

Préparez des données pour un organisme de santé publique

Nettoyage et exploration du jeu de données
openfoodfacts





Objectif : Nettoyer et explorer un jeu de données

Enjeux : Déterminer la faisabilité et le cas échéant suggérer une idée d'application d'auto-complétion

Plan :

1. Description du jeu de données et notion de nutriscore
 - 1.1. Dataset et sélection de variables
 - 1.2. Définition et calcul du nutriscore
2. Démarche de prétraitement de données
 - 2.1. Visualisation et analyse
 - 2.2. Traitement des valeurs aberrantes
 - 2.3. Traitement des valeurs manquantes
3. Analyses univariée et bivariées
 - 3.1. Analyse univariée, distribution des variables
 - 3.2. Analyse bivariées : corrélations et distributions
 - 3.3. ANOVA et test du Chi-2
4. Analyse multivariées et suggestion d'application
 - 4.1. Analyse en composantes principales
 - 4.2. Faisabilité et suggestion d'une application

Conclusion et conformité RGPD



1. Description du jeu de données et notion de nutriscore

1.1. Dataset et sélection de variables

Provenance des données : site d'openfoodfacts - téléchargement du jeu de données mis à disposition

Jeu de données volumineux :

- 320 772 lignes correspondant à des produits
- 162 colonnes correspondant à diverses caractéristiques de ces produits.

code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	quantity	...
3087	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	NaN	1kg	...
4530	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)	NaN	NaN	...
4559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	NaN	NaN	...
16087	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	NaN	NaN	...
16094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	NaN	NaN	...



1. Description du jeu de données et notion de nutriscore

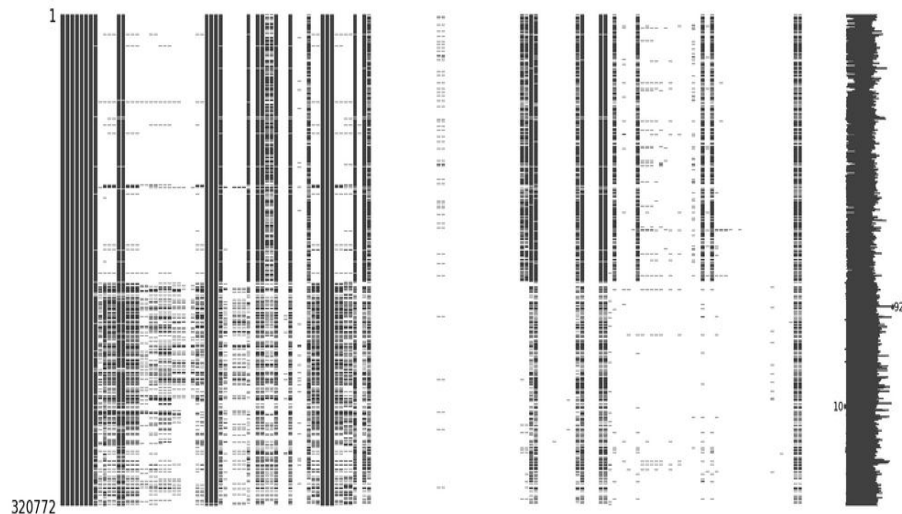
1.1. Dataset et sélection de variables

Sélection de variables selon 2 critères :

- taux de **valeurs manquantes**
- **importance** / problématique

pnns_groups_1	energy_100g	saturated-fat_100g	sugars_100g	salt_100g	sodium_100g	fiber_100g	proteins_100g	nutrition-score-fr_100g	nutrition_grade_fr
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	2243.00	28.57	14.29	0.00	0.00	3.60	3.57	14.00	d
NaN	1941.00	0.00	17.86	0.64	0.25	7.10	17.86	0.00	b
NaN	2540.00	5.36	3.57	1.22	0.48	7.10	17.86	12.00	d
NaN	1552.00	NaN	NaN	NaN	NaN	5.70	8.57	NaN	NaN
...
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	b
Salty snacks	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	2092.00	NaN	0.00	0.00	0.00	NaN	0.00	NaN	NaN

Matrice des valeurs manquantes :










1. Description du jeu de données et notion de nutriscore

1.2. Définition et calcul du nutriscore

Principe :

Attribution d'un logo sur la base d'un score prenant en compte pour 100g (ou 100ml) de produit la teneur :

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

Score nutritionnel			Catégorie	Nutri-Score
Aliments solides	Matières grasses animales et végétales, noix et graines	Boissons		
Minimum à 0	Minimum à -6	Eaux	A	
1 à 2	-5 à 2	Minimum à 2	B	
3 à 10	3 à 10	3 à 6	C	
11 à 18	11 à 18	7 à 9	D	
19 à Maximum	19 à max	10 à Maximum	E	



