Projet 3
Concevez une application au service de la santé publique

Zeynep Erdem

06-04-2023

Santé publique France

Sommaire

Santé publique France

Rappel de la Problématique

L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.

On va utiliser le jeu de données Open Food Facts

Environnement

- Python: 3.8.16
- Pandas: 1.5.3
- Numpy: 1.23.5
- Seaborn: 0.12.2
- Matplotlib: 3.7.1
- Missingno: 0.5.2
- Wordcloud: 1.8.2.2
- Statsmodels: 0.13.5
- Scipy: 1.10.0
- Sklearn: 1.2.1

La situation en France

- Etude L'agence Santé publique
- En 2016 : Elèves 7 à 9 ans
- Selon les seuils de l'International Obesity Task Force
  - 18.7 % des filles
  - 14.4 % des garçons
  - en surpoids (dont obésité)

- Le goûter est le relai
  - indispensable
  - environ 10% de ce qu'un enfant va manger pendant une journée
  - qu'il soit équilibré
  - pas uniquement une source de produits sucrés.

Santé publique France

Le goûter idéal

- L'apport nutritif du goûter doit être
  - 3 à 5 ans,
    - de 200 à 250 kcal
  - 6 à 8 ans,
    - de 250 et 290 kcal
  - 9 à 12 ans,
    - de 290 et 360 kcal
- Il faut éviter des produits
  - ultra transformés,
  - gras ,
  - sucrés ou salés.

Santé publique France

Le principe d'application

GOÛTER SAIN

Entrez l'âge de l'enfant

Scannez ou entrez le nom de produit

Les informations nutritionnelle et nutri grade et goûter score

Recommandation de produits alternatifs de même catégorie selon goûter score

GOÛTER-SCORE

Recommandation de quantité selon le besoin énergétique de l'enfant

Santé publique France

NUTRI-SCORE

A B C D **E**

Additifs

E322 - Lécithines

Analyse des ingrédients

Huile de palme

Tableau nutritionnel	Tel que vendu pour 100 g / 100 ml
Énergie	2 252 kj (539 kcal)
Matières grasses	30,9 g
Acides gras saturés	10,6 g
Glucides	57,5 g
Sucres	56,3 g
Fibres alimentaires	0 g
Protéines	6,3 g
Sel	0,107 g

GOÛTER-SCORE

Pour un enfant de 4 ans, ne passez pas 15g de Noccialata par jour

NUTRI-SCORE

A B C **D** E

Additifs

E322 - Lécithines

E322i - Lécithine

Analyse des ingrédients

Sans huile de palme

Tableau nutritionnel	Tel que vendu pour 100 g / 100 ml
Énergie	2 218 kj (530 kcal)
Matières grasses	30 g
Acides gras saturés	6 g
Glucides	53 g
Sucres	51 g
Fibres alimentaires	4 g
Protéines	7 g
Sel	0,12 g

Présentation du Jeu de Données

- Informations générales
- Tags
- Ingrédients
- Infos nutritionnelles

- Il y a environ 320 000 lignes et 162 colonnes dans notre dataframe
- On a 76 % de valeur null dans notre dataframe
- On a des incohérences par des erreurs des contributeurs

Santé publique France

Les étapes du nettoyage

Étape du Nettoyage	Nb. Lignes	Nb. Colonnes
Jeu de donnée initiale	320772	162
Sélection des indicateurs	320772	29
Modification des noms des colonnes et suppression les accents	320772	29
Sélection la France	97485	26
Suppression des doublons pour le code barre	97483	26
Product name, Caféine, Alcool	88503	24
Sélection des catégories (pnns groups 1) et suppression tous null	32669	18
Vérification des outliers	32352	18
Traitements des valeurs manquantes	32352	18
Deuxième vérification des outliers	32183	18

Santé publique France

Vérification des outliers

- Méthode interquartile
- 6917 outliers 3595 produit
- Elles sont plutôt les valeurs atypiques

- 1 produit < 0
- 7 produits > 100

- Kcal ou kjoule ?
- Le montant energy ne peut pas passer 3795 kjoule
- 17 produits

- Le montant sodium ne peut pas passer 40 % de sel
- Aucun produit RAS

- Le montant matière grasse saturée ne peut pas passer le montant matière grasse
- 48 produits

