

Nettoyage et analyse de données pour l'alimentation d'une application mobile

“Il est où mon su-sucré ?”



Julien LE BOUCHER – février 2023



Plan

1. Genèse et fonctionnalités de l'application.
2. Préparation des données :
 - Le jeu de données d'Open Food Facts.
 - La procédure d'extraction et de nettoyage.
 - Le jeu de données résultant.
3. Exploration des données :
 - Analyse univariée.
 - Analyse multivariée.
 - Observations intéressantes

1. Genèse et fonctionnalités de l'application

Genèse



Un besoin de l'agence de santé publique :

Développer des applications pour sensibiliser les citoyens à une alimentation plus saine.



Ma proposition :

Aider les utilisateurs à détecter les produits inutilement trop sucrés pour se prémunir contre l'obésité, les maladies hépatiques, et les maladies cardiovasculaires, ou encore le diabète.



Fonctionnalité principale de l'application

scan du produit

catégorie peu
sucrée

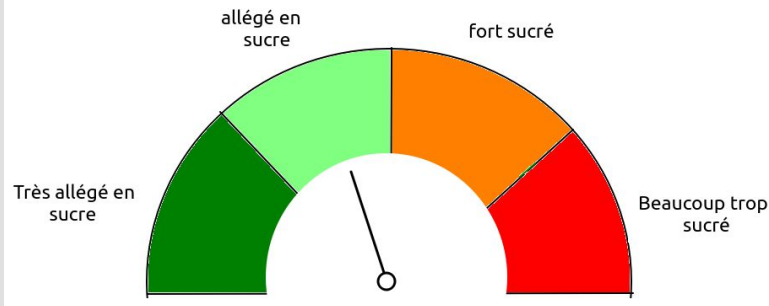
afficher
'catégorie de
produits peu
sucrés'

catégorie très sucrée

afficher 'Tous les
produits de cette
catégorie sont très
sucrés : à éviter ou
consommer avec
modération'

Catégorie très variable vis-à-vis du sucre

afficher la jauge suivante avec le niveau de
sucre du produit par rapport à sa catégorie



+ faire des recommandations de produits
moins sucrés et/ou avec un meilleur nutriscore
(dans la catégorie, voire dans une catégorie
moins spécifique, et si possible dans le
magasin)

Seconde fonctionnalité de l'application

Indiquer :

- La quantité de sucre dans le produit
- Le **pourcentage** de l'apport d'une **ration du produit** par rapport à l'apport journalier maximum recommandé par l'ANSES*.

- 24 g de sucre pour 100 g
- Une ration de ce produit (30 g) représente 7,2 % de l'apport quotidien à ne pas dépasser en sucre .

* : Agence nationale de sécurité sanitaire de l'alimentation de l'environnement et du travail

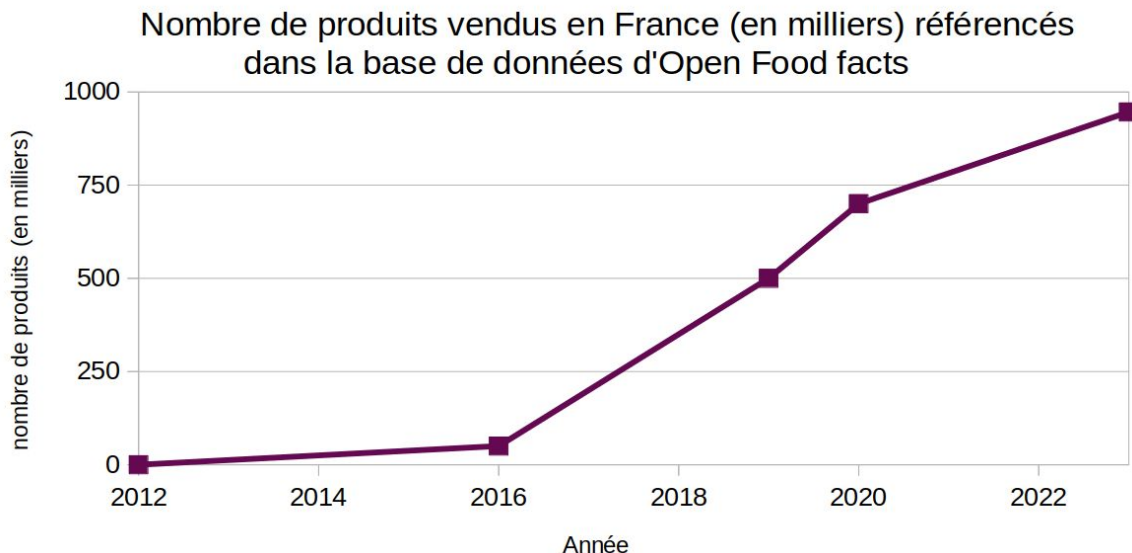
2 - Préparation des données pour l'application

Objectif : avoir un maximum de données fiables concernant la teneur en sucre des produits.

2.1 - La base de données d'Open Food Facts

La base de données d'Open Food Facts

- lancée en 2012 – Pierre Slamiche et Stéphane Gigandet
- mondiale, libre et participative.
- Objectifs majeurs :
 - **nutri-score** (2015)
 - **groupe NOVA** (2018)
 - **eco-score** (2021)



- Actuellement : 2 600 000 produits référencés dans le monde (196 variables mesurées/calculées).

2.2 - Extraction et nettoyage des données

25 variables retenues pour l'application

Pour :

- classer les produits ;
- évaluer la qualité des produits (nutri-score) ;
- connaître les valeurs nutritives pour 100 g de produit ;
- récupérer des images et créer un visuel pour l'application ;
- obtenir des informations complémentaires
 - magasins de vente
 - labels (bio, sans gluten...)
 - masse unitaire de consommation (30g pour une barre chocolatée)

Les étapes du processus d'extraction et de nettoyage des données

- Extraire les produits vendus en France.
- Gérer les duplicatas (code-barres en double).
- Préparer la classification des produits en groupes plus ou moins spécifiques :
 - récupérer de l'information dans certaines variables
 - formater le nom des catégories
- Étiqueter les produits à problème(s). (-> score de fiabilité pour l'app designer)
- Effacer les valeurs aberrantes.
- Corriger les valeurs d'énergie renseignées dans la mauvaise unité.
- Former des groupes et calculer leurs statistiques sur le sucre.
- Imputer les valeurs de sucre dans les groupes grâce aux statistiques.
- Imputer les autres nutriments.
- Nettoyer après imputation.
- Mettre à jour les liens entre variables après imputation.

946 000 produits bruts → 725 000 produits satisfaisants.

Les étapes du processus d'extraction et de nettoyage des données

- Extraire les produits vendus en France.
- Gérer les duplicatas (code-barres en double).
- Préparer la classification des produits en groupes plus ou moins spécifiques :
 - récupérer de l'information dans certaines variables
 - formater le nom des catégories
- **Étiqueter les produits à problème(s). (-> score de fiabilité pour l'app designer)**
- **Effacer les valeurs aberrantes.**
- Corriger les valeurs d'énergie renseignées dans la mauvaise unité.
- Former des groupes et calculer leurs statistiques sur le sucre.
- Imputer les valeurs de sucre dans les groupes grâce aux statistiques.
- Imputer les autres nutriments.
- Nettoyer après imputation.
- Mettre à jour les liens entre variables après imputation.

946 000 produits bruts → 725 000 produits satisfaisants.

Détection des produits aberrants grâce aux nutriments

Vérification de la cohérence des valeurs des nutriments par 100 g.

- Chaque masse de nutriment doit être :
 - inférieur à 100 g.
 - supérieur à 0 g.
- Graisses saturées < Lipides
- Sucre < Glucides
- Glucides + Lipides + Protéines + Sel < 100

liste des nutriments	
<i>Français</i>	<i>Anglais</i>
Lipides	fat
gras saturés	saturated fat
glucides	carbohydrates
sucre	sugars
protéines	proteins
fibres	fiber
sel	salt

