Projet 3 : Préparez des données pour un organisme de santé publique

Maodo FALL

OpenClassrooms

Soutenance du projet

14 Mai 2024

Sommaire

Sommaire

Jeu de données Open Food Facts

Le jeu de données est disponible sur le site officiel et les variables sont documentées sur cette adresse. On a 320772 produits avec 162 colonnes.
But : analyse descriptive des données de nutrition pour Santé Publique France. Nous exploitons principalement les variables :

- energy_100g
- fat_100g
- carbohydrates_100g
- sugars_100g
- fiber_100g
- proteins_100g
- salt_100g
- sodium_100g

Santé Publique France

Premiers nettoyages

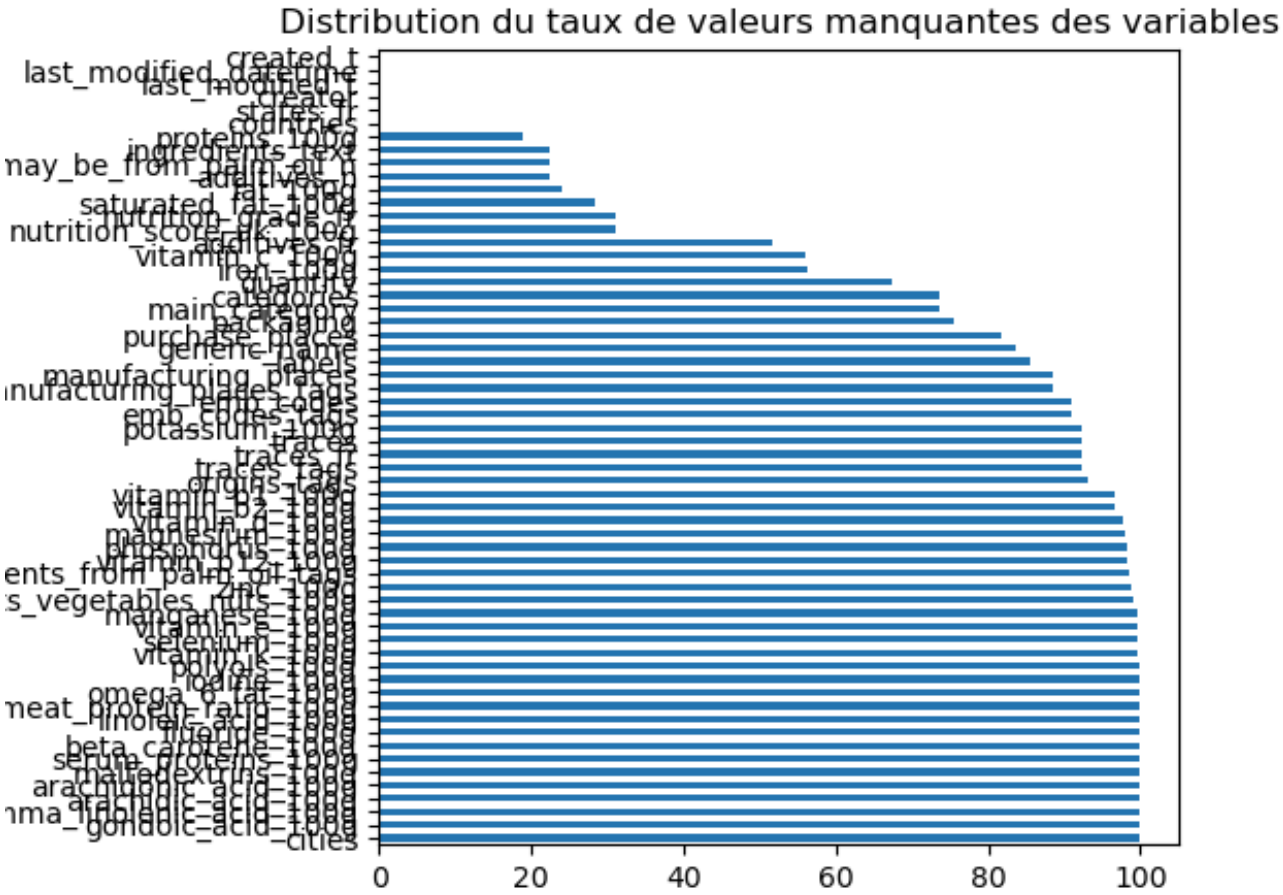
Nous avons procédé aux nettoyages suivants :

- Enlèvement des colonnes vides s'il en existe.
- Enlèvement des duplications de produits par le code barre unique.
- Correction du nom des variables mal orthographiées.
- Enlèvement des variables non renseignées à plus de 75%.

