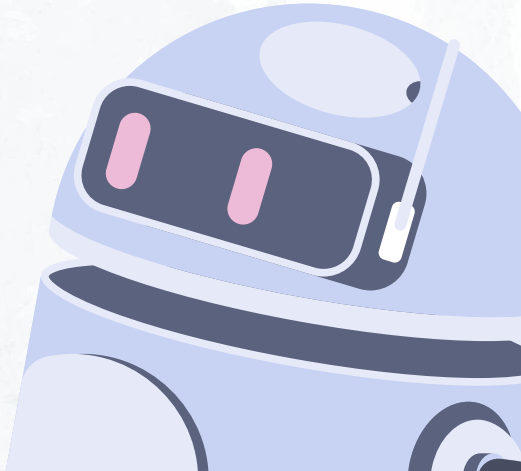
Préparez des données pour un organisme de santé publique →



Sommaire

01 →

Rappel du contexte et de l’objectifs

Rappel du contexte

L'agence Santé publique France souhaite améliorer sa base de données Open Food Facts.
Cette base de données open source est mise à la disposition de particuliers et d’organisations afin de leur permettre de connaître la qualité nutritionnelle de produits.

Rappel de l'objectif

Création d’un système de suggestion ou d’auto-complétion pour aider les usagers à remplir plus efficacement la base de données.

02 →

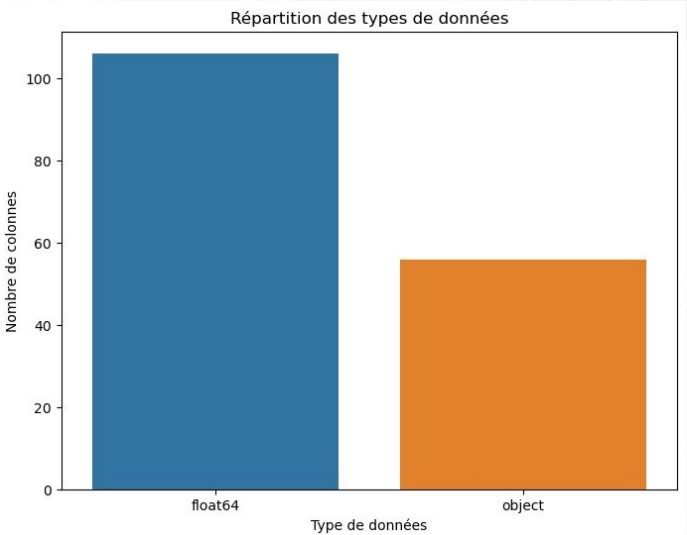
Exploration du jeu de données

(a) Exploration du jeu de données

L'analyse initiale du DataFrame products_data révèle 320 772 entrées réparties sur 162 colonnes, avec une occupation mémoire de 396.5 MB. Les types de données dominants sont float64 (106 colonnes) et object (56 colonnes).

```
print("Nombre de colonnes :", products_data.shape[1])
print("Nombre de lignes :", products_data.shape[0])
```

```
Nombre de colonnes : 162
Nombre de lignes : 320772
```

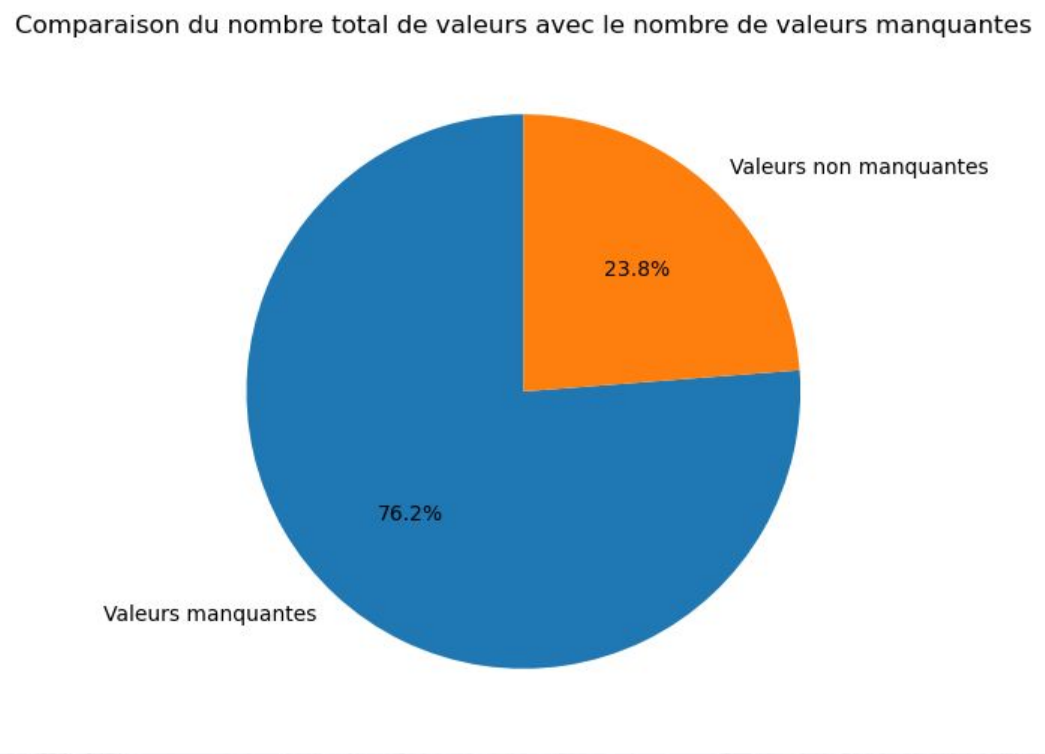


Les statistiques descriptives montrent une variabilité significative dans les colonnes numériques, tandis que l'examen des colonnes catégorielles indique une diversité de valeurs uniques, comme 320 638 valeurs distinctes pour la colonne "code".

```
products_data.describe()
```

	no_nutriments	additives_n	ingredients_from_palm_oil_n	ingredients_from_palm_oil	ingredients_that_may_be_from_pali
count	0.0	248939.000000	248939.000000	0.0	248939
mean	NaN	1.936024	0.019659	NaN	0
std	NaN	2.502019	0.140524	NaN	0
min	NaN	0.000000	0.000000	NaN	0
25%	NaN	0.000000	0.000000	NaN	0
50%	NaN	1.000000	0.000000	NaN	0
75%	NaN	3.000000	0.000000	NaN	0
max	NaN	31.000000	2.000000	NaN	6

Un défi majeur identifié est la présence de 39 608 589 valeurs manquantes sur un total de 51 965 064, nécessitant une attention particulière lors des étapes ultérieures de prétraitement.



(b) Vérification de l'équilibre de notre DataFrame

L'importance d'une distribution équilibrée des données est cruciale pour assurer la fiabilité des analyses et des modèles.

Dans le contexte de notre étude, une exploration approfondie des données nécessaires à la publication d'un produit en ligne a été réalisée.

Nous avons donc extrait les variables des produits déjà présent sur l’application.

La distance de Levenshtein, une métrique de similarité entre deux chaînes, a été employée pour identifier les correspondances entre les colonnes des DataFrames products_data et extrate_data.

Grâce à la bibliothèque fuzzywuzzy, une liste de correspondances a été établie en fixant un seuil de similarité à 70%.

Une analyse de la fréquence des variables a été effectuée, mettant en évidence les occurrences des colonnes correspondantes.

Correspondances trouvées :

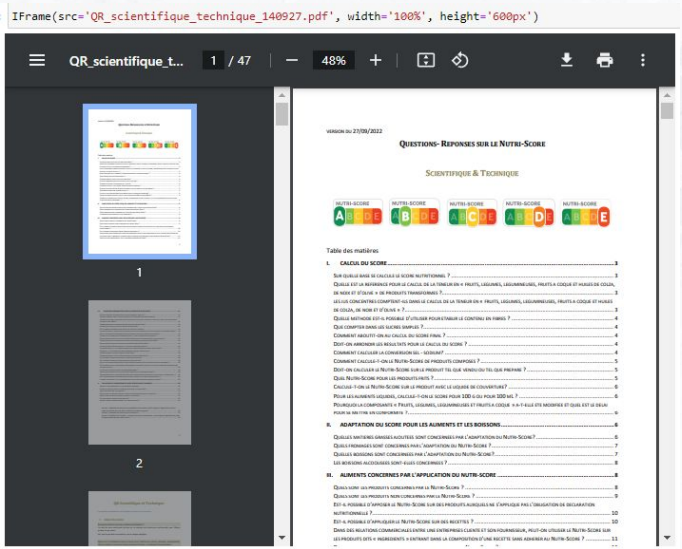
	Extrate Column	Products Column	Extrate Value count	Products Value count	Total Value count
0	Energy	energy_100g	1	1	2
1	Saturated fat	saturated-fat_100g	1	1	2
2	Carbohydrates	carbohydrates_100g	1	1	2
3	Sugars	sugars_100g	1	1	2
4	Proteins	proteins_100g	1	1	2
5	Silica	silica_100g	1	1	2
6	Bicarbonate	bicarbonate_100g	1	1	2
7	Potassium	potassium_100g	1	1	2
8	Chloride	chloride_100g	1	1	2
9	Calcium	calcium_100g	1	1	2
10	Magnesium	magnesium_100g	1	1	2
11	Barcode	code	1	1	2
12	Quantity	quantity	1	1	2
13	Packaging	packaging	2	1	3
14	Packaging	packaging_tags	2	1	3
15	Brands	brands	2	1	3
16	Brands	brands_tags	2	1	3
17	Categories	categories	3	1	4
18	Categories	categories_tags	3	1	4
19	Categories	categories_fr	3	1	4
20	Manufacturing or processing places	manufacturing_places	1	1	2
21	Stores	stores	1	1	2

Une étude détaillée du Nutri-Score a été menée, identifiant les critères et formules utilisés pour sa détermination.

Le score est basé sur des éléments négatifs (énergie, sucres simples, acides gras saturés, sel) et positifs (fruits, légumes, fibres, protéines).

Le score final est dérivé de la soustraction des points positifs des points négatifs.

Nous avons trouvé ce document PDF sur le site de santé publique France intitulé : "questions-réponses sur le nutri-score - scientifique & technique".



Ce document nous as fournis de précieuse informations sur le calcul du Nutri-Score.

Suite à l'extraction de ces informations sur le Nutri-Score, il est apparu pertinent de créer un troisième data frame avec les variables servant au calcul du Nutri-Score afin de refaire une recherche de corrélations et une analyse de fréquences.

Un troisième DataFrame, nutriscore_df, a été créé pour rechercher des corrélations avec les variables utilisées dans le calcul du Nutri-Score.

