Concevez une application au service de la santé publique

Projet n°3 – Parcours Data Scientist – Jérémy VANGANSBERG

Table des matières

L’application

Le nutri-score

Santé publique France

- Label attribuant une note de A à E aux aliments
- Cette classification est dérivée d'un score numérique de **-15 à 40**
- Il est calculé en prenant en compte les éléments suivants :

Eléments défavorables au score	Eléments favorables au score
Apport calorique	fruits, légumes, légumineuses (dont les légumes secs), oléagineux, huiles de colza, de noix et d'olive
Sucre	
Graisses saturées	Fibres
Sel	Protéines

- Il y a **4 formules différentes** : boissons, fromages, matières grasses, autres aliments

Pitch : Générateur de nutri-score

Le nutri-score est un outil qui permet aux consommateurs d'**évaluer la qualité nutritionnelle** d'un aliment…

[…] mais ce score n'est pas utilisé par tous les acteurs de l'agro-alimentaire.

L'idée ?

Créer un outil qui permet **d'évaluer le score** d'un aliment en **scannant l'étiquette** des **données nutritionnelles** à l'aide de l'appareil photo de son smartphone

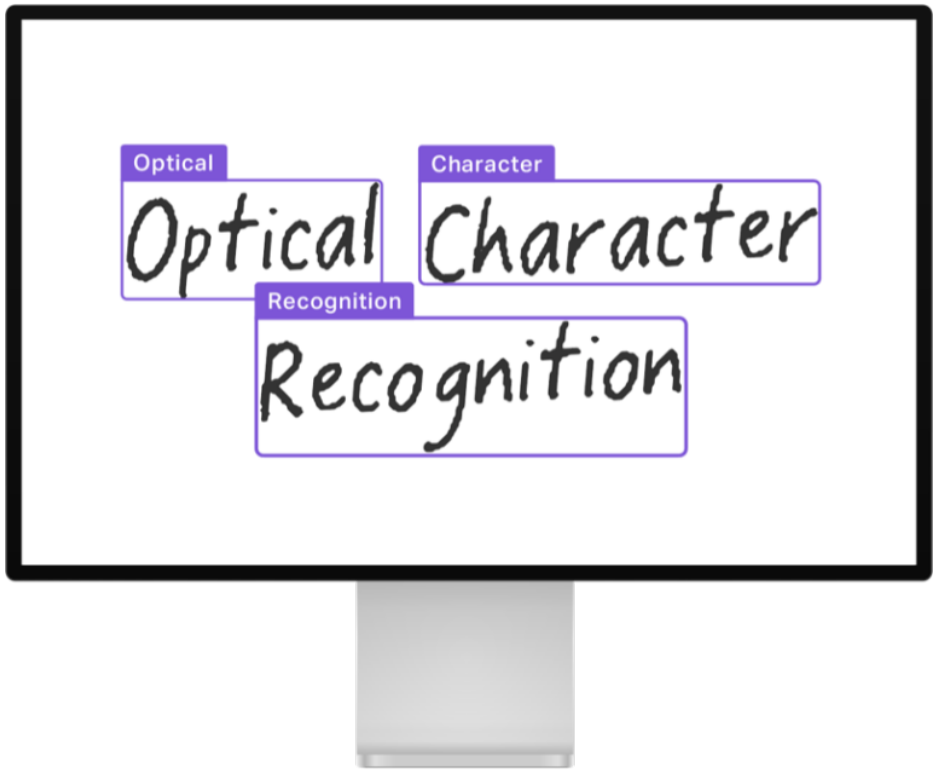
Aspect technique

Comment mettre en œuvre cette idée ?

Deux algorithmes au cœur de l'idée :

- Algorithme d'***O**ptical **C**haracter **R**ecognition*
  - ➔ Lecture des informations sur l'étiquette nutritionnelle
- Régression linéaire multiple
  - ➔ Source : le jeu de données **d'open food fact**
  - ➔ Estimation du score

Optical Character Recognition



Algorithme de deep learning : ConvNet

- *Text detection* : contour du mot
- *Text recognition* :
  - Contour des lettres
  - Preprocessing des lettres
  - Prédiction du mot

Source : https://www.elementai.com/fr/api/ocr

Remarque sur l'approche

La **formule** n'est pas prise en compte car elle implique de **connaître la catégorie du produit**.

Cependant, j'estime que la **catégorie de produit est difficile à identifier par le biais d'un OCR**. En effet, cette information n'est pas standardisée comme c'est le cas pour l'étiquette nutritionnelle.

Il y a également l'alternative où on pourrait imaginer que **l'utilisateur choisisse la catégorie du produit qu'il souhaite scanner**. Cependant ajouter une étape systématique. Ca peut être néfaste pour l'expérience utilisateur. **Je préfère perdre un peu en précision au profit de la facilité d'utilisation**.

Nettoyage

Caractéristiques du jeu de données

Fichier CSV :

- Source : open food fact
- **3Go**
- 1 555 491 lignes
- 183 colonnes

- Le fichier est **trop volumineux** pour le charger en entier sous un jupyter notebook

Stratégie d'import

Deux options envisagées:

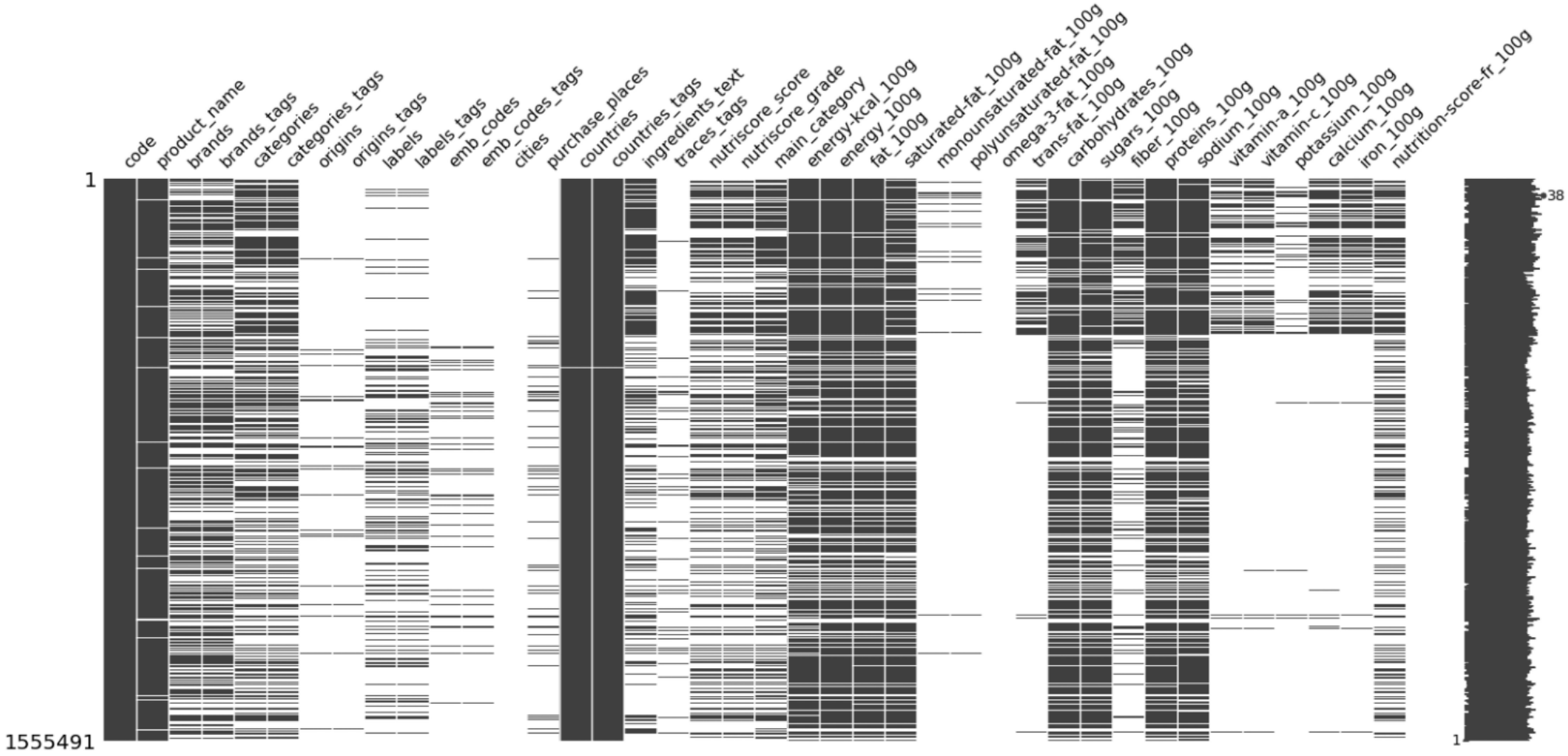
- Découper le fichier en tronçons (*chunks*)
- **Sélectionner** uniquement les variables **intéressantes** parmi les 183 colonnes

Les catégories de variables sélectionnées

- Information relatives aux produits :
  - Nom
  - Marque
  - Catégorie
  - Label
  - Pays d'origine
- Informations relatives à l'apport **nutritionnelle**
  - Nutri-score
  - Macronutriments
  - Micronutriments
  - Apport énergétique

