Préparer des données pour un organisme de santé publique
-
Synthèse

Executive brief – Principes généraux & champ d'application du RGPD

Le RGPD:

- Paquet européen de protection des données (2016)
- Contrôle *a posteriori* avec obligation de documentation de la conformité
- Concerne les données personnelles contenues dans un fichier
- Garantit certains droits
- Interdit certains traitements (sauf exceptions rares)
- Violations à déclarer à la CNIL

Les données personnelles collectées doivent être:

- traitées de manière licite, loyale et transparente
- collectées pour des finalités déterminées, explicites et légitimes
- adéquates, pertinentes et limitées
- exactes et tenues à jour
- conservées pour une durée limitée
- conservées de manière sécurisée
- responsable du traitement a une obligation de documentation

Santé publique France

open FOOD facts

Executive brief – Principes généraux & champ d'application du RGPD

- Données anonymisées : adresses email supprimées (meilleur que la pseudonymisation et offre plus de flexibilité en matière de durée maximale de conservation des données)
- Données retraitées : identification impossible "avec des moyens raisonnables"
- Durée de conservation : données anonymisées donc pas de limite
- Mises à jour : non nécessaires

➢ N.B.: ces règles de gestion ne s'appliquent qu'aux données utilisées pour cette analyse, pas aux données collectées par le site

Santé publique France

open FOOD facts

Sommaire

Santé publique France

open FOOD facts

1 – Téléchargement, vérification & sélection des données

1.1 – Import & exploration

Fichier initial

- Grande taille : plus de 320,000 lignes et 129 colonnes
- Erreurs de formatage du csv ralentissant l'import
- Médiocre qualité : plus de 83% de valeurs manquantes et 20 colonnes entièrement vides

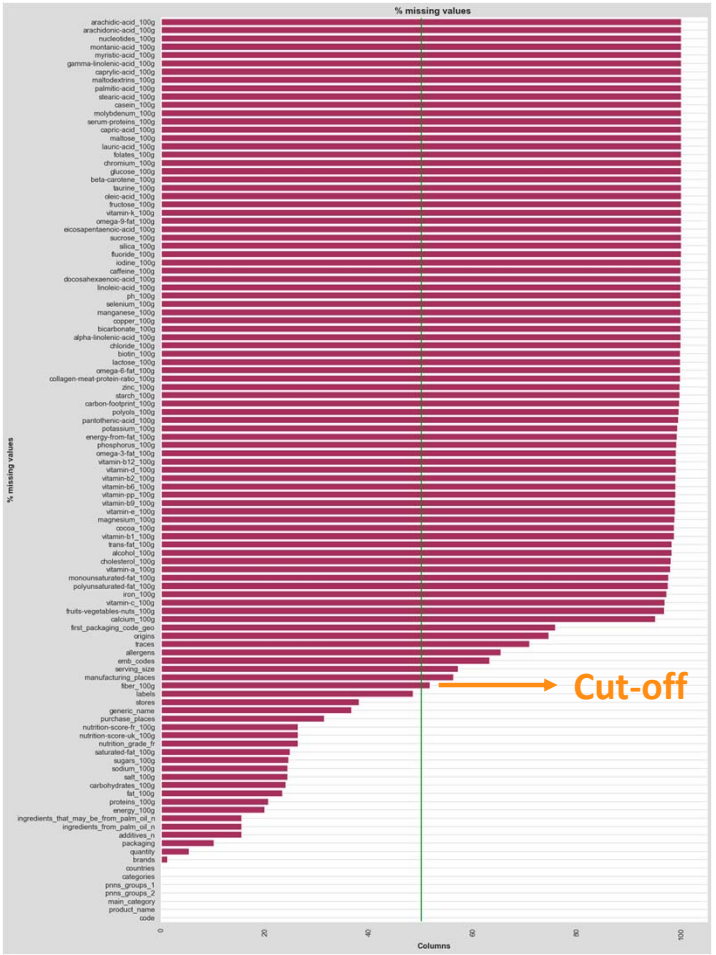
➢ Techniques utilisées pour permettre l'import: spécification du séparateur et du signe de retour à la ligne, dictionnaires de colonnes et de types de données, spécification de l'usage mémoire et option « skip » sur les lignes mal formatées

➢ Gain: réduction de l'utilisation de la mémoire de 847 Mo à un peu plus de 300 Mo.

1 – Téléchargement, vérification & sélection des données

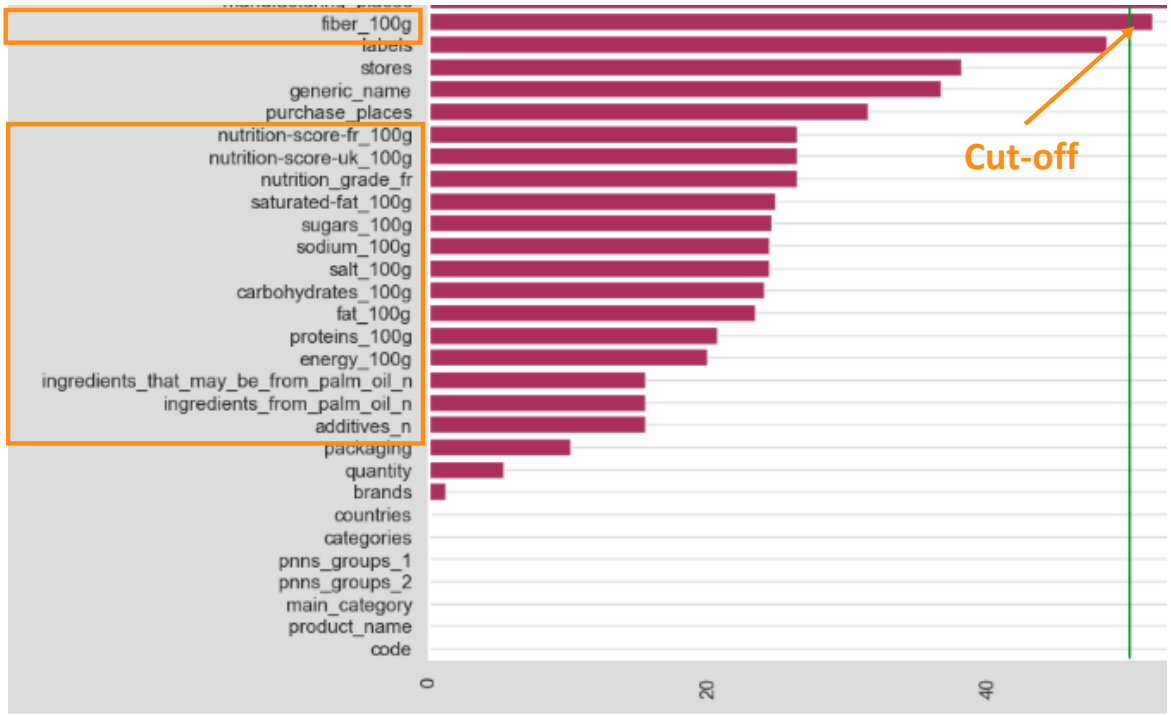
1.2 - Sélection des variables

- Clé primaire: code produit – pas de duplicatas dans le fichier (la suppression des "mauvaises" lignes à l'import a également supprimé 23 doublons dans la colonne ['code'] qui étaient des NULLs dus à des lignes mal formées et décalées vers la droite)
- 1ère sélection : on conserve les lignes ayant un nom de produit et un label PNNS Groups 1 et 2 non-nul et non « inconnu » => reste environ 68,000 lignes
- Sur les données restantes, environ les 2/3 des colonnes sont vides a plus de 80%
  - Cut-off : 53% de valeurs NULL maxi (= taux remplissage fibres)
  - Création d'une **fonction load_food_data** qui synthétise ces étapes



