

Projet 2 : Conception d'un application pour la santé publique française

Rapport d'exploration



Auteur : Antoine Chesnais
Date dernière version : 05/08/2019

Projet réalisé dans le cadre de la formation
« Ingénieur Machine Learning » d'Openclassrooms.

Table des matières

I. Introduction.....	3
II. Idée d'application	4
A) Objectif de l'application.....	4
B) Principe de fonctionnement.....	5
III. Préparation des données	7
A) Opérations de bases	8
B) Zoom sur la France.....	8
C) Sélection des produits dont on connaît la nature	9
D) Aperçu des statistiques descriptives	12
E) Gestion des variables peu représentées	13
F) Gestion des produits ayant des variables essentielles non assignées.....	13
G) Gestion des données aberrantes.....	15
H) Réduction des outliers à partir du Z-score.....	15
IV. Analyse des données.....	16
A) Distribution des variables importantes.....	16
B) Corrélations linéaires entre les variables importantes.....	17
C) Comparaison des différentes marques	18
V. Conclusion.....	22

I. Introduction

L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation. Le but du projet est d'explorer les données Open Food Fact disponible sur internet et d'élaborer une idée d'application à proposer. La base de données Open Food Fact est un projet non lucratif ouvert à tous. Les utilisateurs peuvent s'enregistrer sur le site afin de contribuer en ajoutant des produits et leur description. Les données sont librement accessibles et téléchargeables sur [leur site](#).

Afin d'élaborer une idée d'application, deux notebooks en python ont été réalisés : l'un contenant toute la partie nettoyage du dataset, et l'autre la partie analyse des données. Le premier a permis d'obtenir des données exploitables centrées sur la problématique santé et le second d'explorer le dataset afin de repérer des faits intéressants et pertinents pour la conception d'une application. En complément du rapport un dossier d'annexes est disponible pour avoir accès à un complément d'informations si nécessaire.

Ce rapport se décompose en trois parties principales et une synthèse. Dans un premier temps il sera exposé le concept de l'application identifié, son intérêt et son principe de fonctionnement. Ensuite seront présentées toutes les étapes de nettoyage effectuées sur les données pour les rendre exploitables. La troisième partie elle montrera les faits pertinents pour l'application qui ont pu être identifiés lors de l'étape d'analyse. Enfin s'en suivra la synthèse, qui conclura sur la faisabilité de l'application vis à vis des données d'Open Food Facts.

II. Idée d'application

A) Objectif de l'application

Avant d'aborder l'idée d'application en elle-même, il est intéressant de rappeler les indicateurs nutritionnels qui sont actuellement en place et disponibles sur les emballages :

- La déclaration obligatoire de la quantité de certains composants pour 100g : énergie, matières grasses, acides gras saturés, glucides, sucres, protéines et sel.
- La déclaration (non obligatoire) du Nutri-score, un indicateur donnant une note (A, B, C, D ou E) prenant en compte en majeure partie les données énoncées au premier point.

L'indicateur pour 100g de produit présente certaines limites :

- Il ne prend pas en compte la portion du produit à consommer.
- Il ne donne pas au consommateur une information relative à ses besoins nutritionnels facilement lisible.

Le nutri-score corrige ce dernier point en donnant une lecture claire via sa note de la valeur nutritionnelle. Néanmoins il présente pour moi une autre lacune importante, celle de ne pas positionner le produit par rapport à la catégorie (sandwichs, céréales, jus de fruits ...) à laquelle il appartient. Ainsi par exemple le nutri-score pourrait associer la notation C à un grand nombre de produits de type « chips », mettant sur le même pied tous les fabricants. Néanmoins certains peuvent potentiellement faire des efforts pour réduire les matières grasses ou le sel qui passent inaperçus à cause de ce manque d'étalement par rapport à une catégorie de produit. Quitte à choisir un produit mauvais d'un point de vue nutritionnel, pourquoi ne pas chercher à choisir tout de même le meilleur de sa catégorie ? C'est ce complément d'informations que l'application présentée dans ce rapport cherchera à apporter.

L'application proposée consiste donc un comparateur de marque sur base des données nutritionnelles de leurs produits, cela permettrait ainsi aux marques de se distinguer de manière visible des autres. Cette application aurait donc pour but de mettre en valeur les marques proposant des produits aux qualités nutritionnelles plus élevées. Cela pourrait les encourager soit à poursuivre leur efforts, soit à s'améliorer.

B) Principe de fonctionnement

- (1) L'utilisateur choisit une catégorie de produit et certaines contraintes à inclure dans le comparatif (énergie, sel, matières grasses ...)
- (2) Accès à un classement des différentes marques existantes afin de pouvoir effectuer un choix.

Afin de classer les différentes marques sur chaque critère nutritionnel et dans une catégorie précise, il serait possible de se baser sur la moyenne obtenue sur ce critère par les différents aliments appartenant à la marque.

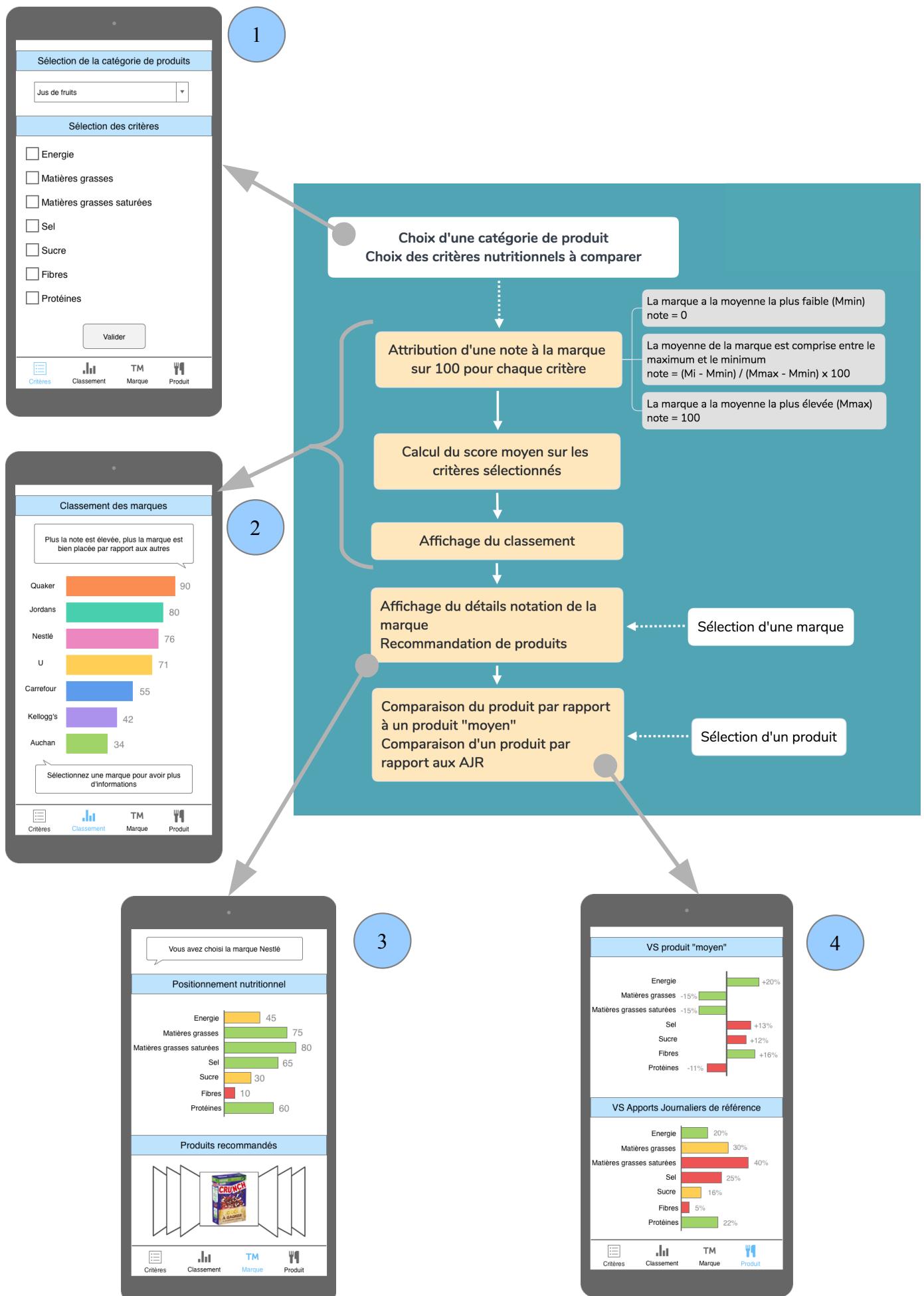
Ainsi pour chacune des composantes pour 100g il est possible d'attribuer une note sur 100 proportionnelle à la moyenne obtenue. 0 correspondant à la moyenne de la marque avec l'intérêt nutritionnel le plus faible et 100 à la moyenne de celle qui a l'intérêt nutritionnel le plus élevé.

- (3) Après avoir sélectionné une marque dans le classement, l'utilisateur accède à la notation de la marque sur tous les critères nutritionnels.