- Le montant sucre ne peut pas passer le glucide
- 44 produits

- La somme des colonnes _100g ne doit pas passer de 100 g
- 317 produits

Supprimées

Remplacées par nan

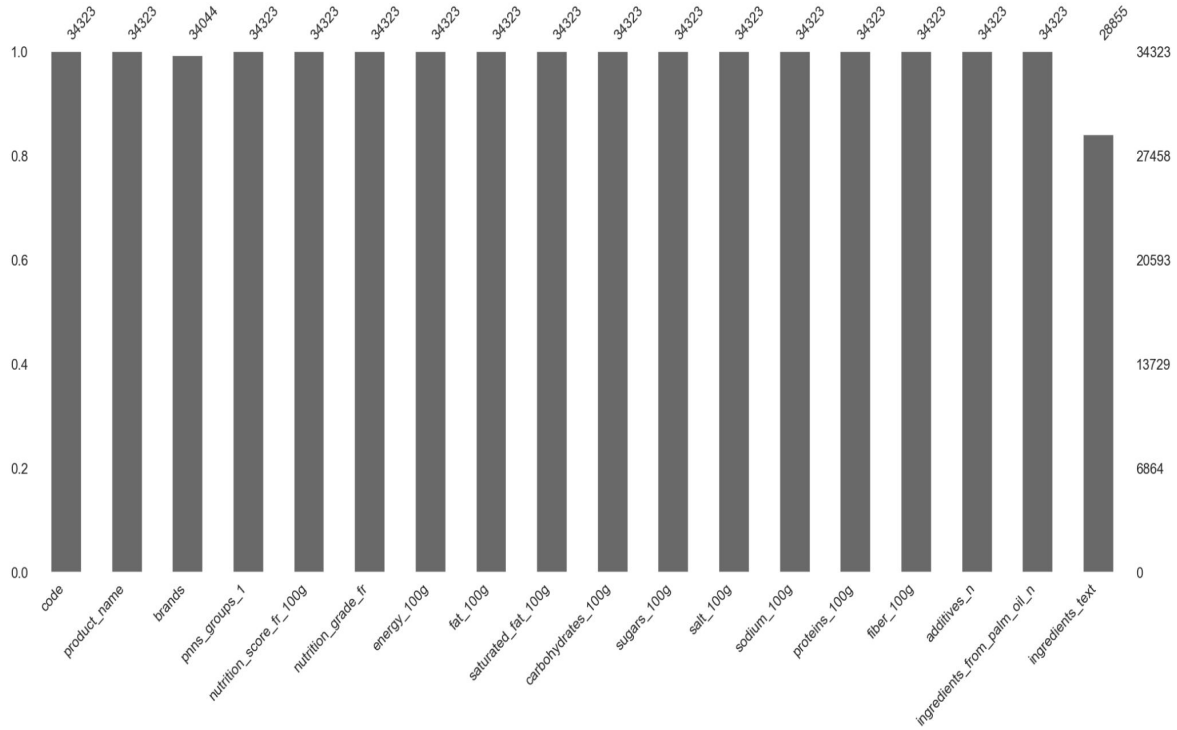
Santé publique France

Imputation des valeurs manquantes

01	Imputation avec zéro	• Si l'un de montant de produit ou la somme des valeurs est 100 g • On évite pour certains produits qu'ils passent le montant total 100
02	Imputation avec Iterative Imputer	• Si les colonnes sont corrélées entre eux • 'fat_100g' et' saturated_fat_100g' ---> 80 % • 'fat_100g' et 'energy_100g' ---> 77 % • On a testé Bayesian Ridge, Random Forest et Knn Imputer
03	Imputation avec médian de chaque catégorie de pnns groups 1	• S'il n'y pas de corrélation entre variables • protein, fiber • additives_n, ingredients_from_palm_oil_n
04	Imputation avec calcul de nutrition score et nutrition grade	• Pour imputation des variables catégoriques • Pour la colonne nutrition grade fr • On a testé le mode de chaque catégorie et calcul nutrition score

Santé publique France

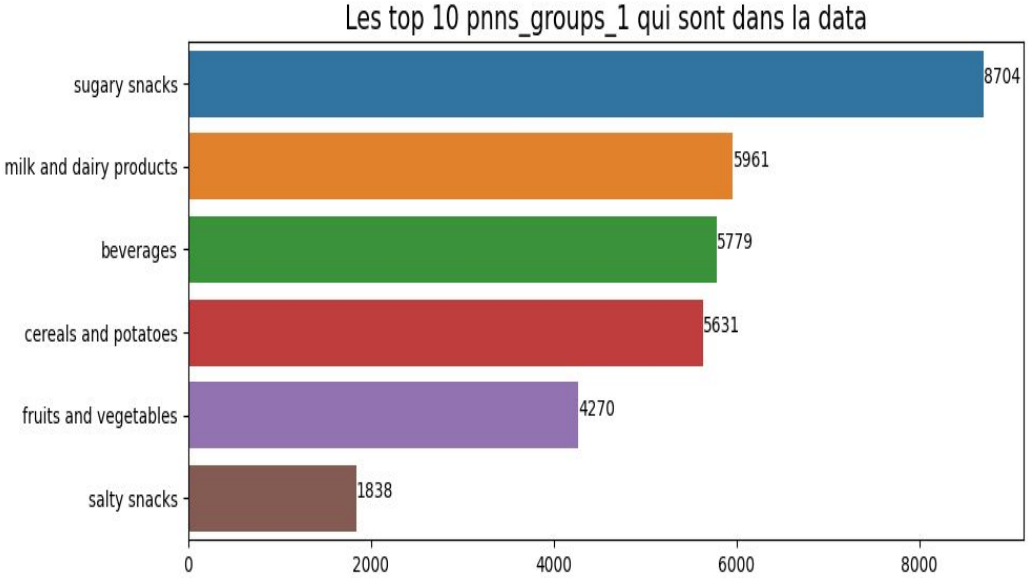
Après deuxième vérification des outliers



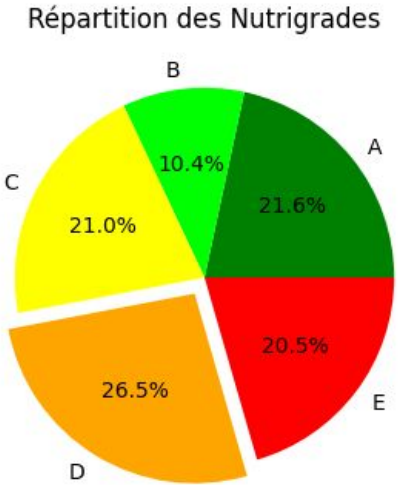
- On a supprimé les produits qui passent 100 g
- On a changé les autres valeurs avec leur max possible comme fat et saturated fat
- On a 32 000 lignes et 18 colonnes
- On a créé deux colonnes
    - additive : sans ou avec additive
    - palm oil : sans ou avec huile de palme

