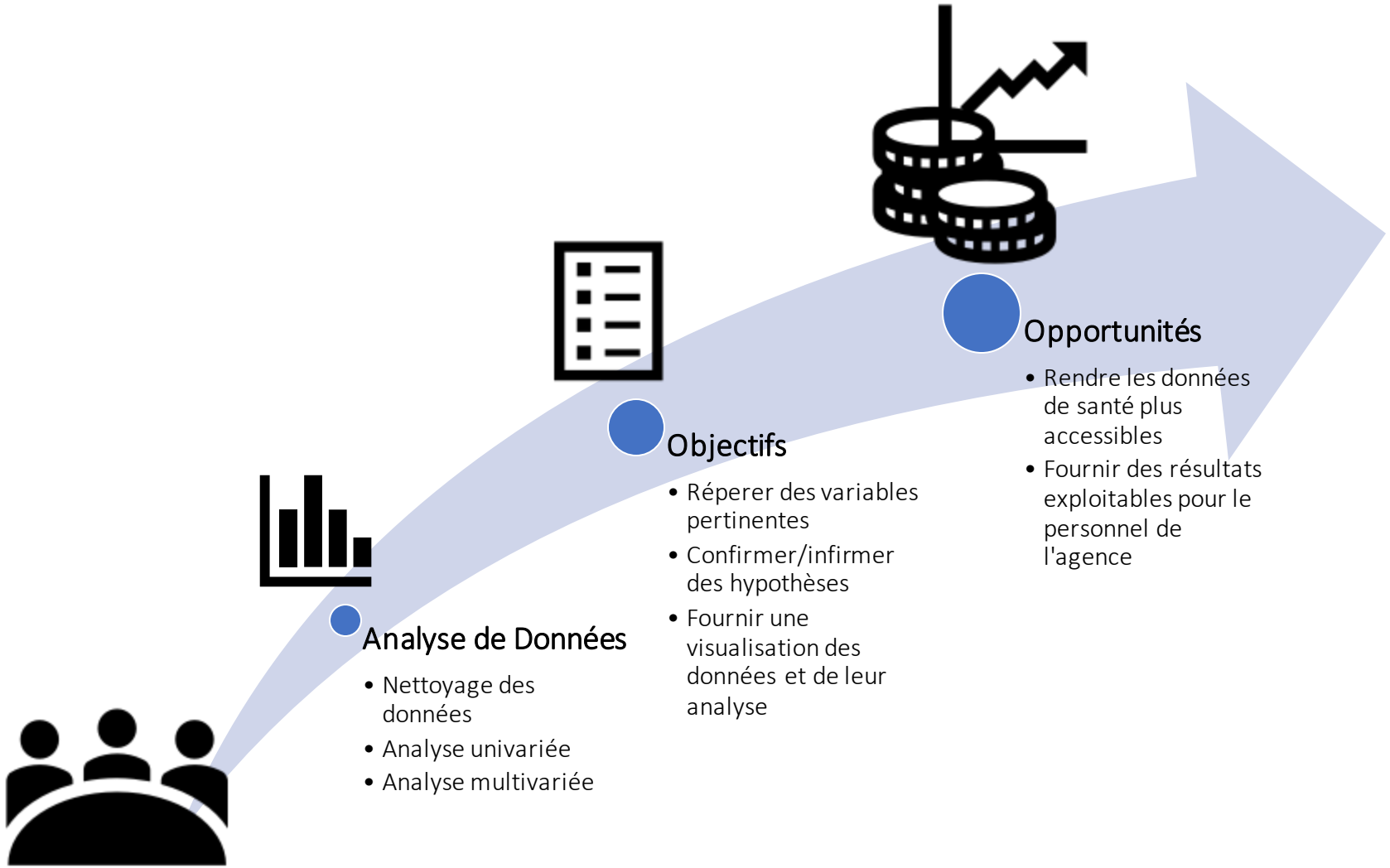
Appel à projet "Rendre vos données de santé plus accessibles"

Analyse du jeu de données Open Food
pour l'Agence Santé publique France

Bailly DIOUNOU – 21/09/2020

Contexte & périmètre du projet



Analyse de Données

- Nettoyage des données
- Analyse univariée
- Analyse multivariée

Objectifs

- Réperer des variables pertinentes
- Confirmer/infirmer des hypothèses
- Fournir une visualisation des données et de leur analyse

Opportunités

- Rendre les données de santé plus accessibles
- Fournir des résultats exploitables pour le personnel de l'agence

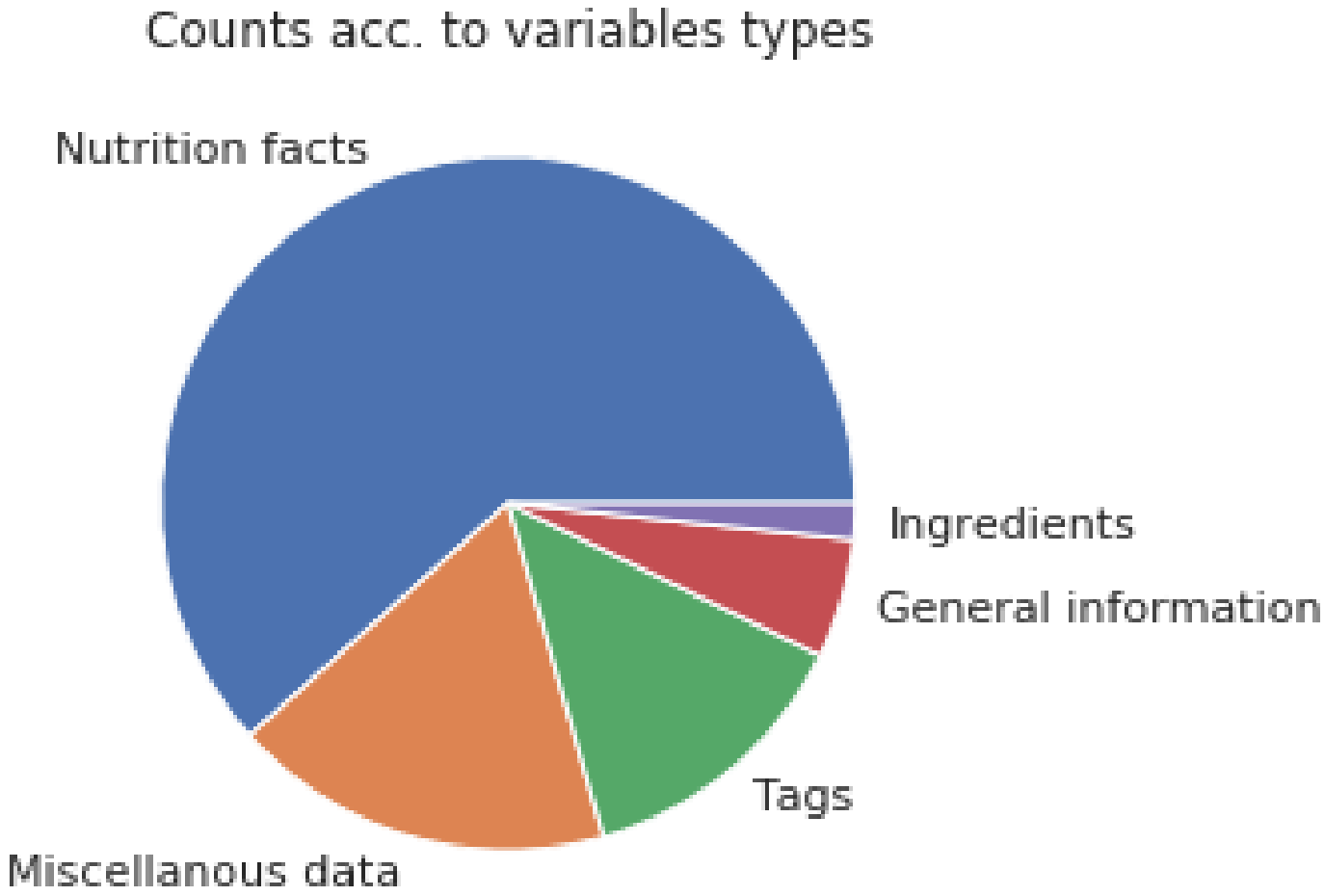
Agence Santé publique France"

Présentation générale du jeu de données

Présentation générale du jeu de données

Données statistiques de base - Données brutes

- Taille de la population **N = 1 437 214** individus
- Nombre de variables: **180 + 1**
- Variables quantitatives **q = 115**



Présentation générale du jeu de données

	carbohydrates_100g	energy_100g	fat_100g	proteins_100g	salt_100g	saturated-fat_100g	sugars_100g
count	1148942	1158319	1149482	1150748	1129792	1102551	1127905
mean	28,55185991	5,75451E+36	13,92499591	8,511698902	2,177489684	123350308,5	13,88769009
std	30,33881954	6,19331E+39	182,3046013	36,80307148	142,6501191	1,29521E+11	20,66323889
min	0	0	0	-500	0	0	-1
25%	3,57	420	0,6	1,2	0,06	0,1	0,7
50%	15,71	1095	6,9	5,8	0,535	1,9	3,9
75%	53,57	1674	21,43	12	1,35382	7,27	20
max	6670	6,66556E+42	153679,4643	31000	105000	1,36E+14	4800

Présentation générale du jeu de données

	carbohydrates_100g	energy_100g	fat_100g	proteins_100g	salt_100g	saturated-fat_100g	sugars_100g
count	1148942	1158319	1149482	1150748	1129792	1102551	1127905
mean	28,55185991	5,75451E+36	13,92499591	8,511698902	2,177489684	123350308,5	13,88769009
std	30,33881954	6,19331E+39	182,3046013	36,80307148	142,6501191	1,29521E+11	20,66323889
min	0	0	0	-500	0	0	-1
25%	3,57	420	0,6	1,2	0,06	0,1	0,7
50%	15,71	1095	6,9	5,8	0,535	1,9	3,9
75%	53,57	1674	21,43	12	1,35382	7,27	20
max	6670	6,66556E+42	153679,4643	31000	105000	1,36E+14	4800

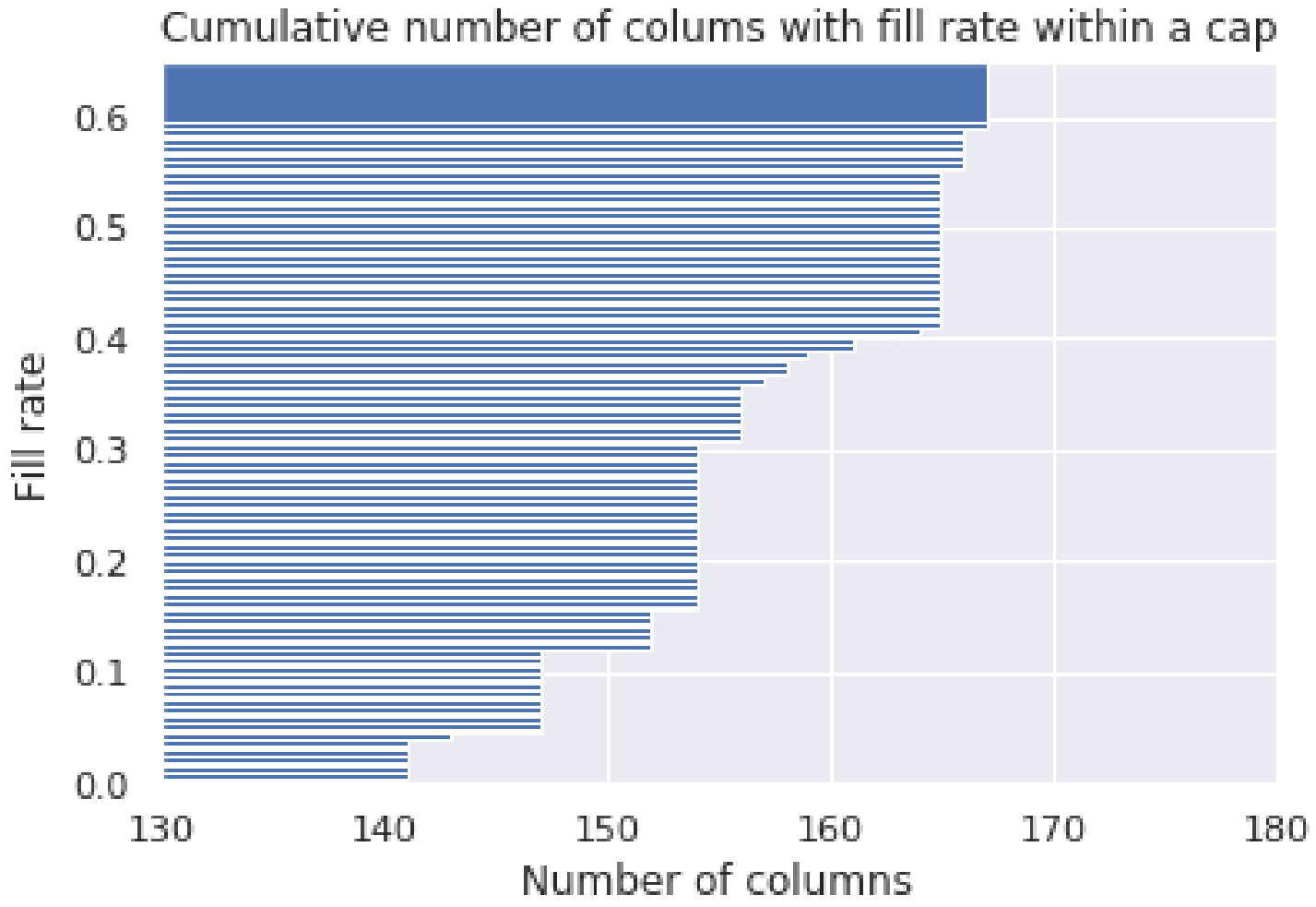
Valeurs très aberrantes !!!

