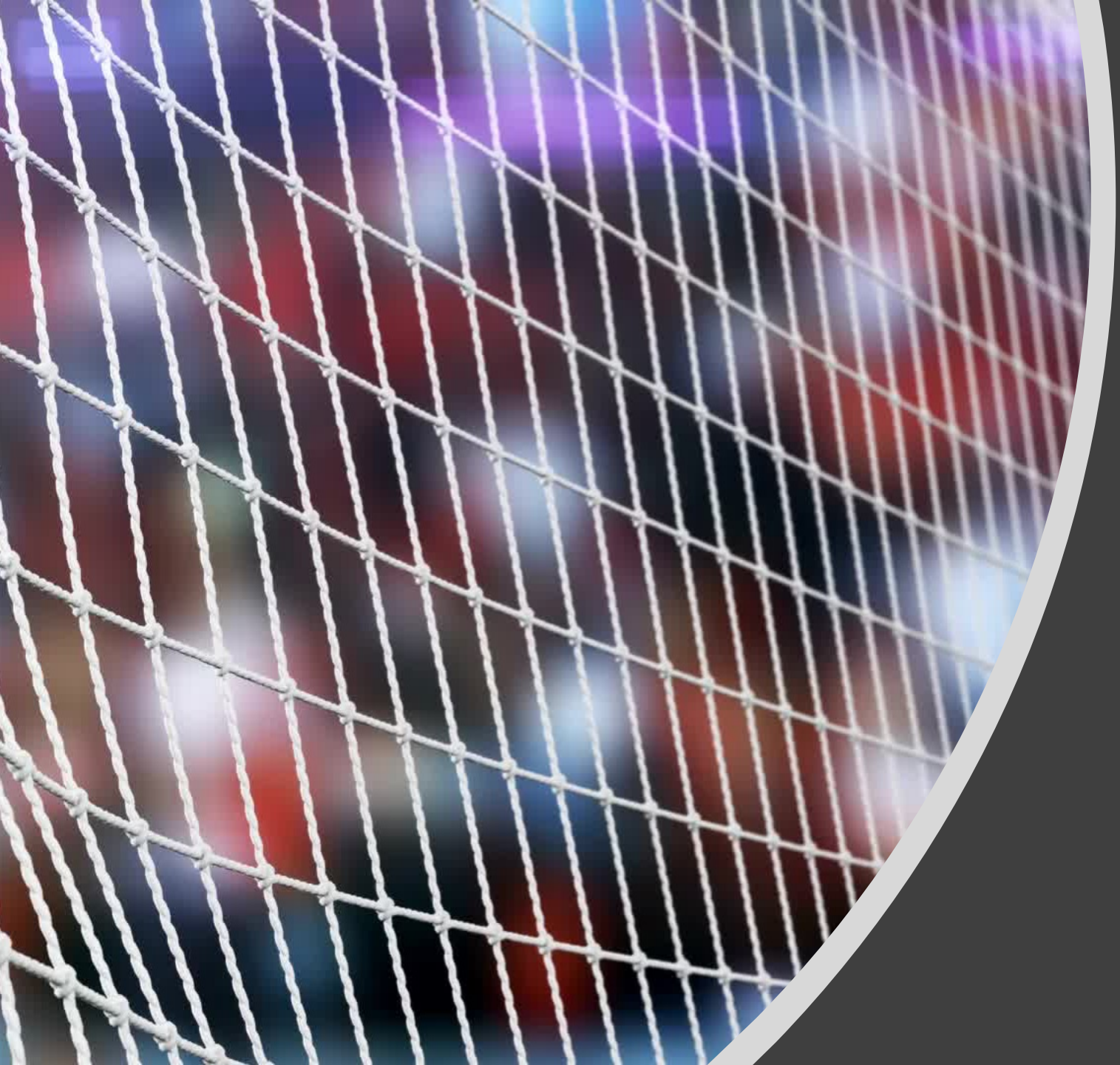


Application digitale dans le domaine de l'alimentation des sportifs de haut niveau.

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.

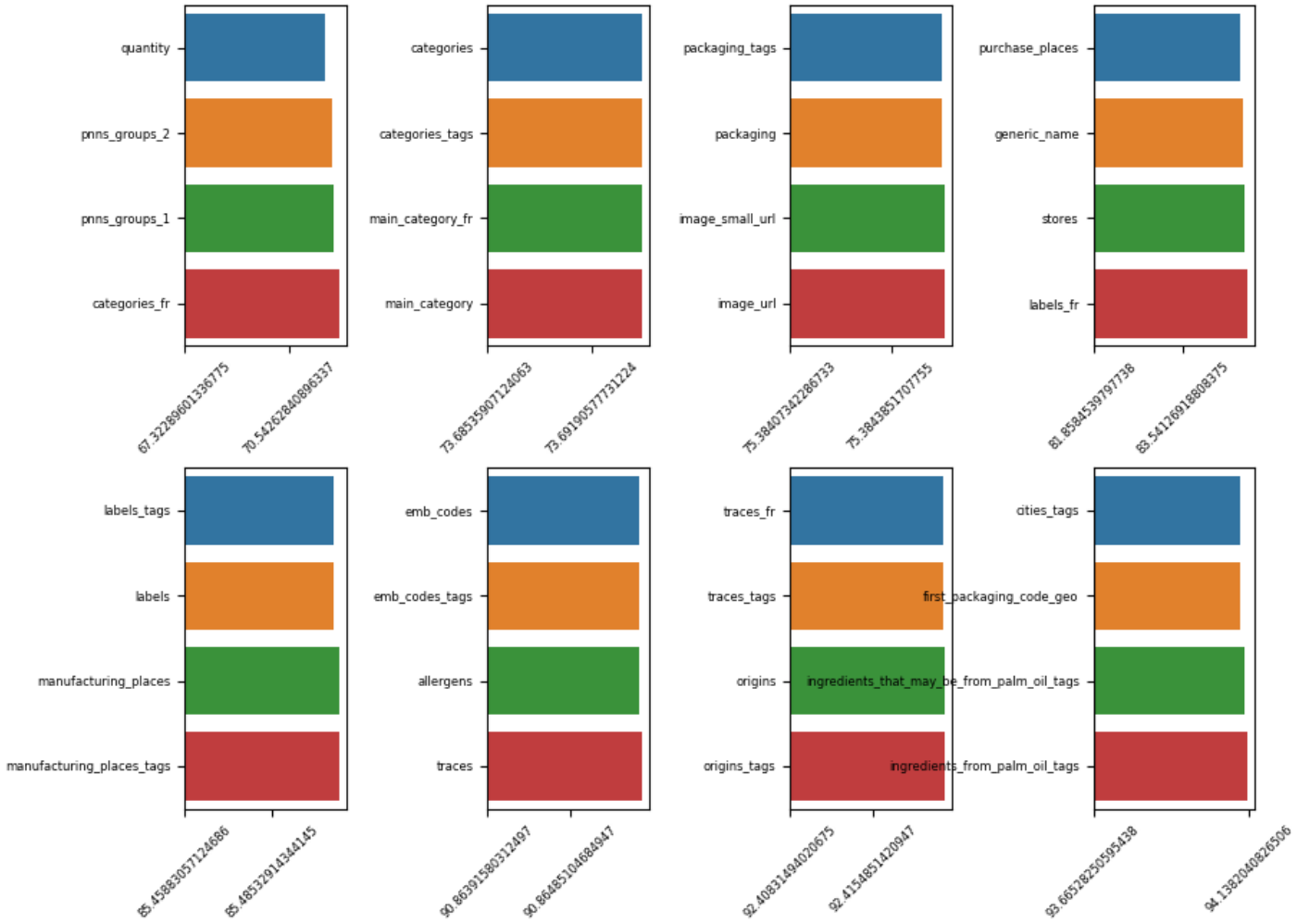
Introduction.

