FORMATION DATA SCIENTIST : PROJET 3

Préparez des données pour un organisme de santé publique

Sommaire

1. Introduction

- **Agence Santé publique France**

acteur majeur en santé publique, dédié à l'amélioration de la santé des citoyens à travers la prévention et la recherche

- **Open Food Facts**

base de données collaborative, libre et ouverte contenant des informations nutritionnelles sur des milliers de produits alimentaires

- **Défi**

Actuellement, l'ajout de données à Open Food Facts est entravé par des erreurs de saisie et des valeurs manquantes, réduisant la fiabilité et la qualité de l'information disponible

- **Impact**

Améliorer cette base de données augmentera significativement la qualité de l'information nutritionnelle accessible, contribuant ainsi à une meilleure santé publique

- **Développement d'un système de suggestion/auto-complétion**

Notre mission est de créer un outil innovant pour assister les utilisateurs dans la saisie des données, rendant le processus plus efficace et réduisant les erreurs.

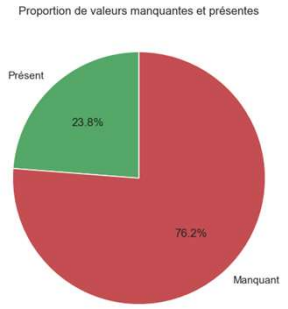
- **Focus sur le nettoyage et l'exploration des données**

La première étape cruciale est de nettoyer et d'explorer la base de données existante, pour évaluer la faisabilité et la pertinence de notre solution proposée.

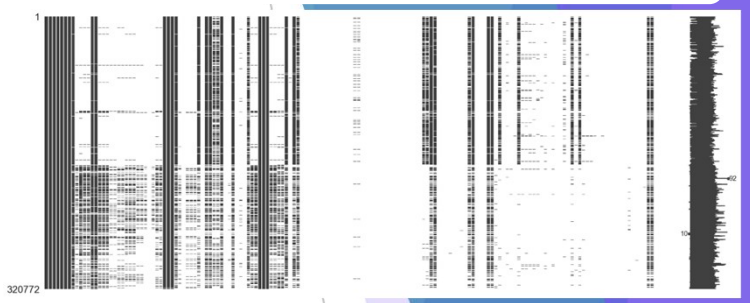
2. Nettoyage des données

- État de la base de données avant nettoyage :

  320772 lignes et 162 colonnes avec beaucoup de valeurs manquantes



Matrice des valeurs présentes/manquantes dans notre jeu de données



- Test de faisabilité de l'application sur une variable cible :

  Nutrition score : catégorielle, beaucoup de valeurs manquantes (31%), très importante pour les objectifs de santé publique.

- Identification des variables pertinentes

Revue des champs disponibles et sélection basée sur la pertinence pour l'application de suggestion/auto-complétion.

Score et groupes :
nutrition_grade_fr, nutrition-score-fr_100g, pnns_groups_1, pnns_groups_2

Nutriments et énergie:
fat_100g, saturated-fat_100g, carbohydrates_100g, sugars_100g, proteins_100g, fiber_100g, alcohol_100g, salt_100g, energy_100g

- Suppression des doublons

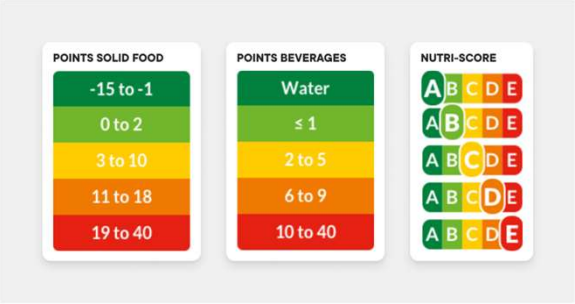
sur code / marque, nom de produit /marque, nom de produit / energy 100g, marque / energy 100g

