

Préparez des données pour un organisme de santé publique

Projet 3 - Julien Agneray



Introduction et objectifs du projet

Contexte du projet :

- Client : Santé publique France
- Base de données : Open Food Facts, open source

Demande du client :

- Améliorer la base de données pour mieux connaître la qualité nutritionnelle des produits
- Créer un système de suggestion ou d'auto-complétion pour aider les usagers à remplir plus efficacement la base de données.

Mission confiée :

- Nettoyer et explorer les données existantes
- Déterminer la faisabilité d'un système de suggestion pour l'auto-complétion



Sommaire

1. Opérations de nettoyage effectuées
2. Analyse univariée des variables
3. Analyse bivariée, multivariée et résultats
4. Observations sur l'application client
5. Respect des principes RGPD
6. Conclusions sur la faisabilité du projet



Opérations de nettoyage effectuées

Dimensions et Aperçu Initial des Données

Dimensions des données :

- Lignes : 320 772
- Colonnes : 162

Aperçu initial des données :

- Vérification des premières lignes : aperçu rapide des données pour vérifier le chargement correct
- Identification des colonnes dans les données : liste le nom des colonnes pour trouver une variable cible ayant plus de 50 % de valeurs manquantes.

Détails des colonnes importantes :

- Informations générales : code produit, URL, créateur, dates de création et modification
- Valeurs nutritionnelles : énergie, graisse, glucides, protéines, etc.
- Étiquettes et catégories : tags, origine, ingrédients

Opérations de nettoyage effectuées

Dimensions et Aperçu Initial des Données

Tableau qui résume les statistiques descriptives des features quantitatives (Aperçu)

	no_nutriments	additives_n	ingredients_from_palm_oil_n	ingredients_from_palm_oil	ingredients_that_may_be_from_palm_oil_n	ingredients_that_may_be_from_palm_oil	nutrition_grade_uk	energy_100g	energy-from-fat_100g	fat_100g	...
count	0.0	248939.000000	248939.000000	0.0	248939.000000	0.0	0.0	2.611130e+05	857.000000	243891.000000	...
mean	NaN	1.936024	0.019659	NaN	0.055246	NaN	NaN	1.141915e+03	585.501214	12.730379	...
std	NaN	2.502019	0.140524	NaN	0.269207	NaN	NaN	6.447154e+03	712.809943	17.578747	...
min	NaN	0.000000	0.000000	NaN	0.000000	NaN	NaN	0.000000e+00	0.000000	0.000000	...
25%	NaN	0.000000	0.000000	NaN	0.000000	NaN	NaN	3.770000e+02	49.400000	0.000000	...
50%	NaN	1.000000	0.000000	NaN	0.000000	NaN	NaN	1.100000e+03	300.000000	5.000000	...
75%	NaN	3.000000	0.000000	NaN	0.000000	NaN	NaN	1.674000e+03	898.000000	20.000000	...
max	NaN	31.000000	2.000000	NaN	6.000000	NaN	NaN	3.251373e+06	3830.000000	714.290000	...

8 rows × 106 columns



Opérations de nettoyage effectuées

Fonctions de Prétraitement Automatisées

Préparation des données :

- Filtrer les produits français
- Supprimer les doublons
- Sélectionner les features pertinentes

Séparation en dataframes :

- Création d'un DataFrame appelé 'X' contenant les variables pertinentes prédictives
- Création d'un DataFrame appelé 'y' contenant la variable à prédire

Dimensions après le prétraitement :

- DataFrame 'X' : 55 313 lignes et 11 colonnes
- DataFrame 'y' : 55 313 lignes et 1 colonne



Opérations de nettoyage effectuées

Variables Cibles et Prédicatives

DataFrame 'y' (Variable cible) :

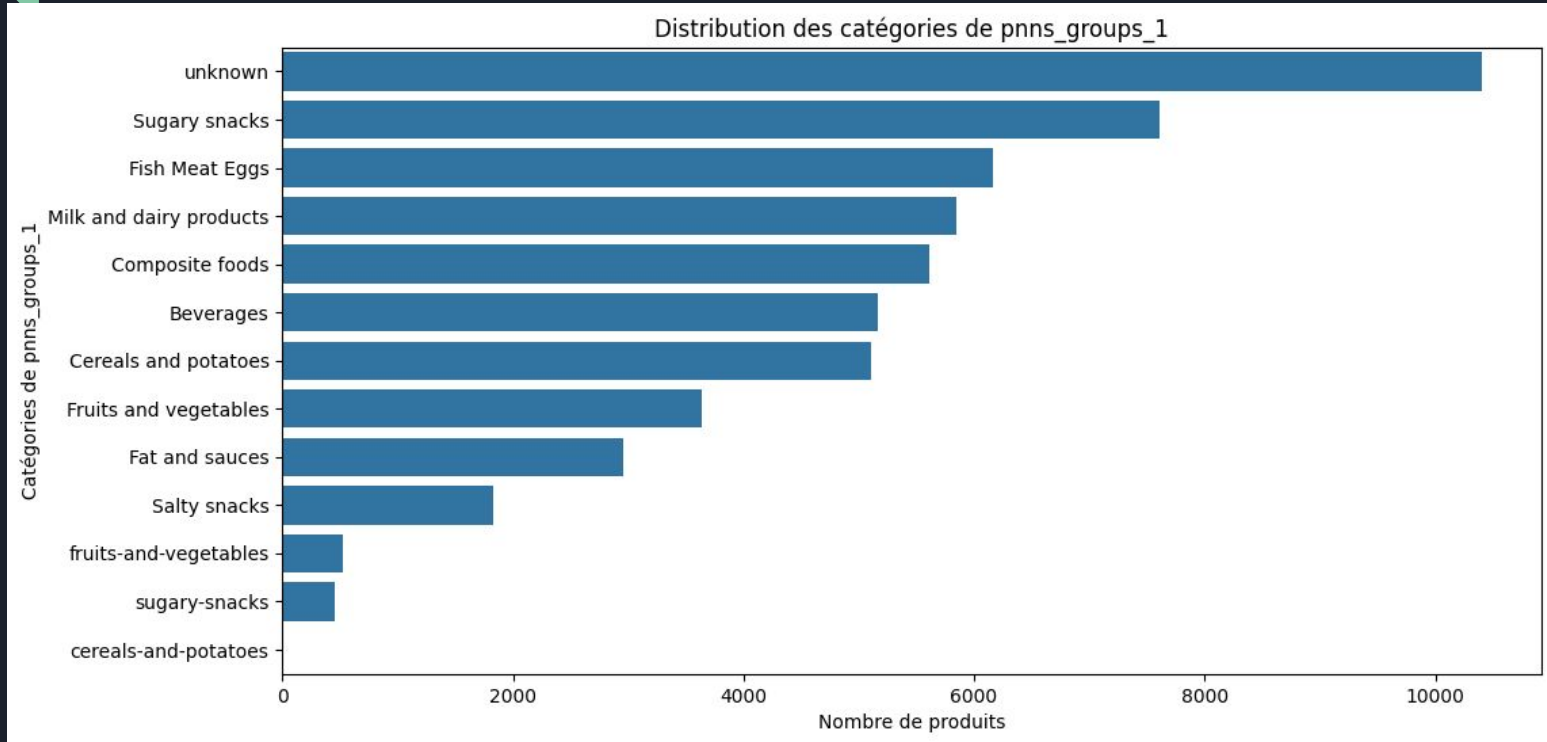
- 'pnns_groups_1' Programme National Nutrition Santé (PNNS) qui classe les produits. Prédicative, car elle regroupe les produits en catégories spécifiques.

DataFrame 'X' (Variables prédictives) :

- 'energy_100g', 'fat_100g', 'saturated-fat_100g', 'carbohydrates_100g', 'sugars_100g', 'proteins_100g', 'salt_100g', 'sodium_100g', et 'fiber_100g' fournissent des informations clés sur le contenu nutritionnel des produits.
- 'brands' peut influencer les caractéristiques des produits en fonction des marques.
- 'pnns_groups_2' fournit une classification secondaire qui enrichit l'analyse prédictive.