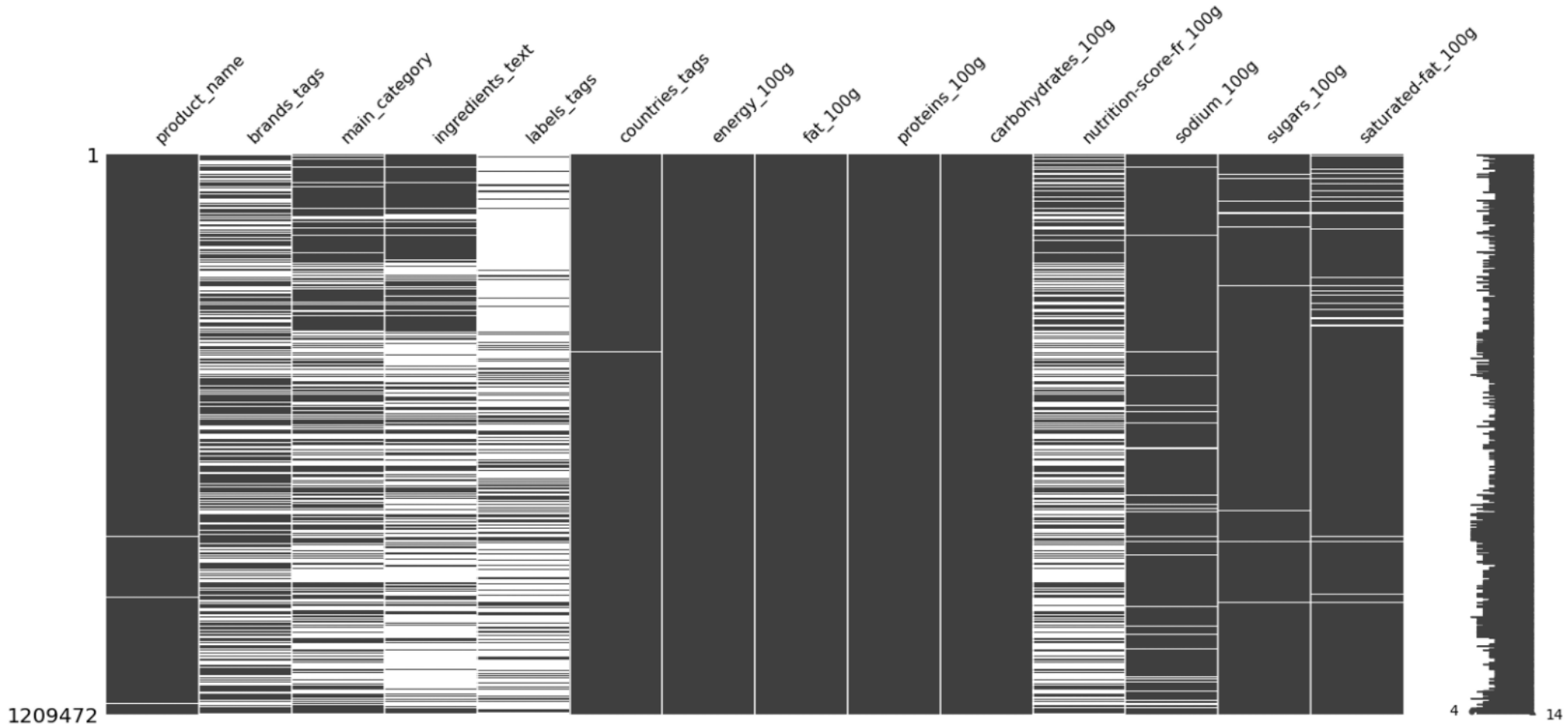
Visualisation à l'aide missingno

Structure des variables sélectionnées



Visualisation à l'aide missingno

Structure des variables sélectionnées



Valeurs dupliquées

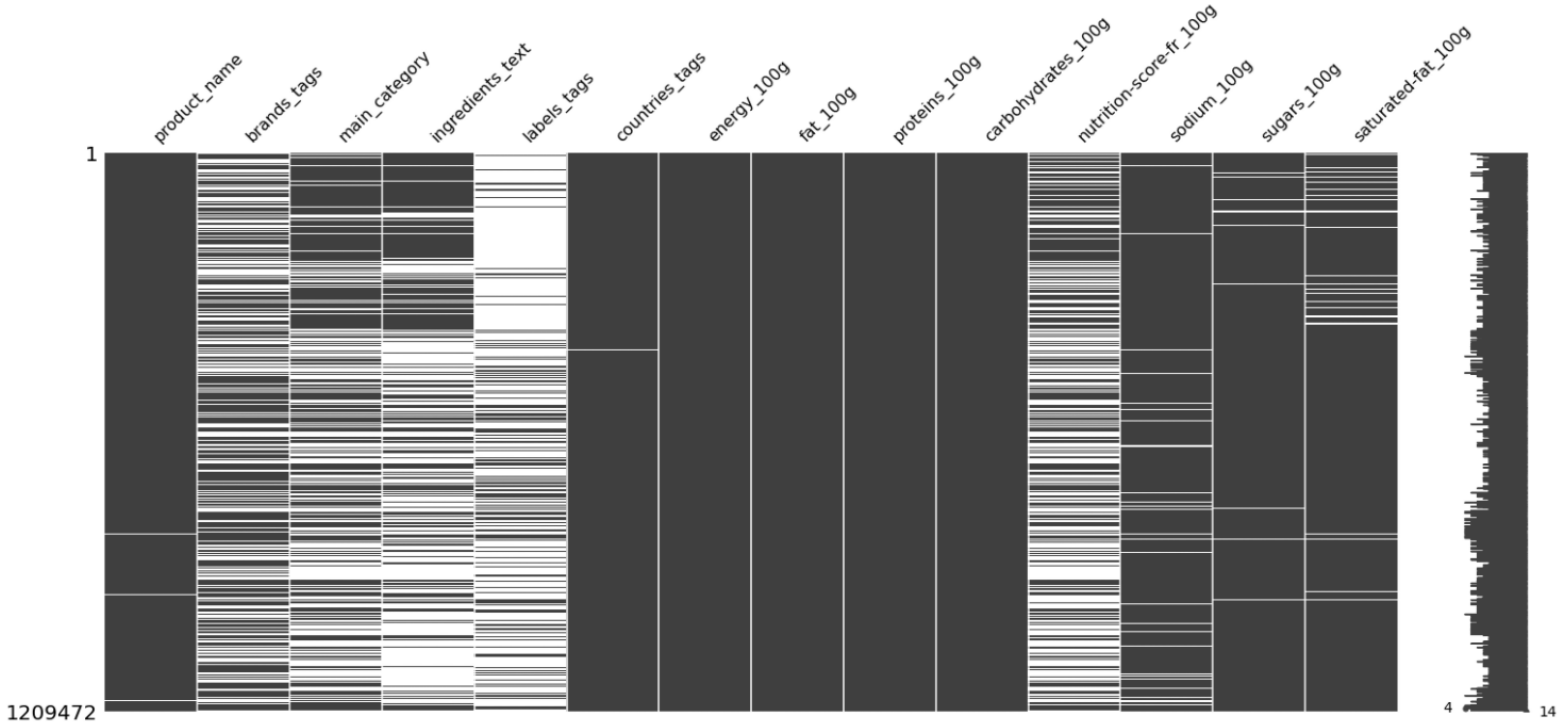
J’ai décidé de supprimer les valeurs dupliquées uniquement sur ces variables :

- *'product_name’, 'nutrition-score-fr_100g’,*
- *'fat_100g’, 'proteins_100g', 'carbohydrates_100g’*
- *'sodium_100g', 'sugars_100g’, ‘saturated-fat_100g’*

Bilan : 73387 entrées ont été supprimées, soit 6% du dataset

Valeurs manquantes

Structure des variables sélectionnées



- 1ère opération :
  - Supprimer les lignes qui ont des NaN dans toutes les colonnes
- 2e opération :
  - Supprimer les lignes qui comportent un NaN dans l'une 3 colonnes renseignant les macronutriments
- 3e opération :
  - Supprimer les lignes qui comportent un NaN dans : sel, sucre ou graisse saturée

Imputation des valeurs manquantes

Variable cible : nutrition-score

% de NaN : 57

Méthodes testées:

- KNN Imputer : 13h d'exécution
- Iterative Imputer : La variable cible (y) est définie comme une fonction des autres variables (X)
- Aucun imputer n'a été retenu. Le nutri-score sera estimé avec la régression multiple

Jeu de données et variable cible:

	energy_100g	fat_100g	proteins_100g	carbohydrates_100g	nutrition-score-fr_100g	sodium_100g	sugars_100g
0	1569.0	7.0	7.8	70.1	NaN	0.560	15.0
3	936.0	8.2	5.1	29.0	18.0	1.840	22.0
5	88.0	0.0	0.2	4.8	NaN	0.816	0.4
6	251.0	3.0	2.0	10.0	NaN	0.460	3.0
13	134.0	0.3	0.9	5.3	1.0	0.168	3.9

Valeurs aberrantes

- Les variables numériques sont exprimées sur **une base de 100g**
- Par conséquent :
    1. Il ne peut pas avoir une valeur supérieur à 100 ou inférieur à 0
    2. Le total des macronutriments ne peut pas être supérieur à 100
- J’ai supprimé toutes les variables qui ne respectaient pas ces critères