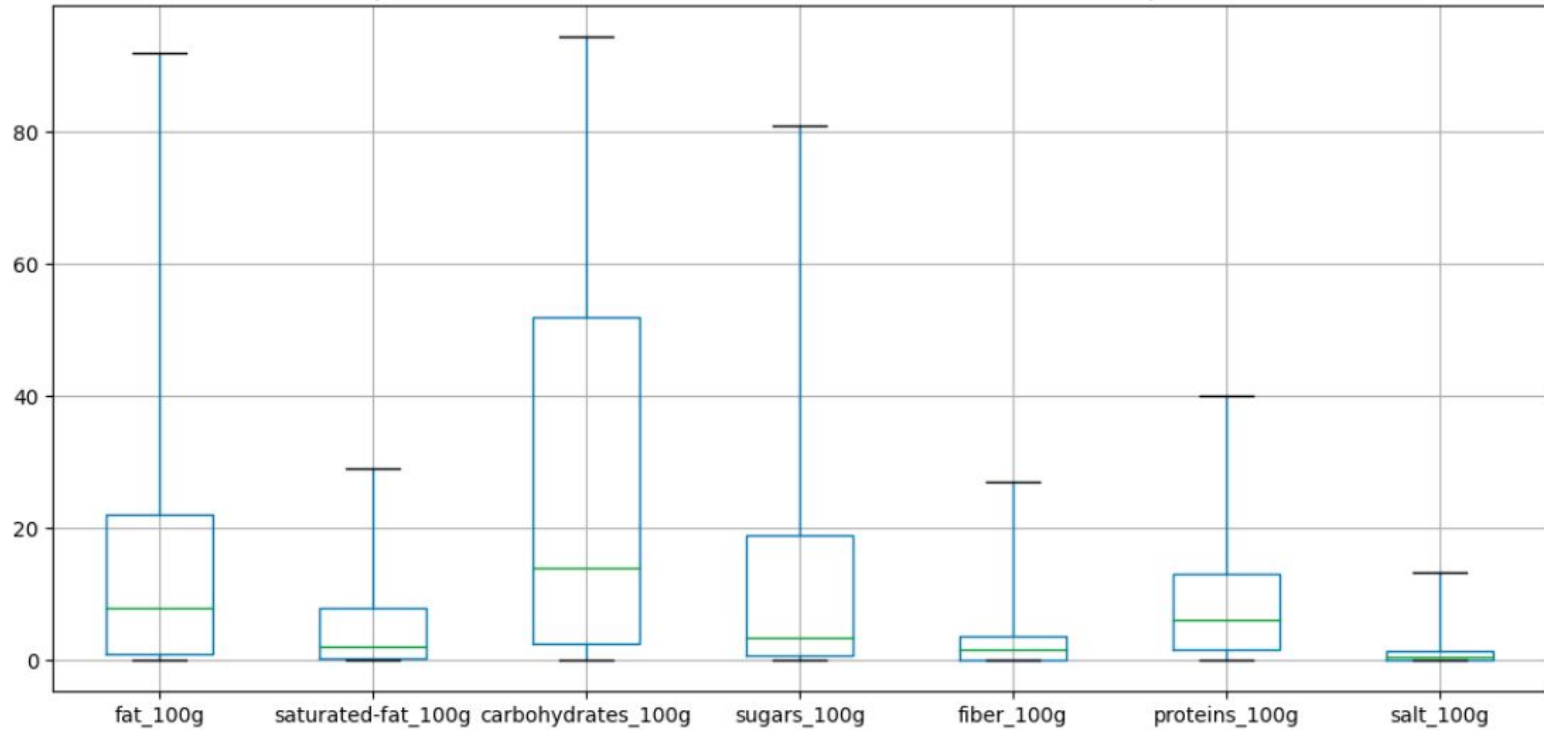
Pratiquement, une petite tolérance de dépassement de certains seuils a été implémentée.

Statistiques et visualisations des valeurs aberrantes au regard des nutriments

Nombre et pourcentages des problèmes identifiés parmi les 946 102 produits vendus en France selon leur type :

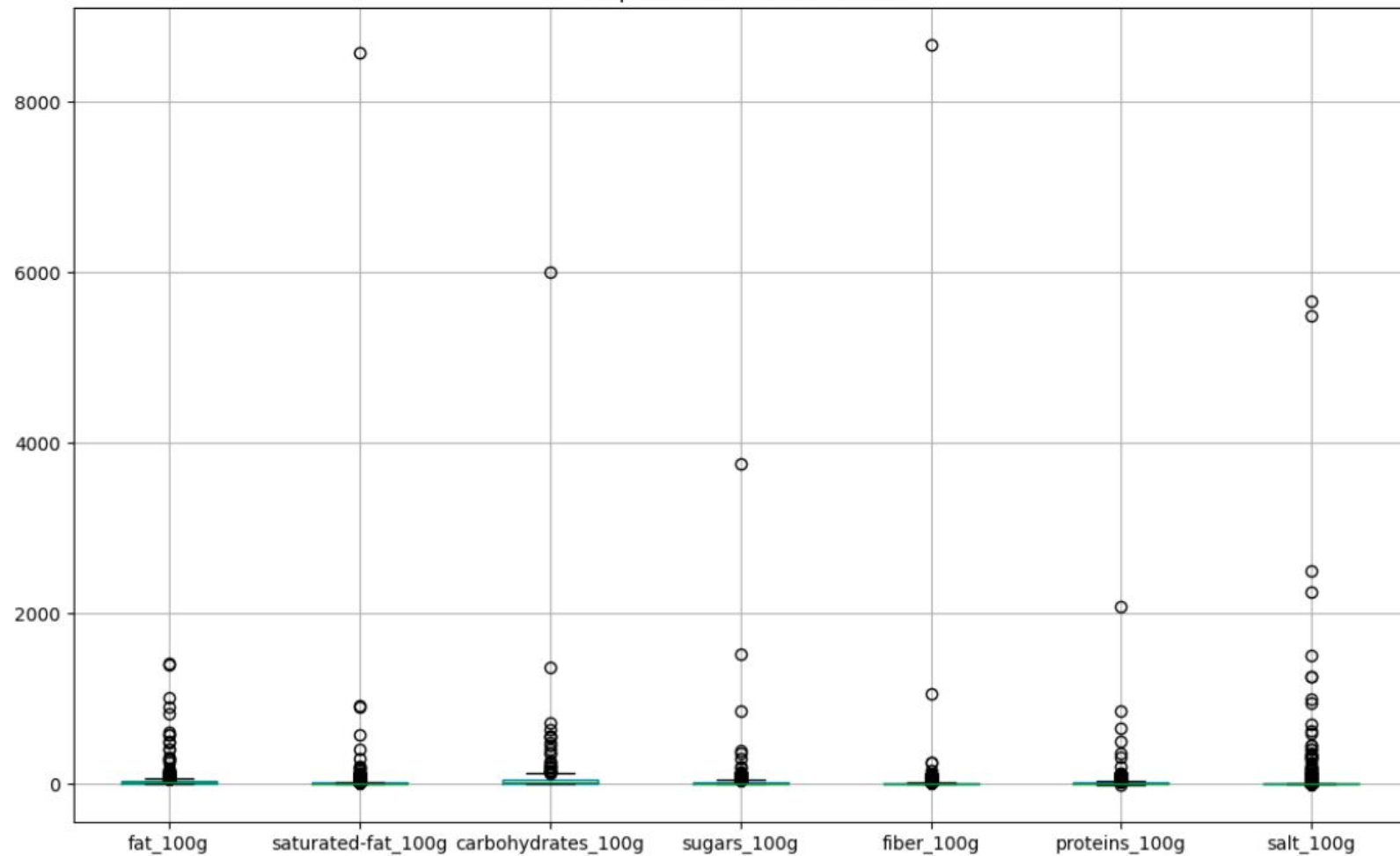
pbs	number	pct
initial_nutrients_problem	4977	0.526
more_saturated_fat_than_fat	341	0.036
more_sugars_than_carbohydrates	871	0.092
value_over_100	123	0.013
negative_value	6	0.001
no_nutrients_information	202197	21.372
nutrients_sum_over_101	3715	0.393

Boxplots of each nutrient with values within the first and the 99th percentile

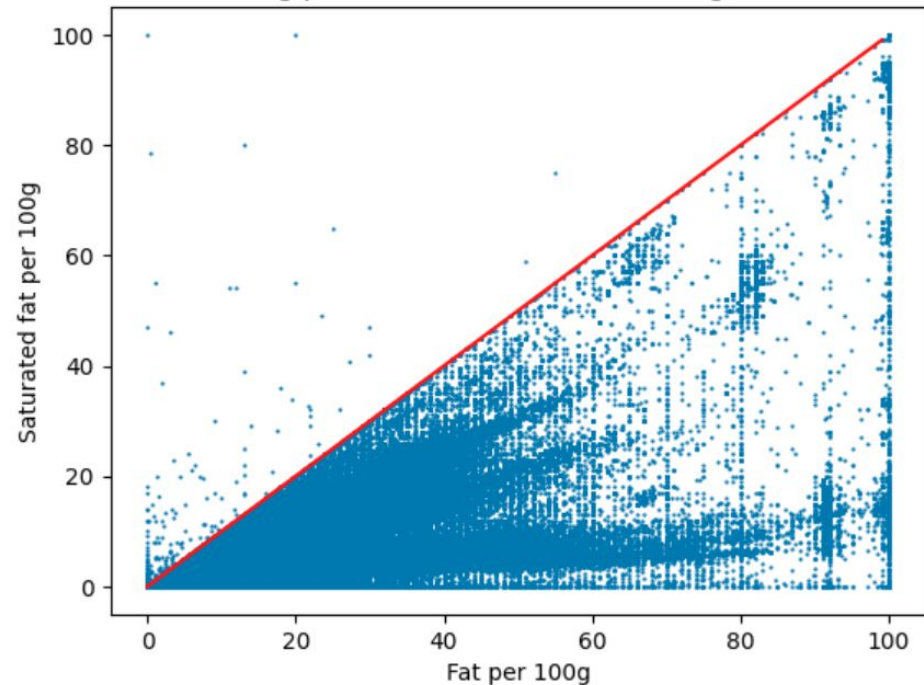


- Pas de problème si on ignore le premier et le dernier centile.
- Les lipides, les glucides et les sucres peuvent prendre des valeurs proches de 100. Pour les fibres et les protéines, ce n'est pas le cas.