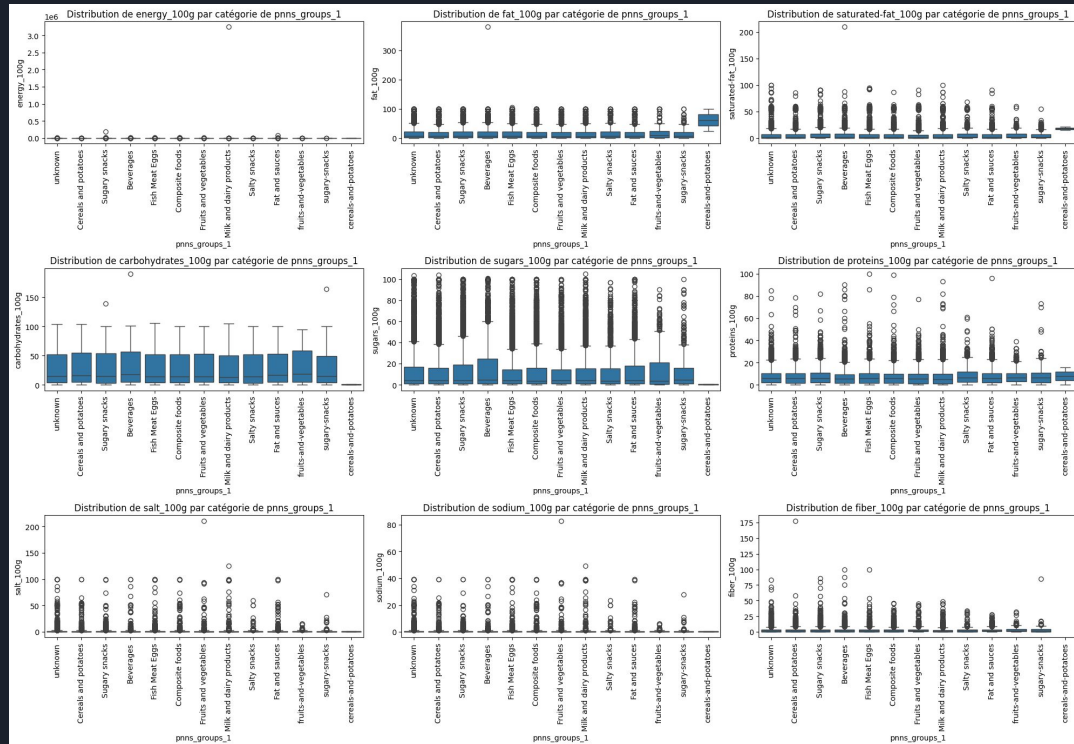
Opérations de nettoyage effectuées

Variables Cibles et Prédicatives



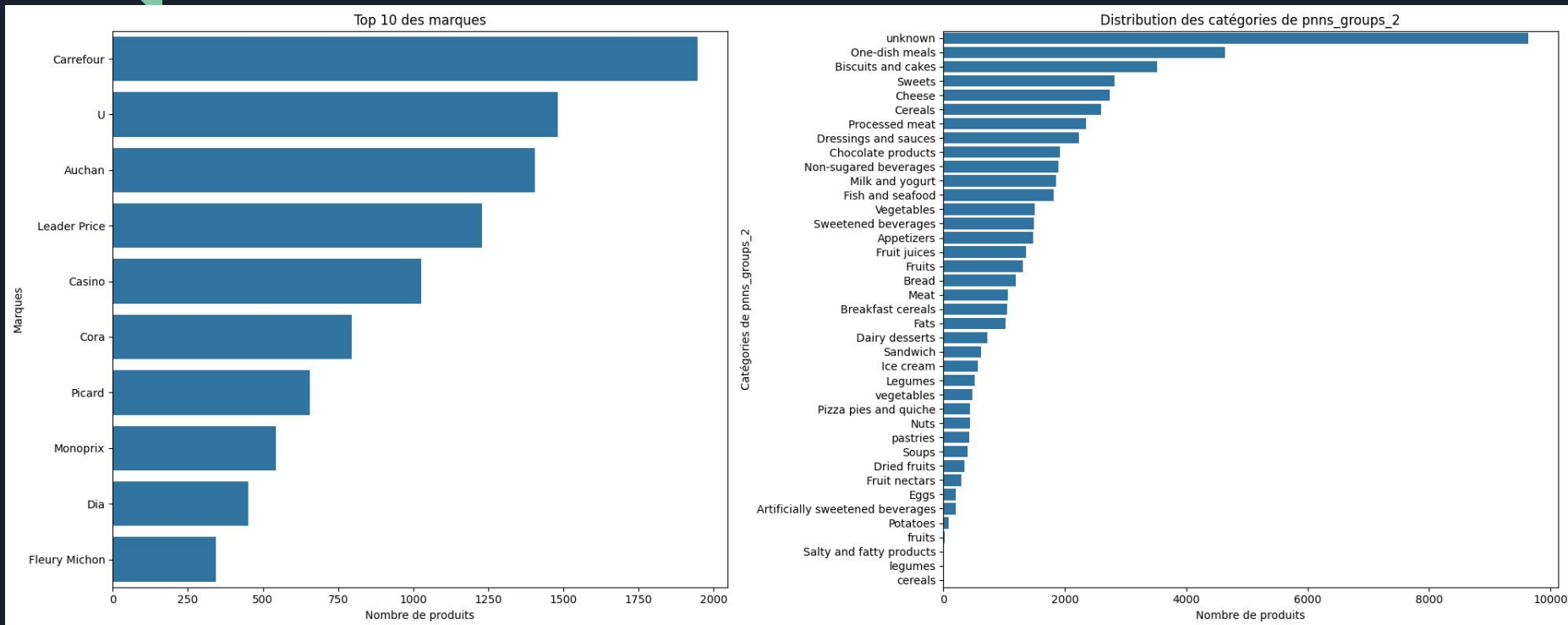
Opérations de nettoyage effectuées

Variables Cibles et Prédicatives



Opérations de nettoyage effectuées

Variables Cibles et Prédicatives





Opérations de nettoyage effectuées

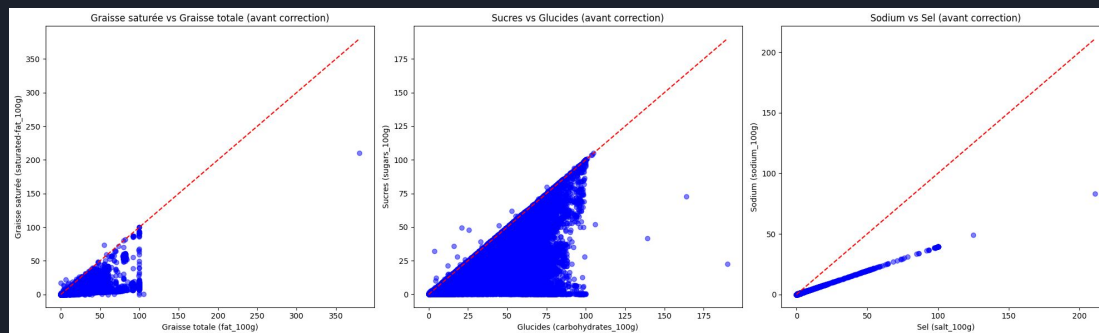
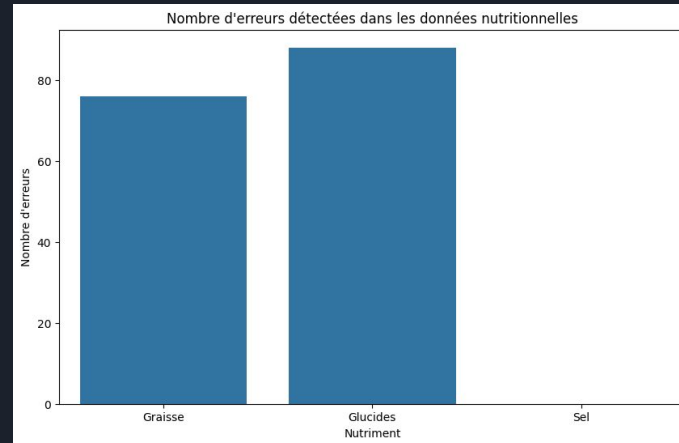
Validation et Correction des Données Nutritionnelles

Détecter et corriger les erreurs selon une approche métier :

- La graisse saturée est une composante de la graisse totale et ne peut donc pas être présente en quantité supérieure. 76 erreurs détectées.
- Le glucose est une composante des glucides. 88 erreurs détectées.
- Le sodium est une composante du sel. Aucune erreur détectée.
- Étant donné le faible nombre d'erreurs, nous supposons fortement que ces erreurs proviennent de valeurs inversées.
- Correction de ces inversions dans les données nutritionnelles.

