

Projet 03

Préparez des données pour un organisme de santé publique

Nettoyage et exploration des données Open Food Facts



David Scanu | Septembre 2024



Parcours **AI Engineer**

Contexte et mission

Contexte du projet

L'agence **Santé publique France** souhaite améliorer sa base de données **Open Food Facts**.

Objectif du projet










Déterminer la faisabilité d'un système de suggestion ou d'auto-complétion pour aider à remplir la base de données d'Open Food Facts.

Notre rôle

Nettoyer et explorer la base de données d'Open Food Facts afin d'évaluer la faisabilité de mettre en place un système de suggestion.



Démarche méthodologique

-  Installation environnement et importation
-  Visualisation générale du jeu de données
-  Sélection de la cible
-  Nettoyages des données
-  Analyses univariées et bivariées
-  Analyses multi-variées
-  Rapport d'exploration
-  Conclusion sur la faisabilité de l'application
-  Notre projet est-il concerné par le RGPD ?

Installation environnement et importation

- **Environnement** : [Notebook Colab](#) 
- Installation et **importation des bibliothèques Python**
 - Pandas, Numpy, Scipy, Matplotlib, Seaborn, Plotly
- **Importation du jeu de données**
 - Jeu de données [Open Food Facts](#)
 - Informations générales
 - Ensemble de tags
 - Ingrédients et additifs
 - Informations nutritionnelles
- **Correction d'irrégularités** dans le fichier CSV
- **Chargement du fichier .csv dans un DataFrame**

```
david@David-PC:~$ python3 -m venv .venv
david@David-PC:~$ source .venv/bin/activate
david@David-PC:~$ pip install pandas numpy matplotlib seaborn
plotly==5.24.0 --no-cache-dir --quiet
```

```
import sklearn
import matplotlib.pyplot as plt
import matplotlib.colors
import numpy as np
import pandas as pd
import seaborn as sns
import scipy.stats as st
import plotly.express as px
```

```
data_filepath =
"/content/data/fr.openfoodfacts.org.products-clean.csv"
data = pd.read_csv(data_filepath , delimiter=";")
```



Visualisation générale du jeu de données

- **Nombre de variables** : 162
- **Nombre de lignes** : 32 0749
- **Valeurs manquantes** : 39 604 863
- **Valeurs manquantes en %** : 76,22%



```
data.head()
```

index	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	quantity	packaging	packaging_tags	brands	brands_tags	categories	categories_tags	categories_fr	origins	origins_tags	manufacturing_places
0	000000003087	http://world-fs.openfoodfacts.org/product/000000003087/farine-de-ble-noir-ferme-fy-r-nao	openfoodfacts-contributors	1474103866	2016-06-17 09:17:46+00:00	1474103893	2016-06-17 09:18:13+00:00	Farine de blé noir	NaN	1kg	NaN	<NA>	Ferme fy r-nao	ferme-fy-r-nao	NaN	<NA>	NaN	NaN	<NA>	NaN
1	0000000004530	http://world-fs.openfoodfacts.org/product/0000000004530/banana-chips-sweetened-whole	usda-ndb-import	1489069957	2017-03-09 14:32:37+00:00	1489069957	2017-03-09 14:32:37+00:00	Banana Chips Sweetened (Whole)	NaN	NaN	NaN	<NA>	NaN	<NA>	NaN	<NA>	NaN	NaN	<NA>	NaN
2	0000000004559	http://world-fs.openfoodfacts.org/product/0000000004559/peanuts-tom-glasser	usda-ndb-import	1489069957	2017-03-09 14:32:37+00:00	1489069957	2017-03-09 14:32:37+00:00	Peanuts	NaN	NaN	NaN	<NA>	Tom & Glasser	tom-glasser	NaN	<NA>	NaN	NaN	<NA>	NaN
3	0000000016087	http://world-fs.openfoodfacts.org/product/0000000016087/organic-salted-nut-mix-grizzlies	usda-ndb-import	1489055731	2017-03-09 10:35:31+00:00	1489055731	2017-03-09 10:35:31+00:00	Organic Salted Nut Mix	NaN	NaN	NaN	<NA>	Grizzlies	grizzlies	NaN	<NA>	NaN	NaN	<NA>	NaN
4	0000000016094	http://world-fs.openfoodfacts.org/product/0000000016094/organic-polenta-bob-s-red-mill	usda-ndb-import	1489055653	2017-03-09 10:34:13+00:00	1489055653	2017-03-09 10:34:13+00:00	Organic Polenta	NaN	NaN	NaN	<NA>	Bob's Red Mill	bob-s-red-mill	NaN	<NA>	NaN	NaN	<NA>	NaN
5	0000000016100	http://world-fs.openfoodfacts.org/product/0000000016100/breadshop-honey-gone-nuts-granola-unifi	usda-ndb-import	1489055651	2017-03-09 10:34:11+00:00	1489055651	2017-03-09 10:34:11+00:00	Breadshop Honey Gone Nuts Granola	NaN	NaN	NaN	<NA>	Unifi	unifi	NaN	<NA>	NaN	NaN	<NA>	NaN
6	0000000016117	http://world-fs.openfoodfacts.org/product/0000000016117/organic-long-grain-white-rice-lundberg	usda-ndb-import	1489055730	2017-03-09 10:35:30+00:00	1489055730	2017-03-09 10:35:30+00:00	Organic Long Grain White Rice	NaN	NaN	NaN	<NA>	Lundberg	lundberg	NaN	<NA>	NaN	NaN	<NA>	NaN
7	0000000016124	http://world-fs.openfoodfacts.org/product/0000000016124/organic-muesli-daddy-s-muesli	usda-ndb-import	1489055711	2017-03-09 10:35:11+00:00	1489055712	2017-03-09 10:35:12+00:00	Organic Muesli	NaN	NaN	NaN	<NA>	Daddy's Muesli	daddy-s-muesli	NaN	<NA>	NaN	NaN	<NA>	NaN

Sélection de la cible

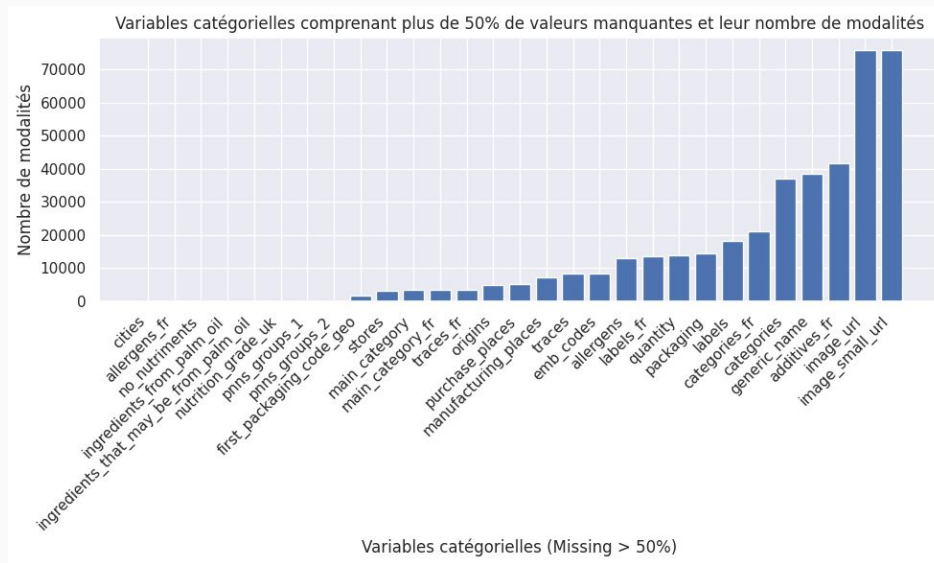
Extrait du mail de briefing :

“Etablir la faisabilité de suggérer les valeurs manquantes pour une variable dont plus de 50% des valeurs sont manquantes.”

