PROJET 3

Préparer des données pour un organisme de santé publique

Yann Héreng

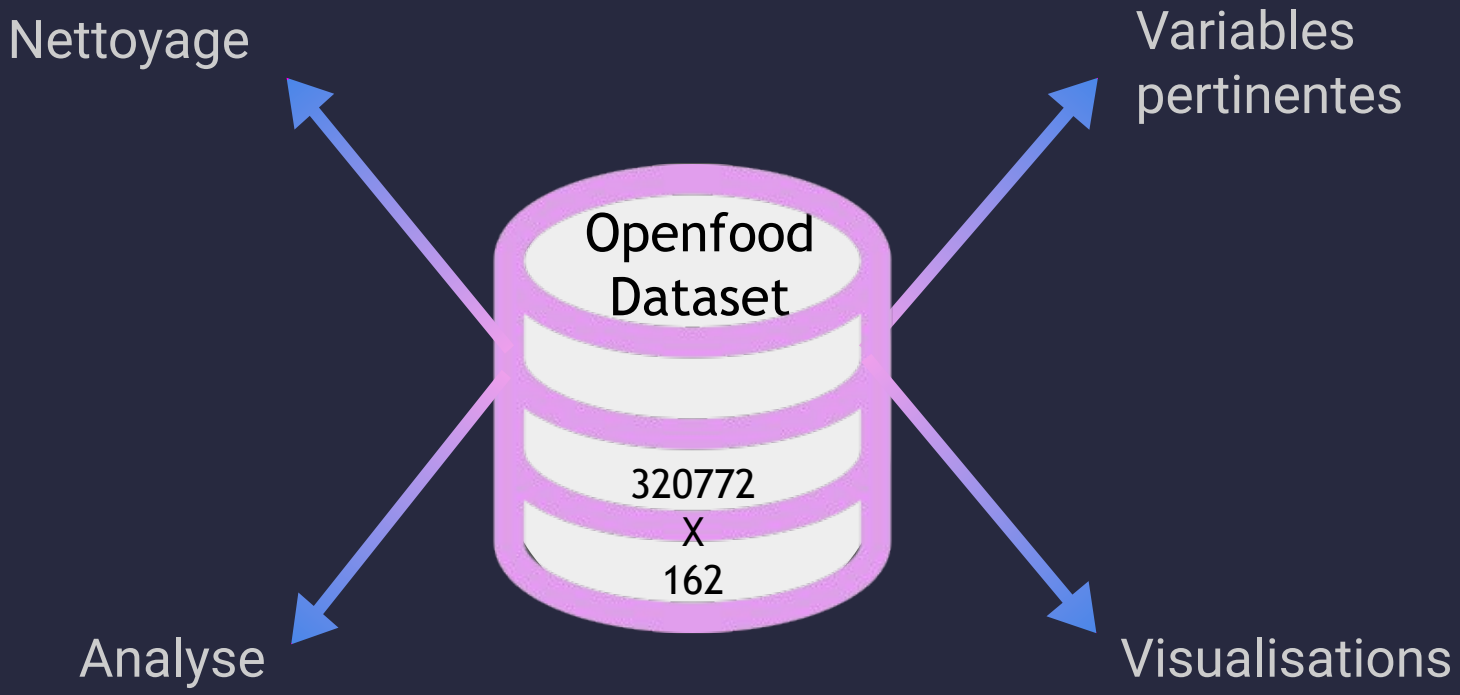
CONTEXTE

Santé publique France

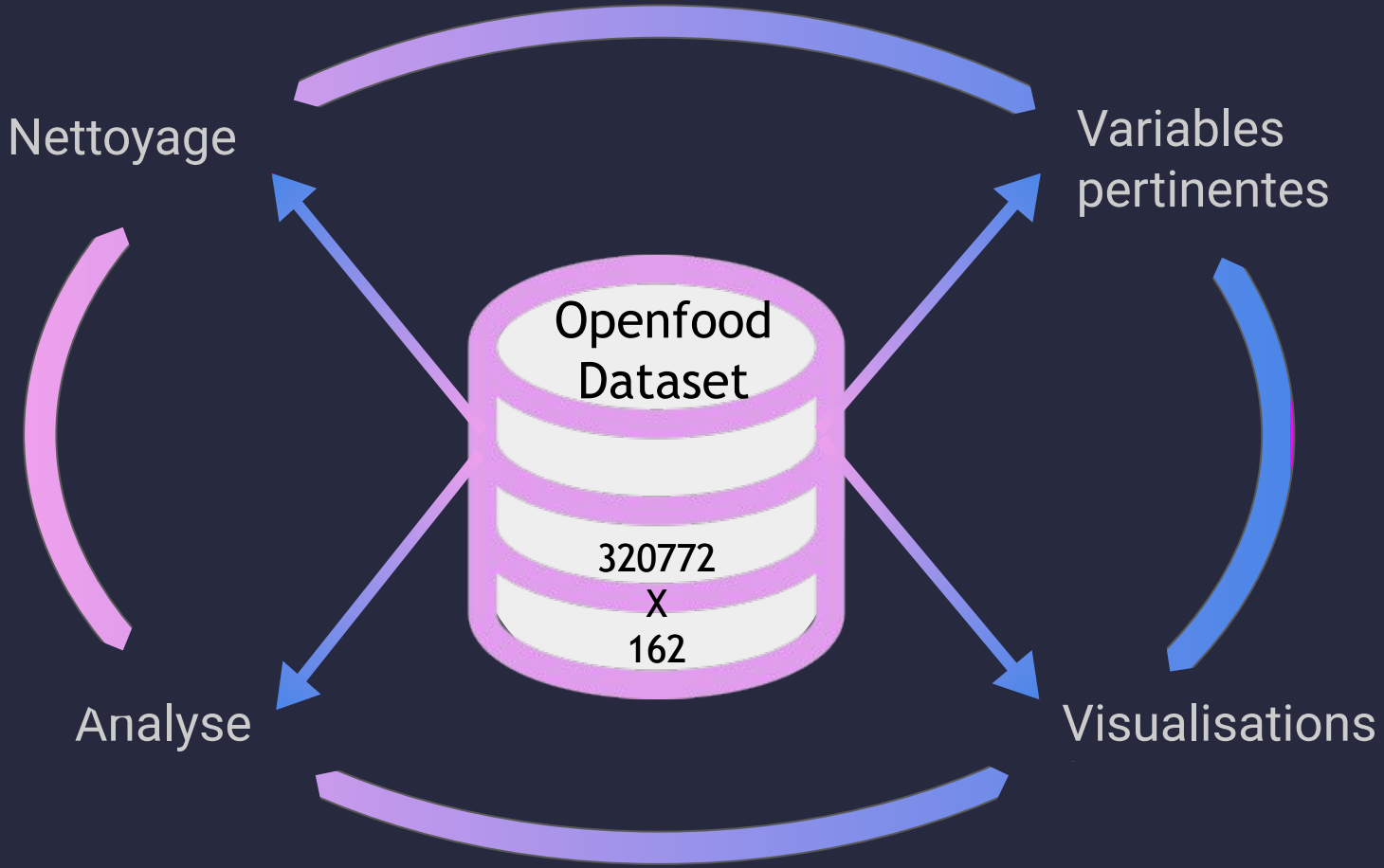
Appel à projet data

Plus accessible

OBJECTIFS


Nettoyage
Variables pertinentes
Openfood Dataset
320772
X
162
Analyse
Visualisations

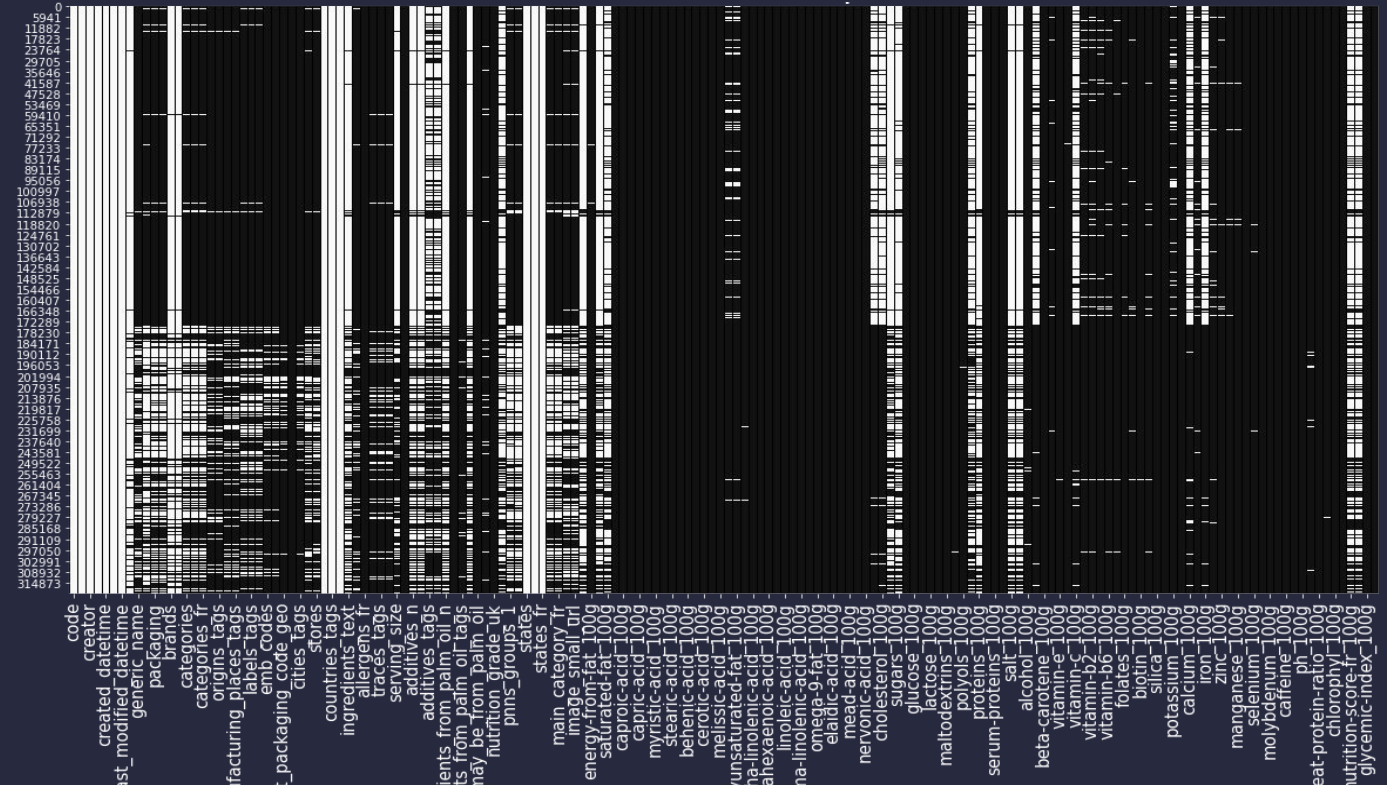
OBJECTIFS


Nettoyage
Variables pertinentes
Openfood Dataset
320772
X
162
Analyse
Visualisations

NETTOYAGE

- Valeurs manquantes
- Doublons
- Variables d’interêt
- Valeurs abérrantes
- Format adéquat

