Concevez une application au service de la santé publique

Santé publique France

Sommaire :

1 : Rappel du projet

Contexte :

- L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d’applications en lien avec l'alimentation.

Données :

- Issues du site Open Food Facts : https://world.openfoodfacts.org/
- Aide explicative pour les variables : https://world.openfoodfacts.org/data/data-fields.txt

2 : Présentation de l'idée d'application choisie

Mieux manger c'est bien mais tout en protégeant l'environnement c'est encore mieux.

1 - Le client scanne le code barre de son produit

Nutriscore :

- Si il est présent dans la base de donnée le Nutriscore est déjà présent.
- Sinon il est déterminé via du Machine Learning (KNN).

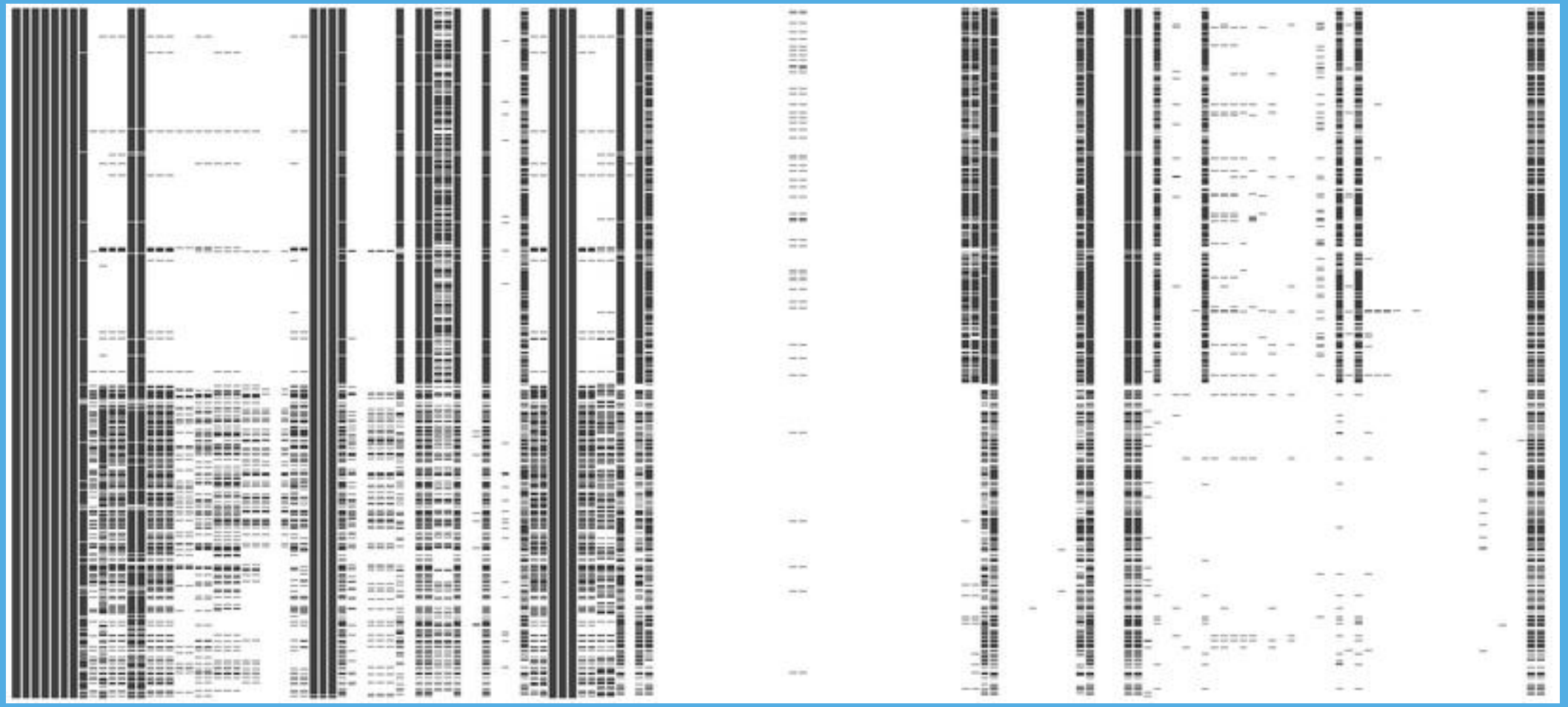
Création d'un Environnement Grade :

- Synthése des données (Huile de Palme, lieux de fabrication, origines des ingrédients, présence d'additifs, caractéristiques de l'emballage et label Bio).
- Si il est présent dans la base de donnée l'Environnement Grade est déjà déterminé via KMeans.
- Si il n'est pas présent il est automatiquement calculé via le même procédé.

2 - Le programme donne au client le meilleur produit (nutritif et environnemental de cette catégorie).

3 : Opérations de nettoyage effectuées

- Données :
  - **320000** produits pour **162** variables informatives (informations textes et données nutritives)
  - **76%** de données non renseignées dans ce fichier.



Opérations de nettoyages :

Etapes	Action	NbColonne	NbLigne	PourcentageNaN
Etape 1	Ouverture du Fichier	162	320772	76
Etape 2	Suppression des colonnes vides	146	320772	74
Etape 3	Suppression des lignes vides	146	265302	71
Etape 4	Etude des Variables Info Générales	139	261822	74
Etape 5	Etude des Variables Tags + Centrage sur produits vendus en France	121	66822	75
Etape 6	Etude des Variables Ingrédients	118	66822	75
Etape 7	Etude des Variables Données Diverses	106	66822	79
Etape 8	Etude des Variables Données Nutritives	97	66820	78
Etape 9	Remplissage des Colonnes Textes + Suppression des produits sans groupes PNNS2	91	40003	70
Etape 10	Création des variables pour l'EnvironnementScore + Recentrage sur les variables utiles au Nutriscore	27	40003	2
Etape 11	Remplissage des colonnes quantitatives par 0	27	40003	0
Etape 12	Suppression des valeurs abérrantes (négatives, +100 et somme sup à 100)	27	39752	0
Etape 13	Suppression manulle des Outliers	27	39741	0
Etape Final	Suppression des variables inutiles ou en doubles	23	39741	0

