# Préparez des données pour un organisme de santé publique

# Sommaire

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

# Objectifs

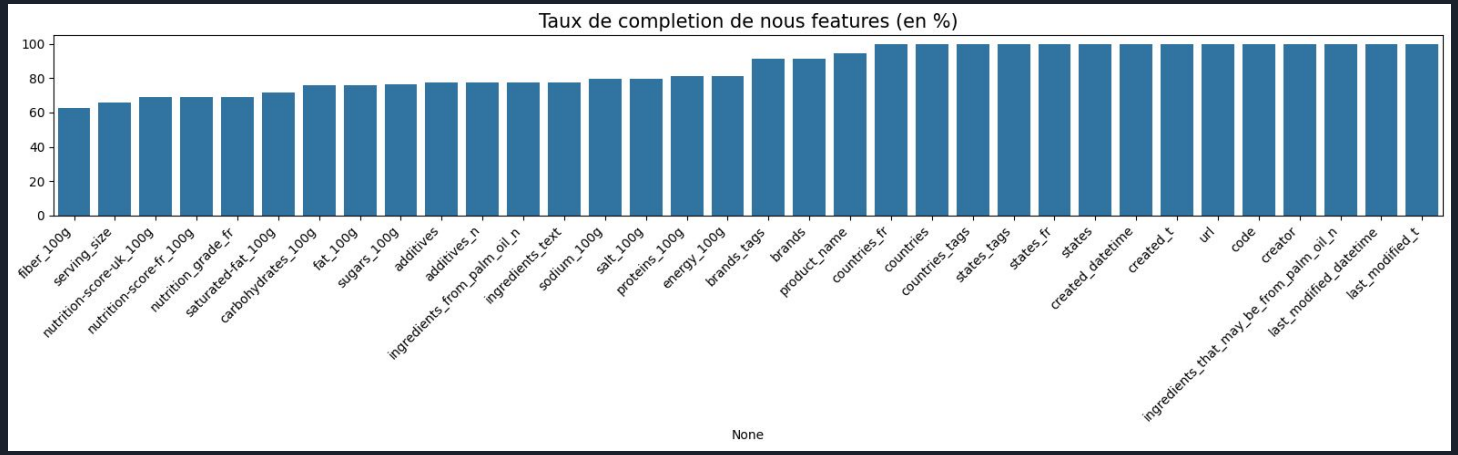
choix de la target

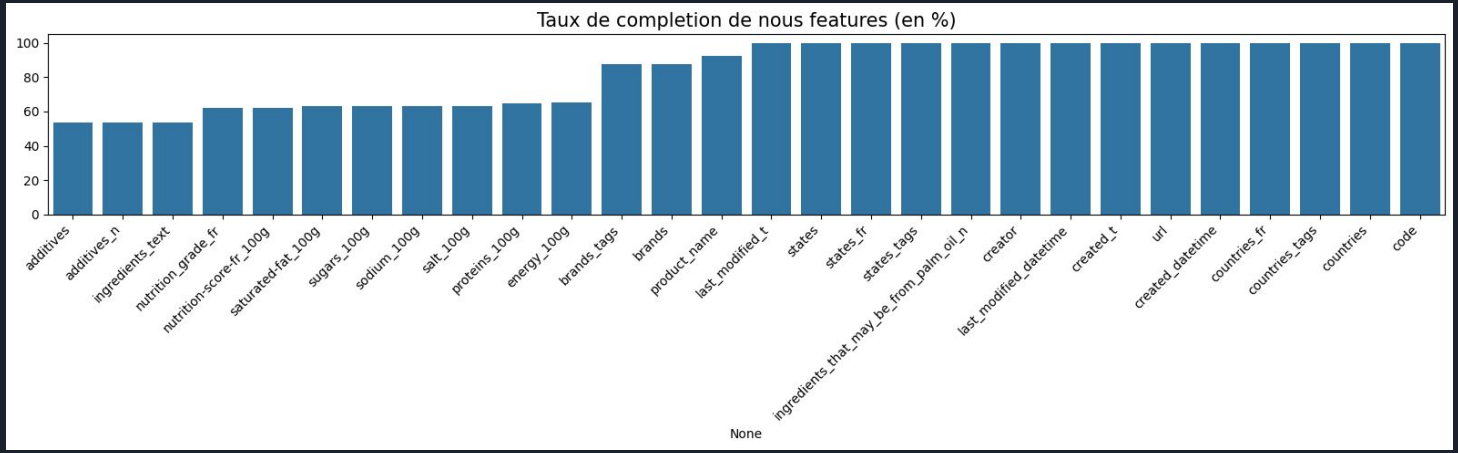
# Dataset

- dimensions
- valeurs Nulles
-

# Réduction des données: Dimensions

- Filtrage les colonnes qui ont moins de 80% 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


Taux de completion de nous features (en %)
100
80
60
40
20
0
fiber_100g
serving_size
nutrition-score-uk_100g
nutrition-score-fr_100g
nutrition_grade_fr
saturated-fat_100g
carbohydrates_100g
fat_100g
sugars_100g
additives
additives_n
ingredients_from_palm_oil_n
ingredients_text
sodium_100g
salt_100g
proteins_100g
energy_100g
brands_tags
brands
product_name
countries_fr
countries
countries_tags
