OpenClassrooms

PROJET 3

Concevez une application au service de la santé publique

Kilian ALLIOT

Parcours Data Scientist

PLAN DE LA PRÉSENTATION

- Présentation du sujet
- La problématique
- Mon application
- Présentation des données
- Nettoyage
- Analyse exploratoire
- Conclusion

Santé publique France

PRÉSENTATION DU SUJET

L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d’applications en lien avec l'alimentation. Vous souhaitez y participer et proposer une idée d’application.

PROBLÉMATIQUE

Proposer une idée d'application pertinente d'un point de vue santé,
et réalisable à partir des données mises à ma disposition (après nettoyage et analyse).

MON APPLICATION, HFC

HFC, Healthy French Courses, permet de rechercher ou scanner un produit pendant vos courses.
L'application vous propose alors des produits alternatifs plus sains, moins gras, moins sucrés, moins salés.

Lorsque c'est possible, nos produits favoris seront mis en avant, ils :

- sont d'origine française
- sont transformés/emballés en France
- ont un label BIO
- ont un emballage recyclable

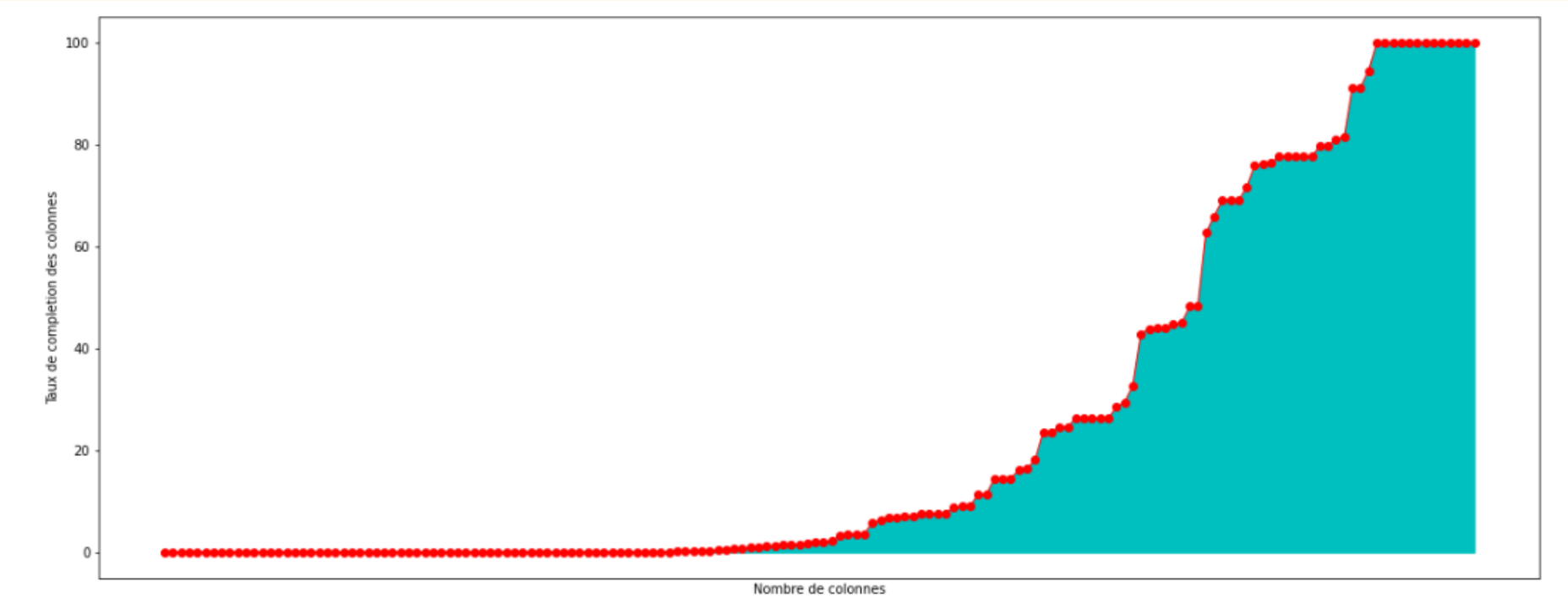
LES DONNÉES

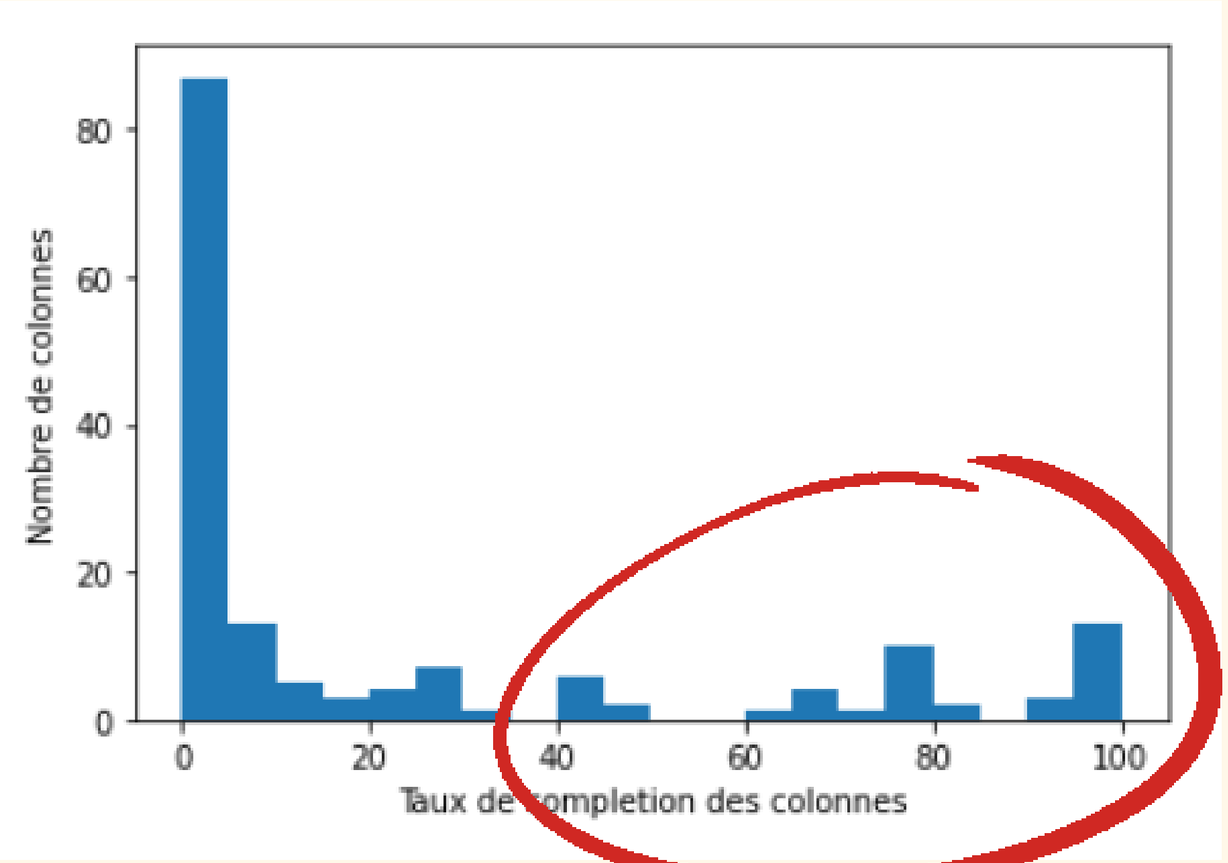
```
data.shape
```

```
(320772, 162)
```

- 320 772 lignes/indiividus
- 162 colonnes/variables

Taux de complétion des colonnes



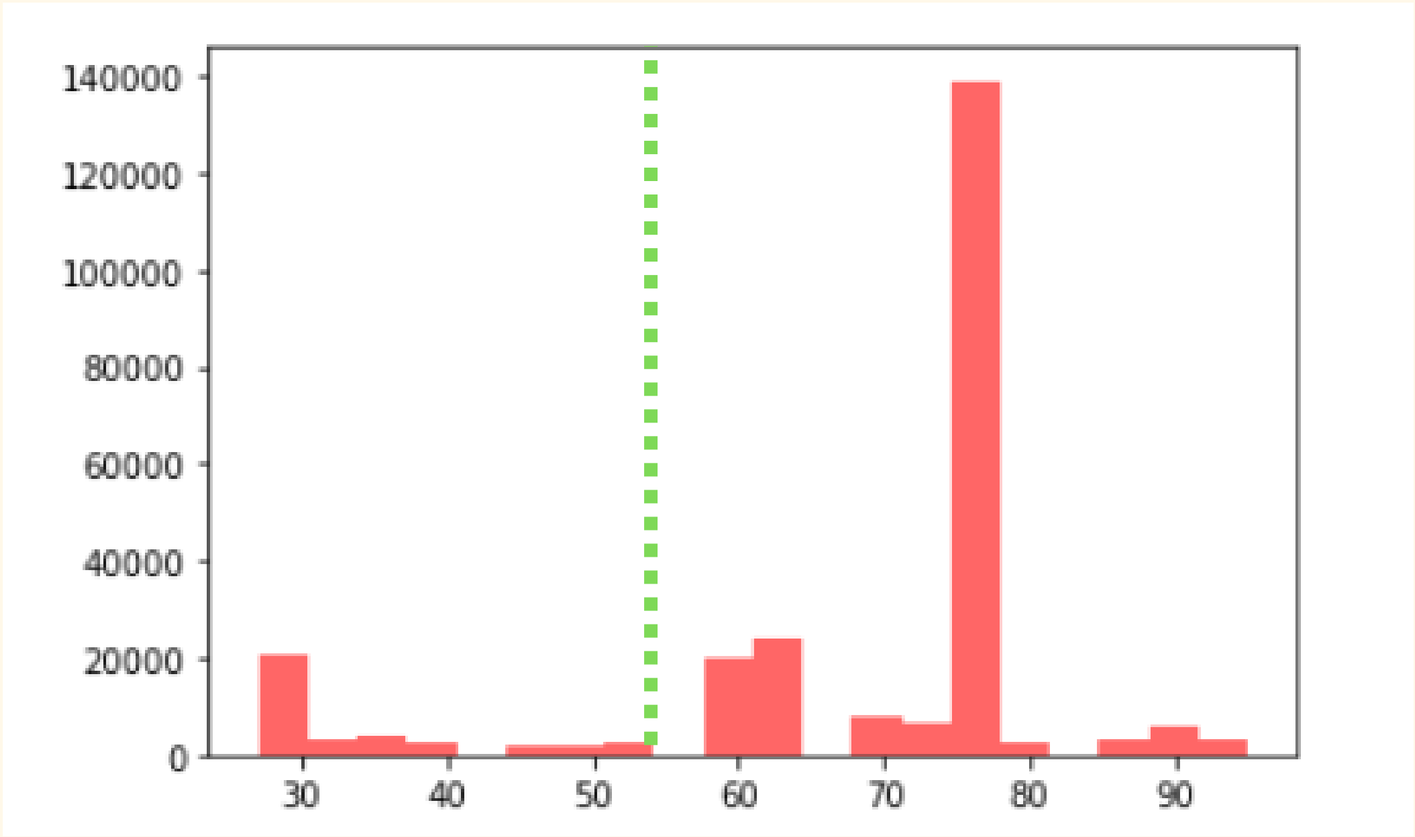


Définir un seuil acceptable, garder des colonnes exploitables pour mon idée d'application

NETTOYAGE DES DONNÉES

- Sélection des colonnes grâce aux taux de remplissage (162 -> 22)
- Suppression des lignes dupliquées
- Mise à nan des valeurs aberrantes (nutrition_100g)
- Imputation simple (manufacturing_place et emb_code)
- Uniformiser certaines modalités (ex: sugary-snacks et Sugary snacks)

