OPENCLASSROOMS

Projet n°2 - Préparer les données pour un organisme de santé publique

Mounira ABDERRAHMANI

Soutenance réalisée devant

Sabrine BENDIMERAD

Sommaire

I. Objectif de la démarche

Qui ?

L'agence

Santé publique France

Quoi ?

Améliorer sa base de données

OPEN FOOD FACTS

Comment ?

Application de suggestion de valeurs manquantes

OPENCLASSROOMS

open FOOD facts

Les données :

Principes :

- Licéité, loyauté et transparence
- Limitation des finalités
- Minimisation des données
- Limitation de la conservation
- Intégrité et confidentialité

Aperçu général

2 965 170 observations et 203 variables

4 catégories d'informations :

CODE, URL, CREATOR

PACKAGING_TAGS, LABELS_TAGS, COUNTRIES_TAGS

INGREDIENTS_TEXT, ALLERGENS, ADDITIVES_N

Valeurs nutritionnelles moyennes	Pour 100 g	Pour ce plat
Énergie	177 Kcal 734 Kj	512 Kcal 2127 Kj
Protéines	5,3 g	15,4 g
Glucides	8,9 g dont sucres : 2,6 g	25,8 g dont sucres : 7,5 g

ENERGY_100G, FAT_100G, SUGARS_100G

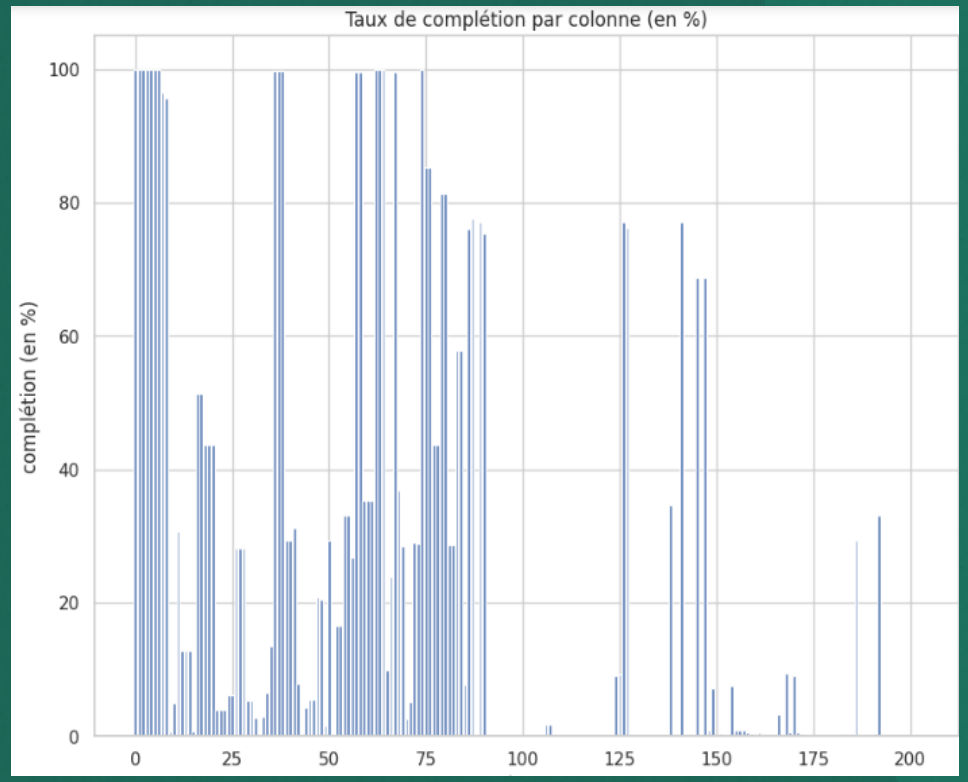
II. Nettoyage des données

Les étapes :

- Sélection des colonnes non vides
- Sélection des produits français
- Choix des variables à analyser
- Traitement des valeurs aberrantes
- Traitement des valeurs manquantes