Il se verrait également recommander les produits les plus intéressants de celle-ci vis à vis des critères sélectionnés.

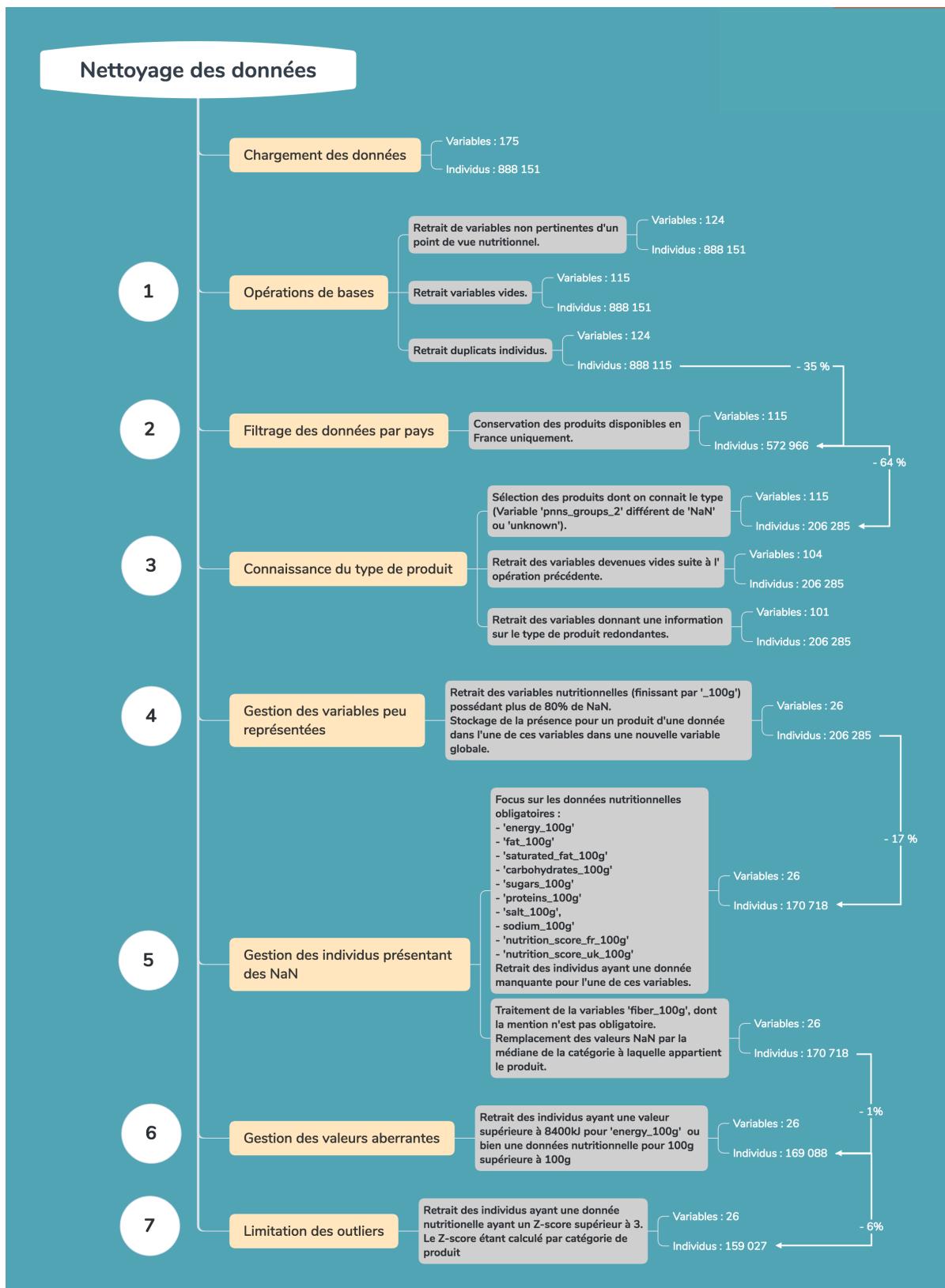
- (4) Il serait ensuite possible d'affiner la recherche en sélectionnant un aliment qui donnera accès à un descriptif positionnant celui ci par rapport à un produit « moyen » de sa catégorie et également par rapport aux Apports Journaliers de Référence (AJR). Le but étant qu'un aliment identifié comme « bon dans sa catégorie » ne soit pas interprété systématiquement comme « Bon pour la santé ».

Ci dessous une illustration du fonctionnement de l'application :



III. Préparation des données

Le dataset est composé de 175 variables et de 888 151 produits. Les variables correspondent à des données générales sur le produit et des données nutritionnelles. Le détail des variables est disponible à [cette adresse](#). Toutes les opérations décrites dans cette partie ont été réalisées de manière séquentielle, l'une se basant sur les résultats issus de la précédente. Ci dessous une visualisation concise du séquençage des opérations de nettoyage. L'état du dataset (nombre de variables et de produits) est affiché à chaque étape.



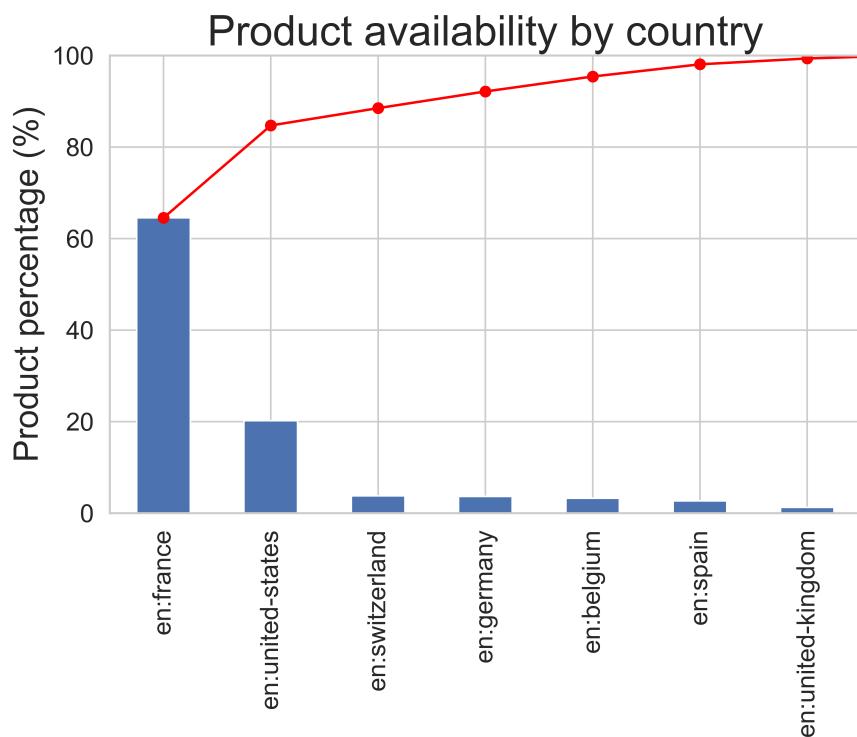
A) Opérations de bases

Dans un premier temps, trois opérations de base ont été effectuées sur le dataset afin de supprimer certaines données non nécessaires pour le projet :

- Retrait de certaines variables qui apportent peu sur la connaissance de la nature du produit et de sa composition (infos database, variables redondantes identifiées via le suffixe 'en' ou 'tags' dont on conservera un seul exemplaire, packaging). La liste des variables retirées est disponible dans l'**annexe 1 du dossier d'annexes**.
- Retrait des variables qui ne possèdent aucune information pour aucun des produits
- Retrait des produits en double (même valeurs dans tous les champs)

B) Zoom sur la France

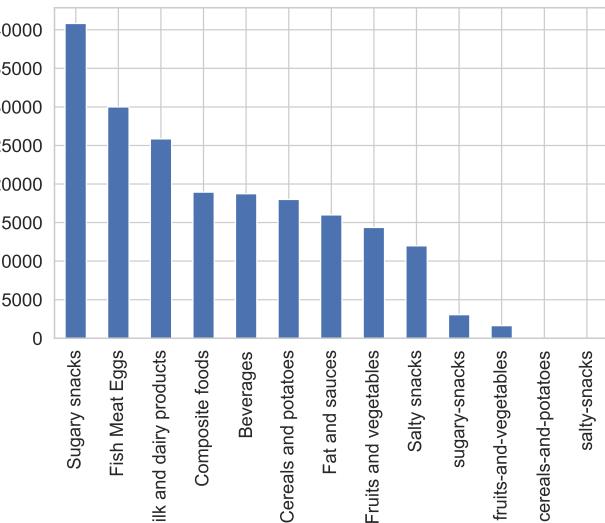
L'objectif du projet étant de proposer une idée d'application aux services de santé publique français, on conserve uniquement les produits disponibles en France. La grande majorité des produits (environ 65%) sont disponibles en France, nous sommes donc assuré de conserver suffisamment de produit pour mener à bien une analyse par la suite.