Nettoyage des données

Nettoyage des données

Traitement des valeurs manquantes

Tracé du nombre de colonnes ayant au plus tel *taux de remplissage*.

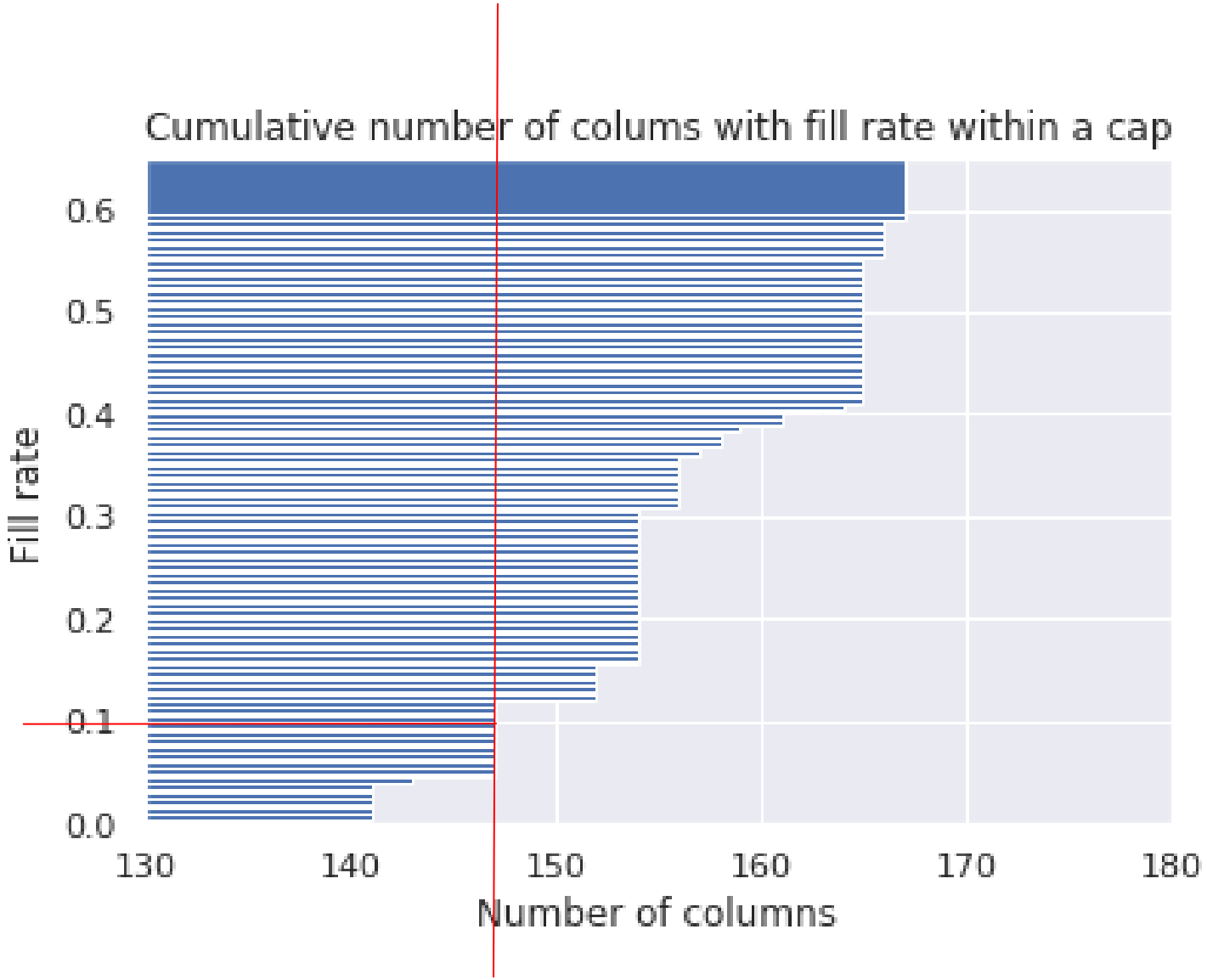


Nettoyage des données

Traitement des valeurs manquantes

Tracé du nombre de colonnes ayant au plus tel *taux de remplissage*.

Cap 10% : 147



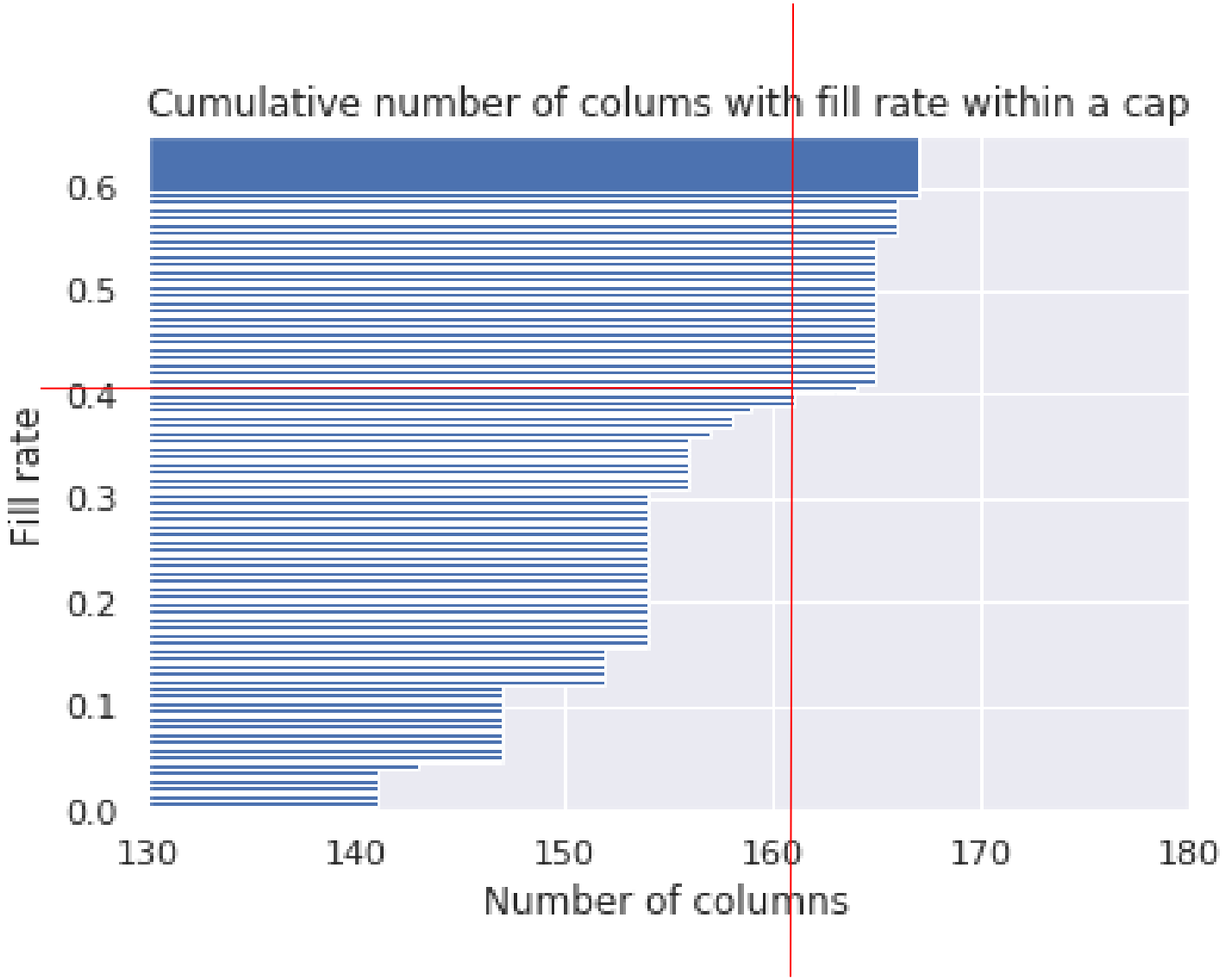
Nettoyage des données

Traitement des valeurs manquantes

Tracé du nombre de colonnes ayant au plus tel *taux de remplissage*.

Cap 10% : 147

Cap 40% : 161



Nettoyage des données

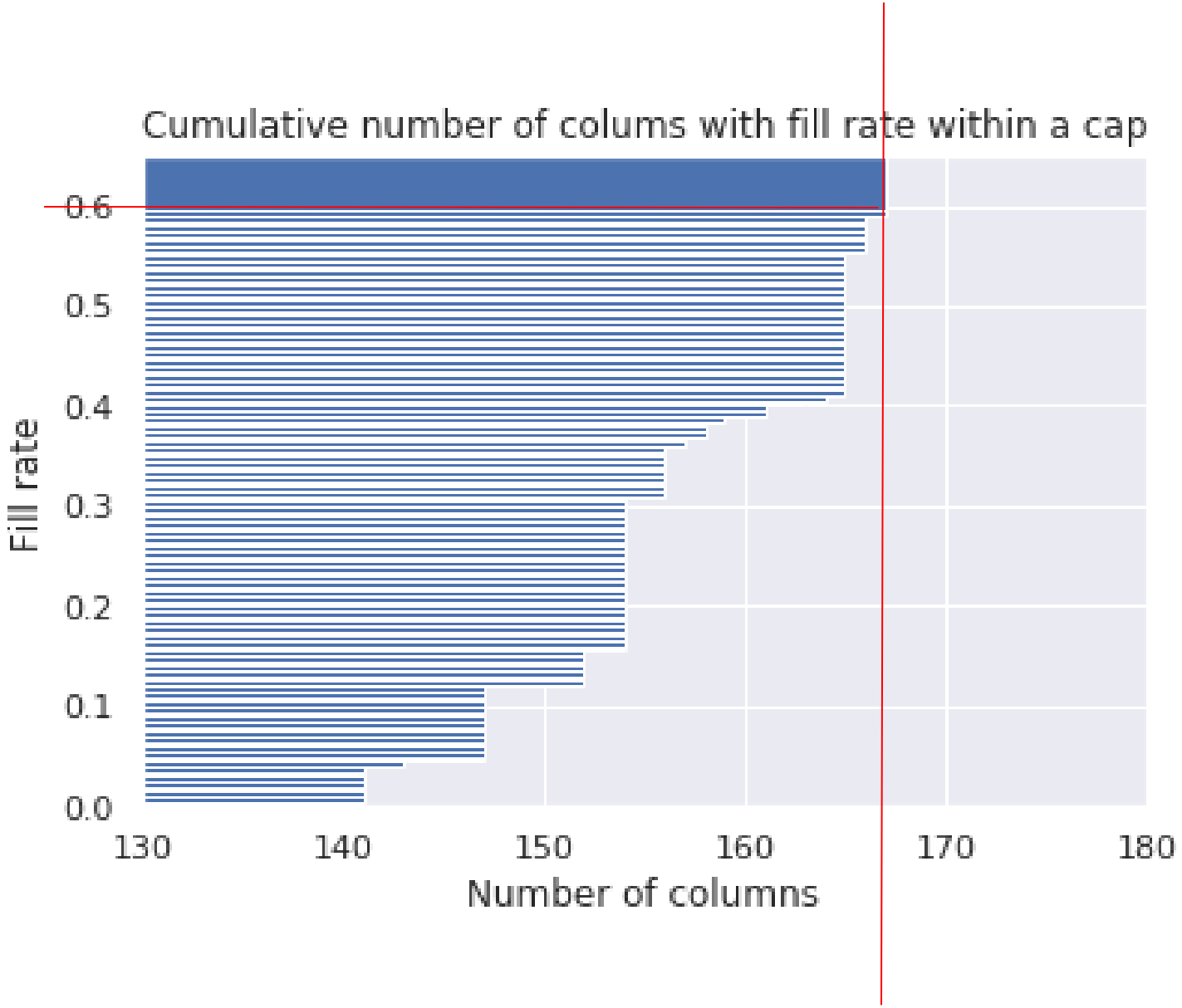
Traitement des valeurs manquantes

Tracé du nombre de colonnes ayant au plus tel *taux de remplissage*.

