Projet 3 : Préparez des données pour un organisme de santé publique

11/06/2024

Soukaina GUAOUA ELJADDI

Parcours Data Scientist
OpenClassrooms

Plan:

- ❏ Problématique
- ❏ Nettoyage et filtration des features et produits
- ❏ Identification et traitement des valeurs aberrantes
- ❏ Identification et traitement des valeurs manquantes
- ❏ Analyses uni-variée et bi-variée
- ❏ Analyse multi-variée
- ❏ Conclusion

Problématique

Contexte : Projet d'**amélioration** de la base de données **Open Food Facts** de l'agence **santé publique france**.

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

Mission : **nettoyage** et **exploration** des données, afin de déterminer la faisabilité de l'idée d'application de Santé publique France.

Étape 1 : Nettoyez et filtrez des features et produits

Étape 1 : Nettoyez et filtrez des features et produits

Analyse exploratoire du jeu de données:

fr.openfoodfacts.org.products.csv : 320772 observations, 162 variables

Fiche produit	Tags	Ingrédients et additifs	Informations nutritionnels
- code - url - creator - created_t - product_name	- packaging_tags - brand_tags - categories_tags - origin_tags	- ingredients_text - allergens - additives	- fat_100g - sugars_100g - fiber_100g - salt_100g

Étape 1 : Nettoyez et filtrez des features et produits

Identification de la cible et des features pertinentes:

Données originales : 320772 lignes, 162 colonnes

Cible = ‘nutrition_grade_fr’

Features pertinentes : Features avec plus de 50% de valeurs présentes ⇒ 43 colonnes

Suppression des produits en double ⇒ 221210 lignes

Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression des colonnes redondantes: 'additives_n', 'additives', 'additives_tags'

'**additives_fr**' plus informative ⇒ 221210 lignes × 40 colonnes

Exemple :

product_name	Peanut Butter Power Chews
additives_n	3.0
additives	[peanut-butter -> en:peanut-butter] [but...
additives_tags	en:e170,en:e322,en:e410
additives_fr	E170 - Carbonate de calcium,E322 - Lécithines,...

Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression des colonnes Temps sous format _t (UNIX timestamp format): 'created_t' et 'last_modified_t'

_datetime (format ISO 8601) : '**created_datetime**', '**last_modified_datetime**' ⇒ 221210 lignes × 38 colonnes

Exemple:

product_name	Peanut Butter Power Chews
created_t	1489138486
created_datetime	2017-03-10T09:34:46Z
last_modified_t	1489138486
last_modified_datetime	2017-03-10T09:34:46Z

Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression des colonnes redondantes _tags : 'brands_tags', 'countries_tags', 'states_tags'

‘brands’, ‘countries’, ‘states’⇒ 221210 lignes × 35 colonnes

Exemple:

product_name	Peanut Butter Power Chews
brands_tags	sunridge
brands	Sunridge
countries_tags	en:united-states
countries	US
states_tags	en:to-be-completed,en:nutrition-facts-complete...
states	en:to-be-completed, en:nutrition-facts-complet...

Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression de la colonne redondante : 'nutrition-score-uk_100g'

'nutrition-score-fr_100g' ⇒ 221210 lignes × 34 colonnes

Exemple :

product_name	Peanut Butter Power Chews
nutrition-score-uk_100g	9.0
nutrition-score-fr_100g	9.0

Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression des colonnes redondantes : 'countries', 'states'

'countries_fr' et 'states_fr' ⇒ 221210 lignes × 32 colonnes

Exemple :

product_name	Peanut Butter Power Chews
countries	US
countries_fr	États-Unis
states	en:to-be-completed, en:nutrition-facts-complet...
states_fr	A compléter,Informations nutritionnelles compl...

Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression des colonnes inutiles pour l'analyse : 'code', 'url', 'creator', 'created_datetime', 'last_modified_datetime','brands', 'origins', 'countries_fr', 'ingredients_text', 'serving_size', 'additives_fr', 'states_fr'

⇒ 221210 lignes × 21 colonnes

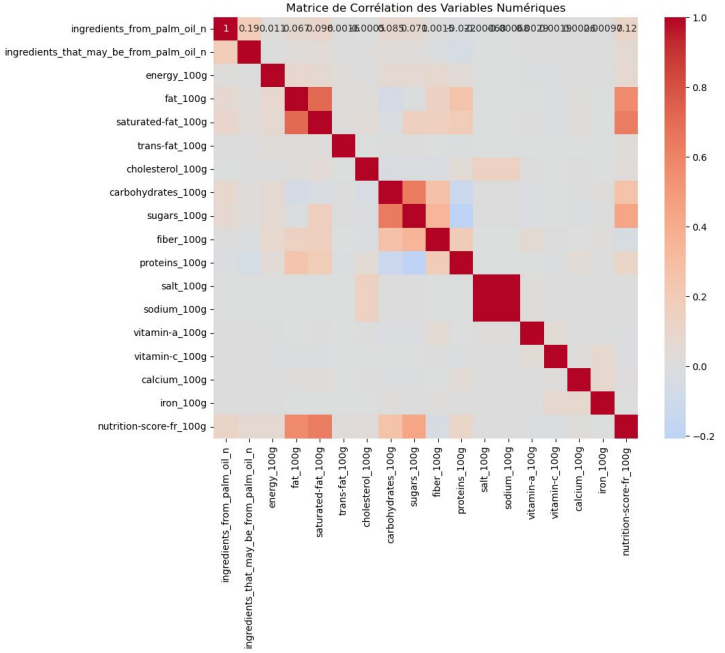
Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression d'une des 2 variables corrélées

- 'sodium_100g' == 'salt_100g'
- 'sugars_100g' == 'carbohydrates_100g'
- 'fat_100g' == 'saturated-fat_100g'

⇒ 221210 rows × 18 columns



Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression des 2 colonnes mal renseignés :
'ingredients_from_palm_oil_n' : 0 et NaN

et 'ingredients_that_may_be_from_palm_oil_n'

⇒ 221210 lignes × 16 colonnes

Étape 1 : Nettoyez et filtrez des features et produits

Filtrage et nettoyage des données :

Suppression de la colonne : 'nutrition-score-fr_100g' correspondante à 'nutrition_grade_fr'

⇒ 221210 lignes × 15 colonnes

nutrition_grade	nutrition-score
A	-15 et -2
B	-1 à +3
C	+4 à +11
D	+12 à +16
E	+17 à +40

Source :Wikipedia

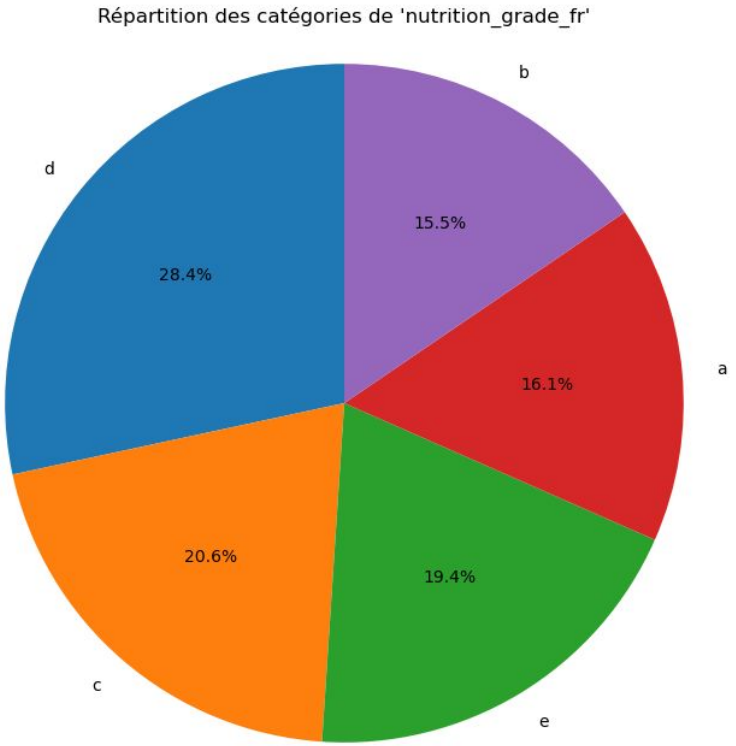
Étape 1 : Nettoyez et filtrez des features et produits

