Concevez une application au service de la santé publique

Benjamin Bouchard

Sommaire

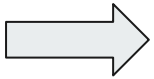
1. Principe de l'application
2. Nettoyage des données
    a. Caractéristiques du dataset.
    b. Nettoyage des colonnes.
    c. Nettoyage des lignes.
    d. Données aberrantes.
    e. Traitement des outliers.
    f. Homogénéisation et regroupement des catégories.
    g. Imputation des données manquantes.
3. Analyse des données
    a. Histogramme des catégories
    b. Analyses univariées
    c. Analyse bivariées
    d. Test statistique
    e. Matrice des corrélations
    f. Analyse en composantes principales
4. Conclusion

ANNEXE: plans des notebooks

1- But de l'étude & principe de l'application

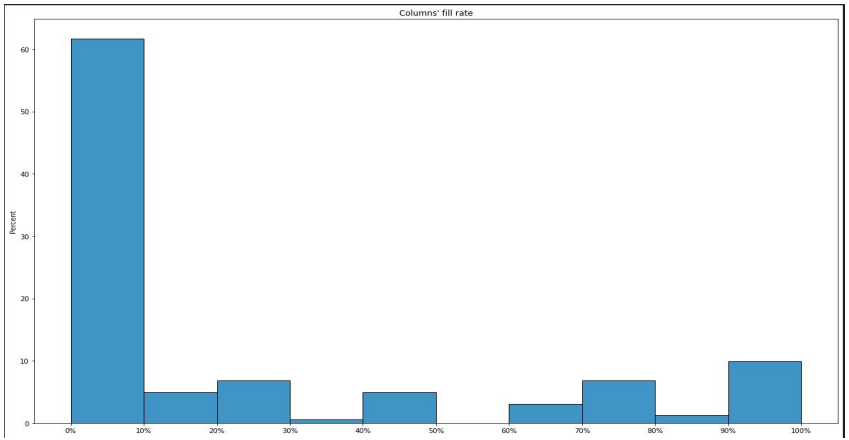
- Etude de la faisabilité d'une application de recommandation.

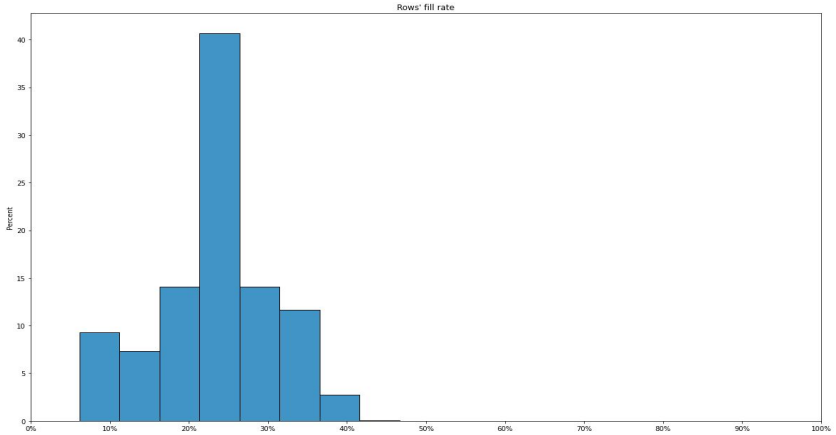
- Application basée sur la composition des produits.

 Ainsi sur la base d'un produit soumis par l'utilisateur, le but serait de proposer un certain nombre de substituts, la recherche de produits alternatifs étant basée sur la composition du produit d'entrée. Des informations sur la qualité nutritionnelle (nutri-score,nutri-grade) seraient également prises en compte pour filtrer les produits présentés.

2-a Caractéristiques du dataset

- Un “grand” dataset: 320 772 lignes et 162 colonnes
- Taux de remplissage global de moins de 24%
- Plus de 60% des colonnes sont remplies à moins de 10% et les lignes ne sont remplies au mieux qu’à 40%

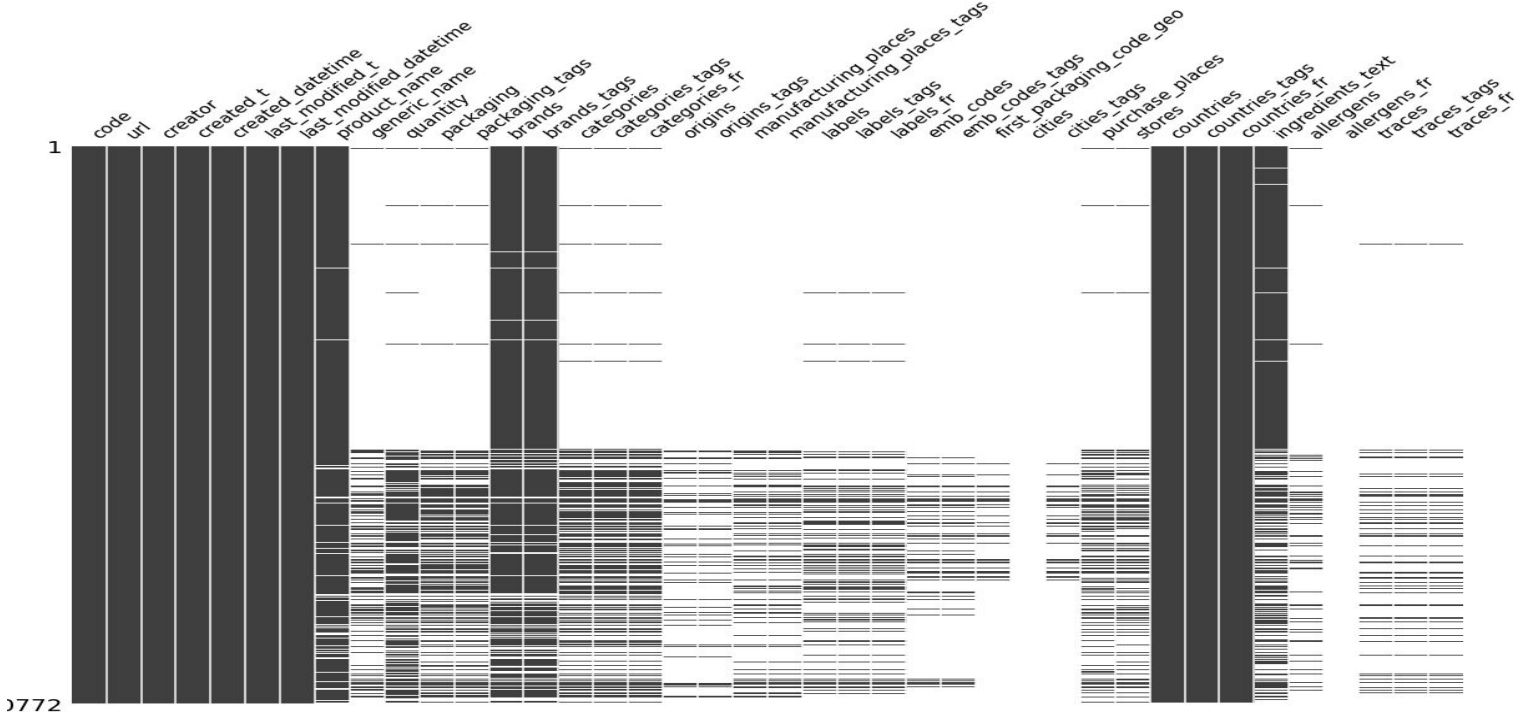




- Un dataset “open source” avec potentiellement beaucoup de saisies manuels (=> erreurs, données non-normalisées)

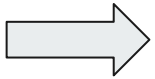
2-a Caractéristiques du dataset (2/2)

- Un remplissage très erratique



2-b Nettoyage des colonnes

- Principes: Les différentes options envisageables pour réduire le nombre de variables d'un dataset sont les suivantes:
  (1) supprimer les variables sans intérêt pour l'étude,
  (2) supprimer l'information redondante,
  (3) supprimer les variables "vides".

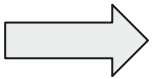


Dans tous les cas, il faut veiller à ne pas perdre d'information utile. En particulier dans le cas (1), il peut être difficile de déterminer à priori les variables qui sont d'intérêt pour le phénomène étudié.

- Traitements mis en place:
  - Suppression de colonnes redondantes: dates dans deux formats différents; colonnes colinéaires (sel/sodium)
  - Suppression de colonnes vides

2-c Nettoyage des lignes

- Principes: Les différentes options envisageables pour réduire les lignes d'un dataset sont les suivantes:
  (1) supprimer les lignes dupliquées (besoin d'un identifiant),
  (3) supprimer les lignes "vides" ou "presque vides".

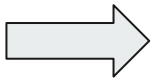


Dans tous les cas, il faut veiller à ne pas perdre d'information utile.

- Traitements mis en place:
  - Suppression des doublons **avec utilisation de l'information des lignes supprimées**
  - Suppression de lignes aberrantes
  - Suppression des lignes sans informations nutritionnelles

2-d Données aberrantes

- Principes: Pour certaines variables, on a une connaissance à priori de leurs domaines de valeur (“les valeurs qu’elles peuvent prendre”)



On met à “NAN” les valeurs hors domaine.

- Traitements mis en place:
  - Apport nutritionnel pour 100g ne peut pas être négatif.
  - Apport nutritionnel pour 100g ne peut pas être supérieur à 100g.
  - Apport énergétique pour 100g supérieur à une limite théorique ().
  - Nutri-score inférieur à -15 ou supérieur à 40.
  - Nutri-grade non inclus dans {'a','b','c','d','e'}.

2-e Traitement des “outliers”

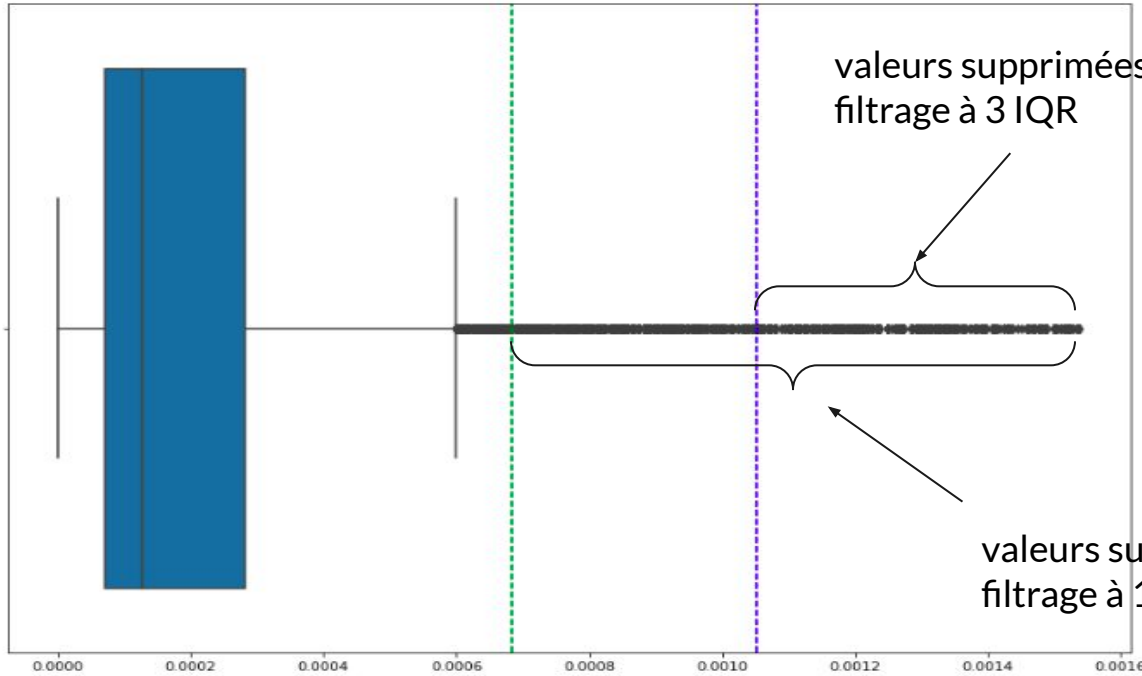
- Principes: Les “outliers” correspondent à des valeurs possibles (donc dans le domaine de valeurs précédent) mais situés dans “les queues de distribution”.
    - Classiquement on traite les outliers en utilisant des limites basées sur les quartiles (Qi) et l’écart interquartile (IQR = Q3-Q1). On considère comme outliers les valeurs en dehors de l’intervalle :

      [Q1 - 1,5xIQR ; Q3 + 1,5xIQR] ou parfois [Q1 - 3xIQR ; Q3 + 3xIQR]
    - Dans notre cas, cette approche aurait aboutit à supprimer beaucoup trop de valeurs.
    - Nous avons donc opté pour une méthode basée sur la contribution à la variance:

      $C_i = \frac{var - var_{wo}(i)}{var}$(i) où var_{wo} est la var calculée sans le point i
    - Une valeur est considérée comme un outlier si elle représente plus de 50% de la variance totale.