Critères de sélection

- **Moins de 50% de valeurs présente**
- Variable **catégorielle**
- Certaines variables dont le nom se termine par **_tags**, contiennent des valeurs de type list. Nous ne les utiliserons pas comme cible
- **Nombre optimal de modalités**

Variable choisie : pnns_groups_1 (14 modalités)



Cible sélectionnée : `pnns_groups_1`

Catégorie d'aliments (en langue anglaise)

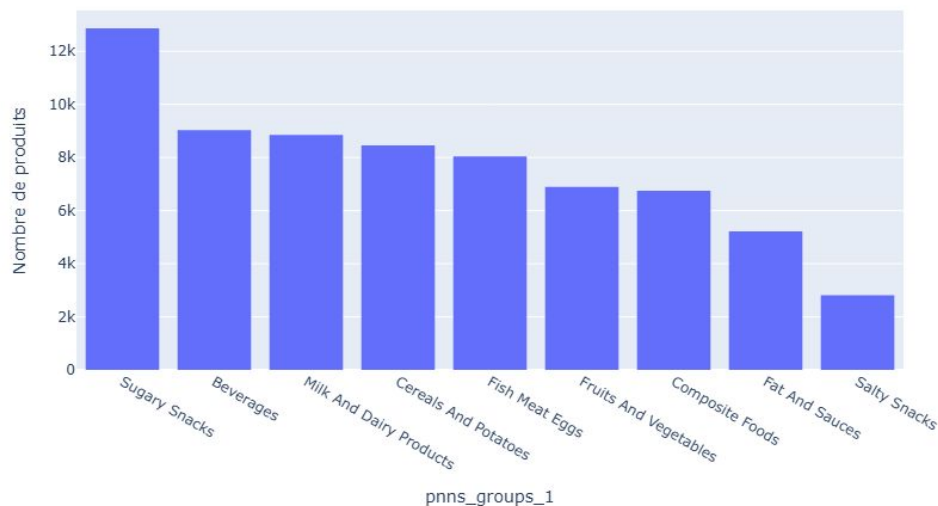
Nombre de modalités : 9

- Beverages
- Cereals And Potatoes
- Composite Foods
- Fat And Sauces
- Fish Meat Eggs
- Fruits And Vegetables
- Milk And Dairy Products
- Salty Snacks
- Sugary Snacks

Nb lignes après supp. : **68 910**

Un traitement a été réalisé pour supprimer les modalités redondantes.

Répartition des modalités de pnns_groups_1



Sélection des variables pertinentes

Objectif

Prédire **pnns_groups_1** (*catégorie d'aliment en langue anglaise*)

Critères de sélection

- **Suffisamment de valeurs remplies** pour pouvoir faire une analyse statistique fiable (> 50% de valeurs non-nulles).
- **Pertinentes** pour prédire la cible

Élimination manuelle de variables

Avec l'aide de la [description des données](#), nous décidons d'**éliminer manuellement les variables suivantes** :

- Identifiant (**code**)
- L'url du produit (**url**)
- Auteur (**creator**)
- Les dates (**created_t**, **created_datetime**, **last_modified_t**, **last_modified_datetime**)
- Les quantités car problème de formatage (**quantity**)
- Etat dans la base de données (**states**, **states_tags**, **states_fr**)
- Variables catégorielles ayant trop de modalités (**product_name**, **generic_name**, **brands**, **purchase_places**, **stores**, **countries**, **ingredients_text**)
- La catégorie d'aliment en langue étrangère (**main_category**).
- Les URL de l'image (**image_url**, **image_small_url**).
- Les variables qui sont des tags (souvent se terminant par **_tags**) sont des listes de valeurs (string). Leur utilisation étant plus difficile, nous ne les utiliserons pas : **packaging**, **packaging_tags**, **brands_tags**, **categories**, **categories_tags**, **categories_fr**, **labels**, **labels_tags**, **labels_fr**, **countries_tags**, **additives**.

Elimination des variables corrélées

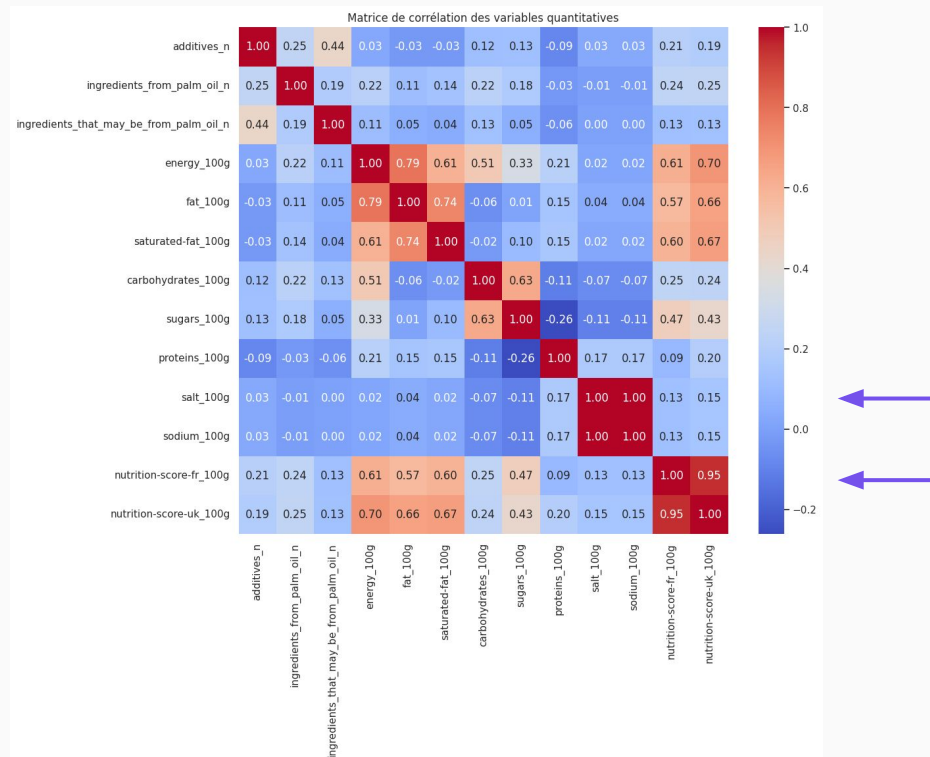
Les variables redondantes transmettent la même information, rendant difficile pour l'algorithme de différencier leur impact individuel.

Conserver des variables **fortement corrélées** dans un modèle de machine learning :

- Problème de **multicolinéarité**
- **Dégrader la performance générale du modèle**

D'après la **matrice de corrélation** suivante, nous décidons d'éliminer les variables suivantes :

- **sodium_100g** : très fortement corrélée à salt_100g.
- **nutrition-score-uk_100g** : très fortement corrélée à nutrition-score-fr_100g



Variables retenues

Nous avons sélectionné **13 variables**.

