Soutenance Projet OC P2 ML:

Concevez une application au service de la santé publique

Par Wael Iskandar
le 10/02/2023

La mission

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

- Je souhaite y participer et proposer une **idée d'application**.

- Nous avons à notre disposition la **base de données Open Food Facts** pour l'application.

L’idée d’application

Aider les consommateurs à choisir un meilleure produit

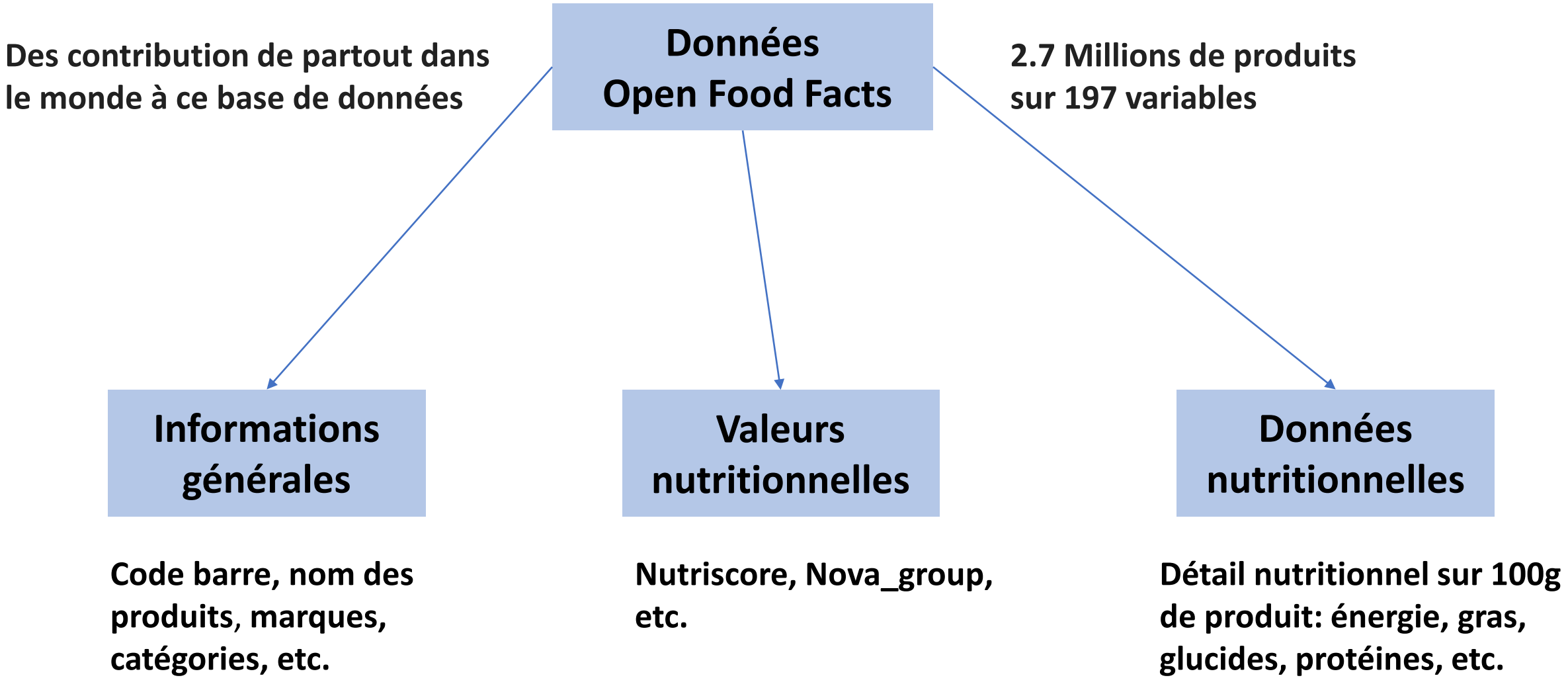
L'application a pour objectif de recommander un produit à l’utilisateur en se basant sur une mesure nutri-grade qui indiquera la qualité nutritive de l'article.

1. Le consommateur **scanne le code barre** du **produit.**

2. L’application cherche le produit dans une base de données de **>1 millions de produits**.

3. L'application propose un **produit similaire** avec une **meilleure note nutri-grade** et un **meilleure score nutri-score.**

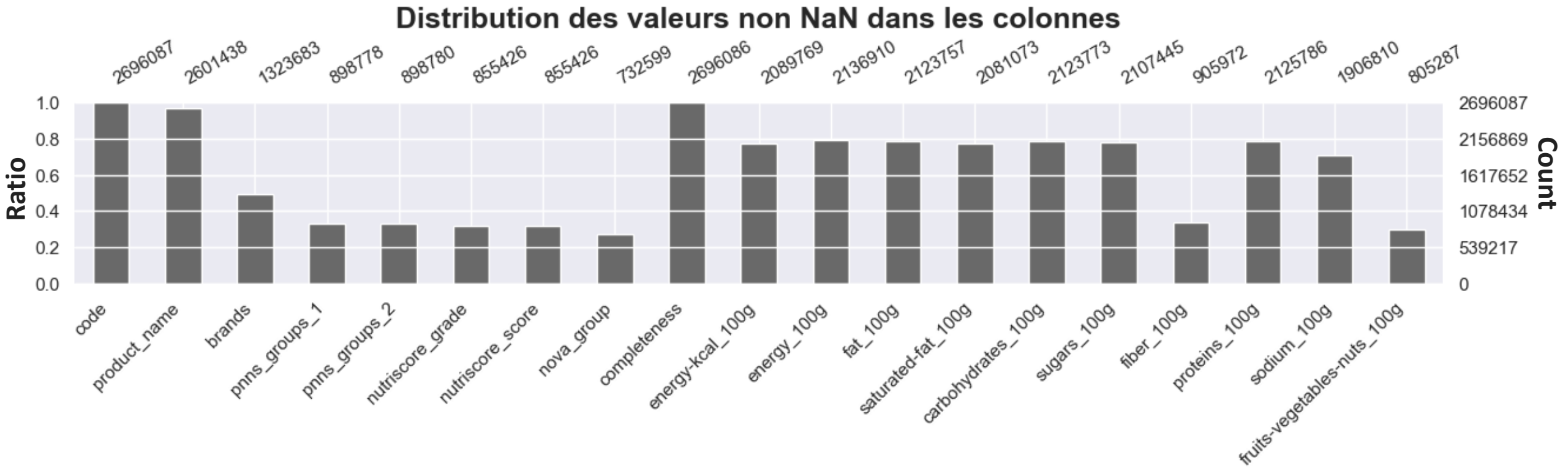
Description du base de données OpenFoodFacts



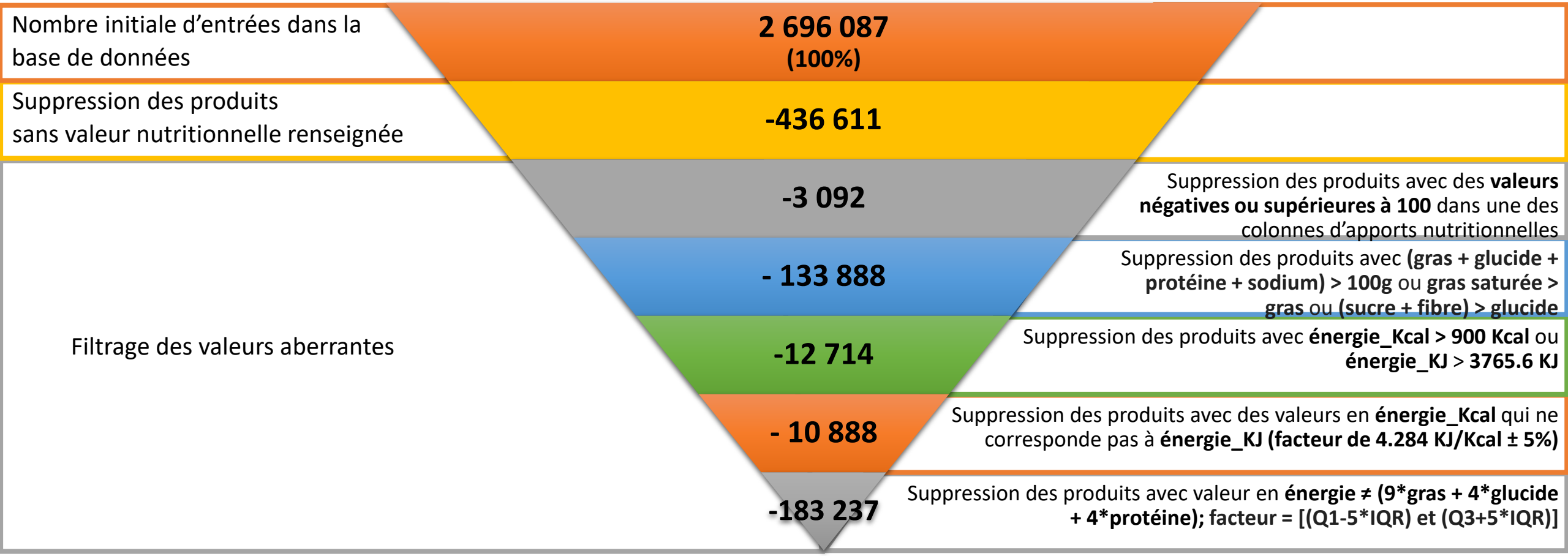
Les opérations de nettoyage effectuées pour l'application:
19 colonnes conservés

- Supprimer les variables selon le taux de remplissage et pertinence:
  - 37 colonnes supprimés remplie que par 0 ou valeurs manquantes NaN.
  - 94 colonnes supprimés remplie au minimum à 80% par des valeurs manquantes NaN.
  - Parmi les 66 variables restantes, sélection de 19 colonnes qui seront pertinentes pour l'application.

En moyenne, **38%** de valeur manquante **NaN** dans la base de données.



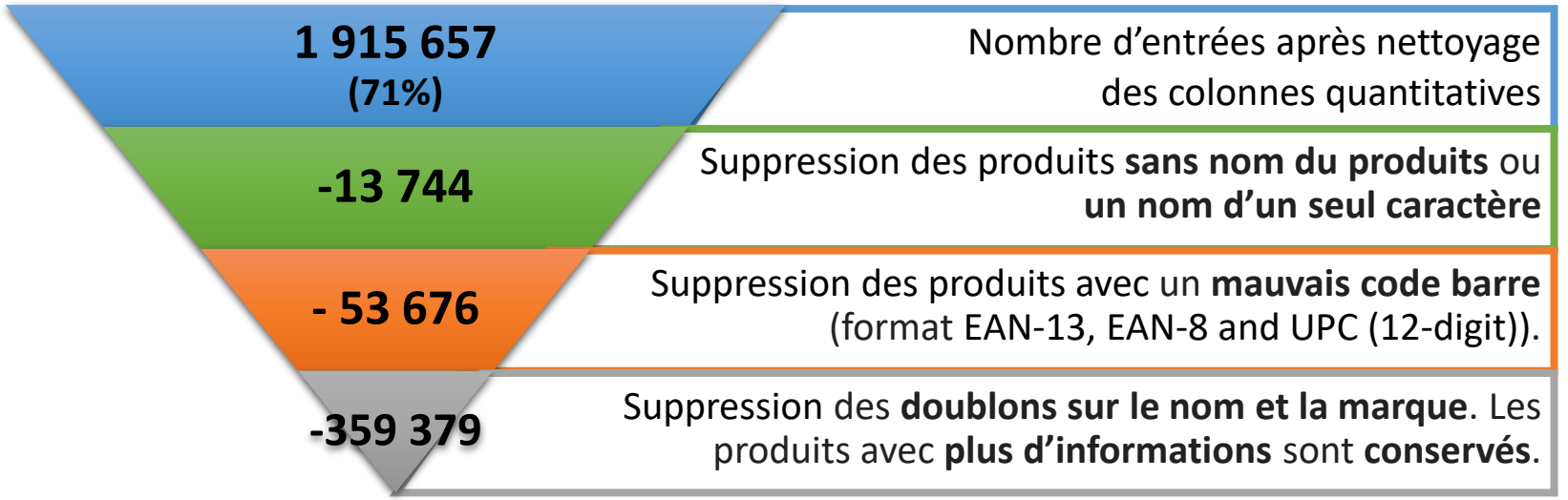
Les opérations de nettoyage effectuées pour l'application:
780k produits supprimés (colonnes quantitatives)



Note: **0** produit supprimé avec un mauvais **nutri-score** (en dehors de **[-15; 40]**) et **nova_group score** (en dehors de **[1; 4]**)

- Totale des produits supprimés sur les colonnes quantitatives = 780k (-29% du 2.7M = 1.9M produits restants).

Les opérations de nettoyage effectuées pour l'application:
427k produits supprimés (colonnes qualitatives)

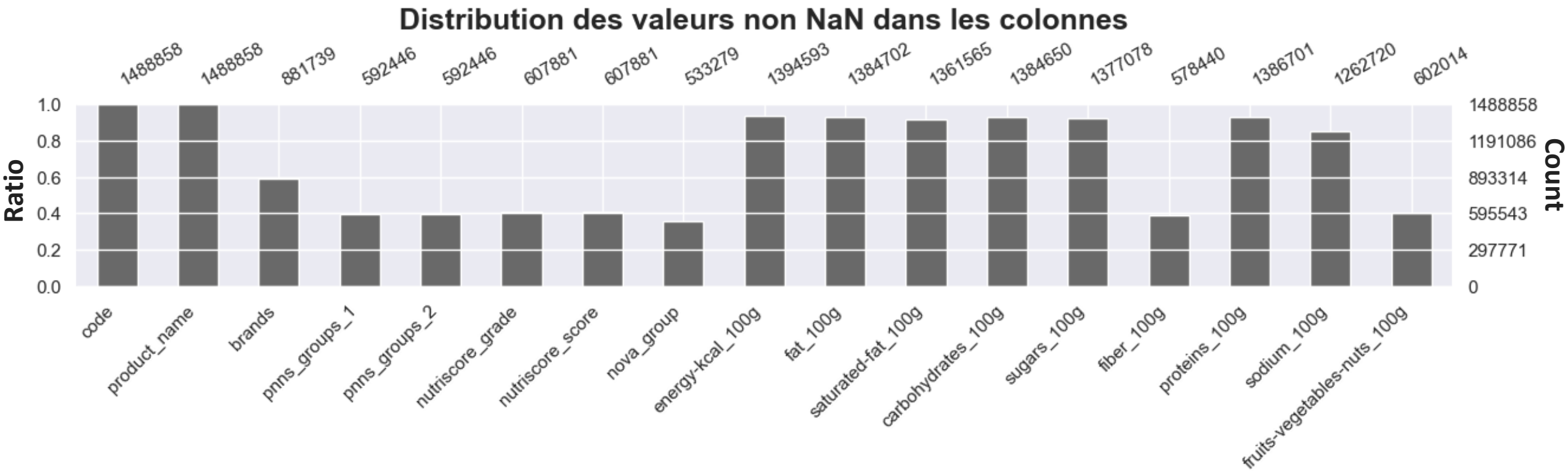


Note: **0** produit supprimé avec un mauvais **nutri-grade** (en dehors des **grades a, b, c, d, et e**):

- Totale des produits supprimés sur les colonnes qualitatives = 427k (-22% du 1.9M = 1.5M produits restants).

