OpenClassrooms – Parcours Data Scientist

Projet 3 :
Application en lien avec l'alimentation pour Santé publique France

Fanjamalala Rajaonalison

14/06/2021

Plan

1. Présentation de l’application
2. Opérations de nettoyage
3. Description et Analyse des variables
4. Analyse multivariée
5. Pertinence et Faisabilité de l'application

QUALECO

Application, qui après avoir identifié le produit, retourne deux scores

QUAL : qualité nutritionnelle et sanité

ECO : impact sur l’environnement

DEUX SCORES

QUAL

- Fournit une note allant de A à E

- Calculé à partir de :
  - → nutriscore
  - → nombre d'additifs
  - → nombre d'ingrédients issus de l'huile de palme

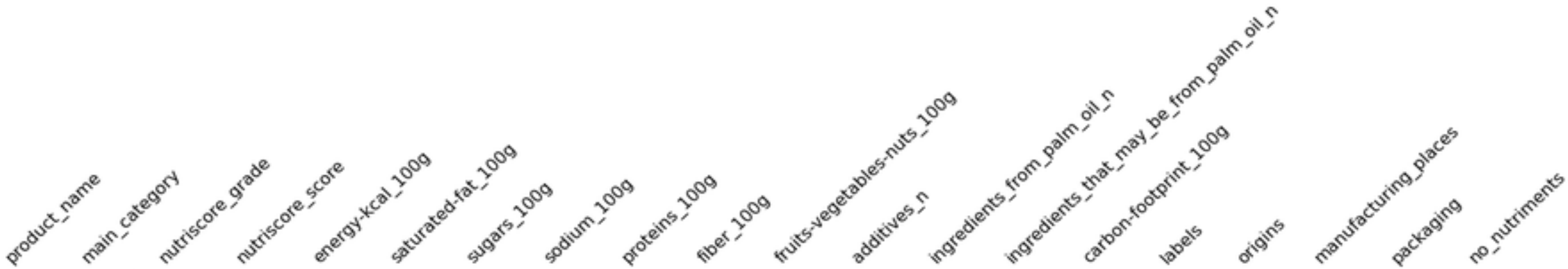
ECO

- Fournit une note allant de A à E

- Calculé à partir de :
  - → présence de label bio
  - → ingrédients locaux ou non
  - → produits transformés/fabriqués localement ou non
  - → empreinte carbone
  - → type d'emballage

1. Sélection de colonnes

186 colonnes → 20 colonnes



Produit

2 colonnes :
2 var. quali

QUAL

12 colonnes :
- 1 var. quali
- 11 var. quanti

ECO

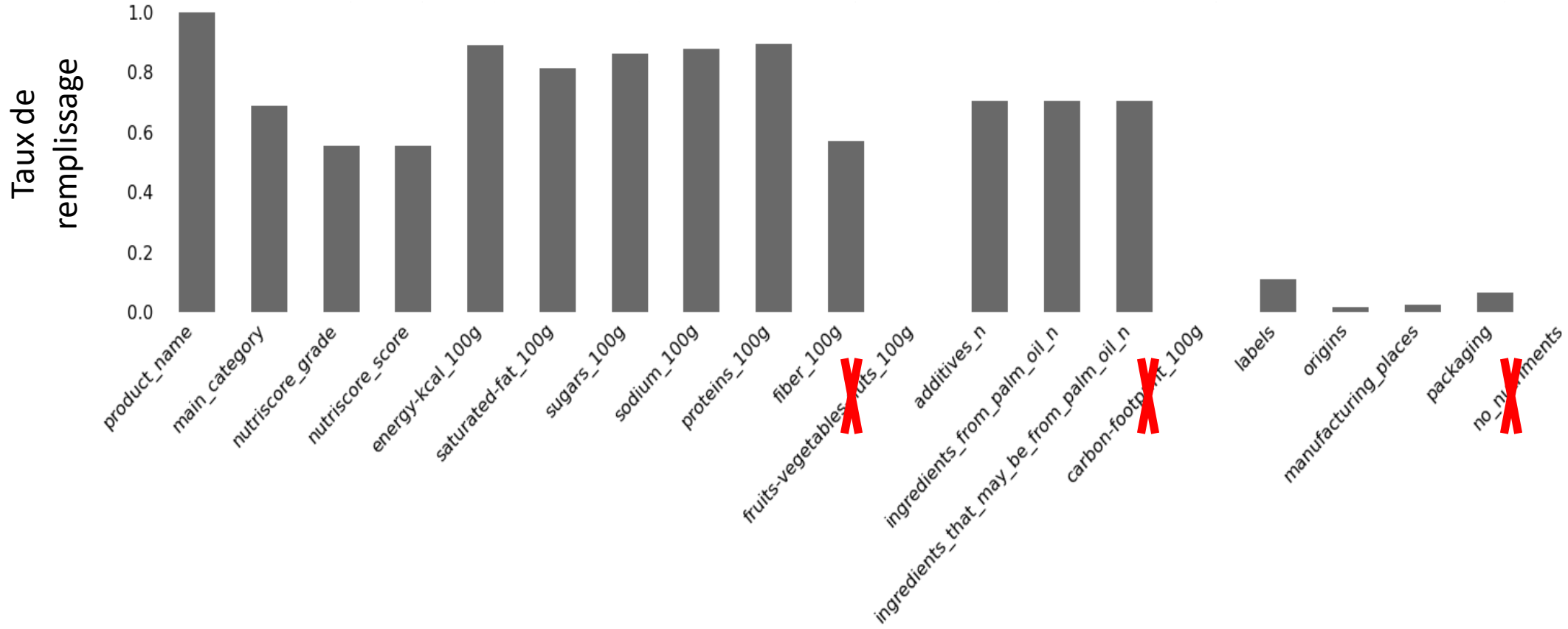
6 colonnes :
- 5 var. quali
- 1 var. quanti