states_tags
states_fr
states
created_datetime
created_t
url
code
creator
ingredients_that_may_be_from_palm_oil_n
last_modified_datetime
last_modified_t
None


Taux de completion de nous features (en %)
100
80
60
40
20
0
additives
additives_n
ingredients_text
nutrition_grade_fr
nutrition-score-fr_100g
saturated-fat_100g
sugars_100g
sodium_100g
salt_100g
proteins_100g
energy_100g
brands_tags
brands
product_name
last_modified_t
states
states_fr
states_tags
ingredients_that_may_be_from_palm_oil_n
creator
last_modified_datetime
created_t
url
created_datetime
countries_fr
countries_tags
countries
code
None

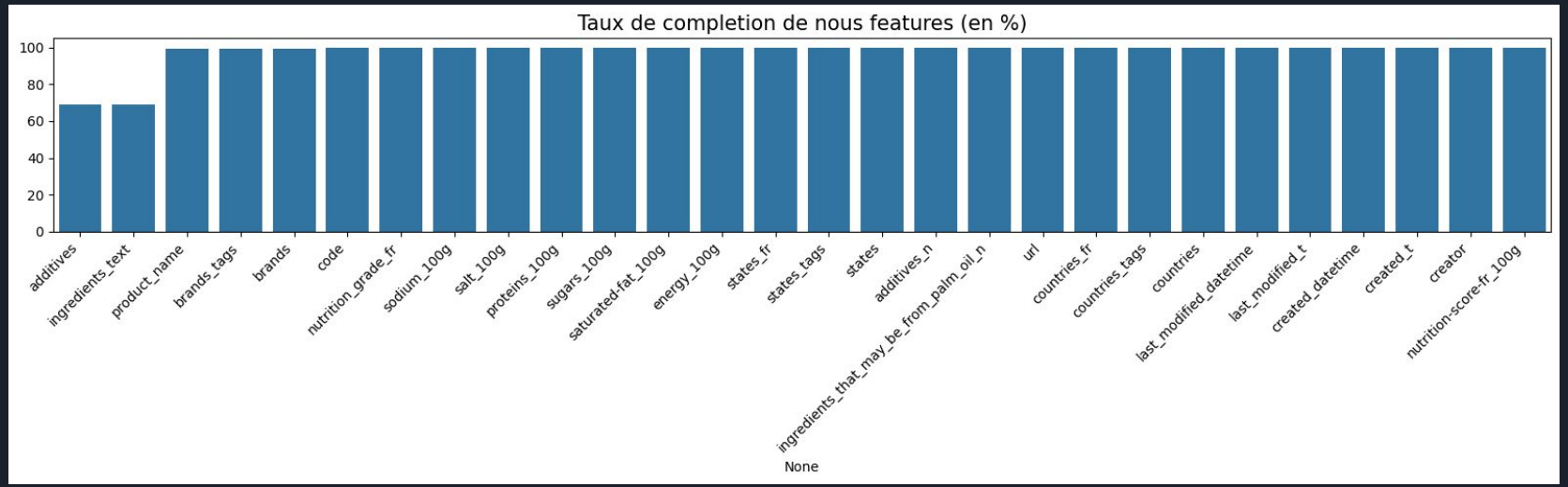
# Réduction des données: Outliers

- On retire les 0.005 quantile  plus grand et plus petit de chaque feature avec des valeurs aberrantes. De sorte à filtrer les outliers
-

