Concevez une application au service de la santé publique

Soutenance Projet 3
OpenClassrooms
Formation Data Scientist
07 Décembre 2022

Photo by Franki Chamaki on Unsplash

open FOOD facts

Base de données libre & ouverte

Image by Freepik

Idée d'application

DEGRÉ DE TRANSFORMATION − **VALEUR NUTRITIONNELLE** +

Produits non 'ultra-transformés'

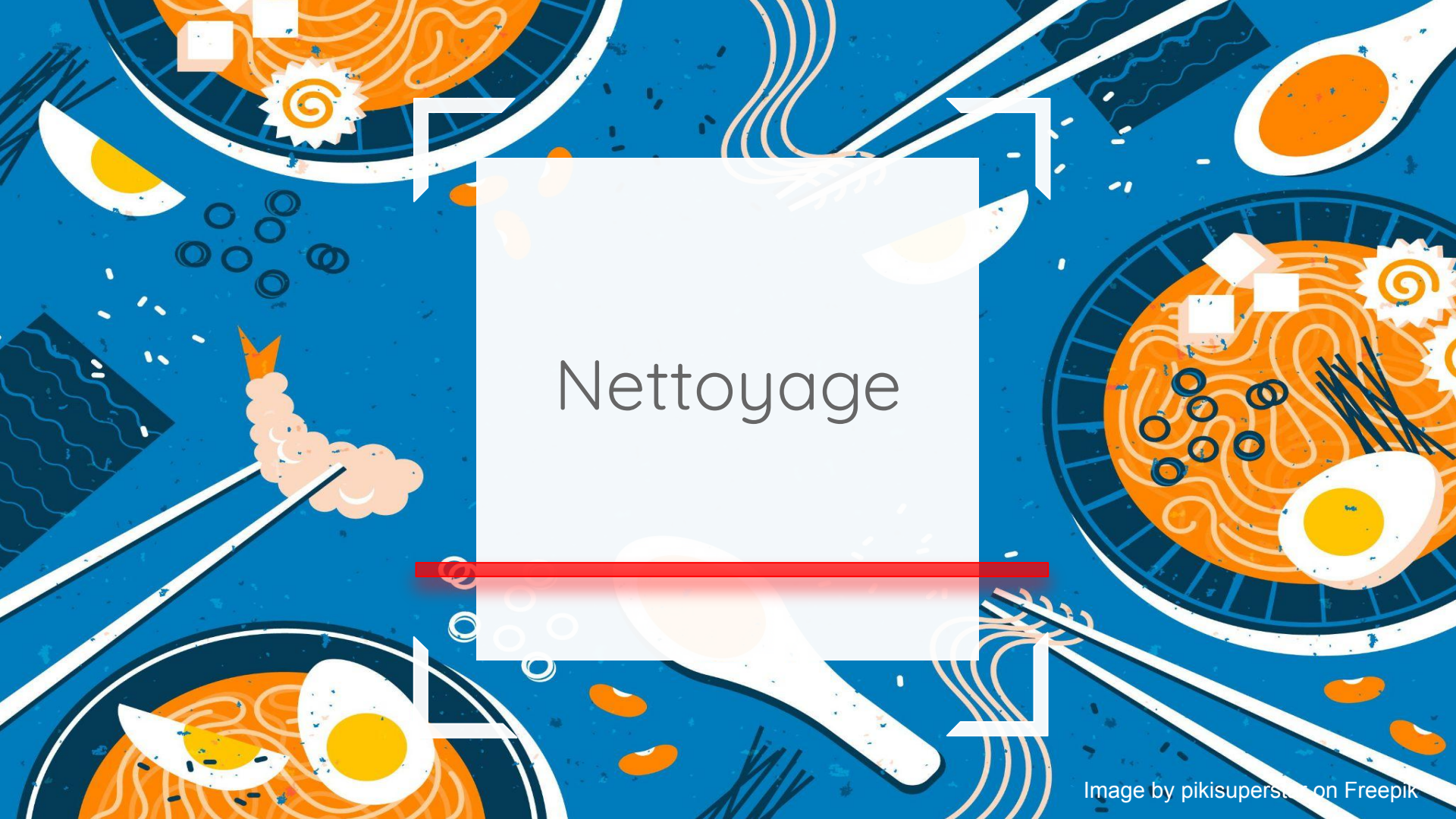
Produits avec une bonne valeur nutritionnelle

Plan de la soutenance

1. Nettoyage des données
2. Analyse des données
3. Conclusion

Image by Freepik and icons8

Nettoyage



Nettoyage

1. Sélection variables

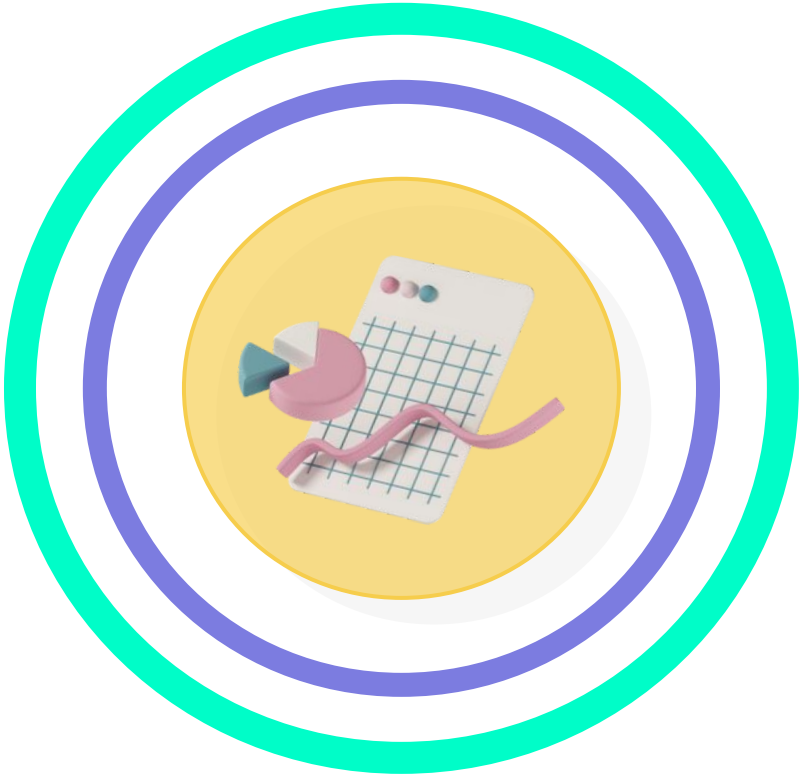
2. Doublons

3. Valeurs aberrantes

4. Valeurs manquantes

Sélection des variables

Sélection variables

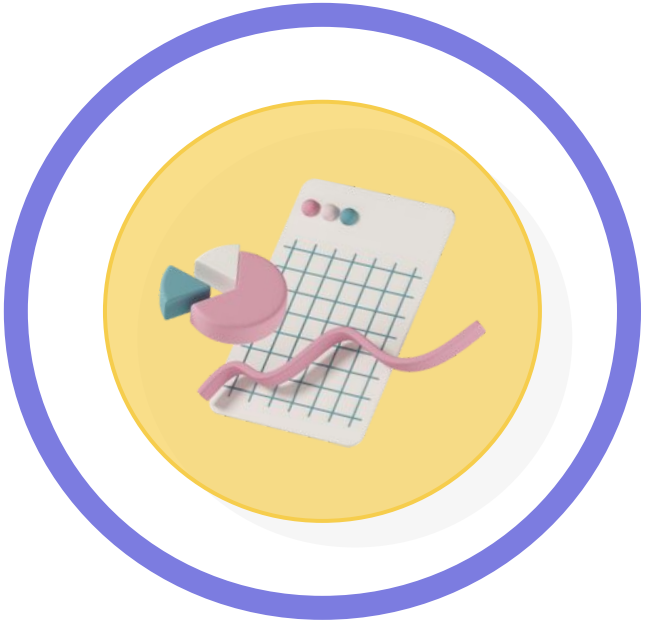


Lignes : 320 772

Colonnes : 106

Illustration by Icons 8 from Ouch!