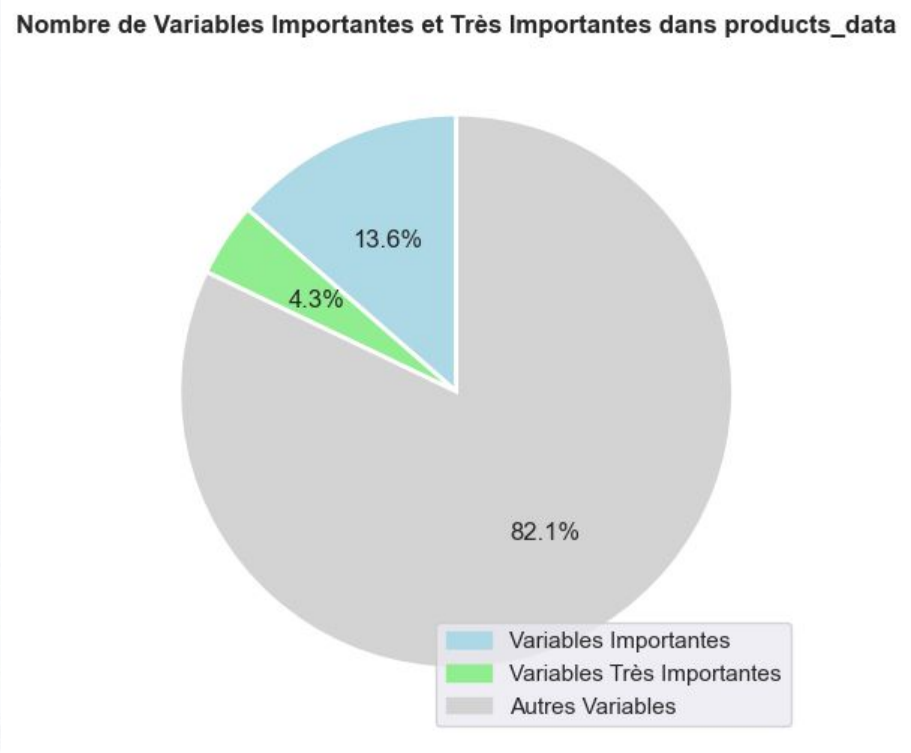
Une nouvelle analyse de correspondance a été effectuée entre les trois DataFrames.

Correspondances trouvées :

	Extrate Column	Products Column	NutriScore Column	Extrate Value count	Products Value count	NutriScore Value count	Total Value count
0	Energy	energy_100g	Energy	1	1	1	3
1	fat	energy-from-fat_100g	saturated fatty acids	1	2	2	5
2	Saturated fat	monounsaturated-fat_100g	saturated fatty acids	1	1	2	4
3	Sugars	sugars_100g	Simple sugars	1	1	1	3
4	Fiber	fiber_100g	fibers	1	1	1	3
5	Proteins	proteins_100g	Proteins	1	1	1	3
6	Salt	salt_100g	Salt	1	1	1	3
7	Fruits, vegetables, nuts and rapeseed, walnut ...	energy-from-fat_100g	rapeseed oils	1	2	1	4

Une visualisation a été générée pour représenter le nombre de variables importantes et très importantes dans le DataFrame products_data.

Environ 17,9% des variables étaient jugées importantes, fournissant une base solide pour le nettoyage des données.



03 →

Nettoyage du jeu de données

(a) Suppression des variables non pertinentes

Pour se concentrer sur les variables pertinentes à nos objectifs, toutes les colonnes non essentielles ont été supprimées du DataFrame products_data.

Seules les colonnes jugées importantes ou très importantes, ainsi que la colonne nutrition_grade_fr, ont été conservées.

Après cette étape de suppression, le DataFrame products_data contient les colonnes suivantes :

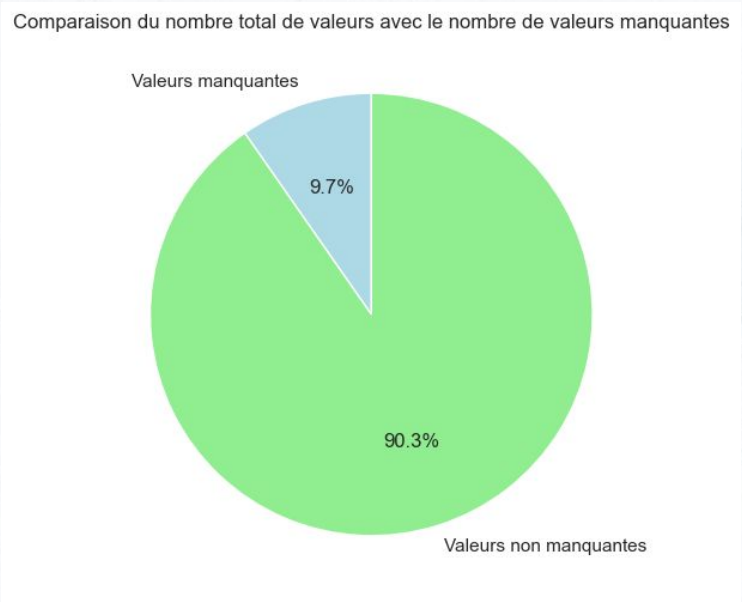
```
list(products_data.columns)
for col in products_data.columns:
    print(col)
```

```
code
quantity
packaging
packaging_tags
brands
brands_tags
categories
categories_tags
categories_fr
manufacturing_places
stores
nutrition_grade_fr
energy_100g
energy-from-fat_100g
saturated-fat_100g
monounsaturated-fat_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
salt_100g
silica_100g
bicarbonate_100g
potassium_100g
chloride_100g
calcium_100g
magnesium_100g
```

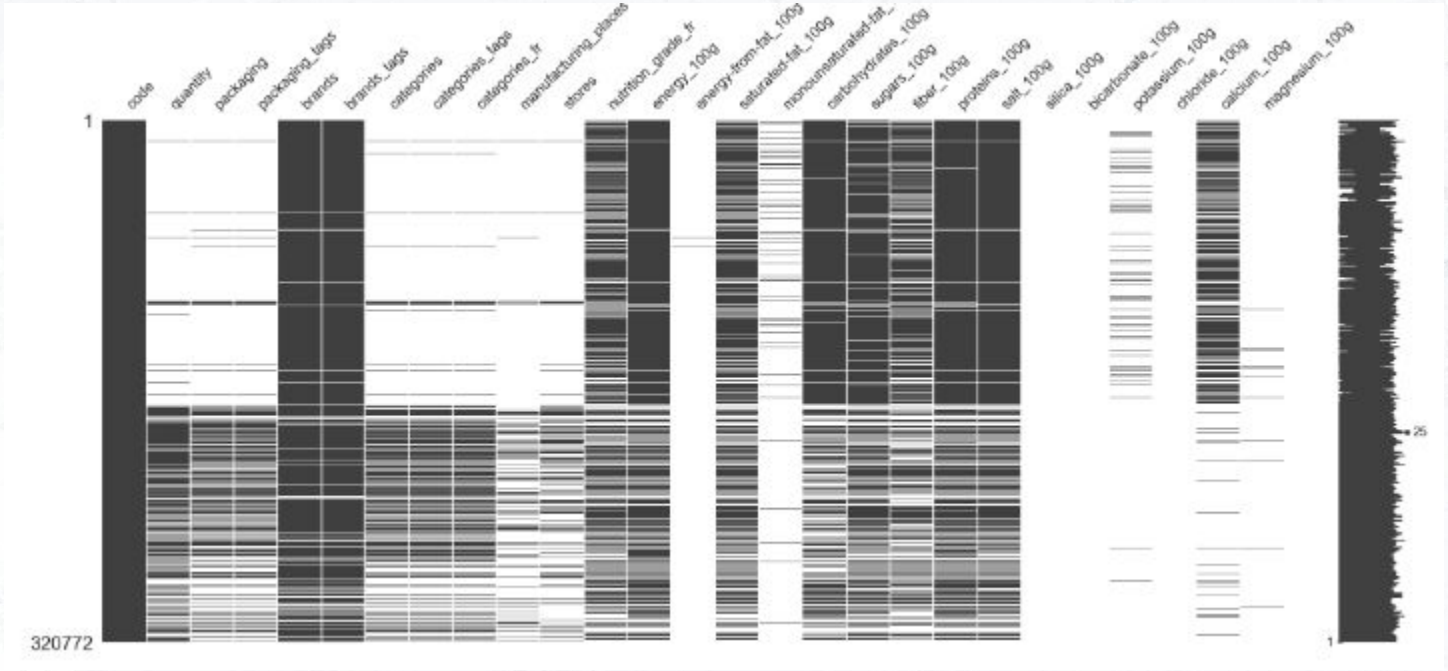
(b) Détection des valeurs manquantes

Dans le DataFrame products_data, il a été constaté qu'il y a un total de 5 038 448 valeurs manquantes.

Afin de mieux comprendre la répartition des valeurs manquantes, plusieurs visualisations ont été générées :

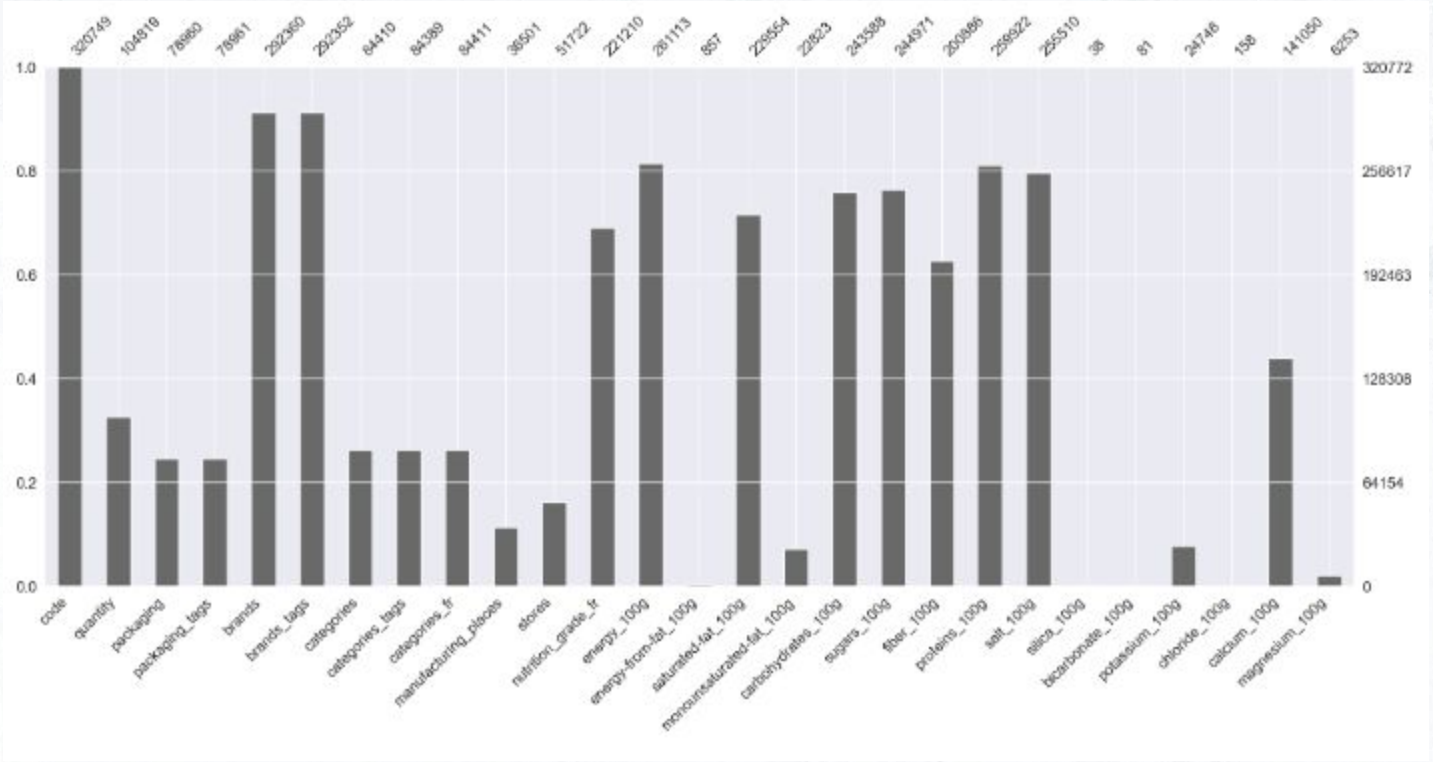


Ce diagramme compare le nombre total de valeurs dans le DataFrame avec le nombre de valeurs manquantes.
Il a été observé que les valeurs manquantes représentent une proportion significative des données.