# 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.

# Imputation: KNN



# Analyse des données : Analyse univariée

# Catégories de features

**2 types de features :**

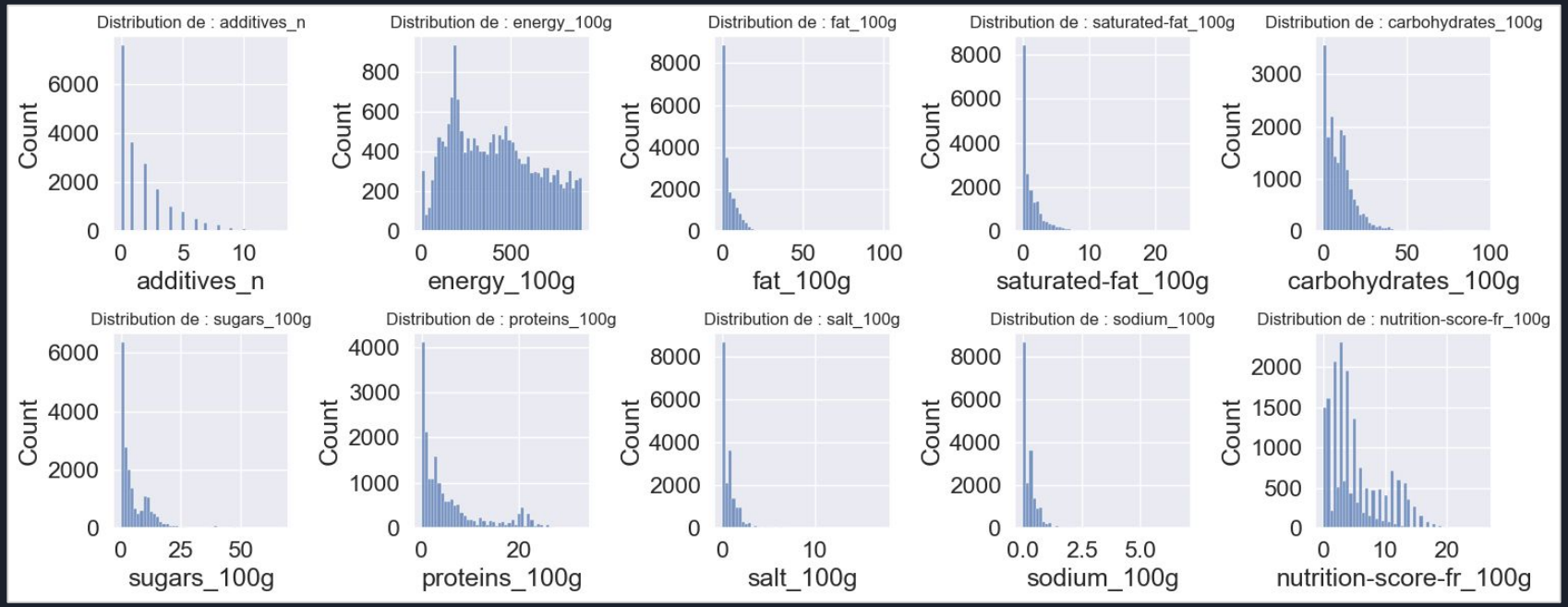