Taux de complétion des lignes APRES nettoyage



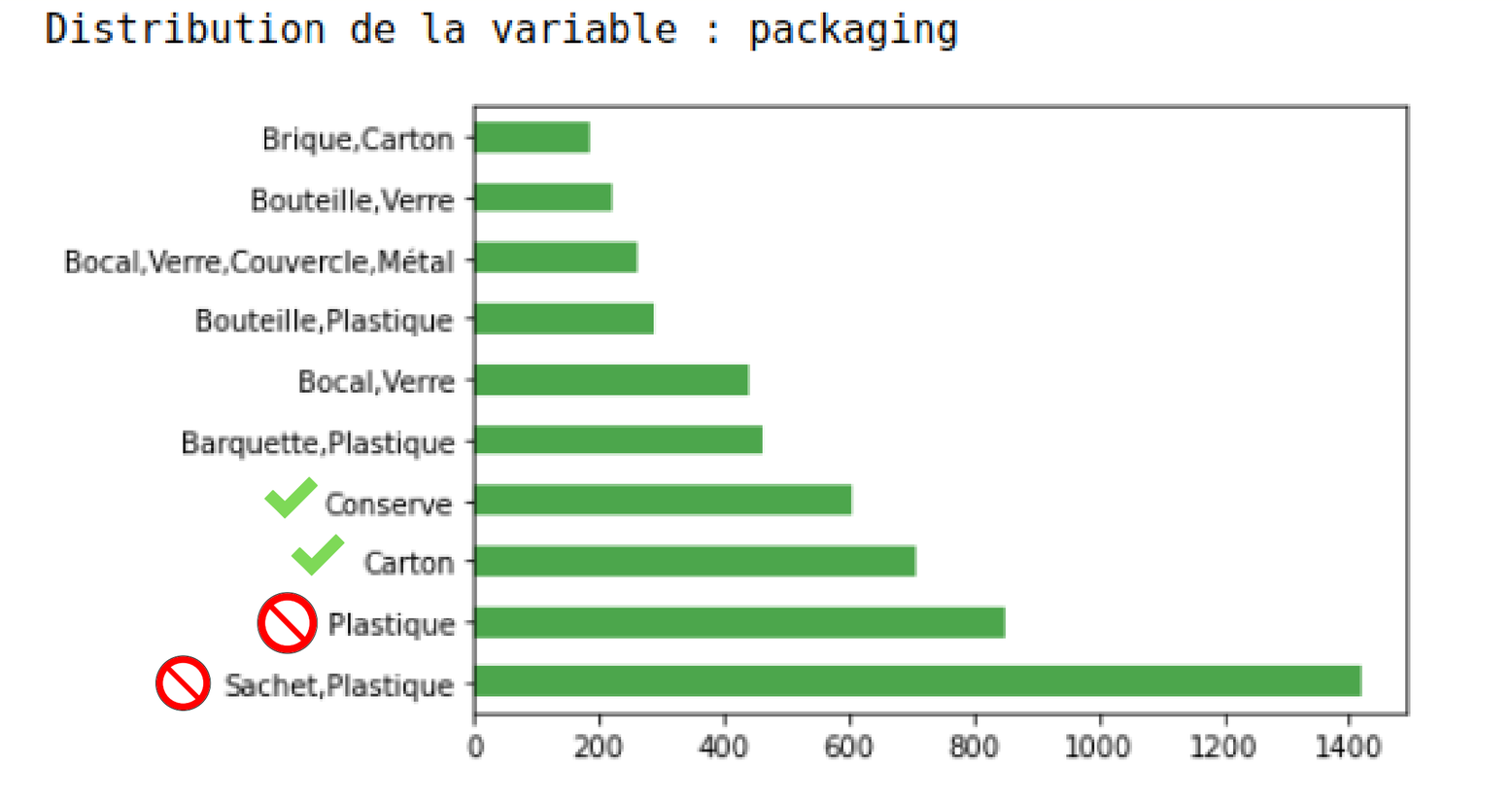
Seuil fixé à 55%

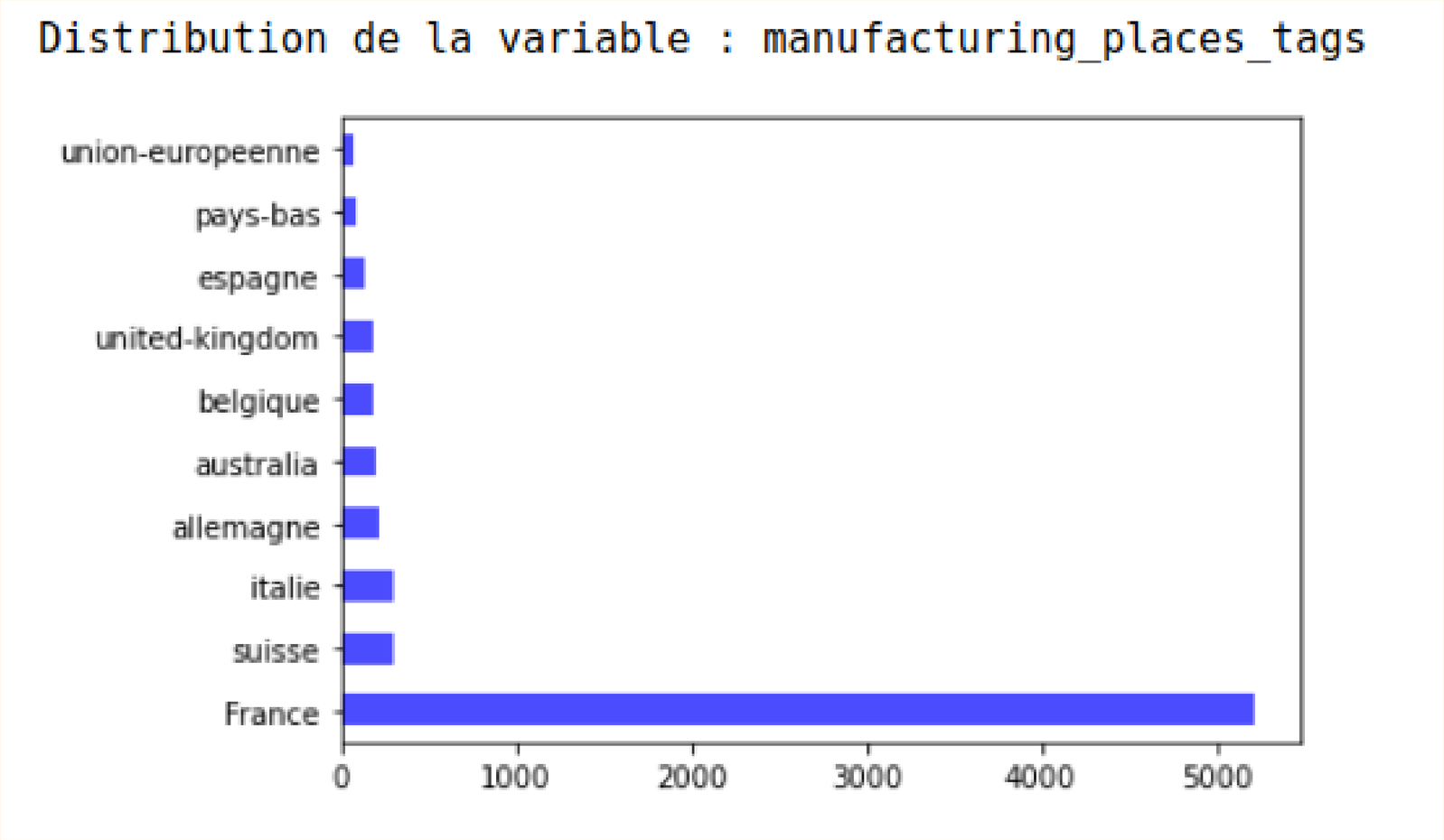
IMPUTATION KNN

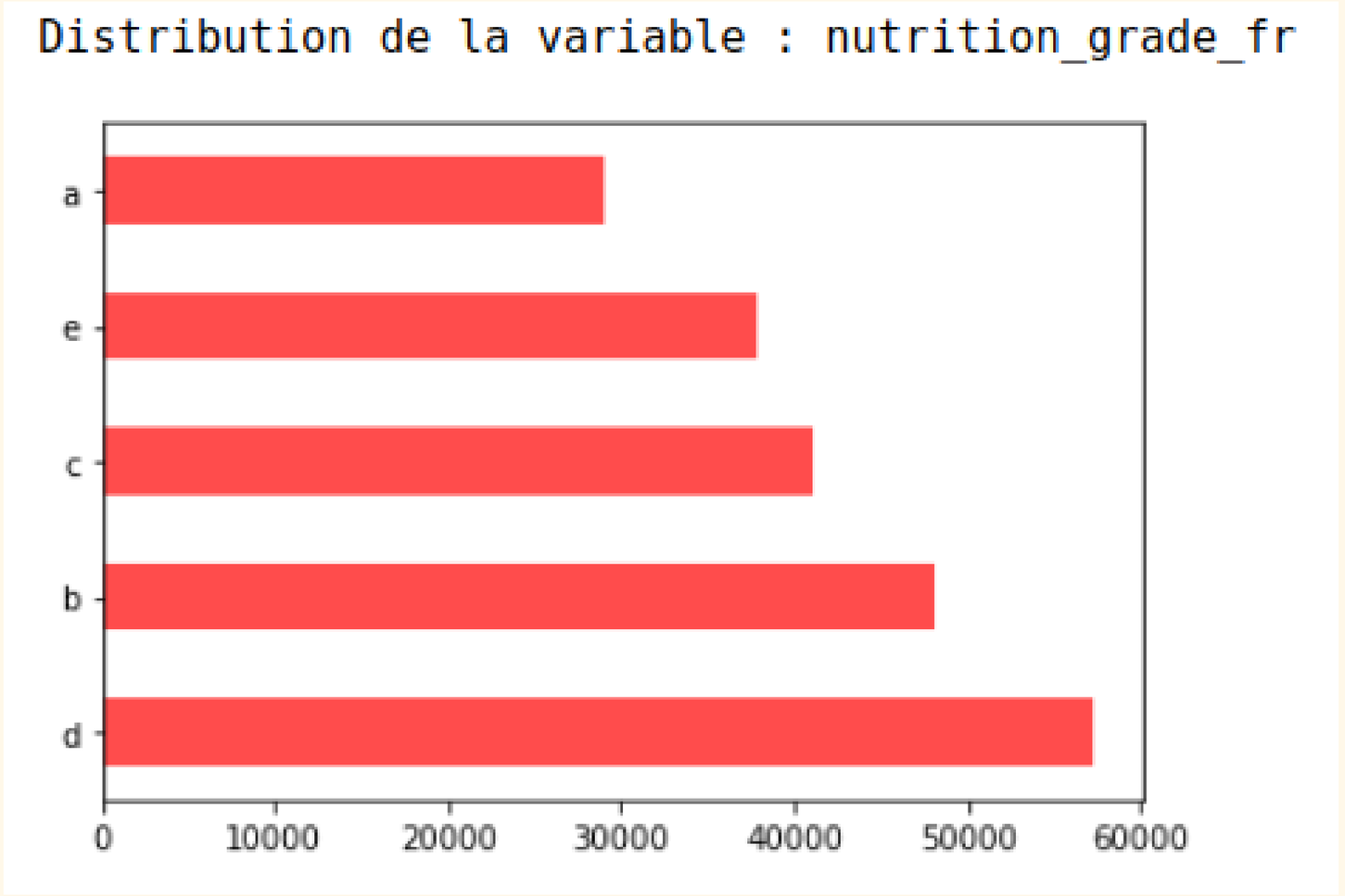
- Variables qualitatives : one hot encoding/dummy variables
- Choix sur pnns groups et nutrition score car nombre de modalités acceptables
- Centrer et réduire les données pour que les colonnes aient le même impact (distance euclidienne)
- Imputation avec k=1
- Décoder les dummies