Divers

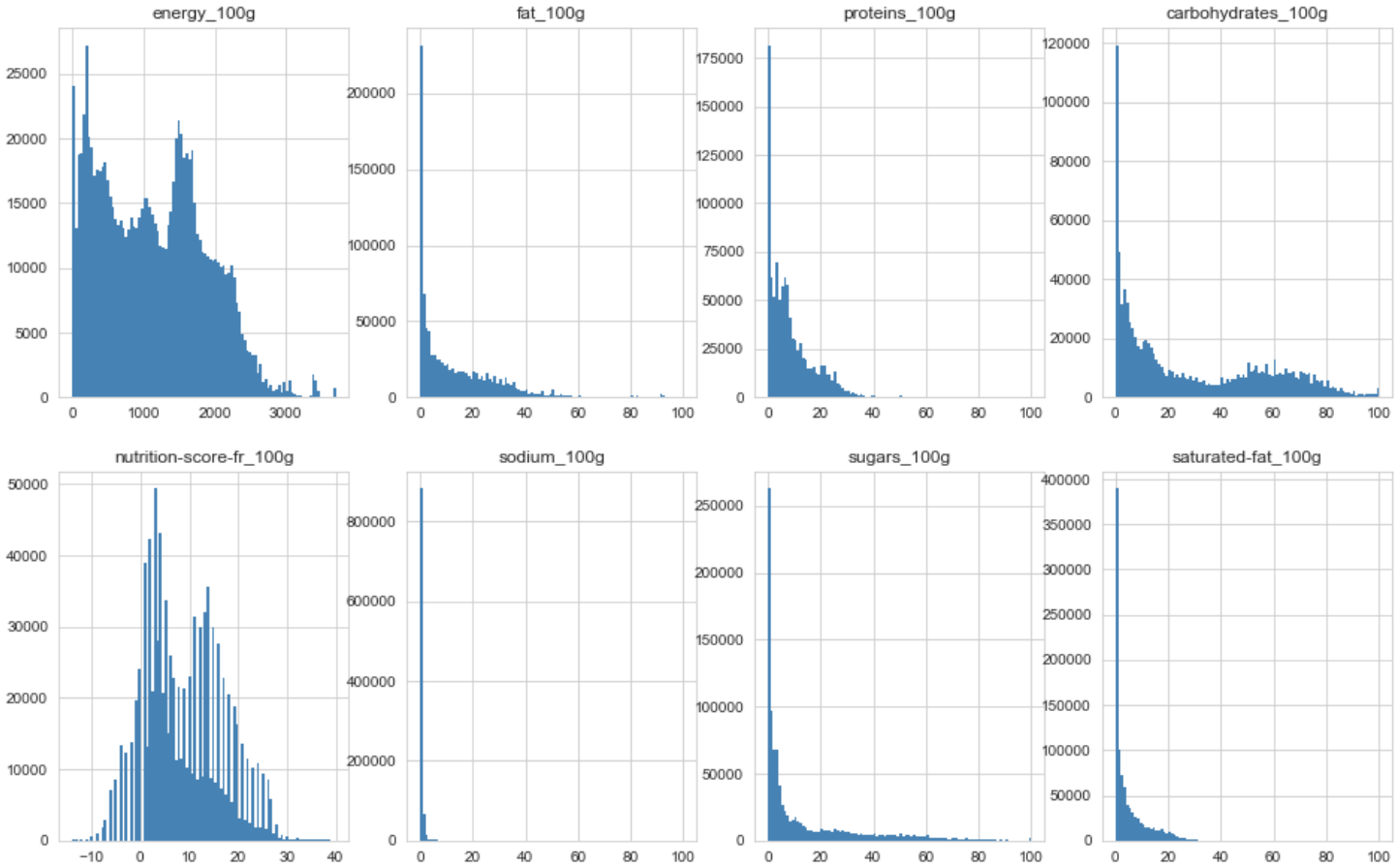
- Calcul de l’apport énergétique par rapport aux macronutriments*
- Calcul des classes nutritionnelles : A, B, C, D, E

*Cette information est présente dans le dataset mais il y a des incohérences. Cette donnée peut être exprimée en KJ ou Kcal.

Phase exploratoire

Analyse univariée (1)

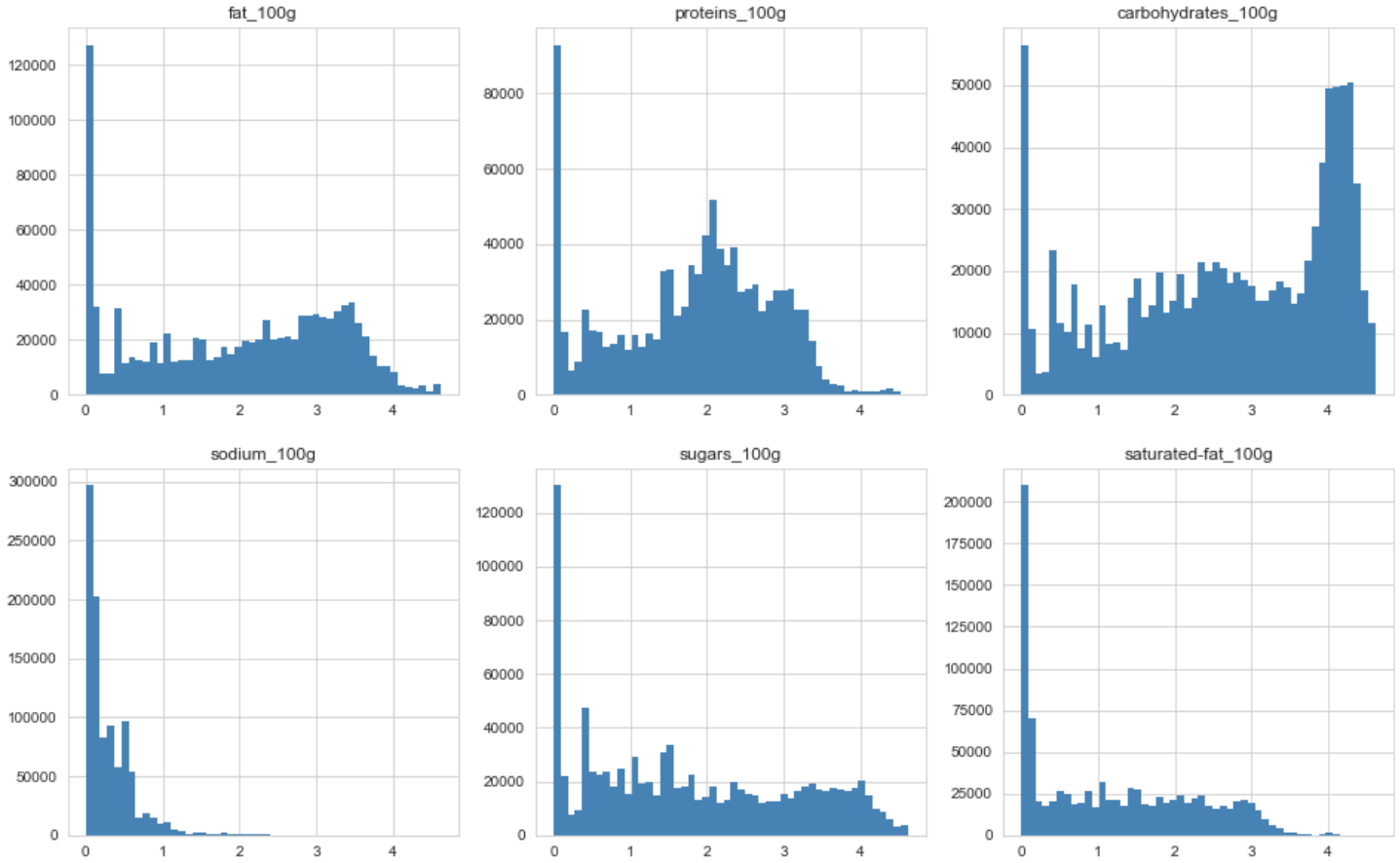
Distribution des valeurs numériques



- Les macronutriments et les micronutriments semblent avoir des distributions exponentielles

