

Projet 3 : Application au service de la santé publique

Goûters enfants en France

Etudiant : Fatma Aidi
Mentor : Kezhan Shi
Evaluateur : Late Lawson
Date : 12/02/2021



Plan de la présentation

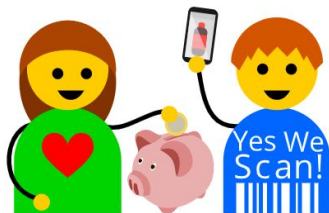
Partie 1: Choix de l'application

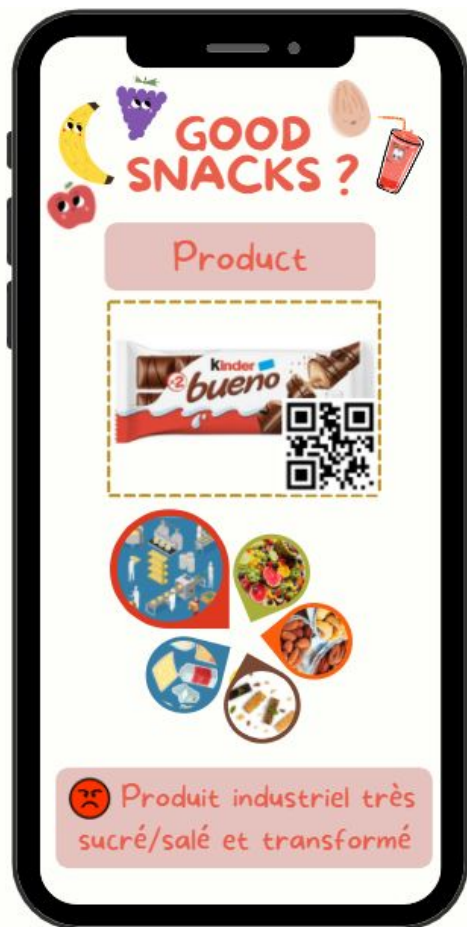
Partie 2: Préparation du jeu de données

Partie 3: Analyse du jeu de données clean

Partie 4: Modélisation

Partie 5: Conclusion





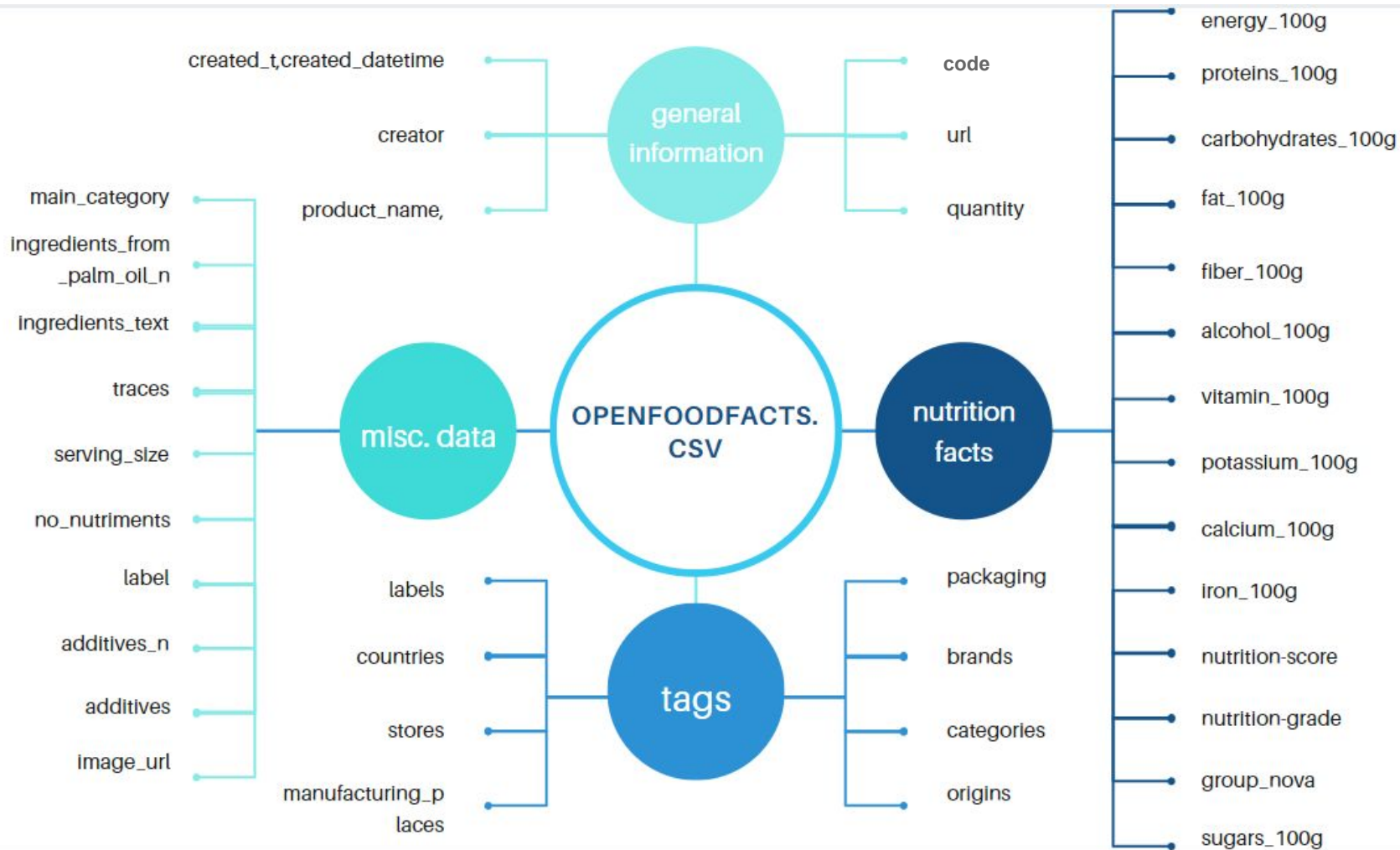
1 Application

- Evaluation du produit scanné par le consommateur suivant :
 - Nutri_score,
 - Nova_classification et
 - Snacks_grade une nouvelle classification goûters (snacks)
- Etude de la qualité du goûter enfant en France
- Aider l'utilisateur pour un meilleur choix (recommandation)



2

Préparation du jeu de données



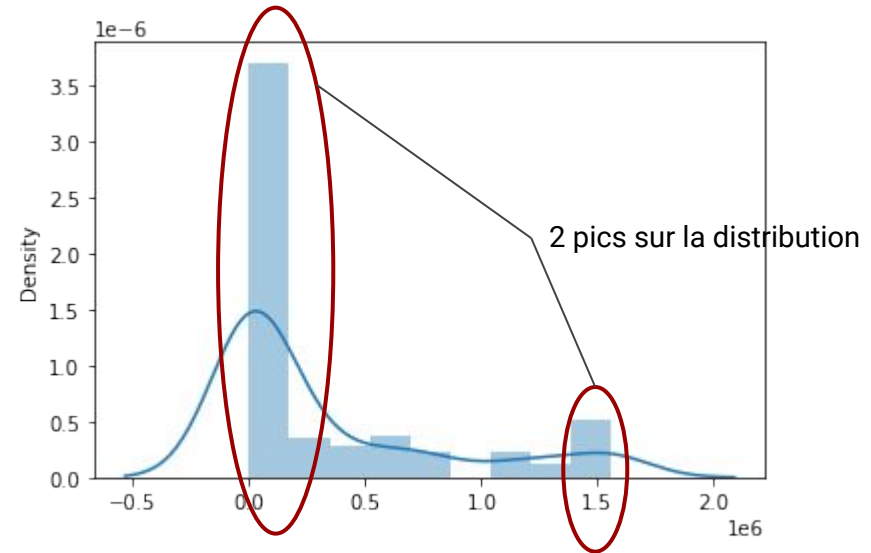
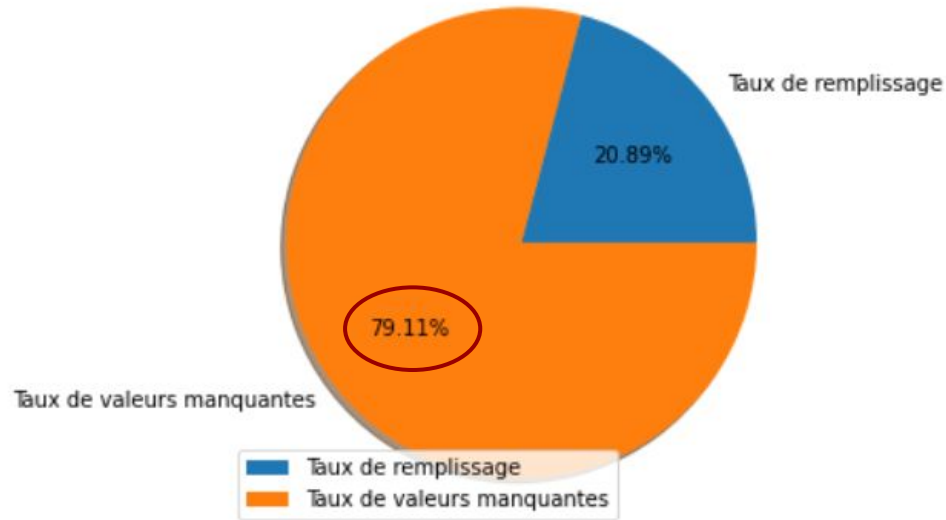
Description de la Dataset

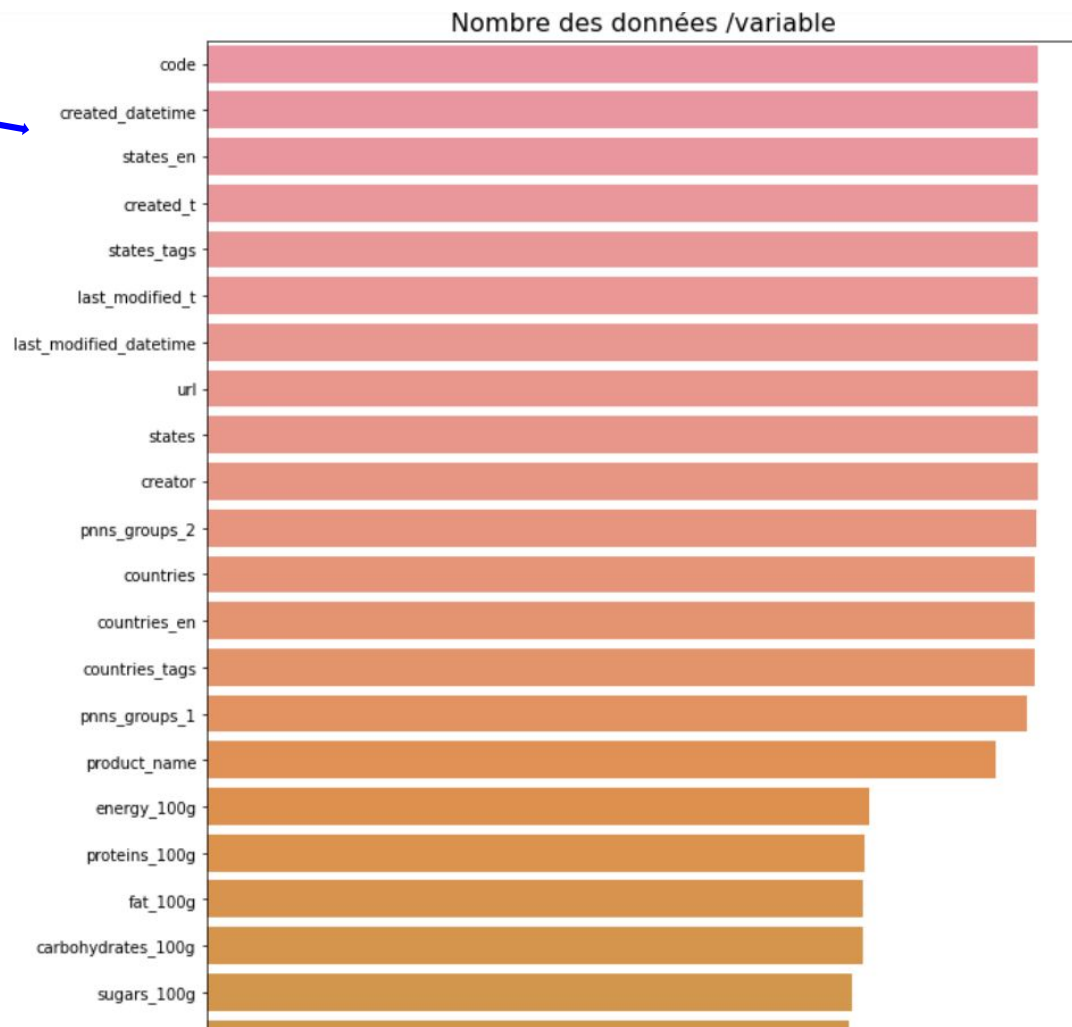
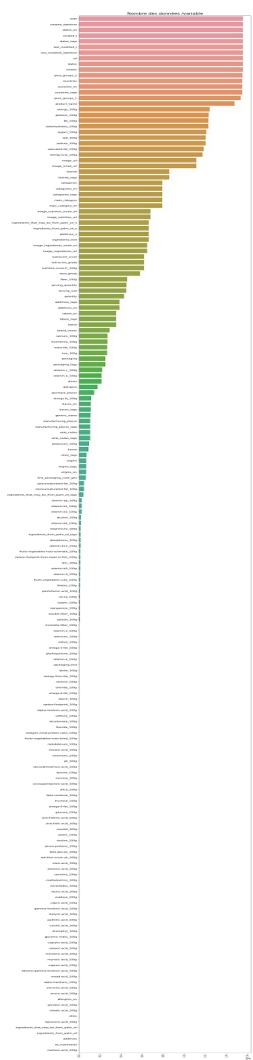
- **Variable target** : Nutri-score (à 5 modalités) et Nova groupe (à 4 modalités)
- **Taille** : 1555491, 183
- **Types de variables** : qualitatives : 58, quantitatives : 125
- **Doublons**: 4
- **Analyse des valeurs manquantes** :
 - beaucoup de NaN (moitié des variables > 98% de NaN)
 - 5 groupes de données: information général, Mots clés, ingrédients, Apports nutritionnels ,donnée divers

