



OPENCLASSROOMS

Formation **Ingénieur Machine Learning**

Projet: Concevez une application au service de la santé publique

5 Janvier 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**

```
graph TD; A["Appel à projet 'Santé Publique France' autour de l'alimentation"] --> B["Proposition d'application"]; B --> C["Vérification de la faisabilité"];
```

Proposition d'**application**

Vérification de la **faisabilité**

CONTEXTE

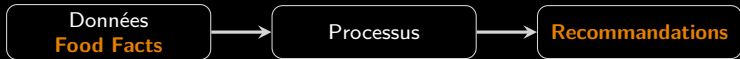
Application proposée

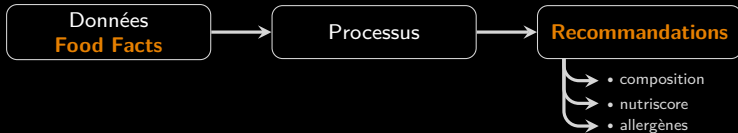
Données
Food Facts

Données
Food Facts



Processus





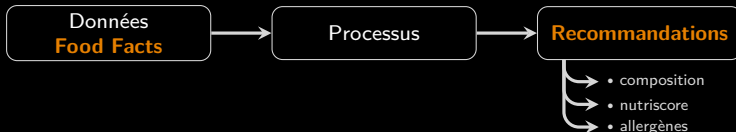
Données
Food Facts

Processus

Recommandations

- composition
- nutriscore
- allergènes

Augmentation des
pathologies cardiaques



Augmentation des
pathologies cardiaques

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

Données
Food Facts

Processus

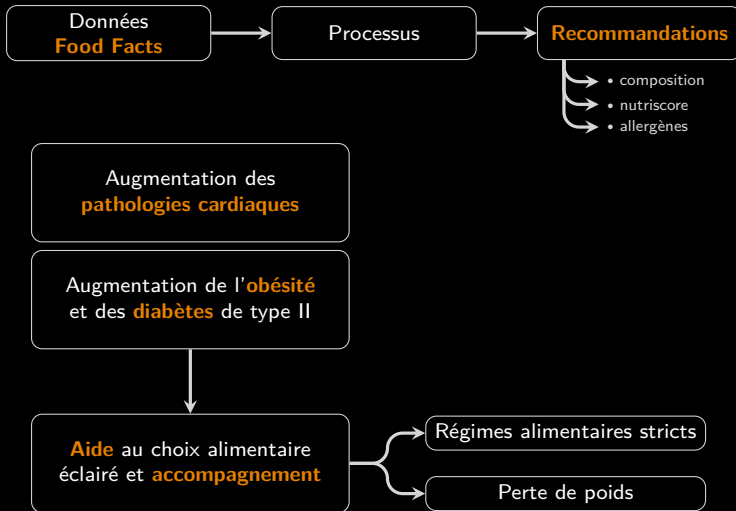
Recommandations

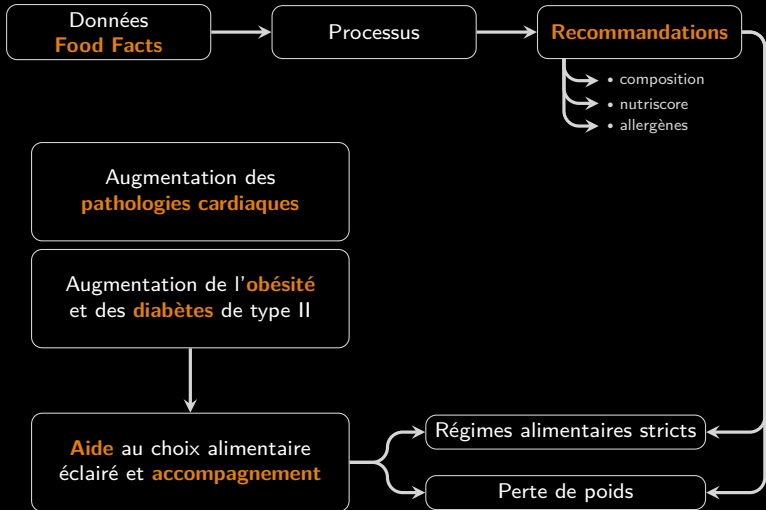
- composition
- nutriscore
- allergènes

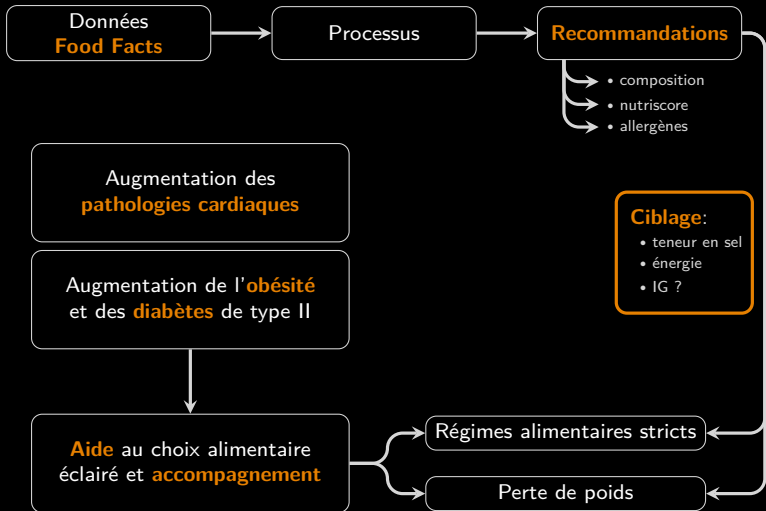
Augmentation des
pathologies cardiaques

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

Aide au choix alimentaire
éclairé et **accompagnement**







Profil utilisateur:

Nom / pseudo:

Pathologies:

Allergènes:

Alim. interdits:

Taille:

Poids:

Poids cible:

Profil utilisateur:

Nom / pseudo:

Pathologies:

Allergènes:

Alim. interdits:

Taille:

Poids:

Poids cible:

Affichage produit:

Nutriscore: **A**

Énergie pour 100g (kJ): **3790**

Sucres pour 100g: **20g**

Graisses pour 100g: **10g**

Teneur en sel: **2%**

Présence d'allergènes: **Aucun**

Labels: **bio, non-OGM**

Produits similaires recommandés

Profil utilisateur:

Nom / pseudo:

Pathologies:

Allergènes:

Alim. interdits:

Taille:

Poids:

Poids cible:

Affichage produit:

Nutriscore: **A**

Énergie pour 100g (kJ): **3790**

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

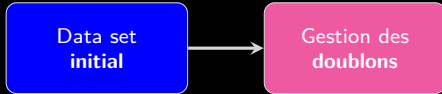
- *catégories*
- *additifs*
- ...

**Informations
nutritionnelles**

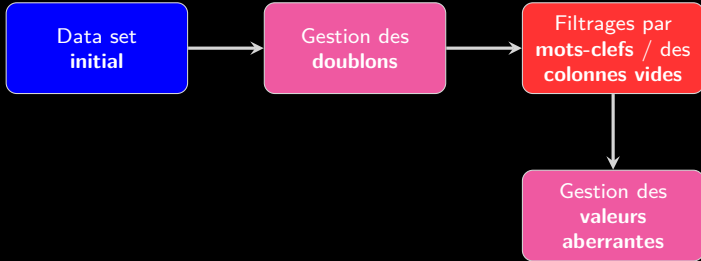
- *protéines*
- *graisses*
- ...

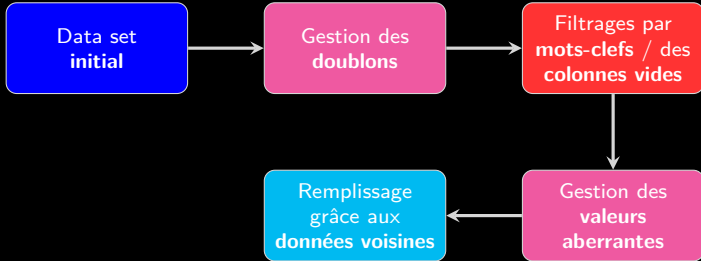
NETTOYAGE

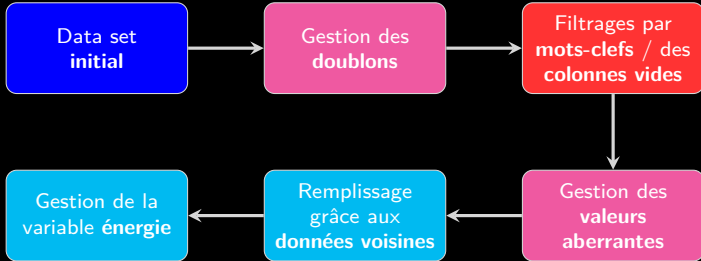
Data set
initial

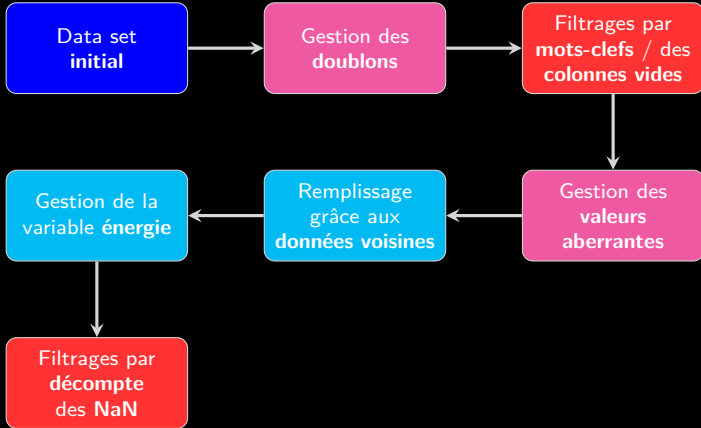


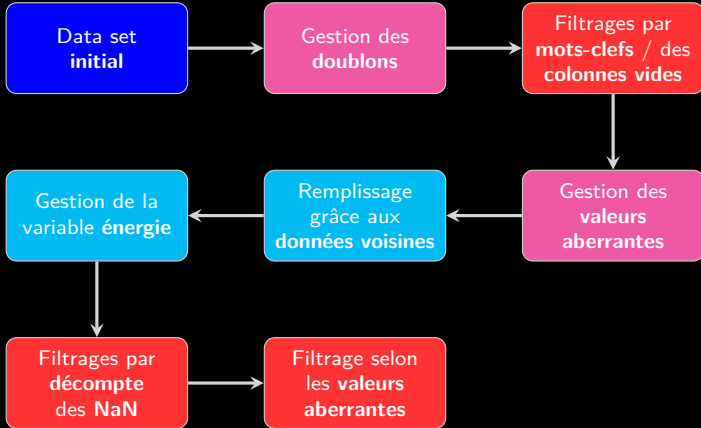


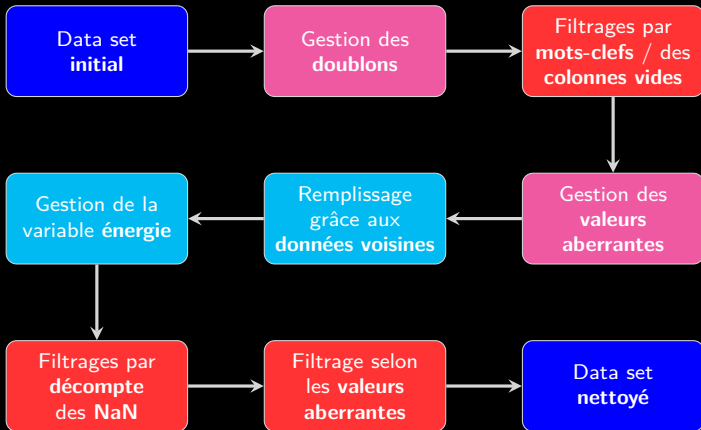


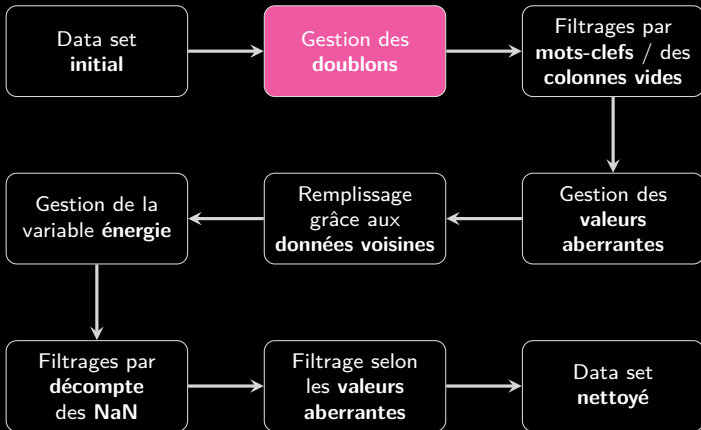












Détection et tri ascendant par valeurs:
(création d'un DataFrame temporaire)

