



# Préparez des données pour un organisme de santé publique

## Présentation

# SOMMAIRE



## Présentation du projet

- Contexte et Problématique
- Mission et Objectif à atteindre

## Analyse des données

- Description du jeu de données
- Nettoyage et préparation des données
- Identification des variables pertinentes
- Identification et traitement des valeurs aberrantes
- Identification et traitement des valeurs manquantes

## Exploration des données

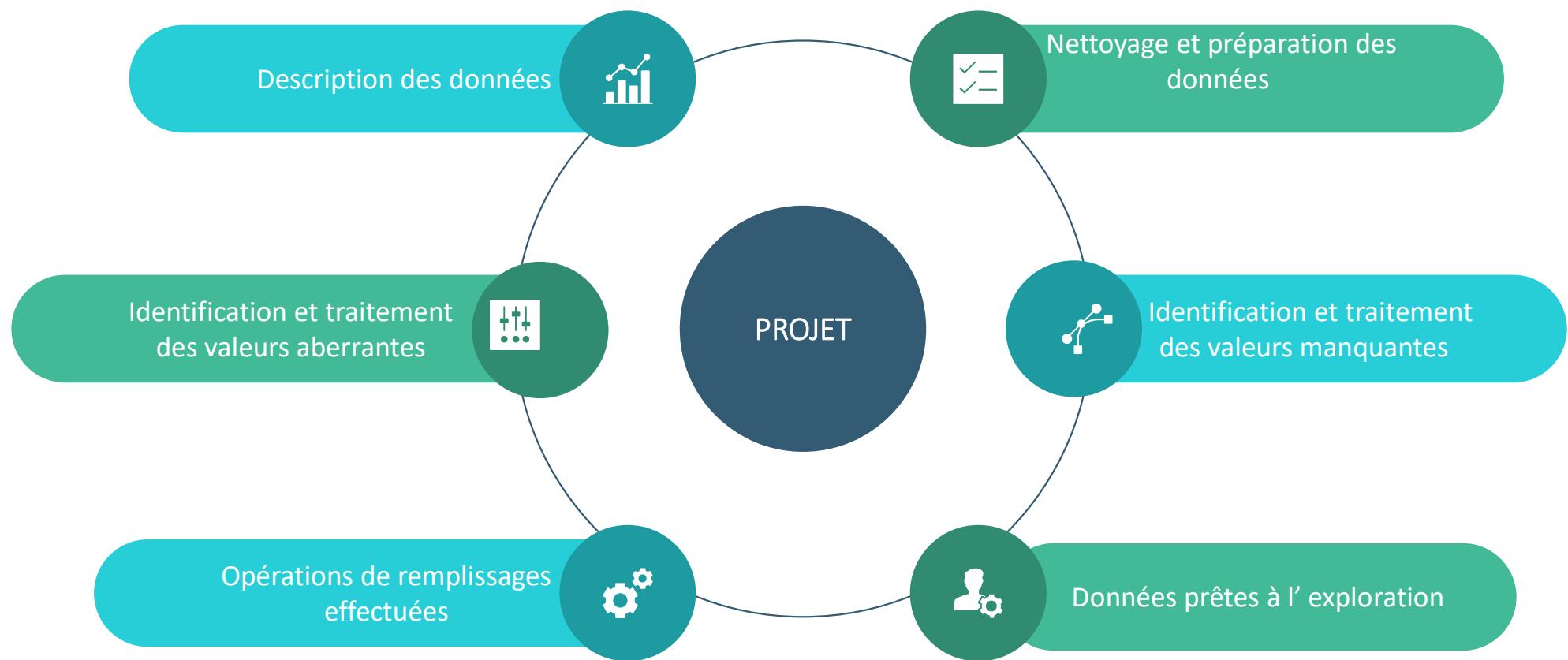
- Description et Analyse univariée des différentes variables
- Analyse bivariée et relations entre les variables
- Analyse multivariée et les résultats statistiques associés
- Analyse en Composante Principale (ACP)
- Analyse de la Variance (ANOVA)

## Synthèse et Perspectives

# PRESENTATION DU PROJET

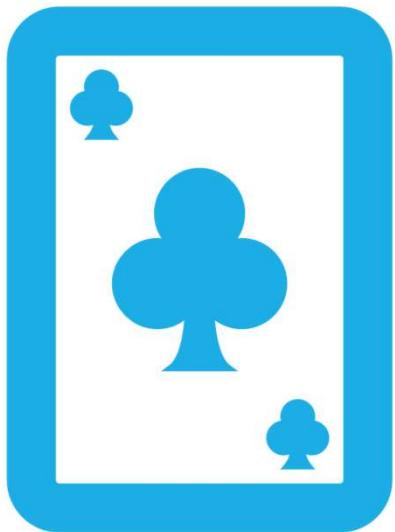
- **Problématique :**
  - La qualité nutritionnelle des produits alimentaires est essentielle pour la santé publique, mais la saisie manuelle des données dans la base Open Food Facts peut entraîner des erreurs et des valeurs manquantes, compromettant ainsi la fiabilité de la base de données.
- **Contexte :**
  - Santé publique France cherche à améliorer la base de données Open Food Facts, un projet open source qui recueille des informations sur les produits alimentaires du monde entier, afin de faciliter l'accès à des données fiables sur la qualité nutritionnelle des aliments.
- **Mission :**
  - Développer un système de suggestion ou d'auto-complétion pour aider les utilisateurs à remplir plus efficacement la base de données Open Food Facts, en commençant par une phase de nettoyage et d'exploration des données existantes.
- **Objectif :**
  - Nettoyer et explorer le jeu de données Open Food Facts pour évaluer la faisabilité d'un système de suggestion de valeurs manquantes pour les champs de données, afin d'améliorer la qualité et la fiabilité de la base de données.

# ANALYSE DES DONNEES



# ANALYSE DES DONNEES

## ➤ Présentation du jeu de données



- **Nature du fichier :** Base de données en format CSV
- **Taille du fichier :**
  - 320,772 lignes
  - 162 colonnes
- **Description des données :**

Chaque ligne représente un produit alimentaire, incluant des identifiants uniques et des informations nutritionnelles détaillées.
- **Remarques sur les données :**
  - Certaines colonnes sont en float et contiennent des dates (années) : à transformer.
  - 16 colonnes totalement vides à supprimer.
  - Beaucoup de valeurs manquantes.
  - En termes de qualité de notre base de données, un taux de remplissage de 76.22% est assez élevé, ce qui indique que la majorité des données sont présentes,
  - Pas de doublons
- **Plan d'action pour la préparation des données :**

Une procédure de nettoyage et d'imputation des données est requise pour optimiser l'utilisation de la base de données avant de procéder à l'analyse.

