

# **Concevez une application au service de la santé publique**

**Appel à projet de Santé Publique France**

**Joyce Kuoh Moukouri,  
P3, Soutenance du 19/02/2023**

# Ordre du jour

## Concevez une application au service de la santé publique

1. La mission
2. L'application ***snack control***
3. Présentation de la base de données
4. Nettoyage des données
5. Exploration des données

Conclusion

# 1. La mission

# **La mission**

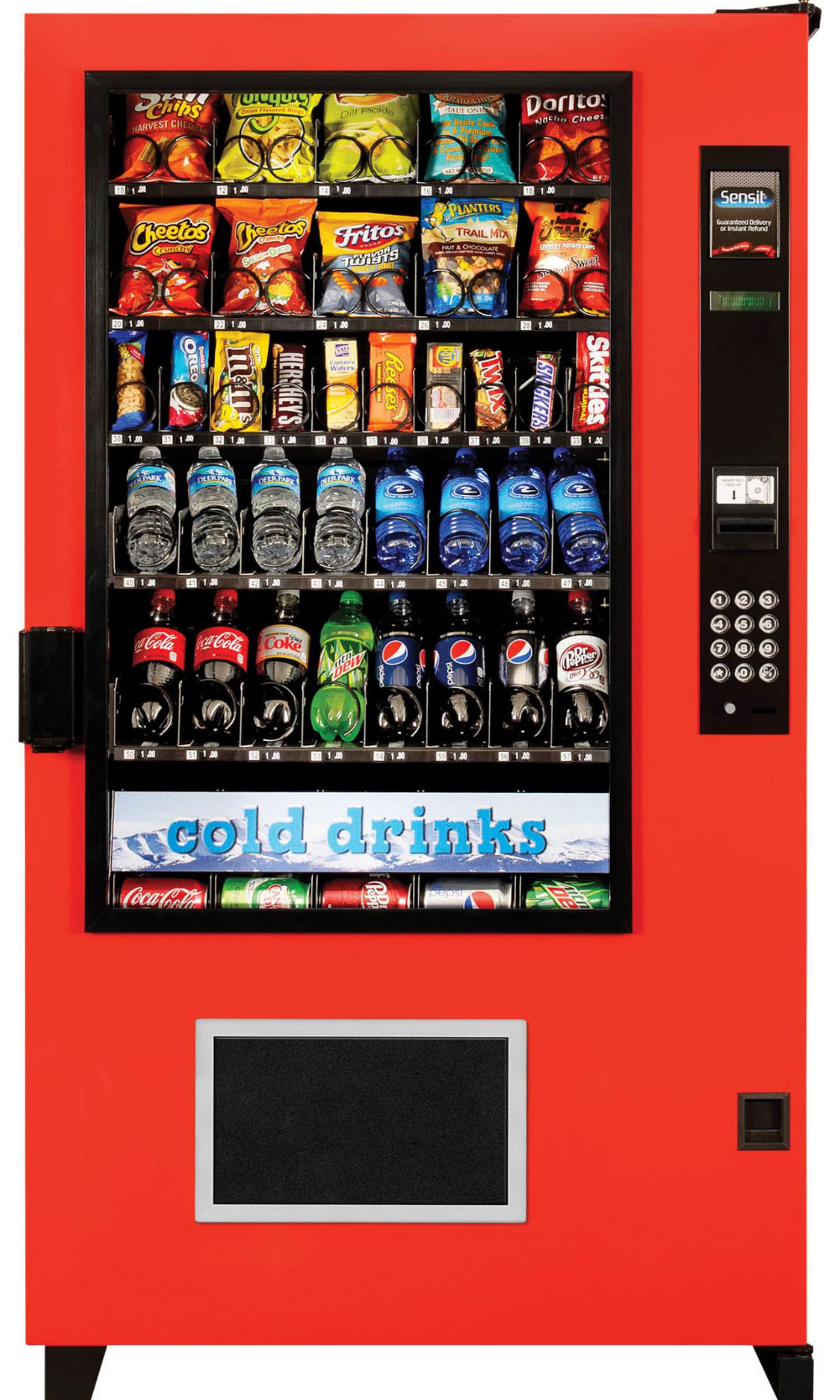
## **Rappel des objectifs fixés par Santé Publique France**

- Proposer une idée d'application en lien avec la nutrition
- Nettoyage et exploration du jeu de données OPEN FOOD FACTS

## **2. L'application *snack control***

## 2. L'application *snack control*

- Recommandation de l'OMS : maximum 50g de sucre libre/j soit 10 morceaux de sucre DADDY n°5.
- Risques associés à l'excès de sucre : diabète de type II, obésité, dépression
- ***Snack control***, permet de surveiller sa consommation de sucre lors des pauses cafés et goûter.



