Présentation de la problématique

Objectif: Créer un **générateur de recettes saines**, *via la BDD OpenFoodFacts*.

Qu'est ce qu'une alimentation saine, un produit sain?

- Pas de produit "magique" sain auto-suffisant:

  *"**Seule une alimentation variée** en protéines, glucides, lipides, sels minéraux, fibres et vitamines, **constitue une "alimentation équilibrée"**. "Cet équilibre s'établit sur plusieurs repas, sur plusieurs jours."*

- Mais il existe des produits malsains:

  *"Il est conseillé de **ne pas consommer trop de produits gras, salés, et sucrés** (principales sources d'**énergie**)".*

- En terme d'énergie, il faut ~10 000 kJ par jour pour un Homme moyen.

Categories "parentes" d'aliments:

- **Lipides**: matières grasses, acides gras, stérols.
- **Glucides**: les carbohydrates, les sucres: fructose, glucose, lactose, saccharose, ainsi que les céréales, l'amidon des féculents...
- **Sels minéraux**: le calcium, le sodium, le fer, etc.
- **Protides**: protéines, polypeptides et acides aminées.
- **Fibres**
- **Vitamines**

Informations nutritionnelles

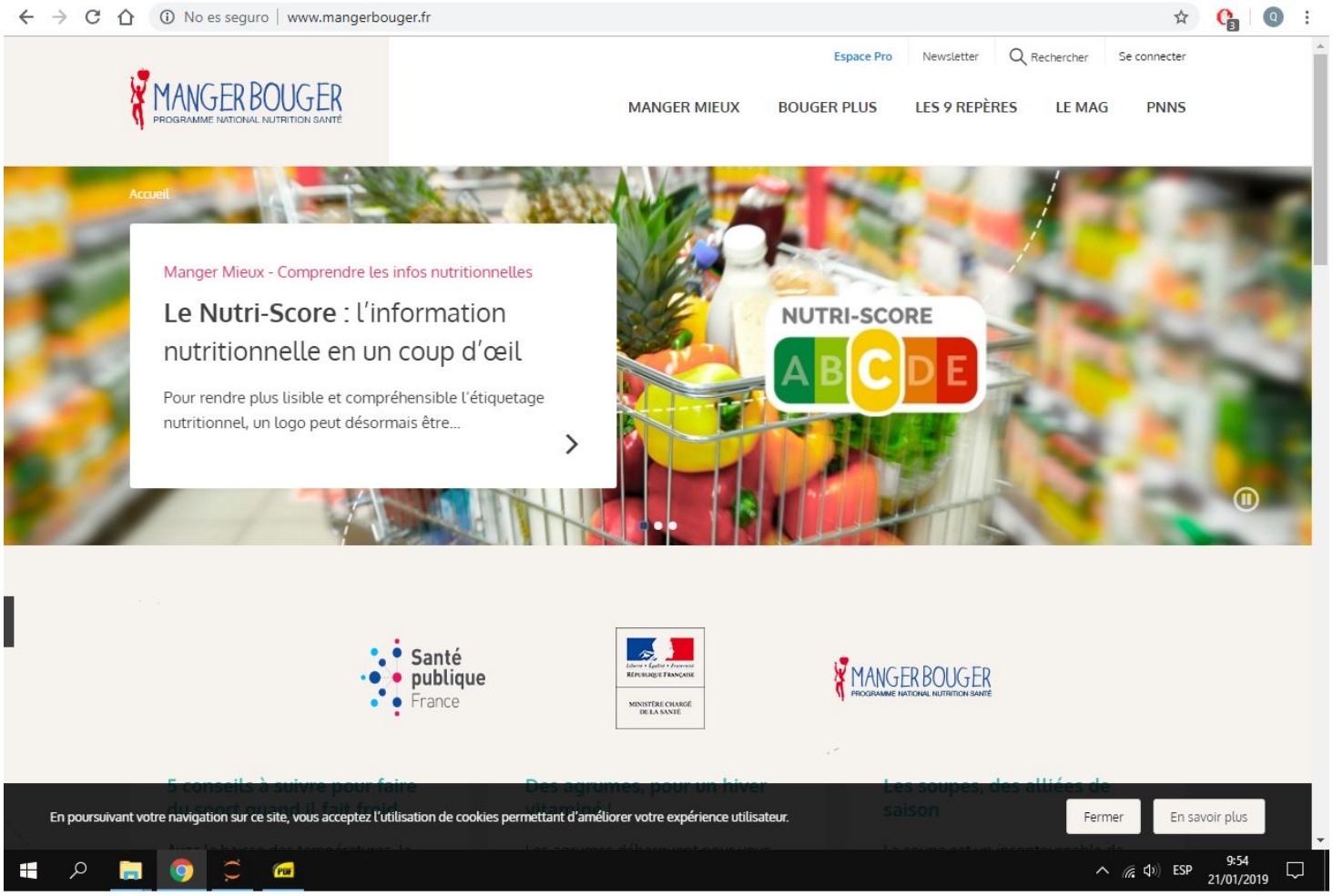
	Pour 100 g	Par moelleux
Valeur énergétique	461 kcal (1923 kJ)	92 kcal (385 kJ)
Matières grasses	27,3 g	5,5 g
dont acides gras saturés	2,8 g	0,56 g
Glucides	38,7 g	7,7 g
dont sucres	29,1 g	5,8 g
Protéines	15 g	3 g
Sel	0,1 g	0,02 g
Vitamine E	9 mg (75 % VNR*)	1,8 mg (15 % VNR*)

(*) VNR : Valeurs Nutritionnelles de Référence selon le Règlement (UE) 1169/2011

Présentation de la problématique

Mesurer la sanité d'un produit?

www.mangerbouger.fr : programme national de nutrition santé



Le Nutri-Score [-15 ; 40]

Le Nutri-Grade [A, B, C, D, E]

Eléments défavorables au score

- Énergie
- Acides gras saturés
- Sucres
- Sel

Eléments favorables

- Fruits
- Légumes
- Légumineuses et oléagineux
- Fibres
- Protéines

Présentation de la problématique

Base de données **OpenFoodFacts**: 320 772 lignes, 162 colonnes

	code	url	creator	created_t	created_datetime	last_modified_t	...	chlorophyl_100g	carbon-footprint_100g	nutrition-score-fr_100g	nutrition-score-uk_100g	glycemic-index_100g
0	0000000003087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	...	NaN	NaN	NaN	NaN	NaN
1	0000000004530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	...	NaN	NaN	14.0	14.0	NaN
2	0000000004559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	...	NaN	NaN	0.0	0.0	NaN
3	0000000016087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	...	NaN	NaN	12.0	12.0	NaN
4	0000000016094	http://world-	usda-ndb-	1489055653	2017-03-	1489055653		NaN	NaN	NaN	NaN	NaN

1. **Nettoyage de la base de données**:
   - Obtenir une base de données ‘clean’
   - et utilisable, contenant les nutri-score & nutri-grade pour chaque produit.
2. **Exploration des données**:
   - Trouver les variables importantes
   - comprendre s’il y a des liens entre des variables
   - les visualiser
3. **Conclusions**

Nettoyage des données

Premier tri fonctionnel:

- Suppression **colonnes non pertinentes** pour la qualification de produit sain. EX: *'url' ; 'created_datetime', etc.*
- Suppression des **colonnes** en **doublons** portant la même information: Ex: *'brands_tags' et 'brands'*
- **Filtre** sur les produits vendus en **France** uniquement
- Suppression de la **colonne 'nutrition-score-uk_100g'**, on utilise 'nutrition-score-fr_100g'
- Suppression des **lignes sans nom de produit** inutilisable pour notre générateur de recettes

  On a désormais **91 247 lignes et 116 colonnes.**

Gestion des valeurs erronées:

- Suppression des lignes de 'nutrition facts' contenant des **valeurs > 100g ou <0g** (les champs se terminant en _100g)
- Suppression lignes de 'nutrition facts' ayant une **valeur de colonne parente < somme des colonnes filles**
- Suppression lignes d'**énergie démesurée**: seuil max fixé à 50 000 kJ (besoin journalier d'un homme moyen est 10 000 kJ)
- **Correction du 'pnns_groups_1'**: suppression des '-' et mise en majuscule.