Traitement des “outliers”

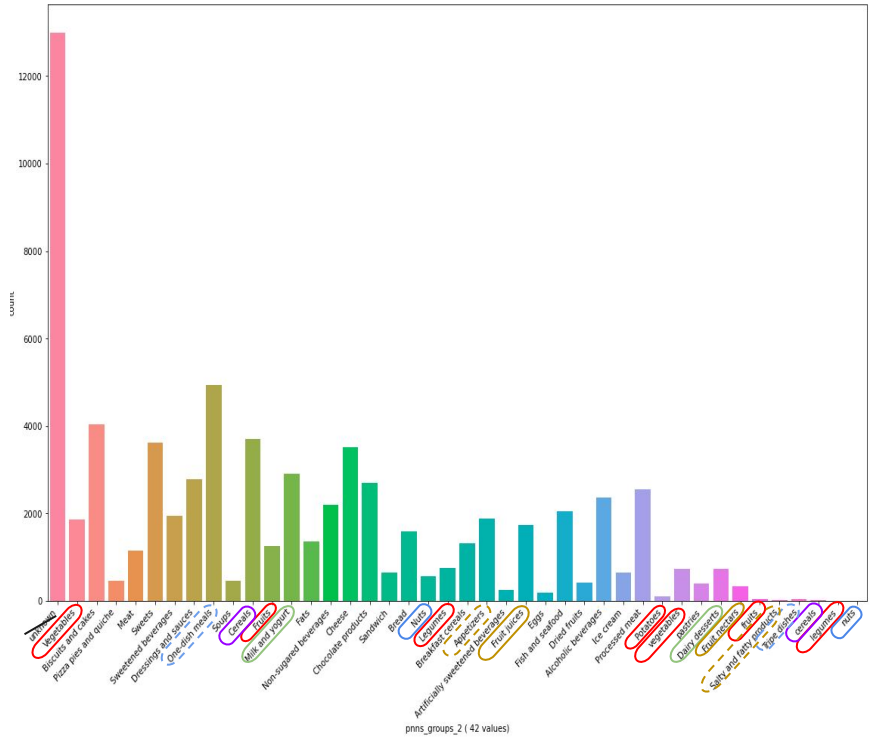
Filtrage en utilisant les seuils liés aux IQR: trop de valeurs supprimées

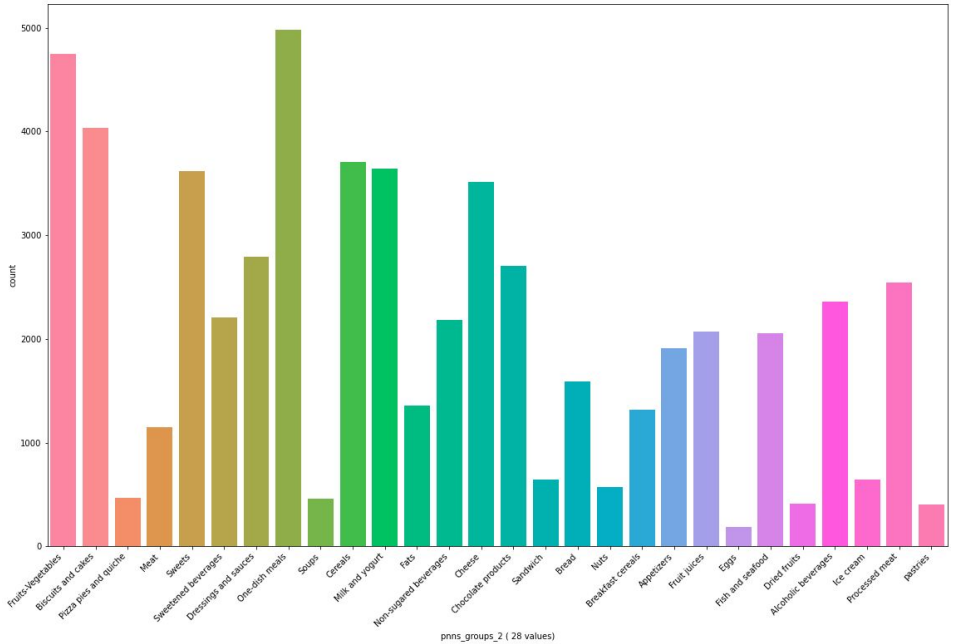


Traitement des “outliers”

contribution à la variance par catégorie

2-g Homogénéisation et regroupement des catégories





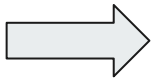
2-h Imputation des données manquantes

- Principes: “Imputer” correspond au fait de renseigner des valeurs manquantes à partir de données présentes dans le dataset. Pour cela, on peut utiliser:
  - des formules/relations qui relient plusieurs variables du dataset,
  - des méthodes statistiques: imputation par la moyenne, la médiane ou en utilisant un algorithme tel le “KNN inputers”

- Traitements mis en place:
  - A partir de la description du calcul du nutri-score (francais):
    - Implémentation en python du nutri-score
    - puis calcul du nutri-score puis du nutri-grade quand ils étaient manquants,
  - Imputation du sel à partir du sodium,
  - Imputation de valeurs manquantes par la médiane en considérant la catégorie des produits.

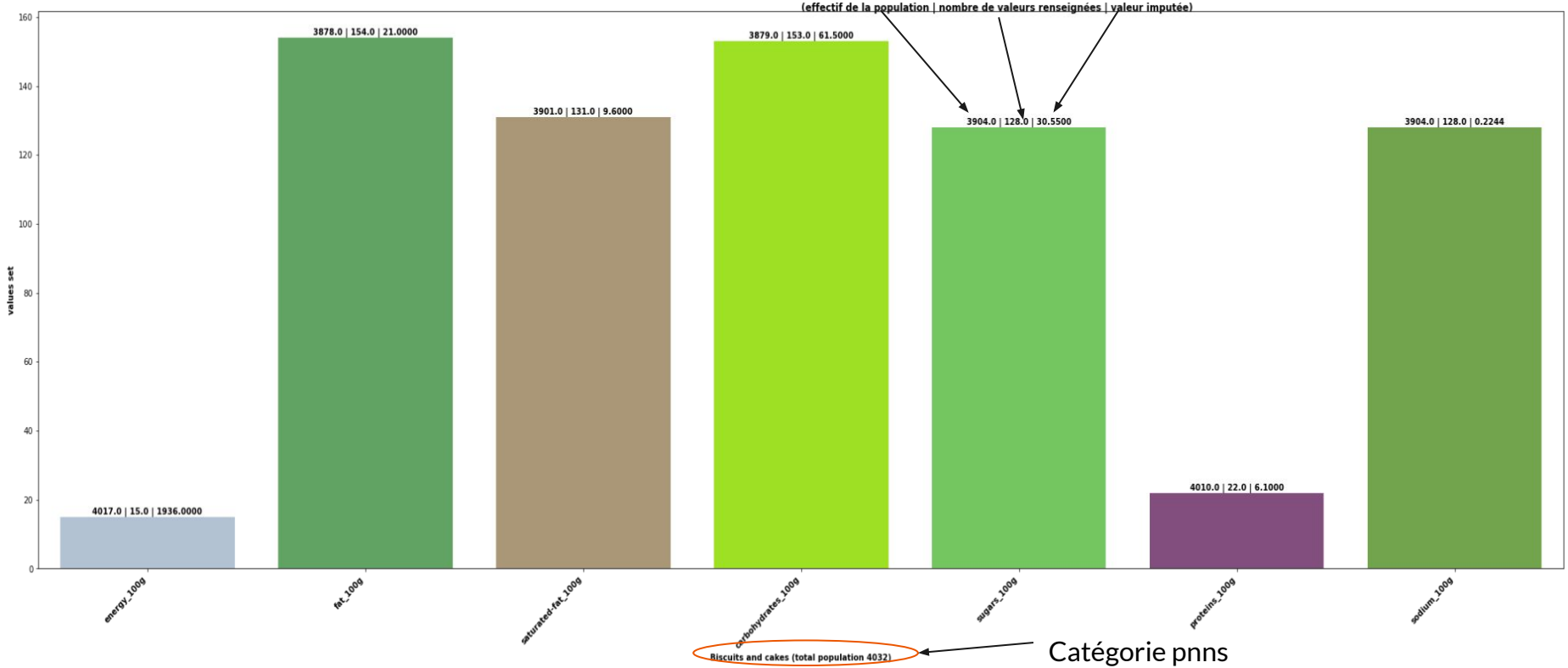
Imputation des données manquantes

- Imputation de valeurs manquantes par la médiane en considérant la catégorie des produits à consister à
  - considérer chaque variable en filtrant par catégorie
  - n'imputer une variable que si elle était remplie par au moins 80% des individus de la catégorie



Ne pas imputer une variable qui n'est pas pertinente pour le type de produit (mais qui ne serait pas totalement vide)

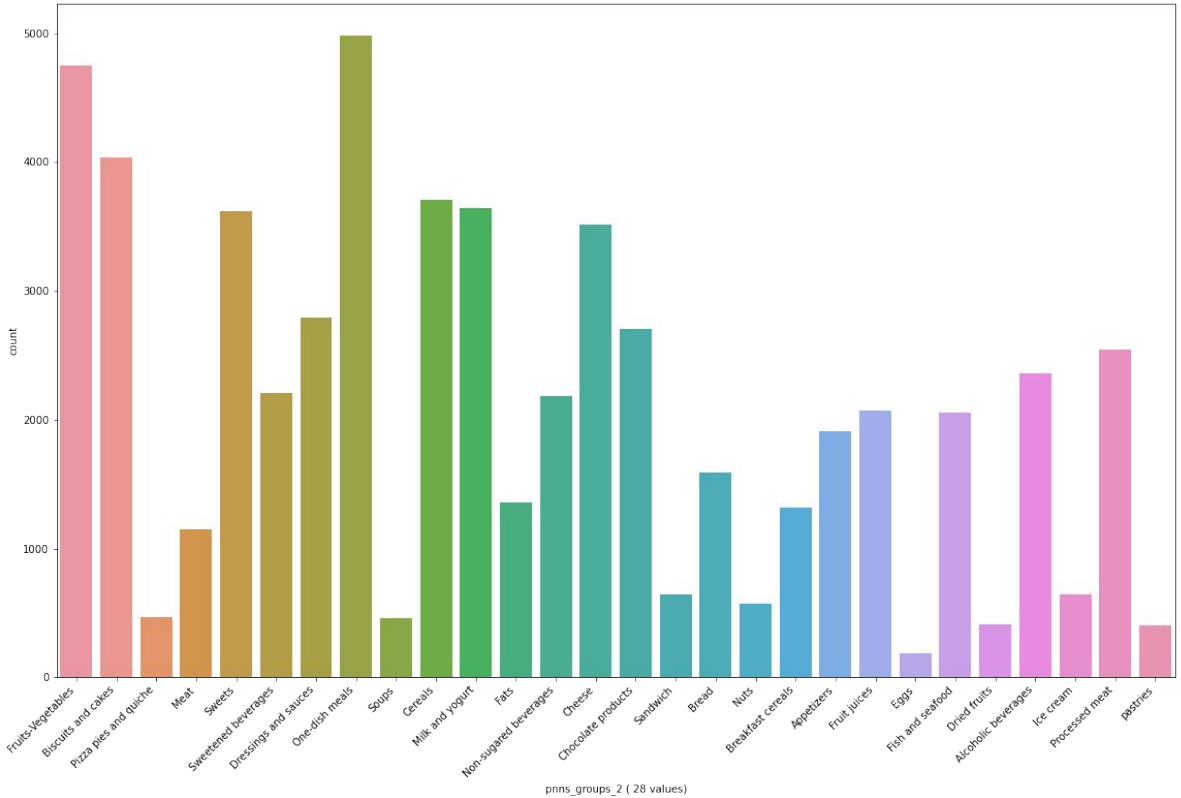
Imputation des données manquantes (exemple des biscuits)



