
Addiscore - Une application au service de la santé publique

Projet 3 du parcours « Data Scientist » d'OpenClassrooms

Mark Creasey

Sommaire

L'agence "Santé publique France"
a lancé un appel à projets pour
trouver des idées innovantes
d'applications en lien avec
l'alimentation.



- 1. Idée de l'application**
 - *Simplifier le choix de plats préparés*
- 2. Les données**
 - *Importation et nettoyage*
- 3. Les contributeurs aux données**
 - *Qui, comment, quand, quoi, pourquoi*
- 4. Les additifs**
 - *Catégorisation, répartitions entre aliments*
- 5. L' ADDISCORE**
 - *Comparaison avec d'autres indicateurs*
- 6. Conclusion**
 - *Axes d'améliorations*

01 Addiscore - Idée de l'application

Simplifier le choix des plats préparés

Carpaccio de viande bovine marinée et sa dosette



Carpaccio parmigiano reggiano aop - Auchan - 197g



Carpaccio de boeuf - L'etal Du Boucher - 210 g



Carpaccio de boeuf - LIDL



(NL) Rundercarpaccio, met dressing en Parmigiano Reggiano op basis van rauwe melk, 30+. Ingrediënten: Carpaccio (2 x 50 g): 95% rundvlees, zout, voedingszuren: kaliumlactaat, natriumacetaten, natriumcitraat; antioxidanten: natriumerythorbaat, natriumcorbaat; stabilisatoren: trifosfaten, polyfosfaten; conservermiddelen: natriumnitriet, kaliumnitraat; natuurlijk aroma. Dressing (10 ml): water, plantaardige oliën (raapzaad, olijf, in wisselende verhoudingen), azijn, suiker, 0,2% witte wijn, zout, gemodificeerd zetmeel (maïs), kruiden, specerijen, gedroogde groenten (ui, knoflook, spinazie), verdikkingsmiddelen: xanthaangom, guarapitmeel. *op eindproduct. Parmigiano Reggiano (10 g): rauwe melk, zout, dierlijk stremsel. Gemiddelde voedingswaarde/100 g: energie 514 kJ/122 kcal; vetten 4,3 g; waarvan verzadigde vetzuren 2,4 g; koolhydraten 1,0 g; waarvan suikers 0,6 g; eiwitonen 21 g; zout 2,1 g. Gekoeld bewaren bij max. +7°C. Na openen beperkt houdbaar. Verpakt onder beschermende atmosfeer. **(Bevat alcohol)**

(FR) Carpaccio de bœuf, avec dressing et Parmigiano Reggiano à base de lait cru, 30+. Ingrédients: Carpaccio (2 x 50 g): 95% viande de bœuf, sel, acidifiants: lactate de potassium, acétates de sodium, citrates de sodium; antioxydants: érythorbate de sodium, ascorbate de sodium; stabilisateurs: triphosphates, polyphosphates; conservateurs: nitrite de sodium, nitrate de potassium; arôme naturel. Dressing (10 ml): eau, huiles végétales (navette, olive, en proportion variable), vinaigre, sucre, 0,2% vin blanc, sel, amidon modifié (maïs), plantes aromatiques, épices, légumes déshydratés (oignon, ail, épinards), épaississants: gomme xanthane, gomme guar. *rapporté au produit fini. Parmigiano Reggiano (10 g): lait cru, sel, préasure. Valeurs nutritionnelles moyennes/100 g: énergie 514 kJ/122 kcal; matières grasses 4,3 g; dont acides gras saturés 2,4 g; glucides 1,0 g; dont sucres 0,6 g; protéines 21 g; sel 2,1 g. À conserver au réfrigérateur à +7°C max. Durée de conservation limitée après ouverture. Conditionné sous atmosphère protectrice. **(Contient de l'alcool)**

Ongewond, ten minste houdbaar tot: / Non consommer de préférence avant le:

NL
114
EG

Geproduceerd in Nederland door: / Élaboré aux Pays-Bas
Menken vleeswarensnijlijn b.v., Edisonstraat 2, 2171 TV, Sassenheim

Total/ 120 g
Carpaccio: 100 g e
Dressing: 10 ml
Kaas/Fromage: 10 g



Carlos
Camionneur

Choisir des plats préparés est trop compliqué

Besoin d'un **indicateur des risques relatifs** pour sa santé.



Le public ciblé et ses besoins

Plats préparés

(NOVA groupes 3 et 4)

pas le temps, l'espace, les moyens ou
l'envie de cuisiner

Pas cher

pas l'argent pour manger bio / restaurant /
acheter tout frais

Facile à identifier les risques

additifs nocifs pour la santé.

Groupes NOVA

NOVA

1

Aliments non transformés



NOVA

2

Ingrédients culinaires



NOVA

3

Aliments transformés



NOVA

4

Produits ultra-transformés



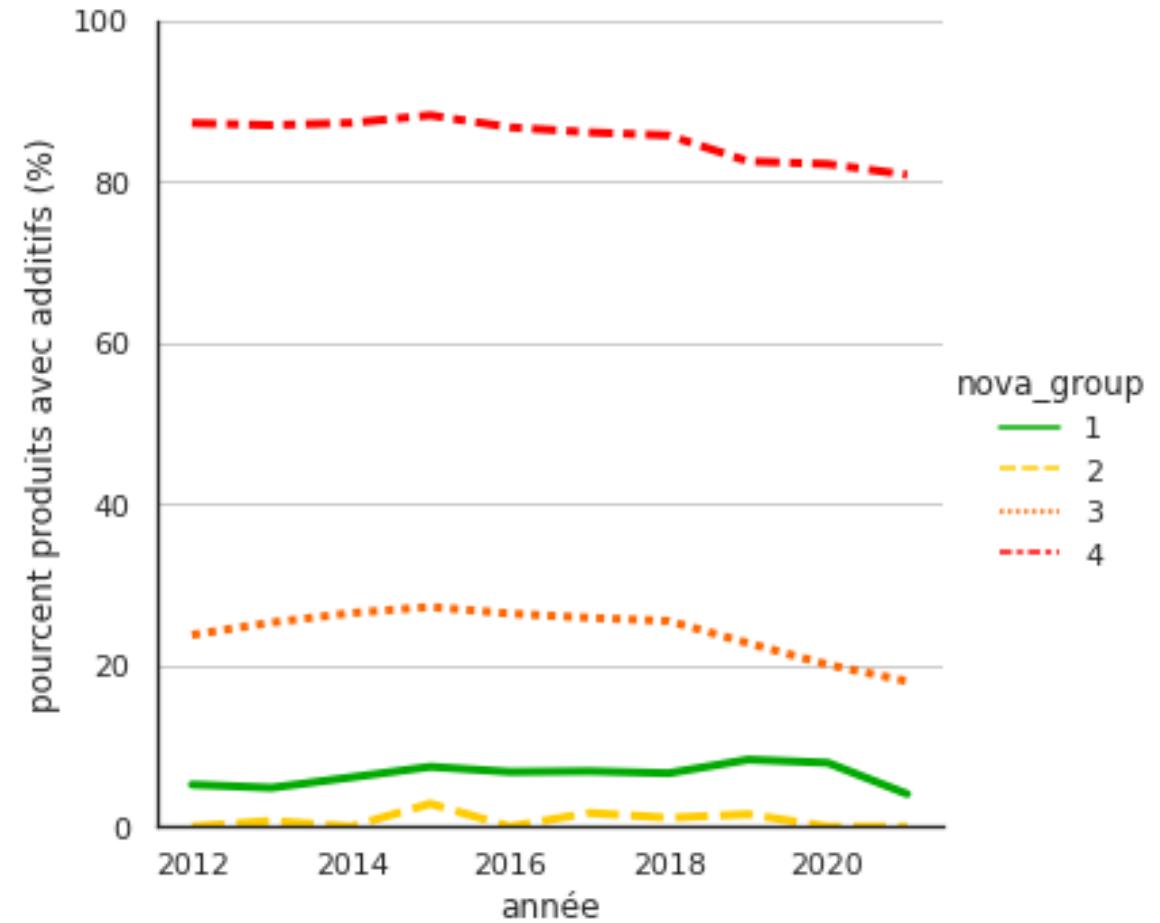
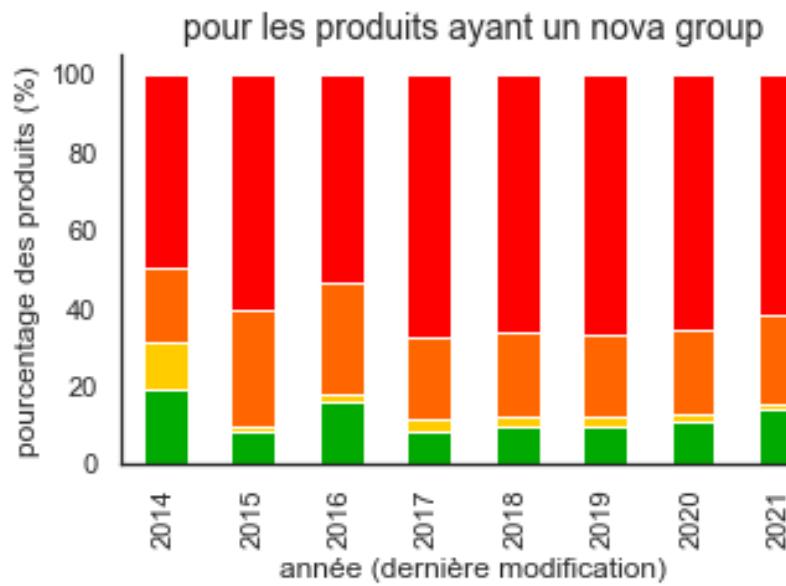
80 % des aliments en groupe 4 contient des additifs

—

> 50% des produits



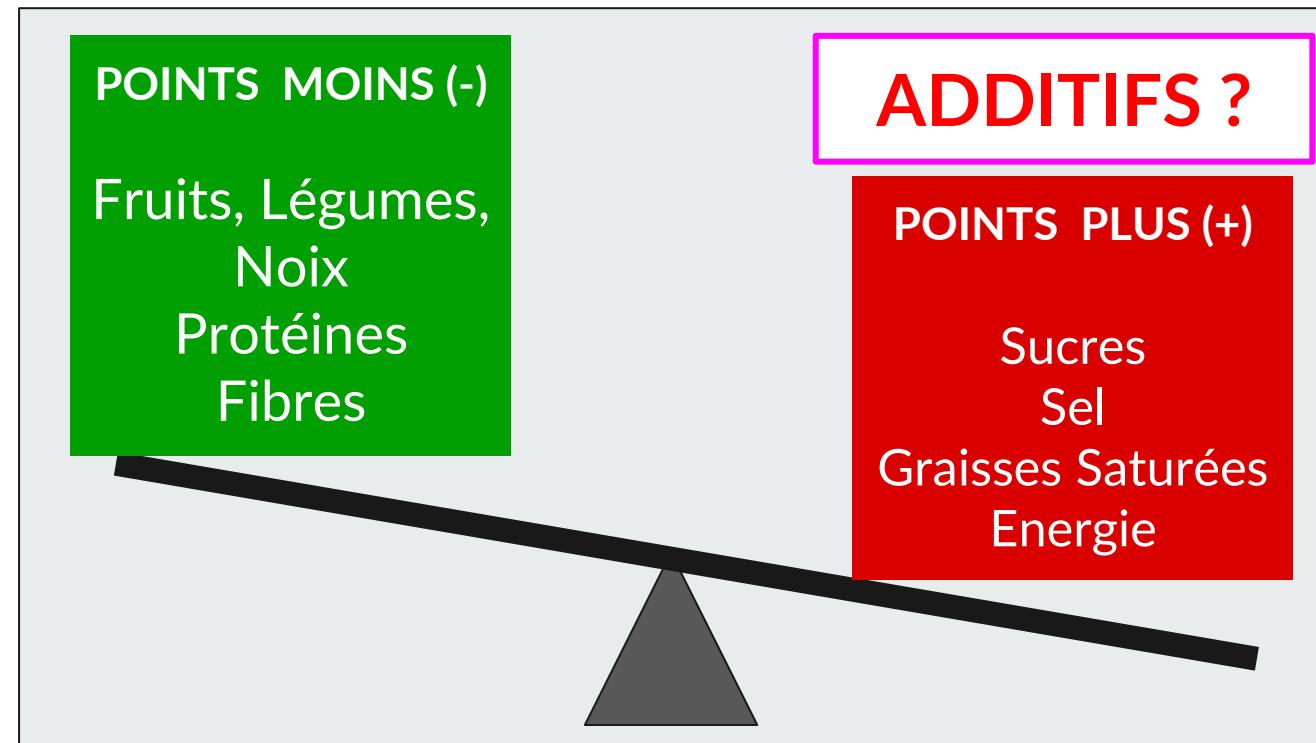
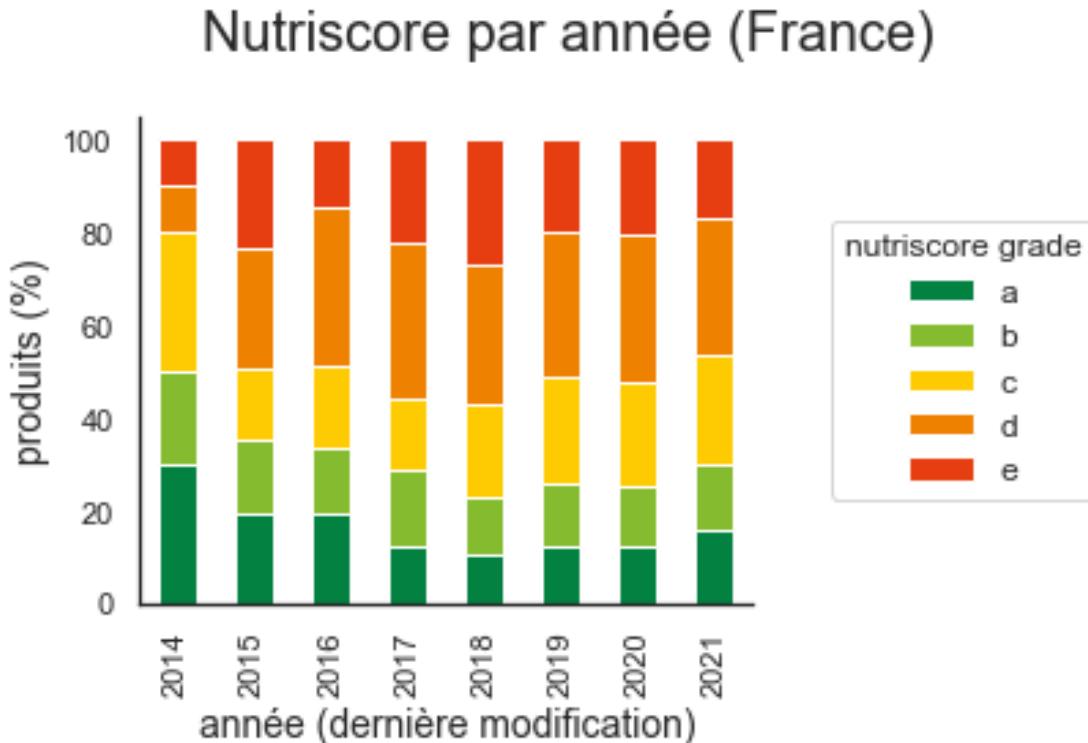
Distribution de groupe NOVA par année (France)