La matrice offre une vue d'ensemble rapide des zones où les données sont manquantes.

Chaque ligne représente une entrée du DataFrame, et les colonnes blanches indiquent les valeurs manquantes.

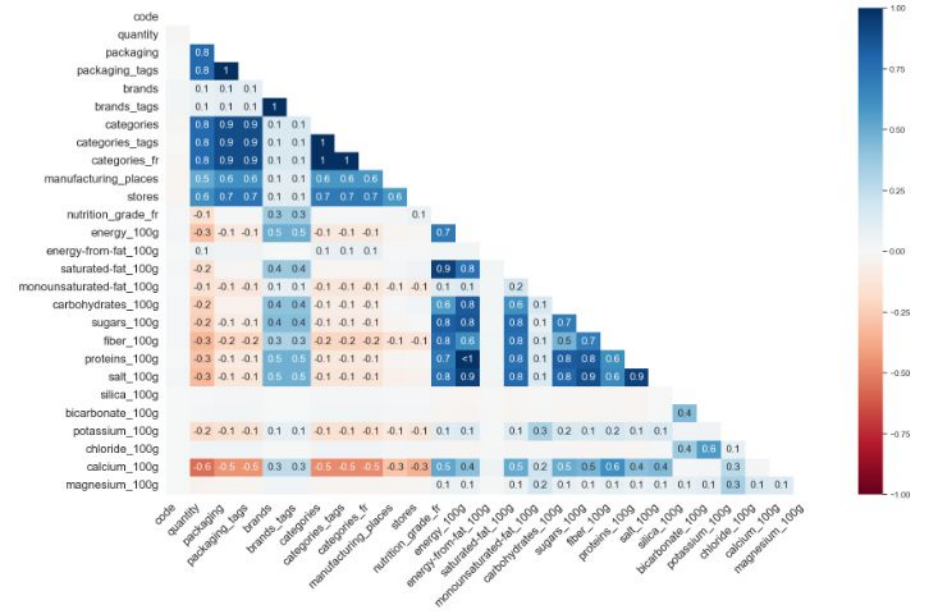


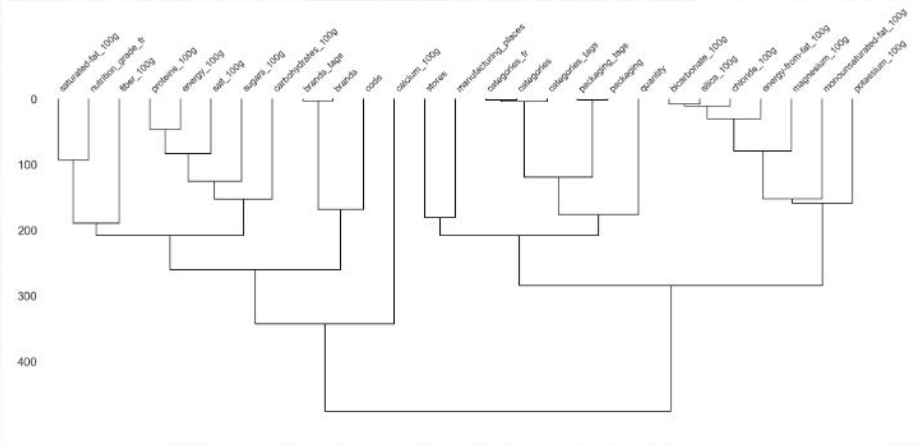
Ce diagramme montre le nombre de valeurs manquantes pour chaque colonne du DataFrame.

(c) Recherche de corrélats pour les valeurs manquantes

Des heatmaps et des dendrogrammes ont été utilisés pour identifier les corrélations entre les valeurs manquantes de différentes colonnes.

Ces visualisations ont aidé à comprendre si l'absence d'une valeur dans une colonne était liée à l'absence d'une valeur dans une autre colonne.





(d) Évaluation du Pourcentage de Valeurs Manquantes:

Un tableau a été généré pour montrer le nombre et le pourcentage de valeurs manquantes pour chaque colonne.

	Missing Values	Percentage
nutrition_grade_fr	99562	31.038245
fiber_100g	119886	37.374210
calcium_100g	179722	56.027958
quantity	215953	67.322896
categories_fr	236361	73.685047
categories	236362	73.685359
categories_tags	236383	73.691906
packaging_tags	241811	75.384073
packaging	241812	75.384385
stores	269050	83.875775
manufacturing_places	284271	88.620890
potassium_100g	296024	92.284863
monounsaturated-fat_100g	297949	92.884977
magnesium_100g	314519	98.050640
energy-from-fat_100g	319915	99.732832
chloride_100g	320614	99.950744
bicarbonate_100g	320691	99.974748
silica_100g	320734	99.988154

Sur la base de cette analyse, il a été décidé de ne pas imputer les données pour les colonnes ayant moins de 30% de valeurs manquantes, en se basant sur des recommandations scientifiques.

Cette approche est soutenue par des recherches dans le domaine, comme le montrent les articles de Dr. Graham et de Dr. Schafer.

- The prevention and handling of the missing data
- When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts

(e) Analyse de Similarité entre les Colonnes :

Une étude détaillée de la similarité entre différentes colonnes a été réalisée.

Cette étude a révélé que certaines colonnes présentaient une forte similarité.

Pour éviter la redondance, certaines de ces colonnes ont été supprimées.

Exemples de similarité entre colonnes :

	Colonne 1	Colonne 2	Score	Valeur 1	Valeur 2
0	brands	brands_tags	97	tesco, encinger	tesco,encinger
1	brands	brands_tags	97	strongbow, heineken	strongbow,heineken
2	brands	brands_tags	97	berchtesgadener	berchtesgadener
3	brands	brands_tags	97	amalattea - granarolo,amalattea	amalattea-granarolo,amalattea
4	brands	brands_tags	96	yoplait,yopa!	yoplait,yopa

Moyenne globale de similarité : 76.46042360305762

Exemples de similarité entre colonnes :

	Colonne 1	Colonne 2	Score	Valeur 1	Valeur 2
0	packaging_tags	packaging	98	пластиковая-бутылка,полиэтилентерефталат	пластиковая бутылка,полиэтилентерефталат
1	packaging_tags	packaging	98	vaschetta,coperchio,plastica,banda-stagnata	vaschetta,coperchio,plastica,banda stagnata
2	packaging_tags	packaging	98	semi-conservas,bandeja-termo-formada,film,term...	semi-conservas,bandeja termo-formada,film,term...
3	packaging_tags	packaging	98	pots,plastique,suremballage-carton,opercule	pots,plastique,suremballage carton,opercule
4	packaging_tags	packaging	98	polyethylenterephthalat,84,tetrapack,tetrapack	Polyethylenterephthalat,84,tetrapack,tetrapack

Moyenne globale de similarité : 96.66047535321037

Exemples de similarité entre colonnes :

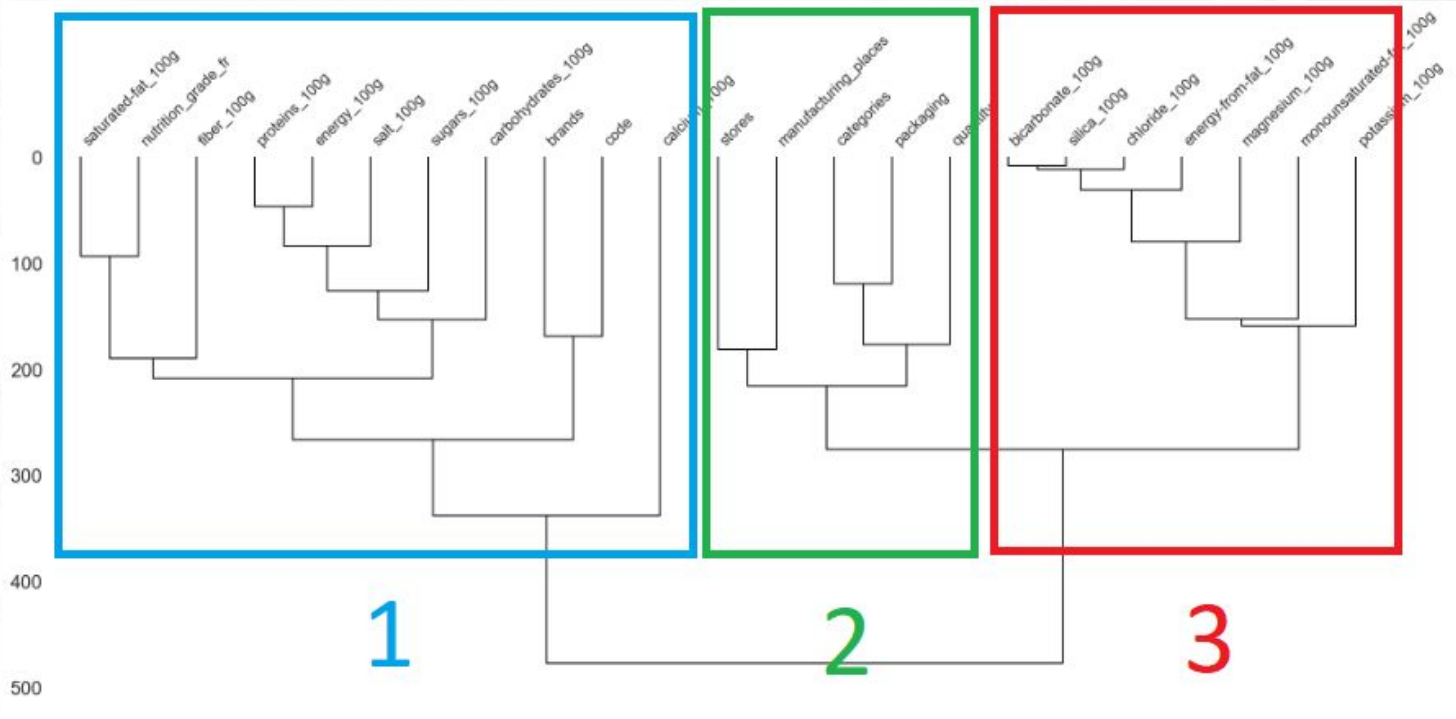
	Colonne 1	Colonne 2	Score	Valeur 1	Valeur 2
0	categories	categories_fr	100	Édulcorants,Sucres,Sucres roux,Sucres de canne...	Édulcorants,Sucres,Sucres roux,Sucres de canne...
1	categories	categories_fr	100	Édulcorants,Sucres,Sucres roux,Sucres de canne...	Édulcorants,Sucres,Sucres roux,Sucres de canne...
2	categories	categories_fr	100	Édulcorants,Sucres,Sucres roux,Sucres de canne...	Édulcorants,Sucres,Sucres roux,Sucres de canne...
3	categories	categories_fr	100	Édulcorants,Sucres,Sucres roux	Édulcorants,Sucres,Sucres roux
4	categories	categories_fr	100	Édulcorants,Sucres,Sucres gélifiant	Édulcorants,Sucres,Sucres gélifiant
5	categories	categories_tags	100	zh:复原乳	zh:复原乳
6	categories	categories_tags	100	pt:matte	pt:matte
7	categories	categories_tags	100	nan	nan
8	categories	categories_tags	100	nan	nan
9	categories	categories_tags	100	nan	nan
10	categories_fr	categories_tags	100	zh:复原乳	zh:复原乳
11	categories_fr	categories_tags	100	xx:煎茶	xx:煎茶
12	categories_fr	categories_tags	100	xx:スナック菓子,xx:ポップコーン	xx:スナック菓子,xx:ポップコーン
13	categories_fr	categories_tags	100	nl:100-kokosnootwater	nl:100-kokosnootwater
14	categories_fr	categories_tags	100	nan	nan

Moyenne globale de similarité : 84.85771098059264

(f) Gestions des valeurs manquantes :

Après avoir analysé les graphiques, il est possible de déduire les variables qui sont corrélées entre elles et celles qui ne le sont pas.