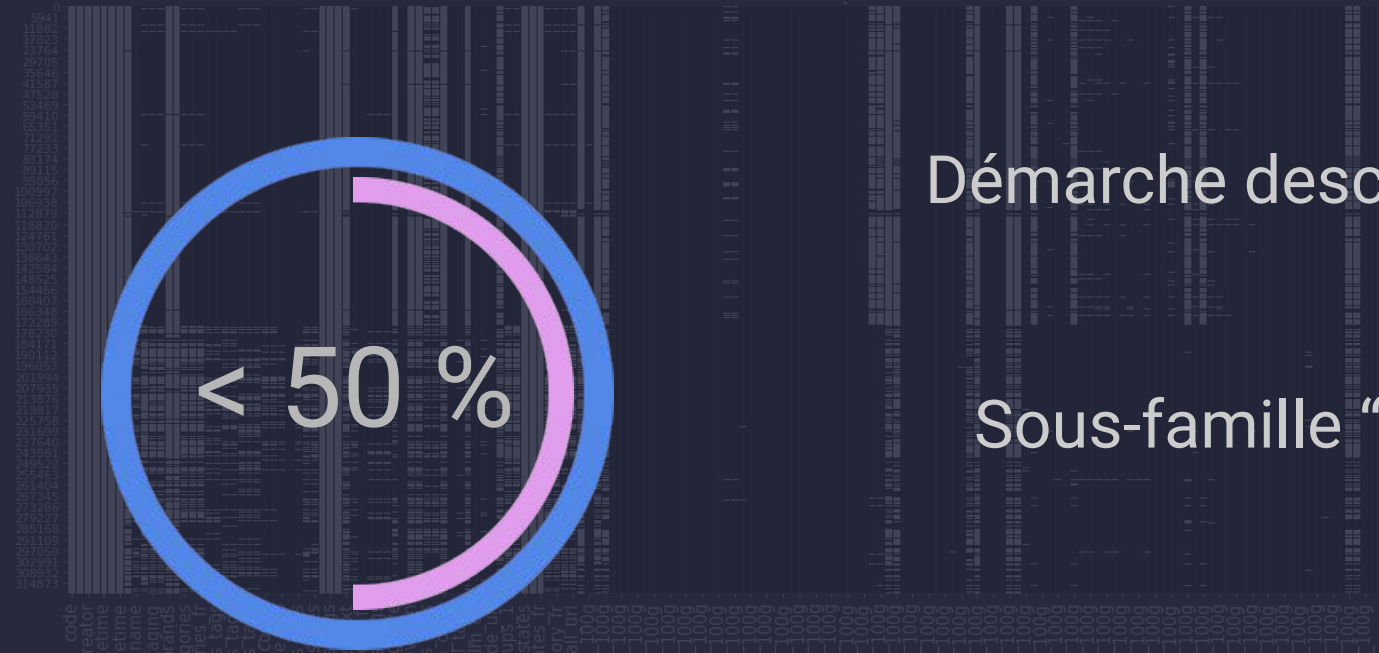
Valeurs manquantes



Valeur présente

Valeur manquante

Valeurs manquantes

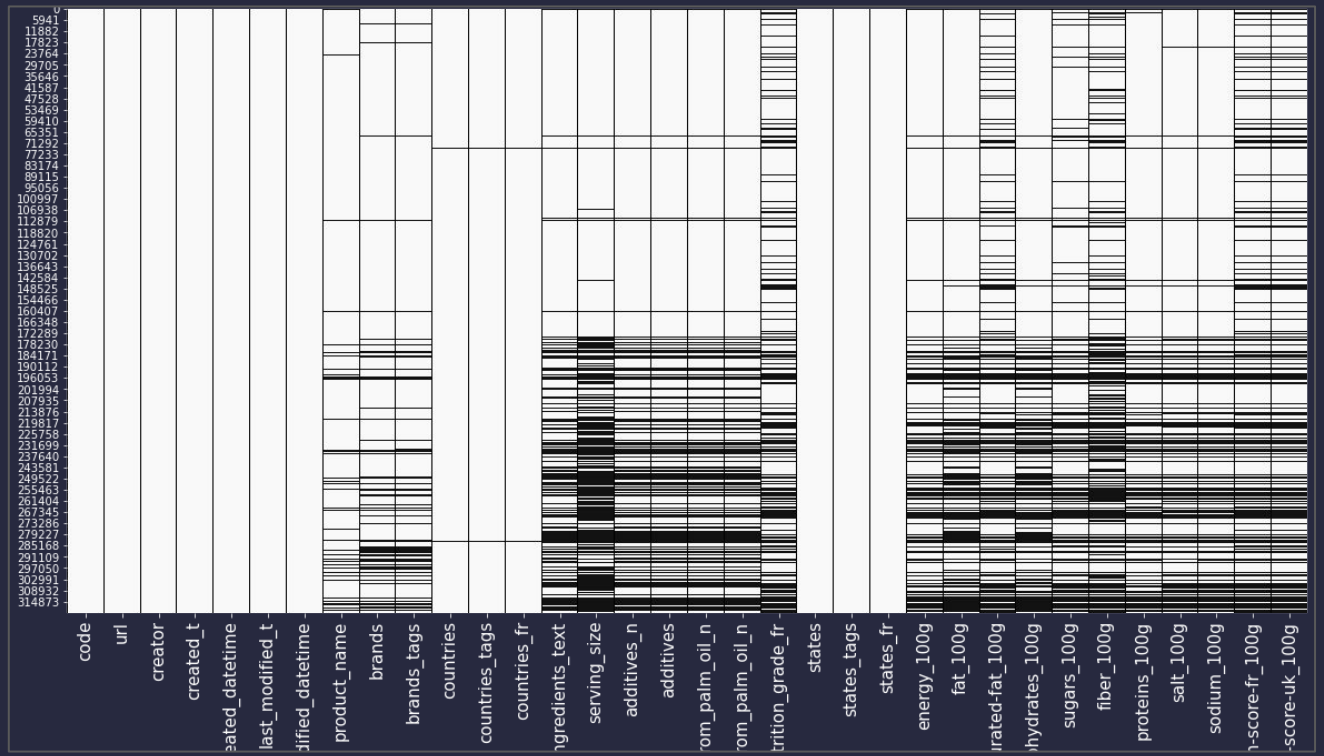


Démarche descriptive

Sous-famille “niche”

Imputations

Valeurs manquantes



Valeur présente

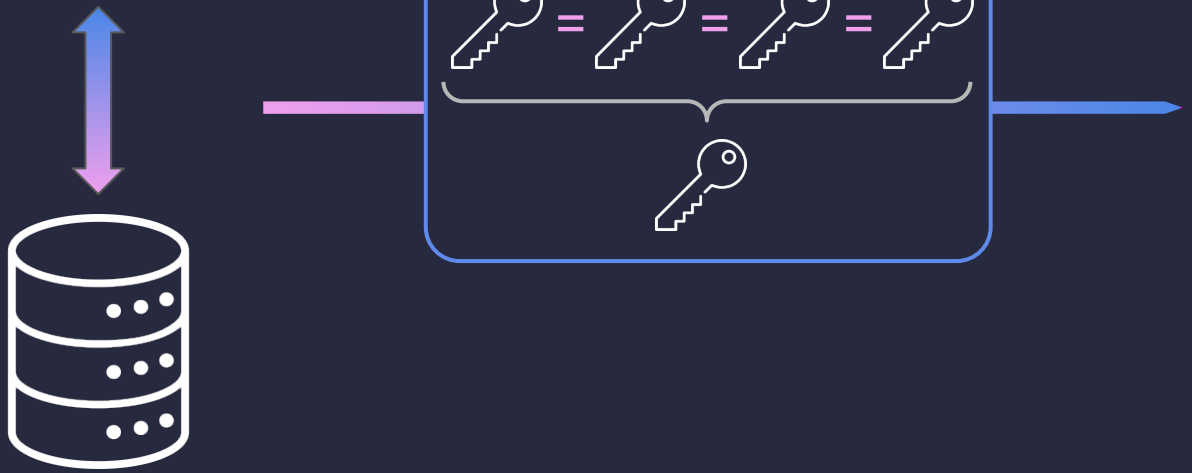
Valeur manquante

Valeurs manquantes

Doublons

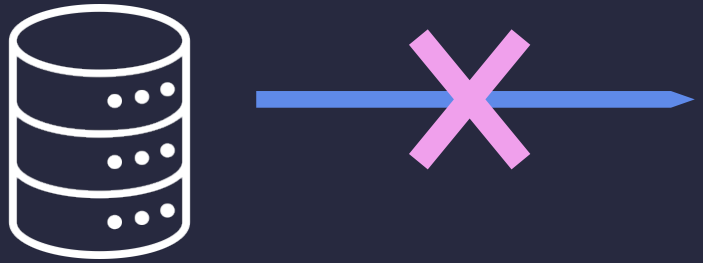
Clé primaire

product_name
brands
energy_100g



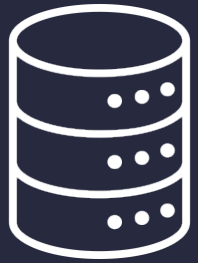
27890 individus

Variables d’interêt



- Redondance linguistique
- Liens vers d’autres sources
- Création de la donnée
- Pauvre en information
- Très fortement corrélées

Variables d’interêt



- Redondance linguistique
- Liens vers d’autres sources
- Création de la donnée
- Pauvre en information
- Très fortement corrélées