11 variables quantitatives :

- additives_n
- ingredients_from_palm_oil_n
- ingredients_that_may_be_from_palm_oil_n
- energy_100g
- fat_100g
- saturated-fat_100g
- carbohydrates_100g
- sugars_100g
- proteins_100g
- salt_100g
- nutrition-score-fr_100g

2 variables qualitatives :

- main_category_fr
- nutrition_grade_fr

Suppression des lignes en double

- Avant suppression : 68 910
- Après suppression : **52 027**

Identifiez et traitez les valeurs aberrantes

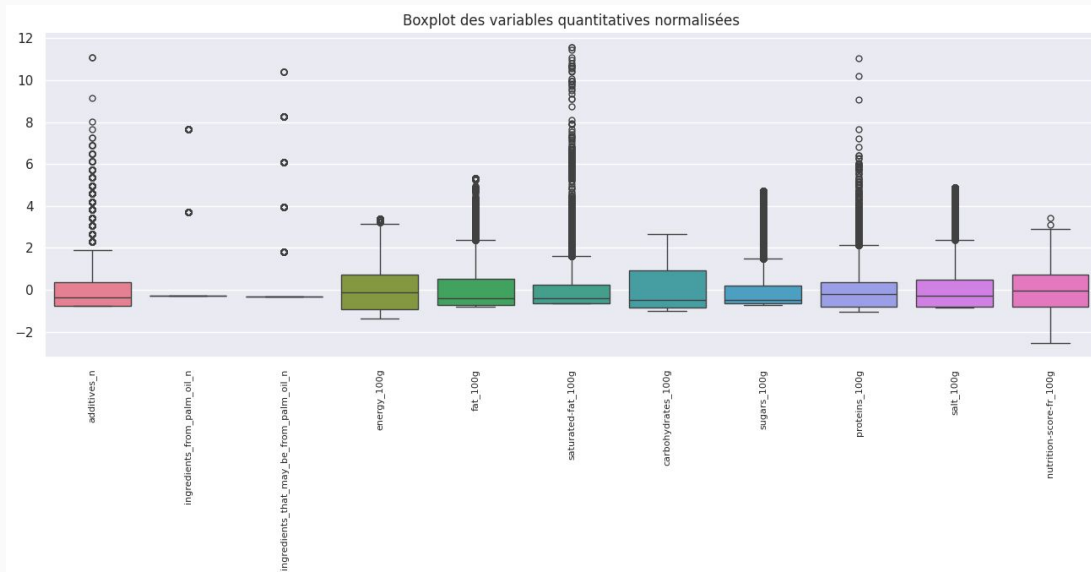
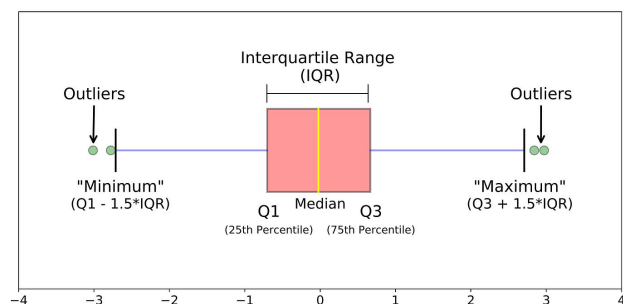
Logique métier	Nb lignes	Correction apportée
Valeurs négatives pour la variable sugars_100g, ce qui est impossible	1	Suppression
Valeurs des marco-nutriments (pour 100 g) qui dépassent 100	10	Suppression
Somme des macro-nutriments dépassent 100	155	Suppression
Valeur maximum pour energy_100g dépasse 3700 Kj	100	Suppression
Valeur de salt_100g dépassant 5 g	649	Suppression

Identification des outliers (valeurs atypiques)

Identification des outliers

- Diagrammes en boîte (boxplot)
- Intervalle interquartile (IQR)

Recherche les outliers à l'aide de la **méthode IQR**.



- additives_n : 2074
- ingredients_from_palm_oil_n : 2952
- ingredients_that_may_be_from_palm_oil_n : 5420
- energy_100g : 5
- fat_100g : 1392
- saturated-fat_100g : 4280

- carbohydrates_100g : 0
- sugars_100g : 5865
- proteins_100g : 2477
- salt_100g : 1674
- nutrition-score-fr_100g : 0

Traitement des outliers (valeurs atypiques)

Stratégies pour traiter les outliers

- **Suppression** : Retirer simplement les observations en dehors de la plage acceptable.
- **Imputation** : Remplacer les outliers par la valeur la plus proche dans l'intervalle acceptable (par exemple, $Q1 - 1.5 \times IQR$ pour les valeurs inférieures ou $Q3 + 1.5 \times IQR$ pour les valeurs supérieures).

Choix de la méthode adaptée

- Si les outliers sont dus à des erreurs de saisie, il est souvent préférable de les **supprimer**.
- Si les outliers représentent des valeurs rares mais valides, les modéliser ou utiliser des **imputation**.

Traitement pour chaque variable

Aucun traitement

- additives_n
- ingredients_from_palm_oil_n
- ingredients_that_may_be_from_palm_oil_n
- carbohydrates_100g

Suppression (car peu nombreux)

- energy_100g
- nutrition-score-fr_100g

Imputation

- fat_100g
- saturated-fat_100g
- sugars_100g
- proteins_100g
- salt_100g

Traitement des valeurs manquantes

Imputation par la médiane de la catégorie

`main_category_fr`

- carbohydrates_100g
- sugars_100g
- proteins_100g
- fat_100g
- saturated-fat_100g
- salt_100g
- nutrition-score-fr_100g

Remplacement des NaNs par 0

- additives_n (5761 NaNs)
- ingredients_from_palm_oil_n (5761 NaNs)
- ingredients_that_may_be_from_palm_oil_n (5761 NaNs)

Ces variables comptent le nombre d'ingrédients ou d'additifs. Nous considérons que si ces valeurs sont vides cela signifie une absence de ces ingrédients.

Calcul de energy_100g (1667 NaNs)

Nous pouvons calculer l'énergie à partir des macro-nutriments : glucides, lipides, protéines.



```
def calculate_energy(carbs, proteins, fats):  
    energy_kcals = (4 * carbs) + (4 * proteins) + (9 * fats)  
    energy_kj = energy_kcals * 4.184  
    return round(energy_kj, 1)
```

Suppression

Nous décidons de supprimer les lignes restantes qui contiennent des NaNs (5727 lignes).

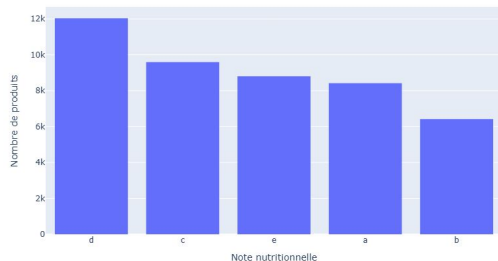


```
subset_data_clean = subset_data_no_outliers.dropna()
```

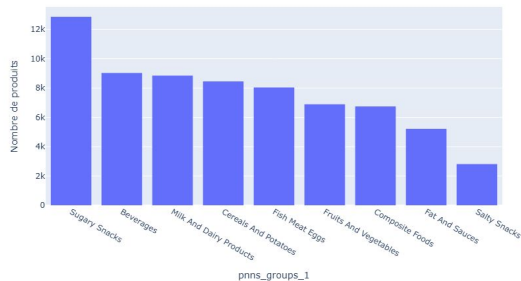
Bilan : aucune variable ne contient de valeurs manquantes

Analyses uni-variées (Variables qualitatives)

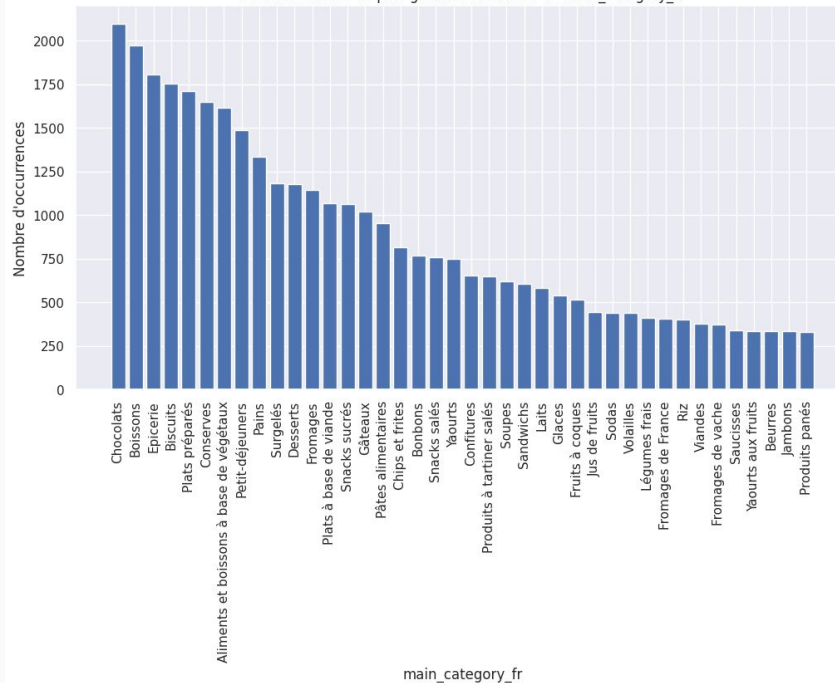
Répartition des notes nutritionnelles



Répartition des modalités de pnns_groups_1



Distribution des 40 plus grandes modalités de main_category_fr



Analyses uni-variées (Variables qualitatives)



```
subset_data_clean.describe()
```