Opérations de nettoyage effectuées

Validation et Correction des Données Nutritionnelles





Opérations de nettoyage effectuées

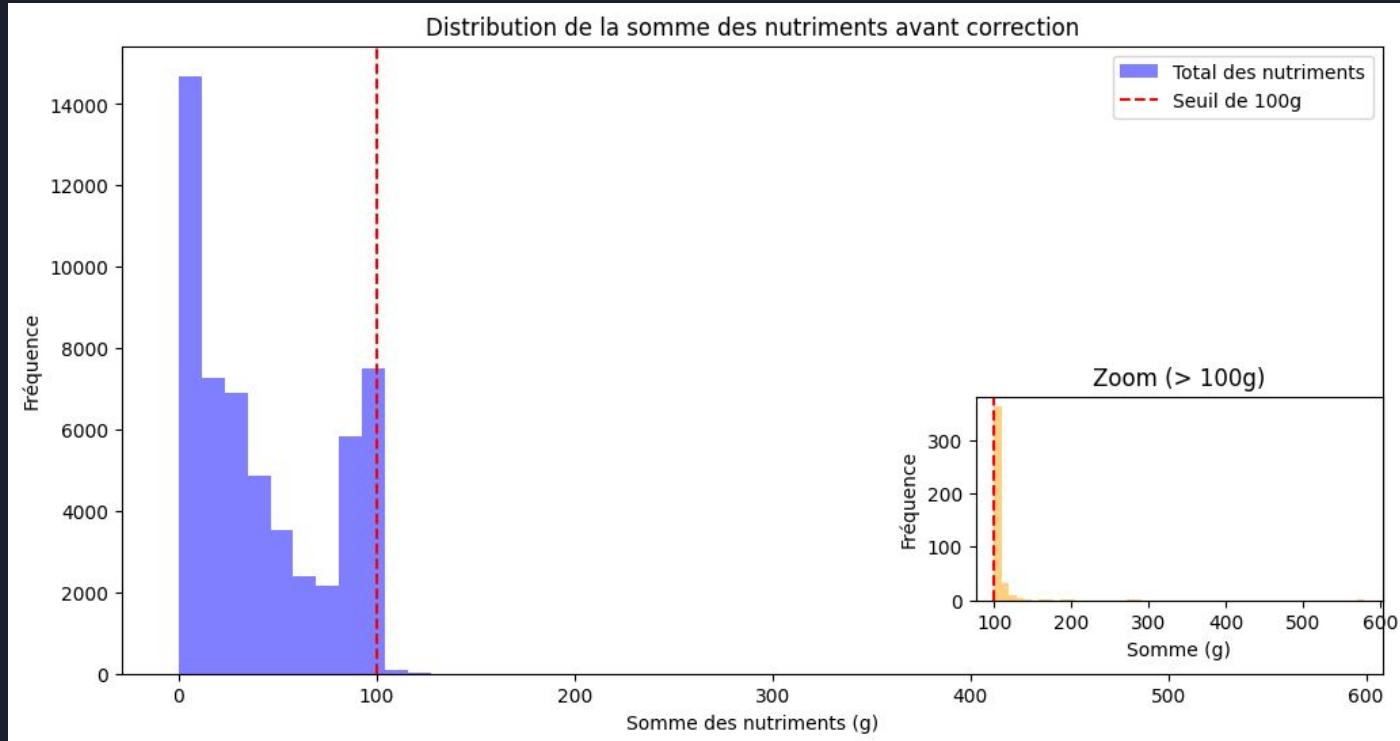
Validation et Correction des Données Nutritionnelles

Validation des données nutritionnelles :

- Pour garantir la cohérence des données, nous vérifions que la somme des nutriments ne dépasse pas 100g.
- Nous sélectionnons les colonnes 'fat_100g', 'carbohydrates_100g', 'proteins_100g', 'fiber_100g', et 'salt_100g' et calculons la somme de ces nutriments pour chaque produit. Les valeurs sont arrondies pour éviter les erreurs de précision des flottants.
- Nous filtrons ensuite les lignes où la somme des nutriments dépasse 100g.
- Il y a 417 erreurs où la somme des nutriments dépasse 100g, avec des valeurs allant de 100.01 à 579.33.
- Notre jeu de données étant constitué de 55 155 lignes, nous excluons ces lignes erronées.

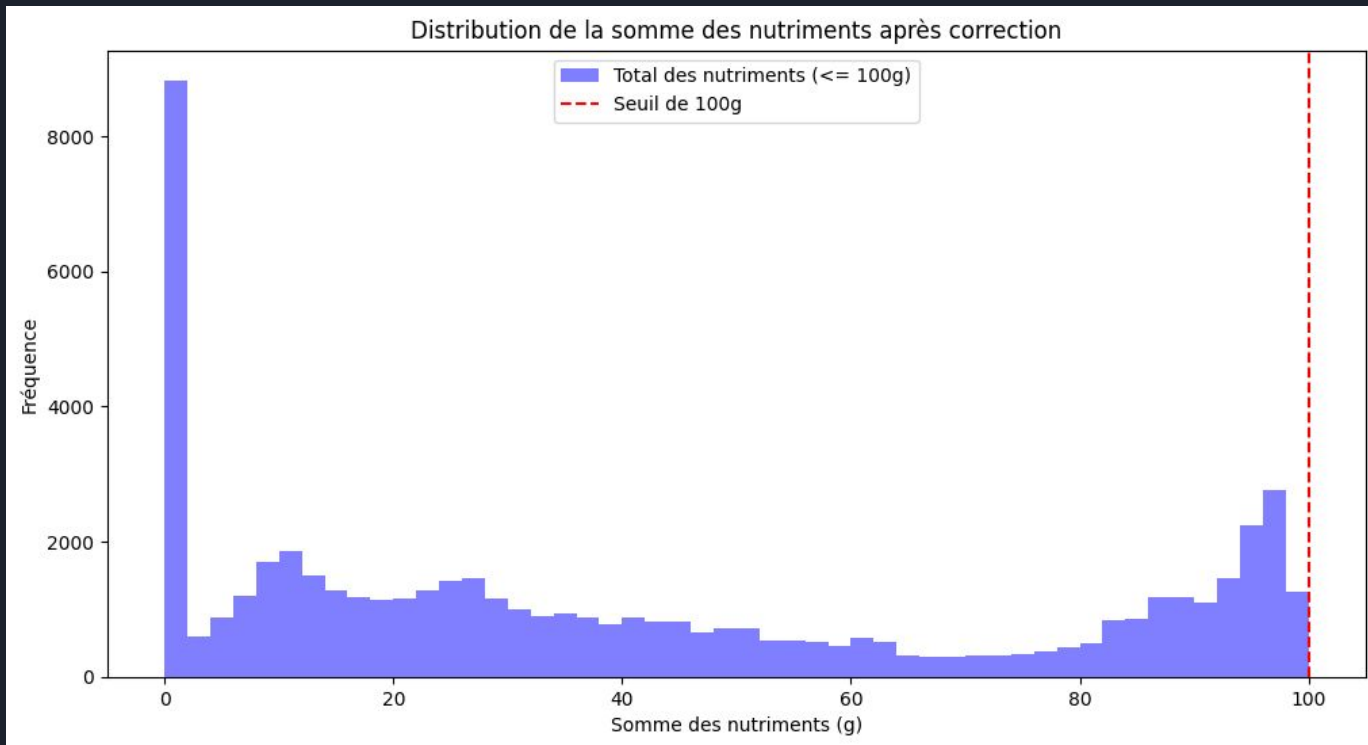
Opérations de nettoyage effectuées

Validation et Correction des Données Nutritionnelles



Opérations de nettoyage effectuées

Validation et Correction des Données Nutritionnelles





Opérations de nettoyage effectuées

Analyse et Traitement des Valeurs Aberrantes

Analyse initiale :

Nous avons calculé les valeurs uniques et leur pourcentage. Cela a révélé des incohérences dues à des différences de casse et l'utilisation de tirets.

