



## EXPOSÉ PROJET 3

---

# PRÉPARER DES DONNÉES POUR UN ORGANISME DE SANTÉ PUBLIQUE

---

Le 31 Août 2020

Zeineb Guizani

# Plan de la présentation

1. Présentation de l'appel à projet
2. Démarche de nettoyage et d'exploration des données

# I- Présentation de l'appel à projet



- L'agence Santé publique France souhaite déterminer la qualité des des produits alimentaires disponibles dans le marché français en fonction de leur valeur nutritionnelle.
- Analyse de données basée sur le jeu de données Open Food Facts.
- Le dataframe a 1 419 694 lignes et 181 colonnes (df.shape).
- OPEN FOOD FACTS
- Il est constitué de 128 entités numériques à savoir 123 de type float64 et 2 de type int64. Les entités de type objet sont au nombre de 56. (df.info())
- Utiliser le système de classement pour déterminer quels éléments nutritifs ont le plus d'impact sur le niveau de qualité du produit

## 2- Démarche de nettoyage et d'exploration des données

### Etape 1 – Pré-traitement

### Etape 2 - Imputation

### Etape 3 – Valeurs aberrantes

Etape	Éléments supprimés	Nature		Néant
		Colonnes	Lignes	
Suppression des colonnes vides à plus de 70%	132		✓	
Suppression des colonnes répétitives non informatives	X			X
Suppression de données non nécessaires (datetime , timestamp _t et image )	10		✓	
Suppression des colonnes similaires (_tags, _fr, etc.)	23		✓	
Suppression de doublons (se fait à la base du code)	337			✓
Suppression de lignes complètement vides	X			X
Suppression de lignes où les valeurs nutritives dépassent le 100g sauf l'énergie	910			✓
Suppression de lignes où les valeurs nutritives essentielles sont nulles (0 ou Nan)	317 394			✓
Suppression des valeurs nan de additives_n et ingredients_from_palm_oil_n	531 892			✓

✓ : propriété satisfaite, X : propriété non satisfaite,

# Démarche méthodologique d'analyse de données

Etape 1 – Pré-traitement

Etape 2 – Imputation

Etape 3 – Valeurs aberrantes

```
percentage_nan = (df_nan.isna().mean()*100)
percentage_nan.sort_values( ascending=False)

fiber_100g           29.240328
nutriscore_grade    23.501830
categories          14.015493
nova_group          11.916949
sugars_100g          2.450786
salt_100g            1.166715
proteins_100g        0.312247
product_name         0.299792
carbohydrates_100g  0.264533
fat_100g              0.259445
countries            0.046837
energy_100g           0.000000
pnns_groups_2        0.000000
| ingredients_from_palm_oil_n  0.000000
| additives_n          0.000000
| code                  0.000000
dtype: float64
```



- Imputation sur valeurs numériques sauf 'nova\_group'.
- Imputation :
  - # define imputer -- BayesianRidge Classifier
  - imputer = IterativeImputer (min\_value=0)

```
nutriscore_grade    23.501830
categories          14.015493
nova_group          11.916949
product_name         0.299792
countries            0.046837
sugars_100g          0.000000
salt_100g            0.000000
proteins_100g        0.000000
ingredients_from_palm_oil_n  0.000000
fiber_100g           0.000000
fat_100g              0.000000
energy_100g           0.000000
carbohydrates_100g  0.000000
additives_n          0.000000
pnns_groups_2        0.000000
code                  0.000000
dtype: float64
```

# Démarche méthodologique d'analyse de données

Etape 1 – Pré-traitement

Etape 2 – Imputation

Etape 3 – Valeurs aberrantes

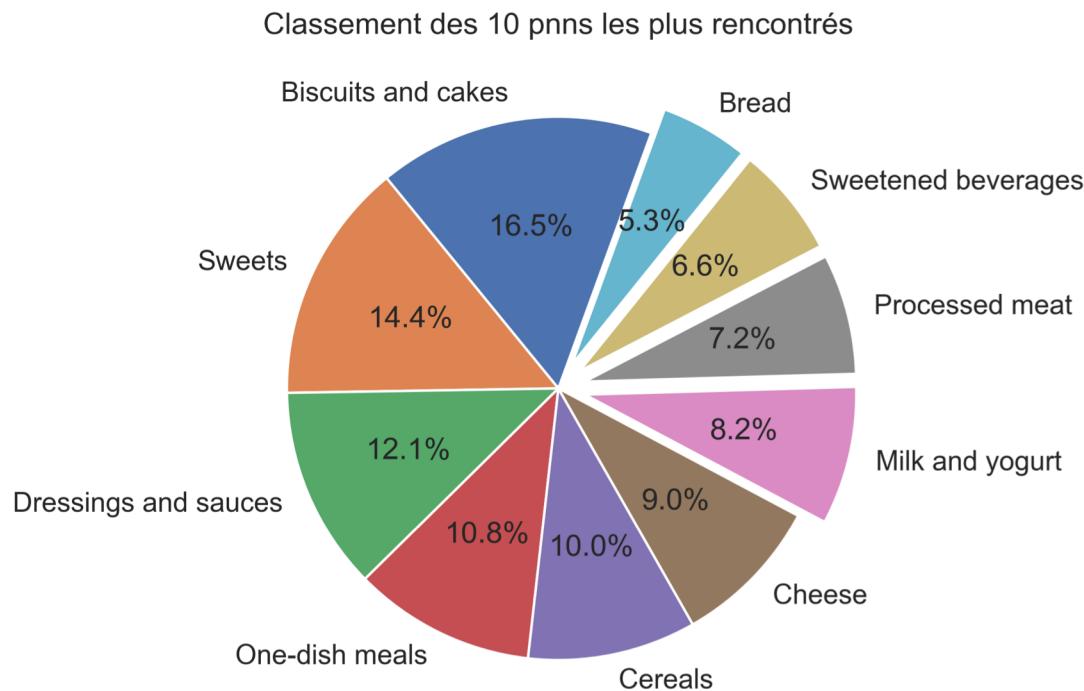
## ❖ Variables nutritionnelles:

- Filtrage après Imputation des valeurs liste\_100g > 100
- Filtrage après Imputation des valeurs liste\_100g < 0
- Filtrage après Imputation des valeurs energy\_100g > 3700 KJ / 100mL
- Suppression des valeurs aberrantes en utilisant l'écart inter quartile.
  - Comme la distribution des variables nutritives n'est pas normale, on se permet d'élargir les critères de suppression au delà de 1.5 \*IQR afin d'enlever le maximum de valeurs aberrantes.
- Après nettoyage, on garde 40% du jeu de données pour effectuer l'analyse.

## ❖ Variable qualitative « countries » :

- Utilisation d'un dictionnaire et ne garder que le premier élément de la colonne pour éviter les redondances.