C) Sélection des produits dont on connaît la nature

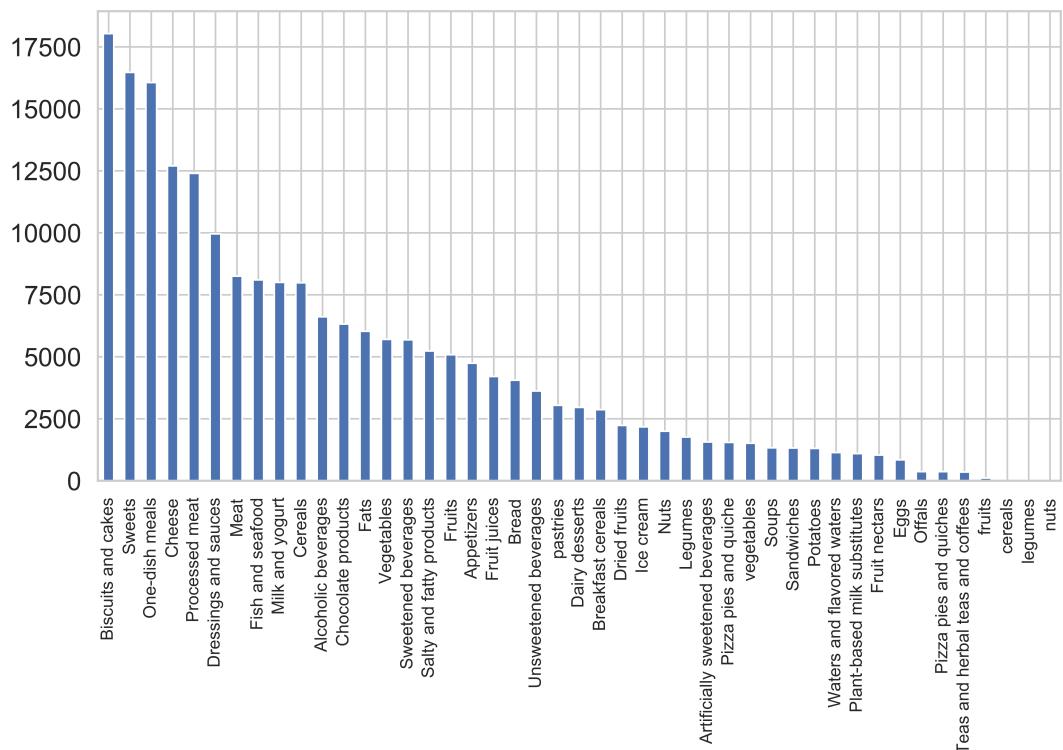
Un point essentiel pour pouvoir correctement analyser le produit est de connaître sa nature, il peut être difficile d'en extraire des informations pertinentes si cela n'est pas le cas. Pour cela on dispose de plusieurs variables : '**pnns_groups_1**', '**pnns_groups_2**' , '**main_category_en**' et '**nova_group**'. On va donc ici étudier ces 4 variables : pour chacune d'entre elles on regardera les différentes modalités prises et le nombre de fois qu'elles apparaissent.

1. le '**pnns_groups_1**'



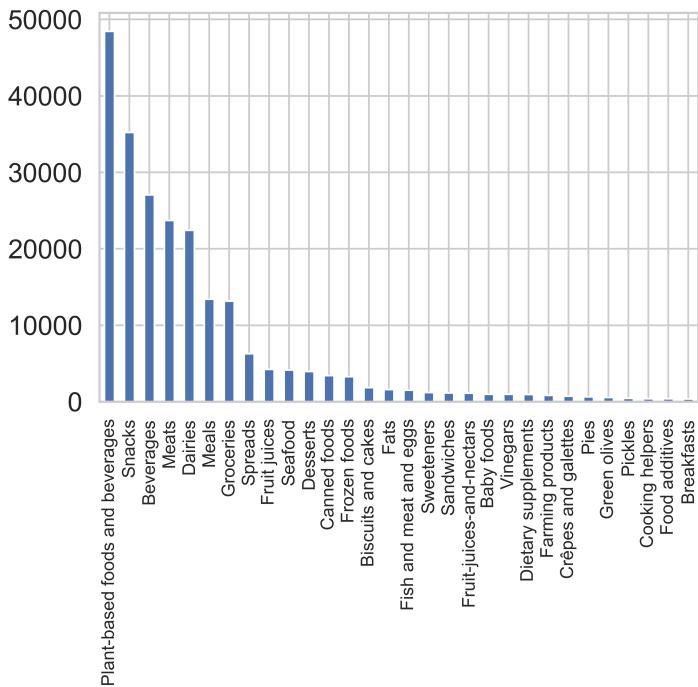
Les catégories du groupe pnns_1 restent larges et permettent de donner uniquement un premier niveau de description. A noter que l'on connaît uniquement **199 305 produits**, car de nombreux produits sont catégorisés comme 'unknown'.

2. Variable '**pnns_groups_2**'



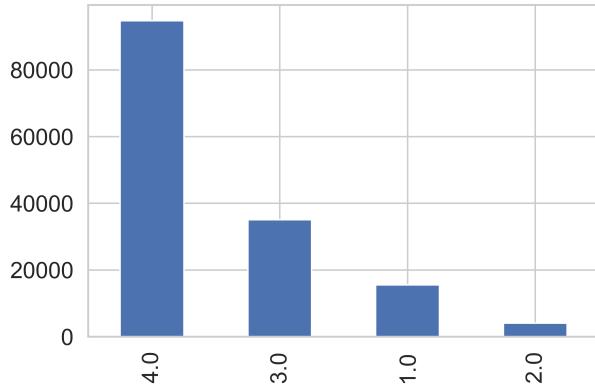
Cette fois-ci la description du type de produit est plus fine et semble plus intéressante à exploiter. On connaît uniquement le type de **206 285 produits** , pour la même raison que pour la catégorie précédente.

3. Variable 'main_category_en'



On trouve une description similaire à 'pnns_groups_2', mais moins homogène avec des redondances (par exemple on trouve la catégorie « Meat » et « Fish and meat and eggs ». Cette fois ci on dispose de l'information pour 233 888 produits, soit environ 13% de plus que pour 'pnns_groups_2'.

4. Variable 'nova_group'



La variable 'nova_group' donne au final peu d'informations sur le type de produit, indiquant uniquement le degré de transformation d'un produit. On voit qu'il s'agit en grande majorité de produits "ultra processés", généralement signe de produits industriels. On n'utilisera pas cette variable pour la description car elle apporte trop peu d'informations en comparaison des autres variables disponibles.

5. Feature engineering

Suite à l'analyse des 4 variables précédentes, on voit que la variable '**pnns_groups_2**' est la plus intéressante d'un point de vue description du produit car la plus détaillée et la plus homogène (car standardisée). Néanmoins ce n'est pas celle qui permet de conserver le plus d'individus.

En effet on voit que l'on aura au mieux **206 285** entrées disponibles sur les **572 966** du dataset filtré sur les produits français avec cette variable.

La variable '**main_category_en**' elle permet d'augmenter ce nombre à **233 888**, permettant un gain de 13% de données disponibles.

Afin d'essayer de tirer au mieux profit de ces différentes catégories descriptives il serait possible d'attribuer à une nouvelle variable '**category**' en priorité la valeur de '**pnns_groups_2**' si elle disponible sinon utiliser la valeur de '**main_category_en**' afin de récupérer un peu d'informations.

Néanmoins après analyse du dataset, on voit que dans la grande majorité là où les variables '**pnns_groups_2**' et '**pnns_group_1**' sont présentes, la variable '**main_category_en**' l'est aussi. En effet seulement **233 896** produits possèdent au moins une valeur sur ces trois variables. Le gain serait quasi nul par rapport à l'utilisation directe de la variable '**main_category_en**'.

Pour ce projet le choix a été fait de s'en tenir uniquement à la variable '**pnns_groups_2**', car plus homogène et donc nécessitant moins de temps traitement, minimisant aussi au passage le risque de redondance dans les catégories. Cela est suffisant pour une première approche.

On conservera donc uniquement les produits possédant une information dans la variable '**pnns_groups_2**'.