34 Variables

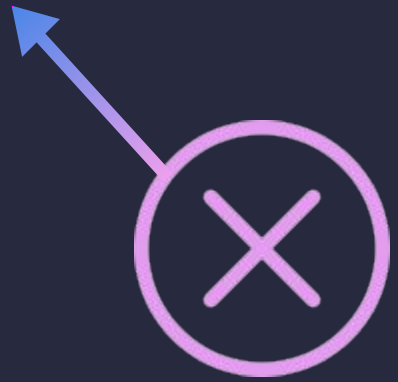
22 Variables

Réintégration future

Valeurs abérrantes

Valeurs abérrantes

Définition de
la variable



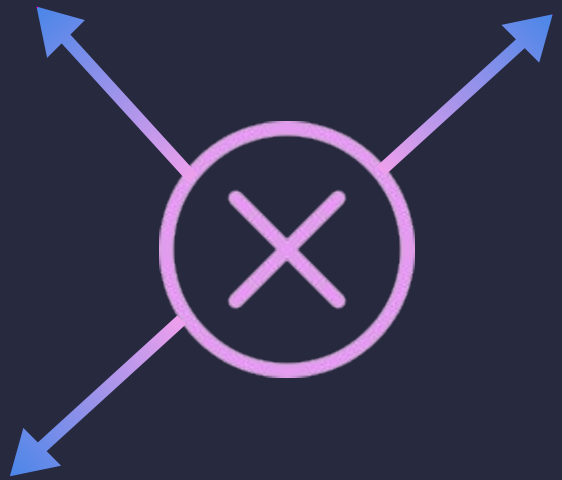
Valeurs abérrantes

Définition de
la variable

Valeurs filles

Valeurs abérrantes

Définition de la variable



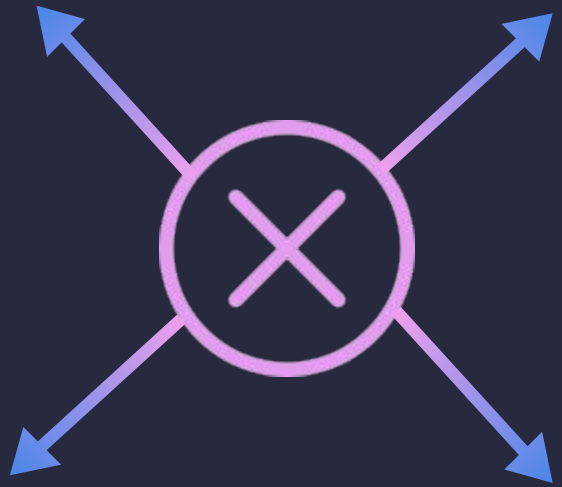
Valeurs filles

Données métier

Valeurs abérrantes

Définition de la variable

Valeurs filles



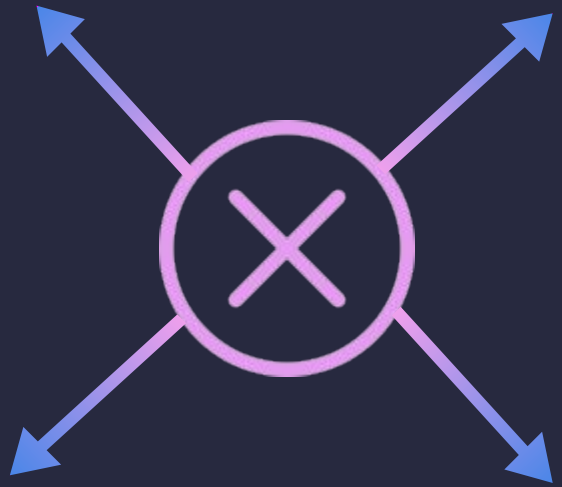
Données métier

Règle des quantiles

Valeurs abérrantes

Définition de la variable

Valeurs filles



Données métier

Règle des quantiles

46781 individus

Format adéquat

/%a
#A Données non structurées

Erreurs de traduction

Codage adéquat

Format adéquat

/%a
#A Données non structurées

Erreurs de traduction

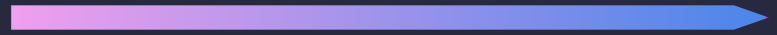
Codage adéquat

Format adéquat

/%a
#A Données non structurées

Erreurs de traduction

Codage adéquat

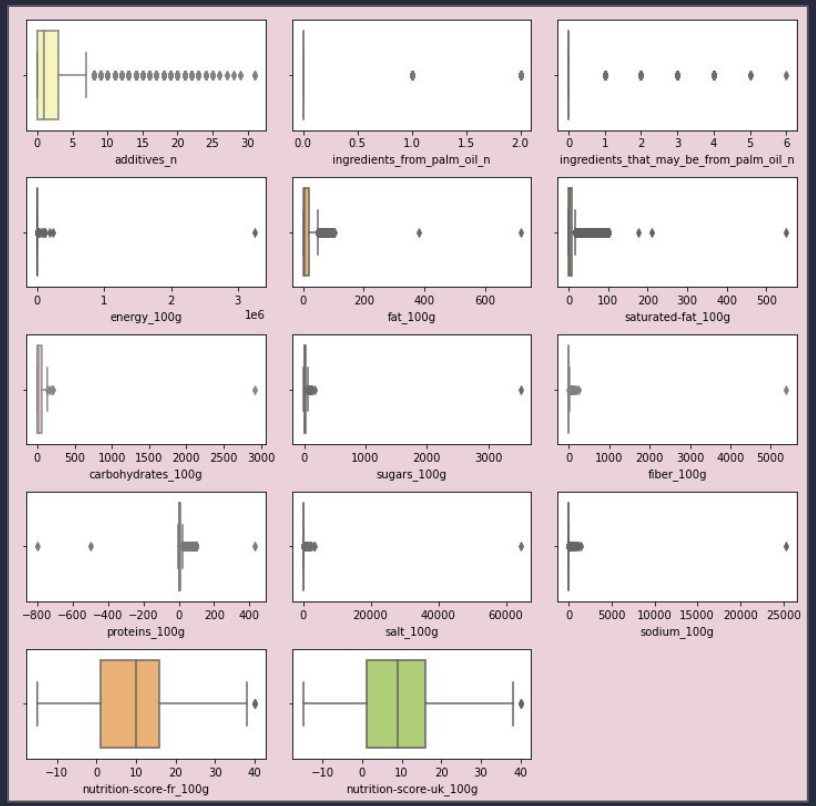


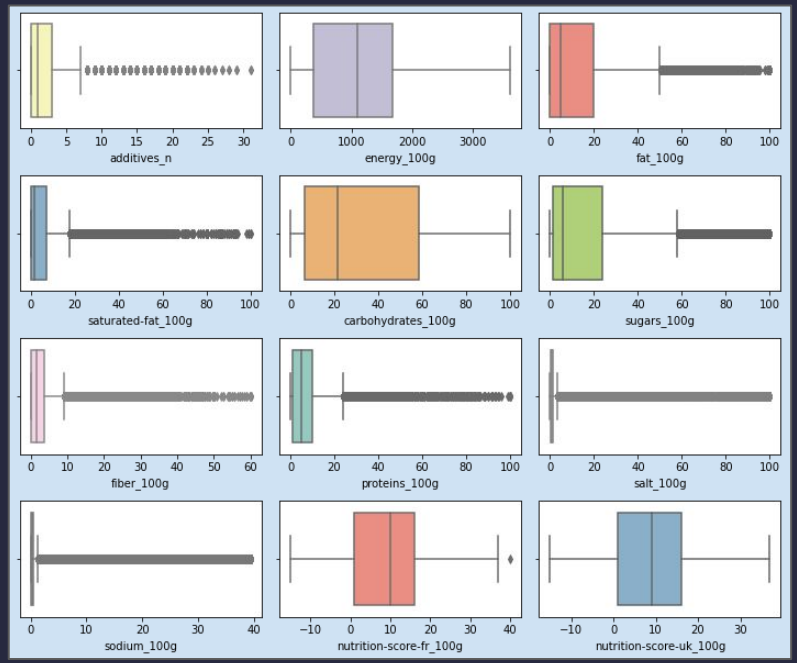
Données homogènes

Ressources optimisées

Bilan Nettoyage

Variables quantitatives





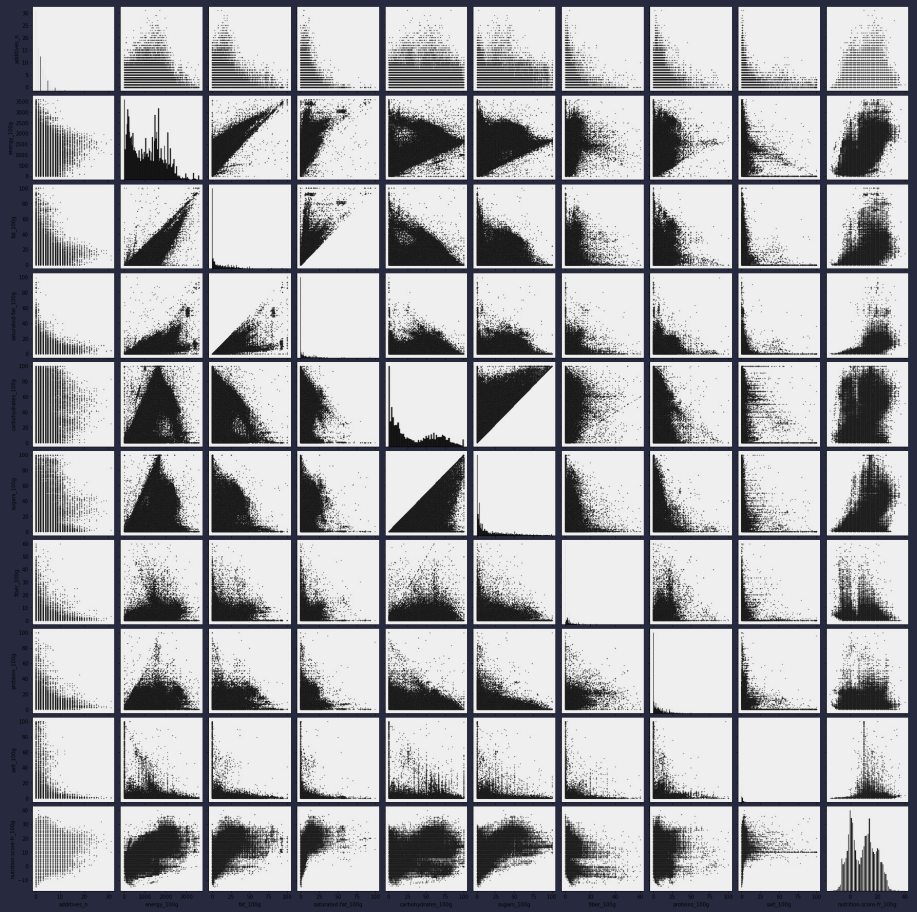
Analyse et Visualisation

Corrélations

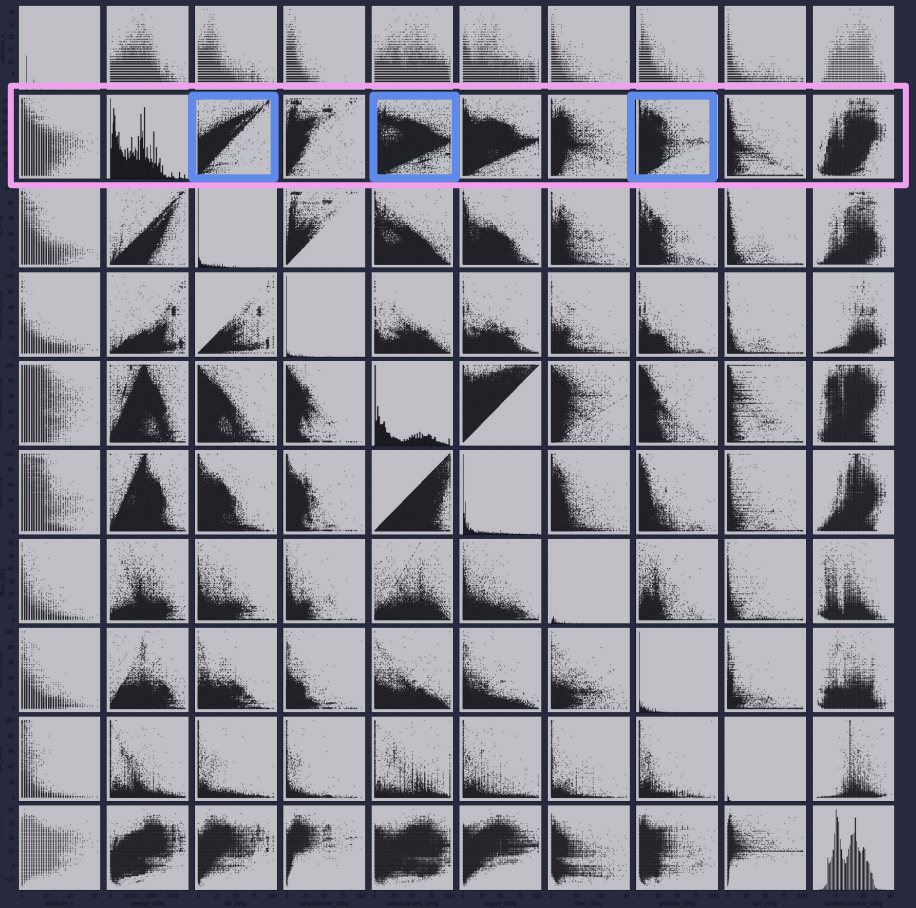
Analyse en Composante Principales

Analyse des variances

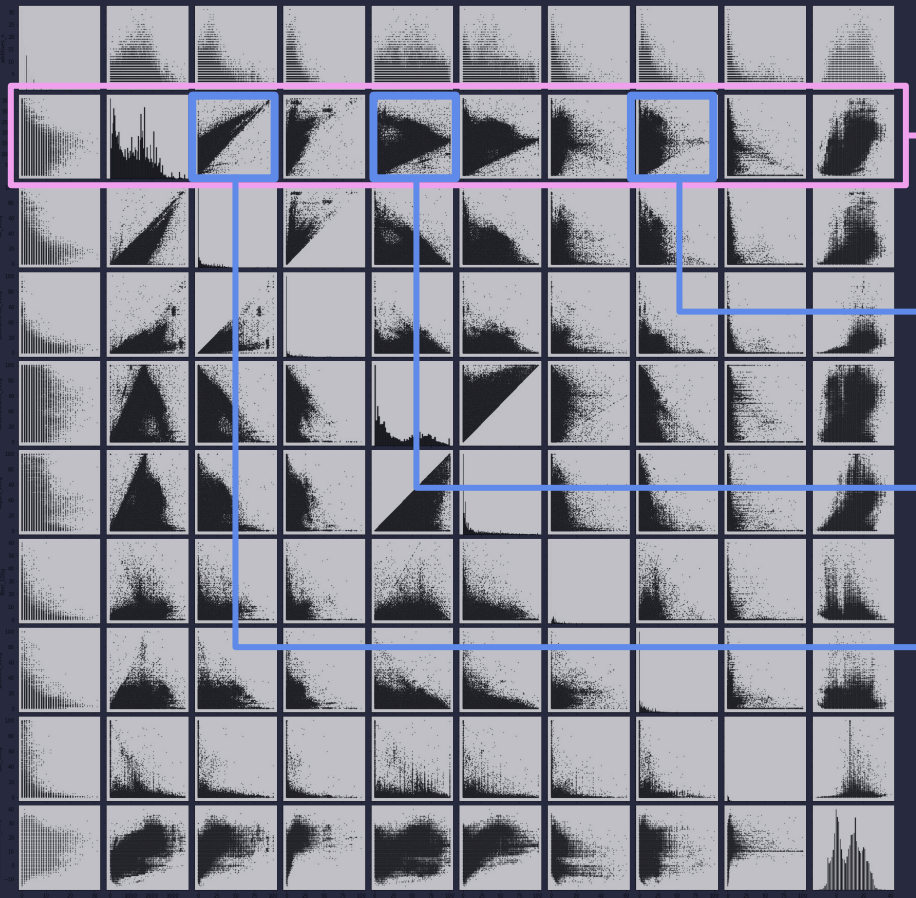
Analyse des corrélations - Régression linéaire



Analyse des corrélations - Régression linéaire



Analyse des corrélations - Régression linéaire



Energy_100g → Y

Proteins_100g → X1

Carbohydrates_100g → X2

Fat_100g → X3

$$Y = a\,X1 + b\,X2 + c\,X3$$

Analyse des corrélations - Régression linéaire

$$Y = a\,X1 + b\,X2 + c\,X3$$

→ [36.2 15.9 16.1]

→ [37.6 16.7 16.7]

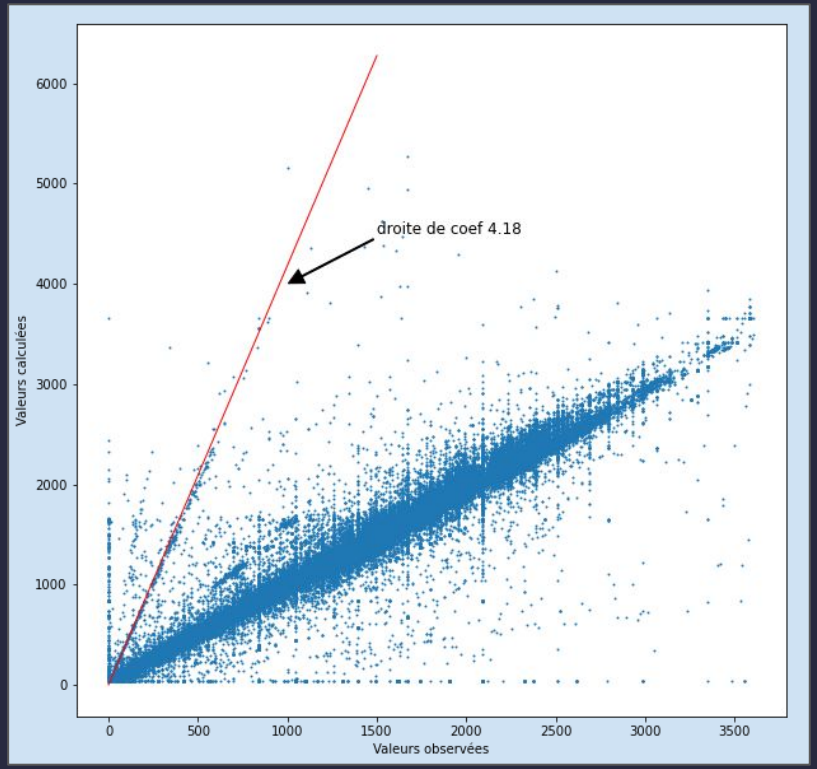
Analyse des corrélations - Régression linéaire

$$Y = a\,X1 + b\,X2 + c\,X3$$

→ [36.2 15.9 16.1]

→ [37.6 16.7 16.7]

Analyse des corrélations - Régression linéaire



1kCal = 4.18 kJ → Erreur d'unité

Analyse des corrélations - Relation linéaire

[37.6 16.7 16.7]

