

Application digitale dans le domaine de l'alimentation des sportifs de haut niveau.

Analyses exploratoire et pre-
processing.

Data set « Open Food Facts ».

Sommaire.



Introduction.



Phase exploratoire.

Chargement des données, vue synthétique.

Analyse valeurs manquantes, data sets numérique et catégoriel.

Distribution des données, valeurs remarquables.

Valeurs aberrantes: volumétrie, remplacement des valeurs, visualisation.



Pre-processing.

Imputation, encodage, remplacement valeurs manquantes.

- KNN.
- Iterative imputing.
- Simple imputing.



Indépendance des variables.



Data set final.

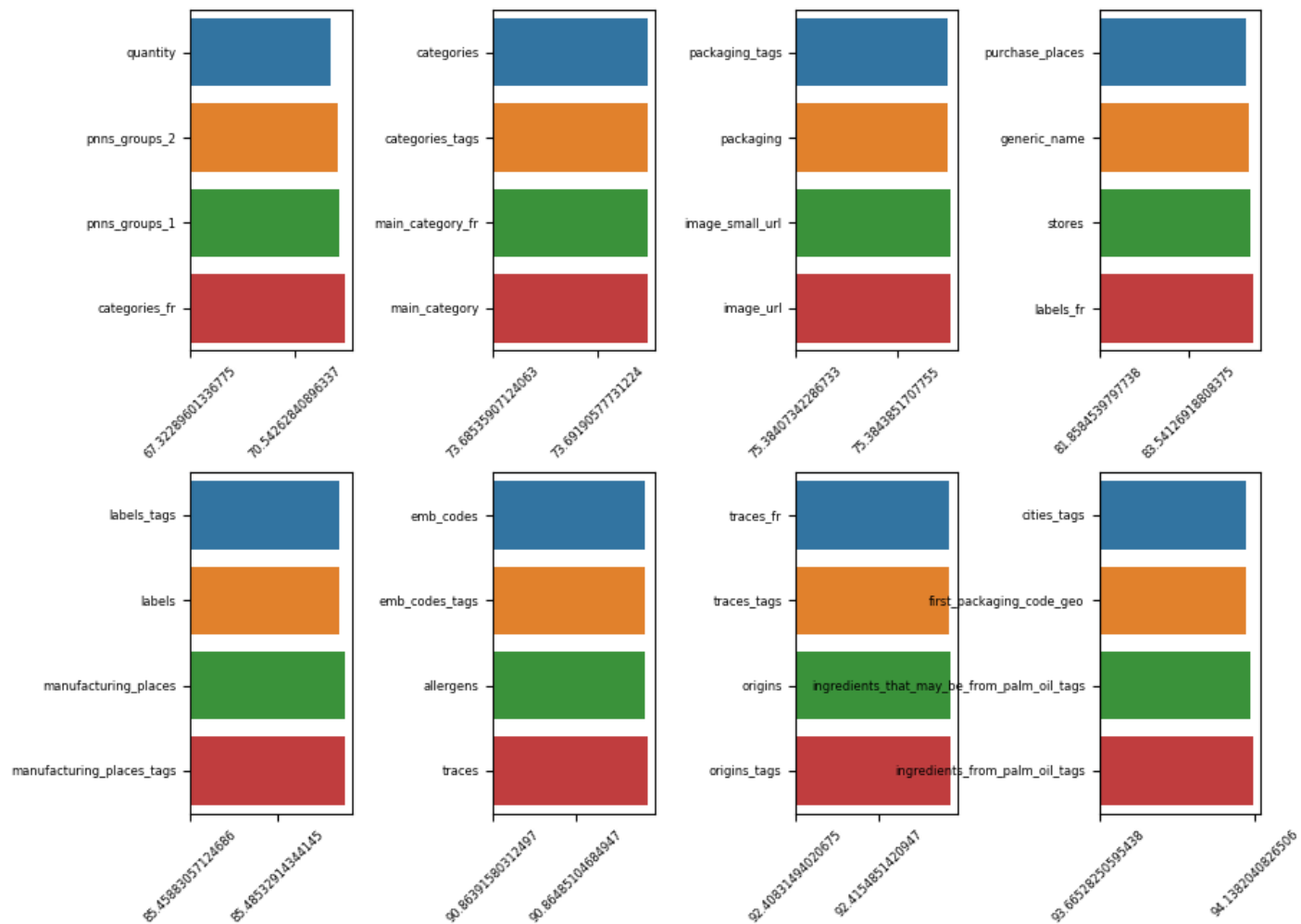
Introduction.

- Le data set Open Food Facts fournit des données précises sur un ensemble de produits de consommation alimentaire.
- Outre les code, nom de produit, marque, pays de production, date de production, le data set comporte aussi les éléments énergétiques et nutritifs.
- Notamment les éléments de teneur en graisse, sucre, carbohydrates, fibre, protéine, sel, sodium, vitamine A, vitamine C, calcium, fer sont essentiels à la nutrition d'un sportif de haut niveau.
- L'objectif ici est de s'appuyer sur ce data set pour fournir à l'athlète une information, le code Nutri-Sport (échelonné de A à E), lorsqu'il soumet à l'application le code barre d'un produit.
- L'application pourra également proposer des produits similaires.
- La qualité du data set n'est pas nécessairement suffisante. Des traitements préalables seront nécessaires à son exploitation.

