Préparer les données pour un organisme de santé publique

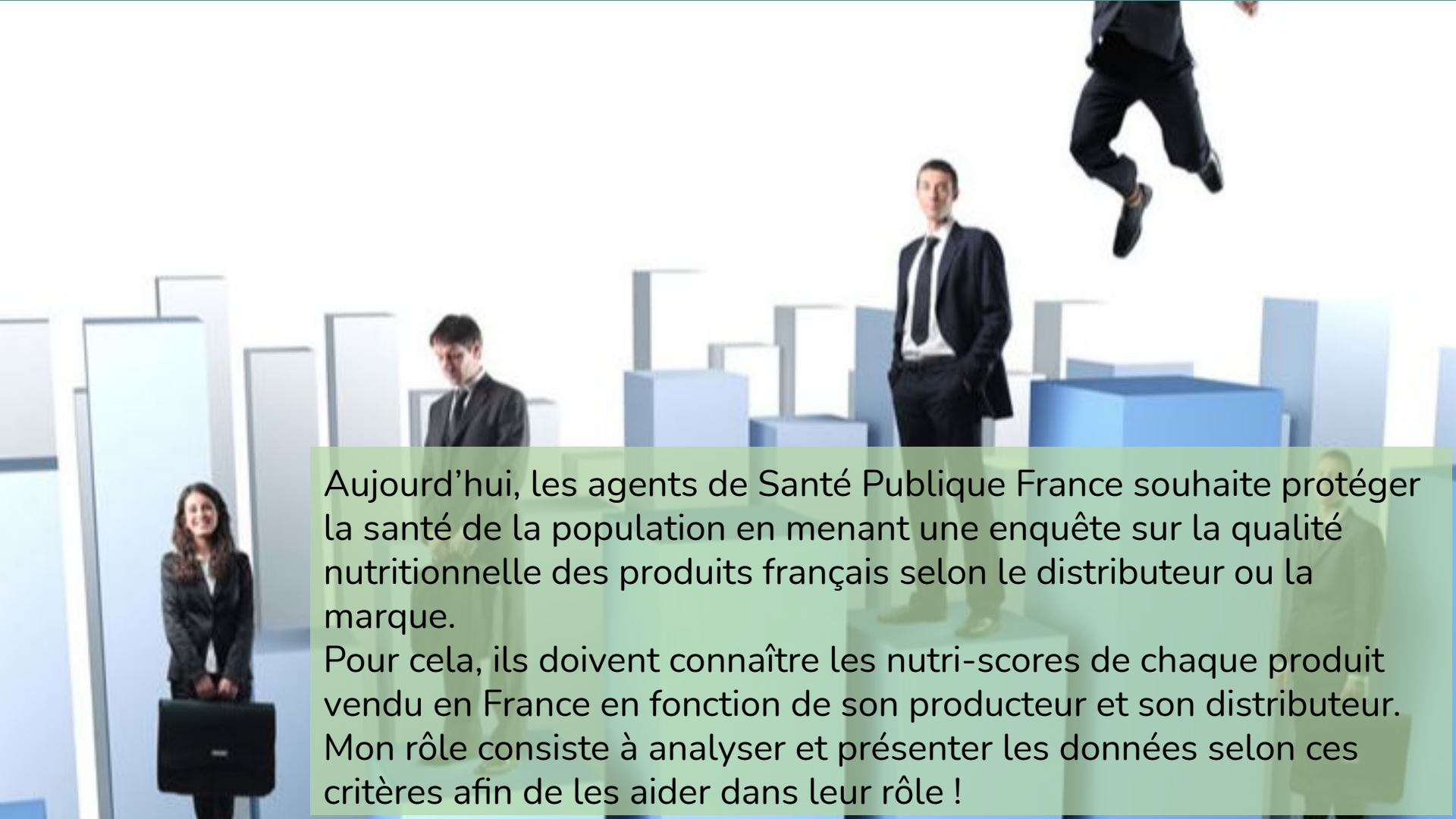
Moussa KIBALY

Santé publique France

SOMMAIRE

1. CONTEXTE DU PROJET

Les données de santé de produits alimentaires doivent être analysées et présentées aux agents de manière simple et compréhensibles.
Pour cela, une analyse exploratoire univariée et multivariée sera faite sur le jeu de données des produits alimentaires français accessibles sur le site Open Food.



Aujourd'hui, les agents de Santé Publique France souhaite protéger la santé de la population en menant une enquête sur la qualité nutritionnelle des produits français selon le distributeur ou la marque.
Pour cela, ils doivent connaître les nutri-scores de chaque produit vendu en France en fonction de son producteur et son distributeur.
Mon rôle consiste à analyser et présenter les données selon ces critères afin de les aider dans leur rôle !

La connaissance des scores nutrition en fonction des distributeurs vont permettre aux agents de Santé Publique France d'élaborer leur enquête sur la qualité nutritionnelle de leurs produits alimentaires et éventuellement communiquer avec les distributeurs.

La connaissance des scores nutrition en fonction des marques alimentaires vont de même permettre aux agents d'étoffer leur enquête sur la qualité nutritionnelle des produits fabriqués en France.

Grâce à notre analyse exploratoire, l'agent de santé sait quels sont les distributeurs et marques avec les meilleurs scores de nutrition.

Les xxx ont les meilleurs scores de nutrition.

2. PRÉSENTATION GÉNÉRALE DU JEU DE DONNÉES

Le jeu de données des produits alimentaires (fichier p2-arbres-fr.csv) contient 143 colonnes et 320772 lignes de produits alimentaires français et américain.

Le fichier contient des informations sur le nom, la marque, le distributeur de chaque produit ainsi que ses données nutritionnelles telles que la quantité de sucre, de graisse, de cholestérol, de vitamine a. b, c, d, e, de caféine, de fibre.

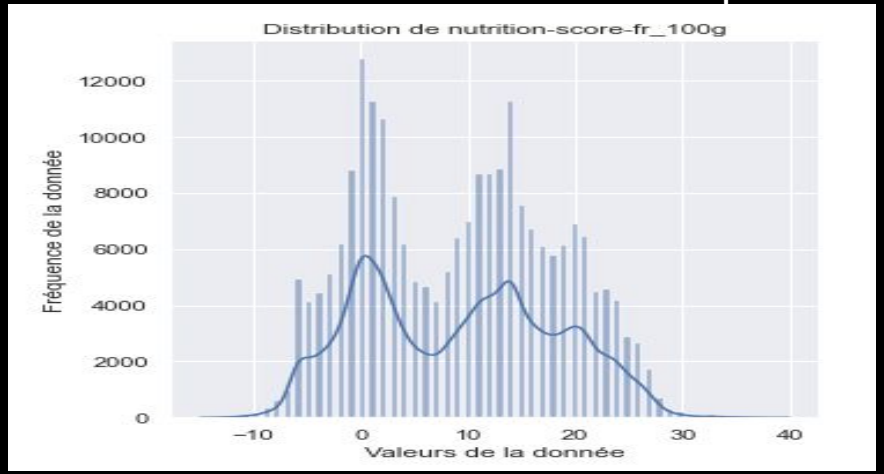
76.6% des cellules de notre jeu de données contiennent des données manquantes NaN ou à 0 qu'il faut traiter pour notre analyse. On ne va retenir que certaines données et le reste des colonnes va être supprimé.