D) Aperçu des statistiques descriptives

Après ces premières étapes de nettoyage sommaires, il est important de regarder les statistiques descriptives afin de mieux comprendre nos données et détecter d'éventuelles anomalies ou particularités dans le dataset. Ci dessous un extrait du résumé statistique permettant d'illustrer les différents points notables de cette étape (le résumé complet est disponible en **annexe 2 du dossier d'annexes.**, les boxplots pour chaque variable numérique en **annexe 3 du dossier d'annexes.**) :

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
code	206285	206273	3.7602e+12	2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
product_name	205613	137855	Comté	598	NaN	NaN	NaN	NaN	NaN	NaN	NaN
quantity	127077	14278	500 g	6139	NaN	NaN	NaN	NaN	NaN	NaN	NaN
brands	164277	38721	Carrefour	4419	NaN	NaN	NaN	NaN	NaN	NaN	NaN
labels_tags	79370	20376	en:green-dot	5385	NaN	NaN	NaN	NaN	NaN	NaN	NaN
countries_tags	206285	743	en:france	183311	NaN	NaN	NaN	NaN	NaN	NaN	NaN
allergens	55348	17018	en:milk	4395	NaN	NaN	NaN	NaN	NaN	NaN	NaN
traces_tags	43087	4140	en:nuts	4438	NaN	NaN	NaN	NaN	NaN	NaN	NaN
serving_size	46887	6885	30 g	2810	NaN	NaN	NaN	NaN	NaN	NaN	NaN
serving_quantity	46899	NaN		NaN	114.305	212.183	0	30	79	150	36575
additives_n	132806	NaN		NaN	1.68557	2.4173	0	0	1	2	30
additives_tags	72958	20154	en:e322,en:e322i	3229	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nutrition_grade_fr	172506	5	d	54527	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pnns_groups_2	206285	40	biscuits and cakes	18031	NaN	NaN	NaN	NaN	NaN	NaN	NaN
energy_100g	182090	NaN		NaN	1154.71	825.188	0	456	1054	1674	22740
energy-from-fat_100g	112	NaN		NaN	380.148	474.3	0	0	181	589	1830
fat_100g	180787	NaN		NaN	15.3713	19.4064	0	1.1	8.2	24	100
saturated-fat_100g	180818	NaN		NaN	5.93294	8.84885	0	0.3	2.2	9	213
-lauric-acid_100g	2	NaN		NaN	47	2.82843	45	46	47	48	49
-arachidic-acid_100g	1	NaN		NaN	0.064	NaN	0.064	0.064	0.064	0.064	0.064
-cerotic-acid_100g	1	NaN		NaN	4	NaN	4	4	4	4	4
-montanic-acid_100g	1	NaN		NaN	61	NaN	61	61	61	61	61
monounsaturated-fat_100g	1768	NaN		NaN	21.1166	23.9348	0	2.8	11	27.5	82
polyunsaturated-fat_100g	1789	NaN		NaN	10.9944	16.2018	0	1.6	3.97	12	75
omega-3-fat_100g	1015	NaN		NaN	3.33039	6.35331	0	0.7	1.9	3.4	105
-alpha-linolenic-acid_100g	102	NaN		NaN	4.54653	10.8279	0.039	0.096175	0.9	4.7	75
-eicosapentaenoic-acid_100g	52	NaN		NaN	2.19829	11.7155	0	0.2	0.5765	0.925	85

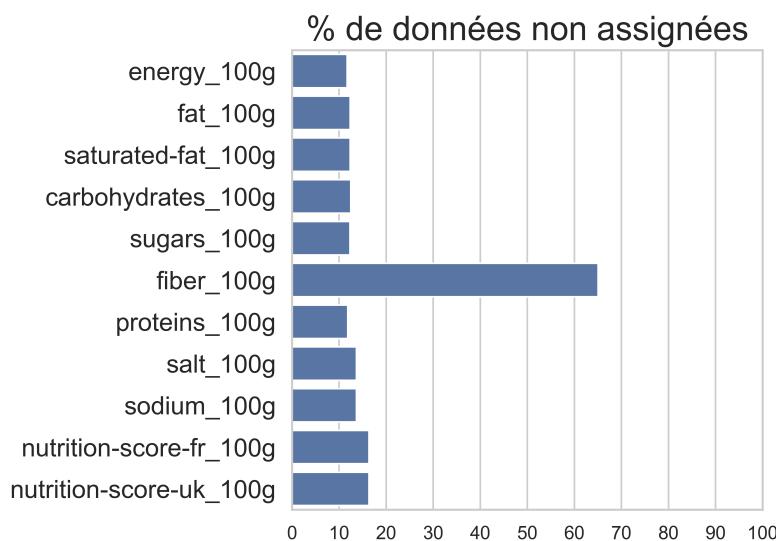
A partir de ce résumé statistique il est possible de constater que :

- Une grande partie des variables possèdent très peu de valeurs non nulles (lauric acid, cerotic acid, montanic acid ...)
- Il existe des valeurs aberrantes dans plusieurs colonnes. Par exemple des données pour 100g supérieures à 100g comme pour 'omega-3-fat_100g' ou bien encore une énergie pour 100g supérieure à trois fois les apports journaliers nécessaires pour un Homme adulte.

Ces deux points, valeurs manquantes et valeurs aberrantes vont être abordés en détails dans les sous parties suivantes.

E) Gestion des variables peu représentées

Comme constaté précédemment, de nombreuses variables possèdent un grand taux de valeurs non assignées. Pour s'assurer que les variables sont vraiment pertinentes, seules les variables ayant un **taux de valeurs non assignées inférieur à 80% ont été retenues**. Celles-ci sont visibles sur le graphique ci-dessous (la graphique avec le taux pour chaque variable est disponible en **annexe 4 du dossier d'annexes.**) :



On retrouve les informations nutritionnelles qui sont conventionnellement renseignées sur les emballages. Les fibres ont tout de même un haut taux de valeurs non assignées (environ 65%) car celles-ci ne font pas partie des éléments qu'il est impératif de renseigner.

Afin de traiter le dataset, seules ces données ont été conservées en l'état, les autres étant jugées comme très spécifiques ont été retirées. Néanmoins afin de ne pas perdre trop d'informations (les variables très peu représentées peuvent être une spécificité du produit et un élément distinctif), la présence de ces éléments « supplémentaires » ont été conservés dans un nouvelle variable nommée '**additionnals_infos_100g**'.

F) Gestion des produits ayant des variables essentielles non assignées

Dans la partie précédente les variables intéressantes suivantes ont été identifiées :

- energy_100g'
- fat_100g
- saturated_fat_100g
- carbohydrates_100g
- sugars_100g
- proteins_100g
- salt_100g
- sodium_100g
- nutrition_score_fr_100g
- nutrition_score_uk_100g
- fiber_100g

Hormis pour les fibres, le reste constitue des données nutritionnelles essentielles qui sont obligatoires lors de l'étiquetage d'un produit, il n'y a donc pas de raison qu'une de ces valeurs ne soient pas présente, si ce n'est une erreur de saisie.

Les produits pour lesquels il manque une de ces valeurs ont donc été retirés.

En ce qui concerne les fibres, les données manquantes sont plus nombreuses (car non soumises à obligation de déclaration) et simplement retirer les produits ne comportant pas cette information diminuerait le nombre de produits disponibles pour l'analyse de plus de 50%, ce qui n'est pas envisageable. Néanmoins il est également possible qu'un aliment ne contienne pas de fibres. Une exploration conjointe des variables 'pnns_groups_2' et 'fiber_100g' permet de voir les catégories d'aliments pour lesquels cette donnée n'est pas renseignée :

Product type	Nan count
biscuits and cakes	9491
processed meat	9361
cheese	9348
sweets	8087
one-dish meals	7613
dressings and sauces	5413
fish and seafood	5112
milk and yogurt	4792
meat	4694
salty and fatty products	3654
chocolate products	3545
fats	3132
cereals	2566
vegetables	2408

On voit que les types d'aliments les plus représentés sont d'origine animale (viande, fromage, poisson, lait ...) ce qui est normal. Néanmoins on observe que la donnée 'fiber_100g' est manquante dans une proportion importante pour des céréales et des légumes qui sont eux normalement riches en fibres et pour lesquels la donnée devrait être présente.

Afin de palier à ces valeurs de la variable 'fiber_100g' manquantes, il a été imputé à celles-ci la valeur de la médiane (car moins sensible aux outliers) pour la catégorie à laquelle appartient le produit.

G) Gestion des données aberrantes

Comme vu précédemment, on constate plusieurs anomalies dans le dataset : par exemple une énergie pour 100g apportant 22 740 kJ, soit presque trois fois les apports journaliers nécessaires pour une homme adulte (8400kJ) ou bien encore une teneur en sucre pour 100g supérieure à 100g.

	energy_100g	fat_100g	saturated_fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutrition_score_fr_100g	nutrition_score_uk_100g
count	170718,00	170718,00	170718,00	170718,00	170718,00	170717,00	170718,00	170718,00	170718,00	170718,00	170718,00
mean	1160,84	15,62	6,10	24,44	12,44	1,85	8,47	0,98	0,39	9,76	9,56
std	816,43	19,03	8,87	26,40	18,26	3,11	8,62	1,99	0,78	8,81	9,39
min	0,00	0,00	0,00	0,00	-0,10	0,00	0,00	0,00	0,00	-15,00	-15,00
25,00%	465,00	1,40	0,30	2,10	0,70	0,00	1,70	0,08	0,03	2,00	1,00
50,00%	1054,00	9,00	2,30	11,80	3,30	1,20	6,20	0,60	0,24	11,00	10,00
75,00%	1690,00	24,00	9,10	49,00	16,00	2,50	12,30	1,30	0,51	16,00	18,00
max	22740,00	100,00	213,00	298,00	240,00	400,00	808,00	125,00	49,21	40,00	37,00