Taux de valeurs manquantes

colonnes	NA_rate
sucrose_100g	99.977554
chlorophyl_100g	100.000000
nutrition-score-fr_100g	31.038245
code	0.007170
emb_codes	90.863916
proteins_100g	18.969860
serum-proteins_100g	99.995012
labels_tags	85.458831
countries	0.087289
nutrition_grade_uk	100.000000

Taux de valeurs manquantes



Sommaire

Critères de sélection

Les variables sont choisies selon les critères suivants :

- Les données nutritionnelles :
  - macronutriments : protéines, glucides et lipides.
  - micronutriments : vitamines et minéraux.
- Les détails sur chaque produit :
  - product_name : Nom du produit
  - pnns_groups_1 et pnns_groups_2 : groupes et sous-groupes d'aliments.
- La grade et le score de nutrition.

Imputation des valeurs manquantes

Les variables de nutrition n'étant pas toutes renseignées sur tous les produits, nous avons effectué une imputation des valeurs manquantes par la méthode des plus proches voisins :

Column	Non-Null Count
product_name	60658 non-null
brands	60658 non-null
pnns_groups_1	60658 non-null
pnns_groups_2	60658 non-null
countries	60605 non-null
countries_fr	60605 non-null
nutrition_grade_fr	60658 non-null
nutrition_score_fr_100g	60658 non-null
energy_100g	60658 non-null
fat_100g	60658 non-null
carbohydrates_100g	60658 non-null
sugars_100g	60658 non-null
fiber_100g	60658 non-null
proteins_100g	60658 non-null
salt_100g	60658 non-null
sodium_100g	60658 non-null

Traitement des valeurs aberrantes - Etapes

- Suppression des valeurs atypiques de score de nutrition par grade.
- Limiter les valeurs des variables de nutrition au seuil de l'écart interquartile(IQR).
- Enlever les valeurs des variables de nutrition qui dépassent 100g.

Jeu de données traité

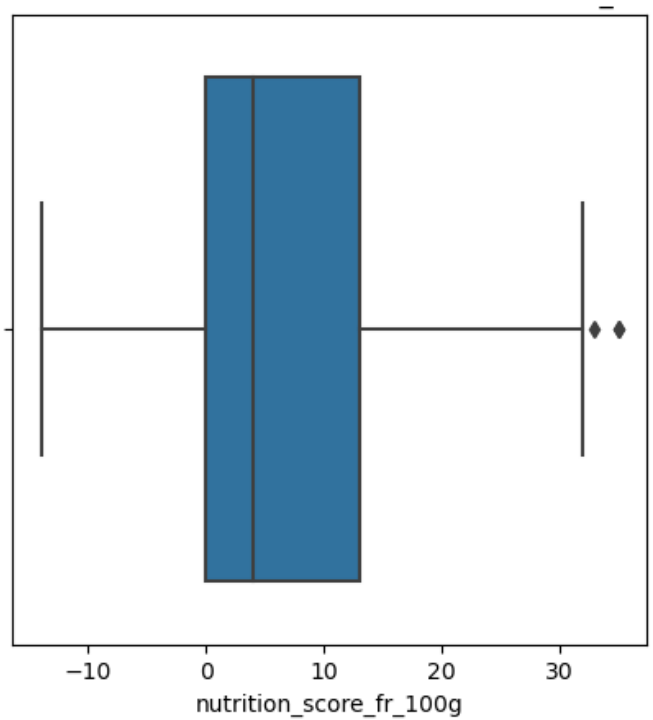
Le jeu de données final que l'on exploite compte 43159 produits uniques :

```
Int64Index: 43159 entries, 174 to 320751
Data columns (total 16 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   product_name            43159 non-null  object
 1   brands                  43159 non-null  object
 2   pnns_groups_1           43159 non-null  object
 3   pnns_groups_2           43159 non-null  object
 4   countries               43125 non-null  object
 5   countries_fr            43125 non-null  object
 6   nutrition_grade_fr      43159 non-null  object
 7   nutrition_score_fr_100g 43159 non-null  float64
 8   energy_100g             43159 non-null  float64
 9   fat_100g                43159 non-null  float64
 10  carbohydrates_100g      43159 non-null  float64
 11  sugars_100g             43159 non-null  float64
 12  fiber_100g              43159 non-null  float64
 13  proteins_100g           43159 non-null  float64
 14  salt_100g               43159 non-null  float64
 15  sodium_100g             43159 non-null  float64
```

Sommaire

Boite à moustache du score de nutrition

```
Description et analyse univariée
mean : 6.567
std : 8.749
median : 4.0
quartile q1: 0.0
quartile q3 : 13.0
min : -14.0
max : 35.0
```



Distribution de la variable protéine

```
mean : 6.959
std : 5.609
median : 6.0
quartile q1: 2.7
quartile q3 : 9.8
min : 0.0
max : 23.21
```

