OPENCLASSROOMS

# Projet n°2 – Concevez une application au service de la santé publique

## Parcours Ingénieur machine learning

Fayz EL RAZAZ

Soutenance réalisée devant

Mohammed Sedki

# Plan de la présentation

- I. Rappel de l’appel à projets et idée d’application
- II. Démarche méthodologique de nettoyage
- III. Démarche méthodologique d’exploration de données
- IV. Présentation des faits pertinents pour l’application

# I. Appel à projets et idée d'application

- **Appel à projet**

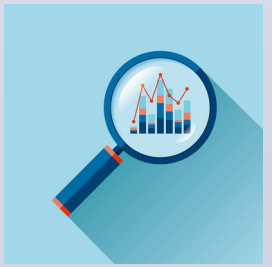
- Qui ? Agence santé publique France

Santé publique France

- Objectifs : Idées innovantes d'applications en lien avec l'alimentation

- **Quelques chiffres globaux**

- Jeu de données provenant d'Open Food Facts :



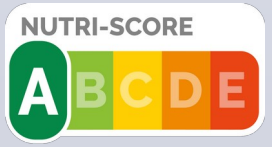
  - 320772 produits
  - 162 features
  - Produit de plus de 100 pays
  - 4 types d'informations dans le data set :
    - Informations générales
    - Ensemble de tags
    - Ingrédients
    - Valeurs nutritionnelles

# I. Appel à projets et idée d'application

- **Idée d'application**

Suggestion d'un produit de même catégorie et de meilleur qualité :

- Meilleur nutriscore




- 

- Packaging plus écologique

# II. Démarche méthodologique de nettoyage

## Les étapes du nettoyage

1. Filtrage des features et produits

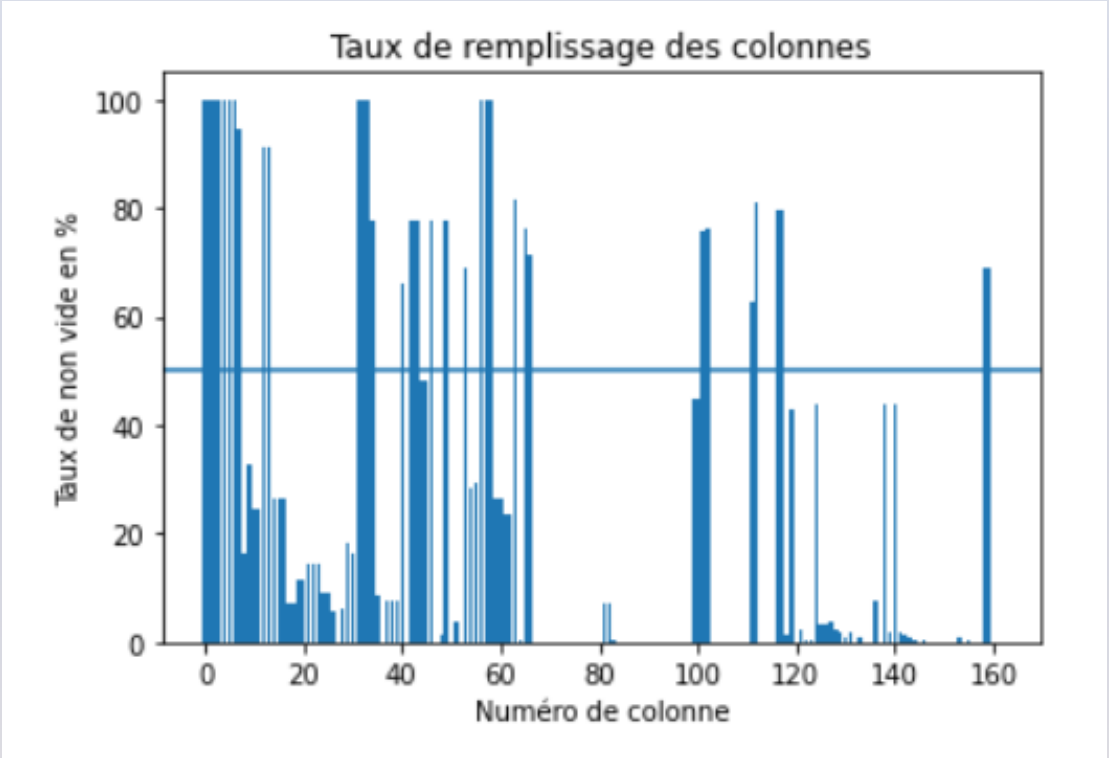
2. Traitement des valeurs aberrantes

3. Traitement des valeurs manquantes

# II. Démarche méthodologique de nettoyage

## 1. Filtrage des features et produits

- Importation des données
- Prise de contact avec les features
- Affichage du taux de remplissage des colonnes



