Projet 3 : Concevez une application au service de la santé publique

Laurent Cagniart

Énoncé de mission

- Traiter le jeu de données
- Produire des visualisations et Effectuer une analyse univariée
- Confirmer ou infirmer les hypothèses à l'aide d'une analyse multivariée. Effectuer les tests statistiques appropriés
- Justifier votre idée d'application

Idée d'application(1/2)

Une des **limites du Nutri-Score** est qu'il donne une indication pour 100g de produit pour chaque produit et ne tient donc pas compte de la **quantité consommée par l'utilisateur**

Une plus-value pour le consommateur serait de donner l'équivalent d'un **Nutri-Score pour le repas**, voire pour la journée et ainsi lui permettre **d'équilibrer son alimentation**

Un produit

Un repas

Idée d'application(2/2)

- Des indicateurs existent aux côtés du Nutri-Score, à savoir **l'Eco-score®** (impact environnemental)

et le **score NOVA** (niveau de transformation des aliments)

- Un **nouvel indicateur** complémentaire au Nutri-Score, basé sur la **présence et la toxicité d'additifs** dans le produit identifié (base évaluation : https://www.quechoisir.org/comparatif-additifs-alimentaires-n56877/)

Jeu de données

```
Entrée [6]: food.head()
```

Out[6]:

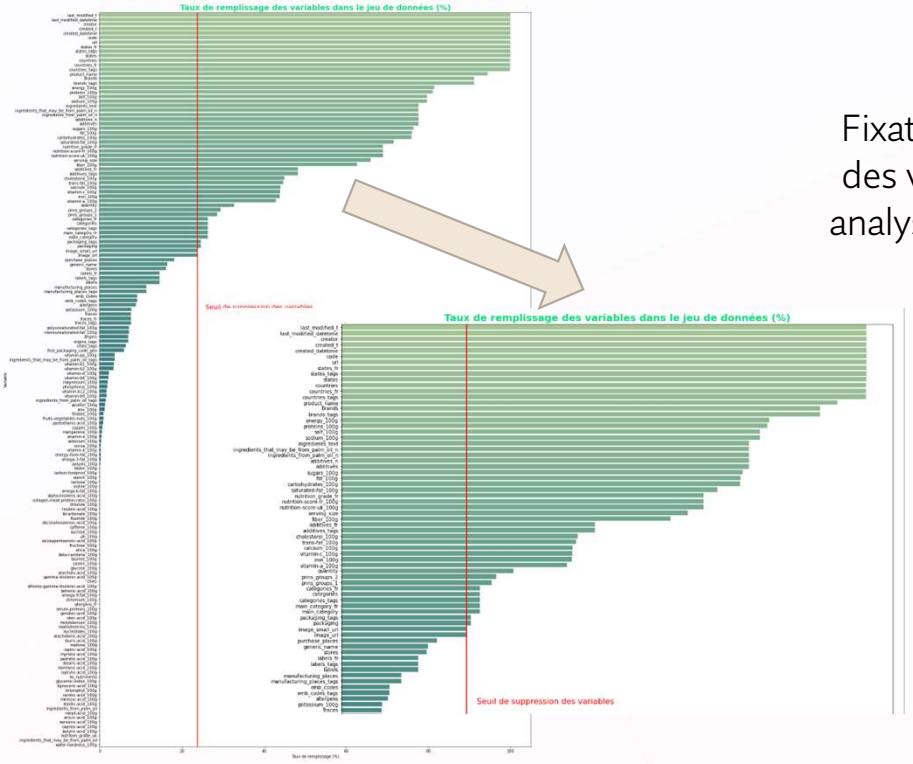
	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_
0	0000000003087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	
1	0000000004530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	
2	0000000004559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	
3	0000000016087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salted Nut Mix	
4	0000000016094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	

5 rows × 162 columns

```
Entrée [7]: food.shape
Out[7]: (320772, 162)
```

160+ variables/features

Travaux de formatage et nettoyage(1/7)



Fixation d'un seuil de remplissage minimum des variables pour être conservées pour les analyses ultérieures (à savoir, moyenne du nb de valeurs nulles)

162

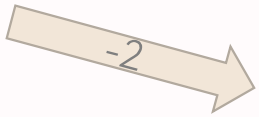
-110

52

Travaux de formatage et nettoyage(2/7)

- Tri du fichier par date de création puis de modification dans l'ordre chronologique (ex. utile pour le retraitement des doublons)

- Suppression de 2 colonnes en doublon (colonnes en _t) et conversion des 2 autres pour garder date de création et date de modification



50

Travaux de formatage et nettoyage(3/7)

- Suppression des doublons : ‘code’ puis par ‘product_name’ & ‘brands’

320,772 → -21,806 → 298,966

- Suppression des lignes pour lesquelles aucun nutriment (ou nutriScore) n'est rempli

298,266 → -55,046 → 243,920

Travaux de formatage et nettoyage(4/7)

Features redondantes :

- ‘states_fr’, ‘states_tags’, ‘states’ : on va conserver la colonne ‘states_fr’ (taux de remplissage identique)

- ‘countries’, ‘countries_fr’, ‘countries_tags’ : on va conserver la colonne ‘countries_fr’ (taux de remplissage identique, pas de valeurs non pays à retraiter) => retraitements à prévoir (plusieurs pays dans une même colonne)

- ‘brands’, ‘brands_tags’ : on va conserver la colonne ‘brands’

- ‘additives_fr’, ‘additives_tags’ : on va conserver la colonne ‘additives_tags’ : conservation uniquement des codages (retraitement chaine de caractères : supprimer les 'en:')

Travaux de formatage et nettoyage(5/7)

Features redondantes :

- ‘categories_fr’, ‘categories’, ‘categories_tags’ : on va conserver la colonne ‘categories_fr’ (légèrement mieux remplie)

- ‘main_category_fr’, ‘main_category’ : idem, on conserve ‘_fr’ (moins de retraitements)

- ‘packaging_tags’, ‘packaging’ : on va conserver la colonne packaging (plus de val. uniques)

➢ La feature "packaging" bien que possiblement utile pour un scoring de type EcoScore® est trop peu remplie et non dérivable des autres variables pour être complétées => à supprimer

Travaux de formatage et nettoyage(6/7)

Autres features :

- Feature "Quantity" et Feature ‘serving_size’ => à supprimer

- "vitamin-a_100g", "iron_100g", "vitamin-c_100g", "calcium_100g", "trans-fat_100g" et "cholesterol_100g" => à supprimer

- On ne conservera pas le nutriscore Uk, on n'utilisera que la définition Fr du NutriScore

Travaux de formatage et nettoyage(7/7)

Autres features :

- ‘Additives’ à supprimer, ‘additives_n’ à étudier avec ‘additives_tags’

- 'ingredients_that_may_be_from_palm_oil_n’ et 'ingredients_from_palm_oil_n’ : le nb d'ingrédients ne donnent pas directement d'infos sur le total huile de palme => impact sur taux de graisses saturées, la feature "saturated_fat" suffira

- On n’ utilisera pas la feature ‘ingredients_text’, son exploitation n'apportant pas des informations analysables

-14

26

Analyse des valeurs aberrantes et nulles (1/8)

Features	nb de val nulles
main_category_fr	182543
categories_fr	182542
pnns_groups_1	178286
pnns_groups_2	178090
additives_tags	105219
fiber_100g	57050
nutrition-score-fr_100g	37742
nutrition_grade_fr	37742
saturated-fat_100g	30241
additives_n	26435
carbohydrates_100g	18810
fat_100g	18500
sugars_100g	16375
sodium_100g	7061
salt_100g	7014
brands	5836
product_name	3181
proteins_100g	2770
energy_100g	1645
countries_fr	67
states_fr	1
url	1
code	1
created_datetime	1
last_modified_datetime	1
creator	0
dtype:	int64