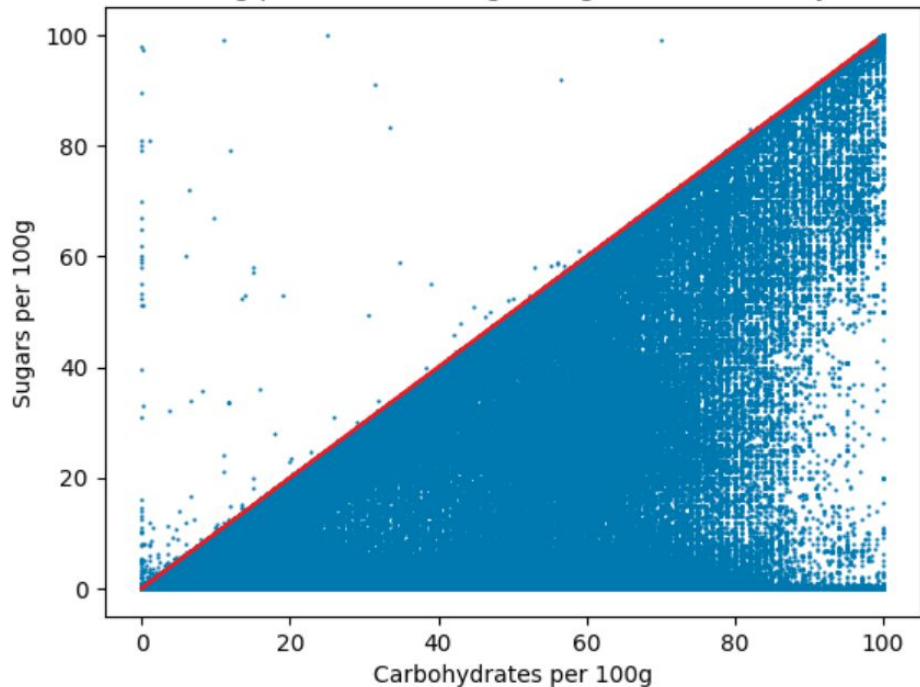
Boxplots of each raw nutrient



Visualizing products with saturated fat higher than fat



Visualizing products with sugars higher than carbohydrates



- Valeurs très éloignées de la ligne rouge considérées comme aberrantes.
- Petite tolérance des valeurs proches.

Seuils choisis pour discriminer les valeurs aberrantes

Minima et maxima choisis par nutriment :

```
extrema = {'fat_100g': (0, 100),  
           'saturated-fat_100g': (0, 100),  
           'carbohydrates_100g': (0, 100),  
           'sugars_100g': (0, 100),  
           'fiber_100g': (0, 60),  
           'proteins_100g': (0, 70),  
           'salt_100g': (0, 100)}
```

Tolérance de dépassement :

- du sucre et des graisses saturées par rapport à leur catégorie respectivement limitante (1 g)
- somme des nutriments (5 g)

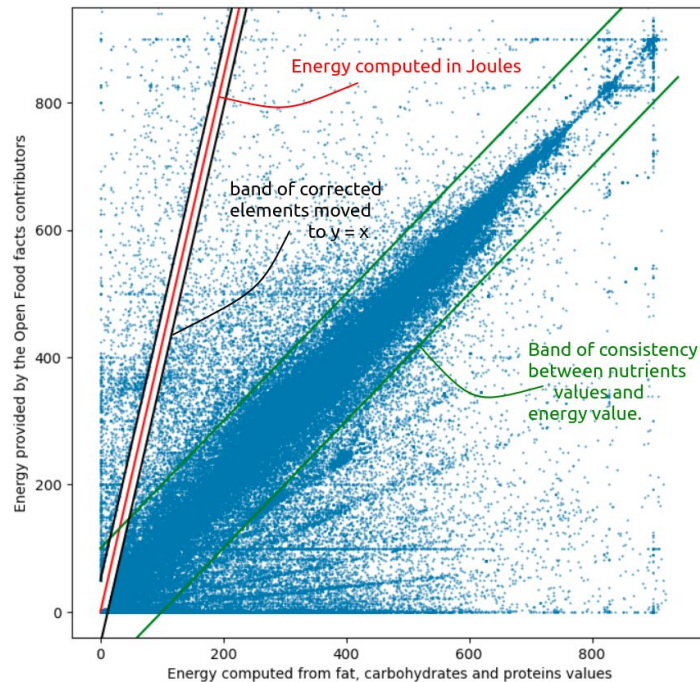
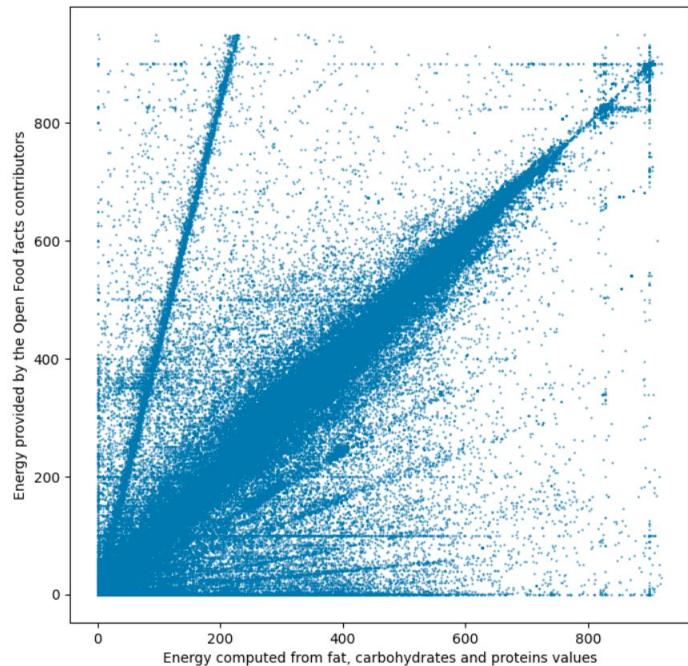
Les étapes du processus d'extraction et de nettoyage des données

- Extraire les produits vendus en France.
- Gérer les duplicatas (code-barres en double).
- Préparer la classification des produits en groupes plus ou moins spécifiques :
 - récupérer de l'information dans certaines variables
 - formater le nom des catégories
- Étiqueter les produits à problème(s). (-> score de fiabilité pour l'app designer)
- Effacer les valeurs aberrantes.
- **Corriger les valeurs d'énergie renseignées dans la mauvaise unité.**
- Former des groupes et calculer leurs statistiques sur le sucre.
- Imputer les valeurs de sucre dans les groupes grâce aux statistiques.
- Imputer les autres nutriments.
- Nettoyer après imputation.
- Mettre à jour les liens entre variables après imputation.

946 000 produits bruts → 725 000 produits satisfaisants.

Correction de l'énergie renseignée en Joules au lieu de kcal

$$\text{Energy-kcal} = 9 \times m_{fat} + 4 \times m_{carbohydrates} + 4 \times m_{proteins}$$



Les étapes du processus d'extraction et de nettoyage des données

- Extraire les produits vendus en France.
- Gérer les duplicatas (code-barres en double).
- Préparer la classification des produits en groupes plus ou moins spécifiques :
 - récupérer de l'information dans certaines variables
 - formater le nom des catégories
- Étiqueter les produits à problème(s). (-> score de fiabilité pour l'app designer)
- Effacer les valeurs aberrantes.
- Corriger les valeurs d'énergie renseignées dans la mauvaise unité.
- **Former des groupes et calculer leurs statistiques sur le sucre.**
- Imputer les valeurs de sucre dans les groupes grâce aux statistiques.
- Imputer les autres nutriments.
- Nettoyer après imputation.
- Mettre à jour les liens entre variables après imputation.

946 000 produits bruts → 725 000 produits satisfaisants.

Les groupes de classification utilisés pour l'imputation

- Non disponibles directement dans la base de données initiale.
- Proviennent d'information récupérée et structurée en 3 couches plus ou moins spécifiques.

Critère de création d'un groupe :

- au moins 3 produits semblables ;
- écart-type calculé sur les valeurs de sucre pour 100 g < 26 g.

- Obtention de 16 145 groupes composés de 3 à 8 000 éléments
- Disponibles pour 51 % des produits

Intérêt par rapport aux groupes du PNNS (plan national nutrition santé) :

- Plus spécifiques
- Information sur un plus grand nombre de produits (5 points en plus).

→ Préférables pour des imputations plus réalistes et plus nombreuses.

Les étapes du processus d'extraction et de nettoyage des données

- Extraire les produits vendus en France.
- Gérer les duplicatas (code-barres en double).
- Préparer la classification des produits en groupes plus ou moins spécifiques :
 - récupérer de l'information dans certaines variables
 - formater le nom des catégories
- Étiqueter les produits à problème(s). (-> score de fiabilité pour l'app designer)
- Effacer les valeurs aberrantes.
- Corriger les valeurs d'énergie renseignées dans la mauvaise unité.
- Former des groupes et calculer leurs statistiques sur le sucre.
- **Imputer les valeurs de sucre dans les groupes grâce aux statistiques.**
- **Imputer les autres nutriments.**
- Nettoyer après imputation.
- Mettre à jour les liens entre variables après imputation.

946 000 produits bruts → 725 000 produits satisfaisants.

Les différentes méthodes d'imputations utilisées

Rappel : avant toute imputation, les valeurs aberrantes ont été rendues inconnues (ou le produit a été éliminé s'il cumulait trop d'aberrations).

