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

Sommaire

I. Idée d'application

- Application pour aider les entreprises à proposer de meilleurs produits pour leurs consommateurs
- Basé sur l'amélioration du nutriscore
- Pistes pour rendre un produit plus 'sain' (sur quelles variables agir pour avoir un meilleur nutriscore)
- Produit déjà existant ou nouveau
- Produits Français

I. Idée d'application

Démarche pour évaluer la faisabilité :

→ Identifier dans le jeu de données quelles sont les variables associées au nutriscore

→ Filtrer le jeu de données pour se focaliser sur ces variables

→ Analyser les variables et essayer de comprendre leur comportement vis à vis du nutriscore

→ Si cela est envisageable, proposer sur quelles variables agir en priorité pour améliorer le nutriscore

→ Faire des tests avec le jeu de données

II. Nettoyage des données

- Base de données CSV disponible sur le site d'openfoodfacts :

  https://static.openfoodfacts.org/data/en.openfoodfacts.org.products.csv

- Gros fichier > 3 Giga Octets
- 1751241lignes et 184 colonnes
- Nettoyage en plusieurs étapes

II. Nettoyage des données

1ère étape :

- Nettoyage des produits dupliqués (même bar-code)
  - 248 produits
- Récupération des produits uniquement disponibles en France
  - Filtrer par la colonne 'countries'
- Récupération des colonnes utiles (28)
  - Qualitatives:

code	product_name	brands	categories	labels	main_category	pnns_groups_1	pnns_groups_2	ingredients_text	serving_size	additives_en	nutriscore_grade

  - Quantitatives:

additives_n	energy_100g	energy-kj_100g	energy-kcal_100g	sugars_100g	saturated-fat_100g	fat_100g	salt_100g	fiber_100g	proteins_100g	carbohydrates_100g	nutrition-score-fr_100g	nutriscore_score	nova_group	carbon-footprint_100g	fruits-vegetables-nuts_100g

- Enregistrement d'un premier fichier

II. Nettoyage des données

2ème étape :

- Nettoyage des produits non-alimentaires
  - Filtrage par la colonne 'categories'
- Supprimer les produits qui ne contiennent pas de valeurs dans nos variables quantitatives
- Pour la partie énergie:

  - Identification des 'outliers' puis on supprime les valeurs aberrantes avec des connaissances métier (>4500kJ)

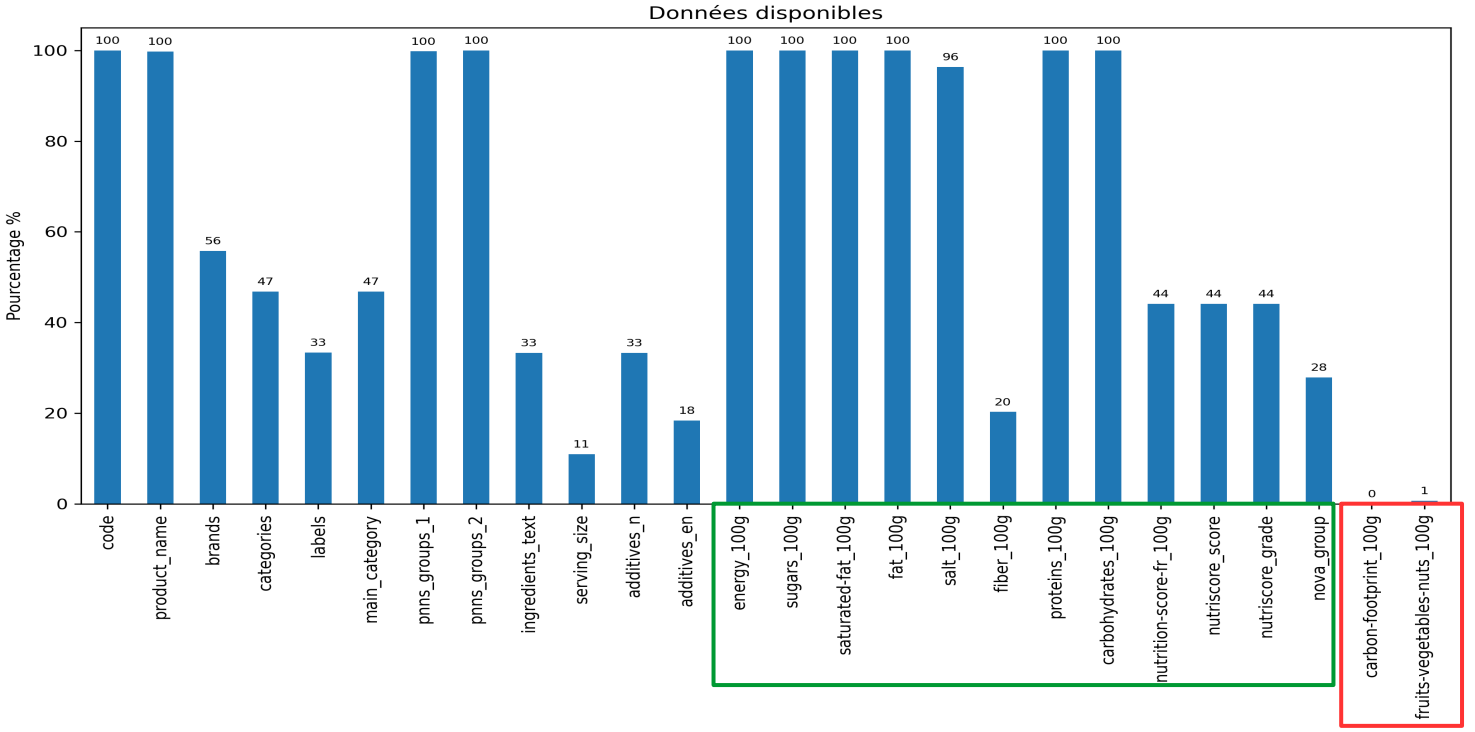
  - On ne garde qu'une seule colonne (energy_100g convertie en kcal)
- Pour la partie nutriments :

  - Produits avec la somme des nutriments principaux > 100g supprimés

  - Identification des 'outliers' puis on supprime les valeurs > 100g
- Enregistrement fichier final nettoyé