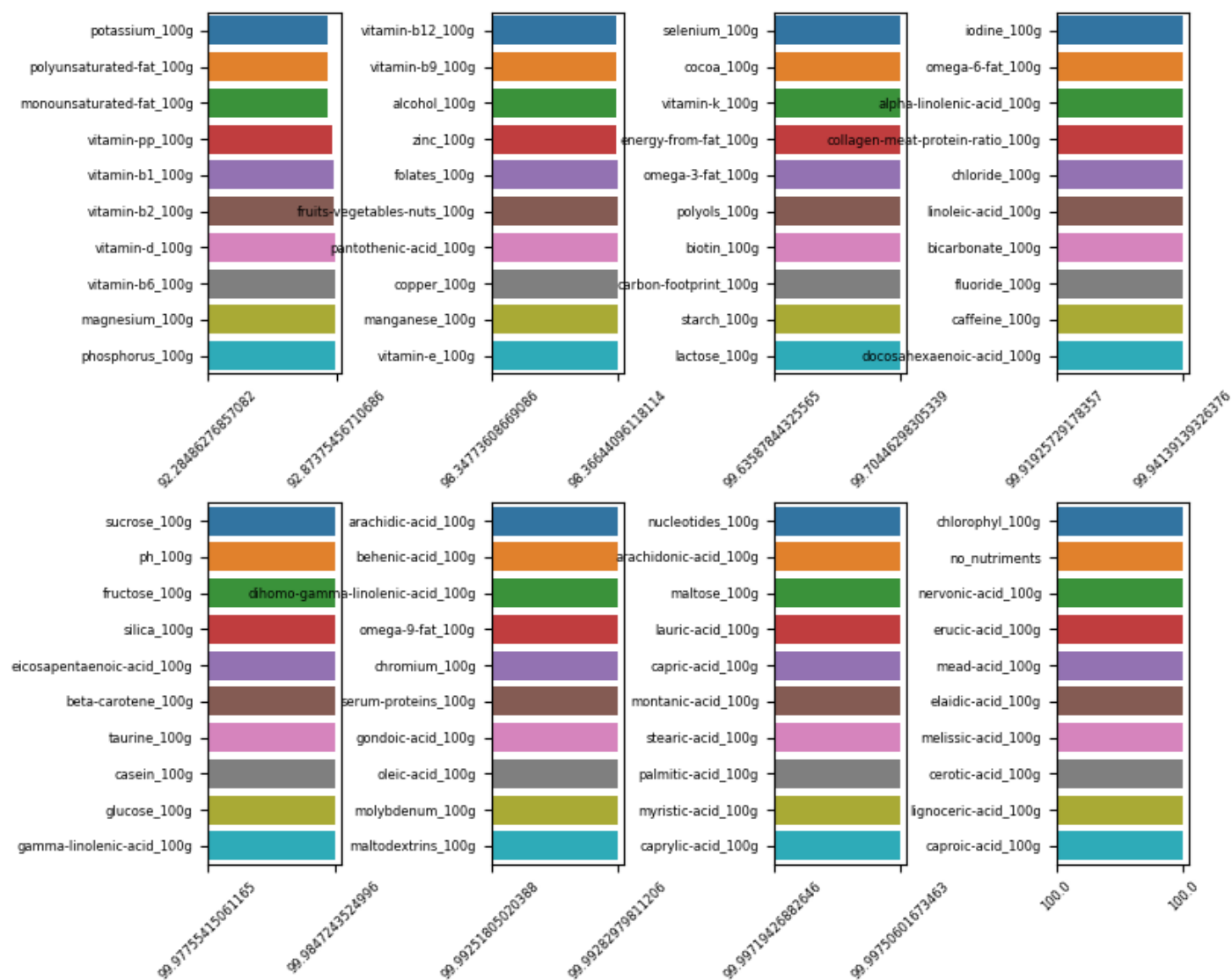
Phase exploratoire.

- Dimensions.
 - 320772 lignes.
 - 162 variables: 106 de type numérique, 22 de type catégoriel.
- Valeurs manquantes.
 - Certaines variables ont des taux de valeurs manquantes supérieurs à 60%. Cf. slides suivants.
 - Pour ces variables, on juge qu'elles sont insuffisamment informatives. Exclusion du data set.

Taux de
valeurs
manquantes
par variable
catégorielle.



Taux de
valeurs
manquantes
par variable
numérique.



Phase exploratoire.

Après exclusion des variables non informatives.

```
df_cat.isnull().mean()*100
```

code	0.007170
url	0.007170
creator	0.000623
created_t	0.000935
created_datetime	0.002806
last_modified_t	0.000000
last_modified_datetime	0.000000
product_name	5.537266
brands	8.857382
brands_tags	8.859876
countries	0.087289
countries_tags	0.087289
countries_fr	0.087289
ingredients_text	22.386617
serving_size	34.118003
additives	22.404387
additives_tags	51.778834
additives_fr	51.778834
nutrition_grade_fr	31.038245
status	0.014240