Pour les produits ayant :	Méthode imputation
un groupe & une valeur de sucre inconnue	moyenne du groupe
une valeur de fibre inconnue	constante : 0
un groupe & un ou plusieurs nutriments inconnus (mais au moins 3 connus sur les 7).	KNN Imputer
une ou deux valeurs inconnues & aucun groupe	constante : 0

2.3 - Le jeu de données résultant

Bilan :

- 946 000 produits → 725 000 produits.
- Tous valident les règles de cohérence des nutriments édictées plus tôt.
- 16 125 produits avec une incohérence (nutriments ; énergie) étiquetés.
- 76 % des produits avec un problème initial corrigés.

pbs	number
-----	-----
initial_nutrients_problem	3772
more_saturated_fat_than_fat	295
more_sugars_than_carbohydrates	815
value_over_100	61
negative_value	5
no_nutrients_information	0
nutrients_sum_over_101	2664

pas de solution
magique

nombre de produits corrigés, par problèmes initiaux, grâce à toutes les imputations.

fat_100g	6124
saturated-fat_100g	5609
carbohydrates_100g	6148
sugars_100g	0
fiber_100g	0
proteins_100g	5553
salt_100g	12629

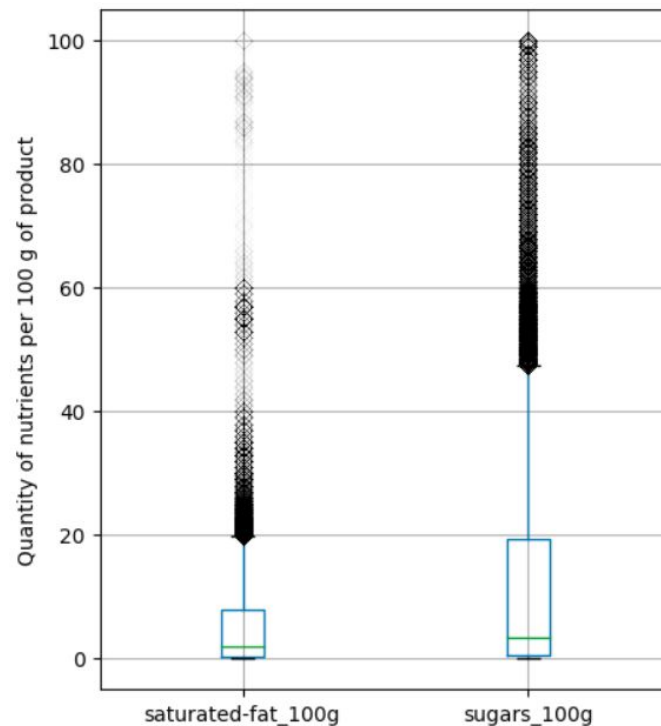
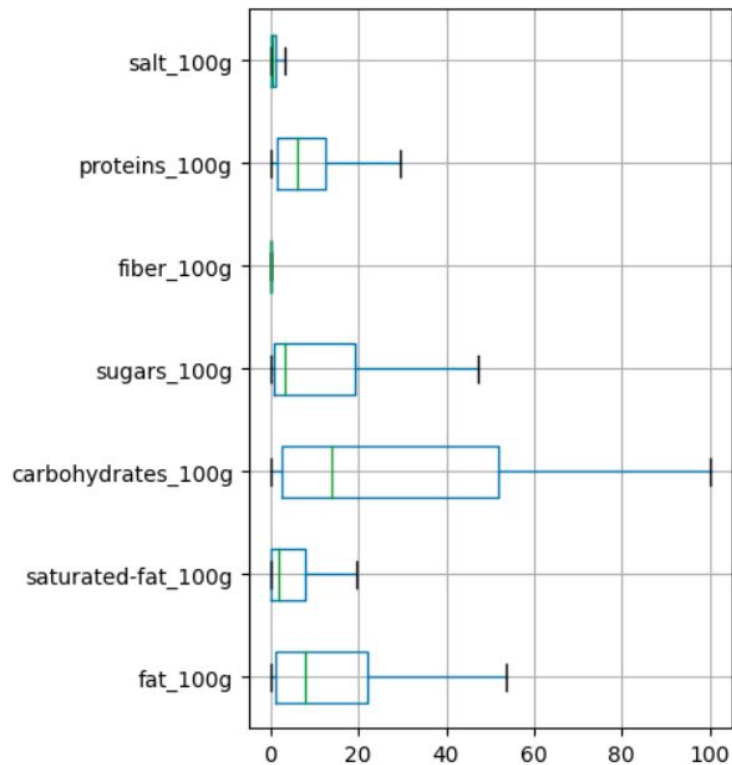
nombre de nutriments imputés grâce au KNN imputer, par type de nutriment.

3 – Analyse exploratoire

3.1 – Analyse univariée et faisabilité de l'application

Les nutriments






Disponibles pour l'intégralité du jeu de données

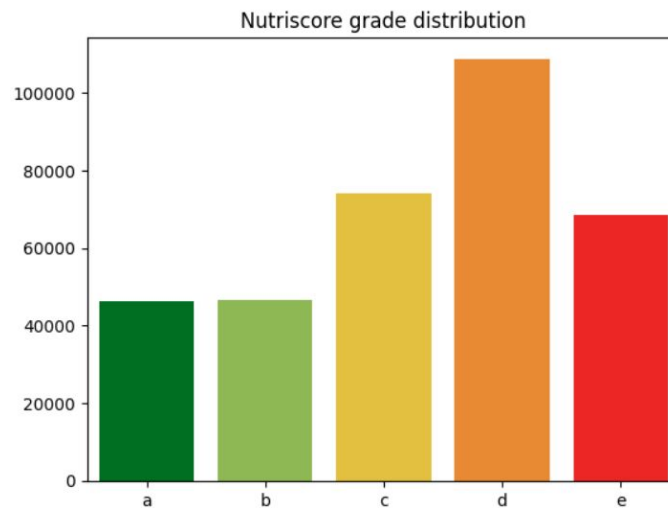
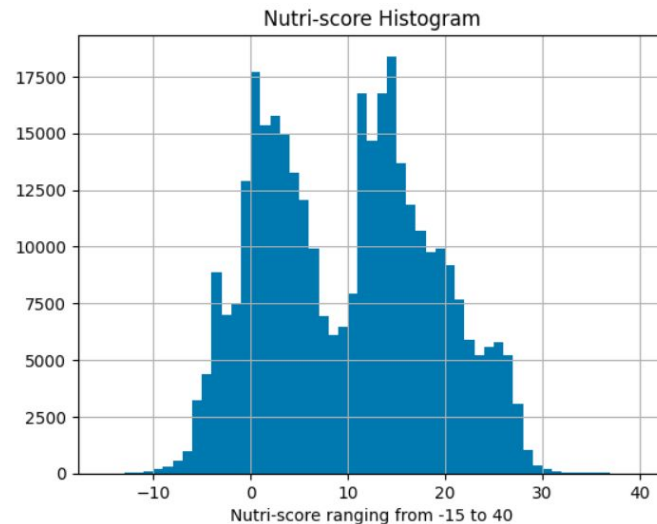


- Les éléments les plus malsains sont les graisses saturées, puis vient le sucre (cf. matrice de corrélation et scatterplot plus loin + articles de la communauté scientifique).
- Leurs distributions sont similaires en forme mais celle du sucre est beaucoup plus dilatée (taux plus élevés).
- Environ un quart des produits contenant du sucre en contiennent au moins 20 g pour 100g.

Le nutri-score

- Disponible pour 48 % du jeu de données
- Permet une graduation de la qualité d'un produit au regard de la santé.

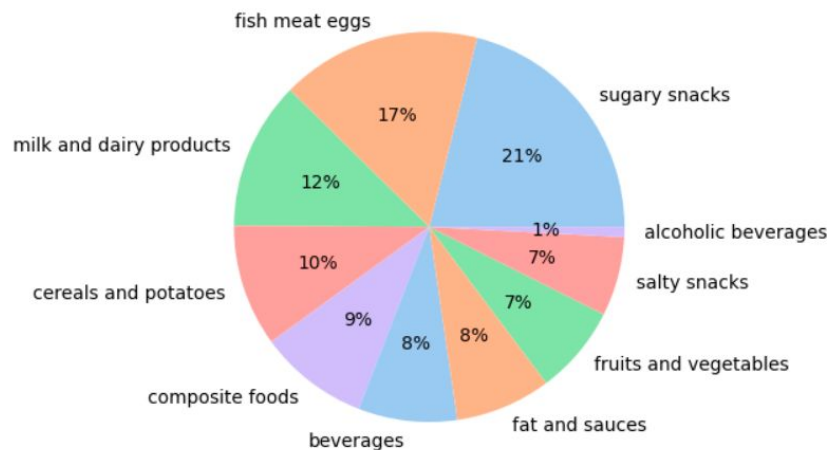
Points		Logo
Solid foods	Beverages	
Min to -1	Waters	
0 - 2	Min - 1	
3 - 10	2 - 5	
11 - 18	6 - 9	
19 - max	10 - max	