**Features continues:**

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

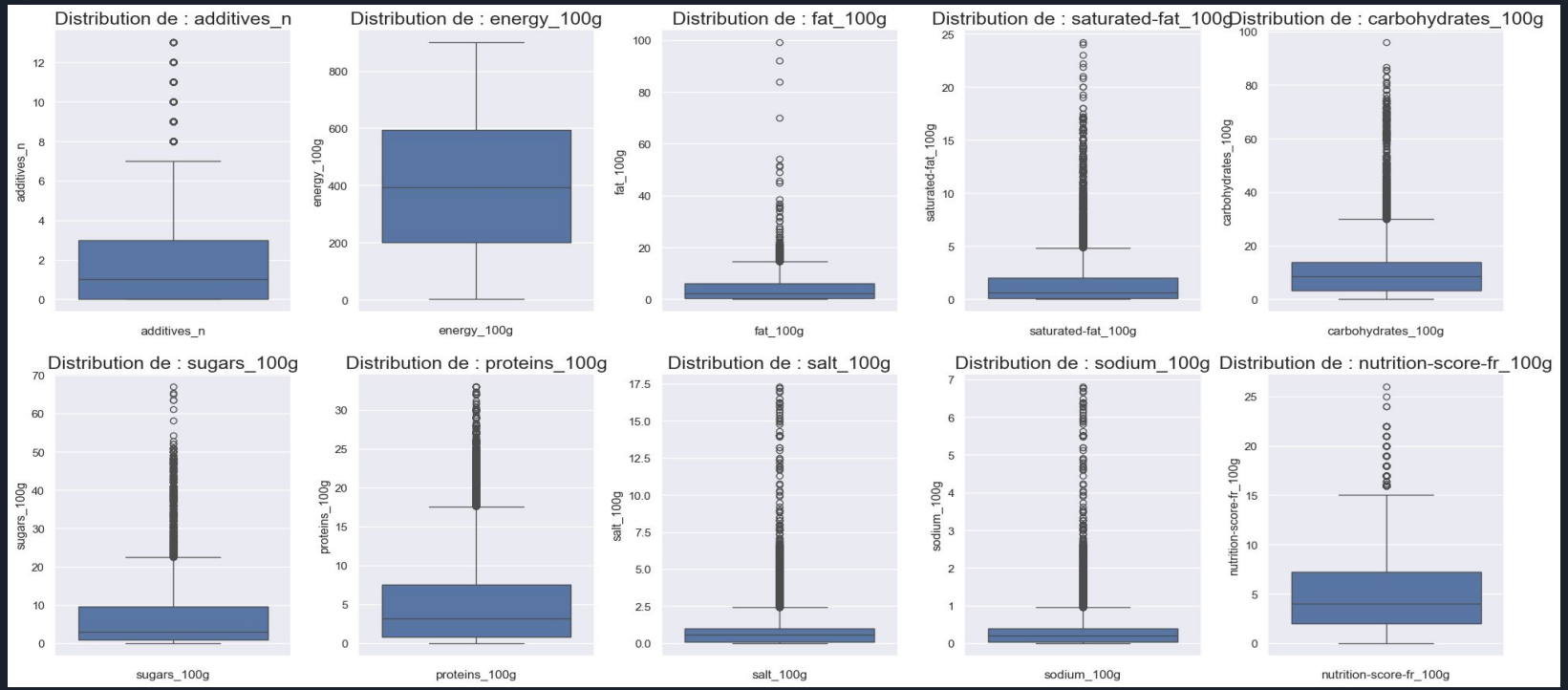
**Features discrètes:**

- additives_n
- nutrition-score_100g

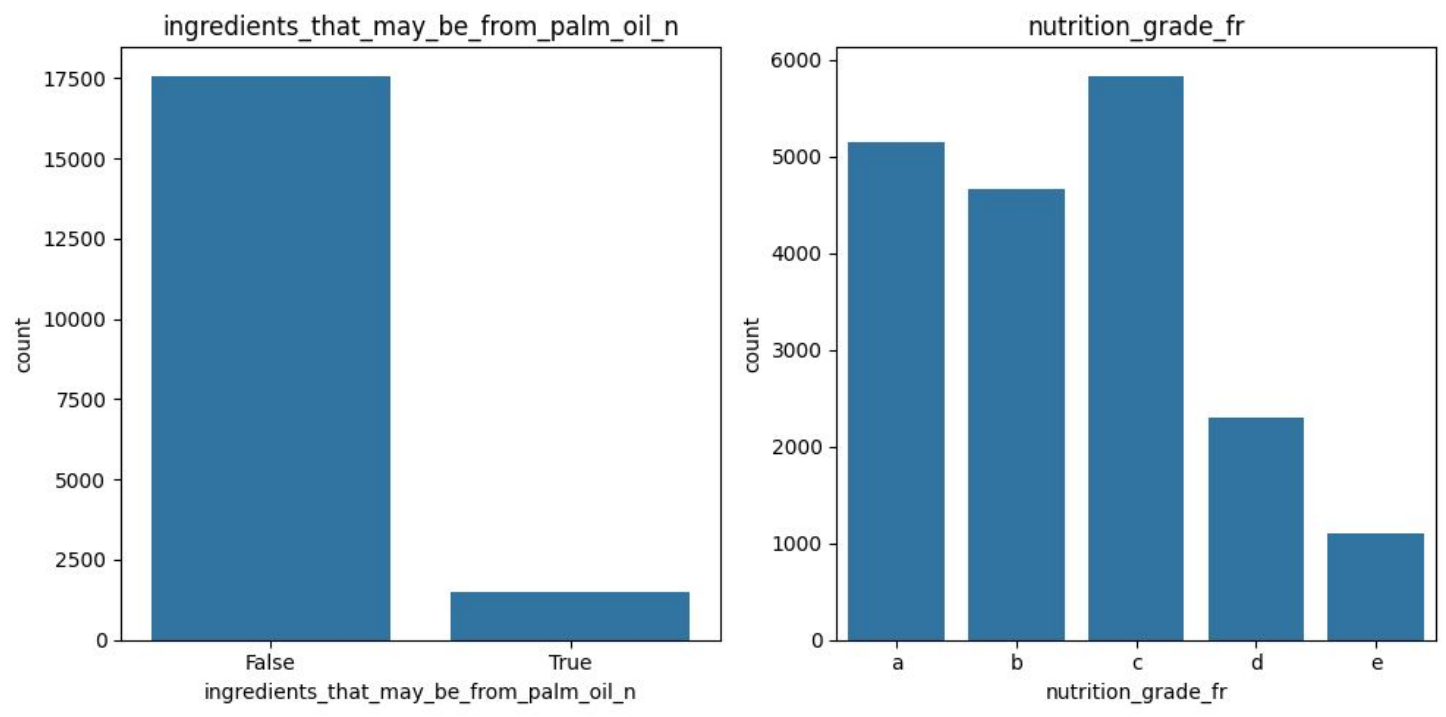