Les opérations de nettoyage effectuées pour l'application:

Moins de valeurs manquantes qu'avant le traitement



En moyenne, **31%** de valeur manquante **NaN** dans la base de données.

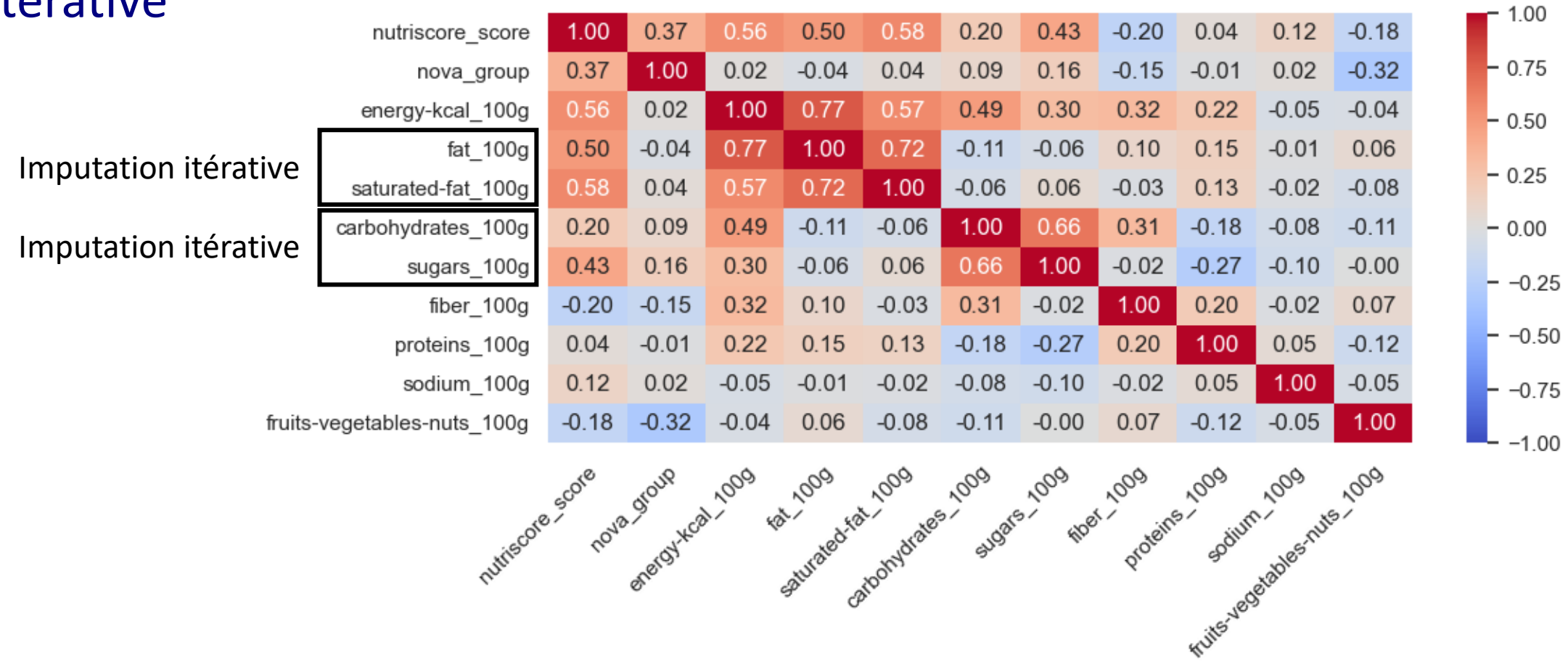
Les opérations de nettoyage effectuées pour l'application:

Traitement des valeurs manquantes (colonnes quantitatives)

Imputation itérative

Imputation itérative

Imputation itérative



- Correlation forte entre fat et energy, entre saturated-fat et fat et entre carbohydrates et sugars:
  - Imputation itérative sur saturated-fat et fat.
  - Imputation itérative sur carbohydrates et sugars.
  - Le traitement de l'énergie pour après.

Les opérations de nettoyage effectuées pour l'application:

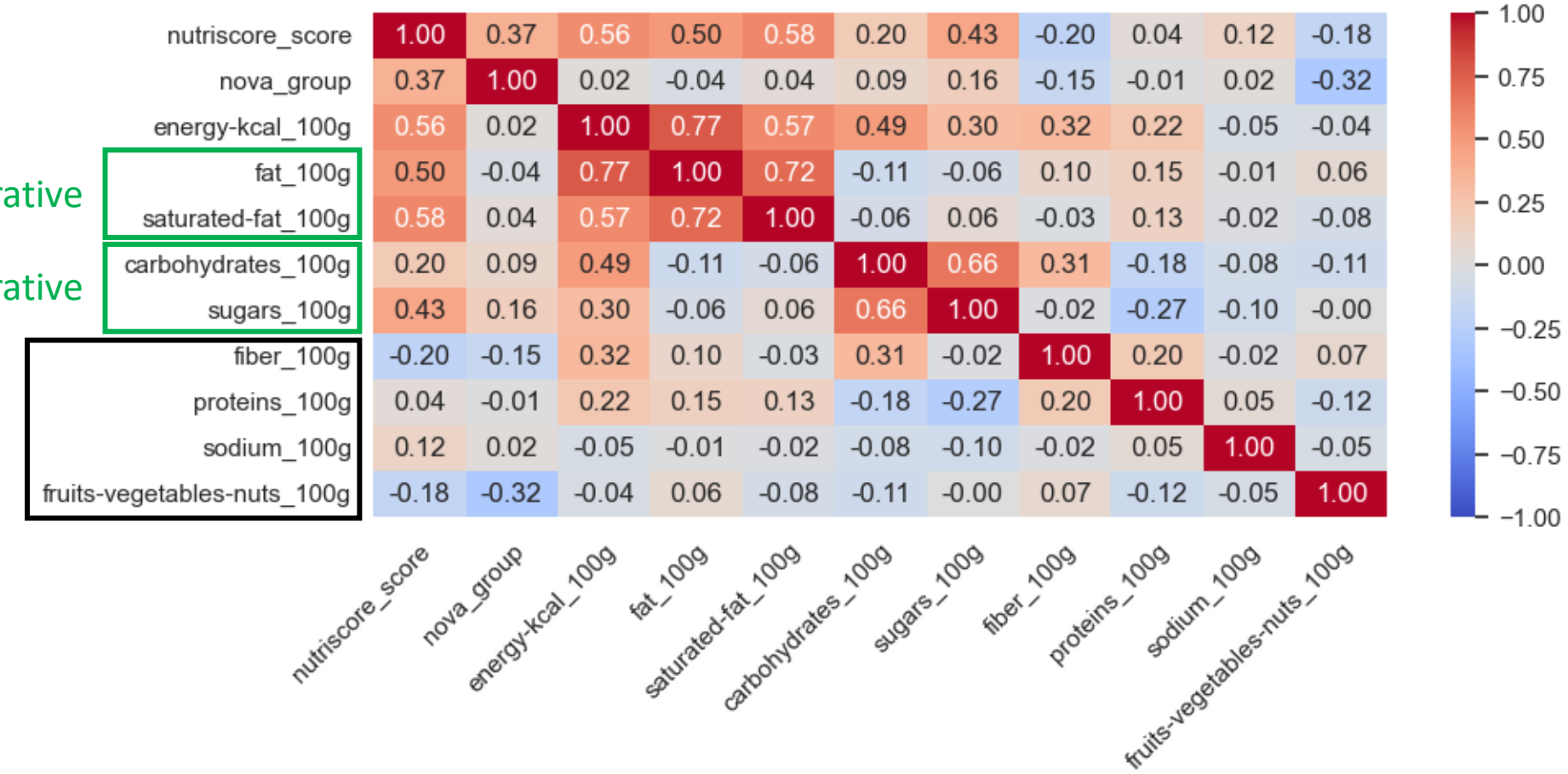
Traitement des valeurs manquantes (colonnes quantitatives)

Médiane

✓ Imputation itérative

✓ Imputation itérative

Médiane



- Pour les colonnes fibre, protéines, sodium, et fruits-vegetables-nuts:
  - Remplissage par la valeur médiane en fonction de la catégorie pnns_group_2 du produit.
  - Pour les produits avec la catégorie du produit pnns_group_2 non renseigné, remplissage par la valeur médiane totale.

Les opérations de nettoyage effectuées pour l'application:

Traitement des valeurs manquantes (colonnes quantitatives)