2. Nettoyage des données

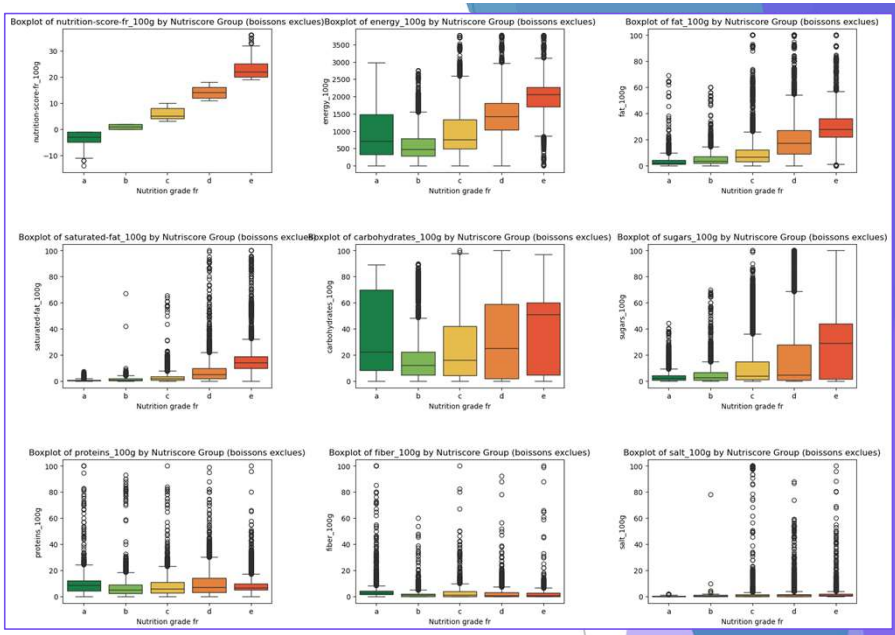
- Gestion des valeurs hors-normes

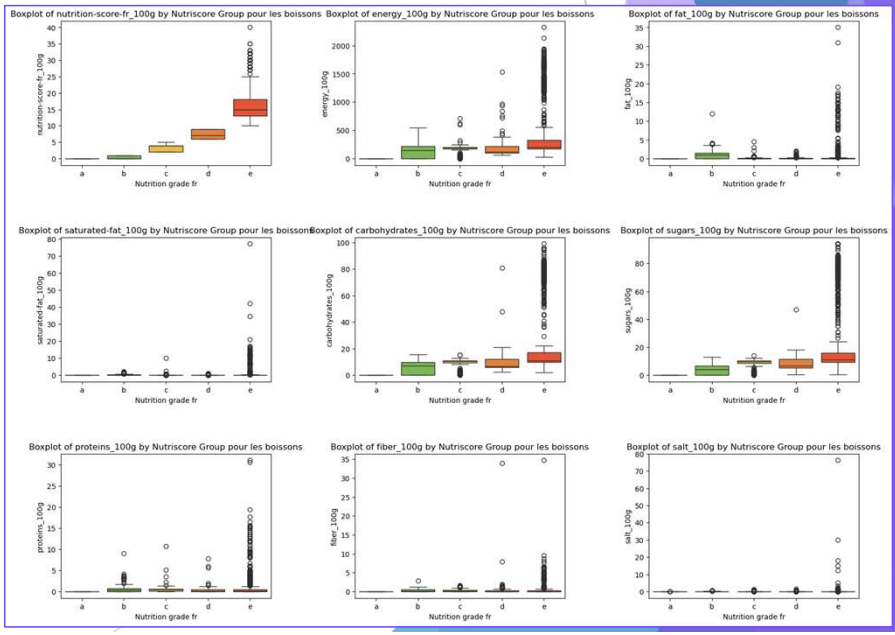
Application de règles métiers :

- Colonnes en _100g : Valeur max fixée à 100
- Total nutriments : Ne doit pas excéder 100g
- Nutrition score : Bornes de -15 à 40
- Correspondance : Nutrition score ↔ Nutrition grade
- Vérification énergétique : Calcul de l'énergie pour 100g
- Cohérence nutriments : Graisses ↔ Graisses saturées, Glucides ↔ Sucres



- Trop de valeurs hors-normes pour être vérifiées à la main, nous partirons du principe qu'il s'agit de valeurs atypiques.