# Distributions: Bar plots



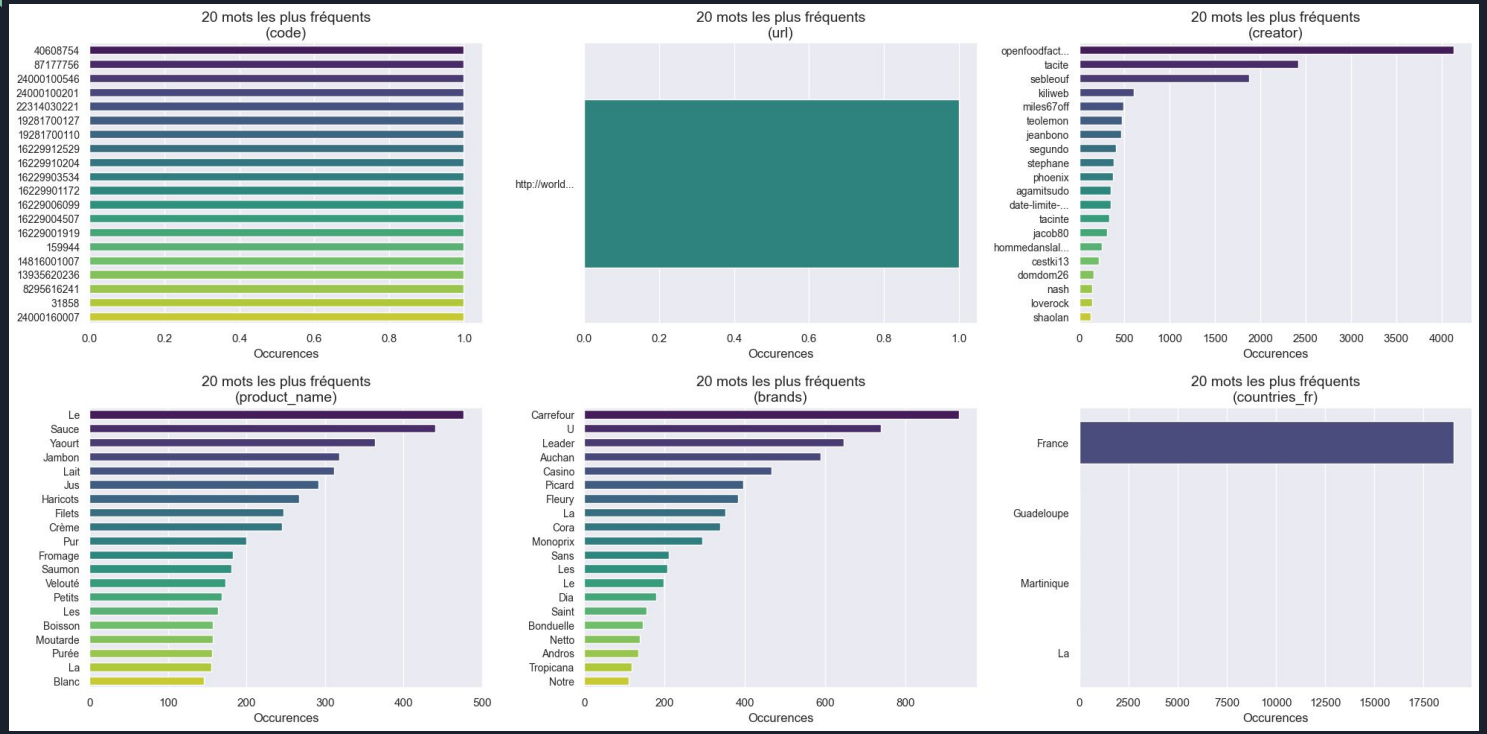
# Distributions: Box plots



# Distribution: Features discrètes

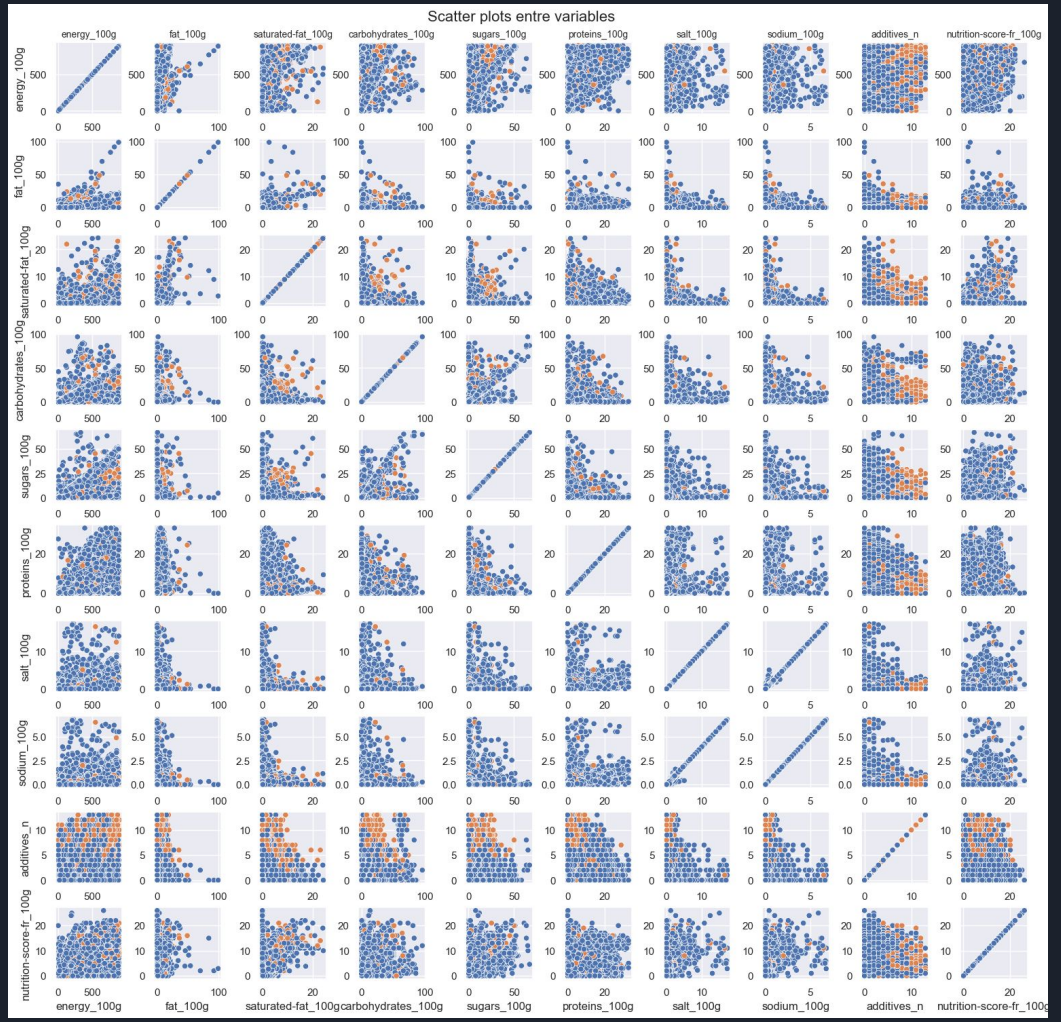


# Occurrences de mots dans les feature textuelles



# Analyse des données : Analyse