Sélection variables



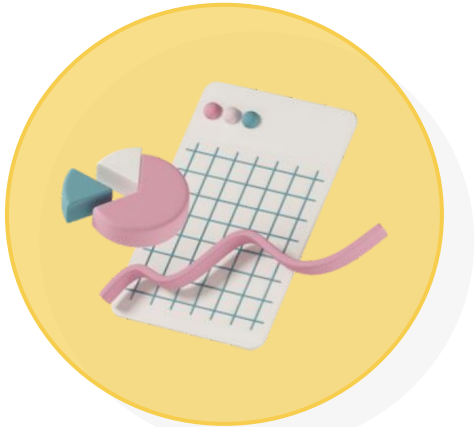
Lignes : 320 772

Colonnes : 54

* Filtre : + 80% valeurs manquantes

Illustration by Icons 8 from Ouch!

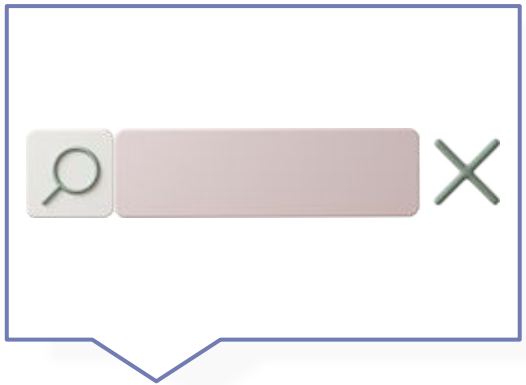
Sélection variables



Lignes : 320 772

Colonnes : 20

Illustration by Icons 8 from Ouch!



Clé de recherche

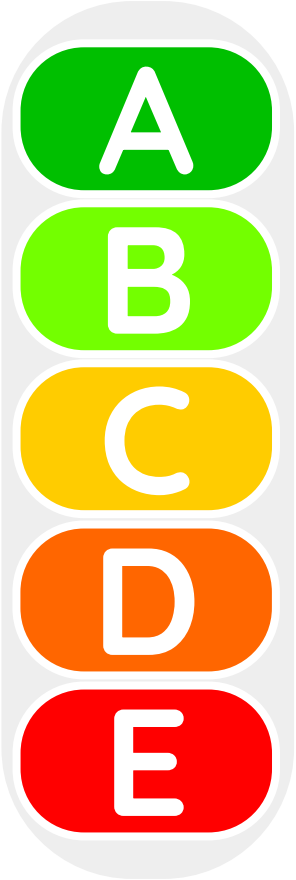
	Valeurs manquantes
Code-barres	0.01 %
Nom du produit	6 %

Qualité des produits

Valeurs manquantes

Additifs	22 %
Nutriscore/Nutrigrade	31 %

NUTRI-SCORE



-15

40

À limiter :
sel
gras
sucre
énergie

À favoriser :
fibres
protéines
fruits/légumes/noix

Familles d'aliments

Valeurs manquantes

Catégories (pnns)	72 %
Sous-catégories (pnns)	71 %

Autres (analyse)

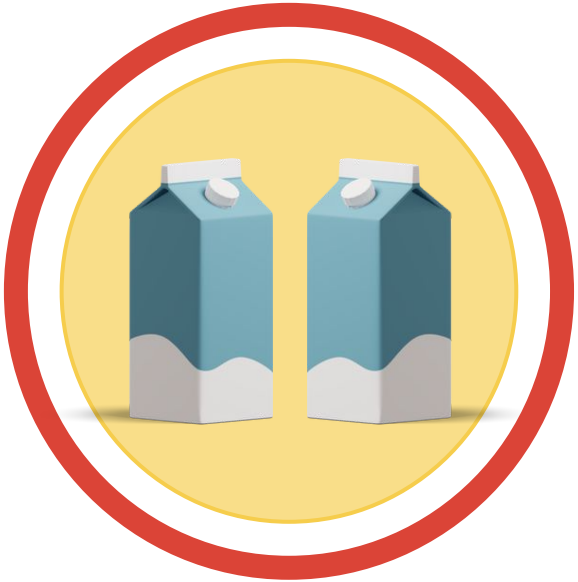
	Valeurs manquantes
Dates (création et modification)	0 %
Pays de vente	0.09 %
Contributeur	0 %
Image	76 %

Traitement des doublons

Traitement des doublons

* Clé d'unicité : code-barres

* Filtre : + rempli + récent



Lignes : 302 987

Doublons : 158

Illustration by Icons 8 from Ouch!

Traitement des valeurs aberrantes

Composantes nutri-score

-100, -5, _ _ _ < MIN _ _ _ _ _ _ MAX < _ _ _ _ _

* Imputation : moyenne sous-catégorie

Lignes : 302 829

Valeurs aberrantes : 723

Nutri-score

- [-15 : 40]
- Score et lettre correspondent

* Supprime

Lignes : 302 211

Valeurs aberrantes : 3 260

Traitement des valeurs manquantes

Fibres

* Imputation : 0

Lignes : 298 951

Valeurs manquantes : 45 080

Illustration by Icons 8 from Ouch!

Autres composantes

* Imputation : moyenne
(sous-catégorie)

Lignes : 298 951

Valeurs manquantes : 83 956

Illustration by Icons 8 from Ouch!

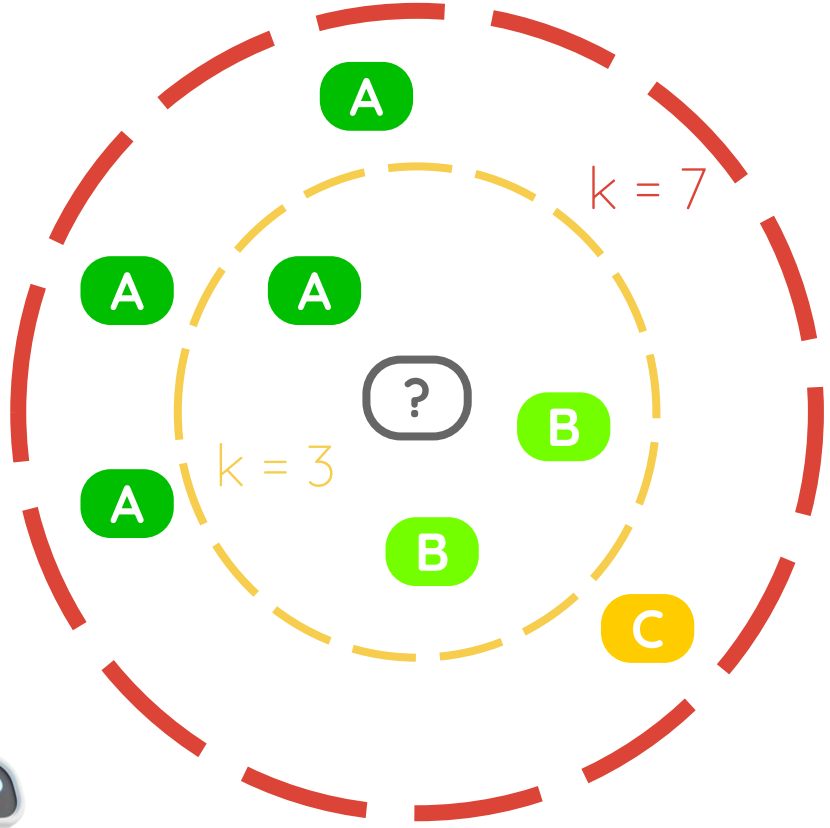
Nutriscore

* Imputation : k-NN

Lignes : 235 562

Valeurs manquantes : 84 201

Illustration by Icons 8 from Ouch!

