1. identification des variables quantitatives et qualitatives:

. création de deux listes : une pour les variables quantitatives (avec suffisamment de diversité), une autre pour les qualitatives.

2. Analyse des valeurs manquantes:

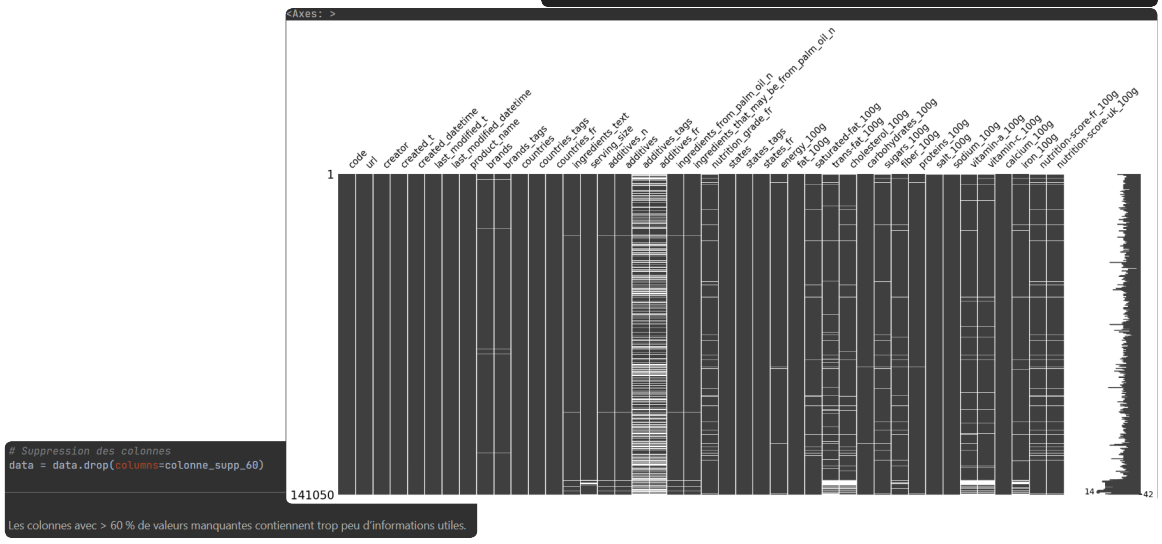
. je calcule pour chaque colonne combien de valeurs sont manquantes, et le pourcentage que cela représente. (En les triant du plus élevé au plus faible, j' obtiens une priorisation claire des colonnes problématiques.)

. je supprime toutes les colonnes ayant plus de 60% de valeurs manquantes, considérées comme trop peu informatives

Choix de la cible:

(quant) calcium_100g - 179722 valeurs manquantes, cela correspond à (56.03%) du dataset

Apport du calcium	Impact sur la qualité
Information souvent manquante	Enrichit la base de données
Non pris en compte dans le Nutri-Score	Permet une évaluation complémentaire
Essentiel à la santé	Aide à mieux informer les consommateurs
Peut différencier des produits similaires	Encourage des choix plus équilibrés




```
# Suppression des colonnes
data = data.drop(columns=colonne_supp_60)
```

Les colonnes avec > 60 % de valeurs manquantes contiennent trop peu d'informations utiles.

3. Correction des doublons (code)

. cela évite les biais lors de l'entrainement des modèles ou dans les statistiques descriptives

---

Sélectionner une **cible dès le début de la phase de nettoyage** a plusieurs intérêts stratégiques, surtout dans un projet d’analyse prédictive ou d’auto-complétion

cela nous a permis de mieux préparer les données, d’anticiper les traitements nécessaires, et d’orienter les analyses et la modélisation de manière plus efficace.

1. Orienter le nettoyage intelligemment :

- Vérifier si elle contient des valeurs manquantes, et décider de **les imputer ou les supprimer** intelligemment.
- Garder uniquement les variables explicatives **pertinentes** pour cette cible (gain de temps et de clarté).
- Appliquer des filtres logiques : par exemple, on peut dire qu’un produit avec 0 calcium et 0 fer est peut-être mal renseigné.

2. **Conserver la bonne cohérence d’échantillon :**
Si tu veux entraîner un modèle, tu as besoin d’une **cible propre et fiable**. Donc, tu peux choisir de ne garder que les lignes où la cible est présente ou crédible dès le départ.

3. **Préparer les analyses statistiques dès le nettoyage :**
Certains tests comme les corrélations ou ANOVA nécessitent que ta **variable cible soit connue**, donc tu anticipes les étapes suivantes.

4. **Faciliter l’auto-complétion en deux phases :**
Dans ton projet, tu fais d’abord un nettoyage + remplissage des données manquantes (phase 1), puis tu construis un modèle pour **prédire la cible dans une seconde phase**. Le fait d’avoir ciblé calcium_100g dès le départ t’a permis de suivre ce fil conducteur.

Variable CIBLE:

Sélectionner une **cible dès le début de la phase de nettoyage** a plusieurs intérêts stratégiques, surtout dans un projet d'analyse prédictive ou d'auto-complétion.

Pourquoi sélectionner une cible au début ?