Le dendrogramme montre trois groupes de variables ayant des corrélations, suggérant que ces données manquantes sont probablement de type MNAR (Missing Not At Random).



Un aperçu des données montre le pourcentage de valeurs manquantes pour chaque colonne.

Certaines variables, comme energy-from-fat_100g et silica_100g, ont presque 100% de valeurs manquantes.

Les colonnes avec plus de 90% de valeurs manquantes ont été supprimées car elles sont peu susceptibles d'apporter des informations utiles.

```
missing_values = products_data.isnull().sum()
missing_percentage = 100 * missing_values / len(products_data)

print(missing_percentage)
```

```
code                          0.007170
quantity                     67.322896
packaging                    75.384385
brands                        8.857382
categories                   73.685359
manufacturing_places         88.620890
stores                       83.875775
nutrition_grade_fr           31.038245
energy_100g                  18.598568
energy-from-fat_100g         99.732832
saturated-fat_100g           28.437021
monounsaturated-fat_100g     92.884977
carbohydrates_100g           24.061951
sugars_100g                  23.630803
fiber_100g                   37.374210
proteins_100g                18.969860
salt_100g                    20.345292
silica_100g                  99.988154
bicarbonate_100g             99.974748
potassium_100g               92.284863
chloride_100g                99.950744
calcium_100g                 56.027958
magnesium_100g               98.050640
dtype: float64
```

Malgré la suppression des colonnes avec un taux élevé de valeurs manquantes, il reste encore des colonnes avec des données manquantes.

Ces colonnes sont divisées en deux groupes basés sur leurs corrélations.

Chaque groupe contient des variables qualitatives et quantitatives.

Pour les variables qualitatives, l'imputation par le mode est utilisée.

Groupe 1 :

Variables qualitatives :

- code
- brands
- nutrition_grade_fr

Variables quantitatives :

- energy_100g
- saturated-fat_100g
- carbohydrates_100g
- sugars_100g
- fiber_100g
- proteins_100g
- salt_100g
- calcium_100g

Groupe 2 :

Variables qualitatives :

- packaging
- categories
- manufacturing_places
- stores

Variables quantitatives :

- quantity

La colonne quantity est particulière car elle est de type object mais contient des informations quantitatives.

Les valeurs de cette colonne sont des chaînes contenant à la fois un nombre et une unité (par exemple "1kg").

Une fonction est utilisée pour extraire la partie numérique de ces valeurs et convertir toutes les valeurs en grammes.

Une exception est faite pour 'nutrition_grade_fr' qui donne le Nutri-score est un système d'étiquetage nutritionnel à cinq niveaux, allant de A à E et du vert au rouge.

La formule exacte n'étant pas disponible et le taux de valeurs manquante étant faible les valeurs contenant des NaN ont été supprimer

Imputations pour les variables quantitatives.

Les variables sont classées en fonction de la méthode d'imputation la plus efficace.

Les variables sont ensuite imputées en utilisant la méthode KNN ou IterativeImputer, selon ce qui est jugé le plus efficace.

	Variable	Méthode	Erreur	Pourcentage d'erreur
0	energy_100g	KNN	240713.877935	2195.092814
1	energy_100g	IterativeImputer	242318.045681	2209.721372
2	saturated-fat_100g	KNN	32.675519	35.907164
3	saturated-fat_100g	IterativeImputer	33.382720	36.684307
4	carbohydrates_100g	KNN	286.335407	286.335407
5	carbohydrates_100g	IterativeImputer	177581.675108	177581.675108
6	sugars_100g	KNN	213.228943	213.228943
7	sugars_100g	IterativeImputer	221.116365	221.116365
8	fiber_100g	KNN	13.635794	27.271588
9	fiber_100g	IterativeImputer	13.966023	27.932046
10	proteins_100g	KNN	47.840583	47.840583
11	proteins_100g	IterativeImputer	48.418950	48.418950

En regardant les résultats de l'imputation, on peut voir que certaines variables ont un taux d'erreur élevé.
Cela suggère que les méthodes d'imputation que nous avons utilisées, à savoir KNN et IterativeImputer, peuvent ne pas être les meilleures pour ces variables spécifiques.
A savoir : energy_100g, carbohydrates_100g, sugars_100g et quantity
Nous avons opté pour une approche par la médiane, sauf pour la variable "quantity", celle-ci ne risquant pas de biaise l'analyse nous obtenons pour une approche par la mise à zéro.

(g) Gestion des valeurs aberrantes :

Approche du métier :

Avant de plonger dans les méthodes statistiques, il ete essentiel de comprendre le contexte métier.

Pour les valeurs qui doivent être comprises entre 0 et 100g, toute valeur négative ou supérieure à 100g est considérée comme aberrante.

Après avoir identifié ces valeurs aberrantes, elles sont remplacées par NaN

Cependant, nous avons exclus la feature 'energy_100g' car elle est généralement exprimée en kilocalories (kcal) ou en kilojoules (kJ), et non en grammes.

Selon l'ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail) en France, il n'y a pas de limite maximale spécifique pour l'énergie (en kilocalories ou kilojoules) dans un produit alimentaire.

Cependant, des recommandations existent pour la consommation énergétique quotidienne. L'examen des statistiques descriptives de 'energy_100g' montre une valeur maximale de 69292.0, ce qui semble être une valeur aberrante.

Approche du métier :

Avant de plonger dans les méthodes statistiques, il ete essentiel de comprendre le contexte métier.

Pour les valeurs qui doivent être comprises entre 0 et 100g, toute valeur négative ou supérieure à 100g est considérée comme aberrante.

Après avoir identifié ces valeurs aberrantes, elles sont remplacées par NaN

Cependant, nous avons exclus la feature 'energy_100g' car elle est généralement exprimée en kilocalories (kcal) ou en kilojoules (kJ), et non en grammes.

Selon l'ANSES (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail) en France, il n'y a pas de limite maximale spécifique pour l'énergie (en kilocalories ou kilojoules) dans un produit alimentaire.

Cependant, des recommandations existent pour la consommation énergétique quotidienne. L'examen des statistiques descriptives de 'energy_100g' montre une valeur maximale de 69292.0, ce qui semble être une valeur aberrante.

Approche du métier :

Nous avons donc procéder à l'identification des valeurs aberrantes en utilisant différentes méthodes, puis nous sélectionnons celle qui se révèle la plus efficace pour la feature 'energy_100g'.

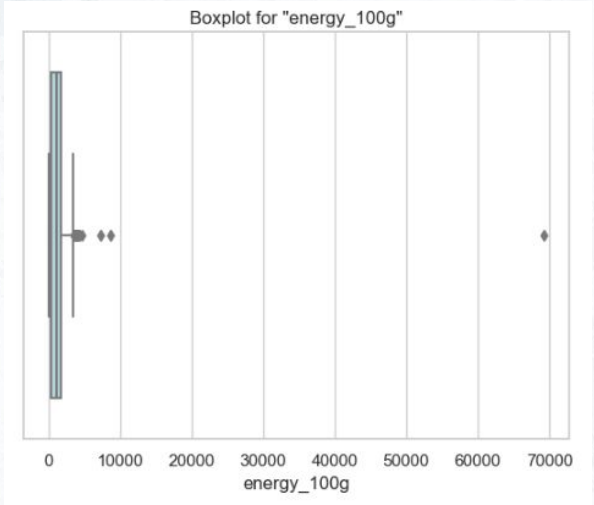
Approche de l'interquartile (IQR) :

Cette méthode utilise les quartiles pour identifier les valeurs aberrantes. Les valeurs en dehors de la plage Q1 - 1,5 * IQR à Q3 + 1,5 * IQR sont considérées comme aberrantes. Pour 'energy_100g', 49 valeurs aberrantes ont été identifiées.

Approche du Z-score :

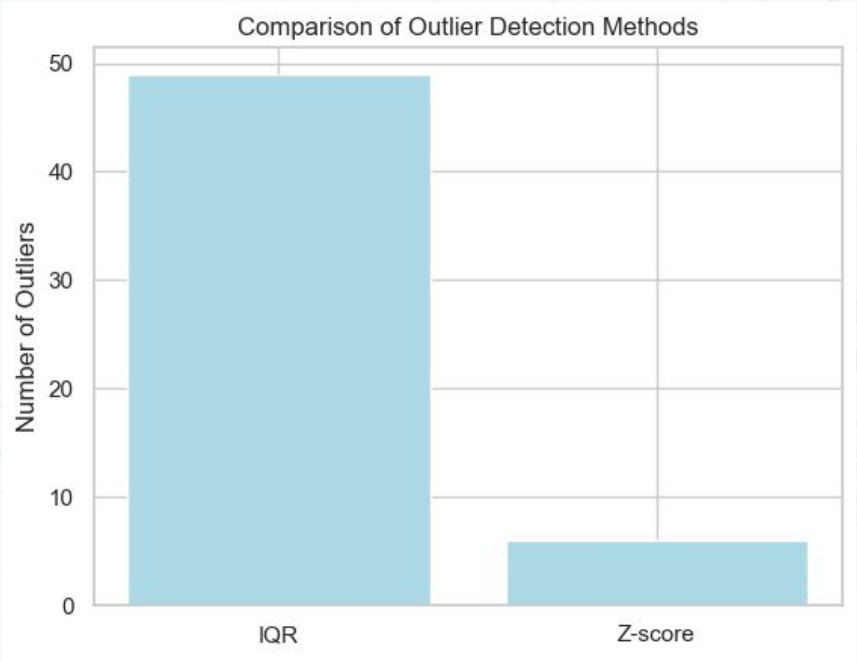
Cette méthode est basée sur l'écart par rapport à la moyenne des données. Les valeurs ayant un Z-score supérieur à un certain seuil sont considérées comme aberrantes.

Pour 'energy_100g', 6 valeurs aberrantes ont été identifiées.



Comparaison des méthodes :

En comparant l'IQR et le Z-score, l'IQR a identifié plus de valeurs aberrantes.



Dans notre analyse, nous avons choisi d'utiliser l'approche de l'intervalle interquartile (IQR) pour détecter et filtrer les valeurs aberrantes.

L'IQR est une méthode robuste qui est moins sensible aux valeurs extrêmes que d'autres méthodes, comme le Z-score.

De plus, l'IQR ne fait pas d'hypothèses sur la distribution des données, ce qui le rend particulièrement adapté à notre ensemble de données.

En utilisant l'IQR, les données aberrantes sont filtrées, laissant 11008 points de données sur les 11057 initiaux.