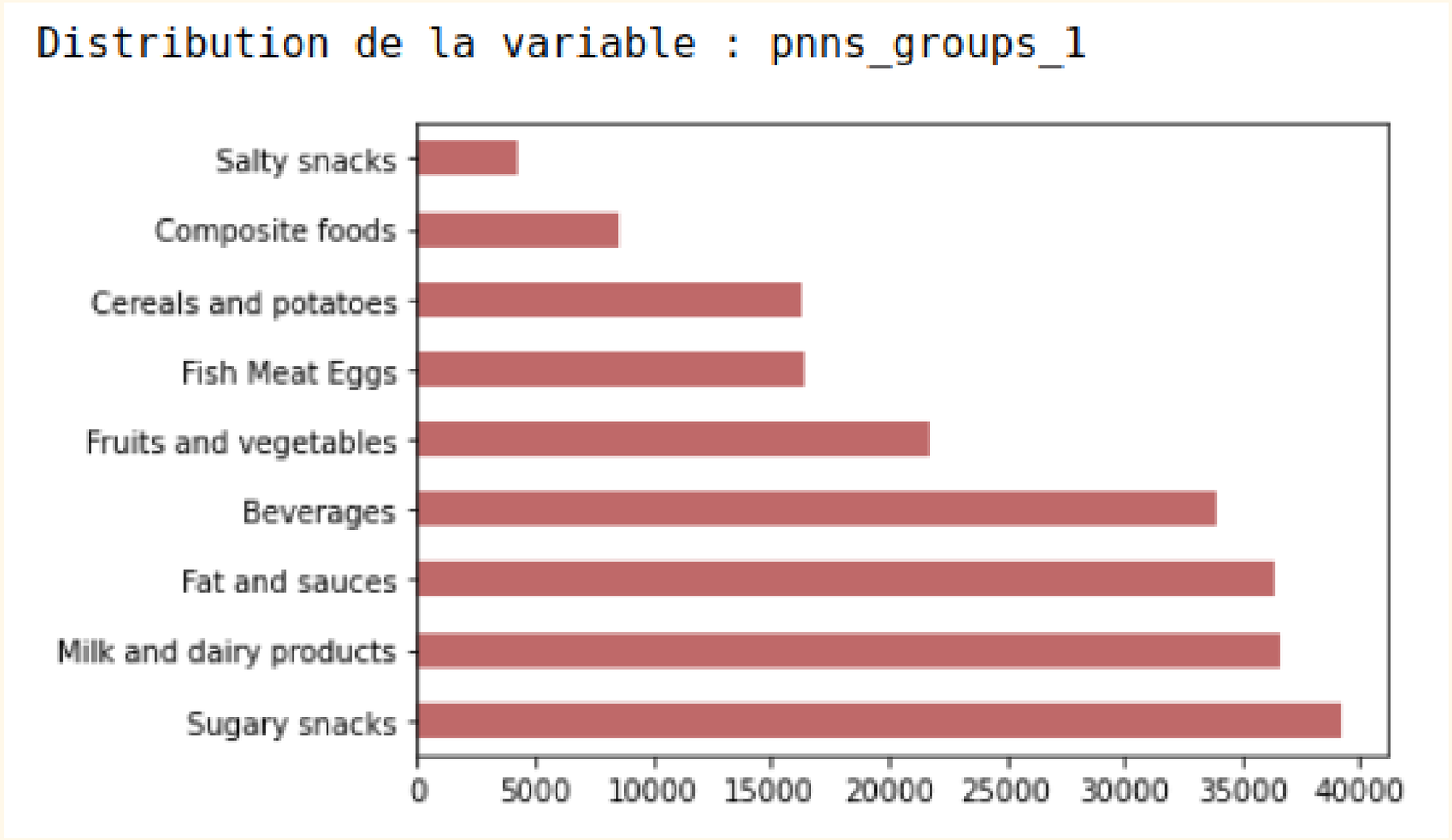
ANALYSE EXPLORATOIRE

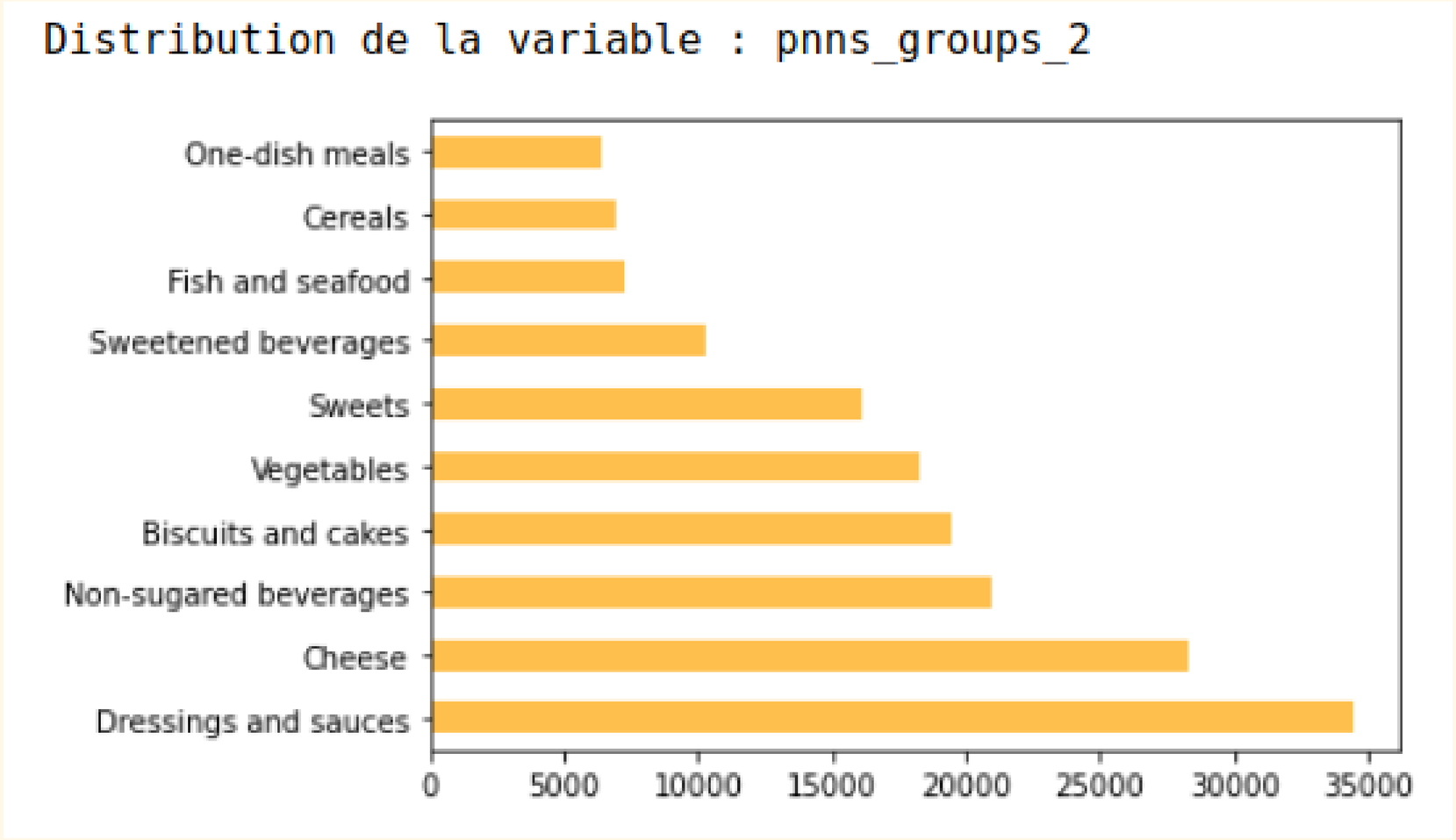
Analyse univariée

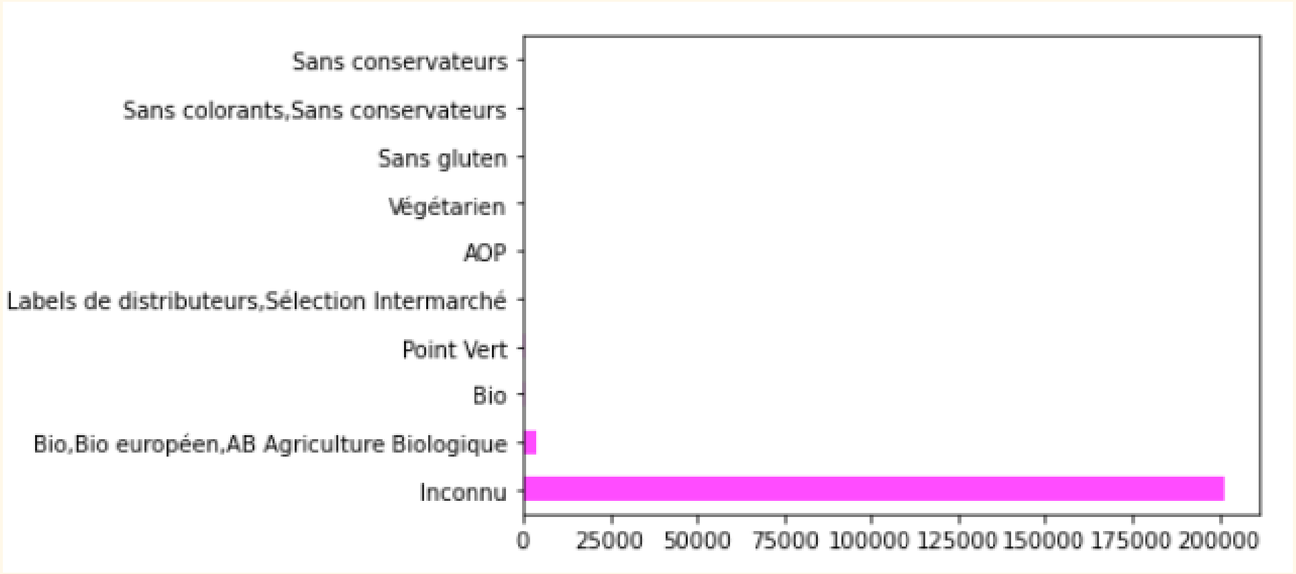


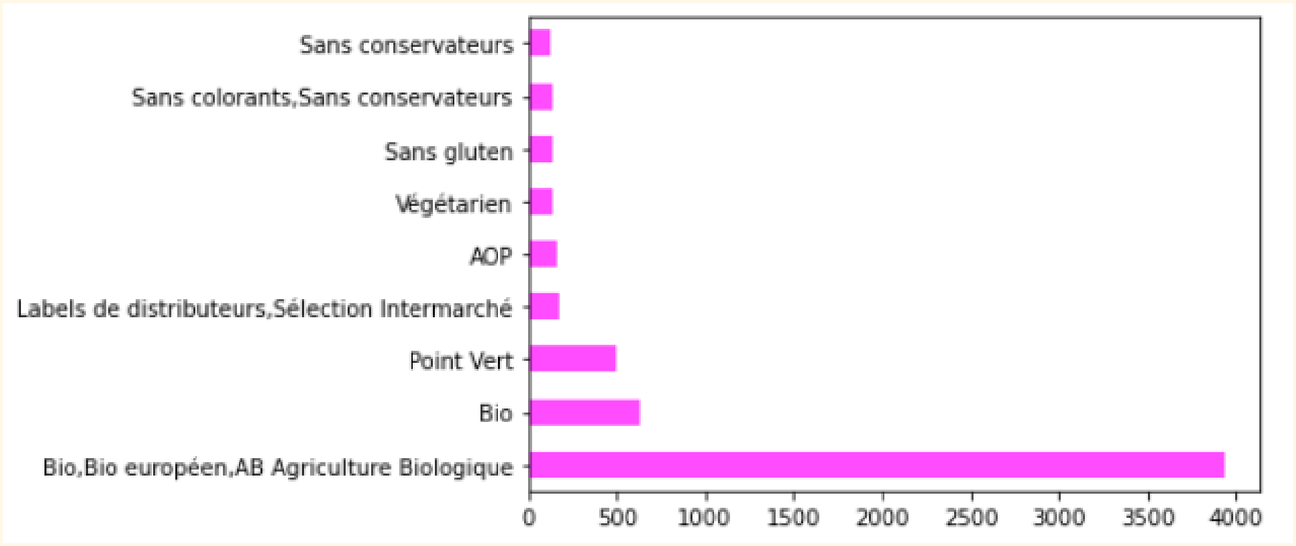

Distribution de la variable : manufacturing_places_tags
union-europeenne
pays-bas
espagne
united-kingdom
belgique
australia
allemagne
italie
suisse
France
0
1000
2000
3000
4000
5000


Distribution de la variable : nutrition_grade_fr
a
e
c
b
d
0
10000
20000
30000
40000
50000
60000

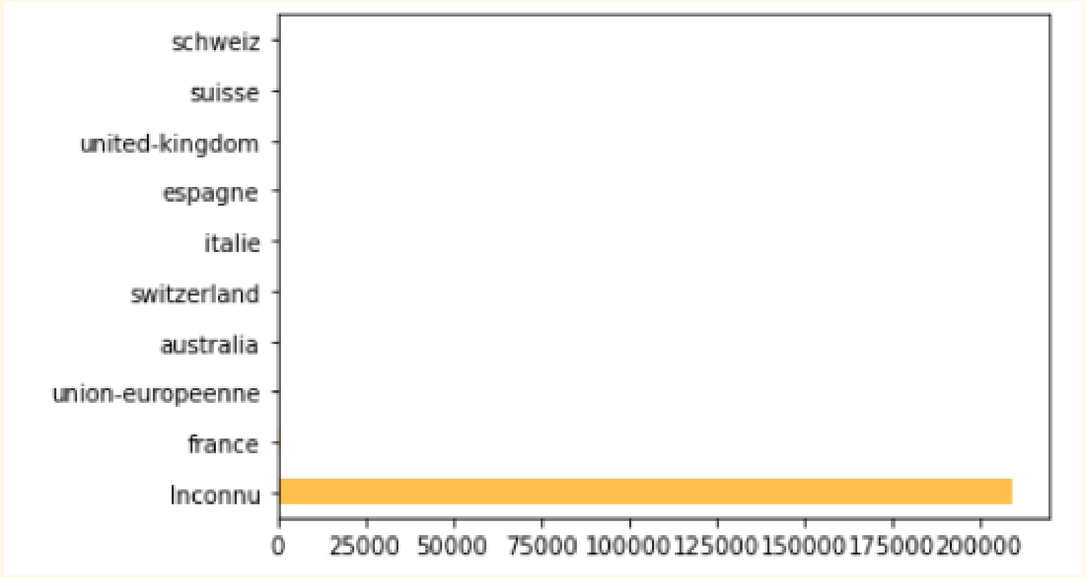

Distribution de la variable : pnns_groups_1
Salty snacks
Composite foods
Cereals and potatoes
Fish Meat Eggs
Fruits and vegetables
Beverages
Fat and sauces
Milk and dairy products
Sugary snacks
0
5000
10000
15000
20000
25000
30000
35000
40000

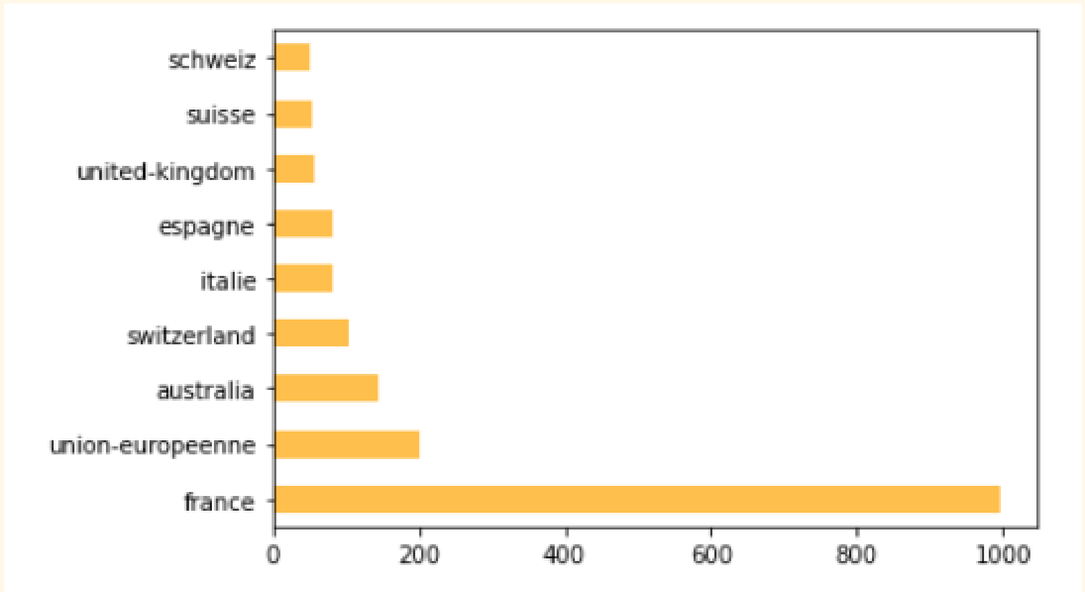

Distribution de la variable : pnns_groups_2
One-dish meals
Cereals
Fish and seafood
Sweetened beverages
Sweets
Vegetables
Biscuits and cakes
Non-sugared beverages
Cheese
Dressings and sauces
0
5000
10000
15000
20000
25000
30000
35000


Sans conservateurs
Sans colorants,Sans conservateurs
Sans gluten
Végétarien
AOP
Labels de distributeurs,Sélection Intermarché
Point Vert
Bio
Bio,Bio européen,AB Agriculture Biologique
Inconnu
0
25000
50000
75000
100000
125000
150000
175000
200000