II. Nettoyage des données

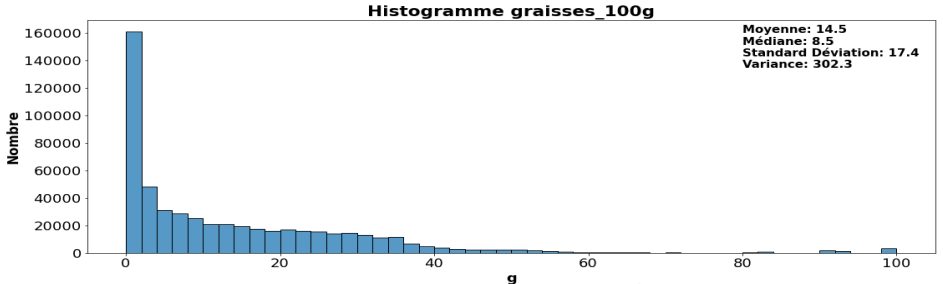
Au final :

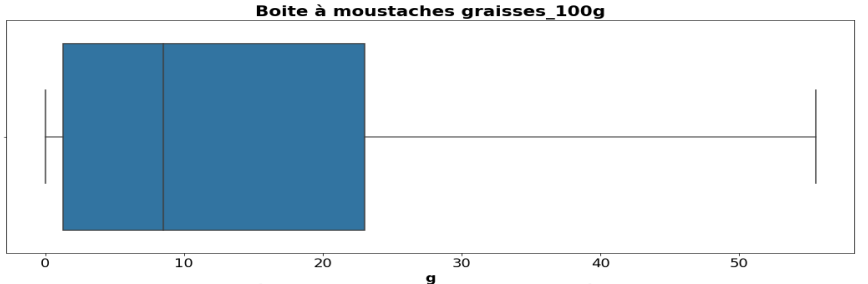


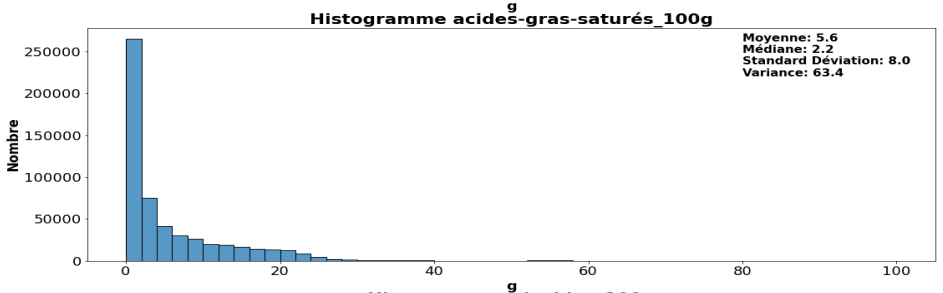
- 556359 lignes et 26 colonnes
- Trop peu de données pour 'carbonfootprint' et 'fruits-vegetables-nuts'
- Pas d'imputations pour les informations quantitatives manquantes
- Certaines erreurs sont toujours présentes (ex : kJ rentré en kcal)

