Amélioration de la base de données

open FOOD facts

Santé publique France

SOMMAIRE

Enjeux

La **nutrition** (alimentation et activité physique) est un **déterminant majeur de la santé** et notamment des **pathologies chroniques** (…) comme les maladies cardiovasculaires et le diabète.

Source : Pr Serge Hercberg - Propositions pour un nouvel élan de la politique nutritionnelle française de santé publique

Santé publique France

OPEN FOOD FACTS

Contexte

- **Améliorer la base de données Open Food Facts avec une aide pour un remplissage plus efficace**
  - Open Food Facts permet de connaître la qualité nutritionnelle de produits.
  - Enrichissement de la base de données nécessitant de remplir de nombreux champs textuels et numériques

    => erreurs de saisie et valeurs manquantes

Mission

- **Déterminer la faisabilité de créer un système de suggestion ou d'auto-complétion**
  - Nettoyage des données, y compris imputations des valeurs manquantes
  - Exploration des données
  - Conclusions sur la faisabilité et les idées d'applications

Quelles applications ?

Les critères

Fibres


Sel


Acides gras insaturés


Sucre


Fruits et légumes


Matières grasses saturées

Proposer des alternatives de produits avec moins sucre / sel / matières grasses

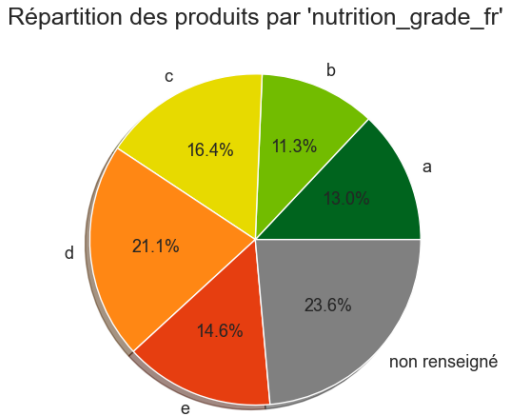
Proposer un panier avec les articles les plus sains à partir d'une liste de courses

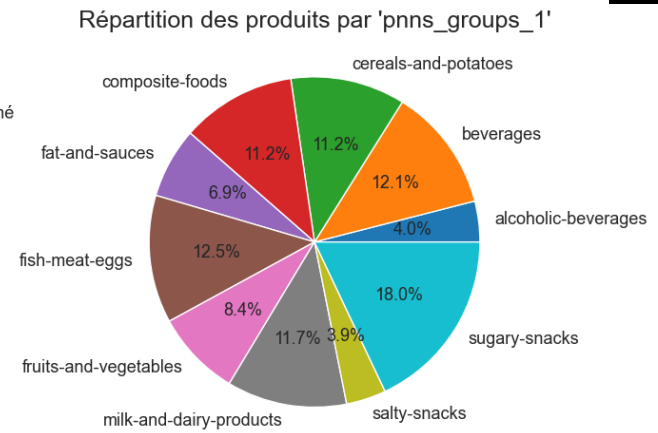
Proposer des menus adaptés aux recommandations médicales.

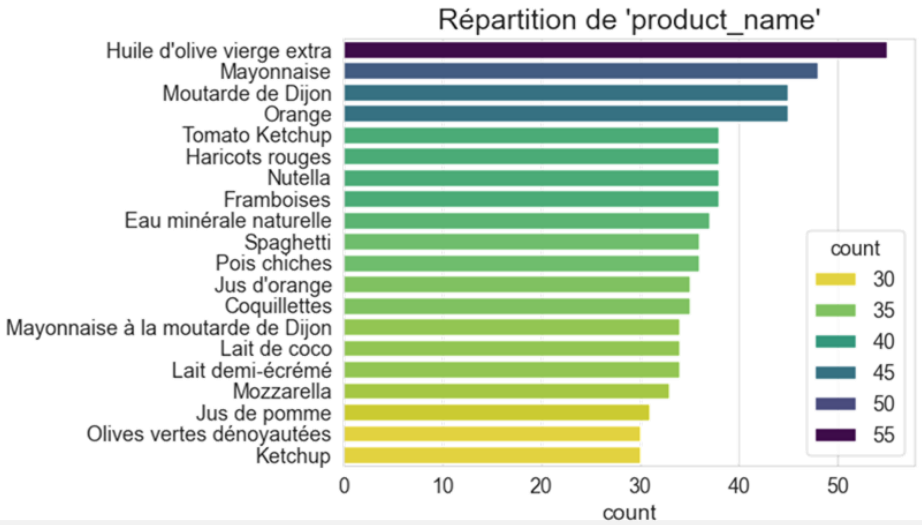
Une application de qualité nécessite des informations correctes et complètes.