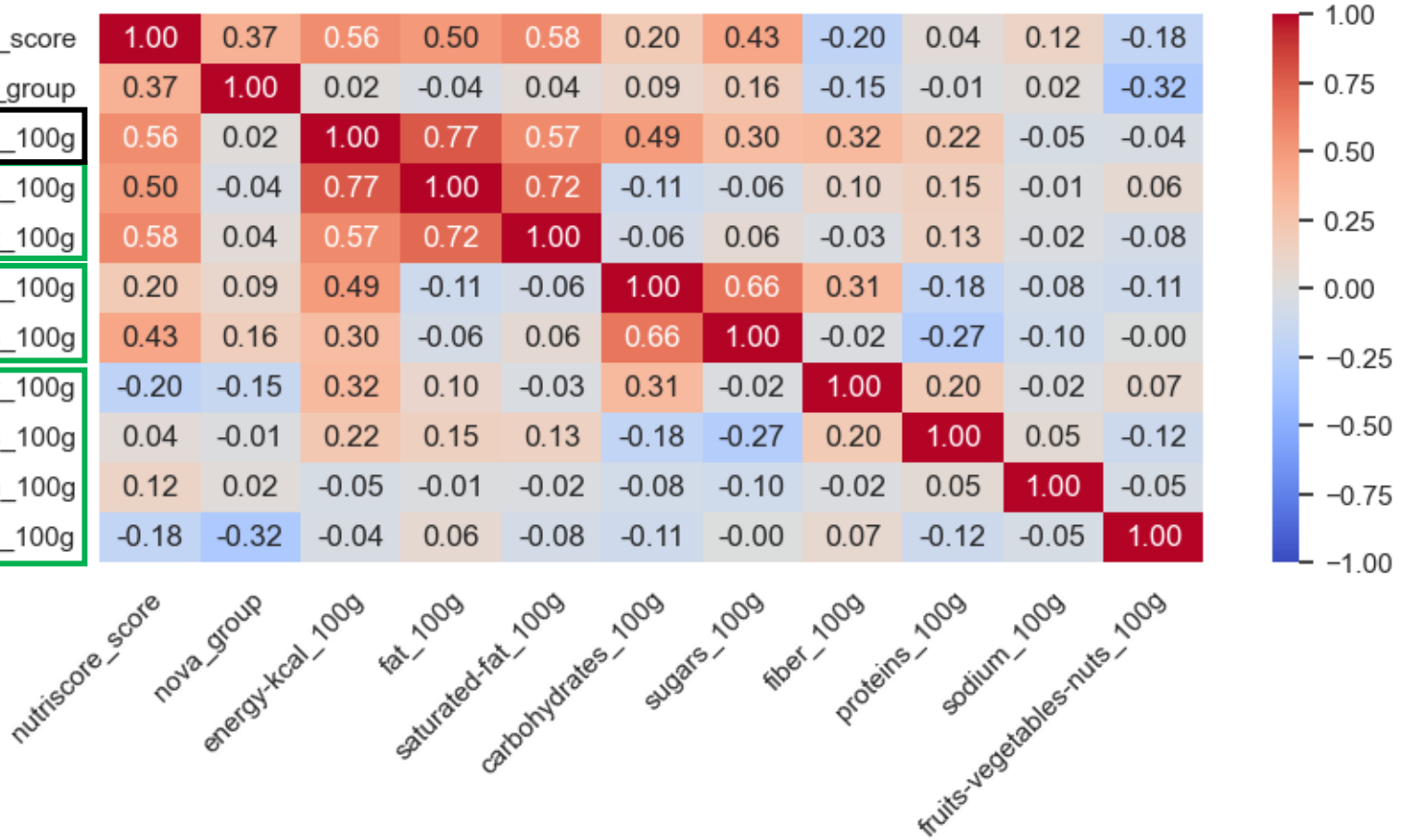
Calcule

Calculé

✓ Imputation itérative

✓ Imputation itérative

✓ Médiane



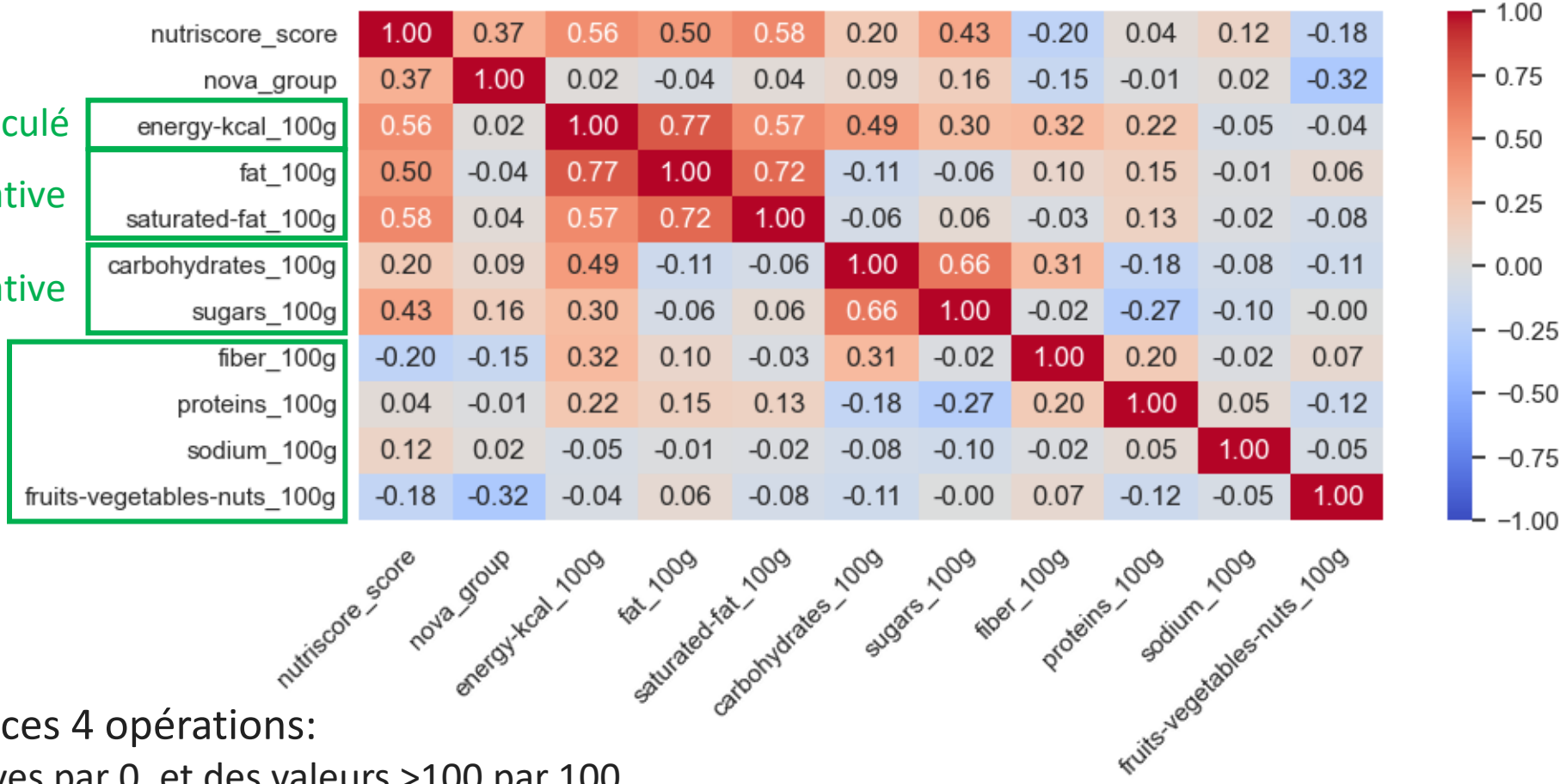
- Pour la colonne énergie, calcule de l'énergie en fonction du fat (9Kcal/g), carbohydrates (4Kcal/g), et proteins (4Kcal/g): **énergie = (9 * gras + 4 * glucide + 4 * protéine)**

Les opérations de nettoyage effectuées pour l'application:

Traitement des valeurs manquantes (colonnes quantitatives)

Nettoyage

- ✓ Calculé
- ✓ Imputation itérative
- ✓ Imputation itérative
- ✓ Médiane



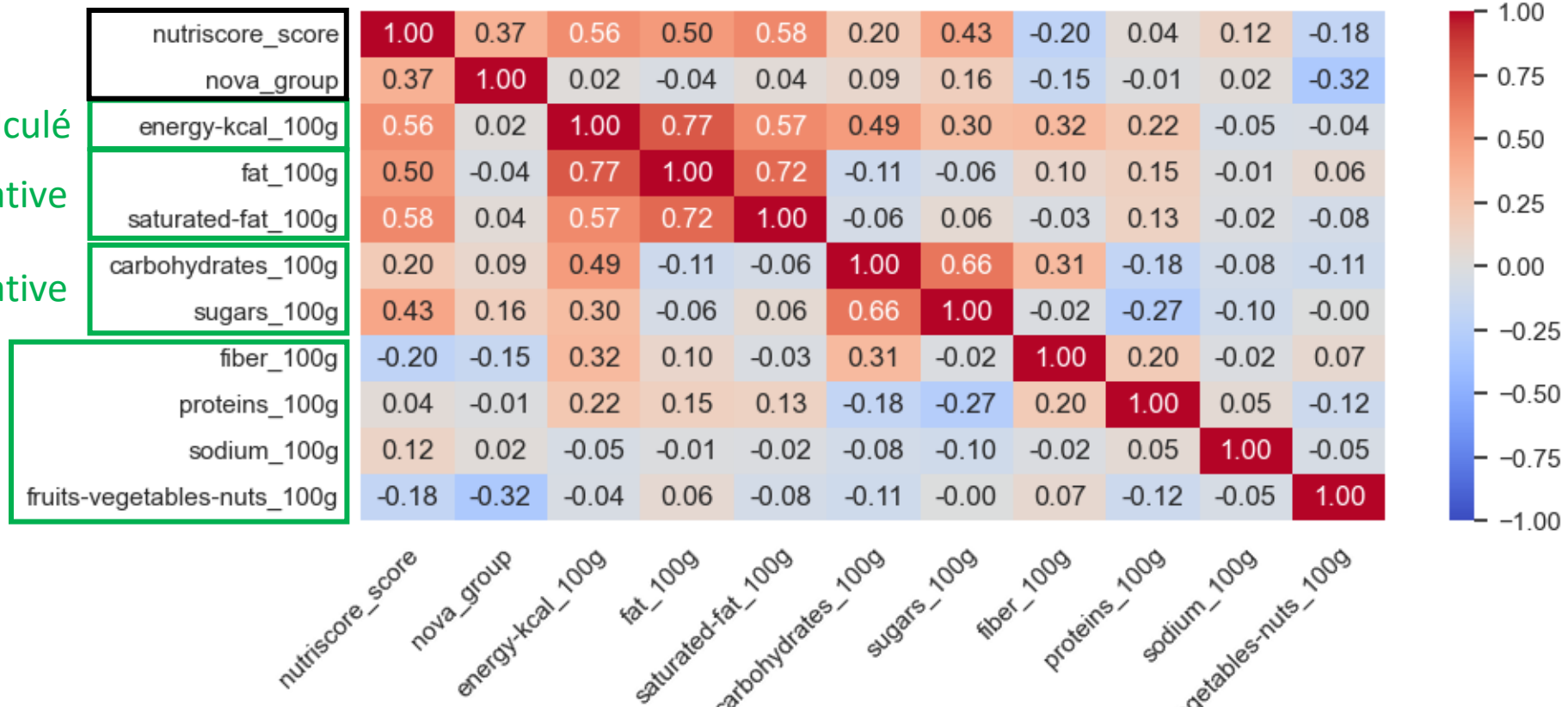
- Traitement des erreurs due à ces 4 opérations:
  - Remplacement des négatives par 0, et des valeurs >100 par 100.
  - Suppression des produits avec la somme de fat, carbohydrate, et proteins > 100.
  - Si saturated_fat > fat, remplacement du valeur de saturated_fat par fat.
  - Si sugars > carbohydrates, remplacement du valeur du sugars par carbohydrates.
  - Si (sugars + fiber) > carbohydartes, remplacement du valeur de fiber = (carbohydrate – sugars).

Les opérations de nettoyage effectuées pour l'application:

Traitement des valeurs manquantes (colonnes quantitatives)

RandomForestClassifier

- ✓ Calculé
- ✓ Imputation itérative
- ✓ Imputation itérative
- ✓ Médiane



- Pour nutri-score et nova_group, j'ai entrainé un modèle basé sur RandomForestClassifier:
  - Sélection des données non-NaN sur nutri-score et division des données en 80% train et 20% test.
  - J'entraine le modèle en utilisant RandomForestClassifier, puis je test le modèle et j'obtiens une erreur de 0.13.
  - J'applique le modèle sur les produits qui ont des valeurs NaN sur nutri-score.
  - Je répète la même démarche pour nova_group (l'erreur obtenu du modèle cette fois ci est de 0.1).
  - Je nettoie ces deux colonnes des valeurs invalide (en dehors de [-15;40] pour nutri-score et [1;4] pour nova-group).

Les opérations de nettoyage effectuées pour l'application:

Traitement des valeurs manquantes (colonnes qualitatives)

RandomForestClassifier

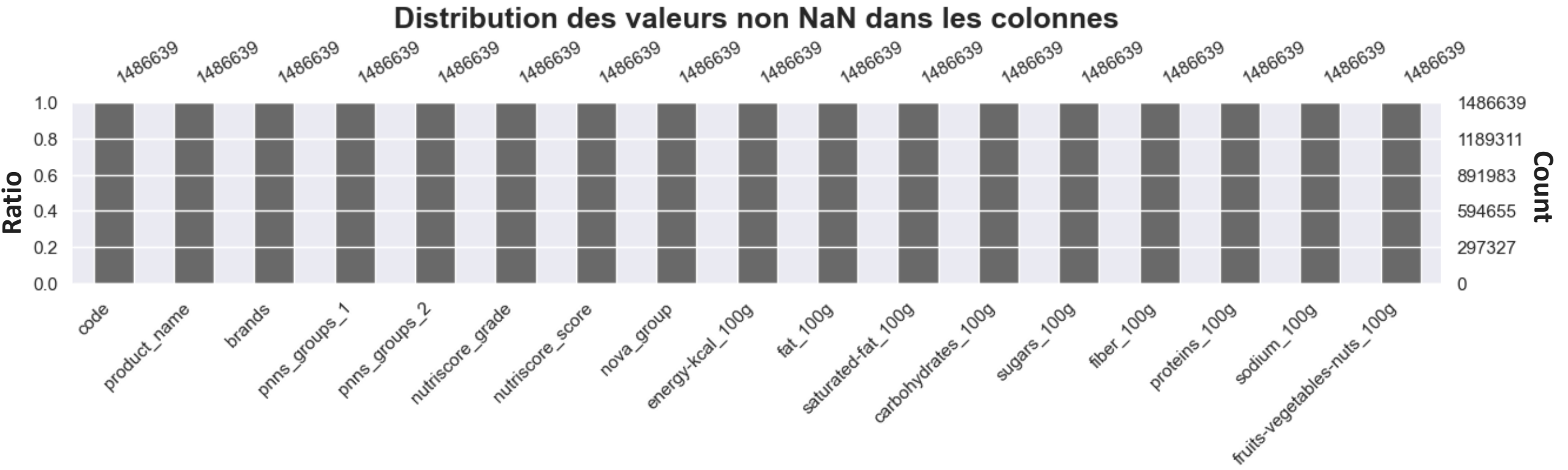
Classifier

	product_name	brands	pnns_groups_1	pnns_groups_2	nutriscore_grade
type	object	object	object	object	object
number null	0.0	606594.0	894833.0	894833.0	878790.0
percentage null	0.0	40.803046	60.191681	60.191681	59.112535
count	1486639	880045	591806	591806	607849
unique	1204765	164869	10	39	5
top	Spaghetti	Carrefour	Sugary snacks	Biscuits and cakes	d
freq	392	10809	129187	57958	196464

- Pour nutri-grade, pnns_group_1 et pnns_group_2, j'ai entrainé un modèle basé aussi sur RandomForestClassifier
  - Sélection des données non-NaN sur nutri-grade et division des données en 80% train et 20% test.
  - Je transforme les valeurs catégorielle en valeur numérique.
  - J'entraine le modèle en utilisant RandomForestClassifier, puis je test le modèle et j'obtiens une erreur de 0.002.
  - J'applique le modèle sur les produits qui ont des valeurs NaN sur nutri-grade.
  - Je répète la même démarche pour pnns_group_1 et _2 (l'erreur obtenu du modèle cette fois-ci est de 0.04).
  - Je remplace les valeurs erronées de pnns_group_2 qui n'appartienne pas à la catégorie pnns_group_1 par unknown.
- Enfin pour la colonne brands, je remplace les NaN par unknown.

Les opérations de nettoyage effectuées pour l'application:

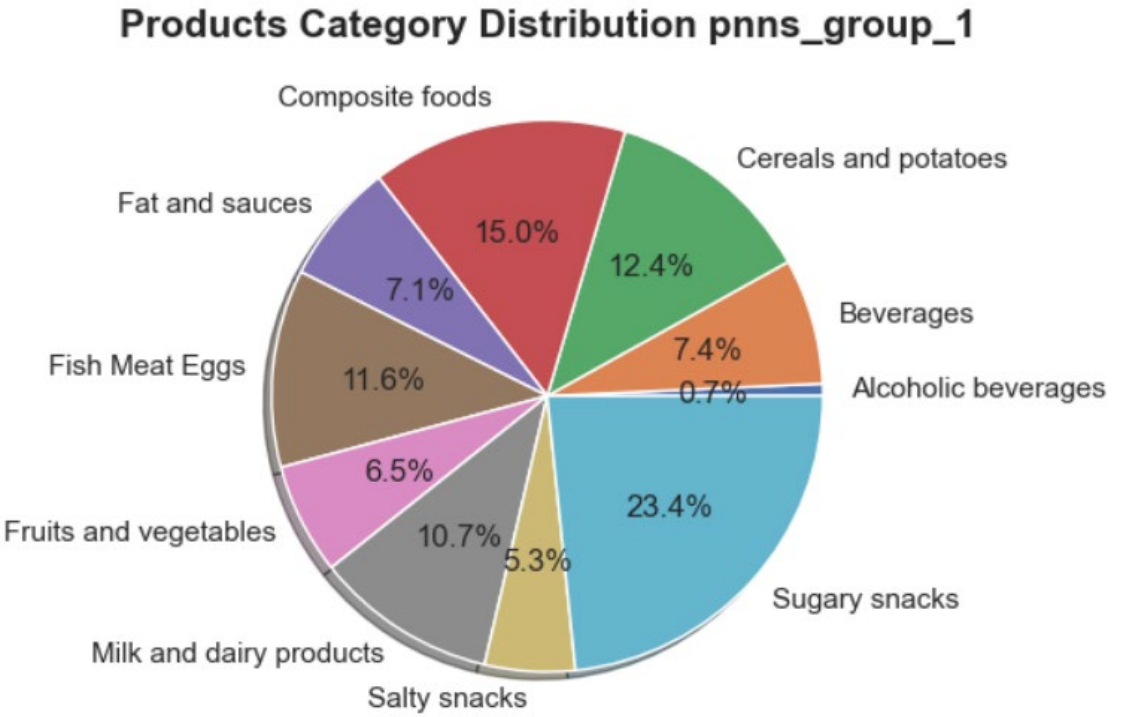
Pas de valeurs manquantes dans la base de données