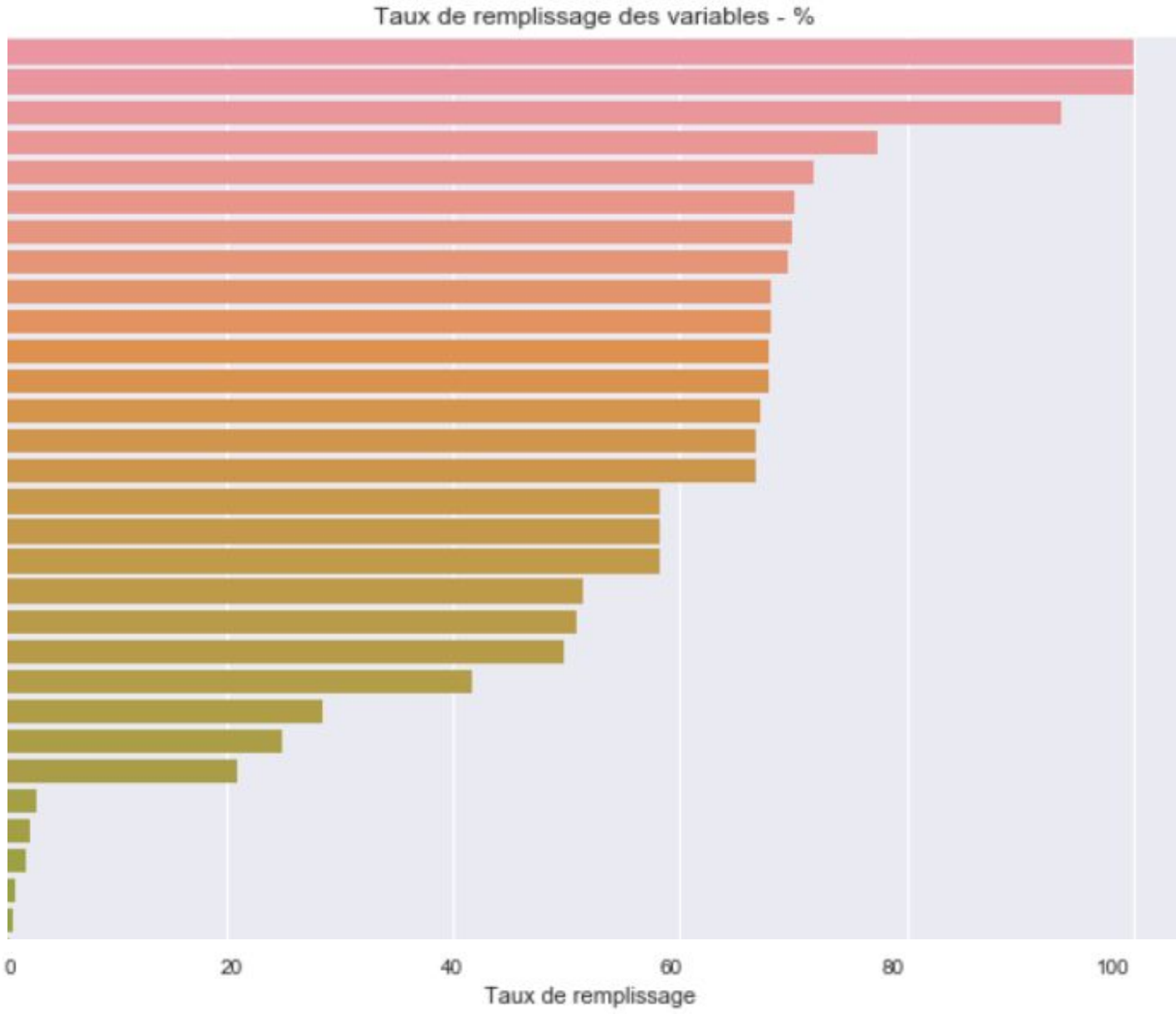
  On a désormais **90 534 lignes et 116 colonnes.**

Réserves: Fiabilité des données: base de données OpenSource, procédure d'ajout? Procédure de contrôle?

Nettoyage des données

Calcul du **taux de remplissage des colonnes**:

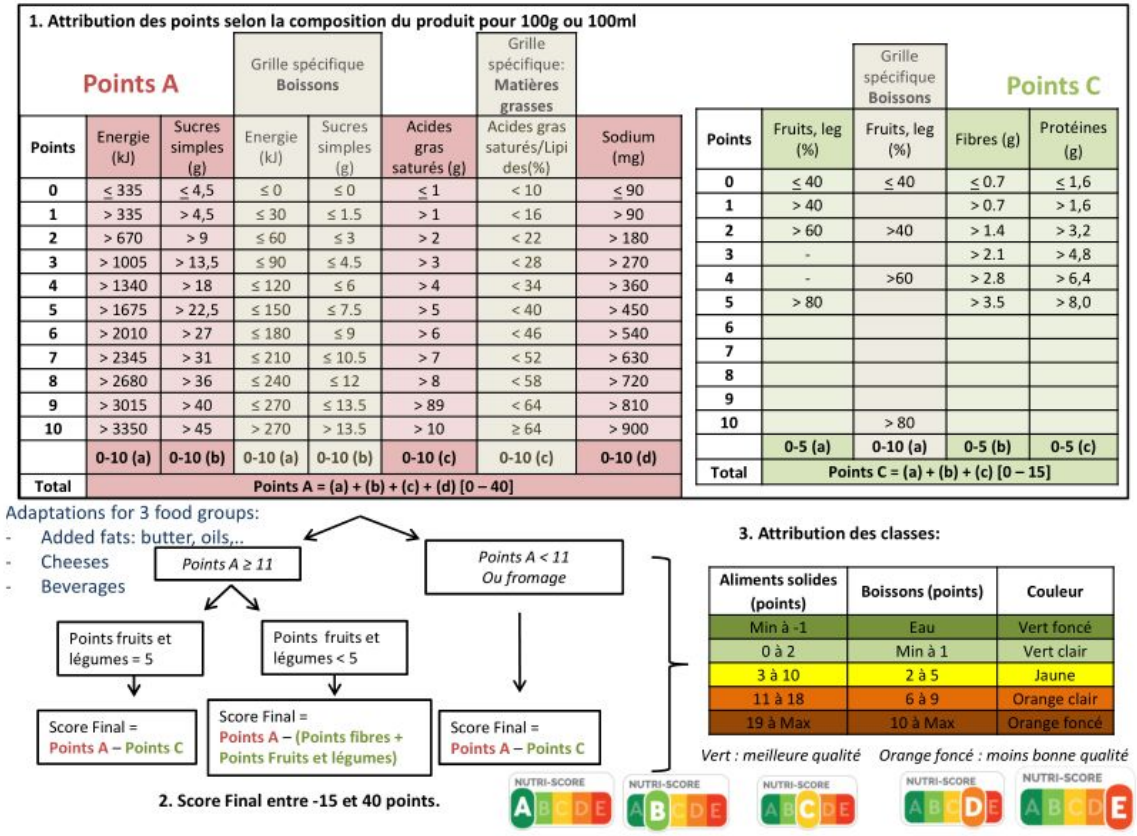
index	Taux
product_name	100.00
countries_fr	100.00
brands	93.54
quantity	77.46
pnns_groups_2	71.74
energy_100g	70.15
proteins_100g	69.86
pnns_groups_1	69.45
sodium_100g	67.96
salt_100g	67.96
sugars_100g	67.90
saturated-fat_100g	67.73
main_category_fr	67.16
nutrition-score-fr_100g	66.72
nutrition_grade_fr	66.72
ingredients_that_may_be_from_palm_oil_n	58.28
ingredients_from_palm_oil_n	58.28
additives_n	58.28
fat_100g	51.45
carbohydrates_100g	50.98
fiber_100g	49.67
generic_name	41.69
serving_size	28.52
allergens	25.00
traces	20.95
fruits-vegetables-nuts_100g	3.25
alcohol_100g	2.63
calcium_100g	2.35
vitamin-c_100g	1.35
iron_100g	1.22
...	...



Nettoyage des données

Gestion des valeurs manquantes:

- Suppression des **colonnes contenant trop peu de valeurs** remplies: (80 colonnes supprimées)
  Le 'nutrition_score' est remplie à ~74%
  Choix: seuil de remplissage minimum des colonnes: 3%
- Suppression des **lignes avec tous leurs 'nutrition facts' vides**. (26 475 lignes supprimés)
- **Remplacement des NaN** par des 0 pour tous les "nutrition facts"
- **Remplacement des Nan du nutrition_score et nutrition_grade par calcul**



1. Attribution des points selon la composition du produit pour 100g ou 100ml

	Points A		Grille spécifique Boissons			Grille spécifique: Matières grasses	
Points	Energie (kJ)	Sucres simples (g)	Energie (kJ)	Sucres simples (g)	Acides gras saturés (g)	Acides gras saturés/Lipides(%)	Sodium (mg)
0	≤ 335	≤ 4,5	≤ 0	≤ 0	≤ 1	< 10	≤ 90
1	> 335	> 4,5	≤ 30	≤ 1.5	> 1	< 16	> 90
2	> 670	> 9	≤ 60	≤ 3	> 2	< 22	> 180
3	> 1005	> 13,5	≤ 90	≤ 4.5	> 3	< 28	> 270
4	> 1340	> 18	≤ 120	≤ 6	> 4	< 34	> 360
5	> 1675	> 22,5	≤ 150	≤ 7.5	> 5	< 40	> 450
6	> 2010	> 27	≤ 180	≤ 9	> 6	< 46	> 540
7	> 2345	> 31	≤ 210	≤ 10.5	> 7	< 52	> 630
8	> 2680	> 36	≤ 240	≤ 12	> 8	< 58	> 720
9	> 3015	> 40	≤ 270	≤ 13.5	> 89	< 64	> 810
10	> 3350	> 45	> 270	> 13.5	> 10	≥ 64	> 900
	0-10 (a)	0-10 (b)	0-10 (a)	0-10 (b)	0-10 (c)	0-10 (c)	0-10 (d)
Total	Points A = (a) + (b) + (c) + (d) [0 – 40]						

		Grille spécifique Boissons	Points C	
Points	Fruits, leg (%)	Fruits, leg (%)	Fibres (g)	Protéines (g)
0	≤ 40	≤ 40	≤ 0.7	≤ 1,6
1	> 40		> 0.7	> 1,6
2	> 60	>40	> 1.4	> 3,2
3	-		> 2.1	> 4,8
4	-	>60	> 2.8	> 6,4
5	> 80		> 3.5	> 8,0
6				
7				
8				
9				
10		> 80		
	0-5 (a)	0-10 (a)	0-5 (b)	0-5 (c)
Total	Points C = (a) + (b) + (c) [0 – 15]			

Adaptations for 3 food groups:
- Added fats: butter, oils,..
- Cheeses
- Beverages

