SANTÉ PUBLIQUE FRANCE

Nom du projet : Préparez des données pour un organisme de santé publique

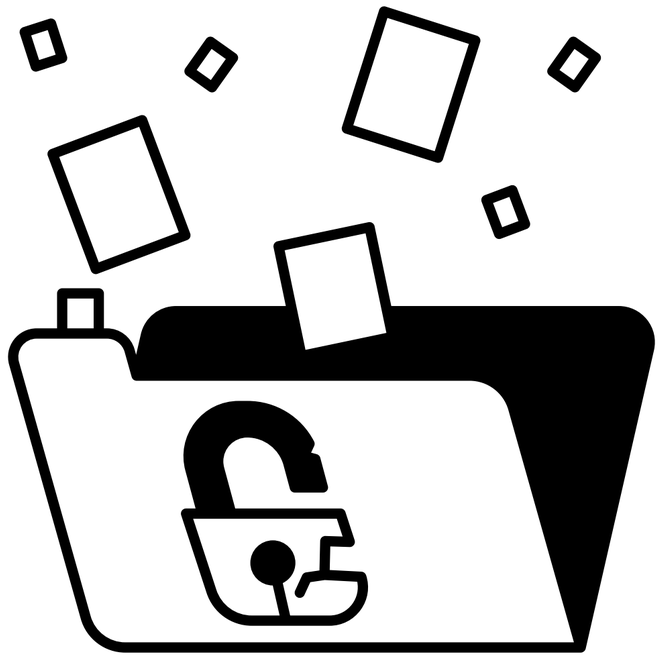
Présenté par : Nathan FARDIN

Le besoin

Traiter les données et mettre en évidence la faisabilité d’une solution permettant de faciliter la complétion de la base de données par les utilisateurs

Plan

RGPD



Finalité

- Les informations sur des individus ne peuvent être enregistrées que dans un but spécifique, légal et légitime.

Proportionnalité et pertinence

- Les données enregistrées doivent être strictement nécessaires et pertinentes par rapport à l'objectif du fichier.

Durée de conservation limitée

lune durée précise doit être fixée en fonction du type de données et de l'objectif du fichier.

Sécurité et confidentialité

- Le responsable du fichier doit assurer la sécurité des informations détenues, en limitant l'accès .

Droits des personnes

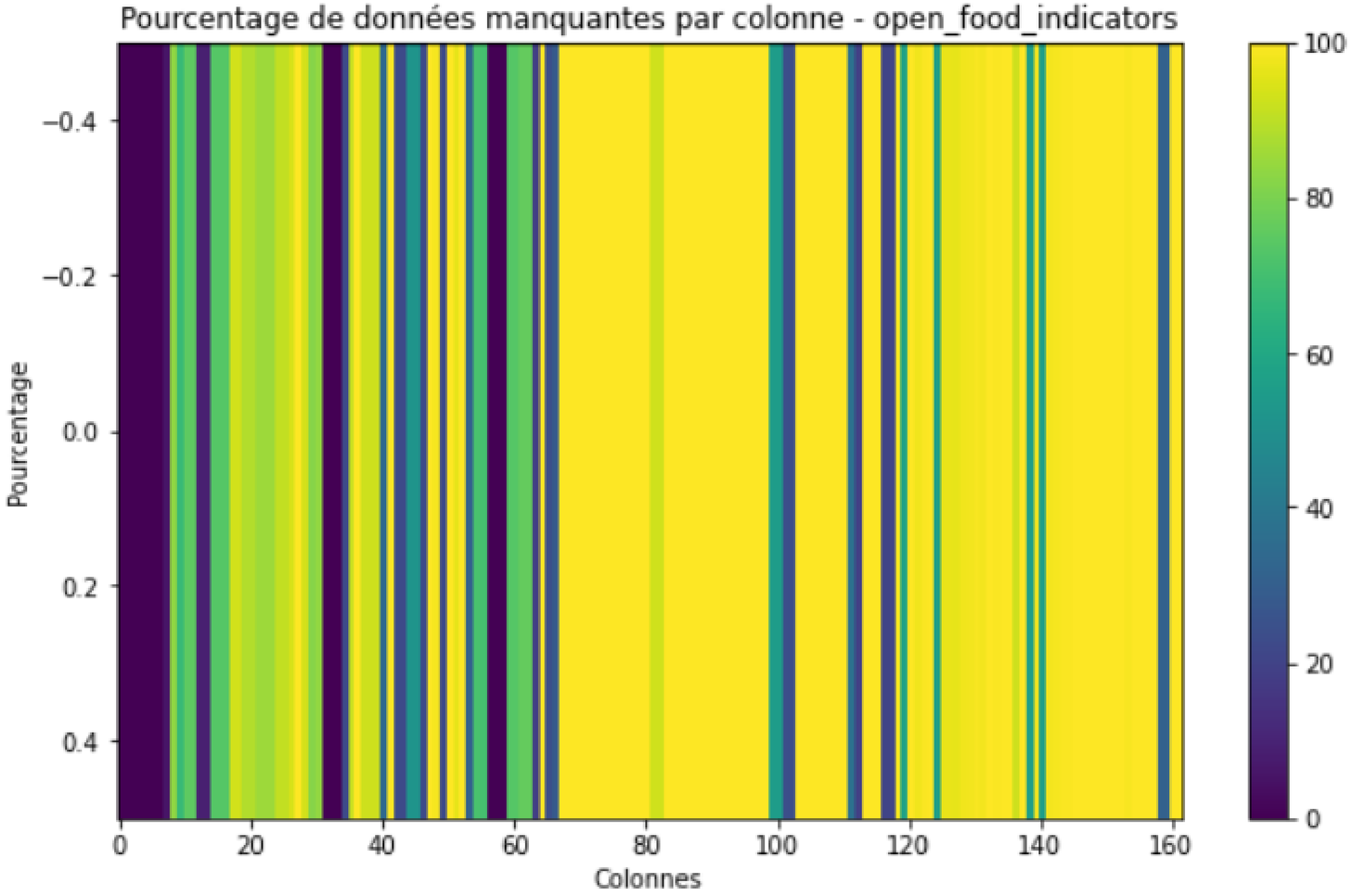
- Les individus ont des droits sur leurs données, incluant le droit d'accès, de rectification, et parfois le droit à l'effacement

Les données à disposition

Un ensemble de données concernant des informations sur des produits alimentaires :

- Des informations générales tel que le nom
- Des tags (catégorie du produit,localisation etc)
- Les ingrédients et additifs
- Les informations nutritionnels

La qualité des données



Nombres de lignes : 320 772
Nombres de colonnes : 162

Enormément de valeurs sont manquantes dans le dataframe principal

76,22%

Des valeurs dupliquées existent, notamment dans la colonne concernant les noms des produits