A propos des données

Base de données Open source







Type	Description
Informations générales	Code barre, nom du produit, contributeur, date de création/modification, etc.
Ensemble de tags	Catégorie du produit, localisation, origine, etc.
Ingrédients	Composants des produits et leurs additifs éventuels
Informations nutritionnelles	Quantité en g, µg ou % d'un nutriment pour 100 grammes du produit

5 grands principes du Règlement Général sur la Protection des Données

FINALITÉ

Informations enregistrées et utilisées dans un but bien précis, légal et légitime

PROPORTIONNALITÉ ET PERTINENCE

Informations enregistrées pertinentes et strictement nécessaires

DURÉE DE CONSERVATION LIMITÉE

Durée de conservation précise fixée en fonction du type d'informations et de la finalité du fichier

SÉCURITÉ ET CONFIDENTIALITÉ

Restrictions d'accès aux informations enregistrées aux seules personnes autorisées

DROITS DES PERSONNES

Droit à l'information, recueil du consentement, droit d'opposition, droits d'accès et rectification, droit à la portabilité

Le contributeur peut avoir utilisé ses nom et prénom comme identifiant : ces données n'étant pas utiles pour notre application, nous ne devons pas les conserver dans la base de données open source.

Nettoyage des données

Démarche du nettoyage

Inspection des données

- **320 772 lignes, 162 colonnes**
- **Clé primaire : « code »**
- **Valeurs aberrantes:**
  - Informations nutritionnelles négatives ou supérieures à 100g ou 100%
- **Beaucoup de valeurs manquantes**
  - 16 colonnes entièrement vides
  - 75% des colonnes avec plus de 56% de valeurs manquantes

Nettoyage

- **80 249 lignes, 27 colonnes**
- **Etapes du nettoyage**
  - Réduire l'échantillon de travail
  - Doublons de « code »
  - Types de variables incorrects
  - Traitement des valeurs aberrantes
  - Traitement des valeurs manquantes

- Produits vendus en **France et DOM/TOM**
- Sélection des données pertinentes
- Suppression des colonnes vides et lignes sans données pour les données pertinentes
- Identifier et traiter les doublons pour la clé primaire « code »

Variables retenues

```
selection = [
 'code',
 'created_datetime',
 'last_modified_datetime',
 'product_name',
 'brands',
 'pnns_groups_1',
 'pnns_groups_2',
 'main_category_fr',
 'energy_100g',
 'fat_100g',
 'saturated-fat_100g',
 'monounsaturated-fat_100g',
 'polyunsaturated-fat_100g',
 'trans-fat_100g',
 'carbohydrates_100g',
 'sugars_100g',
 'starch_100g',
 'polyols_100g',
 'fiber_100g',
 'proteins_100g',
 'salt_100g',
 'sodium_100g',
 'alcohol_100g',
 'fruits-vegetables-nuts_100g',
 'nutrition-score-fr_100g',
 'nutrition_grade_fr'
]
```

Filtrage des données

Réduire la taille du jeu de données pour tester la faisabilité de l'application

Ranges acceptables :

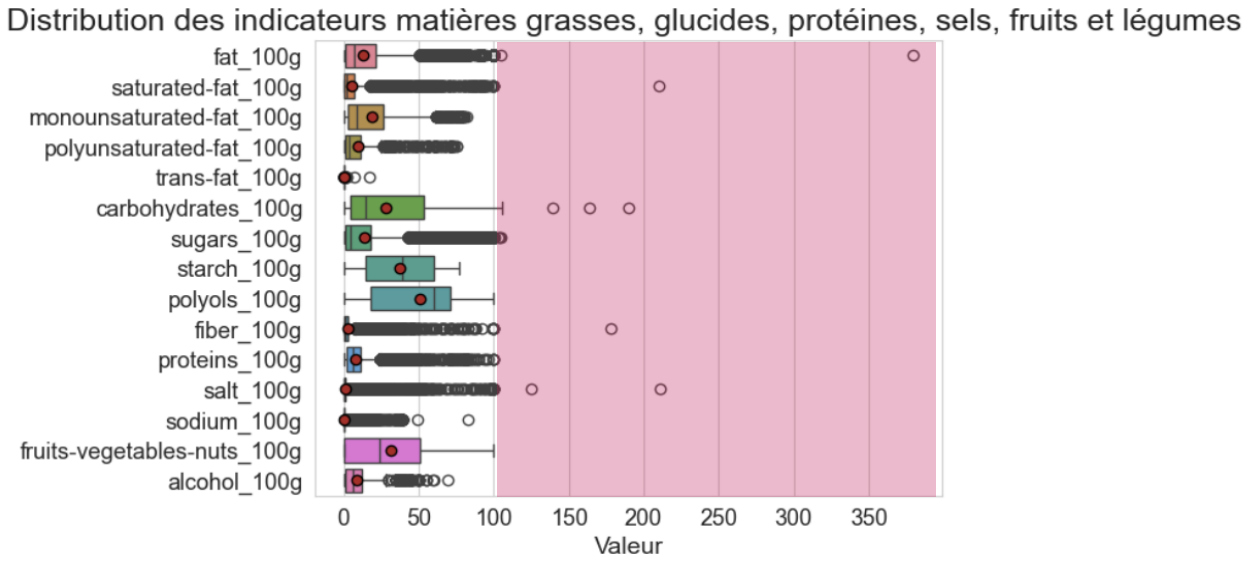
- energy_100g : [0, 5000] kJ/100g
- nutrition-score-fr_100g : [-15, 40]
- sodium [0, 40]
- autres variables « _100g » : [0, 100] g ou %