Points A ≥ 11

Points A < 11
Ou fromage

Points fruits et légumes = 5

Points fruits et légumes < 5

Score Final = Points A – Points C

Score Final = Points A – (Points fibres + Points Fruits et légumes)

Score Final = Points A – Points C

2. Score Final entre -15 et 40 points.

3. Attribution des classes:

Aliments solides (points)	Boissons (points)	Couleur
Min à -1	Eau	Vert foncé
0 à 2	Min à 1	Vert clair
3 à 10	2 à 5	Jaune
11 à 18	6 à 9	Orange clair
19 à Max	10 à Max	Orange foncé

Vert : meilleure qualité *Orange foncé : moins bonne qualité*

NUTRI-SCORE A B C D E

via la formule OpenFoodFacts
(3782 lignes modifiées)

Nettoyage des données

Gestion des valeurs manquantes:

- Gestion des NaN restants du **nutrition_score: prédiction par régression linéaire** : (81 lignes modifiées)

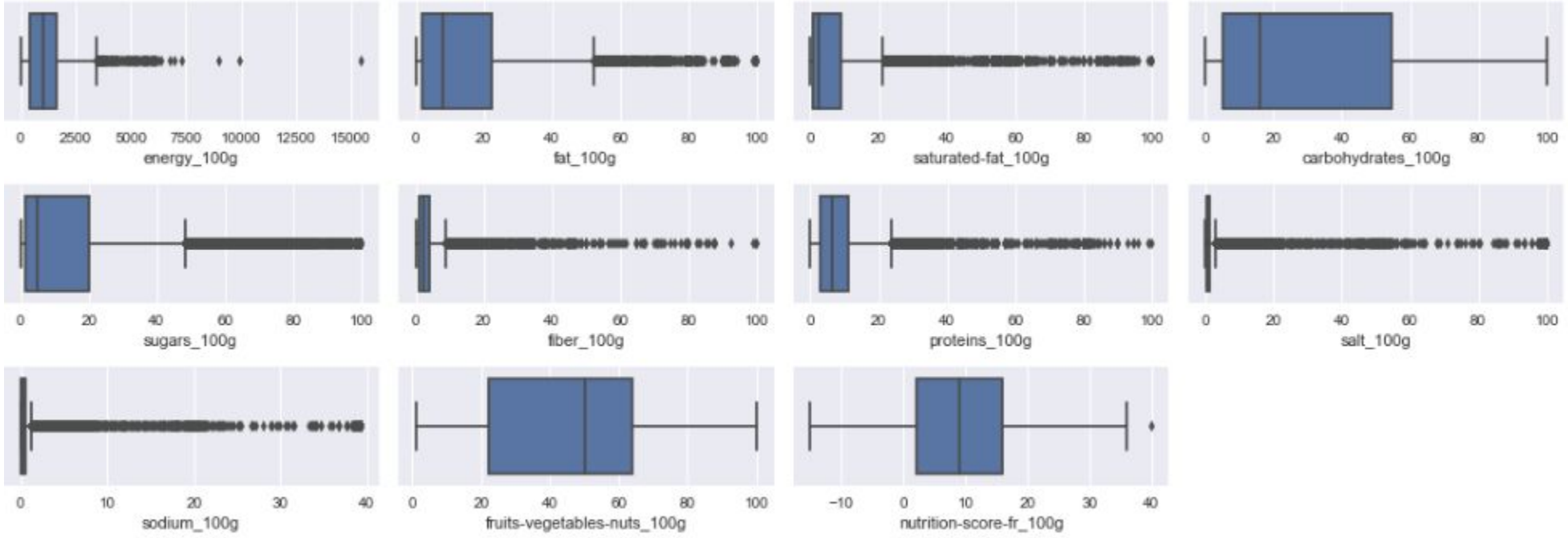
  Bootstrapping d'arbre de décision: BaggingRegressor() ⇒ $R^2 = 0.96$

- Gestion des NaN restants du **nutrition_grade: prédiction par algorithme de classification** : (81 lignes modifiées)

  Forêt aléatoire: RandomForestClassifier() ⇒ Accuracy = 0.91

Gestion des valeurs aberrantes:

- Recherche d'outliers chez les "nutrition facts":
  - via boxplot: **outliers** présents vers les maximums



Nettoyage des données

- Via intervalle: [Q1-1.5*IQR ; Q3+1.5*IQR] ; tableaux triés par ordre décroissant:

Des valeurs attendues:

saturated-fat_100g	product_name
100.0	Jog'Frit
100.0	Huile De Tournesol Végétale 2 Litre
100.0	Frites & Fritures
100.0	Eau
100.0	Végétaline (offre familiale)

salt_100g	product_name
100.0	Sel de table fin
100.0	Sel Marin Ile de Noirmoutier fin naturel
100.0	Fleur de sel de Guérande
100.0	Sel de Guérande
100.0	Fleur de sel de Guérande

fat_100g	product_name
100.0	Huile vierge biologique Chanvre
100.0	Coeur de Tournesol facile à étaler
100.0	Duo Huile & Beurre
100.0	Lesieur Cœur de Tournesol Mini
100.0	Huile d'olive vierge extra

sugars_100g	product_name
100.0	Sucre en Morceaux n°4
100.0	Sucre en poudre
100.0	Sucre en Morceaux 1 kg
100.0	Morceaux Bruns
100.0	Sucre poudre

D'autres inattendues:

proteins_100g	product_name
100.0	Le Hobbit : La Bataille Des Cinq Armés - Versi...
100.0	Eau
100.0	Savarez - Jeu De Cordes
100.0	Harry Potter & The Deathly Hallows Radcliffe /...
100.0	Marshall - Major Noir
99.0	Lingettes pocket pour visage et mains, biodégr...

fiber_100g	product_name
100.0	Eau
100.0	Savarez - Jeu De Cordes
100.0	Noix sèches
100.0	Marshall - Major Noir
100.0	Le Hobbit : La Bataille Des Cinq Armés - Versi...
100.0	Harry Potter & The Deathly Hallows Radcliffe /...

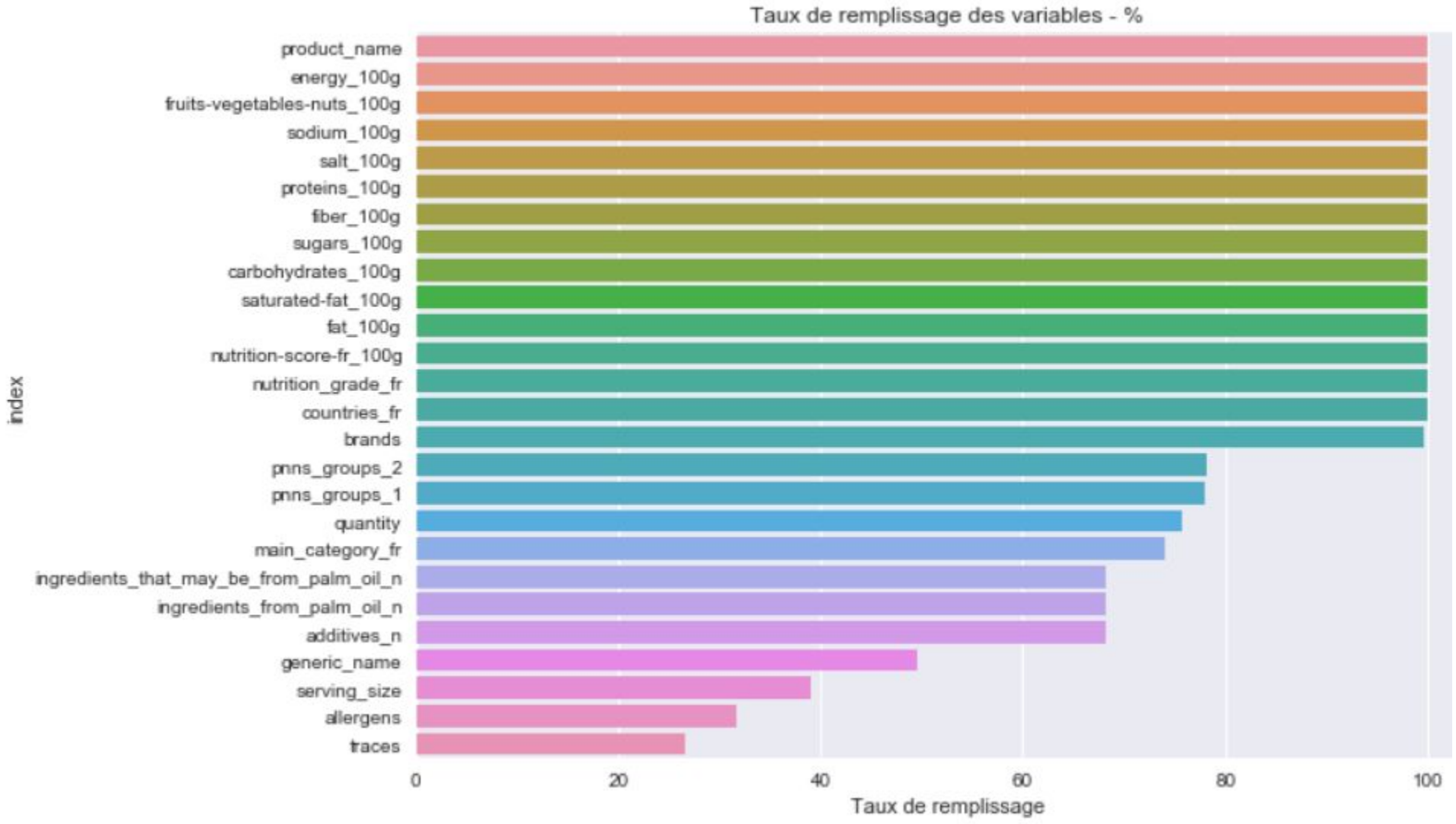
- **Suppression manuelle des outliers erronés**.