- 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.

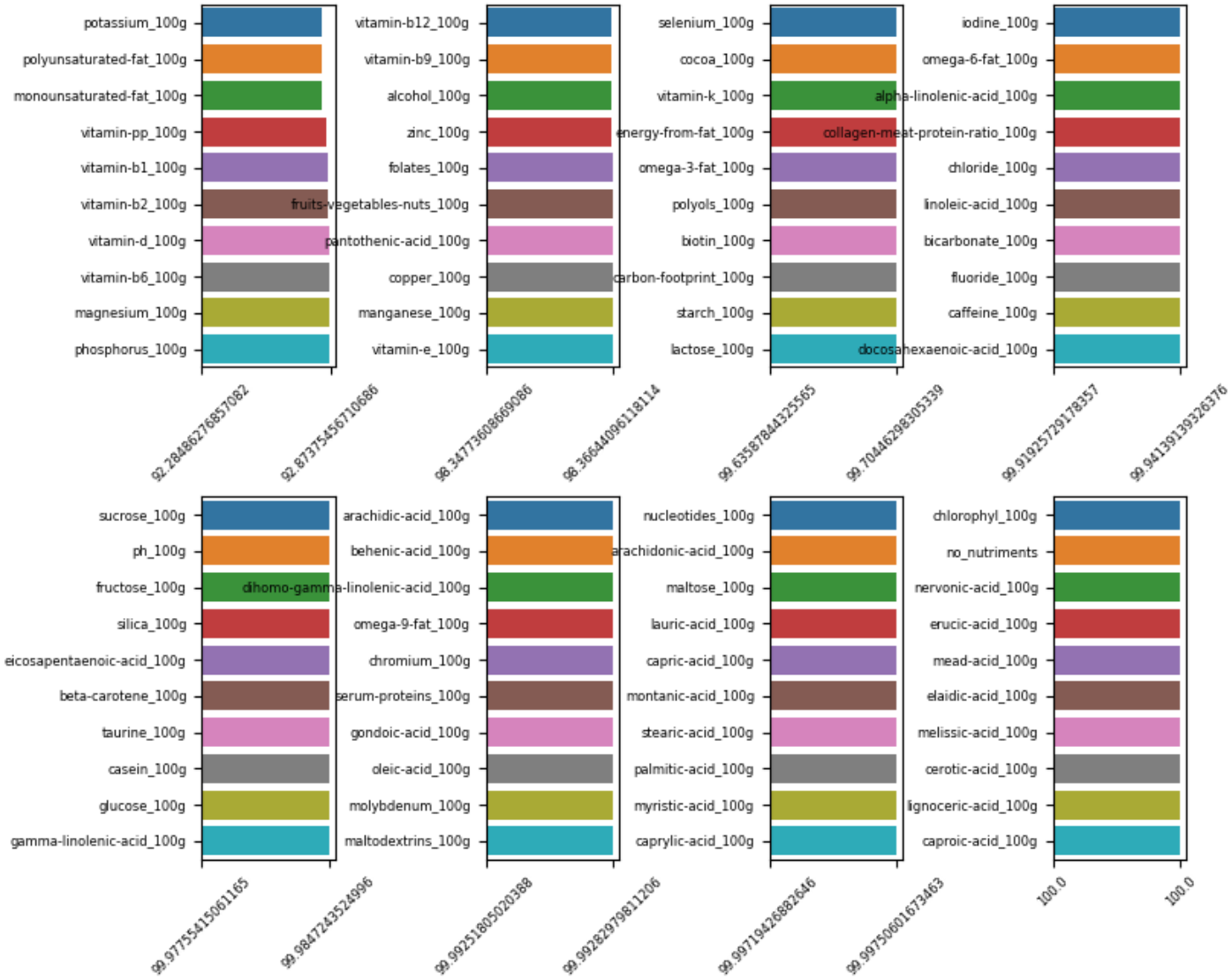
Phase exploratoire.