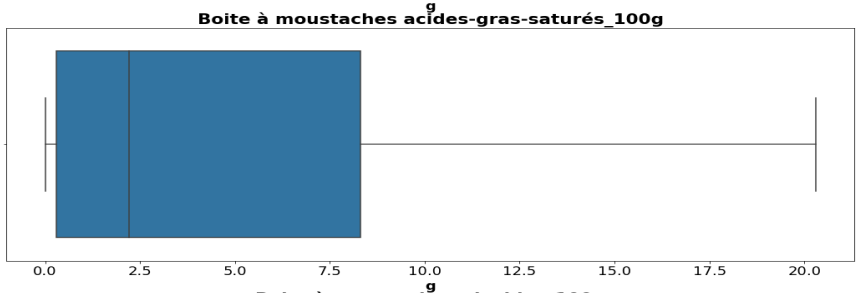
III. Analyse Exploratoire des données

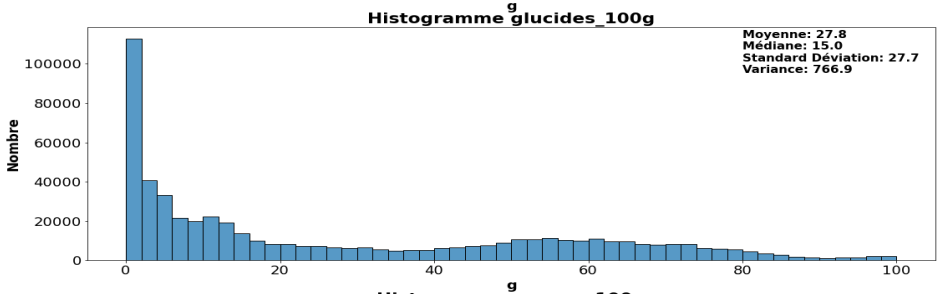
- Analyse univariée nutriments

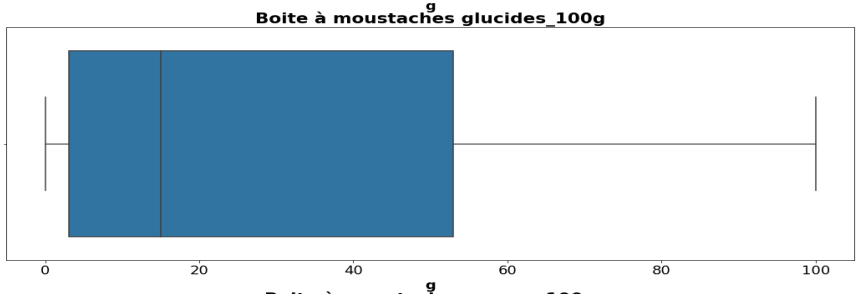


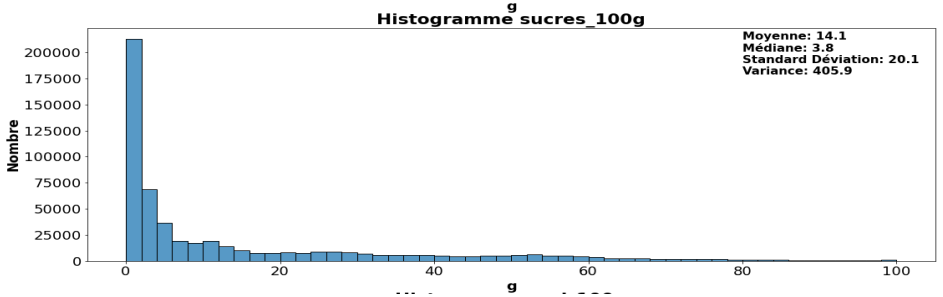


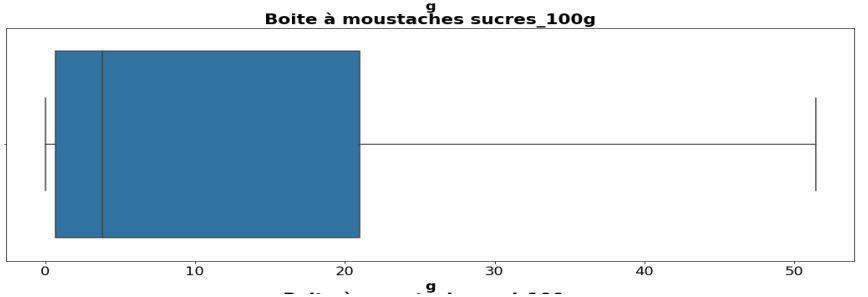






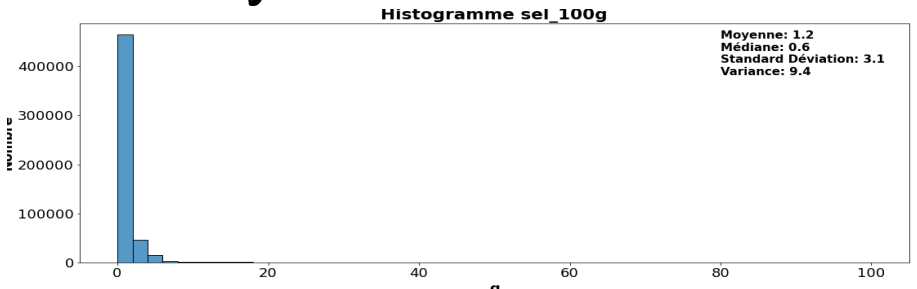


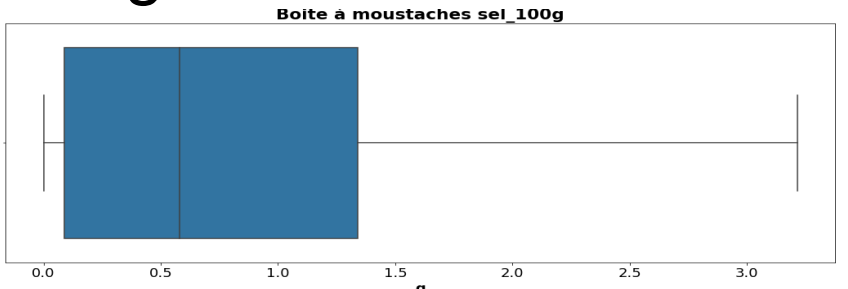


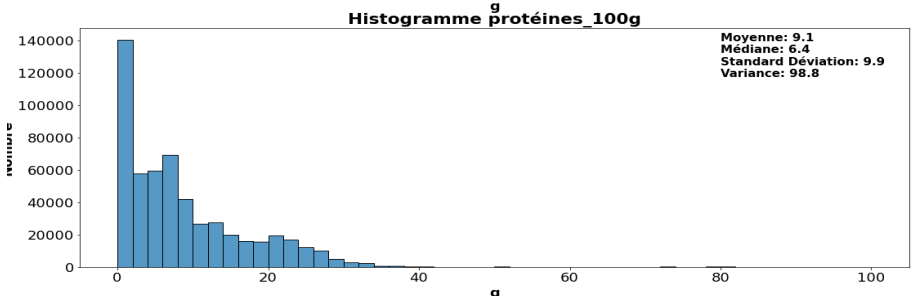


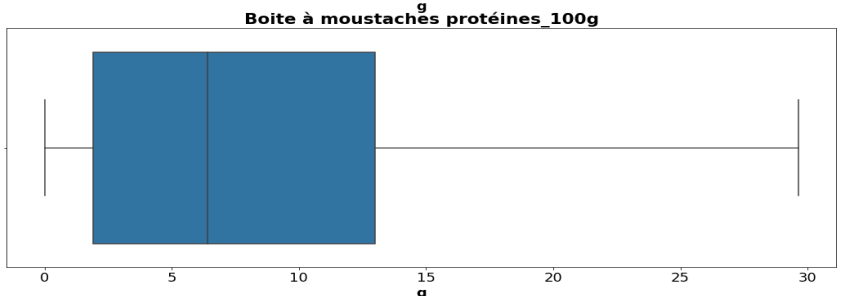
III. Analyse Exploratoire des données

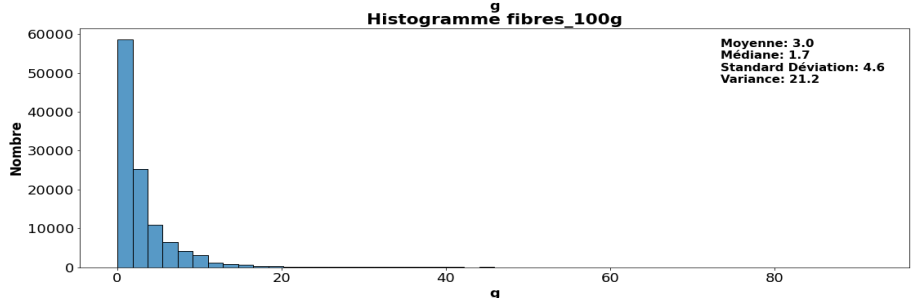
- Analyse univariée nutriments / énergie

