



Concevez une application au service de santé publique

Prédire les nutriscore et groupe nova à l'aide de modèle de machine learning

NUTRI-SCORE



NOVA



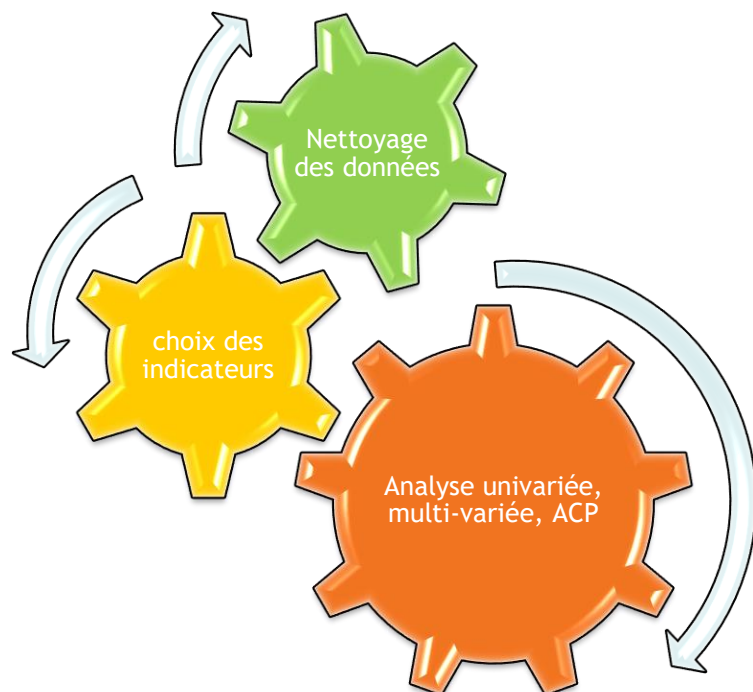


1. Objectifs de la mission

Santé Publique France lance un appel à projet pour trouver des idées innovantes d'application en lien avec l'alimentation

Proposition:

Prédire le nutriscore et le nova groupe à partir de la liste des ingrédients et des informations disponibles sur les étiquettes produits.



Méthodologie:

- Base de données open food fact



- Nettoyage des données
- Choix des indicateurs pertinents
- Analyse uni-variée, multi-variée, Analyse des composantes principales



2. Présentation du jeu de données



Base données base de données sur les produits alimentaires faite par tout le monde, pour tout le monde.

5
catégories

Informations
générales

tags

Ingrédients

Données
diverses

Informations
nutritionnelles

196 variables - 2 610 883 lignes



3. Idée d'application



Score nutritionnel d'un produit

Le logo est attribué sur la base d'un score prenant en compte pour 100 gr ou 100 mL de produit, la teneur :

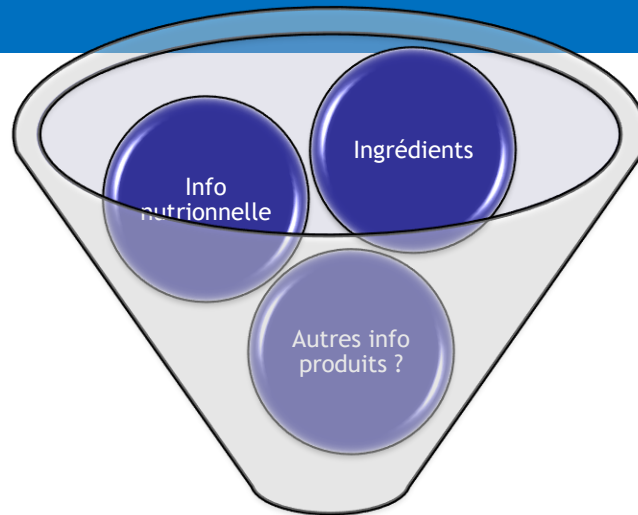
- en nutriments et aliments à favoriser (fibres, protéines, fruits, légumes, légumineuses, fruits à coques, huile de colza, de noix et d'olive),
- et en nutriments à limiter (énergie, acides gras saturés, sucres, sel).

La classification NOVA assigne un groupe aux produits alimentaires en fonction du degré de transformation qu'ils ont subi :

- Groupe 1 - Aliments non transformés ou transformés minimalement
- Groupe 2 - Ingrédients culinaires transformés
- Groupe 3 - Aliments transformés
- Groupe 4 - Produits alimentaires et boissons ultra-transformés