# • ANALYSE DES DONNEES •

## ➤ Nettoyage et préparation des données



### Sélection et nettoyage des colonnes par features

- Suppression des colonnes inutiles
- Suppression des colonnes vides et peu renseignées
- Traitement de la colonne "countries"



### Traitement des colonnes contenant des dates

- Conversion du format des dates



### Traitement de la colonne "countries"

- Conserver uniquement les lignes où la colonne 'countries' a des valeurs plus longues que 3 caractères



### Nettoyage des données par Produit

- Suppression des produits redondants
- Suppression des colonnes peu renseignés
- Traitement des données aberrantes et manquantes

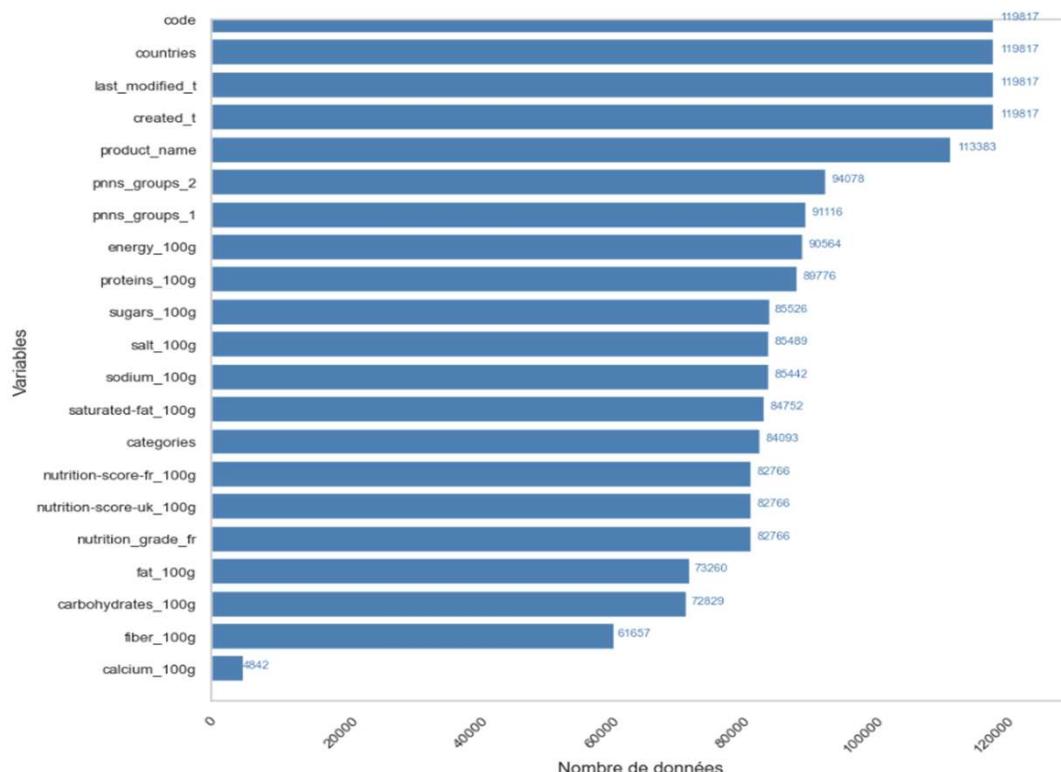
# ANALYSE DES DONNEES

## ➤ Identification des variables pertinentes pour l'Analyse Nutritionnelle:

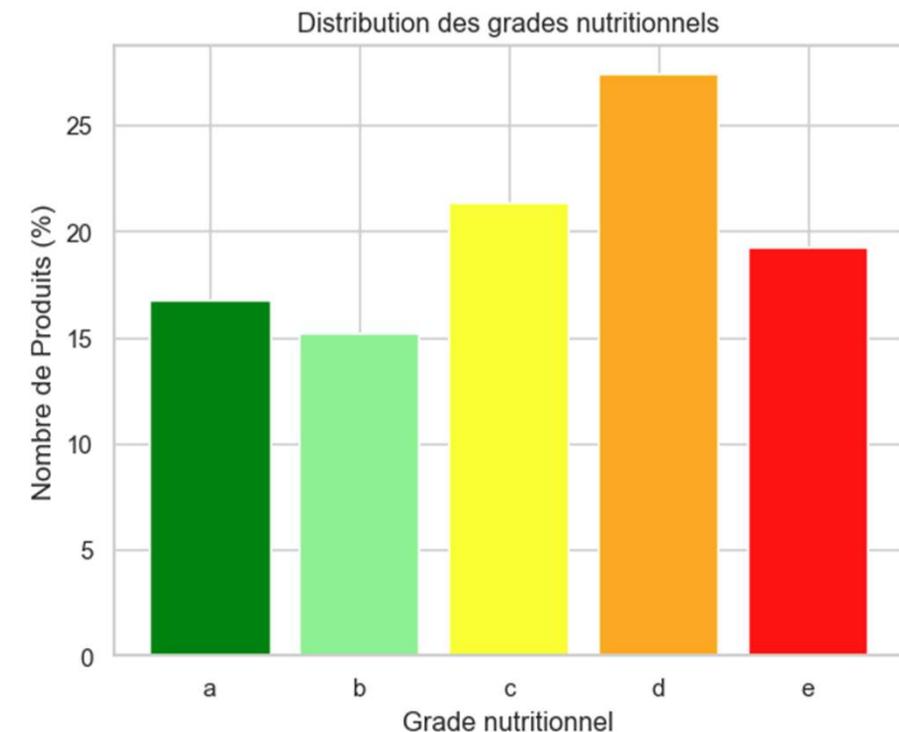
Catégorie	Variables clés	Justification du choix des variables
Exemples des valeurs pertinentes	<ul style="list-style-type: none"><li>le nom du produit, les marques, l'énergie ,les glucides, les protéines, les lipides ,les fibres pour 100g,etc.</li></ul>	Indispensables à l'analyse nutritionnelle pour évaluer la qualité des produits.
Auto-complétion	<ul style="list-style-type: none"><li>Date de creation</li><li>Date de modification</li></ul>	Disponibilité élevée des données permettant de développer des modèles de prédiction pour compléter les valeurs manquantes.
Analyse Statistique	<ul style="list-style-type: none"><li>Nutriscore et nutrigrade (nutrition_grade_fr, nutrition-score-fr_100g, nutrition-score-uk_100g)</li></ul>	Pertinentes pour des analyses univariées et multivariées sur la qualité nutritionnelle.
Contexte du Produit	<ul style="list-style-type: none"><li>Catégories alimentaires et les labels de qualité(pnns_groups_1, pnns_groups_2)</li></ul>	Fournissent un contexte pour l'analyse en catégorisant les produits alimentaires.

# ANALYSE DES DONNEES

## ➤ Visualisation de la complétude des données



Le graphique à barres horizontales présente la complétude des données de notre base de données. Chaque barre indique le nombre de valeurs non manquantes pour chaque variable (ou colonne) du jeu de données. Les variables sont listées sur l'axe vertical (axe des y) et le compte des valeurs non manquantes est indiqué sur l'axe horizontal (axe des x).



"Pour les étapes suivantes, notre analyse se concentrera principalement sur l'analyse du Nutri-Score et du Nutrition Grade, et sur l'identification de corrélations potentielles avec les différentes variables nutritionnelles."

