



P2

Concevez une application au service de la santé publique





Contexte

L'agence "**Santé publique France**" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation. Vous souhaitez y participer et proposer une idée d'application.

Objectifs

- Trouver une idée d'application : sélection des variables pertinentes
- Traitement des données (imputation, gestion des valeurs aberrantes / manquantes)
- Tests hypothèses et analyses exploratoires
- Identifier des arguments justifiant la faisabilité (ou non) de l'application à partir des données Open Food Facts.



Sommaire

01

Les
données

02

Idée
Application

03

Preprocessing

04

Analyse
univariée/
Bivariée

05

Analyse
multivariée

06

Conclusion



01

Les données



Data Open Food Facts

Open Food Facts est un projet collaboratif dont le but est de constituer une base de données libre et ouverte sur les produits alimentaires commercialisés dans le monde entier

Le jeu de données Open Food Facts:

- Les **informations générales sur la fiche du produit** : nom, date de modification, etc.
- **Un ensemble de tags** : catégorie du produit, localisation, origine, etc.
- **Les ingrédients** composant les produits et leurs additifs éventuels.
- **Des informations nutritionnelles** : quantité en grammes d'un nutriment pour 100 grammes du produit.





320772, 162



02
Idée
d'application



Nutri-Score prediction

Utiliser un algorithme de Machine learning afin de déterminer le Nutriscore d'un produit avec uniquement ces valeurs nutritionnelles

Grade nutritionnel : Echelle de A à E présente sur les produits





03

Preprocessing





Sélection des variables

Variables nutritionnelles

La quantité : d'énergie, gras, gras saturés , sucres, sel, fibres , protéines, carbohydrates

Variables descriptives

Product Name, Nutrition Grade, Pnns groupe

	fat_100g	sugars_100g	energy_100g	carbohydrates_100g	saturated-fat_100g	fiber_100g	proteins_100g	salt_100g	pnnsgroups_1	product_name	nutrition_grade_fr
0	3.0	8.5	477.0	8.5	2.0	0.0	13.0	2.000000	Milk and dairy products	Bridelight 3% Les carrés fondants goût Emmental	C
1	0.0	9.7	179.0	10.0	0.0	0.0	0.5	0.030000	Beverages	100 % Pur Jus Pomme	C
2	0.8	56.0	1378.0	76.0	0.1	4.8	0.8	0.050038	Sugary snacks	18 marrons glacés	C



Les opérations de nettoyage



1. Suppression des features avec un taux de valeurs manquantes > 90% (fruits-vegetables-nuts_100g)
2. Filtrage des données sur la feature **Countries** en gardant uniquement la France (fonction Regex)
3. Filtrage des données temporelles de 2012 à 2017 (1 valeur NaN)
4. Suppression de *Unknow* sur le Pnns Group
5. Renomme correctement 3 catégories du Pnns Group (9 cat uniques au lieu de 11)
6. Suppression des duplicated values (sur l'ensemble du Dataframe)
7. Group By sur la feature **product name** avec agrégation par la médiane des valeurs nutritionnelles



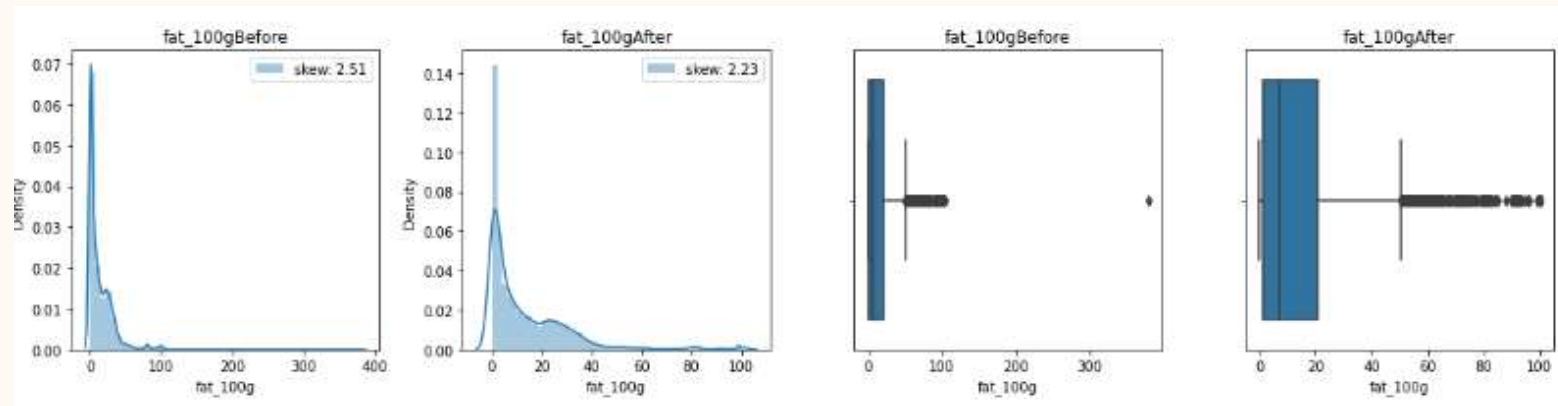


Traitement des données

Vérification des valeurs nutritionnelles en gramme de 0 à 100g pour chaque produit

>suppression des produits hors de l'intervalle

	product_name	fat_100g
25242	Graine de couscous moyen	105.0
63760	mini choux goût fromage de chèvre - poivre	380.0



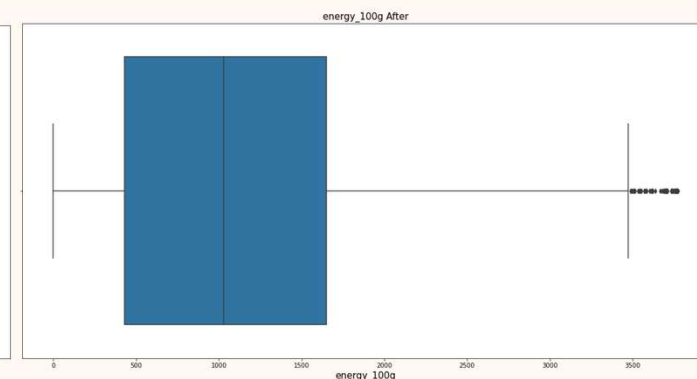
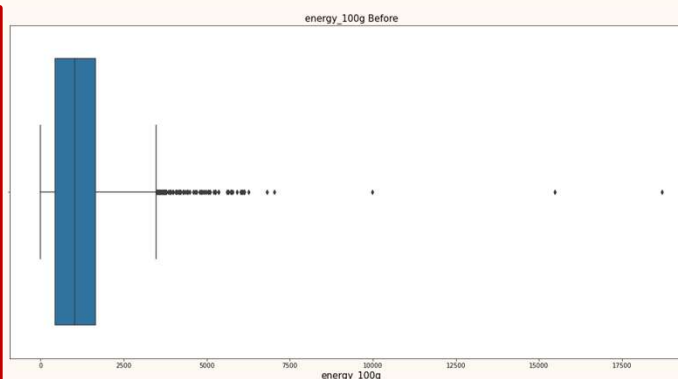


Traitement des données

Vérification des valeurs énergétiques en Joule de 0 à 3800J pour chaque produit

>suppression des produits hors de l'intervalle, il y a un total de 70 produits ou l'énergie est supérieure à 3800Joules.

	product_name	fat_100g	sugars_100g	energy_100g
834	1KG Nuggets Poulet Maitre Coq	NaN	1.30	4289.0
893	2 Eclairs gourmands citron meringué	NaN	35.80	5272.0
3860	Assortiment de roulades	NaN	0.90	4071.0





Valeurs aberrantes

Tous les **nutrition-score-fr_100g** et les **nutrition_grade_fr** ne sont pas correctement associés.
Pour faire ce constat, nous croisons les variables dans un même tableau. **Nutri-Score A : -15 à -1**

nutrition_grade_fr	a	b	c	d	e
nutrition-score-fr_100g					
-14.0	3	0	0	0	0
-13.0	9	0	0	0	0
-12.0	15	0	0	0	0
-11.0	41	0	0	0	0
-10.0	60	0	0	0	0
-9.0	84	0	1	0	0
-8.0	144	1	0	0	0
-7.0	210	2	0	0	0
-6.0	900	4	0	0	0
-5.0	979	12	0	0	0
-4.0	1009	27	0	1	0
-3.0	878	14	2	0	0
-2.0	1196	42	2	3	0
-1.0	1525	18	5	1	0