Cap 10% : 147

Cap 40% : 161

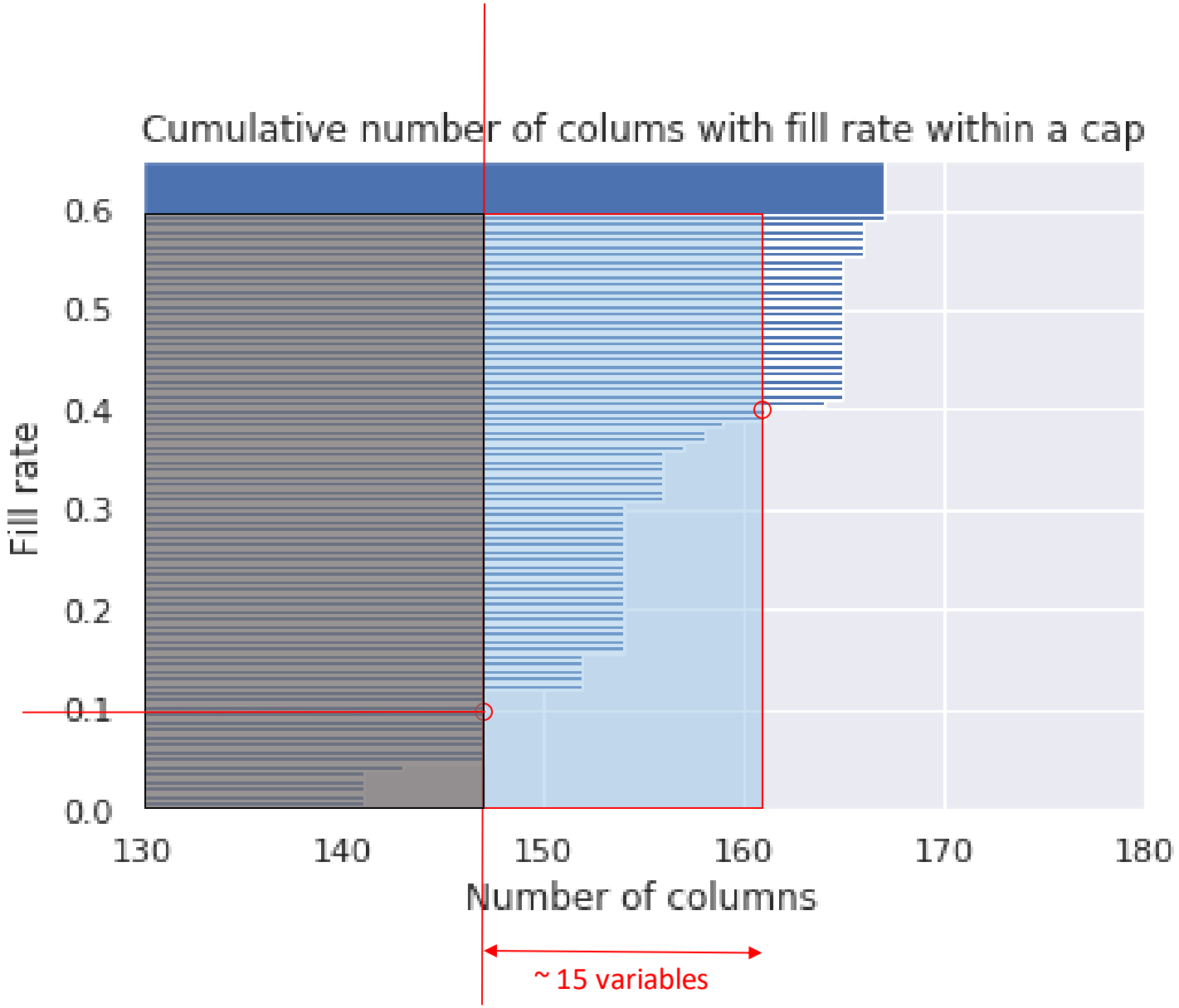
Cap 60% : 167



Nettoyage des données

Traitement des valeurs manquantes

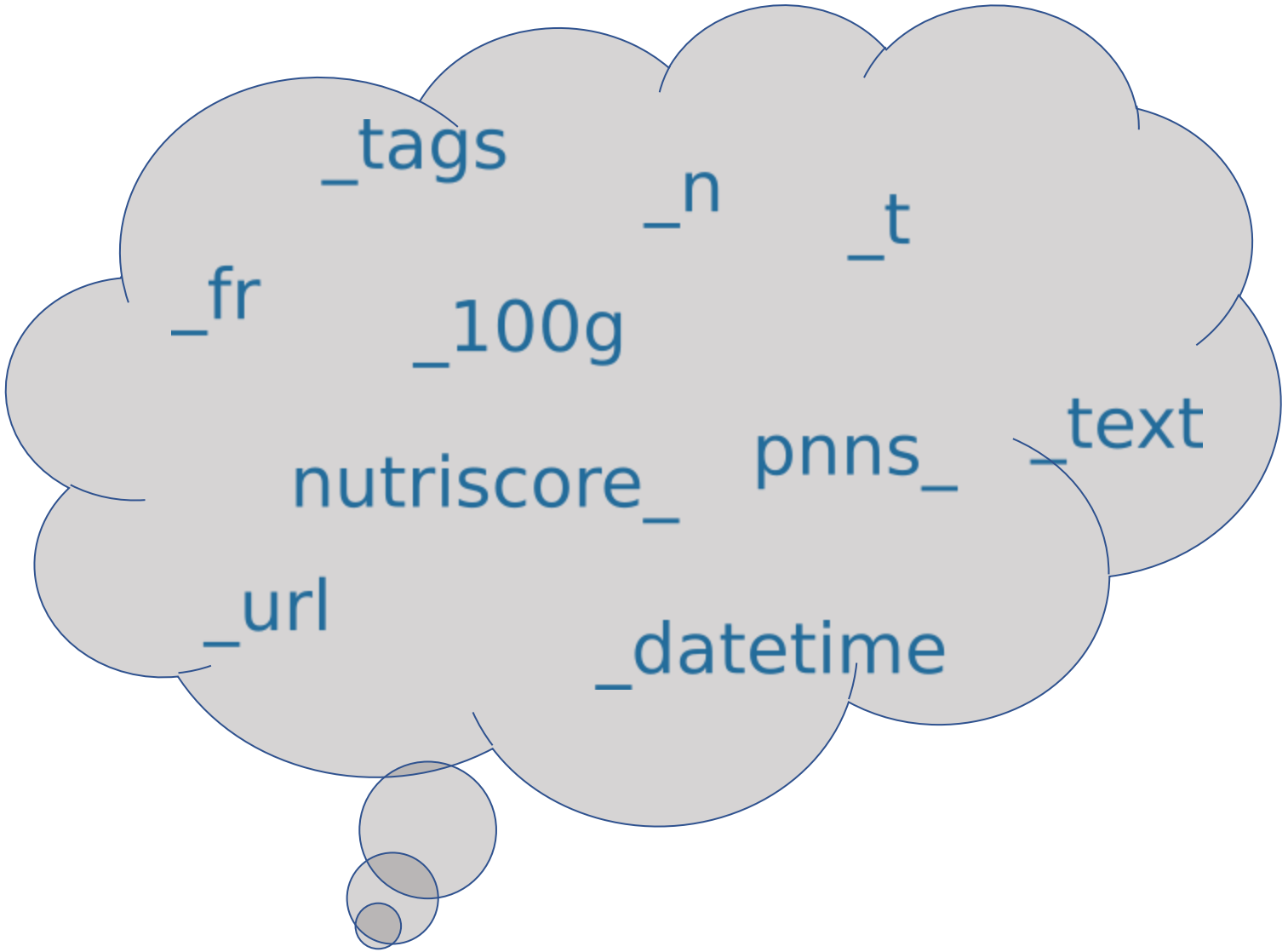
- Plancher de suppression inconditionnelle: **Cap 10%**
  - **Majorité {*nutrition facts*}**
- Plage intermédiaire: suppression conditionnelle --> **Critère de pertinence**
  - 33% {*nutrition facts*, ***misc. data***, ***tags***}



Nettoyage des données

Traitement des variables non-pertinentes

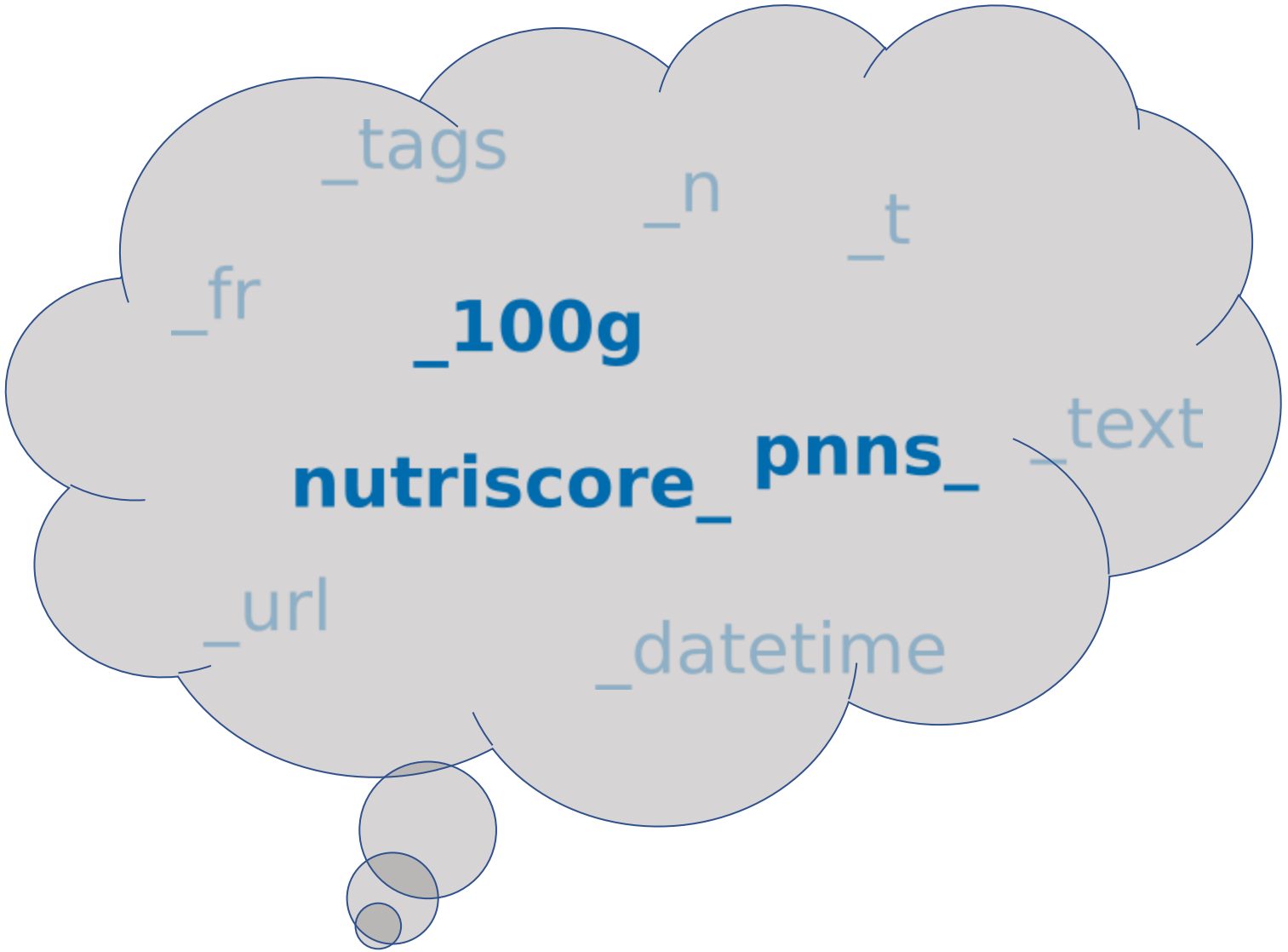
- Variables =
  - Métadonnées &
  - Données nutritionnelles qualitatives