# II. Démarche méthodologique de nettoyage

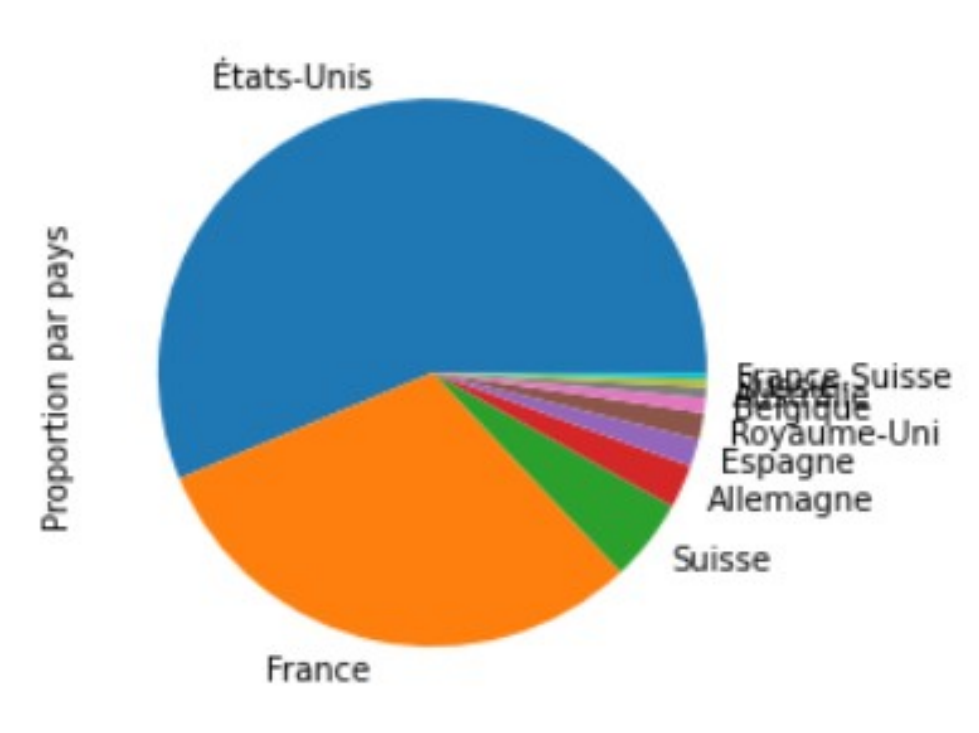
## 1. Filtrage des features et produits

- Suppression des colonnes vides

⇨ Passage de 162 à 146 features

- Filtre sur les données françaises

⇨ Près de 30% de produits français



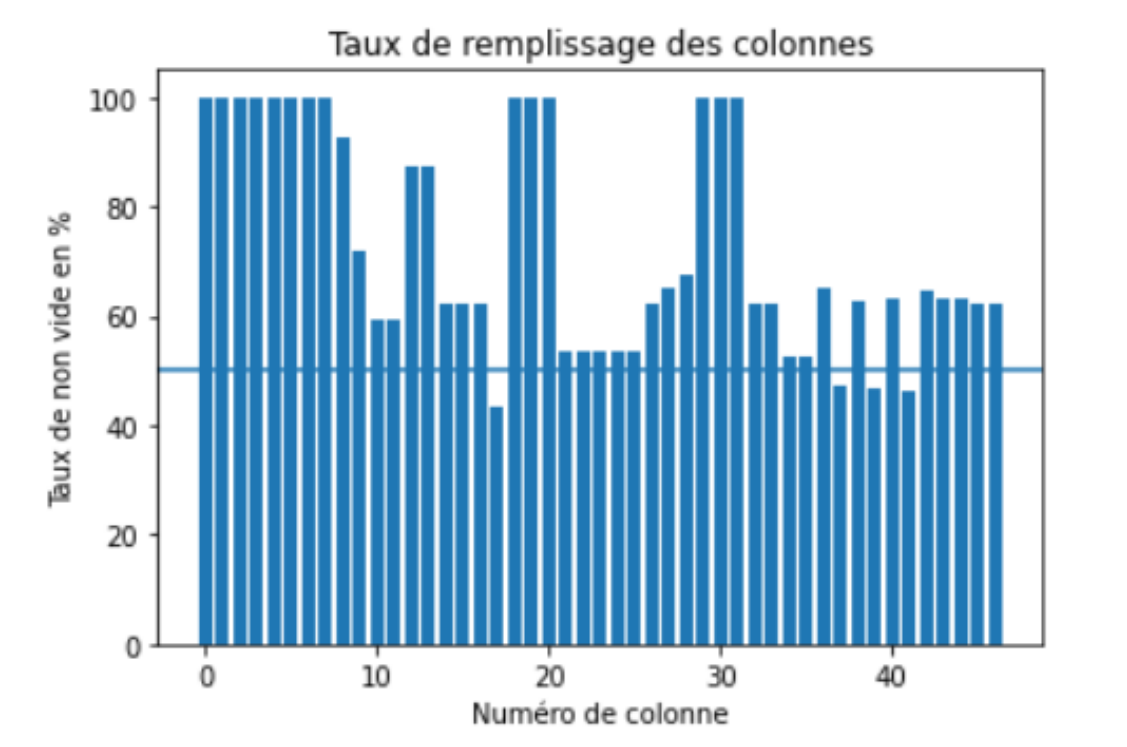
# II. Démarche méthodologique de nettoyage

## 1. Filtrage des features et produits

- Suppression des colonnes inutiles pour l'application

⇨ Première suppression sur le taux de remplissage (plus de 40% de remplie)

⇨ Deuxième suppression sur une logique métier



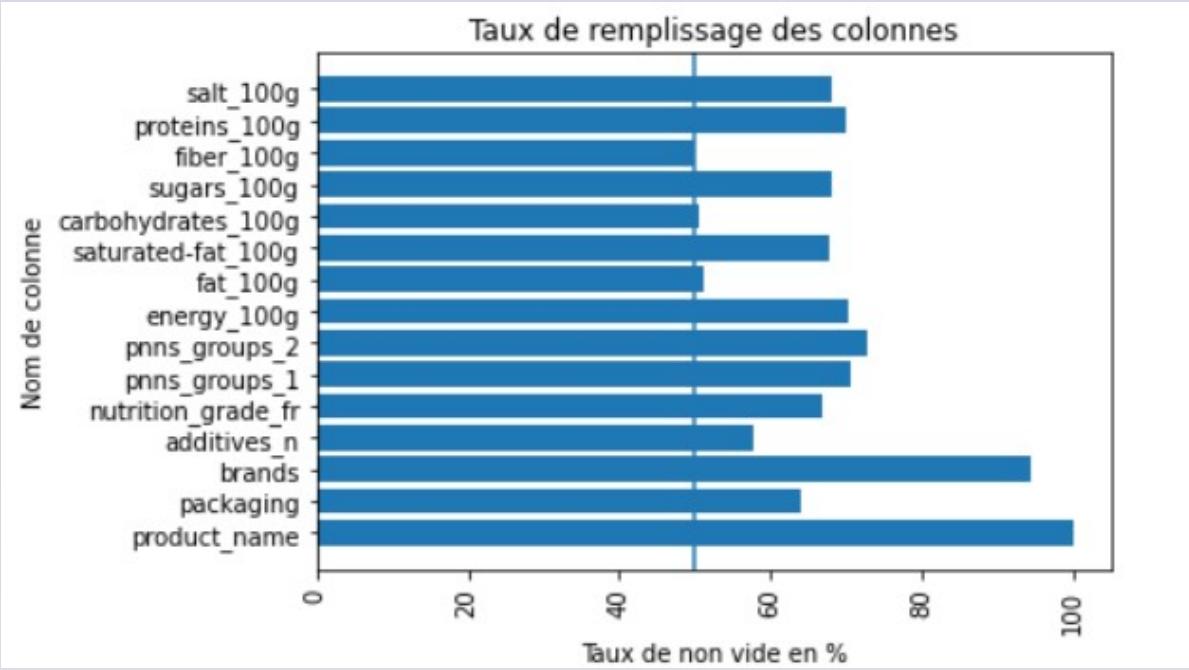
# II. Démarche méthodologique de nettoyage

## 1. Filtrage des features et produits

- Features gardées :

Nom du produit
Packaging
Marque
Nombre d'additifs
Nutriscore
Catégorie n°1
Catégorie n°2
Energie pour 100g

Lipides pour 100g
Acides gras saturé pour 100g
Glucides pour 100g
Sucres pour 100g
Fibres pour 100g
Protéines pour 100g
Sel pour 100g



# II. Démarche méthodologique de nettoyage

## 1. Filtrage des features et produits

- Suppression des produits sans nom
- Suppression des doublons

→ Traitement des valeurs aberrantes

# II. Démarche méthodologique de nettoyage

## 2. Traitement des valeurs aberrantes

### Aperçu global

| | additives_n | energy_100g | fat_100g | saturated-fat_100g | carbohydrates_100g | sugars_100g | fiber_100g | proteins_100g | salt_100g |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 50282.000000 | 6.098000e+04 | 44179.000000 | 58954.000000 | 43767.00000 | 59044.000000 | 43378.000000 | 60731.000000 | 59102.000000 |
| **mean** | 1.858777 | 1.167765e+03 | 13.276896 | 5.387137 | 27.32515 | 13.236850 | 2.525191 | 7.799500 | 1.162860 |
| **std** | 2.567380 | 1.320636e+04 | 16.981752 | 8.540098 | 27.31311 | 19.024517 | 4.643673 | 7.933839 | 4.303963 |
| **min** | 0.000000 | 0.000000e+00 | 0.000000 | 0.000000 | 0.00000 | -0.100000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 0.000000 | 4.270000e+02 | 1.300000 | 0.300000 | 4.00000 | 1.000000 | 0.000000 | 1.800000 | 0.080000 |
| **50%** | 1.000000 | 1.029000e+03 | 6.800000 | 1.900000 | 14.00000 | 4.000000 | 1.300000 | 6.000000 | 0.570000 |
| **75%** | 3.000000 | 1.639000e+03 | 21.000000 | 7.300000 | 52.70000 | 17.000000 | 3.200000 | 11.000000 | 1.250000 |
| **max** | 31.000000 | 3.251373e+06 | 380.000000 | 210.000000 | 190.00000 | 105.000000 | 178.000000 | 100.000000 | 211.000000 |

# II. Démarche méthodologique de nettoyage

## 2. Traitement des valeurs aberrantes

- Traitement global : remplacement des valeurs aberrantes par des NaN

- Traitement métier : Recherche des informations sur chaque feature sur le net et sélection de la valeur approprié

# II. Démarche méthodologique de nettoyage

## 2. Traitement des valeurs aberrantes

- Traitement des lignes aberrantes
    - Sommes des features numériques supérieur à 100g (sauf énergie et additifs)
        - Suppression
    - Sucres supérieurs à glucides
        - NaN à la place des glucides
    - Acides gras saturé supérieur à lipides
        - NaN à la place des lipides

# II. Démarche méthodologique de nettoyage

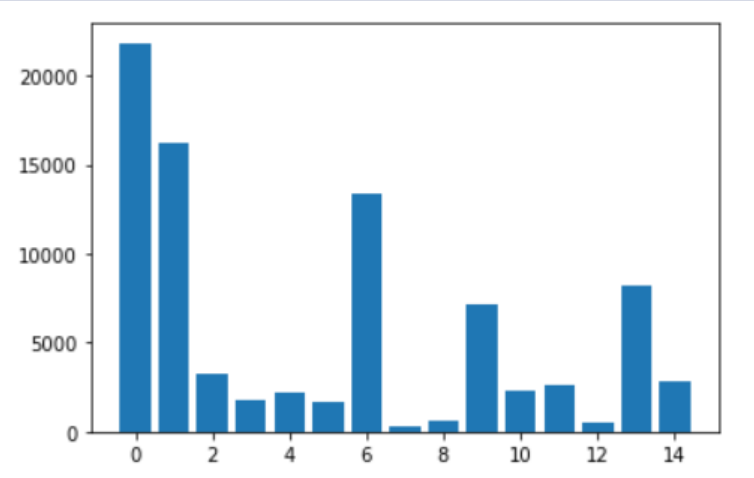
## 3. Traitement des valeurs manquantes

- Suppression de lignes
- Remplissage par une approche métier
- Remplissage par la moyenne de chaque catégorie de produit
- Remplissage par ImputeIterative