  Suppression automatique envisageable en supprimant les lignes ayant les colonnes 'pnns_group' à vide, mais cela supprimerait 20% des données restantes.

Nettoyage des données

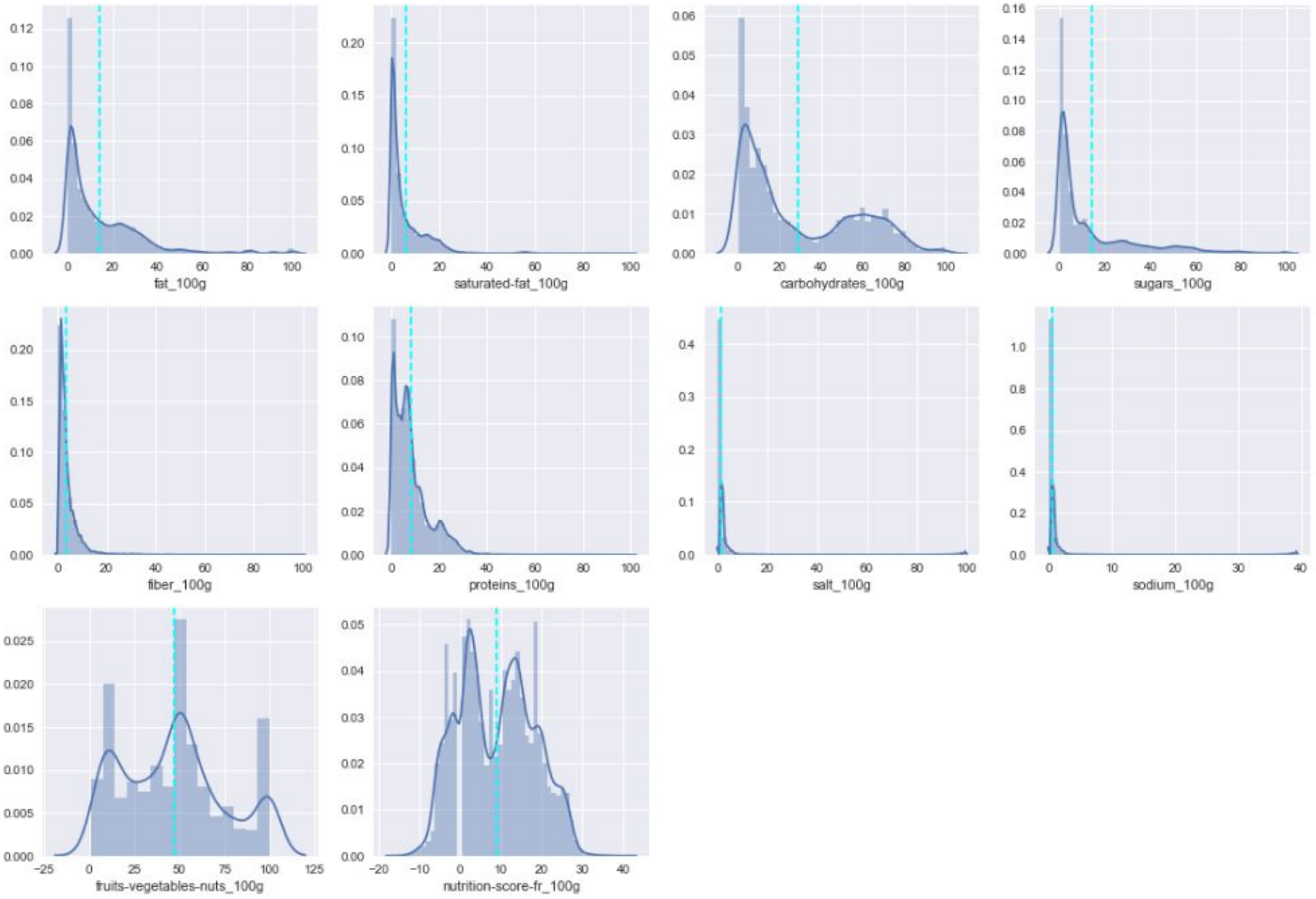
On a désormais **64 055 lignes et 26 colonnes.**

Taux de remplissage final:



Analyse exploratoire

Distribution des “nutrition facts”: **histogrammes et moyennes**:



Analyse exploratoire

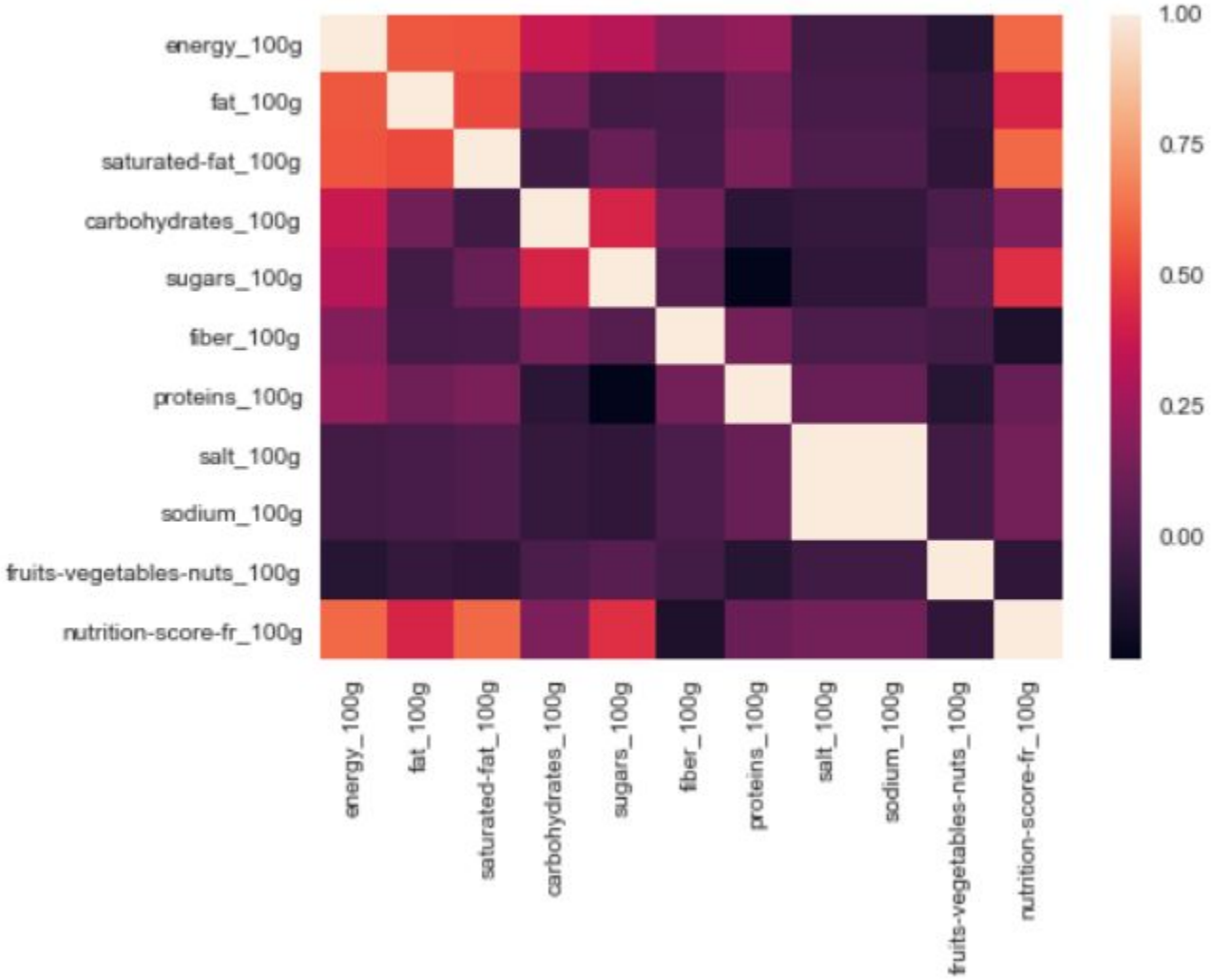
Mesures de forme: **skewness et kurtosis**:

	skewness	kurtosis	total lines != 0
energy_100g	0.819390	2.683782	63175
fat_100g	2.556576	8.633377	42851
saturated-fat_100g	3.575818	21.056781	53703
carbohydrates_100g	1.142625	-0.121452	44280
sugars_100g	1.948997	3.461179	57369
fiber_100g	7.672219	108.854184	31540
proteins_100g	2.224116	10.905484	59623
salt_100g	15.851539	310.362299	56253
sodium_100g	15.851333	310.354178	56251
fruits-vegetables-nuts_100g	7.530126	60.705043	1990
nutrition-score-fr_100g	0.235523	-0.940443	60006

- Toutes les distributions sont **étalées à droite**: **skewness > 0**
- La majorité des colonnes montrent des **exponentielles** vers les valeurs minimums.
  Leur **kurtosis > 1** indique bien que la distribution est très **concentrée** et moins aplatie que la distrib. Normale.
- Seul le 'nutrition_score' et le 'carbohydrates' ont un **kurtosis <0** et montrent des distributions plus **aplaties**, on voit d'ailleurs une **distribution bimodale** pour ceux-ci.

