Application en lien avec l'alimentation

Appel à projet par l'agence "Santé publique France".

Le jeu de données Open Food Fact

- Informations générales du produit : nom, date de modification, etc.
- Tags : catégorie du produit, localisation, marque, etc.
- Valeurs nutritionnelles pour 100 grammes du produit.

Sommaire

- Idée application
- Nettoyage
  - Variables « intéressantes »
  - Taux de remplissage
  - Colonnes redondantes
  - Localisation
- Analyse univariée
- Analyse multivariée
  - corrélation
  - réduction dimensionnelle
  - ANOVA
- Synthèse

Idée d'application

Idée d'application

Informations du data set

- 1 750 000 produits, 184 "informations"
- Code
- Nutriscore
- Categories
- Données nutritionnelles
- Emballages/empreinte Co2
- Photo
- Marque

Idée d'application

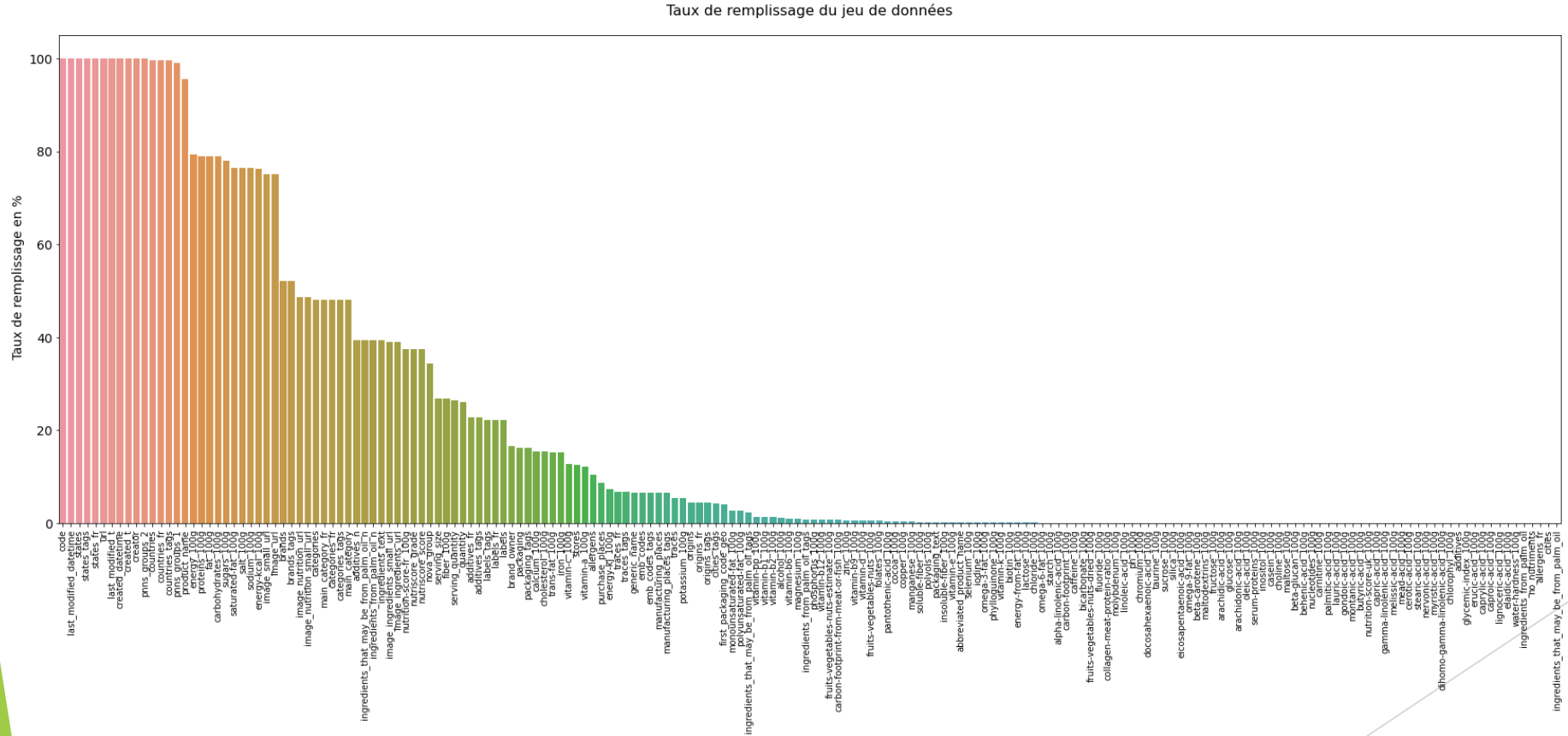
- Scan le produit/photo et envoie sur la fiche produit avec ses information ou bien la créer si le produit n'est pas dans la DataBase
- Evaluation du Nutriscore en fonction des valeurs nutritionnelles si celui n'est pas présent, propositions produits de meme categorie qui a un meilleur score
- Impact Co2

Nettoyage

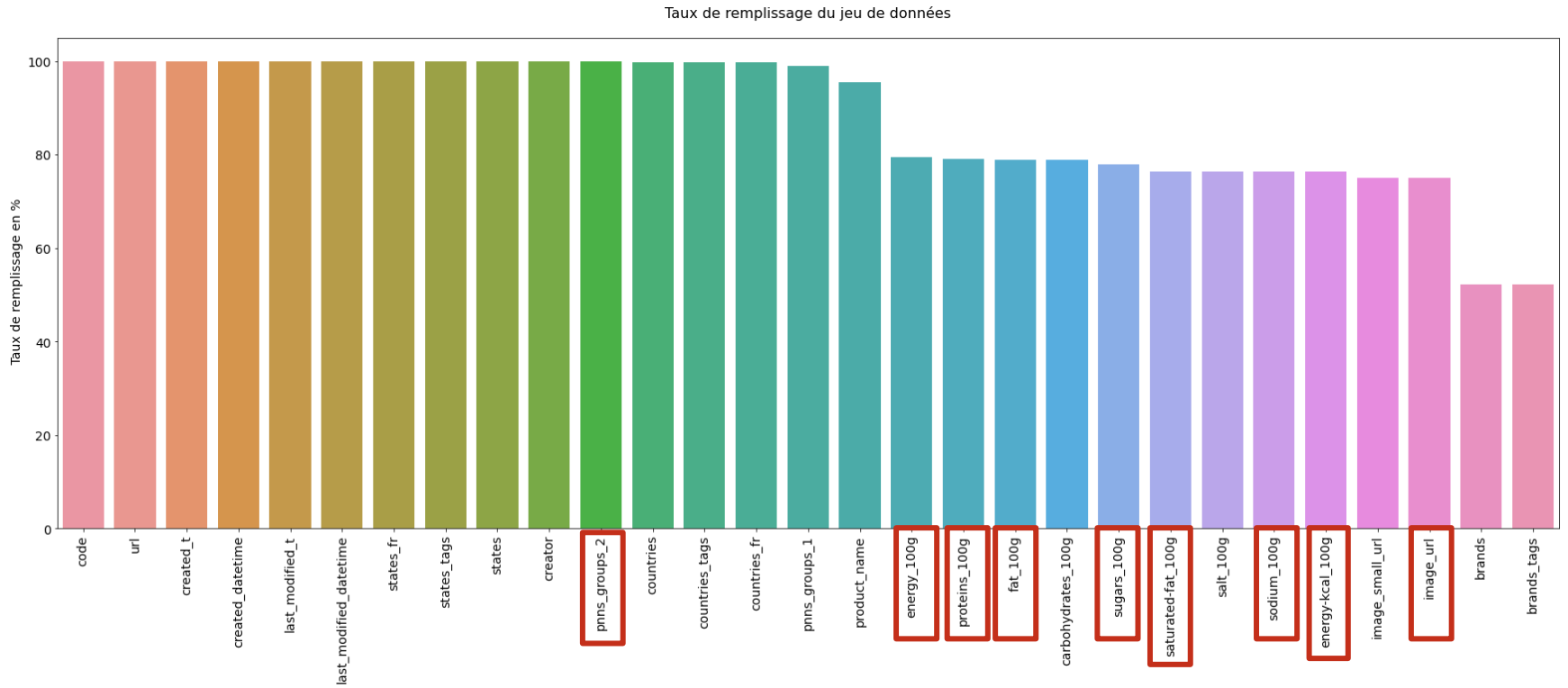
Nettoyage

Identifications de doublons (global/code)

