

NUTRISCORE APP

CHOISISSEZ LES MEILLEURS PRODUITS POUR
VIVRE LONGTEMPS

Présenté par Kokou Sitsopé Sekpona



Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au succès de mon projet et qui m'ont aidée lors de la rédaction de ce mémoire.

Je voudrais dans un premier temps remercier, mon mentor de M.Chakib Belafdil, pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribué à alimenter ma réflexion.

Je remercie également toute la communauté du workplace et les intervenants professionnels, pour leur réponses à mes différentes questions.



Idee d'application

Une application pour prédire le nutriscore des aliments

L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation. La plupart des maladies comme l'hypertension artérielle, le diabète, le cancer et autres sont due en majorité à une mauvaise alimentation Nous allons donc concevoir une application permettant prédire a partir des données mises sur l'etiquette, le score nutritionnelle des produits qui n'en portent pas



Chronologie du projet



Nettoyage des données

Nous allons procéder au nettoyage des données par le traitement des valeurs manquantes, des doublons, des outliers et apparemment des valeurs atypiques

Analyse exploratoire de données

Par une analyse univariée et bivariée et multivariées nous allons traiter les variables et déduire les différentes relations existantes entre eux pour mieux comprendre la base de données

Evaluation de la faisabilité de l'application

Nous allons voir si après le nettoyage et l'analyse, la base de données nous permettra de réaliser notre application.

Nettoyage de données

Traitement des valeurs manquantes

Par 3 méthodes différentes

Traitement des doublons

Suppression d'une partie des
individus qui se répètent dans le jeu
de données

Traitement des Outliers

Valeurs aberrantes et valeurs
atypiques

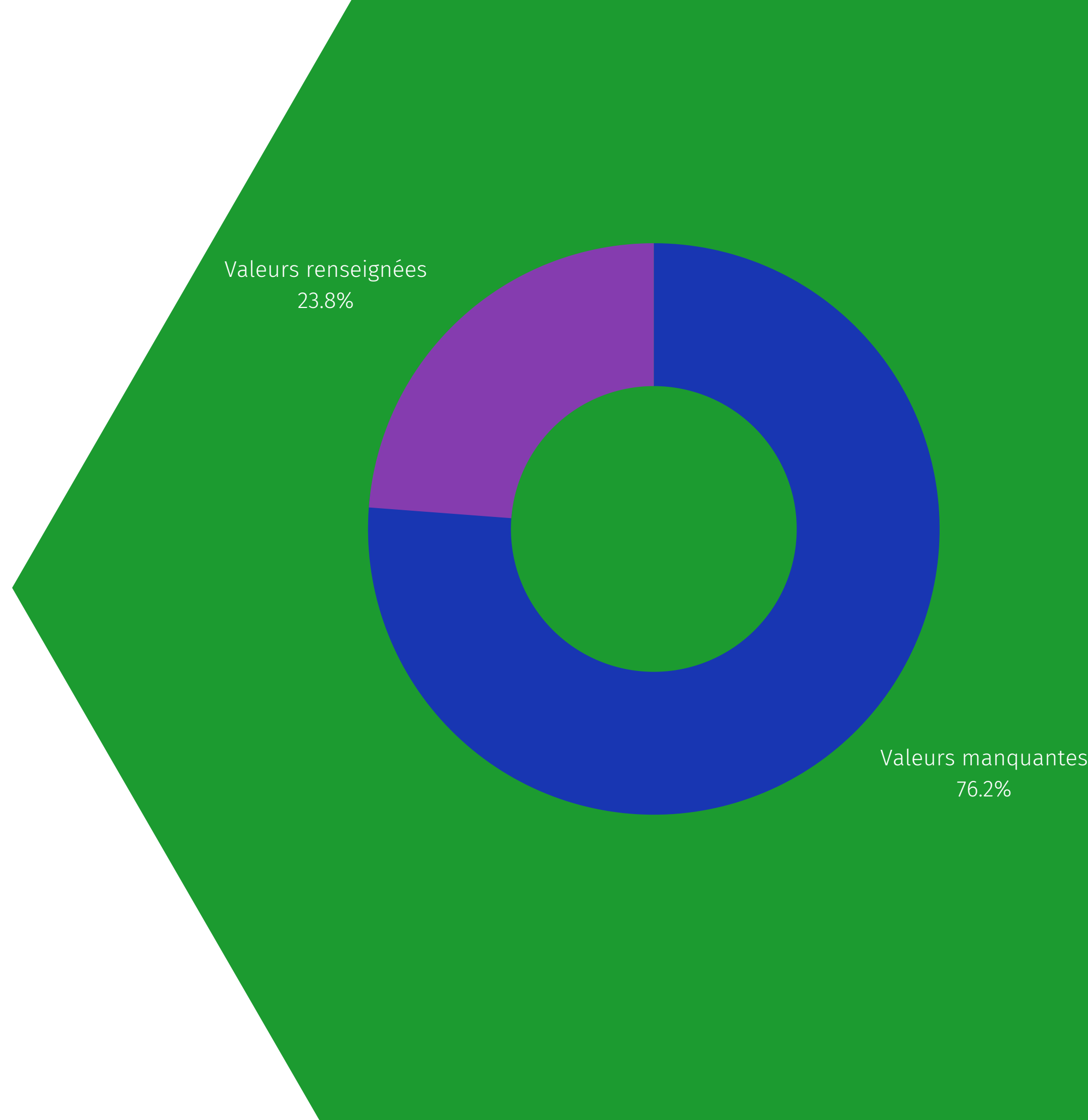
Normalisation de certaines variables

formatage de certaines variables

Taux de valeurs manquantes

Nous avons jusqu'a 76% de valeurs manquantes

Nous allons donc procéder au traitement par les méthodes comme la mise à zéro, la combinaison des colonnes, la suppression des variables à plus de 80% de NaN et l'imputation par la médiane



Tentative de croisement des colonnes

Nous avons parcouru 2 à 2 les 162 colonnes et de créer un binôme, voir si l'un n'est pas renseigné dans le binôme, l'autre l'est. Pour voir la possibilité de les croiser.

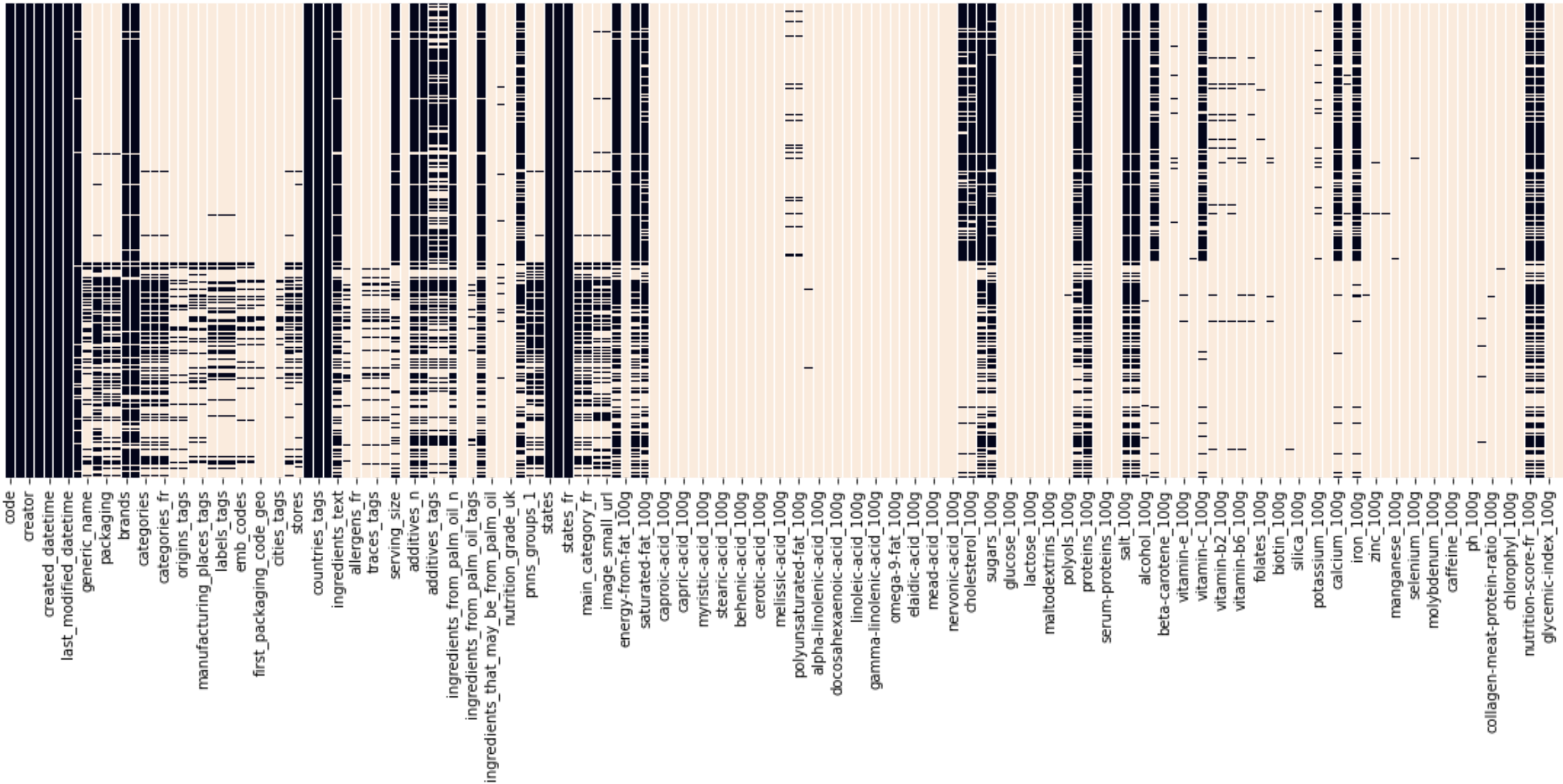
Cela réduira le nombre de valeurs manquantes mais aussi le nombre de colonnes

En observant les colonnes qui remplissent cette condition, il serait impossible de les croiser car laplupart ne sont pas de meme types ni renseignés dans le meme format. ex du code et du cities. On passe donc à la supression des variables ayant plus de 80% de valeurs manquantes

```
162
colonnes ('code', 'cities') à croiser
colonnes ('code', 'behenic-acid_100g') à croiser
colonnes ('code', 'dihomo-gamma-linolenic-acid_100g') à croiser
colonnes ('url', 'cities') à croiser
colonnes ('url', 'behenic-acid_100g') à croiser
colonnes ('url', 'dihomo-gamma-linolenic-acid_100g') à croiser
colonnes ('last_modified_t', 'no_nutriments') à croiser
colonnes ('last_modified_t', 'ingredients_from_palm_oil') à croiser
colonnes ('last_modified_t', 'ingredients_that_may_be_from_palm_oil') à croiser
colonnes ('last_modified_t', 'nutrition_grade_uk') à croiser
colonnes ('last_modified_t', 'butyric-acid_100g') à croiser
colonnes ('last_modified_t', 'caproic-acid_100g') à croiser
colonnes ('last_modified_t', 'lignoceric-acid_100g') à croiser
colonnes ('last_modified_t', 'cerotic-acid_100g') à croiser
colonnes ('last_modified_t', 'melissic-acid_100g') à croiser
colonnes ('last_modified_t', 'elaidic-acid_100g') à croiser
colonnes ('last_modified_t', 'mead-acid_100g') à croiser
colonnes ('last_modified_t', 'erucic-acid_100g') à croiser
colonnes ('last_modified_t', 'nervonic-acid_100g') à croiser
colonnes ('last_modified_t', 'chlorophyll_100g') à croiser
colonnes ('last_modified_t', 'glycemic-index_100g') à croiser
colonnes ('last_modified_t', 'water-hardness_100g') à croiser
colonnes ('last_modified_datetime', 'no_nutriments') à croiser
colonnes ('last_modified_datetime', 'ingredients_from_palm_oil') à croiser
colonnes ('last_modified_datetime', 'ingredients_that_may_be_from_palm_oil') à croiser
colonnes ('last_modified_datetime', 'nutrition_grade_uk') à croiser
colonnes ('last_modified_datetime', 'butyric-acid_100g') à croiser
```

Supression des Colonnes ayant jusqu'à plus de 80% de valeurs manquantes

Pres de 73 colonnes sur les 162 ont plus de 80% de valeurs manquantes. On va donc les supprimer. Mais nous allons converver certaines variables utiles à la classification comme le potassium, l'alcool et autres. Nous n'allons supprimer que 70

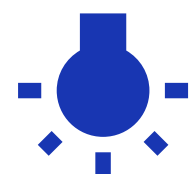


Mise à Zero

Les colonnes ayant _100g dans le nom de leur variables ont été rempli par NaN à la place de 0 car probablement, si l'ingrédient n'y est pas, on laisse la case vide, ce qui se transforme en NaN. On va donc remettre à 0 ces valeurs manquantes

Nous observons qu'apes cette opération, le pourcentage des NaN est passé à 38%

Supression des colonnes inutiles à la classification



Certaines variables comme le code, l'url les autres restent inutiles à la classification, On les supprime.



Au total 19 colonnes inutiles

Notre base de données a une taille de 320772 lignes et 26 colonnes

product_name	origins	origins_tags	labels	emb_codes	emb_code
Farine de blé noir	NaN	NaN	NaN	NaN	
Banana Chips Sweetened (Whole)	NaN	NaN	NaN	NaN	
Peanuts	NaN	NaN	NaN	NaN	
Organicalted Nut Mix	NaN	NaN	NaN	NaN	
Organic Polenta	NaN	NaN	NaN	NaN	

Imputation des autres valeurs manquantes par la mediane

Nous n'avons plus de valeurs
manquantes dans notre jeu de
données

Nous remplaçons les valeurs
manquantes par la médiane

a-3- 100g	cholesterol_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	alcohol_100g	vitamin- a_100g	vitamin- c_100g	potassium_100g	calcium
0.0	0.000	0.00	0.00	0.0	0.00	0.00000	0.000	0.0	0.0	0.0000	0.0	
0.0	0.018	64.29	14.29	3.6	3.57	0.00000	0.000	0.0	0.0	0.0214	0.0	
0.0	0.000	60.71	17.86	7.1	17.86	0.63500	0.250	0.0	0.0	0.0000	0.0	
0.0	0.000	17.86	3.57	7.1	17.86	1.22428	0.482	0.0	0.0	0.0000	0.0	
0.0	0.000	77.14	0.00	5.7	8.57	0.00000	0.000	0.0	0.0	0.0000	0.0	

Traitement des Doublons

Nous allons supprimer les produits qui ont le meme nom

Cette methode a fait remplacer toutes les valeurs manquantes par la médiane pour les valeurs qualitatives et le moyenne pour les variables quantitatives