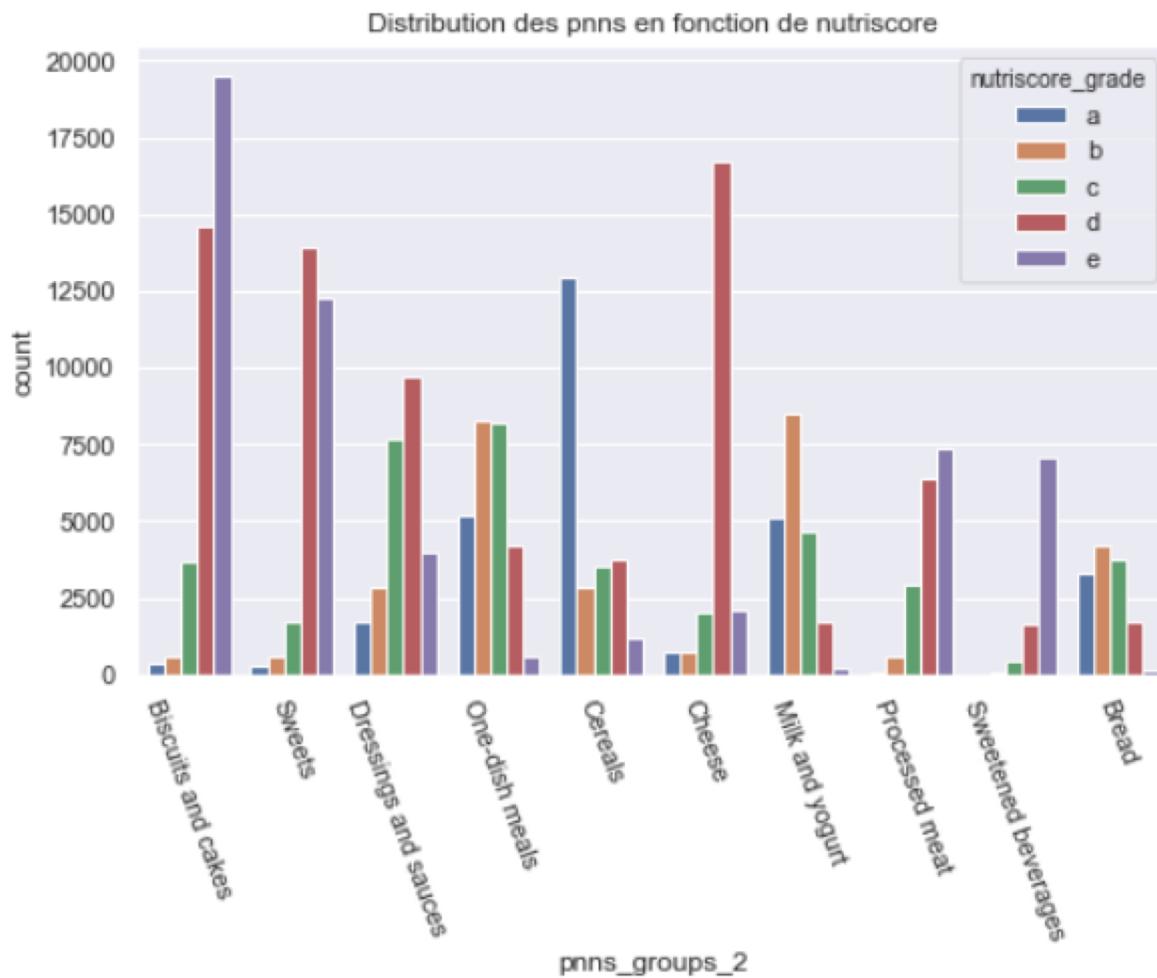
# Synthèse de l'analyse de données



- Sucrerie (36.5%),
- Produits laitiers(17.1%),
- Etc.

→ Que serait la qualité nutritionnelle des produits ?

# Synthèse de l'analyse de données



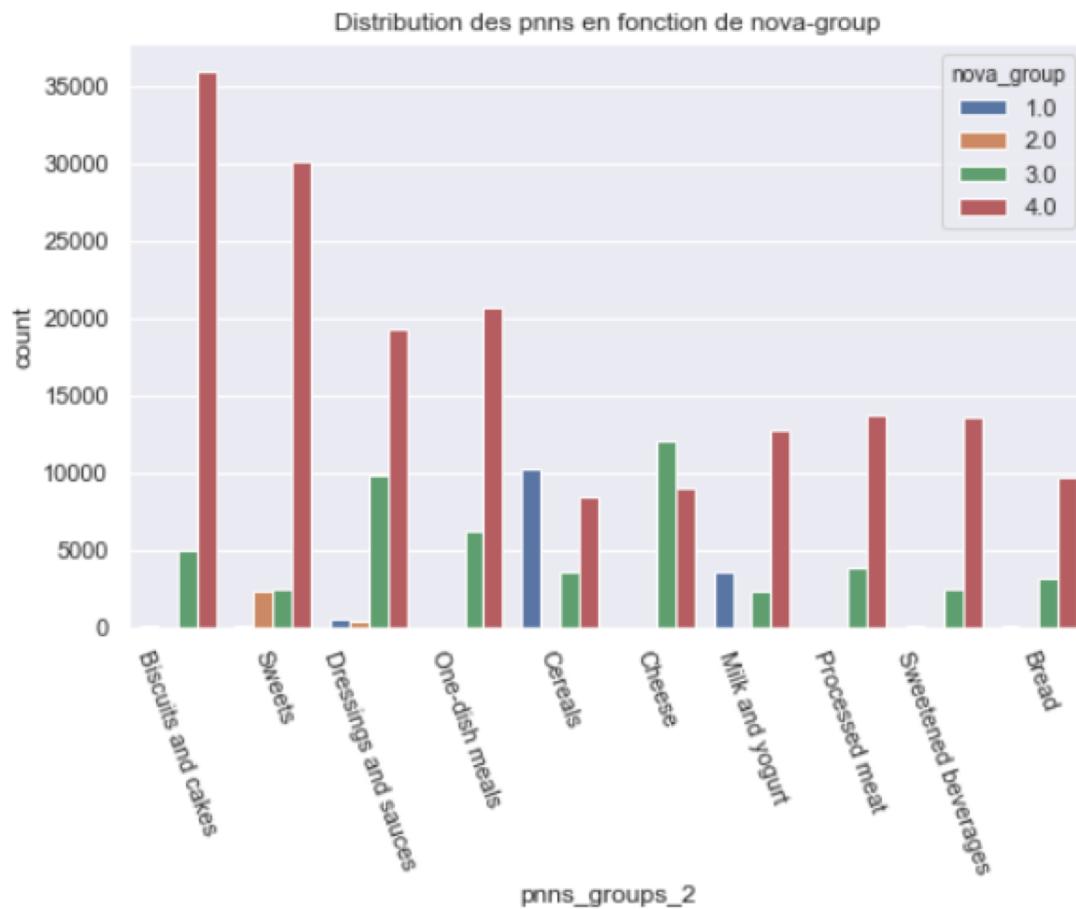
- La majorité des produits les plus répandus sont de mauvaise qualité nutritive.
- Les produits laitiers, les céréales et le pain possèdent de bons grades.