- on voit qu'il y a 4 features qui correspondent aux catégories de produits => à creuser

➢ Ces features sont utiles pour le calcul du NutriScore et la comparaison de produits (pour un projet d'application proposant des alternatives plus saines; projet d'application non retenu dans cette phase du projet)

- pour chaque feature de type '_100g', la valeur ne doit pas dépasser 100g ou présenter des valeurs négatives => à supprimer sinon

➢ Energy <= 3700kJ

➢ saturated-fat (ou trans fat ou cholesterol) < fat

➢ sugars < carbohydrates

➢ sodium < salt

-1,754

242,166

Analyse des valeurs aberrantes et nulles (2/8)

Features	nb de val nulles
main_category_fr	182543
categories_fr	182542
pnns_groups_1	178286
pnns_groups_2	178090
additives_tags	105219
fiber_100g	57050
nutrition-score-fr_100g	37742
nutrition_grade_fr	37742
saturated-fat_100g	30241
additives_n	26435
carbohydrates_100g	18810
fat_100g	18500
sugars_100g	16375
sodium_100g	7061
salt_100g	7014
brands	5836
product_name	3181
proteins_100g	2770
energy_100g	1645
countries_fr	67
states_fr	1
url	1
code	1
created_datetime	1
last_modified_datetime	1
creator	0
dtype:	int64

- Suppression des noms de produits non renseignés

-3,155

239,011

- Retraitements sur les données liées aux additifs:

➢ nettoyage des caractères en surplus ('en:', indicateur de la langue)

➢ définition du nb d'additifs à partir des tags remplis

➢ comparaison du nb d'additifs entre nb saisi et nb recalculé

➢ Les écarts sont de 2 types : '-1' ou 'nan'

➢ On conservera donc, pour le nb d'additifs, la colonne recalculée à partir des tags (si un tag existe, c'est bien qu'il y a présence d'un additif)

Analyse des valeurs aberrantes et nulles (3/8)

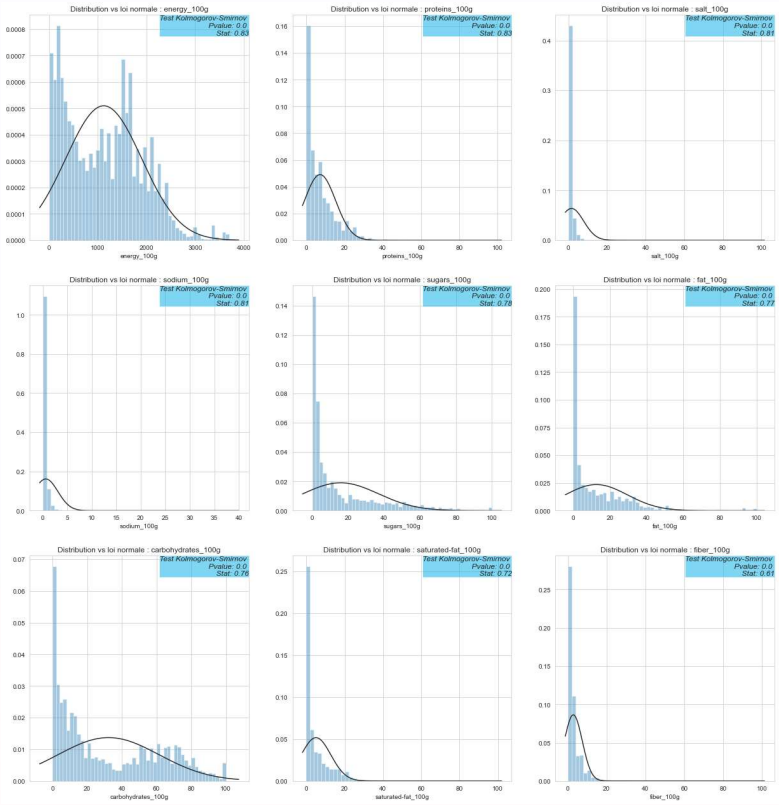
Features	nb de val nulles
main_category_fr	182543
categories_fr	182542
pnns_groups_1	178286
pnns_groups_2	178090
additives_tags	105219
fiber_100g	57050
nutrition-score-fr_100g	37742
nutrition_grade_fr	37742
saturated-fat_100g	30241
additives_n	26435
carbohydrates_100g	18810
fat_100g	18500
sugars_100g	16375
sodium_100g	7061
salt_100g	7014
brands	5836
product_name	3181
proteins_100g	2770
energy_100g	1645
countries_fr	67
states_fr	1
url	1
code	1
created_datetime	1
last_modified_datetime	1
creator	0
dtype:	int64

- Dans une 2e phase du projet, pour améliorer l'utilité de l'application, il faudrait par rapport à la liste des additifs présents dans un produit, croiser cette liste (jonction) avec une liste donnant le niveau de toxicité de ces dits additifs, et appliquer une fonction du type toxicité_produit = max(toxicité_additifs)

➢ Liste niveau de toxicité à créer à partir des données sur le site suivant : https://www.quechoisir.org/comparatif-additifs-alimentaires-n56877/

➢ Par exemple, on a 3100+ produits contenant l'additif e171 (à savoir, le dioxyde de titane), additif classé dans la catégorie "à éviter" (niv 4 sur 4), perturbateur endocrinien, possible impact sur le cancer colorectal…

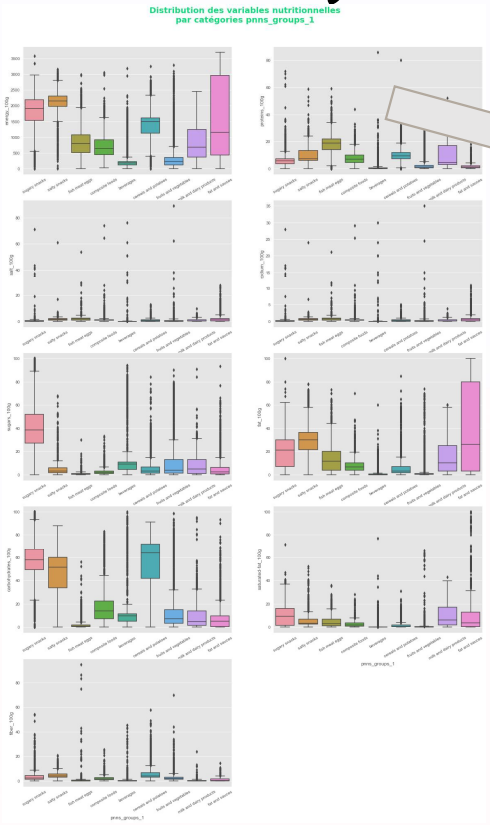
Analyse des valeurs aberrantes et nulles (4/8)