1 – Téléchargement, vérification & sélection des données

1.2 - Sélection des variables



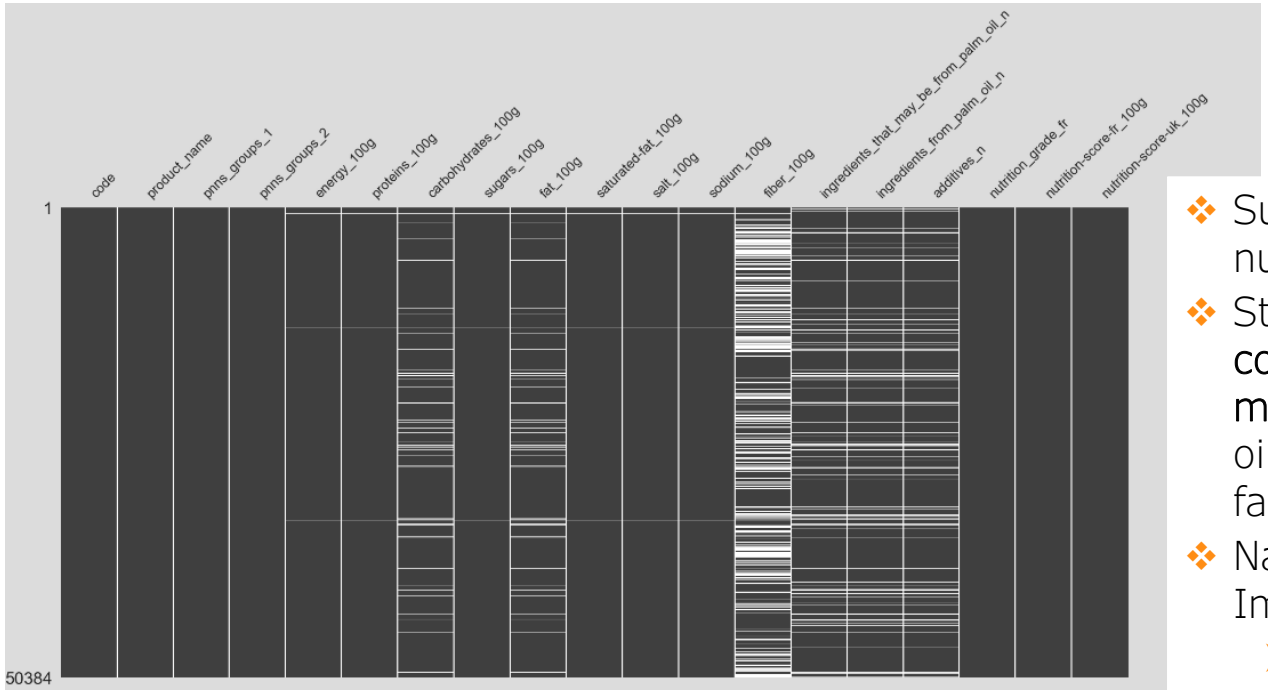
- +68,000 lignes et 29 colonnes, dont 10 seront peu utiles pour notre analyse: labels, stores, generic_name, purchase_places, packaging, quantity, brands, countries et categories, qui seront supprimées
- Variable cible : nutrition_grade_fr (catégorielle [A-E])
- Variables numériques discrètes : nutrition-score-fr_100g, nutrition-score-uk_100g, additives_n, ingredients_that_may_be_from_palm_oil_n et ingredients_from_palm_oil_n
- Les autres variables étant numériques continues

Santé publique France

open FOOD facts

1 – Téléchargement, vérification & sélection des données

1.3 - Nettoyage & retraitements – valeurs manquantes



- Suppression des lignes sans nutriscore : reste +50,000 lignes
- Stratégie **d'imputation** sur la base de **connaissances métier, suppression** ou **mise à zéro** pour les colonnes palm oil, energy_100g, fiber_100g, fat_100g et carbohydrates_100g
- NaNs restants remplis avec l'Iterative Imputer de Scikit-Learn
  - Création d'une **fonction fill_food_data** qui synthétise ces étapes

Santé publique France

open FOOD facts

1 – Téléchargement, vérification & sélection des données

1.3 - Nettoyage & retraitements – valeurs aberrantes

- Correction des anomalies mathématiques par suppression des lignes où:
  - poids total > 100g (259 lignes)
  - saturated_fat_100g > fat_100g (594 lignes)
  - calories totales > calories de graisse pure + marge 20% FDA (12 lignes)
  - carbohydrates_100g > sugar_100g + fiber_100g + marge 20% FDA (3,853 lignes)
  - énergie calculée par 100g est différente de plus de 20% de celle déclarée (3,391 lignes)

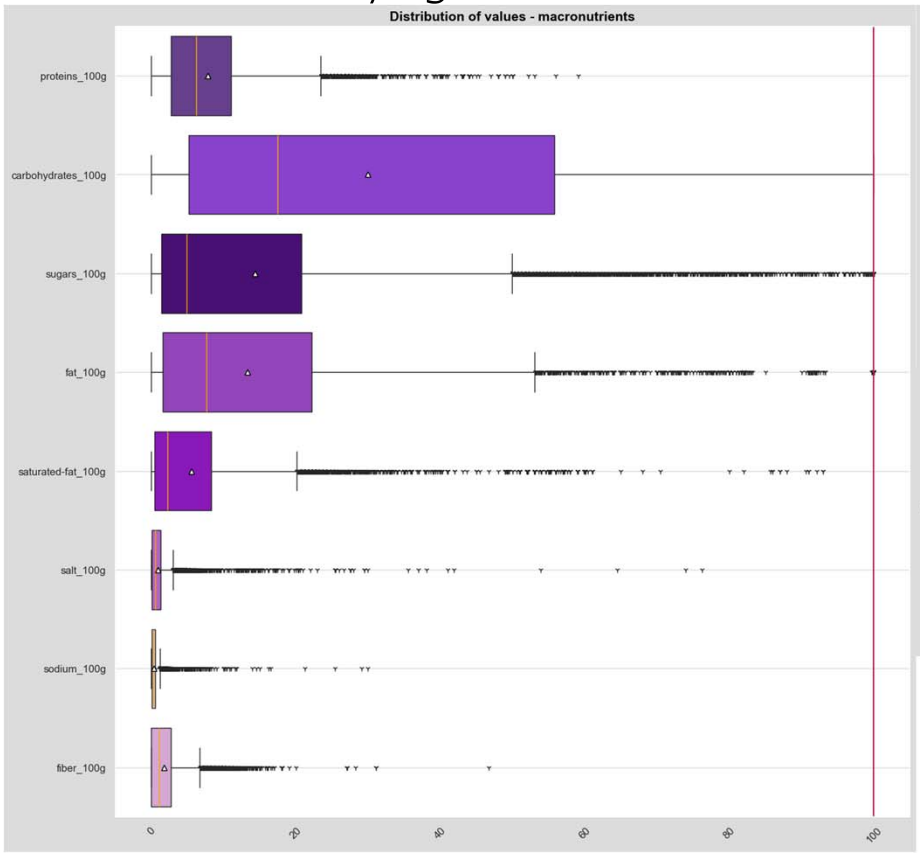
- Remplacement des valeurs négatives par leur valeur absolue (1,749 lignes)