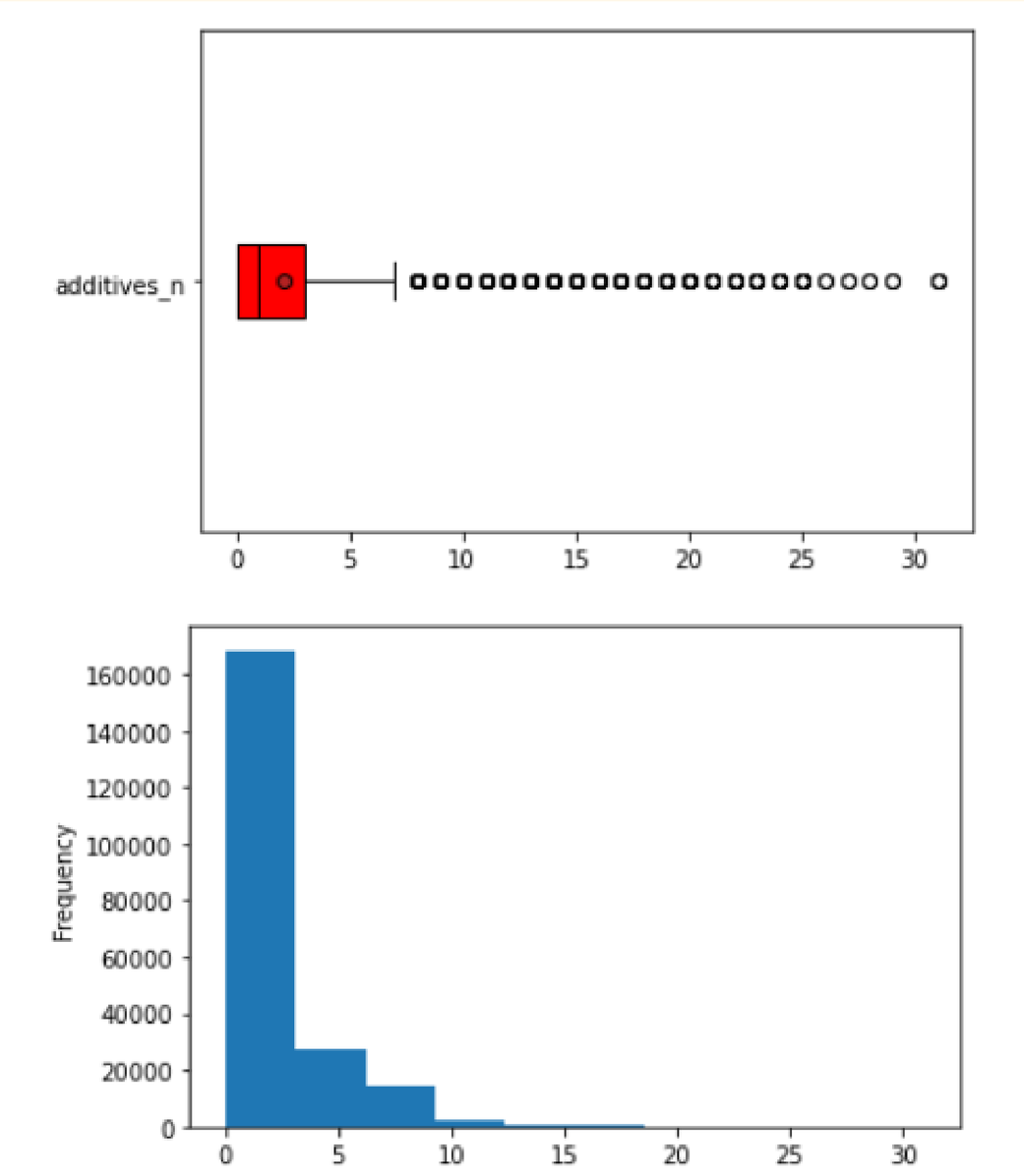

Sans conservateurs
Sans colorants,Sans conservateurs
Sans gluten
Végétarien
AOP
Labels de distributeurs,Sélection Intermarché
Point Vert
Bio
Bio,Bio européen,AB Agriculture Biologique
0
500
1000
1500
2000
2500
3000
3500
4000

Labels


schweiz
suisse
united-kingdom
espagne
italie
switzerland
australia
union-europeenne
france
Inconnu
0
25000
50000
75000
100000
125000
150000
175000
200000

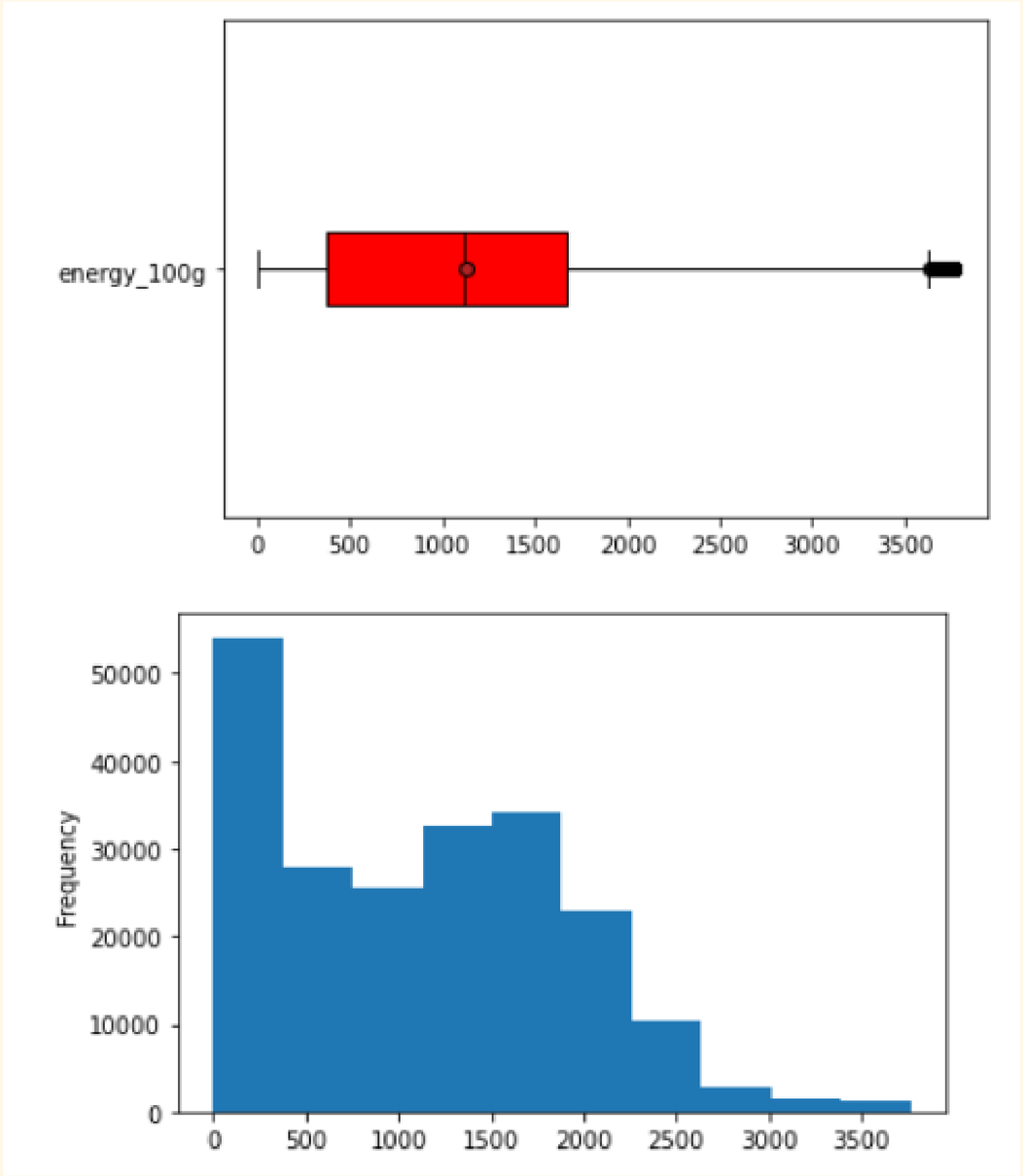

schweiz
suisse
united-kingdom
espagne
italie
switzerland
australia
union-europeenne
france
0
200
400
600
800
1000

Origines



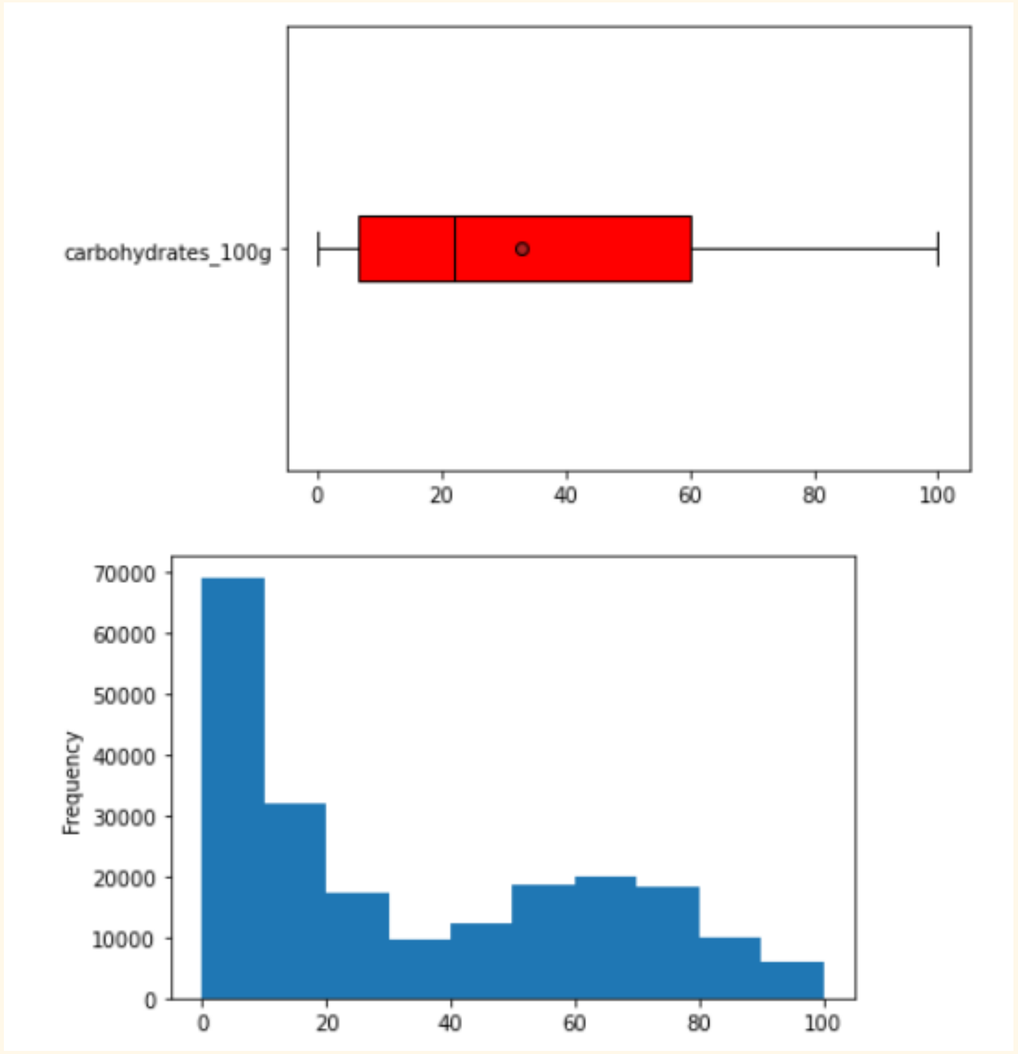
Nombre d'additifs

L'EFSA analyse les risques pour la santé liés à la consommation, entre autres, d'aliments contenant des **additifs**. Certains **additifs** sont responsables d'intolérances et de réactions allergiques. D'autres sont suspectés d'être cancérigènes. C'est pourquoi certains **additifs** sont autorisés mais en dose limitée. 10 sept. 2017

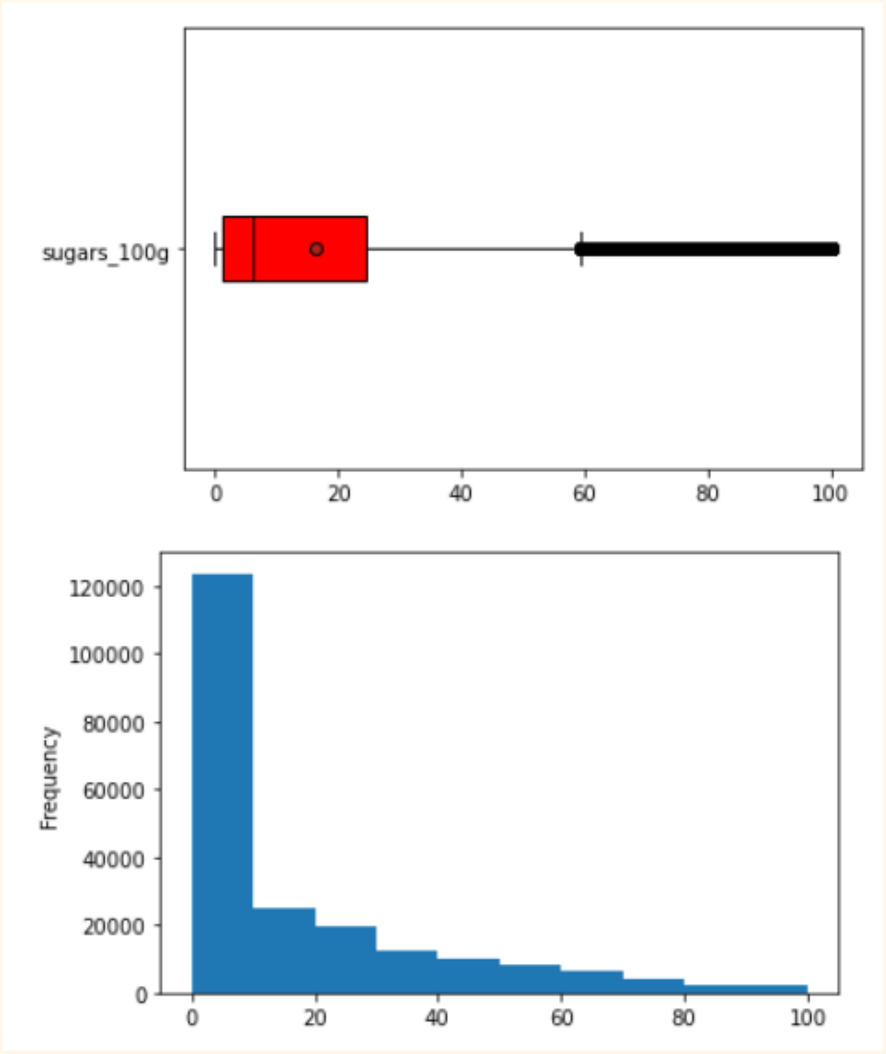


Energie pour 100g (kJ)