2. Démarche de prétraitement de données

2.1. Visualisation et analyse

Statistiques descriptives de nos données numériques :

	energy_100g	saturated-fat_100g	sugars_100g	salt_100g	sodium_100g	fiber_100g	proteins_100g	nutrition-score-fr_100g
count	261113.00	229554.00	244971.00	255510.00	255463.00	200886.00	259922.00	221210.00
mean	1141.91	5.13	16.00	2.03	0.80	2.86	7.08	9.17
std	6447.15	8.01	22.33	128.27	50.50	12.87	8.41	9.06
min	0.00	0.00	-17.86	0.00	0.00	-6.70	-800.00	-15.00
25%	377.00	0.00	1.30	0.06	0.03	0.00	0.70	1.00
50%	1100.00	1.79	5.71	0.58	0.23	1.50	4.76	10.00
75%	1674.00	7.14	24.00	1.37	0.54	3.60	10.00	16.00
max	3251373.00	550.00	3520.00	64312.80	25320.00	5380.00	430.00	40.00

La documentation sur le nutriscore et le site du Ciqua (<https://ciqua.anses.fr/>) nous ont permis de repérer les valeurs maximum aberrantes. Concernant les valeurs négatives, nous avons fait le choix de les supprimer, excepté pour le nutriscore.

2. Démarche de prétraitement de données

2.2. Traitement des valeurs aberrantes

Illustration du traitement des valeurs aberrantes pour la variable concernant les calories :

Tableau 1 : Points attribués pour chacun des éléments défavorables de la composante N

Points	Energie (KJ/100g)	Acides gras saturés (g/100g)	Sucre (g/100g)	Sodium* (mg/100g)
0	≤ 335	≤ 1	≤ 4.5	≤ 90
1	> 335	> 1	> 4.5	> 90
2	> 670	> 2	> 9	> 180
3	> 1005	> 3	> 13.5	> 270
4	> 1340	> 4	> 18	> 360
5	> 1675	> 5	> 22.5	> 450
6	> 2010	> 6	> 27	> 540
7	> 2345	> 7	> 31	> 630
8	> 2680	> 8	> 36	> 720
9	> 3015	> 9	> 40	> 810
10	> 3350	> 10	> 45	> 900

*: le sodium correspond au sel dans la déclaration nutritionnelle obligatoire divisé par 2.5.

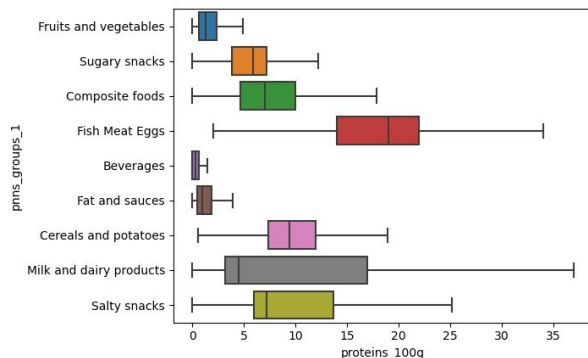
Energie, N x facteur Jones, avec fibres (kJ/100 g)	
Tri croissant/décroissant : Teneurs décroissantes	
Filtrer par le nom de l'aliment Nom	
Nom	Teneur m
Huile combinée, mélange d'huile d'olive et de graines	3700
Huile d'amande	3700
Huile d'amandes d'abricot	3700
Huile d'arachide	3700
Huile d'avocat	3700
Huile d'olive vierge extra	3700
Huile de carthame	3700
Huile de colza	3700
Huile de coton	3700
Huile de foie de morue	3700
Huile de germe de blé	3700
Huile de hareng	3700
Huile de lin	3700
Huile de maïs	3700

L'observation de ces deux éléments nous a conduit à supprimer les valeurs supérieures à 4000KJ/100g.

2. Démarche de prétraitement de données

2.3. Traitements des valeurs manquantes

- Suppression : **ensemble des valeurs “nutritives” manquantes** (44 223 produits);
- Imputation par la médiane : Un **produit** pour lequel la **donnée “protéine”** était **manquante** s’est vu **imputer la médiane** de la valeur en protéines **de sa catégorie PNNS 1** (2150 produits soit une proportion de 0.83 %).
- **Calcul** : 1g protéines = 17KJ, 1g glucides = 17KJ et 1g de lipides = 38KJ
Variable : énergie/100g
- **KNN Imputer (5 voisins)** : autres variables numériques
- **Imputation “conditionnelle”** : nutrigrade en fonction de la valeur du nutriscore



Score nutritionnel			Catégorie	Nutri-Score
Aliments solides	Matières grasses animales et végétales, noix et graines	Boissons		
Minimum à 0	Minimum à -6	Eaux	A	
1 à 2	-5 à 2	Minimum à 2	B	
3 à 10	3 à 10	3 à 6	C	
11 à 18	11 à 18	7 à 9	D	
19 à Maximum	19 à max	10 à Maximum	E	