	additives_n	ingredients_from_palm_oil_n	ingredients_that_may_be_from_palm_oil_n	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	nutrition-score-fr_100g
count	45264.0	45264.0	45264.0	45264.000000	45264.000000	45264.000000	45264.000000	45264.000000	45264.000000	45264.000000	45264.000000
mean	1.789877	0.059738	0.136223	1065.000554	12.776546	5.296996	27.733802	13.255251	7.367780	0.752765	8.317559
std	2.520983	0.240429	0.438167	766.838673	15.715417	7.981378	26.842184	18.290653	7.054752	0.865995	9.198791
min	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-15.000000
25%	0.0	0.0	0.0	385.750000	1.000000	0.200000	4.760000	1.200000	1.700000	0.070000	1.000000
50%	1.0	0.0	0.0	997.000000	6.300000	1.500000	14.500000	4.500000	5.900000	0.500000	8.000000
75%	3.0	0.0	0.0	1658.000000	21.000000	7.400000	53.000000	17.000000	10.000000	1.180000	15.000000
max	31.0	2.0	5.0	3510.000000	100.000000	93.000000	100.000000	100.000000	86.000000	5.000000	35.000000

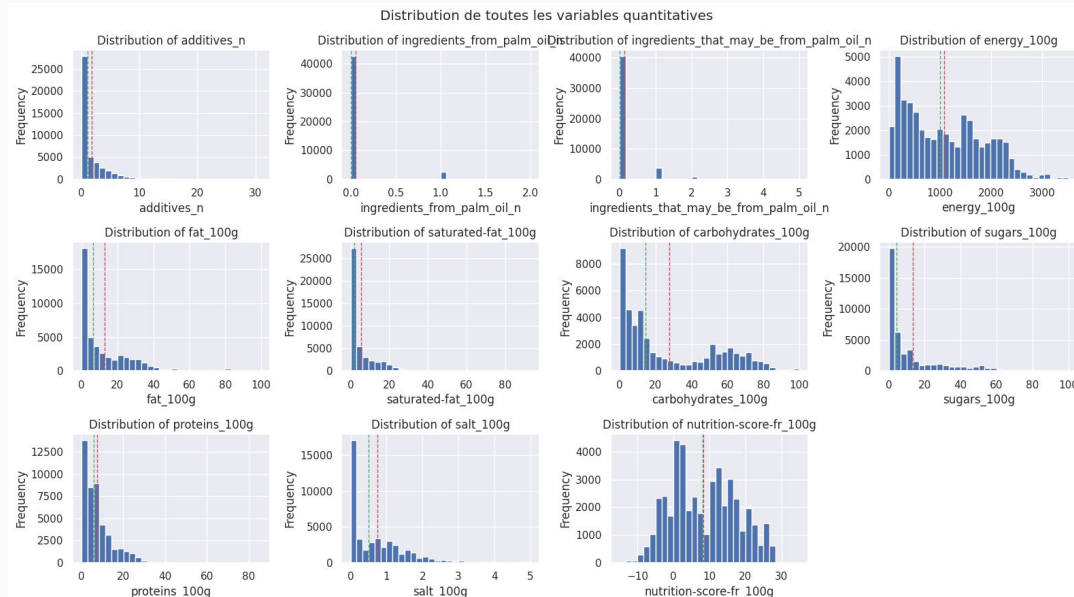
Ecart-types

Écart-types très grands (comparativement à la moyenne), ce qui indique une grande dispersion dans les données.

Écarts importants entre moyennes et médianes

- pour les lipides, les graisses saturées, les glucides, les sucres et les protéines
- distribution des données **asymétrique à droite** (asymétrie positive)

Des valeurs extrêmement élevées qui tirent la moyenne vers le haut.



Analyses bi-variées : pnns_groups_1 et nutrition_grade_fr

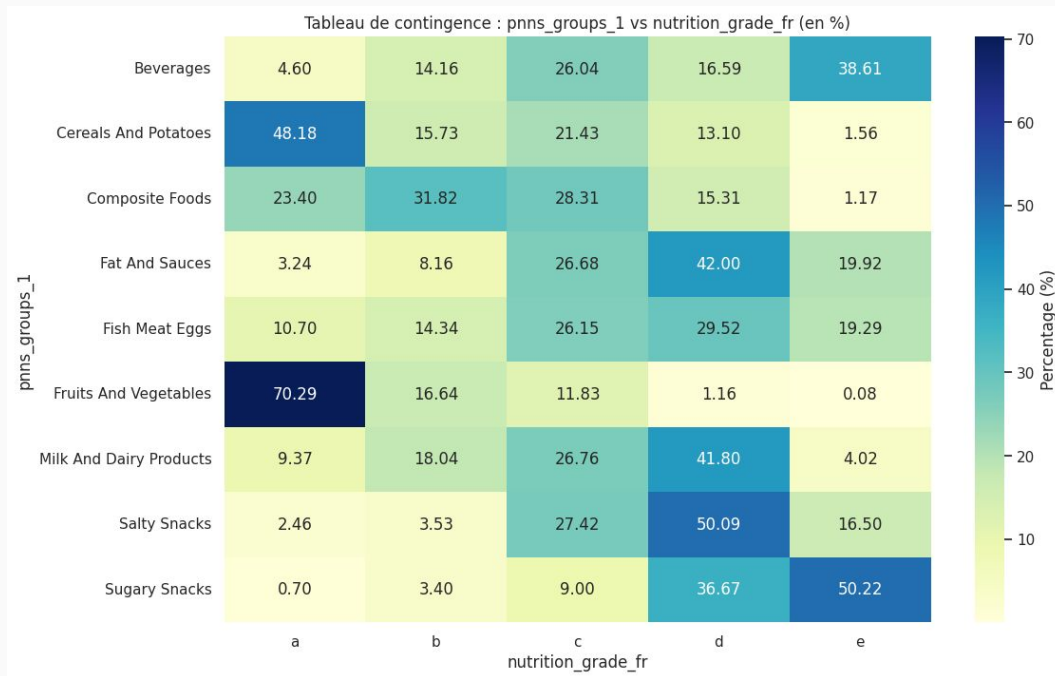
Tableau de contingence

Présente les effectifs croisés entre les catégories des deux variables.

Test du Chi-carré

Permet de tester l'indépendance entre deux variables qualitatives.

- Chi-carré : 27558.05
- P-value : 0.0
- Degrees of freedom : 32



Analyses bi-variées : pnns_groups_1 et main_category_fr

Nous décidons d'étudier la relation entre deux variables qualitatives **pnnns_groups_1** et **main_category_fr**.