**Après la suppression des doublons, nous obtenons
184564 lignes**

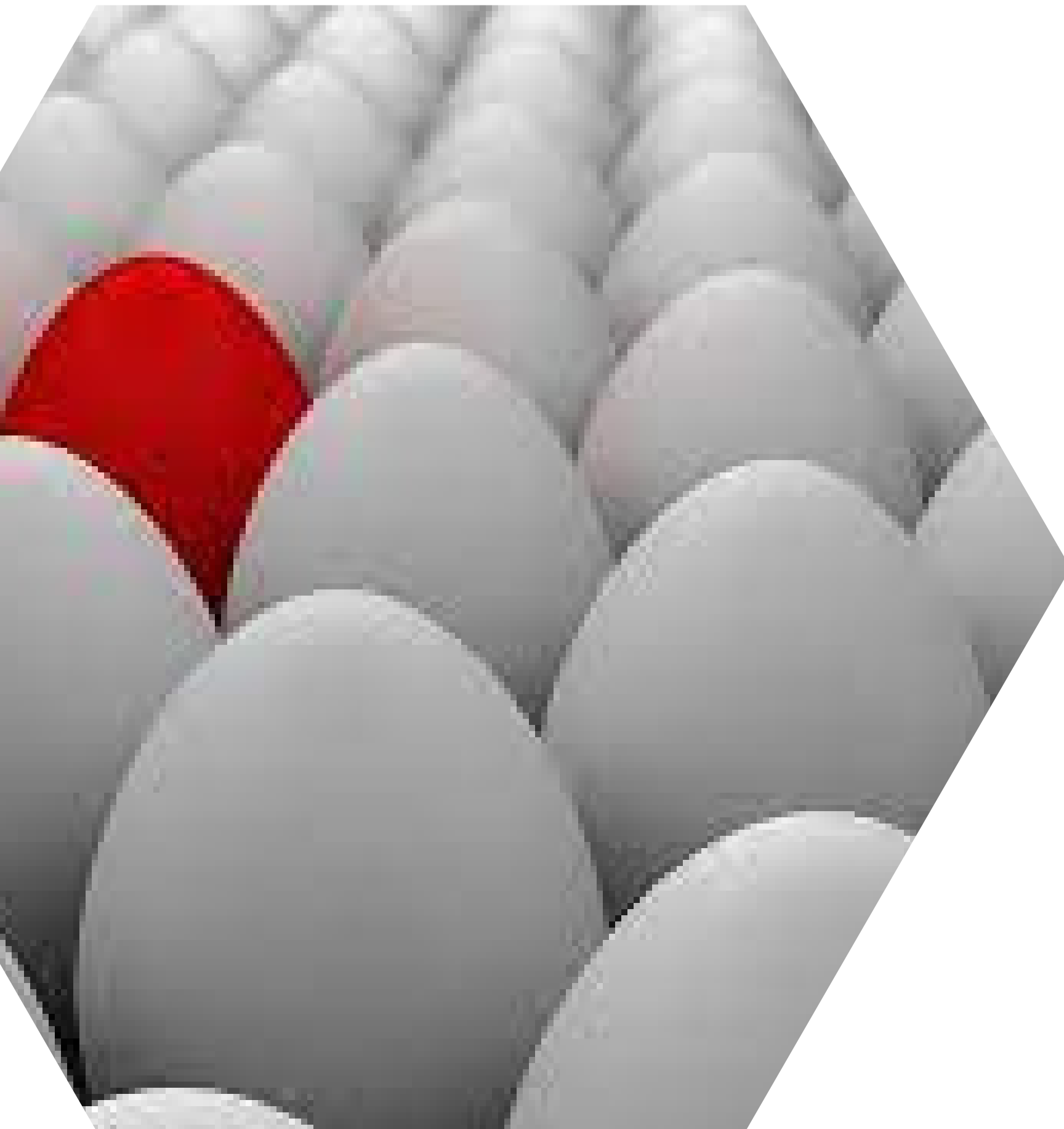




Réécriture des certaines variables

Certaines variables sont comme sous forme de liste. Nous allons sélectionner seulement les premiers éléments de chaque liste.





Valeurs aberrantes

Un monde où n'importe qui peut faire des diaporamas impressionnants

Nous allons identifier dans un
premier temps
les valeurs aberrantes et après les remplacer

Observations et traitement



Traitement de Energy_100g

la variable energy_100g a des valeurs concentrées entre 0 et 0.2×10^6 , elle est censé être inférieur à 100 g. Les valeurs ont été prises probablement dans la mauvaise unité c'est à dire en kilocalorie.

Nous allons donc les ramener en grammes à n'importe quel objectif ou sujet.



Construction de boxplot pour visualiser les outliers

Certains ont des valeurs dépassant les 100g ou inférieures à 0: energy_from_fat, lauric-acid_100g, arachidic acid,

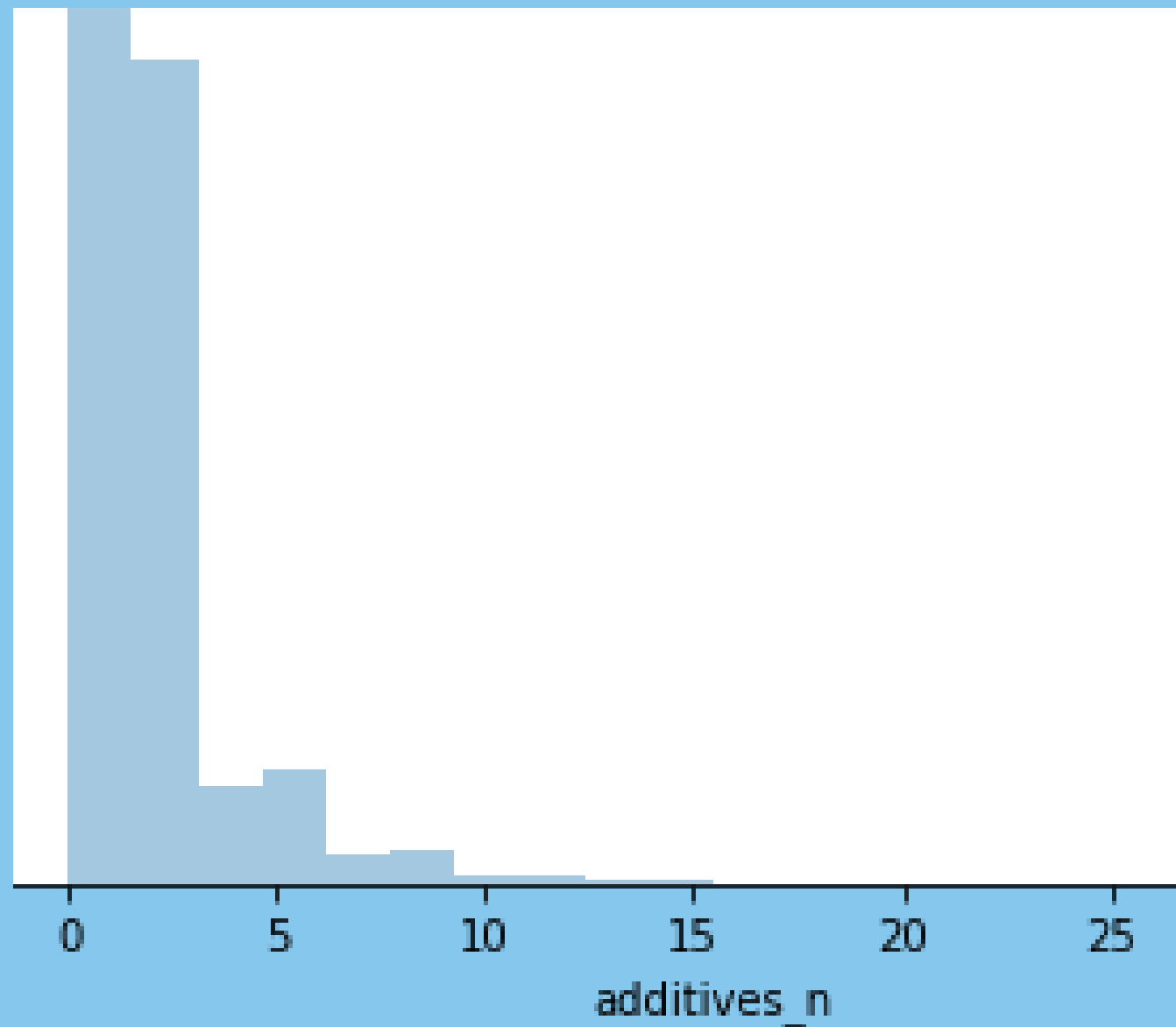
Remplacement de ces valeurs par la valeur médiane des valeurs inférieurs à 100



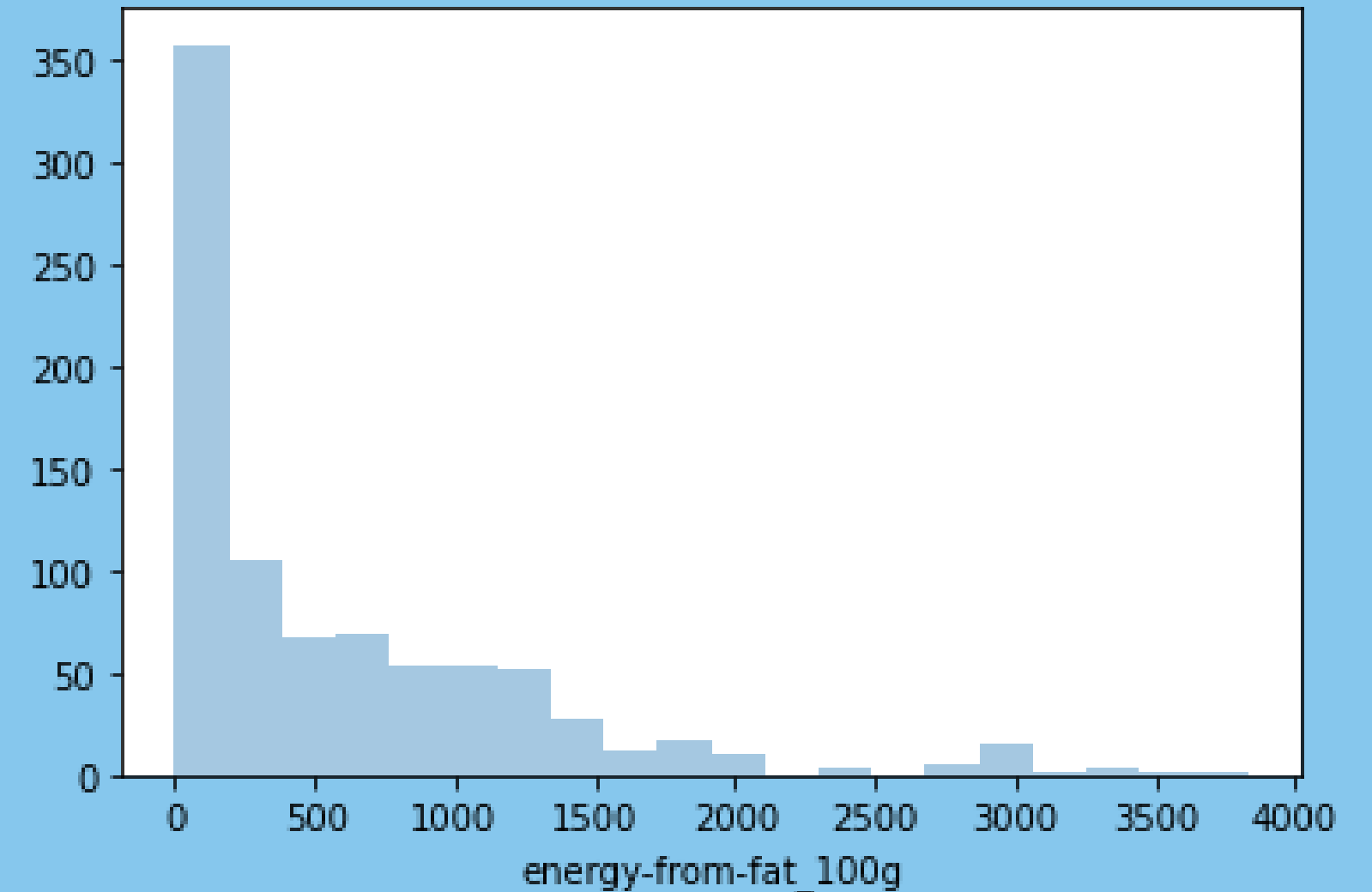
Remplacement des outliers par la médiane

Présentez facilement et épatez tout public avec les diaporamas Canva. Choisissez parmi plus d'un millier de modèles conçus par des professionnels pour s'adapter à n'importe quel objectif ou sujet.

Observons certaines variables

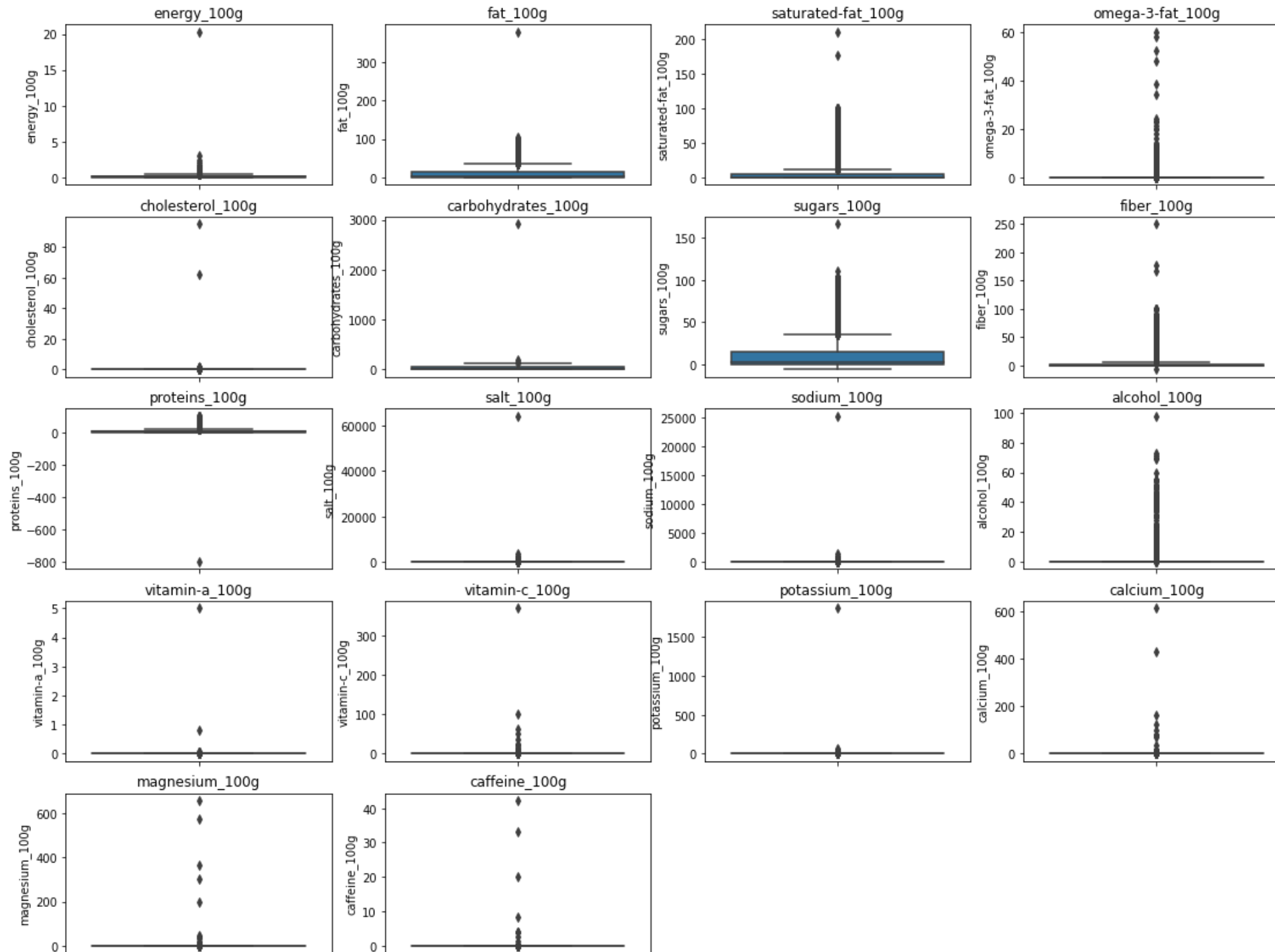


Certaines sont bien renseignées



D'autres dépassent la valeur 100, ce qui est anormale

Boxplot pour voir les valeurs aberrantes



Nous observons
plein
de valeurs
aberrantes
et des valeurs
atypiques
tels des produit qui
ont
100% de sel

**Selction des lignes à
qui ont à la fois le
product_name et le
nutriscore**

**Remplacement des
variables à 100g
dépassant 100g par la
mediane**

**Supprimer les produits
qui ont 100% d'un seul
ingrédient**

Commencez par vous inspirer de milliers
de modèles, collaborez facilement et
engagez votre audience avec une
présentation Canva mémorable.