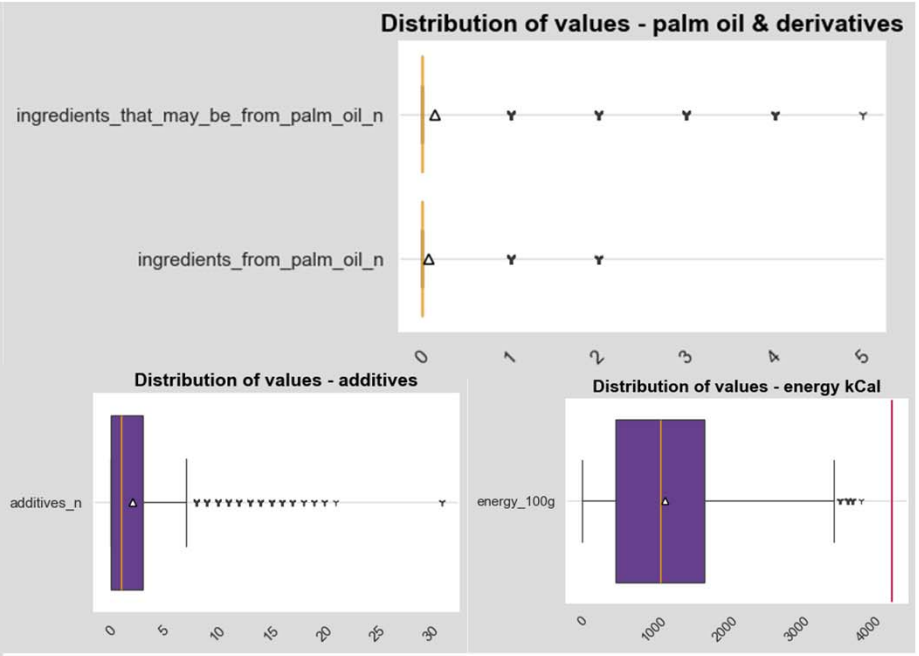
- Remplacement des valeurs décimales par des valeurs entières pour les colonnes palm oil et additives_n

  - Création d'une **fonction food_data_extrim** qui synthétise ces étapes

1 – Téléchargement, vérification & sélection des données

1.3 - Nettoyage & retraitements – valeurs aberrantes





- Autres outliers considérés comme atypiques et non aberrants

1 – Téléchargement, vérification & sélection des données

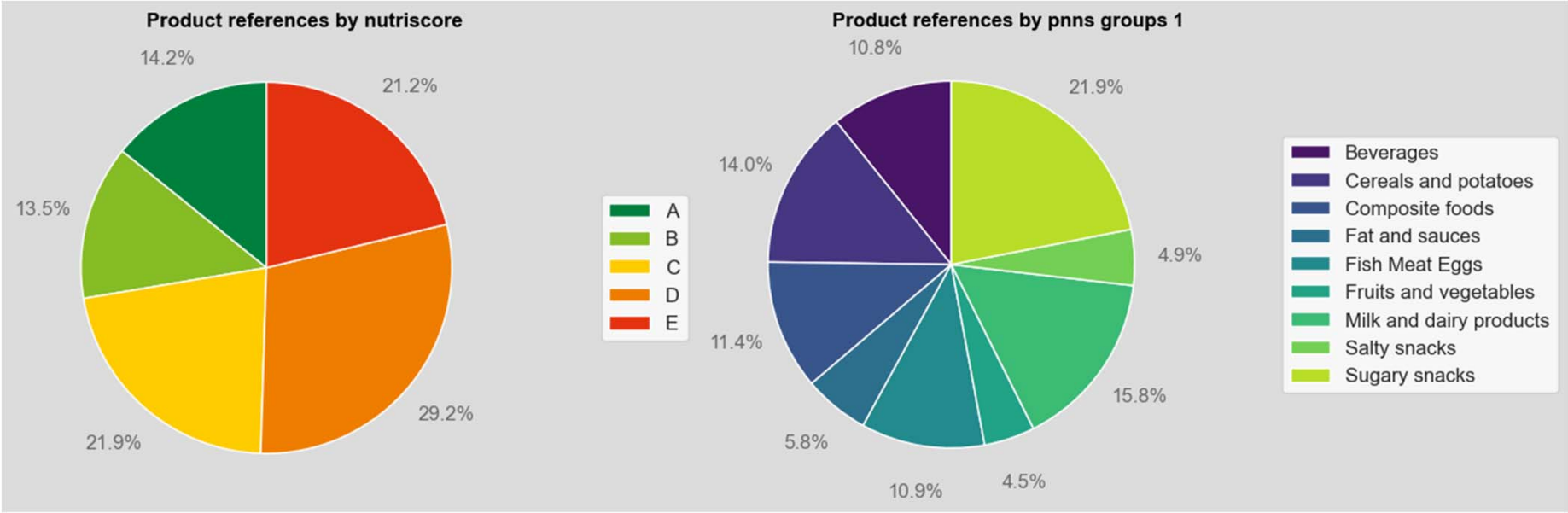
1.3 - Nettoyage & retraitements – feature engineering

- Création de plusieurs variables booléennes à partir du nom des produits:
  - diet_product (produits sans gluten ou allégés en sucre et/ou matières grasses)
  - organic_product (produits bio)
  - contains_palm_oil (contient de l'huile de palme ou un de ses dérivés)
  - can_eat_daily (groupement des nutriscores A et B)

- Utilisées pour les autres analyses multivariées (cf. 2.5 ci-dessous)