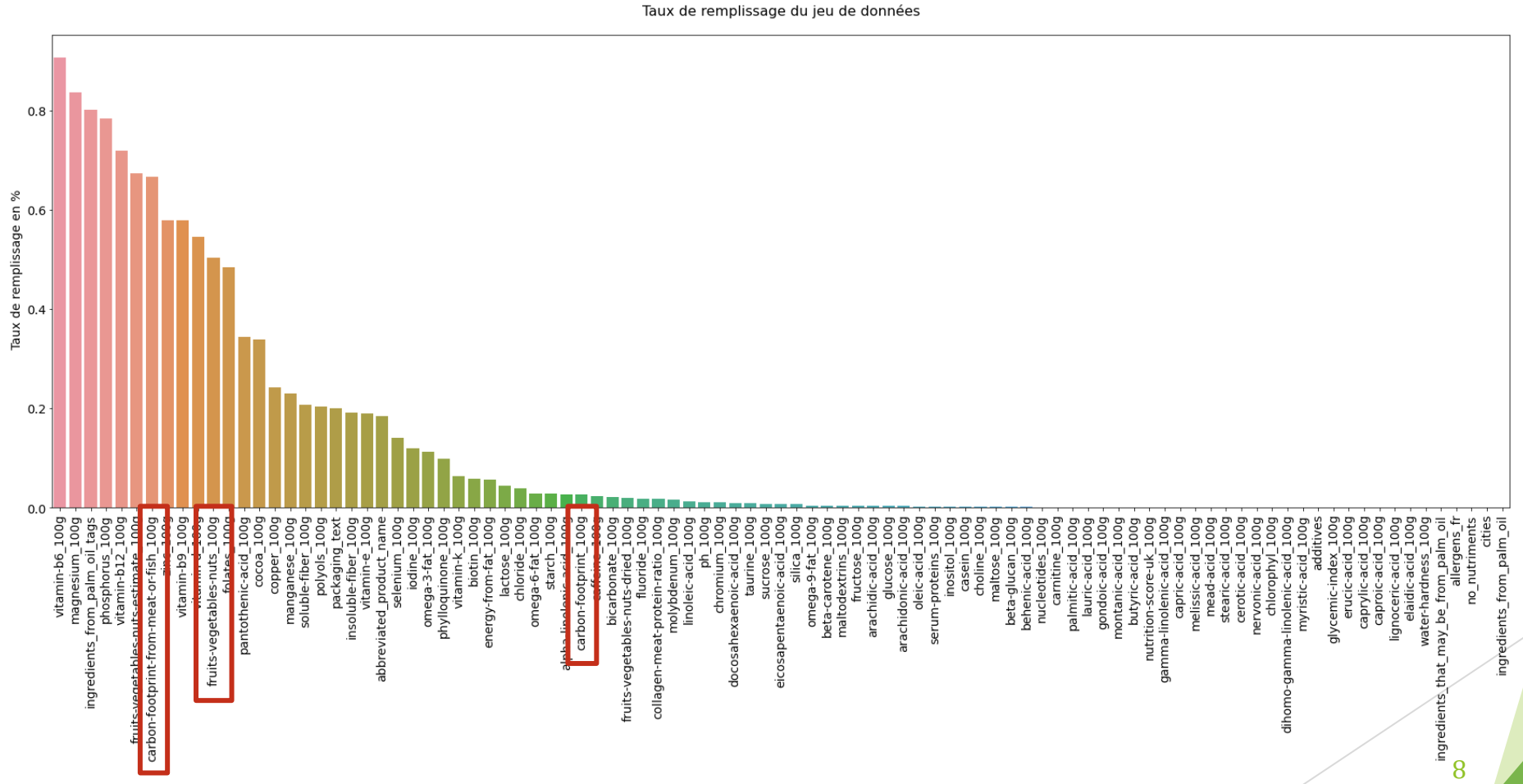
Identification des variables vides ou interessantes



Nettoyage

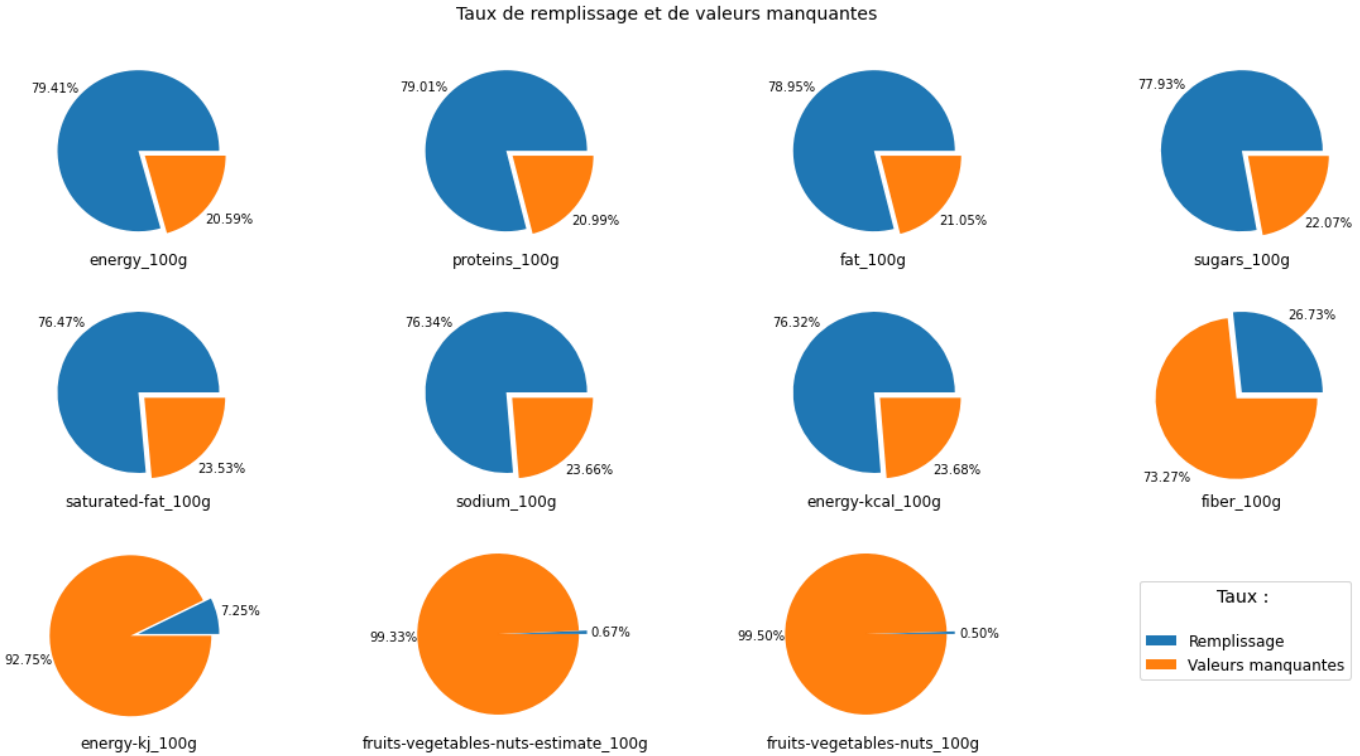

Taux de remplissage du jeu de données
Taux de remplissage en %
0
20
40
60
80
100
code
url
created_t
created_datetime
last_modified_t
last_modified_datetime
states_fr
states_tags
states
creator
pnns_groups_2
countries
countries_tags
countries_fr
pnns_groups_1
product_name
energy_100g
proteins_100g
fat_100g
carbohydrates_100g
sugars_100g
saturated-fat_100g
salt_100g
sodium_100g
energy-kcal_100g
image_small_url
image_url
brands
brands_tags

Nettoyage


Taux de remplissage du jeu de données
Taux de remplissage en %
0.0
0.2
0.4
0.6
0.8
vitamin-b6_100g
magnesium_100g
ingredients_from_palm_oil_tags
phosphorus_100g
vitamin-b12_100g
fruits-vegetables-nuts-estimate_100g
carbon-footprint-from-meat-or-fish_100g
zinc_100g
vitamin-b9_100g
vitamin-d_100g
fruits-vegetables-nuts_100g
folates_100g
pantothenic-acid_100g
cocoa_100g
copper_100g
manganese_100g
soluble-fiber_100g
polyols_100g
packaging_text
insoluble-fiber_100g
vitamin-e_100g
abbreviated_product_name
selenium_100g
iodine_100g
omega-3-fat_100g
phylloquinone_100g
vitamin-k_100g
biotin_100g
energy-from-fat_100g
lactose_100g
chloride_100g
omega-6-fat_100g
starch_100g
alpha-linolenic-acid_100g
carbon-footprint_100g
caffeine_100g
bicarbonate_100g
fruits-vegetables-nuts-dried_100g
fluoride_100g
collagen-meat-protein-ratio_100g
molybdenum_100g
linoleic-acid_100g
ph_100g
chromium_100g
docosahexaenoic-acid_100g
taurine_100g
sucrose_100g
eicosapentaenoic-acid_100g
silica_100g
omega-9-fat_100g
beta-carotene_100g
maltodextrins_100g
fructose_100g
arachidic-acid_100g
glucose_100g
arachidonic-acid_100g
oleic-acid_100g
serum-proteins_100g
inositol_100g
casein_100g
choline_100g
maltose_100g
beta-glucan_100g
behenic-acid_100g
nucleotides_100g
carnitine_100g
palmitic-acid_100g
lauric-acid_100g
gondoic-acid_100g
montanic-acid_100g
butyric-acid_100g
nutrition-score-uk_100g
gamma-linolenic-acid_100g
capric-acid_100g
melissic-acid_100g
mead-acid_100g
stearic-acid_100g
cerotic-acid_100g
nervonic-acid_100g
chlorophyl_100g
dihomo-gamma-linolenic-acid_100g
myristic-acid_100g
additives
glycemic-index_100g
erucic-acid_100g
caprylic-acid_100g
caproic-acid_100g
lignoceric-acid_100g
elaidic-acid_100g
water-hardness_100g
ingredients_that_may_be_from_palm_oil
allergens_fr
no_nutriments
cities
ingredients_from_palm_oil

Nettoyage

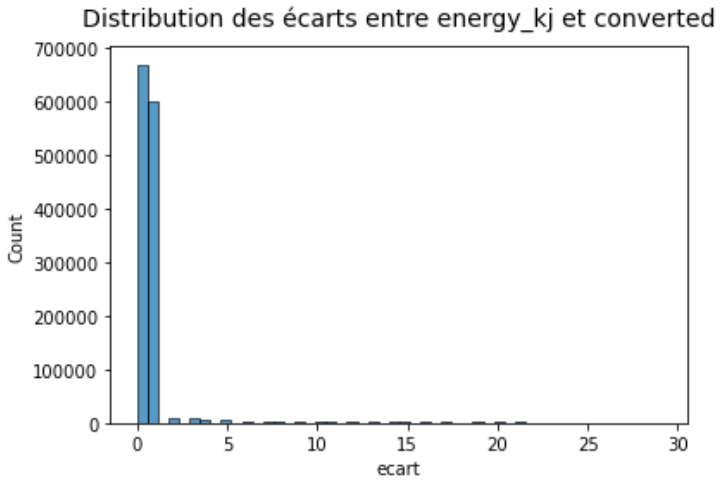
Qu'est-ce qui va nous interesser ? → Valeurs nutritionnelles et nutriscore