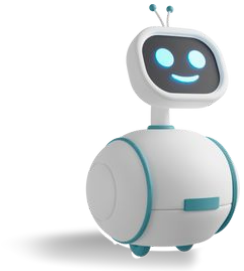


K-NN

k = 15

manhattan

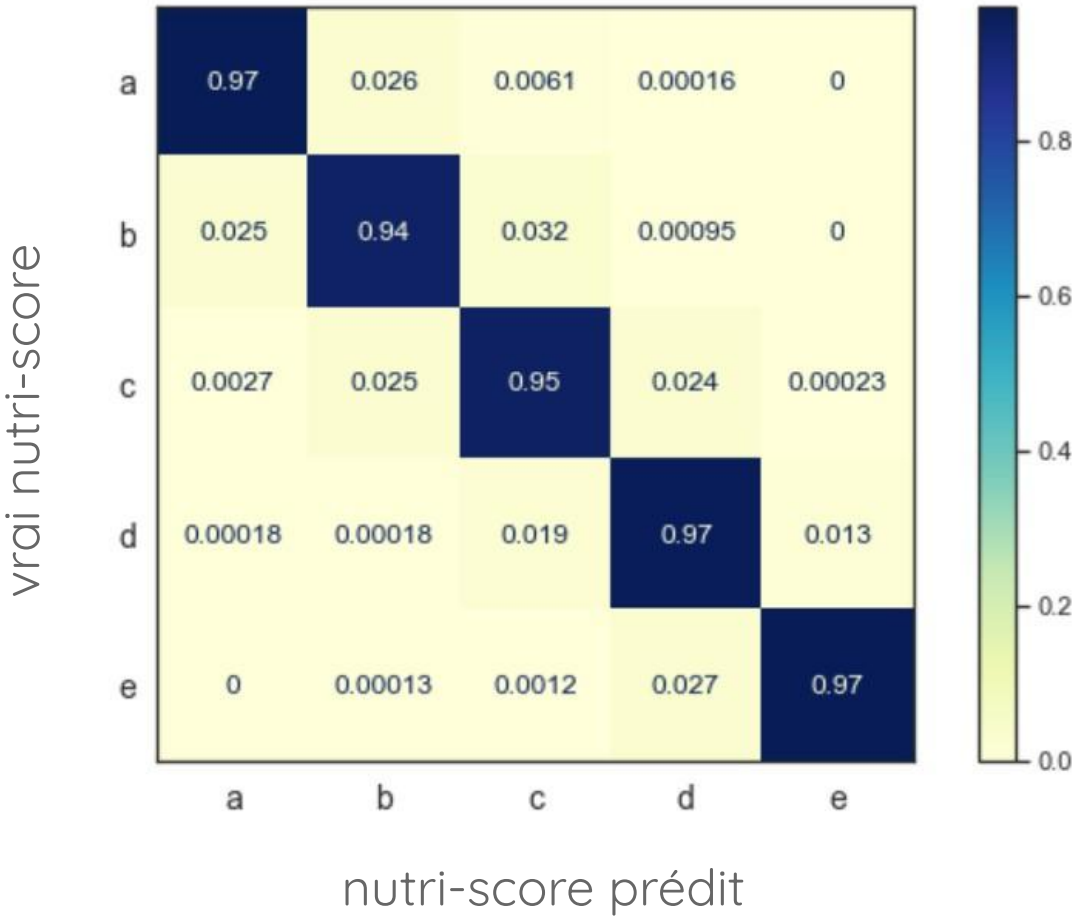
pondération



Matrice de confusion (normalisée)

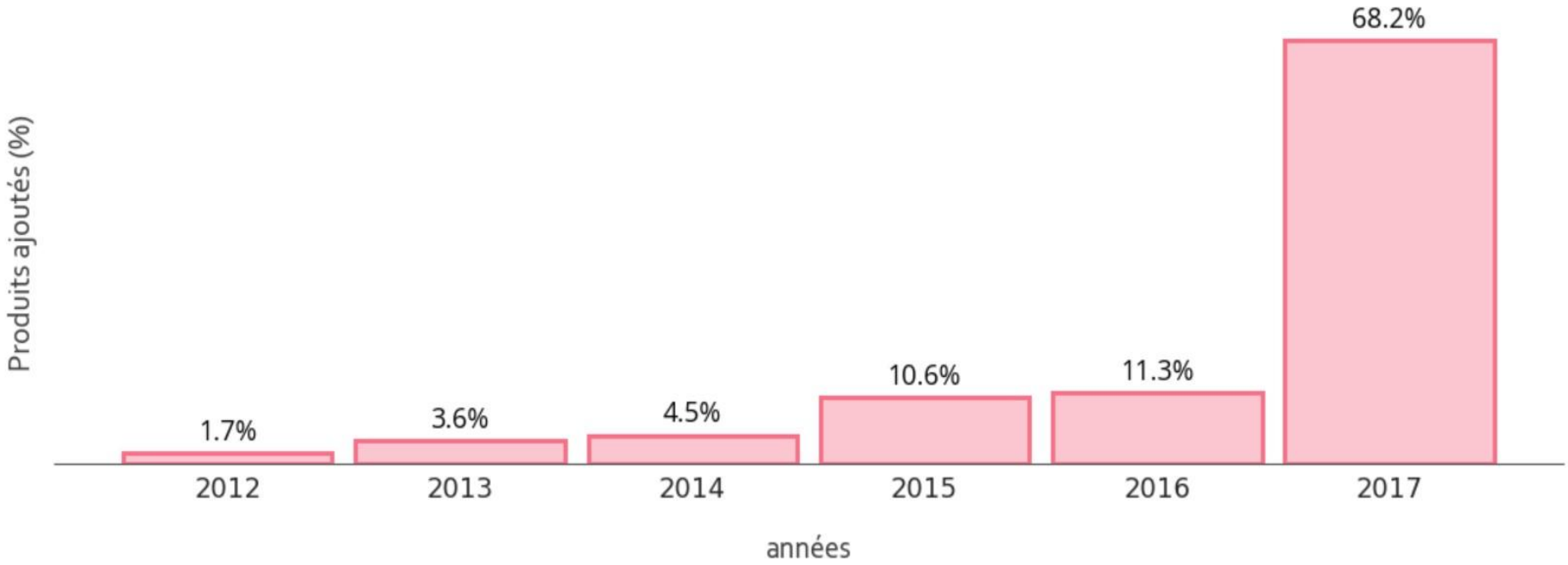
96%

*Prédictions correctes (total)

