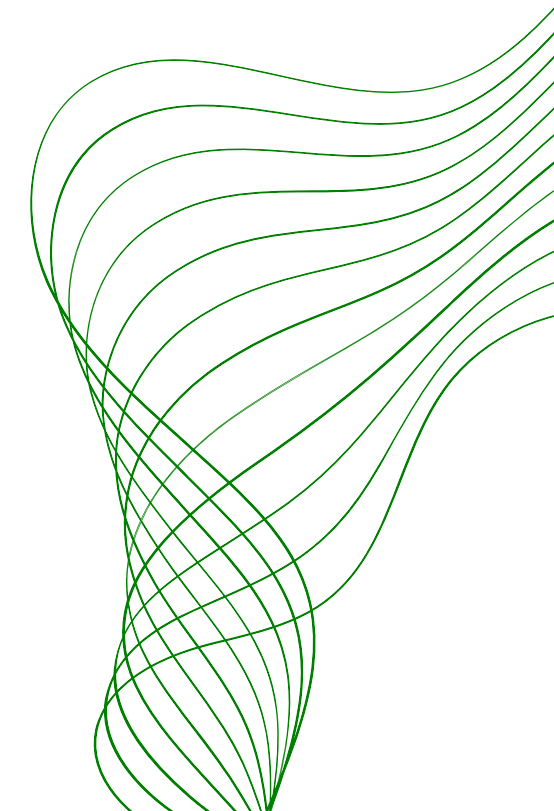


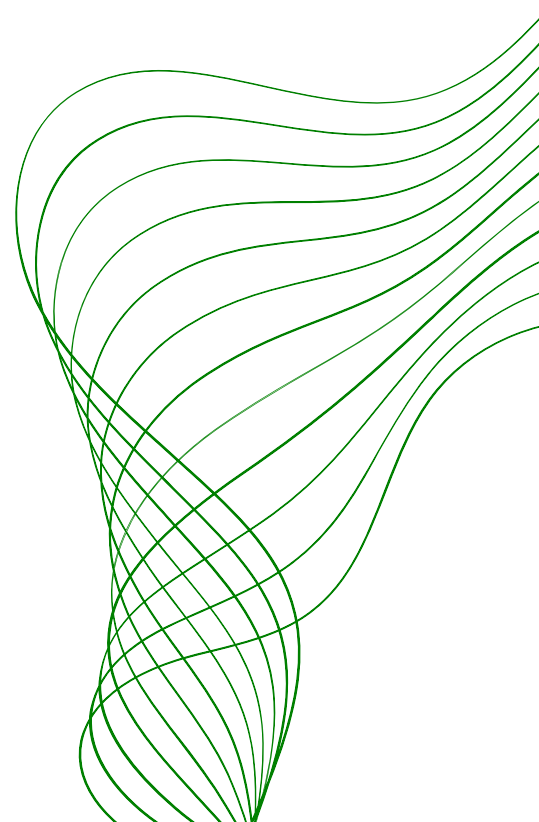


Préparer des données pour un organisme de santé publique





Plan

1. Mission
 2. Nettoyage des données
 3. Exploration des données
 4. Conclusion
 5. Au sujet du RGPD
- 

1. Mission



Évaluer la faisabilité d'une application de suggestion ou d'auto-complétion pour aider au remplissage de la base de données Open Food Facts.

Dans la suite nous ciblons une feature importante :

Le **Nutriscore** (*nutrition_grade_fr*).



2. Nettoyage des données

TRAITEMENT DES OUTLIERS

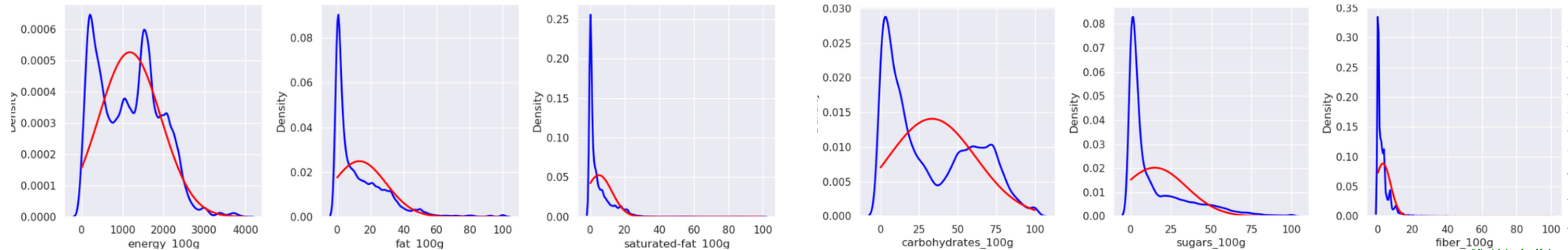
Valeurs aberrantes évidentes :

- Quantité en g pour 100g négatives ou supérieures à 100 ;
- Pour *energy_100g*, valeurs négatives ou supérieures à 4000kj ;
- Valeurs de *saturated-fat_100g* supérieures à celles de *fat_100g* ;
- Valeurs de *sugars_100g* supérieures à celles de *carbohydrates_100g*.

>>> Ces valeurs ont été remplacées par des valeurs manquantes pour être traitées dans la partie consacrée.

Autres outliers (dont valeurs atypiques) :

Les densités estimées des features (en bleu) diffèrent sensiblement des densités normales correspondantes (en rouge) :



>>> La méthode du z-score n'a pas été retenue pour traiter les outliers.

Le taux d'outliers selon la méthode interquartile est significatif pour certaines features :

- supérieur à 5% pour six features (sur 9) ;
- supérieur à 7% pour une feature.

	nbr_outliers	taux_outliers
energy_100g	554	0.27
fat_100g	4874	2.42
saturated-fat_100g	14783	7.33
carbohydrates_100g	0	0.00
sugars_100g	11579	5.74
fiber_100g	12328	6.11
proteins_100g	10312	5.11
salt_100g	10916	5.41
sodium_100g	10800	5.35