Normalisation :

Nous avons converti les valeurs en minuscules pour uniformiser les noms. Ensuite, nous avons remplacé :

- 'cereals-and-potatoes' par 'Cereals and potatoes'
- 'fruits-and-vegetables' par 'Fruits and vegetables'
- 'sugary-snacks' par 'Sugary snacks'

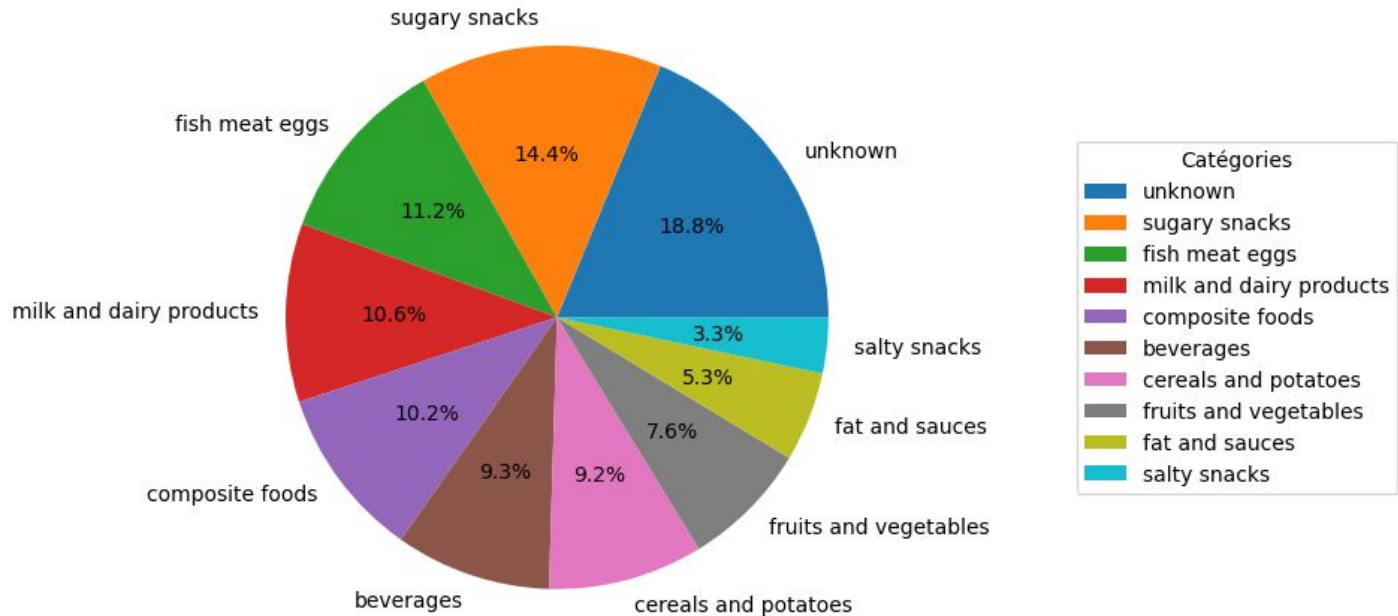
Analyse après normalisation :

Avant la normalisation, il y avait des incohérences. Après, les valeurs sont devenues cohérentes, améliorant ainsi l'analyse des données.

Opérations de nettoyage effectuées

Analyse et Traitement des Valeurs Aberrantes

Répartition des valeurs uniques dans pnns_groups_1 (après normalisation)





Opérations de nettoyage effectuées

Analyse et Traitement des Valeurs Aberrantes

Examen des valeurs uniques :

Analyse des valeurs uniques des variables 'pnns_groups_1' et 'pnns_groups_2' pour les produits ayant une valeur de 0 dans nos variables prédictives.

Catégories légitimes :

Identification des catégories de produits qui peuvent légitimement avoir une valeur de 0 et suppression des autres, car elles représentaient un nombre marginal d'individus.

Établissement de seuils :

Définition de seuils minimums et maximums pour chaque variable prédictive en suivant une approche métier. Suppression des individus en dehors de ces seuils, représentant une quantité négligeable d'individus.

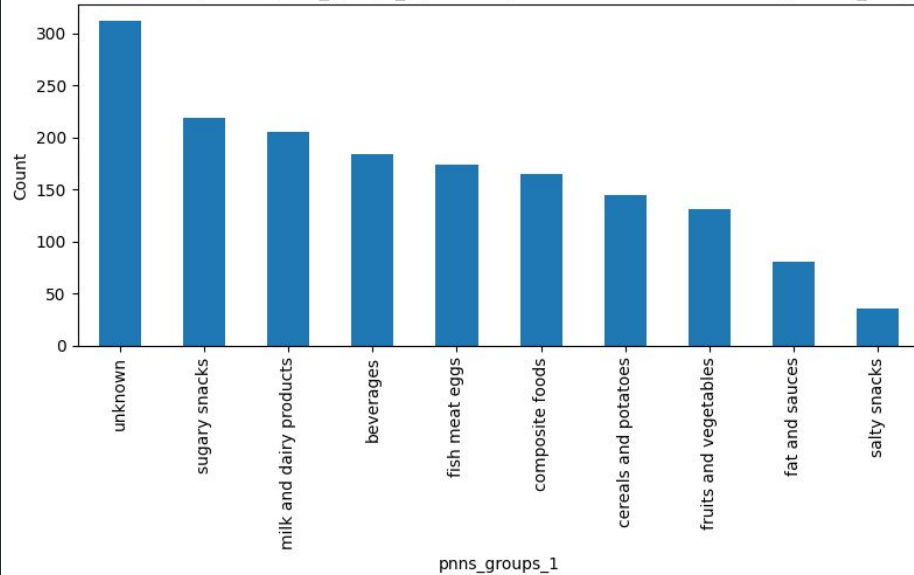
Résultats :

Suppression de 3 279 individus, laissant un total de 51 459 individus restants.

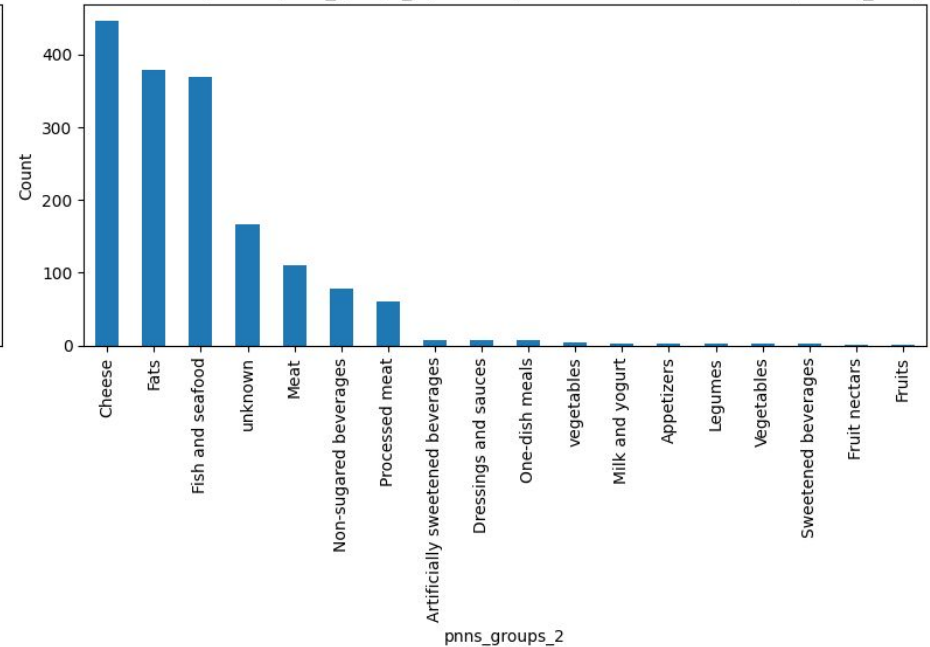
Opérations de nettoyage effectuées

Analyse et Traitement des Valeurs Aberrantes

Valeurs uniques de pnns_groups_1 pour les produits avec 0 dans carbohydrates_100g



Valeurs uniques de pnns_groups_2 pour les produits avec 0 dans carbohydrates_100g





Opérations de nettoyage effectuées

Traitement des valeurs manquantes

Les valeurs manquantes ont été traitées de différentes manières en fonction de la nature des produits.