Analyse univariée (2)

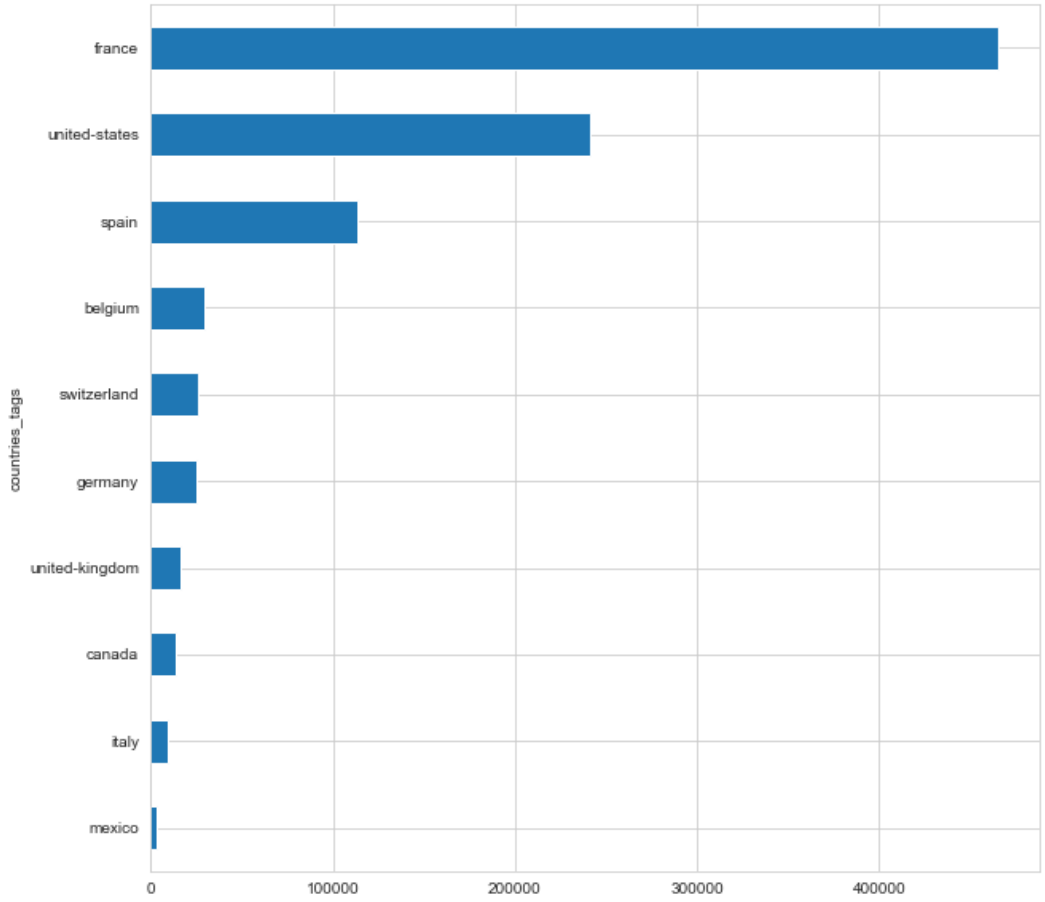
Distribution des valeurs numériques - transformation log(x+1)



- Mode à 0
- Transformation des données d'une distribution exponentielle vers une distribution normale
- Le terme « **x+1** » permet aux valeurs qui prennent 0 de rester à 0 après la transformation logarithmique

Analyse univariée (3)

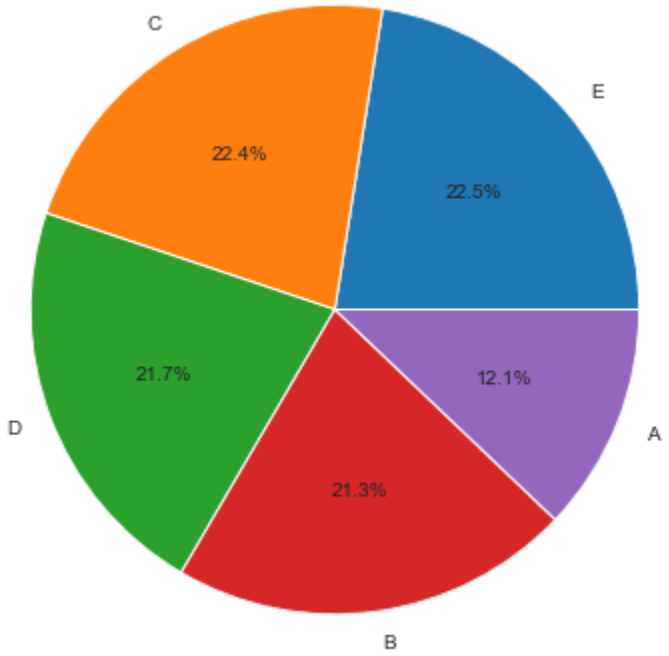
Pays d'origine



- Le pays d'origine le plus important des aliments de ce dataset est la **France**

Analyse univariée (3)