bivariée

# Scatter plots des features 2 à 2



# Statistiques descriptives

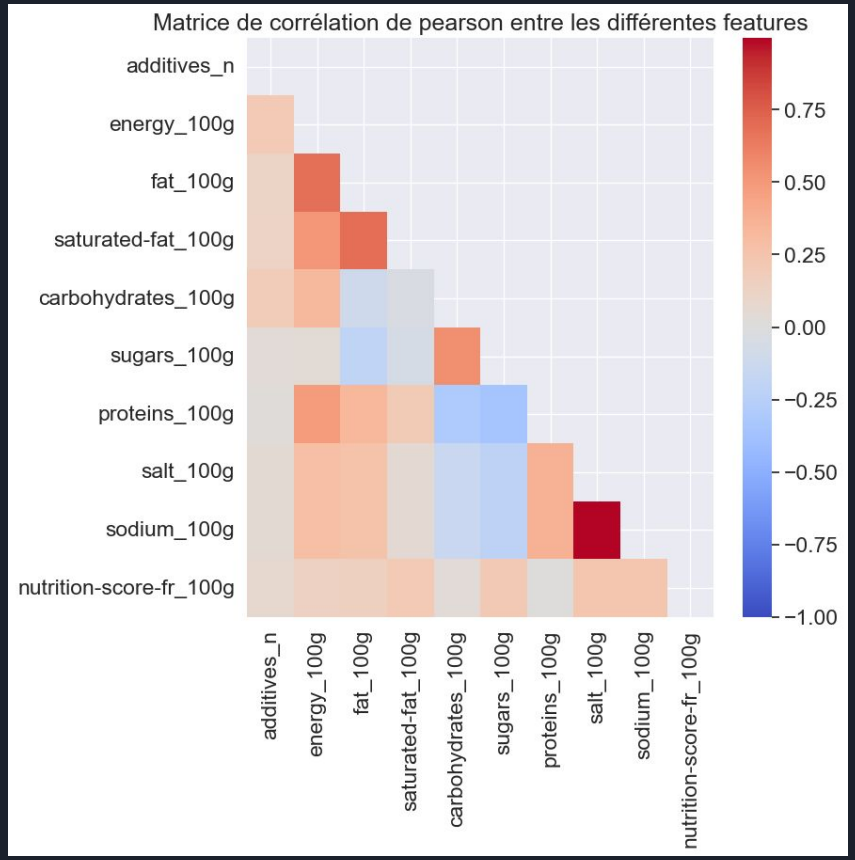
| | additives_n | energy_100g | fat_100g | saturated-fat_100g | carbohydrates_100g | sugars_100g | proteins_100g | salt_100g | sodium_100g | nutrition-score-fr_100g |
|---|---|---|---|---|---|---|---|---|---|---|
| **mean** | 1.791645 | 412.5189 | 3.847811 | 1.386595 | 10.27392 | 5.764333 | 5.777435 | 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 |

