Préparez des données pour un organisme de santé publique

Sommaire

1. Contexte
2. Nettoyage des données
3. Analyse des données
4. Application
5. Conclusions

Contexte

L'agence Santé publique France souhaite améliorer sa base de données **Open Food Facts** et fait appel aux services de notre entreprise.

Cette base de données open source est mise à la disposition de particuliers et d'organisations afin de leur permettre de connaître la qualité nutritionnelle de produits.

Elle met à disposition un ensemble d'informations sur chaques produits tel que la présence d'huile de palme, le nuitriscore, le taux de sel pour 100g etc…

Santé publique France

open FOOD facts

Contexte

- Notre objectif est de réaliser une analyse sur la faisabilité d'une application de prédiction sur un champs du dataset **Open Food Facts**

- Le champs que j'ai choisi renseigne sur la présence d'huile de palme dans un aliment (*ingredients_that_may_be_from_palm_oil_n*)

Contexte

Données du dataset

Informations générales: Nom du produit, code bar, créateur…

Tags: Catégorie, pays de vente, marque …

Ingrédients: Liste d'ingrédients, additifs …

Nutrition: Graisses, sucres, sels, énergie…

Nettoyage des données

Réduction des données

code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name
3087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir
4530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)

energy-from-fat_100g	fat_100g	saturated-fat_100g	butyric-acid_100g	caproic-acid_100g	caprylic-acid_100g	capric-acid_100g	lauric-acid_100g	myristic-acid_100g	palmitic-acid_100g	stearic-acid_100g	arachidic-acid_100g	behenic-acid_100g	lignoceric-acid_100g	ac
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
NaN	28.57	28.57	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

Nettoyage des données

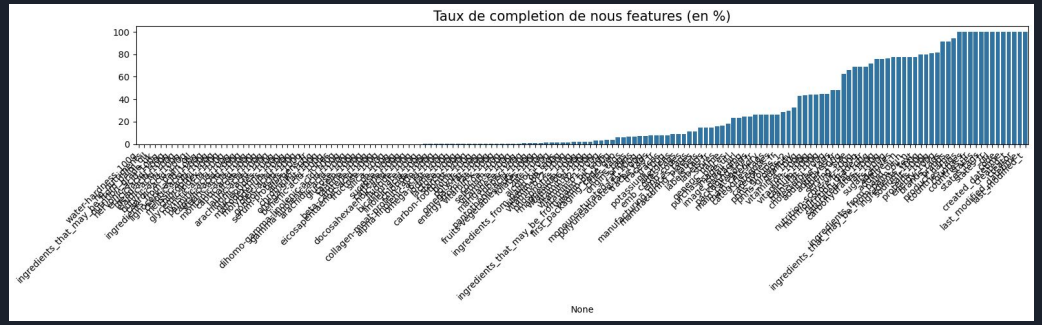
Réduction des données

- Filtrage les colonnes qui ont moins de 50% de valeures définies
    - 162 colonnes -> 34 colonnes
- Garder uniquement les produits Francais
    - 320k lignes -> 94k lignes
- Garder uniquement les lignes ou la feature target est définie
    - 94 lignes -> 50k lignes
- On retire les 0.005 quantile plus grand et plus petit de chaque feature avec des valeurs aberrantes. De sorte à filtrer les outliers

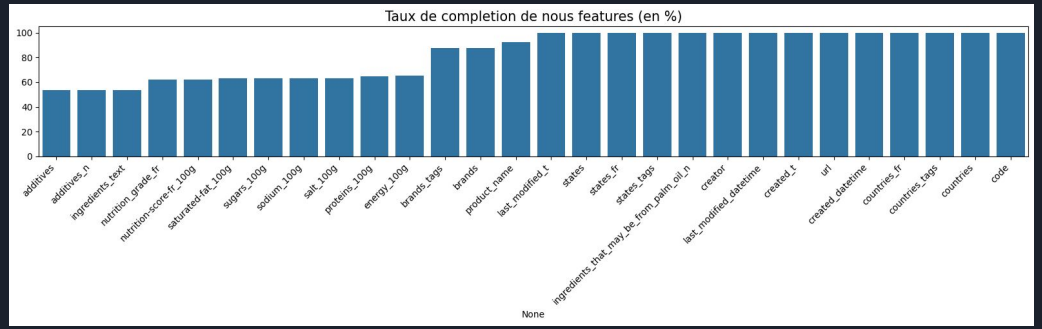
Nettoyage des données

Réduction des données

Avant:

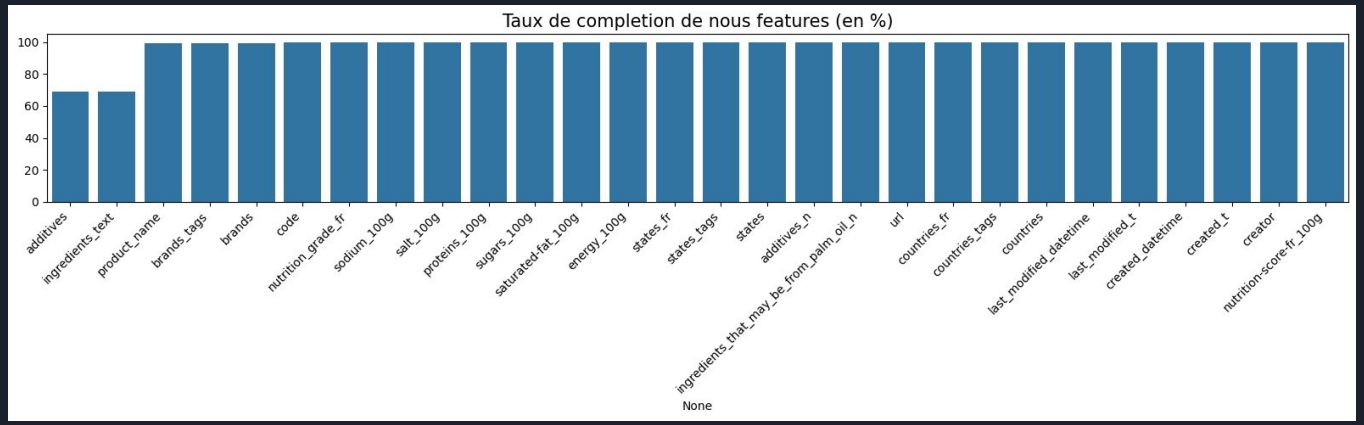


Après:



Imputation: KNN

- 19k lignes imputables
- Pour chaque ligne imputable, on trouve les 10 voisin avec les valeurs les plus proche sur les autres features
- Pour les valeurs catégoriques:
  - On prend la valeur la plus présente parmis les 10. On utilise cette valeur pour l'imputation
- Pour les valeurs numérique
  - On prend la moyenne des 10 voisins.



Analyse des données : Analyse univariée

Catégories de features

2 types de features :

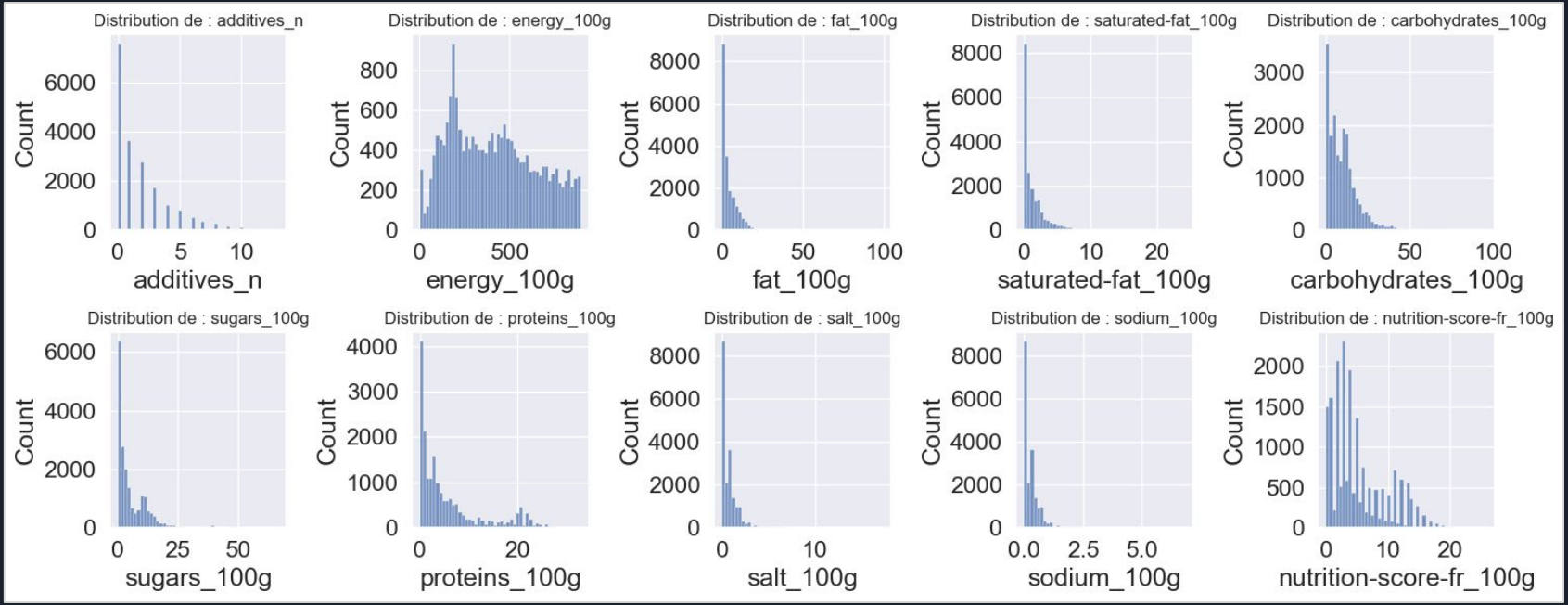
Features continues:

- energy_100g
- fat_100g
- saturated-fat_100g
- carbohydrates_100g
- sugars_100g
- proteins_100g
- salt_100g
- sodium_100g

Features discrètes:

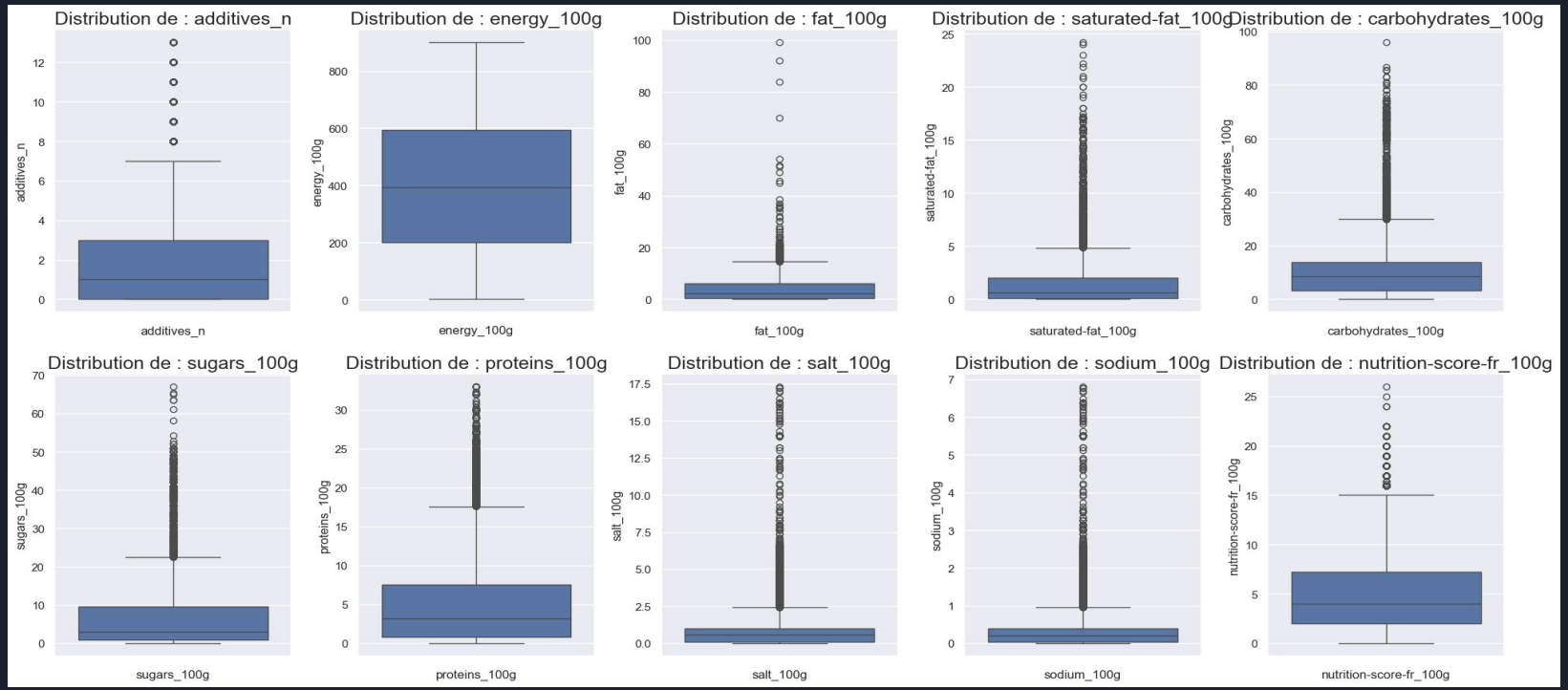
- additives_n
- nutrition-score_100g

Distributions: Bar plots

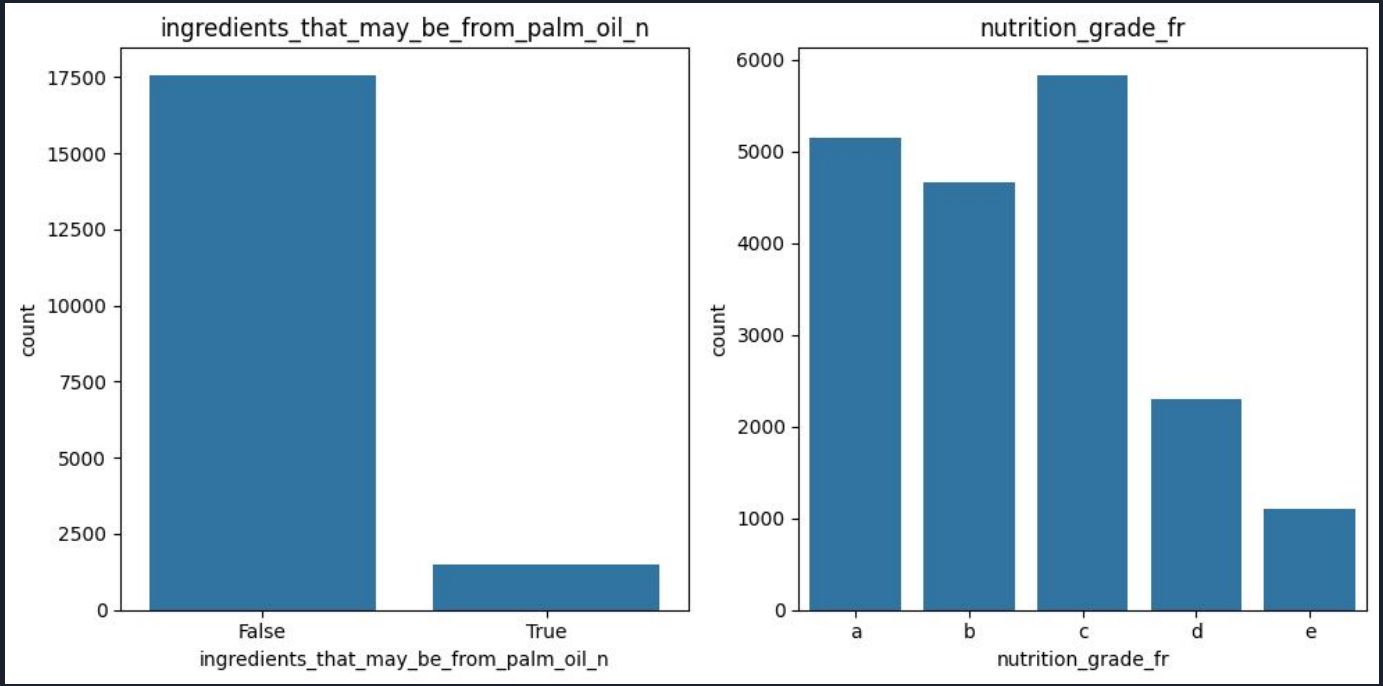


- Forme souvent asymétrique avec étalement à droite

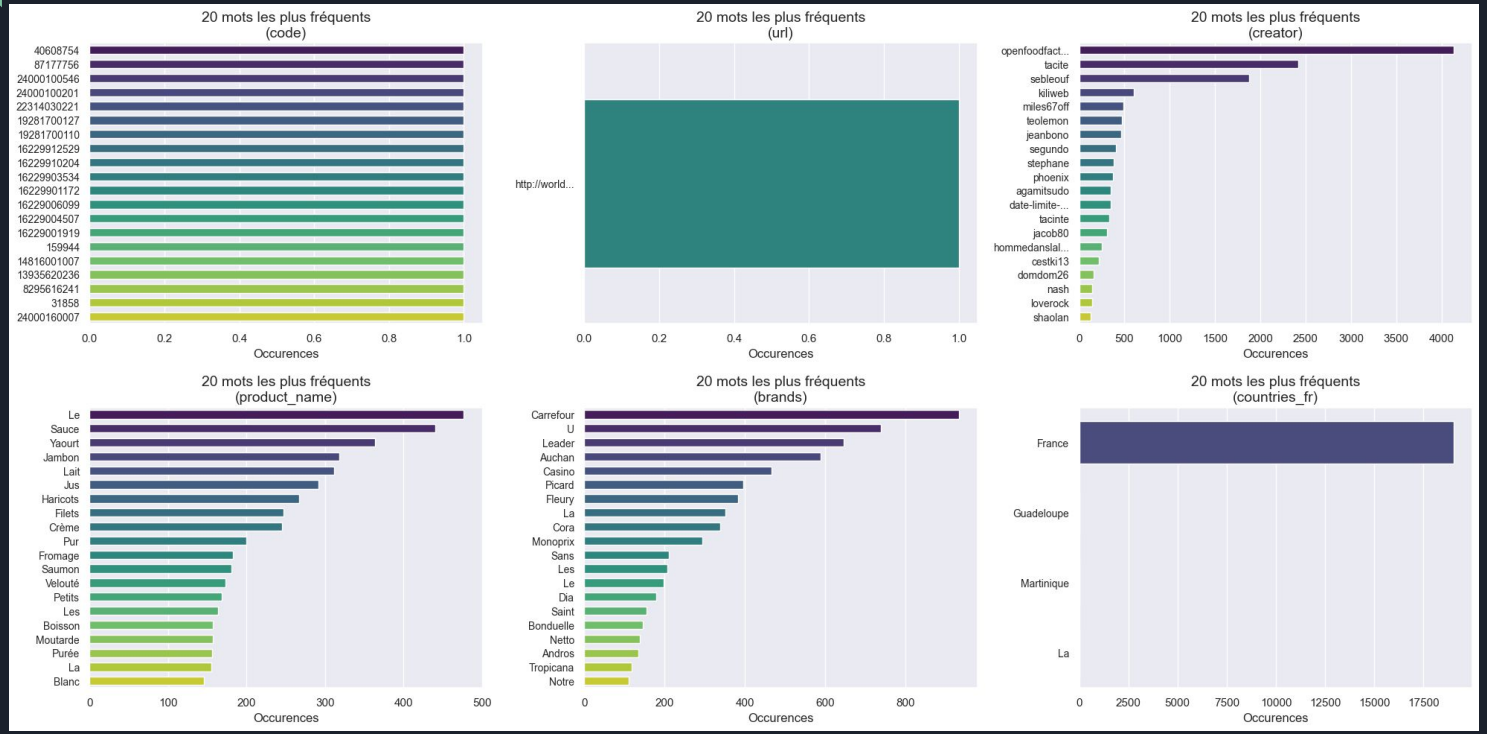