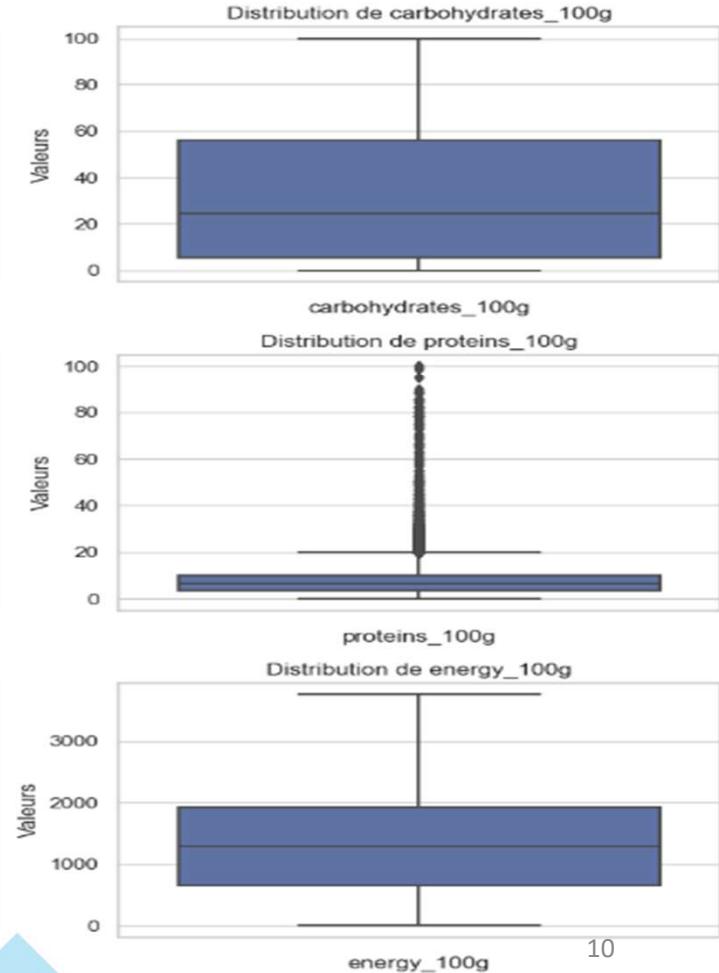
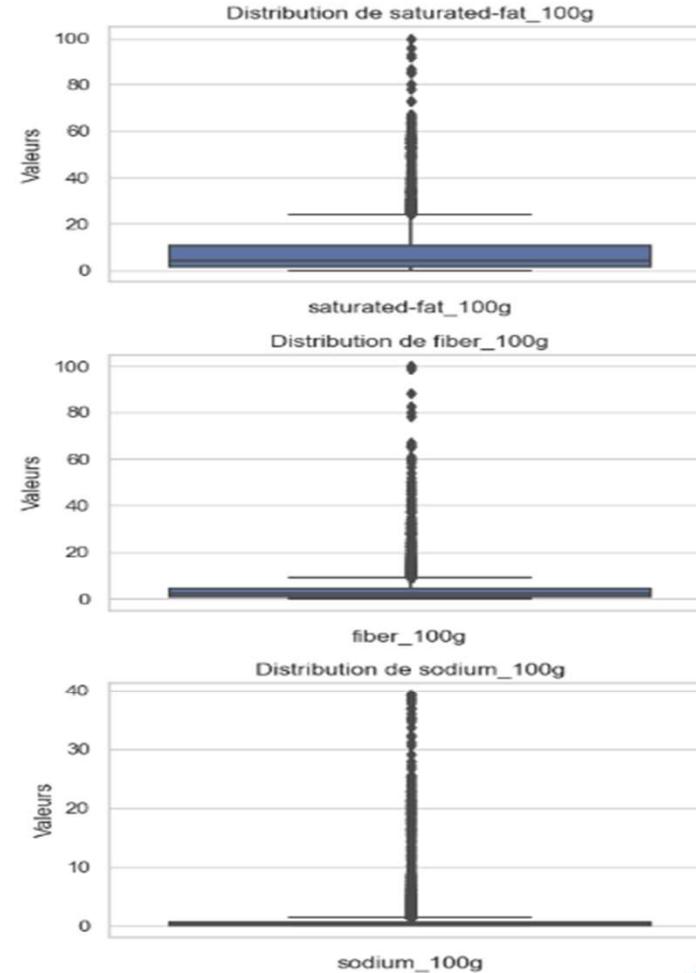
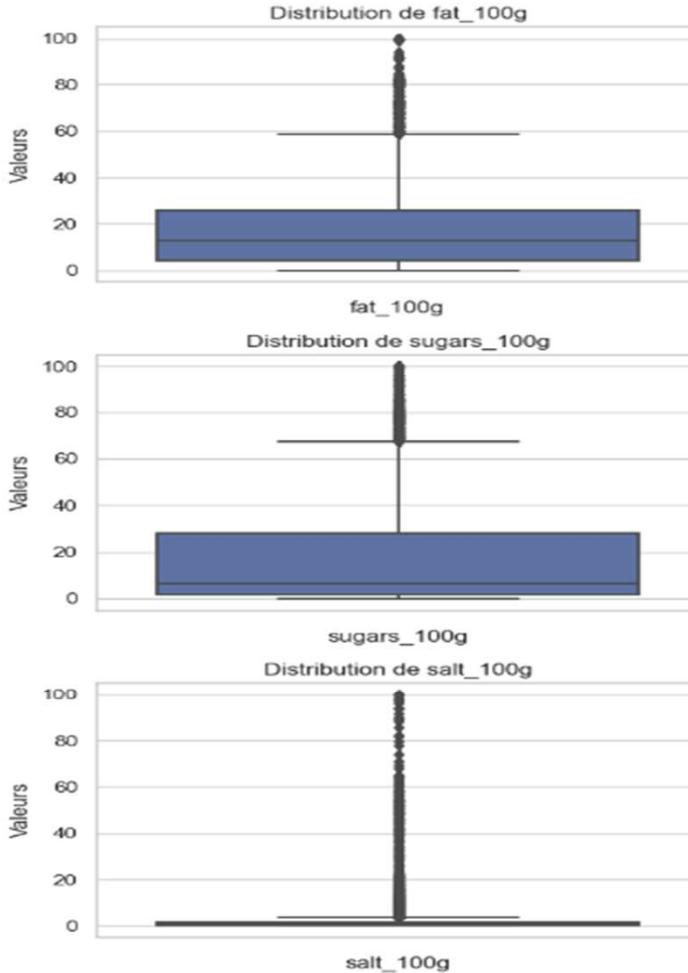
# ANALYSE DES DONNEES

## ➤ Identification et traitement des valeurs aberrantes

Processus	Action	Objectif
Identification	Utilisation des méthodes statistiques pour détecter les valeurs extrêmes.	Déterminer les points de données qui dévient significativement de la tendance générale.
Suppression des valeurs extrêmes	Exclusion des valeurs qui dépassent des seuils logiques (ex: énergie maximale > 3766 kJ pour energy_100g).	Assurer que les analyses subséquentes soient basées sur des données fiables et représentatives.
Comparaison spécifique	Comparaison entre des variables liées (ex: sugars_100g par rapport à carbohydrates_100g).	Vérifier la cohérence des données au sein des groupes de variables similaires.
Traitement des valeurs négatives	Suppression des valeurs négatives pour des scores nutritionnels (nutrition-score-fr_100g et nutrition-score-uk_100g).	Maintenir la validité des mesures de scores nutritionnels.
Résultat post-traitement	Le dataset résultant après traitement.	Disposer d'un dataset nettoyé prêt pour des analyses approfondies.