3. Idée d'application



MACHINE
LEARNING



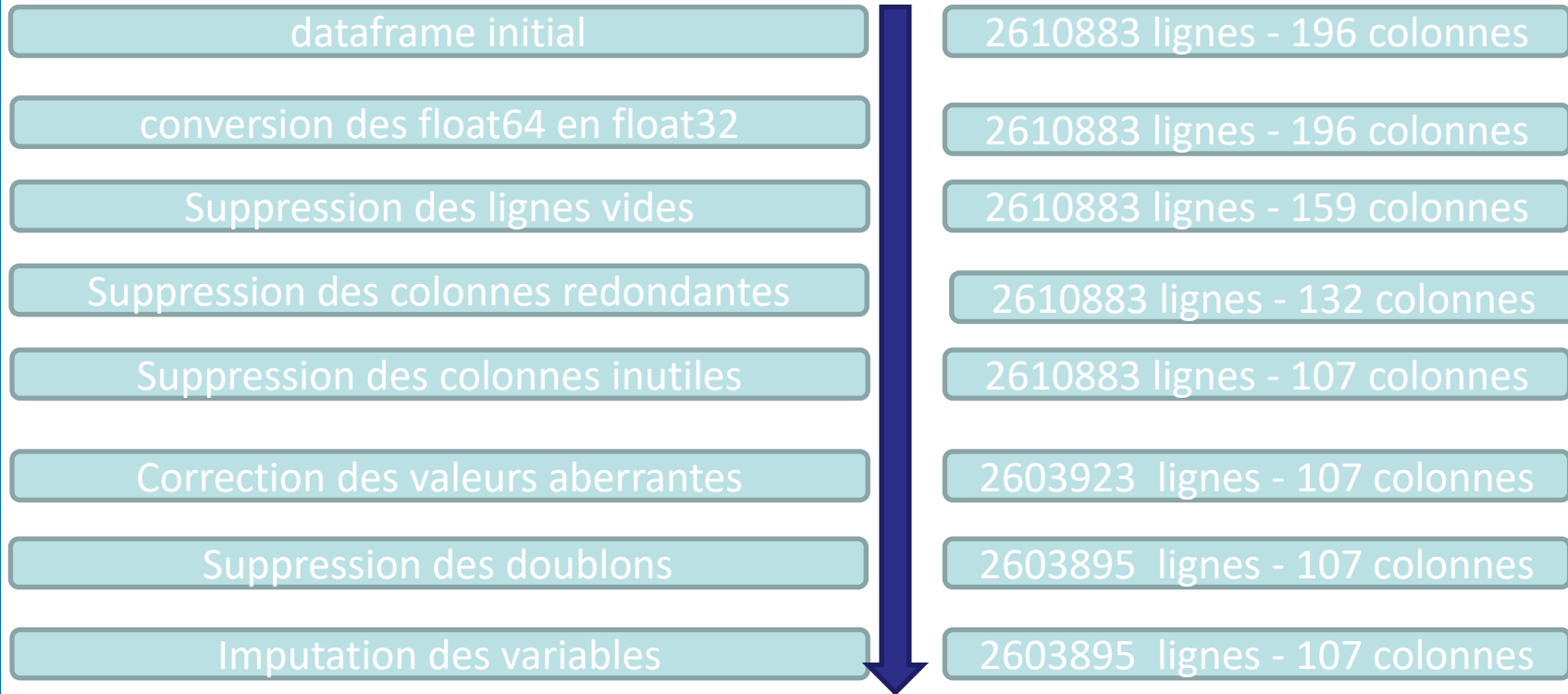
NUTRI-SCORE



NOVA



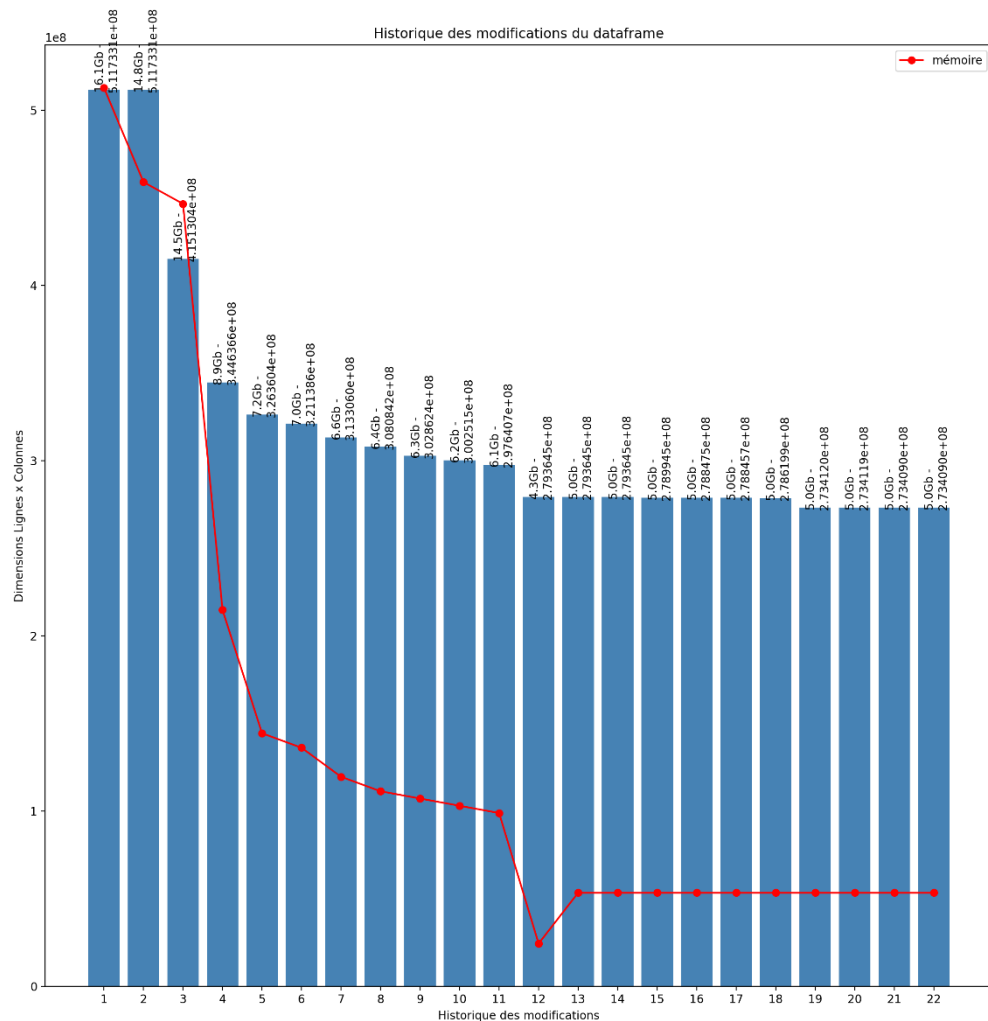
4. Nettoyage du jeu de données



Dataframe final - 2603895 lignes - 107 colonnes



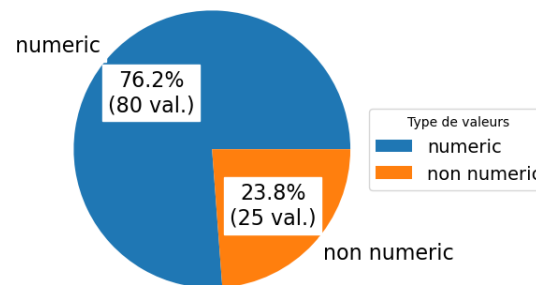
4. Nettoyage du jeu de données



78.3 % de valeurs manquantes
2610883 lignes
196 colonnes
16Gb de mémoire utilisés