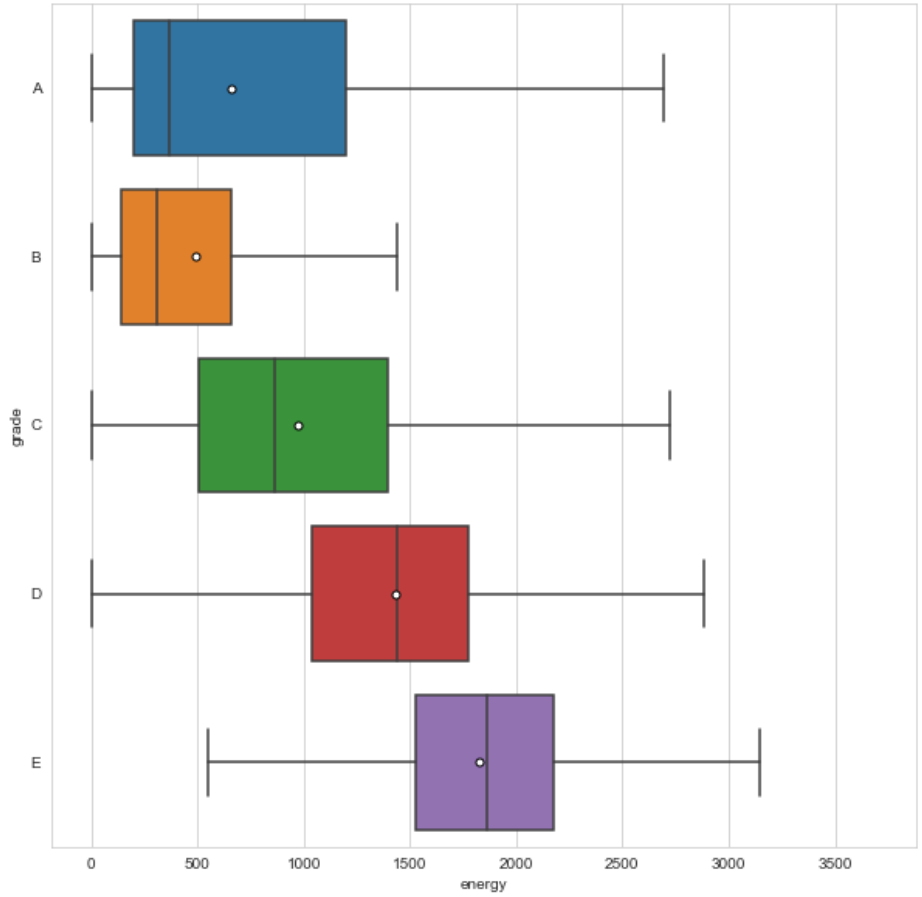
Proportion des nutri-scores



- Le groupe A est sous-représenté
- On peut émettre deux hypothèses
    1. Les utilisateurs d'**open food fact** ont tendance à uploader des aliments de faible qualité nutritionnelle
    2. Ou alors ces proportions sont représentatives du marché

Analyse bivariée

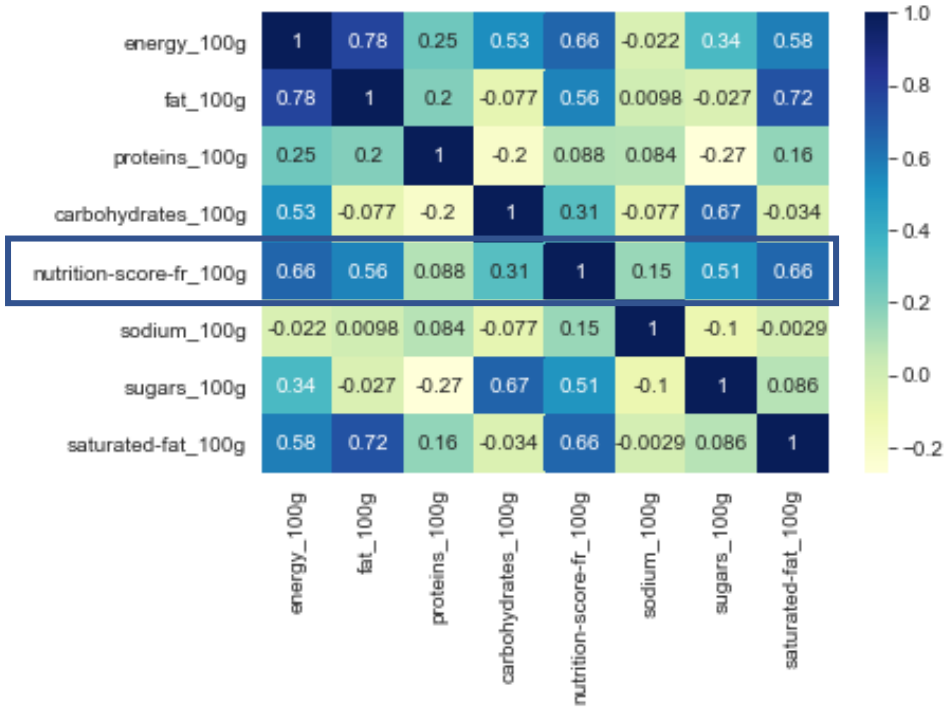
L'apport calorique par score



- Les éléments du boxplot : médiane, moyenne, Q1, Q3,
- L'apport calorique est l'une des composantes du nutri-score
- A l'exception du groupe A, on remarque que l'apport énergétique augmente au fur et à mesure que la note se dégrade
- Le groupe **A** a une **moyenne et une médiane supérieur** au groupe **B**
- Cette moyenne assez élevée est peut être due aux oléagineux, huiles de colza, de noix et d'olive qui sont très caloriques mais favoriser dans le score
- Moyenne > à la médiane dans les groupes A, B et C

Analyse bivariée

Corrélation entre les variables

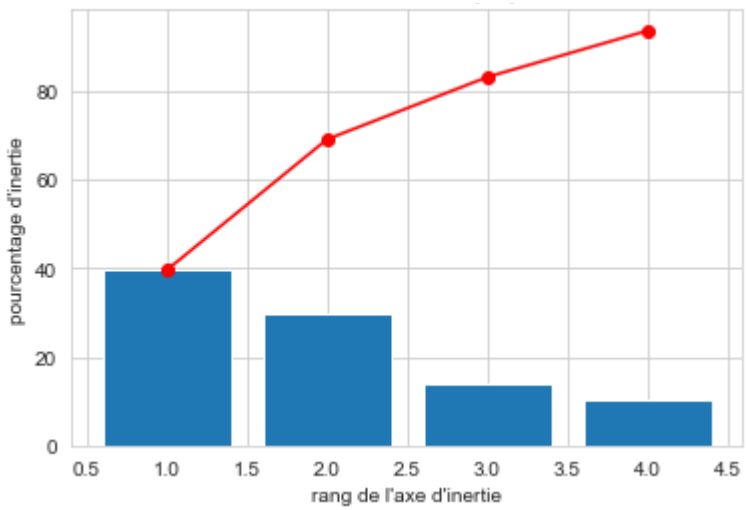


- Les variables qui influencent le plus le nutri-score :
  1. L'apport calorique
  2. Les graisses saturées
  3. Les graisses
  4. Le sucre