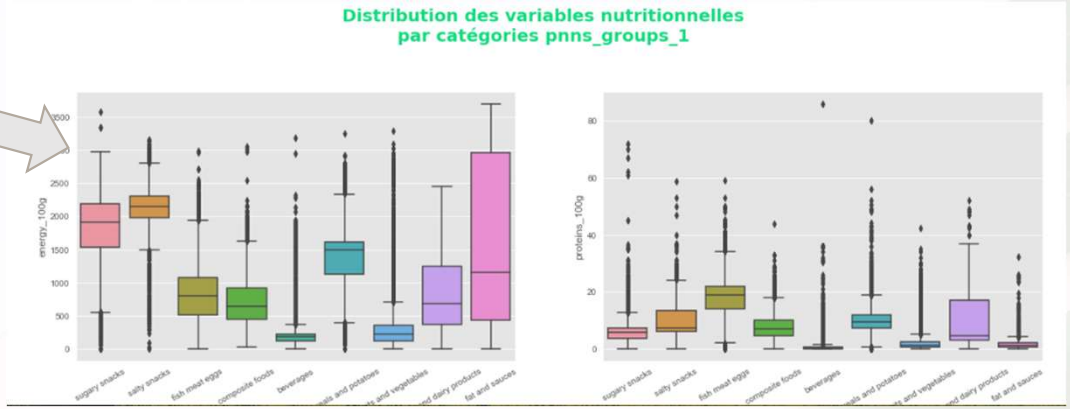


- Analyse de la distribution des variables liés aux nutriments

  - Test Kolmogorov-Smirnov, Distribution vs loi normale
  - on rejette donc l'hypothèse de normalité des distributions de ces variables. Il n'est donc pas souhaitable d'imputer les valeurs manquantes par la moyenne.
  - Pour confirmer cette approche, regardons à présent quelque unes de ces distributions en fonction de la catégorie pnns_groups_1

Analyse des valeurs aberrantes et nulles (5/8)





- Ces représentations nous permettent d'appuyer des résultats attendus, par ex., 'cereals and potatoes' et 'sugary snacks' pour la feature 'carbohydrates', 'sugary snacks' pour la feature 'sugars', proteins pour 'fish meat eggs' ou encore les produits riches en sucres et graisses pour la feature 'energy'

Analyse des valeurs aberrantes et nulles (6/8)

Features	nb de val nulles
energy_100g	1589
proteins_100g	2634
salt_100g	6566
sodium_100g	6605
sugars_100g	15789
fat_100g	18238
carbohydrates_100g	18525
saturated-fat_100g	29592
fiber_100g	55592
dtype:	int64

- Imputation des valeurs nulles sur les features principales liées aux nutriments

➢ La feature 'fiber_100g' est mal renseignée mais nous en aurons besoin pour la suite. Nous allons donc compléter les valeurs nulles par la médiane de la catégorie pnns_groups_2.

➢ Enfin, pour les autres variables, avec peu de null et dont les distributions ne suivent pas la loi normale, nous allons imputer avec l'algorithme kNN

➢ On passe les valeurs nulles ou non renseignées à 'unknown' pour la feature 'brands' et 'no additives' pour 'additives_tags'

	energy_100g	proteins_100g	salt_100g	sodium_100g	sugars_100g	fat_100g	carbohydrates_100g	saturated-fat_100g	nutrition-score-fr_100g	fiber_100g	additives_n2
count	239011.000000	239011.000000	239011.000000	239011.000000	239011.000000	239011.000000	239011.000000	239011.000000	202285.000000	239011.000000	239011.000000
mean	1120.642333	7.121496	1.596630	0.627200	15.224217	12.129956	31.637054	4.577782	9.118175	2.522966	2.215739
std	781.510463	8.101338	6.196646	2.435283	20.646606	16.548964	28.433071	7.440146	9.052964	4.116341	2.189161
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-15.000000	0.000000	1.000000
25%	387.000000	0.740000	0.068580	0.027559	1.000000	0.000000	6.670000	0.000000	1.000000	0.000000	1.000000
50%	1103.000000	4.900000	0.596900	0.236220	5.040000	5.000000	21.100000	1.220000	10.000000	1.500000	1.000000
75%	1674.000000	10.000000	1.381760	0.545000	22.430000	19.548000	57.000000	6.620000	16.000000	3.100000	3.000000
max	3700.000000	100.000000	100.000000	39.370079	100.000000	100.000000	100.000000	100.000000	40.000000	100.000000	31.000000

Analyse des valeurs aberrantes et nulles (7/8)

- Calcul du nutriscore

➢ Ce sont les dernières features à présenter des valeurs nulles, nous allons utiliser les données pour recalculer le NutriScore et le Grade et également comparer aux scores saisis, pour valider la bonne imputation des valeurs que nous avons dû estimées pour remplacer les valeurs nulles précédemment

L'accuracy_score sur les NutriScores calculés est de : 42.04 %.

L'accuracy_score sur les NutriGrades calculés est de : 75.11 %.

➢ Les calculs basés sur trop d'hypothèses n'ont pas permis d'avoir une estimation suffisamment robuste (42%) même si le passage aux grades génèrent moins d'écart (effet lié au regroupement)

➢ On restera donc sur les NutriScores saisis dans la base pour l'analyse exploratoire

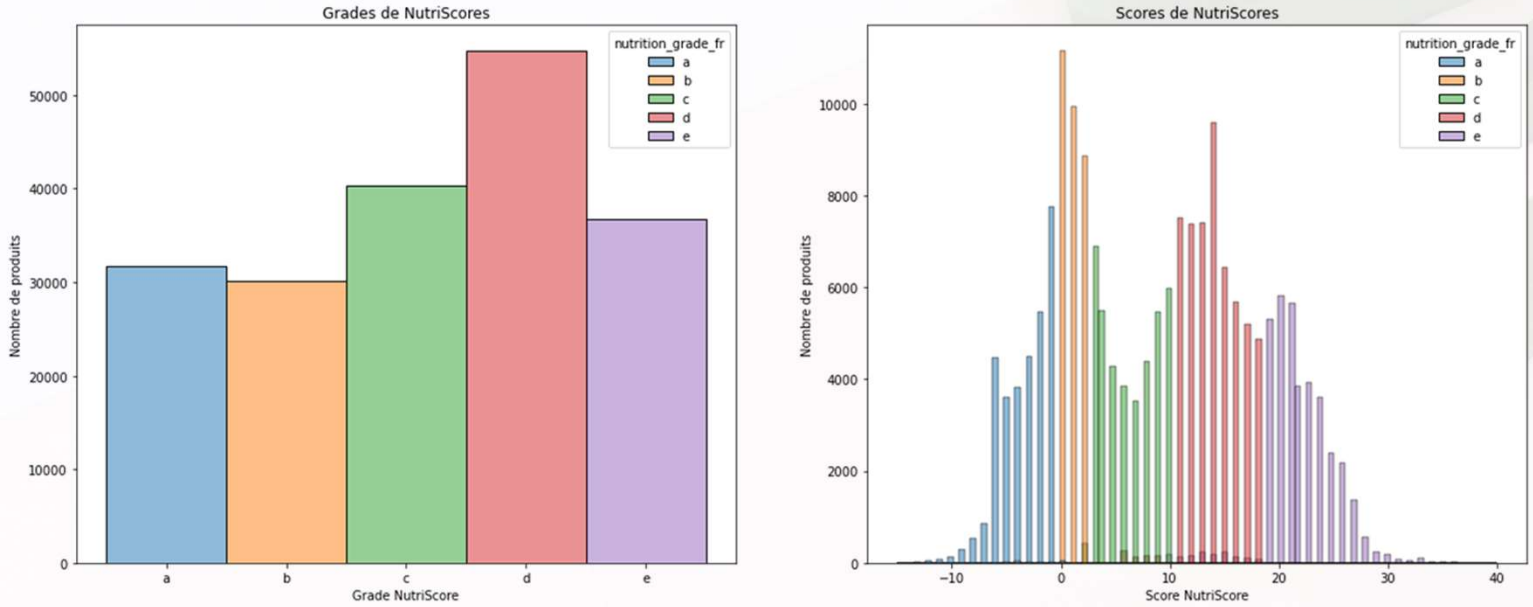
Analyse des valeurs aberrantes et nulles (8/8)

