



# CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

OPENCLASSROOMS - INGÉNIEUR MACHINE LEARNING

THIBAUD GROSJEAN - NOVEMBRE 2021

# PRÉSENTATION DU PROJET

- Appel à projet de l'agence **Santé Publique France** (EPA)
  - « Veille sur les risques sanitaires menaçant les populations »
  - « Promotion de la santé et la réduction des risques pour la santé »
  - « Développement de la prévention et de l'éducation pour la santé »

(*THIERRY CARDOSO, Santé publique France : Missions et Organisation*)
- Conception d'une application innovante en lien avec l'alimentation
- Mise en valeur des données de la base **OpenFoodFacts**
  - « Open Food Facts est un projet collaboratif auquel participent des contributeurs volontaires du monde entier. » (<https://fr.openfoodfacts.org/qui-sommes-nous>)
  - Une base de données (BDD) ouverte

# MÉTHODOLOGIE

- Description du jeu de données
- Nettoyage des données
  - Filtrage
  - Imputation
  - Finalité : obtenir un jeu de données utilisable
- Analyse exploratoire
  - Analyse univariée
  - Analyse bivariée
  - Analyse multivariée
  - Finalité : démontrer la faisabilité de l'application

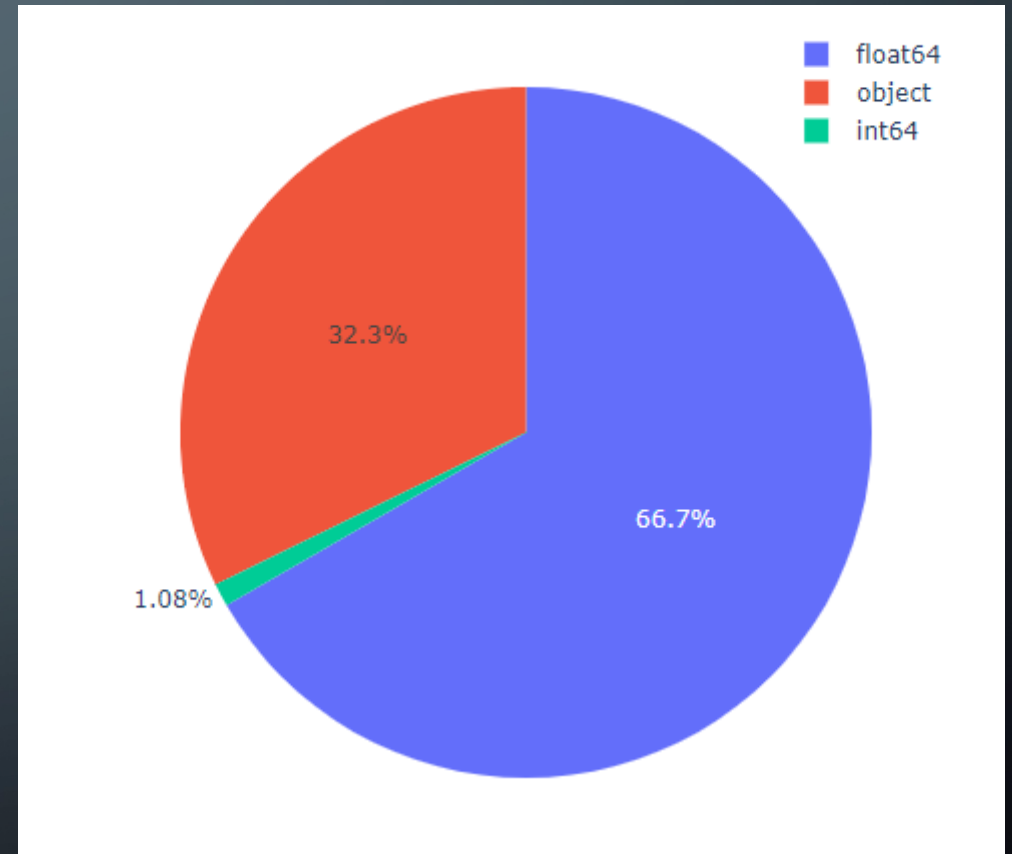
# DESCRIPTION DU JEU DE DONNÉES

- Objectifs
  - Examiner le dictionnaire de données
  - Comprendre la donnée
  - Examiner quelques échantillons
  - Trouver une idée d'application
  - Déterminer la qualité du jeu de données brut
  - Préparer le filtrage des données



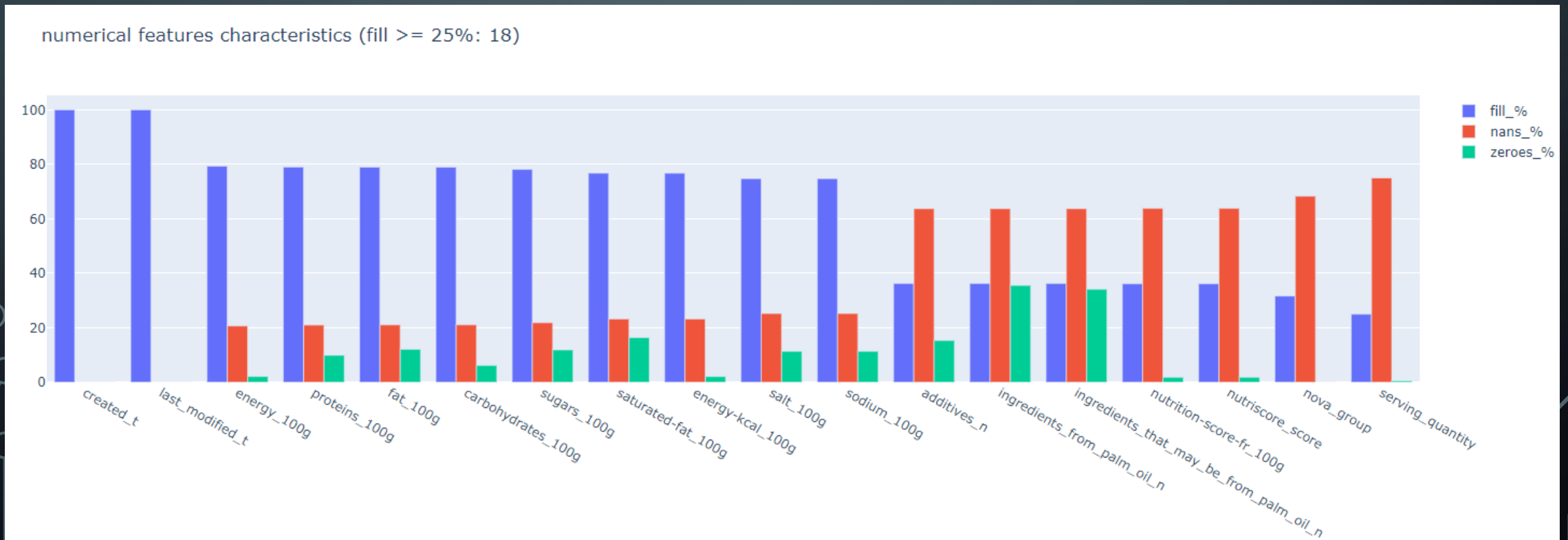
# DESCRIPTION DU JEU DE DONNÉES – VALEURS GLOBALES

- Au format .csv : 4.23 GB
- 2 millions d'individus
- 186 variables
  - Dont 126 variables numériques
    - Dont 124 variables continues
    - Et 2 variables discrètes
  - Et 60 variables qualitatives
- Valeurs manquantes : 79,83 %
- Valeurs uniques : 5.61 %



# DESCRIPTION DU JEU DE DONNÉES – VARIABLES NUMÉRIQUES

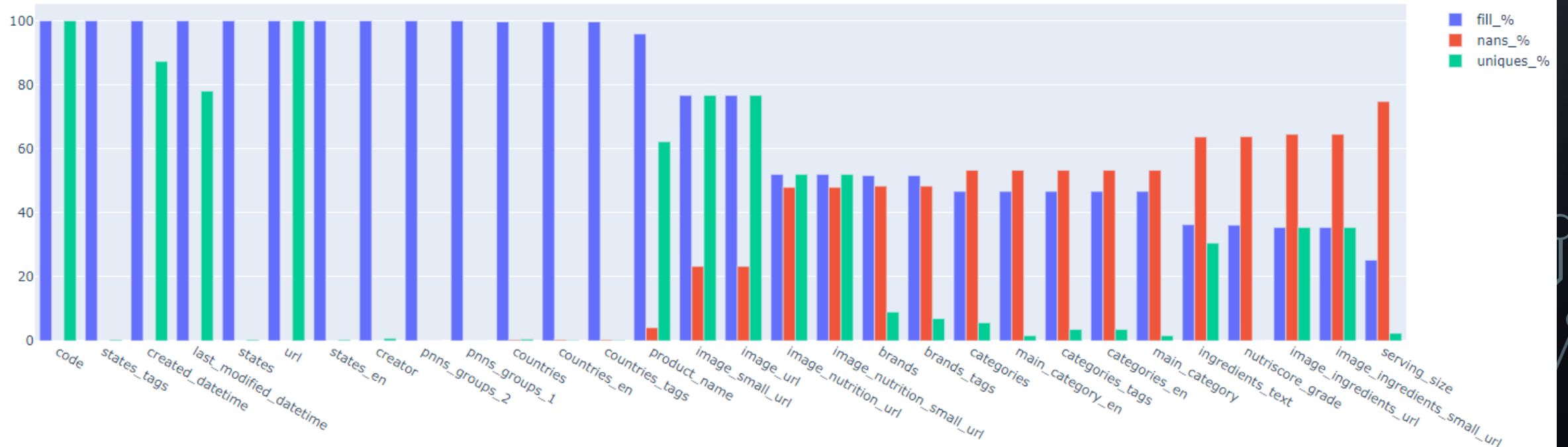
- Variables numériques dont le taux de remplissage est supérieur à 25% : 18



