Santé publique France

open FOOD facts

Projet 2 :

Préparez des données pour un organisme de santé publique

NUTRI-SCORE A B C D E

Matthieu MARQUER – 28 mai 2023

Sommaire

01

Le Contexte

Eviter les erreurs lors d’ajout de produit.

L’objectif est de proposer une première étape pour la mise en place d’un système d'autocomplétion .

Caractéristiques des données

- L'intégralité des données provient de openfoodfacts (https://world.openfoodfacts.org/) version OCR

- Données initiales :
  fr.openfoodfacts.org.products.csv

02

La Démarche

Nettoyage
Visualisation
Analyse en Composantes Principales

Nettoyage

- Suppression de variables non pertinente:
  - url
  - creator
  - created_t
  - created_datetime
  - last_modified_t
  - last_modified_datetime
  - brands_tags
  - countries
  - countries_tags
  - states
  - states_tags
  - states_fr

- Taux de remplissage des variables

	colonne_id	pourc
0	code	99.992830
24	countries_fr	99.912711
1	product_name	94.462734
6	brands	91.142618
51	energy_100g	81.401432
100	proteins_100g	81.030140
104	salt_100g	79.654708
105	sodium_100g	79.640056
25	ingredients_text	77.613383
37	ingredients_from_palm_oil_n	77.606213
40	ingredients_that_may_be_from_palm_oil_n	77.606213

Nettoyage

- Suppression:

  des variables peu rempli (-50%) sauf les pnns groups 1 et 2

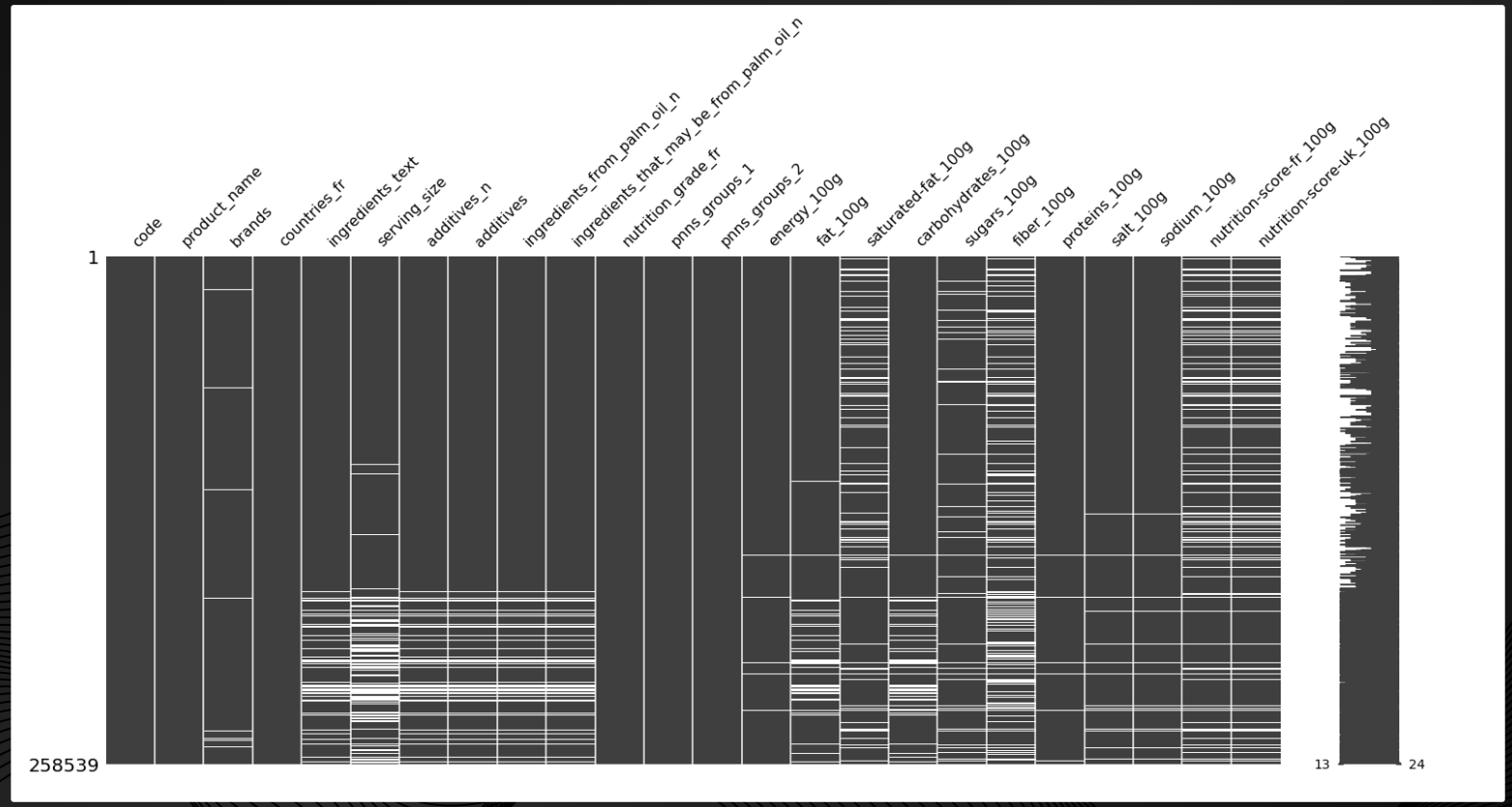
  des codes produit en double

  des produits ayant Nan sur la variable product_name

  des produits ou le total des variables importantes est égal à zéro

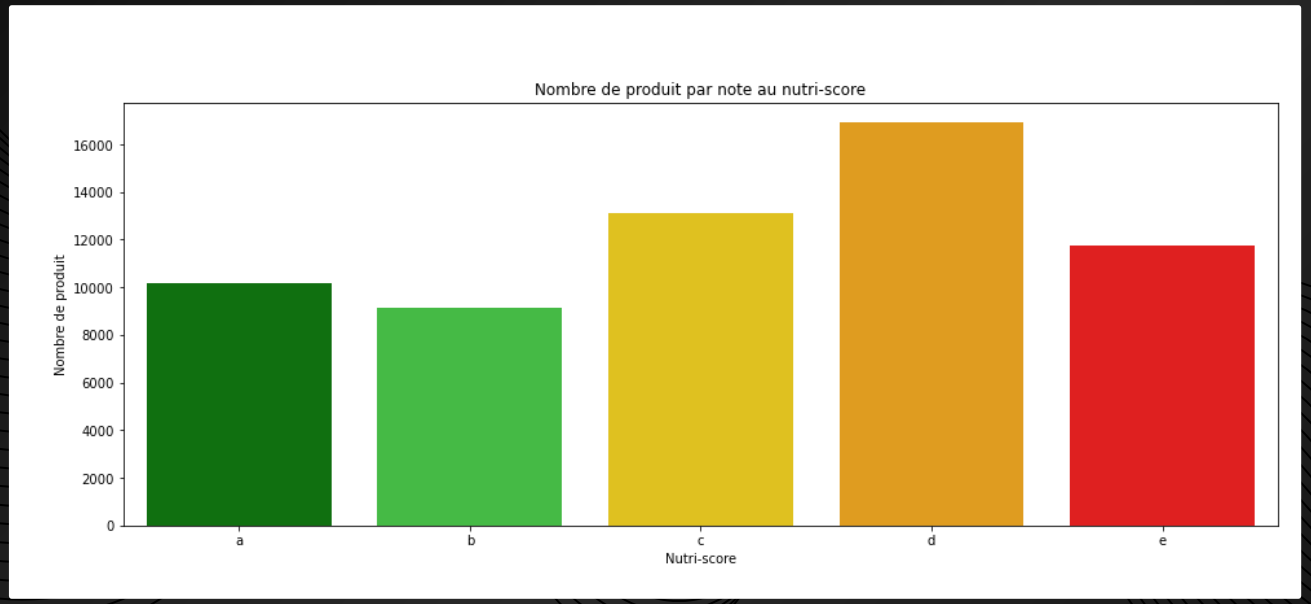
Nettoyage

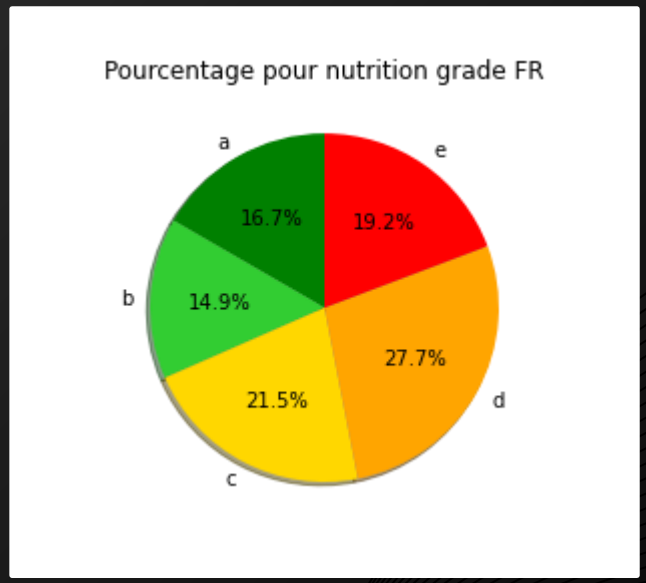
- Utilisation de missingno pour un aperçu graphique des valeurs manquantes



- Sélection de la France et outre-mer

