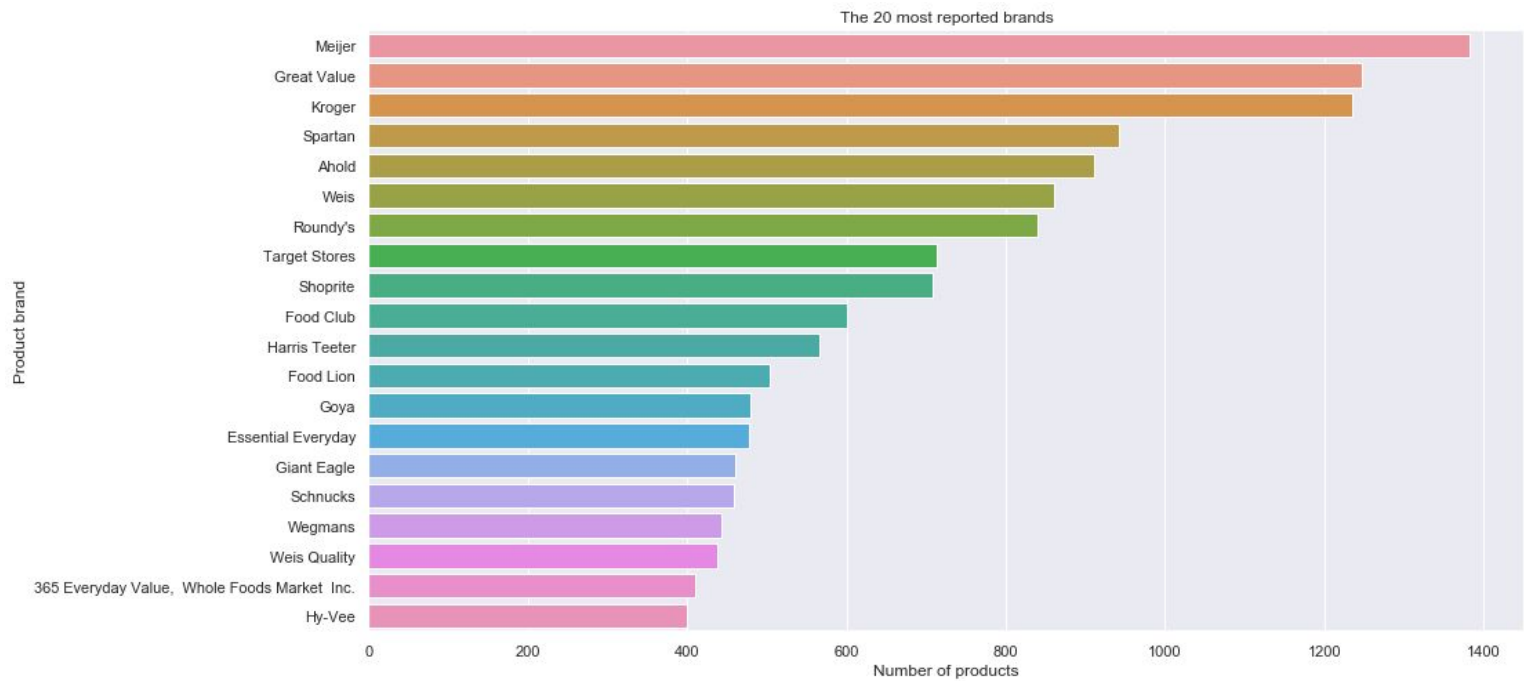
Produits les plus signalés par nom de produit