Analyse exploratoire

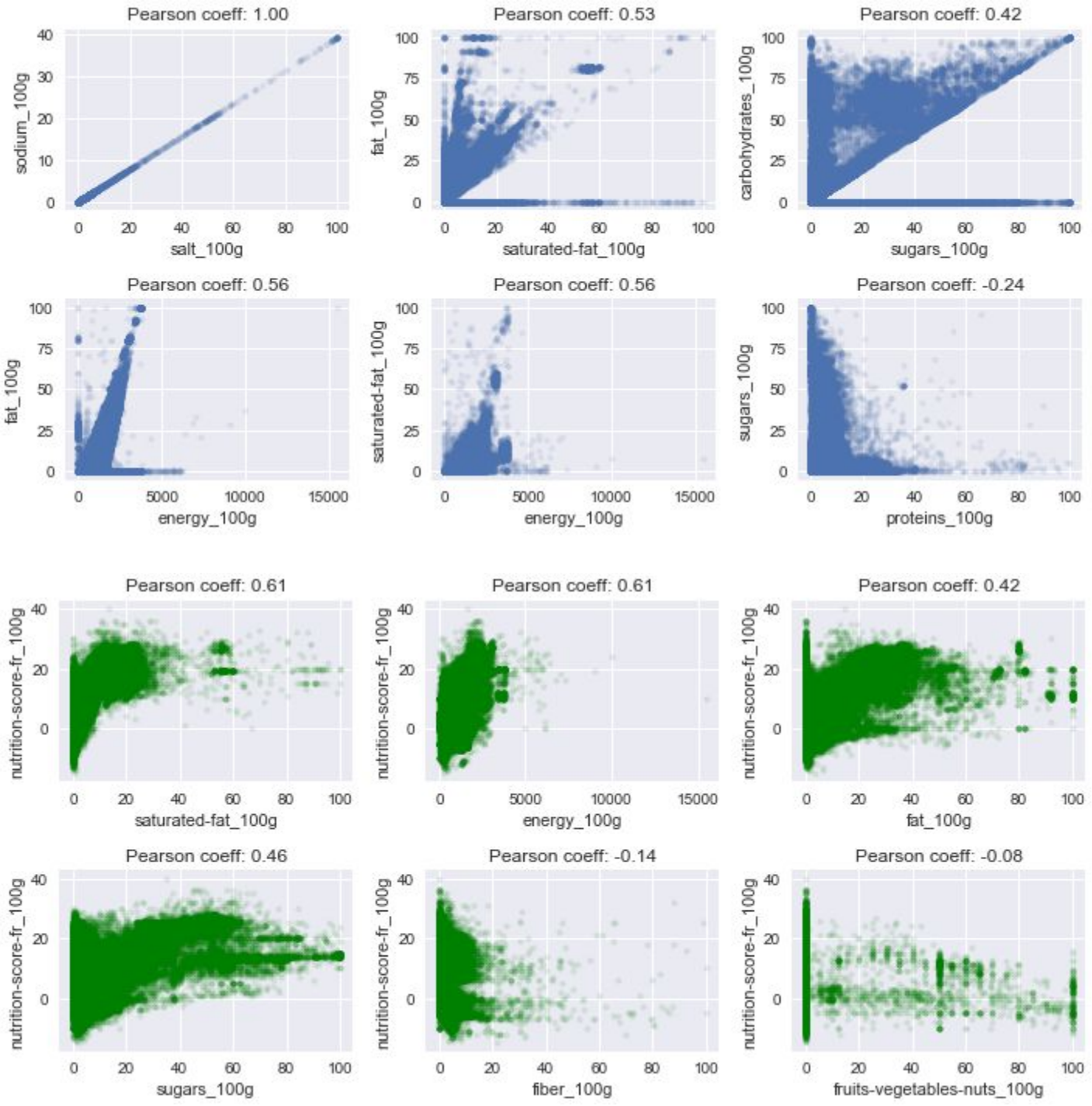
Heatmap: **matrice des corrélations linéaires**:



Analyse exploratoire

Analyse bivariée: **scatter plots**:

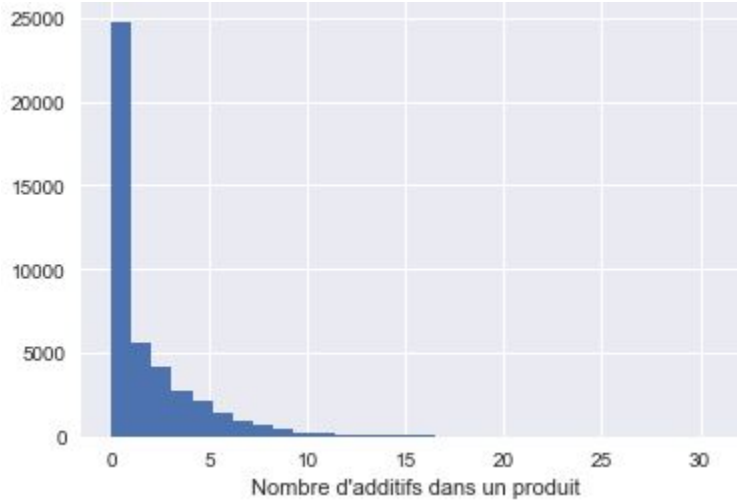
- “**Nutrition facts**” vs “nutrition facts”
- **Nutrition score** vs “nutrition facts”



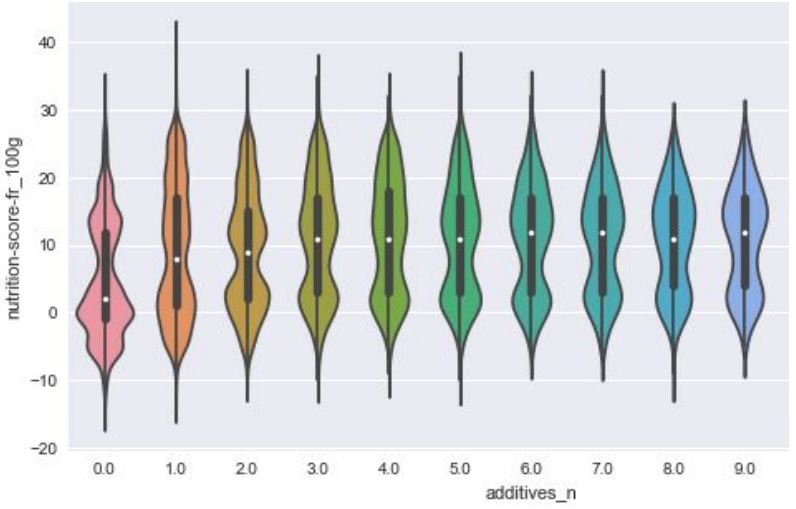
Analyse exploratoire

Nombre d'additifs :

Histogramme:



Nombre d'additifs VS Nutri-score: r = 0.28



⇒ additifs non pris en compte dans le calcul du nutri-score (ce qui est d'ailleurs un reproche fait au nutri-score).

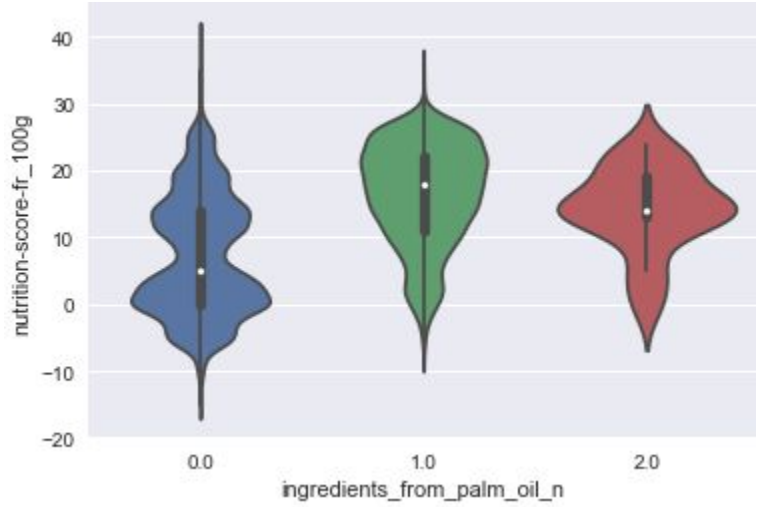
Huile de palme:

Histogramme:

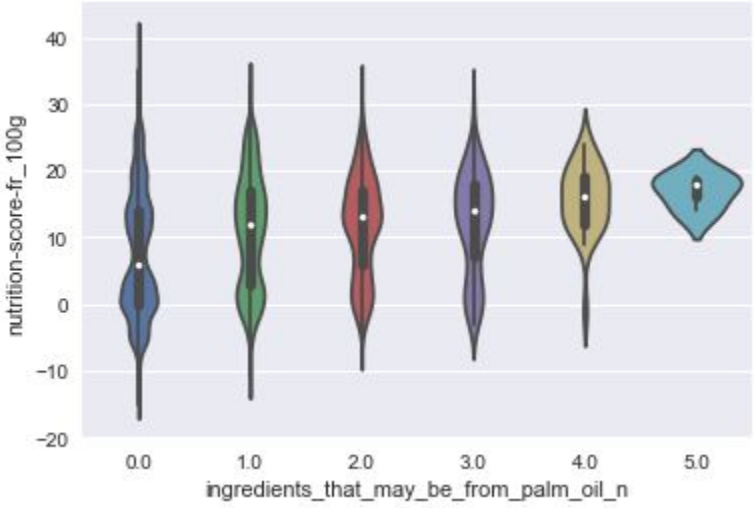
```
0.0    40344
1.0     3283
2.0       45
Name: ingredients_from_palm_oil_n

0.0    38077
1.0     4429
2.0      923
3.0      211
4.0       29
5.0        3
Name: ingredients_that_may_be_from_palm_oil_n
```