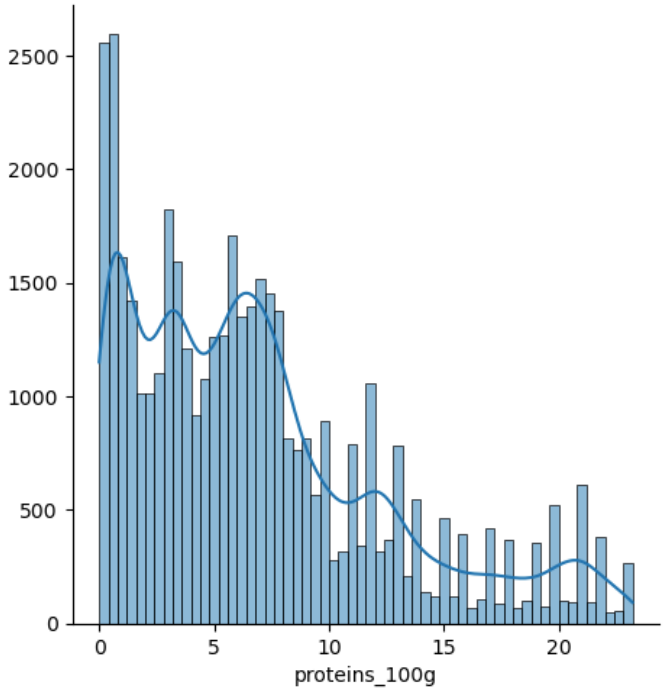
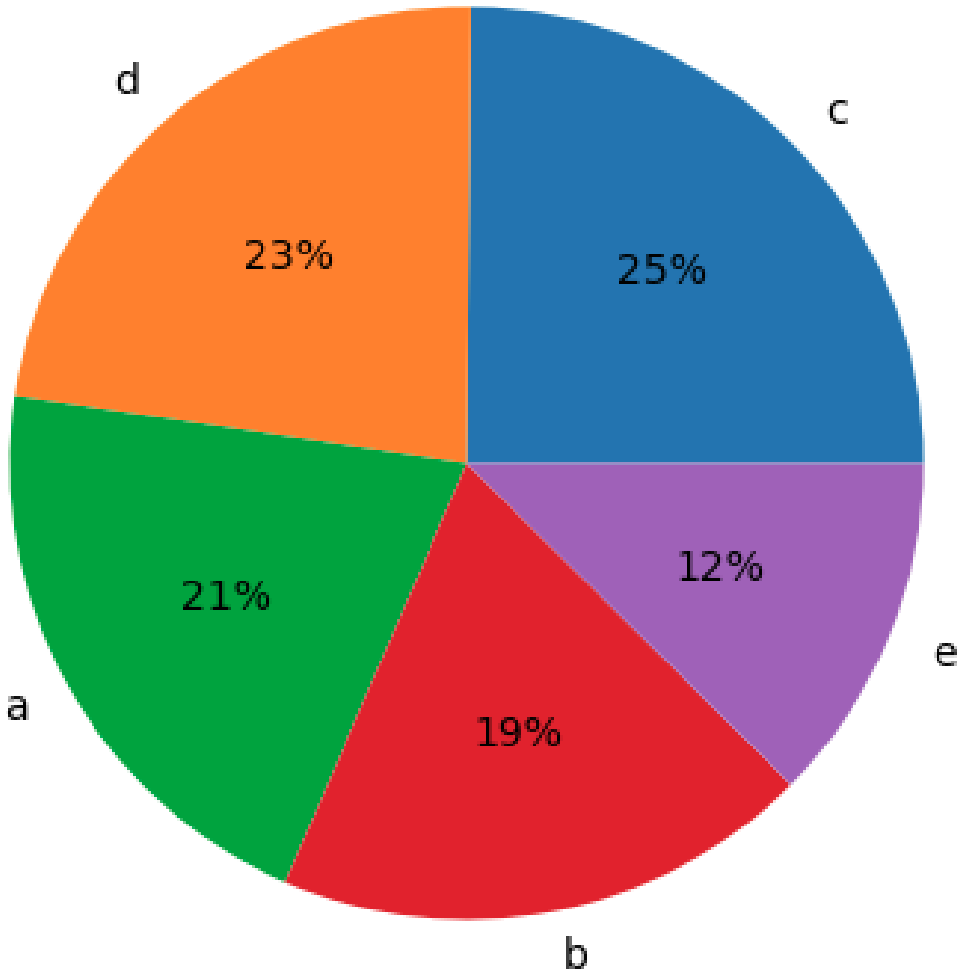