1. Orienter le nettoyage intelligemment
connaitre notre variable cible (calcium_100g par exemple), permet de :

- Vérifier si elle contient des valeurs manquantes, et décider de **les imputer ou les supprimer** intelligemment.
- Garder uniquement les variables explicatives **pertinentes** pour cette cible (gain de temps et de clarté).
- Appliquer des filtres logiques : par exemple, on peut dire qu'un produit avec 0 calcium et 0 fer est peut-être mal renseigné.

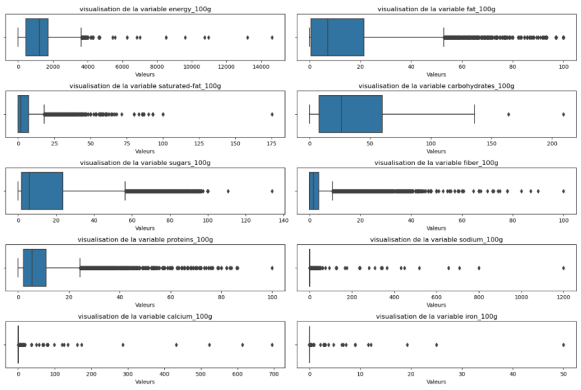
2.Conserver la bonne cohérence d'échantillon :
Si tu veux entraîner un modèle, tu as besoin d'une **cible propre et fiable**. Donc, tu peux choisir de ne garder que les lignes où la cible est présente ou crédible dès le départ.

1. Préparer les analyses statistiques dès le nettoyage :
Certains tests comme les corrélations ou ANOVA nécessitent que ta **variable cible soit connue**, donc tu anticipes les étapes suivantes.

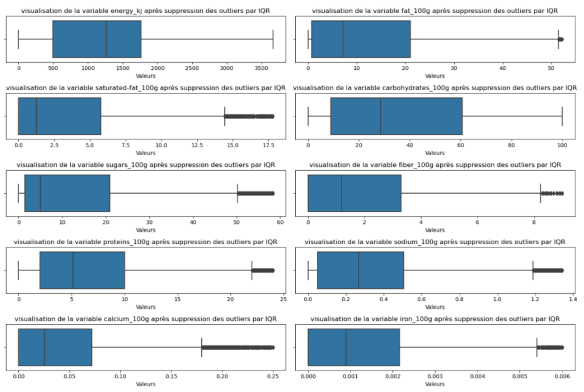
1. Faciliter l'auto-complétion en deux phases :
Dans ton projet, tu fais d'abord un nettoyage + remplissage des données manquantes (phase 1), puis tu construis un modèle pour **prédire la cible dans une seconde phase**. Le fait d'avoir ciblé calcium_100g dès le départ t'a permis de suivre ce fil conducteur.

Résumé simple pour la soutenance :

"Nous avons sélectionné dès le départ une variable cible (calcium_100g) pour structurer notre travail : cela nous a permis de mieux préparer les données, d'anticiper les traitements nécessaires, et d'orienter les analyses et la modélisation de manière plus efficace."



- Une distribution asymétrique signifie que les données sont déséquilibrées autour de la moyenne. Par exemple, certains produits ont beaucoup plus de sucre que la majorité. C'est pourquoi j'ai privilégié des méthodes robustes comme l'analyse par médiane et des scalers adaptés



1. Séléction des variables pertinentes:

. Retrait des colonnes inutiles ou peu informatives
. Conservation des variables clés, en cohérence avec les 7 éléments obligatoires du Nutri-Scores

2. remplacement des valeurs negatives:

. les valeurs numerique negatives (ex; -10g de sel) sont remplacé par NaN

3. définition de limites physiologiques plausibles:

. energy_kj (0-3700), fat_100g (0-100)...
. les valeurs hors limites sont supprimées ou transformé en NaN pour un traitement ultérieur

basées sur des connaissances nutritionnelles standards (ex : 0 à 100g de graisses, 0 à 3700 kJ d'énergie). J'ai aussi utilisé des visualisations comme les boxplots pour confirmer ces bornes. Cela m'a permis d'éliminer des erreurs extrêmes tout en conservant la cohérence des données.

4. suppression statistiques des outliers avec IQR (Interquartile Range)

. methode alternative (Z-score) mais IQR et plus fiable et robuste ici)
. on supprime les valeurs au-dela de 1.5 x IQR
. Permet de réduire les distributions très asymétriques*

Qu'est-ce que l'IQR **?**

L'**Interquartile Range (IQR)** est une méthode **statistique** pour détecter les valeurs extrêmes. Il se base sur la **distribution** des données :

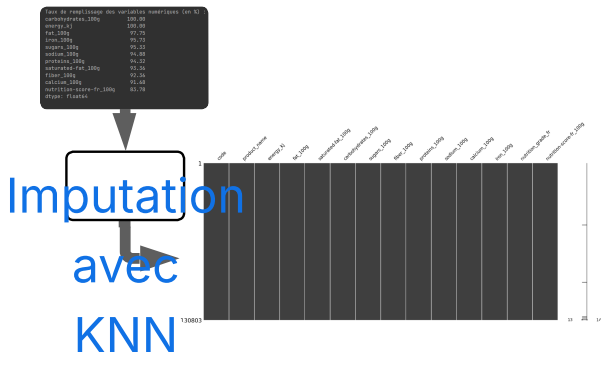
- **Q1** : 1er quartile (25 % des plus petites valeurs)
- **Q3** : 3e quartile (75 % des plus petites valeurs)
- **IQR** = Q3 – Q1 → c'est la "zone normale" où se situent la majorité des données.
-

➕ ➖ **Définir des *bornes* :**

- **Borne basse** = Q1 – 1.5 × IQR
- **Borne haute** = Q3 + 1.5 × IQR

Toutes les valeurs **en dehors de ces bornes** sont considérées comme des **outliers statistiques**.