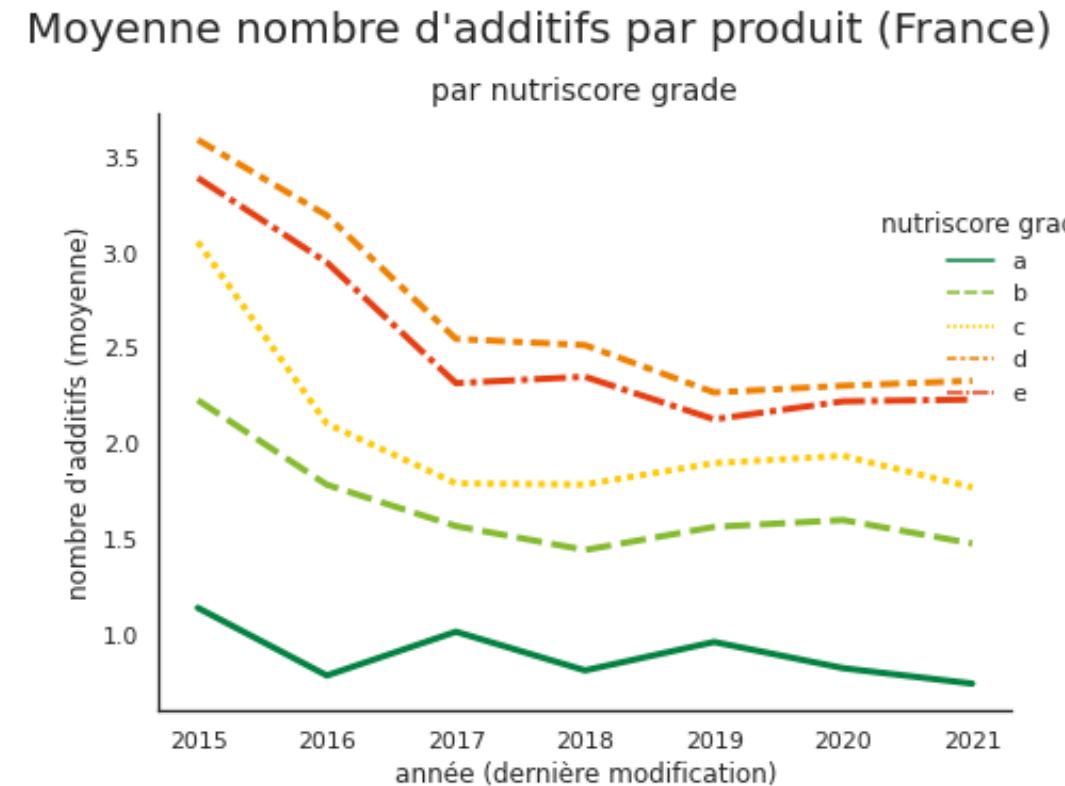
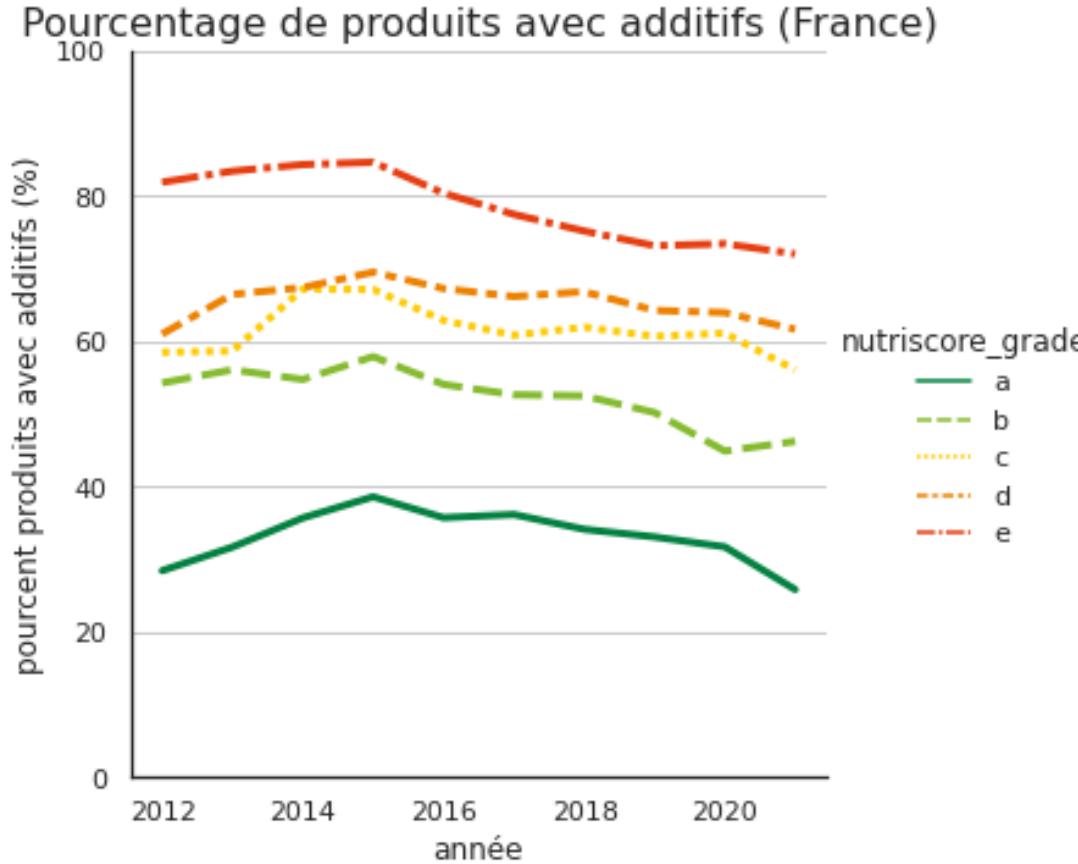
Le nutriscore ne prend pas en compte les additifs

le NUTRISCORE grade A B C D ou E,

le NUTRISCORE score (-20 à + 40)



Les additifs sont omniprésents (nutriscore A à E)



additifs en augmentation
(nutriscore D et E)

Simplifier les indicateurs pour choisir l'alimentation

le NUTRISCORE



la groupe NOVA



le nombre d'additifs



ADDISCORE

nombre d'additifs à risque (* poids)

A: 0 additifs

B: 1 * nb. additifs risque bas

C: 2 * nb. additifs risque modéré

D: 5 * nb. additifs haut risque



ADDISCORE RISK

présence d'additifs à risque:

A: aucun additif

B: avec additifs risque bas

C: avec additifs risque modéré de sur-exposure

D: avec additifs haut risque de sur-exposure

? U: Sans information des additifs

02 Préparation des données utilisées par l'application

Les données - Importation et nettoyage

Les données



> 2 million produits [OpenFoodFacts](#) (4.3Gb)

187 colonnes :

- identification (nom, code, brand, pays..)
- valeurs nutritionnelles (100g)
- ingrédients (incl. additifs)
- tags (labels, catégories, emballage,..)

<https://world.openfoodfacts.org/data>

L'environnement de travail



Jupyter Notebook sur Kaggle,

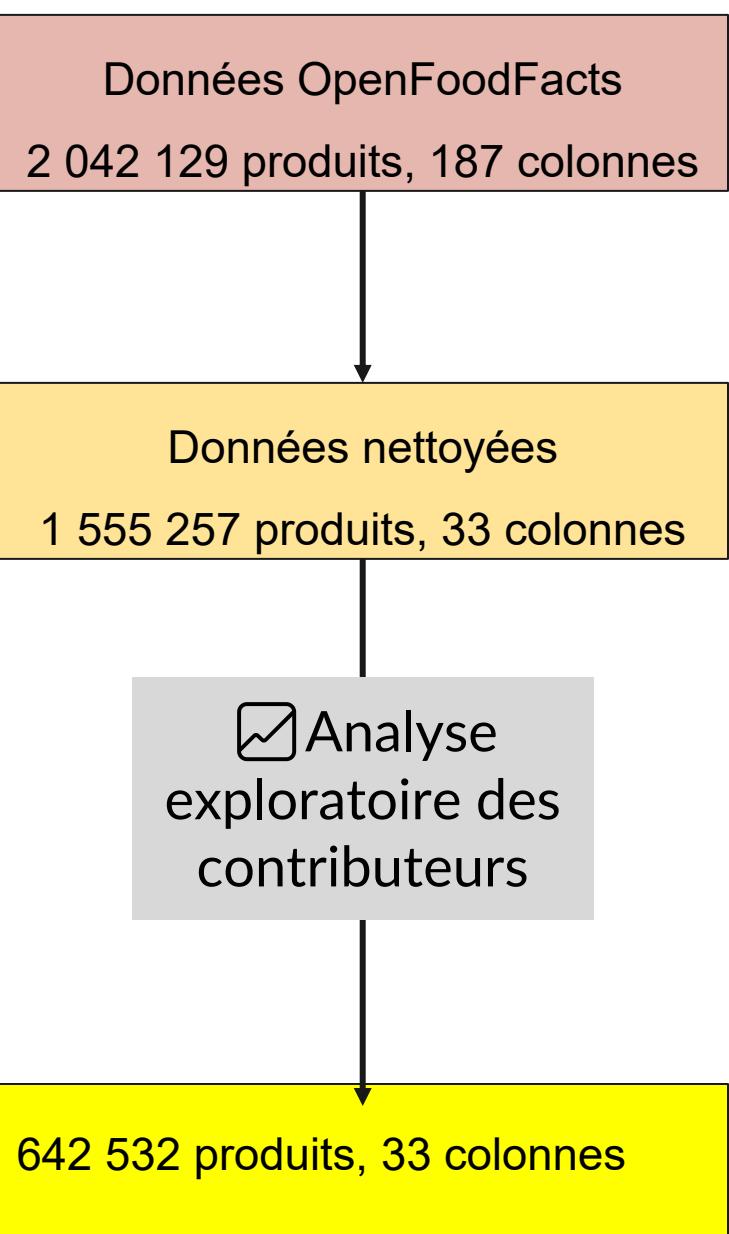
Google Colab, en local VSCode

- Bibliothèques en Python
 - numpy, pandas, missingno
 - matplotlib, seaborn
 - scikit-learn, scipy, statsmodels
 - requests

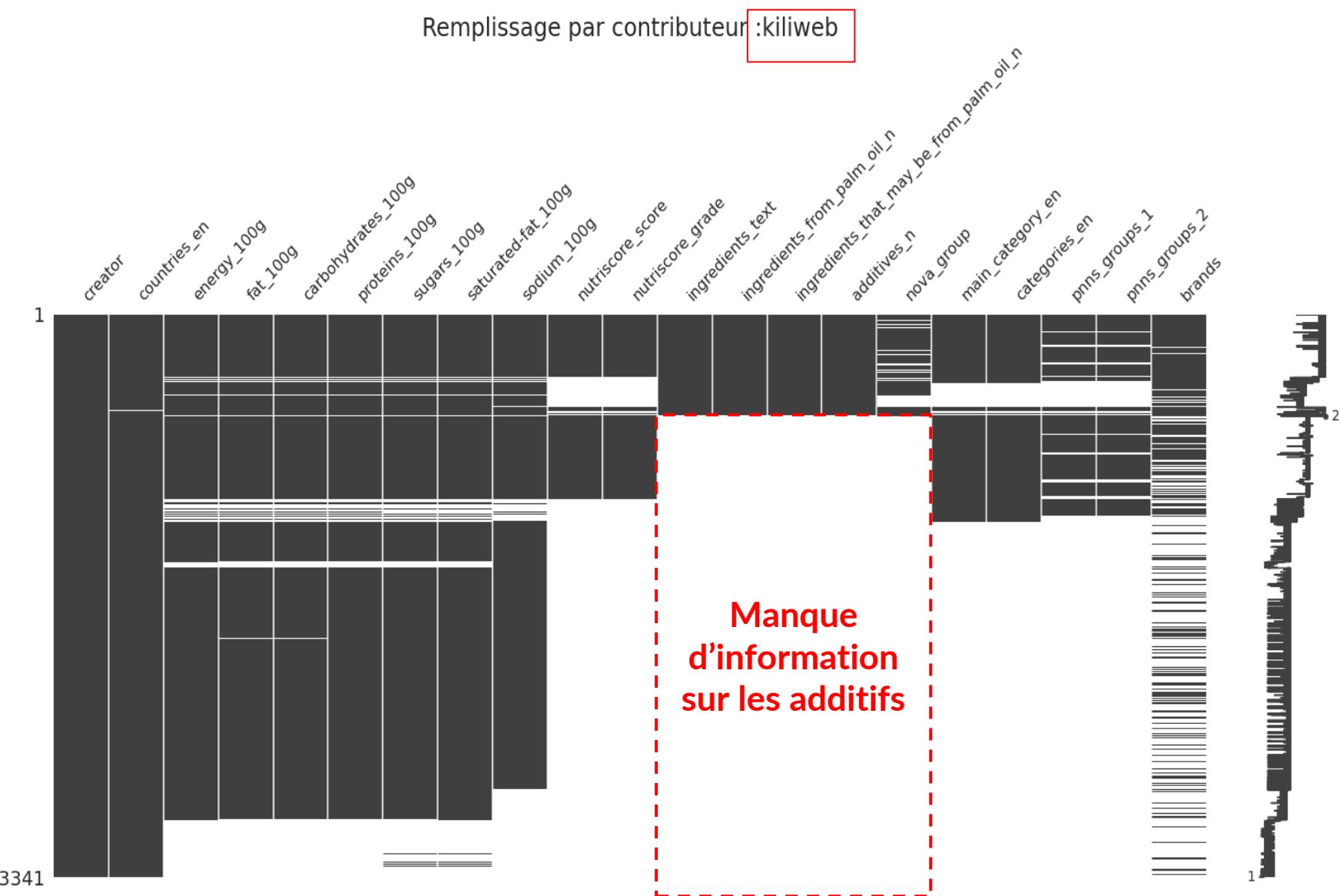
Les étapes du nettoyage

- Importation des données (187 colonnes)
- Correction des types
- Élimination:
 - ✗ 39 colonnes vides
 - ✗ 25 colonnes non-pertinentes
 - ✗ 60000 lignes dupliquées
 - ✗ 106 colonnes avec >80% de valeurs manquantes
- Analyse des valeurs manquantes
- Nettoyage des valeurs aberrantes
- Imputation des valeurs manquantes
- Elimination de 5 colonnes redondantes