4 : Analyse Univariée

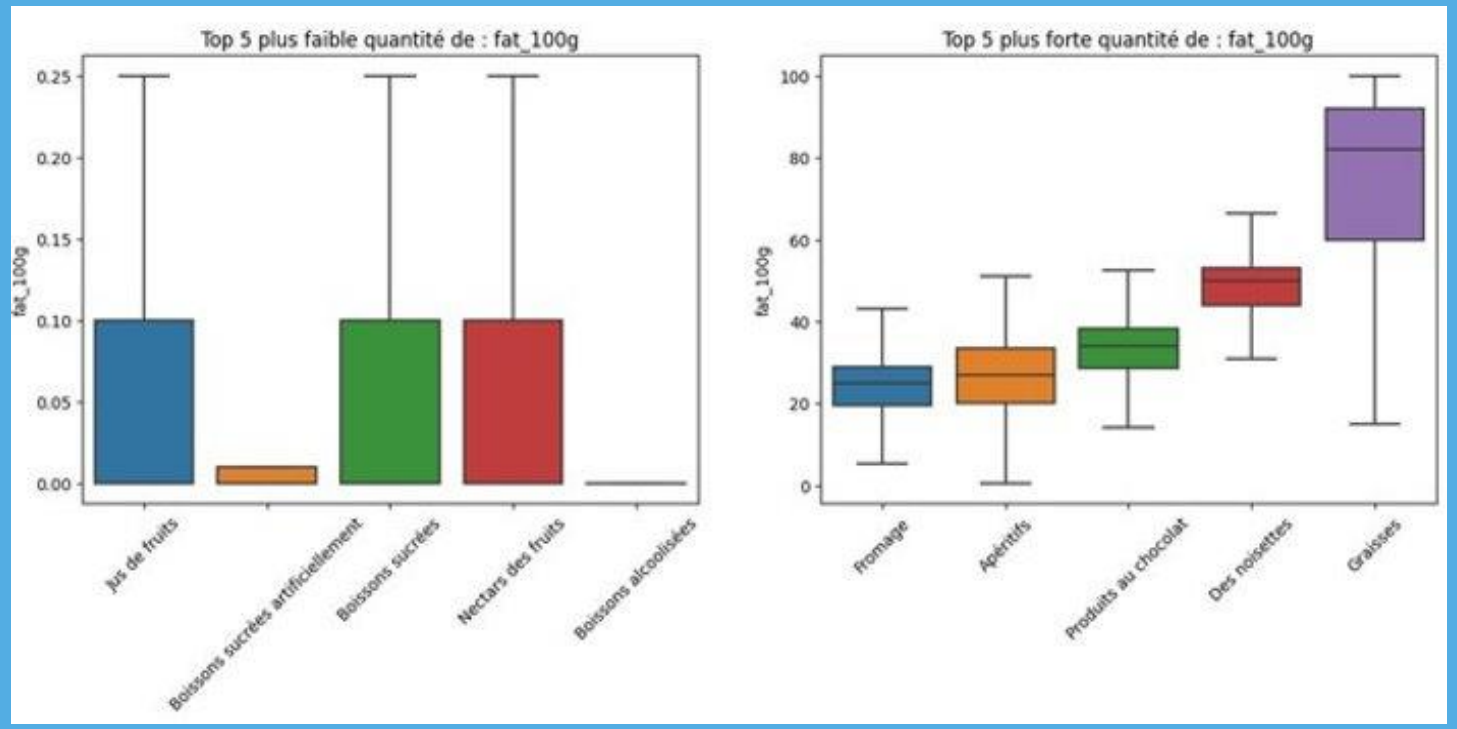
Description des données :

	saturated-fat_100g	energy_100g	sodium_100g	salt_100g	proteins_100g	sugars_100g	fat_100g	carbohydrates_100g	fiber_100g
count	39752.000000	39752.000000	39752.000000	39752.000000	39752.000000	39752.000000	39752.000000	39752.000000	39752.000000
mean	5.346483	1095.645502	0.339570	0.862509	7.663352	12.773031	12.588355	25.656310	1.670217
std	8.283837	787.712964	0.553210	1.405156	7.291520	18.331290	16.766193	27.140387	3.218387
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.300000	402.000000	0.027559	0.070000	1.600000	1.000000	0.600000	2.600000	0.000000
50%	1.900000	1006.000000	0.224409	0.570000	6.000000	4.000000	5.500000	12.300000	0.100000
75%	7.400000	1644.000000	0.472441	1.200000	11.000000	15.800000	20.000000	51.500000	2.300000
max	100.000000	15481.000000	30.000000	76.200000	86.000000	100.000000	100.000000	100.000000	86.200000

- On observe des 0 ce qui est normal on a pas toujours tous les ingrédients pour chaque produit.
- En violet nous avons les ingrédients qui entrent dans la création du Nutriscore puis du Nutrigrade.