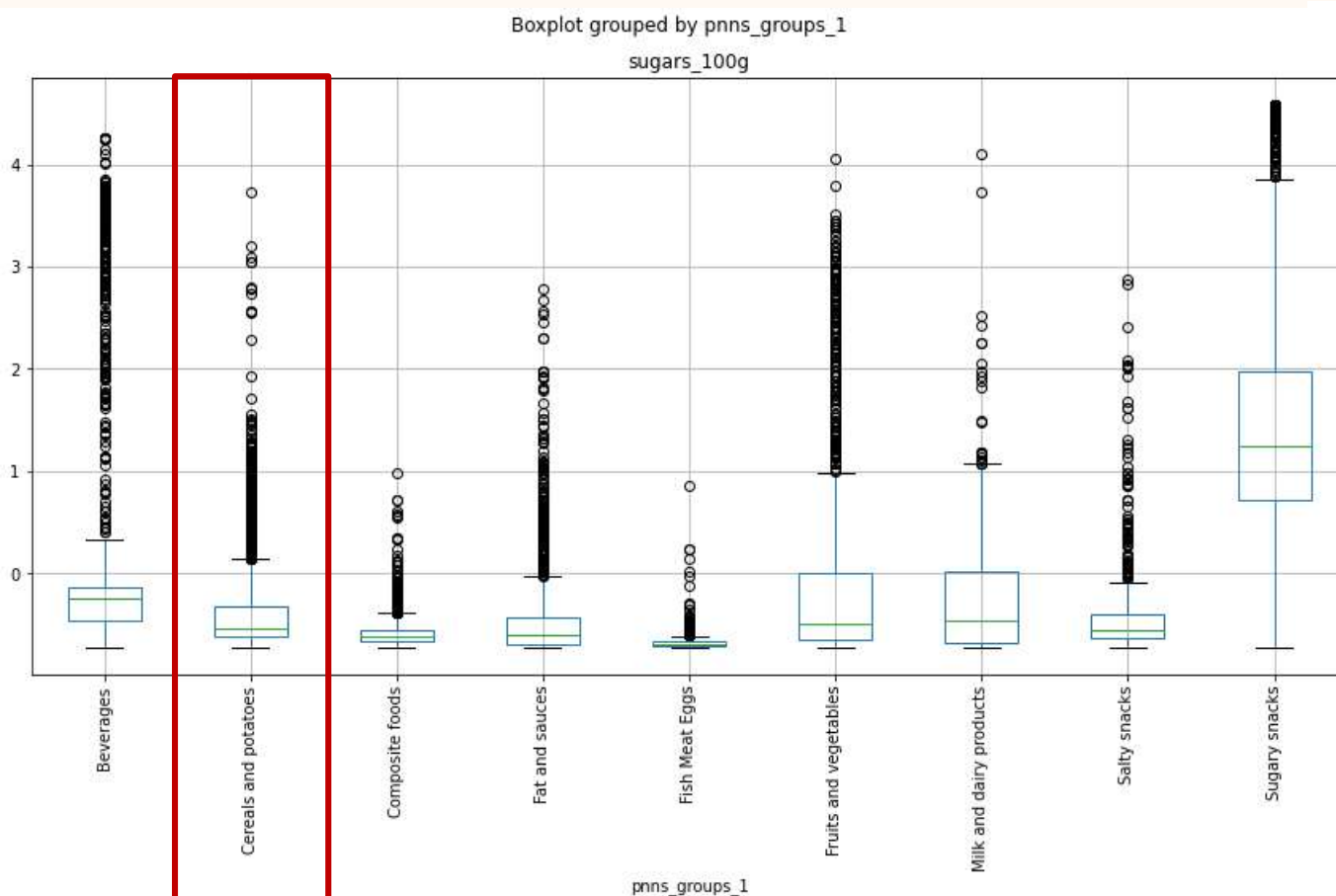
nutrition_grade_fr	A	B	C	D	E
nutrition-score-fr_100g					
-14.0	3	0	0	0	0
-13.0	9	0	0	0	0
-12.0	16	0	0	0	0
-11.0	43	0	0	0	0
-10.0	64	0	0	0	0
-9.0	86	0	0	0	0
-8.0	149	0	0	0	0
-7.0	217	0	0	0	0
-6.0	929	0	0	0	0
-5.0	1024	0	0	0	0
-4.0	1068	0	0	0	0
-3.0	911	0	0	0	0
-2.0	1280	0	0	0	0
-1.0	1577	0	0	0	0





Valeurs atypiques

Outliers – méthode de Interquartile (Tukey)



pnns	nutritionnelle	outliers
Cereals and potatoes	sugars_100g	753
Beverages	salt_100g	637
Sugary snacks	carbohydrates_100g	457
Beverages	saturated-fat_100g	426
Cereals and potatoes	saturated-fat_100g	407
Beverages	fat_100g	401
Fish Meat Eggs	carbohydrates_100g	398
Sugary snacks	fiber_100g	397
Fruits and vegetables	energy_100g	391
Fish Meat Eggs	sugars_100g	374
Beverages	energy_100g	352



Imputation

Describe() sans imputation

	fat_100g	sugars_100g	energy_100g	carbohydrates_100g	saturated-fat_100g	fiber_100g	proteins_100g	salt_100g
count	33478.000000	45736.000000	46978.000000	33144.000000	45651.000000	64989.000000	46885.000000	45764.000000
mean	13.161916	13.519979	1107.763522	27.722506	5.507449	1.313119	7.841506	0.995777
std	16.403927	18.889725	767.329254	27.294064	8.593449	3.475545	7.802238	2.930120

Knn Imputer avec toutes les missing values

	fat_100g	sugars_100g	energy_100g	carbohydrates_100g	saturated-fat_100g	fiber_100g	proteins_100g	salt_100g
count	64989.000000	64989.000000	64989.000000	64989.000000	64989.000000	64989.000000	64989.000000	64989.000000
mean	11.780978	11.913556	1025.727146	27.633396	4.173991	1.313119	7.448868	0.944039
std	12.791154	16.546378	667.554381	21.548974	7.568075	3.475545	6.679092	2.464112

Knn Imputer uniquement les rows avec missing values<50%

	fat_100g	sugars_100g	energy_100g	carbohydrates_100g	saturated-fat_100g	fiber_100g	proteins_100g	salt_100g
count	46909.000000	46909.000000	46909.000000	46909.000000	46909.000000	46909.000000	46909.000000	46909.000000
mean	12.474881	13.509040	1108.554828	27.275873	5.510056	1.818805	7.838918	1.002305
std	15.497559	18.863083	766.824167	26.419368	8.531838	3.976825	7.800604	2.906719