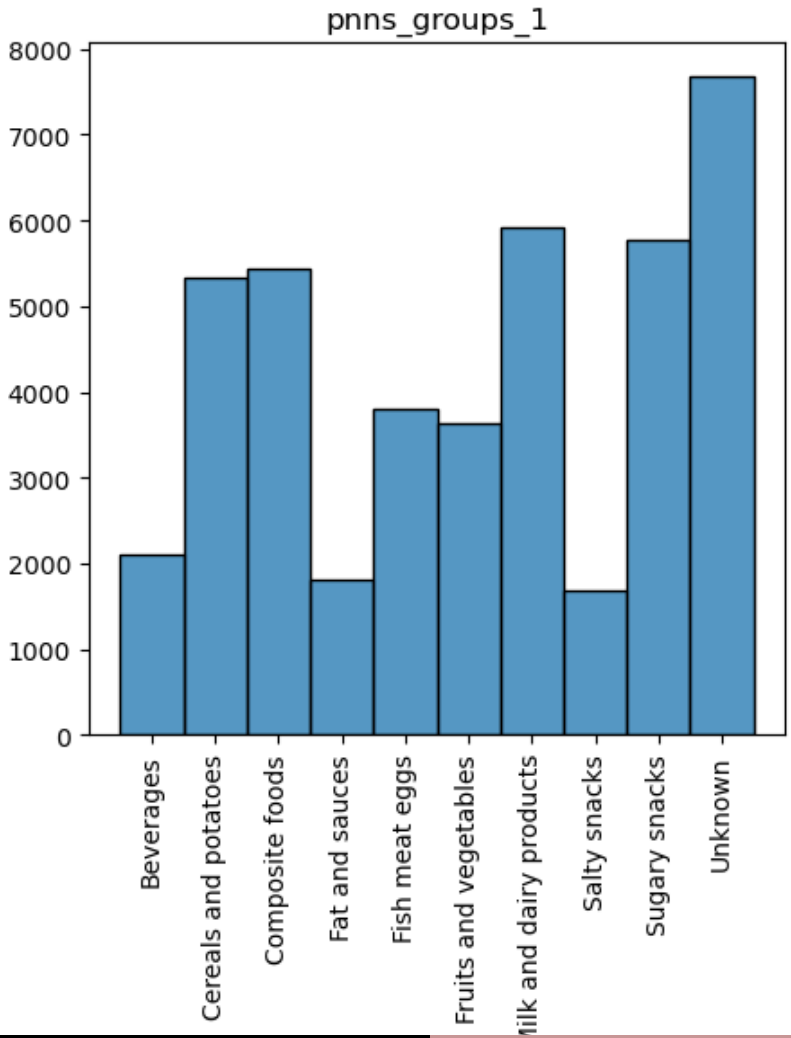