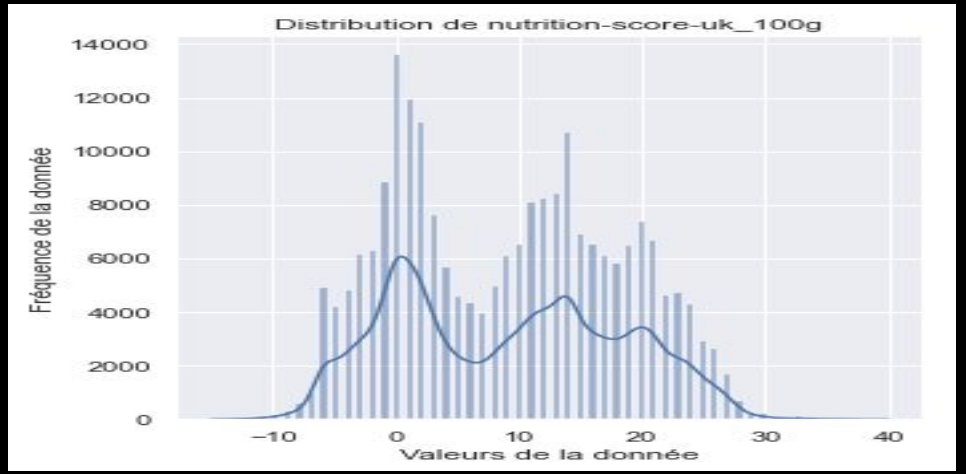
- Indicateurs statistiques du jeu de données avant leur traitement

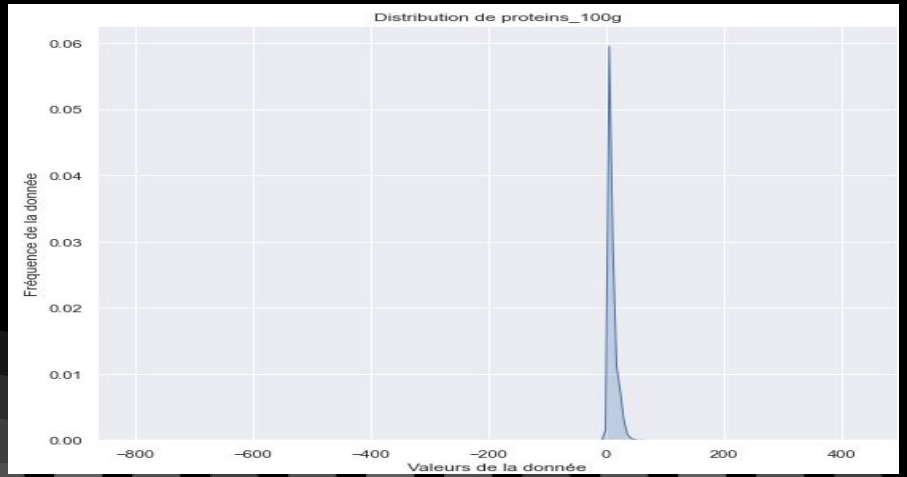
quantity	...	chromium_100g	molybdenum_100g	iodine_100g	caffeine_100g	taurine_100g	ph_100g	fruits-vegetables-nuts_100g	carbon-footprint_100g	nutrition-score-fr_100g	nutrition-score-uk_100g
104784	...	20.000000	11.000000	259.000000	78.000000	29.000000	49.000000	3036.000000	268.000000	221210.000000	221210.000000
13824	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
500 g	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4669	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	...	0.001690	0.000401	0.000427	1.594563	0.145762	6.425698	31.458587	341.700764	9.165535	9.058049
NaN	...	0.006697	0.001118	0.001285	6.475588	0.172312	2.047841	31.967918	425.211439	9.055903	9.183589
NaN	...	0.000007	0.000005	0.000000	0.000000	0.001800	0.000000	0.000000	0.000000	-15.000000	-15.000000
NaN	...	0.000011	0.000020	0.000015	0.015500	0.035000	6.300000	0.000000	98.750000	1.000000	1.000000
NaN	...	0.000023	0.000039	0.000034	0.021000	0.039000	7.200000	23.000000	195.750000	10.000000	9.000000
NaN	...	0.000068	0.000074	0.000103	0.043000	0.400000	7.400000	51.000000	383.200000	16.000000	16.000000
NaN	...	0.030000	0.003760	0.014700	42.280000	0.423000	8.400000	100.000000	2842.000000	40.000000	40.000000

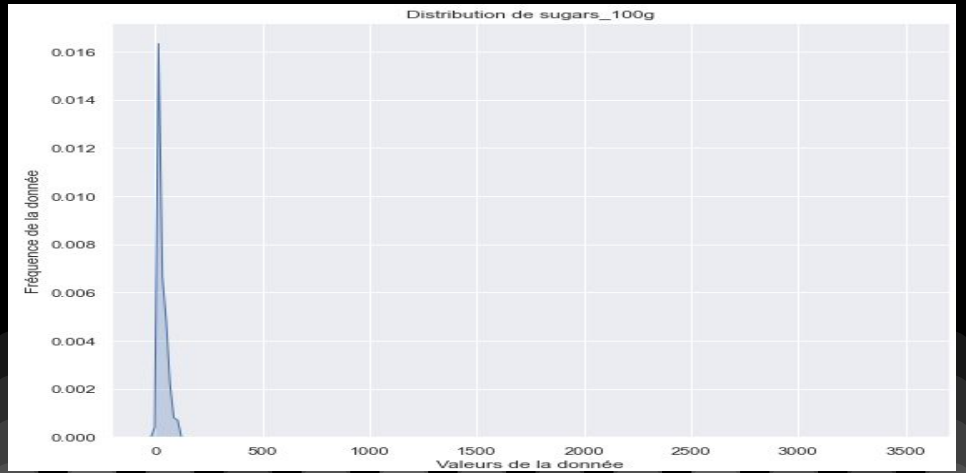
- On observe des valeurs aberrantes pour **la donnée carbon-footprint_100g** avec un maximum de 2842.

- Distribution des variables quantitatives









On observe une symétrie des distributions des scores nutrition fr et uk. Cela est mis en évidence l'indicateur du skewness empirique qui est très faible et égale à 0.11 pour fr. Le kurtosis empirique de ces variables est négatif ce qui signifie un aplatissement important de ces distributions.

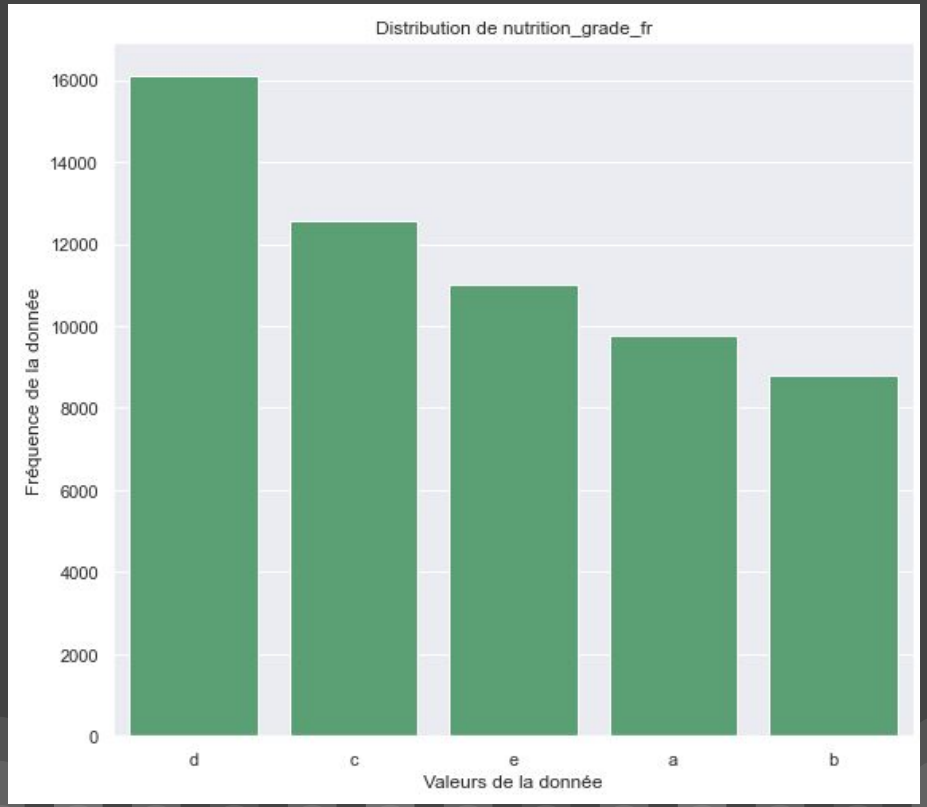
Pour les données **'proteins_100g' et 'sugars_100g'**, on observe une asymétrie de la distribution vers la droite due aux valeurs aberrantes supérieures à 100g. Leurs distributions présentent un pic qui se reflète par un kurtosis empirique très important et positif.