Analyse exploratoire des données

Analyse univariée

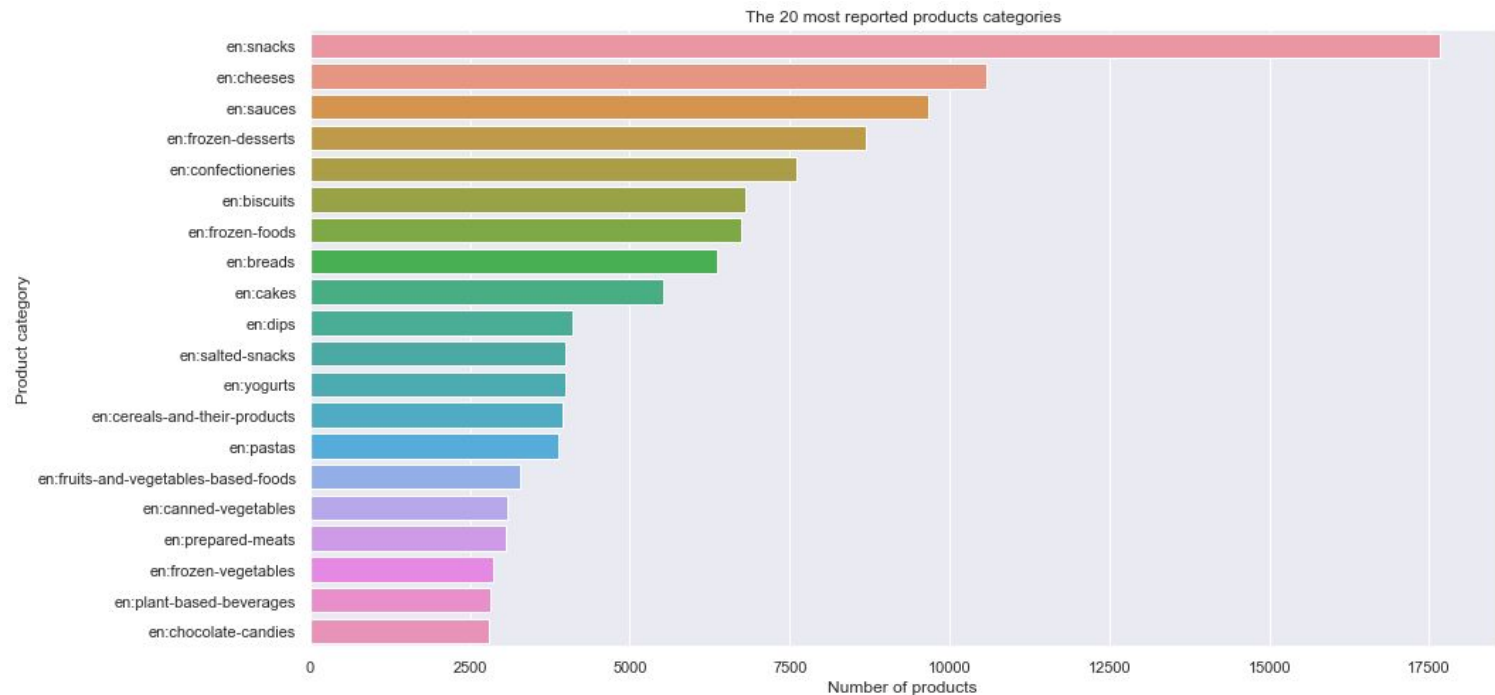
Marques de produits les plus signalées



Analyse exploratoire des données

Analyse univariée

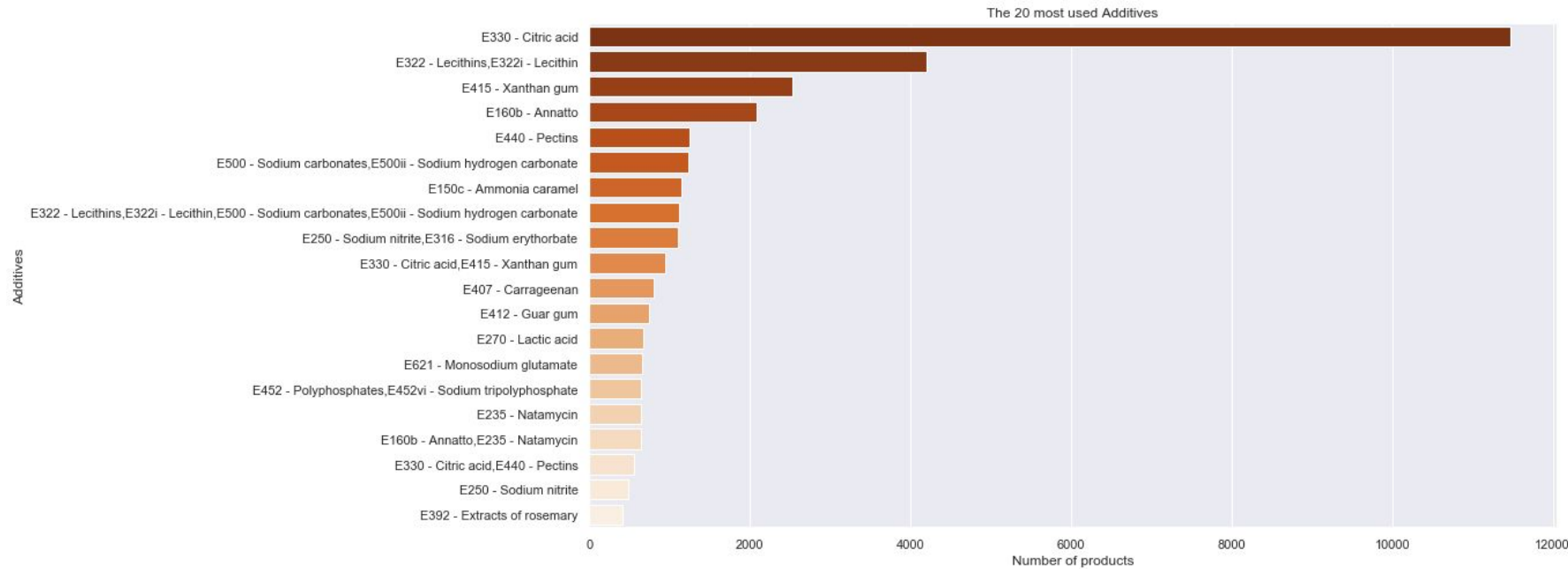
Principales catégories de produits les plus signalées