04 →

Analyse univariée

(a) Variables qualitatives :

Analyse statistiques :

```
Statistiques pour packaging:
Nombre de modalités uniques : 1215
Mode : Carton
```

	packaging	Effectifs	Frequences
0	Carton	8451	0.764312
1	Sachet,Plastique	78	0.007054
2	Plastique	74	0.006693
3	Bouteille,Plastique	52	0.004703
4	Kunststoff	48	0.004341

```
Statistiques pour brands:
Nombre de modalités uniques : 5768
Mode : Carrefour
```

	brands	Effectifs	Frequences
0	Carrefour	370	0.033463
1	Auchan	100	0.009044
2	U	83	0.007507
3	Leader Price	74	0.006693
4	Ahold	66	0.005969

```
Statistiques pour categories:
Nombre de modalités uniques : 2307
Mode : Snacks sucrés,Biscuits et gâteaux,Biscuits
```

	categories	Effectifs	Frequences
0	Snacks sucrés,Biscuits et gâteaux,Biscuits	8227	0.744054
1	Snacks sucrés,Biscuits et gâteaux,Biscuits,Bis...	14	0.001266
2	Biscuits	13	0.001176
3	Snacks salés,Apéritif,Biscuits apéritifs	11	0.000995
4	Snacks sucrés,Chocolats,Chocolats au lait	10	0.000904

```
Statistiques pour manufacturing_places:
Nombre de modalités uniques : 516
Mode : France
```

	manufacturing_places	Effectifs	Frequences
0	France	10088	0.912363
1	Italie	59	0.005336
2	Suisse	35	0.003165
3	Deutschland	31	0.002804
4	Belgique	31	0.002804

```
Statistiques pour stores:
Nombre de modalités uniques : 341
Mode : Carrefour
```

	stores	Effectifs	Frequences
0	Carrefour	9379	0.848241
1	Leclerc	123	0.011124
2	Auchan	118	0.010672
3	Cora	88	0.007959
4	Intermarché	85	0.007687

Il y a une grande disparité dans les variables qualitatives, probablement due à l'imputation des valeurs manquantes.

Le jeu de données nécessite un travail supplémentaire pour améliorer la qualité.

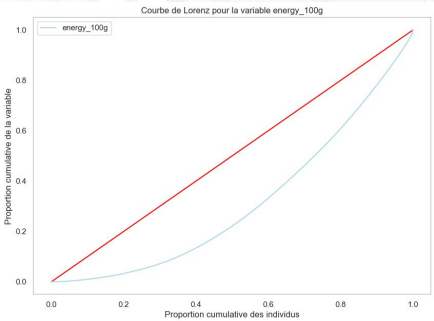
(b) Variables quantitatives :

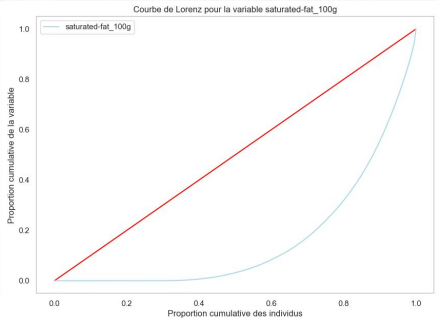
Statistiques descriptives :

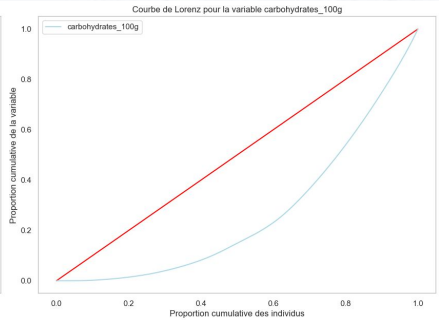
	Variable	Variance empirique	Variance empirique non biaisé	Ecart type	Coefficient de variation	Moyenne	Mediane	Mode	Etendue	Asymetrie	Aplatissement
0	energy_100g	1.010151e+06	1.010060e+06	1005.062631	0.853261	1177.907598	1159.000000	2092.000000	69292.000	28.355100	1906.403045
1	saturated-fat_100g	5.468336e+01	5.467841e+01	7.394820	1.504755	4.914301	1.790000	0.000000	95.000	3.218898	20.485456
2	carbohydrates_100g	7.402521e+02	7.401852e+02	27.207575	0.839483	32.409918	23.880000	23.880000	100.000	0.561069	-0.970861
3	sugars_100g	3.917629e+02	3.917275e+02	19.793002	1.324806	14.940303	5.000000	0.000000	100.000	1.691793	2.421697
4	fiber_100g	1.684129e+01	1.683977e+01	4.103814	1.490537	2.753245	2.000000	0.000000	83.300	4.950825	46.621132
5	proteins_100g	6.374278e+01	6.373701e+01	7.983907	1.037033	7.698799	5.620000	0.000000	90.000	2.034482	7.958028
6	salt_100g	2.004712e+01	2.004531e+01	4.477401	3.464052	1.292533	0.647000	0.000000	100.000	14.667409	262.969838
7	calcium_100g	1.979919e-02	1.979740e-02	0.140710	1.512186	0.093050	0.093049	0.093049	2.105	5.178047	37.172979

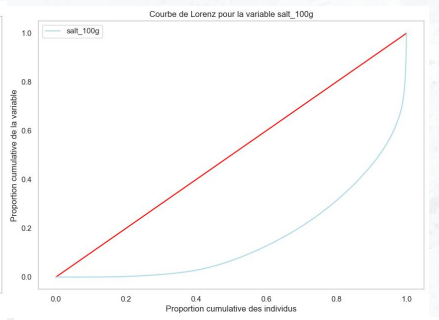
(a) Variables quantitatives :

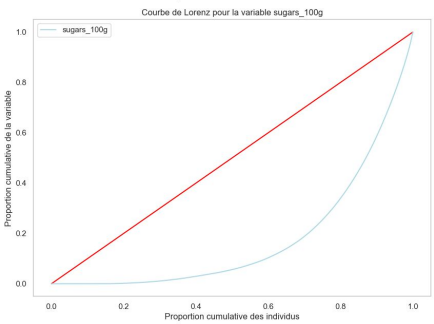
Mesures de concentration :

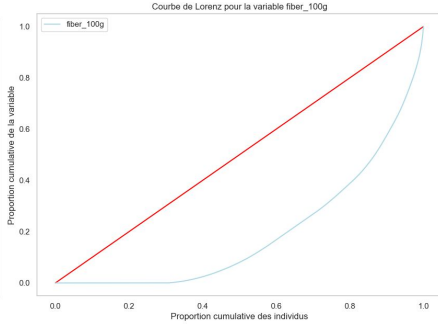


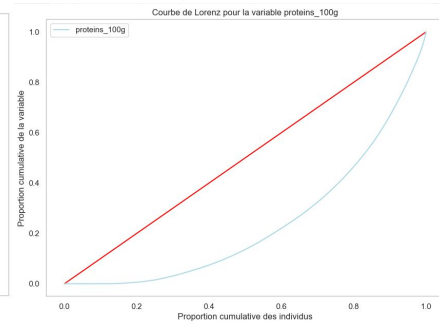


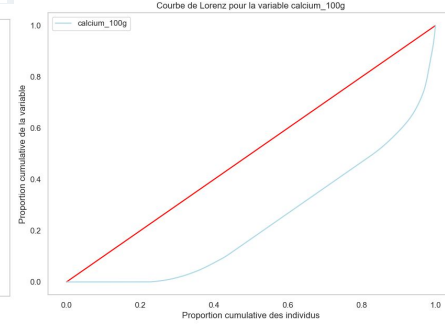












Mesures de concentration :

```
Le coefficient de Gini pour la variable energy_100g est 0.37
Le coefficient de Gini pour la variable saturated-fat_100g est 0.68
Le coefficient de Gini pour la variable carbohydrates_100g est 0.47
Le coefficient de Gini pour la variable sugars_100g est 0.65
Le coefficient de Gini pour la variable fiber_100g est 0.61
Le coefficient de Gini pour la variable proteins_100g est 0.52
Le coefficient de Gini pour la variable salt_100g est 0.69
Le coefficient de Gini pour la variable calcium_100g est 0.53
```

Suite à notre analyse des mesures de concentration, effectuée grâce à la courbe de Lorenz et au calcul du coefficient de Gini, nous avons pu distinguer trois catégories distinctes :

Les variables avec une dispersion élevée : "salt_100g" et "sugars_100g" avec des coefficients de Gini de 0.67 et 0.64 respectivement.

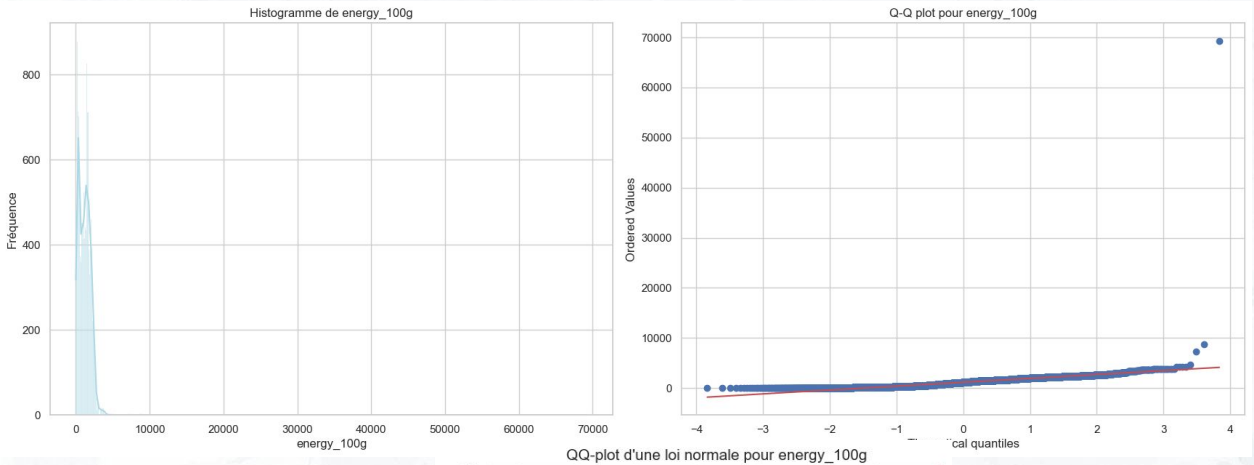
Cela signifie que cesvariables présentent une grande inégalité dans leur distribution, avec probablement quelques valeurs extrêmes.

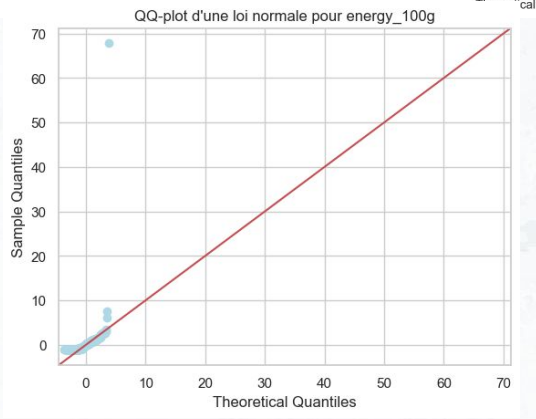
Les variables avec une dispersion modérée : "saturated-fat_100g", "carbohydrates_100g", "fiber_100g", "proteins_100g", "calcium_100g" avec des coefficients de Gini allant de 0.42 à 0.57.

Ces variables présentent une dispersion modérée dans leur distribution. La variable avec une dispersion faible : "energy_100g" avec un coefficient de Gini de 0.36.