Afin de corriger cela les produits dont l'énergie pour 100g dépasse les AJR d'un adulte ou l'une des données pour 100g dépasse les 100g ont été retirés. Ci-dessous les statistiques descriptives pour ces mêmes variables après cette opération :

	energy_100g	fat_100g	saturated_fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutrition_score_fr_100g	nutrition_score_uk_100g
count	169088,00	169088,00	169088,00	169088,00	169088,00	169088,00	169088,00	169088,00	169088,00	169088,00	169088,00
mean	1139,41	14,97	5,98	24,51	12,41	1,85	8,54	0,99	0,39	9,74	9,47
std	783,47	17,55	8,52	26,24	18,01	2,88	8,34	1,91	0,75	8,85	9,39
min	0,00	0,00	0,00	0,00	-0,10	0,00	0,00	0,00	0,00	-15,00	-15,00
25,00%	464,00	1,40	0,30	2,20	0,70	0,00	1,90	0,09	0,04	2,00	1,00
50,00%	1046,00	8,80	2,30	12,00	3,40	1,20	6,20	0,60	0,24	11,00	10,00
75,00%	1674,00	24,00	9,00	49,00	16,00	2,50	12,45	1,30	0,51	16,00	18,00
max	8242,00	99,99	99,90	99,90	99,90	99,00	92,00	99,00	38,98	40,00	37,00

On constate que certaines valeurs très extrêmes restent présentes (92g de protéines pour 100g de produit par exemple). Si l'on regarde de plus près ces valeurs, et en particulier le type de produit qui leur sont associés, on distingue deux cas :

- Des erreurs, comme pour 'fiber_100g' et 'salt_100g' car un steak haché ne contient pas de fibre et la limonade ne contient pas autant de sel.
- Des produits atypiques quasi mono composant comme l'huile d'olive qui est pure matière grasse ou bien encore un complément alimentaire qui est en grande partie de la protéine.

H) Réduction des outliers à partir du Z-score

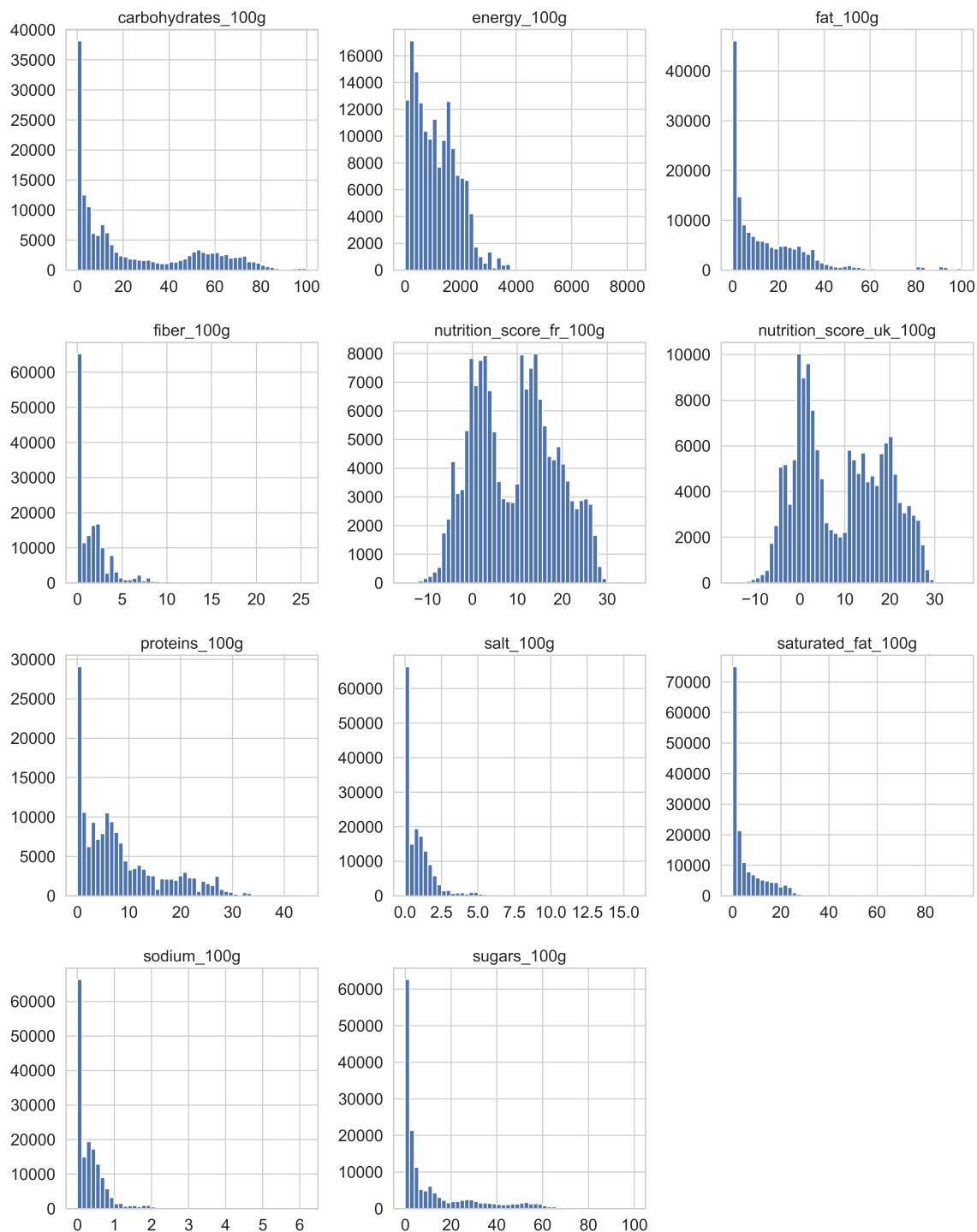
Dans cette partie on essaiera de limiter la présence de produits très atypiques ou bien encore d'erreurs en calculant le Zscore pour chaque composante nutritionnelle d'un produit relativement à la catégorie à laquelle il appartient. De manière conventionnelle, on conservera les produits pour lesquels aucune des données nutritionnelles n'affiche un Zscore supérieur à 3.

	energy_100g	fat_100g	saturated_fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutrition_score_fr_100g	nutrition_score_uk_100g
count	159027,00	159027,00	159027,00	159027,00	159027,00	159027,00	159027,00	159027,00	159027,00	159027,00	159027,00
mean	1118,93	14,66	5,84	24,10	12,15	1,70	8,33	0,89	0,35	9,55	9,30
std	783,96	17,46	8,34	26,13	17,62	2,16	7,95	1,13	0,45	8,84	9,42
min	0,00	0,00	0,00	0,00	-0,10	0,00	0,00	0,00	0,00	-15,00	-15,00
25,00%	448,00	1,30	0,30	2,00	0,70	0,00	1,80	0,09	0,04	2,00	1,00
50,00%	1017,00	8,50	2,20	11,40	3,30	1,20	6,10	0,60	0,24	11,00	9,00
75,00%	1661,00	23,10	8,80	48,90	16,00	2,50	12,00	1,30	0,51	16,00	18,00
max	8242,00	99,99	95,10	99,90	99,90	25,60	44,30	15,70	6,18	36,00	36,00

Après application, on voit que certains produits atypiques comme le complémentaire alimentaire en protéines ou bien les erreurs sur la fibre et le sel mentionnées précédemment ont été supprimées, ce qui devrait permettre de mieux comparer les produits entre eux par la suite.

