Projet 3
-
Préparez des données pour un organisme de santé publique

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 –> avril 2024

Sommaire

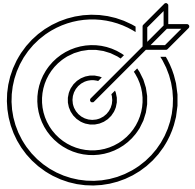
Problématique

Santé publique France est un établissement public à caractère administratif français, placé sous la tutelle du ministère chargé de la santé, crée le 1er mai 2016.

Cet établissement souhaite améliorer sa base de données **Open Food Facts** qui est mise à disposition de particuliers et d'organisations afin de leur permettre de connaître la qualité nutritionnelle de produits. Mise en place d'un système **d'auto-complétion** pour aider les usagers à remplir plus efficacement la base de données.

Missions :

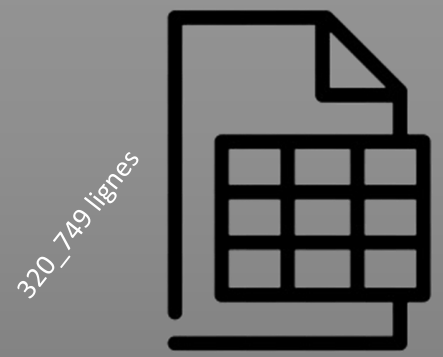
- Nettoyage des données
- Analyses univariées
- Analyse multivariée
- Etudier la faisabilité du système d'auto-complétion
- Rappel RGPD



Sommaire

Présentation du jeu de données

fr.openfoodfacts.org.products.csv



➢ <u>Descriptif des produits alimentaires :</u>

- Les **informations générales** des produits : code barre, nom, date de modification, etc.
- Un ensemble de **tags** : catégorie du produit, localisation, origine, etc.
- Les **ingrédients** composant les produits et leurs additifs éventuels.
- Des **informations nutritionnelles** : quantité en grammes d'un nutriment pour 100 grammes du produit.

➢ <u>Valeurs manquantes :</u>

- 39_604_863 de NaN pour 51_961_338 observations (76.22 %)
- 154 / 162 colonnes concernées

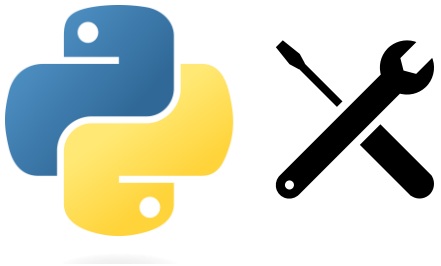
➢ <u>Doublons</u>

Pas de doublons sur la colonne du code barre.

Sommaire

Nettoyage des données

1. Filtrage méthode entonnoir
2. Imputation valeurs manquantes
3. Traitement de variables catégorielles
4. Re-calcul du Nutri-score



Nettoyage des données

Nom	Utilisation	Fonctions spécifiques
Anaconda	Gestion de package Gestion d'environnement virtuel	Conda : installation de package via le terminal
Visual Studio Code 1.83.0	Structurer la démarche Exécuter code par étape Expliquer la démarche (markdown)	
Python 3.11.6	Appel aux librairies Boucles for pour générer plusieurs calculs et graphiques	Boucles, listes, dictionnaires, librairies, méthodes
Pandas 2.1.0	Manipulation de données Représentation des données	Manipulation de Dataframe : création, copie, filtres, tris, description, concaténation
Matplotlib 3.7.2 **Seaborn 0.12.2**	Génération de graphiques de visualisation	Barplot, scatterplot, lineplot, distplot, heatmap
Numpy 1.25.2	Manipulation de matrices et fonctions mathématiques	Histogram, argmax, arange, object, number
Missingno 0.5.2	Représentation graphique pour valeurs manquantes	Matrice de NaN
Sklearn 1.3.0	Apprentissage automatique et modélisation statistique	SimpleImputer, KNNImputer, StandardScaler, PCA
Scipy 1.11.2	Calculs de mathématiques complexes ou de problèmes scientifiques	Stats, chi2_contingency, shapiro, kruskal

Nettoyage des données

1/ Filtrage méthode « entonnoir »

a) Filtrage colonne par % remplissage

=> Suppression des colonnes < 25% de remplissage

Nettoyage des données

1/ Filtrage méthode « entonnoir »

a) Filtrage colonne par % remplissage

=> Suppression des colonnes < 25% de remplissage

b) Filtrage colonnes redondantes

=> Suppression des colonnes identiques:
- Tags *(par ex : « countries » / « countries_tags » / « countries_fr »)*
- Temps *(par ex : « created_t » / « created_datetime »)*

Nettoyage des données

1/ Filtrage méthode « entonnoir »

a) Filtrage colonne par % remplissage

=> Suppression des colonnes < 25% de remplissage

b) Filtrage colonnes redondantes

=> Suppression des colonnes identiques:
- Tags *(par ex : « countries » / « countries_tags » / « countries_fr »)*
- Temps *(par ex : « created_t » / « created_datetime »)*

c) Filtrage des colonnes inutiles

=> Suppression colonne : « url » / « creator » / « Nutri-score_UK »

Nettoyage des données

1/ Filtrage méthode « entonnoir »

a) Filtrage colonne par % remplissage

=> Suppression des colonnes < 25% de remplissage

b) Filtrage colonnes redondantes

=> Suppression des colonnes identiques:
- Tags *(par ex : « countries » / « countries_tags » / « countries_fr »)*
- Temps *(par ex : « created_t » / « created_datetime »)*

c) Filtrage des colonnes inutiles

=> Suppression colonne : « url » / « creator » / « Nutri-score_UK »

d) Filtrage par pays

=> Conservation des produits disponibles en France

Nettoyage des données

1/ Filtrage méthode « entonnoir »

a) Filtrage colonne par % remplissage

=> Suppression des colonnes < 25% de remplissage

b) Filtrage colonnes redondantes

=> Suppression des colonnes identiques:
- Tags *(par ex : « countries » / « countries_tags » / « countries_fr »)*
- Temps *(par ex : « created_t » / « created_datetime »)*

c) Filtrage des colonnes inutiles

=> Suppression colonne : « url » / « creator » / « Nutri-score_UK »

d) Filtrage par pays

=> Conservation des produits disponibles en France

e) Filtrage valeurs aberrantes

=> Application de **règles métiers** :
- 0 ≤ Energie sur 100g ≤ 3_762 kJ
- 0 ≤ Quantité sur 100g ≤ 100g
- Acide gras trans / saturé ≤ Lipide
- Sucres ≤ Glucide

Nettoyage des données

2. Imputation valeurs manquantes

a) Colonnes essentielles

=> **Suppression des produits** où il manque le code-barre, le nom du produit ou le pays

Nettoyage des données

2. Imputation valeurs manquantes

a) Colonnes essentielles

=> **Suppression des produits** où il manque le code-barre, le nom du produit ou le pays

b) Ingrédients mineurs

=> **Imputation par la valeur « 0 »** des valeurs manquantes des ingrédients dont l'affichage sur l'emballage n'est pas obligatoire
(fibres, sodium, calcium, vitamines, etc…)

c) Ingrédients majeurs

=> **Imputation par l'algorithme KNN** des valeurs manquantes des ingrédients dont l'affichage sur l'emballage est obligatoire
(énergie, lipide, acides gras saturés, glucide, sucre, protéine, sel)

Nettoyage des données

3. Traitement de variables catégorielles

a) Labélisation valeurs manquantes

=> Attribution valeur « inconnu »

Nettoyage des données

3. Traitement de variables catégorielles

a) Labélisation valeurs manquantes

=> Attribution valeur « inconnu »

b) Séparation des chaînes de caractère

=> séparation des valeurs par la virgule si nécessaire

categories_fr
Plats préparés,Plats à base de viande,Plats à ...

Nettoyage des données

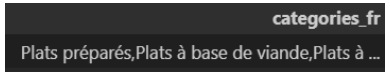
3. Traitement de variables catégorielles

a) Labélisation valeurs manquantes

=> Attribution valeur « inconnu »

b) Séparation des chaînes de caractère

=> séparation des valeurs par la virgule si nécessaire



c) Suppression sources d'erreur

=> Uniformisation bas de casse, suppression des accents, des caractères spéciaux, des espaces vides, des préfixes relatifs aux pays

Nettoyage des données

3. Traitement de variables catégorielles

a) Labélisation valeurs manquantes

=> Attribution valeur « inconnu »

b) Séparation des chaînes de caractère

=> séparation des valeurs par la virgule si nécessaire

c) Suppression sources d'erreur

=> Uniformisation bas de casse, suppression des accents, des caractères spéciaux, des espaces vides, des préfixes relatifs aux pays

Comptage des occurrences

=> Pour chaque valeur unique

Nettoyage des données

Comptage des occurrences

« categories_fr » => 6_518 catégories

« main_categories_fr » => 956 catégories

Nettoyage des données

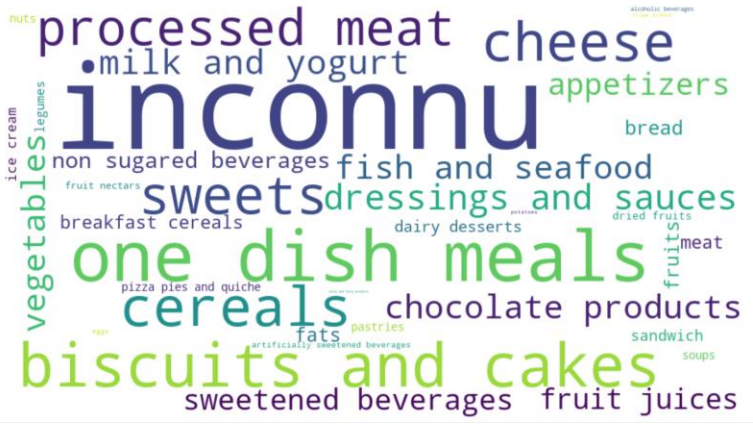
Comptage des occurrences

« categories_fr » => 6_518 catégories

« main_categories_fr » => 956 catégories

« pnns_group_1» => 10 catégories ✓

« pnns_group_2» => 37 catégories ✓



*Pnns : Programme National Nutrition Santé

Nettoyage des données

4. Re-calcul du Nutri-score

Santé publique France

Méthode de calcul de 2017 ✗

⇩

Méthode de calcul de 2023 ✓

Nettoyage des données

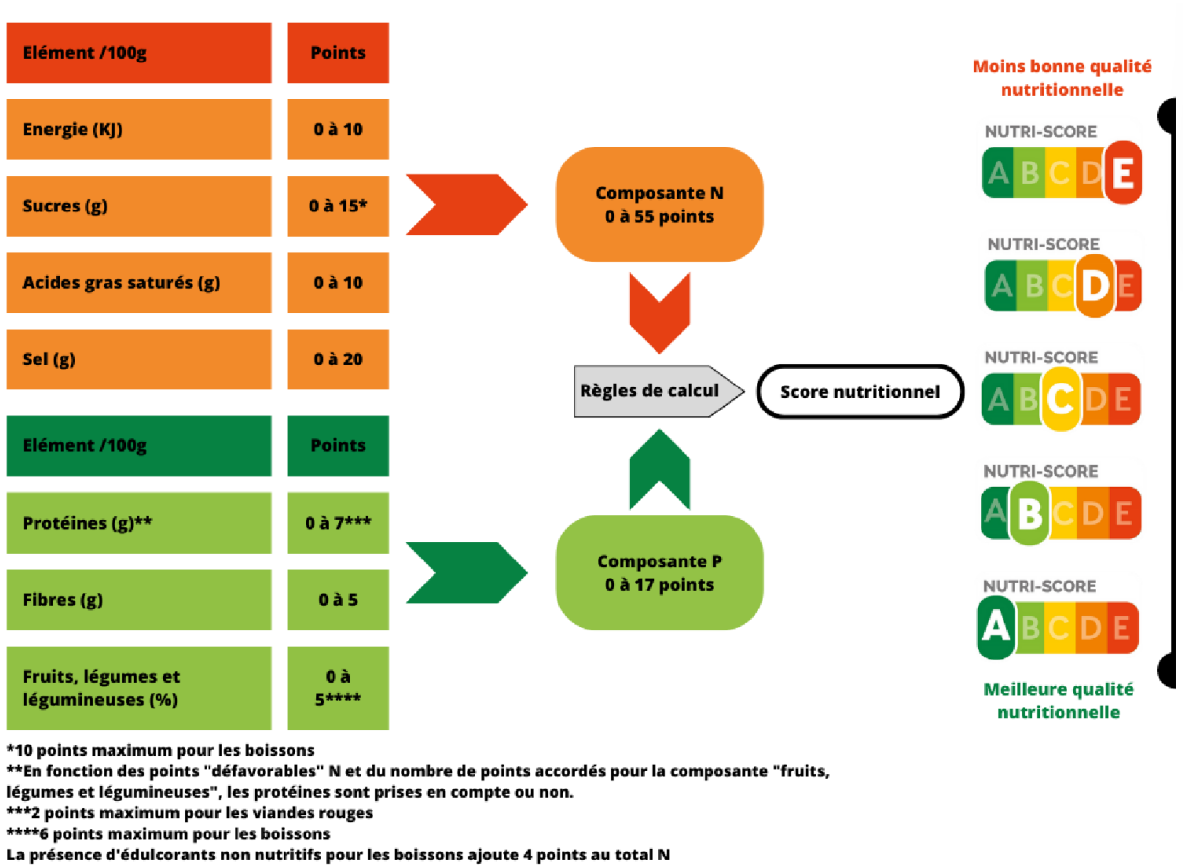
4. Re-calcul du Nutri-score

Santé publique France

Méthode de calcul de 2017 ✗

⇩

Méthode de calcul de 2023 ✓



=> Imputation des valeurs manquantes par les valeurs recalculées :

- Score nutritionnel

- Nutri-score

Nettoyage des données

~~fr.openfoodfacts.org.products.csv~~
dataset_openfoodfacts_cleaned.csv

➢ Descriptif des produits alimentaires :

- Les **informations générales** des produits : code barre, nom, date de modification, etc.
- Un ensemble de **tags** : catégorie du produit, pays de vente.
- Des **informations nutritionnelles** : quantité en grammes d'un nutriment pour 100 grammes du produit.