# DESCRIPTION DU JEU DE DONNÉES – VARIABLES QUALITATIVES

- Variables qualitatives dont le taux de remplissage est supérieur à 25% : 30

categorical features characteristics (fill >= 25%: 30)



# CONCEPT D'APPLICATION

- *BetterSnack*
- Application mobile
- Compare les nutriscores des snacks
- Pour en suggérer des plus sains à l'utilisateur



# FILTRAGE DU JEU DE DONNÉES

- Objectifs:
  - Faciliter les calculs
  - Filtrer les variables
  - Filtrer les individus
  - Sélectionner les données d'intérêt
  - Obtenir une donnée assez qualitative pour être décrite

# FILTRAGE DES VARIABLES – TAUX DE REMPLISSAGE

- Suppression des variables qui présentent un taux de remplissage inférieur à 50%
- Suppression de 132 variables
  - dont X variables numériques
  - Et Y variables qualitatives

# FILTRAGE DES VARIABLES – VARIABLES REDONDANTES

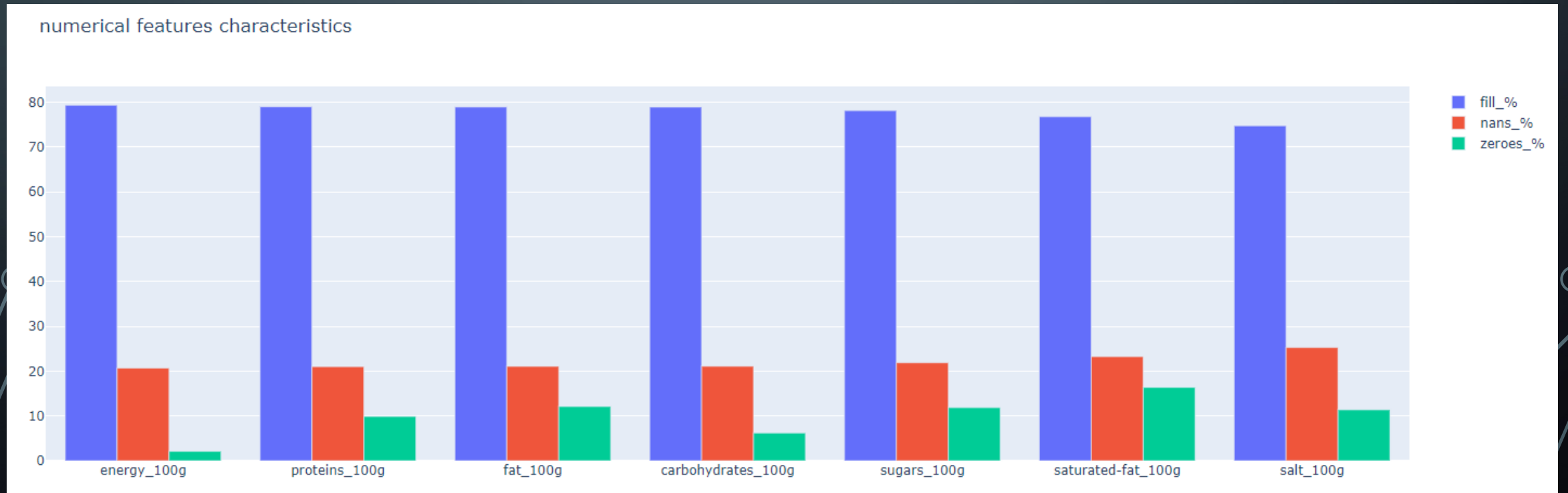
- Variables avec des valeurs équivalentes
- Avec une multitude de suffixes
- Dans des formats différents
- Nous retenons les formats les plus simples (\_tags...)
- Nous supprimons energy-kcal\_100g et sodium\_100g (équivalents à energy\_100g et salt\_100g)
- Au total : 34 variables

# FILTRAGE DES VARIABLES – MÉTA DONNÉES

- Suppression des méta données
- Nous supprimons 8 variables
  - dont X variables numériques
  - Et Y variables qualitatives
- 'last\_modified\_t', 'created\_t', 'url', 'image\_url', 'image\_nutrition\_url', 'code', 'states\_tags', 'creator'

# ETAT DU JEU DE DONNÉES

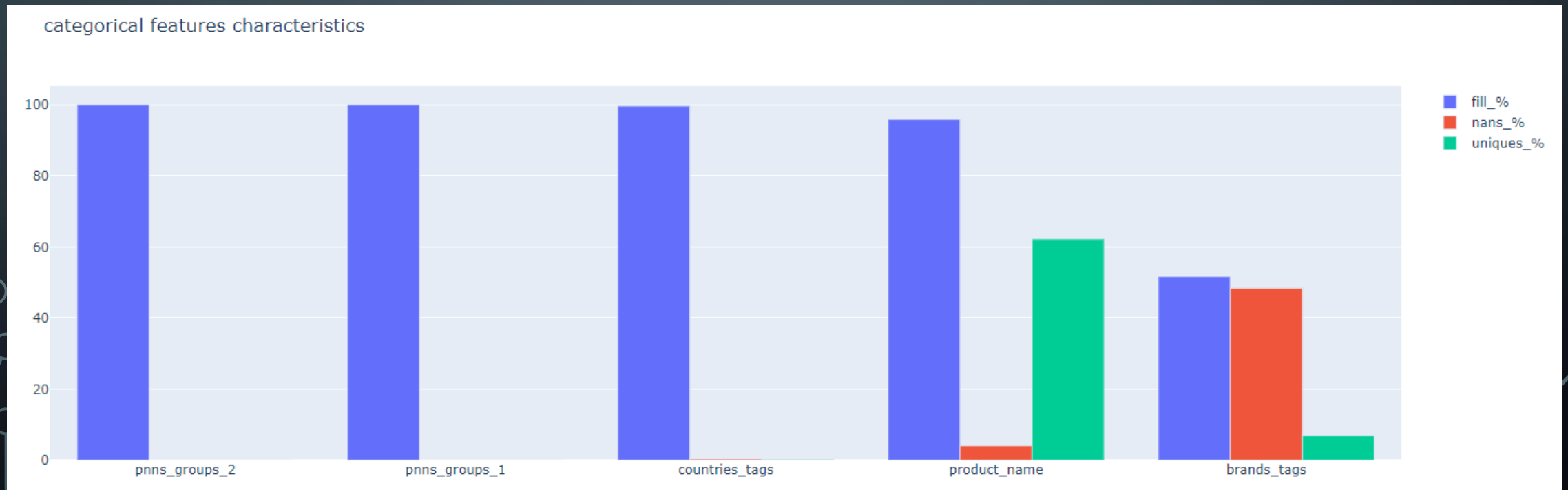
- Sélection de 7 des 126 variables numériques