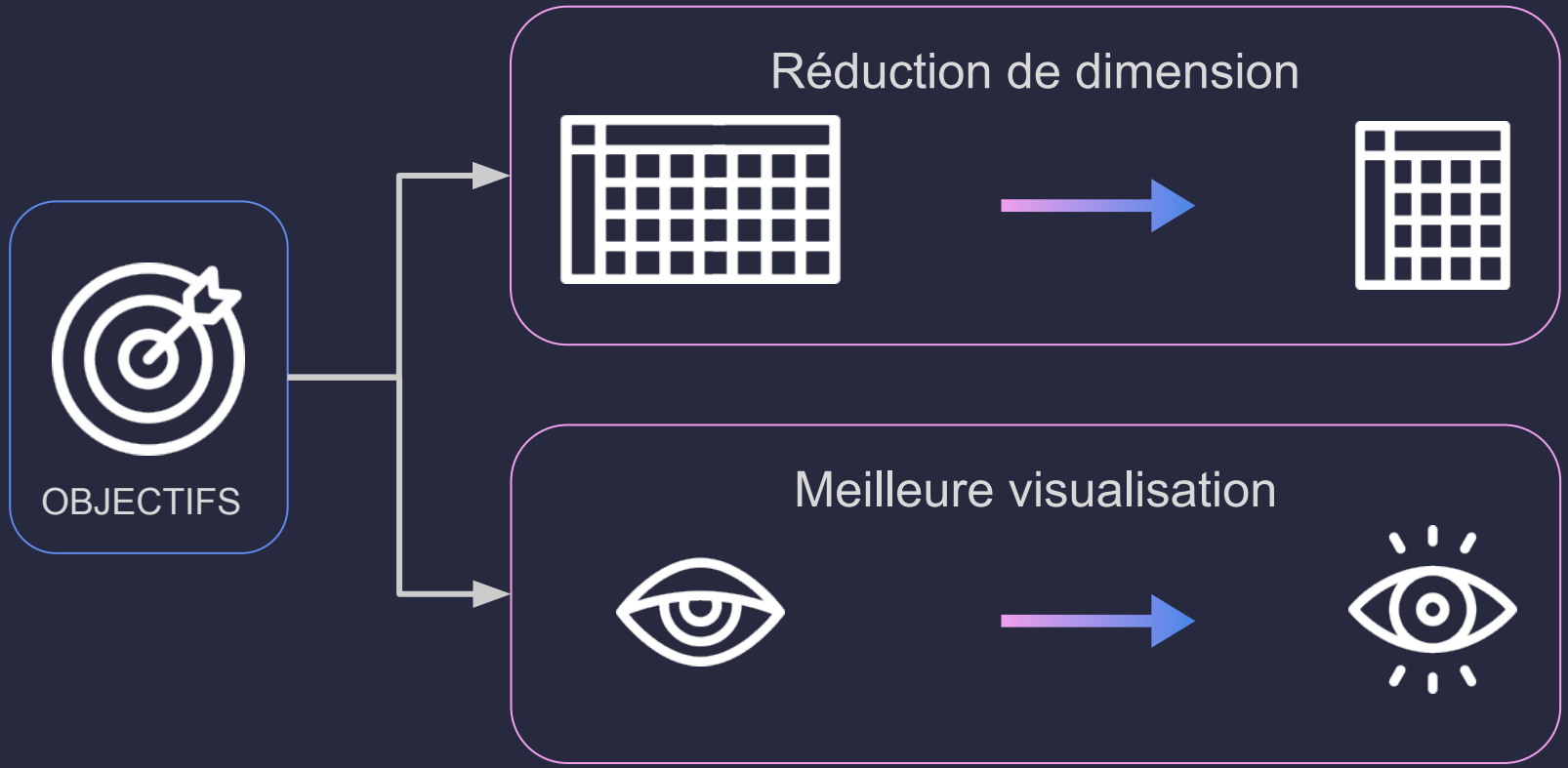
Valeurs abérrantes

seuil: ±25% → 12427 individus

Imputation → 35143 individus

Analyse en Composantes Principales

Analyse en Composantes Principales



Analyse en Composantes Principales

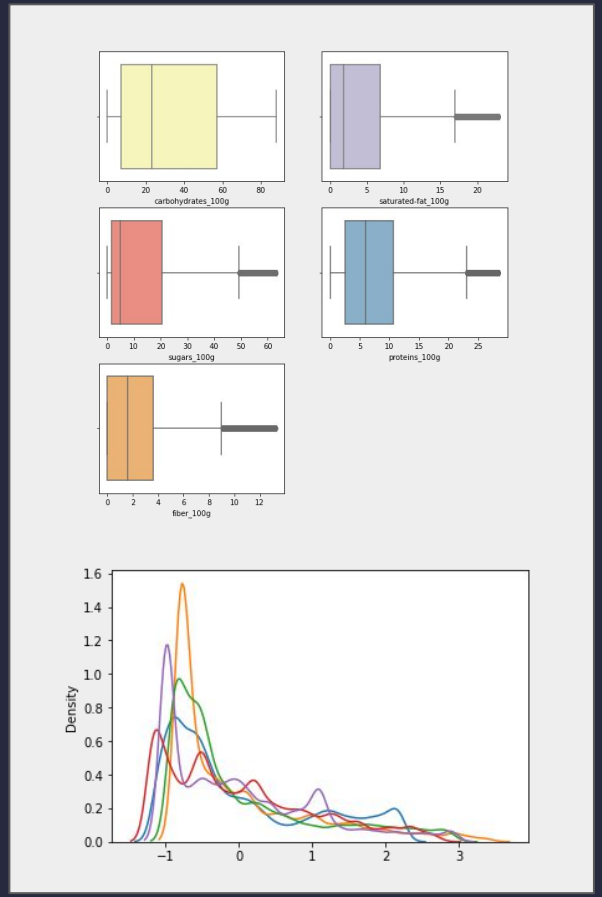
Interprétation des axes → *Choix des variables*

Valeurs extrêmes → *Valeurs < Q98*

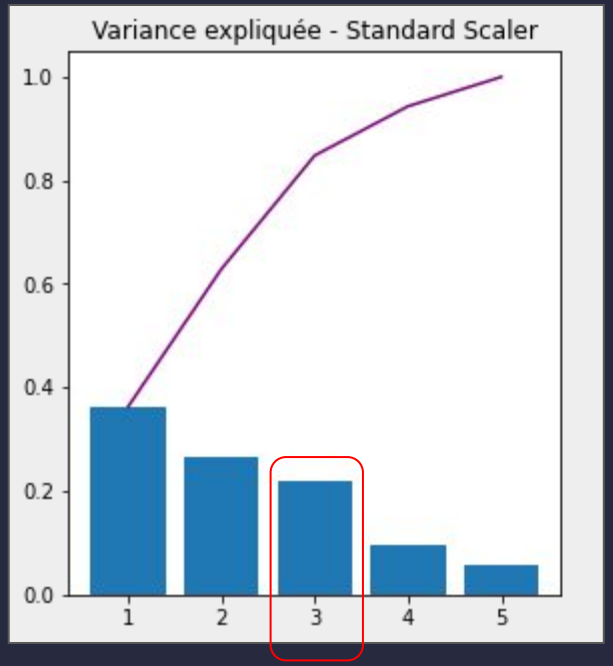
Centrage

Facteur d'échelle

→ *StandardScaler*



Analyse en Composantes Principales



Axes principaux

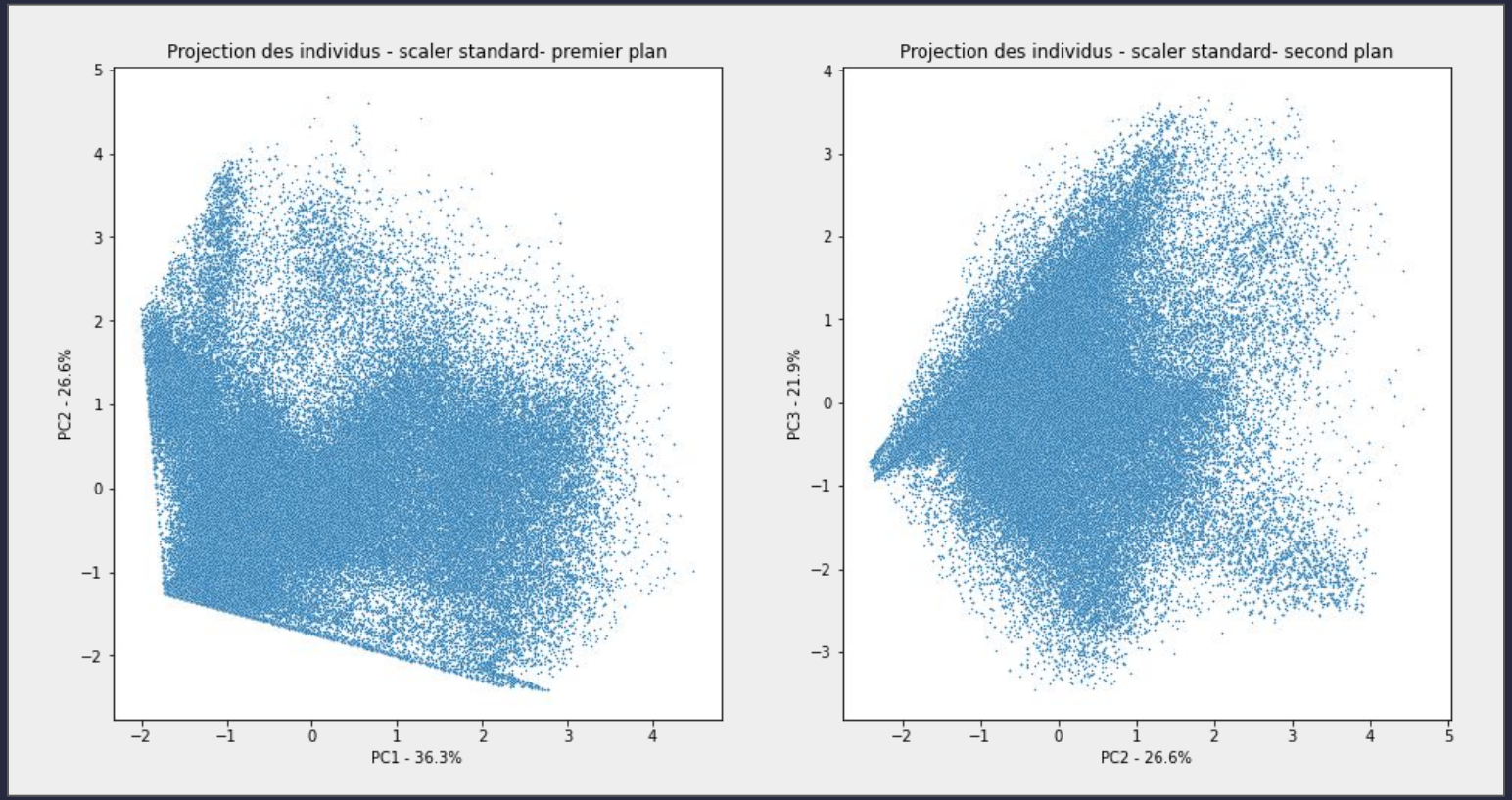
---------- Variance expliquée ----------

[0.363 0.265 0.219 0.095 0.056]

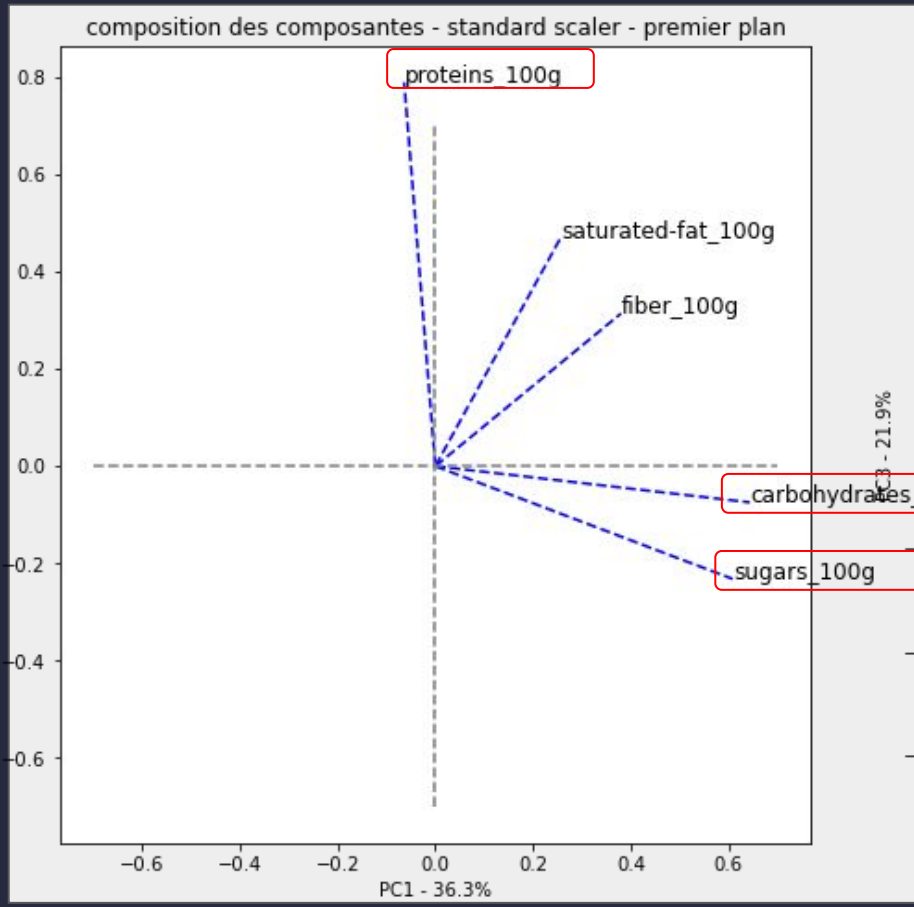
---------- Variance expliquée cumulée ----------

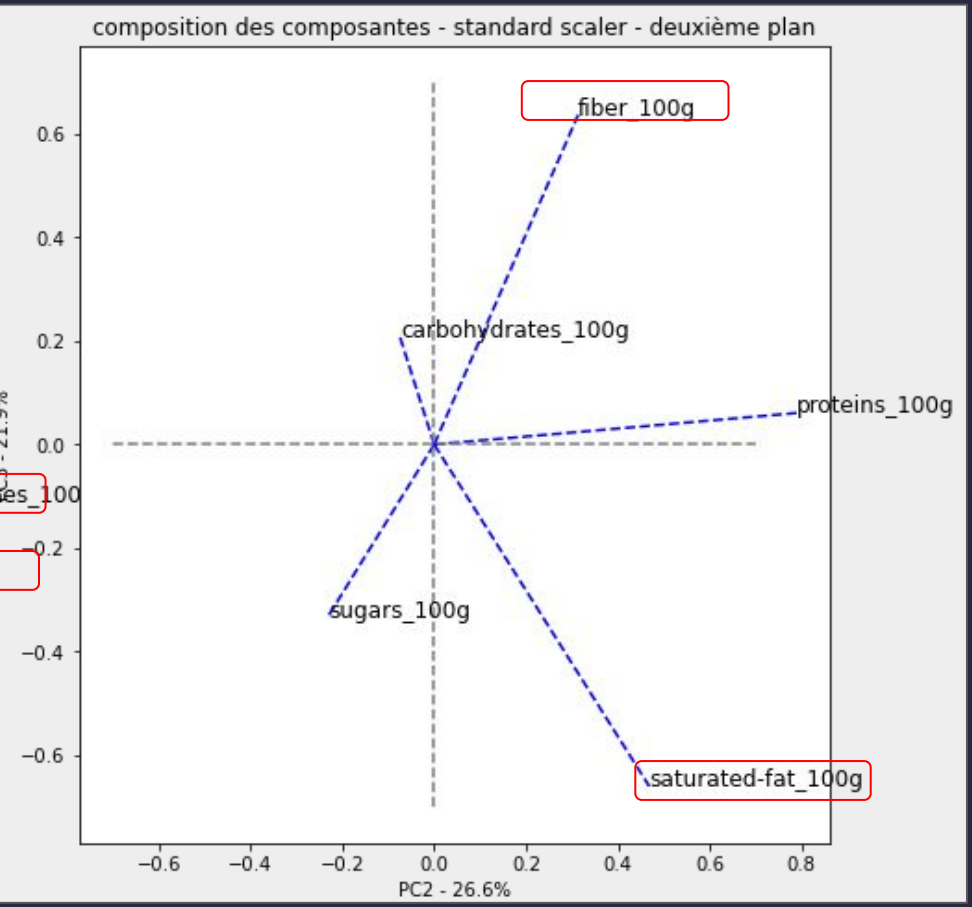
[0.363 0.629 0.848 0.943 1.]