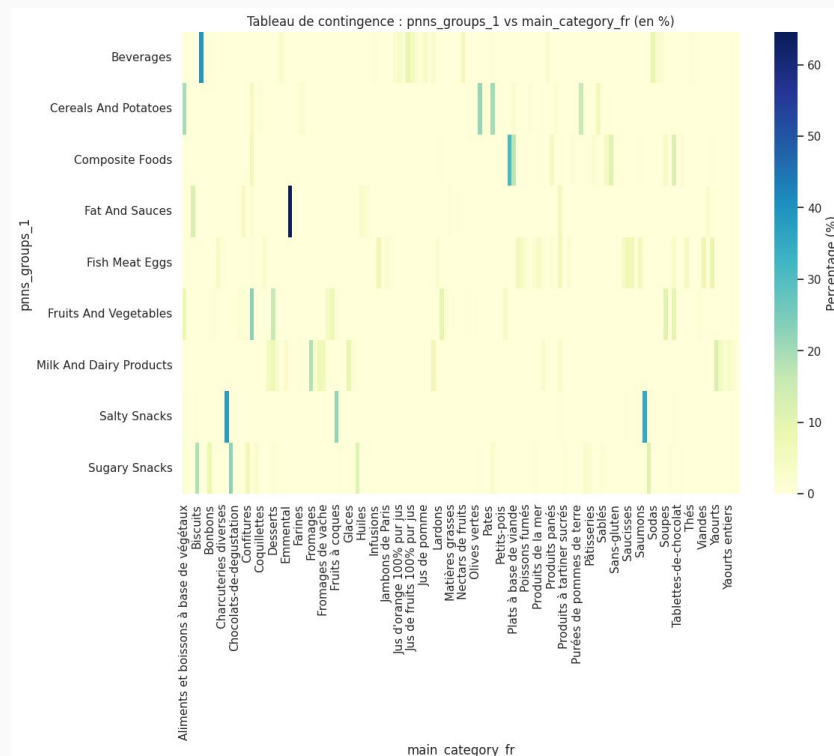
Tableau de contingence

Présente les effectifs croisés entre les catégories des deux variables.

Test du Chi-carré

Permet de tester l'indépendance entre deux variables qualitatives.

- Chi-carré : 304337.07
- P-value : 0.0
- Degrees of freedom : 1048



Analyses bi-variées (variables quantitatives)

Matrice de corrélation

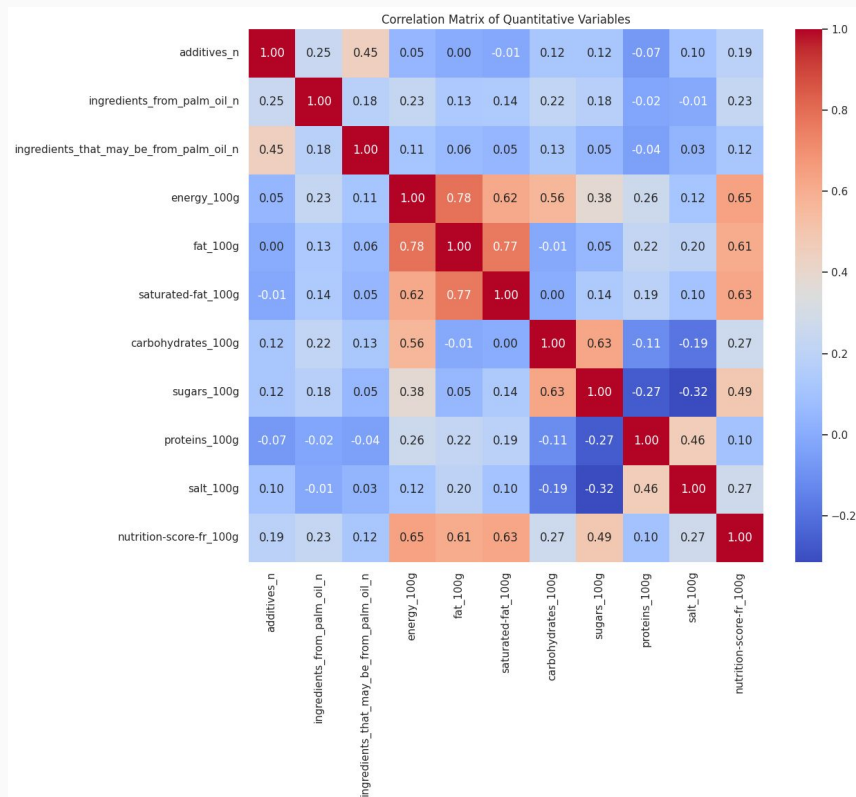
- Multicolinéarité
- Instabilité du modèle
- Diminution de la précision

Paires de variables avec des corrélations élevées

- **energy_100g** et **fat_100g** : corrélation de 0.78
- **fat_100g** et **saturated-fat_100g** : corrélation de 0.77
- **energy_100g** et **saturated-fat_100g** : corrélation de 0.62
- **sugars_100g** et **carbohydrates_100g** : corrélation de 0.63
- **nutrition-score-fr_100g** avec :
 - **energy_100g** : corrélation de 0.65
 - **fat_100g** : corrélation de 0.61
 - **saturated-fat_100g** : corrélation de 0.63

Variables à exclure du futur modèle :

- saturated-fat_100g
- sugars_100g



Relation entre pnns_groups_1 et energy_100g

Observations du graphique

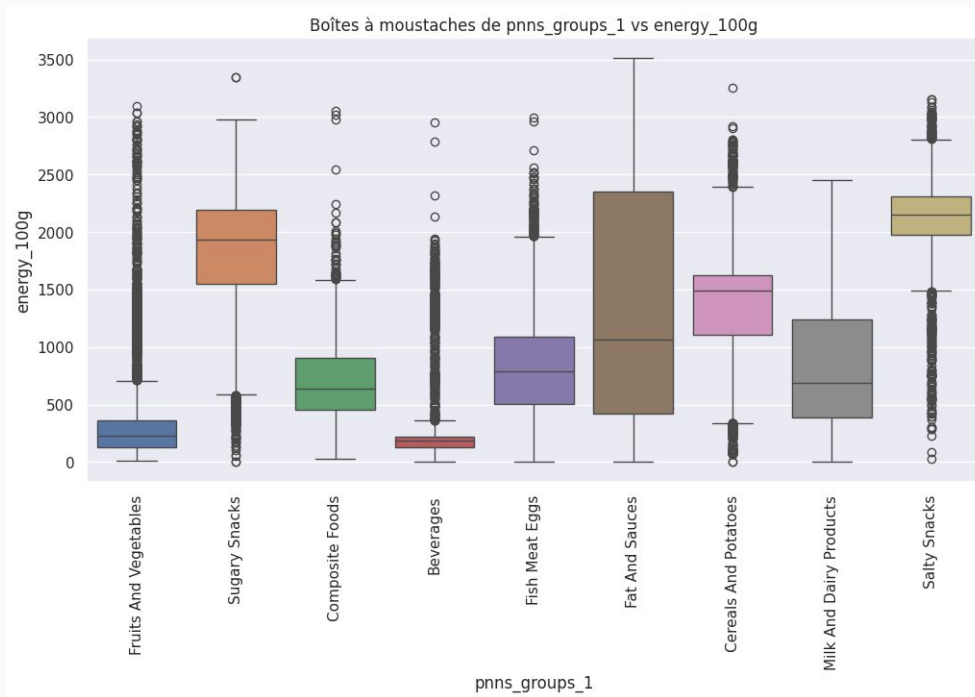
- **Variation** : Les moyennes de **energy_100g** diffèrent de manière significative entre les différents groupes d'aliments.
- **Utilité** : Peut aider à classer certains aliments.
- **Limites** : Ne suffit pas pour tous les groupes.
- **Conclusion** : L'énergie est un indice utile mais pas suffisant pour une classification précise.

Résultat du test ANOVA

- Valeur F : 6860.74
- Valeur P : 0.0 (ou un nombre extrêmement proche de zéro)

Interprétation

- **energy_100g est utile pour différencier les groupes d'aliments**
- pourrait être une variable informative dans un modèle prédictif pour classer les groupes d'aliments.