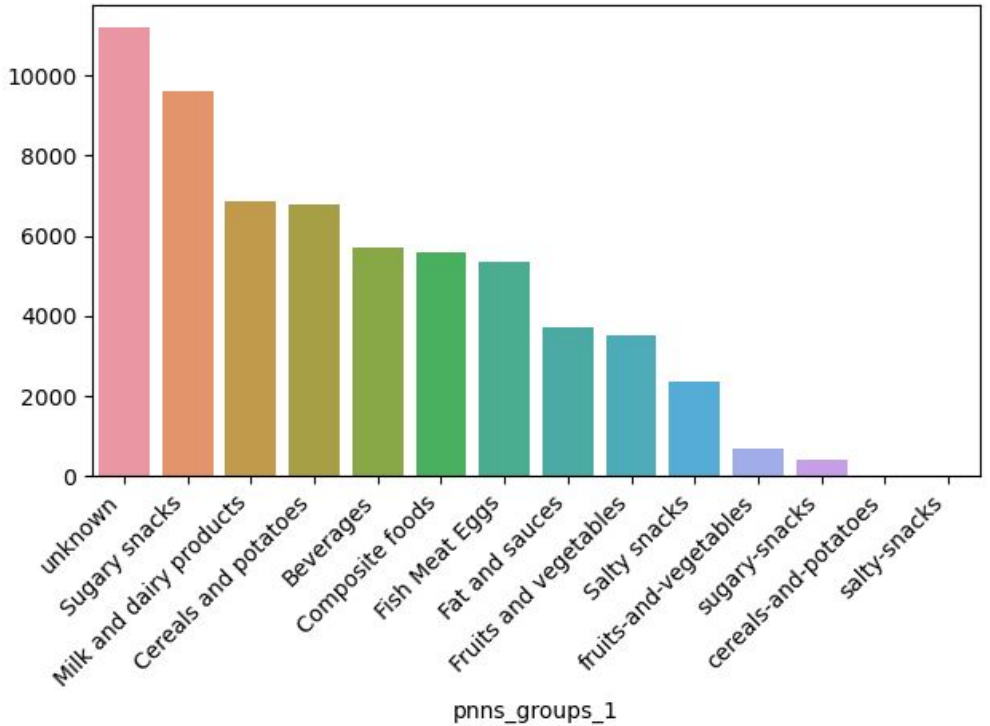
Filtrage et nettoyage des données :

- Suppression des lignes avec **'variables _100g' > 100**, à l'exception du variable energy_100g (en kj)
- Remplacement des valeurs négatives par 0 ⇒ 221133 lignes × 15 colonnes

12 variables numériques	3 catégorielles
energy_100g, saturated-fat_100g, trans-fat_100g, cholesterol_100g, carbohydrates_100g, fiber_100g, proteins_100g, salt_100g, vitamin-a_100g, vitamin-c_100g, calcium_100g, iron_100g.	product_name, pnns_groups_1, nutrition_grade_fr (cible)

Étape 1 : Nettoyez et filtrez des features et produits

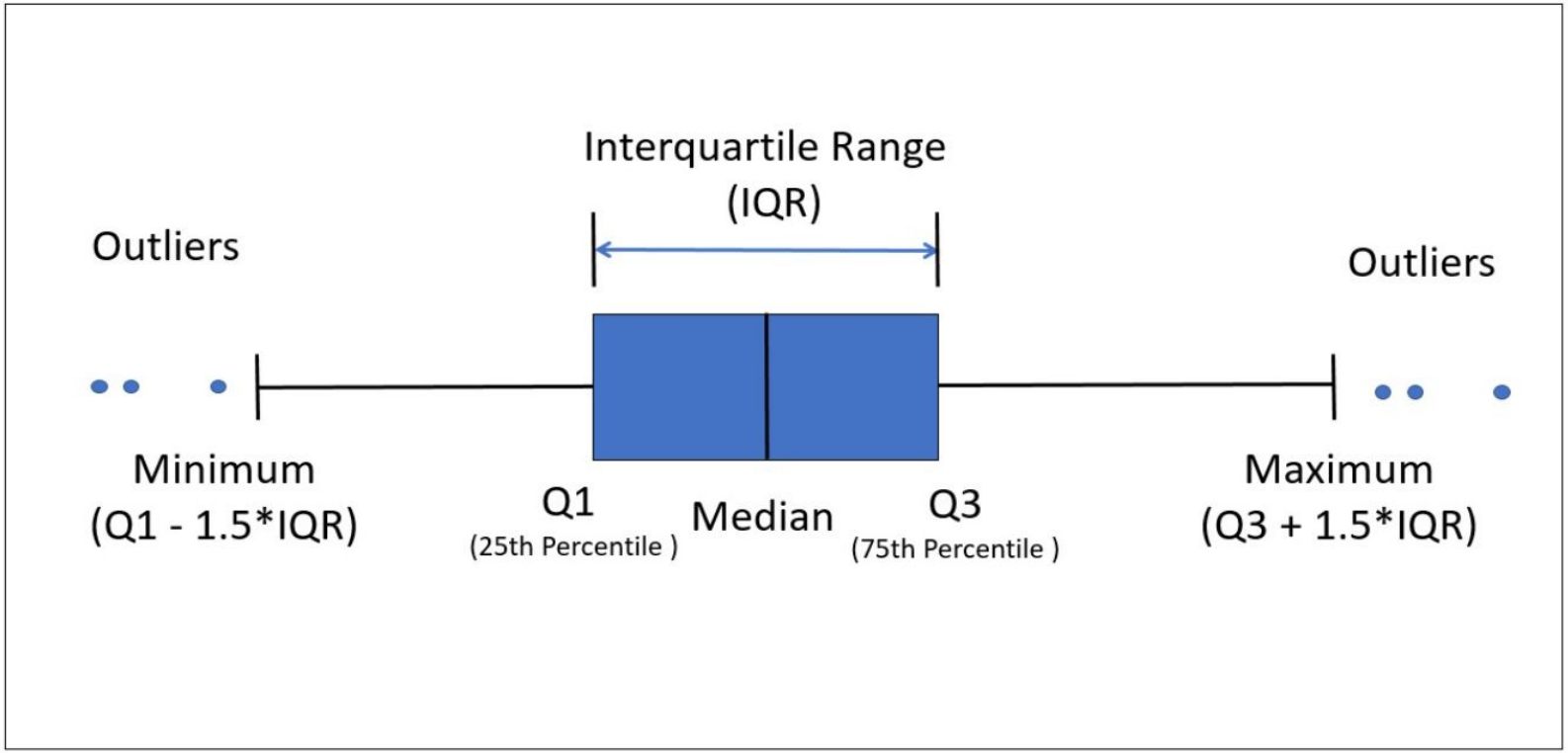




Étape 2 : Identifiez et traitez les valeurs aberrantes

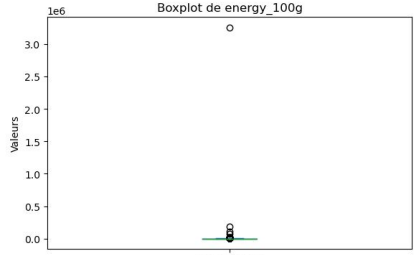
Étape 2 : Identifiez et traitez les valeurs aberrantes