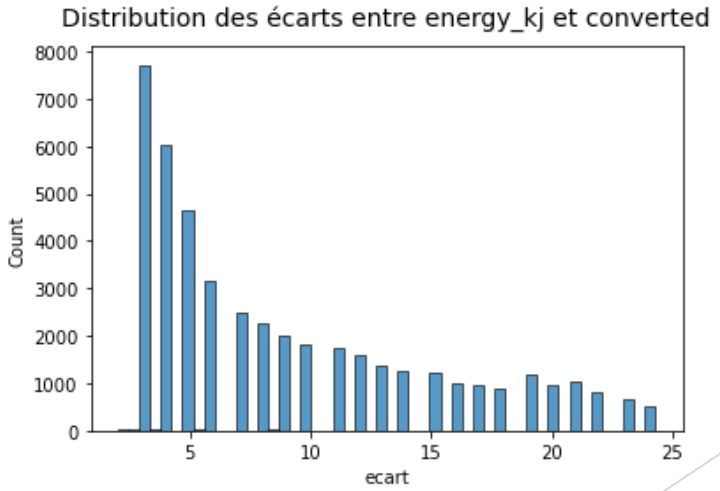


Nettoyage

Colonnes redondantes

- Ernergy
- Kj/kcal/sans unité

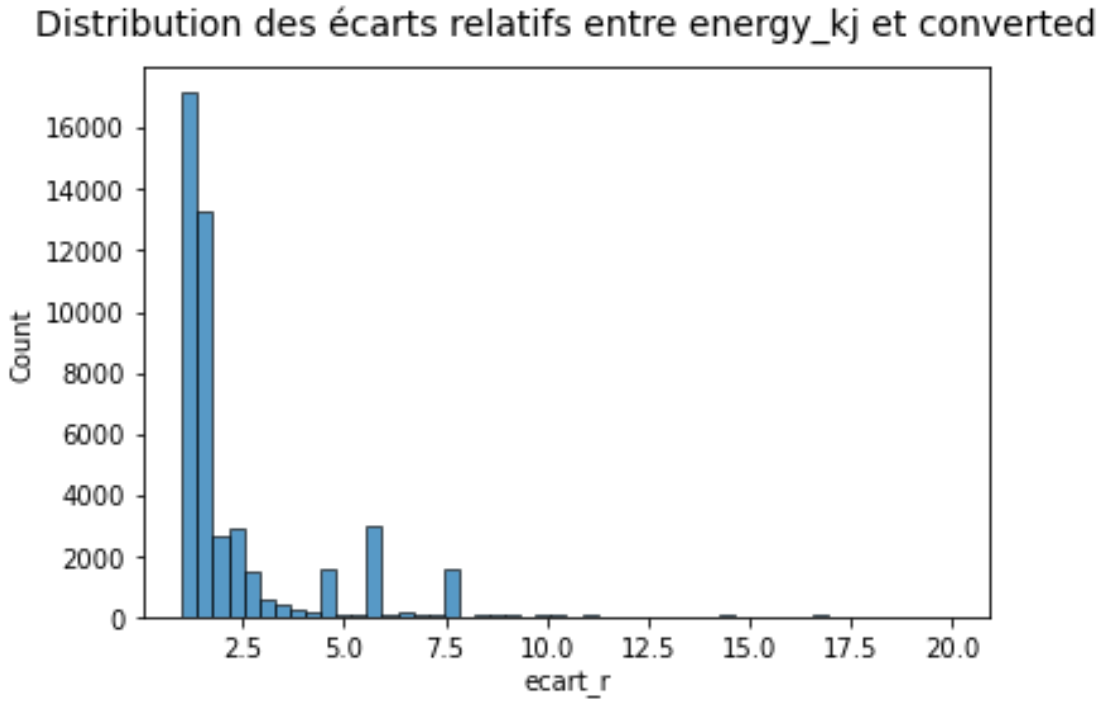




Nettoyage

Colonnes redondantes

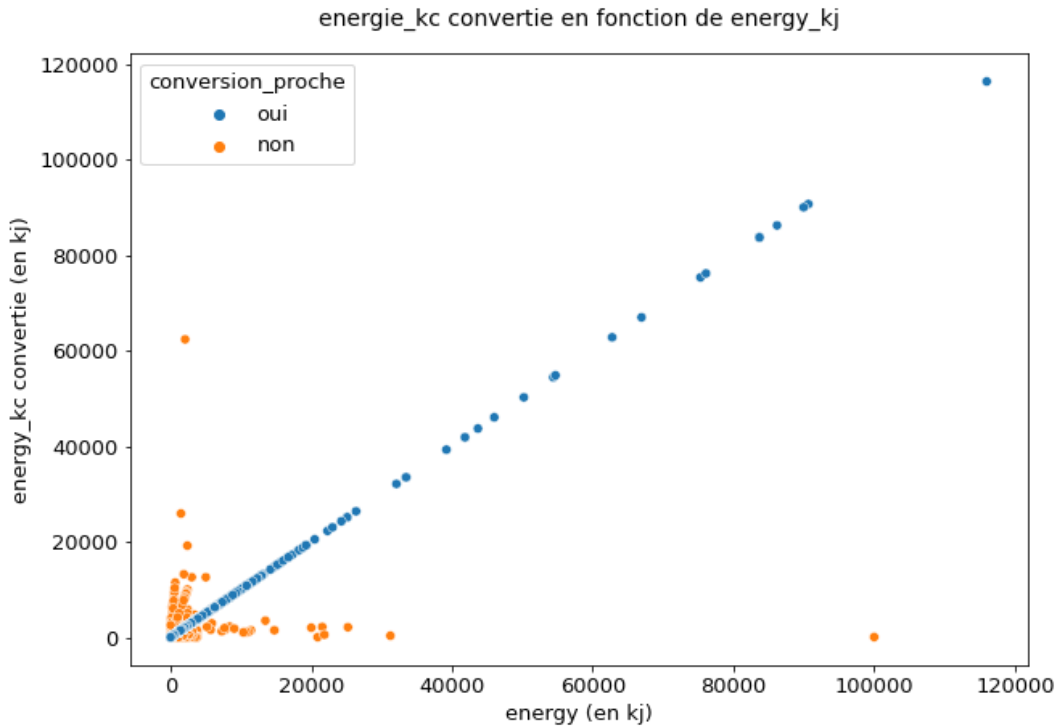
- Ernergy
- Kj/kcal/sans unité



Nettoyage

Colonnes redondantes

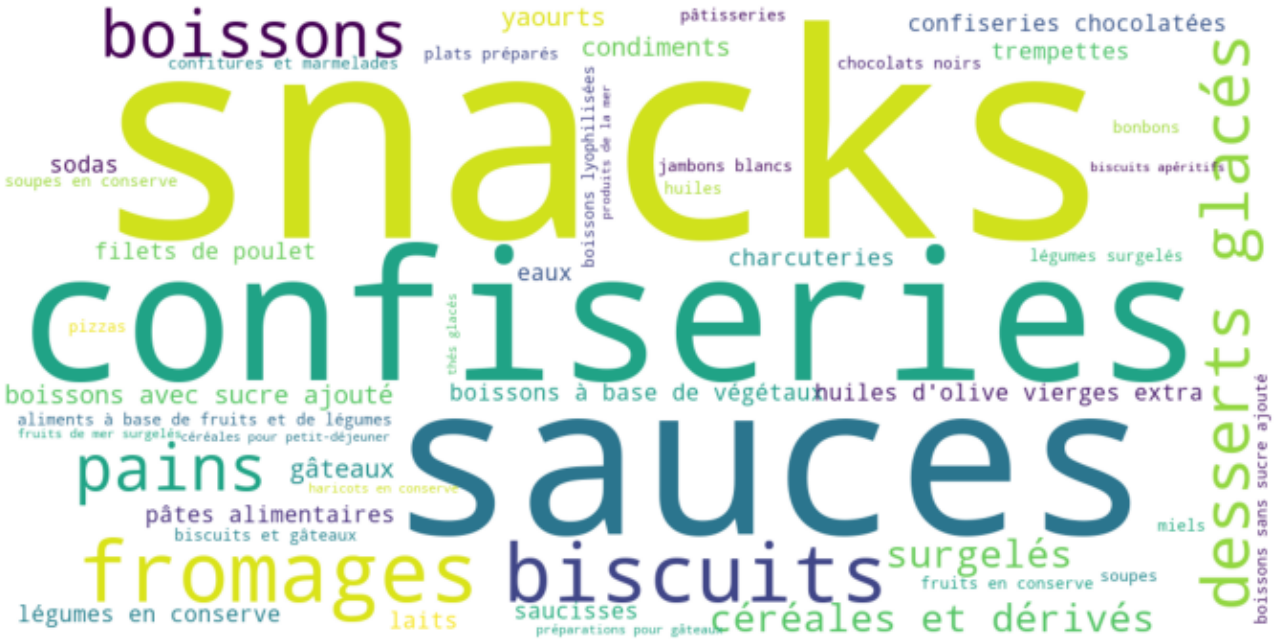
- Ernergy
- Kj/kcal/sans unité
- <u>Elimination 3500 produits</u>