Selection des données

Des colonnes completes à plus de 50%
Une réelle pertinence pour notre objectif

Identification des produits

- url
- creator
- created_t
- product_name

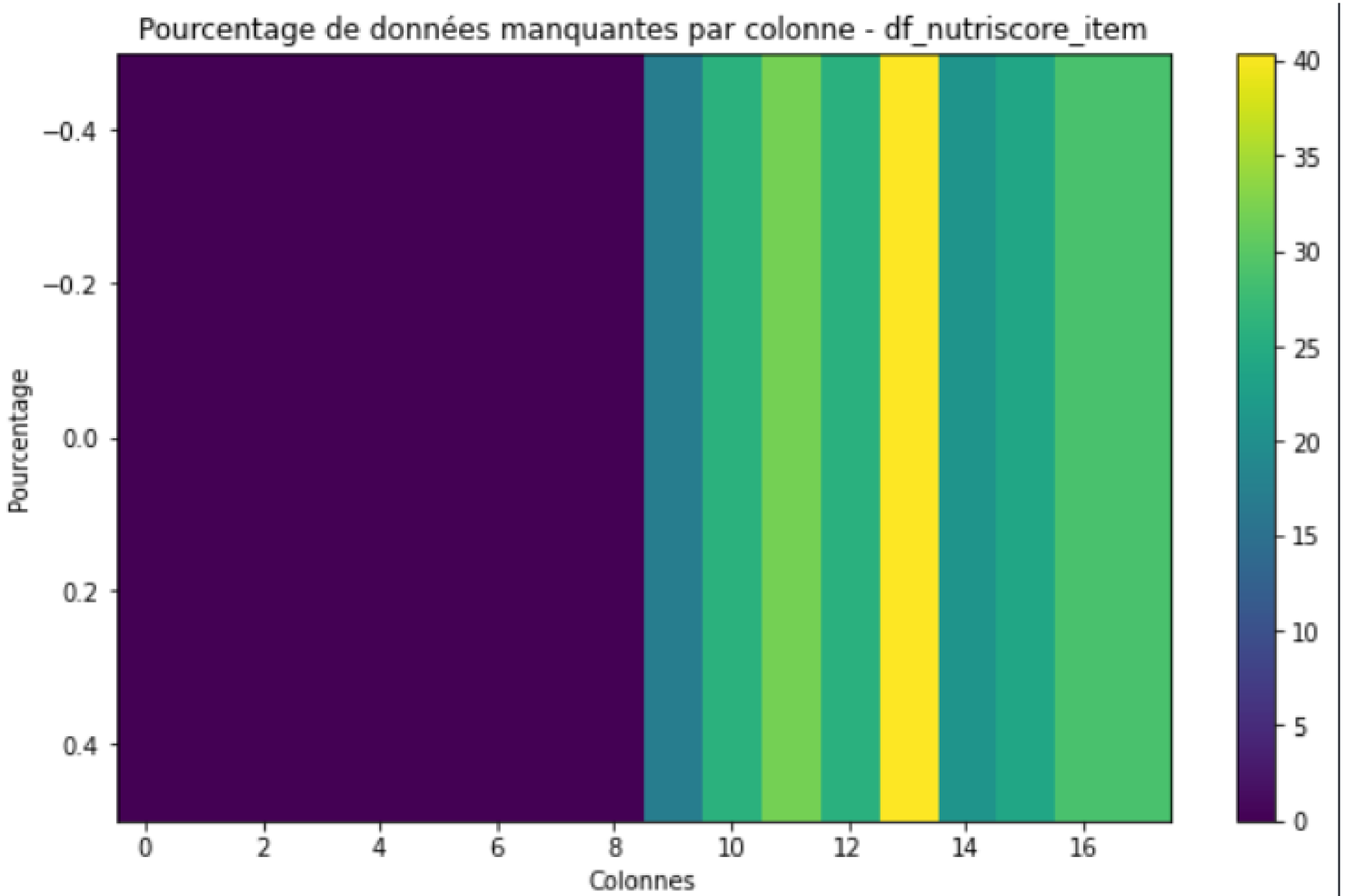
Valeurs nutritionnelles

- energy_100g
- fat_100g
- saturated-fat_100g
- sugars_100g
- fiber_100g
- proteins_100g
- salt_100g

Score nutritionel

- nutrition-score-fr_100g
- nutrition_grade_fr

Nouveau dataframe



Une fois toutes ces sélections effectuées, un nouveau dataframe est obtenu :

Nombre de lignes : 221 348
Nombre de colonnes : 18

Pourcentage du dataframe vide : 13.53%

Nettoyage des données

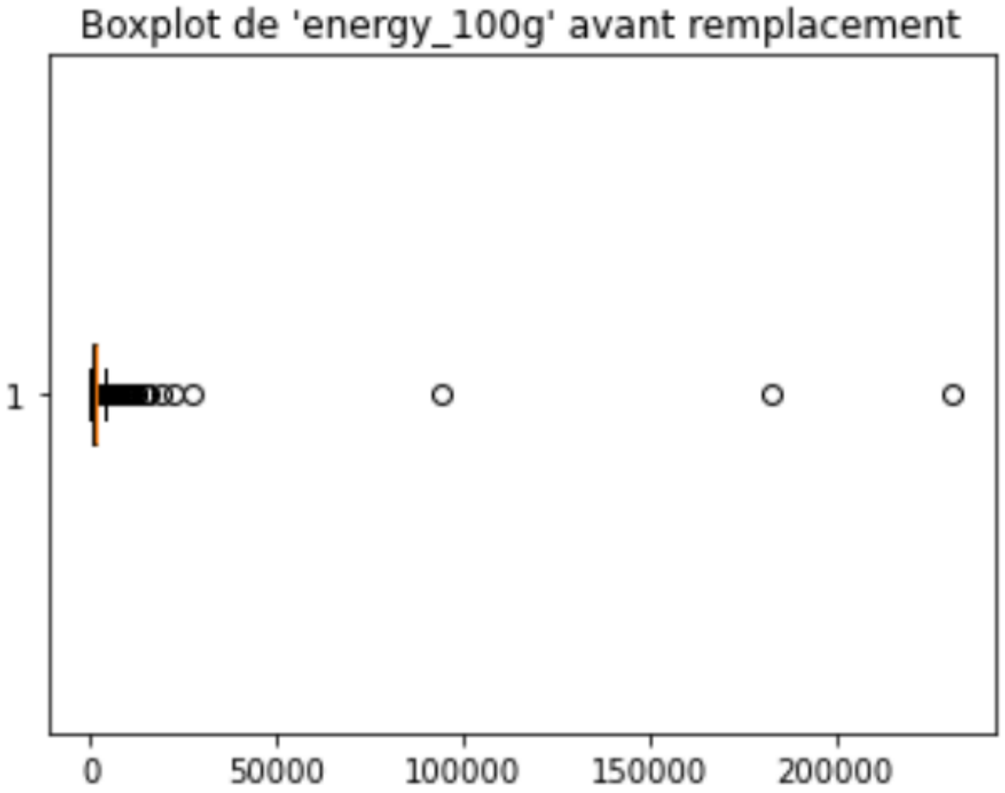
Utilisation de la méthode IQR pour détecter les outliers.
Remplacement par “NaN”

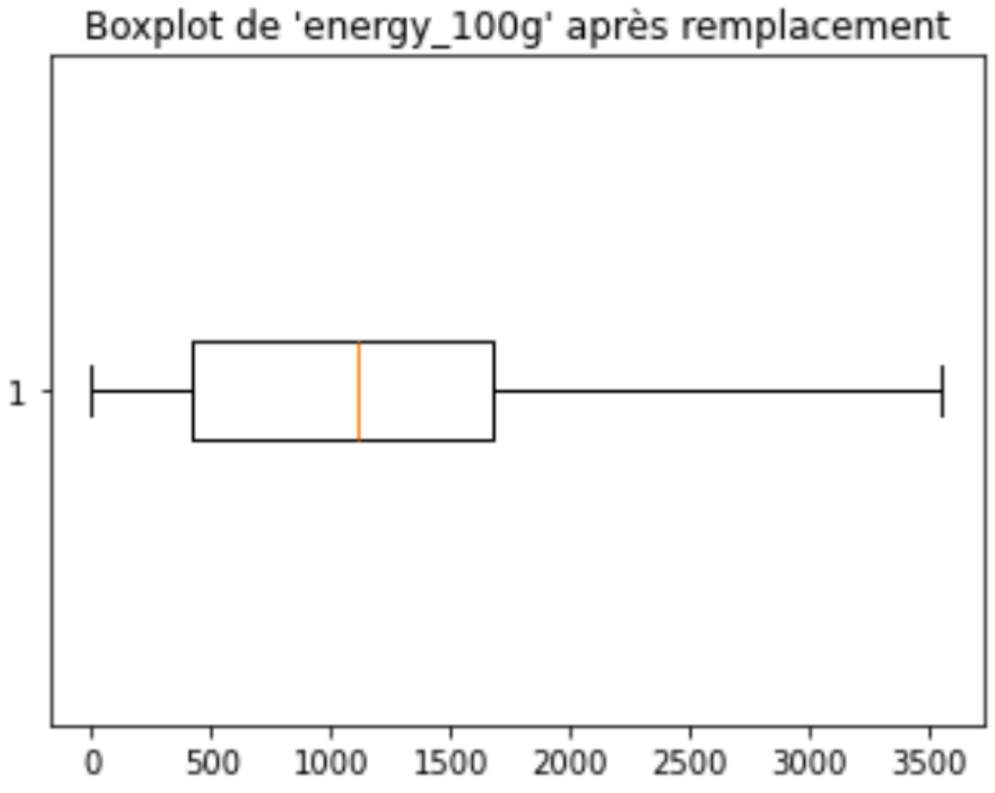
Remplacement par “NaN” des valeurs supérieurs à 100 ou inférieur à 0 pour les colonnes ou cela est justifiée

Remplacement par “NaN” quand la valeur de gras saturés est supérieur a la valeur de gras pour un aliment

Nettoyage des données

Exemple :





Imputation Moyenne

- Méthode simple à réaliser
- Préserve la moyenne globale de la variable

- Ne tiens pas compte des différences de composition selon le type d’aliments
- Sensible aux valeurs aberrantes