Catégorie pnns

Analyse de données

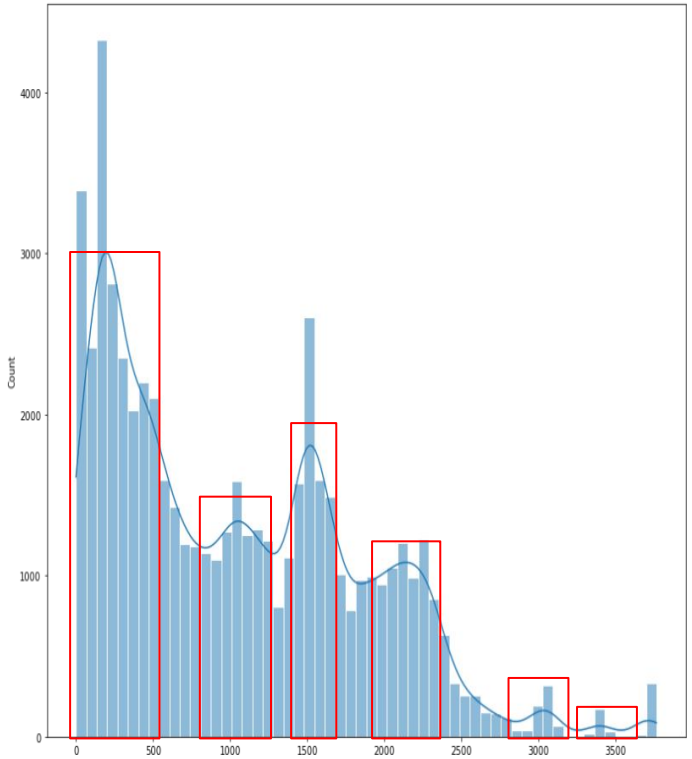
3-a Histogramme des catégories

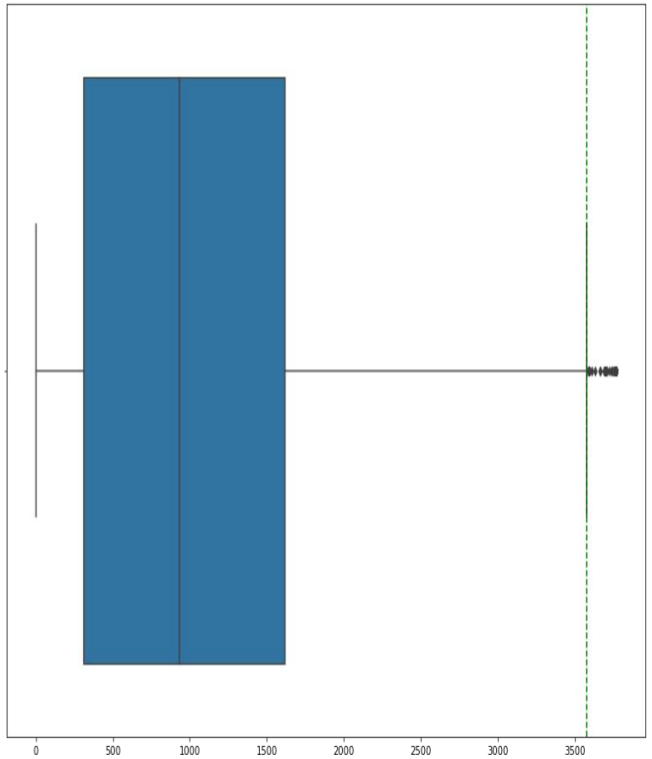


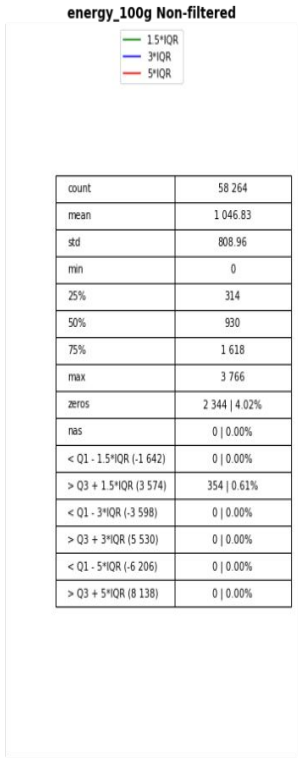
One-dish meals	4 980
Fruits-Vegetables	4 745
Biscuits and cakes	4 032
Cereals	3 706
Milk and yogurt	3 641
Sweets	3 617
Cheese	3 516
Dressings and sauces	2 790
Chocolate products	2 706
Processed meat	2 547
Alcoholic beverages	2 360
Sweetened beverages	2 205
Non-sugared beverages	2 185
Fruit juices	2 071
Fish and seafood	2 055
Appetizers	1 908
Bread	1 591
Fats	1 361
Breakfast cereals	1 314
Meat	1 149
Ice cream	646
Sandwich	641
Nuts	569
Pizza pies and quiche	464
Soups	463
Dried fruits	413
pastries	404
Eggs	185

2-b Analyses univariées (variables d'entrée du nutri-score)

Energy



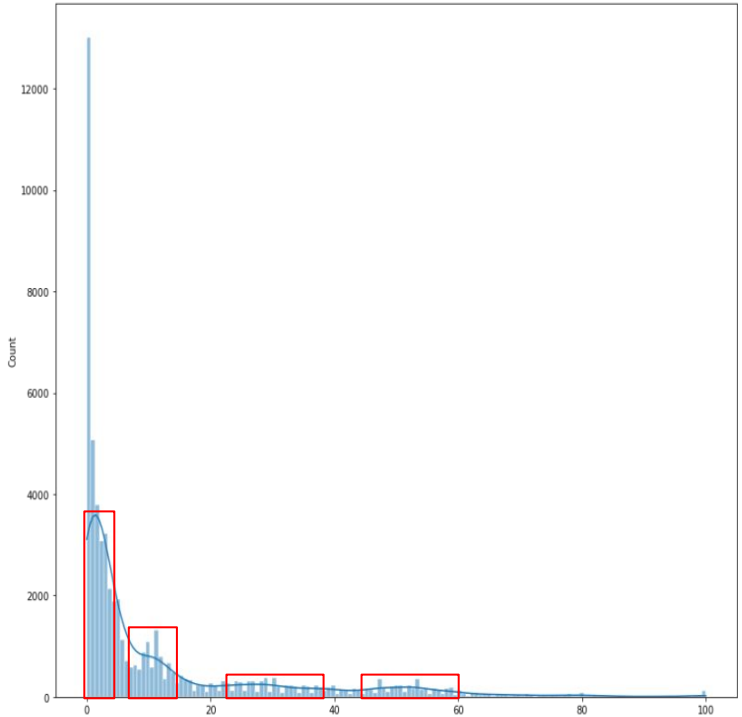


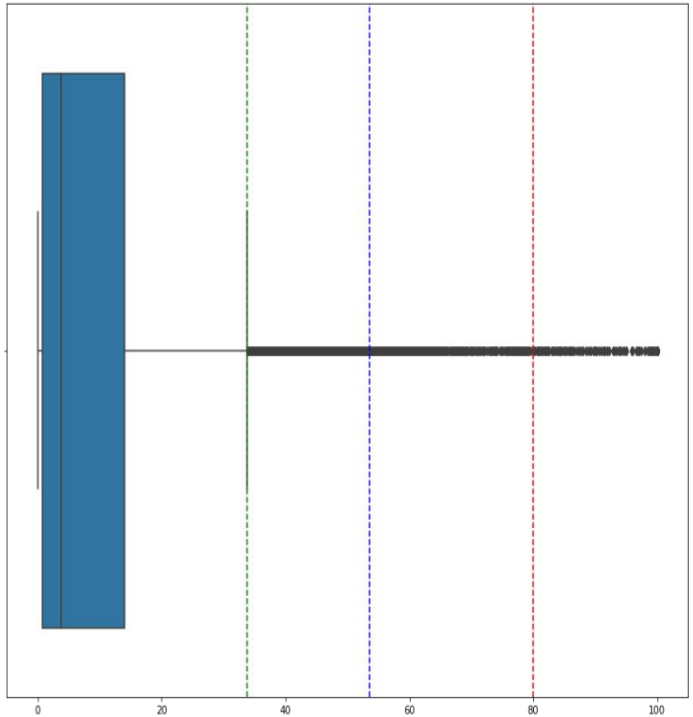


count	58 264
mean	1 046.83
std	808.96
min	0
25%	314
50%	930
75%	1 618
max	3 766
zeros	2 344 | 4.02%
nas	0 | 0.00%
< Q1 - 1.5*IQR (-1 642)	0 | 0.00%
> Q3 + 1.5*IQR (3 574)	354 | 0.61%
< Q1 - 3*IQR (-3 598)	0 | 0.00%
> Q3 + 3*IQR (5 530)	0 | 0.00%
< Q1 - 5*IQR (-6 206)	0 | 0.00%
> Q3 + 5*IQR (8 138)	0 | 0.00%

Analyses univariées (variables d'entrée du nutri-score)

Sugar





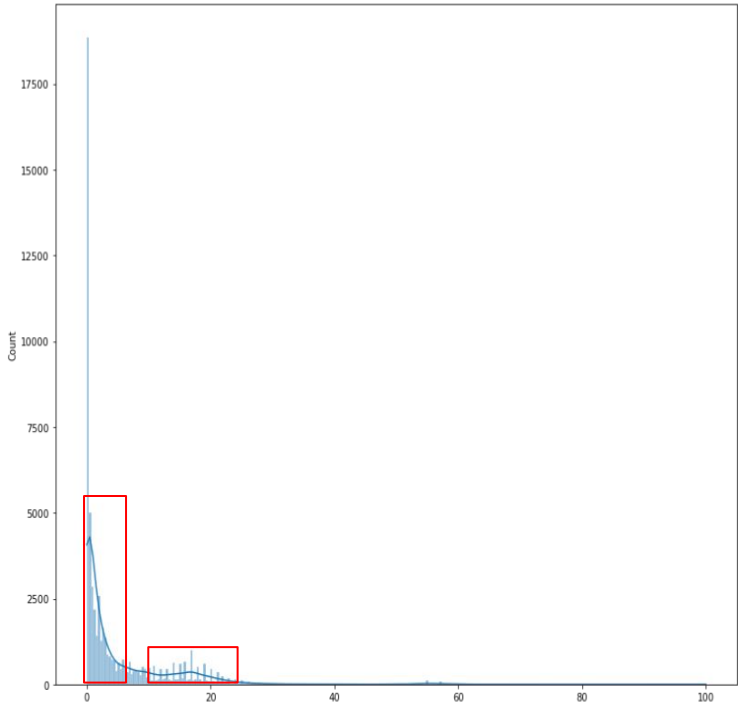
sugars_100g Non-filtered

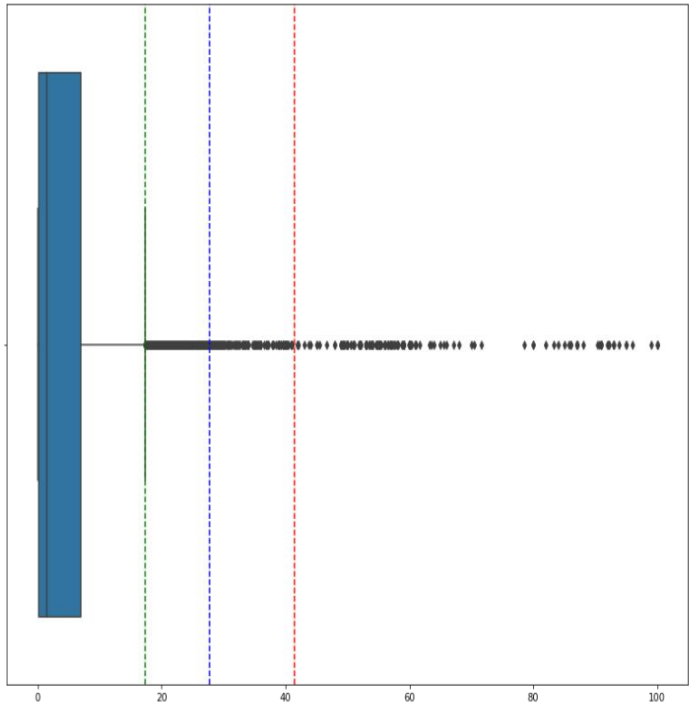
1.5*IQR
3*IQR
5*IQR

count	58 264
mean	12.31
std	18.27
min	0
25%	0.80
50%	3.70
75%	14
max	100
zeros	6 003 | 10.30%
nas	0 | 0.00%
< Q1 - 1.5*IQR (-19)	0 | 0.00%
> Q3 + 1.5*IQR (33.80)	7 730 | 13.27%
< Q1 - 3*IQR (-38.80)	0 | 0.00%
> Q3 + 3*IQR (53.60)	2 865 | 4.92%
< Q1 - 5*IQR (-65.20)	0 | 0.00%
> Q3 + 5*IQR (80)	456 | 0.78%