Nettoyage

Colonnes redondantes

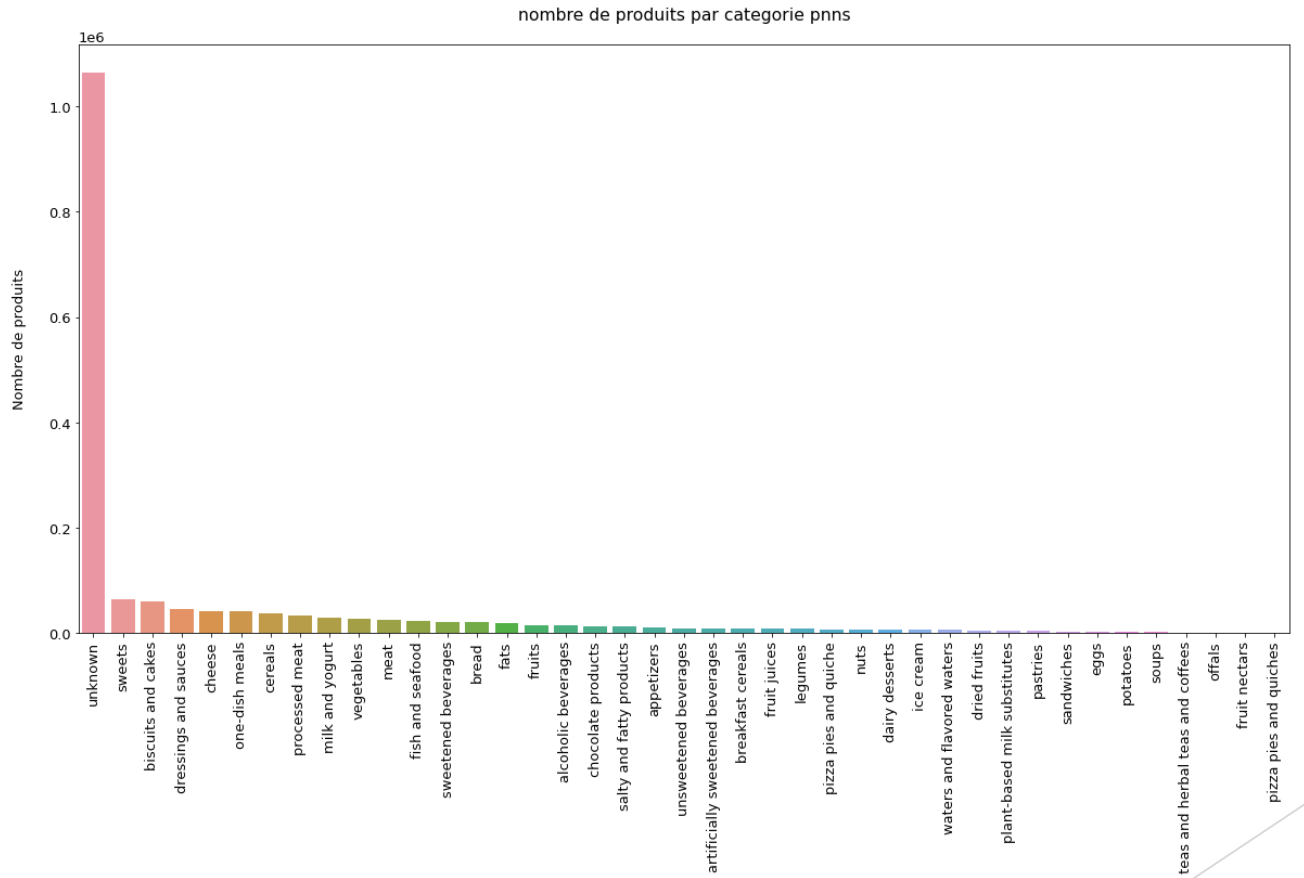
Categories



Nettoyage

Colonnes redondantes

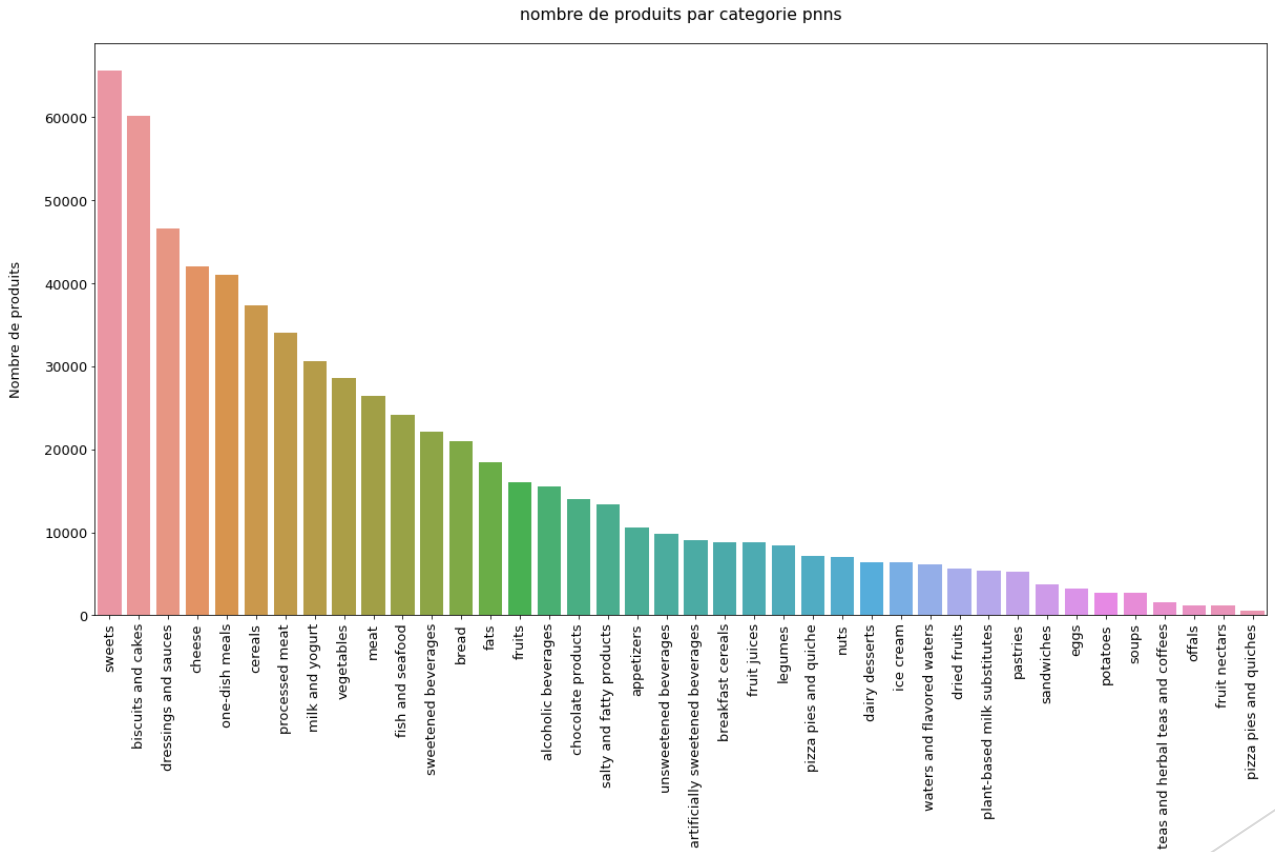
Pnns Groups



Nettoyage

Colonnes redondantes

Pnns Groups



Nettoyage

Colonnes redondantes

Nutriscore (score/grade)

Pays (Selections produits vendus en France)