Cela indique une distribution plus uniforme parmi les produits pour cette variable.

Analyse de normalité:





	Variable	Nom du test	Statistique de test	p-valeur	Normalité
0	energy_100g	Shapiro-Wilk	0.594515	0.000	Non
1	energy_100g	Kolmogorov-Smirnov	0.989296	0.000	Non
2	energy_100g	Anderson-Darling	211.856167	NaN	Non
3	energy_100g	D'Agostino-Pearson	28325.081258	0.000	Non
4	energy_100g	Lilliefors	0.120604	0.001	Non
5	saturated-fat_100g	Shapiro-Wilk	0.681687	0.000	Non
6	saturated-fat_100g	Kolmogorov-Smirnov	0.500000	0.000	Non
7	saturated-fat_100g	Anderson-Darling	976.283579	NaN	Non
8	saturated-fat_100g	D'Agostino-Pearson	8058.010489	0.000	Non
9	saturated-fat_100g	Lilliefors	0.253166	0.001	Non
10	carbohydrates_100g	Shapiro-Wilk	0.905101	0.000	Non
11	carbohydrates_100g	Kolmogorov-Smirnov	0.889320	0.000	Non
12	carbohydrates_100g	Anderson-Darling	372.205794	NaN	Non
13	carbohydrates_100g	D'Agostino-Pearson	2672.051502	0.000	Non
14	carbohydrates_100g	Lilliefors	0.160041	0.001	Non
15	sugars_100g	Shapiro-Wilk	0.756068	0.000	Non
16	sugars_100g	Kolmogorov-Smirnov	0.685545	0.000	Non
17	sugars_100g	Anderson-Darling	982.595269	NaN	Non
18	sugars_100g	D'Agostino-Pearson	3312.298635	0.000	Non
19	sugars_100g	Lilliefors	0.225176	0.001	Non
20	fiber_100g	Shapiro-Wilk	0.609666	0.000	Non
21	fiber_100g	Kolmogorov-Smirnov	0.500000	0.000	Non
22	fiber_100g	Anderson-Darling	927.571043	NaN	Non
23	fiber_100g	D'Agostino-Pearson	11318.169959	0.000	Non
24	fiber_100g	Lilliefors	0.251142	0.001	Non
25	proteins_100g	Shapiro-Wilk	0.824300	0.000	Non
26	proteins_100g	Kolmogorov-Smirnov	0.726352	0.000	Non
27	proteins_100g	Anderson-Darling	462.394173	NaN	Non
28	proteins_100g	D'Agostino-Pearson	5013.116084	0.000	Non
29	proteins_100g	Lilliefors	0.167450	0.001	Non
30	salt_100g	Shapiro-Wilk	0.190836	0.000	Non
31	salt_100g	Kolmogorov-Smirnov	0.500000	0.000	Non
32	salt_100g	Anderson-Darling	2479.558779	NaN	Non

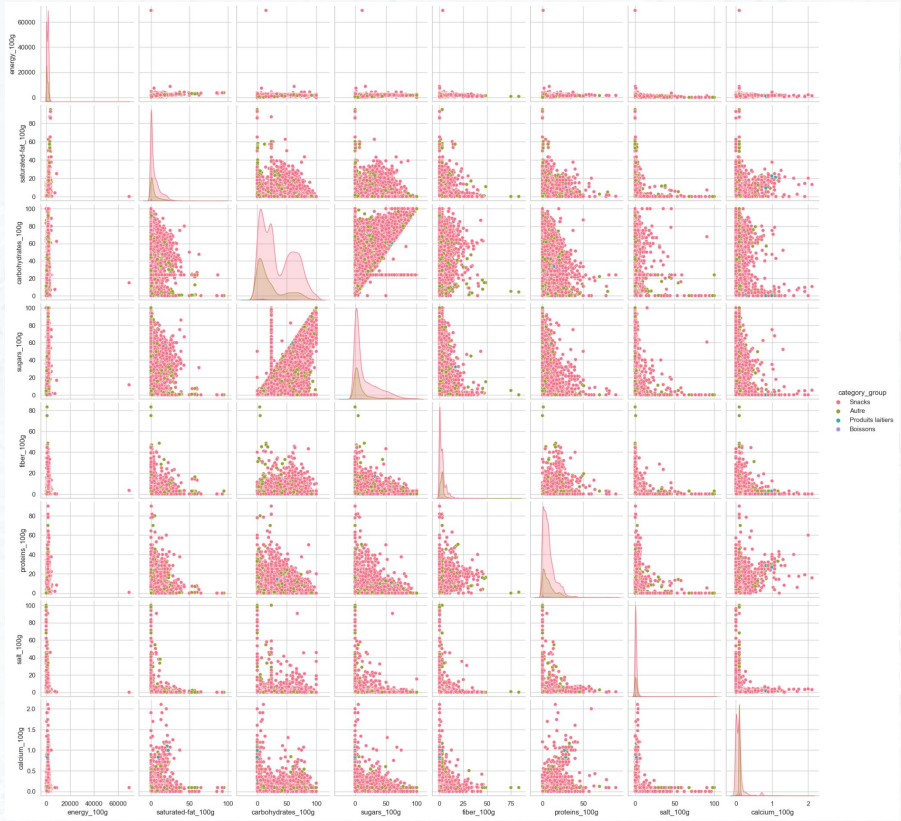
Plusieurs tests et graphiques (histogrammes, boxplots, Q-Q plots) ainsi que des tests statistiques ont montré que la plupart des variables ne suivent pas une distribution normale.

05 →

Analyse bivariée

(a) Variables qualitatives :

Pour rechercher des corrélations ou des relations entre ces variables qualitatives, nous avons utiliser des tableau de contingence, le paitplot, le test de Kruskal-Wallis, Le test du chi-carré.



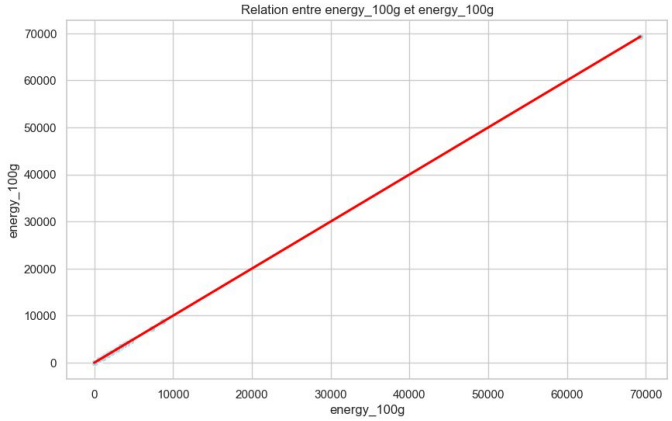
	Correlation	Percentage
(packaging, brands)	34	38
(packaging, categories)	28	28
(packaging, manufacturing_places)	40	44
(packaging, stores)	61	68
(brands, packaging)	34	38
(brands, categories)	25	28
(brands, manufacturing_places)	23	33
(brands, stores)	15	25
(categories, packaging)	28	28
(categories, brands)	25	28
(categories, manufacturing_places)	20	40
(categories, stores)	34	38
(manufacturing_places, packaging)	40	44
(manufacturing_places, brands)	23	33
(manufacturing_places, categories)	20	40
(manufacturing_places, stores)	52	52
(stores, packaging)	61	68
(stores, brands)	15	25
(stores, categories)	34	38
(stores, manufacturing_places)	52	52

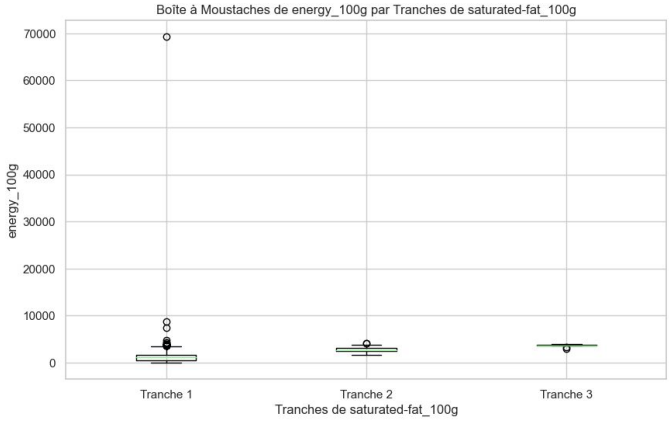
Les variables quantitatives montrent des différences significatives entre les catégories de produits néanmoins celle ci sont associées entre elles, à quelques exceptions près.

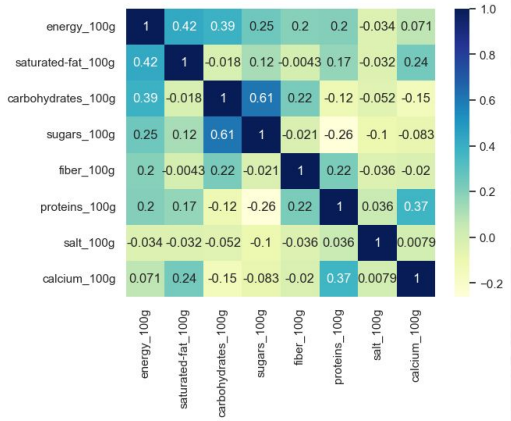
Ces résultats suggèrent des variations dans la formulation des produits et des tendances spécifiques dans leur emballage et leur marque.

(b) Variables quantitatives :

Pour rechercher des corrélations ou des relations entre ces variables quantitatives, nous avons utiliser des diagramme de dispersion, des Boîte à moustaches, Heatmap, nous avons également calculer Covariance empirique et coefficient de corrélation de Pearson.





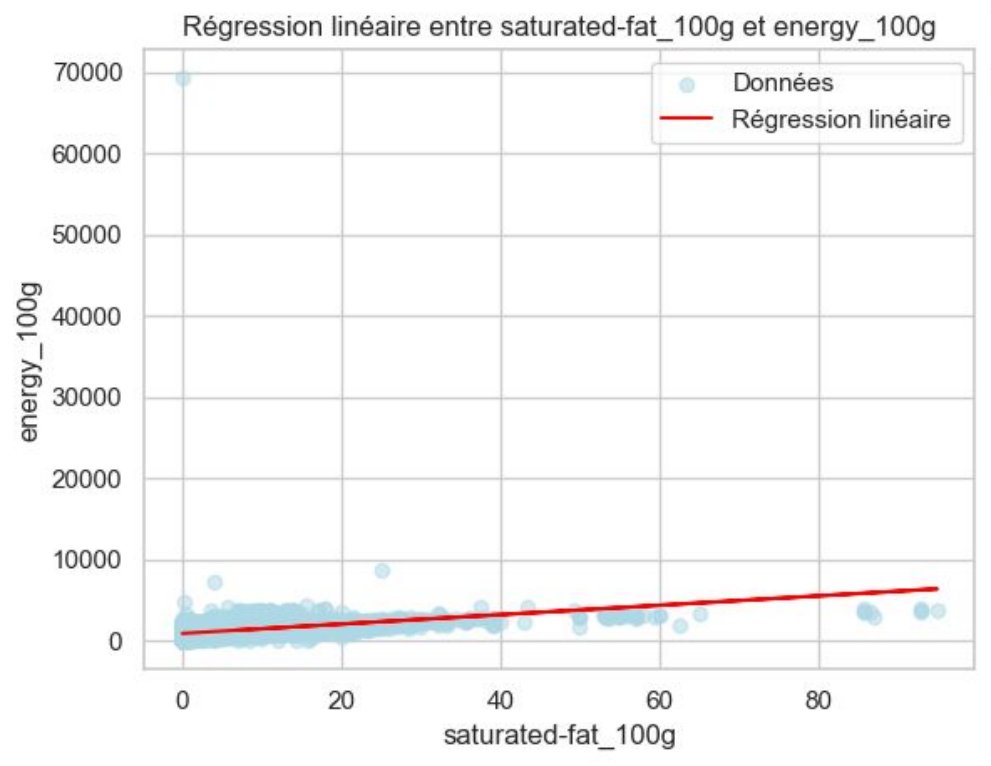


		Pearson Coefficient	Covariance
energy_100g	saturated-fat_100g	0.424966	3158.170039
	carbohydrates_100g	0.388567	10624.516726
	sugars_100g	0.253207	5036.645136
	fiber_100g	0.196211	809.216279
	proteins_100g	0.201584	1617.429200
	salt_100g	-0.034427	-154.908425
	calcium_100g	0.071471	10.106634

Suite à notre analyse, il a été observé que la variable saturated-fat_100g présente le coefficient de Pearson le plus élevé (en valeur absolue) en relation avec energy_100g.

