Conception d’ une application au service de la santé publique

OC Data Scientist
Projet 3

Clément Dubois
27/09/2022

Contexte générale

Contexte :

Santé publique France lance un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.

Données :

Jeu de données sur les produits alimentaires du monde entier

Mission :

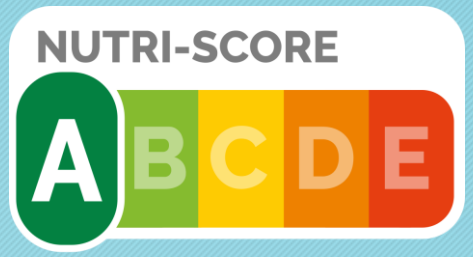
- Proposer une idée d'application pour répondre à l'appel à projet
- Effectuer une analyse exploratoire des données Open Food Facts afin déterminer la pertinence et la faisabilité de l'application.

Sommaire

Partie I : Présentation de l’application

Idée d'application : 1) manger plus sain

1) MANGER MOINS GRAS, MOINS SUCRÉS ET MOINS SALÉS :



Valeurs nutritives pour 100g :

+ Fibres, protéines, fruits légumes, légumineuses, fruits à coques, huile de colza, de noix et d'olive.

- Energie, acides gras saturés, sucres, sel.

2) LIMITER LES ADDITIFS ALIMENTAIRES

Ingrédients d'origine naturel ou de synthèse

- Qualité sanitaire (conservateurs, antioxydants)
- Aspect et Goût (colorants, édulcorants, exhausteurs de goût)
- Texture (épaississants, gélifiants)
- Stabilité (émulsifiants, antiagglomérants, stabilisants).

Qui peuvent être nocifs pour la santé :

- Nitrites et nitrates (E249, E250, E251 et E252) [1]
- Dioxyde de titane (E171) [2]
- Carboxyméthylcellulose (E466) et Polysorbate-80 (E433) (Emulsifiants) [3]

Sources :
1) Évaluation des risques liés à la consommation de nitrates et nitrites. Anses, Rapport d'expertise collective. Juillet 2022
2) Re-evaluation of titanium dioxide (E 171) as a food additive. EFSA Journal, 2016
3) Dietary emulsifiers impact the mouse gut microbiota promoting colitis and metabolic syndrome. Chassaing et al. Nature, 2015

Idée d'application : 2) manger plus responsable

1) PRODUITS D'ORIGINE FRANCE [1] :

- **L'alimentation en France = 140 Mt CO2 eq (2,1 t par personne), soit 22% de l'empreinte carbone totale.**
- **46% de ces émissions sont associées aux importations.**

2) PRODUITS BIO [2] :

- **Préservation de la biodiversité et fertilité des sols (- pesticides de synthèse, utilisation d'engrais vert riche en matière organique)**
- **Préservation de la qualité de l'eau (- de rejets azoté)**
- **Limitation des émissions de GEF (- de rejets, + de séquestration par fixation du carbone dans le sol)**

Sources :
1) Ministère de la transition Ecologique et de la Cohésion des Territoires, DOCUMENT DE TRAVAIL N°59, Juillet 2022.
2) Agence française pour le développement et la promotion de l'agriculture biologique, basée sur le rapport : Sautereau N., Benoit M., 2016. Quantification et chiffrage des externalités de l'agriculture biologique, Rapport d'étude ITAB, 136 p

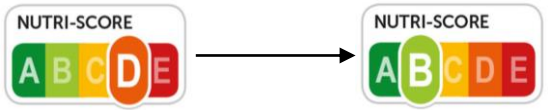
Concept de l’application

Proposition d ’une liste de

produits de la même famille

Manger plus sain

- Nutri-score



- - d’additifs

Manger plus responsable

- Origine France
- Label Biologique

Partie II : Présentation et préparation des données

A) Présentation des données

B) Nettoyage des données

C) Analyse et prédiction des valeurs manquantes

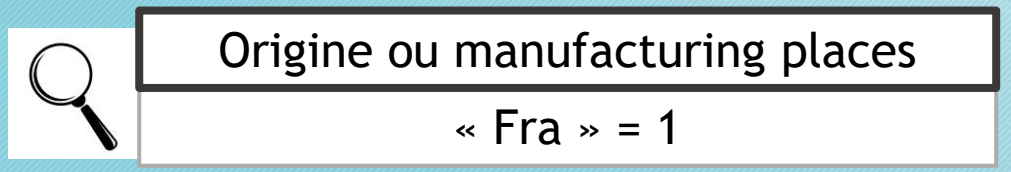
A) Présentation des données et des variables d'intérêts

- **Export du 26/08/2022**
- **2,5 millions de produits**
- **186 variables décrivant les produits (infos générales, tags, ingrédients, valeurs nutritives)**

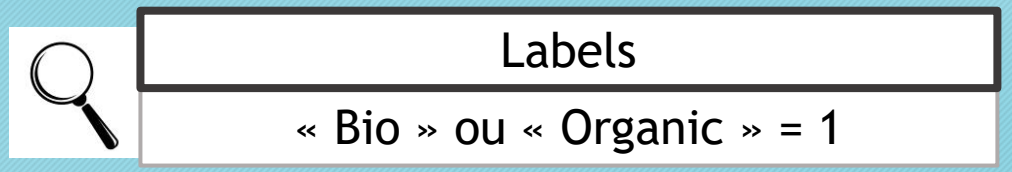
INFOS GÉNÉRALES	ADDITIFS	TAGS	ORIGINE DU PRODUIT	GROUPE DE PRODUITS	ELEMENTS CONSTITUANT LE NUTRISCORE	NUTRI-SCORE
• Code • Product_name • Brands	• additives_n • additives_fr	• Labels_fr ↓ Bio	• Origins_fr • Manufacturing _places ↓ Origine France	• Pnns_groups_1 • Pnns_groups_2	• fibers_100g • Saturated-fat_100g • Fat_100g • Proteins_100g • Sugars_100g • Sodium_100g • Fruits-vegetables-nuts-estimate-from-ingredients_100g	• Nutriscore _score • Nutriscore _grade

B) Nettoyage du jeu de données

- **Etape 1 : Nettoyage des données numériques**
  - ELEMENTS CONSTITUANT LE NUTRISCORE [0;100] (Sinon : NaN)
  - ENERGIE_100g : [0;3700] car les lipides (fat) ont une valeur de 37(kJ/g) (sinon : NaN)
- **Etape 2 : Nettoyage des données textuelles**
  - Création de la colonne « Origine France »

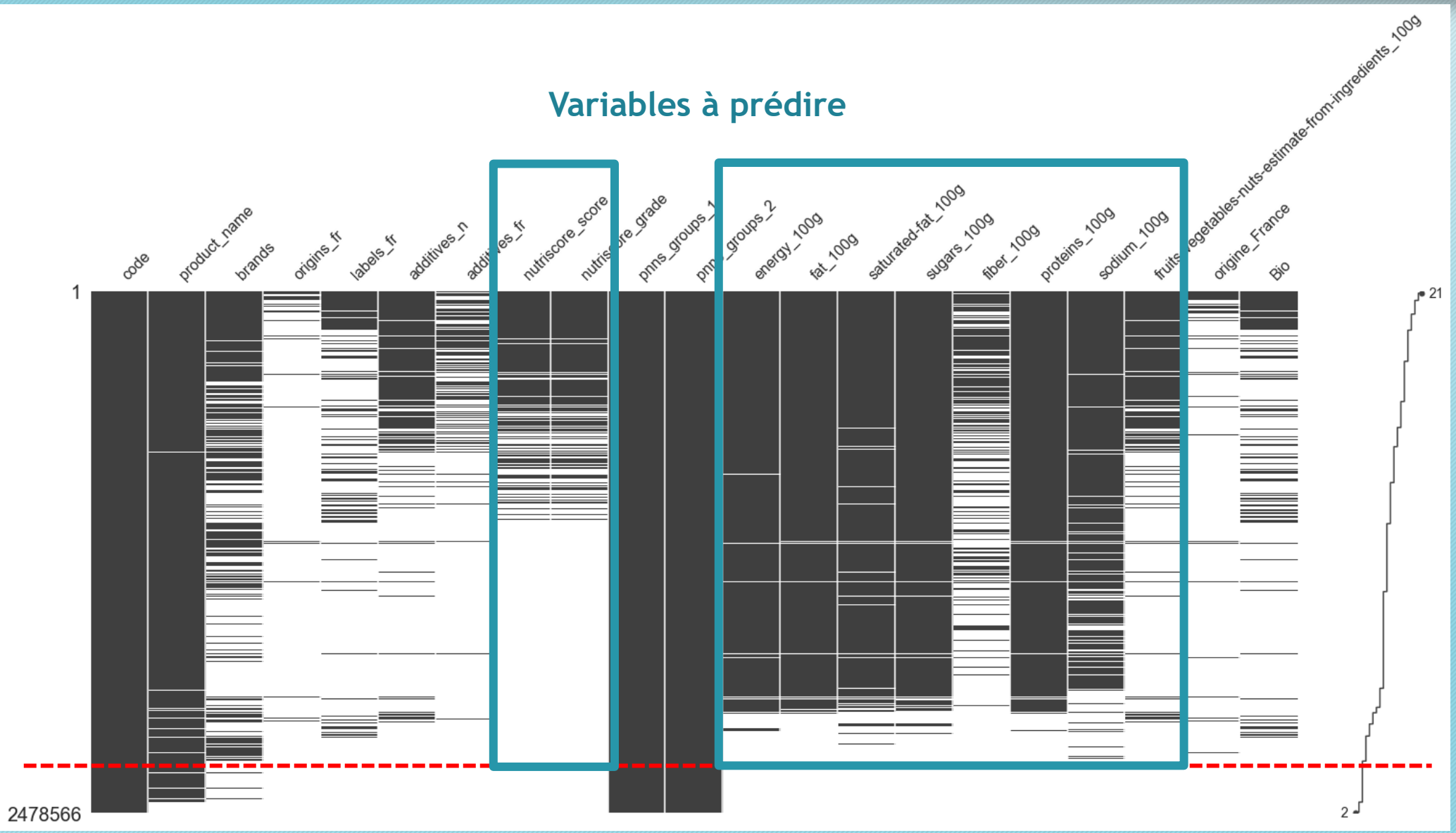


  - Création de la colonne « Bio »



  - Vérification et modification des groupes de produits (exemple : 'sugary-snacks’ > 'Sugary snacks’)
- **Etape 3 : Suppression des doublons**

C) Analyse et prédiction des valeurs manquantes



Retrait des produits avec très peu d'informations (moins de 5 variables connus).