1. Sélection de colonnes
2. Suppression des données non pertinentes
   - Lignes : Données dupliquées & Lignes vides

1. Sélection de colonnes

2. Suppression des données non pertinentes

- Lignes : Données dupliquées & Lignes vides
- Colonnes : Données avec des taux de remplissage < 5%

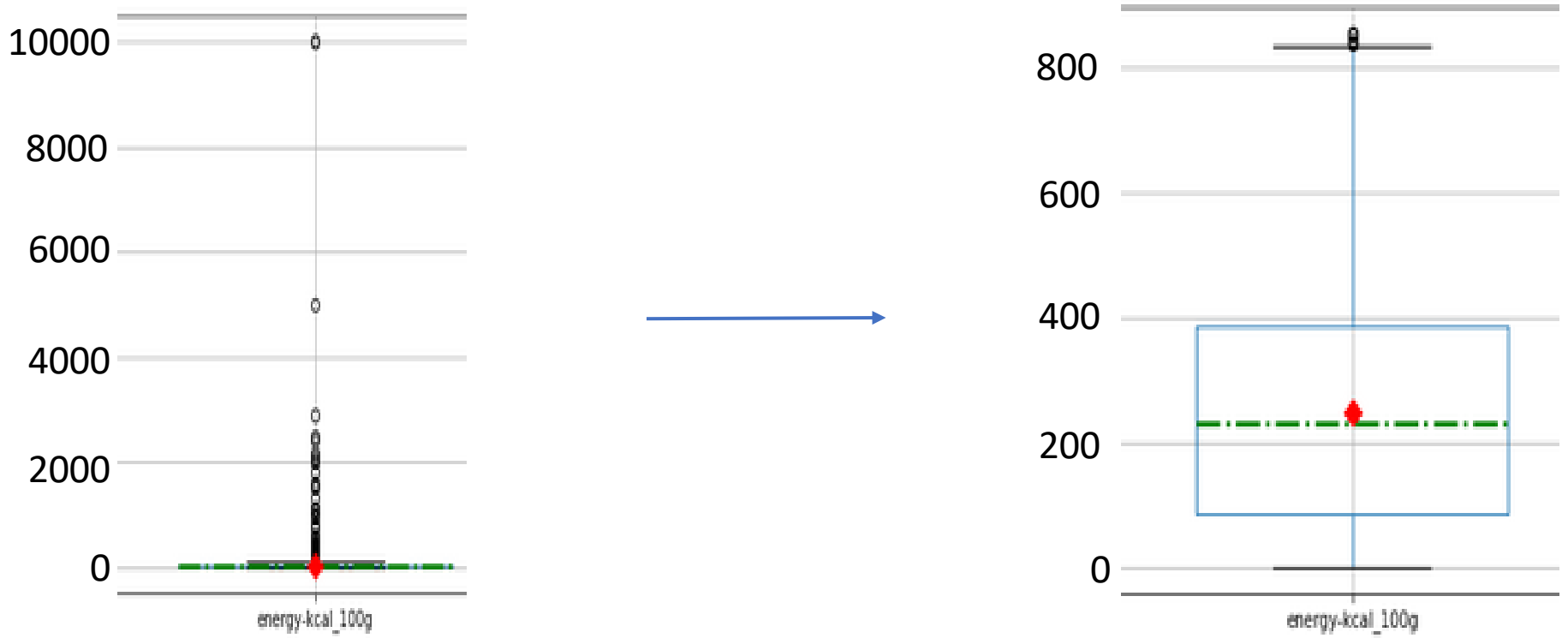


1. Sélection de colonnes
2. Suppression des données non pertinentes
3. Traitement des outliers
   - Identification des outliers : valeurs distantes des extrémités des moustaches

1. Sélection de colonnes
2. Suppression des données non pertinentes

3. Traitement des outliers

- Identification des outliers : valeurs distantes des extrémités des moustaches
- Remplacement des outliers par NaN