- 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

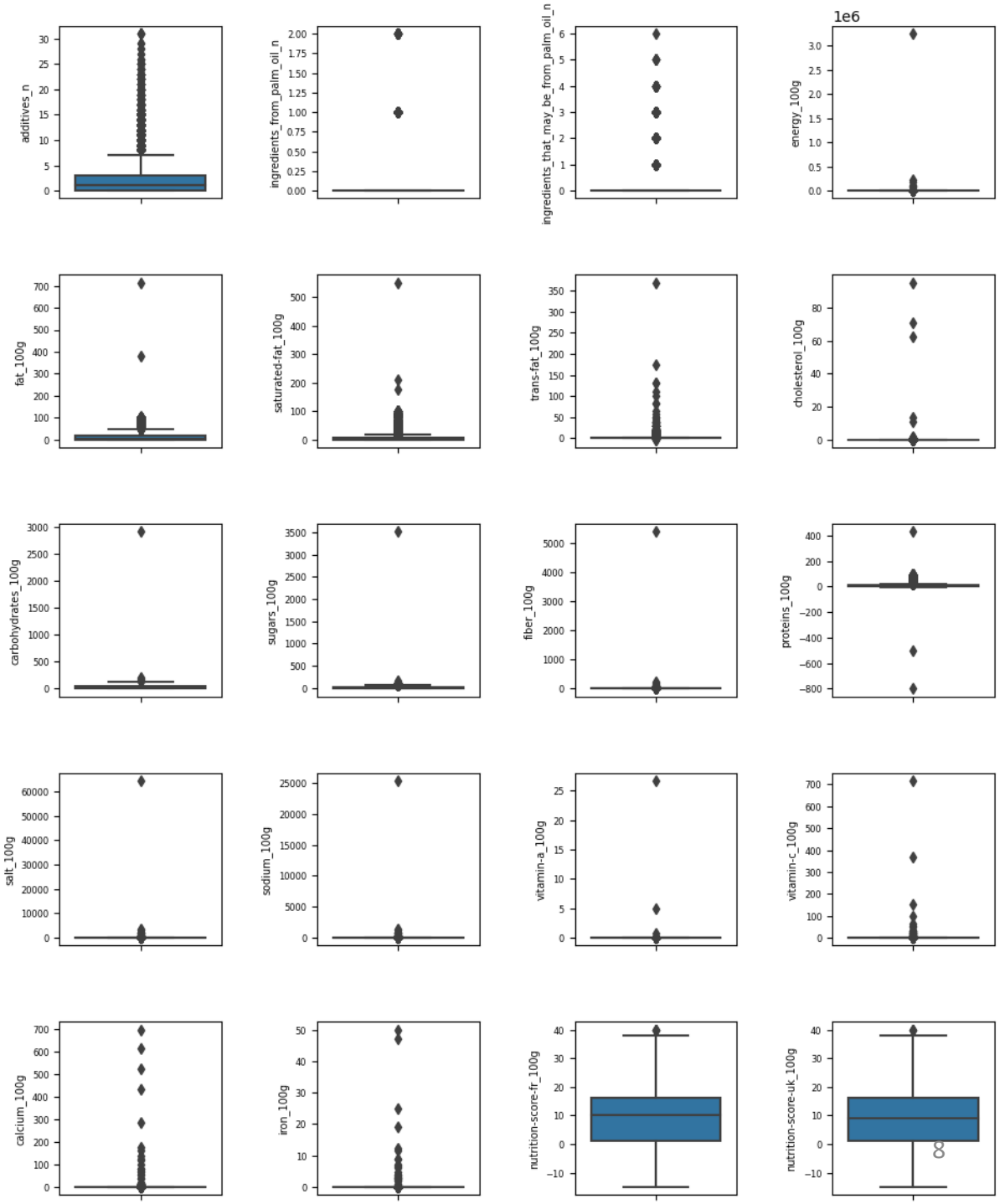
```
df_num.isnull().mean()*100
```

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.

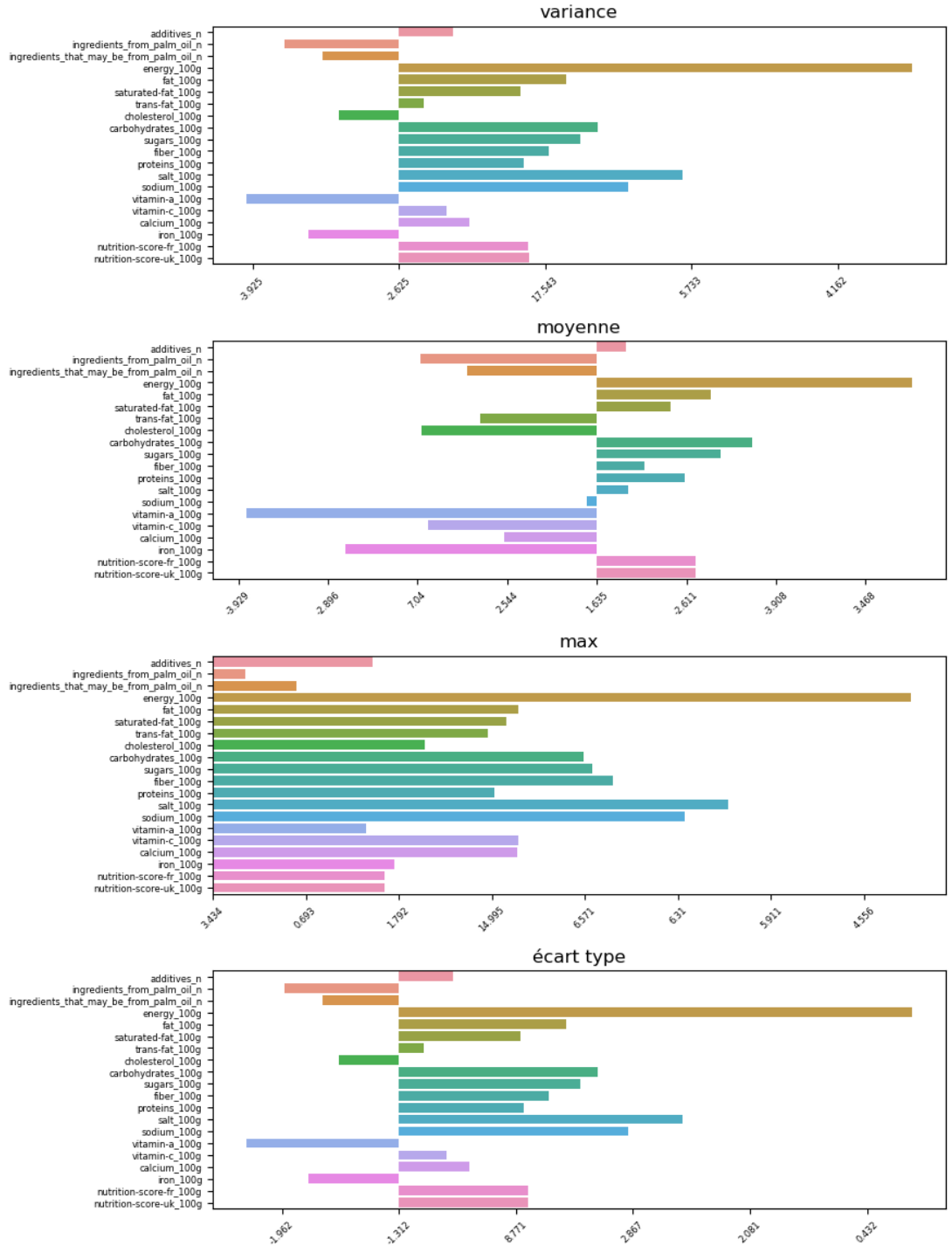
Distribution des données.