- 245 783 produits retirées (~10%)

C) Analyse et prédiction des valeurs manquantes

Objectif : Prédire le Grade Nutri-score manquant sur certains produits

- **Etape 1 : Remplacement des valeurs manquantes**

Groupe pnns 2
« Waters » > valeurs nutritives = 0

Groupe pnns 1 et 2
NaN > « Unknown »

- **Etape 2 : Imputation des valeurs manquantes de Nutriscore-Score, fibers_100g, Saturated-fat_100g, Fat_100g, Proteins_100g, Sugars_100g, Sodium_100g, Fruits-vegetables-nuts-estimate-from-ingredients_100g**

3 formats de données en entrée :

- Données Transformées (log naturel +1), Normalisées (MinMax scaler)
- Données Transformées (log naturel +1), Normalisées (MinMax scaler) + pnns_groups_2 encodées
- Données Normalisées (MinMax scaler)

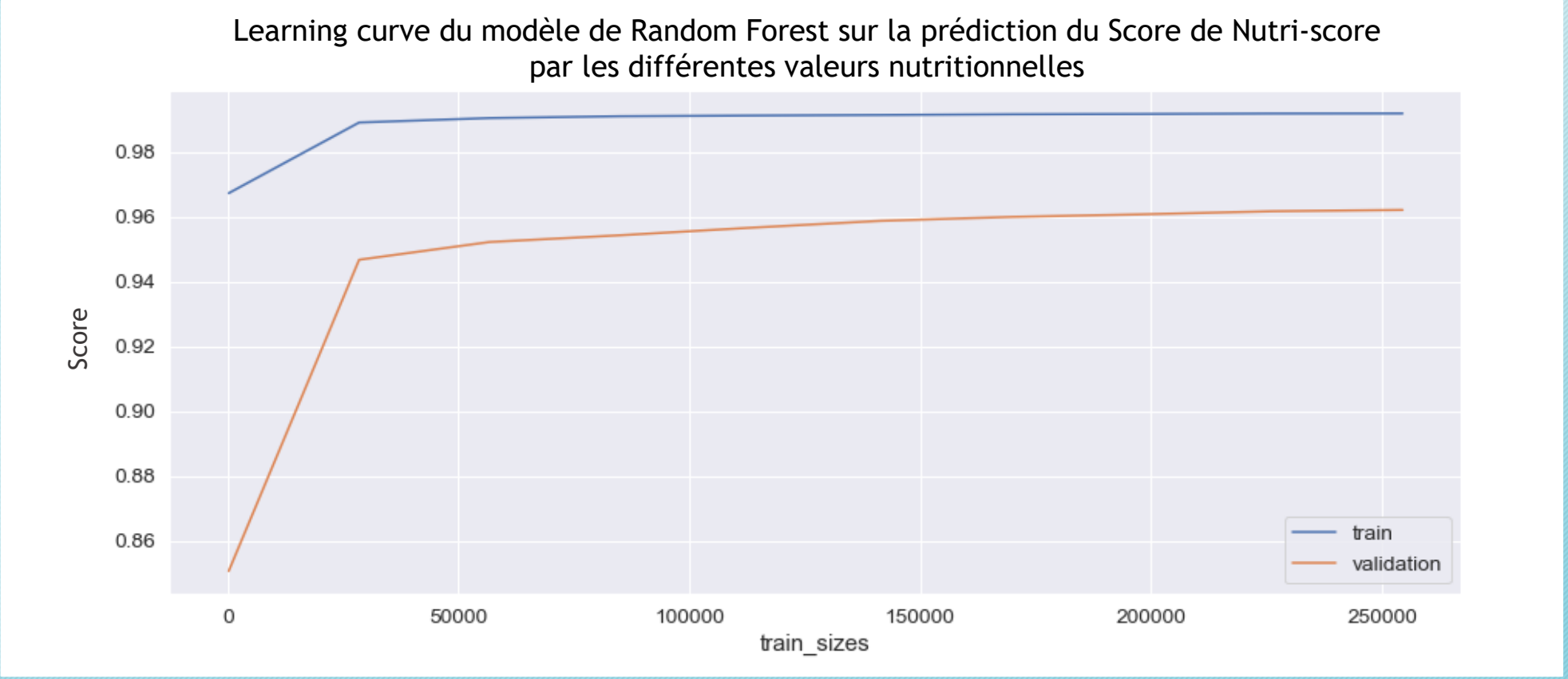
5 méthodes : Régression linéaire; KNN; Random Forest; Moyenne; Médiane

3 Critères de décision :

- Erreur moyenne absolue (Moyenne (Valeurs prédites – Valeurs du test set))
- Score du modèle (Coefficient de détermination R^2)
- Facilité à imputer les données par la suite.

C) Analyse et prédiction des valeurs manquantes

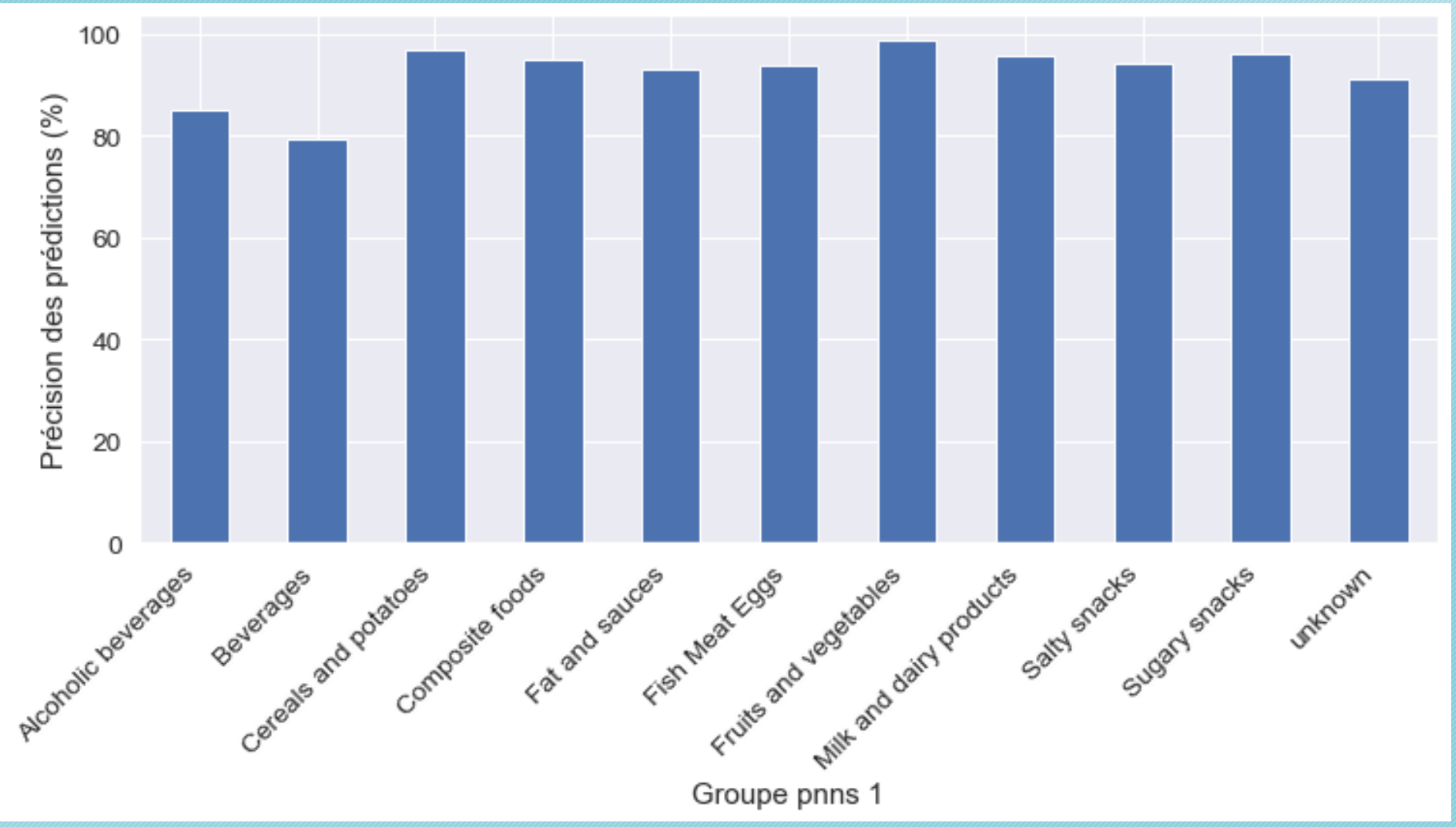
- Optimisation (gridsearchCV et Learning curve) et entraînement du modèle



- Normalisation des données de base (minMax Scaler)
- Imputation des valeurs manquantes (Iterative Imputer utilisant le modèle entrainé sur 100 000 produits)
- Restitution des dimensions d'origines (Inverse Scaling)

C) Analyse et prédiction des valeurs manquantes

- Prédiction du Grade Nutri-score en fonction du Score de Nutri-score prédit (et du type d'aliment).



- Taux d'erreur des prédictions : 6,22%
- Les prédictions sont moins précises pour les boissons (alcoolisées ou non)

Partie III : Analyse Exploratoire des données

A) Analyse multivariée des variables nutritionnelles, du Nutri-score et des groupes de produits

B) Répartition des Grades Nutri-score dans les différentes catégories de produits