Répartition des produits par note au nutri-score pour la France





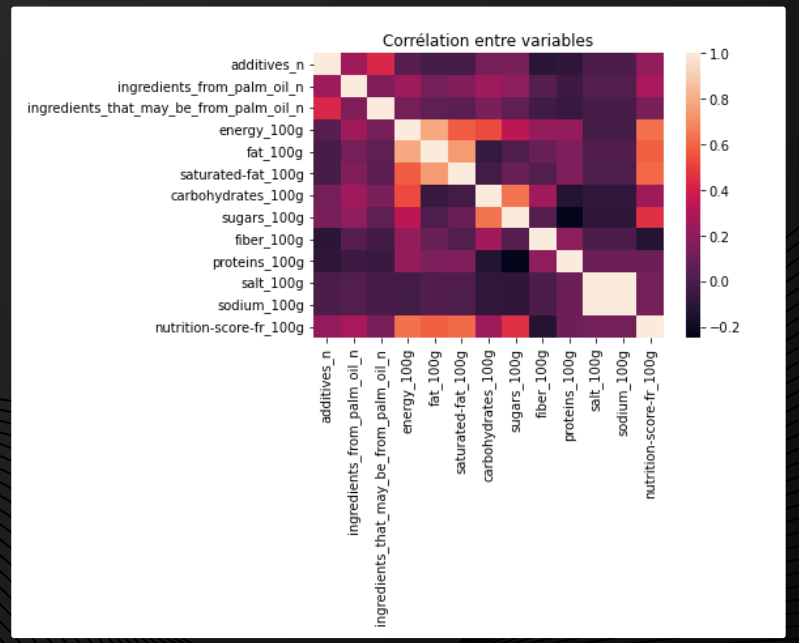
Nettoyage

- Valeur minimum et maximum pour chaque variable:

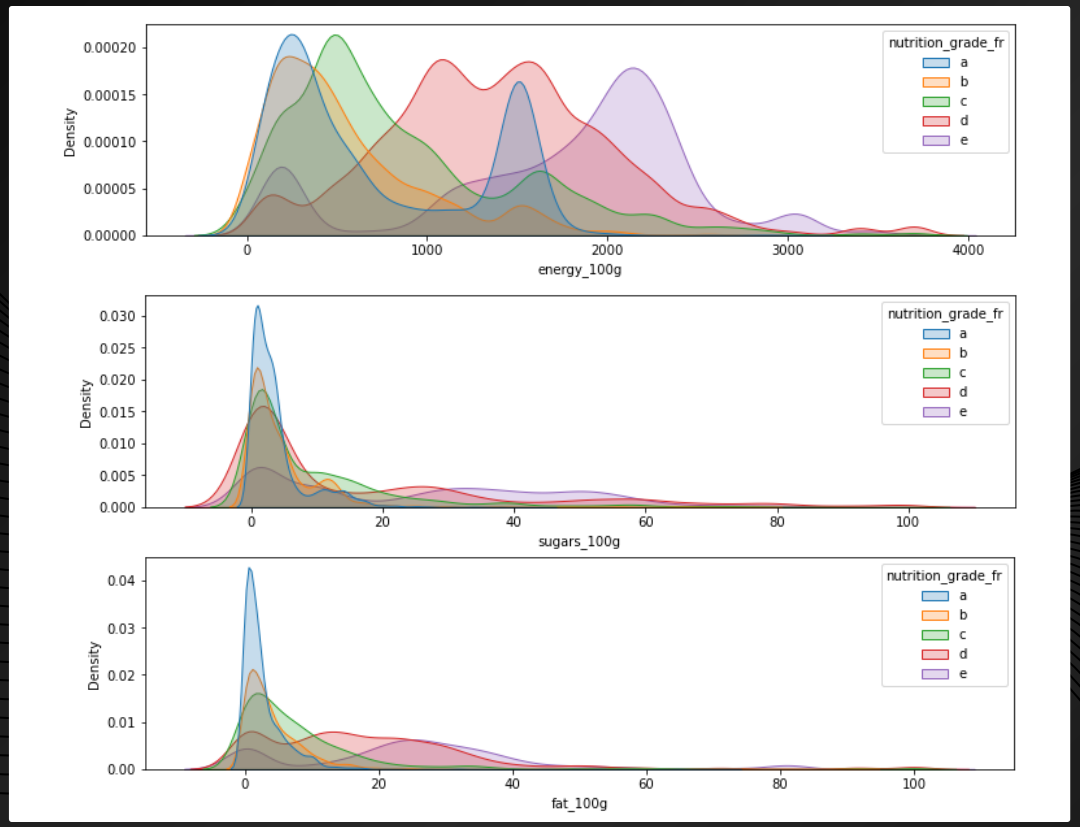
  enregy_100g de 0 à 3700
  fat_100g de 0 à 100
  saturated-fat_100g de 0 à 100
  carbohydrates _100g de 0 à 100
  sugars _100g de 0 à 100
  fiber _100g de 0 à 100
  proteins _100g de 0 à 100
  salt _100g de 0 à 100
  sodium _100g de 0 à 100