Distributions: Box plots



Distribution: Features discrètes



Occurrences de mots dans les feature textuelles

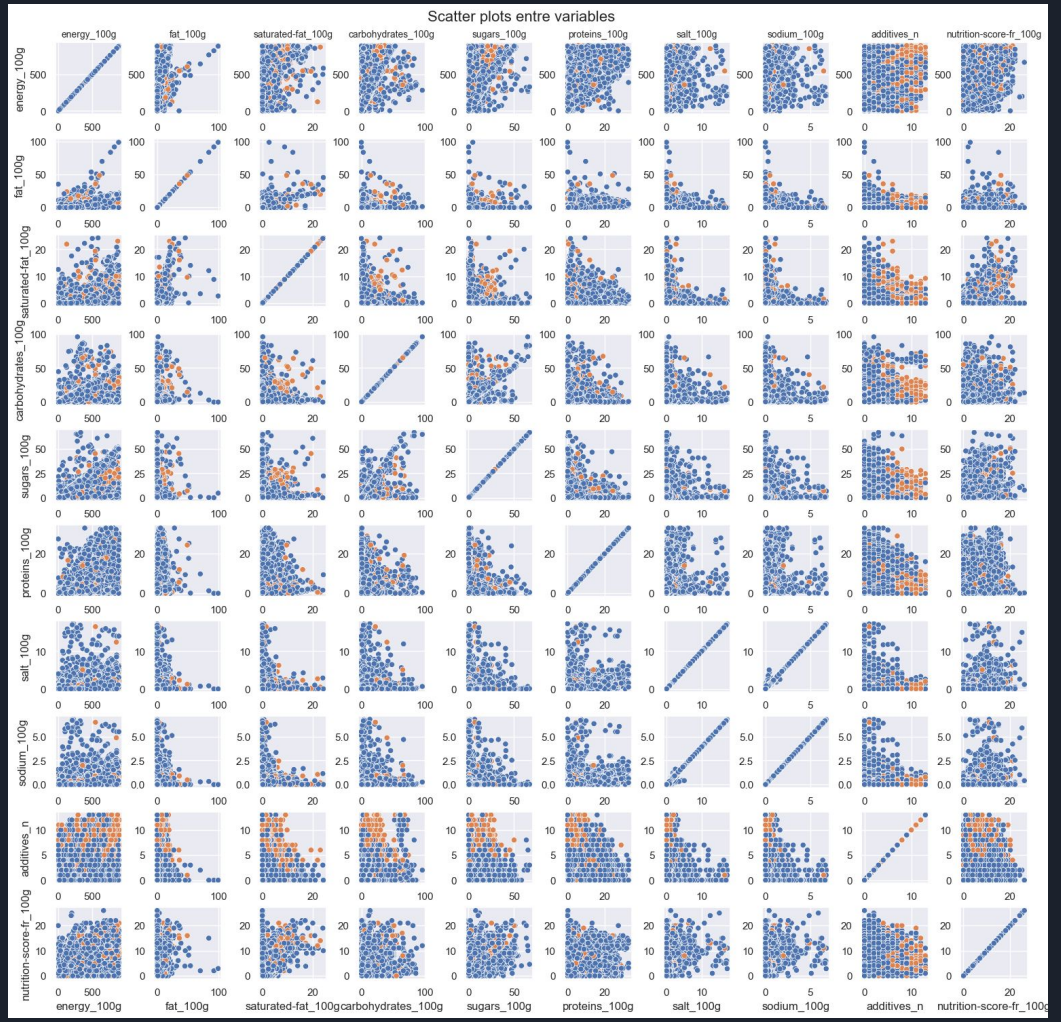


Analyse des données : Analyse bivariée

Scatter plots des features 2 à 2

Seulement quelques variables sont fortement corrélées:

- Graisse et énergie
- Graisses saturées et énergie
- Sel et sodium



Statistiques descriptives

	additives_n	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g
mean	1.791645	412.5189	3.847811	1.386595	10.27392	5.764333	5.777435	0.798958	0.313105	5.160046
std	2.294941	236.1632	4.612705	2.048614	10.04443	7.265970	6.770555	1.215570	0.478068	4.239644
skew	1.776258	0.300534	2.835103	3.200467	2.384330	2.551682	1.497686	5.086778	5.106569	1.044727
kurtosis	3.479100	-0.977531	28.07001	16.15922	9.321812	9.443425	1.246896	44.30677	44.55483	0.372412

- On observe une forte variabilité des valeurs (écarts types élevés), des distributions asymétriques (skew positif, notamment pour les graisses et les acides gras saturés), ainsi que des distributions fortement étalées (kurtosis élevé, notamment pour le sel et le sodium), indiquant la présence de valeurs extrêmes ou de distributions non normales.