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

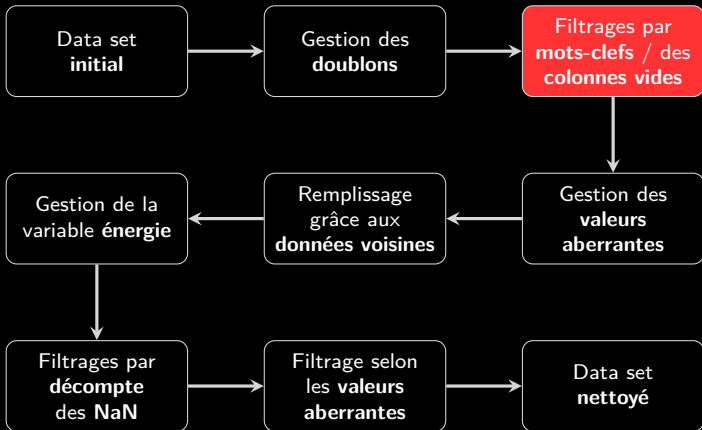
index	code	last modified datetime	n filled
421527	31843340000818	2021-08-17t06:35:03z	28
349035	31843340000818	2022-02-11t08:47:36z	30
61995	3560070278831	2021-04-17t07:44:17z	41
188851	3560070278831	2022-02-10t18:03:06z	47
270028	3700320230572	2021-08-24t12:58:09z	16
749882	3700320230572	2021-08-24t12:58:58z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



191

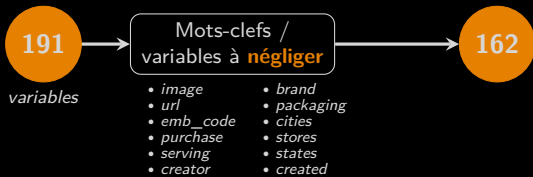
variables

191

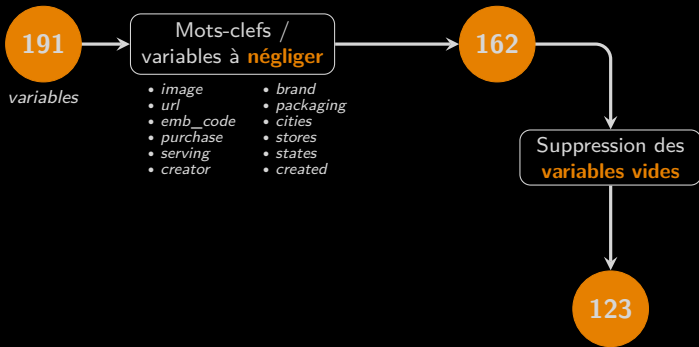
variables

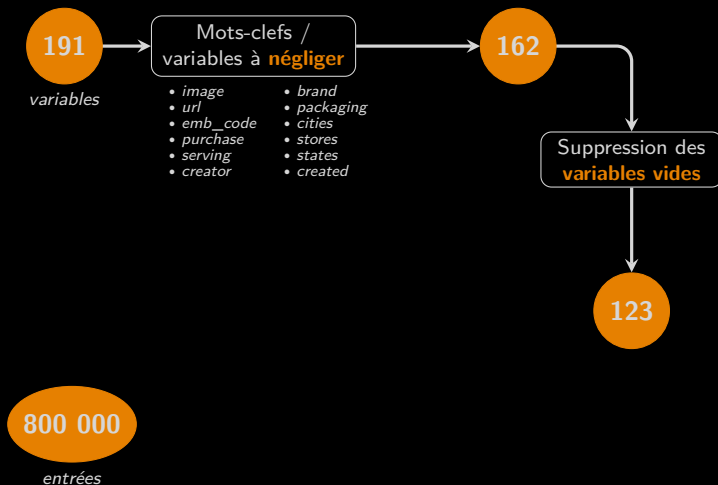
Mots-clefs /
variables à **négliger**

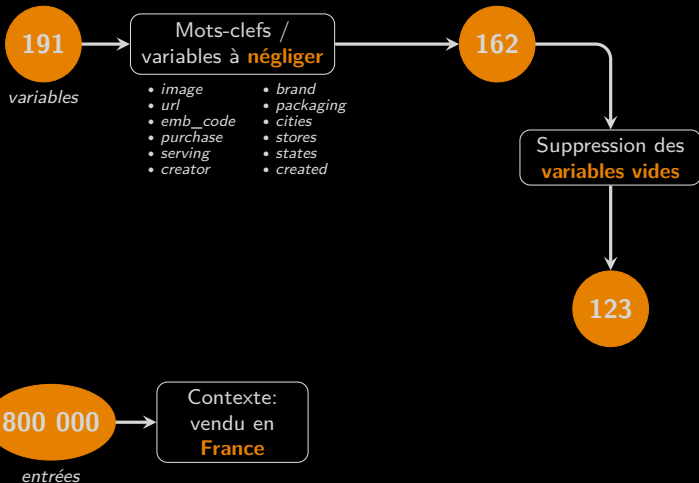
- *image*
- *url*
- *emb_code*
- *purchase*
- *serving*
- *creator*
- *brand*
- *packaging*
- *cities*
- *stores*
- *states*
- *created*

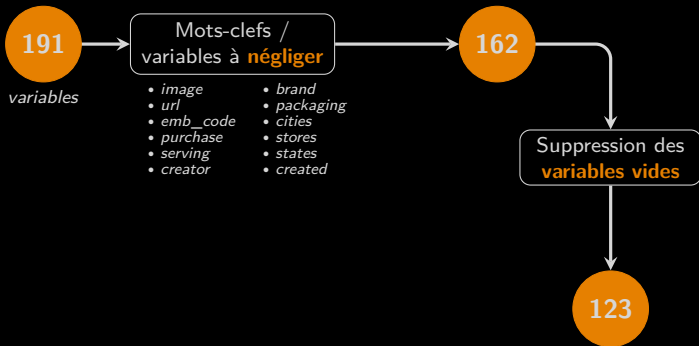












191

variables

Mots-clefs /
variables à **négliger**

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

162

Suppression des
variables vides

123

Suppression des
variables pays

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

800 000

entrées

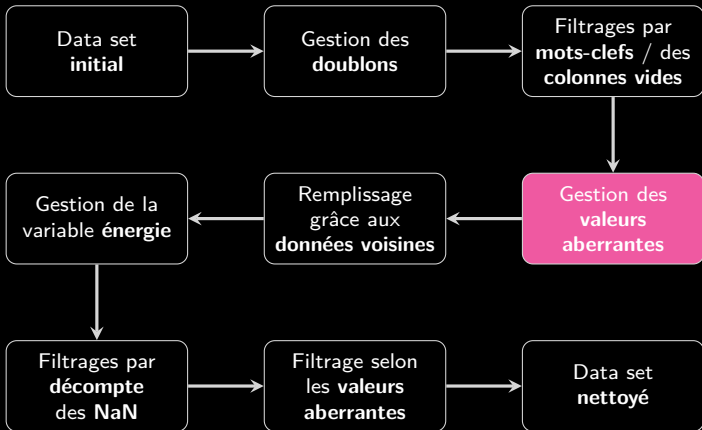
Contexte:
vendu en
France

321 630

entrées
restantes

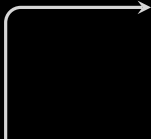
120

variables
restantes



Détection des
valeurs aberrantes

Détection des
valeurs aberrantes



Bornages des
valeurs "pour 100g"

Détection des
valeurs aberrantes

Bornages des
valeurs "pour 100g"

- Général: $0 \leq \text{valeurs} \leq 100$
- Nutriscore: $-15 \leq \text{valeurs} \leq 40$
- ph: $0 \leq \text{valeurs} \leq 14$
- Énergie: $0 \leq \text{valeurs} \leq 3700$

Détection des
valeurs aberrantes

Bornages des
valeurs "pour 100g"

- Général: $0 \leq \text{valeurs} \leq 100$
- Nutriscore: $-15 \leq \text{valeurs} \leq 40$
- ph: $0 \leq \text{valeurs} \leq 14$
- Énergie: $0 \leq \text{valeurs} \leq 3700$

Comparaison des variables
connexes "pour 100g"

Détection des
valeurs aberrantes

Bornages des
valeurs "pour 100g"

- Général: $0 \leq \text{valeurs} \leq 100$
- Nutriscore: $-15 \leq \text{valeurs} \leq 40$
- ph: $0 \leq \text{valeurs} \leq 14$
- Énergie: $0 \leq \text{valeurs} \leq 3700$

Comparaison des variables
connexes "pour 100g"

- *carbohydrates* \geq *sugars*
- *salt* \geq *sodium*
- *fat* $>$ *other fats*

Détection des
valeurs aberrantes

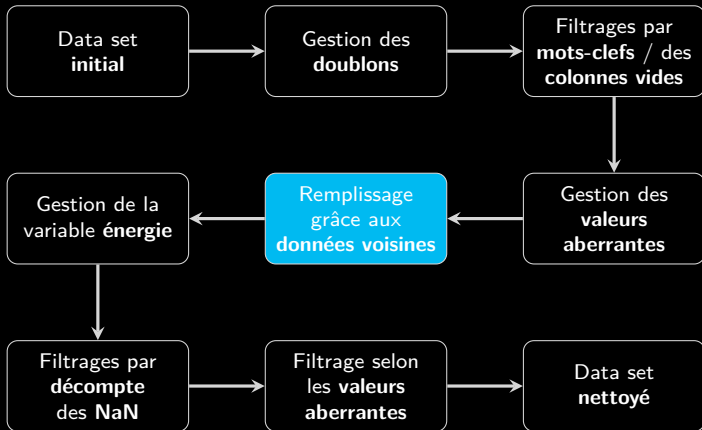
Bornages des
valeurs "pour 100g"