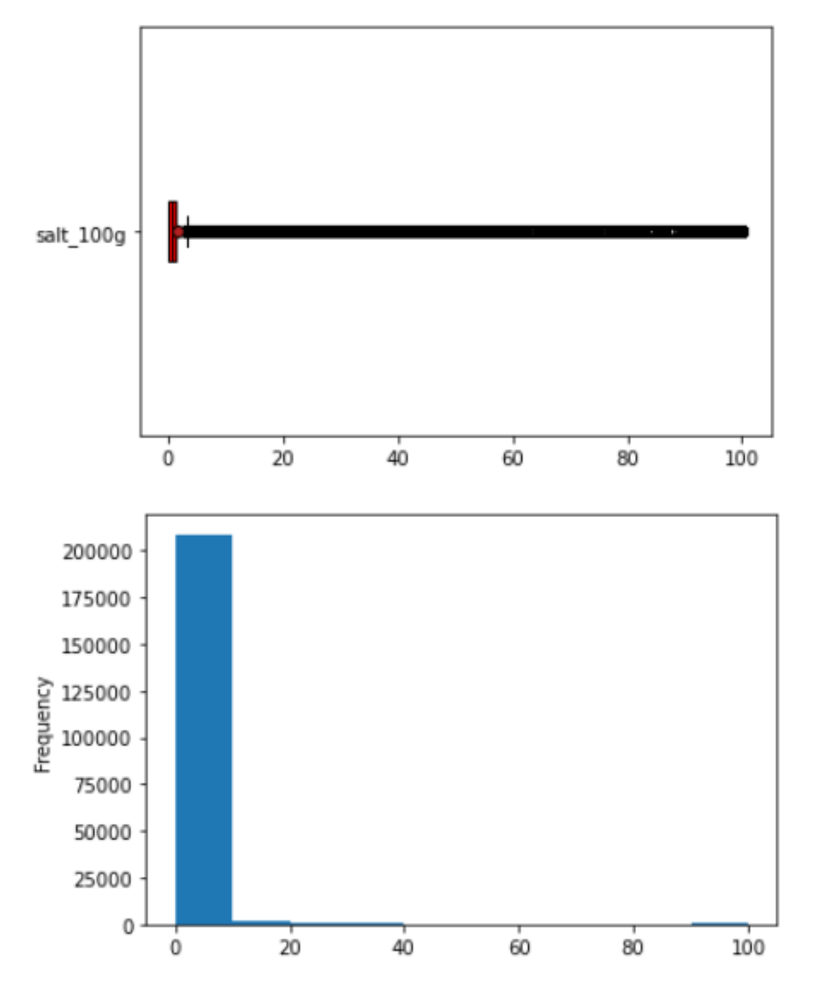
Glucides



Sucre

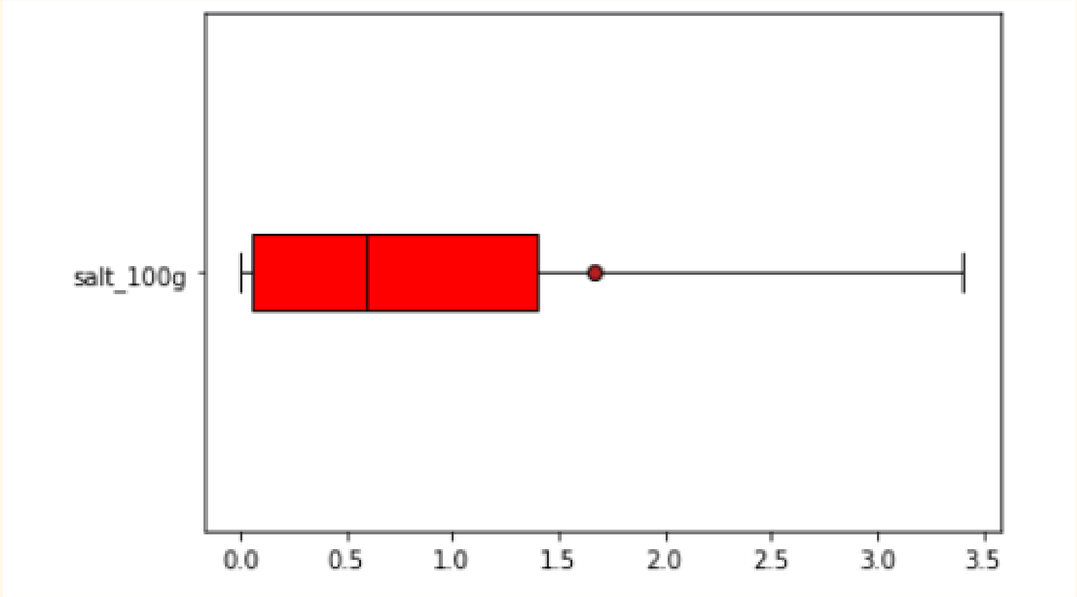


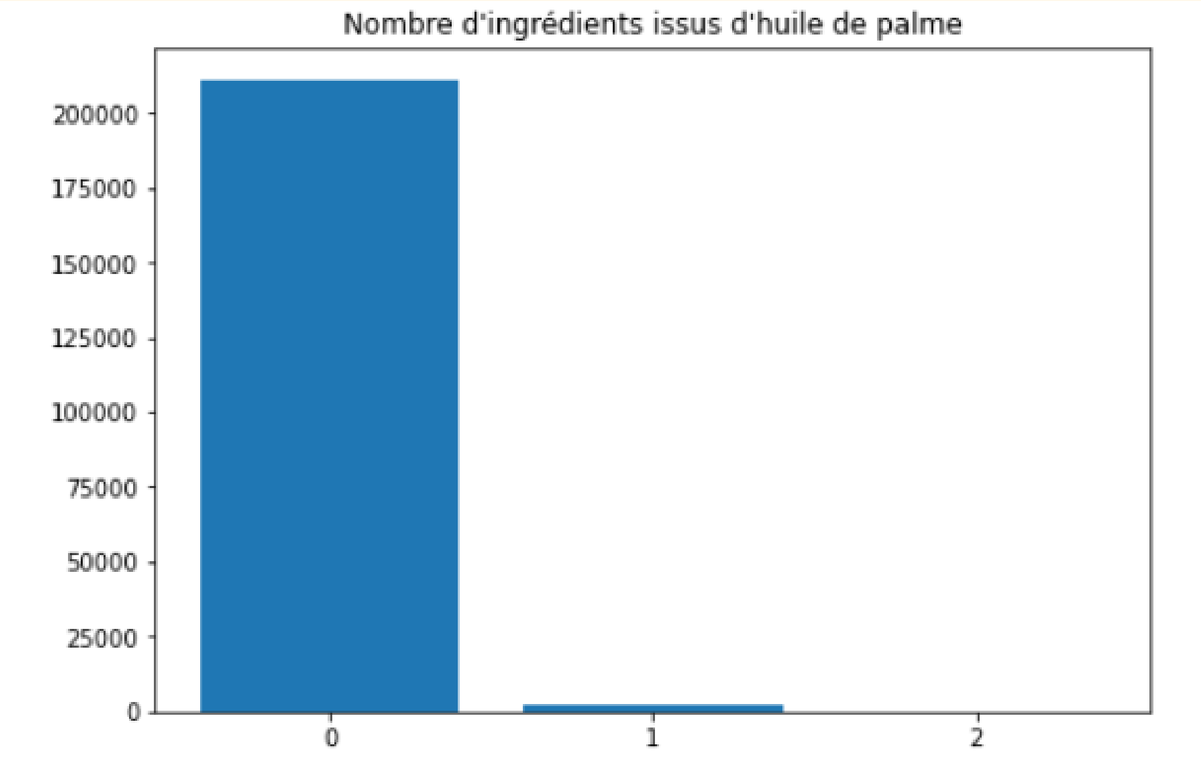
L'hypertension artérielle concerne un adulte sur trois en France



L'excès de sel peut avoir des conséquences néfastes sur la santé. Rétention d'eau, vieillissement de la peau, maladies cardiovasculaires, augmentation des risques de déclin cognitif, hypertension artérielle, risques d'œdèmes, insuffisance cardiaque ou rénale… la liste est longue. 27 juil. 2016

Quel est l'**apport quotidien** en **sel** recommandé chez l'adulte ? L'OMS recommande une consommation inférieure à 5 g par jour (soit l'équivalent d'une cuillère à café de **sel** par jour) afin de prévenir les maladies cardiovasculaires.





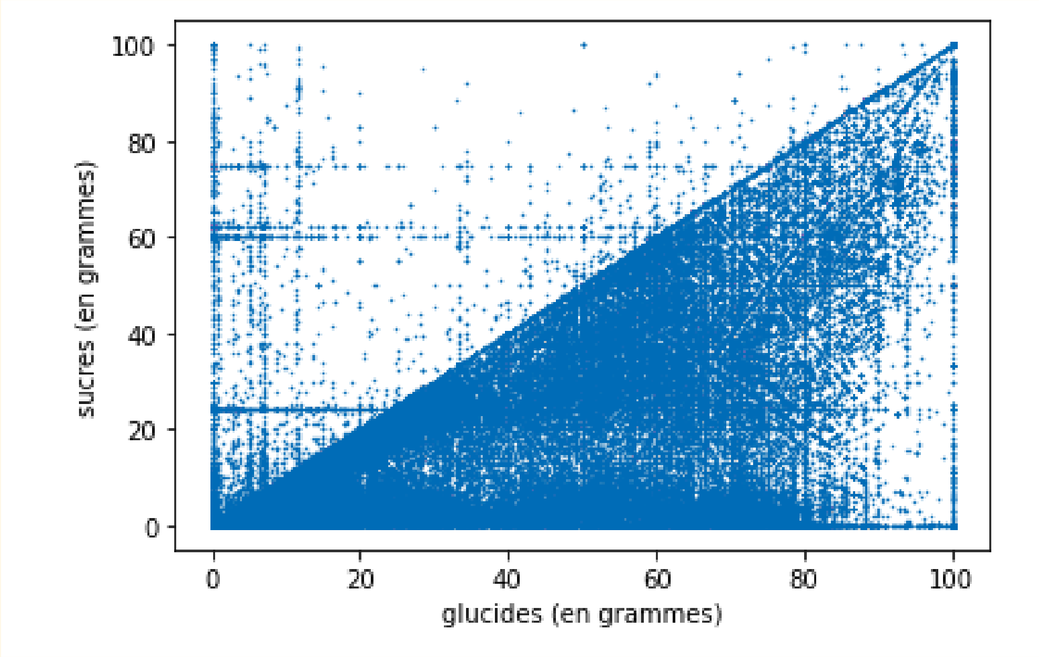
L'**huile de palme est** réputée **mauvaise** pour la santé en raison de son fort taux d'acides gras saturés, qui augmentent le **mauvais** cholestérol.

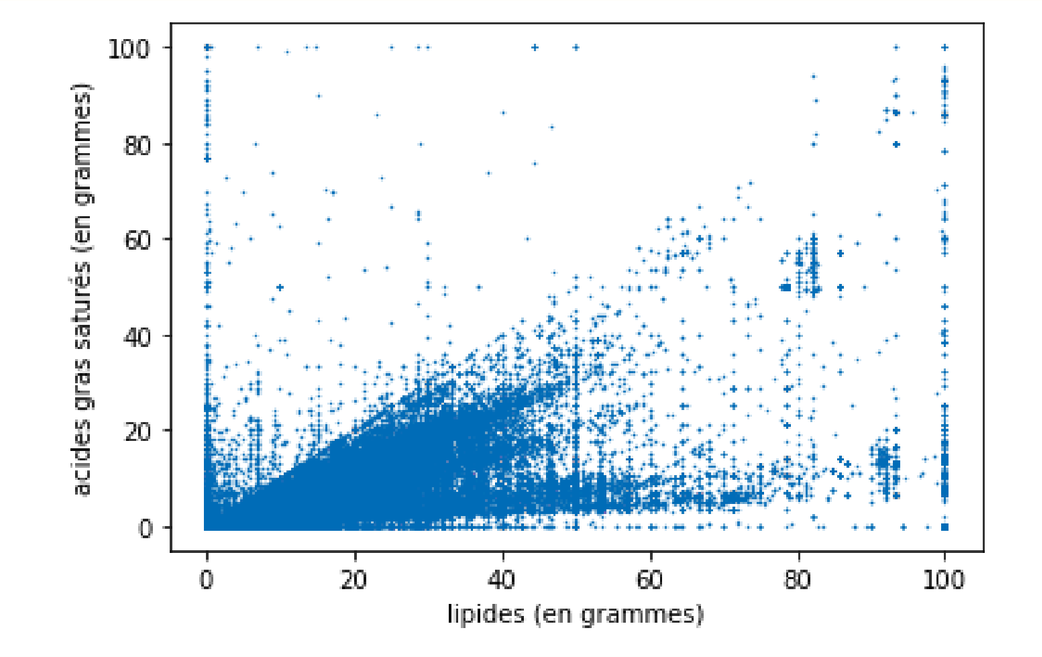
Sous sa forme actuelle, la production d'**huile de palme est** responsable d'une importante déforestation, elle contribue à la disparition de nombreuses espèces à l'image des Orangs-outangs, use de produits hautement toxiques et les conditions de travail dans les plantations y **sont** souvent déplorables.