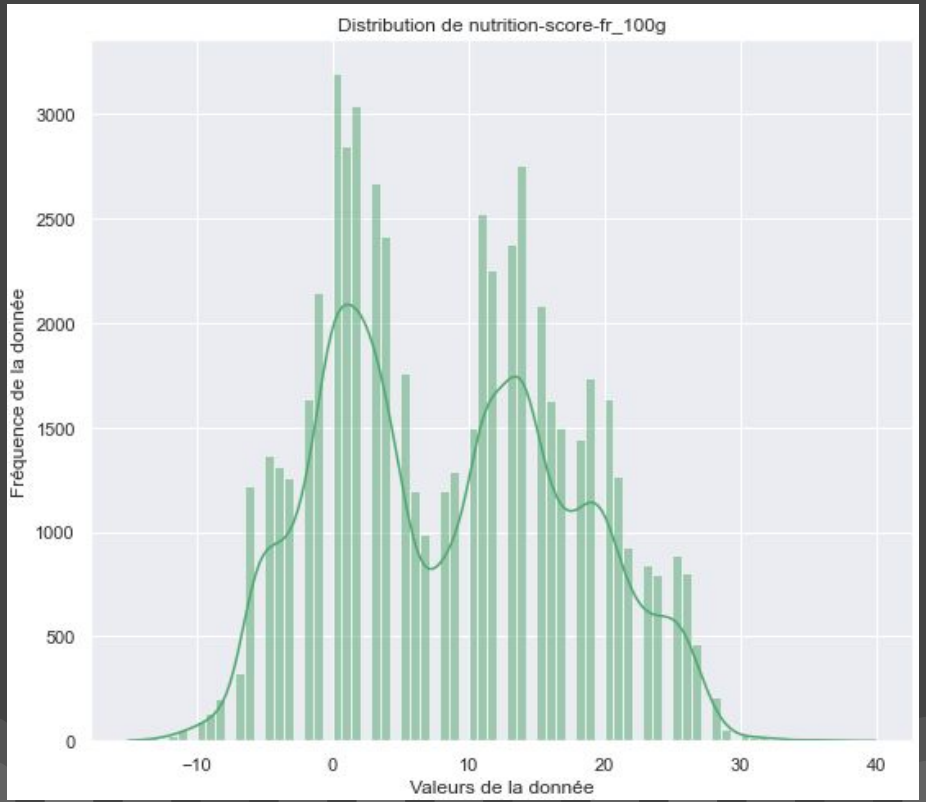
NETTOYAGE ET TRAITEMENT DES DONNÉES

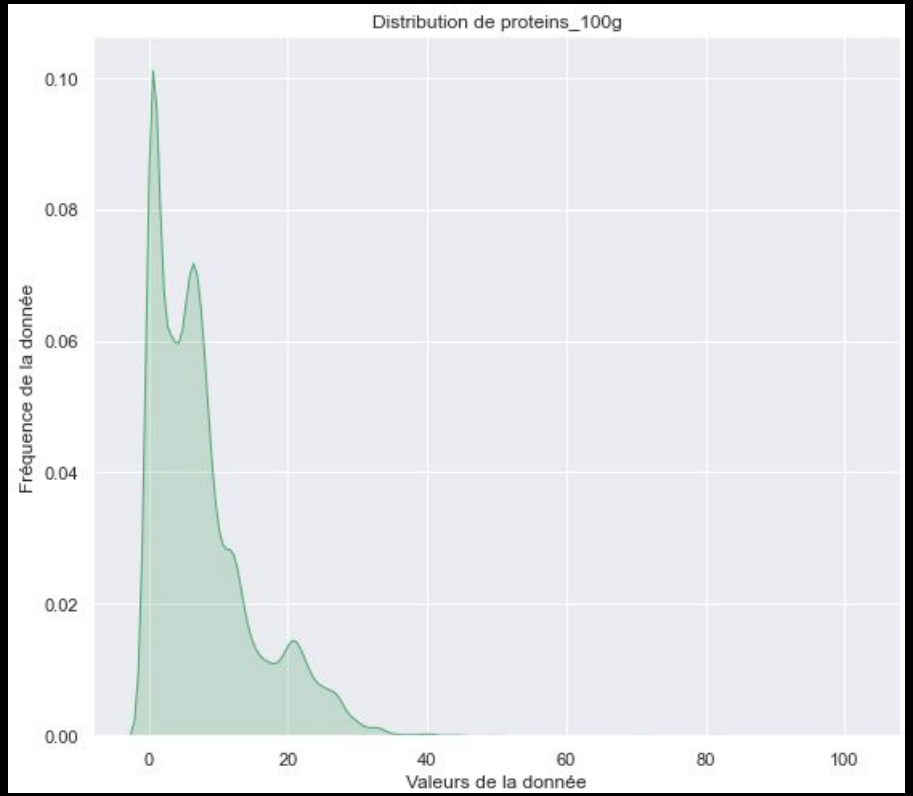
On a supprimé les 23 lignes de doublons du jeu de donnés et on ne sélectionne uniquement les 58197 produits français.

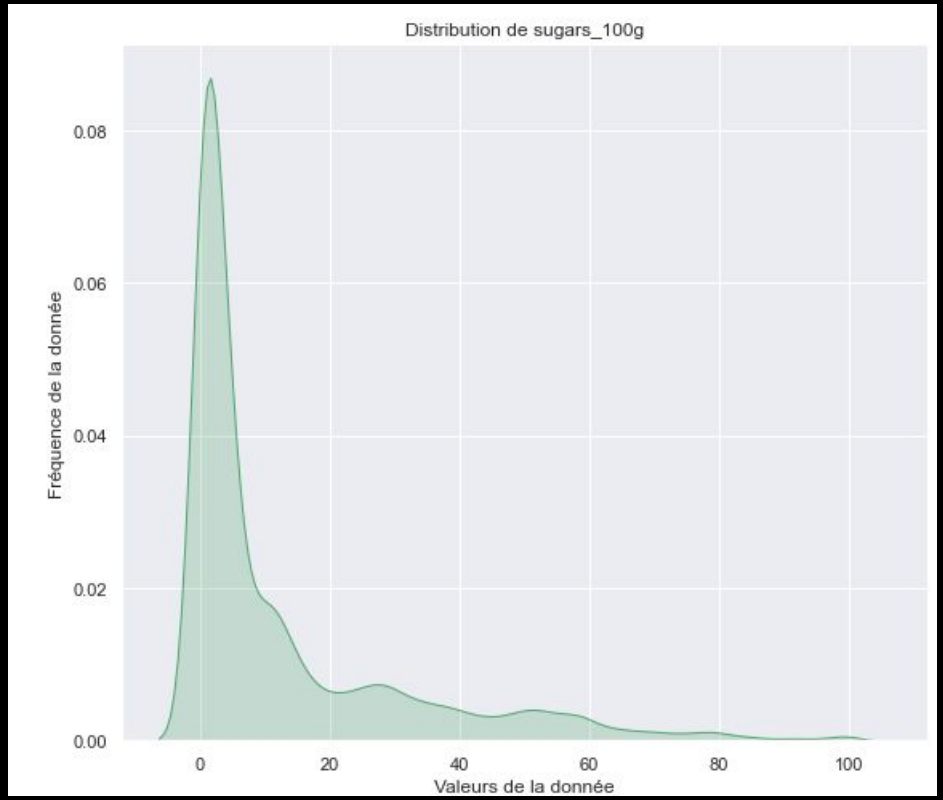
Les lignes dont la colonne product_name n’est pas renseignée représentent 7% des lignes totales donc leur suppression n’influence pas notre analyse.