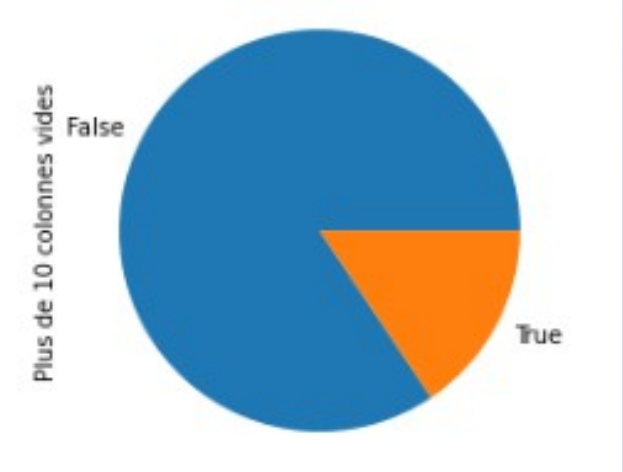
# II. Démarche méthodologique de nettoyage

## 3. Traitement des valeurs manquantes

- Nombre de NaN par lignes




- Proportion de lignes avec plus de 10 colonnes vides




⇨ Suppression de lignes avec une quantité trop important de NaN par lignes

# II. Démarche méthodologique de nettoyage

## 3. Traitement des valeurs manquantes

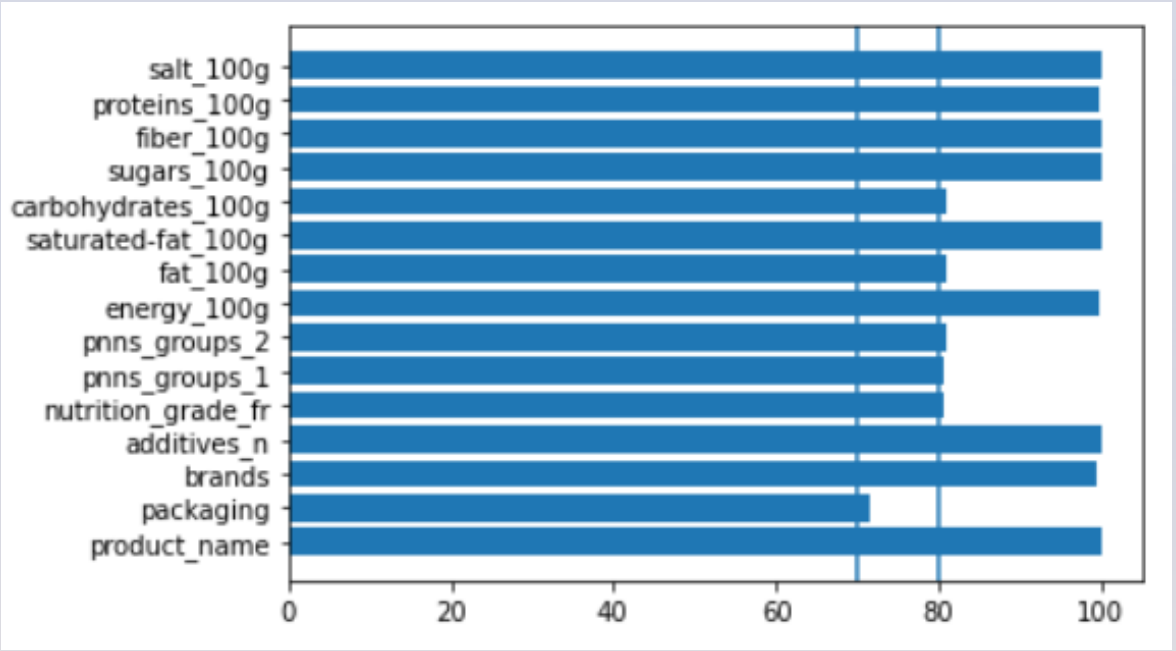
- Remplissage par une approche métier
  - Remplissage par la moyenne sur les additifs
  - Remplissage par des 0 pour les fibres

# II. Démarche méthodologique de nettoyage

## 3. Traitement des valeurs manquantes

- Remplissage par la moyenne de chaque catégorie de produit

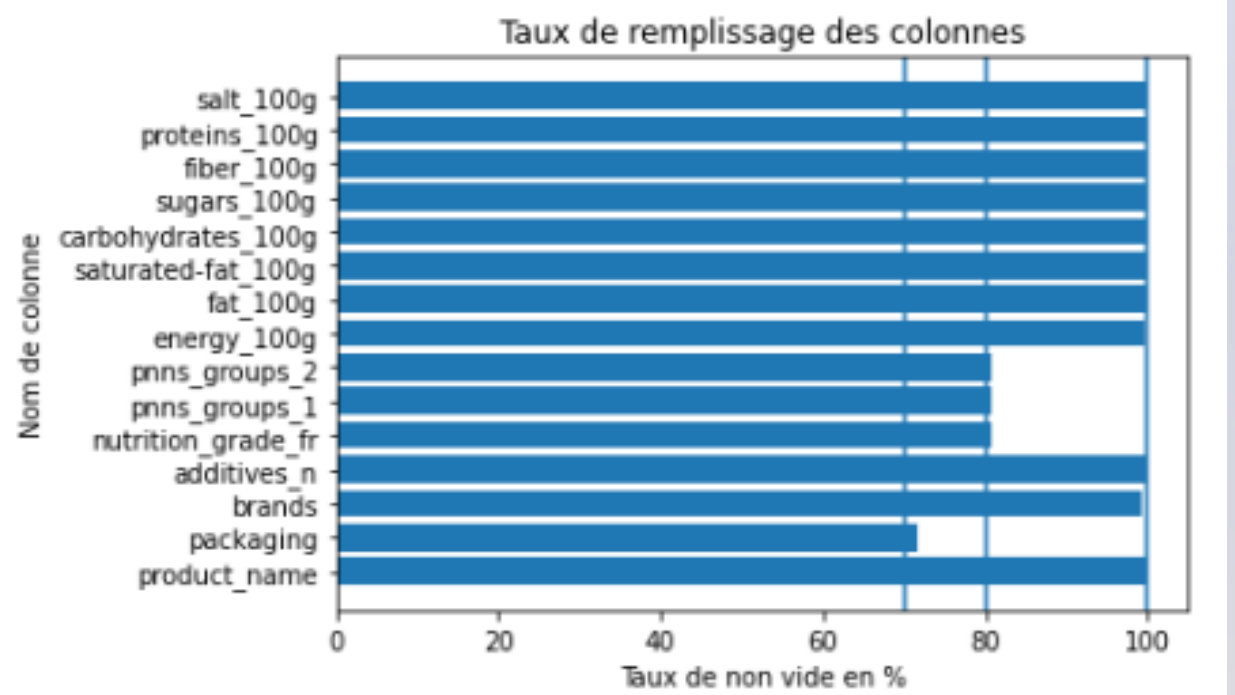
  - Ecriture d'une fonction qui remplis les variables par leur médiane par catégorie
  - Utilisation de la fonction sur les features quantitatives



# II. Démarche méthodologique de nettoyage

## 3. Traitement des valeurs manquantes

- Remplissage par ImputeIterative
  - Remplissage du reste des features
  - Danger :
    Recrée des valeurs aberrantes sur la somme totale et sur les maximums