Relation entre pnns_groups_1 et nutrition-score-fr_100g

Observations du graphique

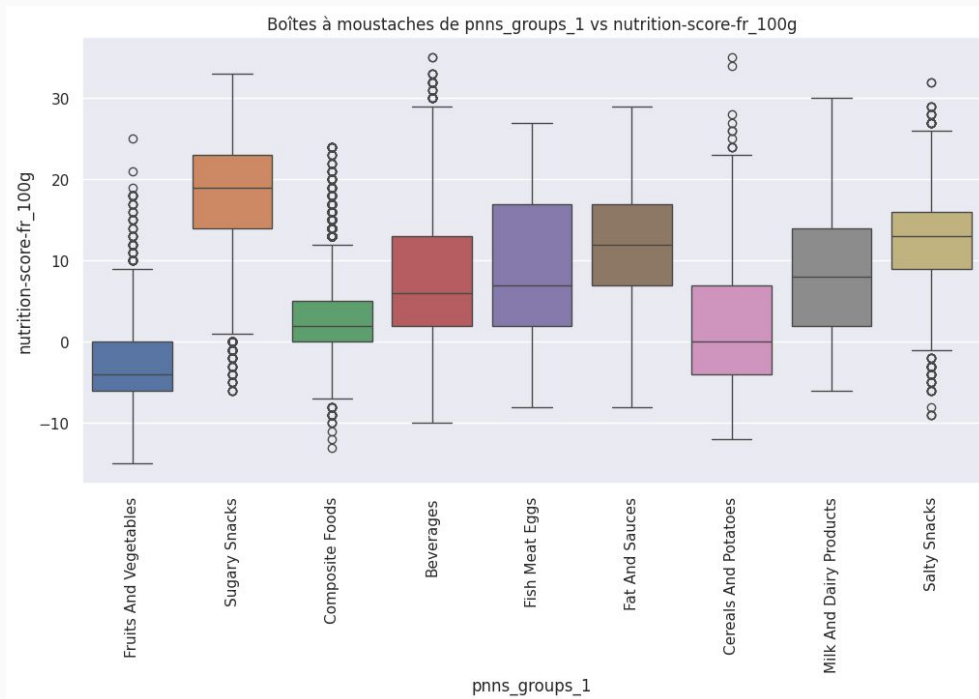
Les groupes ont des distributions de scores nutritionnels distinctes.

Résultat du test ANOVA

- Valeur F : 4963.64
- Valeur P : 0.0

Interprétation

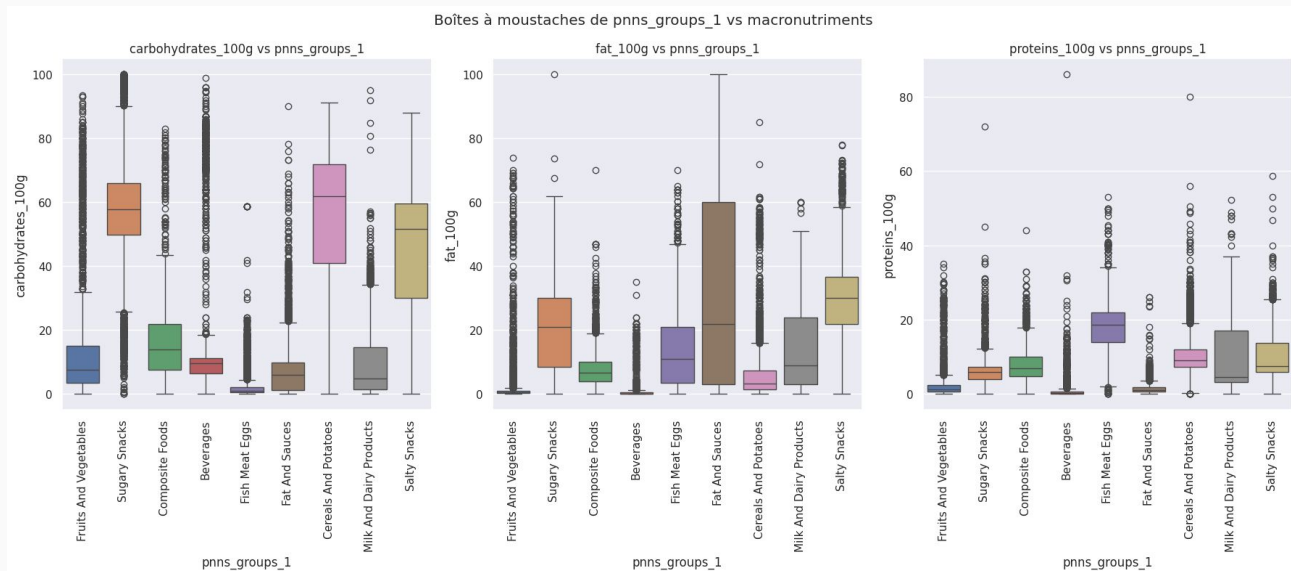
Les résultats de l'ANOVA et le graphique montrent que `nutrition-score-fr_100g` est un indicateur significatif des différences entre les groupes alimentaires.



Relation entre pnns_groups_1 et les macronutriments

Résultat du test ANOVA

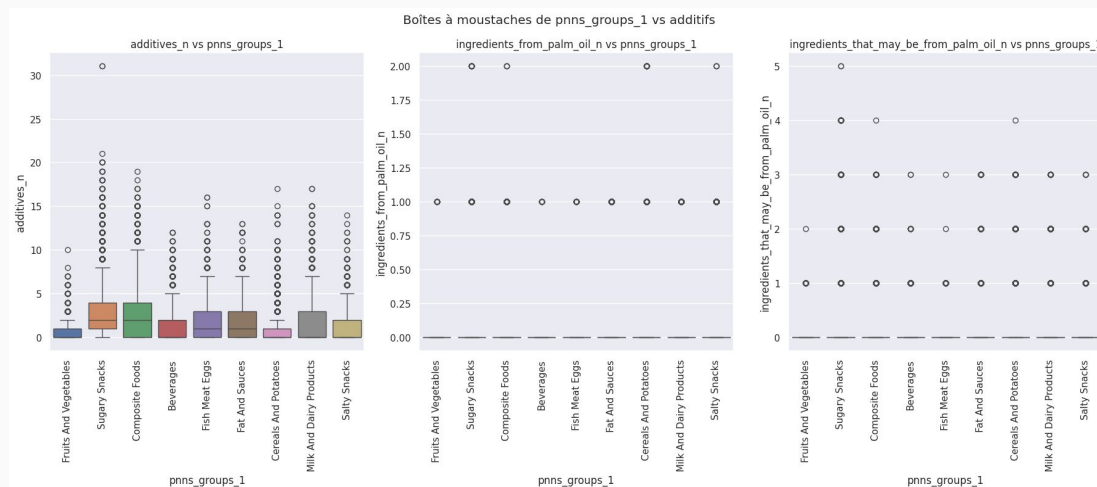
- **carbohydrates_100g**
 - fvalue 11313.52
 - pvalue 0.0
- **fat_100g**
 - fvalue 2914.92
 - pvalue 0.0
- **proteins_100g**
 - fvalue 5009.54
 - pvalue 0.0