1. Traitement des valeurs manquantes numériques:



. utilisation de l'algorithme KNNimputer (k=5 par défaut)

. le principe consiste a combler les valeurs manquantes en m'appuyant sur les produits les plus proches en termes de valeurs nutritionnelles

. certaines valeurs sont manquantes dans nutritions_grade_fr, une variable cible importante

2. Methode:

. Entrainement d'un modèle de régression logistique sur les produits avec une note connue

. prédiction des notes pour les produits sans note

. Remplacement des NaN par la valeur prédite

3. Résultat:

. la colonne nutritions_grade_fr ne contient plus aucune valeur manquantes

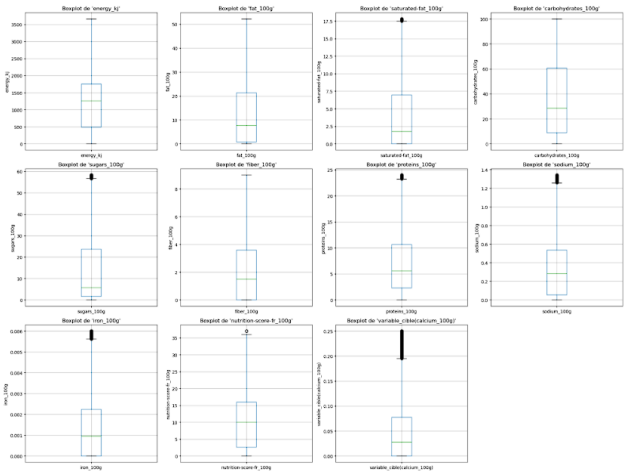
. Toutes les variables numériques sont complète à (quasi) 100%

Data set final:

tailles du data set:	130 803 lignes, 14 colonnes
valeurs manquantes:	0 (numérique et cible)
outliers:	supprimés via IQR
doublons:	supprimés
variables inutiles:	supprimées (>60% manquants)
variables finales:	Conformes au nutri-scores

Mon objectif était de nettoyer les données tout en maximisant la conservation d'information.
J'ai combiné des méthodes statistiques (IQR, KNN) et supervisées (regression logistique) pour traiter les données manquantes et aberrantes.
Le jeu de données est maintenant propre, complet et prêt à être exploité dans une analyse exploratoire ou un modèle predictif.

Étape 1 : Analyse Univariée



But :

. Étudier chaque variable (chaque nutriment) **séparément**, pour comprendre sa distribution, sa dispersion, et identifier d'éventuelles anomalies (outliers).

. Chaque variable numérique (nutriment) a été analysée individuellement à l'aide de **boxplots**.

Résultat:

. La majorité des variables présentent des asymétries positives (longue queue à droite).

. Plusieurs variables, notamment sugars_100g, proteins_100g, calcium_100g, comportent des outliers clairement visibles. (typiques des aliments ultra-transformés.)

Utilité :

. Cela justifie des traitements complémentaires comme la **normalisation** ou la **suppression statistique des outliers**.

V|-----|=====|=====|-----|

min Q1 Q2 Q3 max *plus d'explications sur la fiche boxanalyse

↑

médiane

. **Boîte (rectangle) :**

Délimite les quartiles Q1 (25%) et Q3 (75%).

. La ligne dans la boîte est la **médiane** (Q2, 50%).

. "**Moustaches**" (traits fins verticaux) :

S'étendent jusqu'aux valeurs minimales et maximales non aberrantes.

. **Points noirs / cercles au-dessus** :

Ce sont les valeurs aberrantes (outliers) détectées au-delà de 1.5 × IQR (interquartile range).

Boxplots Analyse:

energy_kj

-
- Distribution **assez étendue**, médiane vers ~1500 kJ.
-
- Quelques outliers très hauts (> 3500 kJ).

🔷 fat_100g / saturated-fat_100g

-
- Médiane faible (autour de 5-10g).
-
- **Asymétrie** vers la droite (valeurs élevées).
-
- **Beaucoup d'outliers** : produits très gras.

🔷 carbohydrates_100g

-
- Grande variabilité : médiane vers 25g.
-
- Produits très sucrés identifiés comme outliers > 100g.

🔷 sugars_100g

-
- **Extrême asymétrie** : médiane basse (~5g), mais des valeurs montent jusqu'à 60g.
-
- Beaucoup d'**outliers** → produits très sucrés (bonbons, sodas...).

🔷 fiber_100g

-
- Médiane basse (~2g).
-
- Valeurs max à ~9g → outliers limités, bonne répartition.

🔷 proteins_100g

-
- Médiane modérée (~6g), certaines valeurs jusqu'à 25g.
-
- Présence d'**outliers** pour les produits très protéinés.

🔷 sodium_100g

-
- La majorité des produits en dessous de 0.5g.
-
- Quelques très salés jusqu'à 1.4g → outliers logiques (sauces, fromages...).