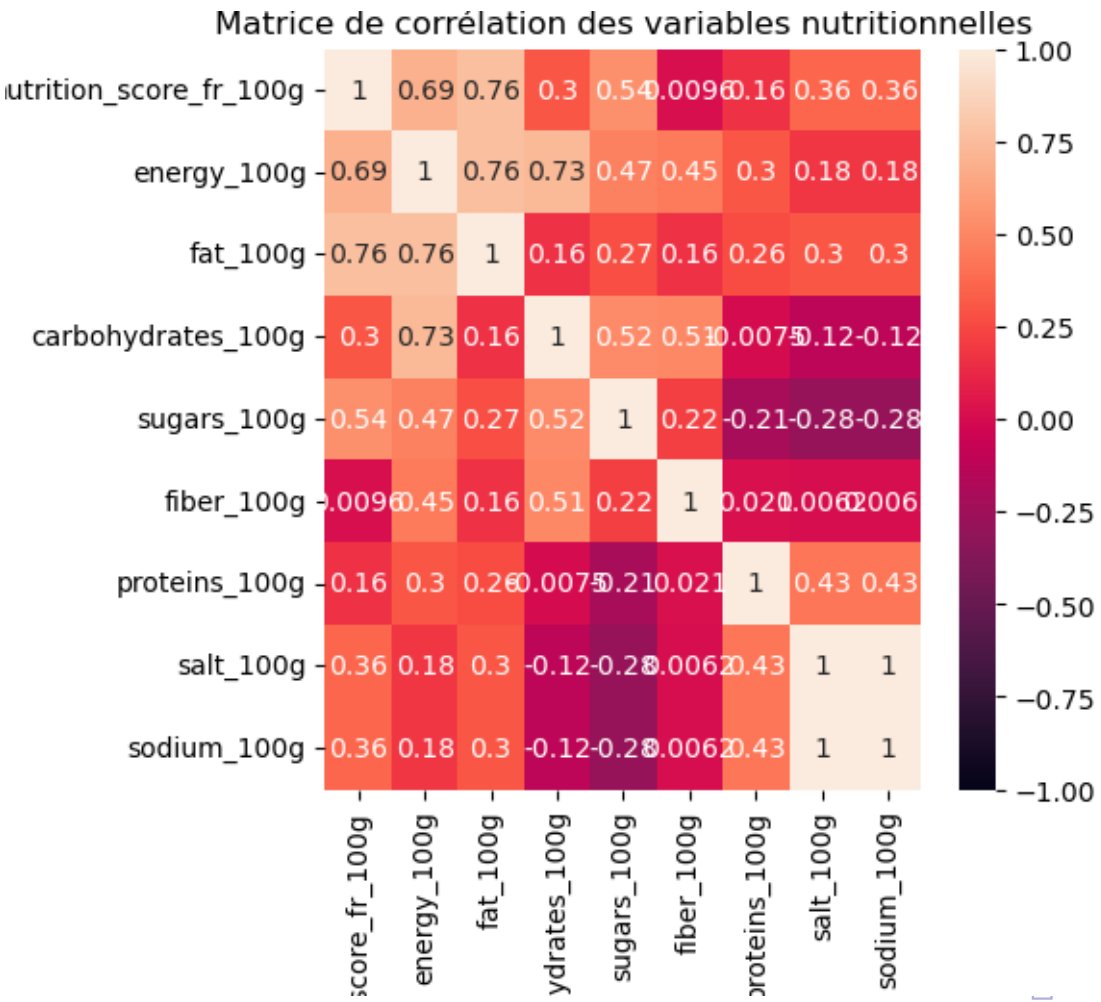
Diagramme en secteurs du grade de nutrition