Imputation Médiane

- Méthode simple à réaliser
- Méthode moins sensible aux outliers

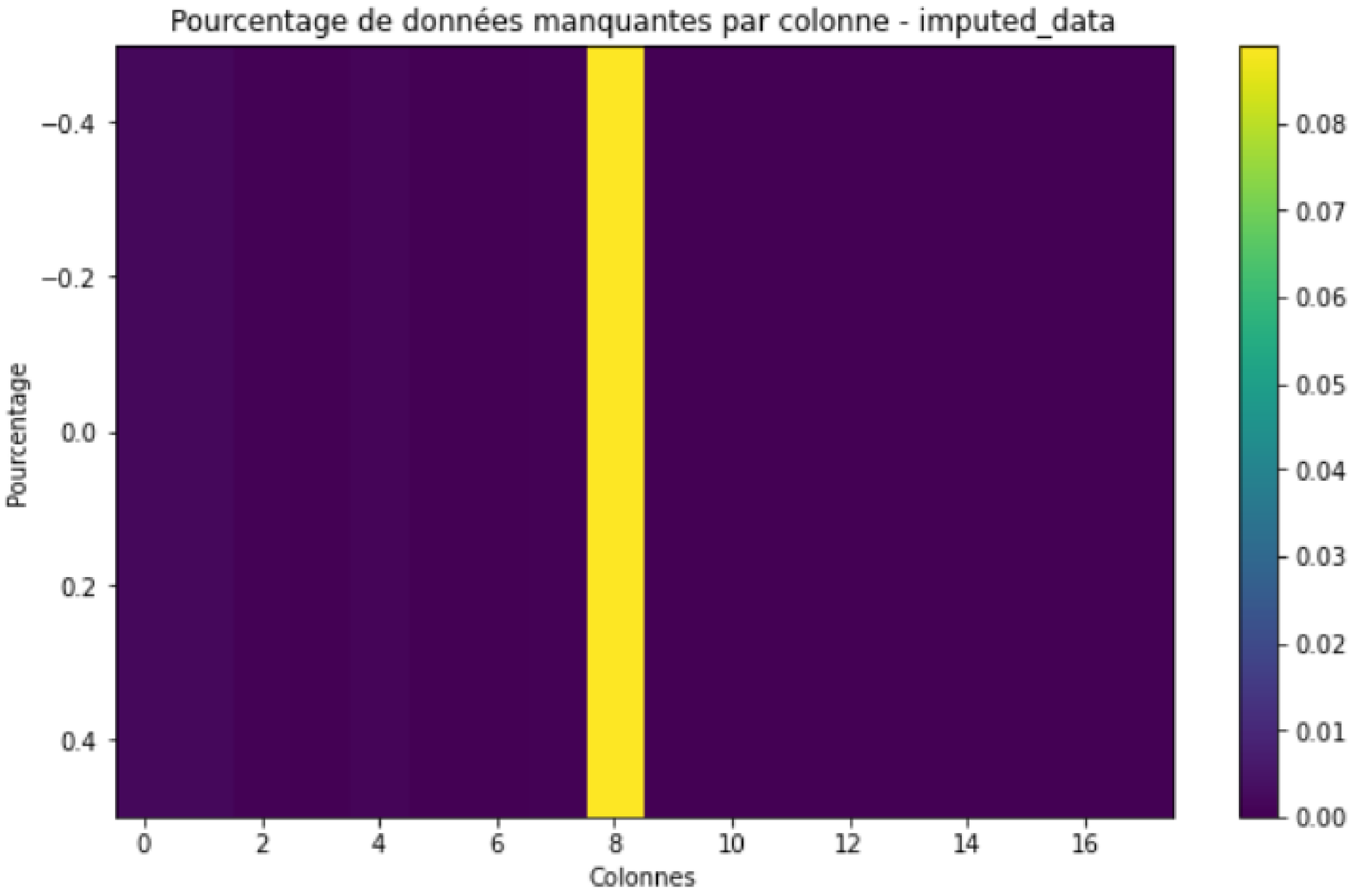
- Ne tiens pas compte des spécificités des aliments
- Peut induire des biais si la distribution entre données manquantes et données complètes est différente

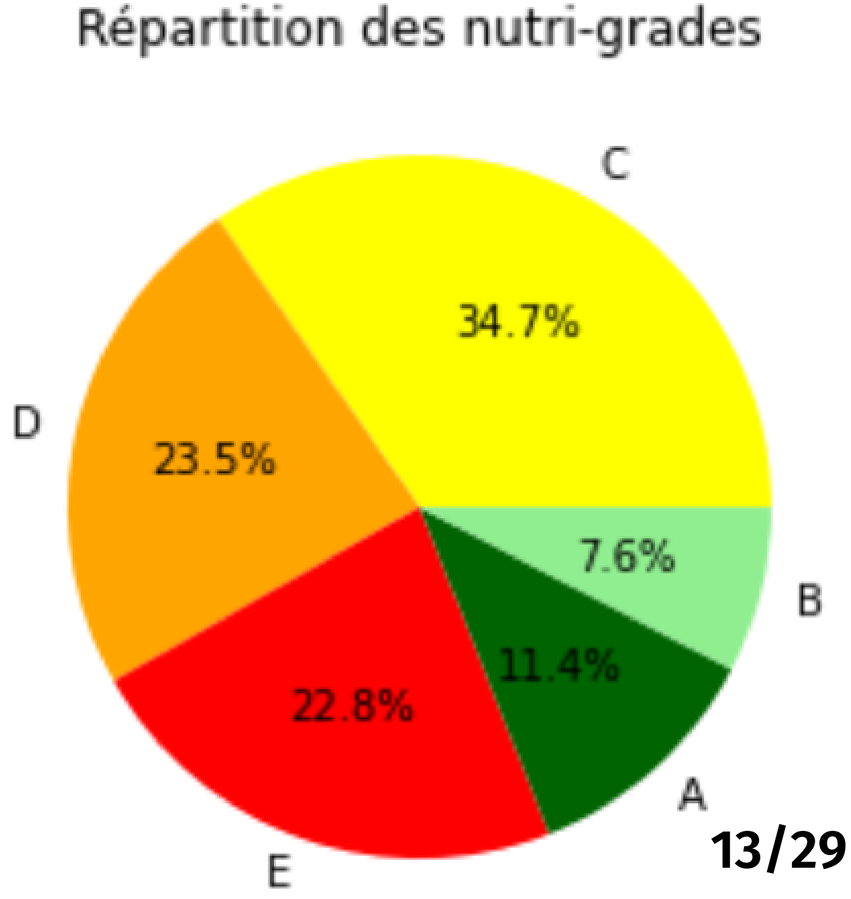
Imputation KNN

- Utilise la proximité de certains aliments pour calculer les valeurs
- Produit donc des imputations plus précises

- Performance dépendant grandement du nombre de voisin définis
- Execution longue selon la quantité de données