Analyse exploratoire des données

Analyse univariée

Les additifs les plus signalés

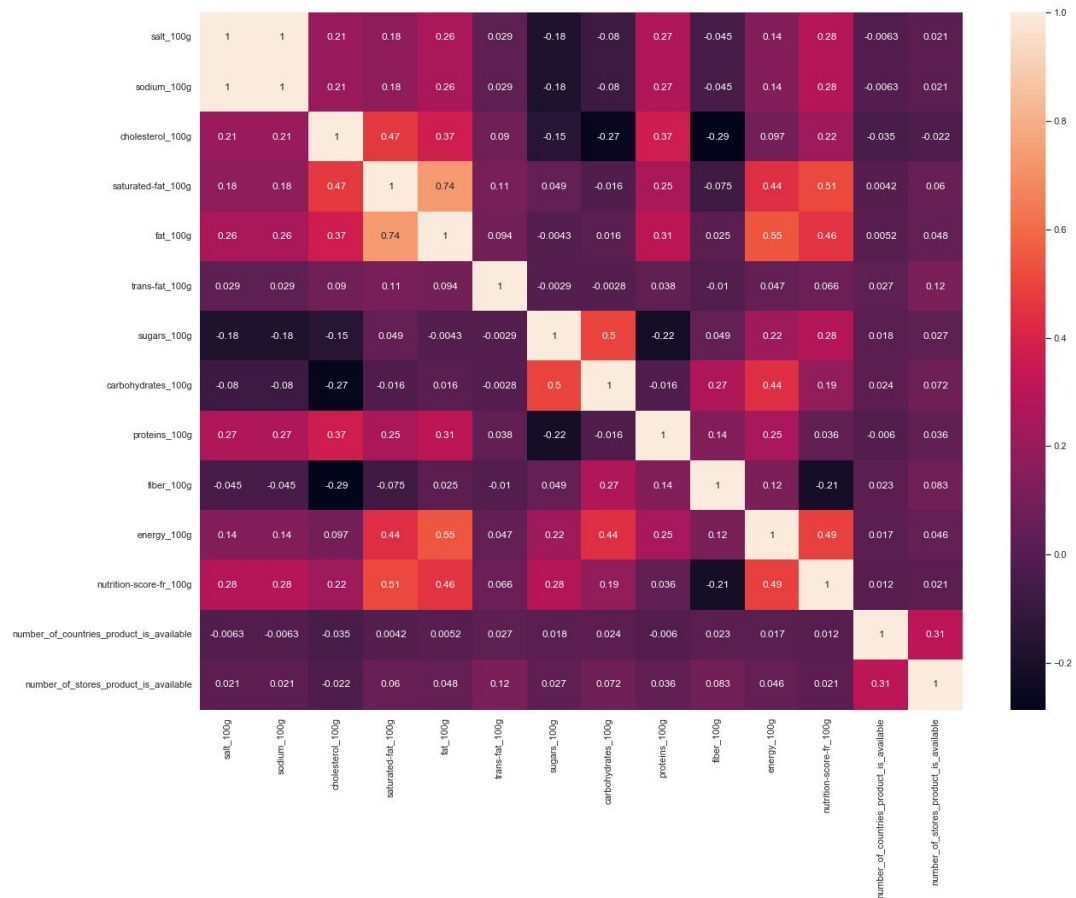


Analyse exploratoire des données

Analyse multivariée et bivariée

Matrice de corrélation

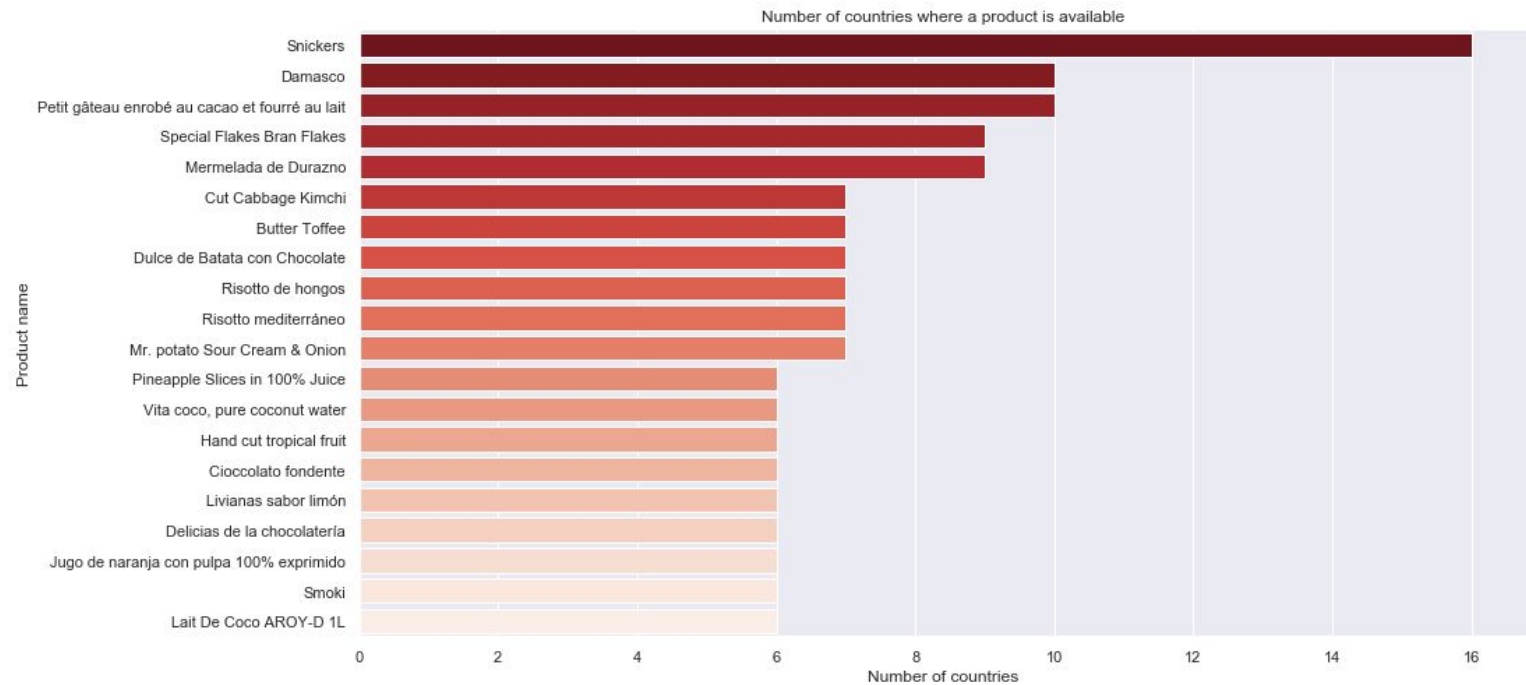
Méthode de Kendall : le coefficient de corrélation mesure la relation monotone entre deux variables. Il n'est pas nécessaire que les variables soient normalement distribuées



