

Concevez une application au service de la santé publique

Soutenance de projet

Nicolas FAUCONNIER
Parcours Ingénieur ML
29/11/2021



Plan

1.

Application

Contexte et idée

2.

Data Cleaning

Nettoyage et sélection des données

3.

Analyse exploratoire

Analyse univariée et multivariée des variables

4.

Conclusions

Faisabilité et possibles améliorations

1. Application



Contexte

Réponse à un appel à projets d'application innovante en lien avec l'alimentation, émis par **Santé Publique France**



Idée


Face à la multitude d'options disponibles et d'informations, aider les consommateurs dans le choix de boissons et proposer des alternatives plus saines et/ou respectant des critères choisis par l'utilisateur



2.

Data Cleaning

Nettoyage et sélection des données pertinentes pour
l'application



Dataset

Export de la base **OpenFoodFacts** à date du 24/10/2021

Disponible au lien suivant:

<https://world.openfoodfacts.org/data>



- Format csv, **4,24 Go**
- **1998448 lignes** et **187 colonnes**
- Dataset **très épars**: beaucoup de NaNs

Outils utilisés:



Données produit

- Informations générales: dates, noms, et différentes url
 - Données plus spécifique: taille, additifs, huile de palme, etc.
 - Tags: marque, catégories de produit, données géographiques, etc,
 - Liste des ingrédients et traces
 - Données nutritionnelles (g/100g)
 - Scores et grades
-

Filtre par pays et produit

On se concentre sur la **France**:

Différentes dénominations dans la colonne: "France" , "Frankreich", etc.

On utilise un dictionnaire de 19 traductions

⇒ *847601 lignes restantes (42% du dataset original)*

On se concentre sur les **boissons**:

⇒ *40114 lignes restantes (2% du dataset original)*

Suppression de colonnes et de lignes

Colonnes vides:

eg. Différents types d'acides, de sucres, de fibres, etc.

On passe de 187 colonnes à 139

Lignes **sans informations nutritionnelles** renseignées:

⇒ *34330 lignes restantes (1,71% du dataset original)*

Suppression de lignes

Observations sans nom de produit:

⇒ *34160 lignes restantes (1,70% du dataset original)*

Doublons et quasi-doublons:

- Pas de lignes entièrement identiques
- 2 doublons sur le code produit
- **8649** sur le nom du produit, mais **beaucoup sont en réalité des produits différents** (marque, packaging et quantités différentes). On supprime **1553 quasi-doublons**

⇒ *32605 lignes restantes (1,6% du dataset original)*

Suppression de colonnes

Colonnes jugées inutiles pour la suite du projet, ou hors descriptif du dataset:
Urls, images, variables temporelles, certaines dénominations produit, quelques variables en _100g, etc.

Colonnes avec **trop d'observations manquantes**:
eg. carbon-footprint_100g (nb. pas un composant alimentaire),

Colonnes **redondantes**:
Données " tags" (_tags, _en)

⇒ *101 colonnes restantes sur les 187 du dataset original*

nb. ce processus de sélection des variables se répètera au cours de la phase de nettoyage des données

Valeurs aberrantes

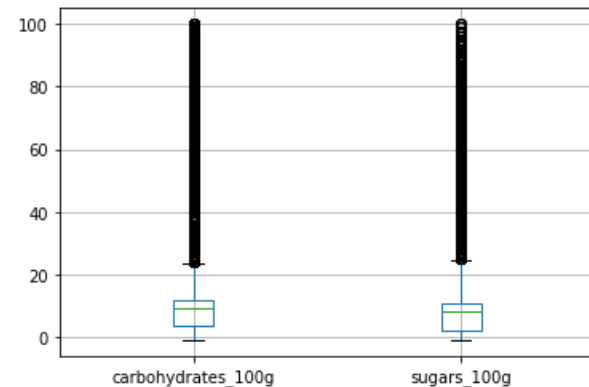
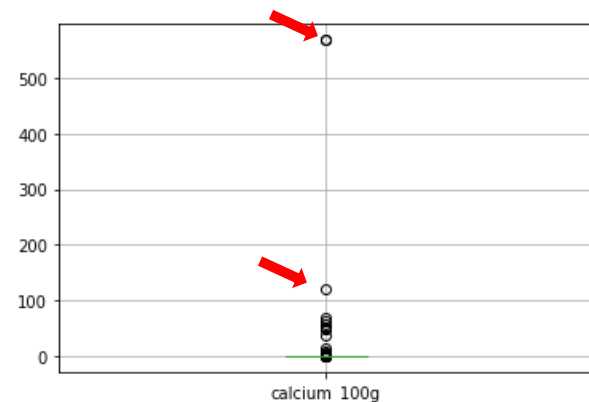
Suppression des valeurs aberrantes:

- **supérieures à 100** et **inférieures à 0 g/100g**
- **hors des bornes** de l'écoscore (0 à 100)

⇒ *32566 lignes restantes (1,62% du dataset original)*

Cas des valeurs atypiques:

Sans connaissance métier sur le sujet, je décide de ne pas les supprimer



Nouvelle suppression de colonnes

Focus sur les colonnes difficilement exploitables:

Les données tags sont **difficilement exploitables**: beaucoup de valeurs uniques, beaucoup de valeurs similaires

exemple: - "categories_en " possède 2390 valeurs uniques
- "labels " en possède 2015

⇒ Il reste 89 colonnes sur les 187 du dataset d'origine

Création de " labels "

Utilisation de la variable tag "labels" :

Recherche de mots/expressions dans la colonne pour créer des *Dummy variables*:

- organic: `"bio|organic"`

- fairtrade: `"fairtrade|fair trade|equitable|équitable"`

- MadeinFrance:

`"made in france|madeinefrance|fabriqué en france|fabrique en france|fabriquéenfrance|fabriqueenfrance"`

Imputations

Valeur énergétique pour 100g:

Fill de la colonne *energy-kcal_100g* avec les colonnes *energy-kj_100g* et *energy_100g*.

Conversion: 1 (kcal) = 4,1868 (kJ)

Fill par la médiane du groupe PNN du reste des valeurs manquantes: 7541

nb. cette méthode fournit une erreur moyenne de 7 kcal sur les lignes sans valeurs manquantes

Imputations

Teneur en fruits pour 100g:

On fill la colonne *fruits-vegetables-nuts_100g* avec *fruits-vegetables-nuts-estimate-from-ingredients_100g* en cas de données manquantes

Marque:

On fill la colonne *brands* avec *brand_owner* en cas de données manquantes

Ecograde et nova_group:

En cas de NaN; fill par une catégorie *inconnu*

Fill par 0:

Toutes les colonnes nutritives en *_100g*, ainsi que quelques autres (eg. nombre d'additifs et ingrédients issu d'huile de palme)

Imputations

Nutriscore et Nutrigrade:

Une méthode de calcul spécifique aux boissons; des valeurs en point différentes des autres aliments

Points	Energy density (kJ/100g or 100mL)	Sugars (g/100g or 100mL)	Saturated fatty acids (g/100g)	Sodium (mg/100g)
0	≤0	≤0	≤1	≤90
1	≤30	≤1.5	>1	>90
2	≤60	≤3	>2	>180
3	≤90	≤4.5	>3	>270
4	≤120	≤6	>4	>360
5	≤150	≤7.5	>5	>450
6	≤180	≤9	>6	>540
7	≤210	≤10.5	>7	>630
8	≤240	≤12	>8	>720
9	≤270	≤13.5	>9	>810
10	>270	>13.5	>10	>900

Points	Fruits, vegetables, pulses, nuts, and rapeseed, walnut	Fibre (g/100g)	Protein (g/100g)
0	≤40	≤0.9	≤1.6
1		>0.9	>1.6
2	>40	>1.9	>3.2
3		>2.8	>4.8
4	>60	>3.7	>6.4
5		>4.7	>8.0
6			
7			
8			
9			
10	>80		

Attribution des couleurs

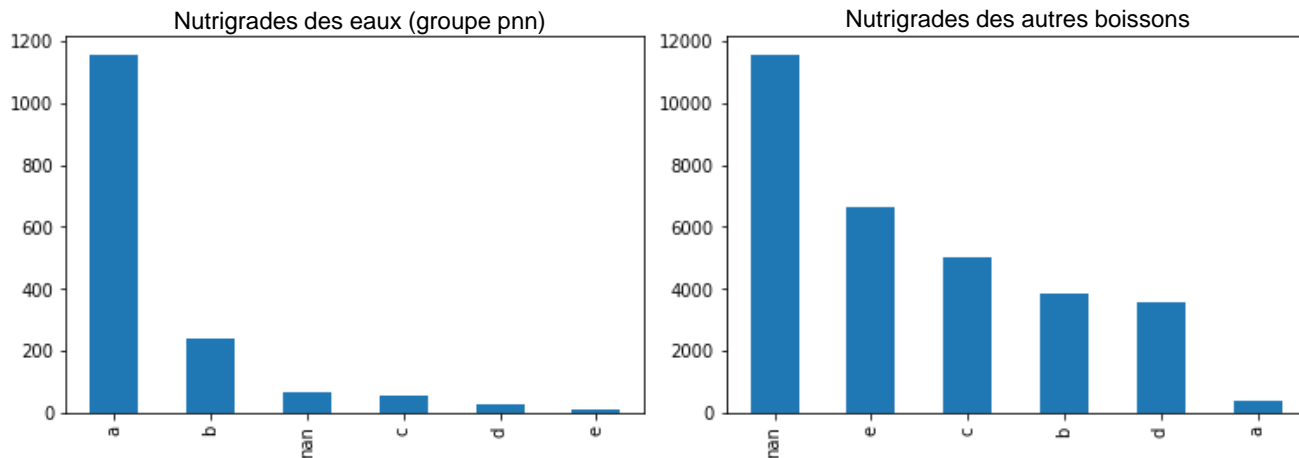
Attribution des couleurs	
Aliments solides (points)	Boissons (points)
Min à -1	Eau
0 à 2	≤ 1
3 à 10	2 à 5
11 à 18	6 à 9
≥ 19	≥ 10



