

# **Concevoir une application au service de la santé publique**

Agence 'Santé publique France' – Appel à projets

# Introduction

**Base de données : OpenFoodFacts**

**Problématiques :**

- Quelle application allons nous concevoir ?**
- Quelles sont les variables pertinentes ? Comment synthétiser leurs comportements ?**
- Quelles analyses multivariées pourront confirmer ou infirmer nos hypothèses ?**
- Comment visualiser les analyses ?**

# La base de données - OpenFoodFacts

- **Contient 186 informations pour chaque produit. Dont :**

- Sa composition nutritionnelle
- Son nutri-score, son éco-score
- Sa catégorie

- **Contient ~2 millions de produits**

Tout pays confondu



# L'application

**- Etant donné un aliment, l'application doit pouvoir proposer des alternatives à celui-ci :**

- Meilleurs d'un point de vue nutritif, mais proche
- D'une même catégorie

# I- Le nettoyage des données

- 1- Préliminaires (doublons, colonnes...)**
- 2- Sélection des colonnes**
- 3- Suppression des valeurs aberrantes**
- 4- Imputation des valeurs manquantes**



# I- 1- Nettoyage préliminaire

- **Suppression des doublons (270)**
  - **Correction des erreurs de casse**
  - **Correction des noms des colonnes**
  - **Sélection des produits qui ont :**
    - Un code barre
    - Une catégorie principale en anglais
- 1908859 → 812217 produits**



SNacks  
→ Snacks

-sugars\_100g  
→ sugars\_100g

## I- 2- Sélection des colonnes

- **On demande au minimum 30 % de remplissage au départ**

186 → 131 colonnes

- **Parmi ces colonnes on garde celles-ci :**

### → **Quantités**

energy-kcal, fat, saturated-fat, carbohydrates, sugars, fiber, proteins, salt, sodium, nutrition-score-fr, additives\_n, ingredients\_from\_palm\_oil\_n, ecoscore\_score\_fr, serving\_quantity, nova\_group

### → **Qualités**

main\_category\_en, pnns\_groups\_2, pnns\_groups\_1, additives\_en

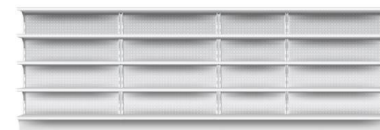
# I- 3-Suppression des valeurs aberrantes

- Strings contenant des caractères non-latins
- Valeurs nutritionnelles entre 0 et 100g
- Cohérence entre énergie et composants
- Limite de poids
- Appartenance à aucune catégorie
- Bornes : Nutri-score, Eco-score, Groupe Nova

200%

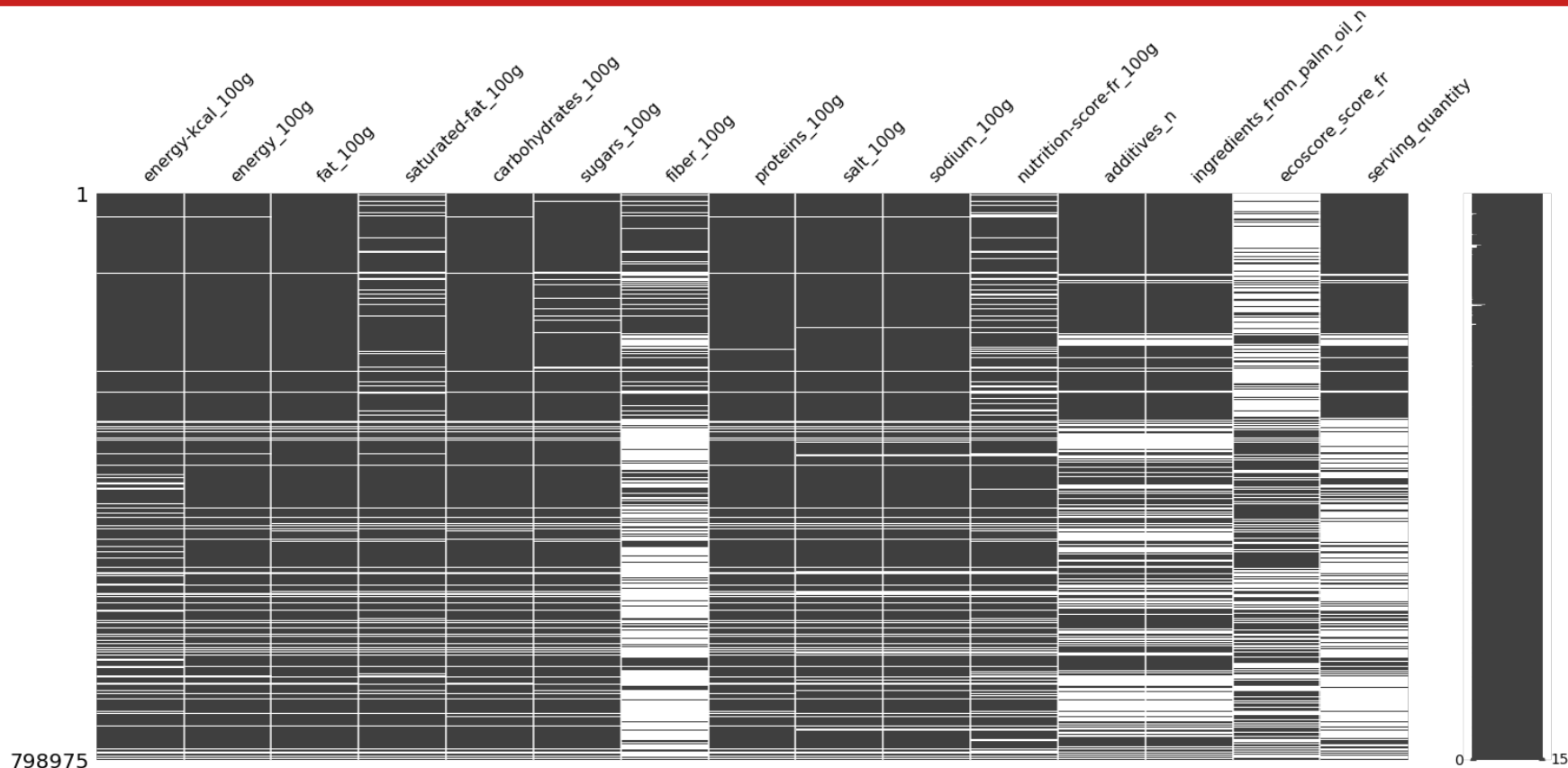


藥





# I- 4- Imputation des valeurs manquantes



## **I- 4- Imputation : Méthodes**

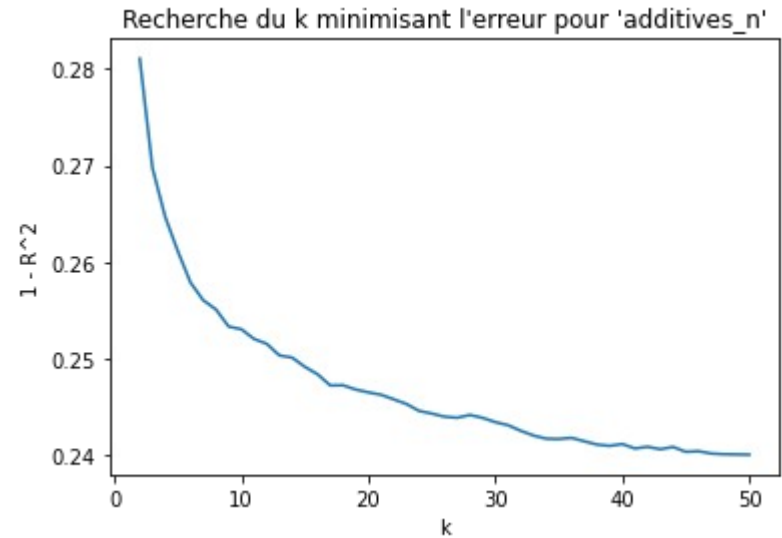
- Régression avec arbre de décision**
- Régression linéaire bayésienne**
- Régression/classification k-nn**
- Moyennes par catégorie**

## I- 4- Imputation : additives\_n

Taille de l'échantillon : **411997**

Erreur par méthode :

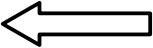
- BayesianRidge : **0.19** ←
- DecisionTreeRegressor : **0.28**
- Means : **0.61**
- k-nn : **0.23**



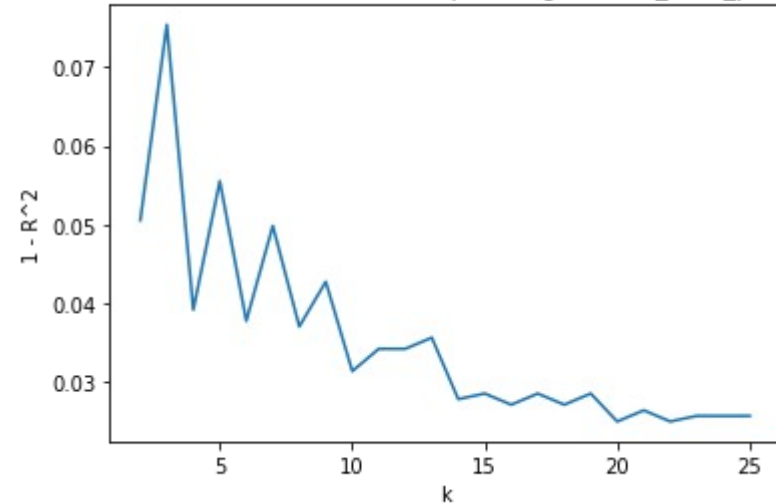
## I- 4- Imputation : ingredients\_from\_palm\_oil

Taille de l'échantillon : **411997**

Erreur par méthode :

- BayesianRidge : **0.005** 
- DecisionTreeClassifier: **0.13**
- Means : **1.0**
- k-nn : **0.02**

Recherche du k minimisant l'erreur pour 'ingredients\_from\_palm\_oil\_n'

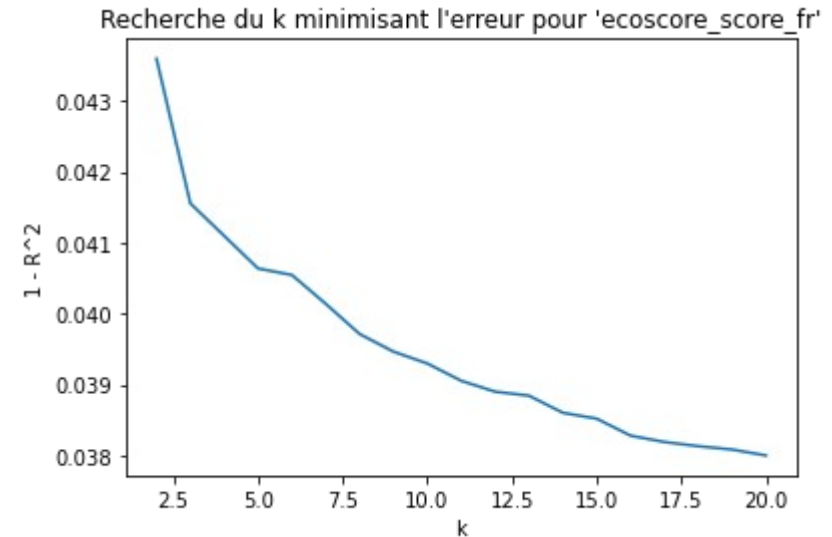


## I- 4- Imputation : ecoscore

Taille de l'échantillon : **293540**

Erreur par méthode :