# ETAT DU JEU DE DONNÉES

- Sélection de 5 des 60 variables qualitatives



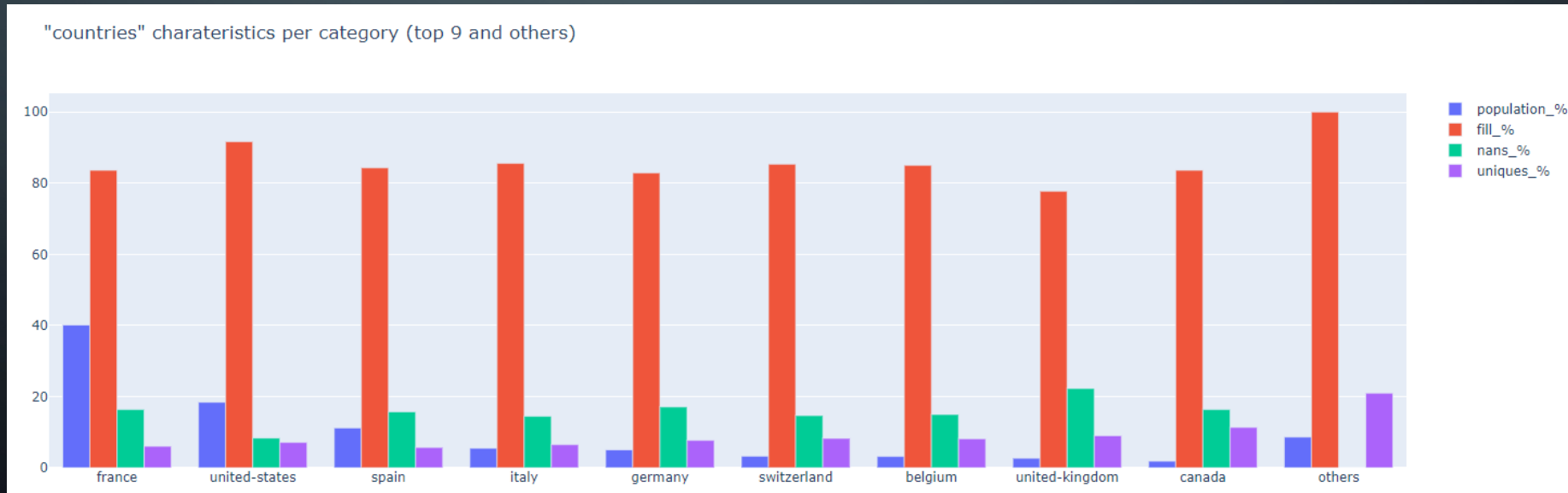
# FILTRAGE DES INDIVIDUS – PAYS

- Nettoyage de chaque ligne grâce à un regex

|        | countries_tags                                    | countries_tags_index | countries   | countries_index |
|--------|---|----------------------|---|-----------------|
| sample |   |                      |   |                 |
| 0      | en:croatia,en:france,en:germany,en:romania,en:... | 1164                 | bulgaria,poland,russia                            | 2855            |
| 1      | en:germany,de:españa,de:francia                   | 2195                 | bosnia-and-herzegovina,france,serbia              | 3835            |
| 2      | en:italy,en:united-states                         | 312                  | belgium,japan,switzerland                         | 2657            |
| 3      | en:france,en:hungary,en:poland,en:romania,en:s... | 2140                 | zagreb-hrvatska                                   | 1769            |
| 4      | en:france,en:italy,en:netherlands,en:united-ki... | 1707                 | belgium,france,portugal,reunion,spain,switzerland | 1611            |
| 5      | en:czech-republic,en:united-states                | 519                  | canada,france,italy,luxembourg,monaco,spain       | 2719            |
| 6      | en:uganda   | 2987                 | italy,russia                                      | 2255            |
| 7      | en:sweden,en:united-kingdom,en:united-states      | 2498                 | australia,france,germany,spain,united-kingdom     | 3938            |
| 8      | en:austria,en:belgium,en:germany,en:netherland... | 3123                 | thailand,united-kingdom                           | 558             |
| 9      | en:togo,en:united-states                          | 484                  | czech-republic,france,germany,united-kingdom      | 1834            |

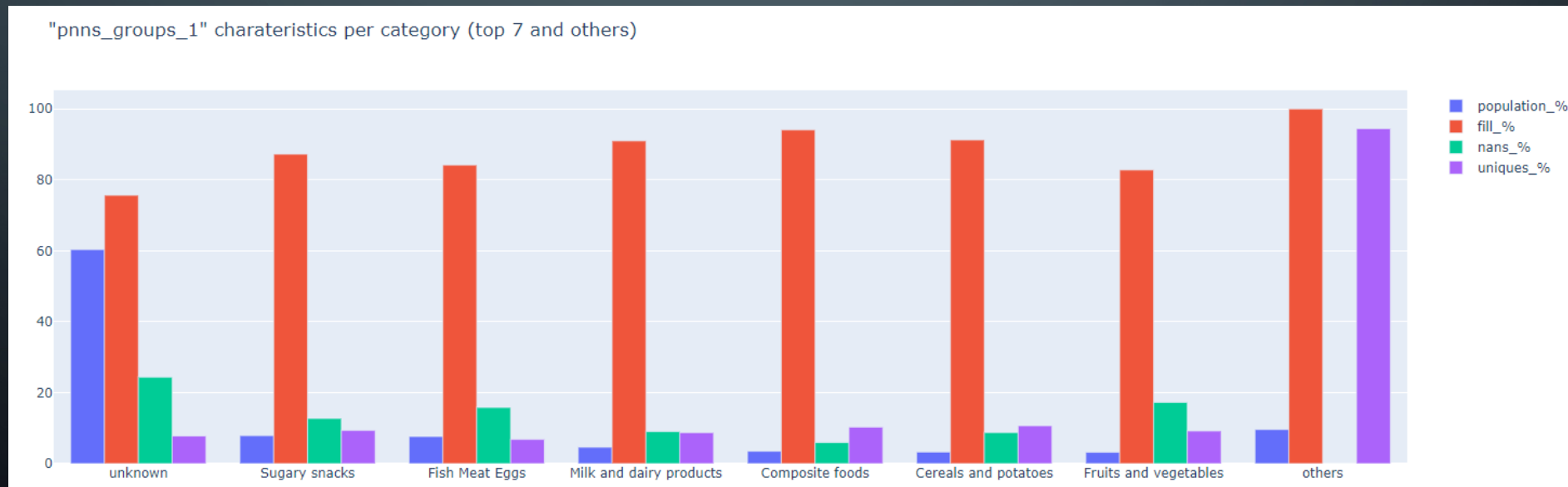
# FILTRAGE DES INDIVIDUS – PAYS

- Unzipping des valeurs de chaque ligne
- Duplication pour chaque individu
- Nous choisissons d'utiliser les données des produits français



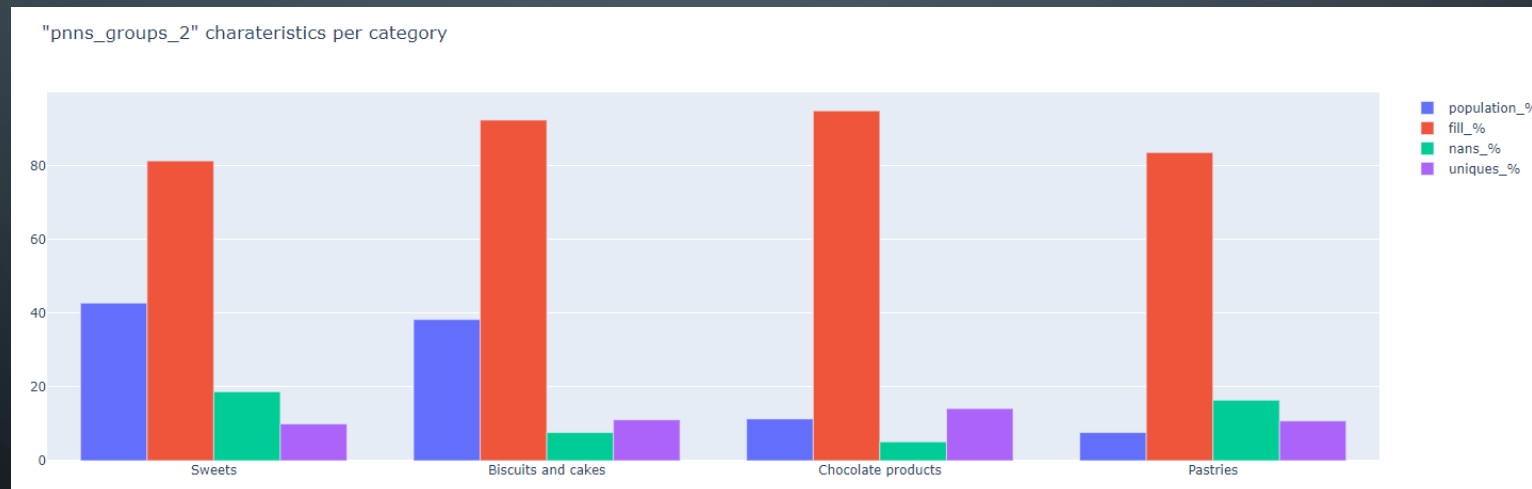
# FILTRAGE DES INDIVIDUS – PNNS GROUPS 1

- La modalité « Unknown » représente 60% de la population
- Pour répondre au besoin de notre application, nous sélectionnons la catégorie « Sugary Snacks »



# FILTRAGE DES INDIVIDUS – PNNS GROUPS 2

- Il apparait que lorsque « PNNS groups 1 » est rempli, « PNNS groups 2 » l'est aussi
- Cela permettra de fournir à l'utilisateur des recommandations plus précises en fonction du type de snack