France	792372	**45.45%**
Non France	951160	54.55%

Nettoyage

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 789472 entries, 0 to 1743530
Data columns (total 16 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   code                  789472 non-null  object
 1   product_name          767443 non-null  object
 2   brand                 418293 non-null  object
 3   category              353482 non-null  object
 4   info_nutri_complete   789472 non-null  int64
 5   pnns_group            789390 non-null  object
 6   nutriscore            275387 non-null  float64
 7   nutriscore_grade      275383 non-null  object
 8   energy_kj             617648 non-null  float64
 9   sugars                613736 non-null  float64
 10  fat                   611579 non-null  float64
 11  saturated_fat         614549 non-null  float64
 12  sodium                598321 non-null  float64
 13  fiber                 137888 non-null  float64
 14  proteins              613557 non-null  float64
 15  fruits_vegetables     16454 non-null   float64
dtypes: float64(9), int64(1), object(6)
memory usage: 102.4+ MB
```

DataSet à la fin du nettoyage

De 1 750 000 à 790 000 produits

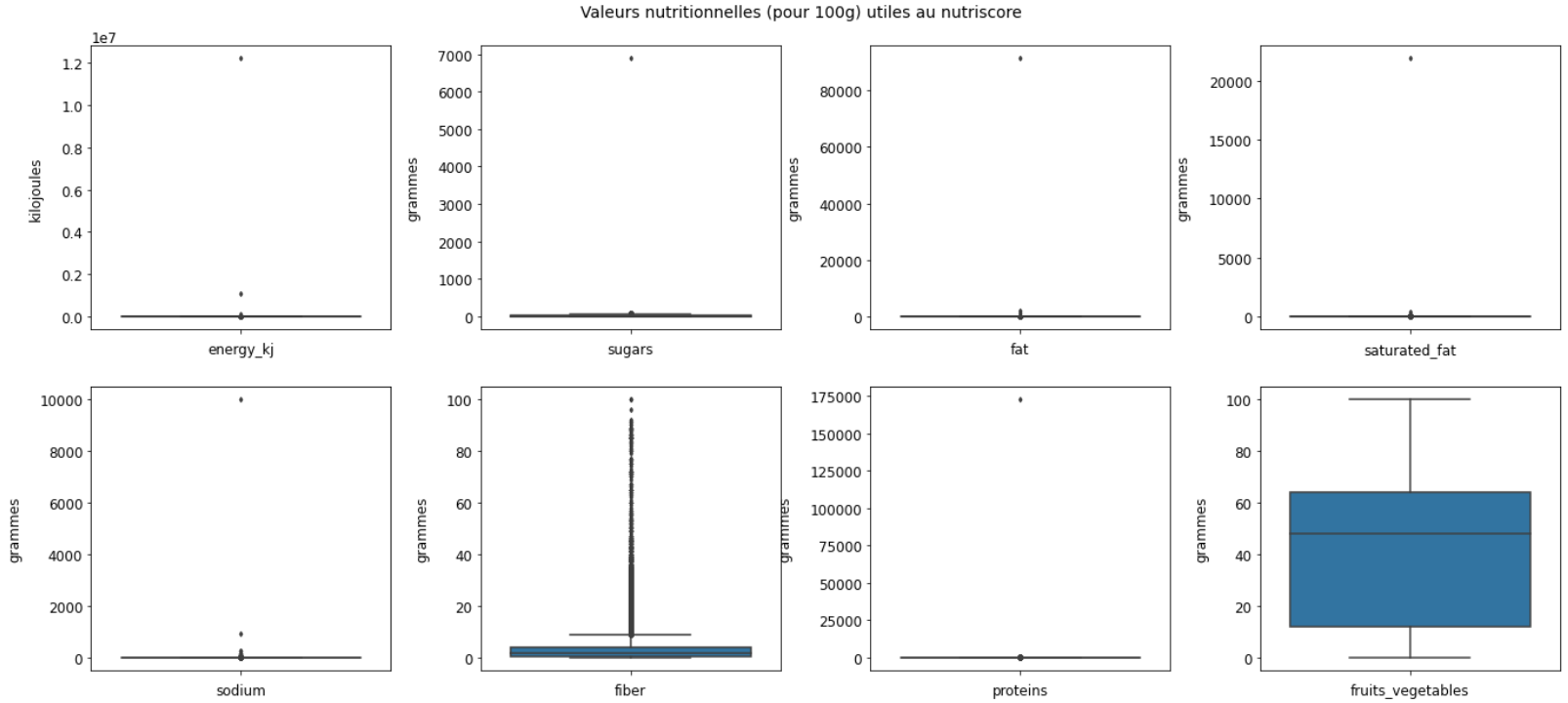
De 184 à 16 colonnes

Conservations des informations relatives au nutriscore et catégories de produits

Analyse univariée

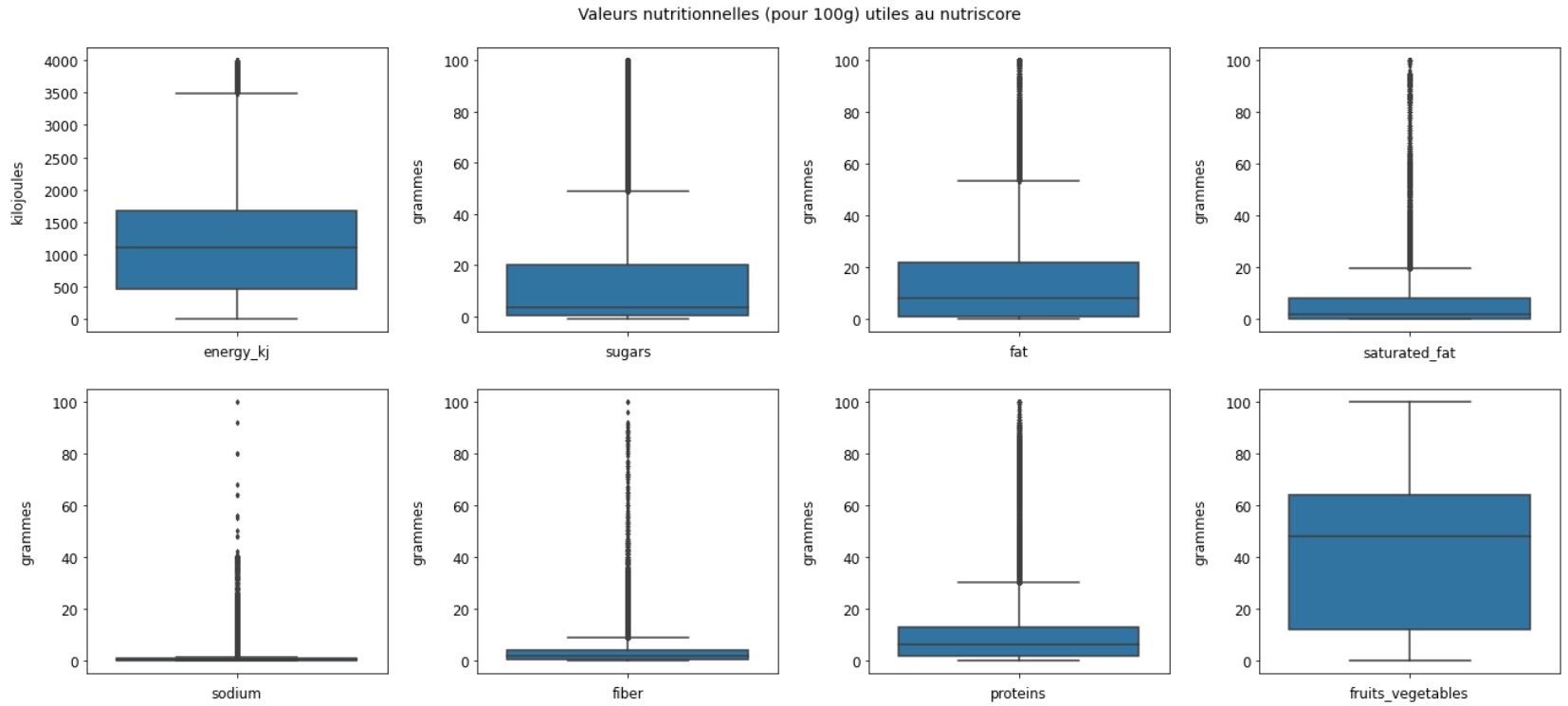
Analyse univariée

Détection derniers outliers sur 8 features → Elimination de 1 800 produits



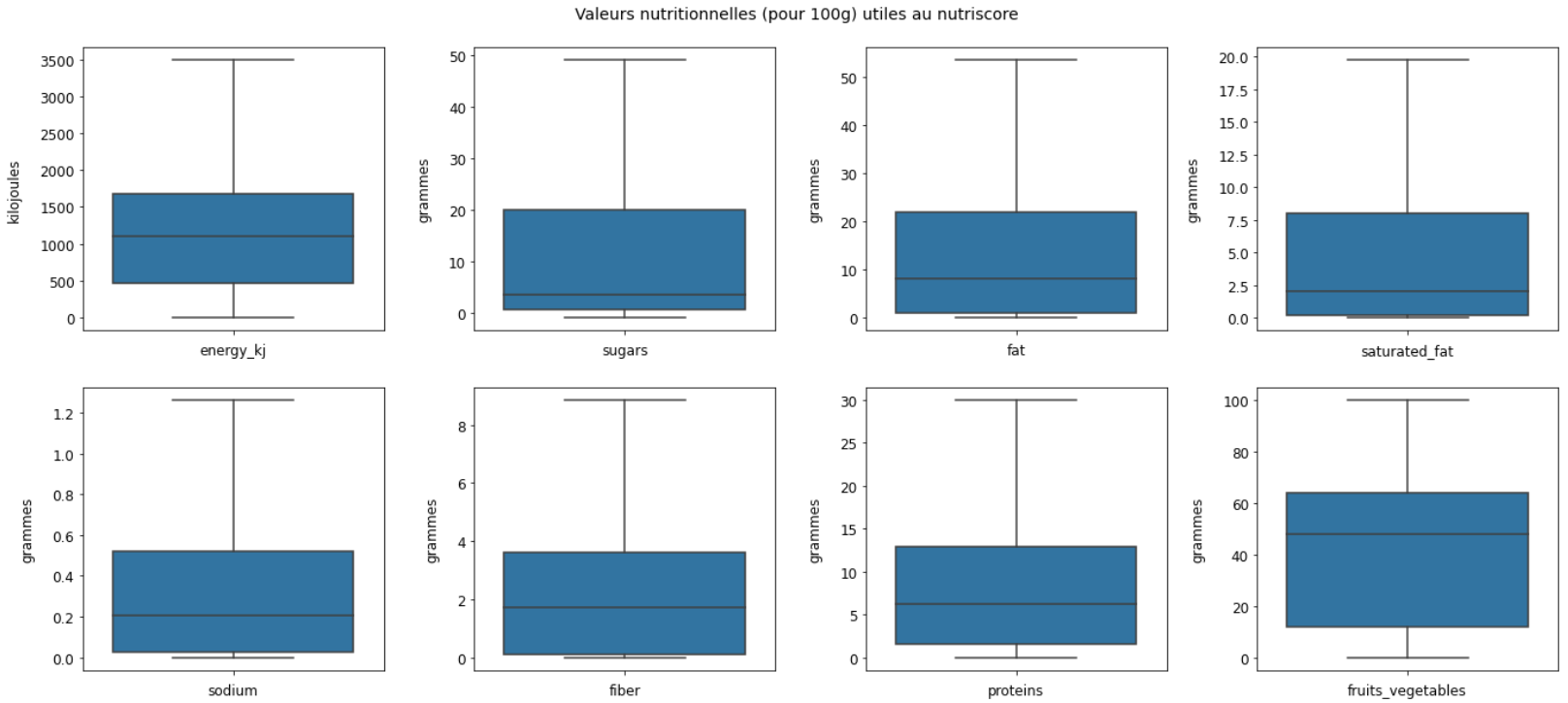
Analyse univariée

Valeurs nutritionnelles



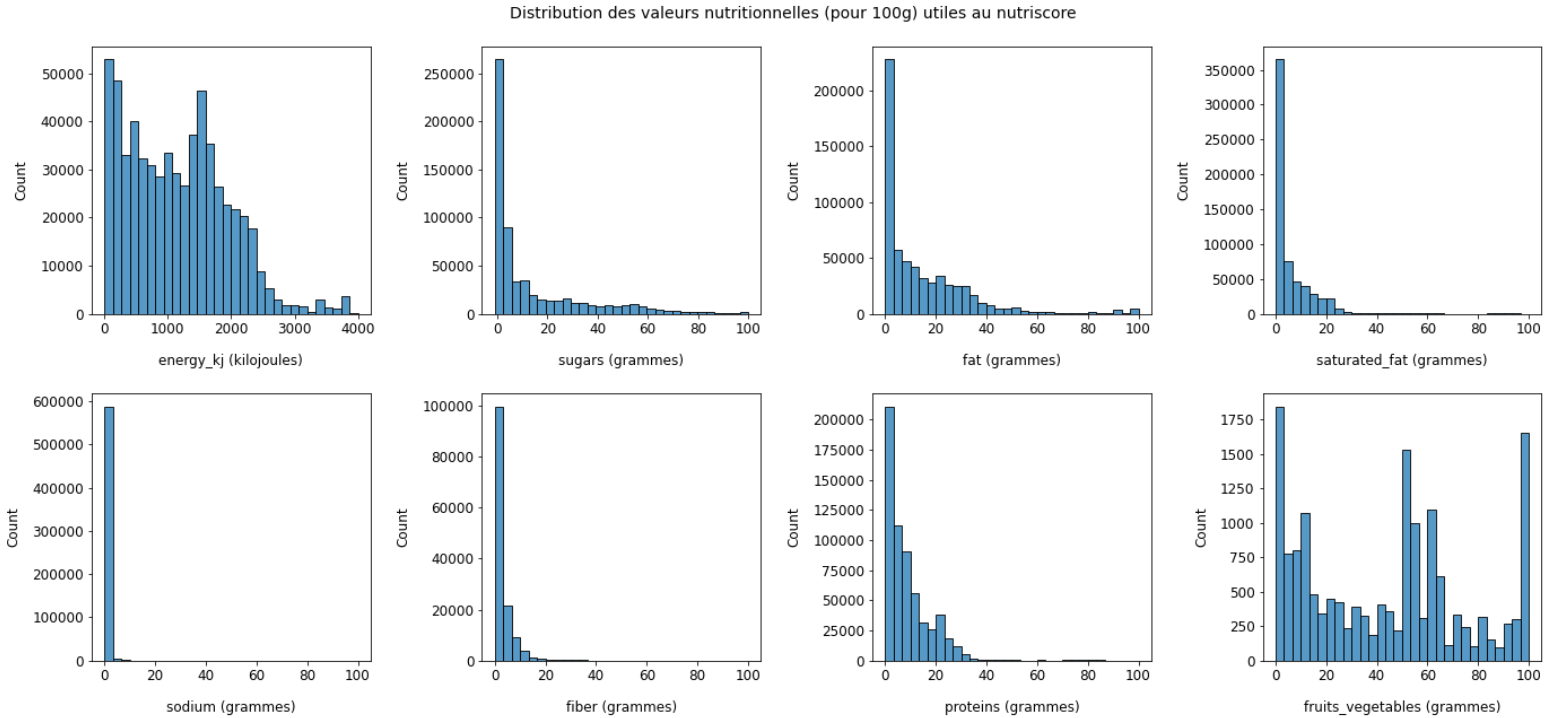
Analyse univariée

Valeurs nutritionnelles



Analyse univariée

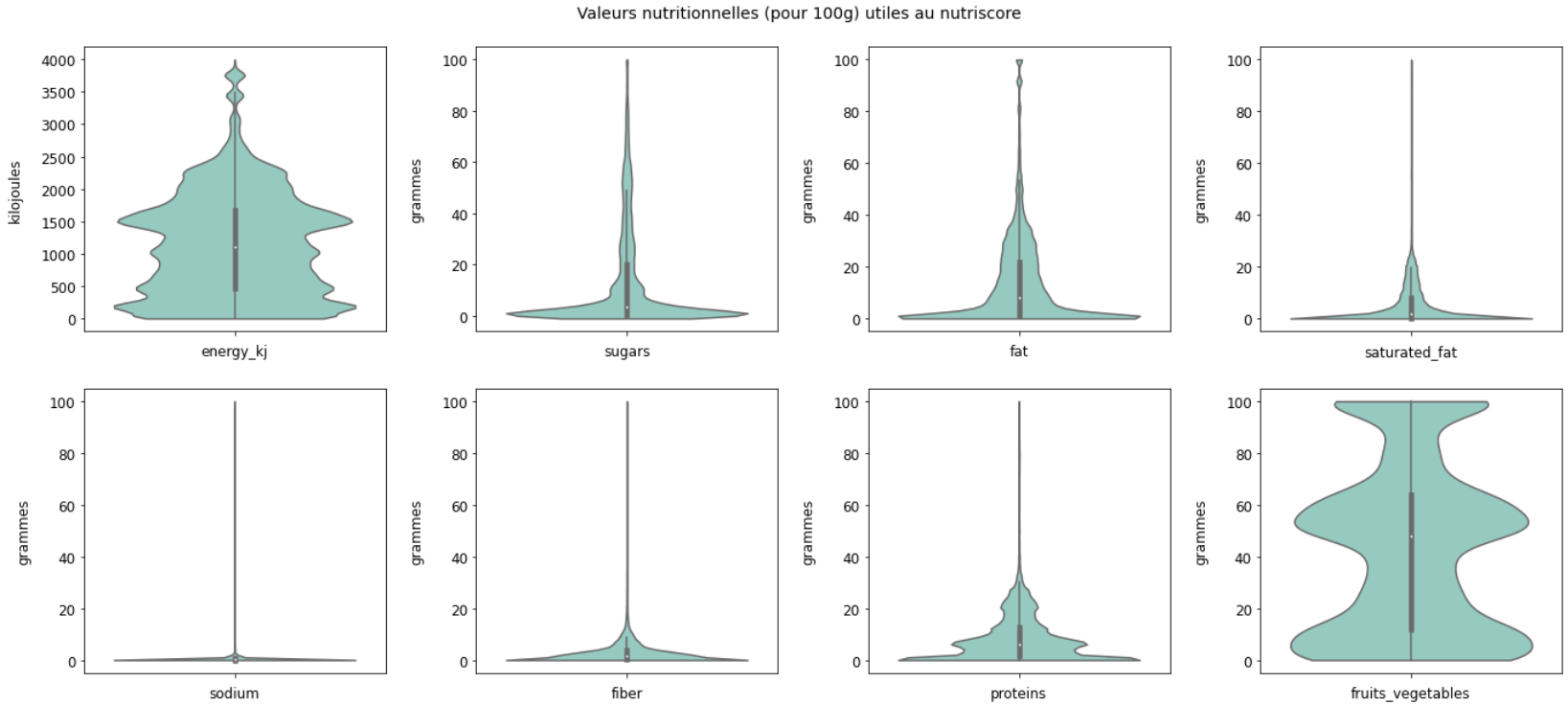
Valeurs nutritionnelles



3 nodes > possibilité de discretiser

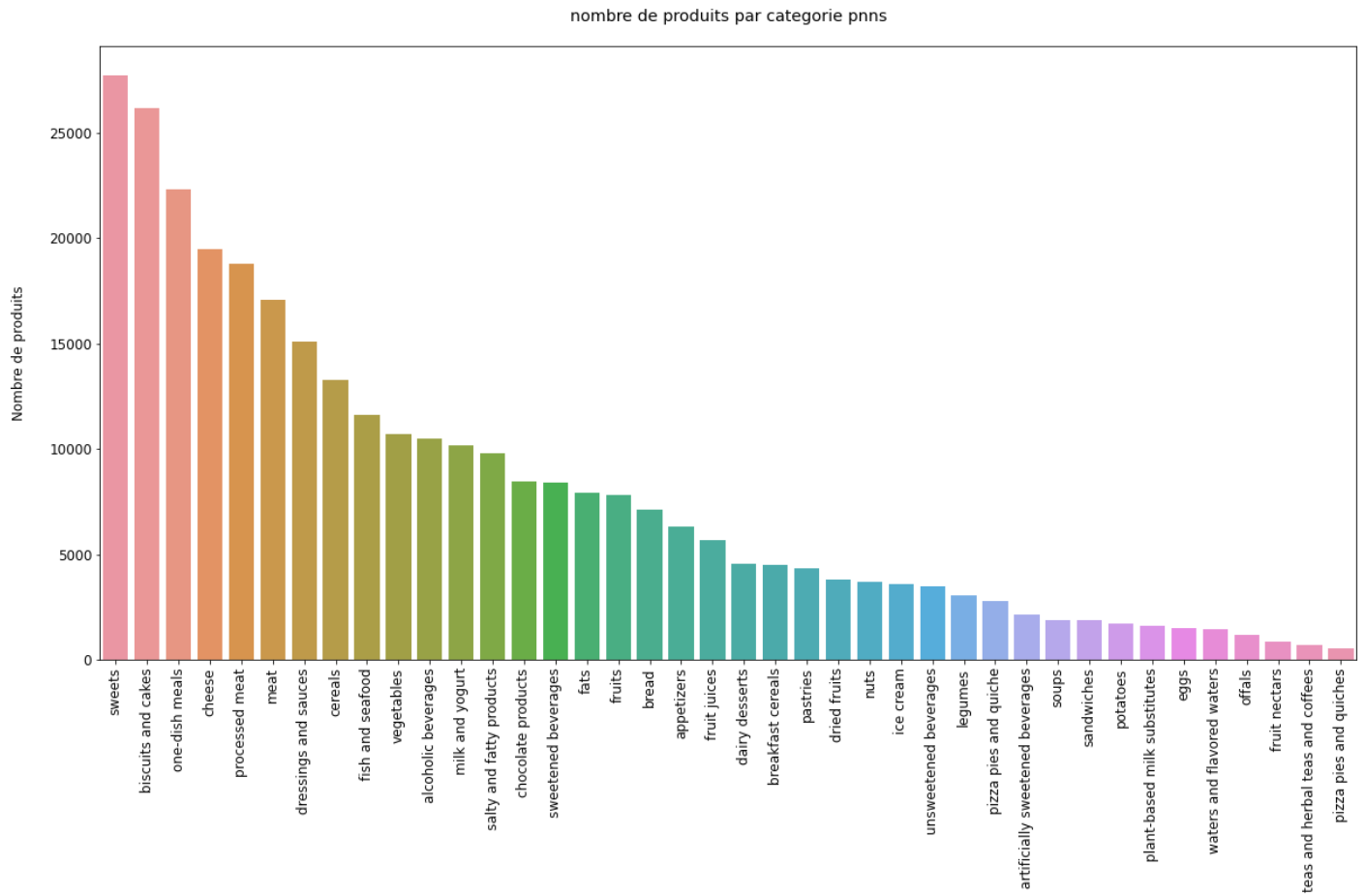
Analyse univariée

Valeurs nutritionnelles



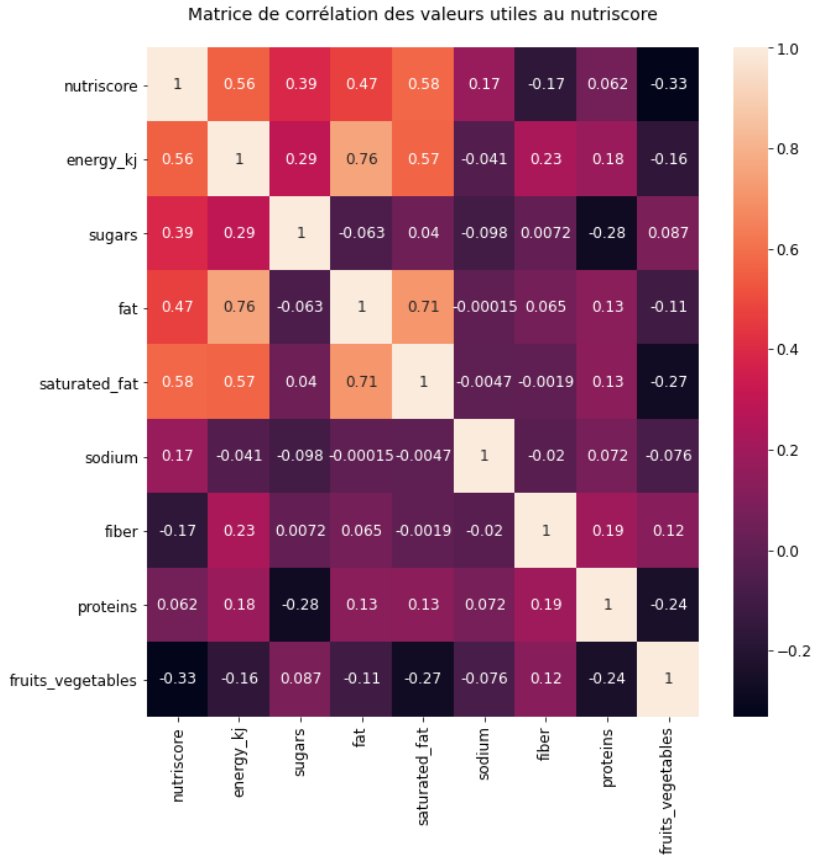
Analyse univariée

Catégories de produits



Analyse multivariée

Analyse multivariée



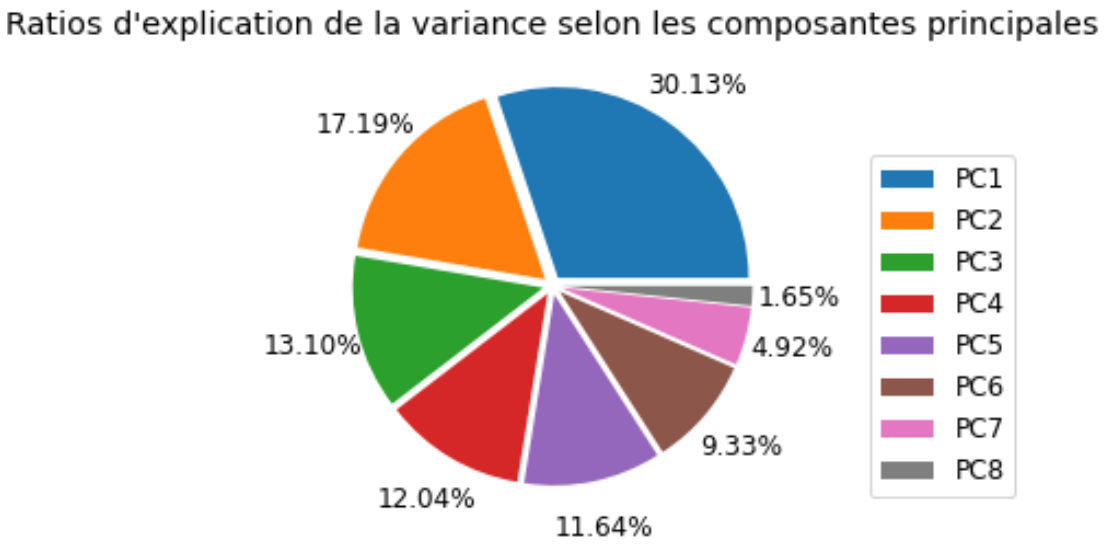
Corrélation

Limitation aux produits "données complètes" ~ 600 000

Remplissage par 0

Positive/Négative