- BayesianRidge : **0.02**
- DecisionTreeRegressor : **0.04**
- Means : **0.12**
- k-nn : **0.03** ←

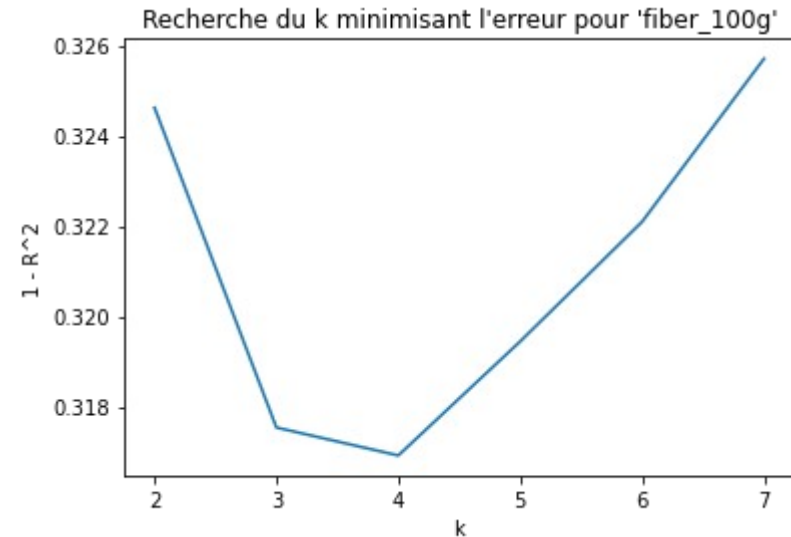


## I- 4- Imputation : fiber

Taille de l'échantillon : **338798**

Erreur par méthode :

- BayesianRidge : **0.7**
- DecisionTreeRegressor : **0.41**
- Means : **0.58**
- k-nn : **0.32** ←



## I- 4- Imputation : serving\_quantity

Taille de l'échantillon : **311328**

Erreur par méthode :

- BayesianRidge : **0.72**
- DecisionTreeRegressor : **0.77**
- Means : **0.47**
- k-nn : **0.45** ←

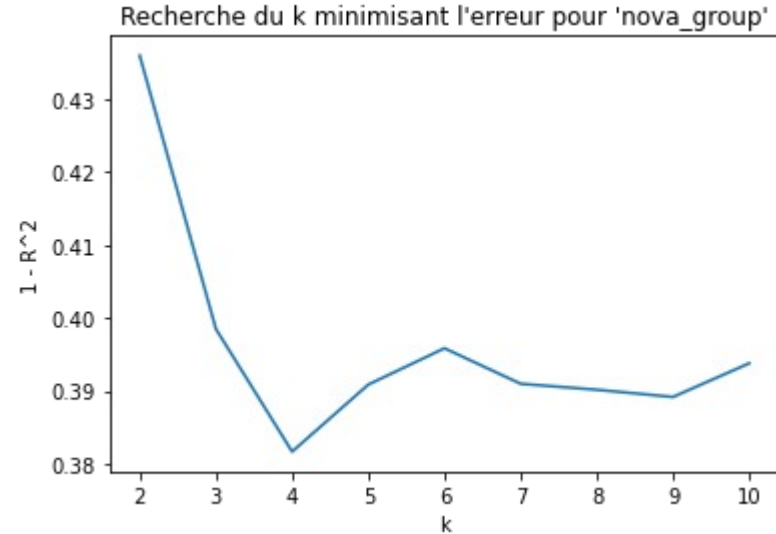


## I- 4- Imputation : nova\_group

Taille de l'échantillon : **388670**

Erreur par méthode :

- BayesianRidge : **0.72**
- DecisionTreeClassifier: **0.22**
- Means : **0.6**
- k-nn : **0.38** ←





# I- Nettoyage - Conclusion

**Sans imputation : 70299 individus**

**Avec imputation : 590357 individus**

## **II- Exploration**

- 1 - Analyses uni-variées**
- 2 - Analyses des corrélations 2 à 2**
- 3 - Analyses des Composantes Principales**
- 4 - Corrélations avec les variables qualitatives**

## II- 1- Analyses uni-variées

- **Diagrammes en boîtes**
- **Modélisation de la distribution**

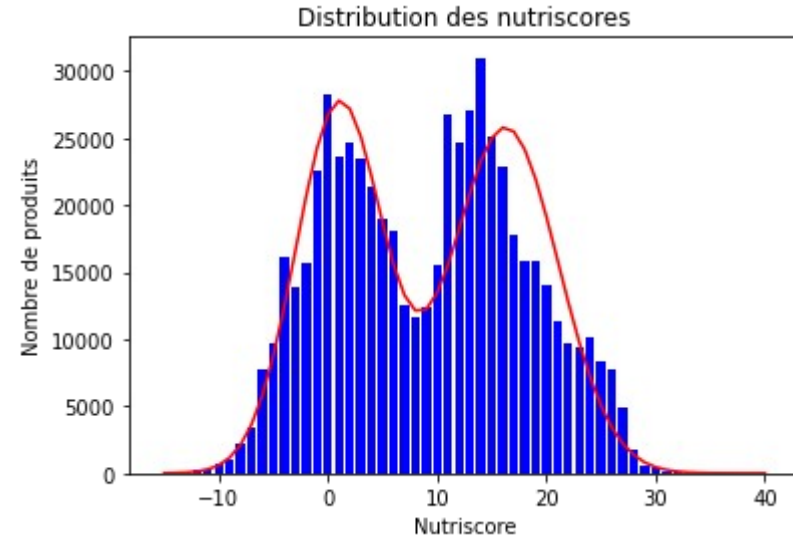
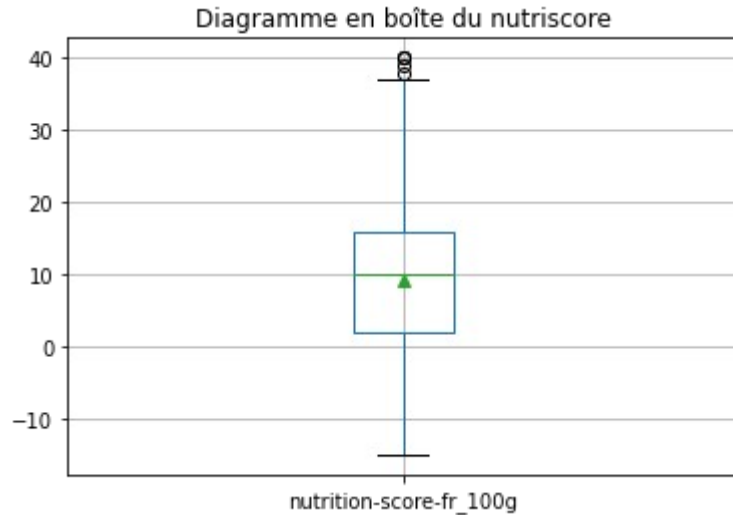
H0 : Le modèle et l'observation suivent la même distribution

H1 : Le modèle et l'observation suivent des distribution différentes

On accepte un risque maximal de 10 %

- **Concentration**

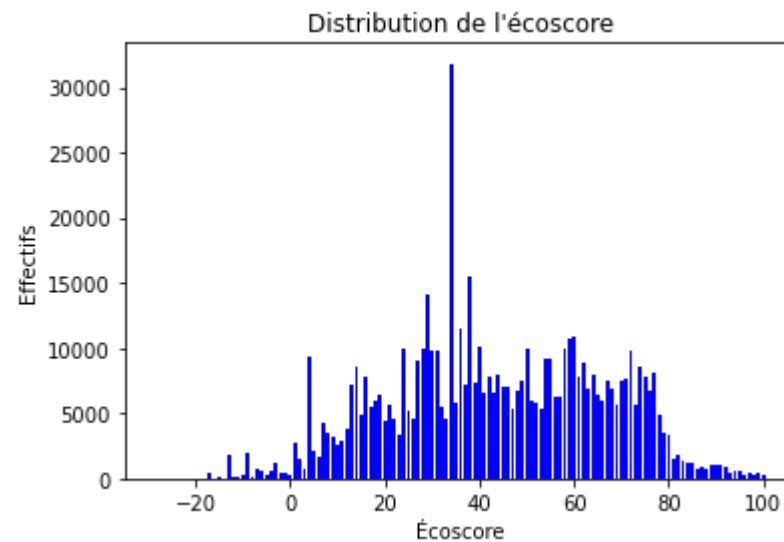
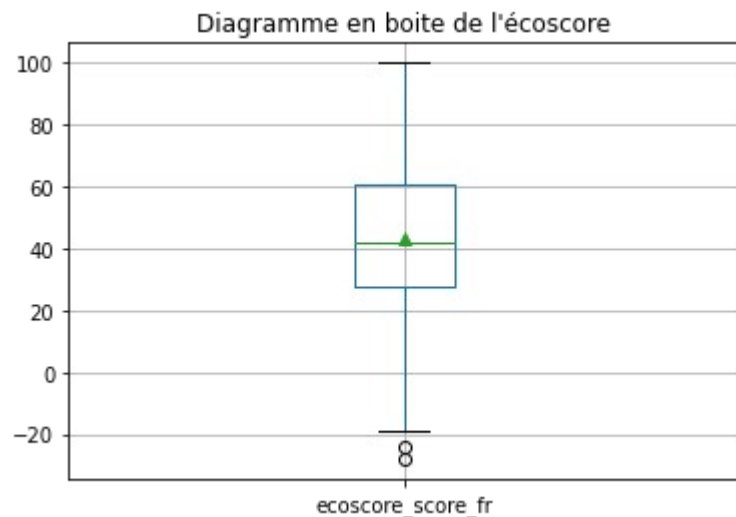
# I- 1- Étude univariée : Le nutriscore