Nettoyage des données

Traitement des variables non-pertinentes

- Variables =
  - Métadonnées &
  - Données nutritionnelles qualitatives



Nettoyage des données

Traitement des variables non-pertinentes

- Variables =

  Métadonnées &
  Données nutritionnelles qualitatives

- Corrélation sur les valeurs nutritionnelles moyennes

  {(energy-kcal, energy), (salt, sodium)}

Nettoyage des données

Traitement des variables non-pertinentes

- Variables =

  Métadonnées &
  Données nutritionnelles qualitatives

- Corrélation sur les valeurs nutritionnelles moyennes

  {(energy-kcal, energy), (salt, sodium)}

Nettoyage des données

Traitement des valeurs aberrantes

Traitement des doublons

Nettoyage des données

Traitement des valeurs aberrantes

- Sur les valeurs nutritionnelles moyennes: *'_100g'*
  - Critères généraux
    - Valeur inférieure: **non-négative**
    - Valeur supérieure: **"< 100 g"**
  - Critère spécifique sur l'énergie
    - Valeur supérieure: [donnée stat+métier]: **IQR + Apport de Référence pour un adulte 8400 kJ**

Traitement des doublons

Nettoyage des données

Traitement des valeurs aberrantes

- Sur les valeurs nutritionnelles moyennes: *'_100g'*
  - Critères généraux
    - Valeur inférieure: **non-négative**
    - Valeur supérieure: **"< 100 g"**
  - Critère spécifique sur l'énergie
    - Valeur supérieure: [donnée stat=métier]: **IQR + Apport de Référence pour un adulte 8400 kJ**

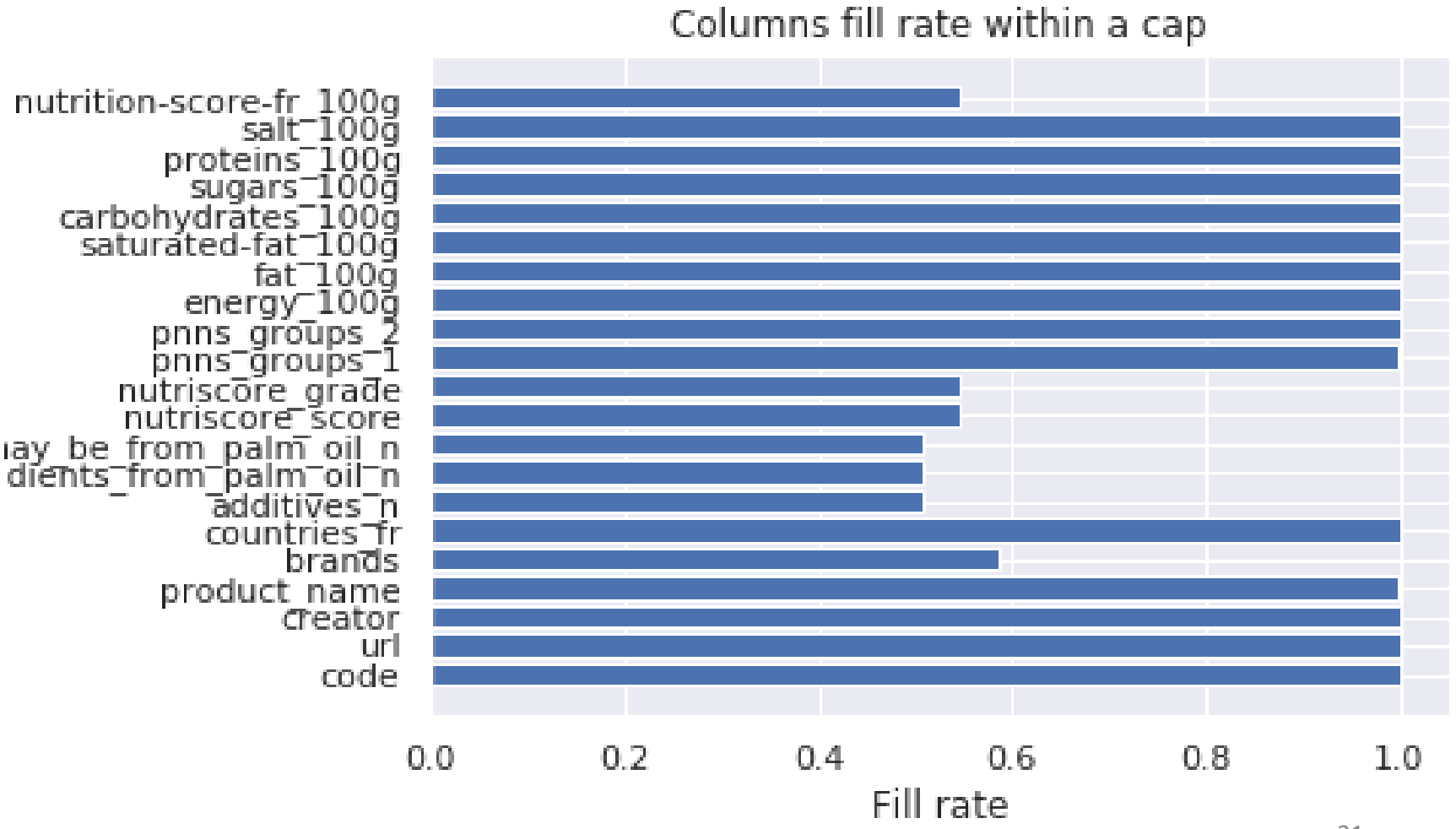
Traitement des doublons

- Prétraitement du tableau de données
  - Génération d'une colonne avec taux de remplissage des lignes
  - Tri du tableau suivant ce taux de remplissage
- Suppression des doublons
  - Rétention de la première occurrence dans le tableau trié
- Restauration de la forme et de l'ordre avant traitement.

Imputation statistique

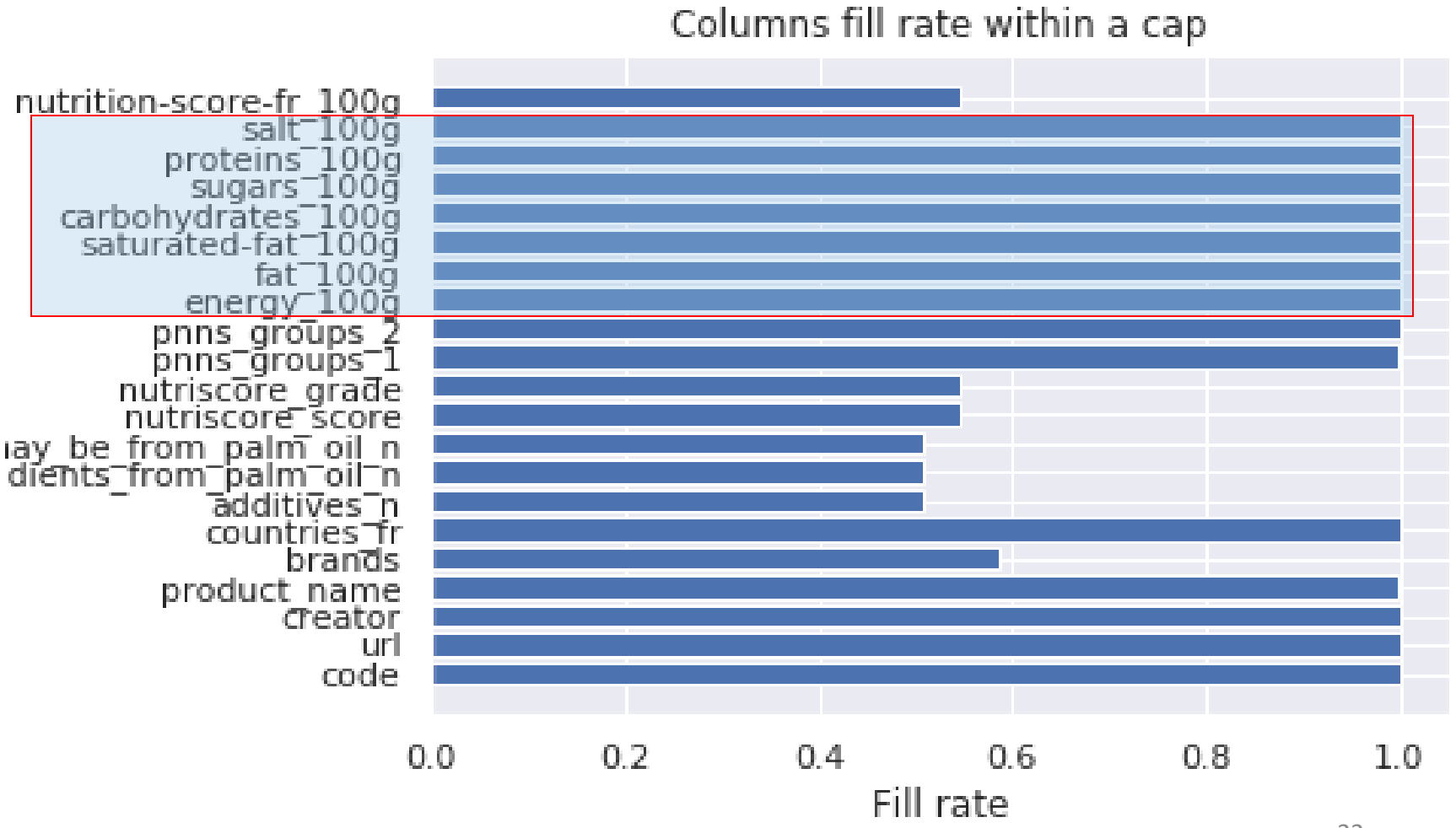
Imputation Statistique

- Iterative imputer
- Sur les **valeurs nutritionnelles** moyennes **uniquement**
  - Risque de génération de biais sur les autres variables
- Suppression pure et simple des autres lignes



Imputation Statistique

- Iterative imputer
- Sur les **valeurs nutritionnelles** moyennes **uniquement**
  - Risque de génération de biais sur les autres variables
- Suppression pure et simple des autres lignes



Synthèse du nettoyage des données

Jeu de données des "Open Food Facts" sur les produits alimentaires

N = 1 437 214 individus, 181 caractéristiques (115 quantitatives)

Démarche de nettoyage

(1) Nettoyage

- Filtre "passe-haut" à taux de remplissage 10%, "passe-bande" conditionnel –pertinence- sur 10-40%.
- Filtre de non-pertinence: Suppression des variables de type métadonnées (date, tags, url, traductions, etc.) et de n-1 variables redondantes dans les n-uplets de variables fortement corrélées.
- Filtre sur valeurs aberrantes de valeurs nutritionnelles moyennes '_100g' : critères basés sur combinaison de données métier et quantiles de distribution

→ **Bilan: Suppression: 88% de lignes, 79% des colonnes**

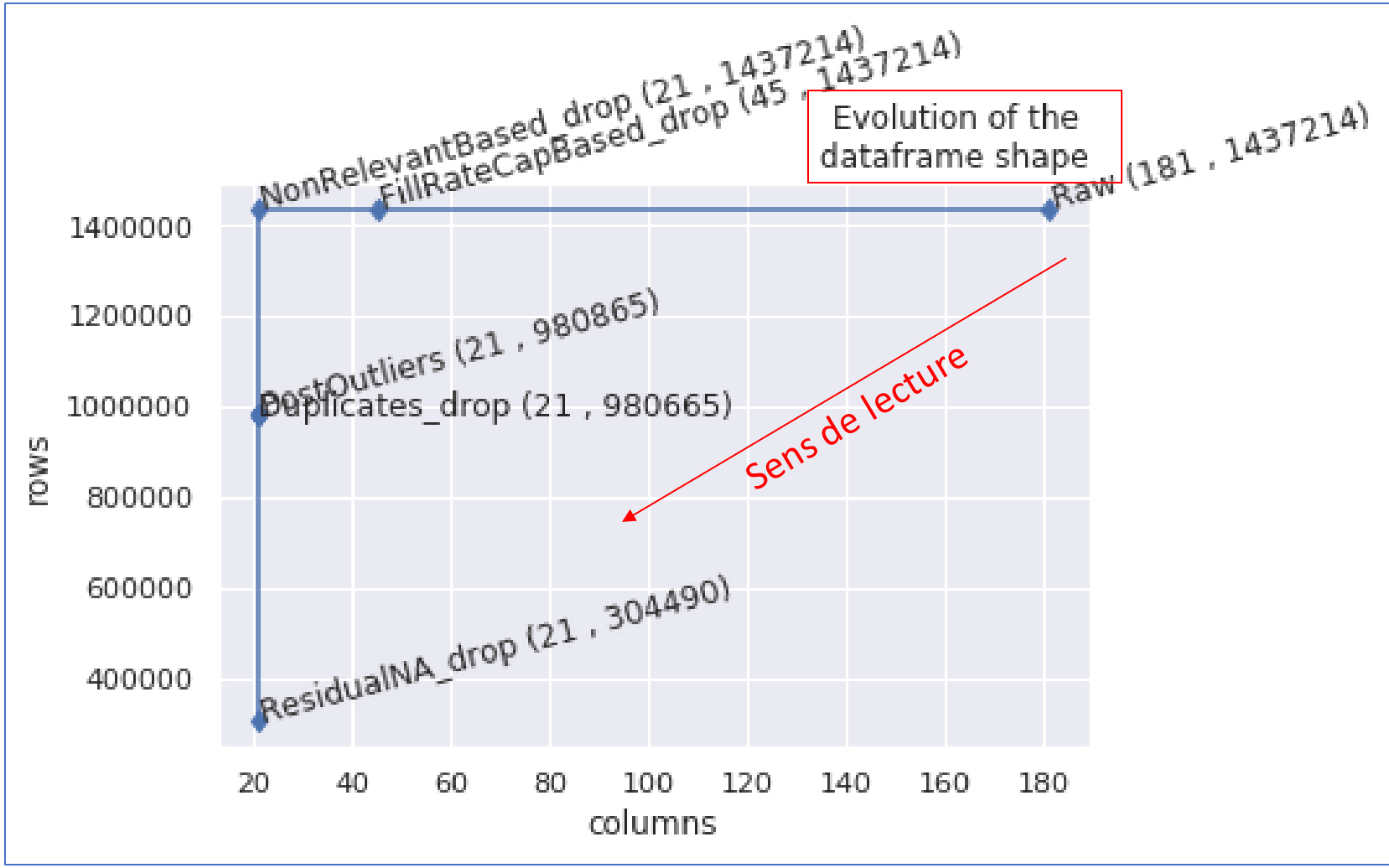
(2) Imputation

- *Iterative imputer* portant sur les valeurs nutritionnelles moyennes
- Gains sur valeurs manquantes: 0%
- Pas d'imputation sur les autres données car risque de biais statistique