Dataframe final

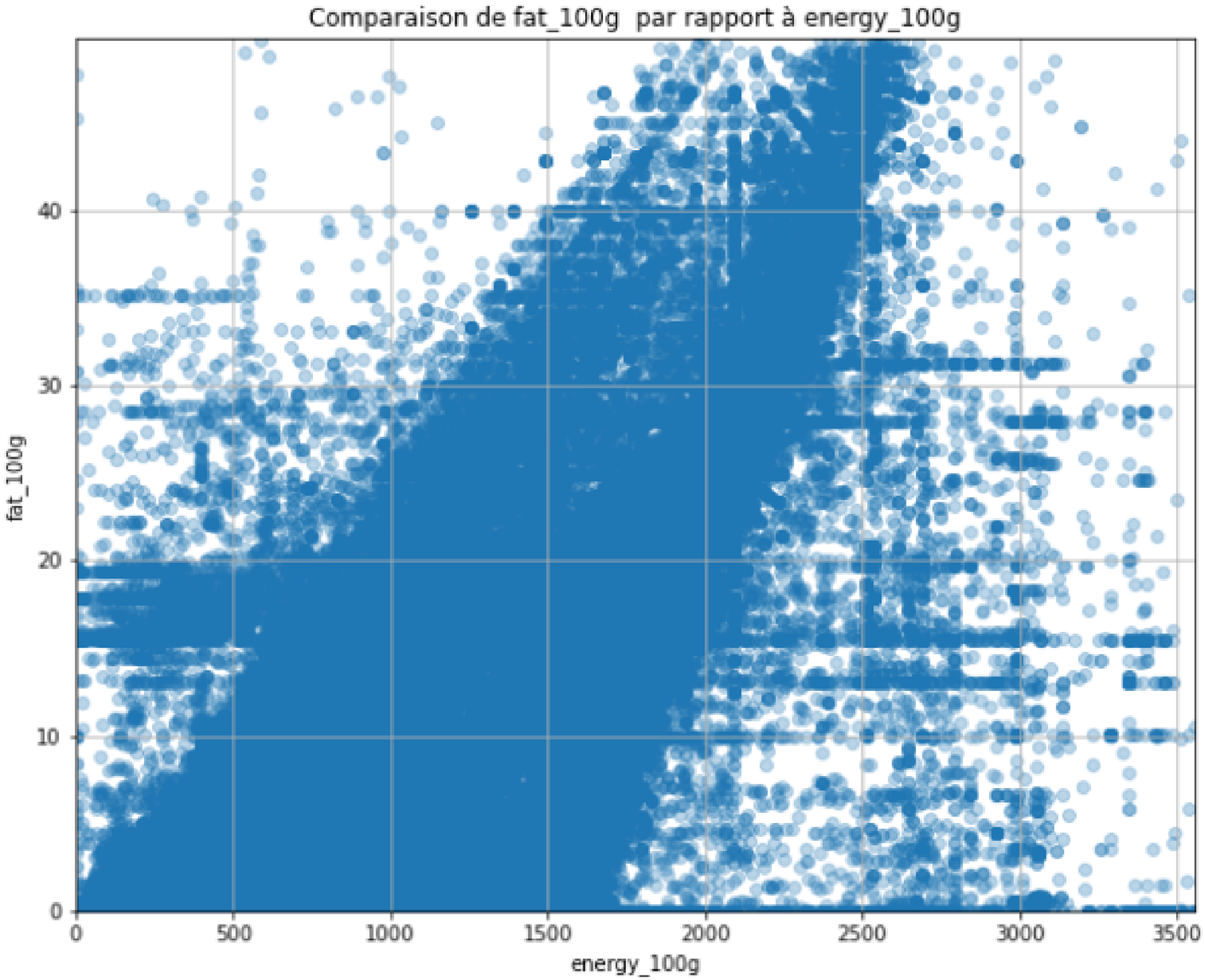




Analyse des données

Scatter plot

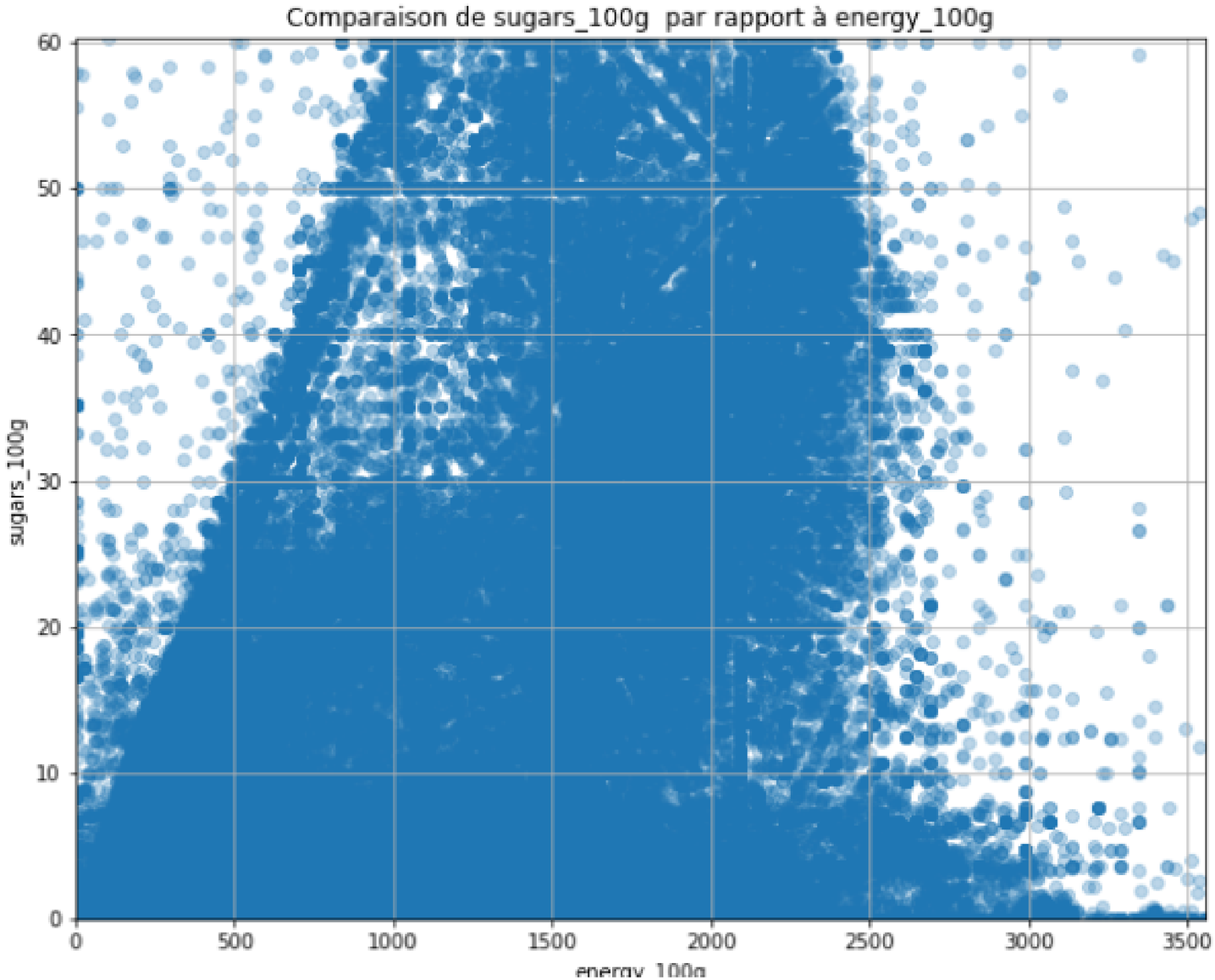
corrélation positive avec un score de pearson égal à : 0.66



Analyse des données

Scatter plot

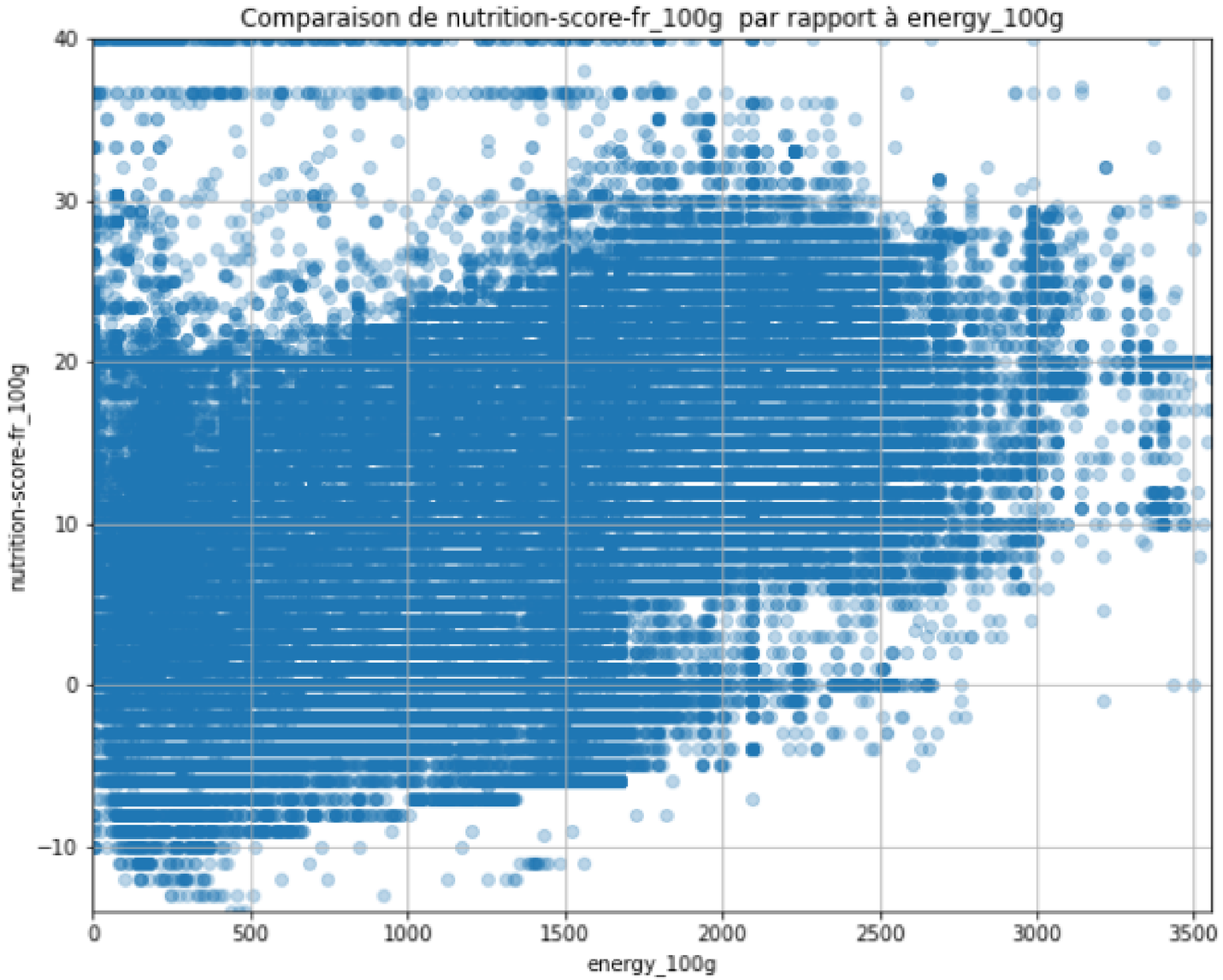
corrélation positive avec un score de pearson égal à : 0.35