Analyse exploratoire

Nombre d'ingrédients provenant d'huile de palme VS Nutri_score: r = 0.26



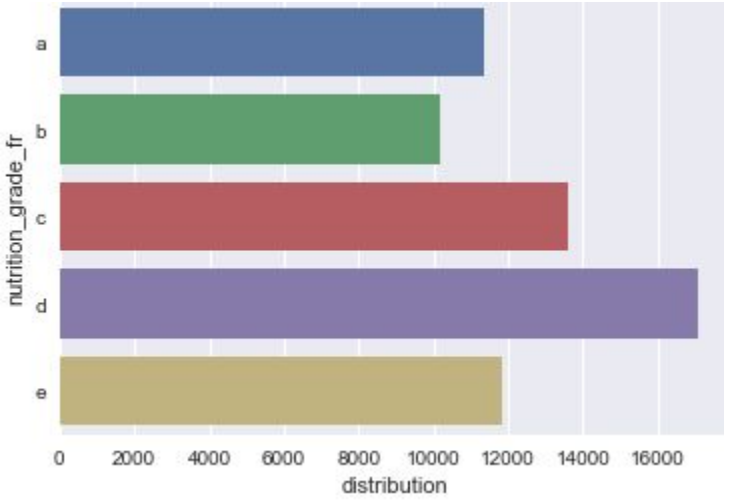
Nombre d'ingrédients pouvant provenir d'huile de palme VS Nutri_score: r = 0.14



⇒ huile de palme non prise en compte dans le calcul du nutri-score.

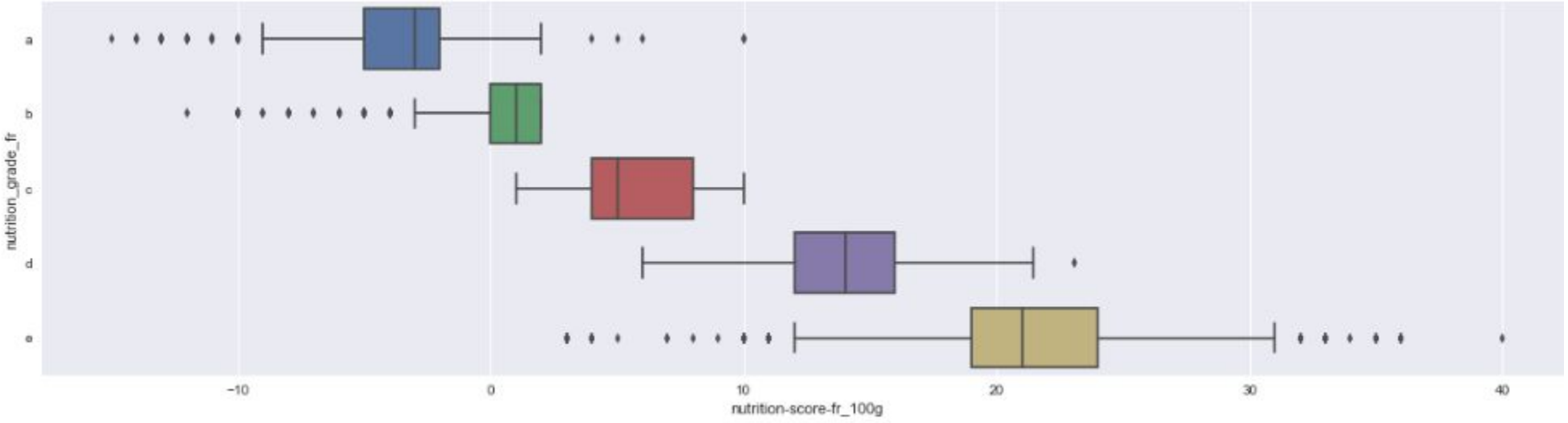
Analyse exploratoire

Nutri-grade: distribution des catégories: barplot



Distribution assez homogène.

Nutri-grade VS nutri-score: box plots splitté par catégorie



Répartition sans surprise. A part quelques outliers.

Analyse exploratoire

Les outliers maximum de catégorie A:

	product_name	nutrition-score-fr_100g
306261	Agua Mineral Natural	10.0
192056	Evian	10.0
304683	Courmayeur	6.0
249676	Agua mineral natural	5.0
220374	Sémillante arômes naturels Citron	4.0

Les outliers minimum de catégorie E:

Aliment solide: E si nutri_score > 18
Boisson: E si nutri_score > 9

	product_name	nutrition-score-fr_100g
251991	Limonade Fraise Des Bois	10.0
212895	Pur jus raisin	10.0
200535	Cola	10.0
222480	Smoothie carotte pomme-thé vert	10.0
191388	Ricqles Menthe	10.0

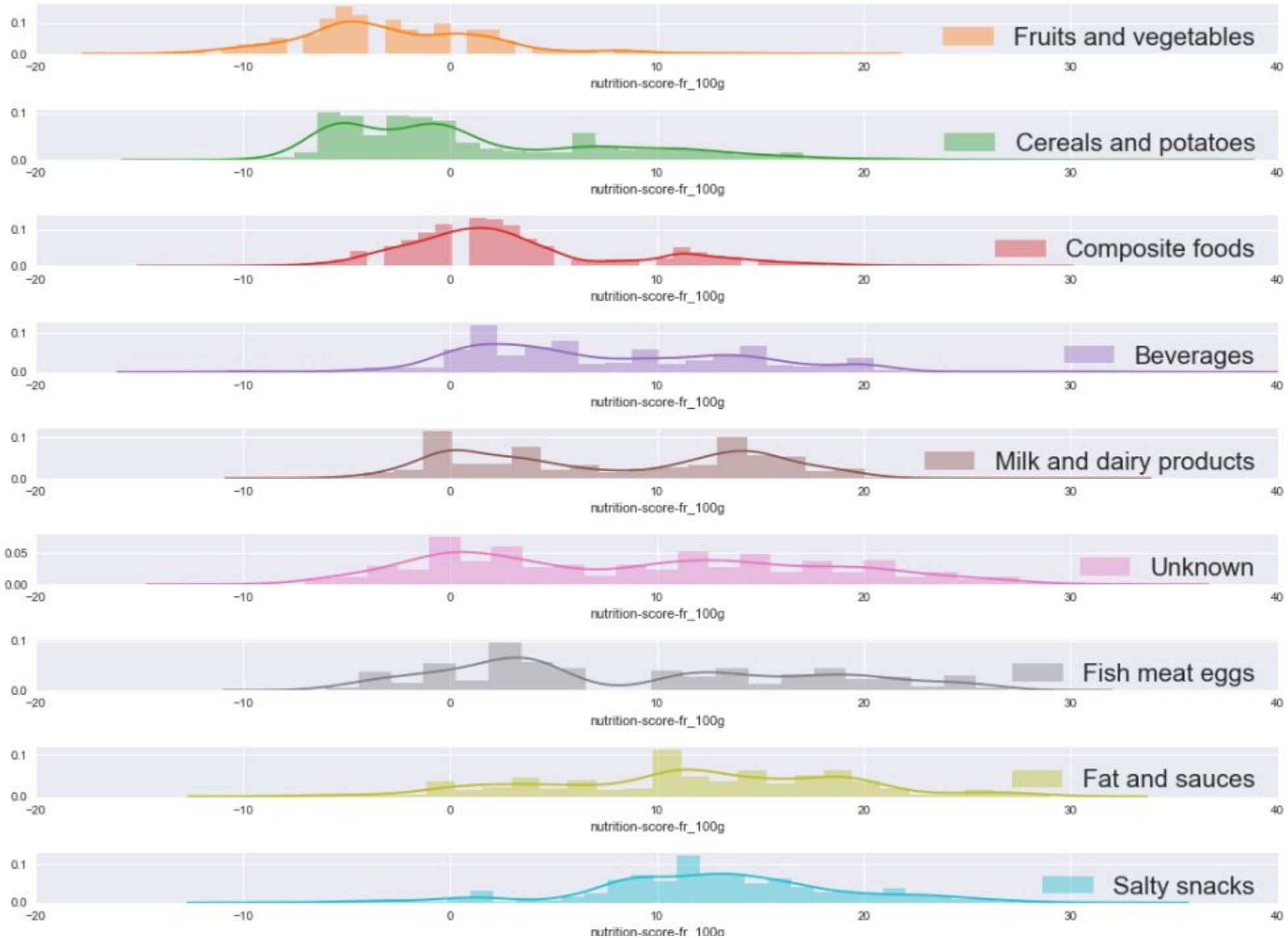
Moyennes du nutri-score des catégories de produit: 'pnns_groups_1' triées par ordre croissant:

pnns_groups_1	nutrition-score-fr_100g
Fruits and vegetables	-2.595844
Cereals and potatoes	1.200999
Composite foods	3.390529
Beverages	7.659552
Milk and dairy products	7.731098
Unknown	8.695548
Fish meat eggs	8.899388
Fat and sauces	11.845887

Résultats sans surprise.

Analyse exploratoire

Histogrammes séparés par groupe ‘pnns_groups_1’:



Analyse exploratoire

Certaines **distributions multimodales** s'expliquent par différentes sous-categories regroupées ensembles.

Exemple distribution bimodale du fromage: ‘pnns_group_2’==’Cheese’

Catégorie contenant des fromages blancs et des fromages plus traditionnels.