Santé publique France

Analyse Univariée



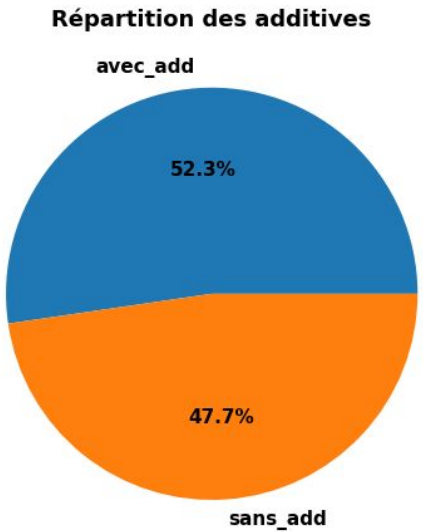
On a plus de produits de catégorie snacks sucrés

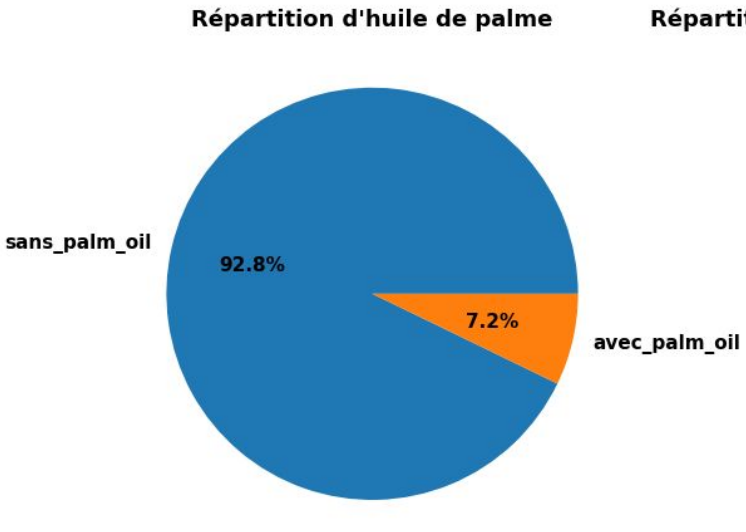


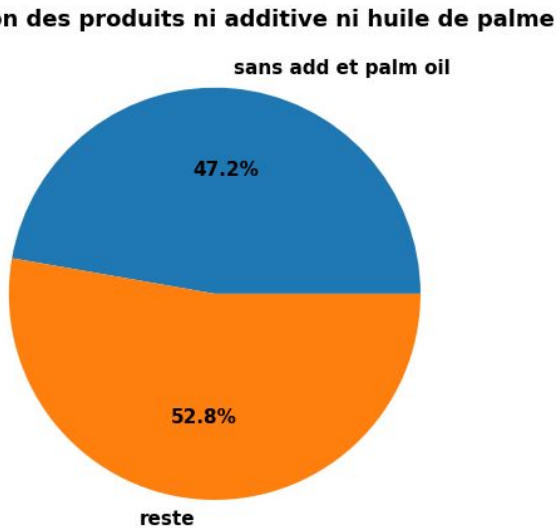
On a plus de produits de score D

Santé publique France

Analyse Univariée