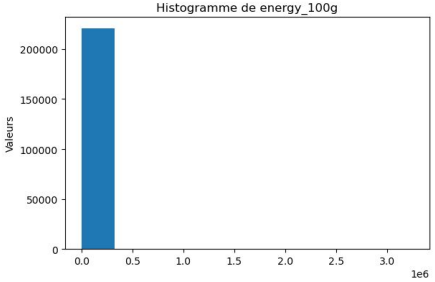
Méthode des plages interquartiles

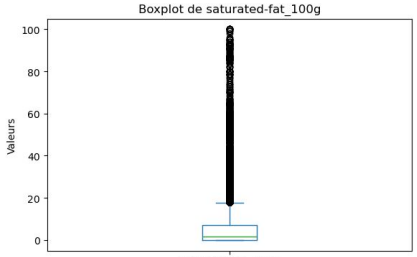


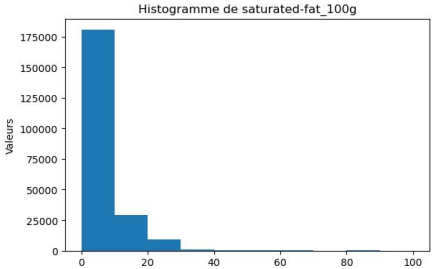
Étape 2 : Identifiez et traitez les valeurs aberrantes

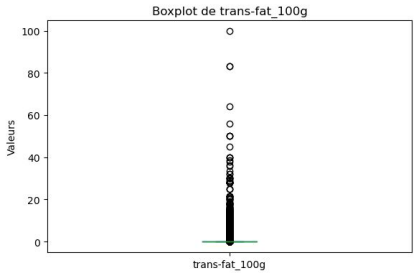
Avant suppression des valeurs aberrantes

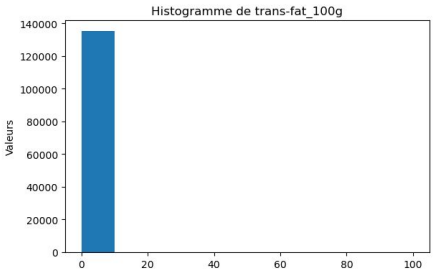




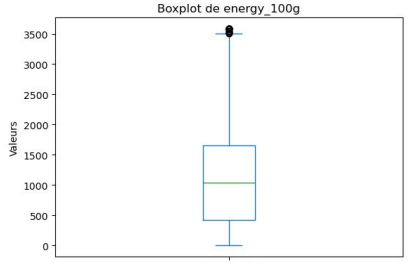


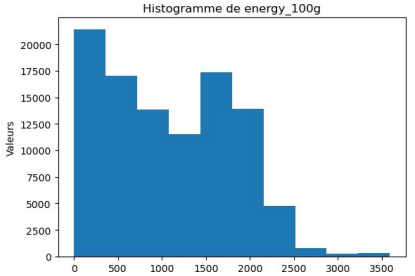


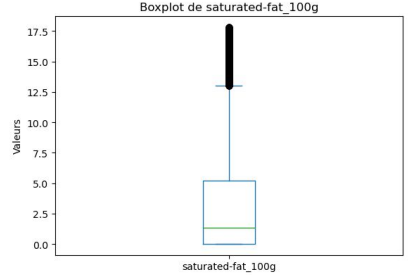


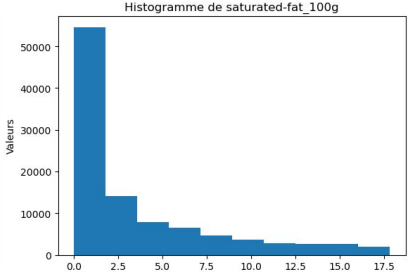


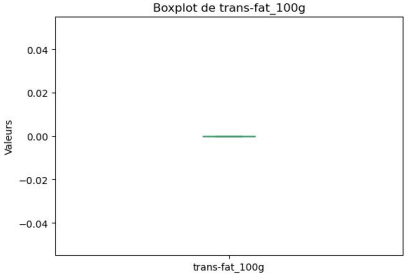
Après suppression des valeurs aberrantes

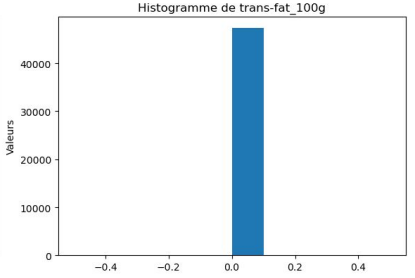






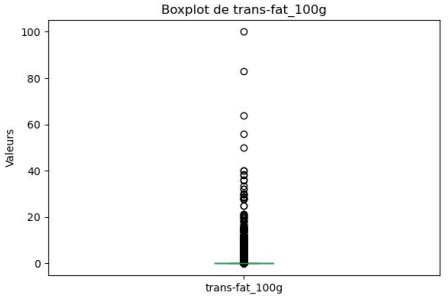


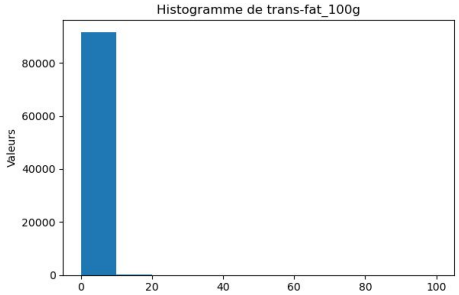


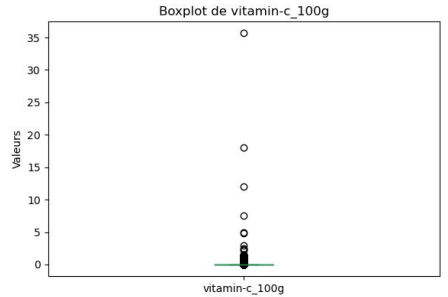


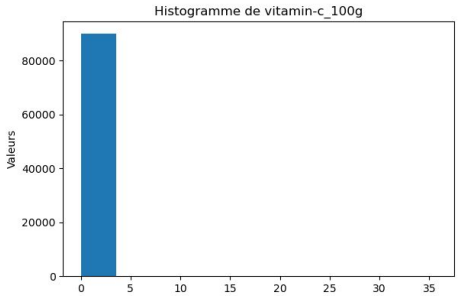
Étape 2 : Identifiez et traitez les valeurs aberrantes

