OPENCLASSROOMS

Formation Ingénieur Machine Learning

Projet:

Concevez une application au service de la santé publique

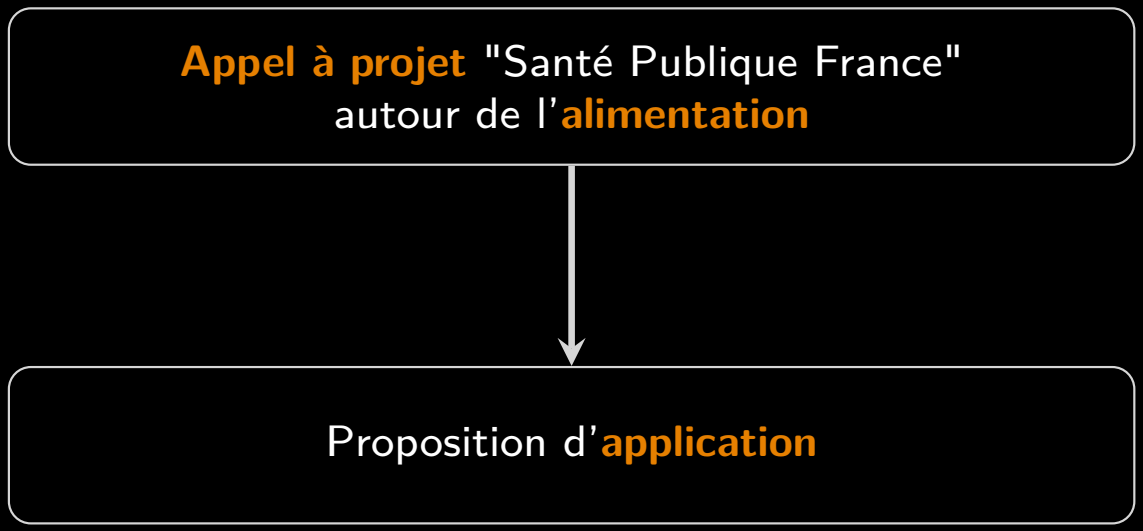
---

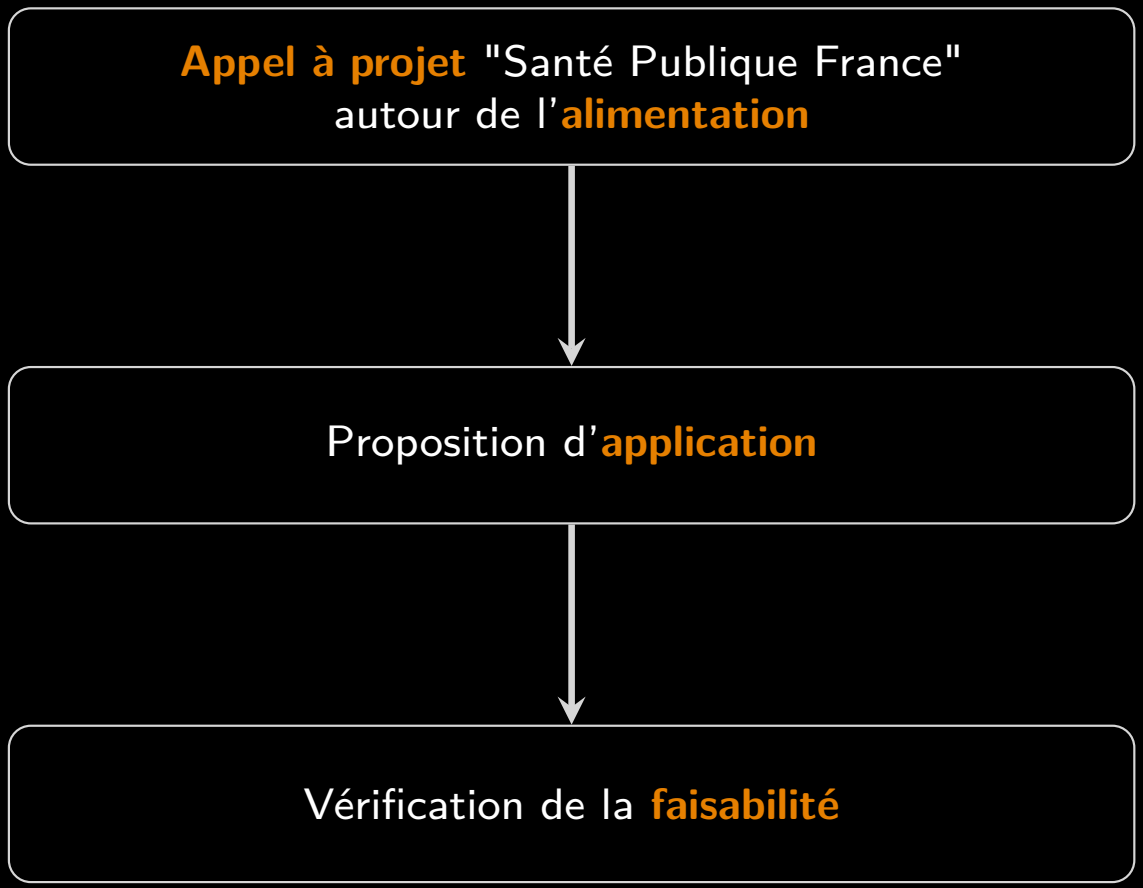
12 Février 2023

thomas.durandtexte@protonmail.com

CONTEXTE

Appel à projet "Santé Publique France"
autour de l'**alimentation**


Appel à projet "Santé Publique France"
autour de l'alimentation
Proposition d'application

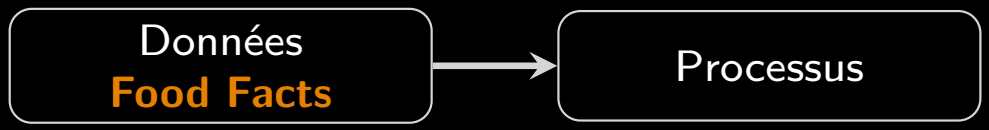

Appel à projet "Santé Publique France"
autour de l'alimentation
Proposition d'application
Vérification de la faisabilité

CONTEXTE

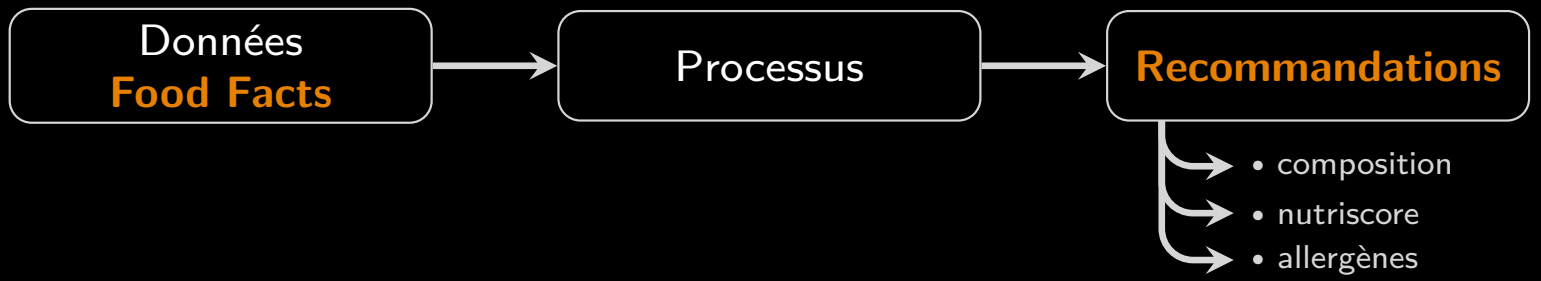
Application proposée

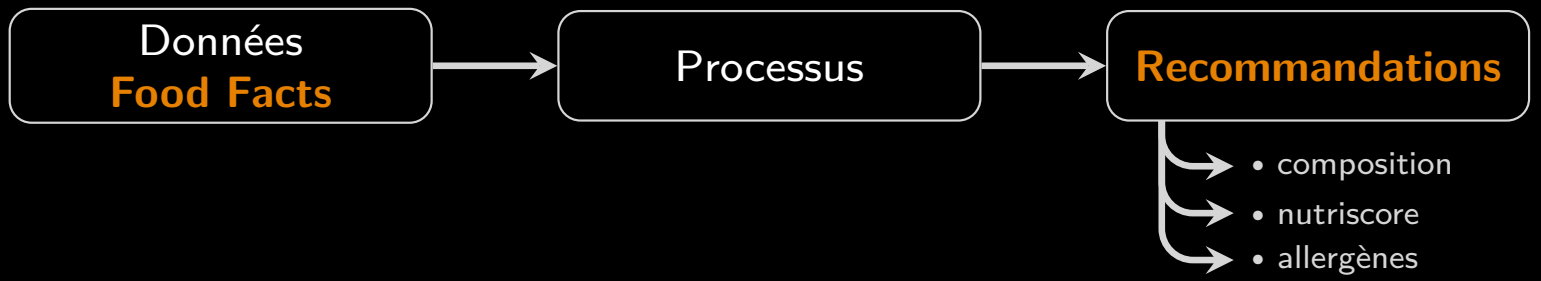
Données

Food Facts

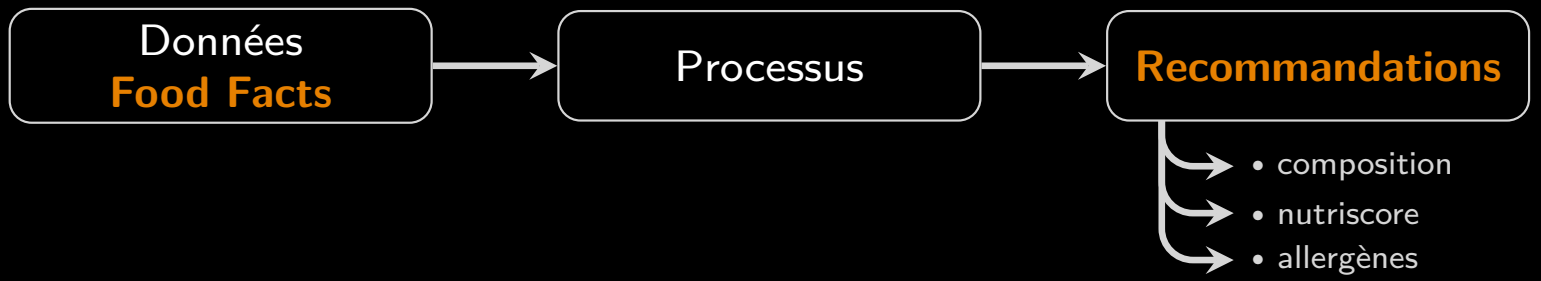

Données
Food Facts
Processus


Données
Food Facts
Processus
Recommandations


Données
Food Facts
Processus
Recommandations
• composition
• nutriscore
• allergènes

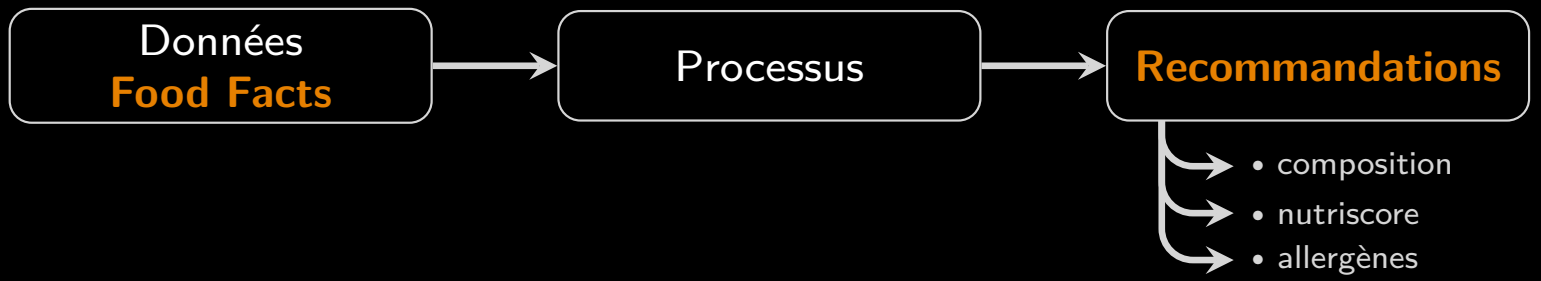


Augmentation des
pathologies cardiaques



Augmentation des
pathologies cardiaques

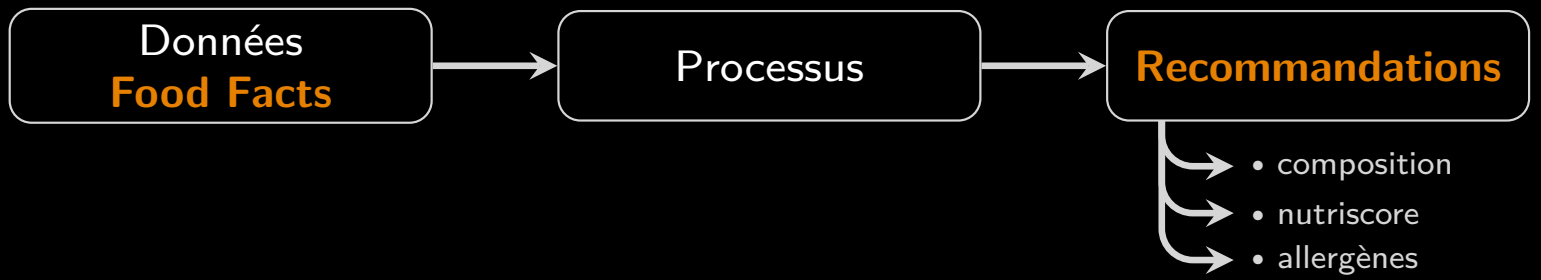
Augmentation de l'**obésité**
et des **diabètes** de type II



Augmentation des
pathologies cardiaques

Augmentation de l'**obésité**
et des **diabètes** de type II

Aide au choix alimentaire
éclairé et **accompagnement**

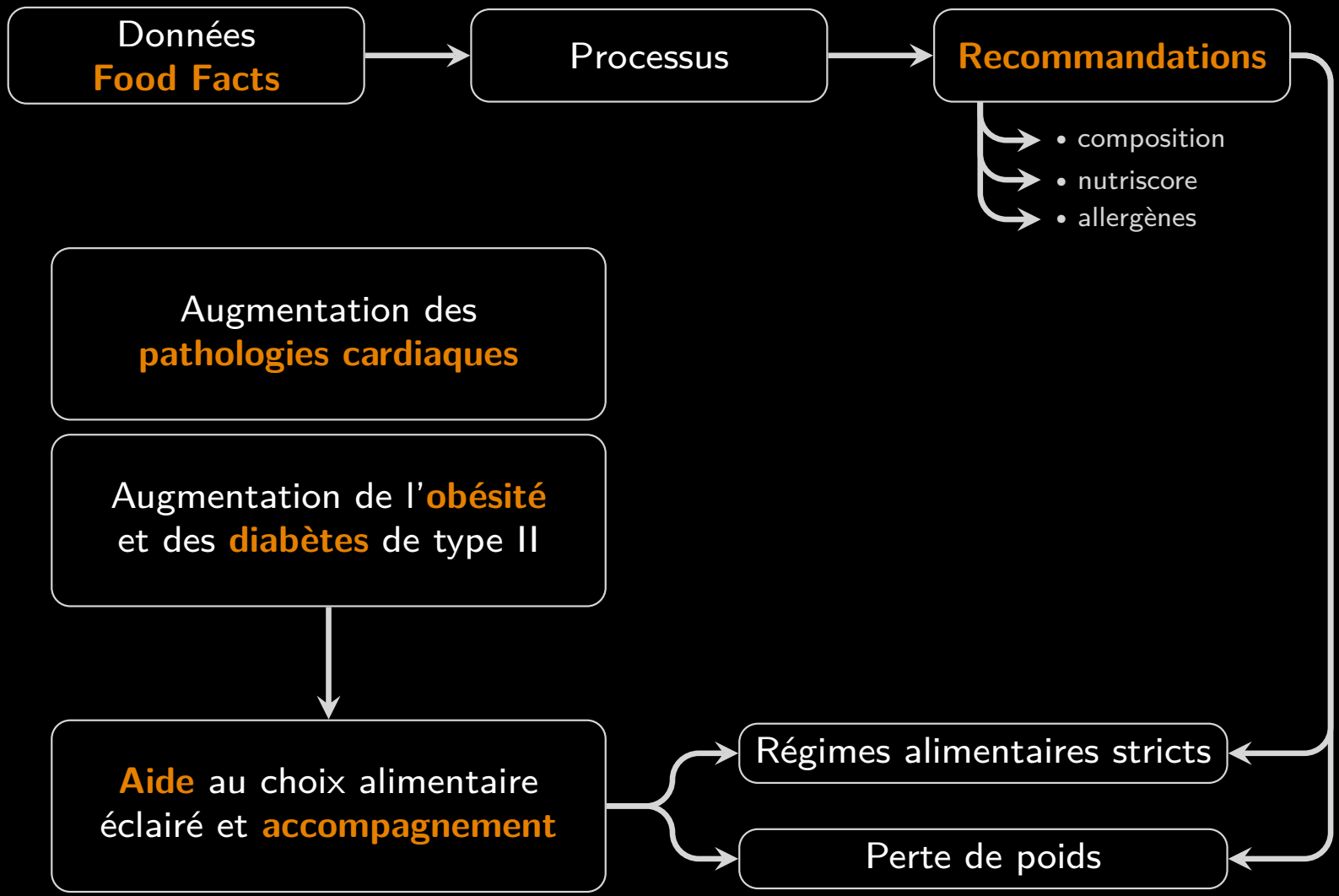


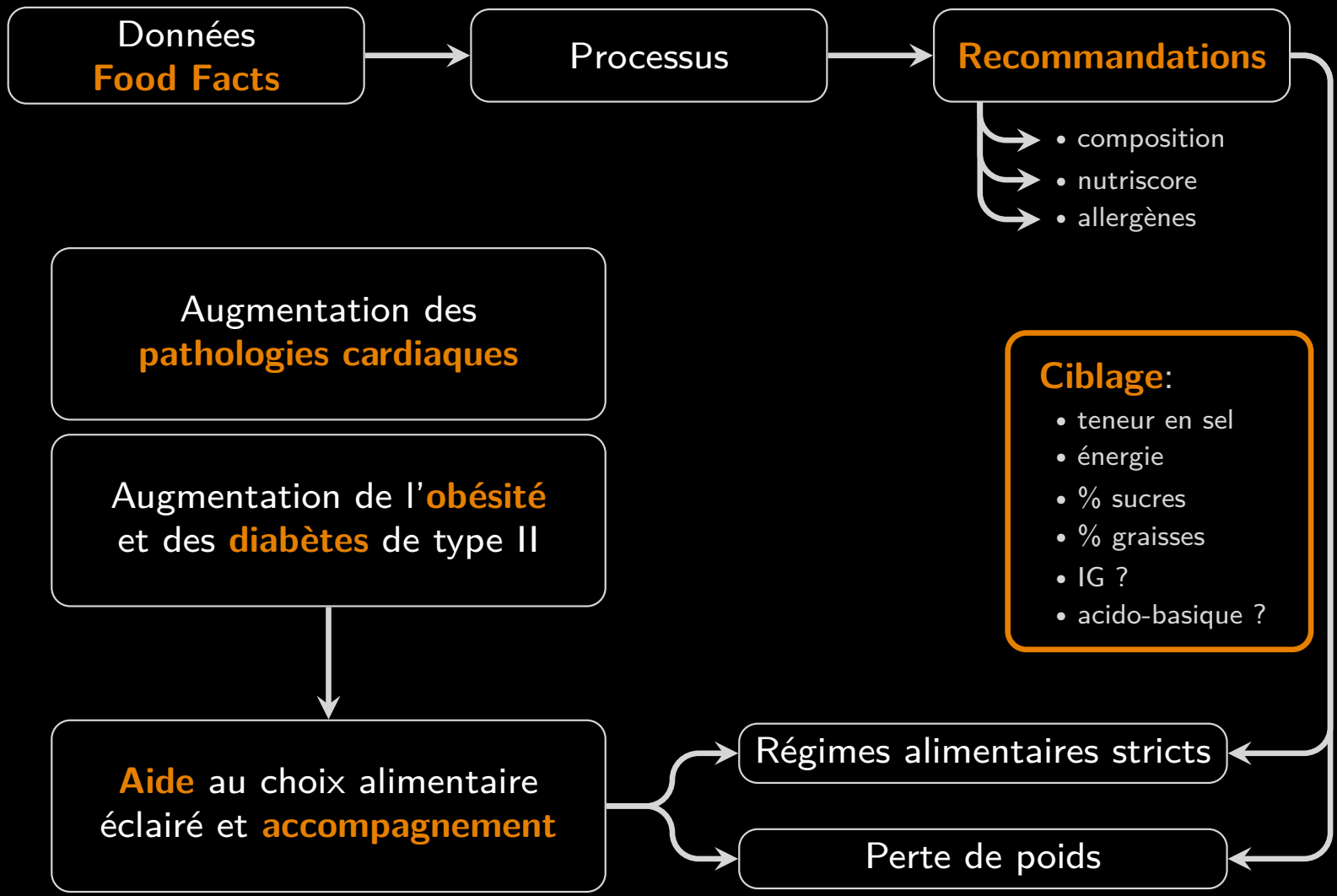
Augmentation des
pathologies cardiaques

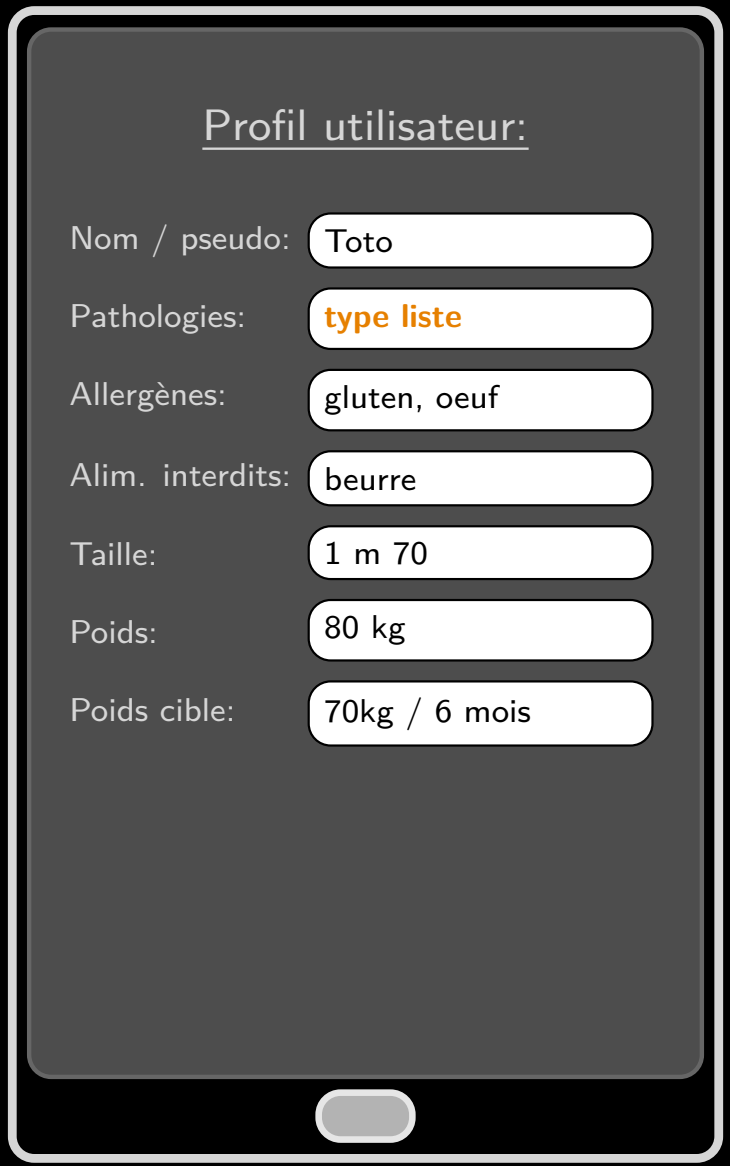
Augmentation de l'**obésité**
et des **diabètes** de type II

Aide au choix alimentaire
éclairé et **accompagnement**

- Régimes alimentaires stricts
- Perte de poids


Données
Food Facts
Processus
Recommandations
• composition
• nutriscore
• allergènes
Augmentation des
pathologies cardiaques
Augmentation de l'obésité
et des diabètes de type II
Aide au choix alimentaire
éclairé et accompagnement
Régimes alimentaires stricts
Perte de poids


Données
Food Facts
Processus
Recommandations
• composition
• nutriscore
• allergènes
Augmentation des
pathologies cardiaques
Augmentation de l'obésité
et des diabètes de type II
Ciblage:
• teneur en sel
• énergie
• % sucres
• % graisses
• IG ?
• acido-basique ?
Aide au choix alimentaire
éclairé et accompagnement
Régimes alimentaires stricts
Perte de poids


Profil utilisateur:
Nom / pseudo:
Toto
Pathologies:
type liste
Allergènes:
gluten, oeuf
Alim. interdits:
beurre
Taille:
1 m 70
Poids:
80 kg
Poids cible:
70kg / 6 mois

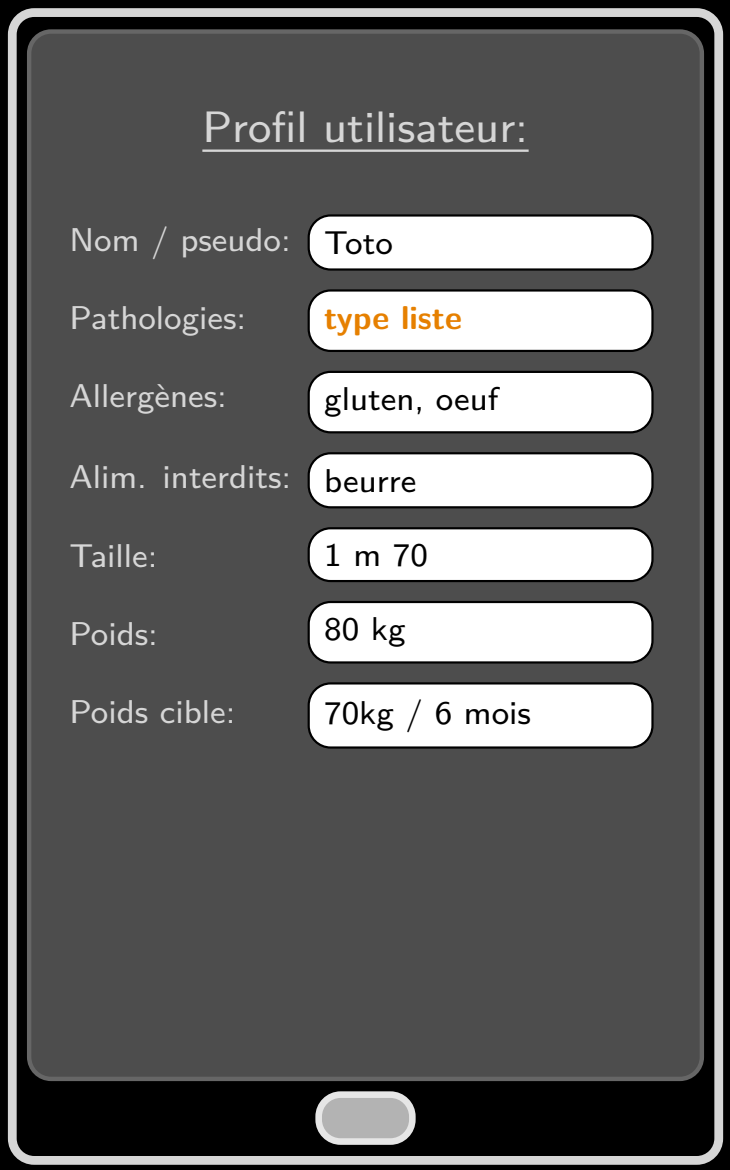
Profil utilisateur:

Nom / pseudo:	Toto
Pathologies:	**type liste**
Allergènes:	gluten, oeuf
Alim. interdits:	beurre
Taille:	1 m 70
Poids:	80 kg
Poids cible:	70kg / 6 mois

Affichage produit:

Nutriscore:	**A**
Énergie pour 100g (kJ):	3700
Sucres pour 100g:	20g
Graisses pour 100g:	10g
Teneur en sel:	2%
Présence d'allergènes:	Aucun
Labels:	bio, non-OGM

Produits similaires recommandés



- **Courbes** IMC/poids fonction du temps
- **Recommandations** de produits en fonction de l'IMC, de la charge calorique et d'un objectif et d'un temps choisi

CONTEXTE

Data set

800 000

800 000

produits référencés (pour les données étudiées)

800 000

produits référencés (pour les données étudiées)

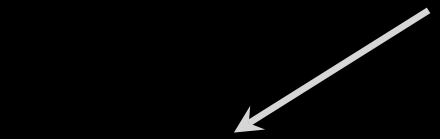
191

800 000
produits référencés (pour les données étudiées)

191
variables

800 000
produits référencés (pour les données étudiées)

191
variables

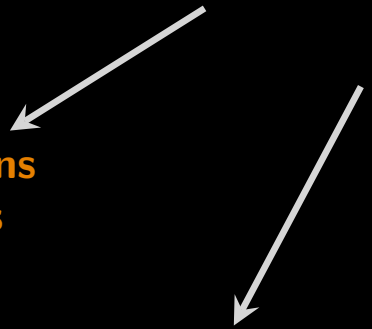


Informations générales

- *code*
- *nom*
- *...*

800 000
produits référencés (pour les données étudiées)

191
variables



Informations générales

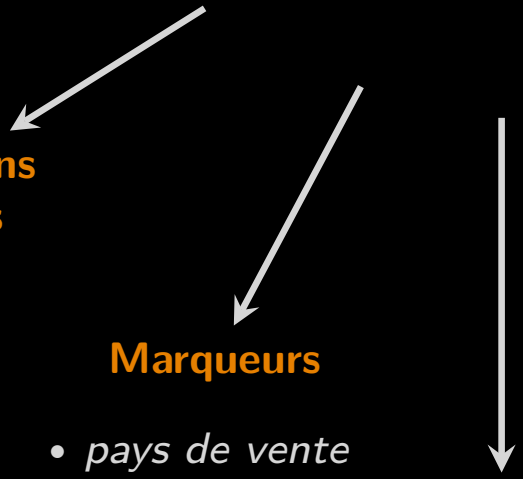
- *code*
- *nom*
- *...*

Marqueurs

- *pays de vente*
- *labels*
- *...*

800 000
produits référencés (pour les données étudiées)

191
variables



Informations générales

- *code*
- *nom*
- *...*

Marqueurs

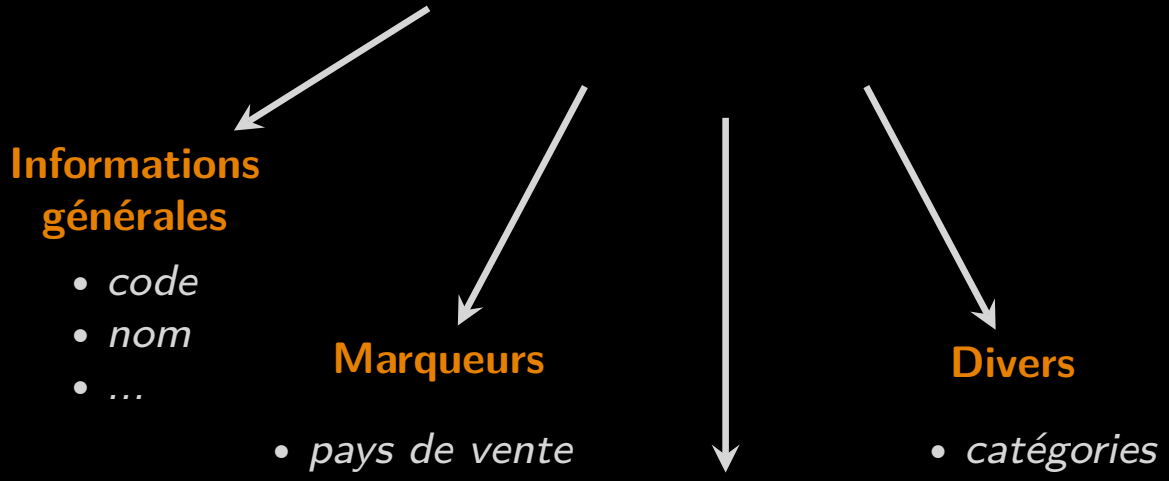
- *pays de vente*
- *labels*
- *...*

Ingrédients et traces

- *ingredients text*
- *traces (allergènes)*
- *...*

800 000
produits référencés (pour les données étudiées)

191
variables



Informations générales
- *code*
- *nom*
- *...*

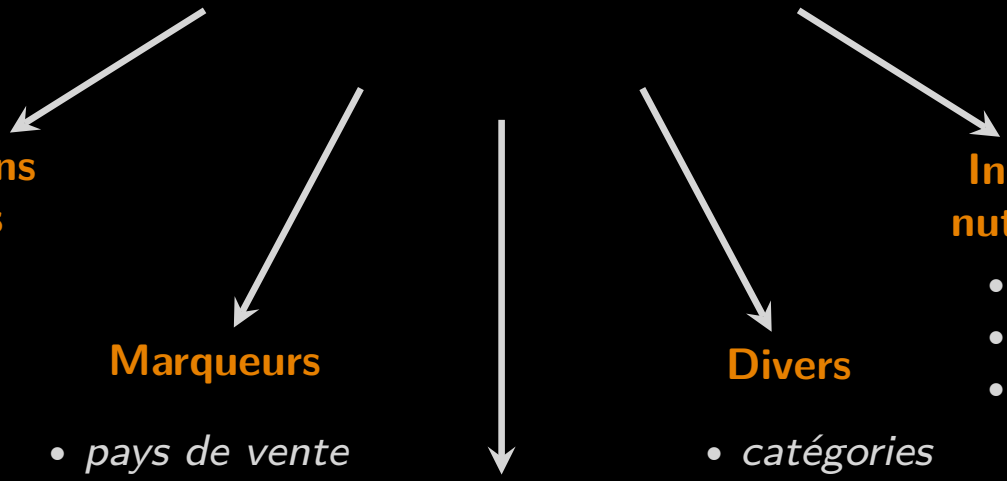
Marqueurs
- *pays de vente*
- *labels*
- *...*

Ingrédients et traces
- *ingredients text*
- *traces (allergènes)*
- *...*

Divers
- *catégories*
- *additifs*
- *...*

800 000
produits référencés (pour les données étudiées)

191
variables



Informations générales
- *code*
- *nom*
- *...*

Marqueurs
- *pays de vente*
- *labels*
- *...*

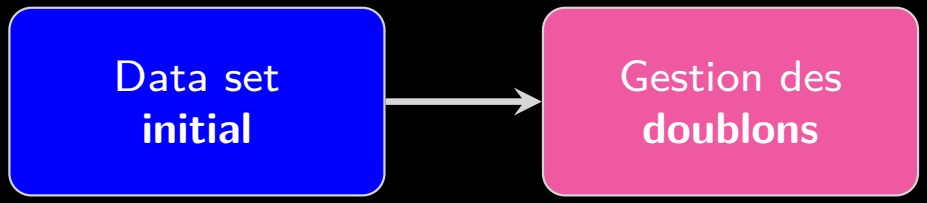
Ingrédients et traces
- *ingredients text*
- *traces (allergènes)*
- *...*

Divers
- *catégories*
- *additifs*
- *...*

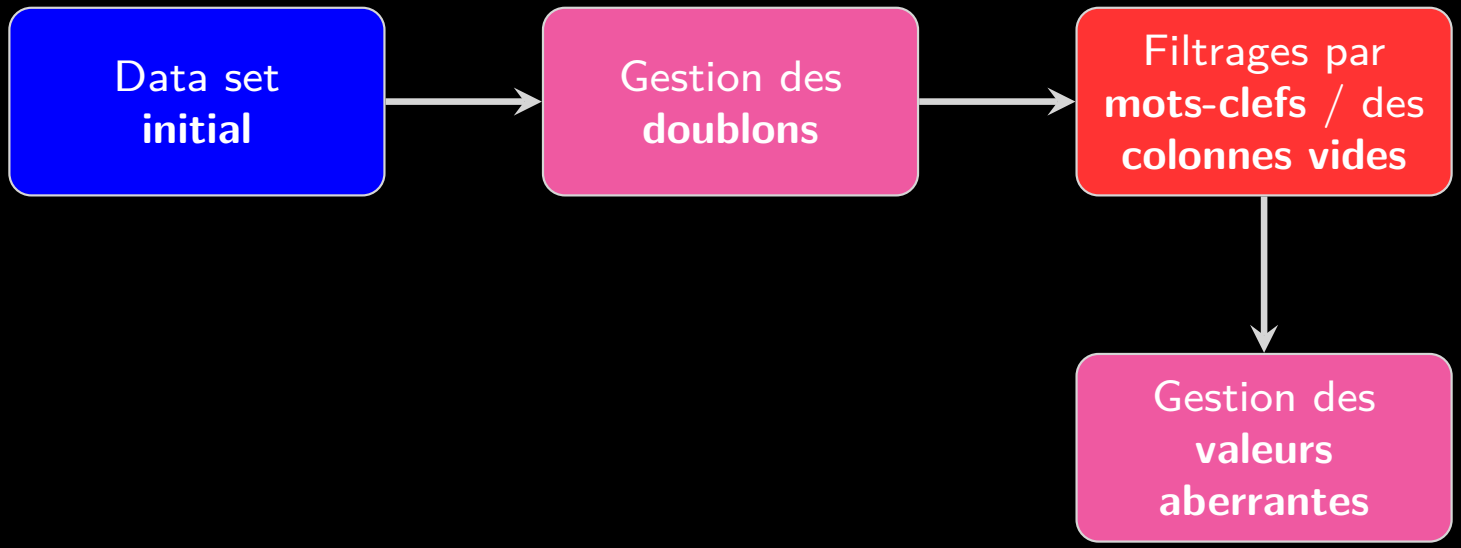
Informations nutritionnelles
- *protéines*
- *graisses*
- *...*

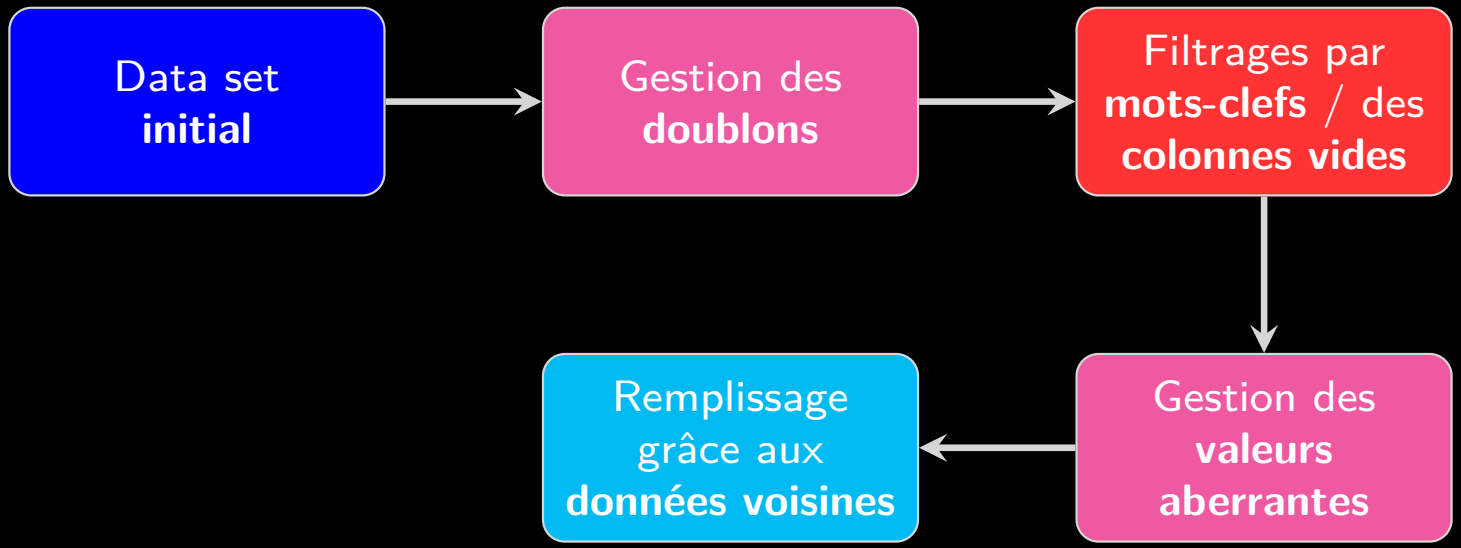
NETTOYAGE

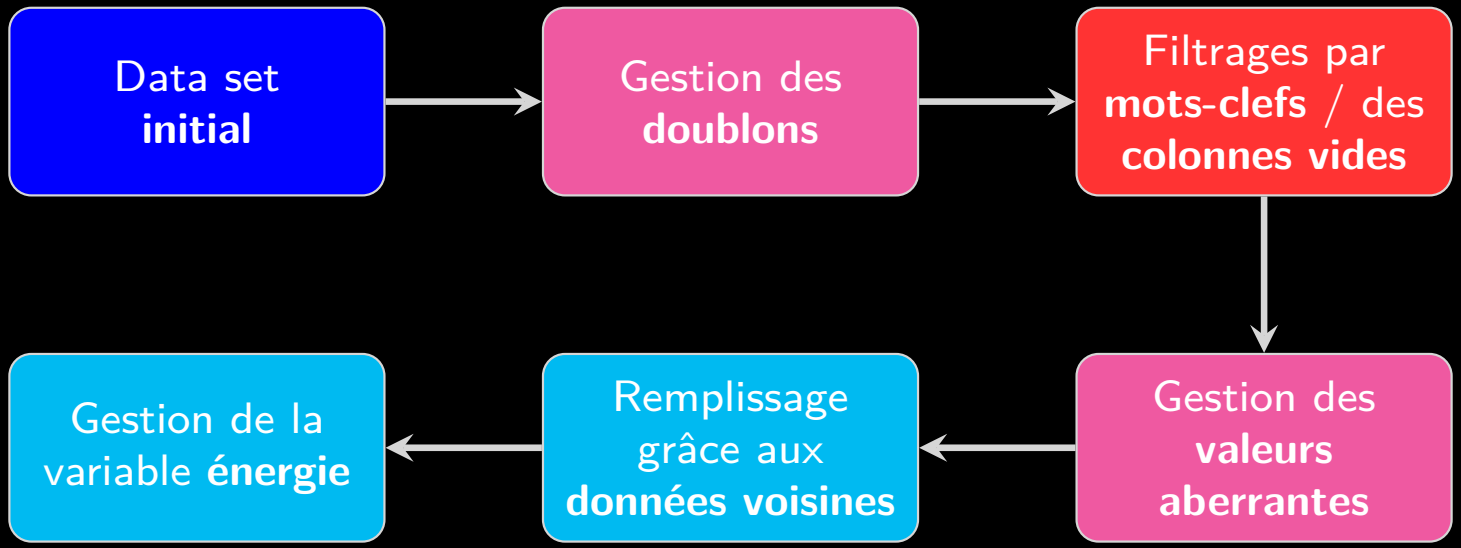

Data set
initial

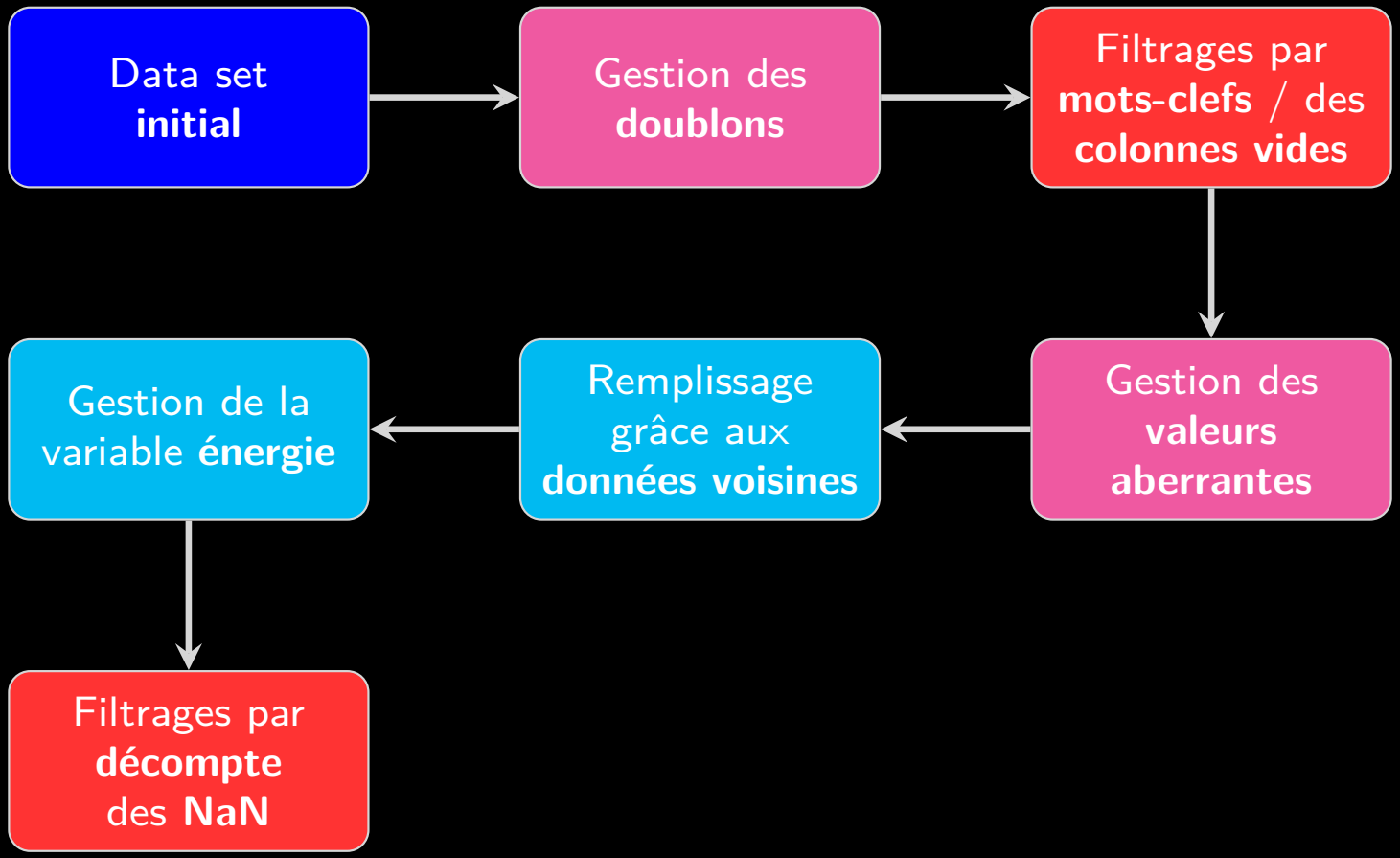

Data set
initial
Gestion des
doublons

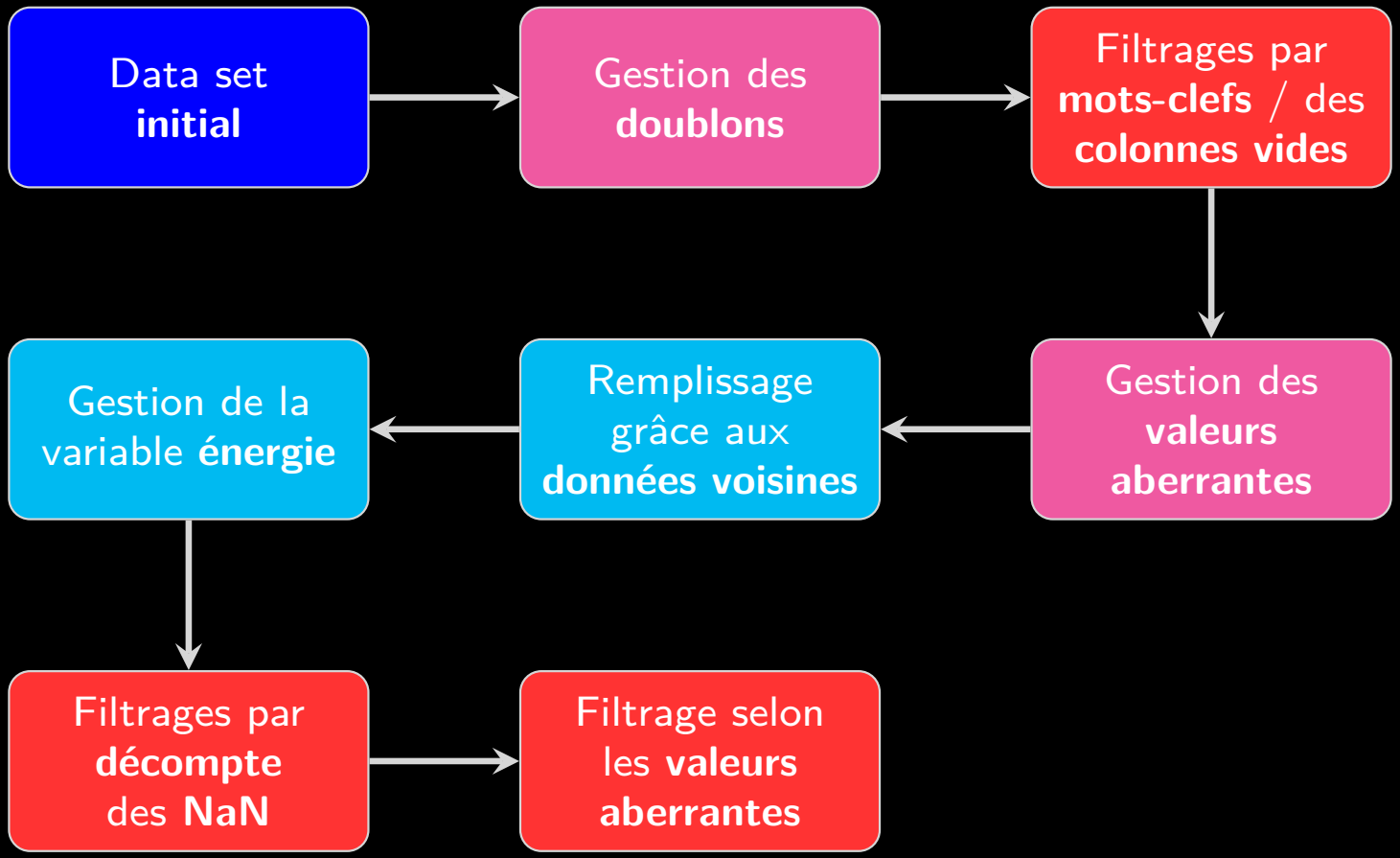

Data set
initial
Gestion des
doublons
Filtrages par
mots-clefs / des
colonnes vides