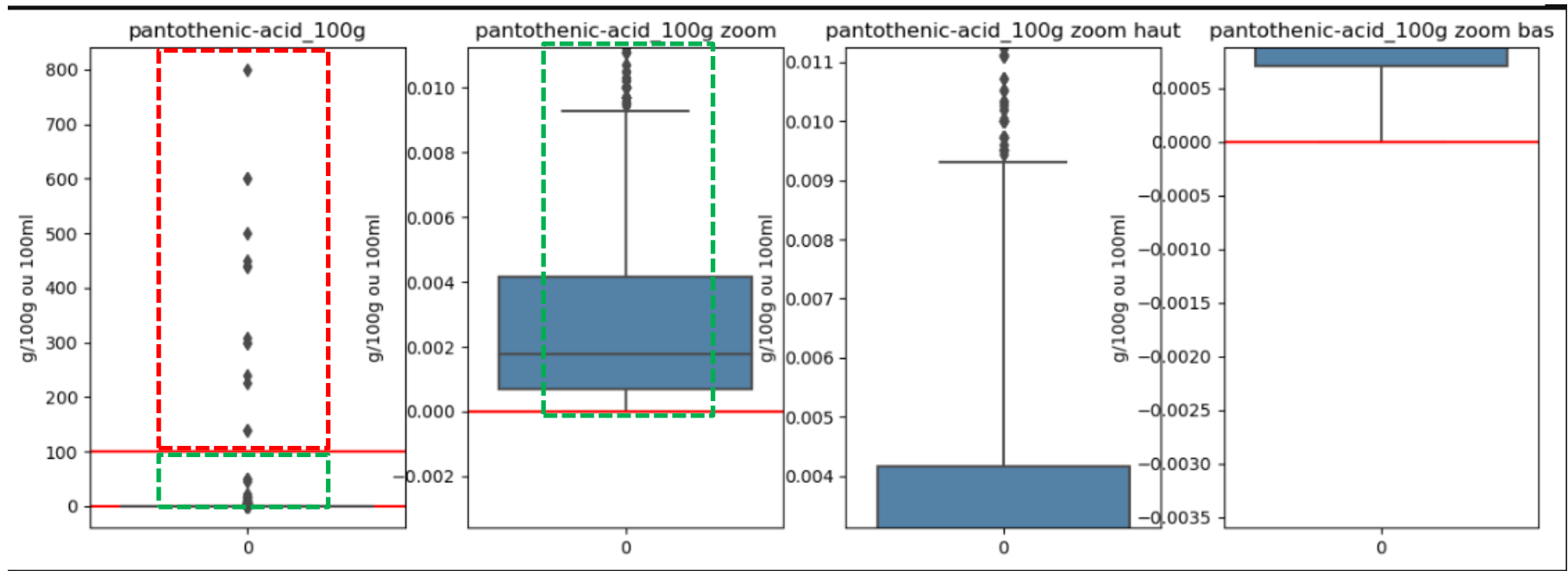


17 % de valeurs manquantes
2603895 lignes
107 colonnes
5Gb de mémoire utilisés



4. Nettoyage du jeu de données : valeurs aberrantes






Des valeurs qui dépassent les 100g par portion de 100g. Plutôt que d'utiliser la méthode interquartile, on remplace toutes les valeurs au dessus de 100g par des 0
On remplace toutes les valeurs négatives par des 0



4. Nettoyage du jeu de données : valeurs aberrantes

beverage						solid food						unknown					
	A	B	C	D	E		A	B	C	D	E		A	B	C	D	E
min	-9	-11	2	6	10	min	-15	-10	2	6	10	min	-14	-10	2	6	10
max	20	2	10	18	40	max	4	2	10	18	40	max	-1	2	10	18	40

Le logo Nutri-Score est ensuite attribué en fonction du score obtenu (cf. tableau ci-dessous).

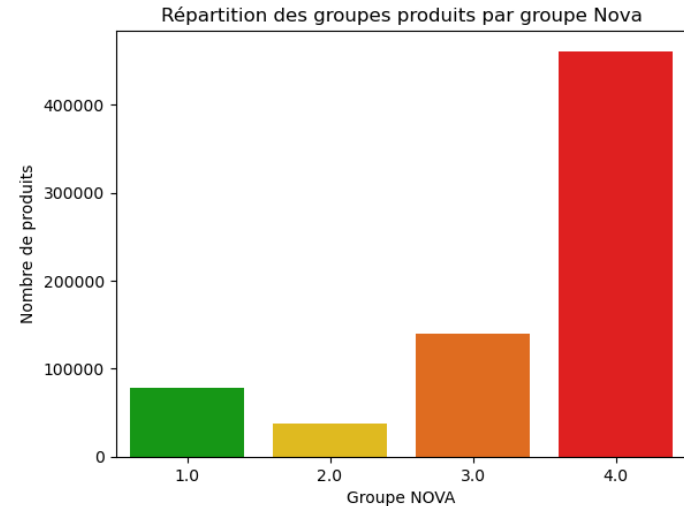
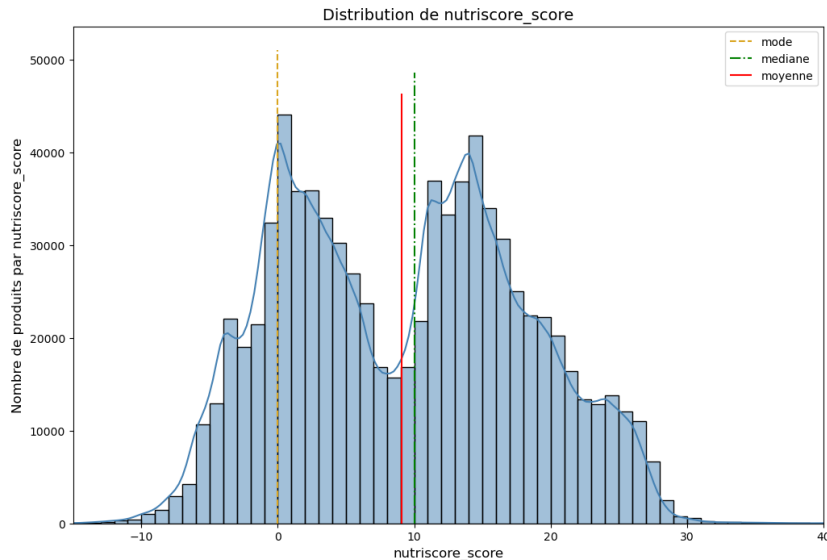
Points		Logo
Aliments solides	Boissons	
Min à -1	Eaux	
0 à 2	Min à 1	
3 à 10	2 à 5	
11 à 18	6 à 9	
19 à Max	10 à Max	

4. Nettoyage du jeu de données : imputation

- On va considérer que les variables de quantité pour 100g non renseignées valent zero
- Pour la variable `energy_kj`, celle-ci sera imputée à l'aide du modèle `iterativeImputer` sur le jeu de données réduit, présenté par la suite
- Pas d'autres imputations, les valeurs non renseignées sont supprimées