Création de la catégorie ‘Fromage blanc’ regroupant tous les produits contenant le mot ‘blanc’ dans ‘product_name’, ainsi que de la catégorie **‘Autres fromages’** qui ne contiennent pas le mot ‘blanc dans le ‘product_name’

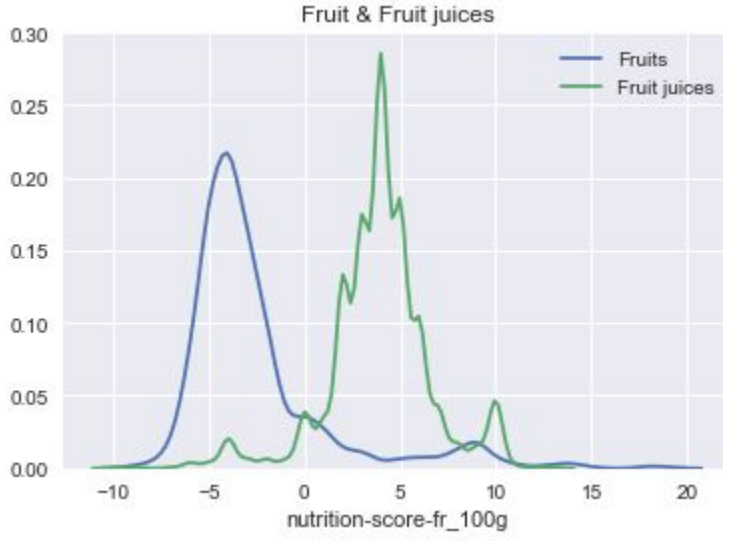
Analyse exploratoire

Tests statistiques: des **catégories soient saines que d'autres**?

H0: hypothèse nulle: les moyennes des nutri-score des 2 catégories sont égales: $\mu_1 = \mu_2$
HA: hypothèse alternative: $\mu_1 < \mu_2$

Conditions remplies: - échantillons > 50 individus
- variances comparables (rapport 1 à 2)

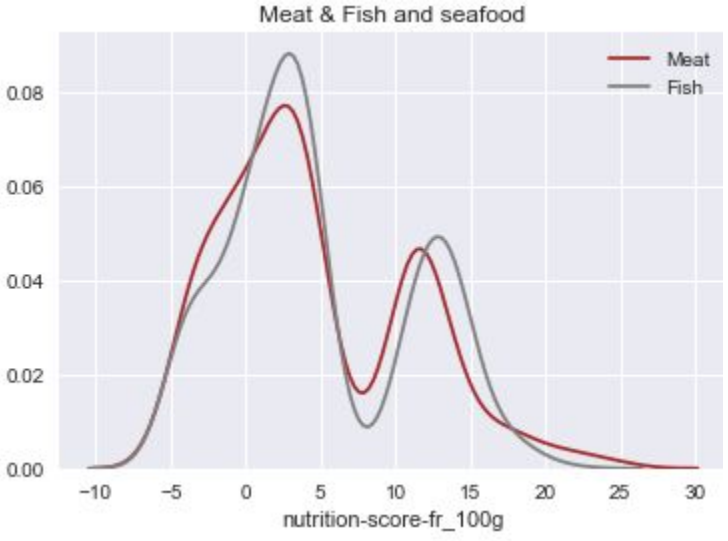
Distributions du nutri_score pour: **'Fruits'** & **'Jus de fruits'**:



T-test: `t-test: p-value = 1.1406171051838952e-279`

Conclusion: p-value << 5% - on rejette H0
Les 'Fruits' sont plus sains que 'Jus de fruits' car ils ont un score nutritionnel plus faible en moyenne. On a une très très faible chance de se tromper ici

ainsi que de **'Viandes'** et **'Poissons et fruits de mer'**:



`t-test: p-value = 0.4911766485873108`

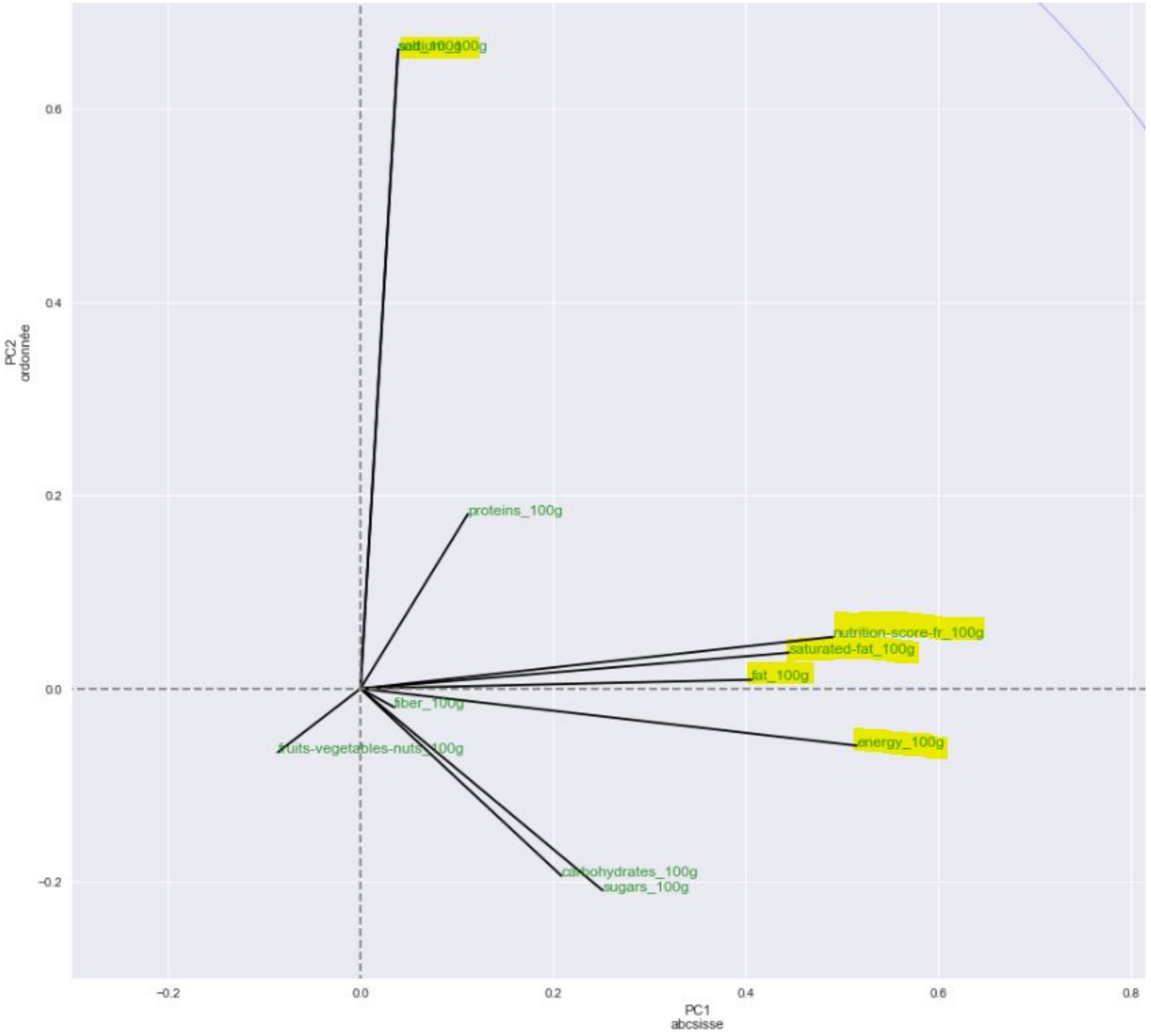
p-value >> 5% - on ne peut rejeter H0
On ne peut pas affirmer qu'il y a une différence significative entre les moyennes des 2 catégories.

Analyse exploratoire

PCA: visualisation: 6 CP

Taux de variance totale expliqué par les 6 premiers CP: **0.86**

Cercle de corrélations:
CP1 et CP2



Analyse exploratoire

PCA: **matrice** des vecteurs propres:

	CP1	CP2	CP3	CP4	CP5	CP6
energy_100g	0.515321	-0.058941	-0.018817	0.186800	0.049606	-0.005369
fat_100g	0.406175	0.009192	-0.261384	-0.076486	0.162293	-0.475657
saturated-fat_100g	0.444876	0.037146	-0.264504	-0.179958	0.111691	-0.096664
carbohydrates_100g	0.208655	-0.193617	0.463884	0.319480	-0.064959	-0.045465
sugars_100g	0.251436	-0.208695	0.560634	-0.049507	-0.114654	0.284269
fiber_100g	0.034955	-0.019290	0.035562	0.761369	0.225651	-0.260992
proteins_100g	0.111433	0.180421	-0.425317	0.379054	0.055813	0.716642
salt_100g	0.039066	0.661021	0.229689	0.013438	0.015554	-0.059661
sodium_100g	0.039067	0.661021	0.229690	0.013438	0.015554	-0.059660
fruits-vegetables-nuts_100g	-0.085807	-0.065847	0.193269	-0.183615	0.939242	0.166173
nutrition-score-fr_100g	0.490652	0.053520	0.082686	-0.254779	-0.068313	0.255103

PCA: interprétation:

- Nutrition_score et 1ère composante (corrélation 0.49)