Analyse des données

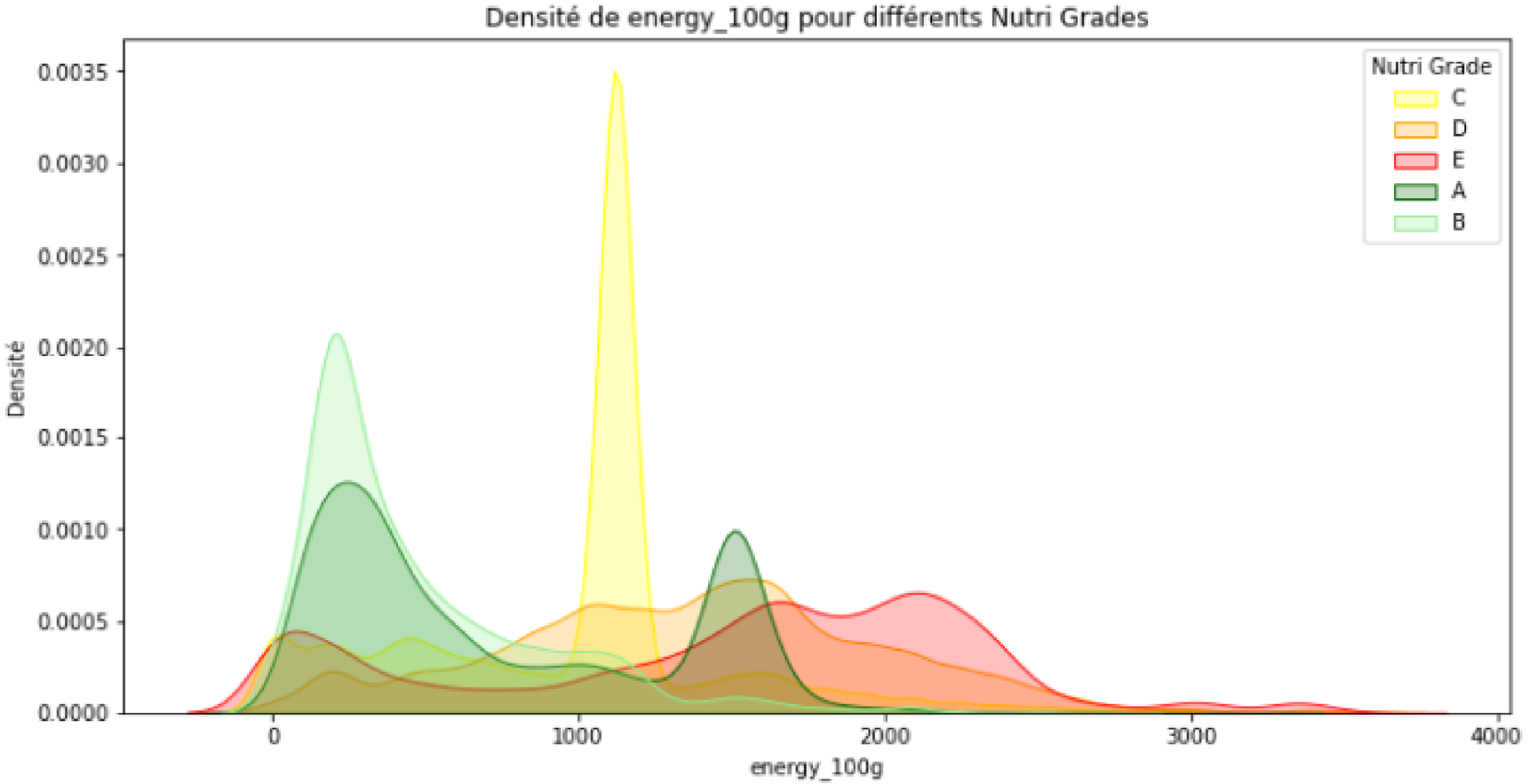
Scatter plot

corrélation positive avec un score de pearson égal à : 0.54



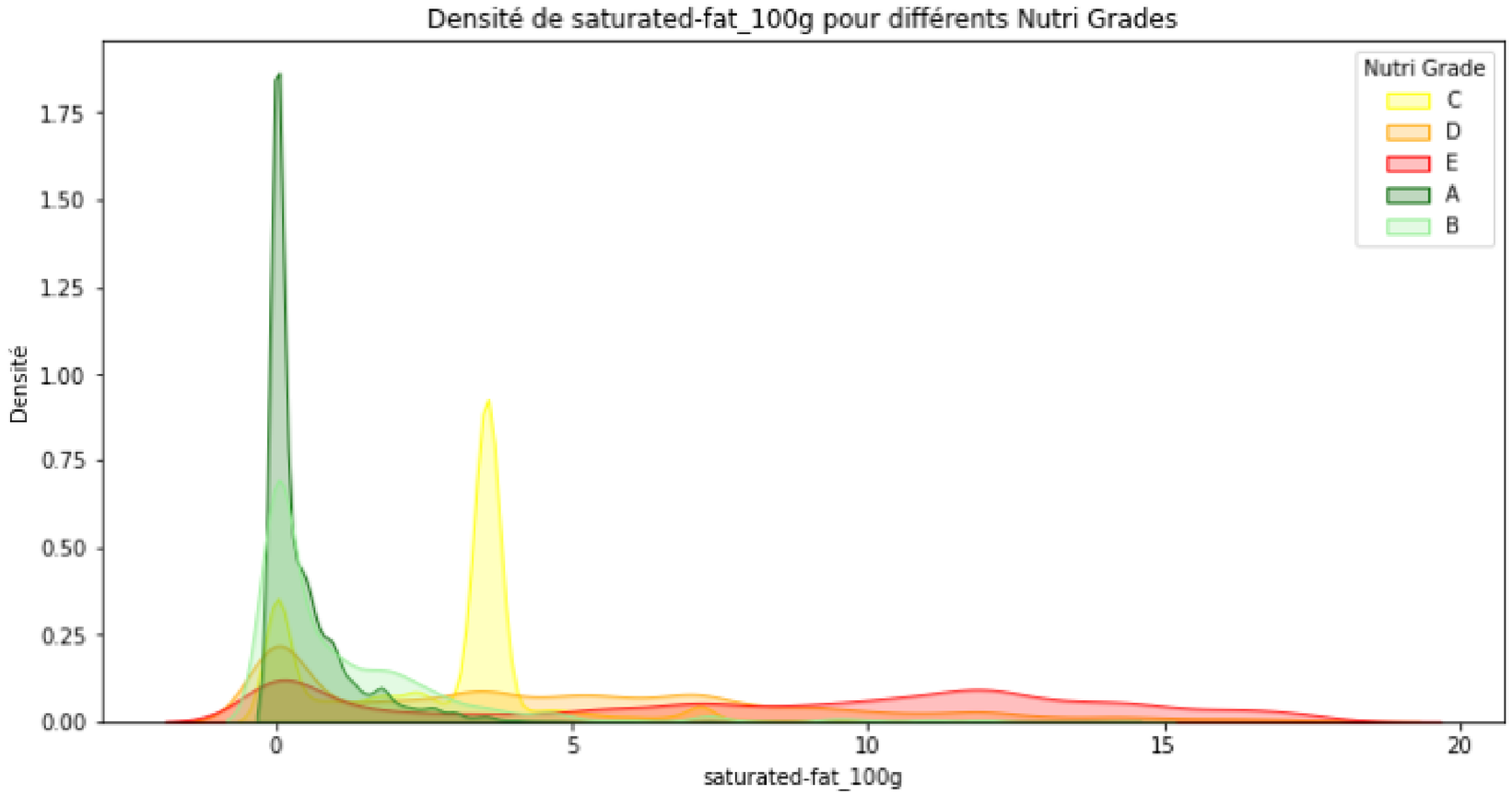
Analyse des données

Graphique de densité par grade



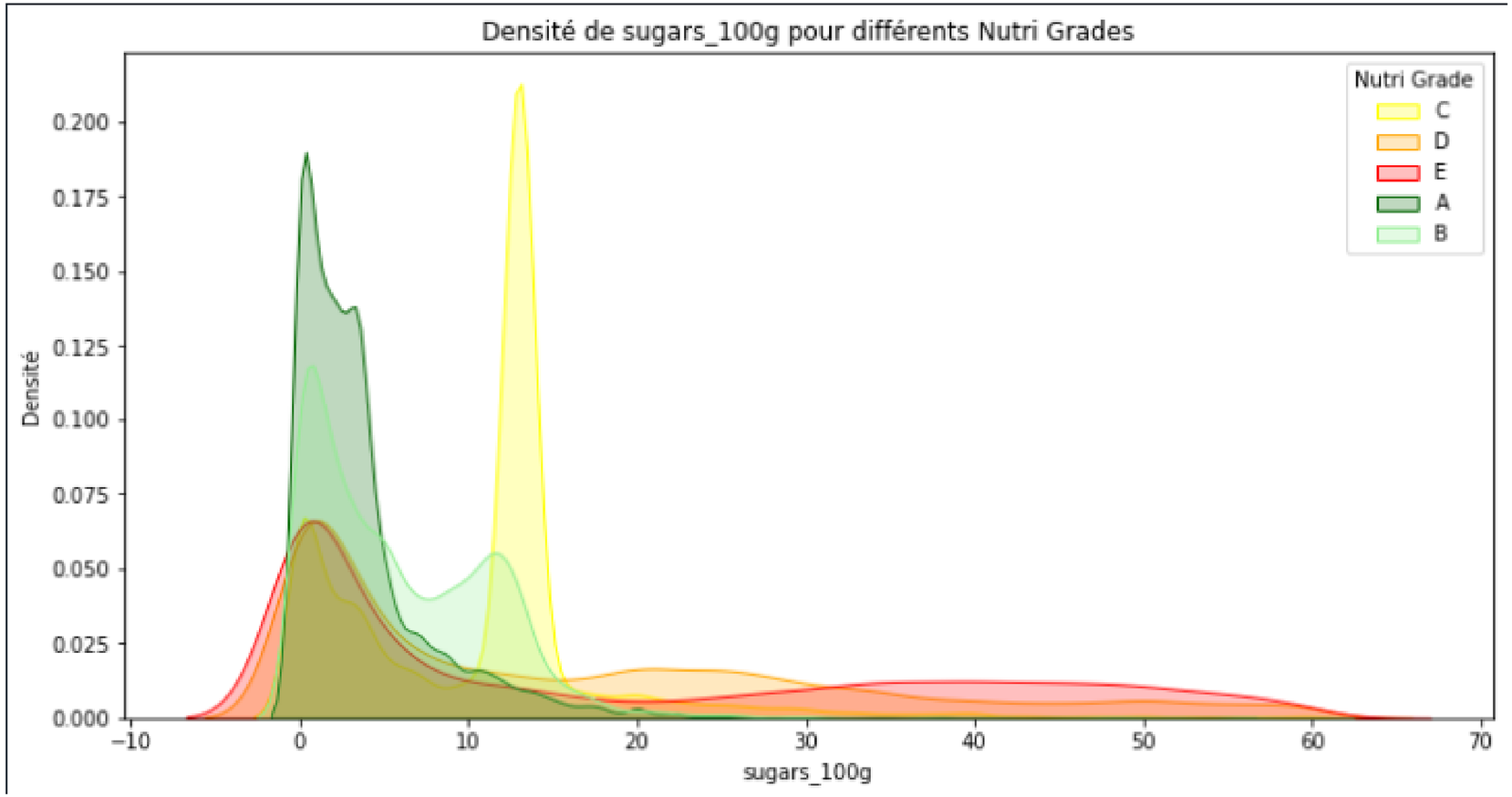
Analyse des données

Graphique de densité par grade



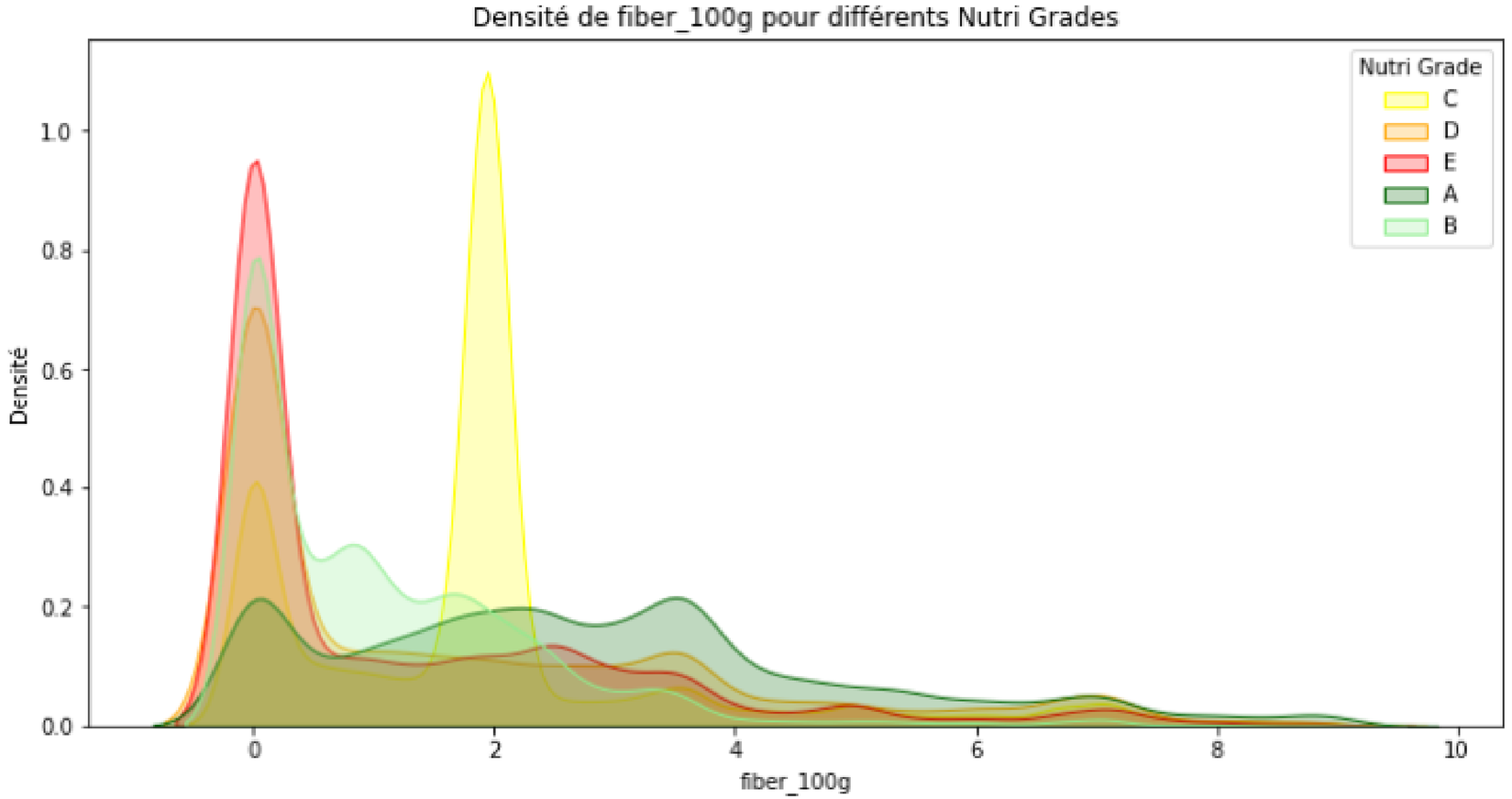
Analyse des données

Graphique de densité par grade



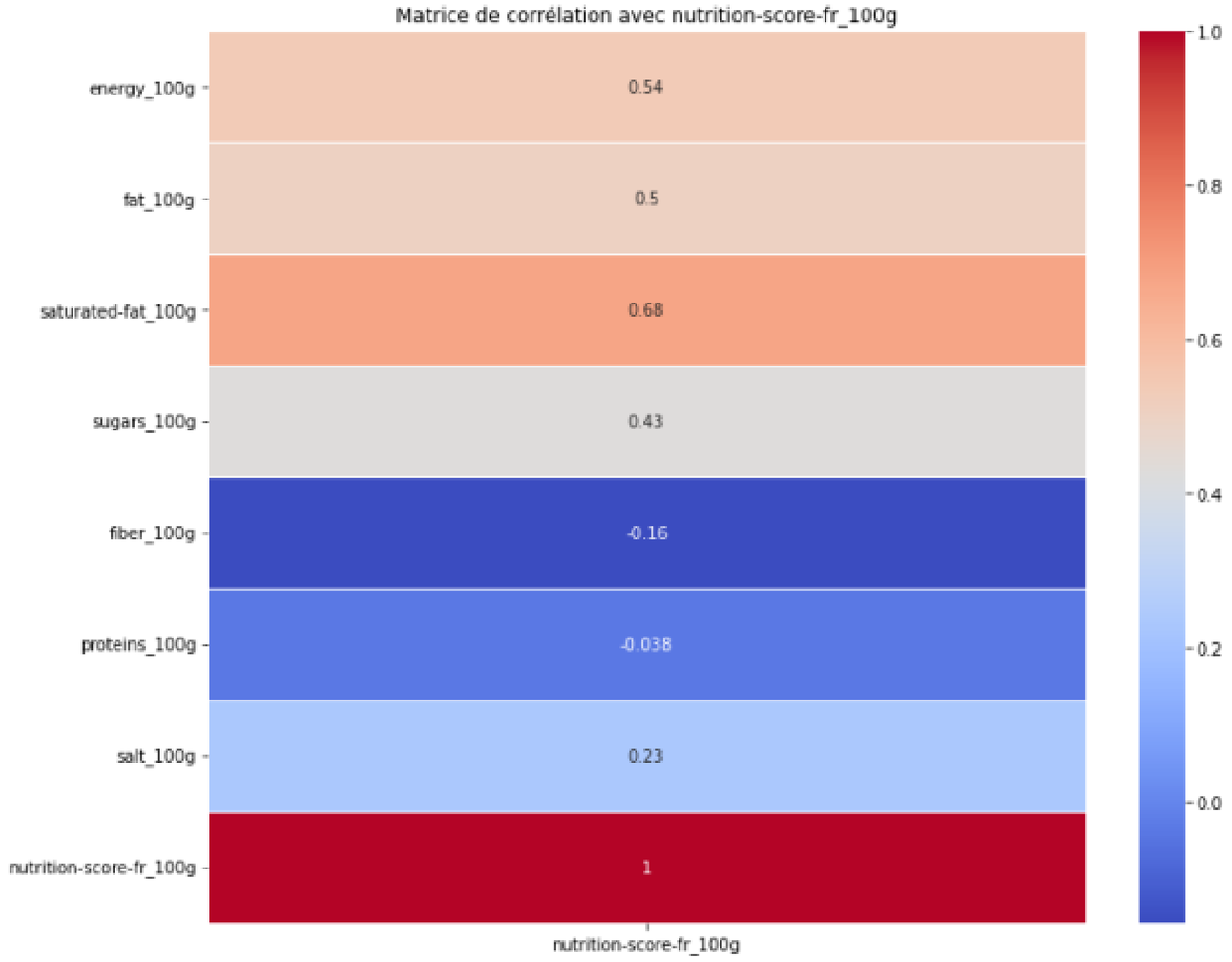
Analyse des données

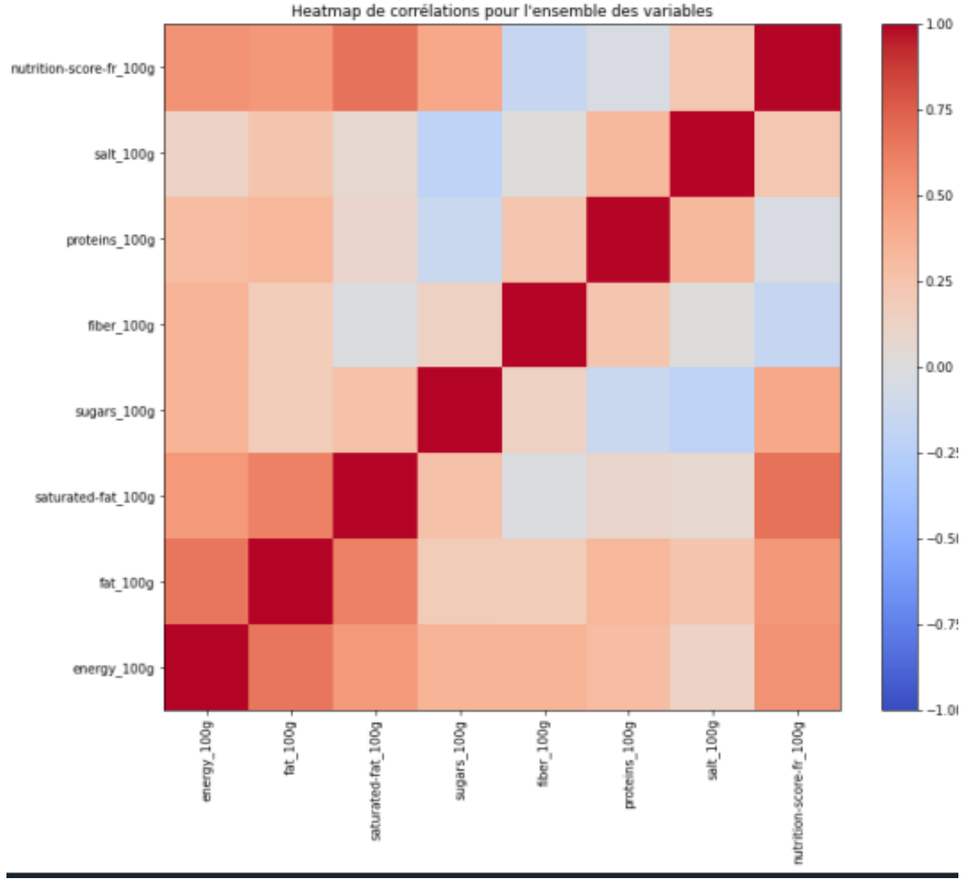