Imputations

Nutriscore et Nutrigrade:

Spécificités: - ne s'applique pas à l'alcool: création du grade *Alcool*
- seul les eaux **sans ajouts** peuvent obtenir le grade A



nb. 72% de grades correctes prédits sur les données présentes avec cette méthode de recalcul

Sélection finale de colonnes

- Drop de colonnes utilisées pour en fill d'autres: date de création, ingrédients venant d'huile de palme
- Colonnes jugées peu exploitables ou utiles dans l'analyse

	étape	n colonnes	n lignes	% lignes del	% lignes restantes
0	import	187	1998448	0.000000	1.000000
1	filtre sur la France	187	847601	57.587038	42.412962
2	sélection des boissons	187	40114	95.267349	2.007258
3	suppression des colonnes vides	139	40114	0.000000	2.007258
4	suppression des lignes sans nutrition facts	139	34330	14.418906	1.717833
5	suppression des lignes sans nom de produit	139	34160	0.495194	1.709326
6	suppression des doublons	139	32605	4.552108	1.631516
7	suppression des colonnes inutiles ou redondantes	101	32605	0.000000	1.631516
8	suppression des valeurs aberrantes	101	32566	0.119614	1.629565
9	suppression de colonnes inexploitables ou inut...	89	32566	0.000000	1.629565
10	sélection finale de colonnes	77	32566	0.000000	1.629565



3.

Analyse exploratoire

Analyse univariée et multivariée



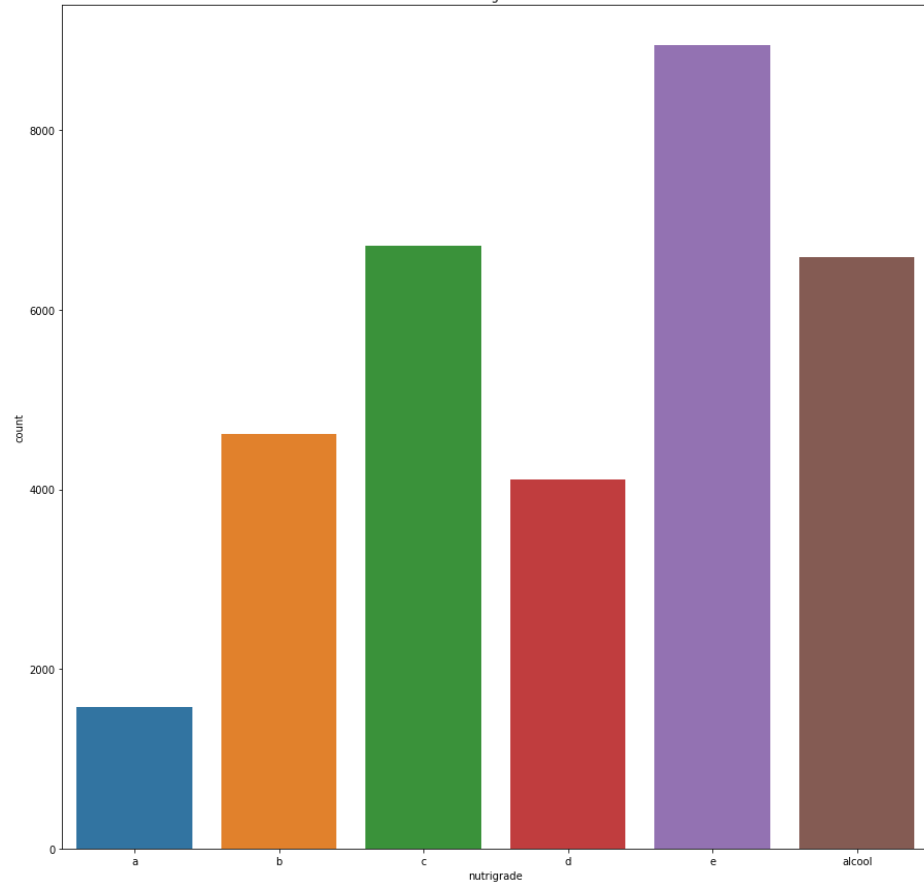
Attentes

Vérifier que les données nous permettent de discriminer des produits de catégories similaires sur des critères définis, notamment des critères liés à la santé

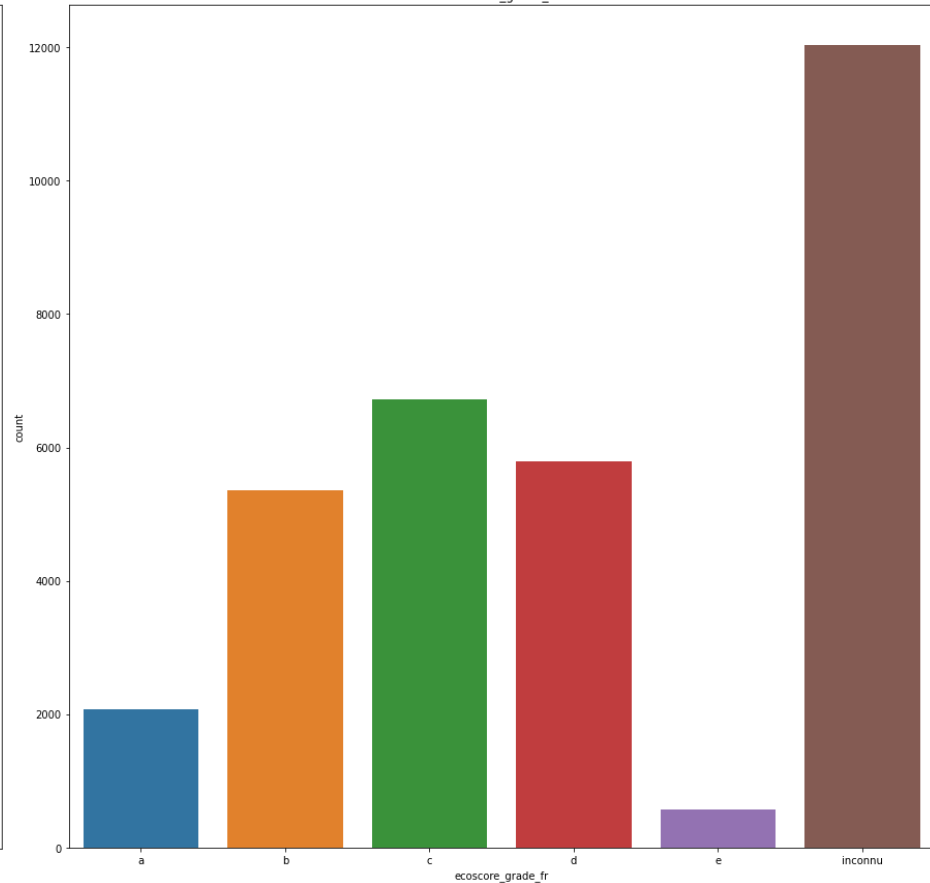
Déceler des relations entre variables potentiellement utiles