3. Analyses univariée et bivariées

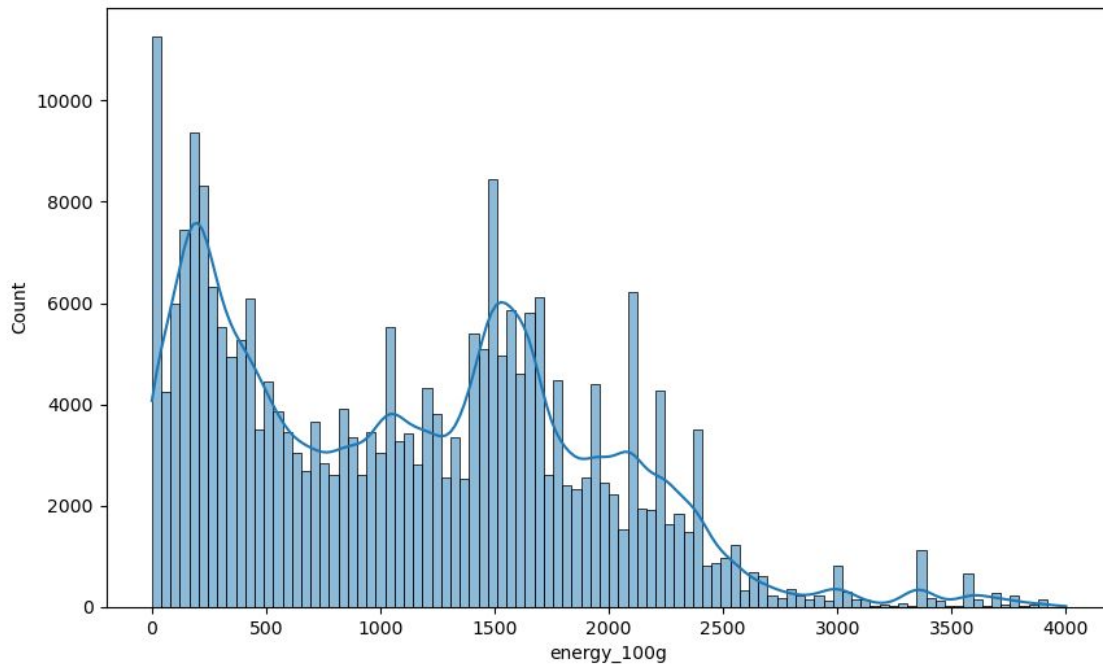
3.1 Analyse univariée, distribution des variables

DataFrame de **258401**
lignes, et donc produits, et
12 colonnes.

Chaque **variable numérique**
a un mode égal à 0.

Visualisation de leur
distribution compliquée,
exceptée pour l'énergie et le
nutriscore :