# ANALYSE DES DONNEES

## ➤ Identification et traitement des valeurs aberrantes



# ANALYSE DES DONNEES

## ➤ Identification et traitement des valeurs manquantes

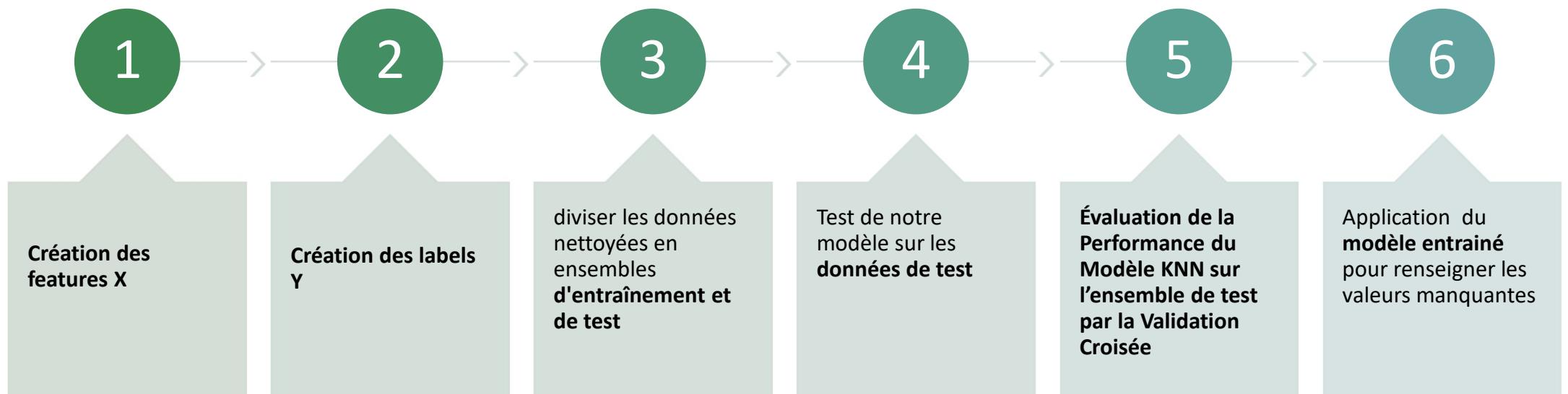
Nous précisons les méthodes utilisées pour quelles variables et en donnant un aperçu de l'étape de validation croisée pour assurer la fiabilité des imputations.

Méthode d'Imputation	Variables	Description
Moyenne	energy_100g, carbohydrates_100g	Remplissage des valeurs manquantes avec la moyenne de la variable, pour <b>les données symétriques</b> .
Médiane	fat_100g, saturated-fat_100g, sugars_100g, fiber_100g, proteins_100g, sodium_100g, salt_100g, calcium_100g	Remplissage des valeurs manquantes avec la médiane, approprié pour <b>les données asymétriques</b> avec des valeurs aberrantes.
Mode (valeur la plus fréquente)	categories, pnns_groups_1, pnns_groups_2	Remplissage des valeurs manquantes avec la valeur la plus fréquente pour <b>les variables catégorielles</b> .
Machine Learning (Régression linéaire et KNN)	nutrition-score-fr_100g, nutrition-score-uk_100g, nutrition_grade_fr	Utilisation de modèles de machine learning pour prédire et imputer les valeurs manquantes pour <b>les scores nutritionnels et grades</b> .

➤ Nous avons évalué la performance du modèle KNN par la « validation croisée » pour assurer une prédiction robuste et fiable des valeurs manquantes.

# ANALYSE DES DONNEES

➤ processus de remplissage des données manquantes avec la Machine Learning



# ANALYSE DES DONNEES

## ➤ Synthèse des Stratégies d'Imputation et Performance des Modèles Associés

Méthode d'Imputation	Variables Imputées	Performance
Moyenne	energy_100g, carbohydrates_100g	—
Médiane	fat_100g, saturated-fat_100g, sugars_100g, fiber_100g, proteins_100g, sodium_100g, salt_100g, calcium_100g	—
Mode (Valeur la plus fréquente)	Variables catégorielles (categories, pnns_groups_1, pnns_groups_2)	—
KNN Regressor	nutrition-score-fr_100g, nutrition-score-uk_100g	Score R <sup>2</sup> de validation croisée : 0.7795
KNN Classifier	nutrition_grade_fr	Score R <sup>2</sup> de validation croisée : 0.6556
Régression Linéaire	Modèle global après imputation	Score R <sup>2</sup> sur données de test : 0.712      13

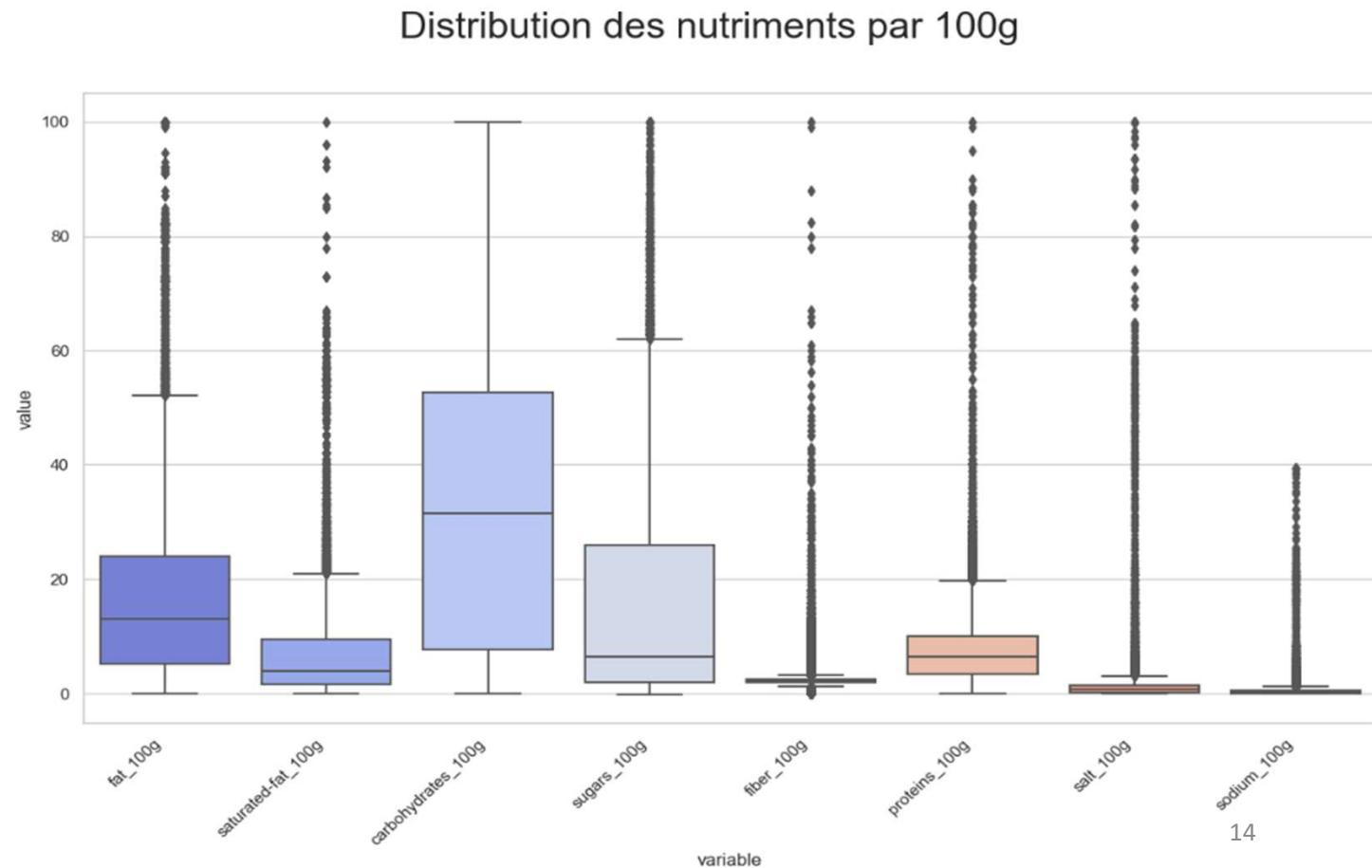
# Exploration des données

## ➤ Description et Analyse univariée des différentes variables

Ce graphique boxplot illustre la distribution des nutriments par 100g dans notre jeu de données.