2 – Analyse exploratoire des données

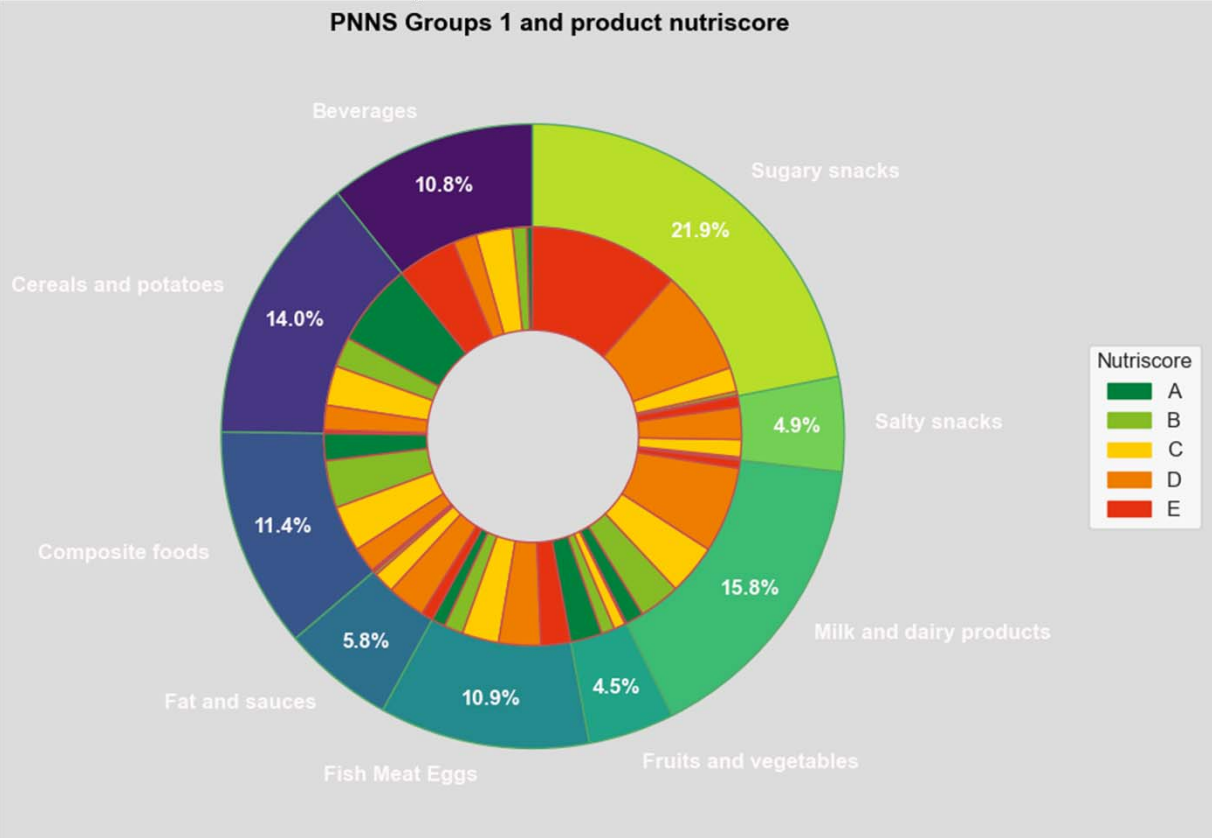
2.1 - Analyses univariées



- Base de données qui recense essentiellement des **produits hautement transformés** (UPF)

2 – Analyse exploratoire des données

2.1 - Analyses univariées



- Catégories Sugary snacks et Salty snacks contiennent presque exclusivement des produits à mauvais nutriscore (C, D ou E) et qui ne doivent pas être consommés quotidiennement (UPF)…
- …contrairement aux catégories Cereals and potatoes et Fruits and vegetables, qui sont dominées par les « bons » nutriscores et peuvent être consommées quotidiennement

2 – Analyse exploratoire des données

2.1 - Analyses univariées

- Test de Kolmogorov-Smirnov sur toutes les variables numériques continues

- Aucune ne suit une distribution gaussienne
  - Utilisation de tests non-paramétriques

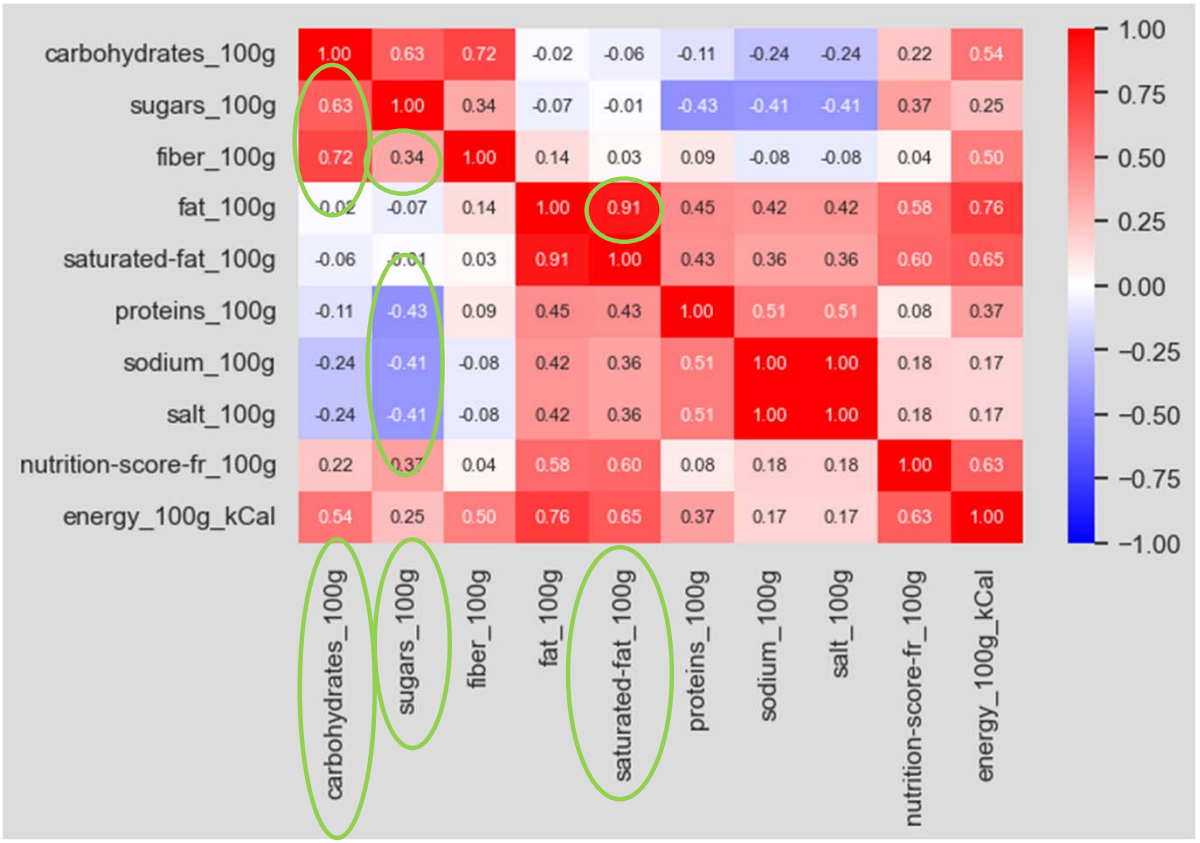
```
********
carbohydrates_100g
Statistic = 0.1804
p-value = 0.0000
Reject H0 : carbohydrates_100g does not follow a normal distribution
********
sugars_100g
Statistic = 0.2262
p-value = 0.0000
Reject H0 : sugars_100g does not follow a normal distribution
********
fiber_100g
Statistic = 0.2162
p-value = 0.0000
Reject H0 : fiber_100g does not follow a normal distribution
********
fat_100g
Statistic = 0.1905
p-value = 0.0000
Reject H0 : fat_100g does not follow a normal distribution
********
saturated-fat_100g
Statistic = 0.2338
p-value = 0.0000
Reject H0 : saturated-fat_100g does not follow a normal distribution
********
proteins_100g
Statistic = 0.1345
p-value = 0.0000
Reject H0 : proteins_100g does not follow a normal distribution
********
sodium_100g
Statistic = 0.2810
p-value = 0.0000
Reject H0 : sodium_100g does not follow a normal distribution
********
salt_100g
Statistic = 0.2809
p-value = 0.0000
Reject H0 : salt_100g does not follow a normal distribution
********
energy_100g_kCal
Statistic = 0.0925
p-value = 0.0000
Reject H0 : energy_100g_kCal does not follow a normal distribution
```

Santé publique France

open FOOD facts

2 – Analyse exploratoire des données

2.2 - Analyse des corrélations



- ρ de Spearman
- mesure **non-paramétrique** de **corrélation de rang** qui mesure la dépendance statistique entre les **classements** de deux variables
- évalue dans quelle mesure la relation entre deux variables peut être décrite à l'aide d'une **fonction monotone**