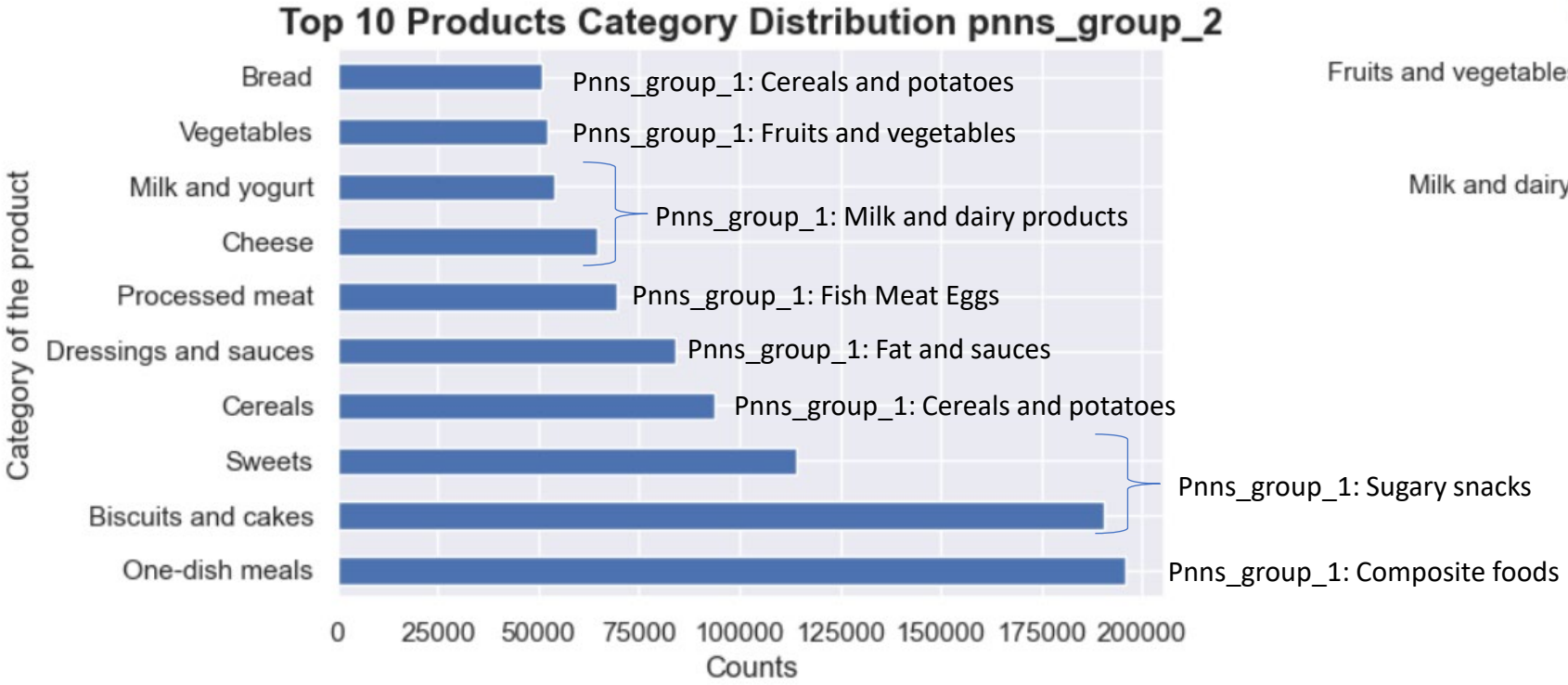


- Base de données final après nettoyage: 1.5M de produits avec 17 variables prêts pour analyse.

Analyse univariée et bivariée:

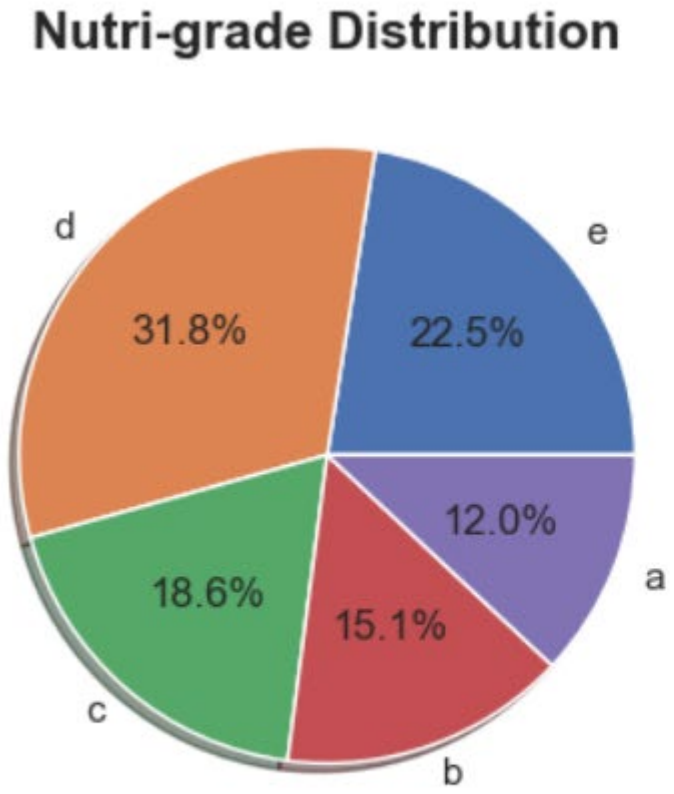
Grande représentation pour la plupart des catégories du produit

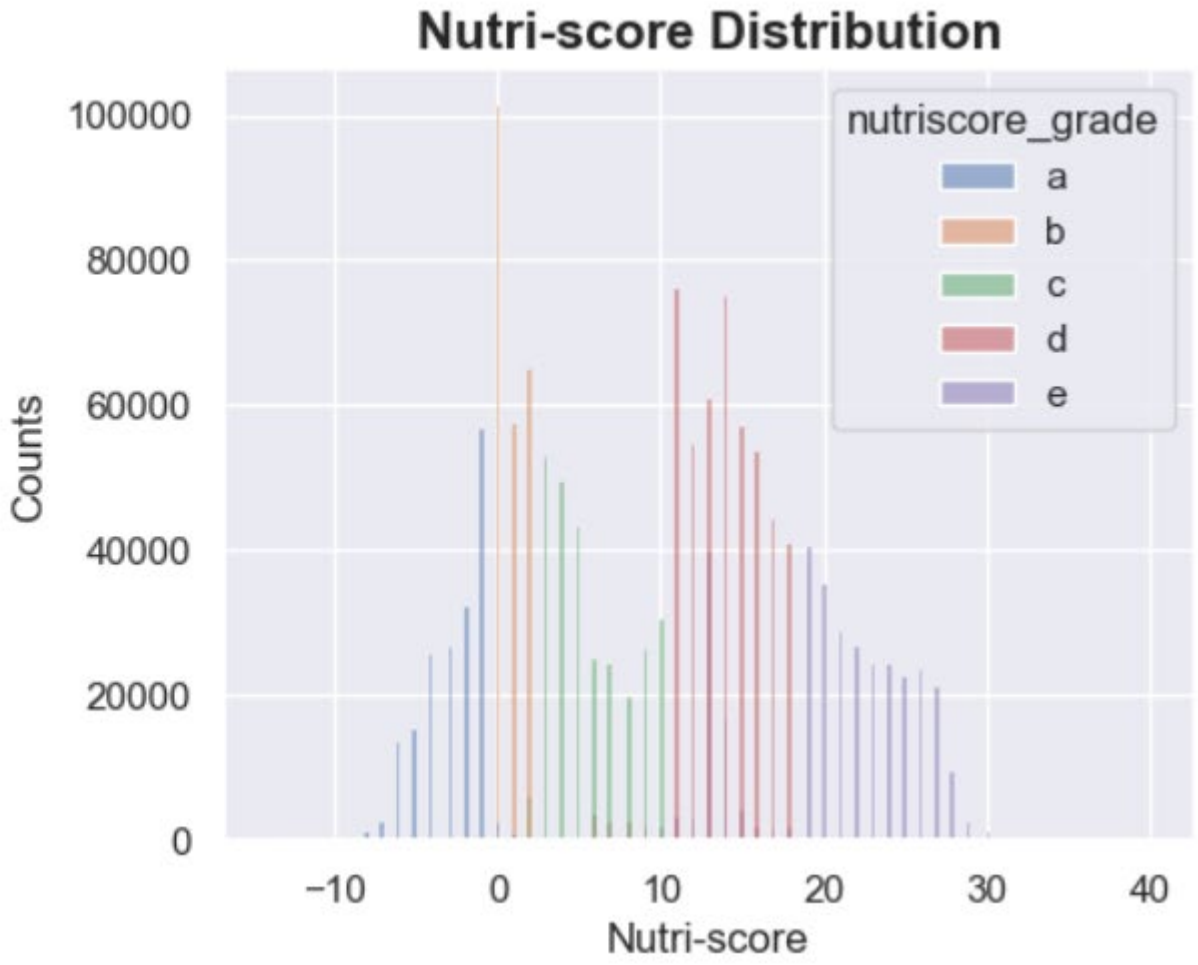




Analyse univariée et bivariée:

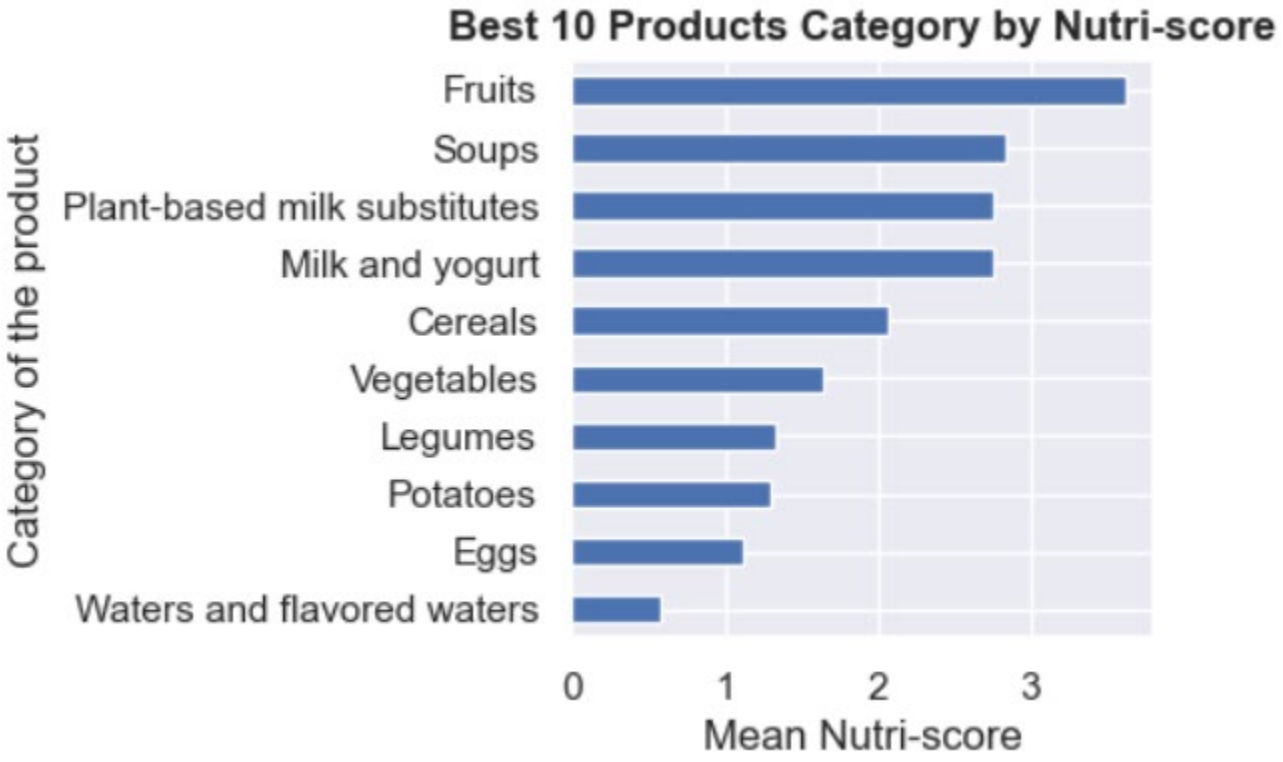
Forte corrélation entre nutri-score et nutri-grade



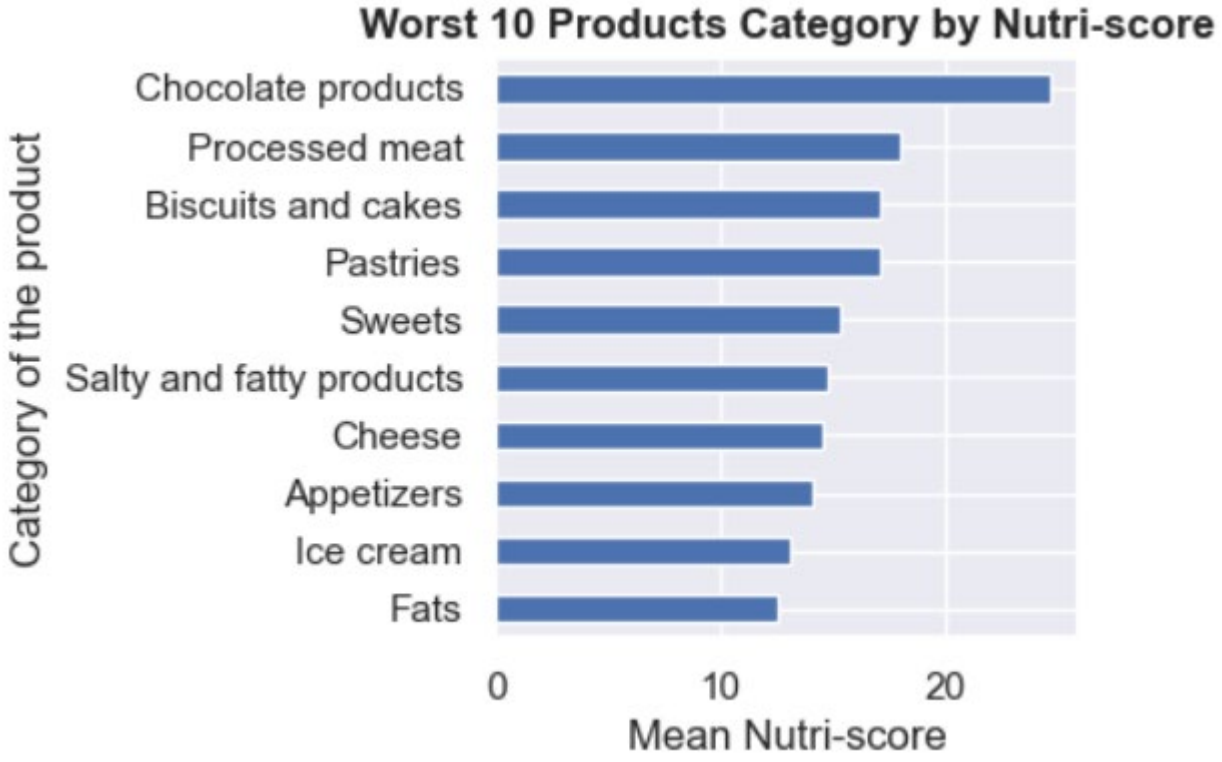


Analyse univariée et bivariée:

Bonne valeur nutritionnelle pour légumes, œufs, céréales, lait, yogourt, soupes, lait à base de plantes, et fruits.