## 2. L'application *snack control*

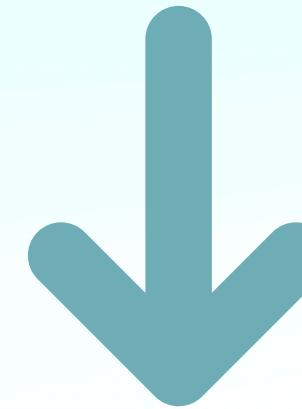


Pour un code bar EAN13 ou GTIN, ou la désignation du produit, l'application renvoie la quantité de sucres simples consommés et son équivalent en nombre de morceaux de sucre

# **3. La base de données OPEN FACTS FOOD**

### **3. La base de données**

- Base de données OPEN FOOD FACTS
- Une table comportant 197 colonnes et 2 718 738 lignes (7Go)
- Clé : Code barre du produit
- Données accessibles en open source et renseignées par une équipe de bénévoles et/ou des utilisateurs lambda



#### **Qualité du jeu de données : médiocre**

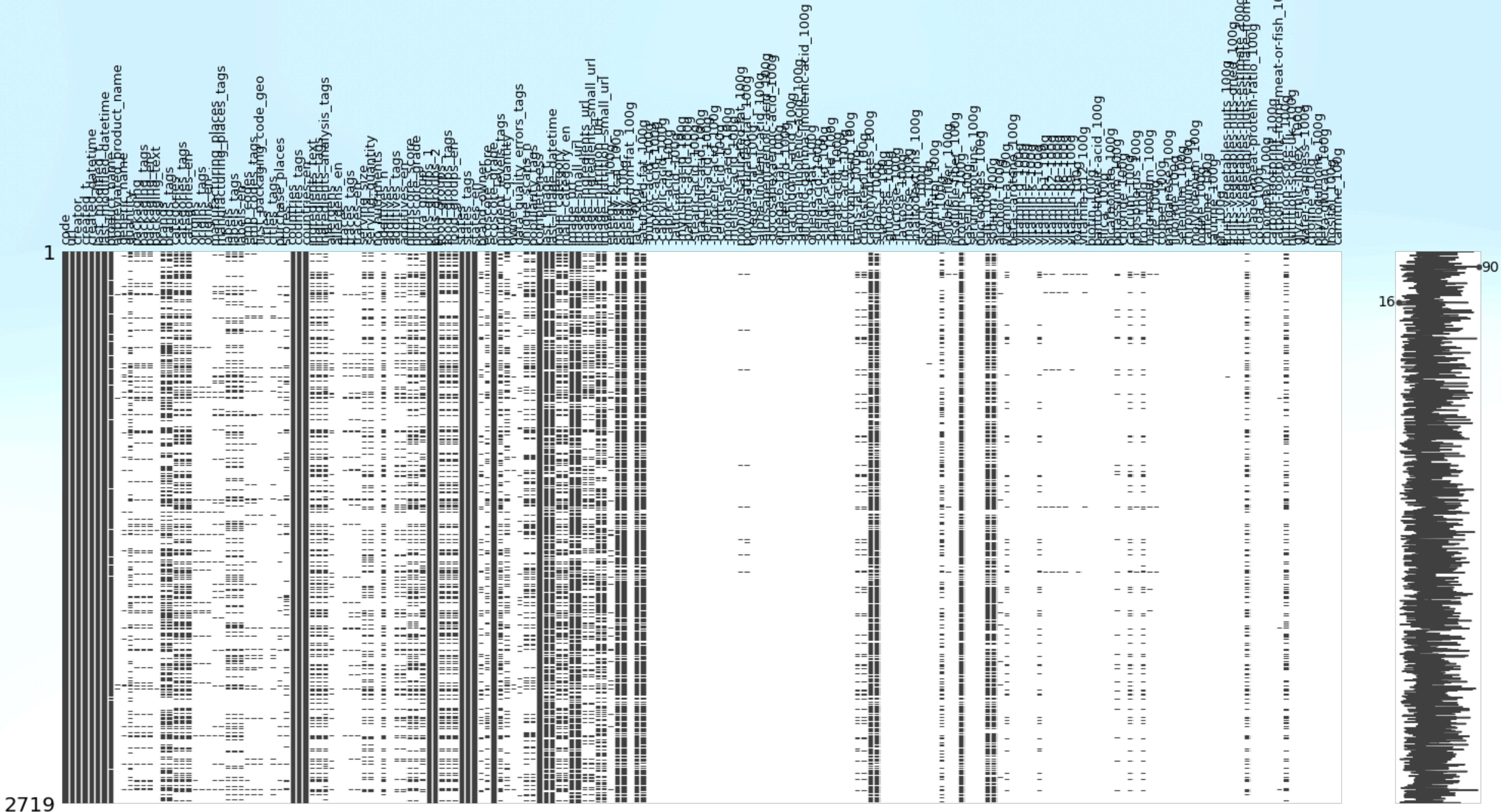
- Valeurs manquantes : taux de valeur manquantes moyen à plus de 70%
- Nombreuses erreurs et valeurs aberrantes
- « Vandalisme »