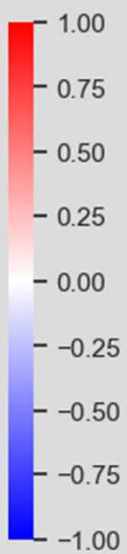
Santé publique France

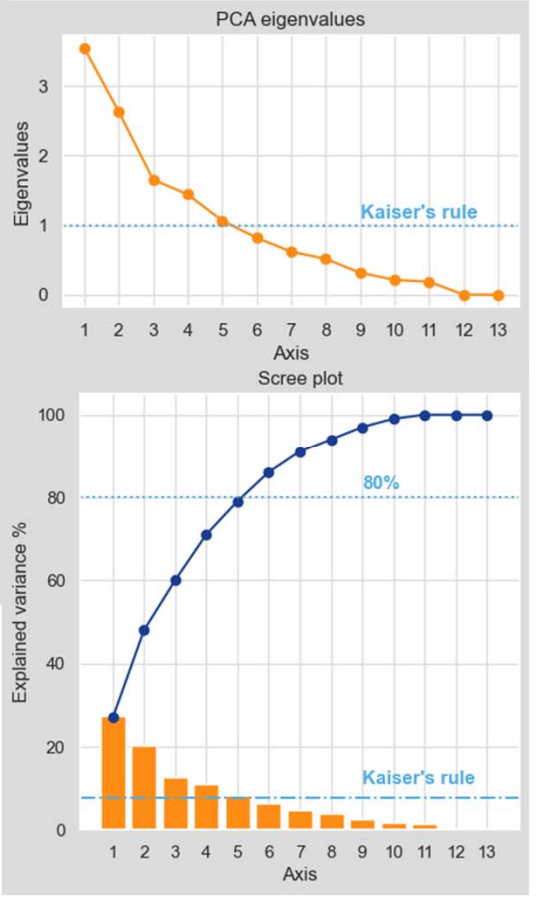
open FOOD facts

2 – Analyse exploratoire des données

2.3 - ACP

	F1	F2	F3	F4	F5
carbohydrates_100g	0.29	-0.34	0.21	0.39	0.03
sugars_100g	0.27	-0.33	0.12	0.07	0.48
fiber_100g	0.20	-0.20	0.06	0.51	-0.42
fat_100g	0.40	0.21	-0.28	-0.18	-0.11
saturated-fat_100g	0.38	0.18	-0.29	-0.26	0.01
proteins_100g	0.07	0.37	-0.10	0.17	-0.42
sodium_100g	0.04	0.50	0.35	0.24	0.19
salt_100g	0.04	0.50	0.35	0.24	0.19
ingredients_that_may_be_from_palm_oil_n	0.10	-0.06	0.43	-0.34	-0.40
ingredients_from_palm_oil_n	0.20	-0.09	0.26	-0.15	-0.15
additives_n	0.09	-0.06	0.51	-0.39	-0.10
energy_100g_kCal	0.49	0.00	-0.09	0.14	-0.14
nutrition-score-fr_100g	0.44	0.09	0.01	-0.15	0.34





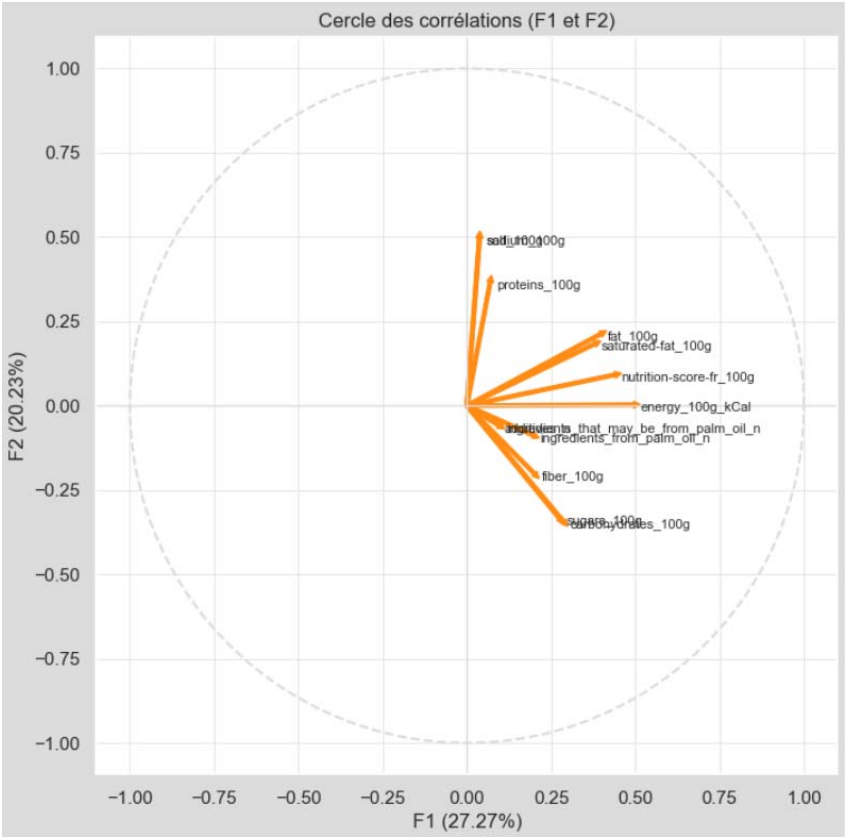
- 1er axe : valeur énergétique des produits (teneur en matières grasses et en sucres)
- 2ème axe : produits riches en sel et en protéines, et pauvres en fibres et en glucides