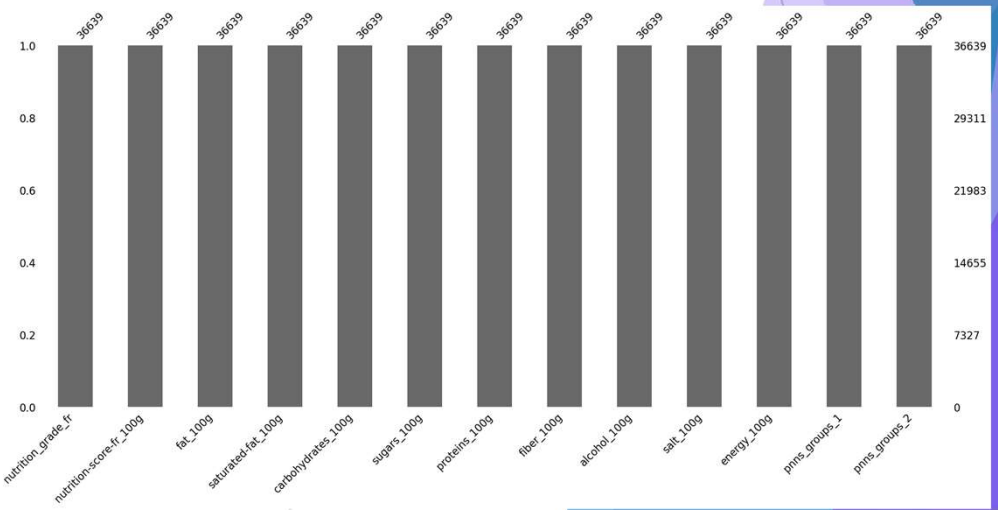




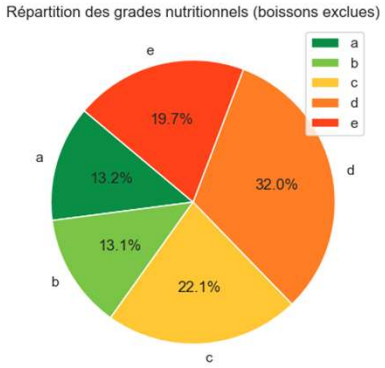
2. Nettoyage des données

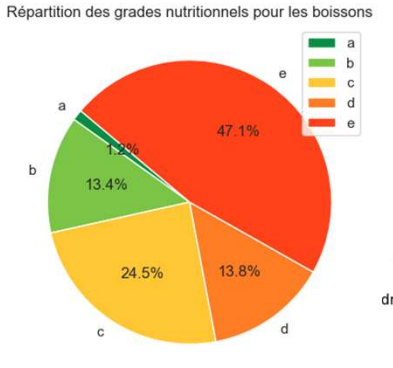
- Méthodes de traitement des valeurs manquantes :
  - Imputation à 0 :
    Fibres, alcool, nutriments : imputation à zéro si cohérent.
  - KNeighborsClassifier :
    Catégories d'aliments ('pnns_groups_1' et 'pnns_groups_2') : prédiction et imputation par KNN.
  - Imputation Médiane par Groupe :
    Variables 'pnns_groups_2' : médiane pour les valeurs manquantes.
  - Suppression et Vérification Post-Imputation :
    Suppression : entrées non imputables.
    Contrôle qualité : retraitement des valeurs aberrantes après imputation.

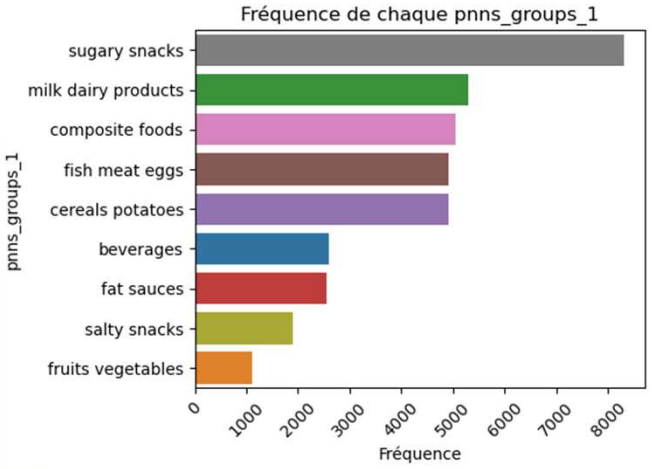
36639 lignes et 13 colonnes
0% de valeurs manquantes