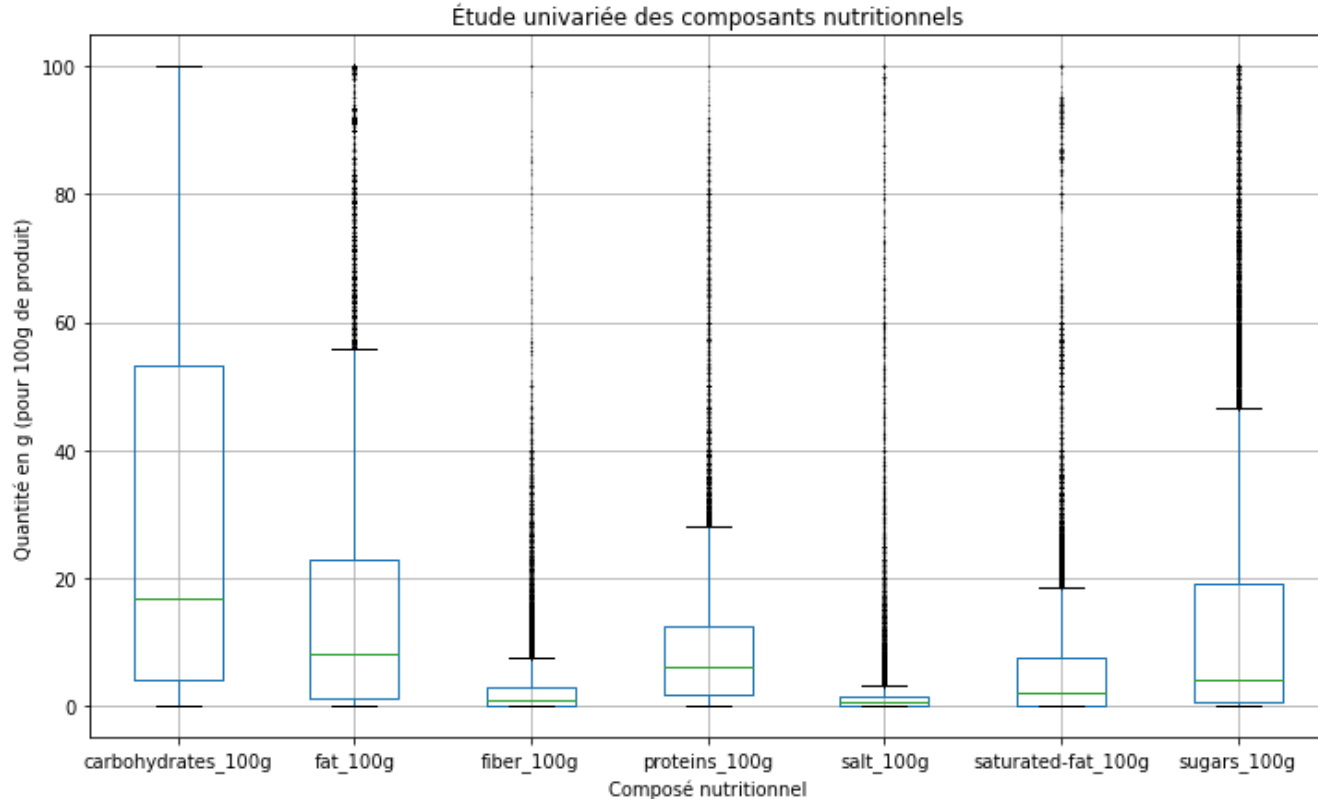
$R^2 = 0.88$

ks\_2samp :  $p=0.99$

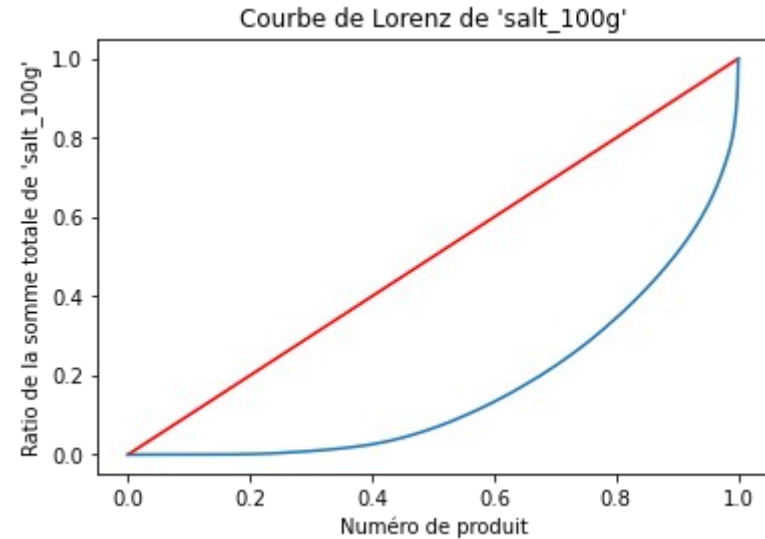
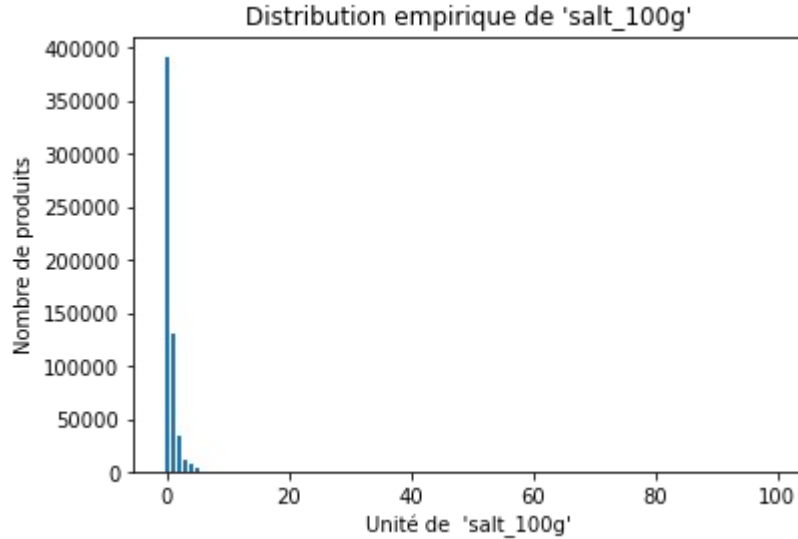
# I- 1- Étude univariée : Écoscore



# I- 1- Étude univariée : Valeurs nutritionnelles



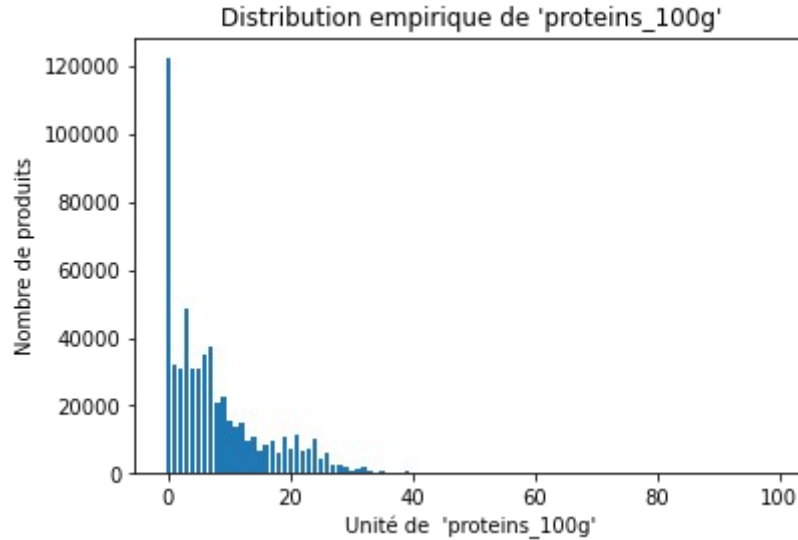
# I- 1- Étude univariée - le sel



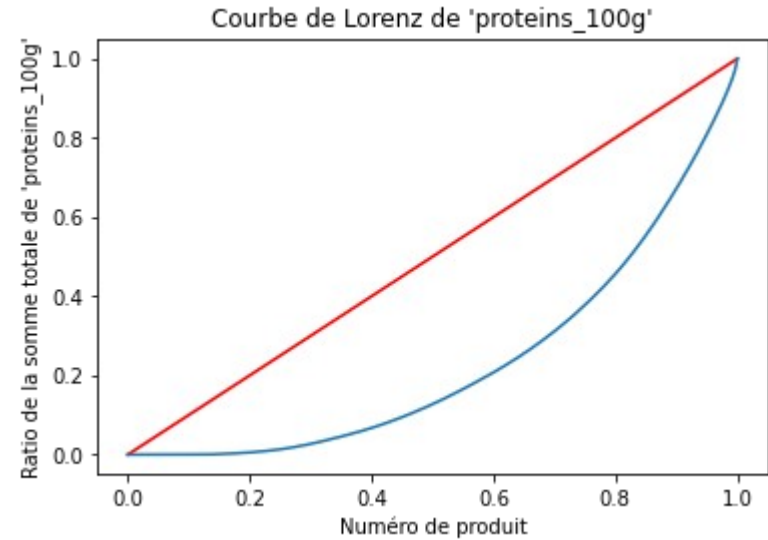
Modèle : Fonction inverse  
 $R^2 = 0.91$   
 $P = [3.79e-66, 9.02e-53]$

Gini = 0.67

# I- 1- Étude univariée - les protéines



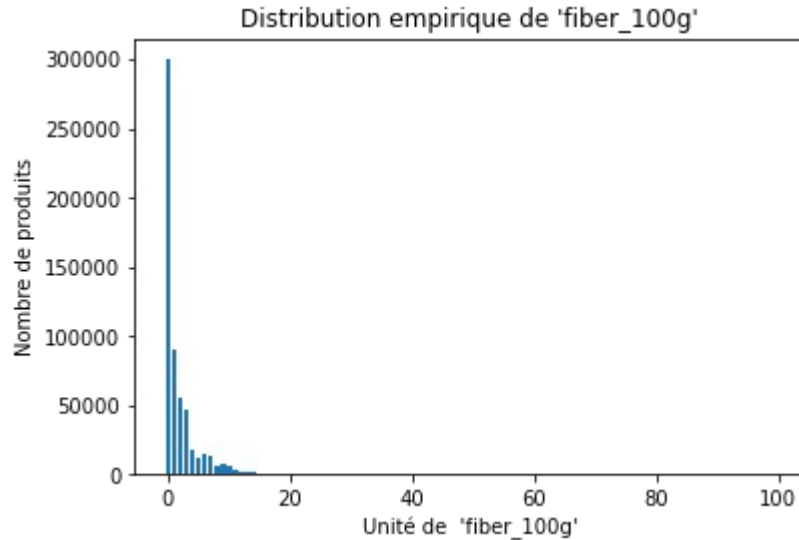
Modèle : Loi géométrique  
 $R^2 = 0.9$   
 $P = [6.08e-77, 1.67e-51]$