C) Analyse des additifs

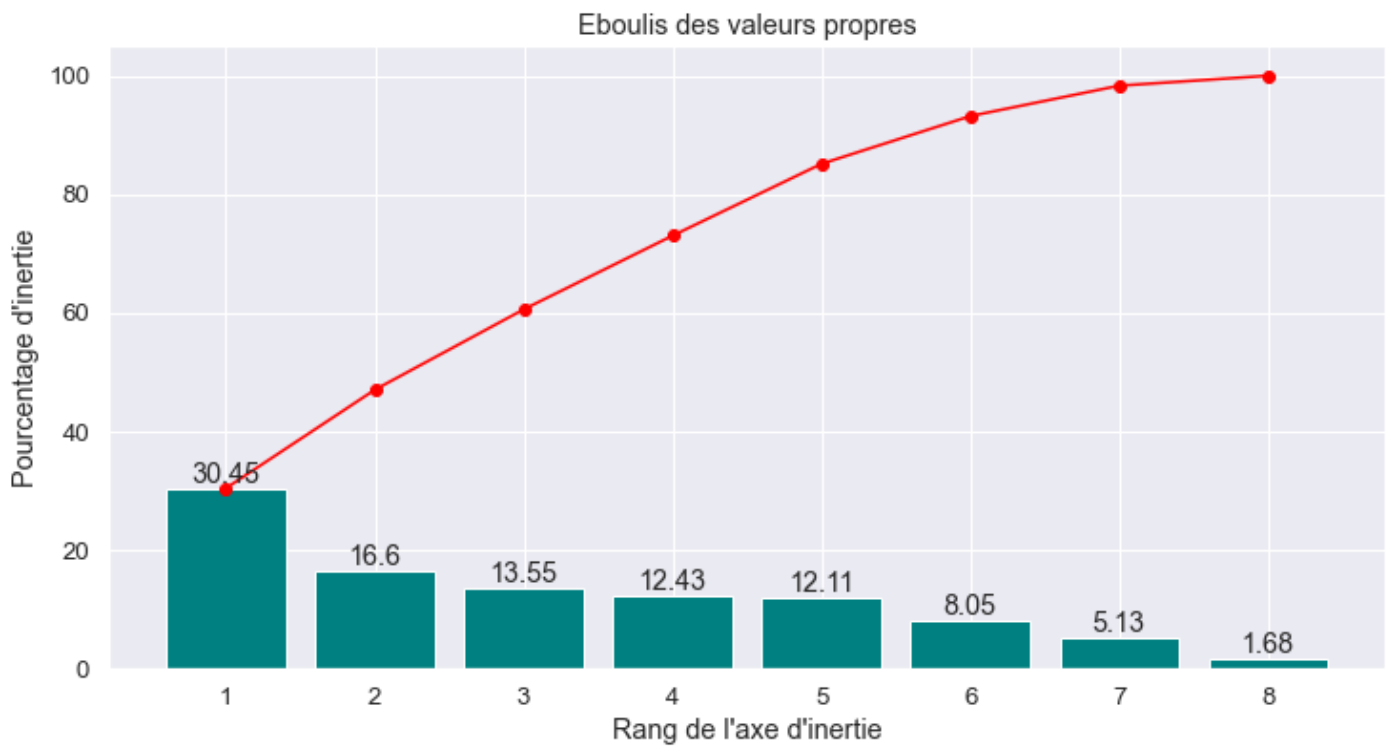
D) Analyses des produits d'origine France

E) Analyses des produits Bio

A) Analyse multivariée des variables nutritionnelles, du Nutri-score et des groupes de produits

Principe : Résumer l'information qui est contenue dans de nombreuses variables en un certain nombre d'axes synthétiques (Composantes principales) en gardant le plus d'information possible.

Détermination du nombre d'axes d'intérêts :



Critère de Kayser : on ne garde que les composantes > (100/*p*)% où *p* est le nombre de variables.

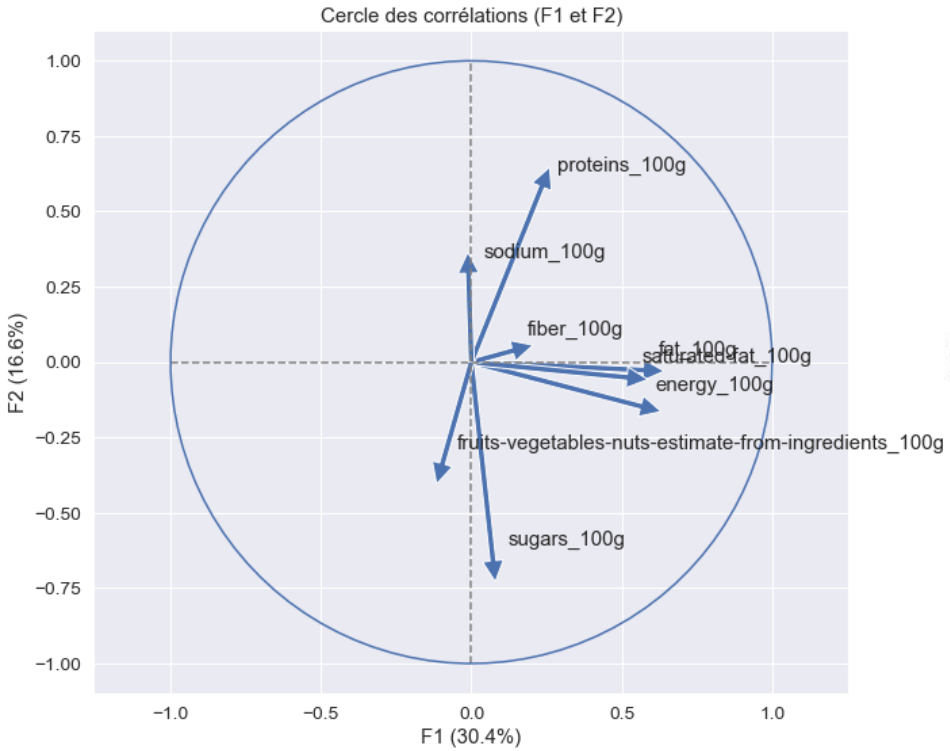
- (100/*p*)% = 12,5 % > Axe 1 à 3

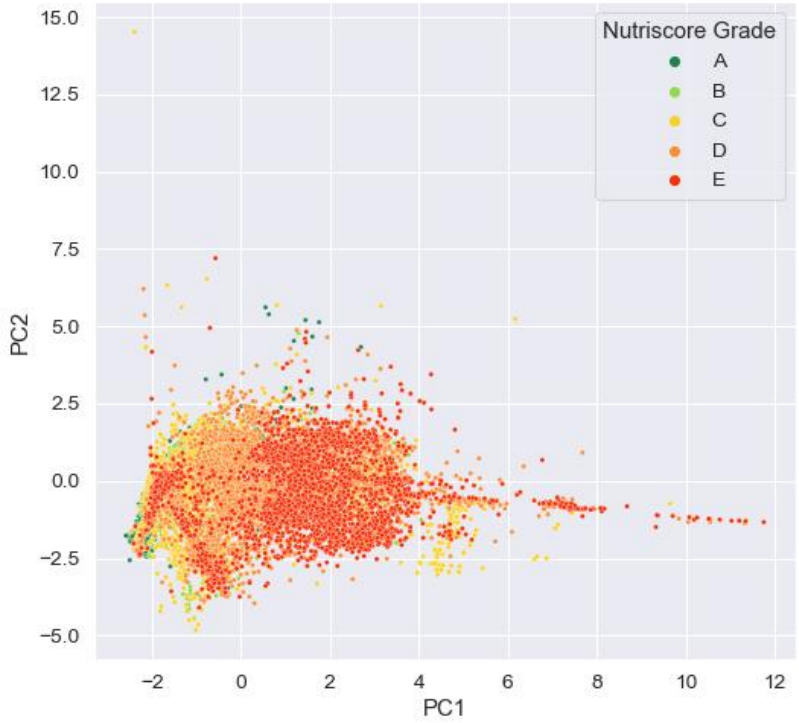
Méthode du coude : méthode visuelle

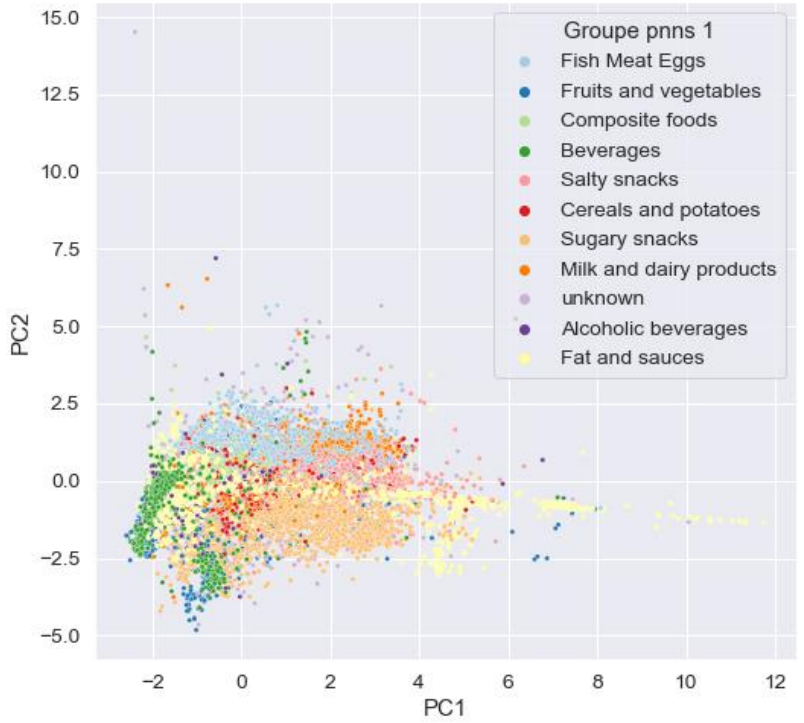
- Axe 1 à 5

Nous analyserons les **3 premières dimensions**, expliquant 61% de la variance totale.

A) Analyse multivariée des variables nutritionnelles, du Nutri-score et des groupes de produits