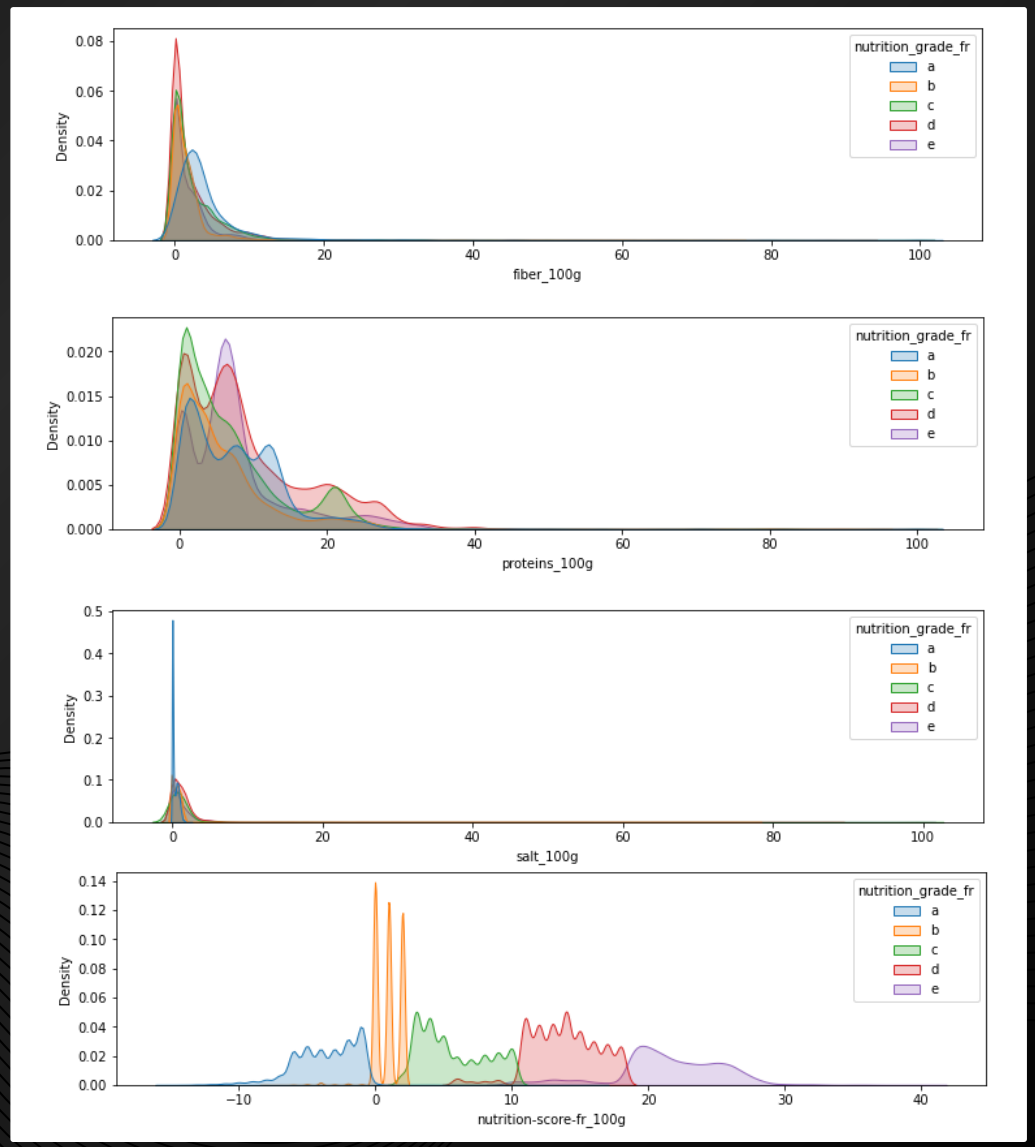
- Corrélation:

  Très forte entre salt et sodium
  Forte entre fat et saturated_fat



Courbe de densité de probabilité

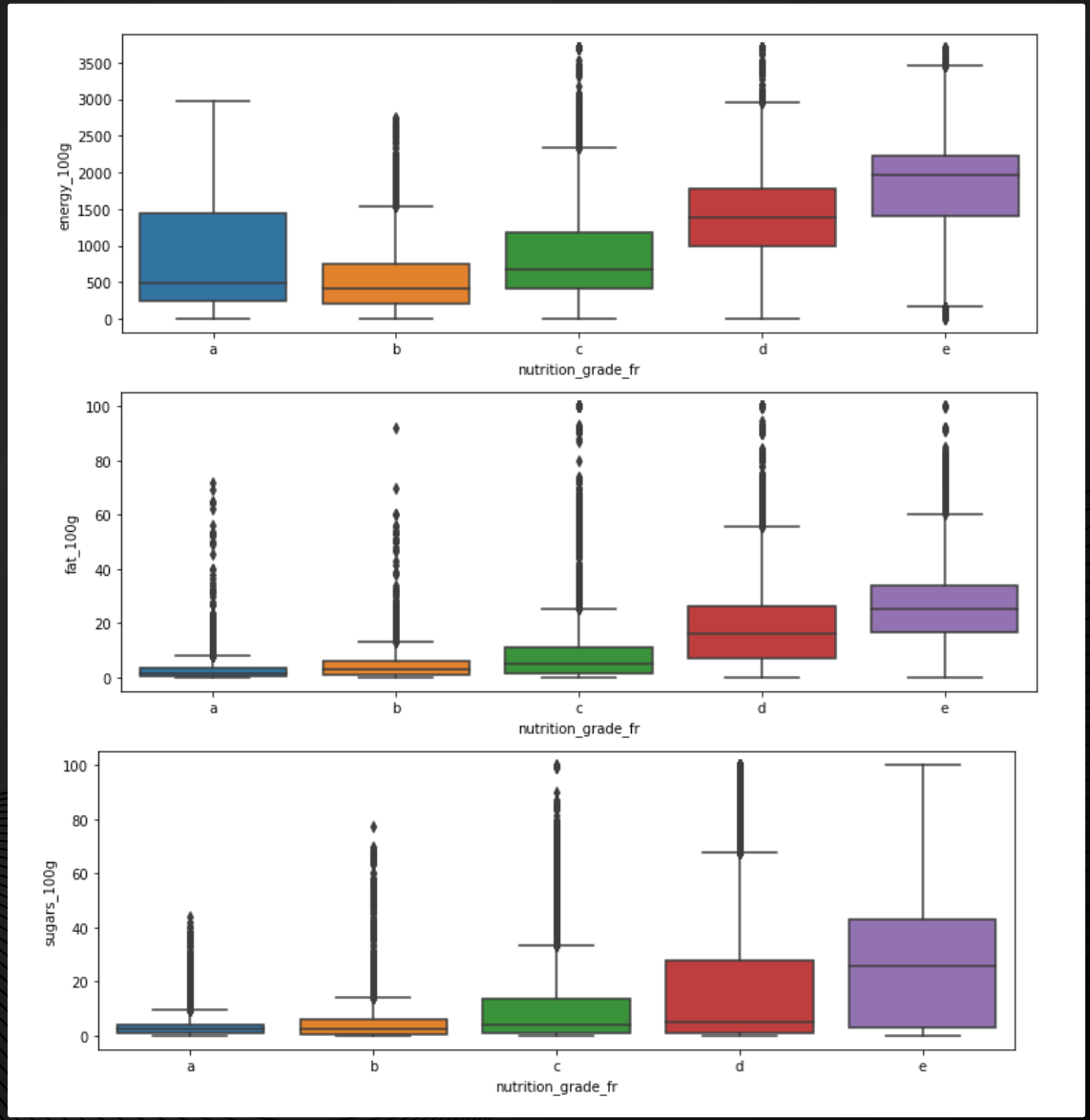




Impact du nutri-score sur la composition des produits

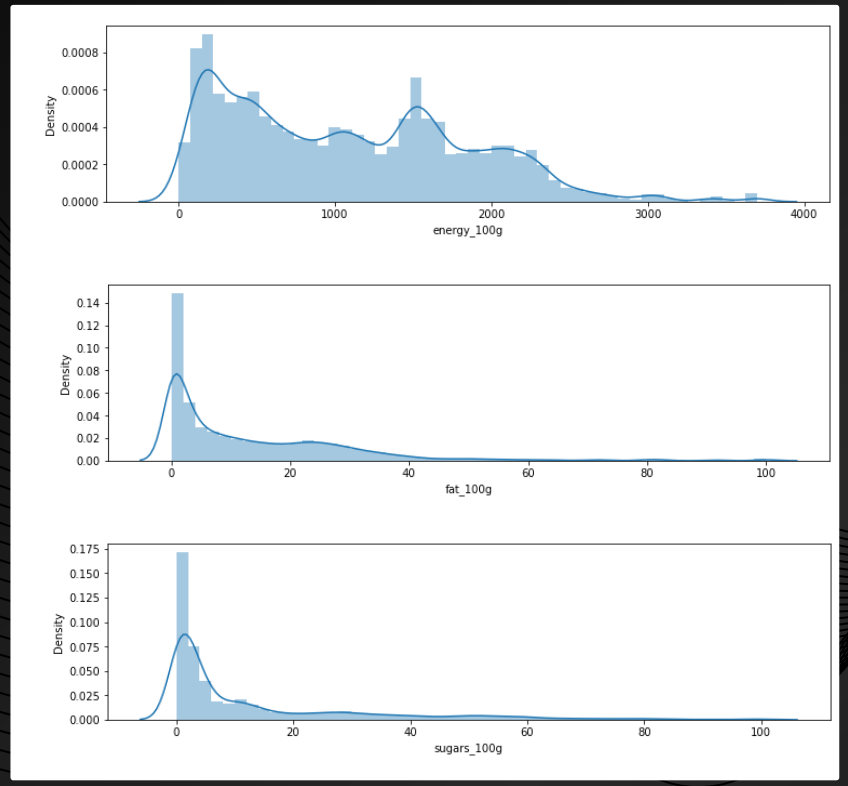
- Visuel des variables par note du nustri-score:

  energy
  fat
  sugars

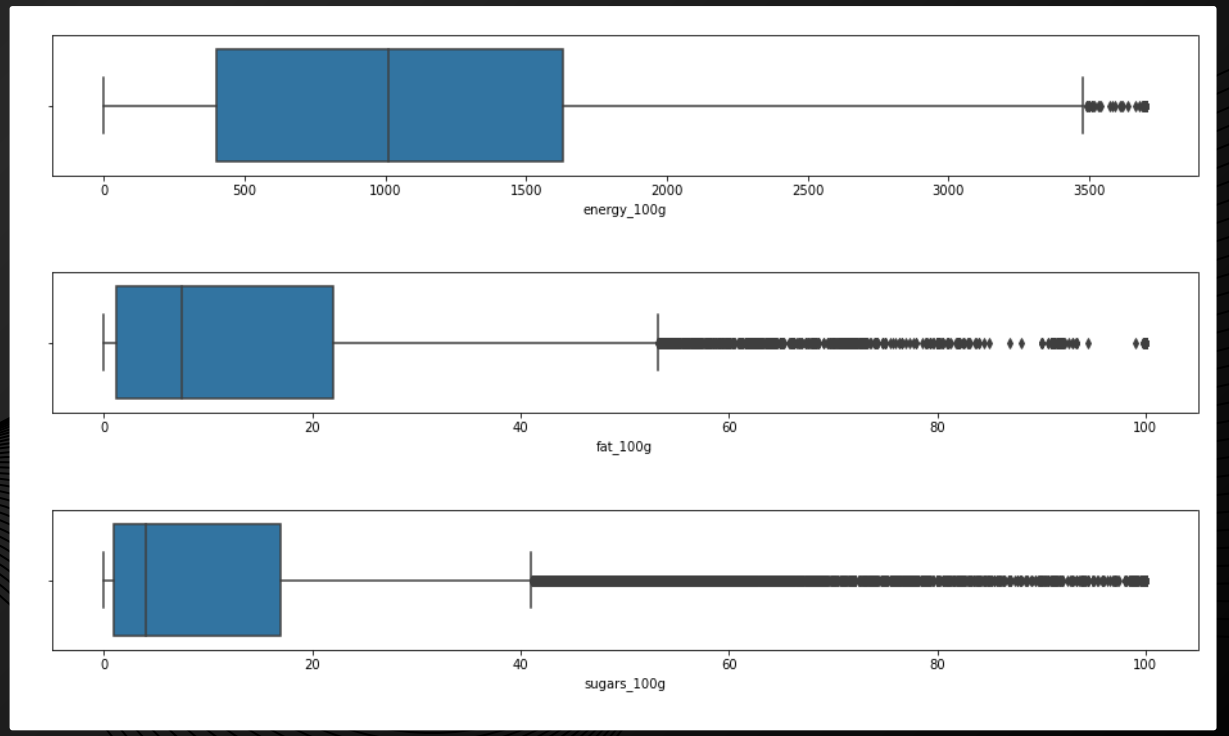


Densité et répartition des variables energy, fat et sugars

- Densité:

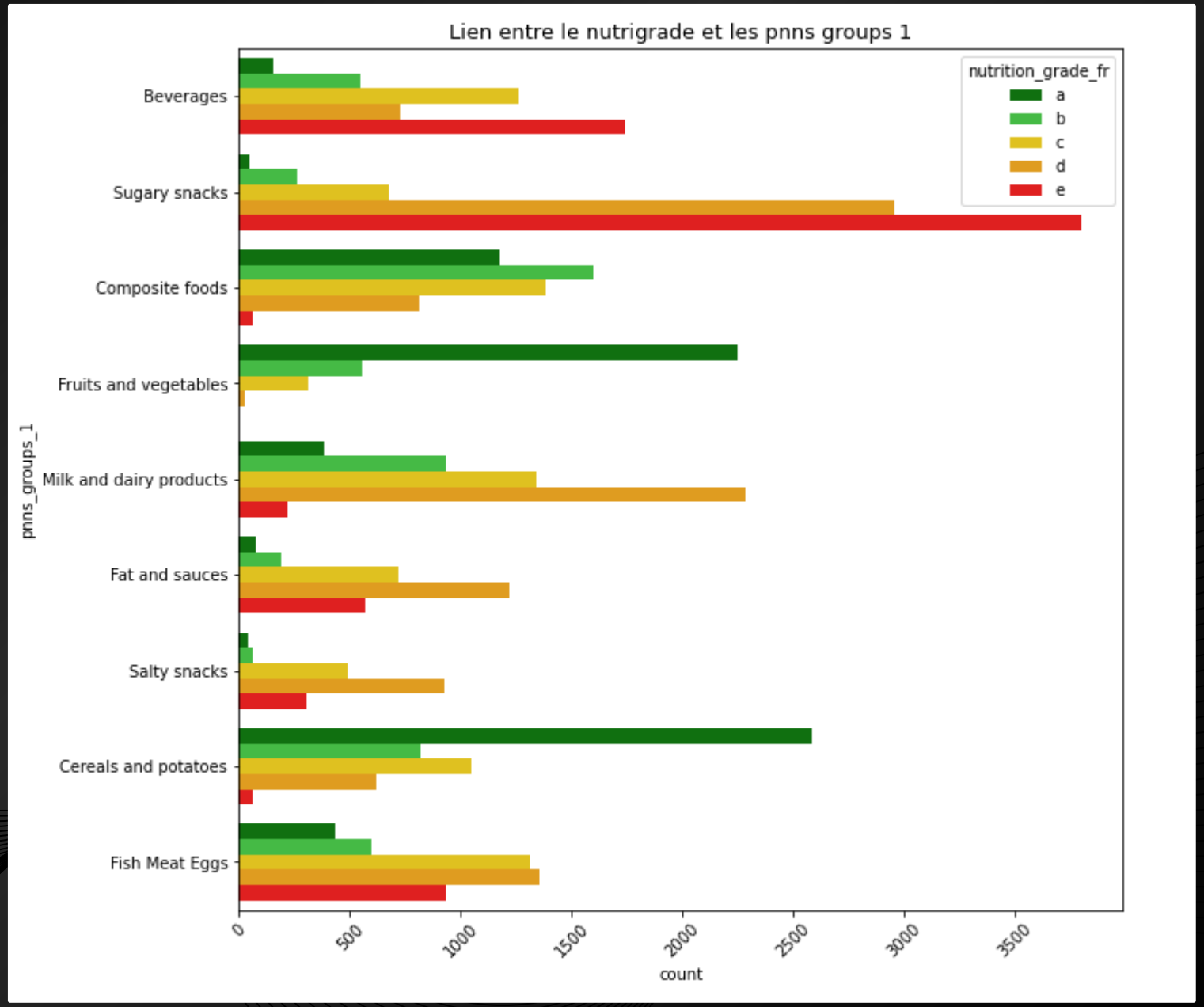


- Répartition:



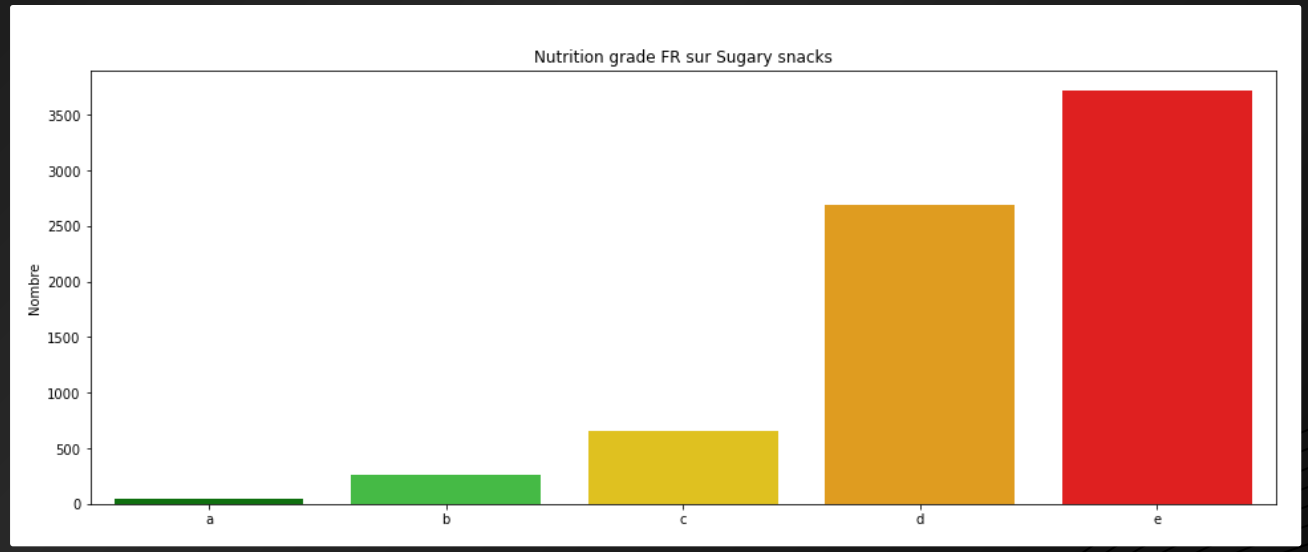
Impact du nutri-score par type de produits