#	Column	Non-Null	Count	Dtype
0	last_modified_datetime	222099	non-null	datetime64[ns]
1	creator	222099	non-null	object
2	created_datetime	222099	non-null	datetime64[ns]
3	code	222099	non-null	object
4	url	222099	non-null	object
5	states_fr	222099	non-null	object
6	countries_fr	222042	non-null	object
7	product_name	222099	non-null	object
8	brands	222099	non-null	object
9	energy_100g	222099	non-null	float64
10	proteins_100g	222099	non-null	float64
11	salt_100g	222099	non-null	float64
12	sodium_100g	222099	non-null	float64
13	sugars_100g	222099	non-null	float64
14	fat_100g	222099	non-null	float64
15	carbohydrates_100g	222099	non-null	float64
16	saturated-fat_100g	222099	non-null	float64
17	nutrition_grade_fr	193684	non-null	object
18	nutrition-score-fr_100g	193684	non-null	float64
19	fiber_100g	222099	non-null	float64
20	additives_tags	222099	non-null	object
21	pnns_groups_2	222099	non-null	object
22	pnns_groups_1	222099	non-null	object
23	categories_fr	222099	non-null	object
24	main_category_fr	222099	non-null	object
25	additives_n2	222099	non-null	int64
26	fruits-legumes-ratio_100g	222099	non-null	int64
27	sat-fat_ratio	222099	non-null	float64
dtypes:	datetime64[ns](2),	float64(11),	int64(2),	object(13)
memory	usage:	49.1+	MB	

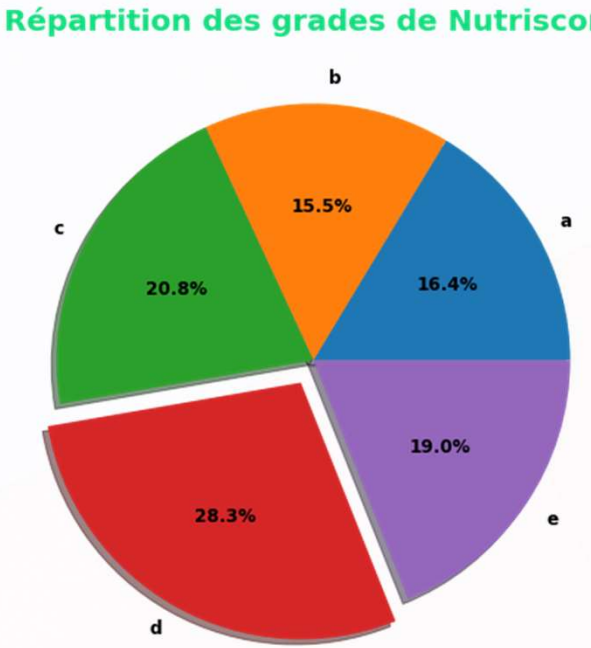
- Le jeu de données est maintenant nettoyé et exploitable, on le sauvegarde donc en format csv pour l'analyse exploratoire

Analyse des variables NutriScore et NutriGrade(1/7)

Répartition des scores NutriScore et de leurs grades



Analyse des variables NutriScore et NutriGrade(2/7)



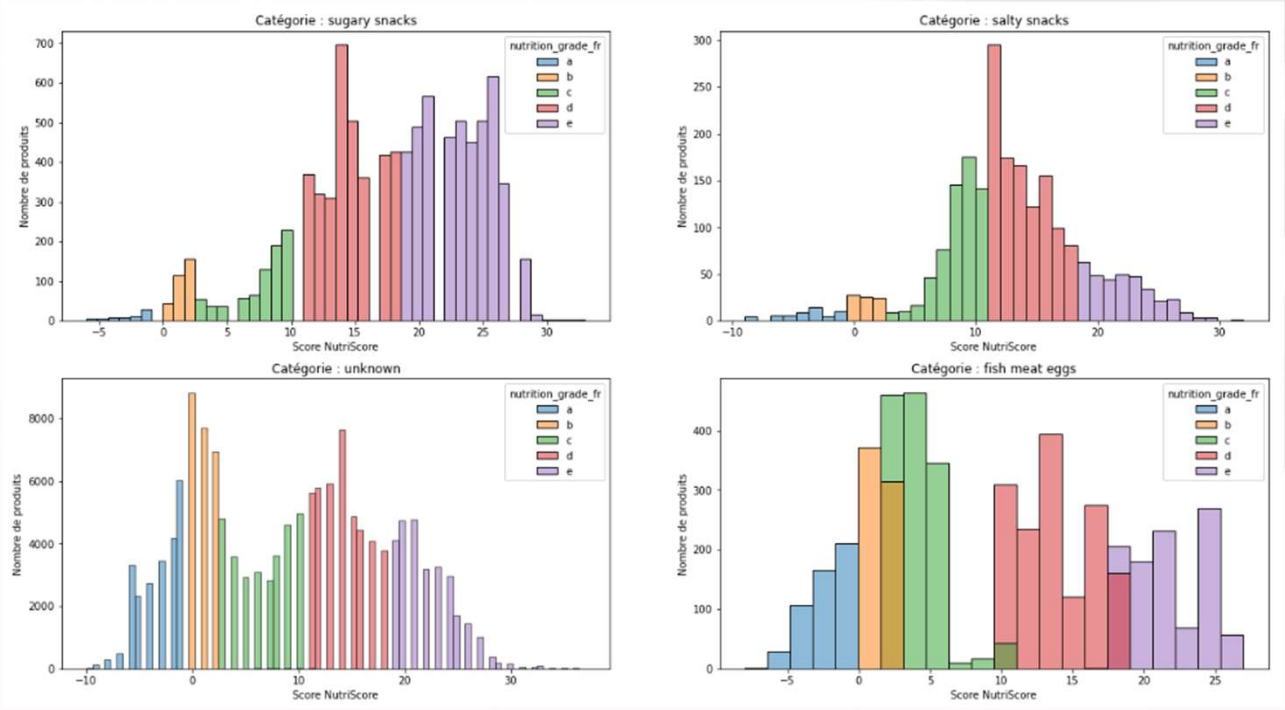
- Les différents graphiques montrent que le grade 'D' est légèrement plus représenté que les autres grades

- Le 2e graphique du slide précédent permet de visualiser la distribution de la variable "NutriScore« , qui laisse apparaître une distribution bi-modale (analyse univariée)