Analyse en Composantes Principales

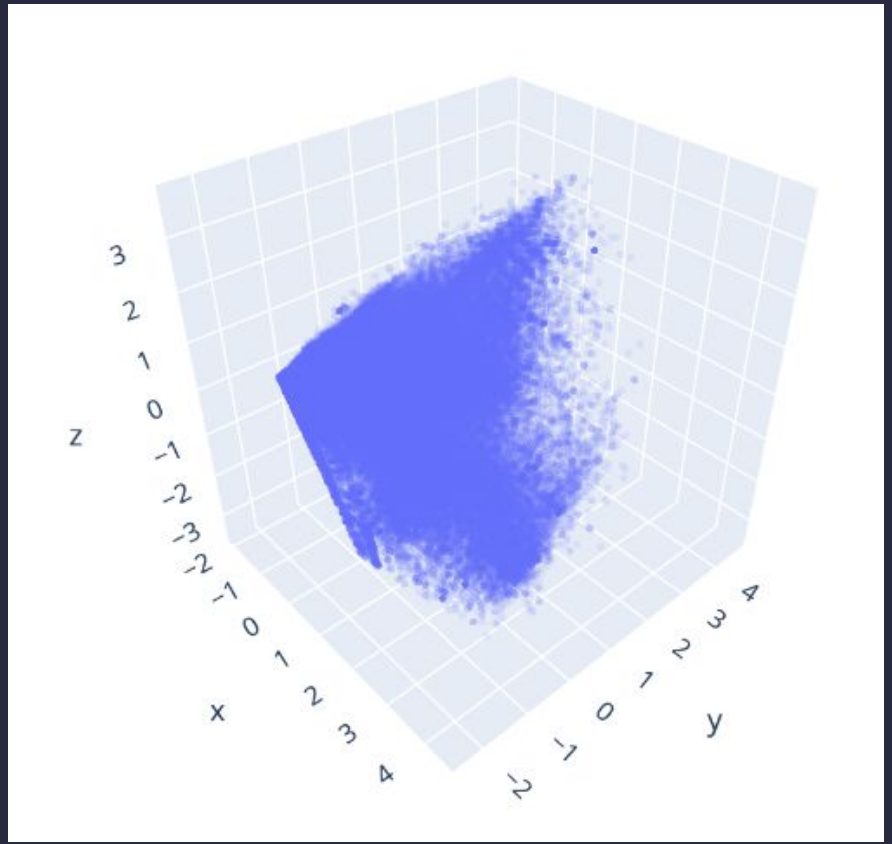


Analyse en Composantes Principales

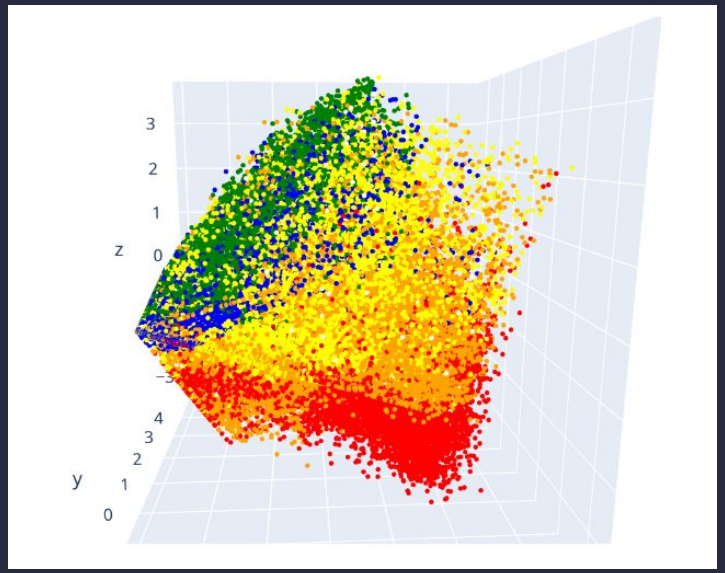


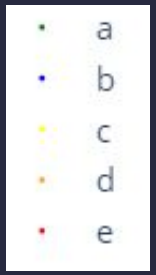


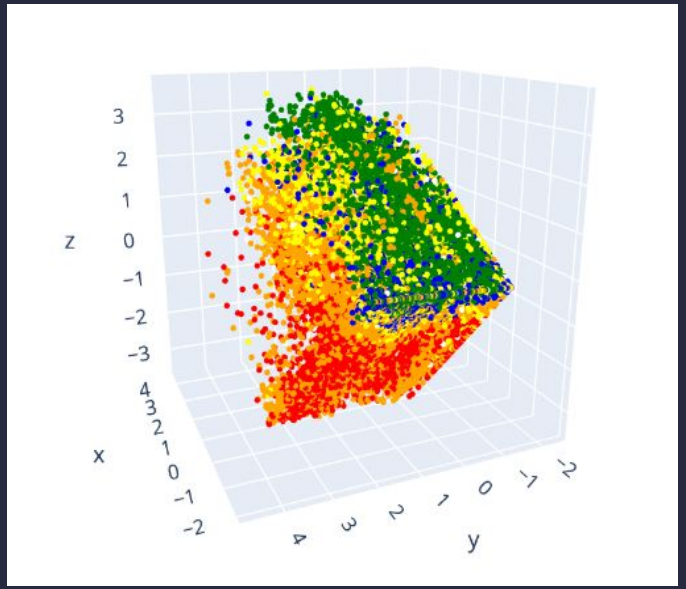
Analyse en Composantes Principales



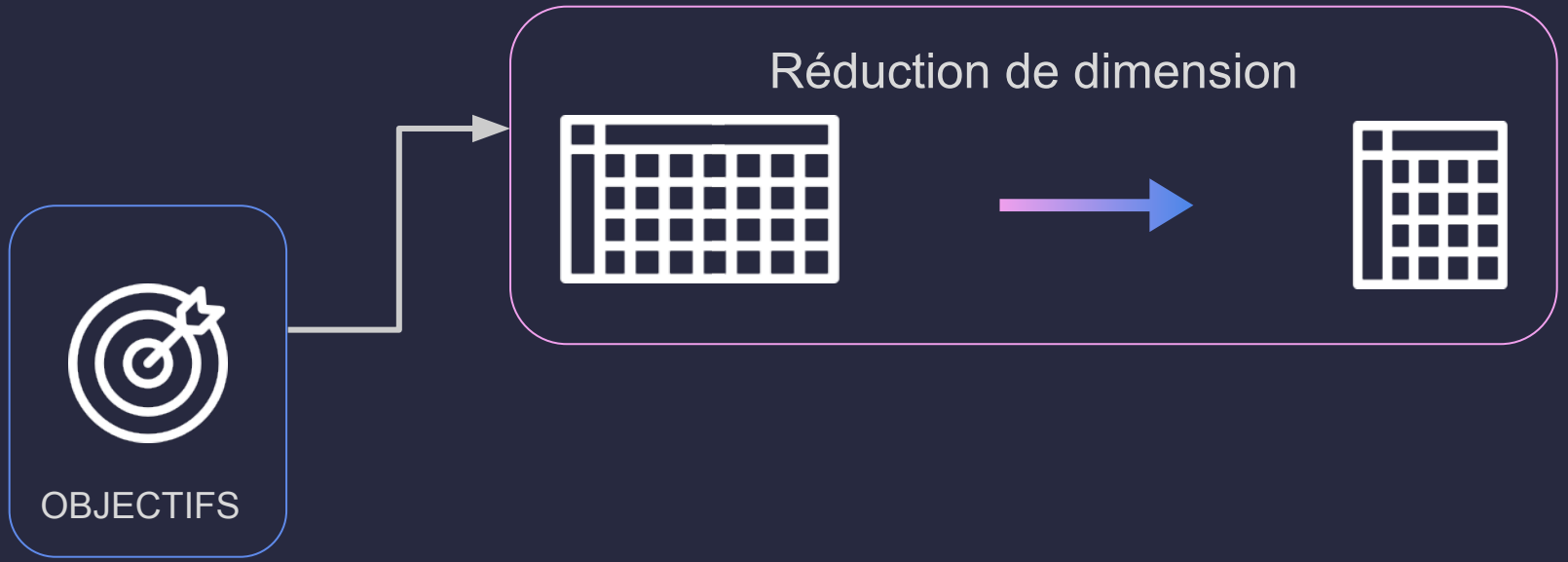
ACP - Capacité à disciminer



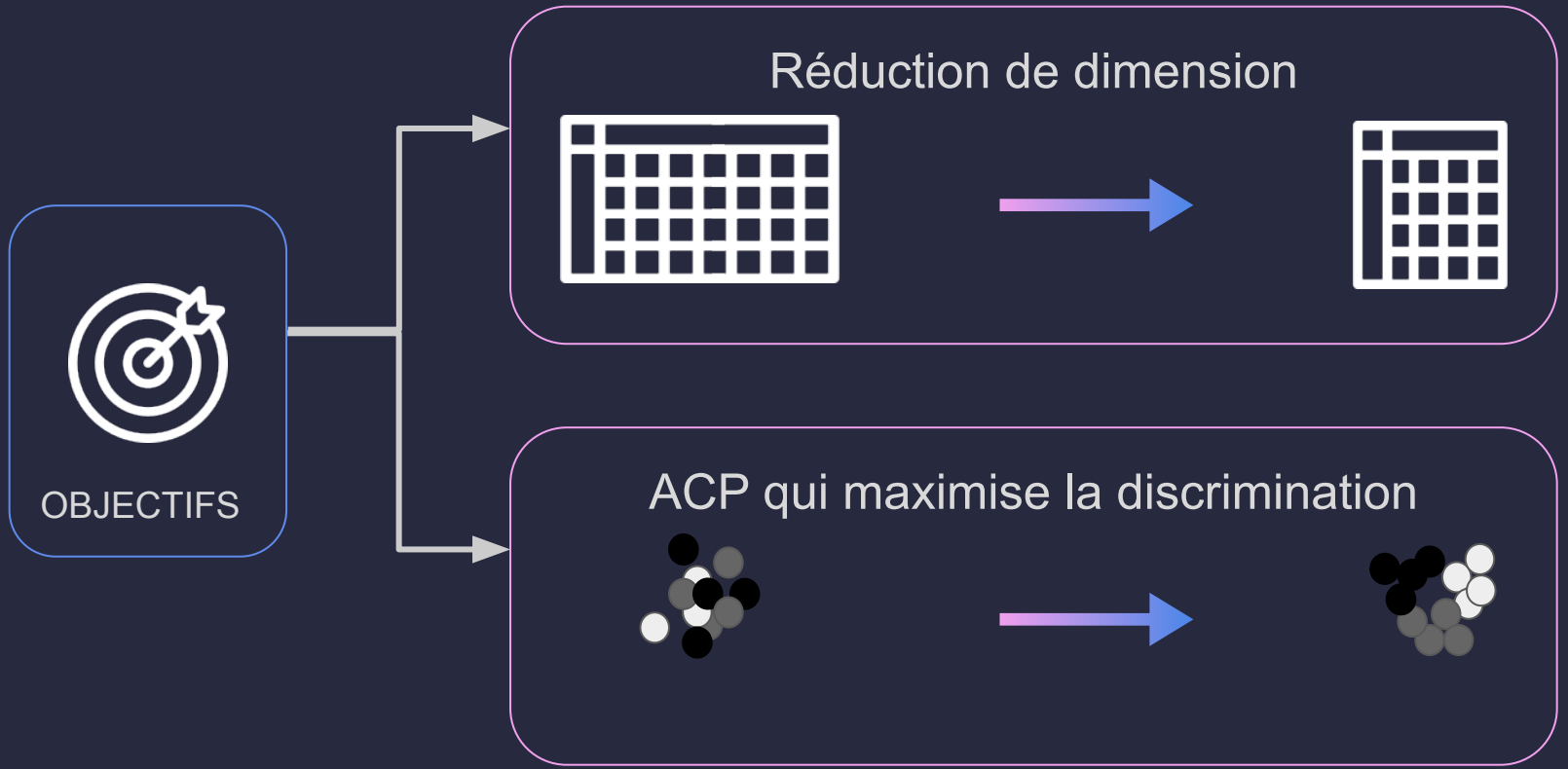




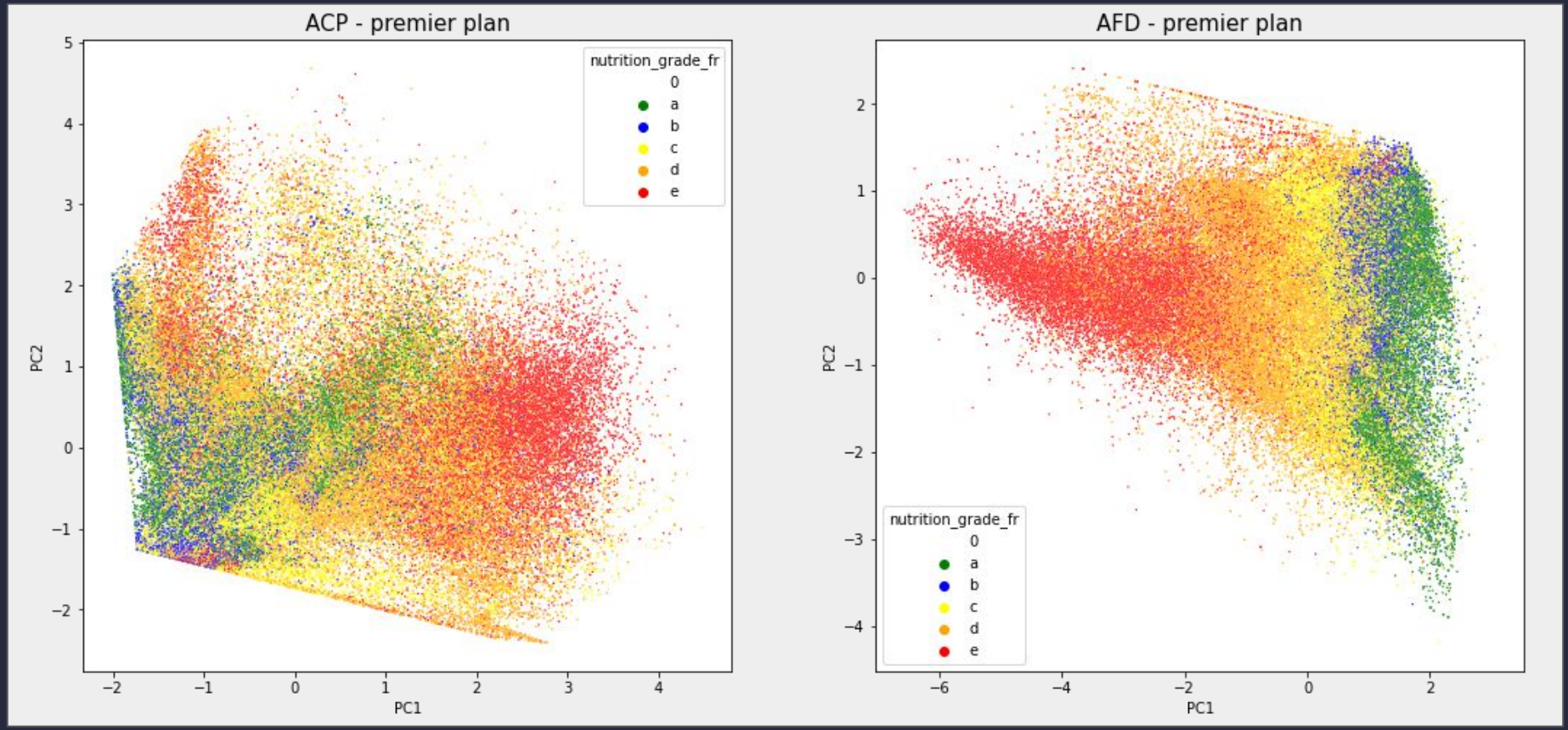
AFD - Analyse Factorielle Discriminante



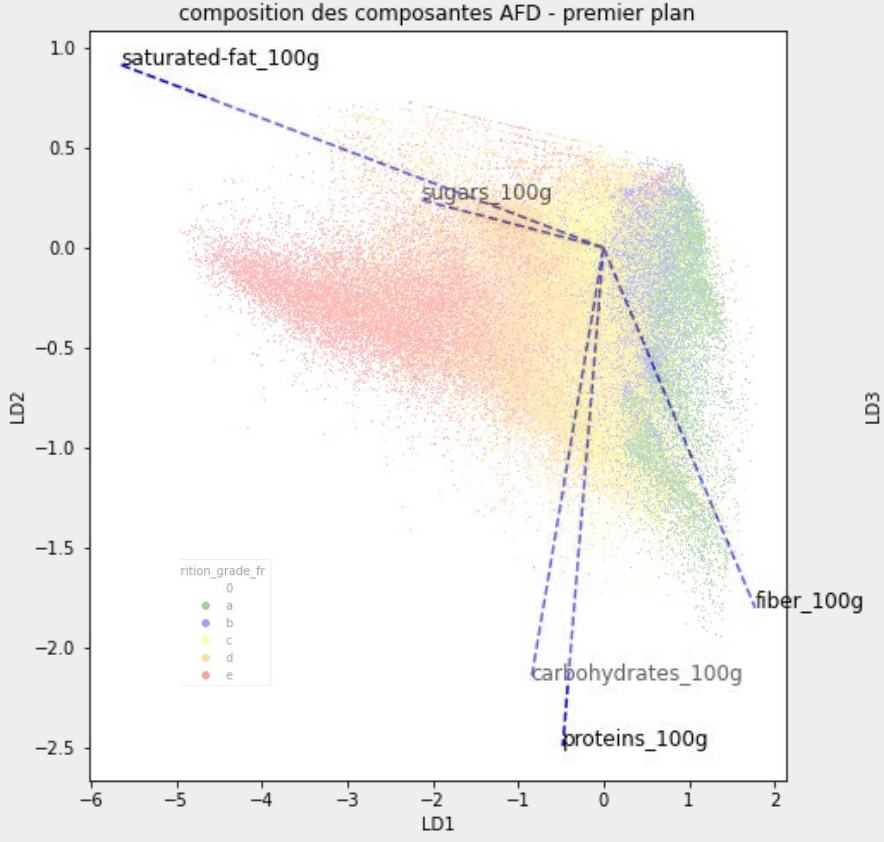
AFD - Analyse Factorielle Discriminante

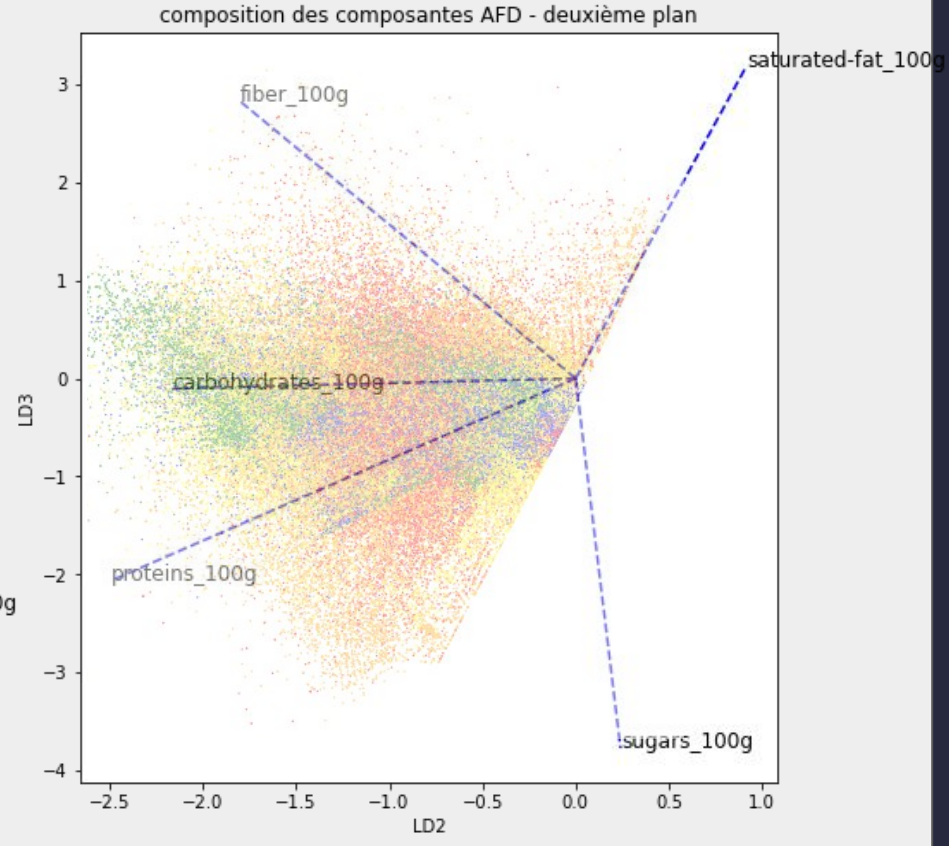


AFD - Analyse Factorielle Discriminante