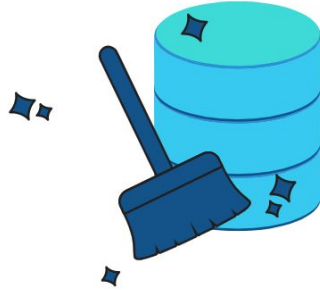
Type des variables

- **Les variables quantitatives**
 - *Discrète* :quantity, nutriscore_score
 - *Continues*:code, energy_100g, fat_100g, carbohydrates_100g, fiber_100g, proteins_100g, potassium_100g, glycemic-index_100g...
- **Les variables qualitatives**
 - *Nominales*: labels_en, categories, categories_tags , categories_en, 'pnns_groups_1', 'pnns_groups_2..
 - *Ordinales*: nova_group, nutrition_grade_fr

Taux de remplissage







Nettoyage

```
def cleaning_data (data)
```

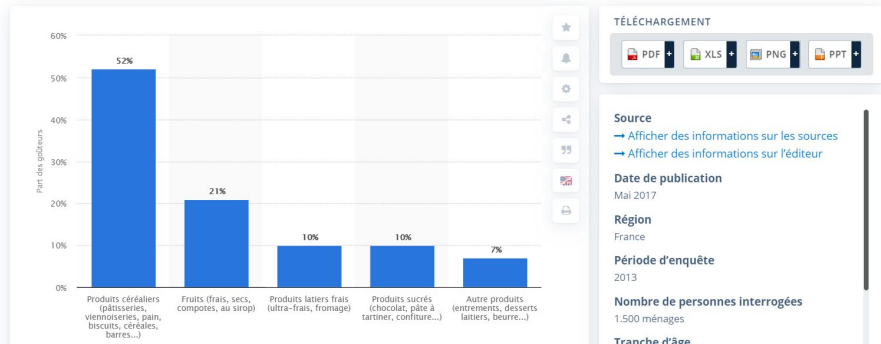
- Supprimer :
 - les observations et variables vides
 - les doublons
 - les observations sans nom et code
 - les codes nulles
- Définir les variables ordinale
- Convertir df_time to datetime.
- Supprimer:
 - valeurs nutritives_100g négatives
 - valeurs nutritives_100g >100g
 - energiekcal_100g >900

Choix catégorie snacks



Biens de consommation > Alimentation et nutrition

Composition des goûteurs pris par les enfants ayant entre 3 et 17 ans en France en 2013, par produit alimentaire



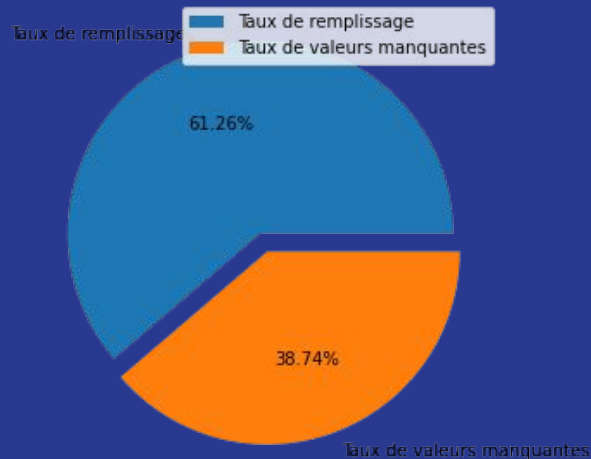
```
def choix_Catg (data,var, catg_choix,  
taux_var):
```

- Standardiser des Str
- Garder que les catégories snacks(pnn2)
- Garder les variables taux 40 % de valeurs
- Traitement des valeurs aberrantes :
 - $\text{sum}(\text{Fat}, \text{proteine}, \text{carbo}) > 100$
 - $\text{sugar} > \text{carbohydrate}$
 - $\text{sat_fat} > \text{fat}$
 - $\text{salt} > 5\text{g}$
 - $\text{fiber} > 25\text{g}$
 - Outliers = $Q3 + 15 * (Q3 - Q1)$
- Imputation energy: suivant l'équation de l'énergie
- Imputation par **IterativeImputer** pour les variables numériques
- Supprimer les variables non intéressantes pour le projet ou en double comme _tag ou datetime

Sauvegarde

- Taille de la base de données :(178351, 47)
- Sauvegarde de la DataFrame nettoyée

Taux de remplissage après imputation

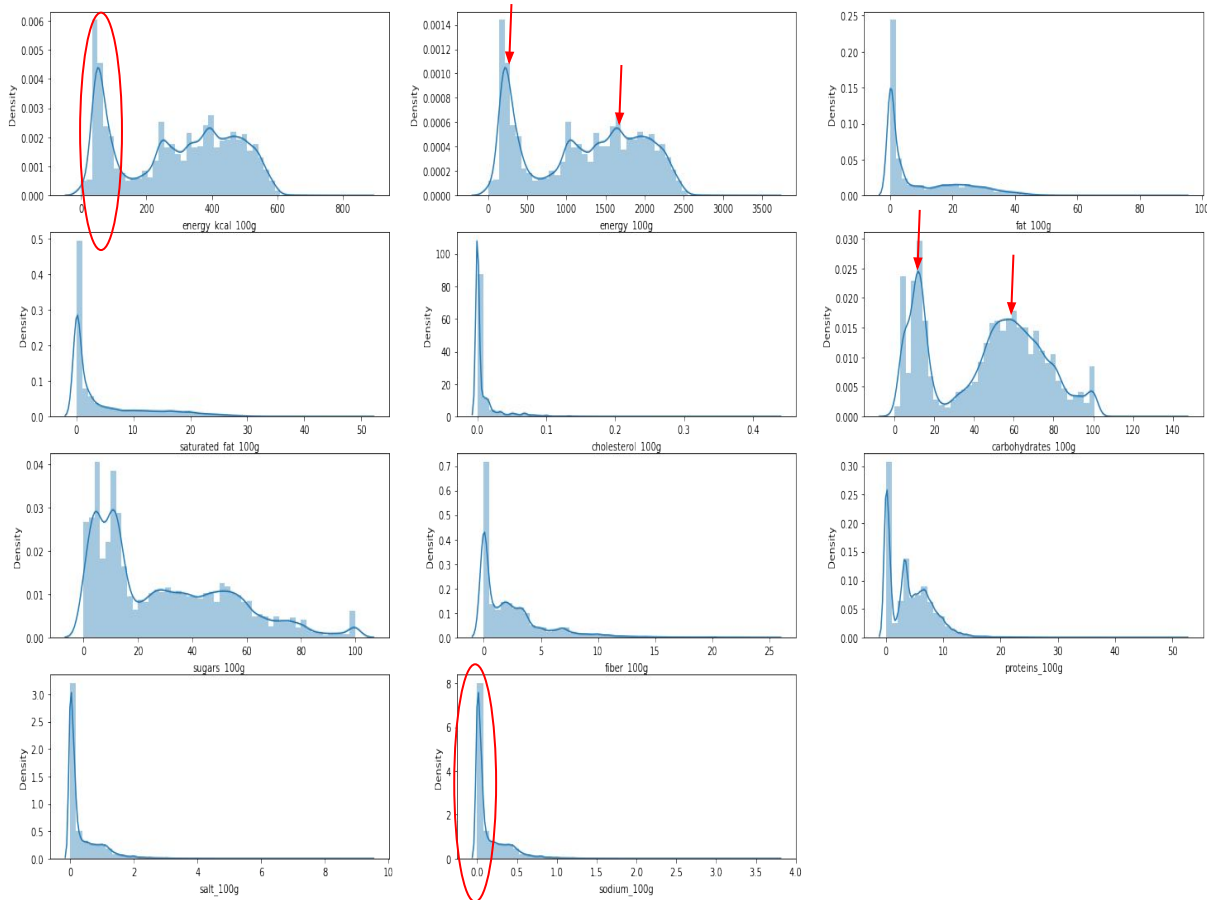


- Analyse descriptive des données
- Exploration et segmentation des données

3

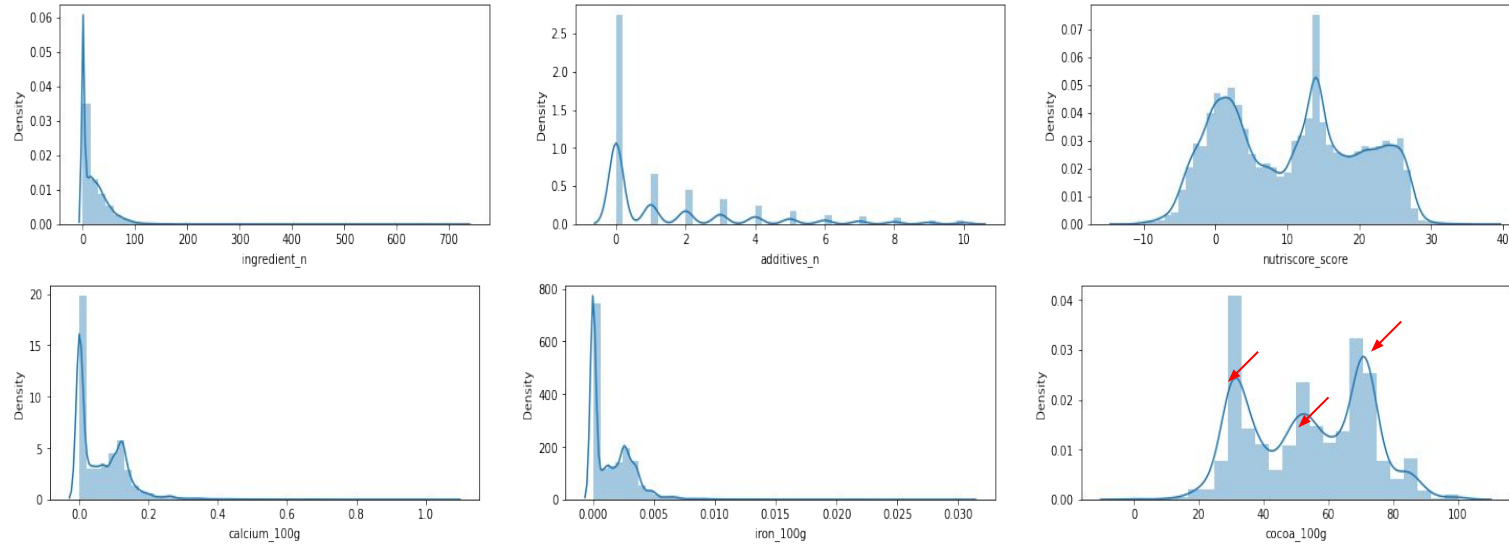
Analyse du jeu de données clean

Analyses univariées : Variables quantitatives distribution



- des pics au alentour de 0 pour la plupart des variables
- Types d'asymétrie des distributions :
 - **asymétrie vers la droite** (positivement biaisées): fat_100g, saturated_fat_100g, cholesterol_100g, fiber_100g, salt_100g, sodium_100g
 - distribution **bimodale** et **positivement biaisée**: energy_100g, carbohydrates_100g, sugars_100g et proteins_100g