# III. Démarche méthodologique d’exploration

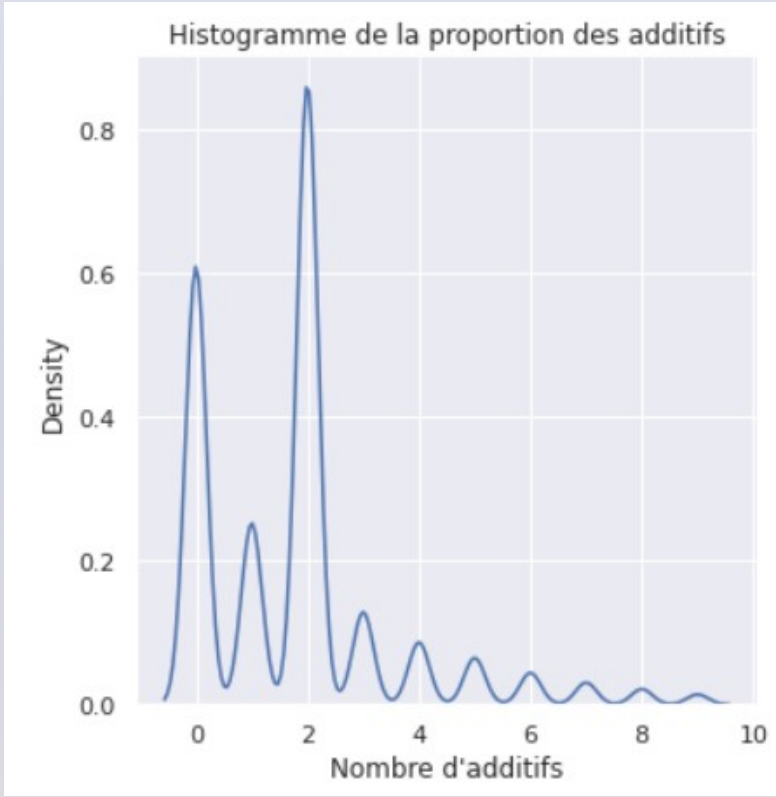
- **Les grandes étapes**

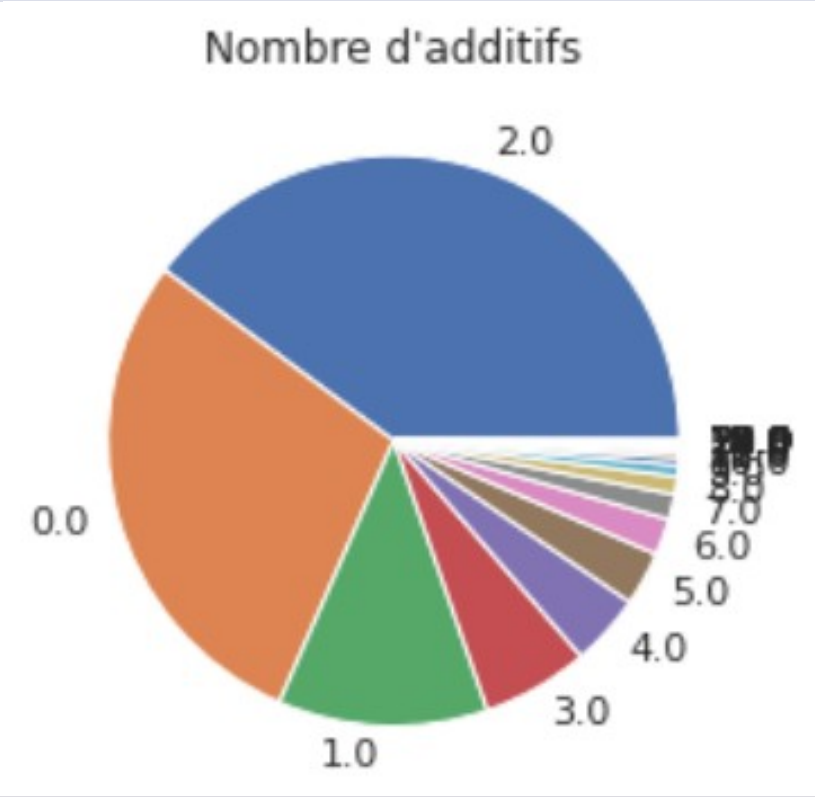
- Analyse univariée
- Analyse bivariée
- Analyse multivariée