- On va affiner l'analyse en intégrant les catégories de produits

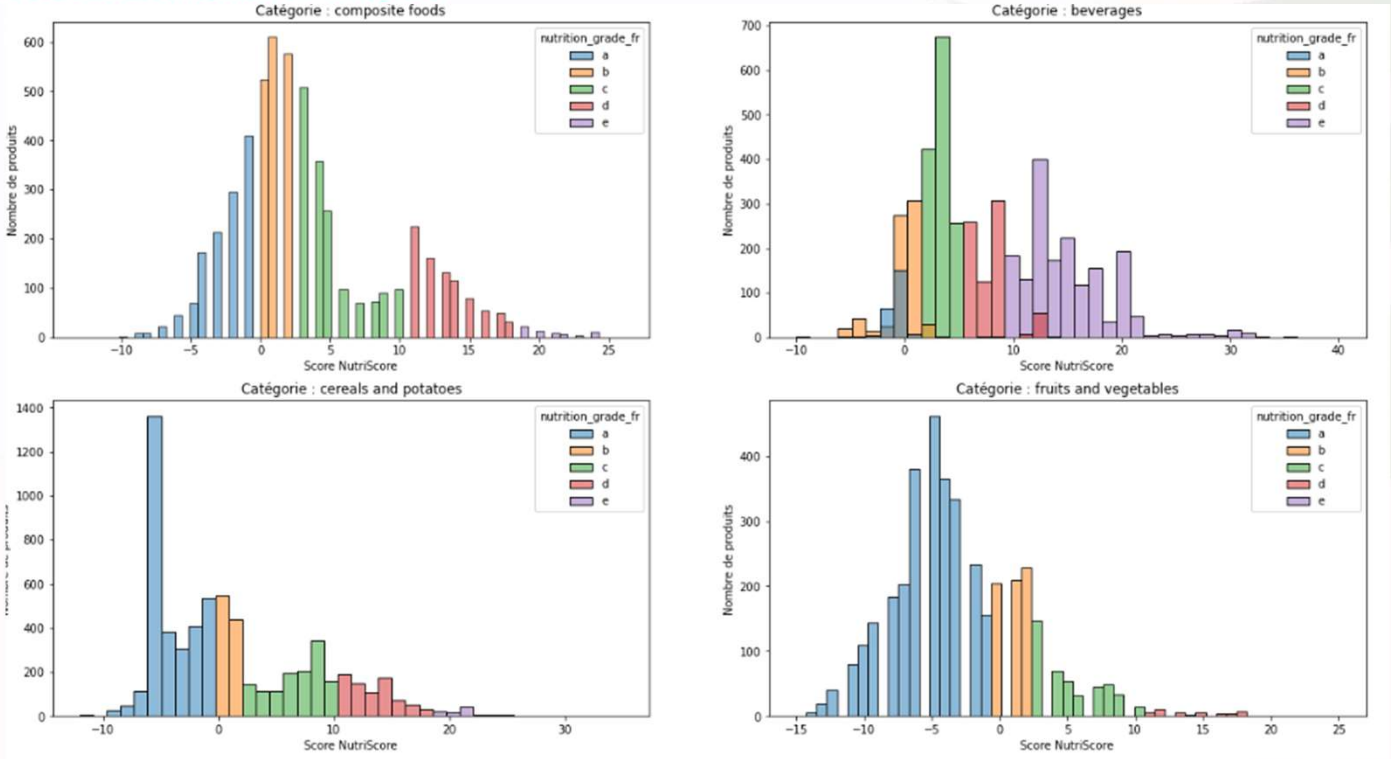
Analyse des variables NutriScore et NutriGrade(3/7)

Distribution des scores NutriScore par catégorie



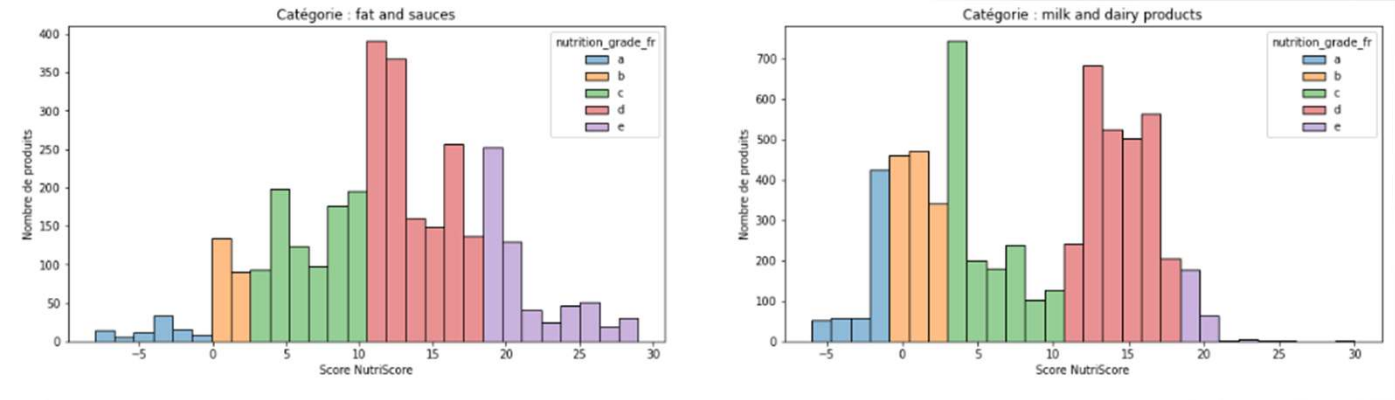
Analyse des variables NutriScore et NutriGrade(4/7)

Distribution des scores NutriScore par catégorie



Analyse des variables NutriScore et NutriGrade(5/7)

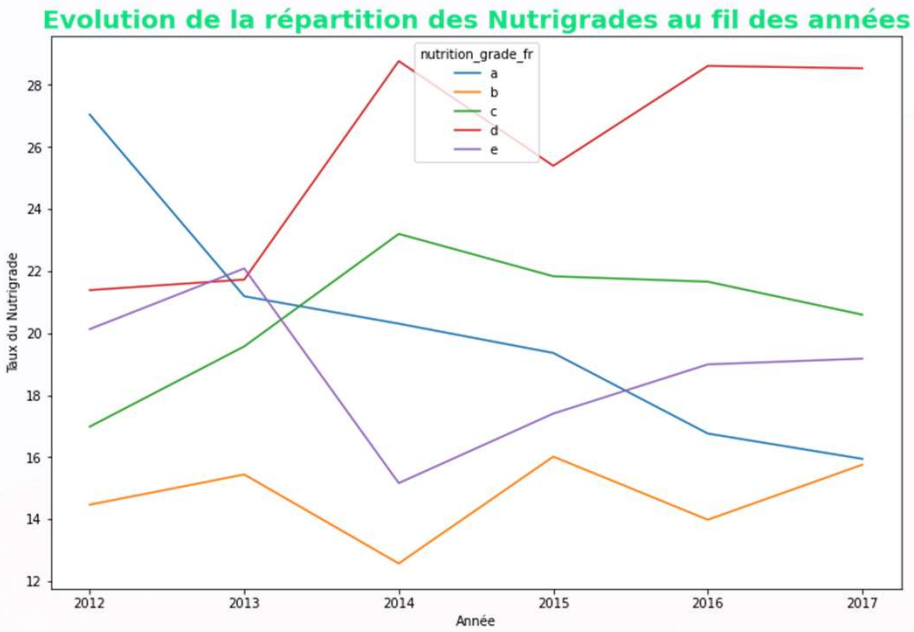
Distribution des scores NutriScore par catégorie



- Intuitivement, il semble qu'il y ait un lien entre catégorie de produits et NutriScore.