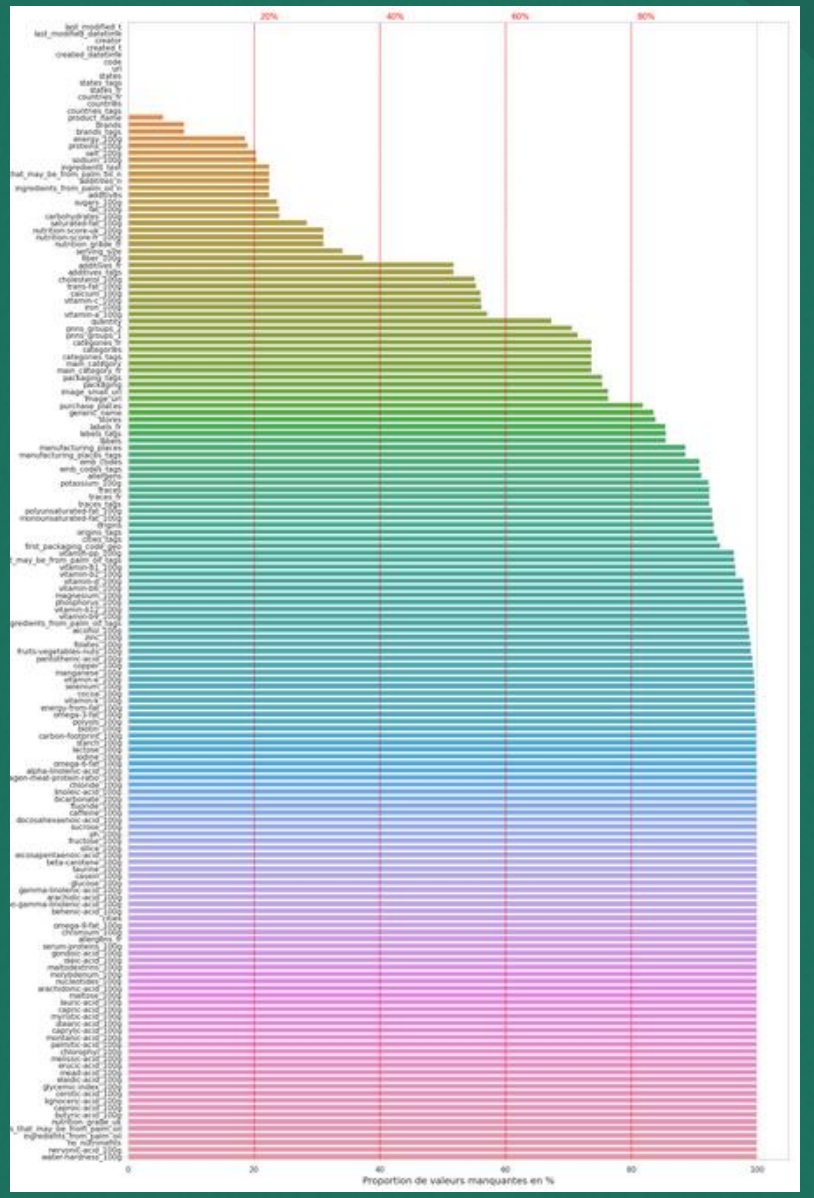
II. Nettoyage des données

1. Sélection des colonnes non vides :



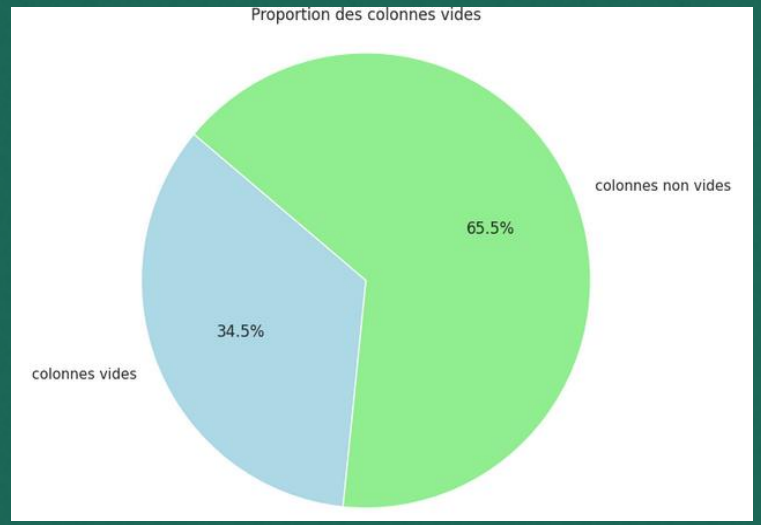
Taux de complétion des colonnes

Proportion de valeur manquantes par colonne



II. Nettoyage des données

1. Sélection des colonnes non vides :

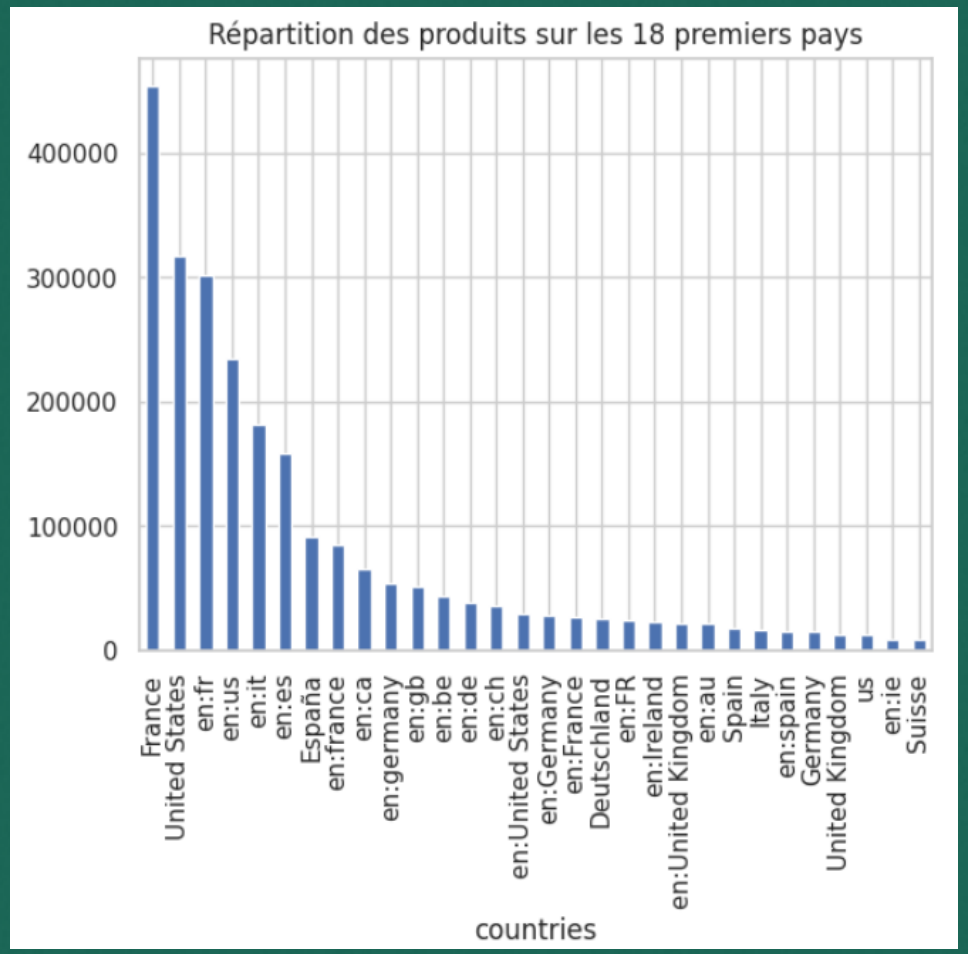


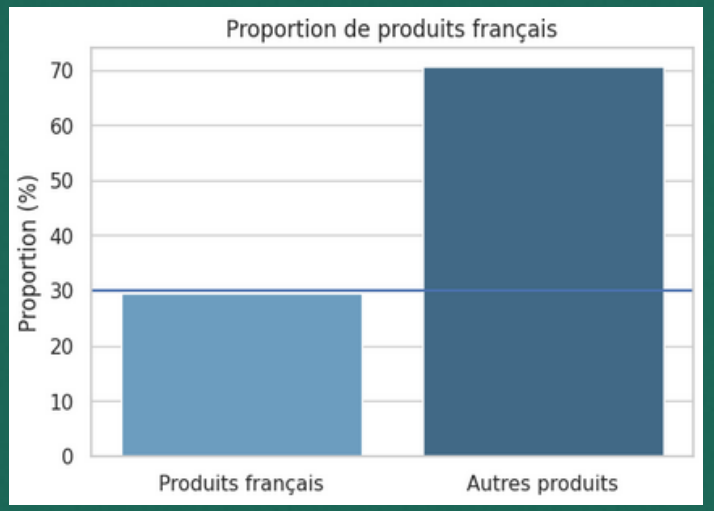
Suppression des 70 colonnes

203 colonnes → 133 colonnes

II. Nettoyage des données

2. Sélection des produits français :

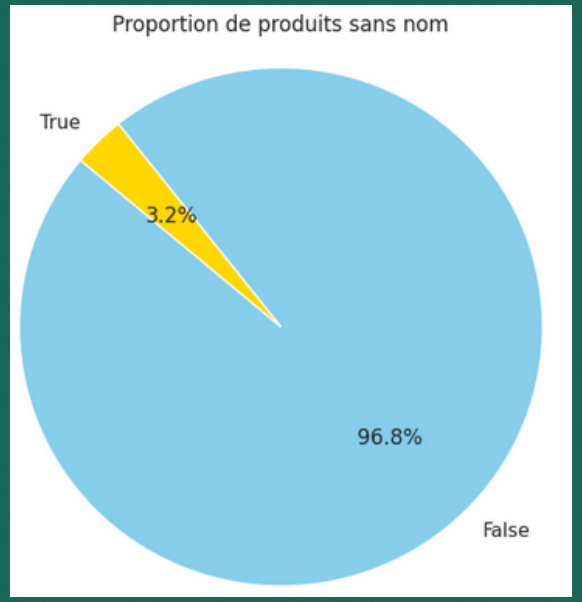




- Produits français : environs 30% des données
- 2 965 170 observations → 874 918 observations

II. Nettoyage des données

➢ Autre sélection : variable product_name



Suppression produit sans noms : 3.2 %

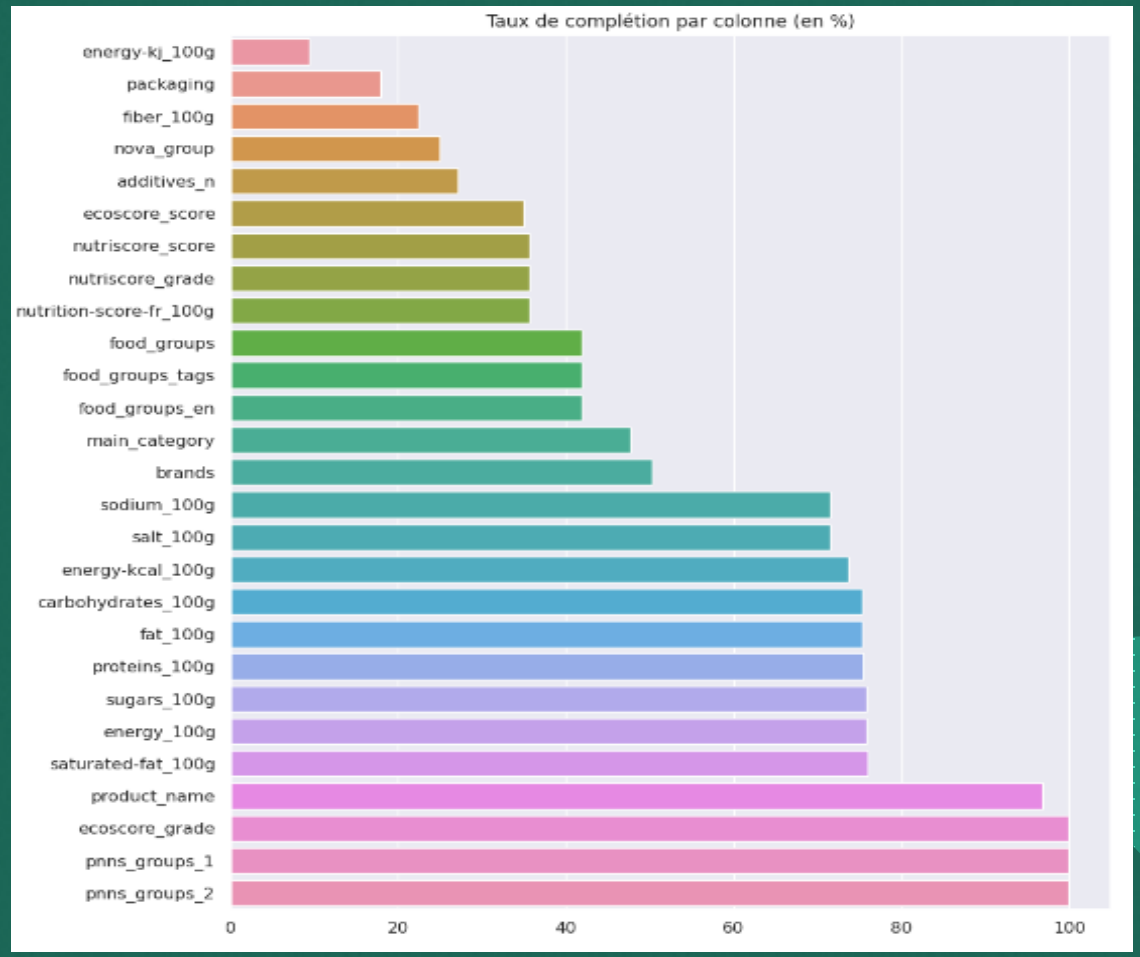
- 874 918 observations
- 846 810 observations

3. Choix des features :

➢ Sélection des colonnes pour l'analyse

133 colonnes

27 colonnes



4. Traitement des valeurs aberantes:

	energy_kj_100g	fat_100g	saturated_fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g
count	8.062300e+04	656434.000000	661933.000000	656258.000000	660995.000000	194009.000000	657880.000000
mean	1.075770e+03	14.278452	5.421087	26.533411	13.497671	3.098052	9.069336
std	3.925083e+03	17.485386	7.950142	27.543451	19.963295	5.841682	11.199501
min	0.000000e+00	0.000000	0.000000	0.000000	-0.100000	0.000000	0.000000
25%	4.200000e+02	1.000000	0.200000	2.300000	0.600000	0.000000	1.500000
50%	9.500000e+02	8.200000	2.000000	13.500000	3.200000	1.500000	6.300000
75%	1.569000e+03	22.400000	8.000000	51.000000	19.000000	3.600000	13.000000
max	1.094259e+06	820.000000	900.000000	966.000000	390.000000	256.000000	2706.000000

Quantité d'énergie exprimée en million

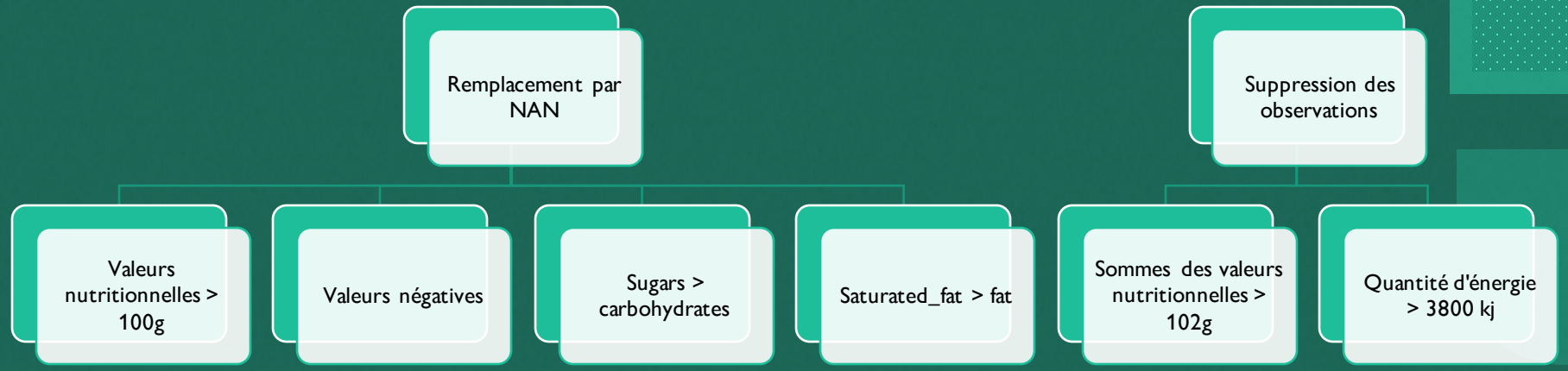
Valeurs nutritionnelles pour 100g > 100g