# III. Démarche méthodologique d'exploration

## Analyse univariée

- Additifs

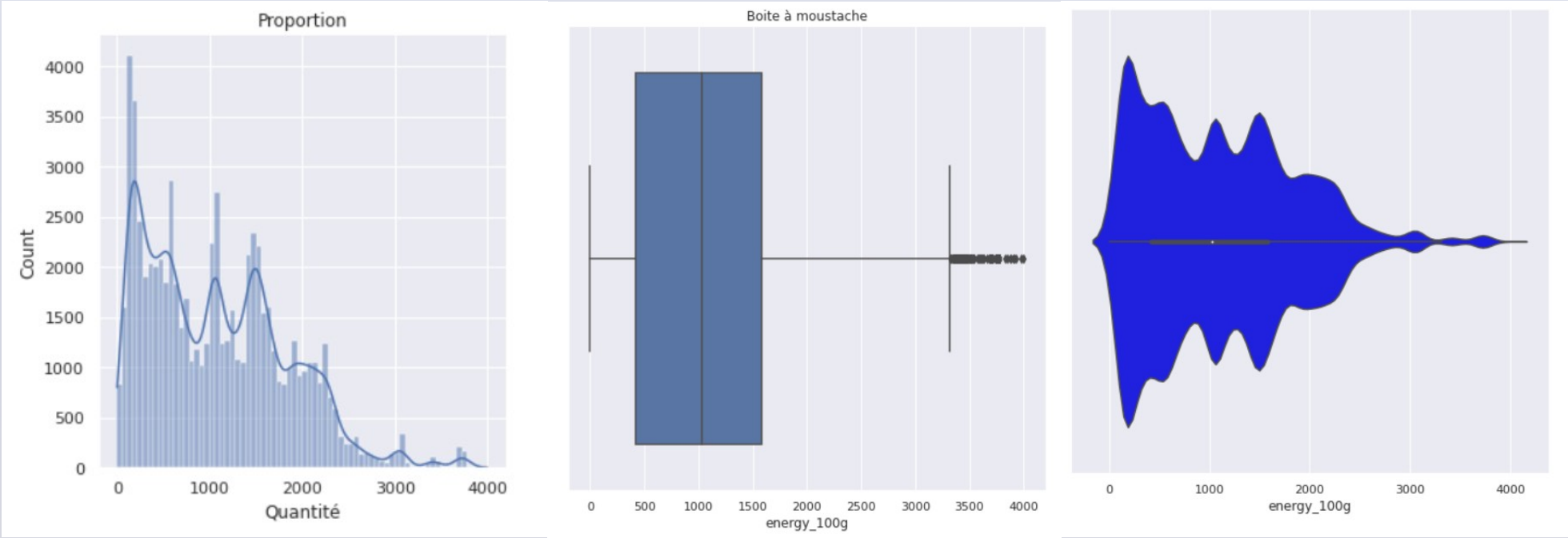


# III. Démarche méthodologique d'exploration

## Analyse univariée

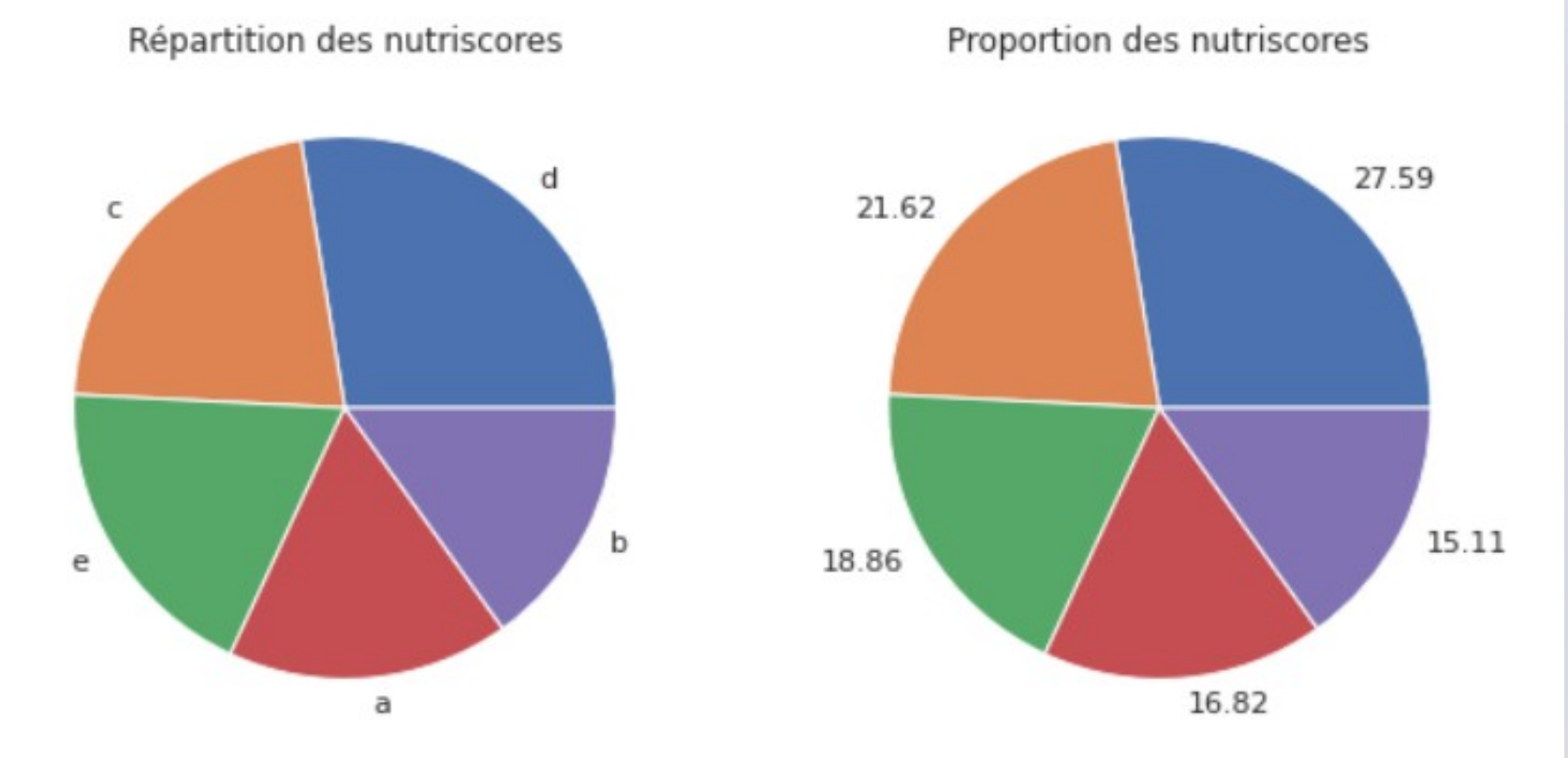
- Energie



# III. Démarche méthodologique d'exploration

## Analyse univariée

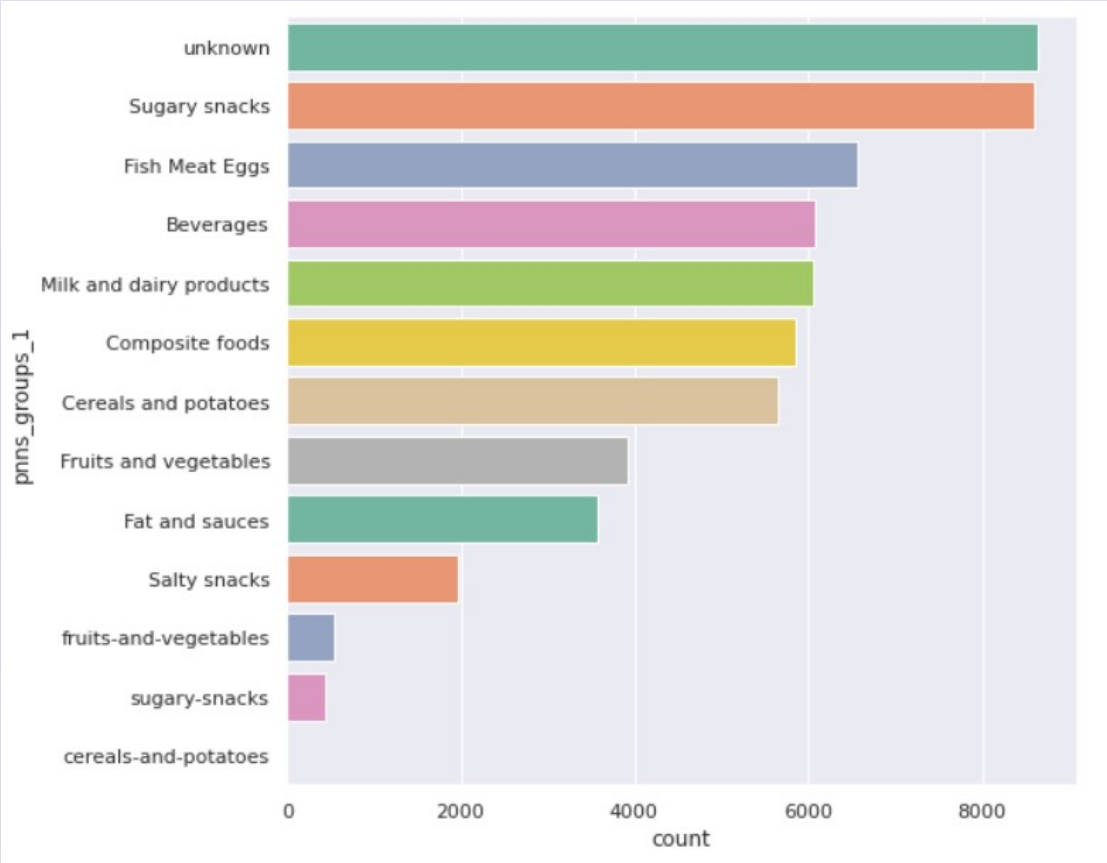
- Nutriscore