Relation entre pnns_groups_1 et les additifs

Résultat du test ANOVA

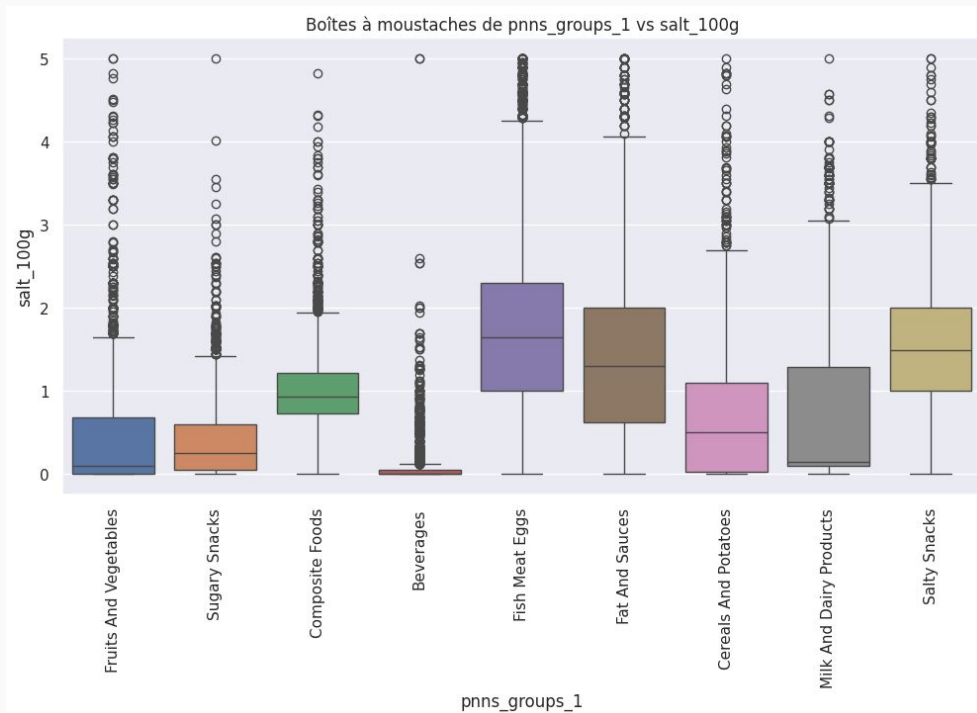
- **additives_n**
 - fvalue 532.38
 - pvalue 0.0
- **ingredients_from_palm_oil_n**
 - fvalue 427.93
 - pvalue 0.0
- **ingredients_that_may_be_from_palm_oil_n**
 - fvalue 183.01
 - pvalue 7.761367589461213e-306



Relation entre pnns_groups_1 et salt_100g

Résultat du test ANOVA

- **salt_100g**
 - fvalue': 183.00
 - pvalue': 7.761367589461213e-306



Analyse multi-variée



```
sns.pairplot(data=subset_data_clean[top_vars], hue='pnns_groups_1')
```

Diagrammes de dispersion pour chaque paire de variables numérique, avec la couleur des points pour chaque catégories d'aliments (**pnns_groups_1**).

Variables

- nutrition-score-fr_100g
- energy_100g
- carbohydrates_100g
- fat_100g
- proteins_100g

Observation

- Séparation des catégories plus difficilement observable



Analyse multi-variée

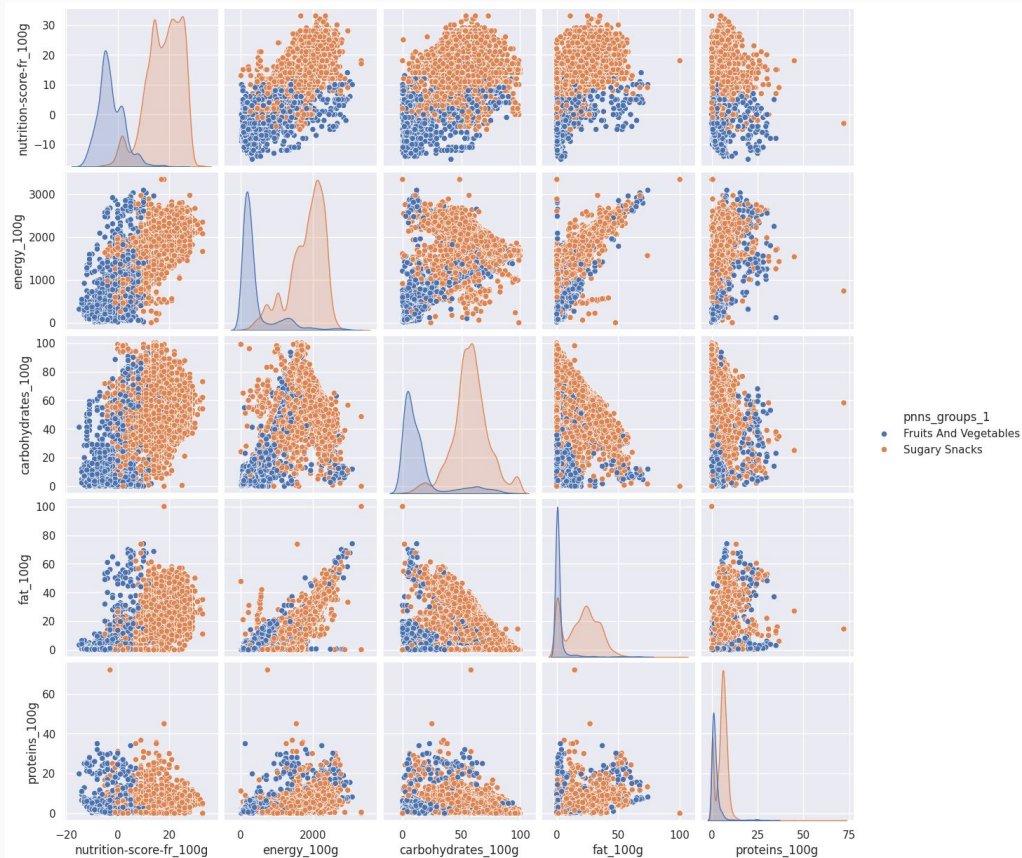


```
sns.pairplot(data=subset_mods[top_vars],  
             hue='pnns_groups_1')
```

Pour mieux visualiser la séparation des points, nous représentons les données pour deux catégories de `pnns_groups_1`: **'Fruits et légumes'** et **'Snacks sucrés'**.

Observations

- Séparations des catégories observable



Analyse en Composantes Principales (ACP)

Utilité

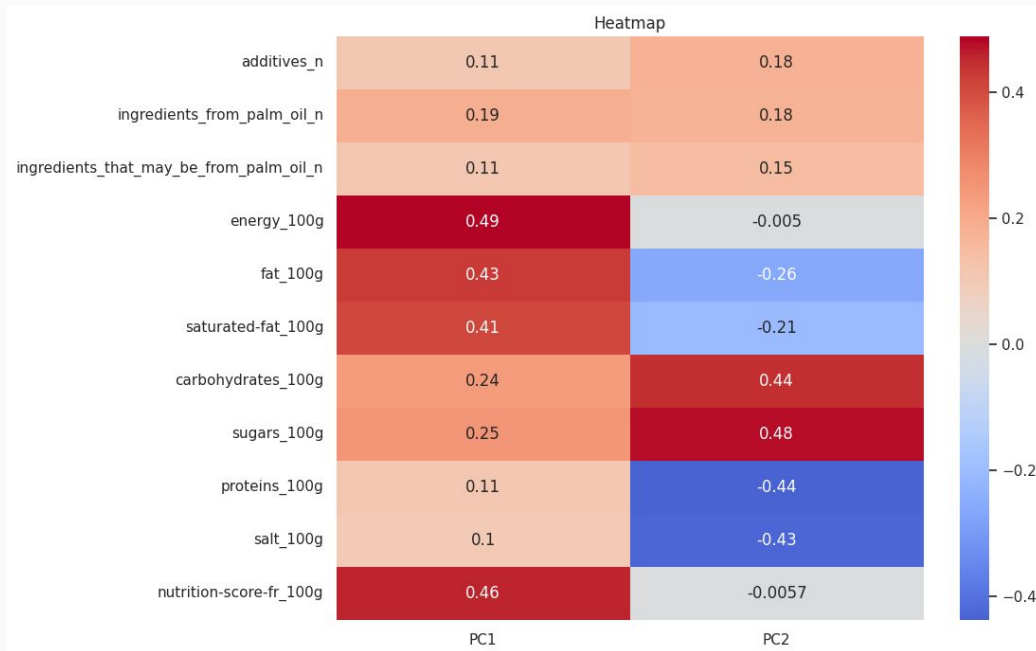
- Identifier les principales dimensions de la variabilité des données
- Réduire la dimensionnalité du problème.