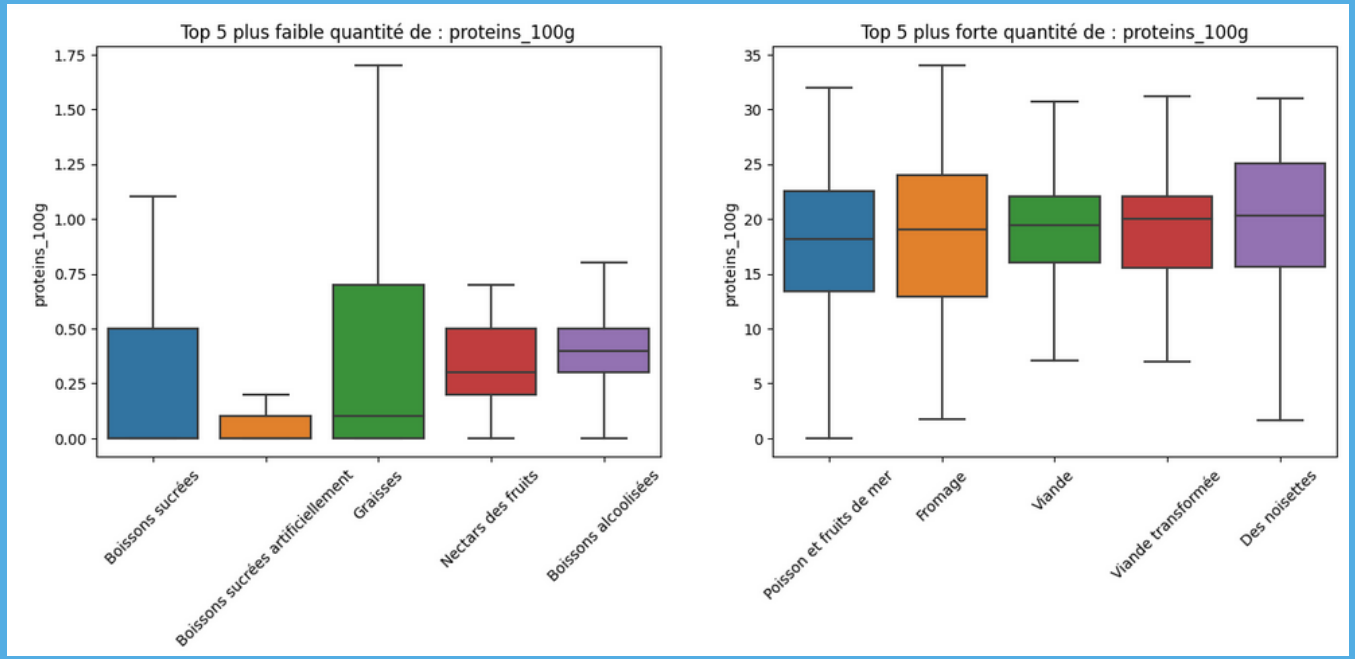
Etude par catégorie :

	saturated-fat_100g	energy_100g	sodium_100g	salt_100g	proteins_100g	sugars_100g	fat_100g	carbohydrates_100g	fiber_100g
MoyenneMaximum	Graisses	Graisses	Viande transformée	Viande transformée	Viande transformée	Bonbons	Graisses	Céréales du petit-déjeuner	Produits salés et gras
MoyenneMinimum	Nectars des fruits	Boissons sucrées artificiellement	Nectars des fruits	Nectars des fruits	Boissons sucrées artificiellement	Graisses	Boissons sucrées artificiellement	Graisses	Graisses



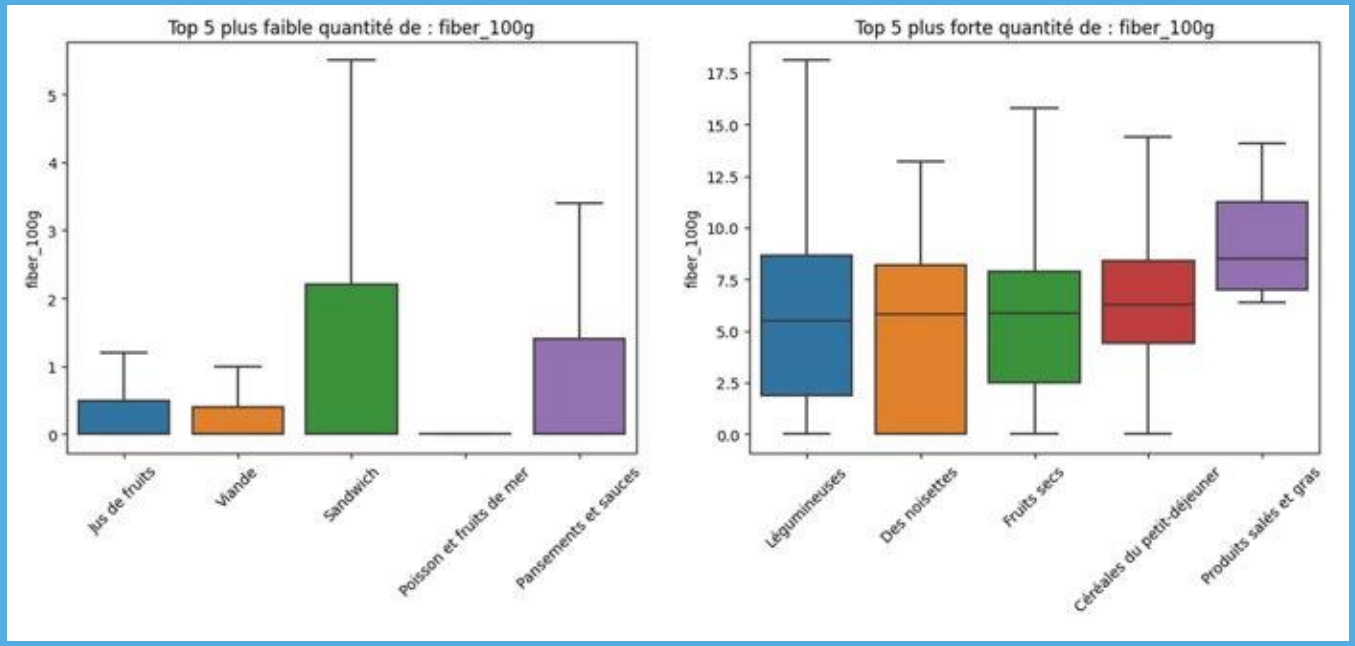
Analyse :

- Les produits les plus gras sont les graisses, les noisettes et le chocolat.
- Les produits les moins gras sont les jus de fruits et les alcools.