# III. Démarche méthodologique d'exploration

## Analyse unvariée

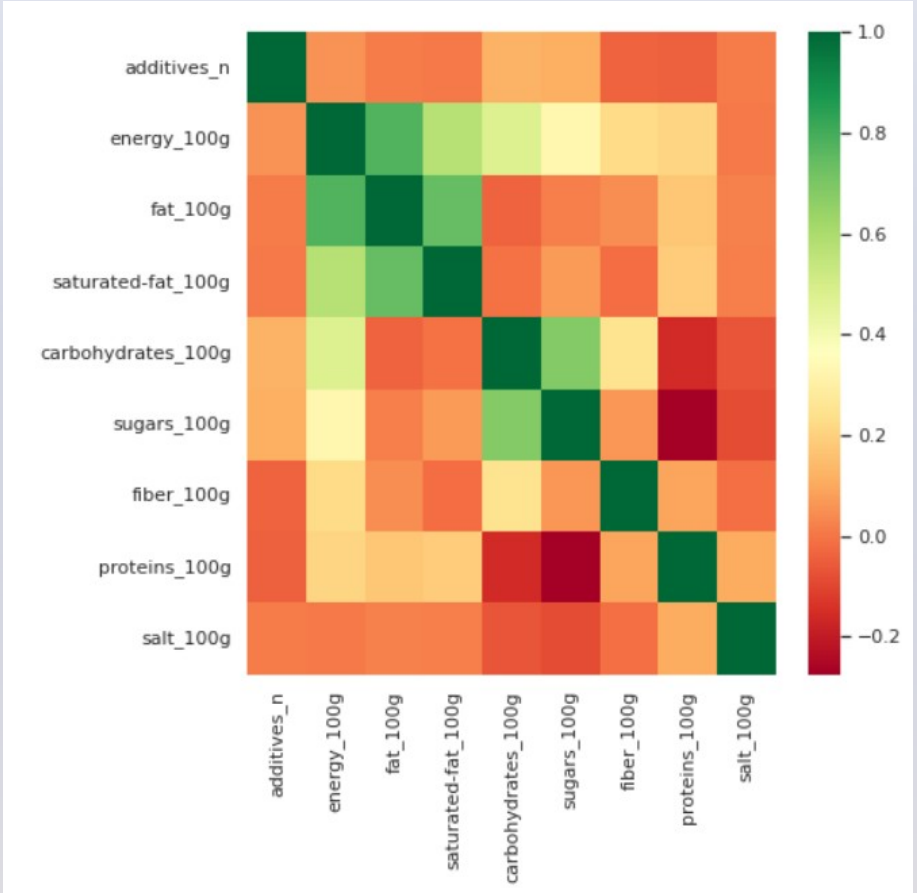
- Catégorie



# III. Démarche méthodologique d'exploration

## Analyse bivariée

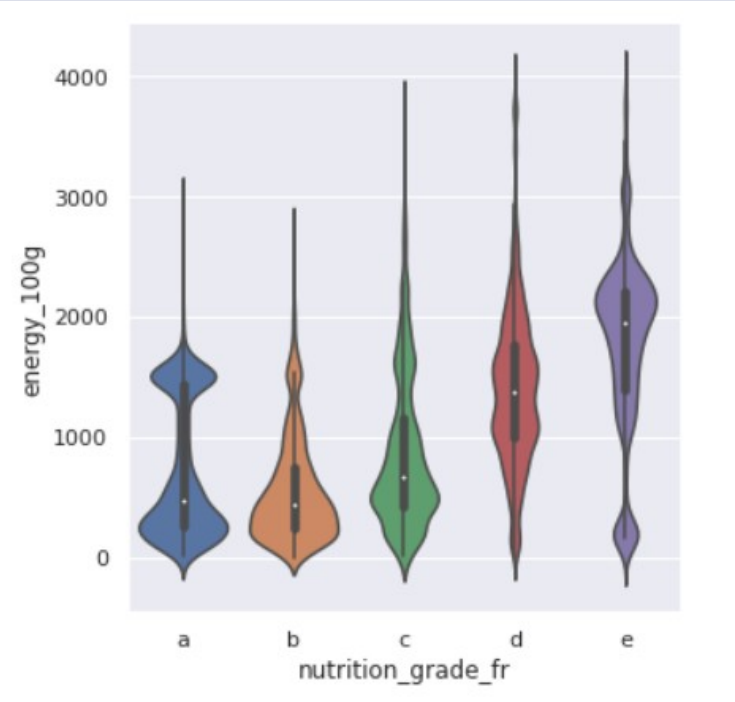
- Heatmap de corrélation

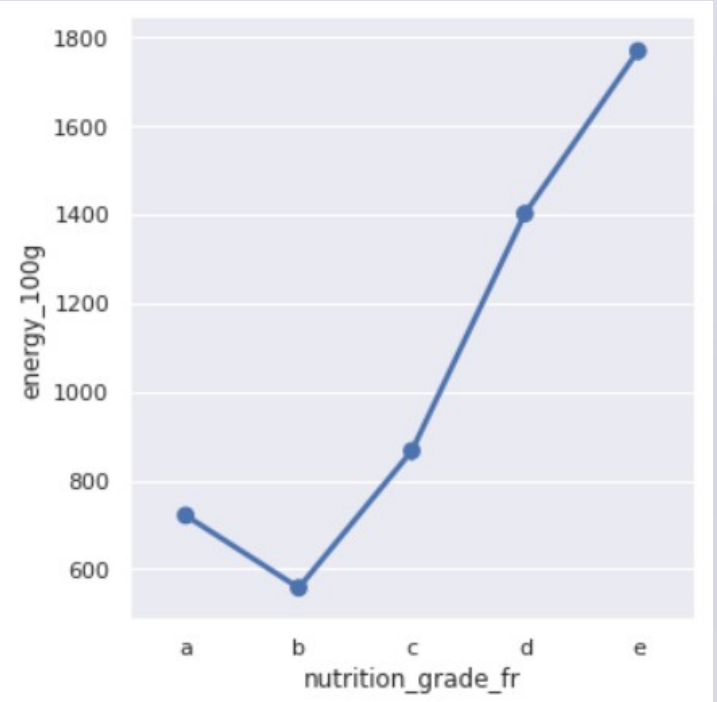


# III. Démarche méthodologique d'exploration

## Analyse bivariée

- Energie/Nutriscore

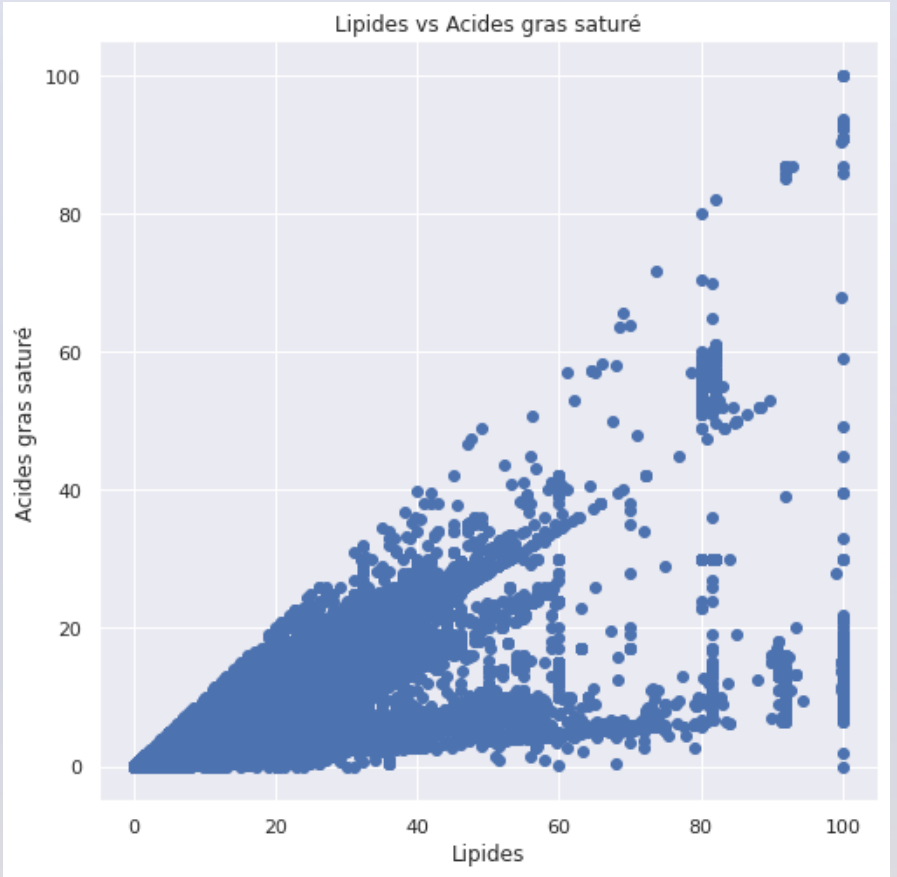