# FILTRAGE DES INDIVIDUS – VALEURS NUTRITIONNELLES

- Suppression des individus
  - Dont au plus 2 des 7 variables sont vides
  - Comportant:
    - Des valeurs aux 100 grammes supérieures à 100
    - Des valeurs de matières grasses saturées supérieures aux matières grasses
    - Des valeurs de sucres supérieures aux glucides
    - Un total de sel, protéines, matières grasses et glucides aux 100 grammes supérieures à 100

|                    | nans | nans_% |
|--------------------|------|--------|
| feature            |      |        |
| salt_100g          | 1201 | 2.38   |
| saturated-fat_100g | 105  | 0.21   |
| energy_100g        | 83   | 0.16   |
| sugars_100g        | 31   | 0.06   |
| proteins_100g      | 23   | 0.05   |
| fat_100g           | 10   | 0.02   |
| carbohydrates_100g | 3    | 0.01   |

# IMPUTATION DES DONNÉES MANQUANTES

- Objectifs:

- Compléter les valeurs manquantes
  - Valeurs nutritionnelles
  - Nutriscore Score
  - Nutriscore Grade

- Finalité:

- Finaliser le traitement des données avant l'analyse exploratoire

# NETTOYAGE DES DONNÉES – IMPUTATION DES VALEURS NUTRITIONNELLES

- Application d'un KNN Imputer
- Application du filtrage des individus une nouvelle fois

| Standard Deviation |            | Standard Deviation |            | nans               | nans_% |
|--------------------|------------|--------------------|------------|--------------------|--------|
| salt_100g          | 0.796125   | salt_100g          | 0.789483   | feature            |        |
| saturated-fat_100g | 8.354746   | saturated-fat_100g | 8.353887   | salt_100g          | 0 0.0  |
| sugars_100g        | 20.150950  | sugars_100g        | 20.148161  | saturated-fat_100g | 0 0.0  |
| carbohydrates_100g | 17.547006  | carbohydrates_100g | 17.540927  | sugars_100g        | 0 0.0  |
| fat_100g           | 13.918258  | fat_100g           | 13.917860  | carbohydrates_100g | 0 0.0  |
| proteins_100g      | 3.957629   | proteins_100g      | 3.957424   | fat_100g           | 0 0.0  |
| energy_100g        | 581.875118 | energy_100g        | 581.835072 | proteins_100g      | 0 0.0  |
|                    |            |                    |            | energy_100g        | 0 0.0  |

# NETTOYAGE DES DONNÉES – IMPUTATION DE NUTRISCORE SCORE






- Application d'un KNN Regressor
- Split des données
  - X : données comportant un nutriscore
    - X\_train : 80% des données de X
    - X\_val : 20% des données de X
  - X\_test : données sans nutriscore
- Jointure de X et de X\_pred

```
Pass 0: 1 neighbor(s), MSE: 5.0
Pass 1: 2 neighbor(s), MSE: 3.89
Pass 2: 3 neighbor(s), MSE: 3.53
Pass 3: 4 neighbor(s), MSE: 3.36
Pass 4: 5 neighbor(s), MSE: 3.31
Pass 5: 6 neighbor(s), MSE: 3.29
Pass 6: 7 neighbor(s), MSE: 3.26
Pass 7: 8 neighbor(s), MSE: 3.28
Pass 8: 9 neighbor(s), MSE: 3.3
Best pass 6: 7 neighbor(s), MSE: 3.26
```



# NETTOYAGE DES DONNÉES – IMPUTATION DE NUTRISCORE GRADE

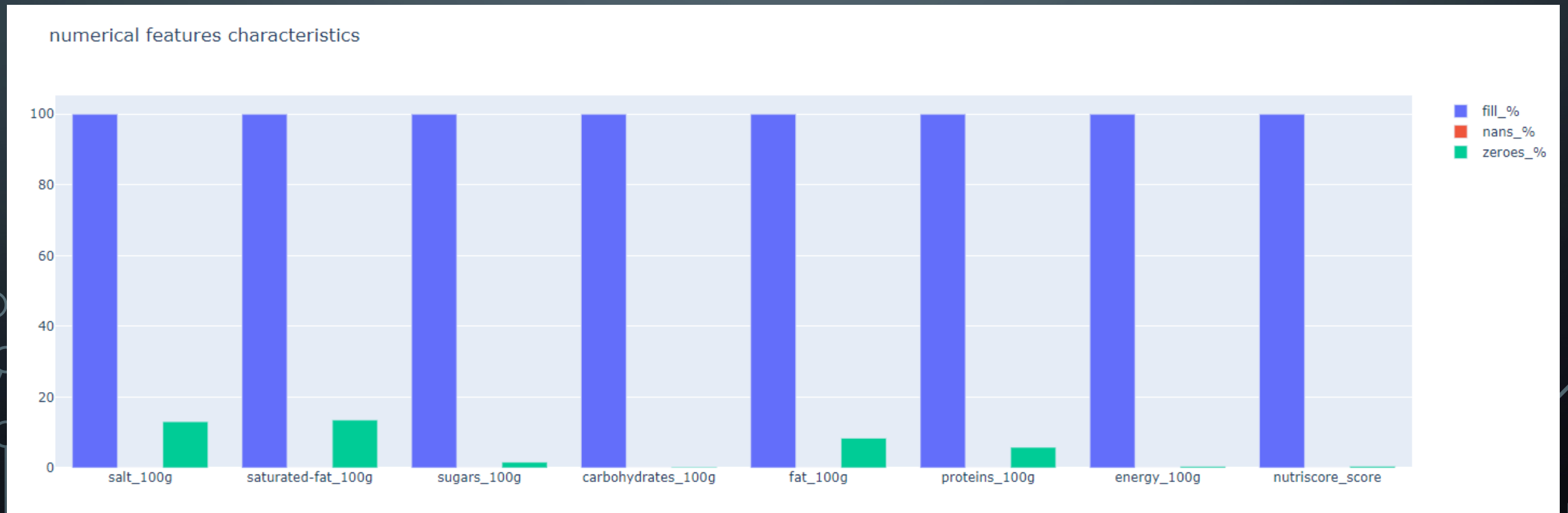
- Remplissage des grades de Nutriscore grâce au tableau d'équivalence

| Points      |           | Logo  |
|-------------|-----------|---|
| Solid foods | Beverages |   |
| Min to -1   | Waters    |    |
| 0 - 2       | Min - 1   |    |
| 3 - 10      | 2 - 5     |   |
| 11 - 18     | 6 - 9     |  |
| 19 - max    | 10 - max  |  |