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

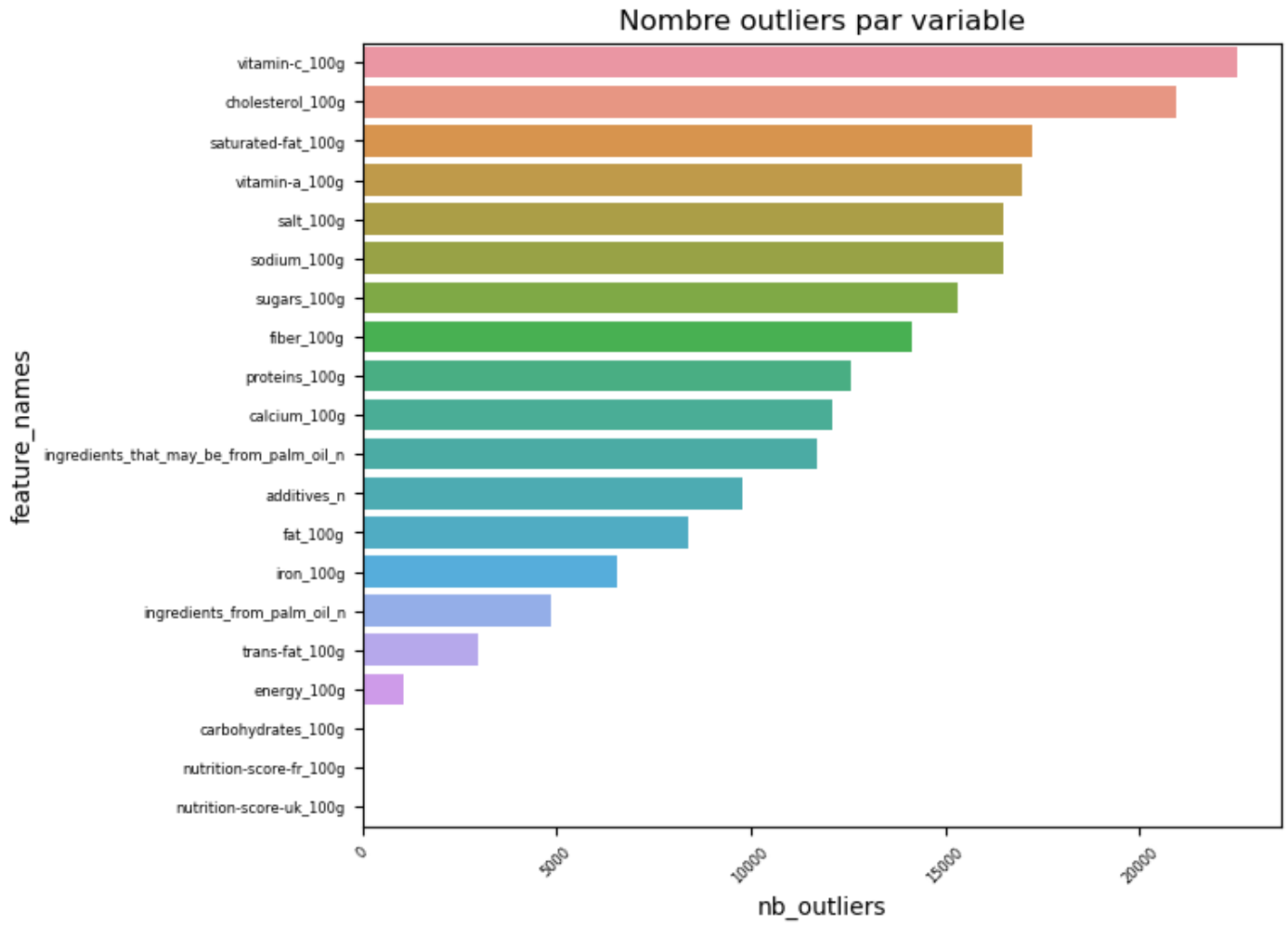
Phase exploratoire.

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


variance
additives_n
ingredients_from_palm_oil_n
ingredients_that_may_be_from_palm_oil_n
energy_100g
fat_100g
saturated-fat_100g
trans-fat_100g
cholesterol_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
salt_100g
sodium_100g
vitamin-a_100g
vitamin-c_100g
calcium_100g
iron_100g
nutrition-score-fr_100g
nutrition-score-uk_100g
-3.925
-2.625
17.543
5.733
4.162
moyenne
-3.929
-2.896
7.04
2.544
1.635
-2.611
-3.908
3.468
max
3.434
0.693
1.792
14.995
6.571
6.31
5.911
4.556
écart type
-1.962
-1.312
8.771
2.867
2.081
0.432

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.)



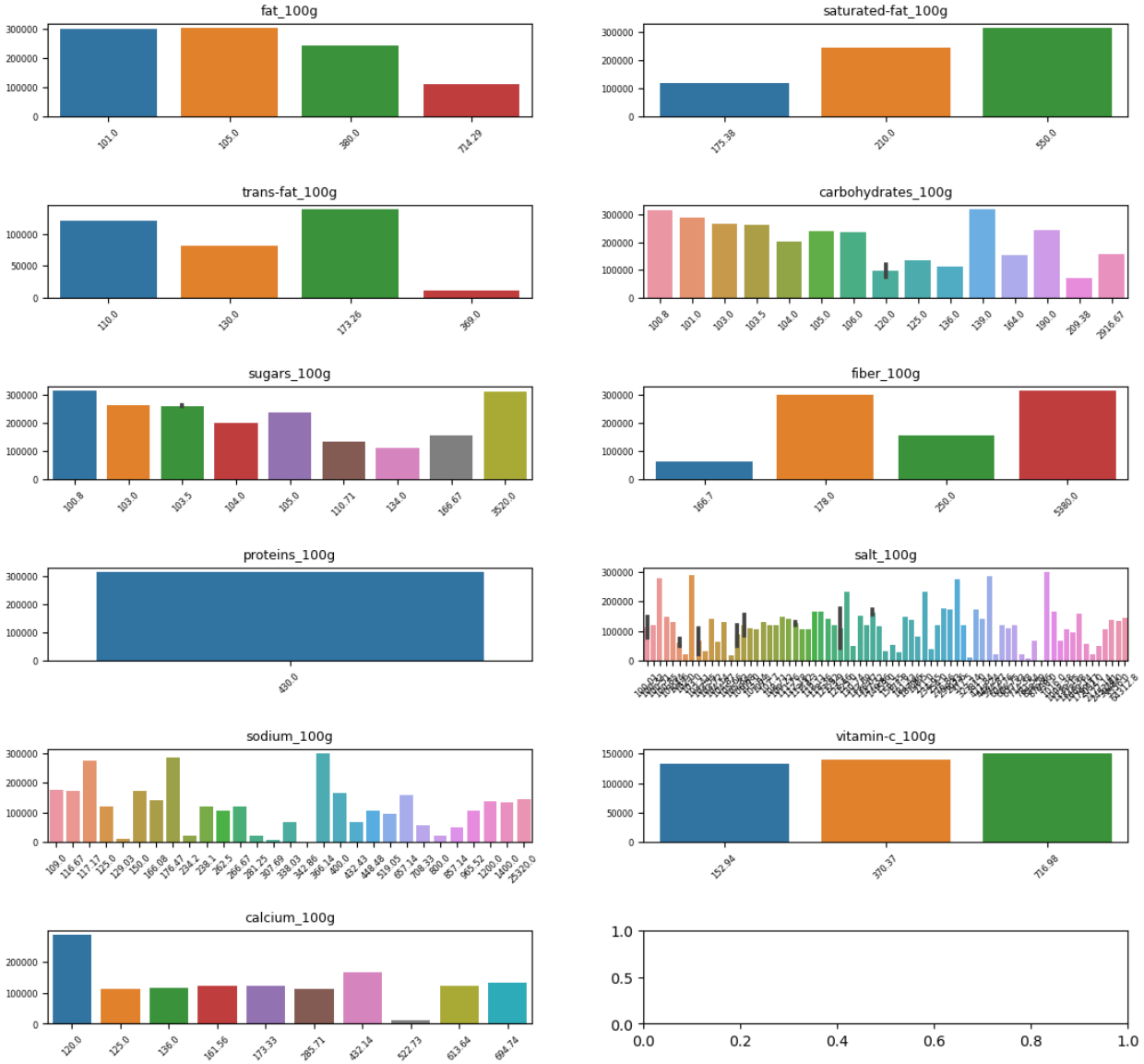
Phase exploratoire.