# 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 |

- Les features ne suivent pas une loi normale

# Matrice de corrélations de pearson



# Test d’indépendance des features : Test de Chi

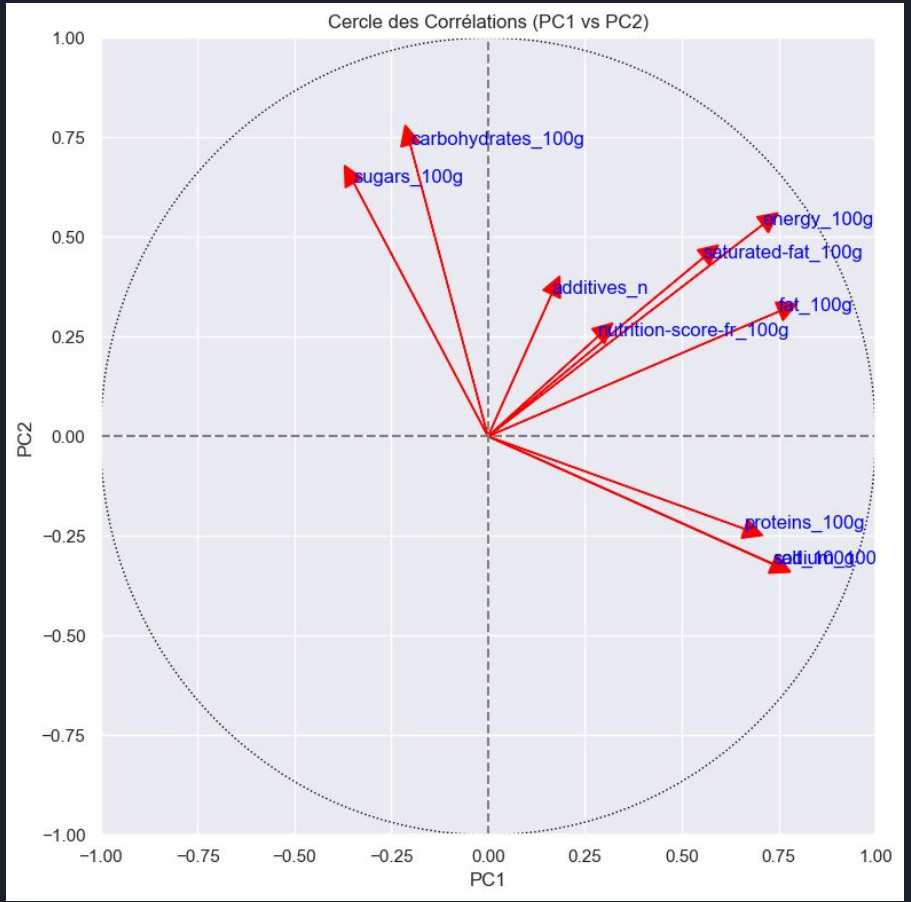
| | additives_n | energy_100g | fat_100g | saturated-fat_100g | carbohydrates_100g | sugars_100g | proteins_100g | salt_100g | sodium_100g | nutrition-score-fr_100g |
|---|---|---|---|---|---|---|---|---|---|---|
| **Chi²** | 2165.9132 | 543.55955 | 464.86169 | 477.8709 | 849.0466 | 342.1521 | 491.16408 | 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) |

**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