➢ Valeurs manquantes :

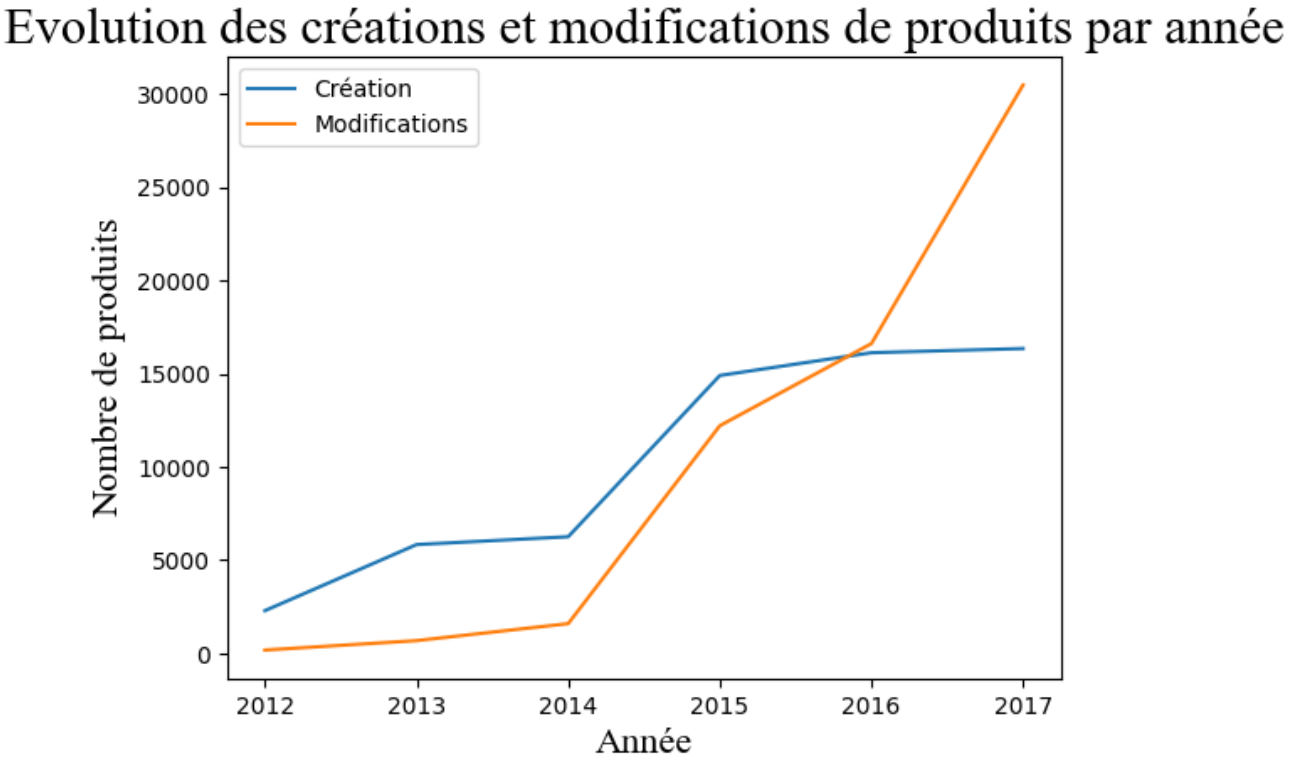
~~- 39_604_863 de NaN pour 51_961_338 observations (76.22 %)~~
- 110_059 de NaN pour 2_222_496 observations (4.95 %)

~~- 154 / 162 colonnes concernées~~
- 5 / 36 colonnes concernées

Sommaire

Analyses univariées

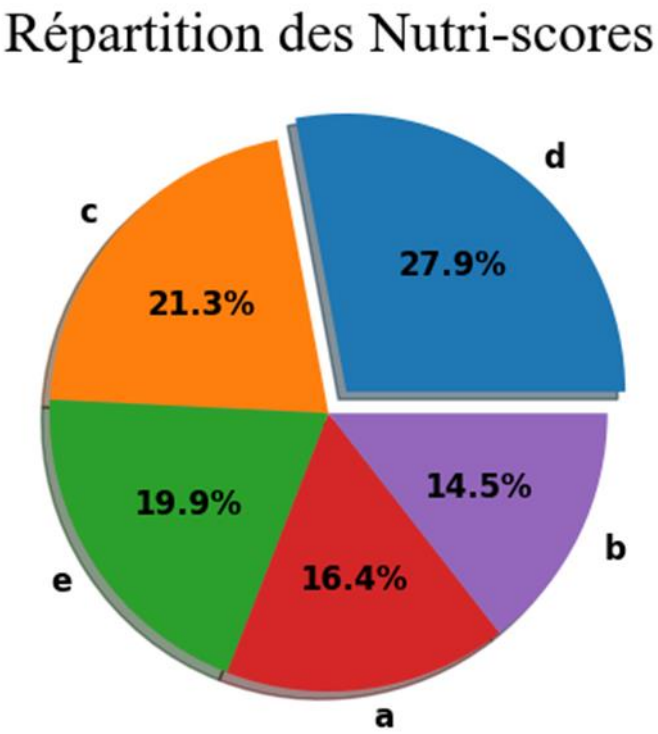
Nombre de création de nouvelle ligne de produit et de modification sur le jeu de données



=> Le jeu de données n'est plus mis à jour depuis 6 ans

Analyses univariées

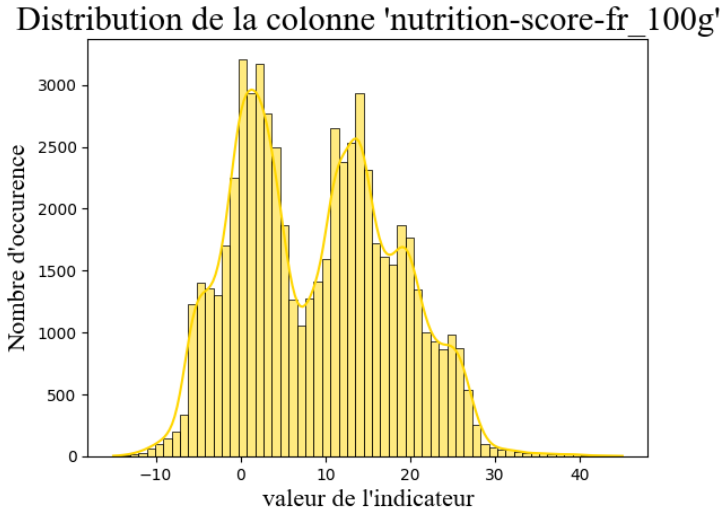
Nombre de produits associés à chaque Nutri-score



=> Les lettres C et D représentent la moitié des produits (49,2%)

Analyses univariées

Distribution du score nutritionnel



	moyenne	médiane	classe modale	ecart-type	CV	skewness	kurtosis
0	8.9	9.0	(-0.2, 0.7]: 3210 élément(s)	9.2	1.03	0.22	-0.8

- **Aspect :** distribution bimodale => 2 zones de fortes concentrations (autour de 0 et 15).

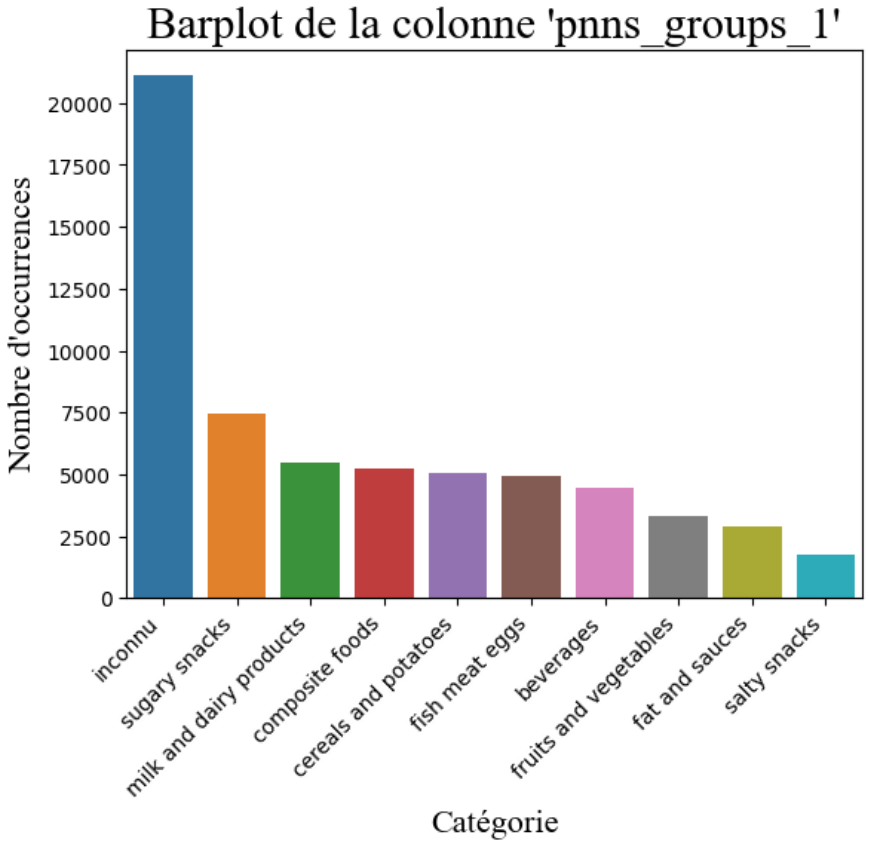
- **Tendance centrale :** La moyenne et la médiane sont proches autour du score 9.

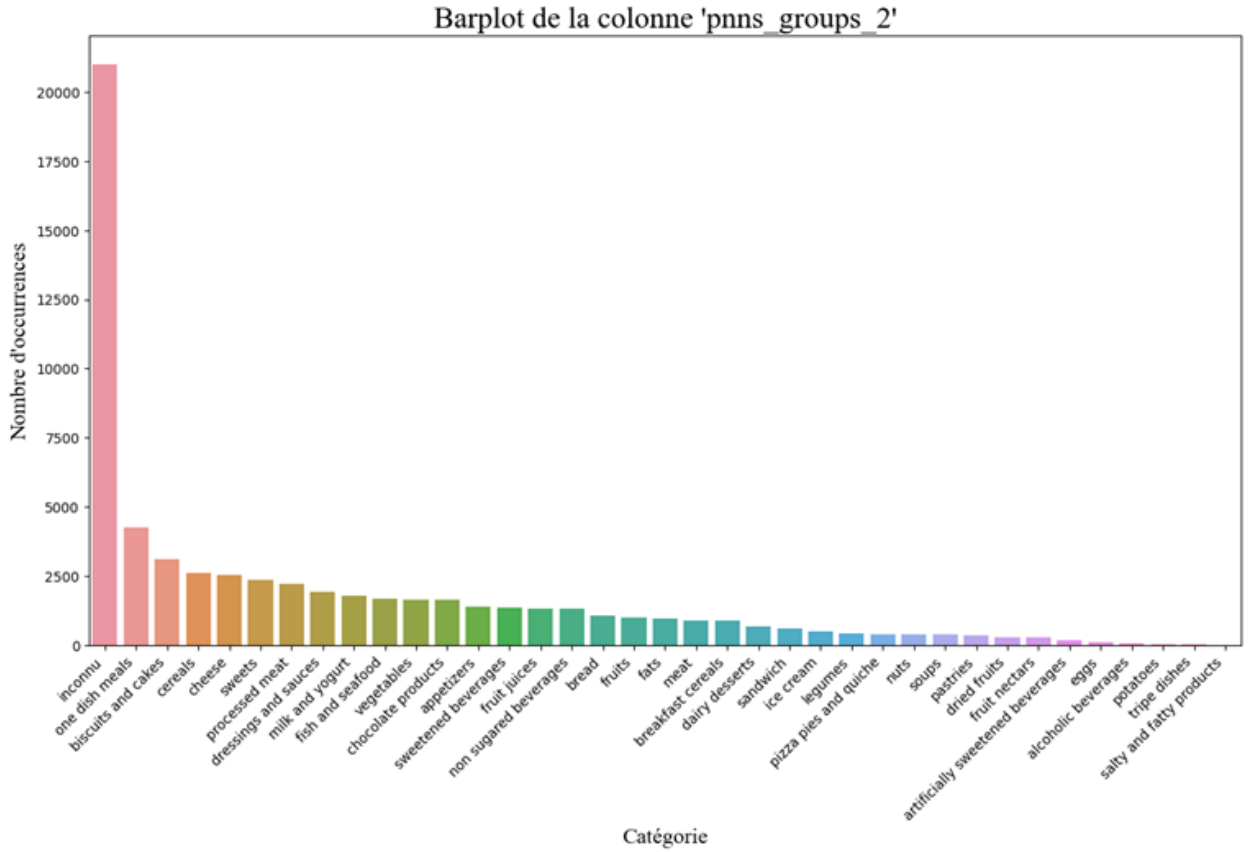
- **Dispersion :** les données forment un groupe très hétérogène (Coeff. variation > 100%).

- **Forme :** données étalées à droite (skewness>0) et plus aplaties que la loi normale (kurtosis<0).