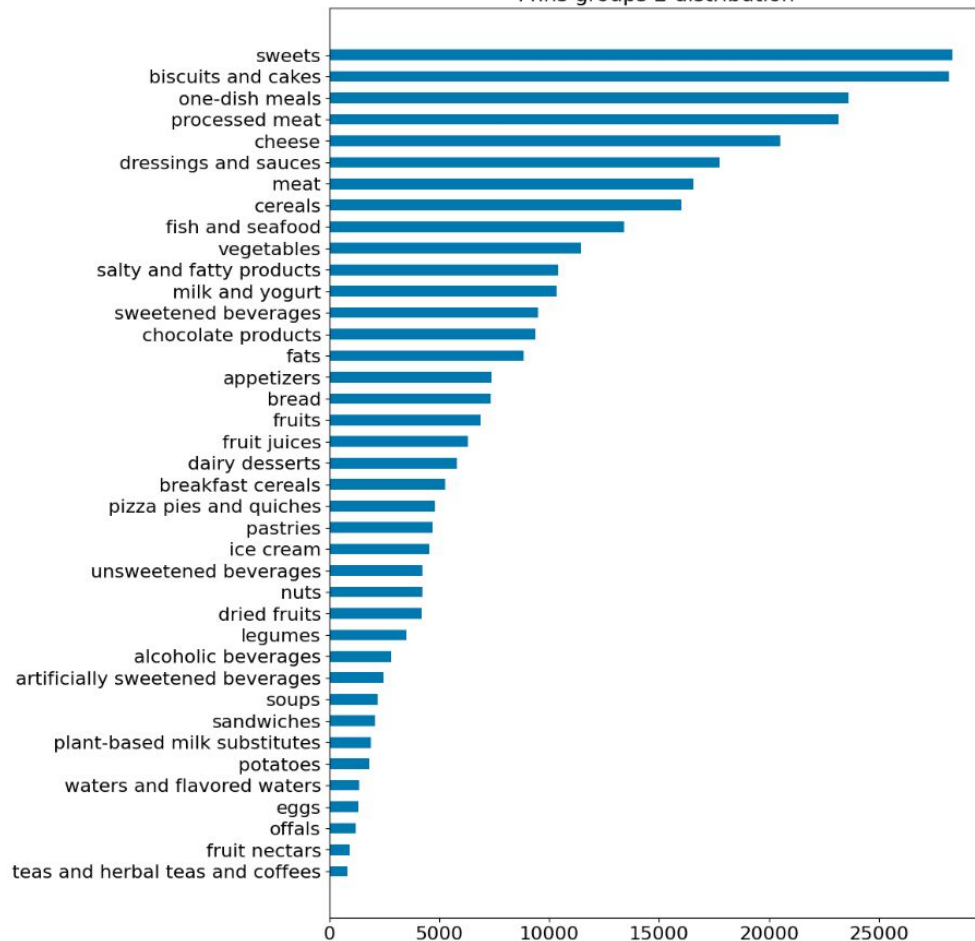
Les groupes PNNS

Disponibles pour 46 % des produits

pnns groups 1 distribution



Pnns groups 2 distribution



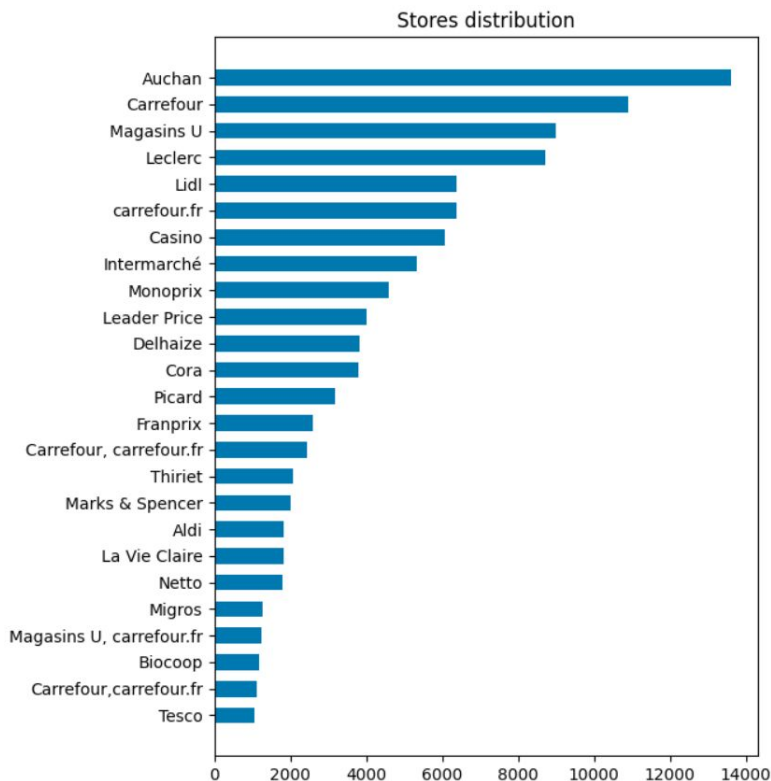
Bilan intermédiaire :

- Parmi les produits avec du sucre, beaucoup possèdent des taux élevés en sucre.
- Parmi les groupes les mieux représentés dans la base de données, beaucoup sont liés à des produits sucrés.
- Ainsi, même si certaines graisses malsaines sont 'pires' que le sucre, le sucre moins malsain est plus abondant. Il faut s'en méfier.

L'application est bien justifiée.

Les lieux de vente

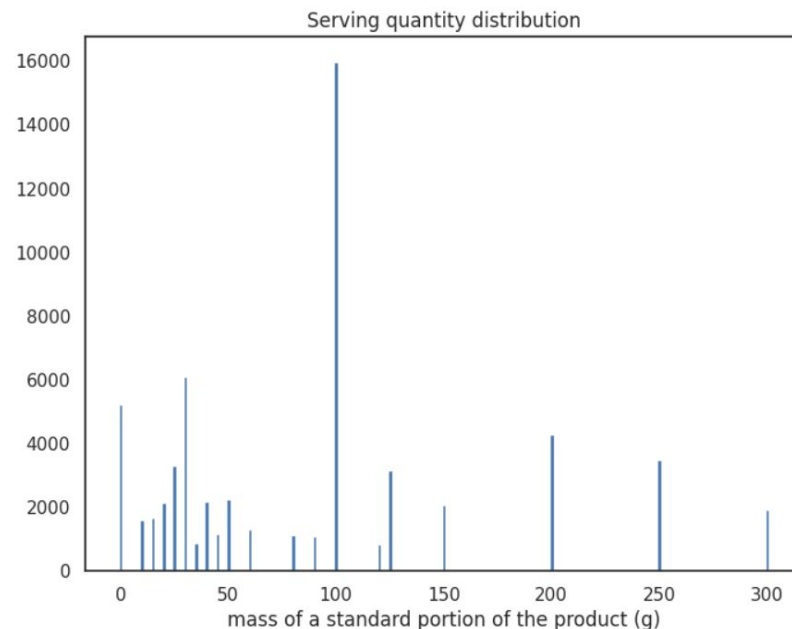
Disponible pour 21 % du jeu de données



Suggestion de produits plus sains dans le magasin envisageable.

Quantité de consommation unitaire

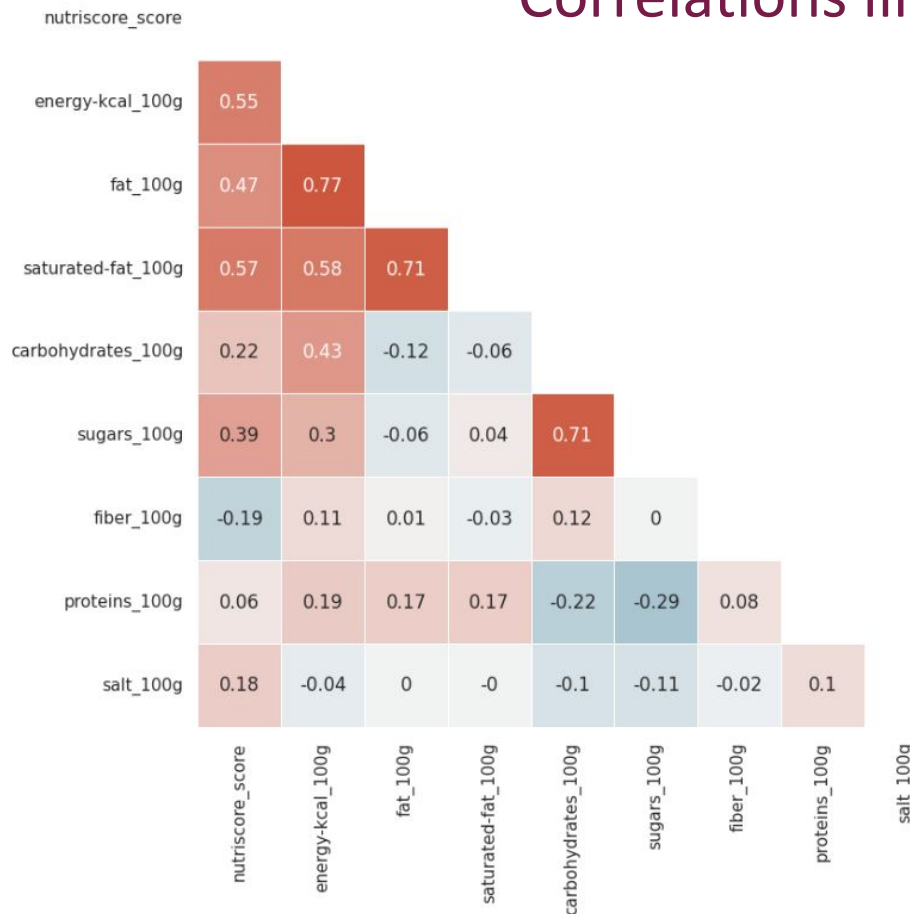
Disponible pour 12 % du jeu de données



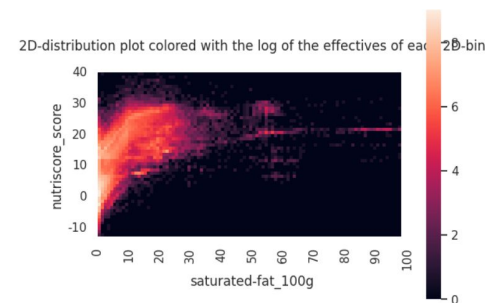
Le remplissage des données semble trop faible pour développer systématiquement le deuxième point de la seconde fonctionnalité de l'application.