🔷 iron_100g

-
- Très **faibles valeurs** en général (< 0.002g).
-
- Quelques **outliers légers**, mais tout reste dans un ordre de grandeur réaliste.

🔷 nutrition-score-fr_100g

-
- Médiane autour de 10.
-
- Outliers importants (>35) : probablement des produits ultra-transformés.

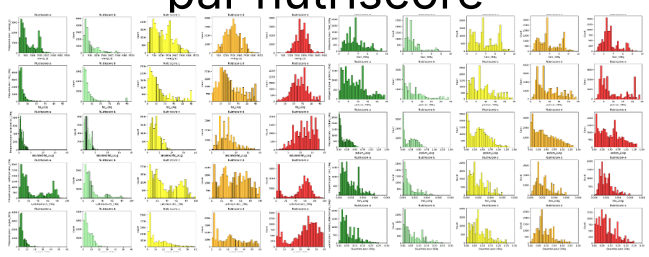
🔷 variable_cible(calcium_100g)

-
- La médiane est très faible (~0.02), mais des outliers vont jusqu'à 0.25.
-
- **Très forte asymétrie**, comme attendu pour des données de type minéral dans les aliments.

conclusion:

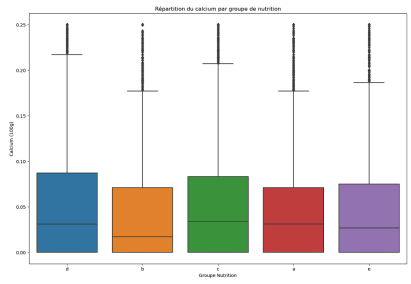
. Les Nutri-Scores A et B sont associés à des produits plus sains, riches en fibres et modérés en énergie, graisses saturées, sucres, et sodium.

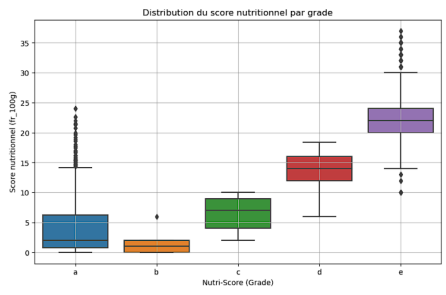
Distribution des nutriments par nutriscore



. Les Nutri-Scores D et E contiennent des produits plus riches en énergie, graisses saturées, sucres, et sodium, avec moins de fibres, indiquant une moins bonne qualité nutritionnelle.

Calcium (calcium_100g) : Les niveaux de calcium restent relativement constants et similaires à travers les différents groupes de Nutri-Score





Sucres (sugars_100g) : Les sucres suivent une tendance similaire aux glucides, avec des niveaux beaucoup plus élevés dans les Nutri-Scores D et E, et les niveaux les plus bas dans le Nutri-Score A.

on réalisé une analyse bivariée pour étudier les relations entre différentes paires de features:

Conclusion pour notre cible le calcium :

. la répartition du calcium ne semble pas avoir de corrélation forte avec les groupes de Nutri-Score

. La médiane du calcium est relativement similaire à travers les différents groupes de Nutri-Score, indiquant que le calcium n'est pas un facteur déterminant dans la classification du Nutri-Score

. Cependant, il existe une variabilité importante dans chaque groupe, avec des outliers significatifs, ce qui montre que certains produits, même dans les groupes moins sains (comme D et E), peuvent avoir des niveaux élevés de calcium.

Conclusion pour le score nutritionnel :

On observe une progression nette et croissante du score nutritionnel en allant du grade a vers e :

Ordre croissant du score: La classification a → e est cohérente avec les scores numériques. Bonne séparation des boîtes: Le score nutrition-score-fr_100g est pertinent pour prédire le grade Nutri-Score. Outliers pour a et e: Certains produits sont atypiques, malgré leur note.

Analyse ANOVA : Score nutritionnel par grade Nutri-Score:

Élément Valeur observée Interprétation
F-statistic 288 993.01 Très élevé → il y a des différences nettes entre les groupes
p-value (PR(>F)) 0.0 Très significatif → on rejette l'hypothèse nulle (pas de différence entre groupes)

Conclusion:

Le grade nutritionnel (A, B, C, D, E) influence significativement le score nutritionnel continu.
Le Nutri-Score est bien aligné avec les scores calculés
Pour confirmer ce qu'on avait visuellement observé avec les boxplots, j'ai utilisé un test statistique appelé ANOVA. Ce test permet de voir s'il y a des différences entre plusieurs groupes – ici les 5 grades A à E. Le résultat était très significatif, ce qui confirme que le Nutri-Score catégorise bien les produits selon leur score nutritionnel. Il y a donc une vraie cohérence entre la version "lettre" (grade) et la version "chiffre" (score). »

Vérifier si la variable cible calcium_100g suit une distribution normale, condition nécessaire pour utiliser des tests paramétriques comme l'ANOVA

La variable cible ne suit pas une distribution normale. Lorsqu'une variable ne suit pas une distribution normale, l'utilisation de tests statistiques qui supposent la normalité, comme l'ANOVA pour comparer les moyennes de plusieurs groupes, n'est généralement pas appropriée.

Mais le test de Shapiro-Wilk est généralement limité à des échantillons de taille plus réduite, typiquement jusqu'à 5 000 observations. Nous allons utilisé un autre test afin d'avoir une seconde confirmation sur la distribution de nos données

Comparer les médianes de la variable cible calcium_100g entre les différents groupes de nutrition_grade_fr (a, b, c, d, e).