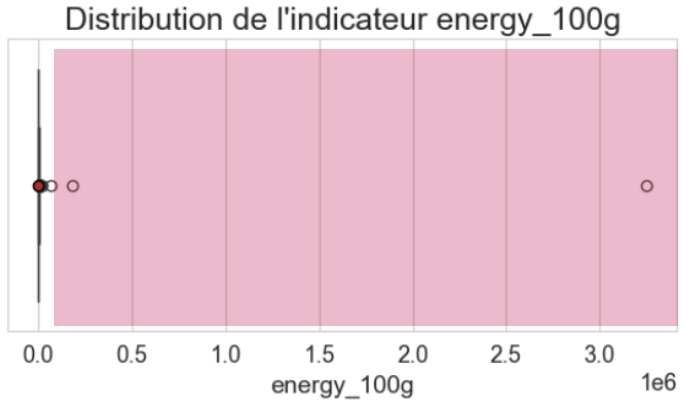
Relations entre les variables :

- salt = sodium × 2,5
- fat = saturated-fat+ monounsaturated-fat+ polyunsaturated-fat+ trans-fat
- carbohydrates= sugars+ starch+ polyols
- Σ informations nutritionnelles <= 100g

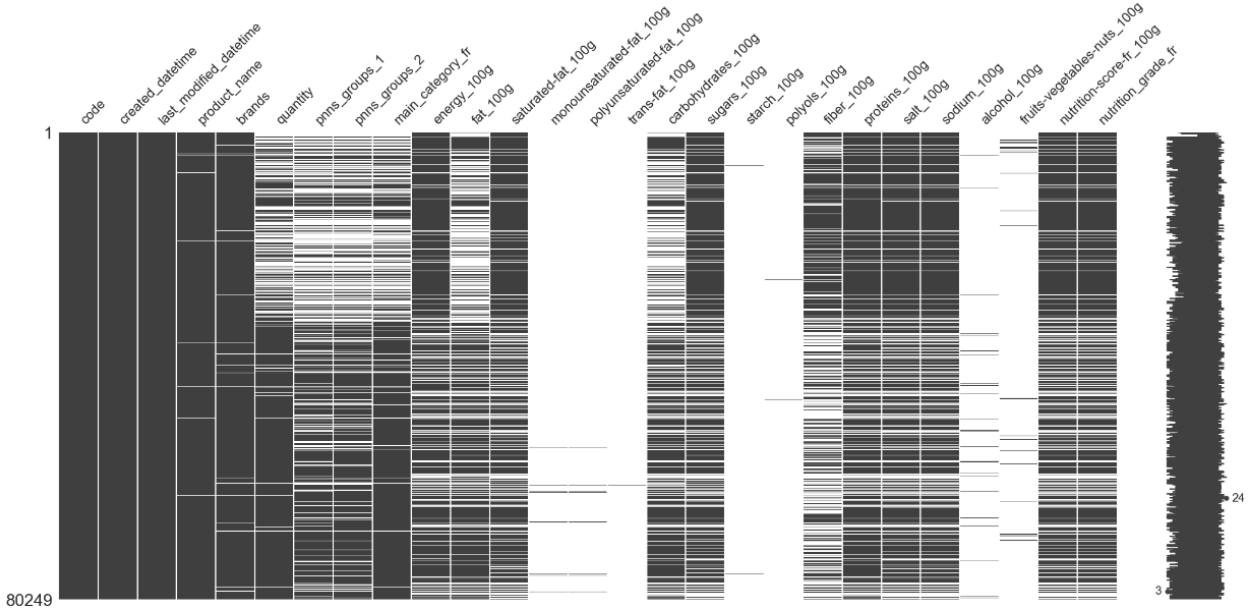
Valeurs aberrantes

Identifier les valeurs atypiques et les traiter





Moitié des variables > 25% valeurs manquantes
dont 7 variables avec + de 95%



2 types de valeurs manquantes :