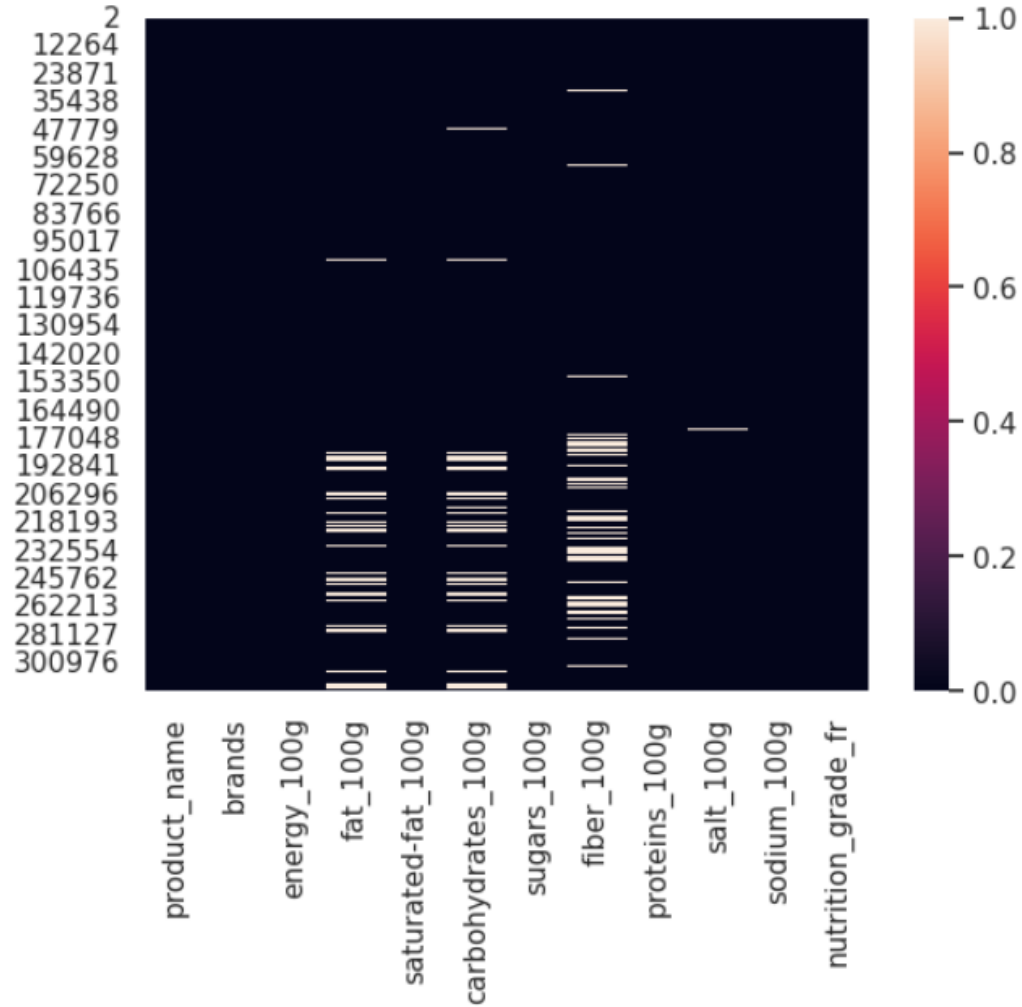
>>> **Ces outliers ont été conservés pour éviter d’avoir trop de valeurs manquantes dans notre jeu de données.**

TRAITEMENT DES VALEURS MANQUANTES

voici le taux de valeurs manquantes par feature ainsi que la heatmap des valeurs manquantes :

- On observe que :
- les taux de valeurs manquantes sont négligeables sauf pour les variables *fat_100g*, *carbohydrates_100g* et *fiber_100g* ;
 - les valeurs manquantes semblent situées aux mêmes lignes pour les features *fat_100g* et *carbohydrates_100g*.

	effectif	taux
product_name	0	0.0
brands	0	0.0
energy_100g	118	0.1
fat_100g	16839	8.3
saturated-fat_100g	286	0.1
carbohydrates_100g	16873	8.4
sugars_100g	589	0.3
fiber_100g	24854	12.3
proteins_100g	1	0.0
salt_100g	45	0.0
sodium_100g	19	0.0
nutrition_grade_fr	0	0.0



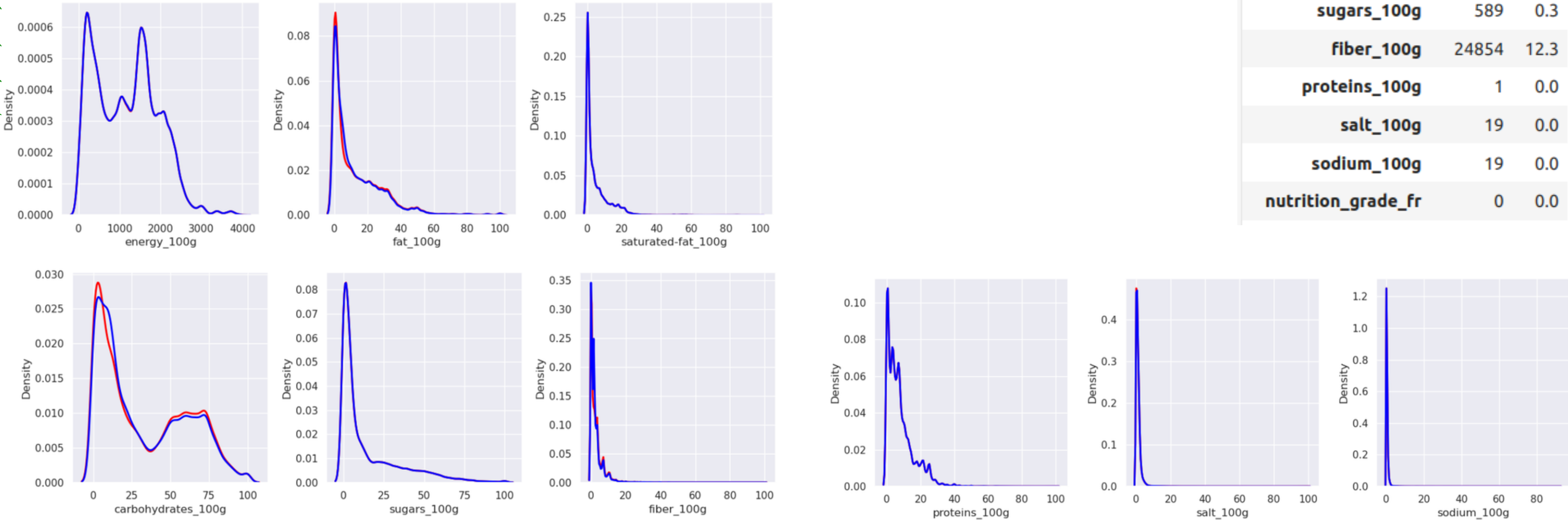
- >>> **Après vérification, il s’agit pour l’essentiel des produits dont seules les valeurs de saturated-fat_100g et sugars_100g sont mentionnées comme indication suffisante du caractère gras et glucidique du produit. Elles ont été complétées à l’aide des features fat_100g et carbohydrates_100g.**
- >>> **La feature fiber_100g n’étant corrélée à aucune autre, nous ne pouvons compléter ses valeurs manquantes à l’aide des autres features.**

Ci-contre, les taux de valeurs manquantes par feature après ce premier traitement :

Les taux sont négligeables excepté pour la feature *fiber_100g*. Nous avons fait une imputation par la médiane qui est peu sensible aux outliers.

Voilà, le nettoyage est terminé ! Voici une comparaison des courbes de densité des features avant (en rouge) et après (en bleu) nettoyage :

	effectif	taux
product_name	0	0.0
brands	0	0.0
energy_100g	118	0.1
fat_100g	286	0.1
saturated-fat_100g	286	0.1
carbohydrates_100g	589	0.3
sugars_100g	589	0.3
fiber_100g	24854	12.3
proteins_100g	1	0.0
salt_100g	19	0.0
sodium_100g	19	0.0
nutrition_grade_fr	0	0.0



Les changements sont imperceptibles ou très légers, ce qui indique que le nettoyage n'a pas dénaturé l'information contenue dans le jeu de données.