Cette variable démontre une corrélation positive modérée avec energy_100g.

Nous avons donc décidé d'initier notre étude par une régression linéaire impliquant ces deux variables.




```
print("Coefficient de détermination (R²):", model.score(X, Y))
```

```
Coefficient de détermination (R²): 0.18059600986643487
```

Le graphique montre clairement une tendance ascendante, indiquant que lorsque la quantité de graisses saturées augmente, l'énergie du produit tend également à augmenter. Cependant, le coefficient de détermination (R^2) est de 0.18.

Cela signifie que seulement 18% de la variabilité de l'énergie peut être expliquée par la quantité de graisses saturées.

06 →

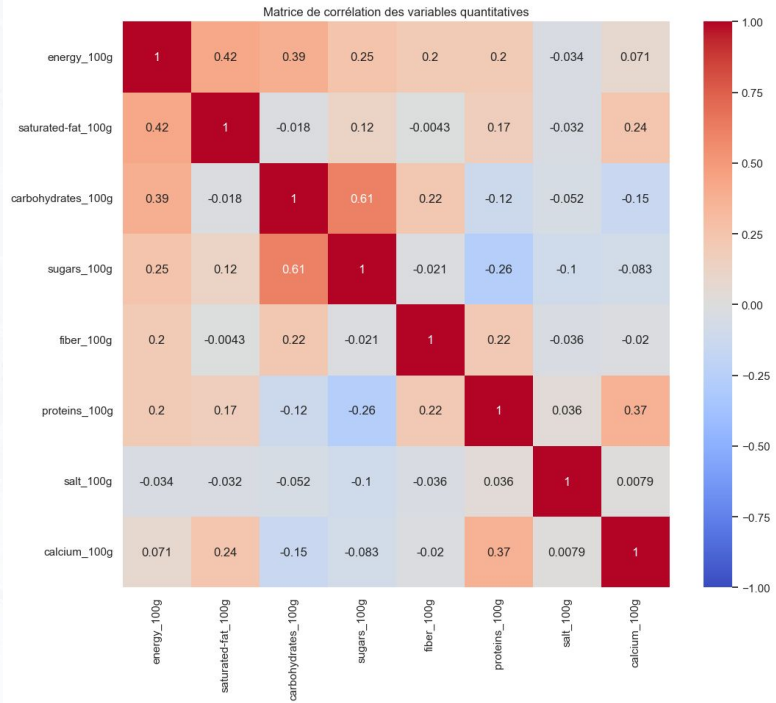
Analyse multivariée

(a) Réduction de Dimensionnalité :

Afin de faciliter l'analyse, nous avons regroupé les catégories en groupes plus larges basés sur des caractéristiques communes.

Suite à l'analyse des coefficients de corrélation entre nos variables quantitatives, il a été observé que les variables carbohydrates_100g et sugars_100g présentent une corrélation significative.

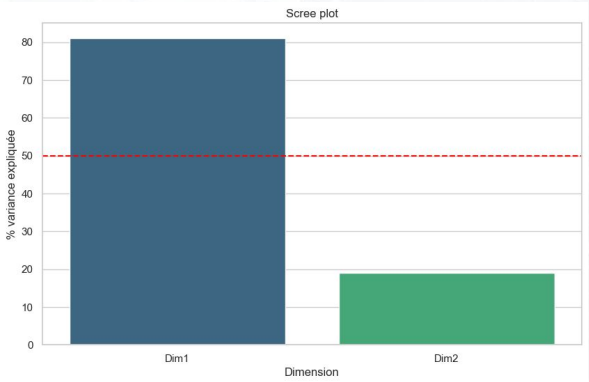
16	carbohydrates_100g	energy_100g	0.388567	Faible
17	carbohydrates_100g	saturated-fat_100g	-0.018265	Nul
19	carbohydrates_100g	sugars_100g	0.610790	Moyen
20	carbohydrates_100g	fiber_100g	0.224116	Faible
21	carbohydrates_100g	proteins_100g	-0.122048	Nul
22	carbohydrates_100g	salt_100g	-0.051846	Nul
23	carbohydrates_100g	calcium_100g	-0.146473	Nul
24	sugars_100g	energy_100g	0.253207	Faible
25	sugars_100g	saturated-fat_100g	0.122076	Nul
26	sugars_100g	carbohydrates_100g	0.610790	Moyen
28	sugars_100g	fiber_100g	0.020662	Nul

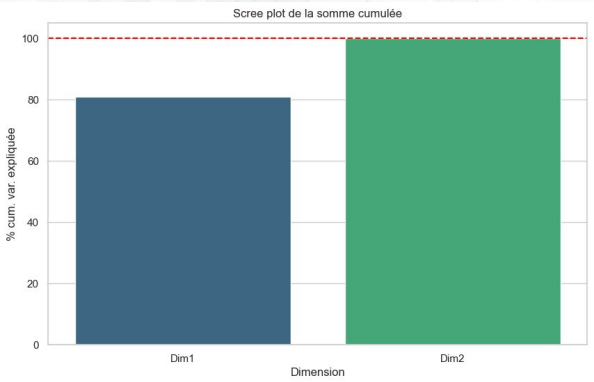


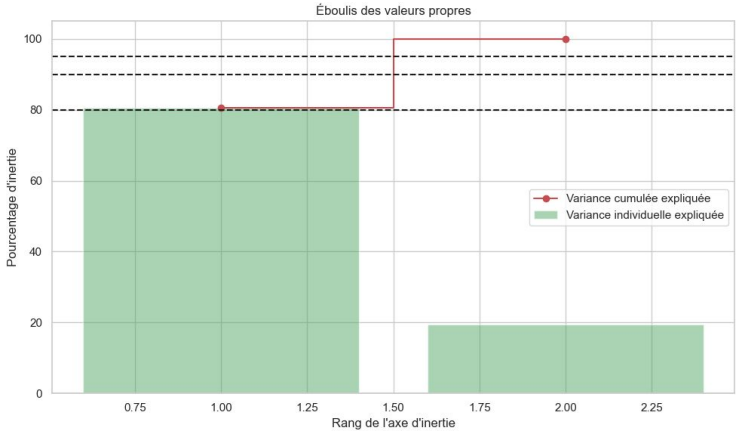
Afin d'optimiser la représentativité de notre jeu de données tout en réduisant sa dimensionnalité, nous avons décidé d'appliquer une Analyse en Composantes Principales (ACP) sur ces deux variables spécifiques.

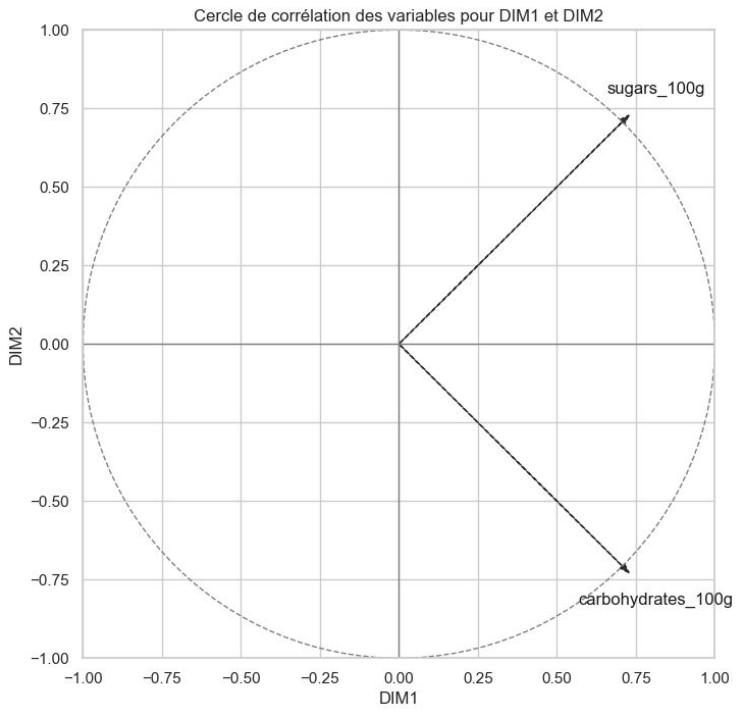
Suite à notre analyse des variances expliquées par les composantes principales, il est clairement observé que la première dimension (Dim1) explique à elle seule 82.70% de la variance totale, tandis que la deuxième dimension (Dim2) apporte un complément pour atteindre 100%.

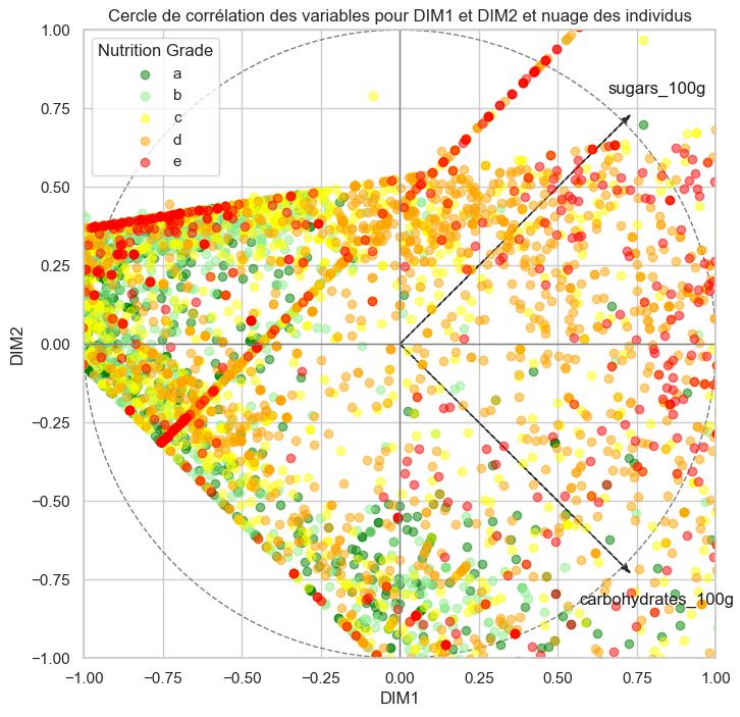
Ainsi, en considérant uniquement ces deux dimensions, nous sommes en mesure de représenter intégralement l'information contenue dans les variables carbohydrates_100g et sugars_100g. Cette réduction dimensionnelle nous offre une représentation synthétique et optimisée de notre jeu de données initial.