- Variables catégorielles : 6
- Variables numériques : 20

Valeurs manquantes

Identifier les valeurs manquantes et les traiter

Traitements des valeurs manquantes

Variables catégorielles

- **Quoi ?**
  - pnns_groups_1, pnns_groups_2, nutrition_grade_fr
- **Comment ?**
  - ☒ mots-clés uniques dans ‘main_category_fr’
  - ☑ catégories uniques et/ou les plus fréquentes
  - ☑ Relation avec les valeurs nutritionnelles
  - ☐ KNN Imputer en encodant les variables catégorielles
  - ☐ Recherche de produits identiques ‘product_name’ et ‘brands’ avec un code barre différent (exemple : ‘quantity’ différente)

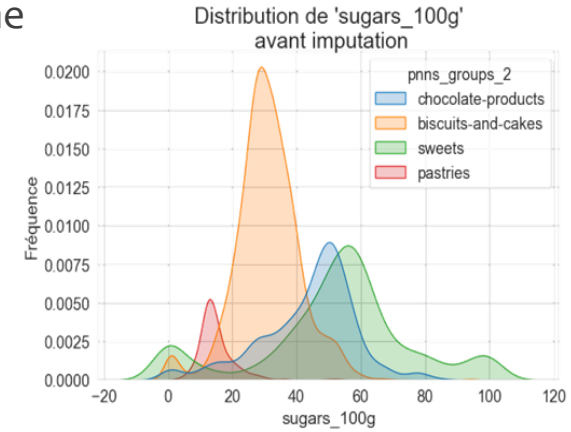
Traitements des valeurs manquantes

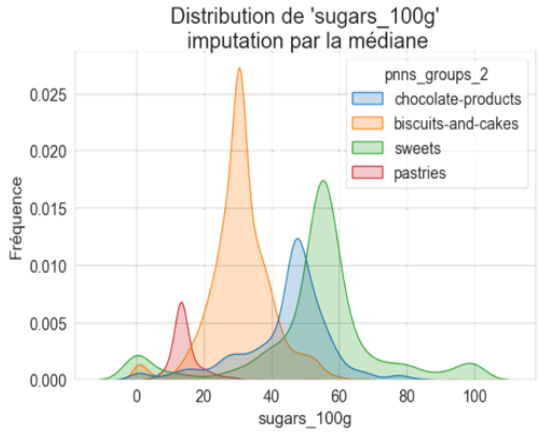
Variables numériques

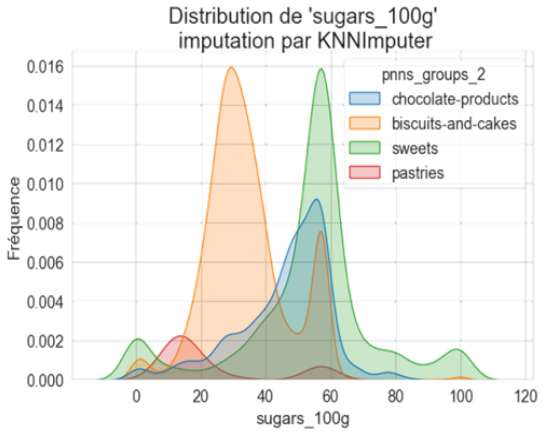
- **Quoi?**
  - energy_100g, saturated-fat_100g, sugars_100g, salt_100g, proteins_100g, fibers_100g, fruits-vegetables-nuts_100g, nutrition-score-fr_100g
- **Comment?**
  - ☑ Mise à 0 des valeurs manquantes
  - ☑ Imputation par la moyenne / médiane
  - ☑ KNN Imputer
  - ☑ Régression linéaire
  - ☐ Iterative Imputer
  - ☐ Relation entre les variables nutritionnelles
  - ☐ Formule de calcul pour le nutriscore

Très bon résultat avec le KNN Imputer :