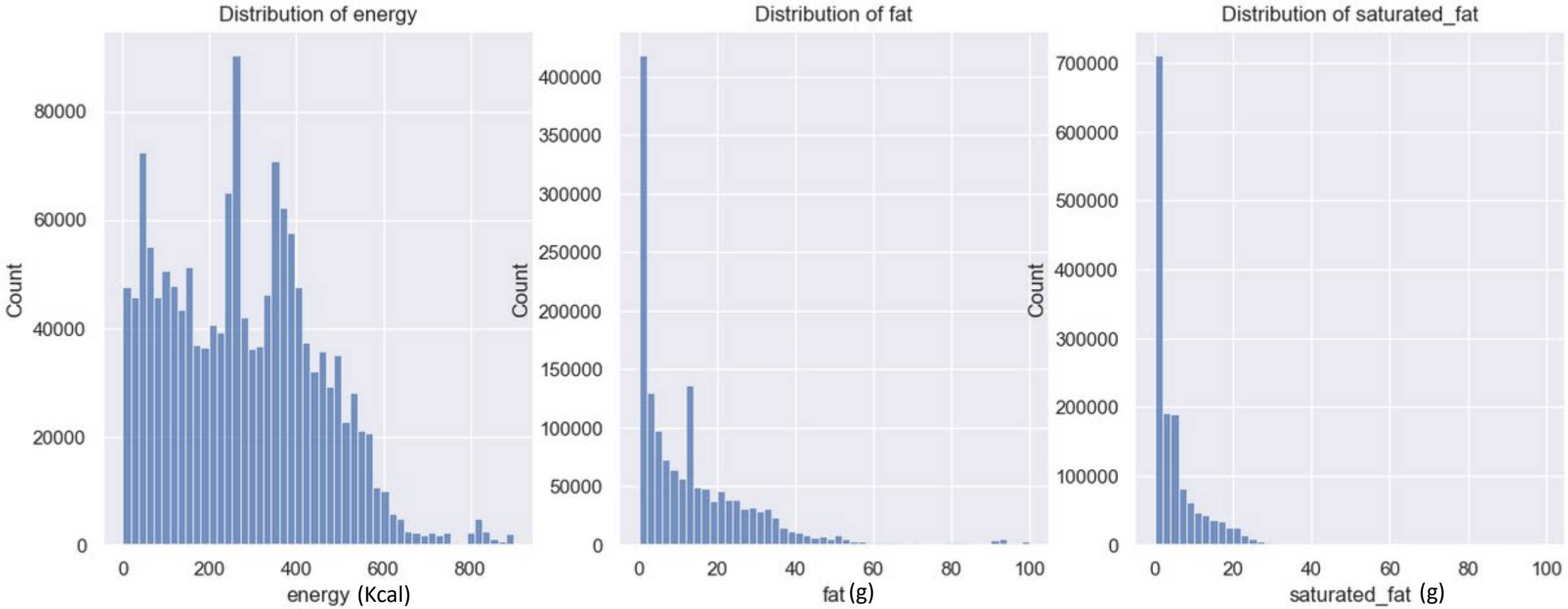


Mauvaise valeur nutritionnelle pour chocolats, viande transformée, biscuits, gâteaux, pâtisseries, snacks sucrées

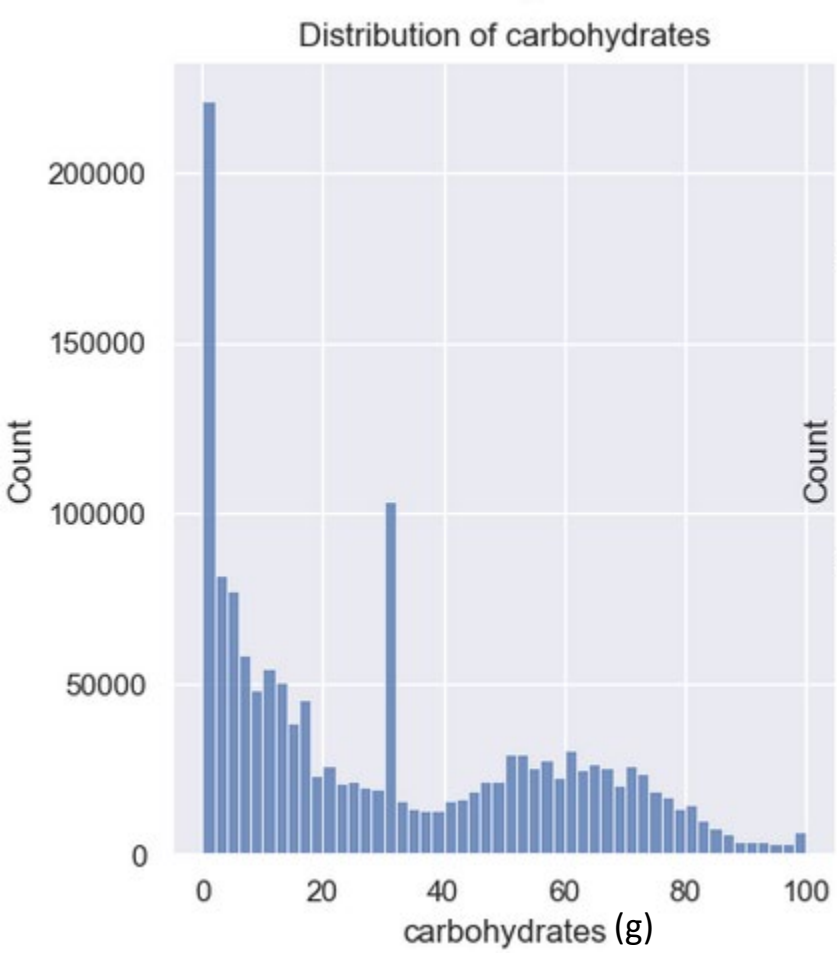


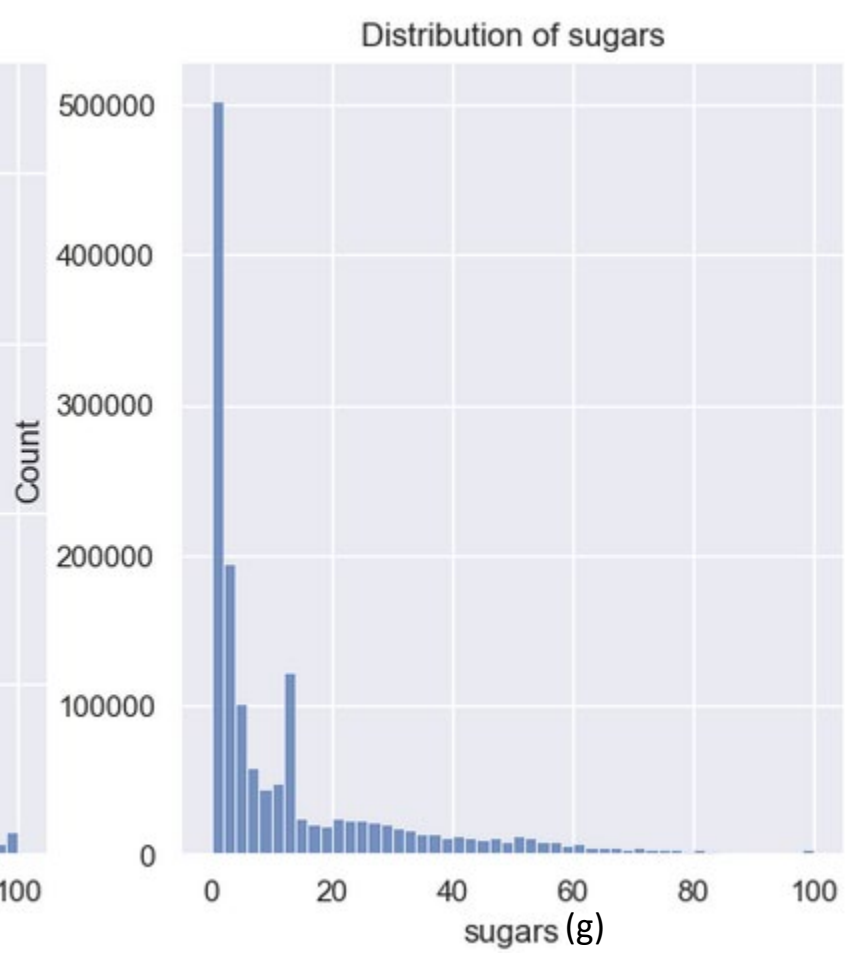
Analyse univariée et bivariée:

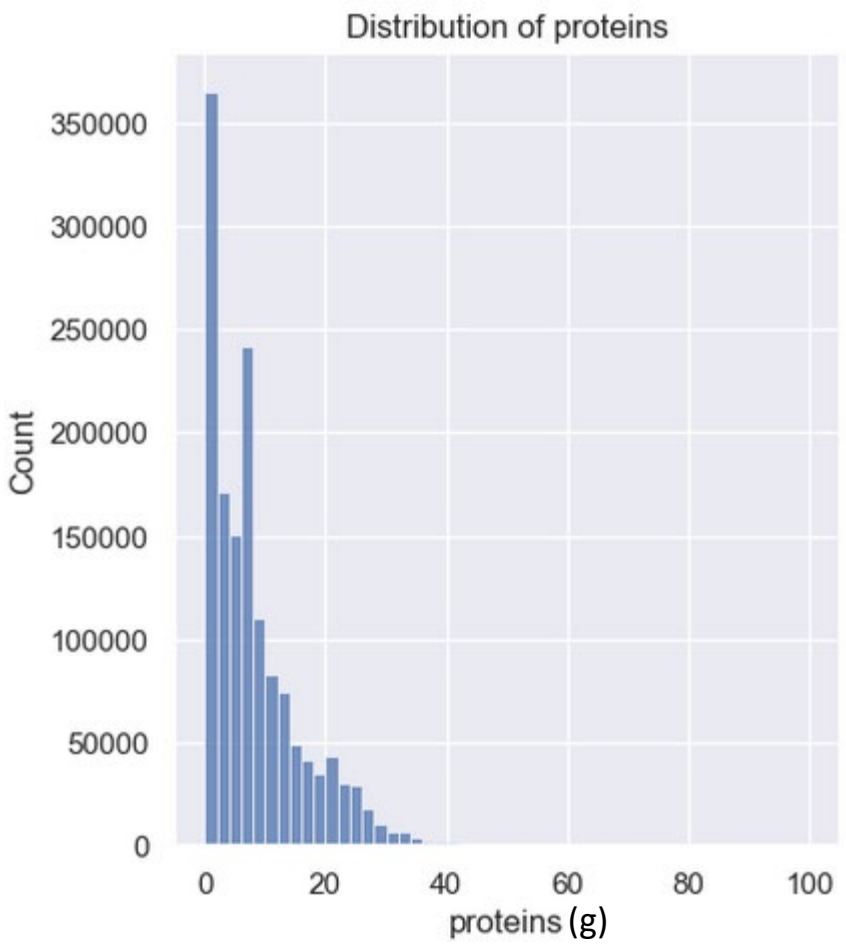
Plupart des produits avec énergie <500Kcal



Analyse univariée et bivariée:
Distribution à 2 modes pour glucides

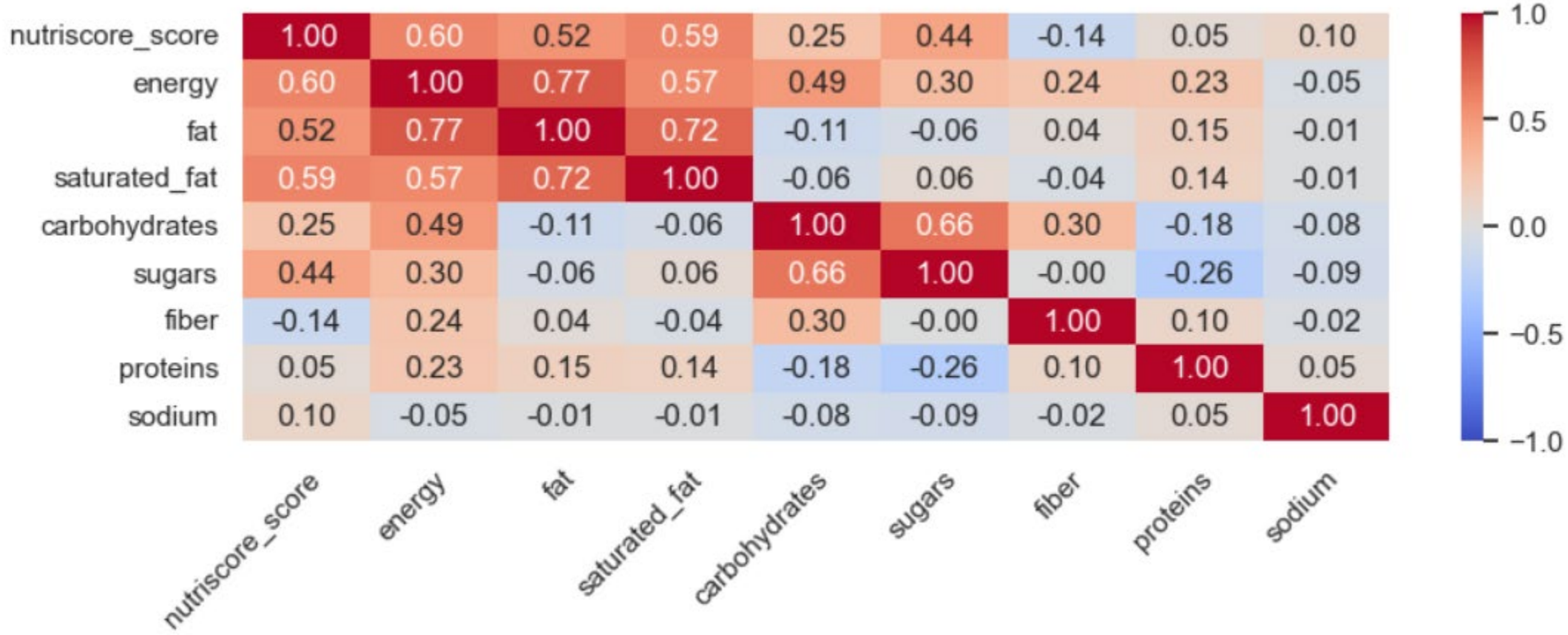






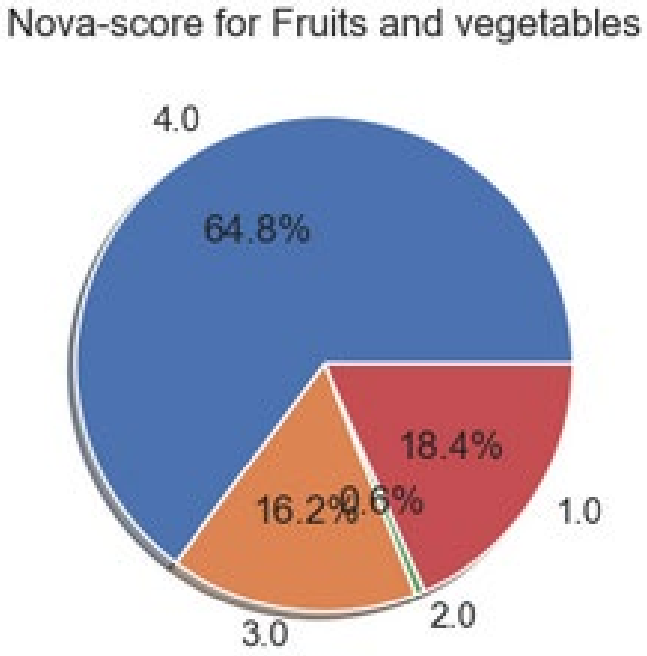
Analyse univariée et bivariée:

Forte corrélation entre nutri-score, énergie, gras et gras saturé

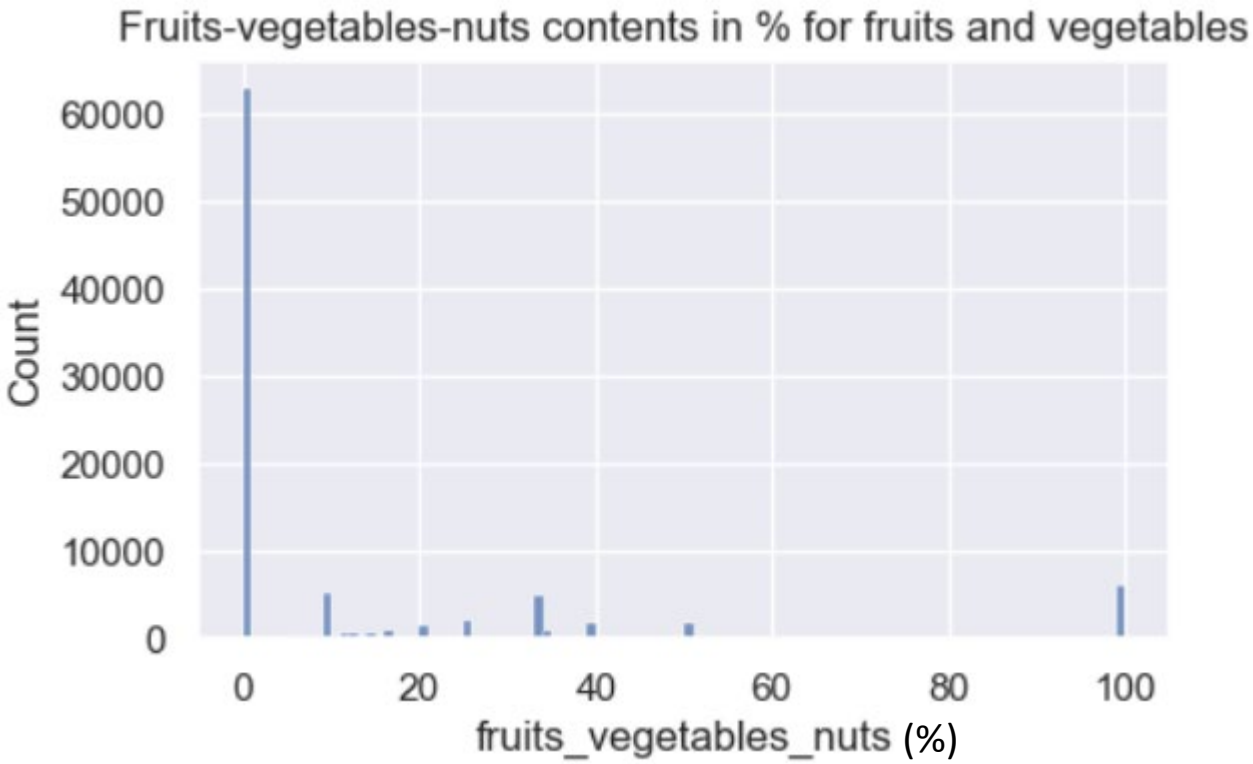


Analyse univariée et bivariée:

Suppression de 2 colonnes avec distribution illogique



Nova-group mesure le degré de transformation des aliments sur une échelle de 1 « minimalement transformé » à 4 « ultra-transformé ».
La colonne Nova-group est supprimée parce que le score ne correspond pas à 1 pour les fruits et légumes. Il n'y a pas de processus de transformation pour les fruits et légumes.



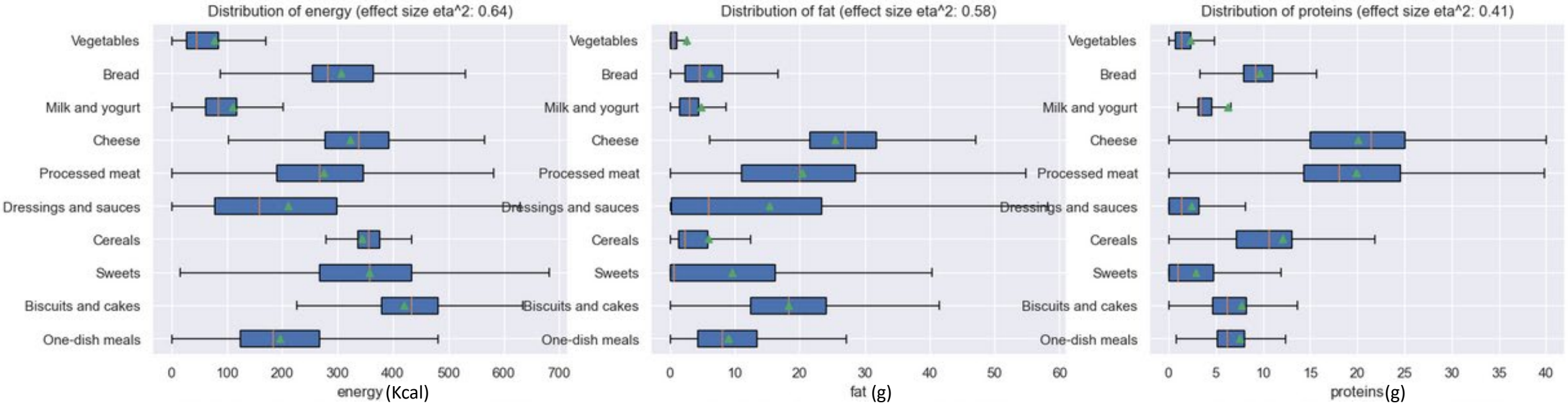
La colonne fruit_vegetables_nuts est supprimée parce que la distribution ne corresponde pas à 100% pour les fruits et légumes.

Analyse univariée et bivariée:

Analyse de variance ANOVA (catégorie du produit)

Peut-on prévoir la **catégorie** du produit à partir des **valeurs nutritionnelles**?

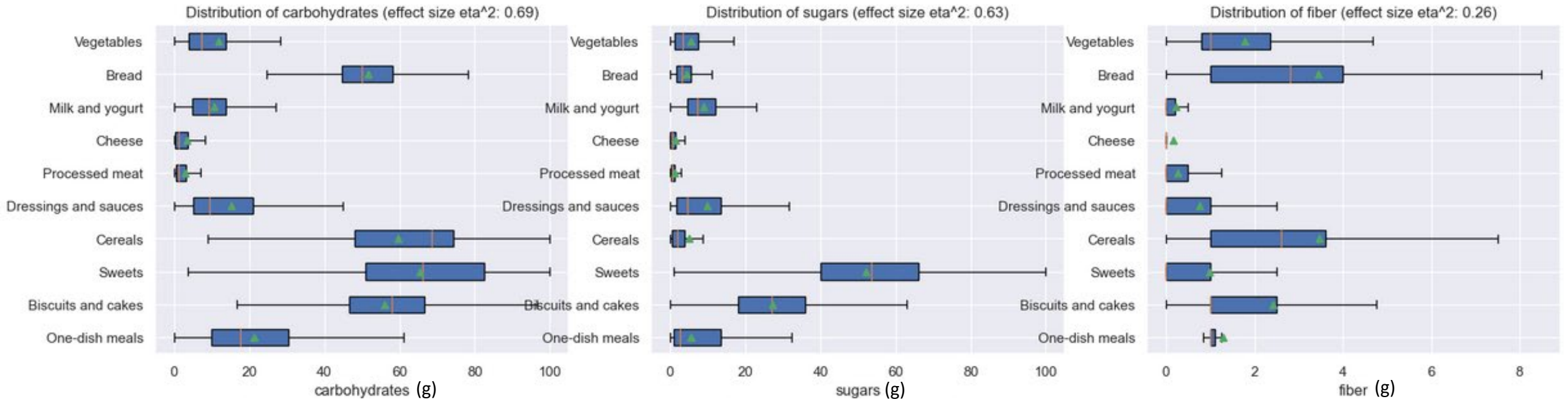
Analyse univariée et bivariée:
Analyse de variance ANOVA (catégorie du produit)



- Les biscuits, gâteaux, et produits sucrées sont denses en énergie, et en gras.
- Les fromages et viande transformée sont denses en énergie, gras et protéines.
- Les légumes sont les plus faibles en énergie, gras et protéines.

- En terme de l'effet de l'énergie et le gras sur la catégorie du produits, on mesure un effet significative avec η^2 = 0.64 et 0.58 respectivement qui est plus grande que l'effet observé avec les protéines.

Analyse univariée et bivariée:
ANOVA (catégorie du produit)



- Les biscuits, gâteaux, et produits sucrées sont denses en glucides, et en sucres.
- Les céréales et pains sont denses en glucides et en fibres mais faible en sucres.
- Les fromages et viande transformée sont faibles en glucides, sucres, et fibres.

- En terme de l'effet de la glucide et sucres sur la catégorie des produit, on mesure un effet bien significative avec $\eta^2 = 0.69$ et 0.63 respectivement qui est un effet proche de l'énergie et gras mais bien plus grande que l'effet observé pour le fibre.

Analyse univariée et bivariée:
ANOVA (catégorie du produit) – vérification statistique

H0 : les moyennes sont égales.
H1 : une ou plusieurs moyennes sont inégales.

F = 35250.49
Influence du facteur PR = 0.0
Rejet H0 : PR = 0 (**<0.05**) et **F ≠ 1**

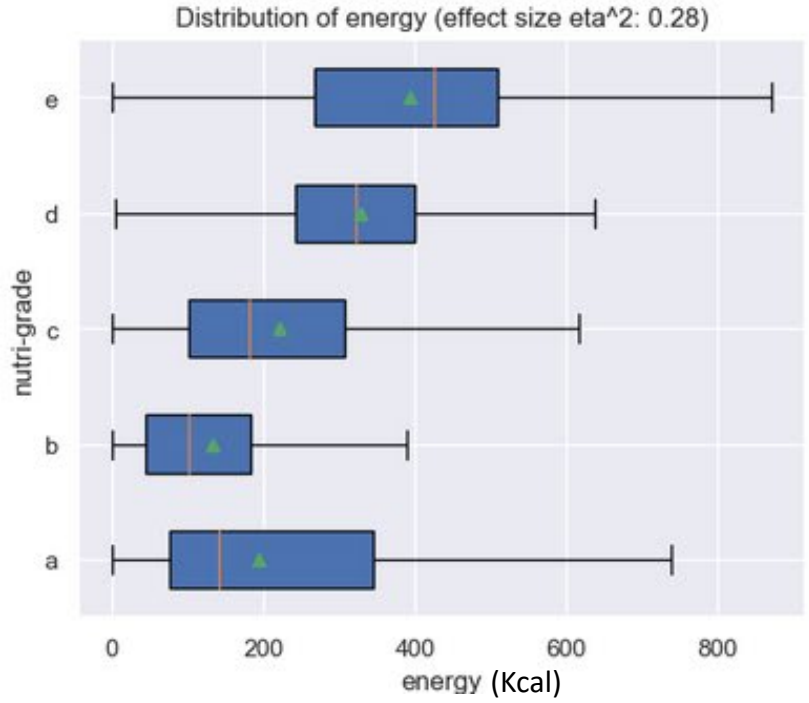
	df	sum_sq	mean_sq	F	PR(>F)
pnns_groups_2	39.0	5.364700e+07	1.375564e+06	35250.499017	0.0
Residual	1486607.0	5.801119e+07	3.902255e+01	NaN	NaN

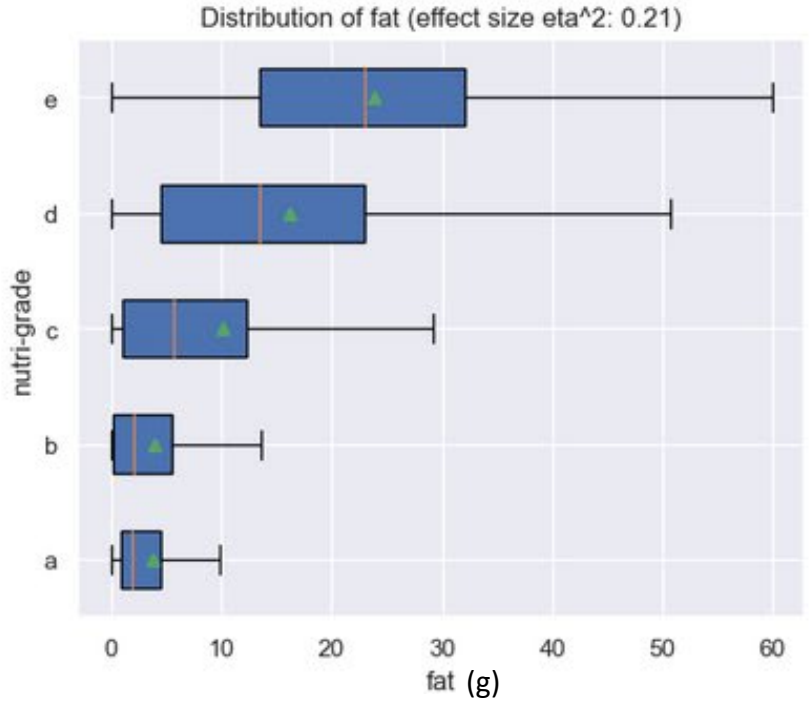
✓ On peux conclure qu'il existe bien un effet significatif entre la catégorie du produit et le nutri-score.