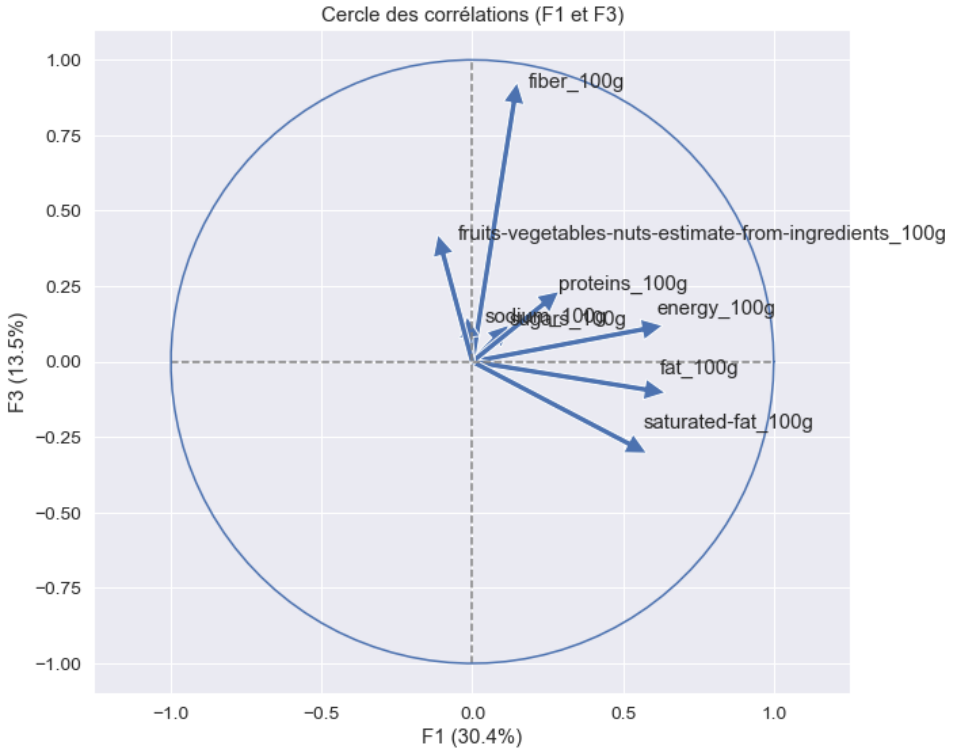


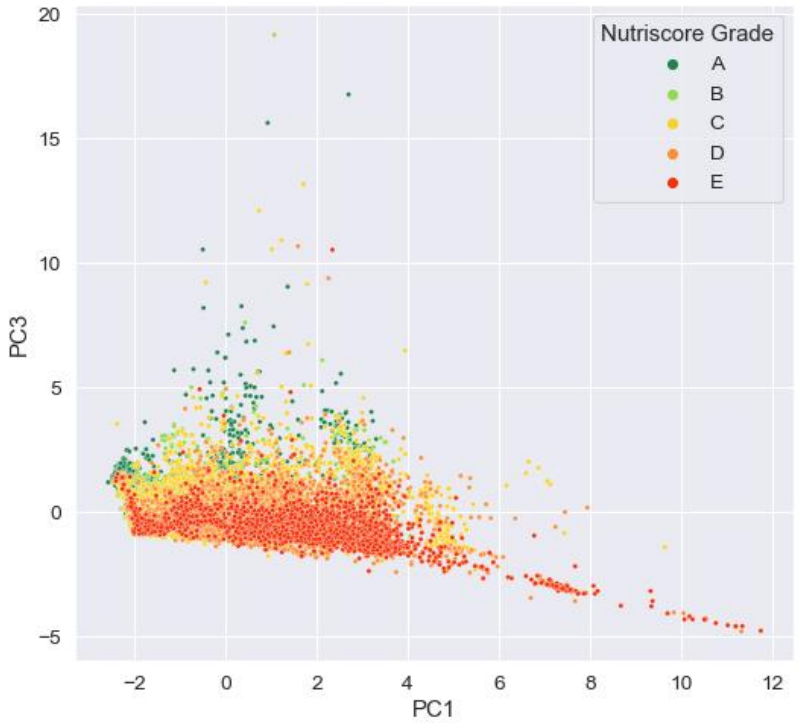


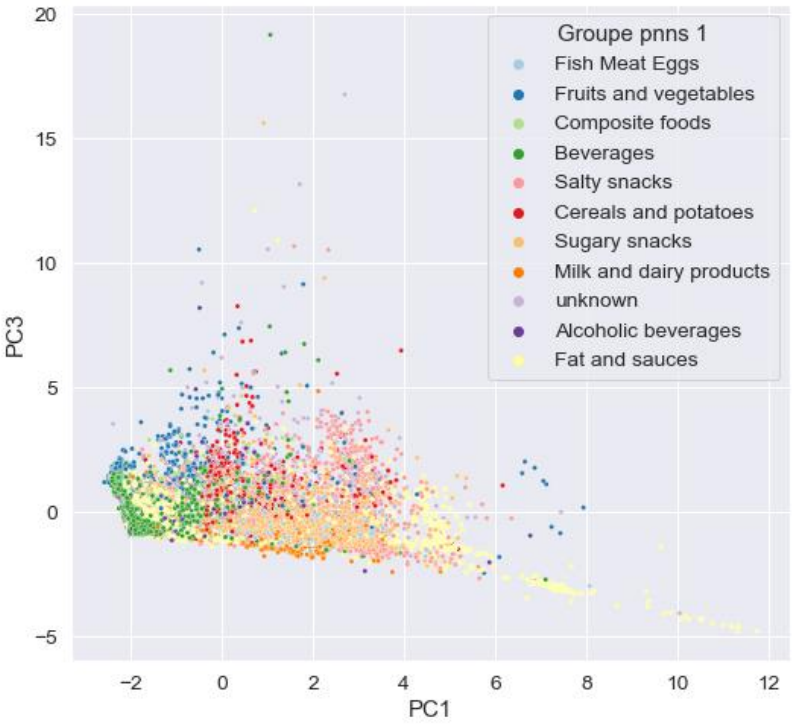
- Les produits de Nutri-score D et E sont plus riches en énergie, lipides et sucres

- Les produits « graisses et sauces » semblent très riches en énergie et lipides
- Les produits « biscuits sucrés » semblent riches en sucre
- Les groupes Poisson-Viande-Œufs et les produits laitiers apparaissent bien comme riches en protéine.

A) Analyse multivariée des variables nutritionnelles, du Nutri-score et des groupes de produits

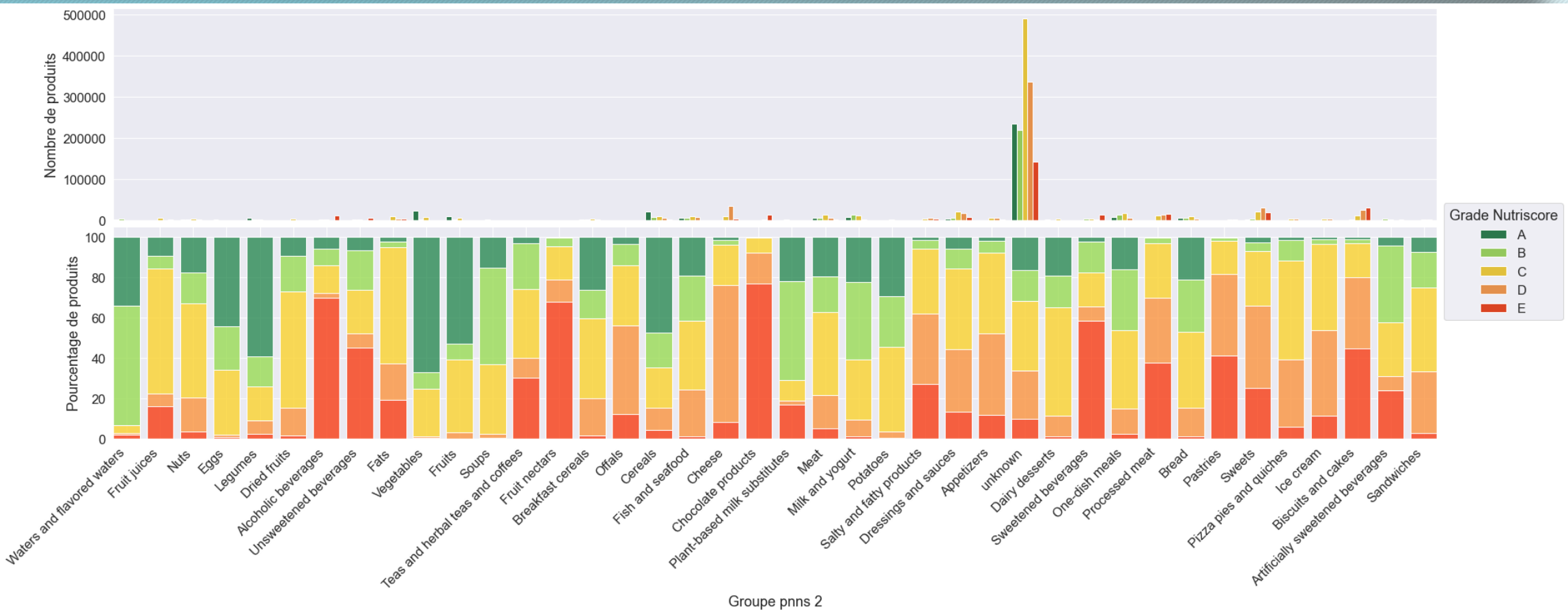






- Les produits de Nutri-score A et B sont plus riches en fibre
- Produits de catégorie Fruits et légumes

B) Répartition des Grades Nutri-score dans les différentes catégories de produits



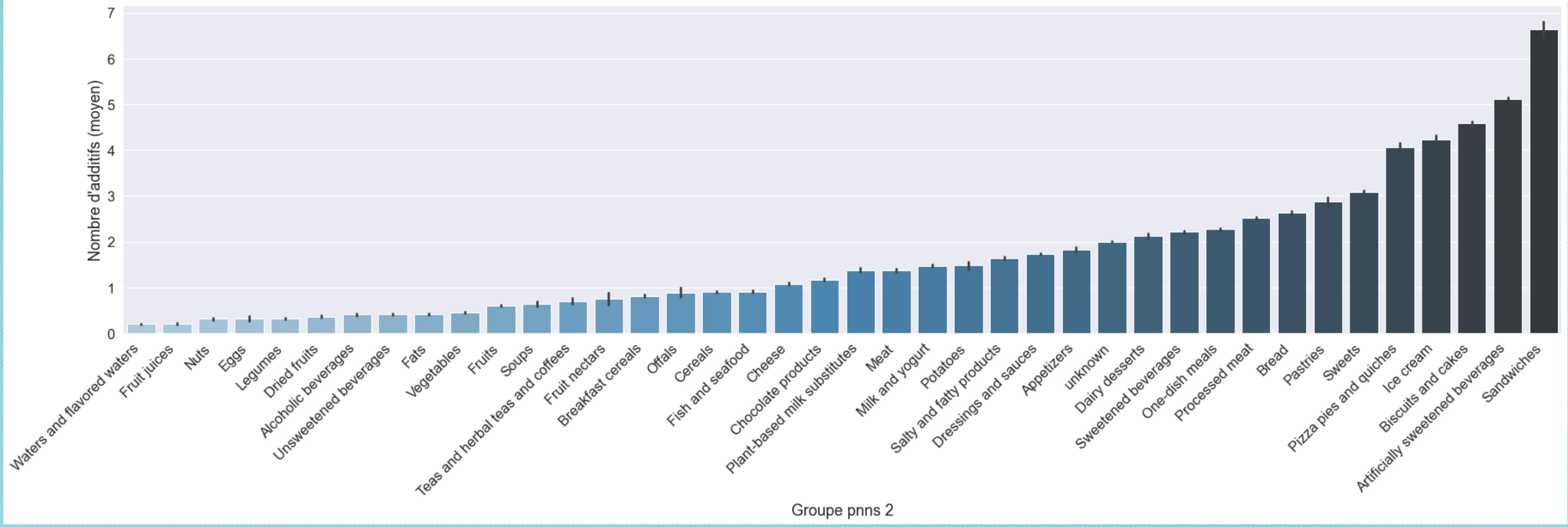
Le Grade Nutri-score est-il réparti équitablement entre les différentes catégories de produits ?

- Non, Il y a une différence significative de grade Nutri-score entre les différents groupes (x^2 = 678746.7 , p-value = 0)

Existe-t-il pour chaque produit une alternative de Grade Nutri-score plus faible ?