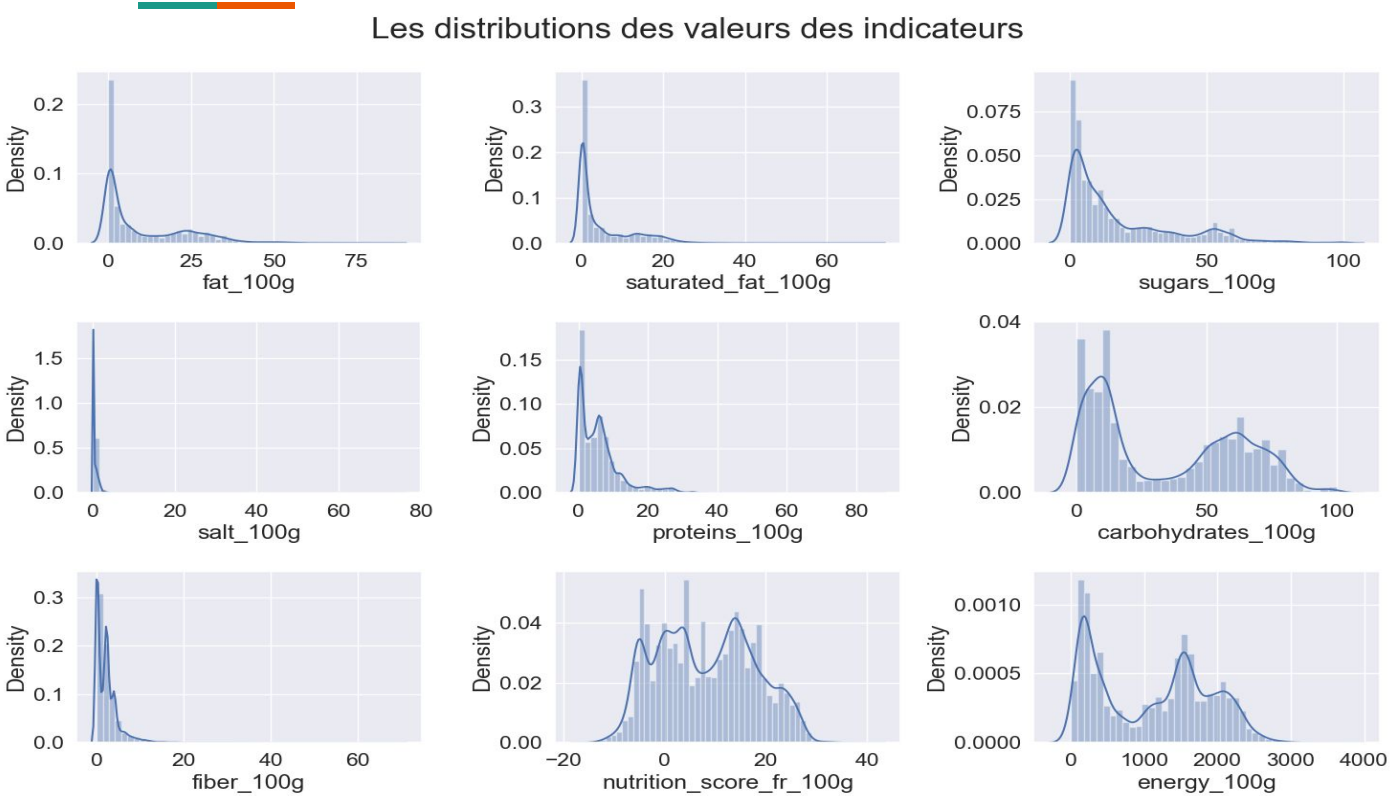




Environ 50 % des produits n'ont ni additifs ni ingrédient huile de palme

Santé publique France

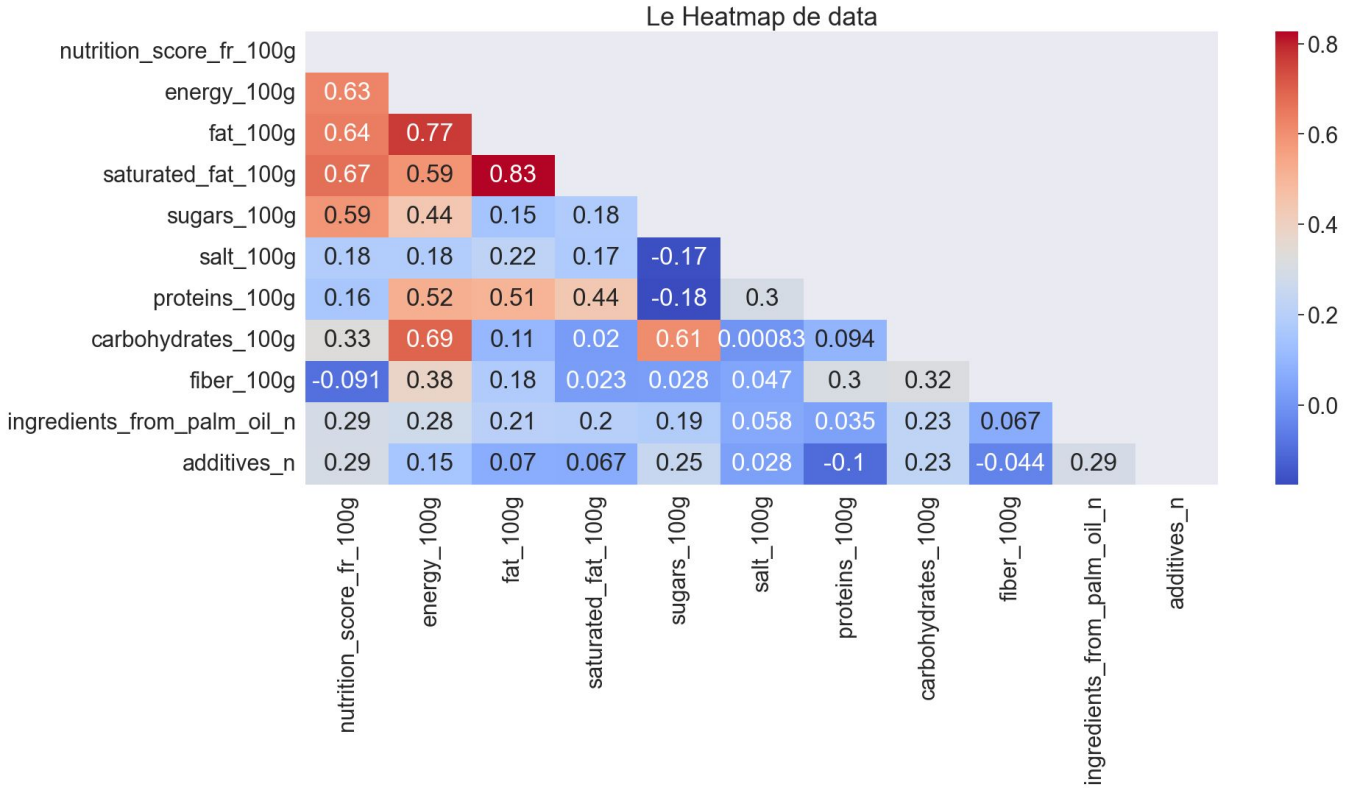
Analyse Univariée



- On a des distributions plus larges pour energy, carbohydrates et nutrition score.
- Les distributions ne suivent pas la loi normale.