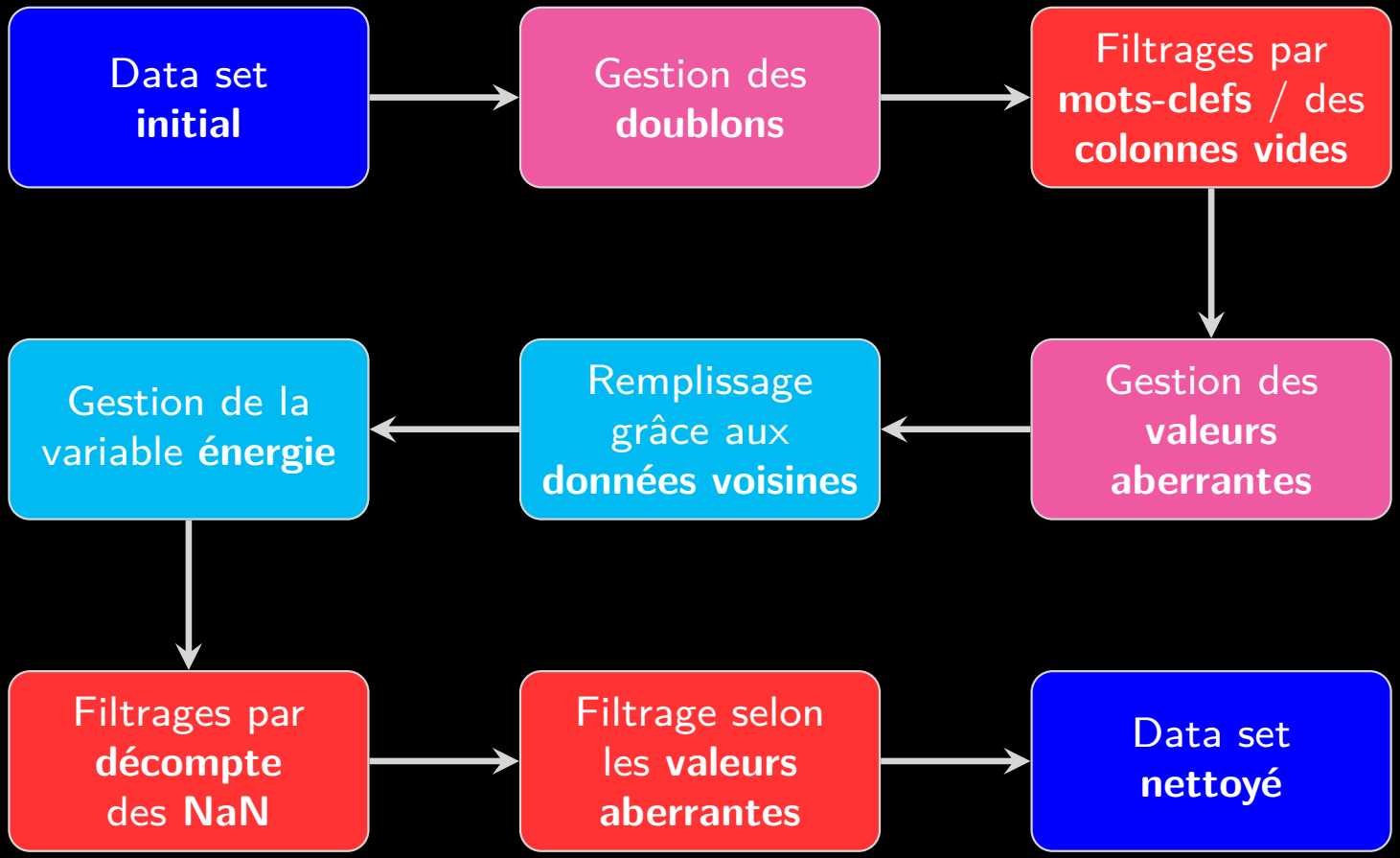

Data set
initial
Gestion des
doublons
Filtrages par
mots-clefs / des
colonnes vides
Gestion des
valeurs
aberrantes

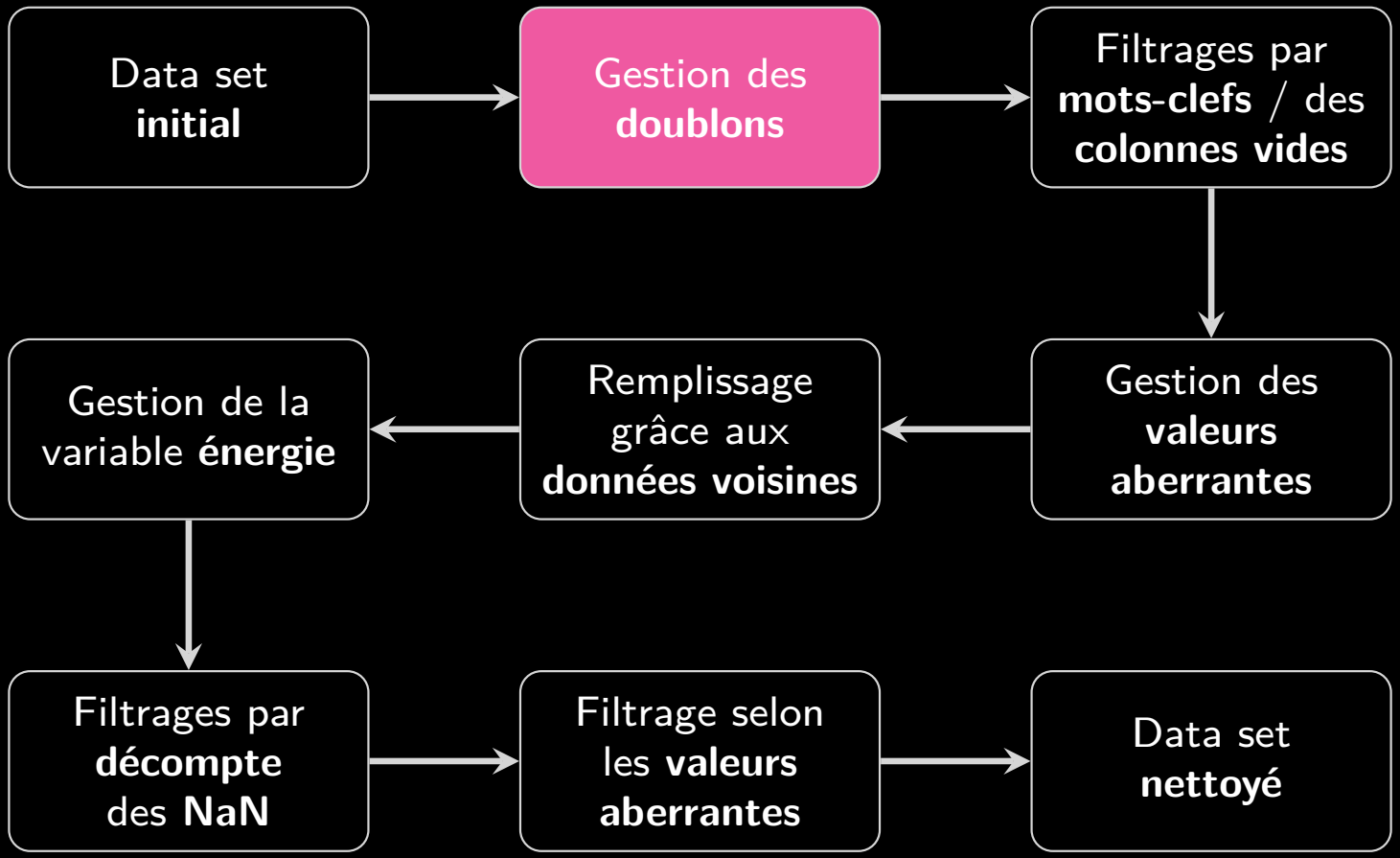

Data set
initial
Gestion des
doublons
Filtrages par
mots-clefs / des
colonnes vides
Gestion des
valeurs
aberrantes
Remplissage
grâce aux
données voisines


Data set
initial
Gestion des
doublons
Filtrages par
mots-clefs / des
colonnes vides
Gestion des
valeurs
aberrantes
Remplissage
grâce aux
données voisines
Gestion de la
variable **énergie**


Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**


Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**


Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**
Data set **nettoyé**


Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**
Data set **nettoyé**

Détection et tri ascendant par valeurs:

(création d'un DataFrame temporaire)

Détection et tri ascendant par valeurs:

(création d'un DataFrame temporaire)

index	code	last modified datetime	n filled
421527	31843340000818	2021-08-17t06:35:03z	28
349035	31843340000818	2022-02-11t08:47:36z	30
61995	3560070278831	2021-04-17t07:44:17z	41
188851	3560070278831	2022-02-10t18:03:06z	47
270028	3700320230572	2021-08-24t12:58:09z	16
749882	3700320230572	2021-08-24t12:58:58z	33
480000	7071688002962	2021-07-13t14:26:35z	40
477267	7071688002962	2021-07-13t14:26:35z	45

Détection et tri ascendant par valeurs:

(création d'un DataFrame temporaire)

même code

index	code	last modified datetime	n filled
421527	31843340000818	2021-08-17t06:35:03z	28
349035	31843340000818	2022-02-11t08:47:36z	30
61995	3560070278831	2021-04-17t07:44:17z	41
188851	3560070278831	2022-02-10t18:03:06z	47
270028	3700320230572	2021-08-24t12:58:09z	16
749882	3700320230572	2021-08-24t12:58:58z	33
480000	7071688002962	2021-07-13t14:26:35z	40
477267	7071688002962	2021-07-13t14:26:35z	45

Détection et tri ascendant par valeurs:

(création d'un DataFrame temporaire)

\+ ancien / + récent

index	code	last modified datetime	n filled
421527	31843340000818	2021-08-17t06:35:03z	28
349035	31843340000818	2022-02-11t08:47:36z	30
61995	3560070278831	2021-04-17t07:44:17z	41
188851	3560070278831	2022-02-10t18:03:06z	47
270028	3700320230572	2021-08-24t12:58:09z	16
749882	3700320230572	2021-08-24t12:58:58z	33
480000	7071688002962	2021-07-13t14:26:35z	40
477267	7071688002962	2021-07-13t14:26:35z	45

Détection et tri ascendant par valeurs:

(création d'un DataFrame temporaire)

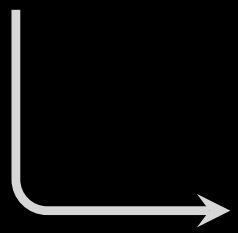
- rempli / + rempli

index	code	last modified datetime	n filled
421527	31843340000818	2021-08-17t06:35:03z	28
349035	31843340000818	2022-02-11t08:47:36z	30
61995	3560070278831	2021-04-17t07:44:17z	41
188851	3560070278831	2022-02-10t18:03:06z	47
270028	3700320230572	2021-08-24t12:58:09z	16
749882	3700320230572	2021-08-24t12:58:58z	33
480000	7071688002962	2021-07-13t14:26:35z	40
477267	7071688002962	2021-07-13t14:26:35z	45

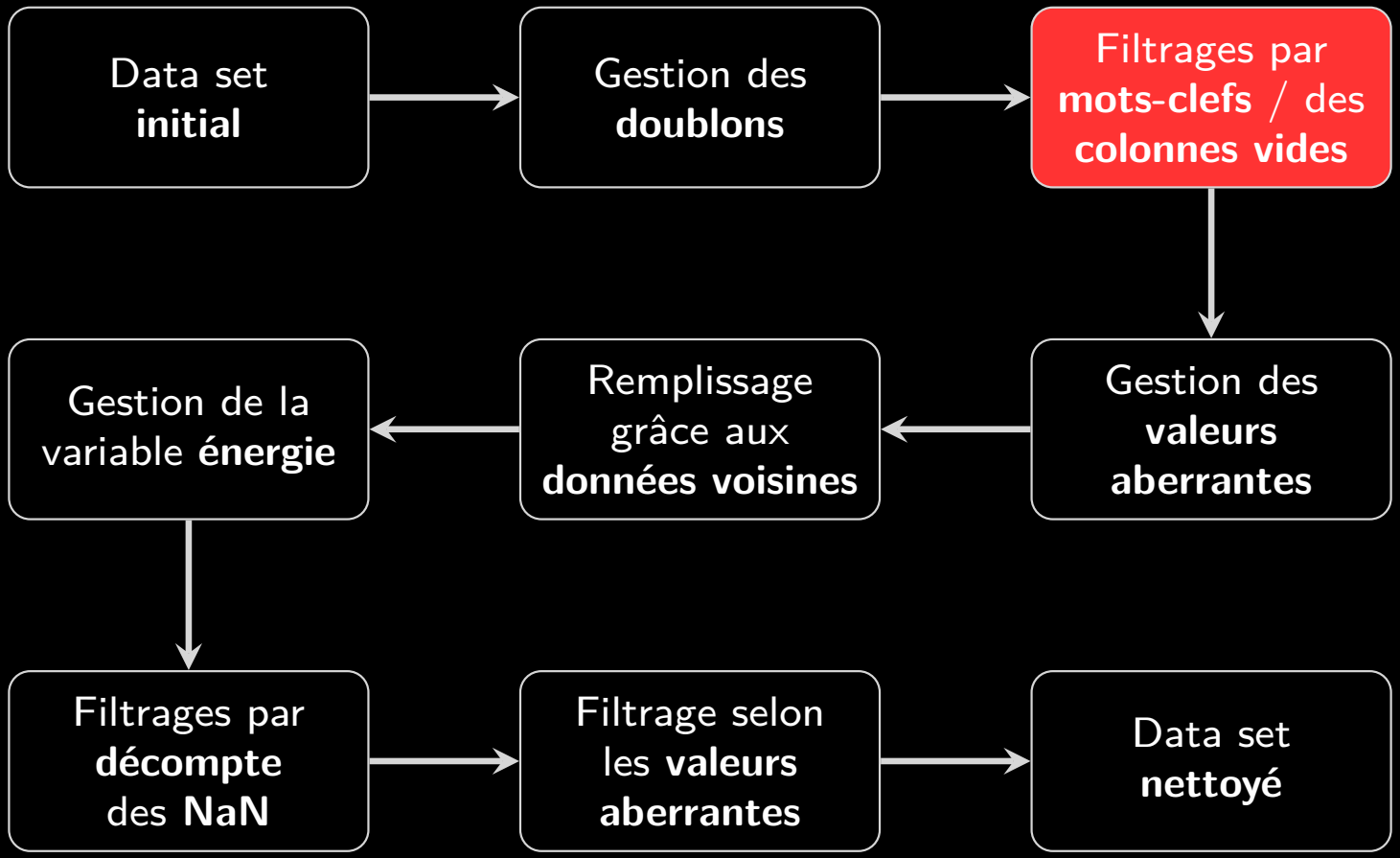
Détection et tri ascendant par valeurs:

(création d'un DataFrame temporaire)

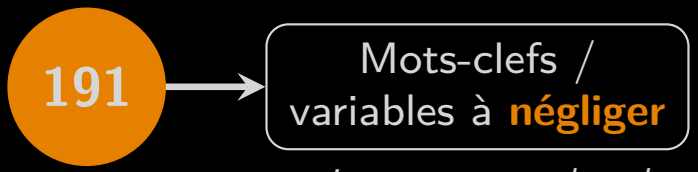
index	code	last modified datetime	n filled
421527	31843340000818	2021-08-17t06:35:03z	28
349035	31843340000818	2022-02-11t08:47:36z	30
61995	3560070278831	2021-04-17t07:44:17z	41
188851	3560070278831	2022-02-10t18:03:06z	47
270028	3700320230572	2021-08-24t12:58:09z	16
749882	3700320230572	2021-08-24t12:58:58z	33
480000	7071688002962	2021-07-13t14:26:35z	40
477267	7071688002962	2021-07-13t14:26:35z	45



4 entrées supprimées


Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**
Data set **nettoyé**

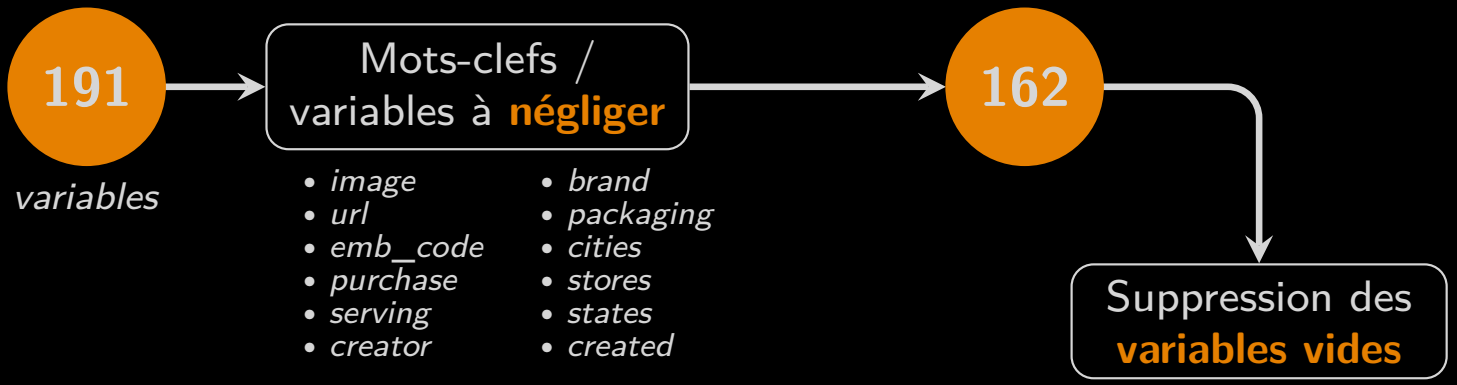

191
variables

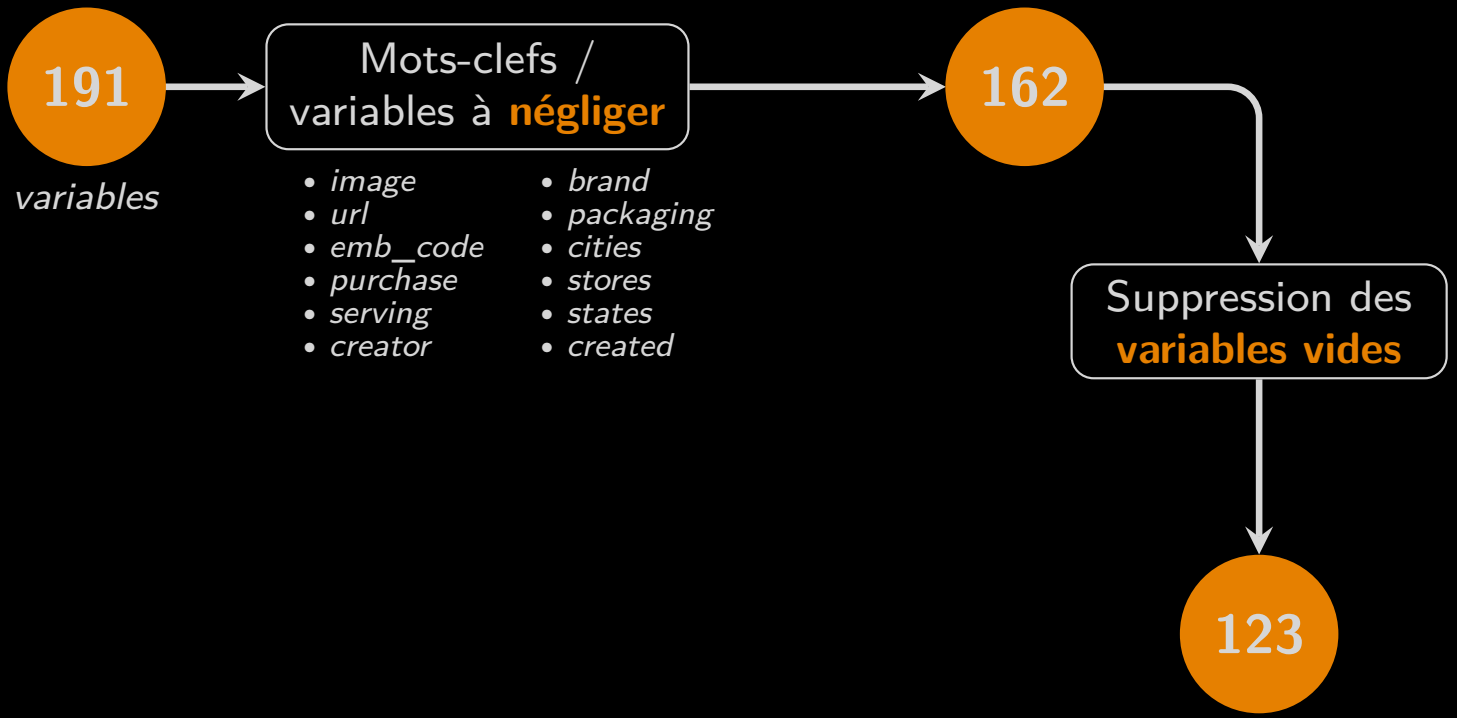


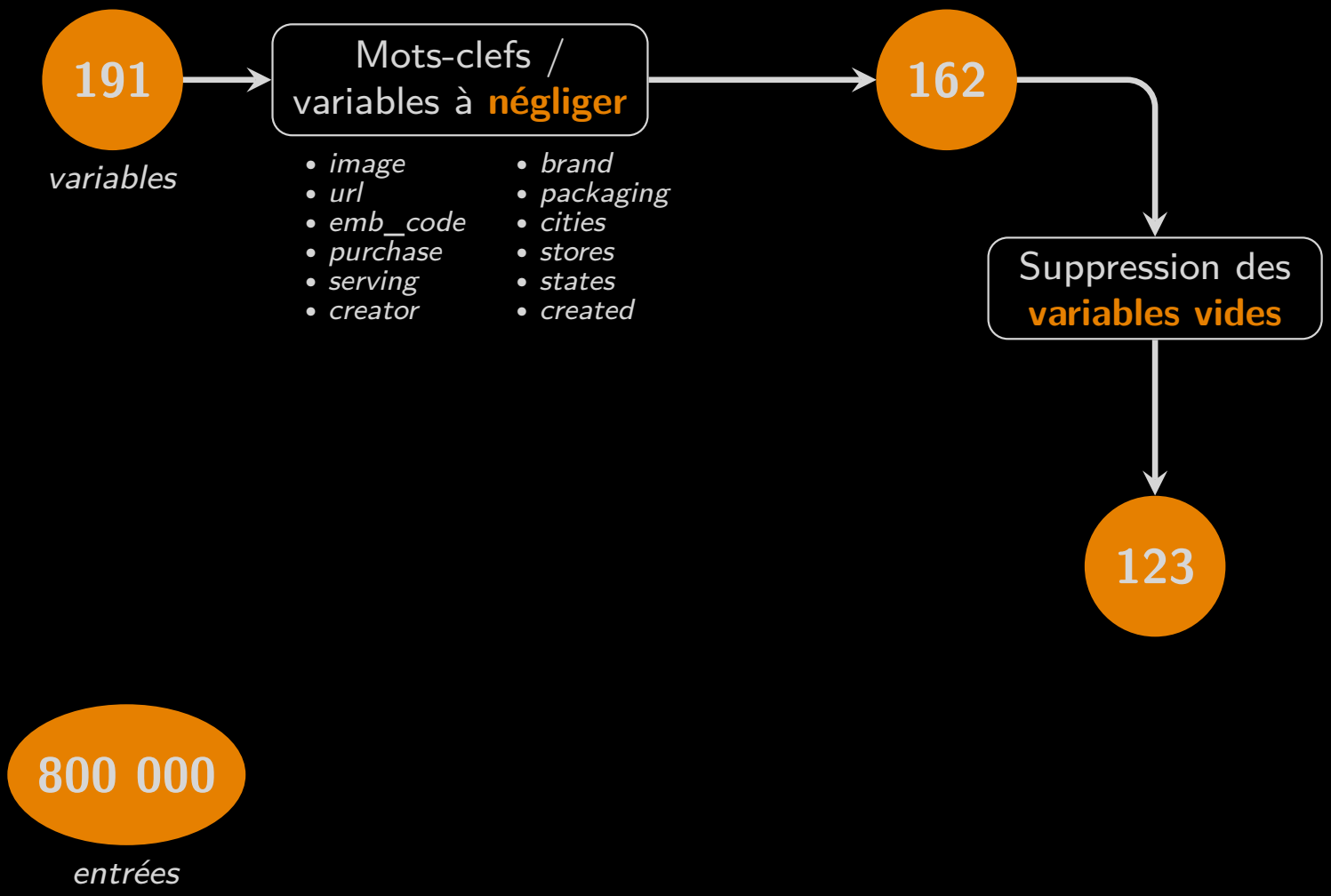
- *image*
- *url*
- *emb_code*
- *purchase*
- *serving*
- *creator*
- *brand*
- *packaging*
- *cities*
- *stores*
- *states*
- *created*

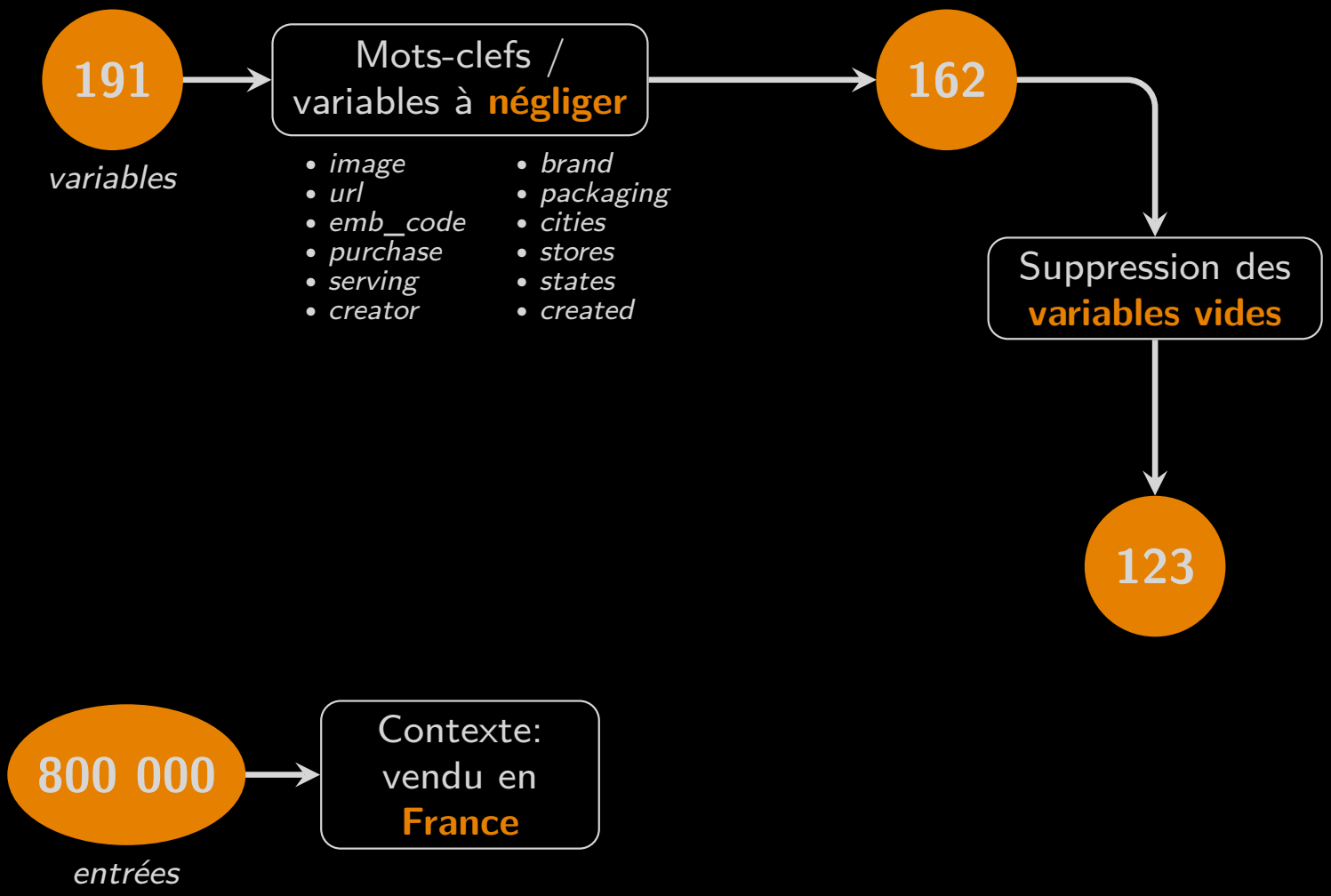
variables

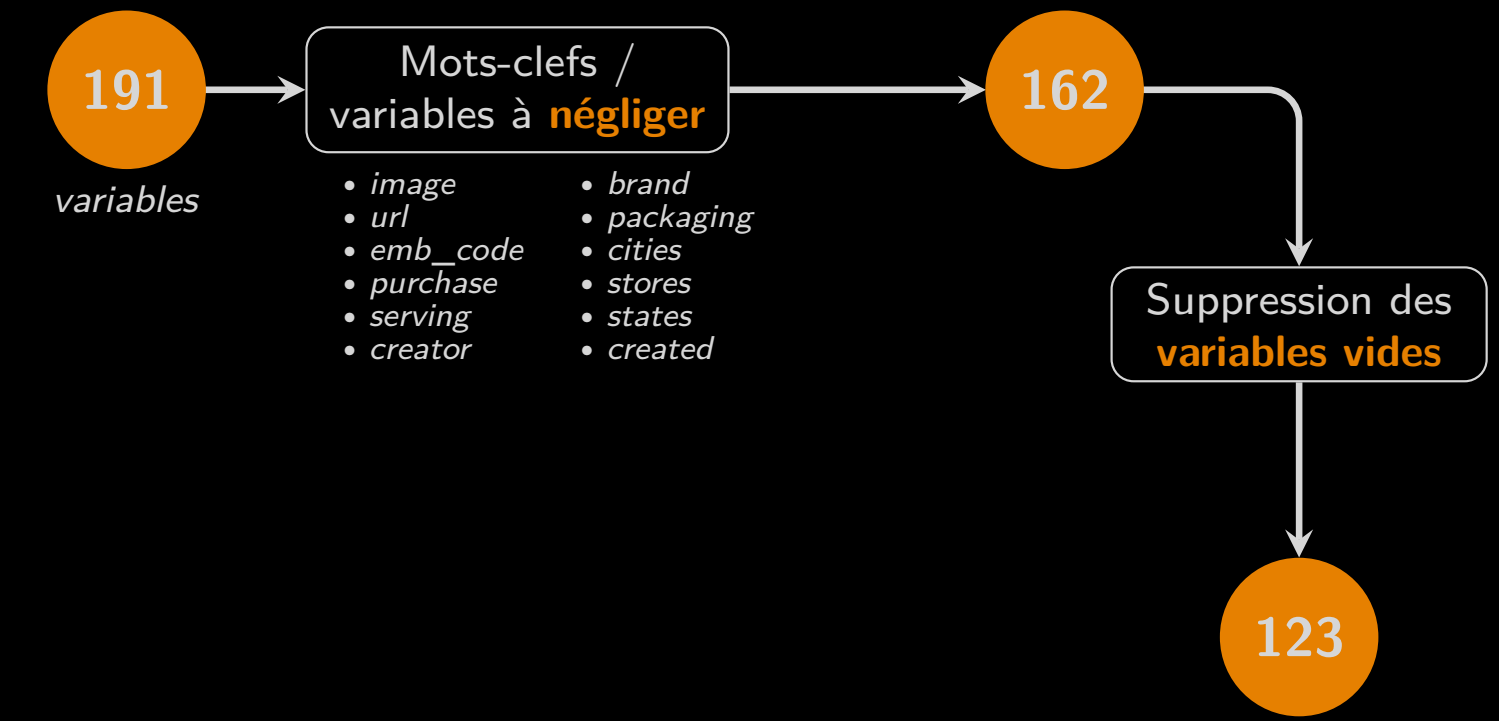
- *image*
- *url*
- *emb_code*
- *purchase*
- *serving*
- *creator*
- *brand*
- *packaging*
- *cities*
- *stores*
- *states*
- *created*

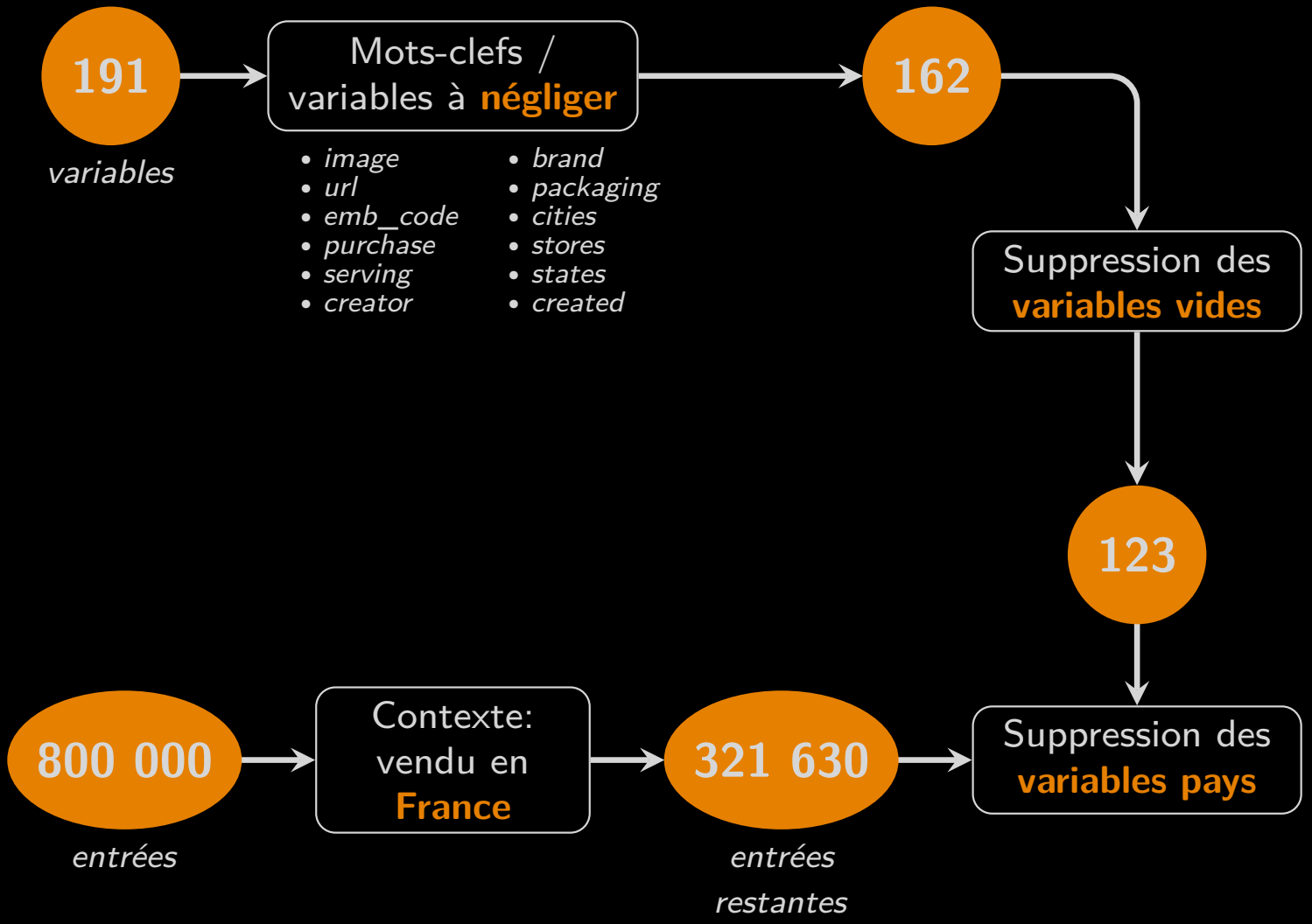

191
variables
Mots-clefs / variables à négliger
image
url
emb_code
purchase
serving
creator
brand
packaging
cities
stores
states
created
162
Suppression des variables vides

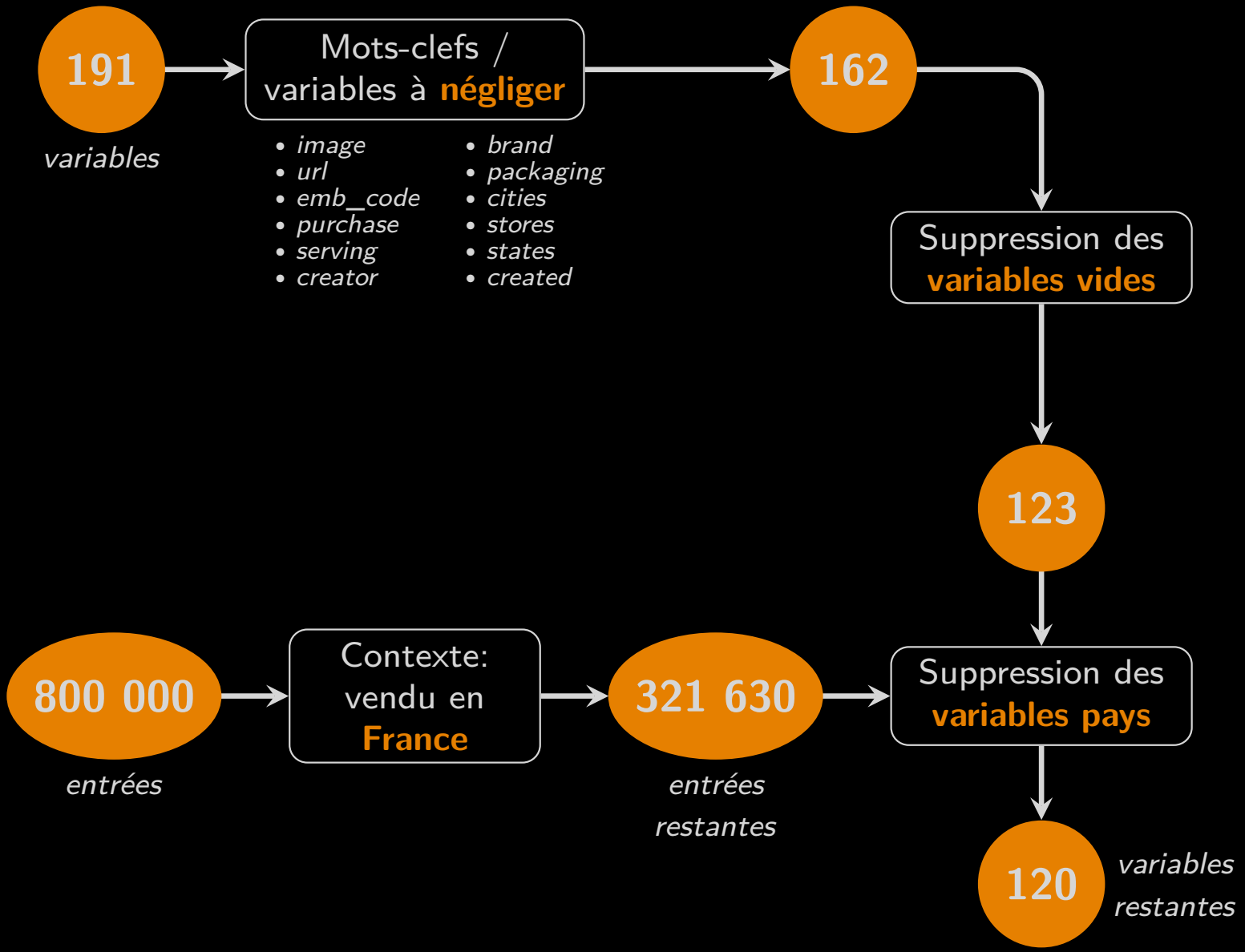

191
variables
Mots-clefs / variables à négliger
image
url
emb_code
purchase
serving
creator
brand
packaging
cities
stores
states
created
162
Suppression des variables vides
123

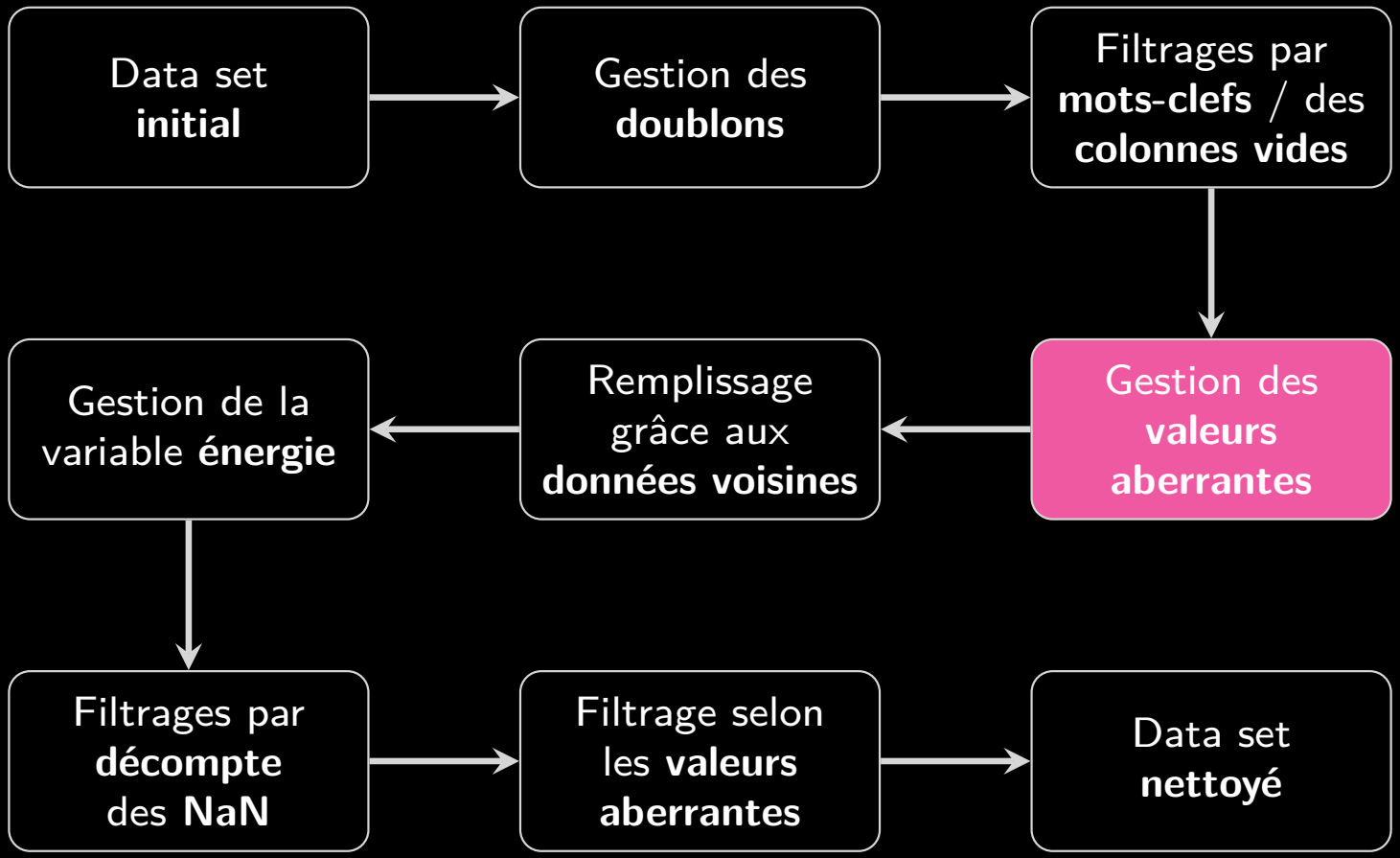

191
variables
Mots-clefs / variables à négliger
image
url
emb_code
purchase
serving
creator
brand
packaging
cities
stores
states
created
162
Suppression des variables vides
123
800 000
entrées


191
variables
Mots-clefs / variables à négliger
image
url
emb_code
purchase
serving
creator
brand
packaging
cities
stores
states
created
162
Suppression des variables vides
123
800 000
entrées
Contexte: vendu en France