Analyses univariées : Variables quantitatives distribution



- Types d'asymétrie des distributions :
 - distribution **bimodale** et positivement biaisée: calcium et fer
 - distribution **trimodale** pour nutriscore_score et cacao

-Absence des distributions normales



```
1 #pour les variable_100g nutritive
2 data_test_norm=df[var_nutr]
3 pg.normality(data_test_norm, method='normaltest').round(3)
```



	W	pval	normal
energy_kcal_100g	4053070.140	0.0	False
energy_100g	3731019.523	0.0	False
fat_100g	21789.950	0.0	False
saturated_fat_100g	35392.929	0.0	False
cholesterol_100g	38132.320	0.0	False
carbohydrates_100g	100065.660	0.0	False
sugars_100g	12977.523	0.0	False
fiber_100g	37151.149	0.0	False
proteins_100g	57696.211	0.0	False
salt_100g	90177.194	0.0	False
sodium_100g	90179.899	0.0	False

Test de normalité

	W	pval	normal
vitamin_a_100g	24564.448	0.0	False
vitamin_c_100g	42394.852	0.0	False

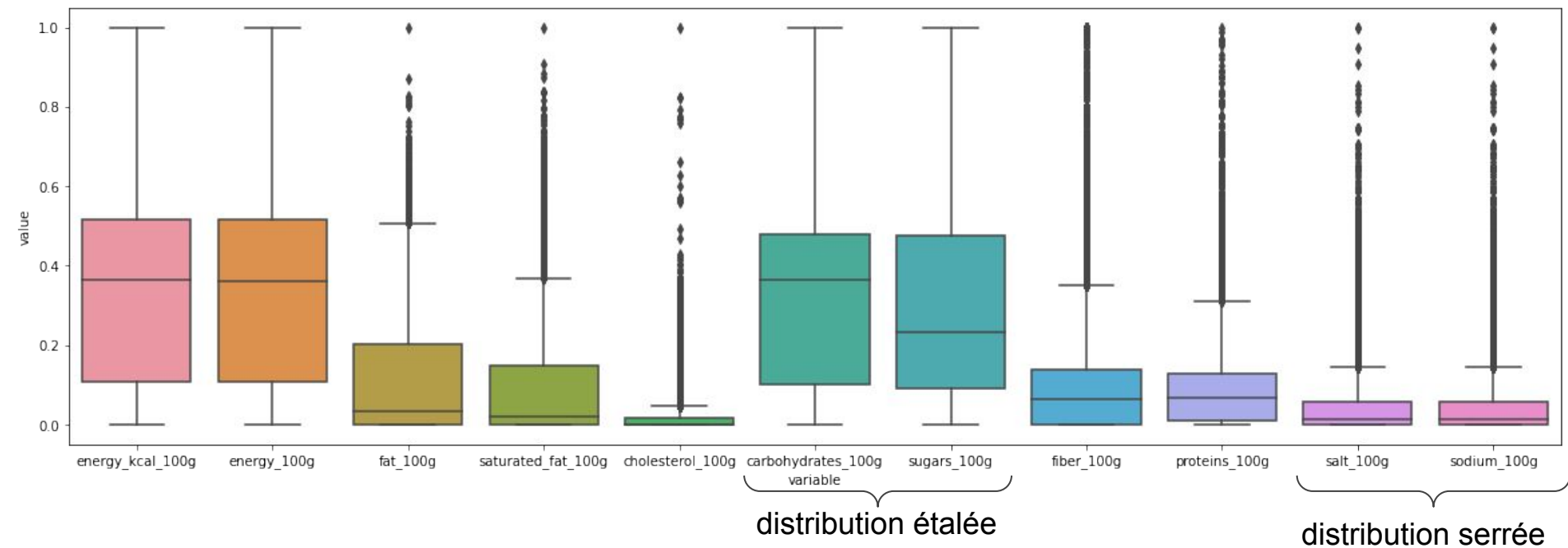
```
[ ] 1 #pour les variable_100g minéraux
2 data_test_norm=df[var_min]
3 pg.normality(data_test_norm, method='normaltest').round(3)
```

	W	pval	normal
calcium_100g	30292.705	0.0	False
iron_100g	20989.090	0.0	False
cocoa_100g	1734.326	0.0	False

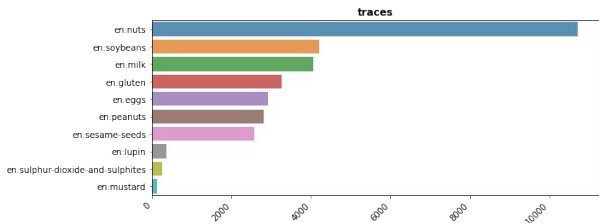
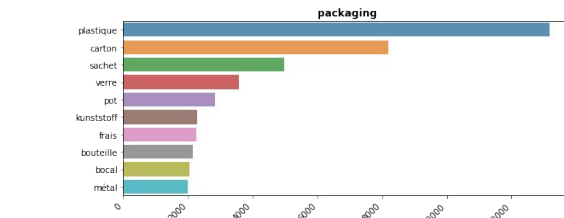
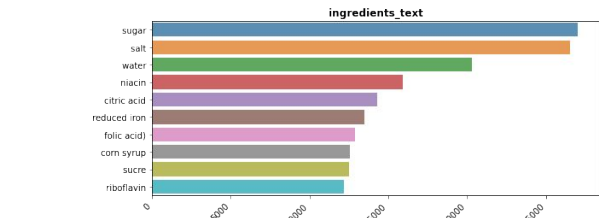
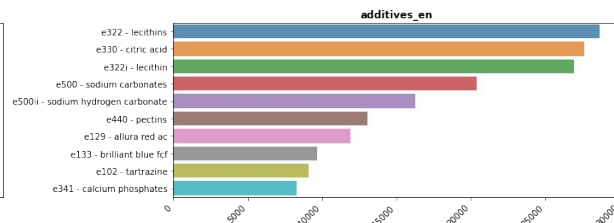
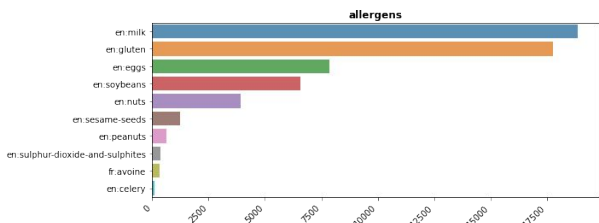
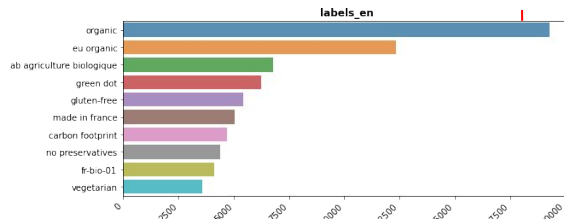
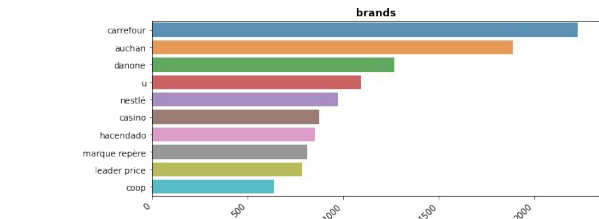
```
[ ] 1 #pour les variable_100g minéraux
2 data_test_norm=df[var_grade]
3 pg.normality(data_test_norm, method='normaltest').round(5)
```

	W	pval	normal
additives_n	54572.70462	0.0	False
nutriscore_score	93716.81922	0.0	False

Boite à moustache des valeurs_100g

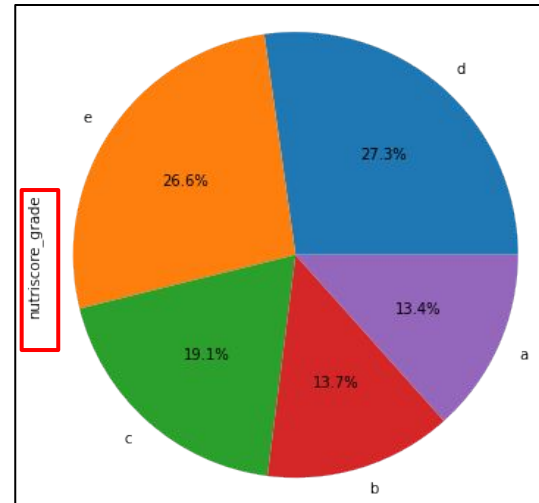
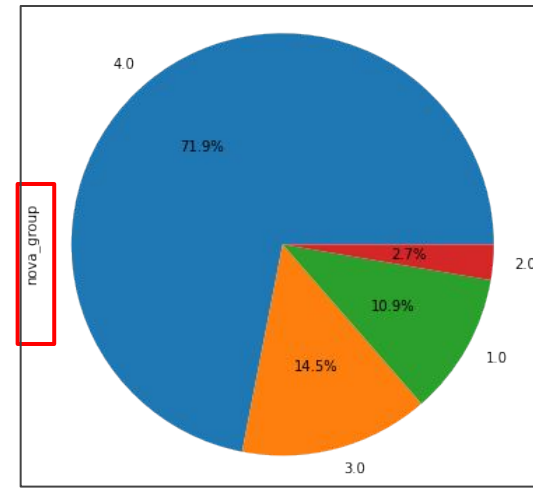
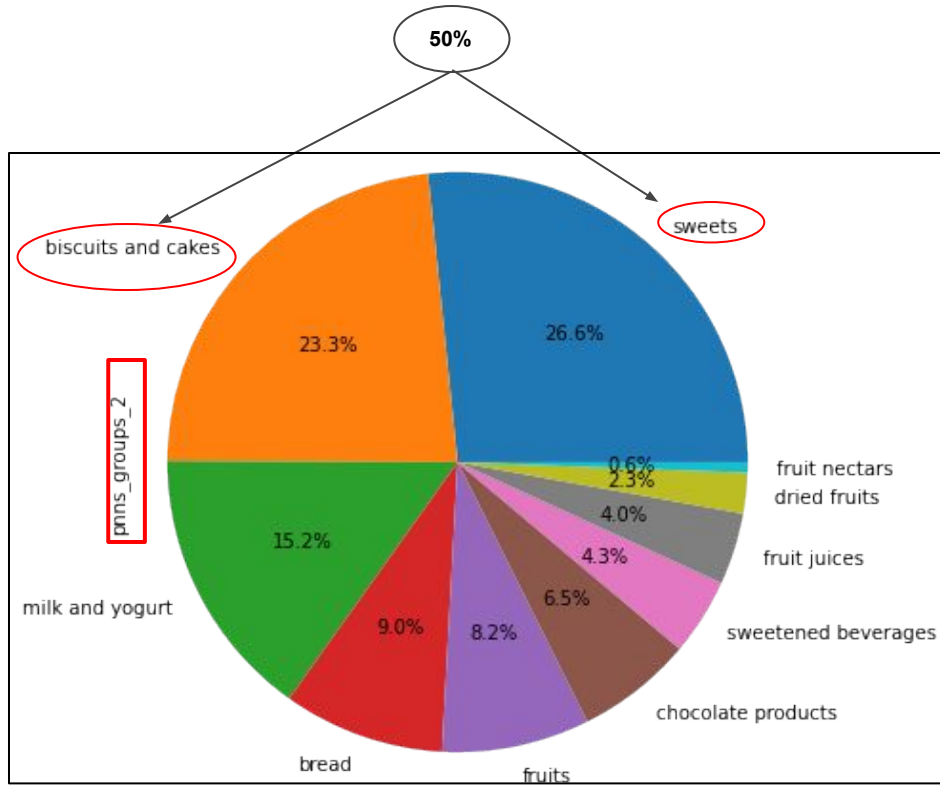