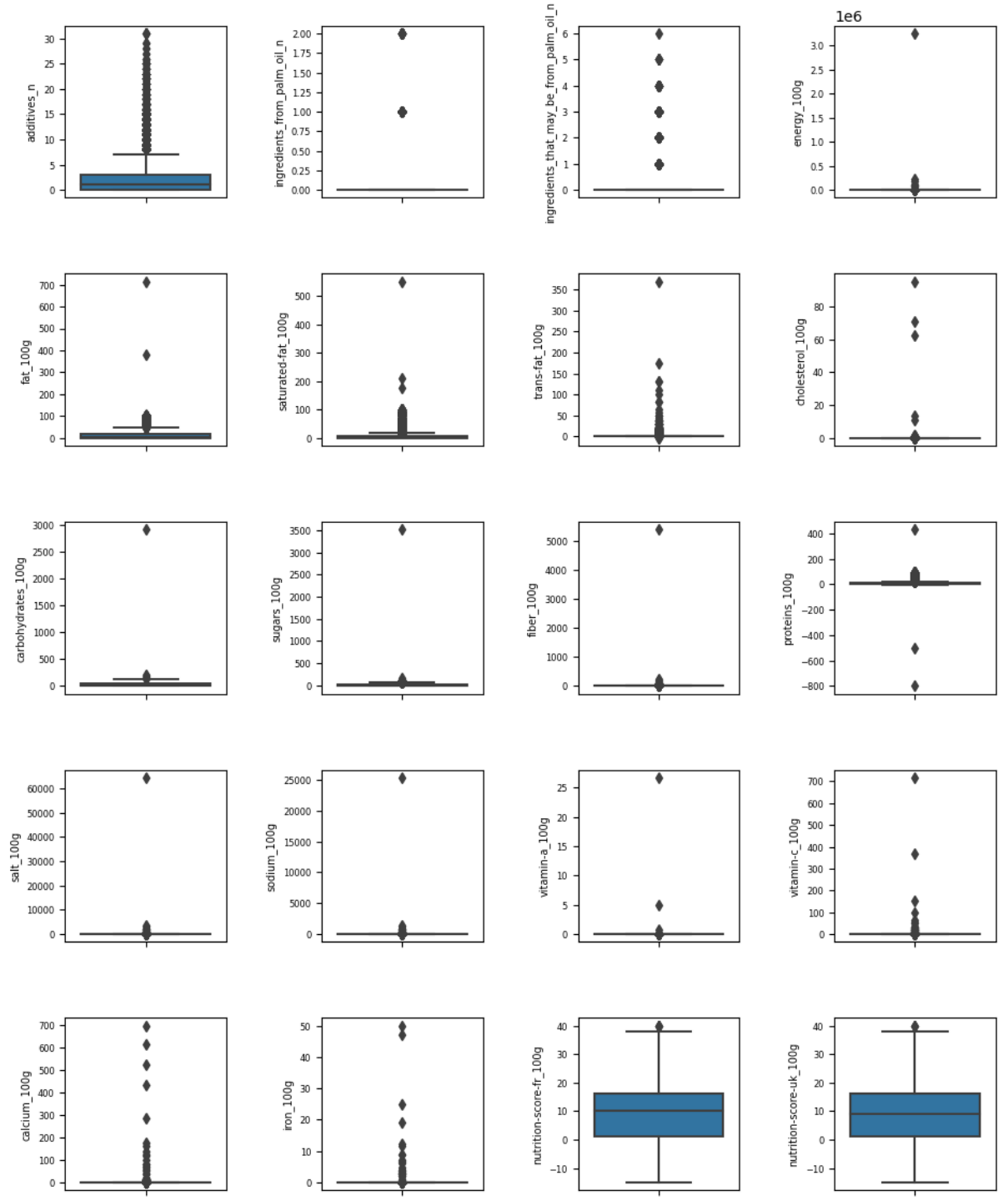
```
df_num.isnull().mean()*100
```

additives_n	22.393787
ingredients_from_palm_oil_n	22.393787
ingredients_that_may_be_from_palm_oil_n	22.393787
energy_100g	18.598568
fat_100g	23.967491
saturated-fat_100g	28.437021
trans-fat_100g	55.327148
cholesterol_100g	55.080244
carbohydrates_100g	24.061951
sugars_100g	23.630803
fiber_100g	37.374210
proteins_100g	18.969860
salt_100g	20.345292
sodium_100g	20.359944
vitamin-a_100g	57.117828
vitamin-c_100g	56.085007
calcium_100g	56.027958
iron_100g	56.211265
nutrition-score-fr_100g	31.038245
nutrition-score-uk_100g	31.038245
dtype: float64	

Phase exploratoire.

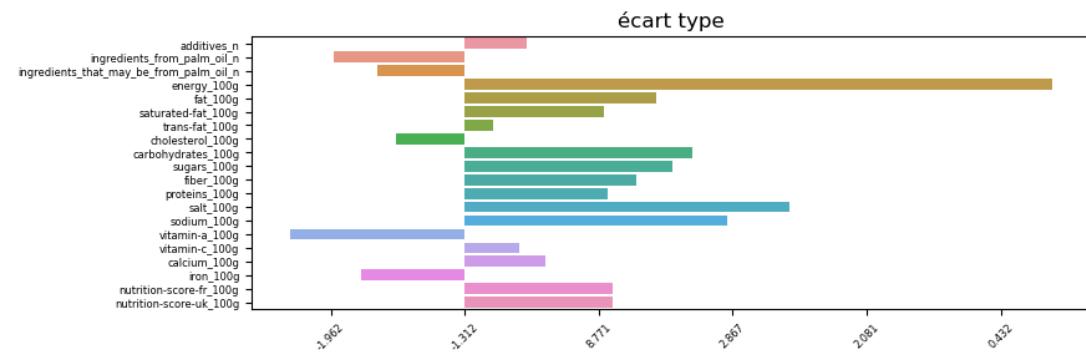
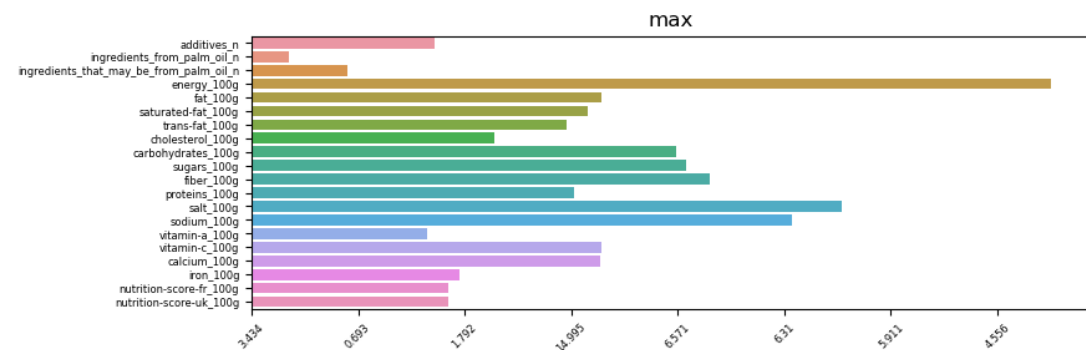
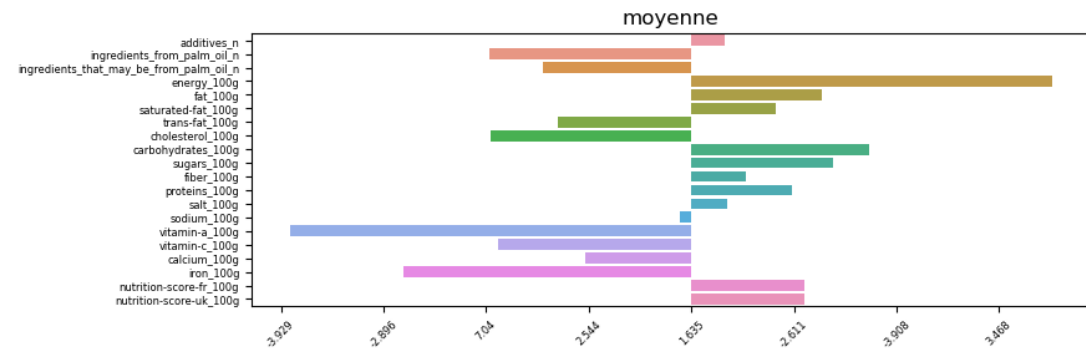
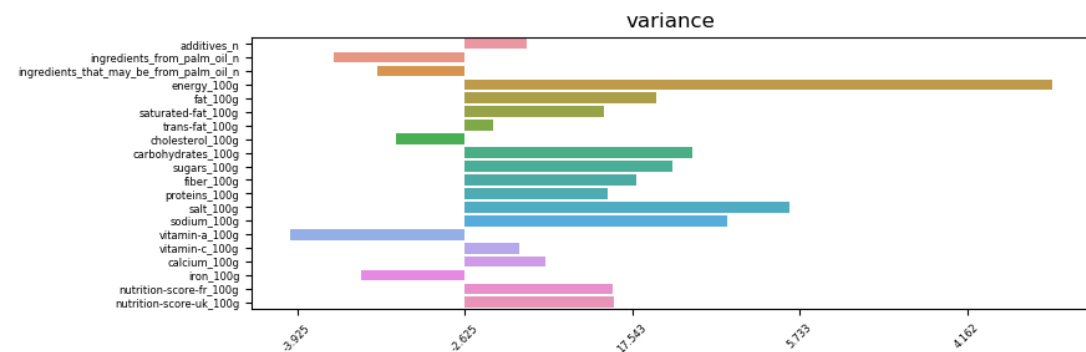
Distribution des données.

- Les données sont écartées de la moyenne.
- Existence de valeurs au delà du dernier quartile (outliers).