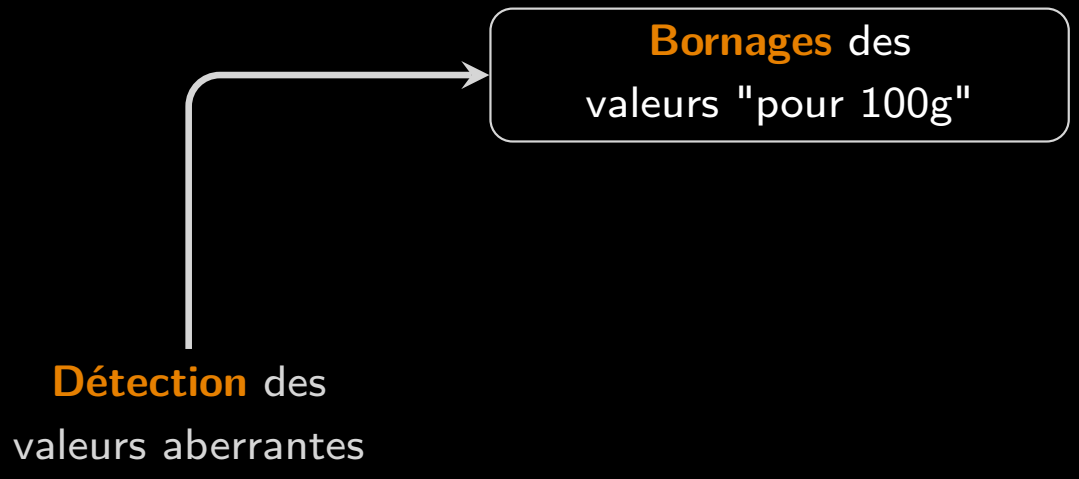

191
variables
Mots-clefs / variables à négliger
image
url
emb_code
purchase
serving
creator
brand
packaging
cities
stores
states
created
162
Suppression des variables vides
123
800 000
entrées
Contexte: vendu en France
321 630
entrées restantes

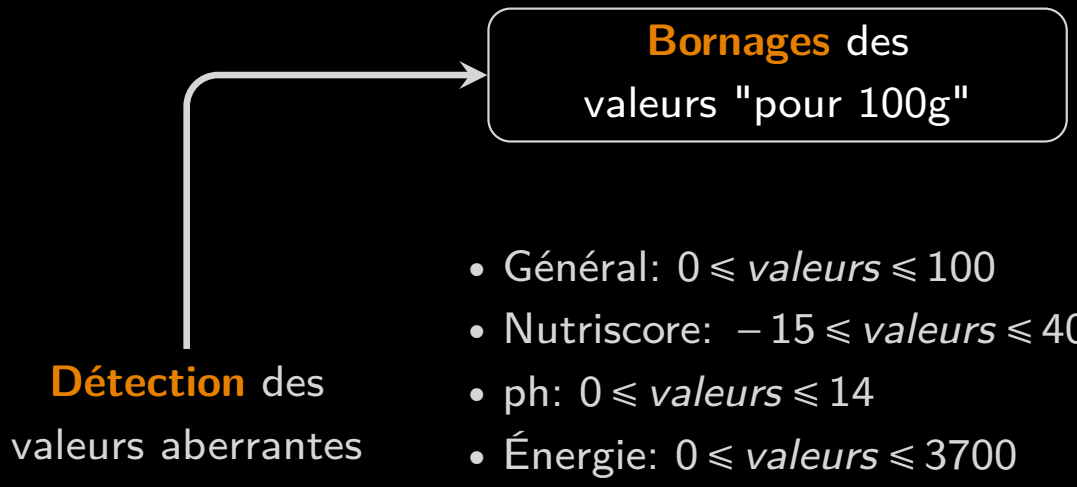

191
variables
Mots-clefs / variables à négliger
image
url
emb_code
purchase
serving
creator
brand
packaging
cities
stores
states
created
162
Suppression des variables vides
123
Suppression des variables pays
321 630
entrées restantes
Contexte: vendu en France
800 000
entrées

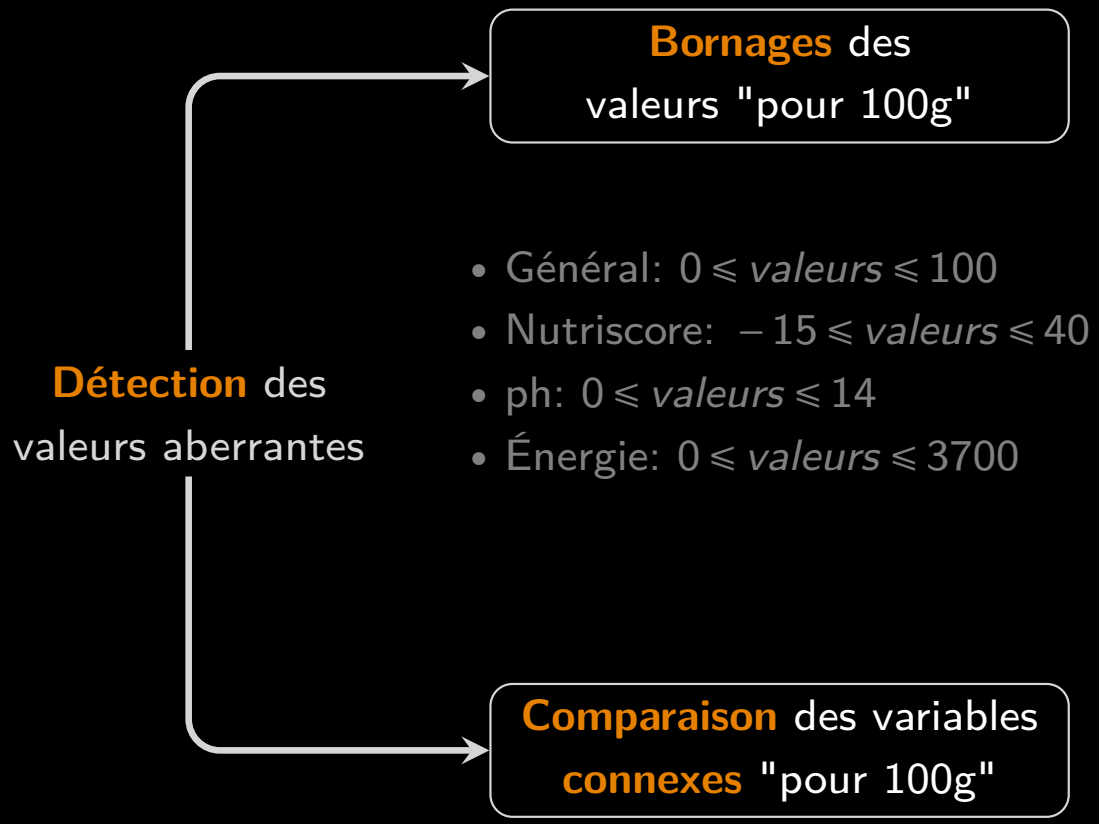

191
variables
Mots-clefs / variables à négliger
image
url
emb_code
purchase
serving
creator
brand
packaging
cities
stores
states
created
162
Suppression des variables vides
123
800 000
entrées
Contexte: vendu en France
321 630
entrées restantes
Suppression des variables pays
120
variables restantes

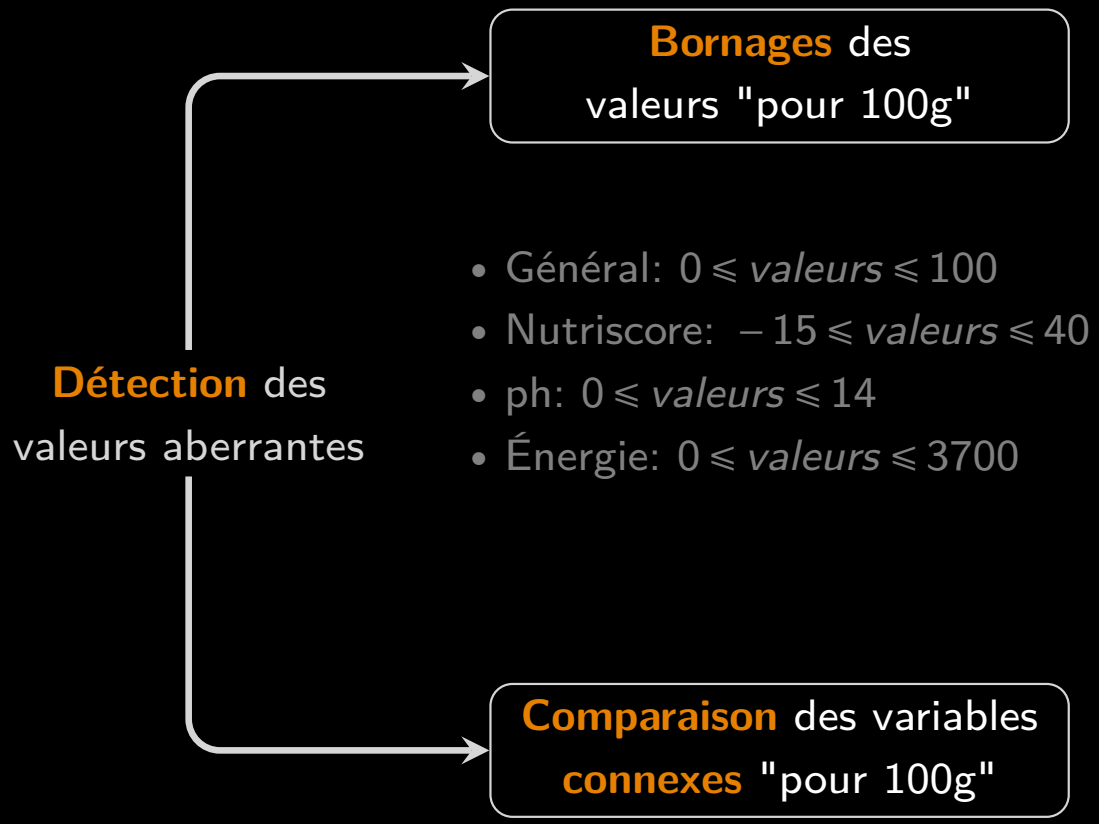

Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**
Data set **nettoyé**

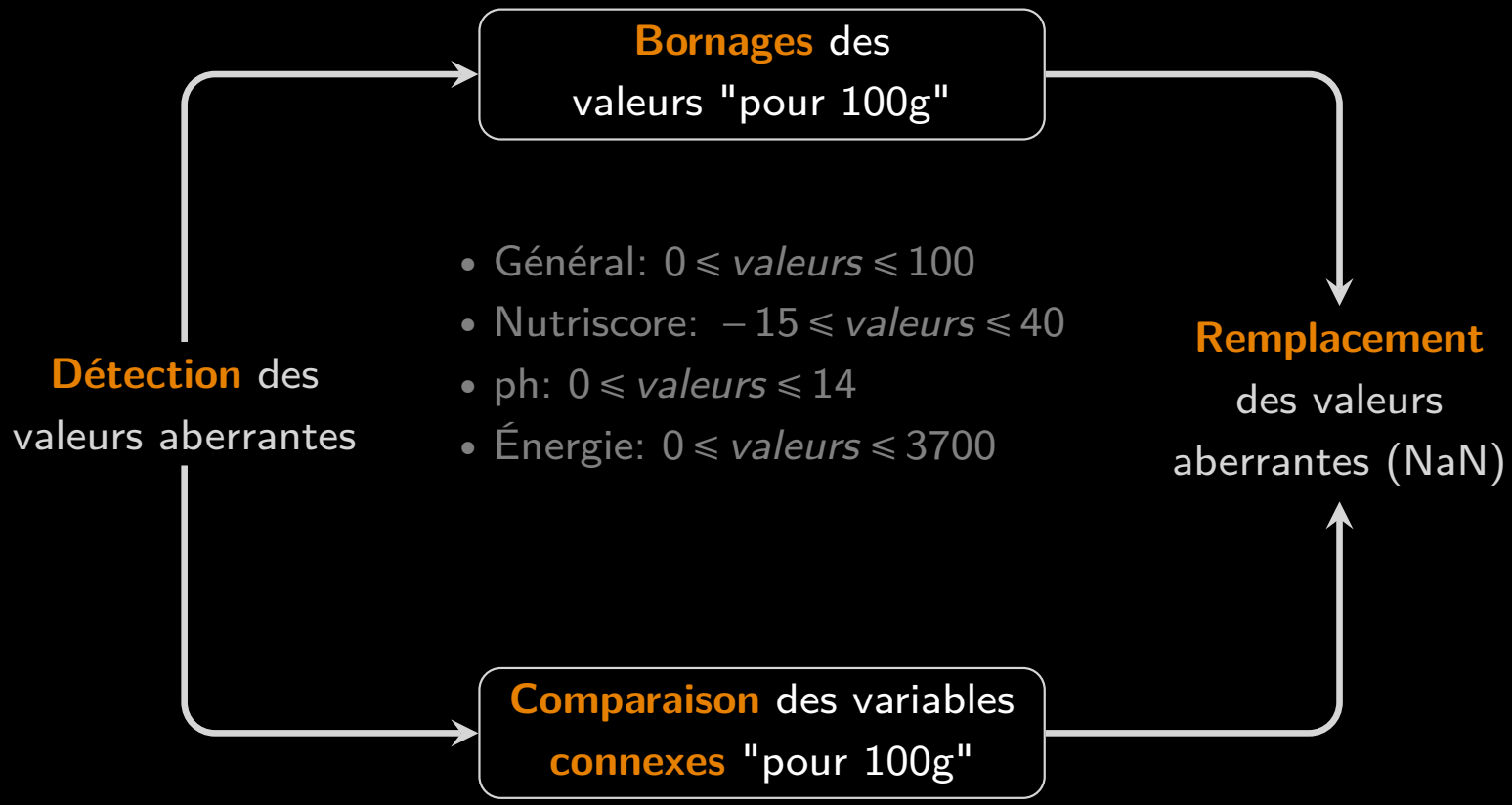
Détection des
valeurs aberrantes

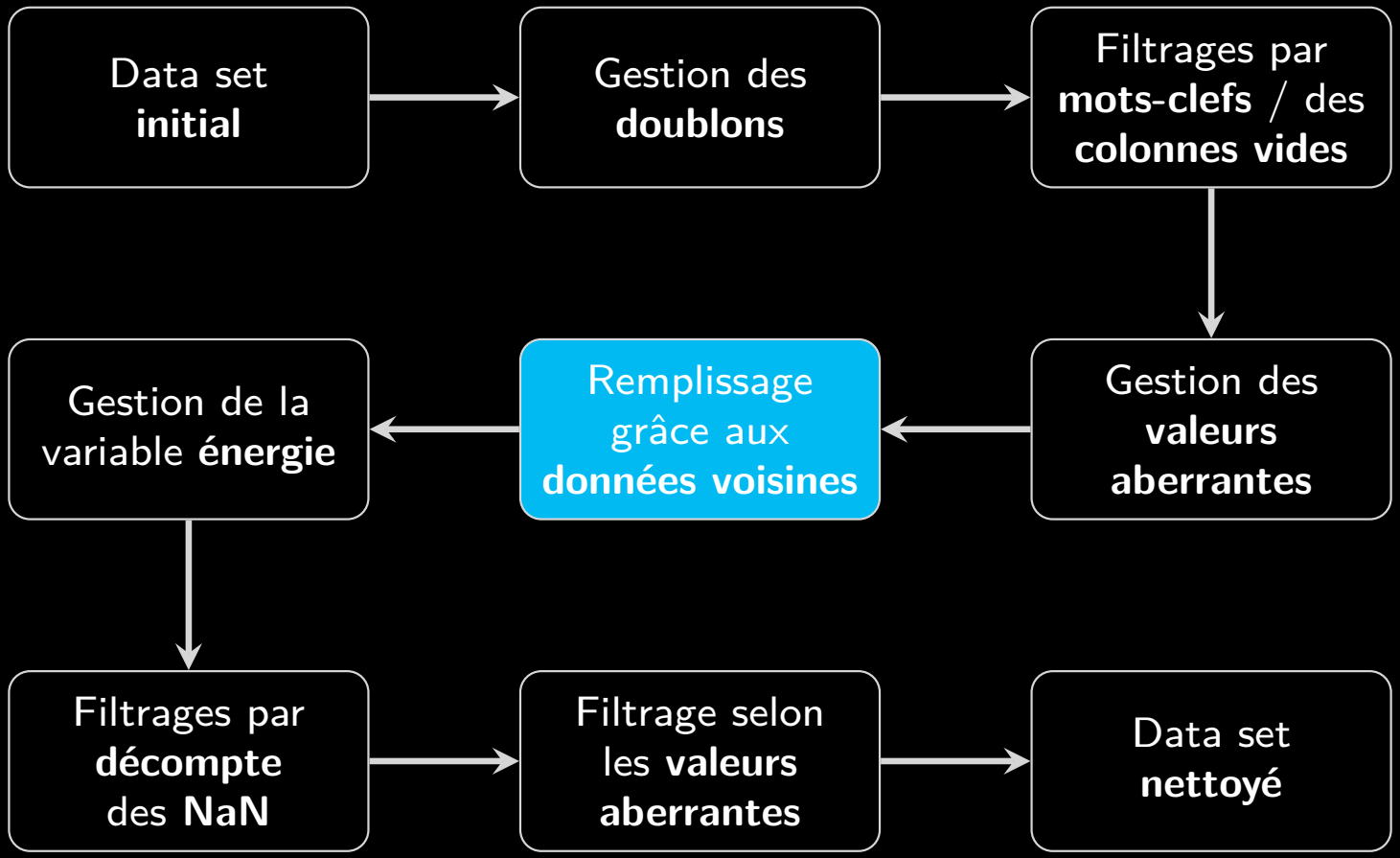

Bornages des
valeurs "pour 100g"
Détection des
valeurs aberrantes


Bornages des
valeurs "pour 100g"
Détection des
valeurs aberrantes
- Général: $0 \leqslant \textit{valeurs} \leqslant 100$
- Nutriscore: $-15 \leqslant \textit{valeurs} \leqslant 40$
- ph: $0 \leqslant \textit{valeurs} \leqslant 14$
- Énergie: $0 \leqslant \textit{valeurs} \leqslant 3700$


Bornages des
valeurs "pour 100g"
- Général: $0 \leqslant valeurs \leqslant 100$
- Nutriscore: $-15 \leqslant valeurs \leqslant 40$
- ph: $0 \leqslant valeurs \leqslant 14$
- Énergie: $0 \leqslant valeurs \leqslant 3700$
Détection des
valeurs aberrantes
Comparaison des variables
connexes "pour 100g"


Détection des
valeurs aberrantes
Bornages des
valeurs "pour 100g"
• Général: 0 ⩽ valeurs ⩽ 100
• Nutriscore: −15 ⩽ valeurs ⩽ 40
• ph: 0 ⩽ valeurs ⩽ 14
• Énergie: 0 ⩽ valeurs ⩽ 3700
Comparaison des variables
connexes "pour 100g"
• carbohydrates ⩾ sugars
• salt ⩾ sodium
• fat > other fats


Bornages des valeurs "pour 100g"
- Général: $0 \leqslant valeurs \leqslant 100$
- Nutriscore: $-15 \leqslant valeurs \leqslant 40$
- ph: $0 \leqslant valeurs \leqslant 14$
- Énergie: $0 \leqslant valeurs \leqslant 3700$
Détection des valeurs aberrantes
Remplacement des valeurs aberrantes (NaN)
Comparaison des variables **connexes** "pour 100g"
- $carbohydrates \geqslant sugars$
- $salt \geqslant sodium$
- $fat > other\ \ fats$


Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**
Data set **nettoyé**

Détection des **NaN**

pour les variables:

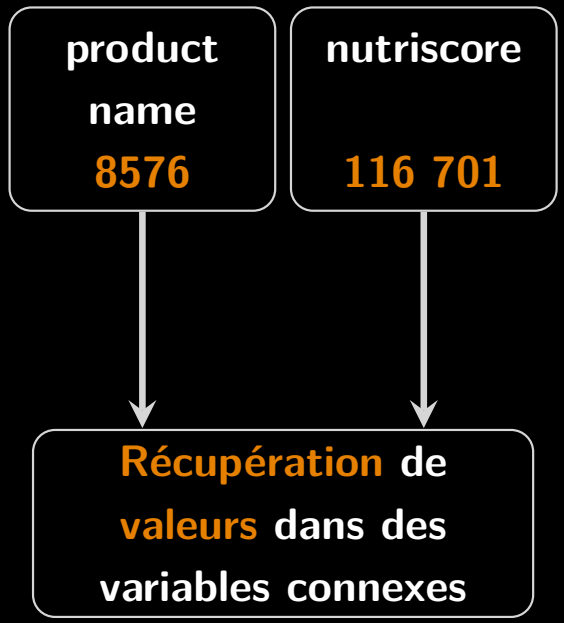
Détection des **NaN**

pour les variables:

product name	nutriscore
8576	116 701

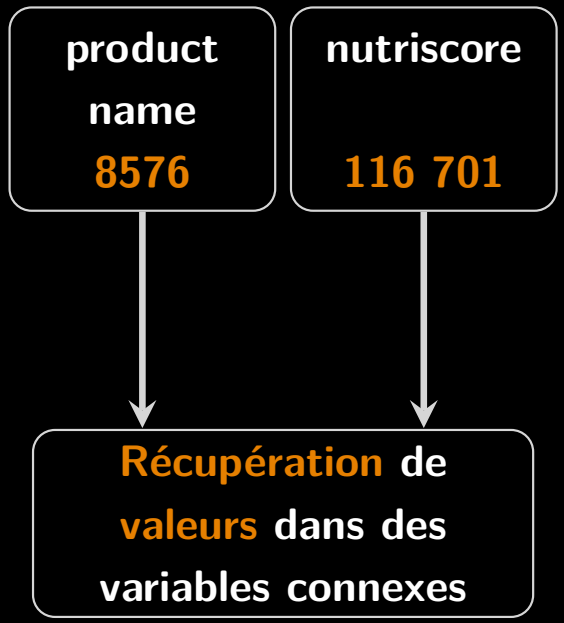
Détection des **NaN**

pour les variables:



Détection des **NaN**

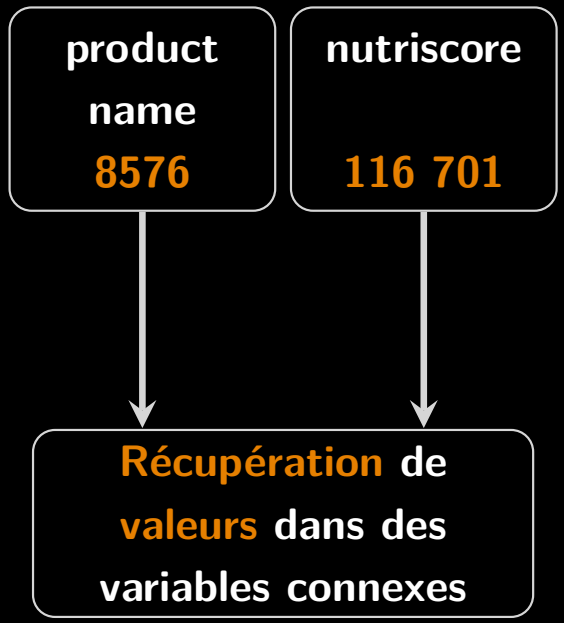
pour les variables:



~isna\isna	product name	abbreviated product name	generic name
product name	0	310337	282502
abbreviated product name	141	0	344
generic name	57	28095	0

Détection des **NaN**

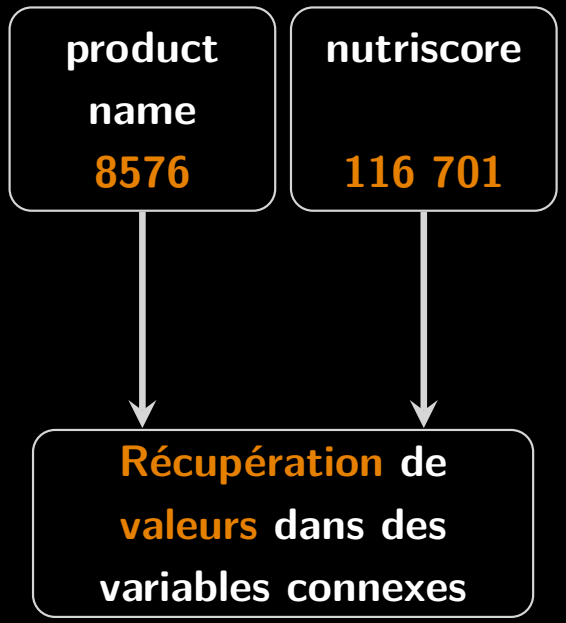
pour les variables:



~isna\isna	product name	abbreviated product name	generic name
product name	0	310337	282502
abbreviated product name	141	0	344
generic name	57	28095	0

Détection des **NaN**

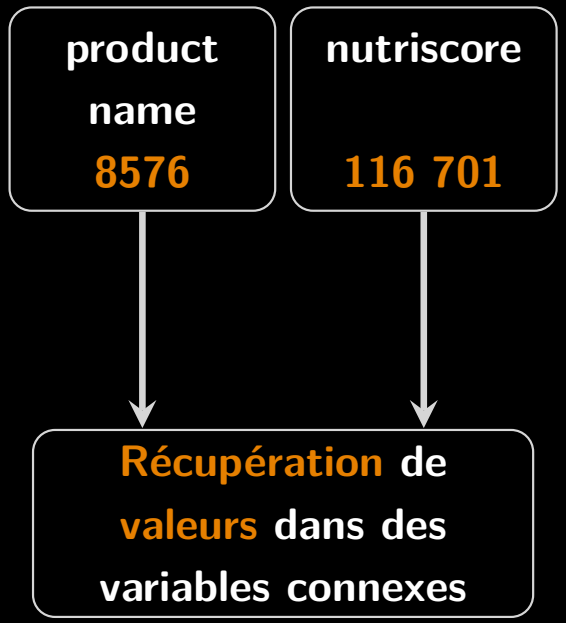
pour les variables:



~isna\isna	nutriscore score	nutriscore grade	nutrition-score-fr 100g	nutrition-score-uk 100g
nutriscore score	0	0	91	117088
nutriscore grade	0	0	91	117088
nutrition-score-fr 100g	1	1	0	116997
nutrition-score-uk 100g	1	1	0	0

Détection des **NaN**

pour les variables:



~isna\isna	nutriscore score	nutriscore grade	nutrition-score-fr 100g	nutrition-score-uk 100g
nutriscore score	0	0	91	117088
nutriscore grade	0	0	91	117088
nutrition-score-fr 100g	1	1	0	116997
nutrition-score-uk 100g	1	1	0	0

Détection des **NaN**
pour les variables:

valeurs **remplies**

~isna\isna	nutriscore score	nutriscore grade	nutrition-score-fr 100g	nutrition-score-uk 100g
nutriscore score	0	0	91	117088
nutriscore grade	0	0	91	117088
nutrition-score-fr 100g	1	1	0	116997
nutrition-score-uk 100g	1	1	0	0

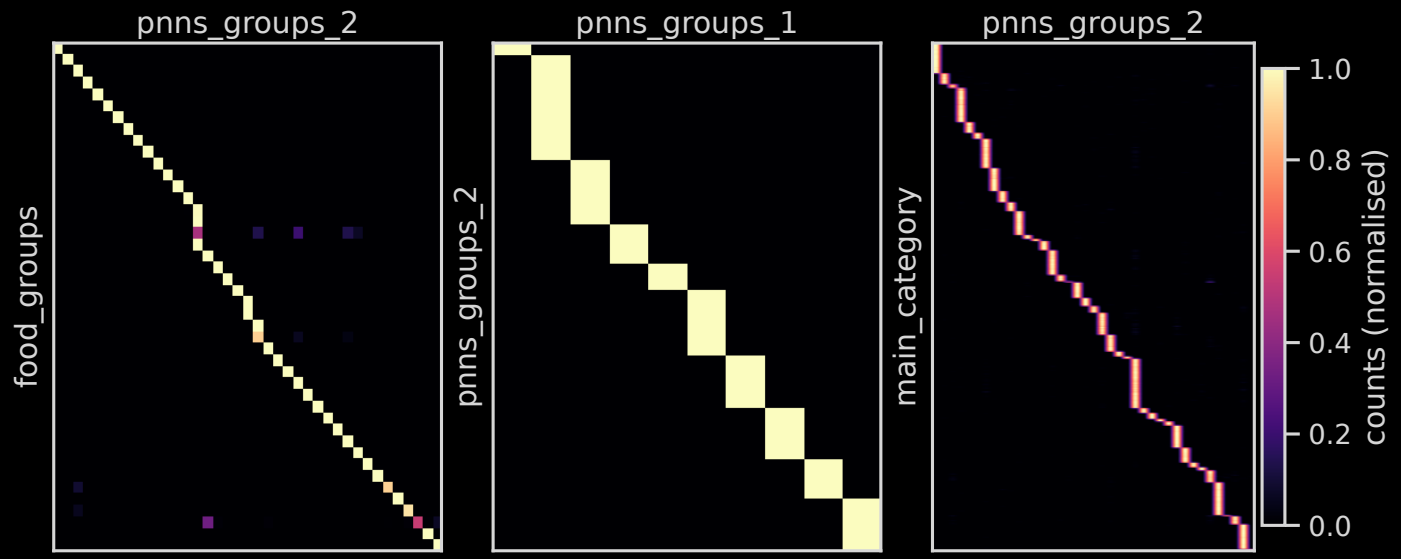
Création de **maps** à partir de **tableaux de contingences**

Création de **maps** à partir de **tableaux de contingences**

179 241 NaN pour la variable pnns groups 1

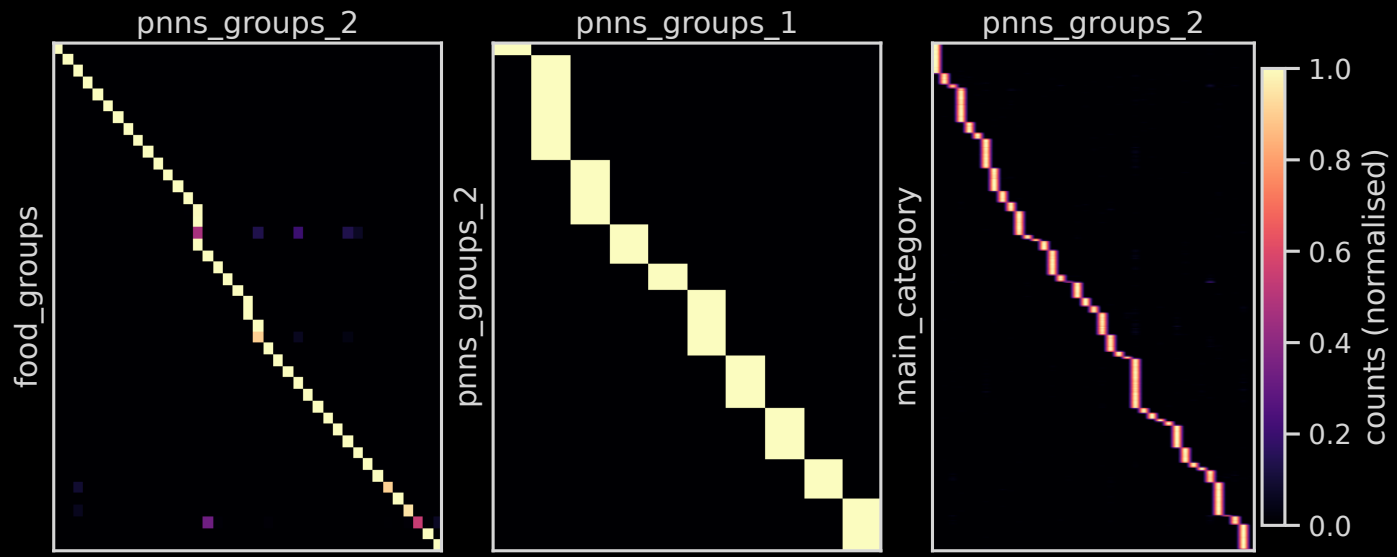
Création de **maps** à partir de **tableaux de contingences**

179 241 NaN pour la variable pnns groups 1



Création de **maps** à partir de **tableaux de contingences**

179 241 NaN pour la variable pnns groups 1

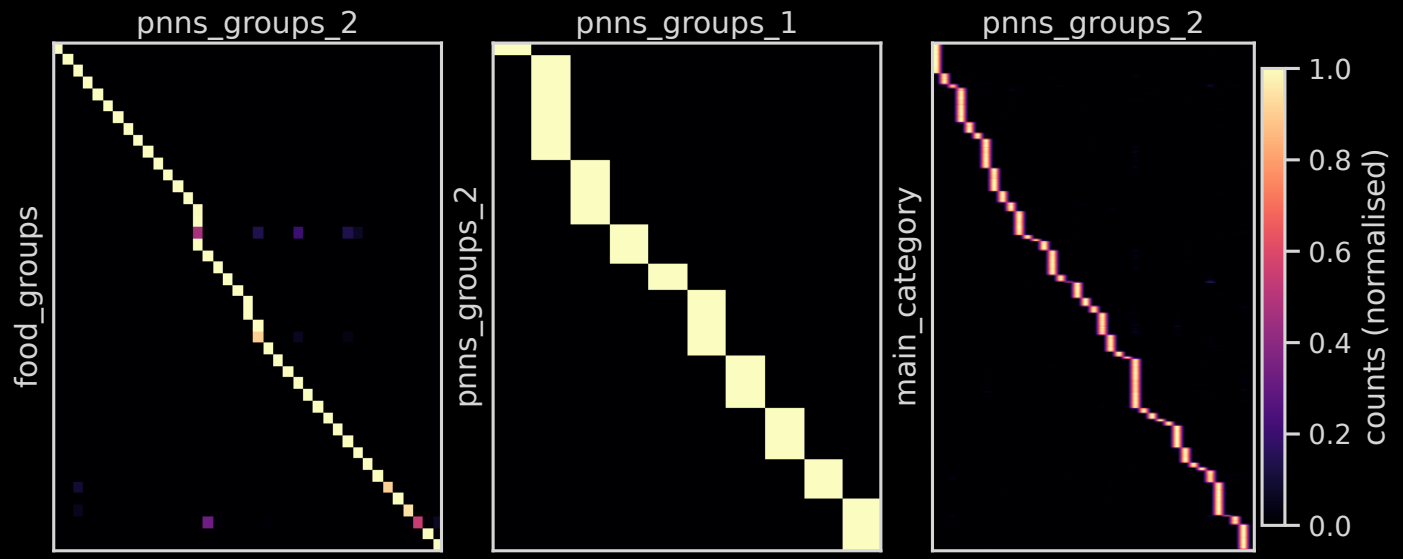


food groups → pnns groups

main category → pnns groups

Création de **maps** à partir de **tableaux de contingences**

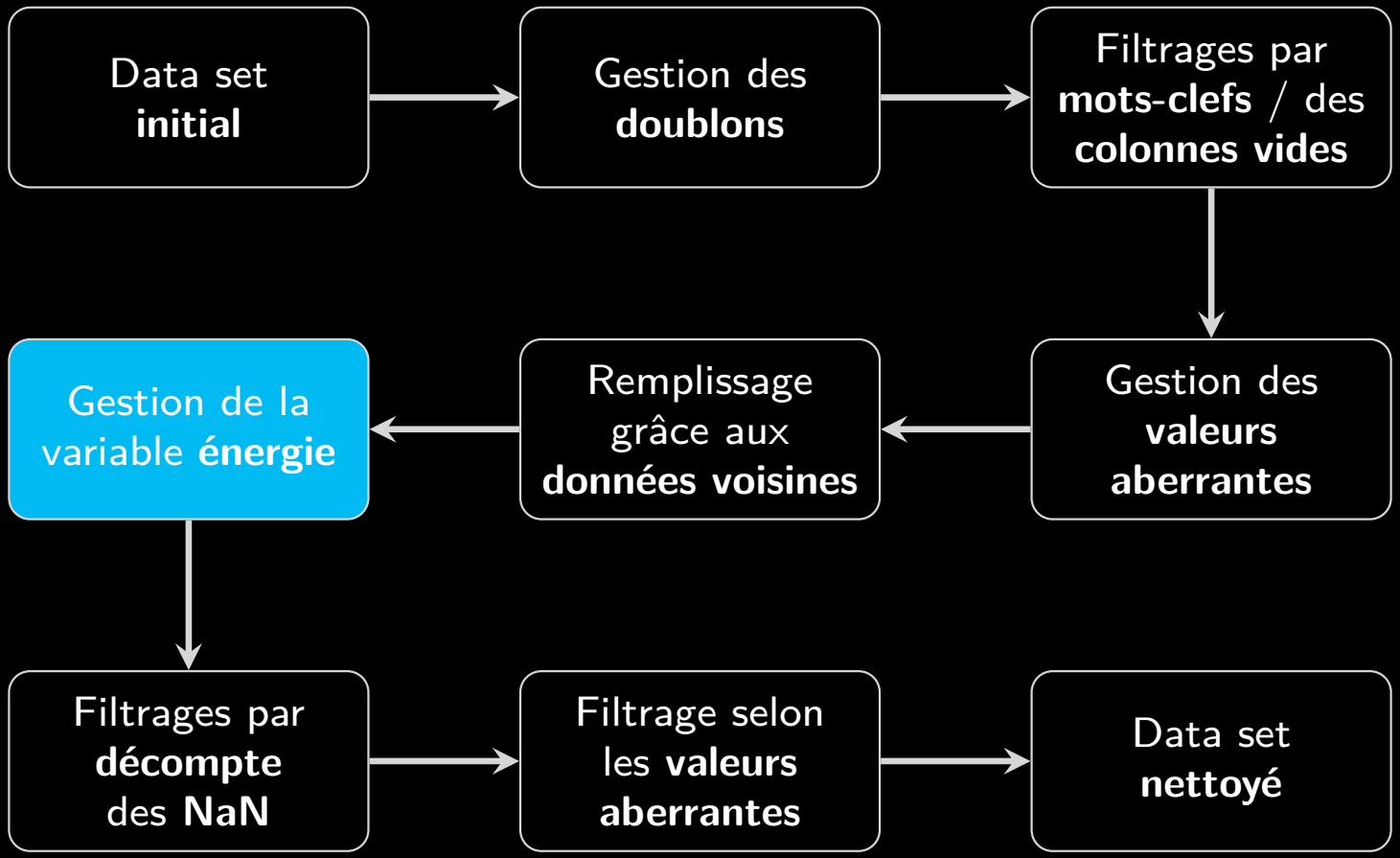
179 241 NaN pour la variable pnns groups 1



food groups → pnns groups

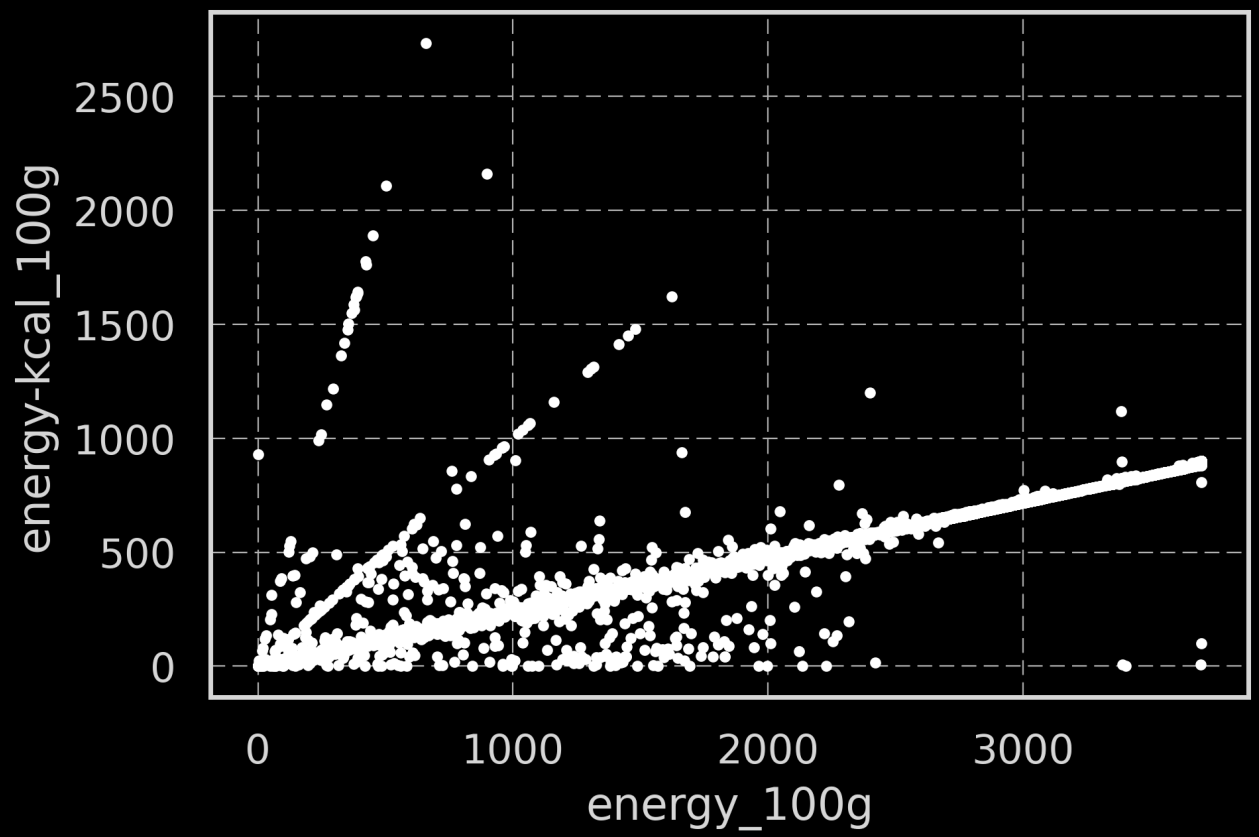
main category → pnns groups

→ Remplissage de **9 401** valeurs


Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**
Data set **nettoyé**

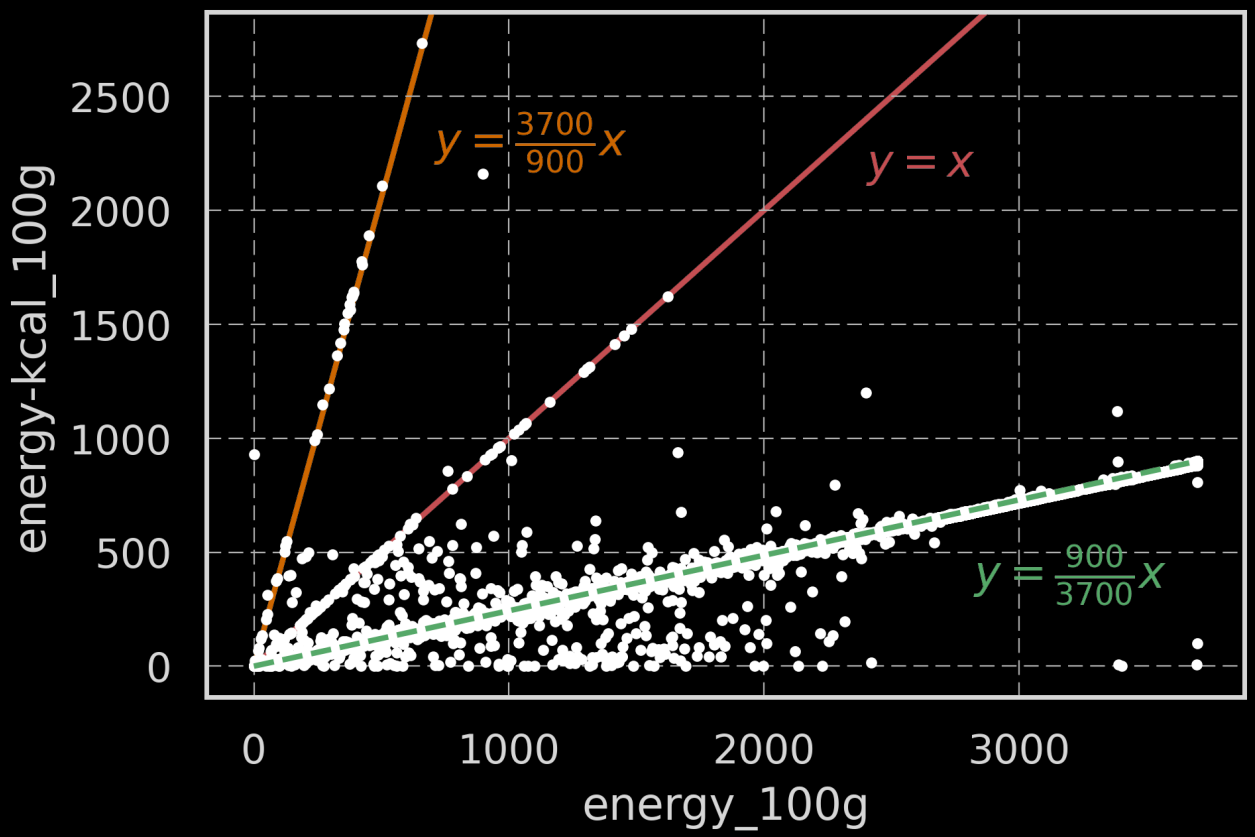
Présence de **valeurs incohérentes**:

idéalement x en kJ, y en kcal et $y \approx \frac{900}{3700} x$



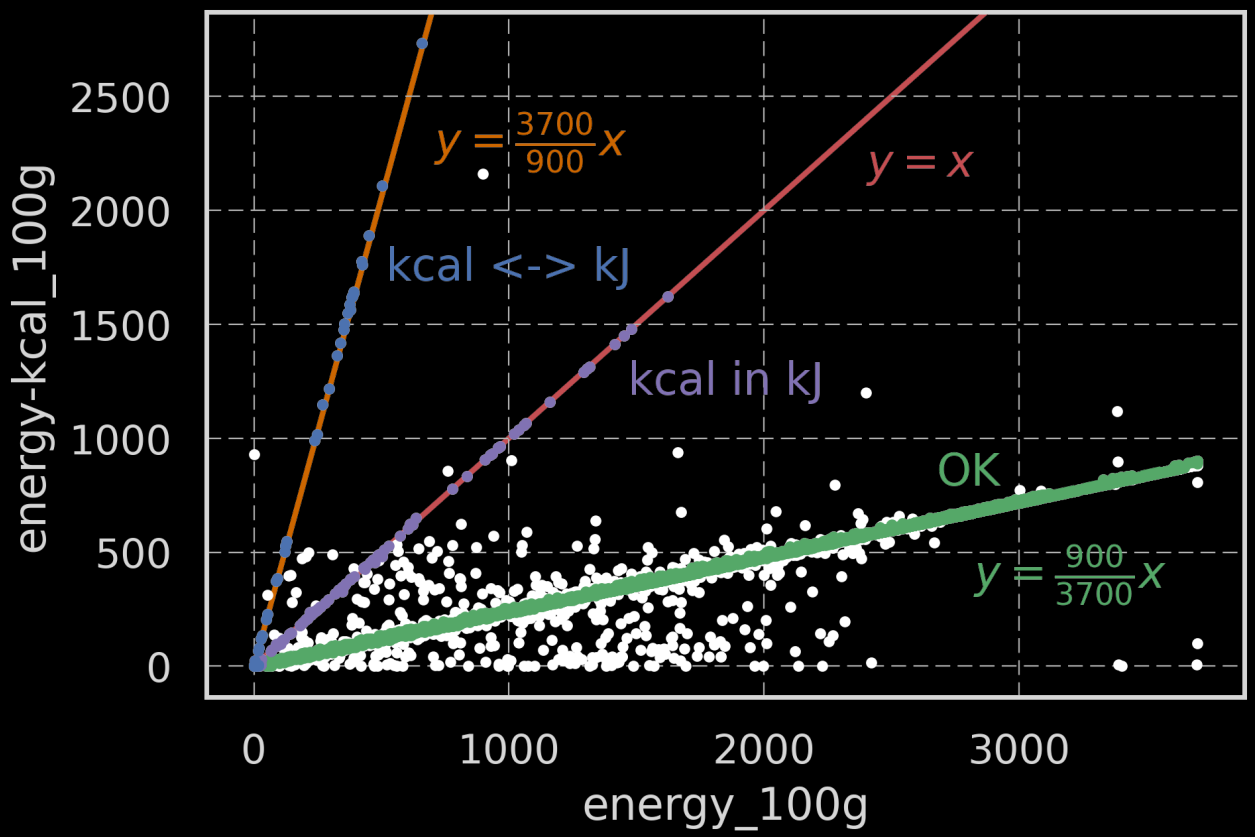
Présence de **valeurs incohérentes**:

idéalement x en kJ, y en kcal et $y \approx \frac{900}{3700} x$



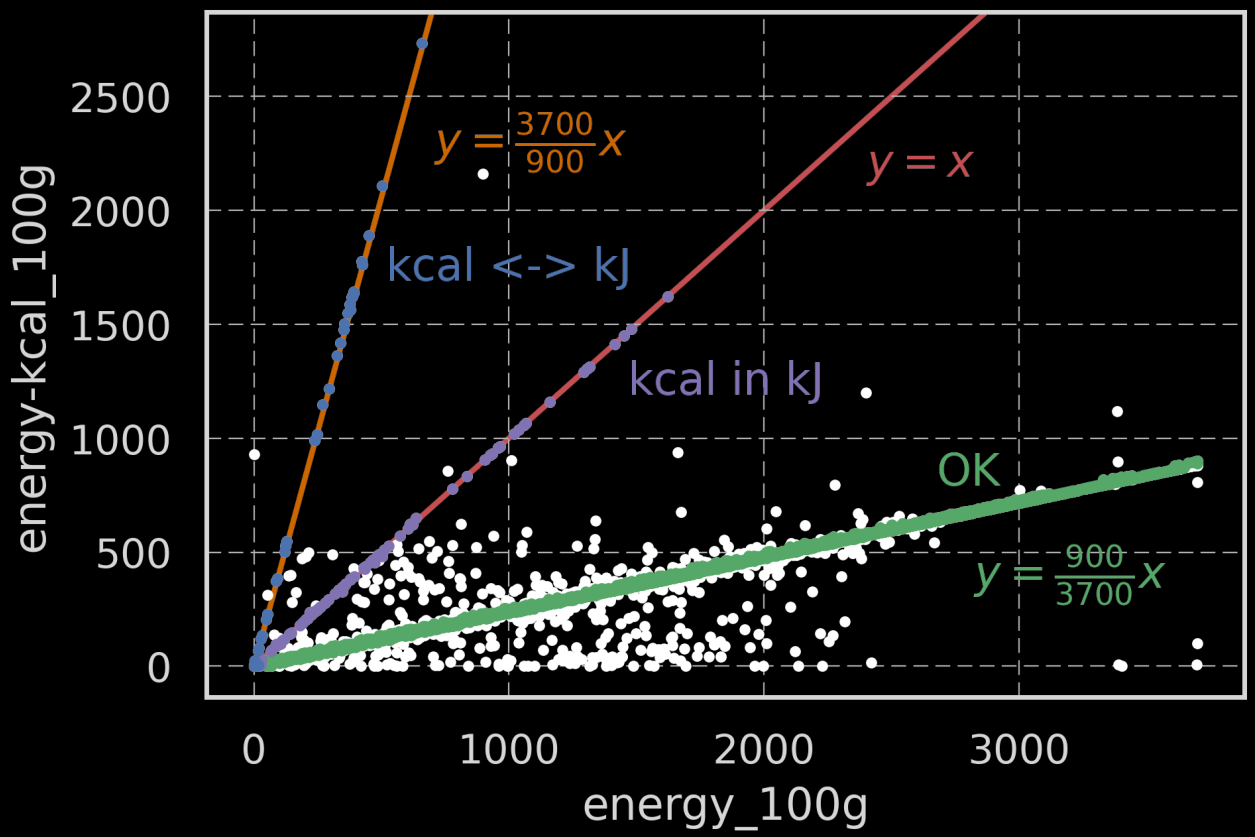
Présence de **valeurs incohérentes**:

idéalement x en kJ, y en kcal et $y \approx \frac{900}{3700} x$

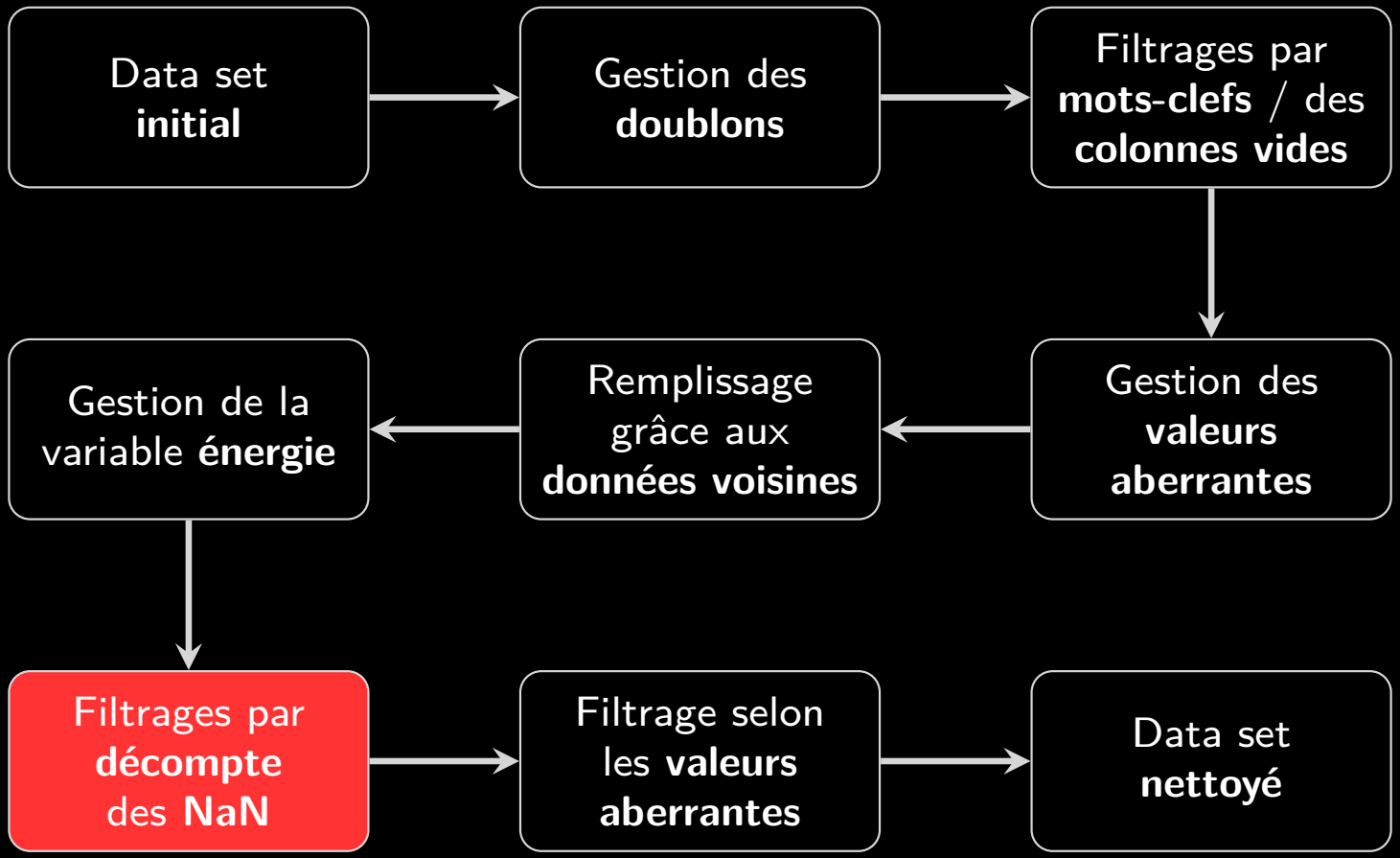


Présence de **valeurs incohérentes**:

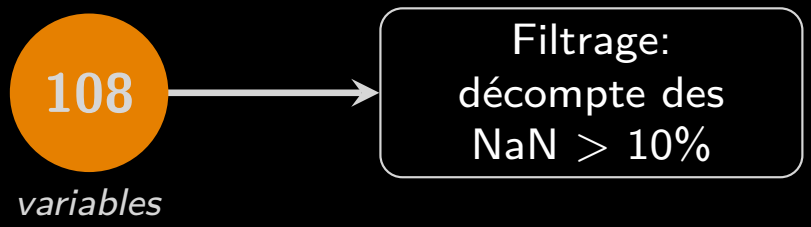
idéalement x en kJ, y en kcal et $y \approx \frac{900}{3700} x$



Solution: garder la **valeur maximale** (en kJ)

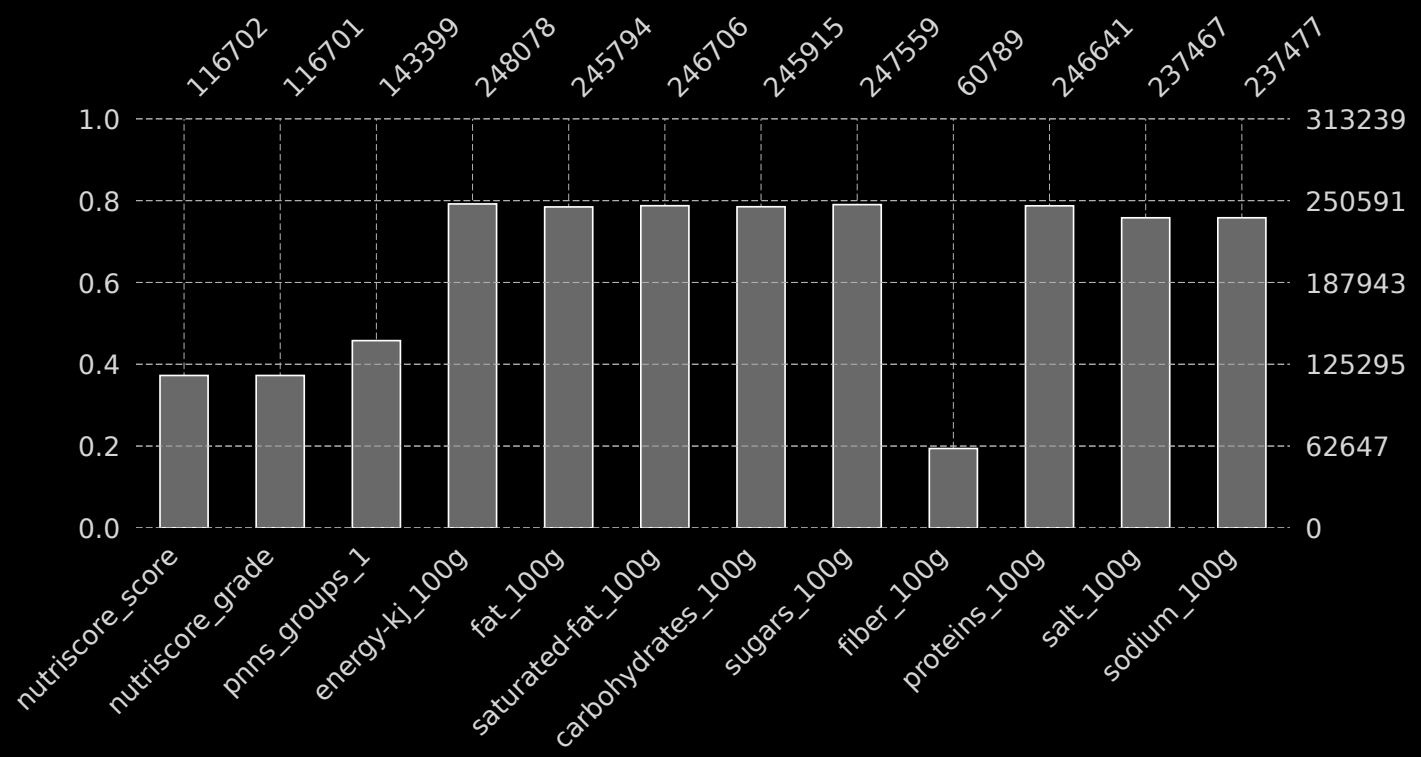

Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**
Data set **nettoyé**


108
variables

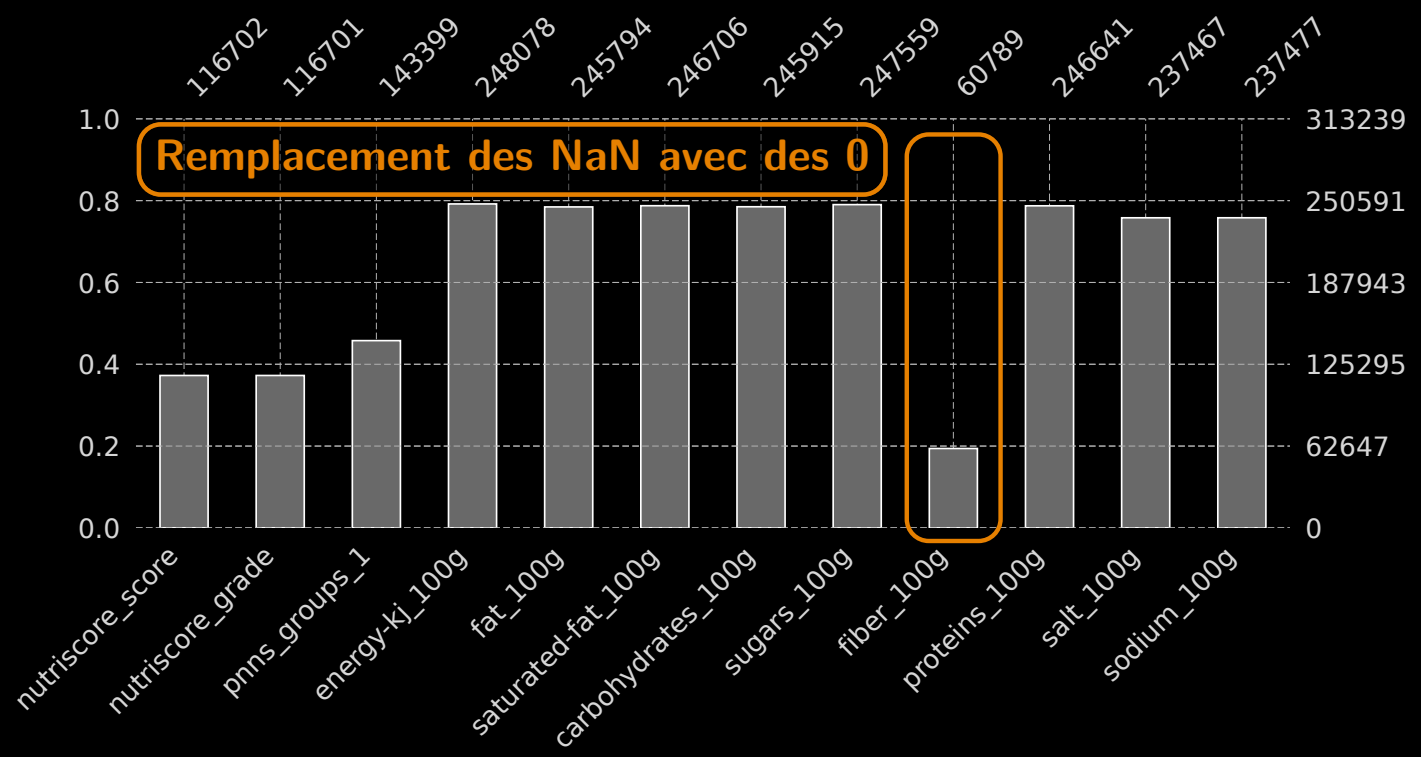

108
variables
Filtrage:
décompte des
NaN $>$ 10%


108
variables
Filtrage:
décompte des
NaN > 10%
33
variables

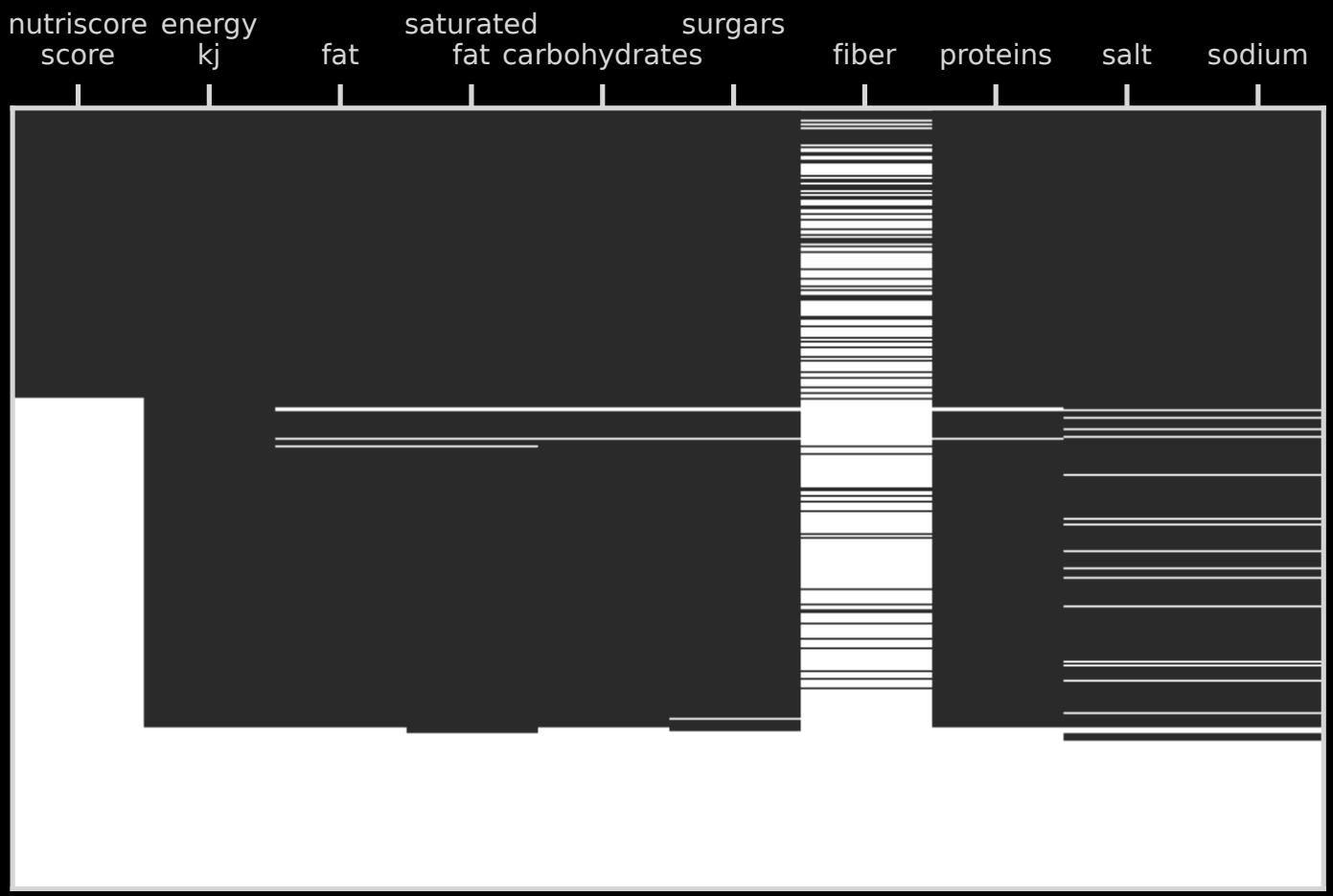
Taux de **remplissage** des **variables sélectionnées**:



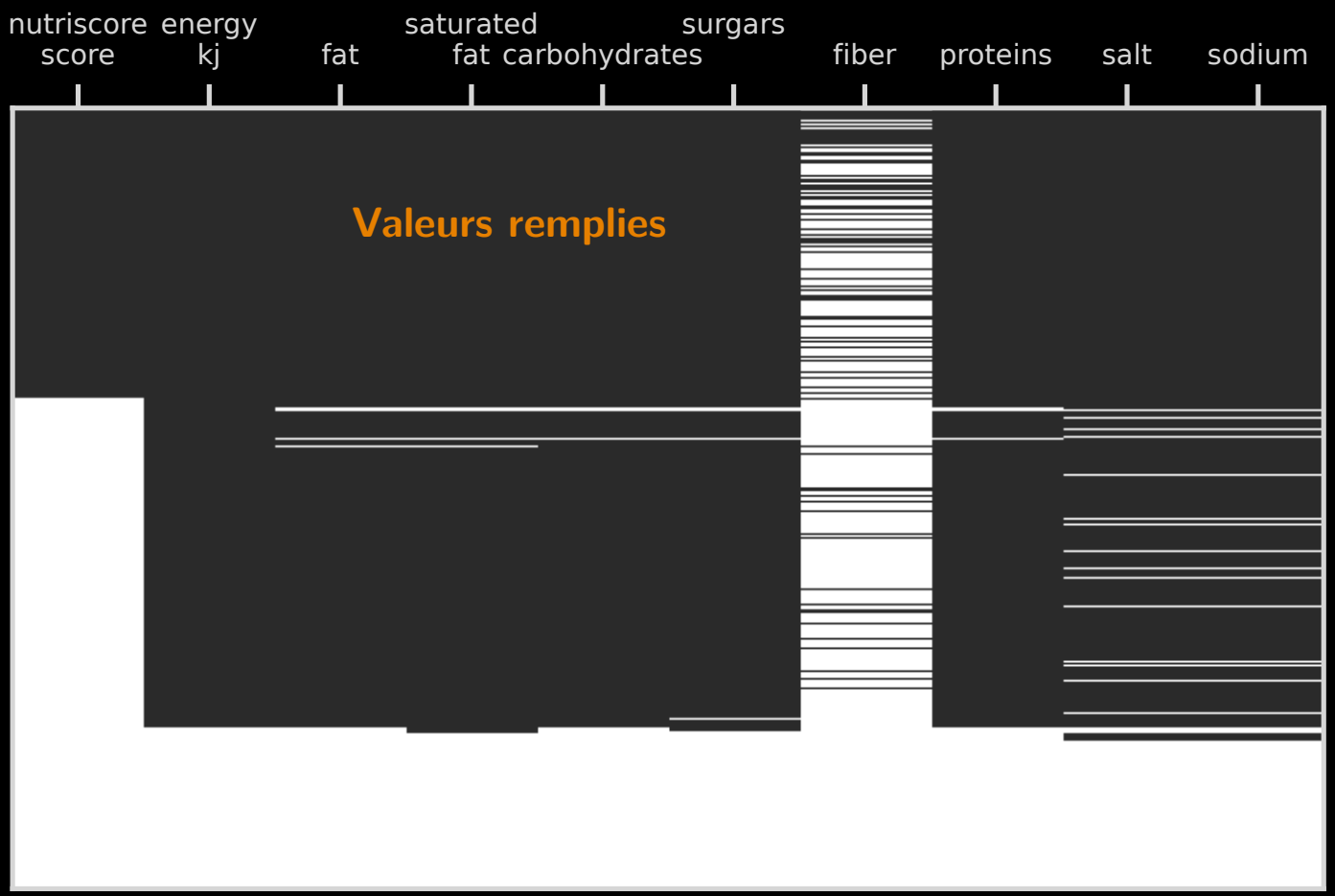
Taux de **remplissage** des **variables sélectionnées**:



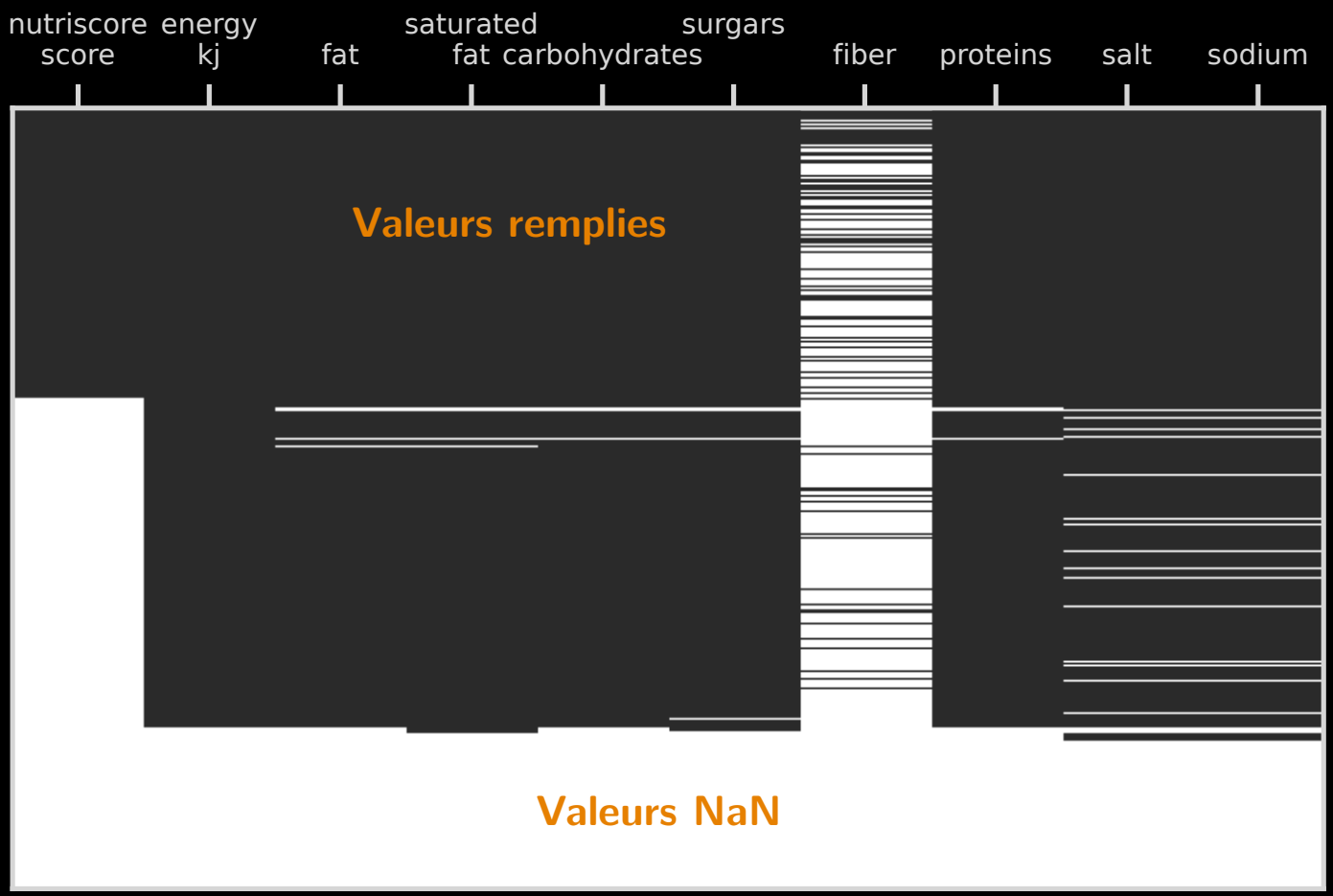
Matrice de **remplissage** des
variables sélectionnées:
(tri ascendant des valeurs)



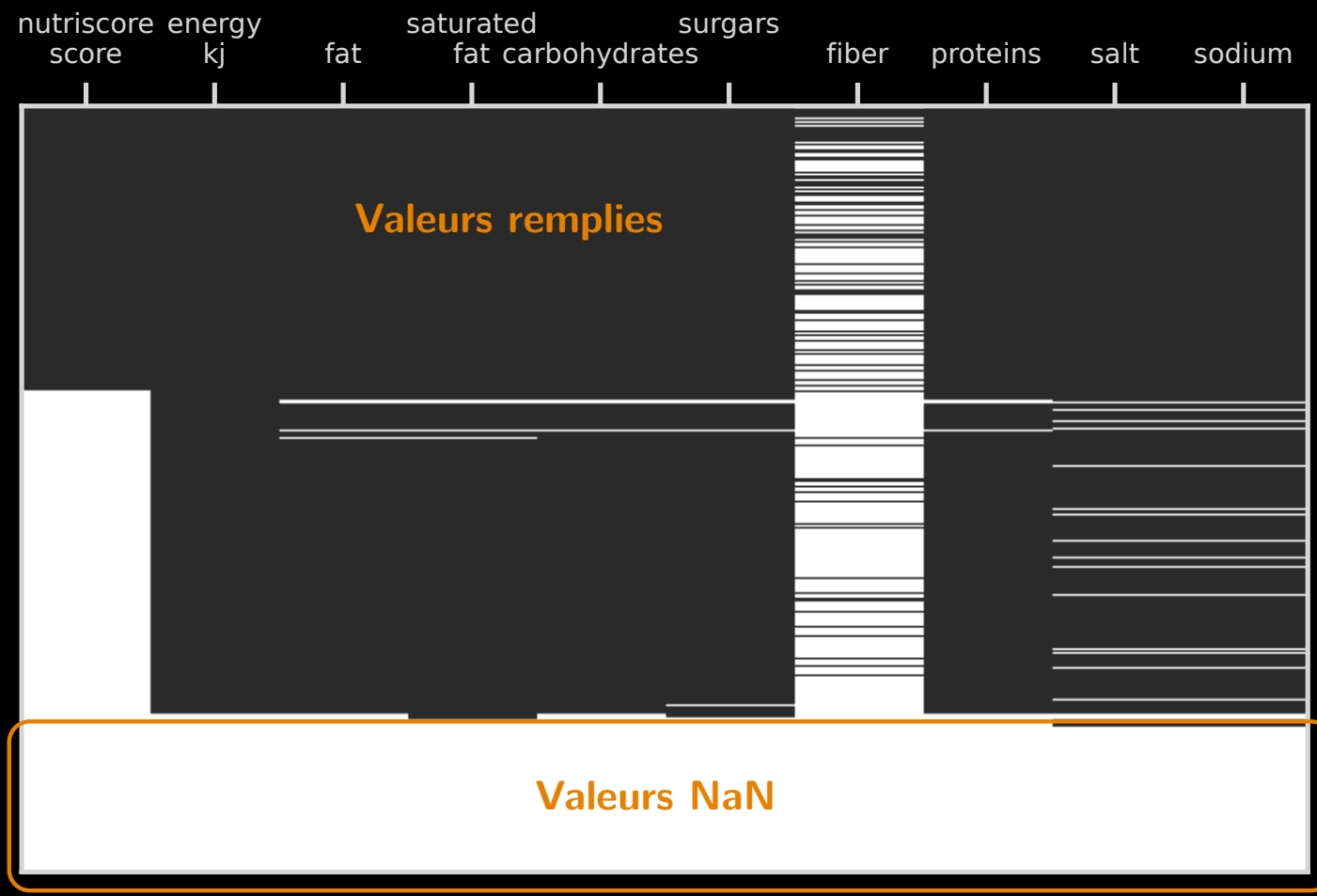
Matrice de **remplissage** des
variables sélectionnées:
(tri ascendant des valeurs)



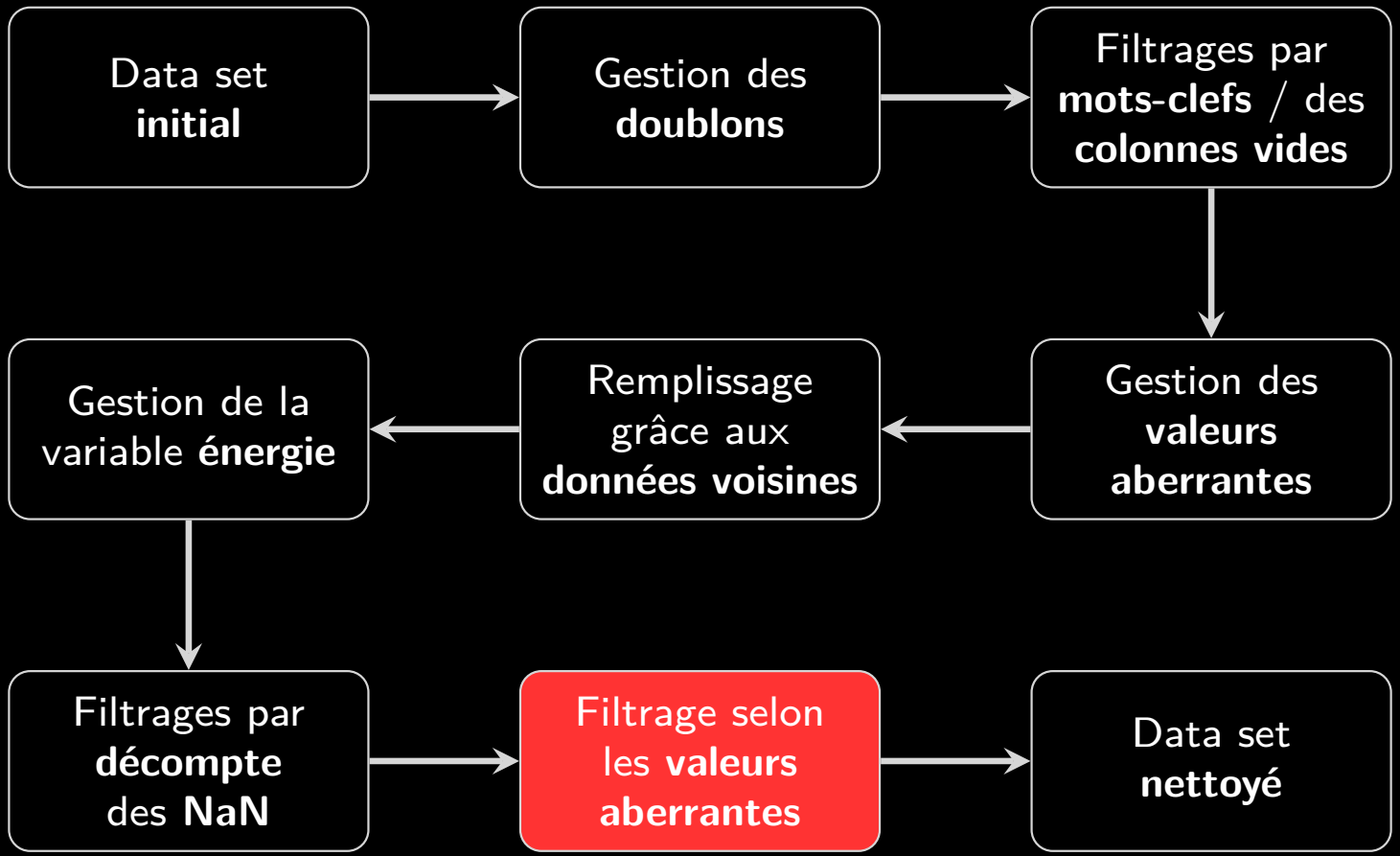
Matrice de **remplissage** des
variables sélectionnées:
(tri ascendant des valeurs)



Matrice de **remplissage** des
variables sélectionnées:
(tri ascendant des valeurs)



Suppression des **entrées vides**


Data set **initial**
Gestion des **doublons**
Filtrages par **mots-clefs** / des **colonnes vides**
Gestion des **valeurs aberrantes**
Remplissage grâce aux **données voisines**
Gestion de la variable **énergie**
Filtrages par **décompte** des **NaN**
Filtrage selon les **valeurs aberrantes**
Data set **nettoyé**

Calcul de la **somme** des **macro-nutriments**

Calcul de la **somme** des **macro-nutriments**

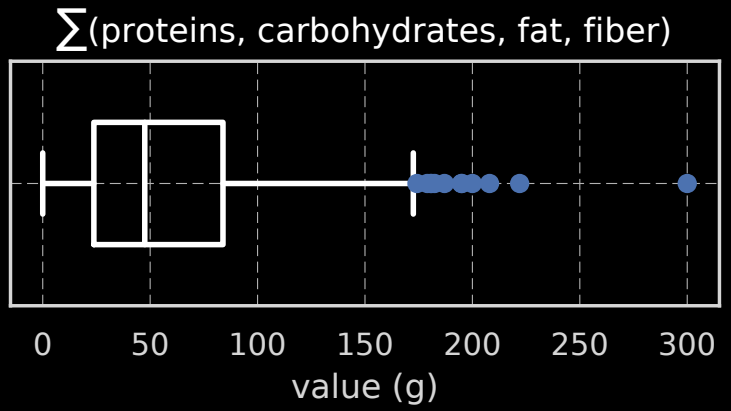
Données **statistiques** sur les **données brutes**:

	count	mean	std	min	25%	50%	75%	max
$\sum$ poids	244025	51.27	31.43	0	23.8	47.5	83.9	300

Calcul de la **somme** des **macro-nutriments**

Données **statistiques** sur les **données brutes**:

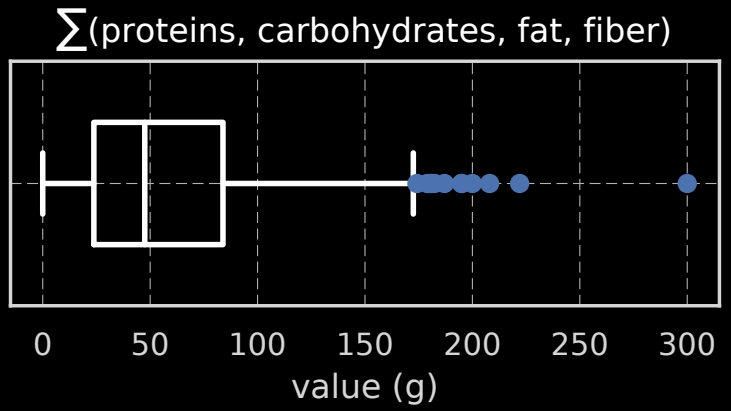
	count	mean	std	min	25%	50%	75%	max
∑ poids	244025	51.27	31.43	0	23.8	47.5	83.9	300



Calcul de la **somme** des **macro-nutriments**

Données **statistiques** sur les **données brutes**:

	count	mean	std	min	25%	50%	75%	max
$\sum$ poids	244025	51.27	31.43	0	23.8	47.5	83.9	300



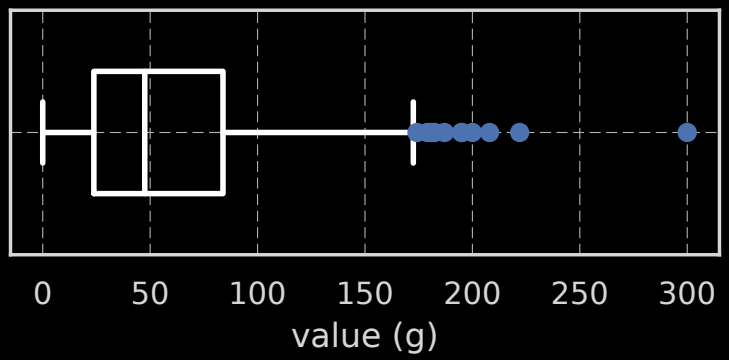
Q3+1.5*IQ = **174.05**
→ 12 outliers

Calcul de la **somme** des **macro-nutriments**

Données **statistiques** sur les **données brutes**:

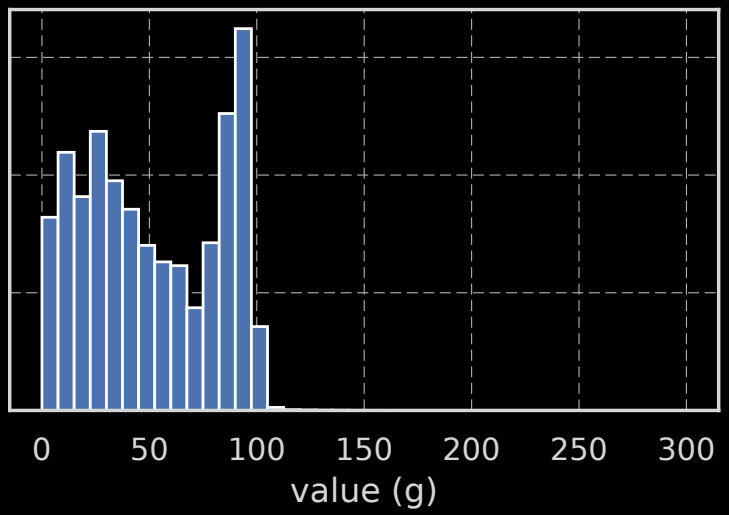
	count	mean	std	min	25%	50%	75%	max
∑ poids	244025	51.27	31.43	0	23.8	47.5	83.9	300

∑(proteins, carbohydrates, fat, fiber)



Q3+1.5*IQ = **174.05**
→ 12 outliers

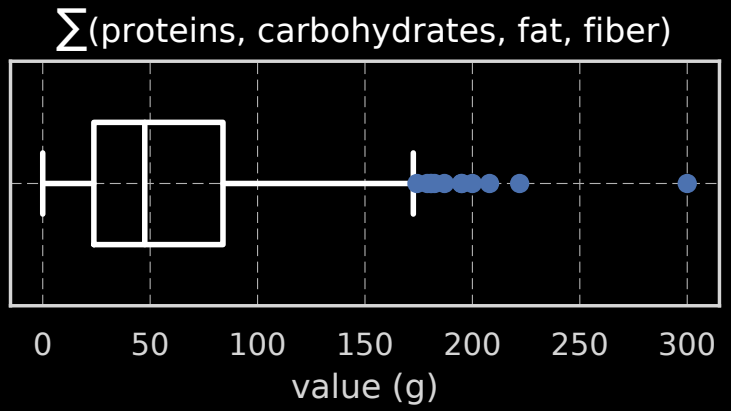
∑(proteins, carbohydrates, fat, fiber)



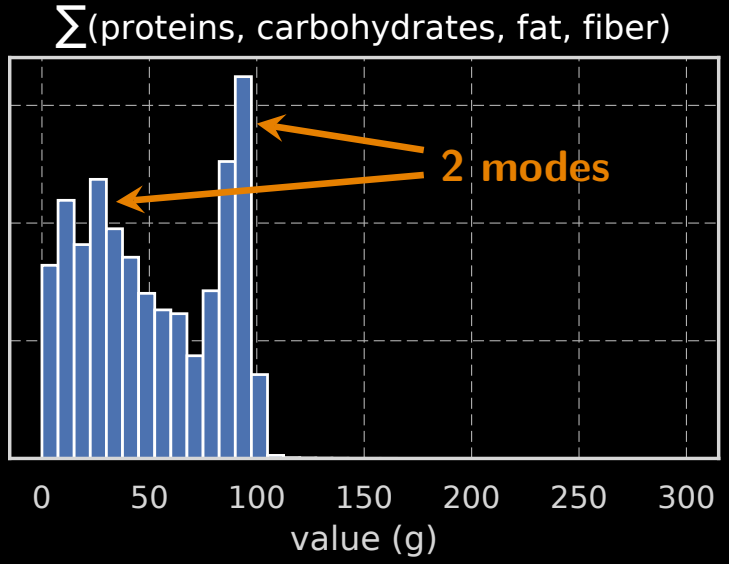
Calcul de la **somme** des **macro-nutriments**

Données **statistiques** sur les **données brutes**:

	count	mean	std	min	25%	50%	75%	max
∑ poids	244025	51.27	31.43	0	23.8	47.5	83.9	300



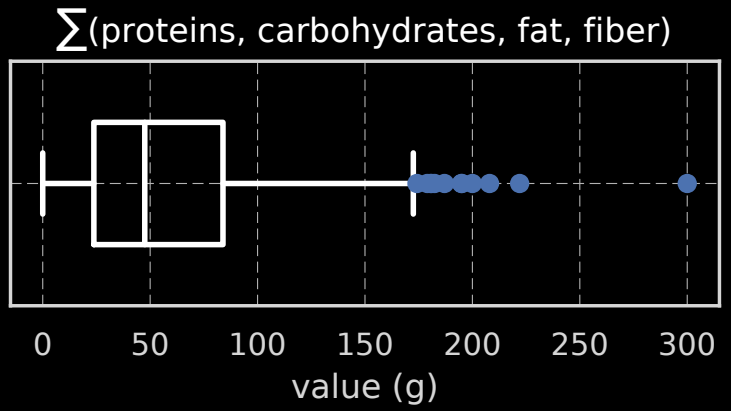
Q3+1.5*IQ = **174.05**
→ 12 outliers



Calcul de la **somme** des **macro-nutriments**

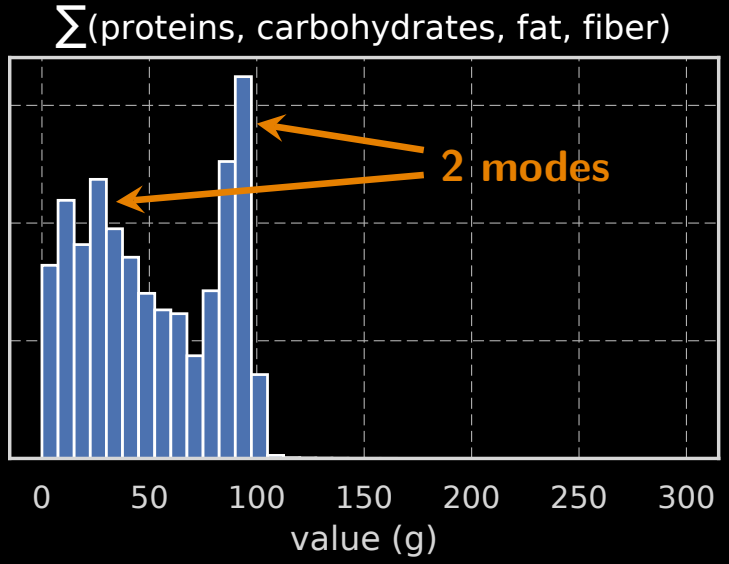
Données **statistiques** sur les **données brutes**:

	count	mean	std	min	25%	50%	75%	max
∑ poids	244025	51.27	31.43	0	23.8	47.5	83.9	300



Q3+1.5*IQ = **174.05**
→ 12 outliers

vmax = 100
→ **1651** outliers



EXPLORATION DES DONNÉES

EXPLORATION DES DONNÉES

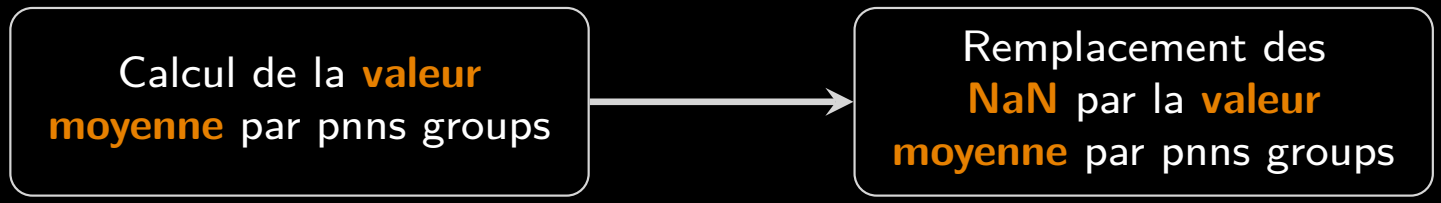
Remplissage des valeurs NaN éparses

Méthode 1: Valeur moyenne par pnns groups et par variable

Méthode 1: Valeur moyenne par pnns groups et par variable

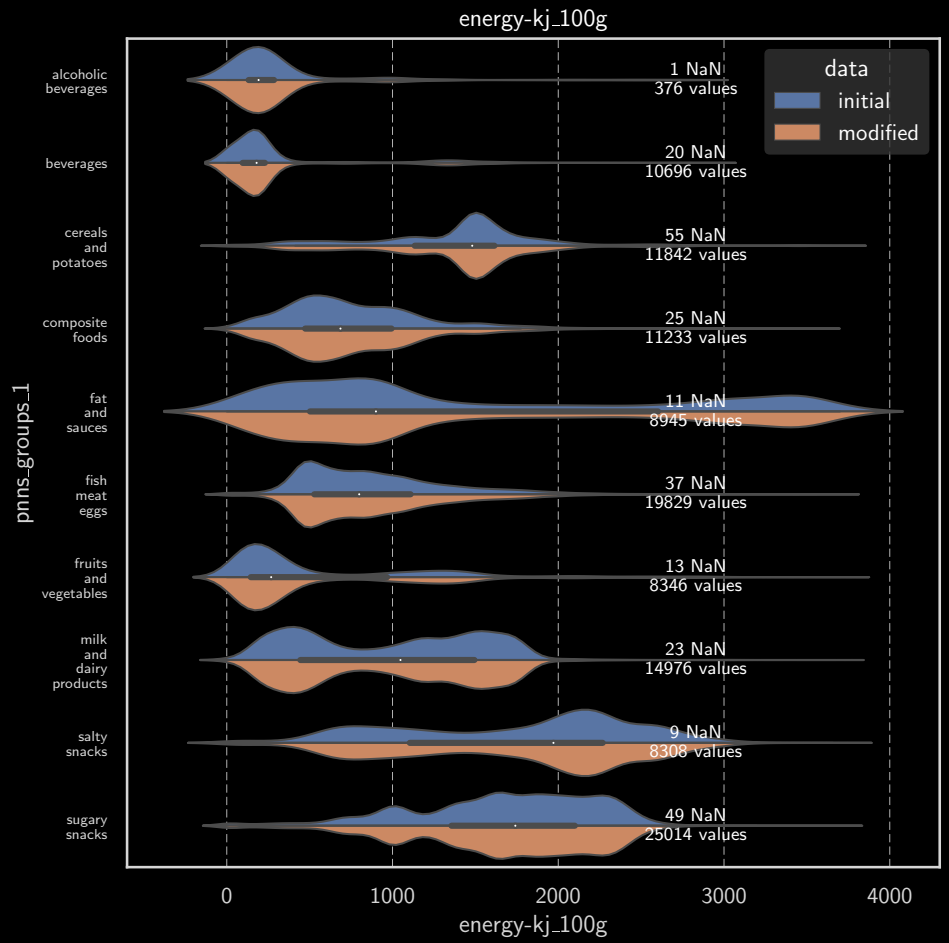
Calcul de la **valeur moyenne** par pnns groups