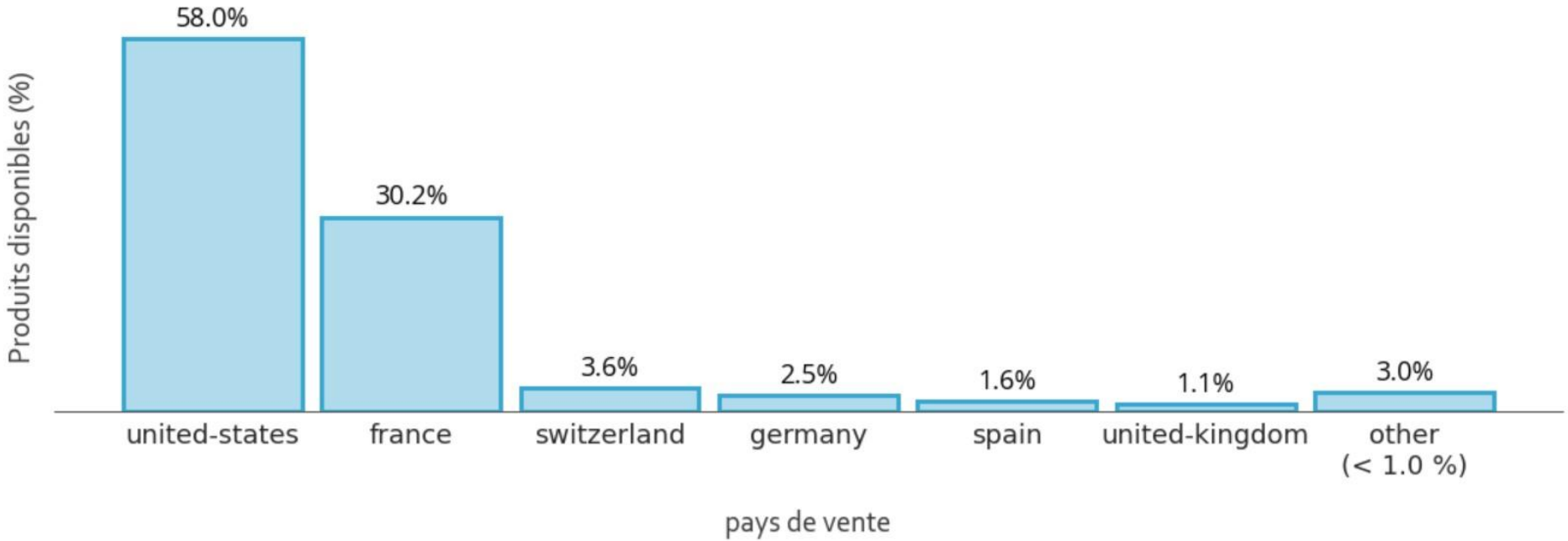


Analyse exploratoire

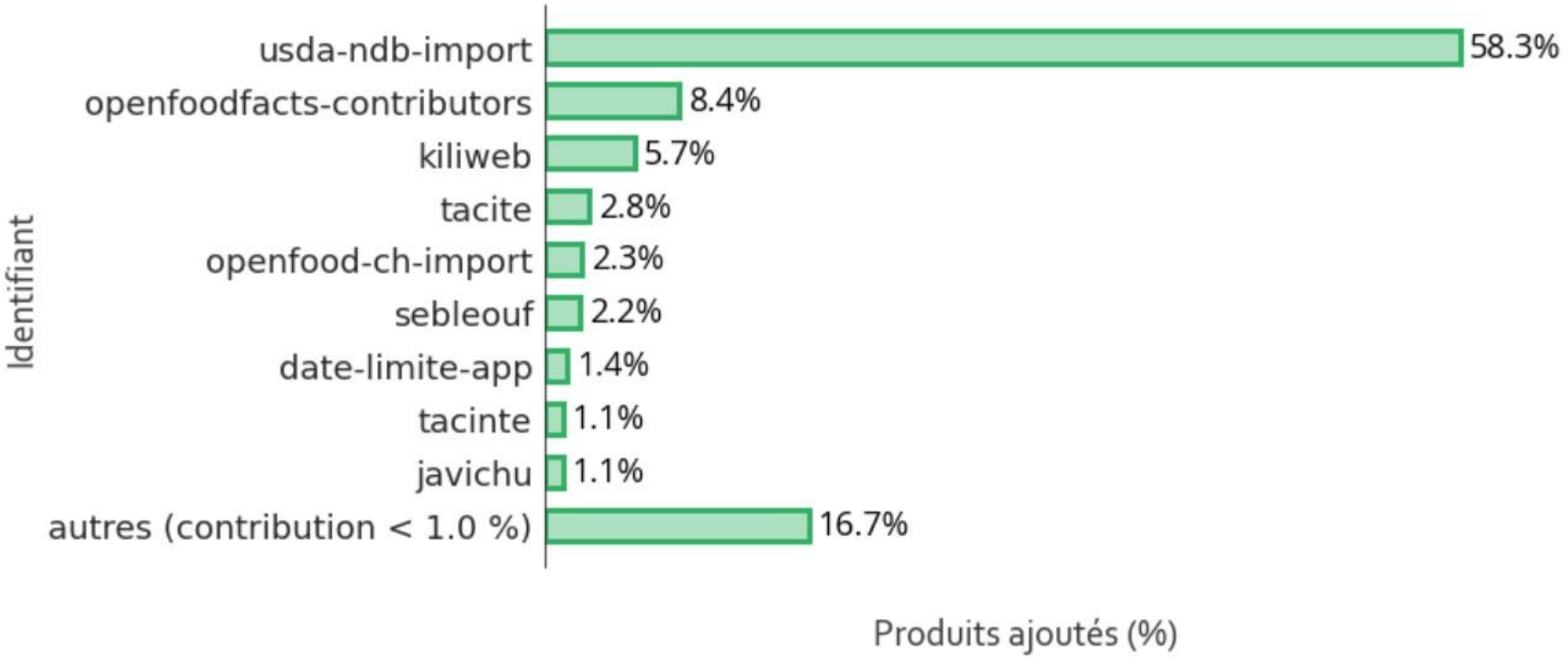
Produits ajoutés par an sur Open Food Facts (%)



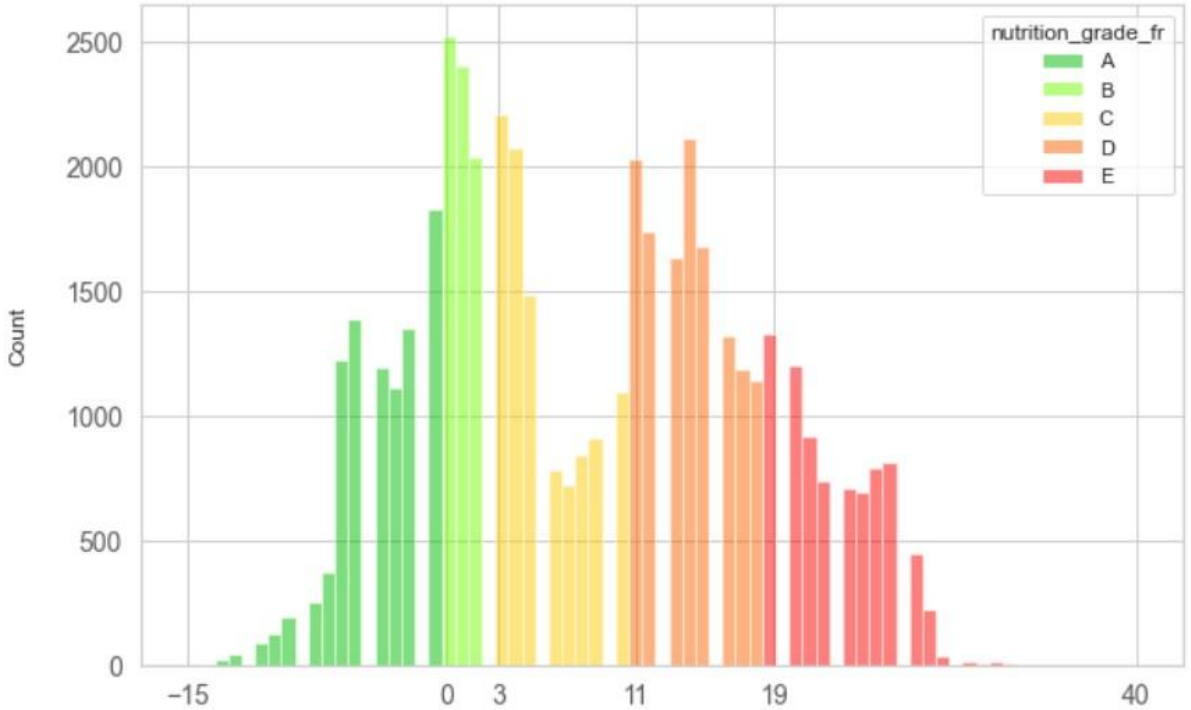
Produits en vente par pays (%)