Mauvais/bon pour nutriscore

Analyse multivariée



Reduction dimensionnelle

8 features d'entrée

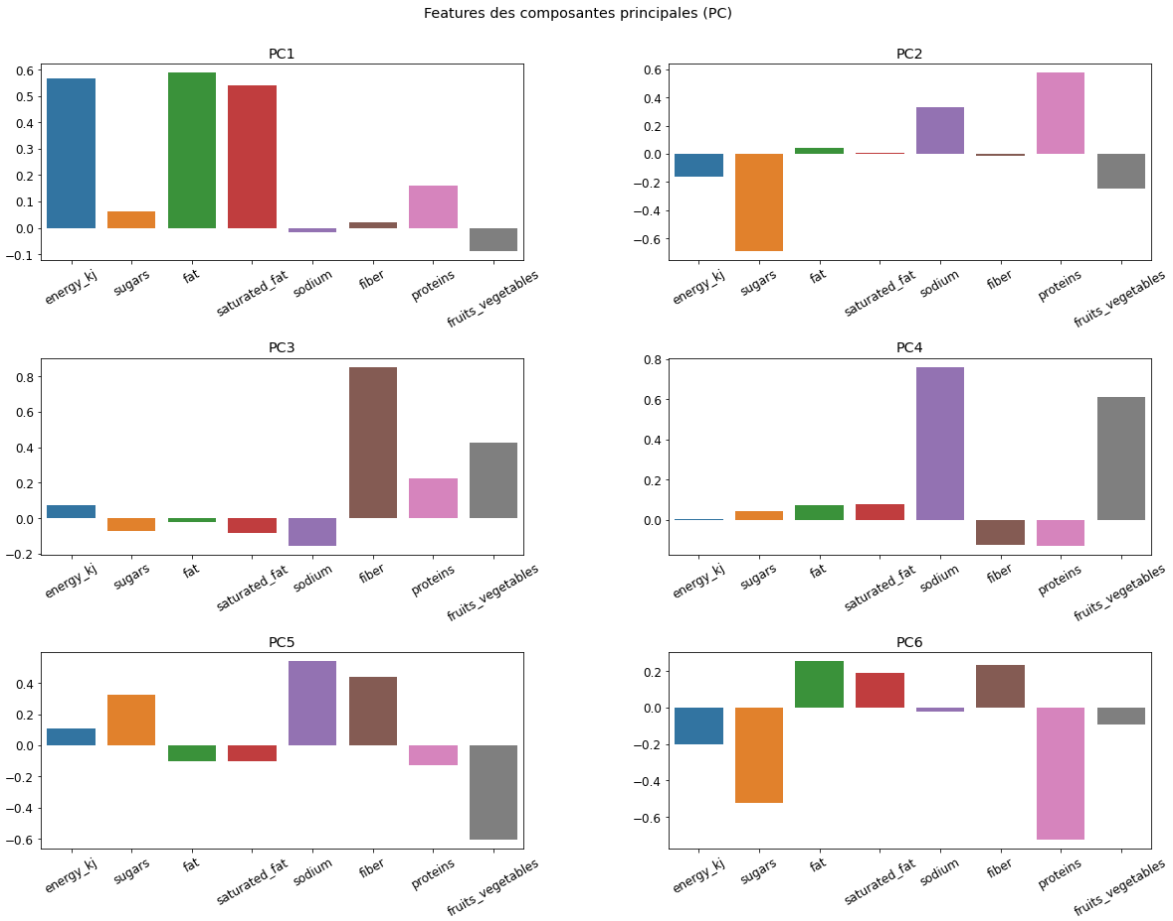
Réduction à 6 pc pour 93% d'explication de variation

ou

Réduction à 5 pc pour 84% d'explication de variation

Analyse multivariée

Composantes nouvelles dimensions



Pc 1

Energy / Fat : corrélation 0.76

Pc 2

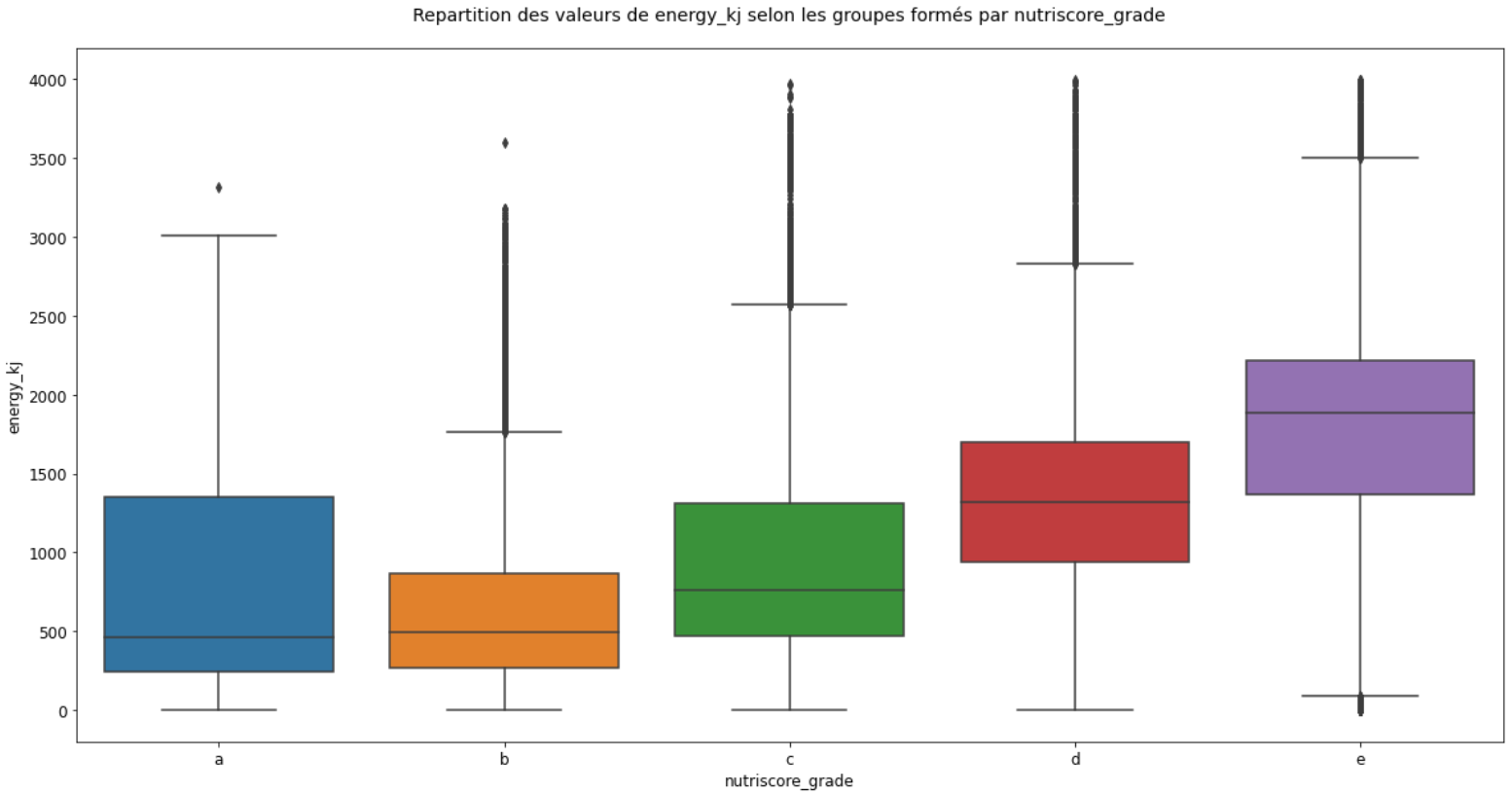
Sucre / Prot : opposés

Pc 3

Fibre / Prot / Fruits : amélioration nutriscore

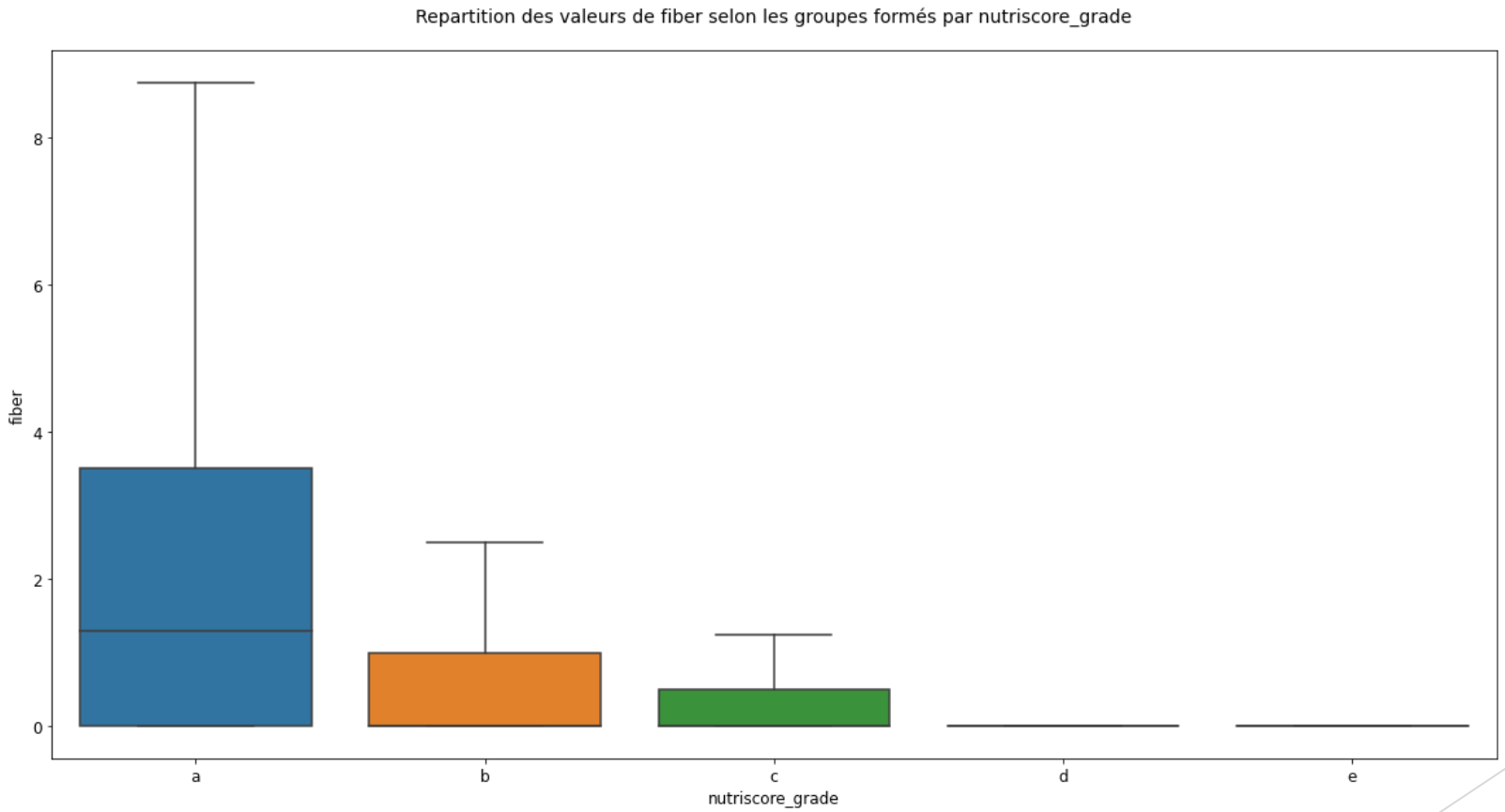
Analyse multivariée

ANOVA : Comparaison des groupes → Limite aux produits qui ont un nutriscore : 275 000



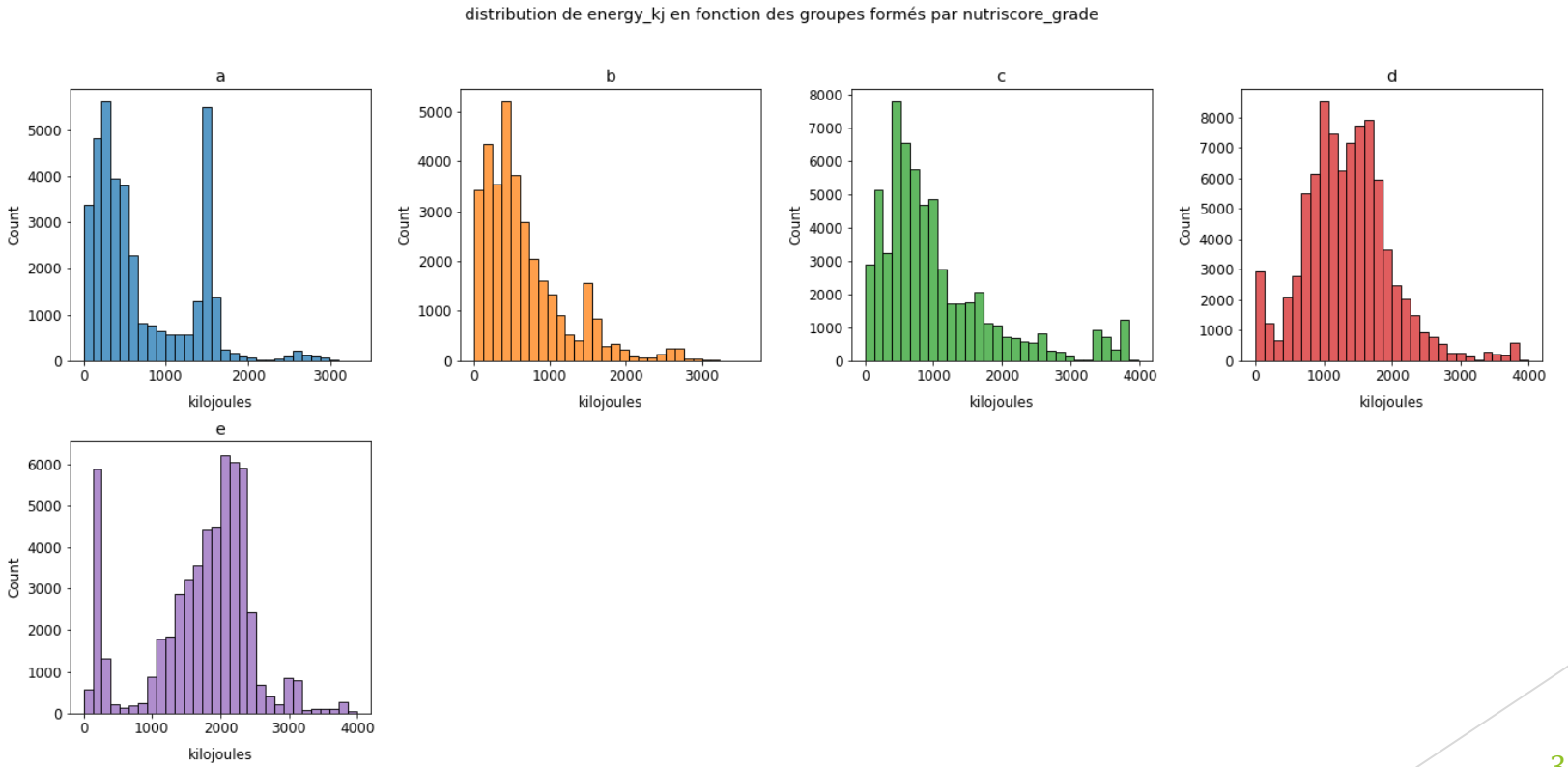
Analyse multivariée

ANOVA : Comparaison des groupes → Limite aux produits qui ont un nutriscore : 275 000



Analyse multivariée

ANOVA : Comparaison des groupes → Limite aux produits qui ont un nutriscore : 275 000



Analyse multivariée

ANOVA : Comparaison des groupes → Limite aux produits qui ont un nutriscore : 275 000

Variable	Resultats du test
Energy_kj	F = 18 900 / Pvalue < 0.05
Sugars	F = 8 900 / Pvalue < 0.05