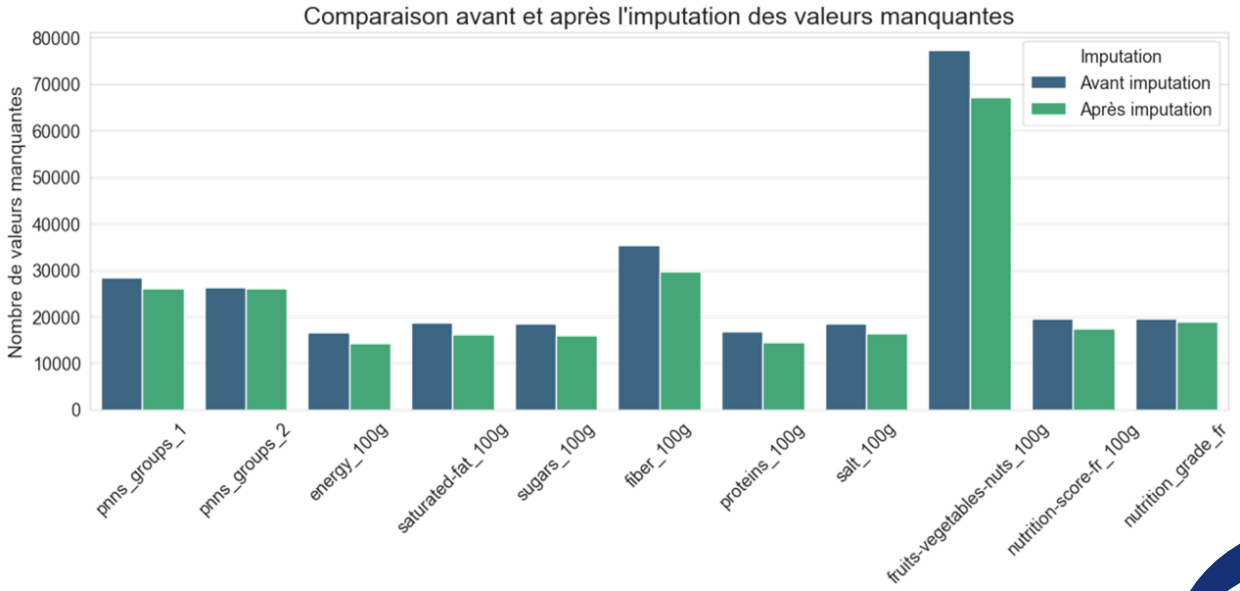
R2 score entre 0,93 et 1

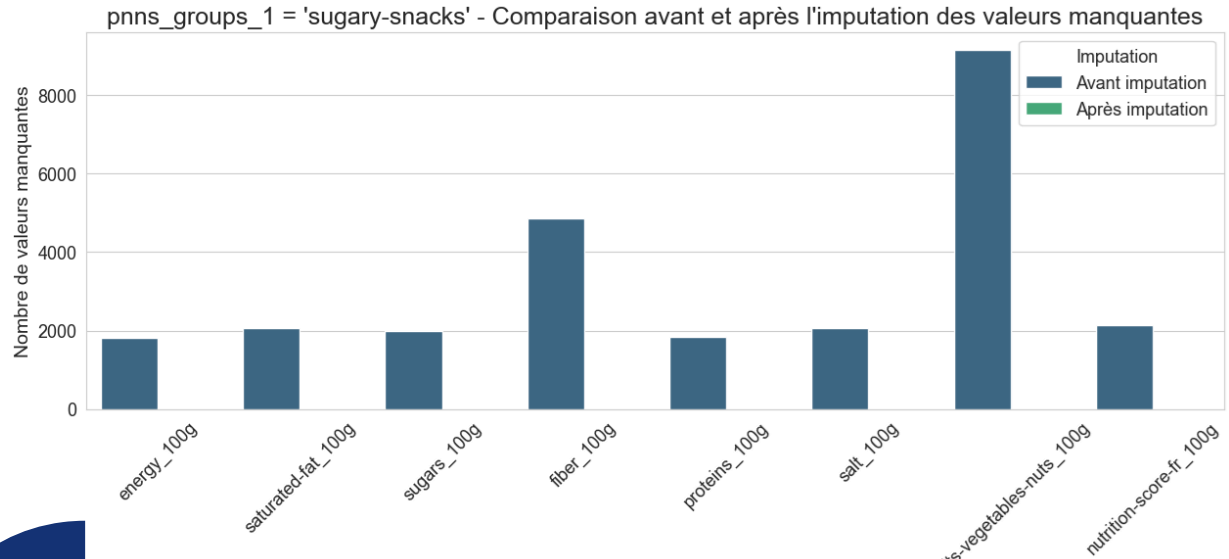




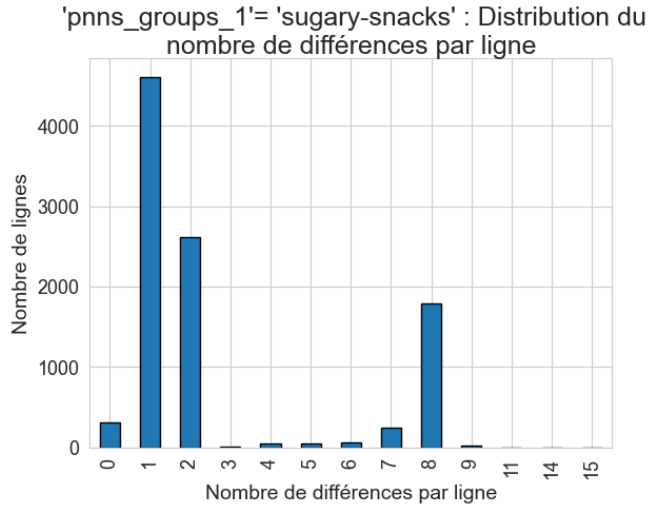


Résultats du traitement des valeurs manquantes



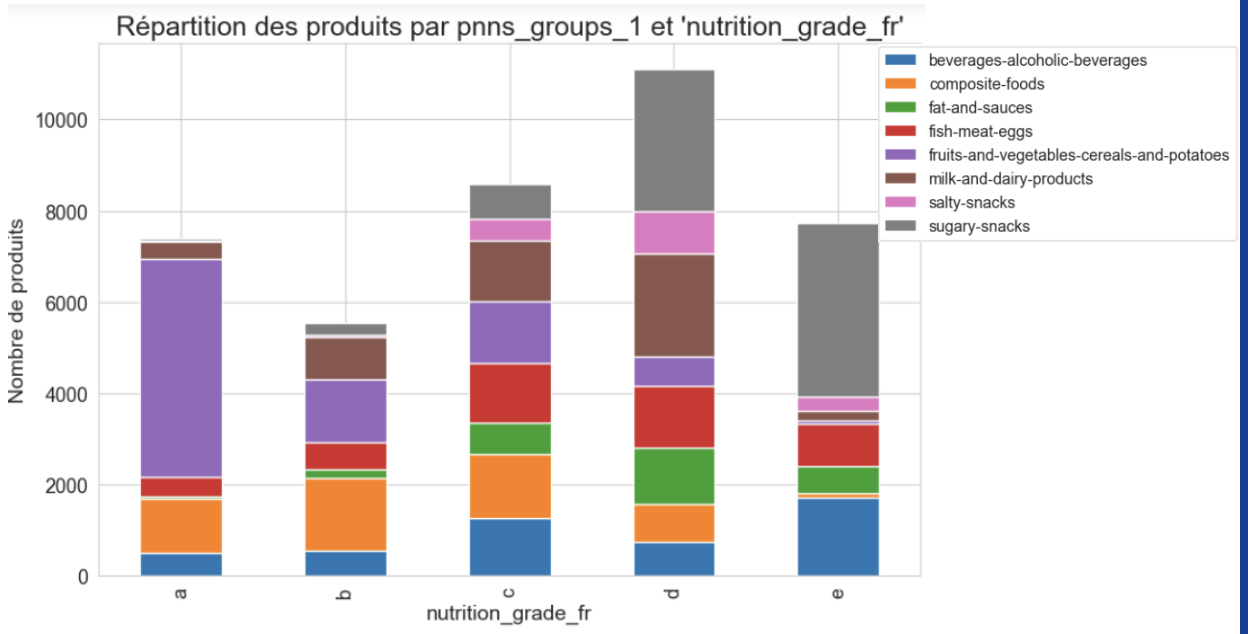


L'auto-complétion entraîne beaucoup de modifications : Fiabilité des valeurs en fonction de l'application



Exploration des données

Le groupe d'aliments influe sur le nutrition grade.



L'Épinard Feuilles Préservées - Bonduelle - 750 g

Les papillotes - Révillon - 360 g

Test statistique :	
Type de variables :	2 variables qualitatives
Test applicable :	Chi-2 d'indépendance
Hypothèse nulle :	Les variables sont indépendantes.
Conditions d'applications du test :	• Variables collectées indépendamment • Aucune valeur attendue = 0 • Valeurs observées et attendues > 5
Résultats :	• Chi^2 = 23915 , degré de liberté = 28 • p-value = 0 < 0,05 -> H0 rejetée • V de Cramer = 0,39 -> lien fort

Interprétation du test statistique :
Il existe une relation forte entre le pnns_groups_1 et le nutrition grade.

Relation entre

`pnns_groups_1` et `nutrition_grade_fr`

Test statistique :

Type de variables :	2 variables quantitatives
Test normalité :	Rejeté, les distribution ne suivent pas une loi normale.
Test applicable :	Coefficient de rang de Spearman (test non paramétrique)
Hypothèse nulle :	Il n'existe pas de relation linéaire entre les variables
Résultats :	• 0,76 -> relation positive et forte • p-value = 0 < 0,05 -> H0 rejetée