3. Exploration des données

ANALYSES UNIVARIÉES

Features quantitatives :

Ci-contre, un tableau donnant pour chaque feature ses principaux indicateurs statistiques :

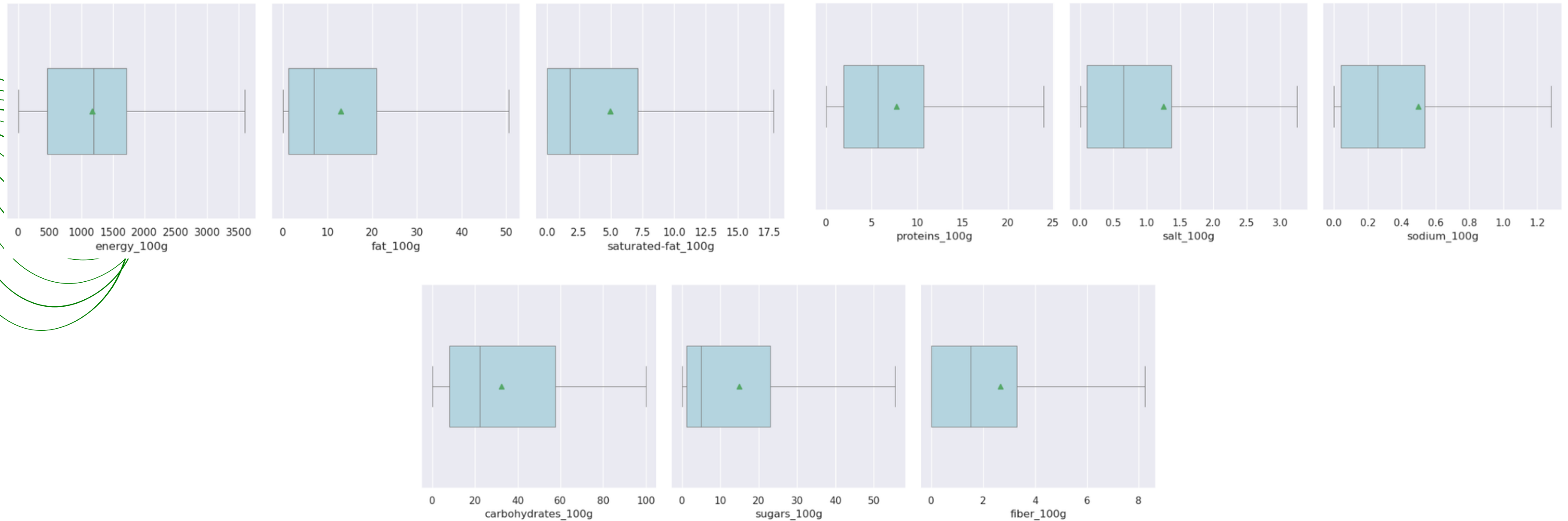
- Indicateurs de tendance centrale :
moyenne et mediane ;
- Indicateurs de dispersion :
écart-type, quartiles et écart interquartile ;
- Indicateurs de forme :
skewness (asymétrie) et kurtosis (aplatissement).

	mean	std	min	Q1	med	Q3	max	etendue	IQR	skew	kurtosis
energy_100g	1177.682	757.723	0.0	460.000	1197.000	1714.000	4000.0	4000.0	1254.000	0.296305	-0.629129
fat_100g	13.041	15.625	0.0	1.220	7.000	21.000	100.0	100.0	19.780	1.910114	5.027429
saturated-fat_100g	4.957	7.559	0.0	0.000	1.790	7.140	100.0	100.0	7.140	3.224456	19.559458
carbohydrates_100g	32.522	27.883	0.0	8.150	22.220	57.500	100.0	100.0	49.350	0.531844	-1.086594
sugars_100g	14.885	19.730	0.0	1.250	5.000	22.998	100.0	100.0	21.748	1.664742	2.274277
fiber_100g	2.649	4.226	0.0	0.000	1.500	3.300	100.0	100.0	3.300	5.249143	54.968820
proteins_100g	7.779	8.042	0.0	1.900	5.710	10.710	100.0	100.0	8.810	2.009230	7.803852
salt_100g	1.257	4.038	0.0	0.100	0.653	1.366	100.0	100.0	1.266	14.870497	281.113835
sodium_100g	0.499	1.687	0.0	0.039	0.257	0.538	92.5	92.5	0.499	16.920554	403.099843

Quelques observations :

- **Skewness positif** : les distributions sont toutes plus étalées à droite qu'à gauche, ce que l'on observe aisément sur les courbes de densité précédentes (Plus ou moins marqué selon les features). Cela signifie que la plupart des produits contiennent des quantités de sel, sucre, fibres, lipides, etc plus proches de 0g que de 100g (pour 100g de produit).
- **moyenne > mediane**, sauf pour *energy_100g* qui est justement celle dont le skewness est le plus proche de 0 : cela est du au fait que sa distribution connaît un fort rebond (cf. courbes de densité précédentes).
- **kurtosis > 0** sauf pour *energy_100g* et *carbohydrates_100g*. Cela qui signifie que les valeurs de chaque feature sont plus concentrées autour de leur moyenne que celles d'un loi normale centrée réduite. Là encore, on peut l'observer sur le premiers graphiques des courbes de densité.

Voici les boxplots de chaque feature :



On peut faire les mêmes observations :

- Les valeurs sont concentrées sur la gauche et étalées sur la droite ;
- La moyenne, représentée par un triangle vert, est toujours largement supérieure à la médiane, sauf pour l'énergie où elles sont très proches.