Analyse exploratoire des données

Analyse multivariée et bivariée

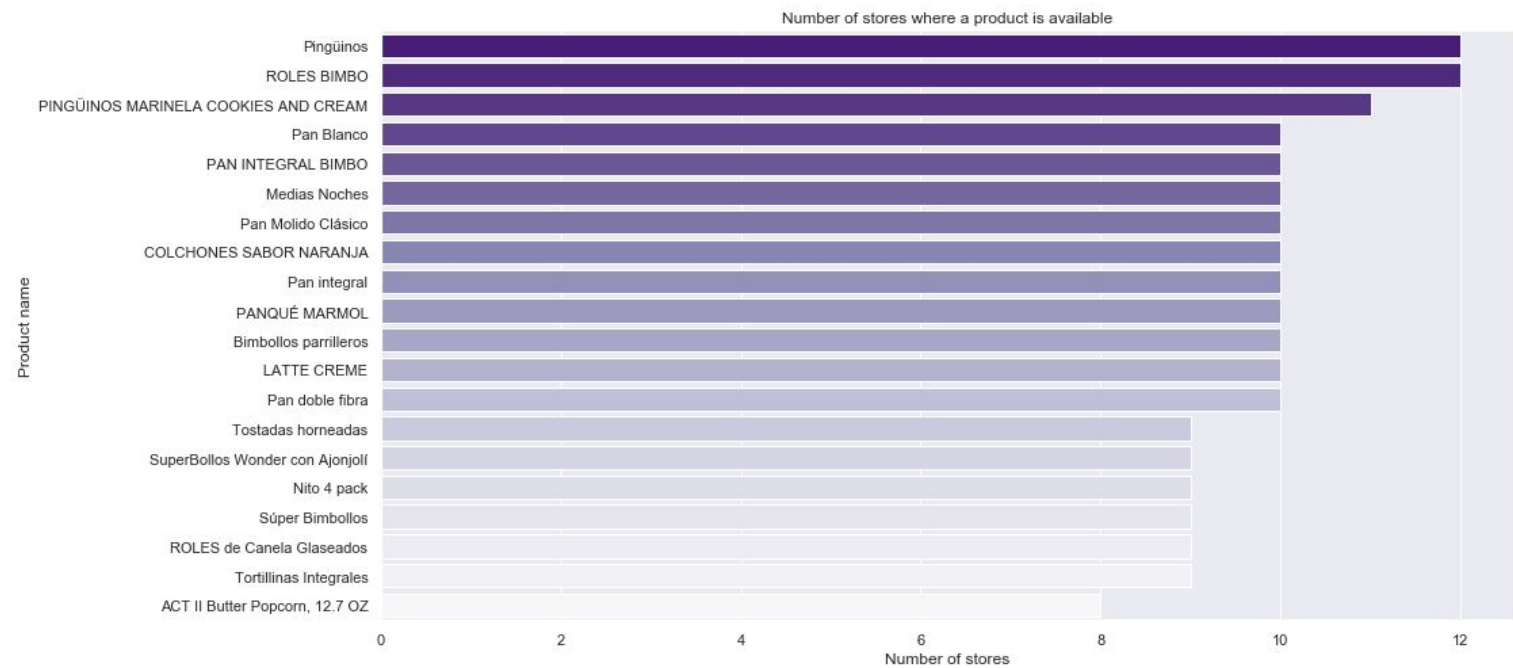
Disponibilité des produits : le top 20 du nombre de pays où un produit est disponible