## Exemple de « vandalisme »

```
Entrée [168]: df[(df['SALT_100G']==100) & (~df['PRODUCT_NAME'].str.contains('SEL')) & (~df['PRODUCT_NAME'].str.contains('SAL'))]
```

78854	889698465823	<a href="https://world.openfoodfacts.org/product/088969...">https://world.openfoodfacts.org/product/088969...</a>	CREPE AU GEL HIDROALCOOLIQUE	FRANCE
760871	7896383053312	<a href="https://world.openfoodfacts.org/product/789638...">https://world.openfoodfacts.org/product/789638...</a>	TUBITOS	FRANCE
656266	4982978702185	<a href="https://world.openfoodfacts.org/product/498297...">https://world.openfoodfacts.org/product/498297...</a>	ENCRE CHIMIQUE	FRANCE
759061	7702354005146	<a href="https://world.openfoodfacts.org/product/770235...">https://world.openfoodfacts.org/product/770235...</a>	AJIACO	FRANCE
758596	7640342961787	<a href="https://world.openfoodfacts.org/product/764034...">https://world.openfoodfacts.org/product/764034...</a>	PEA CHICKEN BOWL	FRANCE
718884	5901646276673	<a href="https://world.openfoodfacts.org/product/590164...">https://world.openfoodfacts.org/product/590164...</a>	IPAD ECOUTEUR SANS FIL	FRANCE
662078	5010994966324	<a href="https://world.openfoodfacts.org/product/501099...">https://world.openfoodfacts.org/product/501099...</a>	POT DE PÂTE À MODELER - PLAY DOH - ASSORTIMENT	FRANCE
715854	5707644231696	<a href="https://world.openfoodfacts.org/product/570764...">https://world.openfoodfacts.org/product/570764...</a>	NICOLAS VAHÉ PARMESAANI- JUUSTO JA BASILIKI SUO...	FRANCE

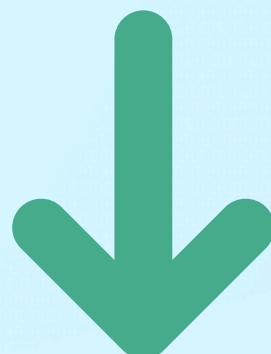
# 3. La base de données



# 4. Nettoyage des données

# 4. Nettoyage des données

## Démarche en 4 phases



PHASE 1

Sélection des variables

Sauvegarde du df dans un .csv



PHASE 2

Mise en forme des variables qualitatives et catégorielles

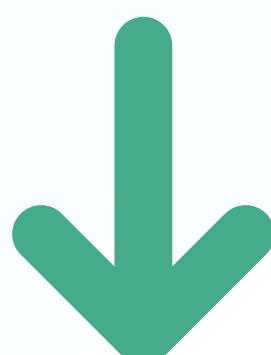
Traitement des variables aberrantes quantitatives