# III. Démarche méthodologique d'exploration

## Analyse bivariée

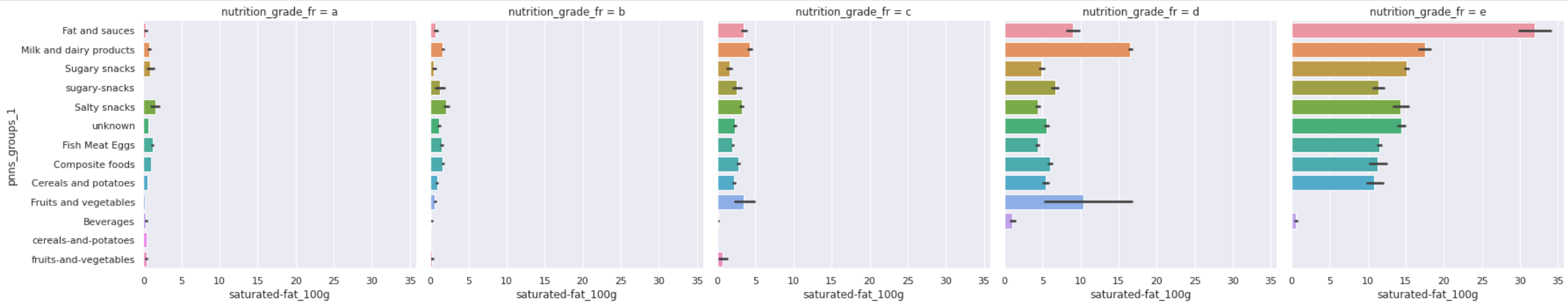
- Lipides/Acides gras saturé



# III. Démarche méthodologique d'exploration

## Analyse multivariée

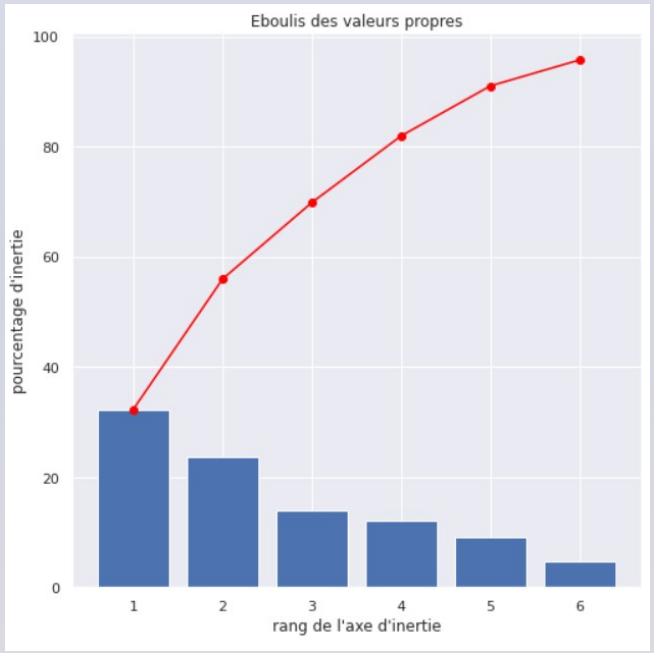
- Acides gras saturé/Catégorie/Nutriscore

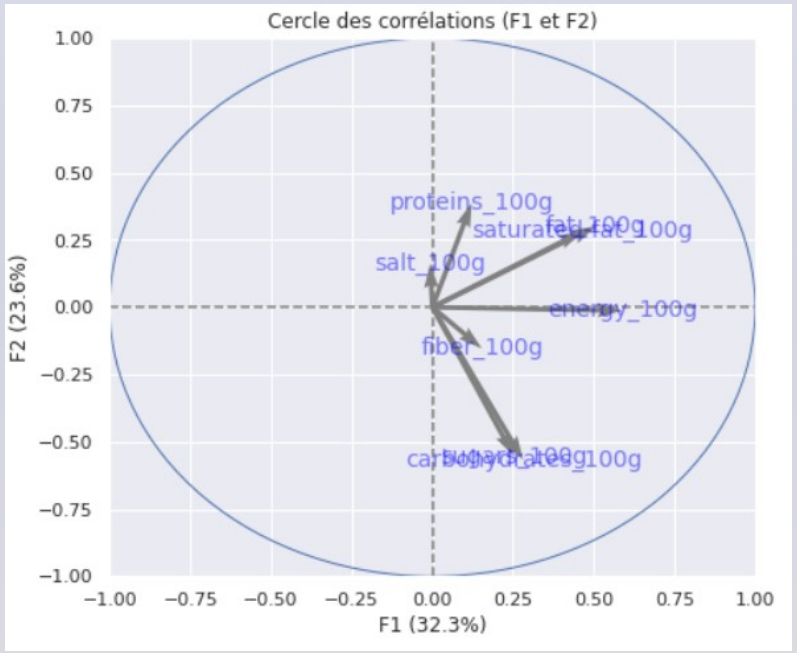


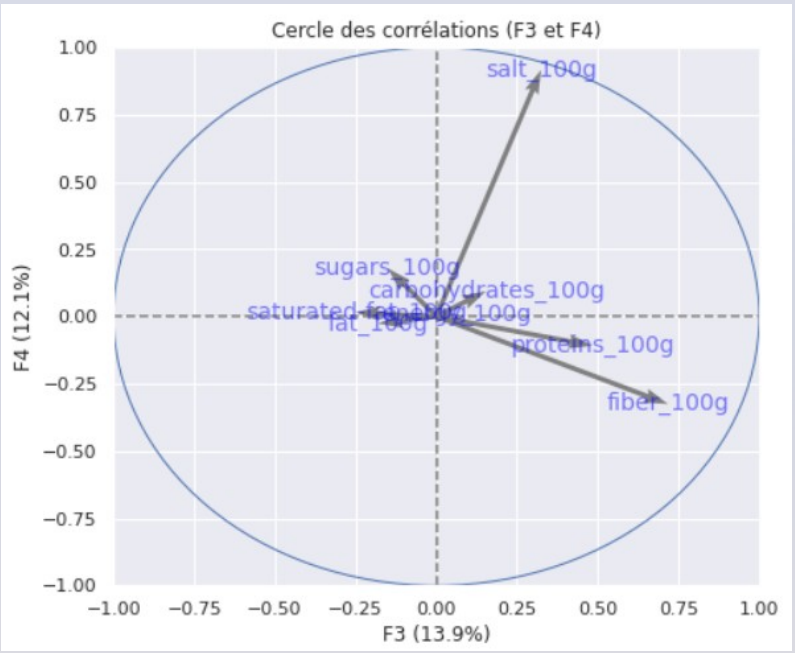
# III. Démarche méthodologique d'exploration