Histogramme de la distribution de la variable énergie/100g

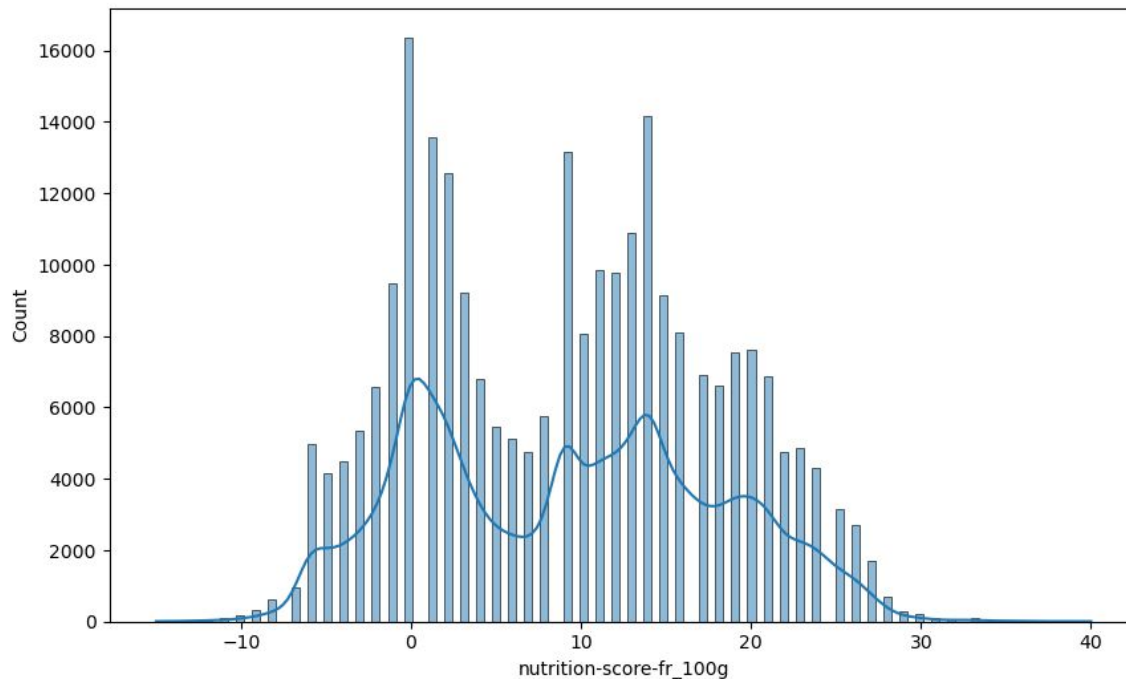




3. Analyses univariée et bivariées

3.1 Analyse univariée, distribution des variables

Histogramme de la distribution de la variable nutriscore



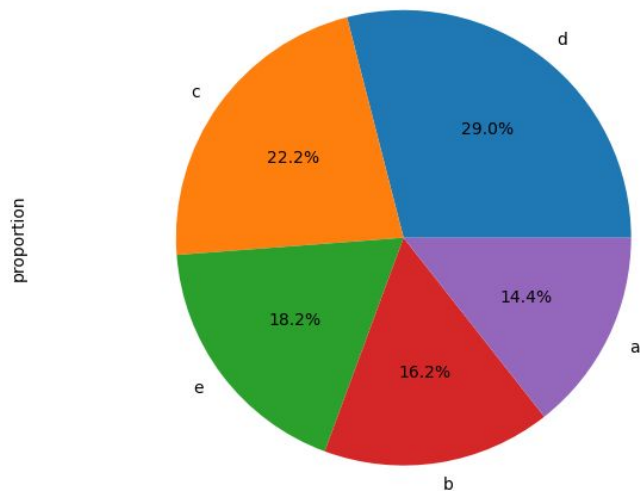


3. Analyses univariée et bivariées

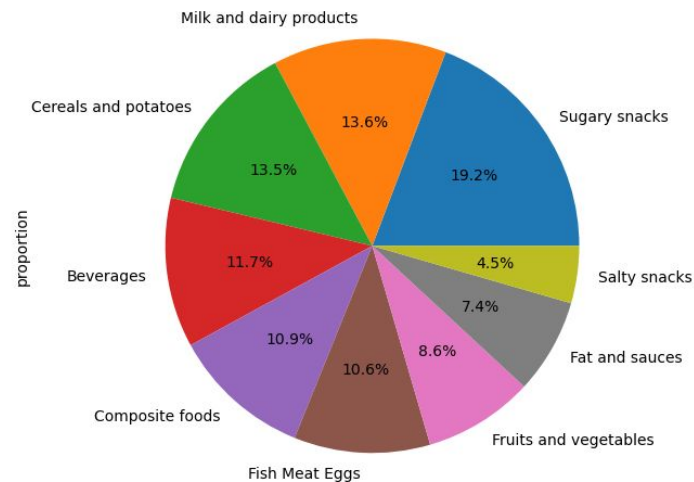
3.1 Analyse univariée, distribution des variables

Variables catégorielles : catégorie PNNS 1 et nutrigrade

Répartition des effectifs dans chaque catégorie de nutrigrade



Répartition des effectifs dans chaque catégorie PNNS (à l'exception de 'unknown')

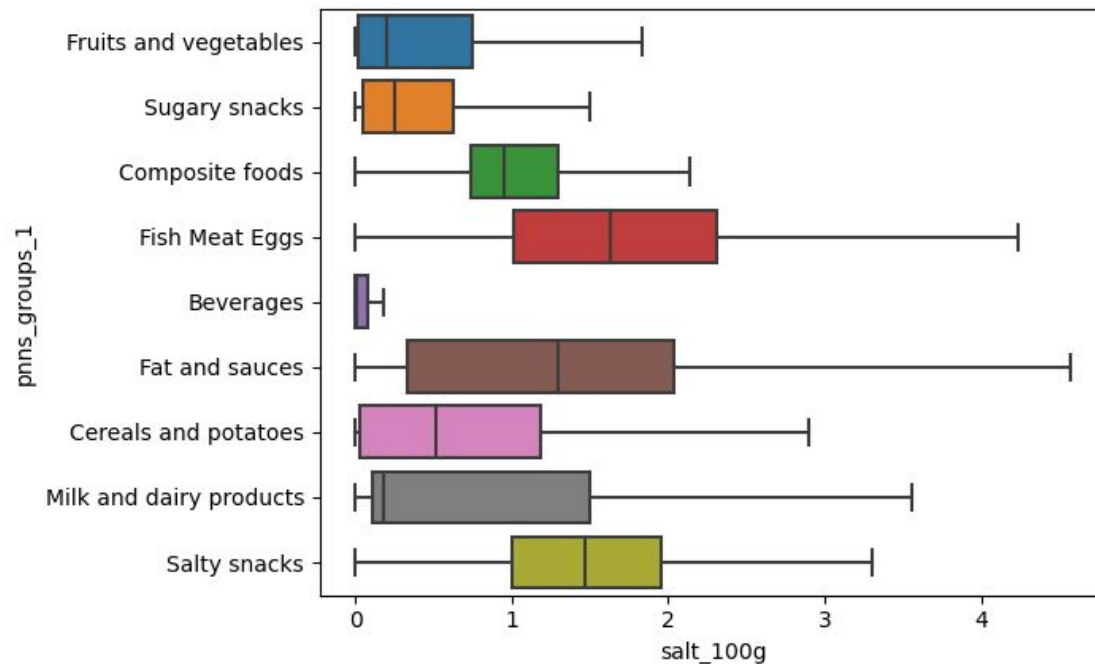


3. Analyses univariée et bivariées

3.2. Analyse bivariées : corrélations et distributions

Utilisation de **boxplots**, associée à des variables catégorielles.