Analyses univariées

Récurrence des catégories





=> env. 30% des produits sont de catégorie « inconnue »

Sommaire

Analyse multivariée

Analyse en Composantes Principales (ACP)

21 composantes expliquent 100% des données

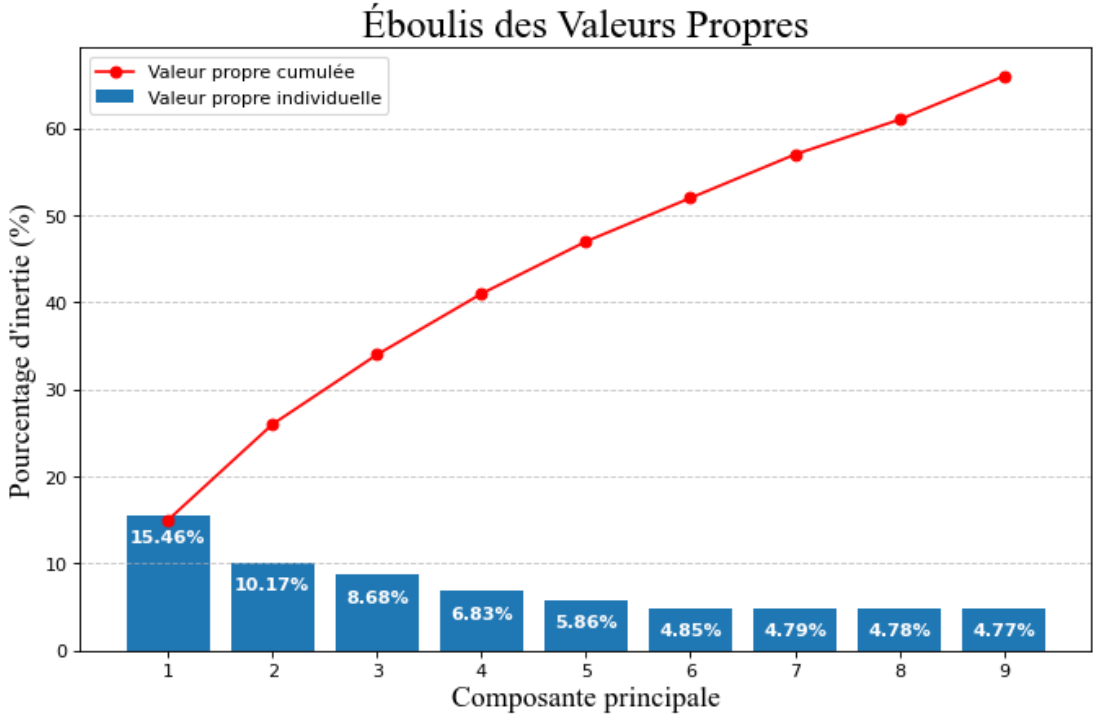
9 composantes > seuil de Kaiser de 4,76

Analyse multivariée

Analyse en Composantes Principales (ACP)

21 composantes expliquent 100% des données

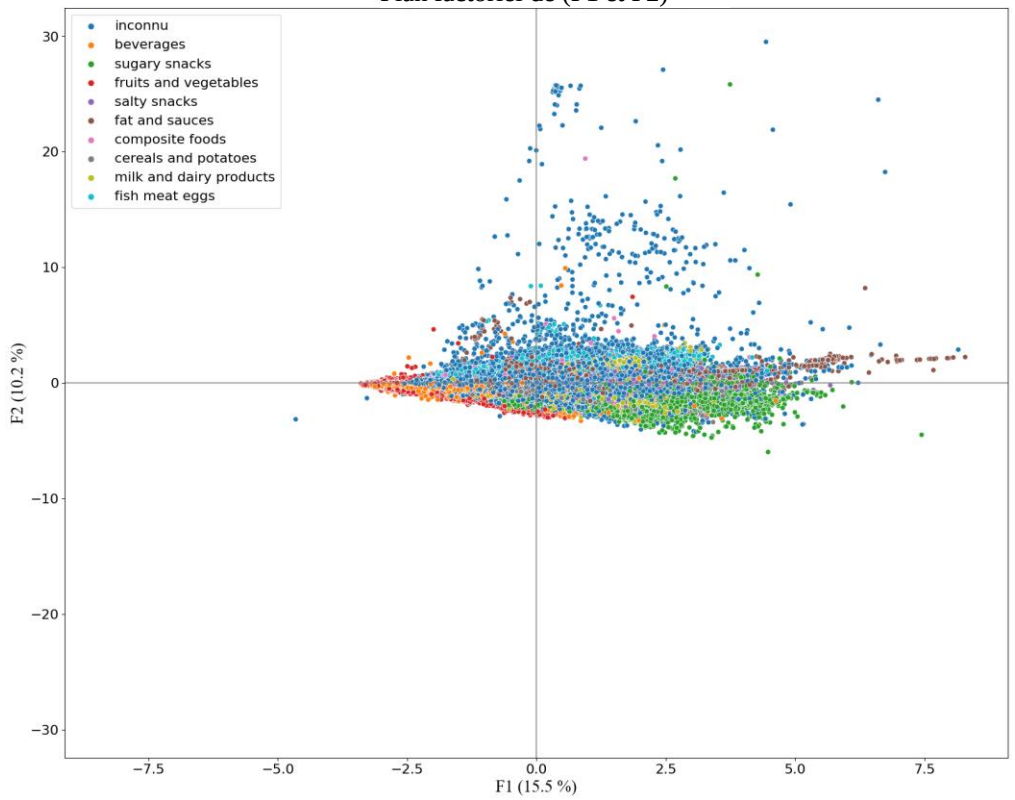
9 composantes > seuil de Kaiser de 4,76



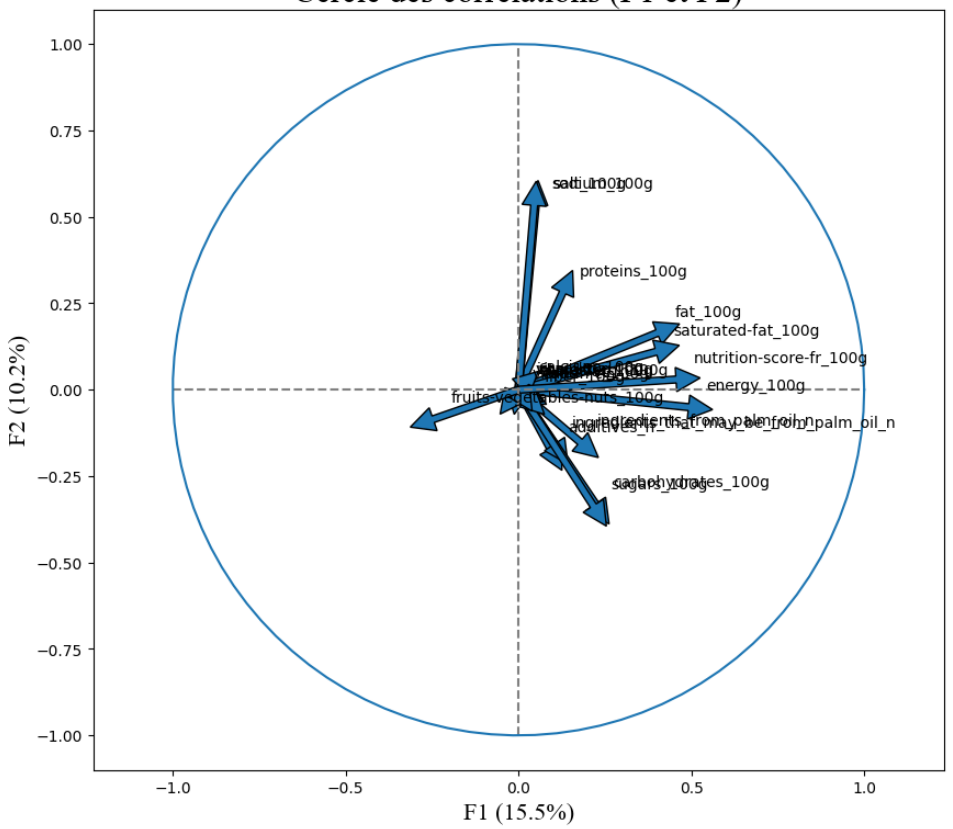
=> Les 5 premières composantes expliquent env. 50% de l'inertie totale

Analyse multivariée

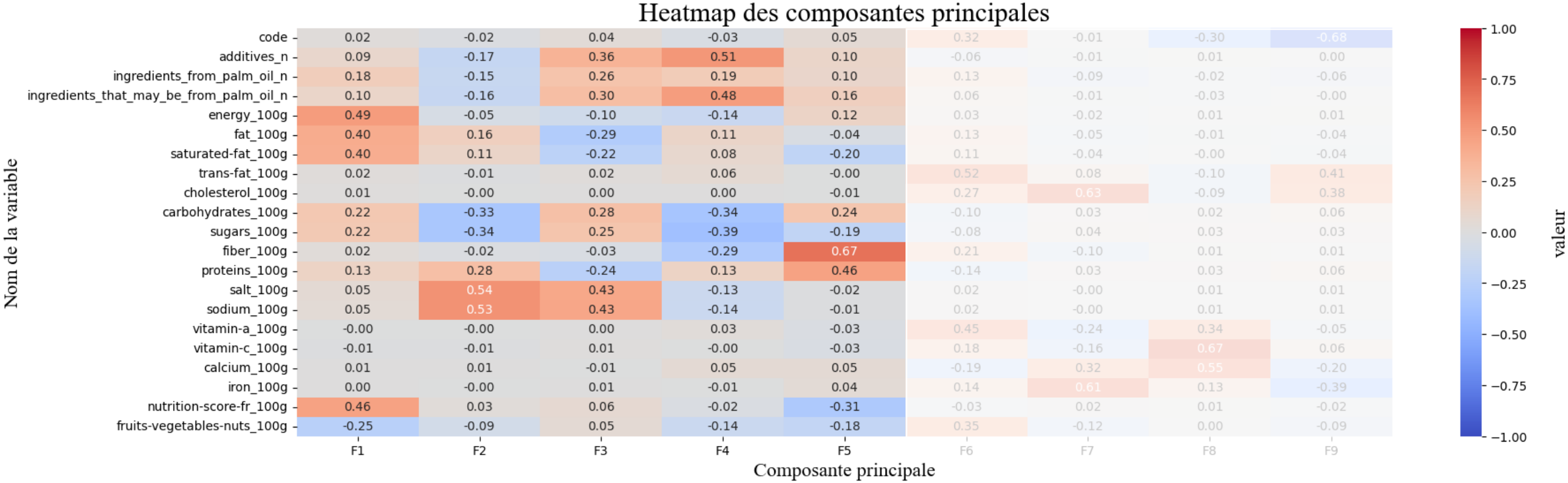
Plan factoriel de (F1 et F2)



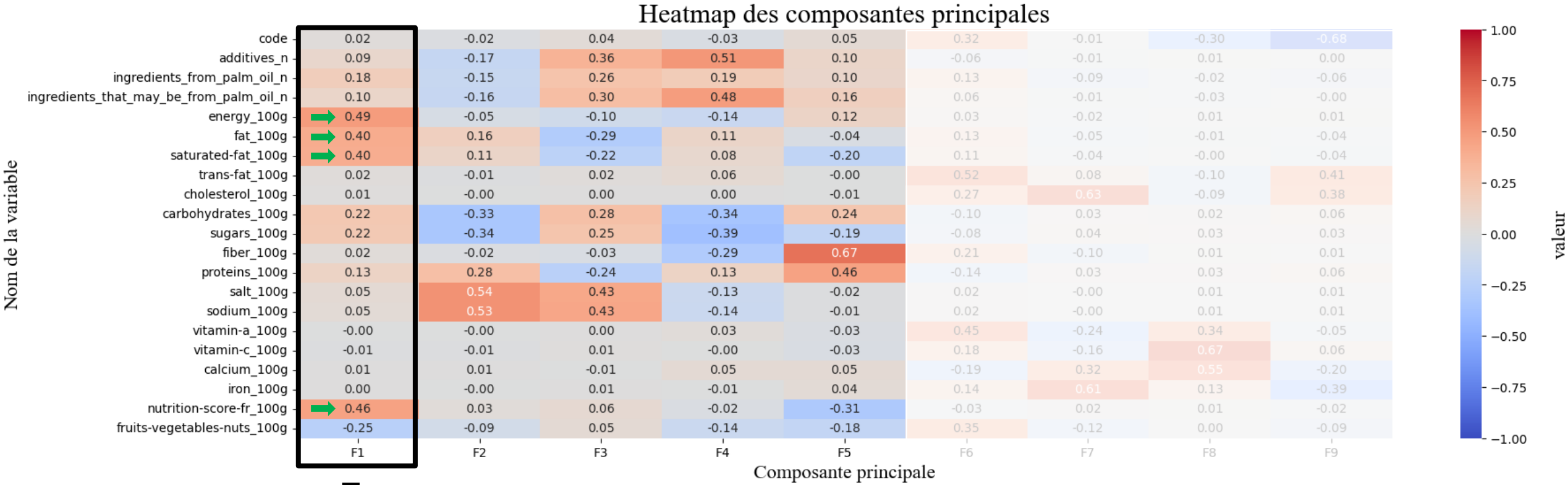
Cercle des corrélations (F1 et F2)