Graphique de densité par grade



Analyse des données

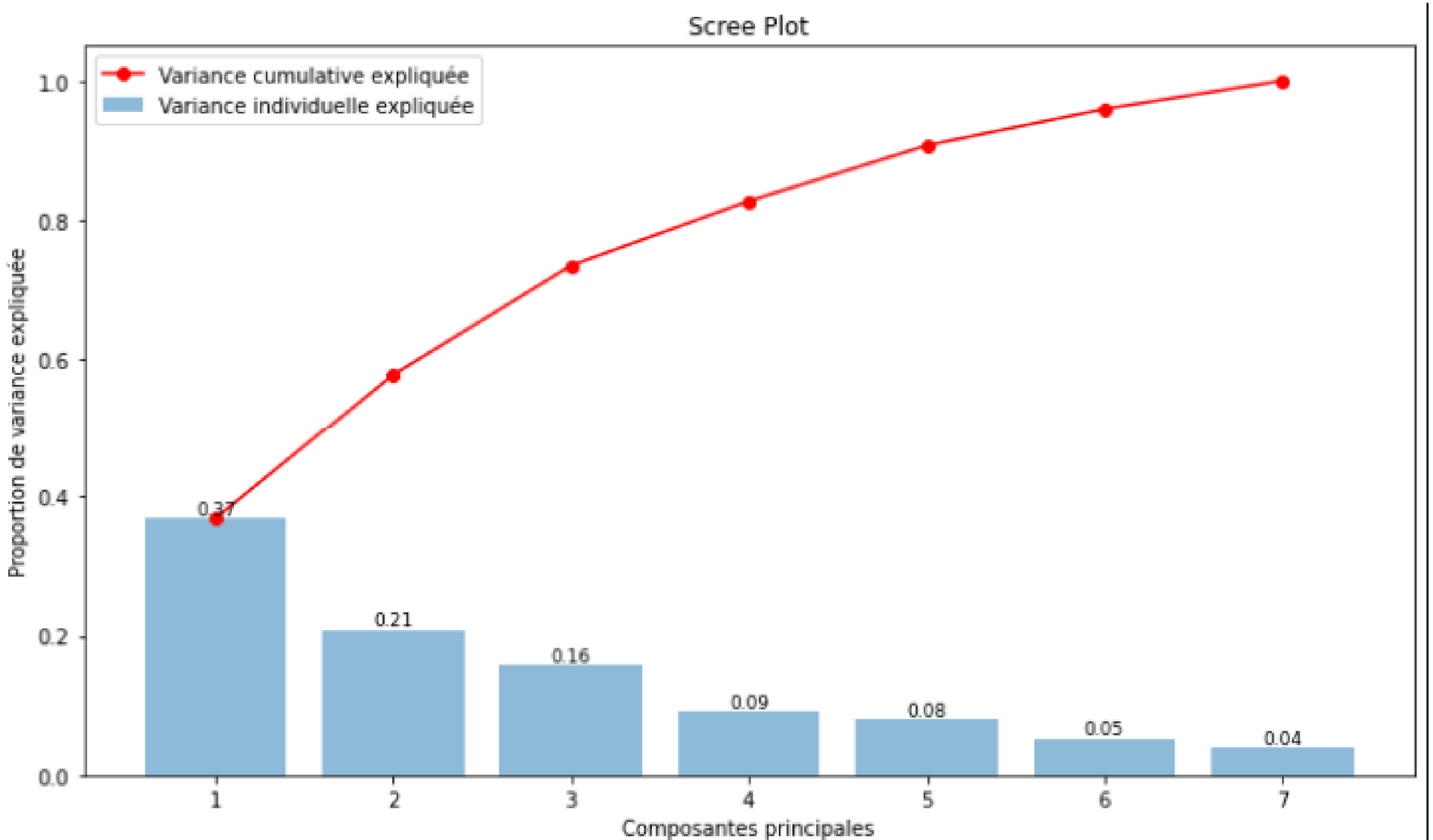
Corrélation avec nutrition score





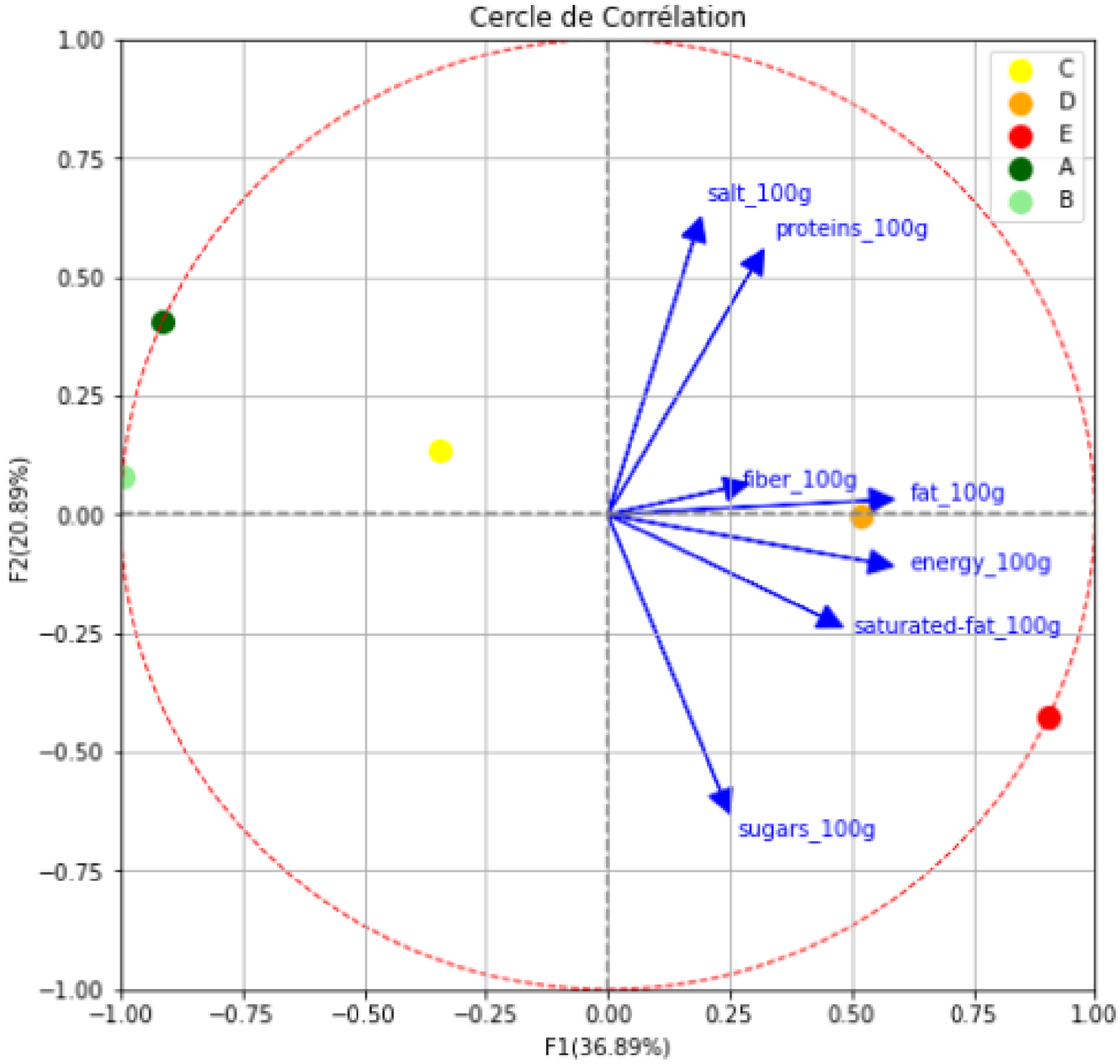
Analyse des données

Scree Plot composantes principales



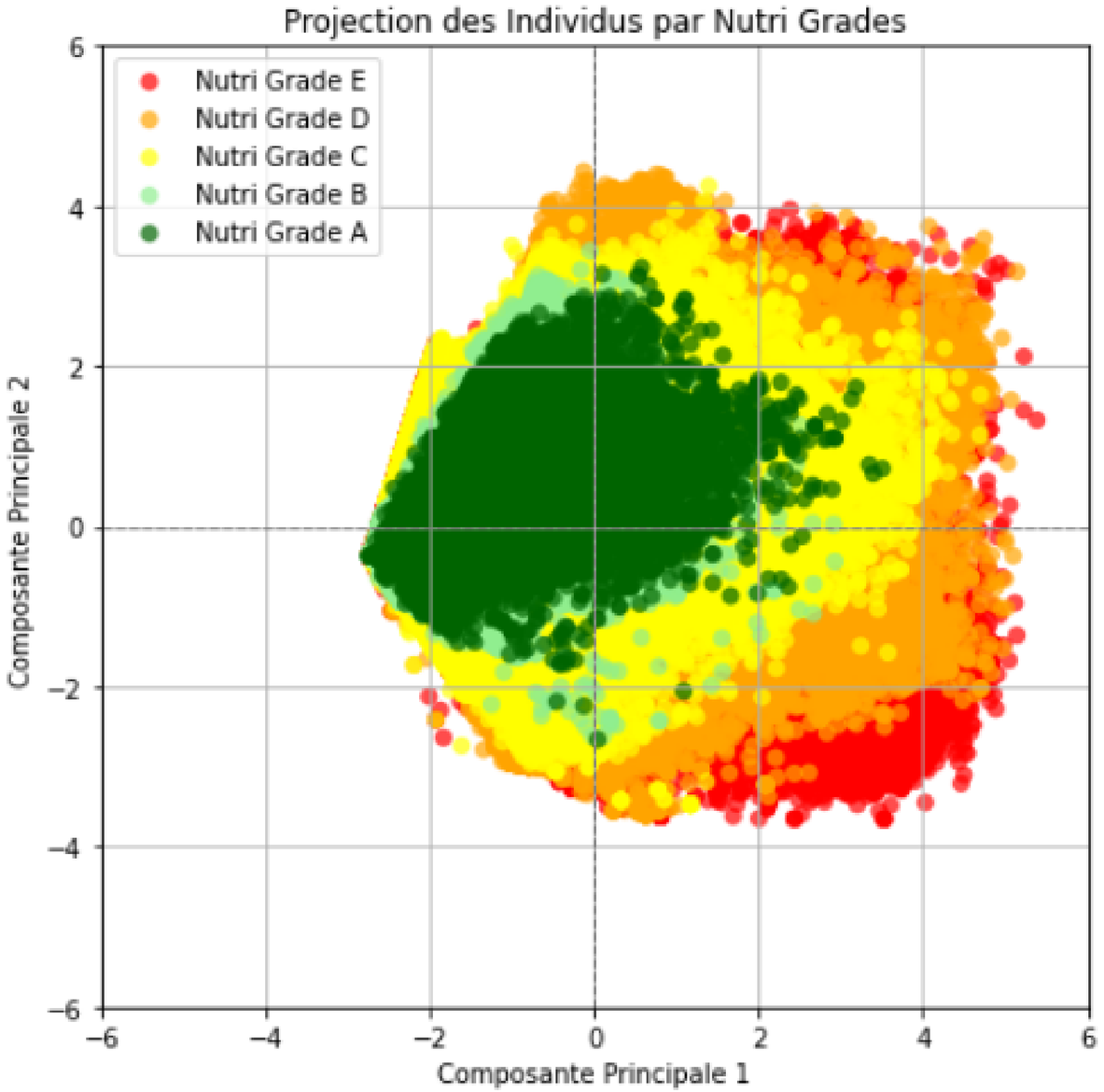
Analyse des données

ACP: cercle de corrélation



Analyse des données

Projection des individus par grade



Anova

eta²

Influence des variables étudiées sur le nutrigrade

energy : 20.54%
Fat : 17.79%
saturated-fat : 33.20%
sugars : 11.29%