ANALYSE STATISTIQUE UNIVARIÉE(APRÈS NETTOYAGE)

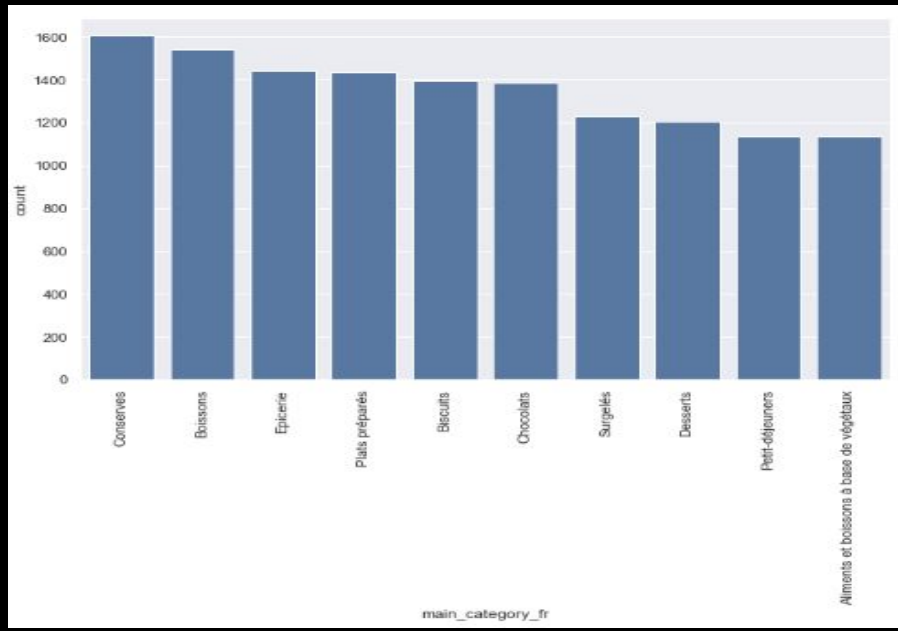


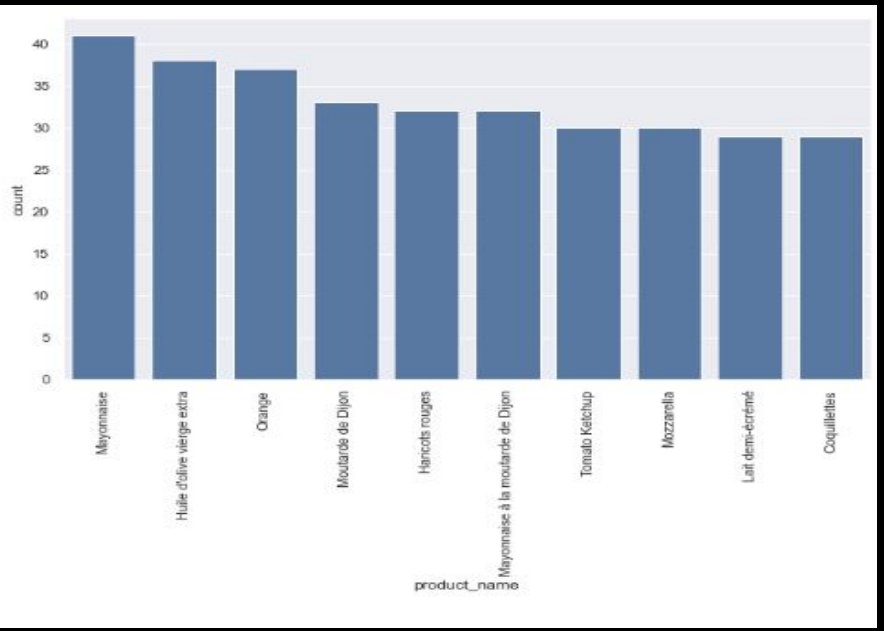



Distribution de proteins_100g
Fréquence de la donnée
0.00
0.02
0.04
0.06
0.08
0.10
0
20
40
60
80
100
Valeurs de la donnée


Distribution de sugars_100g
Fréquence de la donnée
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
Valeurs de la donnée