Variances expliquées par les composantes principales

- PC1 : 31.68 %
- PC2 : 19.47 %
- Variance totale : 64,81%

Deux principales sources de variation dans les données :

- la densité énergétique
- la composition nutritionnelle



Analyse en Composantes Principales (ACP)

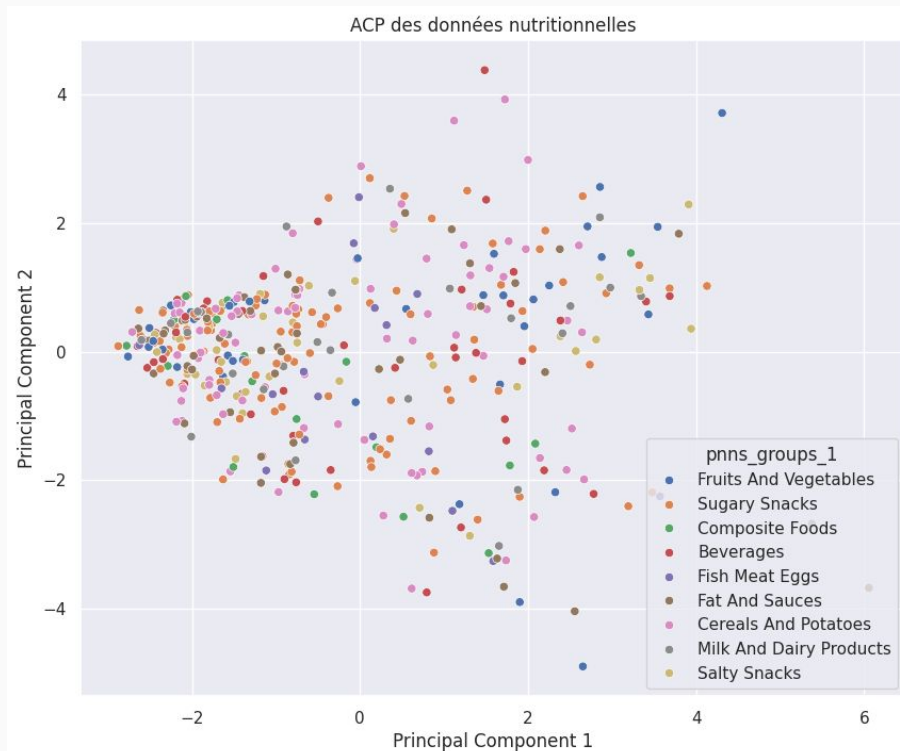
Visualisations

Projection des individus: permet de visualiser la position des individus (aliments) dans l'espace des composantes principales.

Interprétation






L'ACP n'est pas réellement concluante.

- Visuellement, la séparation des données est faible.
- Demandra d'utiliser des modèle plus complexes :
 - XGBoost
 - SVM
 - Deep Learning



Rapport d'exploration

Étapes Réalisées

-  Compréhension des Données
-  Sélection de la cible
-  Traitement des Données
-  Choix des Variables Pertinentes
-  Analyses univariées, bivariées et multivariées



Observations

- Relation variables sélectionnées vs cible (pnns_groups_1)
- Visualisations multi-Variées
- Corrélations significatives entre variables
 - Supprimer saturated-fat_100g et sugars_100g
- Limites de l'Analyse en Composantes Principales



Résultats

Cible : **pnns_groups_1**

9 variables quantitatives

- energy_100g
- fat_100g
- carbohydrates_100g
- proteins_100g
- salt_100g
- additives_n
- ingredients_from_palm_oil_n
- ingredients_that_may_be_from_palm_oil_n
- nutrition-score-fr_100g

2 variables qualitatives

- main_category_fr
- nutrition_grade_fr

Conclusion sur la faisabilité de l'application

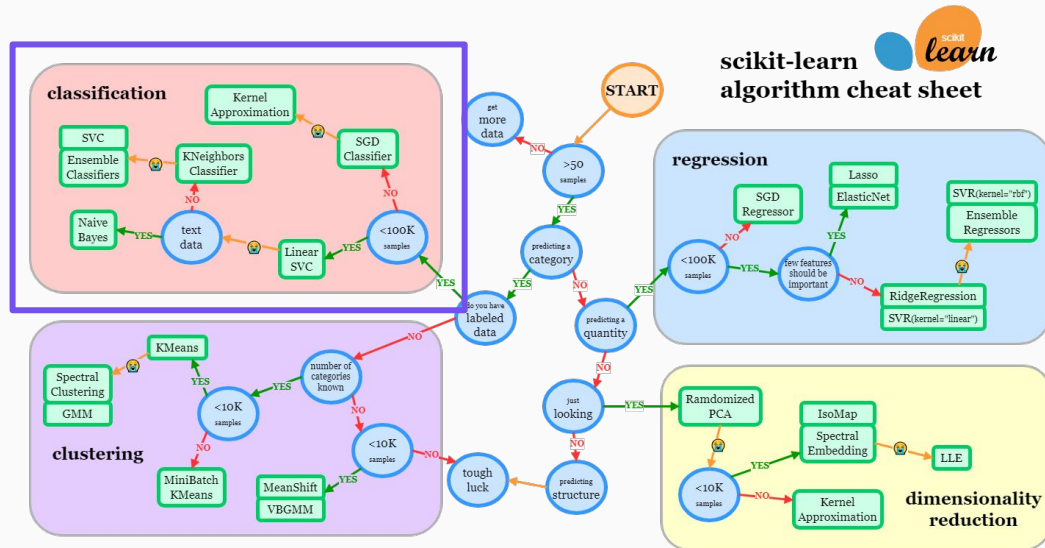
La réalisation d'une application de suggestion ou d'auto-complétion d'aliments basée sur leurs valeurs nutritionnelles apparaît **tout à fait réalisable**.

- Absence de contraintes techniques majeures
- Disponibilité de données suffisantes
- Possibilité d'expérimentation de modèles



Les prochaines étapes

1. Sélectionner les algorithmes
2. Entraîner et évaluer des modèles
3. Optimiser les hyperparamètres
4. Déployer le modèle



Respect du règlement général de protection des données (RGPD)

- Encadre le traitement des données personnelles sur le territoire de l'Union européenne.
- Concerne toute organisation, publique ou privée, qui traite des **données personnelles**.

Les 5 grands principes du RGPD

1. Licéité, loyauté et transparence
2. Limitation des finalités
3. Minimisation des données
4. Exactitude
5. Conservation limitée






La base de données Open Food Facts

- Publiée sous forme de données ouvertes (**open data**) sous la license **Open Database Licence**.
- **Tout le monde peut l'utiliser à n'importe quelle fin.**

Notre projet est-il concerné par le RGPD ?

- Le RGPD **concerne exclusivement les données personnelles**.
- Les données d'Open Food Facts **ne contiennent pas de données personnelles**.
- Il est fréquent que les projets **respectent volontairement les principes du RGPD**.

Notre projet respecte-t-il les 5 grands principes du RGPD ?

1.  **Licéité, loyauté et transparence** : Les données traitées pour ce projet sont obtenues légalement et en toute transparence auprès d'Open Food Facts (qui met ces données à disposition de tous-tes).
2.  **Limitation des finalités** : Les données collectées sont utilisées uniquement pour le développement d'un modèle d'auto-complétion et rien de plus.
3.  **Minimisation des données** : Seules les données nécessaires au développement du modèle sont collectées (données concernant les aliments pour la France).
4.  **Exactitude** : L'exactitude des données est contrôlée lors du traitement des données, notamment car elles sont cruciales pour le développement d'un modèle performant.
5.  **Limitation de conservation** : Les données collectées pour le développement du modèle ne sont pas stockées dans une base de données.

Merci 🙏

Avez-vous des questions ?

Documentation

- [Open Food Facts](#)
- <https://world.openfoodfacts.org/data/data-fields.txt>
- [Open Database License - Wikipedia](#)
- [User guide and tutorial — seaborn 0.13.2 documentation](#)
- [12. Choosing the right estimator — scikit-learn 1.5.2 documentation](#)
- [Nettoyez et analysez votre jeu de données - OpenClassrooms](#)
- [Quels sont les 5 principes du RGPD ?](#)
- [Le règlement général sur la protection des données \(RGPD\), mode d'emploi](#)



Livrables

- [Notebook Colab](#)
- [Dépôt GitHub](#)