Analyse exploratoire des données

Analyse multivariée et bivariée

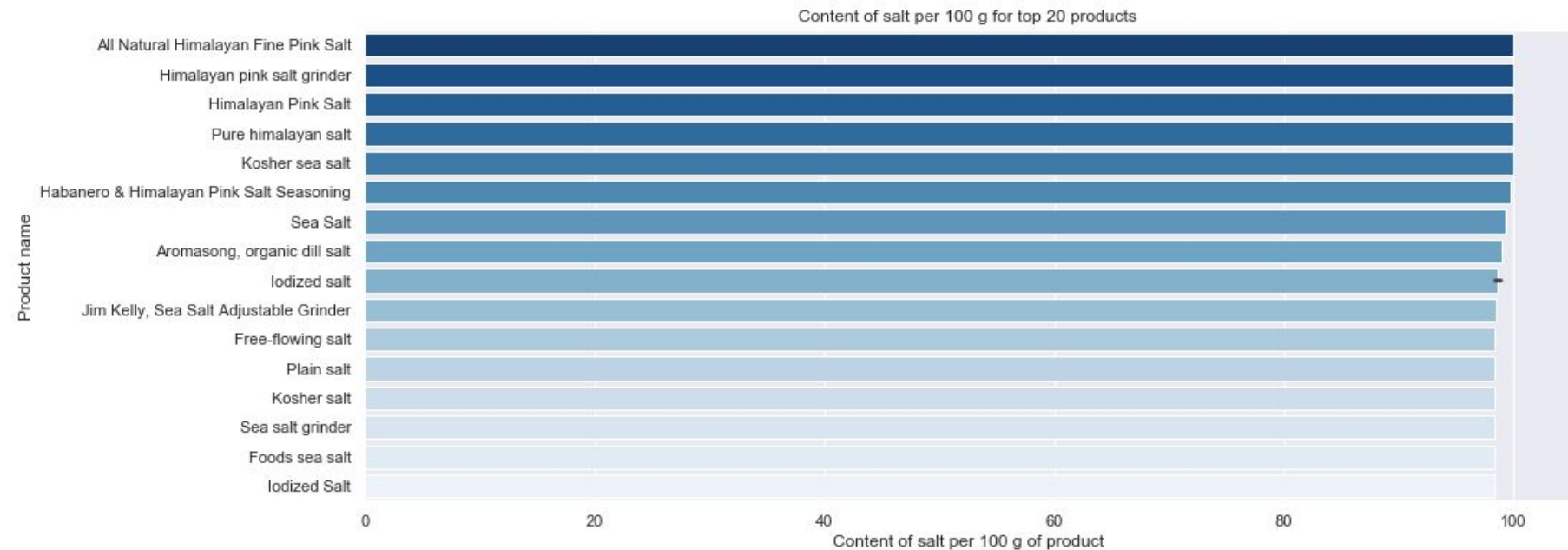
Disponibilité des produits : le top 20 du nombre de magasins où un produit est disponible



Analyse exploratoire des données

Analyse multivariée et bivariée

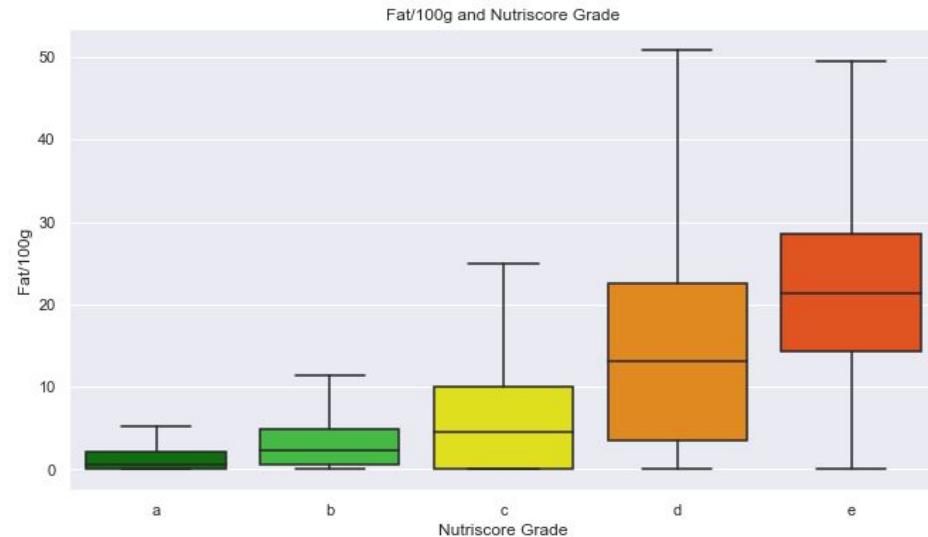
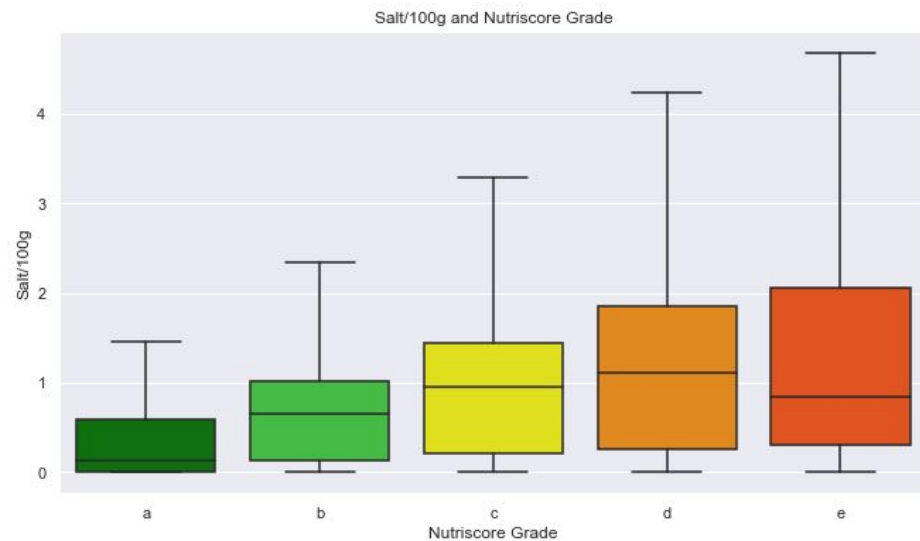
Top 20 des produits les plus signalés avec leur teneur en sel