1. Sélection de colonnes
2. Suppression des données non pertinentes
3. Traitement des outliers

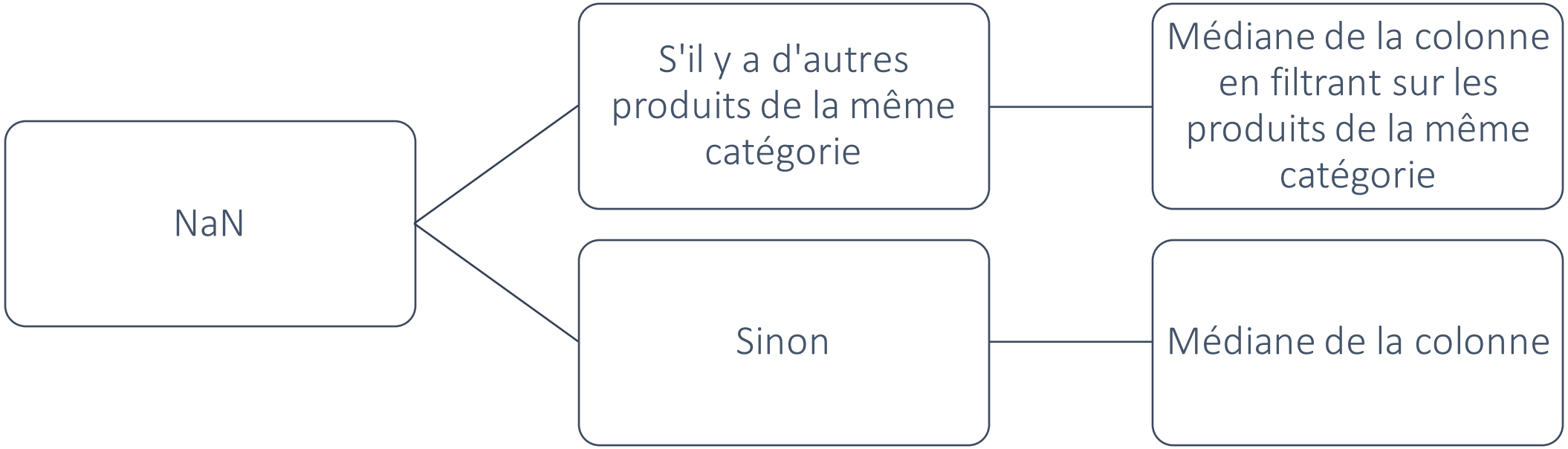
4. Traitement des valeurs manquantes

- Pour les variables qualitatives

1. Sélection de colonnes
2. Suppression des données non pertinentes
3. Traitement des outliers

4. Traitement des valeurs manquantes

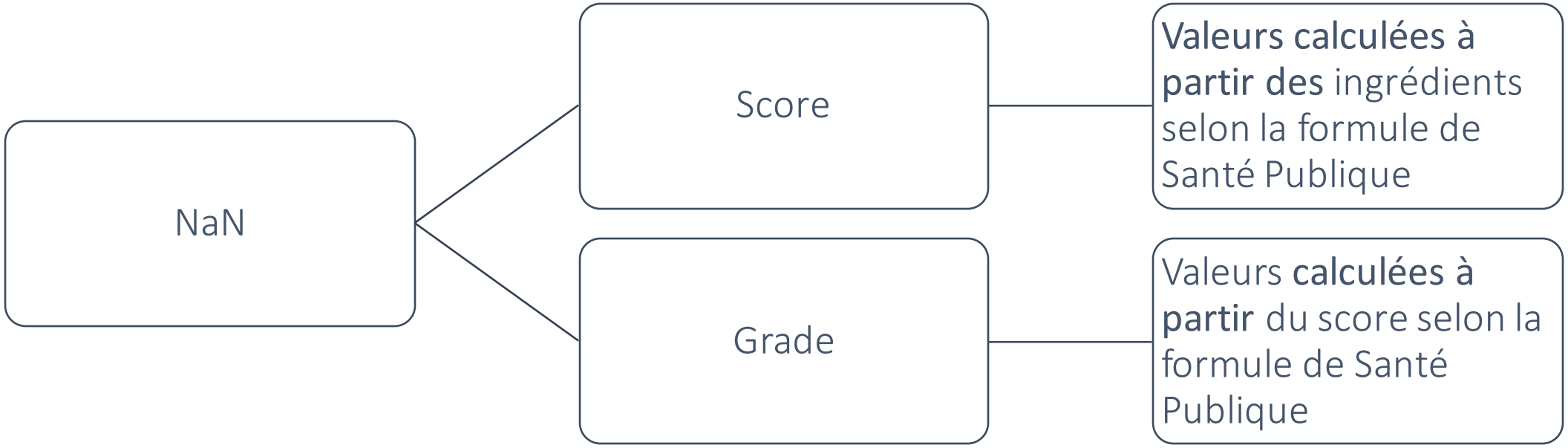
- Pour les variables qualitatives
- Pour les variables quantitatives