Variables qualitatives nominales

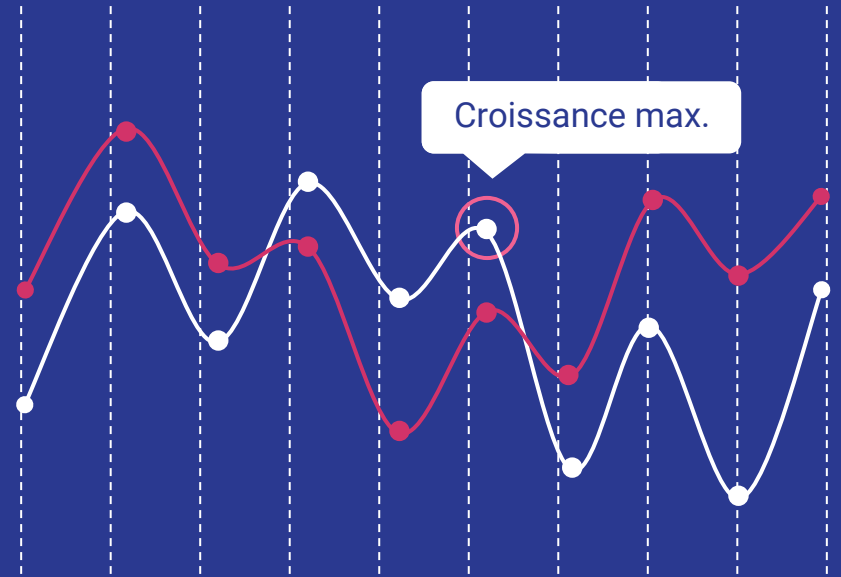


- **Brand:** les principales sont les marques de distributeur(MDD): Auchan , carrefour. Danone est la première marque non MDD
- **Label:** le label Organic domine suivi par de made_in_france, empreinte_carbone.
- **Allergen et Traces:** les 4 premiers allergènes: gluten, lait, œufs, soja
- **Additive:** les principales additives (lectin, acide_citrique,pectines..)
- **Packaging:** les 3 types de packagings(carton, plastique, verre)
- **Ingédients:** les sources de sucre qui dominent(sucre, glucose, fructose) confirme les graphes de distribution du sucre et glucide.

Graphique en secteurs



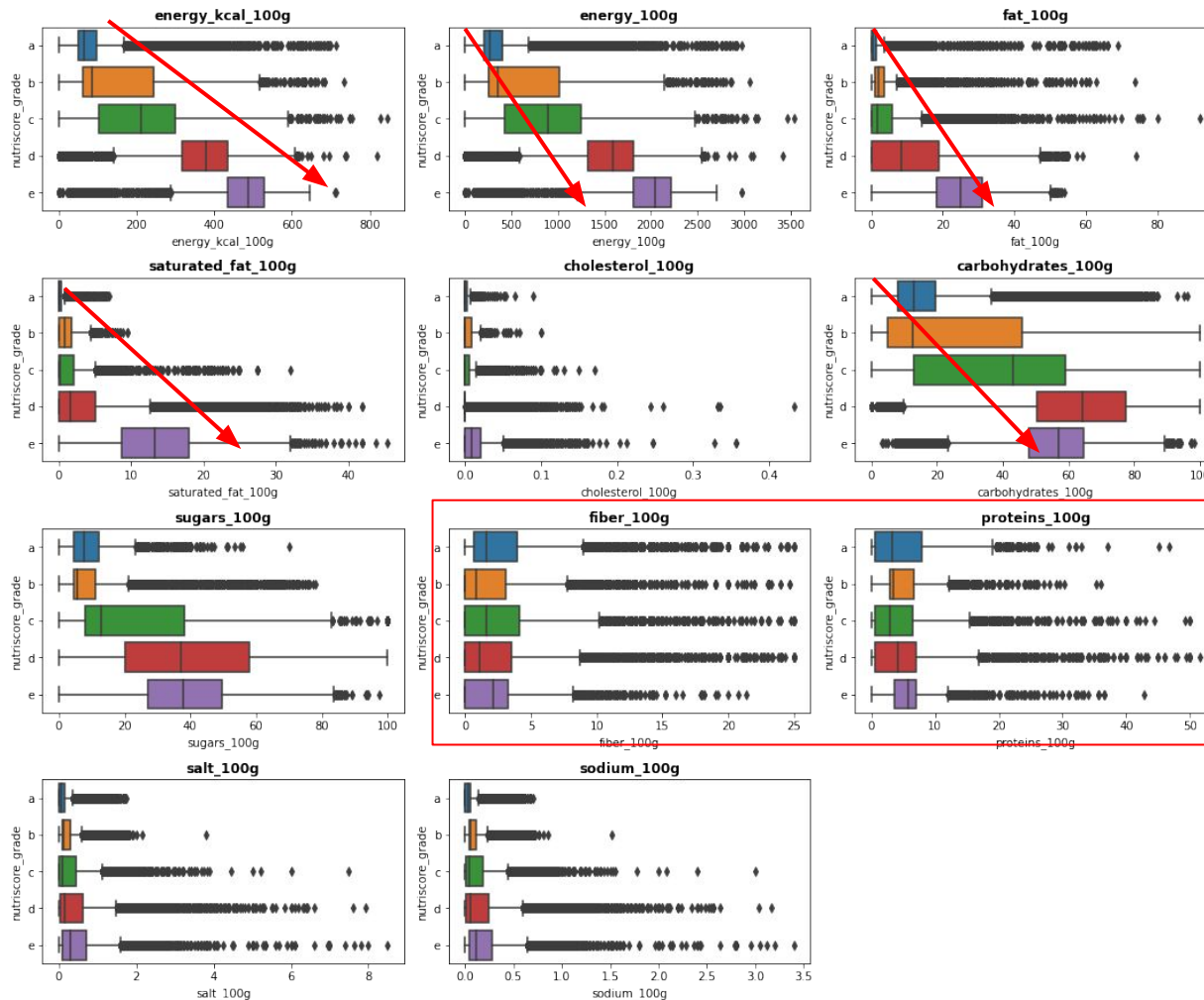
Analyse bivariée



Variables Qualitative(Nutri_grade)

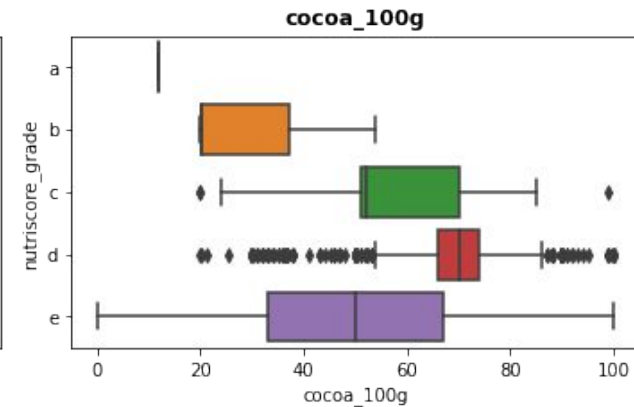
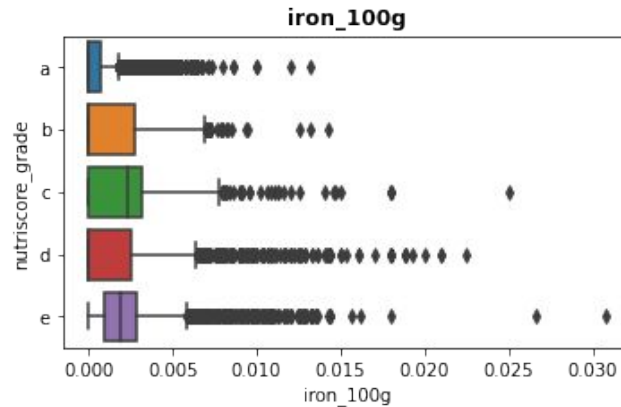
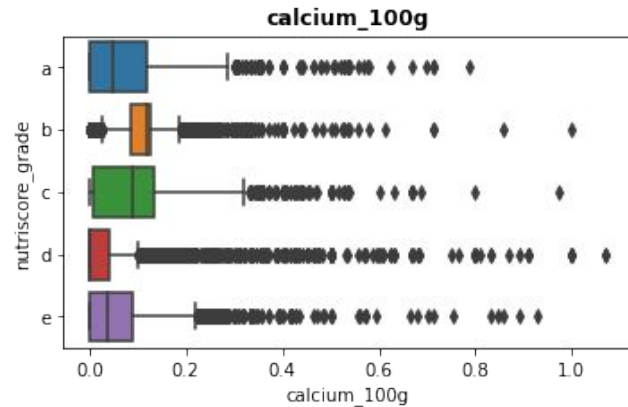
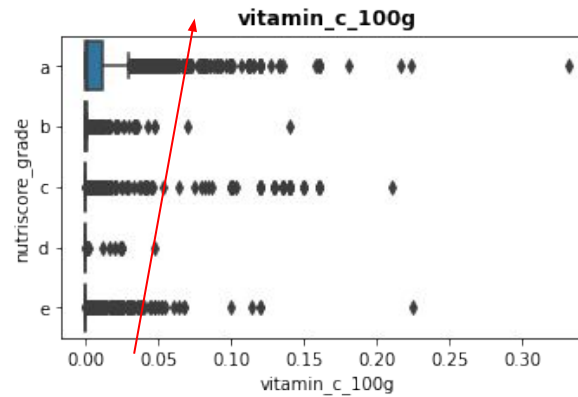
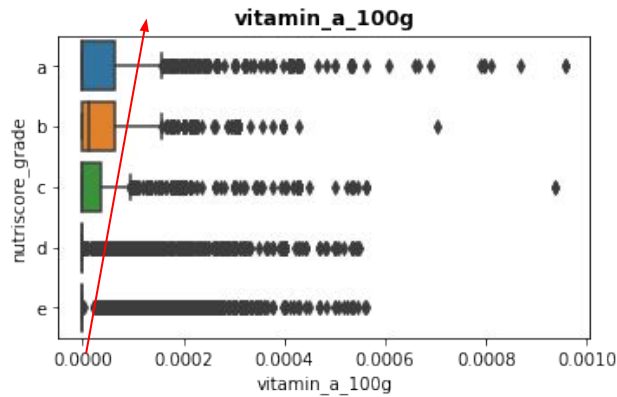
Nutriscore Vs variables de nutrition