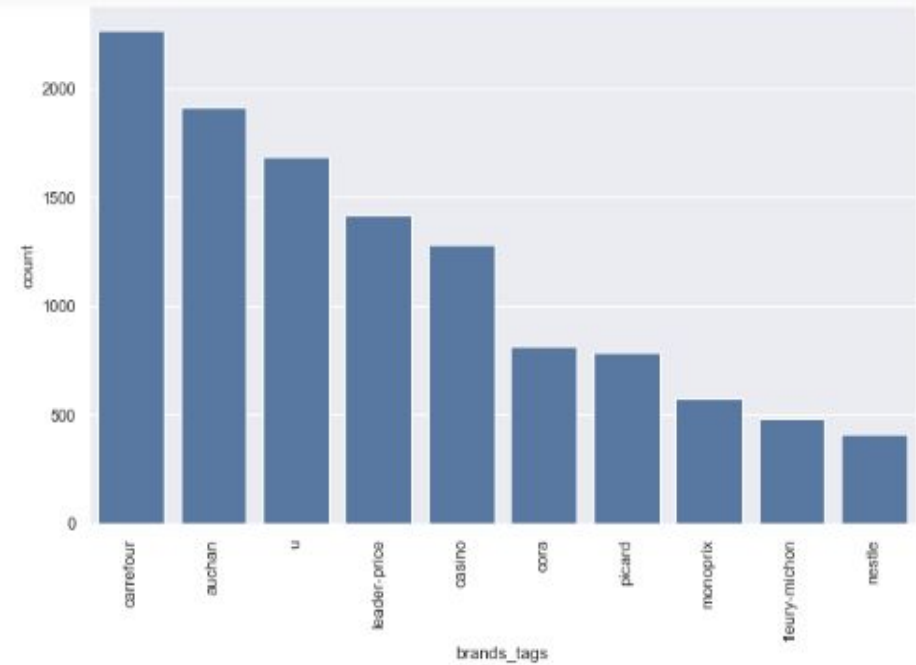
Distribution des variables qualitatives

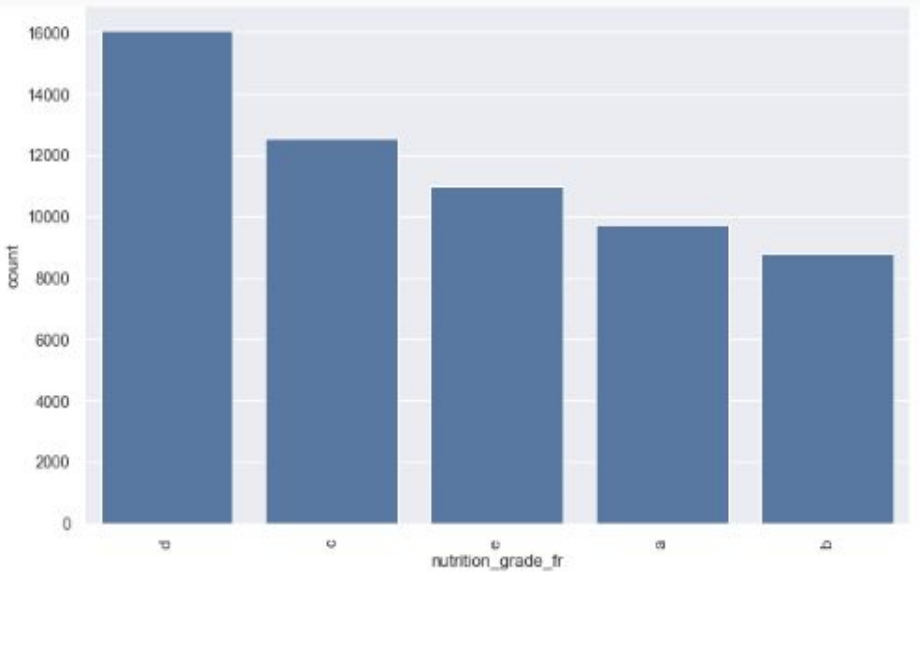




La grande majorité des catégories sont des conserves, boissons et des produits d’ épicerie.

Les produits en plus grand nombre sont la mayonnaise, l’huile d’olive et l’orange.

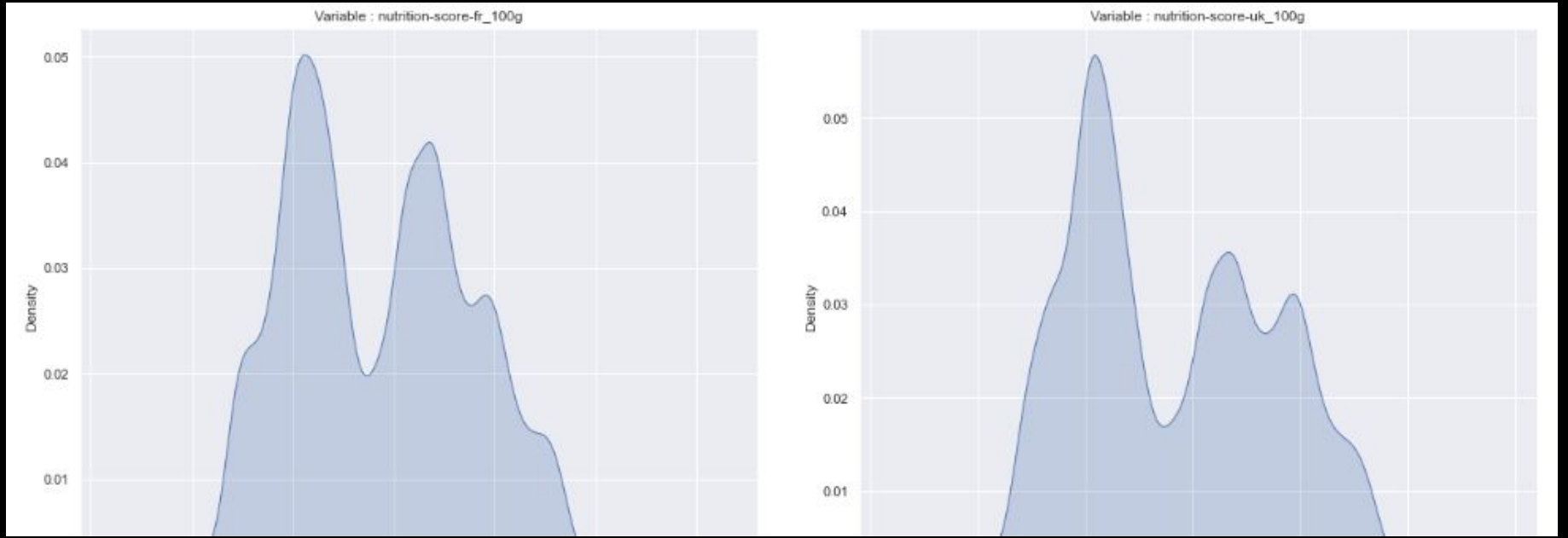




Carrefour, Auchan et U vendent la majorité des produits français.

Beaucoup de produits ont un score nutrition entre D et C ce qui est plutôt une qualité moyenne.

Distribution des variables quantitatives nutrition-score-fr_100g, nutrition-score-uk_100g, proteins_100g et sugars_100g



Les données nutrition-score-fr_100g et nutrition-score-uk_100g ont la même distribution avec une concentration des scores entre -10 et +30.

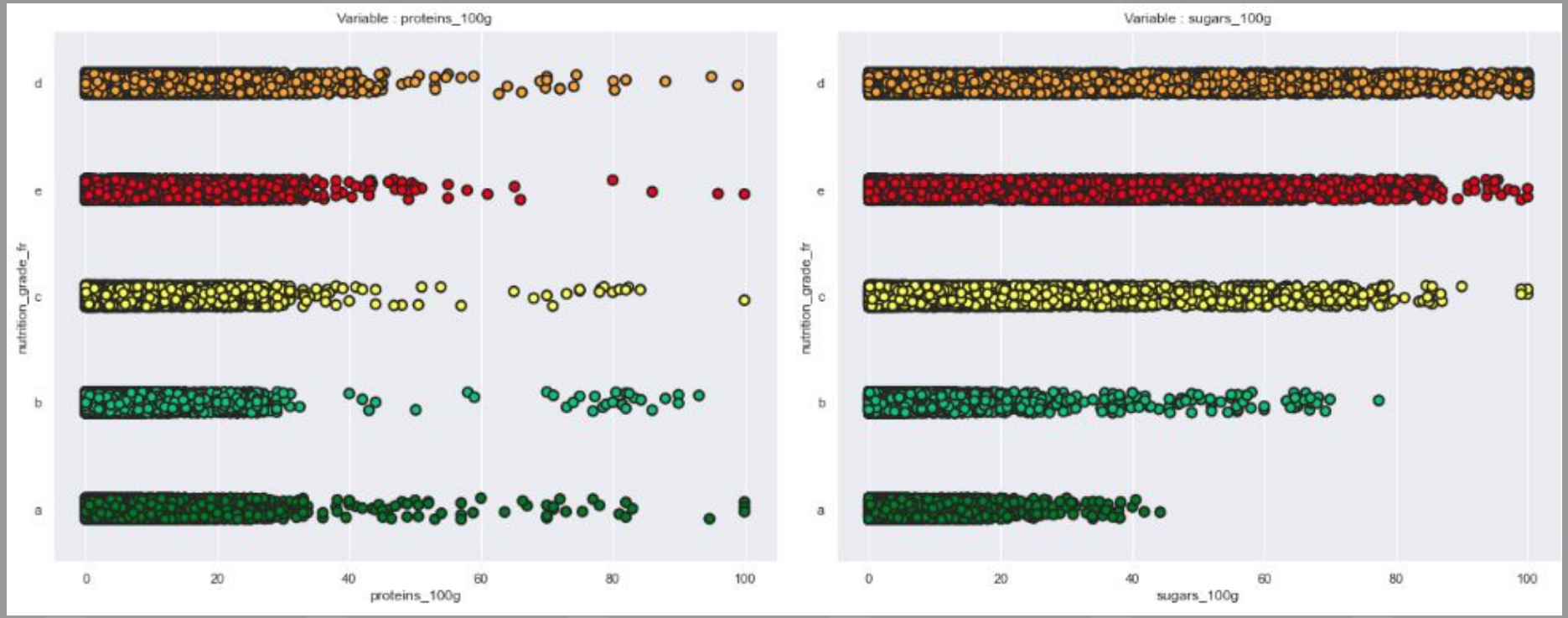
Cherchons les éventuelles corrélations entre les données nutritionnelles et les scores. Le coefficient de corrélation de Pearson est calculé entre chaque donnée.