AFD - Analyse Factorielle Discriminante





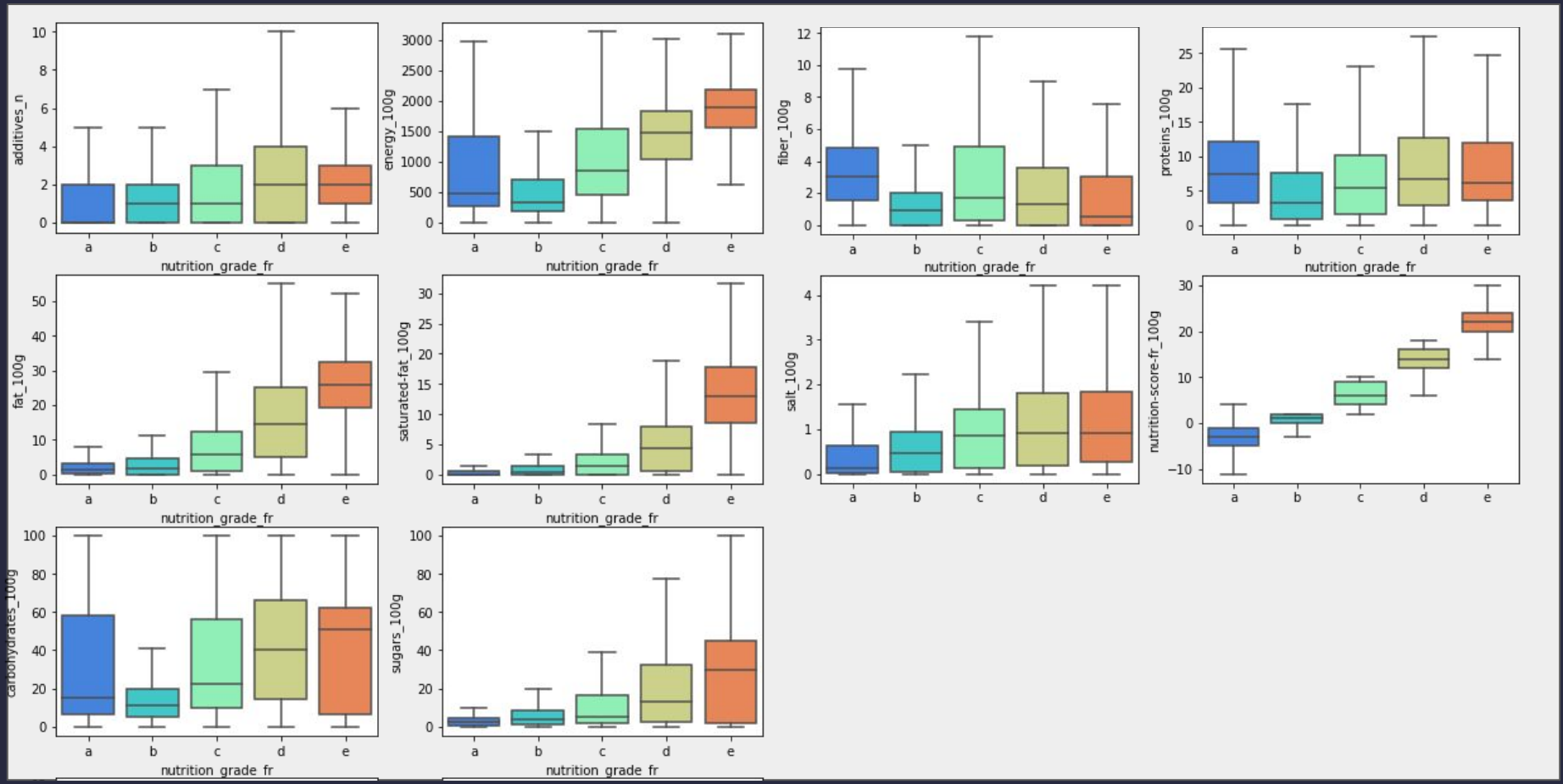
Analyse des variances - ANOVA

Variable nutrition_grade_fr

? Égalité des moyennes des groupes

? Influence entre la variable qualitative et les variables quantitatives

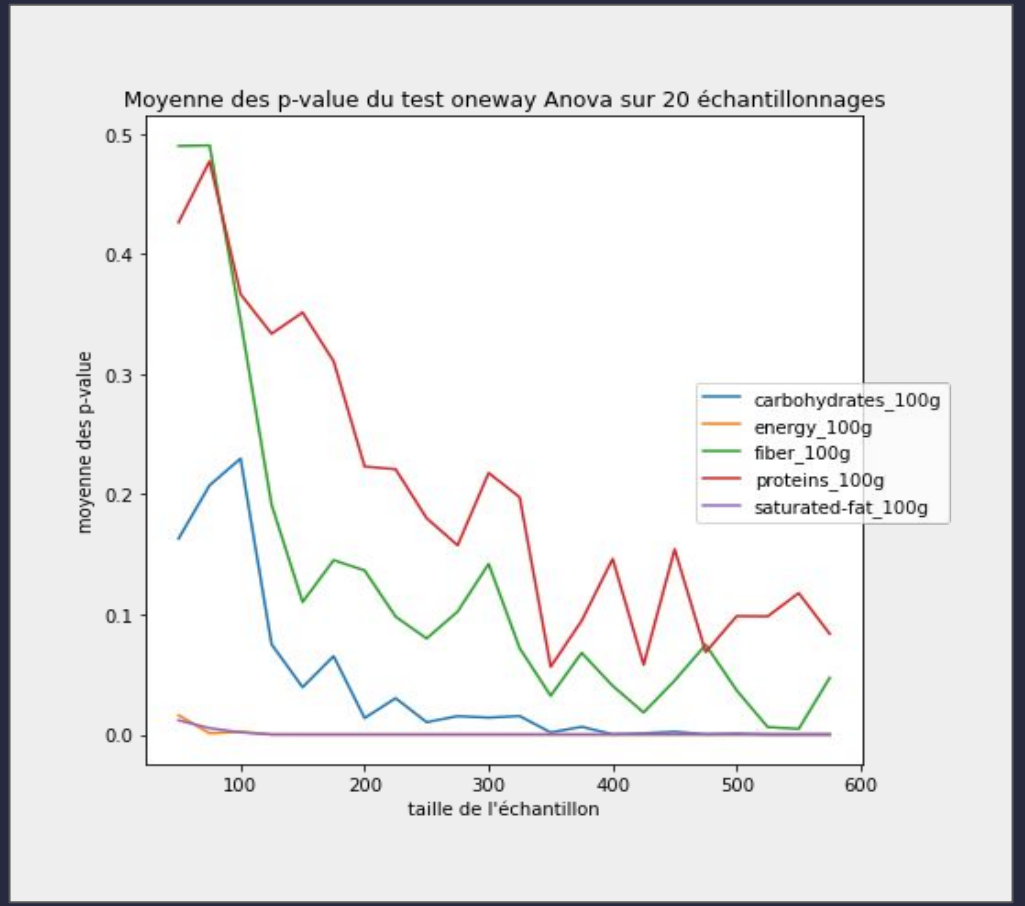
Analyse des variances - ANOVA



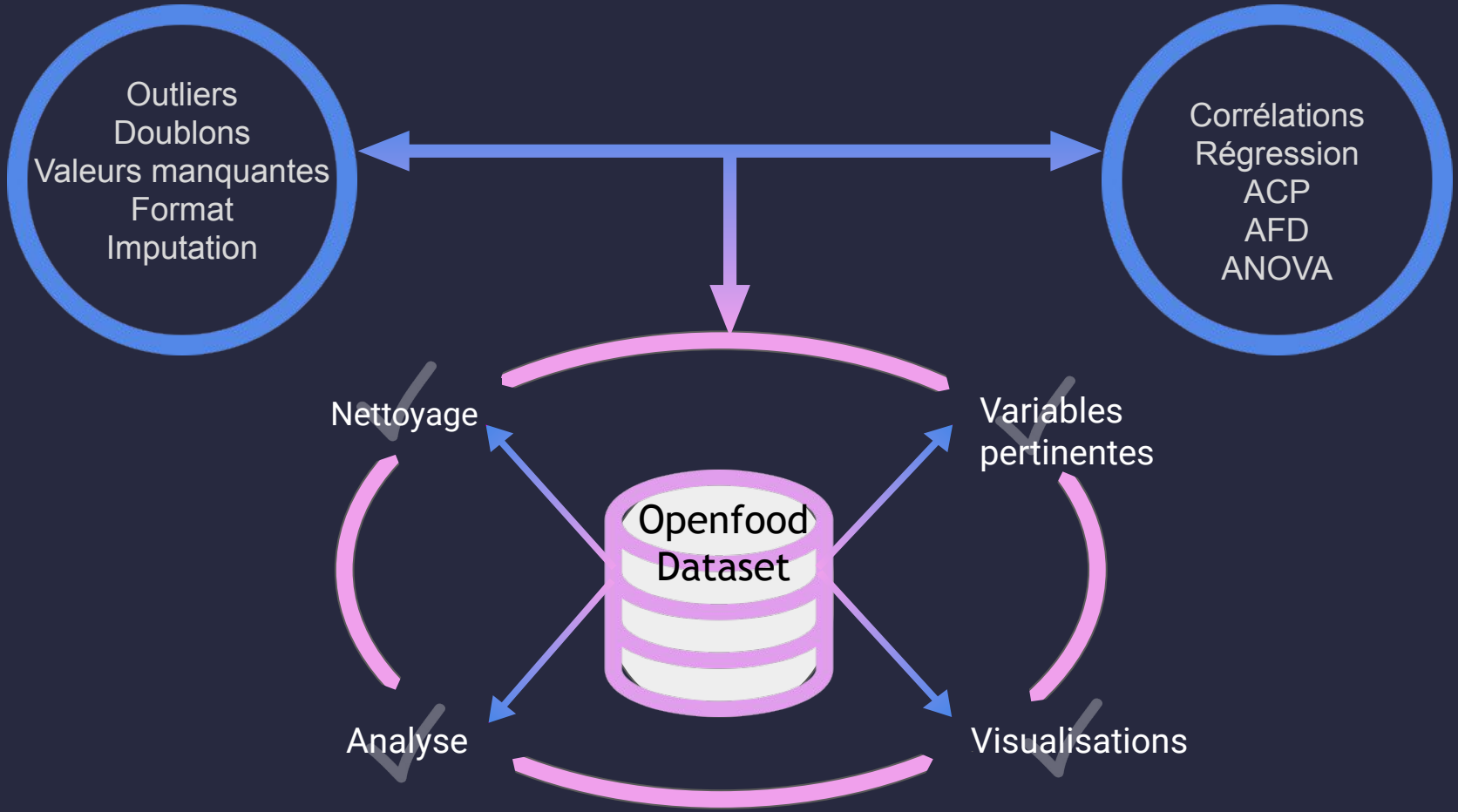
Analyse des variances - ANOVA

ANOVA sensible aux grands échantillons

20 tirages par échantillons



Conclusion


Outliers
Doublons
Valeurs manquantes
Format
Imputation
Corrélations
Régression
ACP
AFD
ANOVA
Nettoyage
Variables pertinentes
Openfood Dataset
Analyse
Visualisations