- Général: $0 \leq \text{valeurs} \leq 100$
- Nutriscore: $-15 \leq \text{valeurs} \leq 40$
- ph: $0 \leq \text{valeurs} \leq 14$
- Énergie: $0 \leq \text{valeurs} \leq 3700$

Comparaison des variables
connexes "pour 100g"

- *carbohydrates* \geq *sugars*
- *salt* \geq *sodium*
- *fat* $>$ *other fats*

Remplacement
des valeurs
aberrantes (NaN)



Détection des NaN

pour les variables:

Détection des NaN

pour les variables:

product

name

8576

nutriscore

116 701

Détection des NaN

pour les variables:

product
name
8576

nutriscore
116 701

Récupération de
valeurs dans des
variables connexes

Détection des NaN

pour les variables:

**product
name
8576**

**nutriscore
116 701**

**Récupération de
valeurs dans des
variables connexes**

~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:

product
name
8576

nutriscore
116 701

Récupération de
valeurs dans des
variables connexes

~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:

product
name

8576

nutriscore

116 701

Récupération de
valeurs dans des
variables connexes

~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:

product
name

8576

nutriscore

116 701

Récupération de
valeurs dans des
variables connexes

~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:

product
name

8576

nutriscore

116 701

Récupération de
valeurs dans des
variables connexes

product
name

185

nutriscore

1

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

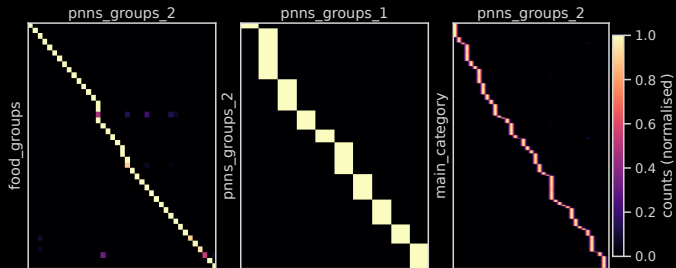
Création de **maps** à partir de **tableaux de contingences**

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

179 241 NaN pour la variable pnns groups 1

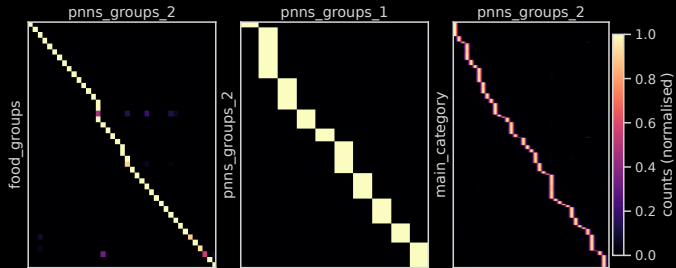
Création de **maps** à partir de **tableaux de contingences**

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

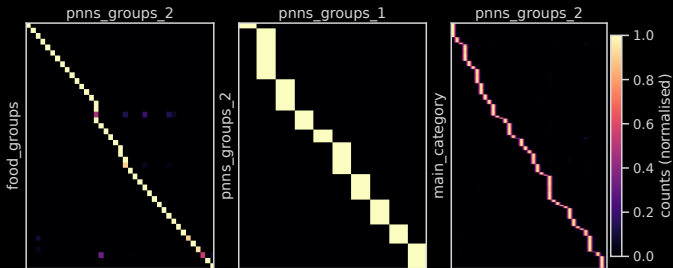


food groups → pnns groups

main category → pnns groups

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

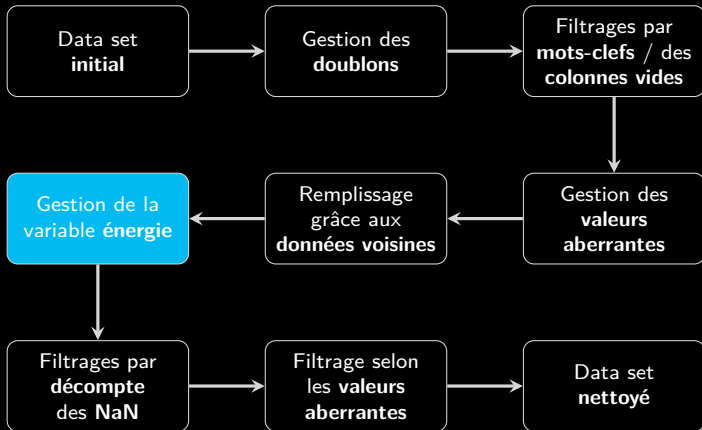
179 241 NaN pour la variable pnnns groups 1



food groups → pnnns groups

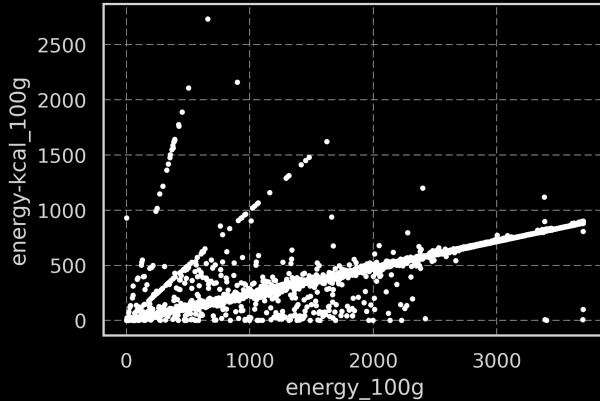
main category → pnnns groups

Remplissage de **9 401** valeurs



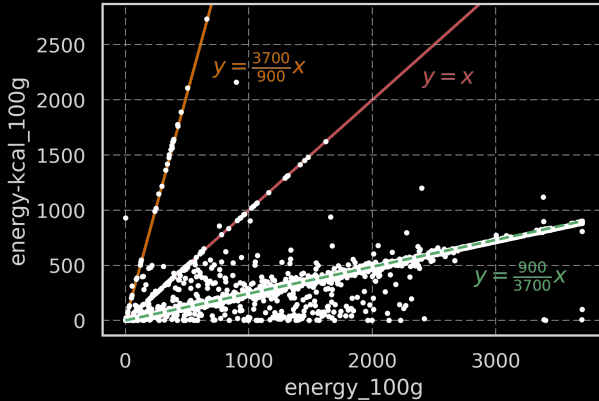
Présence de **valeurs incohérentes**:

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



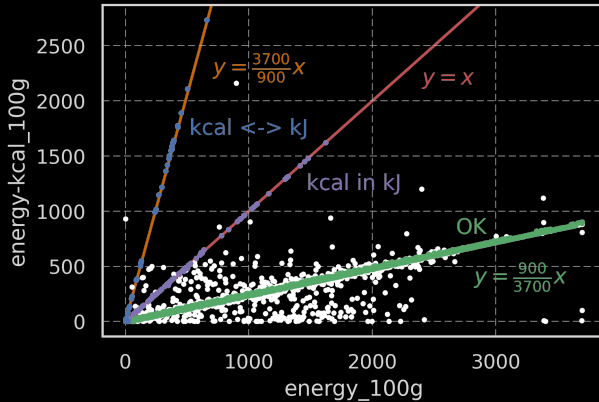
Présence de **valeurs incohérentes**:

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



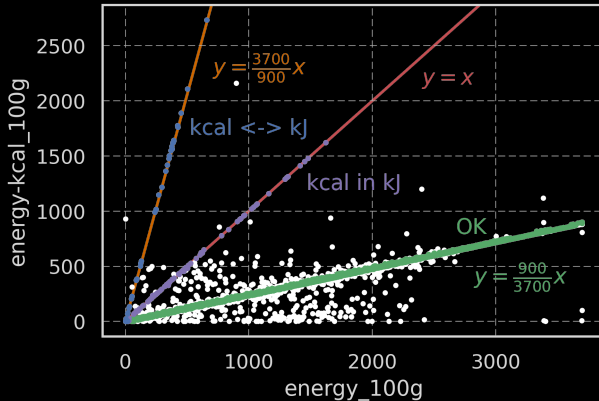
Présence de **valeurs incohérentes**:

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

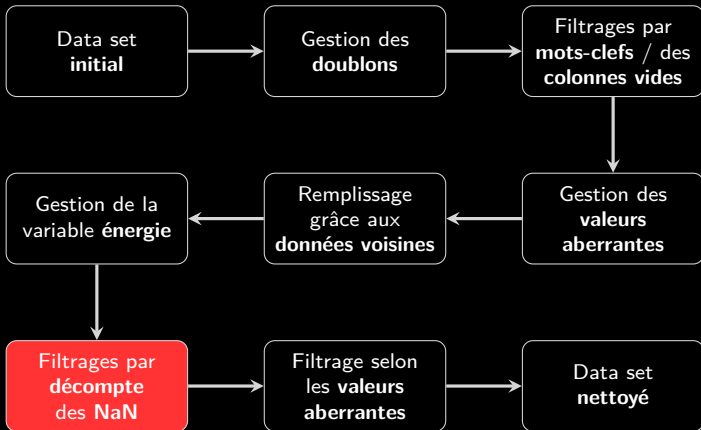


Présence de **valeurs incohérentes**:

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

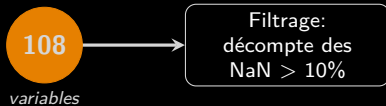


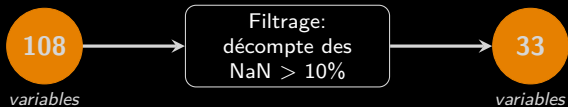
Solution: garder la **valeur maximale** (en kJ)





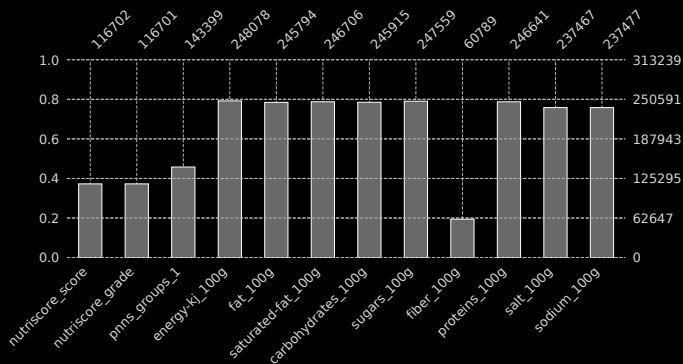
variables





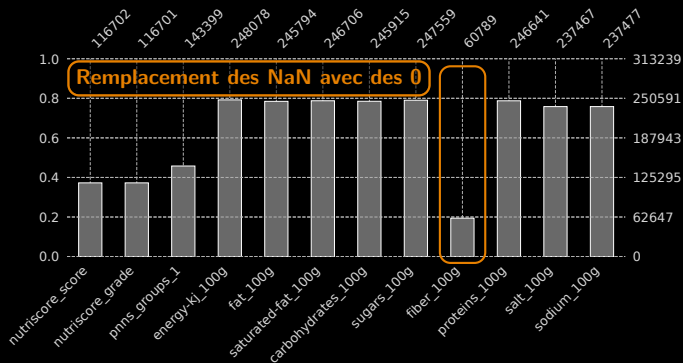


Taux de **remplissage** des **variables sélectionnées**:

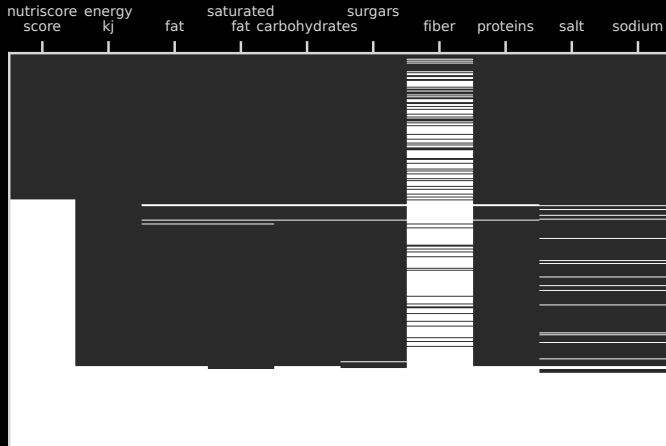




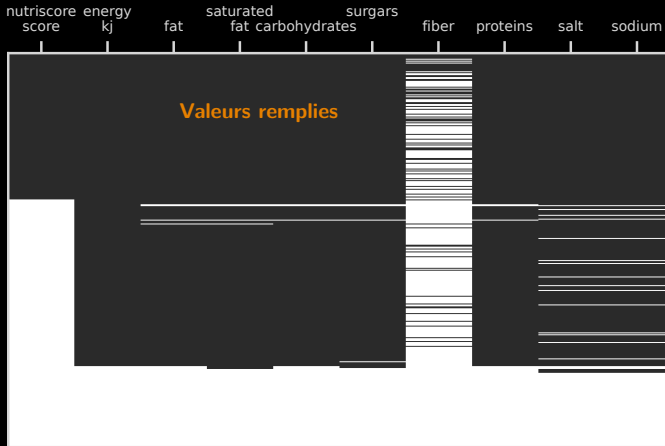
Taux de **remplissage** des **variables sélectionnées**:



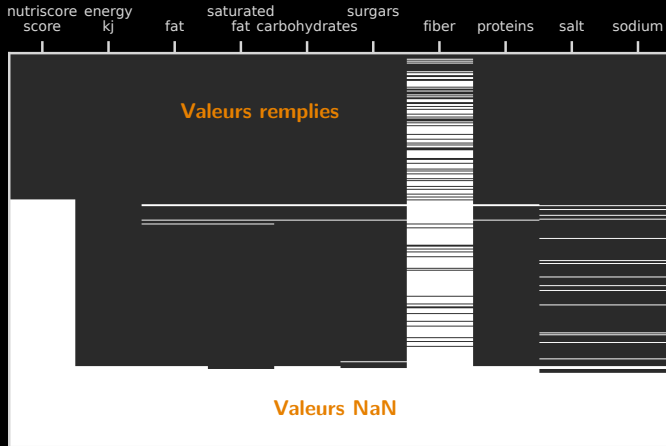
Matrice de **remplissage** des
variables sélectionnées:
(tri ascendant des valeurs)



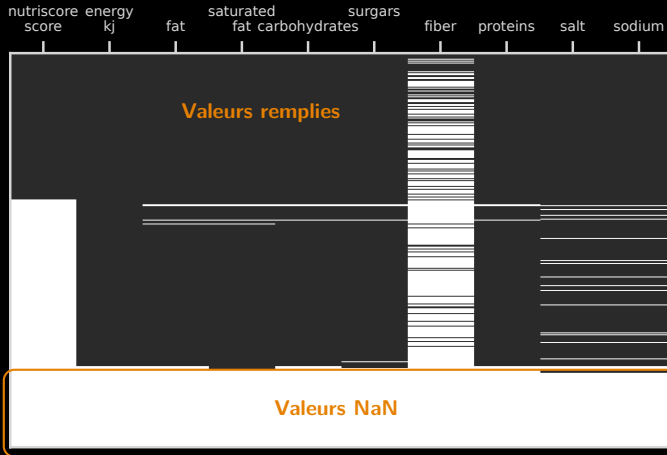
Matrice de **remplissage** des
variables sélectionnées:
(tri ascendant des valeurs)



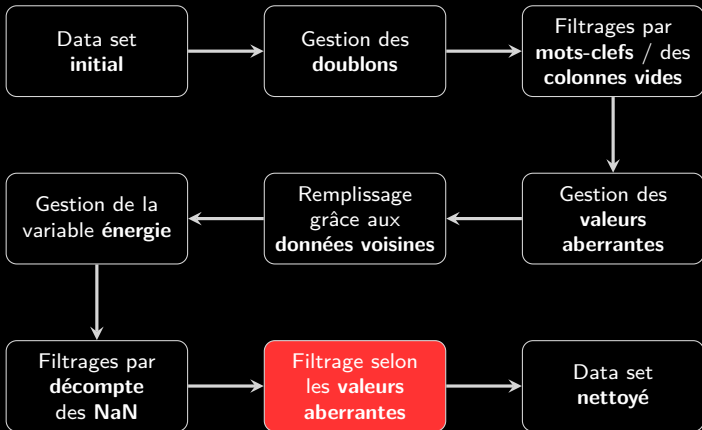
Matrice de **remplissage** des
variables sélectionnées:
(tri ascendant des valeurs)



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



Suppression des entrées vides



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

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

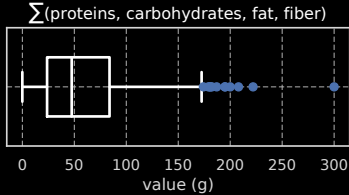
Données **statistiques** sur les **données brutes**:

	count	mean	std	min	25%	50%	75%	max
Σ 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**:

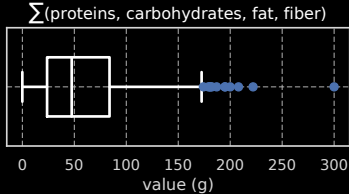
	count	mean	std	min	25%	50%	75%	max
Σ poids	244025	51.27	31.43	0	23.8	47.5	83.9	300



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

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

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



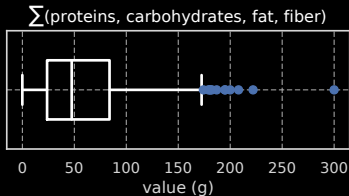
$$Q3 + 1.5 * IQ = \mathbf{174.05}$$

→ 12 outliers

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

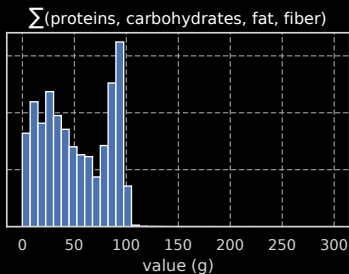
Données **statistiques** sur les **données brutes**:

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



$$Q3 + 1.5 * IQ = \mathbf{174.05}$$

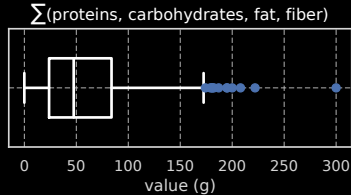
→ 12 outliers



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

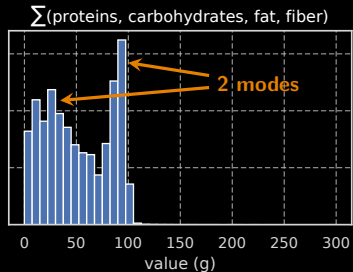
Données **statistiques** sur les **données brutes**:

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



$$Q3 + 1.5 * IQ = \mathbf{174.05}$$

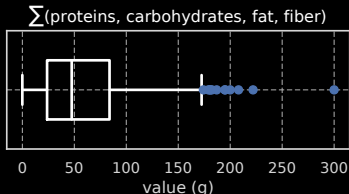
→ 12 outliers



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

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

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

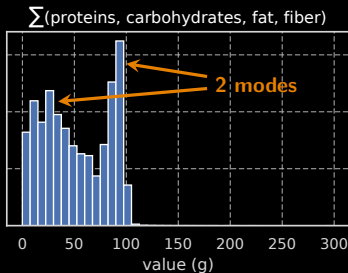


$$Q3 + 1.5 * IQ = \mathbf{174.05}$$

→ 12 outliers

$$vmax = 100$$

→ **1651** outliers



EXPLORATION DES DONNÉES

EXPLORATION DES DONNÉES

Remplissage des valeurs éparées

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

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

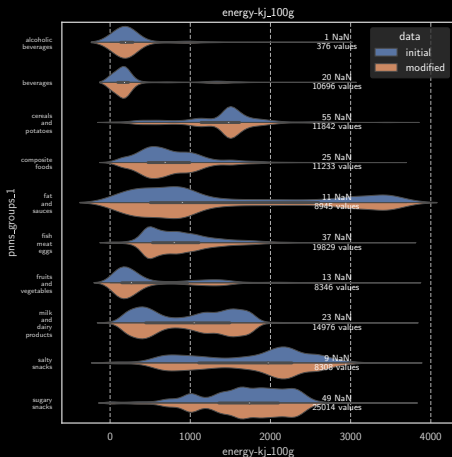


Remplacement des **NaN** par la **valeur moyenne** par pnns groups

Méthode 1: Valeur moyenne par pnns groups et par variable

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

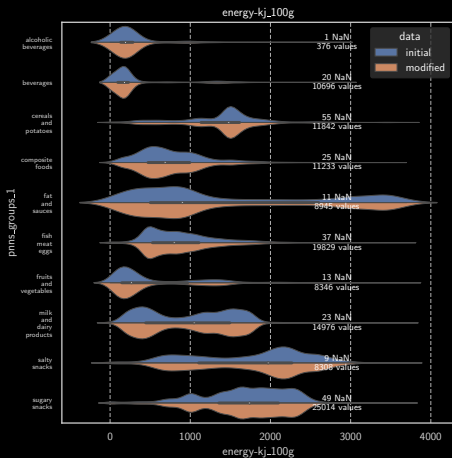
Remplacement des **NaN** par la **valeur moyenne** par pnns groups