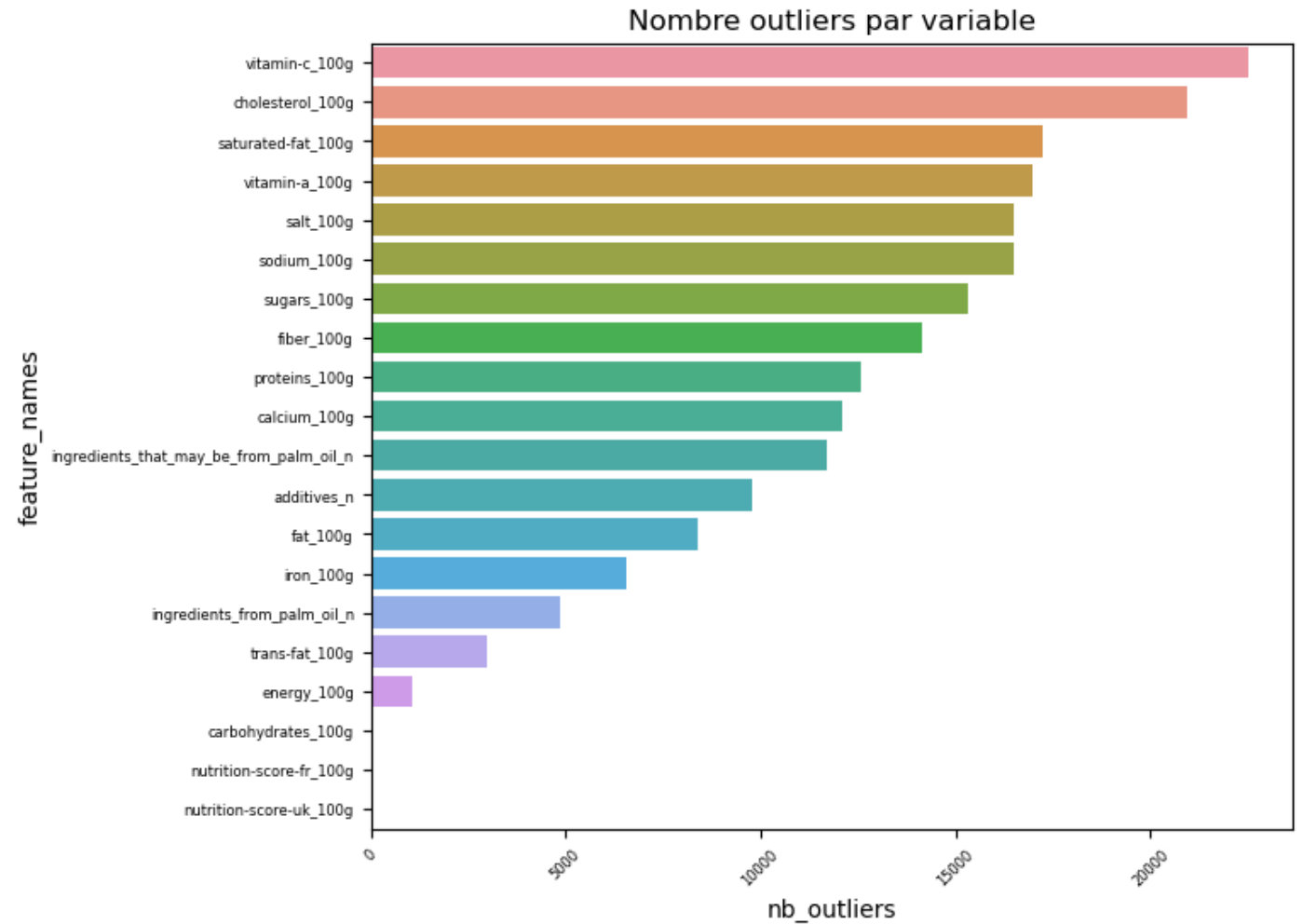
Phase exploratoire.

- Valeurs remarquables: variance, moyenne, max, écart-type.
 - Cf. slide suivant.