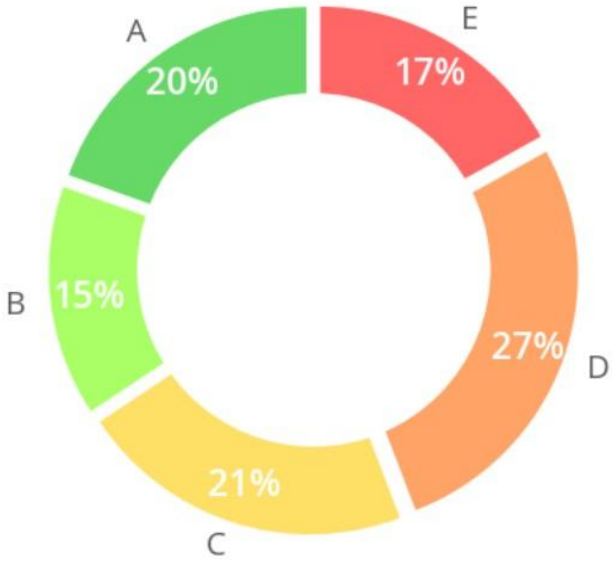


Produits ajoutés par contributeur (%)



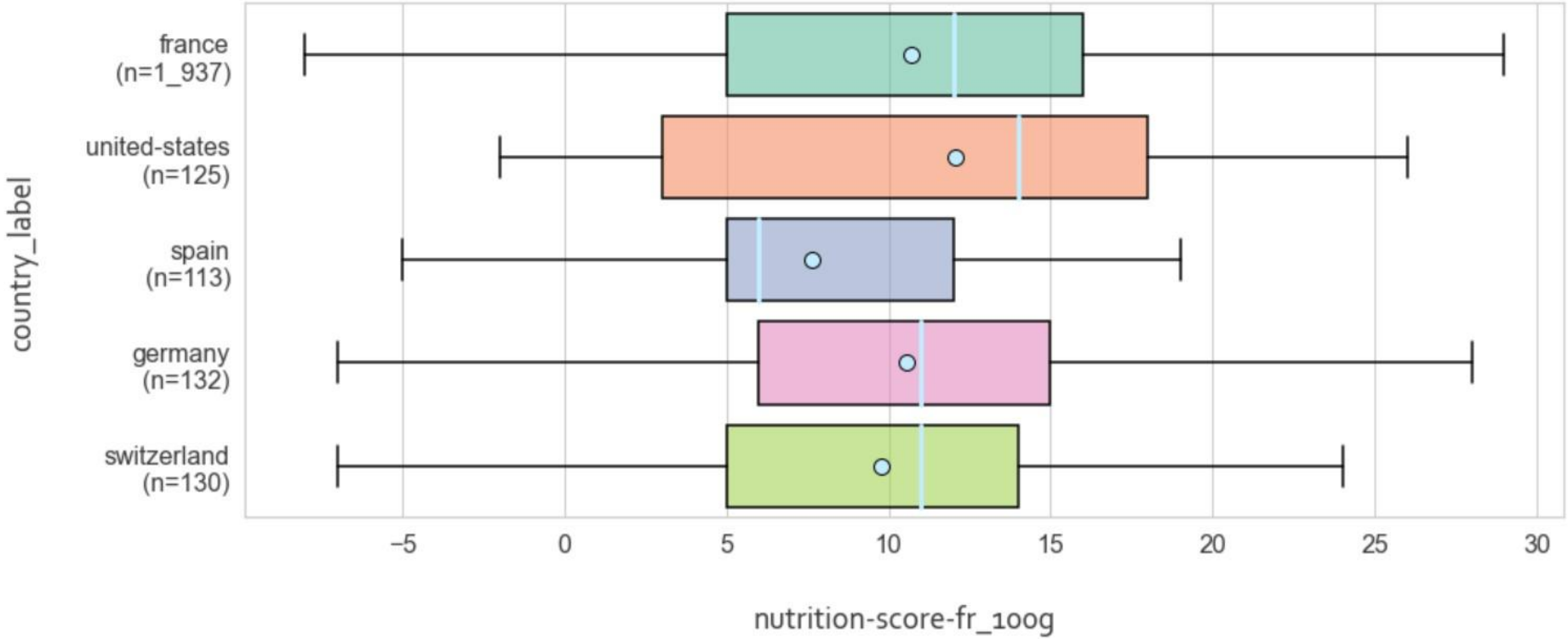
Distribution empirique des nutri-scores





* Répartition des nutri-grades

Nutri-scores par pays (catégorie sauces)



ANOVA (test statistique)

*Niveau Test : 1%

H0 : facteur géographique n'a aucune influence sur la valeur nutritionnelle des produits

H1 : facteur géographique a une influence sur le nutri-score

Veuillez entrer le code-barres :
3250391969098.0

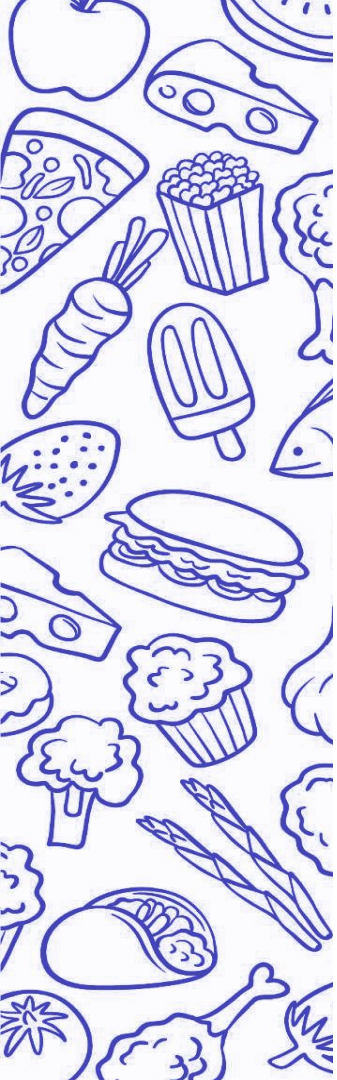
Frutimax aux bons morceaux de fruits

C

Illustration by Icons 8 from Ouch!

Meilleures Alternatives (milk and yogurt)

NutriScore	Nom du produit	Image
A	Yaourts 0% mg, aux fruits avec morceaux, fraise-ceris-pêche-ananas-poire-pruneau	
A	Panier de yoplait aux bons fruits 0%	
A	Spécialité Laitière aux Fruits avec morceaux	



Conclusion :

- données anciennes
- données manquantes
- fiabilité douteuse

- sous-ensemble cohérent
- “application” fonctionne

Annexes

Image by pikisuperstar on Freepik

Plan des Annexes

1. ACP
    a. éboulis des valeurs propres
    b. cercle des corrélations
    c. matrice de corrélation

2. kNN méthodes de calcul distance

3. Analyses bivariées
    a. tableau de contingence
    b. ViolinPlot

4. Analyses univariées
    a. BoxPlot

Analyse en composantes principales

variables synthétiques

Facteur 1

Facteur 2

variables initiales

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5

Illustration by Icons 8 from Ouch!