Analyse multivariée

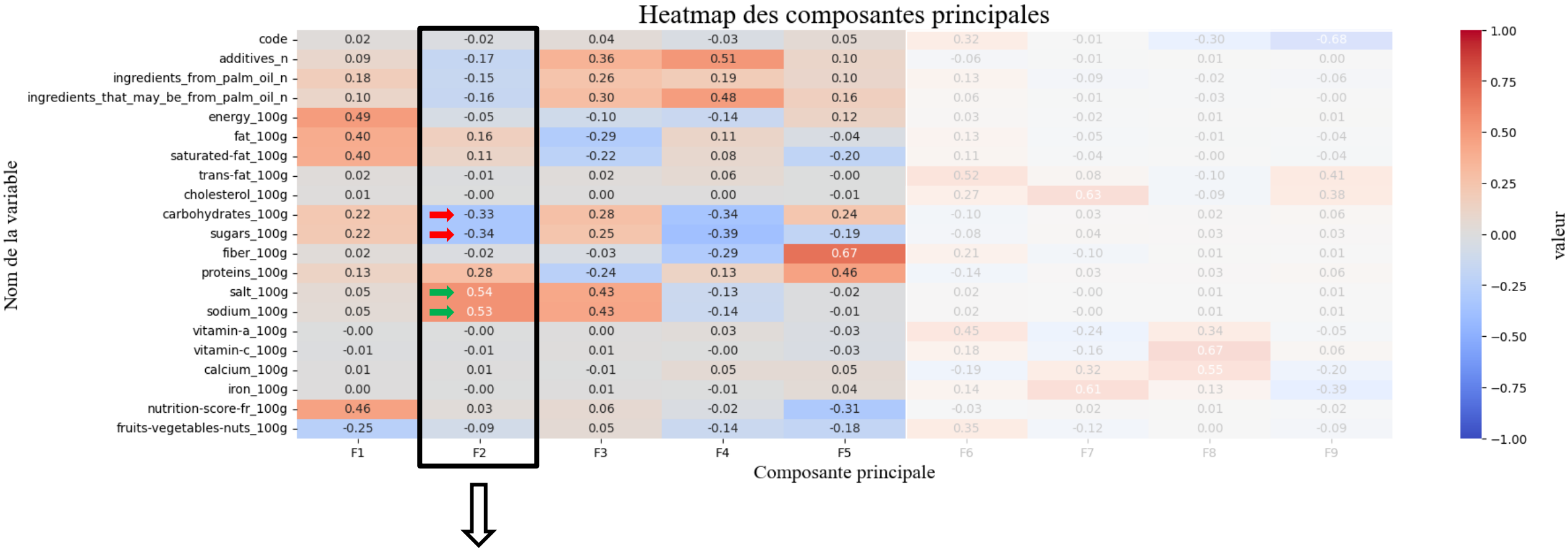


Analyse multivariée



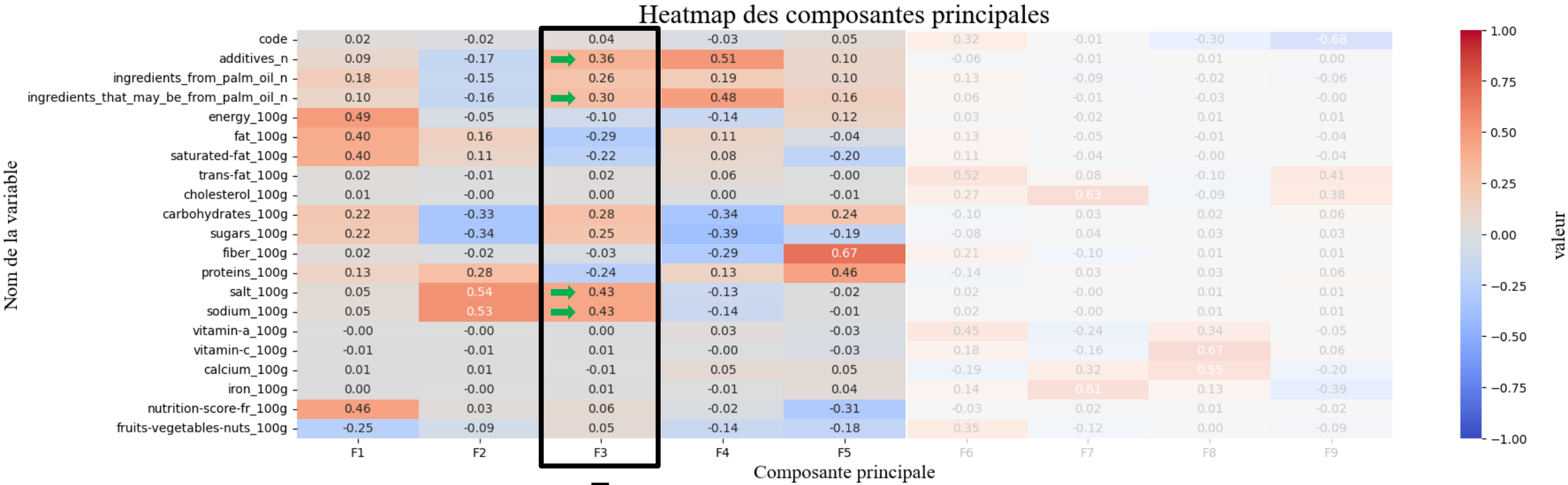
F1 semble
Corrélée à **l'apport calorique**

Analyse multivariée



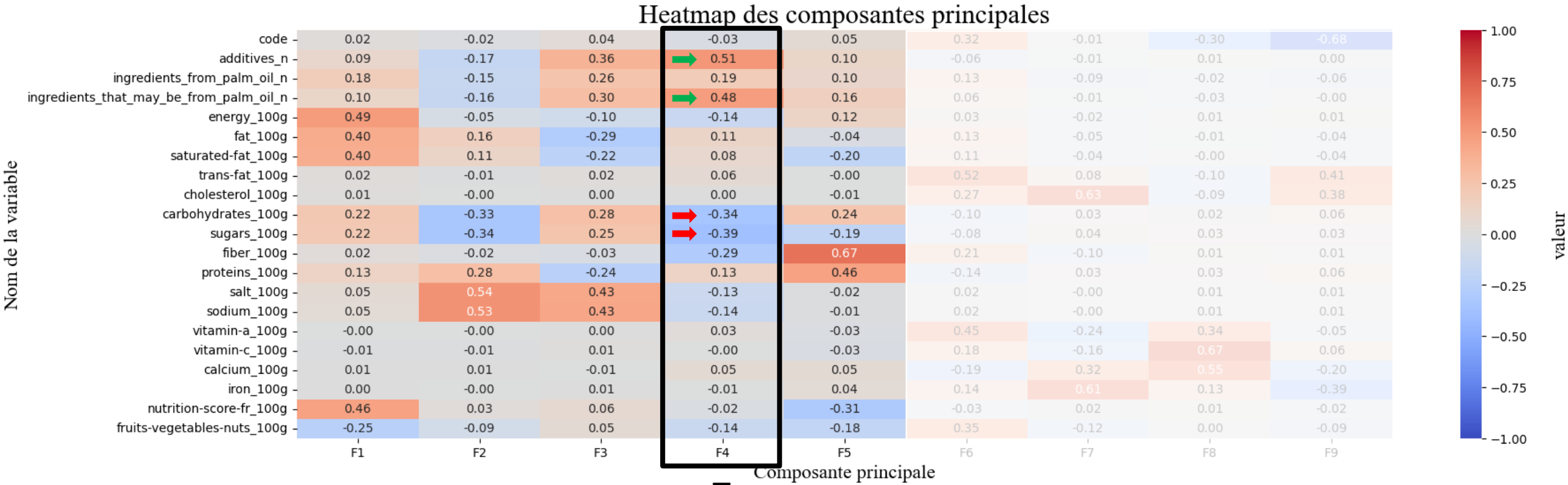
F2 semble
Corrélée à **la saveur salée**

Analyse multivariée



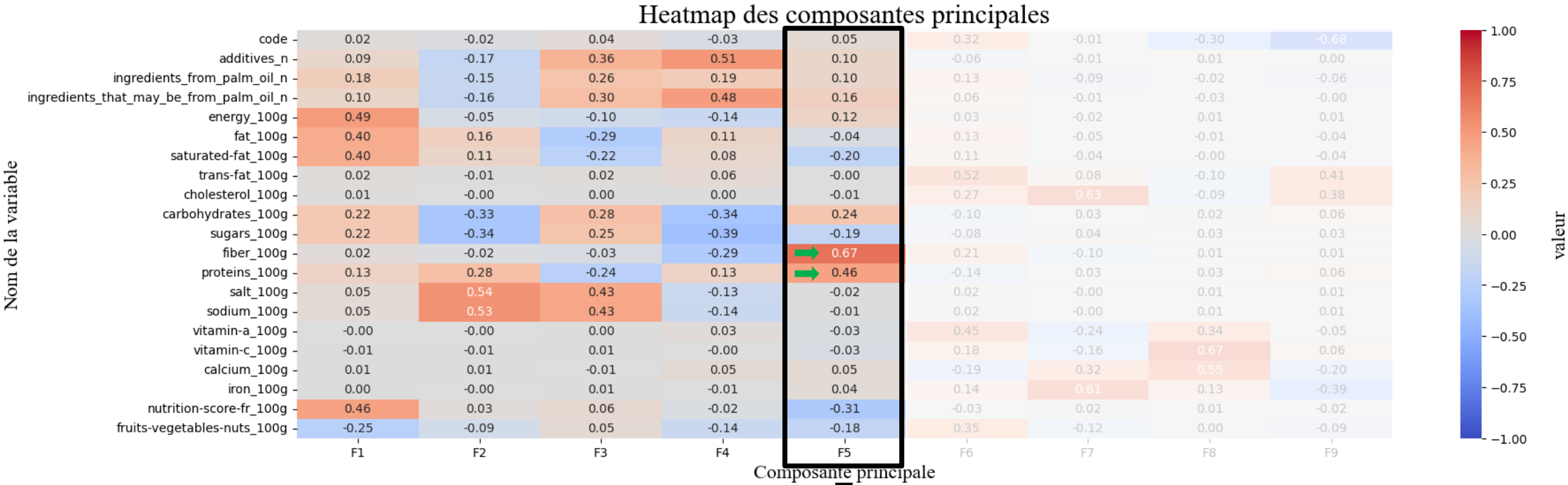
F3 semble
Corrélée à **l'apport d'éléments ajouté artificiellement**

Analyse multivariée



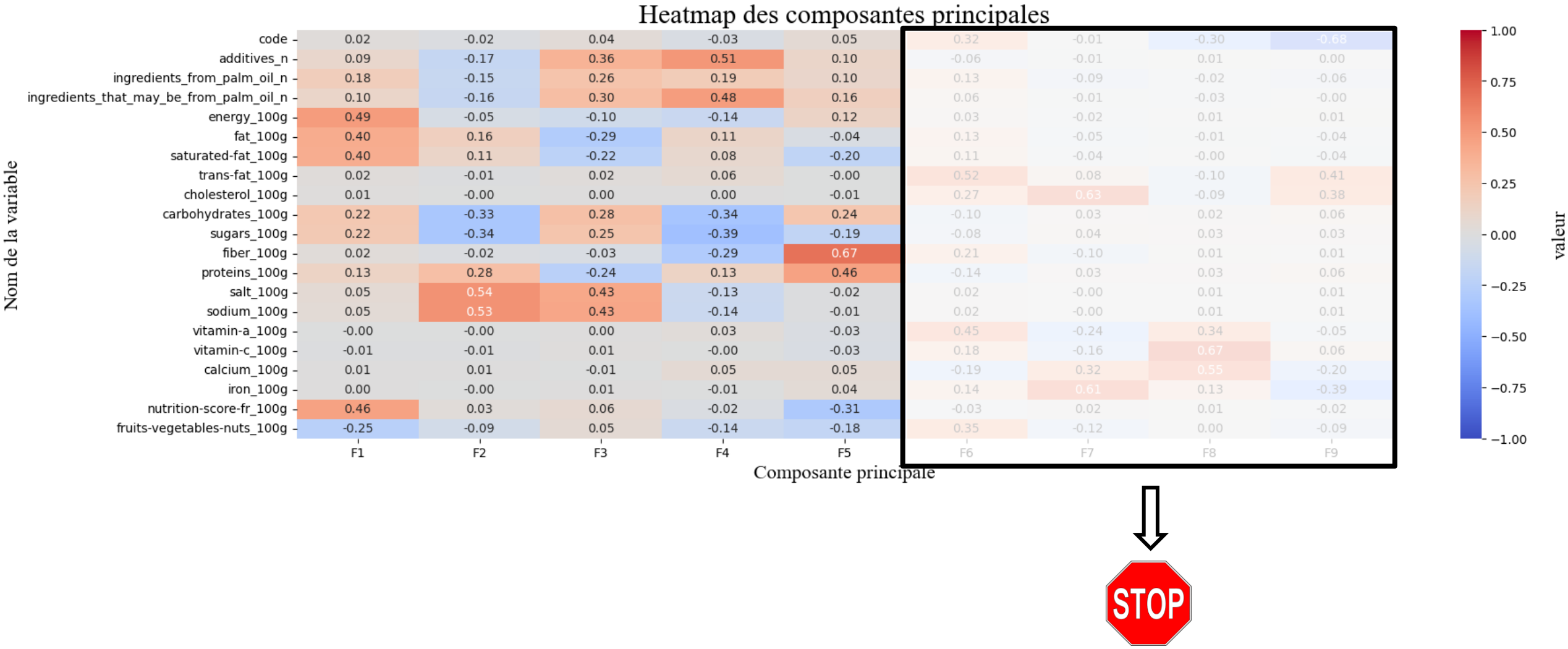
F4 semble
Corrélée à **l'apport d'édulcorants**

Analyse multivariée



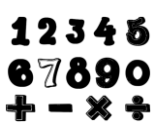
F5 semble
Corrélée à **la quantité des nutriments de la composante P (+) du score nutritionnel**