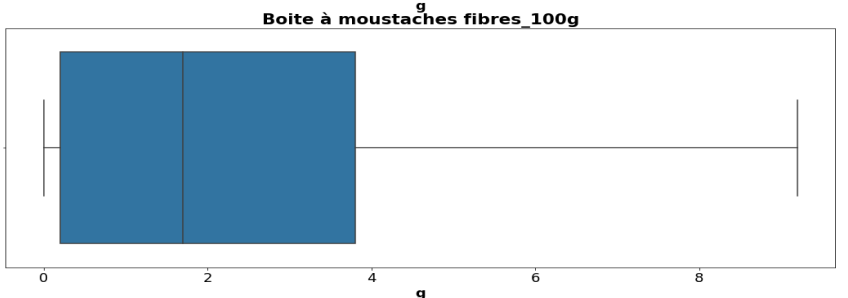


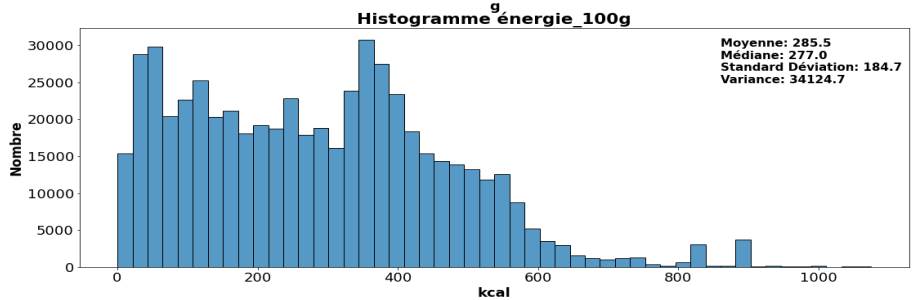


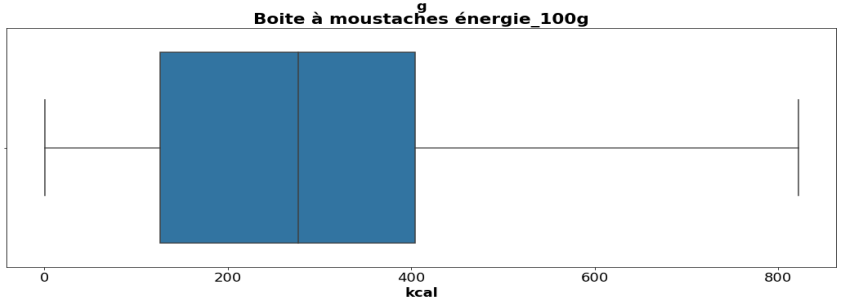






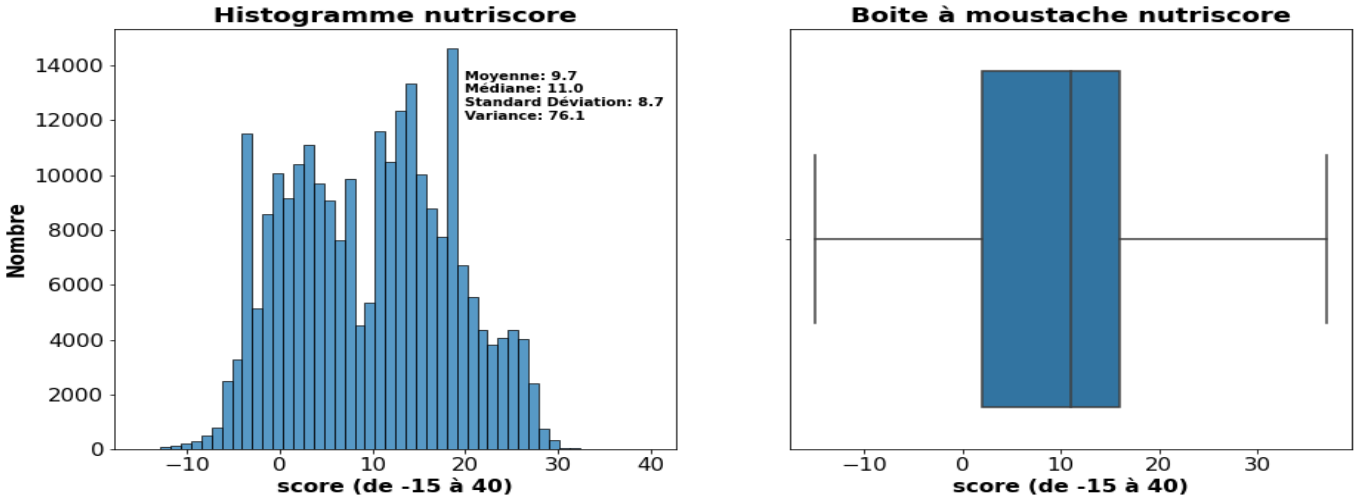




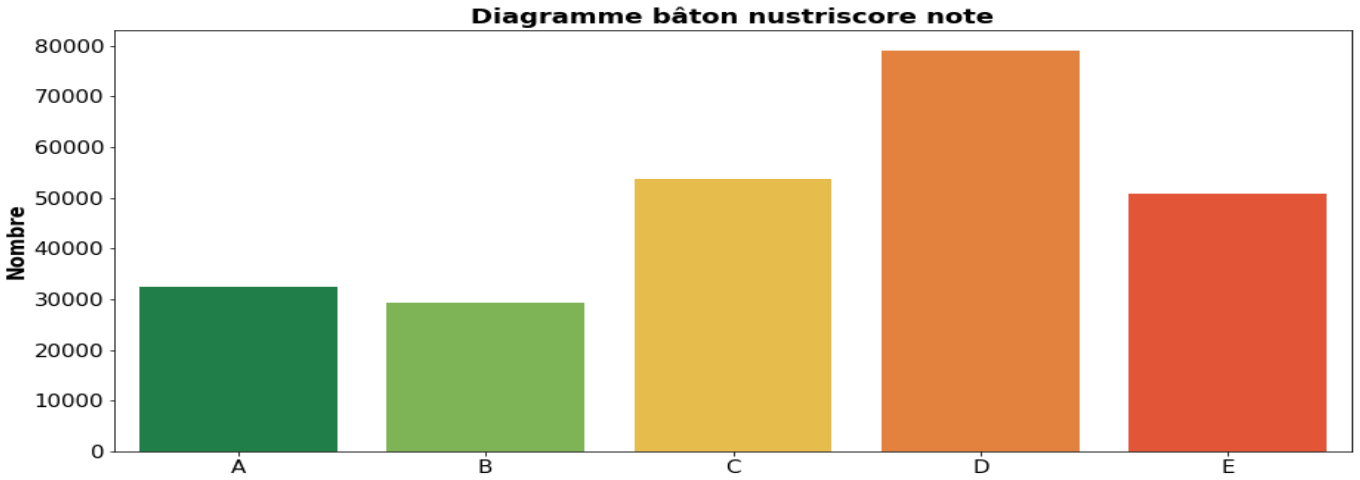


III. Analyse Exploratoire des données

- Analyse univariée Nutriscore



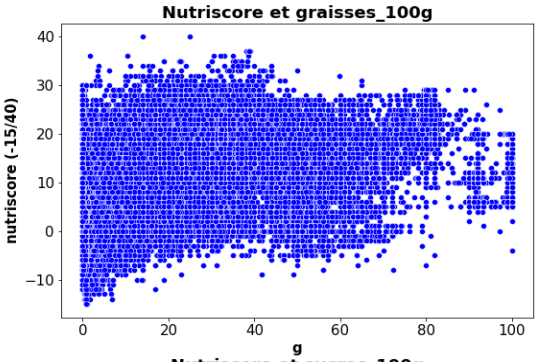
Nutriscore-score[-15, 40]
→ nutriscore grade [A,D]