Analyses univariées (variables d'entrée du nutri-score)

Saturated-fat





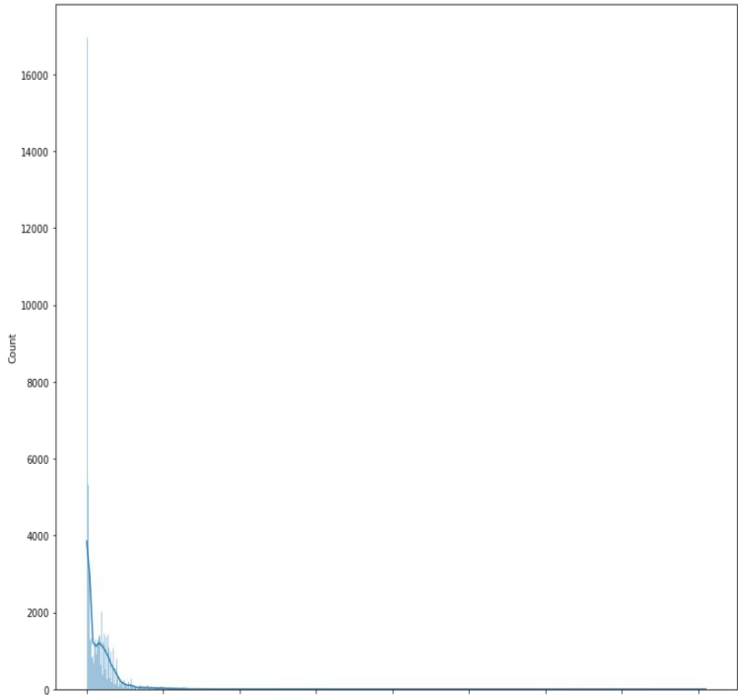
saturated-fat_100g Non-filtered

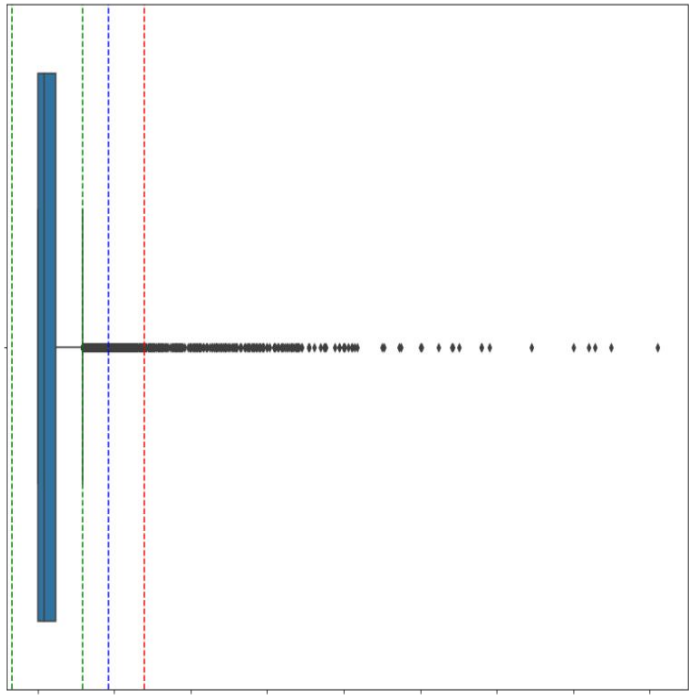
1.5*IQR
3*IQR
5*IQR

count	58 264
mean	5.11
std	8.21
min	0
25%	0.10
50%	1.50
75%	7
max	100
zeros	9 917 | 17.02%
nas	0 | 0.00%
< Q1 - 1.5*IQR (-10.25)	0 | 0.00%
> Q3 + 1.5*IQR (17.35)	4 676 | 8.03%
< Q1 - 3*IQR (-20.60)	0 | 0.00%
> Q3 + 3*IQR (27.70)	804 | 1.38%
< Q1 - 5*IQR (-34.40)	0 | 0.00%
> Q3 + 5*IQR (41.50)	485 | 0.83%

Analyses univariées (variables d'entrée du nutri-score)

Sodium





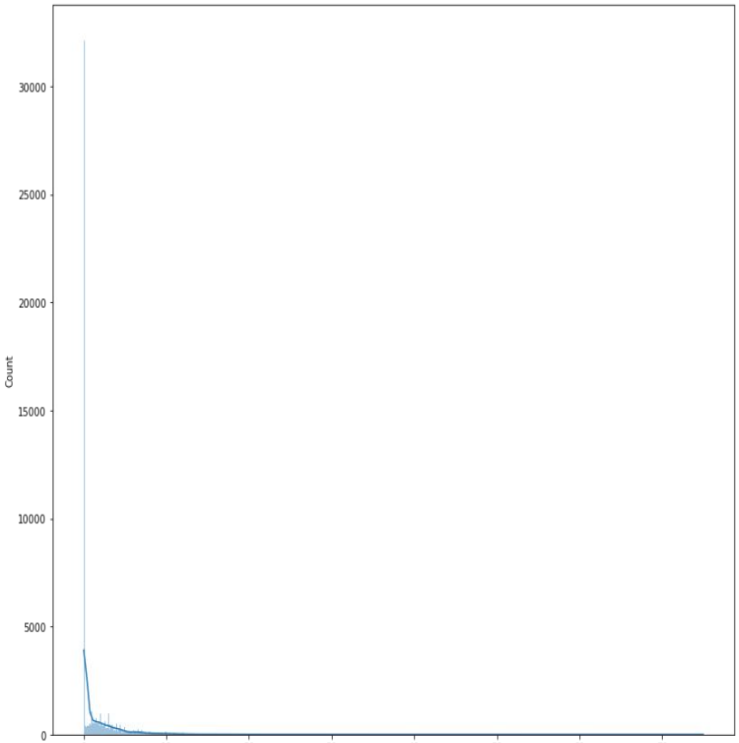
sodium_100g Non-filtered

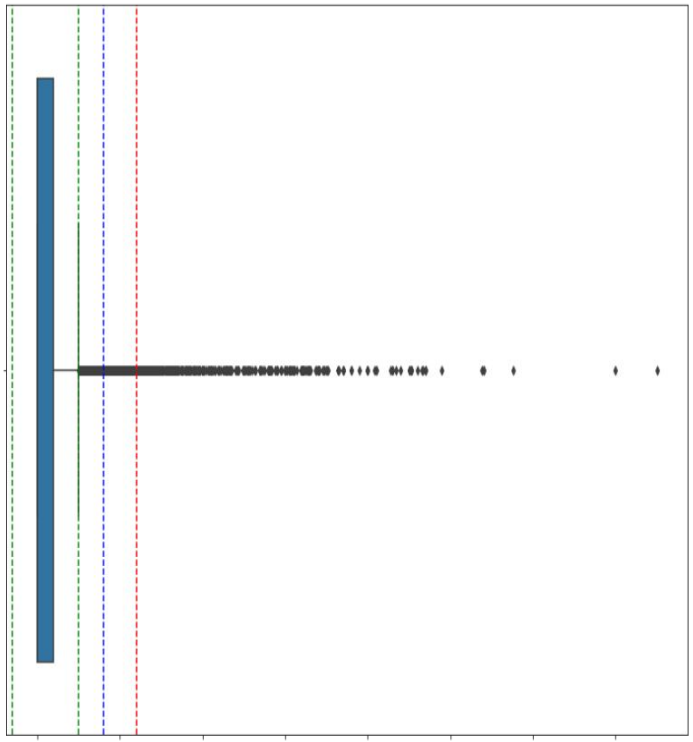
1.5*IQR
3*IQR
5*IQR

count	58 264
mean	0.31
std	0.52
min	0
25%	0.01
50%	0.15
75%	0.47
max	16.20
zeros	7 036 | 12.08%
nas	0 | 0.00%
< Q1 - 1.5*IQR (-0.68)	0 | 0.00%
> Q3 + 1.5*IQR (1.16)	2 225 | 3.82%
< Q1 - 3*IQR (-1.37)	0 | 0.00%
> Q3 + 3*IQR (1.85)	855 | 1.47%
< Q1 - 5*IQR (-2.29)	0 | 0.00%
> Q3 + 5*IQR (2.77)	259 | 0.44%

Analyses univariées (variables d'entrée du nutri-score)

Fiber





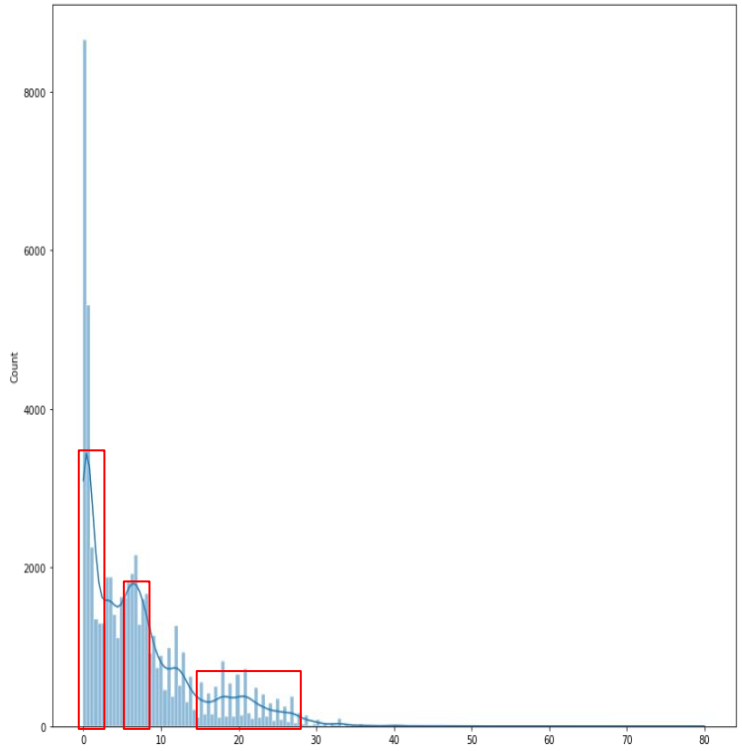
fiber_100g Non-filtered

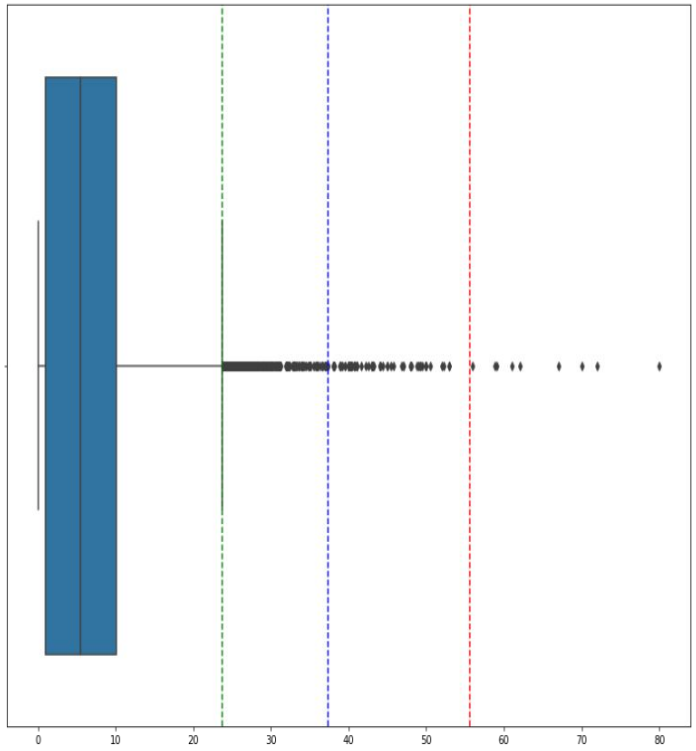
1.5*IQR
3*IQR
5*IQR

count	58 264
mean	1.57
std	3.20
min	0
25%	0
50%	0
75%	2
max	75
zeros	31 356 | 53.82%
nas	0 | 0.00%
< Q1 - 1.5*IQR (-3)	0 | 0.00%
> Q3 + 1.5*IQR (5)	5 019 | 8.61%
< Q1 - 3*IQR (-6)	0 | 0.00%
> Q3 + 3*IQR (8)	2 202 | 3.78%
< Q1 - 5*IQR (-10)	0 | 0.00%
> Q3 + 5*IQR (12)	748 | 1.28%