Eboulis des valeurs propres

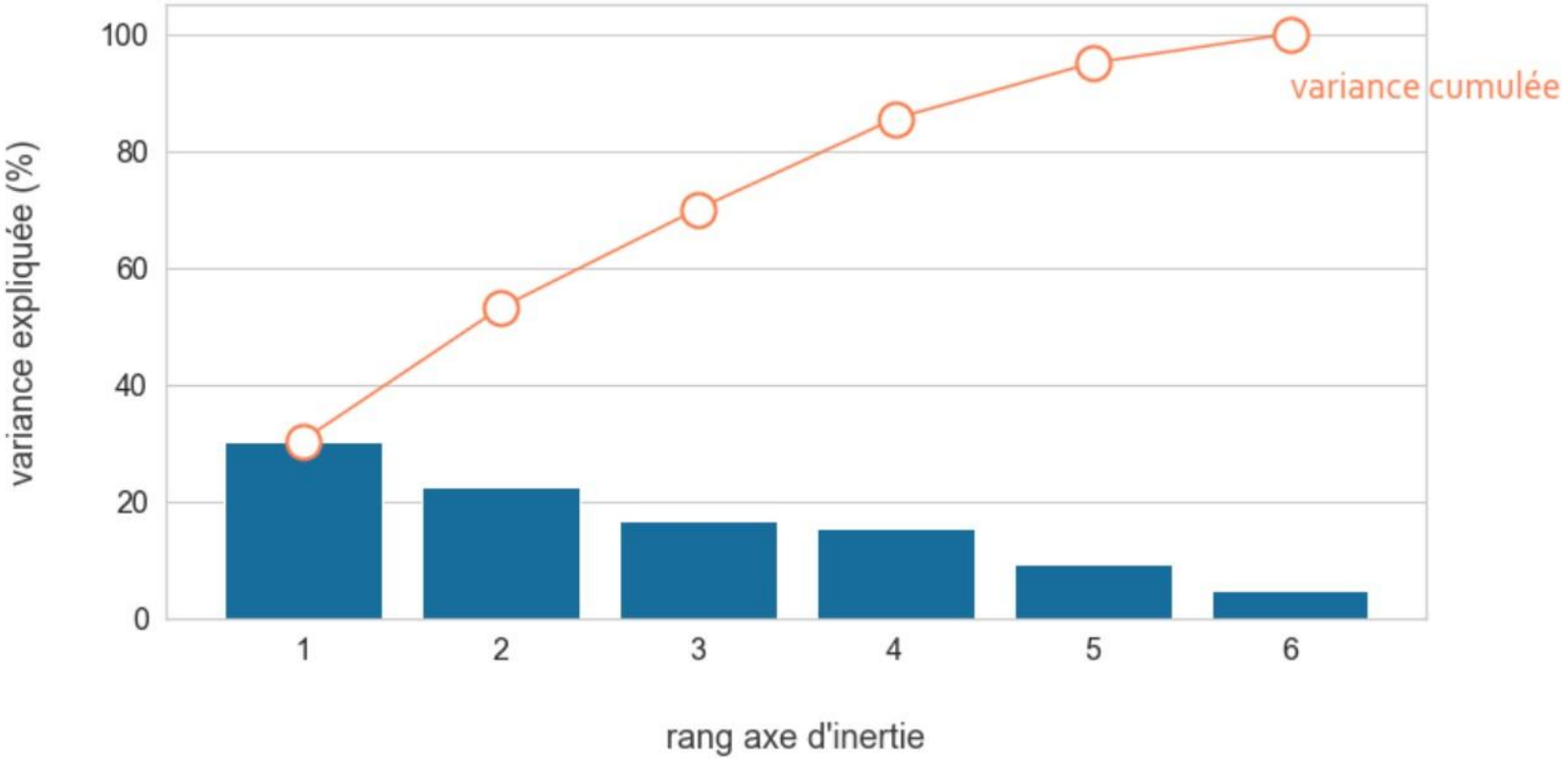


Image by Freepik

Cercle des corrélations

* 1er plan factoriel

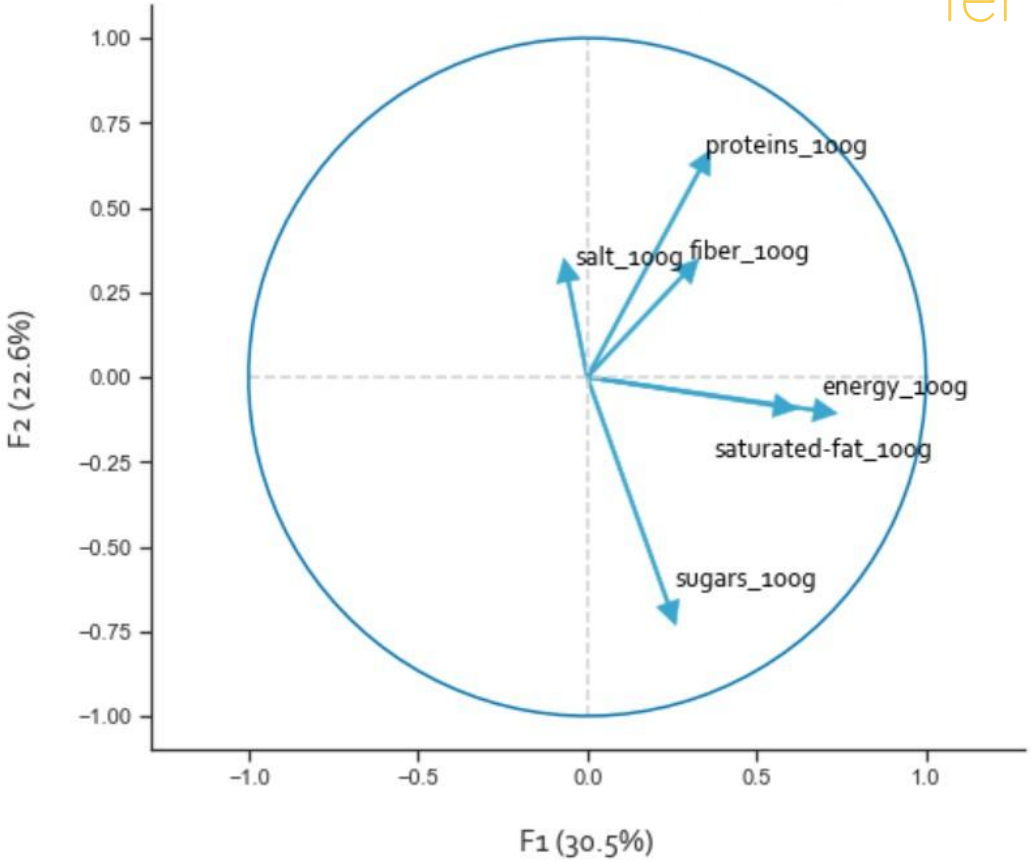
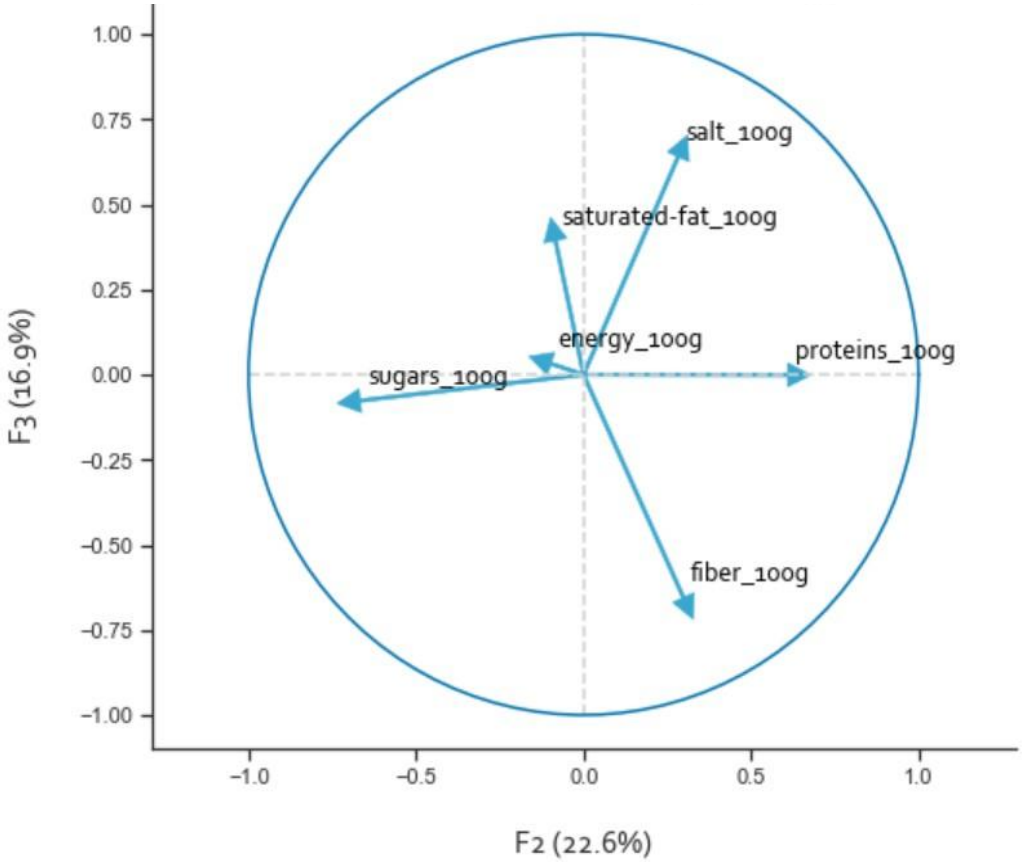