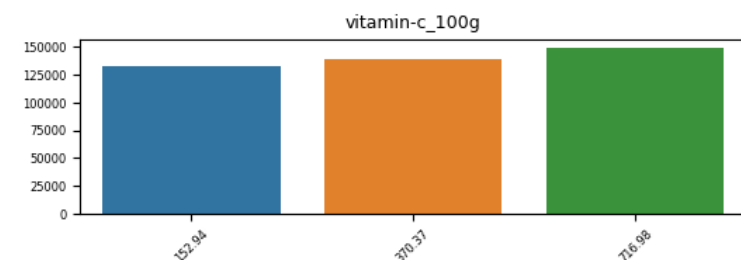
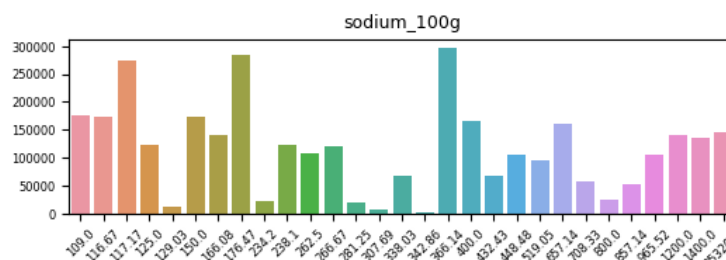
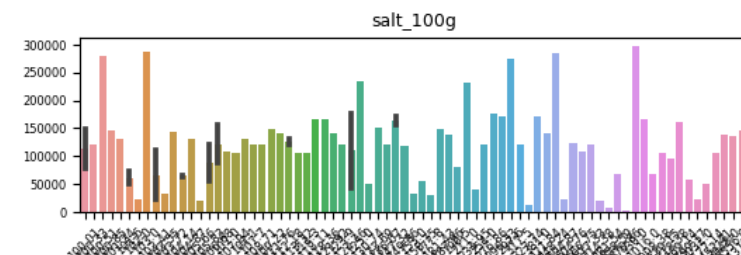
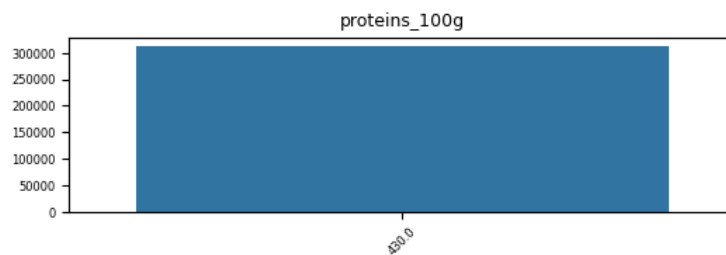
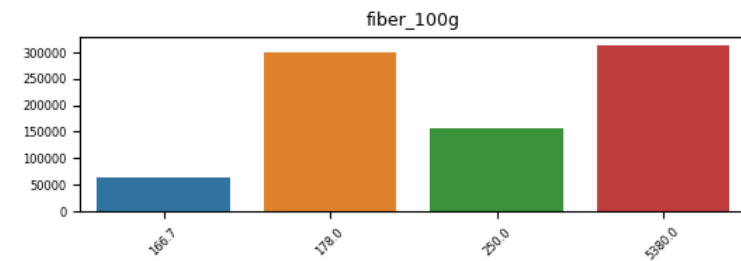
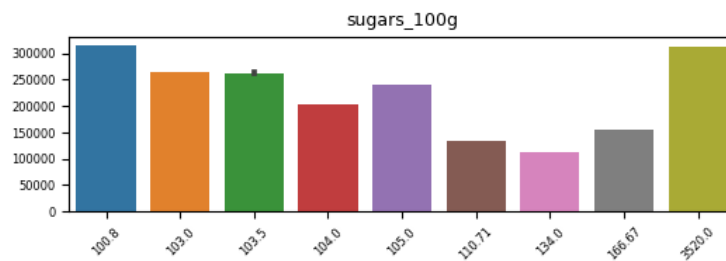
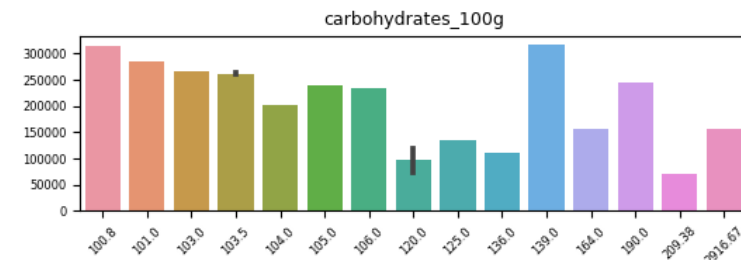
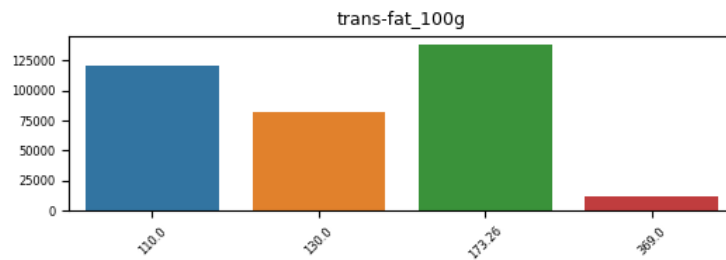
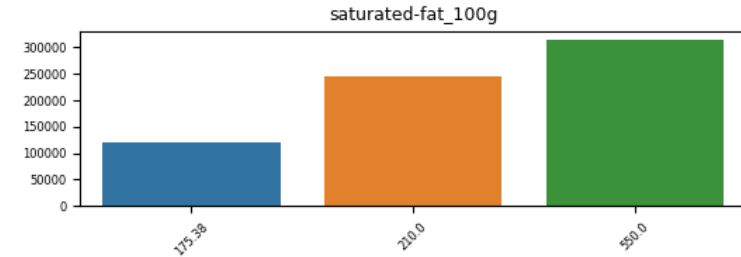
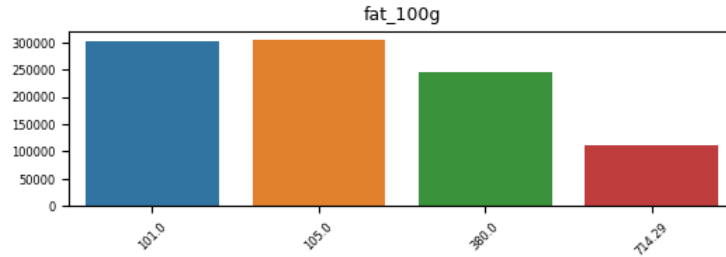
Phase exploratoire.

- Valeurs aberrantes.
 - Volumétrie: volume parfois important de valeurs situées au delà du dernier quartile.
 - Causes: des valeurs erronées (valeur saisie: 350g au lieu de 50g, par ex.)



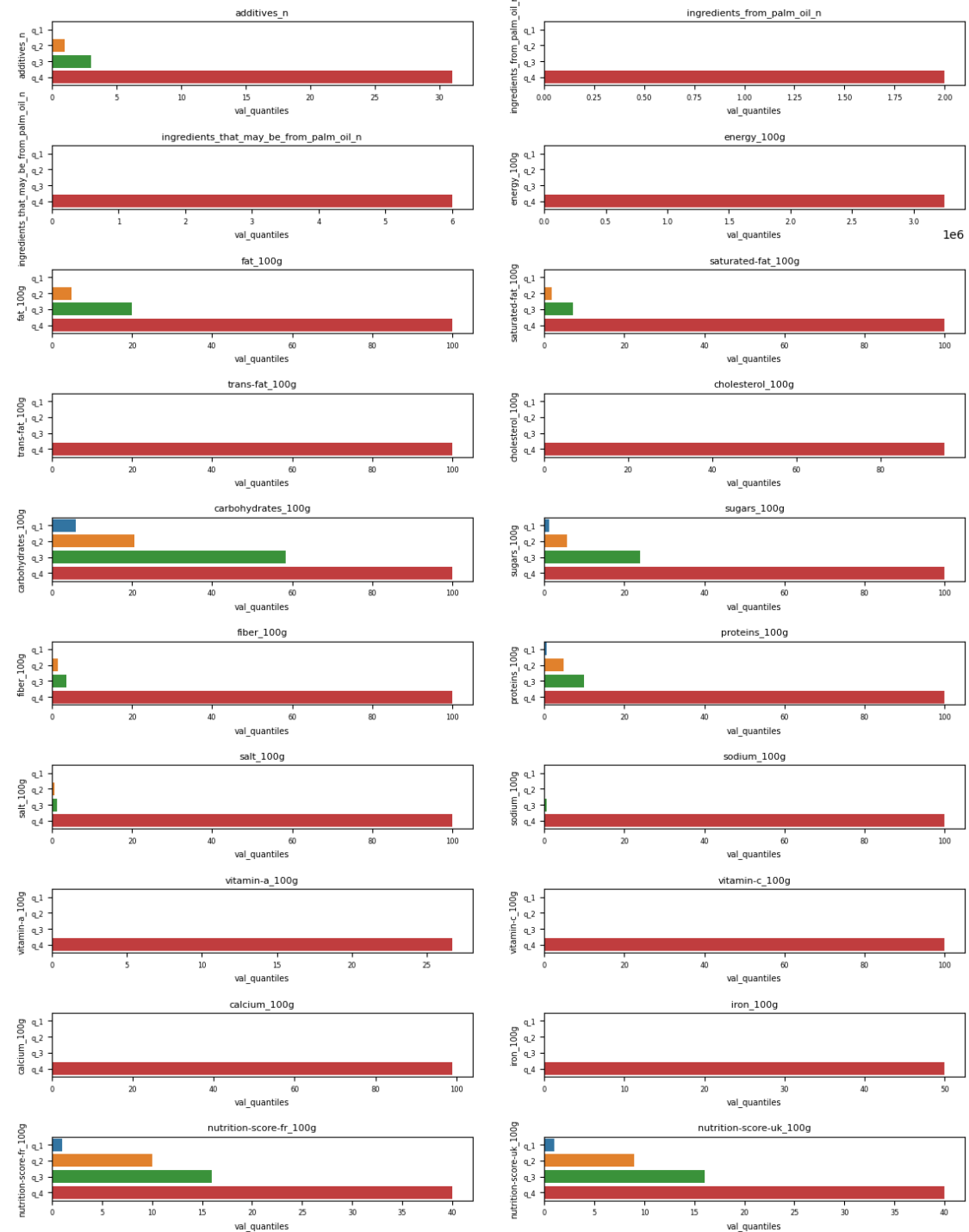
Phase exploratoire.