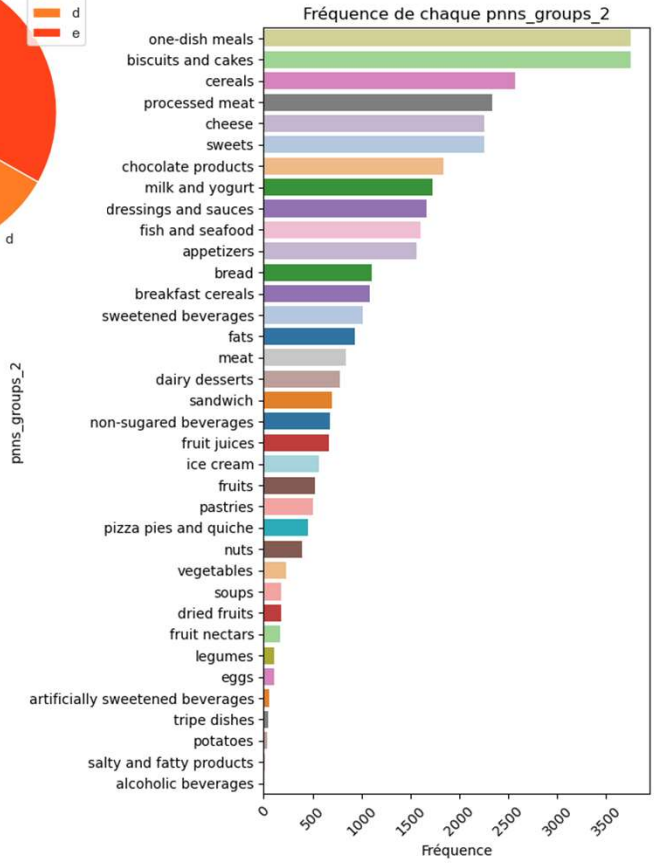


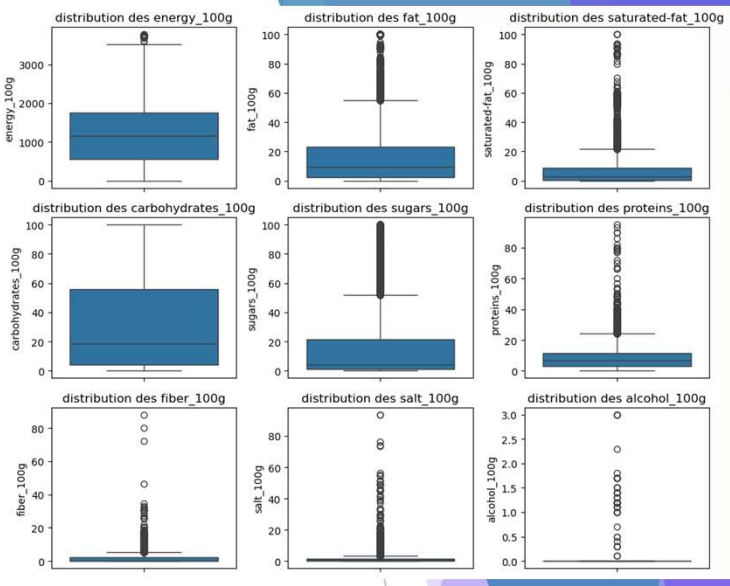
3. Exploration des données

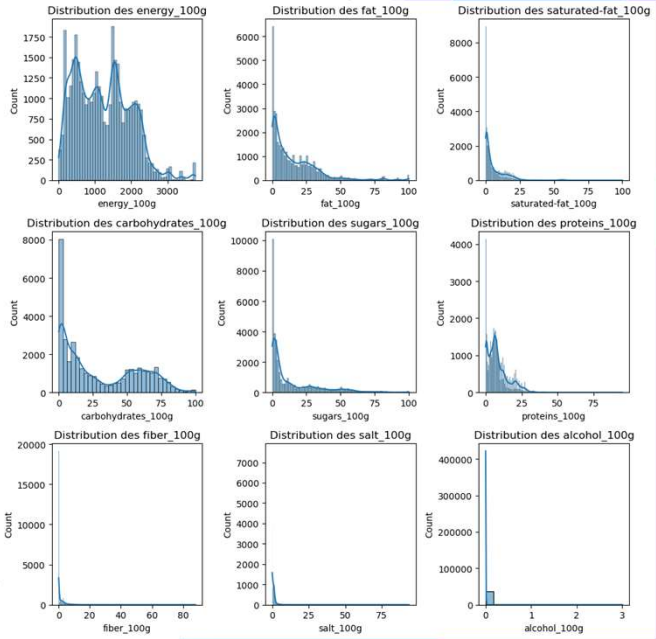




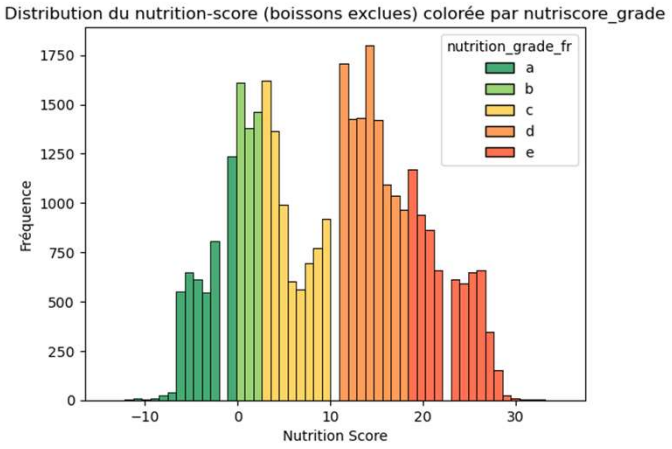


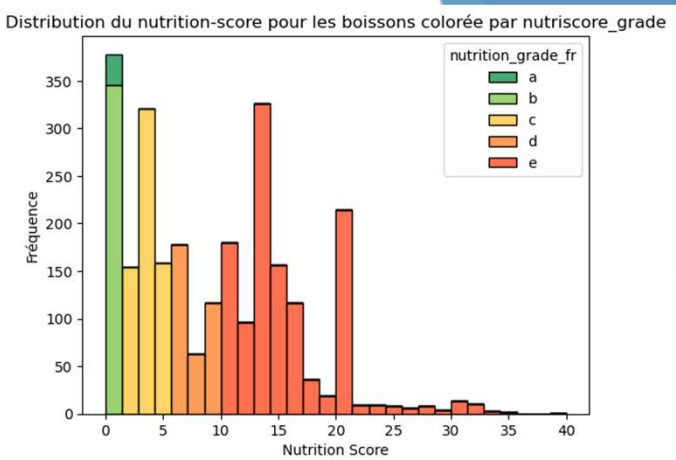


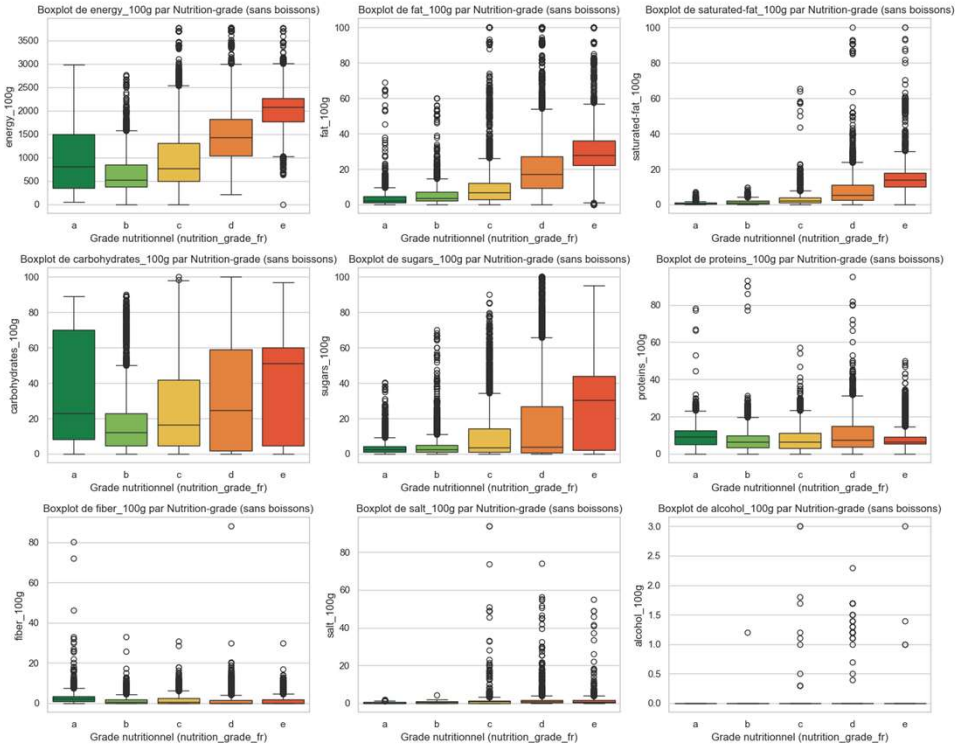


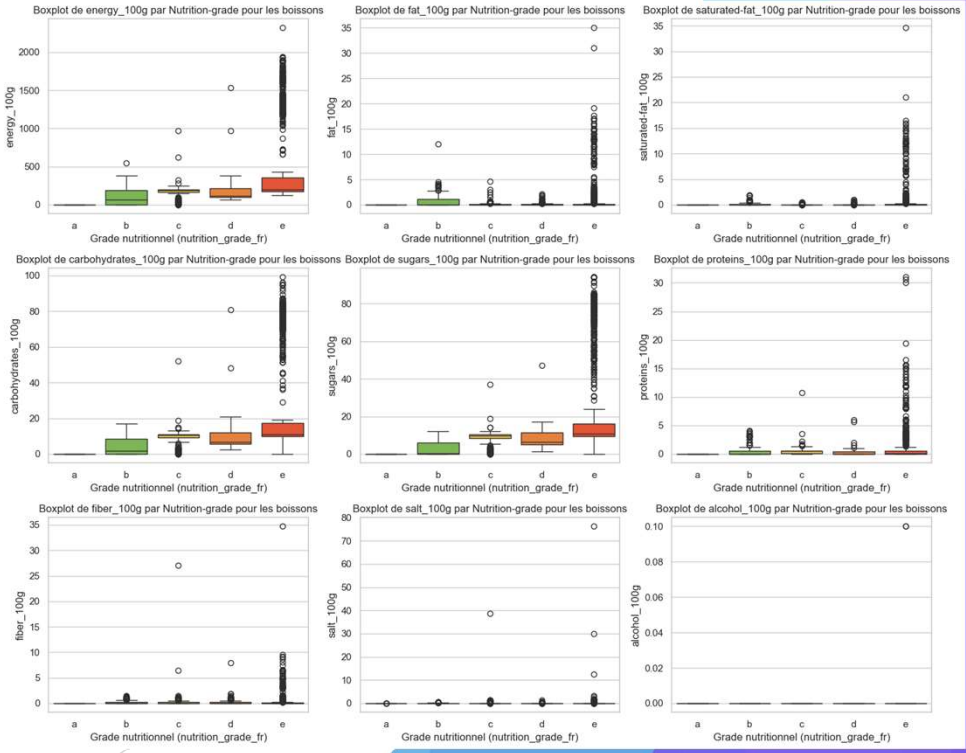


4. Analyse bivariée