	nutrition-score-fr_100g	nutrition-score-uk_100g	proteins_100g	sugars_100g	fat_100g	saturated-fat_100g	omega-3-fat_100g	omega-6-fat_100g	omega-9-fat_100g	cholesterol_100g	...	vitamin-b6_100g	vita b9_
nutrition-score-fr_100g	1.000000	0.962746	0.099778	0.444574	0.425356	0.599359	0.004462	0.009134	0.004155	0.008559	...	0.001579	-0.00
nutrition-score-uk_100g	0.962746	1.000000	0.166291	0.411894	0.493259	0.642493	0.020458	0.031901	0.018051	0.008453	...	0.000871	-0.00
proteins_100g	0.099778	0.166291	1.000000	-0.252772	0.095845	0.145525	0.003144	-0.028341	-0.015258	-0.001122	...	-0.000284	0.01
sugars_100g	0.444574	0.411894	-0.252772	1.000000	-0.029120	0.070375	-0.033507	-0.021018	-0.010760	0.008399	...	0.009533	-0.00
fat_100g	0.425356	0.493259	0.095845	-0.029120	1.000000	0.531817	0.138912	0.157809	0.085423	0.008803	...	0.002437	-0.00
saturated-fat_100g	0.599359	0.642493	0.145525	0.070375	0.531817	1.000000	0.010094	0.021580	0.014971	0.008599	...	-0.004335	-0.00
omega-3-fat_100g	0.004462	0.020458	0.003144	-0.033507	0.138912	0.010094	1.000000	0.271884	0.065345	0.000085	...	-0.000650	-0.00
omega-6-fat_100g	0.009134	0.031901	-0.028341	-0.021018	0.157809	0.021580	0.271884	1.000000	0.205760	-0.000163	...	-0.000378	-0.00

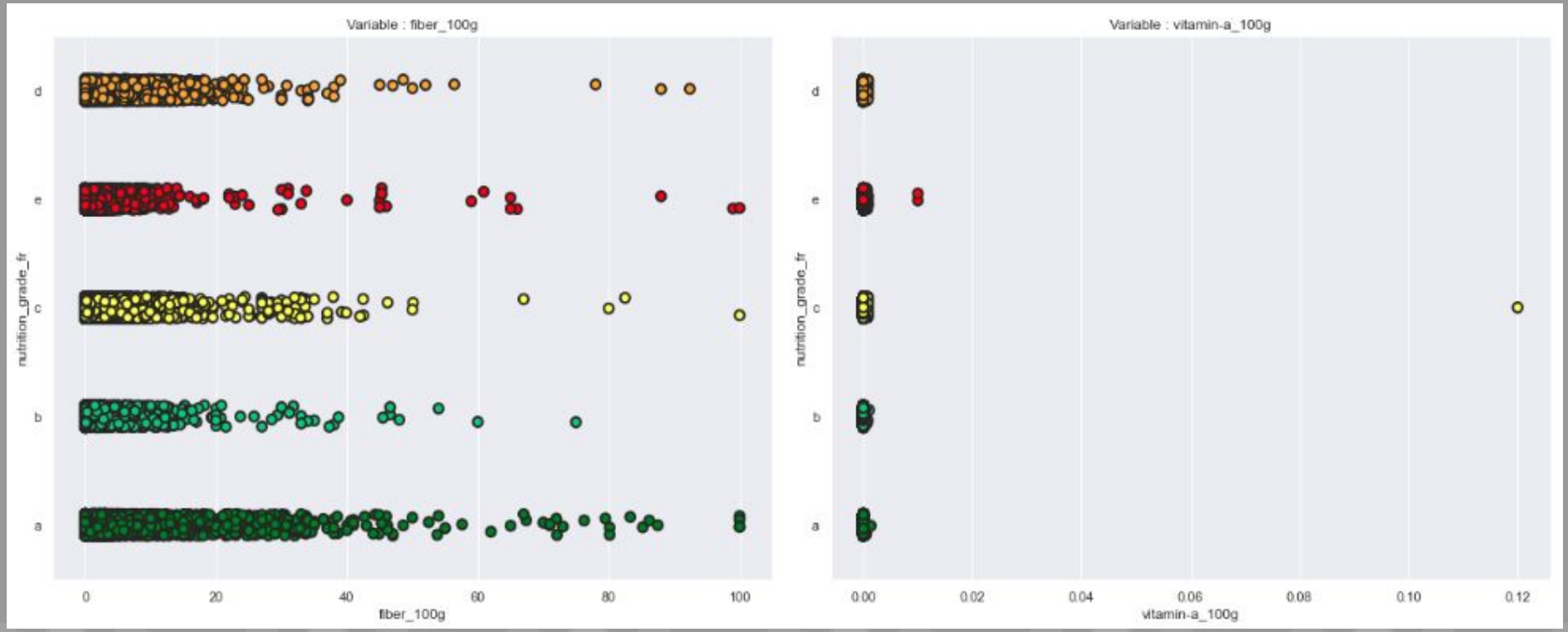
```
Il y a 5 variables corr. avec le nutritionscorefr:
fat_100g                  0.425356
sugars_100g               0.444574
saturated-fat_100g        0.599359
nutrition-score-uk_100g   0.962746
nutrition-score-fr_100g   1.000000
Name: nutrition-score-fr_100g, dtype: float64
```

On constate une forte corrélation entre le score nutrition fr ou uk avec les quantités fat_100g, sugars_100g, saturated-fat_100g,

Sur ces graphiques, on observe la répartition des scores nutrition allant de ‘a’ à ‘e’ suivant la quantité de sucre et de protéine.



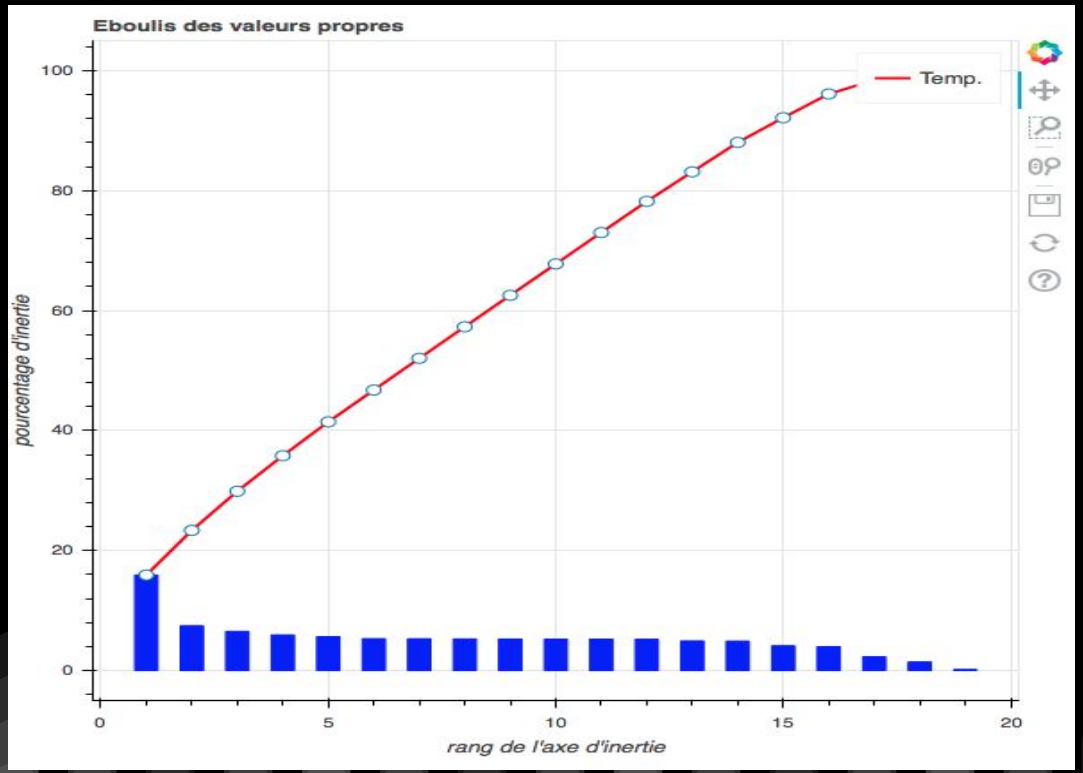
Sur ces graphiques, on observe la distribution des quantités de fibre et de vitamine a en fonction du nutri score.