- Ex. 'fats and sauces' plutôt représentée par des produits aux grades 'D' et 'E' 'fruits and vegetables' plutôt représentée par des produits au grade 'A'

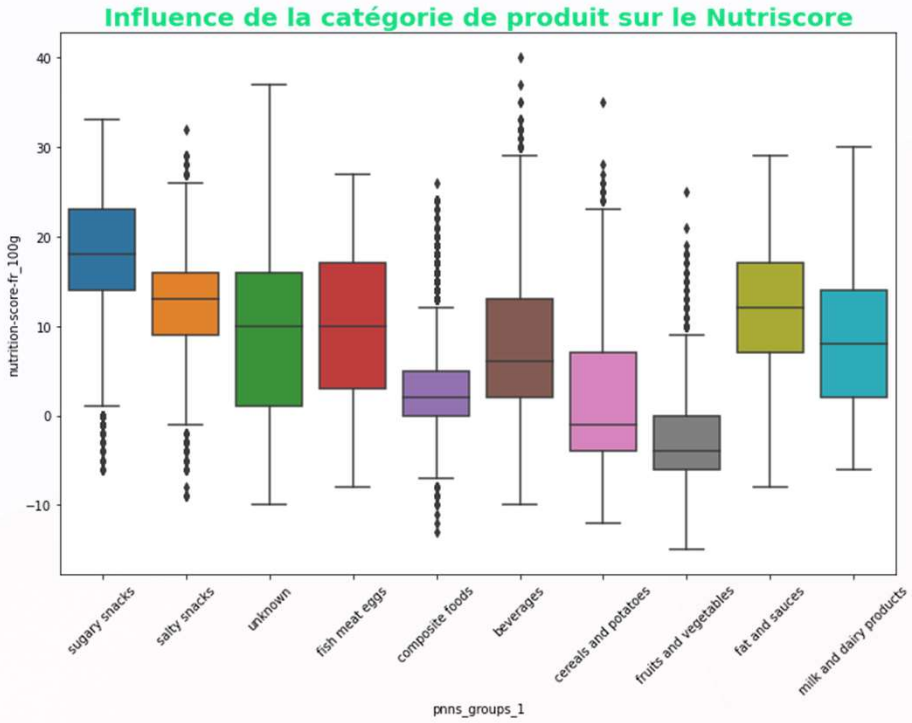
Analyse des variables NutriScore et NutriGrade(6/7)



- De manière contre-intuitive, les "bons" nutriscores ont plutôt stagné ('B') voire fortement diminué ('A') au contraire du 'D' depuis 2012

- Remarque : le jeu de données Fr n'a pas de produits créés depuis 2017 (date du fichier mis à disposition sur OC)

Analyse des variables NutriScore et NutriGrade(7/7)



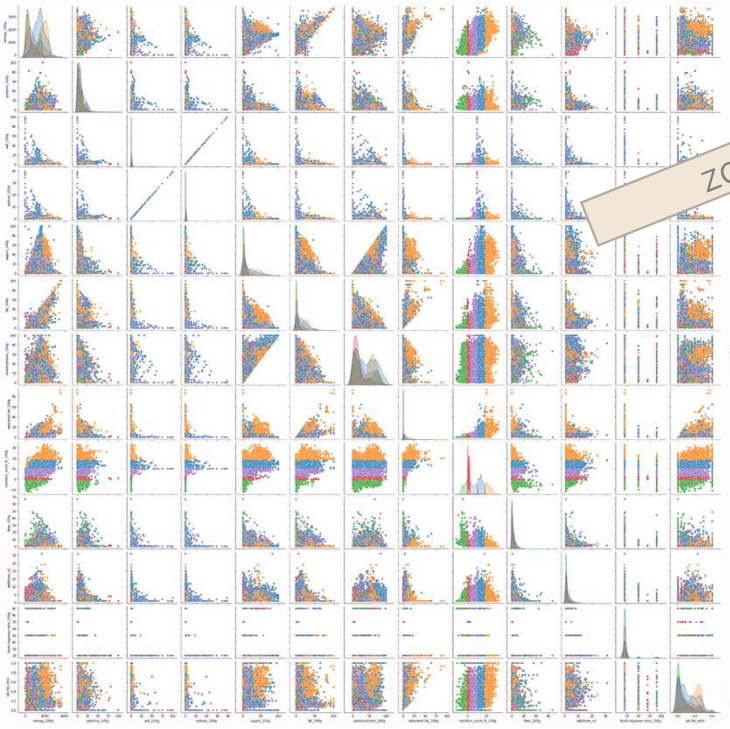
- Analyse de type ANOVA

	sum_sq	df	F	PR(>F)
pnns_groups_1	1.856196e+06	9.0	2865.429279	0.0
Residual	1.394001e+07	193674.0	NaN	NaN

➢ PR(>F) correspond à une p-value de 0, l'hypothèse retenue est bien l'hypothèse alternative à savoir que la catégorie pnns a bien une influence sur le NutriScore

Analyse des corrélations (1/8)

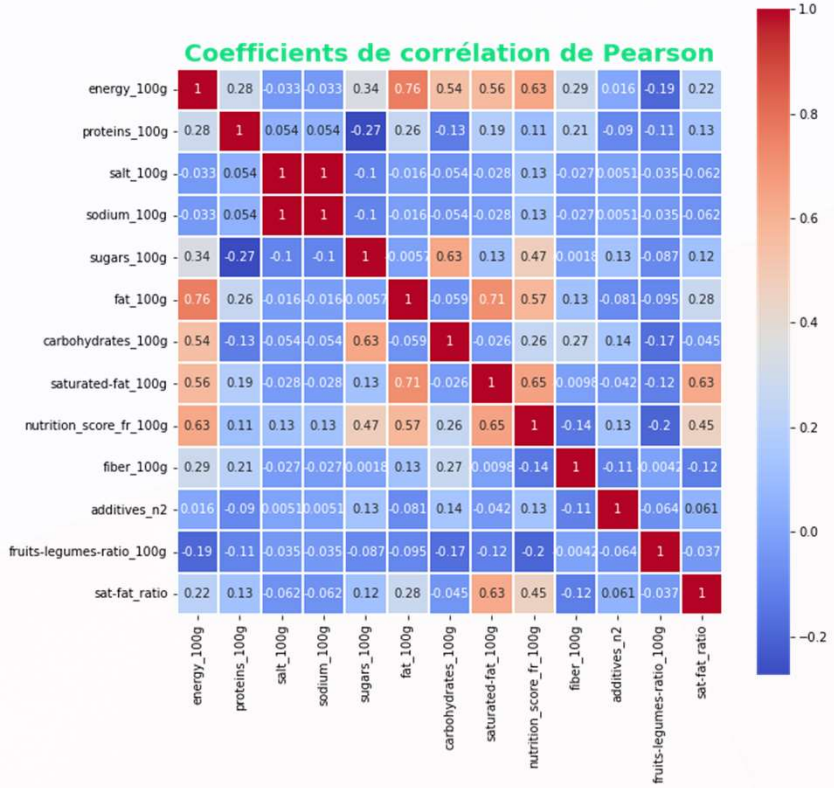
- Analyse bivariée, variable par variable 2 à 2