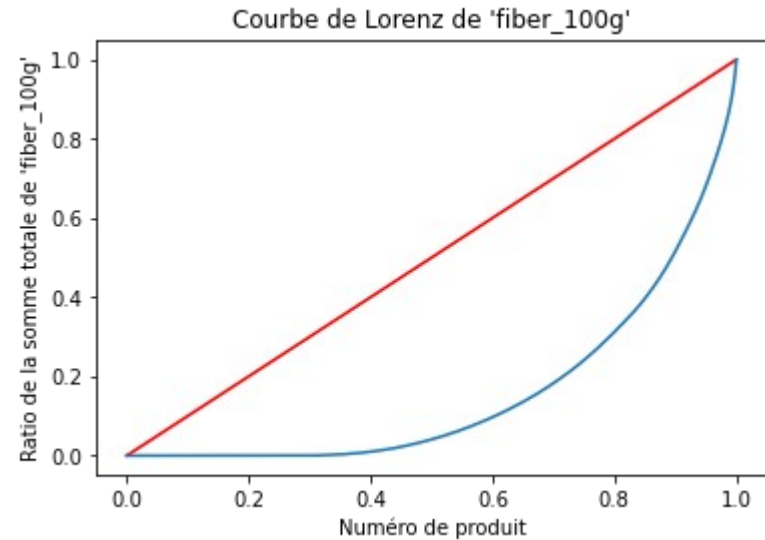
Gini = 0.53



# I- 1- Étude univariée - les fibres alimentaires

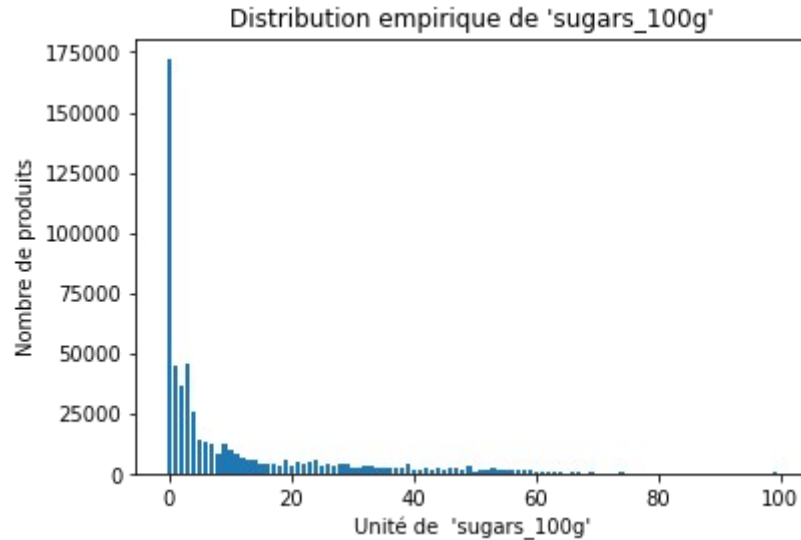


Modèle : Fonction inverse  
 $R^2 = 0.93$   
 $P = [1.5e-67 \ 2.25e-58]$

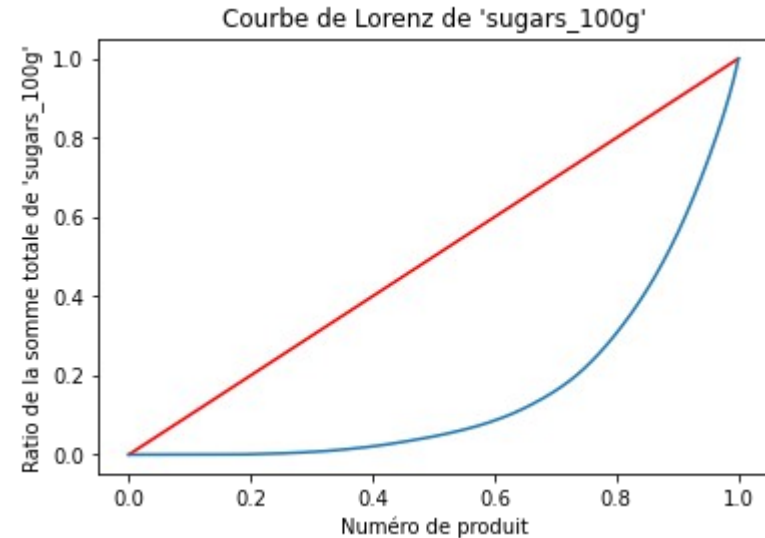


Gini = 0.67

# I- 1- Étude univariée - les sucres

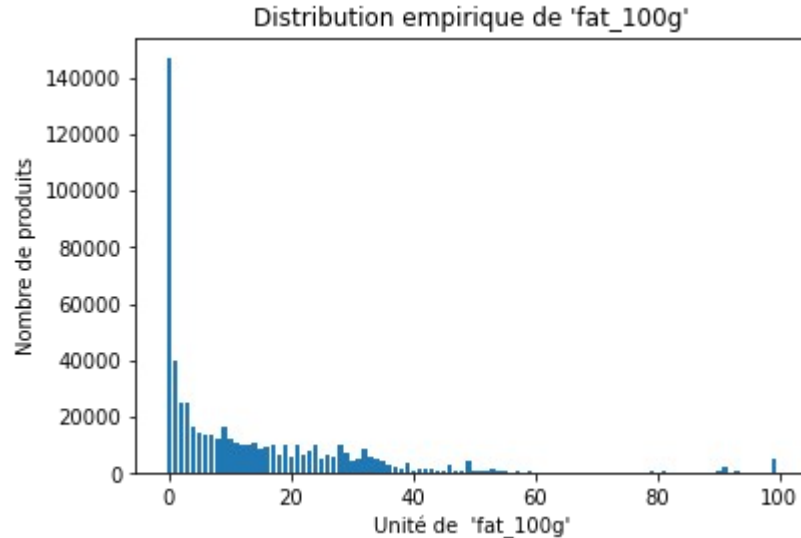


Modèle : Fonction inverse  
 $R^2 = 0.8$   
 $P = [3.54e-66 \ 1.58e-35]$

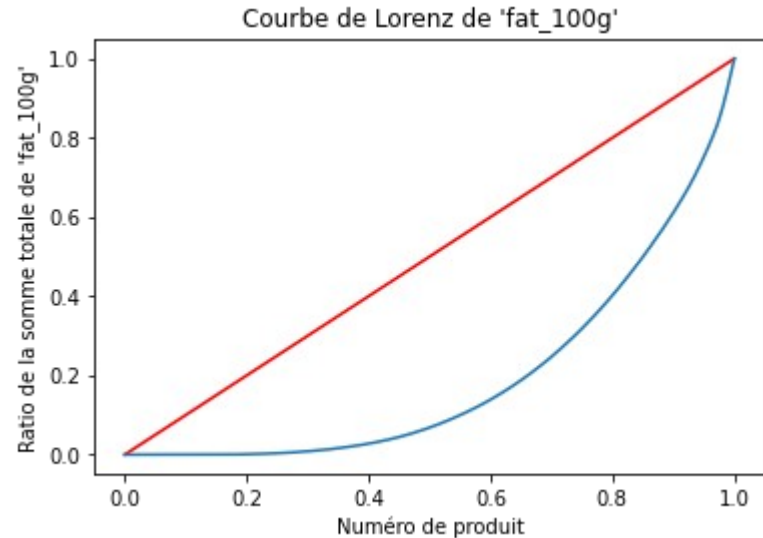


Gini = 0.67

# I- 1- Étude univariée - Matières grasses

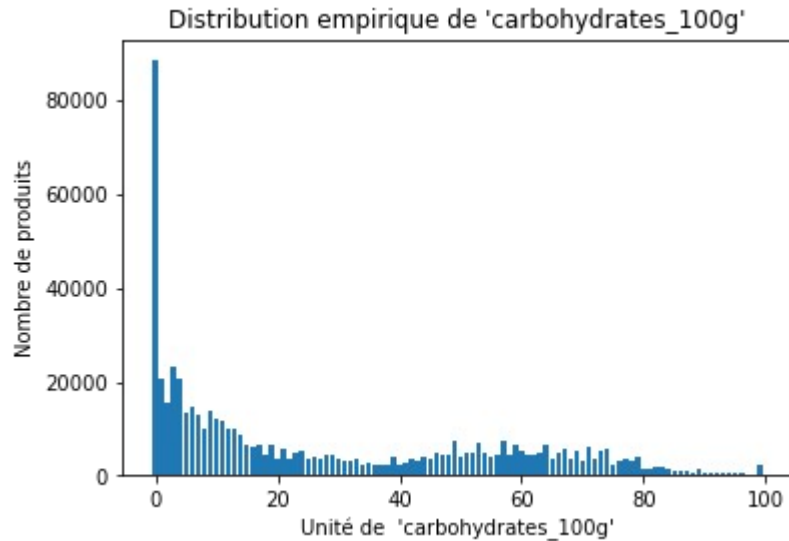


Modèle : Loi géométrique  
 $R^2 = 0.79$   
 $P = [5.78e-76 \ 3.58e-35]$



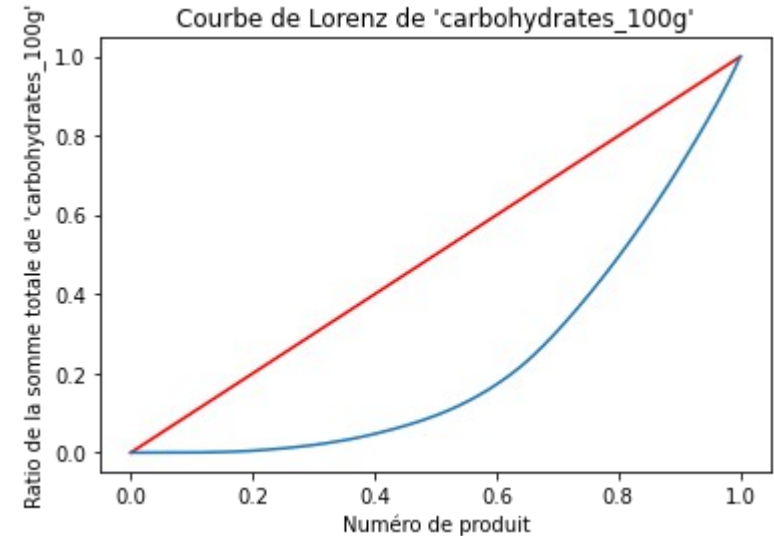
Gini = 0.61

# I- 1- Étude univariée - Carbohydrates



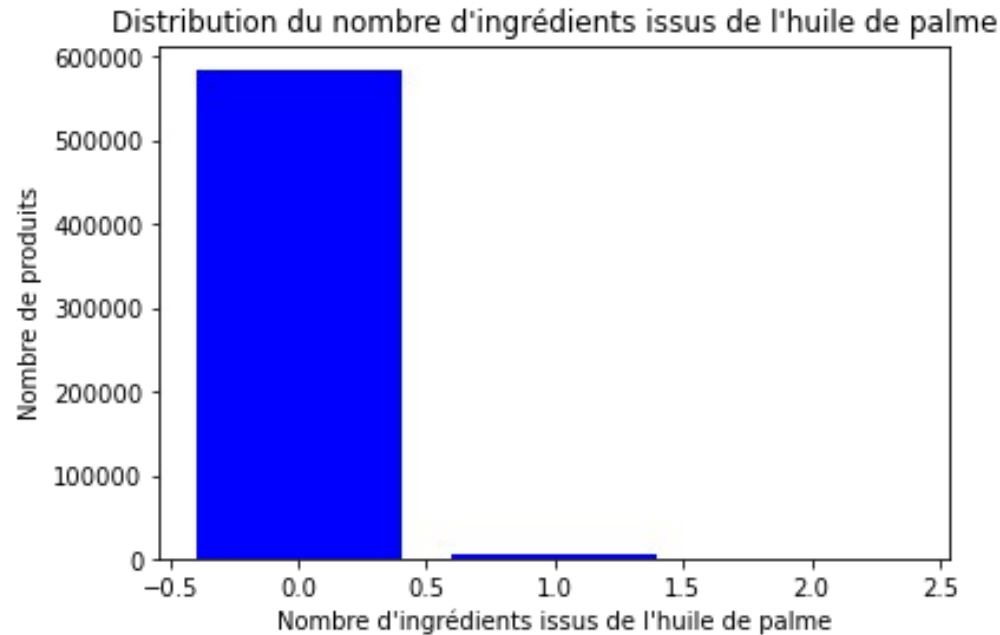
Modèle : Loi inverse  
 $R^2 = 0.9$   
 $P = [7.28e-45 \ 5.97e-21]$