Méthode 1: Valeur moyenne par pnns groups et par variable



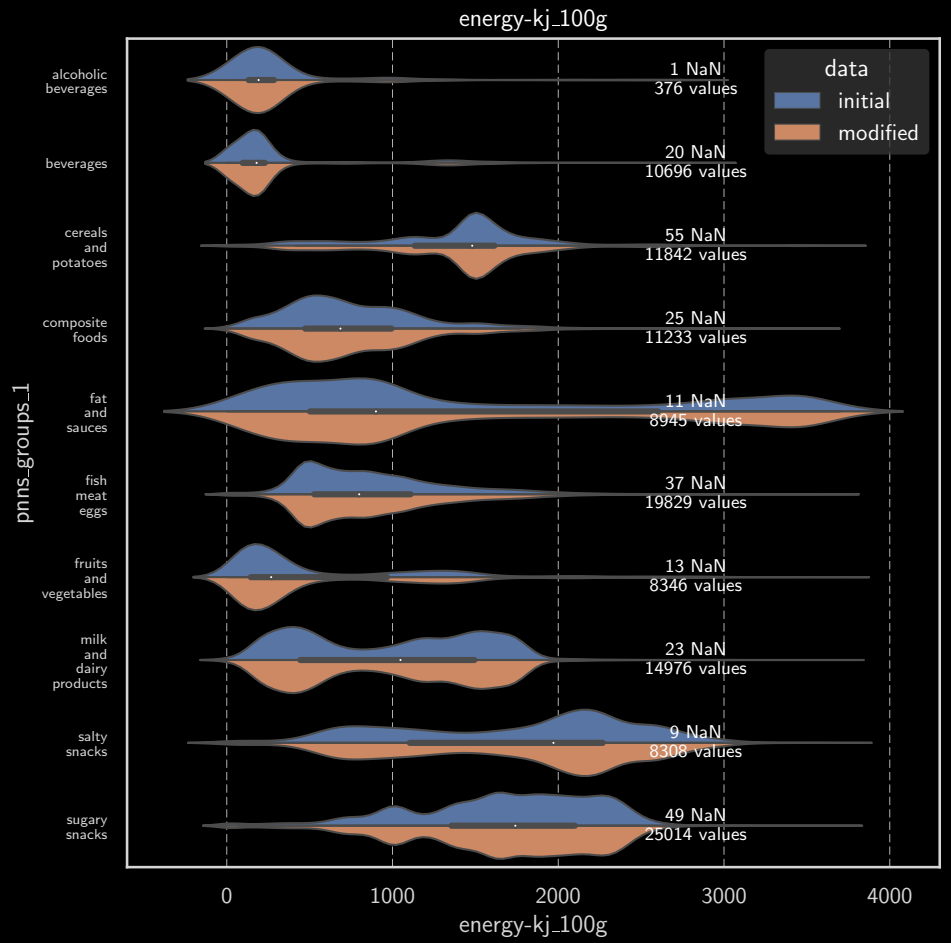
Méthode 1: Valeur moyenne par pnns groups et par variable

Calcul de la **valeur moyenne** par pnns groups → Remplacement des **NaN** par la **valeur moyenne** par pnns groups



Méthode 1: Valeur moyenne par pnns groups et par variable

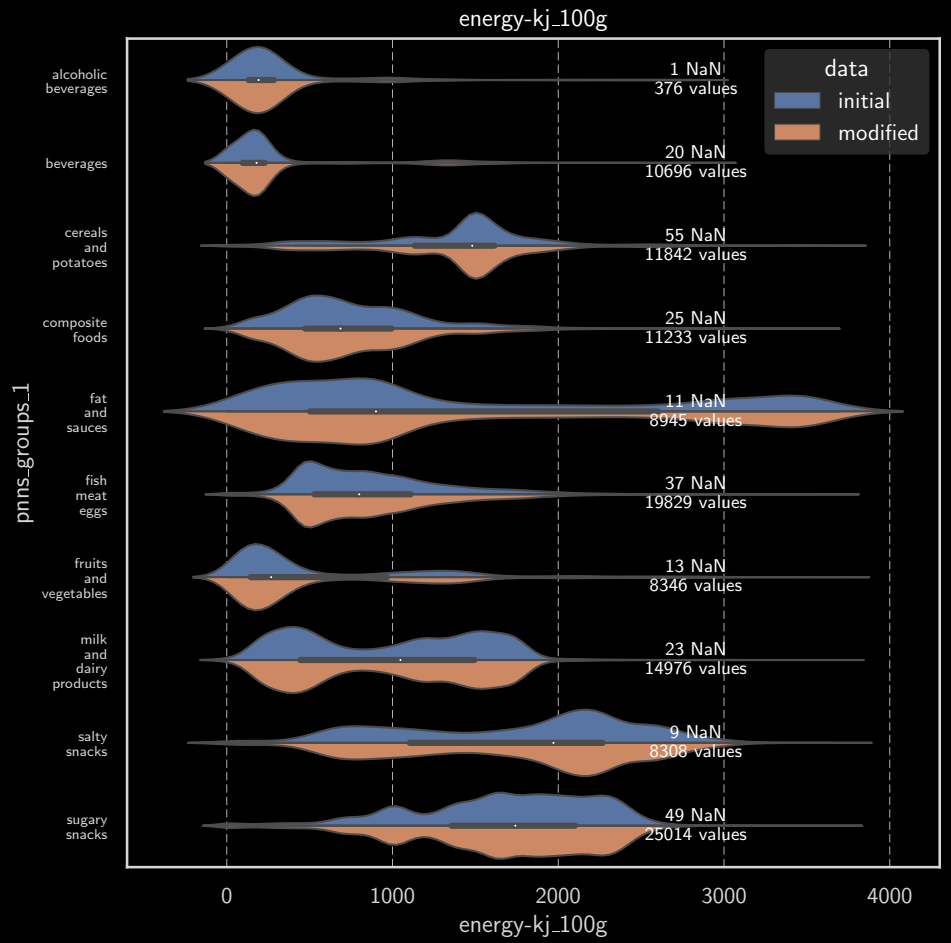
Calcul de la **valeur moyenne** par pnns groups → Remplacement des **NaN** par la **valeur moyenne** par pnns groups



- variations imperceptibles
- distribution **multi-modale**

Méthode 1: Valeur moyenne par pnns groups et par variable

Calcul de la **valeur moyenne** par pnns groups → Remplacement des **NaN** par la **valeur moyenne** par pnns groups



- variations imperceptibles
- distribution **multi-modale**

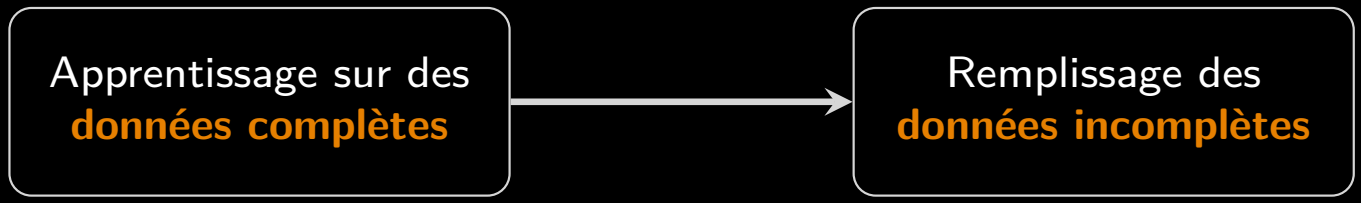
→ Méthode **non adaptée** aux données

Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

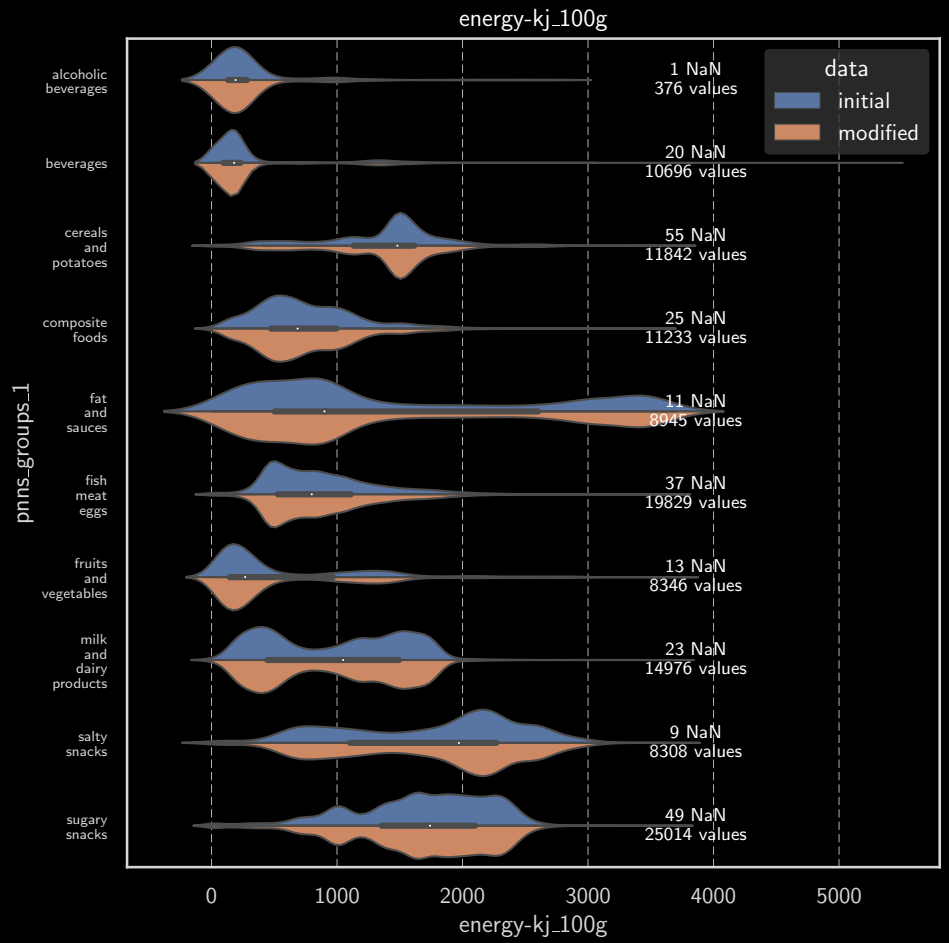
Apprentissage sur des
données complètes

Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn



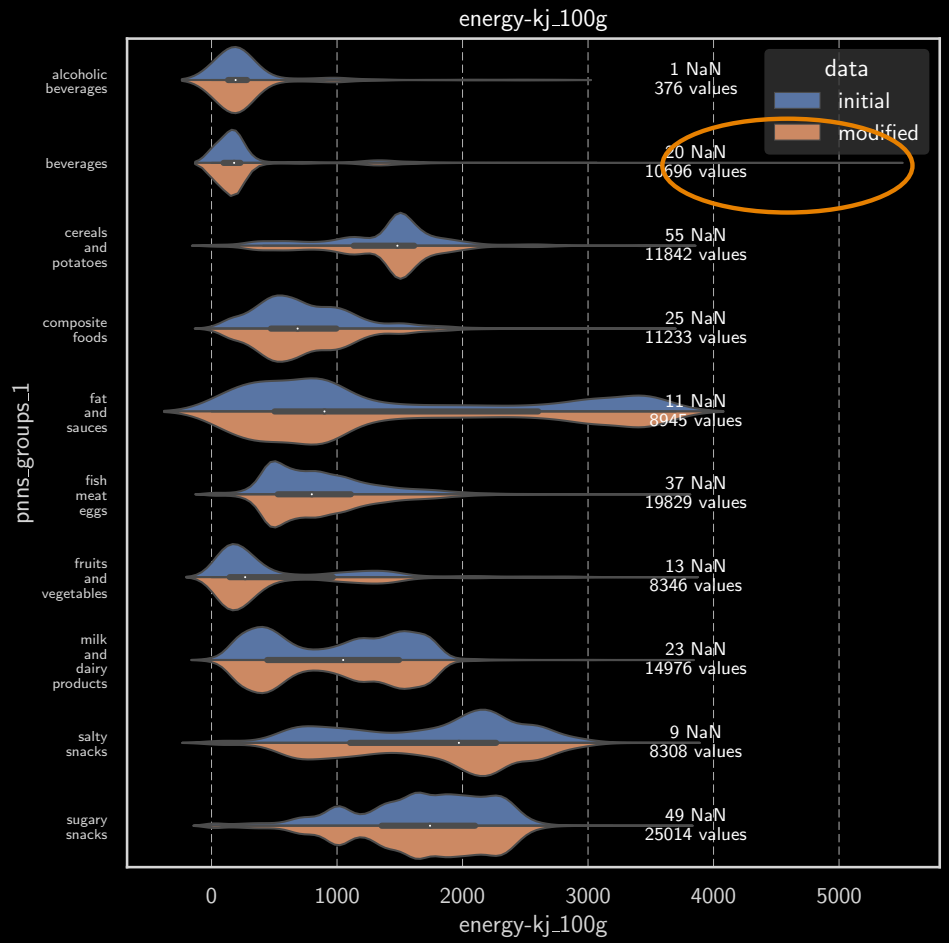
Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

Apprentissage sur des **données complètes** → Remplissage des **données incomplètes**



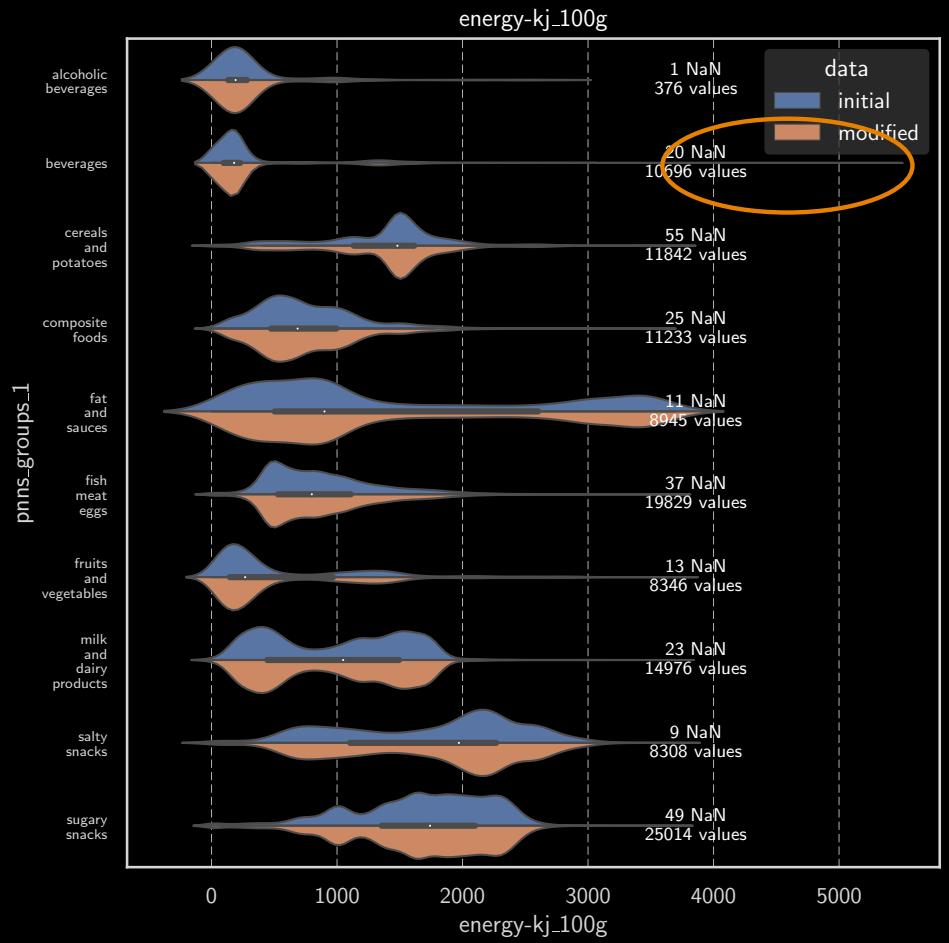
Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

Apprentissage sur des **données complètes** → Remplissage des **données incomplètes**



Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

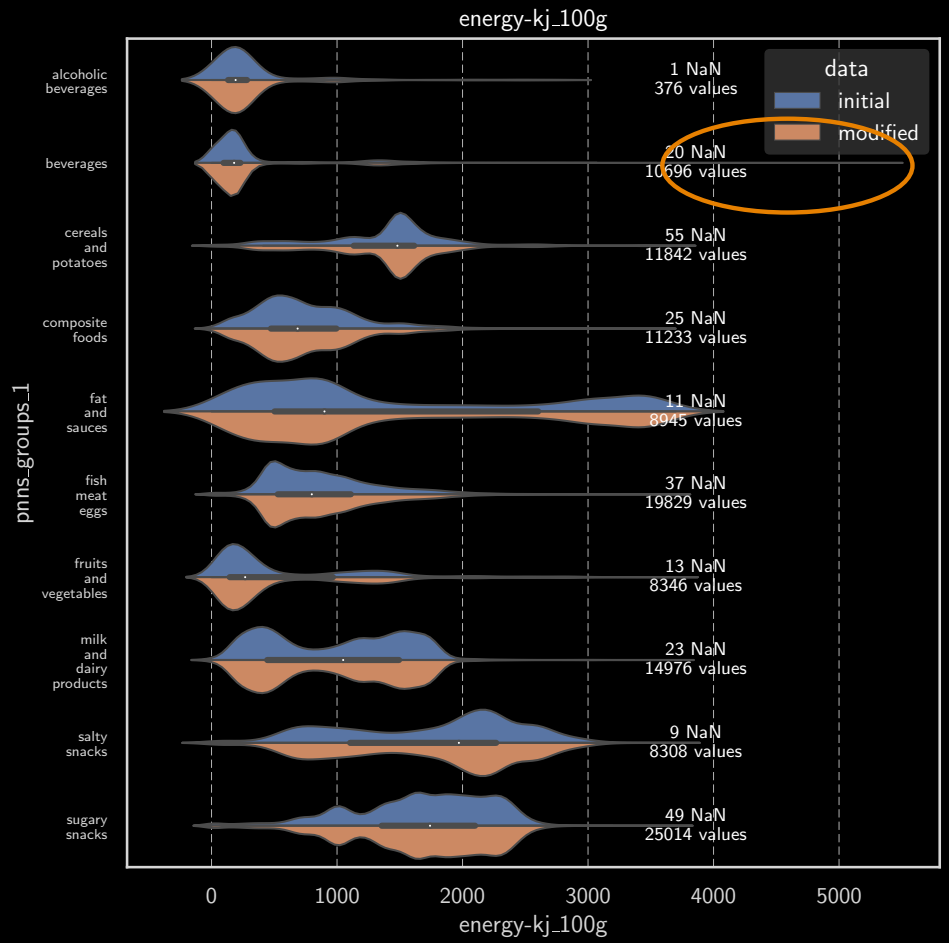
Apprentissage sur des **données complètes** → Remplissage des **données incomplètes**



- génération de **valeurs hors bornes**

Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

Apprentissage sur des **données complètes** → Remplissage des **données incomplètes**



- génération de **valeurs hors bornes**

Pour l'étude, **suppression** des entrées **incomplètes**

EXPLORATION DES DONNÉES

Prédiction du nutriscore (remplissage des NaN)

Principe:

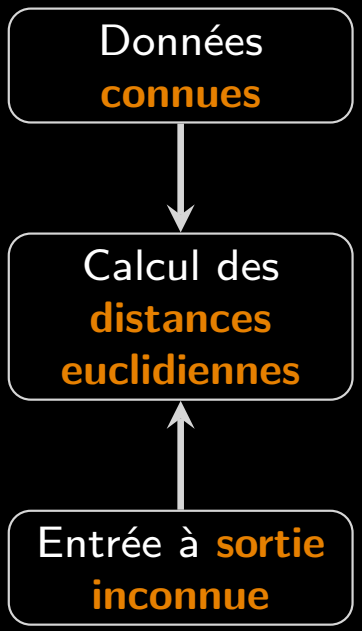
Données
connues

Principe:

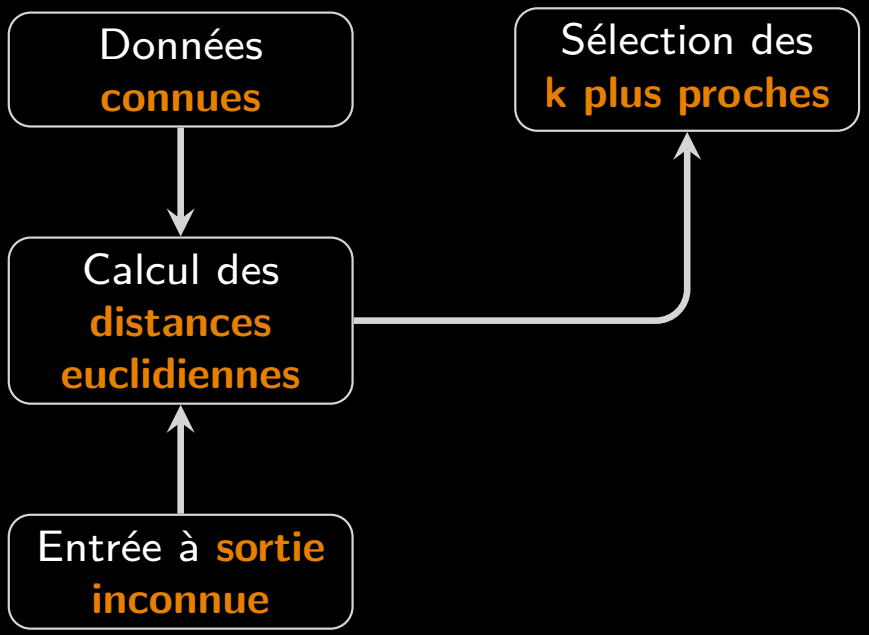
Données **connues**

Entrée à **sortie inconnue**

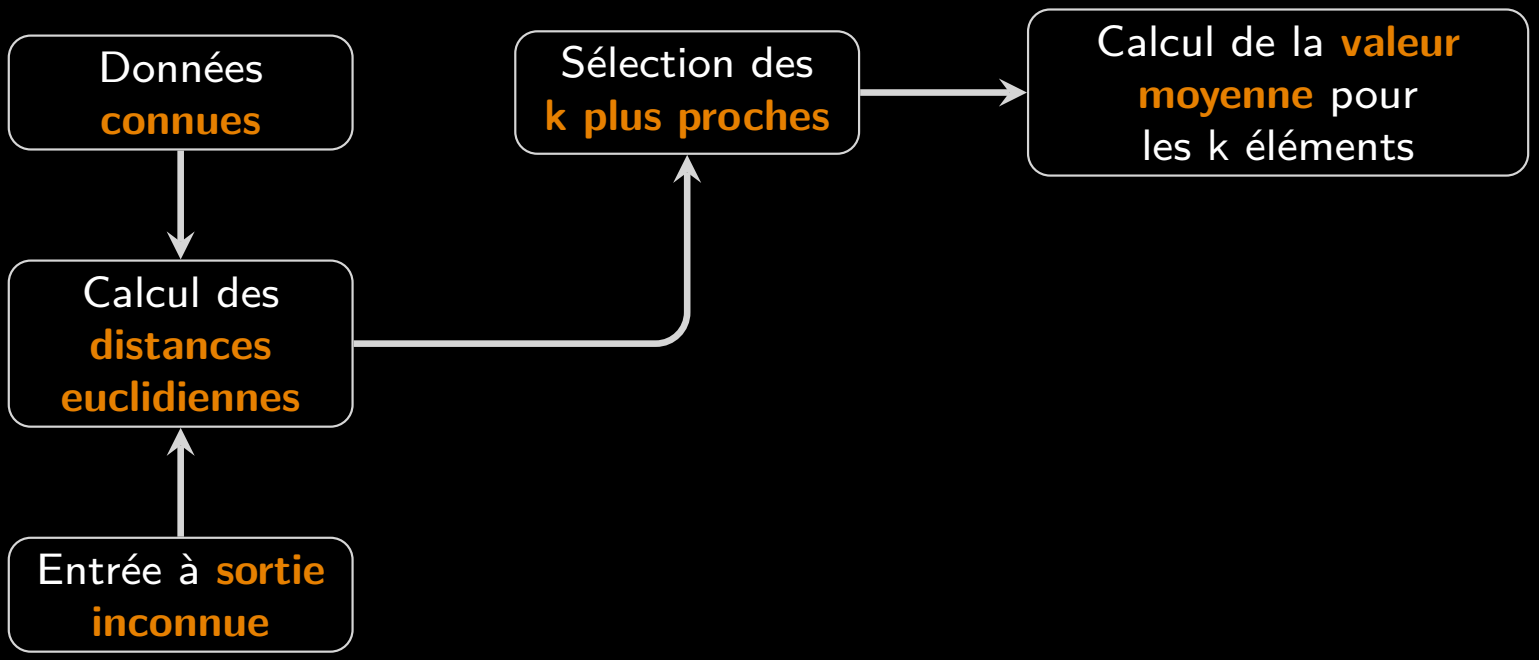
Principe:


Données
connues
Calcul des
distances
euclidiennes
Entrée à sortie
inconnue

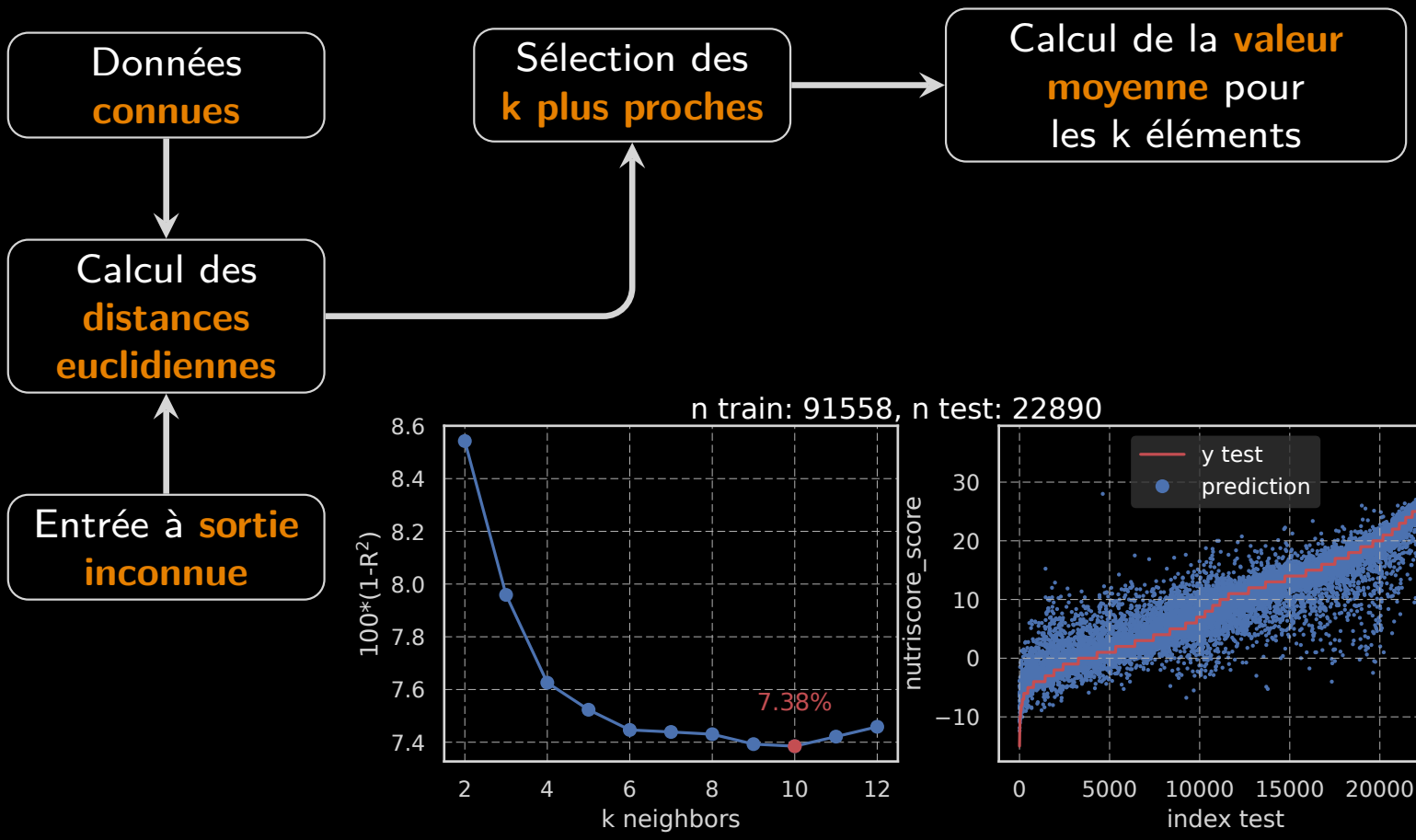
Principe:


Données
connues
Sélection des
k plus proches
Calcul des
distances
euclidiennes
Entrée à sortie
inconnue

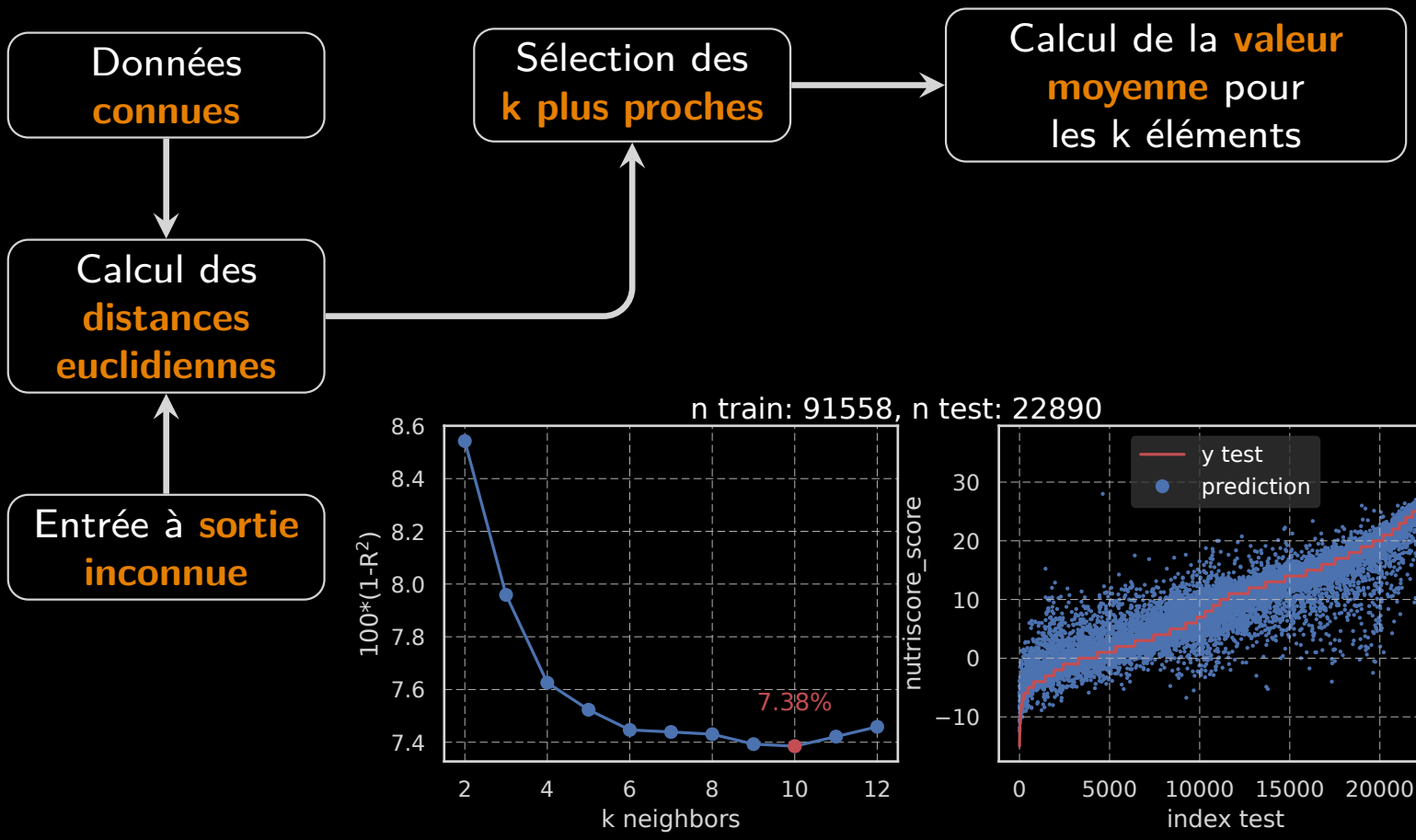
Principe:


Données
connues
Calcul des
distances
euclidiennes
Entrée à sortie
inconnue
Sélection des
k plus proches
Calcul de la valeur
moyenne pour
les k éléments

Principe:

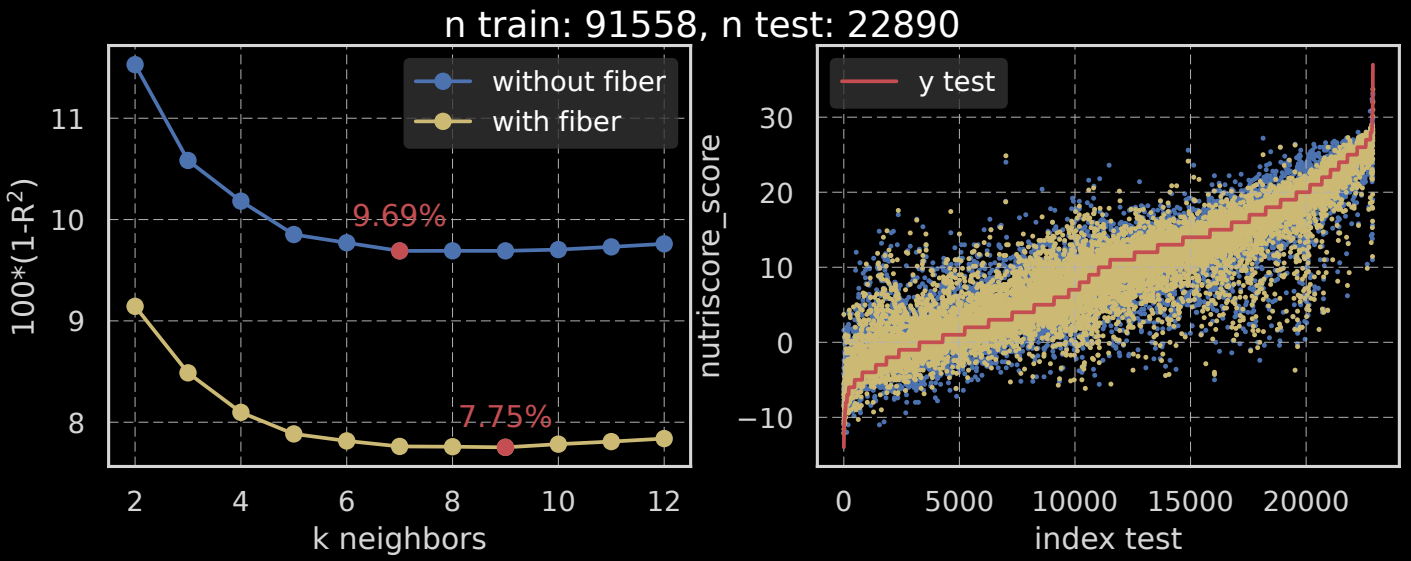

Données connues
Calcul des distances euclidiennes
Entrée à sortie inconnue
Sélection des k plus proches
Calcul de la valeur moyenne pour les k éléments
n train: 91558, n test: 22890
100*(1-R²)
k neighbors
7.38%
nutriscore_score
index test
y test
prediction

Principe:

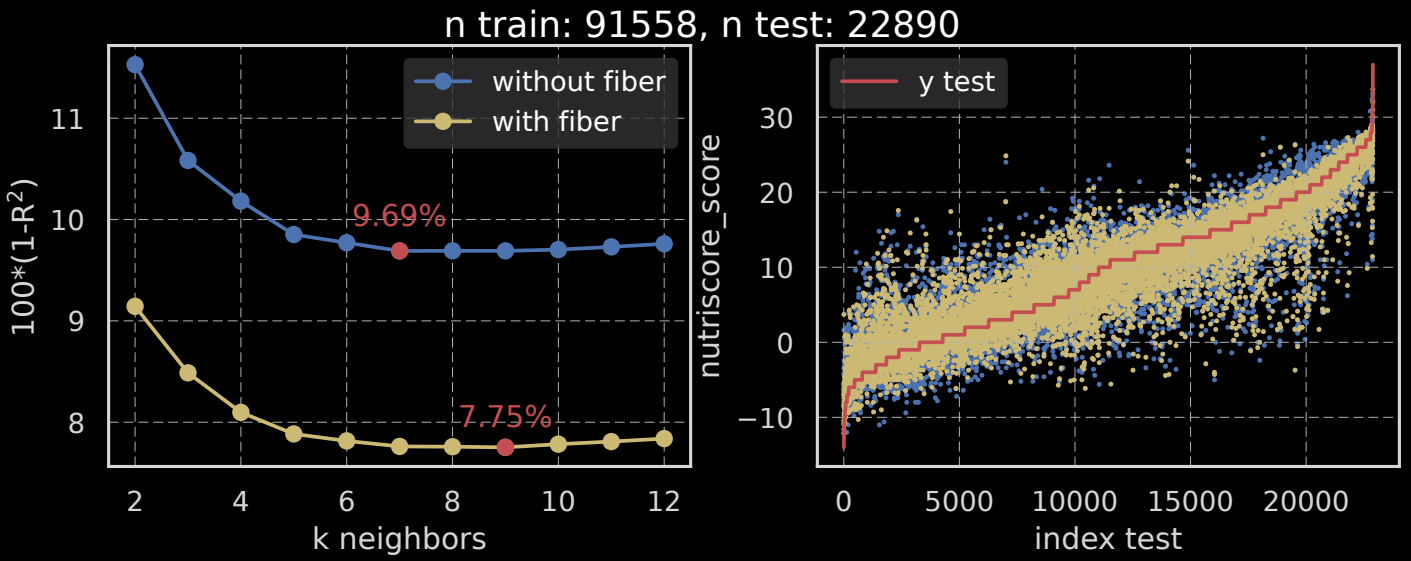


Quid de l'**hypothèse** NaN fibers = 0 ?

Vérification de l'hypothèse NaN fiber = 0:



Vérification de l'hypothèse NaN fiber = 0:



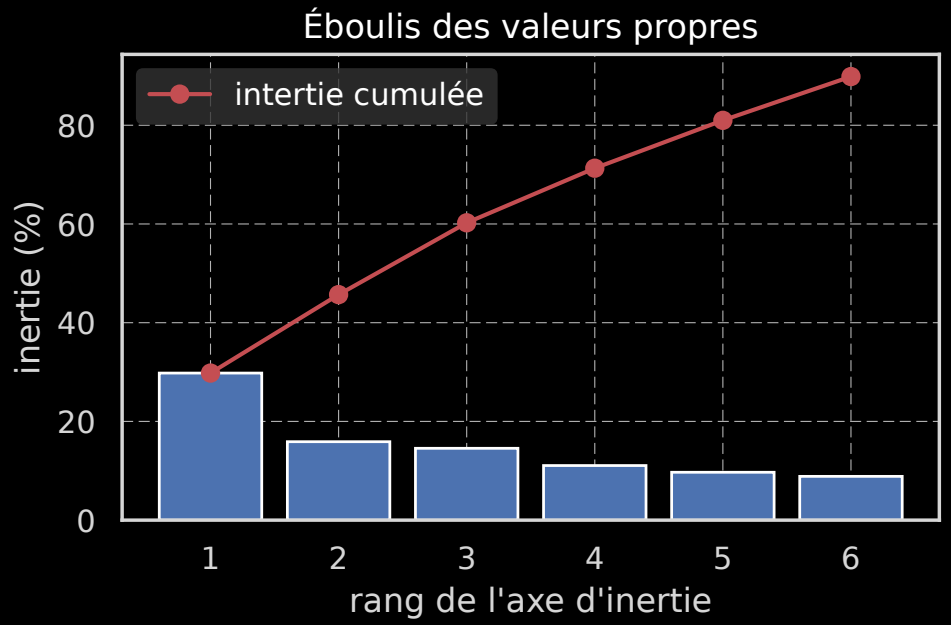
→ Hypothèse à priori **validée**

EXPLORATION DES DONNÉES

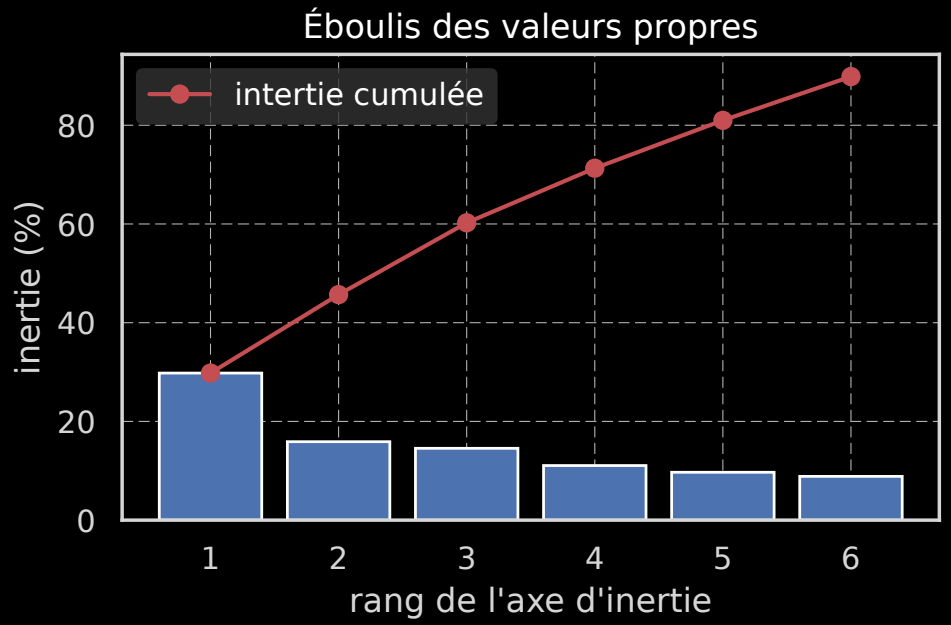
Étude de l'inertie des valeurs - Analyse en Composantes Principales

Principe: Calcul des axes en vue d'**aligner** les **coordonnées** et les **principaux axes d'inertie**

Principe: Calcul des axes en vue d'**aligner** les **coordonnées** et les **principaux axes d'inertie**



Principe: Calcul des axes en vue d'**aligner** les **coordonnées** et les **principaux axes d'inertie**

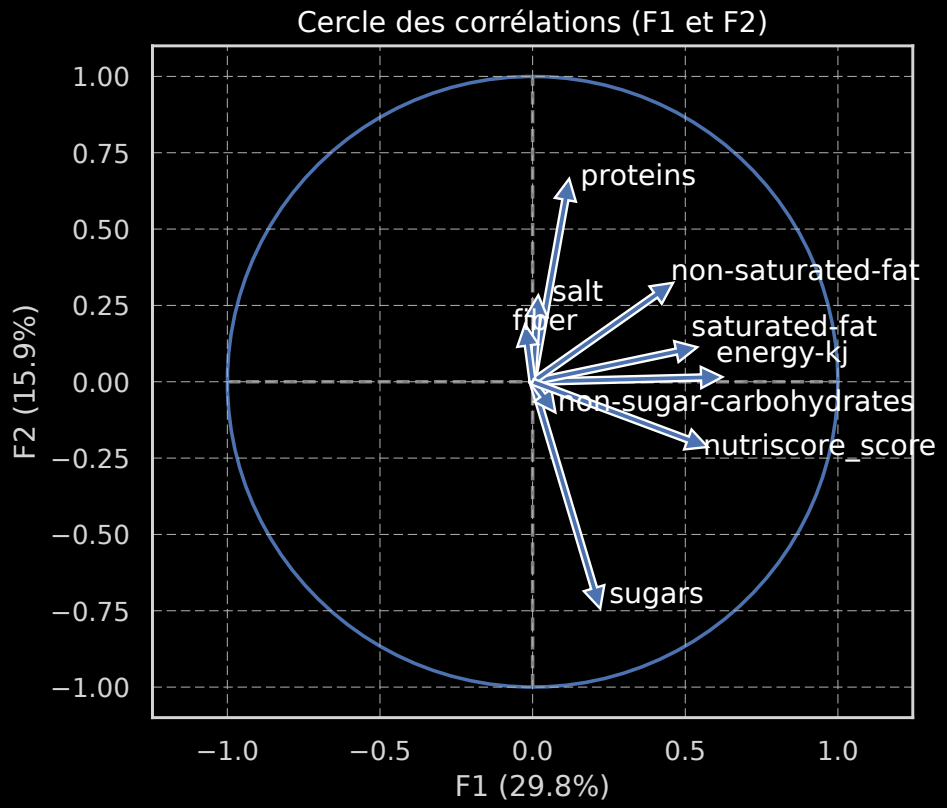


→ **90%** de l'inertie sur **6 axes**

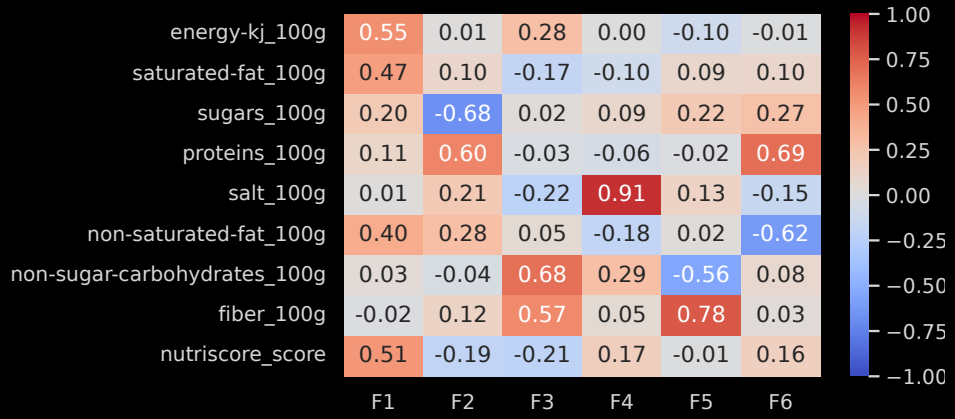
→ Répartition relativement **équilibrée**