1. Sélection de colonnes
2. Suppression des données non pertinentes
3. Traitement des outliers

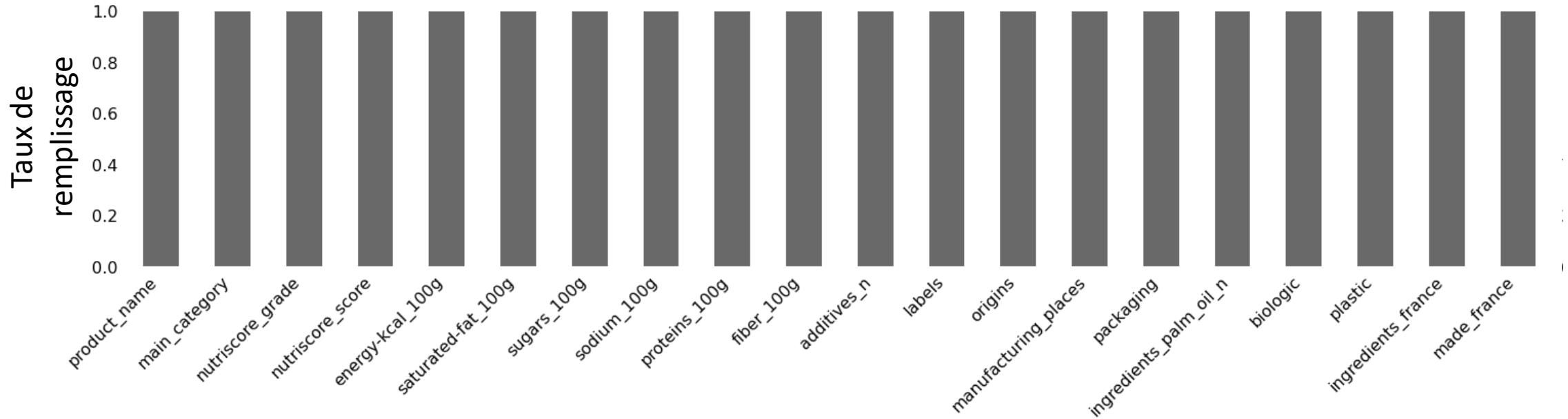
4. Traitement des valeurs manquantes

- Pour les variables qualitatives
- Pour les variables quantitatives
- Pour les variables nutriscores



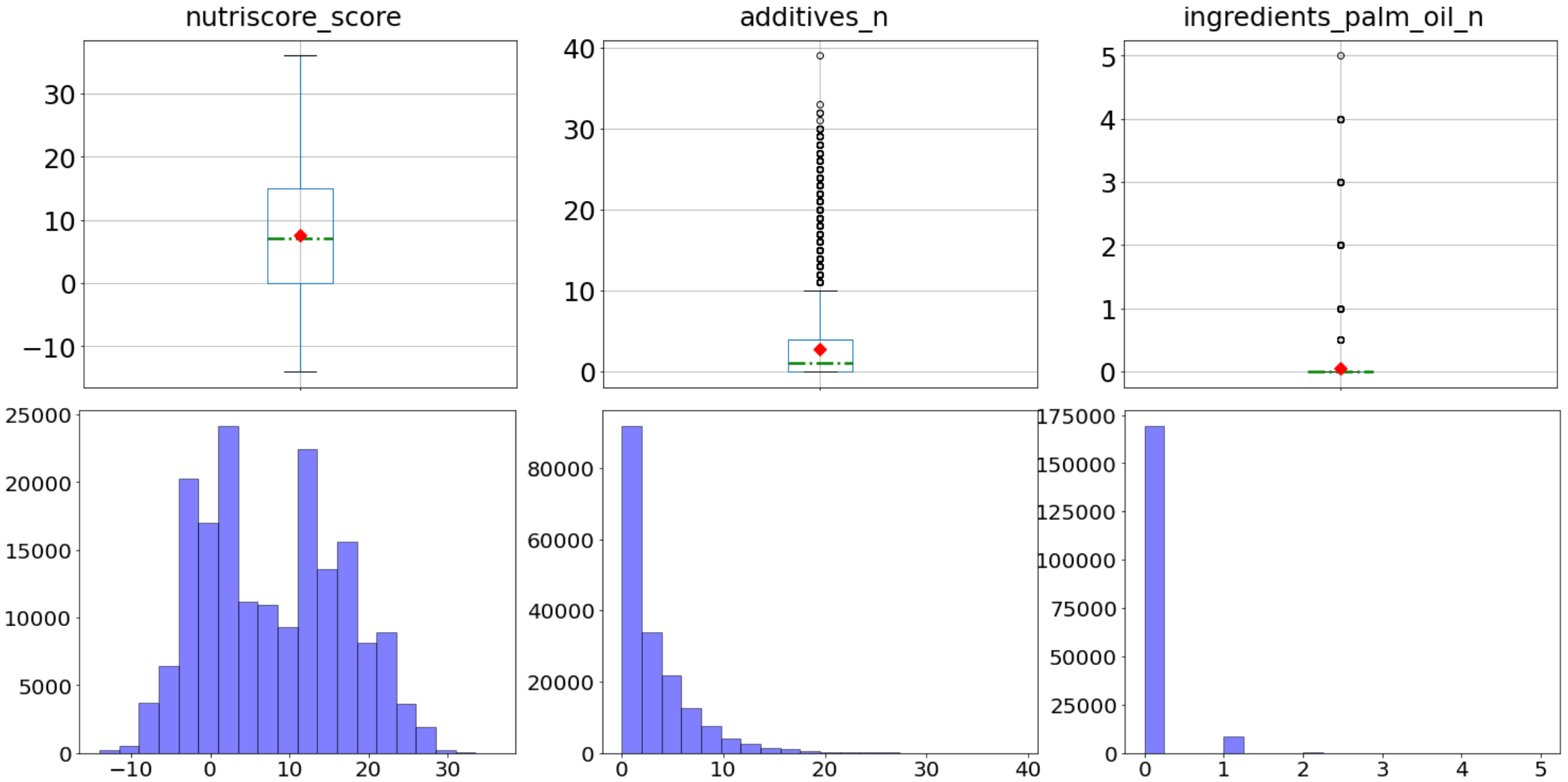
1. Sélection de colonnes
2. Suppression des données non pertinentes
3. Traitement des outliers
4. Traitement des valeurs manquantes