## 3- Synthèse de l'analyse de données

Hypothèses :

- Hypothèse 1: Les produits possèdent le bon score nutritionnels contiennent moins de produits transformés, additifs et l'huile de palme.
- Hypothèse 2: Les produits possèdent le mauvais score nutritionnel proviennent des sucreries, gras.

# Synthèse de l'analyse de données



- Les produits détenant plus de produits transformés sont ceux qui possèdent de mauvais nutriscore et inversement.
- Le graphe indique que les sucreries sont à la tête des produits transformés suivis par tout ce qui est gras (fruits secs, sauces).

→ Hypothèse 2

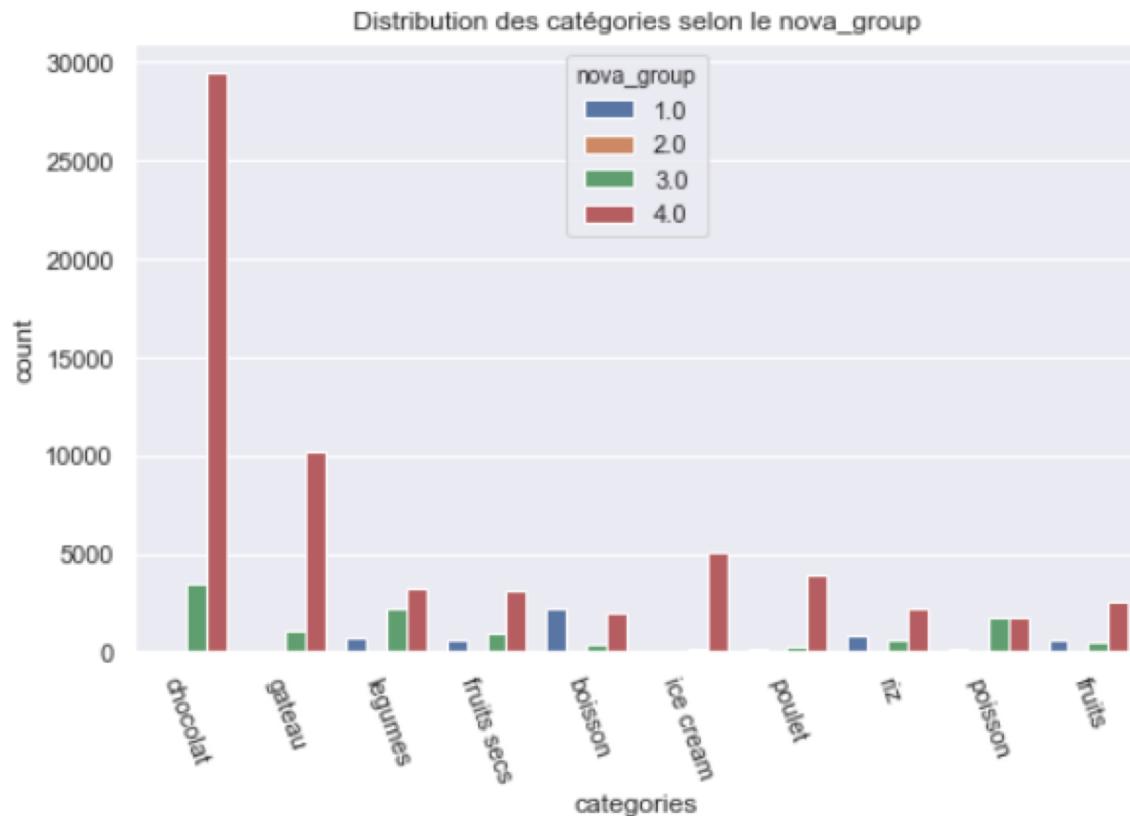
## 3- Synthèse de l'analyse de données

### Création de la fonction catégorie

*Au total, on a créé 16 : legumes - fruits- boisson – alcool – huile – beurre - fruits secs - sucrerie (gateau, chocolat, bonbon, ice cream) – pates - riz - fromage - creme - yaourt - lait – Oeuf - viande – poulet – poisson.*

Nb échantillons de catégories = 110368

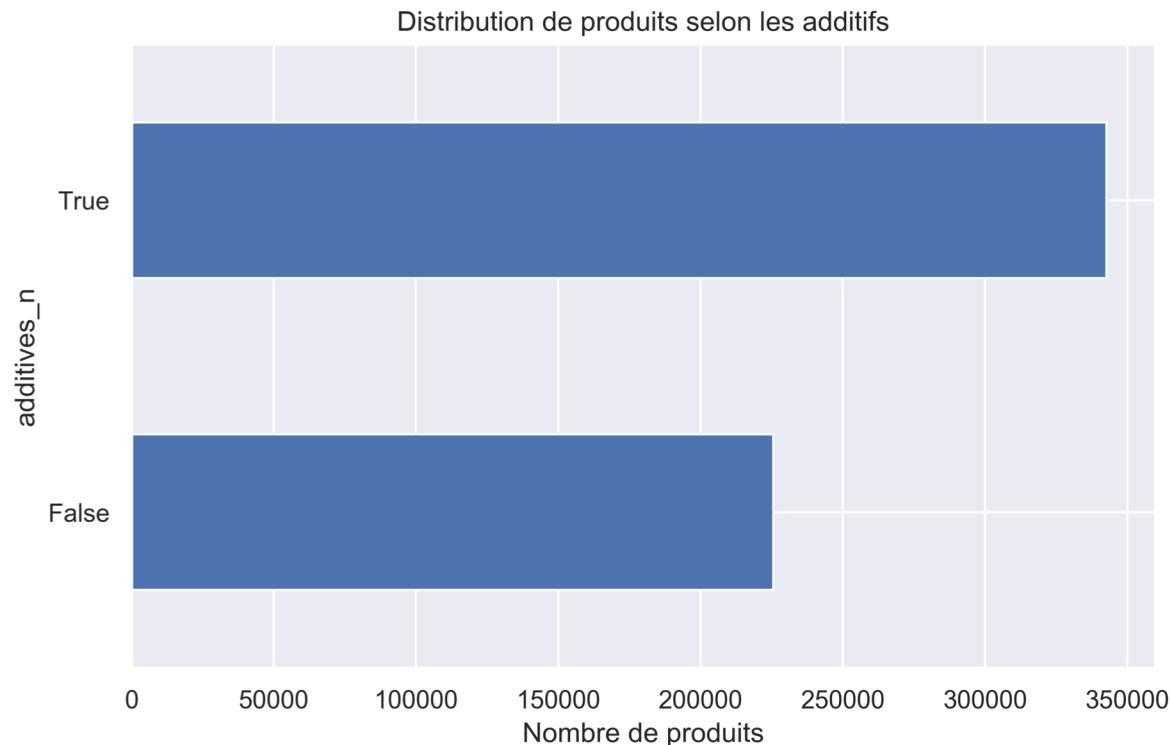
# Synthèse de l'analyse de données



- Résultats similaires qu'avec l'étude des pnns.