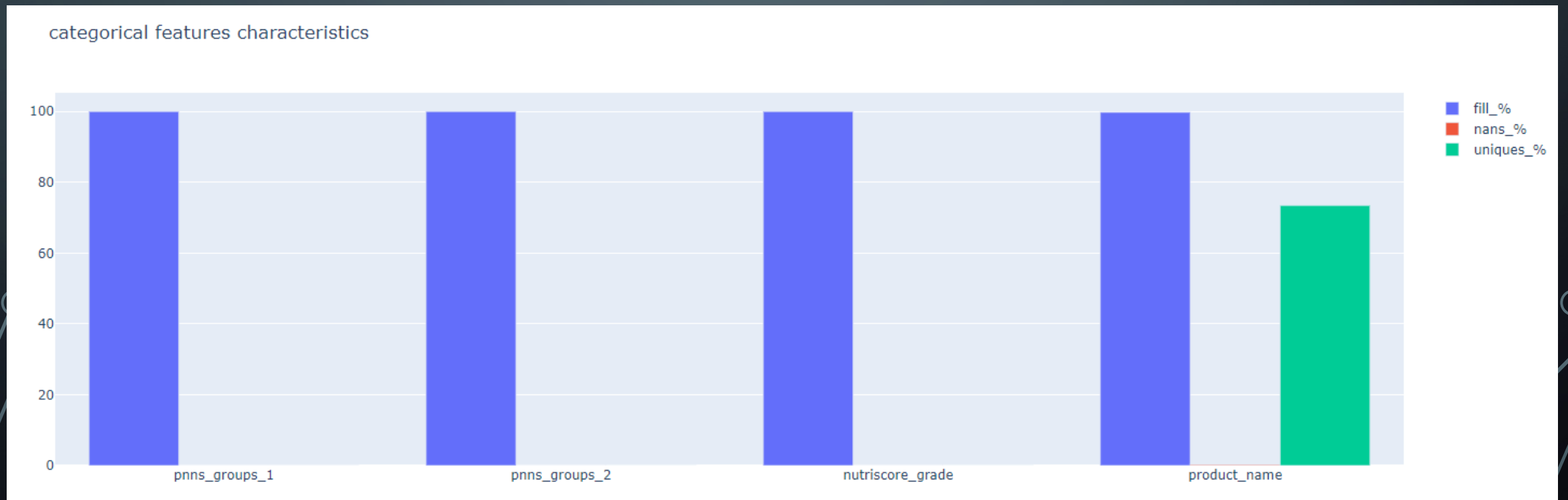
# ETAT DU JEU DE DONNÉES

- Sélection de 50364 des 1993128 individus



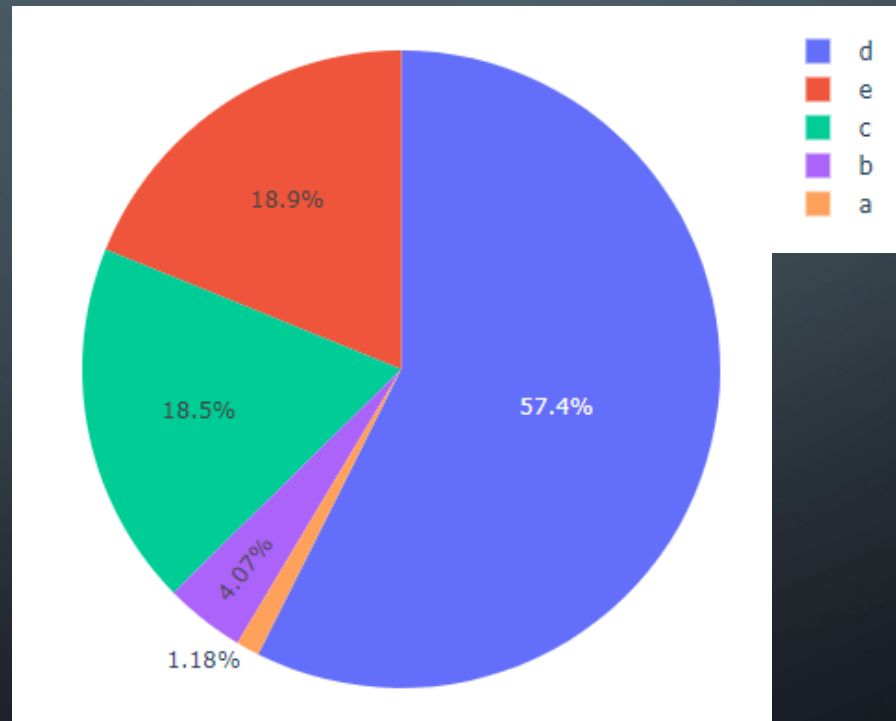
# ETAT DU JEU DE DONNÉES

- Sélection de 50364 des 1993128 individus



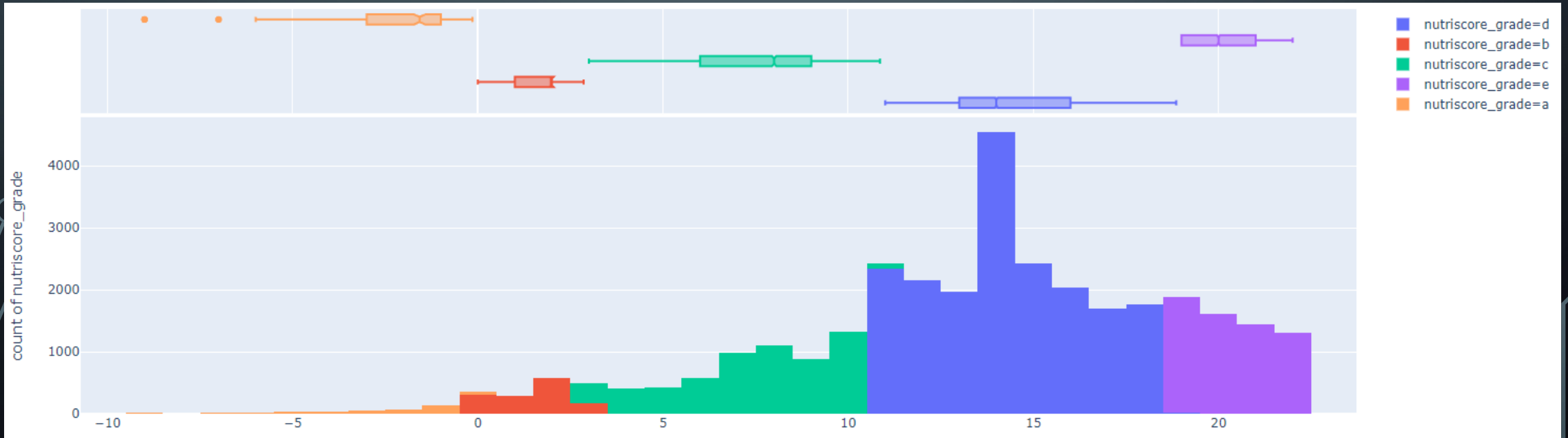
# ANALYSE UNIVARIÉE – NUTRISCORE GRADE

- Les grade de Nutriscore D et E représentent 76,3 % de la population totale



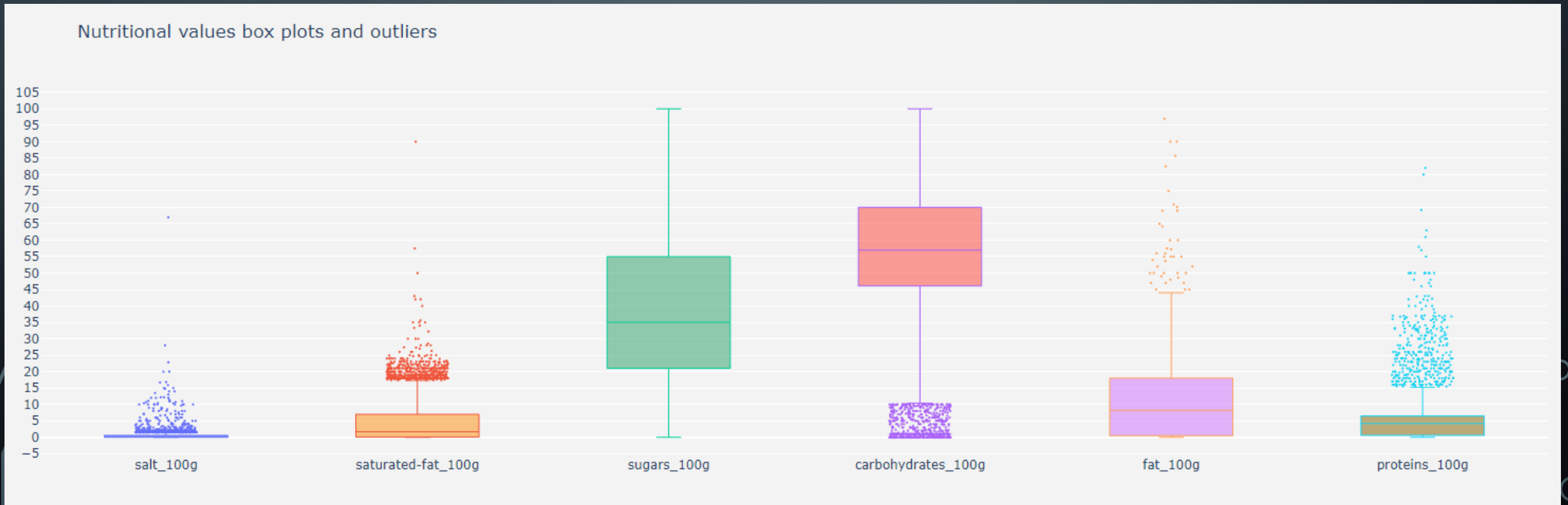
# ANALYSE UNIVARIÉE – NUTRISCORE SCORE

- Distribution des grades de Nutriscore
- L'approche choisie permet d'obtenir une distribution cohérente



# ANALYSE UNIVARIÉE – VALEURS NUTRITIONNELLES

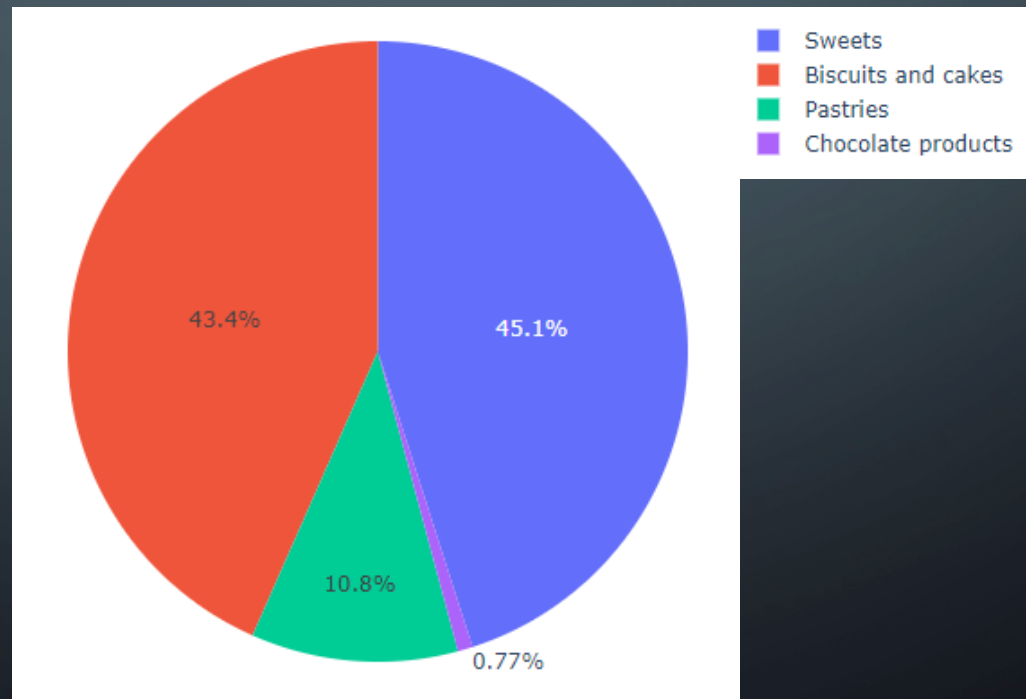
- Boîtes à moustache des valeurs nutritionnelles