- La plupart des produits ont une faible teneur en graisses totales et saturées, avec des exceptions notables.
- Les glucides et les protéines montrent une distribution plus homogène à travers les produits.
- Les sucres, les fibres, le sel et le sodium présentent généralement de faibles valeurs médianes, mais avec des extrêmes significatifs.

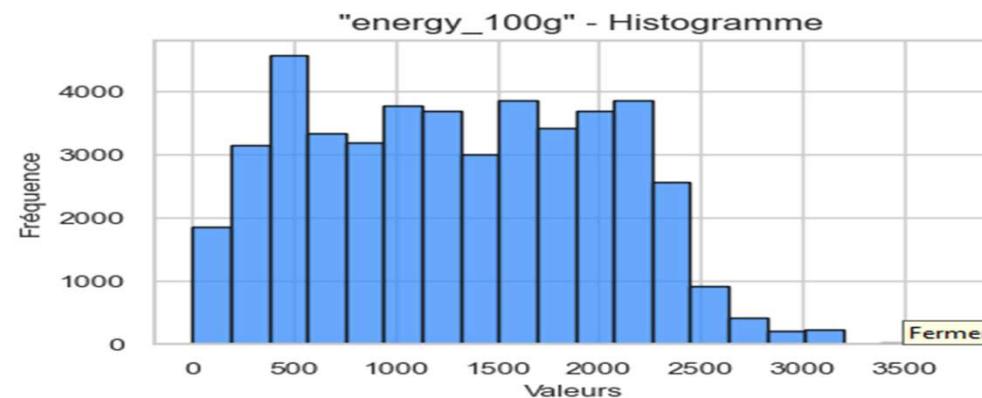
Cette vue d'ensemble révèle la variabilité et la présence de valeurs extrêmes dans les profils nutritionnels des produits.



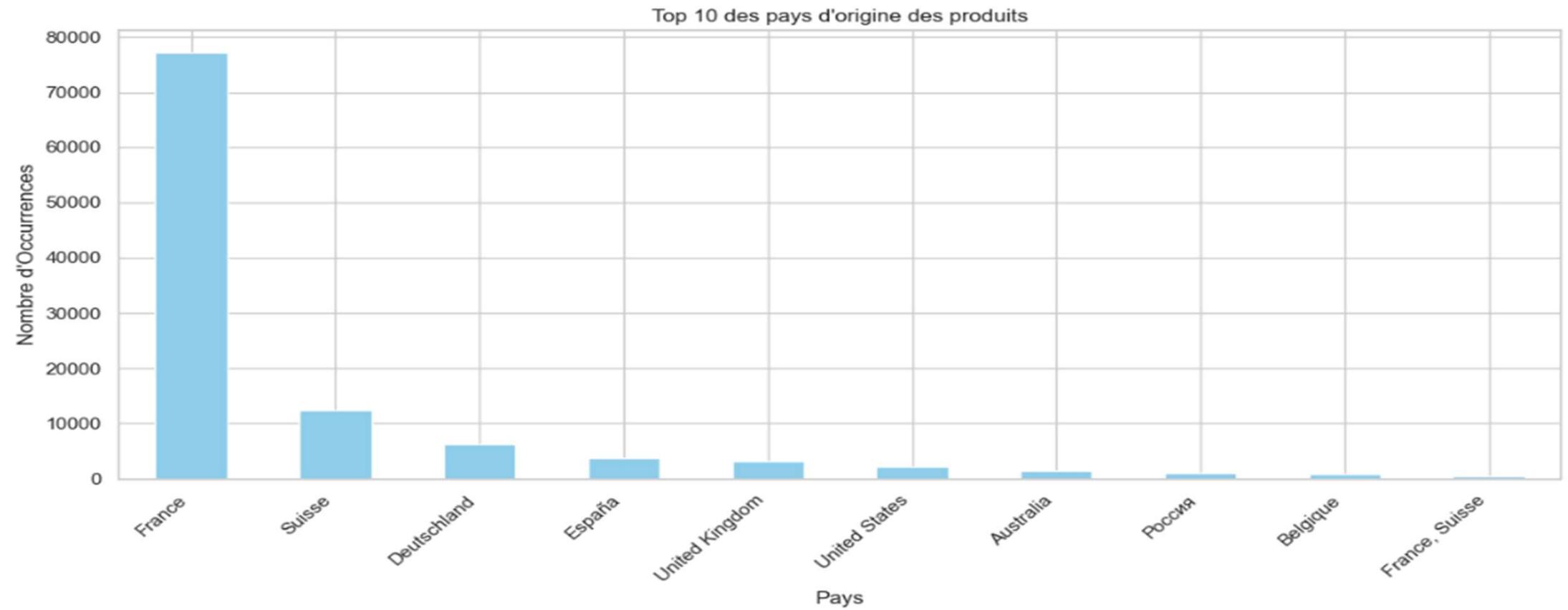
# Exploration des données

- L'histogramme pour **energy\_100g** révèle que la majorité des produits ont une teneur en énergie centrée autour de 1307.25 kJ pour 100g.
- La distribution est relativement symétrique et équilibrée, sans trop d'écart vers les valeurs extrêmement hautes ou basses. Cela suggère une variété modérée dans les niveaux énergétiques des produits alimentaires du jeu de données.

```
Statistiques pour 'energy_100g':  
Moyenne: 1307.25  
Médiane: 1307.25  
Écart Type: 722.98  
Variance: 522701.12  
Minimum: 0.42  
Maximum: 3766.00  
Quartile 25%: 670.00  
Quartile 75%: 1909.00  
Skewness empirique: 0.14  
Kurtosis empirique: -0.91
```



# Exploration des données



Le graphique illustre clairement que la France est le pays d'origine prédominant pour les produits dans notre jeu de données, suivie par une présence significativement moindre de produits en provenance de Suisse, Allemagne, et d'autres pays du top 10.

# Exploration des données

## ➤ Analyse bivariée et relations entre les variables

La heatmap des corrélations nous montre la relation entre les différentes variables nutritionnelles par 100g .

**Énergie et Graisses** : Forte corrélation positive (0.72), indiquant que les produits avec plus de graisses ont généralement plus d'énergie.

•**Graisses et Graisses Saturées** : Très forte corrélation (0.64), suggérant que les graisses dans les produits sont souvent saturées.