PHASE 3

Traitement des doublons

Sauvegarde du df dans un .csv



PHASE 4

Imputation des valeurs manquantes

Sauvegarde du df dans un .csv

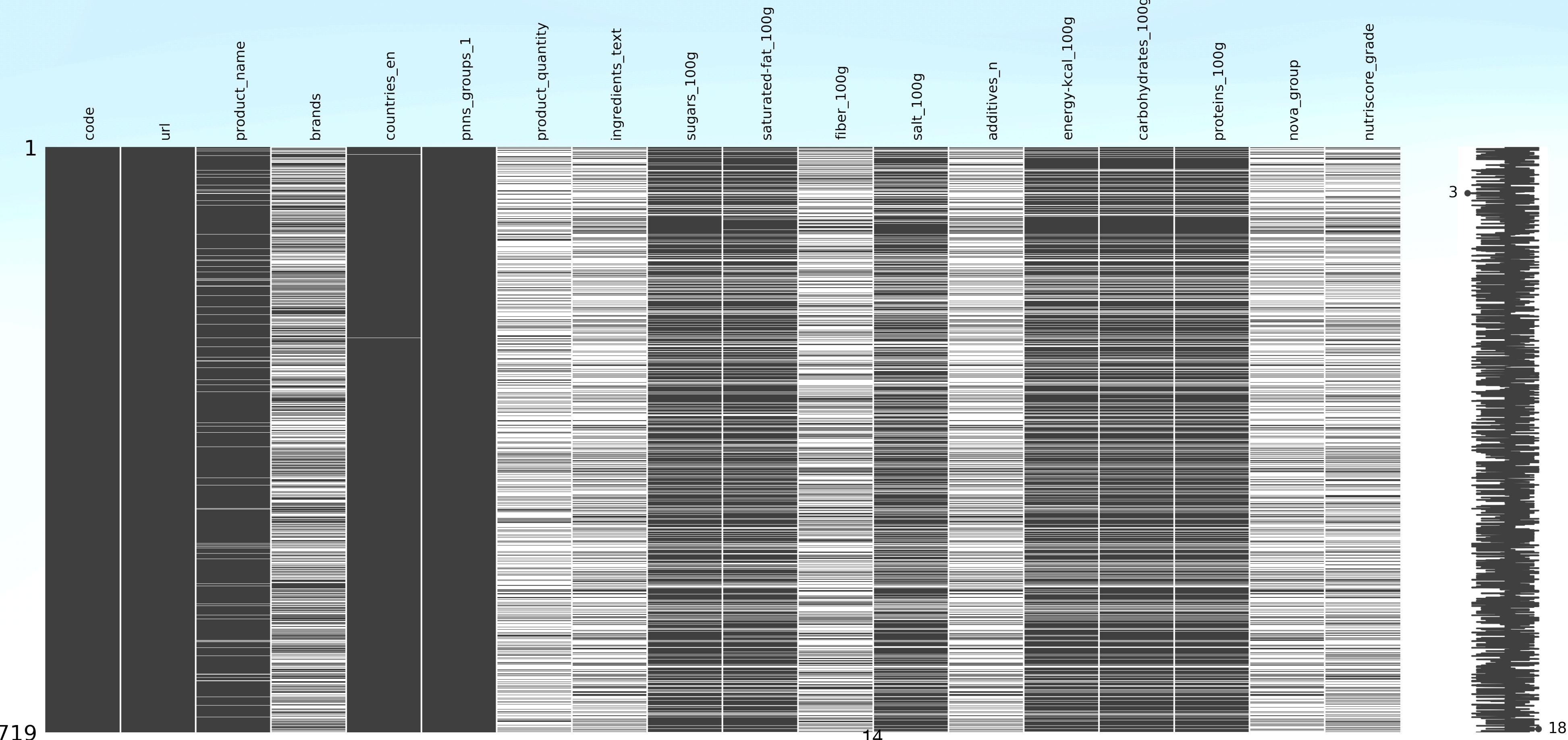
# 4. Nettoyage des données

PHASE 1

Sélection des variables

Sauvegarde du df dans un .csv

- 18 variables sélectionnées



# 4. Nettoyage des données

PHASE 2

Mise en forme des variables qualitatives et catégorielles

```
def homostr(df):
    df.columns = df.columns.str.upper()
    # 1- Sélection des variables de type object (ou str) et passage en lettres capitales
    var_obj = df.dtypes[df.dtypes == object].index
    for i in var_obj:
        print('variable',i)
        df[i] = [str(x) for x in df[i]]
        df[i] = df[i].str.upper()
        df[i] = df[i].str.strip()
    return df
```