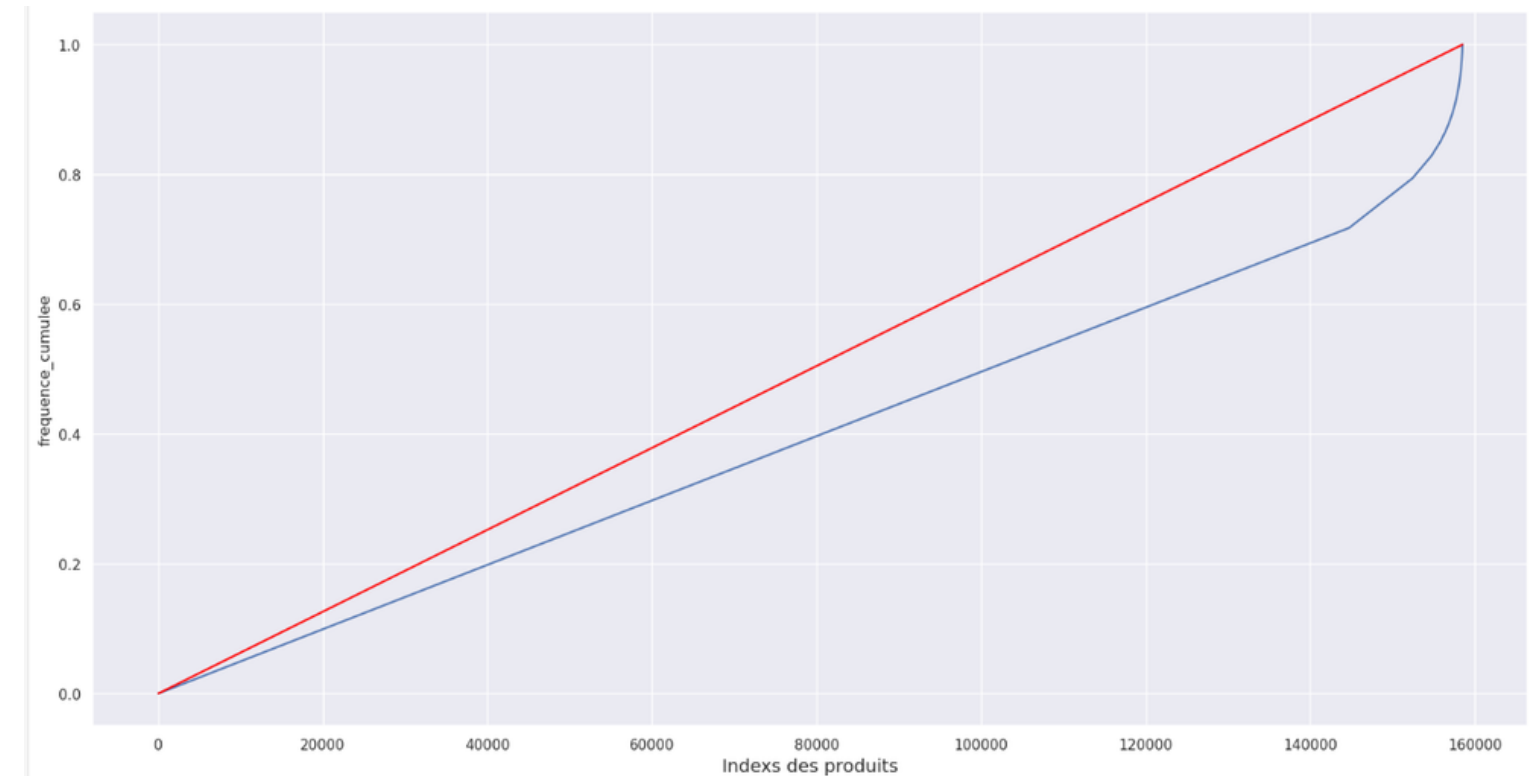
Features *product_name* et *brands* :

Ces deux features sont catégorielles nominales. Voici leurs courbes des fréquences cumulées croissantes (en bleu) :

product_name :

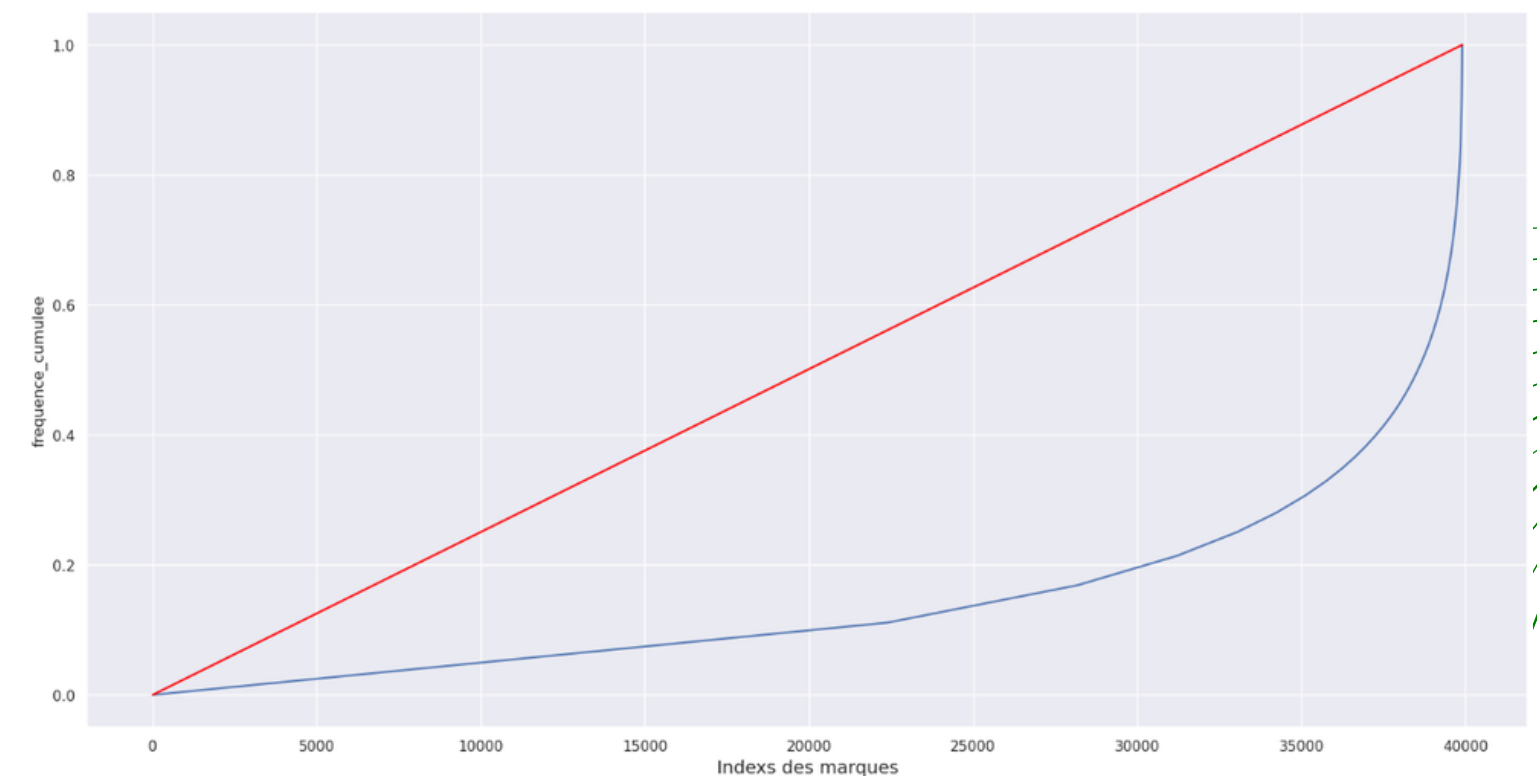
La droite rouge (diagonale) représente le cas où les produits auraient tous le même effectif dans la base de données.