Modèle : Loi normale  
 $P = 0.42$

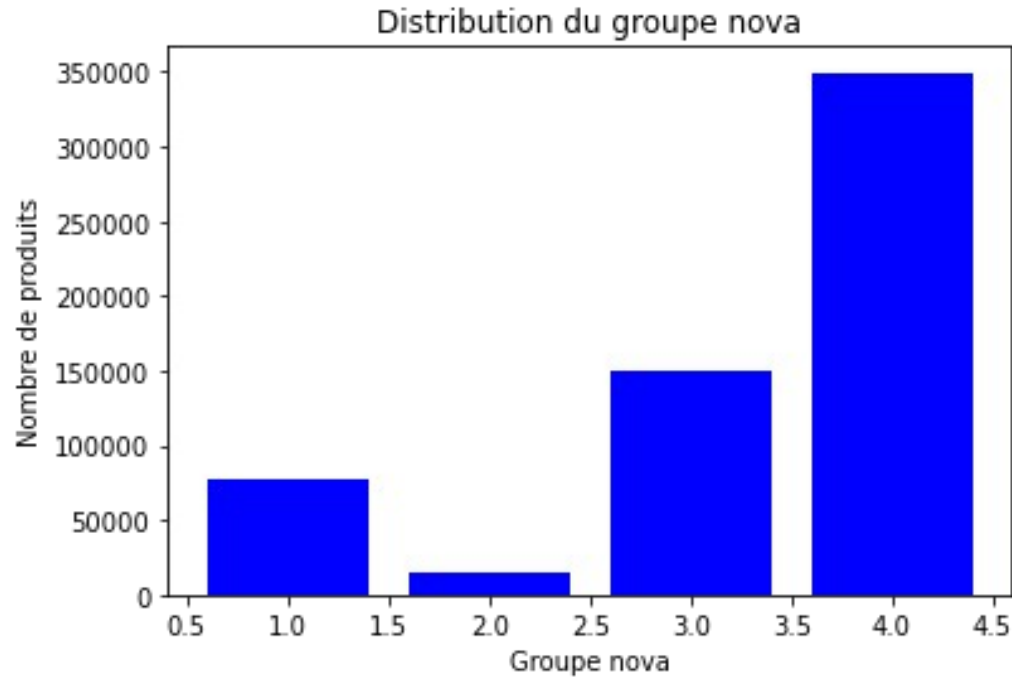


Gini = 0.53

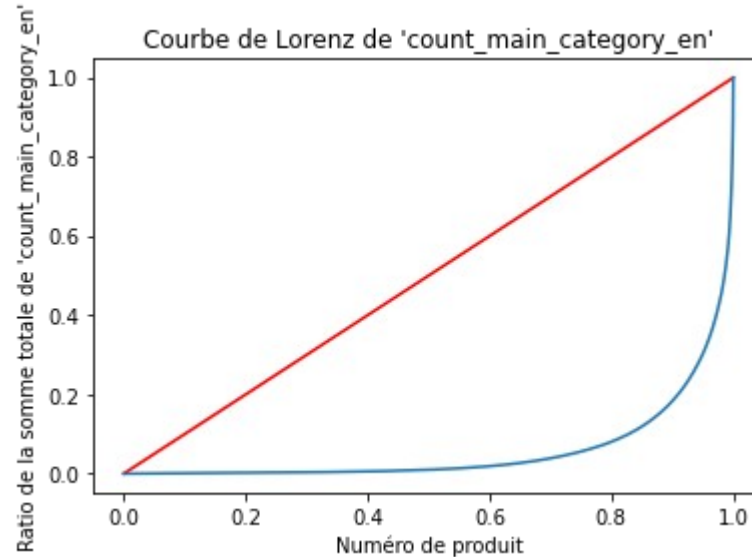
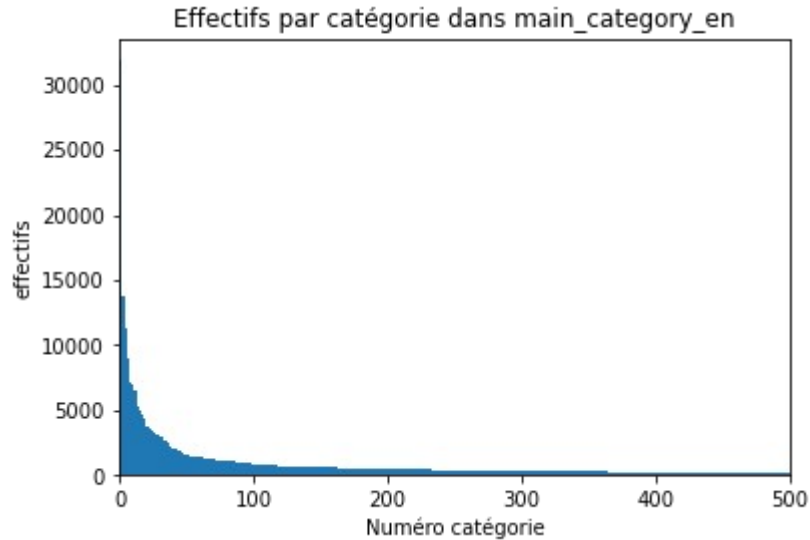
# I- 1- Étude univariée - Ingrédients de l'huile de palme



# I- 1- Étude univariée - Groupe Nova



# I- 1- Étude univariée - Catégorie principale



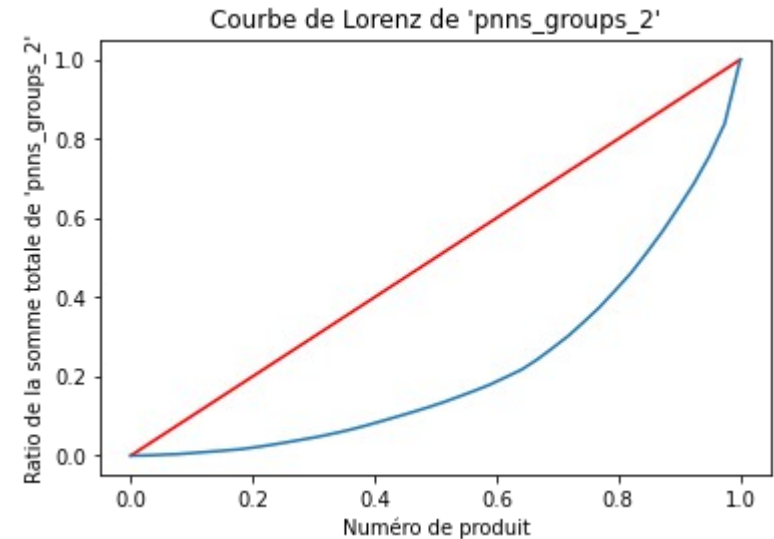
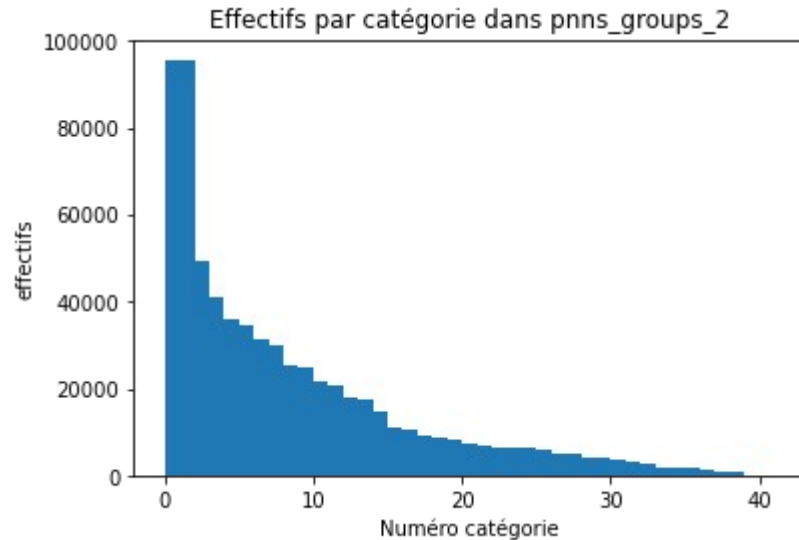
Modèle : Fonction inverse

$R^2 = 0.93$

$P = [0.0, 0.0]$

Gini=0.88

# I- 1- Étude univariée - Catégorie 2

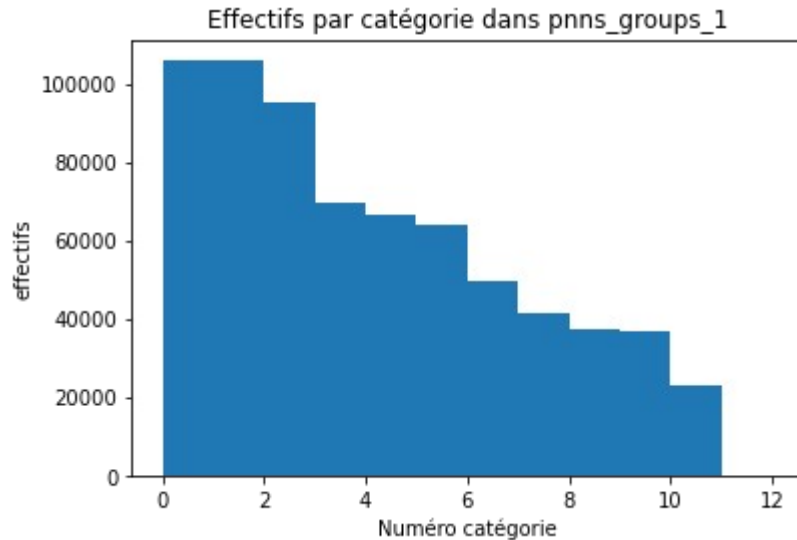


Modèle : Fonction inverse    Modèle : Droite  
 $R^2 = 0.94$      $R^2 = 0.98$   
 $P = [2.75e-21, 3.01e-09]$      $P = [2.09e-25, 2.2e-22]$