# ANALYSE UNIVARIÉE – PNNS GROUPS 2

- « Sweets » et « Biscuits and Cakes » représentent 88,5 % de la population totale



# ANALYSE BIVARIÉE – CORRÉLATIONS LINÉAIRES

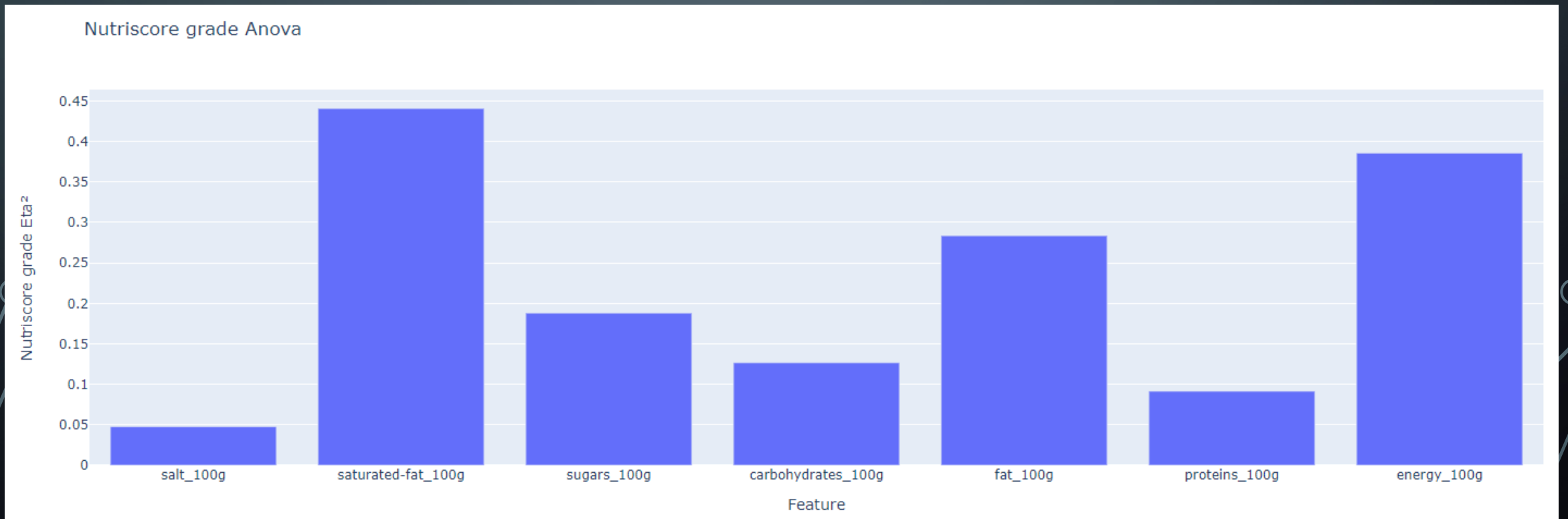
- Corrélation des variables numériques

Correlation Matrix of Nutritional Values

|                    | salt_100g | saturated-fat_100g | sugars_100g | carbohydrates_100g | fat_100g | proteins_100g | energy_100g | nutriscore_score |
|--------------------|-----------|--------------------|-------------|--------------------|----------|---------------|-------------|------------------|
| nutriscore_score   | 0.23      | 0.61               | 0.11        | 0.17               | 0.55     | 0.28          | 0.68        | 1.0              |
| energy_100g        | 0.11      | 0.43               | 0.03        | 0.35               | 0.62     | 0.46          | 1.0         | 0.68             |
| proteins_100g      | 0.21      | 0.42               | -0.42       | -0.24              | 0.56     | 1.0           | 0.46        | 0.28             |
| fat_100g           | 0.23      | 0.78               | -0.5        | -0.36              | 1.0      | 0.56          | 0.62        | 0.55             |
| carbohydrates_100g | -0.15     | -0.36              | 0.66        | 1.0                | -0.36    | -0.24         | 0.35        | 0.17             |
| sugars_100g        | -0.27     | -0.48              | 1.0         | 0.66               | -0.5     | -0.42         | 0.03        | 0.11             |
| saturated-fat_100g | 0.19      | 1.0                | -0.48       | -0.36              | 0.78     | 0.42          | 0.43        | 0.61             |
| salt_100g          | 1.0       | 0.19               | -0.27       | -0.15              | 0.23     | 0.21          | 0.11        | 0.23             |