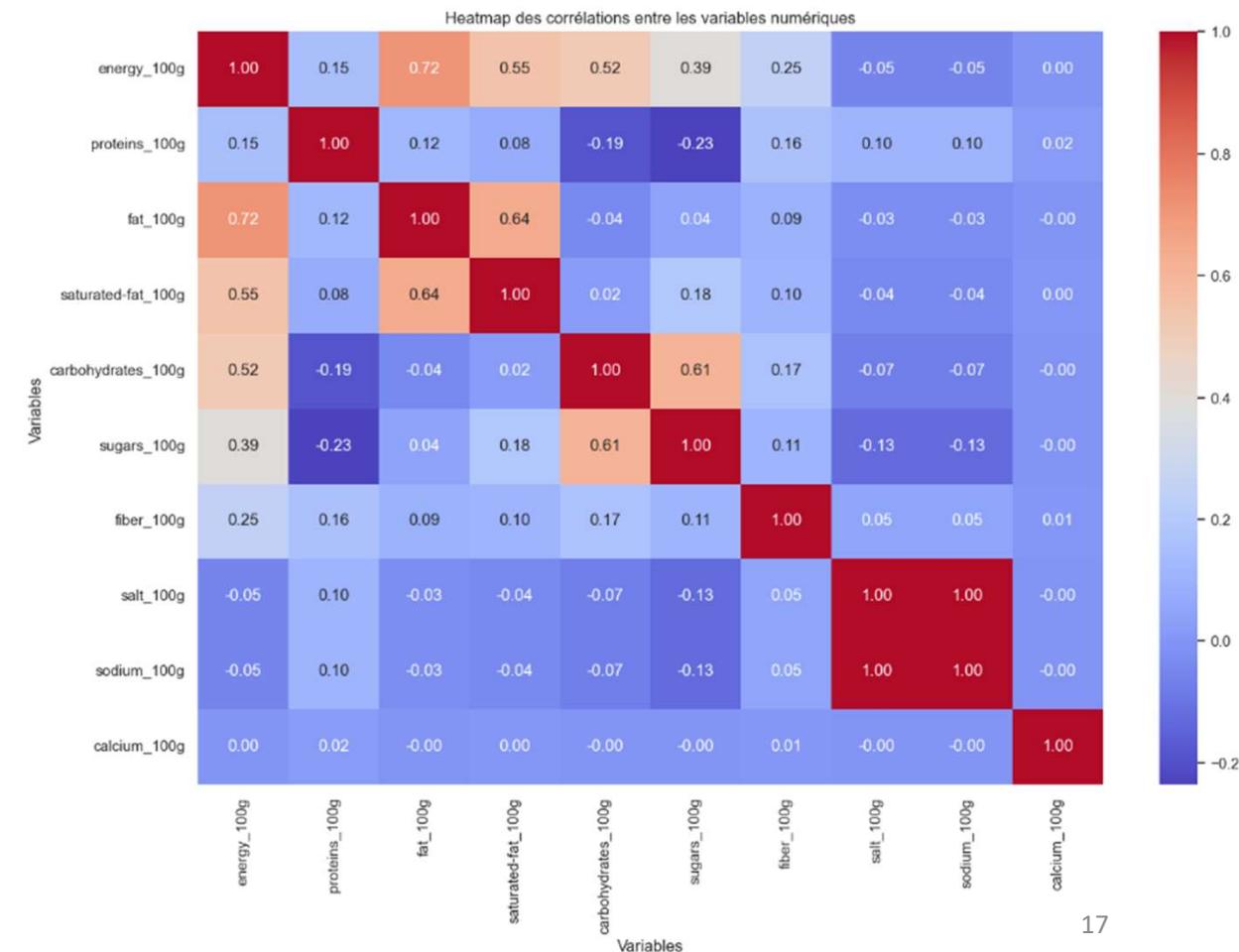
•**Carbohydrates et Sucres** : Corrélation modérée (0.61), ce qui est logique puisque les sucres sont un type de glucide.

•**Fibres** : Pas de corrélation forte avec d'autres nutriments, indiquant une indépendance dans leur contenu par rapport aux autres variables.

•**Sel et Sodium** : Corrélation parfaite (1), ce qui est attendu car le sodium est un composant du sel.

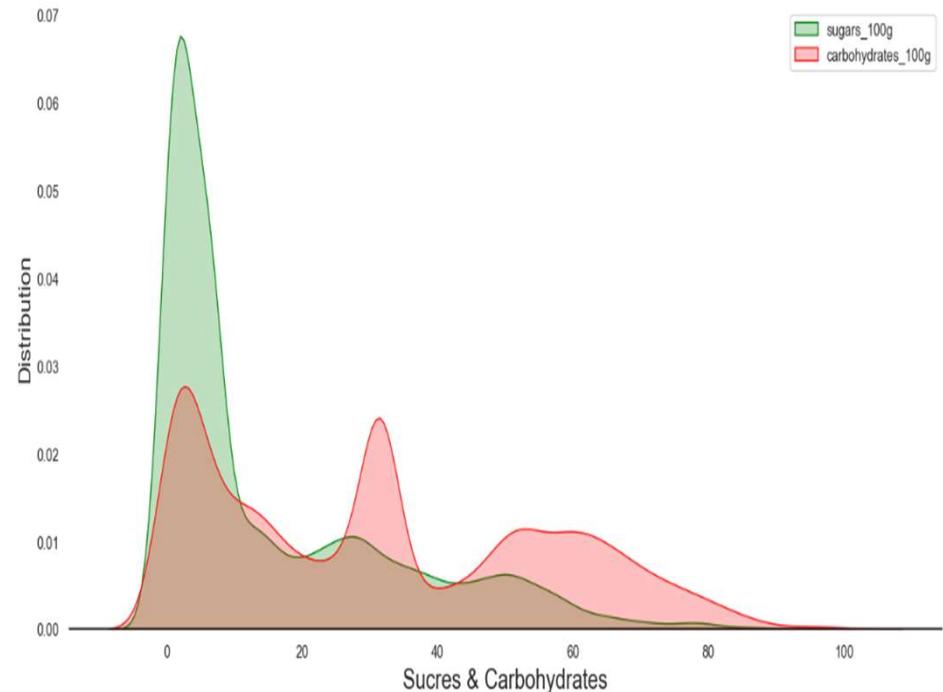
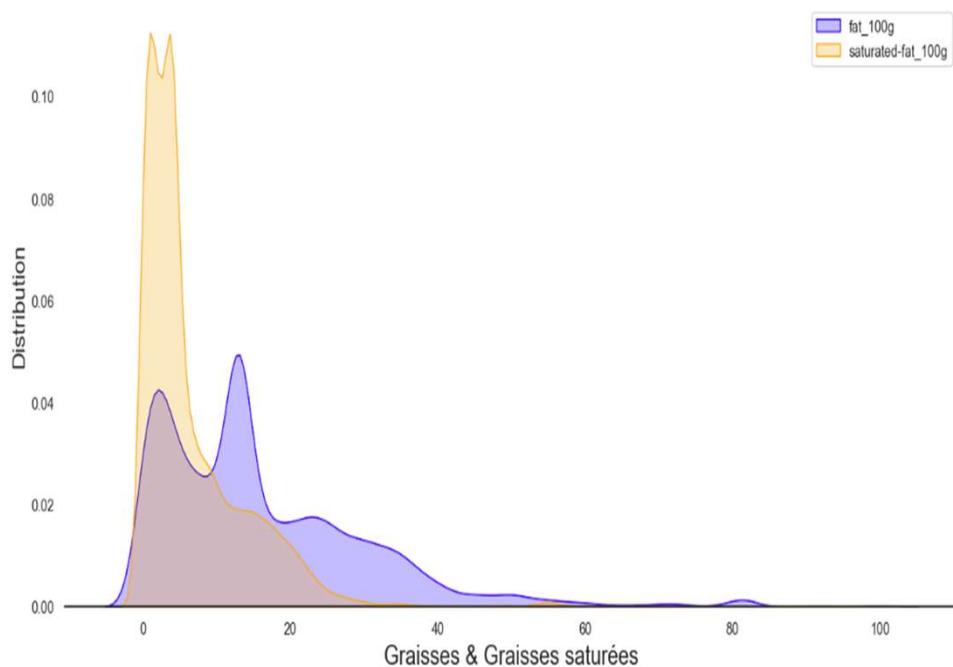
•**Dans le cadre de la modélisation, lorsqu'on identifie deux variables fortement corrélées, il est judicieux d'en sélectionner une et d'exclure l'autre pour éviter la redondance d'informations. Cela optimise la performance et l'interprétabilité du modèle.**

•**Toutefois, dans notre étude actuelle qui est de nature exploratoire, nous conservons toutes les variables pour une analyse exhaustive, sans viser la construction d'un modèle prédictif à ce stade.**



# Exploration des données

## ➤ Visualisation des Distributions des Valeurs Nutritionnelles par Catégorie



- Les graphiques révèlent que la plupart des produits ont de faibles niveaux de graisses et de graisses saturées, mais quelques-uns contiennent des quantités élevées.
- Pour les sucres et les glucides, il y a une plus grande variété de teneurs, suggérant un large éventail de compositions en glucides parmi les produits.