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

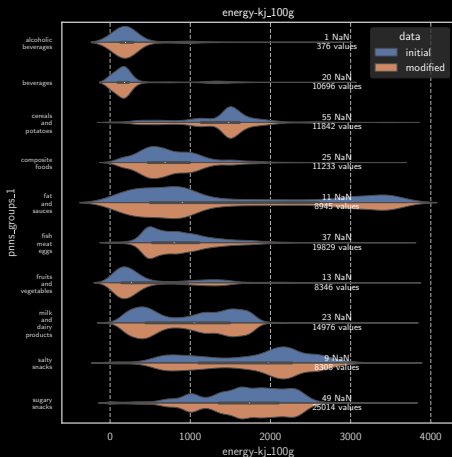


- variations imperceptibles
- distribution **multi-modale**

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

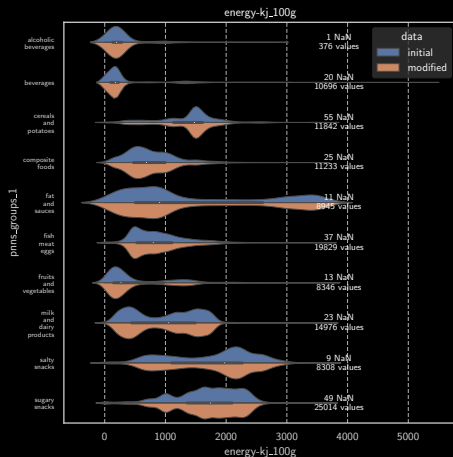


Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

Apprentissage sur des
données complètes



Remplissage des
données incomplètes

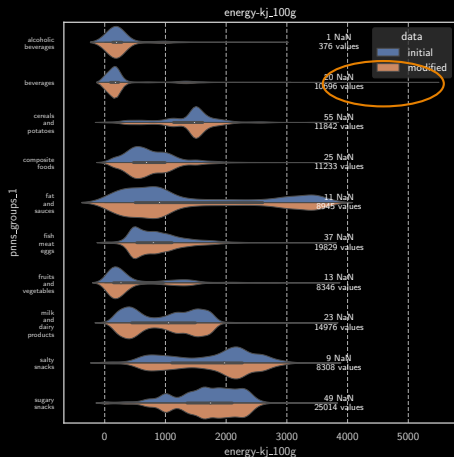


Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

Apprentissage sur des
données complètes



Remplissage des
données incomplètes

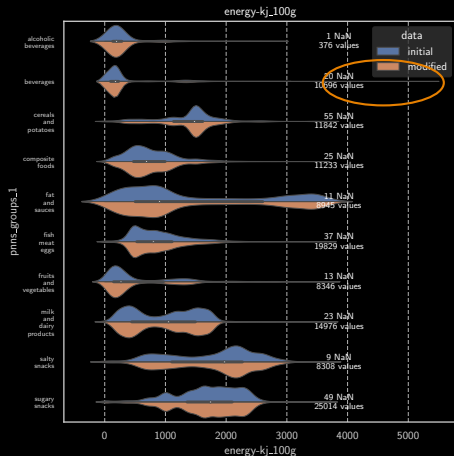


Méthode 2: itérative imputer de la librairie Scikit-learn, basée sur une approche k-nn

Apprentissage sur des
données complètes



Remplissage des
données incomplètes



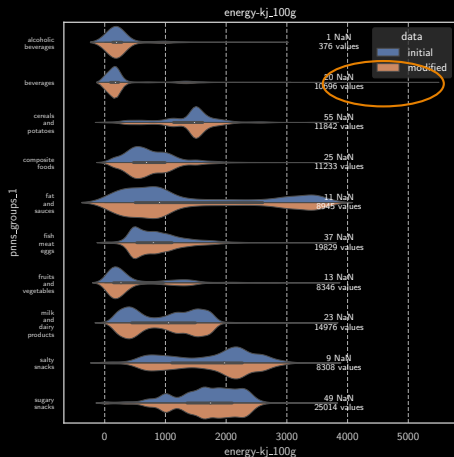
- génération de **valeurs hors bornes**

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:

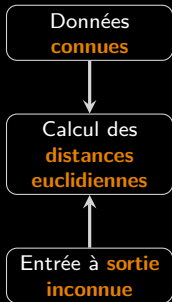
Données
connues

Principe:

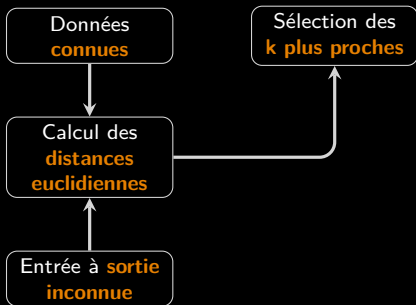
Données
connues

Entrée à **sortie**
inconnue

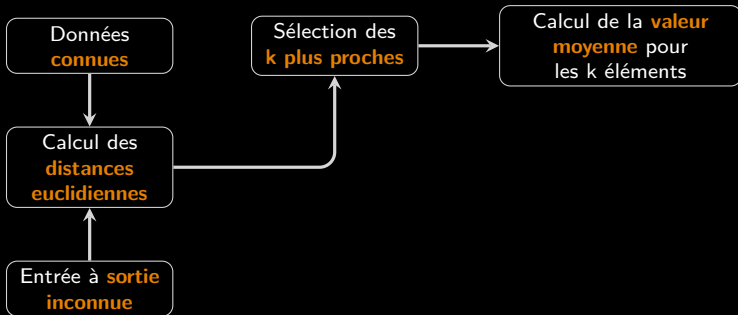
Principe:



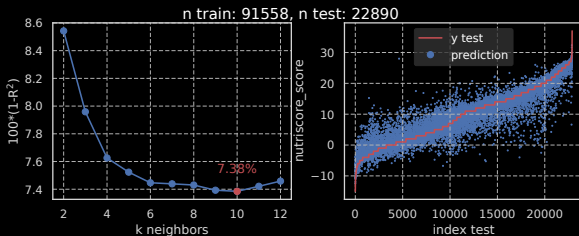
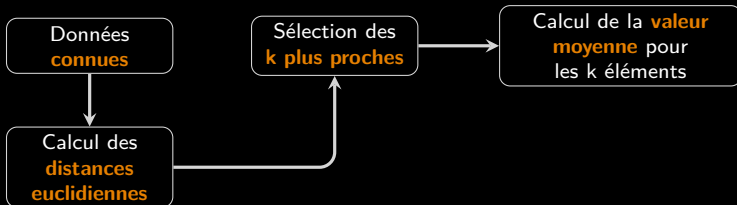
Principe:



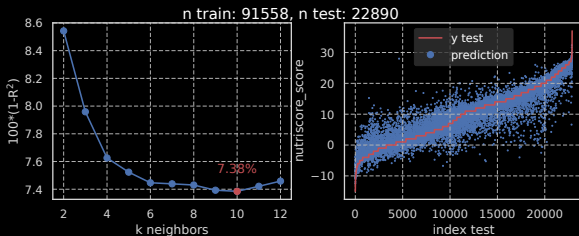
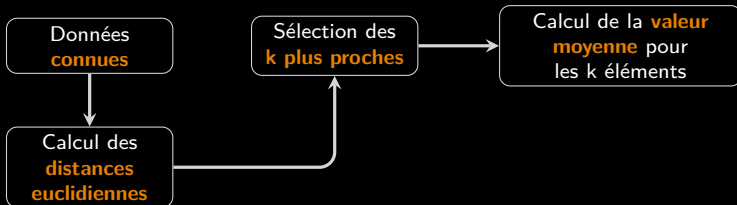
Principe:



Principe:

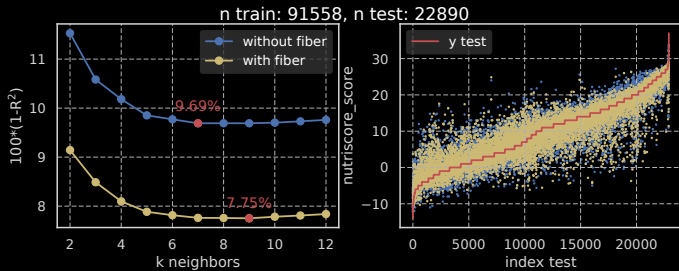


Principe:

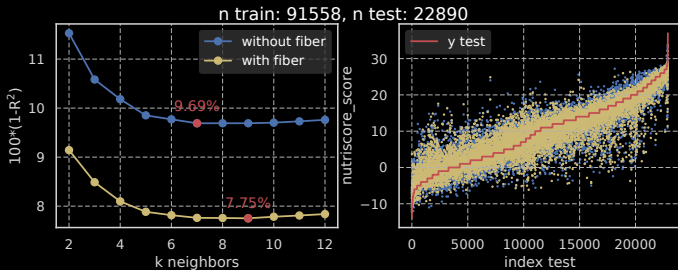


Quid de l'**hypothèse** NaN fibers = 0 ?

Vérification de l'hypothèse NaN fiber = 0:



Vérification de l'hypothèse NaN fiber = 0:



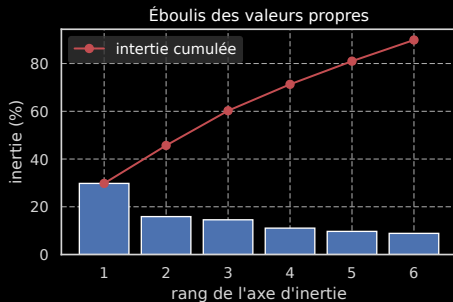
→ Hypothèse à priori **validée**

EXPLORATION DES DONNÉES

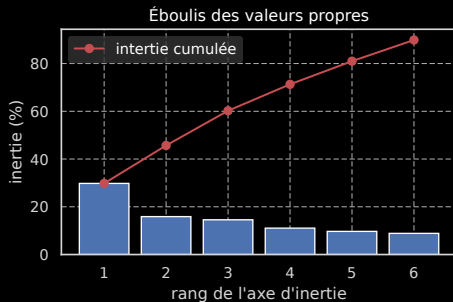
Étude de l'inertie des valeurs - Analyse
en Composantes Principales

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**

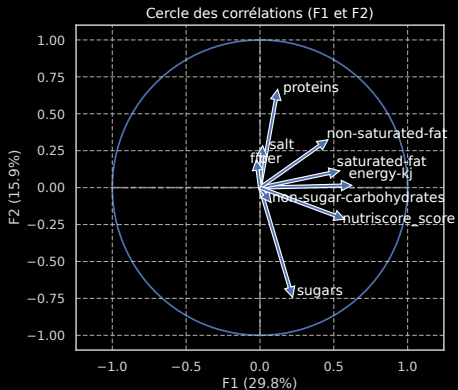


→ **90%** de l'inertie sur **6 axes**

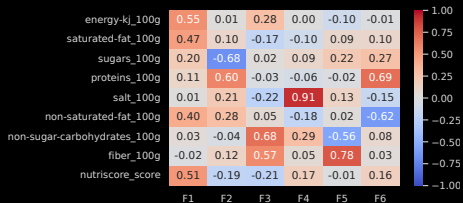
→ Répartition relativement **équilibrée**

Projection des axes initiaux sur les axes estimés

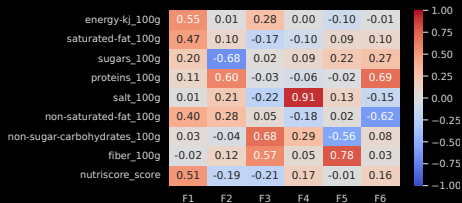
Projection des axes initiaux sur les axes estimés



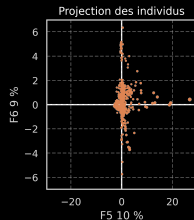
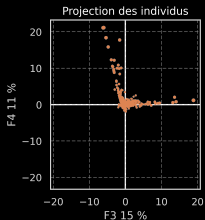
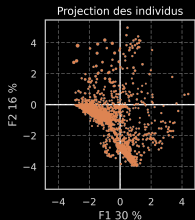
Projection des axes initiaux sur les axes estimés