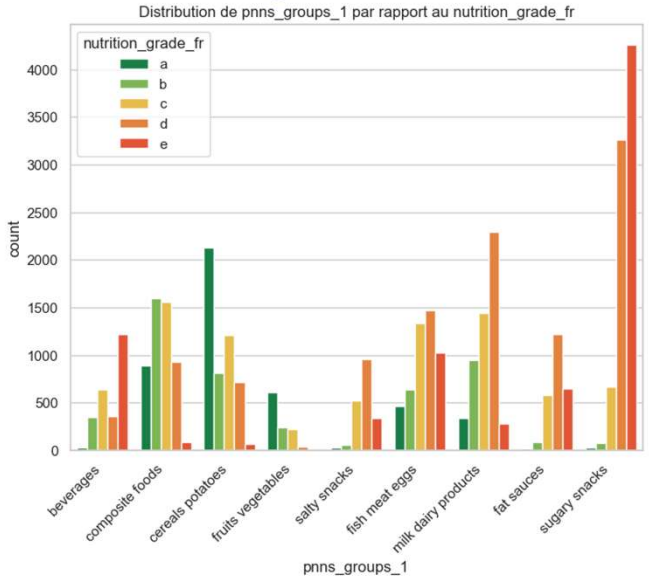


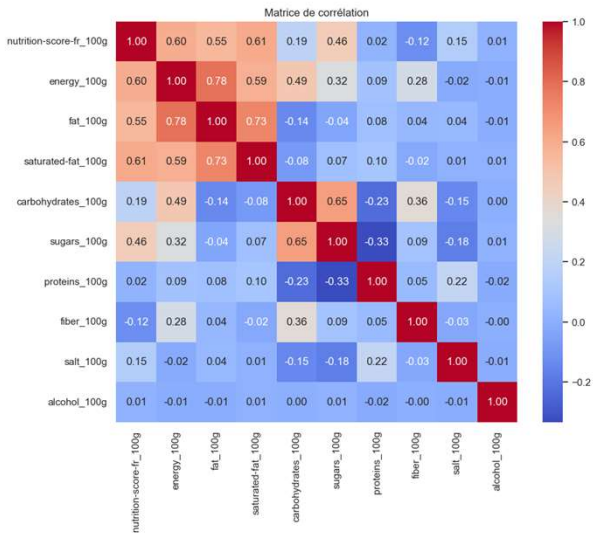


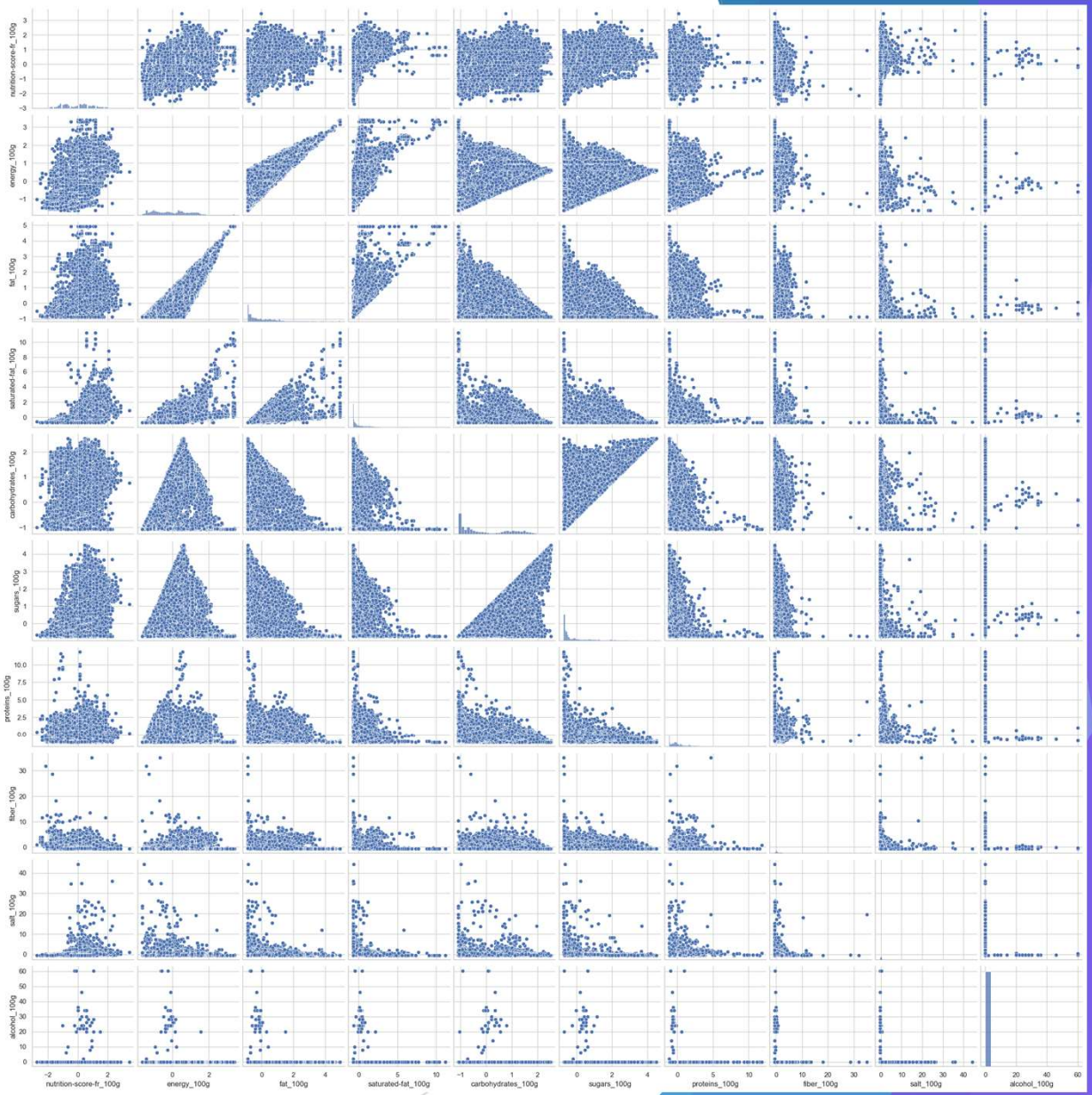




4. Analyse bivariée







5. Analyse Multivariée

- ANOVA

Aliments Solides

- Signification Statistique :
  - P-valeurs < 0.05 pour la plupart des variables nutritionnelles.
  - Forte indication que les grades nutritionnels ne sont pas attribués au hasard.
- Interprétation Globale :
  - Forte Association : "nutrition_grade_fr" corrèle étroitement avec les profils nutritionnels.
  - Exceptions Notables :
    # Protéines, fibres, sel : pas toujours différenciateurs significatifs.
    # Suggère une influence de facteurs nutritionnels plus complexes sur les grades.