On observe que la courbe des fréquences cumulées est assez proche de la diagonale. On en déduit que les différents produits sont assez équitablement représentés dans la base de données. Il n'y a pas de produit réellement sur-représenté.



brands :

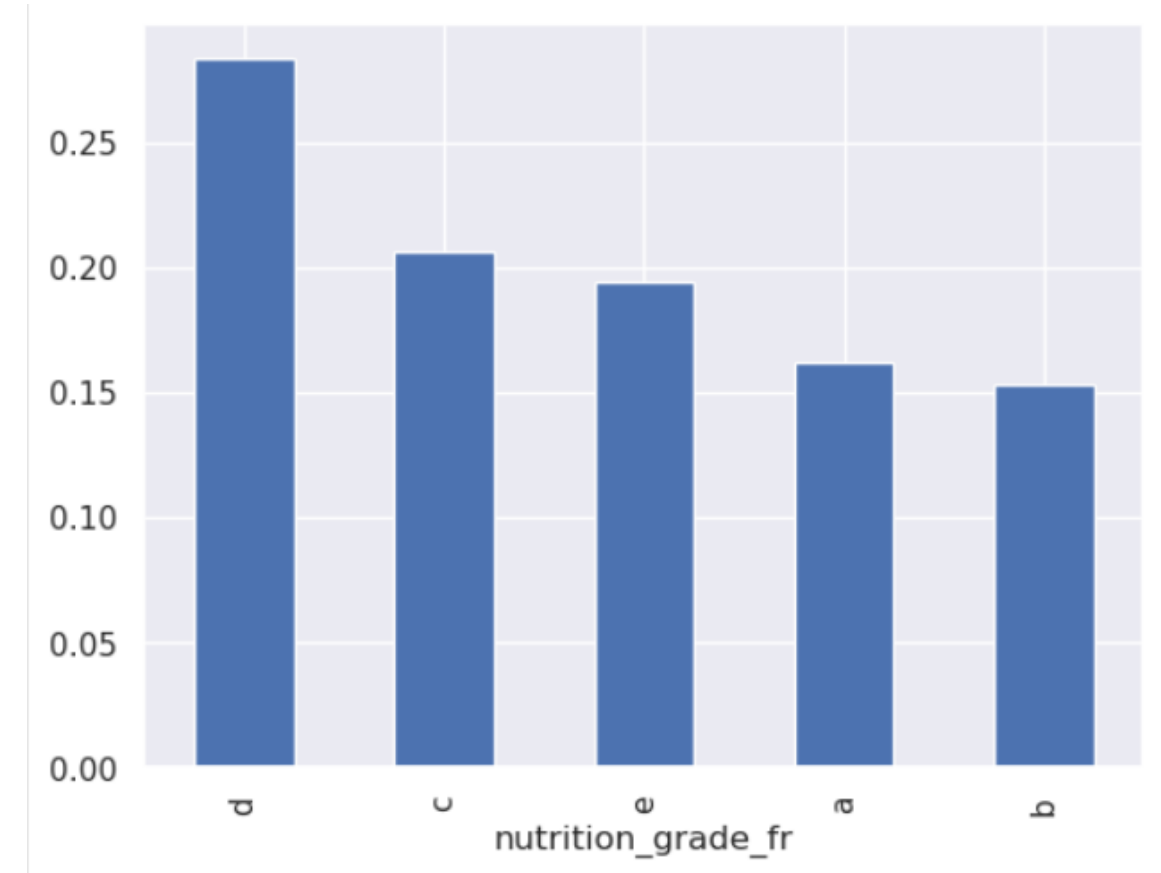
Ici au contraire, on observe que l'aire entre la courbe et la diagonale est grande ; elle est supérieure à l'aire sous la courbe. L'indice de Gini est donc supérieur à 0.5, ce qui indique une grande disparité entre les marques. Cela signifie que certaines marques ont beaucoup de leurs produits enregistrés dans la base de données, et d'autres en ont très peu. Plus précisément, 80% des produits appartiennent à 25% des marques les plus représentées.



La feature cible, *nutrition_grade_fr* :

Ci-contre, sa distribution.

On observe une assez grande disparité dans la répartition des nutri-scores.
La classe D est la plus représentée (~28% des produits) suivie dans l'ordre des classes C, E, A et B.



ANALYSES BIVARIÉES

Nous avons cherché les trois features quantitatives qui permettent de décrire au mieux la cible. Il s'agit des trois features les mieux corrélées à la cible et non corrélées entre elles.

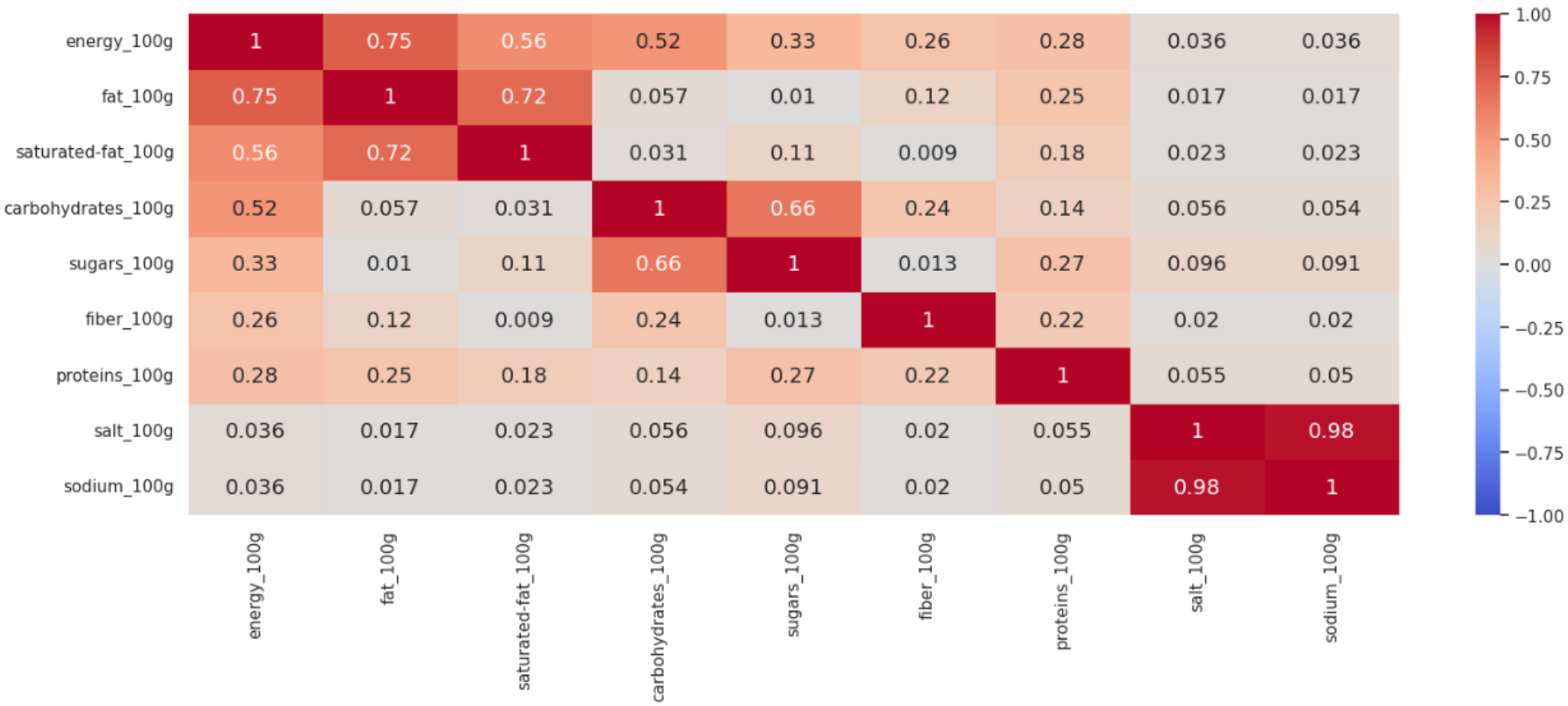
Ci-contre, la heatmap des coefficients de corrélations de Pearson entre les features quantitatives deux à deux.