Projection des axes initiaux sur les axes estimés



beverages

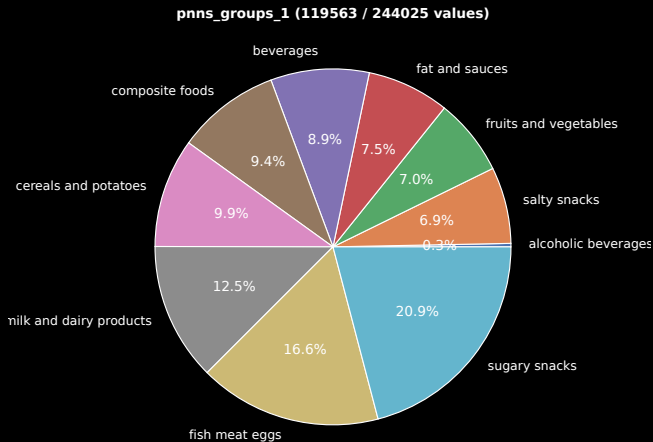


EXPLORATION DES DONNÉES

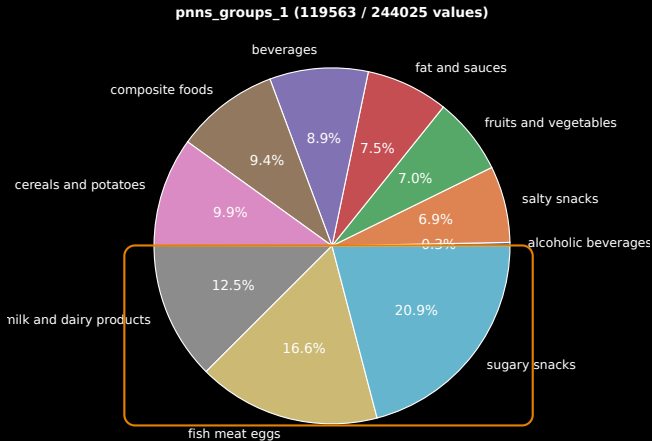
Analyse des données en lien avec
l'application

Répartition pnns groups et nutriscore grade:

Répartition pnns groups et nutriscore grade:

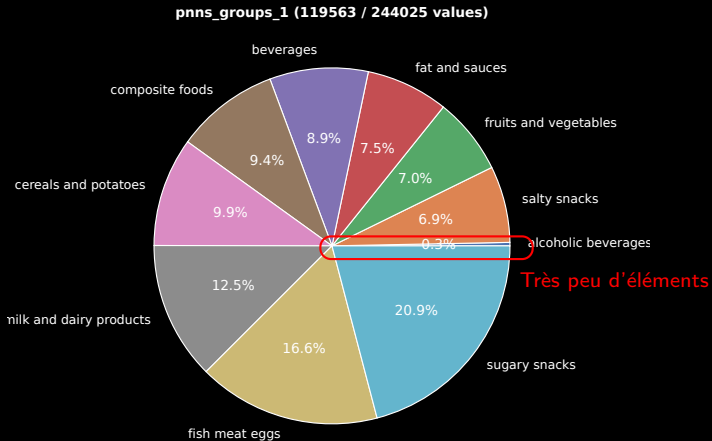


Répartition pnns groups et nutriscore grade:



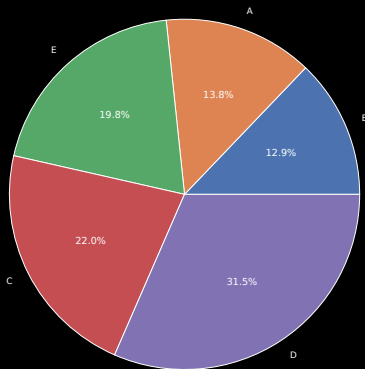
Principaux éléments

Répartition pnns groups et nutriscore grade:



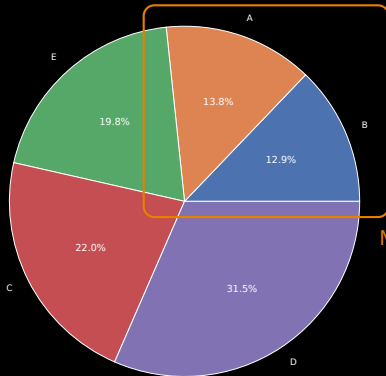
Répartition pnns groups et nutriscore grade:

nutriscore_grade (115813 / 244025 values)



Répartition pns groups et nutriscore grade:

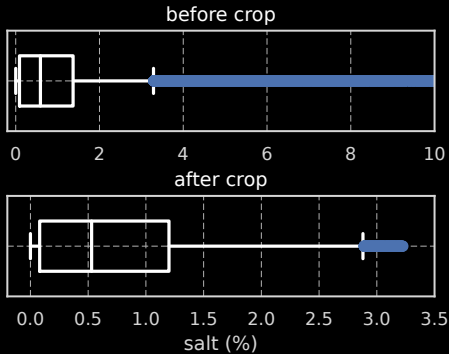
nutriscore_grade (115813 / 244025 values)



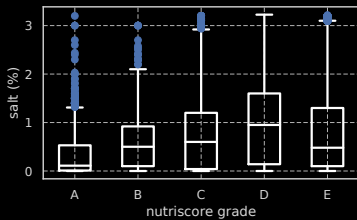
Moindre proportion

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

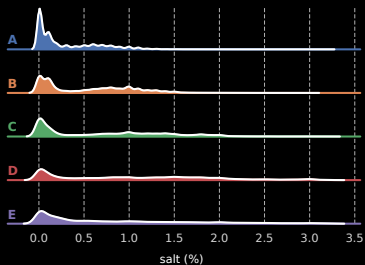
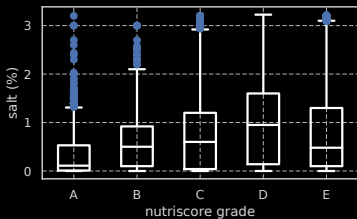
Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



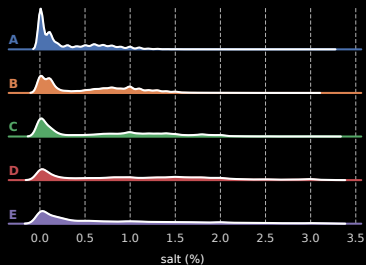
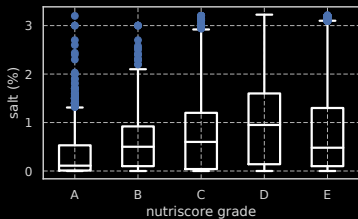
Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



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

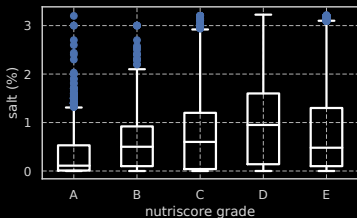


Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



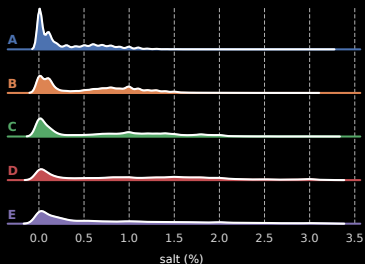
Les groupes sont **différents**

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



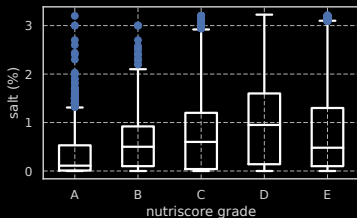
Les groupes sont **différents**

Les moyennes sont
faiblement corrélées
au "grade"



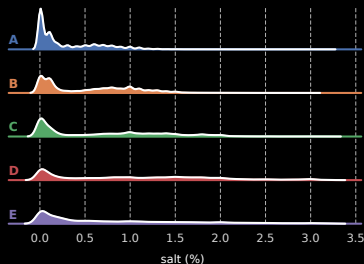
Distributions
multi-modales

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



Les groupes sont **différents**

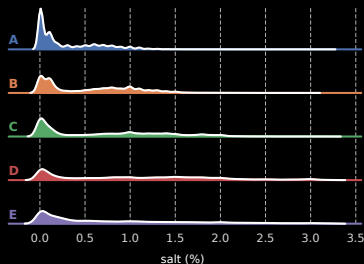
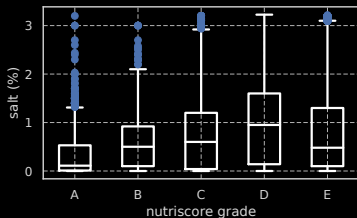
Les moyennes sont
faiblement corrélées
au "grade"



Distributions
multi-modales

	Σy^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**:



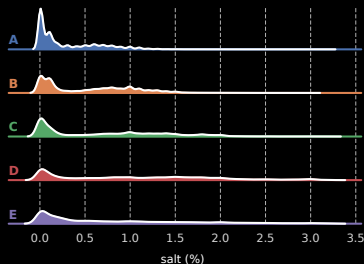
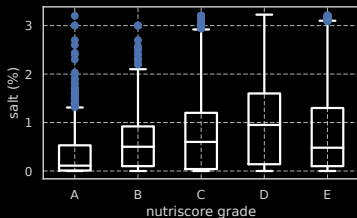
Les groupes sont **différents**

Les moyennes sont
faiblement corrélées
au "grade"

Distributions
multi-modales

	Σy^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**:



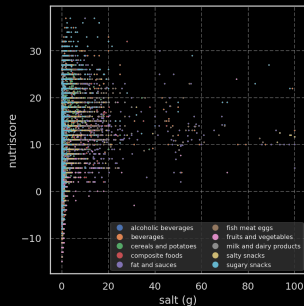
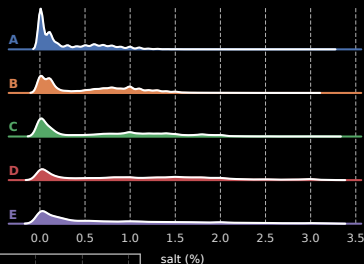
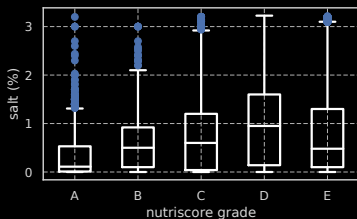
Les groupes sont **différents**

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

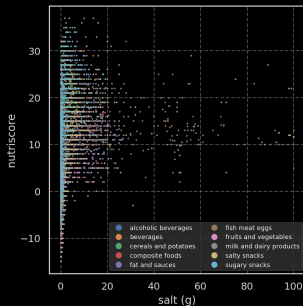
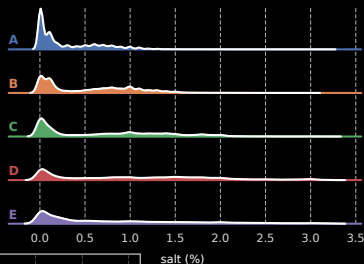
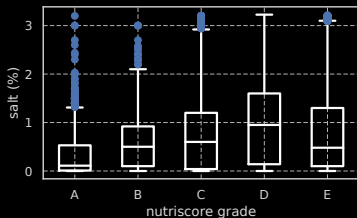
Distributions **multi-modales**

	Σy^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**:



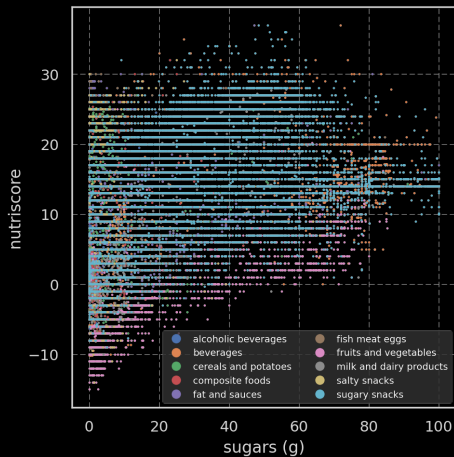
Analyse de la corrélation entre le **nutriscore grade** et la **teneur en sel**:



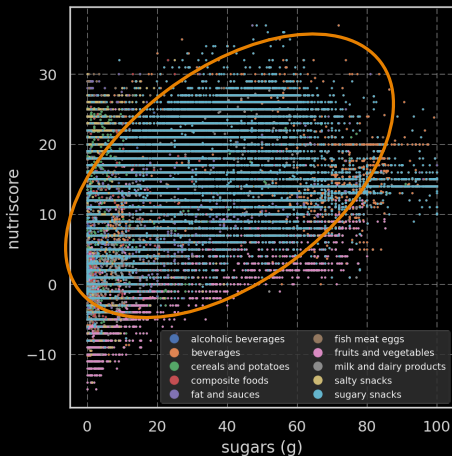
Aucune
corrélation
apparente

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 boîte de garniture: tomates fraîches (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), lait 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 boîte de garniture: tomates fraîches (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), lait 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.