0.0	211386
1.0	1955
2.0	30

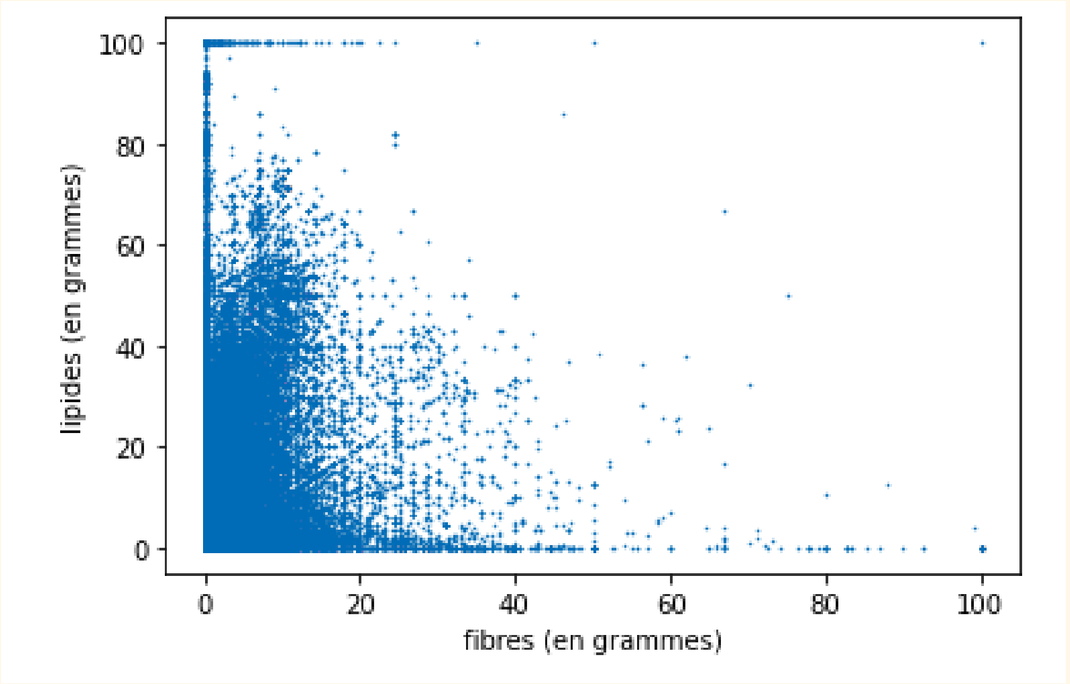
Analyse bivariée

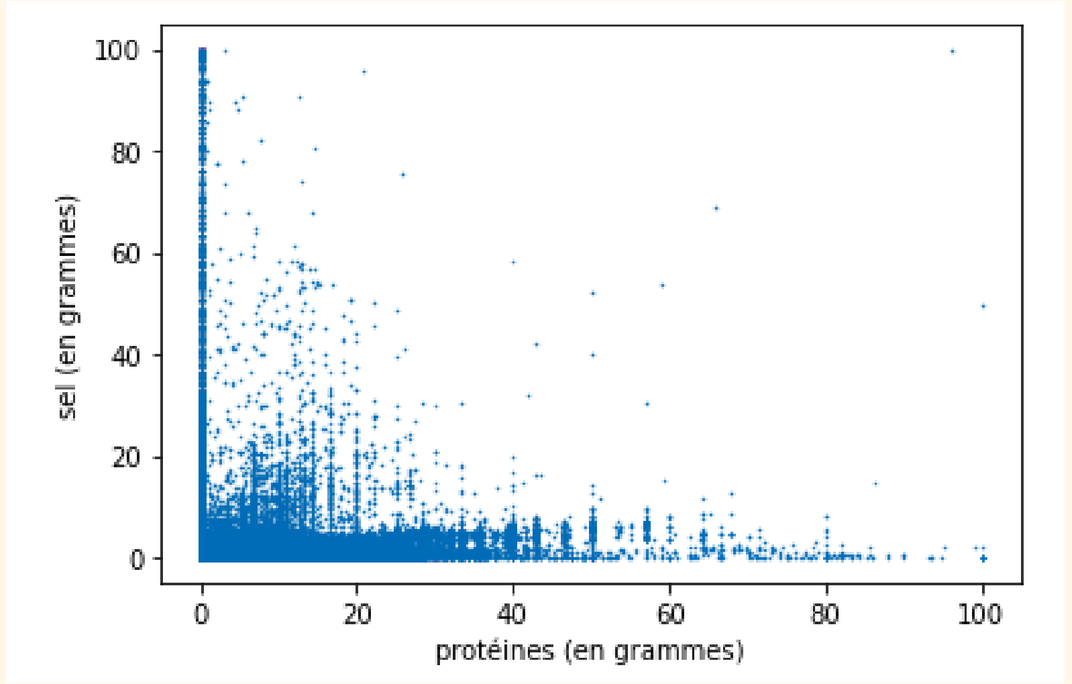
Un produit avec 20g de glucides ne peut avoir plus de 20g de sucres.
C'est pourquoi ce triangle apparaît lors de la data visualisation.

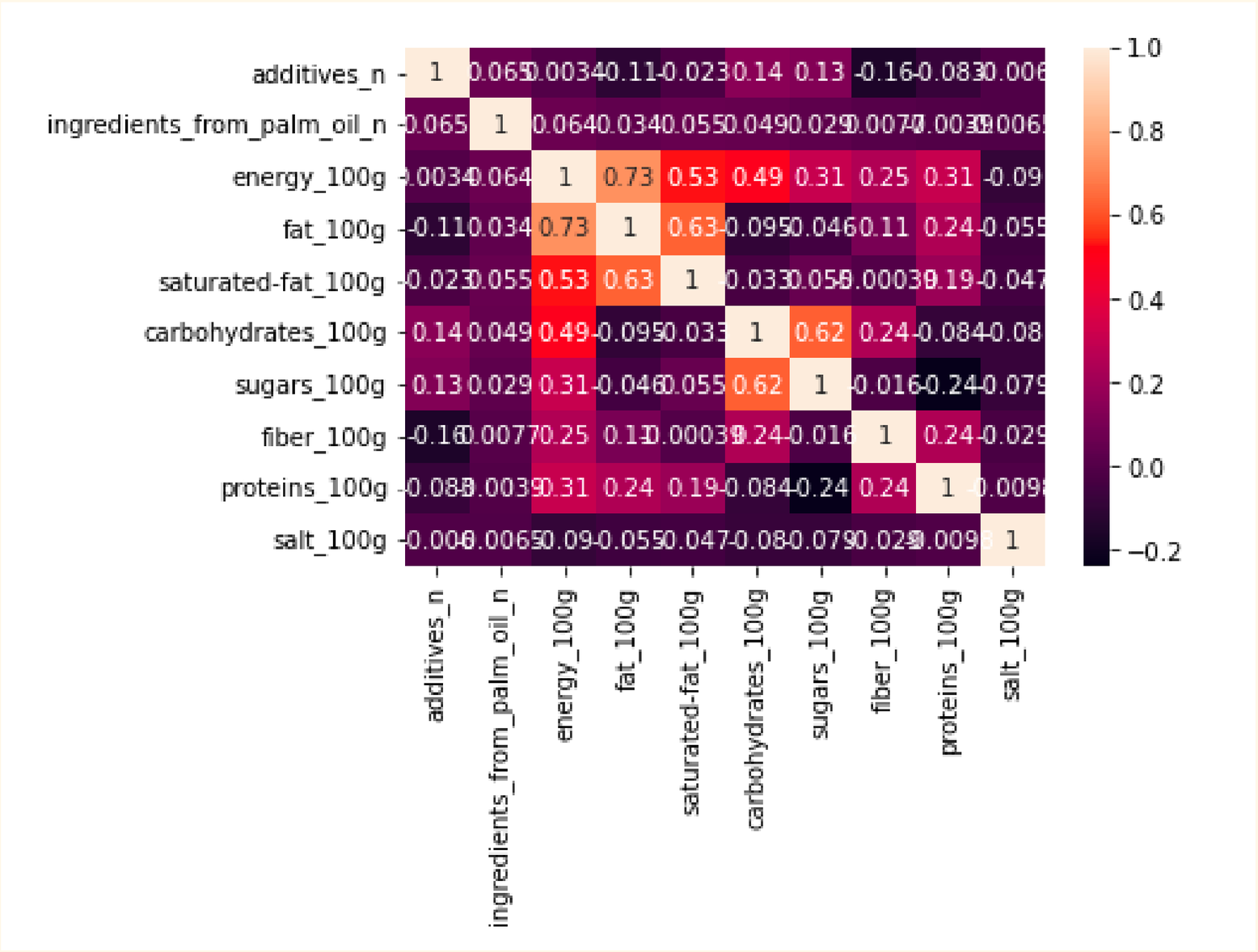




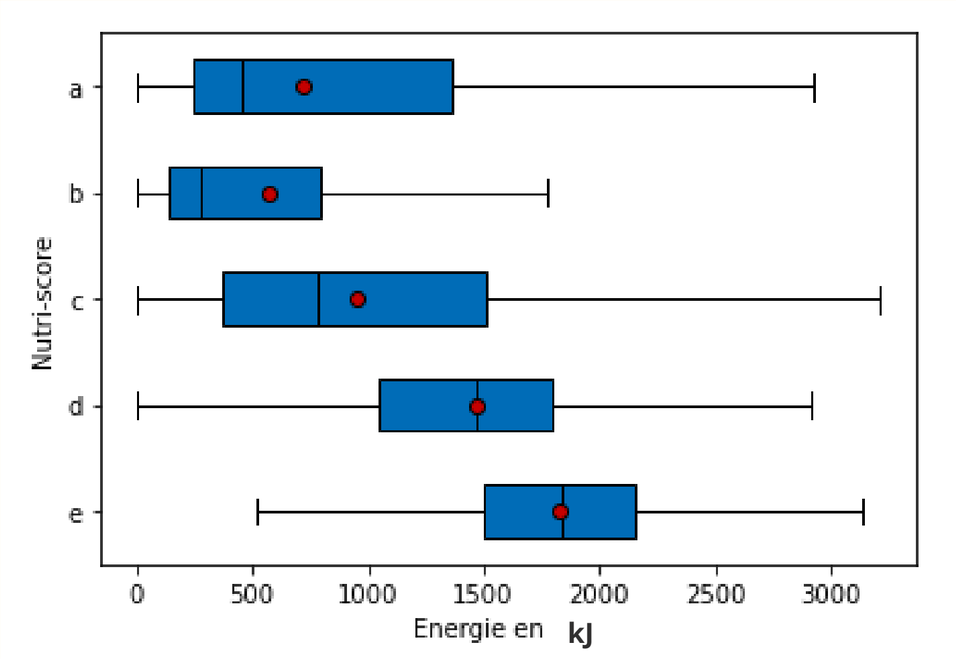
On pourrait penser que les produits à haute teneur en fibres comme les fruits et légumes possèdent peu de lipides.
Il semblerait que les produits à haute teneur en protéines ne soient pas des produits transformés, l'apport en sel serait moindre







Matrice de corrélation
Coef de Pearson

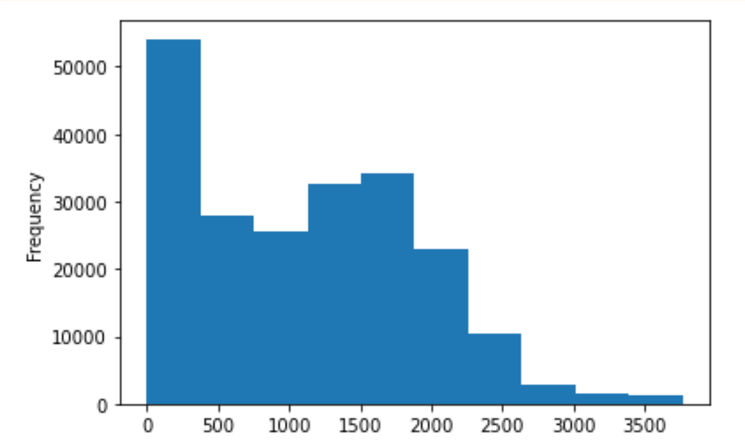


Analyse entre variable qualitative et quantitative:
Anova(paramétrique)/Kruskal-Wallis(non paramétrique)

Les pré-requis ne sont pas respectés:

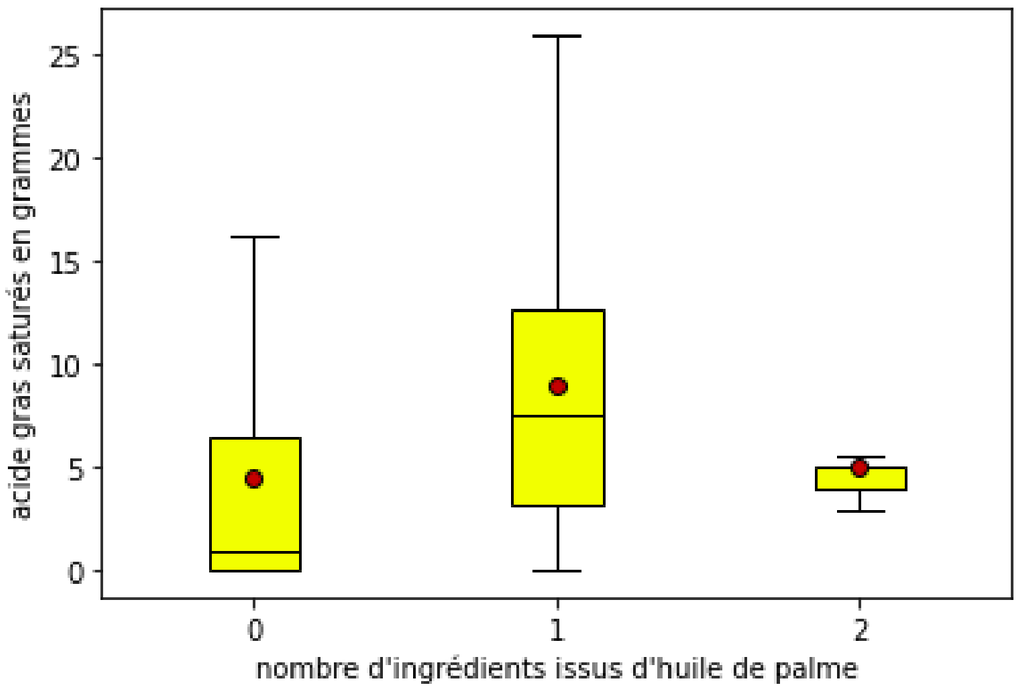
- variances similaires dans les groupes
- distribution normale
- indépendance des observations