Santé publique France

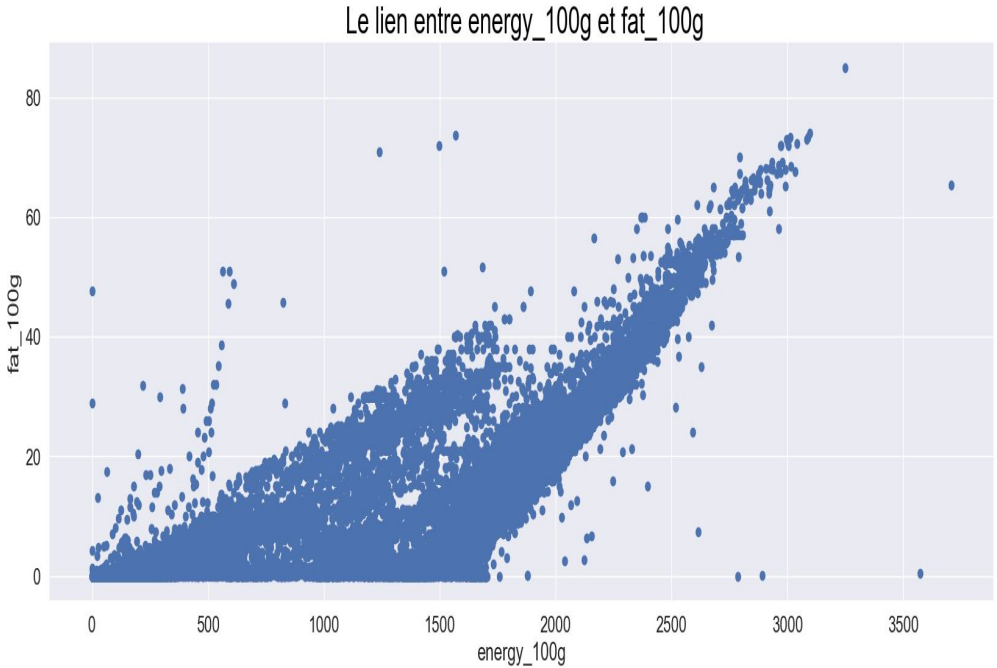
Analyse Bivariée



On a des corrélations entre fat et energy, fat et saturated fat etc.

Santé publique France

Le lien entre énergie et matière grasse - Tests Statistiques non Paramétriques



D'après le test Shapiro-Wilk, aucune de variable a une distribution normale. On va faire les tests non paramétriques.

Spearman's Rank Correlation Test

H0 Pas de corrélation entre deux variables

H1 Il y a une corrélation entre deux variables

alpha = 0.05

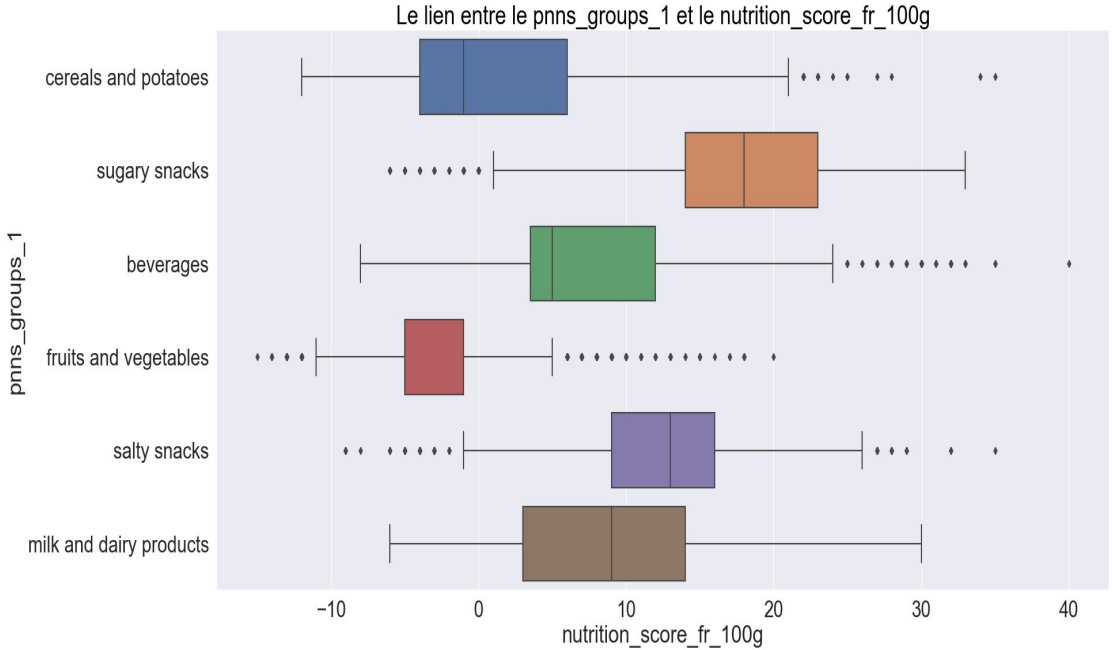
Spearmans correlation coefficient: 0.802
p_value= 0.0

p valeur est inférieur à 0.05

Il y a une corrélation entre deux variables

Santé publique France

Le lien entre le pnns groups 1 et le nutrition score



Test Kruskal Wallis

H0 = il n'y a pas de différence de nutrition_score entre les pnns_groups_1

H1 = il y a une différence de nutrition_score entre les pnns_groups_1

KruskalResult(statistic=14393.967262593873, pvalue=0.0)