IV. Analyse des données

A) Distribution des variables importantes

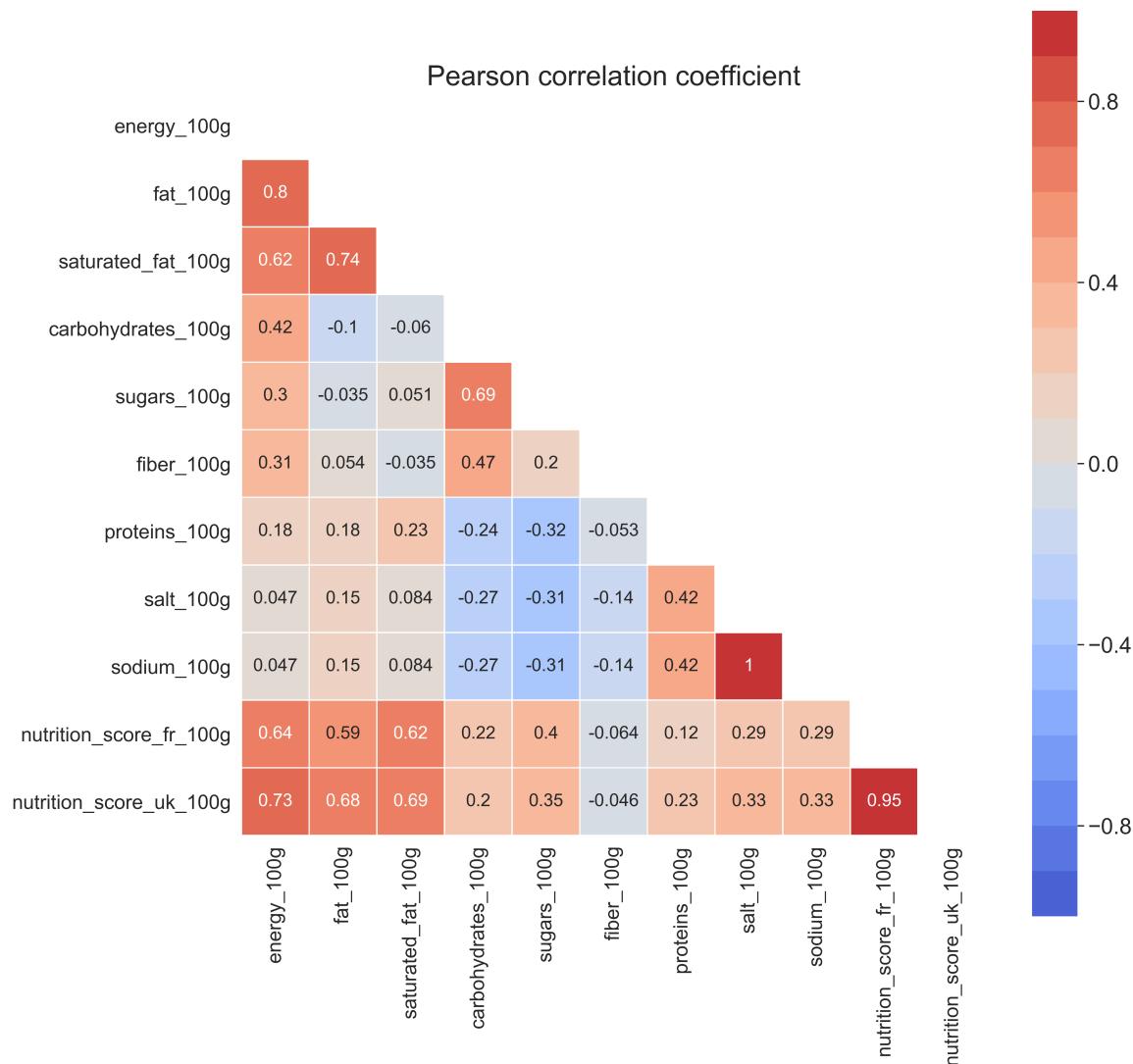


On peut observer plusieurs points à partir de ces histogrammes :

- Les différentes variables ne suivent pas une loi normale
- Le nutri-score a une distribution bi-modale, avec une majorité de produit répartis sur des scores correspondant à des notations à la limite du B et du C pour le premier pic et à D pour le second.
- Pour la quasi totalité des variables (hors nutri score) on note la présence d'outliers (par exemple les produits proches de 100g pour la graisse). Les limites sur l'axe des abscisses du graphique sont définies par rapport à ces valeurs extrêmes.

B) Corrélations linéaires entre les variables importantes

Il est possible d'obtenir un peu plus d'informations en observant le niveau de corrélation des différentes variables entre elles :



On observe les points suivants :

- Le sodium et le sel sont complètement corrélés positivement, ce qui paraît normal étant donné que le sel (Chlorure de sodium) est en partie composé de sodium.
- Les sucres et les glucides sont aussi assez fortement corrélés positivement, les sucres étant une sous catégorie des glucides. On peut faire le même constat pour les graisses saturées et la matière grasse.
- Les fibres sont modérément corrélées aux glucides. Cela s'explique par le fait que l'on trouve des fibres dans les féculents qui sont aussi très riches en glucides mais elles sont également présentes en forte quantité dans les légumes qui eux peuvent contenir moins de glucides.
- L'énergie semble fortement corrélée de manière positive à la présence de graisse et dans une moindre mesure aux sucres et glucides.
- Le nutrition-score lui est fortement corrélé de manière positive à l'énergie et aux matières grasses. Dans une moindre mesure, le sucre et le sel semblent également avoir un impact.

C) Comparaison des différentes marques

Le cœur de l'application réside dans la possibilité de positionner les différentes marques pour une catégorie de produit donnée. Afin d'évaluer l'impact de la marque sur une caractéristique nutritionnelle spécifique au sein d'une catégorie de produit, les étapes suivantes ont été réalisées :

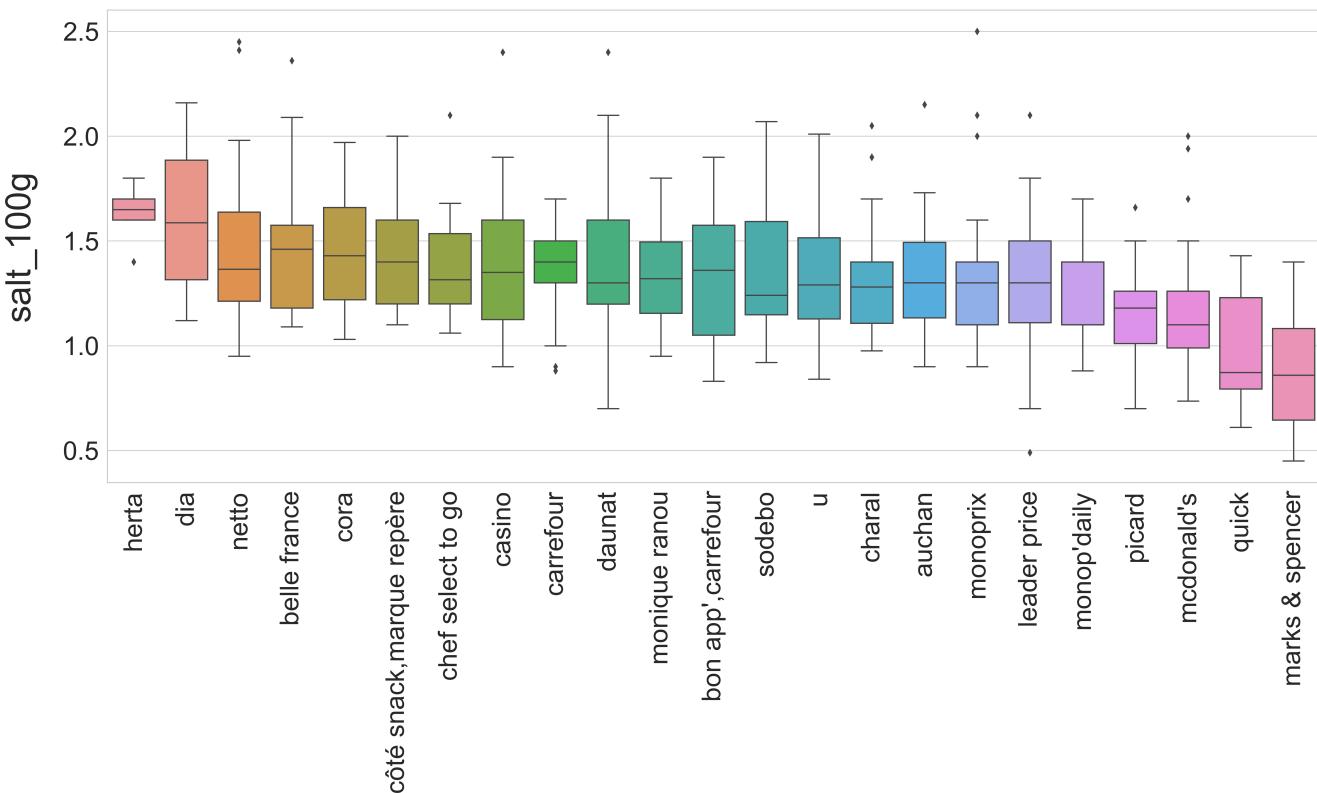
- Visualisation pour une catégorie de produit de la distribution d'une donnée nutritionnelle choisie pour les différentes marques enregistrées. Les marques sont ordonnées de gauche à droite par ordre de moyenne décroissante.
- Validation de l'impact de la marque en tant que critère différenciateur via un test sur les moyennes de Kruskal-Wallis. L'hypothèse nulle de ce test est l'égalité des moyennes et l'hypothèse alternative la différence entre celles-ci. Le test non paramétrique de Kruskal Wallis a été choisi car après plusieurs tests, il a été constaté que les conditions nécessaires à l'application d'une ANOVA n'étaient pas toujours remplies, notamment celle de l'égalité des variances.
- Comparaison visuelle de la distribution d'une information nutritionnelle donnée pour deux marques, avec une catégorie de produits fixée.
- Validation ou non de la significativité de la différence entre les deux marques via un test sur les moyennes de Mann-Whitney. L'hypothèse H_0 est l'égalité des moyennes et l'hypothèse alternative H_1 la supériorité de la moyenne de la première marque par rapport à la seconde. Une fois encore le test non-paramétrique a été préféré à son équivalent paramétrique car certains prérequis comme la normalité de la distribution ou l'égalité des variances n'étaient pas toujours respectés.