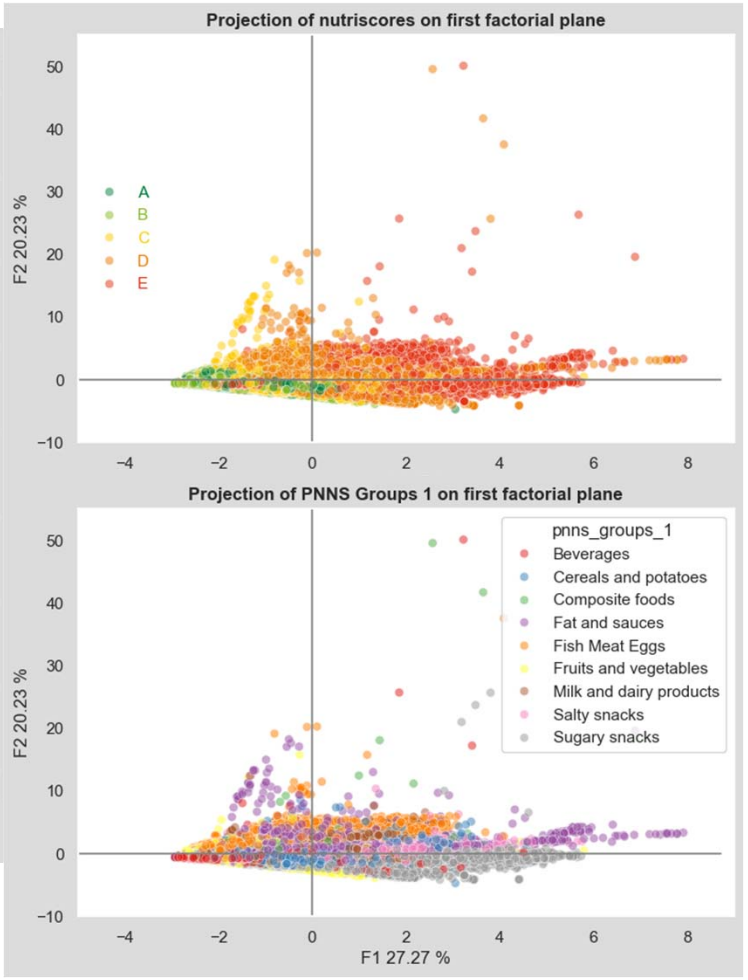
Santé publique France

open FOOD facts

2 – Analyse exploratoire des données

2.3 - ACP

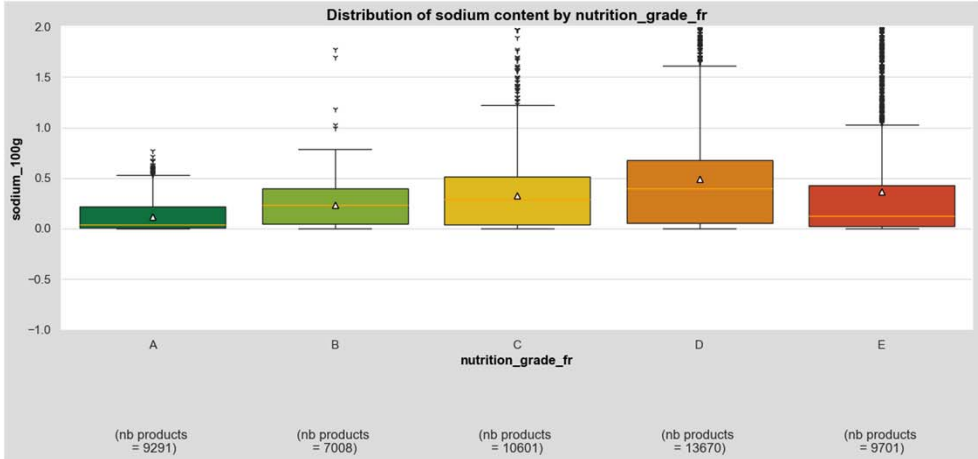


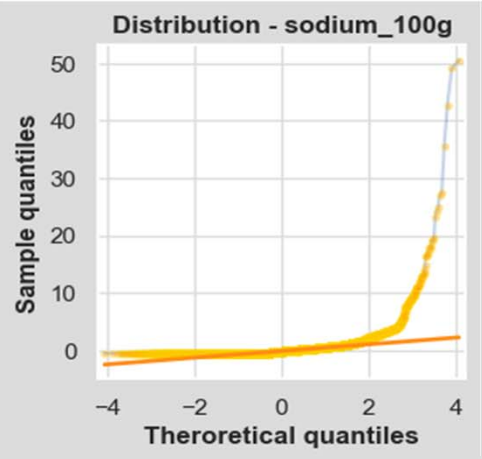


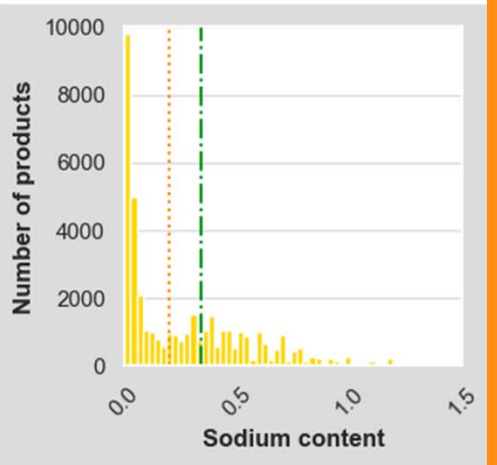
2 – Analyse exploratoire des données

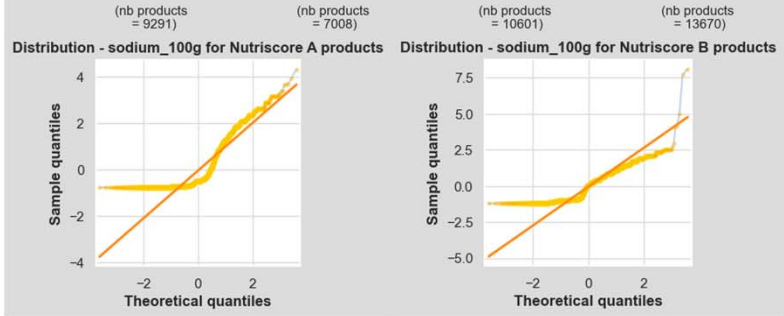
2.4 - ANOVA – cas du sodium

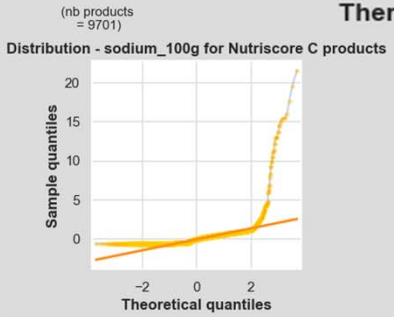
- <u>Hypothèse 1 :</u> distribution Gaussienne – non vérifiée

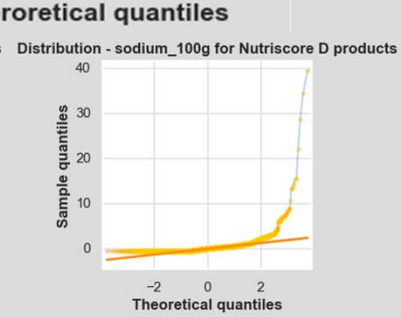


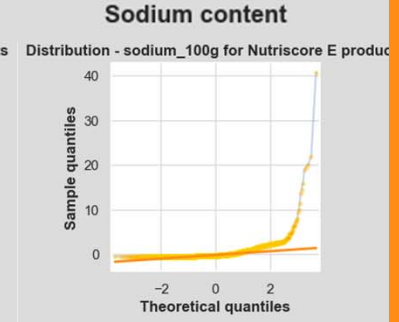












2 – Analyse exploratoire des données

2.4 - ANOVA – cas du sodium

- Hypothèse 2 : homoscédasticité– non vérifiée (test de Levene)

```
nutrition_grade_fr
A    0.150749
B    0.190205
C    0.496272
D    0.727426
E    0.726991
Name: sodium_100g, dtype: float64
```

```
W statistic : 469.55917056586605
p-value : 0.0
Reject the null hypothesis, the variance of the groups in not homogenous.
```

- Hypothèse 3 : observations indépendantes – considérée vérifiée en l'absence d'information contraire

  - Le calcul du η^2 conclut à une **faible** relation entre le contenu en sodium et le nutriscore