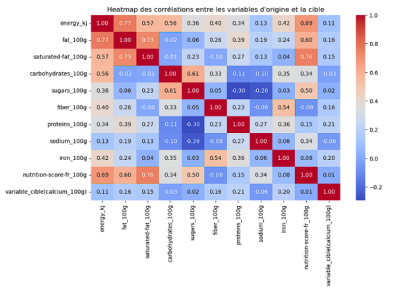
Conclusion Le test de Kruskal-Wallis révèle des différences significatives entre les médianes de la variable cible pour différents groupes. Cela suggère que la qualité nutritionnelle a un impact notable sur la variable cible, qui pourrait être le calcium ou un autre nutriment spécifique.
Pourquoi utiliser Kruskal-Wallis plutôt qu'ANOVA ? Shapiro-Wilk a montré que la distribution de la variable cible ne suit pas une loi normale (p-value ≪ 0.05).
L'ANOVA repose sur l'hypothèse de normalité ⇒ pas adaptée ici.
Le test de Kruskal-Wallis est une alternative non paramétrique à l'ANOVA :
Il ne suppose pas la normalité ;
Il compare les médianes entre plusieurs groupes indépendants.

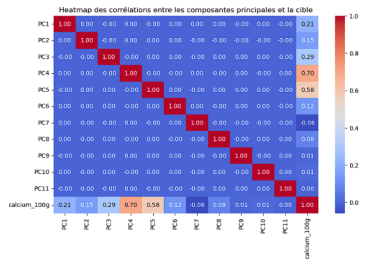
- **Test de Shapiro-Wilk** : a confirmé que le calcium ne suit pas une distribution normale.
-
- **Kruskal-Wallis** : utilisé à la place de l'ANOVA pour comparer les groupes de Nutri-Score (car non-normalité).
-
- **ANOVA** : utilisé entre nutrition-score-fr_100g (quantitatif) et nutrition_grade_fr (qualitatif), a confirmé une **forte cohérence** entre score et grade.

Étape 3 : Analyse Multivariée
But :
Étudier les **interactions entre plusieurs variables** en même temps.
Méthode : Analyse en Composantes Principales (PCA)
- Sert à **réduire la dimension** du dataset
- Permet de **visualiser les regroupements** de produits selon les nutriments

Résultat :
- Les premières composantes PC1, PC3, PC4, PC5 expliquent une grande part de la variance.
- Les produits se regroupent visuellement selon leur Nutri-Score sur les composantes principales.

Conclusion :
La PCA aide à identifier des **profils de produits**, ce qui peut alimenter une **suggestion intelligente** dans l'application.





1. Heatmap des corrélations complètes (variables d'origine + cible)

- **Objectif :** Visualiser toutes les corrélations linéaires entre les variables nutritionnelles et la variable cible calcium_100g.
  - La diagonale est toujours à 1 (autocorrélation).
  - Corrélations fortes entre :
    - energy_kj et fat_100g (0.77)
    - fat_100g et saturated_fat_100g (0.73)
    - sugars_100g et carbohydrates_100g (0.61)
    - nutrition-score-fr_100g et saturated_fat_100g (0.76)
  - **La variable calcium_100g est faiblement corrélée** à toutes les autres variables (correlations ≤ 0.21).

2. Heatmap des corrélations fortes (seuil ≥ 0.3)

- **Objectif :** Filtrer et n'afficher que les corrélations significatives.
  - On retrouve les mêmes liens forts que précédemment.
  - La majorité des variables ont des corrélations > 0.3 entre elles, en particulier celles liées à l'énergie, aux matières grasses, et au score nutritionnel.
  - La cible calcium_100g n'a **aucune** corrélation >= 0.3 ⇒ **prédiction difficile sans transformation**.

3. Heatmap des corrélations entre PCA (PC1 à PC11) et calcium_100g

- **Objectif :** Identifier si des **composantes principales** (issues d'une PCA) capturent mieux la variance de calcium_100g.
  - PC4, PC5, PC3 et PC1 ont des corrélations **modérées à fortes** avec la variable cible :
    - PC4 → 0.70
    - PC5 → 0.58
    - PC3 → 0.29
    - PC1 → 0.21
  - Cela montre que la **PCA améliore la capacité à capturer l'information utile pour le calcium**.

4. Heatmap des corrélations fortes PCA (seuil ≥ 0.2)

- **Objectif :** Épurer la matrice pour ne conserver que les **composantes pertinentes pour calcium_100g**.
  - La cible est bien expliquée par **PC4 (0.70)** et **PC5 (0.58)**.
  - Ces composantes sont donc intéressantes à intégrer dans une modélisation.

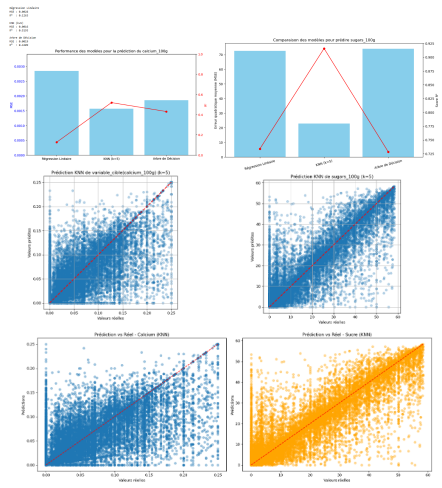
Conclusion générale :

1. Corrélations faibles entre calcium_100g et les variables brutes.
- Cela explique pourquoi les modèles prédictifs sur ces variables seules donnent un R^2 modéré (~0.52 avec KNN).