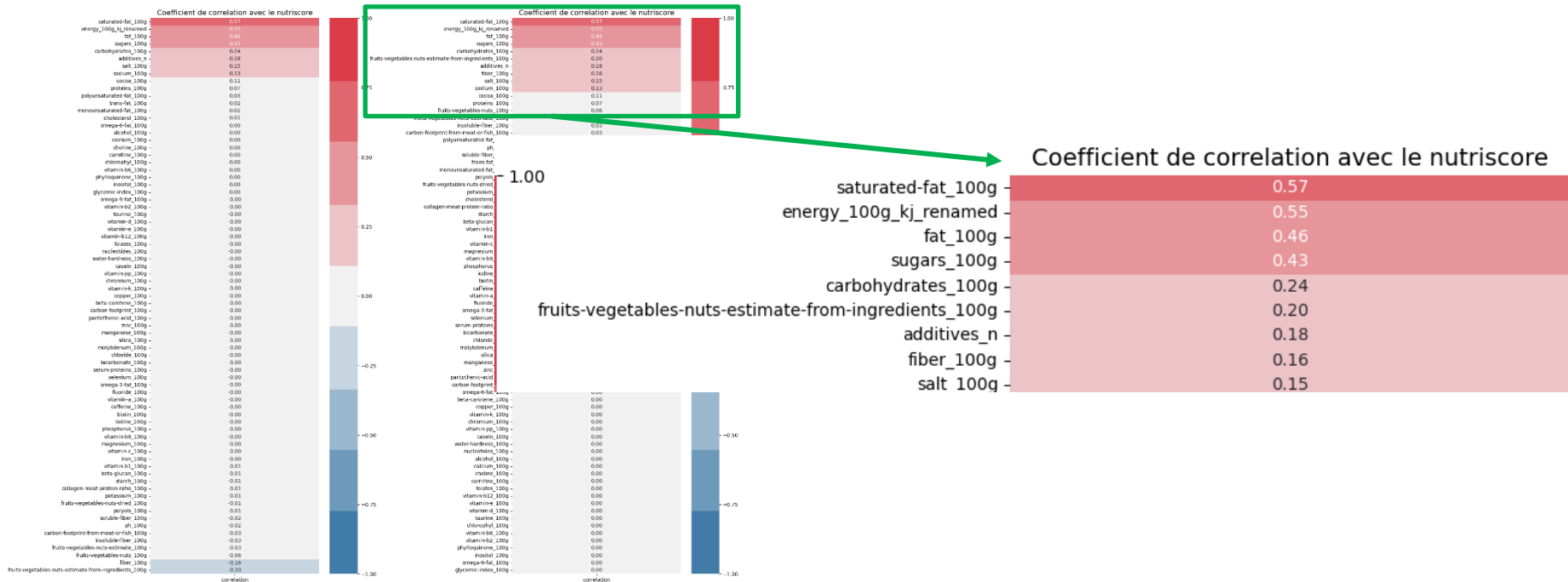
5. Analyse des données



- Les distributions ne suivent pas une loi normale
- La distribution du nutriscore est quasi-bimodale
- Le skewness est légèrement positif, la distribution est étalée sur la droite, mais presque symétrique
- Le kurtosis est négatif, indiquant une distribution plus aplatie que celle que suivrait une loi normale



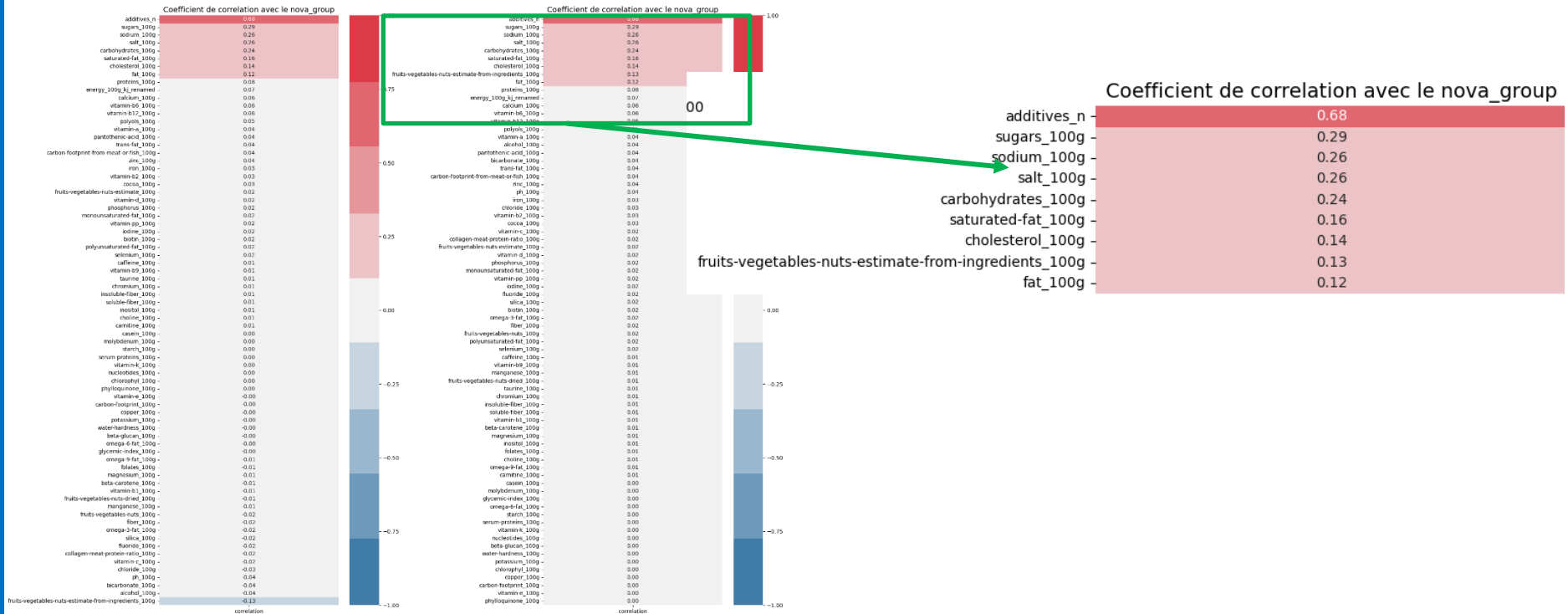
5. Corrélation avec le nutriscore



- 9 variables qui corrént avec le nutriscore
- Corrélations significatives confirmées le test de significativité