Avant suppression des valeurs aberrantes

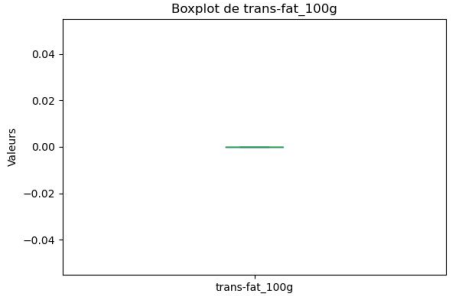


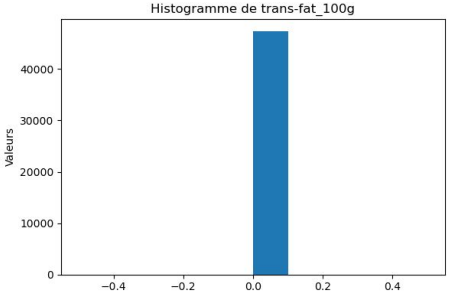


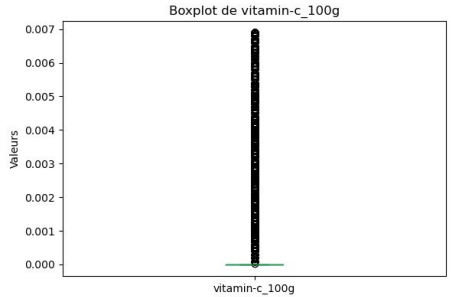


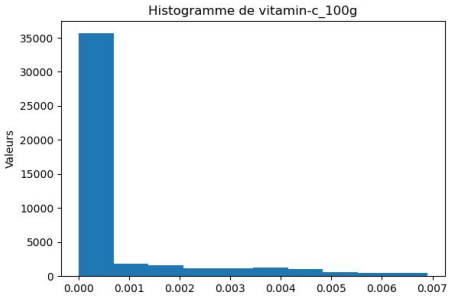


Après suppression des valeurs aberrantes







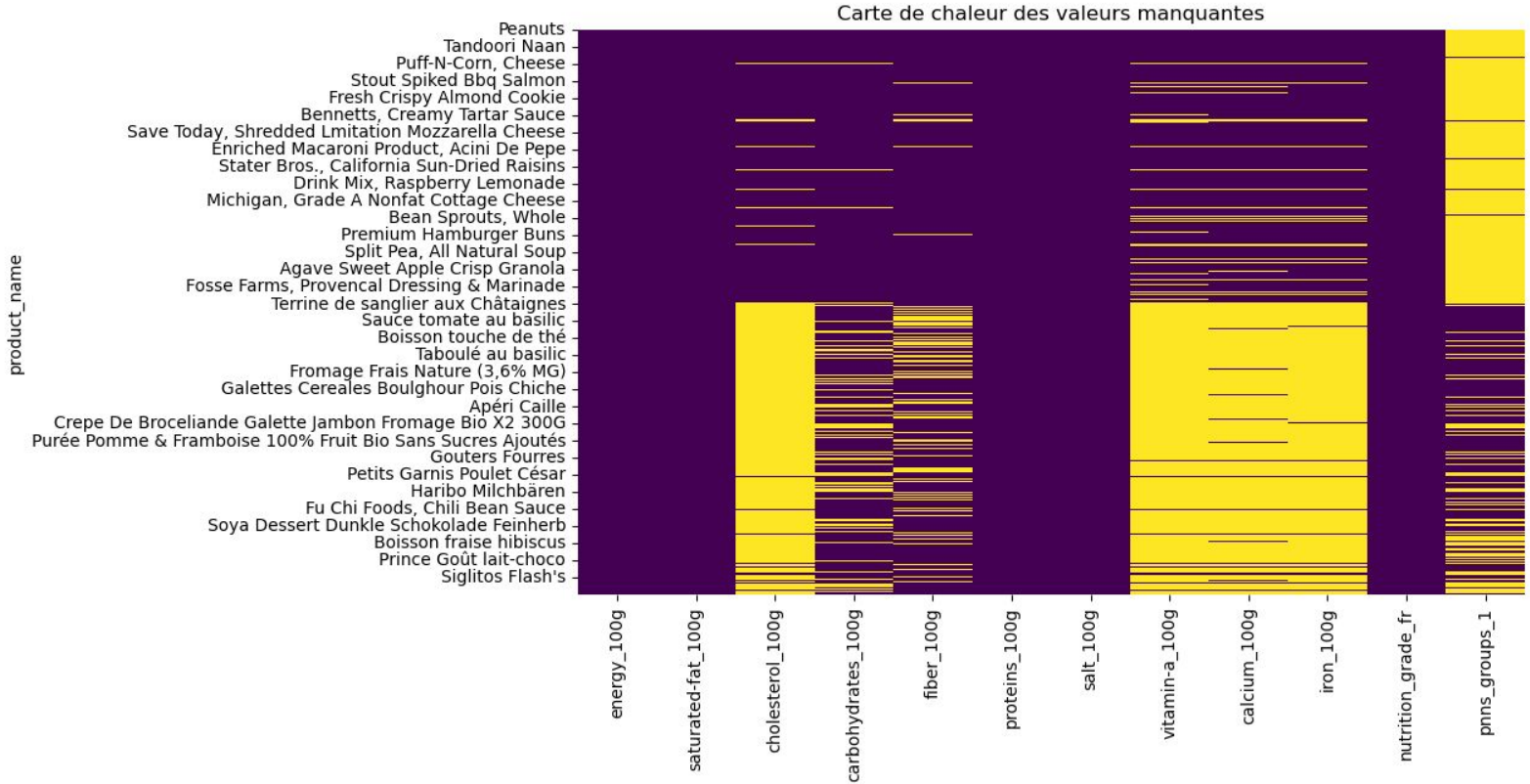


⇒ Suppression de la variable nulle 'trans-fat_100g', et la variable 'vitamin-c_100g' presque nulle avec que des valeurs aberrantes ⇒ 10 variables restantes

Étape 3 : Identifiez et traitez les valeurs manquantes

Étape 3 : Identifiez et traitez les valeurs manquantes

Identification des valeurs manquantes



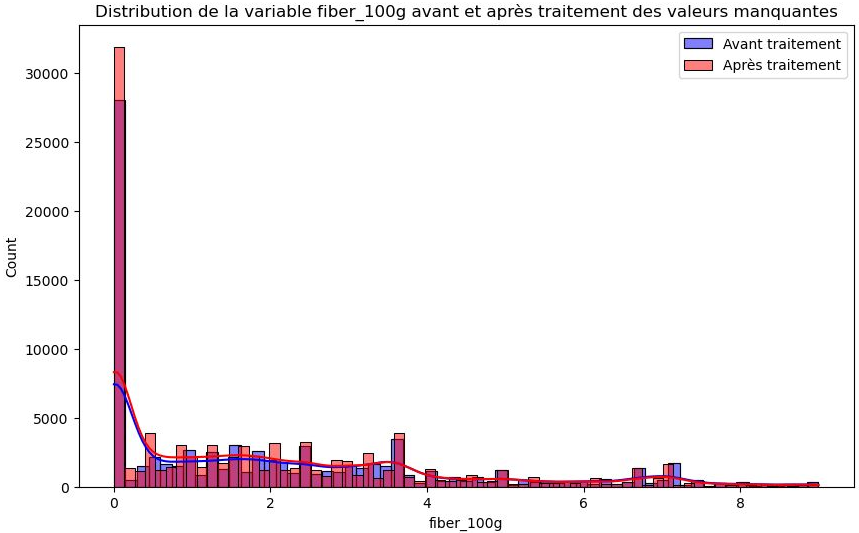
Étape 3 : Identifiez et traitez les valeurs manquantes

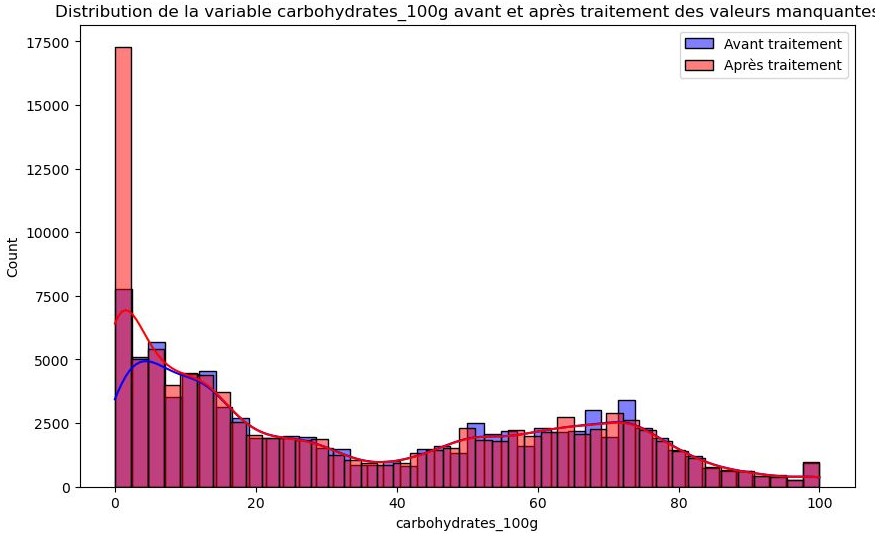
Traitement des valeurs manquantes

Méthode 1	Méthode 2	Méthode 3	Méthode 4
Après suppression des valeurs aberrantes : 101444 lignes, 12 colonnes Suppression des lignes avec que des NaN ⇒ 101281 lignes × 12 colonnes	Suppression des lignes avec que des 0 ⇒ 101080 lignes × 12 colonnes	Imputation des valeurs manquantes qui ont une catégorie pnns_groups_1 avec la méthode KNNI: 101080 lignes × 12 colonnes	Remplissage des valeurs manquantes restantes par 0 ⇒ 101080 lignes × 12 colonnes