- Pour chaque catégorie, il existe des alternatives de produits à Nutri-score plus faible

C) Analyse des additifs : Groupe de produits



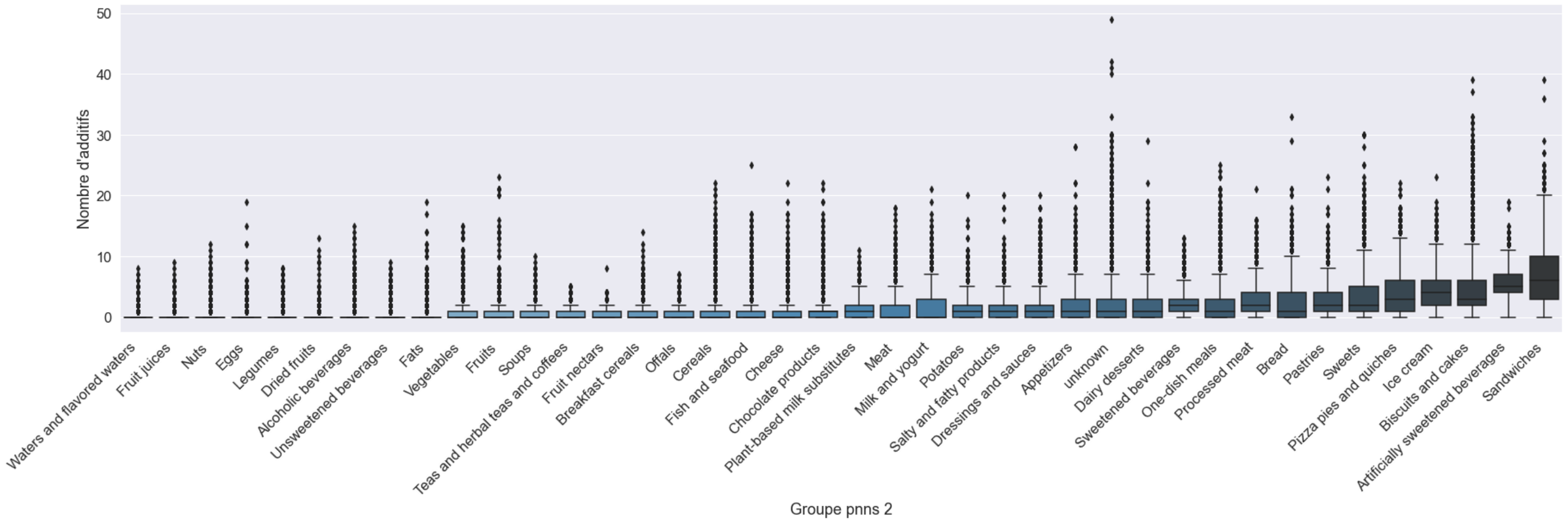
➢ **Quelle est l'effet du groupe de produits sur le nombre d'additifs ?**

La distribution du nombre d'additifs s'approche d'une Loi de Poisson (données de comptage) : Modèle linéaire généralisé et comparaison d'AIC avec un modèle nul.

- Modèle nul : Nombre d'additifs ~ 1 > AIC = 3 951 255.46
- Modèle Grade : Nombre d'additifs ~ Groupe pnns 2 > AIC = 3 460 217.96

➢ Il y a un nombre moyen d'additifs différents entre les différents groupes de produits (validé par Test Post hoc de Dunn)

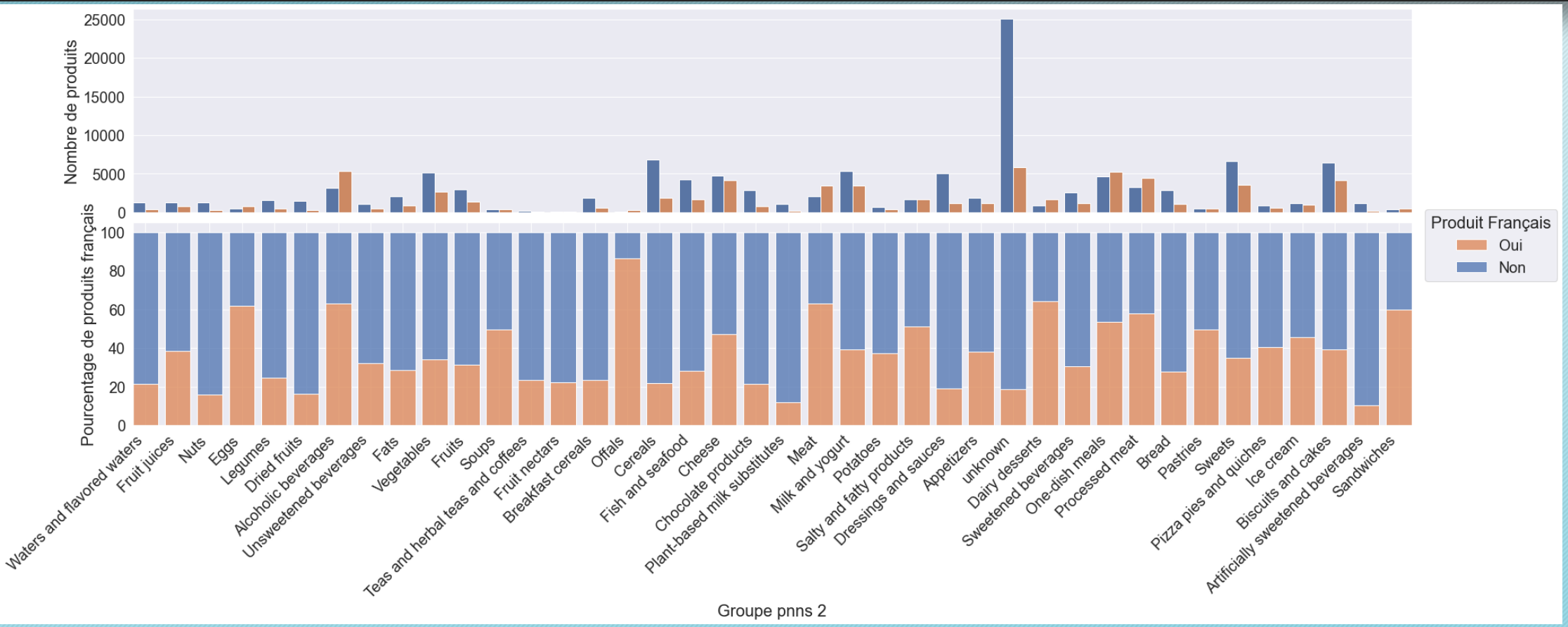
C) Analyse des additifs : Groupe de produits



Existe-t-il des alternatives de produits avec moins d'additifs ?

- Oui, pour chaque groupe de produit, il existe des alternatives avec moins (ou sans) additifs.

D) Analyse des produits d'origine France



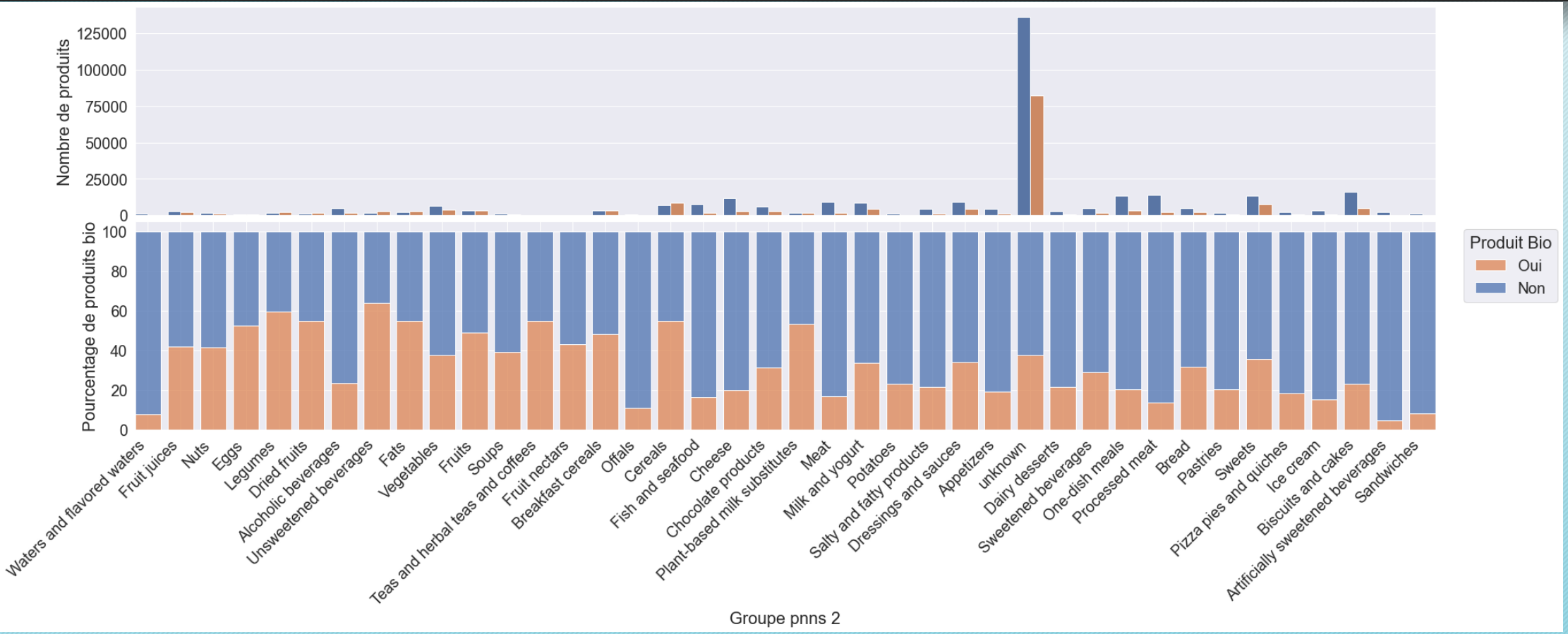
Les produits Français sont-ils répartis équitablement entre les différentes catégories de produits ?

- Non, Il y a une différence significative d'origine des produits entre les différents groupes (x^2 = 18221.4 , p-value = 0)

Existe-t-il des alternatives d'origine France pour chaque groupe de produits ?

- Oui, pour chaque catégorie, il existe des alternatives françaises

E) Analyse des produits bio



Les produits Bio sont-ils répartis équitablement entre les différentes catégories de produits ?

- Non, Il y a une différence significative de répartition des produits bio entre les différents groupes ($x^2 = 25416.9$, p-value = 0)

Existe-t-il des alternatives Bio pour chaque groupe de produits ?

- Oui, pour chaque catégorie, il existe des alternatives bio.