→ Hypothèse 2

# Synthèse de l'analyse de données

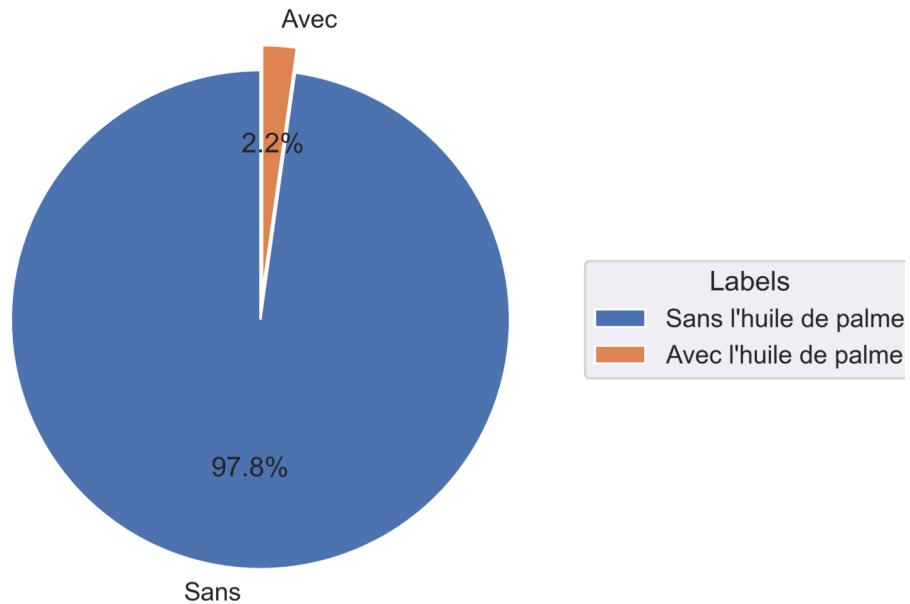


- Le nombre de produits avec additifs est supérieur au nombre de produits sans additifs.
- Le coefficient de corrélation de Pearson est proche de 0.4 donc additives-n et nova\_group sont corrélés linéairement.

→ Hypothèse 1

# Synthèse de l'analyse de données

Distribution des produits contenant l'huile de palme

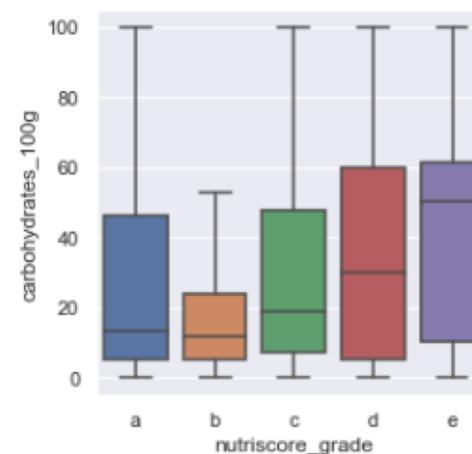
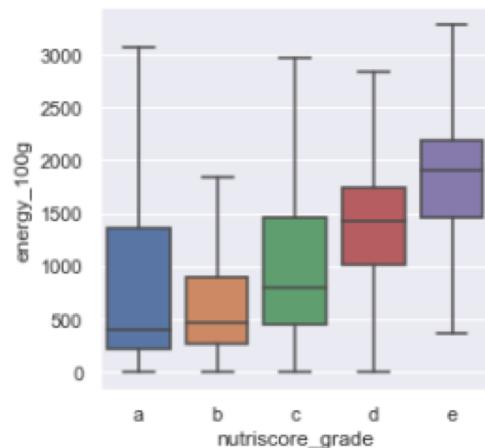
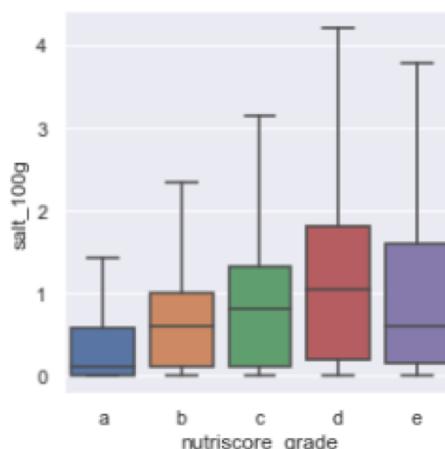
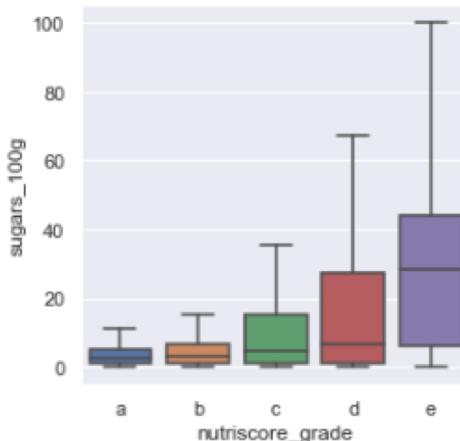


- La majorité des produits ne contiennent pas de l'huile de palme bien qu'ils soient de mauvaise qualité.