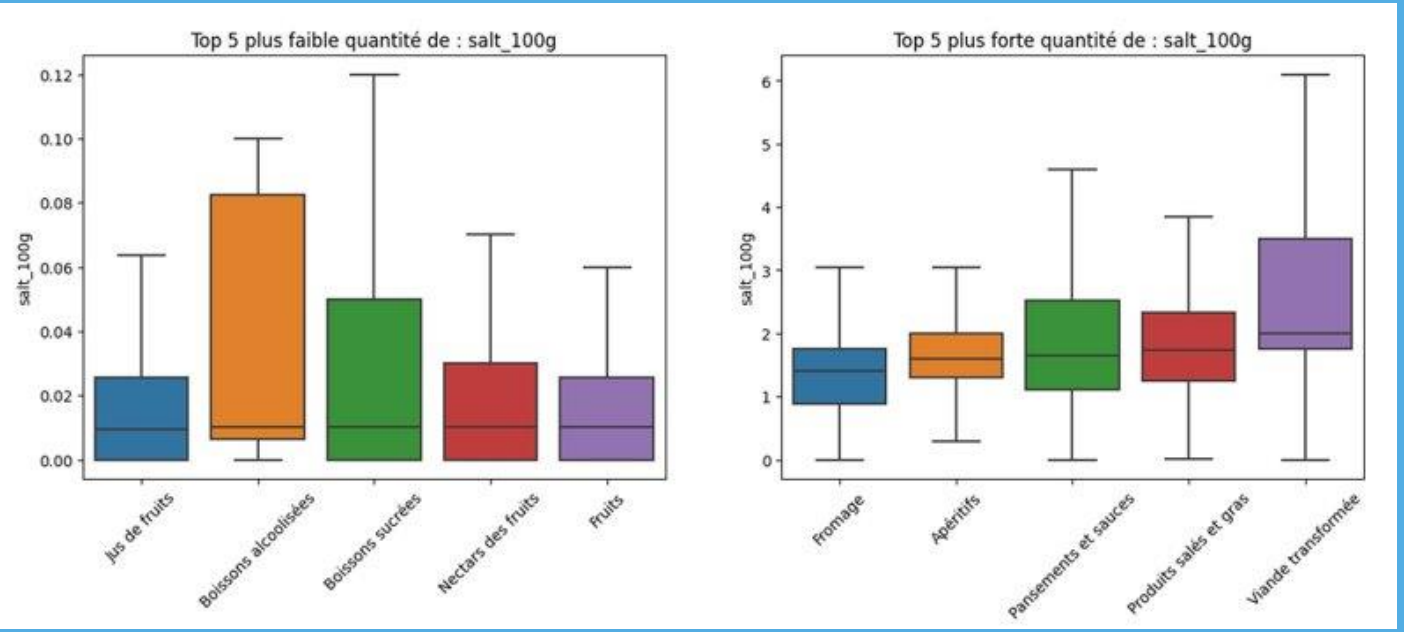
Analyse :

- Les produits avec le plus de protéines sont les noisettes et les viandes.
- Ceux avec le moins de protéines sont les boissons et les graisses.



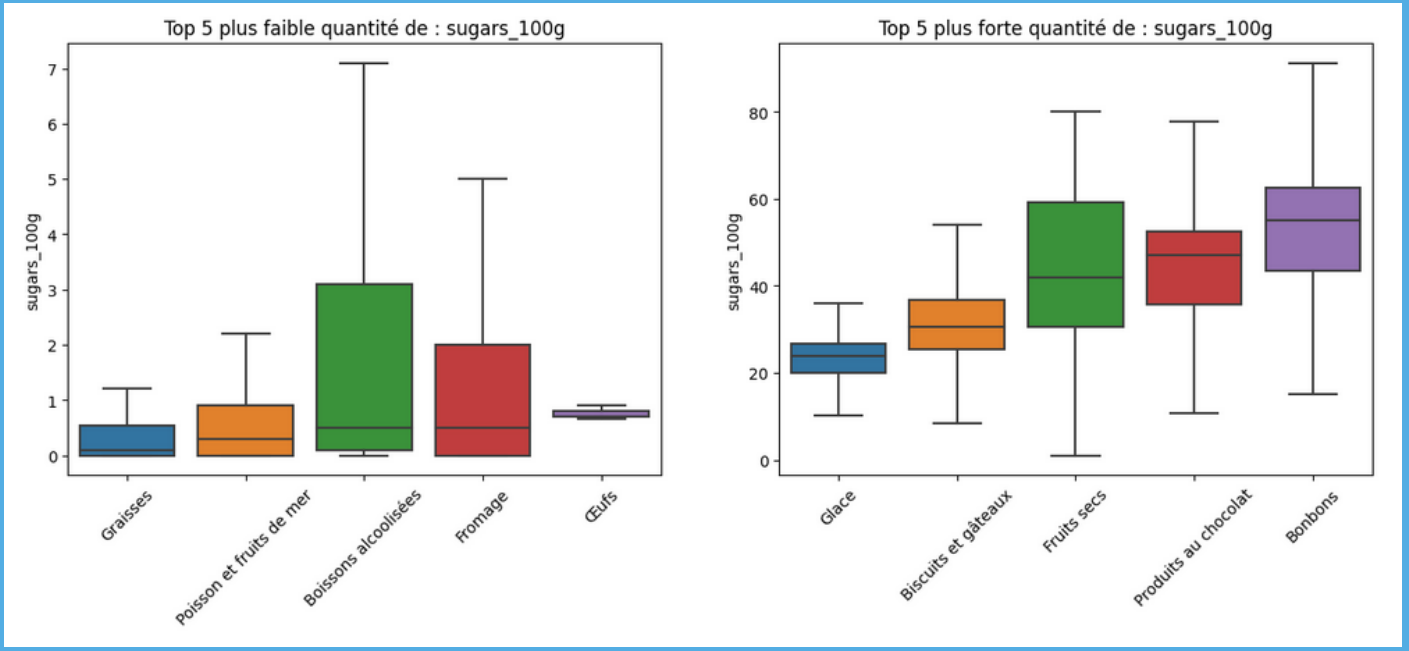
Analyse :

- Les produits avec le plus de fibres sont les produits salées et gras, les céréales et les fruits secs.
- Ceux avec le moins de fibres sont les jus de fruits et les viandes.



Analyse :

- Les produits les plus salés sont les viandes transformées et les produits salés et gras.
- Les produits les moins salés sont les boissons fruités ou alcoolisés.



Analyse :

- Les produits les plus sucrés sont les bonbons, les produits au chocolat et les fruits secs.
- Les produits les moins sucrés sont les graisses, les poissons et les fruits de mer.