zoom

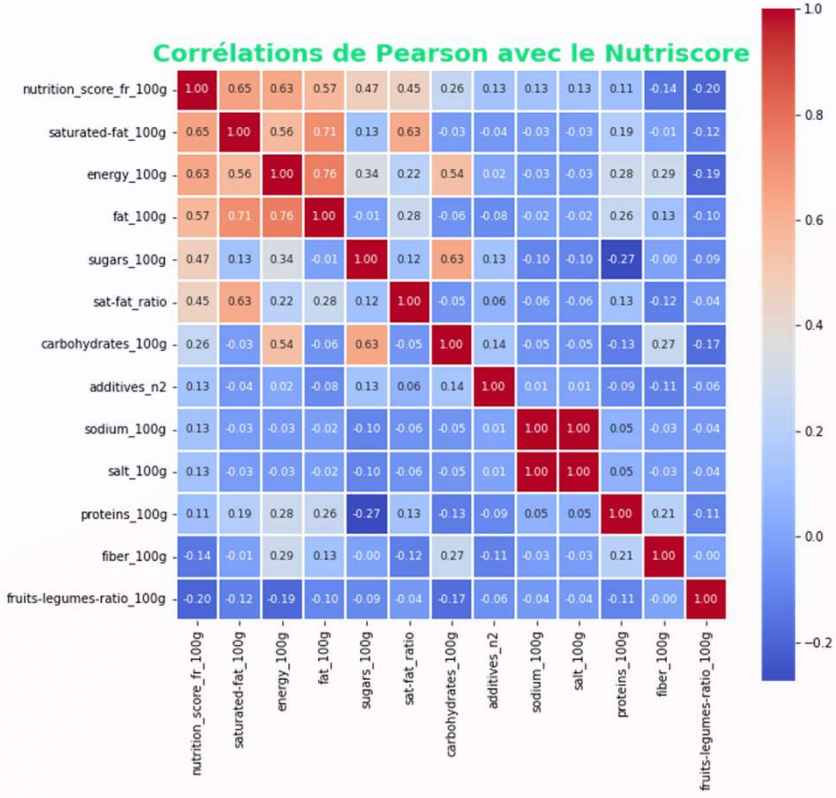
- Par exemple, on va voir si l'on trouve un lien entre 'energy_100g' et 'fat_100g', ce qui devrait être le cas physiologiquement parlant (hors impact des glucides sur le niveau global d'énergie)

Analyse des corrélations (2/8)



Par exemple, le tableau confirme la corrélation directe entre 'salt' et 'sodium' (rapport de 2.5 pour info) et d'autres corrélations fortes

Analyse des corrélations (3/8)

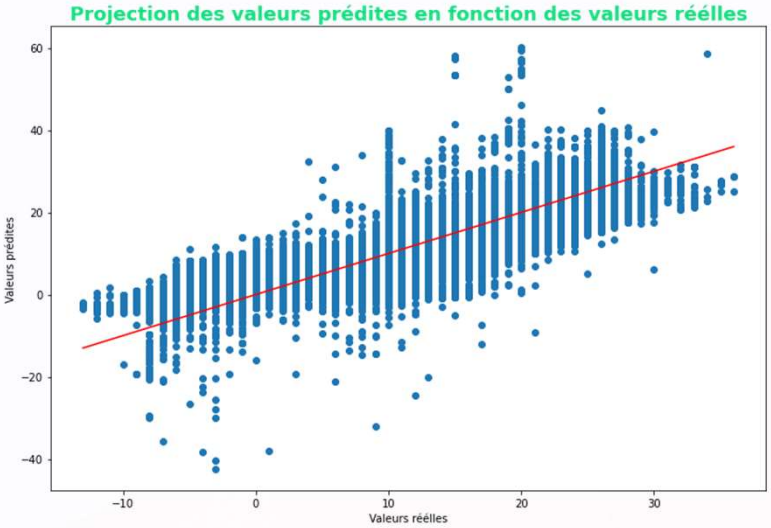


On voit bien ici que le NutriScore est plus fortement corrélé aux niveaux d'énergie et de lipides (saturated-fat davantage) et à un degré moindre au niveau des sucres.

- Cela confirme bien que le NutriScore peut être effectivement utilisé pour limiter la consommation de graisses saturées et de sucre.

Note : En effet, un gramme de lipides renferme plus de deux fois plus d'énergie qu'un gramme de glucides ou de protéines (9 calories pour 1 g de lipides, 4 calories pour 1 g de glucides ou de protéines)

Analyse des corrélations (4/8)



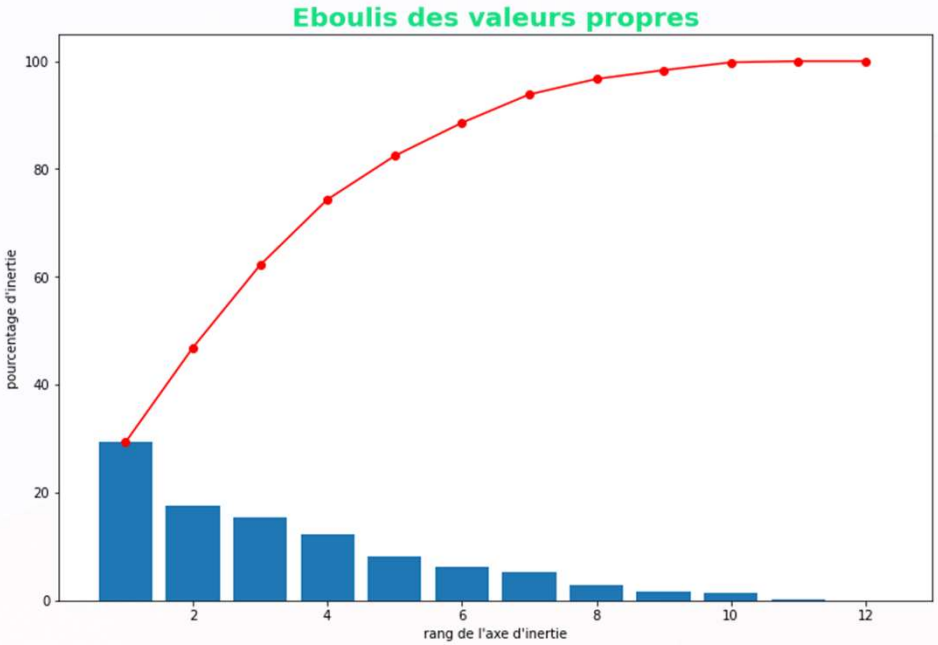
Régression linéaire multivariée.

- Les valeurs restent relativement dispersées. On va essayer d'affiner en intégrant les catégories de produits.

	Métrique	Baseline	LinearRegression	LinearRegression cat
0	**MAE**	7.810103	3.741525	3.382589
1	**MSE**	81.606649	23.868681	19.987402
2	**RMSE**	9.033640	4.885558	4.470727
3	**R^2**	-0.000029	0.707507	0.755069

- R^2 est supérieur dans ce 2e modèle, on a donc bien une amélioration avec la catégorie de produits, ce qui correspond à ce qu'on a vu lors de l'analyse ANOVA

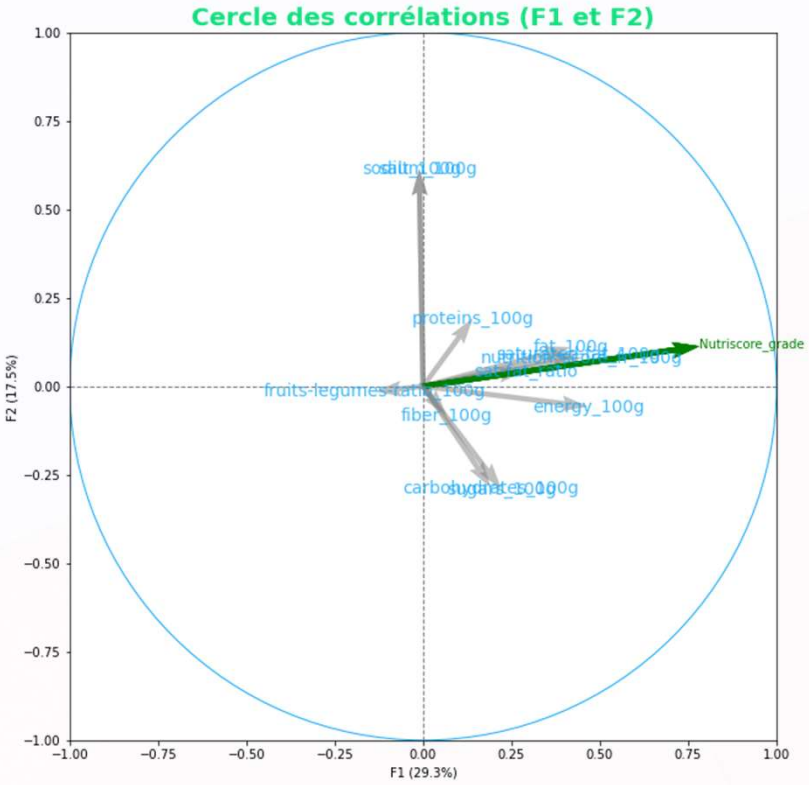
Analyse des corrélations (5/8)



Analyse ACP

Le premier plan factoriel couvrira une inertie de 46.86%, le second plan : 74.34% et le troisième plan : 88.57%.