3. ANALYSE STATISTIQUE MULTIVARIÉE

Analyse en composantes principales (ACP)

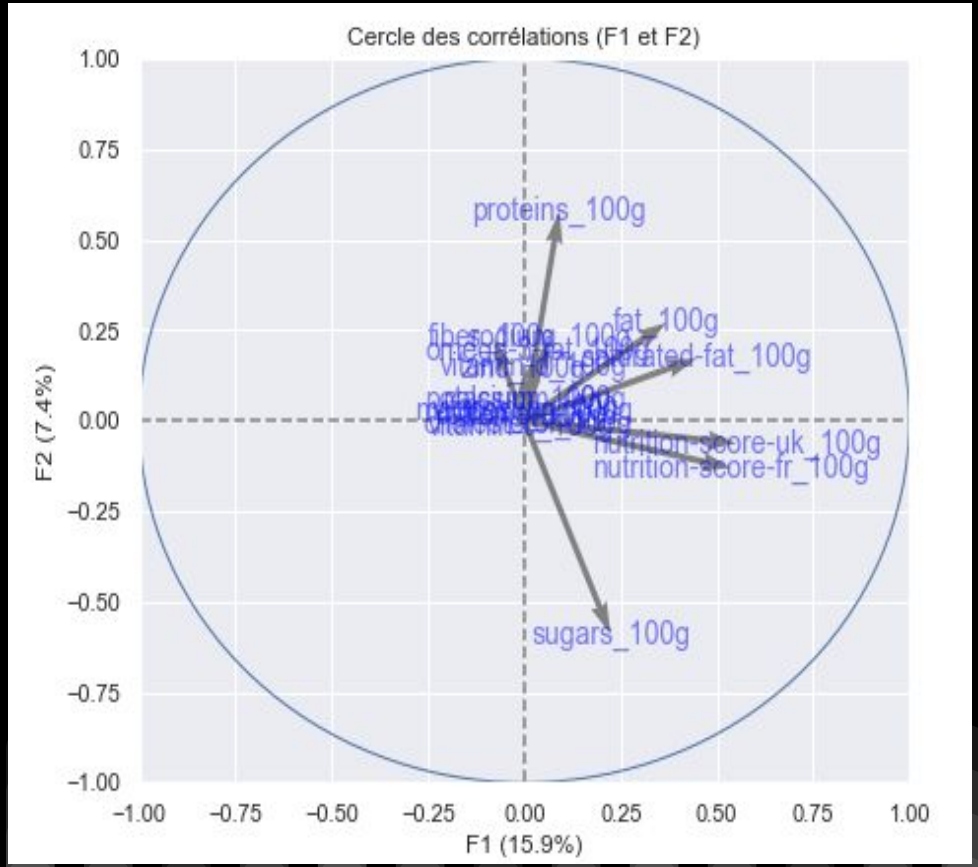
On va partir de 19 variables nutritionnelles pour effectuer l'ACP afin de réduire le nombre de variables, voir d'éventuelles corrélations entre variables et variabilités entre individus.



Les 19 composantes principales sont calculés et on affiche le pourcentage d'inertie des individus que porte chaque axe principal (éboulis des valeurs propres).

On va se limiter aux 12 premiers axes principaux d'inertie.

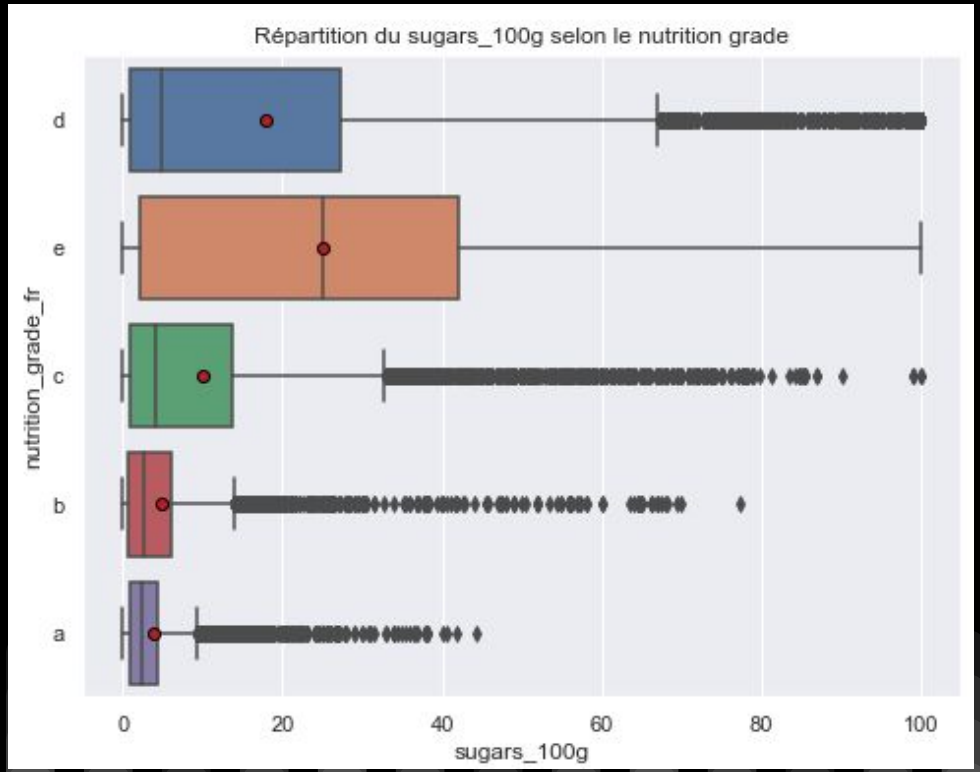
Projection du nuage des 19 variables sur les 6 plans factoriels. On voit ici le cercle de corrélation sur le premier plan factoriel.



Ls variables les plus corrélés positivement à F1 sont : nutrition-score-fr_100g, nutrition-score-uk_100g, fat_100g, saturated-fat_100g. La variable fiber_100g est corrélée négativement à F1. On peut interpréter F1 comme un indicateur de MAUVAISE QUALITÉ NUTRITIONNELLE. F2 peut être interpréter comme un indicateur de BONNE QUALITÉ NUTRITIONNELLE.

Analyse explicative (ANOVA)

Les quantités de sucre sont très différentes d'un grade nutrition à un autre. Pour le score 'e', les quantités sont plus élevées et plus dispersées.



On part de l'hypothèse nulle H0 = les produits avec un nutri score 'e' contiennent peu de sucre.

En calculant la p-value, on va rejeter cette hypothèse nulle.

On affiche la p-value par le biais de la méthode ANOVA appliquée entre les variables nutrition_grade_fr et sugars_100g puis fat_100g.