3.2 – Analyse multivariée et faisabilité de l'application

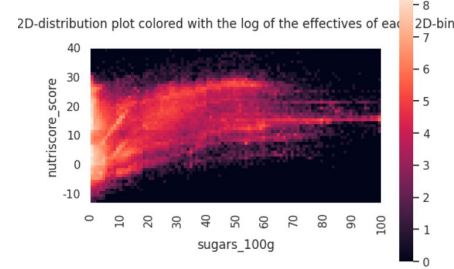
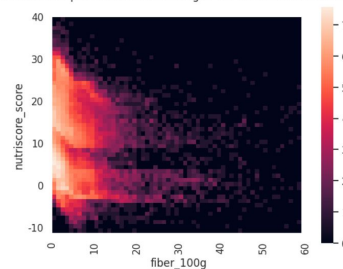
Corrélations linéaires



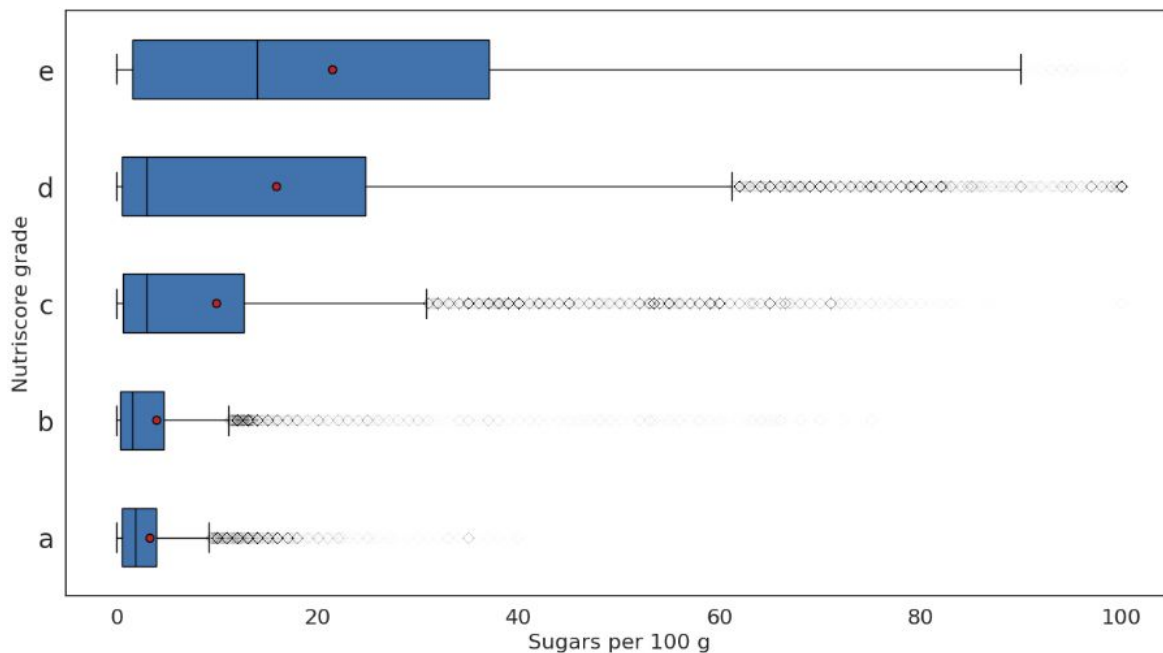
- Énergie très corrélée avec les lipides (cohérent formule).
- Sucres et graisses assez corrélés avec le nutri-score.
Plus il y en a, plus le nutri-score est grand et moins le produit est sain.
- Les fibres seraient le seul facteur qui influencerait plutôt positivement le nutri-score.



2D-distribution plot colored with the log of the effectiveness of each 2D-bin



Corrélation entre la quantité de sucre et les catégories nutri-score.



Coefficient de corrélation non-linéaire

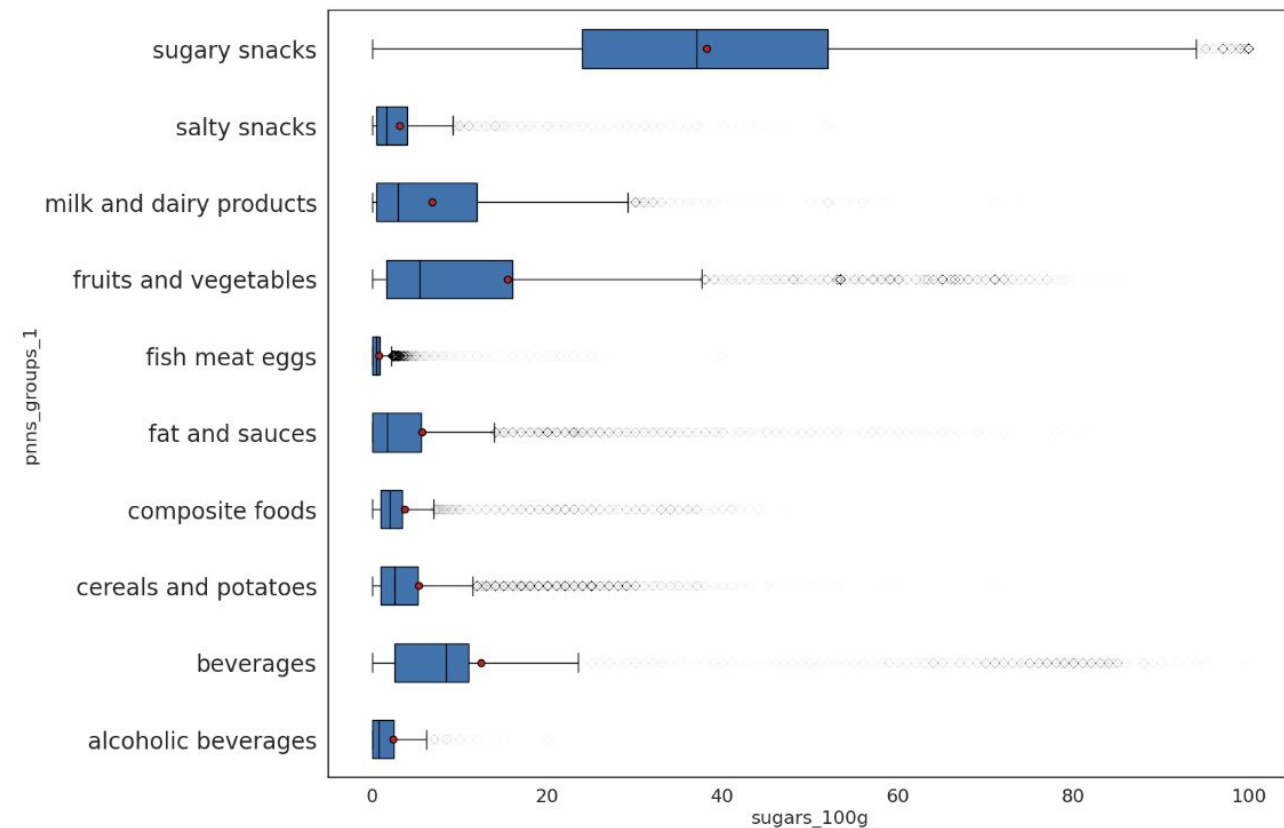
$$\eta^2 = 0.12$$

Règle de Cohen sur l'importance de la corrélation en fonction de la valeur du coefficient.

.01 ~ small
.06 ~ medium
>.14 ~ large

Les groupes construits à partir de la lettre nutri-score ont des moyennes de sucre **significativement** différentes. Confirmé par un test de Kruskal-Wallis (p-value ~ 0) car hypothèses d'application d'ANOVA non vérifiées.

Corrélation entre groupes PNNS 1 et quantité de sucre



$$\eta^2 = 0.5$$

Corrélation non-linéaire
très importante.