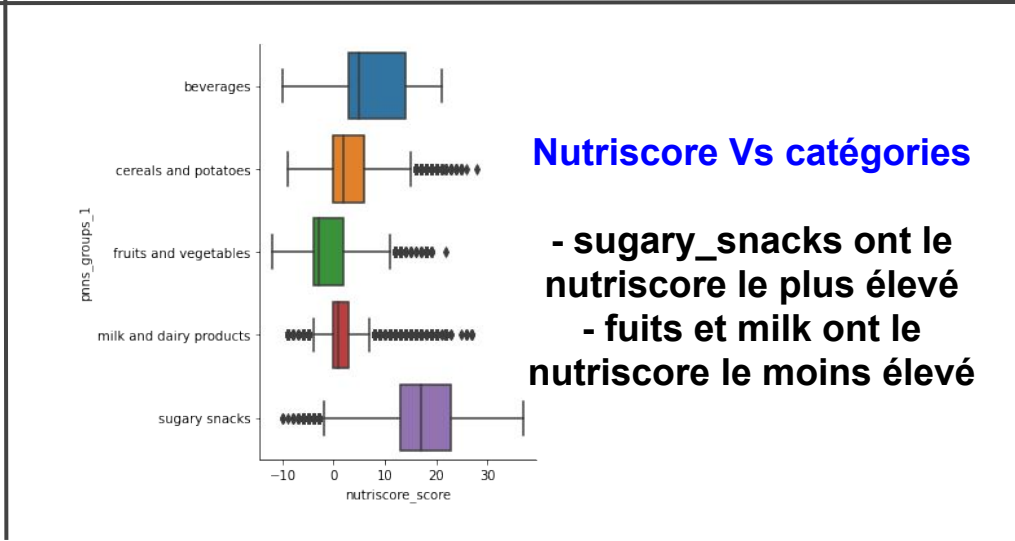
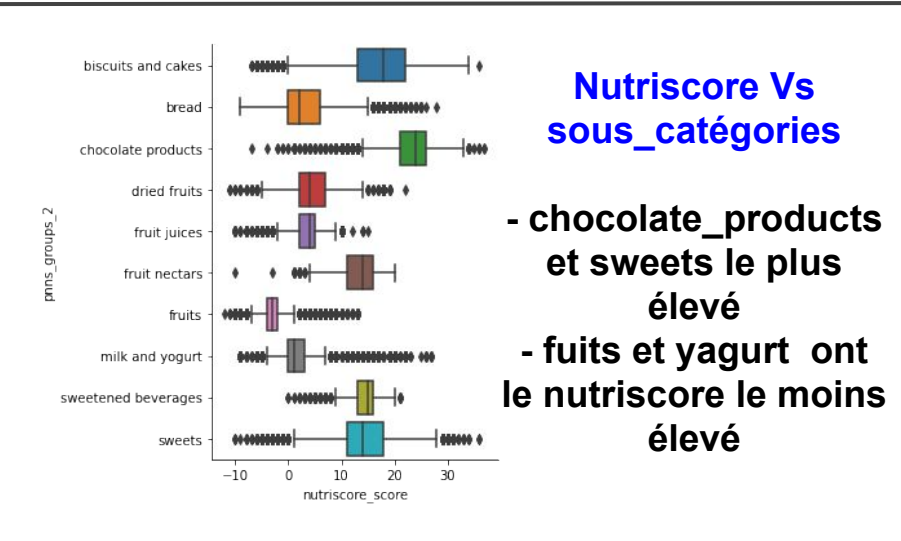
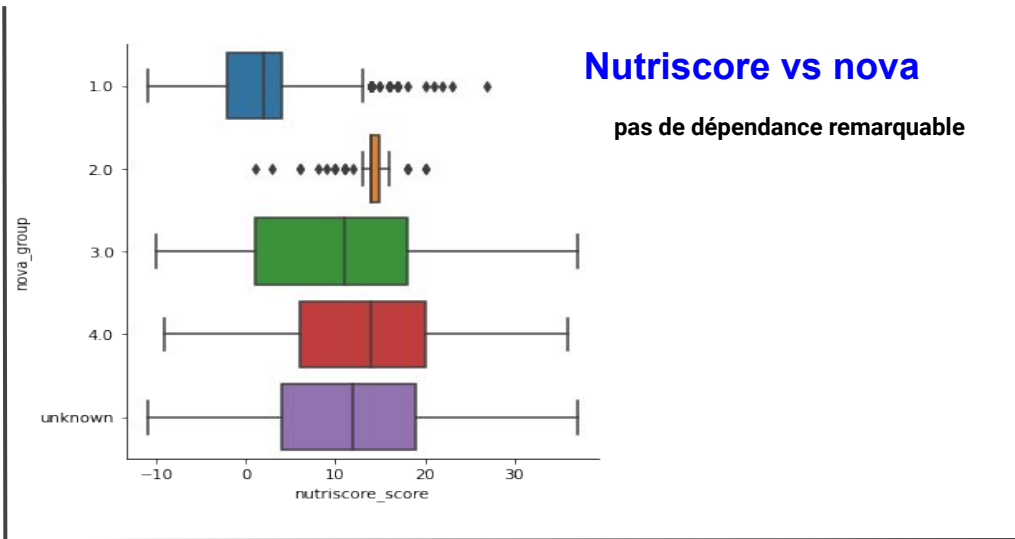
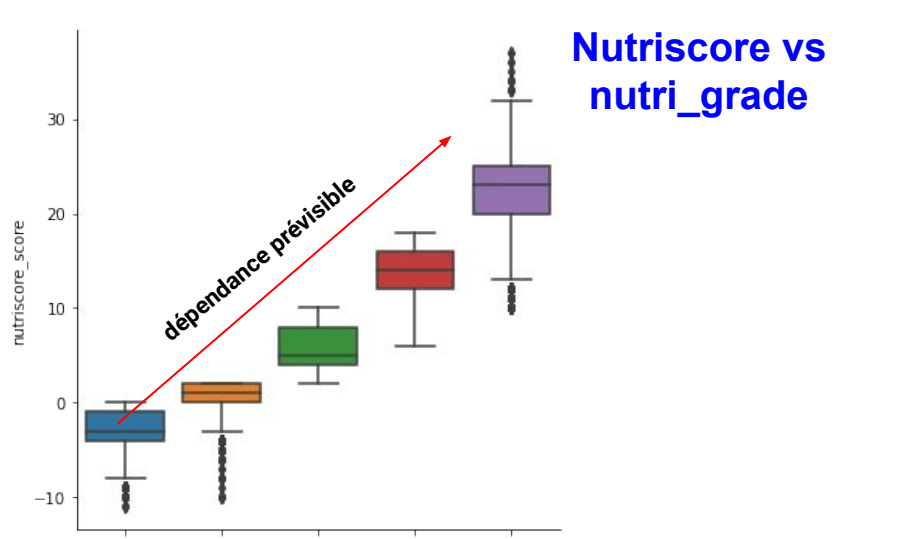
- L'energy,fat,cholesterol, carbohydrates, salt, sodium semblent augmenté lorsque le classement du produit passe de A à E.
- uniforme proteins et fiber c'est uniforme(sans impact)



Nutriscore Vs variables de nutrition

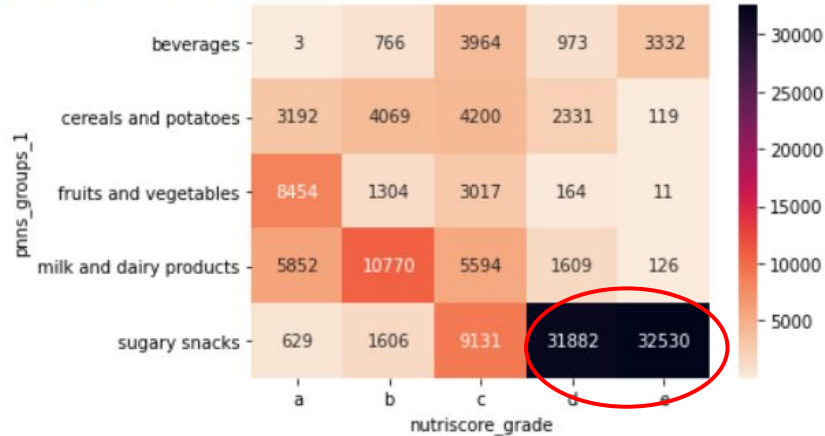
- les vitamines semblent augmentées lorsque le classement du produit passe de E à A.
- uniforme avec le fer, calcium, cacao (sans impact visible)



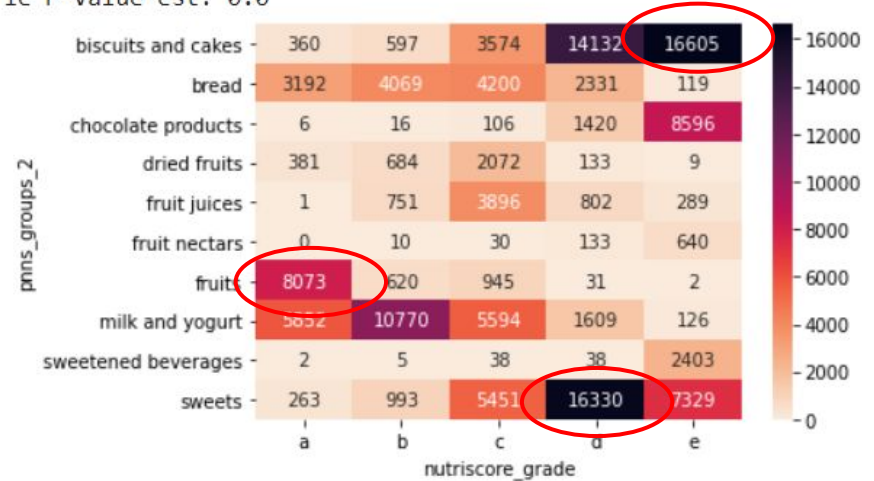


nutri_grade vs pnn

le chi2 est: 108279.18626025991
le Degrés de liberté est: 16
le P-Value est: 0.0



le chi2 est: 149707.23435472805
le Degrés de liberté est: 36
le P-Value est: 0.0



- Les meilleurs(a): fruits and vegetables
- les mauvais(d,e): sugary snacks
- classe (c) est uniforme
- L'hypothèse nulle (H0) de ce test est la suivante : les deux variables X et Y sont indépendantes.
- $p < 0.05$: rejet de H0

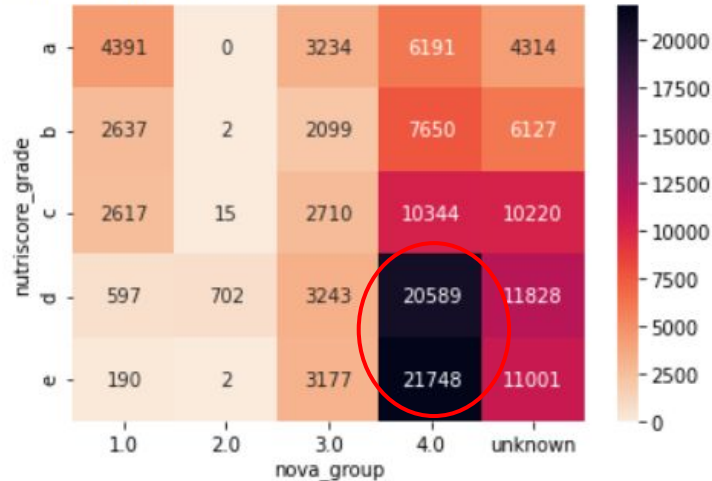
- Les meilleurs(a): fruits
- les mauvais(d,e): biscuits and cakes et sweets
- classe (c) est presque uniforme
- $p < 0.05$: rejet de H0