Analyse multivariée

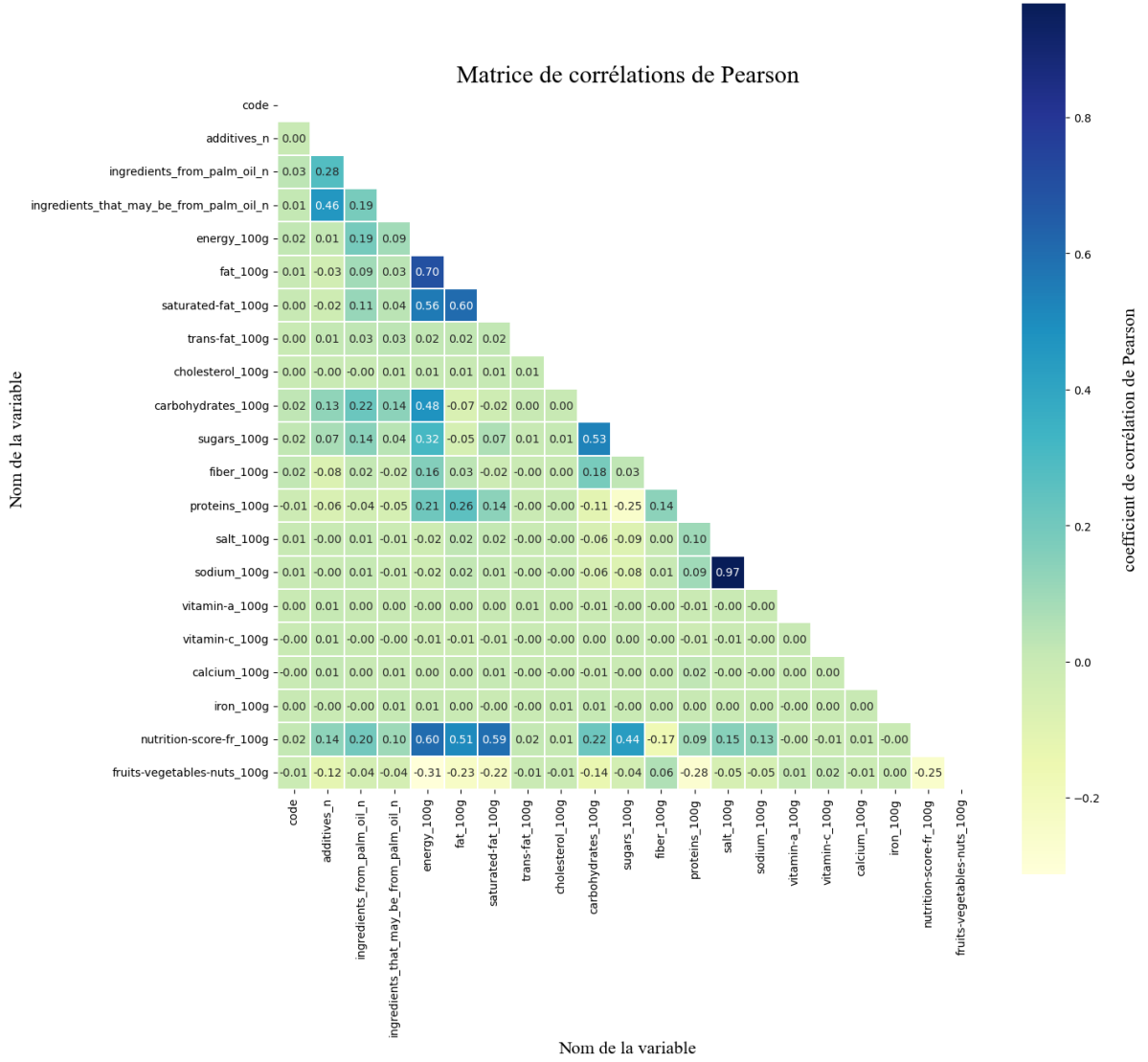


Sommaire

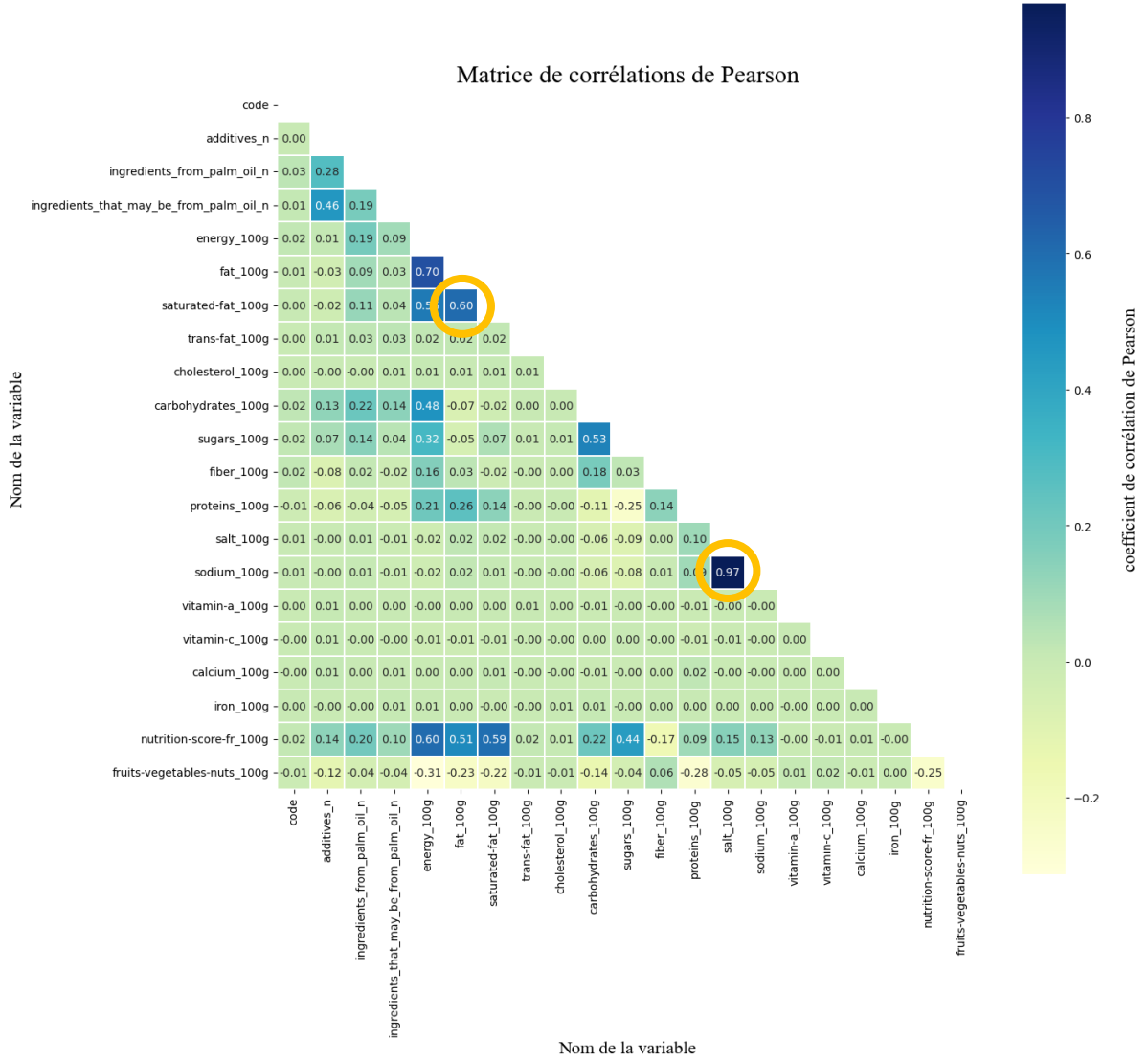
Observations évaluant la pertinence / faisabilité de l’idée application du client



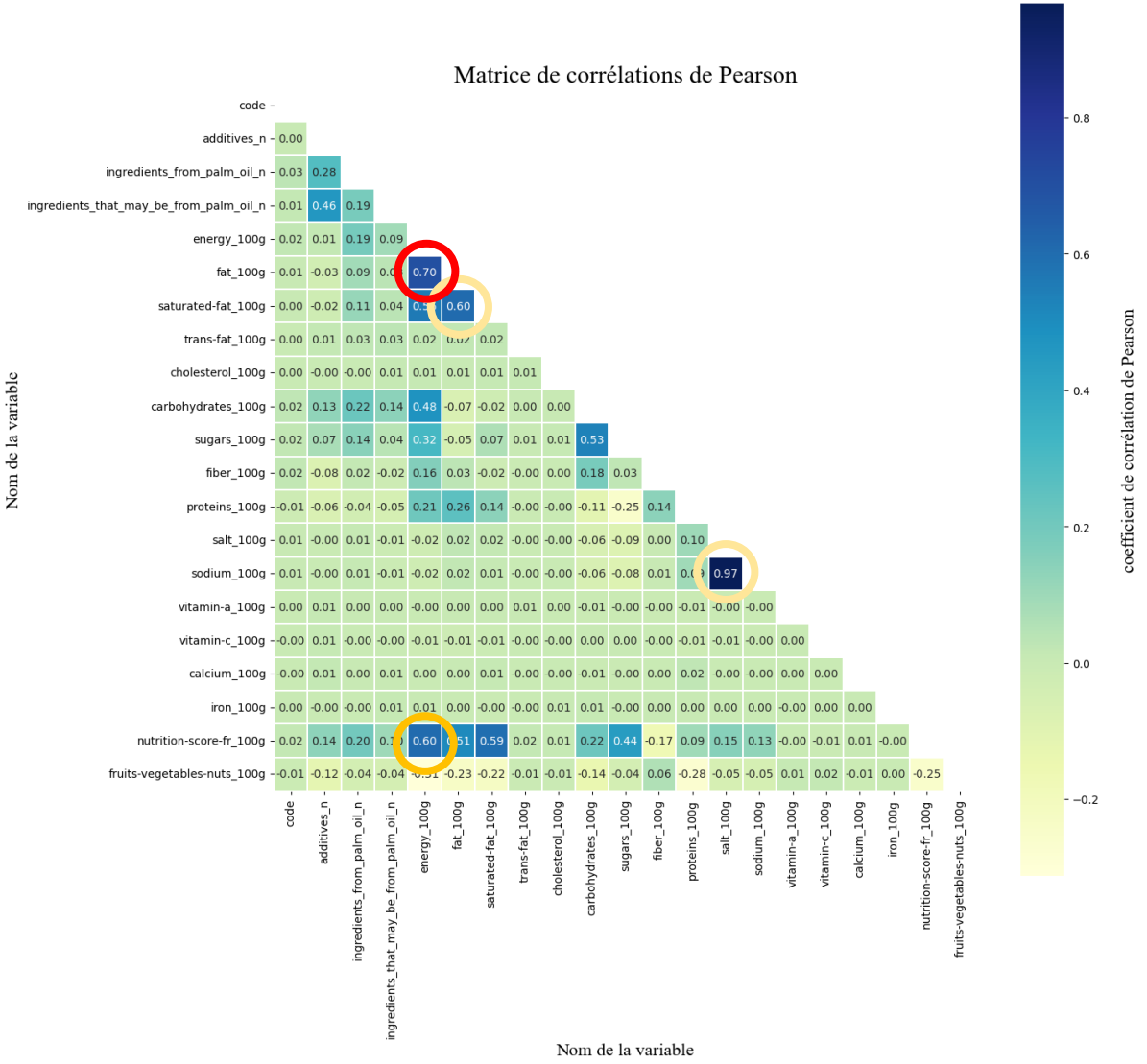
Observations évaluant la pertinence / faisabilité de l'idée application du client



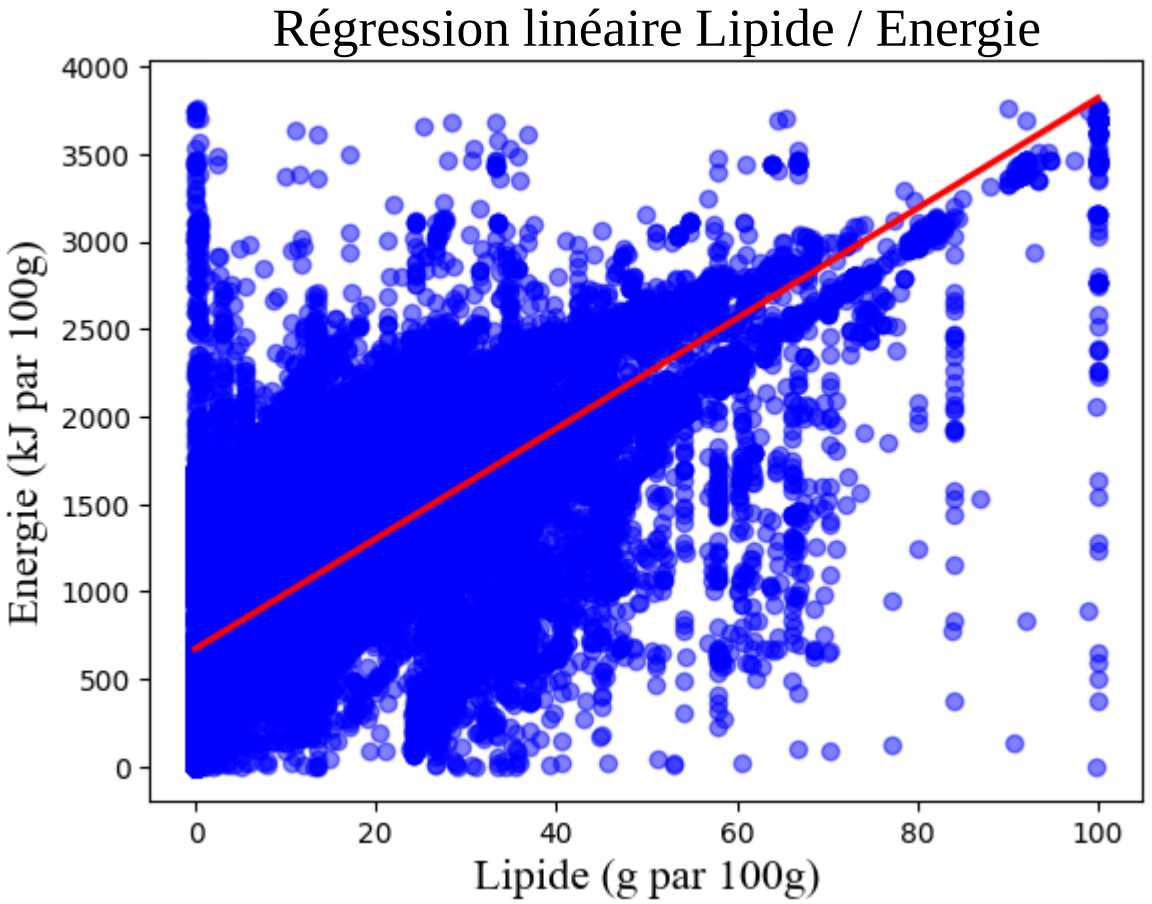
Observations évaluant la pertinence / faisabilité de l'idée application du client