- Variables avec des erreurs de saisie (> 100g).
- Variable ‘energy_100g’: taux de valeurs en erreur de 75%.
  - Taux trop important (trop peu informatif), exclusion feature du dataset.
- Remplacement des valeurs erronées par la médiane (compte tenu de l’étalement des valeurs).

	valeurs
additives_n	0.000000
ingredients_from_palm_oil_n	0.000000
ingredients_that_may_be_from_palm_oil_n	0.000000
energy_100g	0.757803
fat_100g	0.000012
saturated-fat_100g	0.000009
trans-fat_100g	0.000016
cholesterol_100g	0.000000
carbohydrates_100g	0.000059
sugars_100g	0.000037
fiber_100g	0.000012
proteins_100g	0.000003
salt_100g	0.000493
sodium_100g	0.000106
vitamin-a_100g	0.000000
vitamin-c_100g	0.000009
calcium_100g	0.000031
iron_100g	0.000000
nutrition-score-fr_100g	0.000000
nutrition-score-uk_100g	0.000000

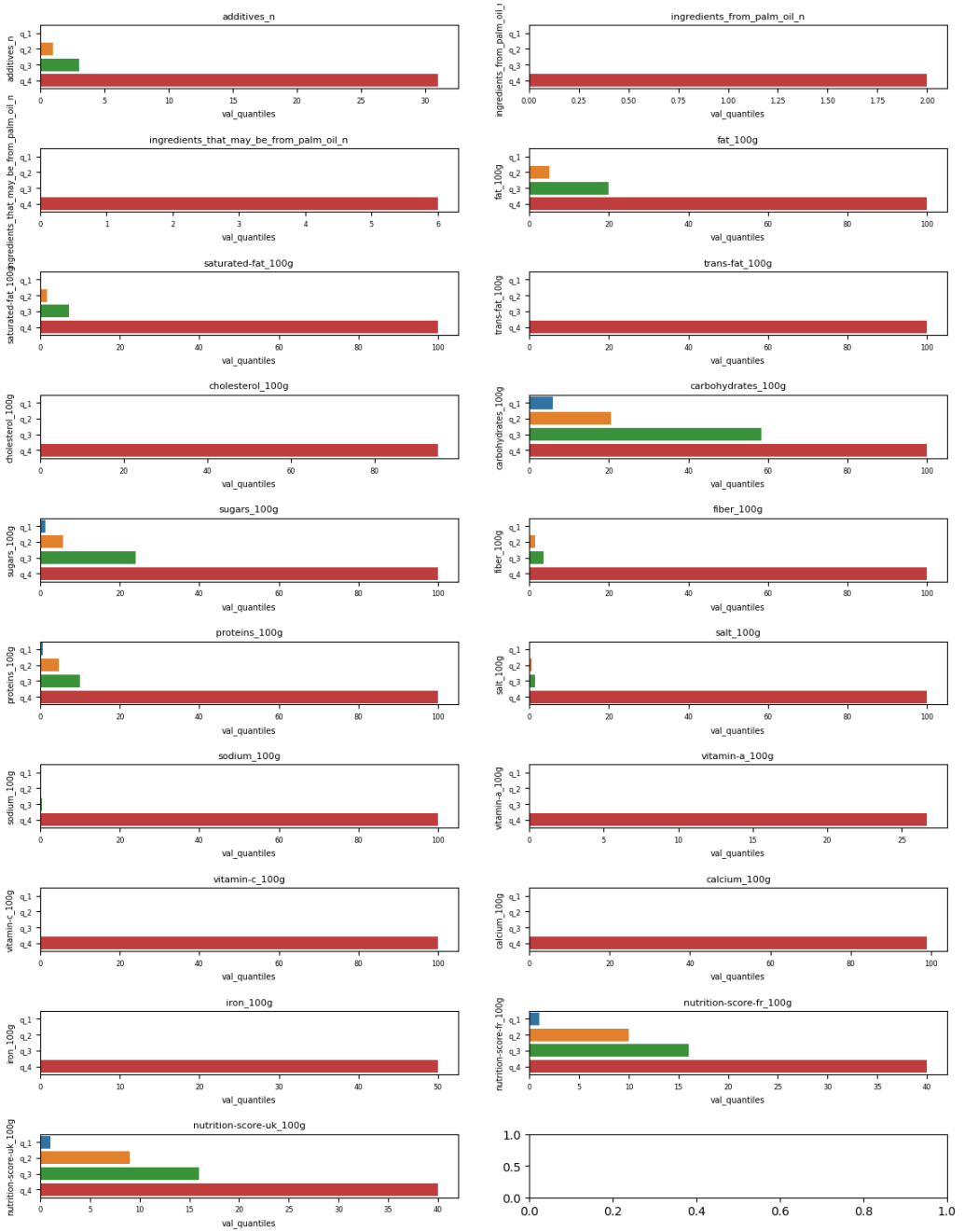
Phase exploratoire.

- Affichage des valeurs aberrantes.



Phase exploratoire.

Visualisation des quantiles après correction des erreurs de saisie.