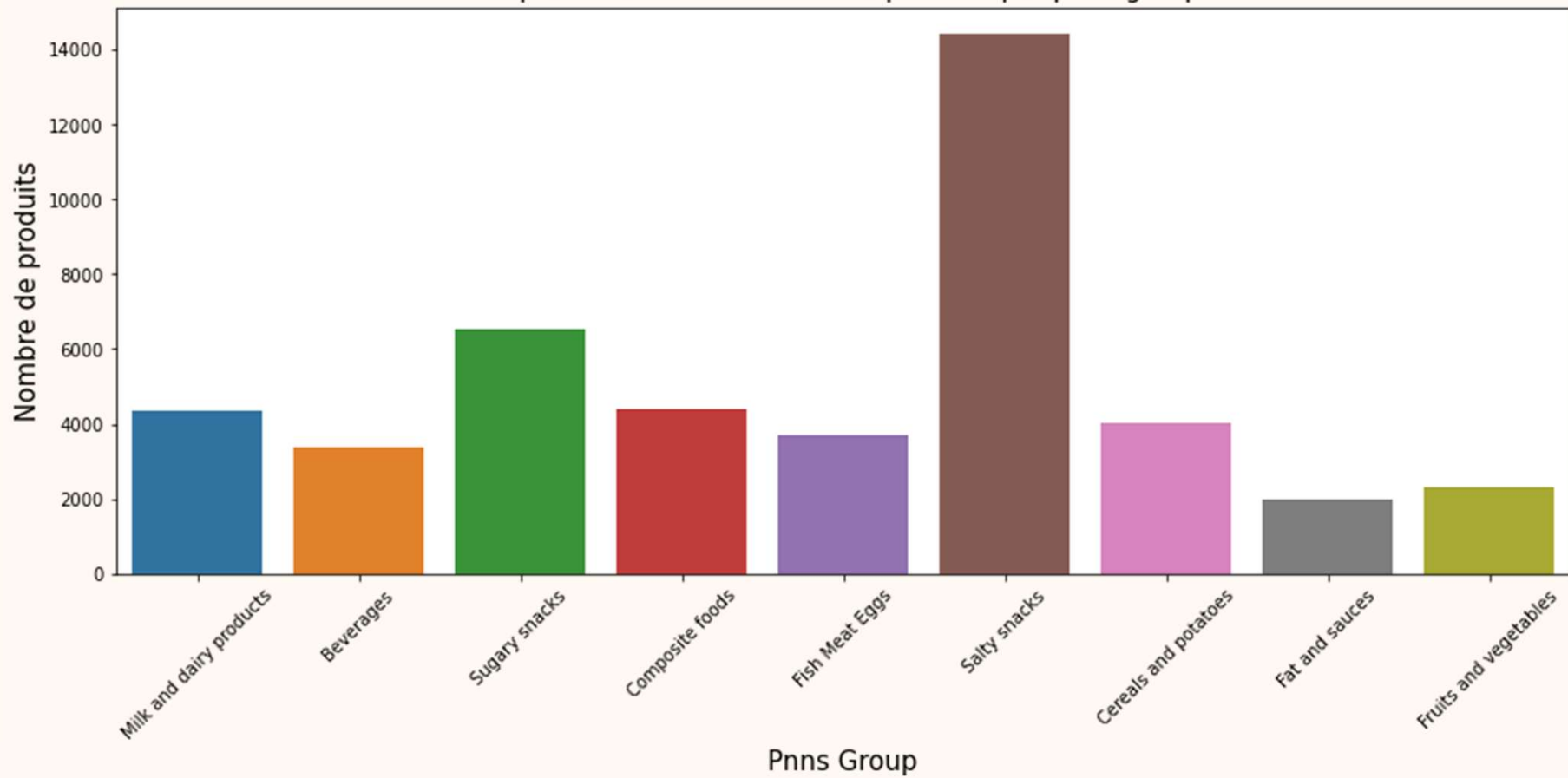
04

Analyse
univariée et
bivariée

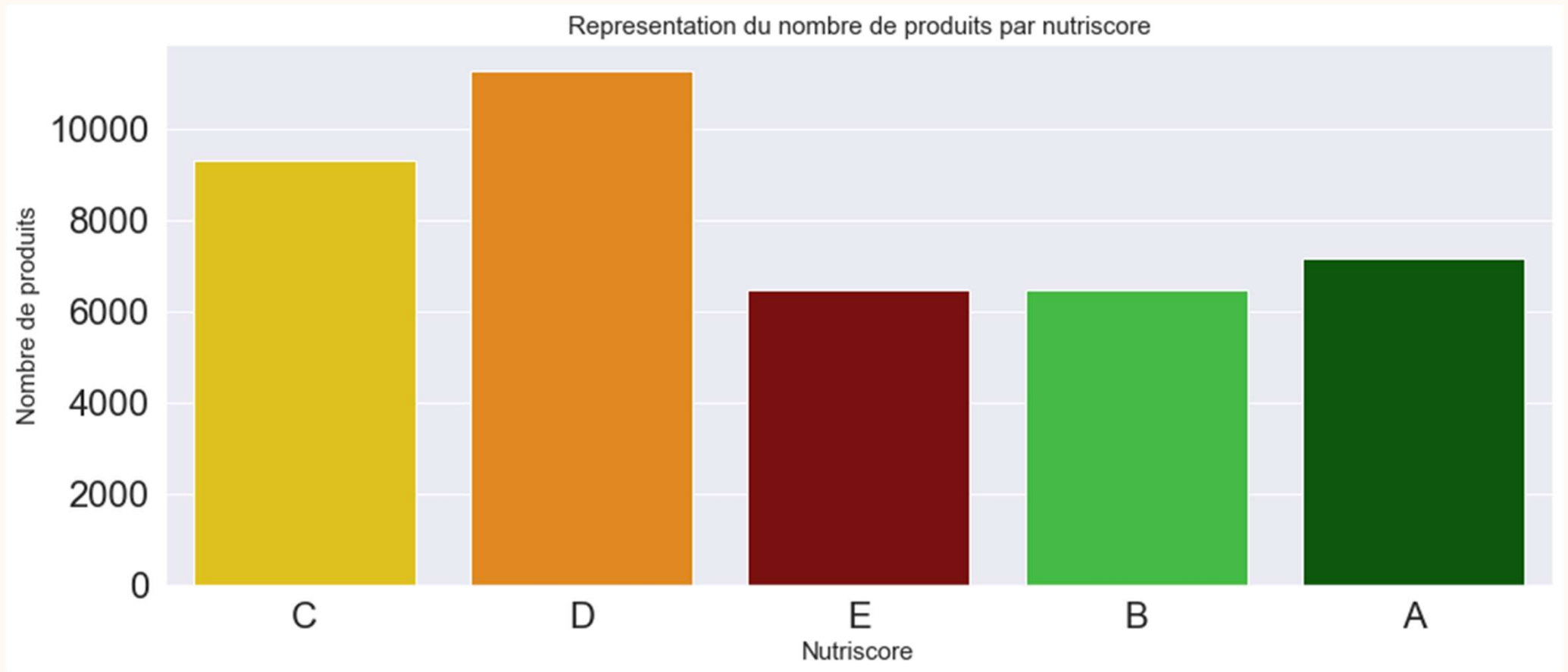


Variables catégorielles

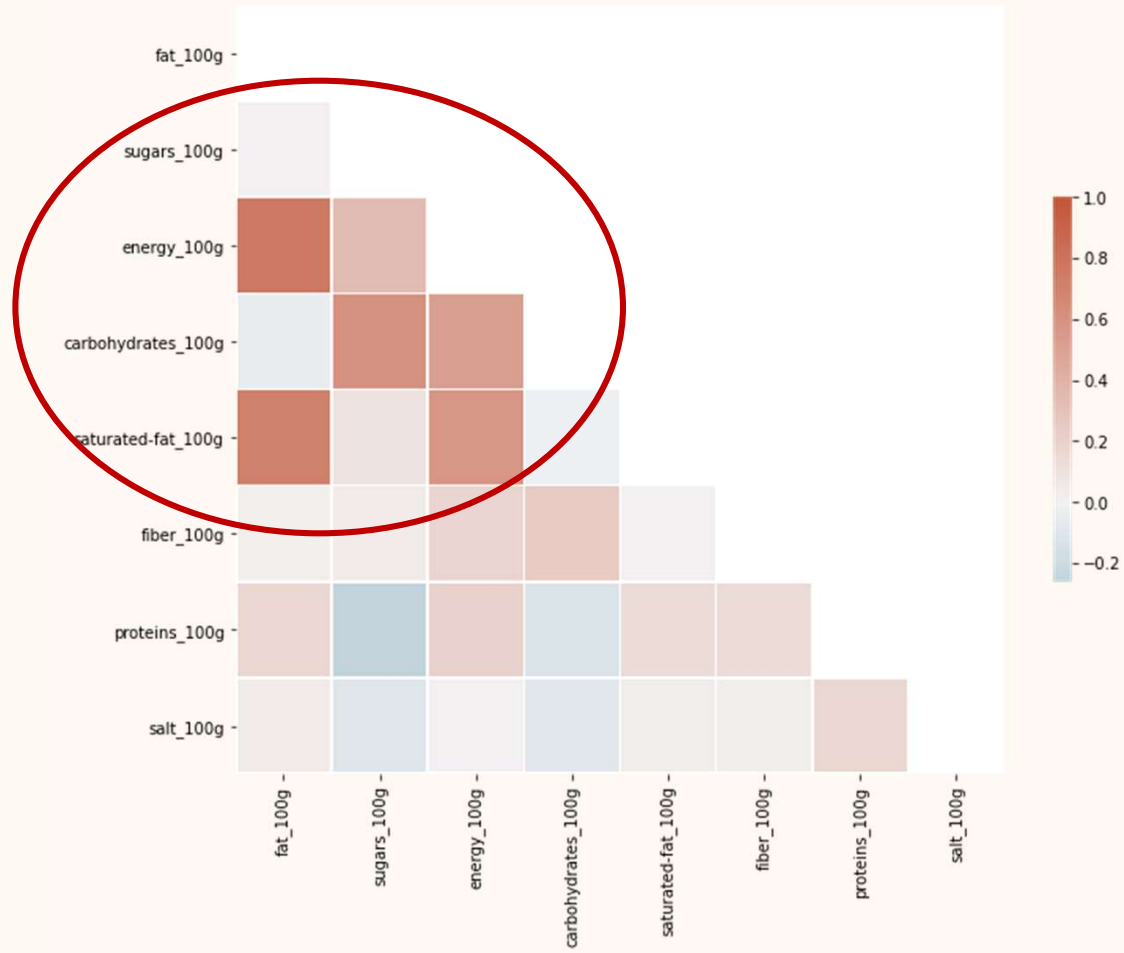
Representation du nombre de produits par pnns group