Observations évaluant la pertinence / faisabilité de l'idée application du client

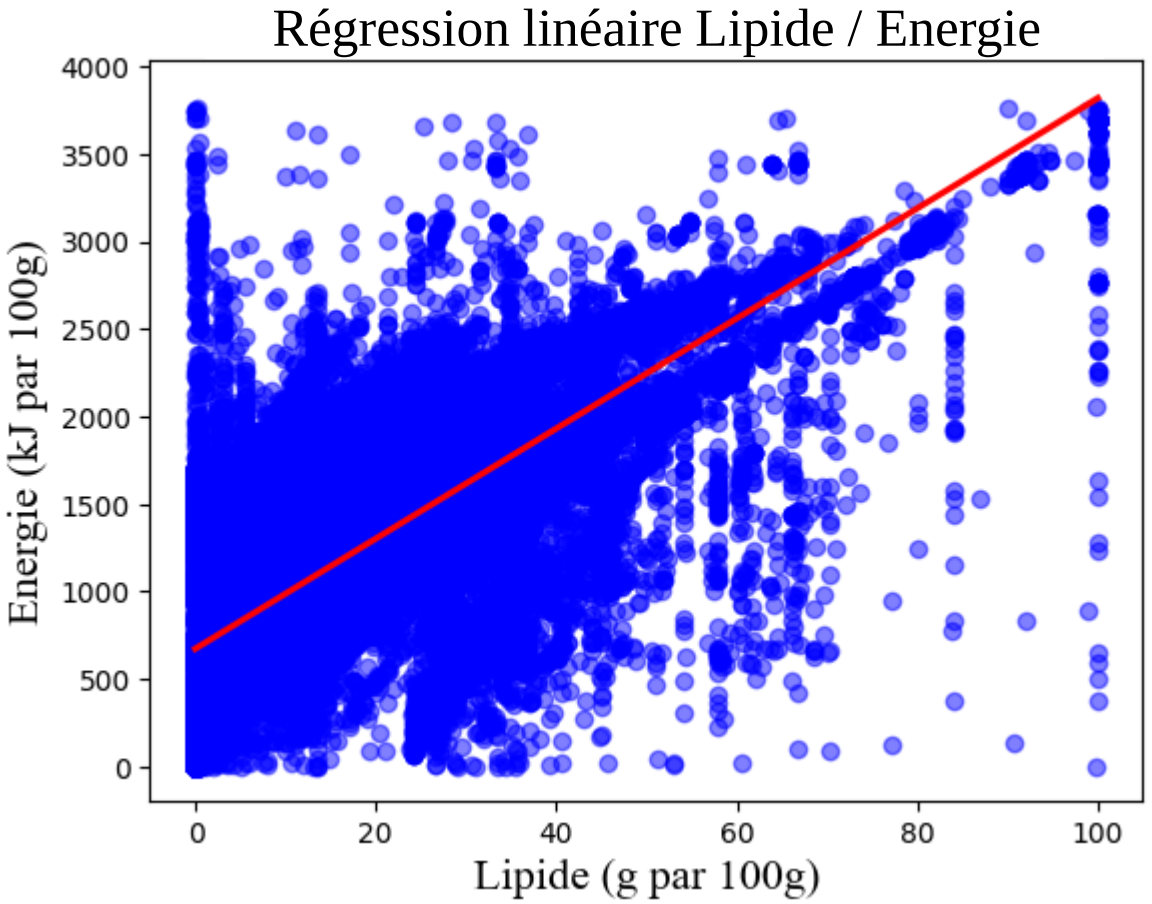


Observations évaluant la pertinence / faisabilité de l'idée application du client



Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.
Hypothèse alternative (HA) : Corrélation entre les 2 variables.

Observations évaluant la pertinence / faisabilité de l'idée application du client

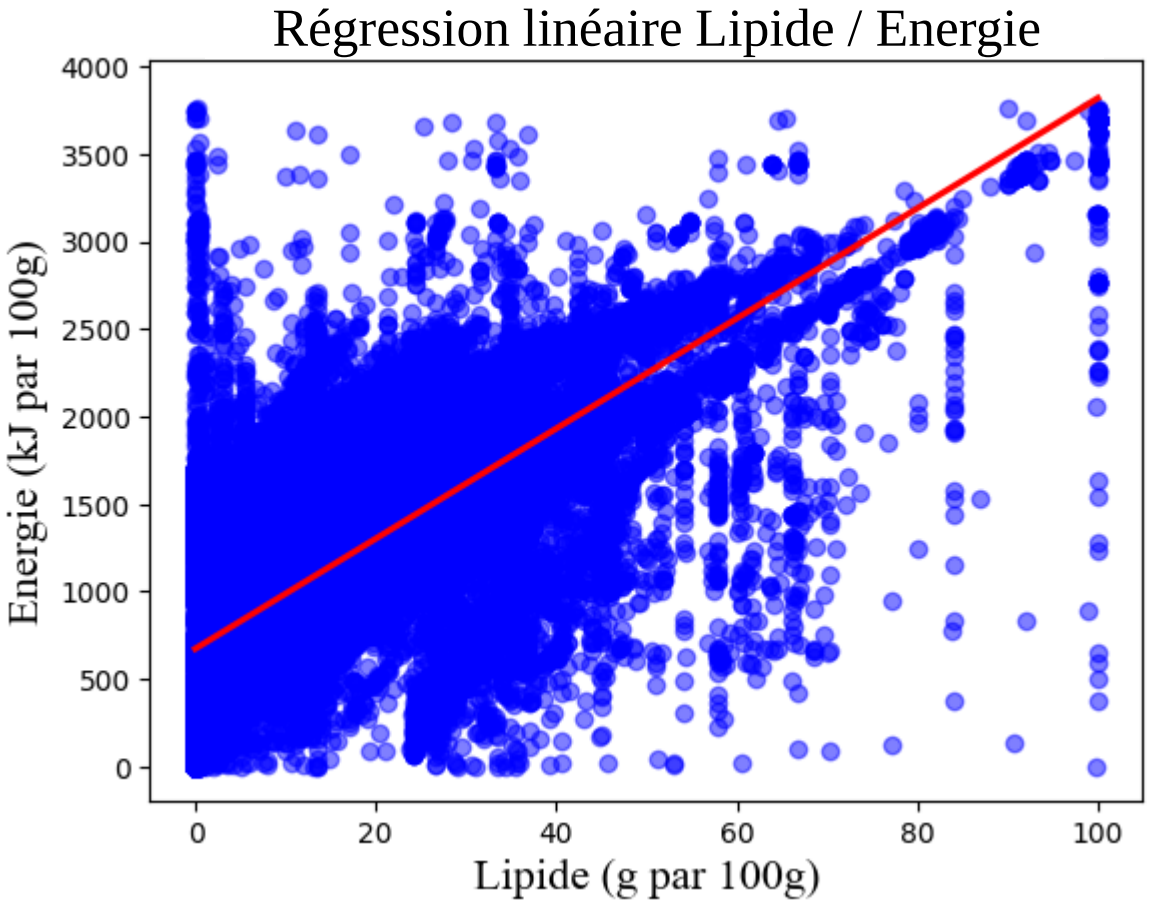


~~Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.~~

Hypothèse alternative (HA) : Corrélation entre les 2 variables.

p_value = 0.00

Observations évaluant la pertinence / faisabilité de l'idée application du client



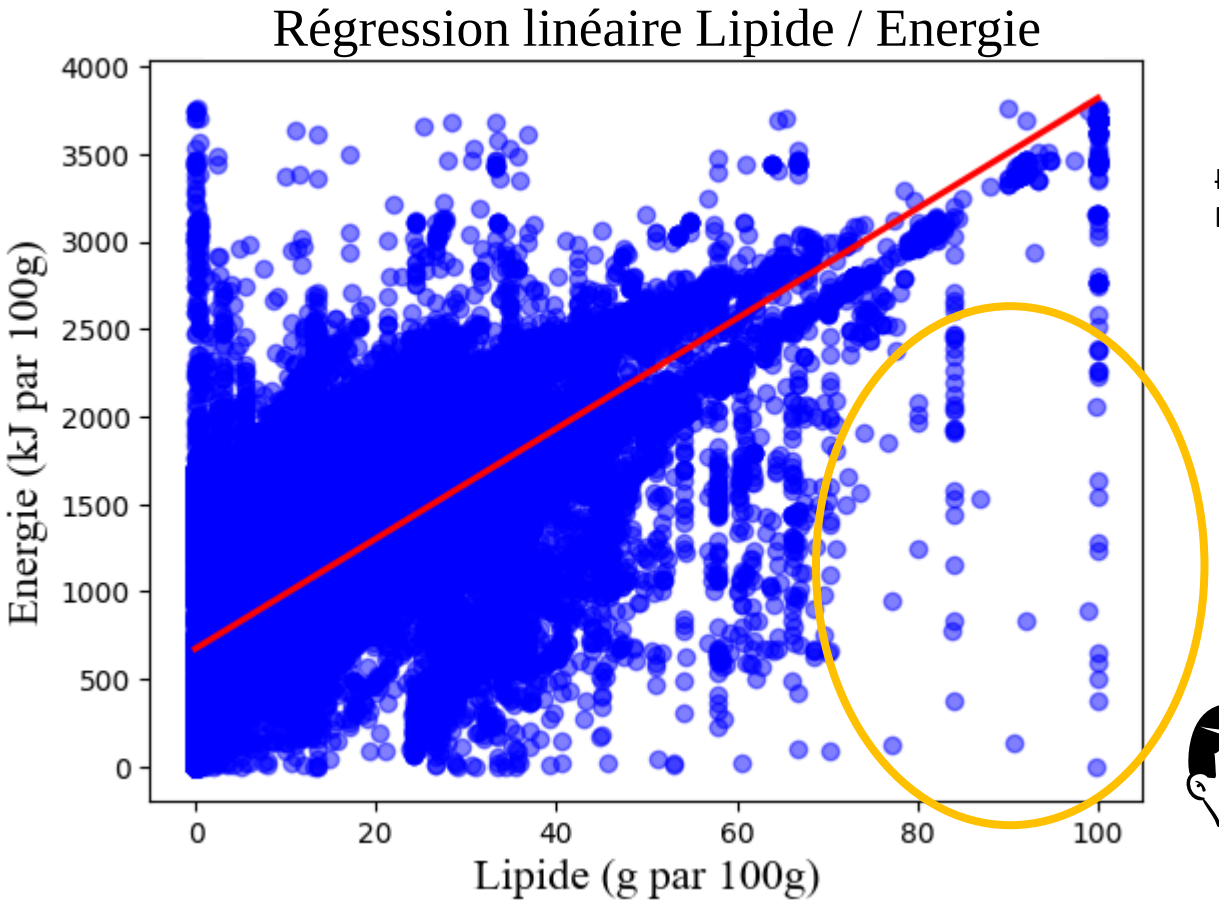
~~Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.~~
Hypothèse alternative (HA) : Corrélation entre les 2 variables.

p_value = 0.00

Energie = 31.5 * lipide + 672.9
Corrélation Pearson (r) = 0.70

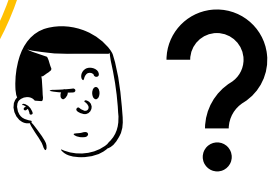
=> Il existe une corrélation positive entre la quantité de lipide et d'énergie (*p-value < 0,05 ; r=0,70*)

Observations évaluant la pertinence / faisabilité de l'idée application du client



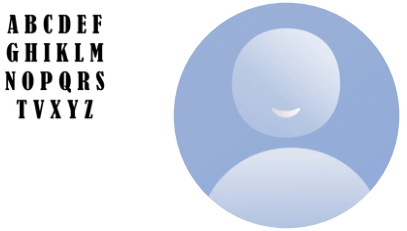
~~Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.~~
Hypothèse alternative (HA) : Corrélation entre les 2 variables.

Energie = 31.5 * lipide + 672.9
Corrélation Pearson (r) = 0.70
p_value = 0.00



=> Il existe une corrélation positive entre la quantité de lipide et d'énergie (*p-value < 0,05 ; r=0,70*)

Observations évaluant la pertinence / faisabilité de l’idée application du client



Observations évaluant la pertinence / faisabilité de l'idée application du client

χ^2

Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.
Hypothèse alternative (HA) : Corrélation entre les 2 variables.

Observations évaluant la pertinence / faisabilité de l'idée application du client

ABCDEF
GHIKLM
NOPQRS
TVXYZ

χ^2

~~Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.~~
Hypothèse alternative (HA) : Corrélation entre les 2 variables.