# Exploration des données

## ➤ Analyse multivariée et les résultats statistiques associés

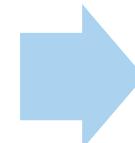
- Analyse descriptive : ACP
- Analyse explicative : ANOVA

### 1. ACP: Réduction, Clarté, Efficacité.

L'ACP, ou Analyse en Composantes Principales, est un algorithme de réduction de dimensionnalité qui permet de transformer des variables possiblement corrélées en un nombre réduit de variables non corrélées appelées composantes principales.



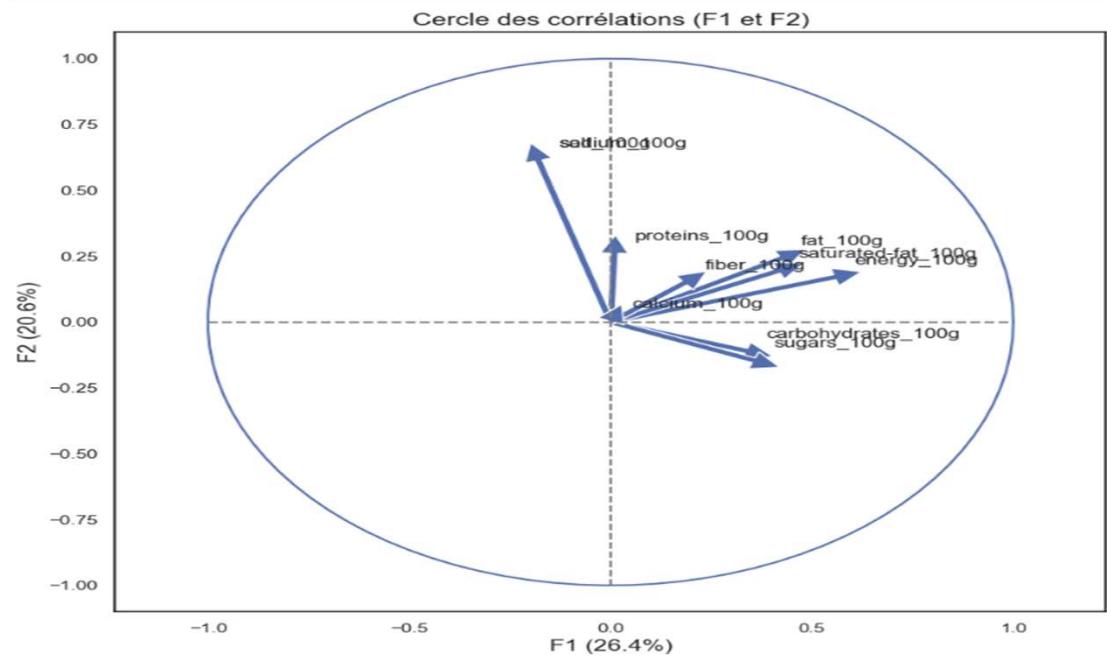
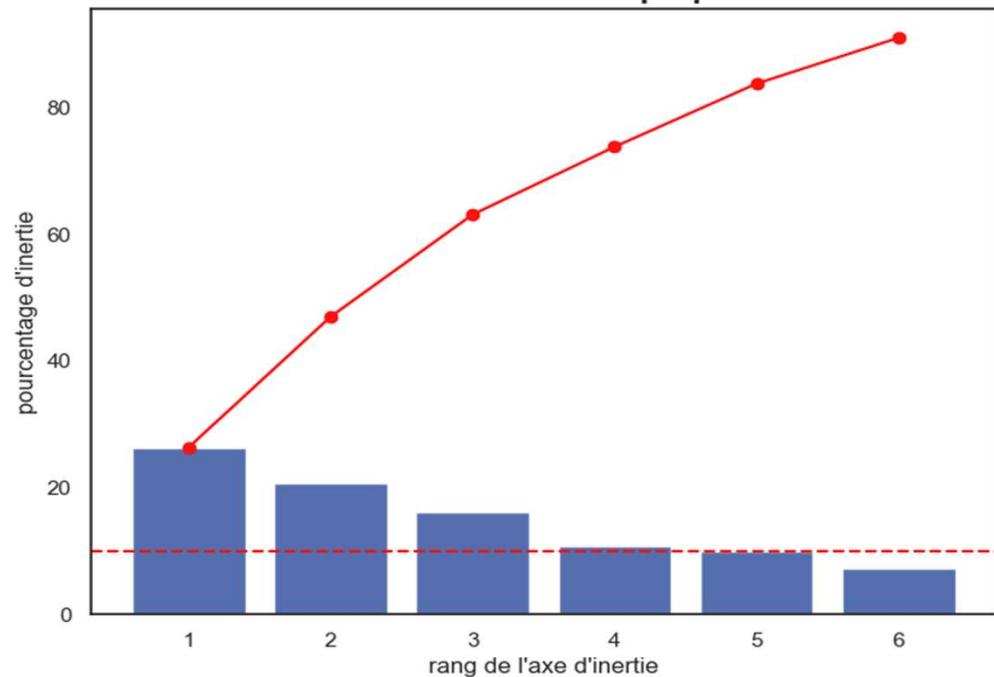
L'**éboulis des valeurs propres** est une représentation graphique qui sert à déterminer le nombre de composantes principales à retenir en visualisant la part de variance expliquée par chaque composante.



"Comment l'Analyse en Composantes Principales nous aide-t-elle à identifier et à créer des variables synthétiques à partir de groupes fortement corrélés dans notre jeu de données ?"

## • Exploration des données •

**Eboulis des valeurs propres**



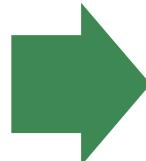
- L'**éboulis des valeurs propres** nous a révélé que les 5 premières composantes principales (F1 à F5) sont significatives, représentant ensemble 86.4% de la variance totale des données. Cela indique que ces 5 composantes capturent la grande majorité de l'information dans le dataset, offrant une représentation réduite mais riche des caractéristiques originales des individus.

- Le cercle de corrélation illustre les relations entre les variables et les deux composantes principales.
- Les variables \*\*\*"saturated\_fat\_100g", "fat\_100g" et "calcium\_100g"\*\*\* sont regroupées et pointent dans une direction similaire vers la droite, ce qui indique qu'elles sont positivement corrélées entre elles et qu'elles ont une forte influence sur la composante principale F1.
- La variable \*\*\*"sodium\_100g" et "salt\_100g" semblent avoir une contribution significative à la composante principale F2, car leurs vecteurs pointent vers le haut.

# Exploration des données

## 2. Analyse explicative : ANOVA (Simplicité, Comparaison, Significativité)

L'ANOVA, ou Analyse de la Variance, est une méthode statistique qui permet de comparer les moyennes de plusieurs groupes pour déterminer si au moins un groupe diffère significativement des autres.