Grades

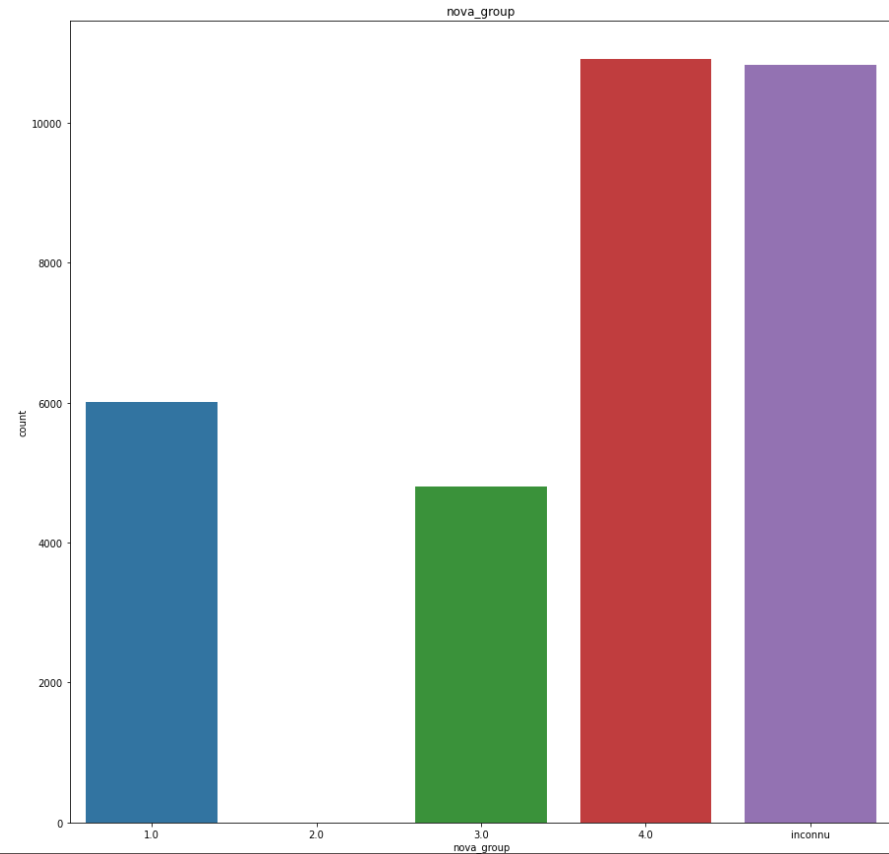
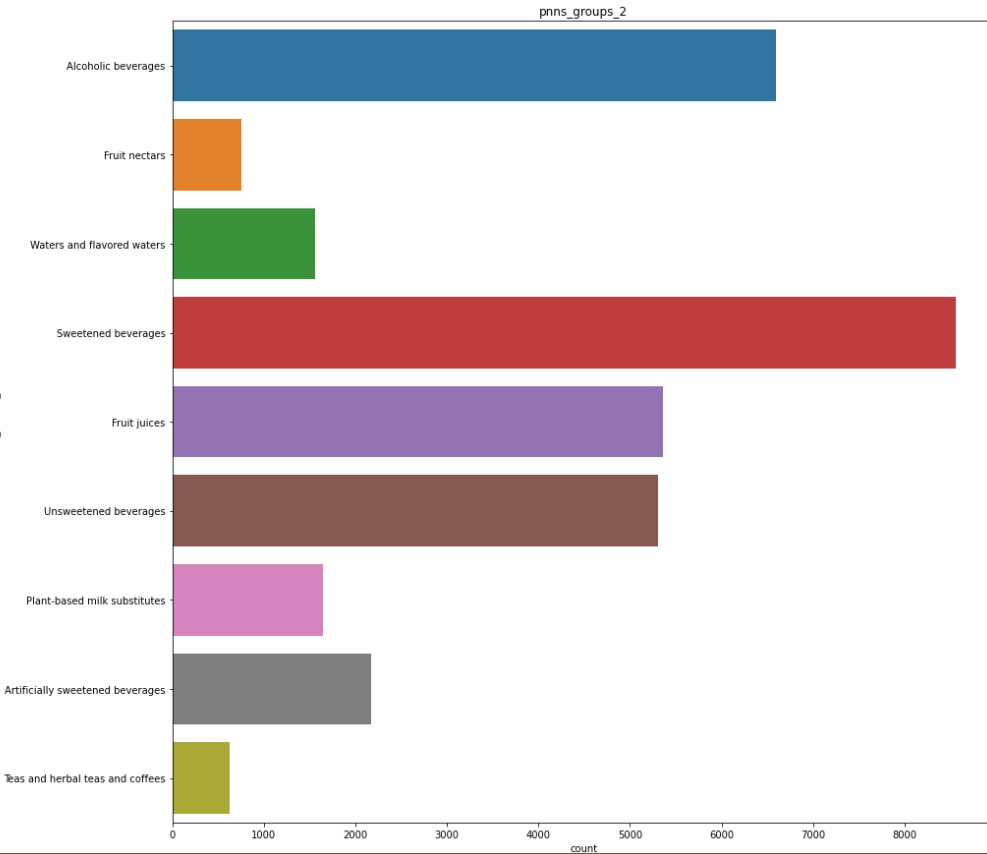
nutrigrade



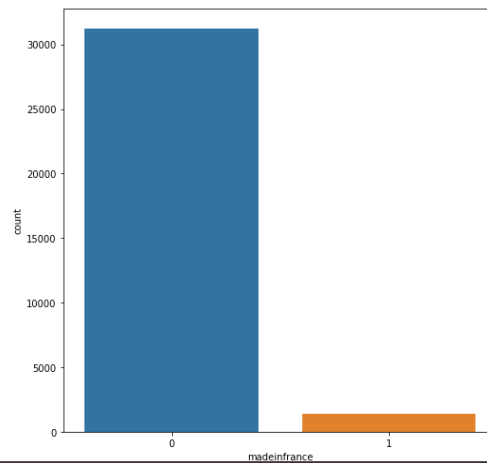
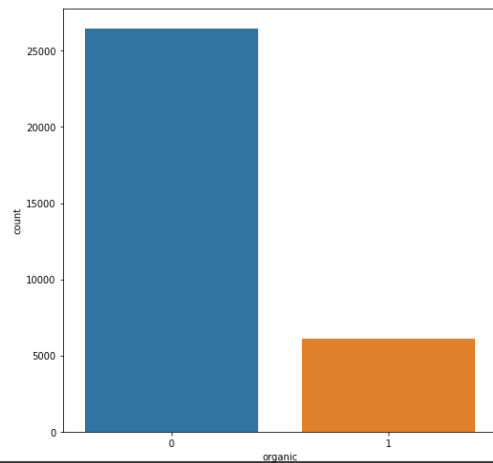
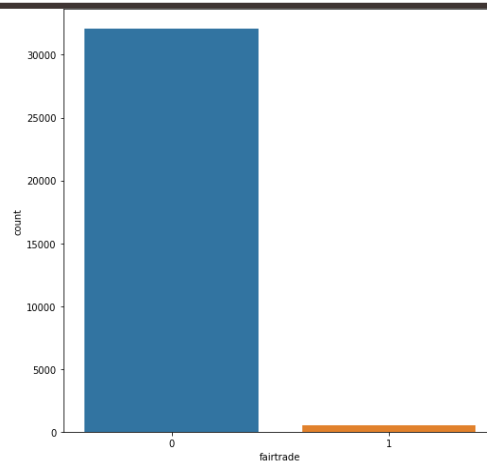
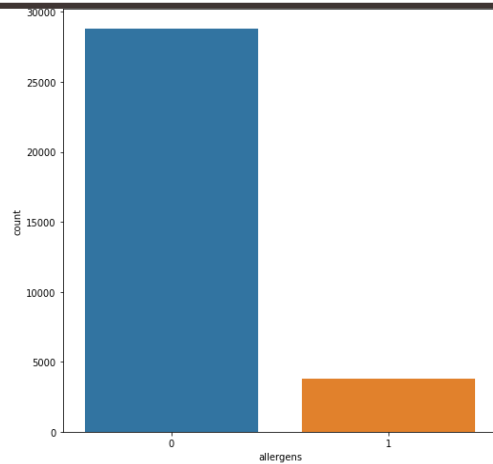
ecoscore_grade_fr



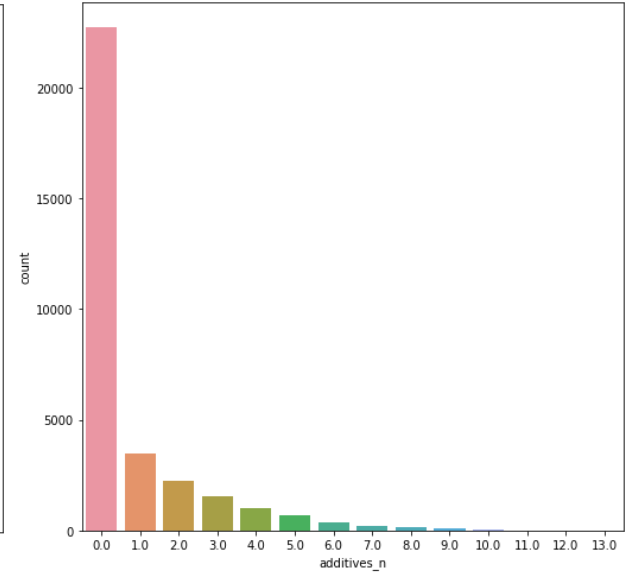
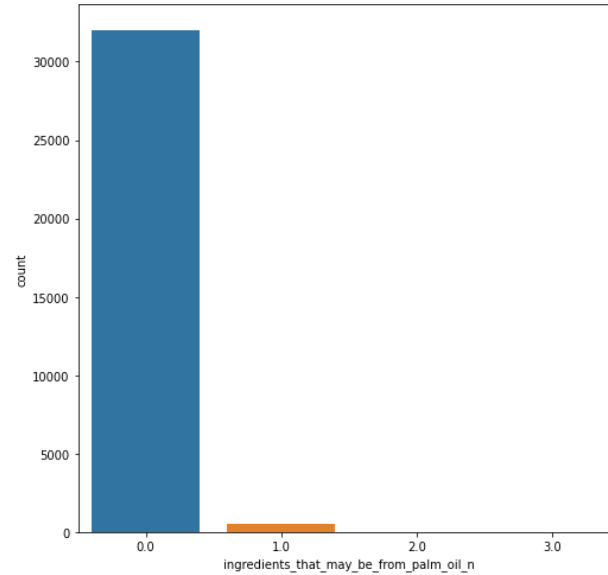
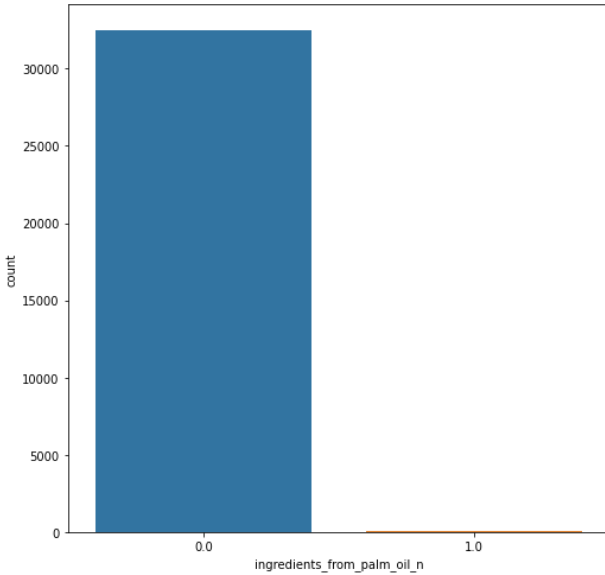
Groupes et catégories de produits