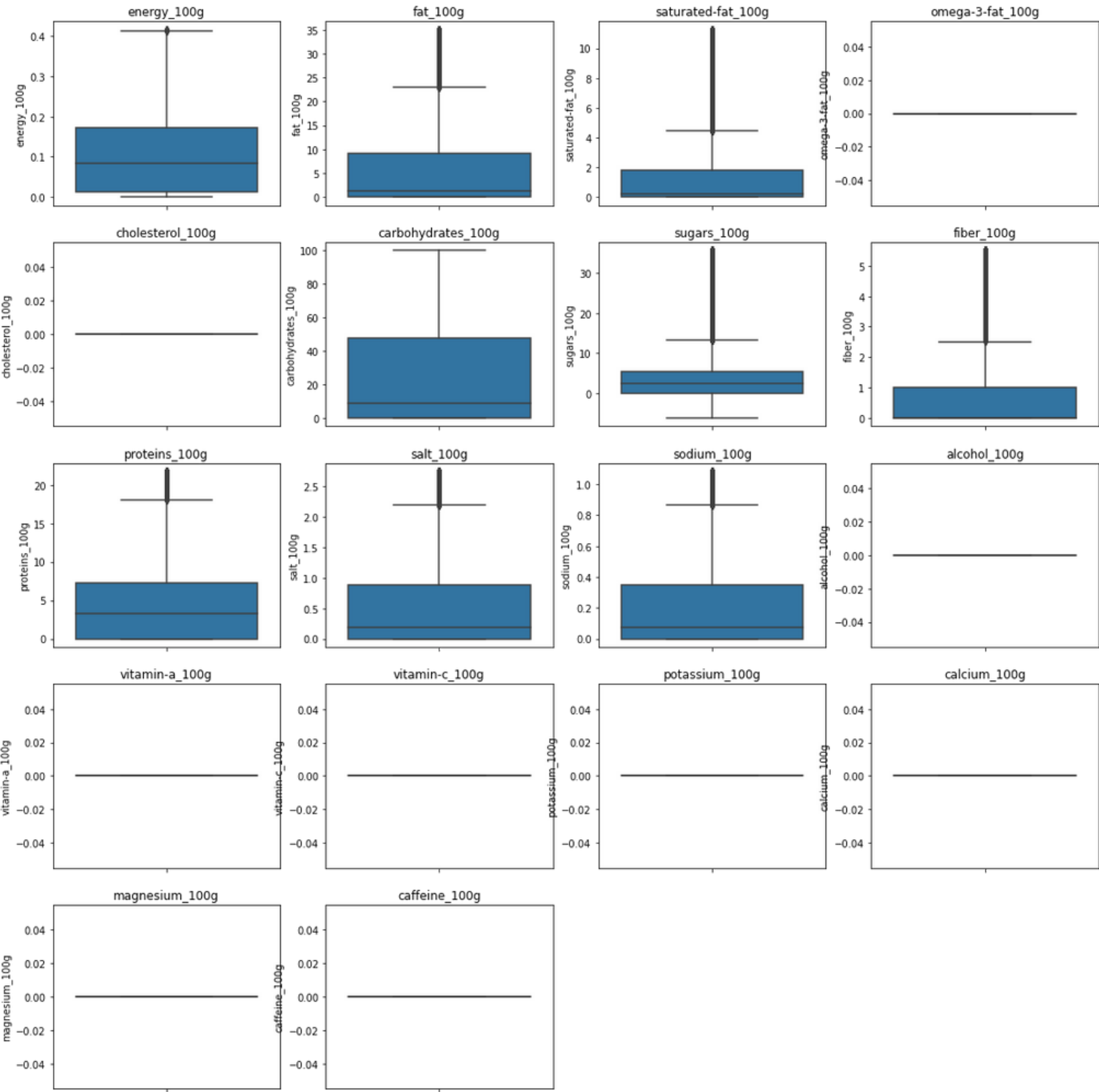
**Pret pour l'analyse
exploratoire de données**



**Notre dataset contient
maintenant 184564 lignes et
26 colonnes**



Apes traitement nous obtenons
l'histogramme suivant où toutes les
valeurs sont nomalisées

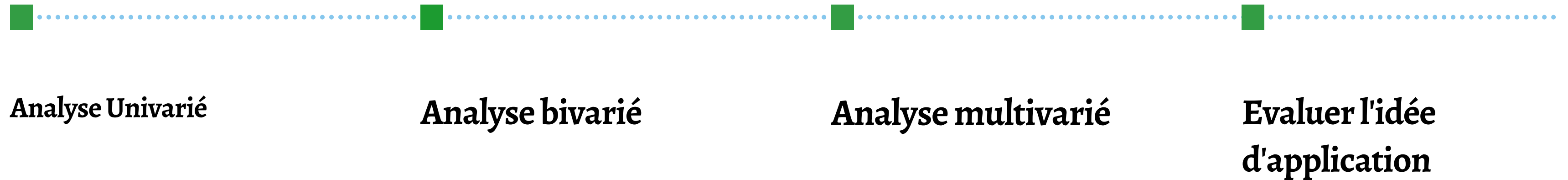




ANALYSE EXPLORATOIRE



Chronologie de l'analyse exploratoire

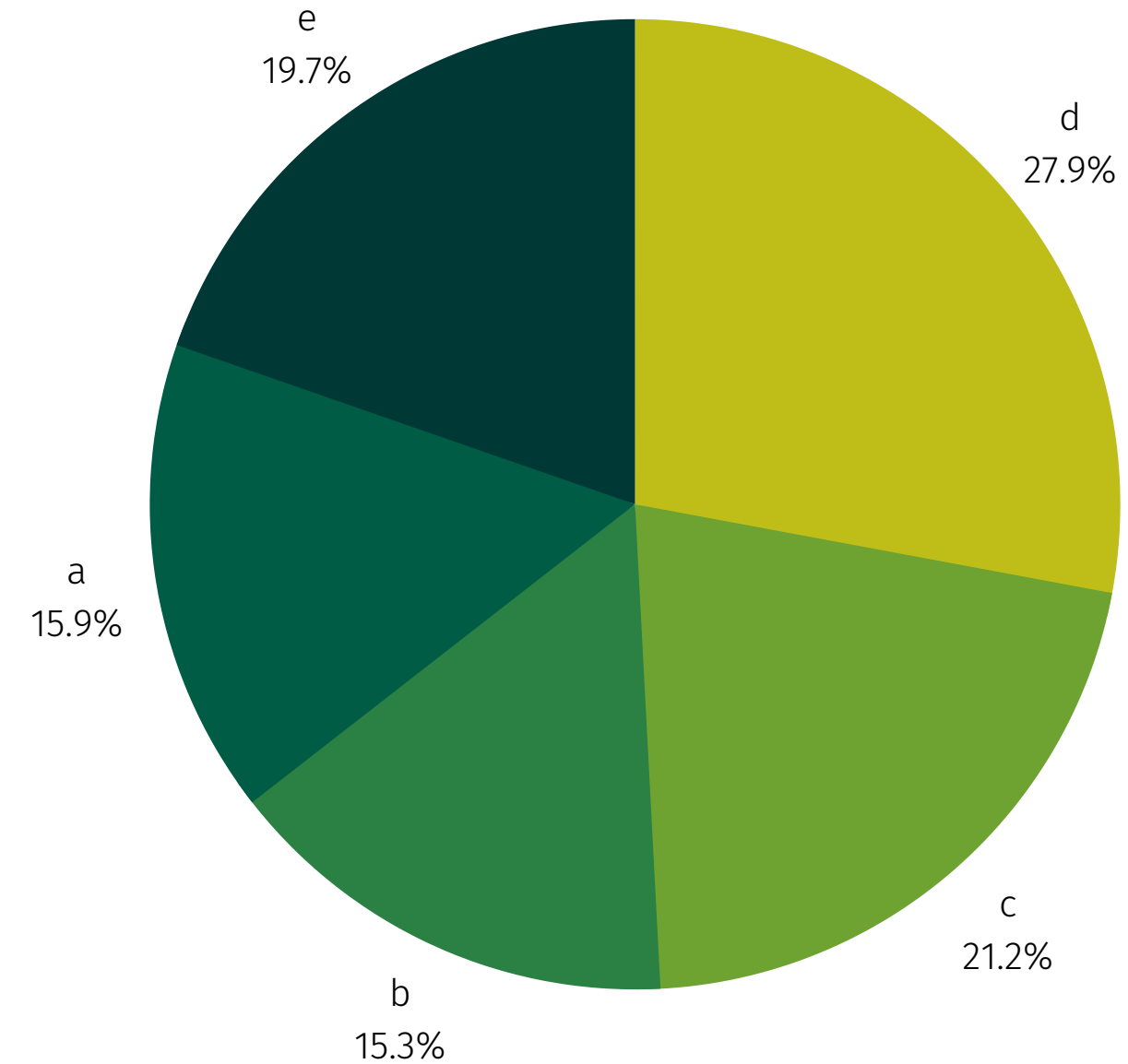


Analyse Univariée

Visualisation du target

Nous constatons que les individus ayant le nutriscore d sont majoritaire. La base de données est donc rempli de données

Seulement pres de 16% des produits sont de score a. La base de données est presque équitablement répartie, ce qui ne vas avantager le modèle à etre plus précis.

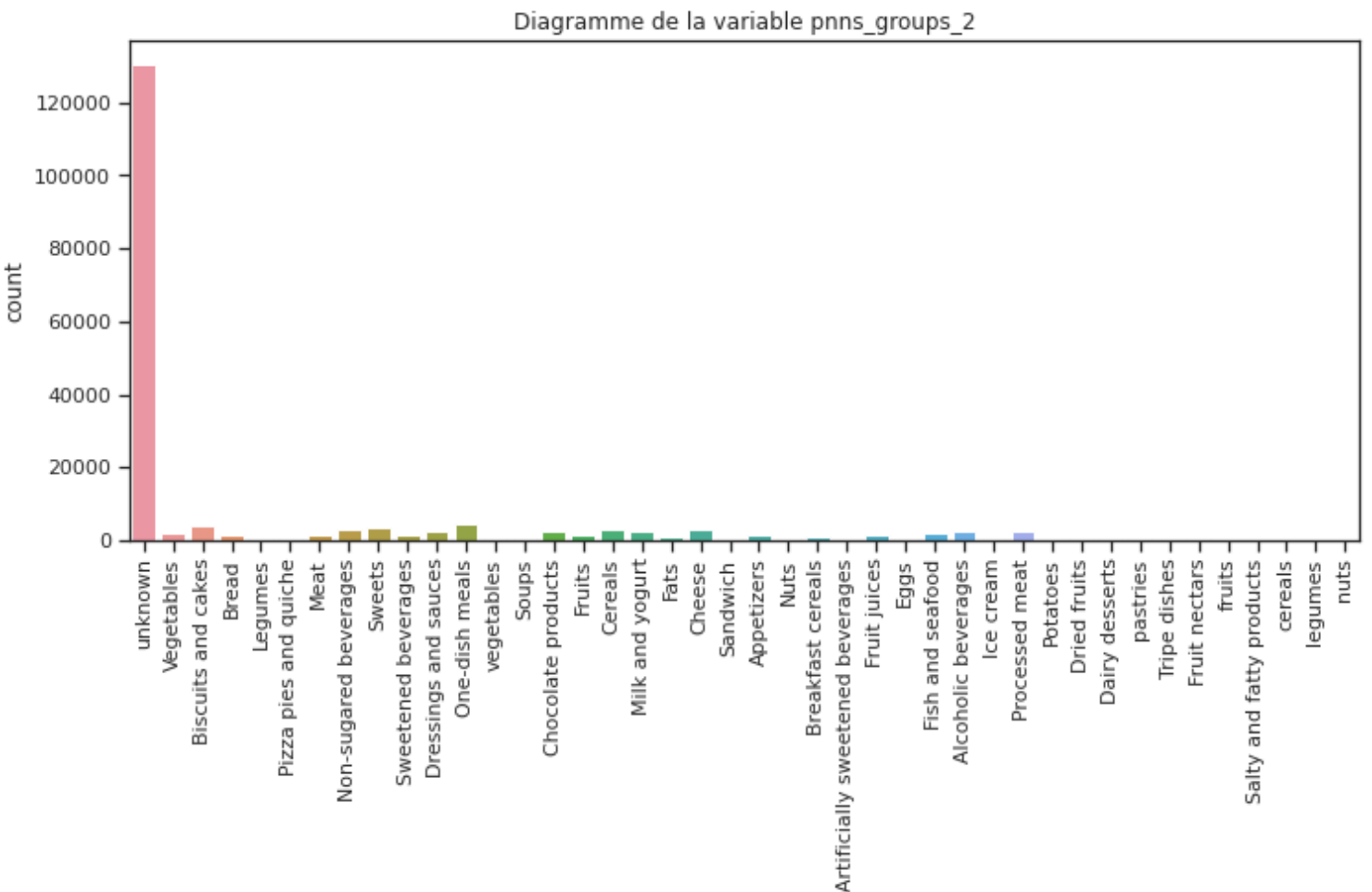
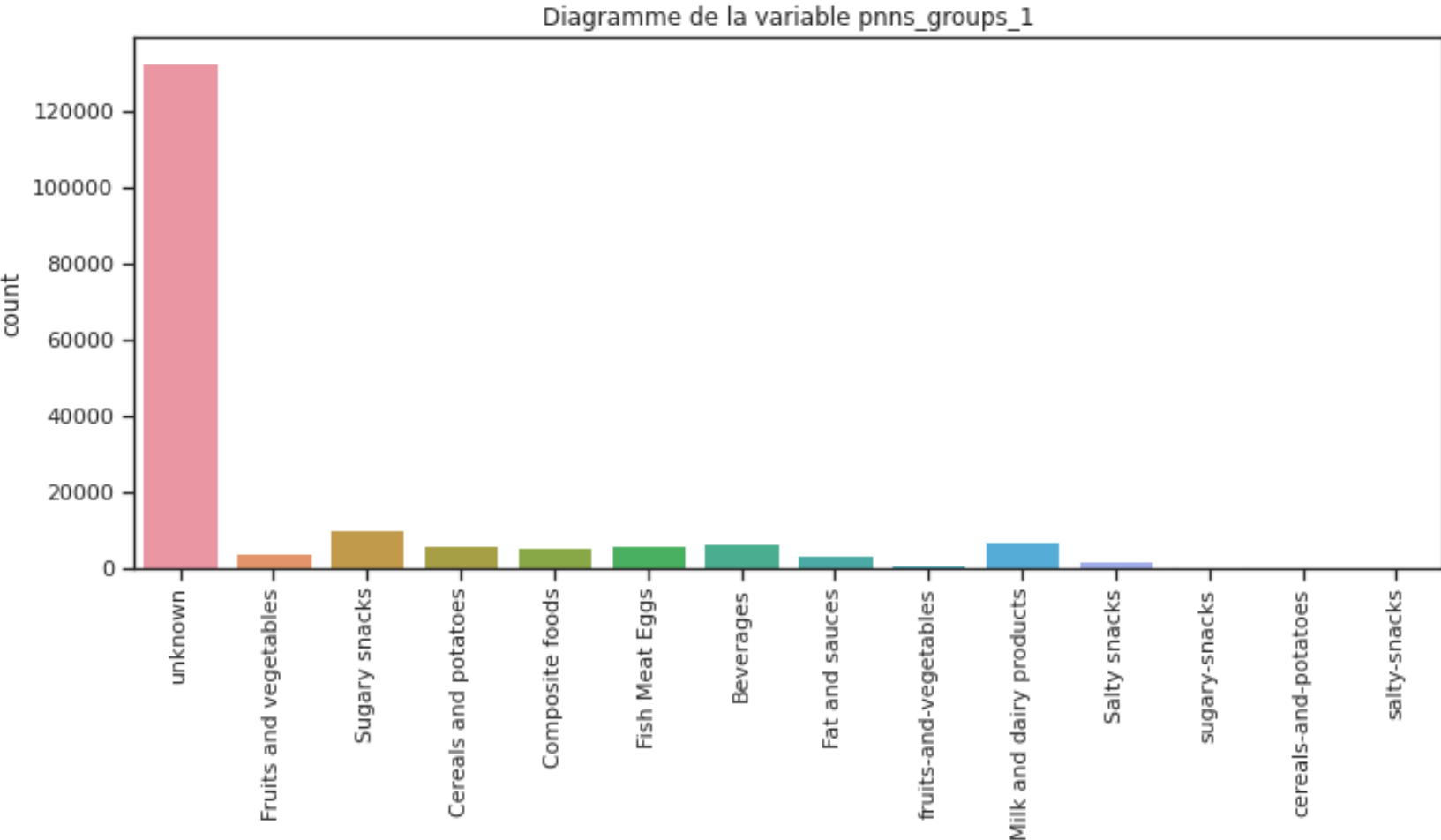