Image by Freepik

Cercle des corrélations



* F2/F3

Image by Freepik

Matrice de corrélation

* Composantes nutri-score

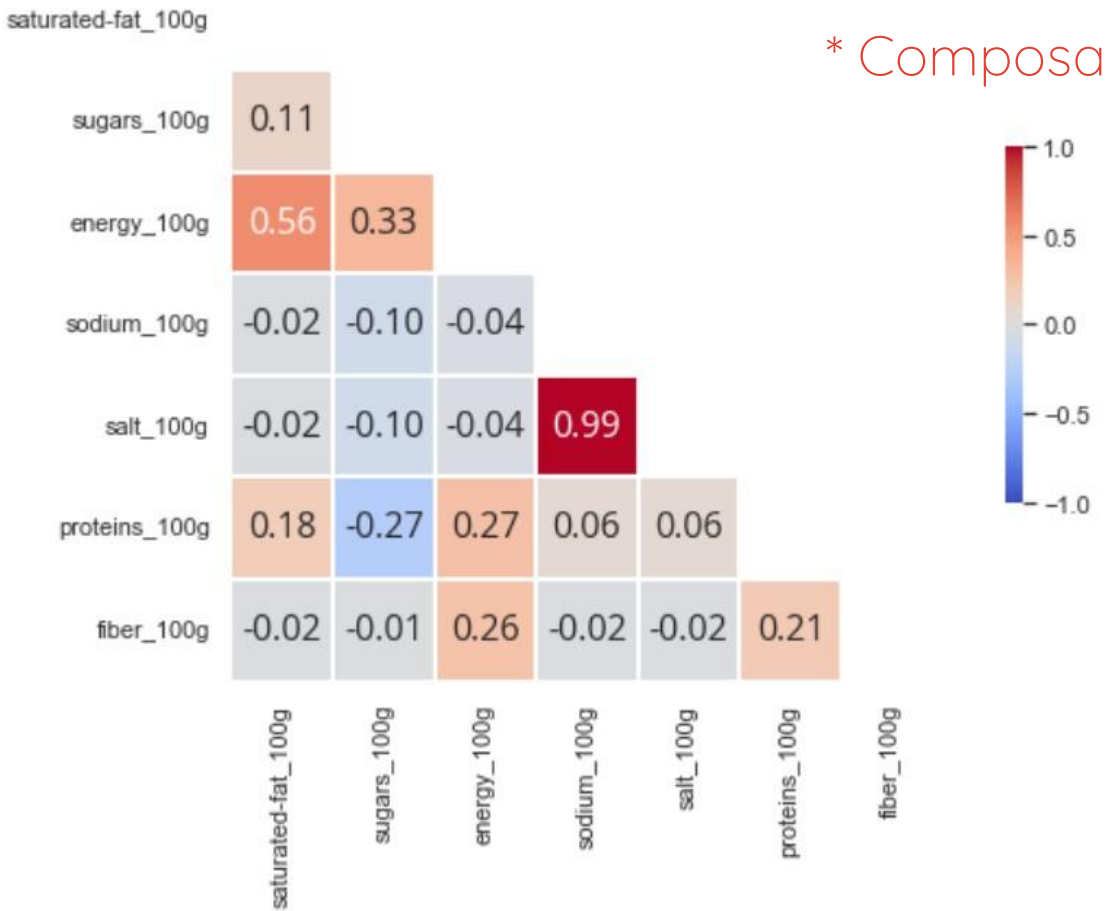


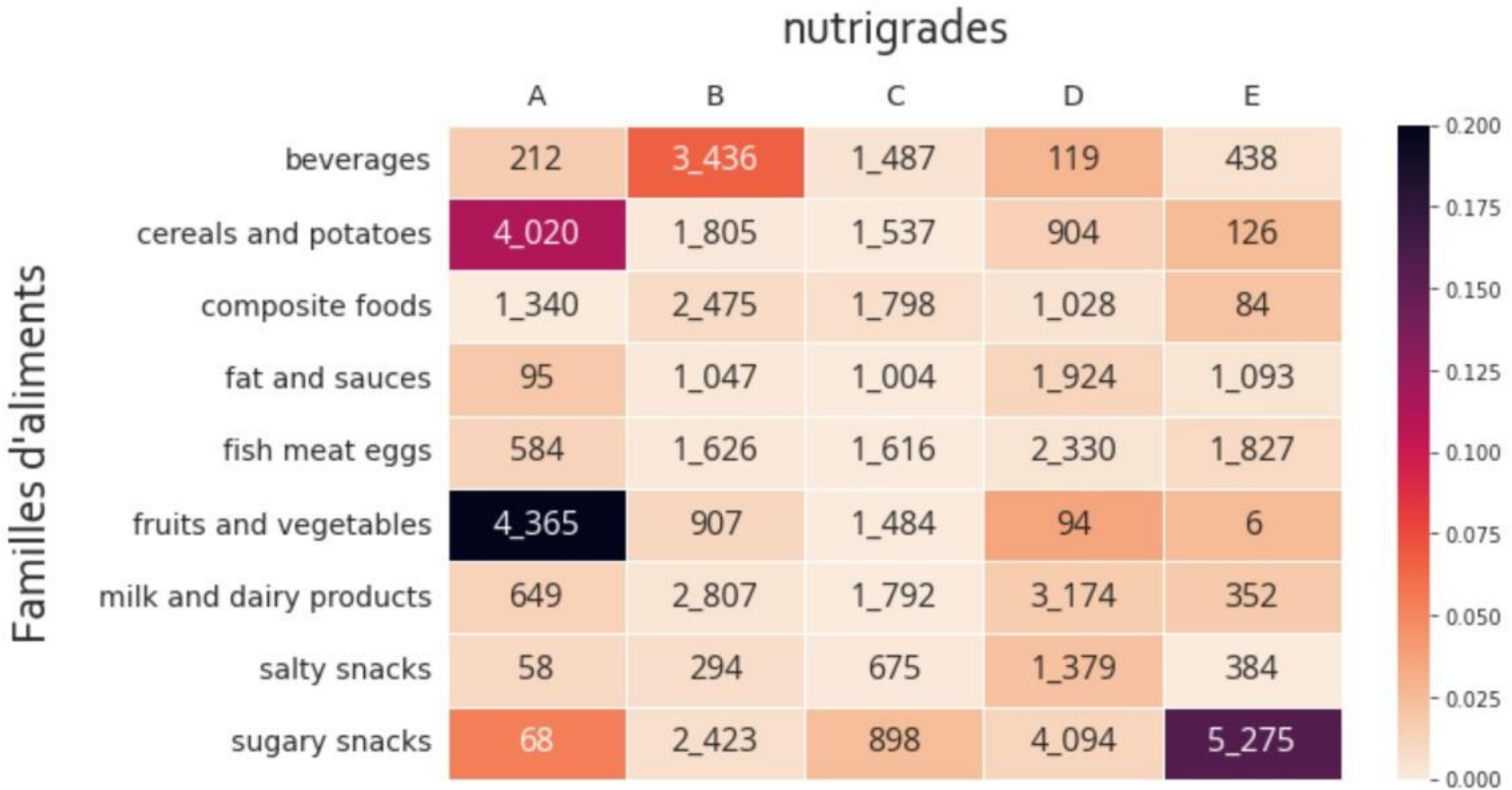
Image by Freepik

Métriques (kNN - sklearn) :

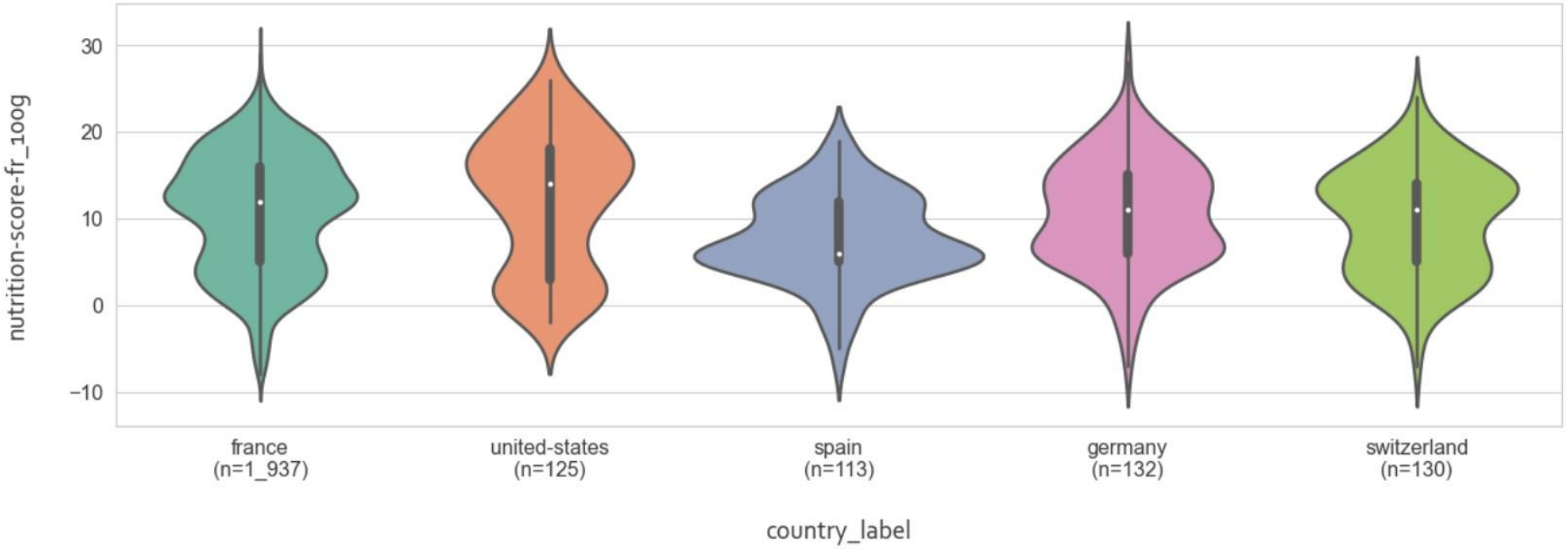
- Euclidean : $\sqrt{\Sigma(xi-yi)^2}$
- Manhattan : $\Sigma|xi-yi|$