2 – Analyse exploratoire des données

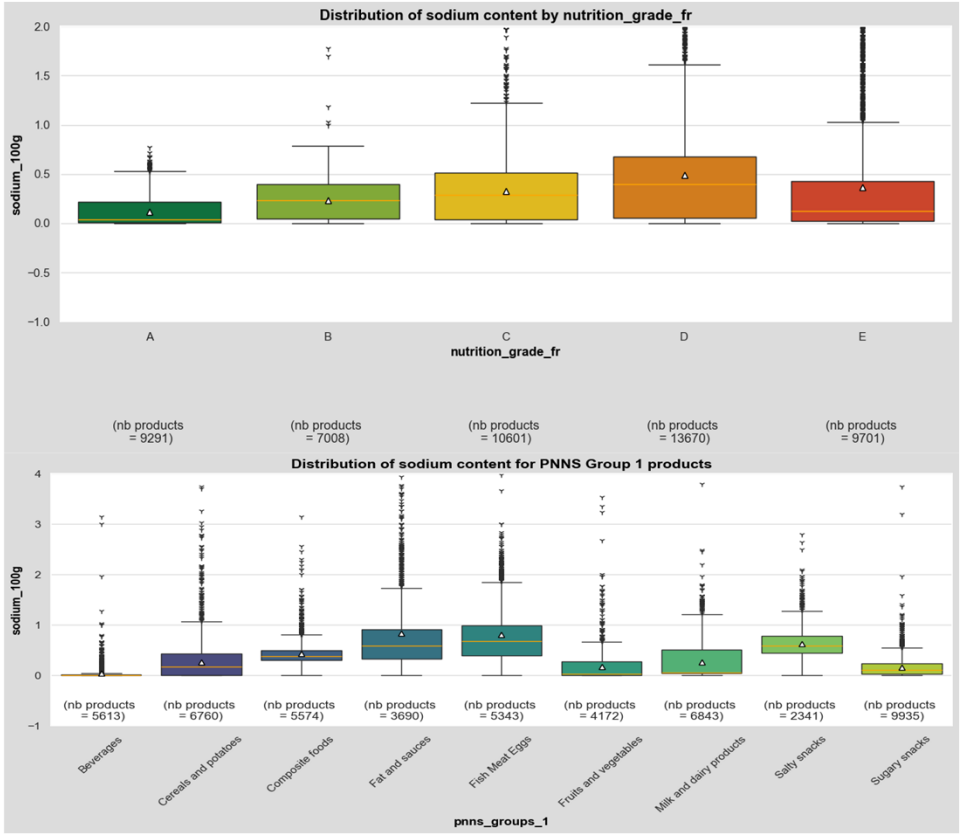
2.4 - Test de Kruskal-Wallis sur le sodium

- Test significatif au seuil α = 5%

```
Stat: 3789.4503984452213
p-value: 0.000000e+00
```

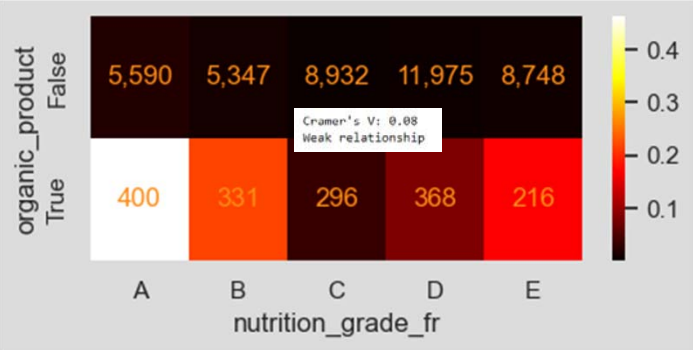
- <u>Test post-hoc de Dunn :</u> tous les nutriscores sauf B et E ont des contenus medians en sodium différents

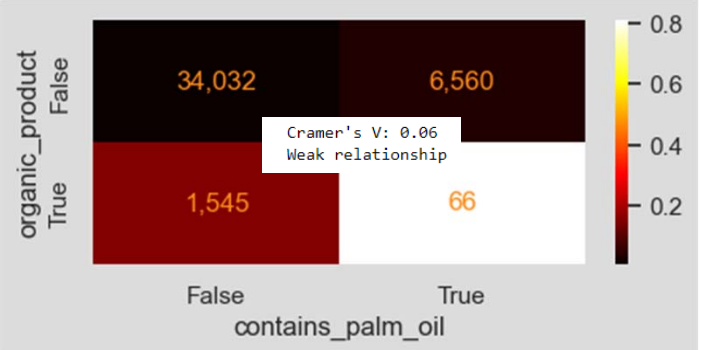
	A	B	C	D	E
A	1.000000e+00	2.092099e-201	0.000000e+00	0.000000e+00	1.232736e-219
B	2.092099e-201	1.000000e+00	9.030128e-11	3.517201e-132	5.229128e-02
C	0.000000e+00	9.030128e-11	1.000000e+00	6.024807e-94	8.611896e-22
D	0.000000e+00	3.517201e-132	6.024807e-94	1.000000e+00	3.415926e-206
E	1.232736e-219	5.229128e-02	8.611896e-22	3.415926e-206	1.000000e+00