avec information
d'additifs



Analyse de la distribution des valeurs manquantes



creator **kiliweb**

=

Yuka App 

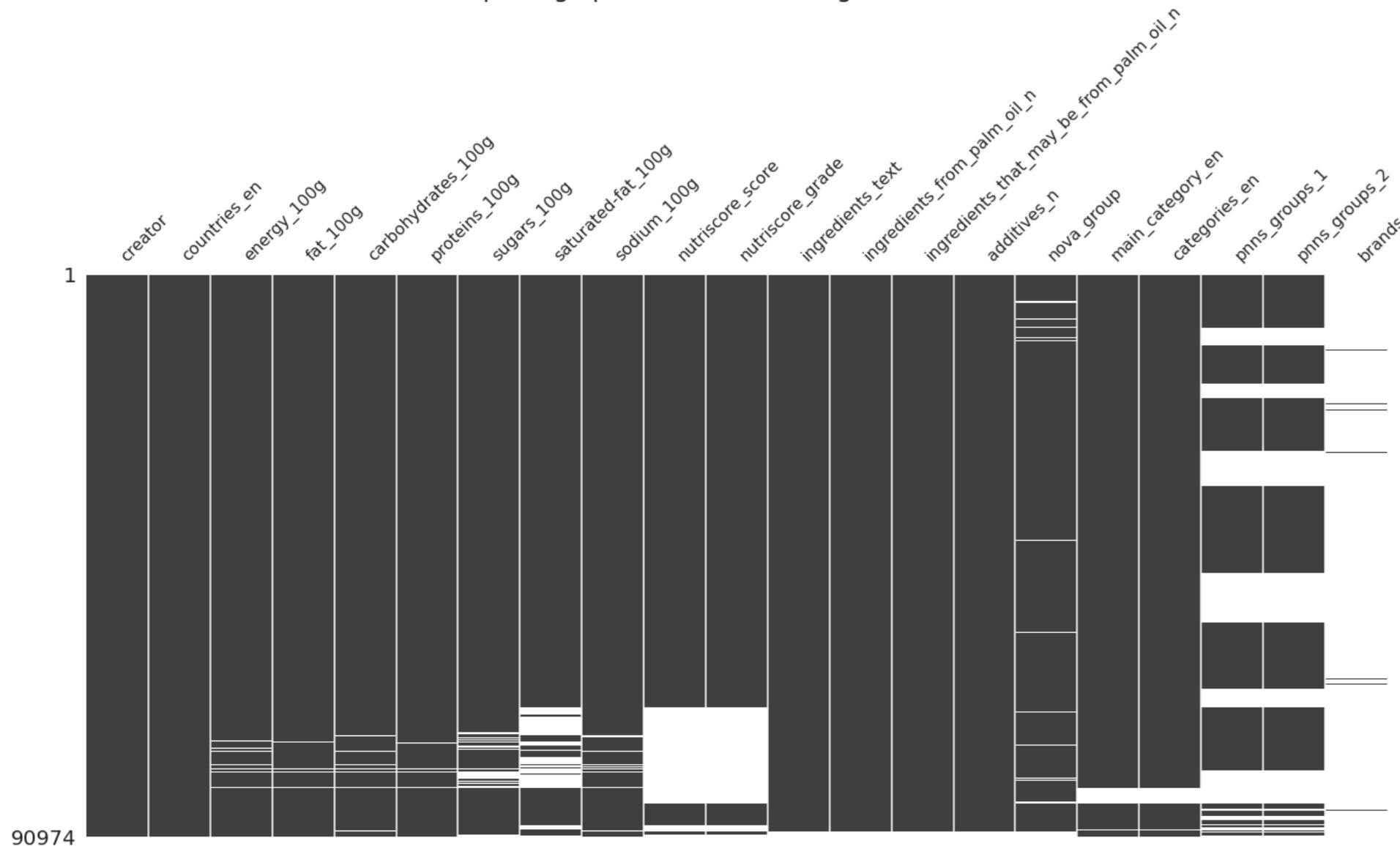
>75%

des produits
enregistrés

France

Analyse de la distribution des valeurs manquantes

Remplissage par contributeur :org-database-usda



Etats Unis

USDA Branded
Food Products
Database

100%

complet

pour les

additifs

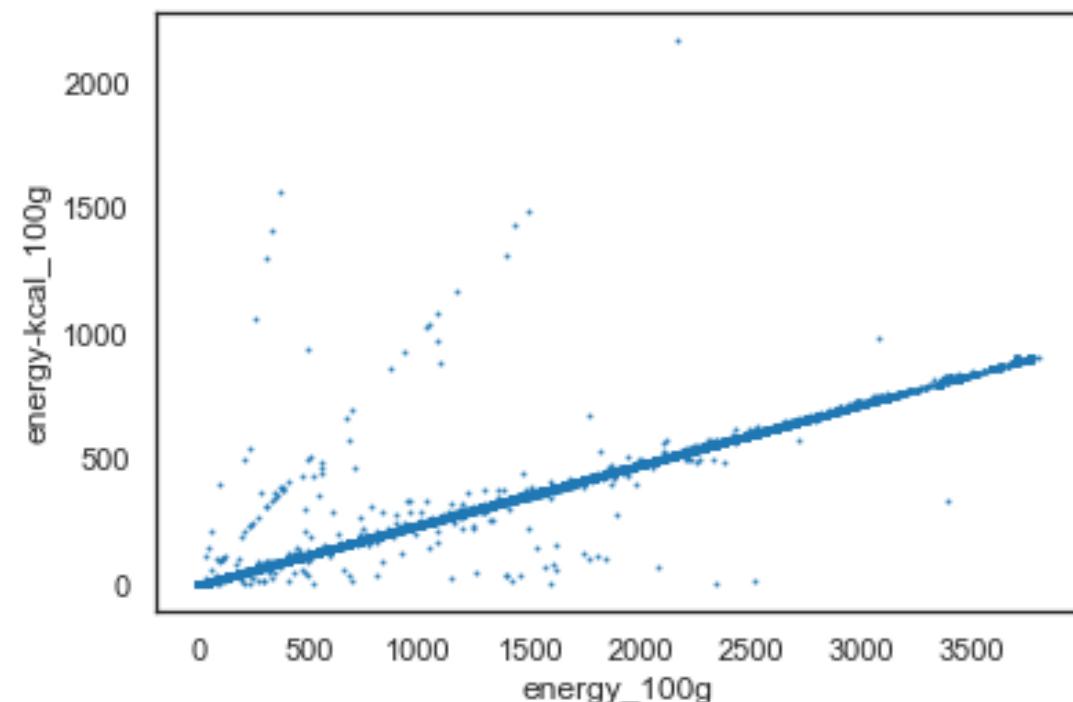


Détection et nettoyage des valeurs aberrantes

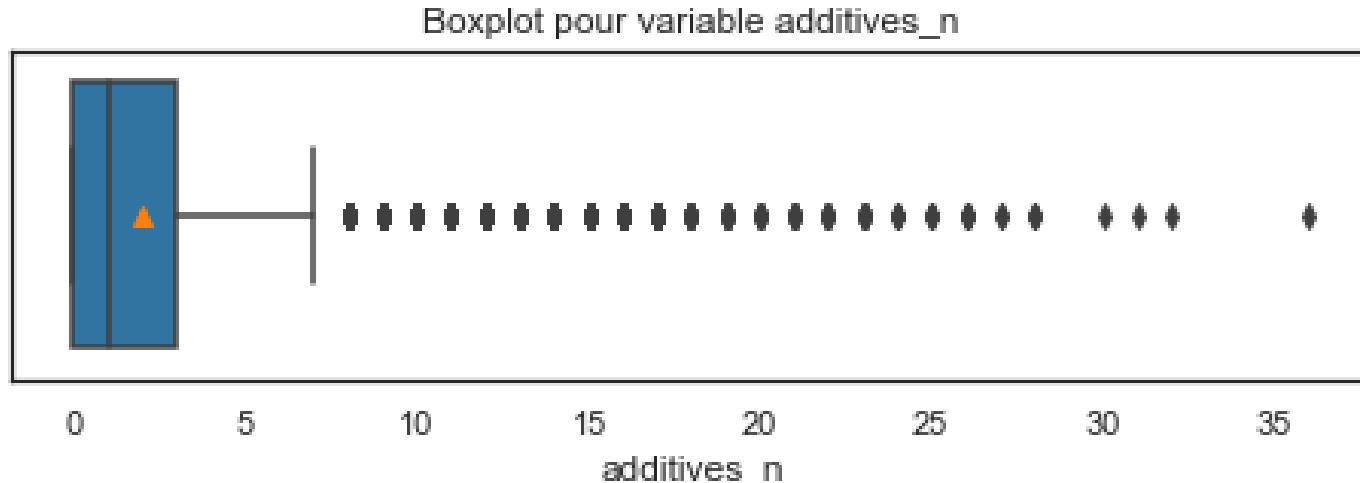
Produits non-alimentaires

main_category_en	
Open Beauty Facts	624
Non food products	512
Tests	112
Toothpaste	110
Cat food	80
Medicine	72
Dog food	55
Open Products Facts	53
Pet food	43
Dry cat food	25

- Valeurs nutritionnelles >100g per 100g
- Valeurs négatives
- Somme valeur nutritionnelle >100g
- Energie > 3800 kJ pour 100g
- Energie en kJ ne correspondant pas aux énergies en kcal



Détection des valeurs atypiques

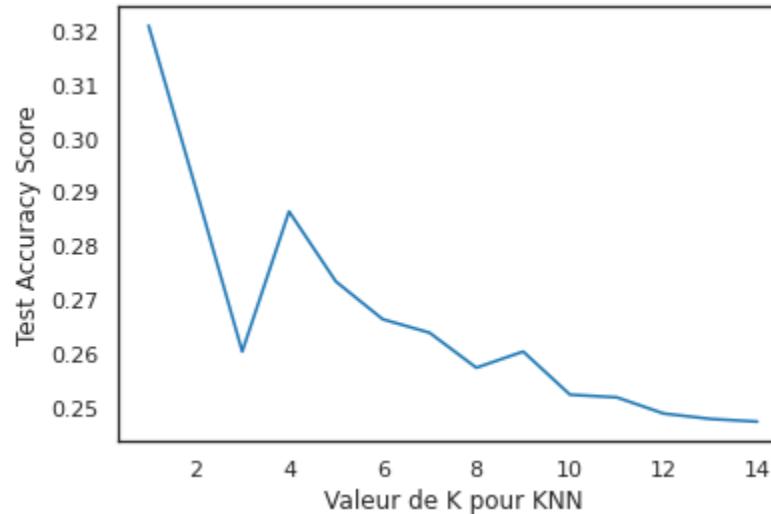
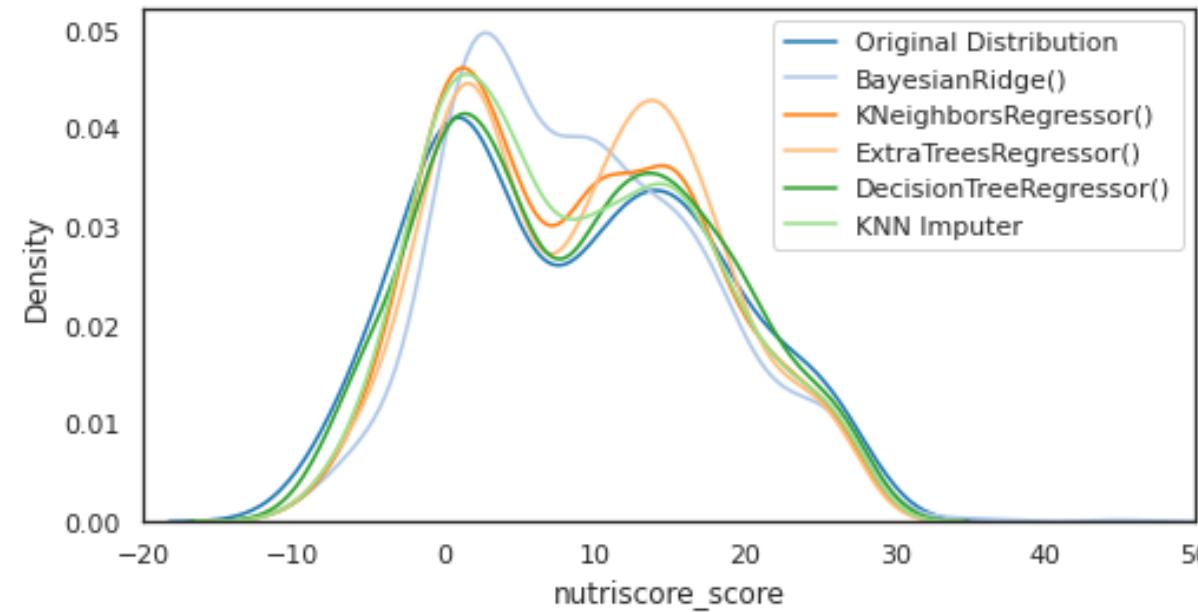
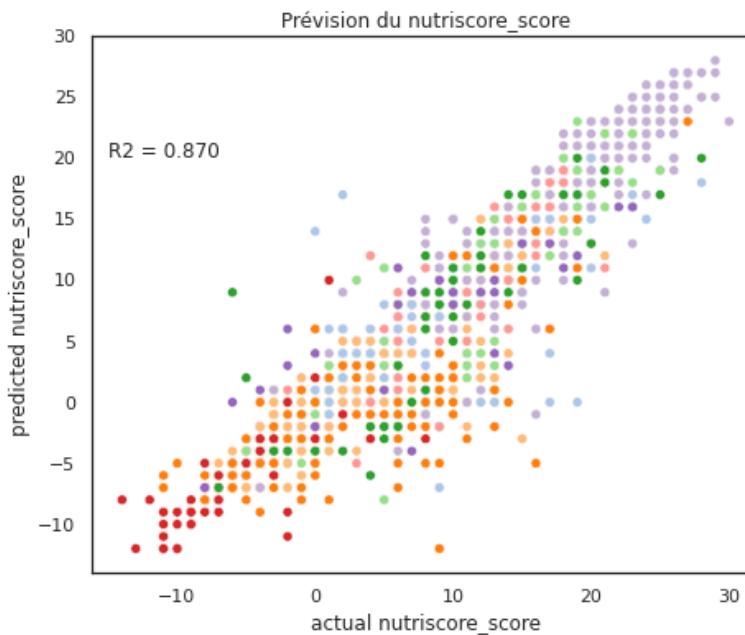


Utilisation d'un box plot pour détecter les outliers

Exemples (nb. additives > 7)

- "Gardetto's **Original Recipe Snack Mix**",
- '**Caffeine free** diet orange flavored soda, orange',
- 'Tortilla Chips',
- 'Calcium fortified **enriched** bread',
- "Udi's **gluten free**, seeded whole grain dinner rolls",
- '**Délice de poulet BBQ**',