- Les valeurs manquantes pour `'fiber_100g'` ont été remplacées par 0, car il est probable que l'absence de cette donnée indique un produit sans fibres.
- Pour les features nutritionnelles `'carbohydrates_100g'`, `'fat_100g'`, `'salt_100g'`, `'proteins_100g'`, les valeurs manquantes ont été initialement remplacées par 0 pour les catégories de produits où cette absence est légitime. Les valeurs restantes ont été imputées par la médiane de chaque catégorie `'pnns_groups_2'`.
- En complément, pour les features `'saturated-fat_100g'`, `'sugars_100g'` et `'sodium_100g'`, une imputation par KNN (K-Nearest Neighbors) a été utilisée.
- Pour la feature `'brands'`, les valeurs manquantes ont été remplacées par une catégorie `'unknown'`, permettant de distinguer clairement les produits sans marque renseignée tout en conservant toutes les observations.



Opérations de nettoyage effectuées

Imputation des Valeurs Manquantes

Pour traiter 'pnns_groups_1', nous avons décidé d'utiliser une méthode d'imputation par KNN (K-Nearest Neighbors) pour combler ces lacunes.

Encodage et Préparation des Données :

- La colonne 'pnns_groups_1' a été encodée en valeurs numériques, en remplaçant les valeurs 'unknown' par NaN.
- Les données ont été synchronisées et fusionnées pour préparer l'imputation.

Analyse de Corrélation :

Nous avons déterminé et sélectionné les features ayant une corrélation supérieure à 0.5 avec 'pnns_groups_1_encoded' pour l'imputation.



Opérations de nettoyage effectuées

Imputation des Valeurs Manquantes

Imputation par KNN :

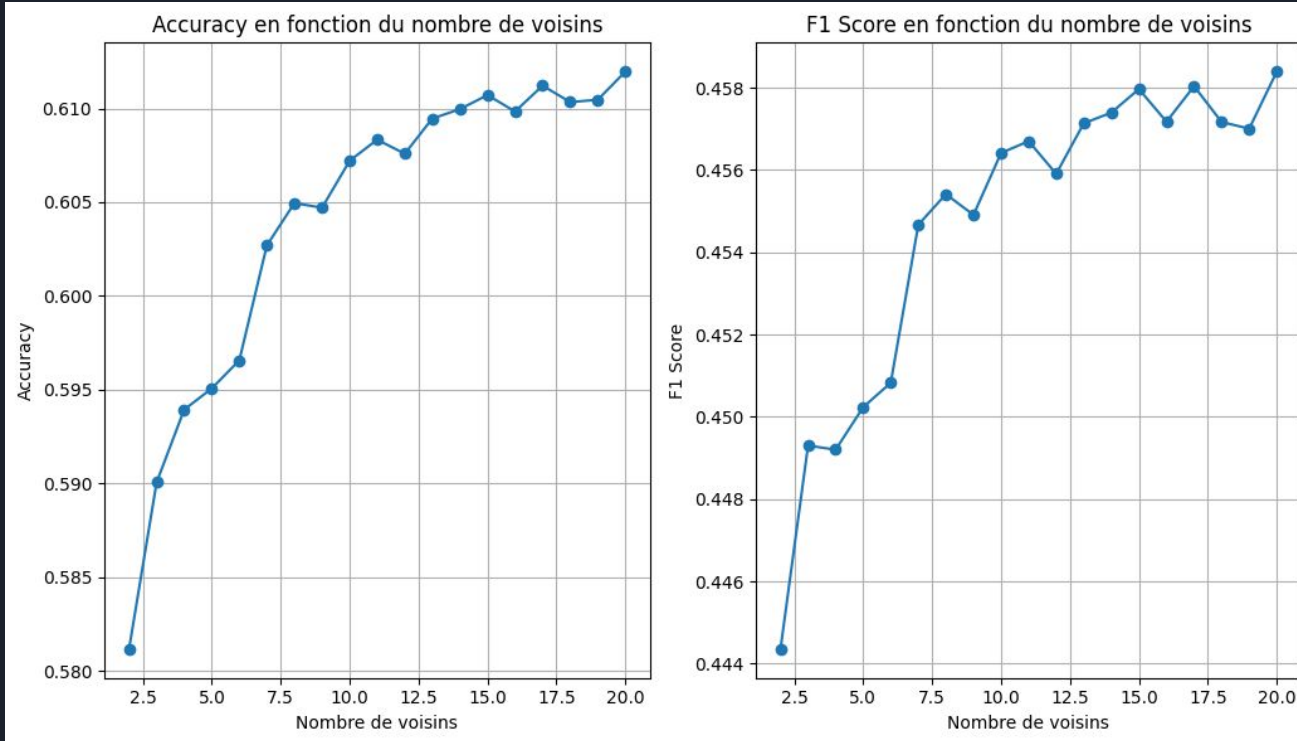
- Nous avons testé différents nombres de voisins (de 2 à 20) pour la méthode KNN afin de déterminer le meilleur paramètre pour l'imputation.
- Nous avons tracé les courbes d'Accuracy et F1 Score en fonction du nombre de voisins pour identifier le nombre optimal de voisins.
- L'analyse a révélé que 5 voisins étaient le nombre optimal pour maximiser l'accuracy et le F1 score.

Résultats de l'Imputation :

- Les valeurs manquantes ont été imputées en utilisant 5 voisins.
- Les distributions des valeurs avant et après imputation ont été comparées pour s'assurer de la cohérence des résultats.

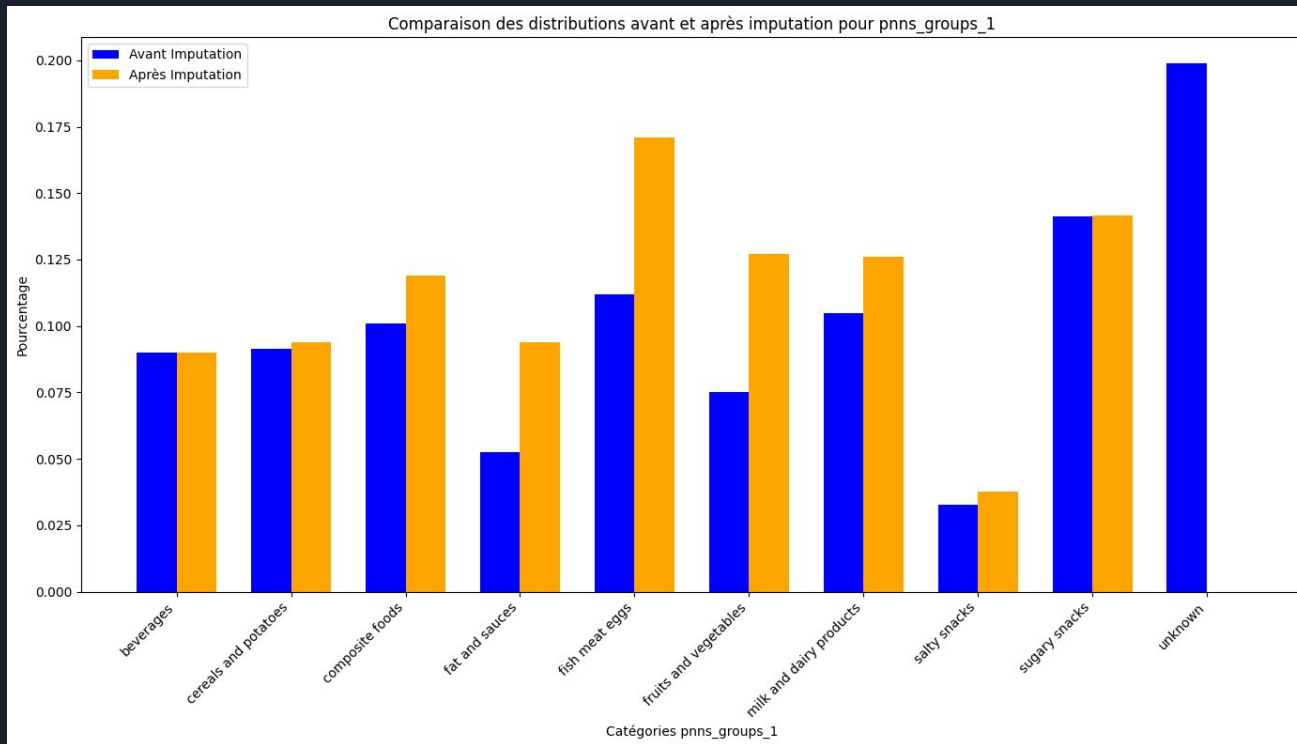
Opérations de nettoyage effectuées

Imputation des Valeurs Manquantes



Opérations de nettoyage effectuées

Imputation des Valeurs Manquantes





Analyse univariée des variables

Statistiques Descriptives des Variables Nutritionnelles :