5. Vérification des données



→ Le jeu de donnée est prêt pour les étapes d'analyses.

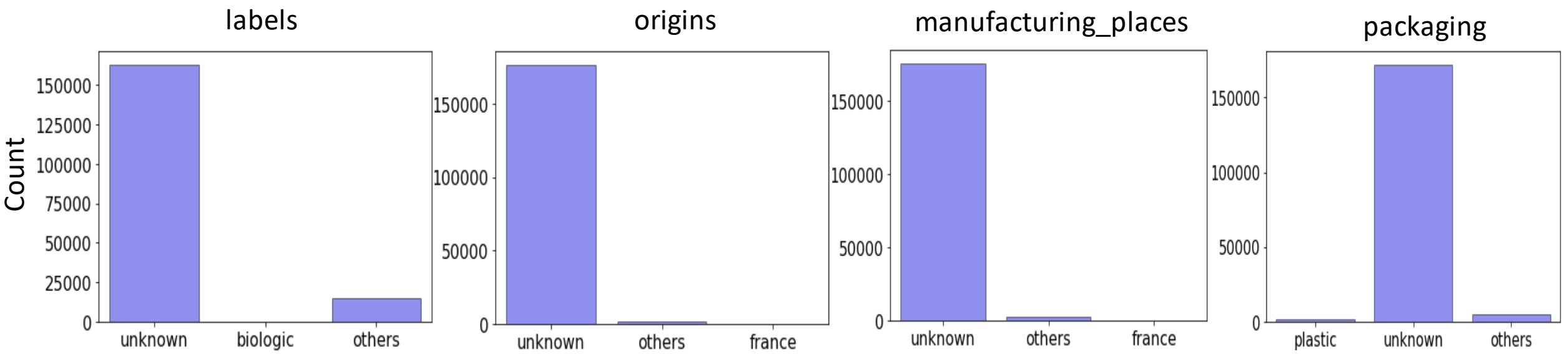
1. Variables quantitatives



→ La distribution des variables ne suivent pas la loi normale

1. Variables quantitatives

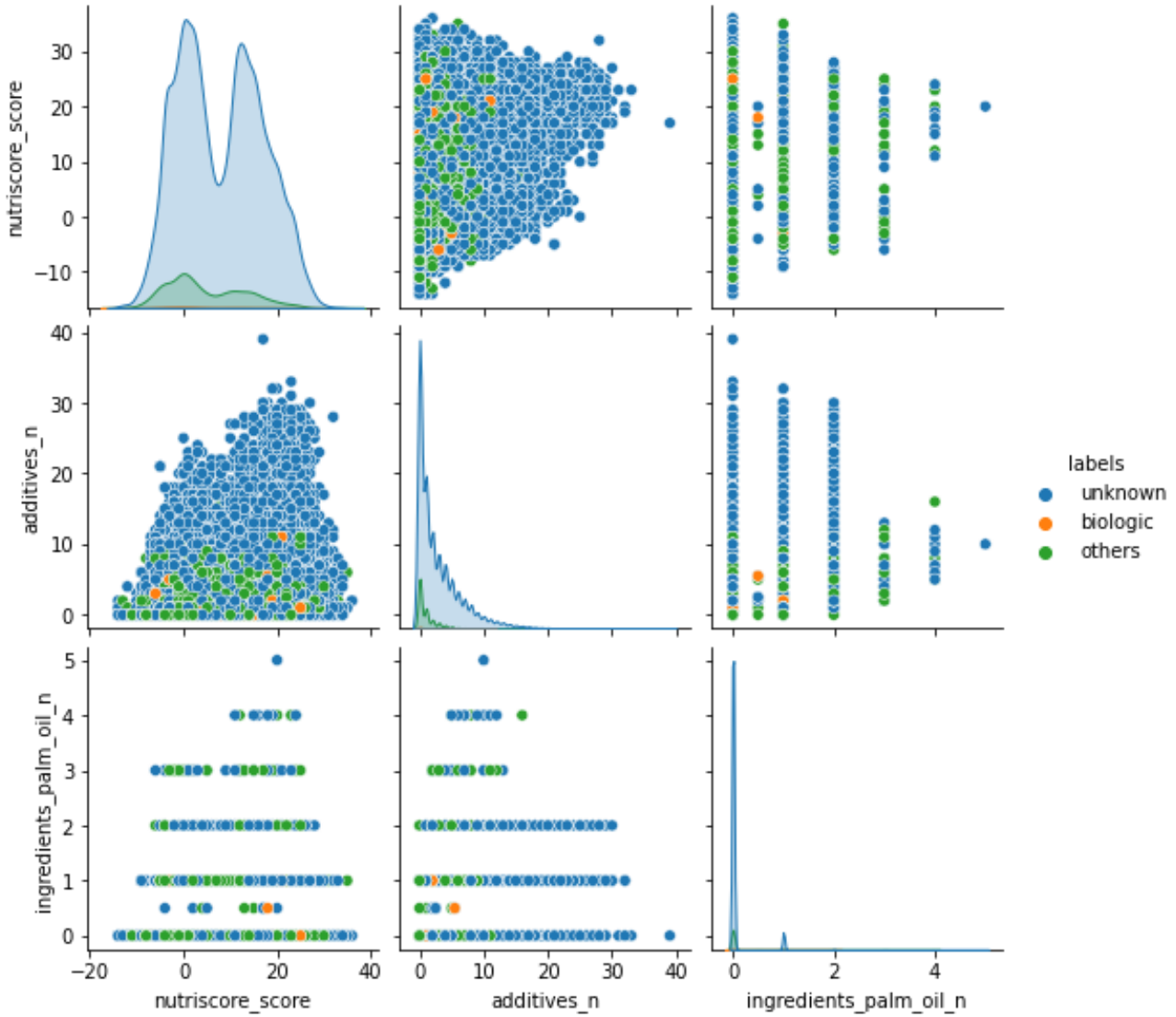
2. Variables qualitatives



→ Il y a beaucoup de données "inconnus" pour les 4 variables, largement plus que les données renseignées

→ Moins bonne qualité des données : réelle manque d'informations

1. Relation entre variables



→ les produits ayant des labels inconnus ont un plus fort nutriscore, un plus grand nombre d'additifs et d'ingrédients provenant de l'huile de palme

→ on ne note pas de relation flagrante entre les variables nutriscore et le nombre d'additifs

→ la variable N ingredients d'huile de palme ne montre aucune relation avec les autres variables