Corrélation entre groupes PNNS 2 et quantité de sucre

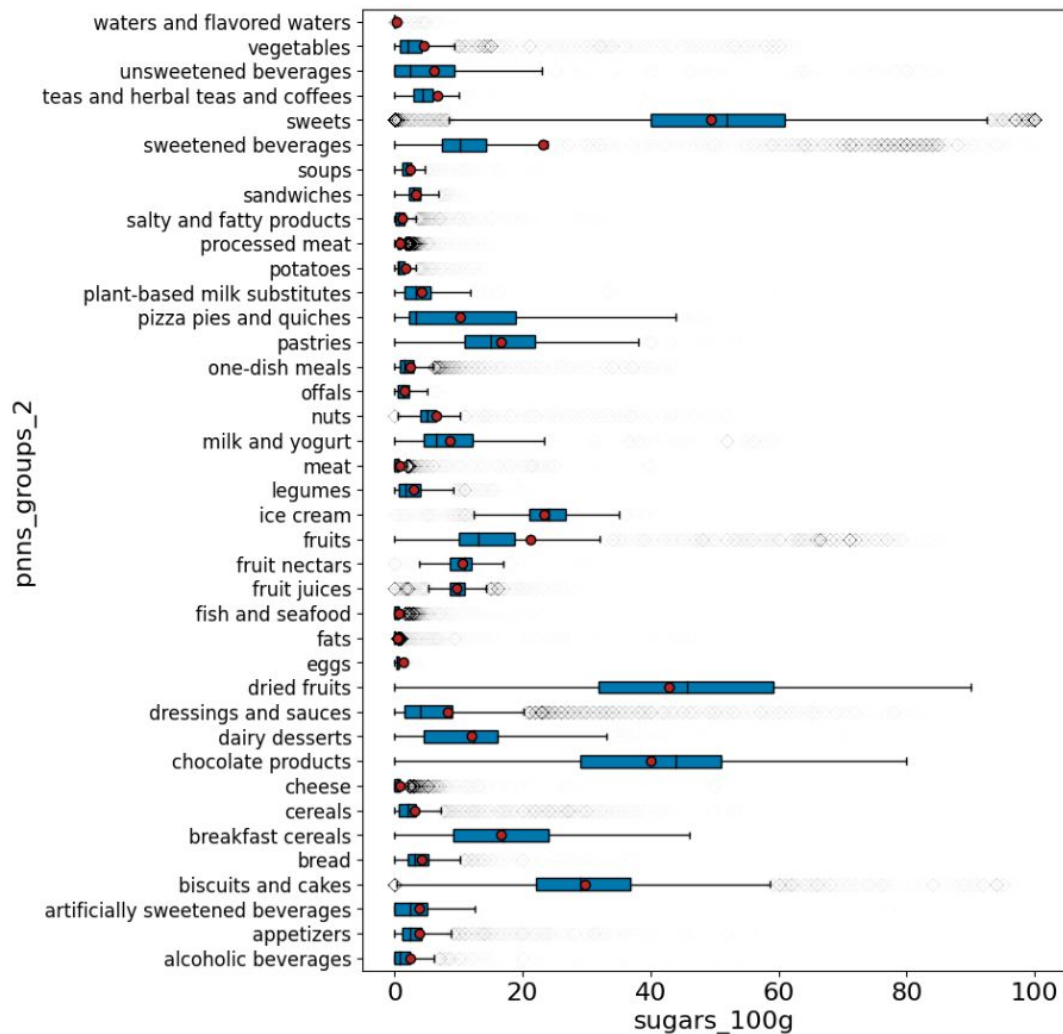
$$\eta^2 = 0.65$$

- Corrélation encore plus importante.
- Les groupes avec les plus grosses moyennes sont plus dispersés que ceux avec une moyenne basse.

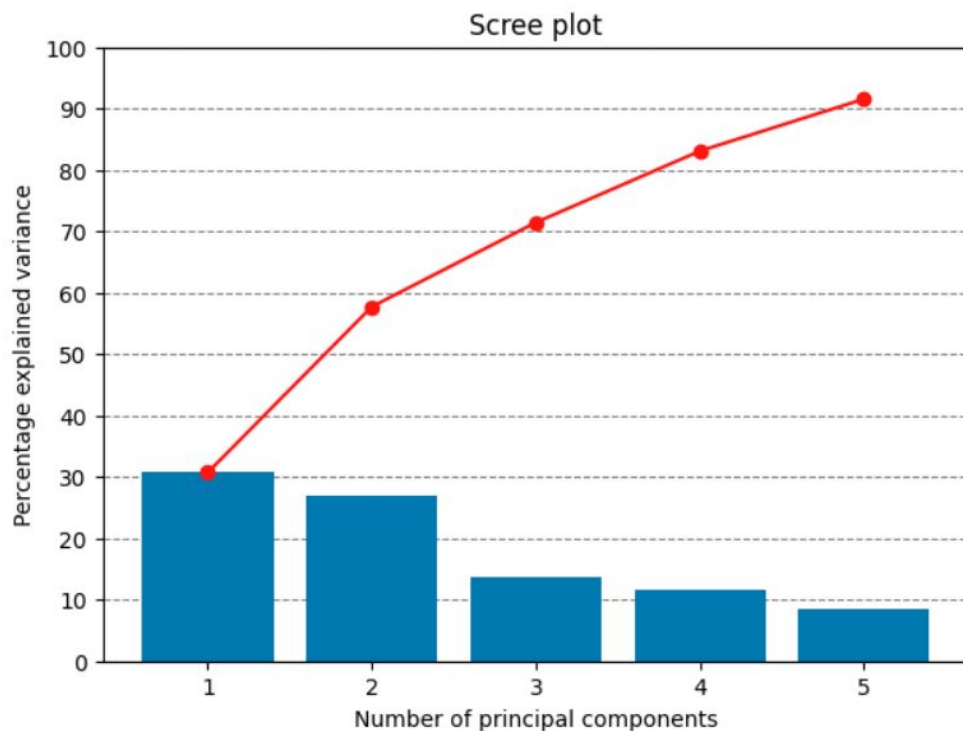
Ainsi, sur le schéma de la première fonctionnalité de l'application, on peut toujours être dans le cas 1 ou le cas 3. Un message de vigilance pourra être ajouté pour les groupes à très grande moyenne. Le cas 2 ne se présenterait que lors d'un travail sur des groupes encore plus spécifiques.

- Il semble étrange que le premier décile soit à 0 pour de nombreuses catégories. Cette division pourrait être un nouveau point de départ pour affiner le traitement des valeurs aberrantes.

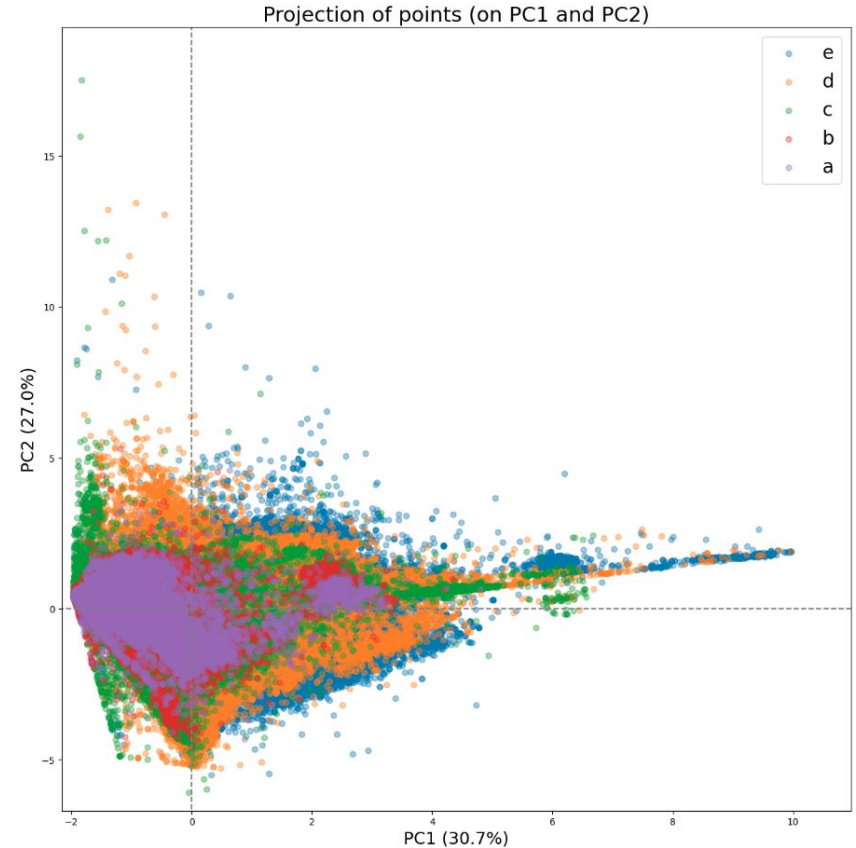
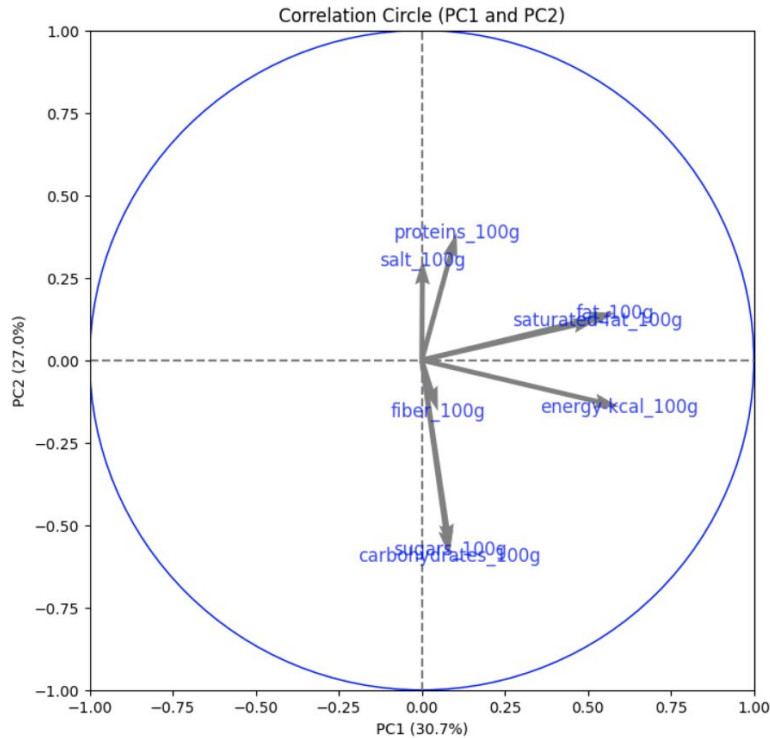
Attention pour les recommandations basses !