"Comment l'analyse ANOVA peut-elle révéler les relations entre 'nutriscore' et nos variables nutritionnelles?"

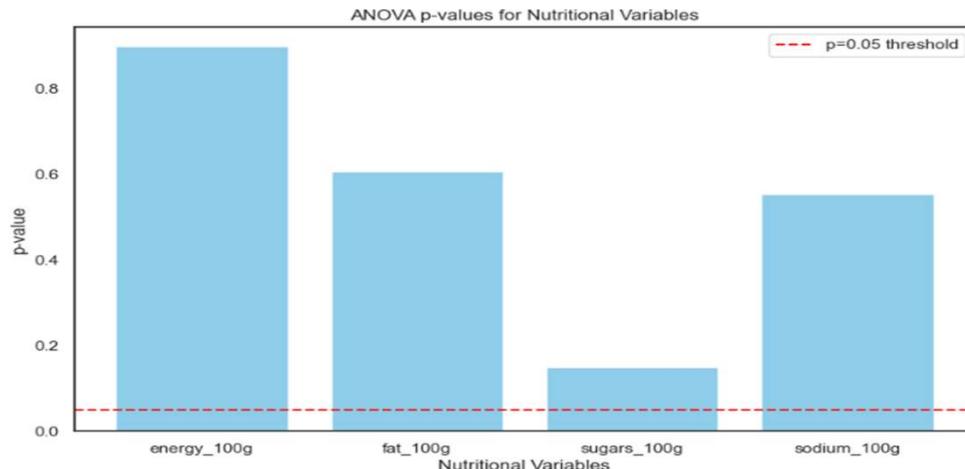
- Etapes de l'analyse explicative ANOVA
  - 1. Sélectionner la variable dépendante qualitative (par exemple, 'nutriscore').
  - 2. Choisir une ou plusieurs variables indépendantes quantitatives (comme les différentes valeurs nutritionnelles).
  - 3. Utiliser une fonction statistique ou un logiciel pour calculer l'ANOVA. En Python, nous pouvons utiliser des bibliothèques comme statsmodels pour réaliser une ANOVA.
  - 4. Interpréter les résultats : Après avoir exécuter l'ANOVA, nous obtiendrons une valeur de p\_value pour chaque variable nutritionnelle testée. Une valeur de p faible (< 0.05 généralement) indique qu'il existe une différence significative entre les moyennes des groupes de 'nutriscore' pour cette caractéristique nutritionnelle.

# Exploration des données

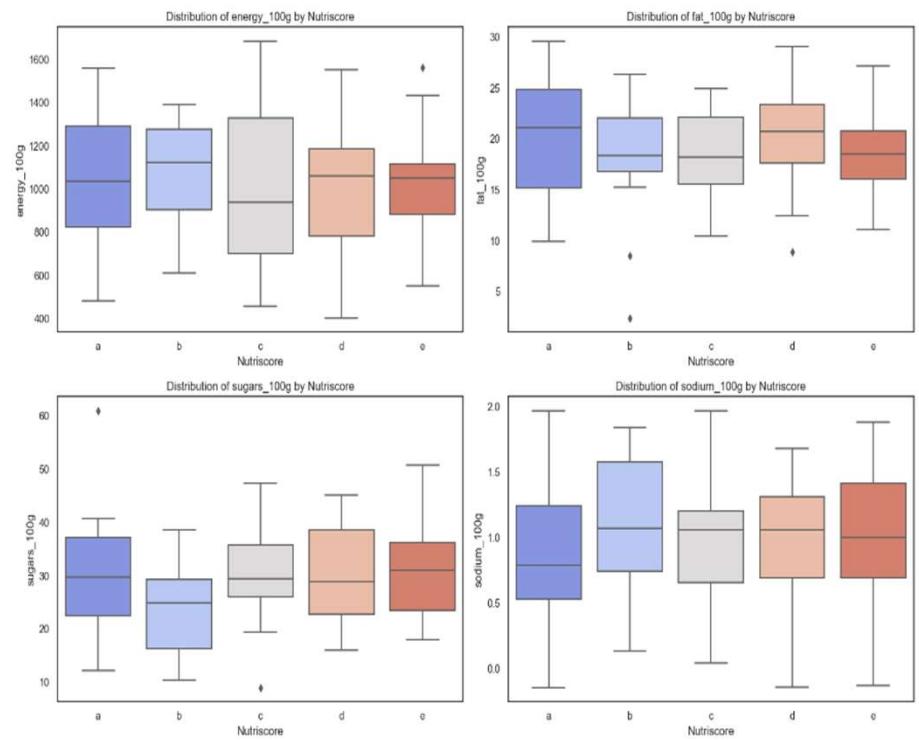
## ➤ Analyse explicative : ANOVA

### Comparaison de la Significativité des Variables Nutritionnelles par "ANOVA"

Le présent graphique représente les valeurs p issues d'une analyse de variance (ANOVA) pour différentes variables nutritionnelles, on observe que toutes **les barres sont au-dessus du seuil de p=0.05**, ce qui signifie qu'aucune des variables testées n'a montré de différence statistiquement significative.



- **Puisque toutes les barres dépassent ce seuil de significativité statistique (généralement admis de  $p = 0.05$ ), cela suggère qu'il n'y a pas de différence statistiquement significative dans la moyenne des groupes de nutriscore pour ces variables nutritionnelles.**
- Les résultats indiquent que, selon l'ANOVA, il n'y a pas de lien significatif entre le nutriscore et ces variables nutritionnelles dans notre jeu de données. Cela peut signifier que le nutriscore n'est pas directement lié à ces mesures nutritionnelles individuelles ou que d'autres facteurs pourraient jouer un rôle dans la détermination du nutriscore.



- Les boxplot illustrent les variations des valeurs nutritionnelles en fonction du nutriscore.
- L'utilisation de l'ANOVA avec ces graphiques nous aide à confirmer si les variations que nous voyons sont dues au hasard ou si elles reflètent des différences réelles et significatives liées au nutriscore.
- Ces analyses ensemble fournissent une compréhension plus profonde de la relation entre le nutriscore et les profils nutritionnels des produits.<sup>22</sup>

# Synthèse et Perspectives

Synthèse du projet	Détails
Points Forts	<ul style="list-style-type: none"><li>Intégrité des données confirmée après nettoyage.</li><li>Méthodes d'imputation améliorant la complétude des données.</li></ul>
Opportunités	<ul style="list-style-type: none"><li>Potentiel pour un système d'auto-complémentation innovant.</li><li>Amélioration possible de la précision des entrées de données.</li></ul>
Points Faibles	<ul style="list-style-type: none"><li>Volume conséquent de données manquantes dans certaines variables.</li><li>Nécessité de techniques d'imputation avancées.</li></ul>
RGPD	<ul style="list-style-type: none"><li>Engagement envers la conformité au RGPD avec une application respectant les principes de minimisation et de sécurité des données.</li><li>Transparence et autonomie des utilisateurs concernant leurs informations personnelles.</li></ul>
Ouverture	<ul style="list-style-type: none"><li>Exploration des possibilités d'élargissement de l'application pour inclure des fonctionnalités de suivi de la qualité nutritionnelle.</li><li>Intégration d'outils de visualisation de données pour les consommateurs soucieux de leur santé.</li><li>Évaluation de partenariats stratégiques pour enrichir la base de données Open Food Facts.</li></ul>



Merci