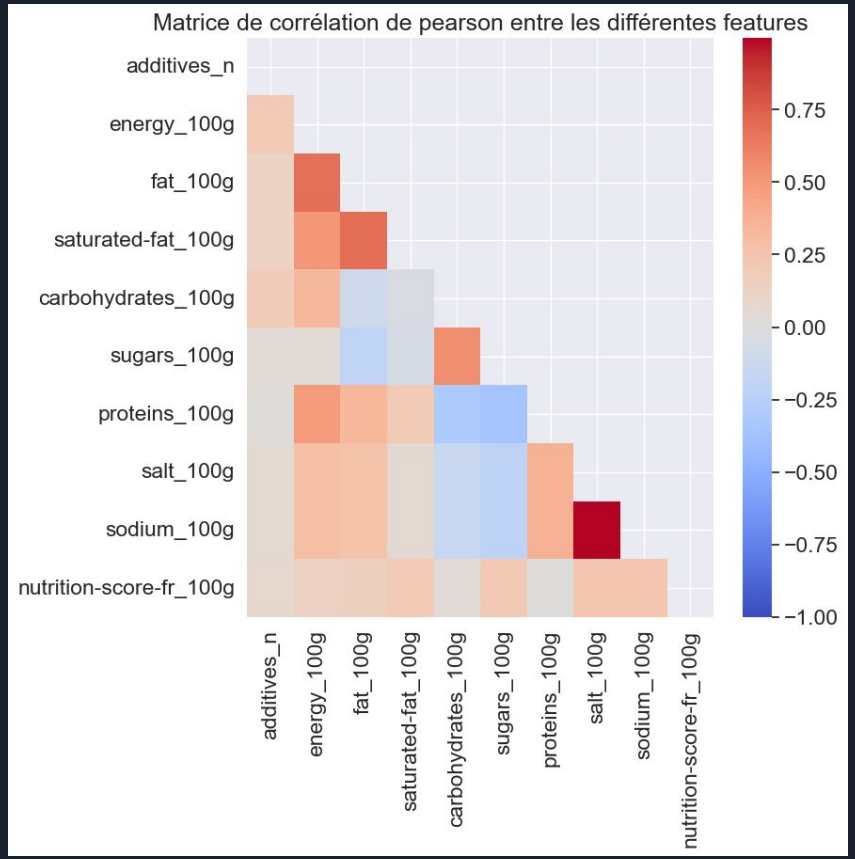
Test de Shapiro-Wilk

additives_n	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g
False	False	False	False	False	False	False	False	False	False

- Le test de Shapiro-Wilk nous permet de vérifier que nos features ne suivent pas une loi normale

Matrice de corrélations de pearson

Même constat que plus tôt, corrélations entre graisses et énergies avec la plupart des autres entre 0 et .40



Test d’indépendance des features

Test de Chi

	additives_n	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g
Chi²	2165.9132	543.55955	464.86169	477.8709	849.0466	342.1521	491.16408	110.91355	110.7863	55.58671
p-valeur	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.00001
DOF	15	19	13	19	19	19	19	19	19	18
Conclusion	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)	Non Indépendantes (H0 rejetée)

- Avec des valeurs de Chi² élevées et des p-valeurs nulles (≈ 0), l'hypothèse nulle (H0) d'indépendance est systématiquement rejetée. Cela indique que toutes les variables testées sont significativement dépendantes entre elles, suggérant de possibles relations ou corrélations fortes entre les caractéristiques nutritionnelles des produits analysés.

ANOVA

Variable	additives_n	energy_100g	carbohydrates_100g	saturated-fat_100g	fat_100g	proteins_100g	salt_100g	sodium_100g	sugars_100g	nutrition-score-fr_100g
p-value	0	2.4913e-99	1.4896e-98	8.6521e-96	6.0927e-51	7.2537e-11	1.1420e-07	1.2673e-07	2.8682e-07	2.0696e-04

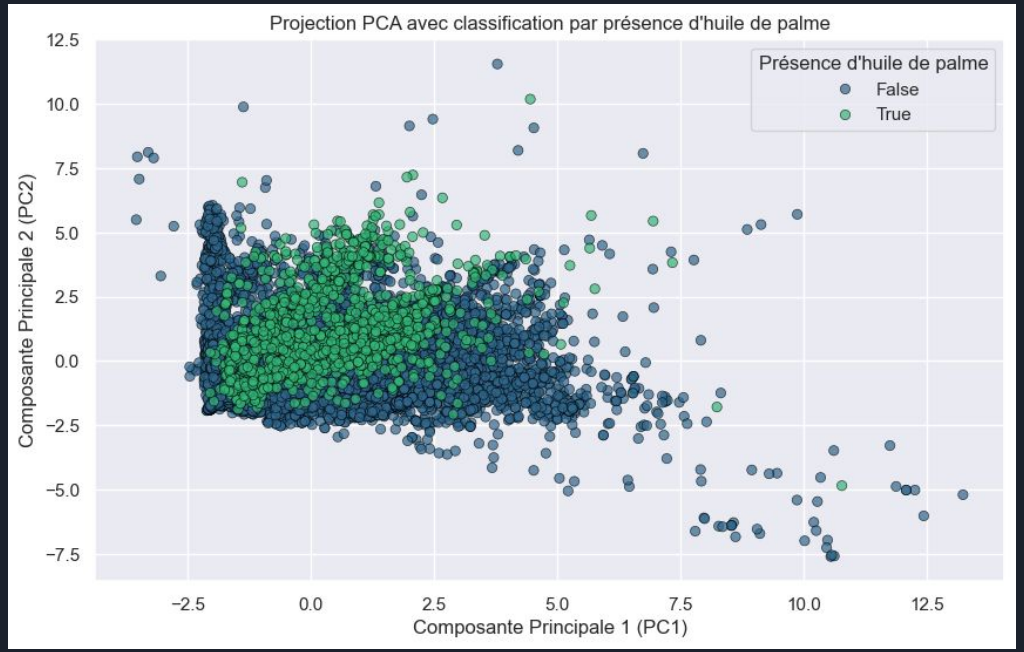
- ANOVA réalisée entre chacune de mes features et ma target *ingredients_that_may_be_from_palm_oil_n*