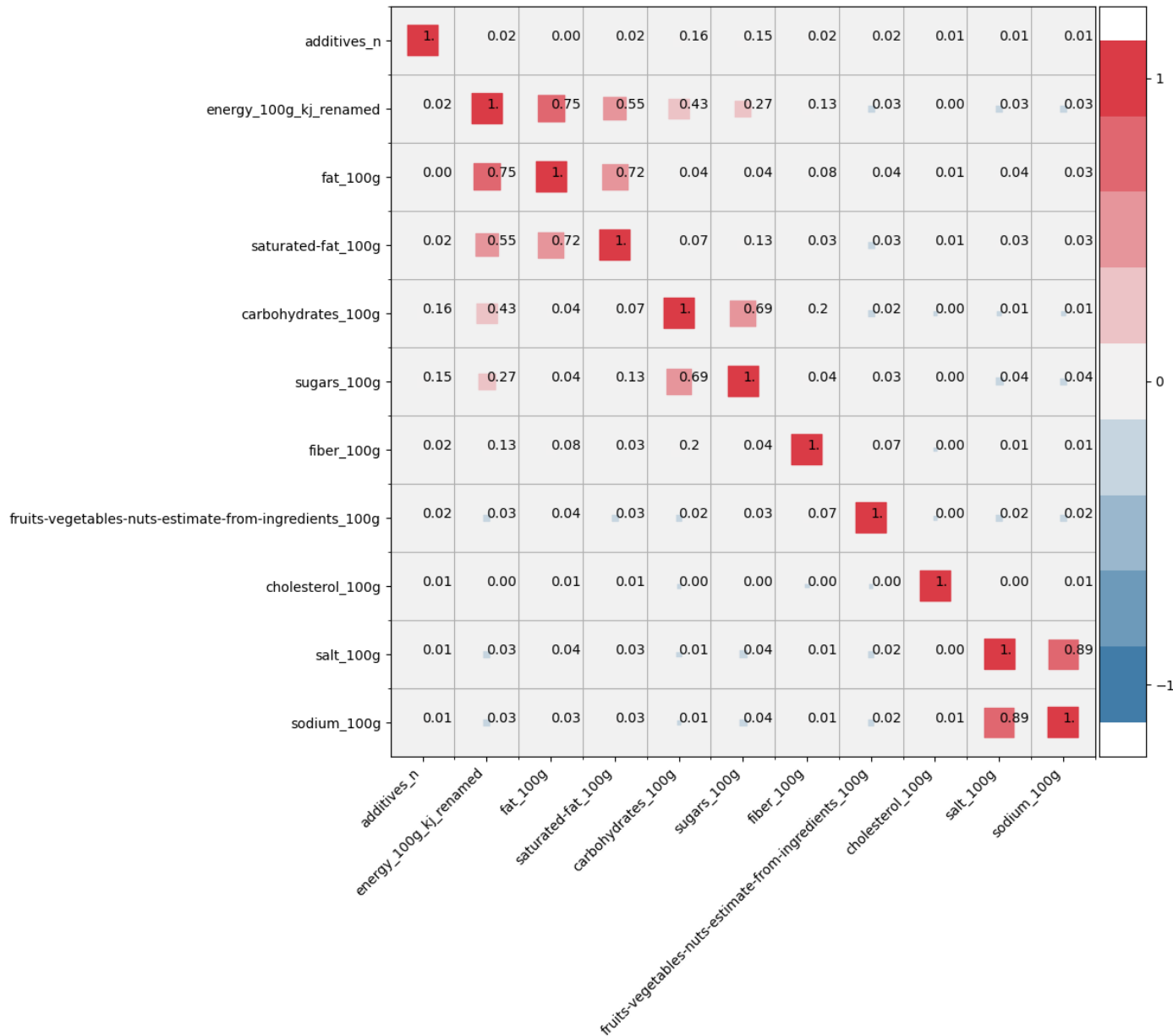
5. Corrélation avec le nova_group



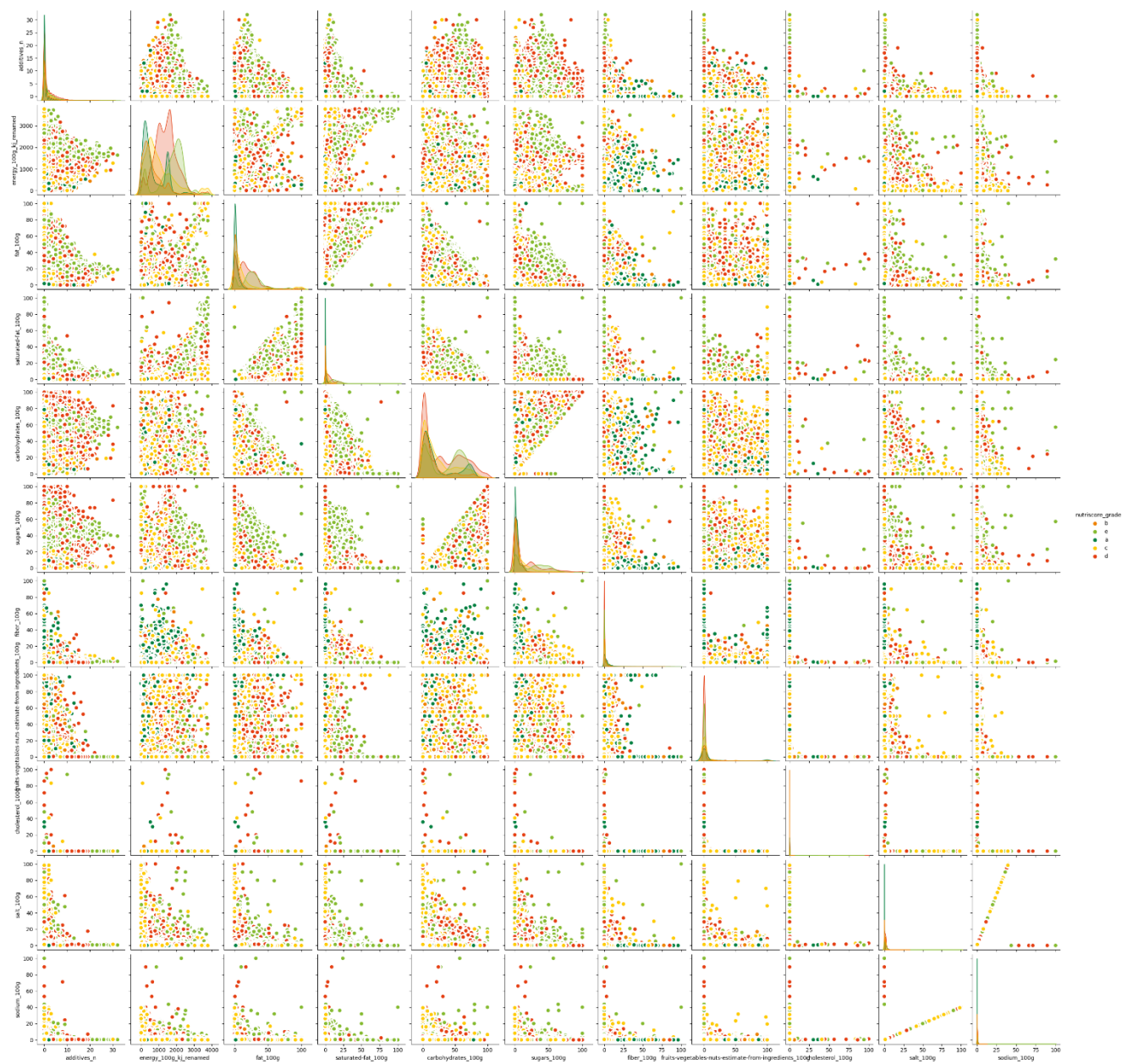
- 9 variables qui corrént avec le group nova
- Corrélations significatives confirmées le test de significativité