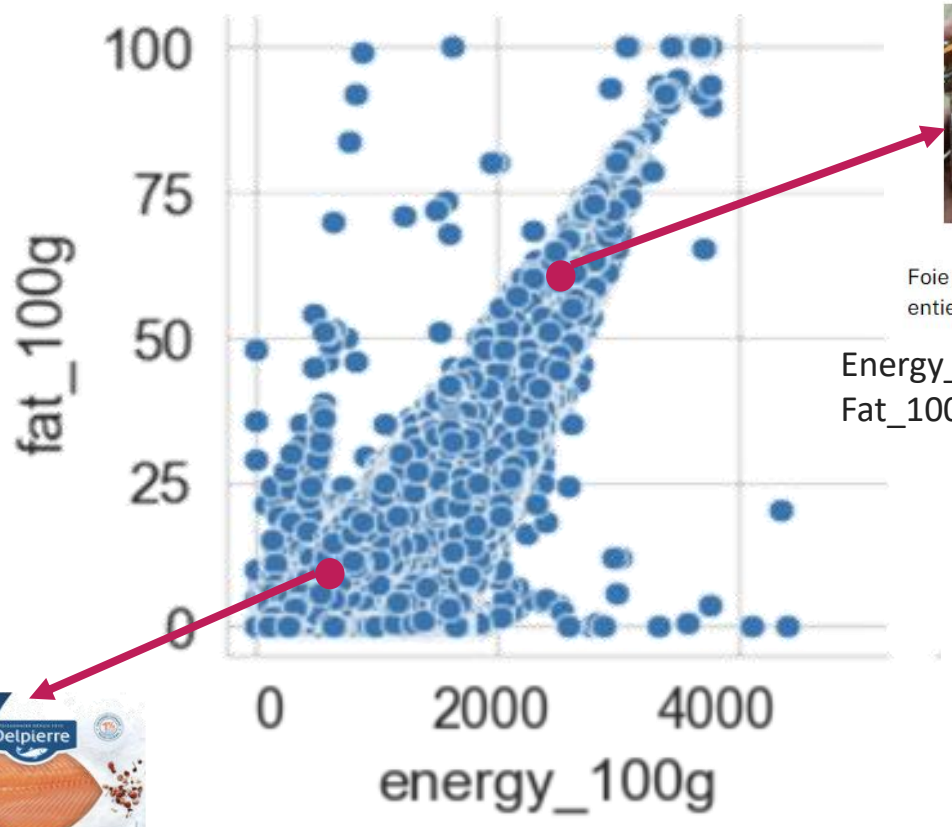
Interprétation du test statistique :

Il existe une corrélation linaire positive forte entre energy_100g et fat_100g.

Relation entre

`energy_100g` et `fat_100g`

Plus un produit est gras, plus il est calorique.



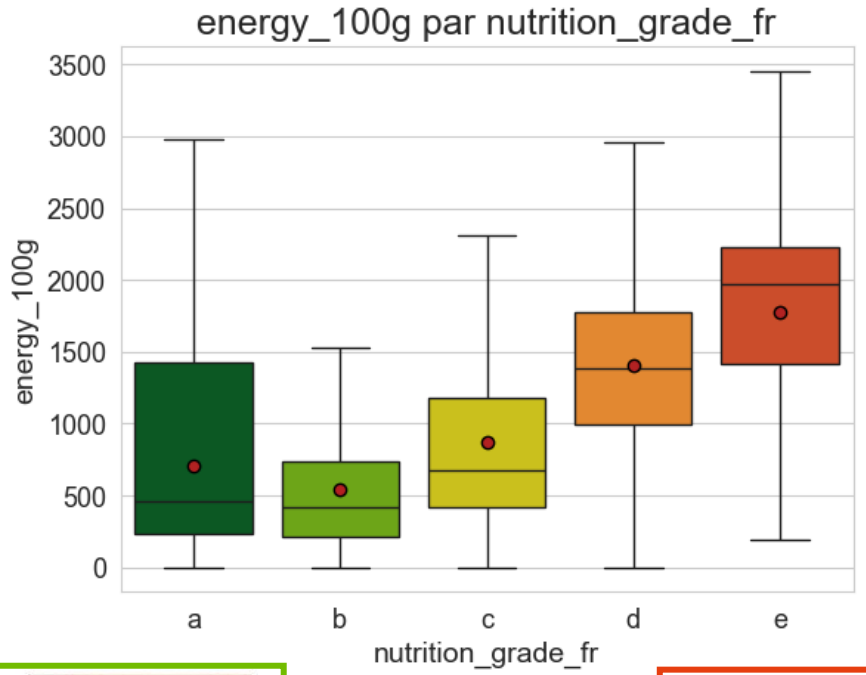
Foie Gras de canard entier - Delpeyrat

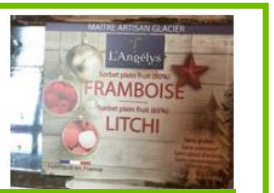
Energy_100g = 2276 kJ
Fat_100g = 56 g

Saumon fumé écosse 4 tranches Delpierre -

Energy_100g = 736 kJ
Fat_100g = 10 g

L’énergie d’un aliment influe sur le nutrition grade.





538 kJ / 100g

Bûche glacée sorbet framboise/litchi - L'Angélys - 750 ml (6/8 parts)

1620 kJ / 100g

Bûche praliné - Colibri - 400 g (8 parts)

Test statistique :	
Type de variables :	1 variable quantitative et 1 qualitative
Tests préalables :	Normalité (Kolmogorov-Smirnov) : rejetée Egalité des variances (Levene) : rejetée
Test applicable :	Test de Kruskal-Wallis : Type ANOVA non paramétrique
Hypothèse nulle :	Les médianes des différents groupes sont égales.
Résultats :	p-value = 0 < 0,05 -> H0 rejetée

Interprétation du test statistique :
Les groupes nutrigrades ont tous des médianes différentes pour l’énergie.