2.**Corrélations élevées entre certaines variables nutritionnelles.**
- Ces redondances (e.g., fat_100g <> saturated_fat_100g) pourraient être compressées via PCA.

3.**La PCA permet de mieux capturer l'information liée à calcium_100g.**
- PC4 et PC5 montrent une forte relation avec la variable cible (corrélations de 0.70 et 0.58).
- Utiliser les composantes principales dans un modèle pourrait améliorer les performances.

Prédiction du calcium (1ère cible – auto-complétion)
Ajout d’une seconde cible : sucre (sugars_100g)



Trois modèles ont été testés pour prédire es variable calcium_100g et sugars_100g:

- **Régression linéaire** : mauvaise performance ($R^2 \approx 0.12$)
- **Arbre de décision** : correct ($R^2 \approx 0.43$)
- **kNN (k=5)** : le plus performant ($R^2 \approx 0.52$)

Une **visualisation** a permis de comparer les valeurs prédites à la réalité (y_test vs y_pred).
. cela montre une dispersion, mais une tendance bien captée pour le calcium
. Les points sont bien alignés sur la diagonale (y_test ≈ y_pred) pour le sucre

Le calcium est modérément prévisible. La performance est correcte mais pas parfaite.

mais Le **kNN** s’est encore avéré excellent avec **$R^2 \approx 0.85$**, révélant que le sucre est une variable **fortement prédictible**.

Ce tableau met en valeur la **supériorité du sucre** comme cible d’auto-complétion.
Et à l'évidence son importance dans la **qualité nutritionnelle**.

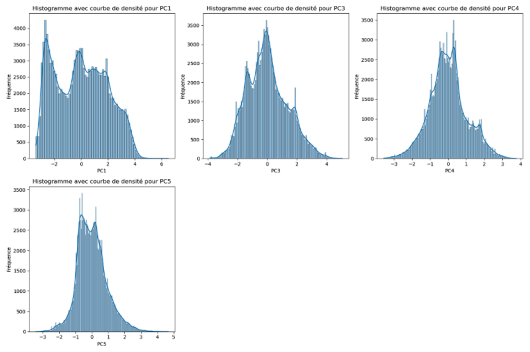
conclusion:

- Les analyses ont validé que les **nutriments influencent clairement le Nutri-Score**, notamment énergie, graisses, sucres, sodium et fibres.

- La prédiction via **machine learning** (notamment avec KNN) permet d’envisager une **auto-complétion** des données nutritionnelles.

- Nous avons d’abord utilisé le calcium comme cible didactique pour illustrer notre démarche, puis nous avons comparé avec le sucre, qui s’est révélé beaucoup plus prédictif.

la conclusion globale est que les composantes principales **PC1, PC3, PC4 et PC5** ne suivent pas une **distribution normale**. Compte tenu de la **non normalité** des distribution je m'oriente vers le **Test de Spearman** car ce test n'exige pas que les données soient normalement distribuées. Il est donc particulièrement adapté aux données qui montrent des écarts significatifs par rapport à la **distribution normale**.



Concentration de la Variance :

Les premières composantes principales (PC1, PC2, et PC3) capturent une grande partie de la variance totale (environ 68,4% combiné). Cela signifie que les premières composantes principales contiennent la majorité de l'information contenue dans les données.

La première composante principale seule explique déjà environ **33,8% de la variance**, ce qui est un pourcentage **significatif**. Elle pourrait représenter la direction dans laquelle les données varient le plus.

- *plus d'explication sur la fiche stats

PC1 : montre une distribution assez asymétrique avec plusieurs pics. Cela suggère que les données projetées sur cette composante sont hétérogènes, possiblement composées de plusieurs sous-groupes. La distribution semble décalée vers la gauche, indiquant une majorité de valeurs négatives ou proches de zéro.

PC3 : montre une distribution plus symétrique, proche d'une distribution normale, avec un pic central marqué autour de zéro. Cela pourrait indiquer que PC3 capte une variance importante des données avec une distribution équilibrée autour de la moyenne.

PC4 : la distribution semble aussi relativement symétrique avec un léger biais à gauche. On observe un pic principal autour de zéro, mais avec quelques fluctuations indiquant la présence de sous-structures ou de variabilité supplémentaire dans cette composante.

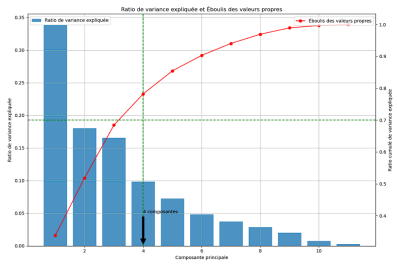
PC5 : la distribution est similaire à celle de PC4, avec une symétrie relative et un pic central autour de zéro. Cette composante semble également capturer une partie de la variance qui présente une distribution proche de la normale.

conclusion:

PC4 et PC5 : Les plus pertinentes pour la modélisation du calcium.
PC3 : Apporte une valeur ajoutée modérée, avec une distribution saine.
PC1 : Faible corrélation mais une distribution qui pourrait cacher des clusters intéressants.
Les composantes principales PC3, PC4, et PC5 montrent des distributions relativement symétriques, proches d'une distribution normale. PC1, par contre, présente une distribution plus complexe avec plusieurs pics, suggérant une structure plus hétérogène des données dans cette direction. Ces résultats indiquent que les composantes principales capturent différentes facettes de la variabilité des données, certaines plus simples et d'autres plus complexes.