Quantités exprimées en nombres négatifs

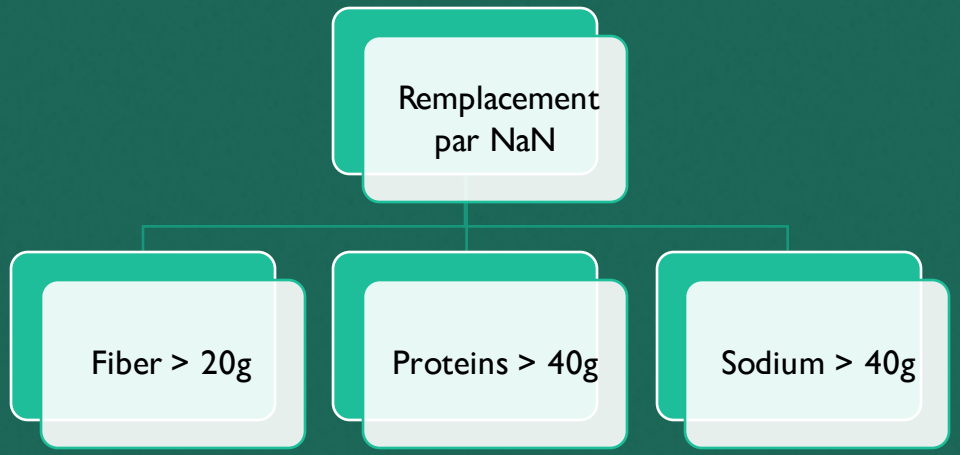
II. Nettoyage des données

4. Traitement des valeurs aberantes :

a. Valeurs aberrantes



b. Valeurs atypiques

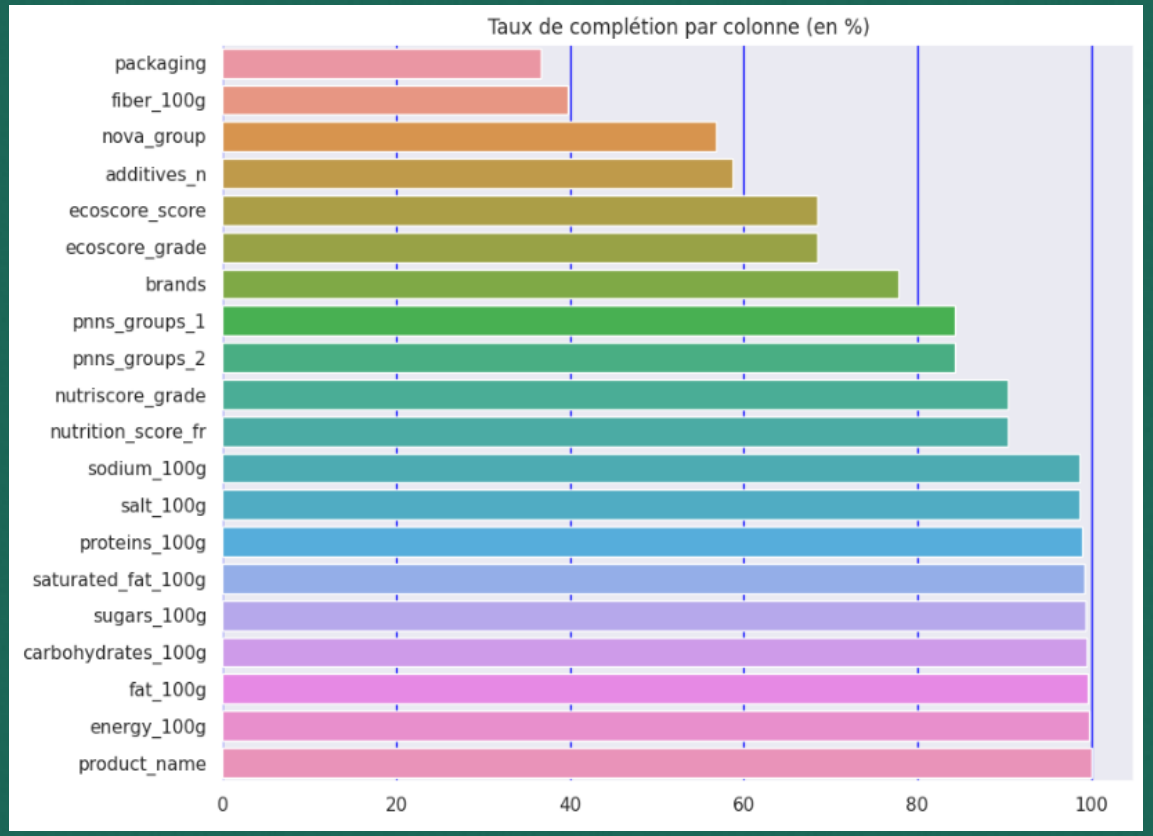


- 20 variables
- 685 340 observations

5. Traitement des valeurs manquantes : les lignes

Suppresion des observations ayant :

- ✓ Toutes les variables nutritionnelles manquantes
- ✓ Plus de 10 variables manquantes
- ✓ Des variables catégorielles manquantes (pnns_groups_1 - pnns_groups_2 - nova_group - nutriscore_grade - ecoscore_grade)



- 330 945 observations
- 20 variables

Taux de complétion à la fin des traitements

5. Traitement des valeurs manquantes : imputation

✓ Méthodes d'imputation sur les variables nutritionnelles :

- La médiane
- La médiane par catégorie pnns_groups_1
- Knn Imputer
- Iterative Imputer

✓ Méthode d'imputation ciblée :

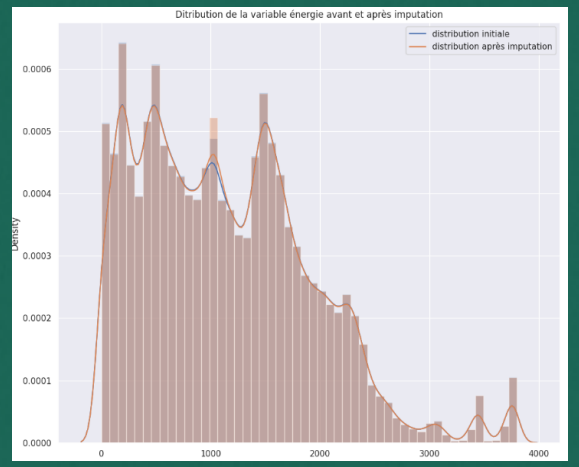
- Imputation par 0 : fiber_100g, proteins_100g
- Arbre de décision : nova_group

II. Nettoyage des données

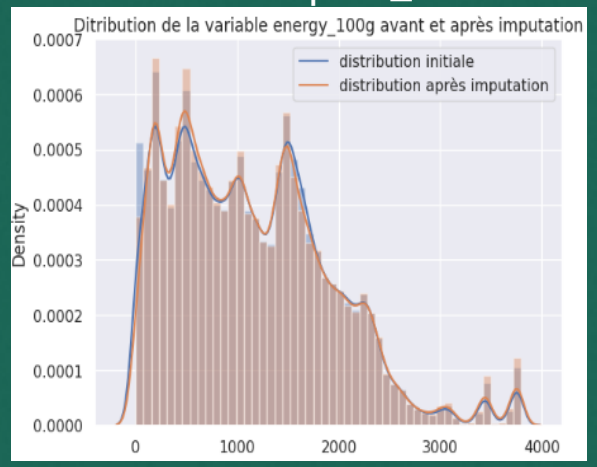
5. Traitement des valeurs manquantes :

Energy_100g

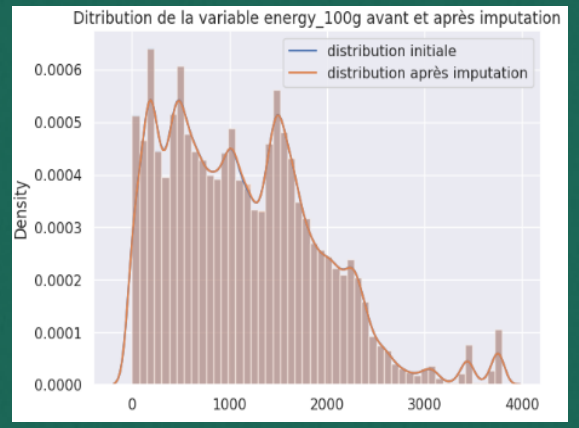
Médiane



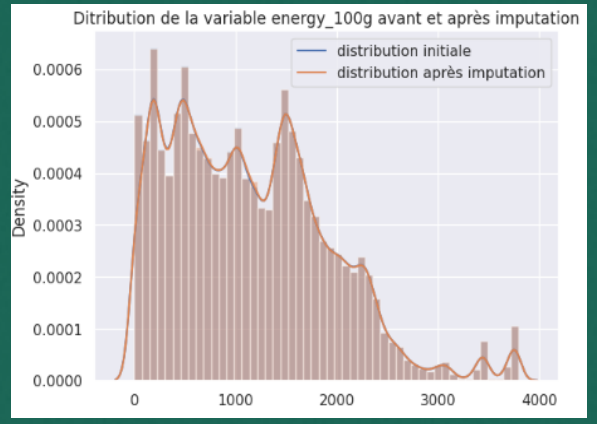
Médiane pnns_1



Knn (k=2)