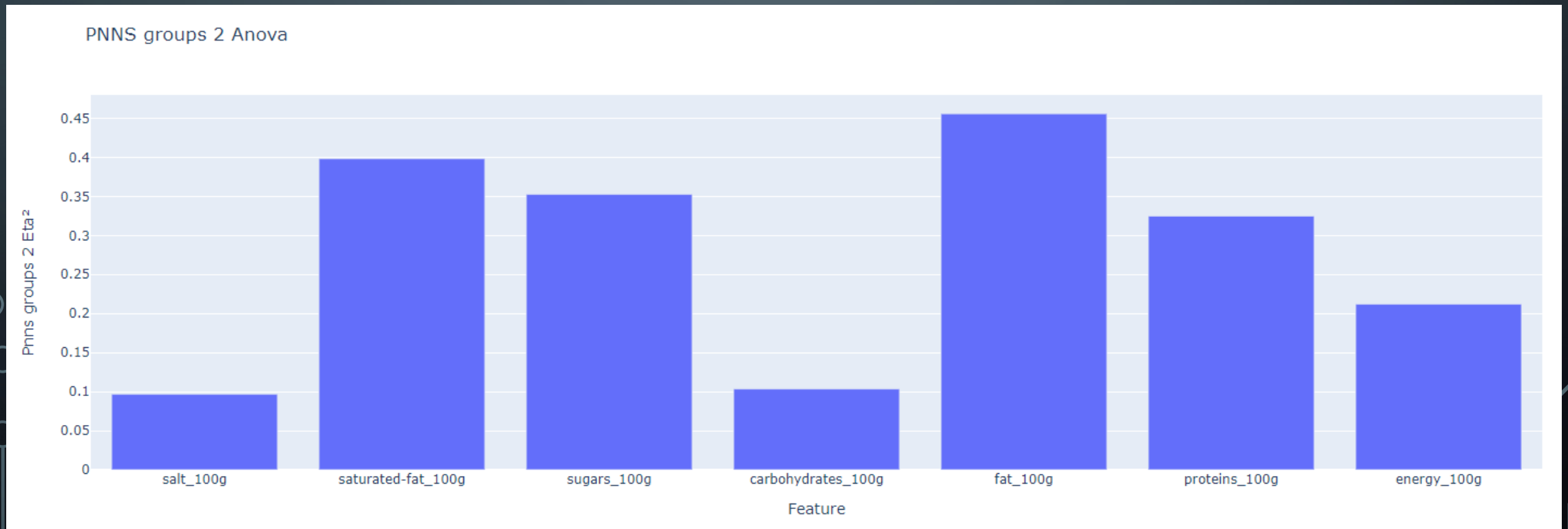
# ANALYSE BIVARIÉE – ANOVA

- Le grade de Nutriscore a un  $\text{Eta}^2$  de 0.22 en moyenne



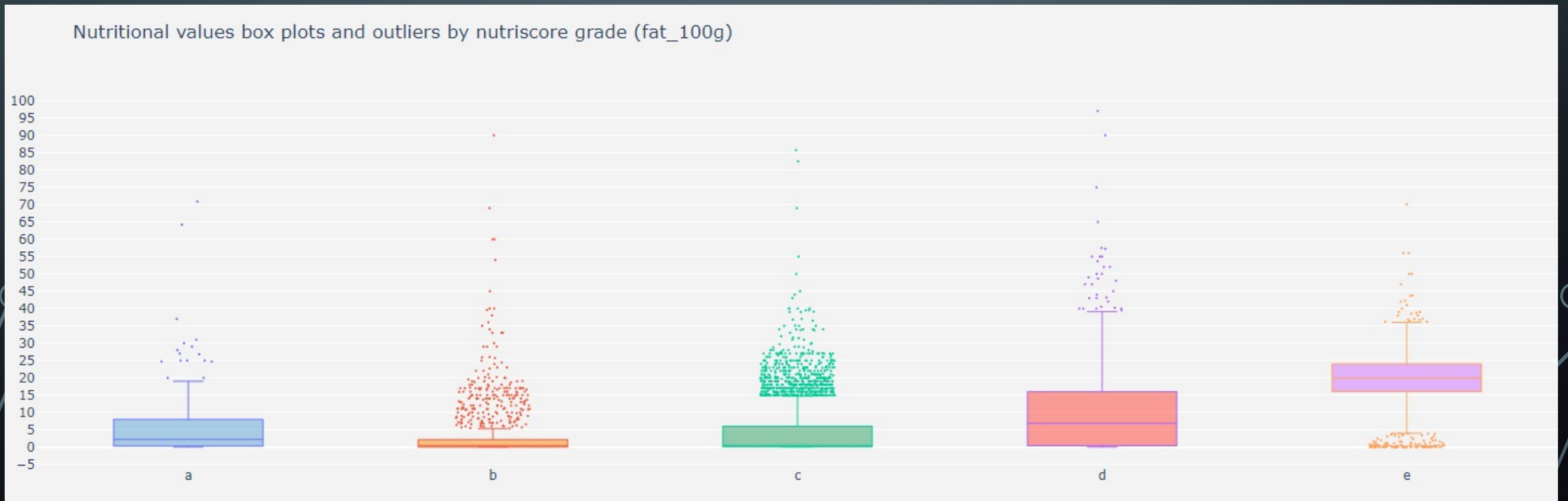
# ANALYSE BIVARIÉE – ANOVA

- Le groupe PNNS a un  $\eta^2$  de 0.27 en moyenne



# ANALYSE BIVARIÉE – NUTRISCORE GRADE & VALEURS NUTRITIONNELLES

- Boîtes à moustache des valeurs nutritionnelles par grade de nutriscore





# ANALYSE BIVARIÉE – NUTRISCORE GRADE & PNNS GROUPS 2

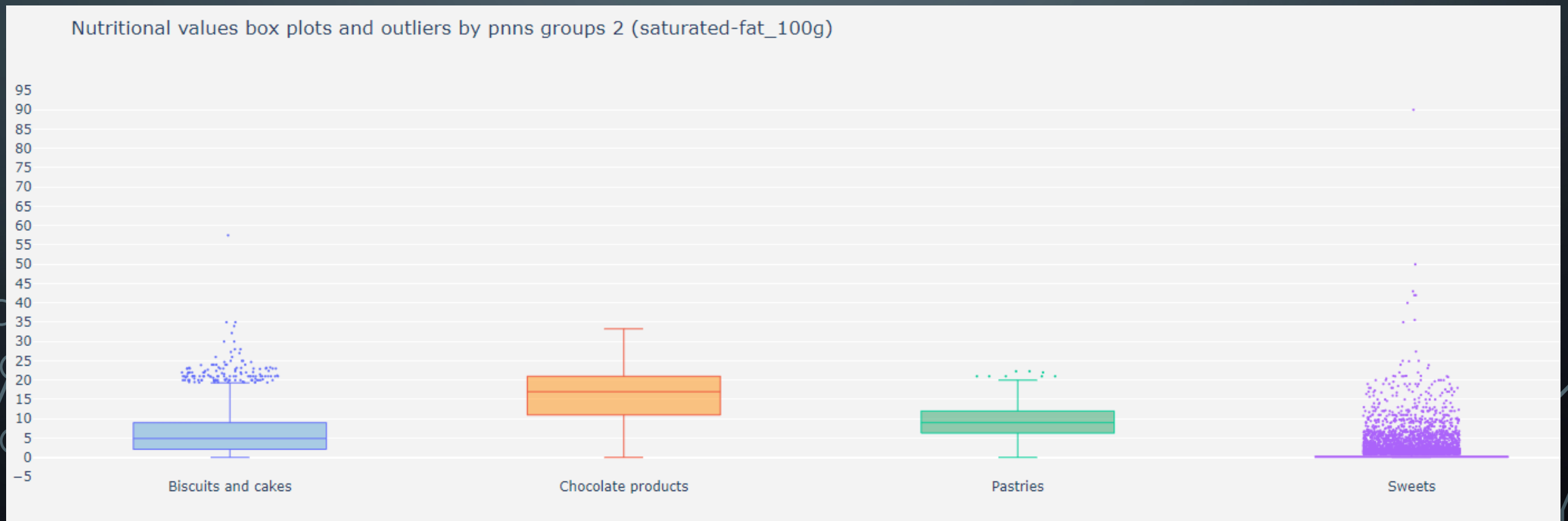
- Tableau de contingence des groupes PNNS par grade de nutriscore

Nutriscore grade repartition per PNNS groups 2 class

|                    | a   | b    | c    | d    | e    |
|--------------------|-----|------|------|------|------|
| Sweets             | 246 | 1056 | 4135 | 8854 | 581  |
| Pastries           | 8   | 42   | 199  | 1806 | 1504 |
| Chocolate products | 1   | 5    | 29   | 113  | 106  |
| Biscuits and cakes | 129 | 244  | 1724 | 8185 | 4037 |

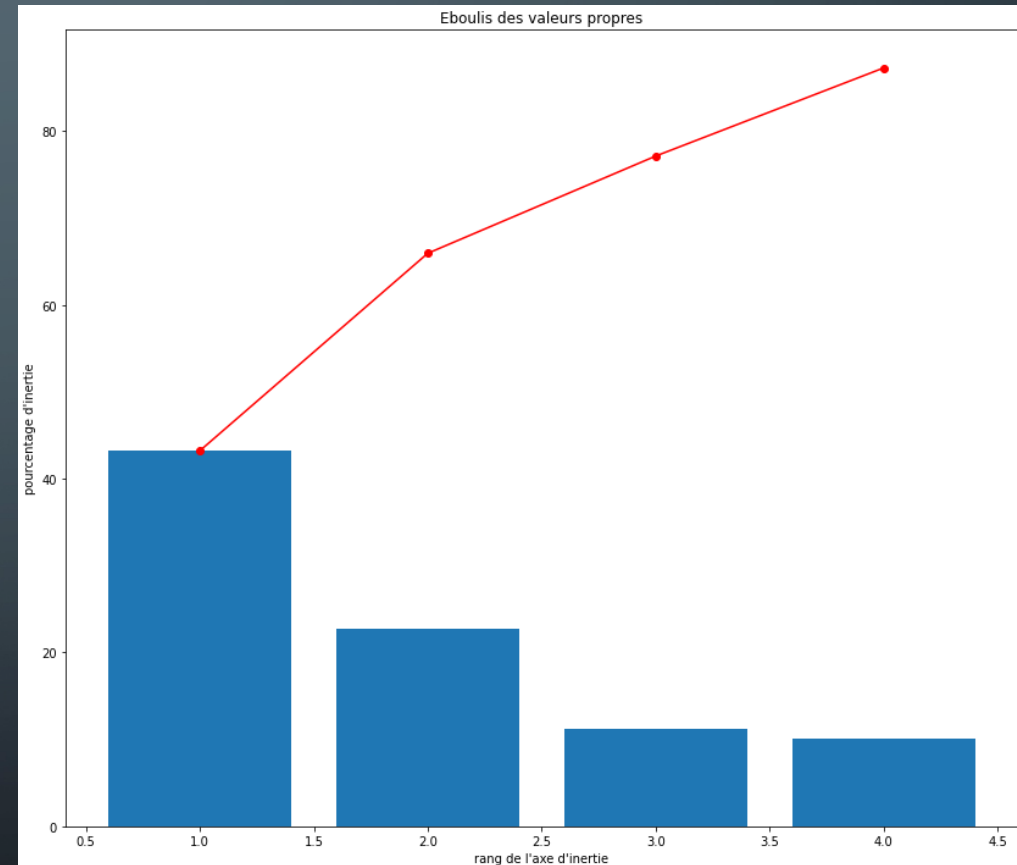
# ANALYSE BIVARIÉE – VALEURS NUTRITIONNELLES & PNNS GROUPS 2