. **PC1 et calcium_100g Corrélation: 0.1716**
. Statistiquement significative
. Interprétation: Bien que significative, la corrélation est relativement faible, ce qui indique que PC1 explique une petite partie de la variabilité de la teneur en calcium.
. **PC3 et calcium_100g Corrélation: 0.2900**
. Statistiquement significative
. Interprétation: Une corrélation modérée qui suggère que PC3 a une influence plus marquée sur la variation de la teneur en calcium que PC1.
PC4 et calcium_100g Corrélation: 0.6703
. Statistiquement significative
. Interprétation: Une forte corrélation indiquant que PC4 est très associée avec les variations de la teneur en calcium. PC4 pourrait capturer des aspects des données qui sont particulièrement pertinents pour prédire ou comprendre les variations du calcium.
. **PC5 et calcium_100g Corrélation: 0.5003**
. Statistiquement significative
. Interprétation: Cette corrélation significative et modérée à forte suggère que PC5 est également un bon prédicteur de la teneur en calcium, bien que moins influente que PC4.
Toutes les composantes principales testées montrent une corrélation statistiquement significative avec la cible (calcium_100g), ce qui signifie que ces relations ne sont probablement pas dues au hasard ou en tout cas par rapport aux aliments que nous disposons dans notre dataset.
Analyse en composante principale (ACP).

Éboulis des valeurs propres & Variance expliquée (PCA)



Barres bleues : Variance expliquée par chaque composante principale (PC1, PC2...)

Courbe rouge : Somme cumulée des variances expliquées

Ligne verte horizontale : Seuil de 70% de variance expliquée

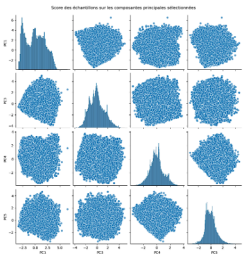
Ligne verte verticale : Le nombre de composantes nécessaires pour atteindre ce seuil

Annotation fléchée : Indique le nombre de composantes retenues selon le critère

Résultat:
. Seuil de variance expliquée : 70%
. Nombre de composantes nécessaires : 4 composantes principales
. Variance expliquée par ces 4 PCs :
PC1 : ~34%
PC2 : ~18%
PC3 : ~17%
PC4 : ~10%
Total (PC1 à PC4) ≈ 70%

Ce graphique nous permet de déterminer combien de composantes principales sont nécessaires pour capturer une proportion substantielle de la variance des données d'origine.
Dans ce cas, le graphique montre que les quatre premières composantes principales expliquent environ 70% de la variance totale

. ici on a la distribution individuelle de chaque composante principale (histogrammes sur la diagonale)



. et Les relations bivariées entre les composantes (nuages de points en dehors de la diagonale)

. Les composantes sélectionnées (PC1, PC3, PC4, PC5) représentent différentes dimensions des données sans redondance.

. PC1 semble la plus discriminante, possiblement utile pour détecter des groupes ou catégories cachées dans les données.

. La PCA a bien réduit la dimensionnalité tout en conservant des dimensions informatives et orthogonales.

L’analyse PCA révèle une structure sous-jacente dans les données, avec plusieurs pics dans PC1 suggérant des sous-groupes potentiels.
La bonne indépendance entre composantes indique une réduction de dimension réussie, utile pour simplifier l’analyse tout en conservant l’information essentielle.
Ces résultats ouvrent la voie à des méthodes de clustering ou classification, facilitant la segmentation, la personnalisation ou des analyses avancées.