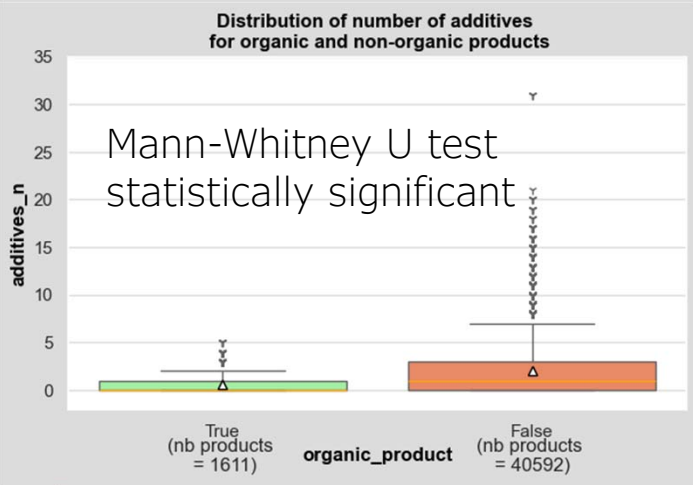


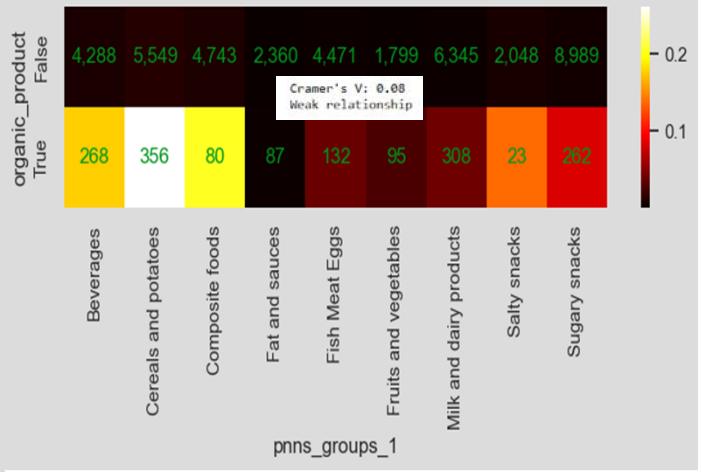
2 – Analyse exploratoire des données

2.5 - Autres analyses multivariées – produits biologiques









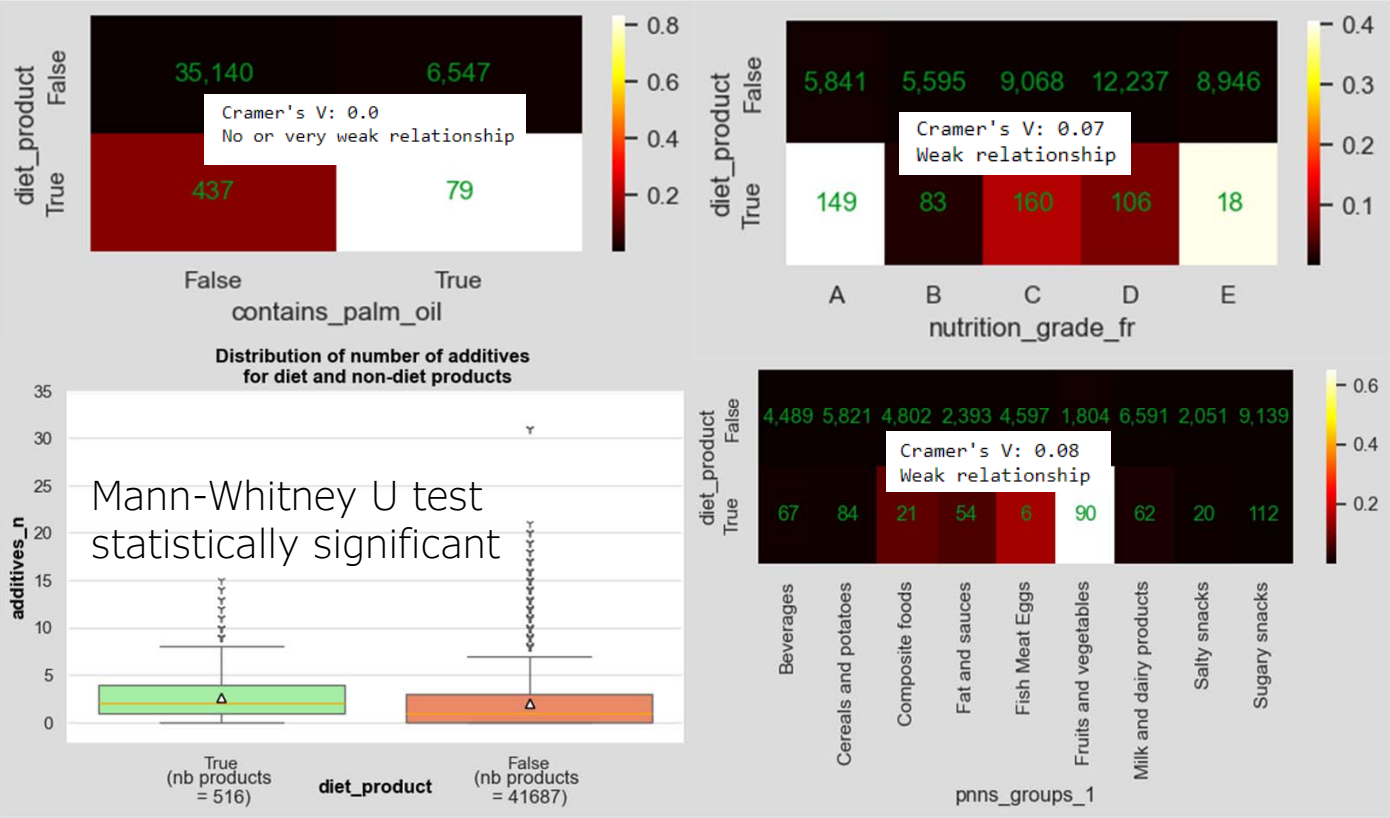
Le caractère biologique d'un produit n'est **pas indépendant**:

- de son contenu en huile de palme
- de son nutriscore
- de son contenu en additifs et
- de son PNNS Groups 1

➢ Même si elles sont d'intensité faible, ces relations sont **statistiquement significatives**

2 – Analyse exploratoire des données

2.5 - Autres analyses multivariées – produits allégés



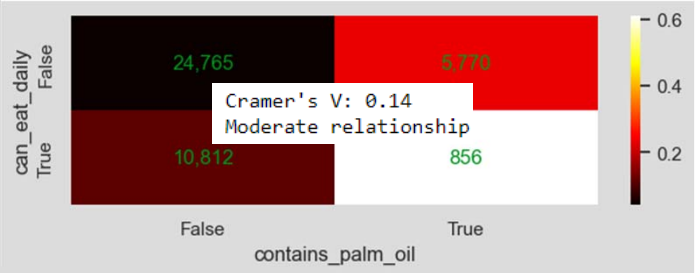
Le caractère allégé d'un produit est indépendant de son contenu en huile de palme mais n'est **pas indépendant**:

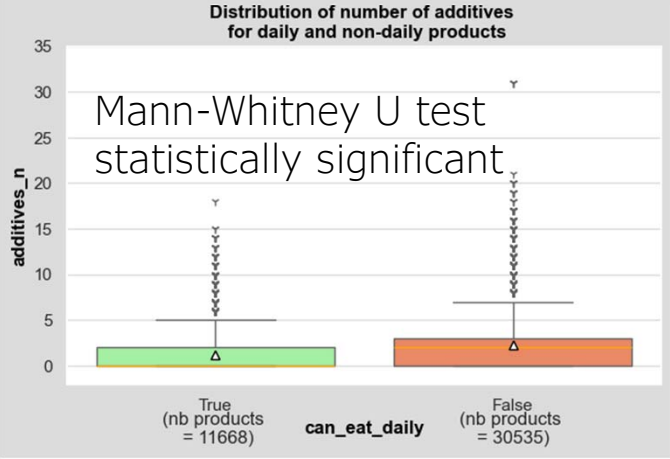
- de son nutriscore
- de son contenu en additifs et
- de son PNNS Groups 1

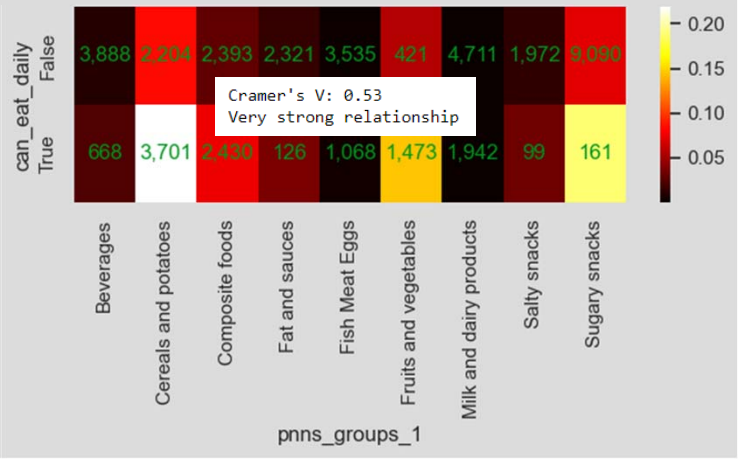
➢ Même si elles sont d'intensité faible, ces relations sont **statistiquement significatives**

2 – Analyse exploratoire des données

2.5 - Autres analyses multivariées – produits à conso. quotidienne







Consommation quotidienne recommandée d'un produit n'est **pas indépendante**:

- de son contenu en huile de palme
- de son caractère allégé
- de son contenu en additifs et
- de son PNNS Groups 1
  - Même si elles sont d'intensité variable, ces relations sont **statistiquement significatives**

Santé publique France

open FOOD facts

Recommandations & Conclusion

- Création d'un système de suggestion ou d'autocomplétion possible compte-tenu des relations mathématiques et statistiques entre nos variables et le nutriscore…
- … mais approche principale devrait consister à améliorer la complétion des données en amont (data validation, formulaires incomplets etc…)