→ Ceci indique qu'il y a d'autres nutriments qui y sont responsables

# Synthèse de l'analyse de données

- Distribution des nutriments en fonction de nutrigrade:

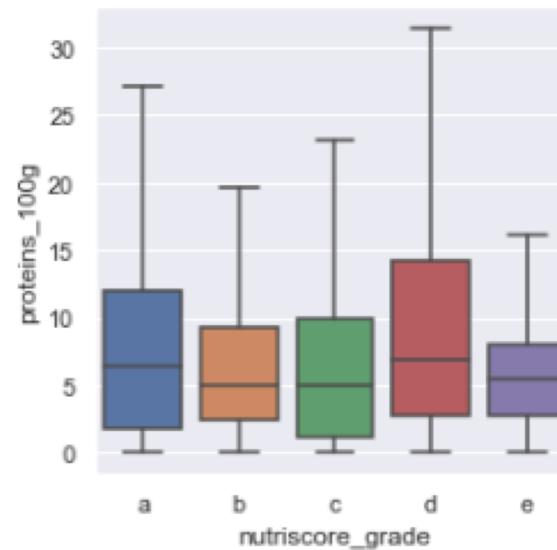
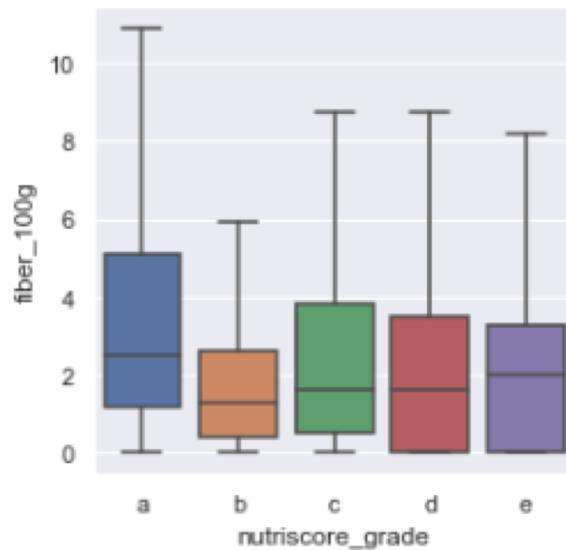


- Tout ce qui est sucré, gras, salé présente un mauvais indice nutrigrade.
- Les glucides présentent autant d'indices sains que mauvais ce qui confirme qu'il y a de bons et de mauvais glucides.

→ Hypothèse 2

# Synthèse de l'analyse de données

- Distribution des nutriments en fonction de nutrigrade:



- Les indicateurs nutritionnels qui présentent une bonne qualité comportent les fibres et les protéines.  
→ Hypothèse 2

# Synthèse de l'analyse de données

- **Vérification de la corrélation des ingrédients avec nutrigrade**

- Analyse multivariée explicative (ANOVA)

- Rapport de corrélation entre:

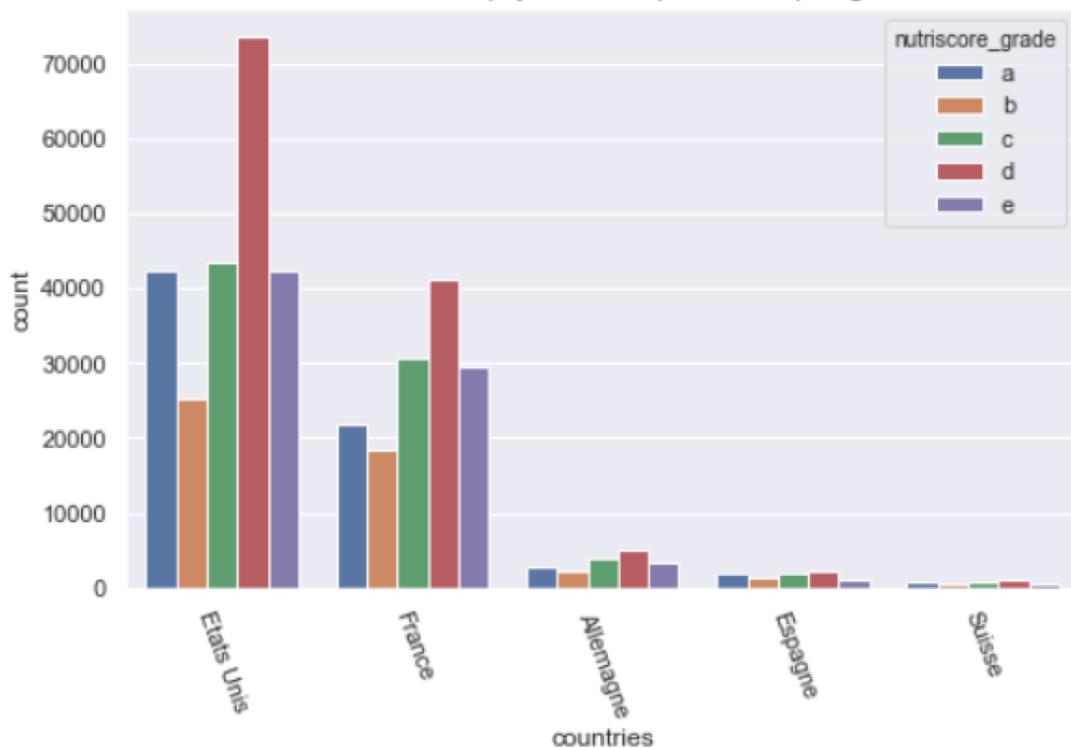
- nutriscore\_grade & fat\_100g = 0.21006534344926198,
      - nutriscore\_grade & carbohydrates\_100g = 0.058356750951968434,
      - nutriscore\_grade & sugars\_100g = 0.18132652479727404 ,
      - nutriscore\_grade & fiber\_100g = 0.036322071585198186
      - nutriscore\_grade & proteins\_100g = 0.014643269651213799,
      - nutriscore\_grade & salt\_100g = 0.03336684032895511.