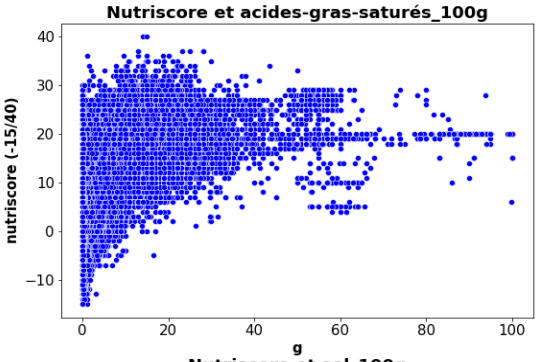


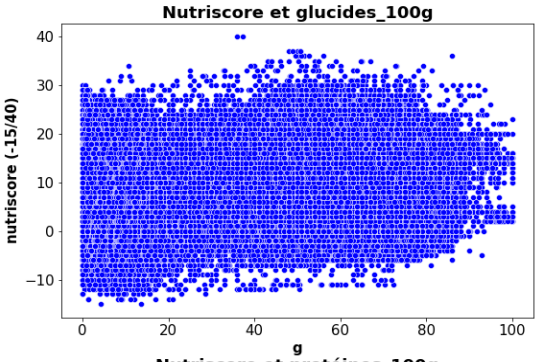
<u>Rappel:</u>
A [-15, -1]
B [0, 2]
C [3, 10]
D [11, 18]
E [19, 40]

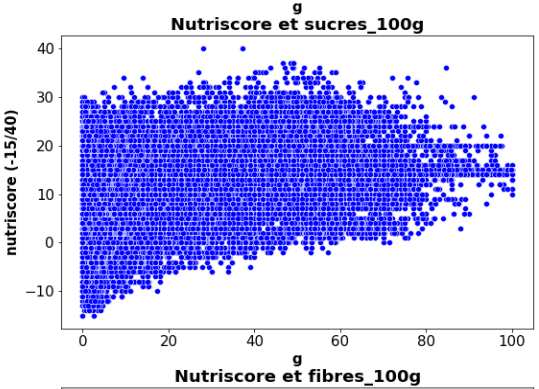
III. Analyse Exploratoire des données

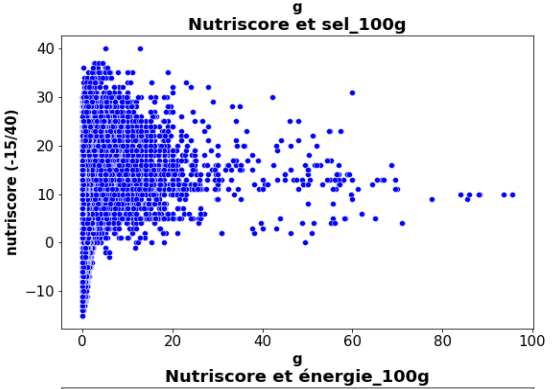
- Analyse bivarié nutriments/énergie → Nutriscore