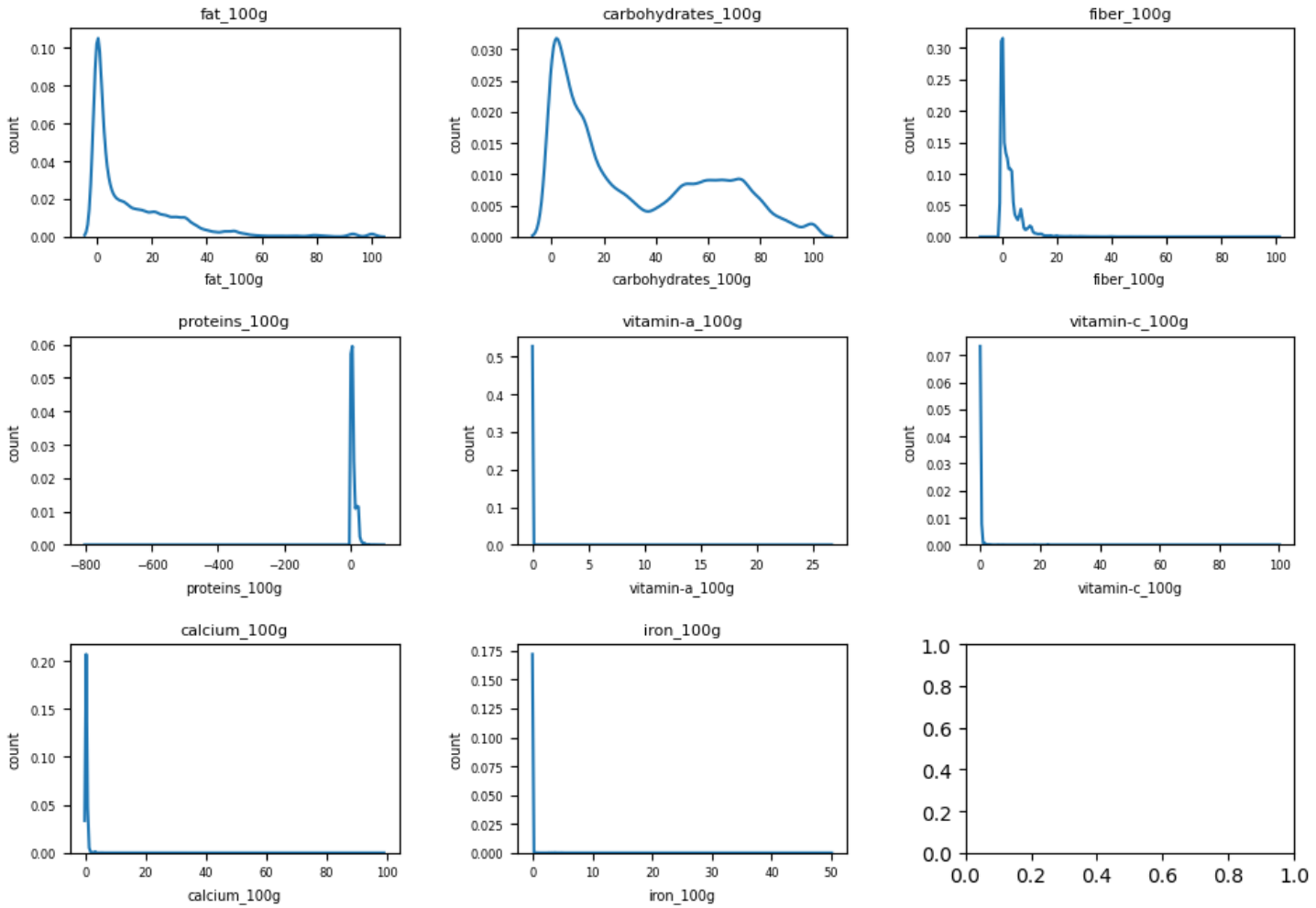
Phase exploratoire.

Pour notre analyse, seules certaines variables sont fondamentales.

Variables numériques fondamentales:

- 'fat_100g', 'carbohydrates_100g’,
- 'fiber_100g', 'proteins_100g’
- 'vitamin-a_100g’
- 'vitamin-c_100g', 'calcium_100g’
- 'iron_100g'

Visualisation des courbes de fréquence par variable niumérique fondamentale.



Pre-processing.

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

Trois méthodes utilisées: KNN, SimpleImputer, IterativeImputer.

```
X_num.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   fat_100g            9714 non-null   float64
 1   carbohydrates_100g  9730 non-null   float64
 2   fiber_100g          7860 non-null   float64
 3   proteins_100g       9719 non-null   float64
 4   vitamin-a_100g      7575 non-null   float64
 5   vitamin-c_100g      7702 non-null   float64
 6   calcium_100g        7720 non-null   float64
 7   iron_100g           7732 non-null   float64
dtypes: float64(8)
memory usage: 625.1 KB
```

Pre-processing.

Pour remplacer les valeurs manquantes dans le data-set, on va utiliser la méthode d'imputation basée KNN.

Les étapes suivantes sont nécessaires au préalable:

Ces étapes permettent de remplacer les valeurs « np.nan » (dtype= float) par des valeurs dtype= string reconnues par l'encodeur OHE.

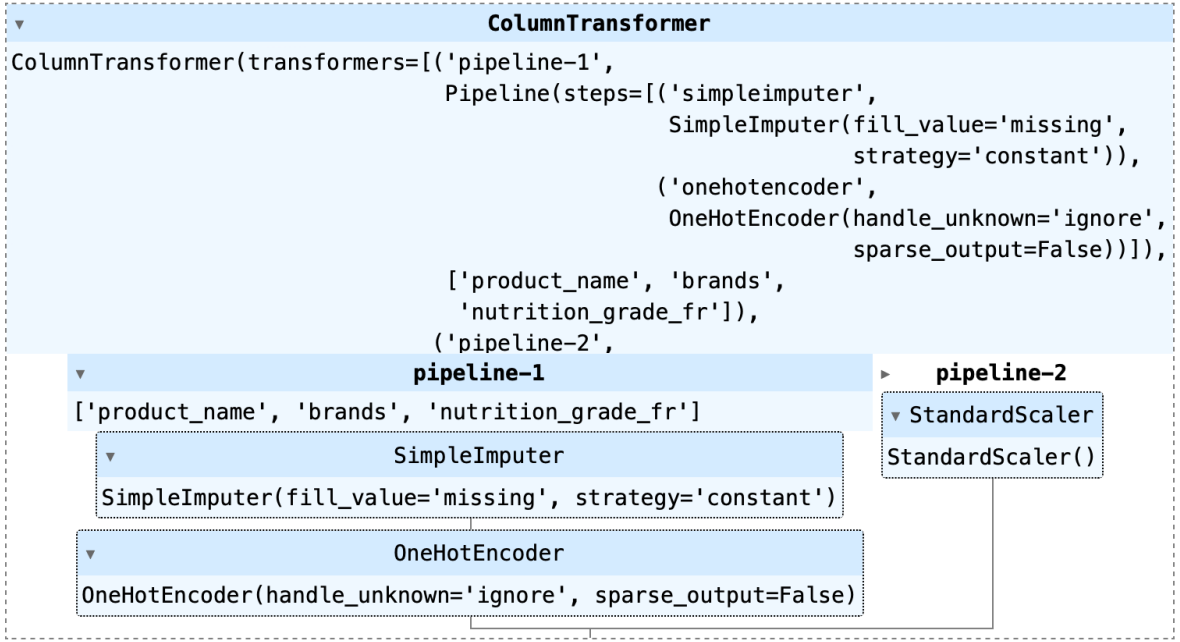
Séparer le data-set en catégoriel et numérique.

Méthode SimpleImputer() sur X_cat.

But: remplacer les « np.nan » par valeur 'missing'

Méthode OHE() sur X_cat.