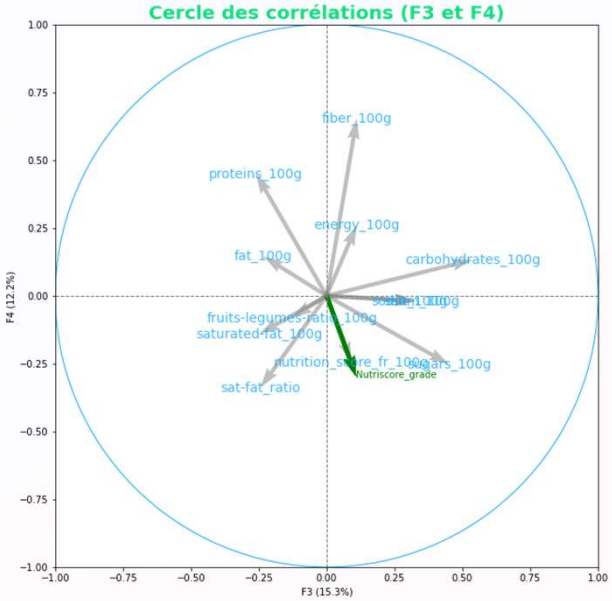
Analyse des corrélations (6/8)

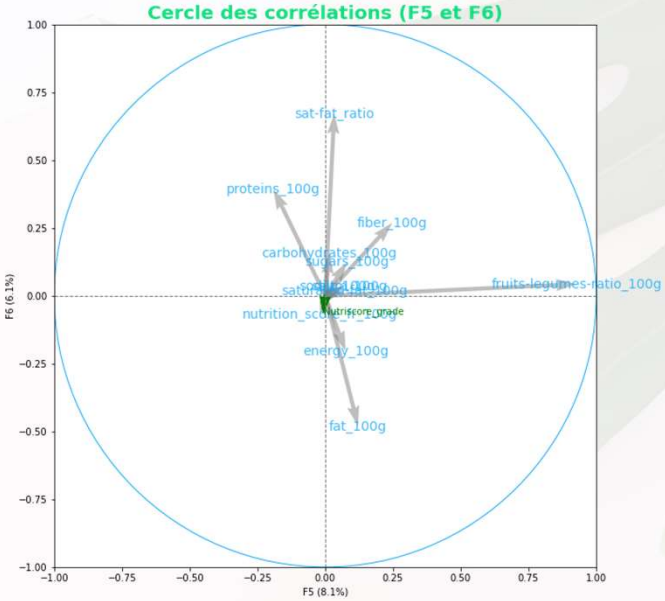


- L'axe **F1** est davantage représenté par la composante **énergétique et matières grasses**. D'ailleurs, on voit le lien de corrélation fort entre ces mêmes variables

- L'axe **F2** fait ressortir les variables **'salé' – 'sucré'** (positivement pour le salé et négativement pour le sucré) et où l'on remarque également les corrélations fortes (salt/sodium et sugars/ carbohydrates)
  - ➢ Prépondérance pour la composante 'salé', le sucré ressort davantage sur l'axe F3 (cf. slide suivant)

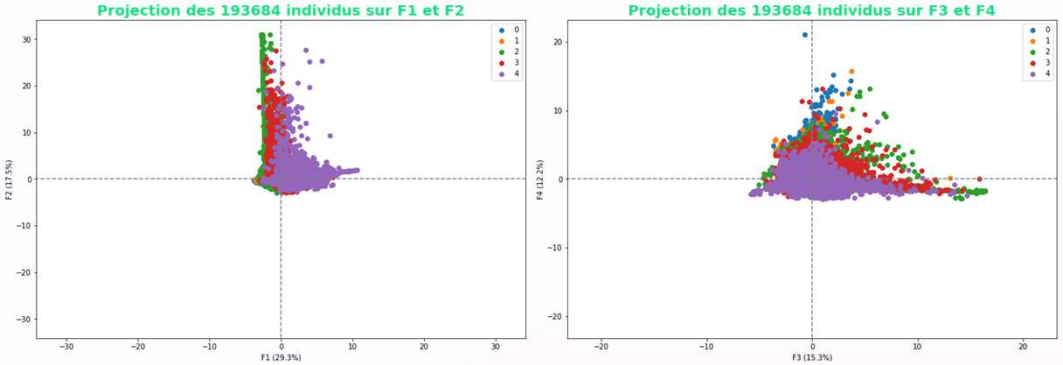
Analyse des corrélations (7/8)





- L'axe **F3** fait ressortir à nouveau la composante **carbohydrates/sugars**
- L'axe **F4** est marqué par la variable '**fiber_100g**'
- L'axe **F5** fait ressortir la composante '**fruits-légumes-ratio**'
- L'axe **F6** est marqué par la variable '**sat-fat ratio**'

Analyse des corrélations (8/8)



- Projection des individus sur les plans factoriels:
  - Les groupes d'individus ont tendance à se superposer, il est par conséquent difficile d'en tirer des conclusions probantes

Métrique	Baseline	LinearRegression	LinearRegression cat	LinearRegression PCA
MAE	7.810103	3.741525	3.382589	6.968916
MSE	81.606649	23.868681	19.987402	69.494500
RMSE	9.033640	4.885558	4.470727	8.336336
R^2	-0.000029	0.707507	0.755069	0.148397

- R^2 est faible, il n'est donc pas pertinent d'utiliser les variables synthétiques
- Ceci s'explique notamment par le fait que nous avons déjà réduit le nb de variables aux plus pertinentes pour le calcul du NutriScore

Conclusion pour l'application (1/2)

Nous avons redéfini la formule de calcul du NutriScore et du NutriGrade à partir des variables disponibles, il est donc possible de réutiliser ce modèle et d'y intégrer les notions de **portions** pour pondérer le calcul et ainsi déterminer un **NutriGrade « récapitulatif » pour l'ensemble du repas**

De plus, compte tenu des corrélations identifiées, nous avons pu confirmer l'intérêt du NutriScore dans un cadre **préventif pour améliorer/éviter la consommation de produits trop gras/ trop sucrés**

Conclusion pour l'application (2/2)

Dans un souci de simplification, nous avons réalisé une Analyse en Composantes Principales qui a permis de confirmer ces dites corrélations mais pas de réduire le nombre de variables étudiées (finalement déjà limitées aux variables numériques pertinentes pour le calcul du NutriScore)

Nous avons vu également qu'il serait possible dans un 2e temps du projet d'y intégrer la possibilité d'analyser la présence d'additifs, d'identifier ceux considérés comme nocifs et ainsi de prévenir l'utilisateur

Merci

Laurent Cagniart