- Nutri-score sur l'ensemble du pnns group 1

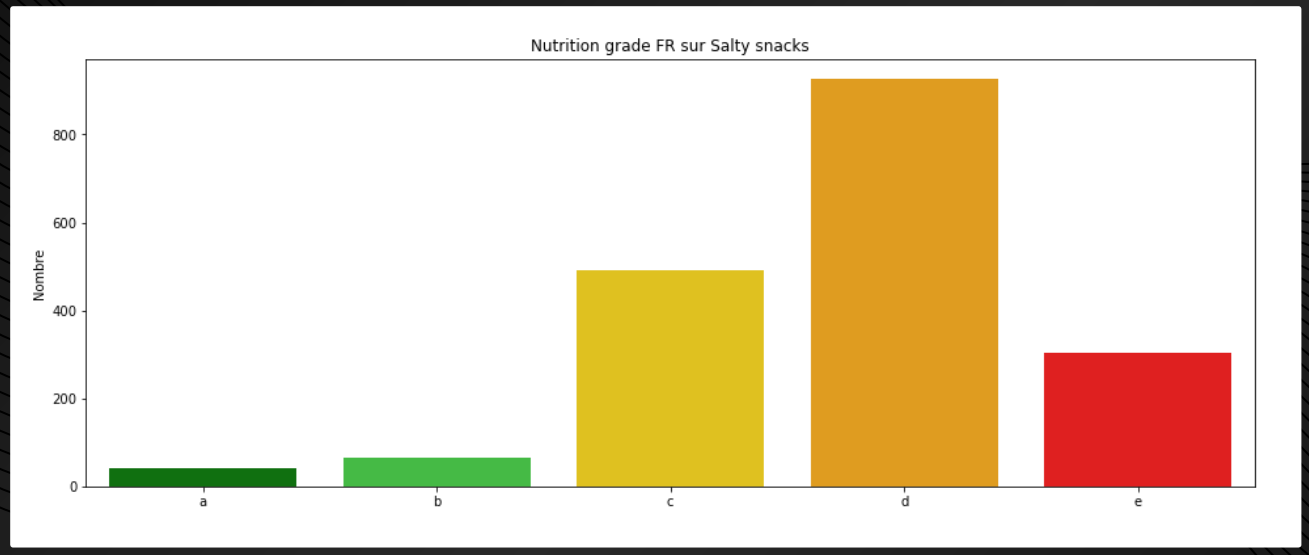


Impact du nutri-score par type de produits

- Nutri-score sur Sugary snacks



- Nutri-score sur Salty snacks

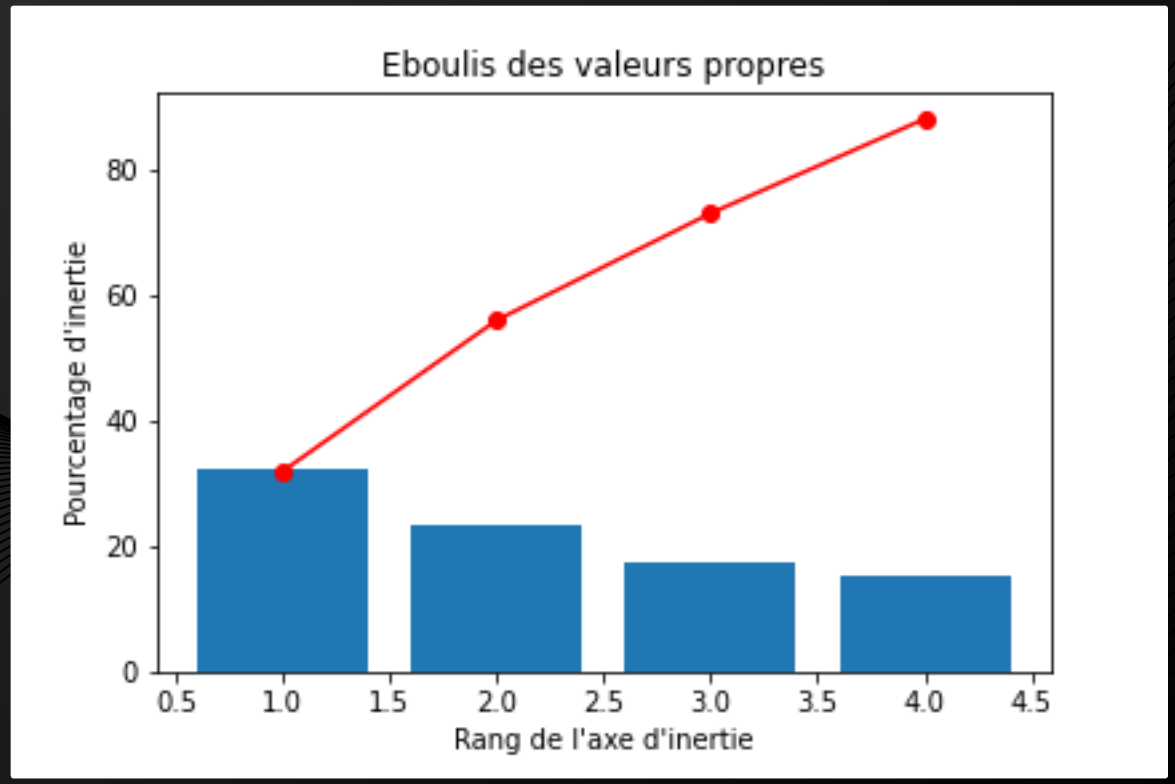


Analyse en composantes principales

Eboulis des valeurs propres

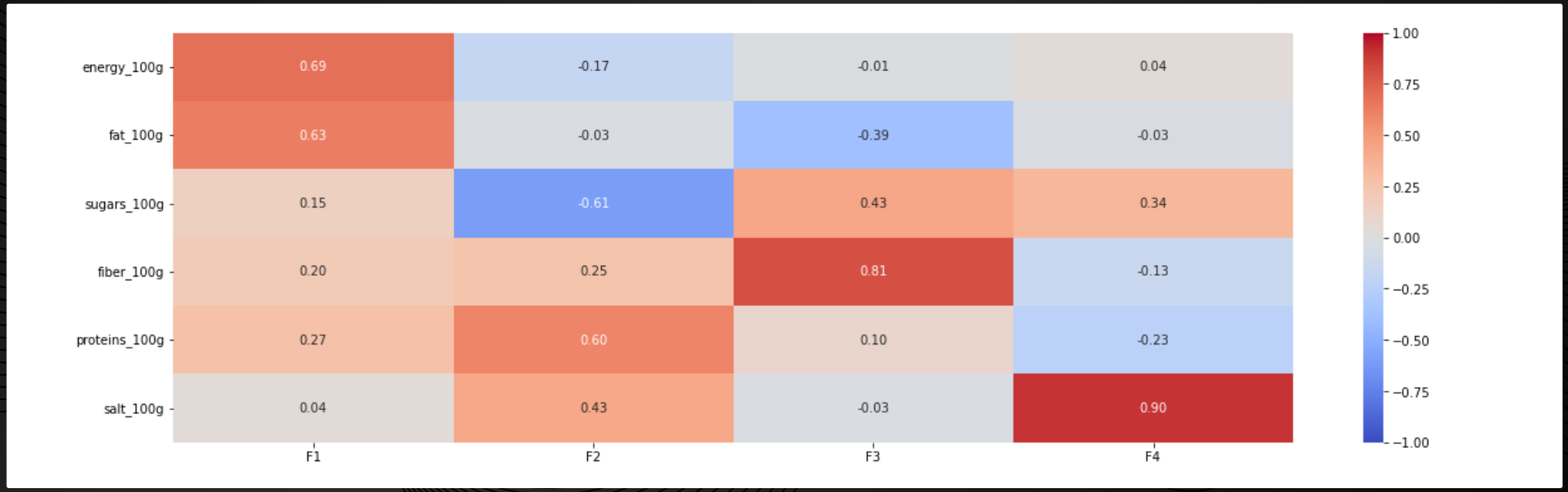
- Les 4 premières composantes représentent 88,2% de l'ensemble.