5 : Analyse Multivariée

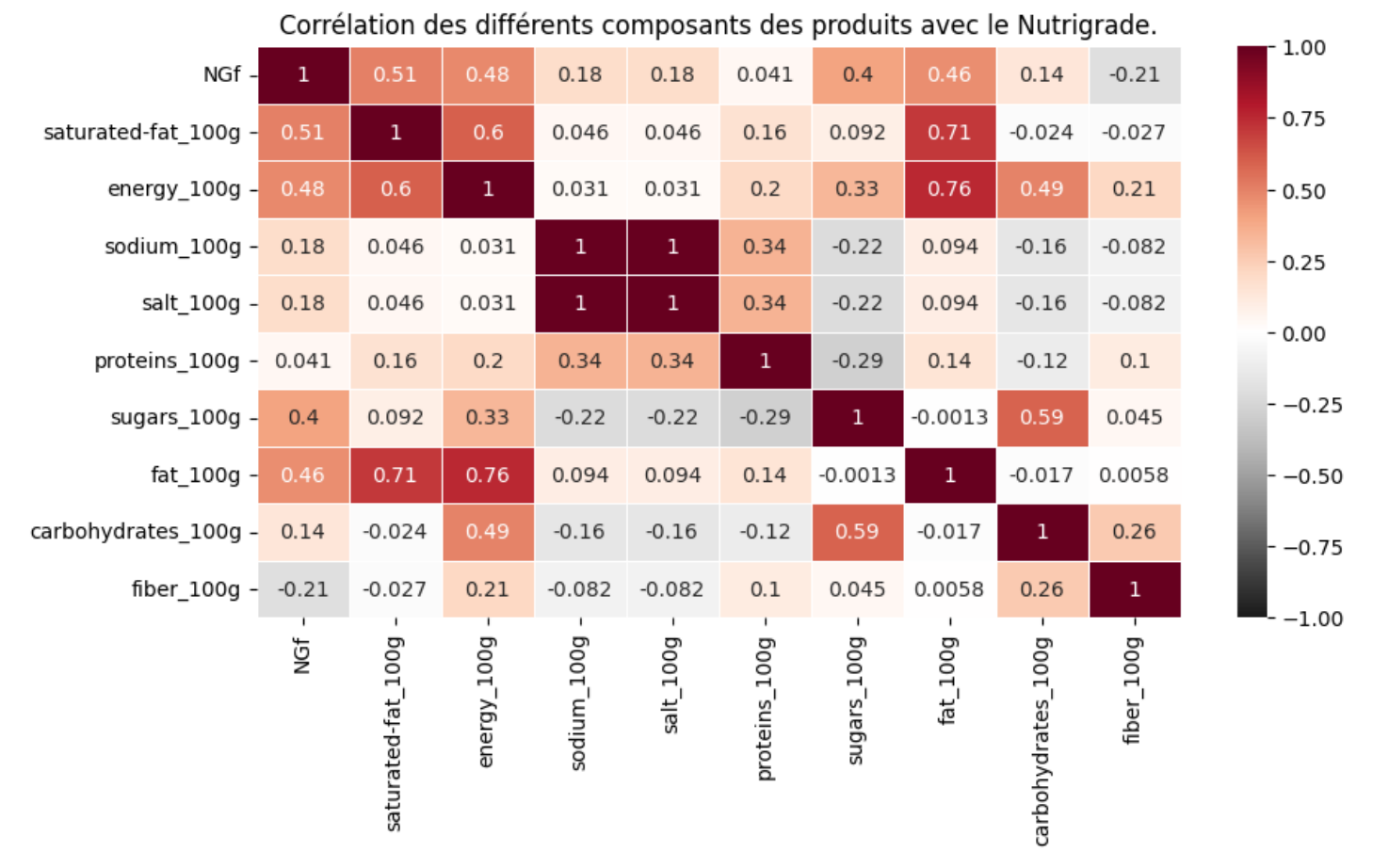
Etude de la corrélation des variables nutritives vs le Nutrigrade :

NGf	1.000000
saturated-fat_100g	0.507537
energy_100g	0.480634
fat_100g	0.464708
sugars_100g	0.401326
sodium_100g	0.184367
salt_100g	0.184366
carbohydrates_100g	0.137637
proteins_100g	0.040820
fiber_100g	-0.208878

Analyse :

- Le Nutrigrade est principalement corrélé avec les graisses et le sucre.
- Il est aussi corrélé avec l'énergie, mais cette variable est un calcul qui prend en compte principalement les valeurs de gras.
- Alors que le calcul du NutriScore prend en compte les fibres et les protéines il semble ne pas y avoir de corrélation avec le NGf.

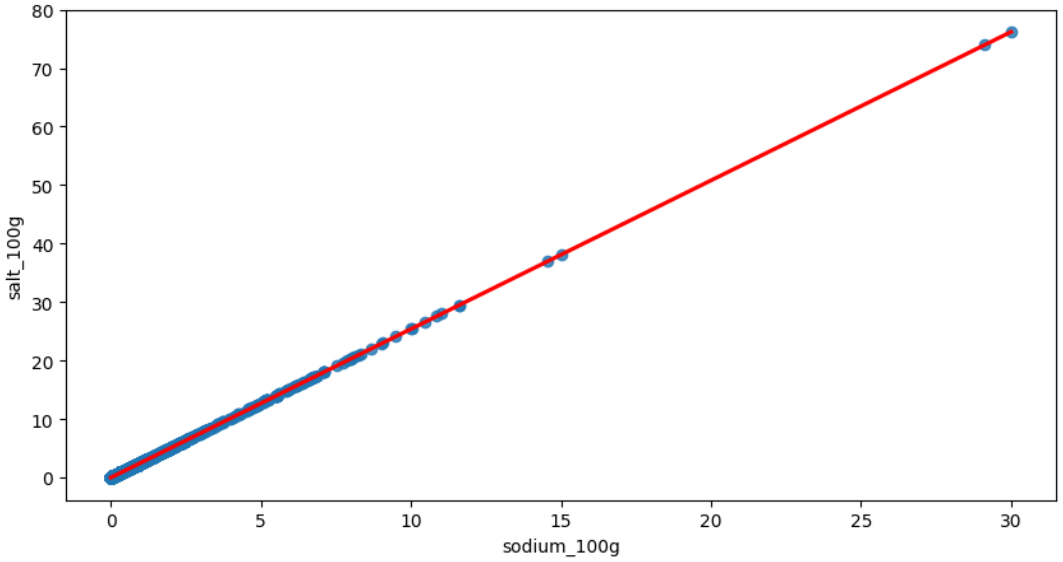
Etude de la corrélation entre toutes les variables :