Analyse exploratoire des données

Analyse multivariée et bivariée

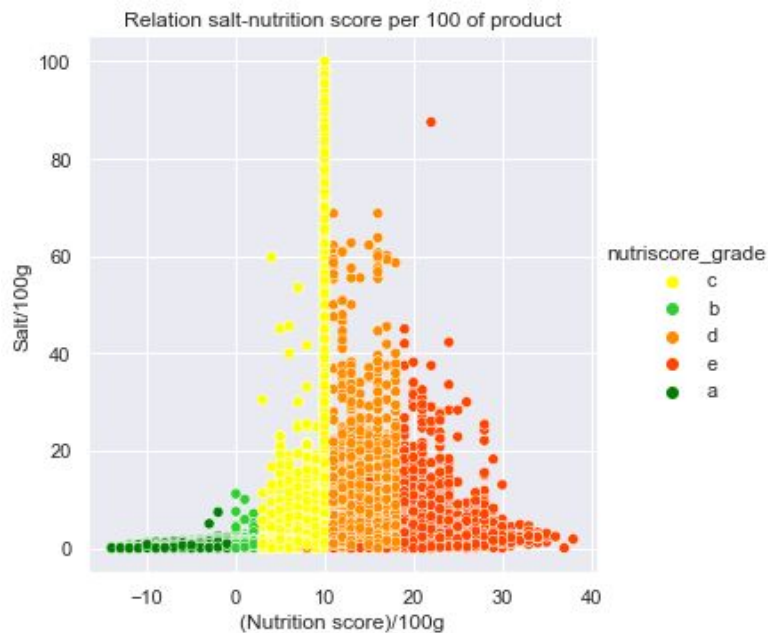
Nutriscore_grade et sel et graisses pour 100g



Analyse exploratoire des données

Analyse multivariée et bivariée

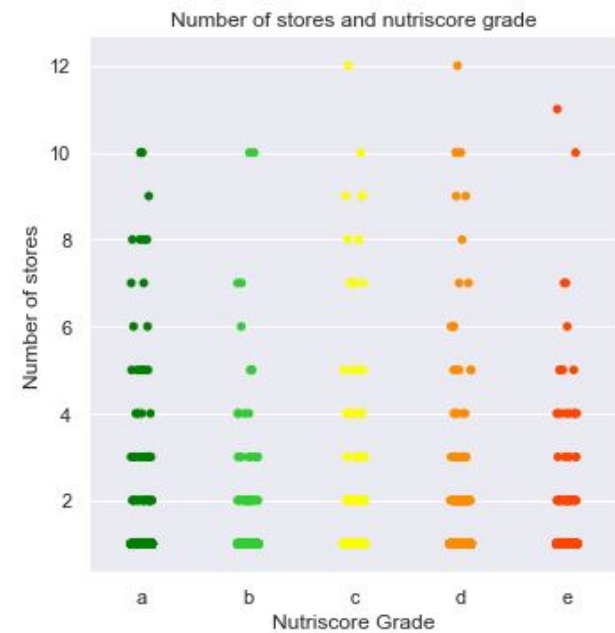
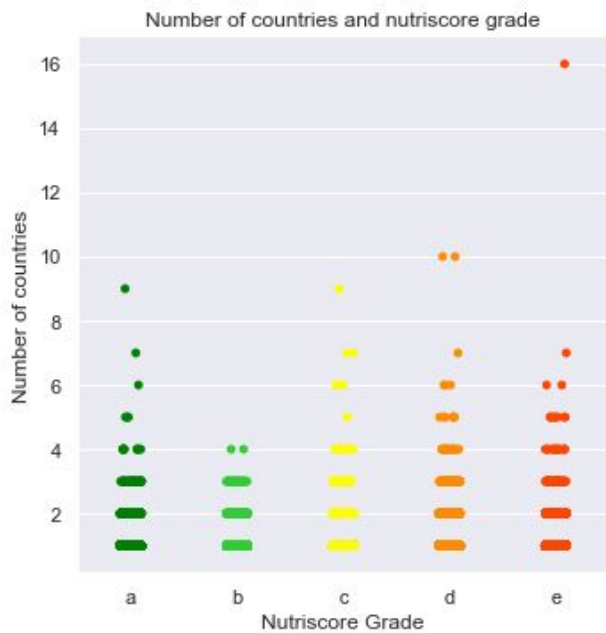
Nutriscore_grade et sel et graisses pour 100g