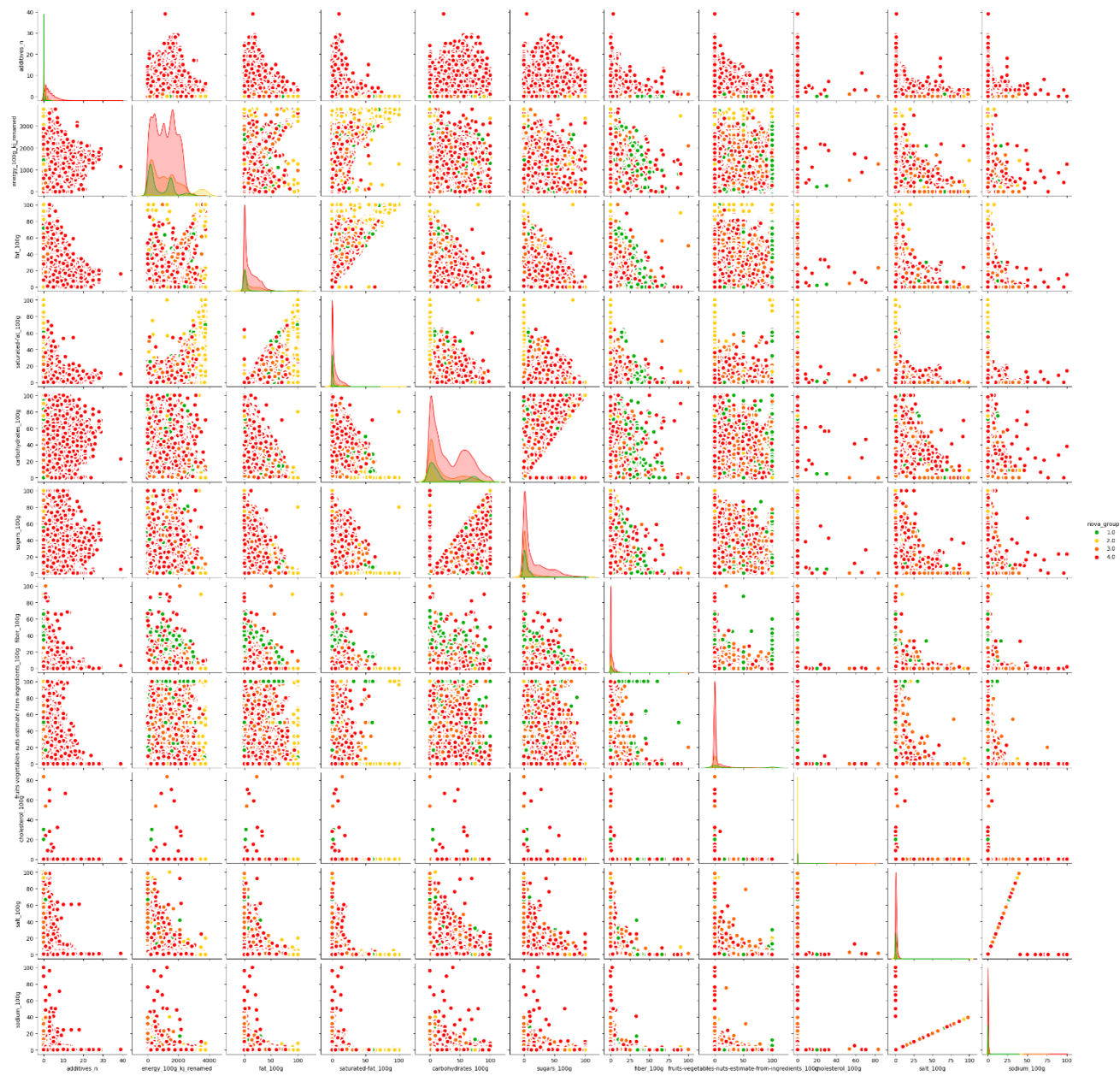
5. Matrice des corrélations



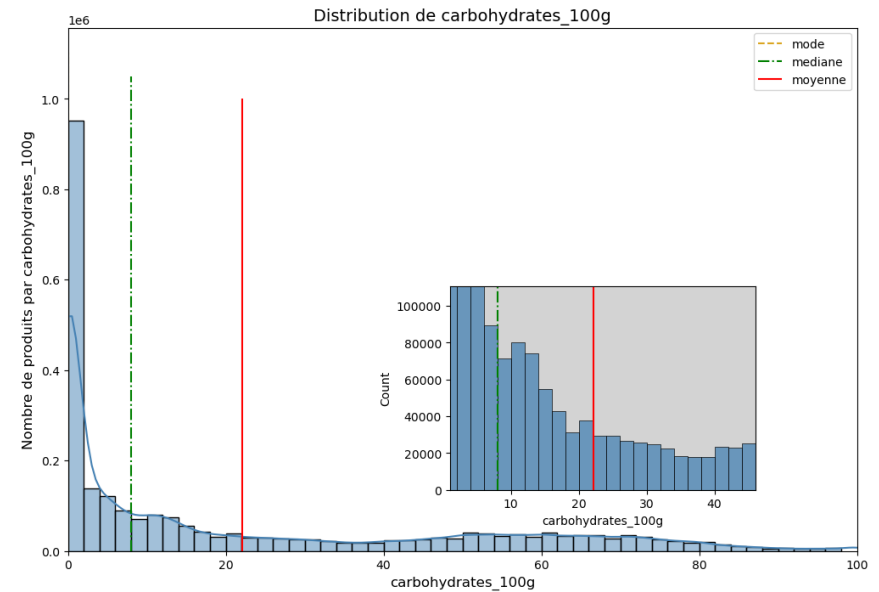
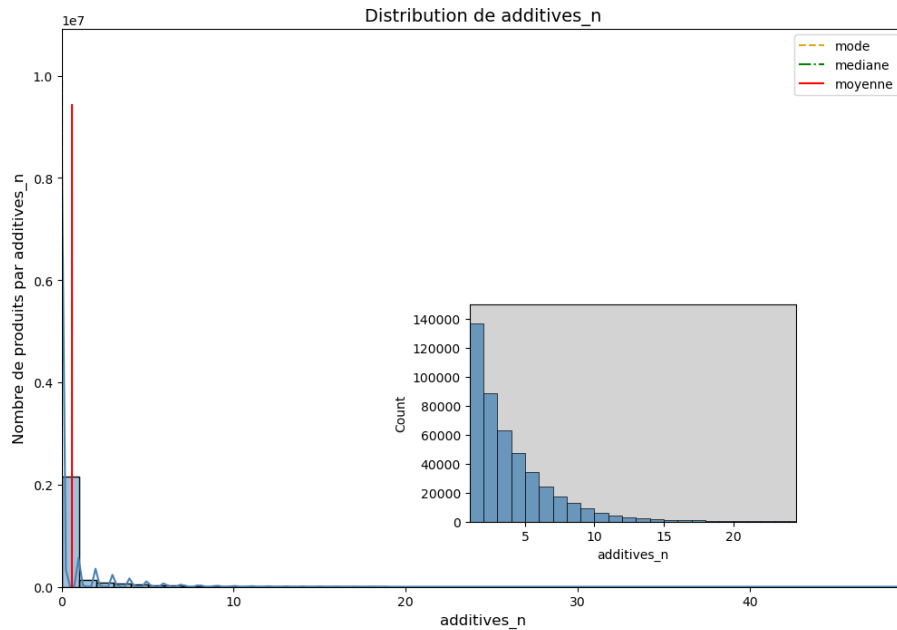
5. Pair plots par nutriscore