Exemple du **sel**, en fonction de la **catégorie PNNS 1** :



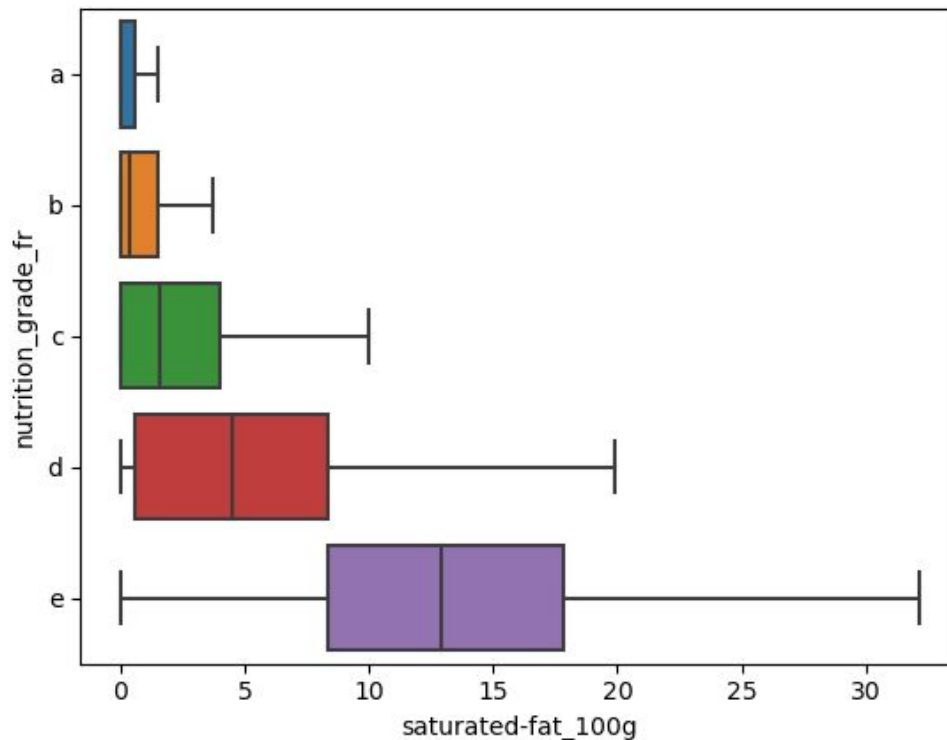


3. Analyses univariée et bivariées

3.2. Analyse bivariées : corrélations et distributions

Distributions en fonction du **nutrigrade** associé.

Exemple des **acides gras saturés** :



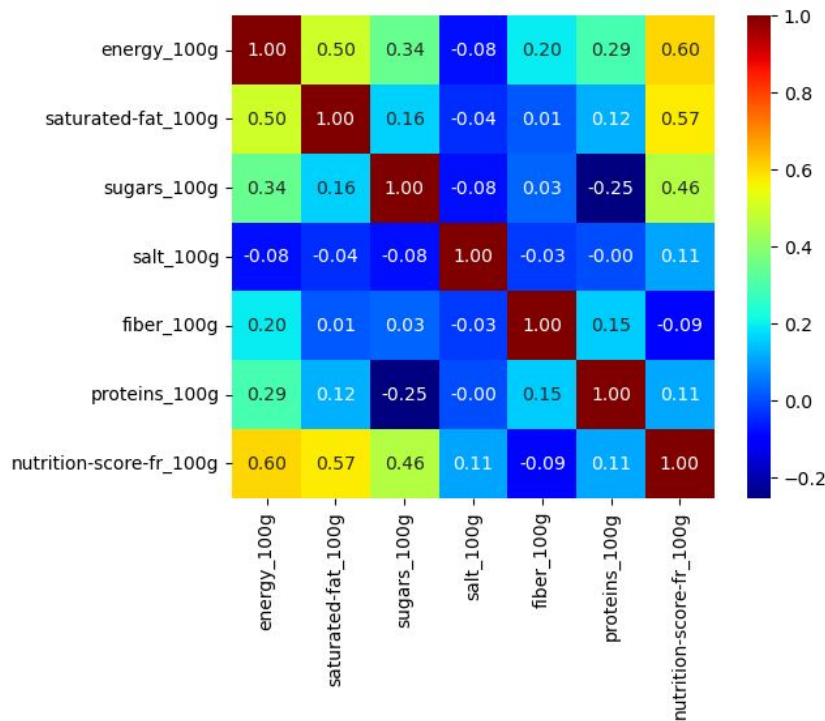
3. Analyses univariée et bivariées

3.2. Analyse bivariées : corrélations et distributions

Existe-t-il des **corrélations**
linéaires entre nos variables ?

Visualisation par une **heatmap**
des corrélations :

Heatmap de corrélation linéaire (Pearson) des variables de notre Data Frame



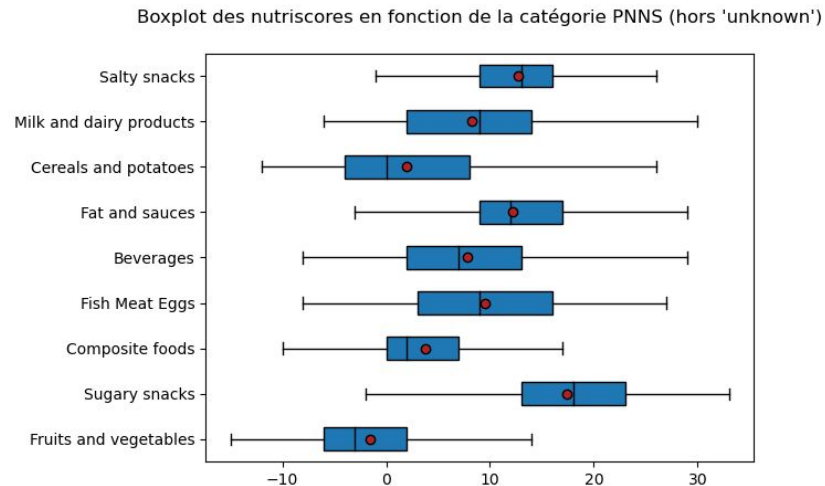
3. Analyses univariée et bivariées

3.3. Anova et test du Chi-2

Objectif de l'**Anova** : tester la présence d'une **dépendance** entre la **catégorie PNNS** 1 d'un produit et son **nutriscore**.