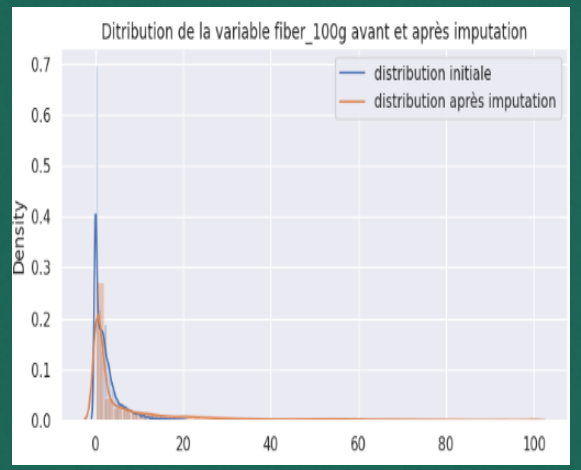


Iterative Imputer

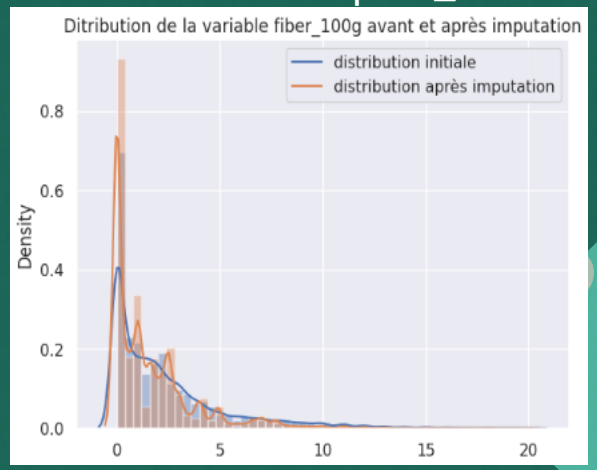


Fiber_100g

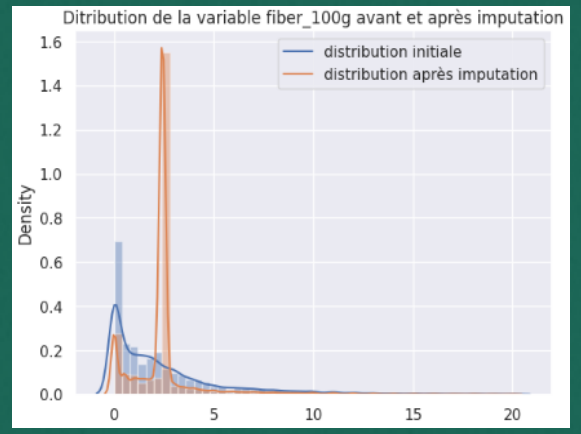
Médiane



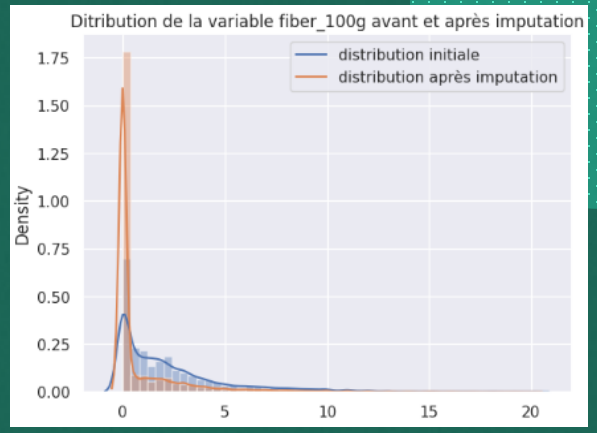
Médiane pnns_1



Knn (k=2)



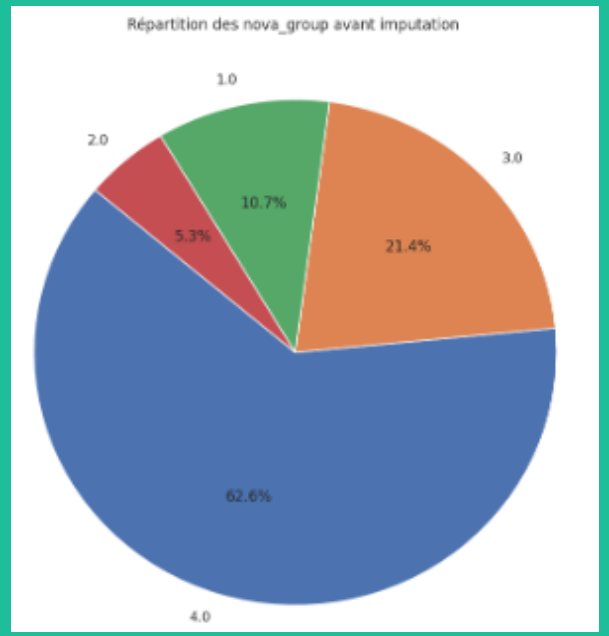
Imupation par 0

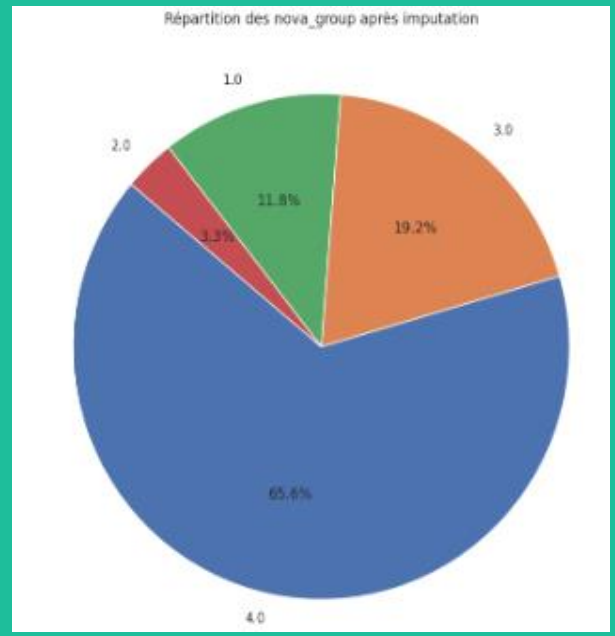


5. Traitement des valeurs manquantes :

Arbre de décision : Nova_group

variables caractéristiques : pnns_groups_2 et nutriscore_grade

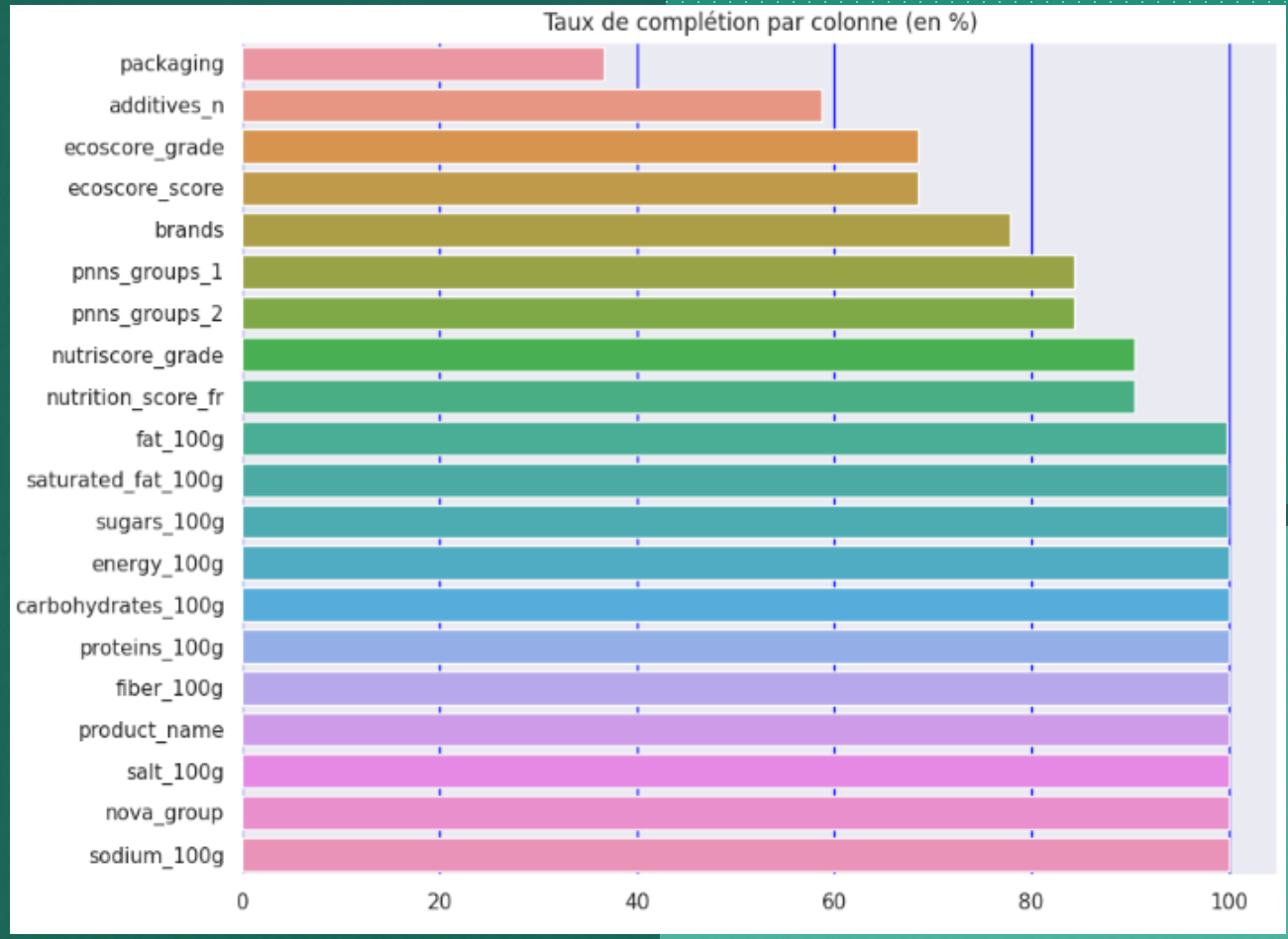




II. Nettoyage des données

5. Traitement des valeurs manquantes :

- ✓ Imputation de :
  - Variables nutritionnelles
  - Nutrition_score
  - Nova_groupe
- ✓ Test de wilcoxon
- ✓ Vérification des valeurs aberrantes après imputation