Synthèse du nettoyage des données

Variables pertinentes pour les analyse descriptive et explicative

- **Valeurs nutritionnelles moyennes (ou *Nutrifacts: '*_100g'*)**
- **Pnns_groups***
- **Nutriscore_***

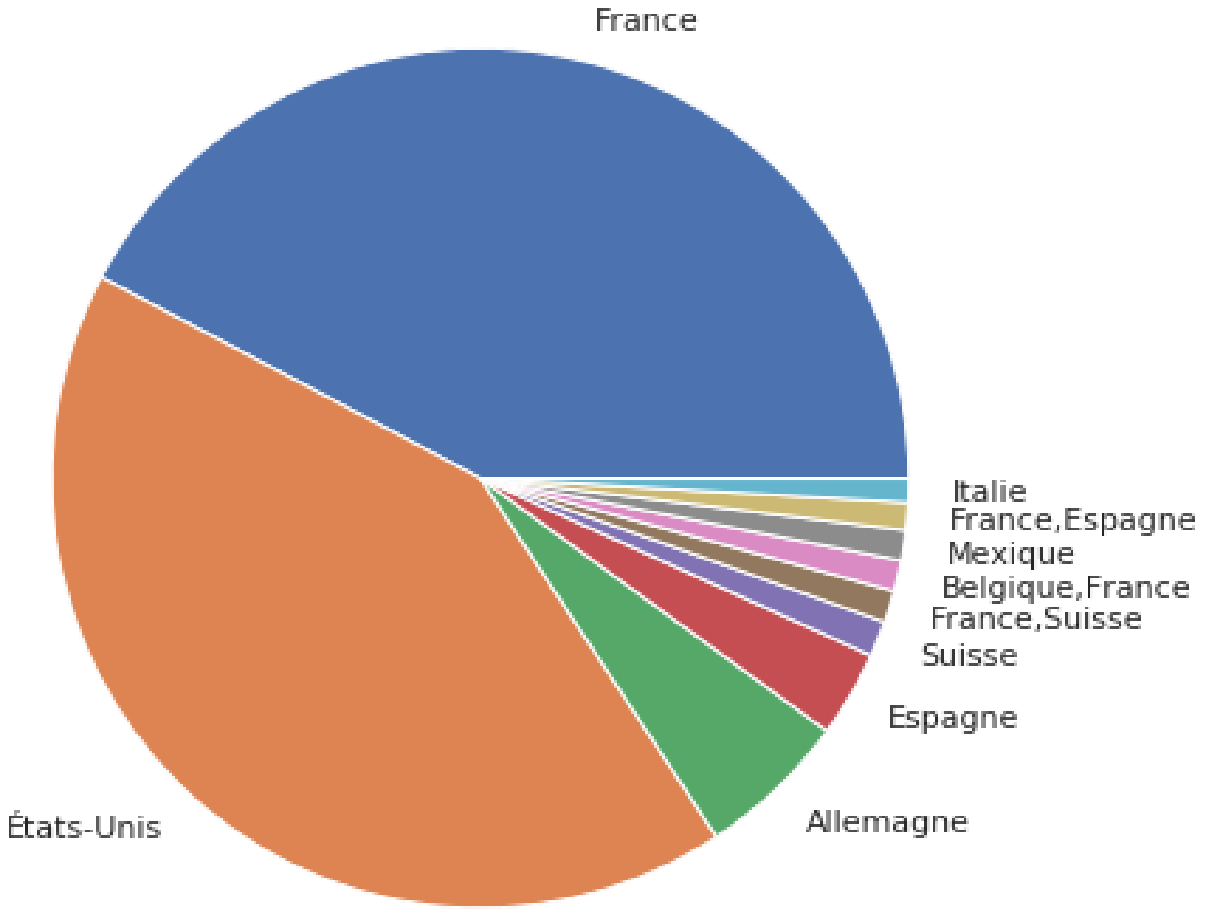


Analyse Univariée

Analyse Univariée

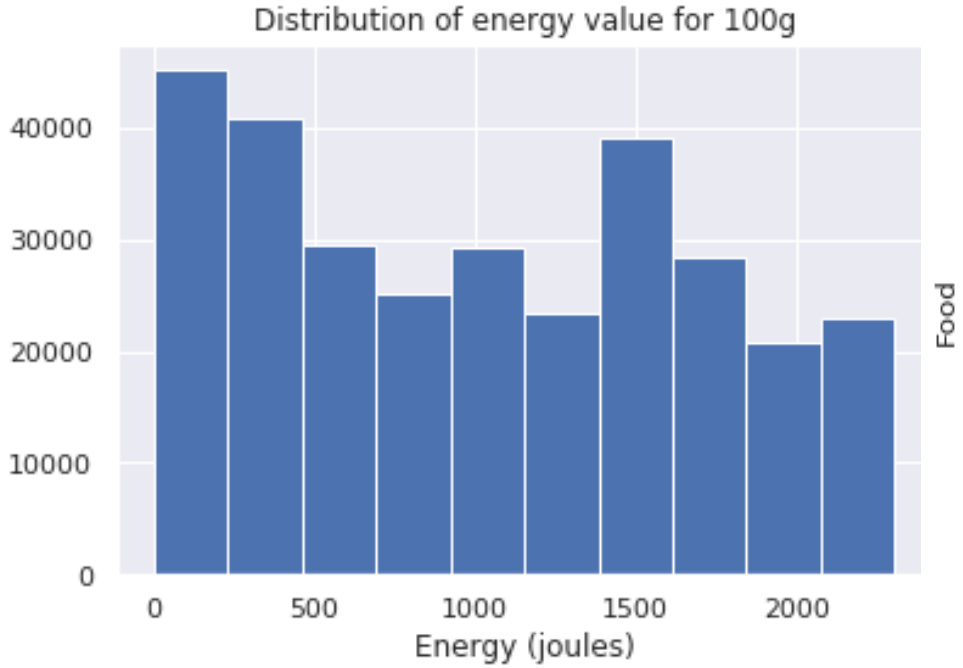
Pays de revente du produit

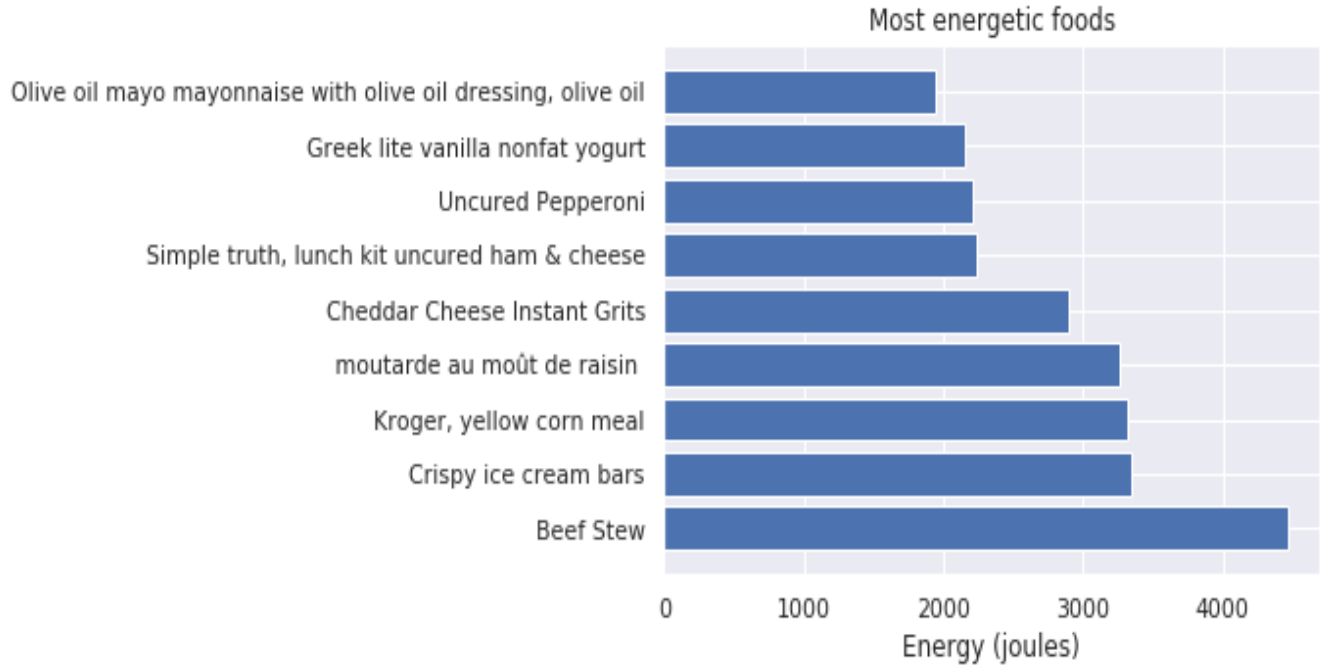
- Liste des pays de reventes, plus précisém.
  - Indicateur de la provenance du contributeur ? -> croisement avec métadonnées
- 622 pays: donnée basée sur l'état actuel (~2% de croissance depuis)
- France: pays principal, sur différentes listes -> Base de données **pertinente pour l'agence de Santé Publique France**