Analyse exploratoire des données

Analyse multivariée et bivariate

Nutriscore_grade et pays et magasins



Analyse exploratoire des données

Analyse multivariée et bivariée

Analyse ACP

Nous avons d'abord remplacé les valeurs NaN des variables numériques par 0 et nous n'avons conservé que le sous-ensemble de variables suivant pour l'analyse ACP

| | saturated-fat_100g | fat_100g | carbohydrates_100g | energy_100g | sugars_100g |
|-------|--------------------|---------------|--------------------|---------------|---------------|
| count | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 | 213996.000000 |
| mean | 4.075873 | 10.715881 | 32.021267 | 1052.934768 | 14.329648 |
| std | 6.299592 | 13.521522 | 27.762173 | 688.246798 | 19.087392 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 0.000000 | 0.220000 | 7.140000 | 402.000000 | 1.300000 |
| 50% | 1.250000 | 5.360000 | 22.350000 | 1046.000000 | 5.000000 |
| 75% | 6.160000 | 17.020000 | 56.250000 | 1582.000000 | 21.840000 |
| max | 100.000000 | 100.000000 | 100.000000 | 13213.000000 | 100.000000 |

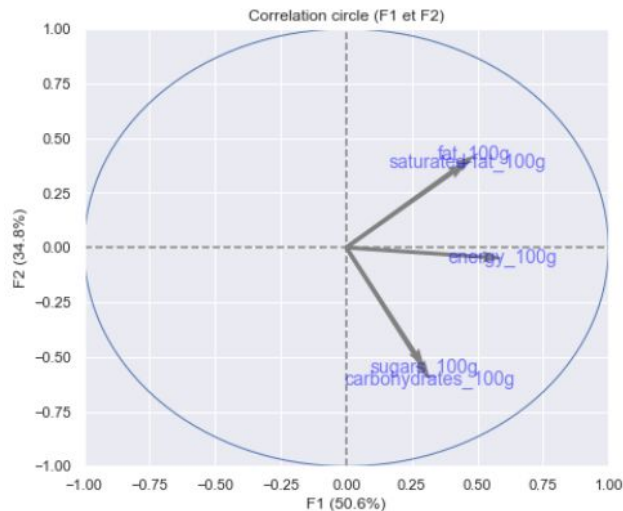
Analyse exploratoire des données

Analyse multivariée et bivariée

Analyse ACP

→ Aucun cluster n'a été observé, mais une certaine tendance dans F1 est observée

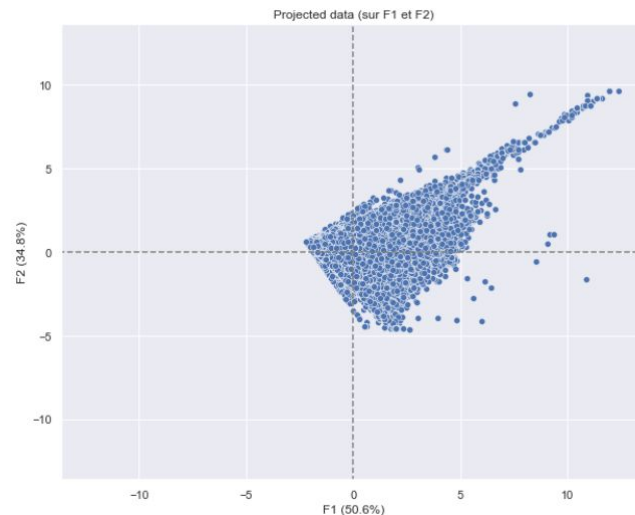
→ La première composante seule capture 50.64% de la variabilité de l'ensemble de données et la deuxième composante seule capture 34.83% de la variabilité de l'ensemble de données



```
print(pca.explained_variance_ratio_ * 100)
```

[144] ✓ 0.5s

... [50.64387975 34.83644335 9.31494433 4.53449842 0.67023416]



Analyse exploratoire des données

Analyse multivariée et bivariable

Analyse ACP

- Aucun cluster n'a été observé, mais une certaine tendance dans F1 est observée. Selon la classification des couleurs nutriscore, F1 a tendance à séparer les données, c'est-à-dire que plus la valeur de F1 est grande, plus le nutriscore est grand, et plus la valeur de F1 est petite, plus le nutriscore est petit