Analyses univariées (variables d'entrée du nutri-score)

Proteins





proteins_100g Non-filtered

1.5*IQR
3*IQR
5*IQR

count	58 264
mean	7.18
std	7.32
min	0
25%	1
50%	5.50
75%	10.10
max	80
zeros	5 914 | 10.15%
nas	0 | 0.00%
< Q1 - 1.5*IQR (-12.65)	0 | 0.00%
> Q3 + 1.5*IQR (23.75)	2 337 | 4.01%
< Q1 - 3*IQR (-26.30)	0 | 0.00%
> Q3 + 3*IQR (37.40)	106 | 0.18%
< Q1 - 5*IQR (-44.50)	0 | 0.00%
> Q3 + 5*IQR (55.60)	9 | 0.02%

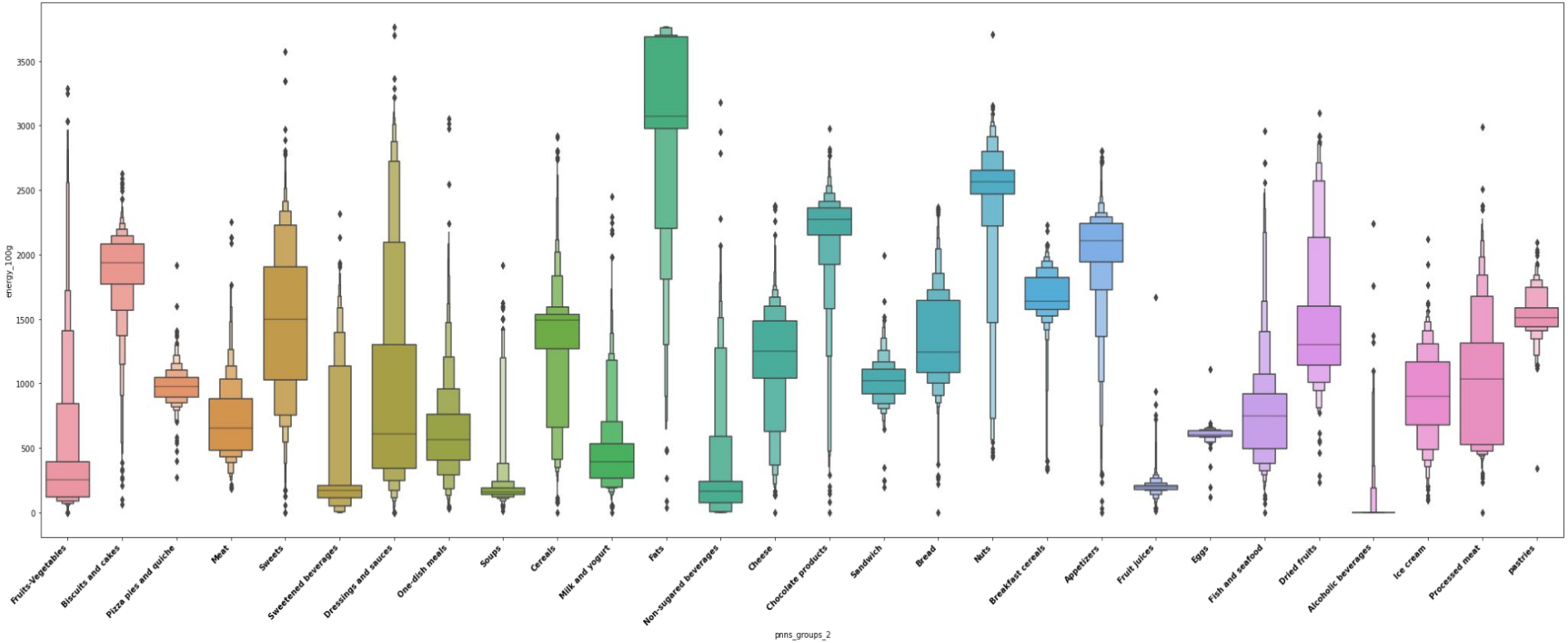
Analyses univariées (variables d'entrée du nutri-score)

- Dans la plupart des cas, on observe plusieurs modes mais pas toujours flagrant
- Beaucoup de zéros
  - rendent la lecture des histogrammes difficiles
  - "décalent vers la gauche les quartiles"
- Pas d'outliers (sauf sodium)

⇨ Etude des variables en fonction des catégories pour s'assurer que les modes sont bien présents et qu'ils permettent de distinguer la catégorie des produits.

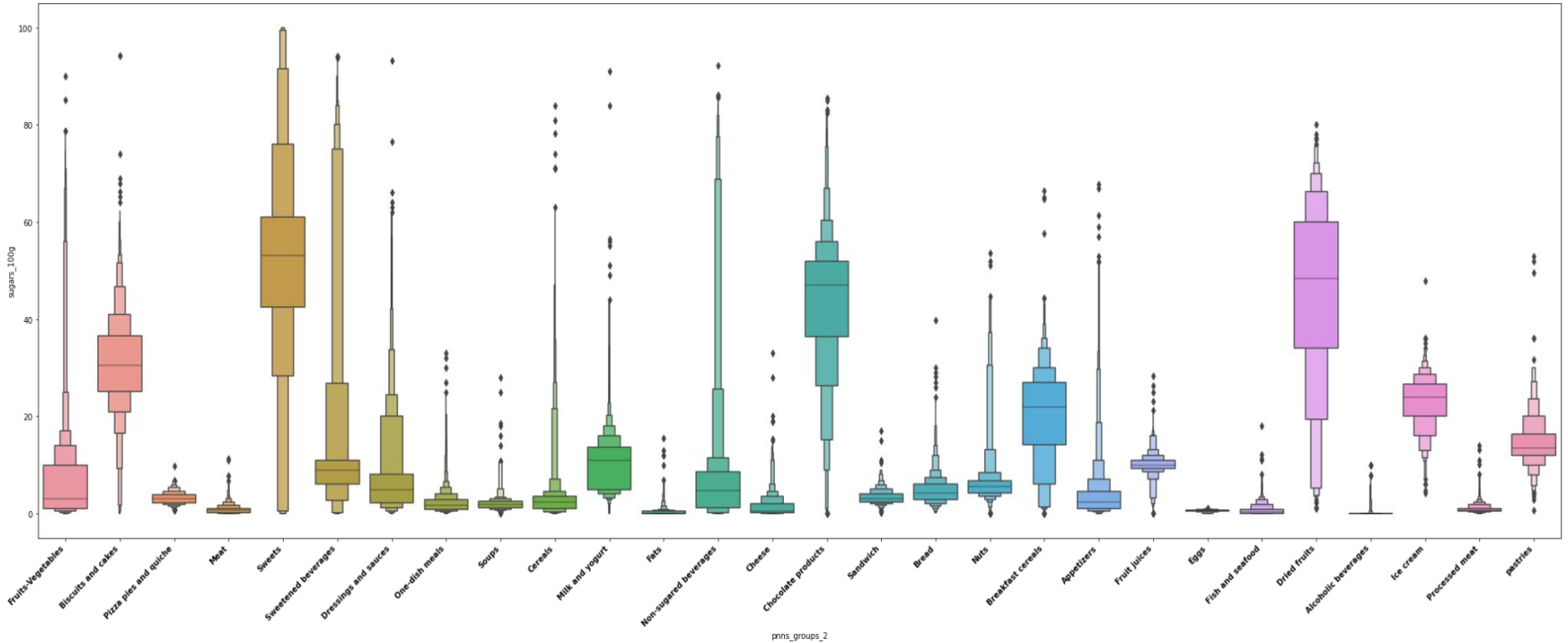
2-c Analyse bivariées (variables d'entrée du nutri-score en fonction de la catégories)

Energy



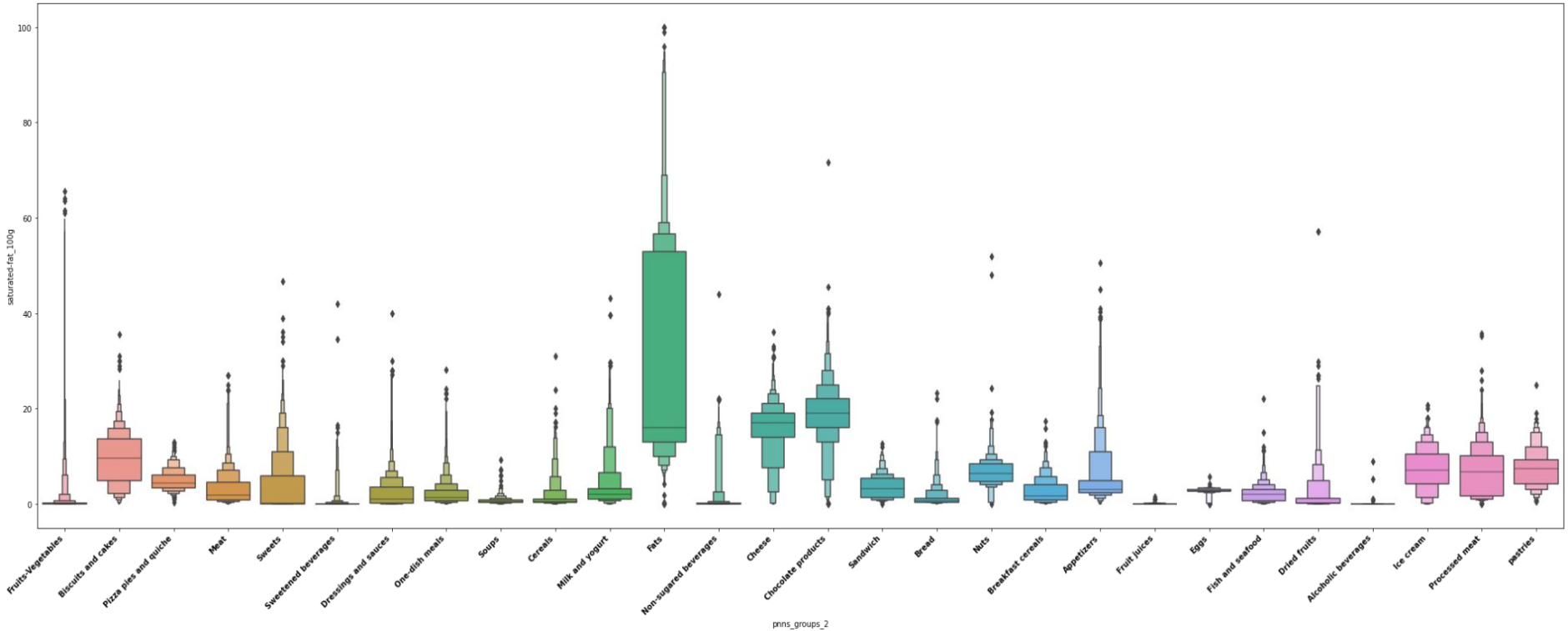
Analyse bivariées (variables d'entrée du nutri-score en fonction de la catégories)