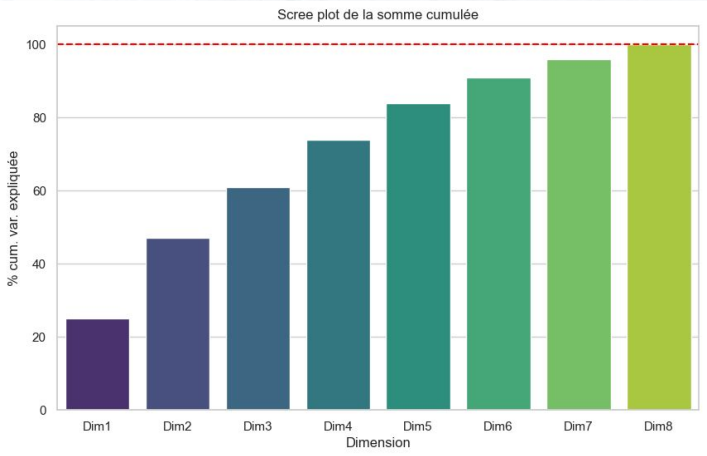


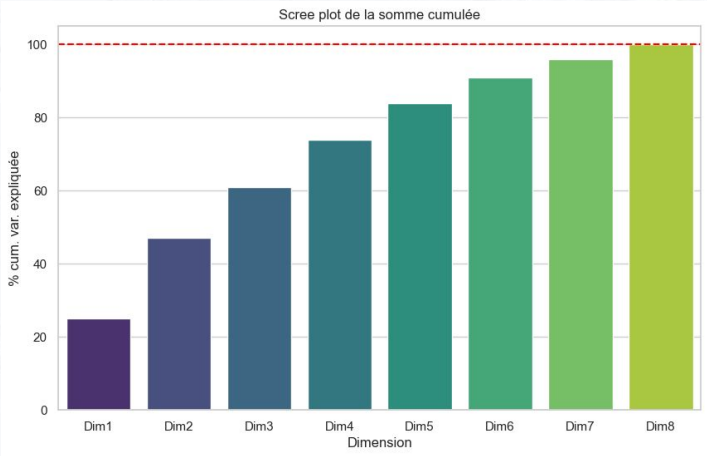


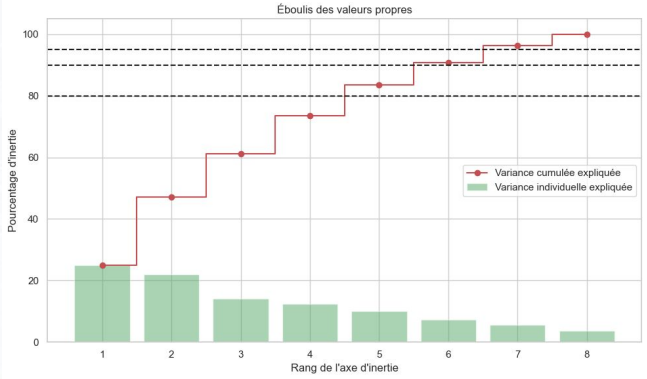

Cercle de corrélation des variables pour DIM1 et DIM2
sugars_100g
carbohydrates_100g
DIM1
DIM2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Cercle de corrélation des variables pour DIM1 et DIM2 et nuage des individus
Nutrition Grade
a
b
c
d
e
sugars_100g
carbohydrates_100g
DIM1
DIM2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

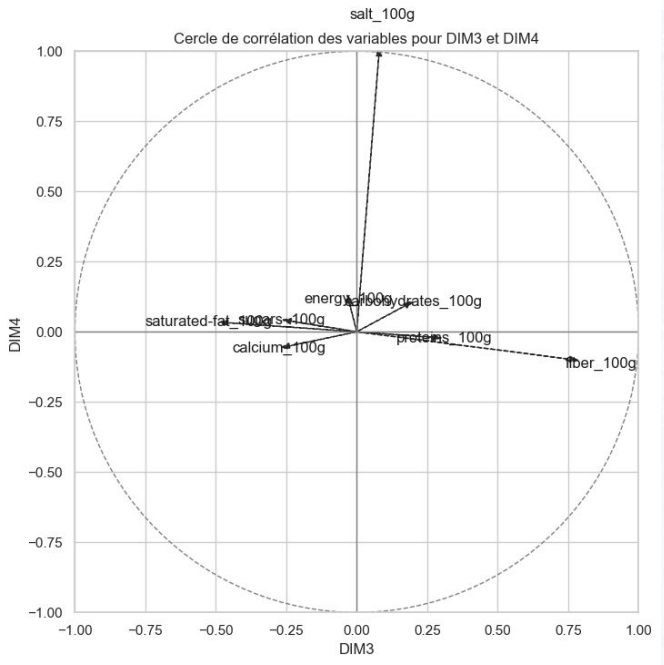
Nous avons ensuite réalisé une ACP sur l'ensemble de nos features

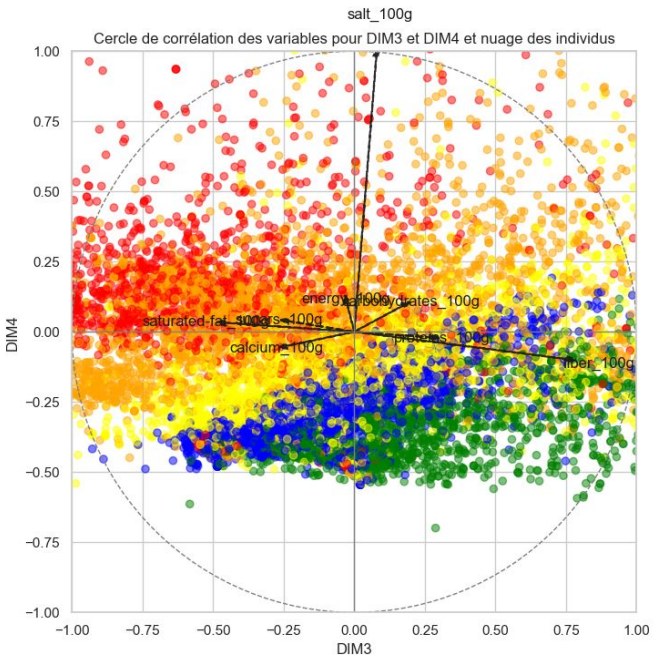






La première dimension (Dim1) explique 25.03% de la variance totale. La deuxième dimension (Dim2) ajoute 22.03%, portant la somme cumulée à 47.06%. Si nous ajoutons la troisième dimension (Dim3), qui explique 14.17% de la variance, la somme cumulée atteint 61.23%.
En considérant uniquement les trois premières dimensions, nous sommes en mesure de représenter plus de 60% de l'information contenue dans notre jeu de données initial.





Les deux premières dimensions, Dim1 et Dim2, contiennent à elles seules près de la moitié (47.06%) de l'information (variance) du jeu de données original. Cependant,il est judicieux de considérer au moins jusqu'à Dim4 ou Dim5, car elles permettent de couvrir plus de 73% à 83% de la variance totale.

Suite à l'analyse en composantes principales et à l'examen des cercles de corrélation des variables, conjugué au nuage des individus classés selon le grade de Nutri-Score, plusieurs observations pertinentes émergent :

La projection sur les dimensions Dim3 et Dim4 révèle une structuration significative des données.

Ces dimensions mettent en évidence les relations intrinsèques entre les caractéristiques étudiées et leur influence sur le Nutri-Score. La distribution des individus dans le nuage, en fonction de leur grade de Nutri-Score, suggère que certaines caractéristiques ont un impact prépondérant sur la classification nutritionnelle

(b) Variables Qualitatives et Quantitatives:

Nous avons exploré la relation entre une variable qualitative et une variable quantitative à l'aide de l'ANOVA.

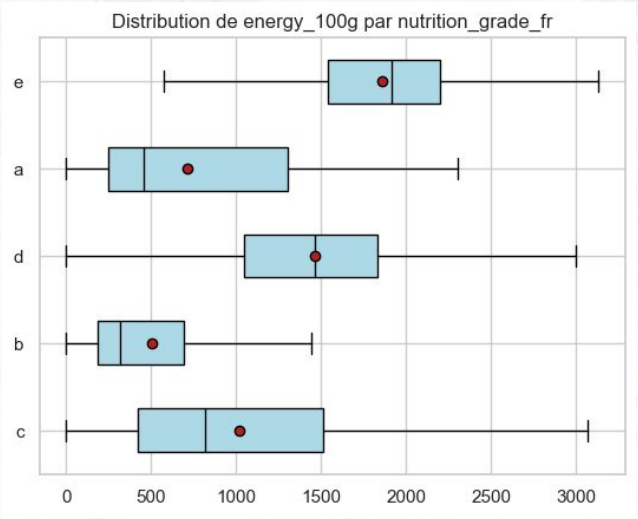
Cela nous a permis de déterminer si une relation significative existait entre les modalités de la variable qualitative et les valeurs de la variable quantitative.

Cette étape a renforcé les constatations de l'ACP.

Nous avons commencé par créer un sous-échantillon de nos données et renommé certaines colonnes pour faciliter l'analyse.

Des boîtes à moustaches ont été utilisées pour visualiser la distribution de chaque variable quantitative par rapport à la variable qualitative.

Les points rouges au milieu de chaque boîte représentent la moyenne des valeurs.



Nous avons utilisé l'ANOVA pour modéliser la relation entre "nutrition_grade_fr" et chaque variable quantitative.

Les résultats montrent la valeur F, la valeur p et le rapport de corrélation η^2 pour chaque variable.

Le rapport de corrélation η^2 a été utilisé pour évaluer la qualité du modèle.

Ce rapport varie entre 0 et 1 et représente le degré de corrélation entre la variable qualitative et la variable quantitative.

	Variable	F-value	p-value	η^2
0	energy_100g	779.555763	0.000000e+00	0.220055
1	saturated_fat_100g	1961.564290	0.000000e+00	0.415184
2	carbohydrates_100g	246.945621	1.633366e-203	0.082043
3	sugars_100g	667.253570	0.000000e+00	0.194520
4	fiber_100g	163.939512	1.447051e-136	0.056011
5	proteins_100g	79.080307	2.924041e-66	0.027825
6	salt_100g	51.866599	2.393066e-43	0.018426
7	calcium_100g	134.393174	2.680211e-112	0.046384

	Variable	η^2	Corrélation
0	energy_100g	0.220055	Forte
1	saturated_fat_100g	0.415184	Forte
2	carbohydrates_100g	0.082043	Modérée
3	sugars_100g	0.194520	Forte
4	fiber_100g	0.056011	Faible
5	proteins_100g	0.027825	Faible
6	salt_100g	0.018426	Faible
7	calcium_100g	0.046384	Faible

Les variables "energy_100g", "saturated_fat_100g" et "sugars_100g" montrent une forte corrélation avec "nutrition_grade_fr".

La variable "carbohydrates_100g" présente une corrélation modérée. Les autres variables montrent une faible corrélation.

L'analyse a confirmé certaines des constatations déduites de l'Analyse en Composantes Principales (ACP).

Plusieurs variables quantitatives montrent une corrélation significative avec la variable qualitative "nutrition_grade_fr".

Merci

Avez-vous des questions ?

GitHub

Cette présentation est en liens avec le noteBook :

“NoteBook_P3_Préparez_des_données_pour_un_organisme_de_santé_publique”