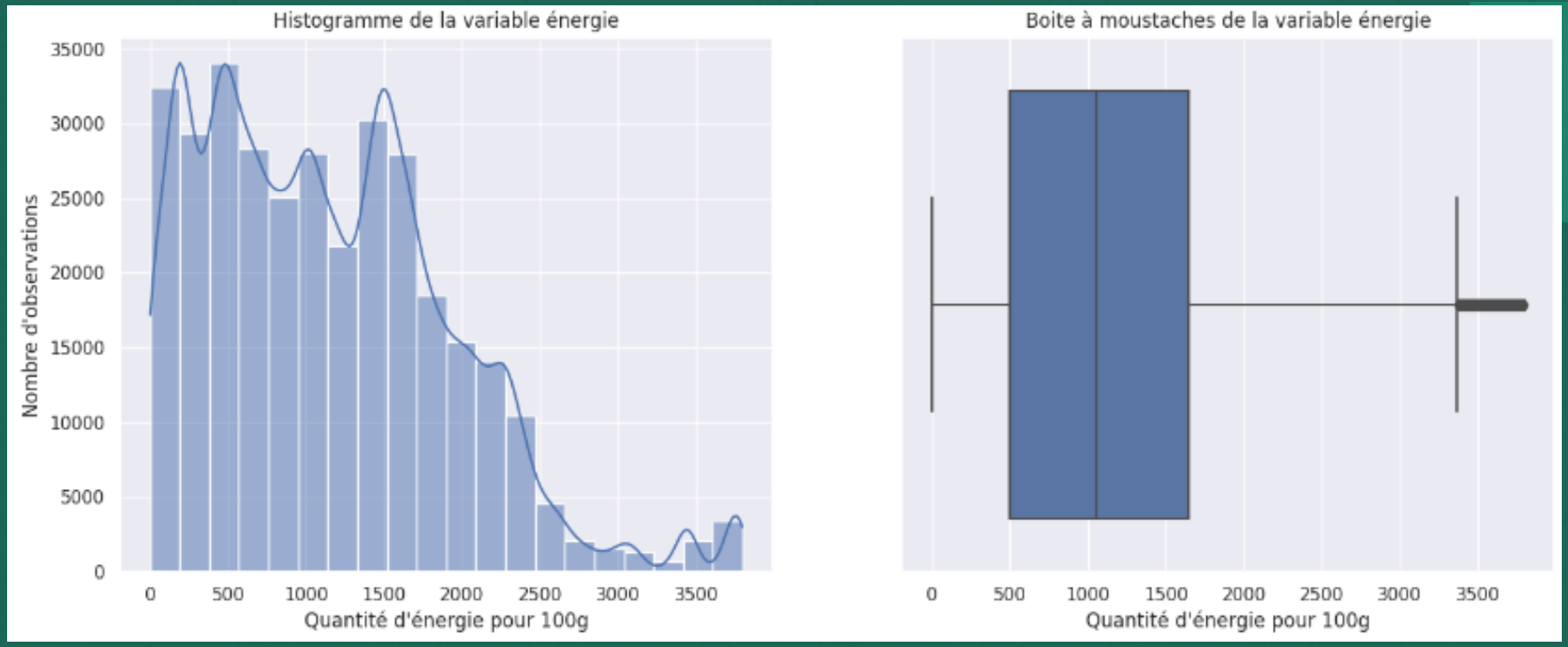
III. Analyses statistiques :