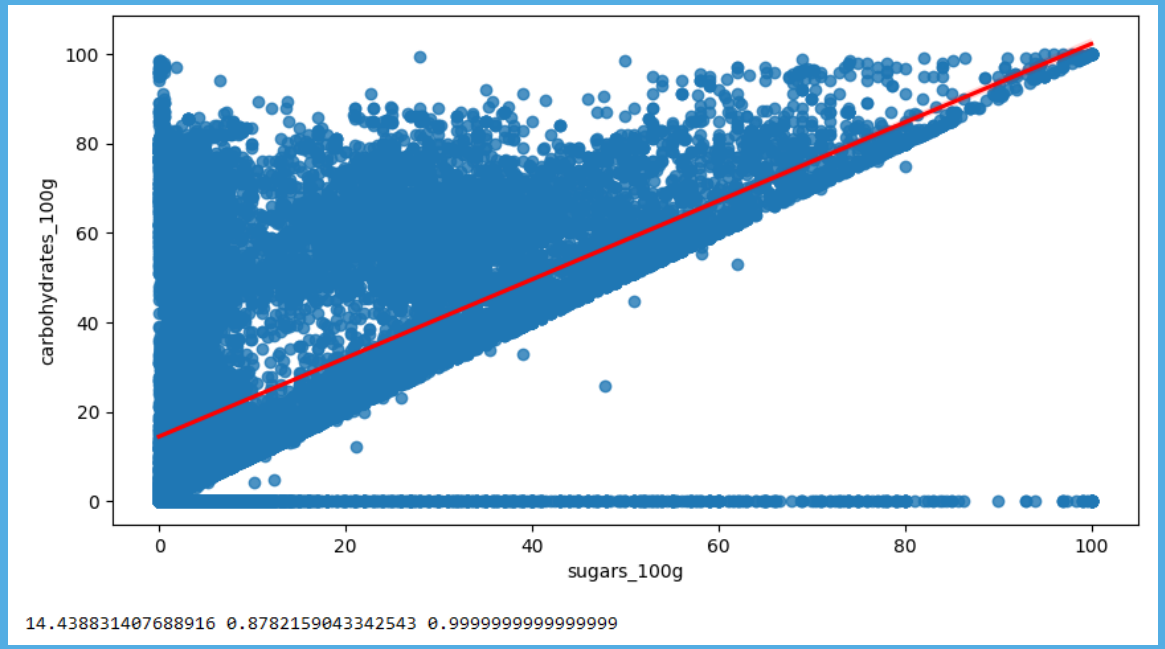
Analyse :

- En plus des corrélations avec le NGf.
- Il y a une forte corrélation entre le gras et le gras saturé (0.71).
- Une corrélation parfaite entre le sel et le sodium (1).
- Une corrélation entre le gras et l'énergie.
- Entre les fibres et les glucides.

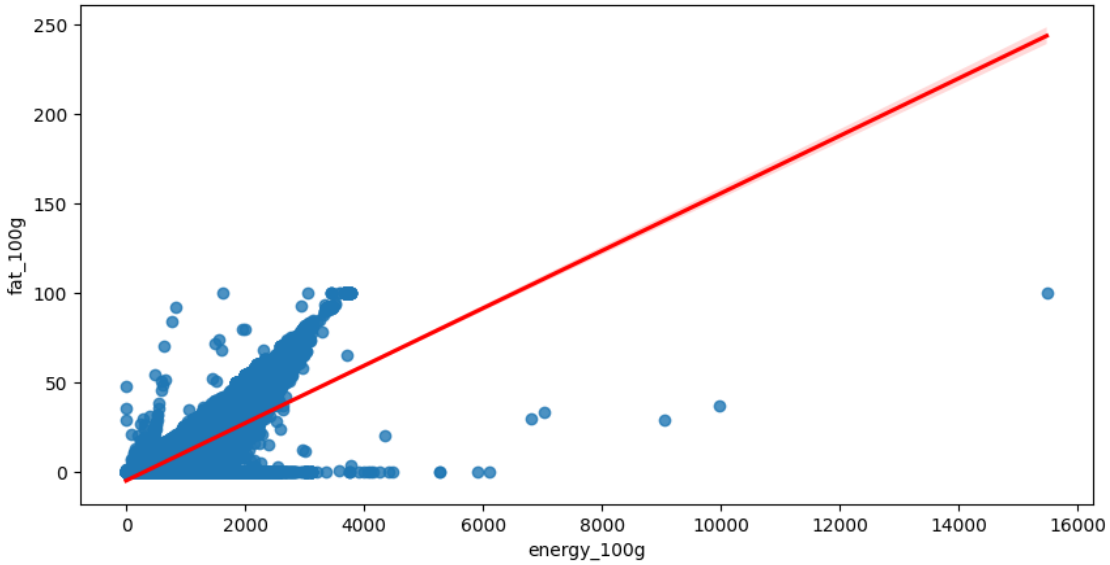
Etude graphique de ces variables :