fiber : 5.76%
proteins : 2.57%
salt : 5.48%

Les variables liées à la composante négative N du nutrigrade contribuent à 88,3 % au nutrigrade

Anova

H0 et H1

H0 : Les moyennes des différentes variables ne changent pas selon le nutrigrade

H1 : Les moyennes des différentes variables changent selon le nutrigrade

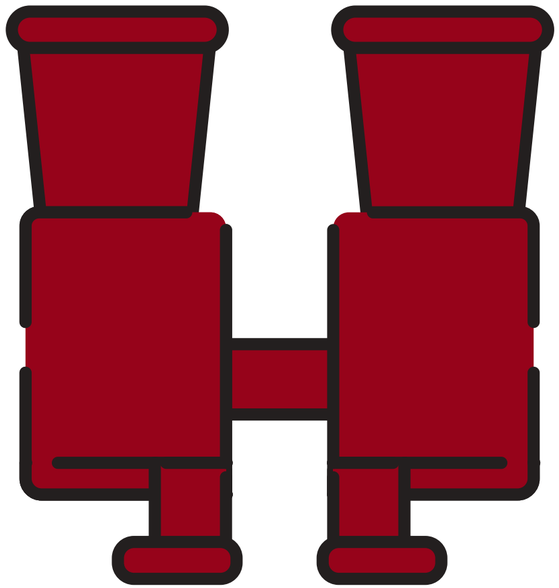
Les valeurs moyennes diffèrent selon le Nutri-Grade (H1 acceptée) avec une p value proche de 0.

Conclusion

Faciliter la complétion de la base de données par les utilisateurs en se concentrant sur les valeurs impactant négativement le nutrigrade et en utilisant le KNN pour suggérer les valeurs qui viendraient à manquer.

Déterminer le nutrigrade en se basant sur les nutriscore et en appliquant le système de correspondance score/grade pour l'auto-complétion.

Pour aller plus loin



- **Des variables plus complètes concernant les fruits, légumes et légumineuses**
- **Plus de données disponibles concernant les catégories de produits, permettant de calculer précisement le nutriscore**