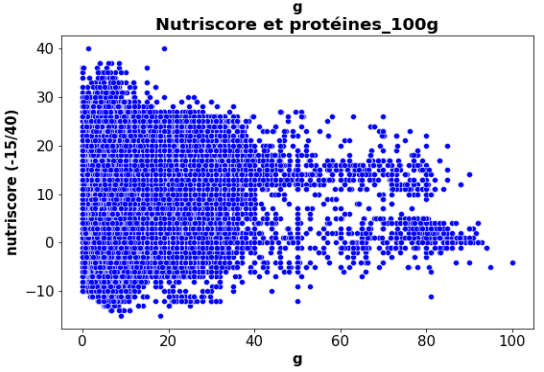


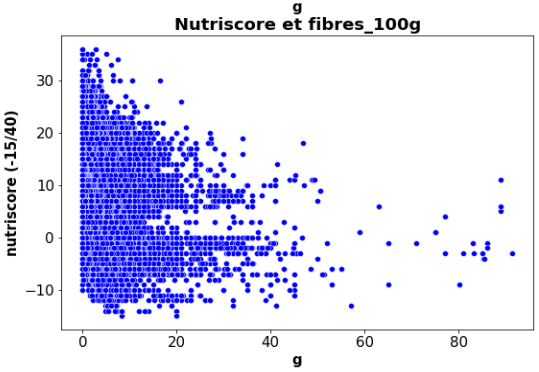


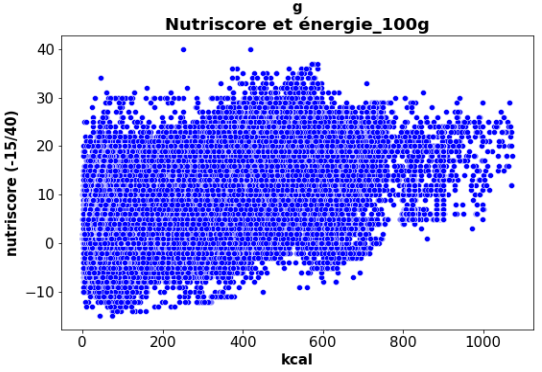








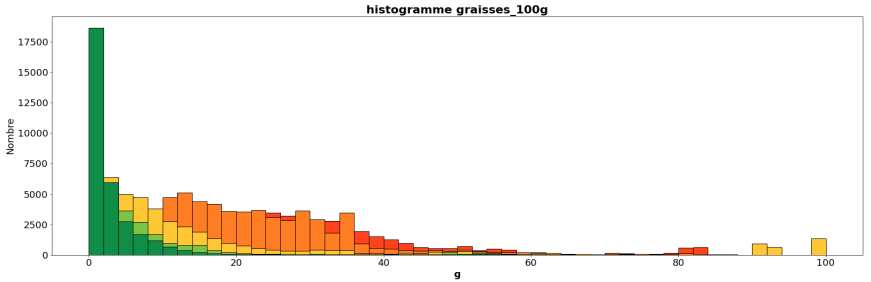


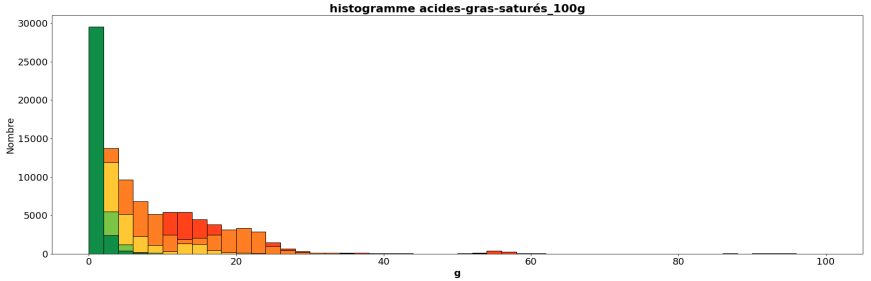


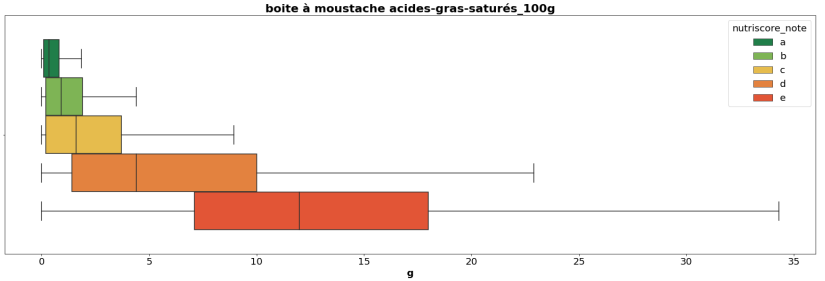
→ Difficile d'interpréter des tendances

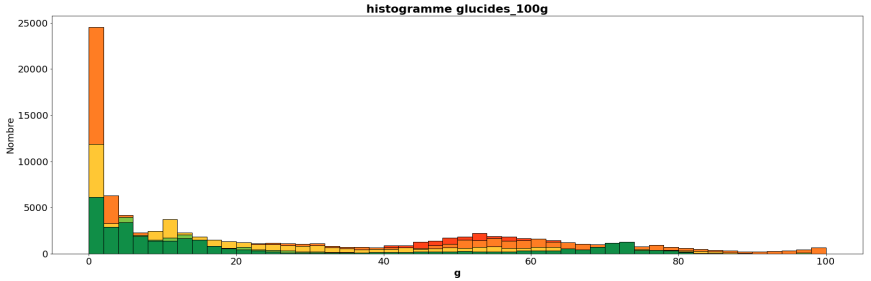
III. Analyse Exploratoire des données

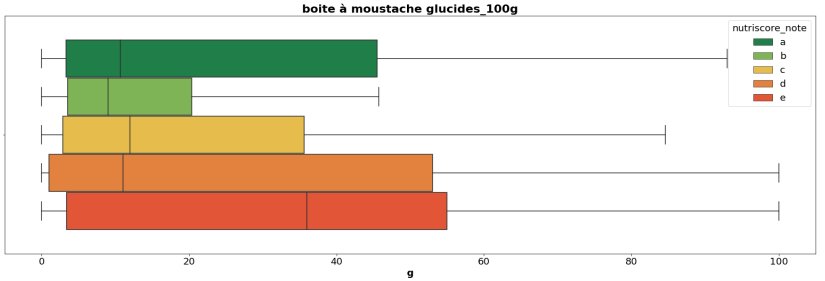
- Analyse bivariée nutriments → Nutriscore