Variables qualitatives

Représentation des variables avec countplot

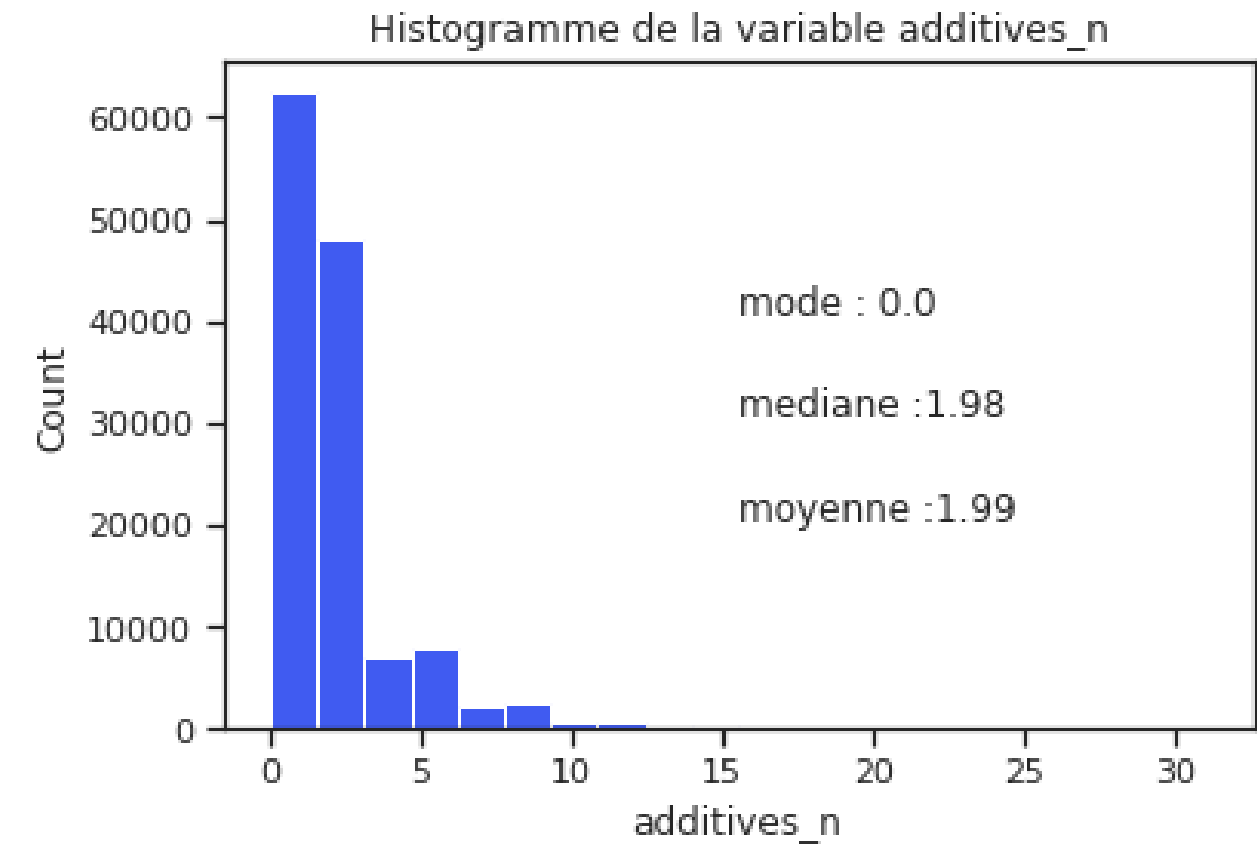
Nous représentons ici les variables qualitatives ayant moins de 100 valeurs différentes pour éviter le plantage du notebook

Le groupe nommé "almonds" prime sur l'ensemble des modalités.

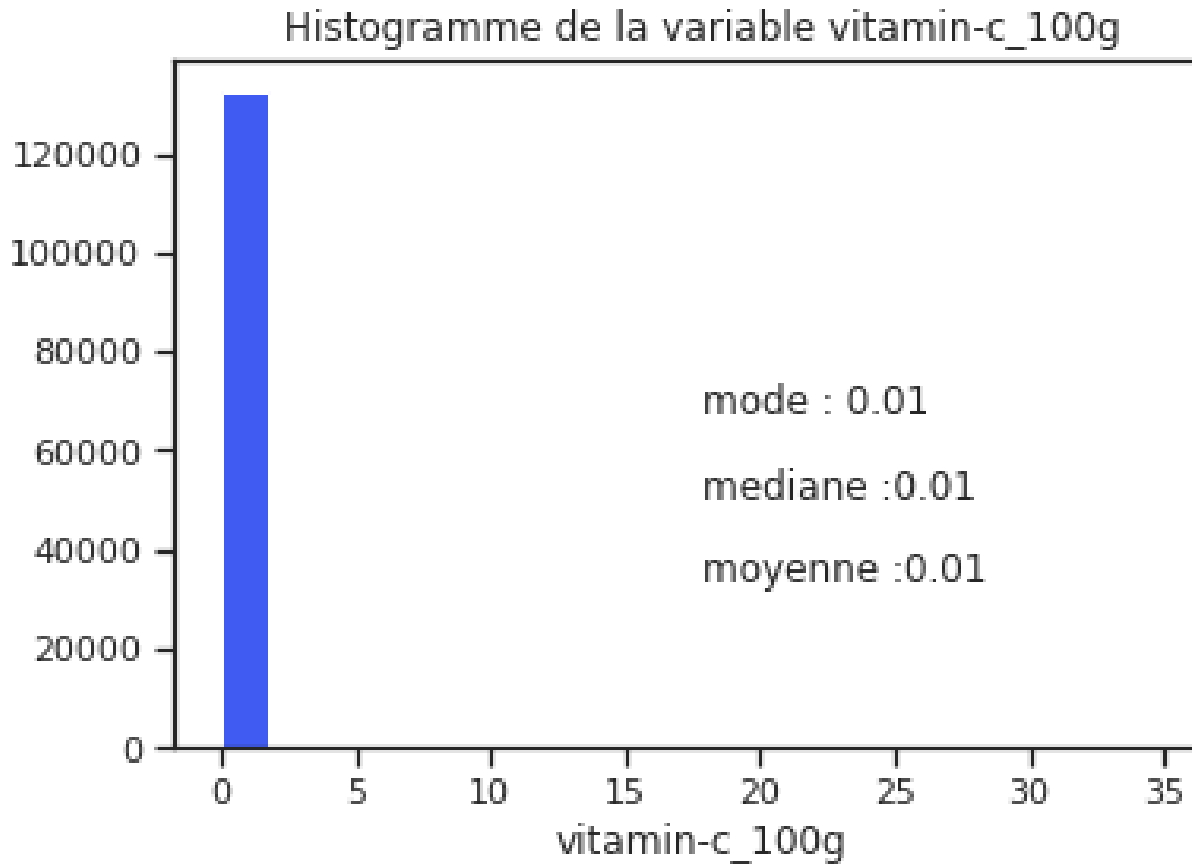
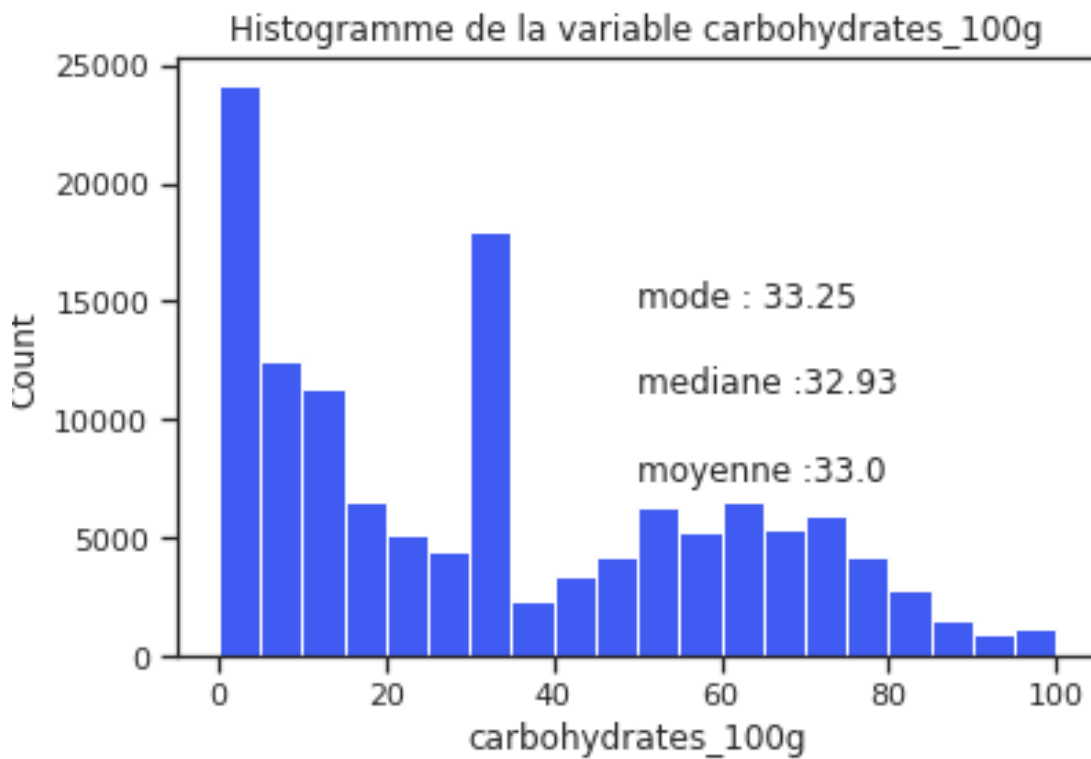


Variables quantitatives

Nous représentons les variables par des histogrammes et nous affichons aussi la moyenne, le mode et la médiane



Distribution étalée vers la gauche



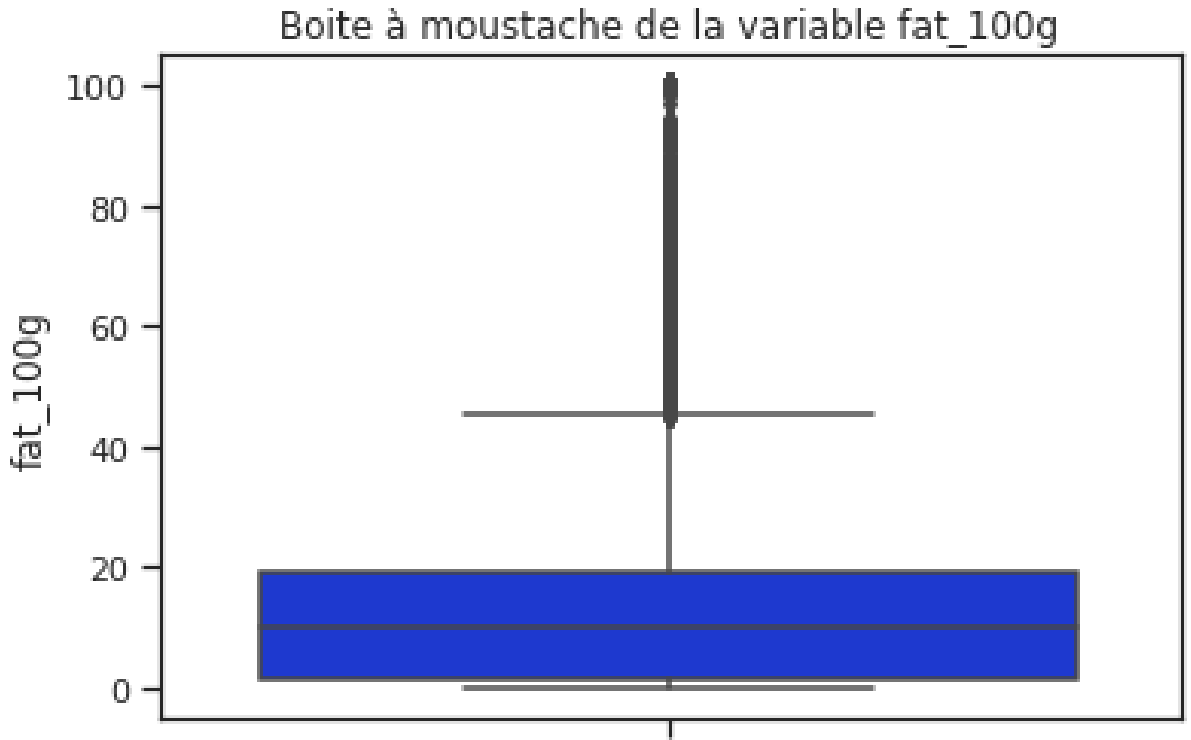
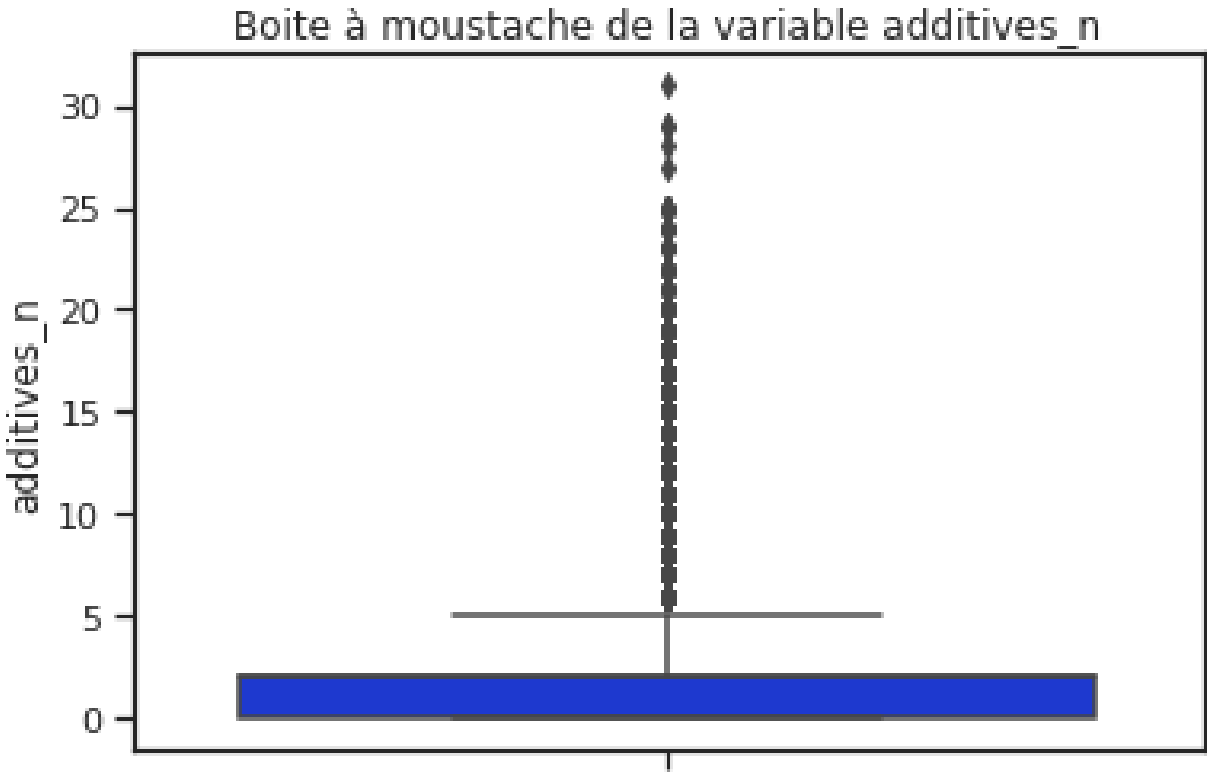
Certains concentrés sur zeros