Imputation des valeurs manquantes



L'imputation change les distributions des variables

- il changera les conclusions
- on testera plus tard l'hypothèse de l'influence sur l'ADDISCORE

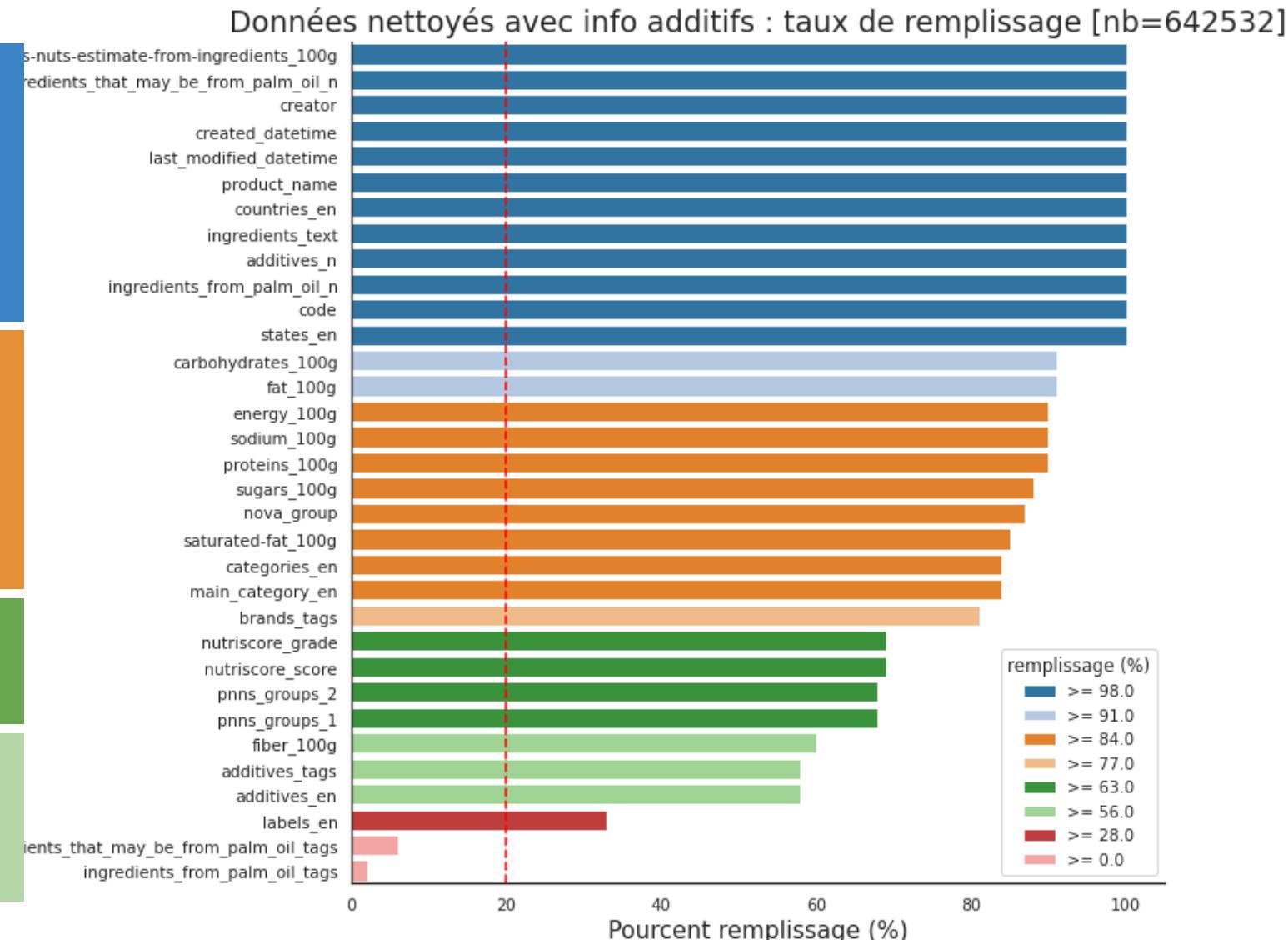
Additifs nettoyés- remplissage des colonnes

ingrédients et additifs
100%

valeurs nutritionnelles
>80%

nutriscore >70%

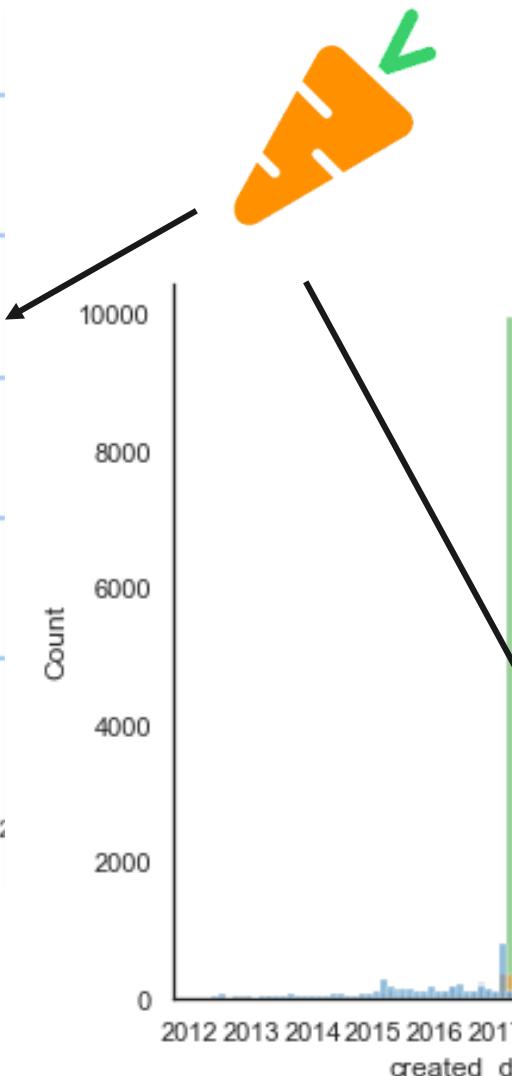
> additif tags 100%
pour les 56% de produits qui contiennent des
additifs



03 Les contributeurs : Qui, comment, quand, quoi, pourquoi

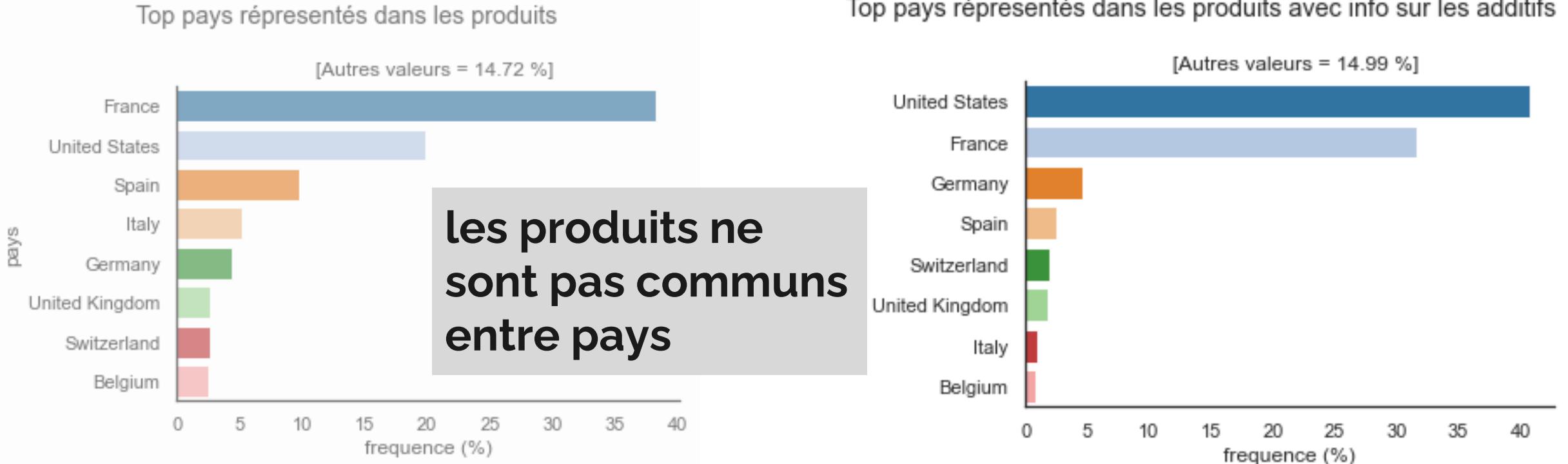
Comprendre d'où viennent les informations

>75% des données créées depuis 2018



Yuka App
56%
des données

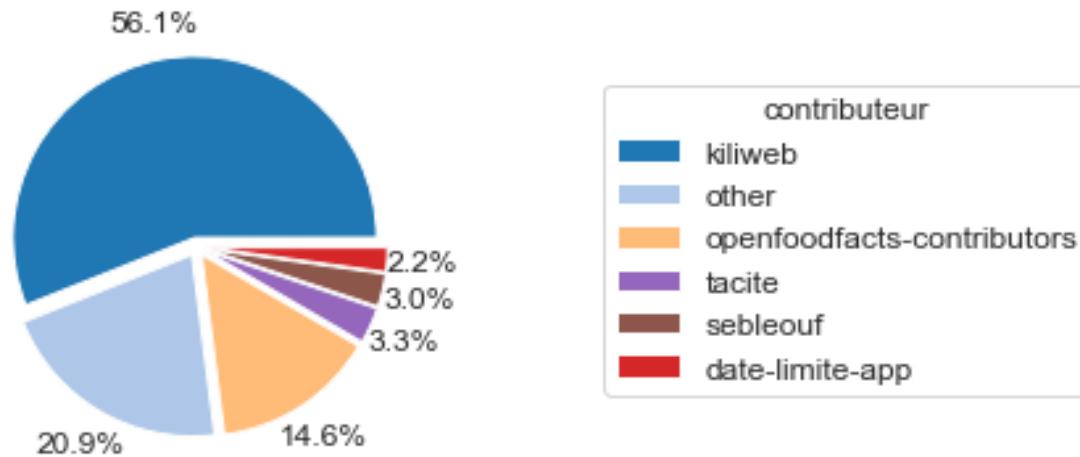
65% des informations des additifs sont hors France



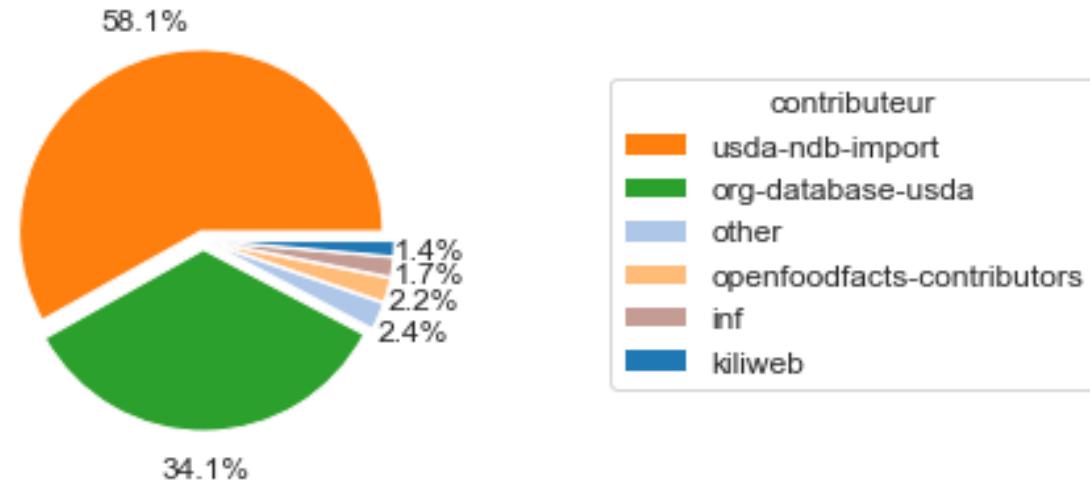
Est-ce que les
données entre
sources peuvent
être intégrées?

Intégrer les données de tous les pays ?

Principaux contributeurs aux colonnes des additifs (France)



Principaux contributeurs aux colonnes des additifs (Etats Unis)



>56%
France



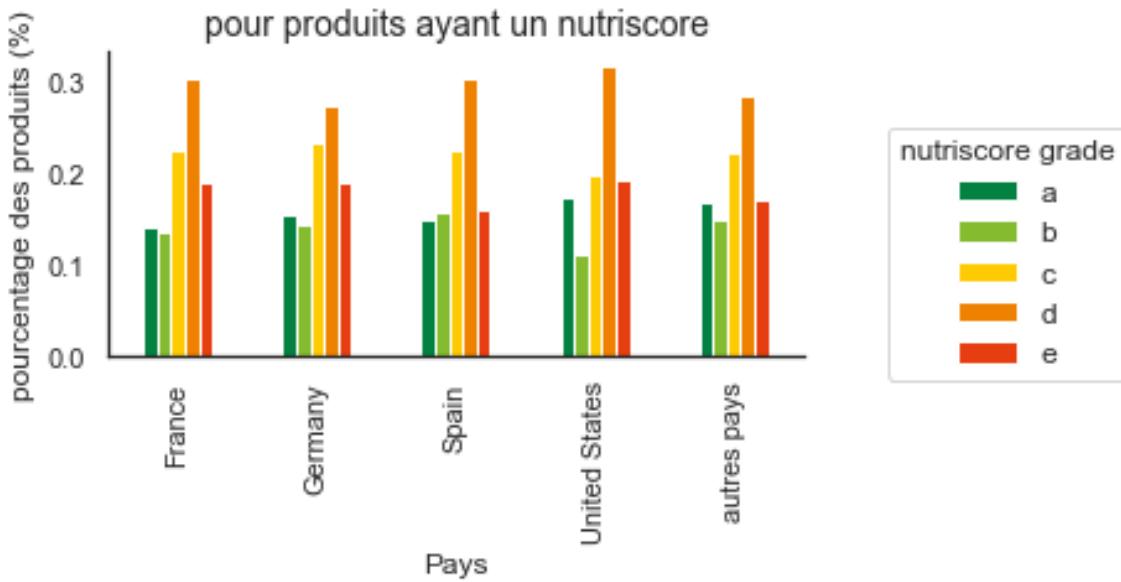
Yuka App

<1.4%
l'etats unis

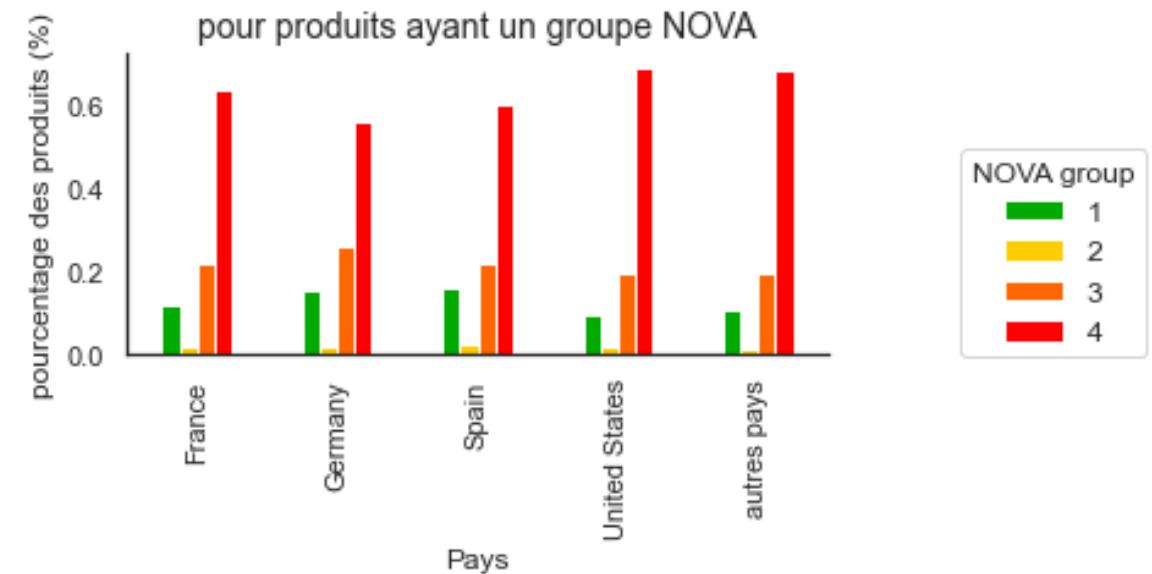
Intégrer les données de tous les pays?