Coefficient proche de -1 ou 1 --> Forte corrélation

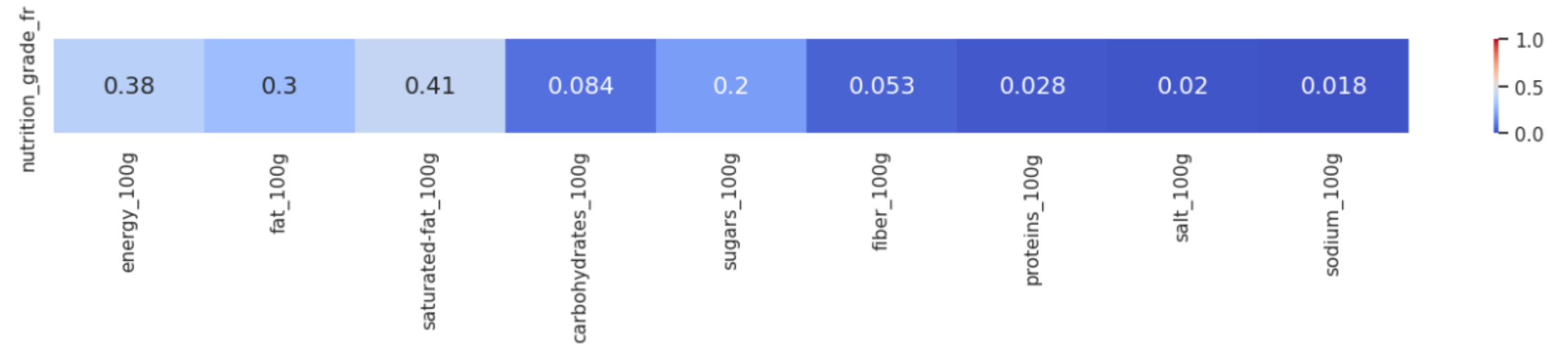
Coefficient proche de 0 --> Faible corrélation

En observant la heatmap, on peut regrouper les features en cinq groupes :

- 1. *energy_100g*, *fat_100g* et *saturated-fat_100g* ;
- 2. *carbohydrates_100g* et *sugars_100g* ;
- 3. *salt_100g* et *sodium_100g* ;
- 4. *fiber_100g* ;
- 5. *proteins_100g*.



Voici la heatmap des rapports de corrélations η^2 entre chaque feature et la cible :



η^2 est compris entre 0 et 1 :

- η^2 très proche de 0 --> faible corrélation ;
- η^2 proche de 1 --> très forte corrélation.

On constate que les trois features les mieux corrélées à la cible sont *energy_100g*, *fat_100g* et *saturated-fat_100g*. Or ces trois features sont corrélées entre elles, donc on ne retient que *saturated-fat_100g* qui est la mieux corrélée à la cible des trois.

Ensuite, les deux features les mieux corrélées à la cible parmi les restantes sont *sugars_100g* et *carbohydrates_100g*. Or ces deux features sont corrélées entre elles, donc on ne retient que *sugars_100g* qui est la mieux corrélée à la cible des deux.

Enfin, la feature la mieux corrélée à la cible parmi les restantes est *fiber_100g*, on la retient donc.

>>> Les trois features permettant de décrire au mieux la cible sont *saturated-fat_100g*, *sugars_100g* et *fiber_100g*.

ANALYSE MULTIVARIÉE

Nous avons réalisé une ACP (Analyse en composantes principales).

Centrage et réduction des données :

Données centrées et réduites avec le Robust Scaler qui est peu sensible aux outliers (que nous avons en grande partie conservé)

Cette méthode consiste à remplacer chaque feature X par la feature $(X - Med)/IQR$, où Med est la médiane des valeurs de X et IQR leur écart-interquartile.

Les features ainsi centrées et réduites ont toutes pour médiane 0 et pour écart-interquartile 1 (cf. tableau ci-contre).

	med	IQR
0	0.0	1.0
1	0.0	1.0
2	0.0	1.0
3	0.0	1.0
4	0.0	1.0
5	0.0	1.0
6	0.0	1.0
7	0.0	1.0
8	0.0	1.0

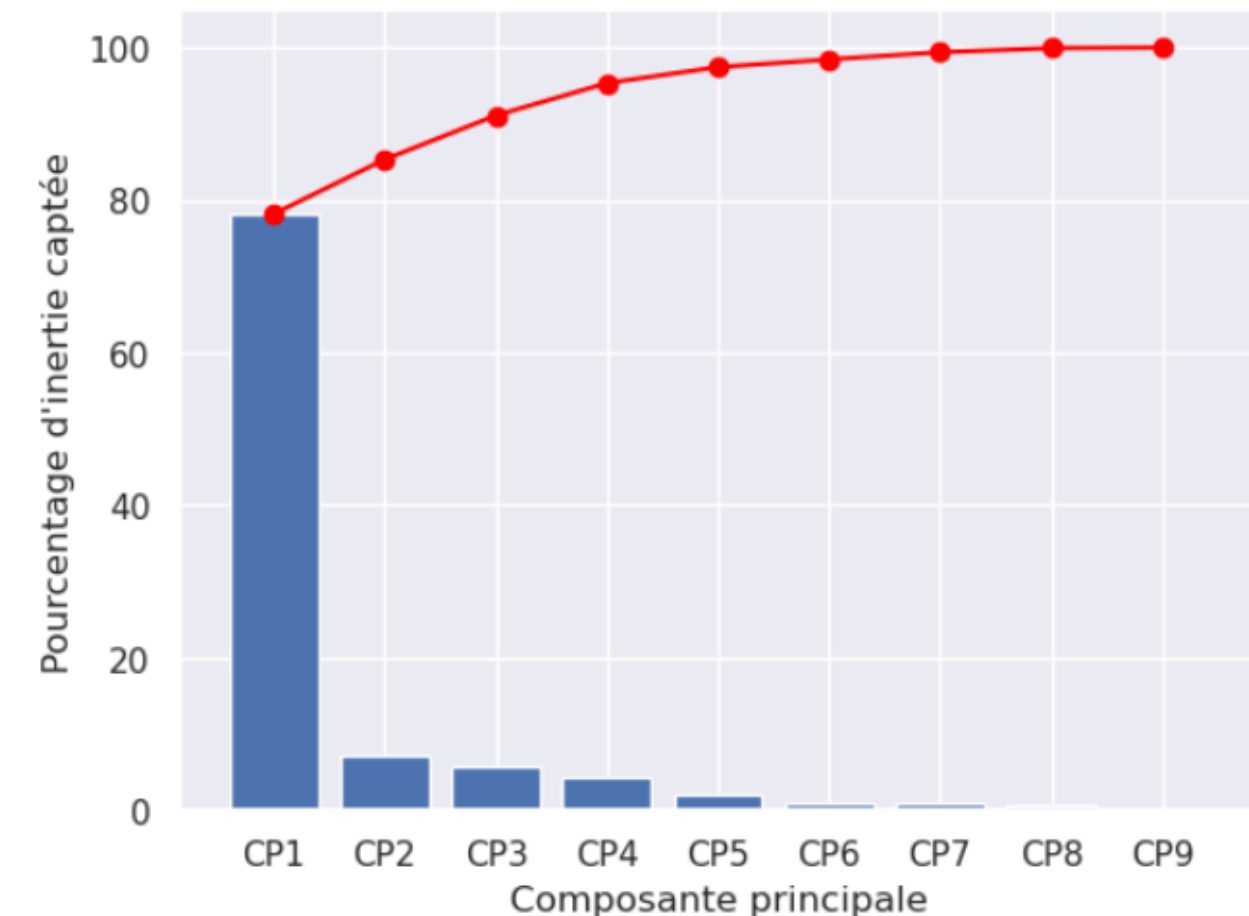
Inertie captée par les composantes principales :

Sur le graphique ci-contre, sont représentés :

- en bleu, les pourcentages d'inertie captée par chaque composantes principales ;
- en rouge, les pourcentages d'inertie captée cumulés.