Distribution double: une étalée à gauche et l'autre normalisée

Analyse descriptive des variables

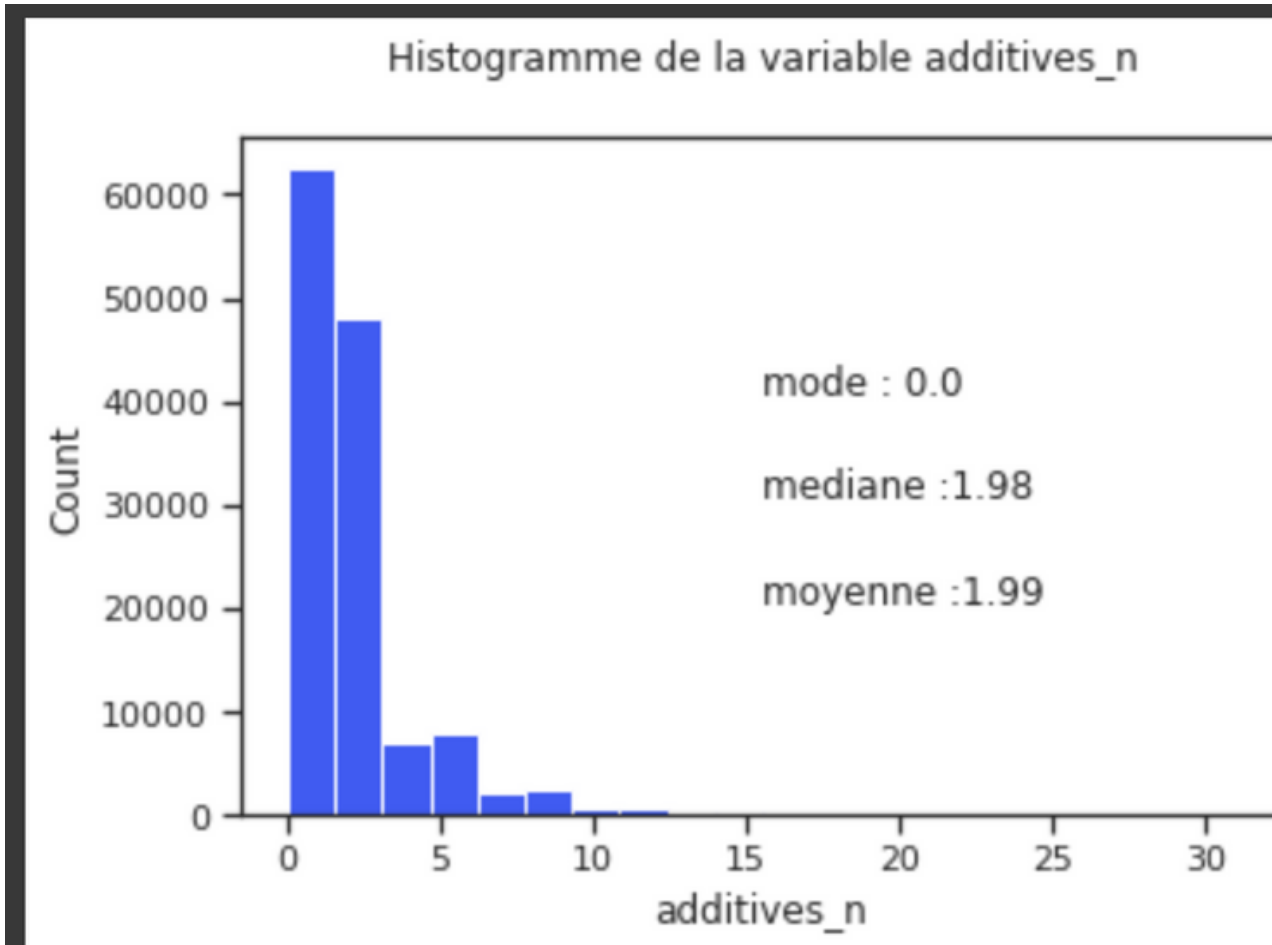
Une partie de cette étape a été faite. En faisant les boîtes à moustaches, nous allons bien les visualiser. Le tableau de description nous donne plus d'information sur la moyenne, le mode, le min, le max et autres de chaque variable

additives_n	ingredients_from_palm_oil_n	ingredients_that_may_be_from_palm_oil_n	energy_100g	fat_100g	saturated-fat_100g	omega-3-fat_100g
9.000000	132359.000000	132359.000000	132359.000000	132359.000000	132359.000000	132359.000000
1.992605	0.029386	0.069619	0.036086	13.170079	5.087569	3.189578
2.370846	0.161577	0.289061	0.028030	14.793543	7.753925	0.377686
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.014671	1.600000	0.000000	3.190612
1.982698	0.000000	0.000000	0.035893	10.000000	1.800000	3.190612
2.000000	0.000000	0.000000	0.052224	19.230000	7.140000	3.190612
1.000000	2.000000	6.000000	5.621133	100.000000	100.000000	60.000000
mns						



Mesure de forme

- Les histogrammes montrent l'asymetrie et l'applatissage des variables
- Pour chaque variable nous estimons si la distribution est asymetrique ou non et si elle est aplatie au sommet ou pointue
- Nous affichons egalement en description le coefficient d'asymetrie et d'applatissage



Coefficient d'asymetrie de additives_n: 2.4697653586605908
Distribution étalée vers la droite

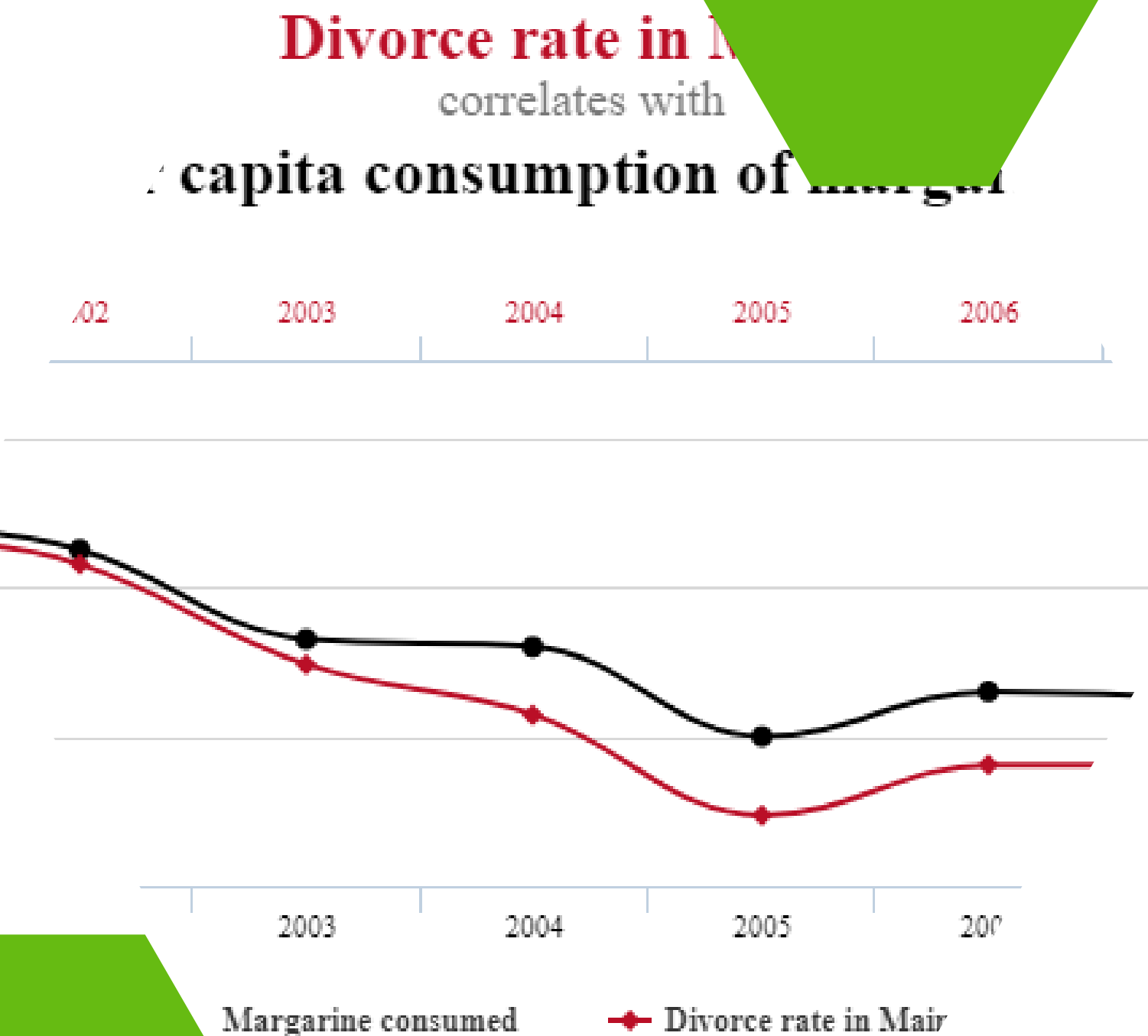
Coefficient d'applatissage de additives_n: 9.952027826598785
Distribution au sommet pointu

Analyse bivariee

Etude bivariee entre les variables qualitatives, quantitatives et entre les variables qualitatives et quantitatives.

Le jeu de donnees contient plusieurs types de variables Nous allons essayer de savoir:

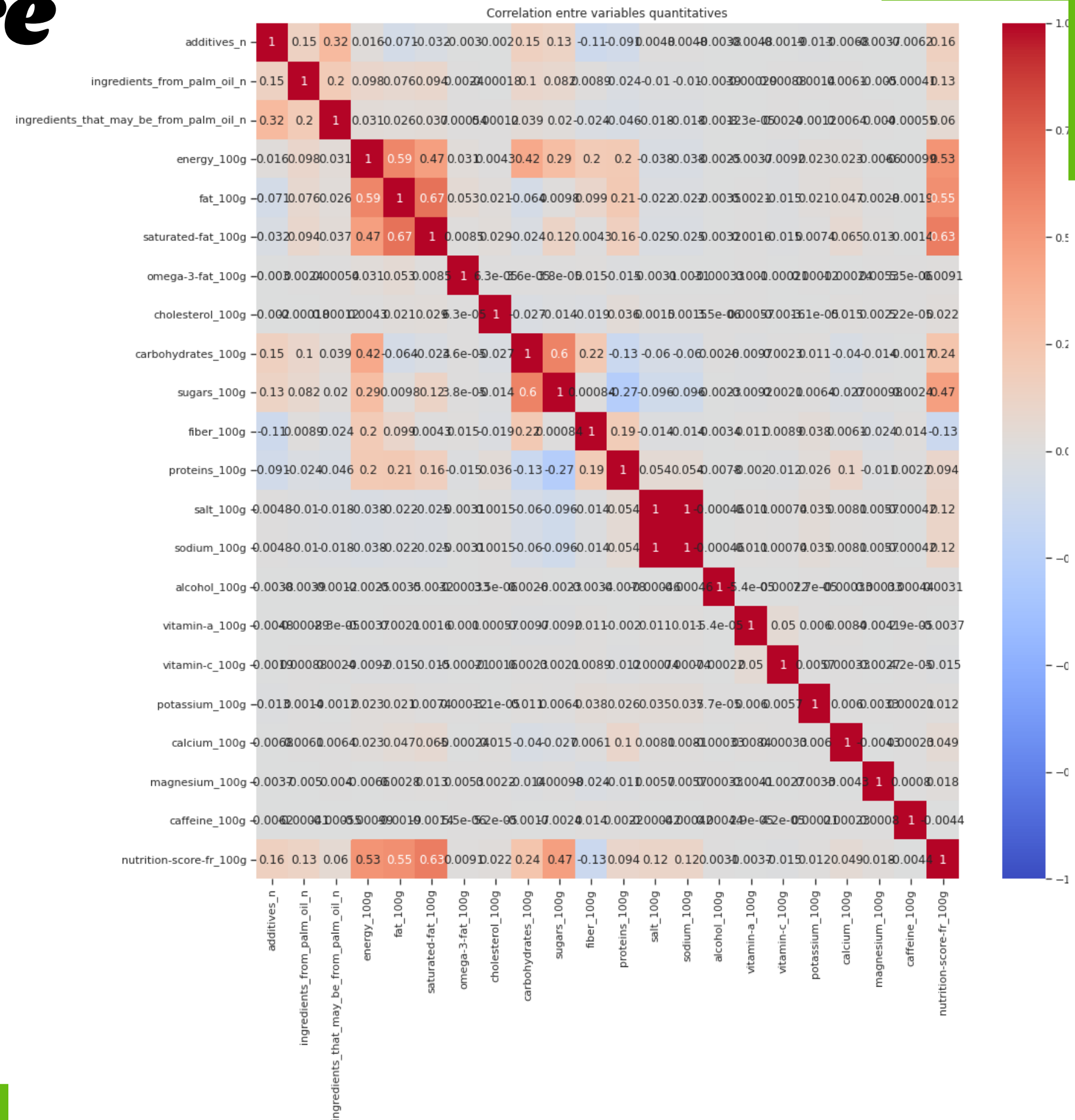
- Quelle est la relation entre le nutriscore pour 100g et le grade nutriscore? (ANOVA)
- si certaines variables sont independantes ou si elles sont deduit a partir d'autres existants
- quels sont les variables qui influent le plus sur le nutriscore des produit?
- Comment les variables d'energie sont correlés entre eux et comment ils influent sur le nutriscore



Correlation entre les variables

Nous remarquons :

- une forte corrélation(1) entre le sodium et le salt
- une corrélation moyenne entre le fat et le saturated_fat: 0.67
- une corrélation moyenne entre le sugars et le carbohydate: 0.6
- une corrélation négative entre le sugars et les protéines: -0.27
- Que les variables sugars, saturated_fat, fat et energy influent de manière significative le nutriscore_fr_100g