Existe-t-il ou non des **différences significatives** entre les **moyennes** de nutriscore pour chaque catégorie ?

Représentation graphique et résultats de test :



Nous avons obtenu un résultat de test élevé : 5216.49 pour une p-value = 0 ce qui nous **confirme** une **dépendance** entre ces deux variables.

3. Analyses univariée et bivariées

3.3. Anova et test du Chi-2

Objectif du **test de Chi-2** : voir s'il y existe ou non une **association significative** entre nos deux variables qualitatives : la **catégorie PNNS 1** d'un produit, et son **nutrigrade**.

Pour cela on peut d'abord établir un **tableau de contingence**, puis utiliser un test.

nutrition_grade_fr	a	b	c	d	e	Total
pnns_groups_1						
Beverages	374	739	1736	1244	2344	6437
Cereals and potatoes	3462	1104	1843	884	128	7421
Composite foods	1323	1848	1827	924	90	6012
Fat and sauces	96	235	1200	1785	767	4083
Fish Meat Eggs	545	754	1817	1661	1077	5854
Fruits and vegetables	2965	768	882	100	15	4730
Milk and dairy products	641	1384	2199	2950	286	7460
Salty snacks	56	82	753	1201	369	2461
Sugary snacks	73	339	1265	3834	5072	10583
Total	9535	7253	13522	14583	10148	55041

Nous avons obtenu un résultat de test élevé : 33798.2 pour une p-value = 0 ce qui nous **confirme** une **association entre ces deux variables**.

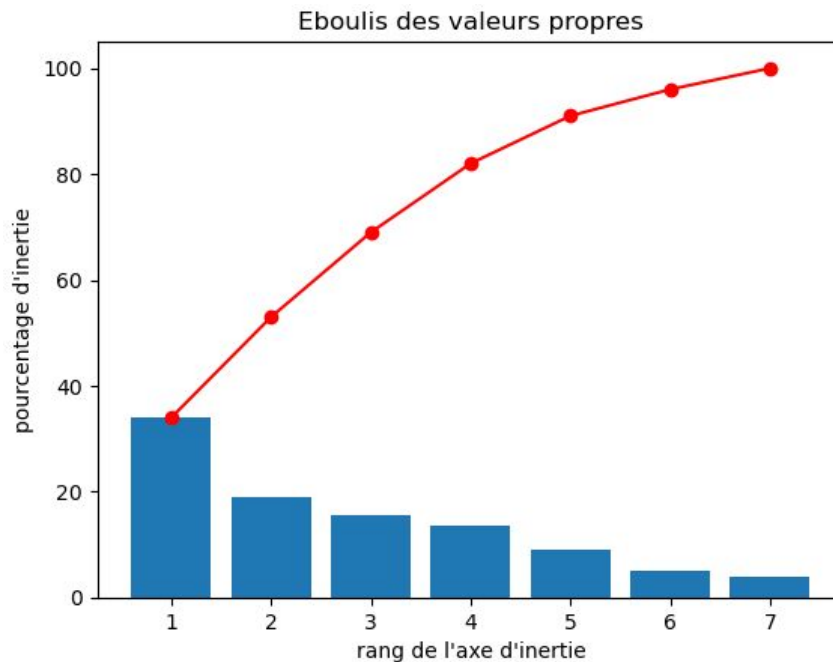
4. Analyse multivariées et suggestion d'application

4.1. Analyse en composantes principales

Objectif de l'ACP ici :

- identifier les **variables corrélées** entre elles et leur impact sur le nutriscore.
- Identifier les **variables importantes**
- Identifier certains **motifs**

On voit que les 4 premières composantes capturent environ 80% de la variation contenue dans les données.