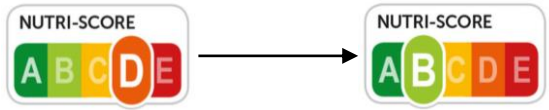
Synthèse et faisabilité de l'application

Proposition d'aliments

de la même famille

Manger plus sain

- Nutri-score

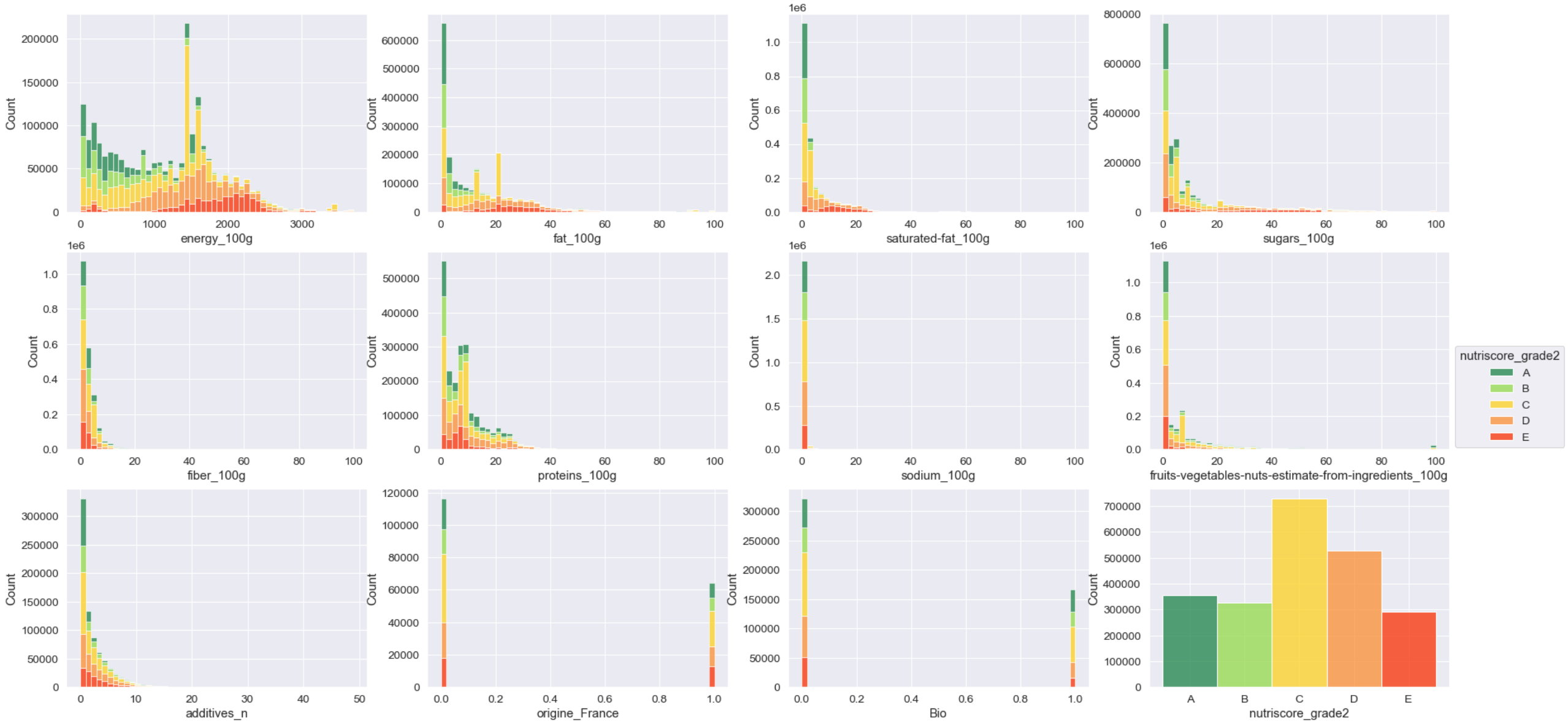


- - d'additifs

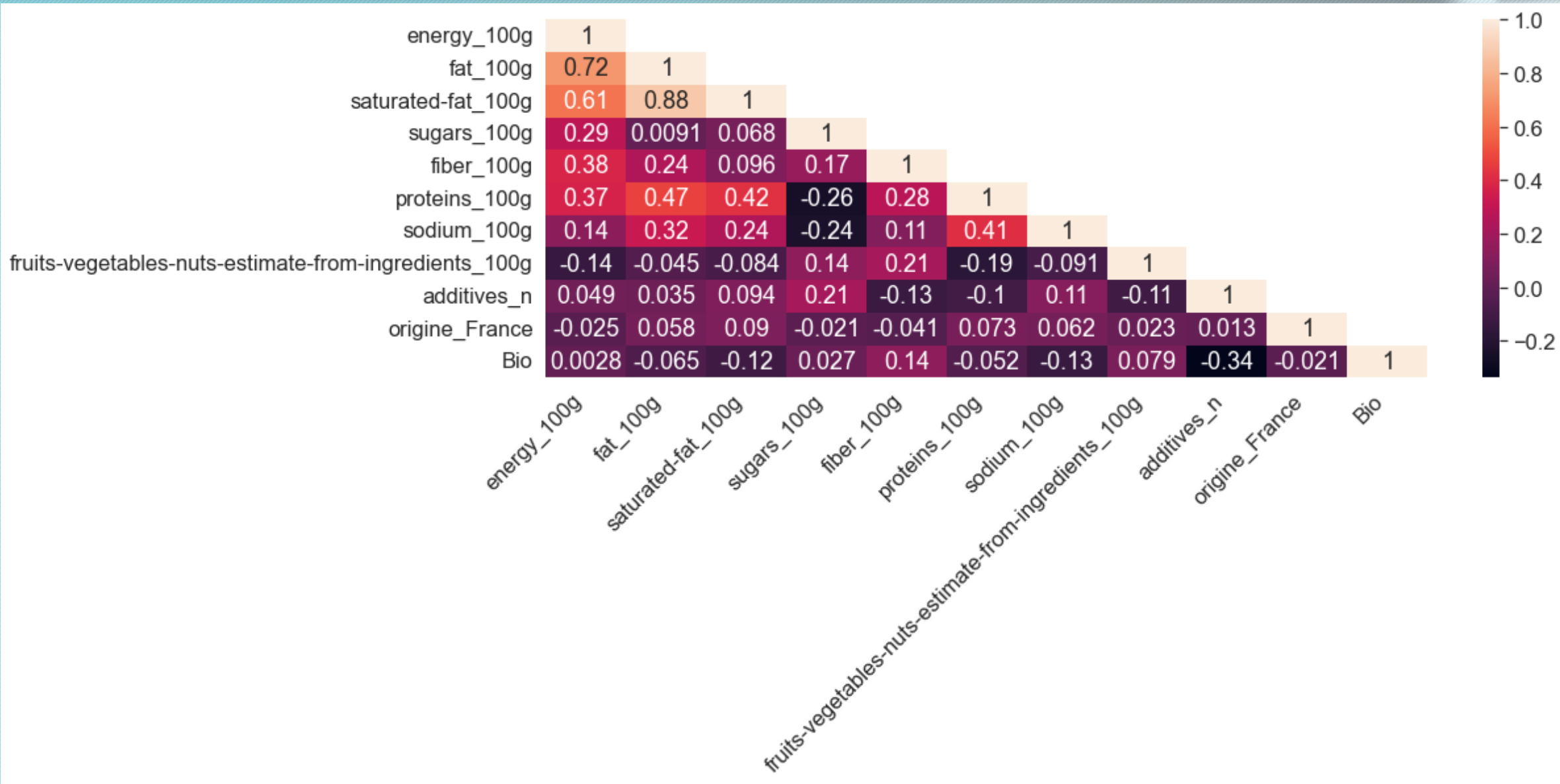
Manger plus responsable

- Origine France
- Label Biologique

Annexe 1 : Répartition empiriques des données

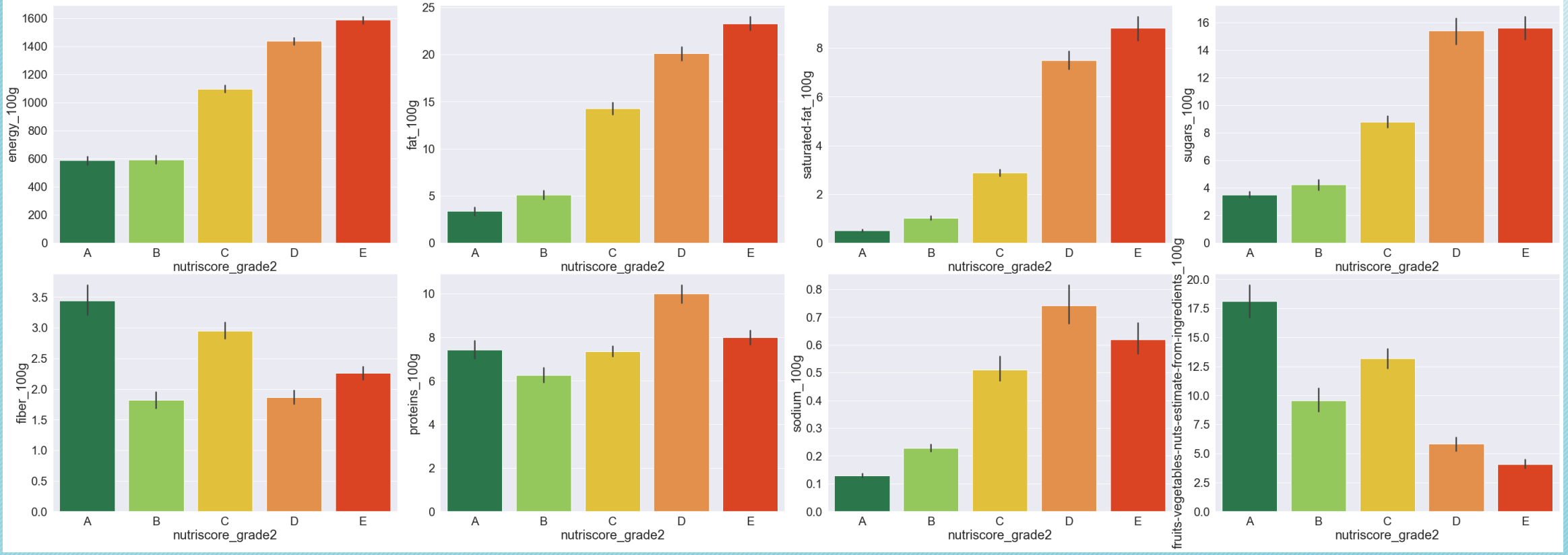


Annexe 2 : Coefficients de corrélation (Spearman) entre les différentes variables



Annexe 3 : Coefficients de corrélation (Spearman) entre les différentes variables