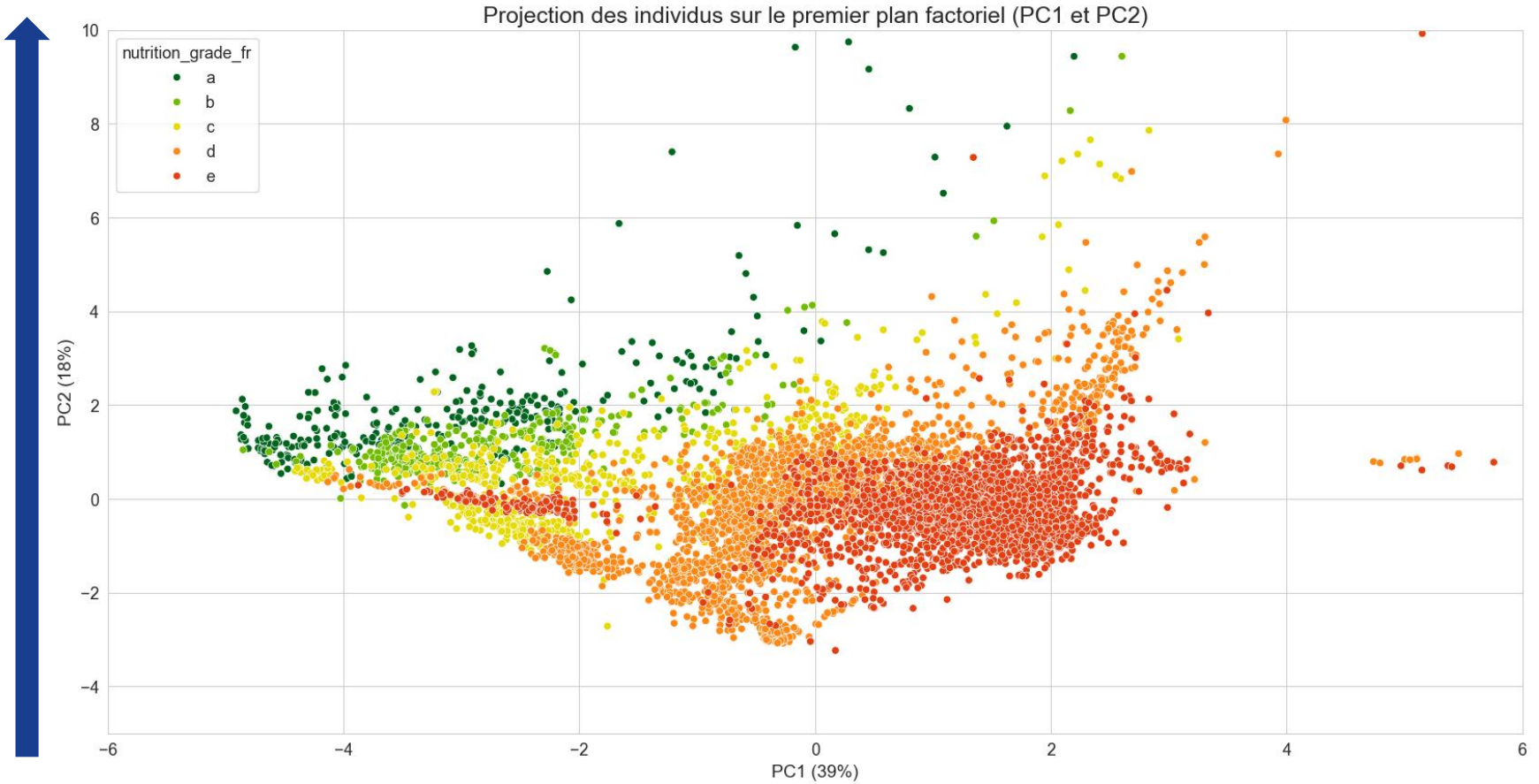
Relation entre

`energy_100g` et `nutrition_grade_fr`

Analyse en Composantes Principales

Sugars ↘

Fibres, protéines ↗



Nutrition_score, energy, saturated fat ↗

Fruits-and-vegetables ↘

Conclusion & perspectives

Conclusion

Faisabilité de la suggestion ou l'auto-complétion

product_name	pnns_groups_1	pnns_groups_2	fat_100g	sugars_100g	proteins_100g	salt_100g
sucre|	sugary-snacks	sweets	0	100	0	0

Valeurs suggérées

Perspectives

Recettes à partir de liste d'ingrédients et calculs de l'énergie

Panier avec les produits les sains à partir d'une liste course

Aliments pour la prévention des maladies cardiovasculaires

Merci de votre attention