Nutriscore et NOVA sont comparables entre pays

Distribution de nutriscore grade par pays

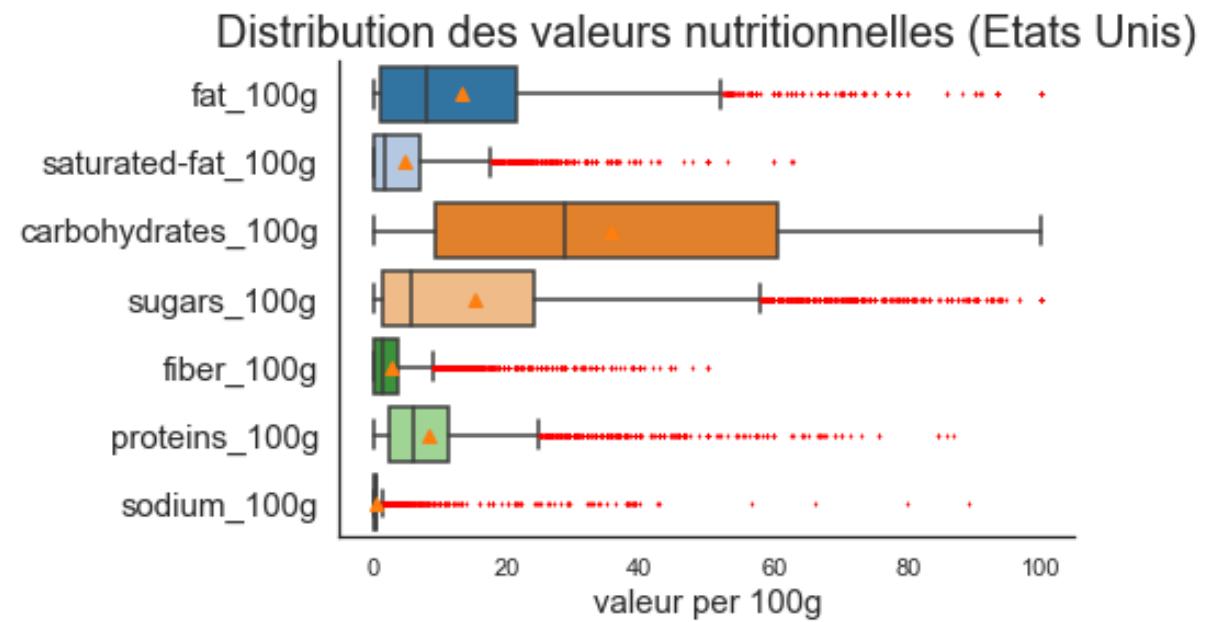
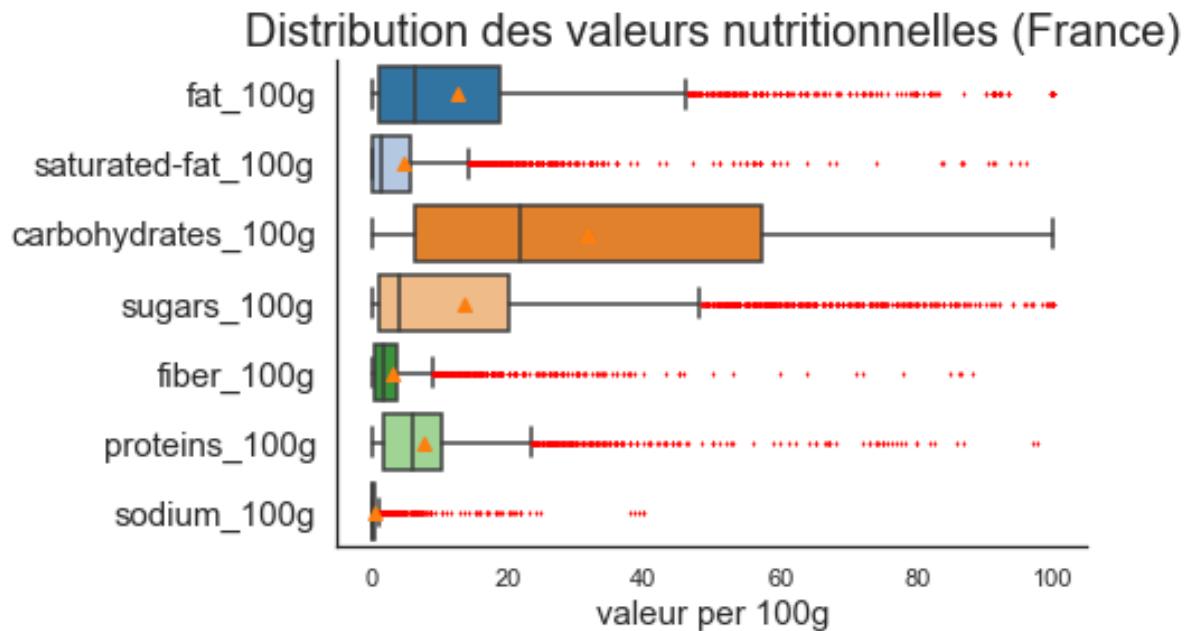


Distribution de groupe NOVA par pays



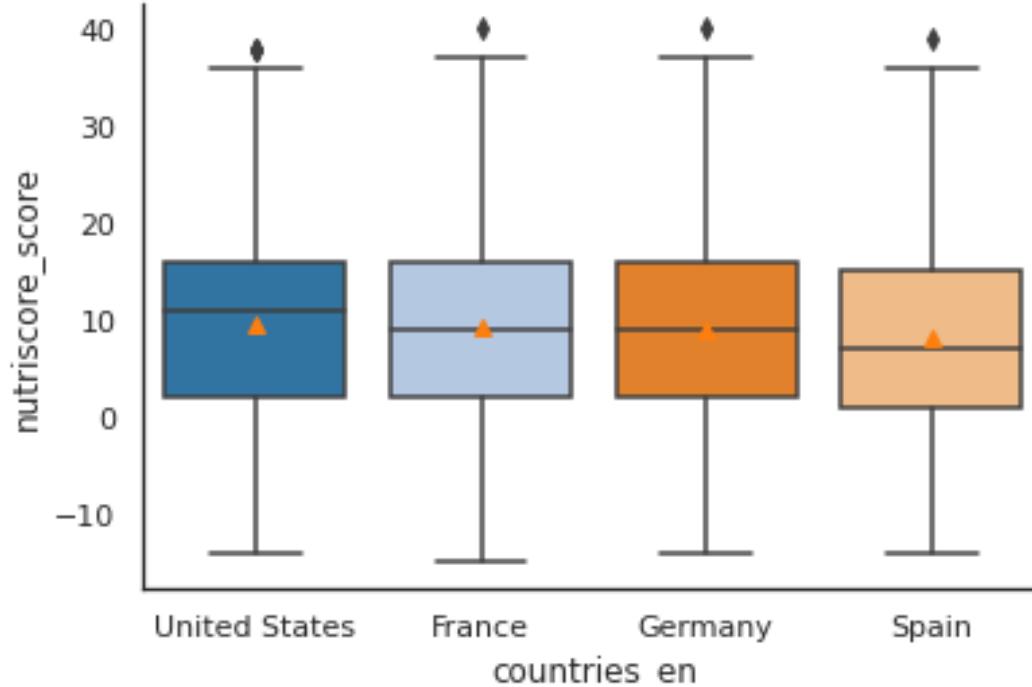
Intégrer les données de tous les pays?

Valeurs nutritionnelles sont comparables entre pays



L'analyse d'additifs se fait sur tous les pays

Nutriscore score par pays



pour ne pas perdre
65% des données

Tests de normalité des distributions

France (mean:9.16, std:8.95): Kolmogorov-Smirnov : stat=0.732, p=0.000***
Germany (mean:9.08, std:8.98): Kolmogorov-Smirnov : stat=0.729, p=0.000***
Spain (mean:8.07, std:8.91): Kolmogorov-Smirnov : stat=0.676, p=0.000***
United States (mean:9.53, std:8.95): Kolmogorov-Smirnov : stat=0.733, p=0.000***
La variable dependant (nutriscore_score) n'est pas distribuée normalement pour chaque groupe dans countries_en.

Test de homoscédasticité

Levene homoscédasticité : stat=15.694, p=0.000***
Levene test, reject H0: variance of groups is not equal

Test H-ANOVA (Kruskal-Wallis)

Kruskal-Wallis H-test : stat=575.906, p=0.000***
les médianes sont significativement différentes (reject H0)

Pairwise Tukey HSD test (honestly significant difference)

[France, Spain] p=0.001;
[France, United States] p=0.001;
[Germany, Spain] p=0.001;
[Germany, United States] p=0.001;
[Spain, United States] p=0.001;

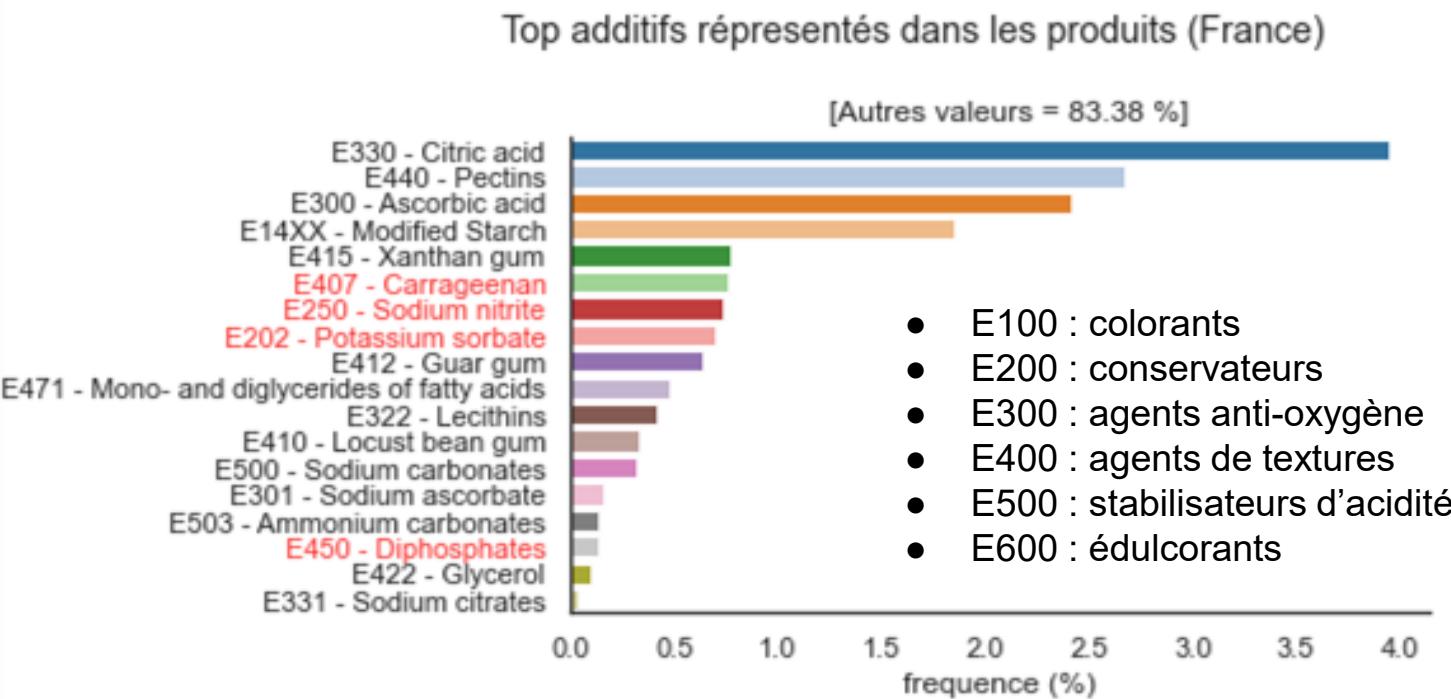
04 Les additifs : Catégorisation, répartitions entre groupes d'alimentation

Cibler les plats préparés

Quels additifs ajouteraient des points sur le Nutriscore ?