Sugar



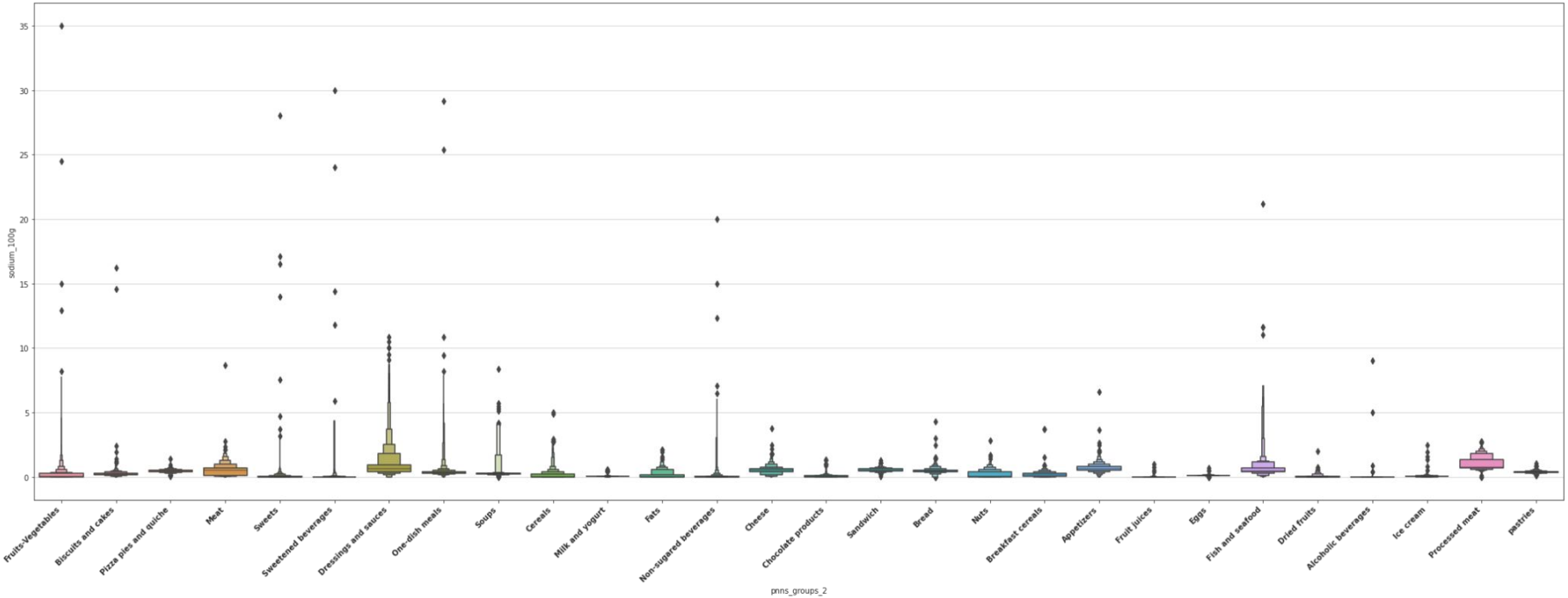
Analyse bivariées (variables d'entrée du nutri-score en fonction de la catégories)

Saturated-fat



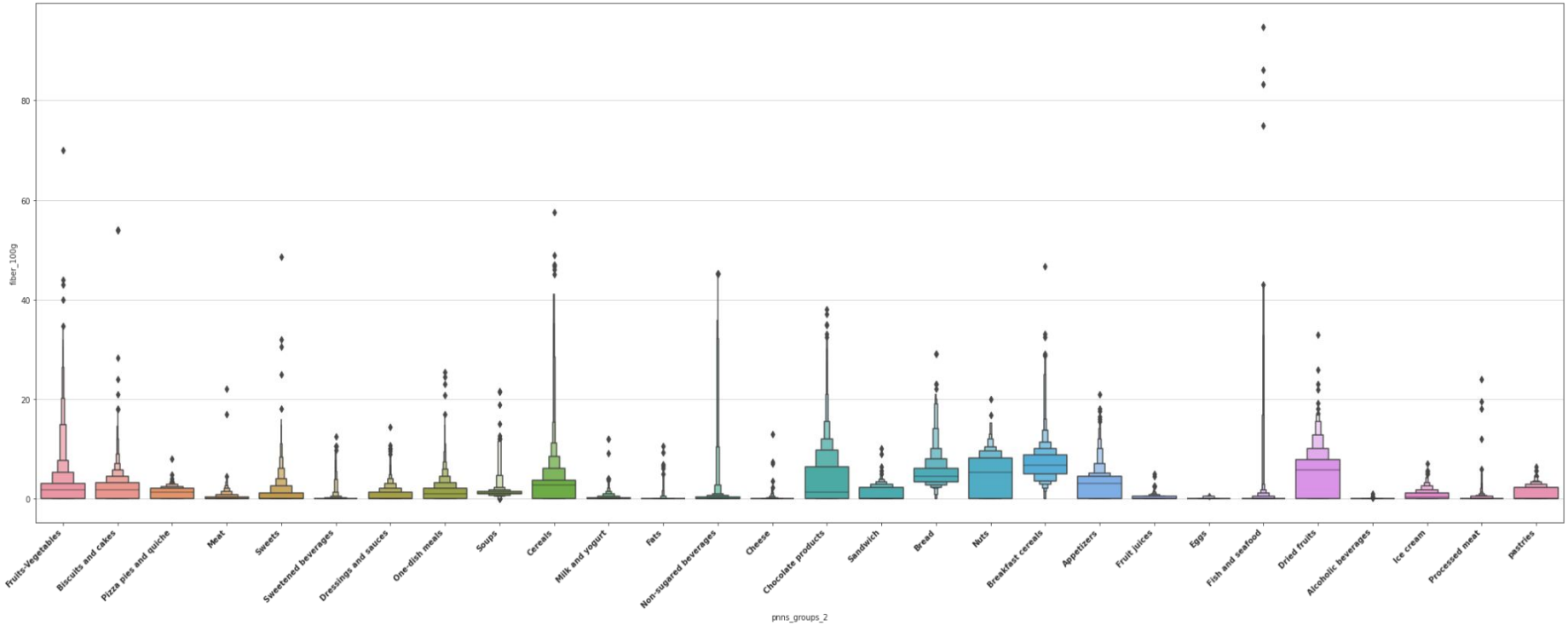
Analyse bivariées (variables d'entrée du nutri-score en fonction de la catégories)

Sodium



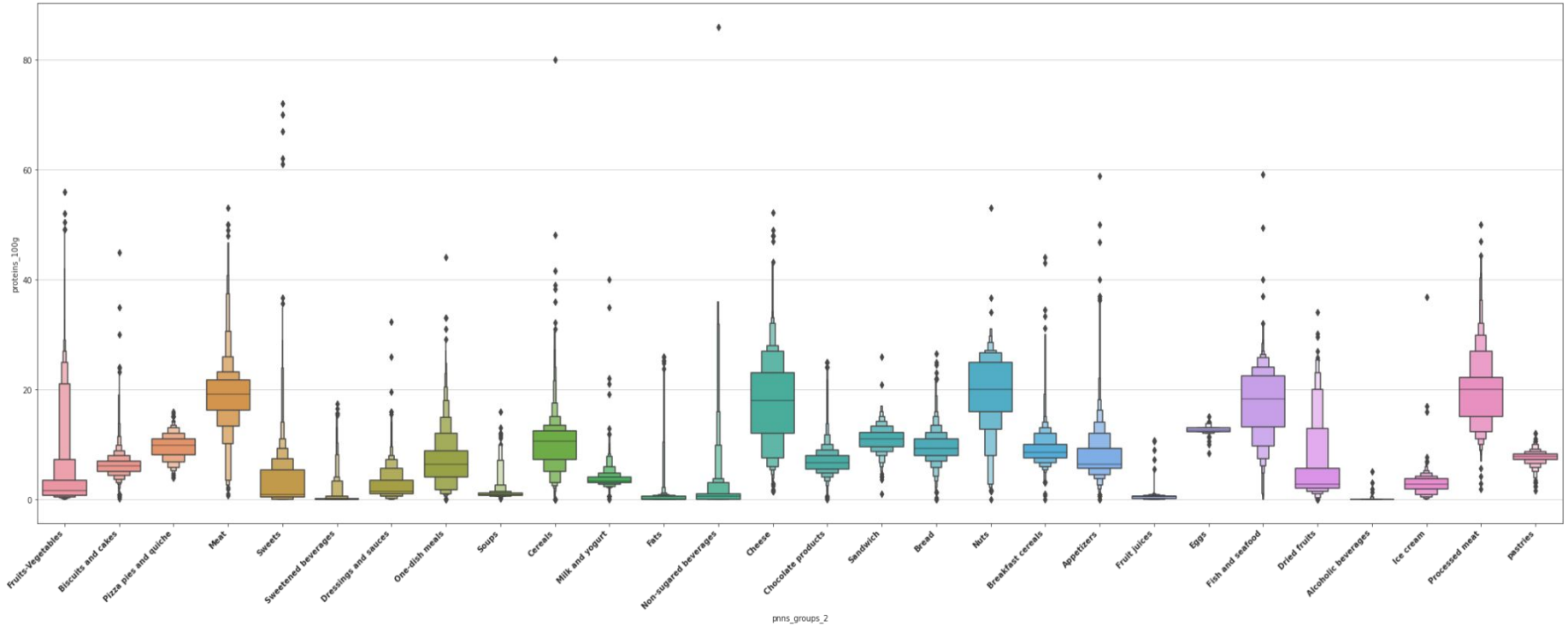
Analyse bivariées (variables d'entrée du nutri-score en fonction de la catégories)

Fiber

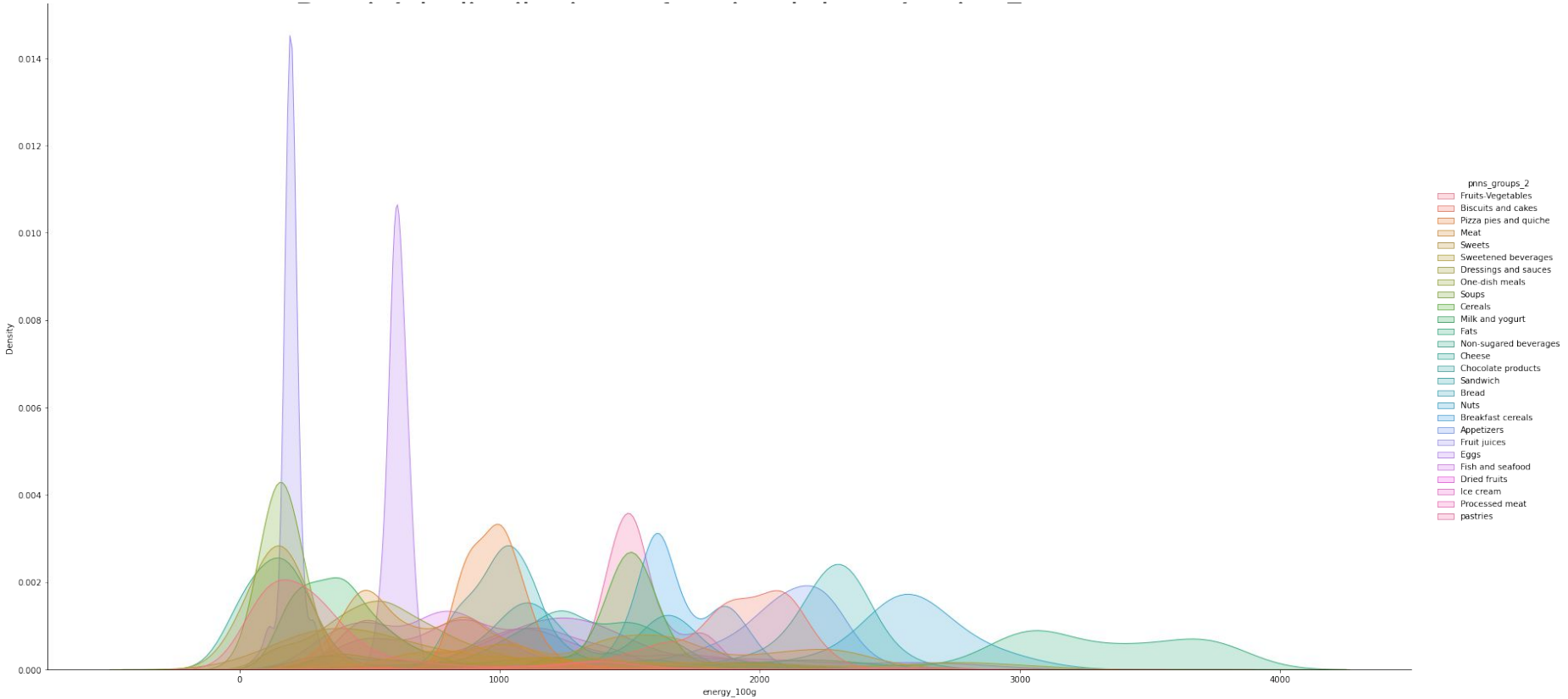


Analyse bivariées (variables d'entrée du nutri-score en fonction de la catégories)