Projection des axes initiaux sur les axes estimés

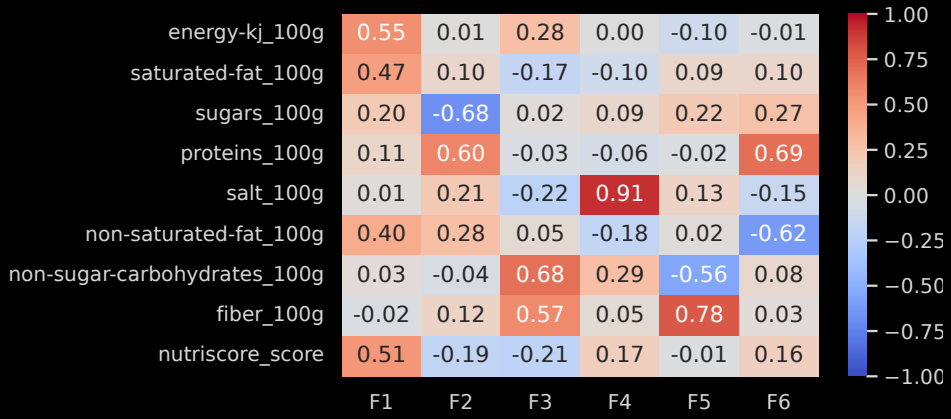
Projection des axes initiaux sur les axes estimés

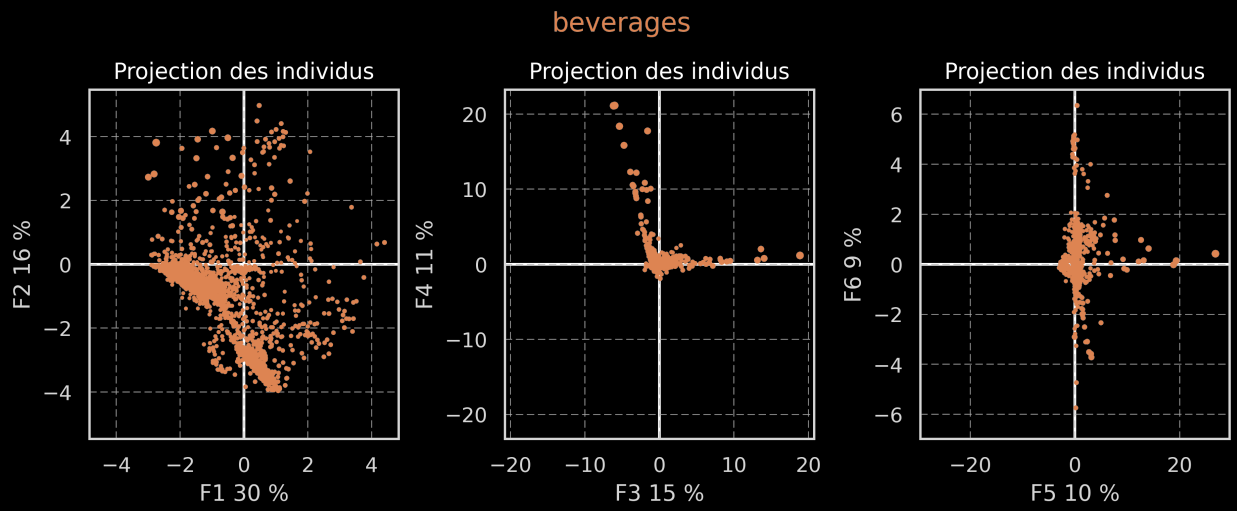


Projection des axes initiaux sur les axes estimés



Projection des axes initiaux sur les axes estimés



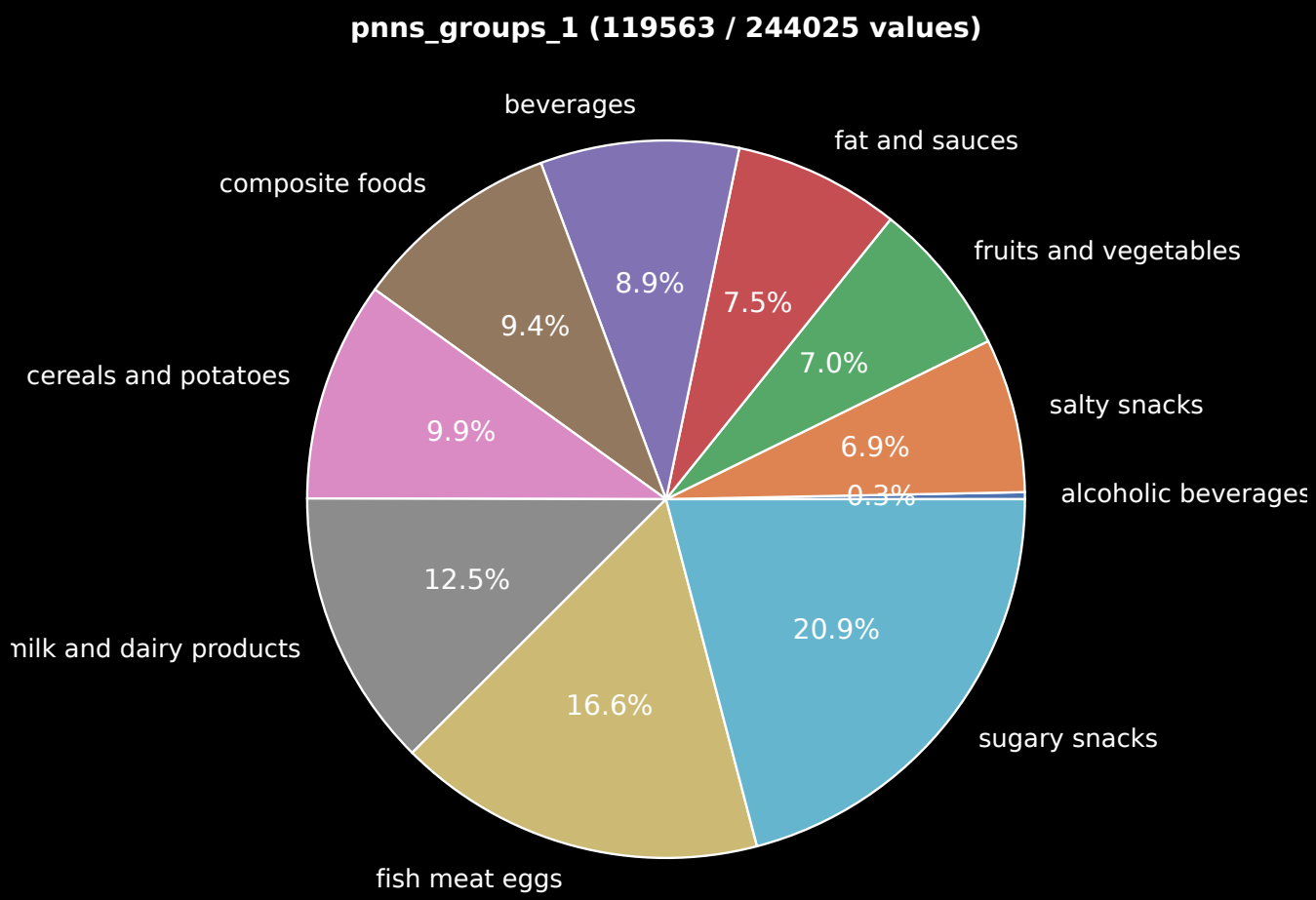


EXPLORATION DES DONNÉES

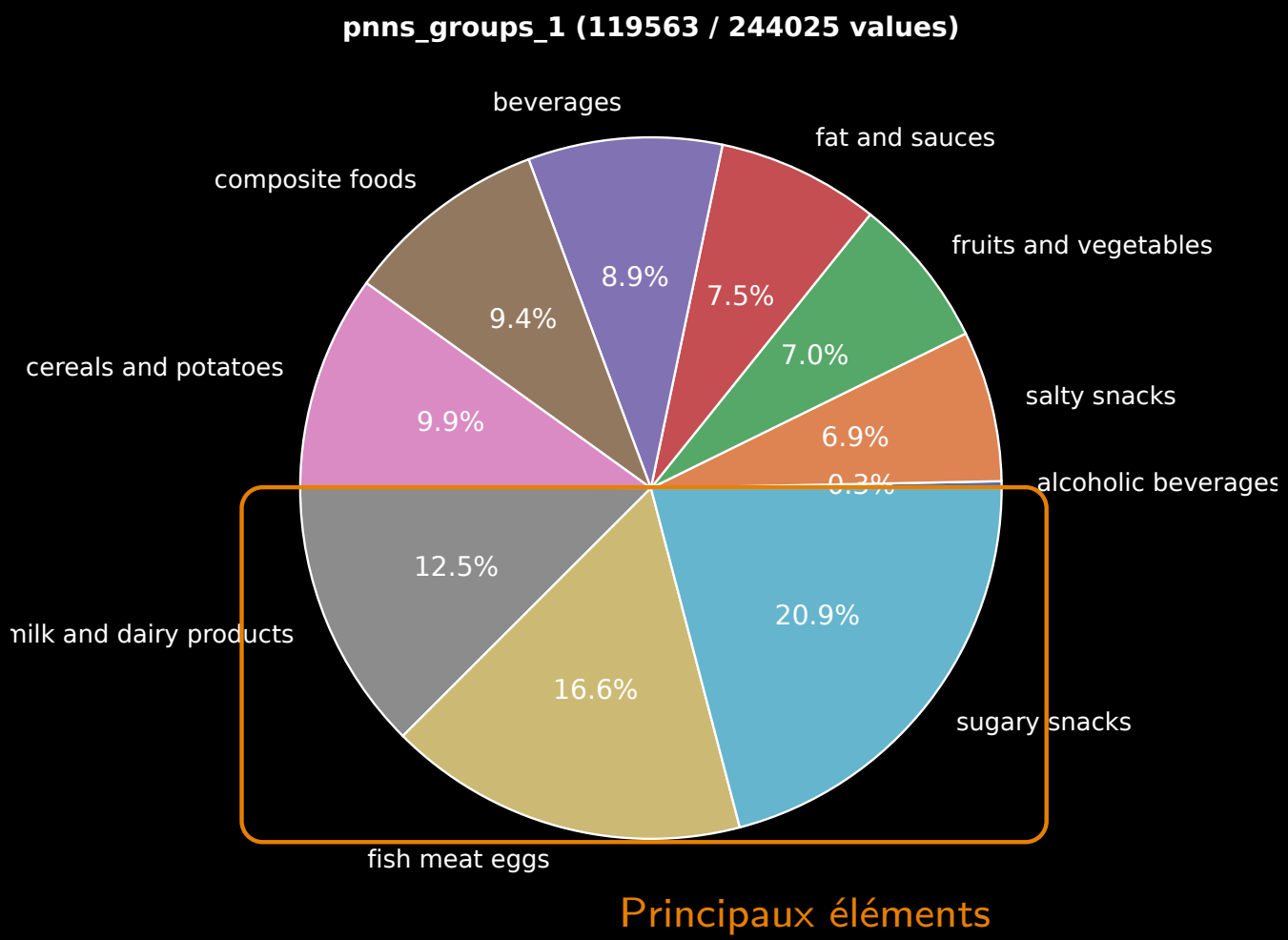
Analyse des données en lien avec l'application

Répartition pnns groups et nutriscore grade:

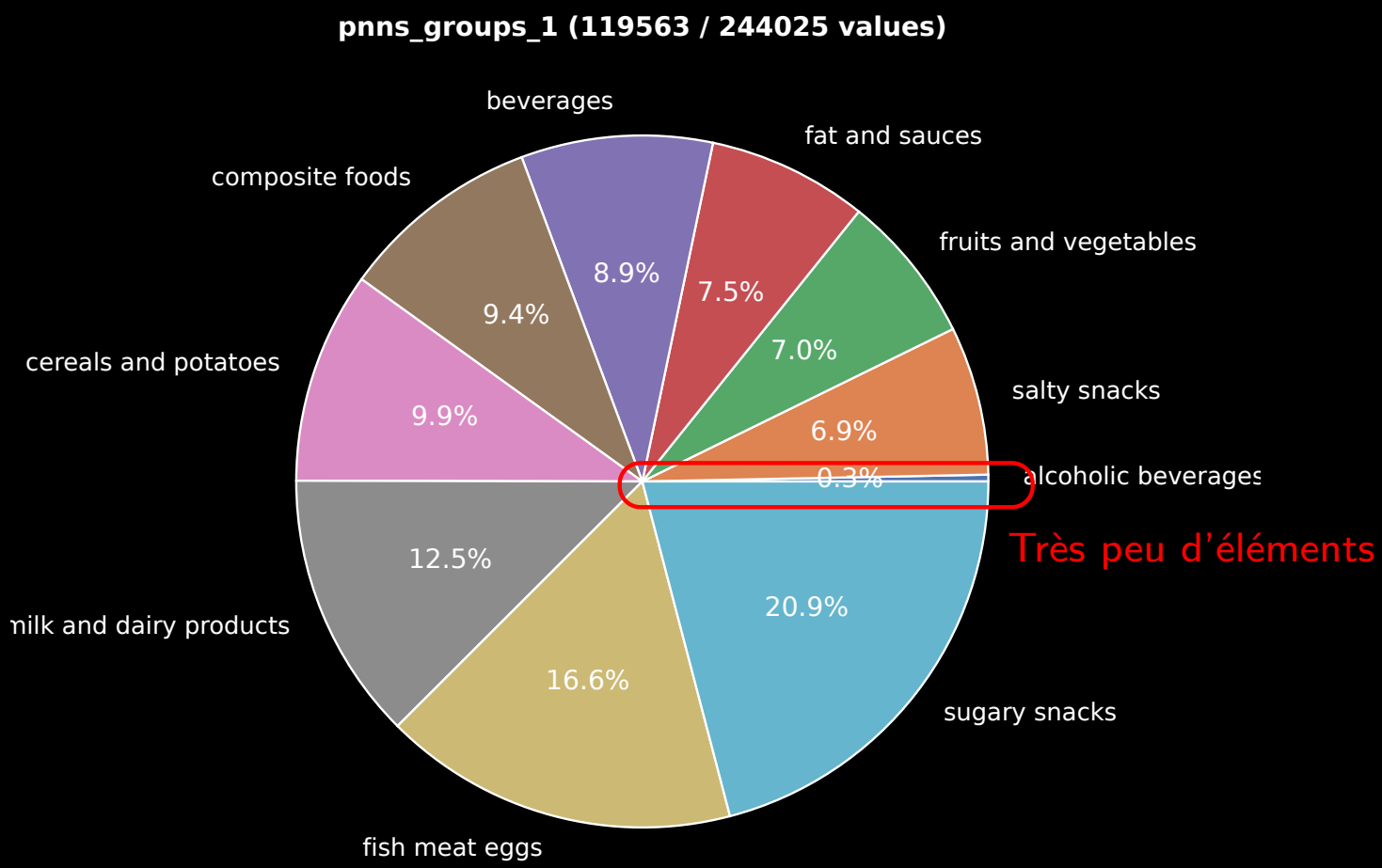
Répartition pnns groups et nutriscore grade:



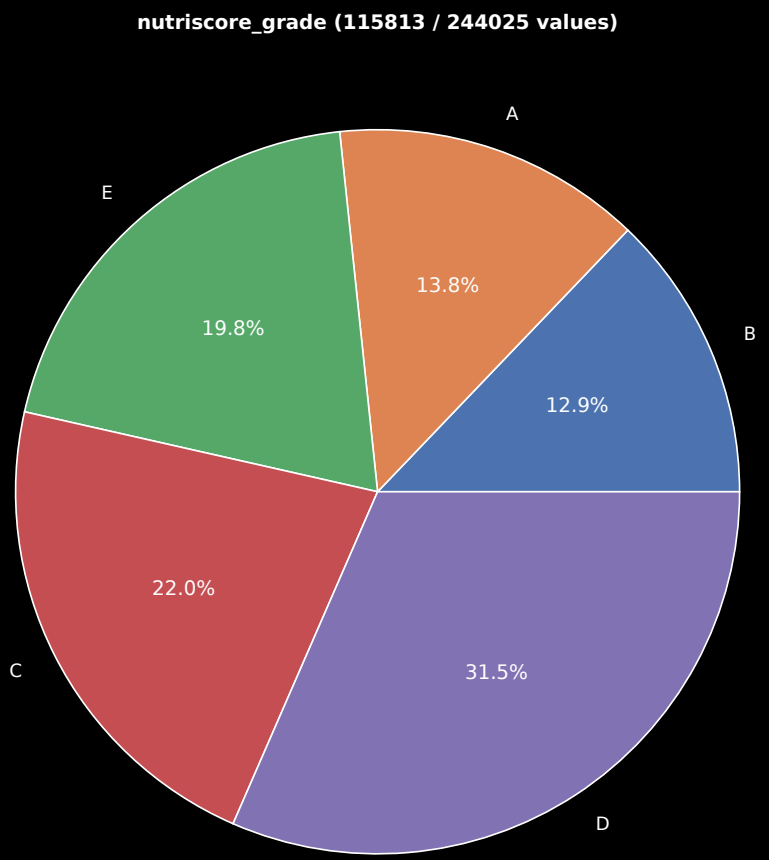
Répartition pnns groups et nutriscore grade:



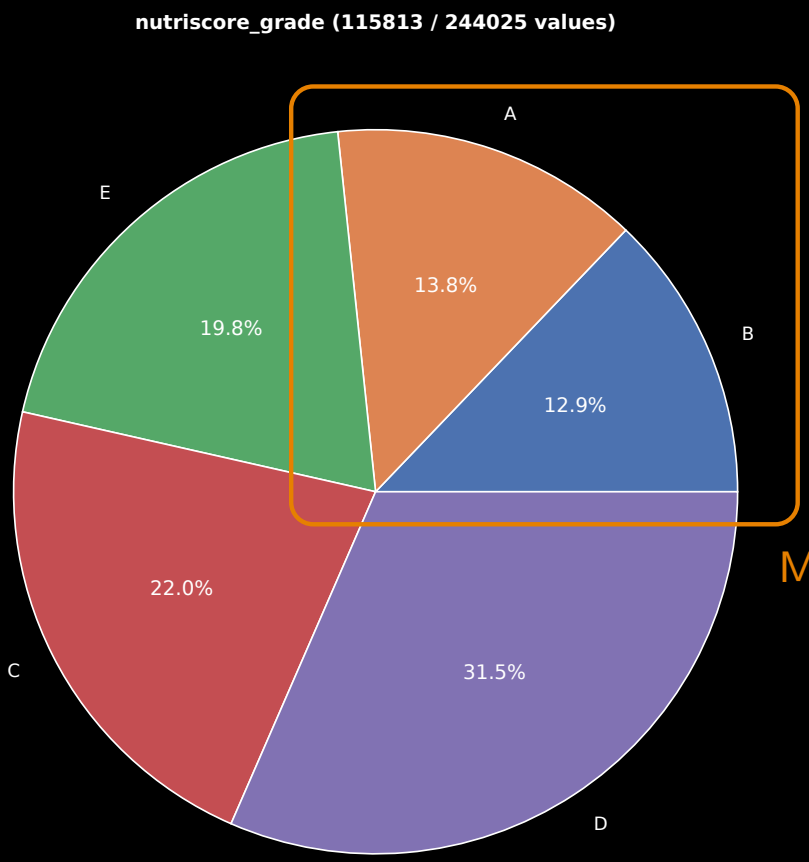
Répartition pnns groups et nutriscore grade:



Répartition pnns groups et nutriscore grade:



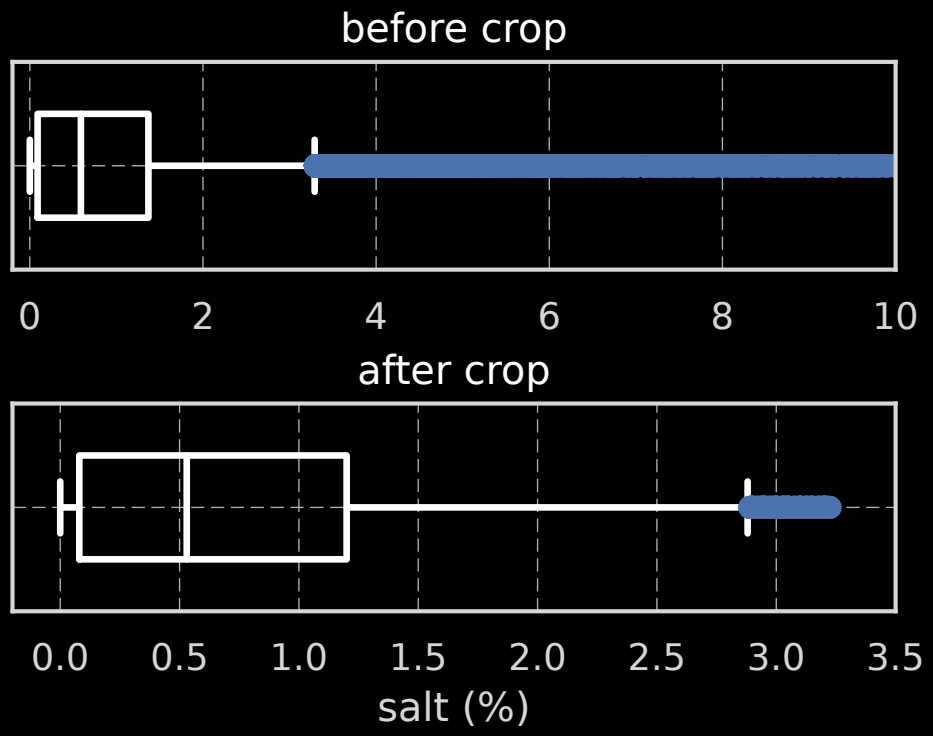
Répartition pnns groups et nutriscore grade:



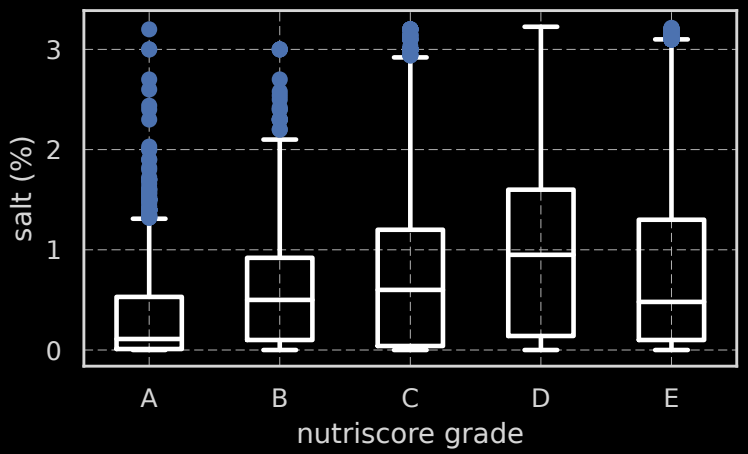
Moindre proportion

Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:

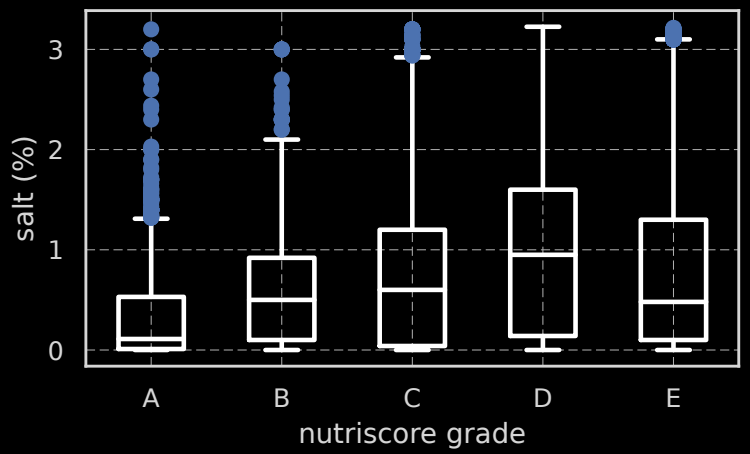
Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:

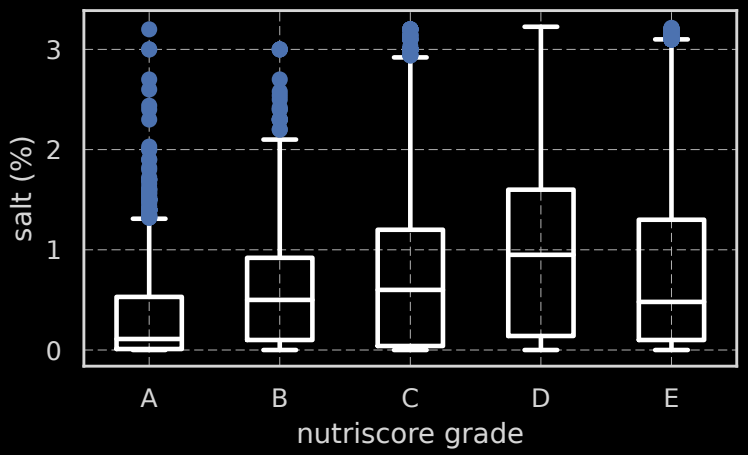


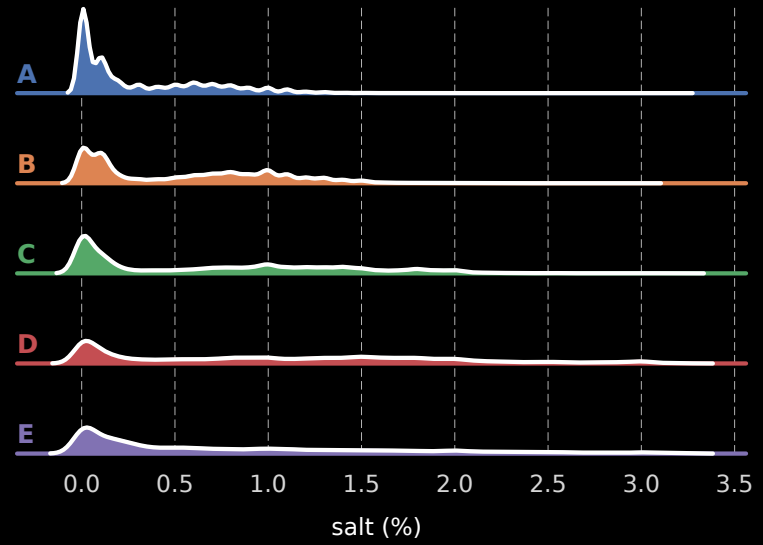
Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



Les groupes sont **différents**

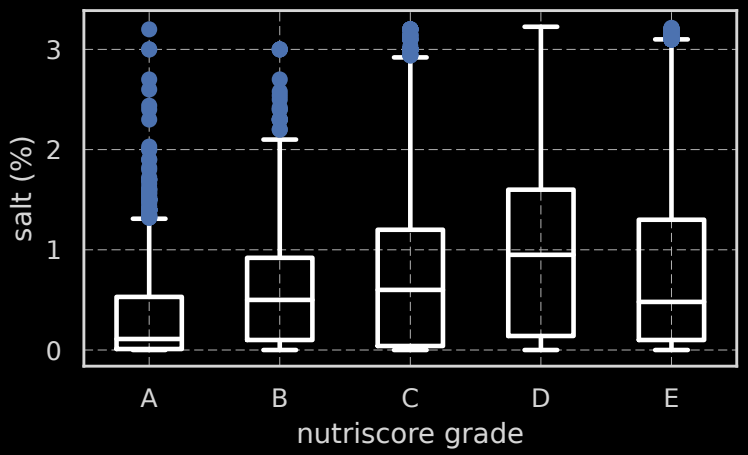
Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



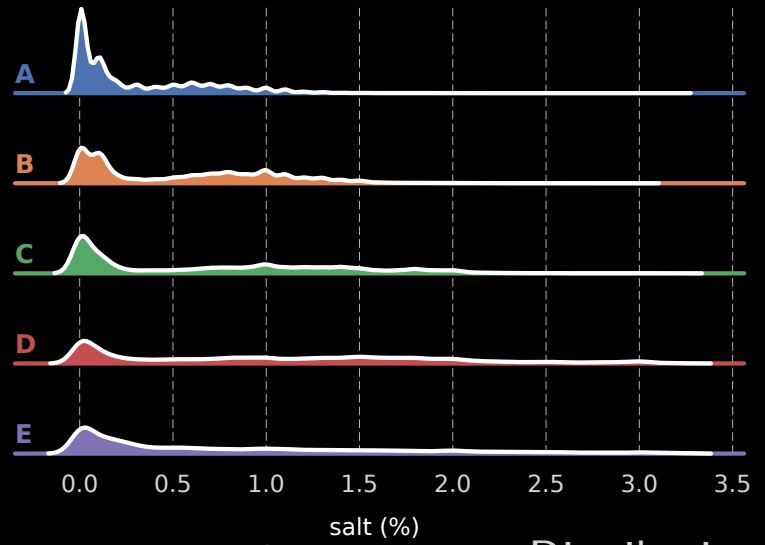


Les groupes sont **différents**

Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



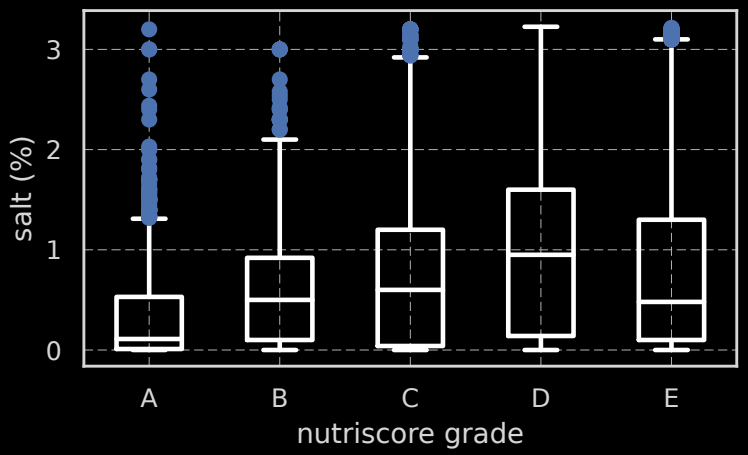
Les groupes sont **différents**



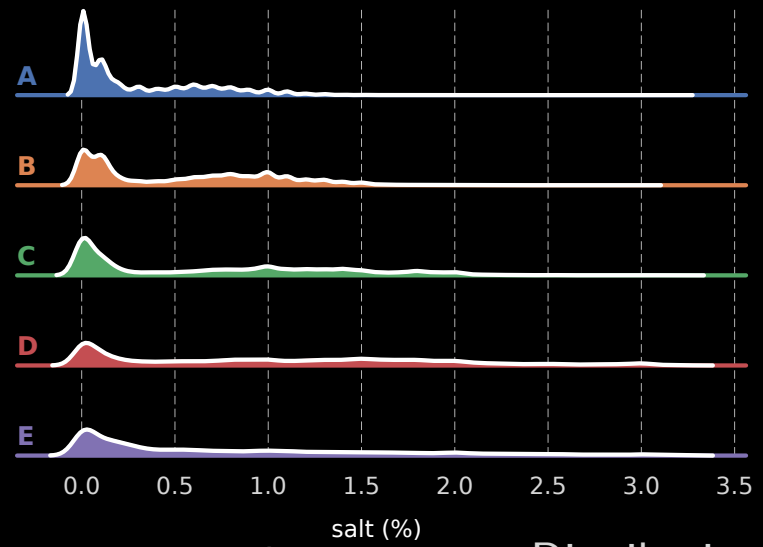
Les moyennes sont **faiblement** corrélées au "grade"

Distributions **multi-modales**

Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



Les groupes sont **différents**

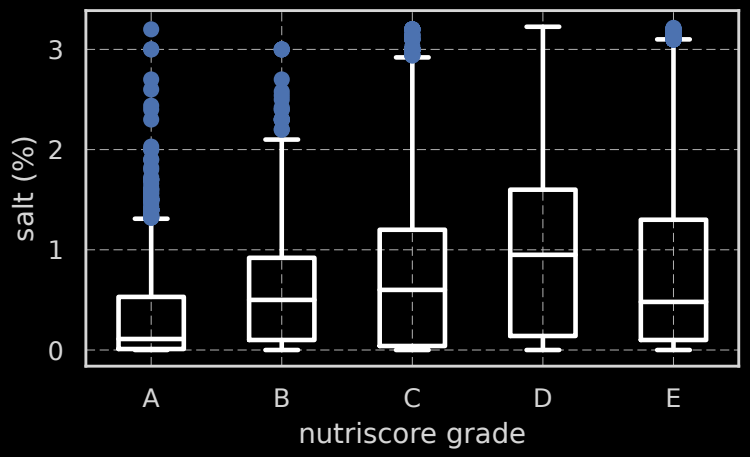


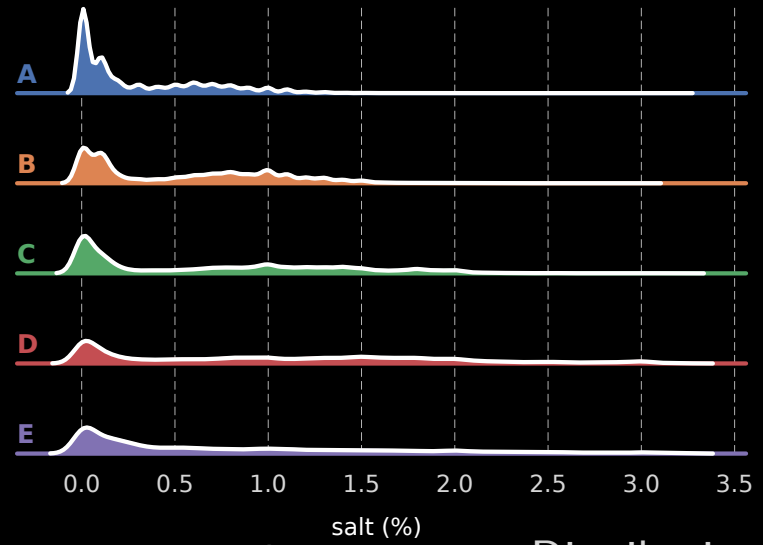
Les moyennes sont **faiblement** corrélées au "grade"

Distributions **multi-modales**

	$\sum y^2$	DF	F	p-value	η^2	ω^2
nutriscore grade	6456.376	4	3247.42	0.0	0.106	0.106
Residual	54512.732	109675				

Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:





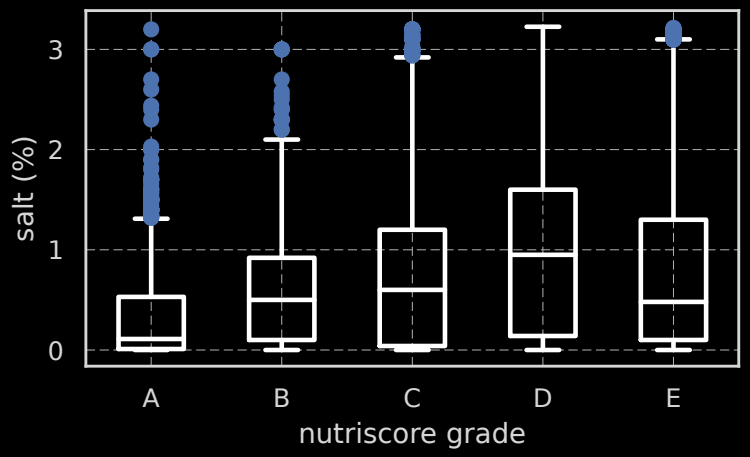
Les groupes sont **différents**

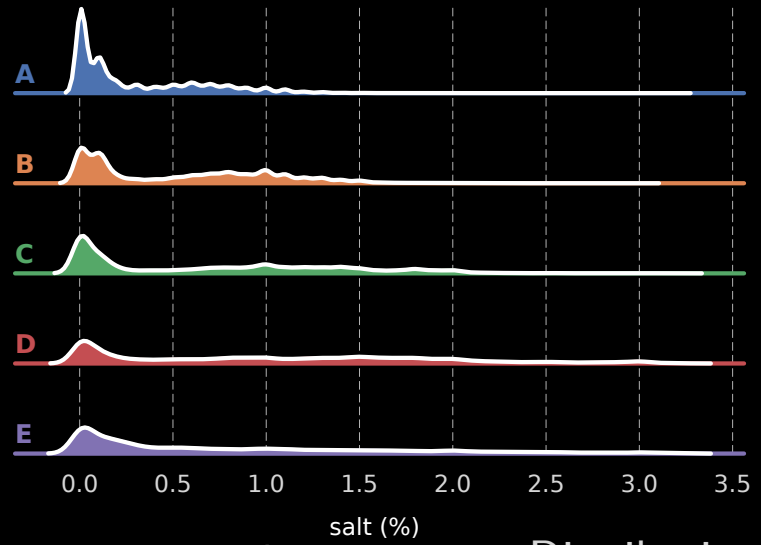
Les moyennes sont **faiblement** corrélées au "grade"

Distributions **multi-modales**

	$\sum y^2$	DF	F	p-value	η^2	ω^2
nutriscore grade	6456.376	4	3247.42	0.0	0.106	0.106
Residual	54512.732	109675				

Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:





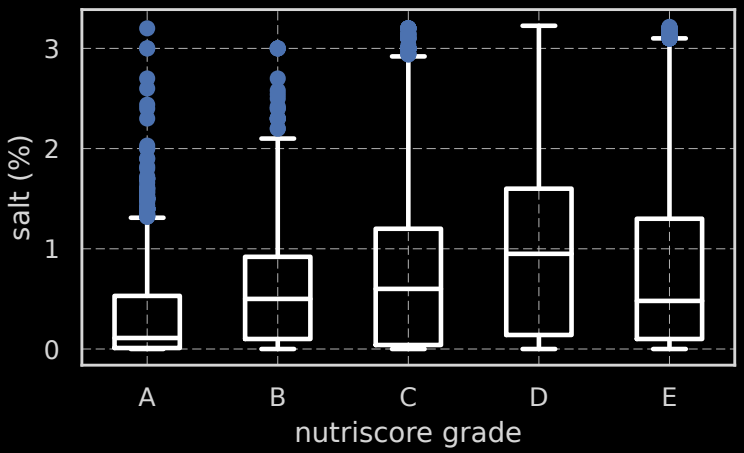
Les groupes sont **différents**

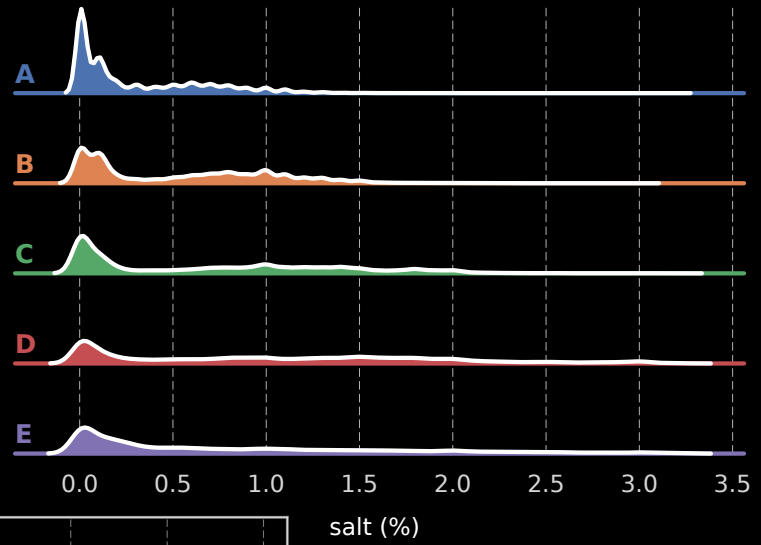
Les moyennes sont **faiblement** corrélées au "grade"

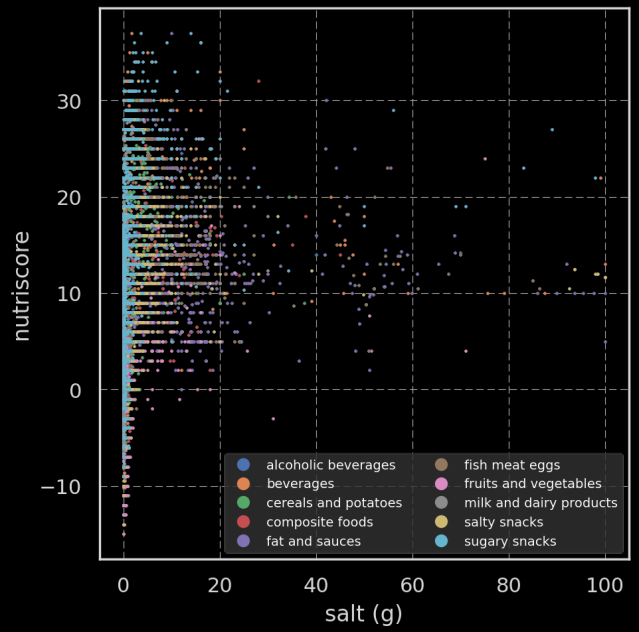
Distributions **multi-modales**

	$\sum y^2$	DF	F	p-value	η^2	ω^2
nutriscore grade	6456.376	4	3247.42	0.0	0.106	0.106
Residual	54512.732	109675				

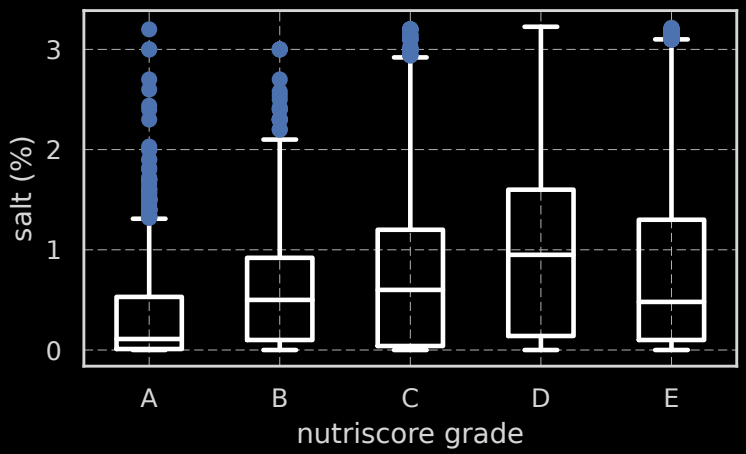
Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:

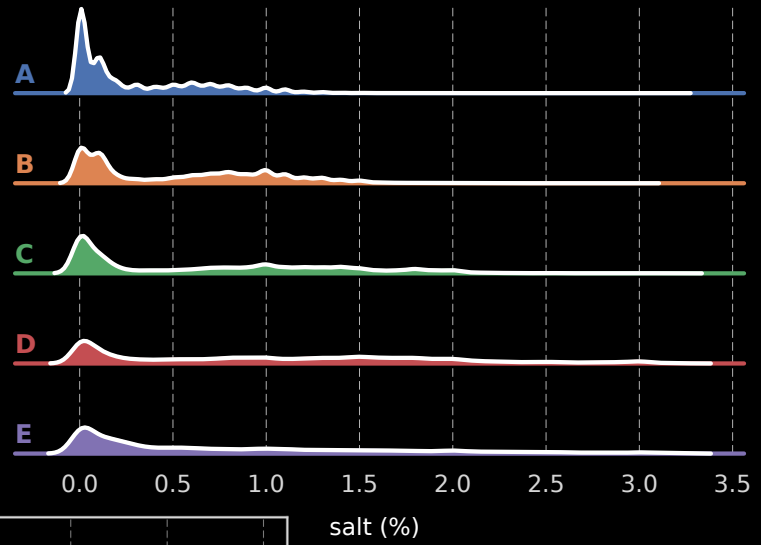


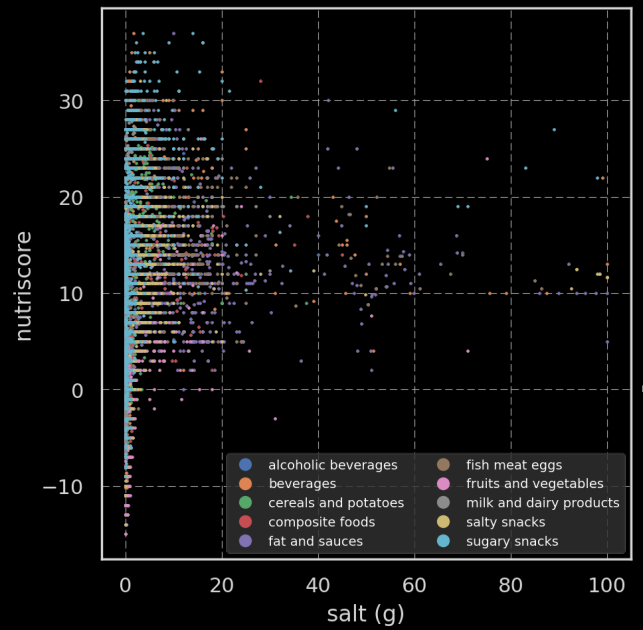




Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



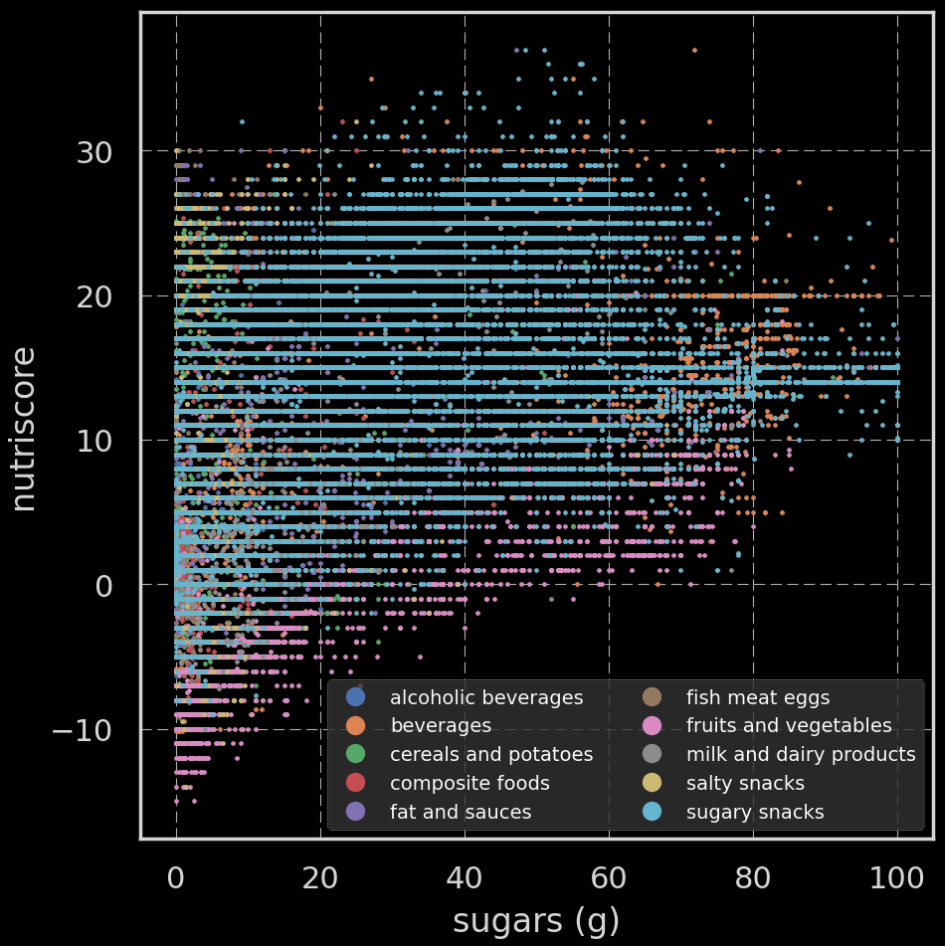




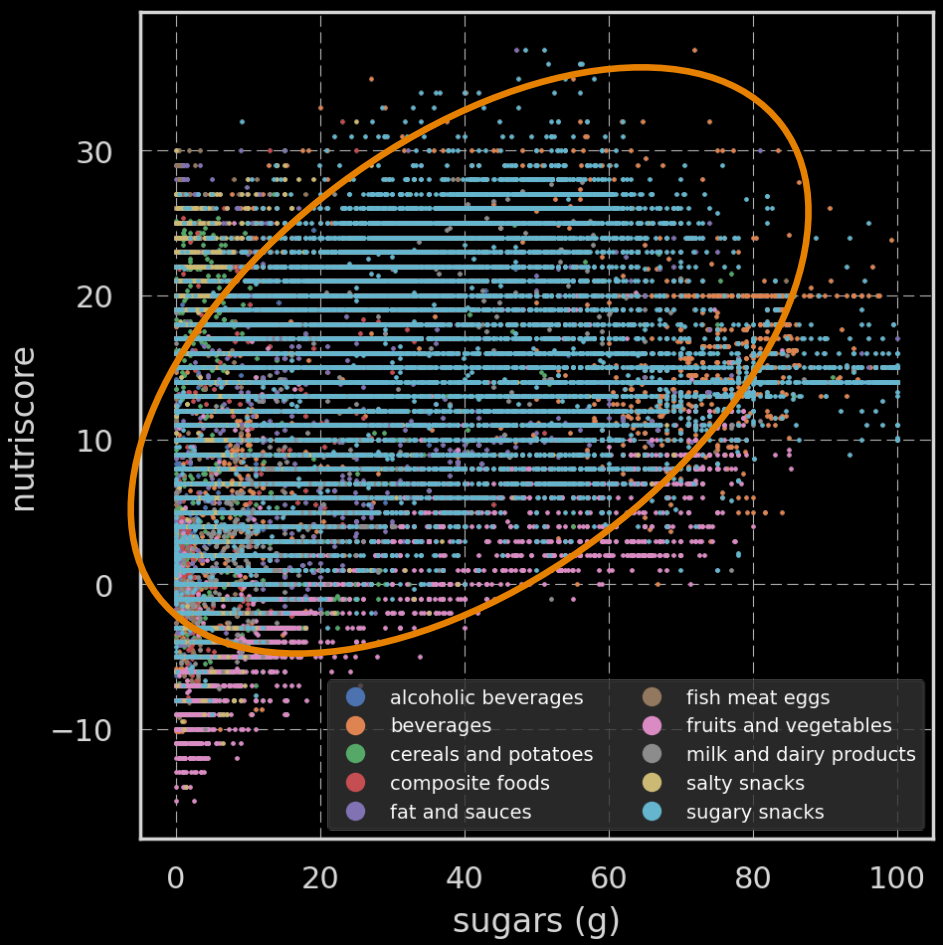
Aucune corrélation apparente

Analyse de la relation **nutriscore** et la **teneur en sucres**:

Analyse de la relation **nutriscore** et la **teneur en sucres**:



Analyse de la relation **nutriscore** et la **teneur en sucres**:



Une **corrélation** est **visible**, mais d'**autres variables** influencent aussi le nutriscore

Peut-on calculer l'indice glycémique ?