nutri_grade vs nova

le chi2 est: 18367.768827369844

le Degrés de liberté est: 16

le P-Value est: 0.0



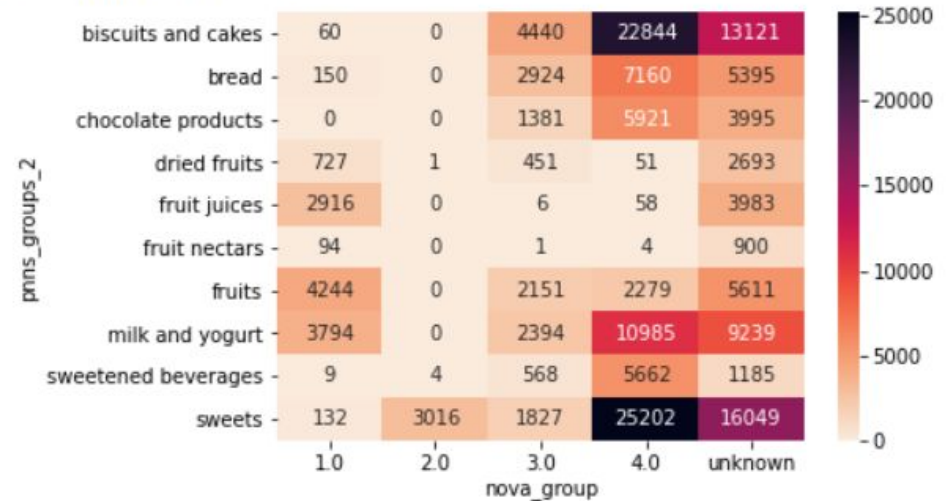
- $p < 0.05$ rejet de H_0

pnn vs nova

le chi2 est: 61303.29346070653

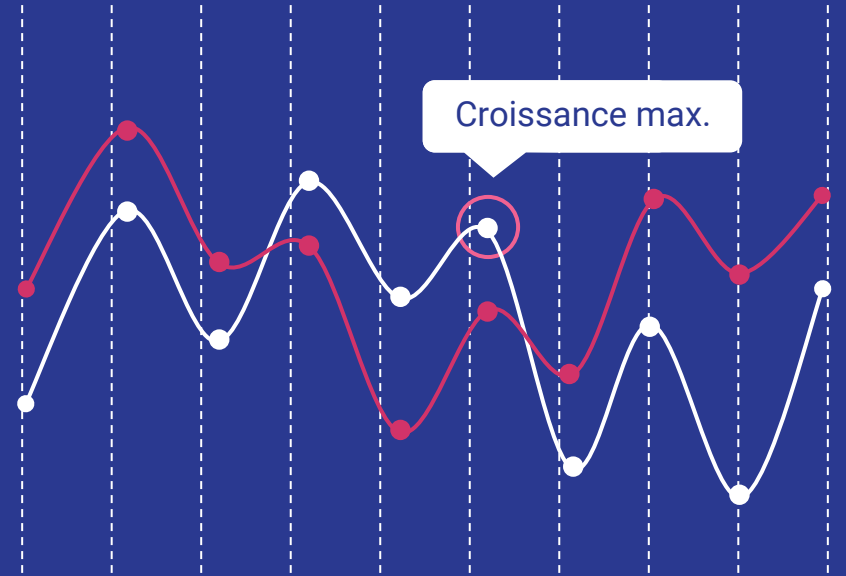
le Degrés de liberté est: 36

le P-Value est: 0.0



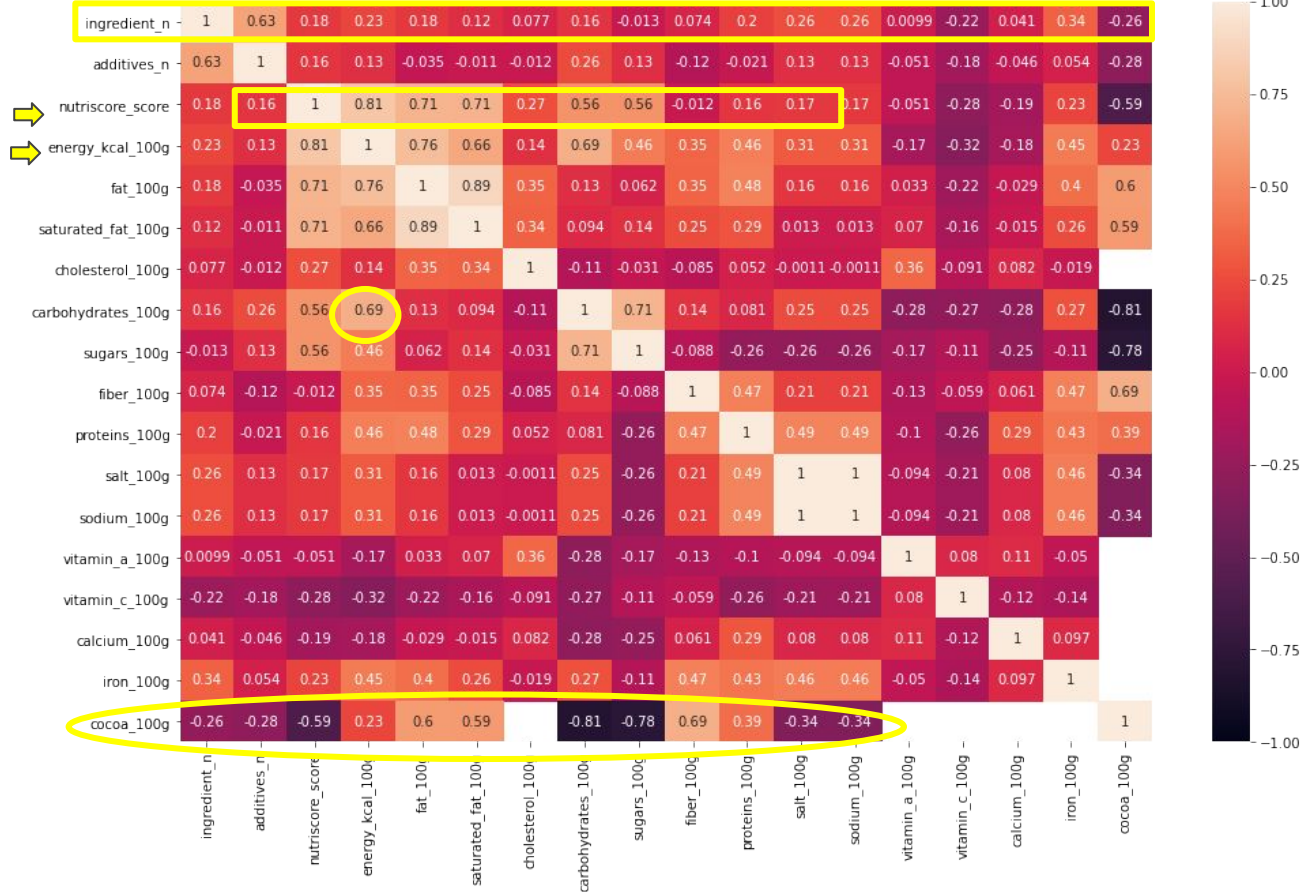
- Les moins transformés(1): fruits
- les plus transformés(4): biscuits and cakes et sweets
- Nova 3 est presque uniforme
- $p < 0.05$ rejet de H_0

Analyse multivariée



Matrice de corrélation

Correlation Heatmap

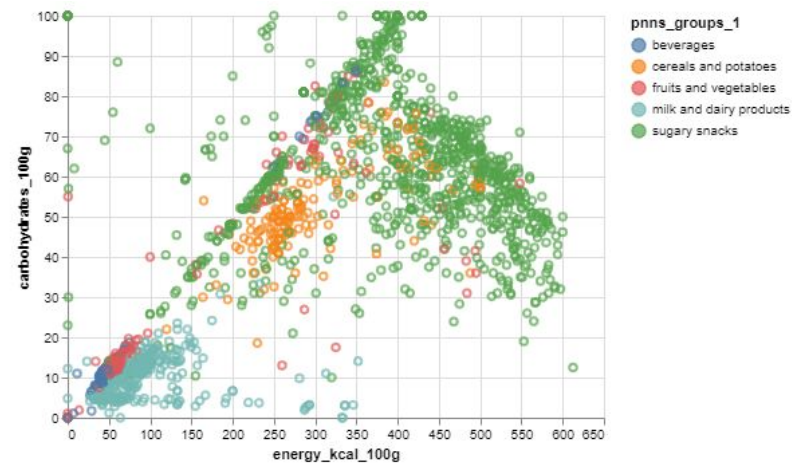
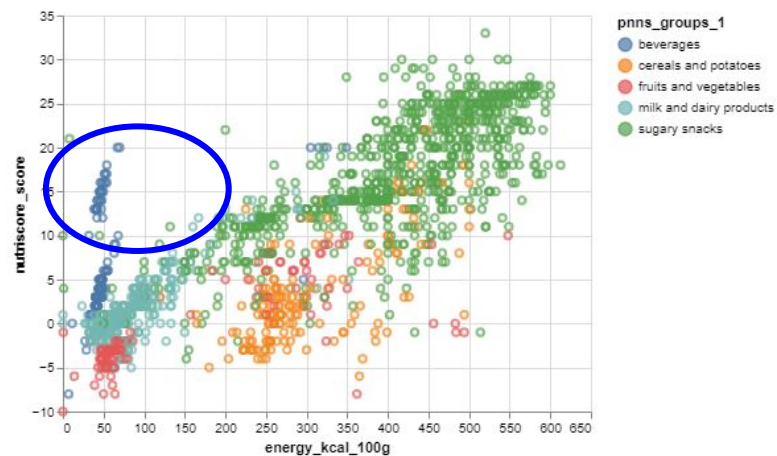
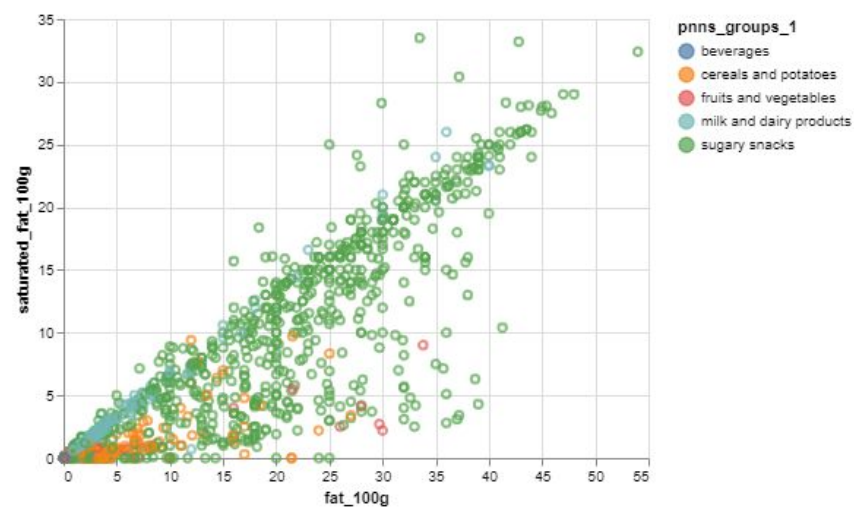
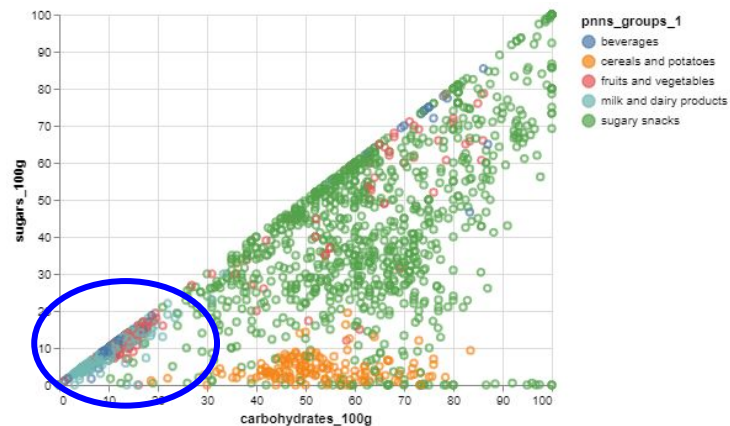