Gini=0.56



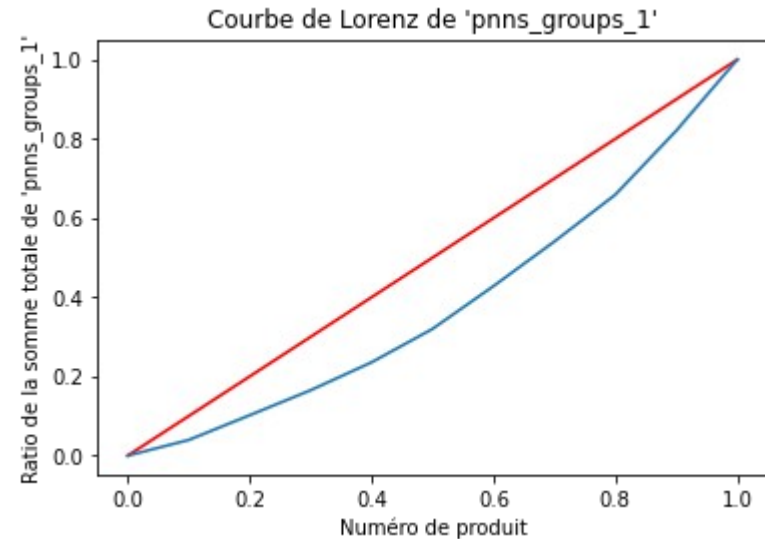
# I- 1- Étude univariée - Catégorie 1



Modèle : Droite

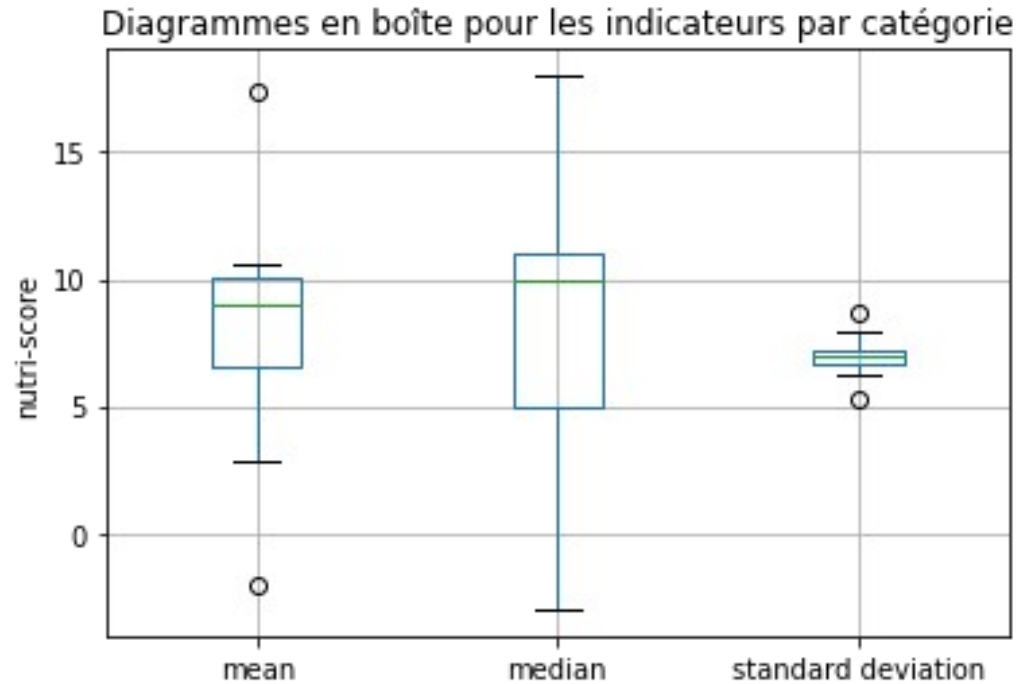
$R^2 = 0.95$

$P = [2.44e-09, 3.40e-07]$



Gini=0.31

# I- 1- Étude univariée - Nutriscore par catégorie



# I - 2- Analyses des corrélations 2 à 2

## - Sel - Sodium

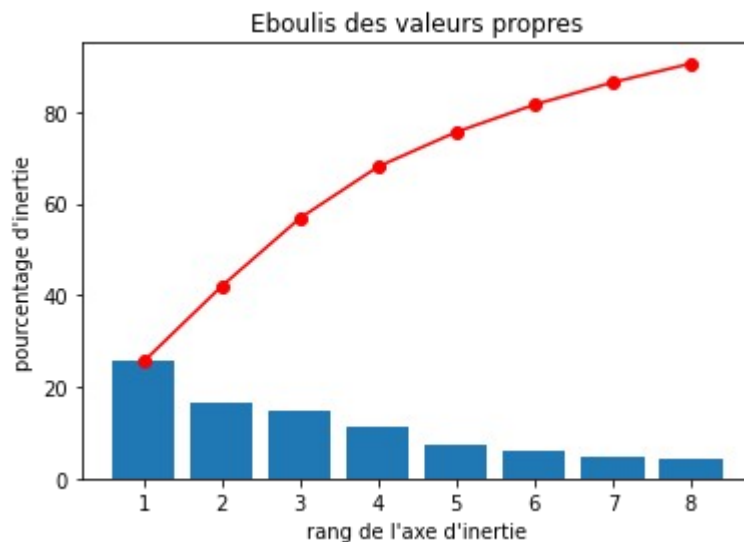
Pearson = 0.999995

## - (gras, sucre, protéïne, fibre, carbohydrates) - Énergie

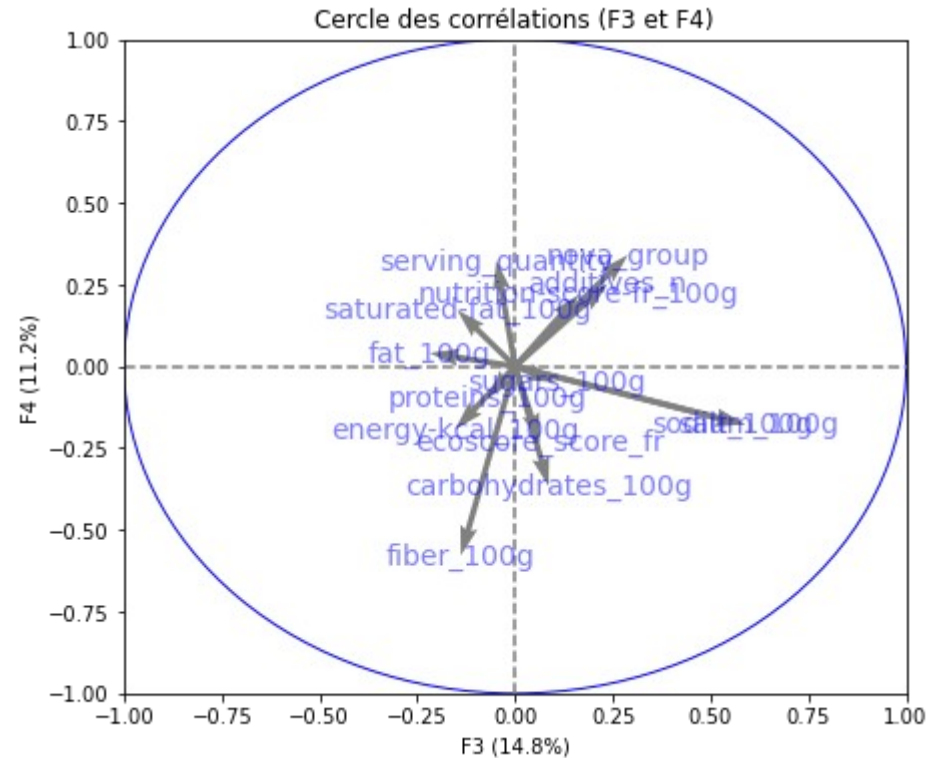
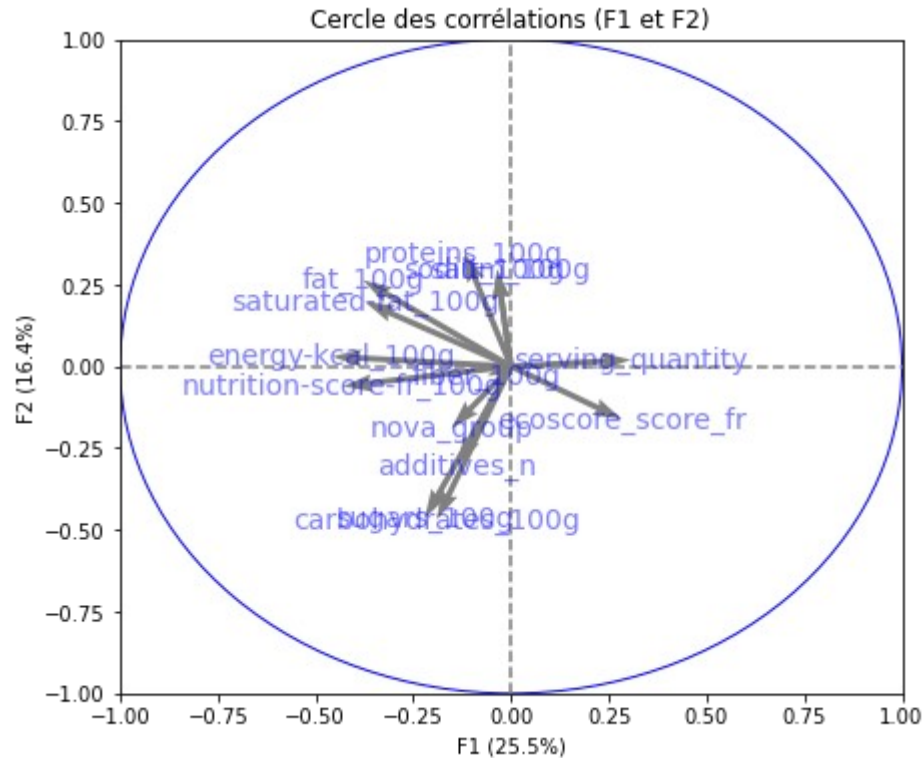
Pearson = 0.93

# I- 3- Analyse des composantes principales

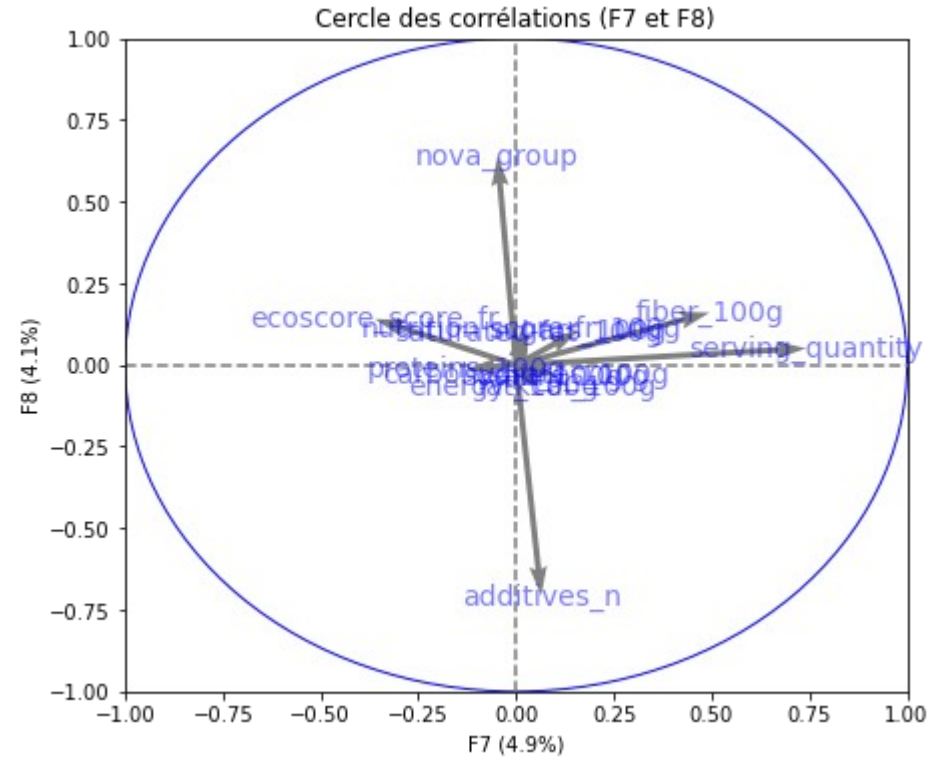
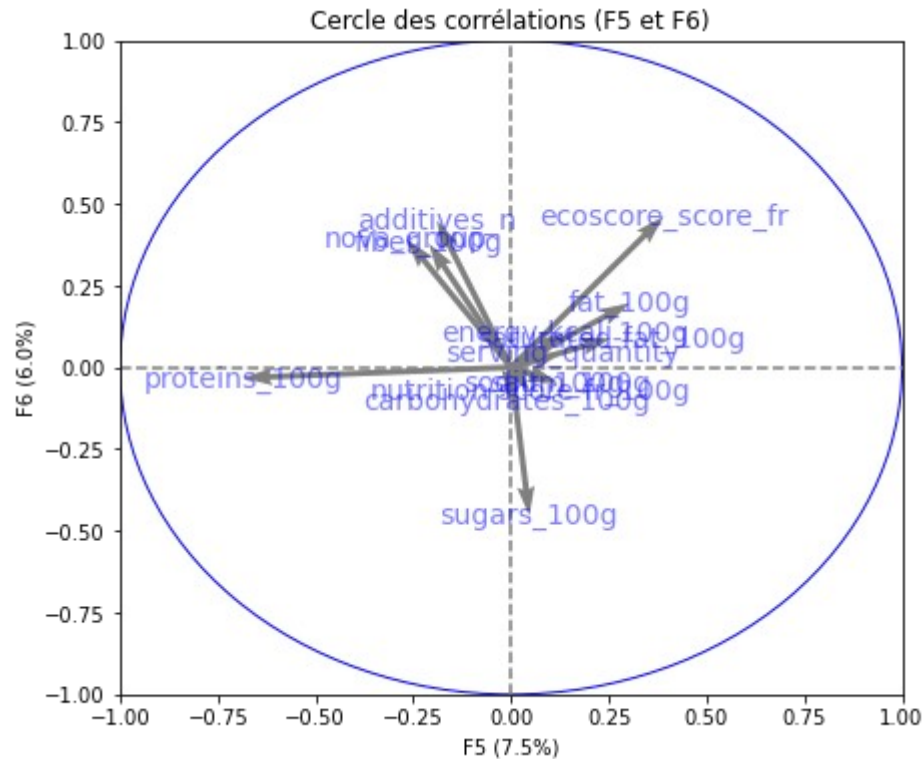
	3	4	5	6	7	8	9	10	11	12	13	14
Pourcentage de l'inertie totale	56.81704	68.053607	75.597792	81.564692	86.418153	90.531764	93.730955	96.637903	98.331556	99.783819	99.999961	100.0