PCA

- Nombre de composants : 4

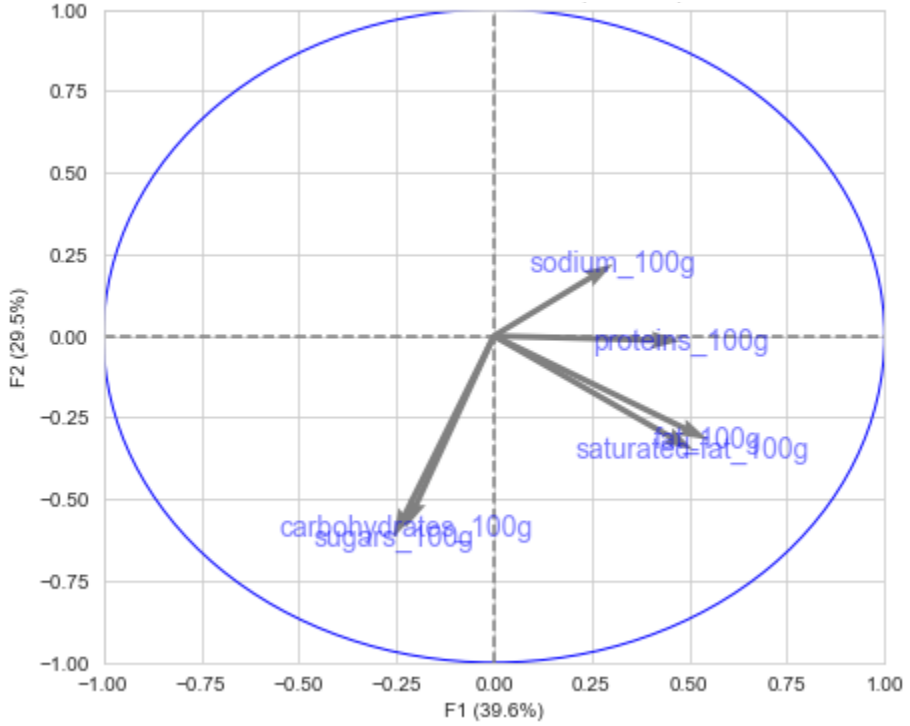
Eboulis des valeurs propres



- Le premier plan factoriel (F1 et F2) explique près de 75 % du dataset
- Je vais restreindre mon analyse à ces deux axes

PCA

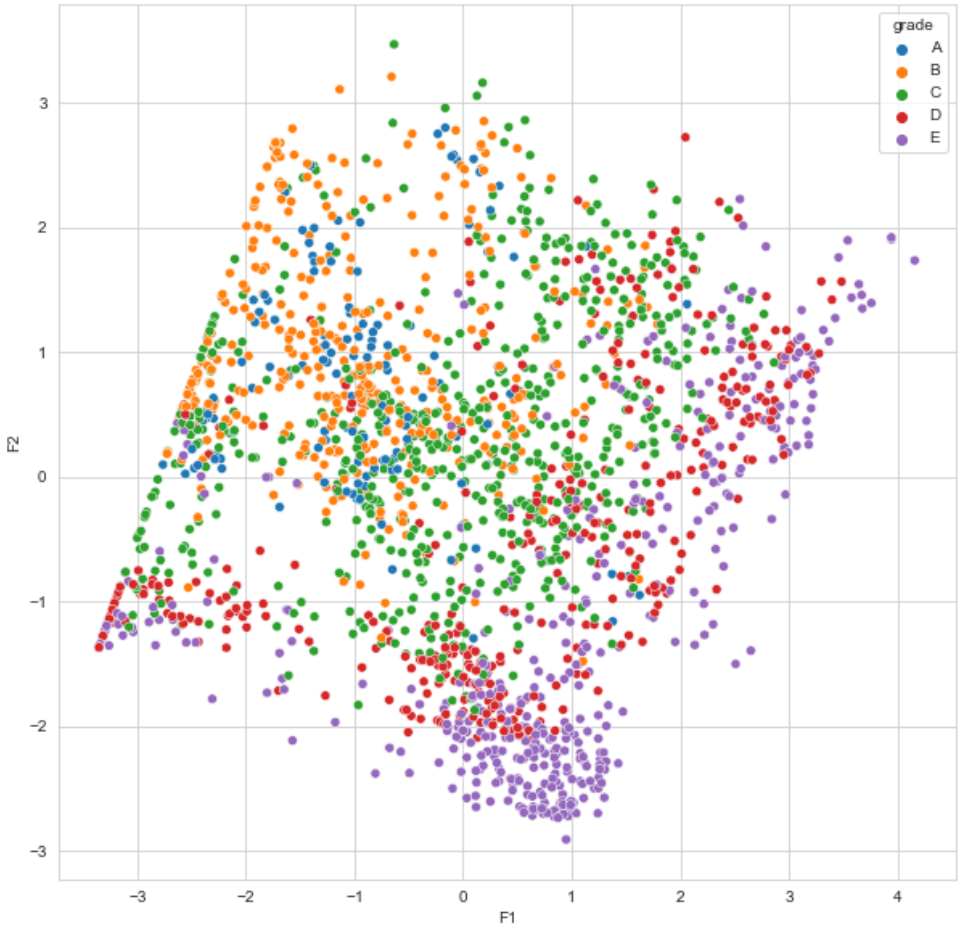
Cercle des corrélations (F1 et F2)



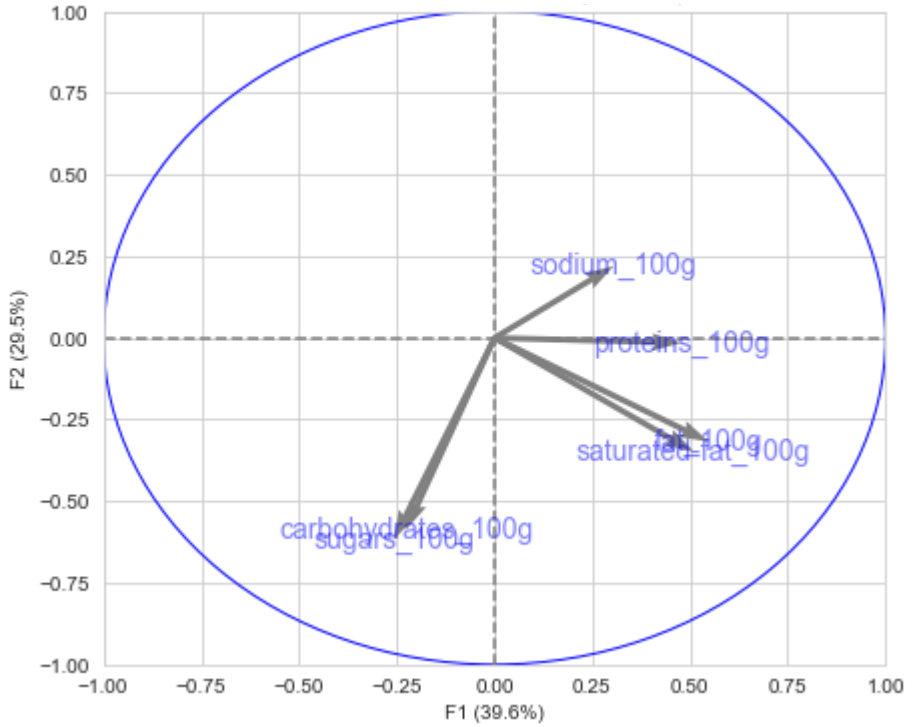
- Les longueurs des vecteurs indiquent que les graisses et les glucides sont les variables les mieux représentées
- Interprétation des axes:
  - F1 : les aliments sucrés et salés
  - F2 : l'apport calorique