Labels



Variables quantitatives discrètes



Variables quantitatives continues

On se concentre sur un set restreint de variables nutritives:

	energy-kcal_100g	fat_100g	saturated-fat_100g	proteins_100g	carbohydrates_100g	alcohol_100g	sugars_100g	sodium_100g	fruits-vegetables-nuts_100g	fiber_100g
count	32566.000000	32566.000000	32566.000000	32566.000000	32566.000000	32566.000000	32566.000000	32566.000000	32566.000000	32566.000000
mean	58.924325	0.673905	0.342794	0.675460	10.239082	1.584972	9.084760	0.065486	12.929426	0.170985
std	83.780966	3.513410	2.206056	3.898502	18.501014	5.681424	17.240794	0.896575	29.076723	1.566103
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	22.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	45.000000	0.000000	0.000000	0.000000	6.000000	0.000000	4.600000	0.000000	0.000000	0.000000
75%	49.000000	0.140000	0.010000	0.500000	11.000000	0.000000	10.000000	0.008000	3.200000	0.000000
max	1489.000000	100.000000	60.000000	100.000000	100.000000	100.000000	100.000000	56.000000	100.000000	89.000000

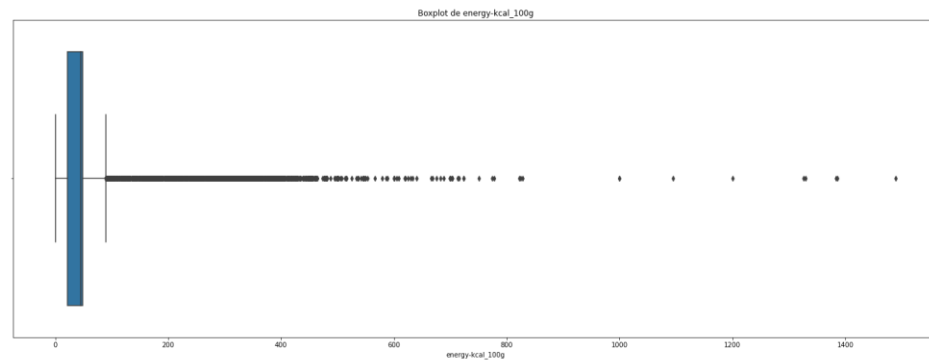
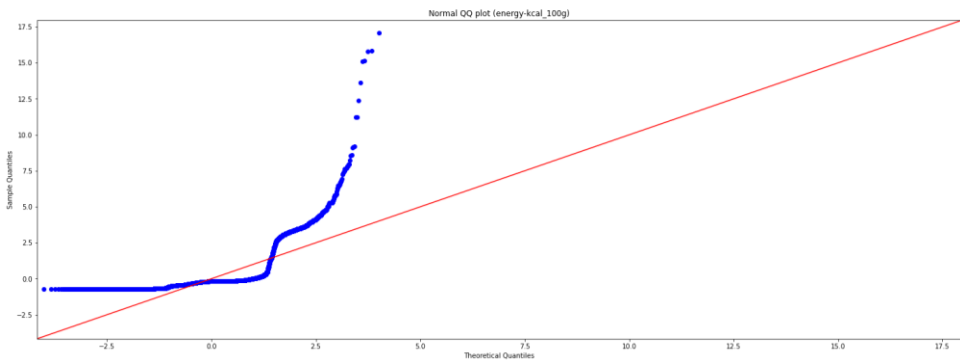
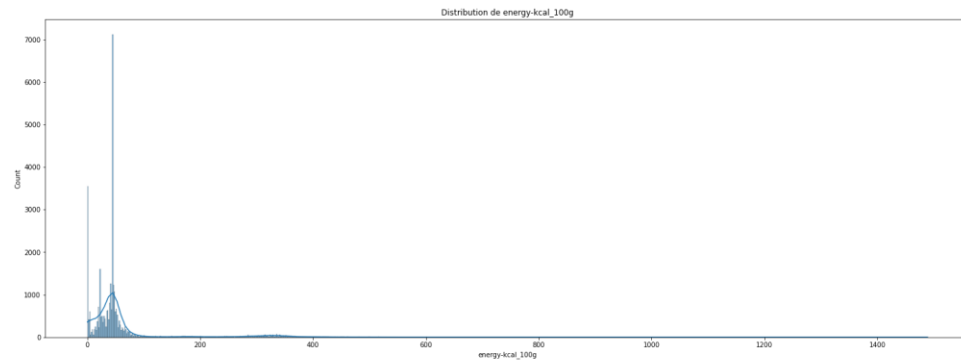
Variables quantitatives continues

Exemple d'un cas typique du set:

- Beaucoup de 0
- N'est pas normalement distribuée

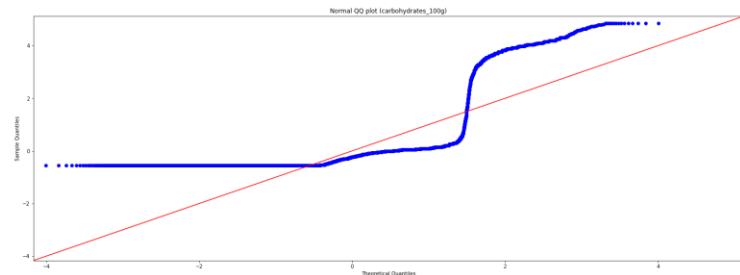
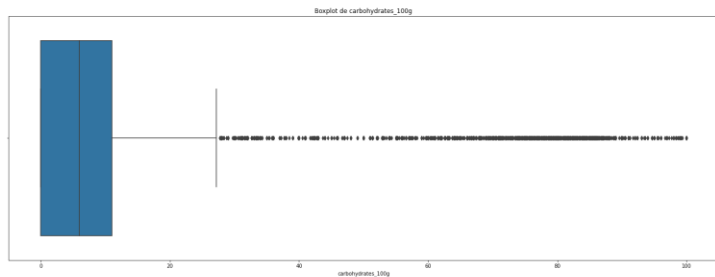
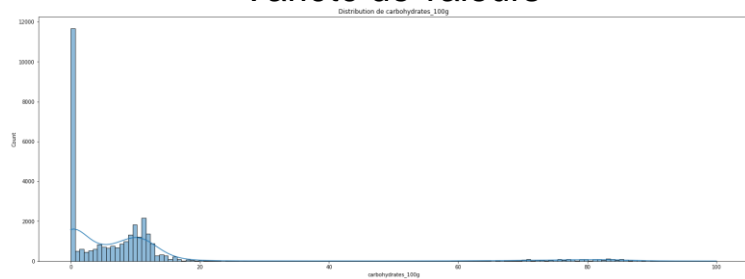
Test de Shapiro-Wilk :