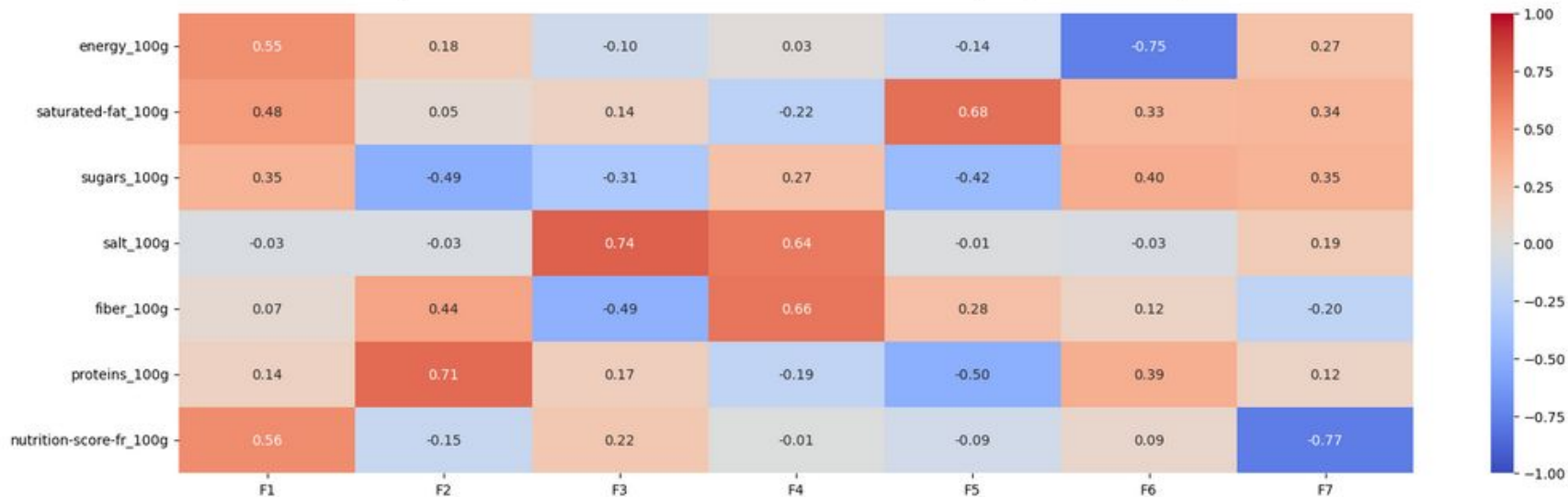


4. Analyse multivariées et suggestion d'application

4.1. Analyse en composantes principales

Regardons d'abord la **corrélation** entre nos **variables** et les différentes **composantes** :

Heatmap et matrice de corrélation de nos variables numériques, avec les composantes