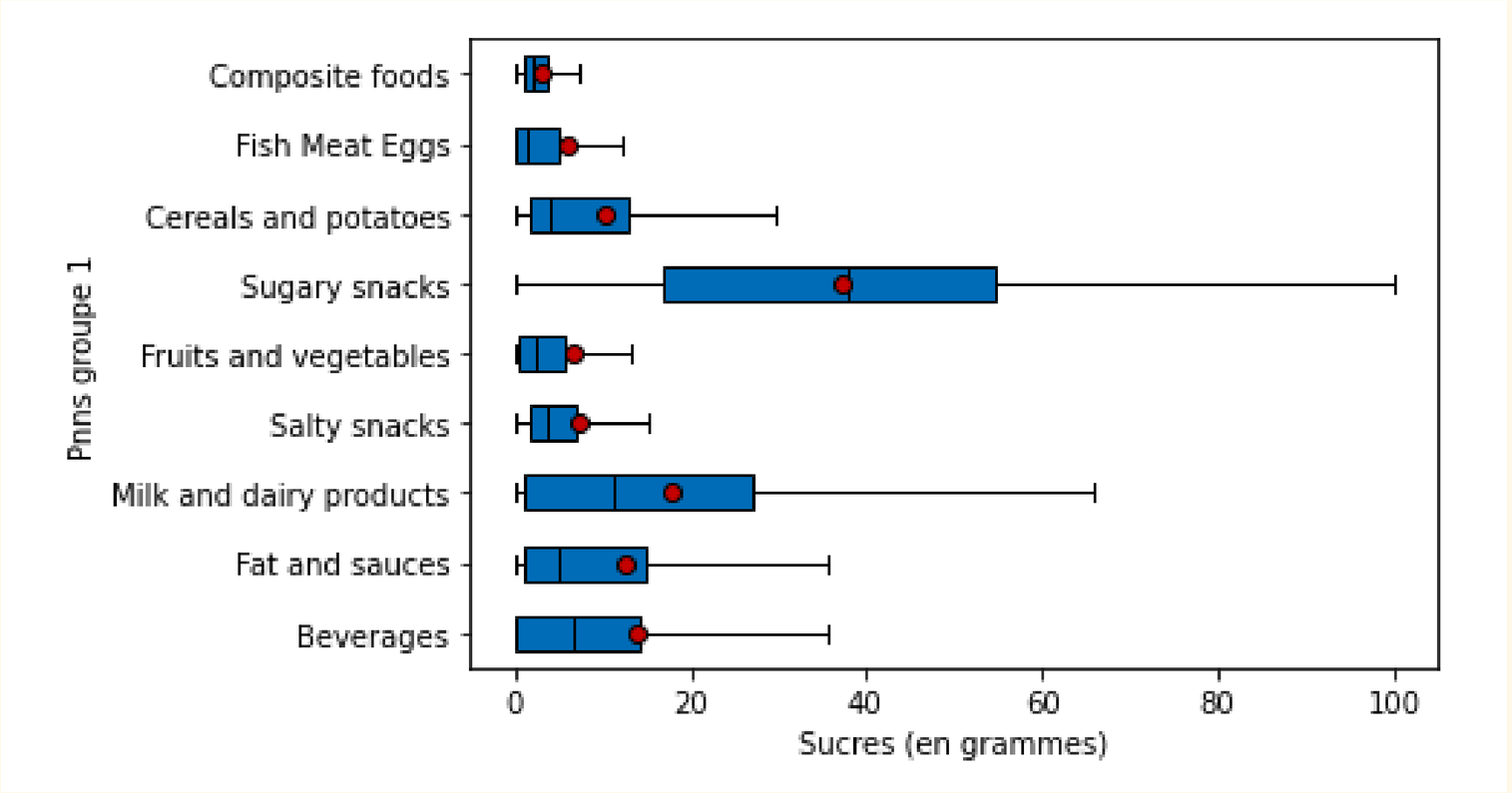
Pvalue de 0 suite au test de Kruskal-Wallis, on peut donc conclure qu'au moins un groupe ne provient pas de la même population que les autres



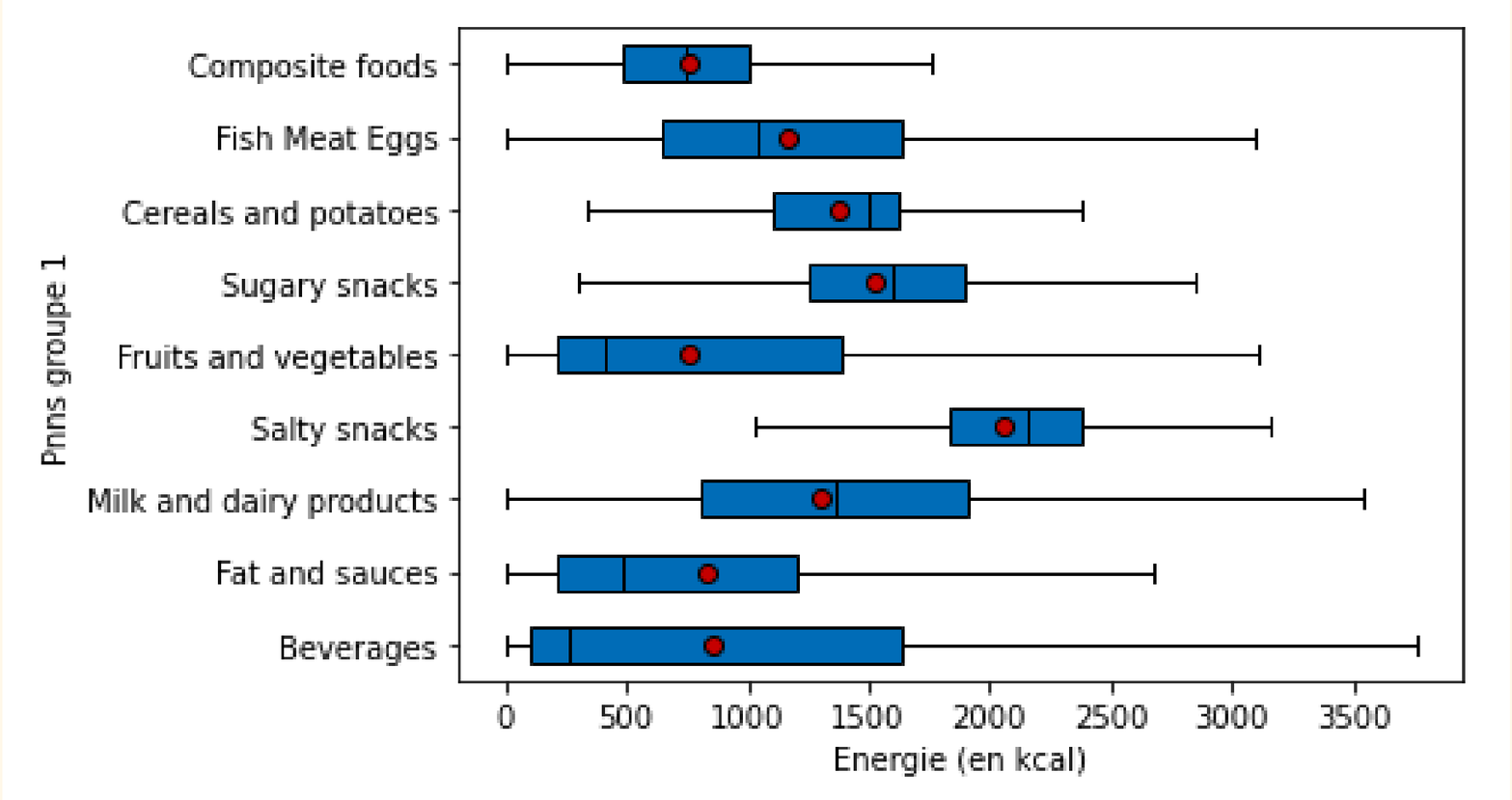

```
variance pour le groupe e
371979.5959034175
variance pour le groupe d
353735.04949712206
variance pour le groupe c
529594.8867714842
variance pour le groupe b
481511.6406760018
variance pour le groupe a
313625.89532218047
```

KruskalResult(statistic=1634.3692086449796, pvalue=0.0)

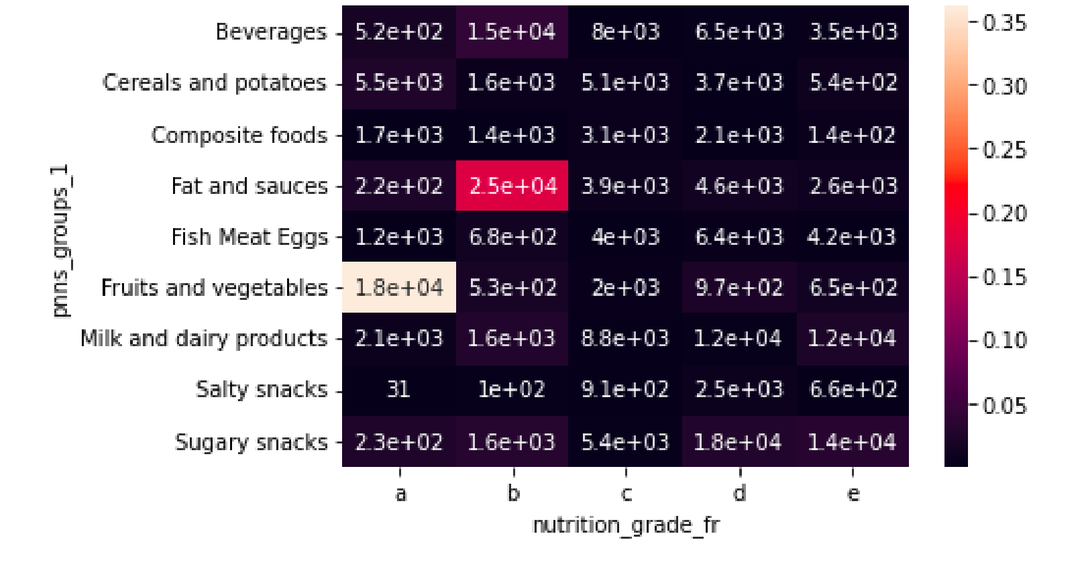


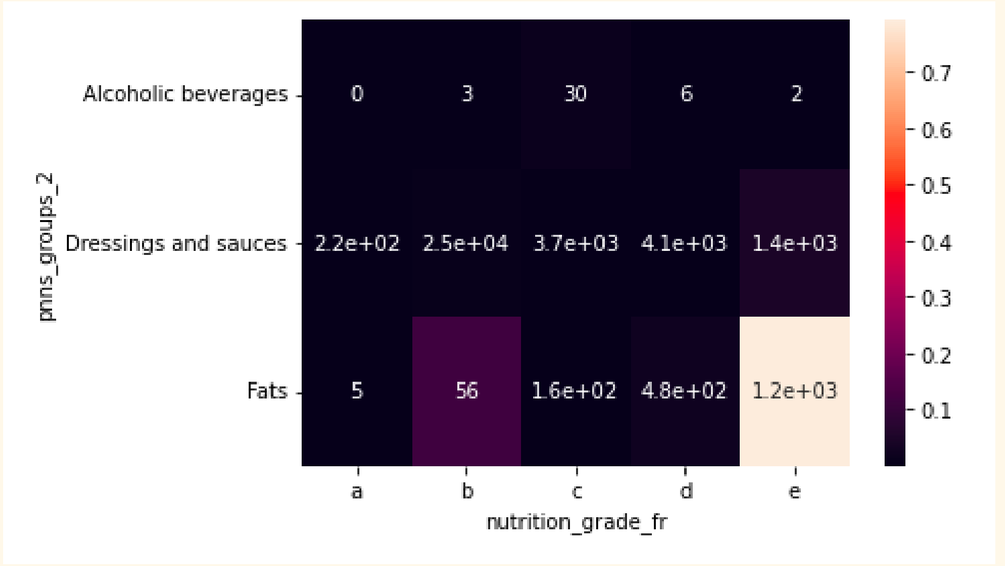


Pvalue de 0 suite au test de Kruskal-Wallis, on peut donc conclure qu'au moins un groupe ne provient pas de la même population que les autres



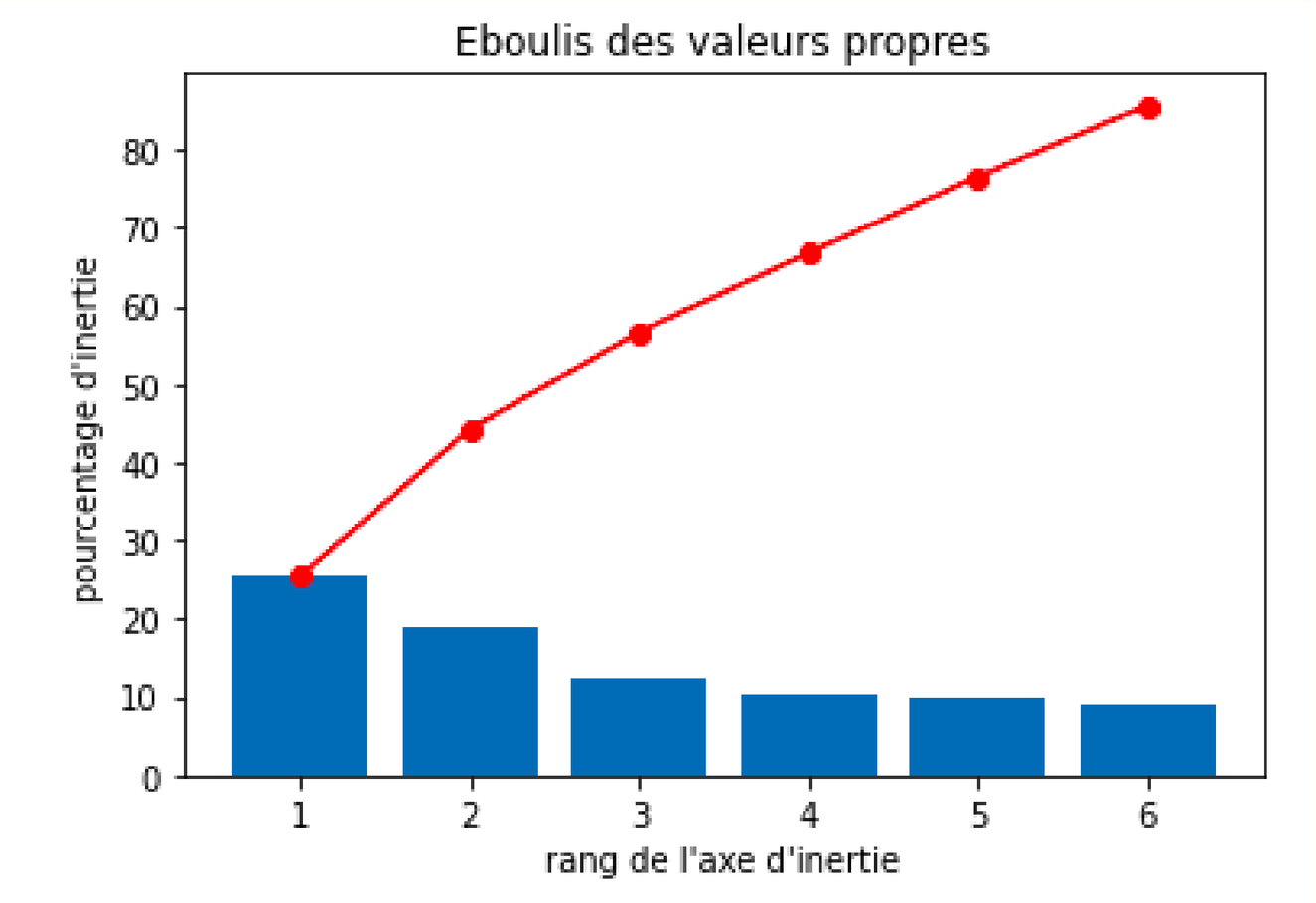
Pvalue de 0 suite au test de Kruskal-Wallis, on peut donc conclure qu'au moins un groupe ne provient pas de la même population que les autres