Illustration by Icons 8 from Ouch!

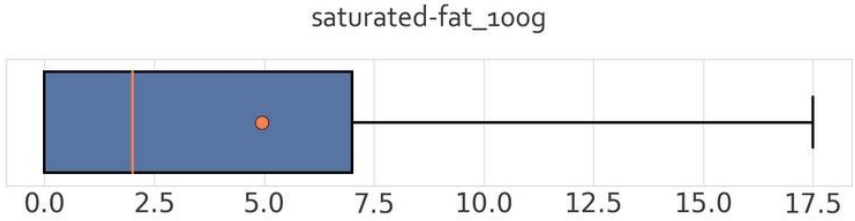
Tableau de contingence

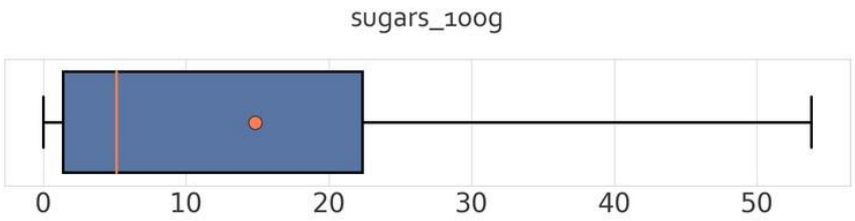


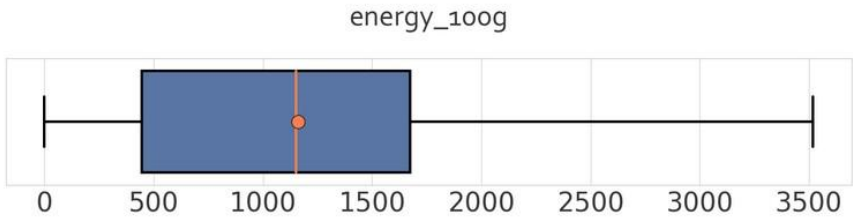
Nutri-scores par pays (catégorie sauces)

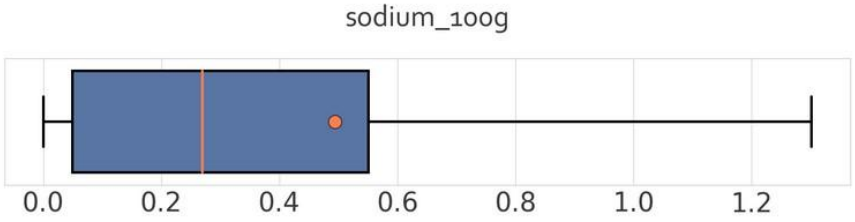


Boxplot composantes du nutri-score :









Boxplot composantes du nutri-score :