	W	pval	normal
energy-kcal_100g	0.527433	0.0	False

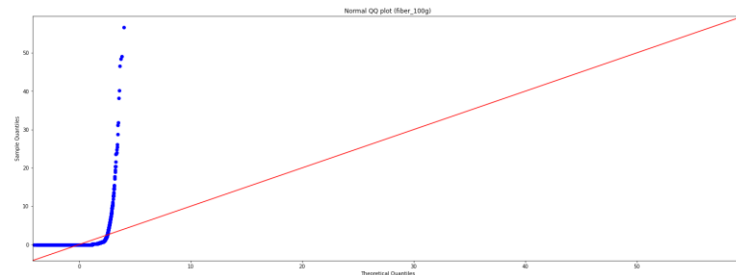
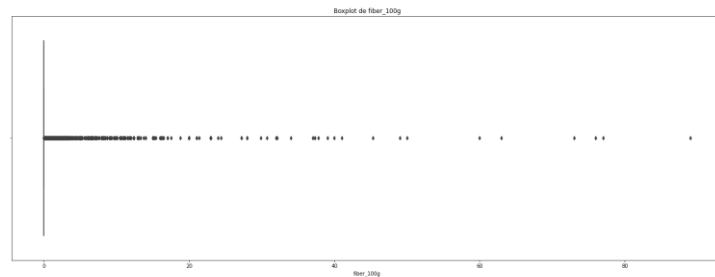
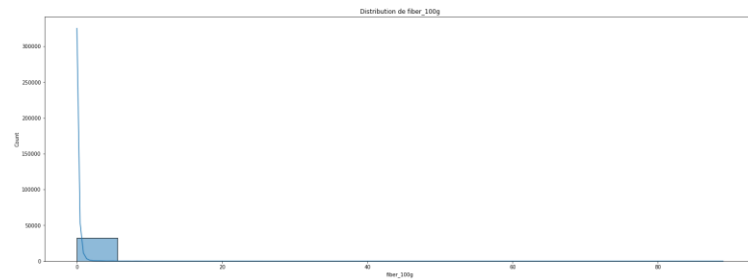


Variables quantitatives continues

Variété de valeurs

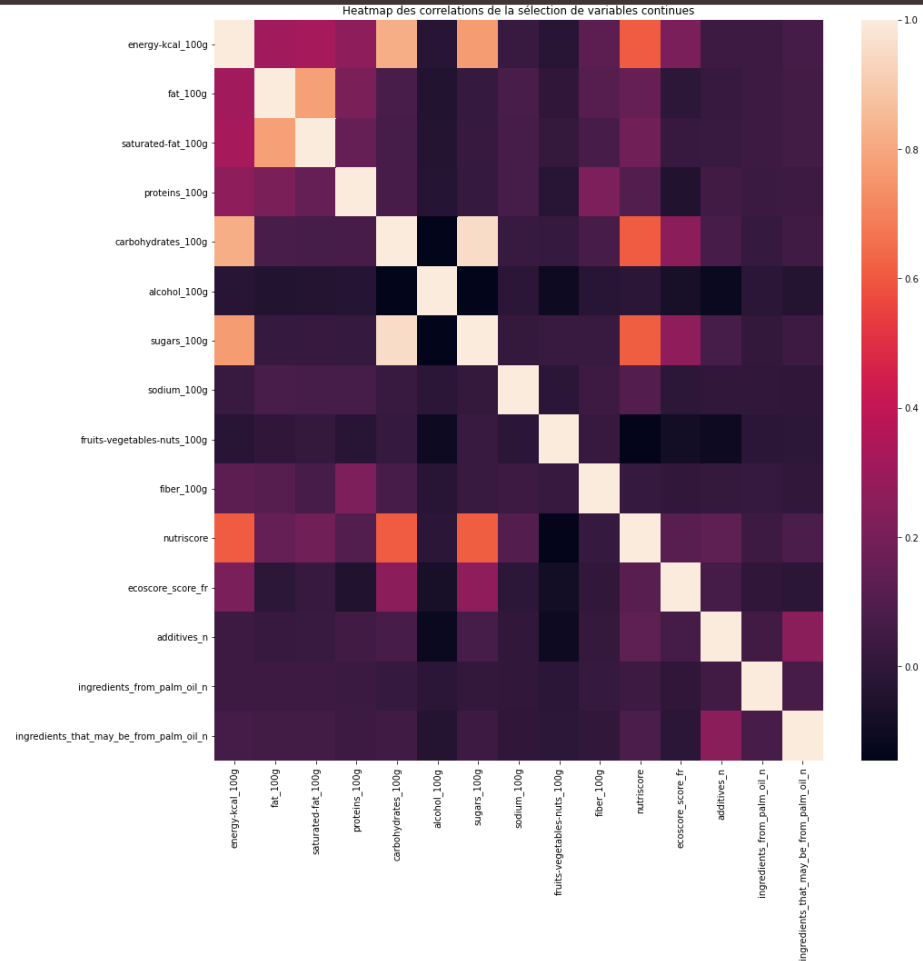


Quasi-totalité de valeurs faibles



Analyse bivariée

heatmap des corrélations:



ANOVAs

Réalisation d'une ANOVA pour chaque couple variable quantitative / variable catégorielle

Mais les résultats ne sont pas fiables, les hypothèses de bases ne sont pas respectées:

- ─ Aucune variable distribuée normalement: cf. tests de Shapiro-Wilk et QQ plots

	W	pval	normal
energy-kcal_100g	0.527433	0.0	False
fat_100g	0.175411	0.0	False
saturated-fat_100g	0.137812	0.0	False
proteins_100g	0.134004	0.0	False
carbohydrates_100g	0.523483	0.0	False
alcohol_100g	0.307187	0.0	False
sugars_100g	0.503339	0.0	False
sodium_100g	0.039961	0.0	False
fruits-vegetables-nuts_100g	0.495334	0.0	False
fiber_100g	0.071408	0.0	False
nutriscore	0.952514	0.0	False
ingredients_that_may_be_from_palm_oil_n	0.102241	0.0	False
additives_n	0.577292	0.0	False

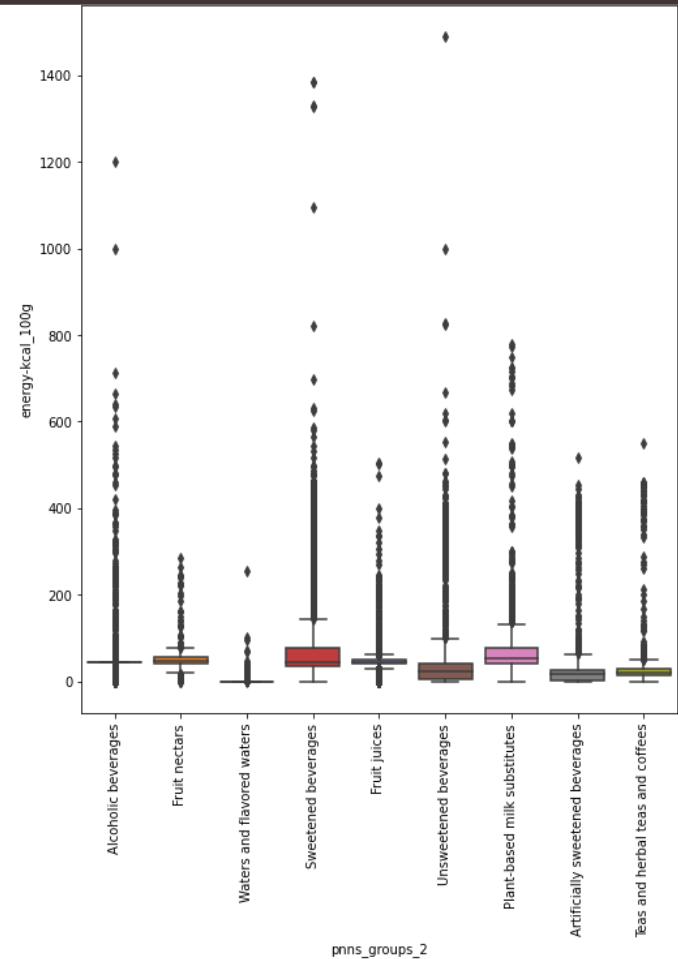
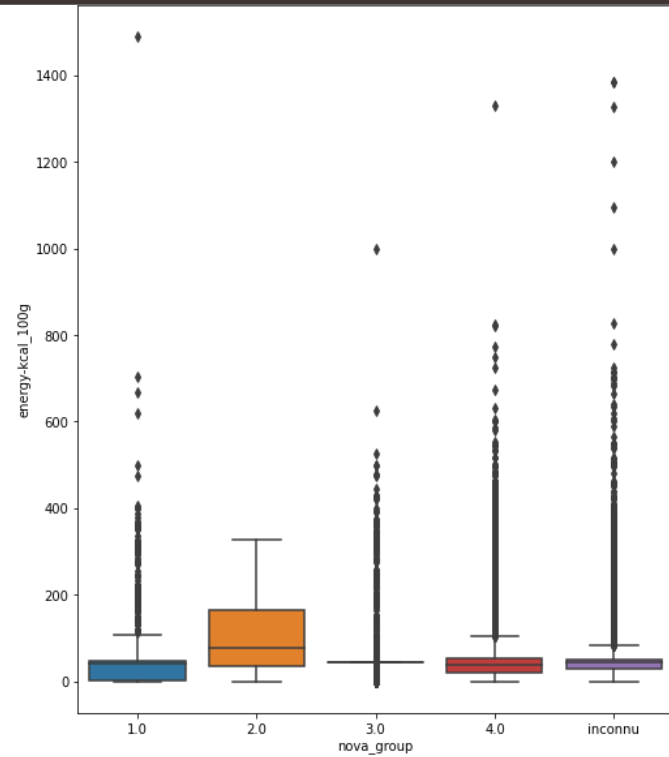
- ─ Hétéroscédasticité: test de Levene (moins sensible aux écarts de normalité)