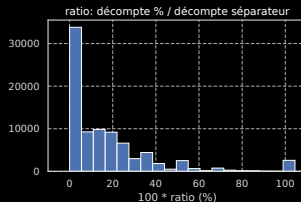
Non, car les % sont assez peu renseignés

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 boîte de garniture: tomates fraîches (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), lait 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:

Faisables (à priori):

Régime pauvre en sel

Gestion de poids/IMC

Vérification de la
présence d'allergènes

Non faisables en l'état:

Gestion de l'indice
glycémique

Se rapprocher de **spécialistes** du domaines
est **nécessaire** pour apporter une
meilleure **connaissance métier**

Pistes d'améliorations des traitements:

Améliorer la gestion des NaN



déterminer le pnnns groups à
partir de la liste des ingrédients

Mettre en place des mo-
teurs de rocommandation



basé sur
le cotenu



filtrage
collaboratif

Merci pour votre **attention**.

knn

Principe:

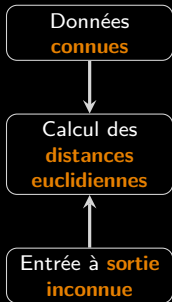
Données
connues

Principe:

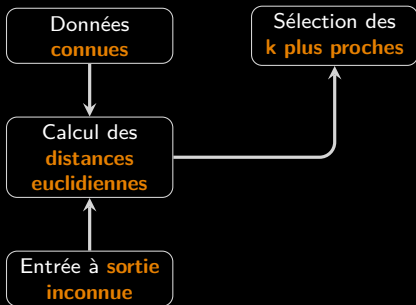
Données
connues

Entrée à **sortie**
inconnue

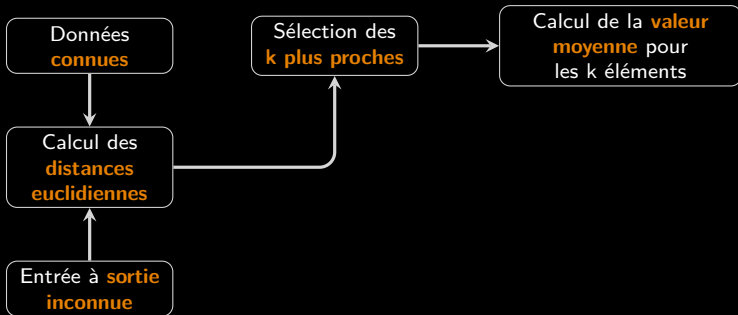
Principe:



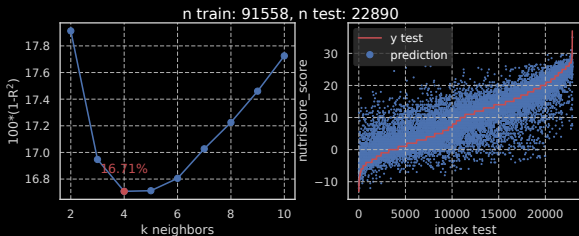
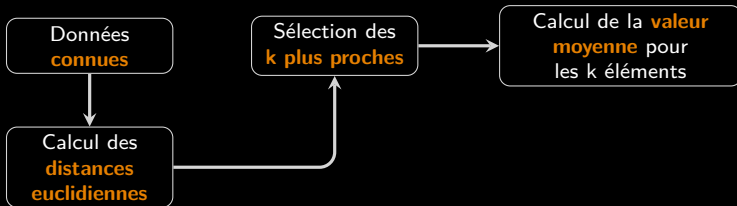
Principe:



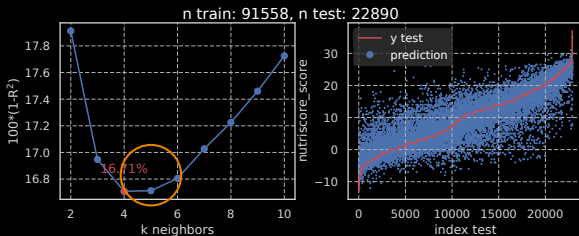
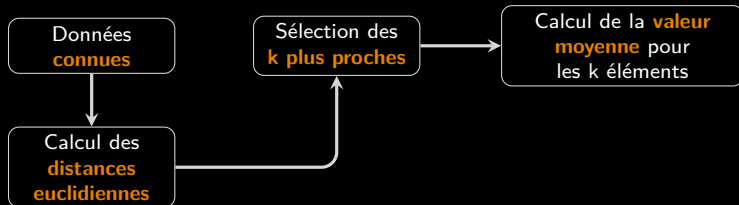
Principe:



Principe:



Principe:

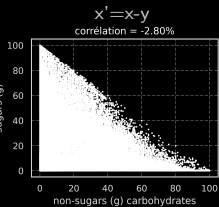
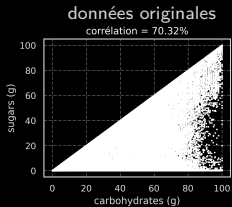


Avec les **données brutes**, **17%** d'erreur → nécessité d'optimiser

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

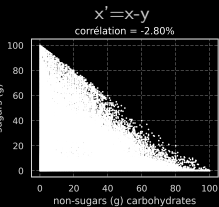
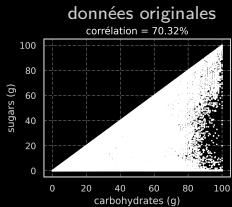
Analyse de la **corrélation** entre les variables:

carbohydrates
/ sugars



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

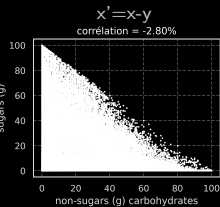
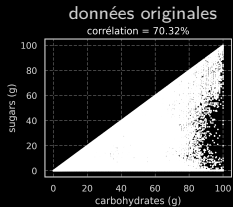
carbohydrates
/ sugars



Décorrélation **OK**

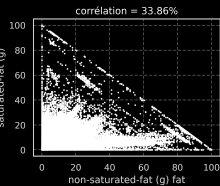
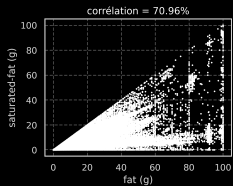
Analyse de la **corrélation** entre les variables:

carbohydrates
/ sugars



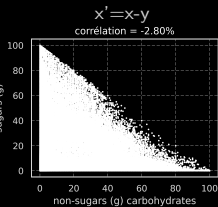
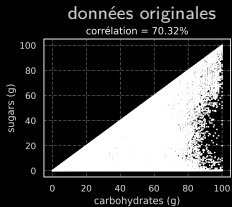
Décorrélation **OK**

fat /
saturated-
fat



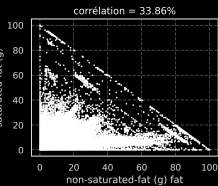
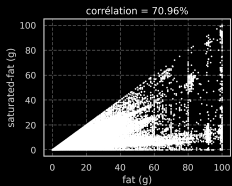
Analyse de la **corrélation** entre les variables:

carbohydrates
/ sugars



Décorrélation **OK**

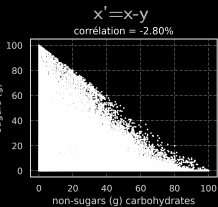
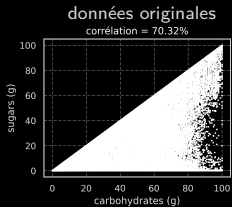
fat /
saturated-
fat



Décorrélation
partielle

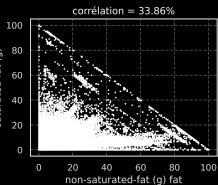
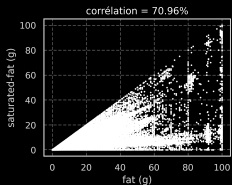
Analyse de la **corrélation** entre les variables:

carbohydrates
/ sugars



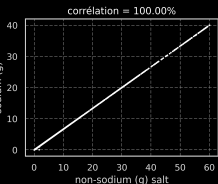
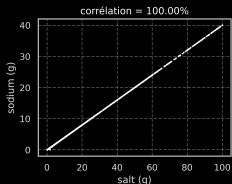
Décorrélation **OK**

fat /
saturated-fat



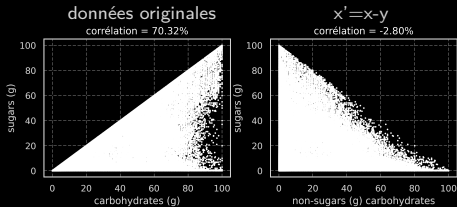
Décorrélation
partielle

sodium /
fat



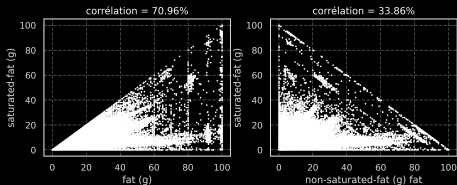
Analyse de la **corrélation** entre les variables:

carbohydrates
/ sugars



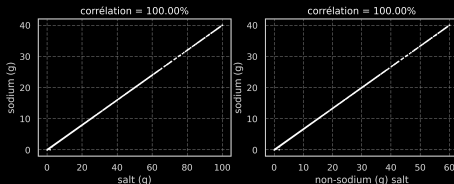
Décorrélation **OK**

fat /
saturated-fat



Décorrélation
partielle

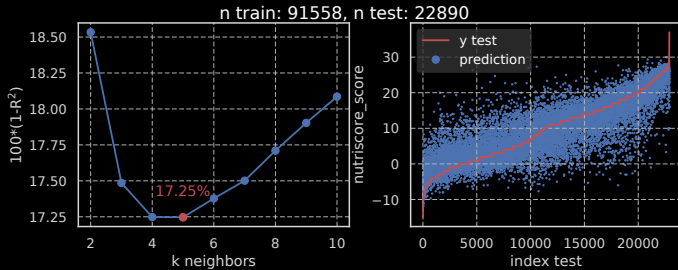
sodium /
fat



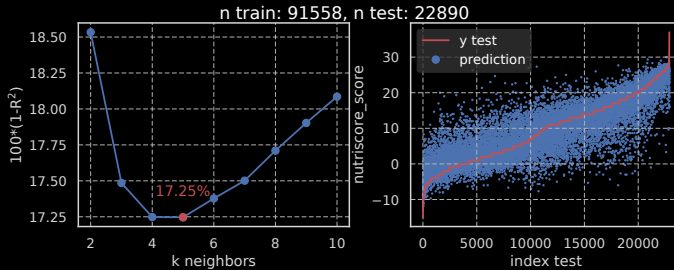
Décorrélation
nulle
→ **suppression**
d'une variable