→ Hypothèse 2

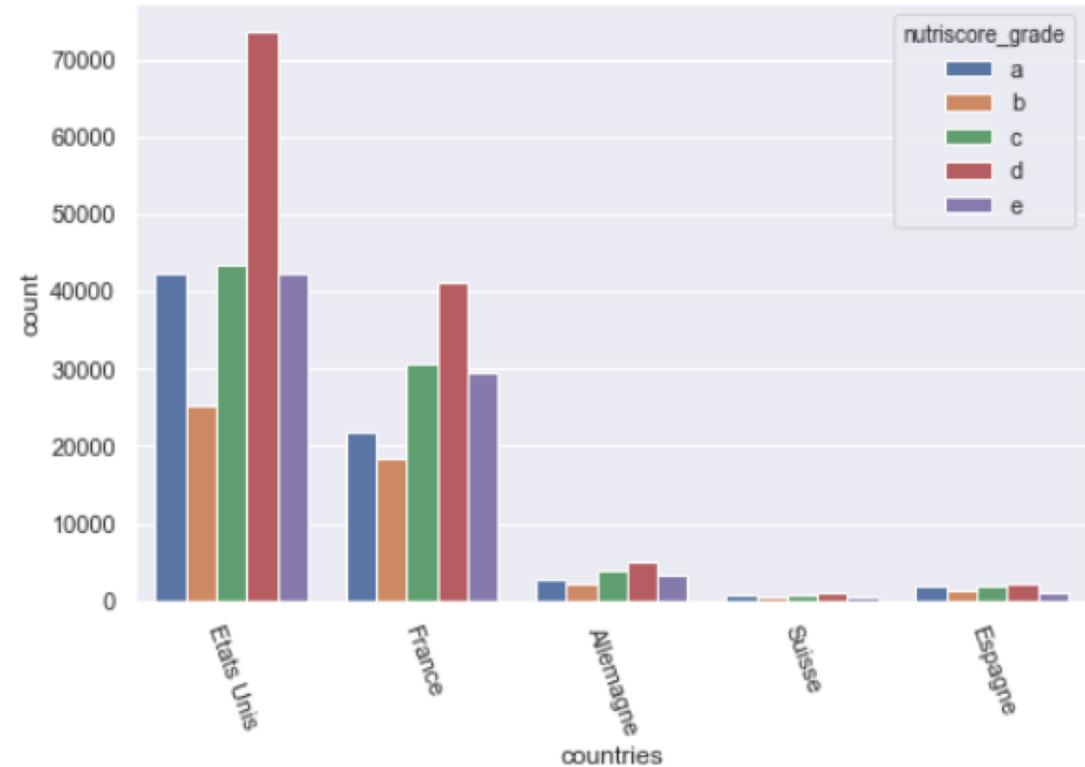
# Synthèse de l'analyse de données

- Pays produisant le plus de gras et de sucre

Distribution des pays selon les produits les plus gras



Distribution des pays selon les produits les plus sucrés



- Les Etats-Unis et la France sont à la tête des pays produisant le plus de produits alimentaires nocifs à la santé

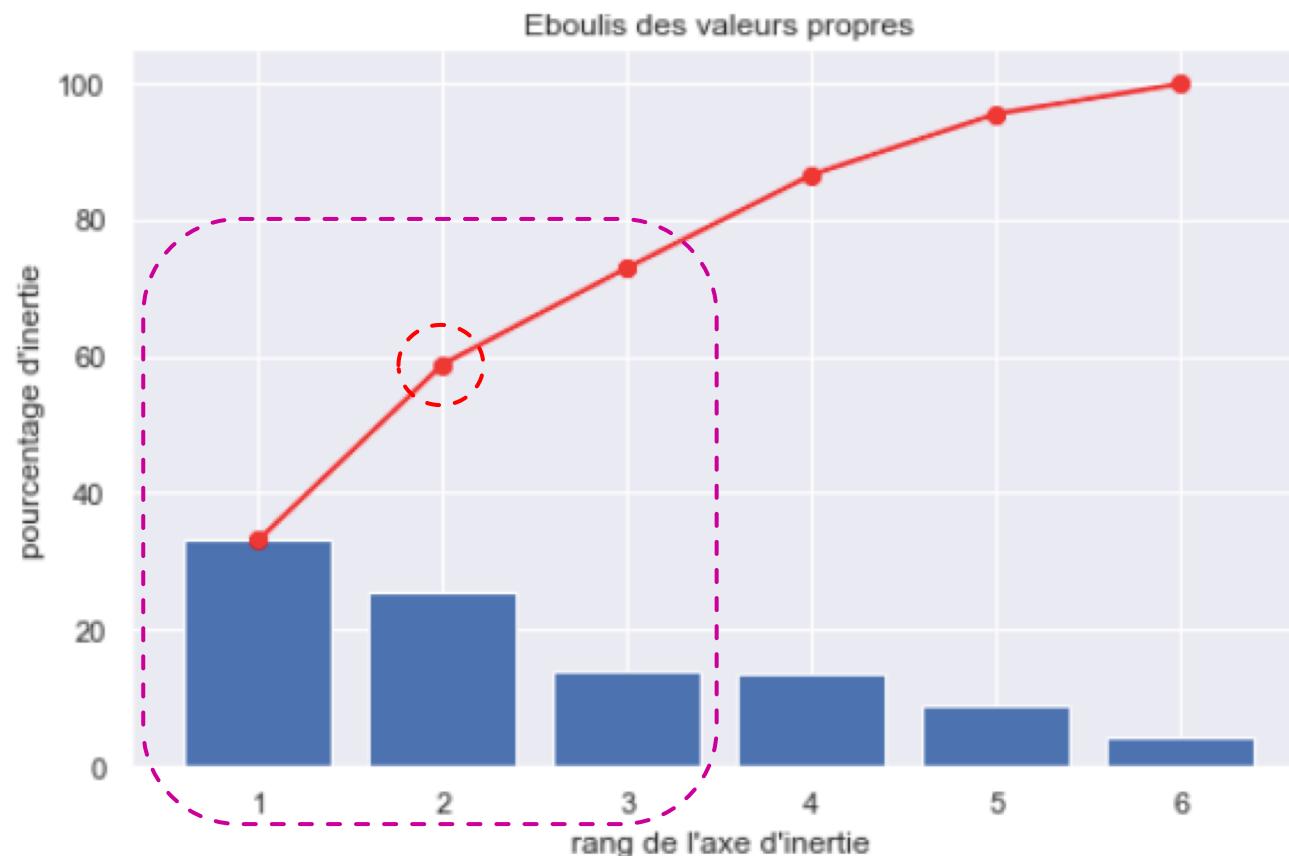
# Synthèse de l'analyse de données

## La méthode d'analyse descriptive (ACP)

Etape 1 – Ebloui des valeurs propres

Etape 2 - Cercle des corrélations

Etape 3 – Projection des individus



- Les trois premières composantes principales ( $F_1$ ,  $F_2$  et  $F_3$ ) dépassent 70% du pourcentage d'inertie.
- Le choix s'effectue la plupart du temps sur les deux premières composantes principales.
- Un taux d'inertie de 60% avec le premier plan factoriel.