Étape 3 : Identifiez et traitez les valeurs manquantes

Visualisation des variables avant et après traitement des valeurs manquantes





Étape 4 : Effectuez les analyses uni-variée et bi-variée

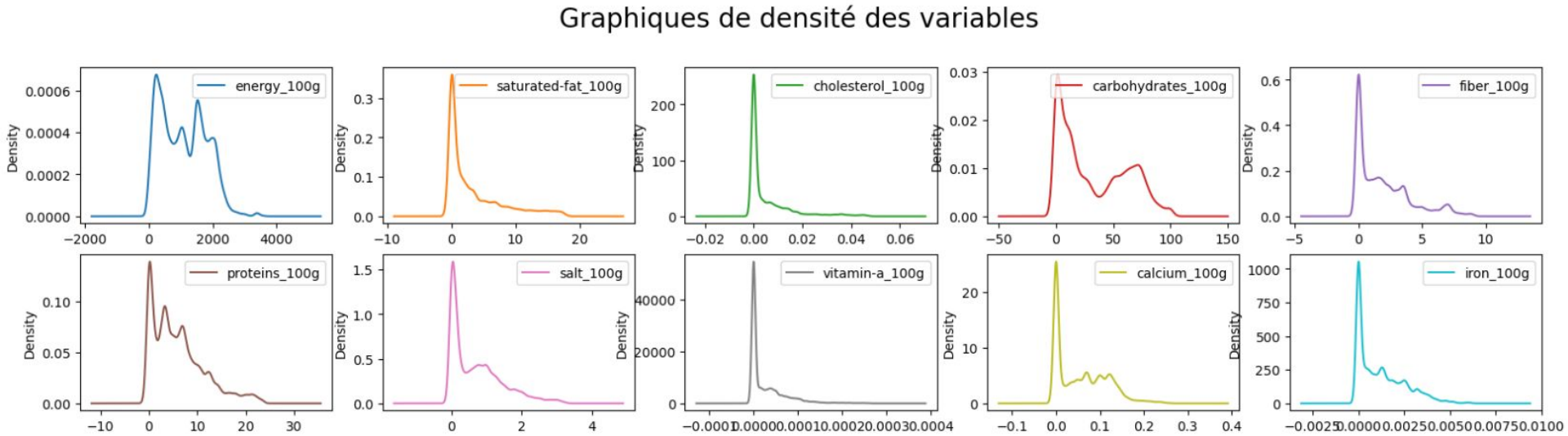
Étape 4 : Effectuez les analyses uni-variée et bi-variée

Analyse uni-variée : statistiques descriptives

	energy_100g	saturated-fat_100g	cholesterol_100g	carbohydrates_100g	fiber_100g	proteins_100g	salt_100g	vitamin-a_100g	calcium_100g	iron_100g
count	101080.000000	101080.000000	101080.000000	101080.000000	101080.000000	101080.000000	101080.000000	101080.000000	101080.000000	101080.000000
mean	1088.438248	3.385542	0.005141	32.909335	1.931544	6.028504	0.707040	0.000025	0.057136	0.001196
std	710.960931	4.455492	0.009197	29.384576	2.131785	5.327604	0.725666	0.000041	0.057313	0.001279
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	423.000000	0.000000	0.000000	5.830000	0.000000	1.670000	0.091440	0.000000	0.000000	0.000000
50%	1046.000000	1.400000	0.000000	23.000000	1.340000	5.000000	0.508000	0.000000	0.048000	0.000840
75%	1653.000000	5.200000	0.007000	60.470000	3.100000	8.850000	1.115060	0.000040	0.102445	0.002030
max	3590.000000	17.840000	0.047000	100.000000	9.000000	23.640000	3.251200	0.000259	0.261000	0.006260

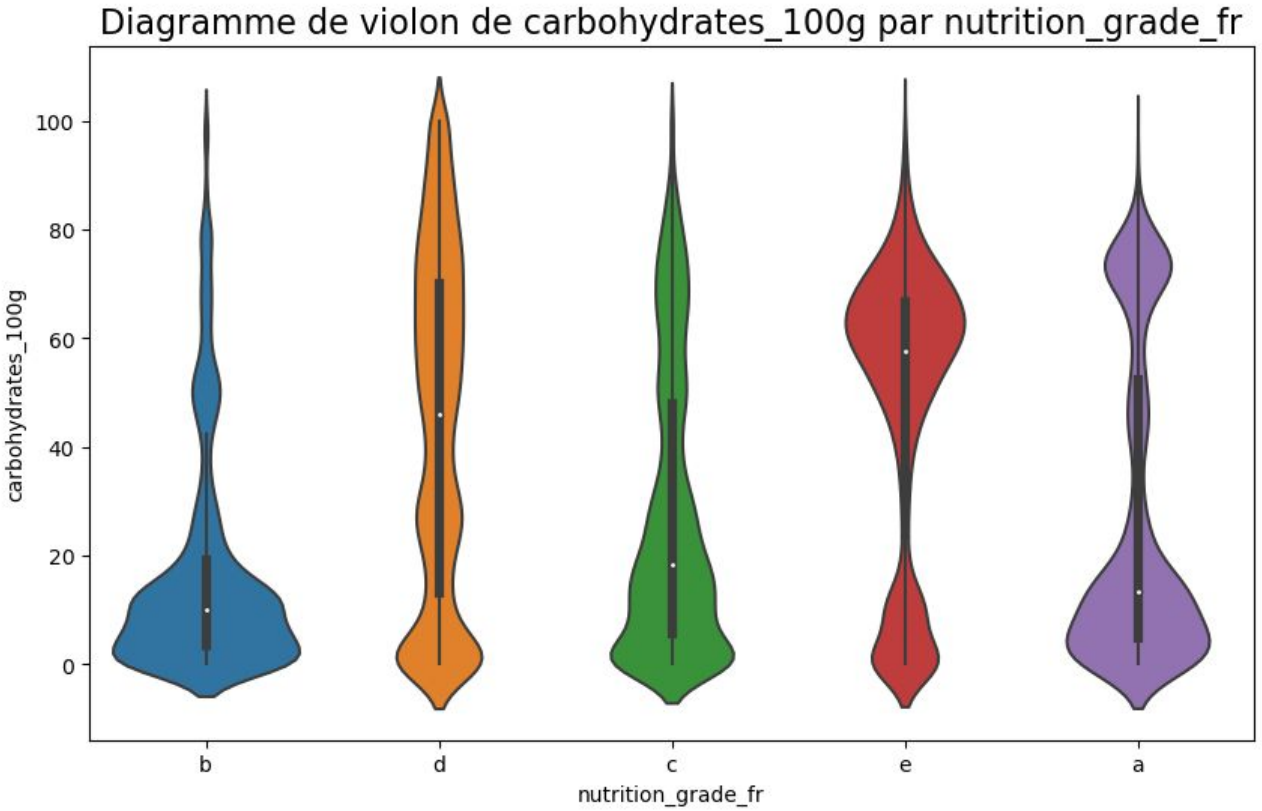
Étape 4 : Effectuez les analyses uni-variée et bi-variée