Fat	F = 14 300 / Pvalue < 0.05
Saturated_fat	F = 26 300 / Pvalue < 0.05
Sodium	F = 2 160 / Pvalue < 0.05
Fibers	F = 4 360 / Pvalue < 0.05
Proteins	F = 1 720 / Pvalue < 0.05
Fruits	F = 650 / Pvalue < 0.05

Groupes par nutriscore

Analyse multivariée

ANOVA : Comparaison des groupes → Limite aux produits qui ont un pnns-group : 270 000

Variable	Resultats du test
Energy_kj	F = 12 500 / Pvalue < 0.05
Sugars	F = 13 400 / Pvalue < 0.05
Fat	F = 12 900 / Pvalue < 0.05
Saturated_fat	F = 6 700 / Pvalue < 0.05
Sodium	F = 1 400 / Pvalue < 0.05
Fibers	F = 1 200 / Pvalue < 0.05
Proteins	F = 12 400 / Pvalue < 0.05
Fruits	F = 550 / Pvalue < 0.05

Groupes par pnns-groups

Pertinence/Faisabilité

Pertinence / Faisabilité

ANOVA :

Des groupes plutôt bien séparés, on peut donc estimer le groupe en fonction des variables retenues

Modélisation KNN :

Algorithme de classification utilisé
Prédiction correcte du nutriscore à ~ 73%
Prédiction correcte de catégorie à ~ 69%

Qualité du dataset:

Nombre de valeurs et de produits « erronées »
Beaucoup d'informations manquantes