1. Relation entre variables

2. Corrélation

QUAL

ECO

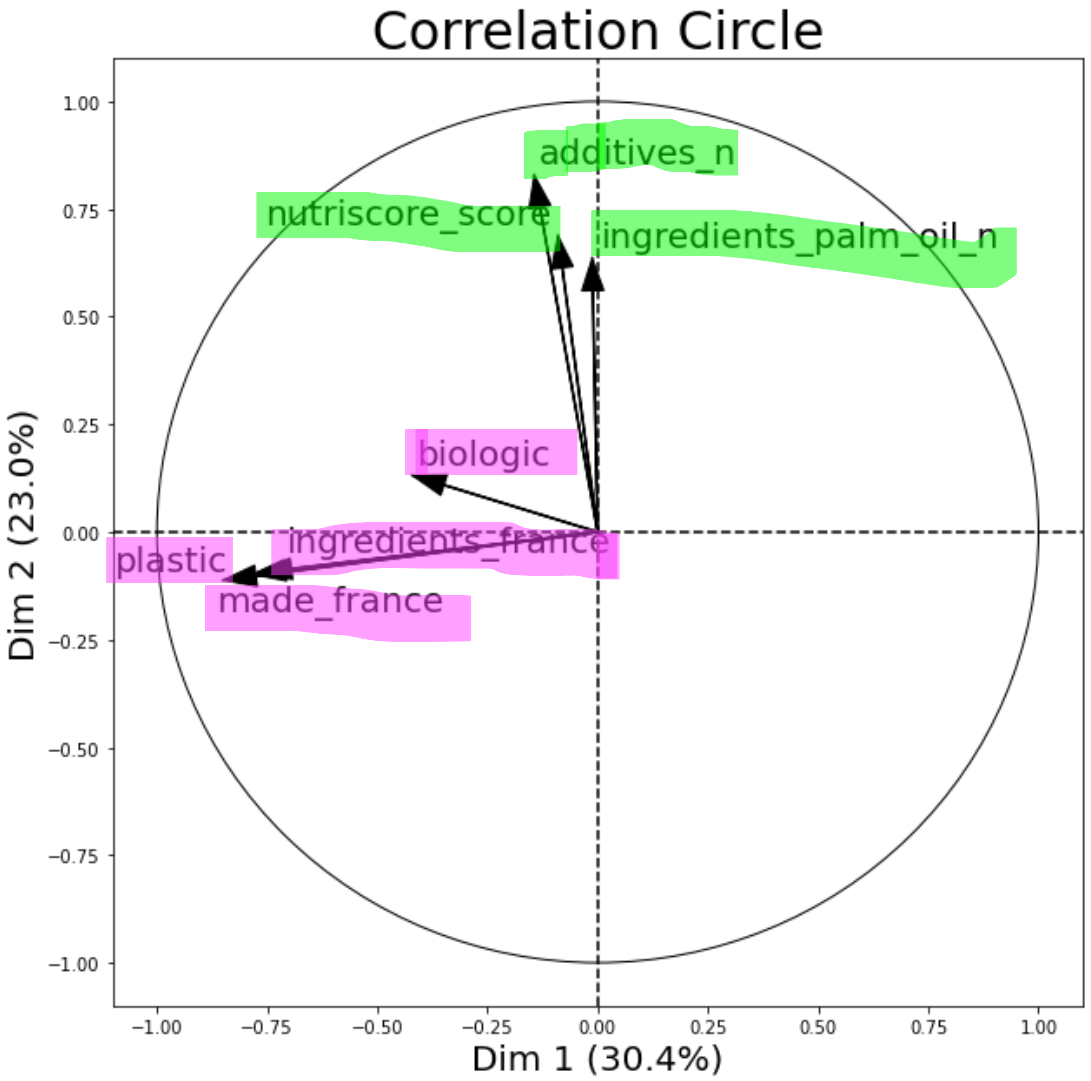
	nutriscore_score	additives_n	ingredients_palm_oil_n	biologic	plastic	ingredients_france	made_france
nutriscore_score	1.00	0.37	0.12	0.10	0.03	0.03	0.02
additives_n	0.37	1.00	0.31	0.14	0.05	0.04	0.04
ingredients_palm_oil_n	0.12	0.31	1.00	0.02	-0.03	-0.01	-0.02
biologic	0.10	0.14	0.02	1.00	0.26	0.15	0.21
plastic	0.03	0.05	-0.03	0.26	1.00	0.46	0.58
ingredients_france	0.03	0.04	-0.01	0.15	0.46	1.00	0.60
made_france	0.02	0.04	-0.02	0.21	0.58	0.60	1.00

→ on note une relation faible pour les variables définissant QUAL

→ les relations sont plus significatives pour les variables définissant ECO

→ il semble ne pas avoir de relation entre les variables définissant QUAL et celles définissant ECO