Peut-on calculer l'indice glycémique ?

Exemple de listes d'ingrédients:

• lait entier (99%); poudre de lait (1%), ferments lactiques, présure.

• jus d'orange 40% - jus de pomme 40% - jus d'ananas 9% - purée de banane - jus de raisin blanc - jus de pamplemousse - purée d'abricot - purée de pêche.

• 1 boite de garniture: tomates fraiches (51%), eau, oignons frais, huile d'olive vierge extra, huile de colza, sel, persil, concentré de tomate, jus concentré de citron, menthe, épaississants: farine de graines de caroube et gomme guar, arômes. 1 coupelle de semoule de blé dur précuite à la vapeur.

• farine de blé, sucre, beurre 12% (lait), sirop de sucre inverti, poudres à lever: carbonates de sodium, diphosphates, lactosérum en poudre lait), lat entier en poudre, sel, émulsifiant: lécithine de soja; acidifiant : acide citrique; arôme (oeufs entiers en poudre).

• palette de porc avec os 90 %, eau, sel, dextrose, conservateur : nitrite de sodium.

Peut-on calculer l'indice glycémique ?

Exemple de listes d'ingrédients:

- lait entier (99%); poudre de lait (1%), ferments lactiques, présure.

- jus d'orange 40% - jus de pomme 40% - jus d'ananas 9% - purée de banane - jus de raisin blanc - jus de pamplemousse - purée d'abricot - purée de pêche.

- 1 boite de garniture: tomates fraiches (51%), eau, oignons frais, huile d'olive vierge extra, huile de colza, sel, persil, concentré de tomate, jus concentré de citron, menthe, épaississants: farine de graines de caroube et gomme guar, arômes. 1 coupelle de semoule de blé dur précuite à la vapeur.

- farine de blé, sucre, beurre 12% (lait), sirop de sucre inverti, poudres à lever: carbonates de sodium, diphosphates, lactosérum en poudre lait), lat entier en poudre, sel, émulsifiant: lécithine de soja; acidifiant : acide citrique; arôme (oeufs entiers en poudre).

- palette de porc avec os 90 %, eau, sel, dextrose, conservateur : nitrite de sodium.

Non, car les % sont assez peu renseignés

Peut-on calculer l'indice glycémique ?

Exemple de listes d'ingrédients:

• lait entier (99%); poudre de lait (1%), ferments lactiques, présure.

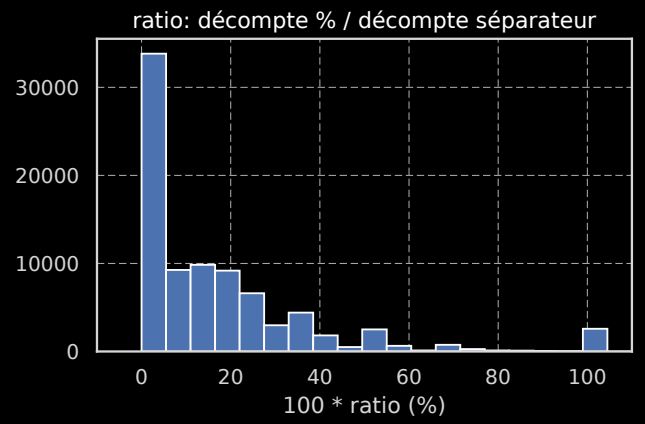
• jus d'orange 40% - jus de pomme 40% - jus d'ananas 9% - purée de banane - jus de raisin blanc - jus de pamplemousse - purée d'abricot - purée de pêche.

• 1 boite de garniture: tomates fraiches (51%), eau, oignons frais, huile d'olive vierge extra, huile de colza, sel, persil, concentré de tomate, jus concentré de citron, menthe, épaississants: farine de graines de caroube et gomme guar, arômes. 1 coupelle de semoule de blé dur précuite à la vapeur.

• farine de blé, sucre, beurre 12% (lait), sirop de sucre inverti, poudres à lever: carbonates de sodium, diphosphates, lactosérum en poudre lait), lat entier en poudre, sel, émulsifiant: lécithine de soja; acidifiant : acide citrique; arôme (oeufs entiers en poudre).

• palette de porc avec os 90 %, eau, sel, dextrose, conservateur : nitrite de sodium.

Non, car les % sont assez peu renseignés



CONCLUSIONS

Résultats de la faisabilité des applications:

Résultats de la faisabilité des applications:

Données récupérables:

Résultats de la faisabilité des applications:

Données récupérables:

- Teneur en sel
- Charge calorique
- Présence d'ingrédients dont allergènes
- Pourcentage de macro-nutriments notamment graisses et sucres

Résultats de la faisabilité des applications:

Données récupérables:

Applications:

- Teneur en sel
- Charge calorique
- Présence d'ingrédients dont allergènes
- Pourcentage de macro-nutriments notamment graisses et sucres

Résultats de la faisabilité des applications:

Données récupérables:

- Teneur en sel
- Charge calorique
- Présence d'ingrédients dont allergènes
- Pourcentage de macro-nutriments notamment graisses et sucres

Applications:

- Régime pauvre en sel
- Gestion de poids/IMC
- Régime sans/avec ...

Résultats de la faisabilité des applications:

Données récupérables:

- Teneur en sel
- Charge calorique
- Présence d'ingrédients dont allergènes
- Pourcentage de macro-nutriments notamment graisses et sucres

Applications:

- Régime pauvre en sel
- Gestion de poids/IMC
- Régime sans/avec ...
- Gestion de l'indice glycémique
- Gestion équilibre acido-basique

Résultats de la faisabilité des applications:

Données récupérables:	Applications:
Teneur en sel	Régime pauvre en sel
Charge calorique	Gestion de poids/IMC
Présence d'ingrédients dont allergènes	Régime sans/avec ...
Pourcentage de macro-nutriments notamment graisses et sucres	Gestion de l'indice glycémique
	Gestion équilibre acido-basique

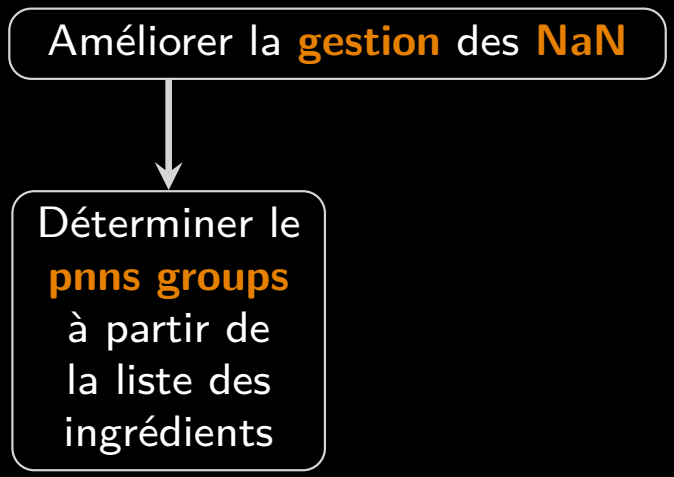
Se rapprocher de **spécialistes** du domaine
est **nécessaire** pour apporter une
meilleure **connaissance métier**
(identifier/valider les contraintes
alimentaires liées aux états de santé)

Pistes d'améliorations des traitements:

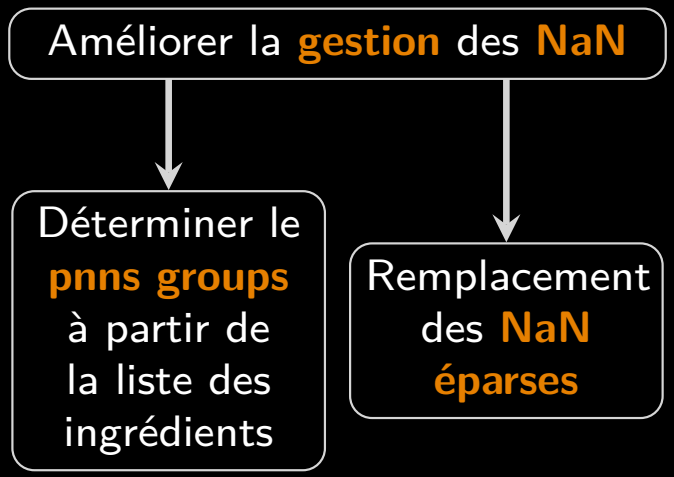
Pistes d'améliorations des traitements:

Améliorer la **gestion** des **NaN**

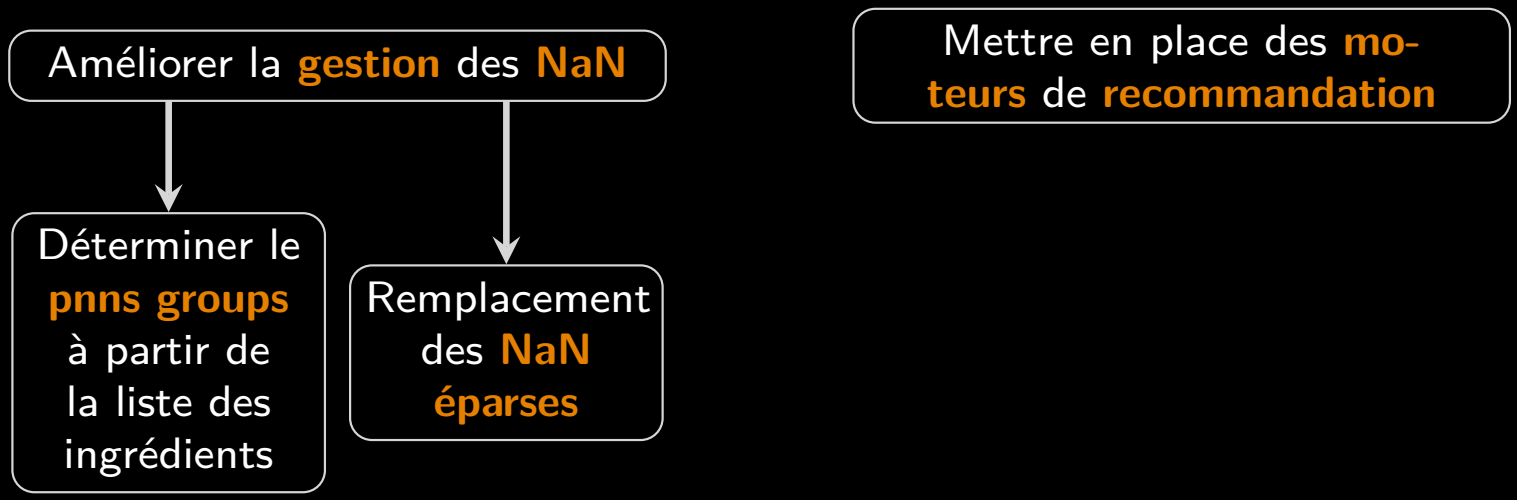
Pistes d'améliorations des traitements:



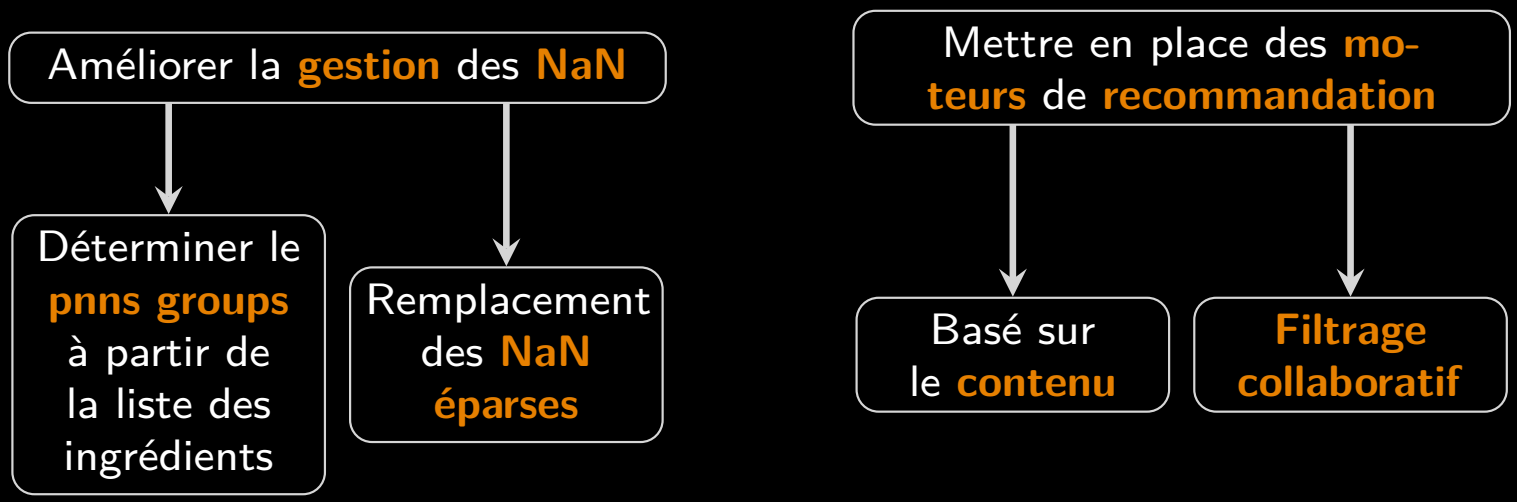
Pistes d'améliorations des traitements:



Pistes d'améliorations des traitements:



Pistes d'améliorations des traitements:



Merci pour votre **attention**.

knn

Principe:

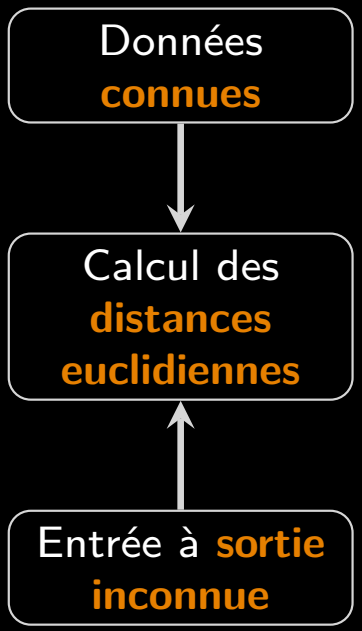
Données
connues

Principe:

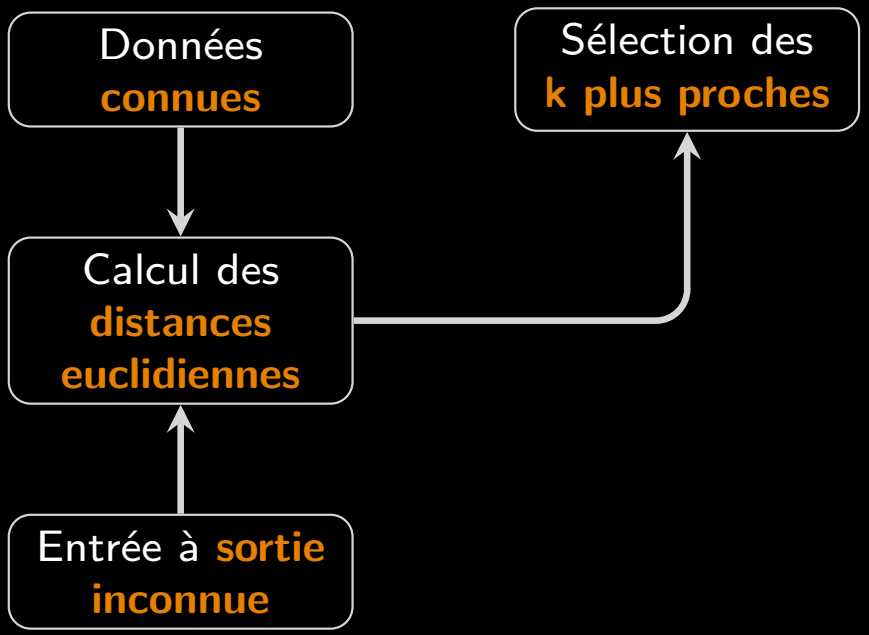
Données **connues**

Entrée à **sortie inconnue**

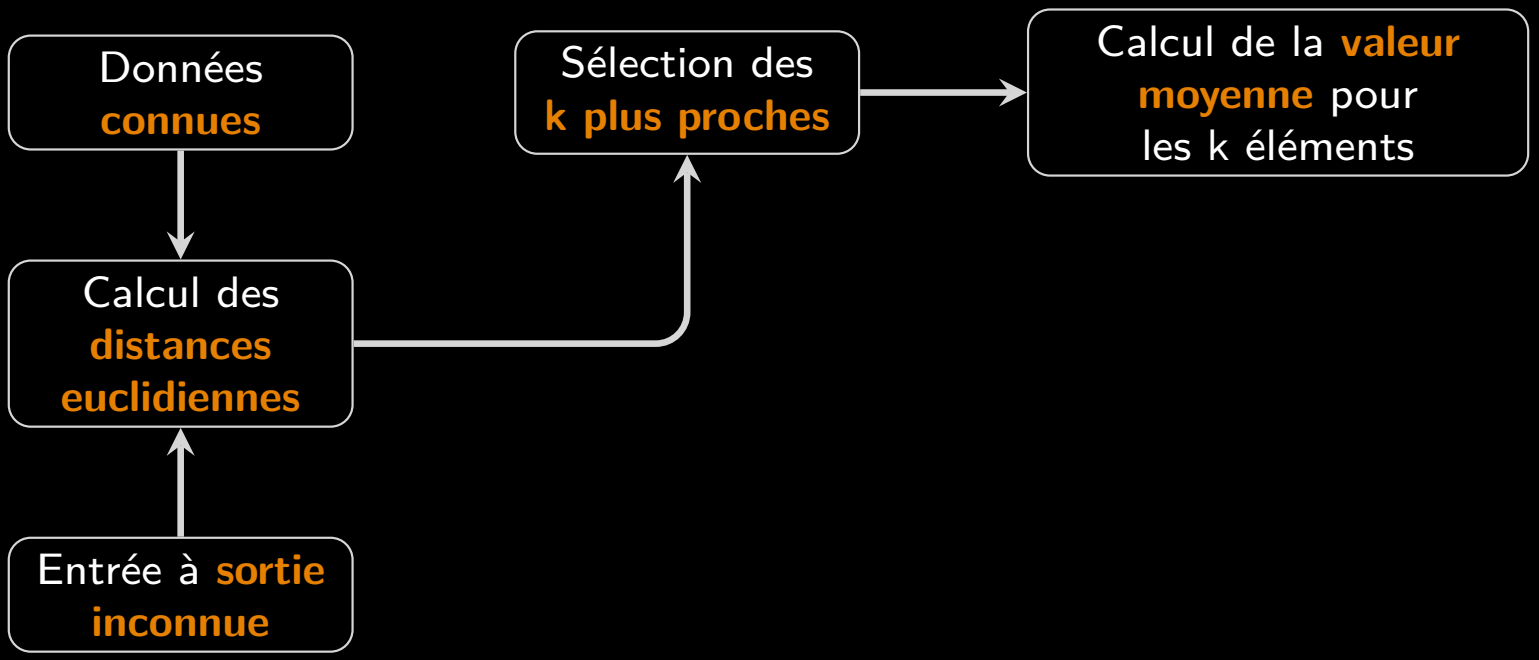
Principe:


Données
connues
Calcul des
distances
euclidiennes
Entrée à sortie
inconnue

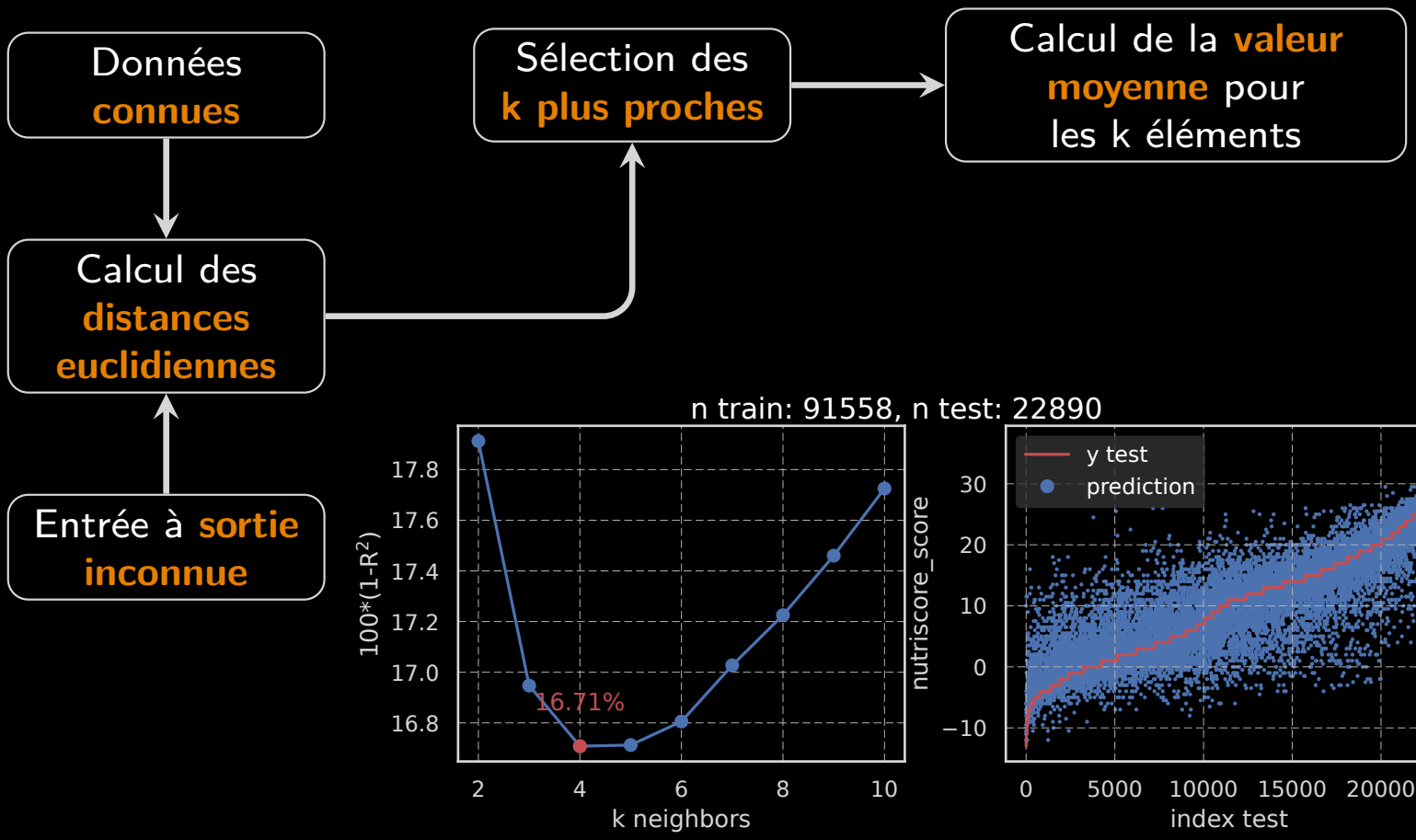
Principe:


Données
connues
Sélection des
k plus proches
Calcul des
distances
euclidiennes
Entrée à sortie
inconnue

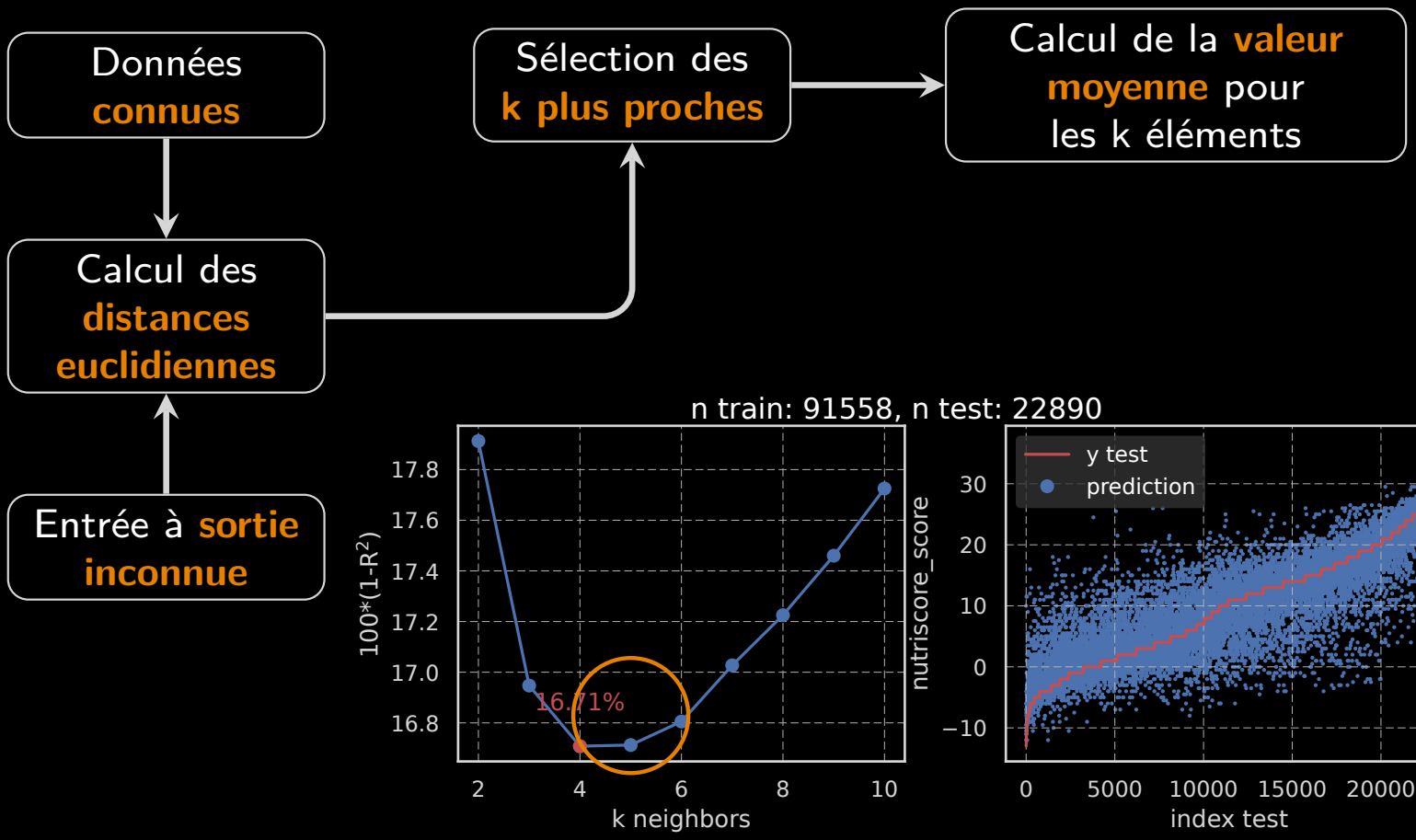
Principe:


Données
connues
Calcul des
distances
euclidiennes
Entrée à sortie
inconnue
Sélection des
k plus proches
Calcul de la valeur
moyenne pour
les k éléments

Principe:


Données connues
Calcul des distances euclidiennes
Entrée à sortie inconnue
Sélection des k plus proches
Calcul de la valeur moyenne pour les k éléments
n train: 91558, n test: 22890
100*(1-R²)
17.8
17.6
17.4
17.2
17.0
16.8
16.71%
2
4
6
8
10
k neighbors
nutriscore_score
y test
prediction
30
20
10
0
−10
0
5000
10000
15000
20000
index test

Principe:

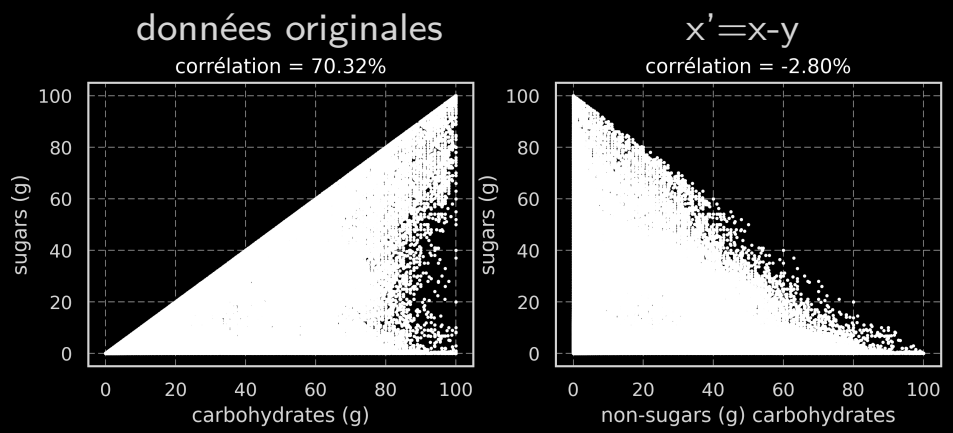


Avec les **données brutes**, **17%** d'erreur → nécessité d'optimiser

Analyse de la **corrélation** entre les variables:

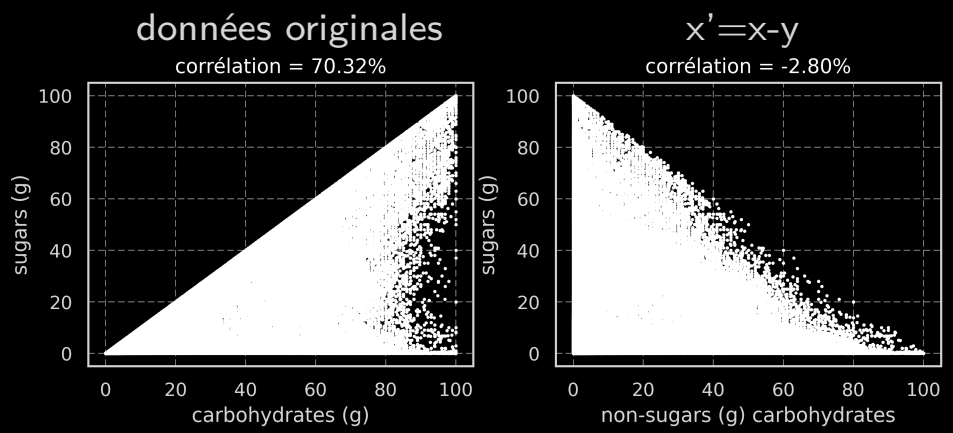
Analyse de la **corrélation** entre les variables:

carbohydrates / sugars



Analyse de la **corrélation** entre les variables:

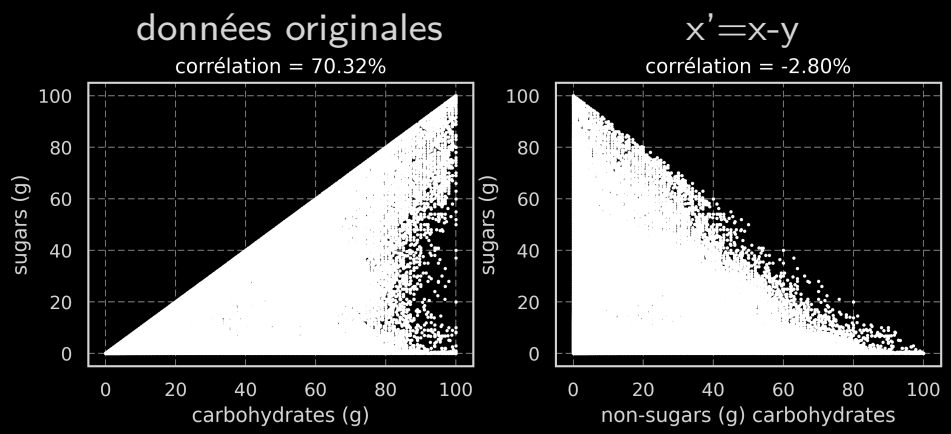
carbohydrates / sugars



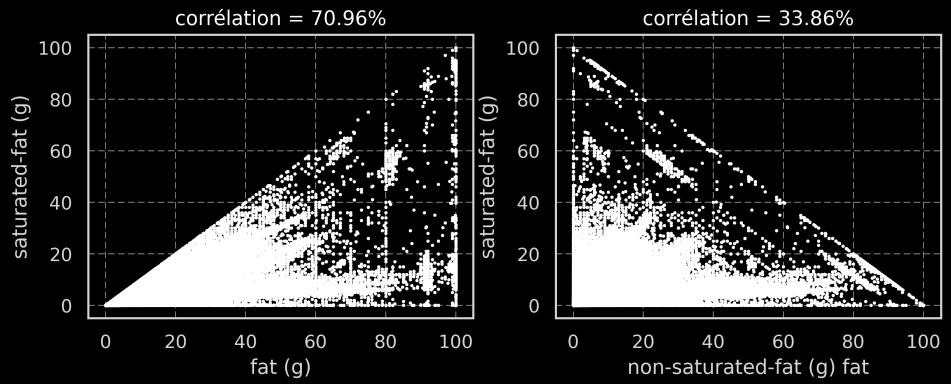
Décorrélation **OK**

Analyse de la **corrélation** entre les variables:

carbohydrates / sugars



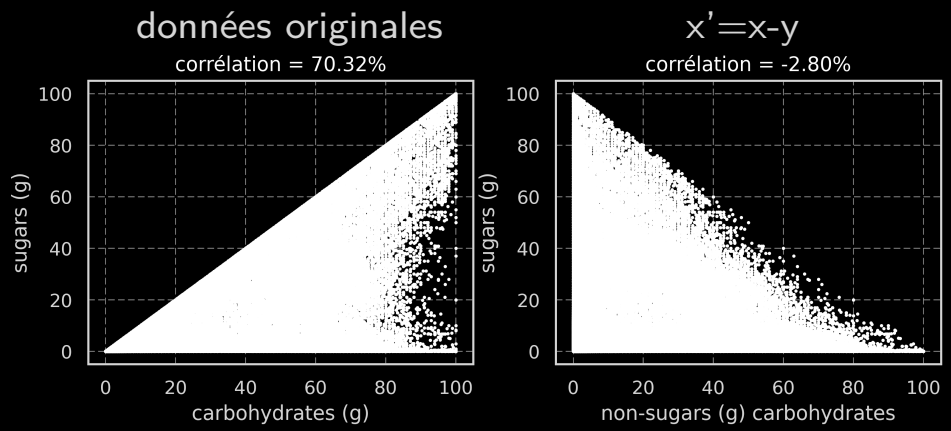
fat / saturated-fat



Décorrélation **OK**

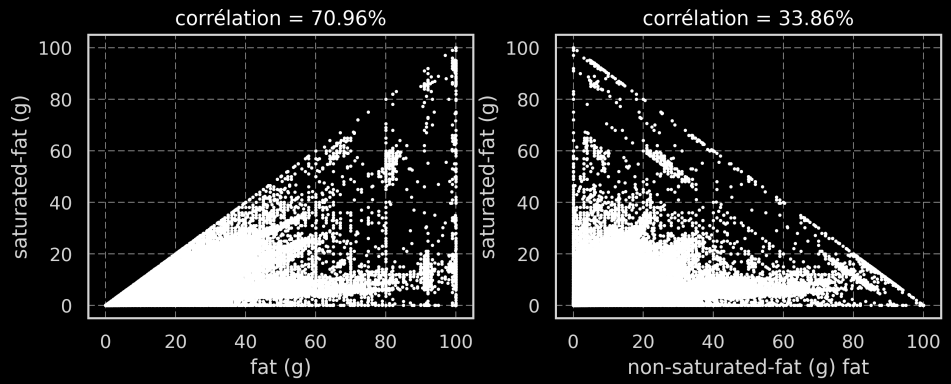
Analyse de la **corrélation** entre les variables:

carbohydrates / sugars



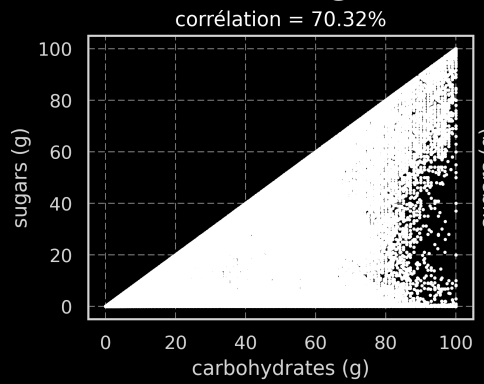
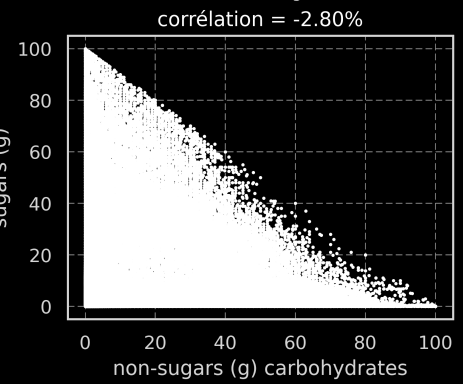
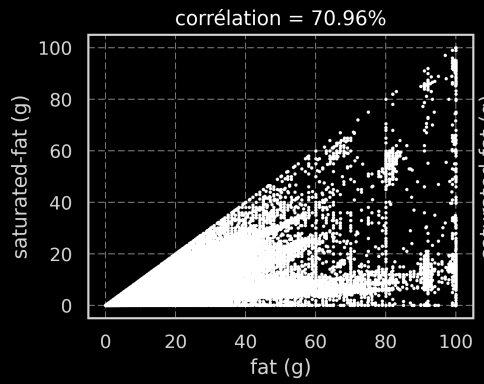
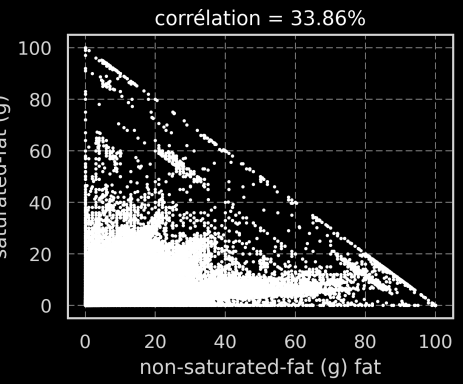
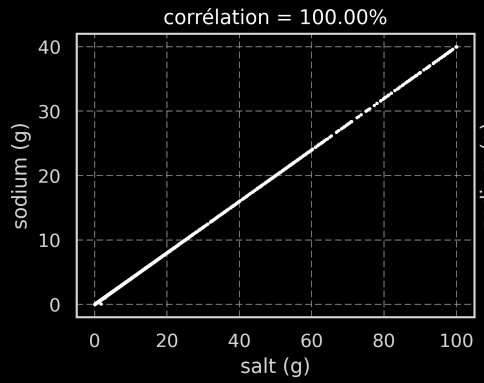
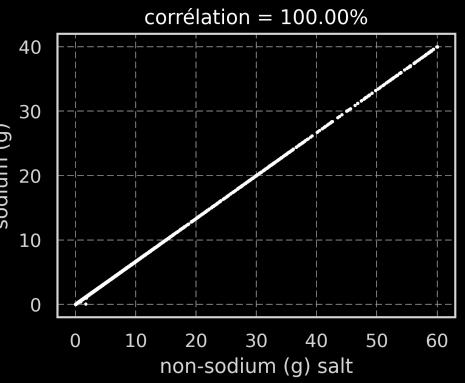
Décorrélation **OK**

fat / saturated-fat

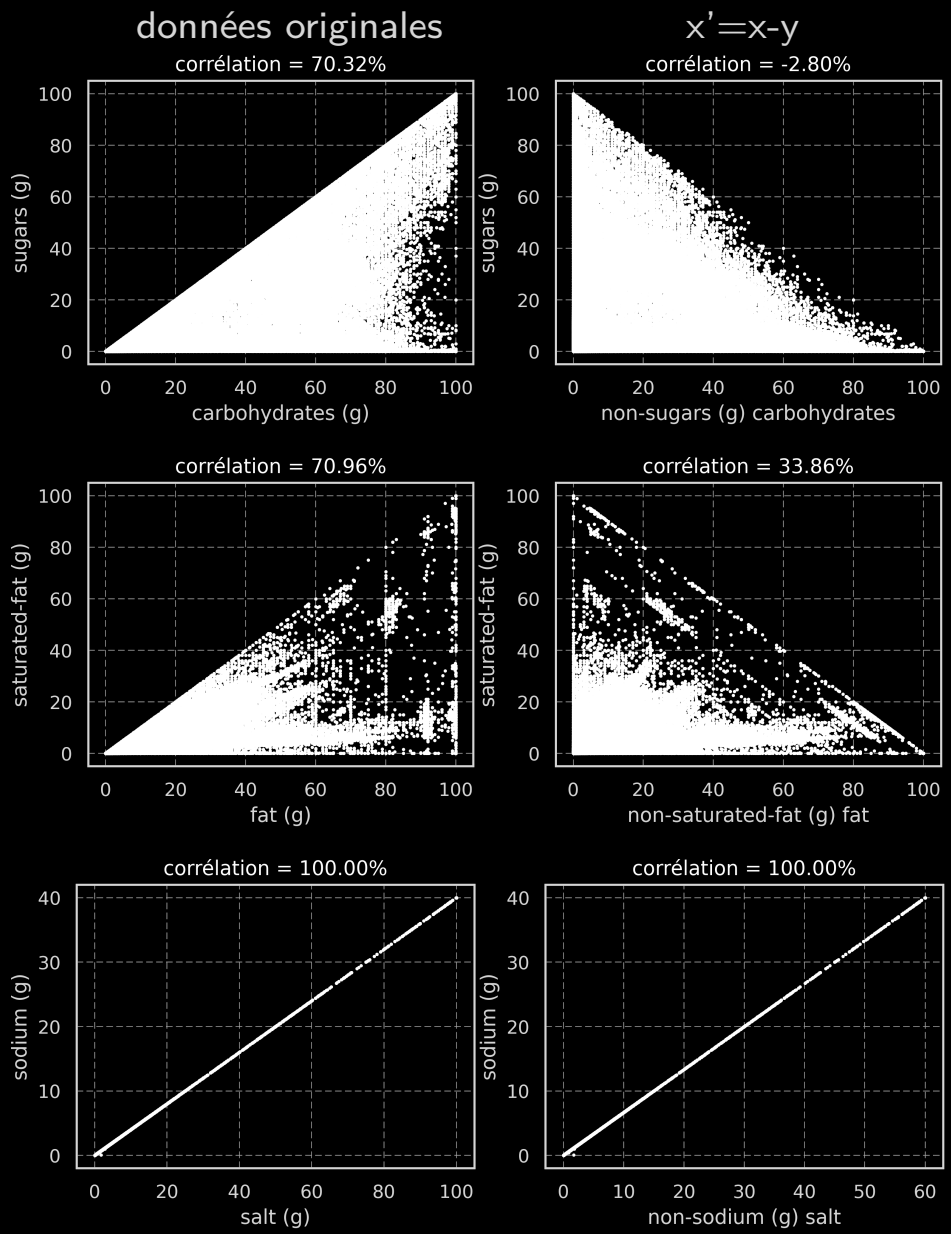


Décorrélation **partielle**

Analyse de la **corrélation** entre les variables:

	données originales	x'=x-y	
carbohydrates / sugars			Décorrélation **OK**
fat / saturated-fat			Décorrélation **partielle**
sodium / fat			

Analyse de la **corrélation** entre les variables:



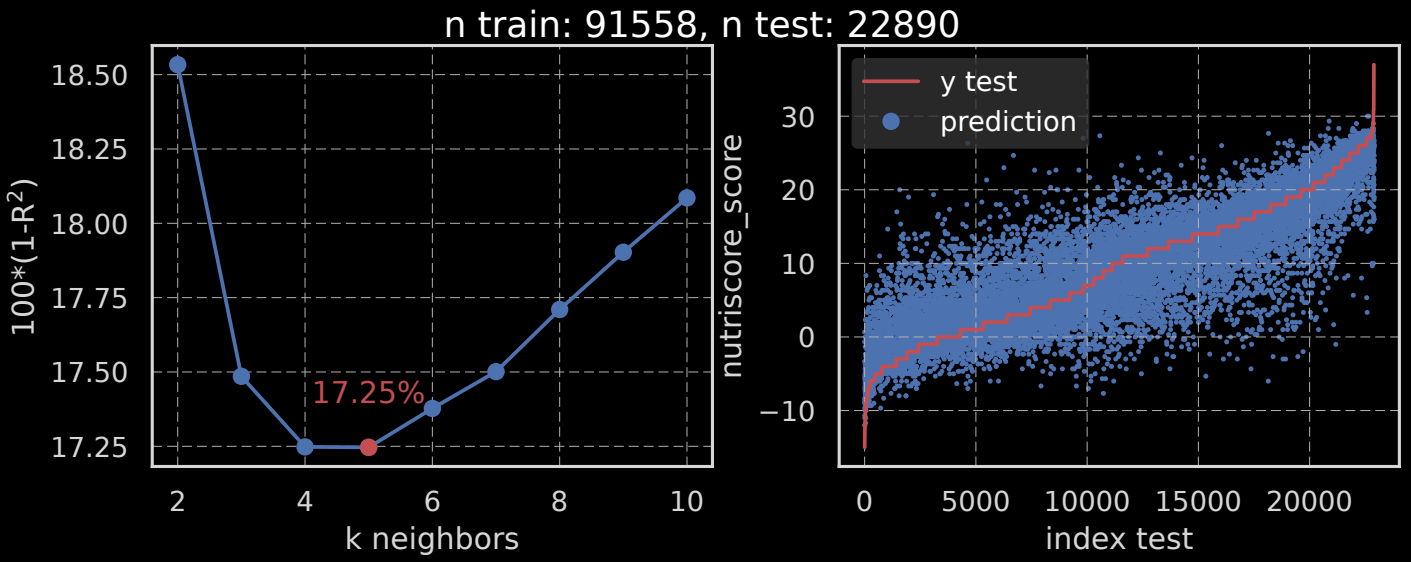
Décorrélation **OK**

Décorrélation **partielle**

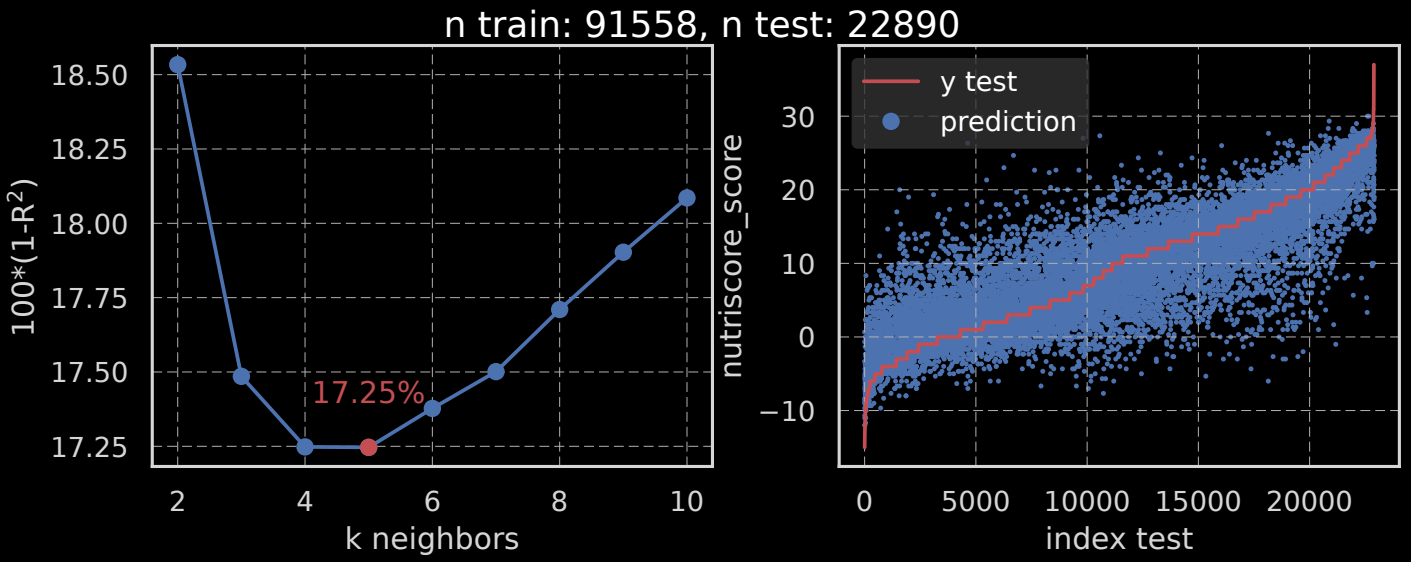
Décorrélation **nulle** → **suppression** d'une variable

Prédiction en utilisant la **base améliorée**:

Prédiction en utilisant la **base améliorée**:



Prédiction en utilisant la **base améliorée**:



→ Amélioration **non significative**

→ **Centrage** et **mise à l'échelle** des données

Centrage et mise à l'échelle des données (par variable):

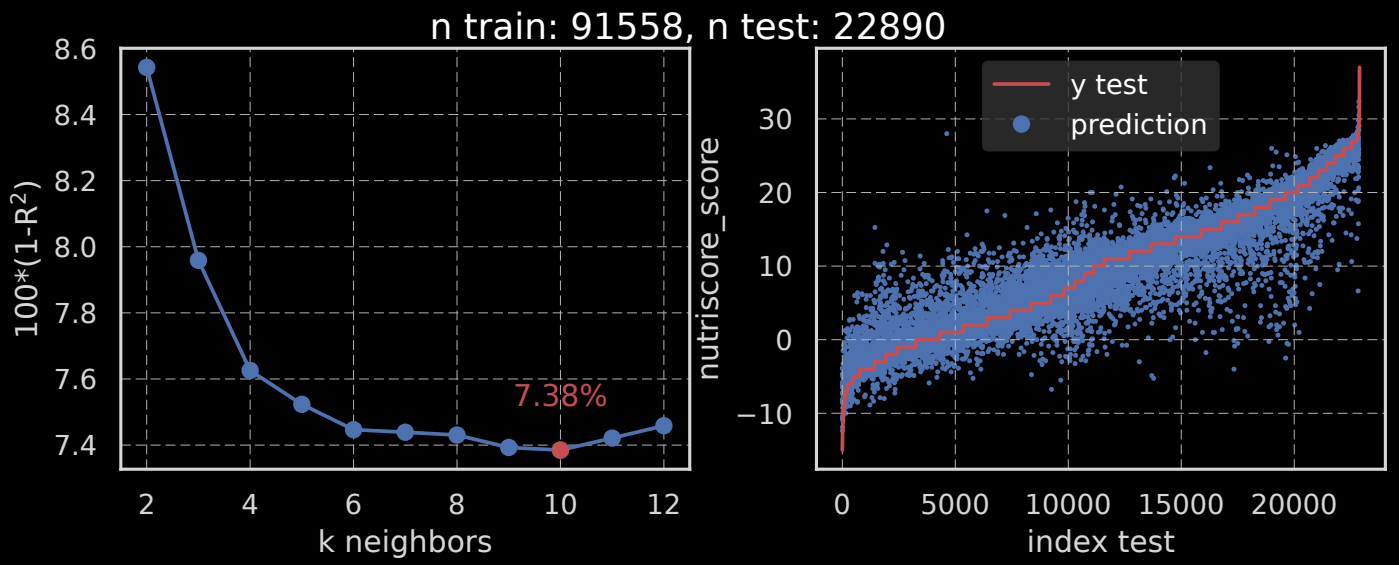
Centrage et mise à l'échelle des données (par variable):

Soustraction de
la valeur moyenne

Centrage et mise à l'échelle des données (par variable):

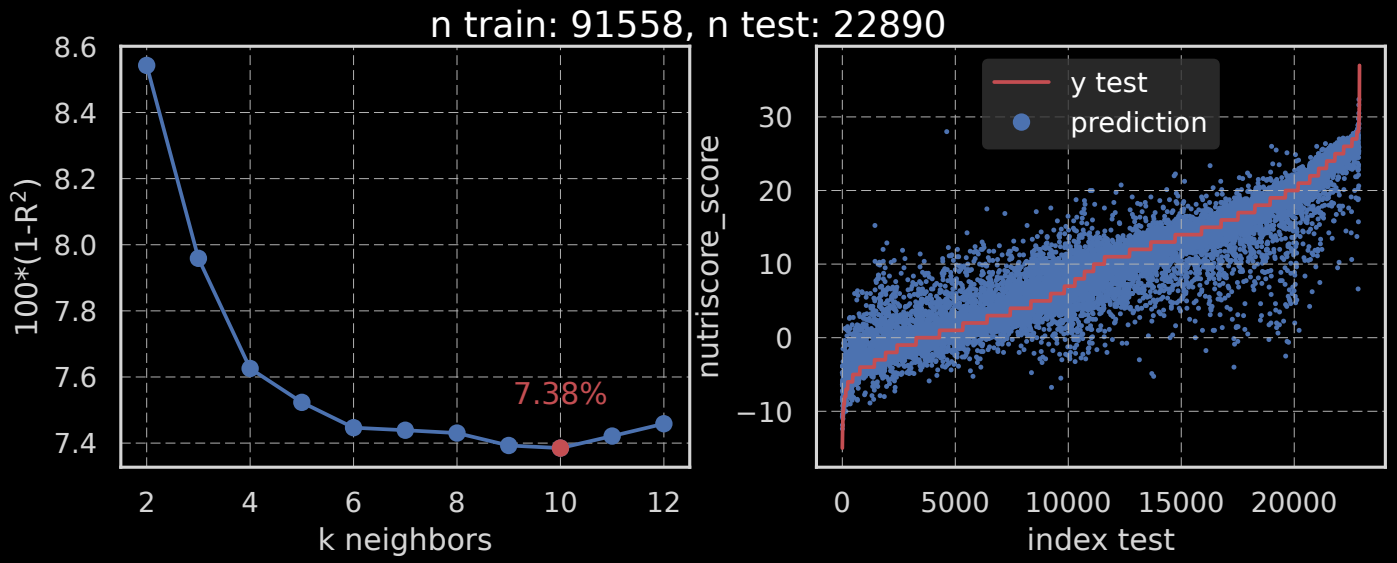
Centrage et mise à l'échelle des données (par variable):

Soustraction de la valeur moyenne → **Normalisation de la variance**



Centrage et mise à l'échelle des données (par variable):

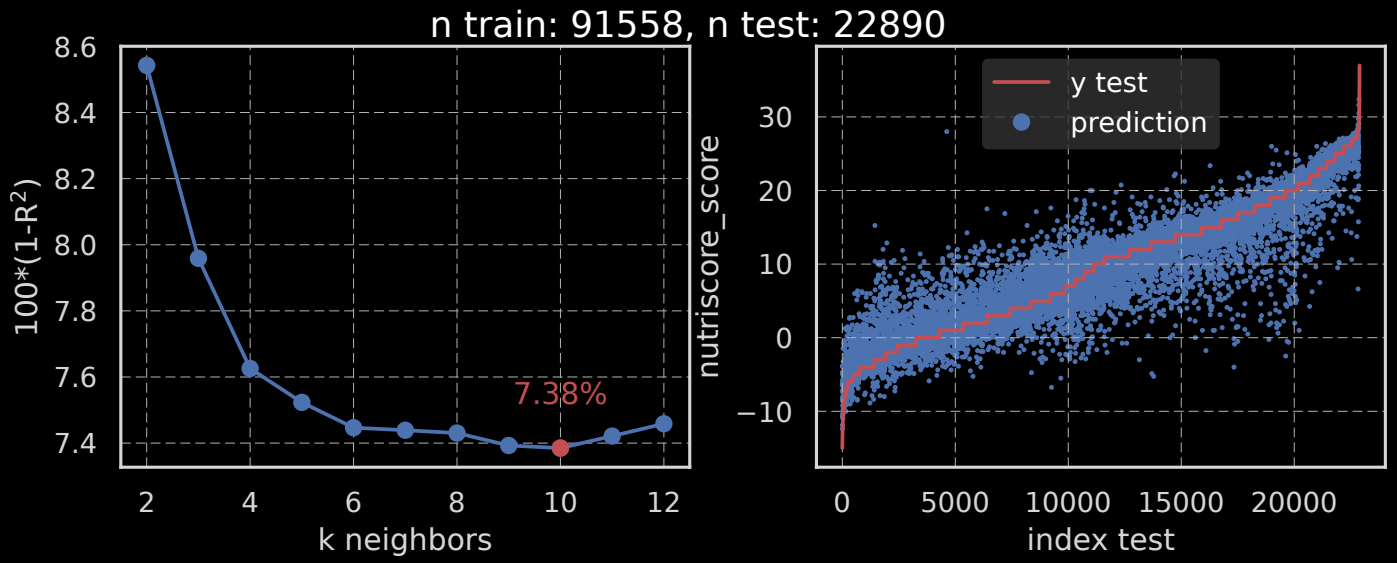
Soustraction de la valeur moyenne → **Normalisation de la variance**



→ **Nette** amélioration

Centrage et mise à l'échelle des données (par variable):

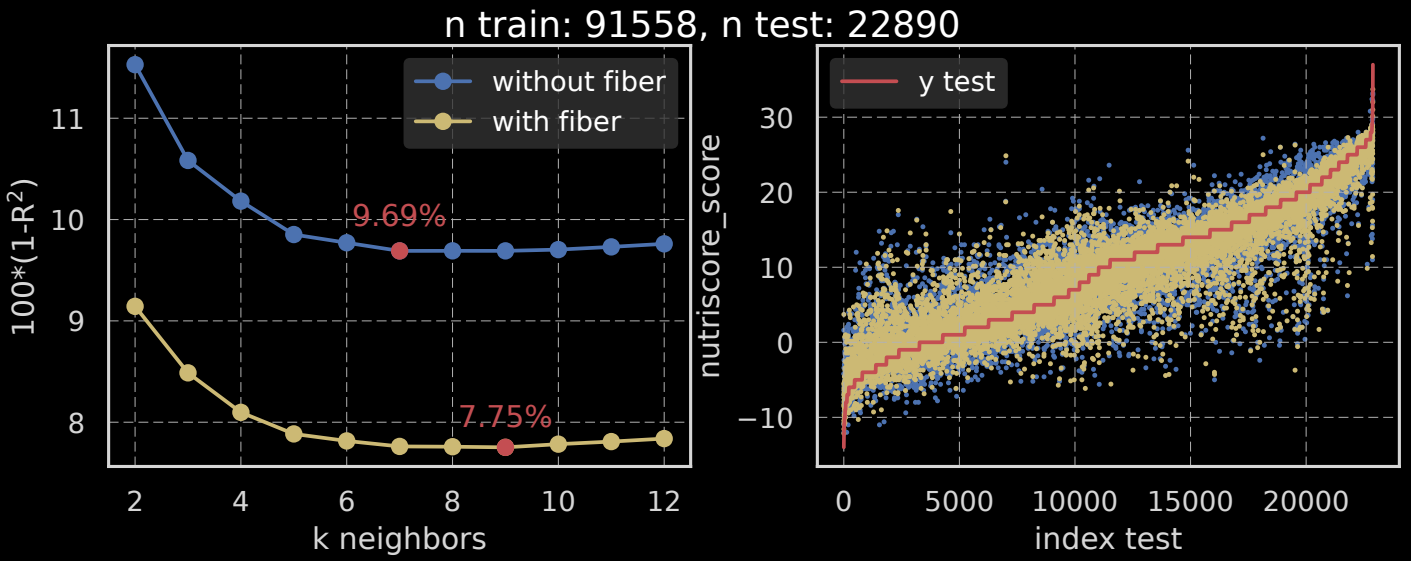
Soustraction de la valeur moyenne → **Normalisation de la variance**



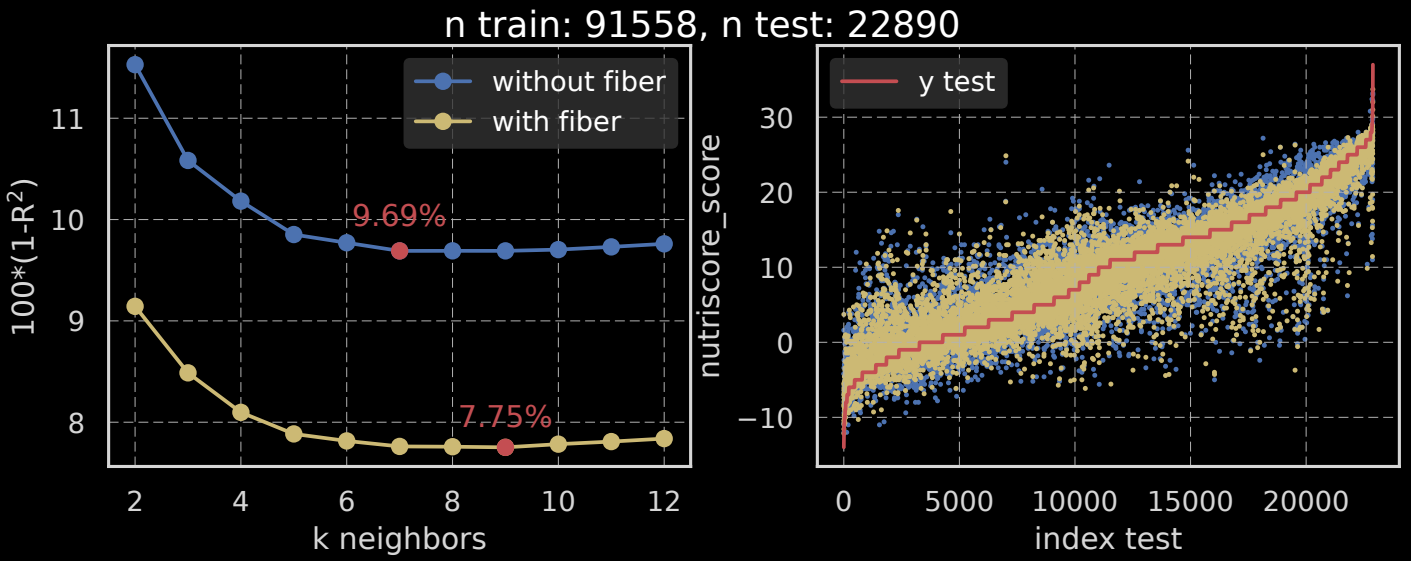
→ **Nette** amélioration

Quid de l'**hypothèse** NaN fibers = 0 ?

Vérification de l'hypothèse NaN fiber = 0:

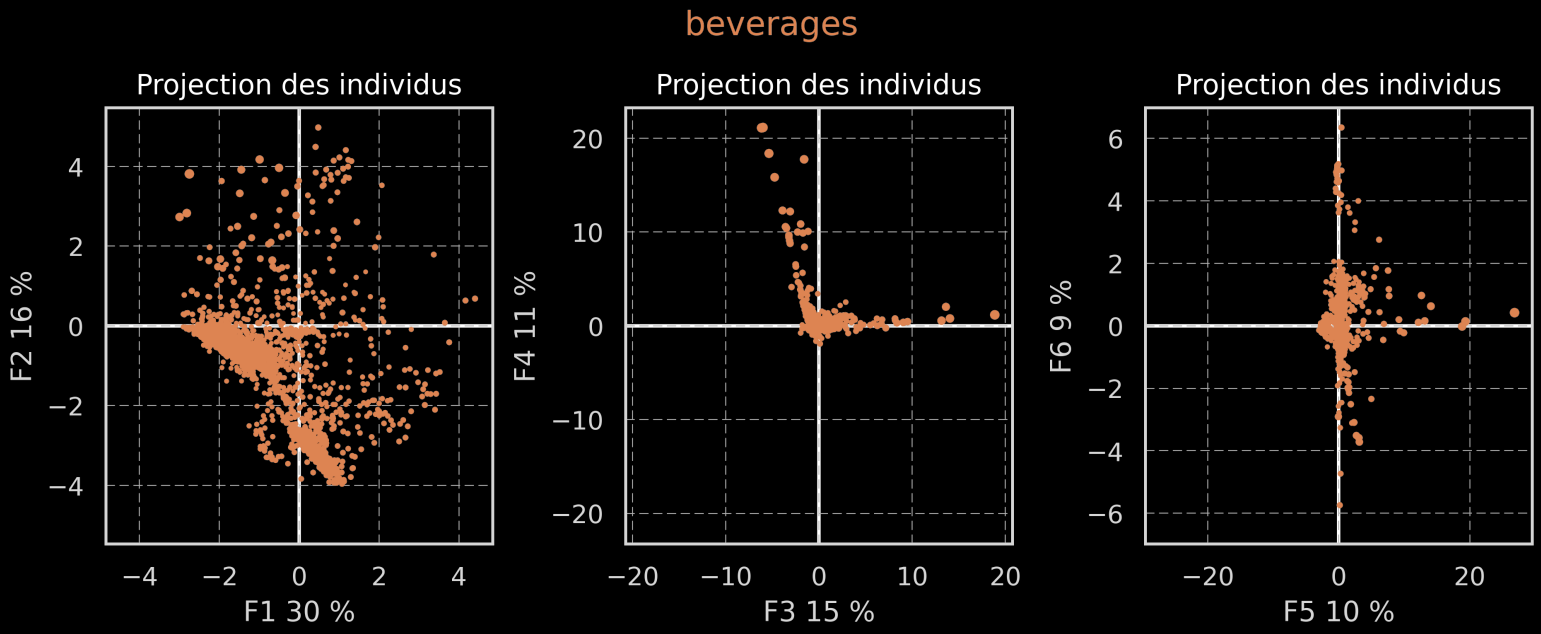


Vérification de l'hypothèse NaN fiber = 0:

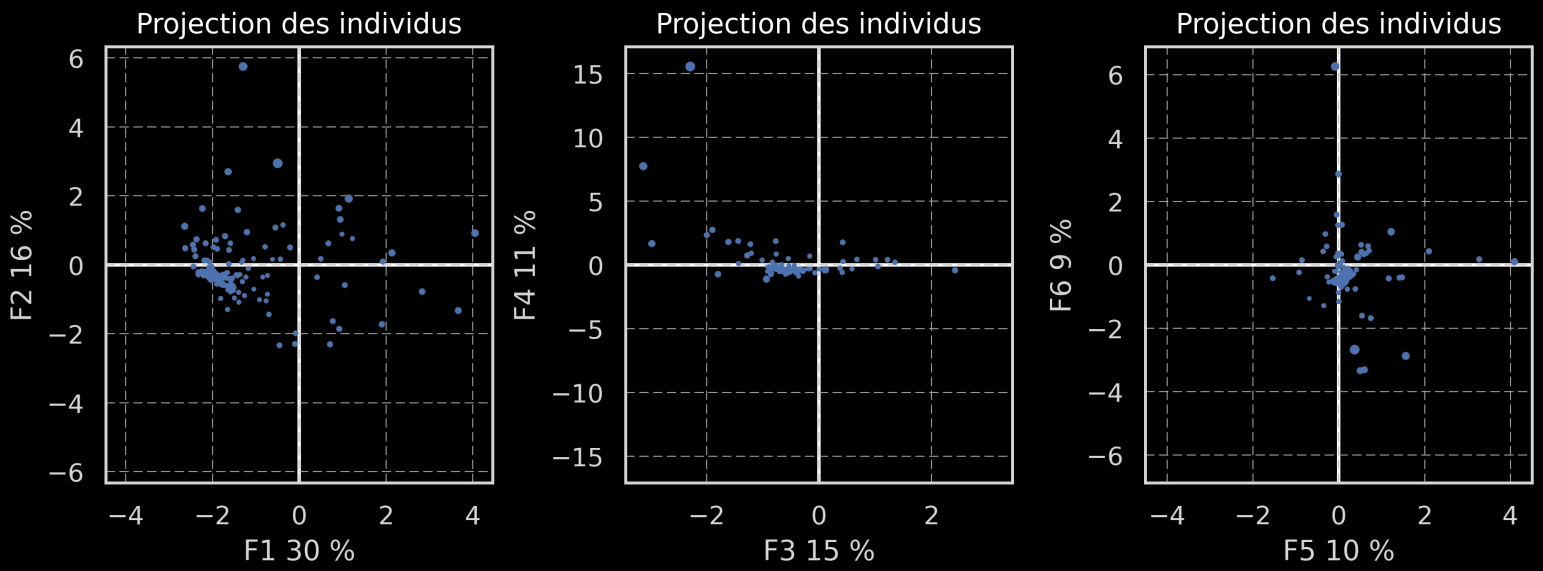


→ Hypothèse à priori **validée**

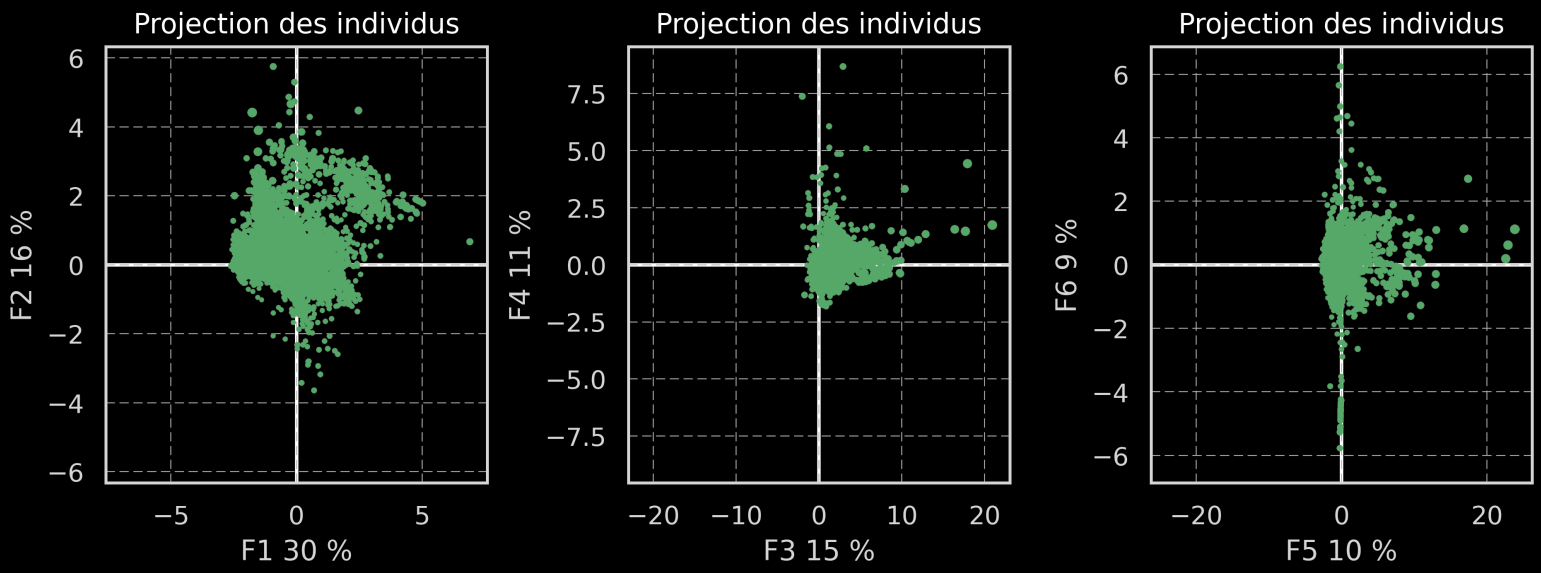
PCA - projection


beverages
Projection des individus
F2 16 %
F1 30 %
Projection des individus
F4 11 %
F3 15 %
Projection des individus
F6 9 %
F5 10 %

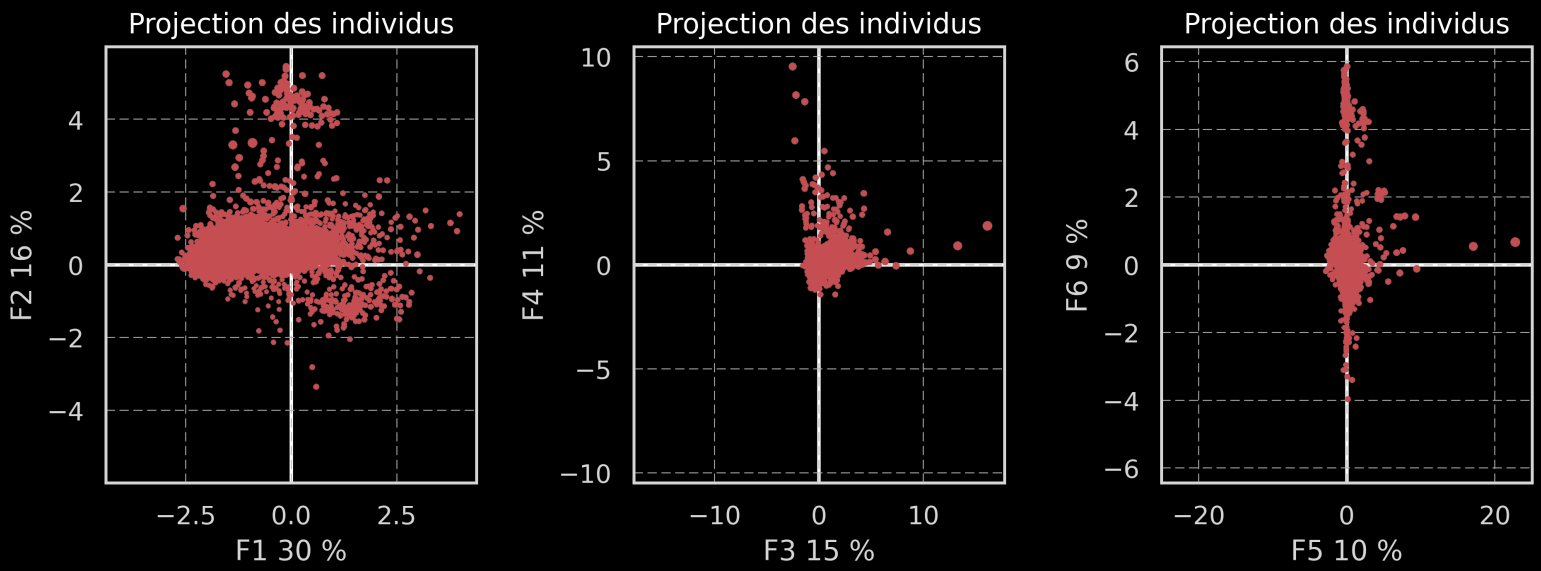
alcoholic beverages



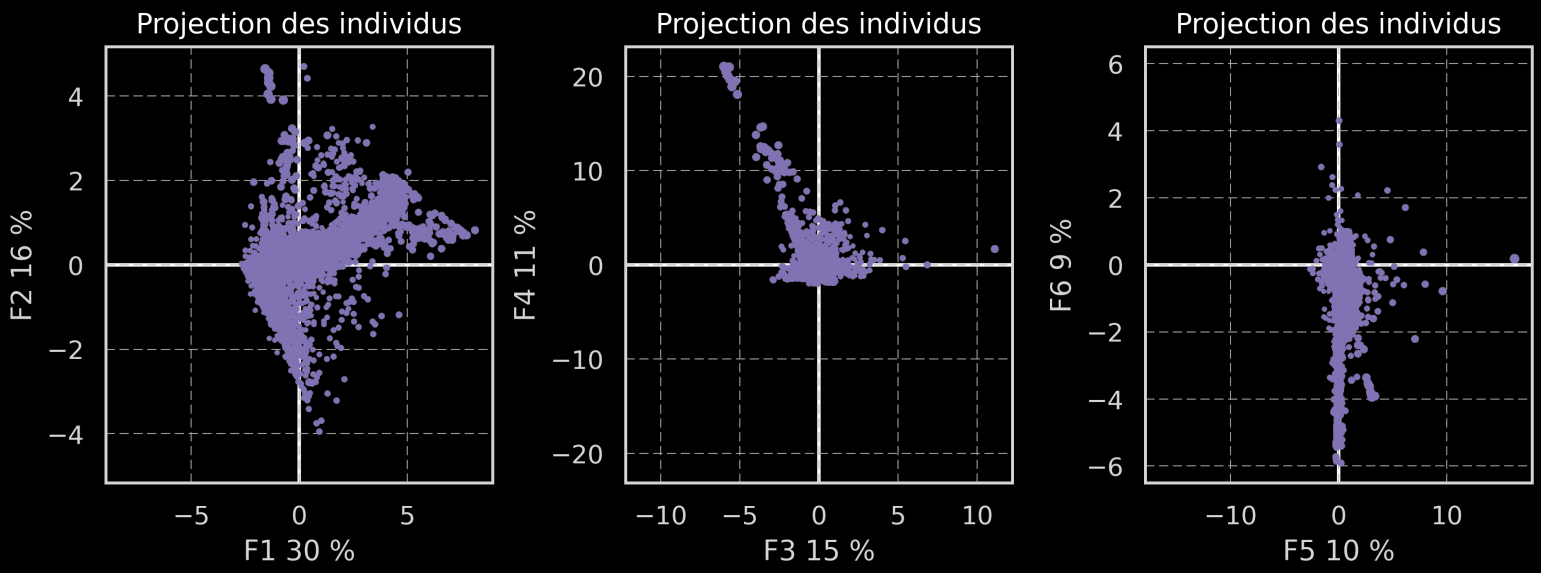
cereals and potatoes

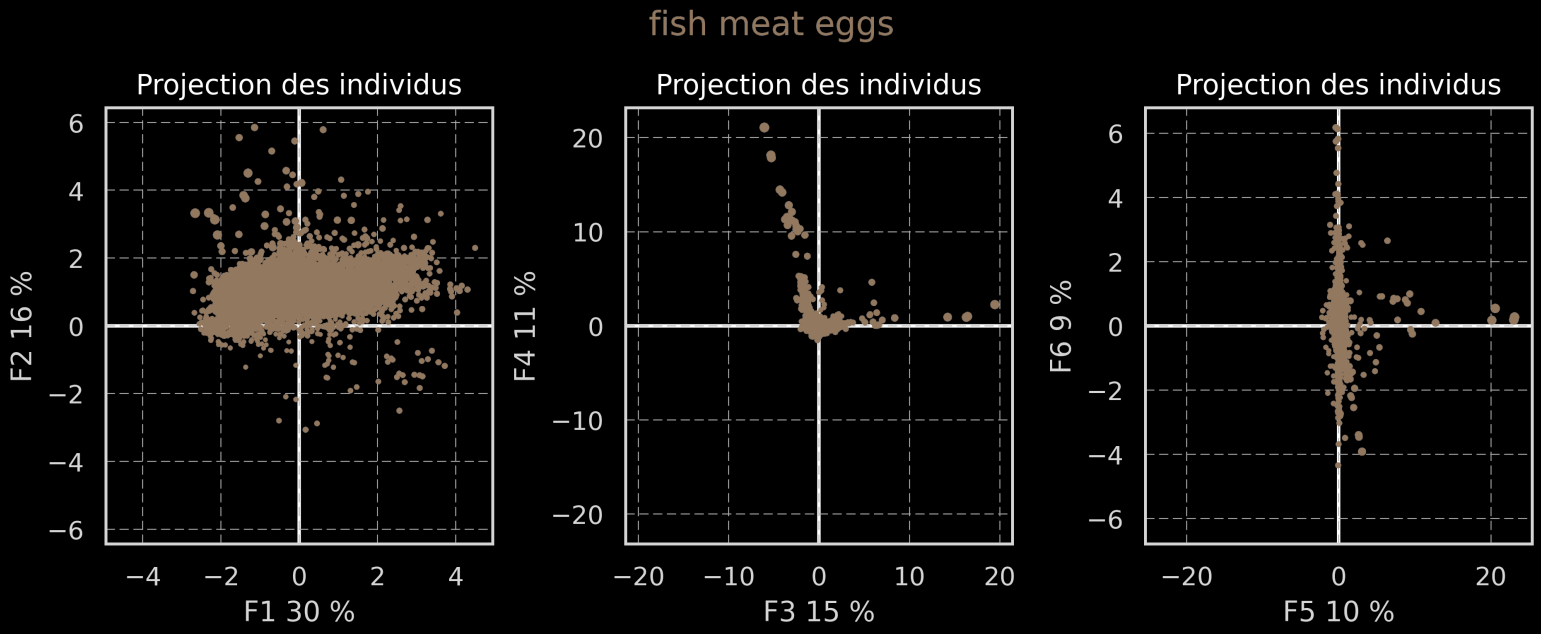


composite foods

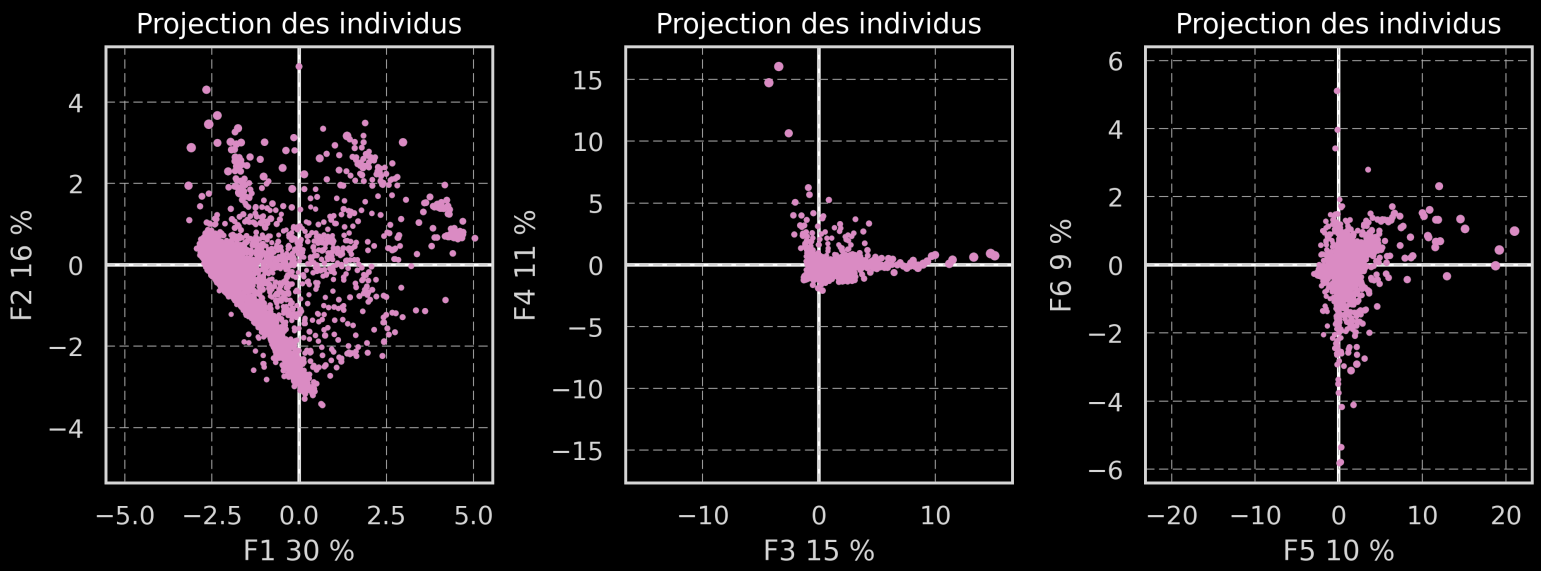


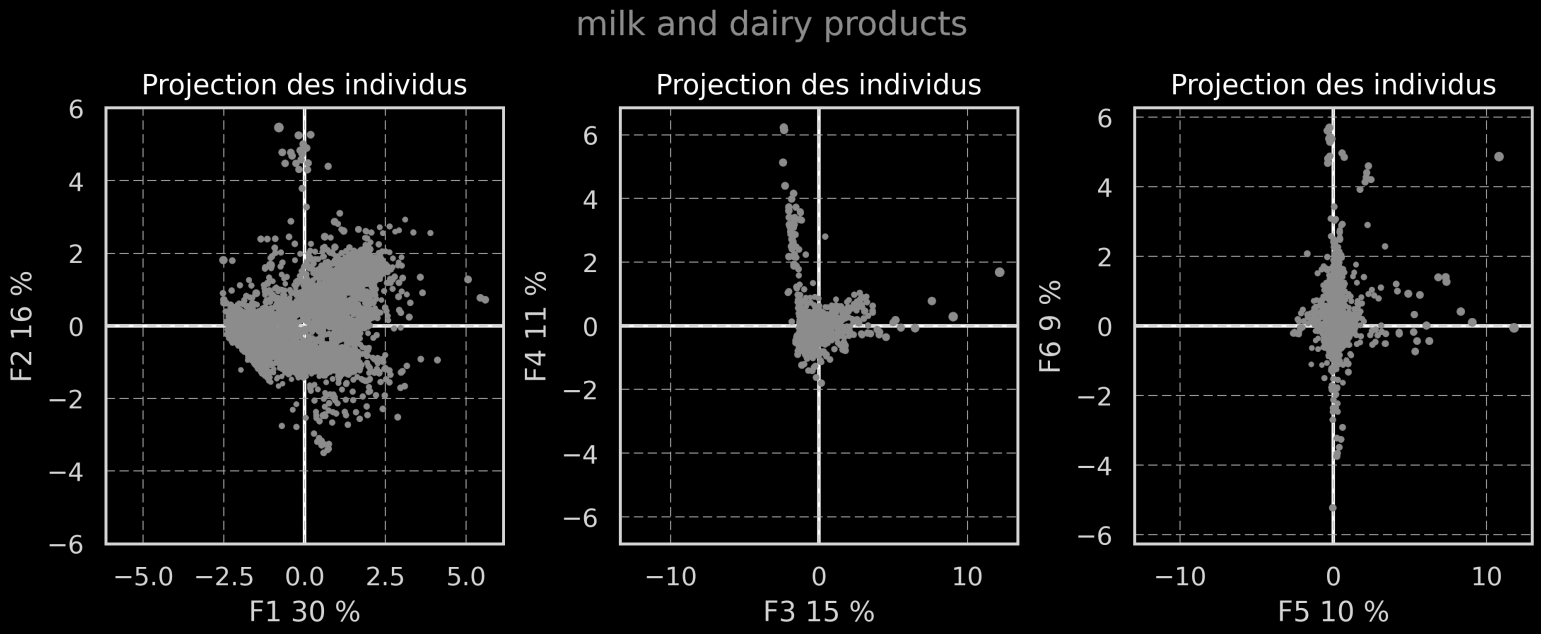
fat and sauces

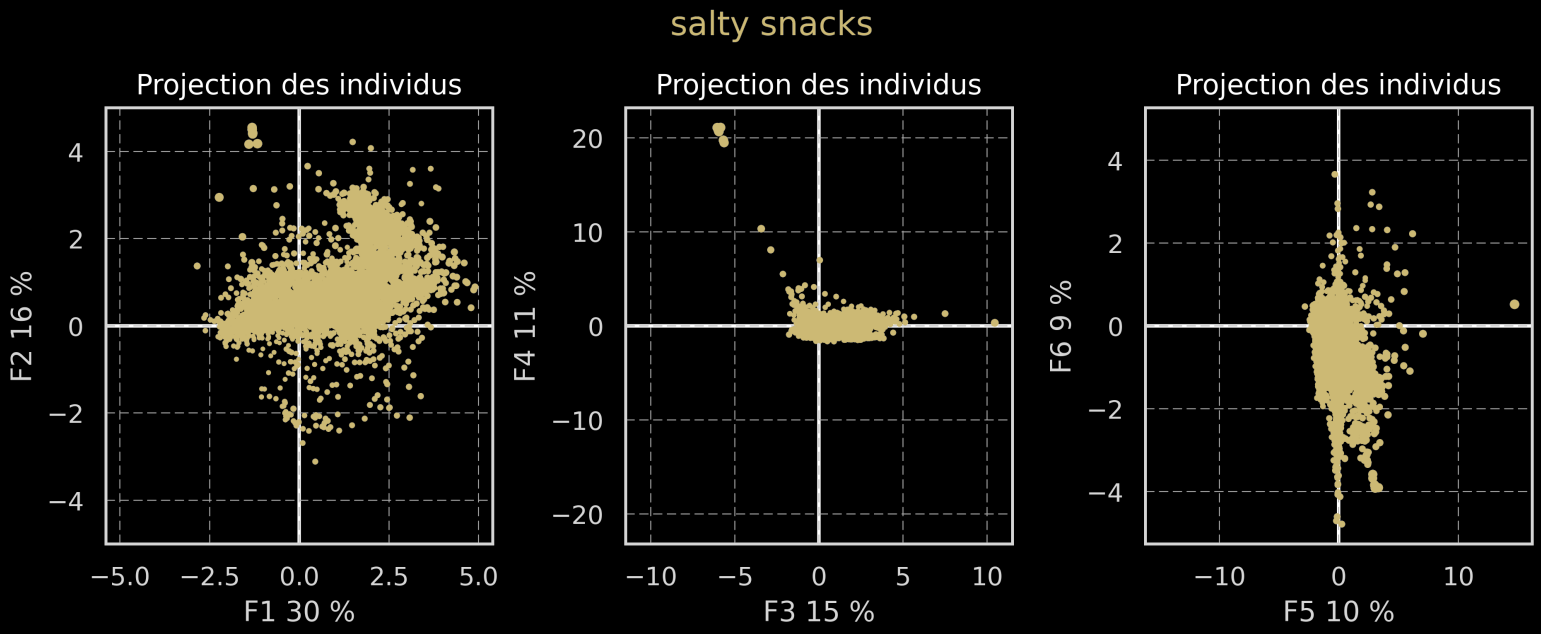



fish meat eggs
Projection des individus
F2 16 %
F1 30 %
Projection des individus
F4 11 %
F3 15 %
Projection des individus
F6 9 %
F5 10 %

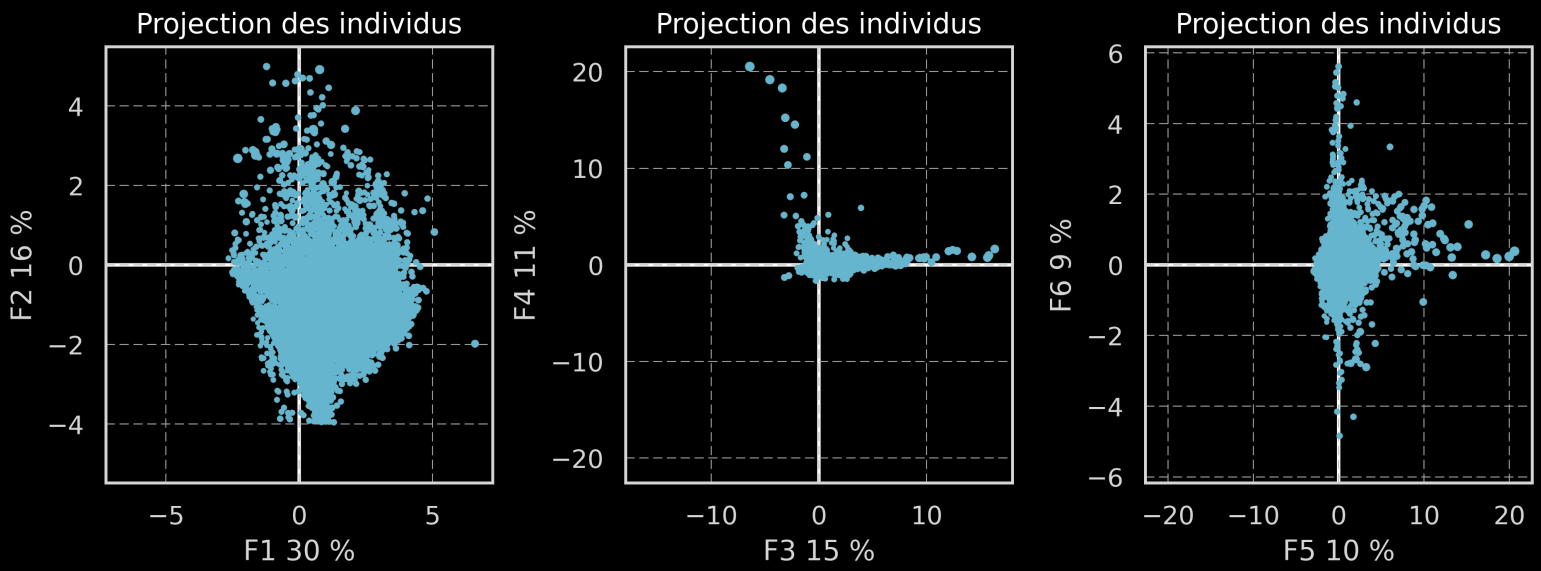
fruits and vegetables




milk and dairy products
Projection des individus
F2 16 %
F1 30 %
Projection des individus
F4 11 %
F3 15 %
Projection des individus
F6 9 %
F5 10 %


salty snacks
Projection des individus
F2 16 %
F1 30 %
Projection des individus
F4 11 %
F3 15 %
Projection des individus
F6 9 %
F5 10 %

sugary snacks



Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page ha been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, bec LaTeX now knows how many pages to expect for this document.