	W	pval	equal_var
levene	6022.62773	0.0	False

- ─ Le nombre d'observations d'une catégorie à l'autre est quasi-toujours différent

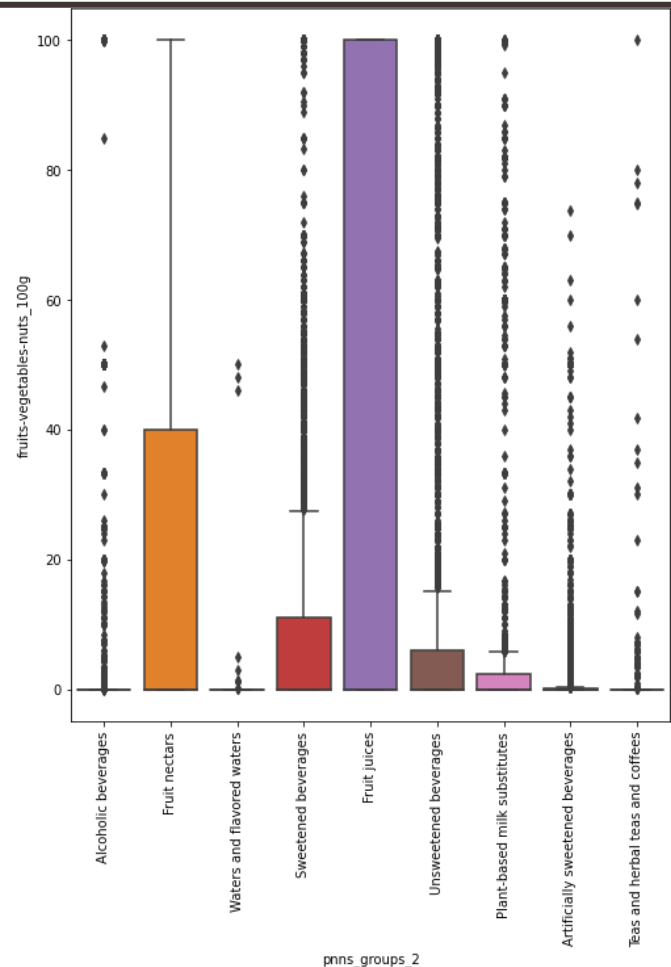
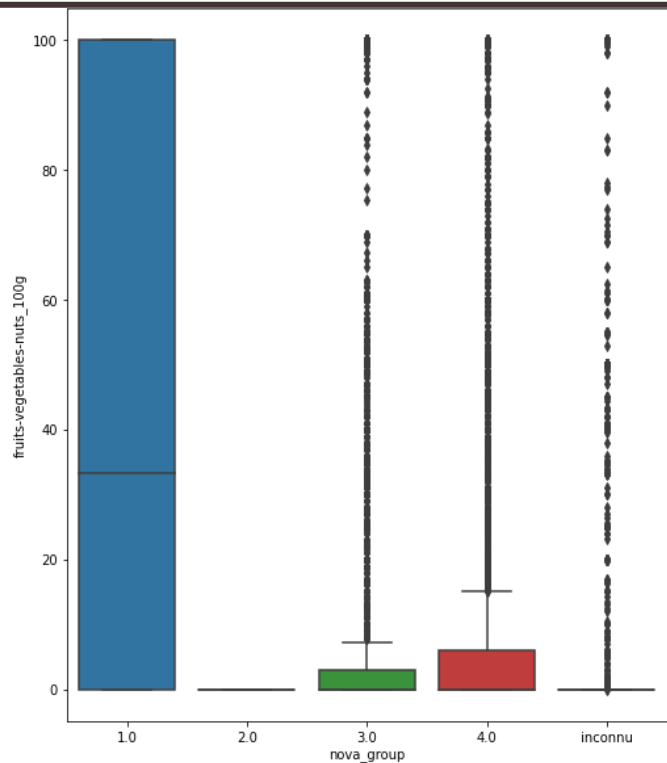
Analyse bivariable

Énergie et catégories de produits:



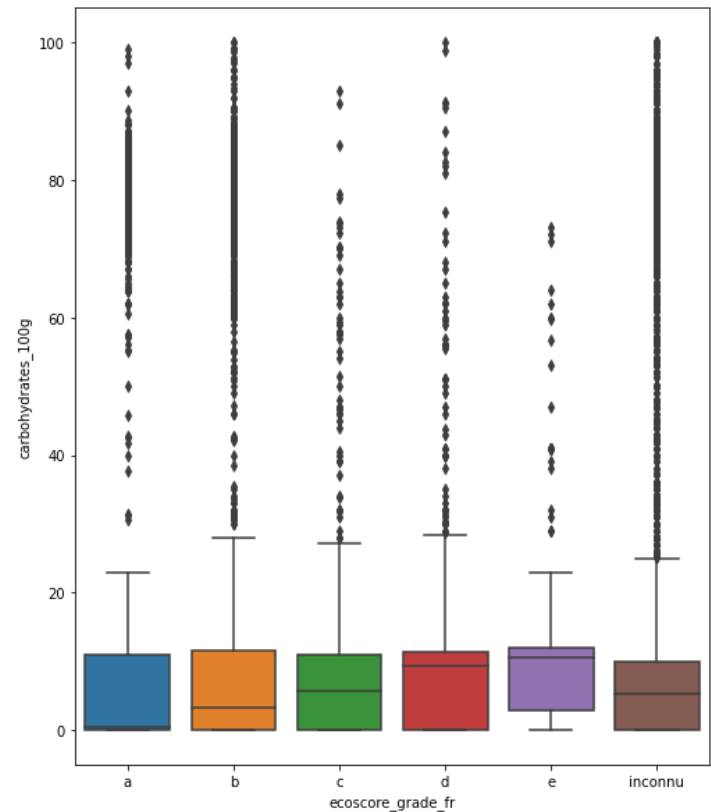
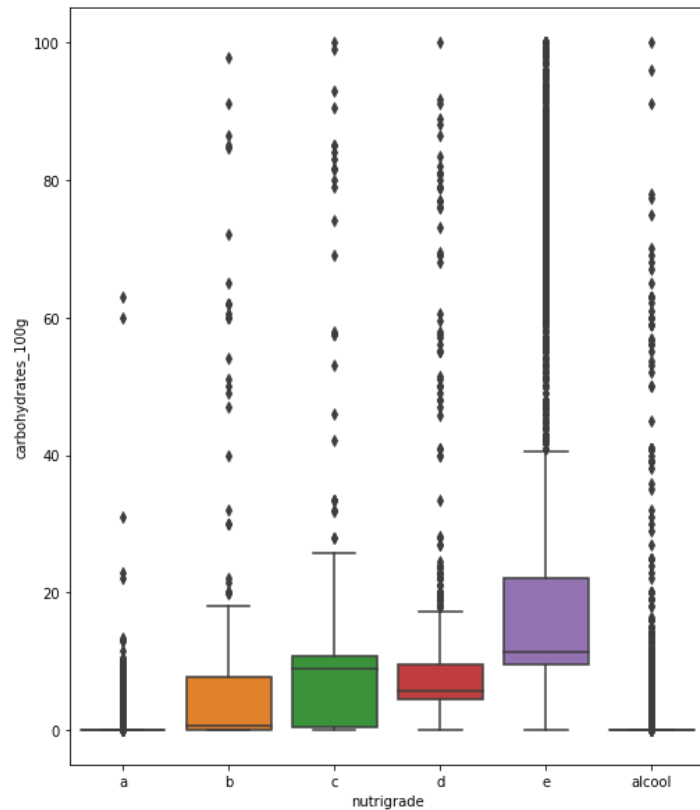
Analyse bivariable

Teneur en fruits/légumes et catégories de produits:



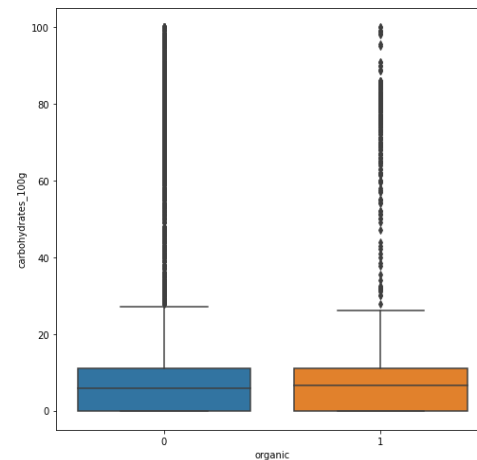
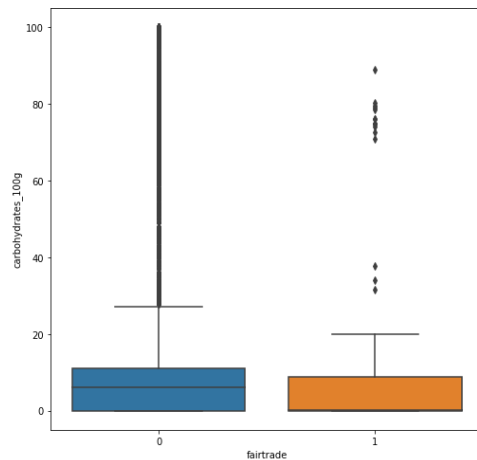
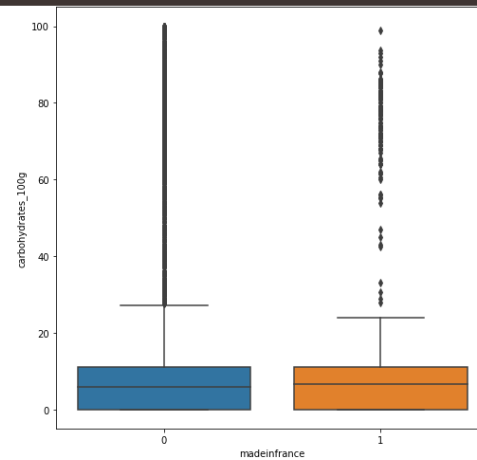
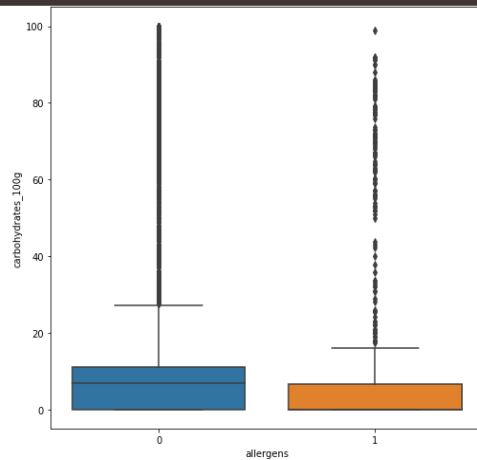
Analyse bvariée

Glucides et grades:



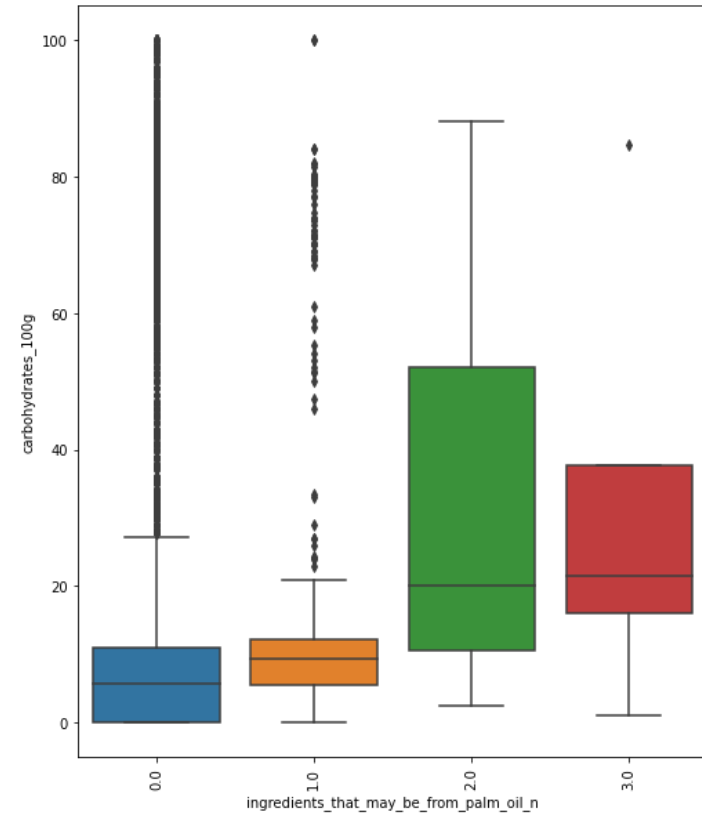
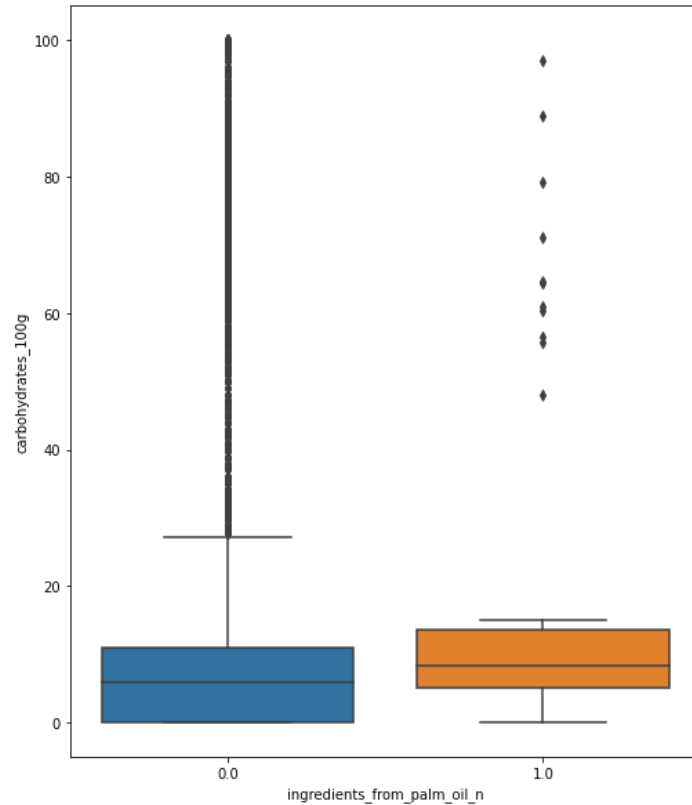
Analyse bivariable

Glucides et labels:



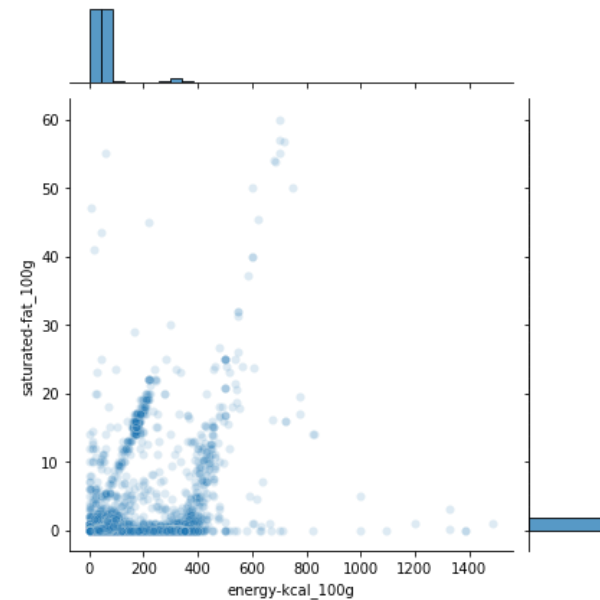
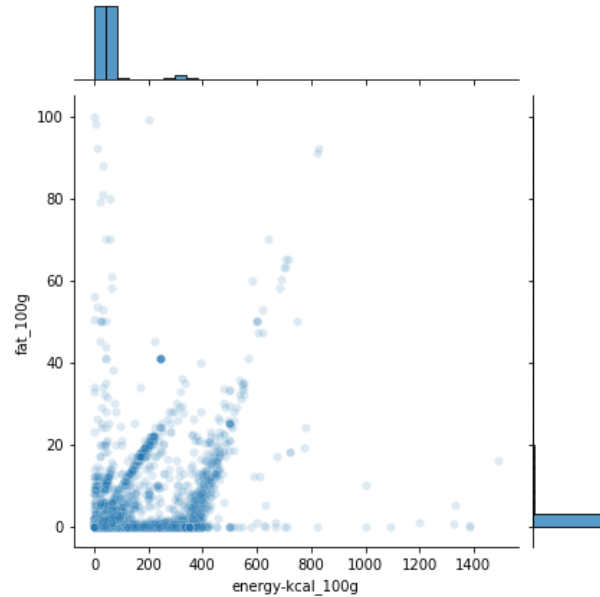
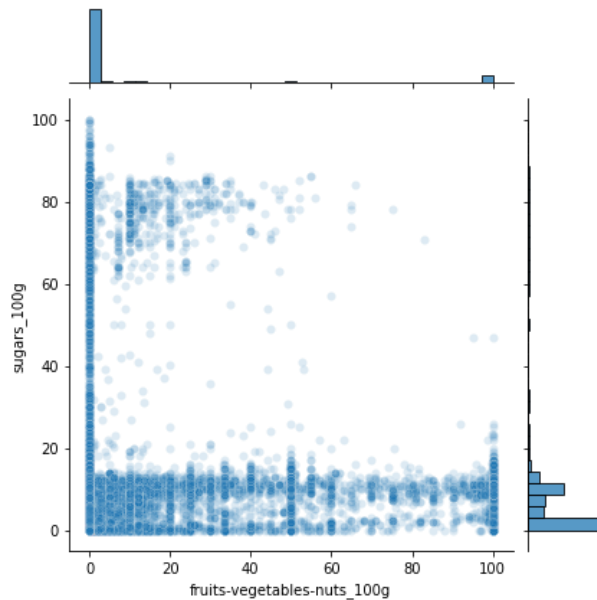
Analyse bivariable