corrélation Prévisibles

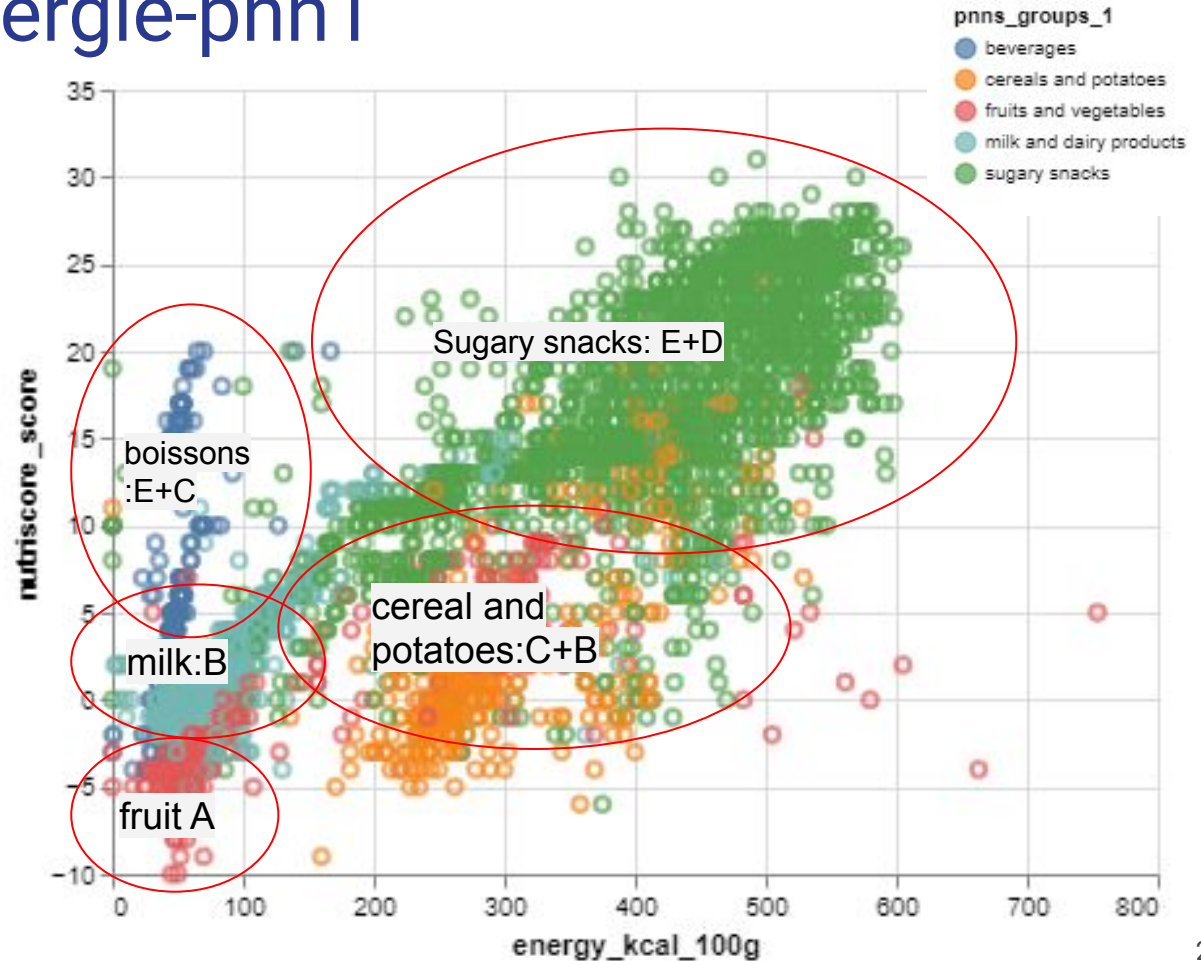
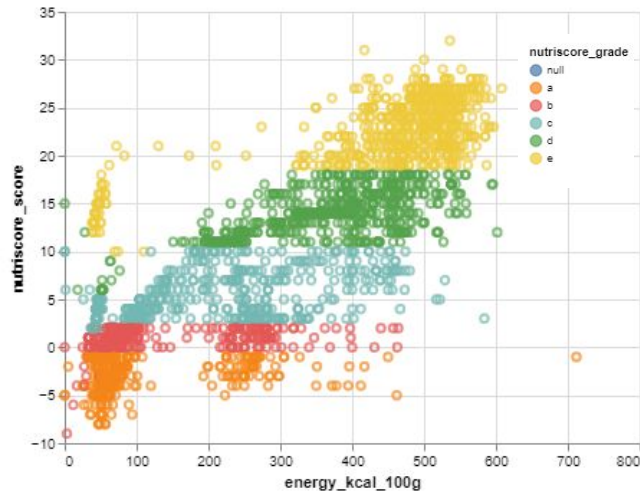
- nutriscore avec glucide, lipide et énergie
- énergie avec glucide, lipide, protéine et sel et sodium
- sucre et glucide

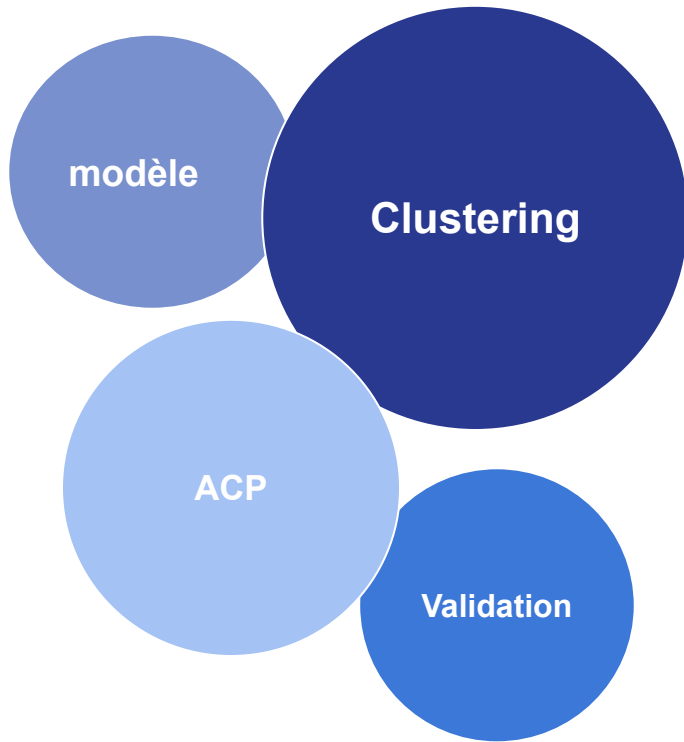
Non Prévisibles

- cacao et les vitamines ont des corrélations négatives
- pas de corrélation entre nombres d'ingrédients et nutriscore



Nutriscore-énergie-pnn1





Partie 4: Modélisation

Approche méthodologique

Appliquer une Analyse en Composantes Principales pour réduire les variables et déterminer les nouvelles composante qui peuvent nous aider à créer ce nouveau classement.

- Interprétation des résultats
- comparaison avec les classification Nutriscore et Nova

1

2

3

4

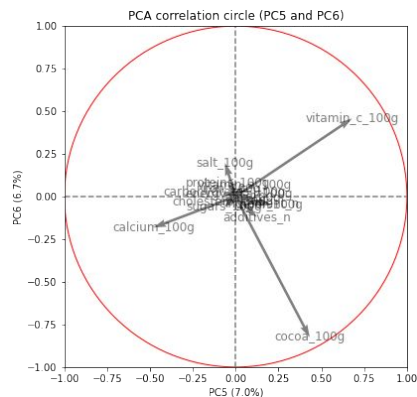
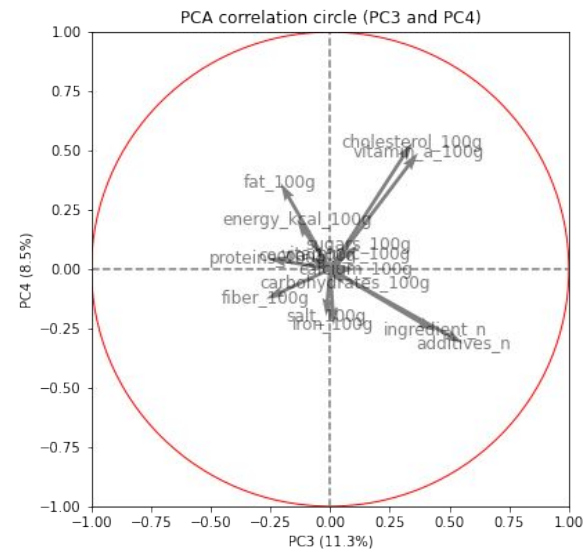
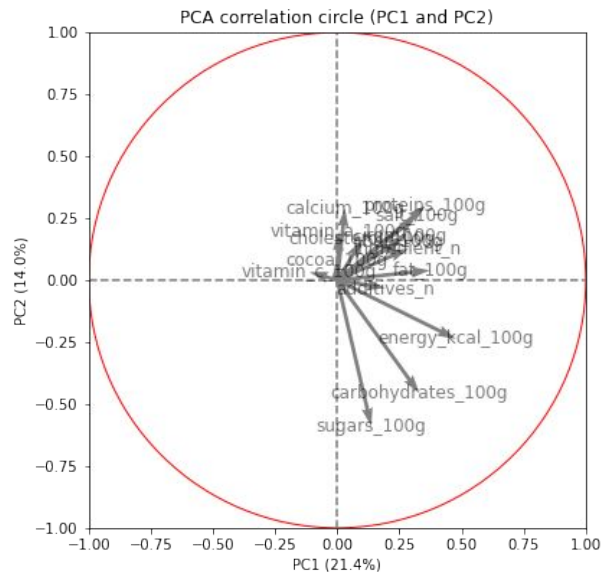
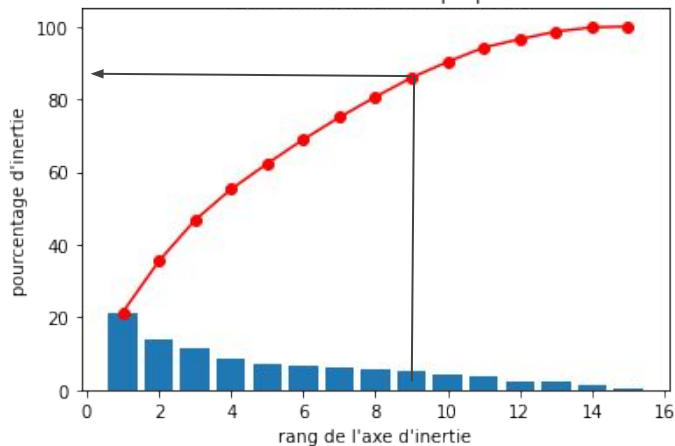
Choisir les 15 variables pour la classification en prenant compte des vitamines et nombre d'ingrédients

Modèle de classification non supervisée : **K_means** et créer **Snack_Grade** en France

features = ['ingredient_n', 'additives_n', 'energy_kcal_100g', 'fat_100g', 'cholesterol_100g', 'sugars_100g', 'fiber_100g', 'proteins_100g', 'salt_100g', 'vitamin_a_100g', 'vitamin_c_100g', 'calcium_100g', 'iron_100g', 'cocoa_100g']