- Analyse univariée
- Analyse bivariée
- Analyse multivariée

III. Analyses statistiques

1. Analyse univariée : variables quantitatives continues

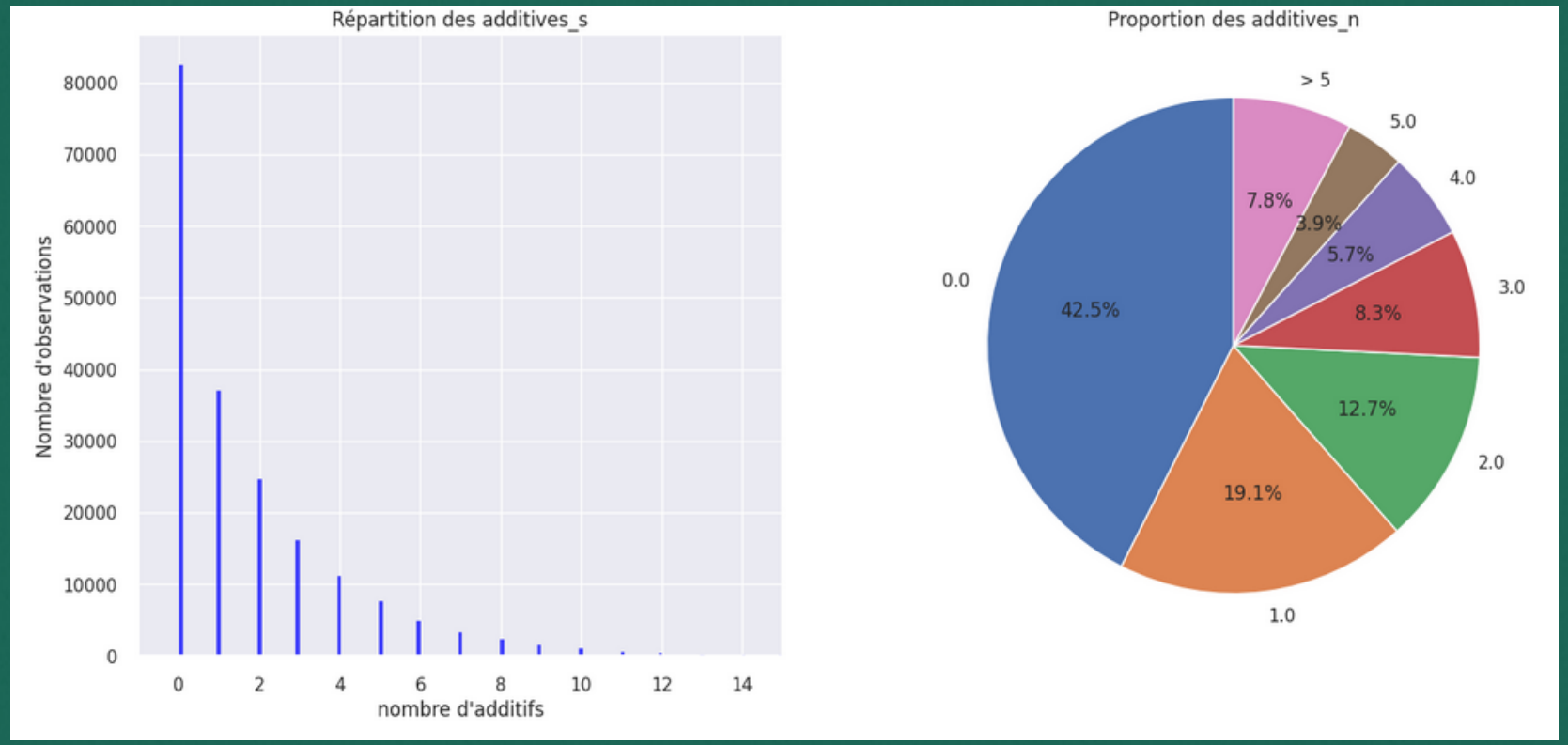
Energy_100g



III. Analyses statistiques

1. Analyse univariée : variables quantitatives discrètes

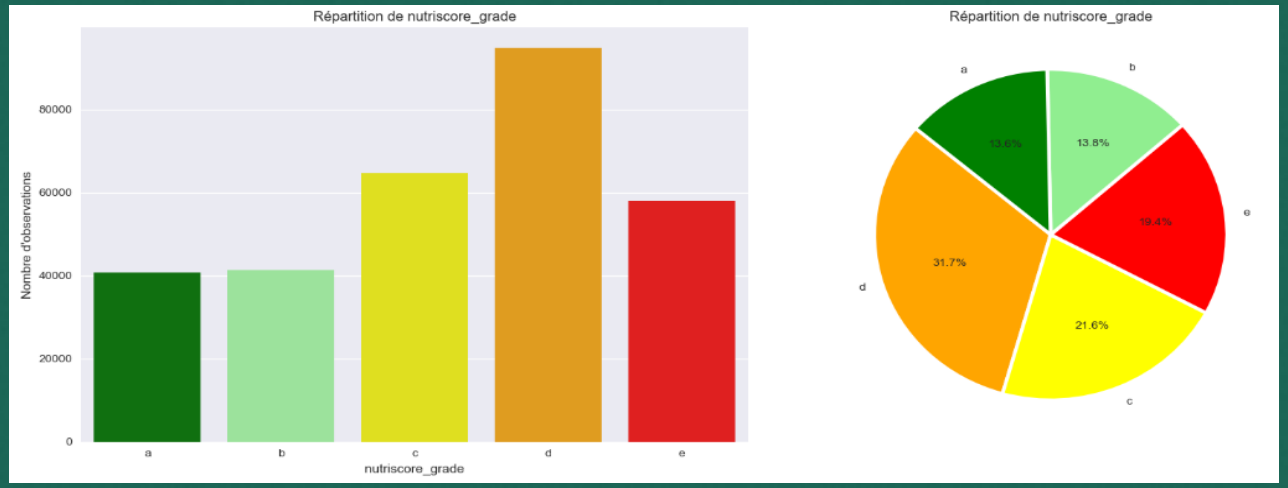
Additives_n



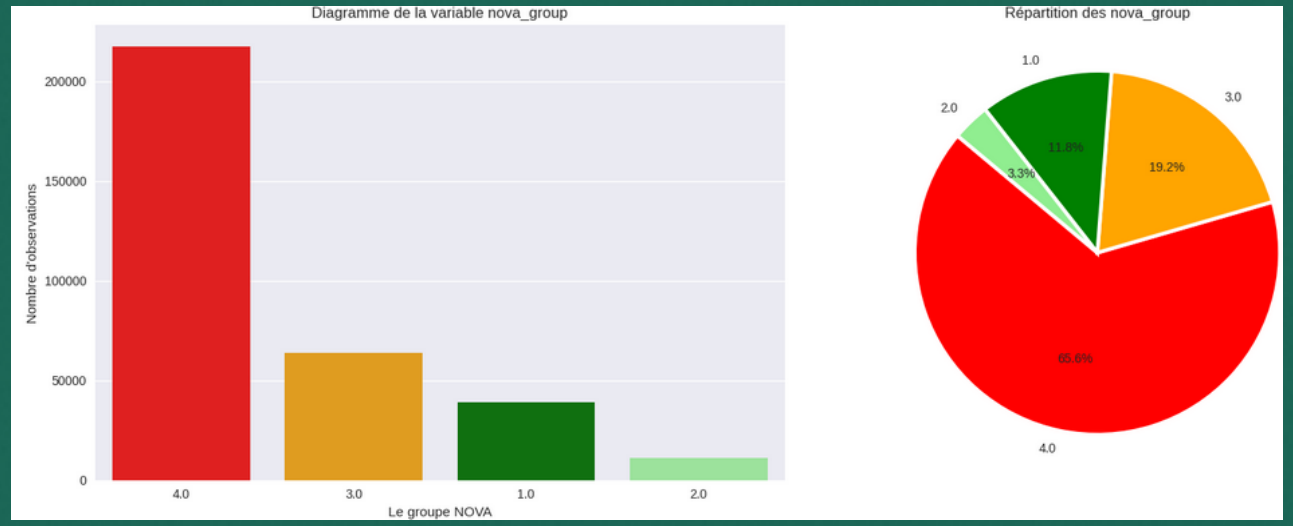
III. Analyses statistiques

1. Analyse univariée : variables qualitatives ordinales

Nutriscore_grade