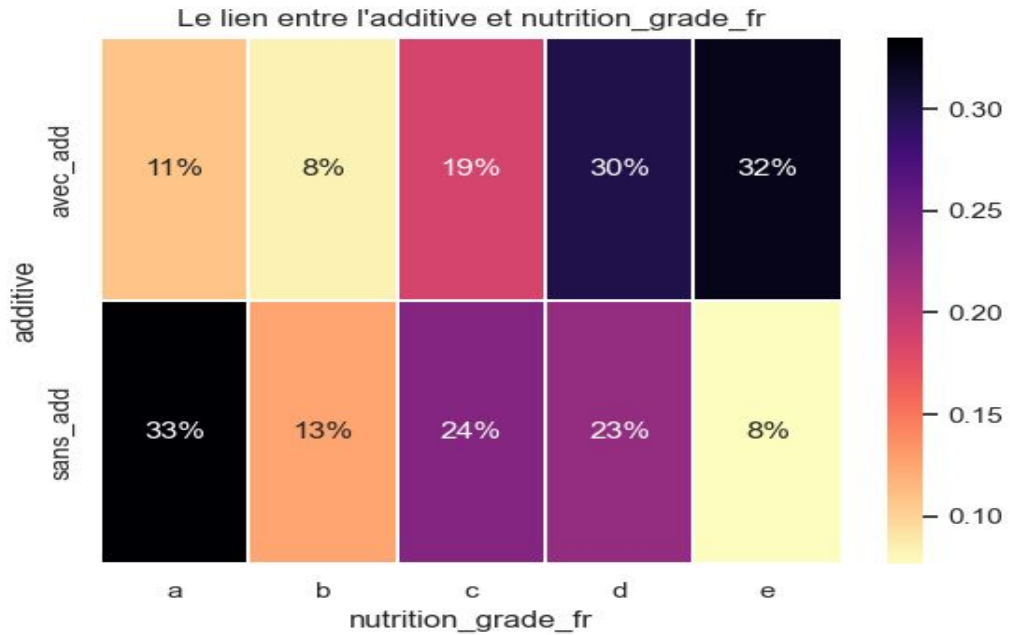
p valeur est inférieur à 0.05 et on rejette H0

Il y a une corrélation entre pnns_groups_1 et le nutrition_score_fr_100g

Le nutri score est meilleur pour les fruits et légumes

Santé publique France

Le lien entre l'additif et nutrition grade



On a le meilleur score -a- avec les produits sans additifs
On a le mauvais (score) -e- avec les produits avec additifs

Test Chi-2

H0= Ils sont indépendants additive et nutrition_grade_fr

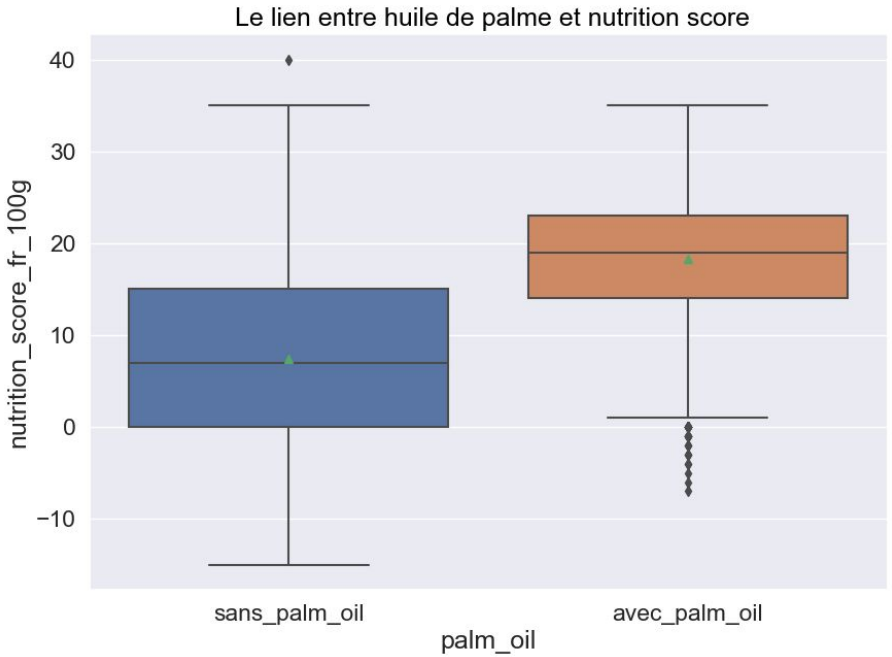
H1= Corrélation entre deux variables

p valeur est inférieur à 0.05 alors on rejette H0

Il y a une corrélation entre additive et nutrition_grade_fr

Santé publique France

Le lien entre huile de palme et nutrition score



Test de Mann-Whitney

H0= il n'y a pas de différence de nutri score moyen entre les produits sans ou avec huile de palme

H1= il y a une différence de nutri score moyen entre les produits sans ou avec huile de palme