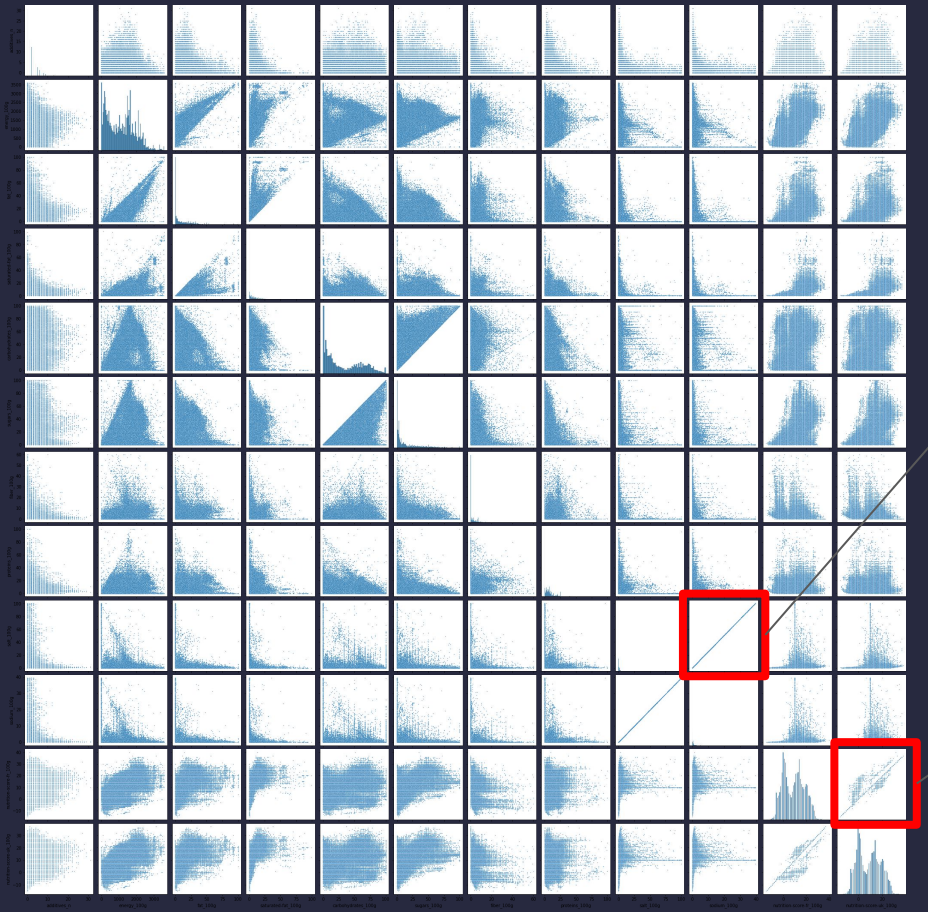
Merci

ANNEXES

☞ Variables d'intérêt - pauvres en information

	additives_n	ingredients_from_palm_oil_n	ingredients_that_may_be_from_palm_oil_n
count	236438.000000	236438.000000	236438.000000
mean	1.938534	0.019853	0.055900
std	2.508246	0.141183	0.271499
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	1.000000	0.000000	0.000000
75%	3.000000	0.000000	0.000000
max	31.000000	2.000000	6.000000

Variables d'intérêt - corrélations



Salt_100g
Vs
Sodium_100g

Spearman
correlation=0.9999997452306634

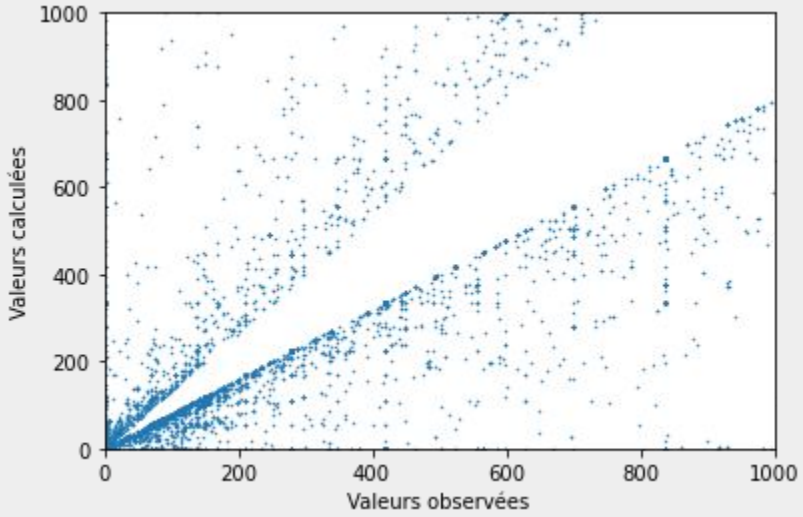
nutrition-score-fr_100g
Vs
Nutrition-score-uk_100g