-1.142885096783175e-06 2.5400054930570626 1.0



14.438831407688916 0.8782159043342543 0.9999999999999999

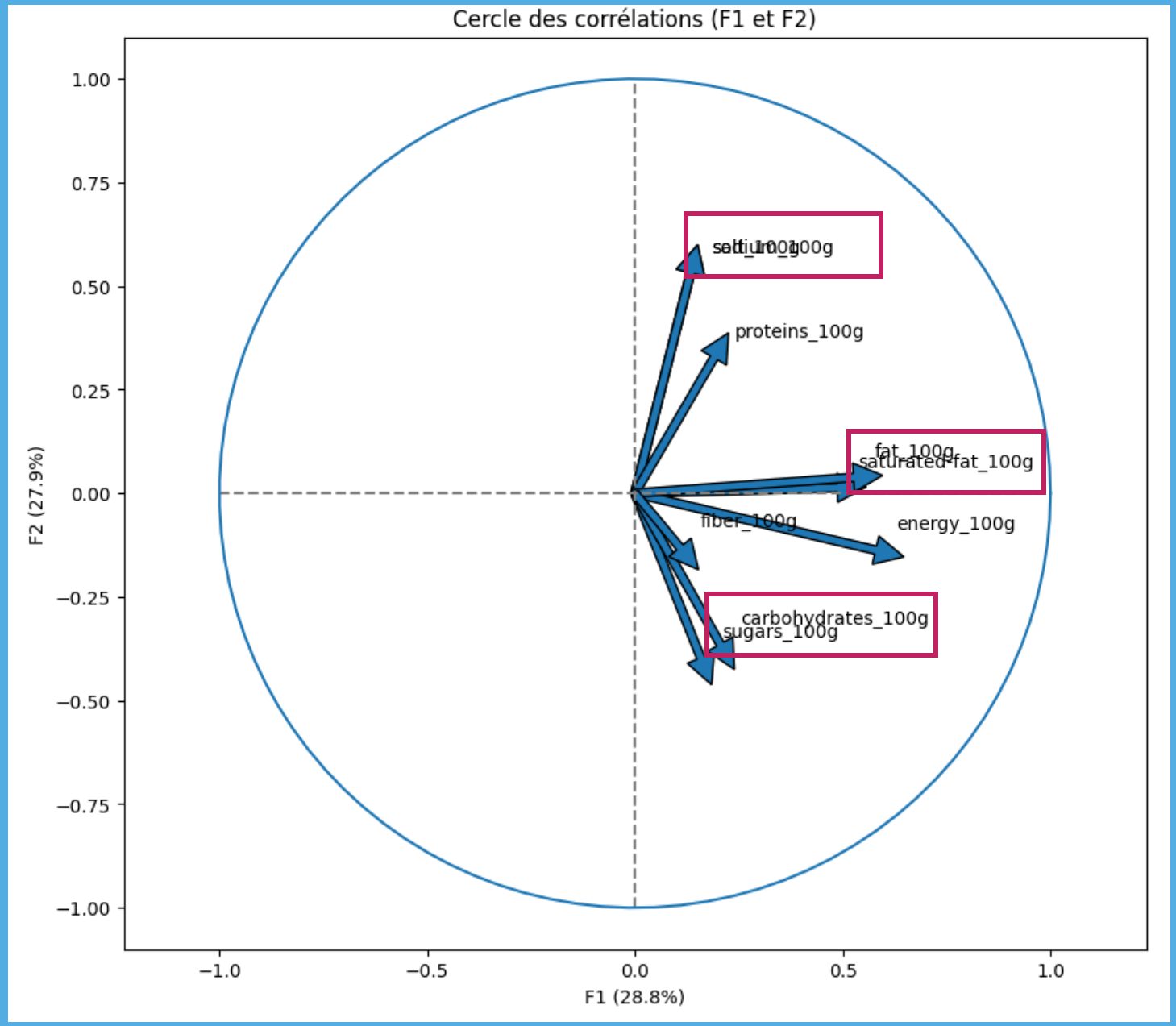


-5.021314474097451 0.016072415164441446 0.9999999999999994

Analyse :

- Ces 3 graphiques nous montrent qu'il y a bien une très forte corrélation entre:
  - Le sel et le sodium
  - Le gras et l'énergie
  - Le sucre et les glucides

Etude de la corrélation des variables via le cercle des corrélations :



Analyse :

- Confirmation des différentes corrélations entre variables.
- Corrélation opposée entre le sucre et le sel.

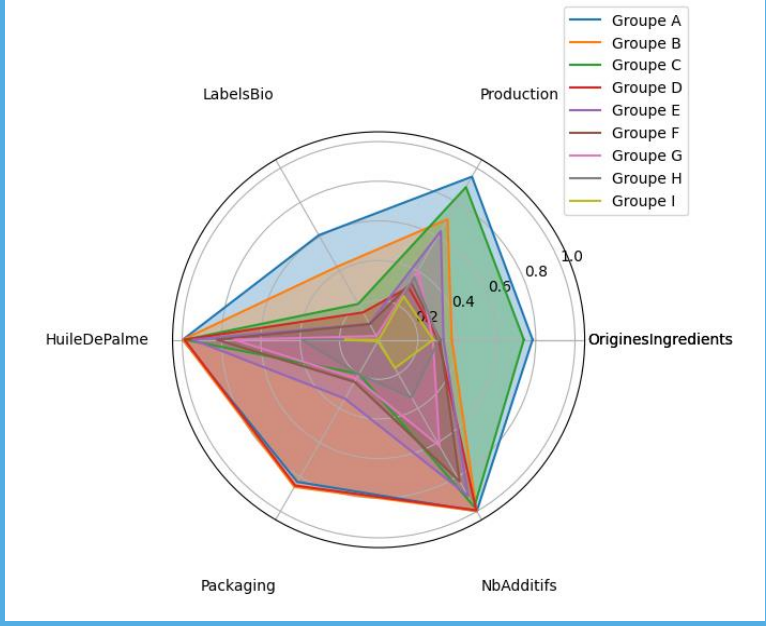
6 : Pertinence et faisabilité de l'application