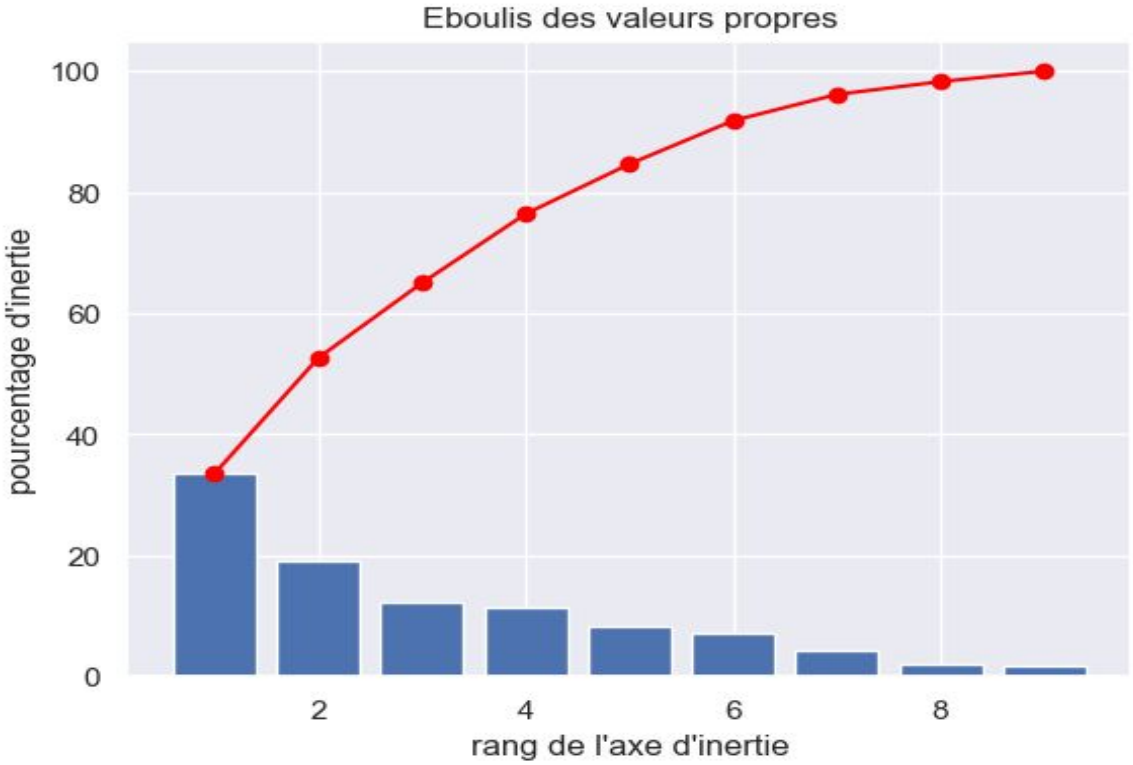
p valeur est inférieur à 0.05

Il y a de différence de nutri score moyen entre les produits sans ou avec huile de palme

On a des meilleurs scores avec les produits sans huile de palme

Santé publique France

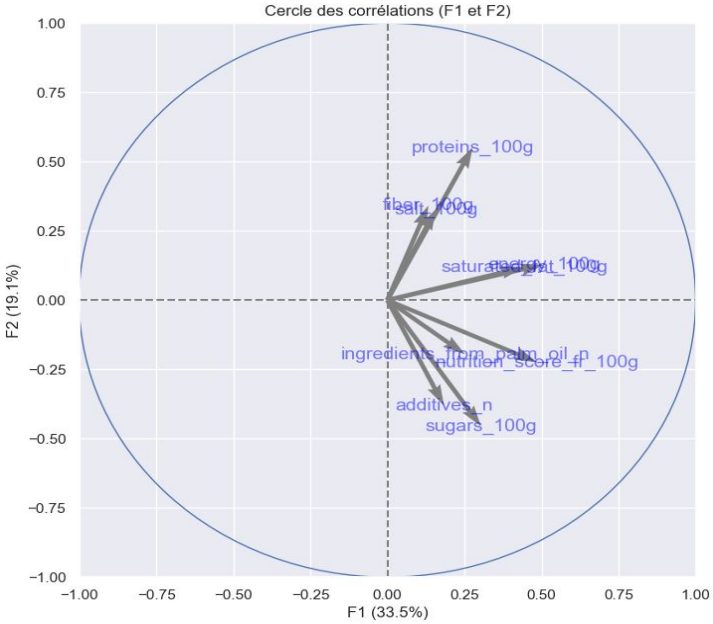
Analyse en Composantes Principales - ACP

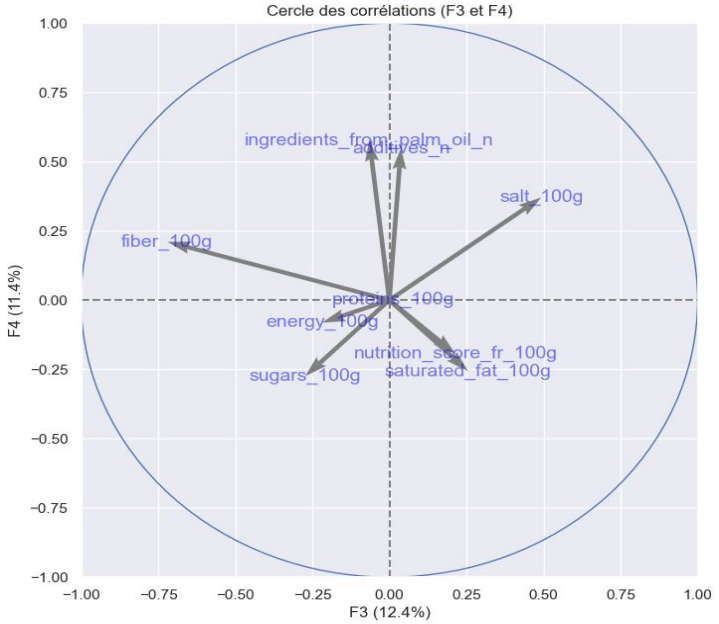


On peut réduire le dataset à 5 composantes avec près de 85 % des données.