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



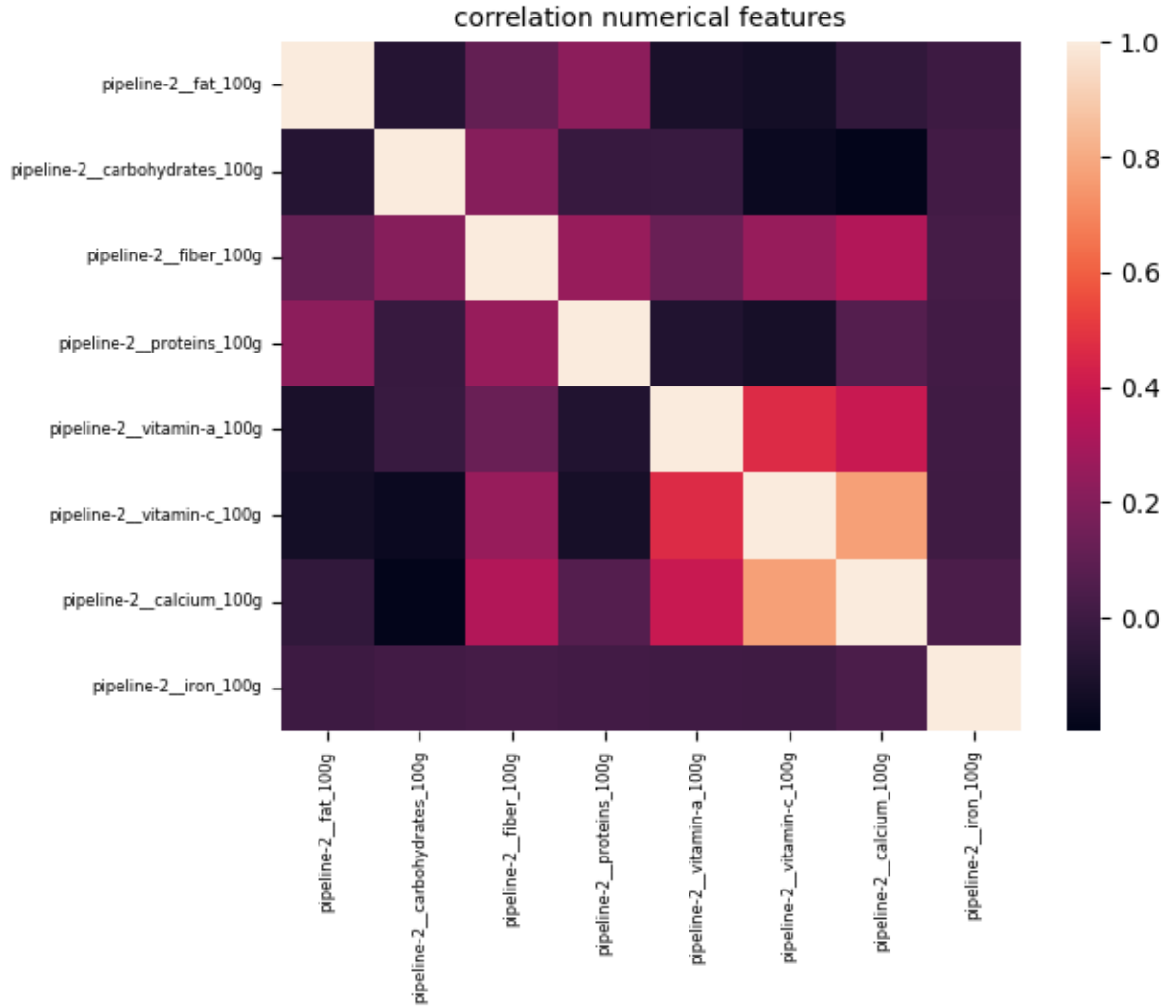
Pre-processing.

- 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.

Pre-processing.

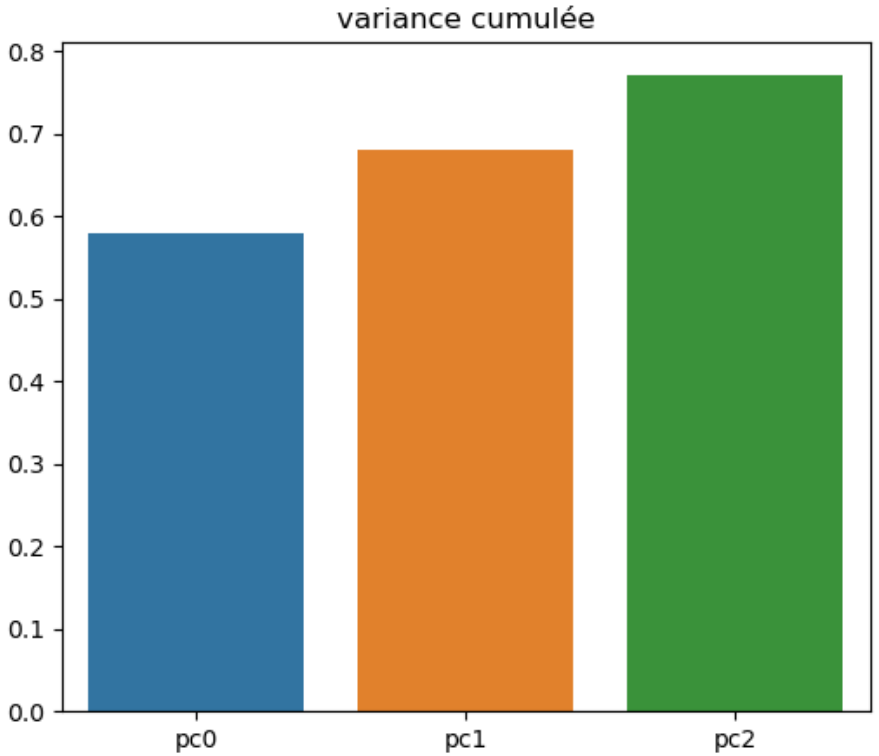
Test d’indépendance des variables.