Spearman correlation=0.9854351387180611

Spearman: corrélation pas forcément linéaire

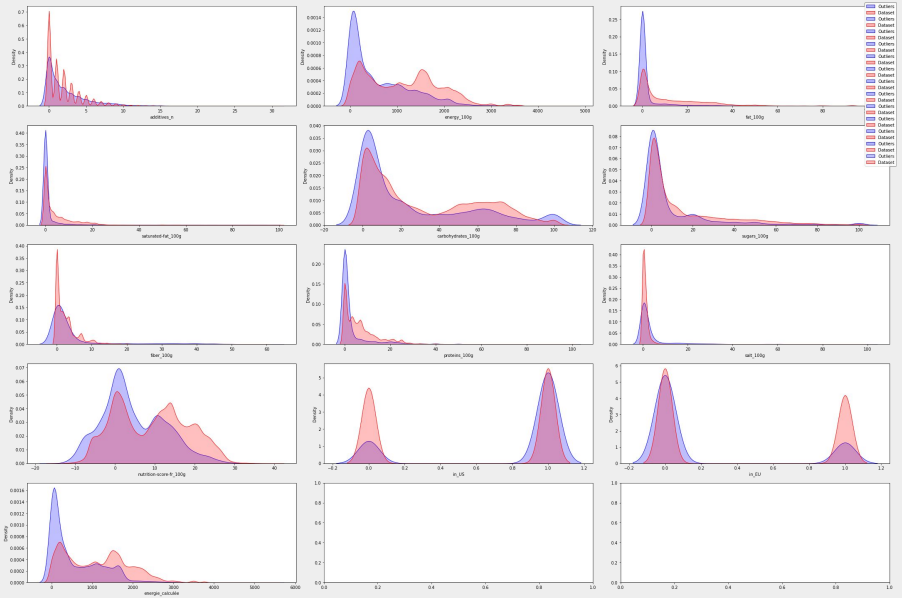
Régression linéaire - Outliers

Outliers identifiés par la régression

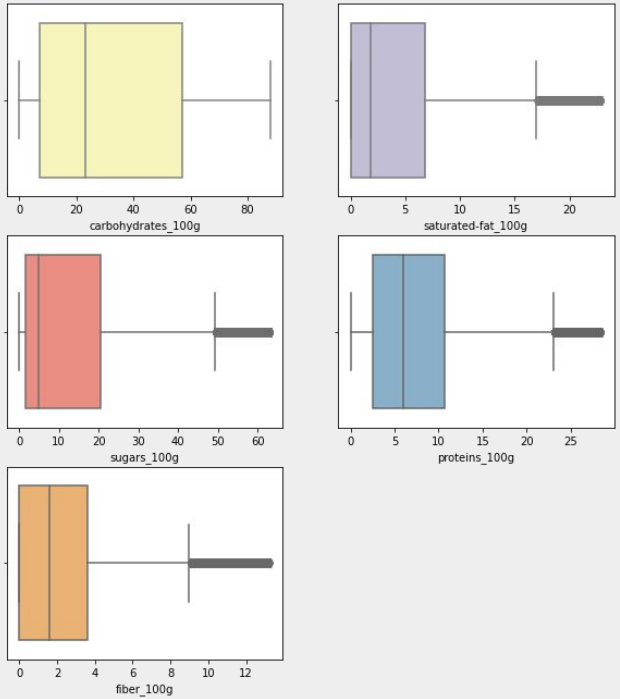


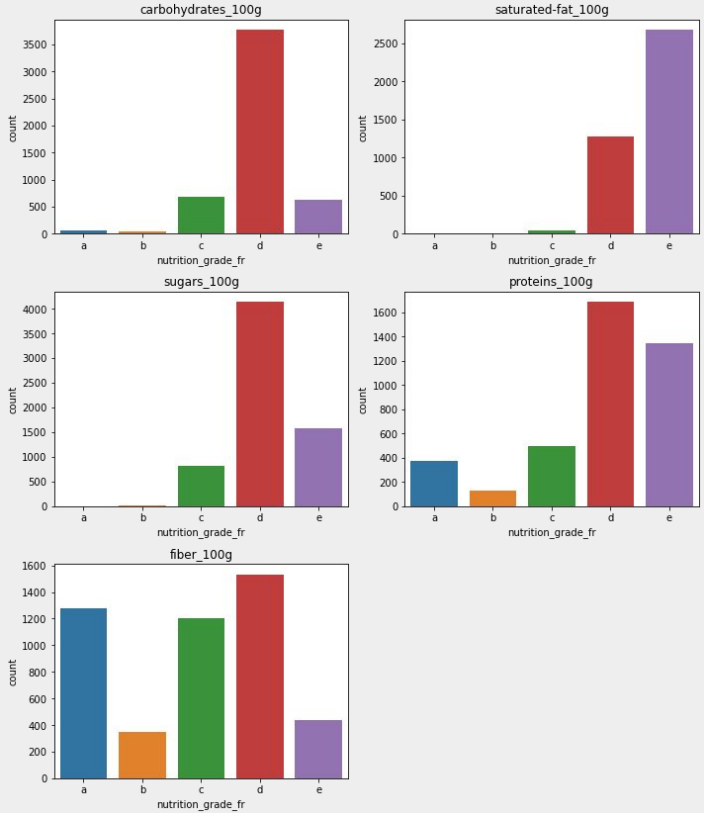
12427 individus

Distributions des outliers et du dataset entier



Valeurs extrêmes





ACP - Calcul du nutri-score

Nutriments à limiter

Points N		Seuils pour les boissons				Seuils pour les matières grasses	
Points	Energie (kJ)	Sucres (g)	Energie (kJ)	Sucres (g)	Graisses saturées (g)	Graisses saturées (%)	Sodium (mg)
0	≤ 335	≤ 4,5	≤ 0	≤ 0	≤ 1	< 10	≤ 90
1	> 335	> 4,5	≤ 30	≤ 1,5	>1	< 16	> 90
2	> 670	> 9	≤ 60	≤ 3	>2	< 22	> 180
3	> 1005	> 13,5	≤ 90	≤ 4,5	>3	< 28	> 270
4	> 1340	> 18	≤ 120	≤ 6	>4	< 34	> 360
5	> 1675	> 22,5	≤ 150	≤ 7,5	>5	< 40	> 450
6	> 2010	> 27	≤ 180	≤ 9	>6	< 46	> 540
7	> 2345	> 31	≤ 210	≤ 10,5	>7	< 52	> 630
8	> 2680	> 36	≤ 240	≤ 12	>8	< 58	> 720
9	> 3015	> 40	≤ 270	≤ 13,5	>9	< 64	> 810
10	> 3350	> 45	> 270	> 13,5	>10	≥ 64	> 900
Gamme (points)	0 à 10	0 à 10	0 à 10	0 à 10	0 à 10	0 à 10	0 à 10
Total	Somme des points pour l'énergie, les sucres, les graisses saturées et le sodium						

Nutriments, aliments à encourager

Points P		Seuils pour les boissons		
Points	Fruits, légumes (%)	Fruits, légumes (%)	Fibres (g)	Protéines (g)
0	≤ 40	≤ 40	≤ 0,7	≤ 1,6
1	> 40	-	> 0,7	> 1,6
2	> 60	> 40	> 1,4	> 3,2
3	-	-	> 2,1	> 4,8
4	-	> 60	> 2,8	> 6,4
5	> 80		> 3,5	> 8,0
6	-	-	-	-
7	-	-	-	-
8	-	-	-	-
9	-	-	-	-
10	-	> 80	-	-
Gamme (points)	0 à 5	0 à 10	0 à 5	0 à 5
Total	Somme des points pour les consommations de fruits et légumes, les fibres et les protéines			

02 Choix de la méthode de calcul du score final

- Points Négatifs N ≥ 11
  - Score des fruits et légumes = 5 → Points N - Points P
  - Score des fruits et légumes < 5 → Points N - (points fibres + points fruits et légumes)
- Points Négatifs N < 11 ou pour les fromages → Points N - Points P

03 Attribution d'une couleur et d'une lettre

Score final variant de **-15 (qualité nutritionnelle élevée)** à **40 (faible qualité nutritionnelle)**

Aliments solides	Boissons	Logo
Min à -1	Eaux toujours en A	A
0 à 2	Min à 1	B
3 à 10	2 à 5	C
11 à 18	6 à 9	D
19 à max	10 à max	E

Quels produits concernés ?

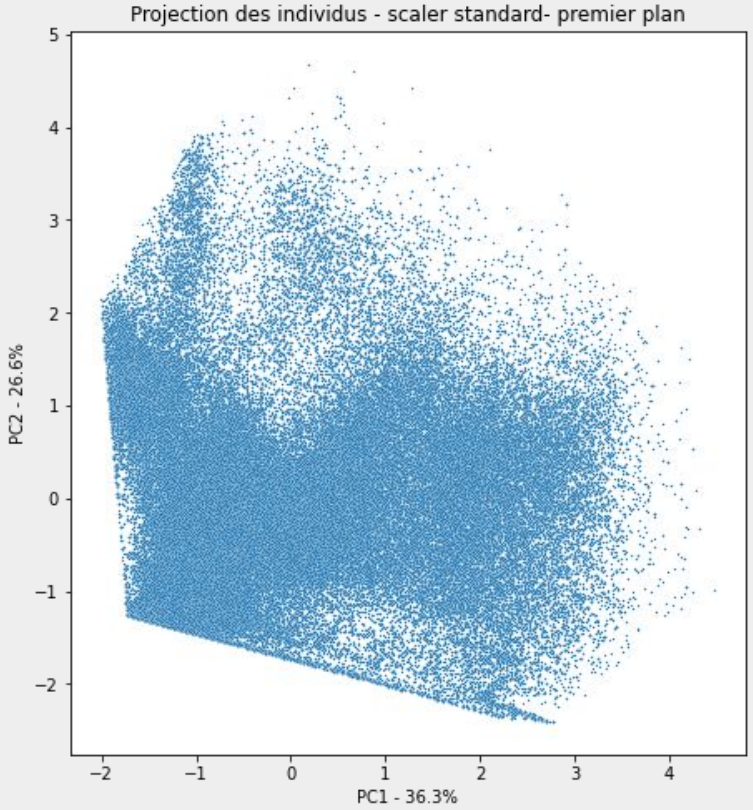
- Tous les aliments transformés, excepté les herbes aromatiques, thés, cafés, levures...
- Toutes les boissons, excepté les boissons alcoolisées
- Excepté les produits dont la face la plus grande a une surface inférieure à 25 cm²

www.quoidansmonassiette.fr T.Fiolet

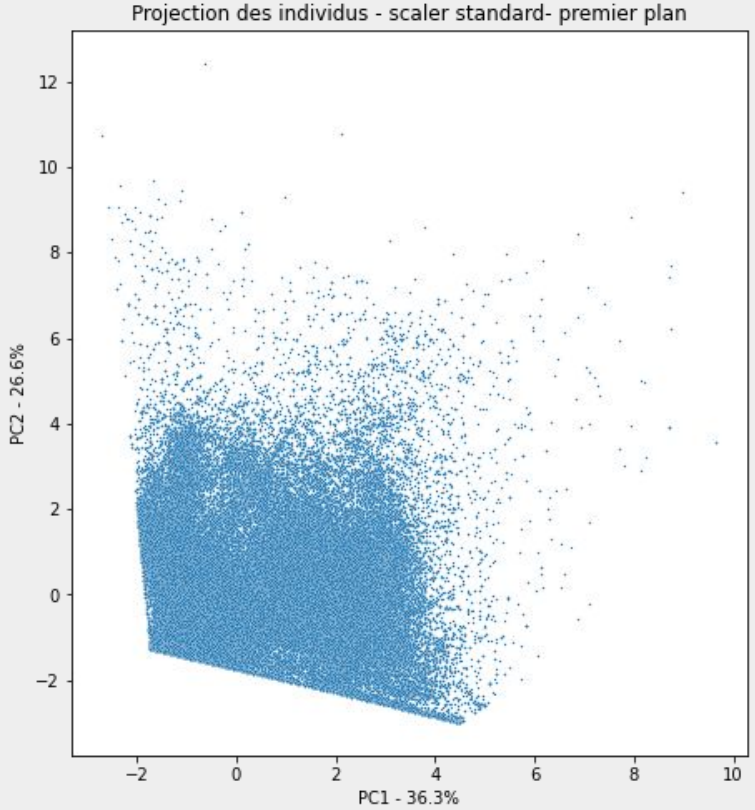
Adapté de Julia C, Hercberg S (2017) Nutri-Score: evidence of the effective-ness of the French front-of-pack nutrition label. Ernahrungs . Umschau 64(12): 181–187

https://solidarites-sante.gouv.fr/

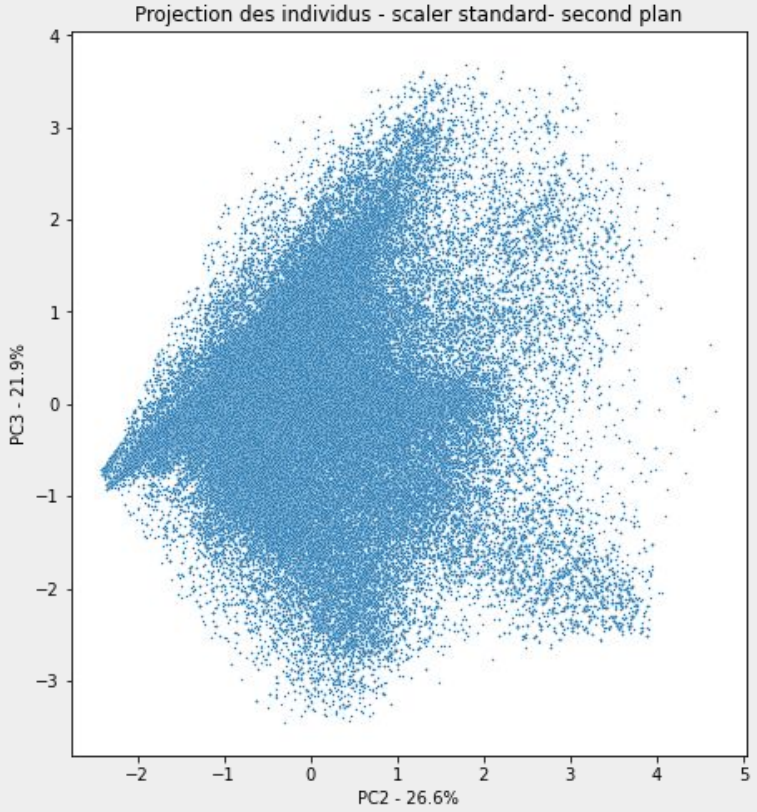
Analyse en Composantes Principales



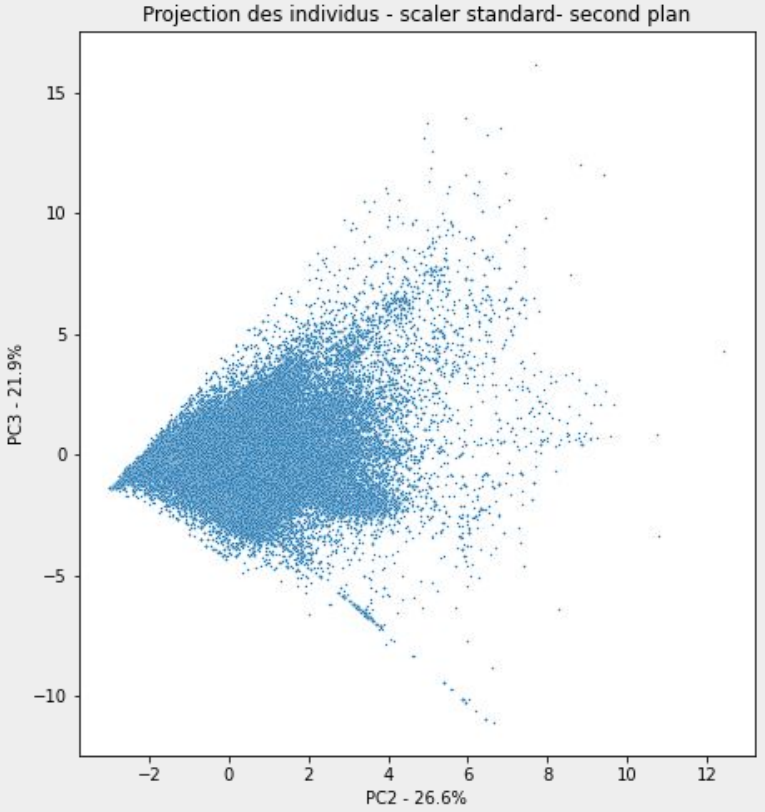
Tous les individus



Analyse en Composantes Principales



Tous les individus



ACP - Capacité à disciminer