Proteins

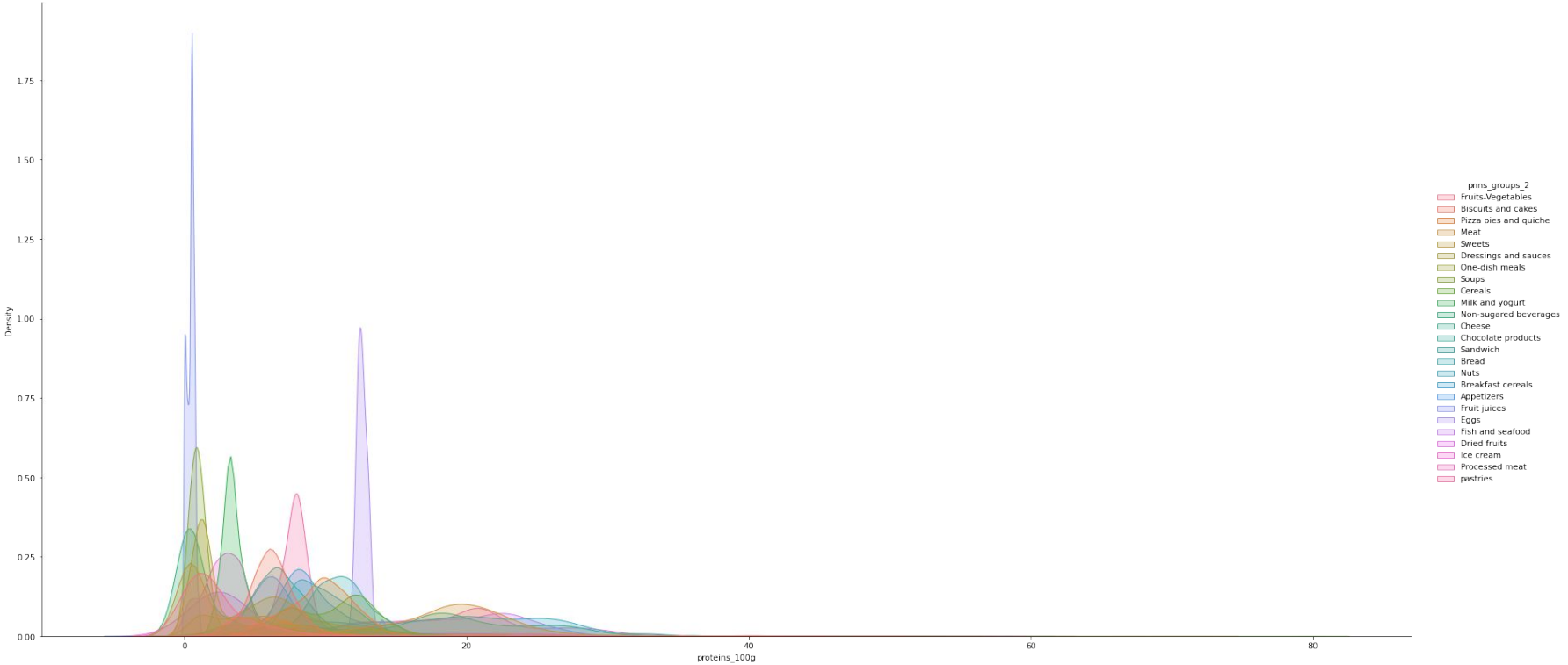


Analyse bivariées (variables d'entrée du nutri-score en fonction de la catégories)



Analyse bivariées (variables d'entrée du nutri-score en fonction de la catégories)

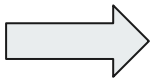
Densité de distribution en fonction de la catégorie - Proteins



Analyses bivariées

- Les boîtes à moustaches tracées en fonction des catégories permettent de visualiser les différents distributions correspondant aux catégories de produit.

- Cependant les densités de distributions estimées en fonction de la catégorie permettent de voir que la distinction entre catégorie au sein d'une même variable n'est toujours évidente.



Pour s'assurer que les variables vont permettre de distinguer la catégorie, un test statistique est nécessaire.

2-d Test statistique

- Test ANOVA impossible car pas de normalité (cf. tests de Shapiro)

- Utilisation d'un test de Kruskal:

    - Pour chaque catégorie, on crée un groupe d'échantillons, chaque échantillon correspondant aux valeurs des apports/indications pour 100g filtrées par catégorie
    - Pour chacun des groupes ainsi constitués, on effectue un test de Kruskal dont l'hypothèse nulle est la suivante:
        - H0: M0=M1=...=Mi= ...=Mn (les médianes des populations dont sont issues les échantillons sont les mêmes)
        - H1: au moins deux médianes sont différentes.

    Les résultats sont examinés au seuil de 5%.

Test statistique

- Pour chaque catégorie (et pour les deux typologies pnns1 et pnns2), **on rejette H0**.

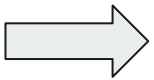
⇨ Au sein de chaque groupe d'échantillons, au moins l'un d'eux est "stochastiquement dominant".

⇨ Pour chaque catégorie, il existe un apport/indication pour 100g qui se distingue des autres et qui peut être utilisé comme **“marqueur” de la catégorie**.

Analyses multivariées

- Beaucoup de colonnes dans le jeu de données

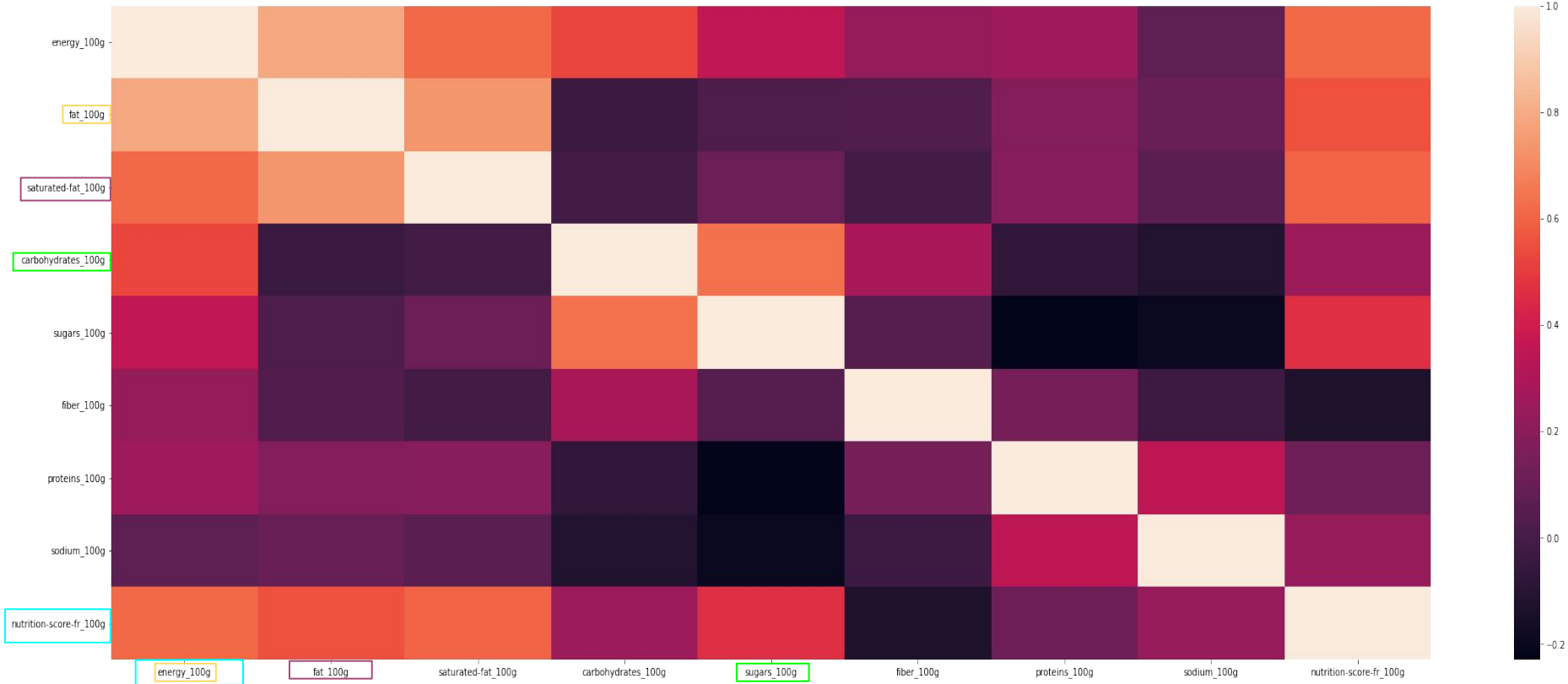
- Analyse des liens/relations de toutes ces variables