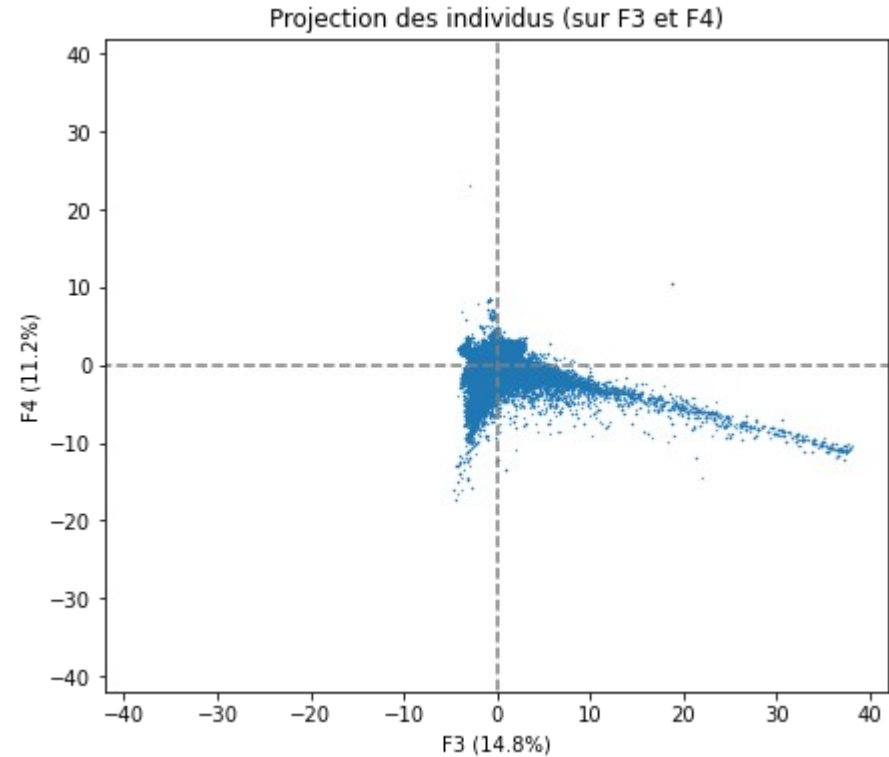
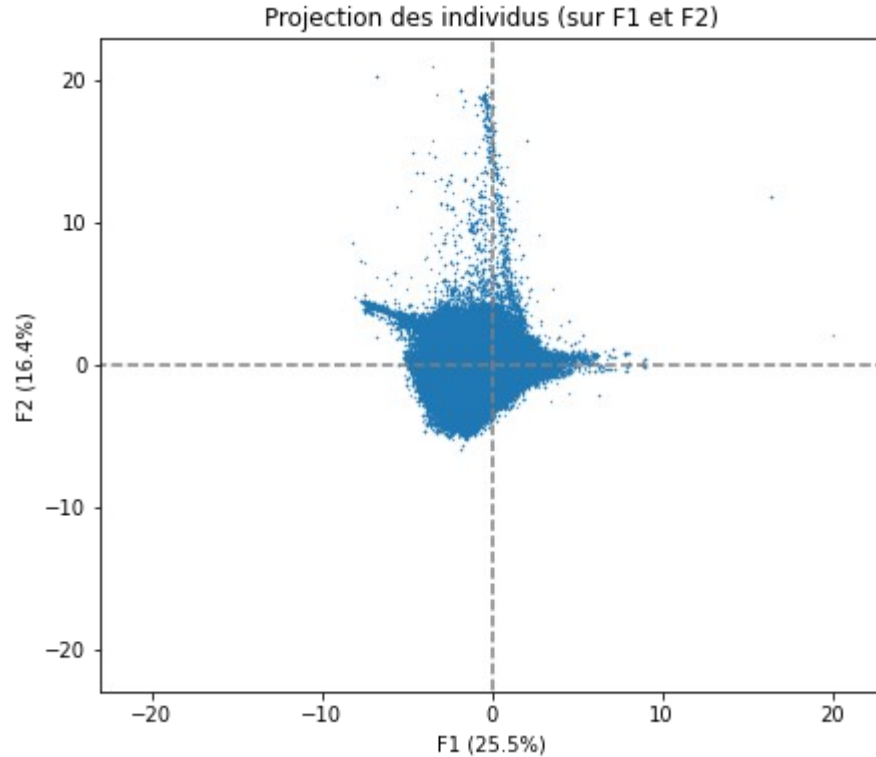
# I- 3- ACP - Cercles des corrélations



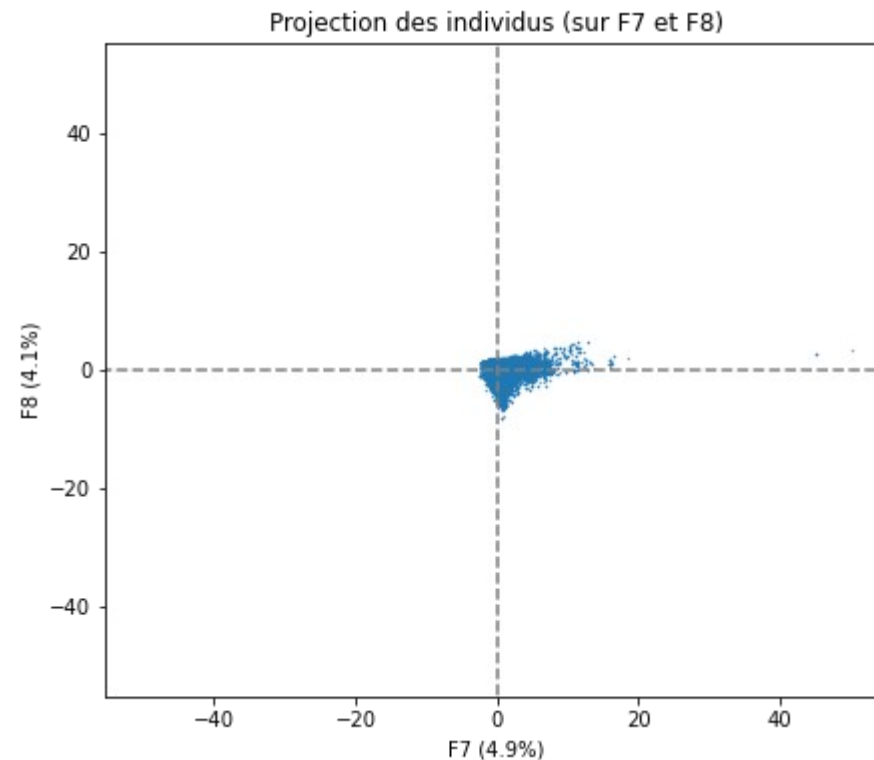
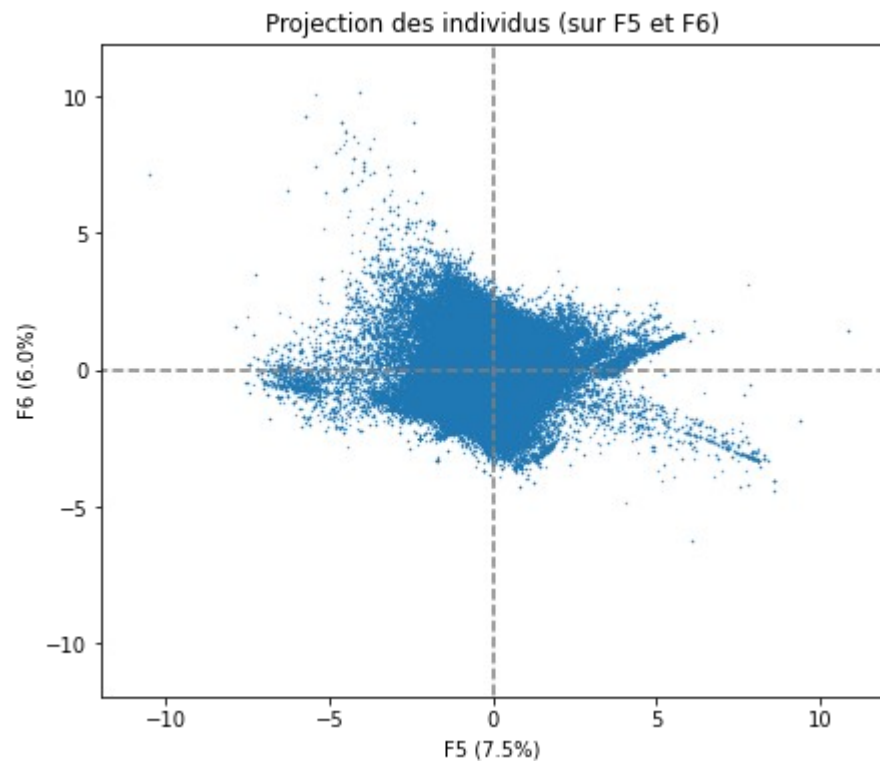
# I- 3- ACP - Cercles des corrélations