5. Pair plots par groupe Nova



5. Distributions



- Les distributions des variables retenues ne suivent pas une loi normale
- Les 3 tests de normalité (Shapiro, D'Agostino et Anderson-Darling) confirment le résultat visuel



5. Tests non paramétriques

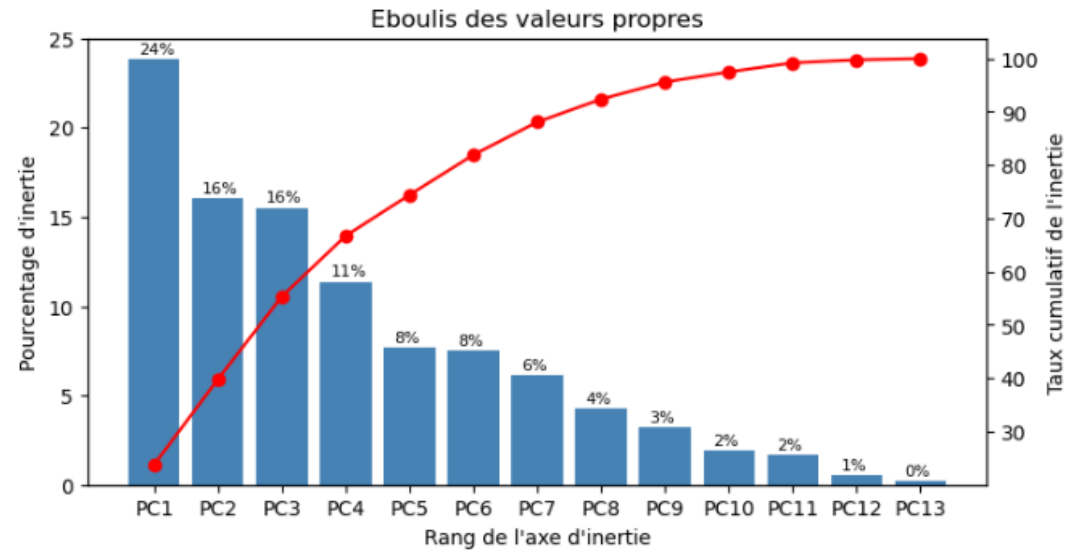
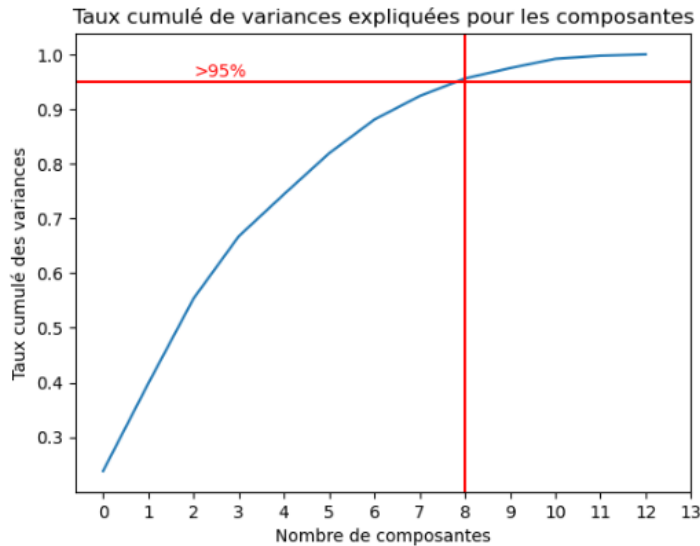
- Test de Krustal-Wallis avec le nova groupe (variable qualitative)

	K_stat	P_val	p value < 0.05	bilan
additives_n	293393.50	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
energy_100g_kj_renamed	43195.61	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
fat_100g	39621.61	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
saturated-fat_100g	38198.44	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
carbohydrates_100g	37364.99	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
sugars_100g	52541.32	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
fiber_100g	22180.58	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
fruits-vegetables-nuts-estimate-from-ingredients_100g	20996.77	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
cholesterol_100g	12719.54	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
salt_100g	116618.56	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
sodium_100g	116588.96	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique

- Test de Wilcoxon avec le nutriscore (variable quantitative)

	K_stat	P_val	p value < 0.05	bilan
additives_n	15968782248.00	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
energy_100g_kj_renamed	8999151.00	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
fat_100g	41464145181.00	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
saturated-fat_100g	28298319807.50	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
carbohydrates_100g	17850345577.50	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
sugars_100g	49135146112.00	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
fiber_100g	20491064370.00	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
fruits-vegetables-nuts-estimate-from-ingredients_100g	42104394645.50	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
cholesterol_100g	9252822861.50	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
salt_100g	13990874796.00	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
sodium_100g	11968841686.50	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique
nova_group	22300433638.00	0.00	True	H0 acceptée - les deux échantillons sont significativement différents d'un point de vue statistique