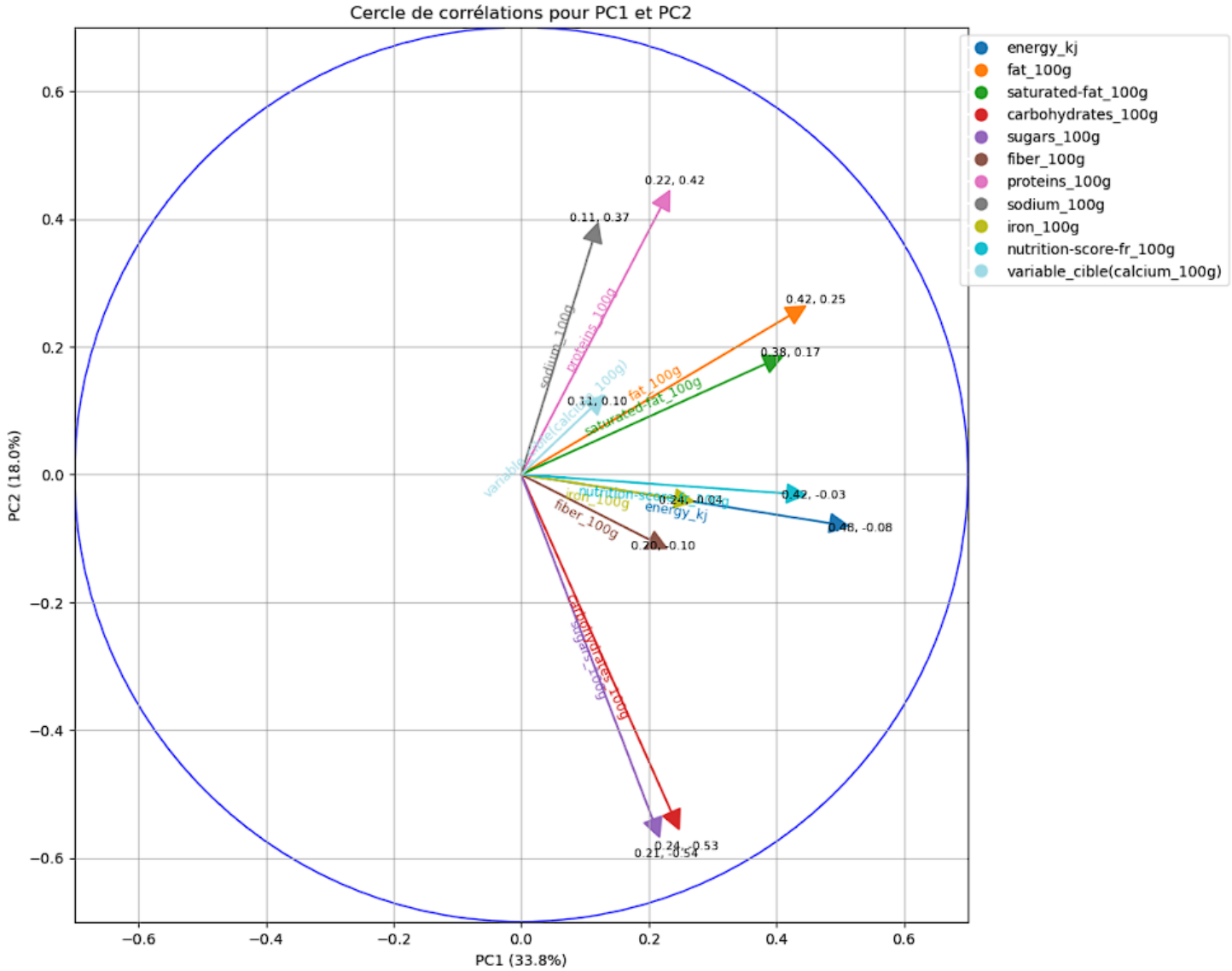
L’analyse montre que les données contiennent probablement différents types de produits ou groupes cachés.
Cela signifie qu’on pourrait les classer en catégories naturelles.
L’étude a aussi permis de résumer les informations sans rien perdre d’important, ce qui facilite leur compréhension.
Grâce à cela, on pourra plus facilement analyser, comparer, ou organiser les produits, par exemple pour mieux cibler les besoins, faire des recommandations, ou détecter des différences importantes.

PC1 (33.8% de variance expliquée) semble être un axe d'opposition entre densité nutritionnelle (fat, saturated fat, energy) et éléments bénéfiques (fiber).

PC2 (18.0%) capte des différences liées à sugars_100g, iron_100g, calcium avec des effets modérés.

Les variables qui doivent être rempli pour la futur applications:

. **Energy_Kj**
. **Carbohydrates**
. **Sugar**
. **Saturated-fat**
. **Fat**
. **Proteins**
. **Sodium**



- **PC1 (33.8 %)** oppose des produits riches en énergie/graisses à ceux riches en fibres : c’est un axe de qualité nutritionnelle.

- **PC2 (18 %)** capte des variations liées au sucre, fer, calcium mais avec moins d'impact.

- Le **calcium** est faiblement représenté sur PC1/PC2 mais fortement sur **PC4 et PC5** → utiles pour le modéliser.

- **ANOVA** montre un effet significatif du Nutri-Score sur le calcium, mais **$R^2 = 0.002$**, donc un effet très faible.

- Le **calcium** n’est pas bien prédit par les autres nutriments classiques → il faut des approches spécifiques.

- La **PCA** est utile pour réduire la complexité, identifier des groupes naturels (clustering), ou améliorer la suggestion automatique de valeurs manquantes.

- Les variables **obligatoires** à remplir pour un bon calcul du Nutri-Score : energy_kj, fat_100g, saturated_fat_100g, carbohydrates_100g, sugars_100g, proteins_100g, sodium_100g.

- Les produits peuvent être classés selon leur **qualité nutritionnelle**, surtout en fonction des graisses, sucres, fibres, etc.

- Le **calcium** ne suit pas bien cette logique : il varie de façon plus indépendante.

- L’analyse montre qu’on peut **résumer les données complexes** en quelques indicateurs essentiels (grâce à la PCA).

- Le Nutri-Score dépend bien des principales infos nutritionnelles (graisses, sucres...), mais pas du calcium.

- Pour **bien remplir la fiche d’un produit**, il faut surtout indiquer : l’énergie, les graisses (et saturées), les sucres, les glucides, le sel, les protéines.

- D’autres infos comme le calcium peuvent être **déduites plus tard**, mais de façon moins fiable.

- L’objectif est de **simplifier la saisie** et **améliorer la qualité** des données sur Open Food Facts.