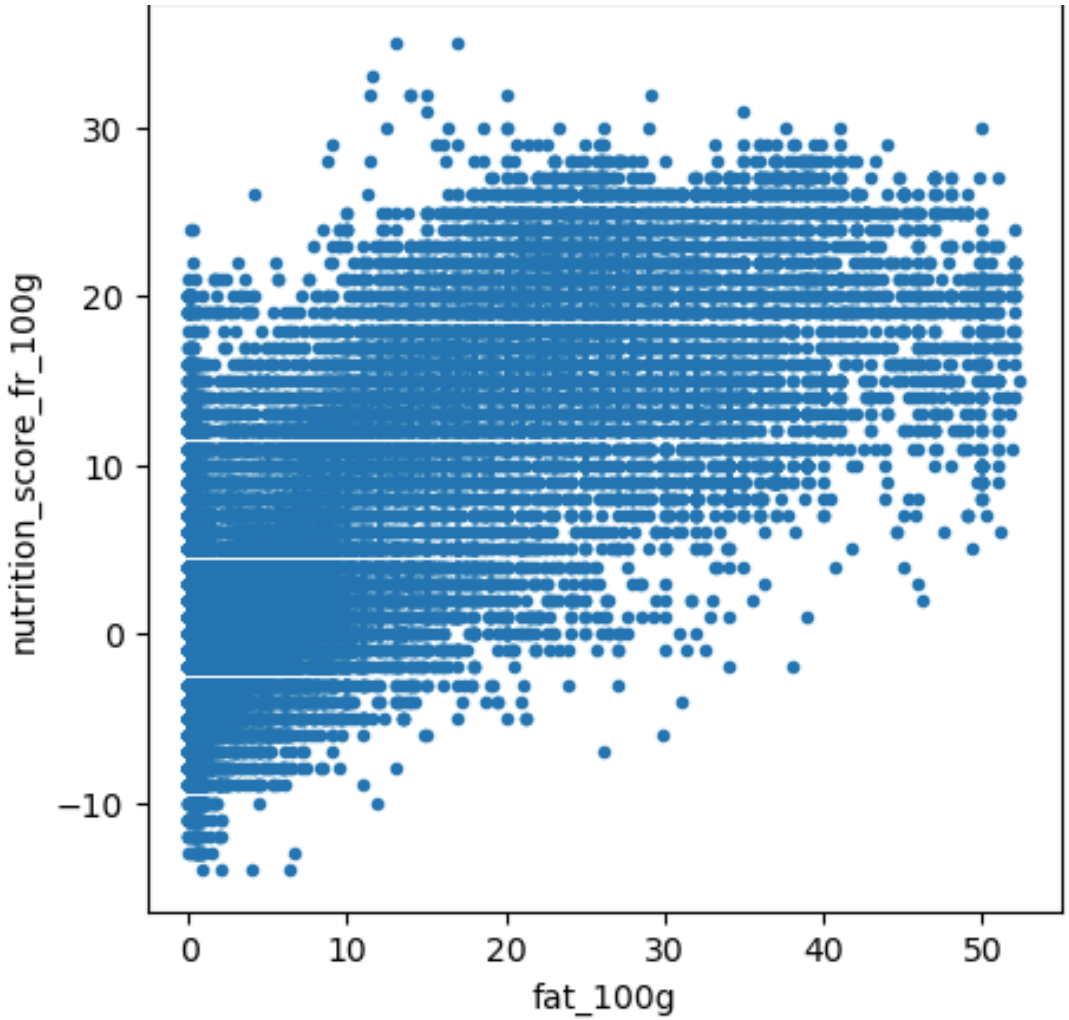
Répartition des groupes d'aliment



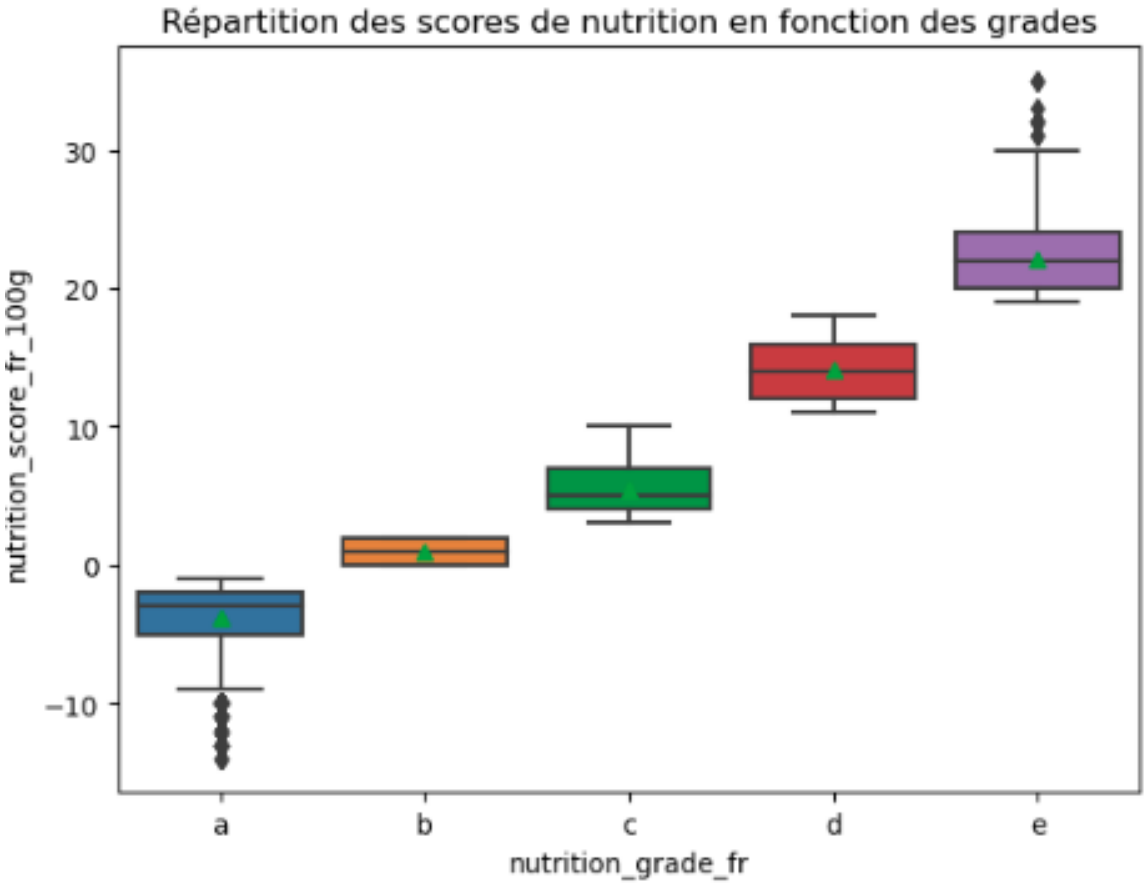