Ces quatre étapes seront réalisées à titre d'exemple sur les catégories et données nutritionnelles suivantes :

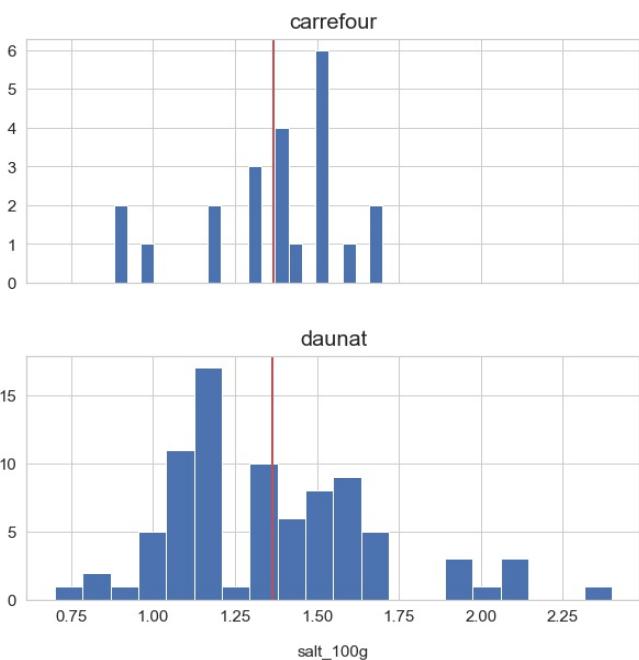
- Impact de la marque sur le taux de sel dans les sandwichs (marques ayant au moins 10 produits enregistrés) avec comparaison des marques Carrefour et Daunat.
- Impact de la marque sur le taux de matières grasses dans les apéritifs (marques ayant au moins 20 produits enregistrés) avec comparaison des marques U et Pringles.
- Impact de la marque sur le taux de sucre dans les céréales (marques ayant au moins 20 produits enregistrés) avec comparaison des marques Auchan et Quacker.

1. Impact de la marque sur le taux de sel dans les sandwichs

Comparison of salt in sandwiches



On observe que la marque 'Herta' semble avoir le taux moyen de sel le plus élevé et la marque 'marks & spencer' le plus bas. Visuellement, il est difficile de voir si une marque est significativement moins salée qu'une autre. Néanmoins, la p-value de l'ordre de 10^{-12} obtenue lors du test de Kruskal-Wallis indique que la marque a bien un effet sur le taux de sel dans les sandwiches.



Comparaison des marques **Carrefour** et **Daunat** pour la variable **salt_100g** :

Nombre de produits dans chaque échantillon :
Carrefour : 22
Daunat : 84

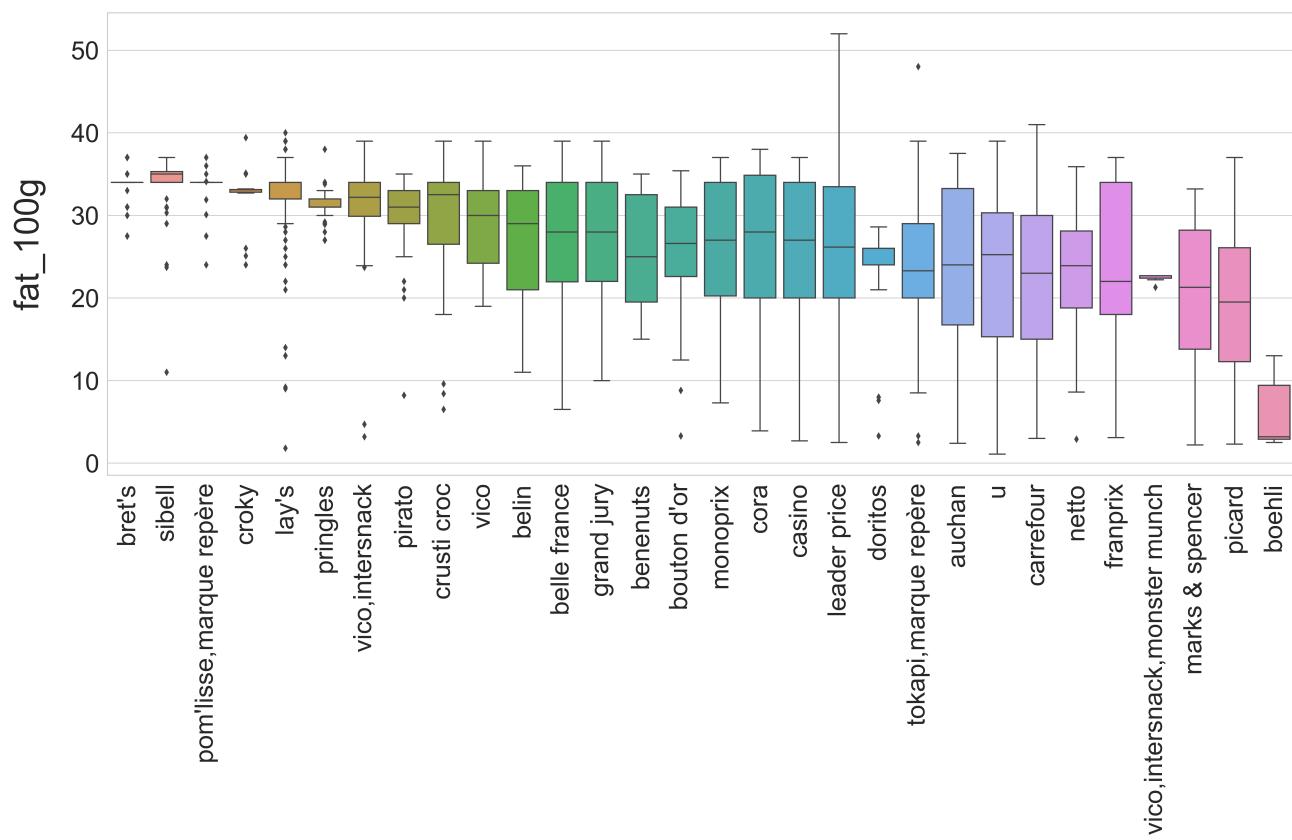
Moyenne et écart type pour chaque échantillon :
Pour la marque Carrefour
moyenne : 1.37
écart type : 0.22

Pour la marque Daunat :
moyenne : 1.36
écart type : 0.32

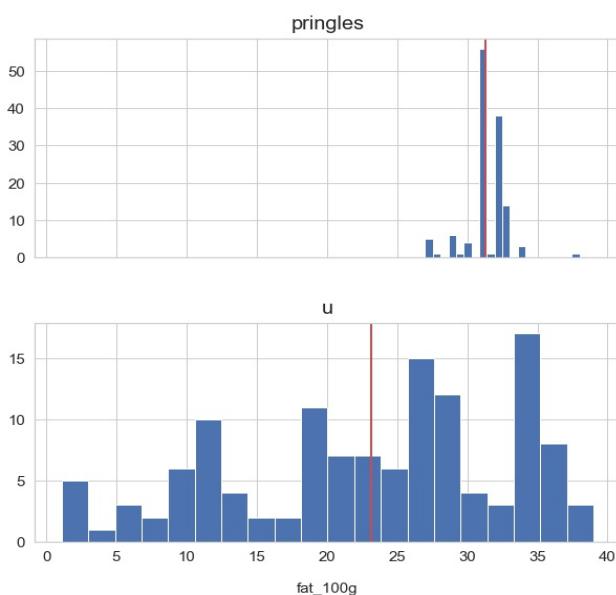
Visuellement on voit que les moyennes sont très proches et cela se confirme par la p-value de 0.225 pour le test de Mann Whitney qui ne permet pas de rejeter l'hypothèse d'égalité de ces dernières. Le taux de sel moyen est similaire pour les sandwichs de la marque Carrefour et Daunat.

2. Impact de la marque sur le taux de matières grasses dans les apéritifs

Comparison of fat in appetizers



On observe un groupe de marques (de Bret's à Pirato) qui semblent avoir un taux moyen de matières grasses plutôt élevé et une faible dispersion. S'en suit un groupe avec un taux qui baisse de plus en plus et surtout une dispersion beaucoup plus importante (hors Doritos) pour lesquelles il est difficile de savoir si l'une a une tendance à être plus grasse que l'autre. La marque 'Boehli' elle se démarque par son taux de matière grasse très faible. Au global, le test de Kruskal-Wallis indique bien que la marque a une influence sur le taux de matières grasses dans les apéritifs (p-value de l'ordre de 10^{-94}).