➢ **Quelle est la relation entre chaque variable nutritionnelle et le grade Nutri-score ?**



Pour chaque variable, les conditions d'applications de l'ANOVA ne sont pas respectées

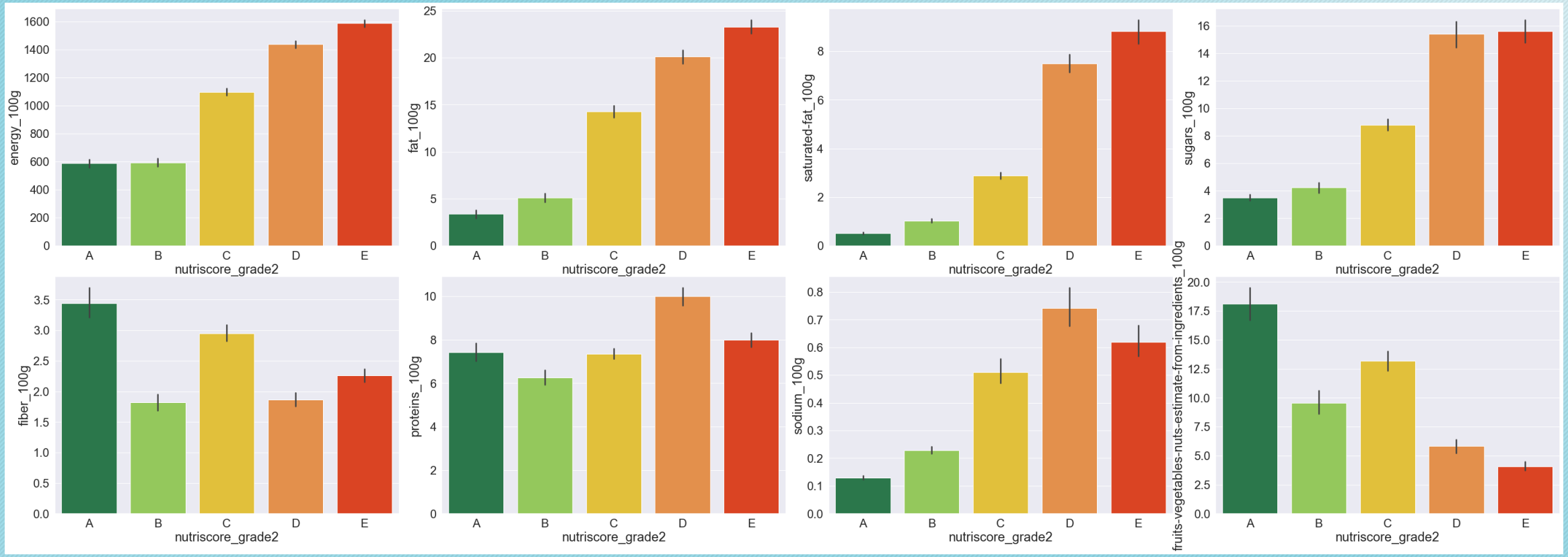
- Normalité des résidus : Test de Kolmogorov Smirnov = X
- Homoscédasticité : Test de Bartlett = X

➢ Equivalent de l'ANOVA en non paramétrique : Test de Kruskall Wallis

➢ Comparaisons 2 à 2 : Test post-hoc de Dunn

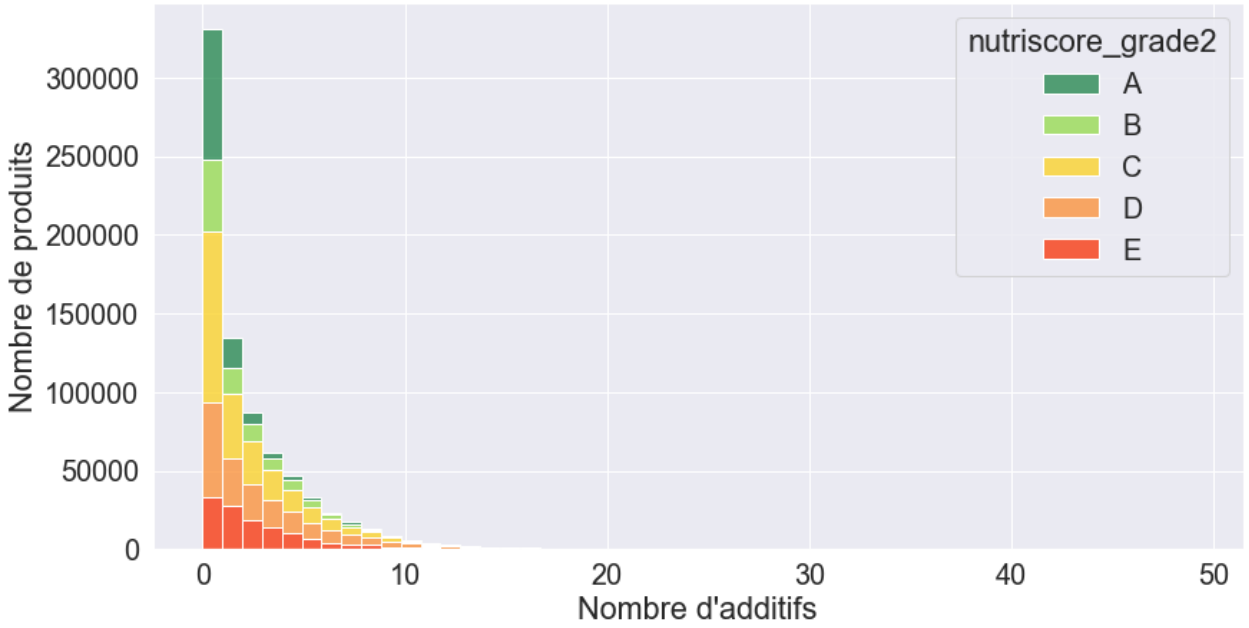
Annexe 3 : Coefficients de corrélation (Spearman) entre les différentes variables

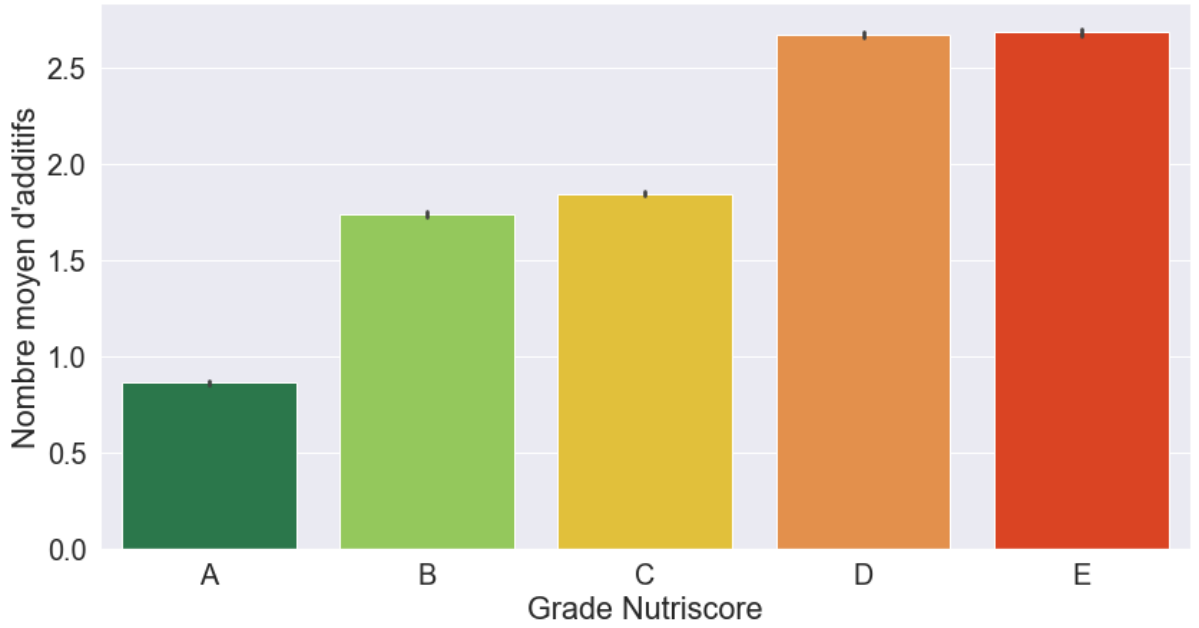
- **Quelle est la relation entre chaque variable nutritionnelle et le grade Nutri-score ?**



- L'effet de chaque variable de valeur nutritive sur le Grade Nutri-score est statistiquement significatif (p-value = 0)
- Les aliments de grade Nutris-score D et E sont plus riches en énergie, acides gras (saturés + insaturés) sucres et sel.
- Les aliments de grade Nutris-score A sont plus riches en fibres, fruits et légumes.

Annexe 4 : Analyse des additifs et du Grade Nutri-score





Quelle est la relation entre le nombre d'additifs et le Grade Nutri-score ?

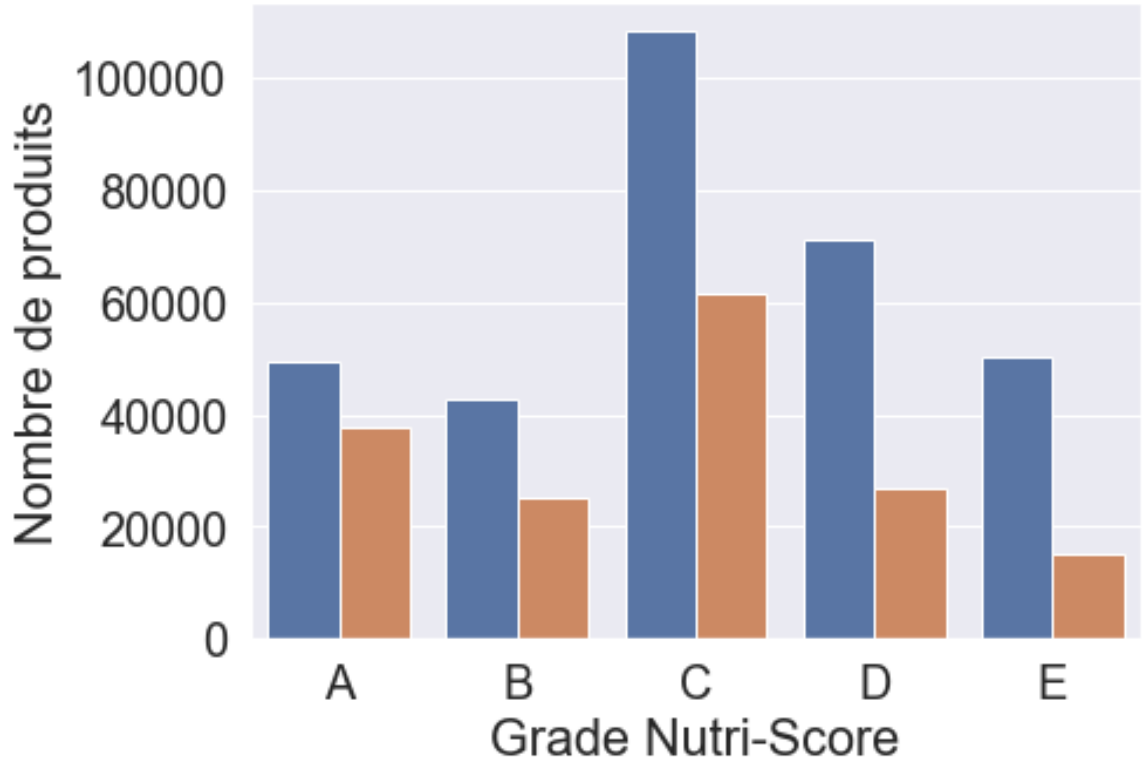
La distribution du nombre d'additifs s'approche d'une Loi de Poisson (données de comptage) : Modèle linéaire généralisé et comparaison d'AIC avec un modèle nul.