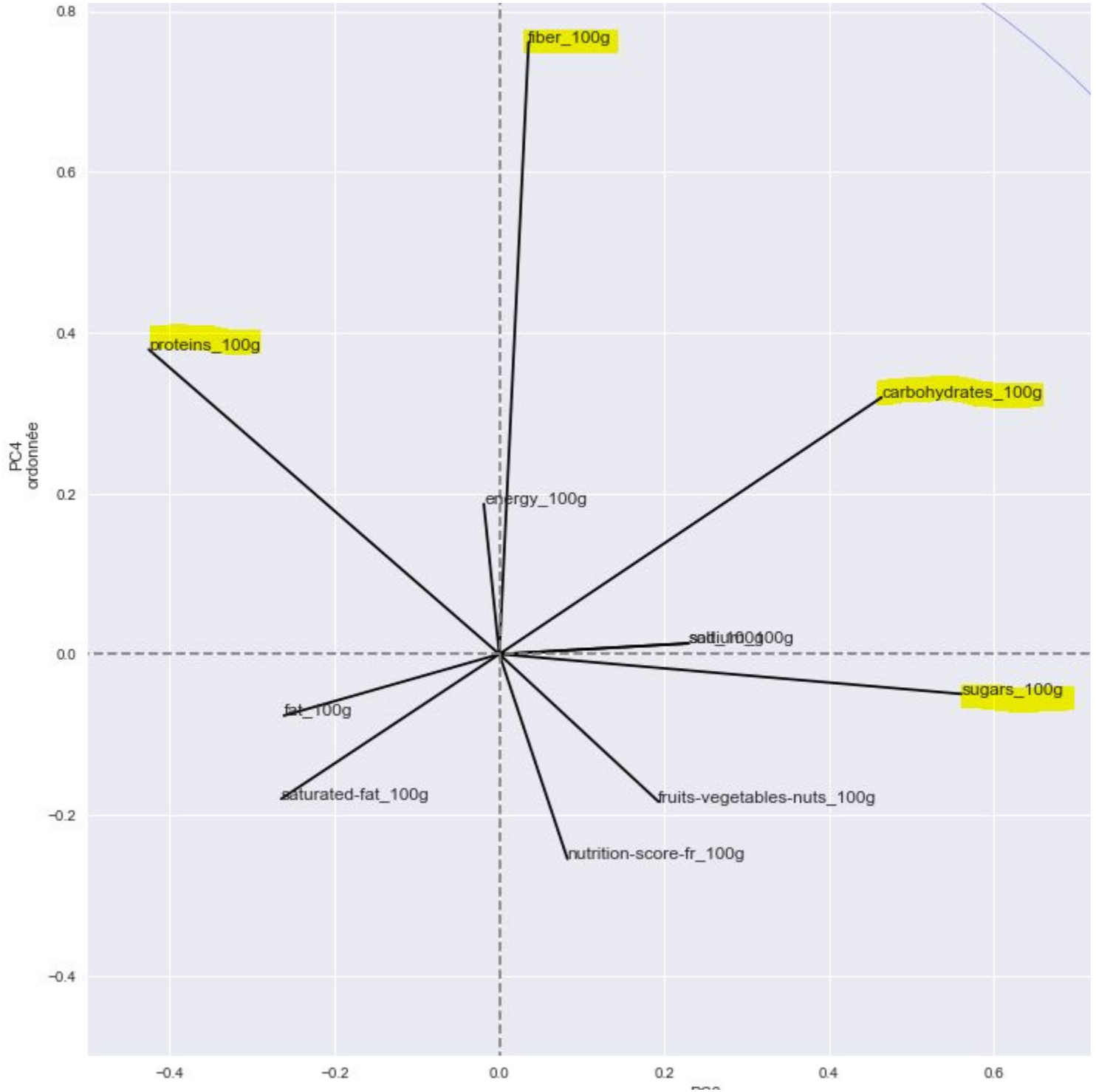
- Variables contribuant à un nutrition_score sont:
  - saturated_fat, fat et energy,
  - carbohydrates (dont les sucres) et protéines
- Variables ne contribuant pas à un nutrition_score:
  - 'fruits-vegetables-nuts'
- Variables sans impact (axe perpendiculaire)
  - sel

Remarques:

- CP1 et CP2 = 45% de la variance totale
- Variables loin du cercle de corrélation sont peu interprétables.

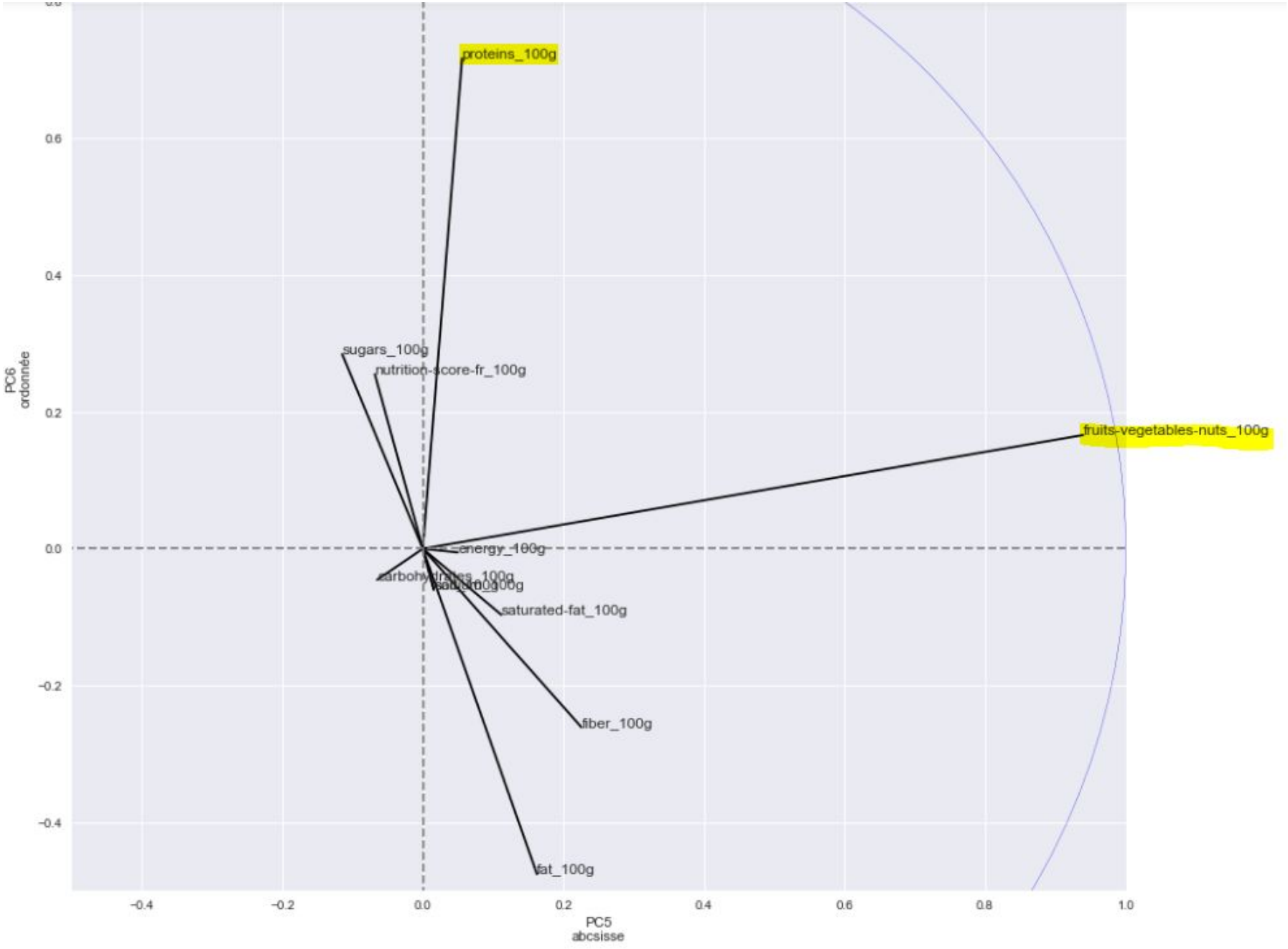
Analyse exploratoire

PC3 et PC4:



Analyse exploratoire

PC5 et PC6



Conclusions

- **Mesure de la sanité** d'un produit via:
  - Nutrition_score
  - Nutrition_grade

- Base de **données répondant au besoin**:
  - 64 000 produits vendus en France
  - Contenant le nutri_score
  - Nouveaux produits ajoutés journalièrement

- Des **catégories plus saines que d'autres**:
  - À privilégier lors de des choix d'ingrédients des recettes
  - À approfondir: beurre VS huile ; sucre VS miel VS agave ; etc.

- À investiguer:
  - Fiabilité des données, champs obligatoires
  - Proposition d'améliorations de la BDD:
    - Normes: bio, label rouge, etc.
    - Catégories: OGM, pesticides, etc.

Générateur de recettes

Next steps: générateur de recettes saines

- Lier : ingrédients des recettes ⇐⇒ avec les produits de la BDD
- Comment ?
    - Soit **définir** des valeurs moyennes par **produit**: énergie moyenne, sucres moyen, etc.
      Soit permettre à l'utilisateur de sélectionner ses produits
    - **Calculer le nutri-score** de chaque **recette** à partir de ces valeurs et des quantités.
    - **Classer les recettes** par ordre de sanité.

Propositions:

- Recherche de recette: proposer un **tri par sanité** (nutri-score ascendant)
- Recherche de recette: proposer un **filtre par nutri-grade** [A,B,C,D,E]
- Dans une recette: **Montrer les produits les plus sains**: le top 5 des farines saines, top 5 des oeufs les plus sains...
- **Générateur de recette**: proposer les recettes les plus saines en page d'accueil