Glucides et ingrédients
venant potentiellement
d'huile de palme:



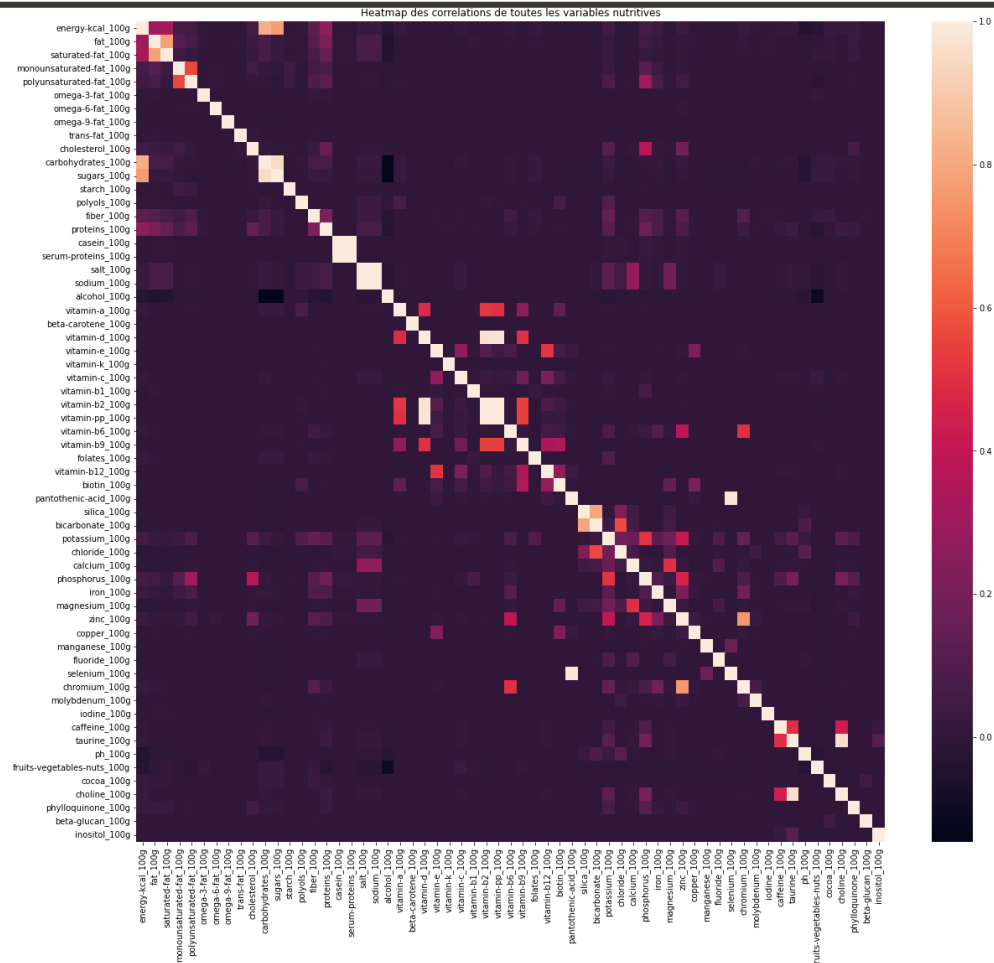
Analyse bivariée

Variables quantitatives:



Analyse bivariée

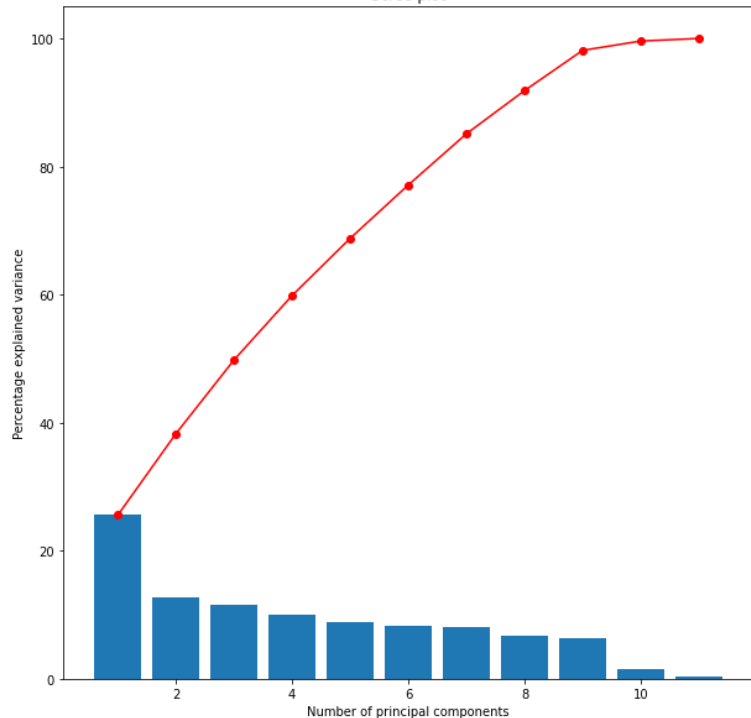
Valeurs nutritives éparses:



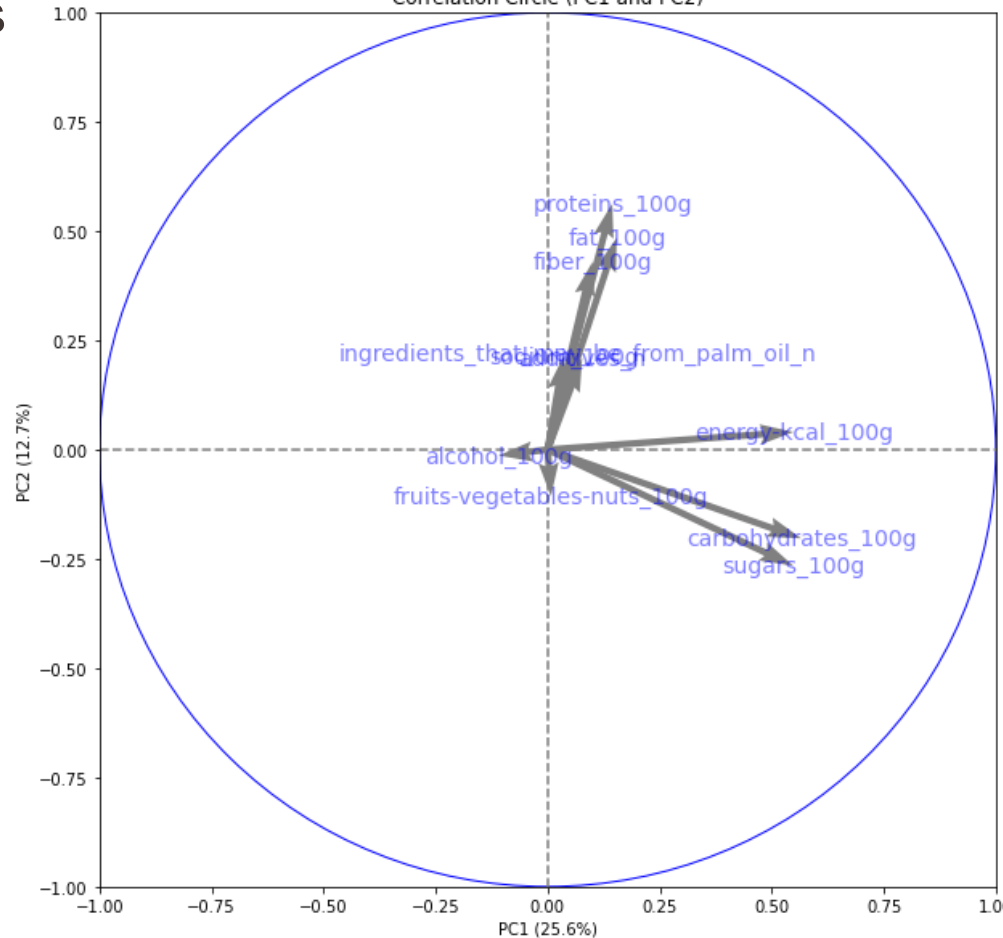
Analyse en Composantes Principales

8 composantes

Scree plot



Correlation Circle (PC1 and PC2)



4. Conclusions

- Possibilité de classer et différencier les produits de manière relative selon plusieurs variables
 - Problèmes de fiabilité et de qualité des données: c'est un point important à améliorer dans la suite du projet
 - Manque de granularité dans les catégories de produits
 - L'apport d'autres données serait bénéfique pour faire une application réellement fonctionnelle (eg. préférences des utilisateurs via leurs achats ou leurs avis)
-

Merci

Avez-vous des questions?