La plupart des 500 additifs sont inoffensifs ou même bons pour la santé

Les additifs sont ajoutés pour augmenter les ventes:



- E100 : colorants
- E200 : conservateurs
- E300 : agents anti-oxygène
- E400 : agents de textures
- E500 : stabilisateurs d'acidité
- E600 : édulcorants

- améliorer le nutriscore
 - aspartame (moins de sucre)
 - lécithines (moins de gras)
- combattre les bactéries / gaspillage
 - nitrites (viandes)
- améliorer le goût / texture
 - diphosphates (gâteaux, biscuits)
 - carrageenan (desserts)

Les additifs à risque de surexposition

Additifs à risque élevé :

E224 - Potassium metabisulphite E223 - Sodium metabisulphite
E200 - Sorbic acid E451 - Triphosphates
E340 - Potassium phosphates E210 - Citric acid
E202 - Potassium sorbate
E473 - Sucrose esters of fatty acids E492 - Sorbitan tristearate
E450 - Diphosphates
E481 - Sodium stearoyl-2-lactylate E491 - Sorbitan monostearate
E228 - Potassium bisulphite E407a - Processed eucheuma seaweed
E407 - Carrageenan E220 - Sulphur dioxide
E452 - Polyphosphates E211 - Sodium benzoate
E252 - Potassium nitrate E338 - Phosphoric acid
E250 - Sodium nitrite E221 - Sodium sulphite
E339 - Sodium phosphates E222 - Sodium bisulphite
E341 - Calcium phosphates E251 - Sodium nitrate

Additifs à risque modéré :

E960 - Steviol glycosides E435 - Polyoxyethylene sorbitan monostearate
E508 - Potassium chloride
E150c - Ammonia caramel
E142 - Green s E131 - Patent blue v
E133 - Brilliant blue FCF
E509 - Calcium chloride E433 - Polyoxyethylene sorbitan monooleate

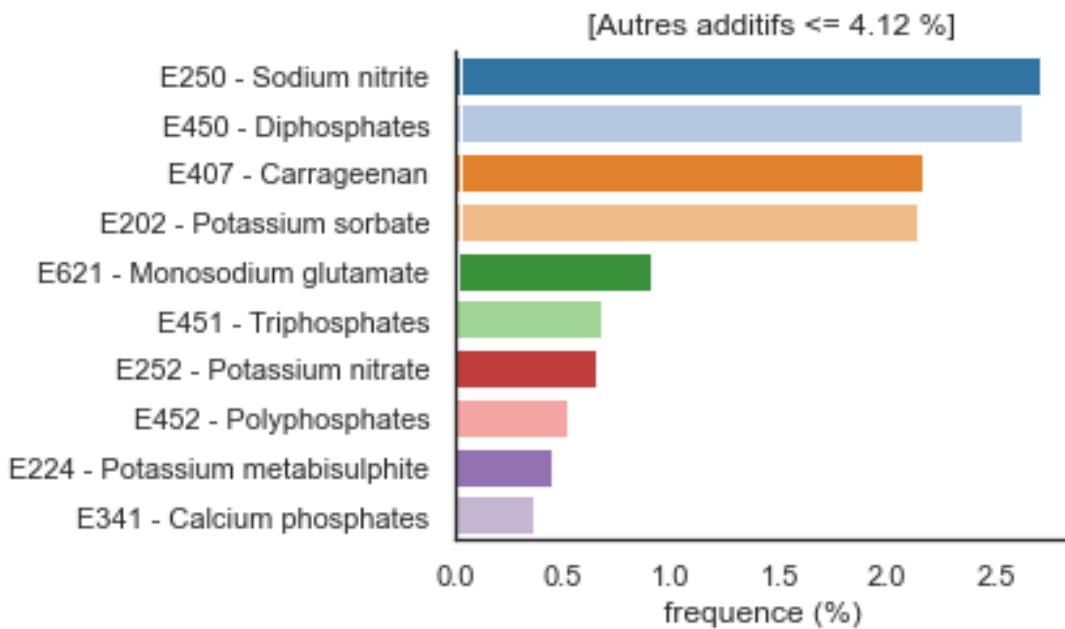
Additifs à bas risque :

E392 - Extracts of rosemary E420 - Sorbitol E120 - Cochineal
E410 - Locust bean gum E160 - Carotenoids
E150a - Plain caramel E955 - Sucralose
E322 - Lecithins
E100 - Curcumin E440 - Pectins
E500 - Sodium carbonates
E330 - Citric acid
E471 - Mono- and diglycerides of fatty acids
E450i - Disodium diphosphate E428 - Gelatine
E262 - Sodium acetates
E14XX - Modified Starch
E503 - Ammonium carbonates E270 - Lactic acid
E422 - Glycerol E331 - Sodium citrates
E300 - Ascorbic acid
E950 - Acesulfame k E160c - Paprika extract E412 - Guar gum
E301 - Sodium ascorbate E414 - Acacia gum
E322i - Lecithin
E500ii - Sodium hydrogen carbonate

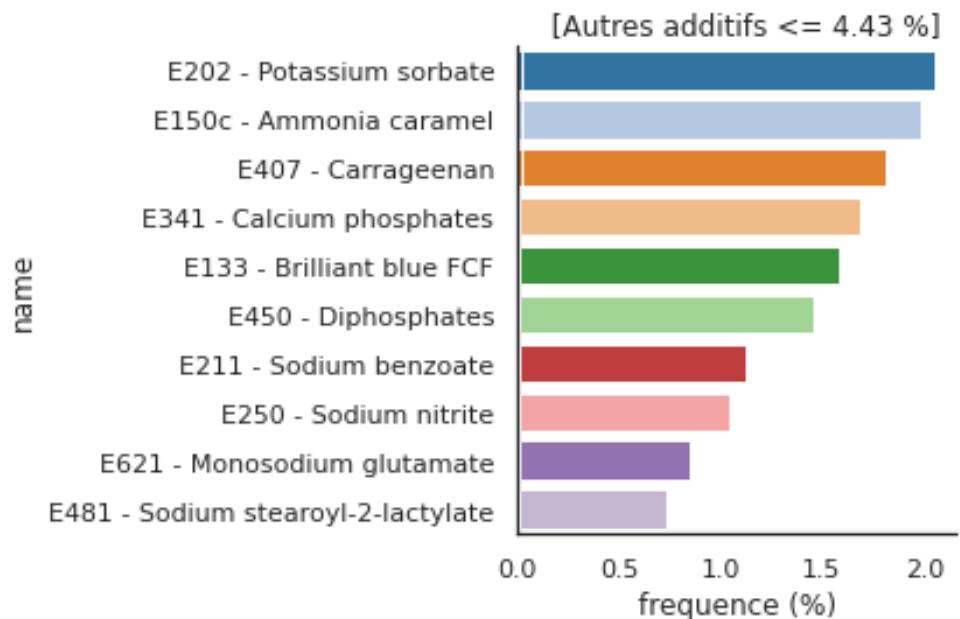
Fréquence d'additifs à risque élevé dépend du pays



Produits avec additifs à risque - France



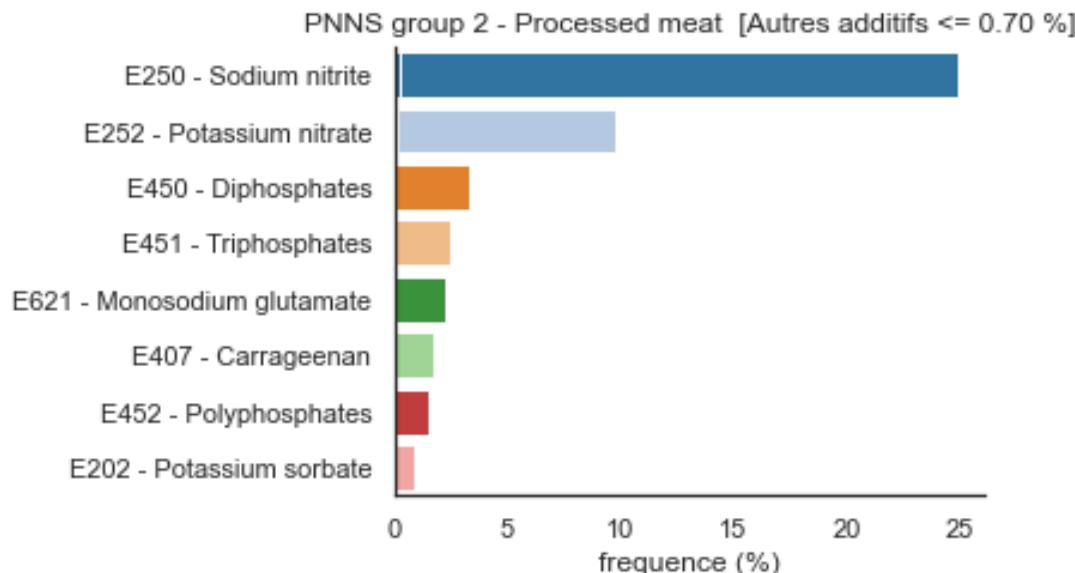
Produits avec additifs à risque - United States



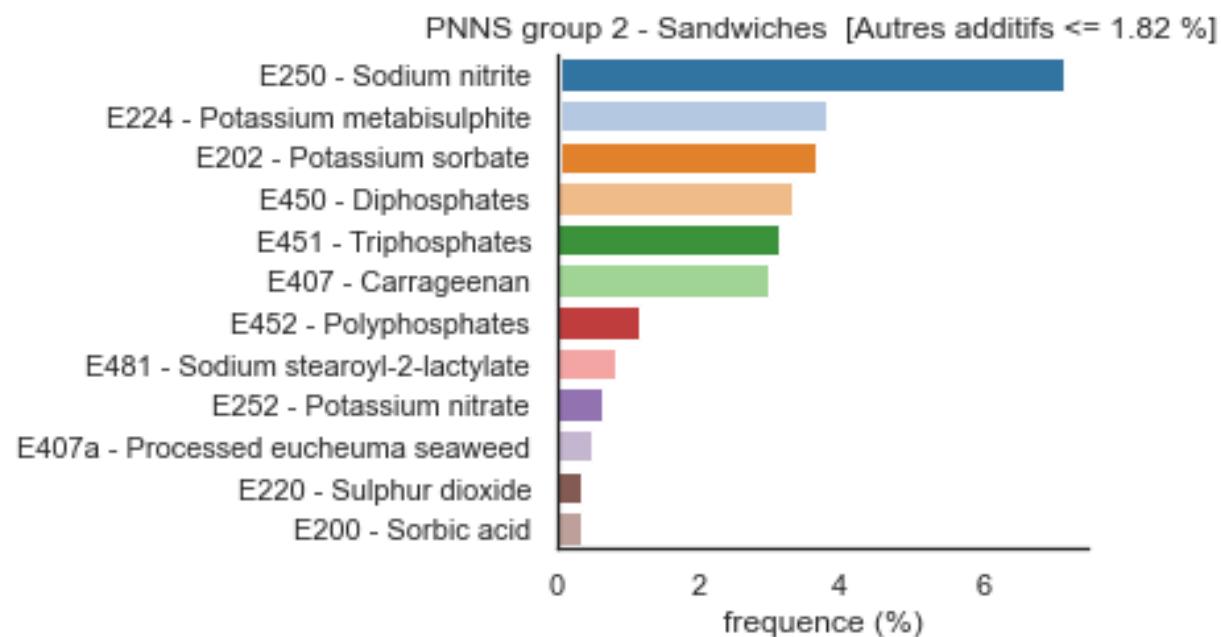
Fréquence d'additifs à risque élevé dépend du groupe



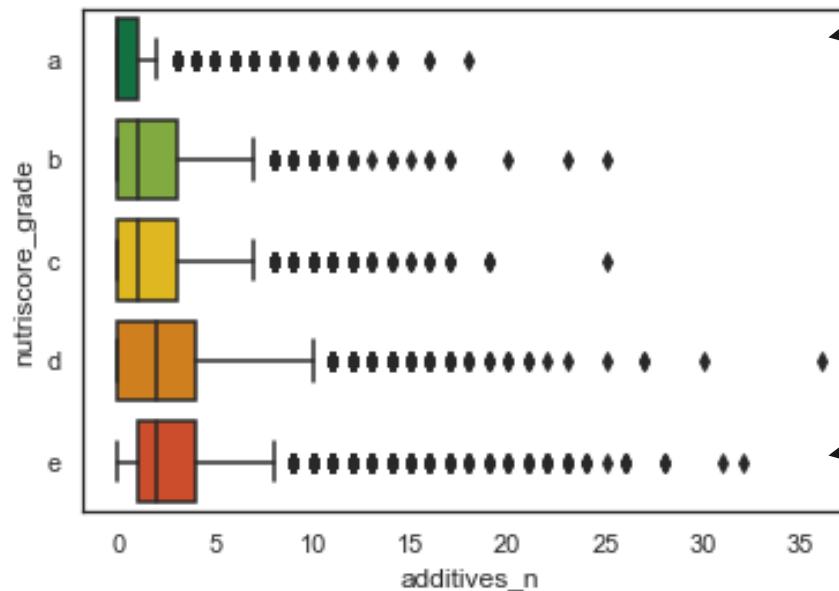
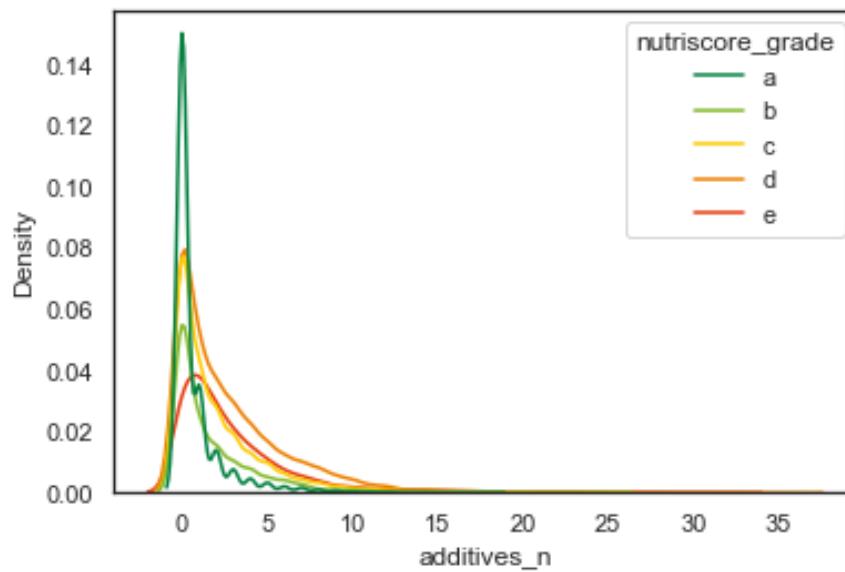
Produits avec additifs à risque - France