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):

**Soustraction de
la valeur moyenne**

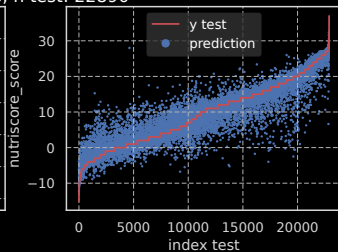
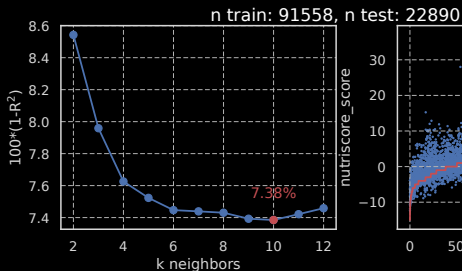


**Normalisation
de la variance**

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

Soustraction de
la **valeur moyenne**

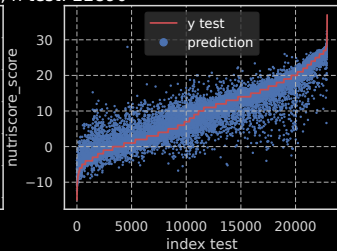
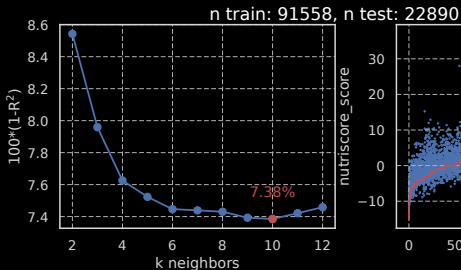
Normalisation
de la **variance**



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

Soustraction de
la **valeur moyenne**

Normalisation
de la **variance**

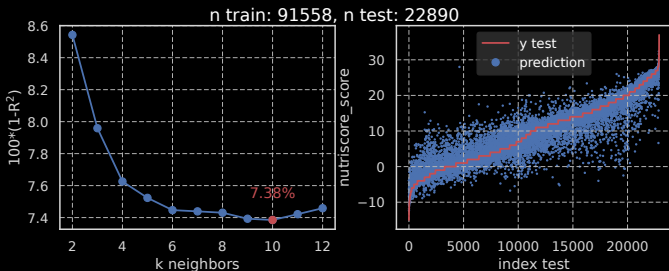


→ **Nette** amélioration

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

Soustraction de
la **valeur moyenne**

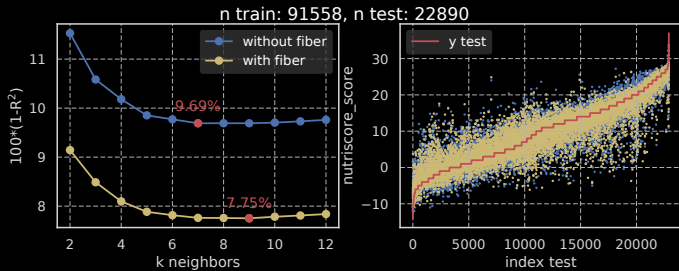
Normalisation
de la **variance**



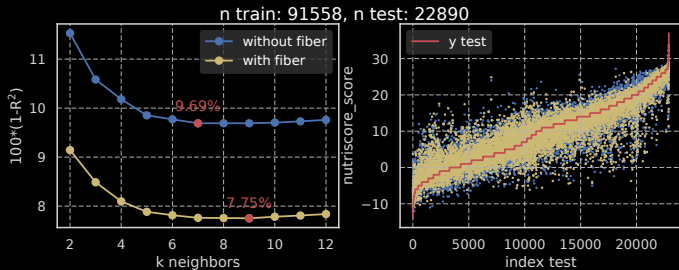
→ **Nette** amélioration

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

Vérification de l'hypothèse NaN fiber = 0:



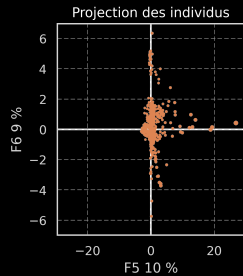
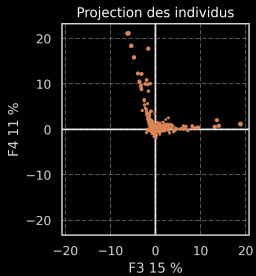
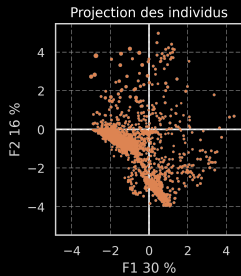
Vérification de l'hypothèse NaN fiber = 0:



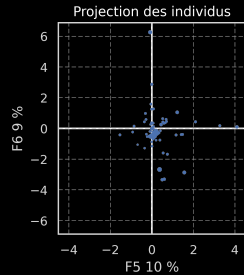
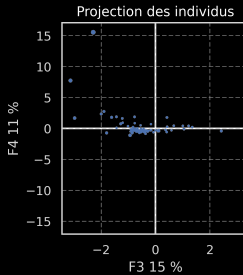
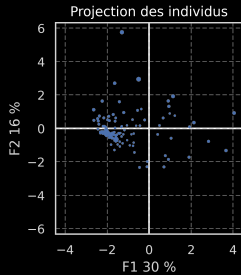
→ Hypothèse à priori **validée**

PCA - projection

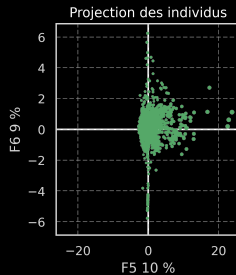
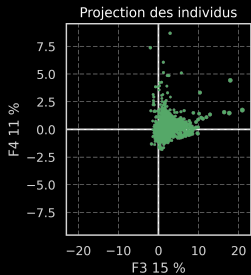
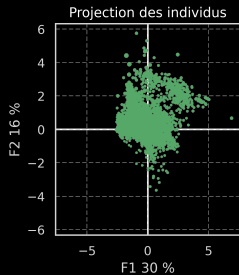
beverages



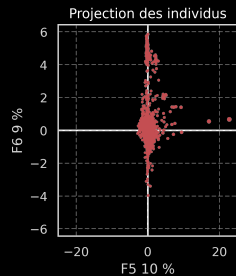
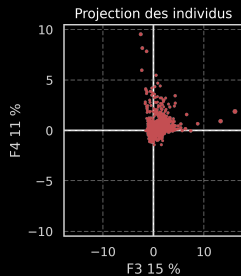
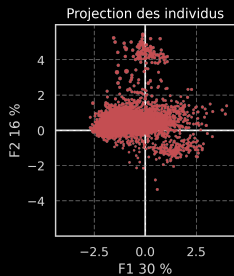
alcoholic beverages



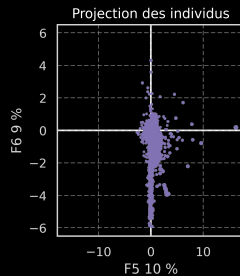
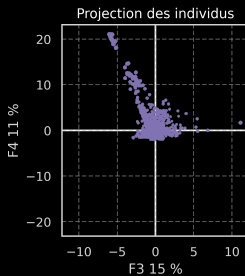
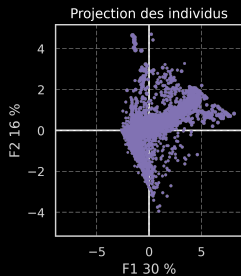
cereals and potatoes



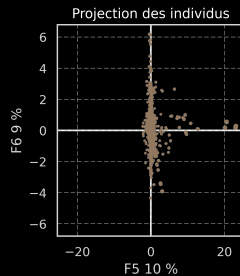
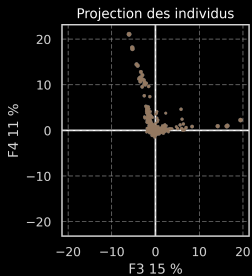
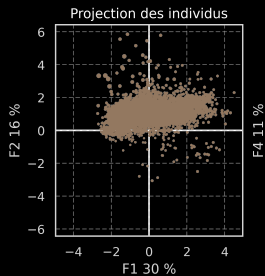
composite foods



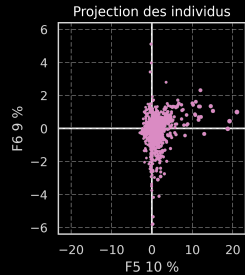
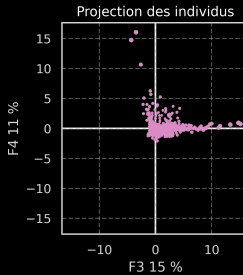
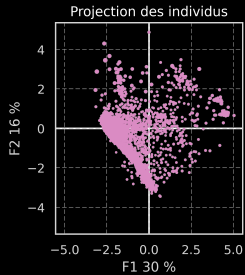
fat and sauces



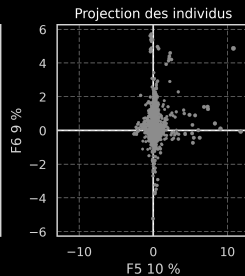
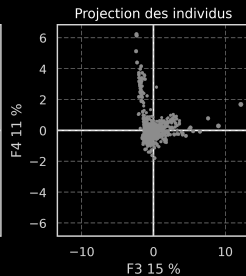
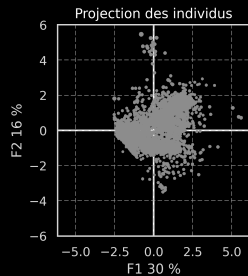
fish meat eggs



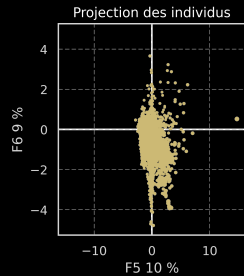
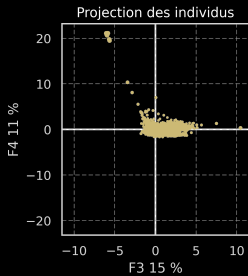
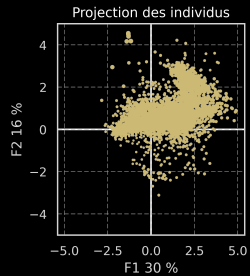
fruits and vegetables



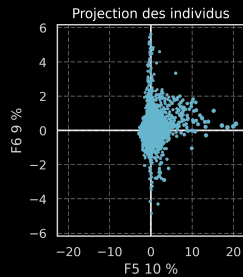
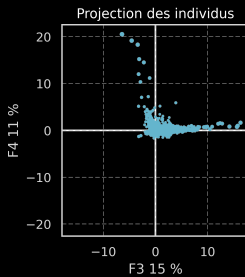
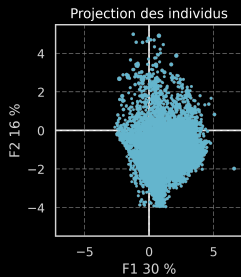
milk and dairy products



salty snacks



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 has been added to receive it.

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