Analyse Univariée

Valeur énergétique

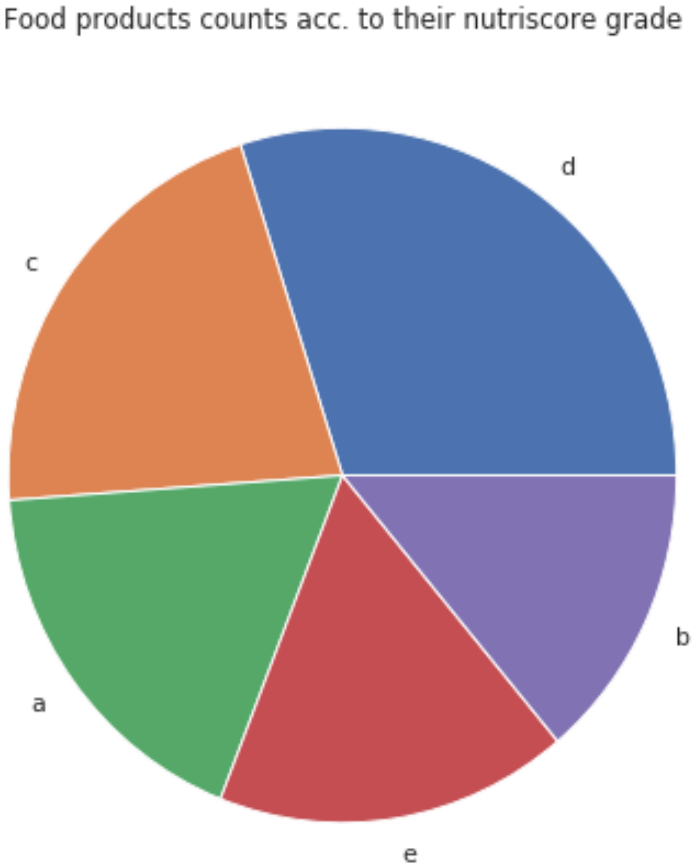


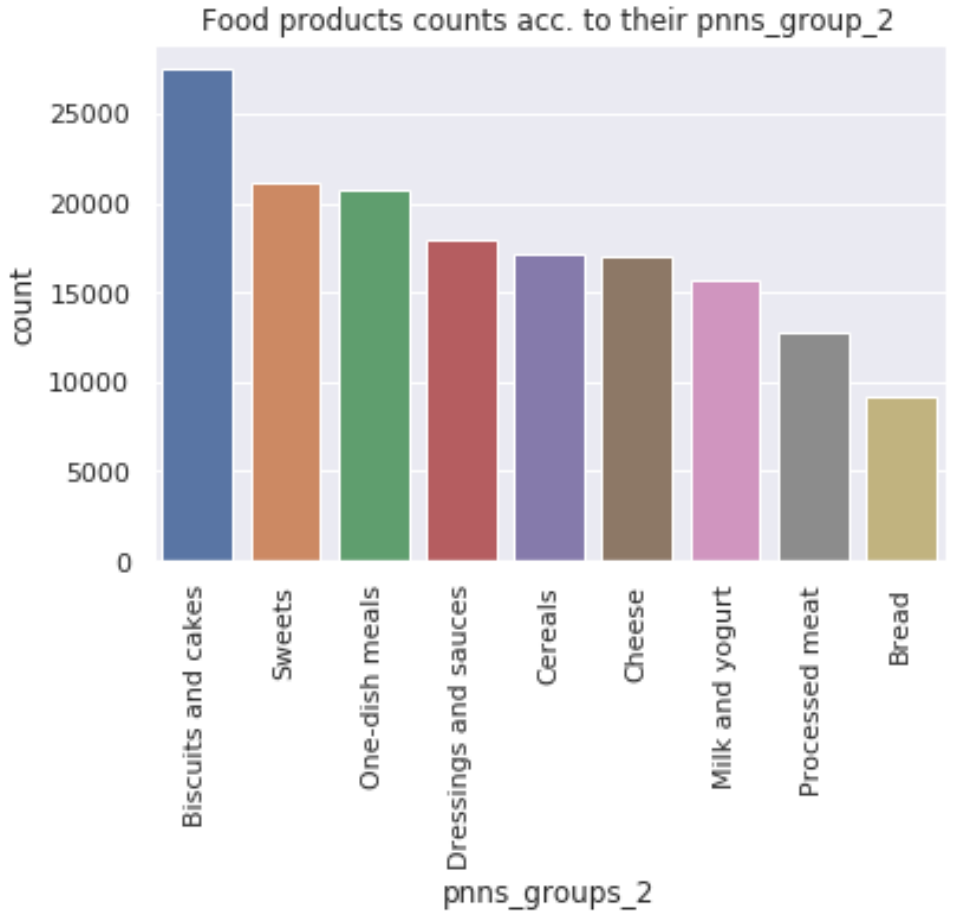


Analyse Univariée

Classifications diverses

- Distributions quasi equimodales
- Intérêt pour des études inclusives



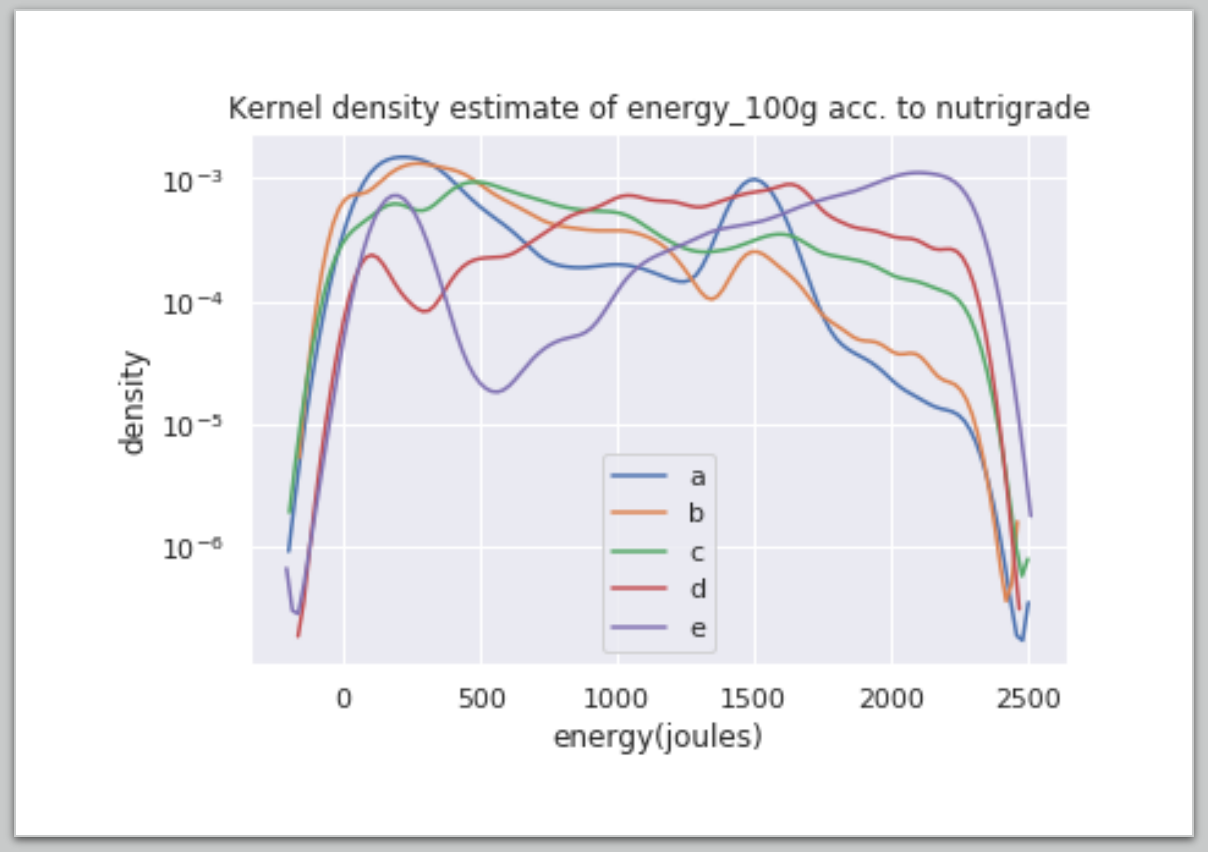


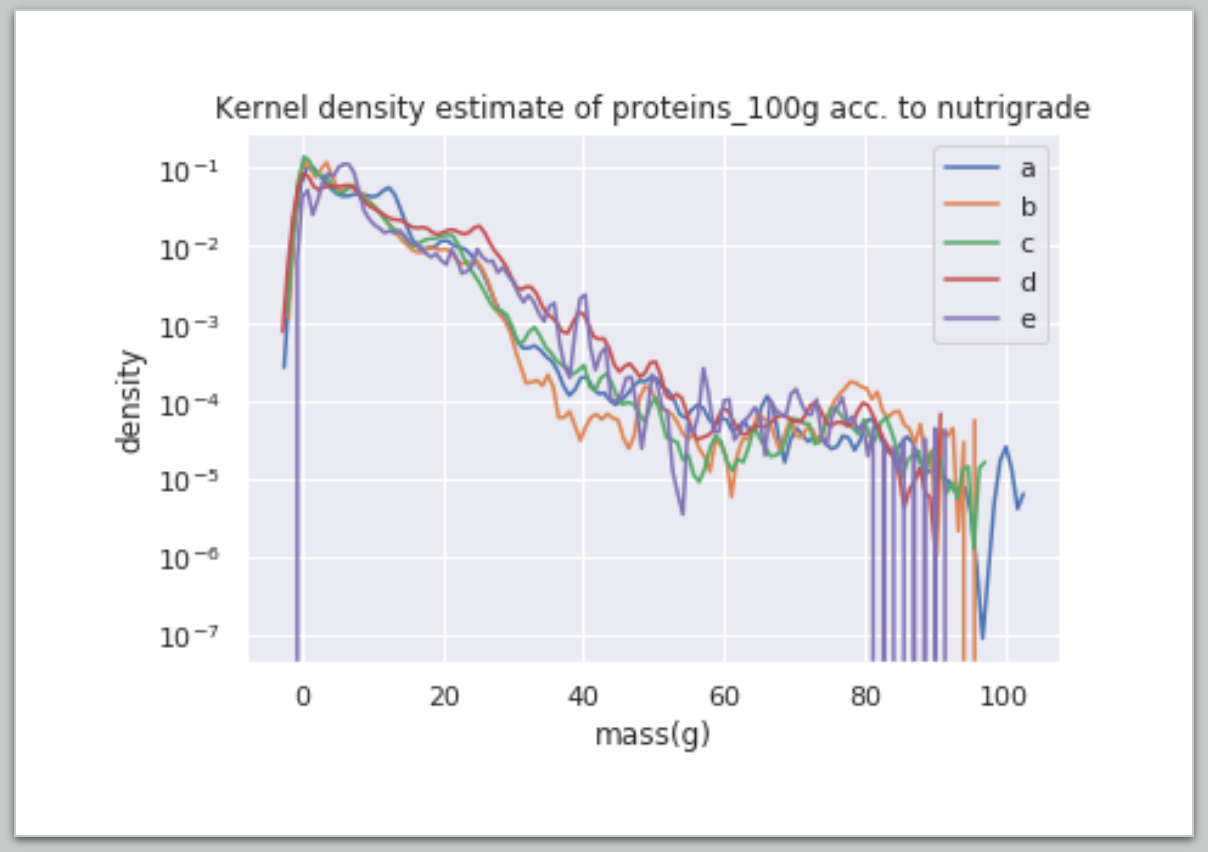
Analyse Multivariée

Analyse Bivariée

Kernel density estimates

- Différences quantitatives nettes (forme, tendances centrales)
- Asymétrie claire de densité de population 'a' vs 'e' pour les valeurs énergétiques
- Creux de densité de population plus important pour les hautes valeurs de nutrigrade 'a', 'b' pour les protéines