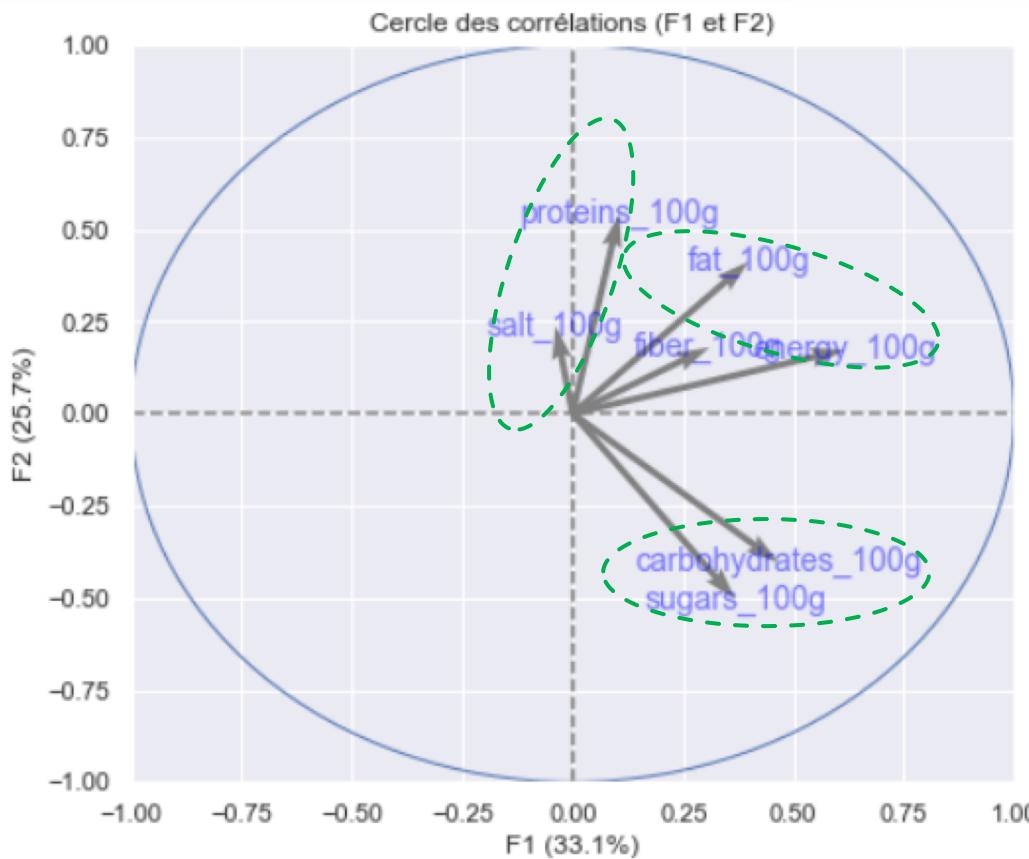
# Synthèse de l'analyse de données

## La méthode d'analyse descriptive (ACP)

Etape 1 – Ebloui des valeurs propres

Etape 2 - Cercle des corrélations

Etape 3 – Projection des individus



- Les flèches entre les paires (glucides/sucres) ou (protéines/sels) ou (graisse / energie) sont dans le même sens et sont proches.
  - Donc ces variables sont corrélées positivement.
- Les flèches des fibres sont plus ou moins perpendiculaires à celles du sel.
  - Donc ces variables ont une corrélation nulle.

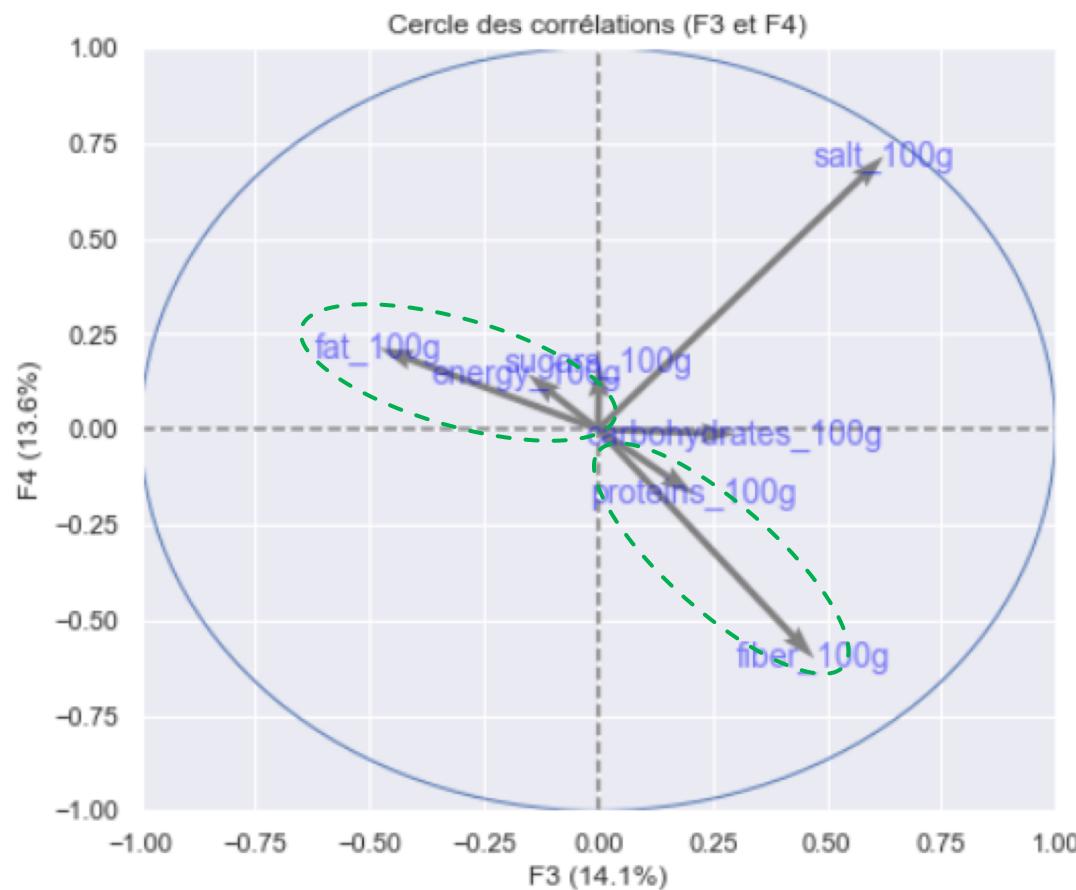
# Synthèse de l'analyse de données

## La méthode d'analyse descriptive (ACP)

Etape 1 – Ebloui des valeurs propres

Etape 2 - Cercle des corrélations

Etape 3 – Projection des individus



- Les flèches entre la teneur en Fibre et en Protéine contre la teneur en Energie et Graisse sont de sens opposé.

→ Donc ces variables sont corrélées négativement.