Matrice de corrélation



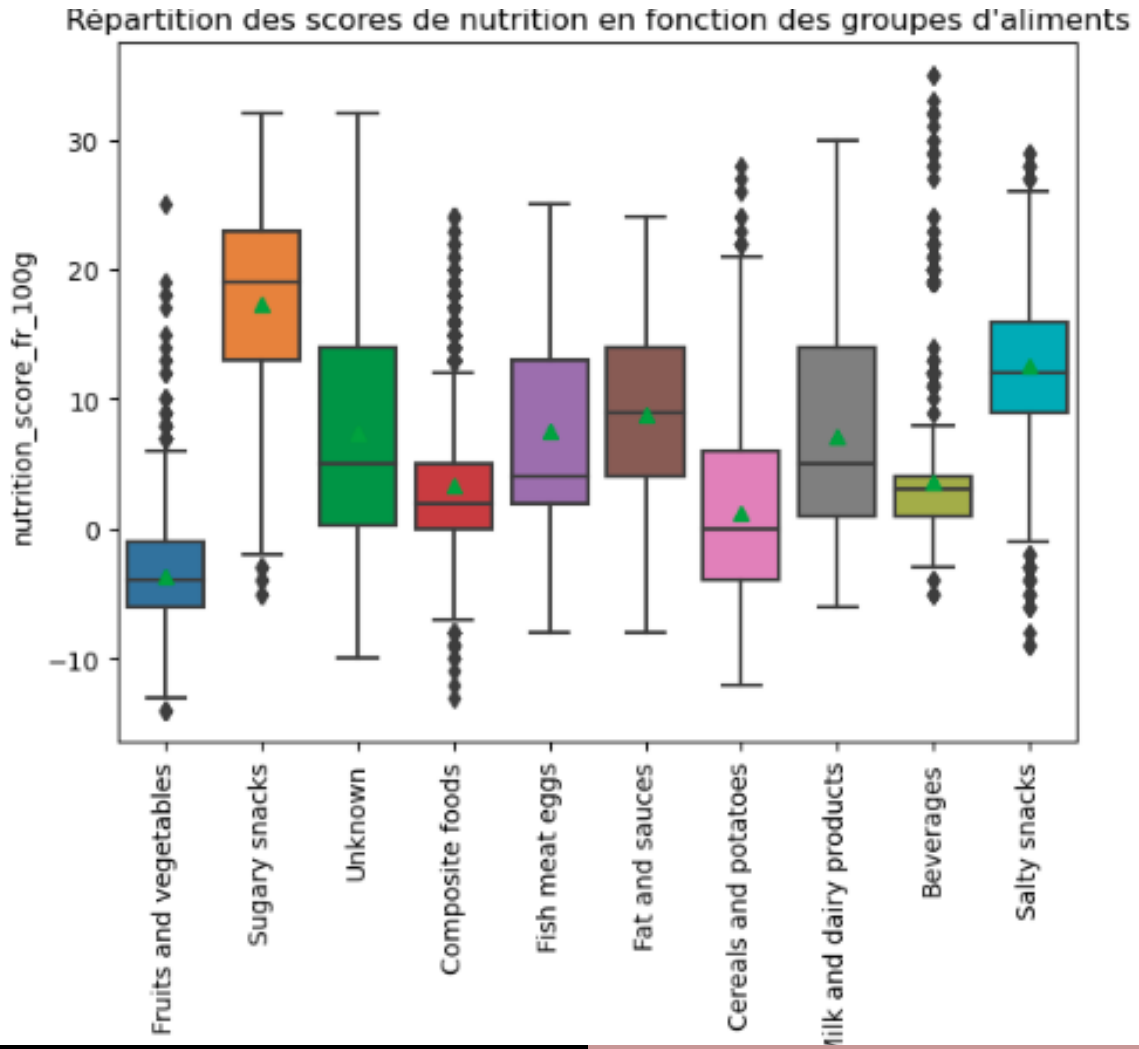
Lien avec le score de nutrition - un exemple