Boissons

- Modèle Distinct:
  - Moins de variations significatives entre grades, spécialement entre a et b.

```
ANOVA pour tous les aliments sauf les boissons
Variable: nutrition-score-fr_100g
Table ANOVA:
                          sum_sq       df             F  PR(>F)
C(nutrition_grade_fr)  2.491134e+06      4.0  125548.924436     0.0
Residual               1.688797e+05  34045.0           NaN     NaN

Résultats du test de Tukey:
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj  lower   upper  reject
-------------------------------------------------
     a      b   4.1335   0.0  4.0049   4.262   True
     a      c   9.0821   0.0  8.9675  9.1966   True
     a      d  17.2737   0.0 17.1659 17.3815   True
     a      e  25.6385   0.0 25.5213 25.7558   True
     b      c   4.9486   0.0  4.8337  5.0635   True
     b      d  13.1402   0.0 13.0321 13.2483   True
     b      e  21.5051   0.0 21.3876 21.6226   True
     c      d   8.1916   0.0  8.1006  8.2827   True
     c      e  16.5565   0.0 16.4544 16.6586   True
     d      e   8.3648   0.0  8.2705  8.4592   True
-------------------------------------------------
```

```
Variable: energy_100g
Table ANOVA:
                          sum_sq       df            F  PR(>F)
C(nutrition_grade_fr)  6.965100e+09      4.0  5523.452804     0.0
Residual               1.073273e+10  34045.0          NaN     NaN

Résultats du test de Tukey:
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
group1 group2  meandiff p-adj   lower     upper   reject
-----------------------------------------------------
     a      b -251.0424   0.0 -283.4427 -218.6421   True
     a      c   61.2146   0.0   32.3282    90.101   True
     a      d  563.5801   0.0  536.4096  590.7506   True
     a      e 1094.1734   0.0 1064.6192 1123.7276   True
     b      c  312.257    0.0  283.2996  341.2145   True
     b      d  814.6225   0.0  787.3765  841.8686   True
     b      e 1345.2158   0.0 1315.5922 1374.8394   True
     c      d  502.3655   0.0  479.4092  525.3218   True
     c      e 1032.9588   0.0 1007.2255 1058.6921   True
     d      e  530.5933   0.0  506.8022  554.3844   True
-----------------------------------------------------
```

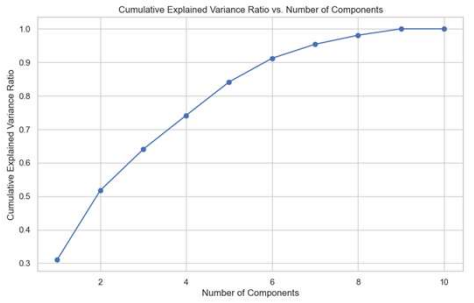
```
ANOVA pour les boissons
Variable: nutrition-score-fr_100g
Table ANOVA:
                          sum_sq      df          F  PR(>F)
C(nutrition_grade_fr)  101606.956419     4.0  2464.60388     0.0
Residual                26632.309706  2584.0         NaN     NaN

Résultats du test de Tukey:
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=================================================
group1 group2 meandiff p-adj   lower   upper  reject
-------------------------------------------------
     a      b   0.2775 0.9902 -1.3418  1.8967  False
     a      c   3.5836    0.0  1.9958  5.1714   True
     a      d   7.4888    0.0  5.8719  9.1057   True
     a      e  15.6259    0.0 14.0566 17.1953   True
     b      c   3.3061    0.0  2.7204  3.8919   True
     b      d   7.2114    0.0  6.5507   7.872   True
     b      e  15.3485    0.0 14.8147 15.8823   True
     c      d   3.9052    0.0  3.3259  4.4846   True
     c      e  12.0423    0.0 11.6132 12.4714   True
     d      e   8.1371    0.0  7.6103  8.6639   True
-------------------------------------------------
```

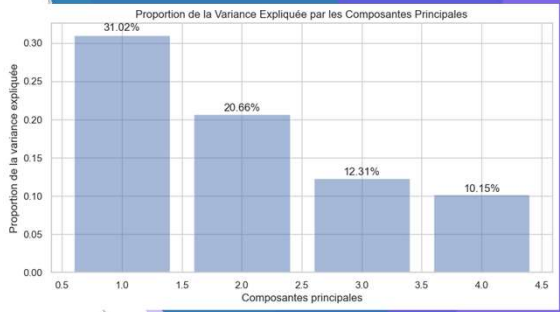
5. Analyse Multivariée

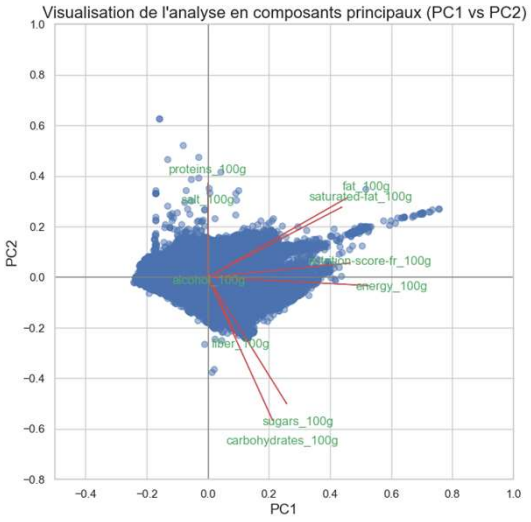
Analyse en Composantes Principales (ACP)