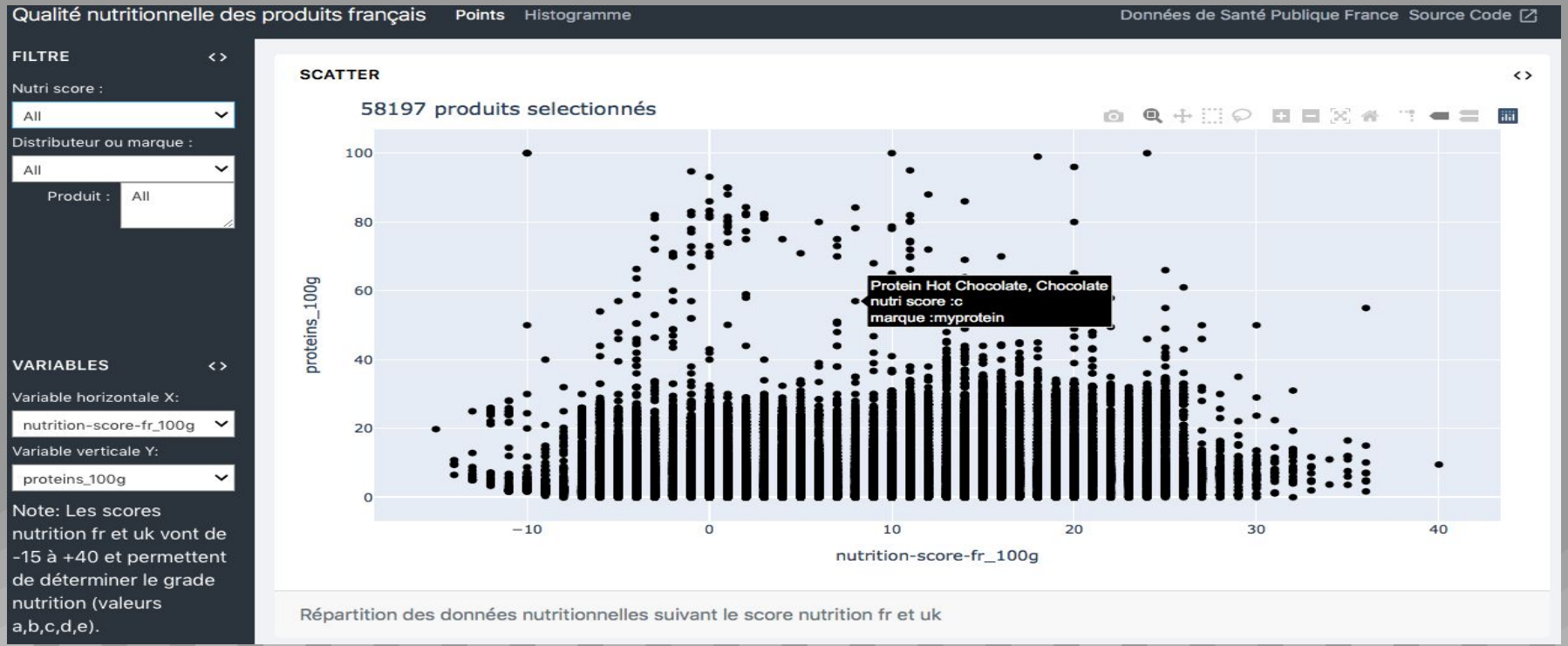
	sum_sq	df	F	PR(>F)
C(nutrition_grade_fr)	5.462335e+06	4.0	5388.480834	0.0
Residual	2.949605e+07	116389.0	NaN	NaN

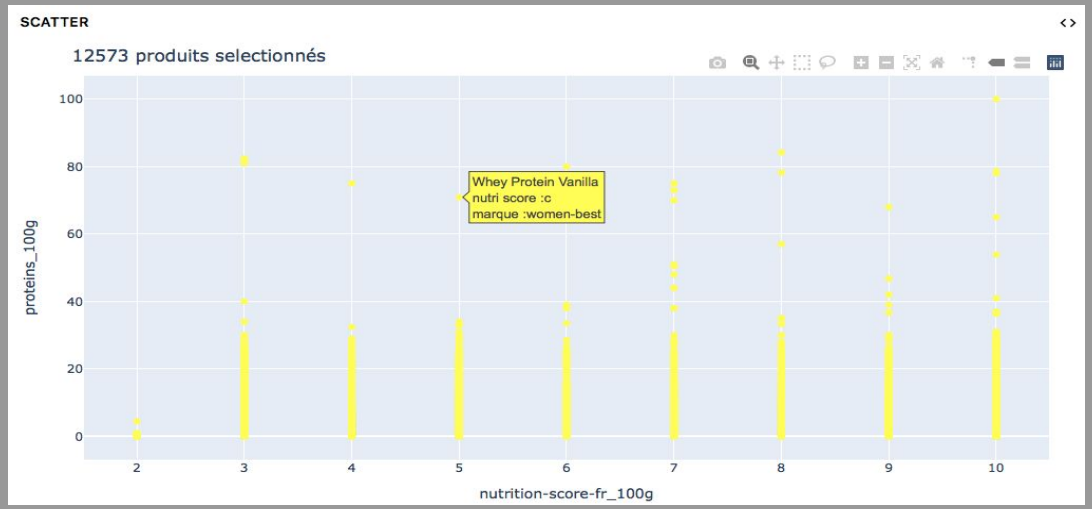
Le premier tableau montre une p-value < 0.05. On rejette l'hypothèse nulle H0.

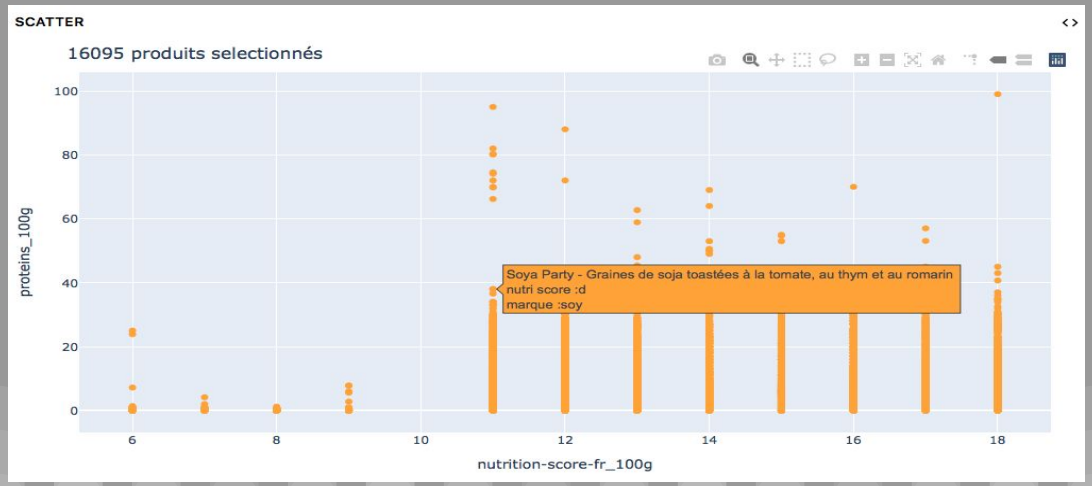
Le second graphique met bien en évidence l'influence négative de la quantité de sucre et de graisse sur le grade nutrition.

4. SYNTHÈSE DE L'ANALYSE DE DONNÉES

Sur ce graphique dynamique, on observe la distribution des scores nutrition fr en fonction de la marque, du nom du produit, du distributeur, des données nutritionnelles et cela apporte les informations dont les agents de Santé Publique France ont besoin.







On constate que 21.6% des produits français ont un grade nutrition de C et 27.6% un grade de D. L'agence doit communiquer aux fabricants d'améliorer la qualité de leur produit.

Carrefour et Auchan regroupe près de 4% des produits de nutrition score C et D.