Le tableau ci-dessous résume les statistiques descriptives des principales variables nutritionnelles

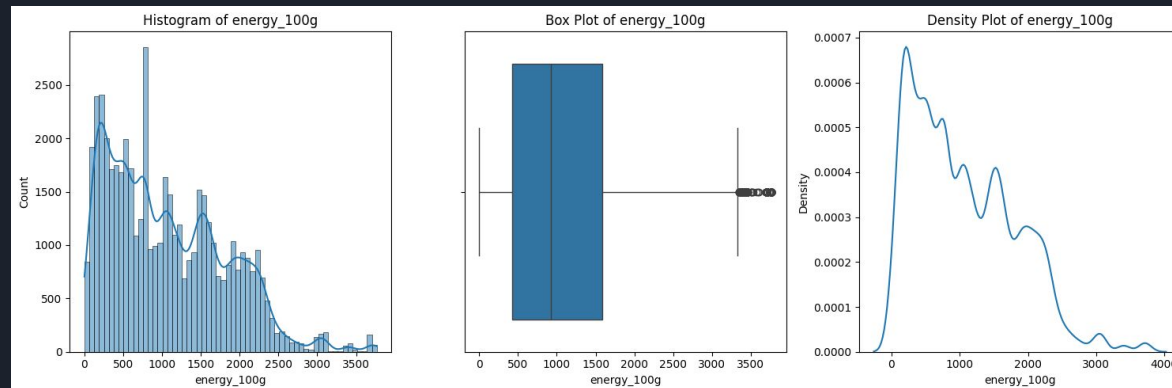
Variable	Moyenne	Médiane	Écart-type	Minimum	Maximum
`energy_100g`	1058.73	920	752.03	0	3766
`fat_100g`	12.70	8	16.01	0	100
`saturated-fat_100g`	5.07	2.10	7.73	0	100
`carbohydrates_100g`	25.90	14.70	25.58	0	100
`sugars_100g`	13.77	4.30	18.30	0	100
`proteins_100g`	7.41	5.80	7.08	0	93.10
`salt_100g`	1.03	0.61	3.63	0	100
`sodium_100g`	0.39	0.20	1.43	0	39.37
`fiber_100g`	1.48	0.00	3.13	0	100

Analyse univariée des variables

Visualisations des Variables Nutritionnelles :

Pour mieux comprendre la distribution de chaque variable nutritionnelle, nous allons utiliser trois types de graphiques : un histogramme, un box plot et un density plot. Ces graphiques nous aideront à visualiser la fréquence, la dispersion et la densité des données.

Chaque variable sera représentée par un histogramme pour montrer la fréquence des valeurs, un box plot pour visualiser la dispersion et les valeurs extrêmes, et un density plot pour voir la distribution de densité.





Analyse univariée des variables

Conclusion de l'Analyse Univariée

L'analyse univariée des variables nutritionnelles nous a permis de mieux comprendre la distribution des valeurs dans notre jeu de données. Les visualisations montrent clairement les tendances et les caractéristiques des différentes variables, ce qui est crucial pour la prochaine étape de notre projet.

Base Solide pour l'Auto-Complétion :

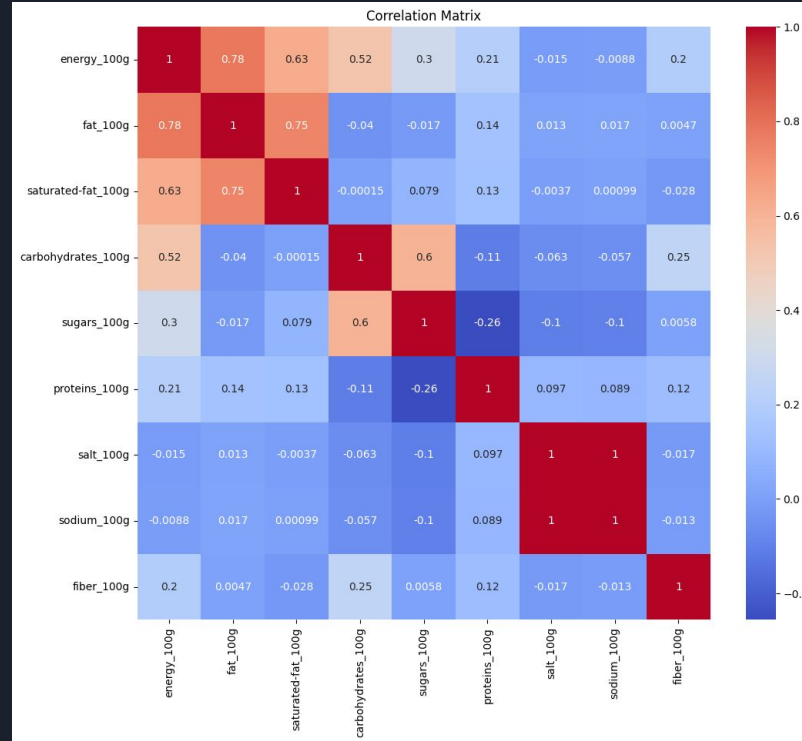
Comprendre les distributions des variables nutritionnelles nous aide à créer des modèles de prédiction plus précis pour l'auto-complétion des données manquantes. Cela est essentiel pour améliorer l'efficacité du système de suggestion.

Conclusion :

En conclusion, l'analyse univariée a validé la qualité de nos données et posé les bases pour des modèles de prédiction robustes, contribuant directement à l'amélioration de la base de données Open Food Facts pour des initiatives de santé publique plus efficaces.

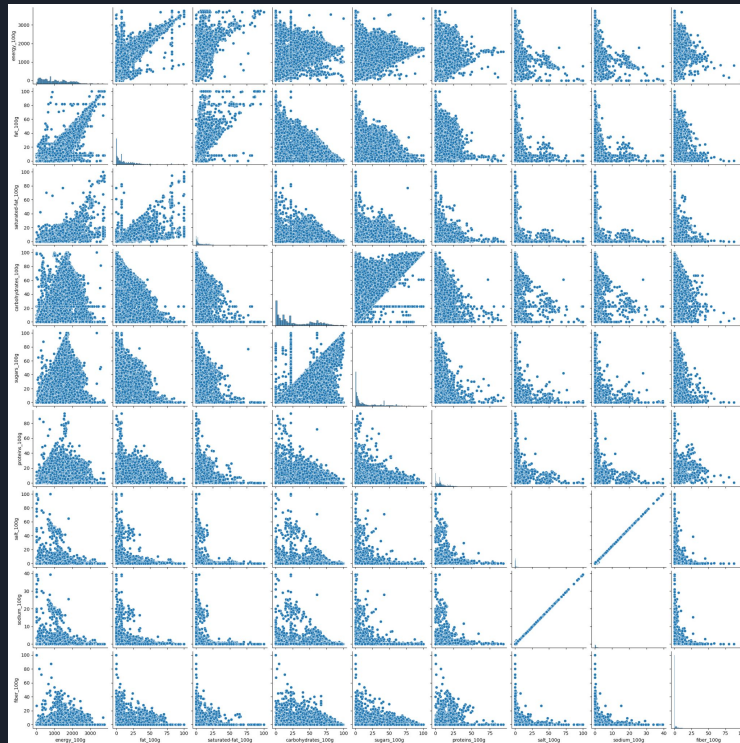
Analyse bivariée, multivariée et résultats