On observe que la première composantes capte à elle seule plus de trois quarts de l'inertie totale et que les trois premières captent plus de 90% de l'inertie totale.

>>> Les trois premières composantes suffisent amplement pour l'analyse des corrélations et la projection sur les plans factoriels.



Analyse des corrélations :

Voici la heatmap des coefficients de corrélation de Pearson entre chacune des trois premières composantes principales et chaque feature :



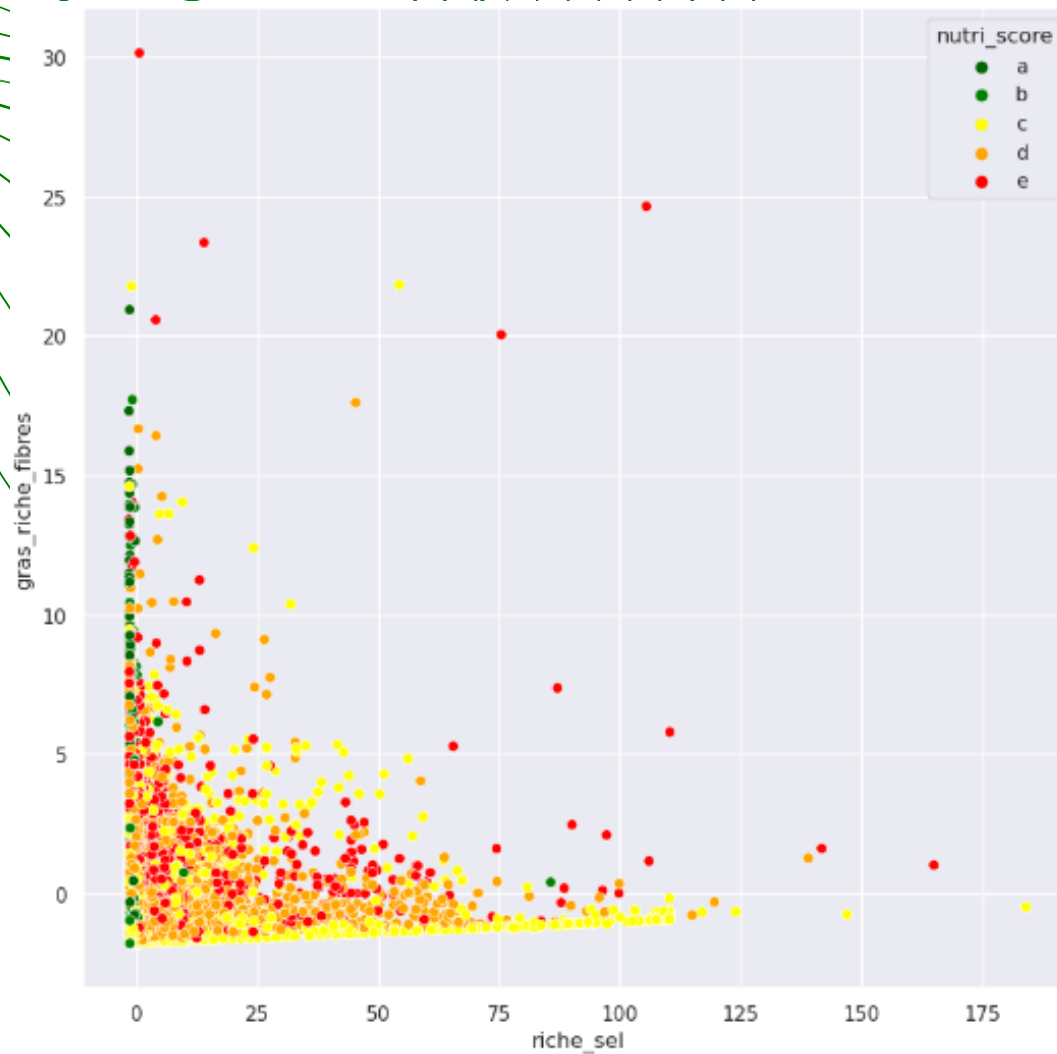
On observe que :

- La CP1 est très fortement corrélée aux features salt_100g et sodium_100g et très faiblement corrélée aux autres features.
 >>> La CP1 représente le caractère **salé** d'un produit.
- La CP2 est fortement corrélée à fiber_100g, energy_100g, fat_100g et saturated-fat_100g. Elle est moins ou peu corrélée aux autres features.
 >>> La CP2 représente le caractère **gras et riche en fibres** d'un produit.
- La CP3 est fortement corrélée à fibers_100g et fortement anti-corrélée à saturated-fat_100g et fat_100g. Elle est moins ou peu corrélée aux autres features.
 >>> La CP3 représente le caractère **sec et riche en fibres** d'un produit.