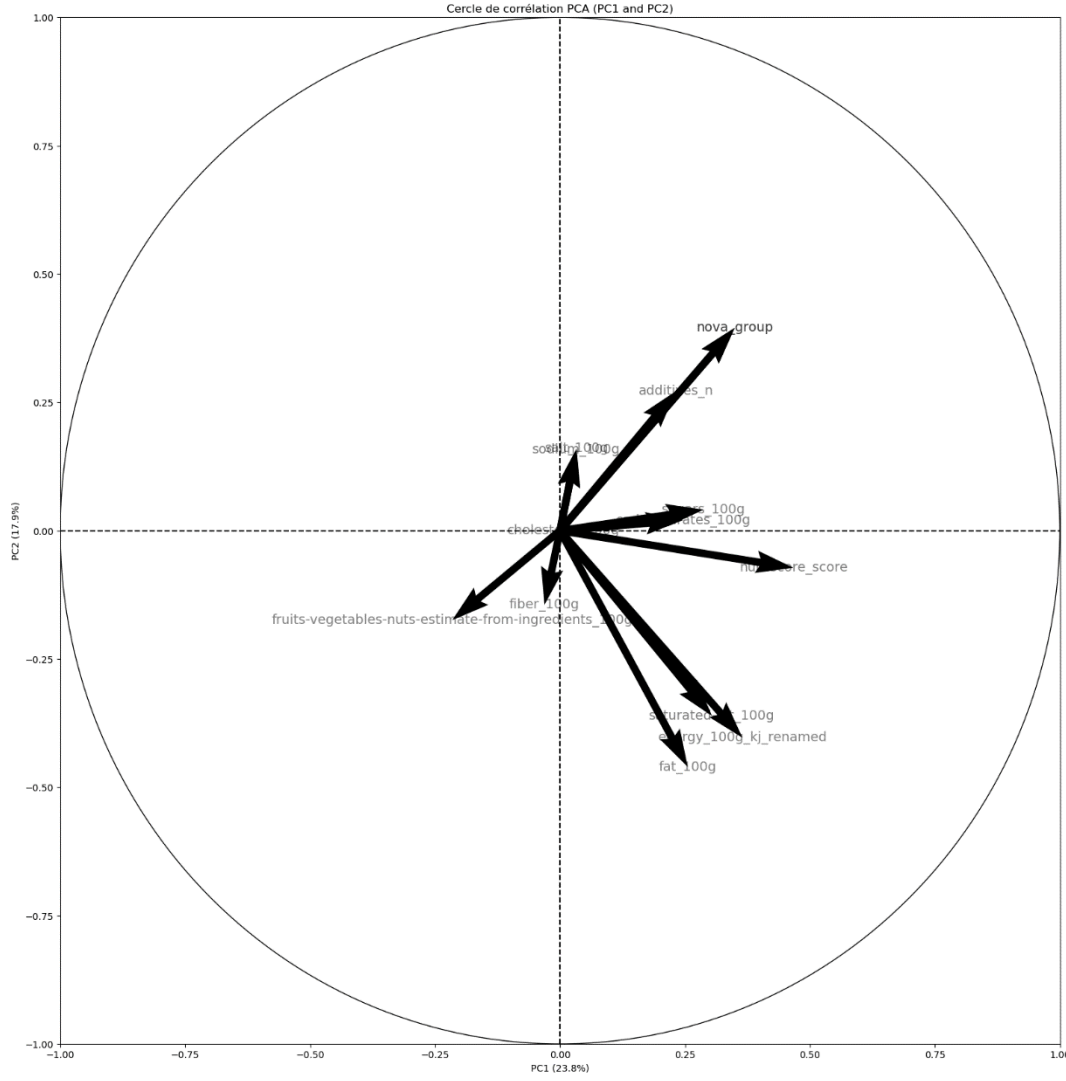
5. Analyse des composantes principales



- Les 8 premières composantes principales expliquent 95% de la variance de l'échantillon de données
- F1 explique à elle seule $\frac{1}{4}$ de la variance



5. Analyse des composantes principales



F1 représente les apports journaliers recommandés en lipides et acides gras saturés, et dans une moindre mesure en sucre.

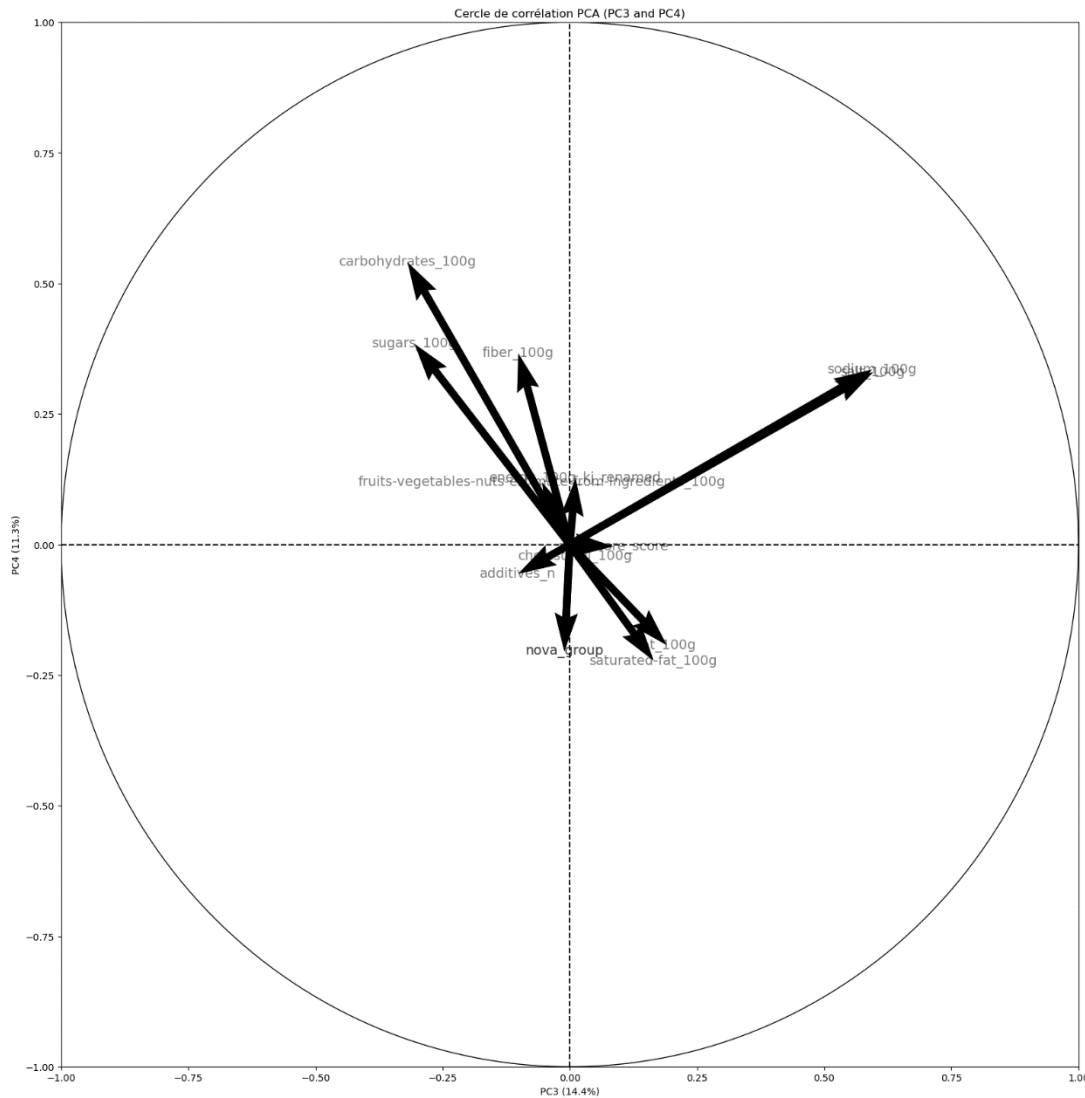
F1 encode les lipides, acides gras saturés, sucre

Le nombre d'additifs et le groupe NOVA corrélient fortement de façon positive avec F2. Les variables sels et sodium, fortement corrélées, corrélient positivement avec F2. Les aliments ultra-transformés sont riches en sucre et en graisse.

F2 encode ces 4 variables



5. Analyse des composantes principales



F3 encode le sel

F4 encode le sucre et les glucides



6. Conclusion

Le nettoyage de données a permis d'obtenir des données de qualité
Ces données ne suivent pas de loi normale
Certaines données devraient permettre de prédire le nutriscore et le groupe nova à l'aide de modèles de machine learning



Recommandations et prolongement

- Extension à l'eco-score
- Tester les modèles de machine learning
- Sélectionner le modèle le plus adapté et optimiser les paramètres.



Des questions ?