Score de nutrition par grade



Score de nutrition par groupe d'aliment



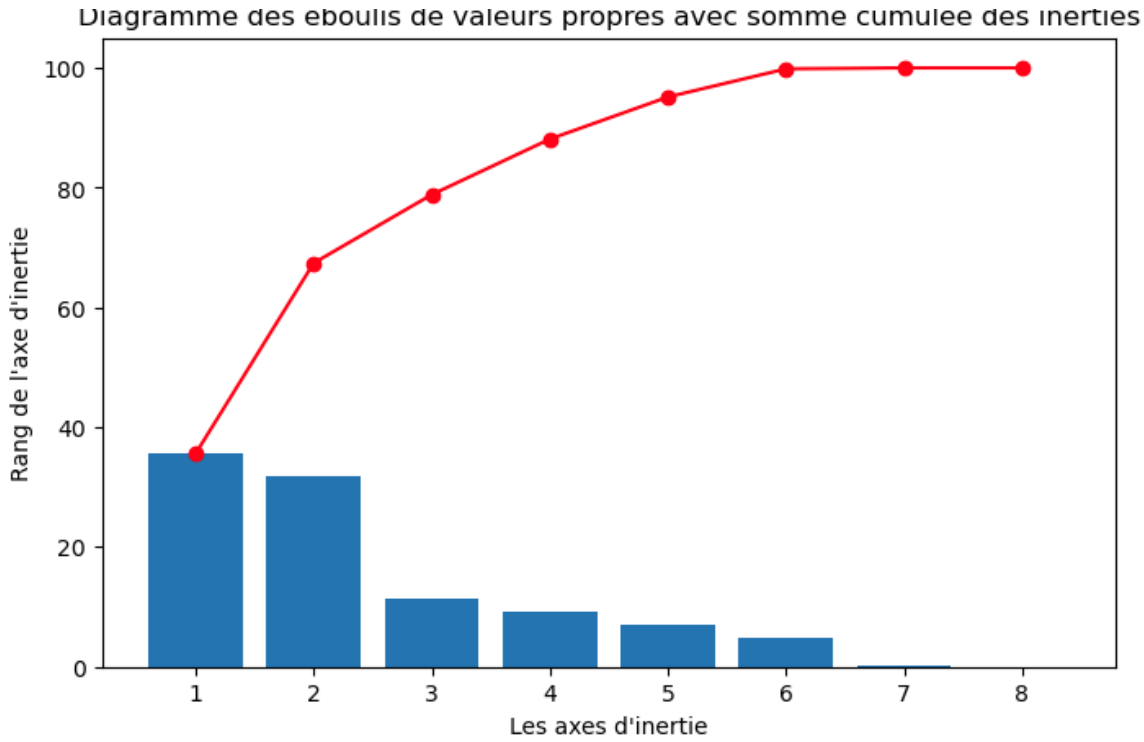
Sommaire

ANOVA

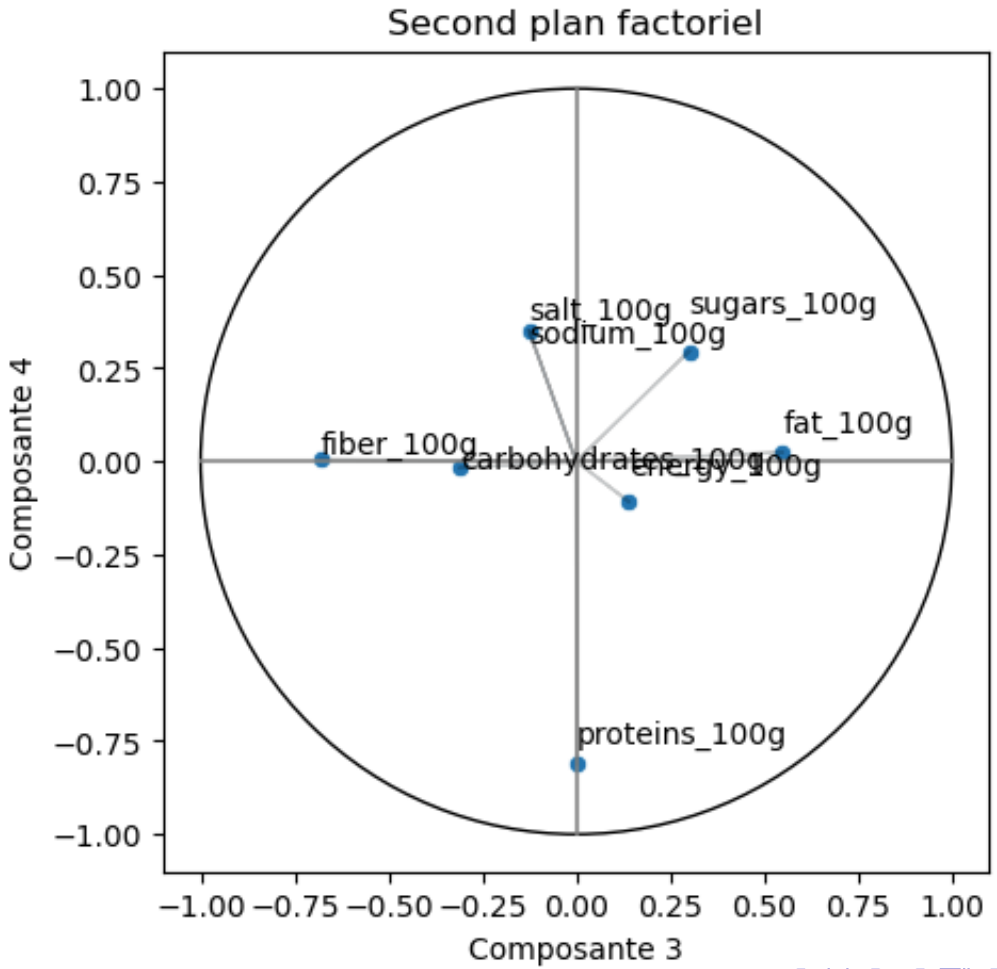
- Le test d'analyse de variance réalisé entre le score de nutrition et le groupe d'aliment nous a confirmé que les écarts des scores de nutrition par groupe d'aliment observés ci-dessus sont significatifs. Autrement dit, la groupe d'aliment a un impacte sur le score de nutrition.
- Nous avons les mêmes conclusions concernant les grades de nutrition.

Sommaire

Sélection du nombre de composantes à analyser



Cercle des corrélation



MERCI BEAUCOUP !