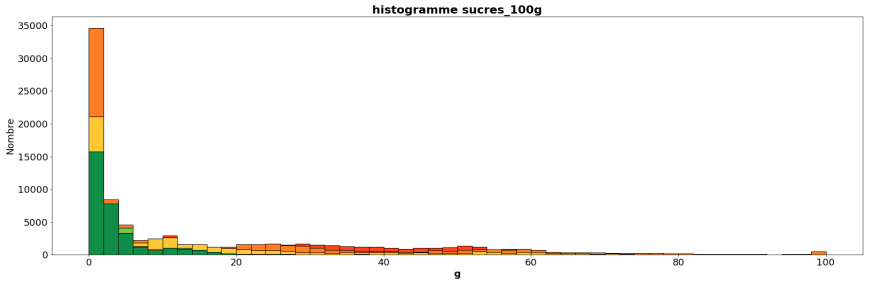


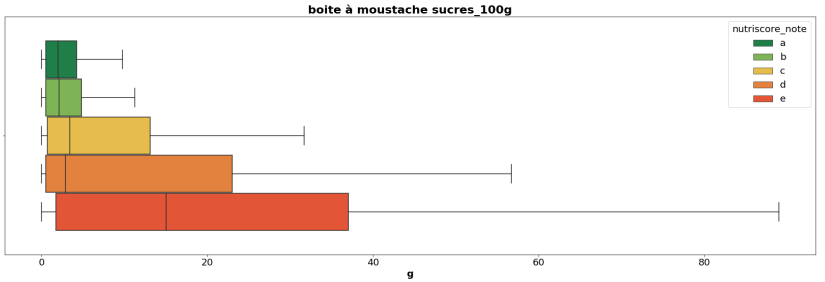






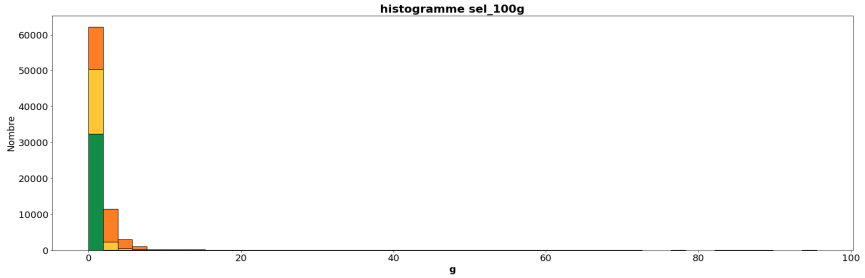


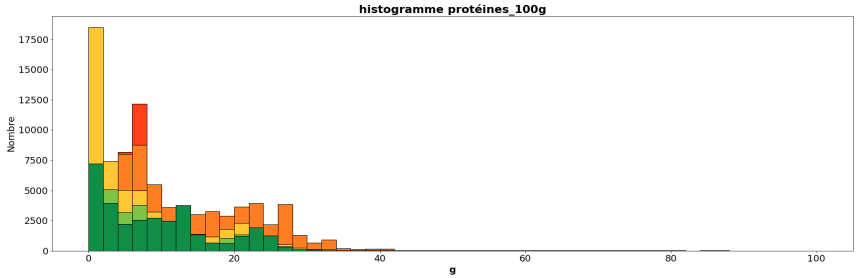


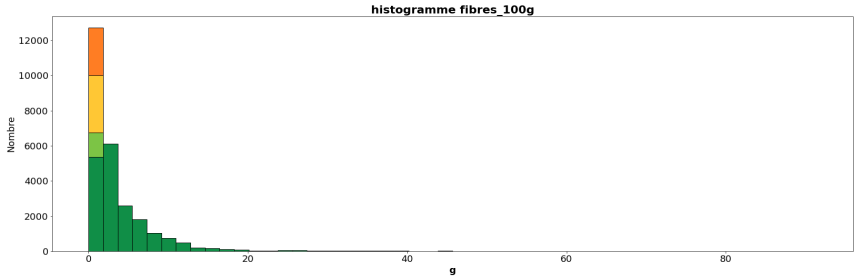


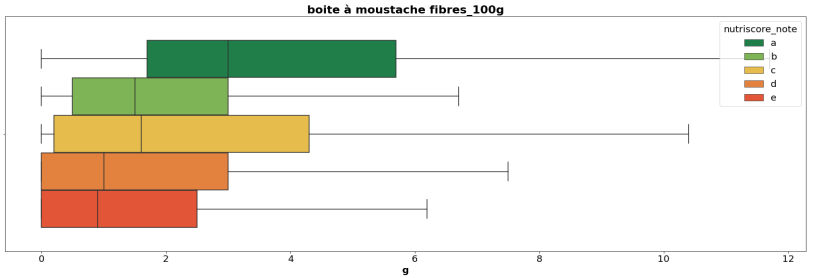
III. Analyse Exploratoire des données

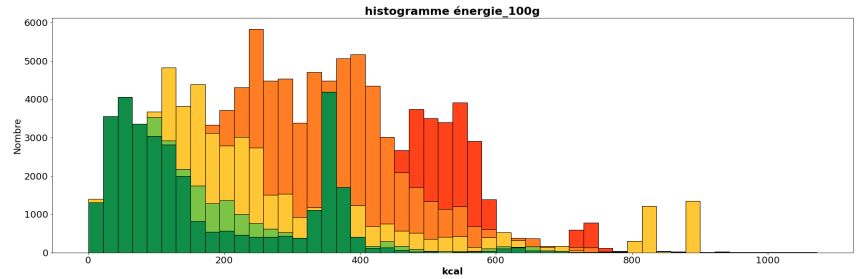
- Analyse bivariée nutriments / energie → Nutriscore