Comparaison des marques Pringles et U pour la variable fat_100g :

Nombre de produits dans chaque échantillon :
Pringles : 130
U : 128

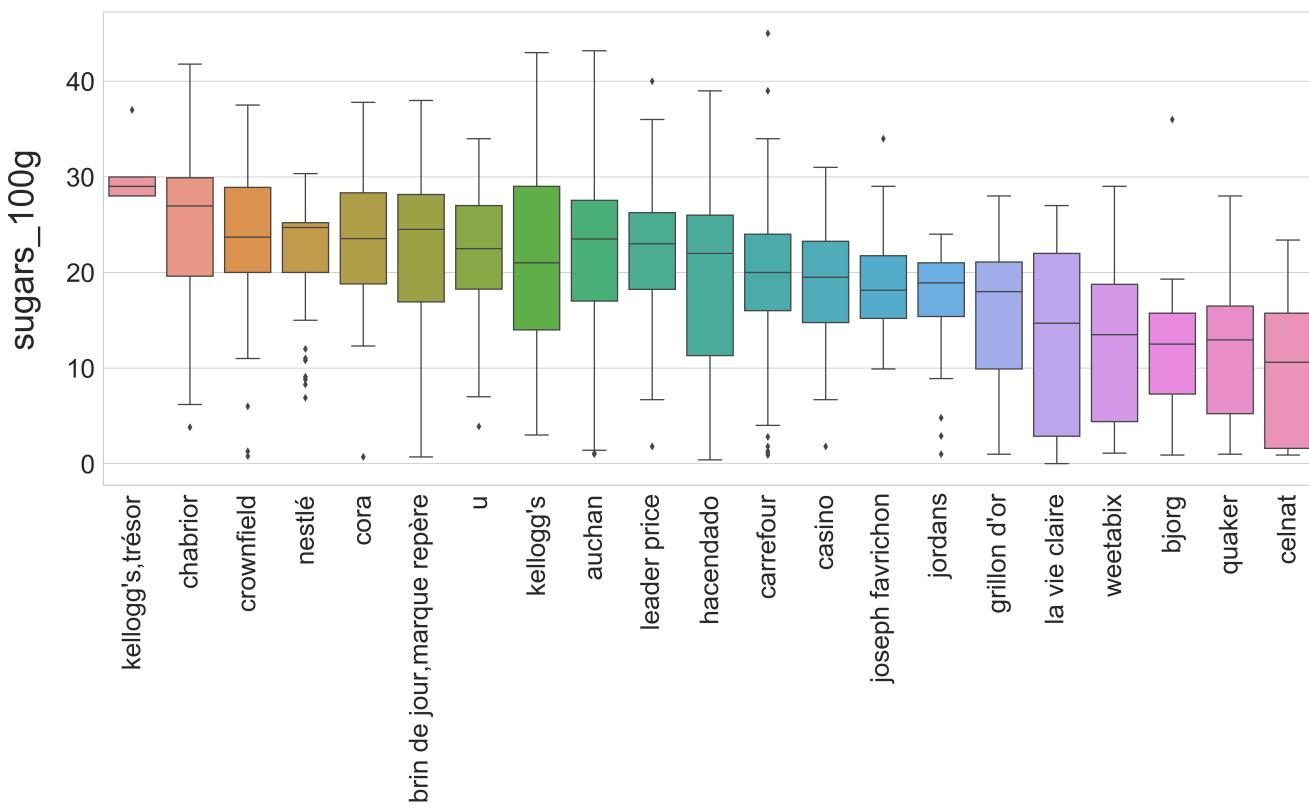
Moyenne et écart type pour chaque échantillon :
Pour la marque Pringles
moyenne : 31.32
écart type : 1.47

Pour la marque U :
moyenne : 23.12
écart type : 9.94

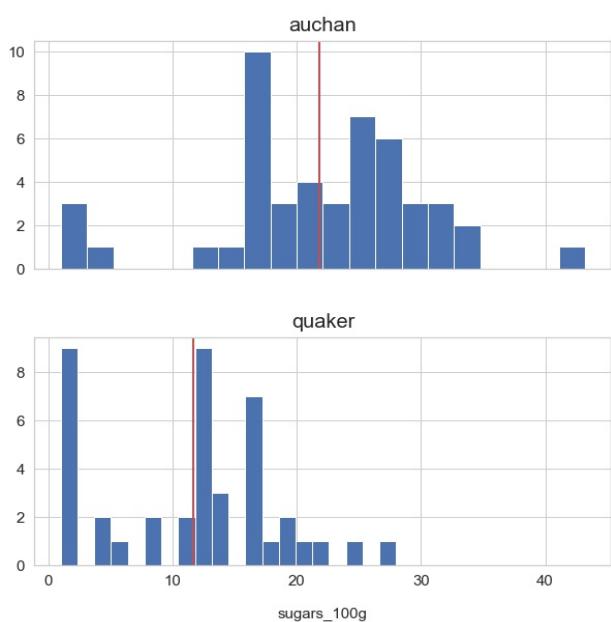
Visuellement il est flagrant que la marque Pringles a tendance à être plus grasse que la marque U. Cela se vérifie avec la p-value du test de Mann Whitney qui est de l'ordre de 10^{-13} .

3. Impact de la marque sur le taux de sucre dans les céréales

Comparison of sugars in breakfast cereals



A part peut être pour les trois dernières marques (Bjorg, Quaker, Celnat), il est visuellement difficile de voir si une marque est significativement moins salée qu'une autre si l'on les compare une à une. Néanmoins on observe un taux qui a tendance à baisser régulièrement. La p-value du test de Kruskal-Wallis est de l'ordre de 10^{-35} confirme cette observation, la marque a donc bien une influence sur le taux de sucre des céréales.



Comparaison des marques **Auchan** et **Quaker** pour la variable **sugars_100g** :

Nombre de produits dans chaque échantillon :

Auchan : 48

Quaker : 42

Moyenne et écart type pour chaque échantillon :

Pour la marque Auchan :

moyenne : 21.80

écart type : 8.64

Pour la marque Quaker :

moyenne : 11.63

écart type : 7.15

Visuellement il est flagrant que la marque de céréales Auchan a tendance à être plus sucrée que la marque Quaker. Cela se vérifie avec la p-value du test de Mann Whitney qui est de l'ordre de 10^{-8} .

V. Conclusion

Au travers de la partie « Analyse des données » de ce rapport, on constate que mettre en place un comparateur de marque sur ses données nutritionnelles fait sens car de réelles différences existent. Les exemples présentés mettent en avant le fait que pour une même catégorie de produit, certaines marques possèdent de réels atouts nutritionnels à faire valoir par rapport à leur concurrents. Cet indicateur serait un complément intéressant au Nutri-Score pour les consommateurs afin de leur permettre de faire des choix avisés concernant leur alimentation.

Pour son fonctionnement, l'application aurait au minimum besoin des données suivantes :

- Nom du produit
- Catégorie du produit
- les données pour 100g dont la mention est obligatoire : énergie, matières grasses, acides gras saturés, glucides, sucres, protéines, sel
- La masse de fibre pour 100g, dont la mention n'est pas obligatoire

Ces données sont supposées être disponibles sur Open Food Facts.

Il pourrait également être intéressant d'inclure dans les critères de notation d'autres aspects non abordés dans ce rapport, comme par exemple les allergènes et les additifs.

Néanmoins il existe des freins à la mise en place d'une telle application à partir du set de données disponible sur Open Food Facts. Ceux ci sont en grande partie dû à la manière dont est complétée la base de données, rendant nécessaires un grand nombre d'opérations de traitement pour obtenir un résultat exploitable. En effet l'ajout d'un produit sur Open Food Facts s'effectue de la manière suivante (extrait du site de l'organisme) :

Ajoutez des produits

Utilisez notre app [Android](#), [iPhone](#) ou
[Windows Phone](#) pour scanner le code
barre des produits que vous possédez ou
de vos magasins préférés et envoyer des
photos de leur étiquette.

Pas de smartphone ? Pas de problème :
vous pouvez tout aussi bien utiliser un
appareil photo pour ajouter des produits
directement sur le site web.

Sur le site web, vous pourrez également
remplir les informations des produits que
vous ajoutez ou que d'autres ont ajoutés.

Avec ce mode de fonctionnement, un aliment peut exister dans la base sans ses informations nutritionnelles qui doivent être insérées par un utilisateur à partir d'une photographie de l'emballage. On constate également que les données sont saisies manuellement par un utilisateur, ce qui est une source d'erreurs importante.

Les limites qui en découlent, observées lors de l'utilisation de ce dataset, sont rassemblées dans le tableau ci-dessous, avec des solutions pour permettre l'exploitation des données pour l'application :

Limites	Solutions
Doublons des produits qui sont difficiles à identifier car le nom ou la description changent légèrement.	Considérer les produits ayant des similarités dans leur description et des données nutritionnelles quasi identiques comme un même produit.
Les marques sont parfois nommées de manière différentes, les contributeurs de la database OpenFoodFacts ajoutant parfois le nom de la gamme de produit, des labels ou le groupe auquel la marque appartient	Pour les nouvelles données : Guider l'utilisateur en lui suggérant une marque lorsqu'il rempli le champ en fonction de ce qui existe déjà dans la base de données. Pour les données déjà enregistrées : Attribuer une catégorie au produit en se basant sur les autres données de marque disponibles. (Utilisation de techniques de NLP)
Le nombre de produits disponibles pour certaines marques sont parfois très faibles (inférieurs à 5) limitant la possibilité de réaliser des tests statistiques paramétriques pertinents (perte de puissance).	Utiliser des test statistiques non paramétriques pour les échantillons de petite taille.
Un quart des produits (et environ 30% pour les produits disponibles en France) ne possèdent pas d'information sur sa catégorie rendant le positionnement du produit difficile.	Pour les nouvelles données : Guider l'utilisateur en lui suggérant une catégorie ou en lui proposant une liste définie. Pour les données déjà enregistrées : Attribuer une catégorie au produit en se basant sur les autres données (ingrédients, données nutritionnelles ...).