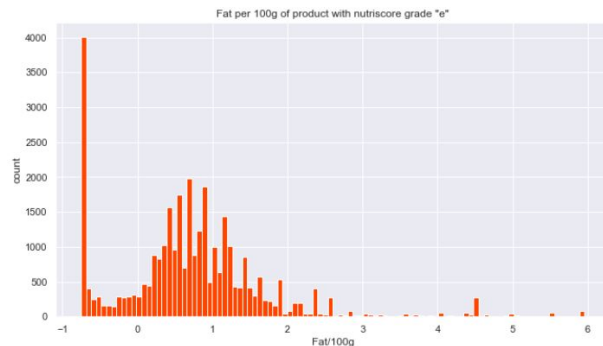
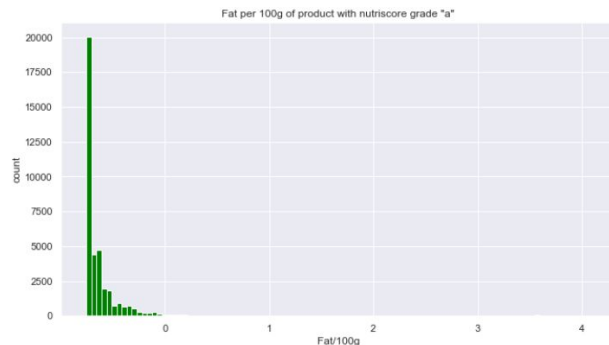


Analyse exploratoire des données

Analyse multivariée et bivariée

Analyse Anova

Distributions de “Fat par 100g” de produit pour les grades "a" et "e" de nutriscore



- L'analyse de variance à sens unique (ANOVA) est utilisée pour déterminer s'il existe des différences statistiquement significatives entre les moyennes de trois groupes indépendants (non liés) ou plus
- La différence entre une statistique d'échantillon et une valeur hypothétique est statistiquement significative si un test d'hypothèse indique qu'elle est trop improbable pour être le fruit du hasard
- En quoi cela est-il pertinent dans les données d'Open Food Facts ? L'utilité de l'ANOVA est de voir si H_0 est vrai : La teneur en graisses est égale entre les produits des groupes "a" et "e" des classifications nutriscore - la valeur p confirmera ou écartera H_0

<https://www.technologynetworks.com/informatics/articles/one-way-vs-two-way-anova-definition-differences-assumptions-and-hypotheses-306553>

<https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/anova/how-to/one-way-anova/interpret-the-results/key-results/>

Analyse exploratoire des données

Analyse multivariée et bivariée

Analyse Anova

Fat per 100g of product for the nutriscore grades "a" and "e"

Valeur $P \leq$ "significance level $\alpha = 0.05$ " : Les différences entre certaines des moyennes sont statistiquement significatives

Valeur $P >$ "significance level $\alpha = 0.05$ " : Les différences entre les moyennes ne sont pas statistiquement significatives

pingouin.anova

`pingouin. anova (data=None, dv=None, between=None, ss_type=2, detailed=False, effsize='np2')`

One-way and N-way ANOVA.

Returns: `aov` : *pandas.DataFrame*

ANOVA summary:

- 'Source' : Factor names
- 'ss' : Sums of squares
- 'DF' : Degrees of freedom
- 'MS' : Mean squares
- 'F' : F-values
- 'p-unc' : uncorrected p-values
- 'np2' : Partial eta-square effect sizes

```
[139] ✓ 0.1s
```

```
...
=====
ANOVA SUMMARY
=====
```

| Source | SS | DF | MS | F | p-unc | np2 |
|------------------|-----------|-------|-----------|-----------|-------|-------|
| nutriscore_grade | 31506.803 | 1 | 31506.803 | 56410.645 | 0.000 | 0.441 |
| Within | 39859.197 | 71365 | 0.559 | nan | nan | nan |

Quelques conclusions

- Les données proviennent d'un projet de collaboration avec des utilisateurs du monde entier
- Une idée pour corriger les données sur le sodium et le cholestérol serait de trouver tous ces produits dans l'image d'information nutritionnelle et d'appliquer un algorithme de reconnaissance d'image pour obtenir les bonnes valeurs de sodium et de cholestérol



Ingredients: Wild Caught Sardines Fish, Extra Virgin Olive Oil, Salt

Net 120g

| NUTRITION INFORMATION | | |
|--|-------------|----------|
| Servings per package: 2 | | |
| Serving size: 33g (1/2 can) Drained Weight | | |
| | Per serving | Per 100g |
| Energy (kcal = 4.2kJ) | 73kcal | 187kcal |
| Protein | 9.1g | 23.4g |
| Total fat | 4.0g | 10.3g |
| - saturated fat | 1.1g | 2.9g |
| - Omega 3 | 0.7g | 1.9g |
| - trans fat | 0.0g | 0.1g |
| Cholesterol | 23mg | 58mg |
| Carbohydrate | 0.0g | 0.1g |
| Dietary fibre | 0.5g | 1.2g |
| Sodium | 100mg | 257mg |
| Calcium | 127mg | 325mg |

- Il est dangereux de développer une application pour les utilisateurs qui suivent un régime low sodium et dont les valeurs du contenu de ces produits sont erronées, car cela pourrait mettre leur vie en danger
- Nous avons trouvé des indicateurs et des variables intéressants qui ont été classés par lieux de vente ou par catégories
- Afin d'effectuer des analyses multivariées telles que l'ACP ou l'ANOVA, nous pouvons aussi classer les produits par catégories, comme les snacks, les biscuits, les boissons, etc. pour voir s'il existe des regroupements