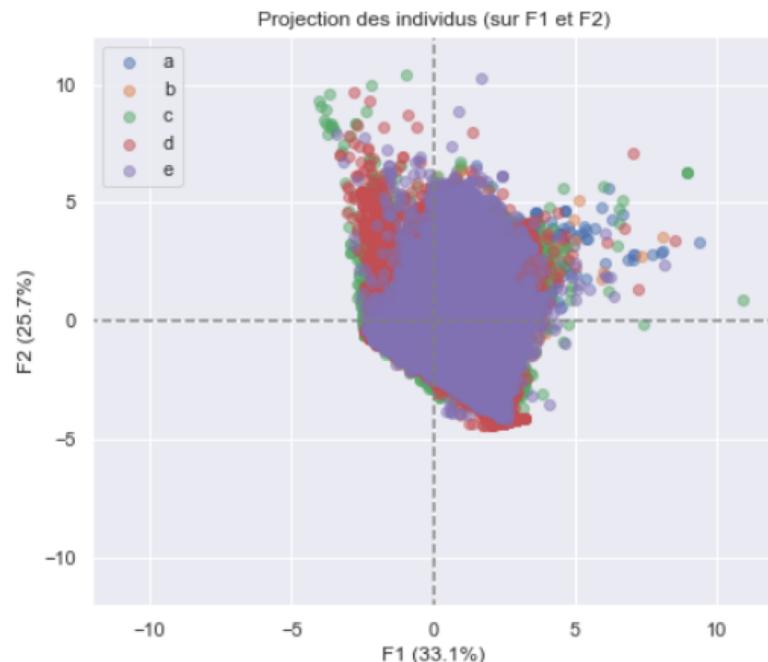
# Synthèse de l'analyse de données

## La méthode d'analyse descriptive (ACP)

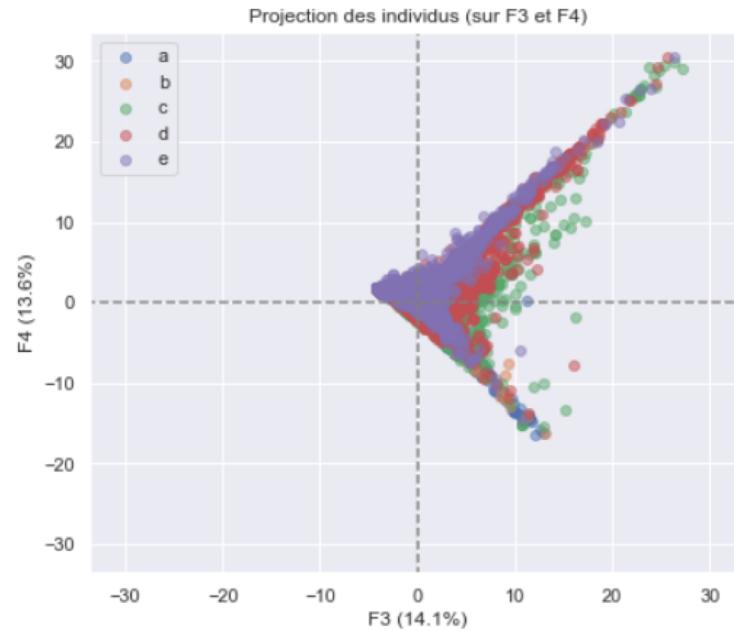
Etape 1 – Ebloui des valeurs propres

Etape 2 - Cercle des corrélations

Etape 3 – Projection des individus



- Plan P1 : La majorité des valeurs nutritives entre e puis d et c .
  - la valeur b autour du fibre et des protéines qui sont effectivement des nutriments plus sains.

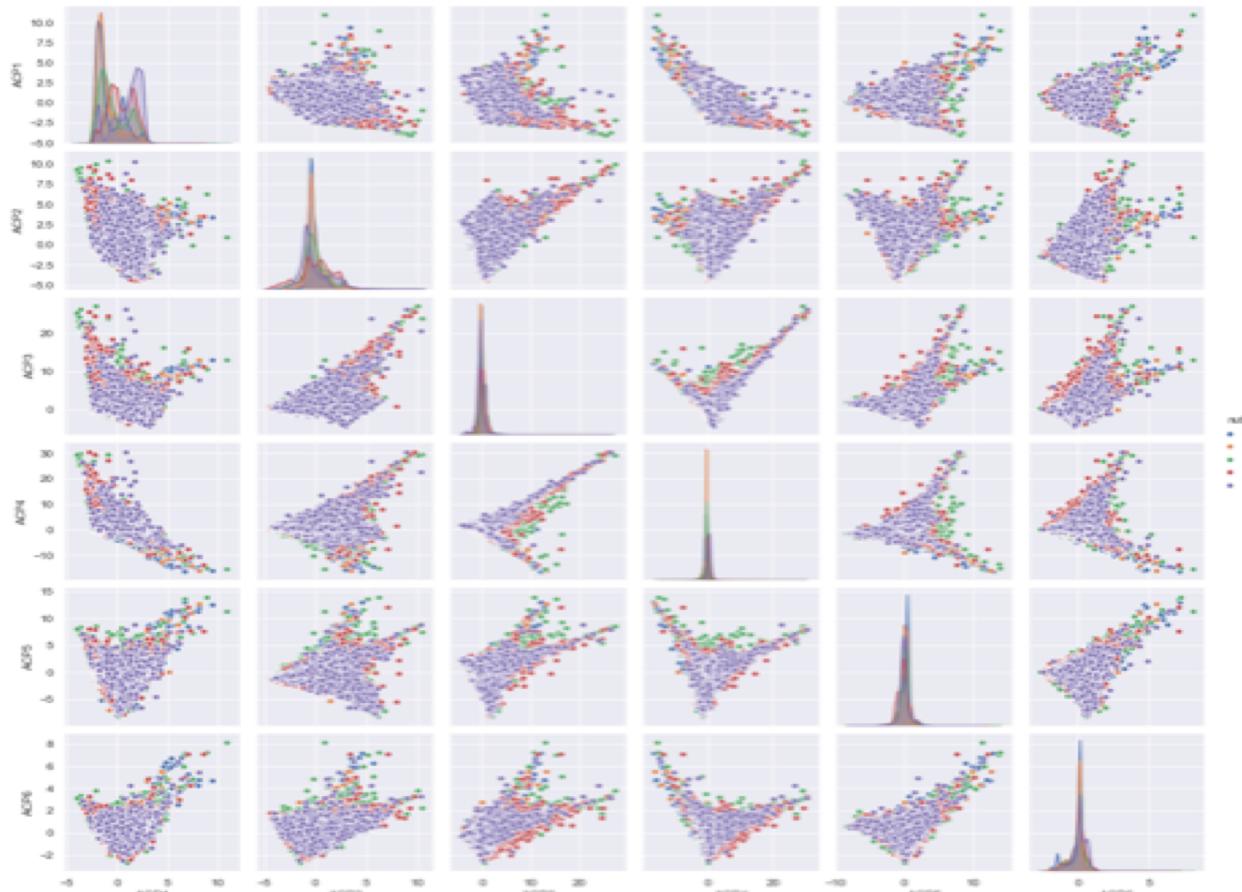


- Plan P2 : L'indice nutritif e est le plus répandu suivi par d (le gras, l'énergie, le sucre et les glucides).
  - La fibre présente des indices nutritifs entre a et b.
  - Le sel a l'indice d

# Synthèse de l'analyse de données

La méthode d'analyse descriptive (ACP)

- **Visualisation avec pairplot**



- Ce pairplot des composantes de l'ACP illustre la formation de groupes selon les cinq catégories nutritionnelles "nutriscore\_grade".
- Le groupe avec la classification e est le plus répandu, suivi de d et c

MERCI DE VOTRE ATTENTION