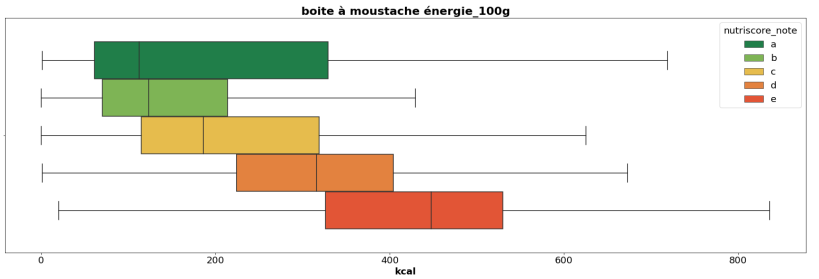










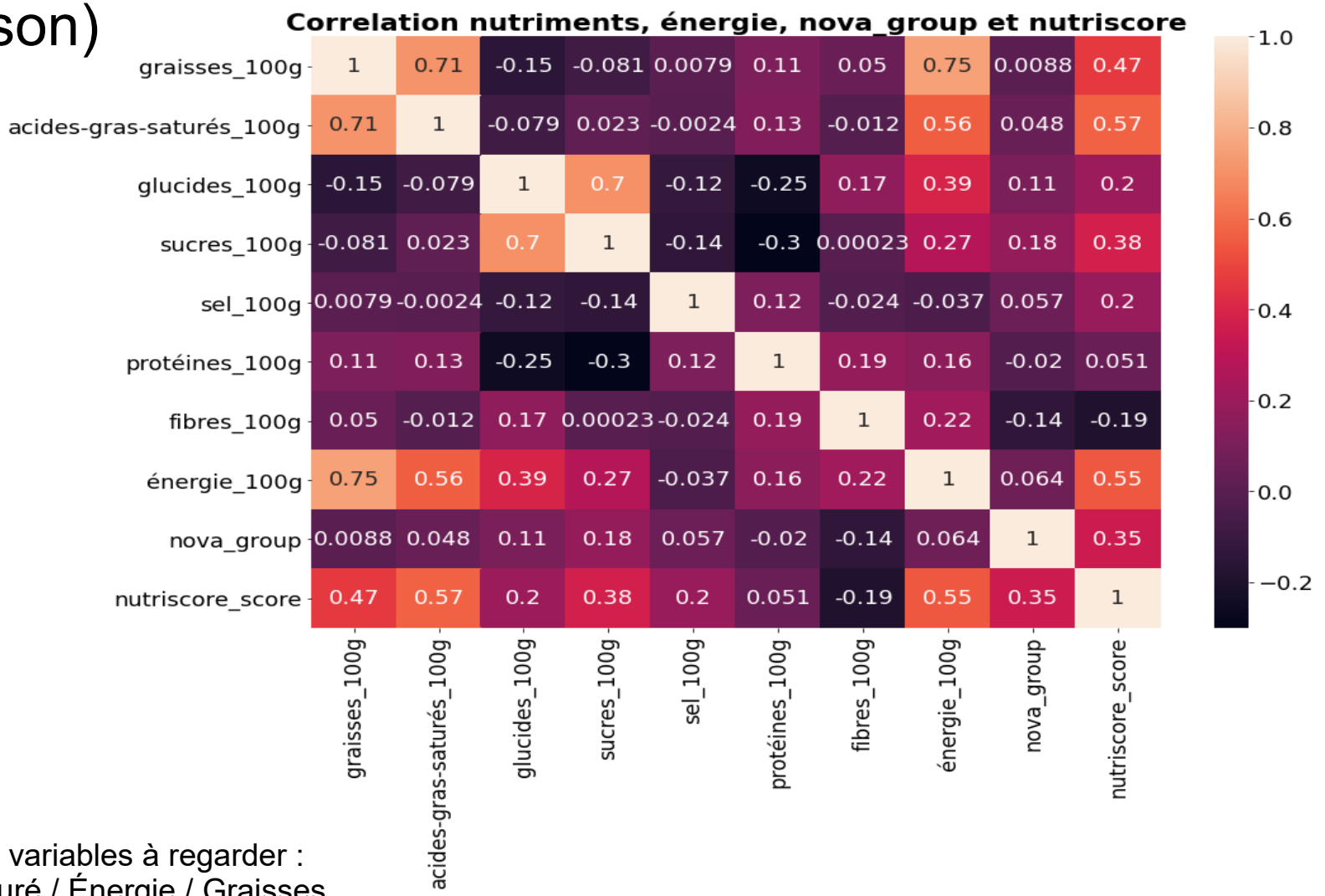


III. Analyse Exploratoire des données

- Analyse bivariée Nutriscore / Nova_Group

III. Analyse Exploratoire des données

- Corrélation linéaire variables quantitatives (coefficients de Pearson)



→ Proposition de variables à regarder :
1) Acide-gras-saturé / Énergie / Graisses
2) Sucres / Sel / Fibres

III. Analyse Exploratoire des données

- Test Statistique : Niveau de significativité des coefficients de corrélation de Pearson obtenus (p_value)

  Hypothèse 0 : 'La variable choisie et le nutriscrore sont indépendantes'

  Hypothèse 1 : 'La variable choisie et le nutriscore sont dépendantes'

  Résultats :

	acides-gras-saturés_100g	énergie_100g	graisses_100g	sucres_100g	sel_100g	fibres_100g
Coefficient de corrélation linéaire	0.57207	0.546125	0.466386	0.376334	0.196741	-0.190247
p_value	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000

  → p_value < 1%, l'hypothèse 0 peut être rejetée avec une confiance de 99%

  → Il est peut probable que nos coefficients de corrélation soient dus au hasard avec notre jeu de données

  → On peut raisonnablement penser à ces dépendances pour d'autres échantillons de produits (ajout d'entrées dans le jeu de données)

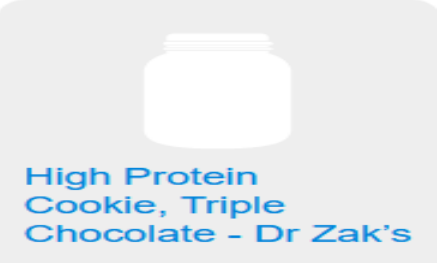
IV. Exemples

- Produit existant

Proposition des meilleurs produits de même catégorie

A

A



B

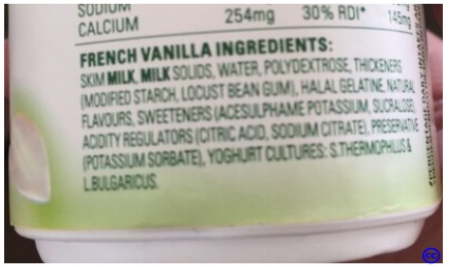
product_name	brands	saturated-fat_100g	fat_100g	energy_100g	sugars_100g	salt_100g	fiber_100g	nutriscore_grade	nova_group
Cioccofrolle	Barilla,Mulino Bianco	2.8	11.5	396.029876	1.4	0.65	7.0	a	4.0
Mini Cookies aux Pepites de Chocolat	Dukan	4.3	16.5	389.098755	1.0	0.10	11.0	a	4.0
High Protein Cookie, Triple Chocolate	Dr Zak's	2.7	10.1	349.902075	2.6	0.74	3.3	b	4.0
Goûter pépites de chocolat	Gerblé	3.3	17.0	458.887967	21.0	0.24	6.0	c	4.0

Analyse des variables à étudier

IV. Exemples

- Produit existant

Proposition des meilleurs produits de même catégorie



product_name	brands	saturated-fat_100g	fat_100g	energy_100g	sugars_100g	salt_100g	fiber_100g	nutriscore_grade	nova_group
Yaourt à la Grecque Vanille	Yoplait	0.1	0.1	37.045643	4.2	0.05	NaN	a	NaN
Yaourt au lait de brebis vanille	Auchan	1.7	2.7	77.915353	8.6	0.12	NaN	a	NaN
Skyr vanille	Arla	0.0	0.0	74.091286	7.6	0.13	NaN	a	4.0
Les 2 vaches Vanille de Madagascar	Les 2 Vaches	2.1	3.0	93.928631	12.5	0.13	NaN	c	3.0

Analyse des variables à étudier

IV. Exemples

- Nouveau produit

On cherche dans la catégorie 'pizza'

Proposition des meilleurs produits dans cette catégorie

product_name	brands	saturated-fat_100g	fat_100g	energy_100g	sugars_100g	salt_100g	fiber_100g	nutriscore_grade	nova_group
Spinaci Ricotta	Mamma Roma	3.9	8.3	196.939419	1.5	0.20	NaN	a	4.0
Pizza XXL fajitas les Américaines	Auchan	1.5	3.9	183.077178	2.9	0.90	2.6	a	4.0
Pizza Primavera - légumes grillés et mozzarella	Auchan	2.4	6.9	219.644813	2.2	0.84	3.1	a	3.0

Analyse des variables à étudier

V. Conclusion

- Identification de variables à modifier pour améliorer le nutriscore d'un produit :

  - Acide-gras-saturés/Graisses/ Énergie

  - Sucres/Sel/Fibres

- Pour aller plus loin :

  - Proposition de calcul du nutriscore intégré dans l'application

  - Proposition d'amélioration via le nova group

- Attention la base de données contient des erreurs de saisie

Merci pour votre attention !

Questions ?