Nutri-score / catégorie pnns 2

```
Test du chi-deux :
```

```
Chi² = 40894.46
```

```
p-value = 0.00
```

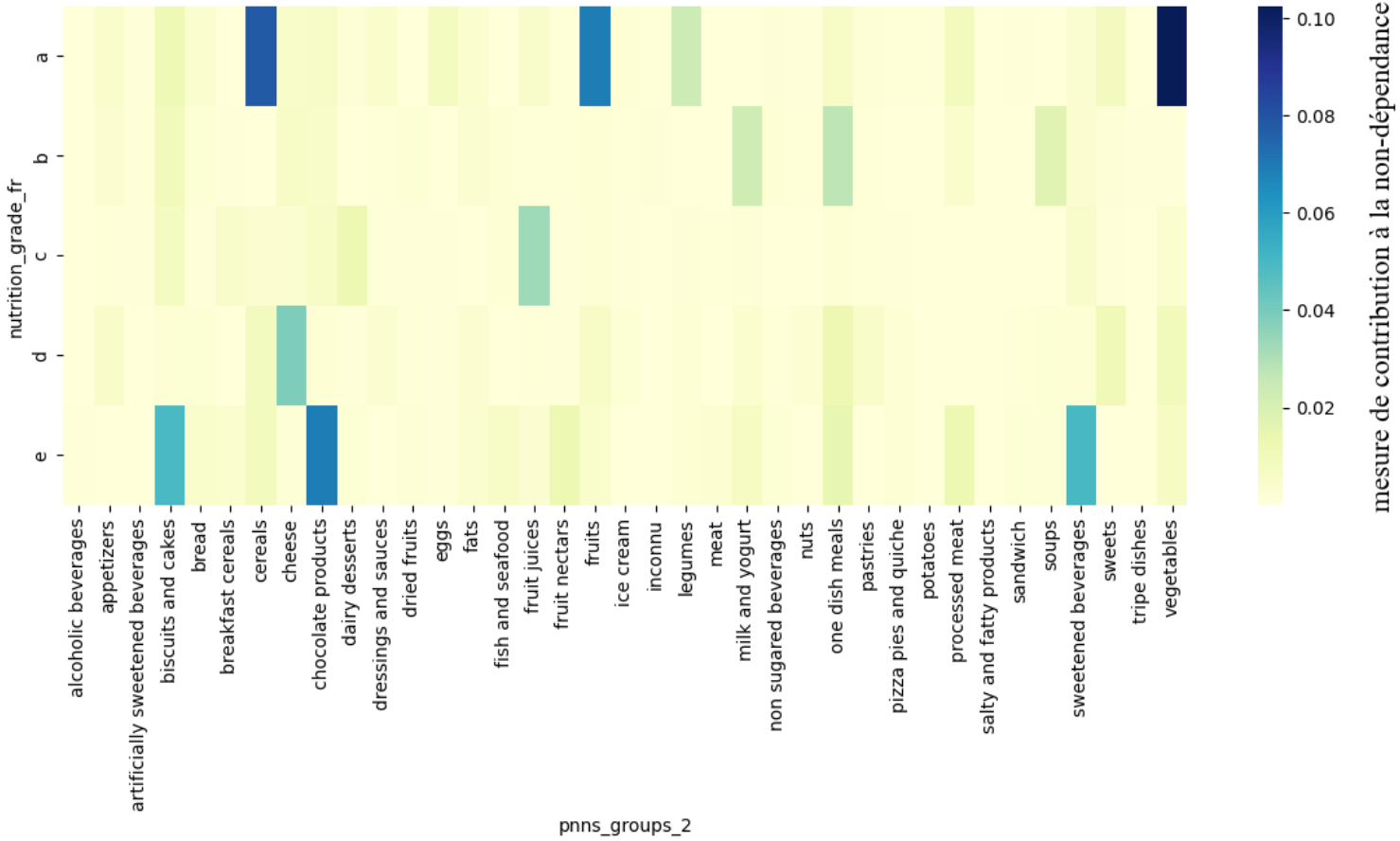
=> Il existe une corrélation entre le Nutri-score et la catégorie pnns 2 (*p-value < 0,05*)

Observations évaluant la pertinence / faisabilité de l'idée application du client

χ^2

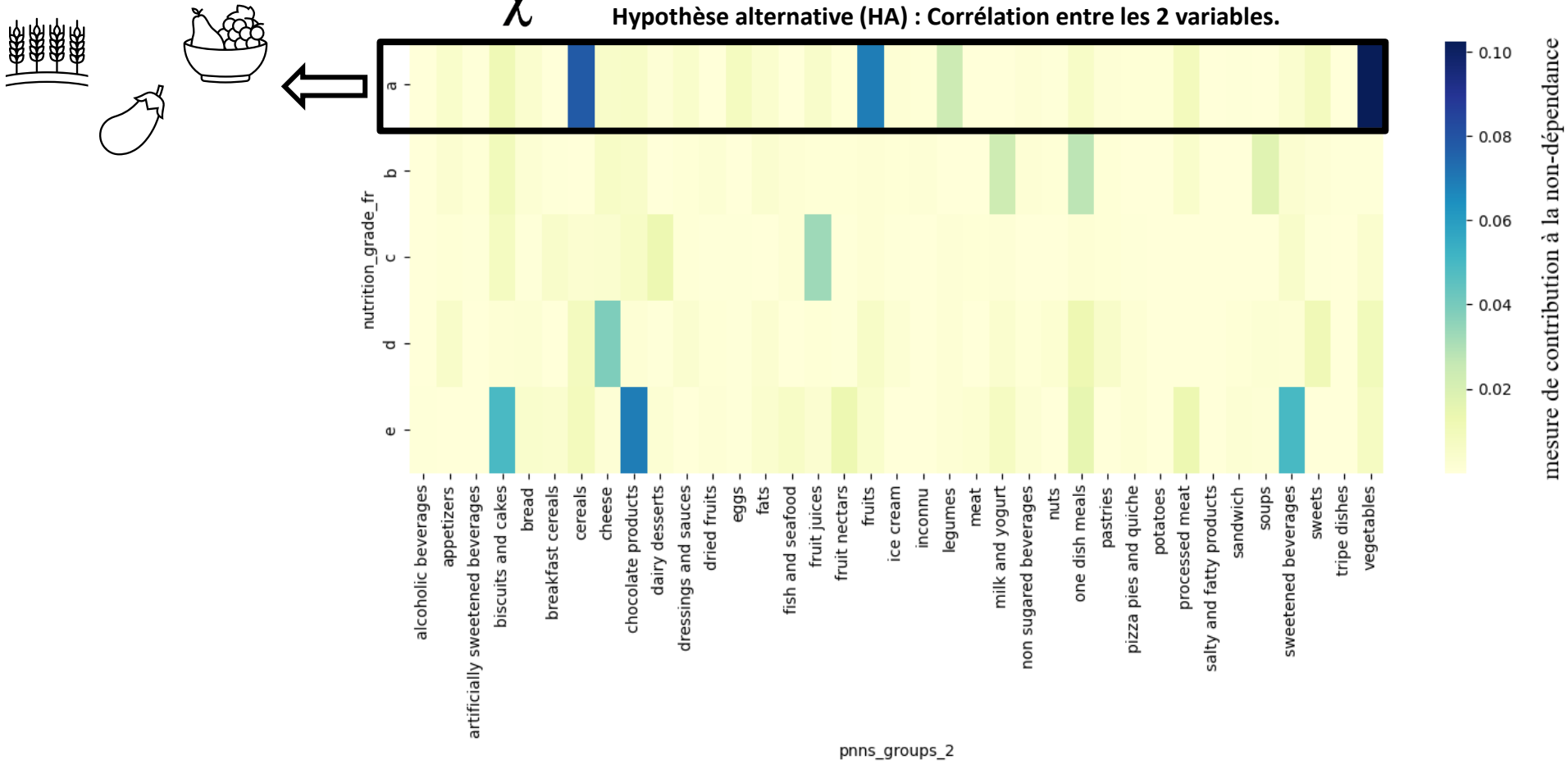
~~Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.~~

Hypothèse alternative (HA) : Corrélation entre les 2 variables.



=> Il existe une corrélation entre le Nutri-score et la catégorie pnns 2 (*p-value < 0,05*)

Observations évaluant la pertinence / faisabilité de l'idée application du client



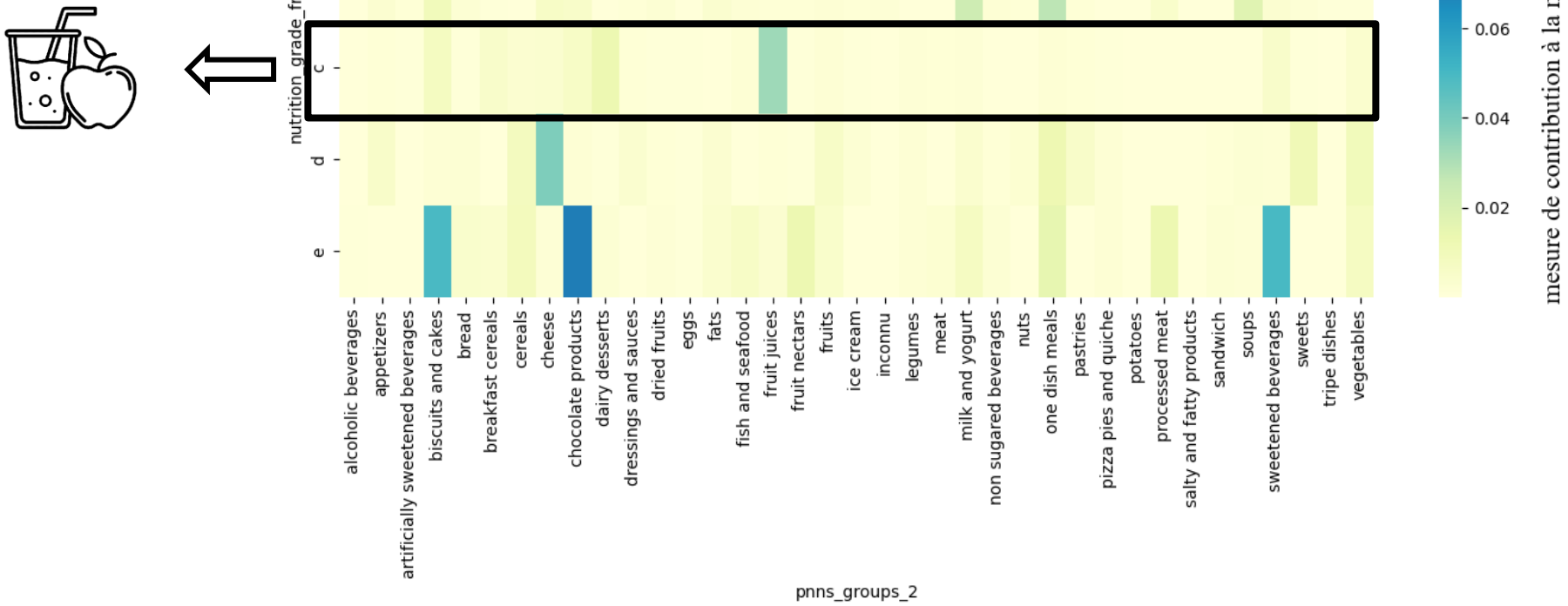
=> Il existe une corrélation entre le Nutri-score et la catégorie pnns 2 (*p-value < 0,05*)

Observations évaluant la pertinence / faisabilité de l'idée application du client

χ^2

~~Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.~~

Hypothèse alternative (HA) : Corrélation entre les 2 variables.



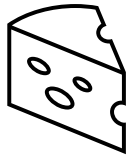
=> Il existe une corrélation entre le Nutri-score et la catégorie pnns 2 (*p-value < 0,05*)

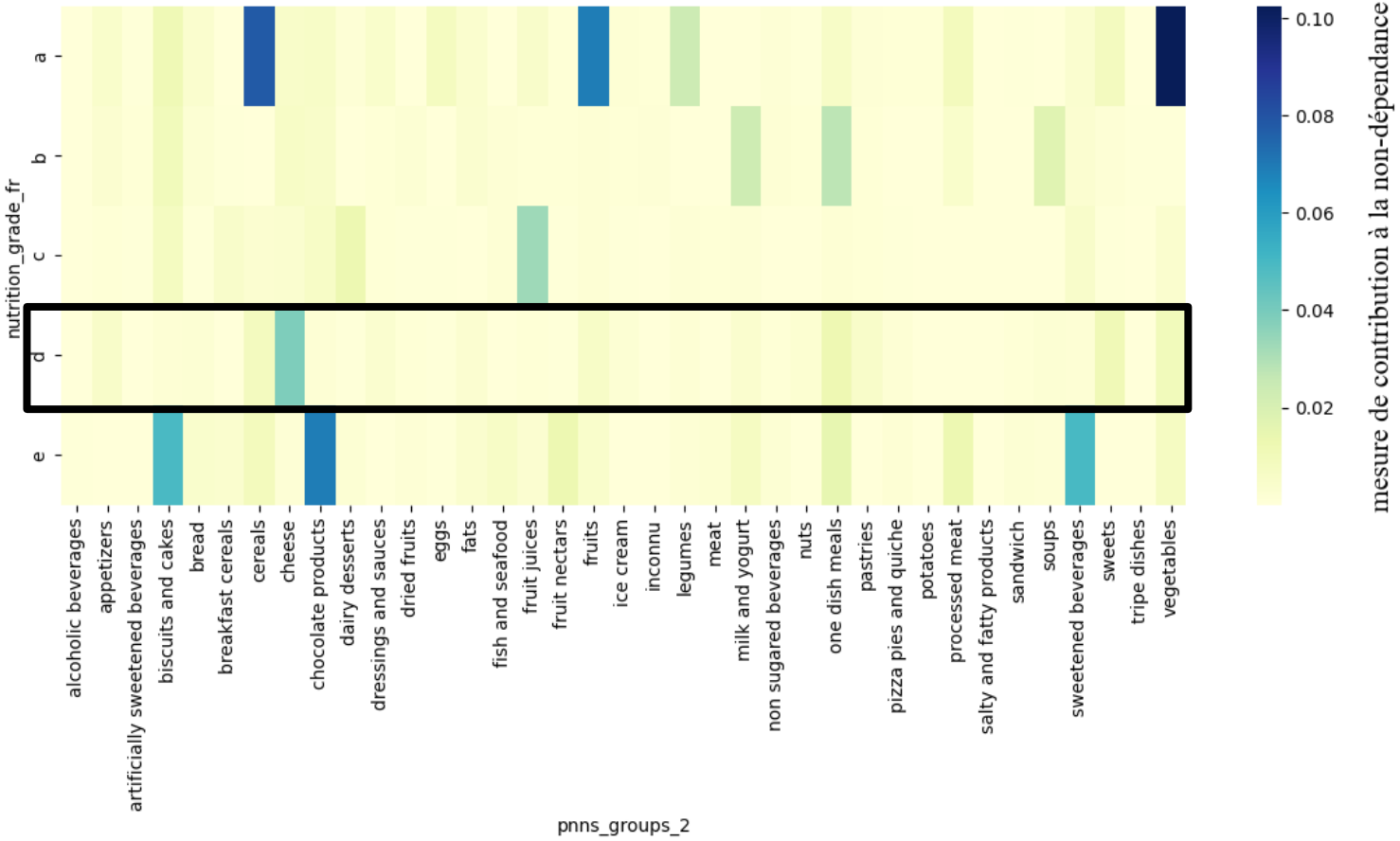
Observations évaluant la pertinence / faisabilité de l'idée application du client

χ^2

~~Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.~~

Hypothèse alternative (HA) : Corrélation entre les 2 variables.