Produits avec additifs à risque - France



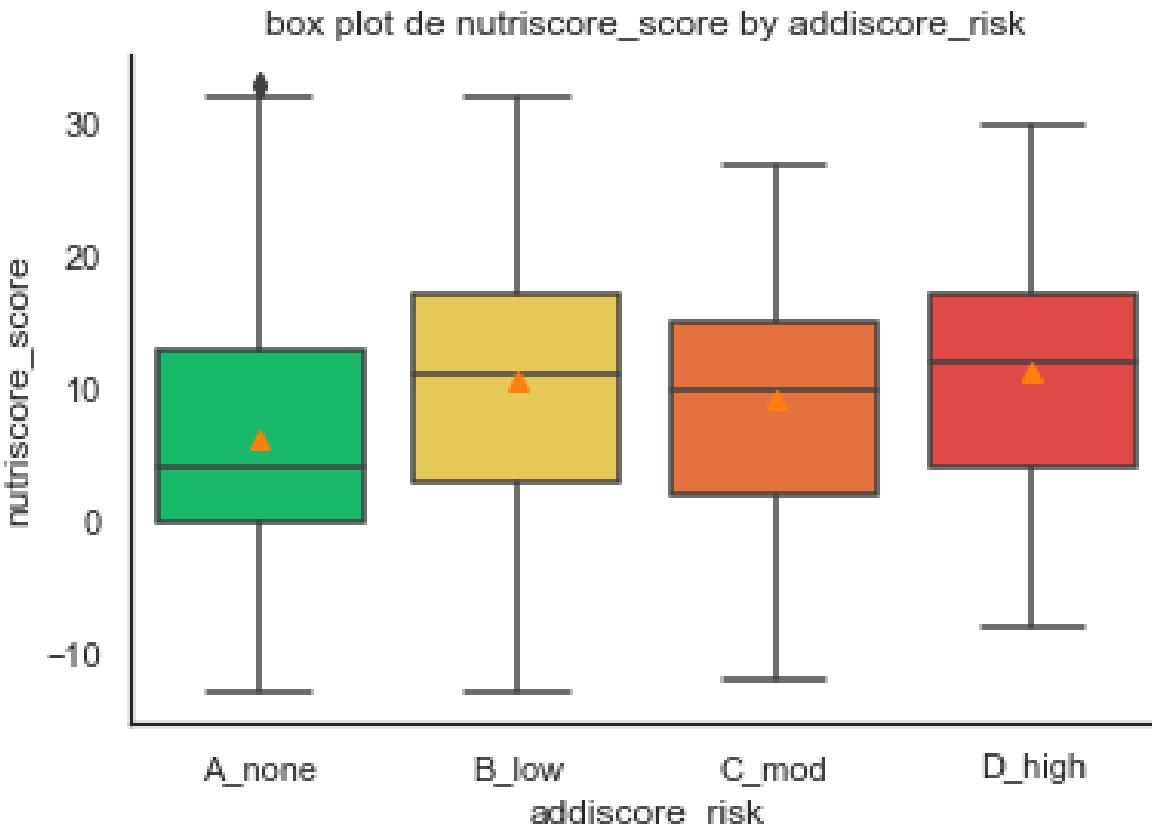
Distribution d'additifs par Nutriscore



05. L' addiscore : Comparaison avec d'autres indicateurs existants

Nutriscore n'est pas un bon indicateur de risque

Le nutriscore n'est pas très dépendant des additifs de haut risque



Tests de normalité des distributions

A_none : Kolmogorov-Smirnov : stat=0.601, p=0.000***

B_low : Kolmogorov-Smirnov : stat=0.796, p=0.000***

C_mod : Kolmogorov-Smirnov : stat=0.757, p=0.000***

D_high : Kolmogorov-Smirnov : stat=0.850, p=0.000***

La variable dépendante (nutriscore_score) n'est pas distribuée normalement pour chaque groupe dans addiscore_risk.

Test de homoscédascité

Levene homoscédascité : stat=8.031, p=0.000***

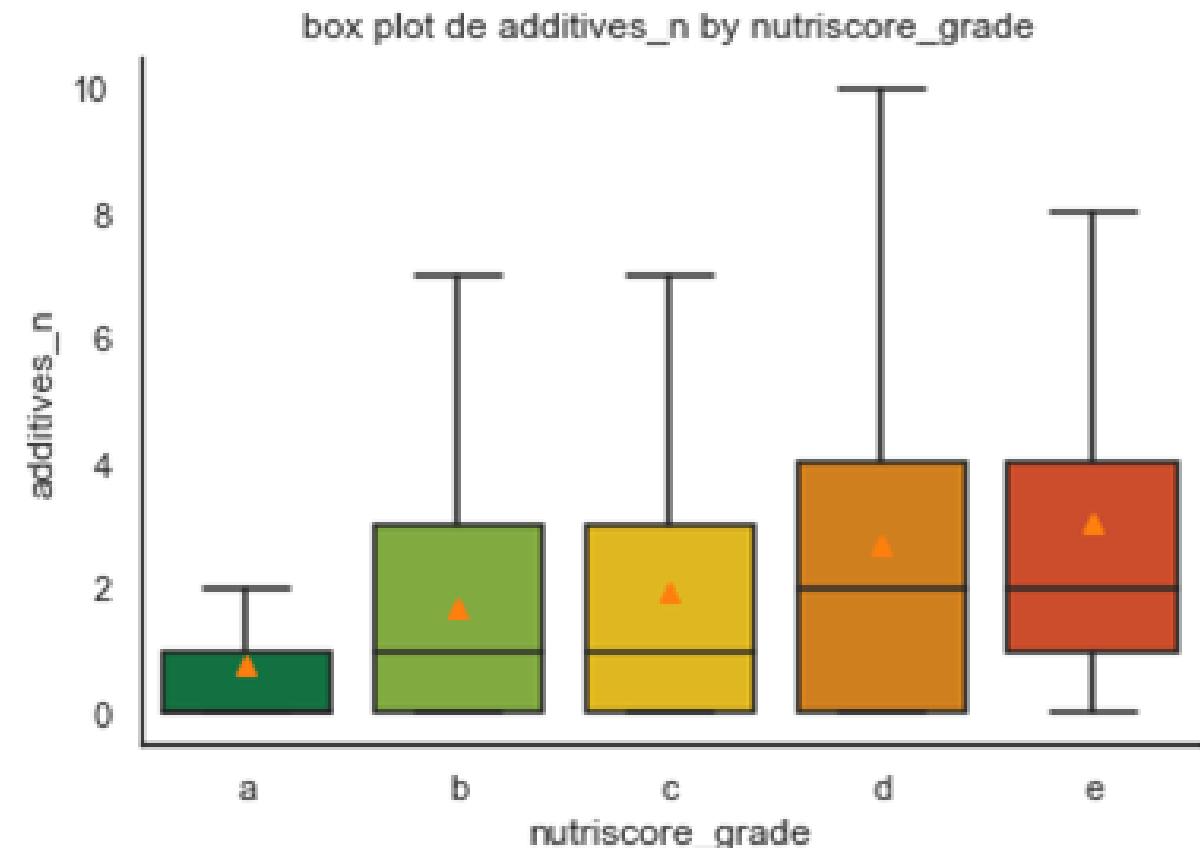
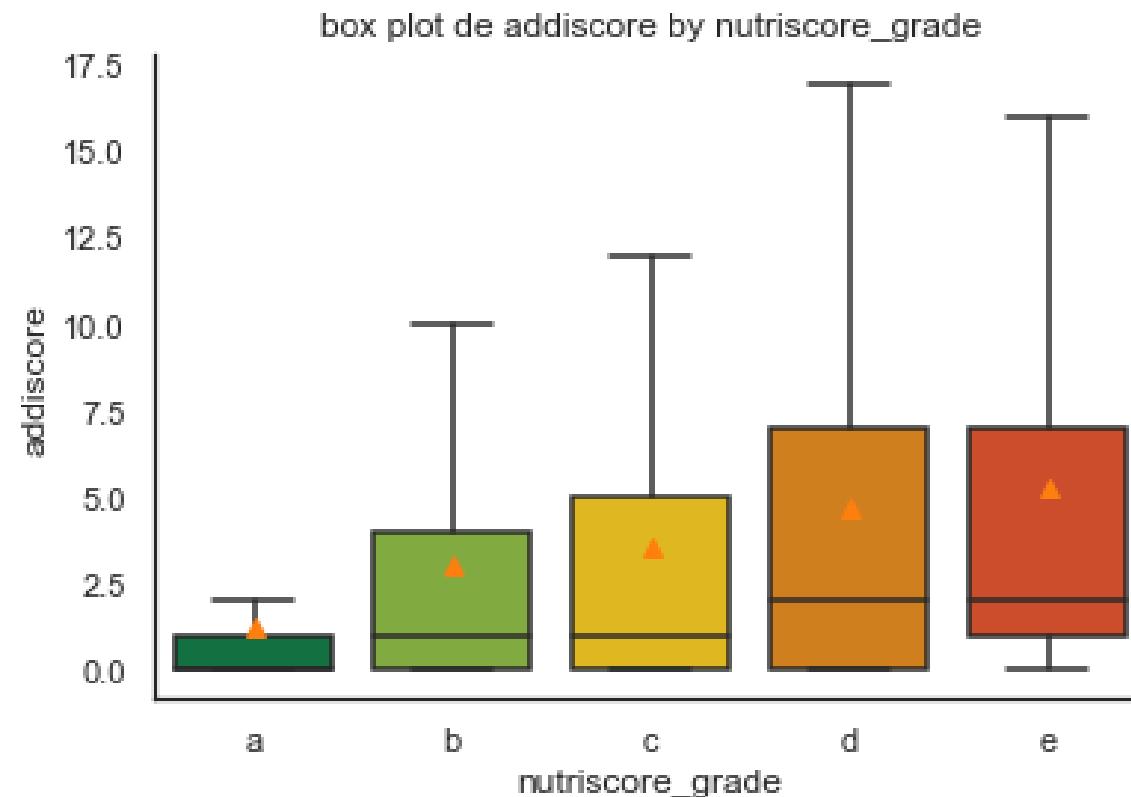
Levene test, reject H0: variance of groups is not equal

Test H-ANOVA (Kruskal-Wallis)

Kruskal-Wallis H-test : stat=719.864, p=0.000***

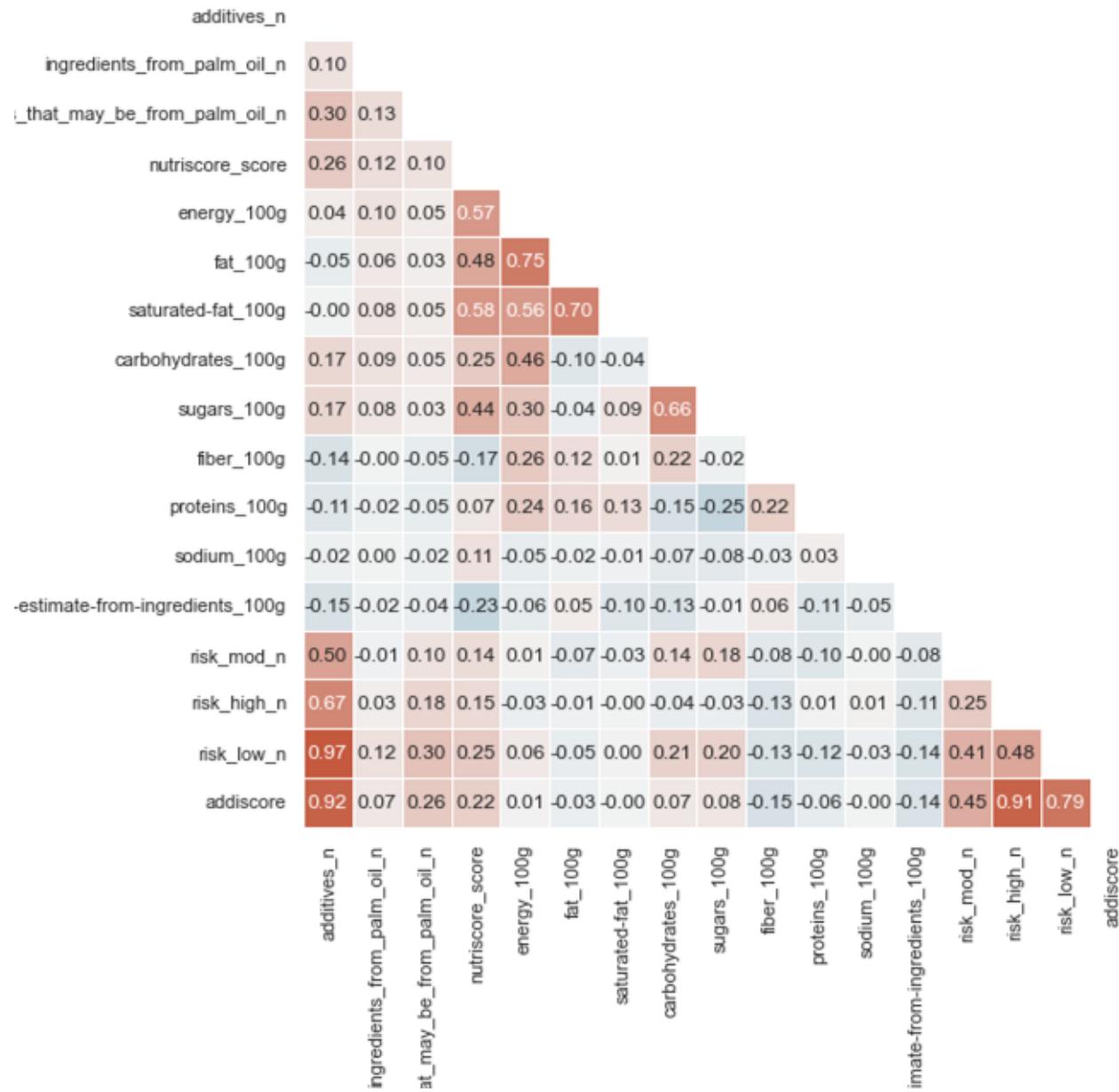
les médianes sont significativement différentes (reject H0)

Distribution de l'ADDISCORE par nutriscore



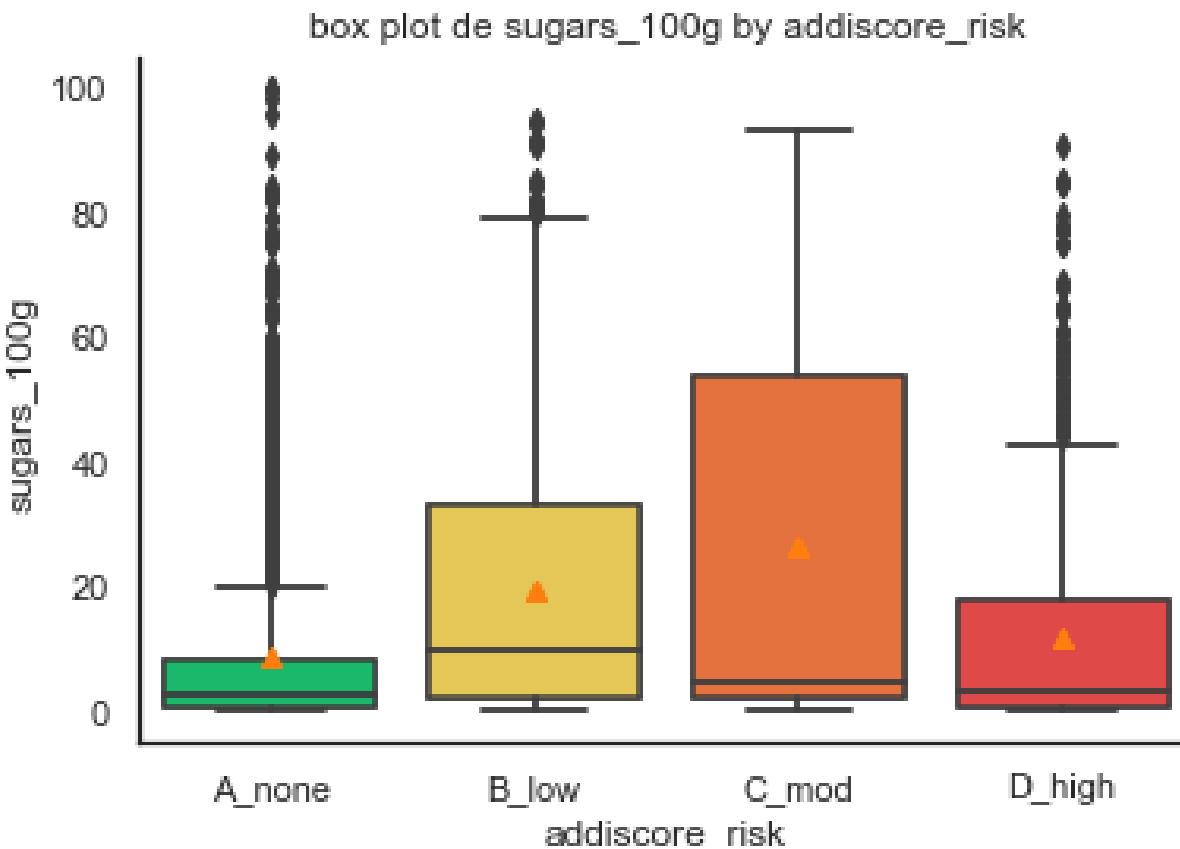
Peu de corrélation entre additifs et valeur nutritionnelle

(Pearson) correlations entre les variables numériques



	addiscore_risk	A_none	B_low	C_mod	D_high
nutriscore_grade					
a	68.27	21.93	1.44	8.36	
b	46.76	28.56	2.22	22.46	
c	39.82	32.41	2.26	25.51	
d	32.30	32.77	5.44	29.49	
e	21.45	42.37	3.90	32.29	
All	39.86	32.09	3.38	24.67	

Les risques sont dans les produits moins sucré



Tests de normalité des distributions

A_none : Kolmogorov-Smirnov : stat=0.538, p=0.000***

B_low : Kolmogorov-Smirnov : stat=0.741, p=0.000***

C_mod : Shapiro-Wilk : stat=0.779, p=0.000***

D_high : Kolmogorov-Smirnov : stat=0.586, p=0.000***

La variable dépendante (sugars_100g) n'est pas distribuée normalement pour chaque groupe dans addiscore_risk.