Variables catégorielles

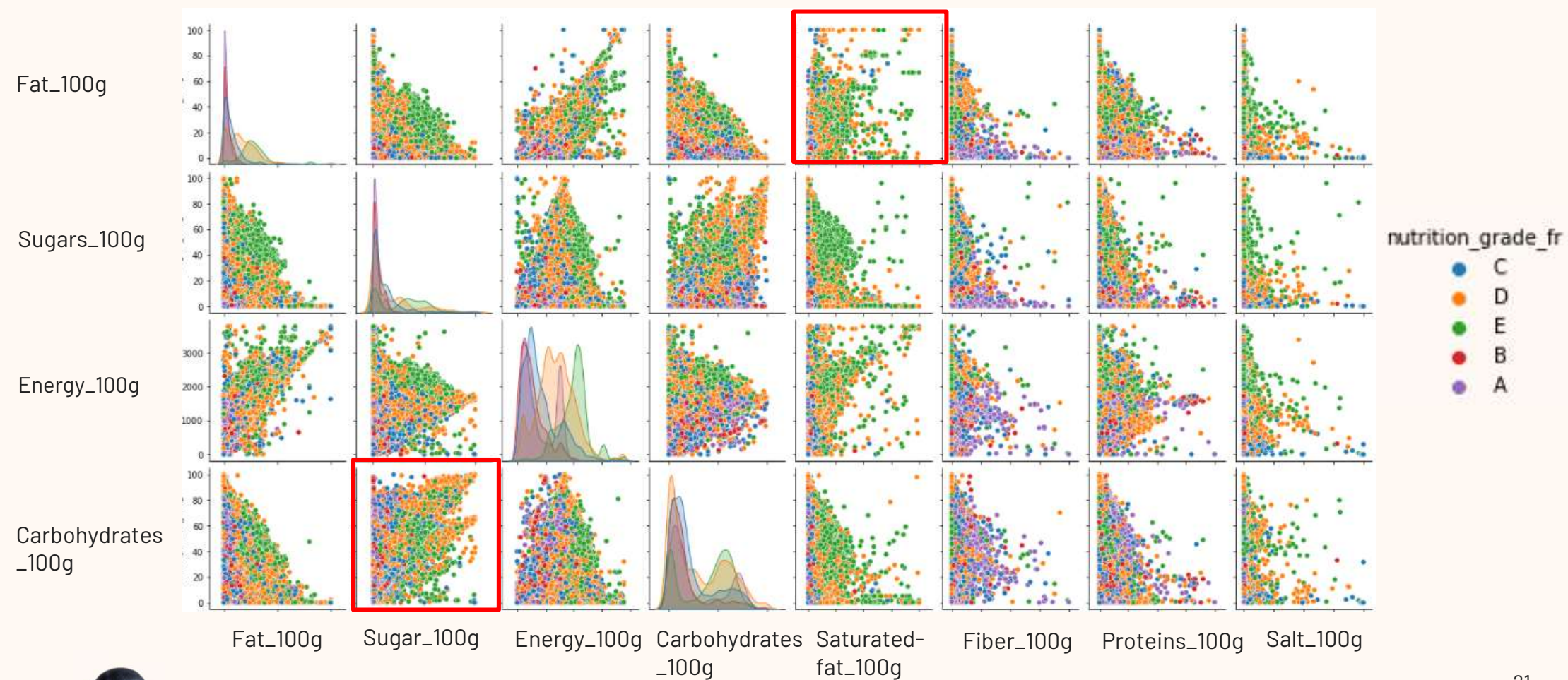


Matrice de corrélation



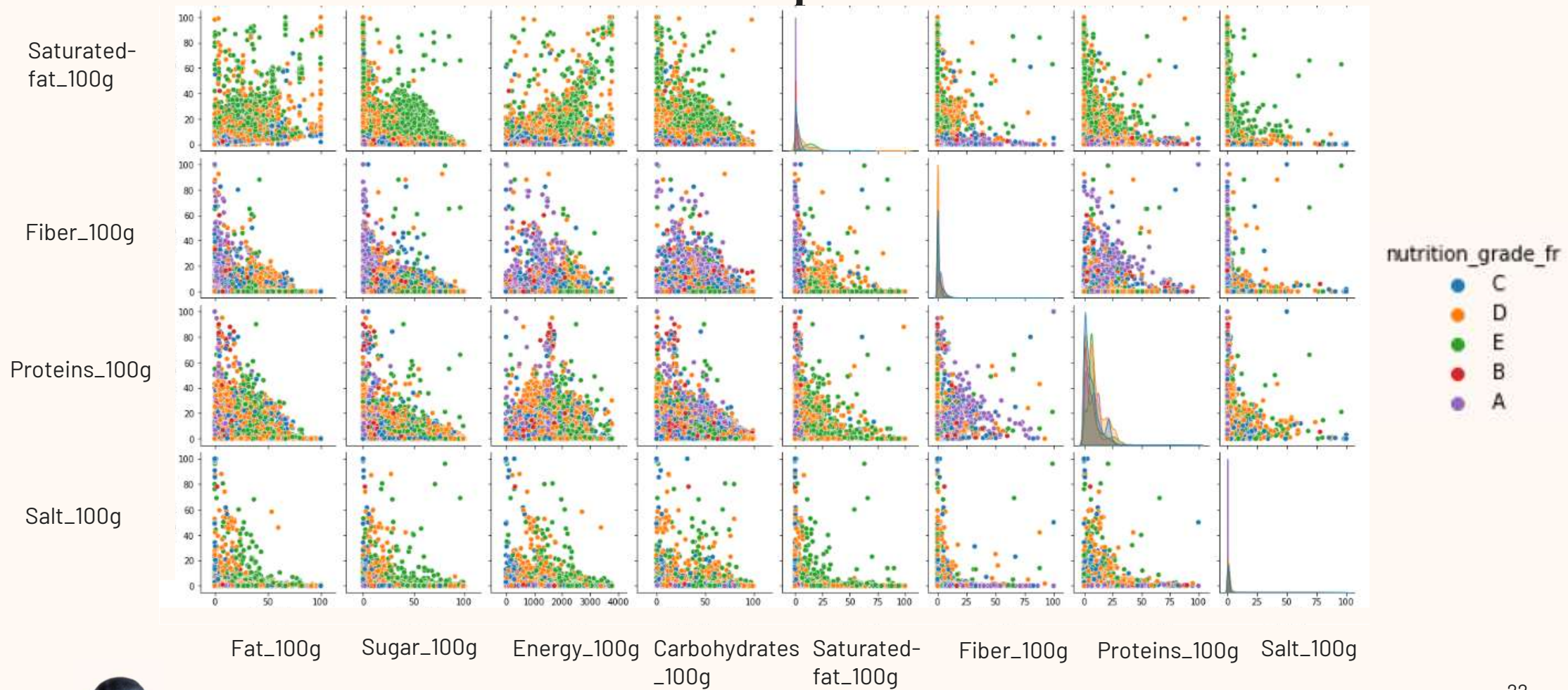


Pairplot



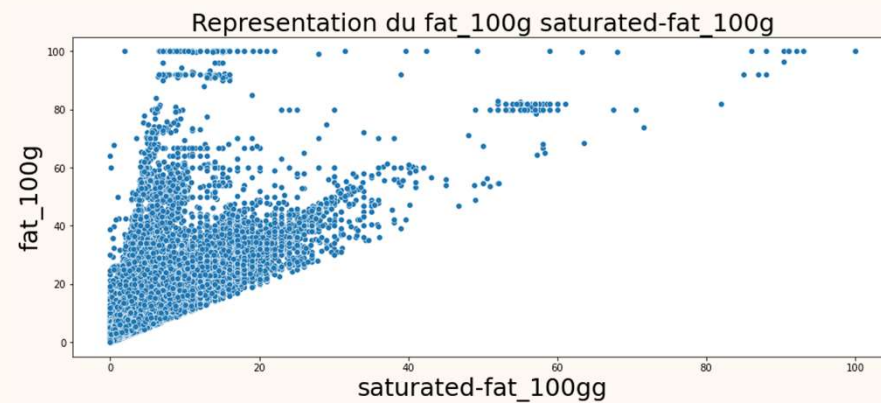
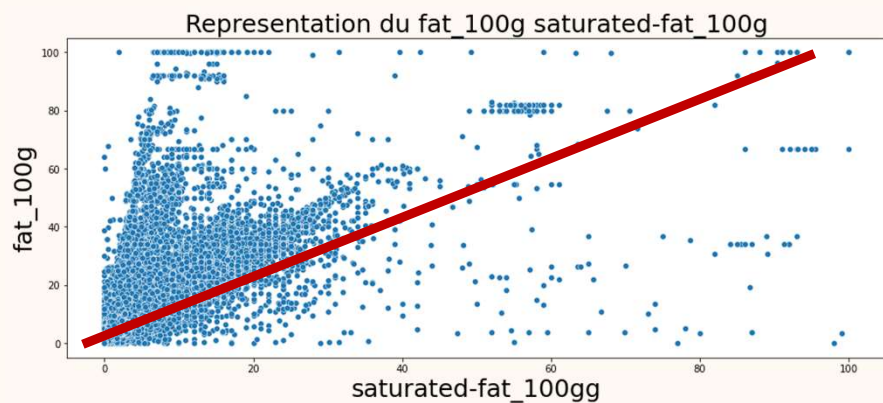
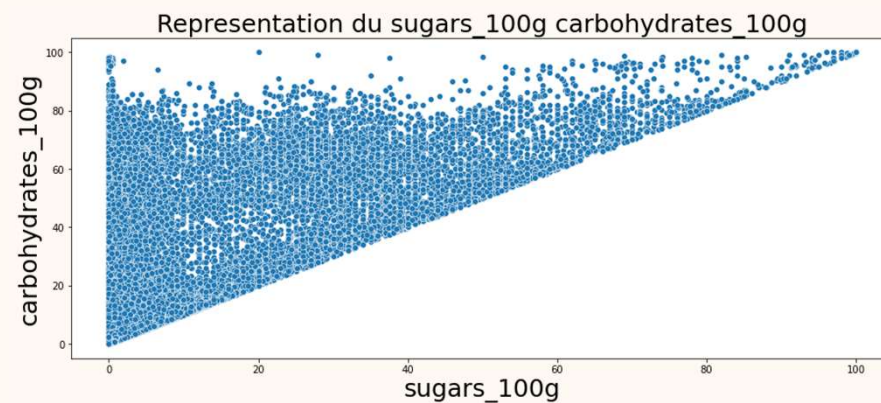
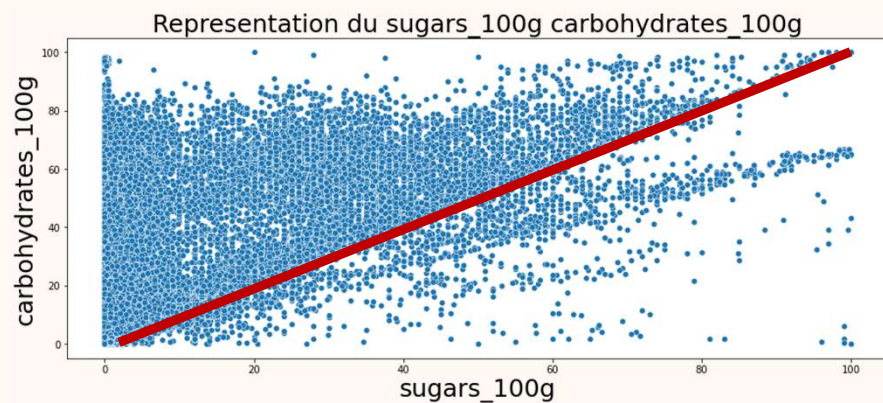


Pairplot





Relation entre les valeurs nutritionnelles





Les valeurs nutritionnelles en fonction du nutriscore

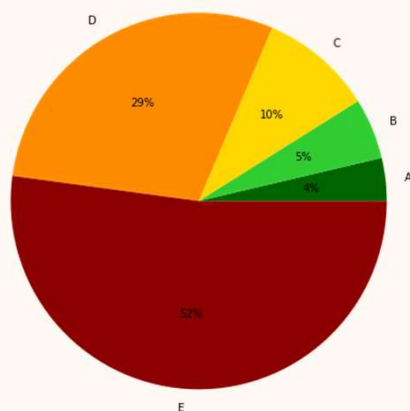
nutrition_grade_fr	fat_100g	sugars_100g	energy_100g	carbohydrates_100g	saturated-fat_100g	fiber_100g	proteins_100g	salt_100g
A	1.9	2.5	517.0	14.000000	0.4	2.555	7.35	0.1280
B	2.7	2.3	438.0	10.266667	0.8	0.500	4.00	0.5600