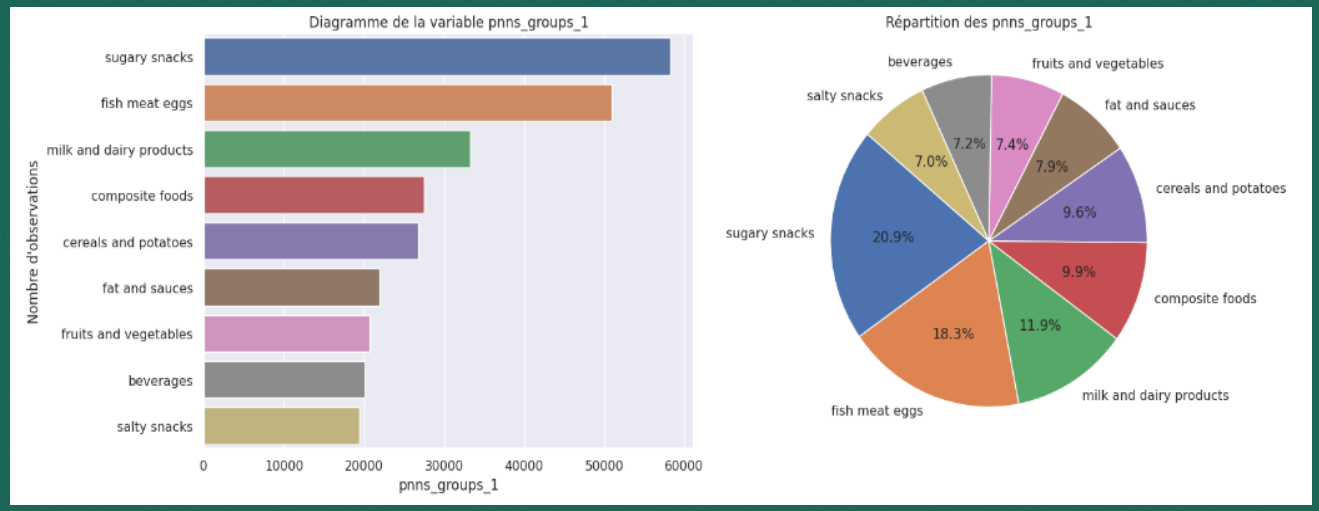
Nova_group



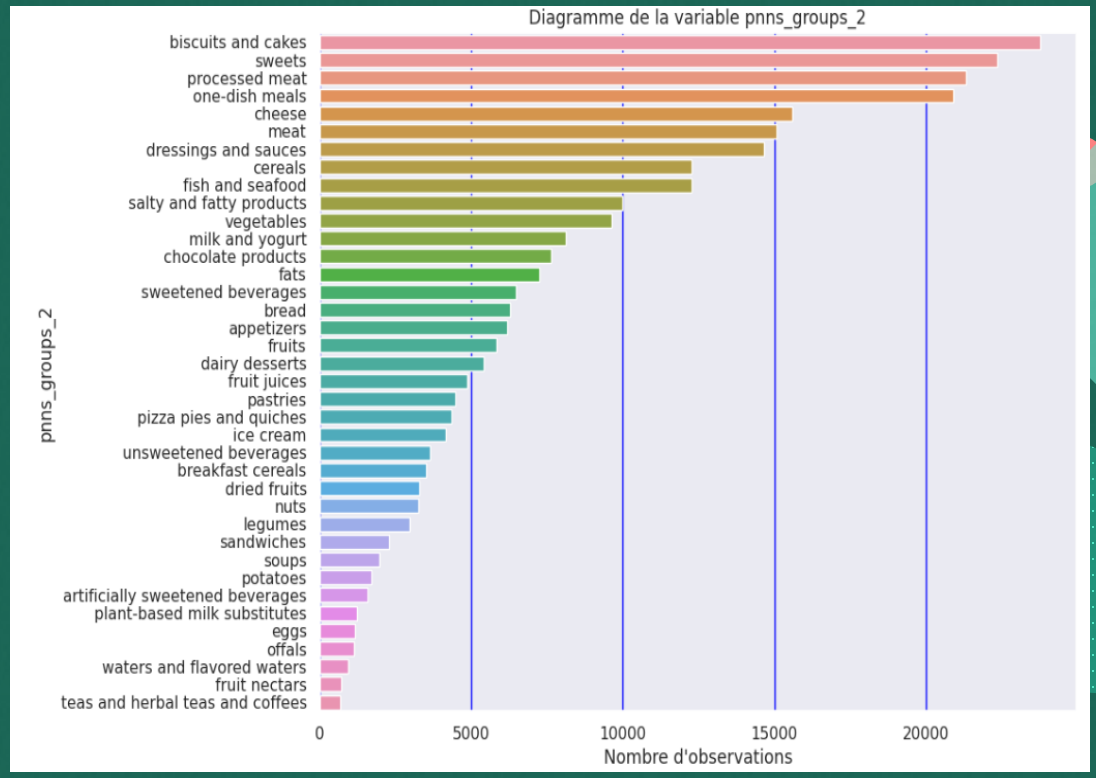
III. Analyses statistiques

1. Analyse univariée : variables qualitatives nominales

Pnns_groups_1

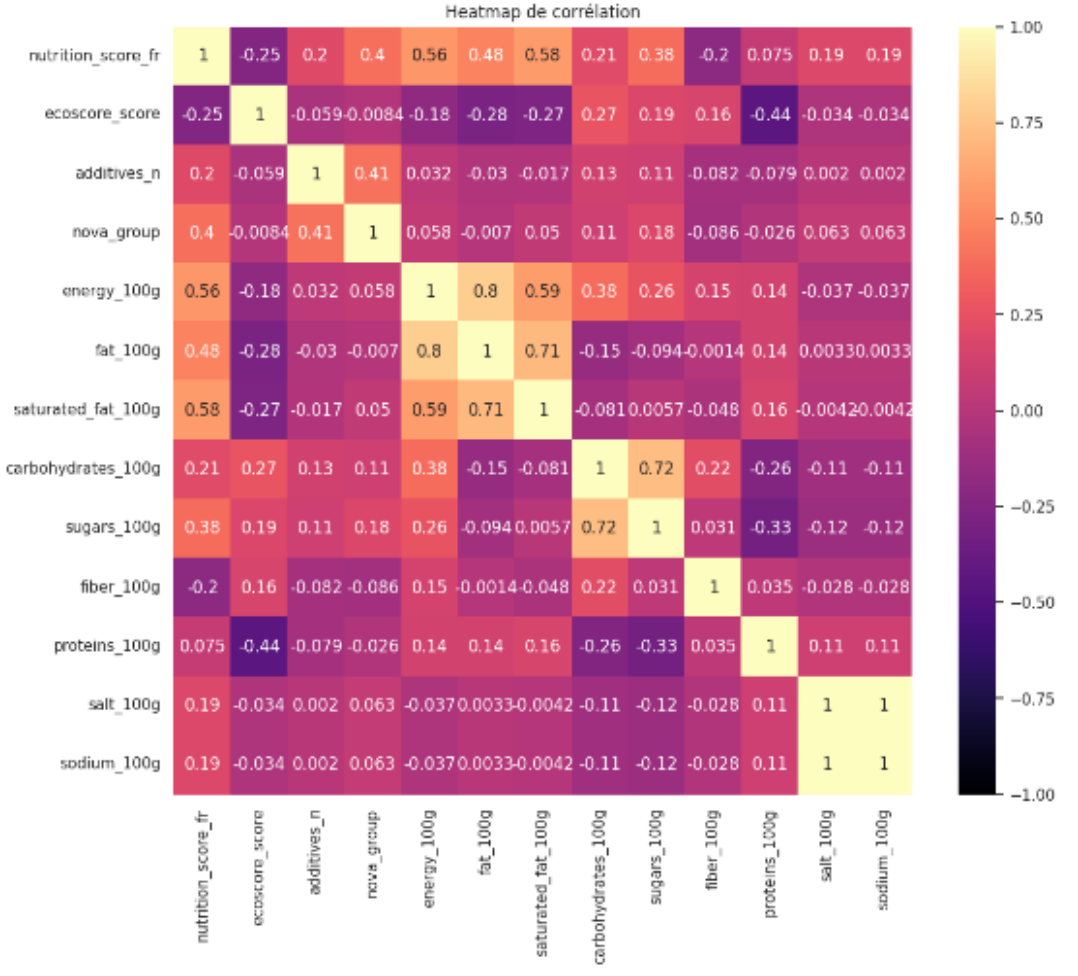


Pnns_groups_2



III. Analyses statistiques

2. Analyse bivariée : heatmap de corrélation



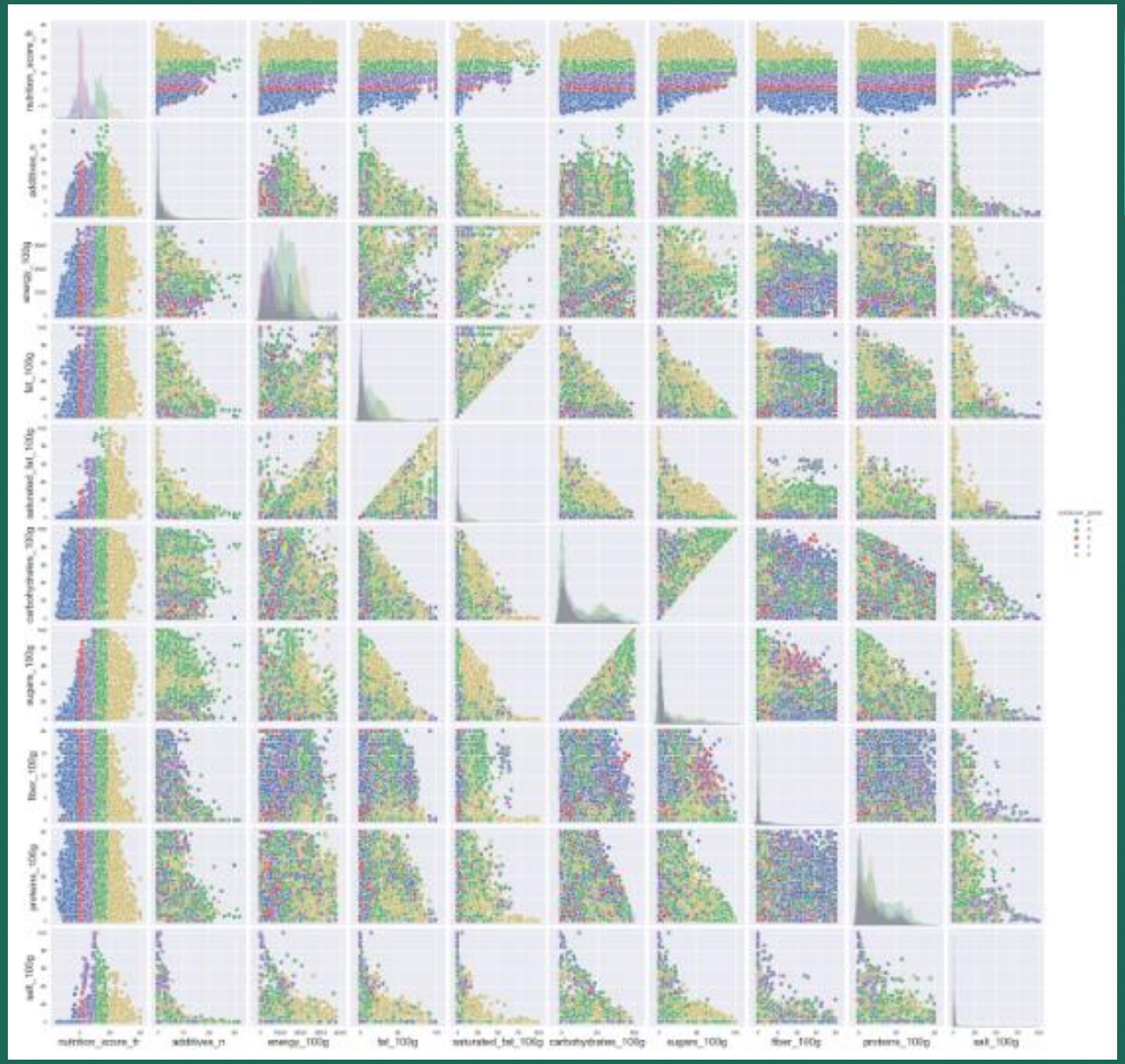
Variables corrélées :

- Energy_100g et fat_100g
- Fat_100g et saturated_fat_100g
- Carbohydrates_100g et sugars_100g