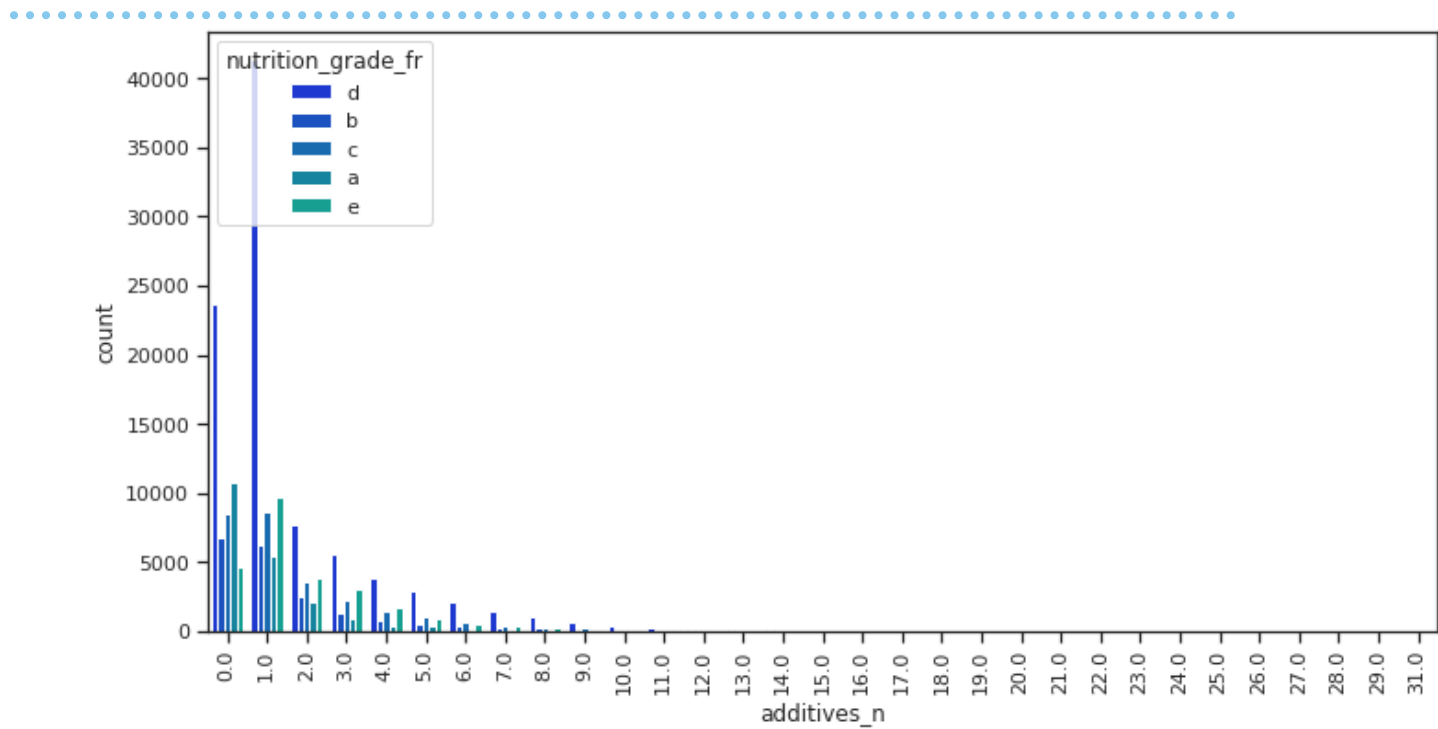


Relation entre les variables et le target

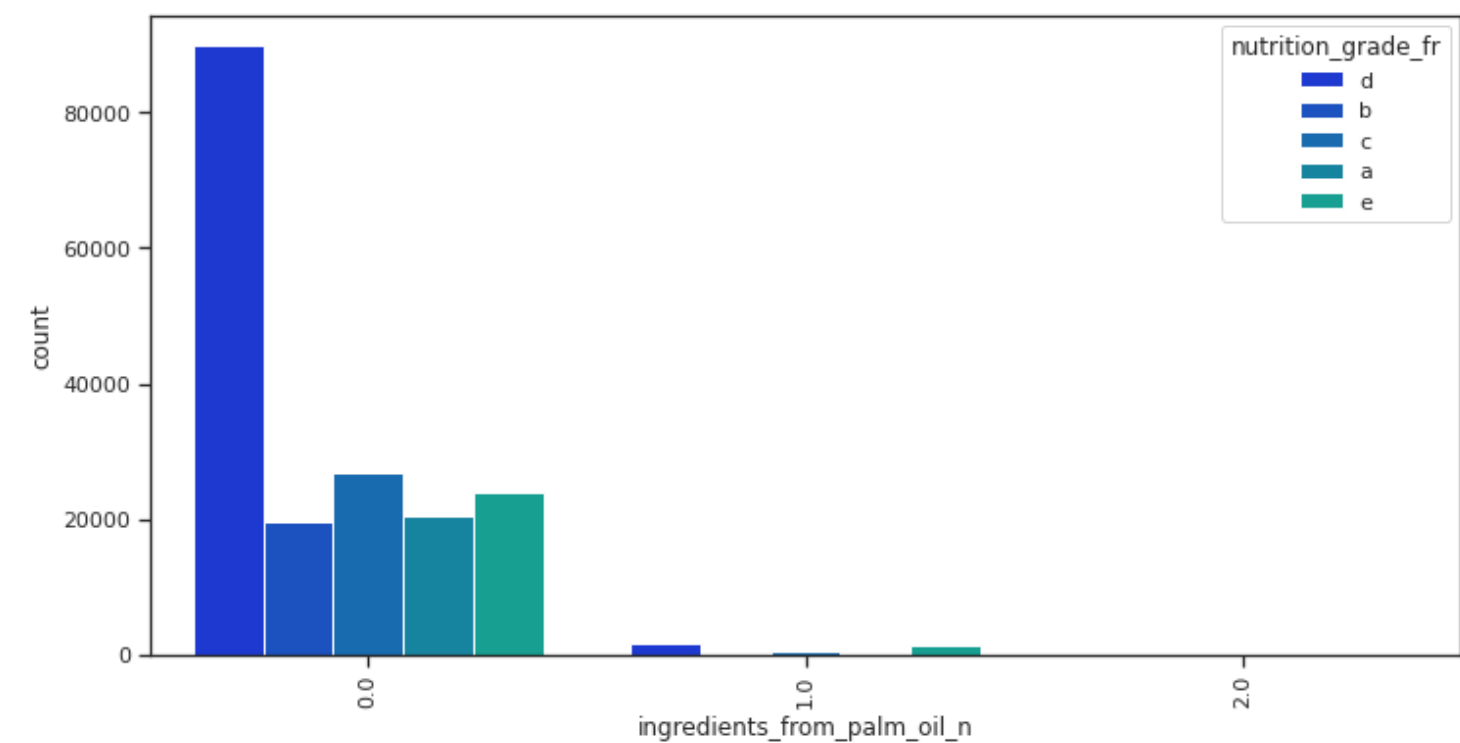
Entre le target et les variables
quantitatives

Entre le target et les variables
qualitatives

Entre le target et les variables quantitatives



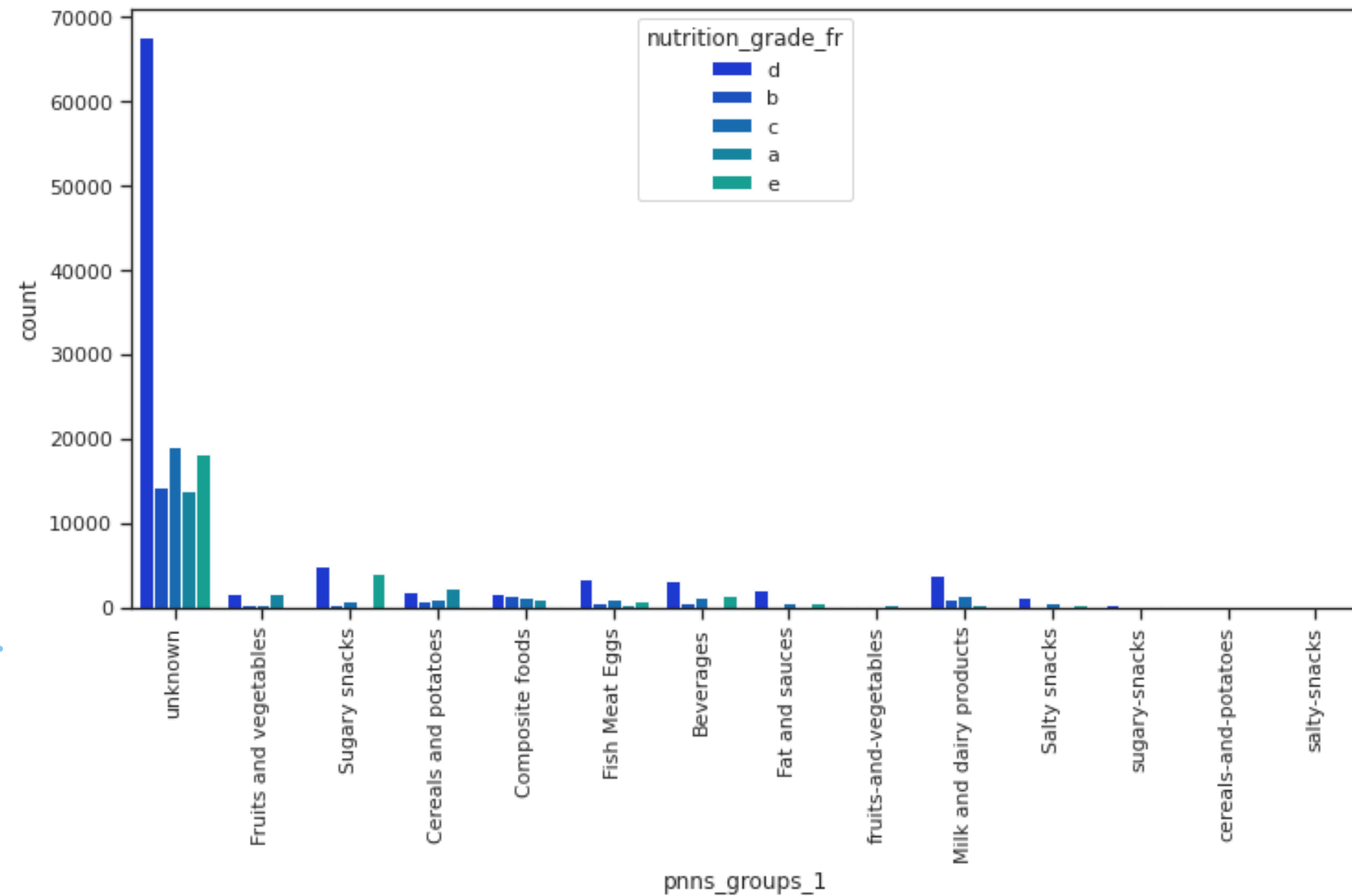
Nous visualisons ici la relation entre le target et ces variables



Entre le target et les variables Qualitatives

Nous allons visualiser comment les rapport entre les variables qualitatives et le target

Ici, les ba sont coloriées en fonction du target



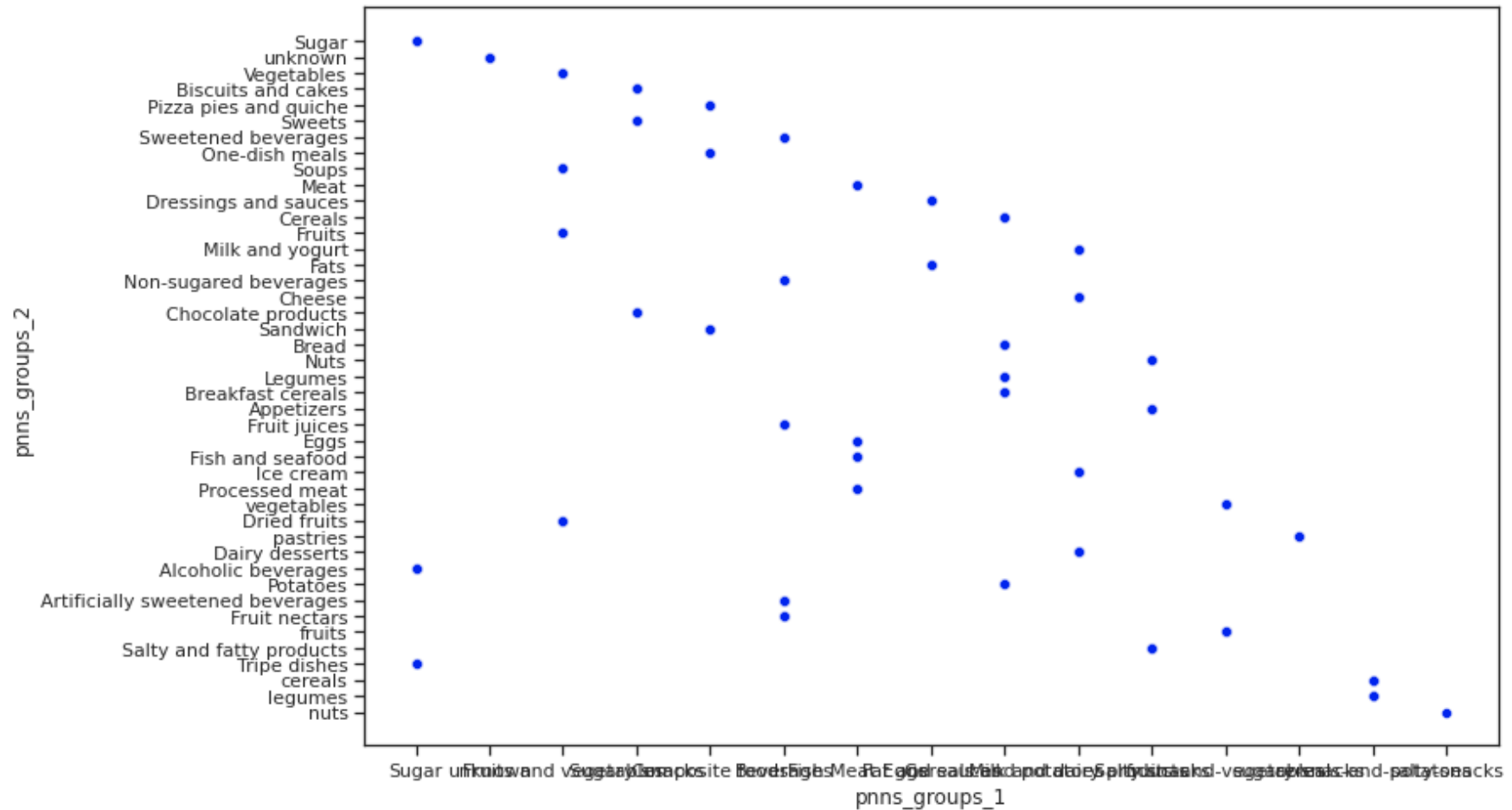
Relation entre les variables qualitatives: test chi2

**Regroupement des variables 2
a 2**

Formulation des hypotheses Formulons les
hypotheses: Ho(Hypothèse nulle): les deux variables
sont indépendantes
Ha(Hypothèse alternative): Les 2 variables sont
correlés

**Tableau de contingence et
deduction d'indépendance par
chi2**

Nuage de points entre les pnns_group et test de chi2



Par tableau de contingence et test de chi2, nous obtenons P_value ('nutrition_grade_fr', 'pnns_groups_1') : 0.0 Les variables ('nutrition_grade_fr', 'pnns_groups_1') sont corrélées, H1 validée P_value ('pnns_groups_1', 'pnns_groups_2') : 0.0 Les variables ('pnns_groups_1', 'pnns_groups_2') sont corrélées, H1 validée P_value ('pnns_groups_2', 'nutrition_grade_fr') : 0.0 Les variables ('pnns_groups_2', 'nutrition_grade_fr') sont corrélées, H1 validée