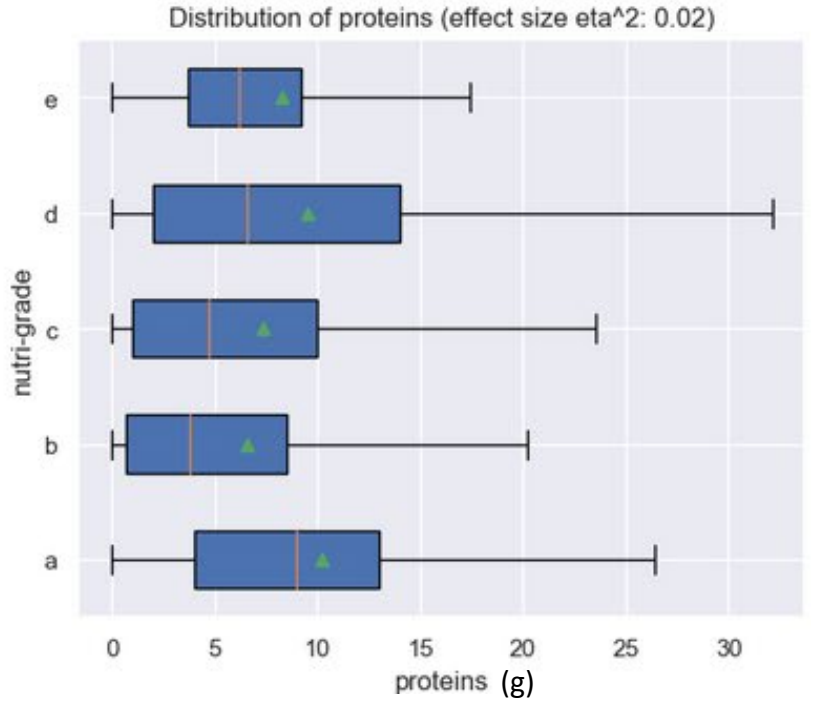
Analyse univariée et bivariée:
ANOVA (Note nutri-grade)

Peut-on prévoir la **note nutri-grade** du produit à partir des **valeurs nutritionnelles**?

Analyse univariée et bivariée:
ANOVA (note nutri-grade)

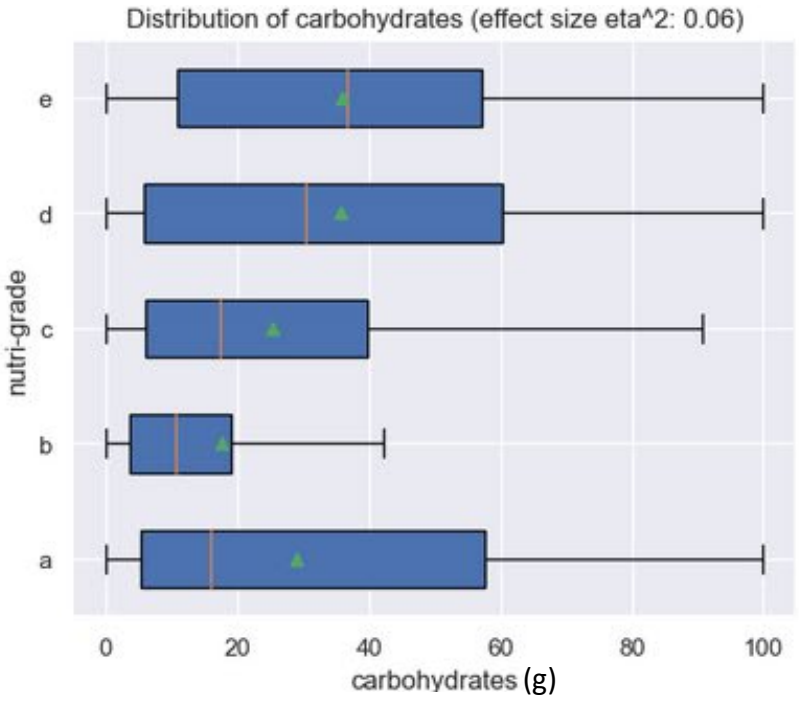


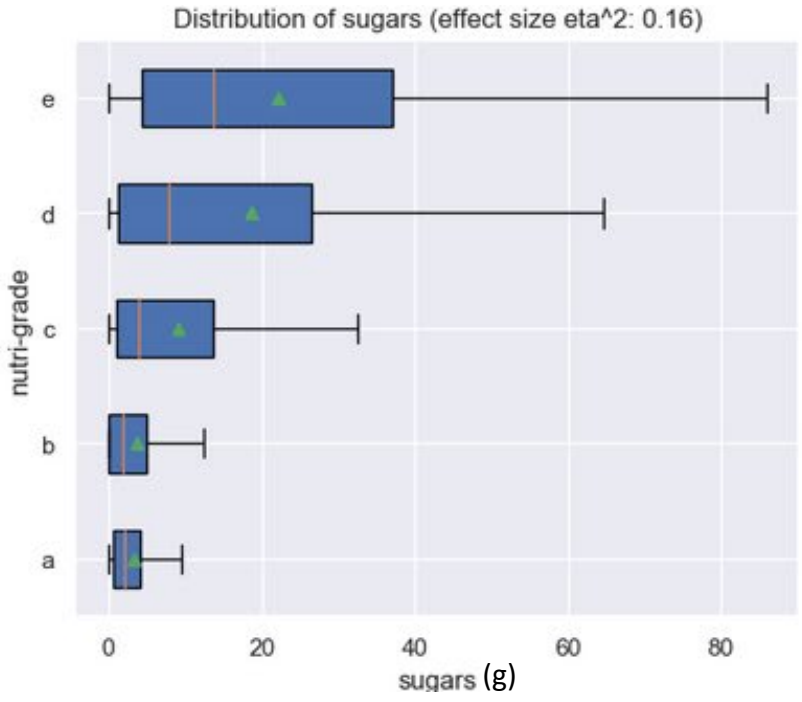


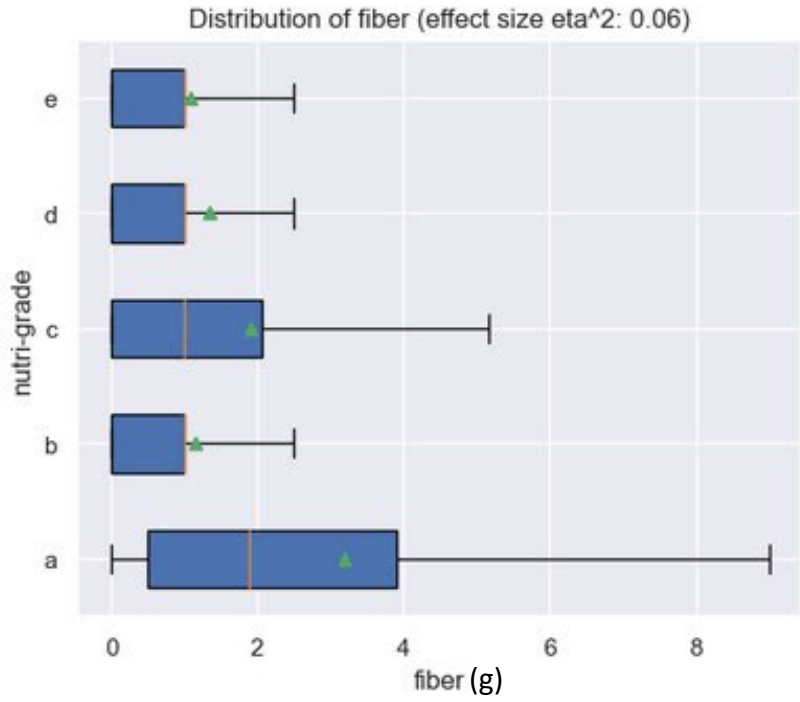


- Les produits avec une bonne nutri-grade sont généralement plus faibles en énergie et gras que les produits avec une mauvaise nutri-grade.
- Energie et gras expliquent le mieux la valeur du nutri-grade. On mesure un effet η^2 de 0.28 pour l'énergie et 0.21 pour le fat.
- Pas d'effet entre les protéines et la nutri-grade du produits. On mesure faible effet η^2 de 0.02.

Analyse univariée et bivariée:
ANOVA (note nutri-grade)







- Les produits avec une bonne nutri-grade sont généralement plus faibles en sucres que les produits avec une mauvaise nutri-grade.
- Pas d'effet entre les glucides et la nutri-grade du produits. On mesure faible effet η^2 de 0.06.
- Pas d'effet entre les fibres et la nutri-grade du produits. On mesure faible effet η^2 de 0.02.

Analyse univariée et bivariée:
ANOVA (catégorie du produit) – vérification statistique

H0 : les moyennes sont égales.
H1 : une ou plusieurs moyennes sont inégales.

➢ For nutriscore_score:

	df	sum_sq	mean_sq	F	PR(>F)
nutriscore_grade	4.0	9.896212e+07	2.474053e+07	2.896982e+06	0.0
Residual	1486642.0	1.269608e+07	8.540105e+00	NaN	NaN

➢ For fat:

	df	sum_sq	mean_sq	F	PR(>F)
nutriscore_grade	4.0	7.904199e+07	1.976050e+07	100833.46134	0.0
Residual	1486642.0	2.913397e+08	1.959716e+02	NaN	NaN

➢ For sugars:

	df	sum_sq	mean_sq	F	PR(>F)
nutriscore_grade	4.0	8.358140e+07	2.089535e+07	70566.088687	0.0
Residual	1486642.0	4.402101e+08	2.961104e+02	NaN	NaN

Influence du facteur **PR = 0.0** (**<0.05**) et facteur **F ≠ 1** → **Rejet H0**

✓ On peux conclure qu'il existe bien une corrélation forte entre la nutri-grade et le nutri-score (la plus grande F).
✓ Il existe aussi un effet significatif entre gras et sucre avec la nutri-grade.

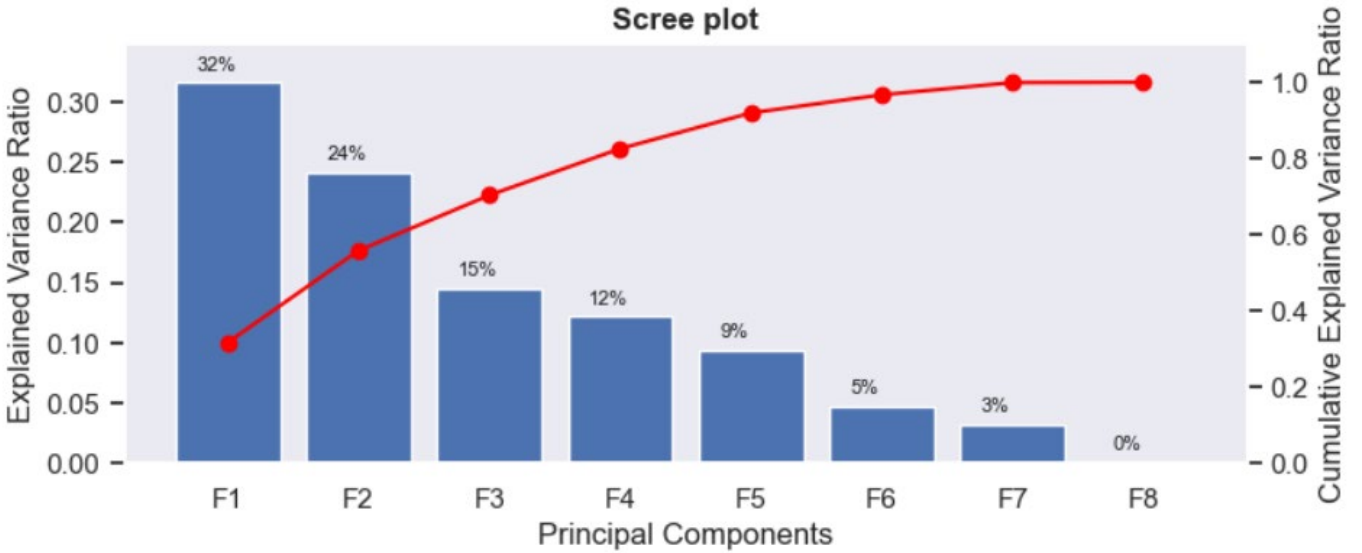
Analyse multivariée:
Analyse en composantes principales ACP

Existe t’il des groupes de variables très corrélées entre elles qui peuvent être regroupées en de nouvelles variables synthétiques ?

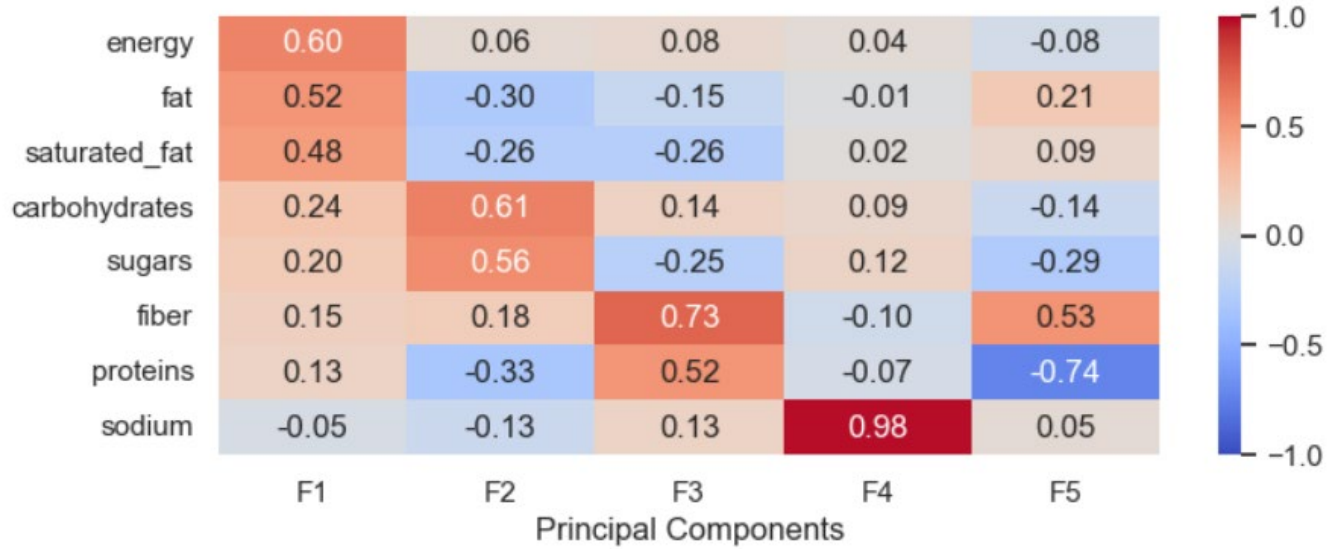
Analyse multivariée:

Analyse en composantes principales ACP – réduction de dimension

- 92 % de l'inertie totale est associée aux 5 premiers axes d'inertie.
- Sélection des 5 premiers composantes.