ACP: Analyse en Composantes Principales

Eboulis des valeurs propres



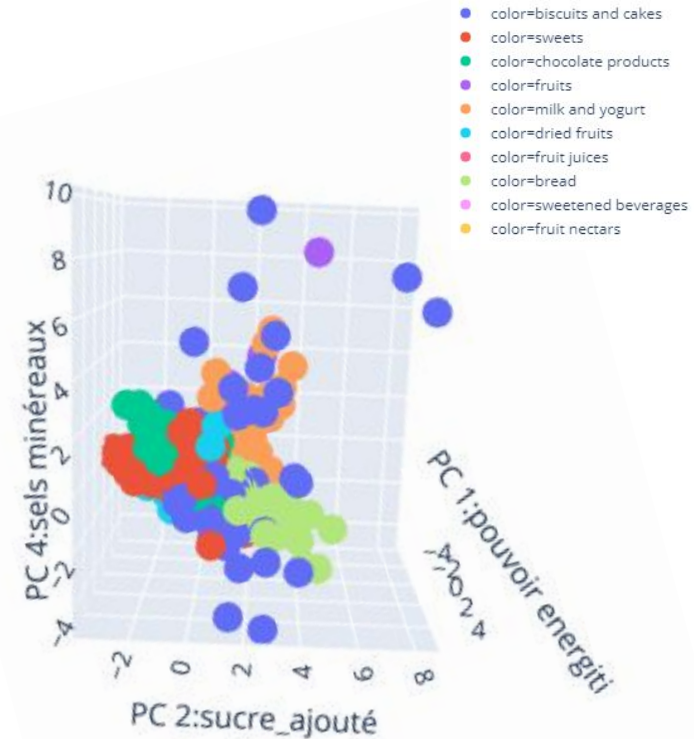
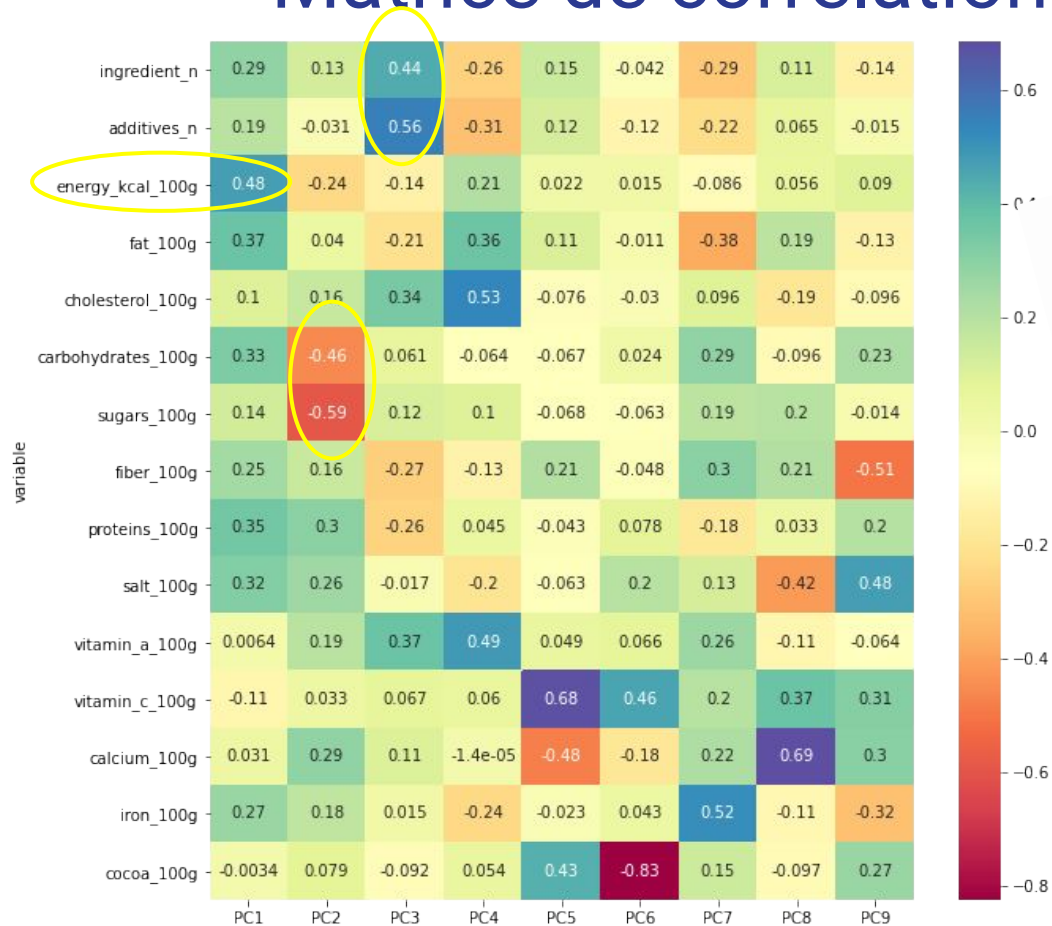
PC1: Pouvoir énergétique

PC2: l'absence du sucre

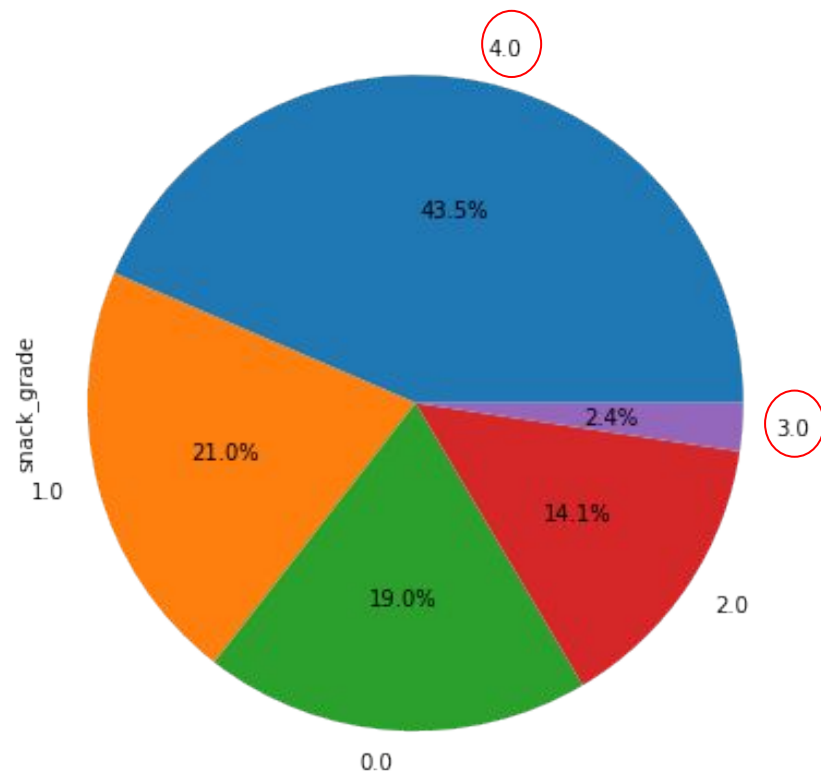
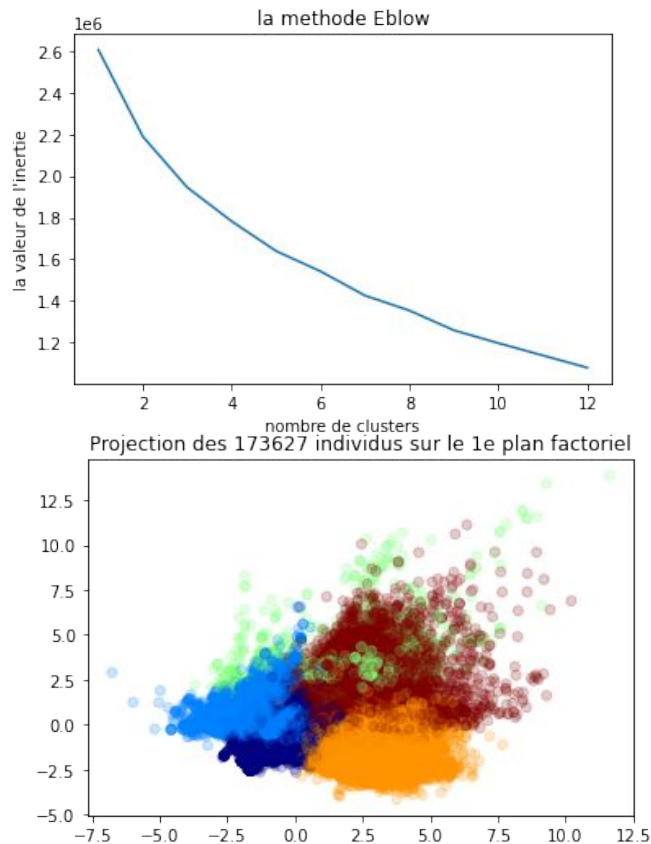
PC3: additives synthétiques (non naturelle) sucre, sel et additives non naturels

PC4: fluide ou solide

Matrice de corrélation des plans factorielles



Classification non supervisée: K_means



Snack_grade:

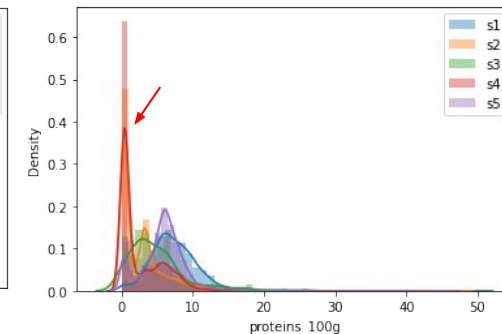
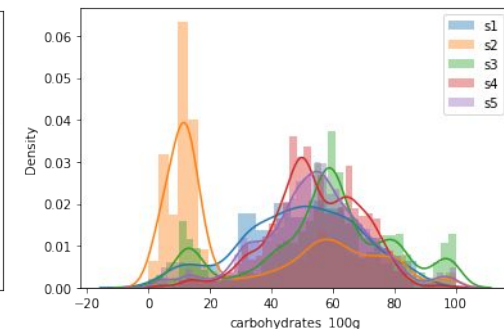
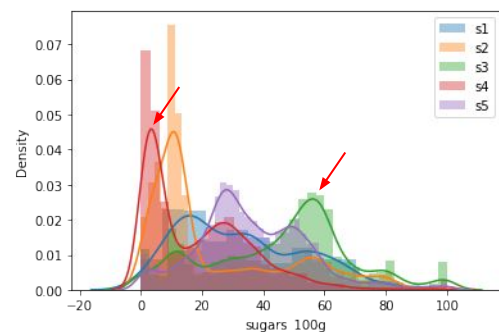
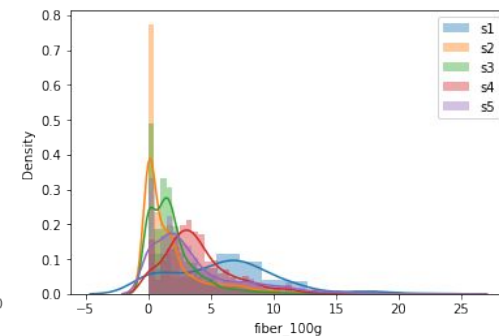
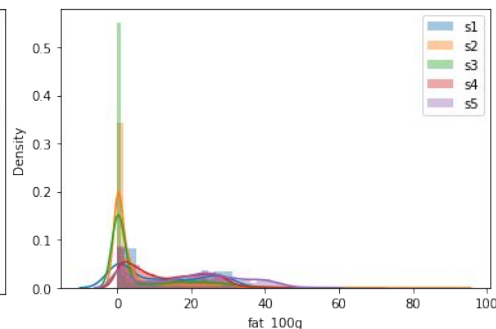
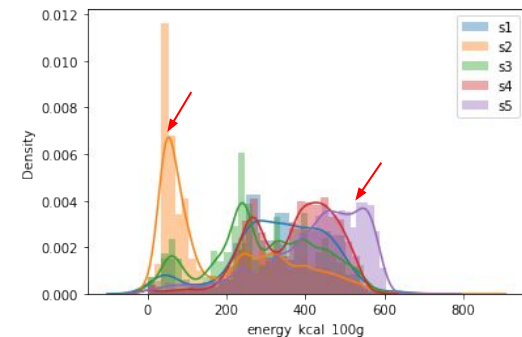
s1: plus de **fibre** et **protéine** et le nombre d'ingrédient faible (produit naturelle simple comme fruit)

s2: le moins énergétique et moins **sucré**(produit laitier)

s3: plus d'additif et transformé (sel et sucre)

s4: Moins de **sucré** ajouté avec **gras** et **fibre** élevés(fruits sec)

s5: plus **énergétique**, **gras** et **protéiné**



Conclusion



- Jeux de données avec plusieurs valeurs manquantes et aberrantes (facteur humain)
- Faire interagir d'autres variables pour la nouvelle classification (label, empreinte carbone.)
- manque d'avis d'un professionnelle nutritionniste
- le snack_grade est-il le mélange de nova et nutriscore?
- est-ce que un autre modèle de prévision serait plus fiable que K_means?



Merci