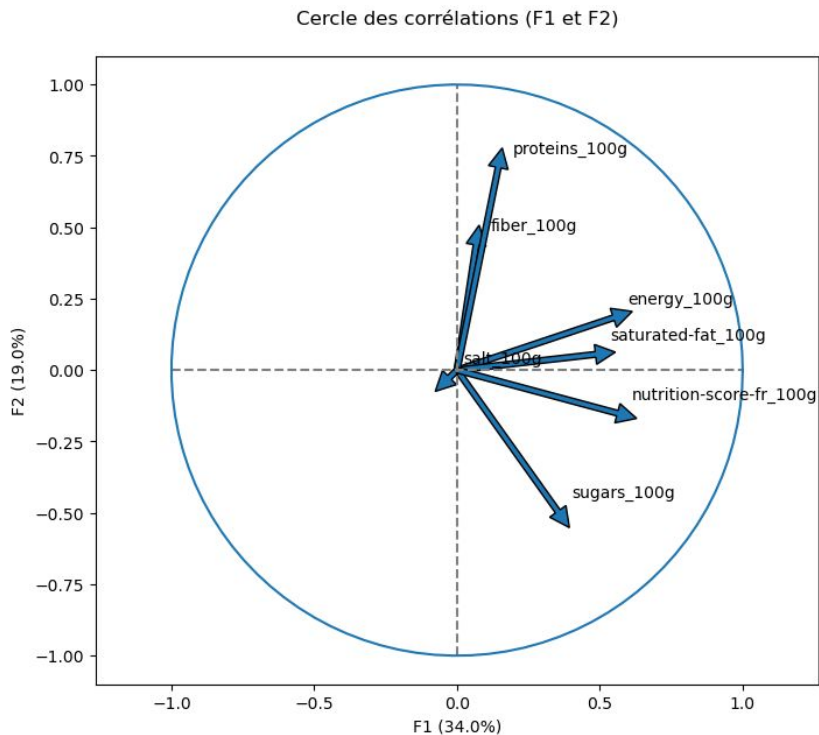




4. Analyse multivariées et suggestion d'application

4.1. Analyse en composantes principales

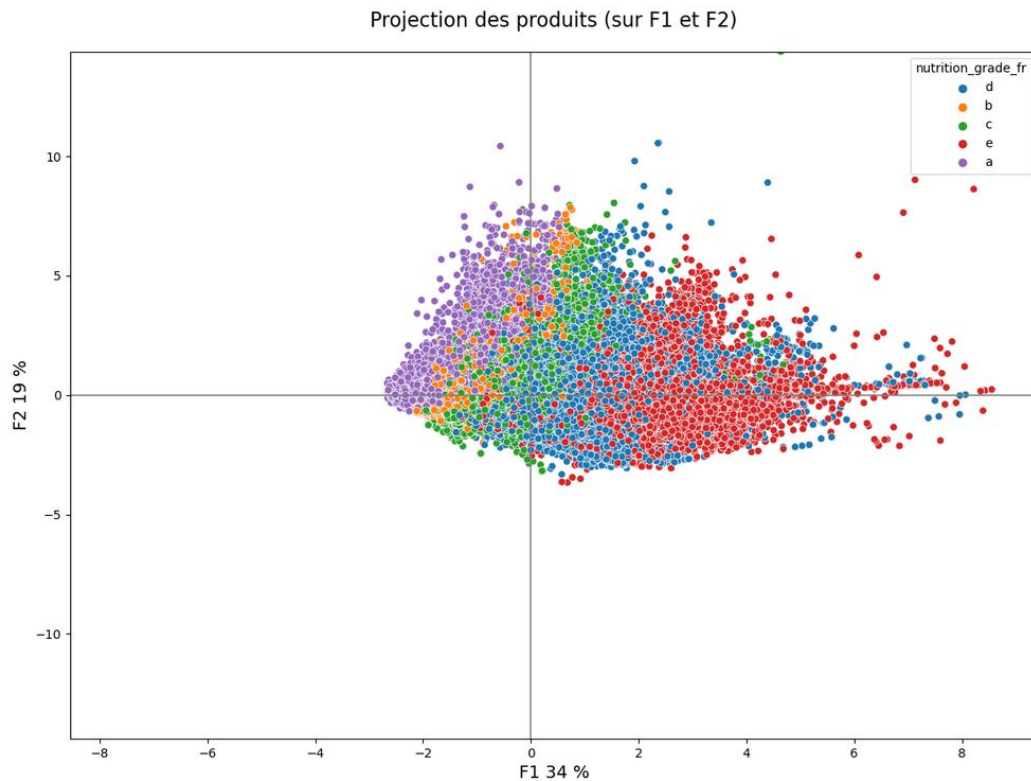
Voyons maintenant cela graphiquement :



4. Analyse multivariées et suggestion d'application

4.1. Analyse en composantes principales

Et si nous regardons la
projection des produits
dans cet espace :





4. Analyse multivariées et suggestion d'application

4.2. Faisabilité et suggestion d'une application

L'**exploration** de nos données nous a permis de mieux comprendre leur **comportement**, leurs **corrélations** ou non, et les liens existants entre les caractéristiques nutritives et le nutriscore (et nutrigrade) lui-même.

Les différents **tests statistiques** ont mis en évidence certaines liaisons significatives. Il semble possible de réaliser certaines estimations d'indicateurs en fonction de la valeur de certaines variables.

Si, lors de la saisie d'informations sur un produit par un utilisateur, certaines données et caractéristiques manquent, il semble possible de déterminer un algorithme ou un calcul qui permettrait de les remplir automatiquement.



Conclusion et conformité RGPD

Nous trouvons sur le site de la CNIL les 6 grands principes du **RGPD**, à savoir :

- 1) Ne collectez que les données vraiment nécessaires pour atteindre votre objectif
- 2) Soyez transparent
- 3) Organisez et facilitez l'exercice des droits des personnes
- 4) Fixez des durées de conservation
- 5) Sécurisez les données et identifiez les risques
- 6) Inscrivez la mise en conformité dans une démarche continue

Par ailleurs la CNIL définit la donnée personnelle comme : « toute information se rapportant à une personne physique identifiée ou identifiable ».

Notre travail ne concerne pas ce type de données.



Conclusion et conformité RGPD

Ce travail a permis d'établir un jeu de données nettoyées et sur lesquelles différentes analyses peuvent être réalisées.

L'exploration des données nous a montré leur comportement et certains liens existants entre elles. Il nous semble possible, après la réalisation de plusieurs autres tests et la comparaison de leurs performances et robustesses, de réfléchir à la création d'une application d'autocomplétion, facilitant le remplissage de la base par les utilisateurs.

Cela nécessiterait de retravailler ce jeu de données et probablement d'aller plus loin dans l'analyse de ces dernières.



Librairies utilisées

Python - Version 3.9.18

Pandas - Version 2.1.1

Numpy - Version 1.26.1

Matplotlib - Version 3.8.0

Seaborn - Version 0.13.0

Missingno - Version 0.5.2

Sklearn - Version 1.3.2

Statsmodels - Version 0.14.0

Scipy - Version 1.11.3



Ressources

Provenance des données : <https://world.openfoodfacts.org/>

Informations sur le nutriscore :

<https://www.santepubliquefrance.fr/determinants-de-sante/nutrition-et-activite-physique/articles/nutri-score>

Informations sur la composition de différents aliments : <https://ciqual.anses.fr/>

Informations sur le RGPD :

<https://www.cnil.fr/fr/comprendre-le-rgpd/les-six-grands-principes-du-rgpd>

<https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on>



Annexe

Régression linéaire en vue d'une détermination du nutriscore

Régression linéaire avec :

- nutriscore en variable expliquée
- apports nutritifs et énergie en variables explicatives

Peut-on déterminer correctement la valeur du nutriscore à partir de nos variables explicatives ?

Résultats significatifs mais coefficients très faibles

R-squared:	0.598
Adj. R-squared:	0.598

	coef	std err	t	P> t	[0.025	0.975]
const	0.8266	0.024	34.644	0.000	0.780	0.873
energy_100g	0.0038	2.05e-05	184.938	0.000	0.004	0.004
saturated-fat_100g	0.3892	0.002	216.629	0.000	0.386	0.393
sugars_100g	0.1350	0.001	199.820	0.000	0.134	0.136
salt_100g	0.2379	0.002	120.703	0.000	0.234	0.242
fiber_100g	-0.3428	0.003	-123.937	0.000	-0.348	-0.337
proteins_100g	0.0836	0.002	48.622	0.000	0.080	0.087