- Variables avec des erreurs de saisie.
- Remplacement des valeurs erronées par la médiane (compte tenu de l'étalement des valeurs).



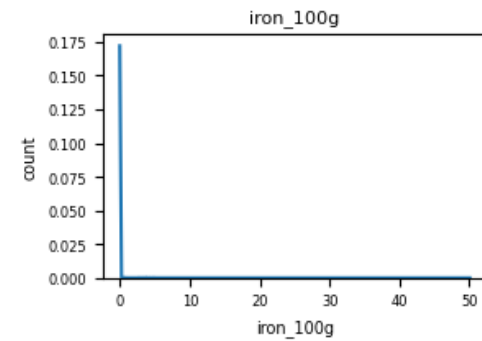
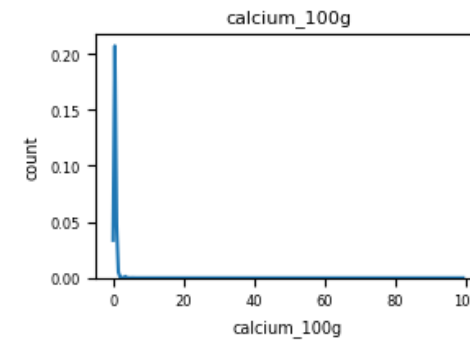
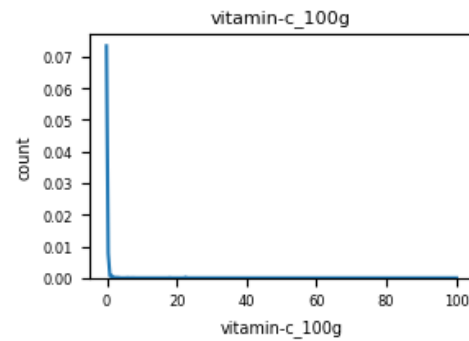
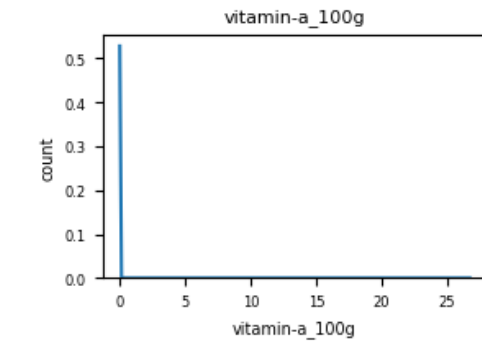
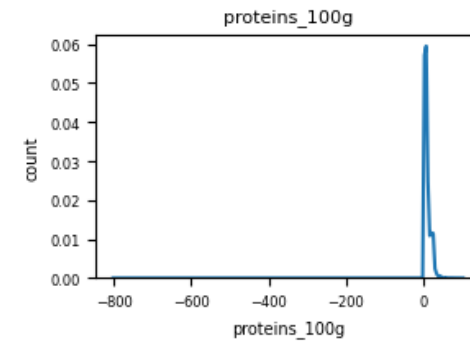
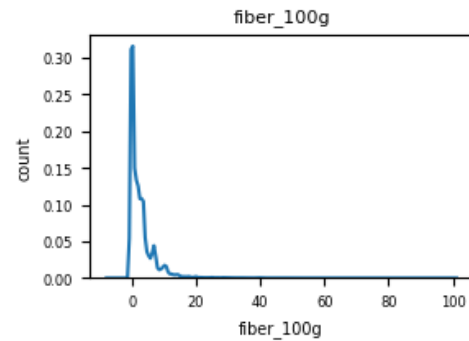
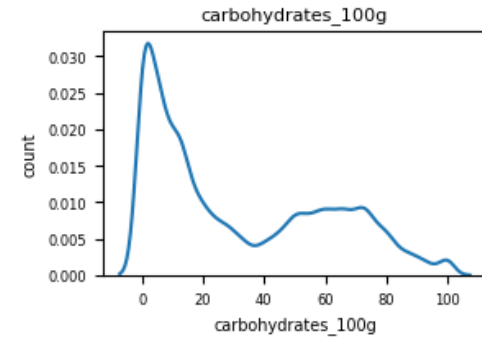
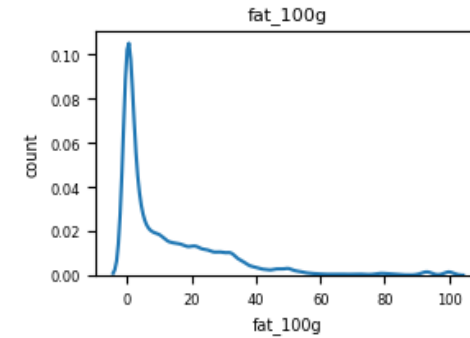
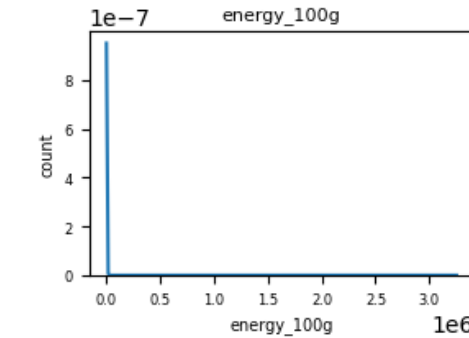
Phase exploratoire.

Visualisation des quantiles après correction des erreurs de saisie.



Phase exploratoire.

Visualisation des courbes de fréquence par variable numérique.



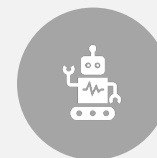
Pre- processing.