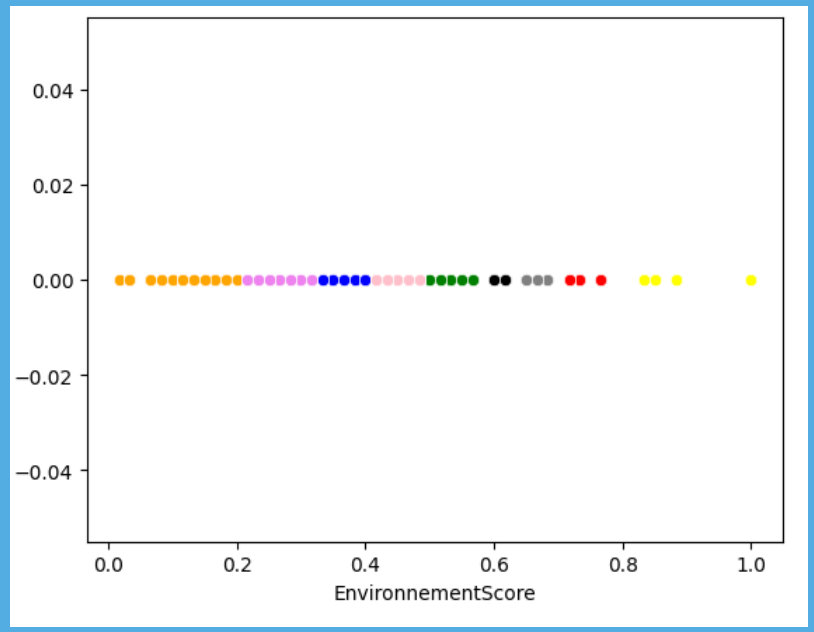
Pour chacune des 6 variables **(Label Bio, Nombre Additifs, Nombre Huile Palme, Provenance des ingrédients, Origine de fabrication des produits, Emballage)** j'ai créé un score **:**

- 1 : Bon
- 0.3 : Non Communiqué
- 0 à 0.2 : Mauvais

Les 6 variables sont ensuite rassemblées en un score unique.

Ensuite j'ai utilisé du Machine Learning avec le KMeans pour créer des groupes significativement différents (ici 9).





One Way Anova : **Tuckey**

group1	group2	meandiff	p-adj	lower	upper	reject
A	B	-0.136	0.0	-0.1511	-0.121	True
A	C	-0.1953	0.0	-0.2116	-0.179	True
A	D	-0.2614	0.0	-0.2762	-0.2466	True
A	E	-0.3186	0.0	-0.3325	-0.3046	True
A	F	-0.4231	0.0	-0.4373	-0.4089	True
A	G	-0.4811	0.0	-0.4957	-0.4666	True
A	H	-0.5948	0.0	-0.6091	-0.5805	True
A	I	-0.7151	0.0	-0.7311	-0.699	True
B	C	-0.0592	0.0	-0.0722	-0.0463	True
B	D	-0.1254	0.0	-0.1363	-0.1144	True
B	E	-0.1825	0.0	-0.1924	-0.1726	True
B	F	-0.287	0.0	-0.2972	-0.2769	True
B	G	-0.3451	0.0	-0.3558	-0.3344	True
B	H	-0.4588	0.0	-0.469	-0.4485	True
B	I	-0.579	0.0	-0.5916	-0.5664	True
C	D	-0.0661	0.0	-0.0788	-0.0535	True
C	E	-0.1233	0.0	-0.135	-0.1116	True
C	F	-0.2278	0.0	-0.2398	-0.2158	True
C	G	-0.2858	0.0	-0.2983	-0.2734	True
C	H	-0.3995	0.0	-0.4116	-0.3874	True
C	I	-0.5198	0.0	-0.5339	-0.5057	True

EnvironnementGrade pnns_groups_2	A	B	C	D	E	F	G	H	I	Total
Un plat de repas	131	444	299	382	1008	512	601	477	178	4032
Biscuits et gâteaux	45	152	68	147	408	479	426	922	581	3228
Céréales	118	309	113	314	574	449	408	141	103	2529
Bonbons	121	288	30	341	221	284	213	522	254	2274
Fromage	124	252	265	101	567	342	365	149	65	2230
Viande transformée	56	69	484	47	645	243	380	190	58	2172
Pansements et sauces	61	228	29	365	322	339	181	319	86	1930
Produits au chocolat	70	277	62	333	296	283	191	205	68	1785
Lait et yaourt	153	112	245	129	410	257	209	183	56	1754
Poisson et fruits de mer	36	186	118	183	536	87	275	107	23	1551
Légumes	102	262	58	375	384	162	135	43	16	1537
Apéritifs	15	61	50	106	228	326	187	336	127	1436
Boissons sucrées	33	96	34	227	179	405	111	245	55	1385
Jus de fruits	53	170	39	230	179	346	74	178	29	1298
Boissons non sucrées	44	134	57	166	124	246	86	290	48	1195
Pain	22	80	40	97	123	224	154	254	112	1106
Céréales du petit-déjeuner	19	115	23	155	117	179	124	200	94	1026
Fruits	58	106	46	158	256	210	60	44	18	956
Graisses	55	152	53	57	125	147	75	90	104	858
Viande	52	42	159	38	182	121	95	84	23	796

Nous vous proposons les choix suivants :

Nom du produit: Spaghetti longs aux œufs frais
Marque du produit: Lustucru,pastacorp
Composition du produit : Semoule de _blé_ dur de qualité supérieure, _œufs_ frais (13,5 %).
Allergènes repertoriés : blé, œufs
Traces rerpertoriés : Non Communiqué
Nutrigrade : a
EnvironnementGrade : B
Pour de plus amples renseignements : http://world-fr.openfoodfacts.org/produit/3660861025658/spaghetti-longs-aux-oeufs-frais-lustucru

Céréales

nutrition_grade_fr EnvironnementGrade	a	b	c	d	e
A	78	28	9	3	0
B	211	60	26	11	1
C	84	10	14	5	0
D	222	44	23	22	3
E	405	82	47	37	3
F	316	54	39	37	3
G	250	69	58	25	6
H	59	22	33	20	7
I	29	21	28	15	10

Pâtes alimentaires

nutrition_grade_fr EnvironnementGrade	a	b	c	d	e
A	14	1	1	1	0
B	57	16	12	1	0
C	44	4	6	1	0
D	48	8	7	3	0
E	198	29	24	10	0
F	200	20	25	22	0
G	150	19	36	9	1
H	27	10	19	8	0
I	18	4	7	4	0

Spaghetti

nutrition_grade_fr EnvironnementGrade	a	b	c	d
B	13	0	0	0
C	9	0	0	0
D	5	1	0	1
E	29	1	0	0
F	29	2	1	0
G	25	1	0	0
H	8	0	0	0
I	1	0	0	0

7 : Applications

NutriGrade = A
EnvironnementGrade = D

Produit Scanné par le client

NutriGrade = A
EnvironnementGrade = B

Produit suggéré par l'application