La Pvalue est de 0.0, elle est inférieure au seuil de 1%,les variables nutriscore et pnns group 1 ne sont pas indépendantes

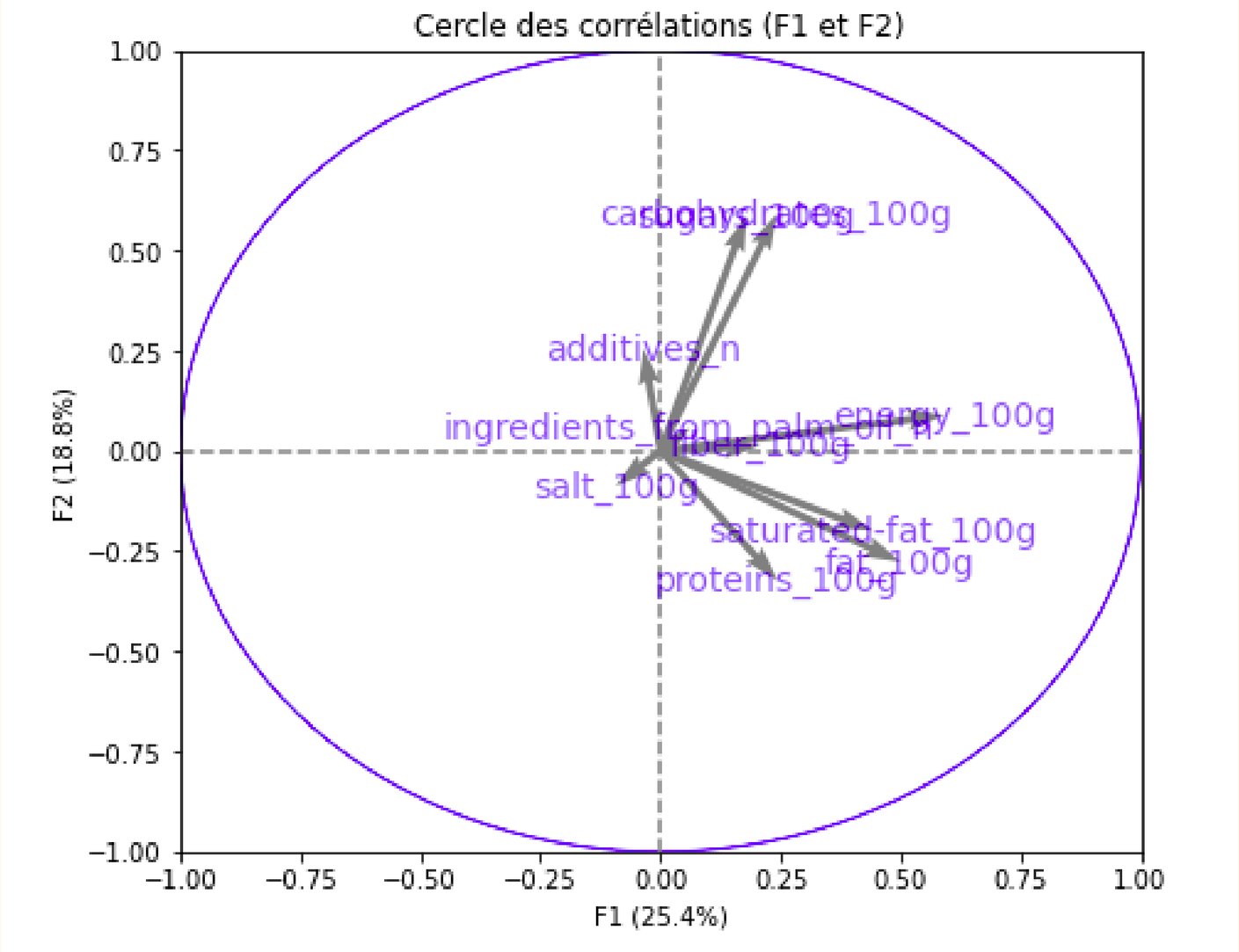
Analyse multivariée, ACP



Nombre de composantes à calculer: 6

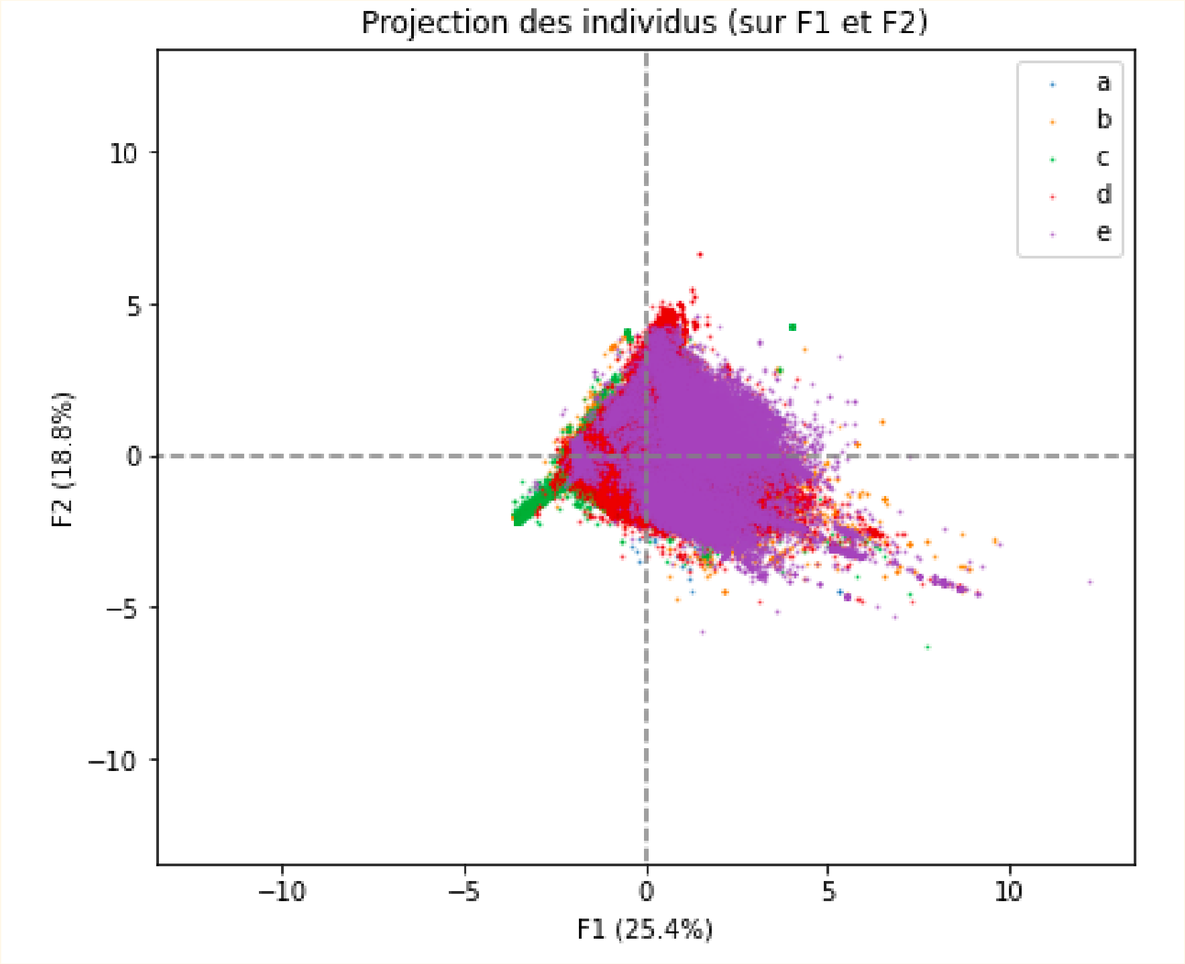
Les deux premiers axes F1 et F2 représentent 45% de la variance.

Nous allons donc étudier uniquement ces deux axes.

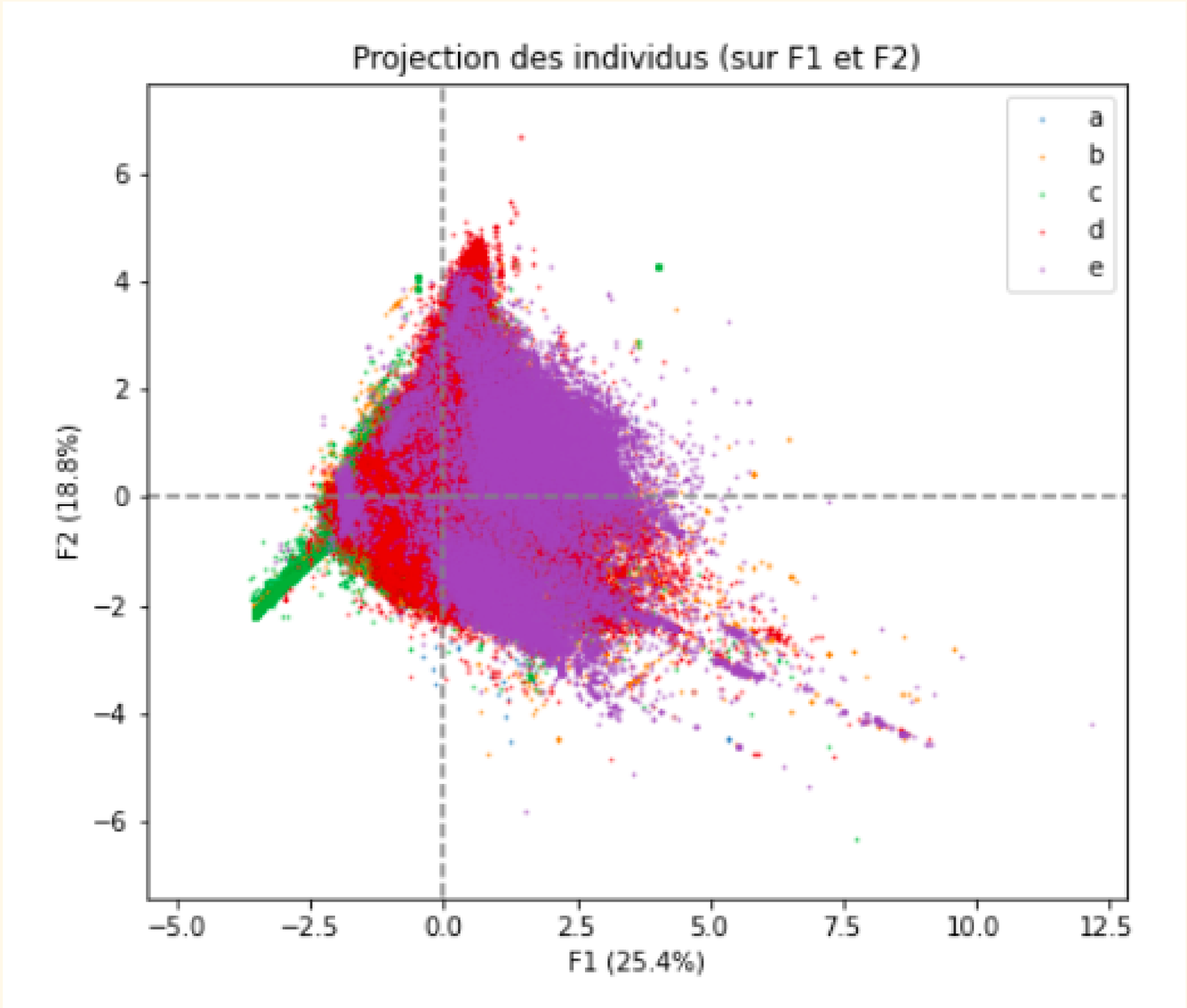


On peut interpréter F1 comme une variable synthétique des variables energie, lipides et acides gras saturés.

On peut assimiler F2 aux variables glucides et sucres.



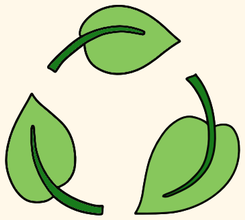
Zoomons davantage


Projection des individus (sur F1 et F2)
a
b
c
d
e
6
4
2
0
−2
−4
−6
F2 (18.8%)
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
12.5
F1 (25.4%)

Conclusion

Les produits que nous consommons, qui font partie de notre alimentation, sont une composante principale de notre santé.
Il convient donc, pour des problématiques de santé publiques, de s'efforcer de mieux consommer.

Malgré un jeu de données peu remplis, j'estime que les données à disposition nous permettent de justifier la faisabilité de notre application HFC, d'un point de vue métier et d'un point de vue des besoins.



En effet, selon un sondage Opinion Way (réalisé en octobre 2020), 61 % des Français déclarent acheter le plus souvent possible des produits fabriqués en France depuis le début de la pandémie et 64 % estiment avoir augmenté leur consommation. Les raisons principales évoquées : la volonté de soutenir les producteurs locaux et les entreprises françaises, de préserver le tissu économique national, et aussi de diminuer la pollution liée aux transports.

Thank you

Merci de votre attention