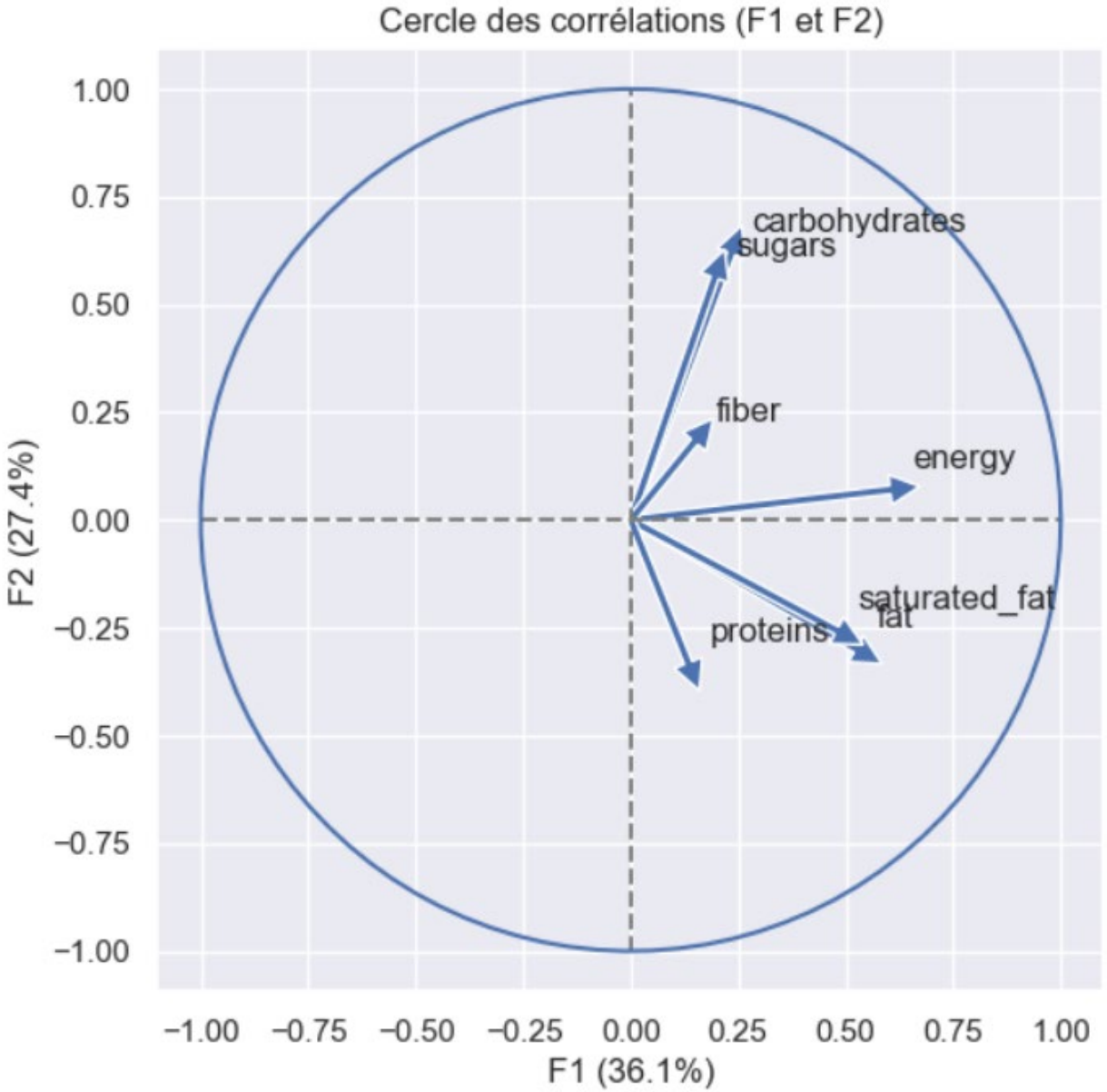
- Sodium occupe une dimension – enlever de l'ACP.
- 4 composantes ACP restantes décrivant 7 variables.



Analyse multivariée:
ACP – Cercle des corrélations

- Projection sur F1:
  - Axe de l'énergie.
  - Gras et gras saturé contribuent le plus à l'énergie.
  - Glucides, sucres, fibres, et protéines contribuent moins à l'énergie.

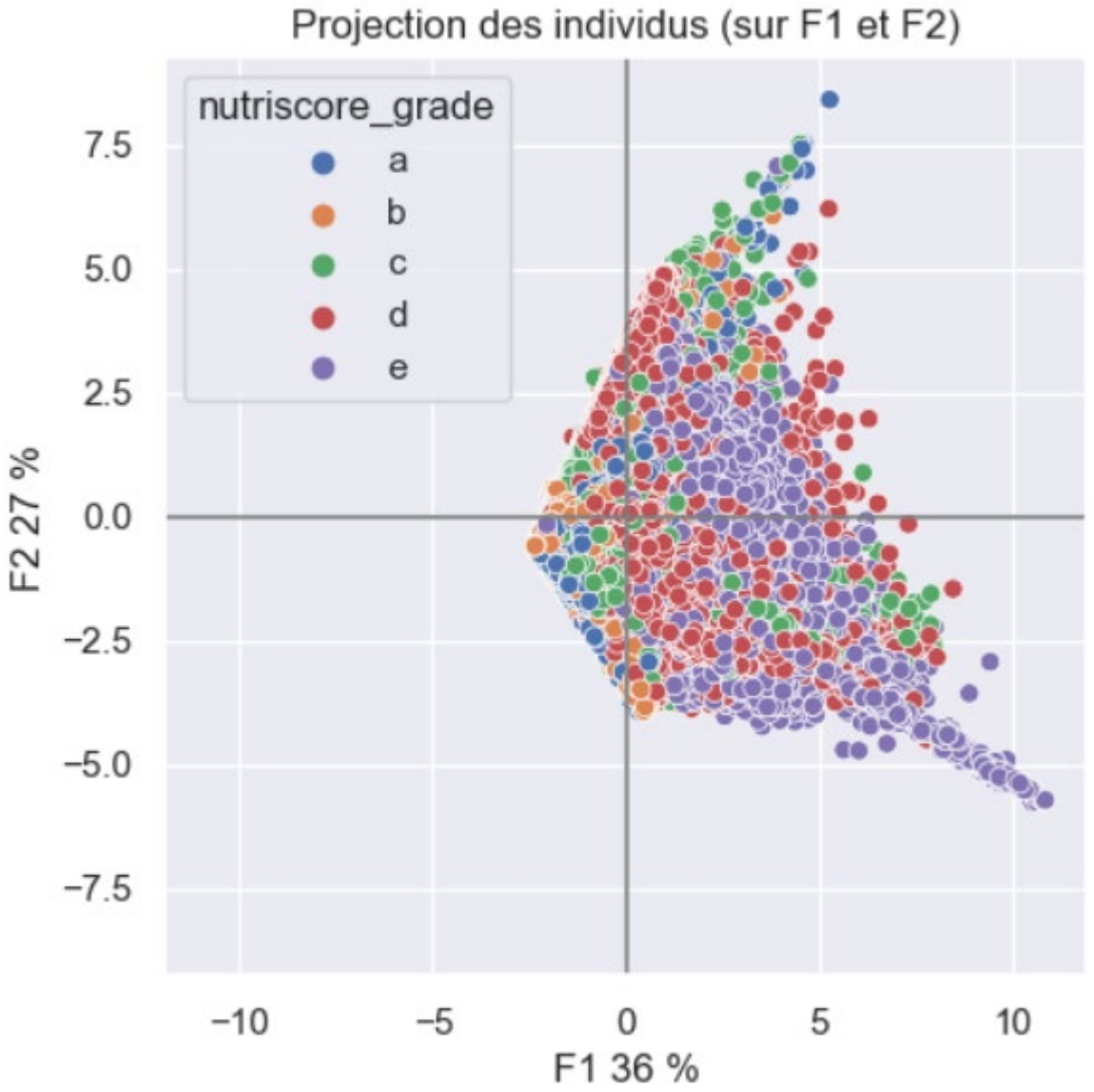
- Projection sur F2:
  - Glucides et sucres sont groupés ensemble sur les valeurs positives. Le fibre est aussi positif.
  - Gras et gras saturé sont groupés ensemble dans les valeurs négatives. On vois que protéines est mal catégorisé sur l'axe de F2.



Analyse multivariée:

ACP – Projection des produits en fonction de leurs nutri-grade

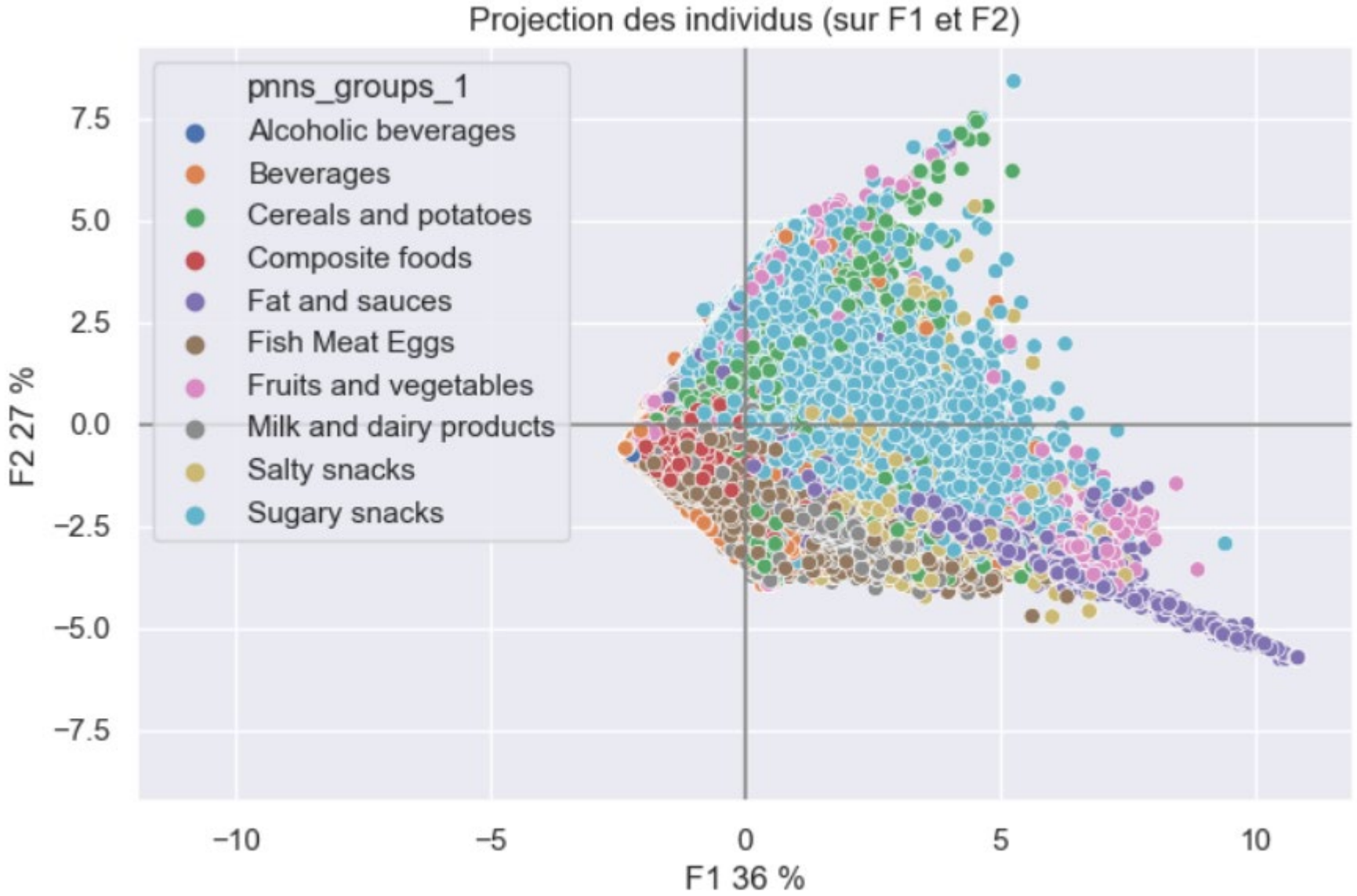
- Les produits avec mauvaise note de nutri-grade (d et e) ont plus grandes valeur de F1 (plus d'énergie) que les produits avec bonne note de nutri-grade (a et b).



Analyse multivariée:

ACP – Projection des produits en fonction de leurs catégorie

- Les fat and sauces sont parmi les plus positives sur l'axe de F1 (plus d'énergie).
- Les Fish Meat Eggs, Milk and dairy products, et fat and sauces ont des valeurs négatives sur l'axe de F2 (protéine, gras, et gras saturée). On voit sugary snacks, et cereals and potatoes avec des valeurs positives de F2 (glucide, sucres et fibres).



Faits pertinents pour l'application:

Recommandation d'un **produit similaire** à celle rechercher par l'utilisateur mais avec une **meilleure note nutri-grade** et un **meilleure score nutri-score.**

- 3 variables importantes au cœur de l'application: la note nutri-grade, la catégorie du produit pnns_group_2, et le score nutri-score.

- Toutes ces 3 variables ont été complété à l'aide de RandomForestClassifier avec des erreurs < 13%.

- Observations sur ces 3 variables:
  - ✓ Forte corrélation entre une bonne note nutri-grade et un bon score nutri-score.
  - ✓ Corrélation significatif entre un bon score nutri-score avec la catégorie des produits qui sont meilleure pour la santé comme les fruits, légumes, céréales, œufs, laits et yogourt.
  - ✓ En parallèle, on a vue une corrélation entre une mauvaise note nutri-grade avec les sucres et les gras.

- Compte tenu des arguments ci-dessus, il est pertinent pour l'application de proposer la meilleure produit à l'utilisateur avec un meilleur nutri-score et nutri-grade.

Conclusion:

- ✓ 1.5 Millions de produits dans la base de données.
- ✓ 15 variables sélectionnés à la fin qui peuvent être réduits à 12 variables via l'aide de l'ACP.

- ✓ Plusieurs observables nous ont permis de tester la pertinence et la faisabilité de l'application.

- ❖ Des idées de développement pour améliorer le système de recommandation:
  - ➢ Proposer un produit vendu dans le pays où l'utilisateur est base.
  - ➢ Proposer un produit avec une meilleure note ecoscore (indicateur représentant l'impact environnemental des produits alimentaires).