# I- 3- ACP - Projection des individus

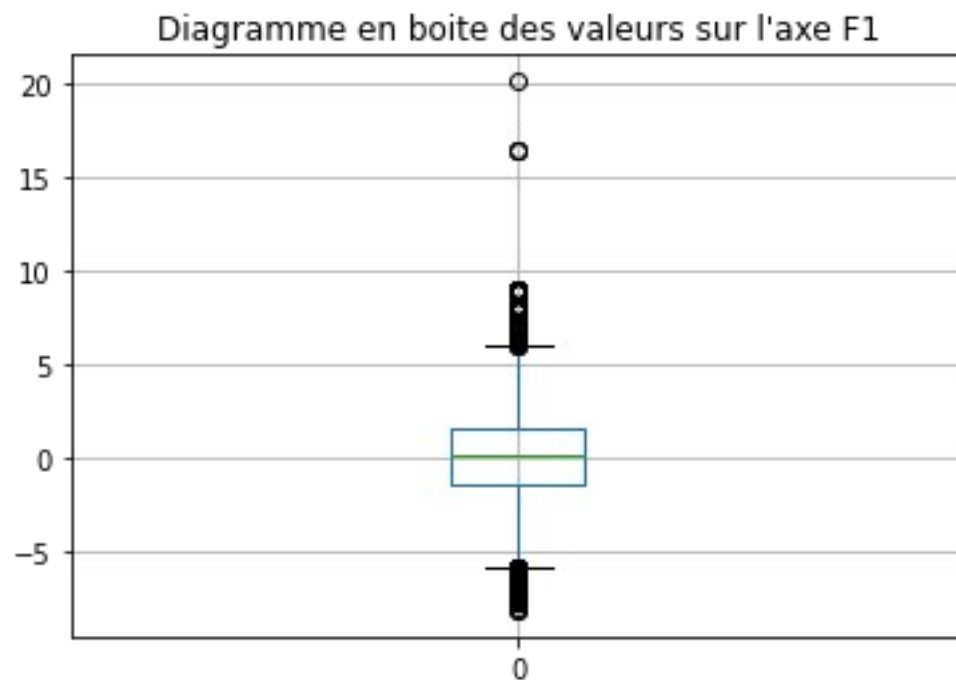


# I- 3- ACP - Projection des individus



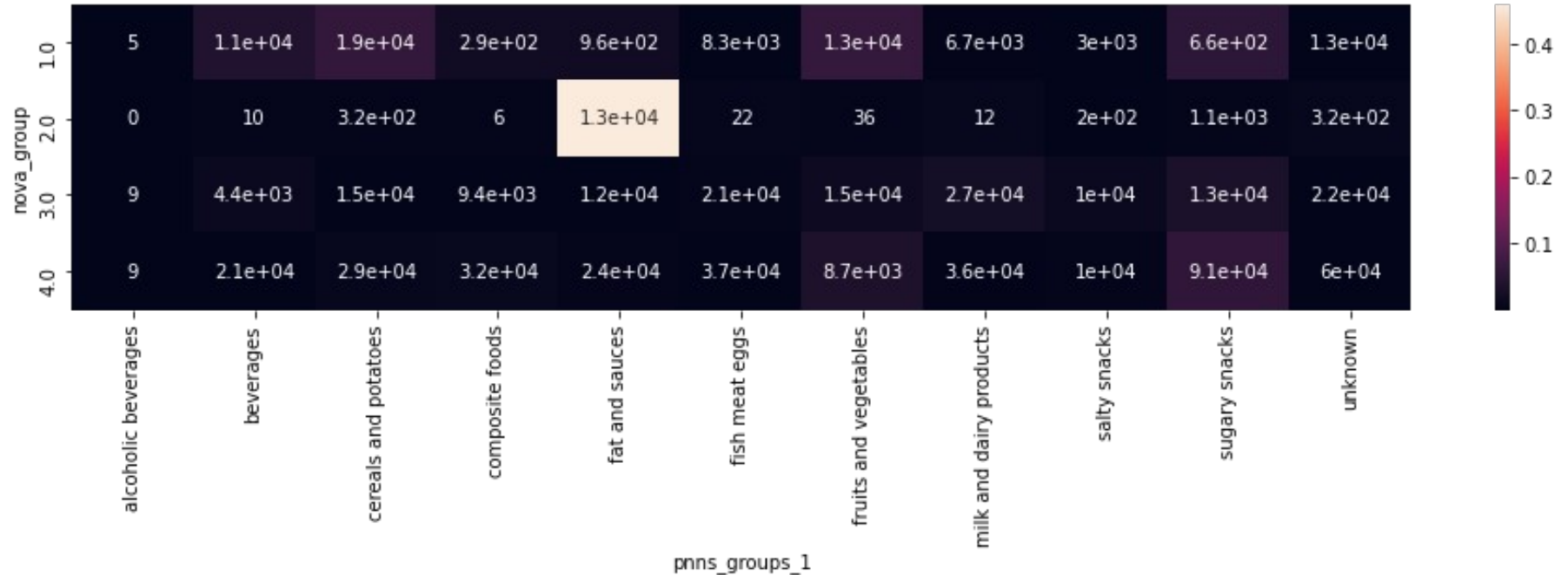


## I- 3- ACP - Score avec F1



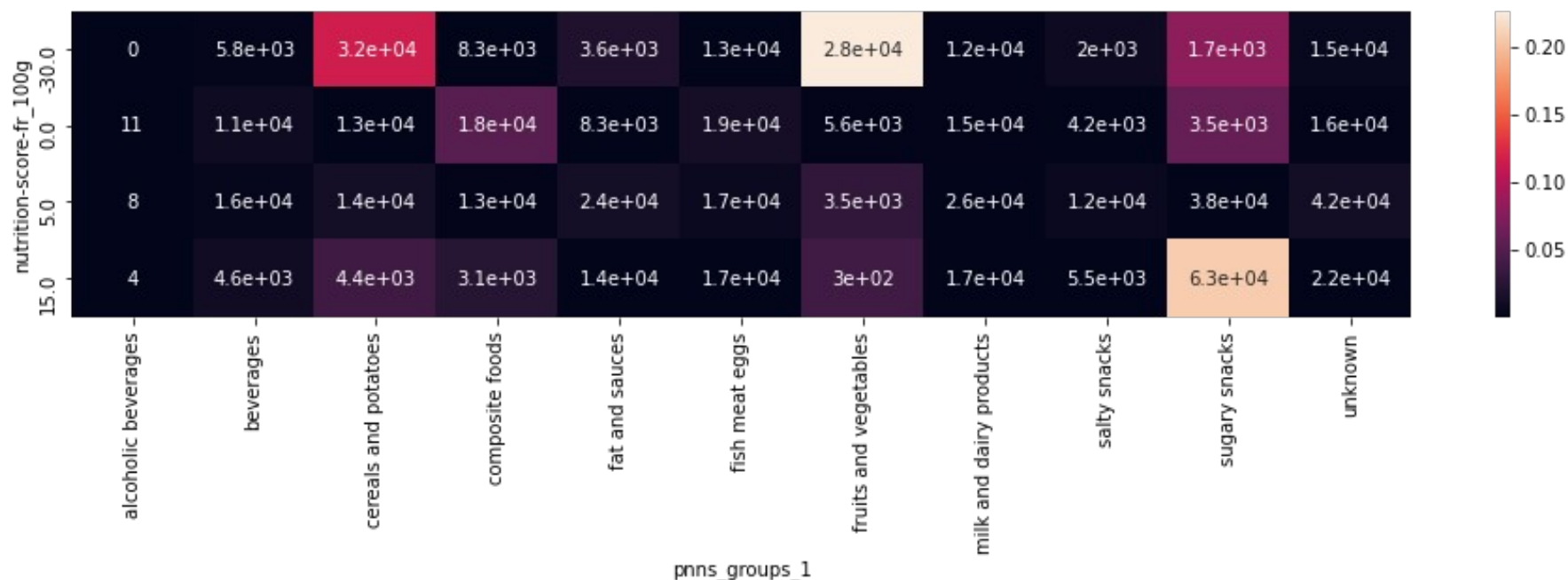
# I- 4- Variables qualitatives

## Groupe Nova - Catégorie



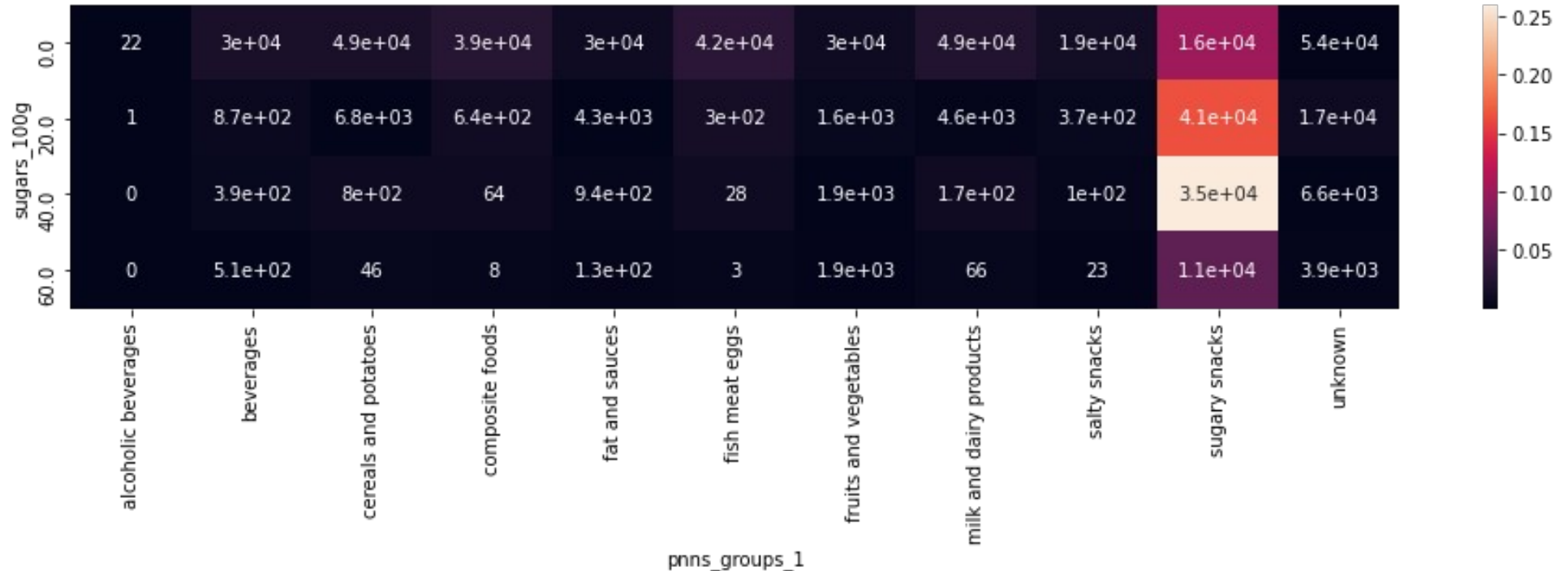
# I- 4- Variables qualitatives

## Nutri-score - Catégorie



# I- 4- Variables qualitatives

## Sucres - Catégorie



# Les faits marquants

- **Les variables sont globalement concentrées : Trouver des alternatives en dehors des valeurs centrales est difficile**
- **L'axe d'inertie principal des données permet de comparer les produits selon plusieurs facteurs à la fois.**
- **Il peut être difficile de trouver des alternatives selon la catégorie.**
  - Car il existe des corrélations entre les indicateurs et la catégorie
  - Car plus la catégorie est précise, moins il est probable que les effectifs soient grands

## Limites de l'étude

- Ne prend pas en compte le pourcentage de fruits / légumes.
- Ne prend pas en compte la liste des ingrédients
- Ne s'appuie sur aucune étude médicale