Santé publique France

ACP - Les Cercles des Corrélations





F1 Les produits ayant plus matière grasse saturée et plus d'énergie

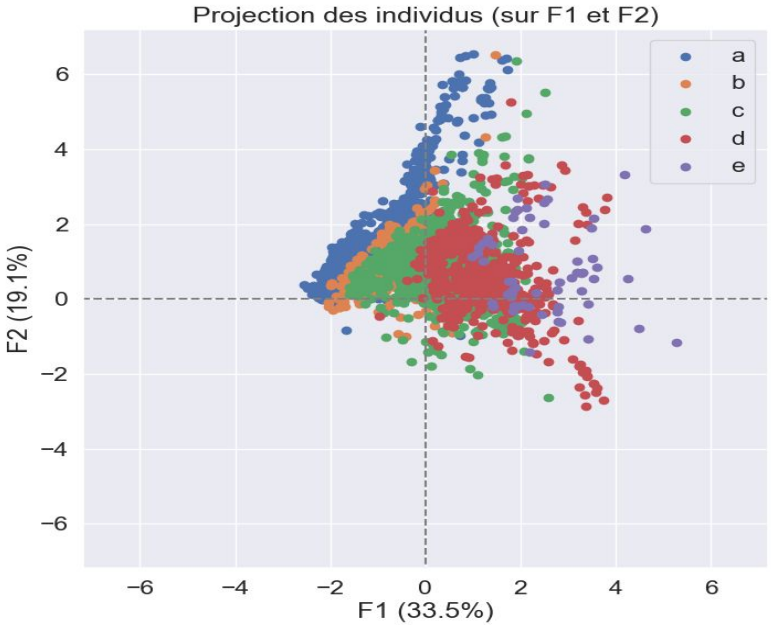
F2 Les produits riche en protéine et le sucre est négativement corrélée

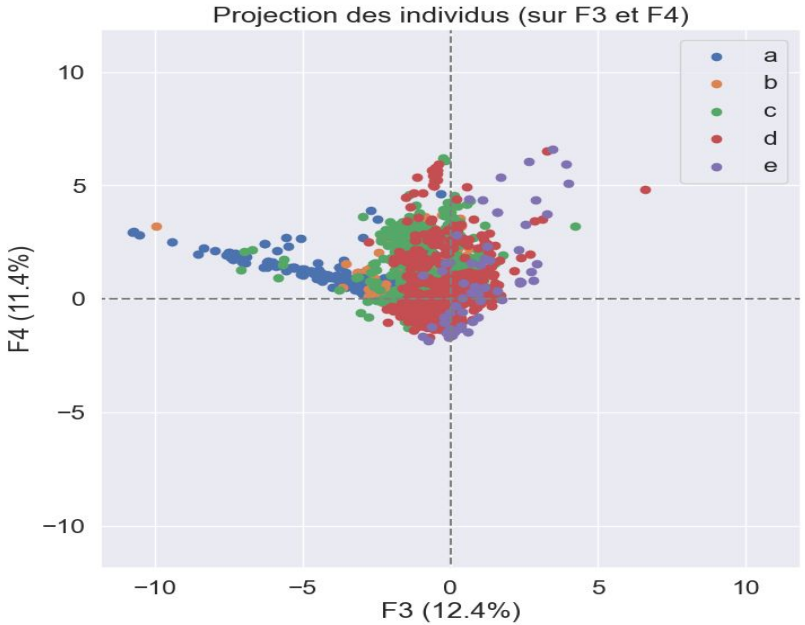
F3 Le fibre est négativement corrélé

F4 Les produits ayant plus d'additifs et huile de palme et sel

Santé publique France

ACP - Projection de la catégorie céréales et pomme de terre





F1 On voit la nutrition grade "a" à "e" et aussi ils ont plus d'énergie et de matière grasse.

F2 On voit les céréales ayant plus de protéine avec nutrigrade "a" et plus sucrés avec nutri grade "d" et "e".

F3 On voit les céréales ayant plus de fibres avec nutrigrade "a".

F4 On voit les produits ayant plus d'additifs et huile de palme et sel avec nutri grade "c", "d" et "e".

Santé publique France

Faisabilité de l'application

- On a constaté qu'il y a bien un lien entre énergie et matière grasse ainsi que la nutrition score et les nombres additifs et les ingrédients d'huile de palme.
- Environ la moitié de nos produits n'ont ni additive ni huile de palme.
- On peut bien réaliser un score goûter pour trier les meilleurs produits.
- On peut aussi recommander des produits dans la même catégorie selon le besoin énergétique de l'enfant.

Santé publique France

Conclusion et Recommandations

- Même si on a nettoyé le dataset, on peut toujours avoir des incohérences parce qu'il a été créé par les contributeurs publics.
- On peut améliorer l'imputation des données manquantes avec les méthodes machine learning plus sophistiquées.
- On peut aussi ajouter une variable Nova qui donne l'information de situation transformée des produits.
- Malgré tous nos efforts, il n'y a pas un produit qui peut remplacer le fait maison. Alors, on peut encourager les parents à donner des produits fait maison à leurs enfants pour leur goûter.

Santé publique France

MERCI

Santé publique France