C	5.0	4.5	671.0	14.000000	1.7	0.300	5.00	0.7112
D	15.1	5.0	1381.0	22.000000	5.0	0.000	7.30	1.0000
E	26.9	30.0	2060.0	51.000000	14.0	0.000	6.60	0.6100



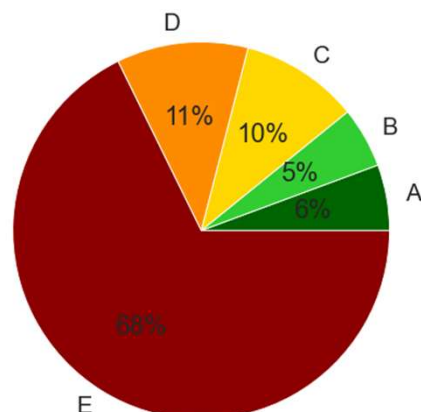


Les valeurs nutritionnelles en fonction du nutriscore

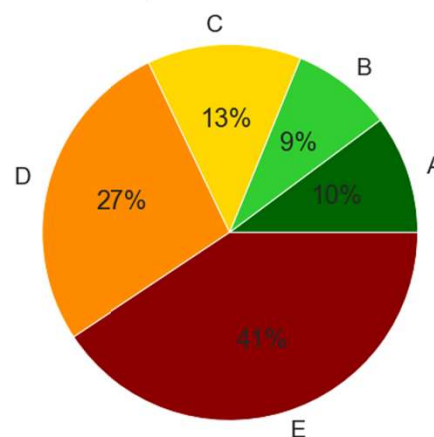
Fat_100g en fonction du nutriscore



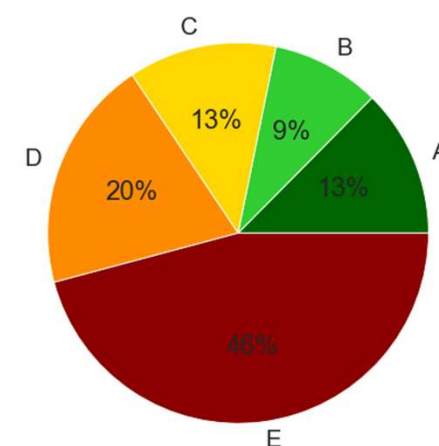
Sugars_100g en fonction du nutriscore



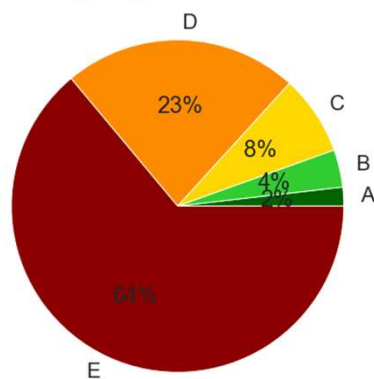
Energy_100g en fonction du nutriscore



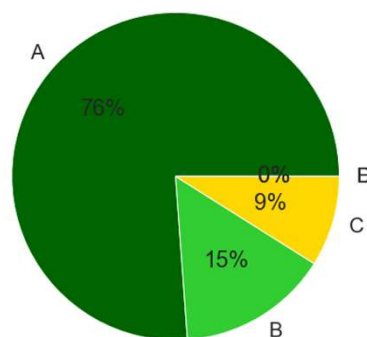
Carbohydrates_100g en fonction du nutriscore



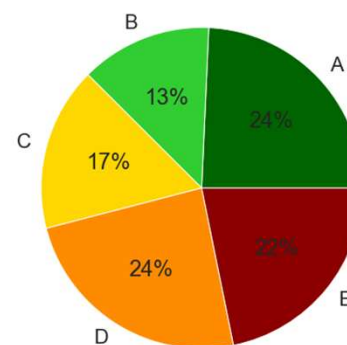
Saturated-fat_100g en fonction du nutriscore