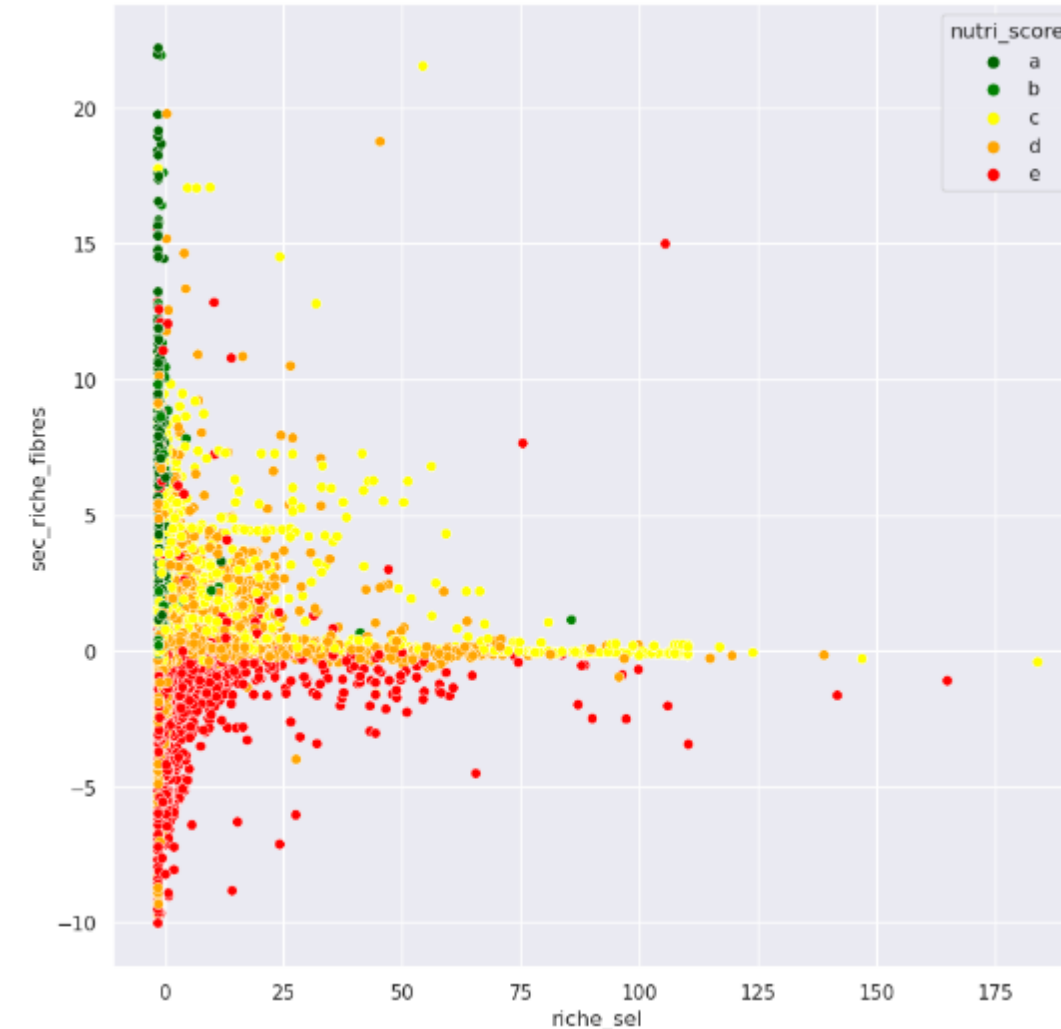
Projection du nuage des produits sur les plans factoriels :

Chaque point a été projeté et est coloré selon son Nutri-Score. Voici les trois projections :

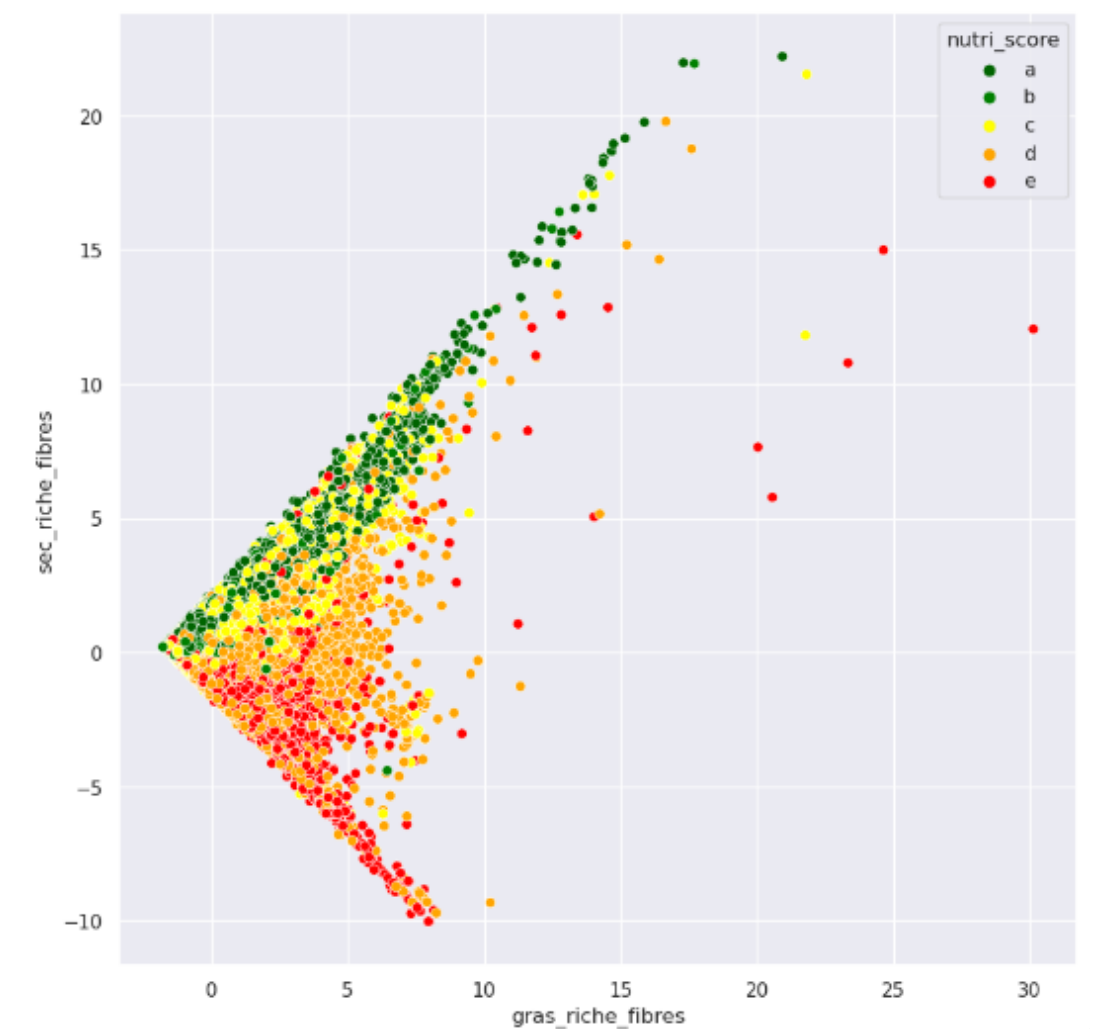
Plan (riche_sel, gras_riche_fibres)



Plan (riche_sel, sec_riche_fibres)



Plan (gras_riche_fibres, sec_riches_fibres)



Les classes nutritionnelles sont assez bien regroupées et clairement discernables.

- Les produits classés A ou B (en vert) sont généralement les produits riches en fibres et pauvre en sel ;
- Les produits classés C (en jaune) sont généralement les produits plutôt salés, pauvres en graisses et en fibres ou assez gras mais riches en fibres ;
- Les produits classés D (en orange) sont généralement les produits moyennement salés, moyennement gras et pauvres en fibres ;
- Les produits classés E (en rouge) sont généralement les produits moyennement salés, gras et très pauvres en fibres.

4. Conclusion

Une application suggérant les valeurs manquantes pour une feature semblable au Nutri-Score est un projet réaliste qui peut-être raisonnablement envisagé et réalisé.

Nous entendons par 'semblable au Nutri-Score' une feature donnant une bonne indication de la qualité nutritive d'un produit, qu'elle soit quantitative ou qualitative avec un nombre restreint de modalités et que son taux de valeurs manquantes soit inférieur à 50%.

4. Au sujet du RGPD

Ce travail respecte le Règlement Général sur la Protection des Données (RGPD). Il a été réalisé sur des données libres et ouvertes mises à disposition par l'organisation à but non lucratif Open Food Facts sur son site internet.

Ces données ne contiennent pas de données personnelles ou privées et sont extraites des informations inscrites sur les emballages des produits alimentaires.

Ces données sont mises à disposition de tous pour un usage libre et conformes aux lois en vigueur.