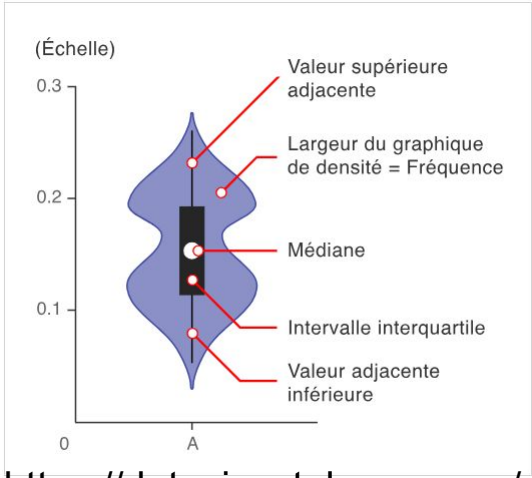
Visualisation uni-variée : graphiques de densité



Étape 4 : Effectuez les analyses uni-variée et bi-variée

Visualisation bi-variée : diagrammes de violon



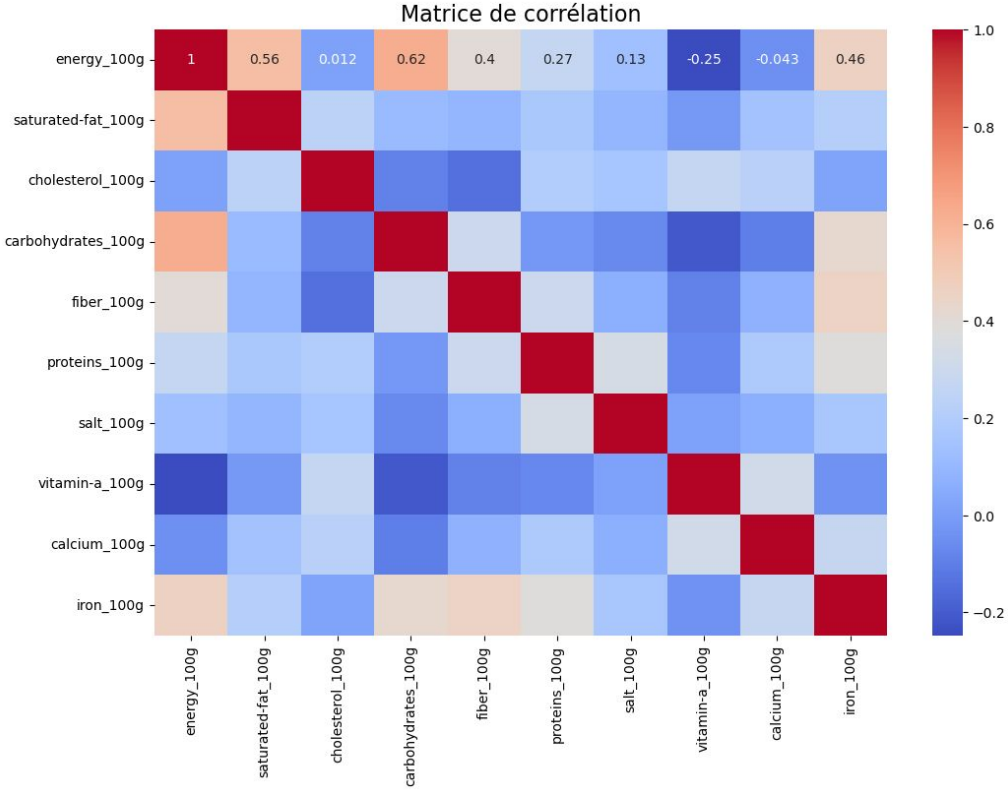


https://datavizcatalogue.com/

Étape 5 : Réalisez une analyse multi-variée

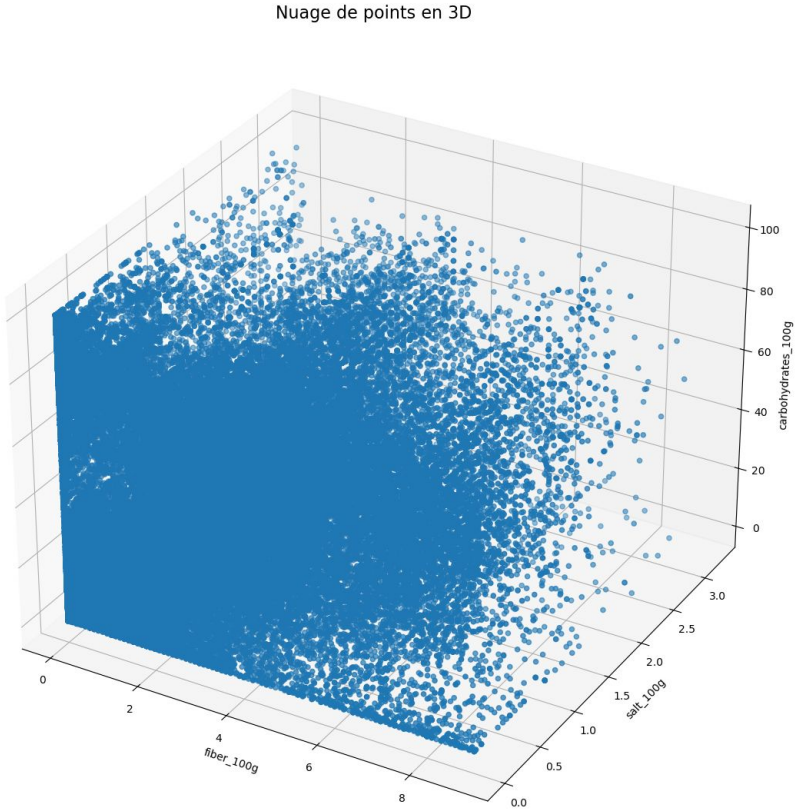
Étape 5 : Réalisez une analyse multi-variée

Visualisation avancée: carte de chaleur



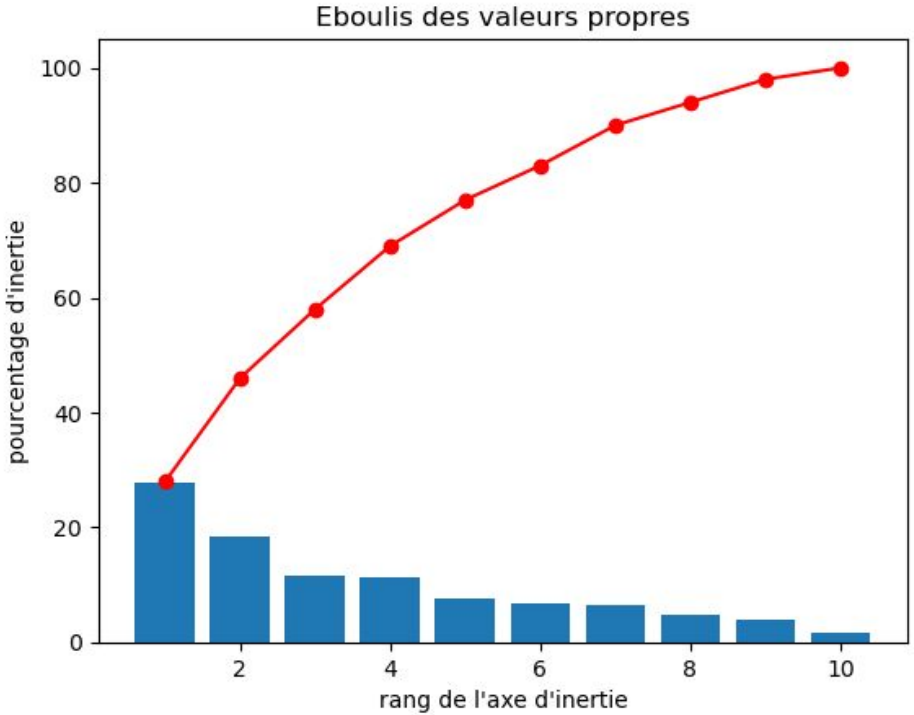
Étape 5 : Réalisez une analyse multi-variée