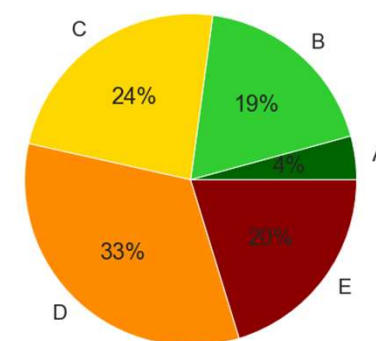
Fiber_100g en fonction du nutriscore



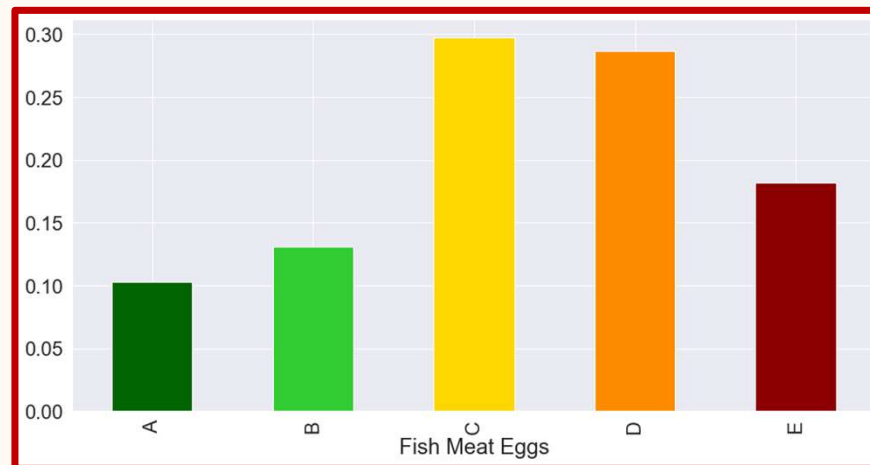
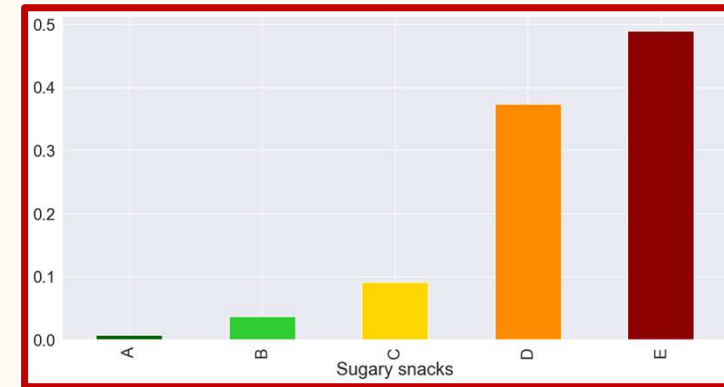
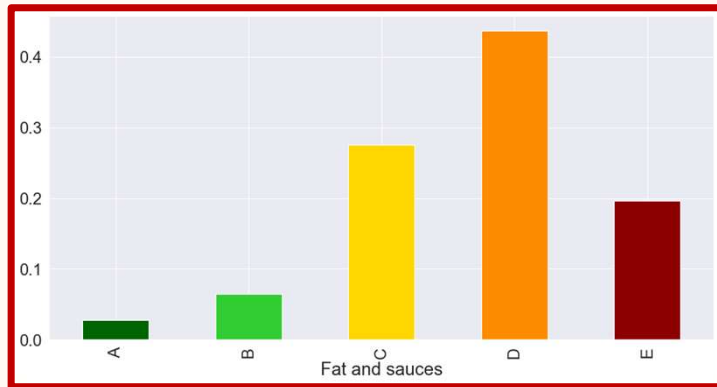
Proteins_100g en fonction du nutriscore



Salt_100g en fonction du nutriscore

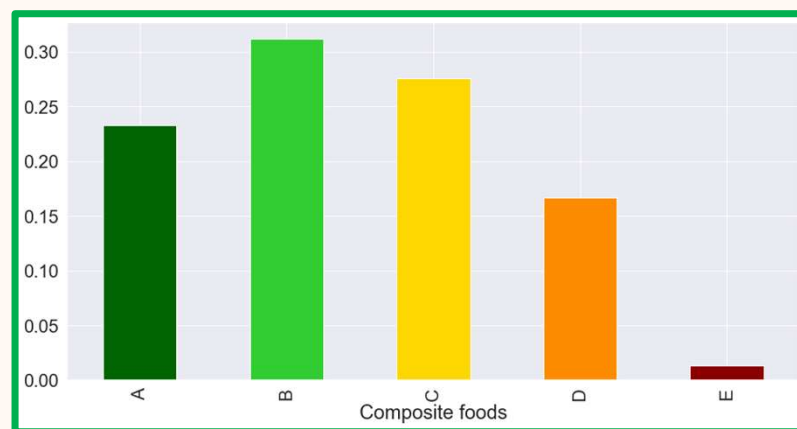
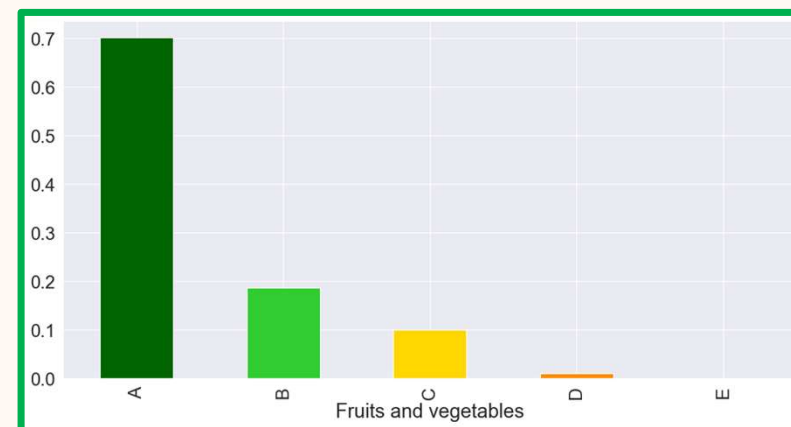
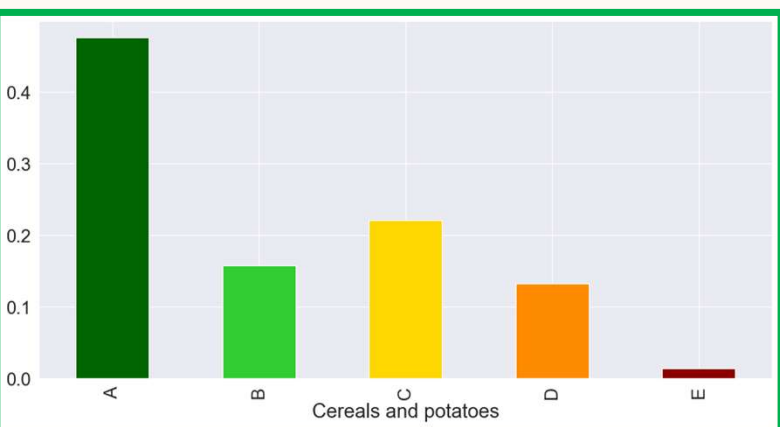


Le Pnns Group en fonction du nutriscore





Le Pnns Group en fonction du nutriscore



05

**Analyse
multivariée**





ANOVA

Hypothèses:

H0: l'hypothèse nulle: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont indépendantes