Matrice de Corrélation



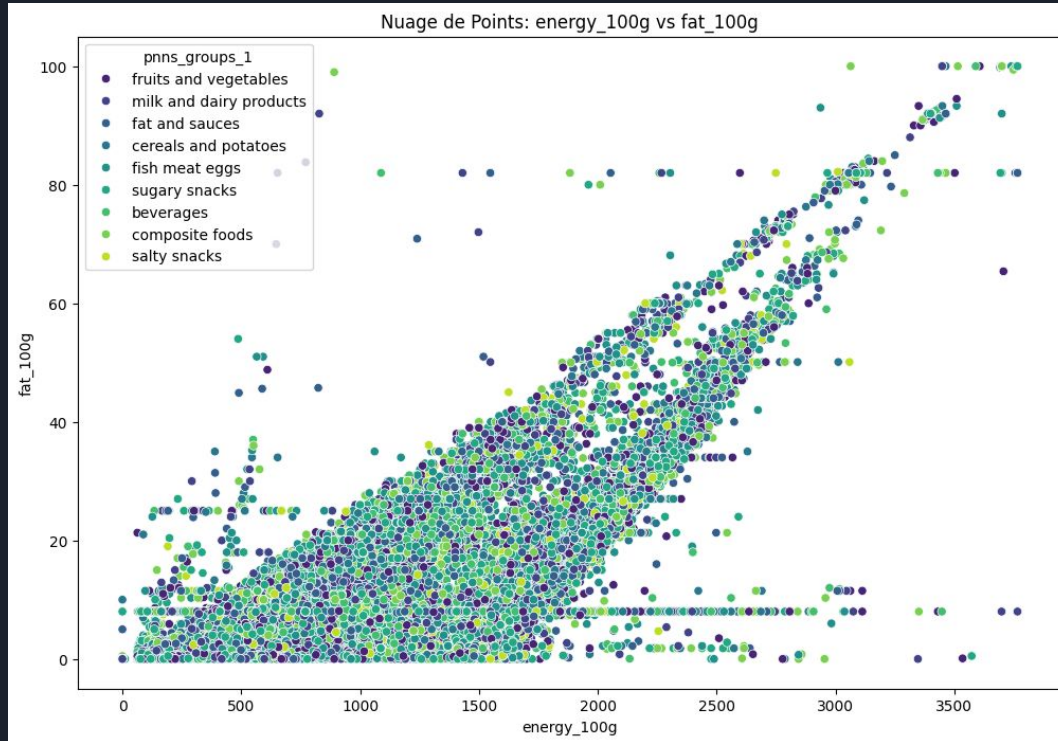
Analyse bivariée, multivariée et résultats


Analyse bivariée



Analyse bivariée, multivariée et résultats

Analyse bivariée





Analyse bivariée, multivariée et résultats

Importance de PC1 et PC2

Nous avons **standardisé nos données** avant de réaliser l'**Analyse en Composantes Principales (ACP)** pour nous assurer que chaque variable contribue de manière égale à l'analyse.

Variance Expliquée :

La première composante principale (PC1) explique 29% de la variance totale des données.

La deuxième composante principale (PC2) explique 23% de la variance totale des données.

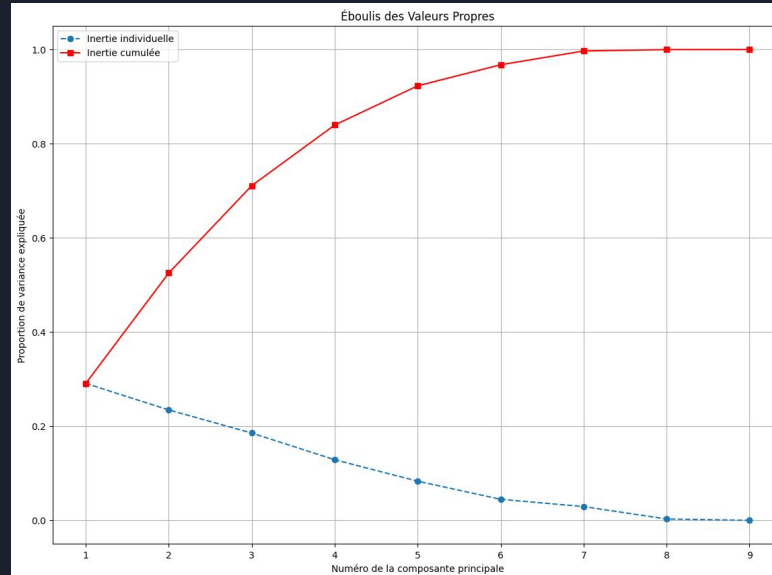
Importance de PC1 et PC2 :

Ensemble, PC1 et PC2 capturent 52% de la variance totale des données. Cela signifie que plus de la moitié des informations présentes dans les données initiales sont représentées par ces deux composantes. Cela simplifie l'analyse en réduisant la dimensionnalité tout en conservant l'essentiel de l'information.

Analyse bivariée, multivariée et résultats

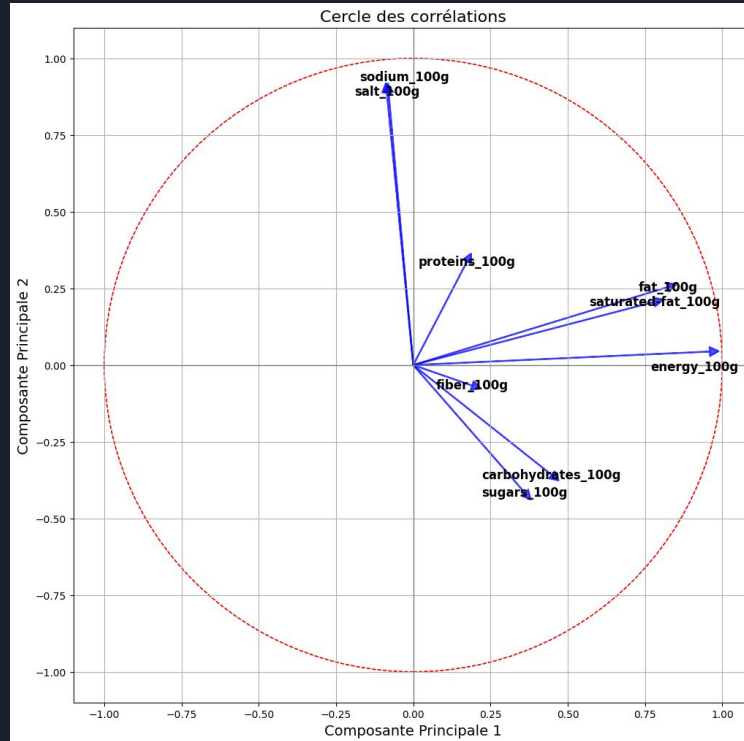
Éboulis des Valeurs Propres

Le graphique ci-dessous montre l'éboulis des valeurs propres pour nos données. Les points bleus indiquent l'inertie individuelle de chaque composante, et la ligne rouge montre l'inertie cumulée.



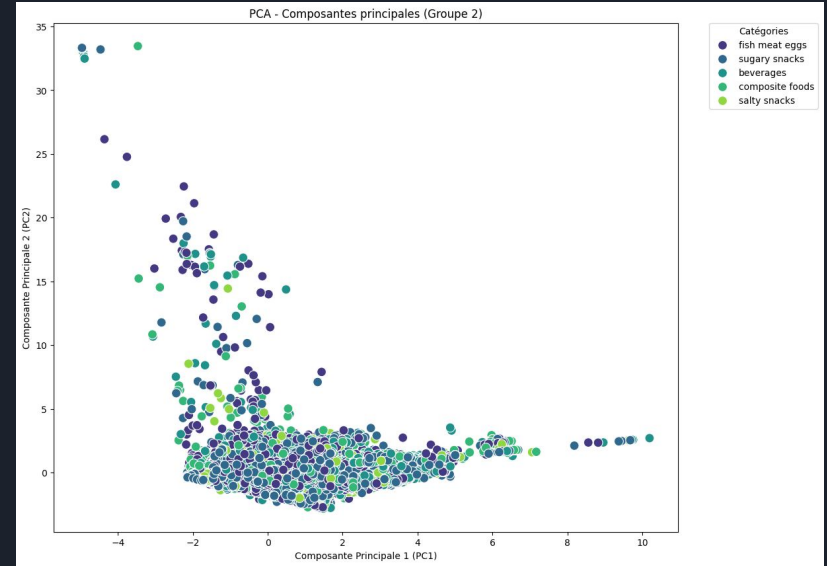
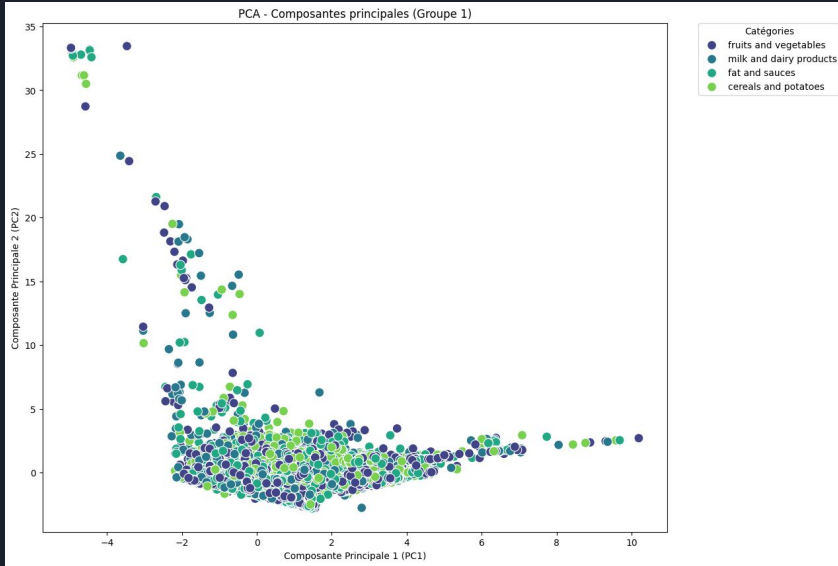
Analyse bivariée, multivariée et résultats