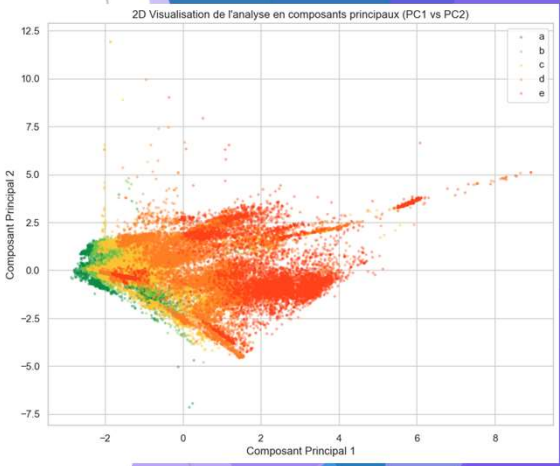
- PC1 (31% de la variance) :
  - Teneur énergétique et Matières Grasses
- PC2 (20,7% de la variance) :
  - contraste entre glucides et protéines/sel
- PC3 (12,3% de la variance) :
  - oppose aliments riches en fibres à ceux riches en sucres
- PC4 (10,2% de la variance) :
  - contraste entre la teneur en sel et en graisses

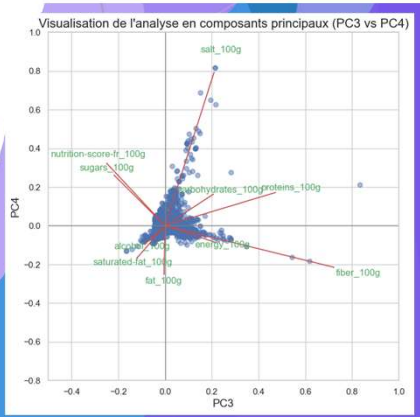
Quatre composants capturent ~84% de la variance totale.











6. Résultats et Observations

- Principales Découvertes :

Analyse Univariée :

Distribution inégale des grades nutritionnels, avec une prédominance du grade 'e' chez les boissons.

Concentration élevée de produits dans certaines catégories PNNS, notamment « sugary snacks " et " milk dairy products ".

Analyse Multivariée :

ANOVA met en lumière des différences nutritionnelles significatives entre les grades.

L'ACP révèle des axes principaux capturant les variations nutritionnelles, avec une distinction claire entre les nutriments associés à chaque axe.

- Évaluation de la Faisabilité de l'Application

Les analyses suggèrent une corrélation forte entre les variables nutritionnelles et le grade, indiquant une bonne base pour la suggestion/auto-complétion.

La complexité des patterns nutritionnels justifie le besoin d'un système avancé pour assister la saisie des données.

- Observations Clés

Nutrition Grade :

Corrélé significativement avec les profils nutritionnels, impliquant une potentialité pour la prédiction automatique.

Variables Distinguantes :

Des nutriments clés tels que les graisses, sucres, et fibres montrent des variations importantes, indiquant des critères solides pour la suggestion de données.

Contraste dans les Boissons :

Les boissons affichent des tendances distinctes par rapport aux aliments solides, soulignant la nécessité d'approches spécifiques pour ces catégories.

7. Respect des Principes du RGPD

Les principes du RGPD :

- Le principe de finalité
- Le principe de proportionnalité et de pertinence
- Le principe d'une durée de conservation limitée
- Le principe de sécurité et de confidentialité
- Les droits des personnes.

Notre projet:

- Finalité claire : aider à la saisie de données nutritionnelles précises
- Collecte de données limitée aux informations nutritionnelles requises.
- Les données sont conservées le temps nécessaire à l'objectif de qualité de la base
- Protection des données conformément aux meilleures pratiques.
- le projet ne traite pas de données personnelles.

Conclusion et Recommandations

- **Synthèse des Résultats**
  Corrélations significatives : validation de l'approche de suggestion/auto-complétion.
  Analyse multivariée : révèle des facteurs nutritionnels clés pour l'orientation des suggestions.

- **Viabilité de l'Application**
  Confirmée par l'analyse des données.
  Prédiction fiable de grades nutritionnels possible.

- **Prochaines Étapes**
  Intégration de l'outil de suggestion dans la base de données Open Food Facts.
  Tests utilisateurs pour affiner l'interface et les fonctionnalités.

- **Recommandations pour Améliorations**
  Continuer l'enrichissement des données pour affiner les modèles de prédiction.
  Évaluer l'impact de l'outil sur la qualité des saisies des utilisateurs.

Merci pour votre attention

Des questions ?