Visualisation avancée: graphique en 3D



Étape 5 : Réalisez une analyse multi-variée

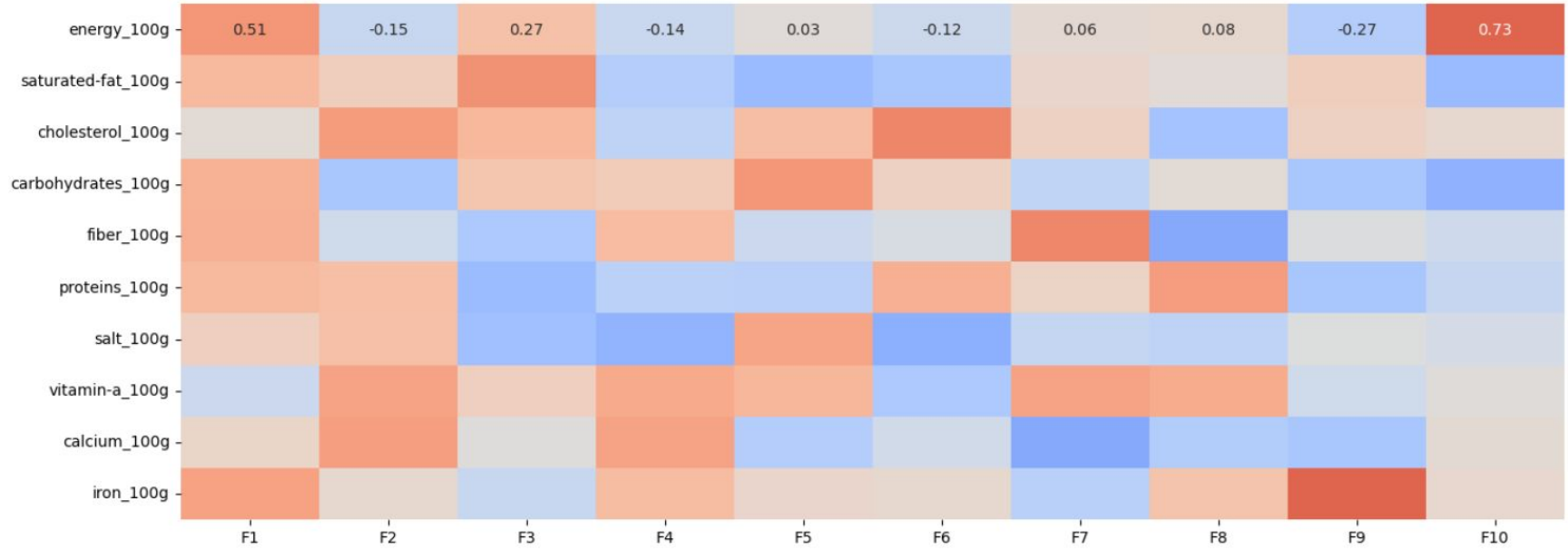
Analyse statistique : analyse en composante principale (ACP)



Étape 5 : Réalisez une analyse multi-variée

Analyse statistique : analyse en composante principale (ACP)

Carte de chaleur des variables en fonction des 10 composantes



Étape 5 : Réalisez une analyse multi-variée

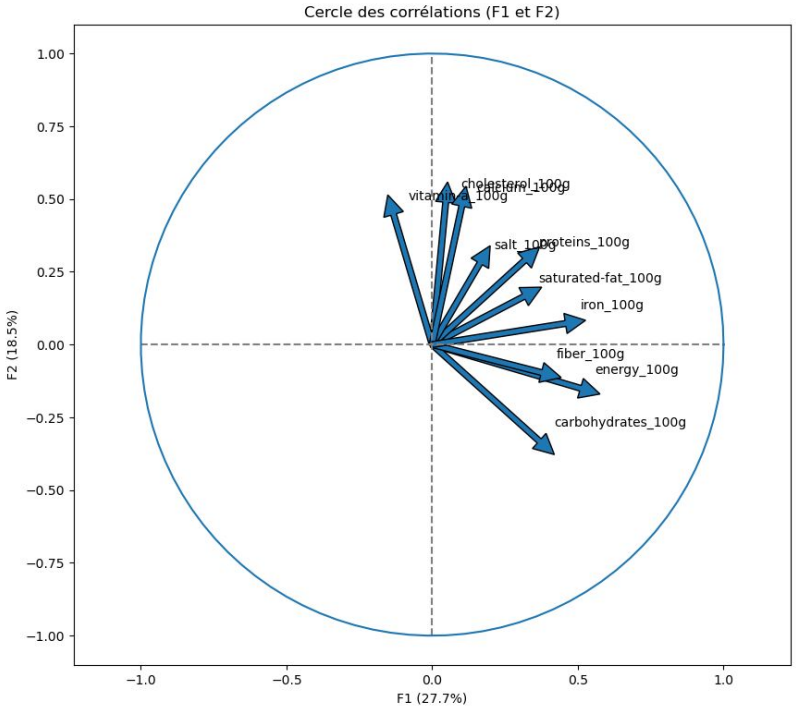
Analyse statistique : analyse en composante principale (ACP)

- **energy_100g** et **iron_100g** : les plus corrélées positivement à **F1**

- **vitamin-a_100g**, **cholesterol_100g** et **calcium_100g** : les plus corrélées positivement à **F2**

- **energy_100g** et **fiber_100g** : très corrélées entre eux

- **energy_100g** et **cholesterol_100g** : indépendantes



Étape 5 : Réalisez une analyse multi-variée

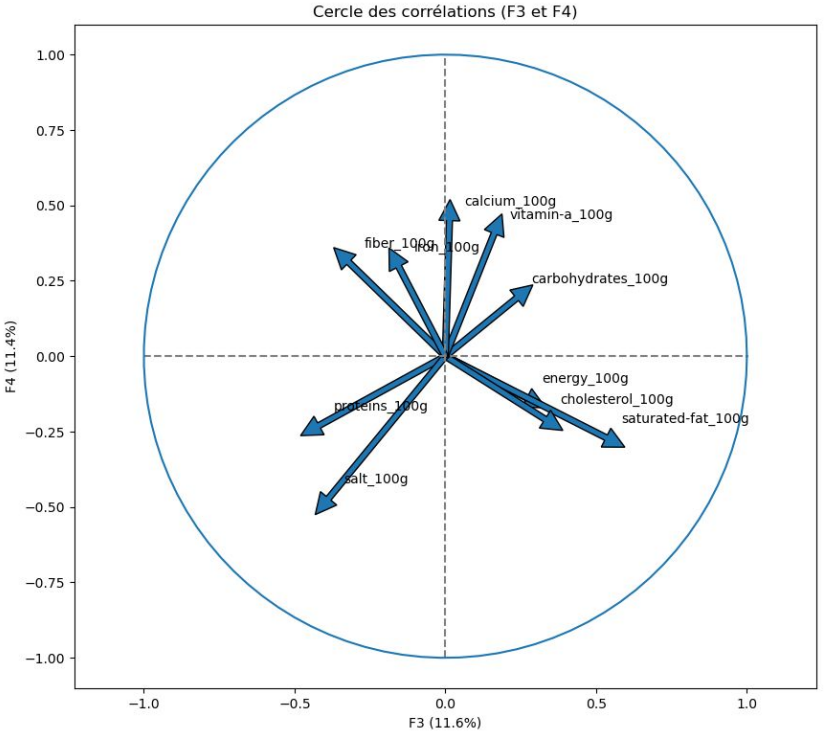
Analyse statistique : analyse en composante principale (ACP)