1. Existence de 2 groupes de variables



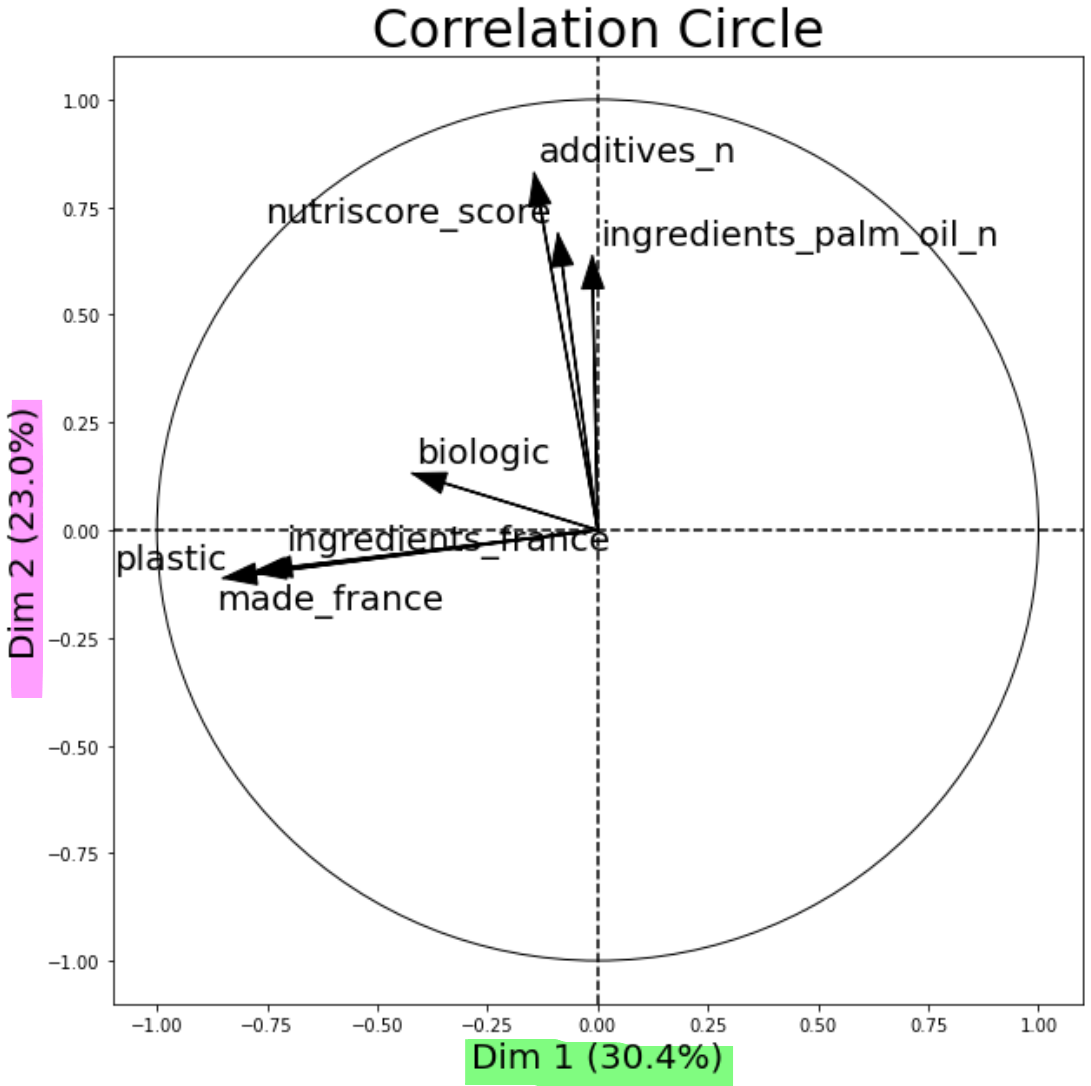
→ on note 2 groupes de variables corrélées entre elles

→ possibilité de résumer ces 2 groupes de variables par 2 variables synthétiques qui pourront être les 2 scores de l'application : QUAL et ECO

(1ère possibilité de définir de QUAL et ECO)

1. Existence de 2 groupes de variables

2. Pertinence de 2 axes principales



→ les 2 groupes de variables identifiés précédemment sont respectivement corrélés aux 2 axes principaux

→ ces 2 axes expliquent plus de la moitié (53,4%) de la variance des données

→ ces 2 axes pourraient être les 2scores de l'application : QUAL et ECO

(2ème possibilité de définir de QUAL et ECO)

1. Existence de 2 groupes de variables
2. Pertinence de 2 axes principales

3. Contribution des variables à ces scores

	ECO	QUAL
nutriscore_score	-0.107672	0.536689
additives_n	-0.130236	0.639554
ingredients_palm_oil_n	-0.026127	0.487845
biologic	-0.294329	0.144872
plastic	-0.529990	-0.103531
ingredients_france	-0.528233	-0.123589
made_france	-0.569398	-0.134021

→ les variables qui contribuent le plus aux scores ECO et QUAL sont ceux attendus

→ ces 2 scores ECO et QUAL peuvent être définis par la combinaison linéaire des variables y participant

QUAL = 0.54 * nutriscore
+ 0.64 * N additifs
+ 0.49 * N ingredients huile de palme

ECO = - 0.29 * bio
- 0.53 * plastic
- 0.53 * ingredients de France
- 0.57 * produit en France

Synthèse & Conclusion

Conception de l'application QUALECO permettant de calculer 2 scores QUAL et ECO

Synthèse & Conclusion

Conception de l'application QUALECO permettant de calculer 2 scores QUAL et ECO

Faisabilité à partir des données de Santé Publique?

Synthèse & Conclusion

Conception de l'application QUALECO permettant de calculer 2 scores QUAL et ECO

Faisabilité à partir des données de Santé Publique?

- Variables pertinentes et liées en 2 groupes
- Définition de 2 axes possible
- Contribution des variables aux 2 axes

Synthèse & Conclusion

Conception de l'application QUALECO permettant de calculer 2 scores QUAL et ECO

Faisabilité à partir des données de Santé Publique?

- Variables pertinentes et liées en 2 groupes
- Définition de 2 axes
- Contribution des variables aux 2 axes

Calcul des 2 scores de l'application QUALECO faisable

...mais peut encore être amélioré

MERCI

Questions

Incompréhensions