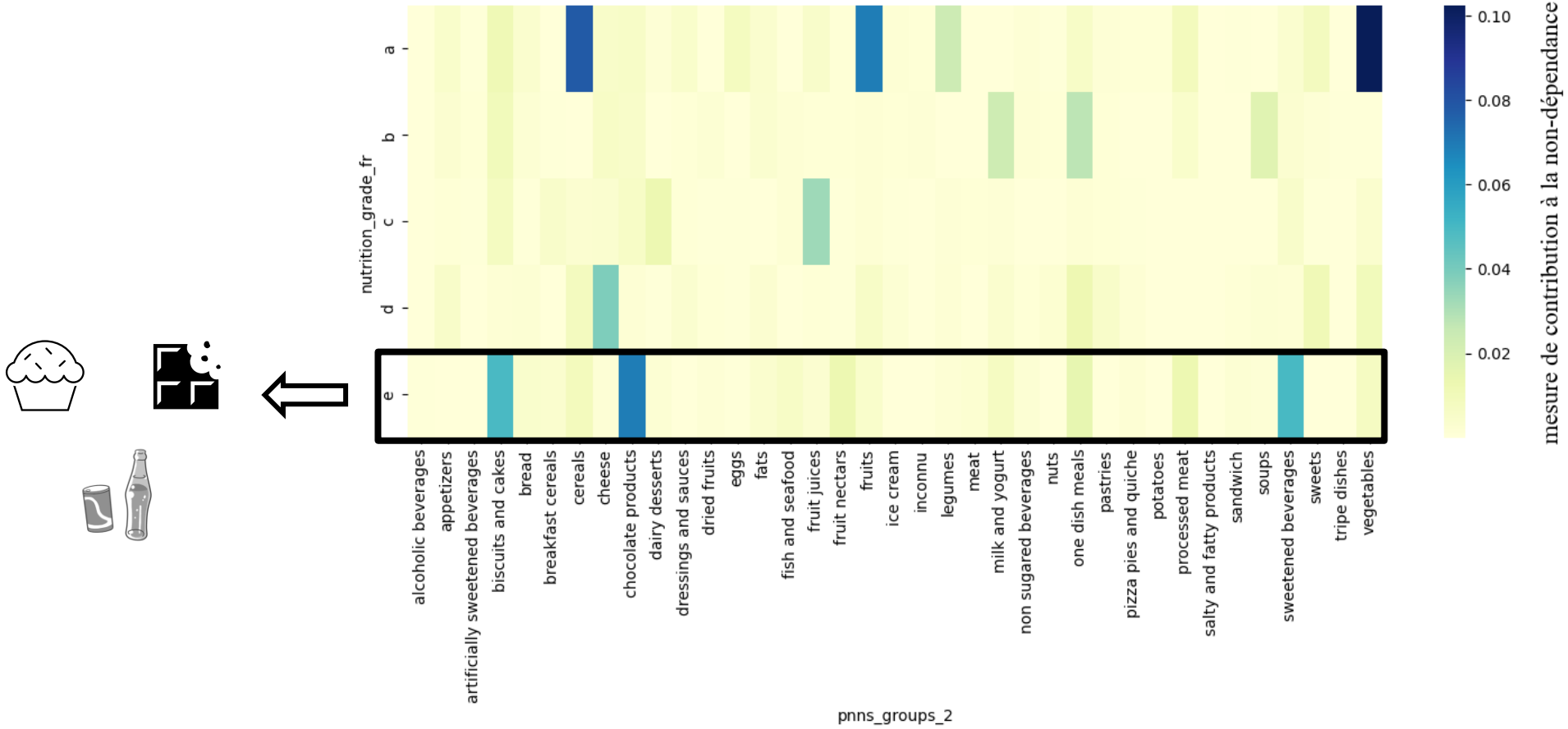
=> Il existe une corrélation entre le Nutri-score et la catégorie pnns 2 (*p-value < 0,05*)

Observations évaluant la pertinence / faisabilité de l'idée application du client

χ^2

~~Hypothèse nulle (H0) : Pas de corrélation entre les 2 variables.~~

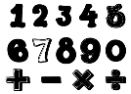
Hypothèse alternative (HA) : Corrélation entre les 2 variables.



=> Il existe une corrélation entre le Nutri-score et la catégorie pnns 2 (*p-value < 0,05*)

Observations évaluant la pertinence / faisabilité de l'idée application du client

Observations évaluant la pertinence / faisabilité de l'idée application du client



Test de normalité des données => **Test de Shapiro-Wilk**

⇩

Hypothèse nulle (H0) : la population est distribuée normalement selon la loi gaussienne.
Hypothèse alternative (HA) : la population n'est pas distribuée normalement selon la loi gaussienne.

Observations évaluant la pertinence / faisabilité de l'idée application du client

Test de normalité des données => **Test de Shapiro-Wilk**

~~Hypothèse nulle (H0) : la population est distribuée normalement selon la loi gaussienne.~~

Hypothèse alternative (HA) : la population n'est pas distribuée normalement selon la loi gaussienne.

	Column	SW-statistic	P-value
0	code	0.12436	0.0
1	additives_n	0.64575	0.0
2	ingredients_from_palm_oil_n	0.22780	0.0
3	ingredients_that_may_be_from_palm_oil_n	0.30616	0.0
4	energy_100g	0.94936	0.0
5	fat_100g	0.77444	0.0
6	saturated-fat_100g	0.64348	0.0
7	trans-fat_100g	0.00292	0.0
8	cholesterol_100g	0.00055	0.0
9	carbohydrates_100g	0.86016	0.0
10	sugars_100g	0.71868	0.0
11	fiber_100g	0.44373	0.0
12	proteins_100g	0.81852	0.0
13	salt_100g	0.18533	0.0
14	sodium_100g	0.17332	0.0
15	vitamin-a_100g	0.00113	0.0
16	vitamin-c_100g	0.00447	0.0
17	calcium_100g	0.00815	0.0
18	iron_100g	0.00087	0.0
19	nutrition-score-fr_100g	0.97473	0.0
20	fruits-vegetables-nuts_100g	0.73058	0.0

Observations évaluant la pertinence / faisabilité de l'idée application du client

Test de normalité des données => **Test de Shapiro-Wilk**

~~Hypothèse nulle (H0) : la population est distribuée normalement selon la loi gaussienne.~~

Hypothèse alternative (HA) : la population n'est pas distribuée normalement selon la loi gaussienne.

Test non-paramétrique de Kruskall-Wallis

kw

	Column	SW-statistic	P-value
0	code	0.12436	0.0
1	additives_n	0.64575	0.0
2	ingredients_from_palm_oil_n	0.22780	0.0
3	ingredients_that_may_be_from_palm_oil_n	0.30616	0.0
4	energy_100g	0.94936	0.0
5	fat_100g	0.77444	0.0
6	saturated-fat_100g	0.64348	0.0
7	trans-fat_100g	0.00292	0.0
8	cholesterol_100g	0.00055	0.0
9	carbohydrates_100g	0.86016	0.0
10	sugars_100g	0.71868	0.0
11	fiber_100g	0.44373	0.0
12	proteins_100g	0.81852	0.0
13	salt_100g	0.18533	0.0
14	sodium_100g	0.17332	0.0
15	vitamin-a_100g	0.00113	0.0
16	vitamin-c_100g	0.00447	0.0
17	calcium_100g	0.00815	0.0
18	iron_100g	0.00087	0.0
19	nutrition-score-fr_100g	0.97473	0.0
20	fruits-vegetables-nuts_100g	0.73058	0.0

Observations évaluant la pertinence / faisabilité de l'idée application du client

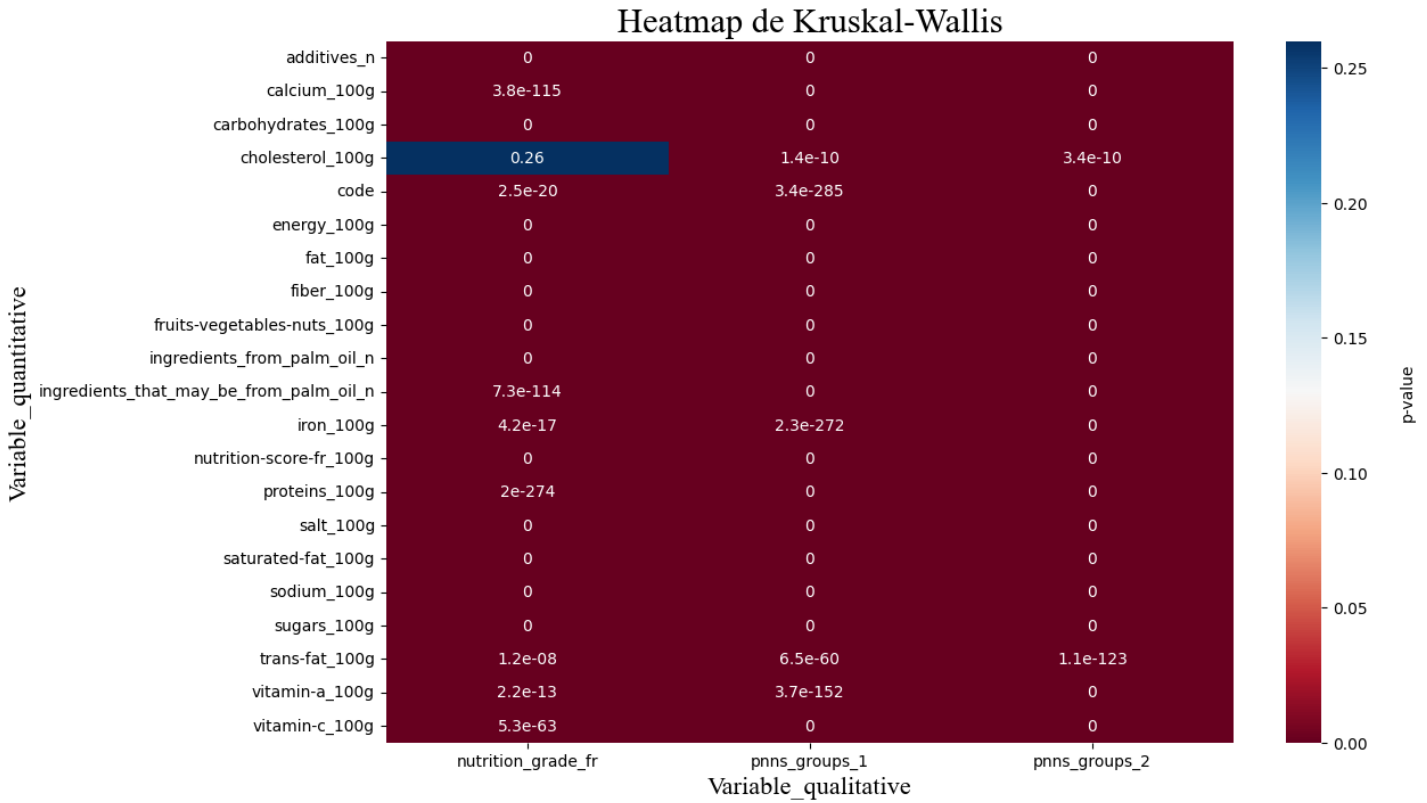
kw

Hypothèse nulle (H0) : la valeur de la variable qualitative n'influence pas les valeurs de la variable quantitative.
Hypothèse alternative (HA) : la valeur de la variable qualitative influence les valeurs de la variable quantitative.

Observations évaluant la pertinence / faisabilité de l'idée application du client

kw

~~Hypothèse nulle (H0) : la valeur de la variable qualitative n'influence pas les valeurs de la variable quantitative.~~
Hypothèse alternative (HA) : la valeur de la variable qualitative influence les valeurs de la variable quantitative.



=> Les valeurs des variables quantitatives sont influencées par les variables qualitatives relatives au Nutri-score et aux pnns des catégories d'aliment, à l'exception de la quantité de cholestérol qui n'est pas influencé par le Nutri-score.

Sommaire

RGPD

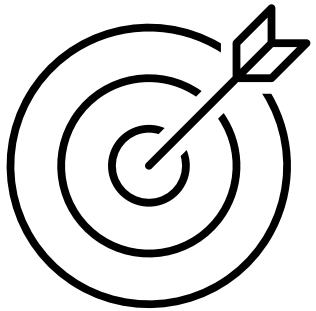
Voici les 5 grand principe RGPD (Règlement sur la Protection des Données Personnelles) :

1. Finalité
2. Pertinence
3. Durée limitée de conservation
4. Sécurité
5. Droits des personnes

Sommaire

Conclusion

- Nettoyer le jeu de données ✓
- Analyse univariée des variables importantes ✓
- Analyse multivariée et résultats statistiques ✓
- Conclusion sur la faisabilité du système d'auto-complétion ✓
- Rappel RGPD ✓



Conclusion

Faisabilité système d'auto-complétion:

- Corrélation données quantitatives
- Corrélation données qualitatives
- Influence des données qualitatives sur les données quantitatives

Limites :

- Données obsolètes (pas mis à jour depuis 2017)
- 30% des produits sont de catégorie « inconnu »
- Application de règle métier pour détecter des variables aberrantes (lipide / énergie)
- Mise à jour des Nutri-score (méthode de calcul 2023)

La création d'un système d'auto-complétion fonctionnel pour notre client "Santé Publique France" n'est pas encore possible avec ce jeu de données. Plusieurs travaux d'amélioration de sa qualité sont à prévoir encore en amont.

Merci pour votre attention

Pierrick BERTHE

Formation Expert en Data Science
Openclassrooms – CentraleSupélec

août 2023 –> avril 2024