- Boîtes à moustache des valeurs nutritionnelles par groupes PNNS

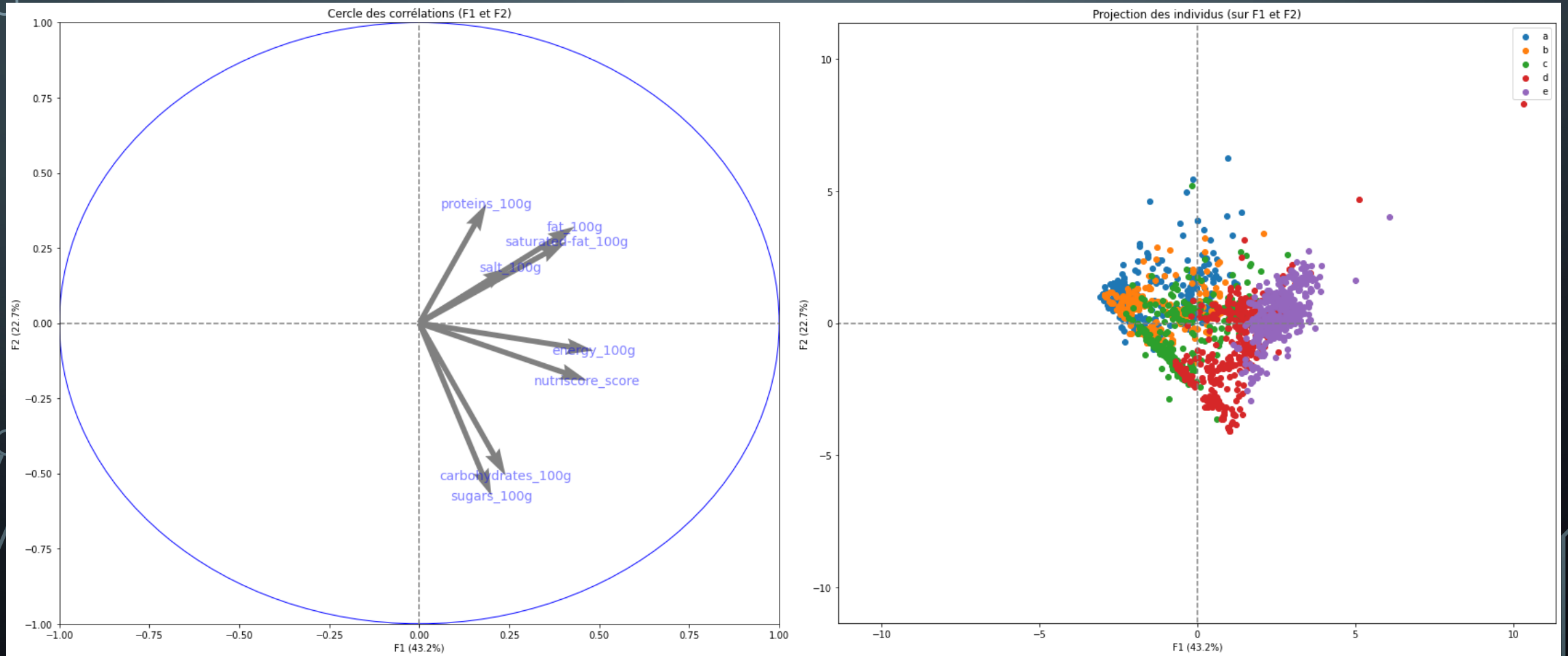


# ANALYSE MULTIVARIÉE – RÉDUCTION DIMENSIONNELLE

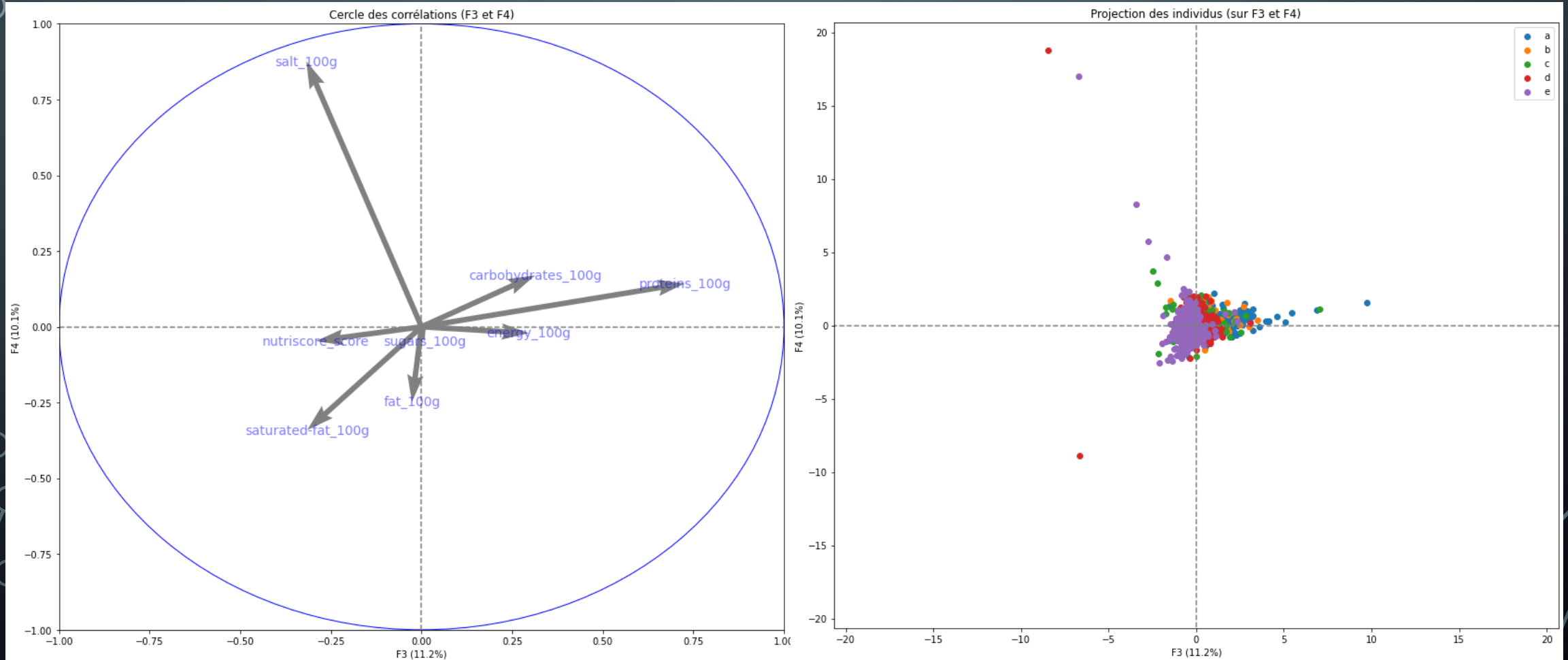
- Les 2 premiers axes représentent plus de 80 % de l'inertie



# ANALYSE MULTIVARIÉE – PROJECTION



# ANALYSE MULTIVARIÉE – PROJECTION





# ANALYSE MULTIVARIÉE – CORRÉLATIONS

- Matrice de corrélation aux axes par variable

| Features Axis Correlations |       |       |       |       |
|----------------------------|-------|-------|-------|-------|
|                            | COR_1 | COR_2 | COR_3 | COR_4 |
| nutriscore_score           | 0.87  | -0.26 | -0.28 | -0.04 |
| energy_100g                | 0.9   | -0.12 | 0.28  | -0.02 |
| proteins_100g              | 0.35  | 0.54  | 0.69  | 0.13  |
| fat_100g                   | 0.8   | 0.43  | -0.03 | -0.22 |
| carbohydrates_100g         | 0.45  | -0.68 | 0.3   | 0.15  |
| sugars_100g                | 0.37  | -0.77 | 0.01  | -0.04 |
| saturated-fat_100g         | 0.76  | 0.36  | -0.3  | -0.31 |
| salt_100g                  | 0.47  | 0.25  | -0.3  | 0.79  |

# ANALYSE MULTIVARIÉE – QUALITÉ DE REPRÉSENTATION

- Qualité de représentation des variables par axe

Quality of features representation

|                    | COS2_1 | COS2_2 | COS2_3 | COS2_4 |
|--------------------|--------|--------|--------|--------|
| nutriscore_score   | 0.75   | 0.07   | 0.08   | 0.0    |
| energy_100g        | 0.81   | 0.02   | 0.08   | 0.0    |
| proteins_100g      | 0.12   | 0.29   | 0.47   | 0.02   |
| fat_100g           | 0.64   | 0.19   | 0.0    | 0.05   |
| carbohydrates_100g | 0.2    | 0.47   | 0.09   | 0.02   |
| sugars_100g        | 0.14   | 0.6    | 0.0    | 0.0    |
| saturated-fat_100g | 0.58   | 0.13   | 0.09   | 0.09   |
| salt_100g          | 0.22   | 0.06   | 0.09   | 0.62   |

# ANALYSE MULTIVARIÉE – CONTRIBUTIONS AUX AXES

- Contribution des variables aux axes

| Features axis contributions |       |       |       |       |
|-----------------------------|-------|-------|-------|-------|
|                             | CTR_1 | CTR_2 | CTR_3 | CTR_4 |
| nutriscore_score            | 0.22  | 0.04  | 0.08  | 0.0   |
| energy_100g                 | 0.23  | 0.01  | 0.09  | 0.0   |
| proteins_100g               | 0.03  | 0.16  | 0.53  | 0.02  |
| fat_100g                    | 0.19  | 0.1   | 0.0   | 0.06  |
| carbohydrates_100g          | 0.06  | 0.26  | 0.1   | 0.03  |
| sugars_100g                 | 0.04  | 0.33  | 0.0   | 0.0   |
| saturated-fat_100g          | 0.17  | 0.07  | 0.1   | 0.12  |
| salt_100g                   | 0.06  | 0.03  | 0.1   | 0.77  |

# CONCLUSION

- Base de Donnée nettoyée
- Preuve de concept
- Axes d'amélioration