- Certaines variables, comme le nombre d'additifs, l'énergie et la teneur en lipides/glucides, sont fortement liées à la présence d'huile de palme. Cela peut indiquer que les produits contenant de l'huile de palme ont des profils nutritionnels distinct.

PCA

Projections

- La majorité des observations sont concentrées autour de (0,0), ce qui est typique d'une PCA normalisée.
- Il y a quelques valeurs extrêmes, notamment vers des PC1 élevées (> 7.5) et PC2 élevées (> 10), suggérant des outliers.
- La séparation limitée entre les 2 classes suggère qu'une classification ne serait pas optimale. Un travail plus important sur la donnée peut donc être nécessaire



PCA

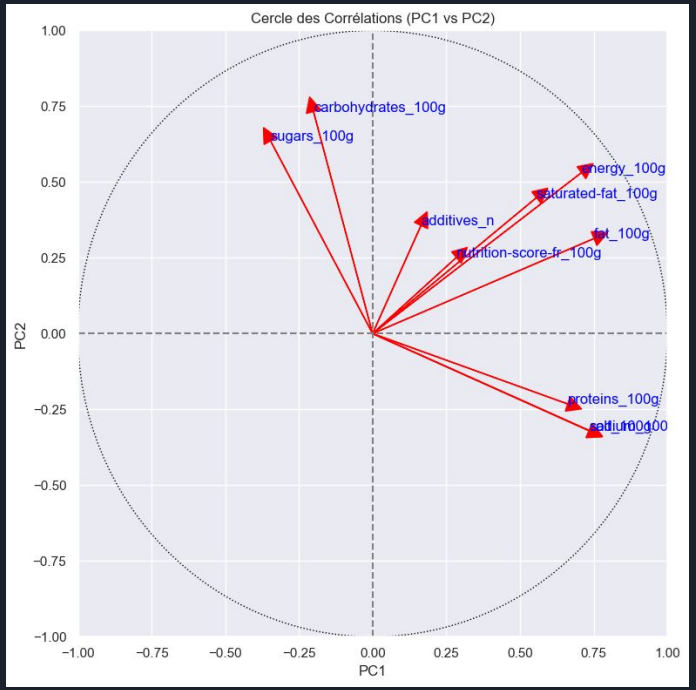
Cercle des corrélations

1. Qualité de représentation des variables (longueur des flèches)

- Les flèches longues qui touchent presque le cercle (ex: **energy_100g, fat_100g, carbohydrates_100g, proteins_100g, sodium_100g**) indiquent que ces variables sont **bien représentées** par les deux premières composantes principales.
- En revanche, si certaines flèches étaient courtes et proches du centre, cela signifierait que ces variables ne sont **pas bien expliquées** par PC1 et PC2.

2. Corrélations entre variables (directions des flèches)

- **energy_100g, fat_100g et saturated-fat_100g** pointent dans une direction similaire → Elles sont **fortement corrélées entre elles**.
- **sodium_100g et proteins_100g** sont opposés aux autres variables → Ils ont une relation **inverse** avec ces dernières.
- **carbohydrates_100g et sugars_100g** sont très proches → Logique, car les sucres sont une sous-catégorie des glucides.



Conclusions

- Faisabilité
- Utilisation des variables PCA
- Limites