PCA

Visualisation des clusters



Cercle des corrélations (F1 et F2)



ANOVA

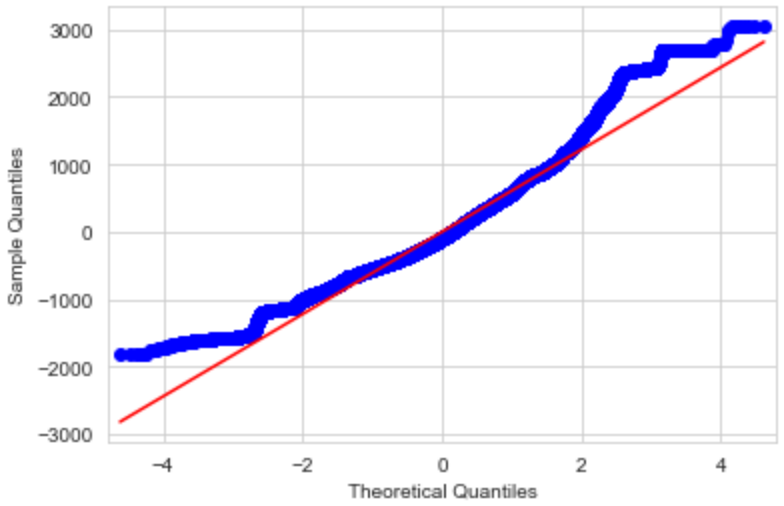
Problématique : l'apport calorique est-il réellement différent en fonction du nutri-score ?

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$
- H_1: Au moins l'une des moyennes est différente
- $R^2 = 0{,}339$
- F-statistic = 66827
- P-value $\simeq 0$

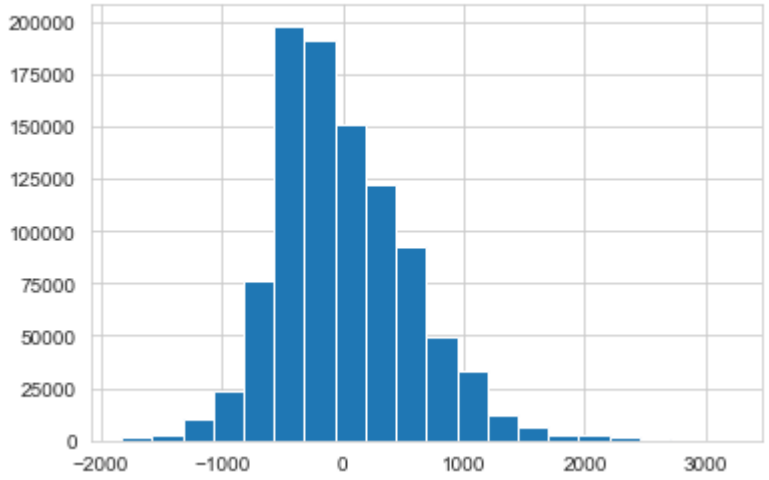
ANOVA : conditions d'application

La normalité des résidus

QQ plot des résidus



Histogramme des résidus



ANOVA : conditions d'applications

L'homoscédascitié

- H_0: les variances sont égales
- H_1: Au moins l'une des variances est différente
- Test de Levene
- Statistic = 3281,23
- P-value ≃ 0
- Rejet de H_0

Les données sont **i.i.d** :

Etant donné la nature du jeu de données, on peut supposer que les enregistrements soient indépendant et identiquement distribués.

ANOVA : conditions d'applications

Bilan

- ⊗ Normalité des résidus
- ⊗ Homoscédascité
- ✓ Echantillons i.i.d

- En principe, l'ANOVA peut être **robuste** à certaines de ces vérifications à condition que les groupes soient de tailles égales
- Cependant ici les groupes ont des tailles relativement différentes
- Je vais donc effectuer une ANOVA non-paramétrique

ANOVA non-paramétrique Kruskal Wallis

Test

- Les hypothèses sont inchangées
- Statistic = 427851
- P-value ≃ 0
- Analyse posthoc (dunn) : **toutes les p-values sont significatives**

Conditions d'applications

- ✓ Echelle ordinale de la variable dépendante (A, B, C, D, E)
- ✓ Echantillons i.i.d

On peut donc rejeter l'hypothèse nulle H_0, les moyennes ne sont pas toutes égales. De plus l'analyse post-hoc, nous montre que toutes les paires de variables sont différentes

Pertinence de l’application

Régression linéaire multiple (1)

Variables **indépendantes** :

- Graisse
- Protéines
- Glucides
- Sel
- Sucre
- Graisse saturée

Variable **à expliquer** :

- Nutri-score

Régression linéaire multiple (2)

1ere observation pertinente

R^2 ajusté = 73,1%

2e observation pertinente

Toutes les variables indépendantes utilisées dans ma régression multiple ont une **p-value** ≃ **0**. Elles sont donc toutes significatives

3e observation pertinente

RMSE = 4,59

RMSE (approche naïve : moyenne) = 8,89

Synthèse d'analyse

- Le jeu de données nous montre que les informations concernant les macronutriments, le sel, le sucre et les graisse transformées **sont très répandues** sur les étiquettes
- Ces informations **influent le plus** sur le nutri-score (*heatmap*)
- Le nutri-score sépare les aliments en plusieurs groupes ayant des apports caloriques significativement différents
- La régression linéaire multiple effectue des prédictions de qualité acceptable

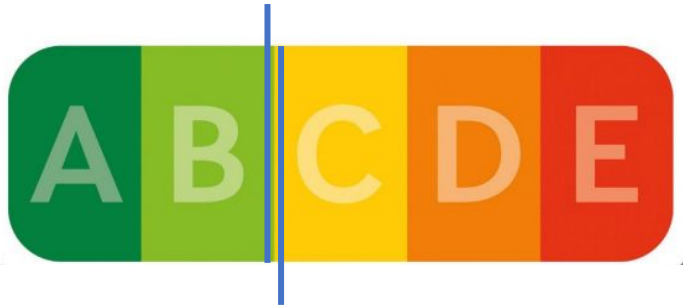
Conclusion

Il est possible de fournir un outil au consommateur pour lui permettre de connaître le nutri-score sur n'importe quel aliment qui comporte une étiquette nutritionnelle.

Représentation graphique :

Groupes différents malgré une estimation proche

Valeur réelle : -2



Estimation : -1,5

Alternative à la représentation classique

Score du produit