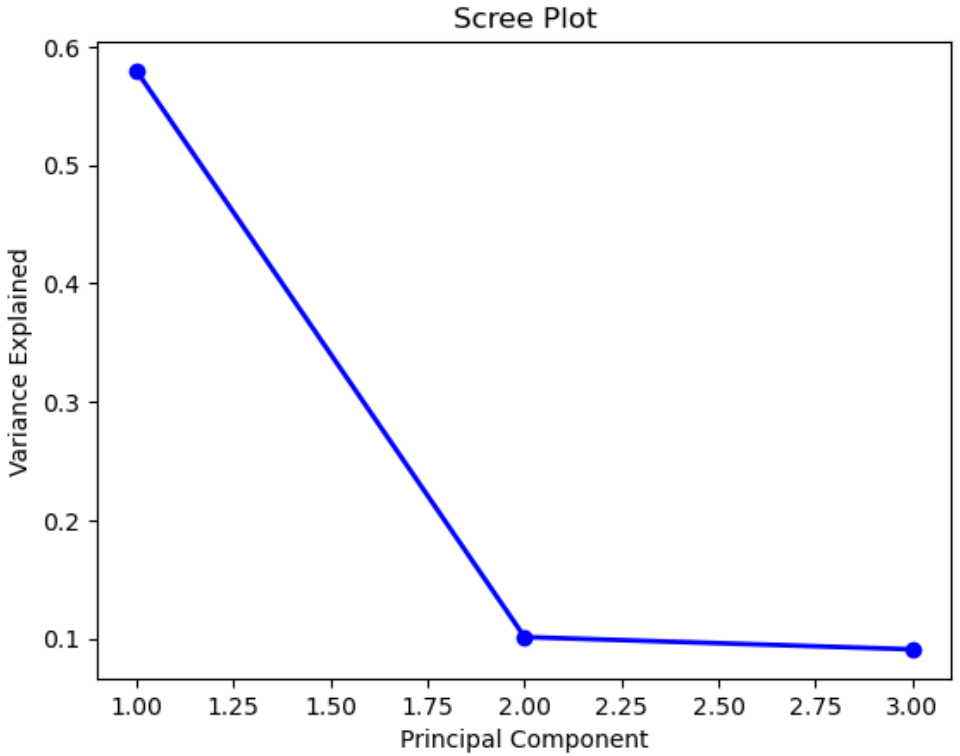
- Variables probablement corrélées:
  - “vitamin-a”
  - “vitamin-c”



Pre-processing.

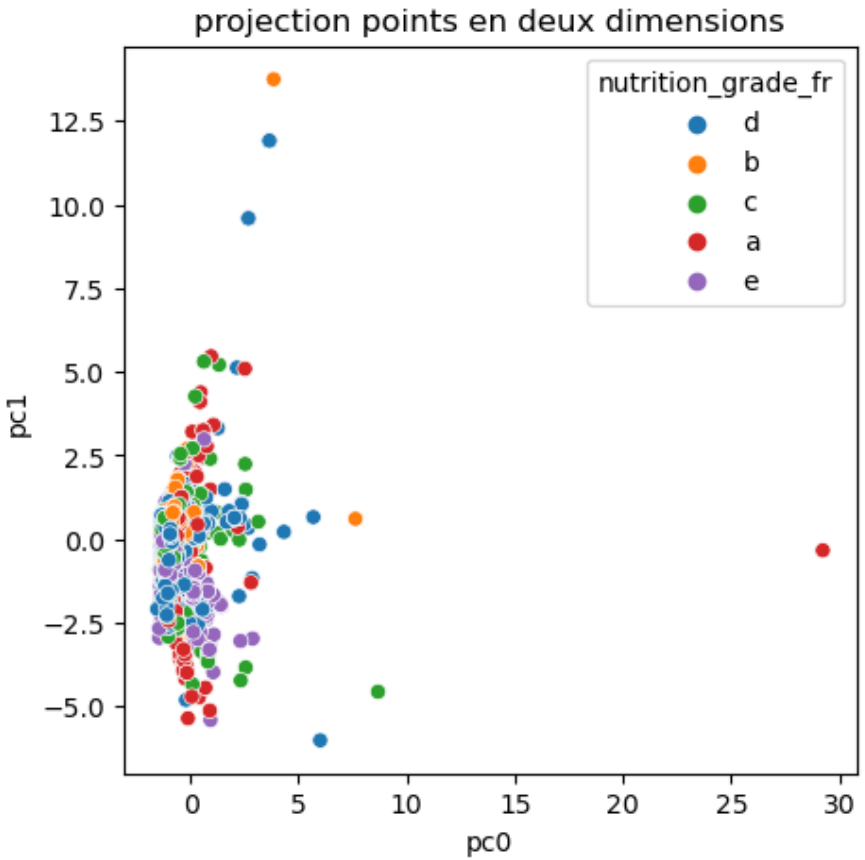
- Utilisation de PCA pour constitution de groupes ayant des caractéristiques similaires.
- Choix de trois composantes principales.
  - Captation de la variance et variance cumulée.

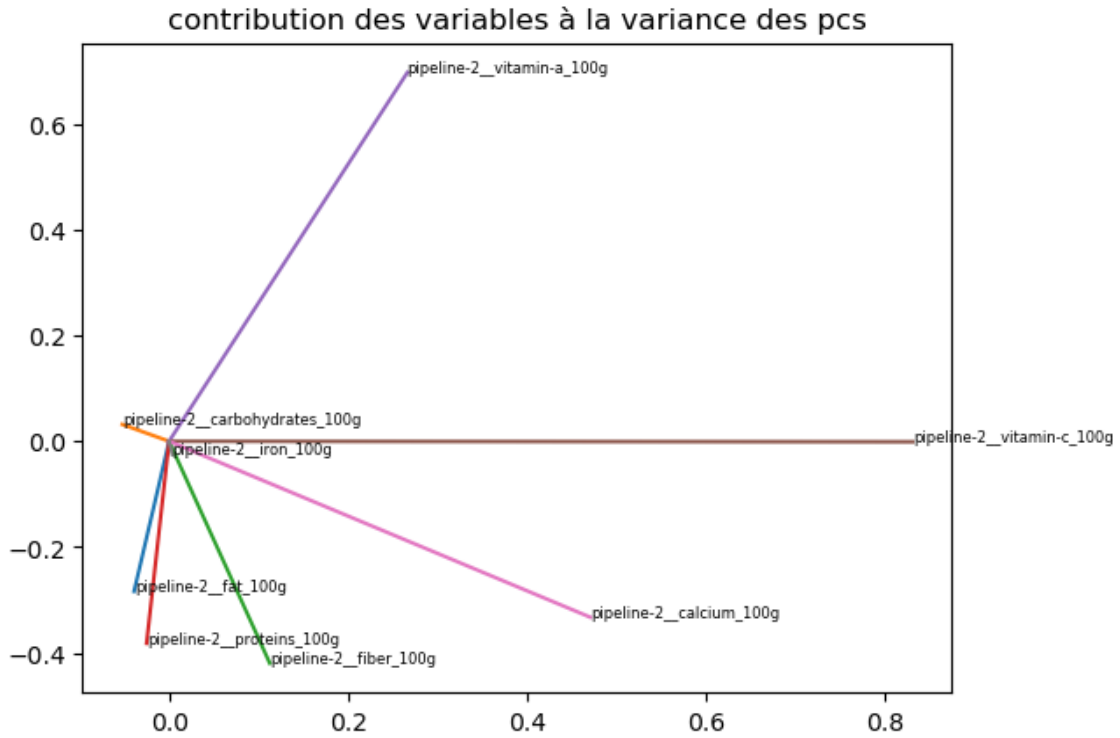




Pre-processing.

Projection de l'échantillon sur les deux premières composantes principales (pcs) et relation entre pcs et features.





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.