Relation entre les variables quantitatives: Regression linéaire

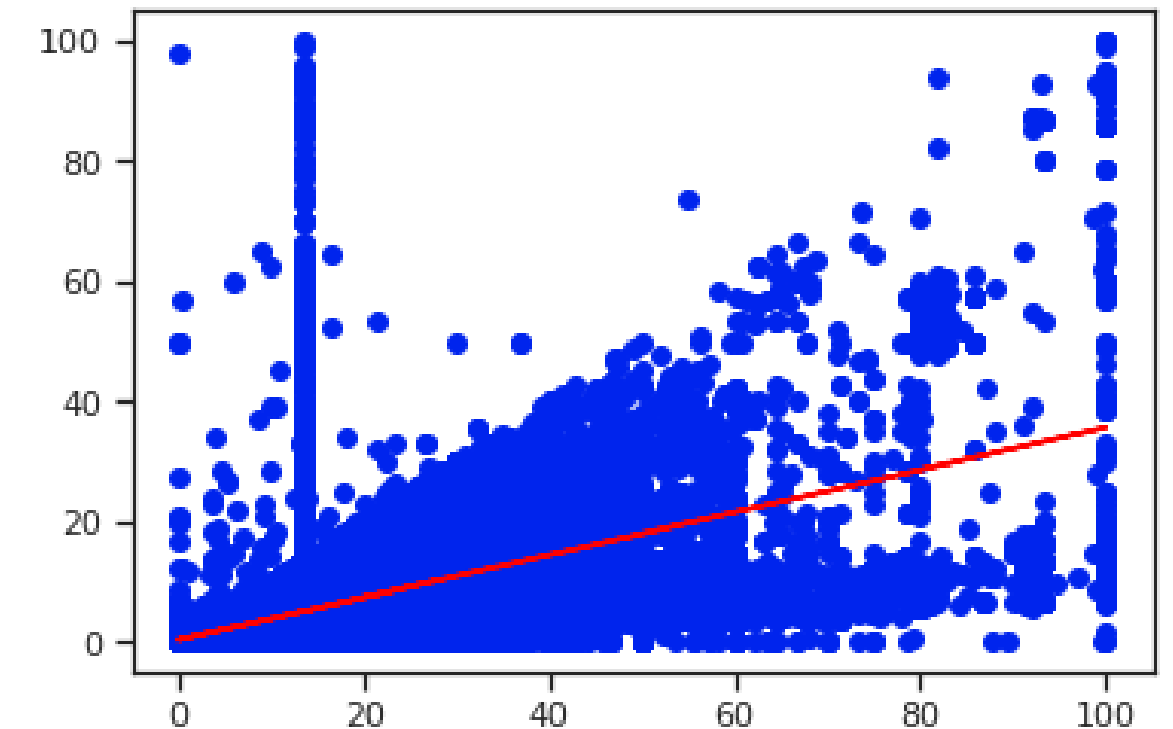
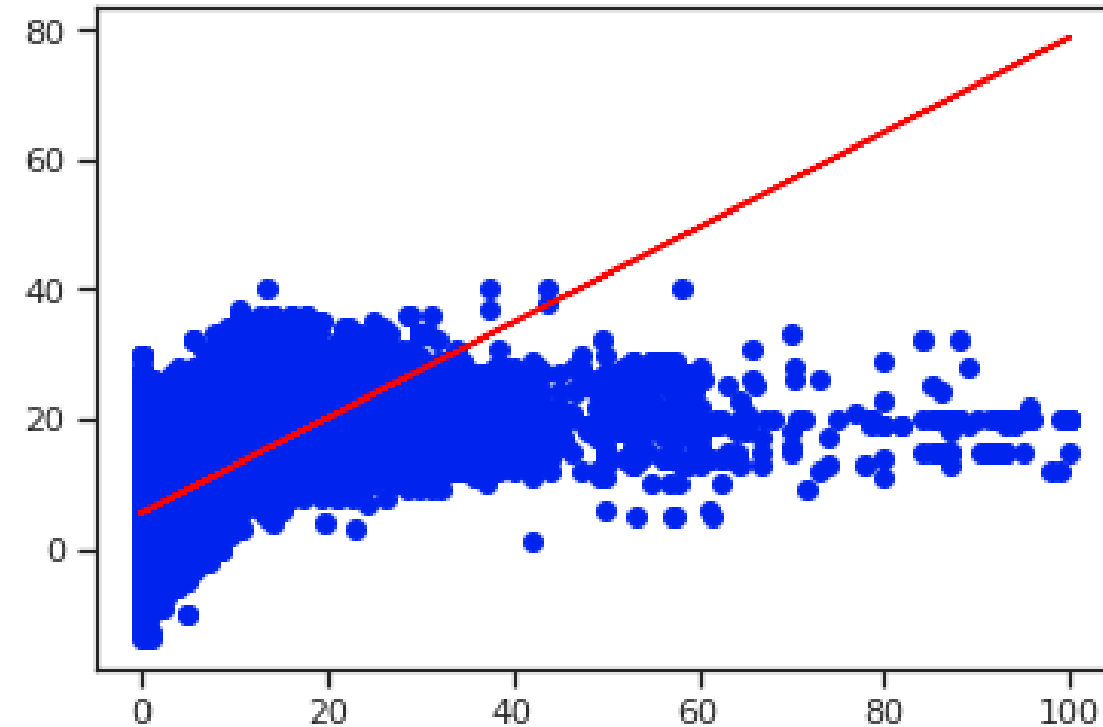
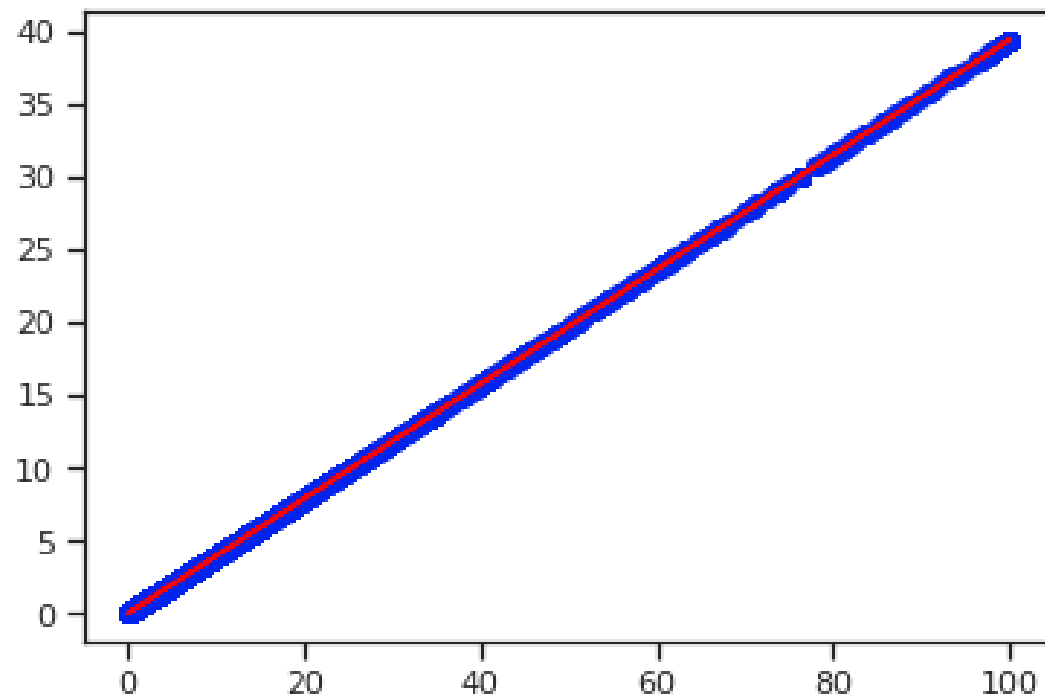
Nous etudions les colonnes dont la correlation est superieure a 0,6

```
[('fat_100g', 'saturated-fat_100g'),  
( 'saturated-fat_100g', 'nutrition-  
score-fr_100g'),  
( 'carbohydrates_100g',  
'sugars_100g'), ('salt_100g',  
'sodium_100g')]
```

Nuages de points et
analyse par regression
linéaire

Fusion ou suppression des
colonnes

Analyse par regression lineaire



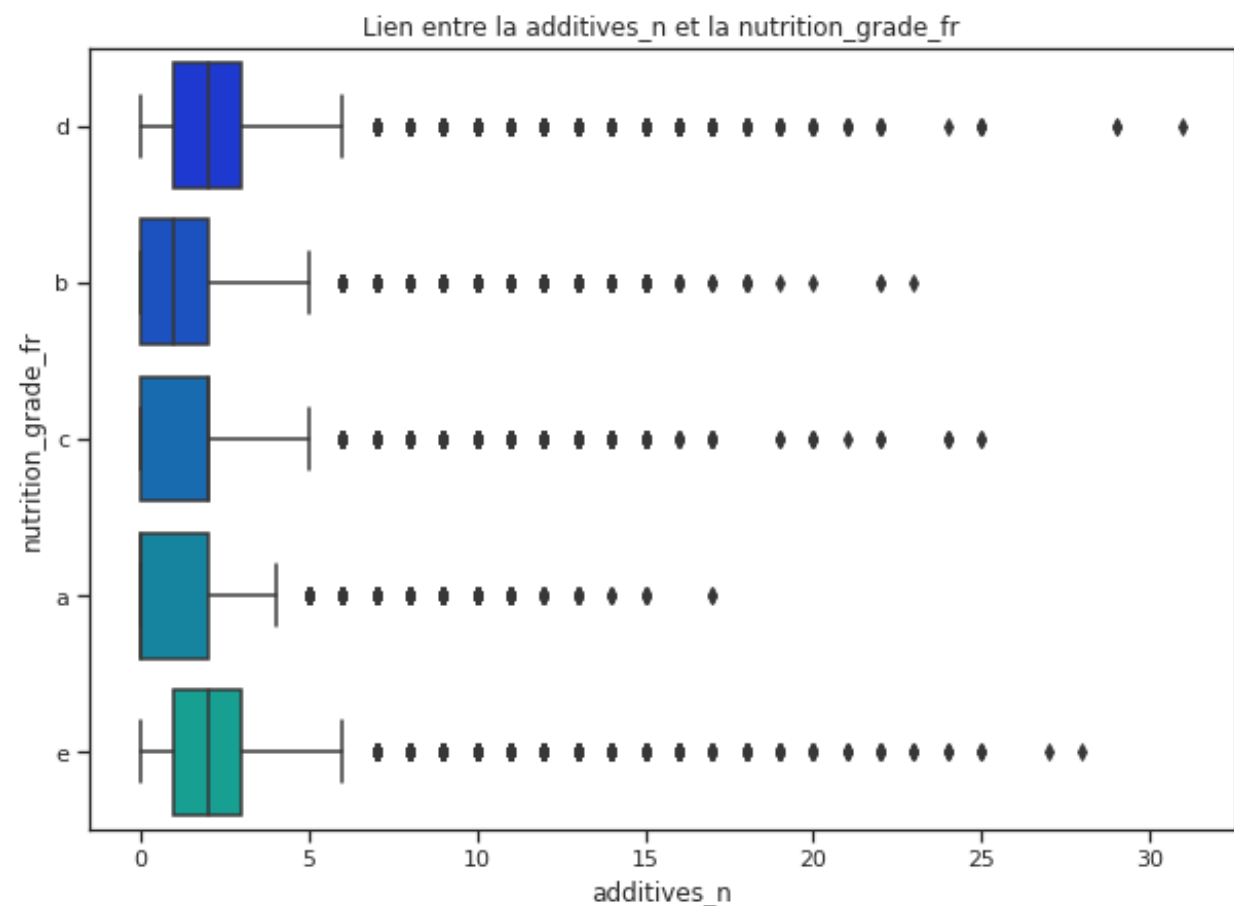
- Pour le 2er graphe, les valeurs ne suivent pas une correlation linéaire, bien que ces derniers soient correlés.
- Pour le graphe 1 et le 3 eme graphe, les données sont pas trop correlés. Et l'erreur commise est un peu importante
- Pour le graphe 4, la correlation est tres forte.
on procede par fusion des colonnes et on revient a 22 colonnes restantes

Variables qualitatives et quantitatives

Visualisation par boxplot

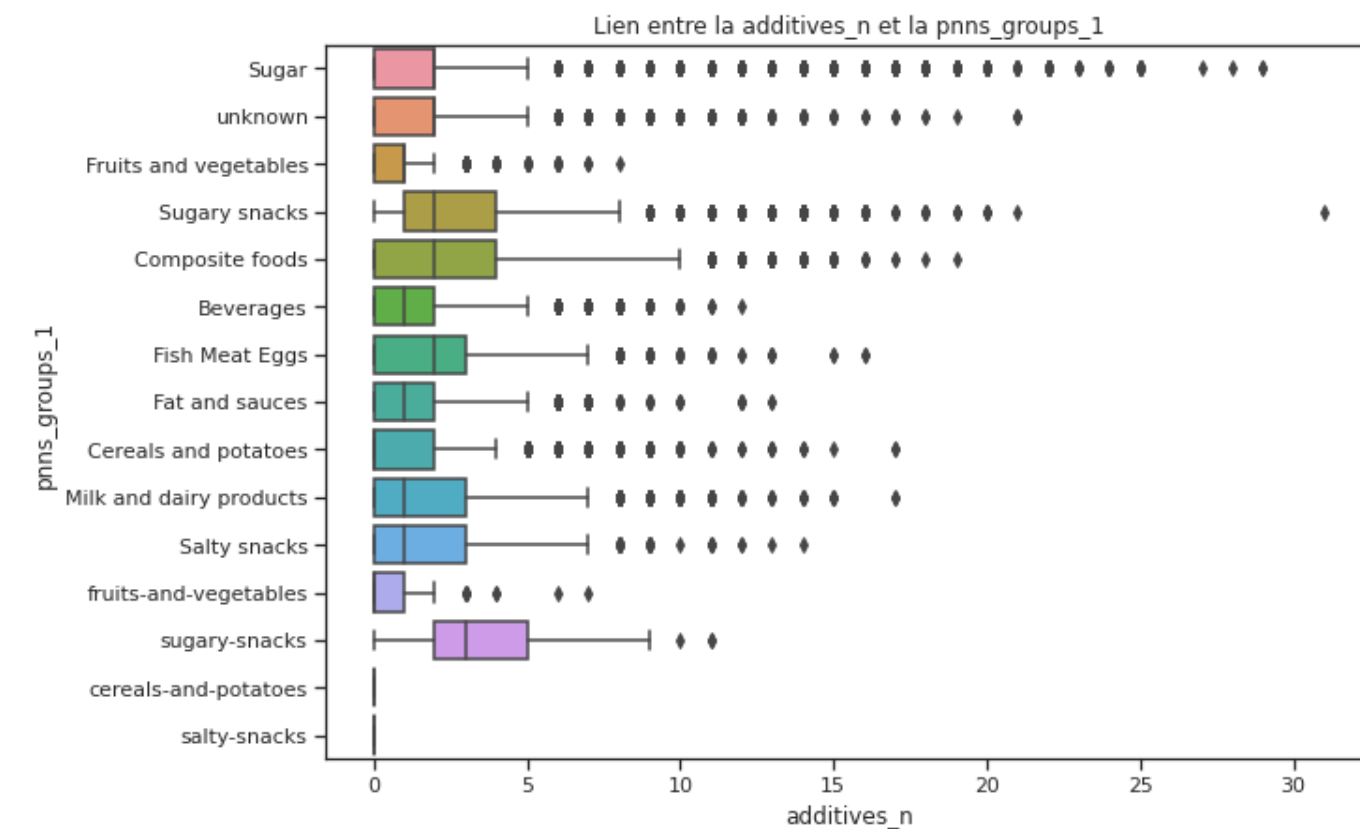
Test anova pour connaitre la p
value

Visualisation par boxplot



relation entre les additives et le nutrigrade_fr

Les produit à meilleur score nutritionnel
ont moins de composants donc sont plus bio
Les mauvais produit ont plus d'additives donc transformés

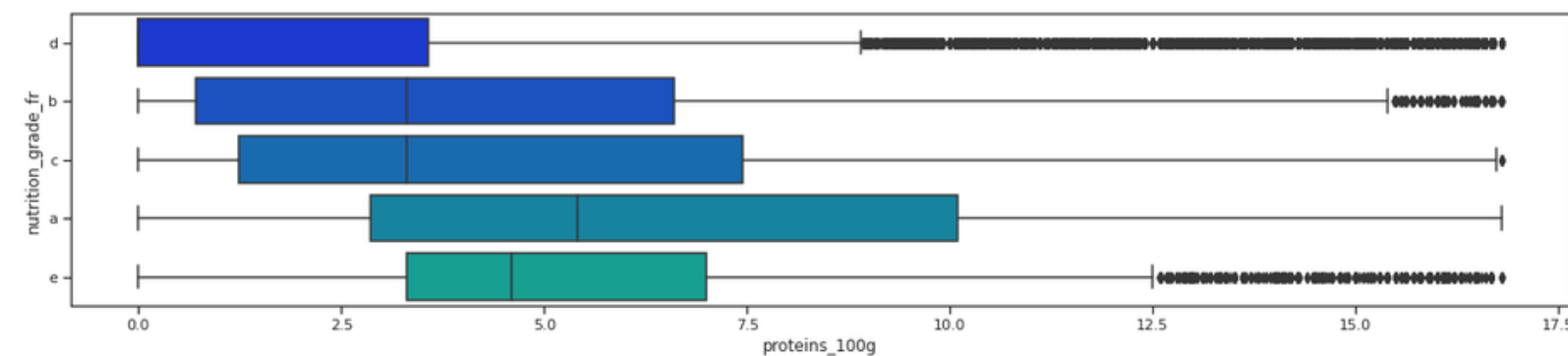


Relation entre le pnns_group et le nombre d additives

Plus on a du sucre, plus le produit
paraît transformé que bio,

Test anova

Nous allons etudier l'anova entre
l'additive n et le nutrition_grade_fr



L'objectif de l'ANOVA,est de montrer si les moyennes des groupes sont significativement différentes. On a les hypothèses suivantes :

- H0 : Les moyennes de chaque groupe sont égales si $p\text{-value} > 5\%$
- H1 : Les moyennes de chaque groupe ne sont pas toutes égales si $p\text{-value} < 5\%$

la pertinence de ce test repose sur la validation de plusieurs hypothèses :

- * l'indépendance entre les échantillons de chaque groupe
- * homoscedasticité (l'égalité des variances) que avec le test de Bartlett.
- * la normalité des résidus avec un test de Shapiro.