Implémentation de la fonction **homostr**

- Pour toutes les variables qualitatives, met toutes les valeurs en lettres capitales
- Supprime les espaces indésirables

# 4. Nettoyage des données

PHASE 2

Mise en forme des variables qualitatives et catégorielles

	Anomalies	Traitement	Aller plus loin
URL	<ul style="list-style-type: none"><li>Aucun des URLs ne fonctionnent.</li><li>Certains URLs ne comprennent pas de code</li></ul>	<ul style="list-style-type: none"><li>Remplacement '<u>WORLD-EN.OPENFOODFACTS.ORG/PRODUCT</u>' par '<u>world.openfoodfacts.org/product</u>' (suppression du '-en')</li><li>Suppression des lignes (product_name vide)</li></ul>	Vérifier la page de chacune des URLs
CODE	<ul style="list-style-type: none"><li>Certains codes sont erronés</li></ul>	<ul style="list-style-type: none"><li>Création de la variable LEN_CODE (longueur du code)</li><li>Sélection d'un échantillon tel que <math>\text{LEN\_CODE} \in ]7; 14[</math></li></ul>	Vérifier la validité de chaque code barre

# 4. Nettoyage des données

PHASE 2

Mise en forme des variables qualitatives et catégorielles

	Anomalies	Traitement	Aller plus loin
<b>URL</b>	<ul style="list-style-type: none"> <li>Aucun des URLs ne fonctionnent.</li> <li>Certains URLs ne comprennent pas de code</li> </ul>	<ul style="list-style-type: none"> <li>Remplacement '<u>WORLD-EN.OPENFOODFACTS.ORG/PRODUCT</u>' par '<u>world.openfoodfacts.org/product</u>' (suppression du '-en')</li> <li>Suppression des lignes (product_name vide)</li> </ul>	Vérifier la page de chacune des URLs
<b>CODE</b>	<ul style="list-style-type: none"> <li>Certains codes sont erronés</li> </ul>	<ul style="list-style-type: none"> <li>Création de la variable LEN_CODE (longueur du code)</li> <li>Sélection d'un échantillon tel que <math>\text{LEN\_CODE} \in [7; 14[</math></li> </ul>	Vérifier la validité de chaque code barre
<b>PRODUCT_NAME</b>	<ul style="list-style-type: none"> <li>Désignation manquantes</li> <li>Désignation longues</li> </ul>	<ul style="list-style-type: none"> <li>Remplacement des désignations manquantes par np.nan</li> <li>Utilisation du langage regex pour supprimer les caractères spéciaux, les chiffres, et conservation des 3 premiers mots</li> </ul>	
<b>BRANDS</b>	Idem	Idem	Lier le code barre à la marque

# 4. Nettoyage des données

PHASE 2

Mise en forme des variables qualitatives et catégorielles

	Anomalies	Traitement	Aller plus loin
<b>URL</b>	<ul style="list-style-type: none"> <li>Aucun des URLs ne fonctionnent.</li> <li>Certains URLs ne comprennent pas de code</li> </ul>	<ul style="list-style-type: none"> <li>Remplacement '<u>WORLD-EN.OPENFOODFACTS.ORG/ PRODUCT</u>' par '<u>world.openfoodfacts.org/ product</u>' (suppression du '-en')</li> <li>Suppression des lignes (product_name vide)</li> </ul>	Vérifier la page de chacune des URLs
<b>CODE</b>	<ul style="list-style-type: none"> <li>Certains codes sont erronés</li> </ul>	<ul style="list-style-type: none"> <li>Création de la variable LEN_CODE (longueur du code)</li> <li>Sélection d'un échantillon tel que <math>\text{LEN\_CODE} \in ]7; 14[</math></li> </ul>	Vérifier la validité de chaque code barre
<b>PRODUCT_NAME</b>	<ul style="list-style-type: none"> <li>Désignation manquantes</li> <li>Désignation longues</li> </ul>	<ul style="list-style-type: none"> <li>Création de la variable LEN_PRODUCT</li> <li>Remplacement des désignation manquantes par np.nan</li> <li>Utilisation du langage regex pour supprimer les caractères spéciaux, les chiffres, et conservation des 3 premiers mots</li> </ul>	
<b>BRANDS</b>	Idem	Idem	Lier le code barre à la marque
<b>INGREDIENT_TXT</b>	Idem	Idem	