ACP sur les nutriments et l'énergie en kcal normalisés



Premier plan factoriel

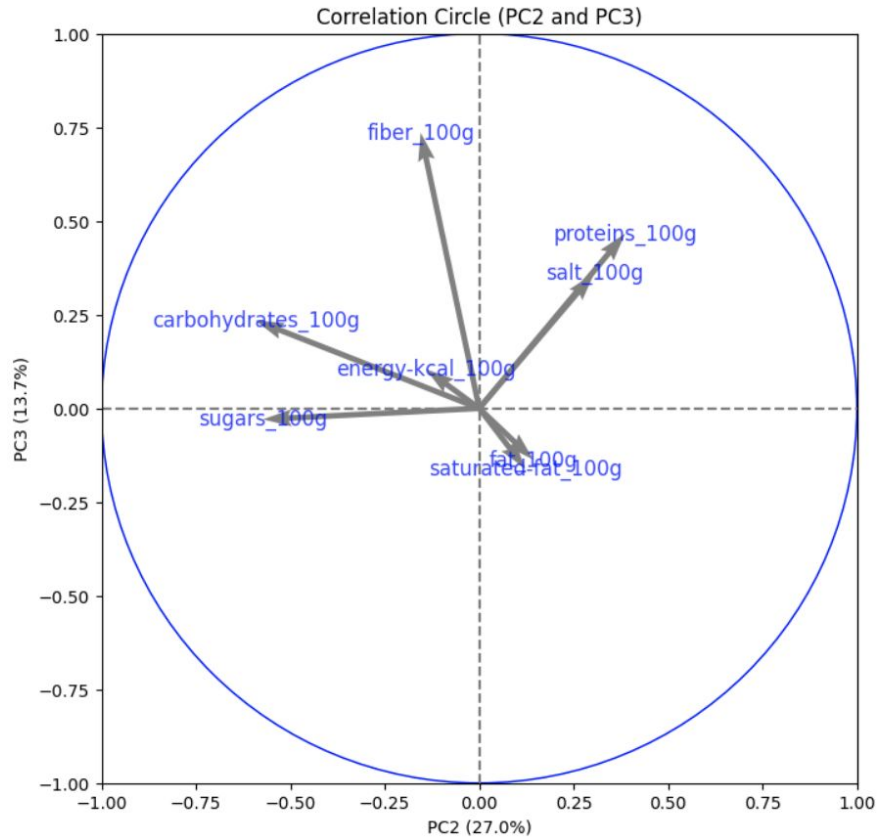


Variables moyennement représentées.

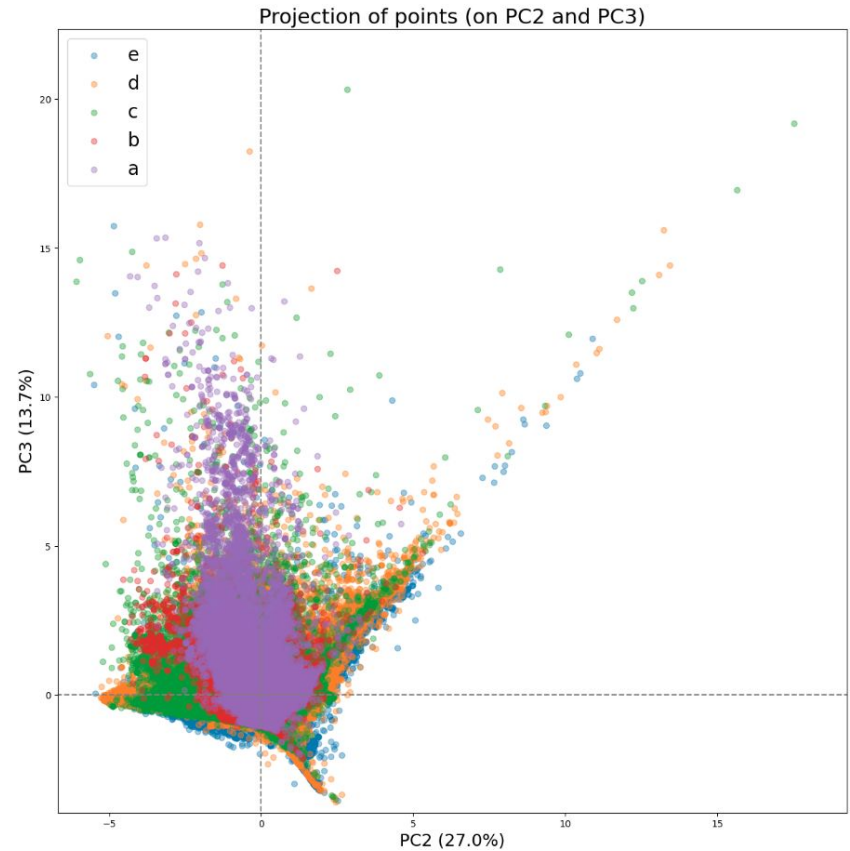
Axe 1 : Graisses et énergie.

Axe 2 : Glucides s'opposant aux protéines.

Deuxième plan factoriel

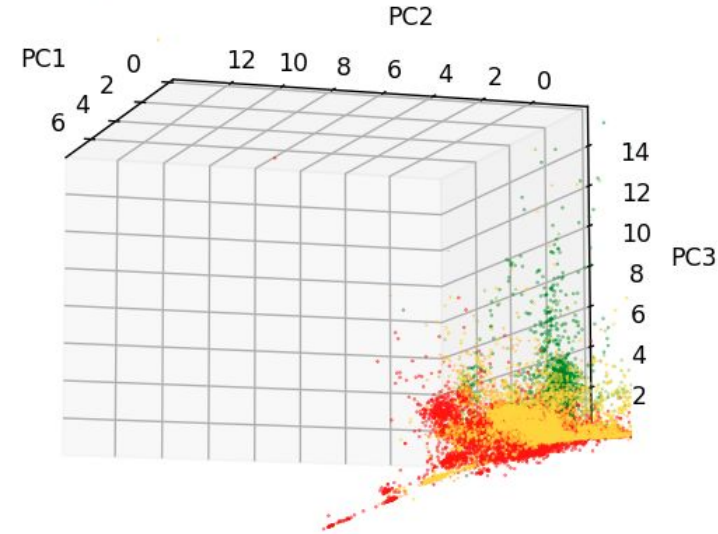
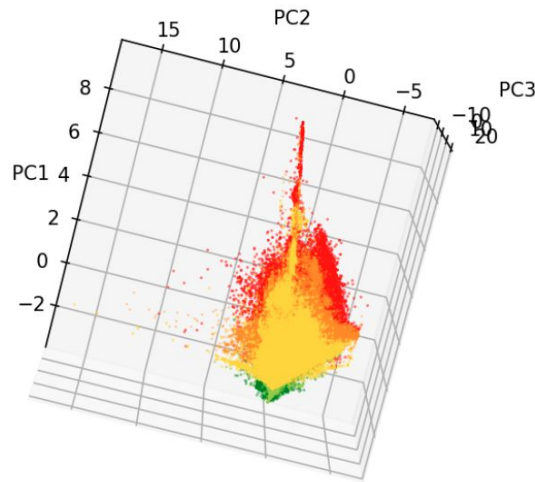
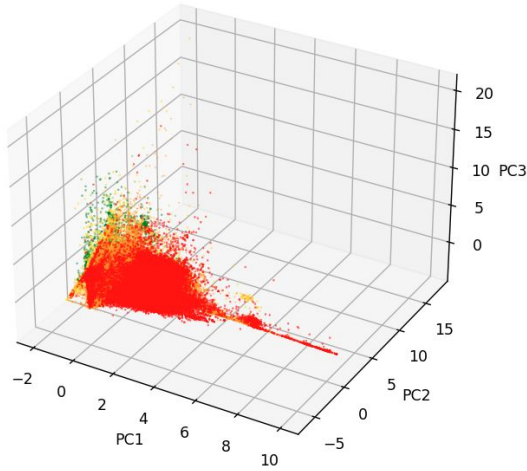


Axe 3 : fibre. Bien représentée.



Projection dans le premier espace factoriel 3D

différentes vues des produits colorisés avec leur couleur nutri-score.



seulement 25 000 produits
échantillonnés au hasard.