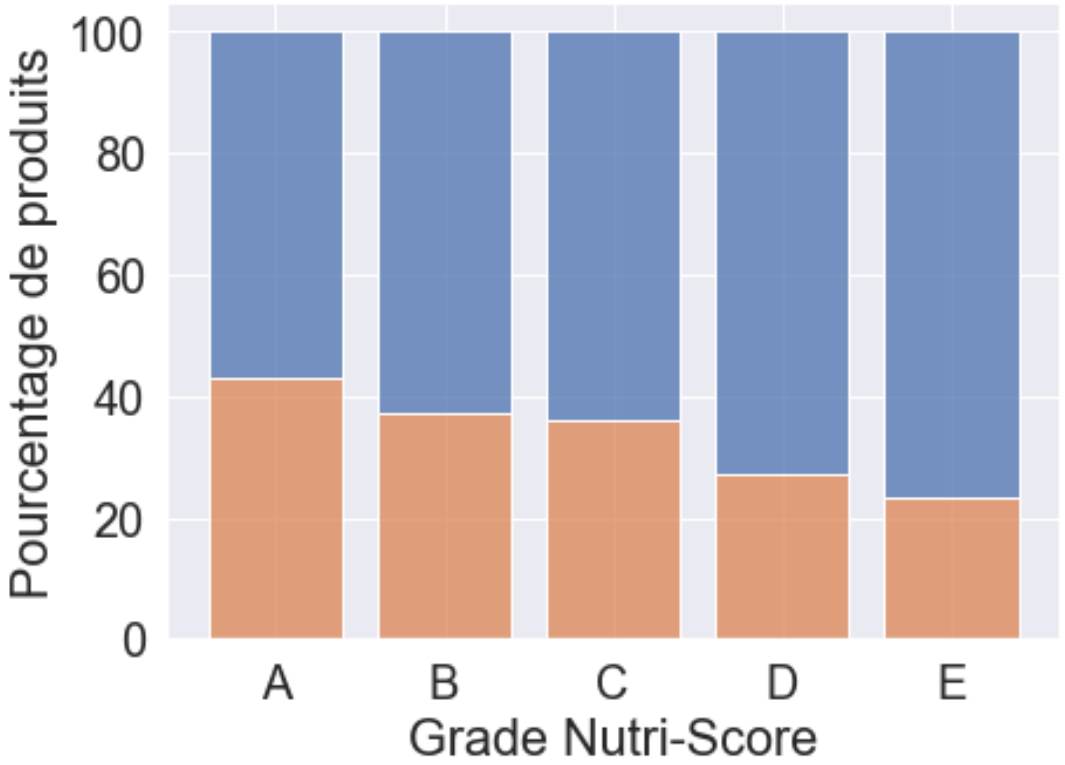
- Modèle nul : Nombre d'additifs ~ 1 > AIC = 3 951 255.46
- Modèle Grade : Nombre d'additifs ~ Grades Nutri-scores > AIC = 3 779 254.99

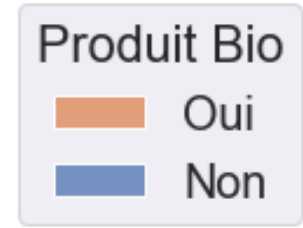
➢ Il y a un nombre d'additifs différents entre les grades Nutri-score (validé par Test Post hoc de Dunn)

Annexe 5 : Analyse des produits bio et du Nutri-score

- **Quelle est la relation entre les produits Bio et le Grade Nutri-score ?**





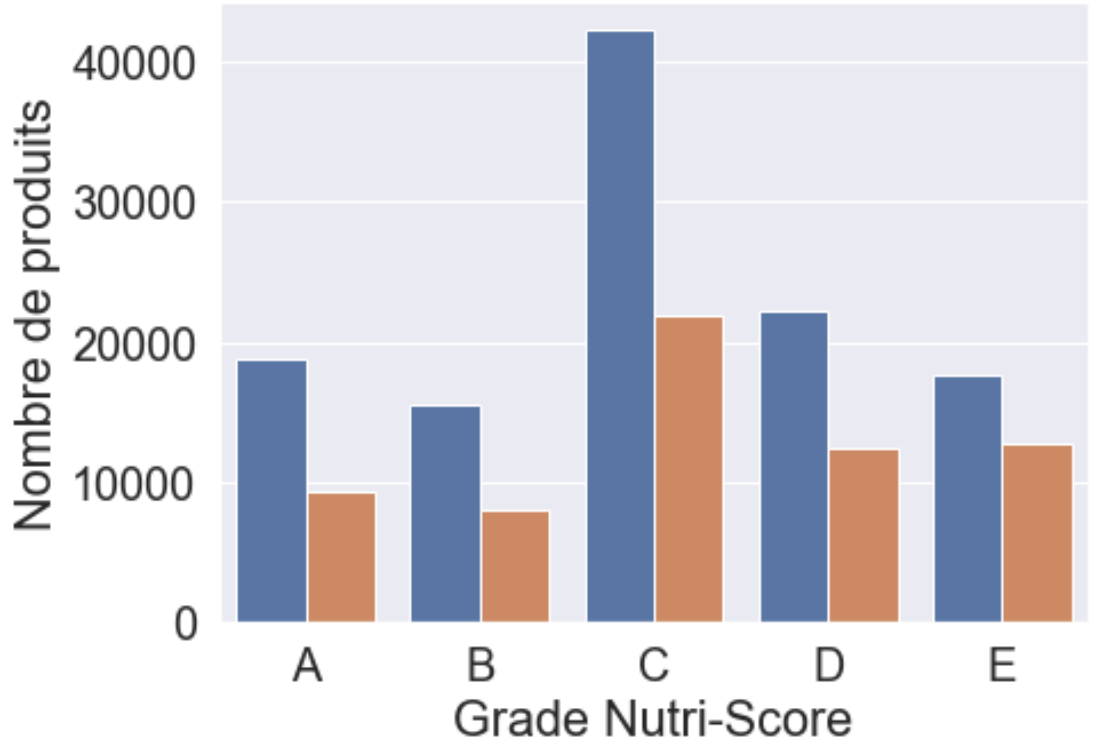


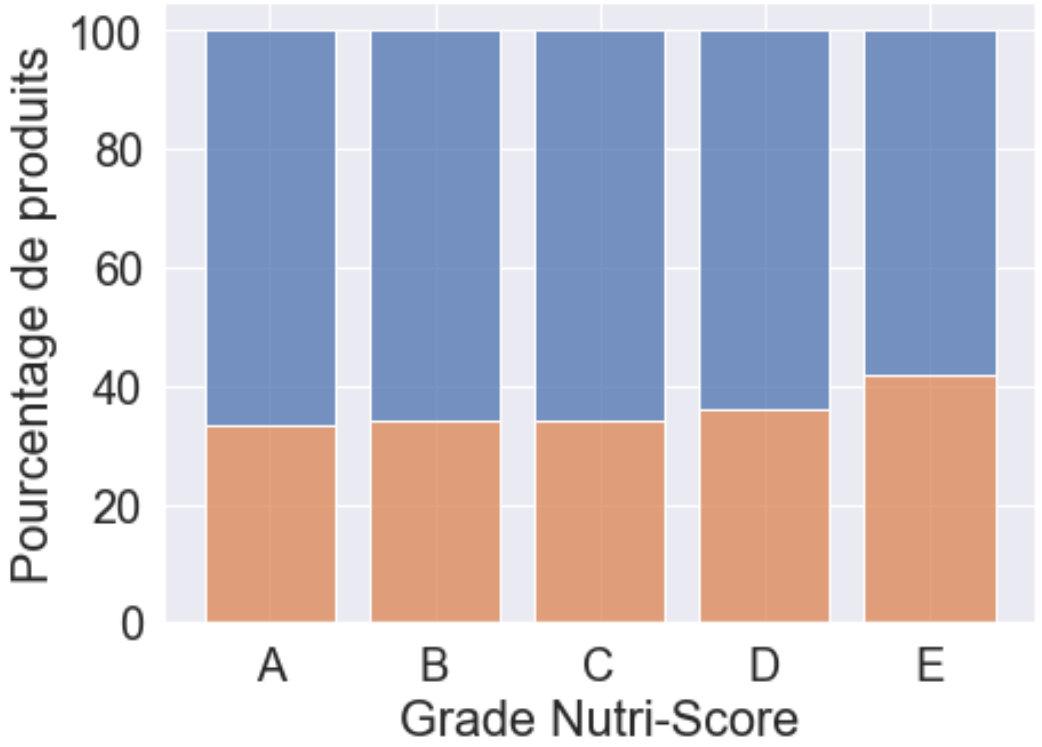
- Il y a une différence significative de grade Nutri-score entre les différents groupes ($\chi^2 = 9166$, p-value=0)

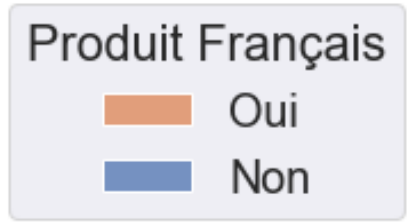
- Pour chaque catégorie, il existe des alternatives de produits à Nutri-score plus faible

Annexe 6 : Analyse des produits d'origine France et du Nutri-score

- **Quelle est la relation entre les produits Français et le Grade Nutri-score ?**







- Il y a une différence significative de grade Nutri-score entre les différents groupes ($\chi^2 = 669.4$, p-value $<10^{-142}$)

- Pour chaque catégorie, il existe des alternatives de produits à Nutri-score plus faible