- **saturated-fat_100g** : la plus corrélée positivement à **F3**

- **salt_100g** : la plus corrélée négativement à **F4**

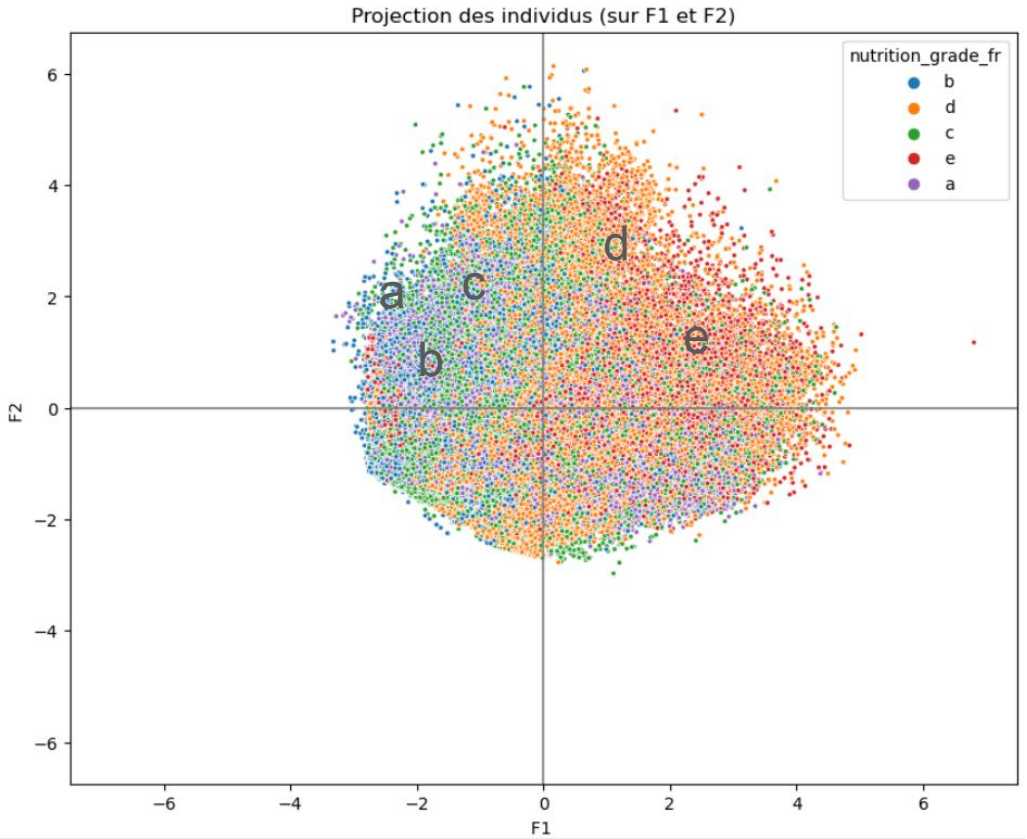
- **calcium_100g** est à **0.5** sur **F4** et à **0** sur **F3** : bien corrélée avec **F4** et elle n'a pas d'impact sur **F3**

- **vitamin-a_100g**, **energy_100g** : indépendantes



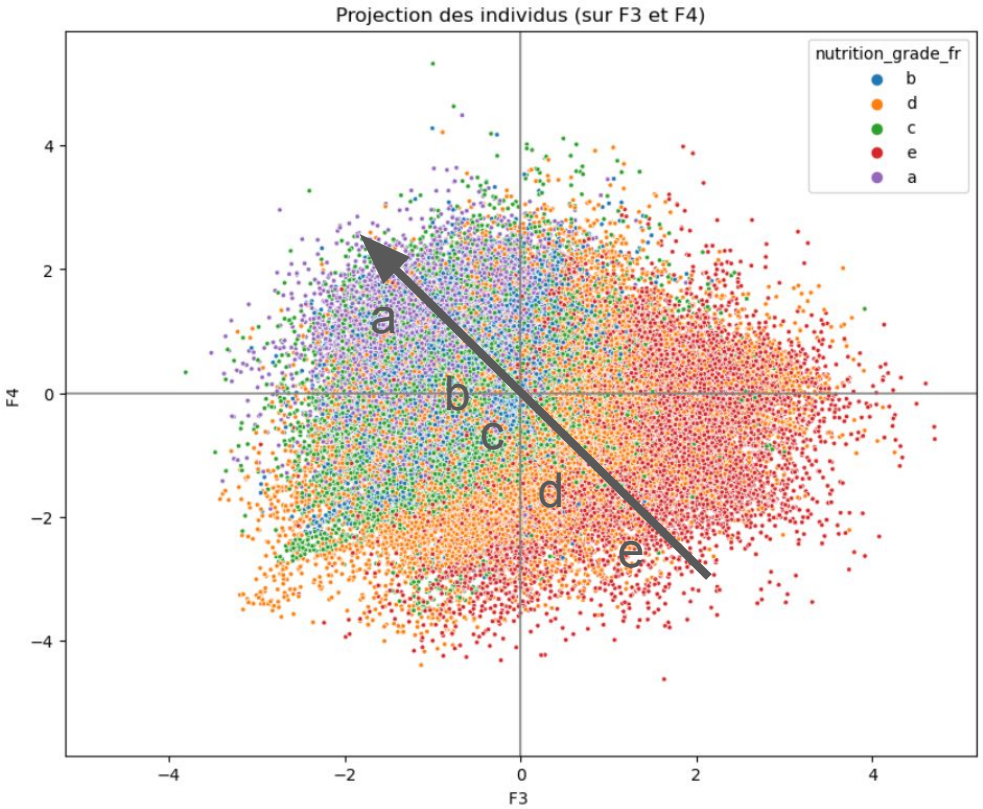
Étape 5 : Réalisez une analyse multi-variée

Analyse statistique : analyse en composante principale (ACP)



Étape 5 : Réalisez une analyse multi-variée

Analyse statistique : analyse en composante principale (ACP)

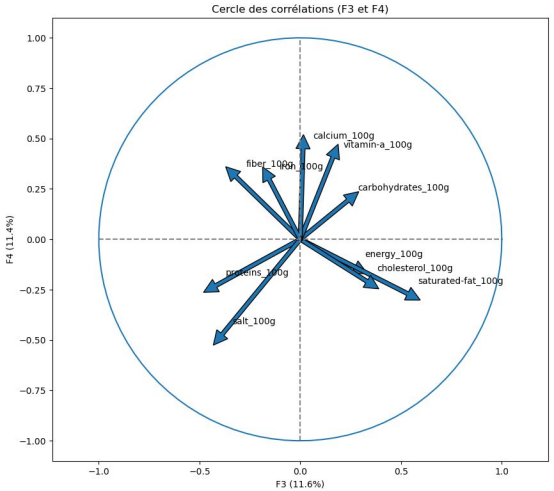


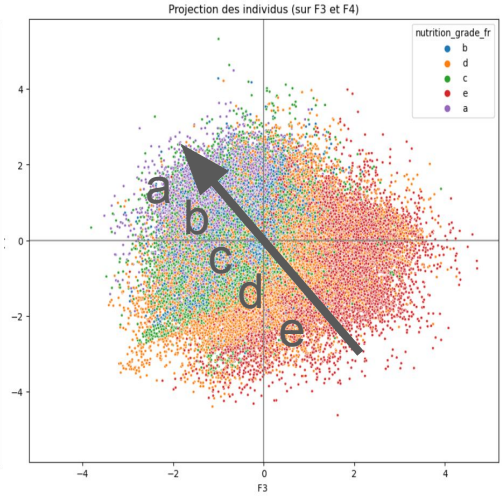
Conclusion

Conclusion :

Le nettoyage et l'exploration des données montre que ces 10 variables peuvent contribuer pour prédire un Nutri-Score automatique:

- saturated-fat_100g
- energy_100g
- cholesterol_100g
- proteins_100g
- salt_100g
- vitamin-a_100g
- carbohydrates_100g
- calcium_100g
- iron_100g
- fiber_100g





Conclusion :

Le Règlement Général sur la Protection des Données (**RGPD**) est basé sur **cinq grands principes** qui encadrent la protection des données personnelles au sein de l'Union Européenne. Voici ces principes :

1. Licéité, loyauté et transparence

2. Limitation des finalités

3. Minimisation des données

4. Exactitude

5. Limitation de la conservation