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 320772 entries, 0 to 320771  
Data columns (total 42 columns):
```

#	Column	Non-Null Count	Dtype
0	additives_n	248939 non-null	float64
1	ingredients_from_palm_oil_n	248939 non-null	float64
2	ingredients_that_may_be_from_palm_oil_n	248939 non-null	float64
3	energy_100g	261113 non-null	float64
4	fat_100g	243891 non-null	float64
5	saturated-fat_100g	229554 non-null	float64
6	trans-fat_100g	143298 non-null	float64
7	cholesterol_100g	144090 non-null	float64
8	carbohydrates_100g	243588 non-null	float64
9	sugars_100g	244971 non-null	float64
10	fiber_100g	200886 non-null	float64
11	proteins_100g	259922 non-null	float64
12	salt_100g	255510 non-null	float64
13	sodium_100g	255463 non-null	float64

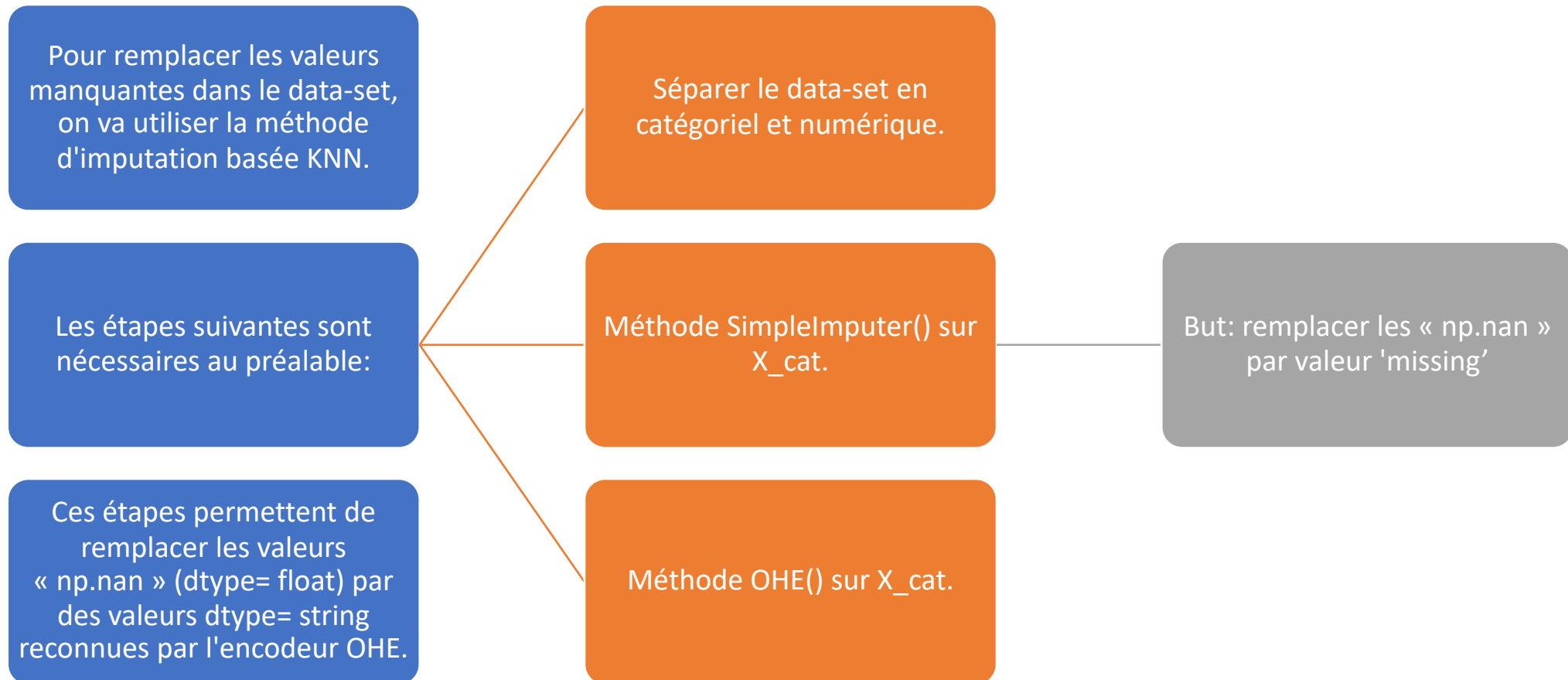


Après ces premières étapes,
le data set doit être traité
pour le remplacement des
valeurs manquantes.

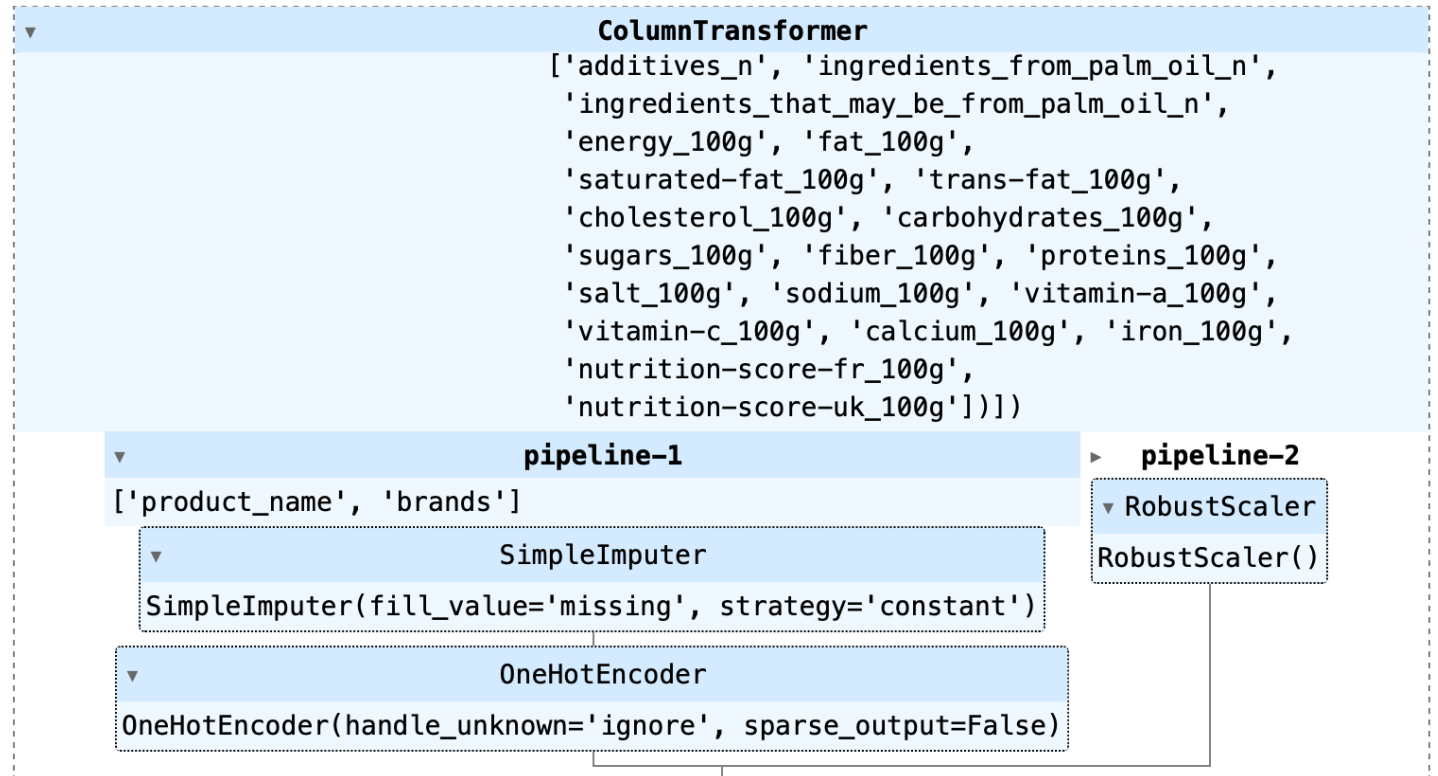


Trois méthodes utilisées:
KNN, SimpleImputer,
IterativeImputer.

Pre-processing.



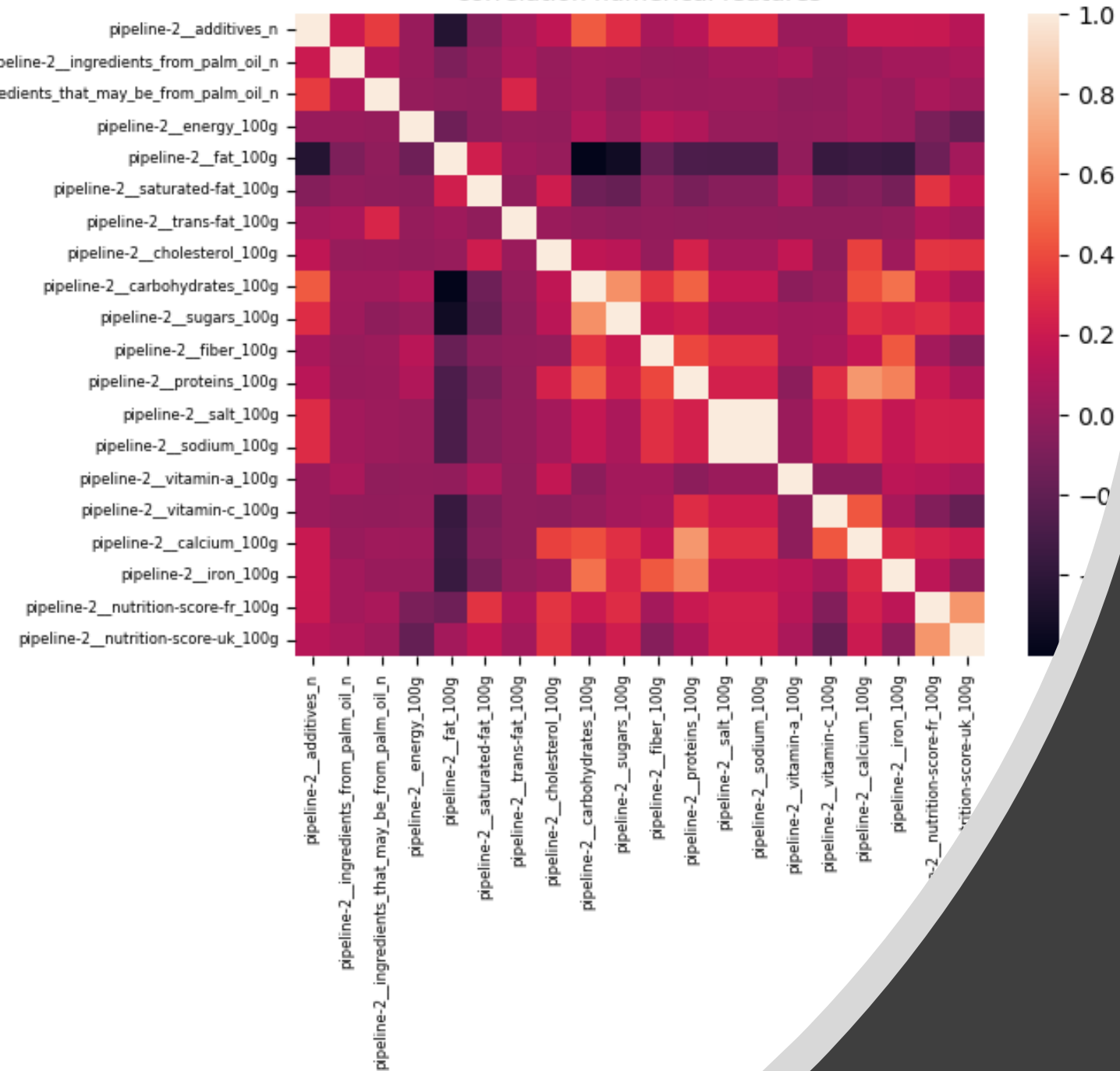
Pre-processing.
Etapas de
transformation
communes aux
trois méthodes.



Pre- processing.

- Méthodes de remplacement des valeurs manquantes s'appliquent sur le data-set transformé (slide précédent).
- Méthode KNN.
 - Utilise le data-set complet.
- Méthode Simple Imputer.
 - Utilise la variable seule.
- Méthode Iterative Imputer.
 - Utilise le data-set complet.

correlation numerical features



Pre-processing.

Test d'indépendance des variables.

- Variables corrélées:
 - "salt"
 - "sodium"

Pre-processing.

- Le jeu de données est maintenant correct. On a développé les étapes ci-dessous:
 - Identification des variables pour lesquelles le taux de valeurs manquantes est au delà du seuil admissible.
 - Analyse de la distribution des données: centrées ou non ?
 - Outliers: volumétrie par variable, valeurs aberrantes et leur remplacement par médiane.
 - Remplacement valeurs manquantes (selon différentes méthodes)
 - Vérification de l'indépendance des variables
 - Exclusion valeurs corrélées.
- Les données correspondant aux variables fondamentales sont maintenant correctement calibrées pour générer la réponse attendue par l'utilisateur: le code NUTRISPORT (A à E) du produit présenté par l'utilisateur.



FIN.

Application digitale dans le domaine de l'alimentation des sportifs de haut niveau.