Test de homoscédascité

Levene homoscédascité : stat=68.016, p=0.000***

Levene test, reject H0: variance of groups is not equal

Test H-ANOVA (Kruskal-Wallis)

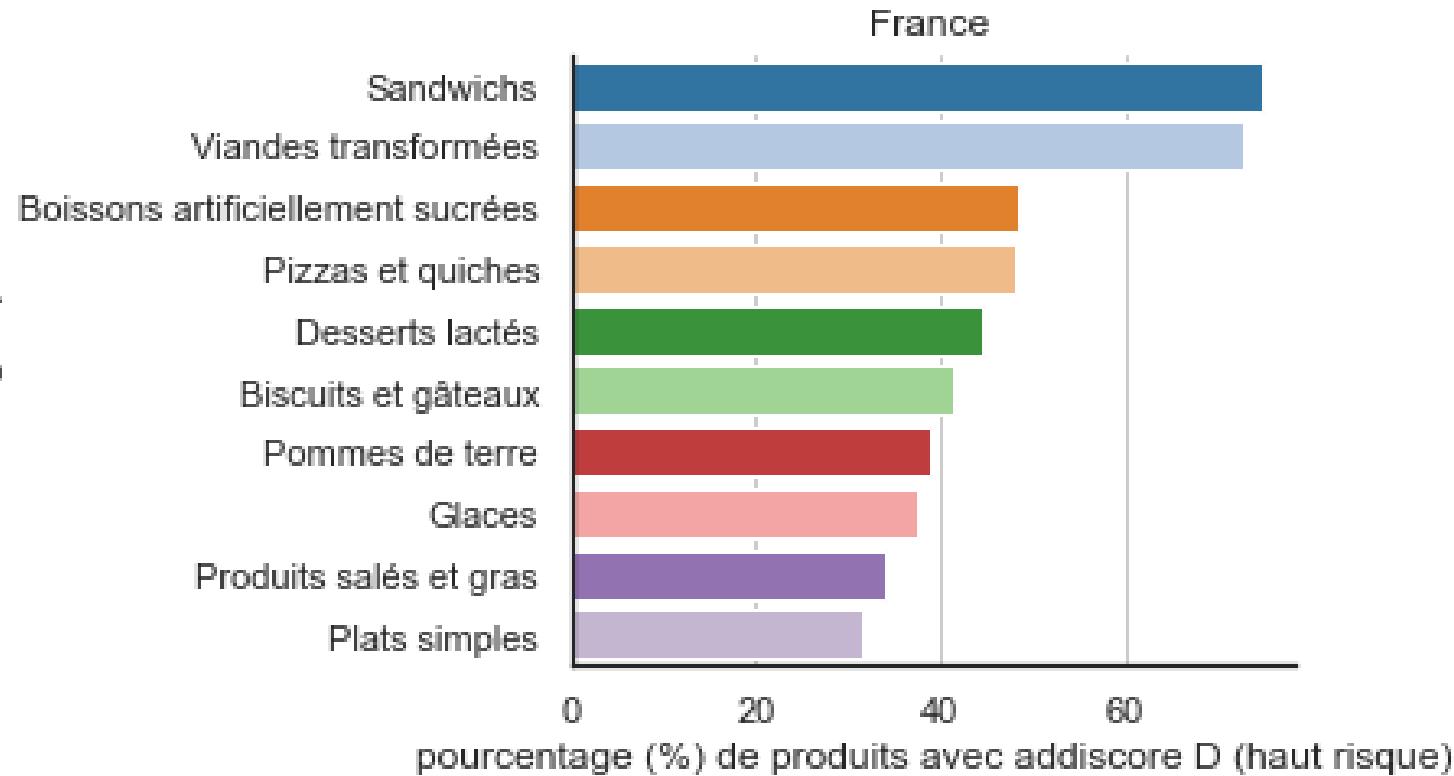
Kruskal-Wallis H-test : stat=245.355, p=0.000***

les médianes sont significativement différentes (reject H0)

Les catégories de produit plus à risque

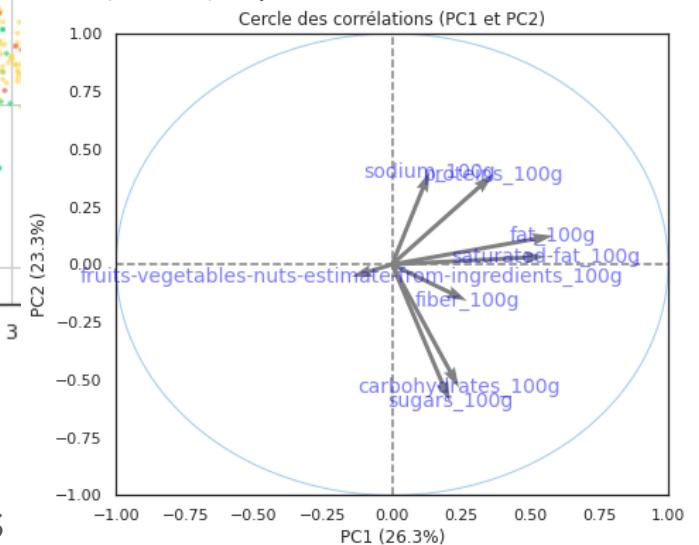
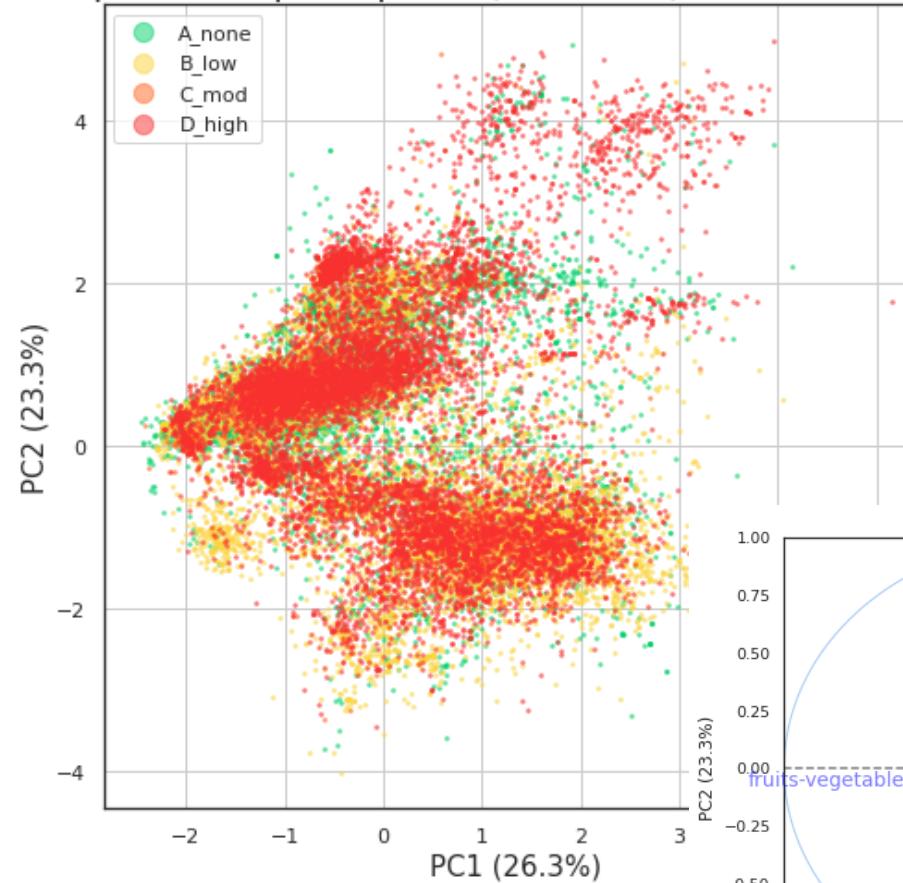
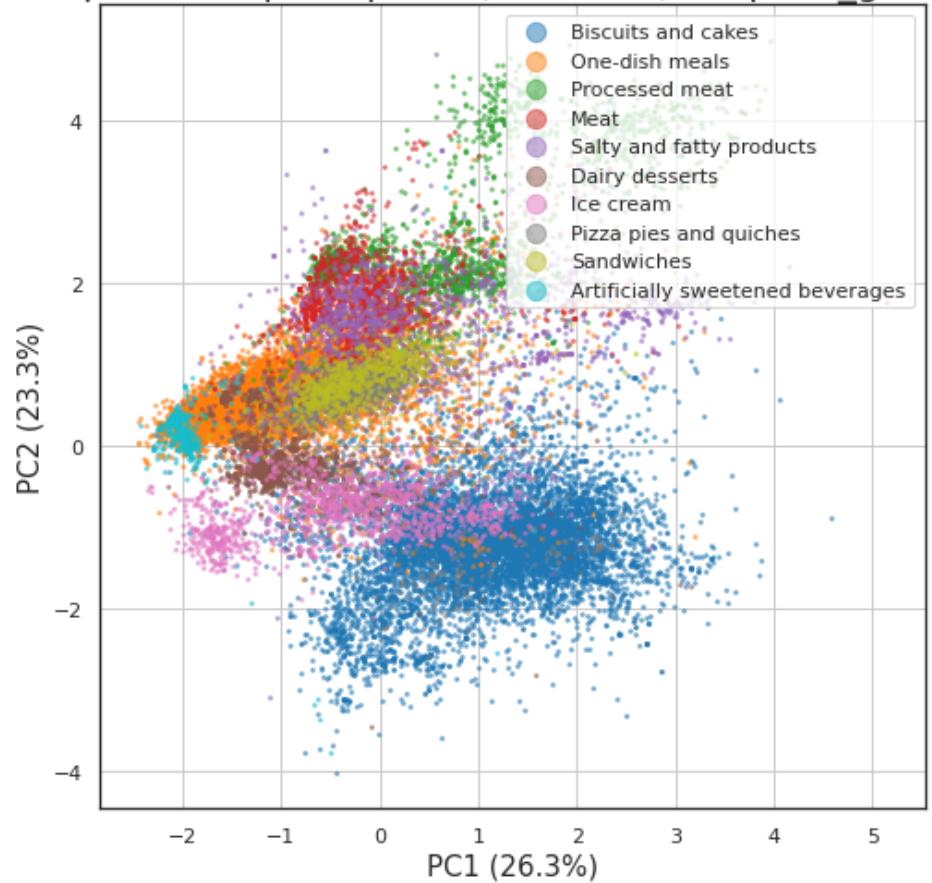
Top 10 catégories avec additifs de risque élevés de sur-exposition

PNNS groupe 2



FAST FOOD sont identifiables mais pas ses risques

Composantes principales (PC1, PC2) vs. pnns_groups_2 Composantes principales (PC1, PC2) vs. addiscore_risk



Composants principaux basée sur les valeurs nutritionnelles des produits

ADDISCORE classification

Viandes transformés

Jambon de porc potassium lactate sodium phosphates flavorings
sodium erythorbate
épices antioxydant : ascorbate de sodium ferments lactic acid starter culture
paprika sodium phosphate
sirop de glucose ail Viande de porc Speisesalz
sodium nitrite
spices sodium diacetate
arômes citric acid Gewürze conservateur : nitrite de sodium
dextrose Pork corn syrup
flavoring beef arômes naturels Kaliumjodat



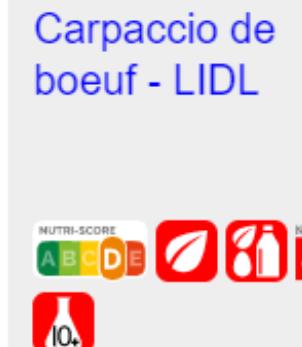
5



0



1



20

ADDISCORE



NUTRISCORE

A

C

B

D

06. Conclusion : Axes d'améliorations

Conclusion

- On ne sait pas ce qu'on mange (*80% ingrédients manquants*)
- > 50 % produits contient additifs à risque (*>75% des sandwichs*)
- Aucun façon de distinguer les produits avec des additifs à risque
- ADDISCORE score - pour des points d'additifs dans le NUTRISCORE
- ADDISCORE risk
 - pour identifier les produits avec additifs à risque de sur exposition

Axes d'amélioration

ADDISCORE (nombre)

- Optimiser les poids des additifs

ADDISCORE_RISK (catégorie)

- Revoir les additifs à inclure
 - huile de palme,
 - titanium dioxide,
 - ...

NUTRISCORE

- Inclure des points pour ADDISCORE et ADDISCORE RISK

Système de recommandation

- Développer un application pour recommander des produits similaires avec moins d'additifs à risque de sur exposition

Questions ?

images des produits alimentaires (brands): OpenFoodFacts (creative commons)

autres images: Mark Creasey

mrcrcreasey@gmail.com

Merci !