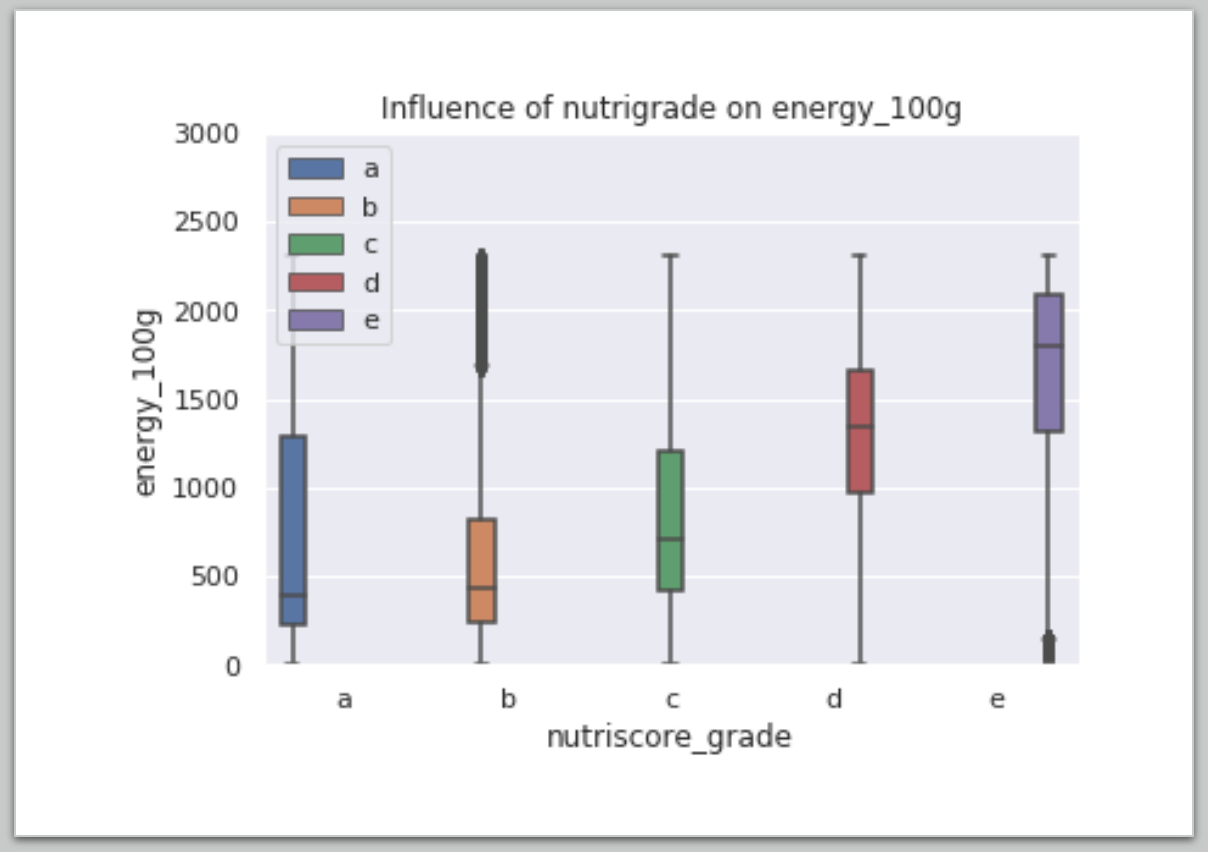


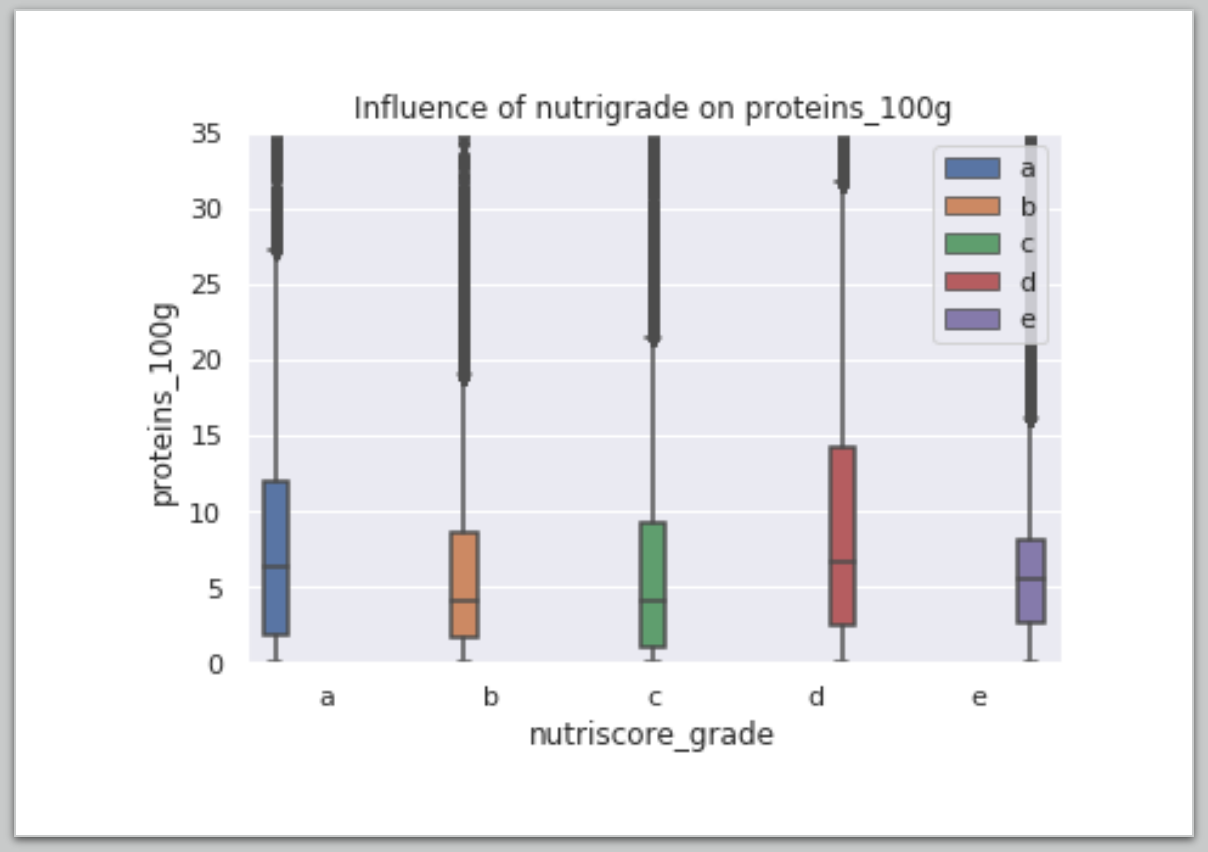


Analyse Bivariée

Grouped boxplot

- Médianes bien distinctes
- Influence des valeurs moyennes nutritionnelles sur le nutrigrade.

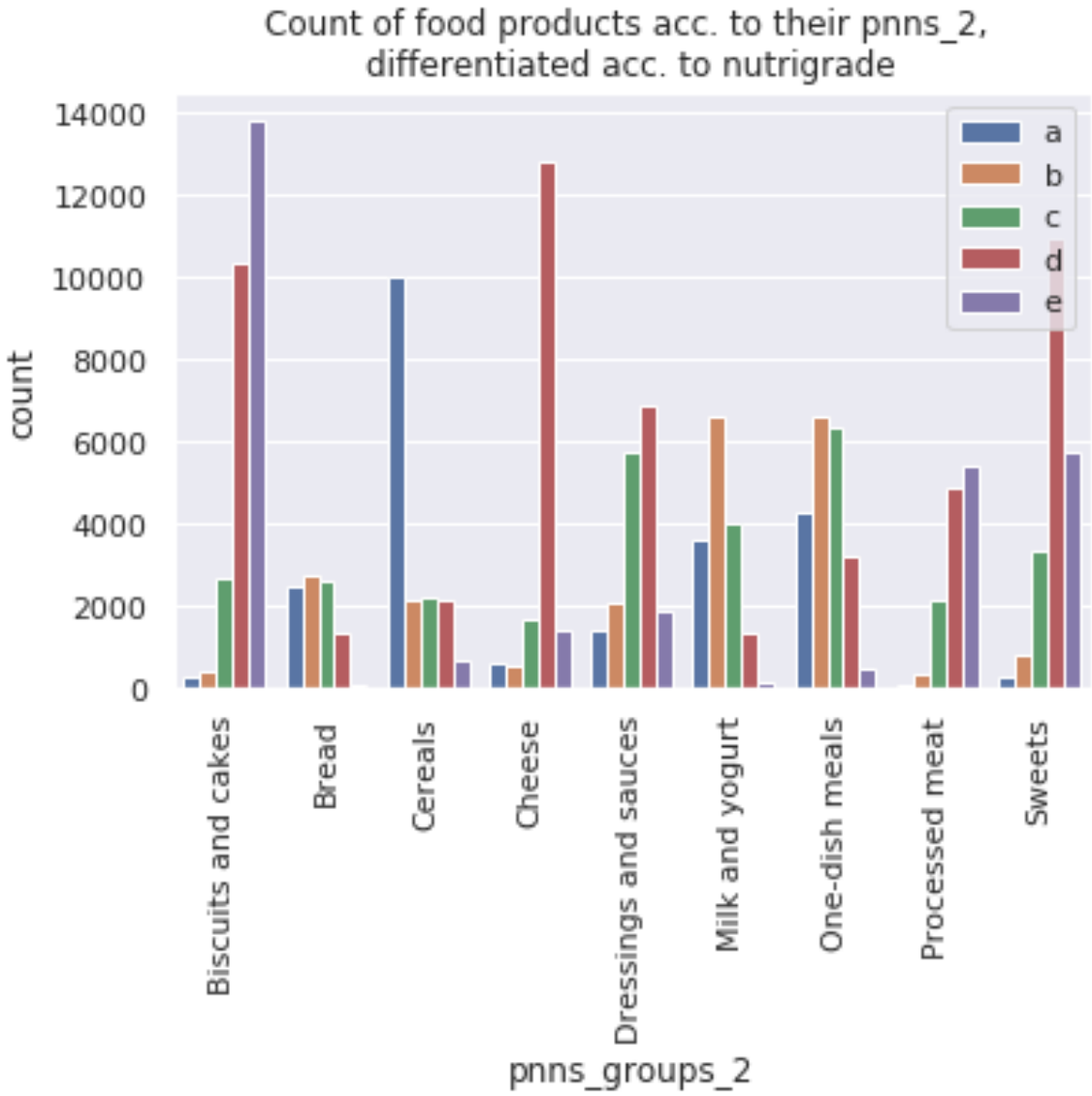




Analyse Bivariée

Classe d'aliments & Nutrigrade

- Différence bien marquée entre:
  - •les catégories à grandes proportions de score **'d' et 'e'** : **Biscuits & cakes, cheese, sweets**, et
  - •celles à score élevés en **'a' et 'b': cereals, milk & yogurt.**
- La catégorie 'processed meat' n'a pas de produit classé 'a' ou 'b'!
- La catégorie 'bread' se disperse assez sur tout le spectre de score.