II. Analyses statistiques

2. Analyse bivariée :

Pairplot

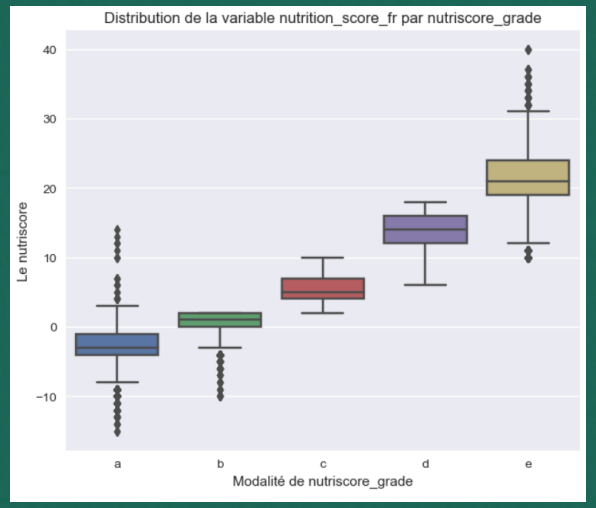


Etiquette:
Nutriscore_grade

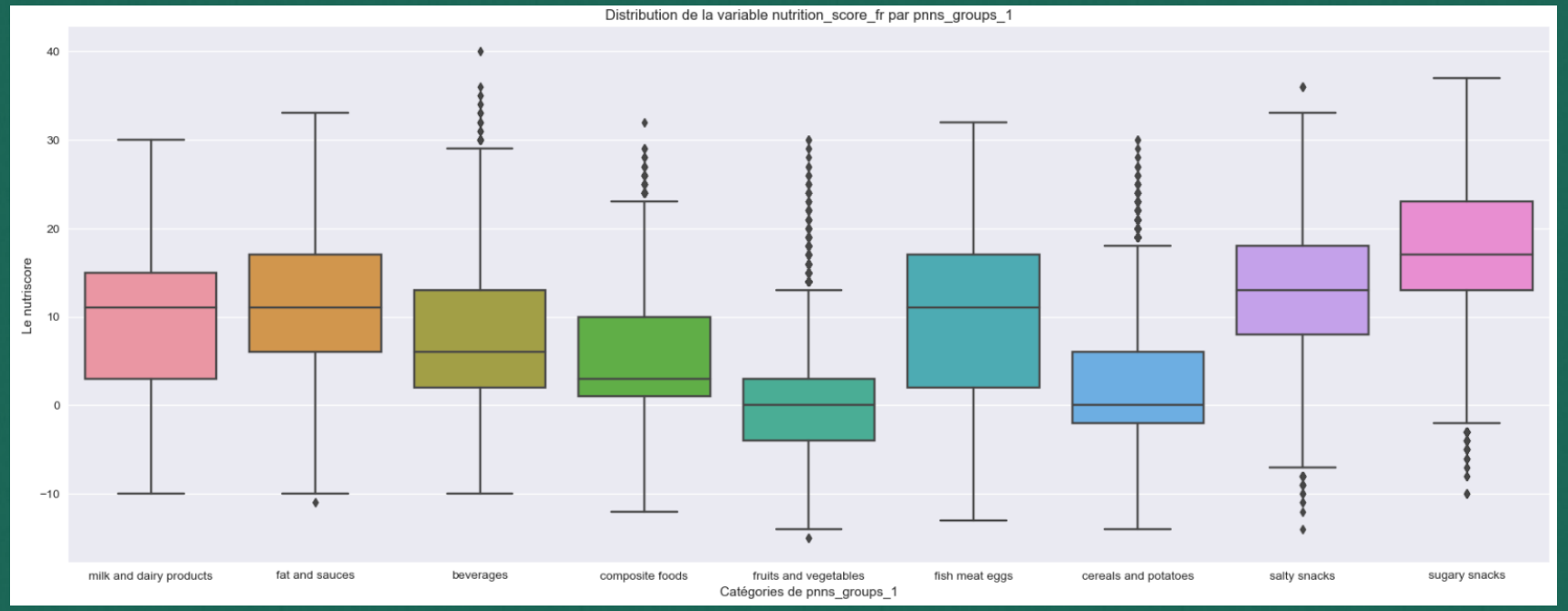
III. Analyses statistiques

2. Analyse bivariée : qualitative - quantitative

Nutriscore_grade
–
Nutrition_score



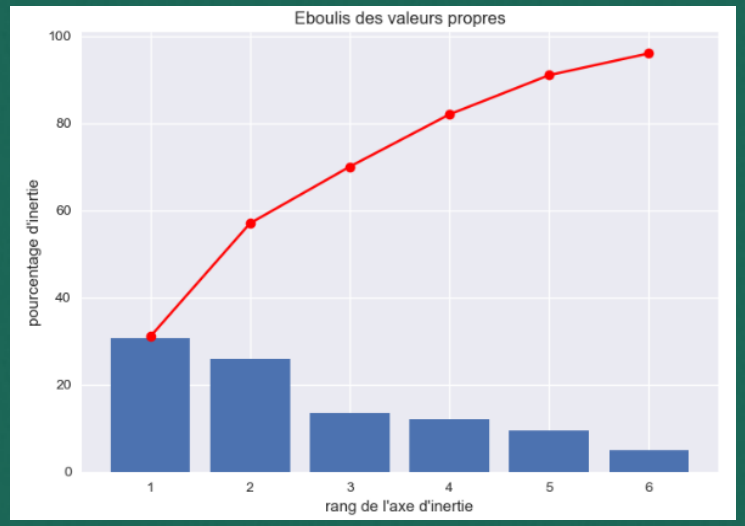
Pnns_groups_1
–
Nutrition_score

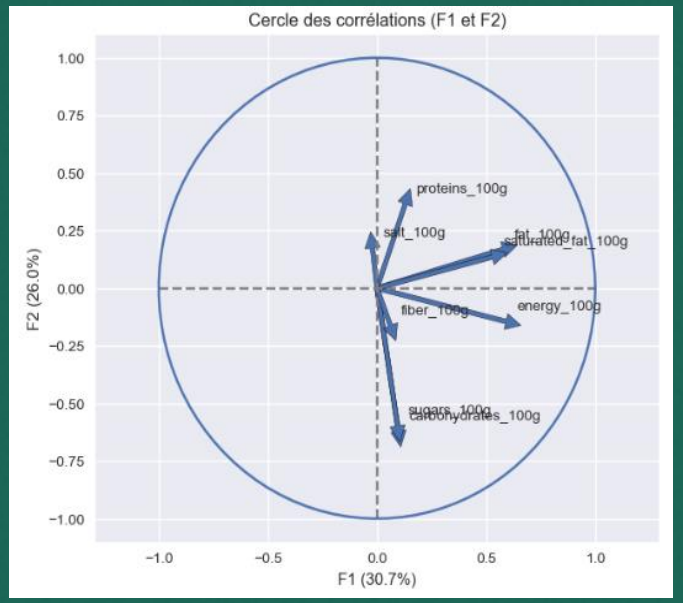


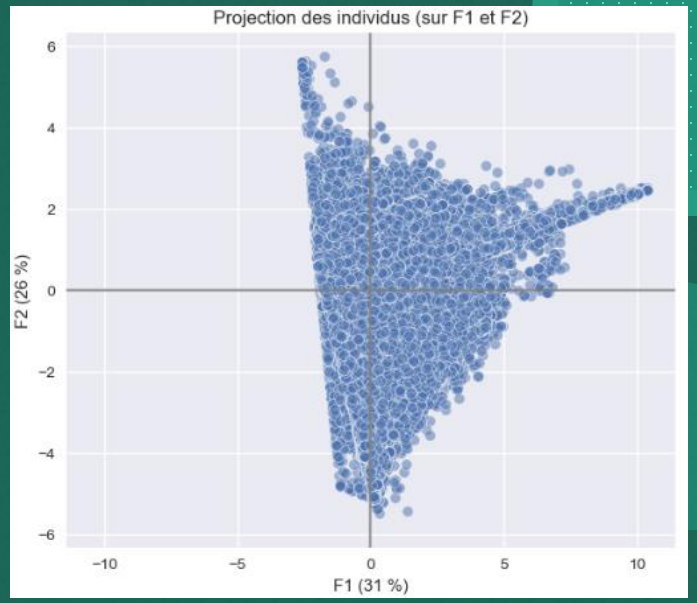
III. Analyses statistiques

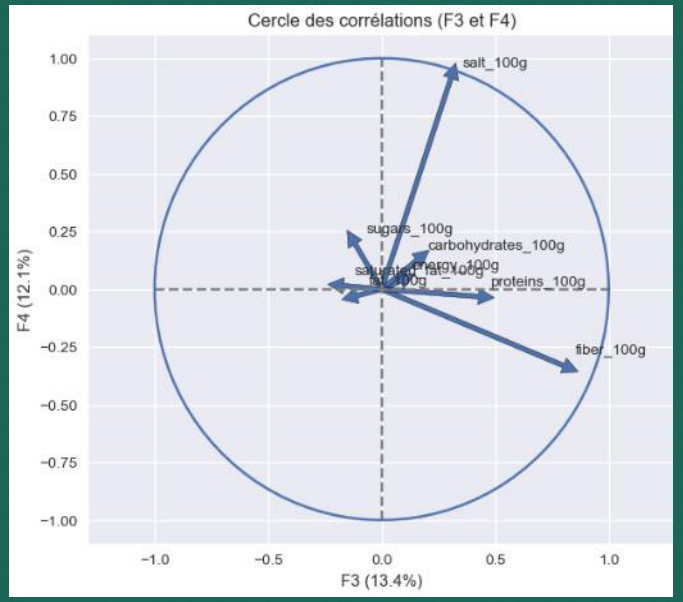
3. Analyse multivariée :

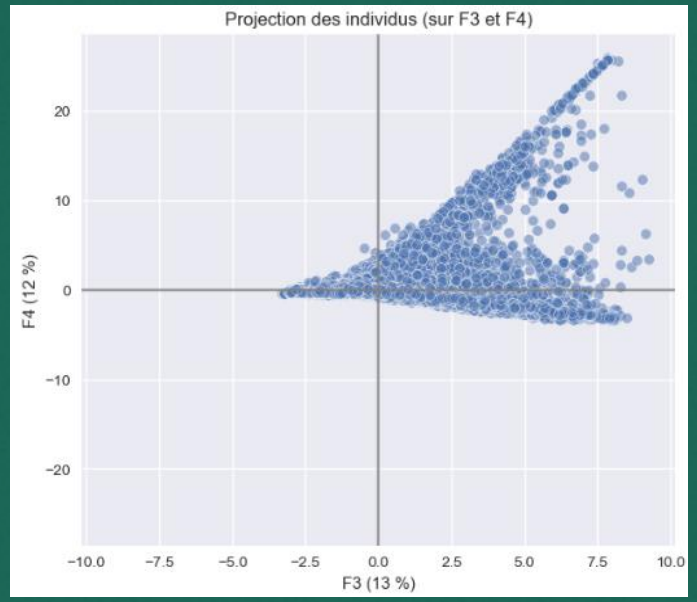
Analyse en composantes principales :











III. Analyses statistiques

3. Analyse multivariée : ANOVA

Peut-on prédire les variables nutritionnelles à partir du nutriscore_grade ?

Pourquoi l'ANOVA ?

Etudier l'impact de la catégorie du nutriscore_grade sur les variables nutritionnelles

Hypothèse de l'ANOVA ?

- Normalité : test de Shapiro-Wilk
- Homogénéité des variances : test de Levene

III. Analyses statistiques

3. Analyse multivariée : ANOVA non paramétrique

Test de Shapiro-Wilk

- Les distributions ne suivent pas une loi normale

Test de Levene

- Variances non homogènes

Test de Kruskal-Wallis :

Les variables nutritionnelles présentent des différences significatives entre les catégories nutriscore_grade

La catégorie du nutriscore_grade a une influence sur les variables nutritionnelles

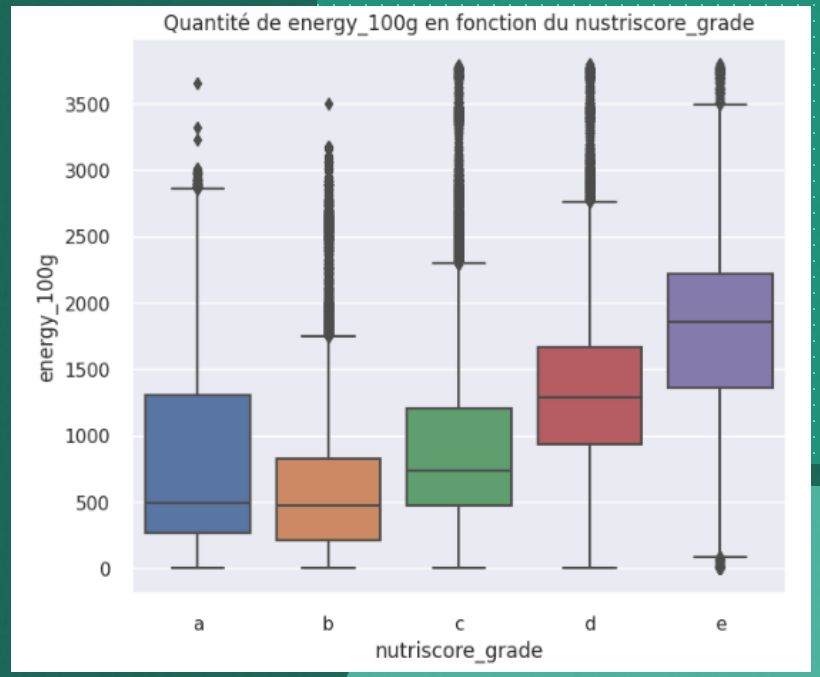
III. Analyses statistiques

3. Analyse multivariée : ANOVA

Test de l'ANOVA :

Energy – nutriscore_grade

```
Test ANOVA - Analyse de la variance
------------------------------------------------------------
energy_100g
Statistique : 23060.60
Valeur p :0.0000
Il y a des différences significatives entre les groupes.
------------------------------------------------------------
```



L'ANOVA confirme le test de Kruskal-Wallis

Régression linéaire

Variable cible : energy_100g

Nutriscore_grade	$R^2 = 0.23$	
	RMSE = 684	
Nutriscore_grade – fat_100g	$R^2 = 0.68$	
	RMSE = 445	
Nutriscore_grade – fat_100g - nutrition_score_fr - saturated_fat_100g	$R^2 = 0.71$	
	RMSE = 421	

Conclusion

- ✓ La précision des prédictions dépend des valeurs déjà renseignées par l'utilisateur.
- ✓ En utilisant les corrélations, on peut prédire les variables nutritionnelles manquantes.
- ✓ L'application proposant des valeurs manquantes est faisable.