- N°1: 32,3%
- N°2: 23,2%
- N°3: 17,5%
- N°4: 15,2%

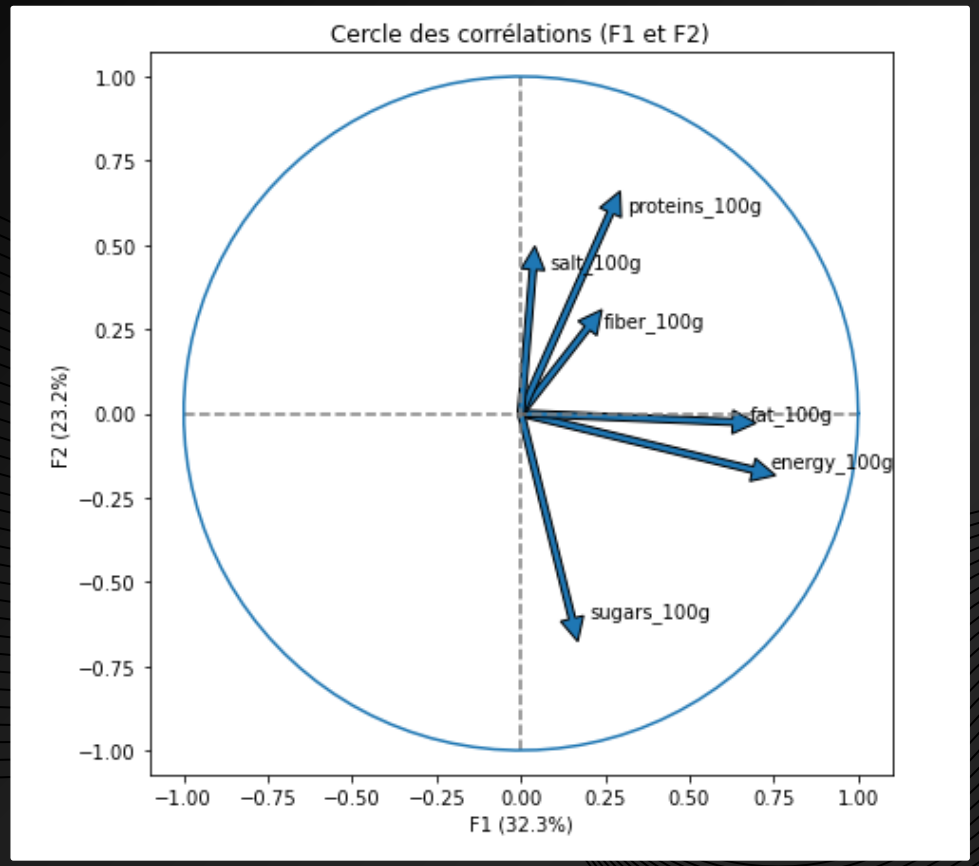


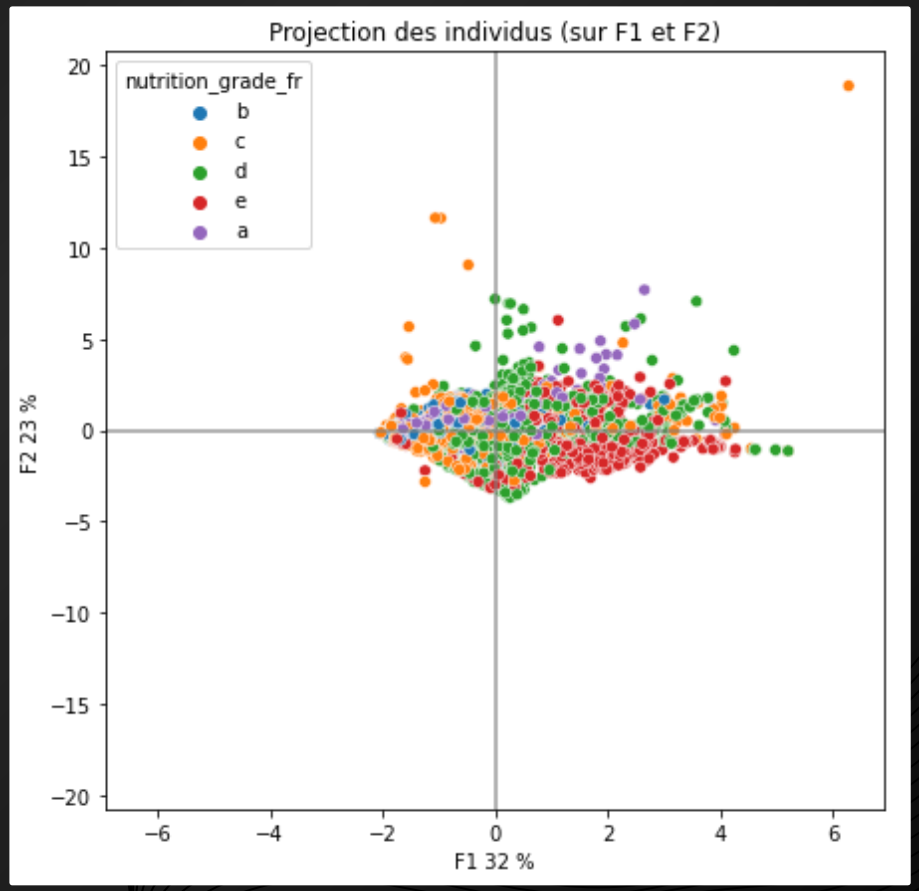
Analyse en composantes principales

Matrice de corrélation



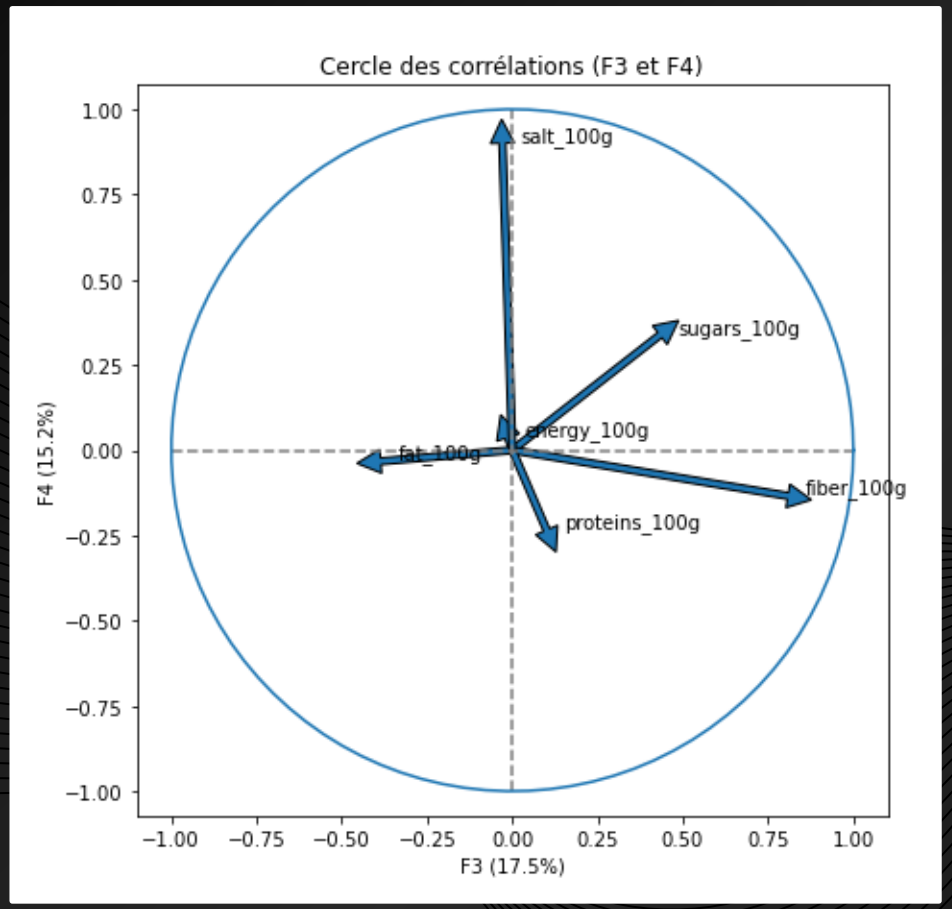
Analyse en Composantes Principales

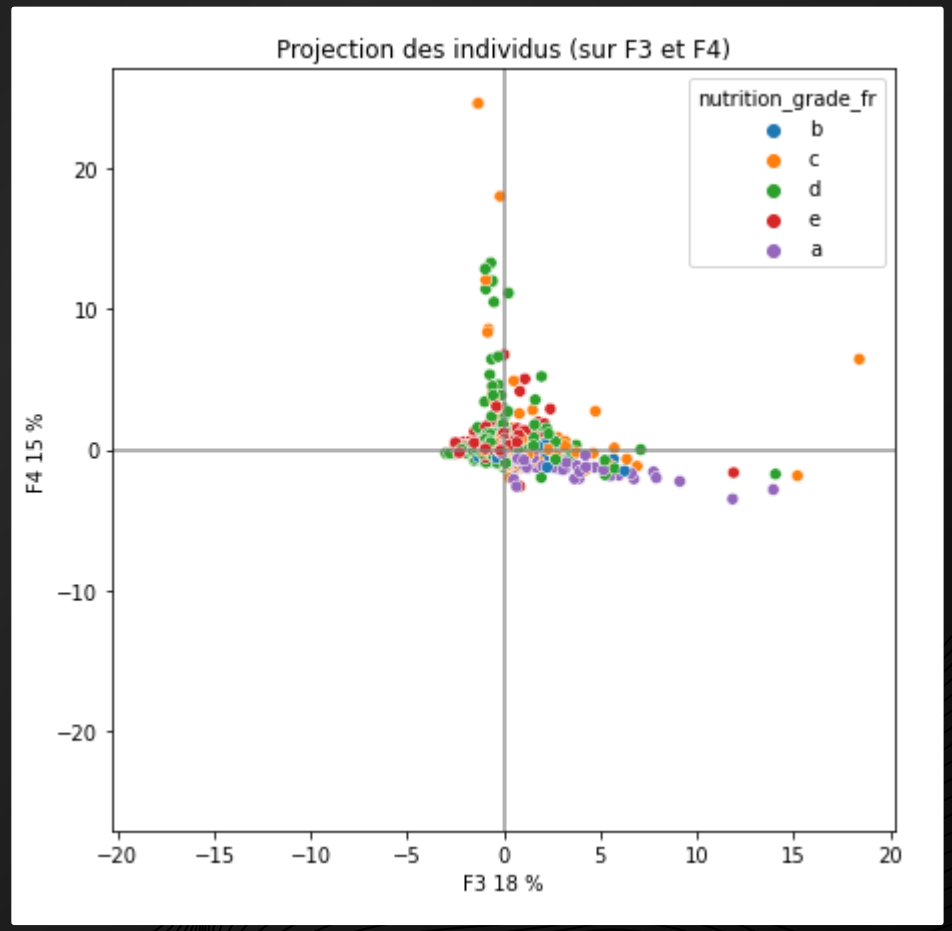




- Corrélation entre produits gras et énergie
- Corrélation négative entre produits salées et sucrées

Analyse en Composantes Principales





Kruskal-Wallis

- P-values:
  Zéro pour energy_100g
  Proche de 0 pour fiber_100g

- Statistique de Kruskal bien plus élevée pour energy

- Résultat:
  - Energy_100g:
    - KruskalResult(statistic=6645.56, pvalue=0.0)
  - Fiber_100g:
    - KruskalResult(statistic=945.95, pvalue=1.844523539468639e-203)

Rapport d'exploration:

- Au vu des résultats obtenu certaines variables sont plus facile à définir que d'autres, cependant il serait préférable de bloquer toutes données considérer comme trop extrême plutôt que de vouloir définir une donnée exact.

RGPD

Les 5 grands principes des règles de protection des données personnelles sont les suivants :

- Le principe de finalité : le responsable d'un fichier ne peut enregistrer et utiliser des informations sur des personnes physiques que dans un but bien précis, légal et légitime ;

- Le principe de proportionnalité et de pertinence : les informations enregistrées doivent être pertinentes et strictement nécessaires au regard de la finalité du fichier ;

- Le principe d'une durée de conservation limitée : il n'est pas possible de conserver des informations sur des personnes physiques dans un fichier pour une durée indéfinie. Une durée de conservation précise doit être fixée, en fonction du type d'information enregistrée et de la finalité du fichier ;

- Le principe de sécurité et de confidentialité : le responsable du fichier doit garantir la sécurité et la confidentialité des informations qu'il détient. Il doit en particulier veiller à ce que seules les personnes autorisées aient accès à ces informations ;

- Les droits des personnes

Merci...

Avez-vous des questions?