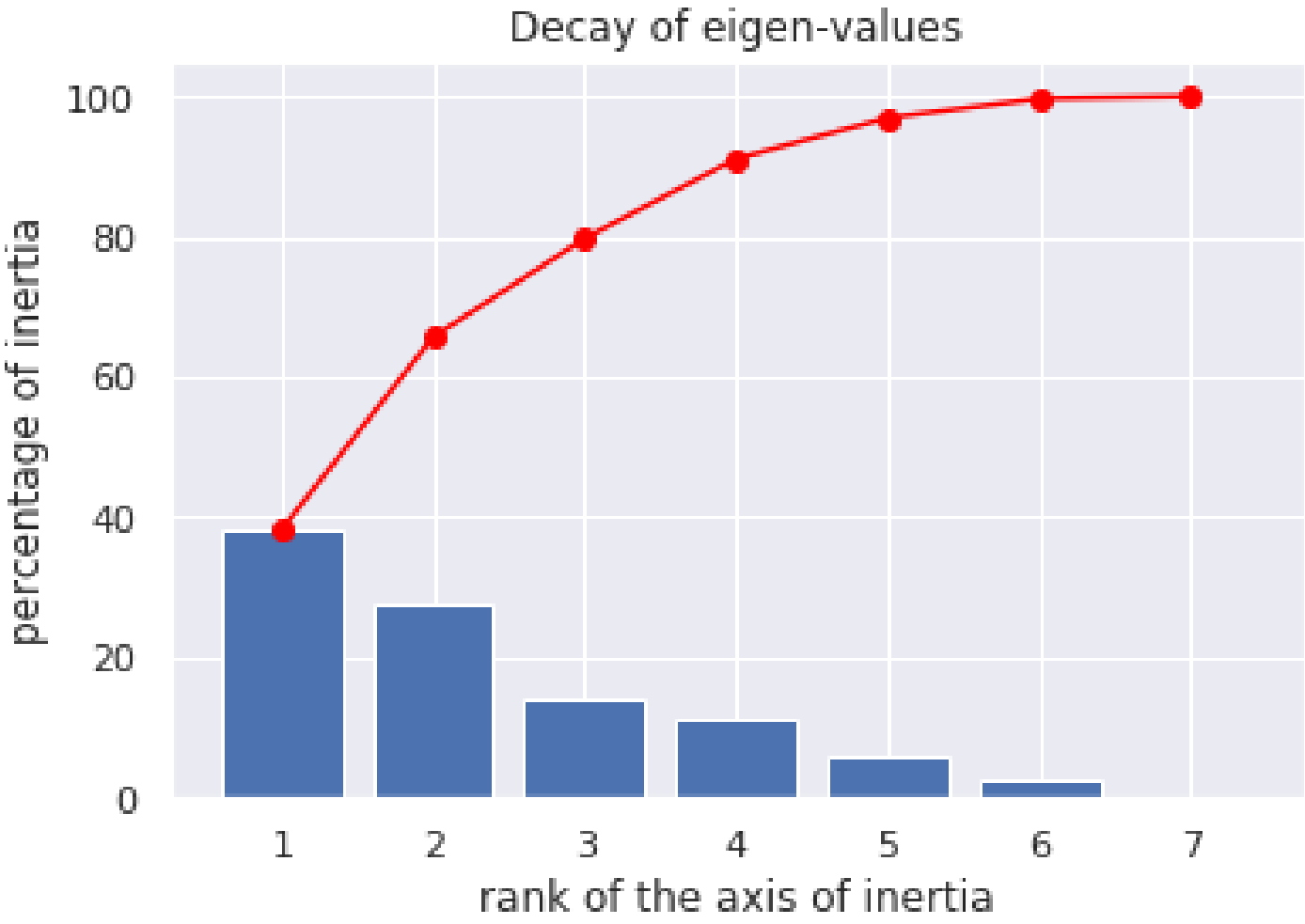
Analyse desriptive

Methode ACP

Méthode ACP

Eboulis des valeurs propres

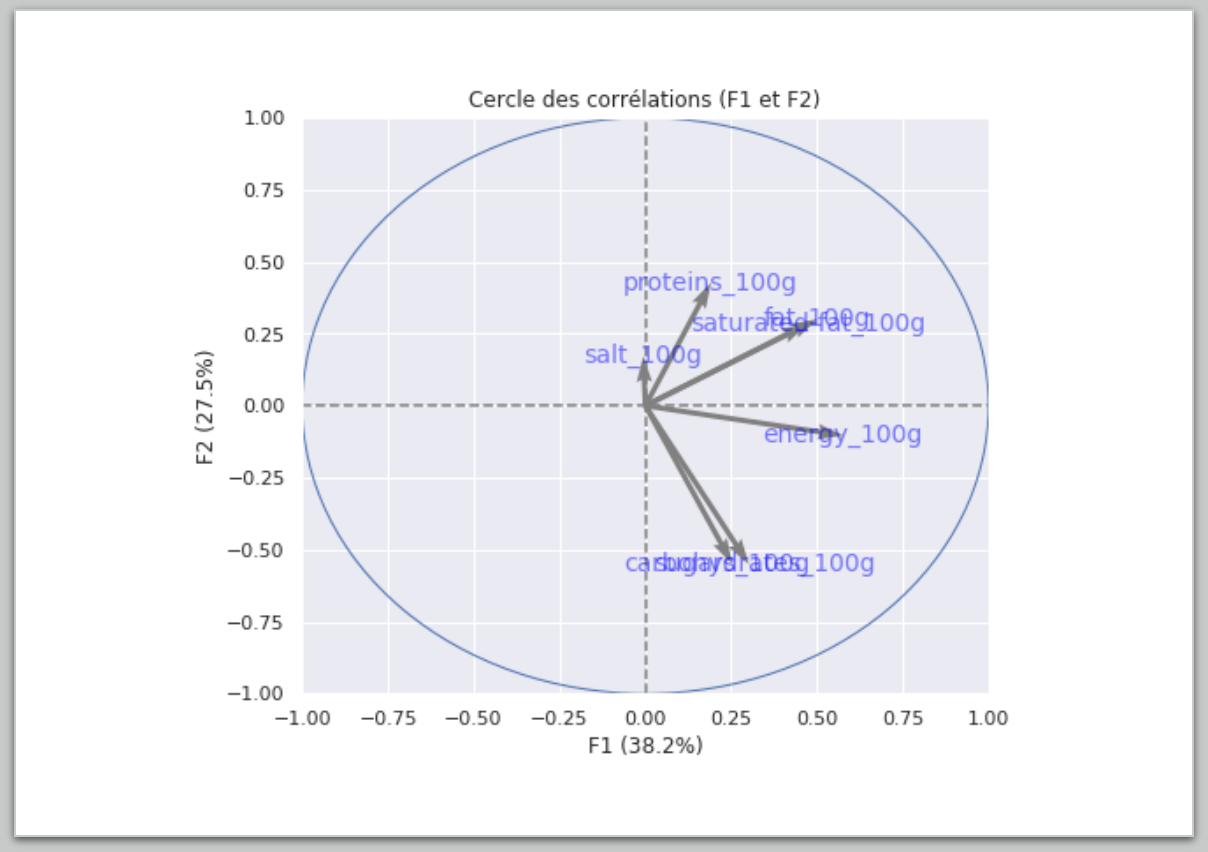
- Les 2 premiers axes principaux d'inertie absorbent 2/3 de l'inertie totale du nuage d'individus.
- Il y a un très léger coude entre le 2nd et le 3e rang.

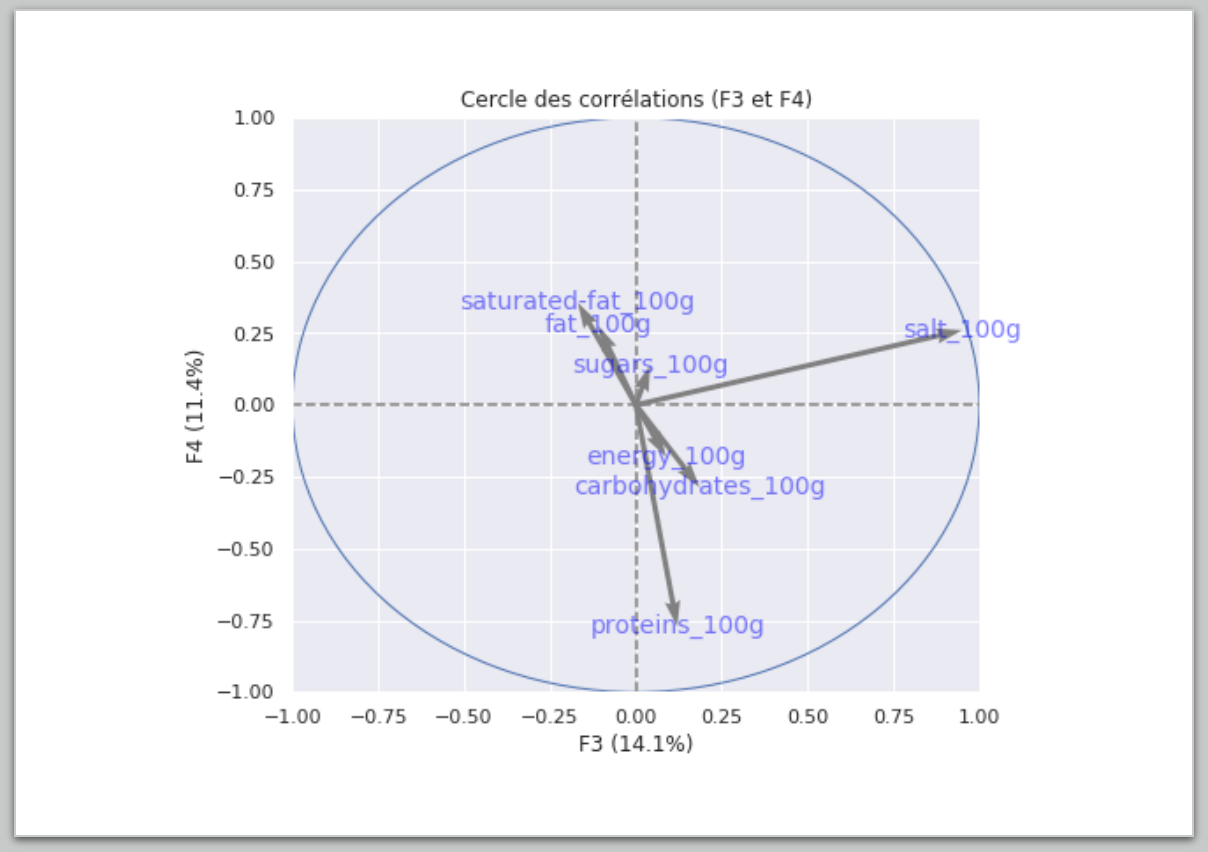


Méthode ACP

Cercles de corrélation, plans factoriels

- (F1,F2): 66%, (F3,F4): 25%.
- Représentation moyennement bonne de 5/6 variables dans (F1,F2).
- Représentation très bonne de 2/6, et très faible du reste, dans (F3, F4).

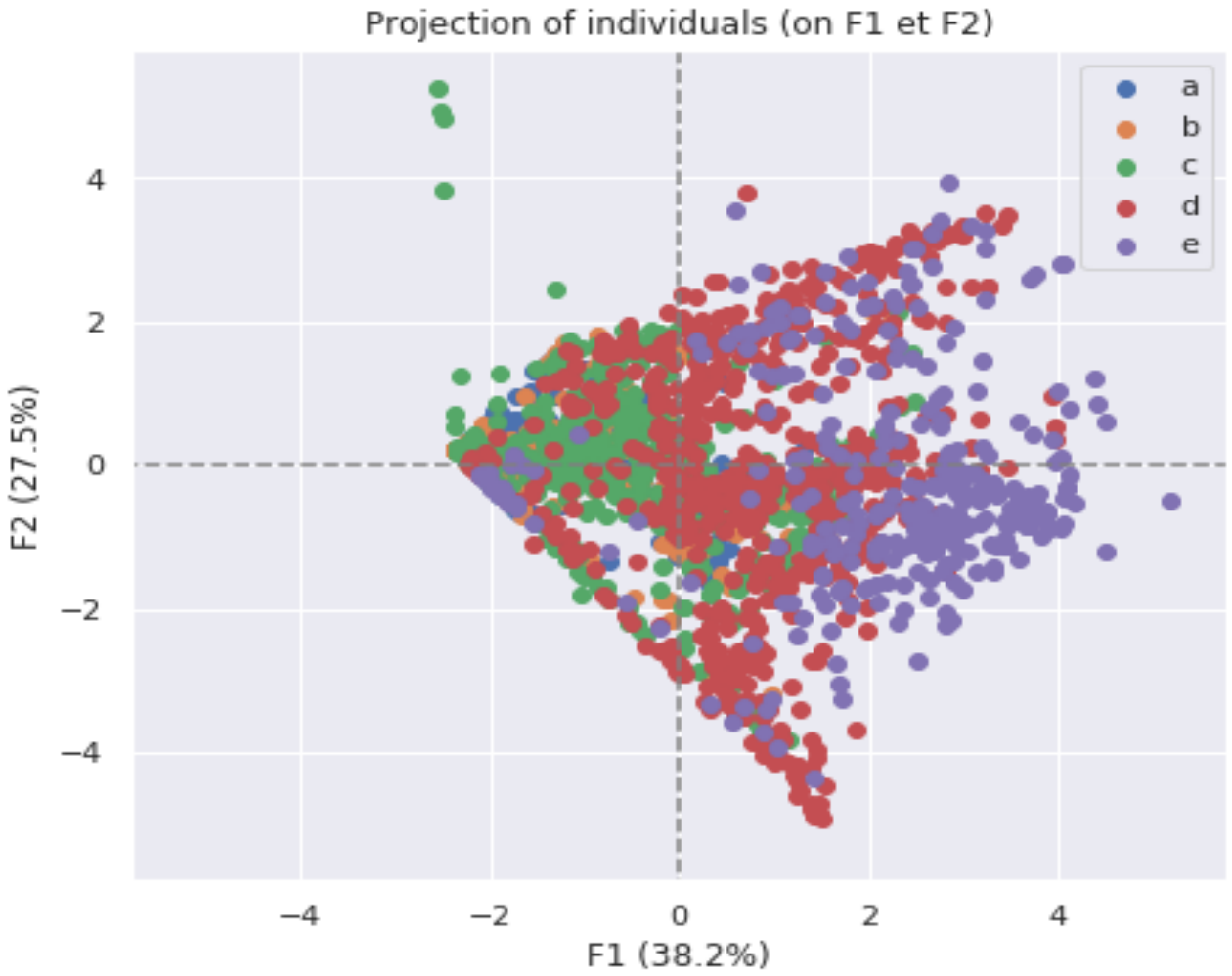




Méthode ACP

Projection des individus sur (F1,F2)

- Echantillonnage aléatoire de 2000 individus.
- Séparation bien disctincte des produits en fonction de leurs nutrigrades.
- Quid de la représentativité des individus par le plan ?



Analyse explicative

Test de Kruskal Wallis

Test de Kruskal-Wallis

Données initiales

- Variable aléatoire: valeur nutritionnelle moyenne d'un produit.
- Variable illustrative: **nutrigrade.**
- Hypothèse h0**: Pour une valeur nutritionnelle donnée, les médianes des distributions par nutrigrade sont toutes égales.**

nutrifact	tstat	pvalue
carbohydrates_100g	11902,75429	0,000
energy_100g	86345,68756	0,000
fat_100g	76882,34626	0,000
proteins_100g	5646,813472	0,000
salt_100g	37006,04175	0,000
saturated-fat_100g	89258,44105	0,000
sugars_100g	40356,20629	0,000

Test de Kruskal-Wallis

Résultats

- 7 nutrifacts testés
- P-value <=0.001

--> Hypothèse nulle rejetée: Les médianes des distributions par nutrigrade sont différentes, et ce quelque soit le nutrifact.

Ce résultat quantifie bien ce que nous avions déjà observé graphiquement:

La quantité de la valeur nutritionnelle moyenne influence le nutrigrade du produit alimentaire.

nutrifact	tstat	pvalue
carbohydrates_100g	11902,75429	0,000
energy_100g	86345,68756	0,000
fat_100g	76882,34626	0,000
proteins_100g	5646,813472	0,000
salt_100g	37006,04175	0,000
saturated-fat_100g	89258,44105	0,000
sugars_100g	40356,20629	0,000

Synthèse de l'analyse multivariée

Jeu de données des "Open Food Facts" sur les produits alimentaires

6 variables quantitatives sélectionnées, Hypothèse d'influence de la valeur moyenne nutritionnelle sur le nutrigrade

Démarche d'analyse

(1) Analyse bivariée

- Mise en évidence de l'influence du nutrigrade sur les valeurs nutritionnelles à travers des kdeplot, grouped boxplot, et countplot.

(2) Analyse descriptive : Méthode ACP

- 66% de l'inertie absorbée par le premier plan factoriel.
- Bonne représentation de la majorité des variables dans les 2 premiers plans factoriels.

(2) Analyse explicative: Test de Kruskal Wallis

- Hypothèse d'égalité des médianes des distribution par nutrigrade rejetée sur toutes les valeurs nutritionnelles avec un p-value < 0.001.

Merci pour votre attention

Temps de questions/réponses