Répartition des Données dans les Nouvelles Dimensions Réduites



Analyse bivariée, multivariée et résultats

Visualisations 2D et 3D des Composantes Principales

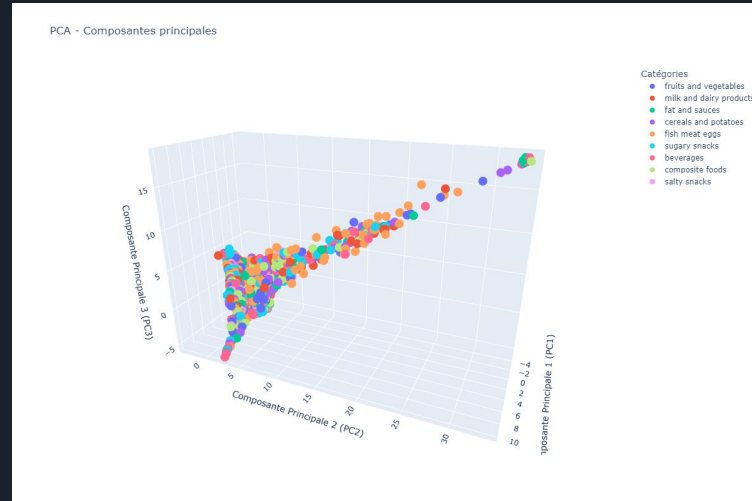


Analyse bivariée, multivariée et résultats

Visualisations 2D et 3D des Composantes Principales

Visualisation 3D des Composantes Principales :

Le graphique suivant montre les trois premières composantes principales (PC1, PC2, et PC3) en 3D, colorées selon les groupes de données ('pnns_groups_1'). Cela représente 71% de la variance totale des données.





Analyse bivariée, multivariée et résultats

Analyse de la Variance (ANOVA)

Pour chaque variable nutritionnelle sélectionnée ('energy_100g', 'fat_100g', 'saturated_fat_100g', 'carbohydrates_100g', 'sugars_100g', 'proteins_100g'), une ANOVA a été réalisée :

- energy_100g : F-value : 2.16 / p-value : 0.0269 (significatif)
- fat_100g : F-value : 5.40 / p-value : 8.005336e-07 (très significatif)
- saturated_fat_100g : F-value : 6.31 / p-value : 3.299738e-08 (très significatif)
- carbohydrates_100g : F-value : 8.28 / p-value : 2.771683e-11 (très significatif)
- sugars_100g : F-value : 10.65 / p-value : 4.411682e-15 (très significatif)
- proteins_100g : F-value : 2.26 / p-value : 0.0203 (significatif)

Les résultats montrent que les différences observées entre les groupes de produits alimentaires sont significatives pour toutes les variables nutritionnelles analysées. Cela signifie que les catégories de produits alimentaires (pnns_groups_1) influencent significativement les valeurs nutritionnelles, ce qui est crucial pour le développement de modèles prédictifs précis. Cependant, les variables ne suivent pas une loi normale et les résultats de l'ANOVA en sont donc affecté.



3 observations sur l'application client

1 - Forte Corrélation entre Variables Nutritionnelles :

- **Observation** : Nos analyses ont révélé des corrélations significatives entre plusieurs variables nutritionnelles, comme la forte corrélation positive entre 'energy_100g' et 'fat_100g' (0.78) et entre 'fat_100g' et 'saturated-fat_100g' (0.75).
- **Preuve** : La matrice de corrélation montre que ces variables sont étroitement liées, ce qui signifie qu'il est possible de prédire 'energy_100g' à partir de 'fat_100g', et 'saturated-fat_100g' à partir de 'fat_100g'.
- **Impact** : Cela indique que l'auto-complétion des données manquantes est faisable avec un niveau de précision acceptable, car les valeurs manquantes de ces variables peuvent être estimées de manière fiable à partir des variables fortement corrélées.



3 observations sur l'application client

2 - Réduction de la Dimensionnalité et Visualisation en Composantes Principales :

- **Observation** : L'Analyse en Composantes Principales (ACP) a montré que les cinq premières composantes principales capturent environ 90% de la variance totale des données.
- **Preuve** : L'éboulis des valeurs propres (scree plot) et les visualisations en 2D et 3D des composantes principales démontrent que les principales variations des données sont capturées par ces composantes.
- **Impact** : En réduisant la dimensionnalité, nous simplifions l'analyse tout en conservant l'essentiel de l'information, rendant la modélisation plus efficace et les visualisations plus compréhensibles.



3 observations sur l'application client

3 - Répartition et Corrélation des Données dans les Nouvelles Dimensions :

- **Observation** : Le cercle des corrélations a montré comment les variables nutritionnelles se répartissent dans les nouvelles dimensions définies par les composantes principales. Par exemple, 'sodium_100g' et 'salt_100g' sont fortement corrélées avec PC2, tandis que 'carbohydrates_100g' et 'sugars_100g' montrent une corrélation négative avec PC2.
- **Preuve** : Le graphique du cercle des corrélations montre clairement la répartition et les relations entre les variables dans l'espace des composantes principales.
- **Impact** : Cela aide à identifier les variables les plus importantes pour expliquer la variance dans les données, facilitant ainsi l'amélioration des modèles prédictifs et validant la faisabilité de l'auto-complétion des données nutritionnelles.



Respect des principes RGPD

Licéité, Loyauté et Transparence :

Toutes les données utilisées proviennent de sources publiques et ouvertes, en particulier du site Open Food Facts. Le traitement des données se fait de manière transparente, en respectant les droits des utilisateurs et en s'assurant que les données sont utilisées uniquement à des fins de recherche et d'amélioration de la qualité nutritionnelle.

Limitation des Finalités :

Les données sont collectées et traitées exclusivement pour le but spécifique de ce projet, qui est de vérifier la faisabilité d'un modèle d'auto-complétion des données nutritionnelles. Aucune donnée n'est utilisée à des fins autres que celles explicitement définies dans le cadre du projet.



Respect des principes RGPD

Minimisation des Données :

Seules les données nécessaires pour atteindre les objectifs du projet sont collectées et traitées. Nous nous assurons de ne collecter et traiter que les variables nutritionnelles pertinentes, sans inclure de données superflues.

Exactitude :

Il est crucial que les données traitées soient exactes et mises à jour. Les données nutritionnelles utilisées sont vérifiées pour leur exactitude et nettoyées avant l'analyse afin de garantir leur fiabilité.

Conservation Limitées des Données :

Les données ne doivent pas être conservées plus longtemps que nécessaire pour atteindre les objectifs du projet. Les données sont supprimées après la fin du projet et une fois que les analyses et rapports nécessaires ont été produits, garantissant ainsi une durée de conservation minimale.



Conclusions sur la faisabilité du projet

- **Pertinence des Données :** Les données analysées montrent des **relations** suffisamment **fortes** entre les variables nutritionnelles pour soutenir la faisabilité d'un modèle d'auto-complétion. Les corrélations significatives et les analyses de variance **confirment** que les **valeurs nutritionnelles** peuvent être **prédites** avec une bonne précision en utilisant les catégories de produits.
- **Conformité aux Normes de Qualité des Données :** En nettoyant et en standardisant les données avant l'analyse, nous nous sommes assurés de la **fiabilité** et de la **précision** des résultats, respectant ainsi les normes de qualité des données nécessaires pour des **modèles prédictifs robustes**.
- **Impact Potentiel pour Santé Publique France :** Un **système d'auto-complétion** basé sur ces analyses permettrait à Santé Publique France de compléter automatiquement les données nutritionnelles manquantes, **réduisant les erreurs de saisie** et **améliorant la qualité et la cohérence des données publiées**.

Avez-vous des questions?