1. l'indépendance entre les échantillons de chaque groupe

Nous obtenons 0.1 montre que les 2 variables ne sont pas corrélés.

2. homoscedasticité(Egalité des variances) avec le test de bartlett

```
BartlettResult(statistic=1349.6036722849096,  
pvalue=5.849322309947615e-291)
```

Nous obtenons la p value inférieure à 0.05 donc les variances ne sont pas équivalentes

3. Normalité des résidu

Il s'agit de s'assurer que les résidus suivent une loi normale
Nous allons utiliser le test de Shapiro-Wilk pour tester la normalité:

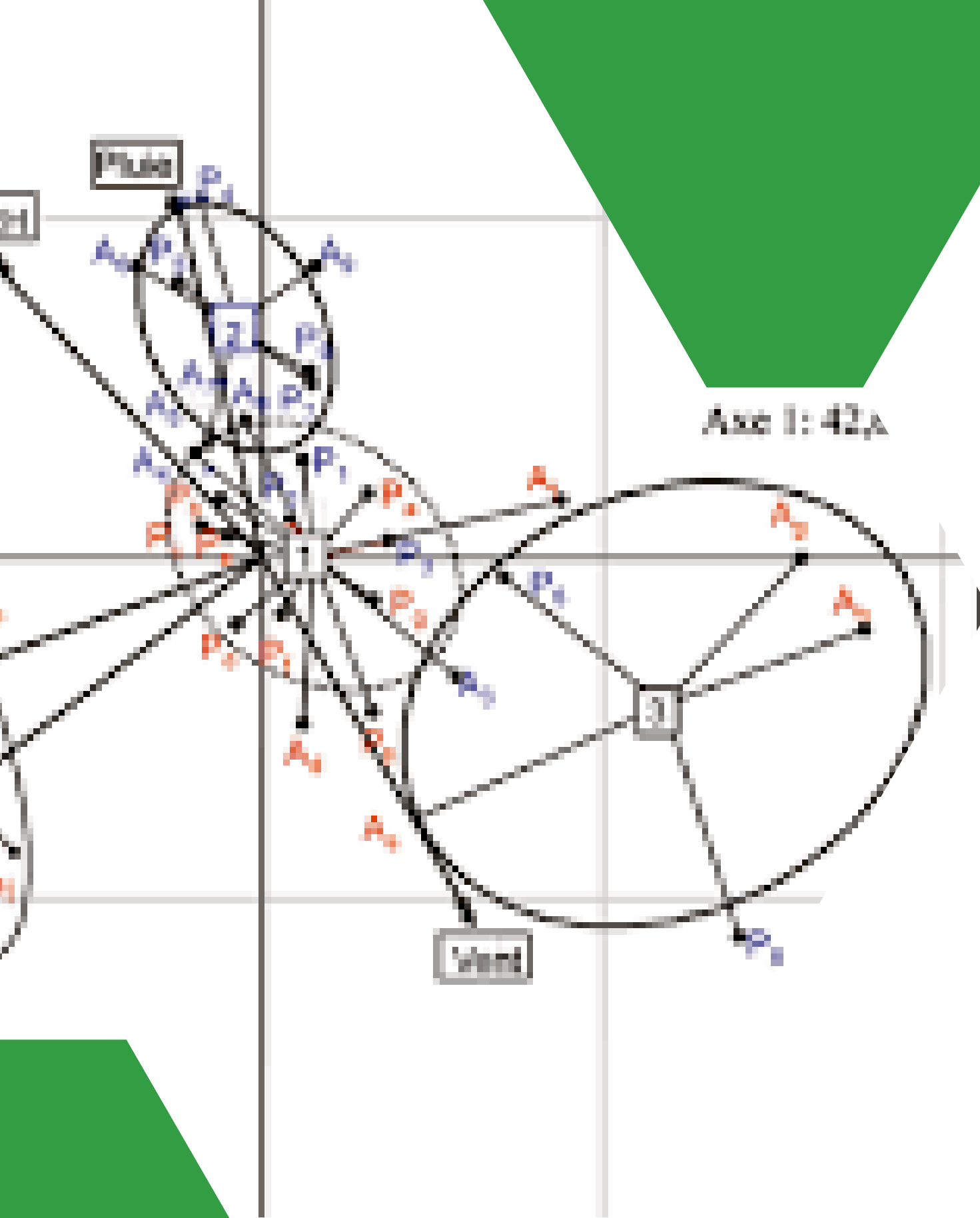
- * H0 : Les résidus suivent une loi normale si p-value > 5%
- * H1 : Les résidus ne suivent pas une loi normale si p-value < 5%

Ici, la p_value est inférieure à 5%, donc les résidus ne suivent pas une loi normale



**Les hypothèses ne sont pas vérifiées, l'anova
n'est donc pas possible dans notre cas.**





Analyse Multivariee

Analyse multivariee par acp

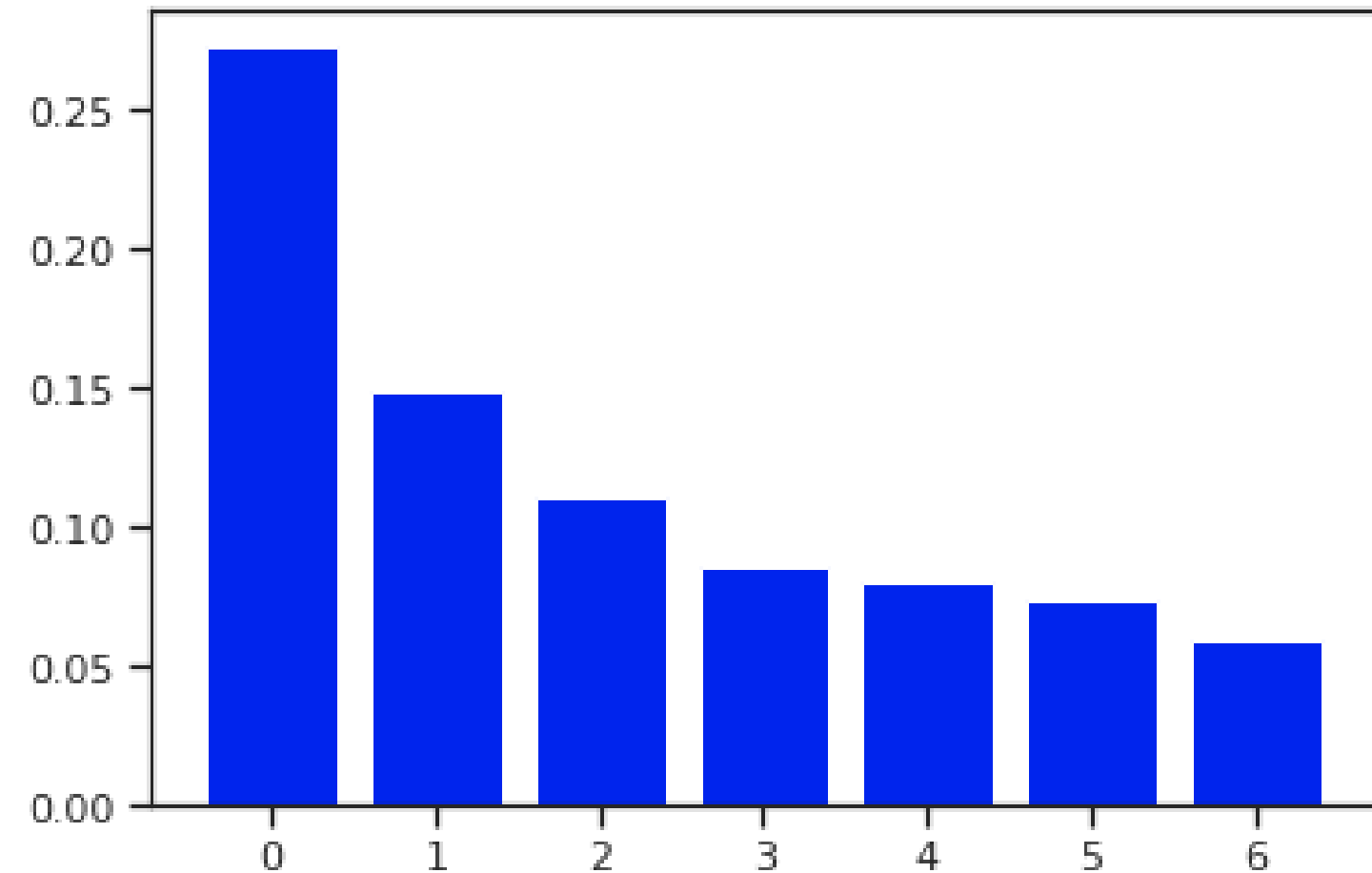
Nous allons par la réduction de dimensions effectuer une analyse multivariée

Meme si on essaie l'ANOVA sans
tenir compte de ces test sur les
variables 2 à 2 , toutes les P_values
sont inférieures à 0.05

```
pval: 0.0  
pval: 0.0  
pval: 0.0  
pval: 0.0  
pval: 7.081349854080156e-142  
pval: 0.0  
pval: 0.0  
pval: 0.0  
pval: 0.0002207109760767863  
pval: 1.3641344379451094e-83  
pval: 2.4061061369818593e-19  
pval: 0.9999999999754239  
pval: 0.0  
pval: 0.0  
pval: 0.0  
pval: 0.0  
pval: 0.0  
pval: 0.0  
pval: 0.0  
pval: 1.2266154465945129e-297  
pval: 0.007043211290410931  
pval: 0.036472961756692124  
pval: 0.03671978630733081  
pval: 8.125486155910161e-07  
pval: 6.889198180220193e-07  
pval: 0.10329112250305725  
pval: 5.2832716377544963e-20  
pval: 0.00012889547748073937  
pval: 5.836904981354637e-101  
pval: 2.7146342767470083e-21  
pval: 2.1469885005855638e-12  
pval: 9.472935790913714e-07  
pval: 0.1984065077820041  
pval: 0.597840607136972  
pval: 0.0  
pval: 0.0
```

Reduction de dimesion avec ACP

On se decide de cincerver au moins 80% de la variance. On construit l histogramme des ratios des composantes du pca



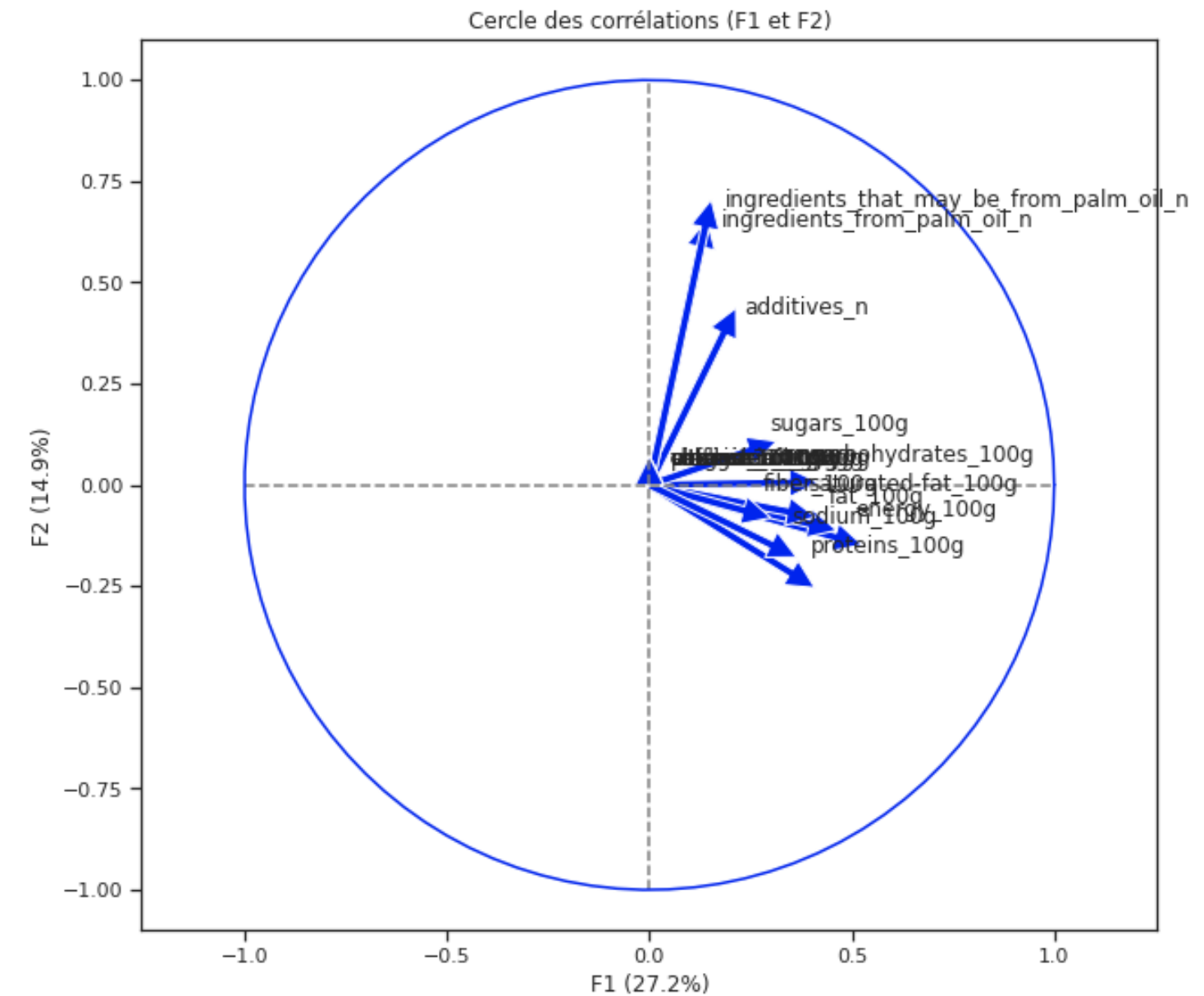
Il nous faut donc au total
7 composantes pour avoir
une variance de 80%
soient 4plans factoriels.

Le cercle des corrélations

Certaines variables sont mieux représentées sur le premier plan factoriel comme l'energy, le nutriscore pour 100g, et le nombre d'ingrédients provenant de l'huile de palm.

Pour avoir au moins 80% de la variance, l'acp nous affirme qu'il nous faut au moins les 3 plans factoriels.

On est passé de 20 variables à 7 composantes principales



Evaluation de l'idée d'application

Notre application peut être très bien réalisée car nous disposons, après le nettoyage et l'analyse, des variables sur lesquelles peuvent s'entraîner notre modèle de KNN pour faire les prédictions.



Merci !