## Analyse multivariée

- Analyse en composante principale



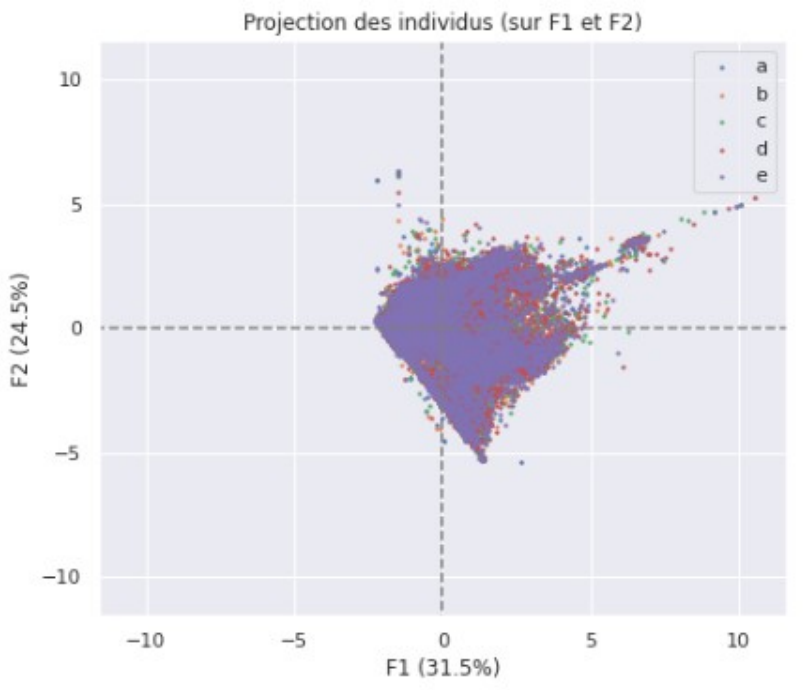



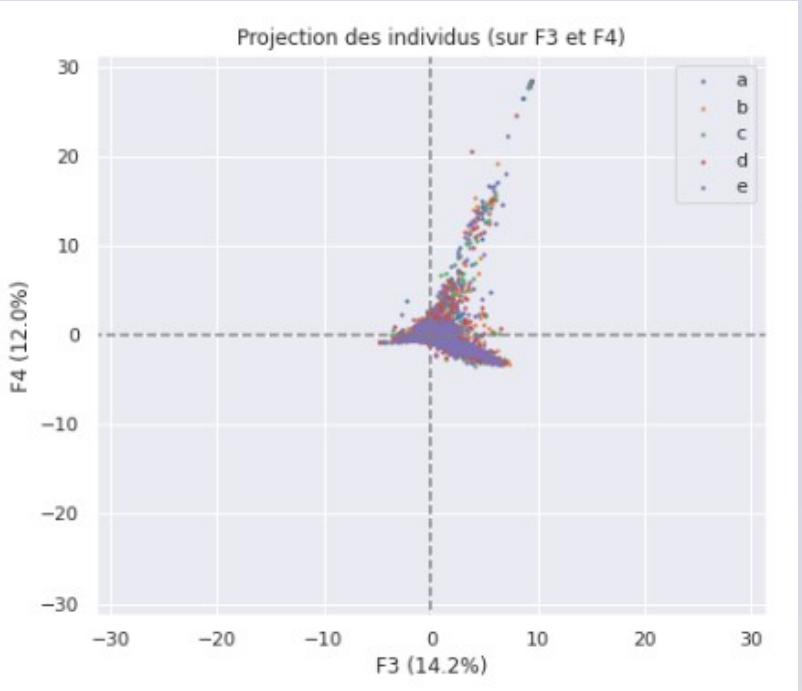



# III. Démarche méthodologique d'exploration

## Analyse multivariée

- Analyse en composante principale

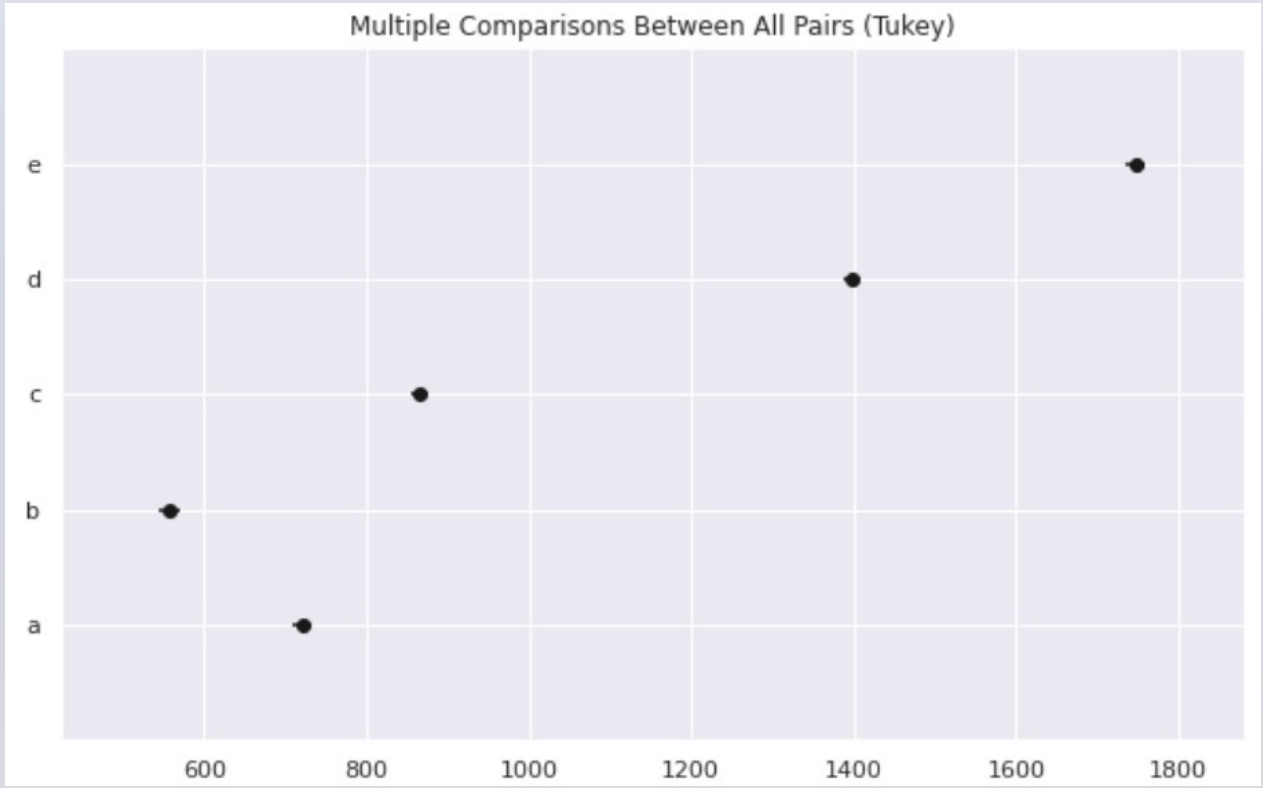


# III. Démarche méthodologique d'exploration

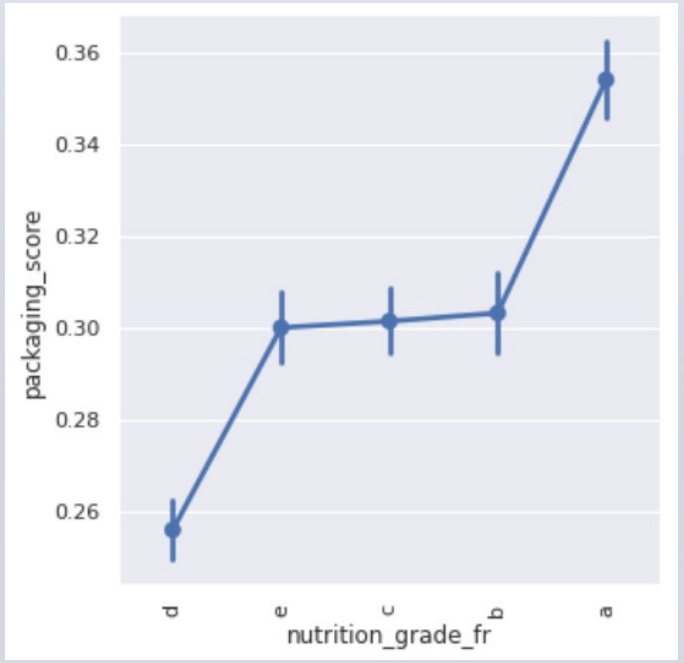
## Analyse multivariée

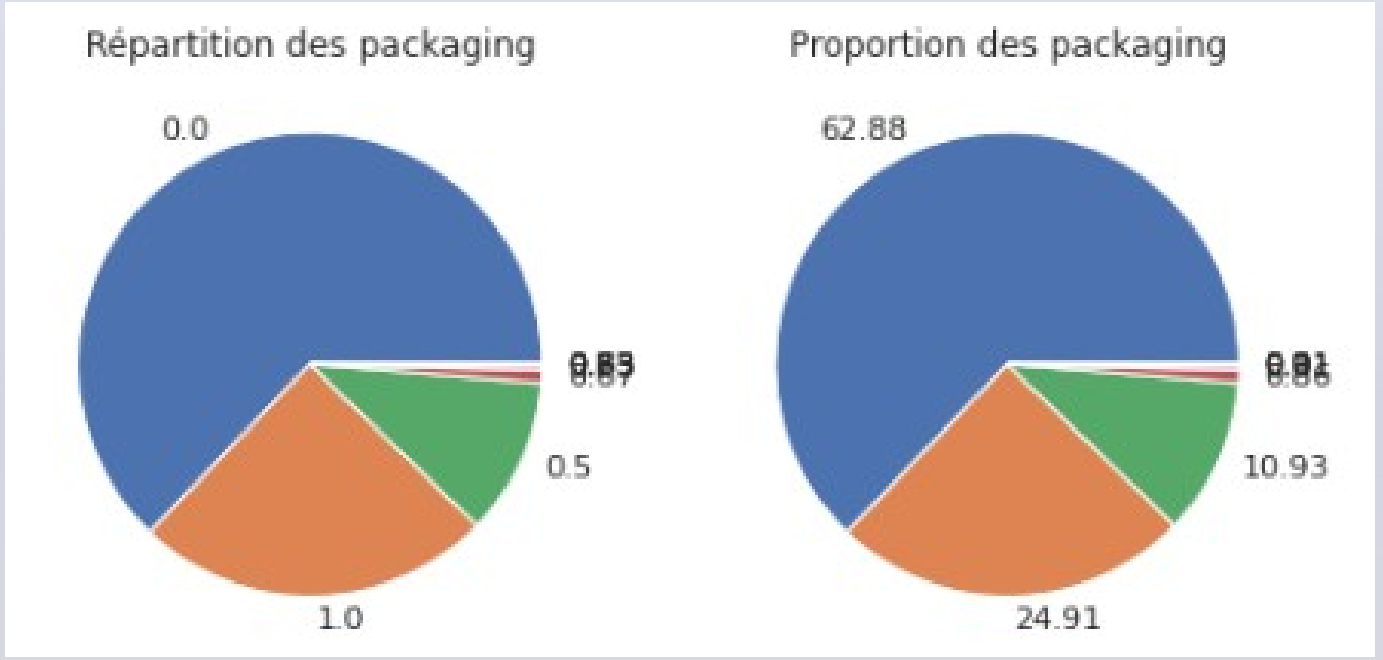
- ANOVA



# IV. Présentation des faits pertinents pour l'application

- Création d'un packaging score



# Conclusion

- L’application semble bien réalisable
- Les données ont été nettoyées
- Les données ont été analysés