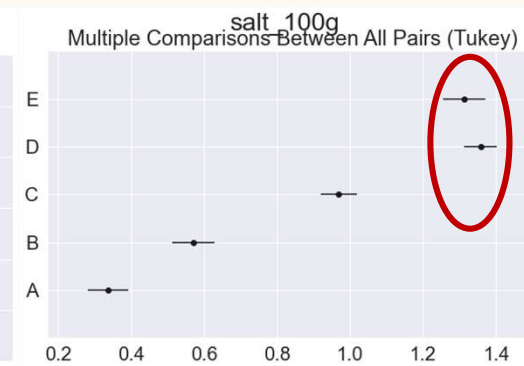
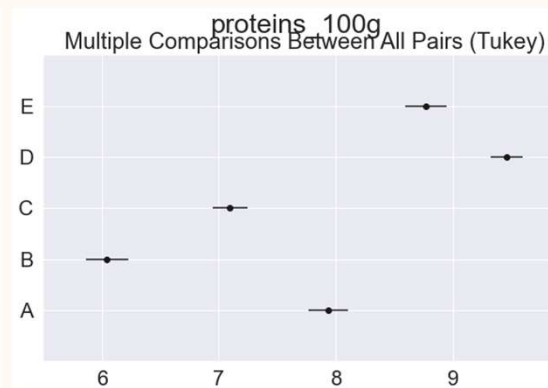
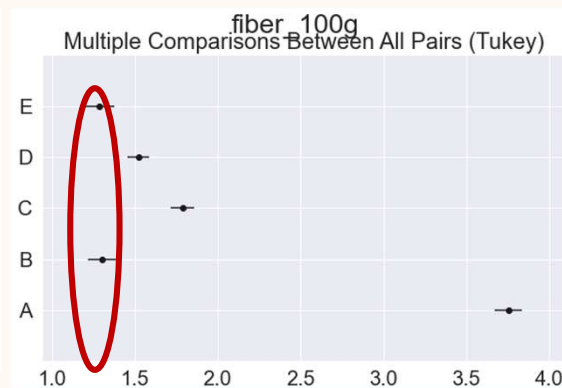
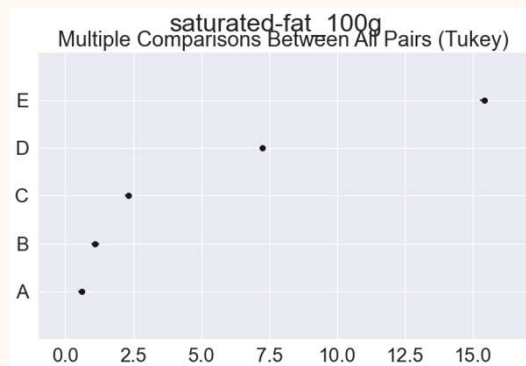
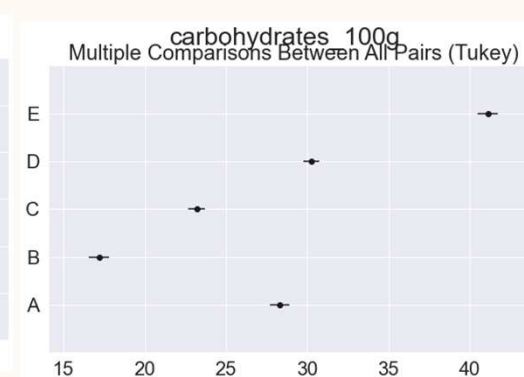
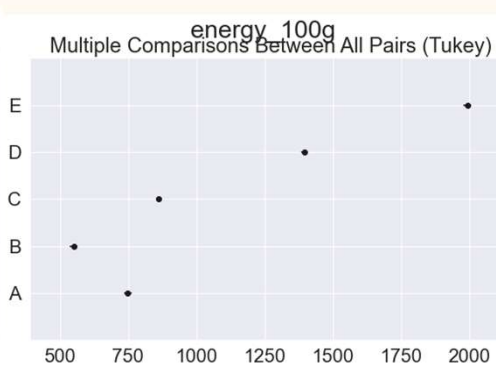
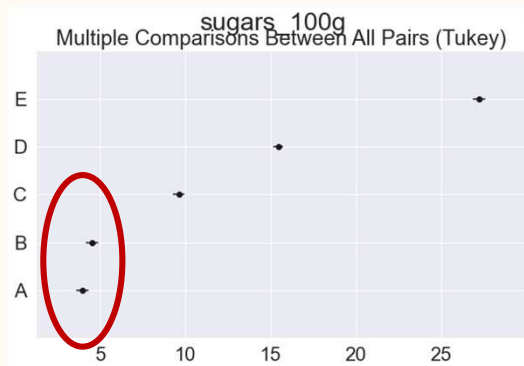
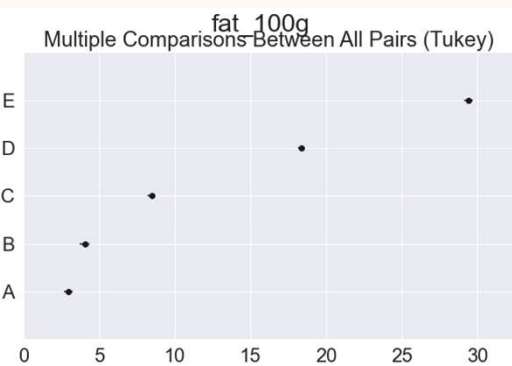
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées

```
fat_100g-----0.0
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées
sugars_100g-----0.0
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées
energy_100g-----0.0
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées
carbohydrates_100g-----0.0
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées
saturated-fat_100g-----0.0
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées
fiber_100g-----0.0
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées
proteins_100g-----8.639923604757663e-233
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées
salt_100g-----1.3709222427292858e-228
Ha: l'hypothèse alternative: Les variables nutritionnelles d'un produit et le nutriscore obtenu sont corrélées
```