Matrice de corrélation

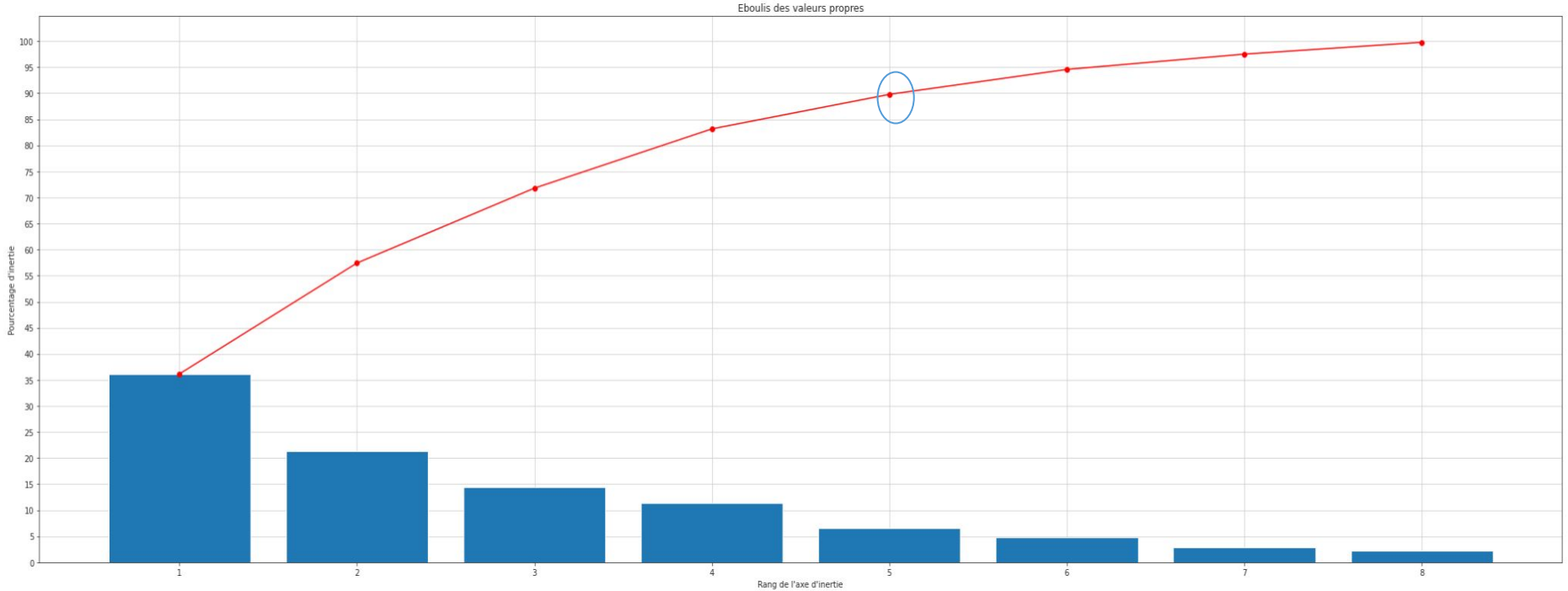
Analyse en composantes principales

2-e Matrice des corrélations



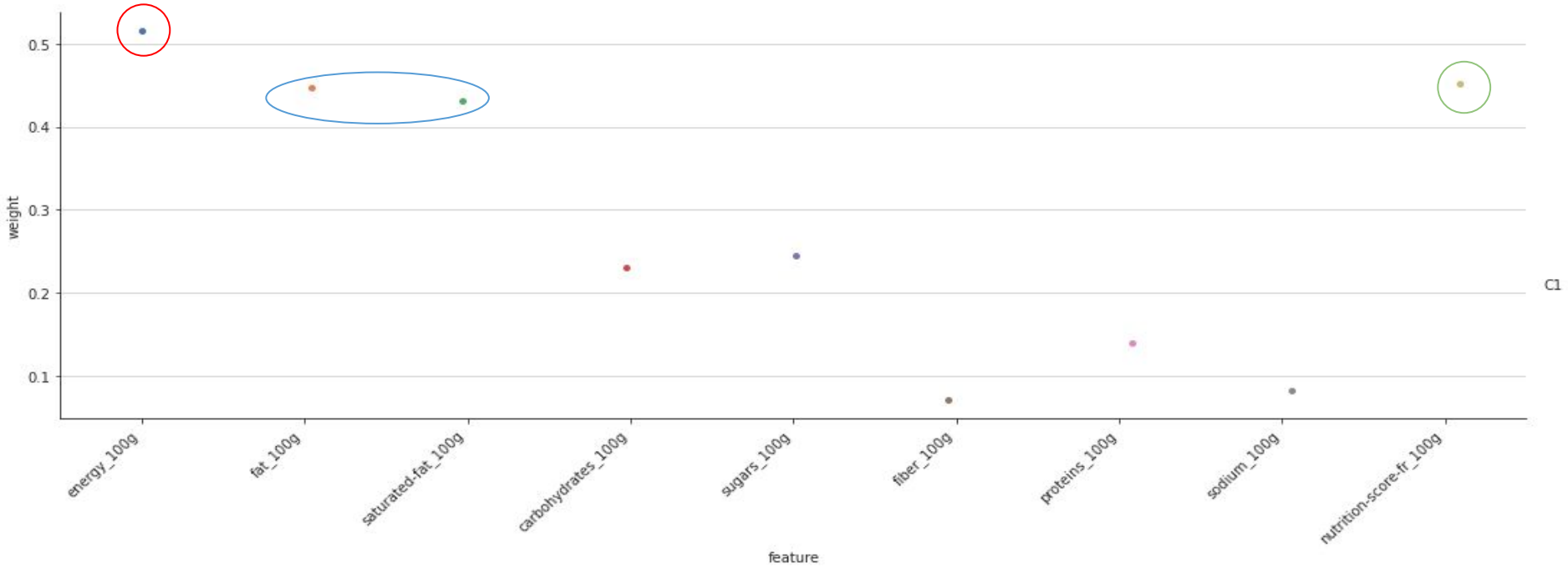
2-f Analyse en composantes principales

Réduction de dimensions possible



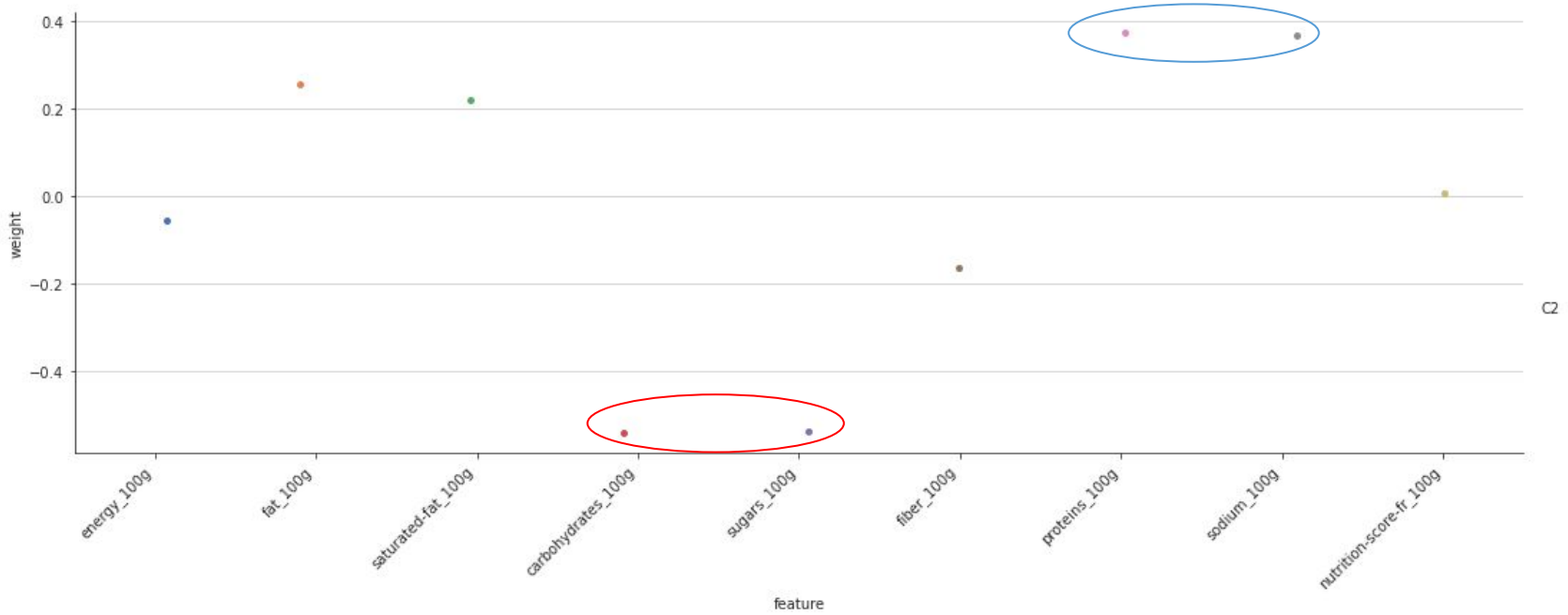
Analyse en composantes principales

- Poids des features dans la première composante



Analyse en composantes principales

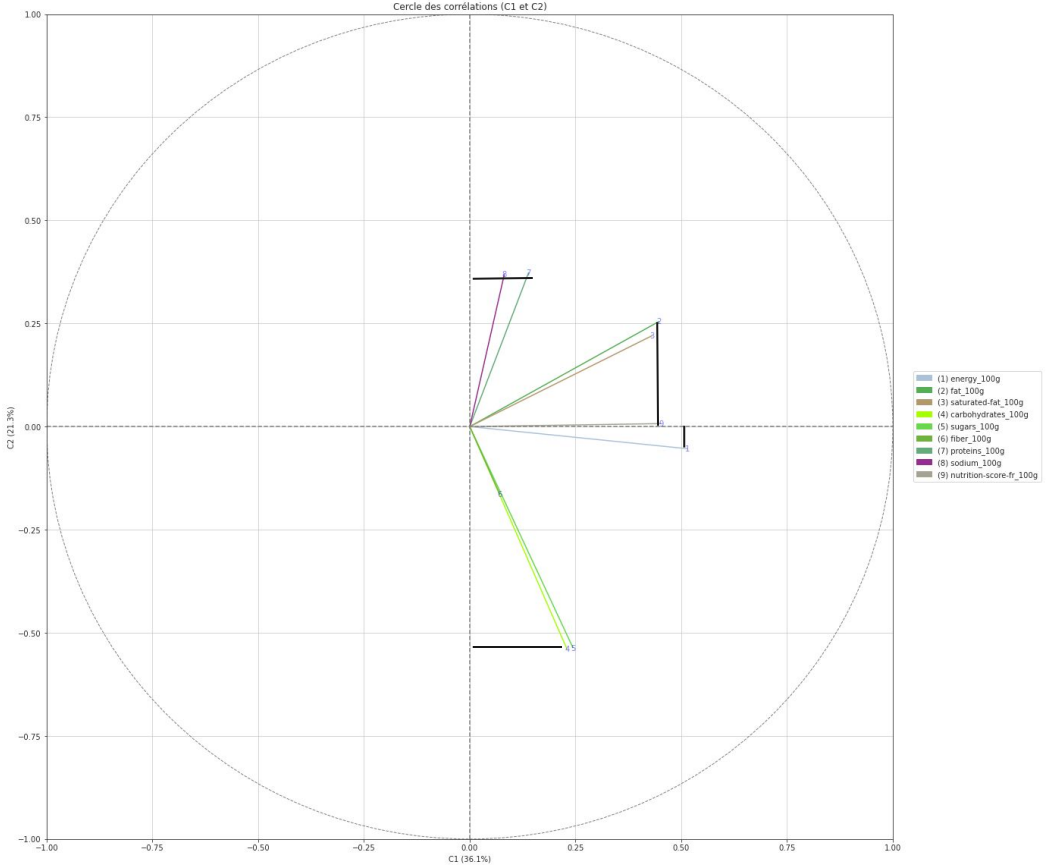
- Poids des features dans la deuxième composante



Analyse en composantes principales

- Corrélations entre:
  - fat et saturated-fat
  - sugar et carbohydrates
  - energy et nutri-score

- Projections sur les axes C1 et C2 permettent de retrouver les résultats concernant le poids des features dans les composants.



Conclusion

- Conclusion: Application de recommandation possible sur la base d'un algorithme de type KNN
    - analyses uni et bi-variées et les tests statistiques ont permis de démontrer que les composants étaient des marqueurs des produits.

- Principe de mise en oeuvre
    - Basé sur un exemple de système de Recommandation de livres
    - Transformation à appliquer au dataset:
        - on ne conserve que les colonnes correspondant aux apports pour 100g
        - transformation de l'identifiant des produits en index

Annexe: plans notebook

Plan complet du notebook de nettoyage

1-Travail préliminaire

- Chargement et aperçu des données
- Taux de remplissage
- Principe de l'application et sélection des colonnes utiles
  - informations générales sur les produits
  - informations sur la catégorie des produits
  - informations sur la composition et nutritives des produits

2-Nettoyage des données

- 2-1 Nettoyage des colonnes:
  - Principes pour la suppression des colonnes
  - Suppression de colonnes avec information dupliquée
  - Suppression de colonnes vides
- 2-2 Nettoyage des lignes:
  - Suppression des lignes dupliquées
  - Suppression des lignes erronées
  - Suppression des lignes sans informations nutritives/de composition
- 2-3 Traitement des données aberrantes (en dehors du domaine de valeurs)
  - quantités négatives
  - quantités pour 100g supérieures à 100 g
  - apport énergétique pour 100g supérieur à 3766.5 KJ
  - nutri score francais inférieure à -15 ou supérieure à 40
  - nutri grade différent de {a,b,c,d,e}
  - homogénéisation des catégories
- 2-4 Traitements des valeurs extrême ("outliers")
  - critère de filtrage des outliers (contribution à la variance)
  - filtrage des "outliers"
- 2-5 Traitements intermédiaires
  - Suppression de colonnes vides
  - Suppression des lignes sans informations nutritives/de composition
- 2-6 Imputation des données manquantes
  - imputation du sodium à partir du sel
  - calcul des nutri score/nutri grade manquants
  - imputation par la médiane par catégorie
- 2-7 Sauvegarde dataset nettoyé:
  - suppression des colonnes trop peu remplies
  - imputation par zéro

Plan complet du notebook d’analyse

1- Analyses univariées

- Histogrammes des catégories
- Analyses univariées composants alimentaires (histogrammes, boxplot)
- Analyse univariée - filtrage des données nulles
- Analyse Fibre - filtrage du max et des valeurs nulles
- Analyse Sodium - filtrage du max et des valeurs nulles

2- Analyses bivariées

- Tests de normalité (Shapiro) des composants alimentaires
- Test de Kruskal-Wallis)
- Affichage des distributions des variables en fonction des catégories
- Affichage des densités de distributions estimées par catégories
- Distribution estimée des fruits et légumes (filtrée 0)
-

3- Analyses multivariées des composants

- Matrice de corrélation
- Analyse en composantes principales
- PCA et projection (données pnns_groups_2)
- Eboulis des valeurs propres
- Poids des features dans chaque composant
- Affichage des cercles de corrélation