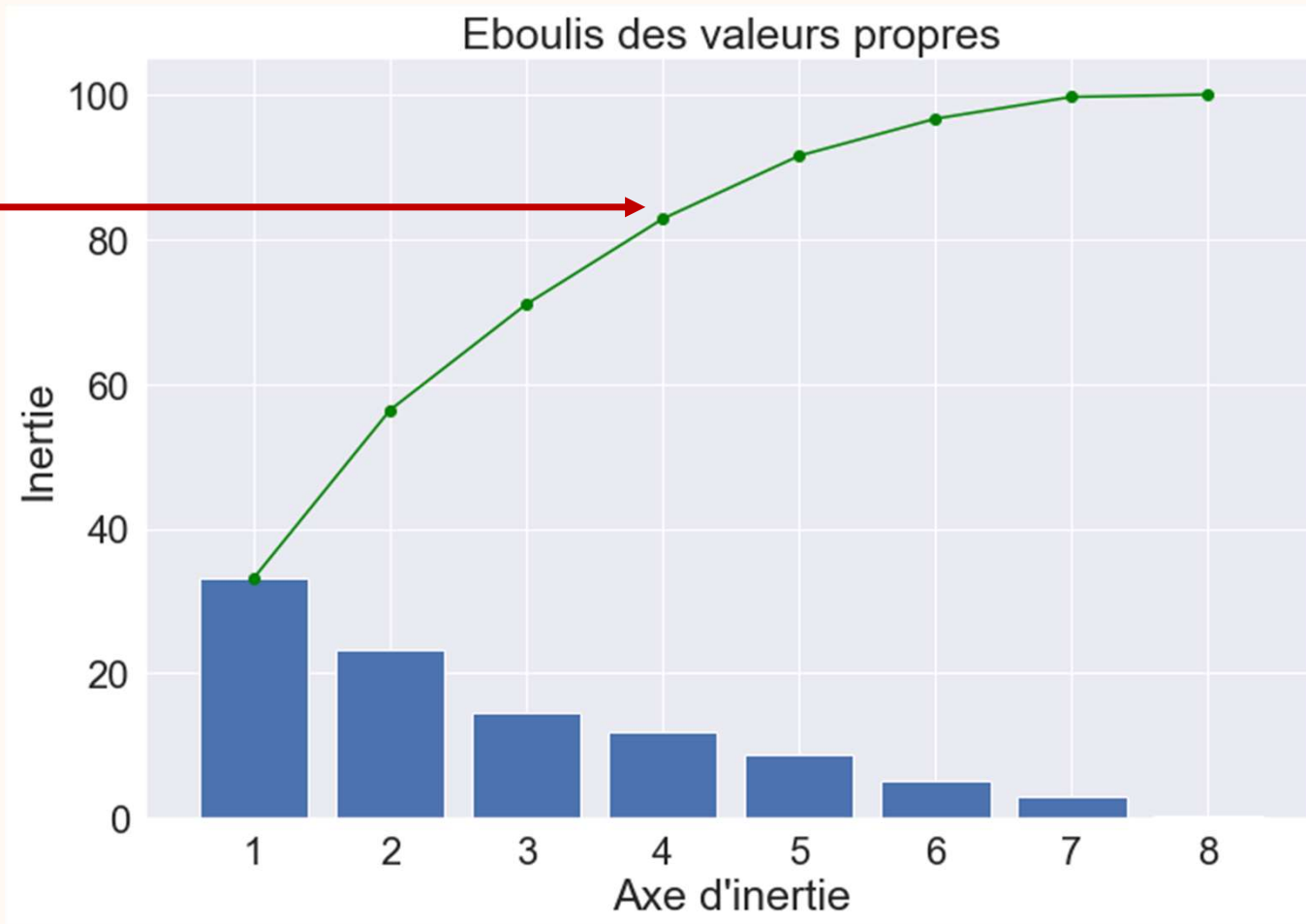


ANOVA

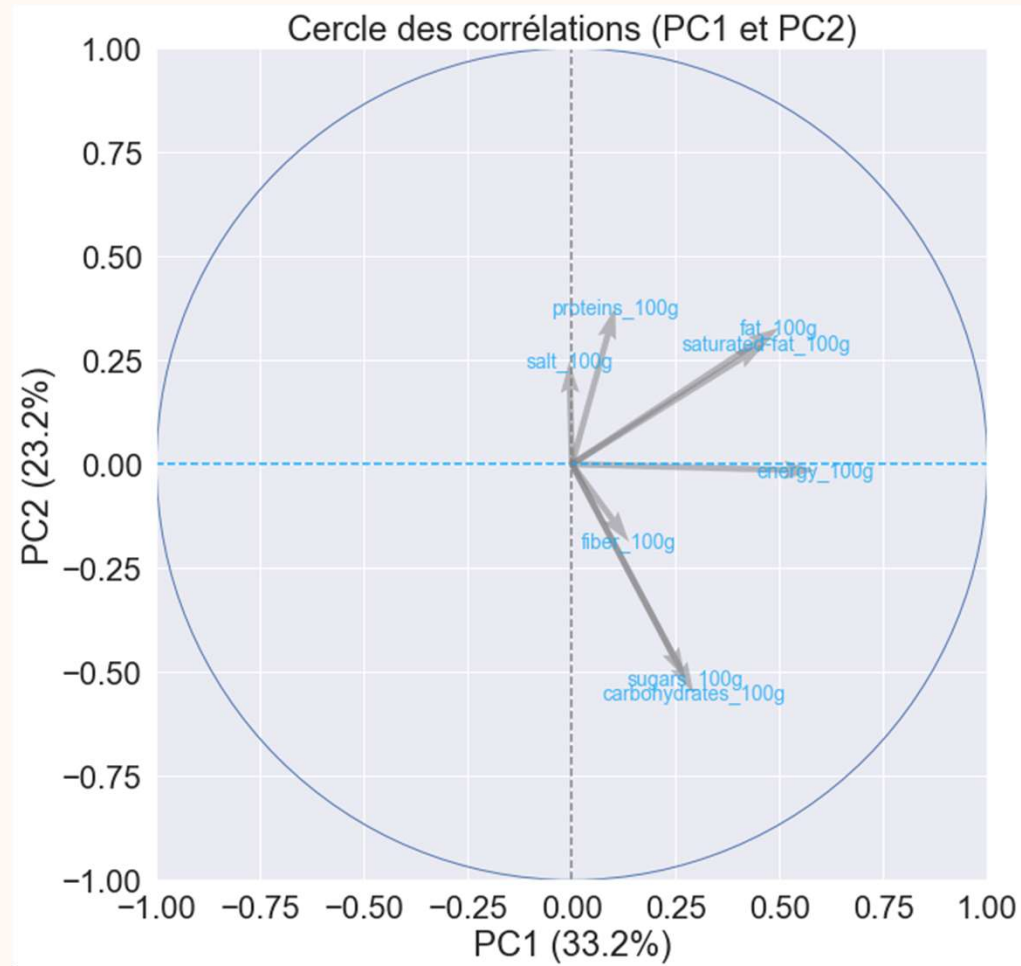


Analyse en Composantes Principales

83% de la
variances
totale des
données



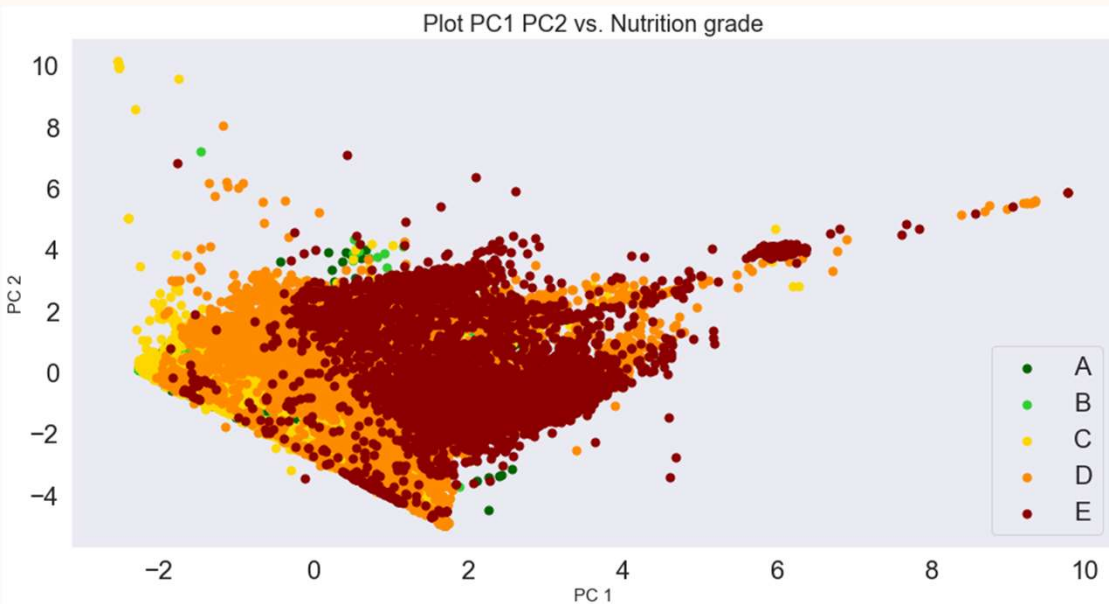
Analyse en Composantes Principales



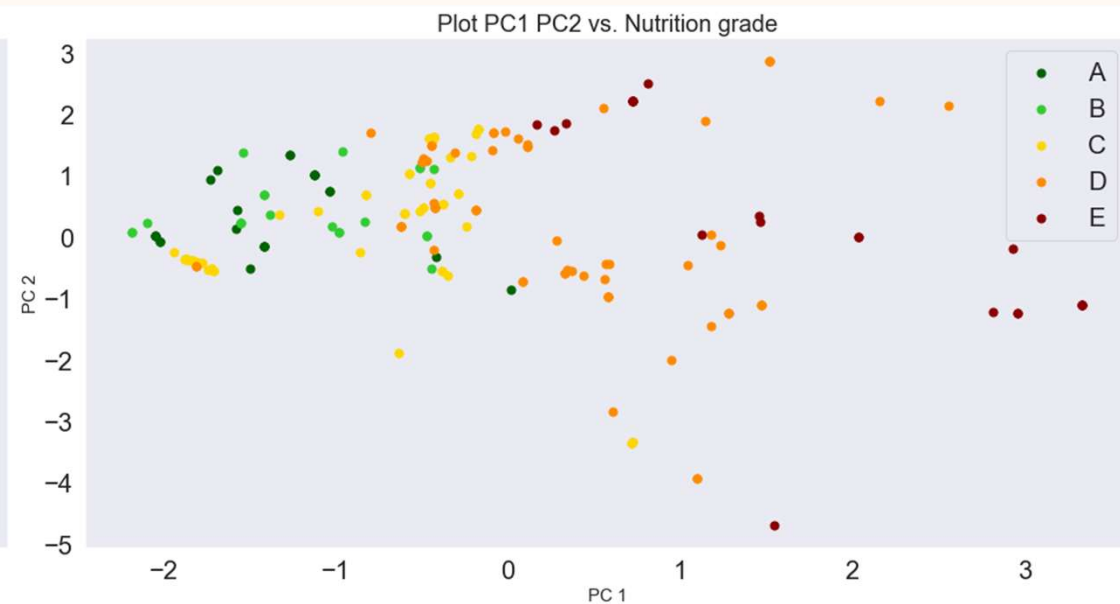


Représentation des individus

Avec l'ensemble des produits



200 produits avec la fonction Random



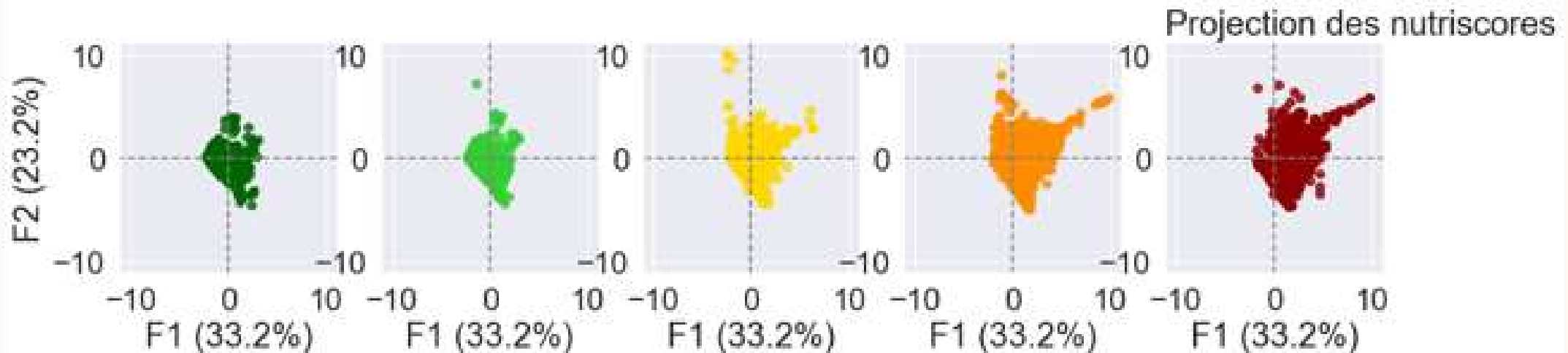


06

Conclusion



Interprétation







Les limites

- Données manquantes pour les fruits et légumes, légumineuses et fruits à coque (g/100g)
- Données également manquantes pour les fibres (imputation des valeurs manquantes par 0)

Les Perspectives

- Compléter l'analyse avec le **Test du Chi-deux** entre le nutriscore et le pnns group, tests non paramétriques Le test de Mann-Whitney ou Kruskal-Wallis entre l'énergie et les autres valeurs nutritionnelles
-  • Déployer un model de machine learning et analyser la performance
-  • Réduire nos variables à 5 (fat,energy,sugar,carbohydrates et saturated fat) , faire l'ACP et entrainer notre model

Merci!