# 4. Nettoyage des données

PHASE 2

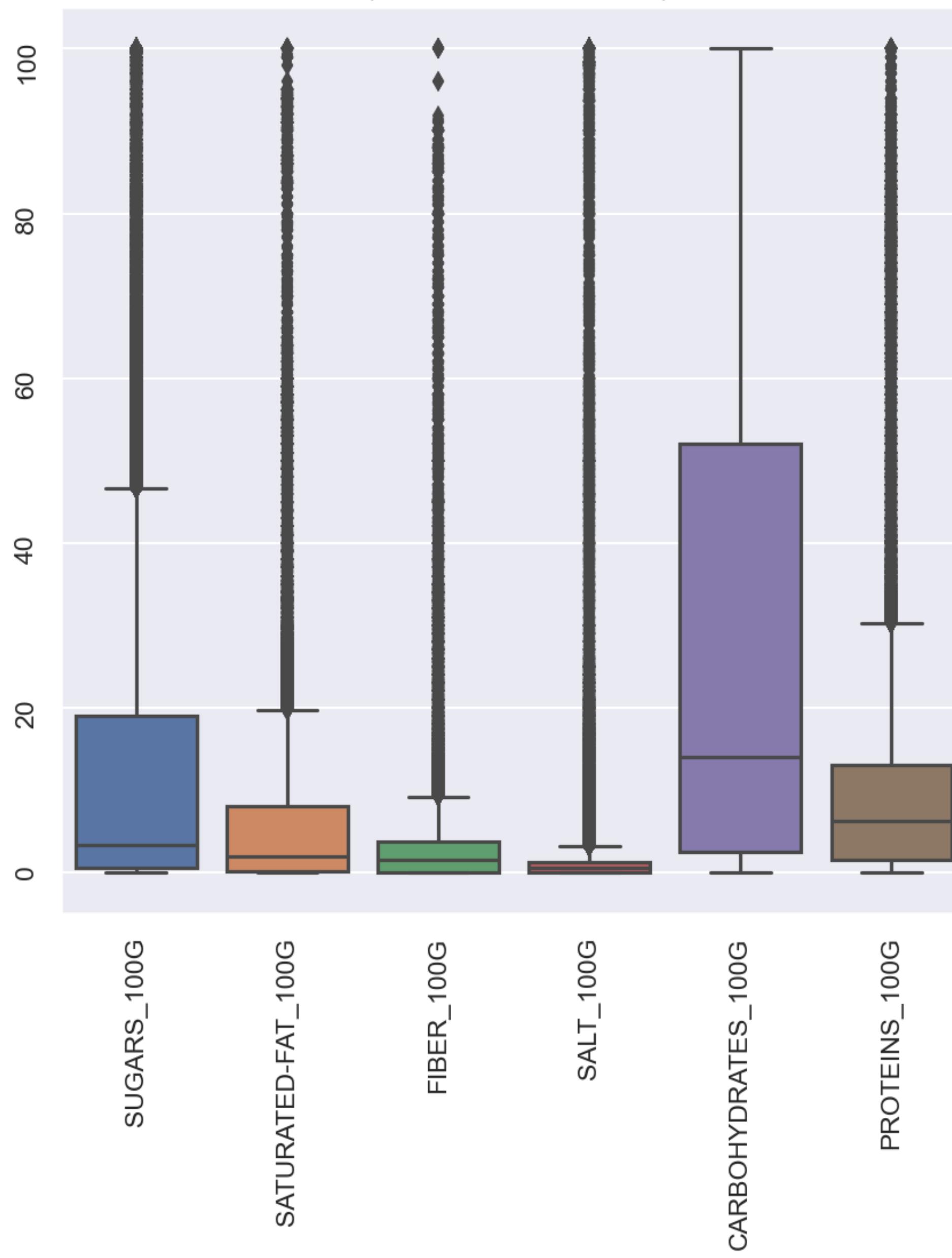
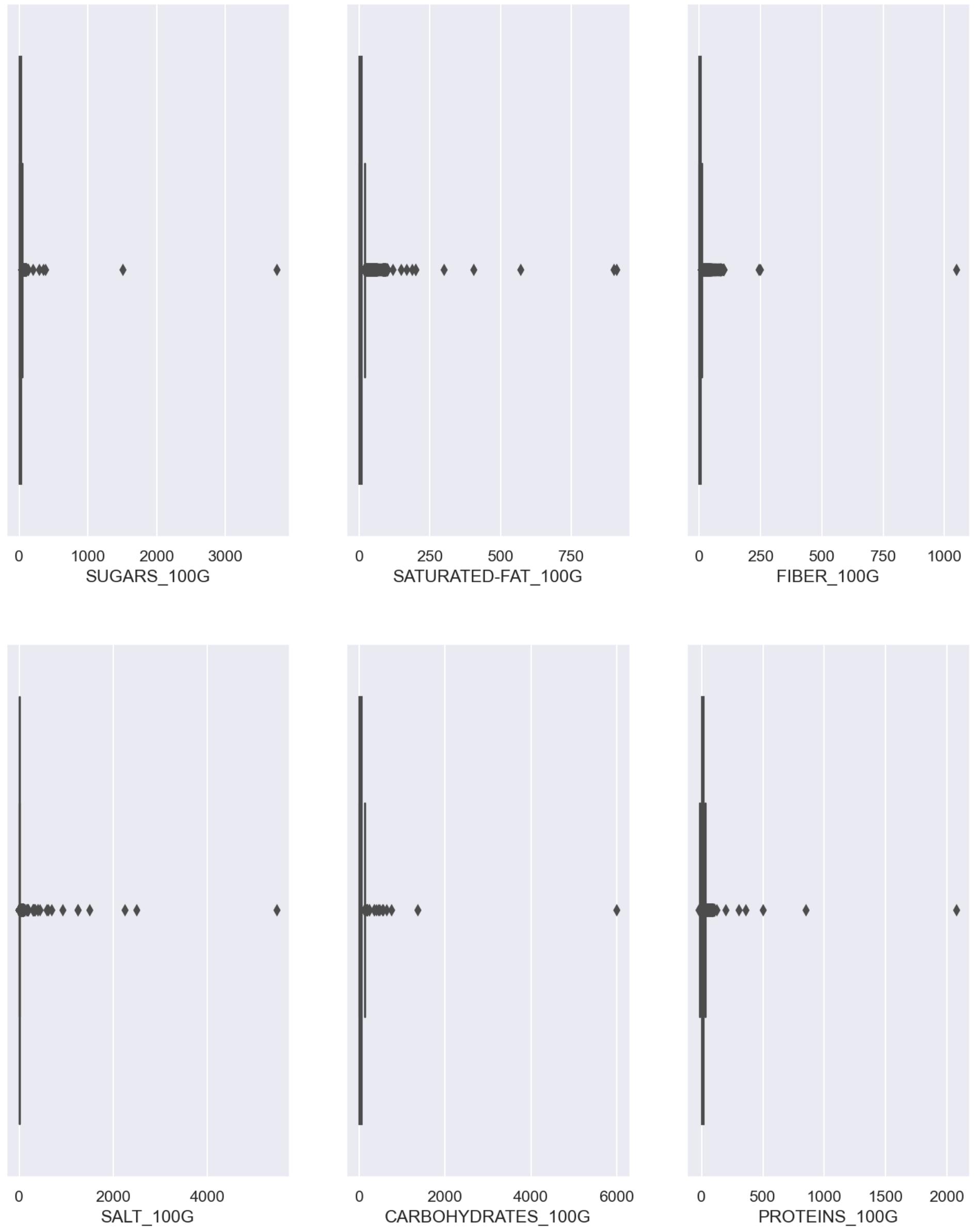
Traitement des variables aberrantes quantitatives

- S'assurer que les variables nutritives sont comprises entre 0 et 100g
- La variable ENERGY\_KCAL\_100G doit être comprise entre 0 et 900kcal
- Les valeurs aberrantes sont remplacées par ***np.nan***

```
#-----
#Traitement des valeurs aberrantes
def val_aberrante(df,col,val_min, val_max):
    if (df[col].max() > val_max):
        ind = df[df[col] > val_max].index
        for i in ind:
            df.loc[i,col] = np.nan
    else:
        print(col,'Pas de valeurs >', val_max)
    if (df[col].min() < val_min):
        ind = df[df[col] < val_min].index
        for i in ind:
            df.loc[i,col] = np.nan
    else:
        print(col,'Pas de valeurs <', val_min)

return df
```

Visualisation de la dispersion des variables comprises entre 0 et 100  
(avec valeurs extrêmes)



# 4. Nettoyage des données

PHASE 3

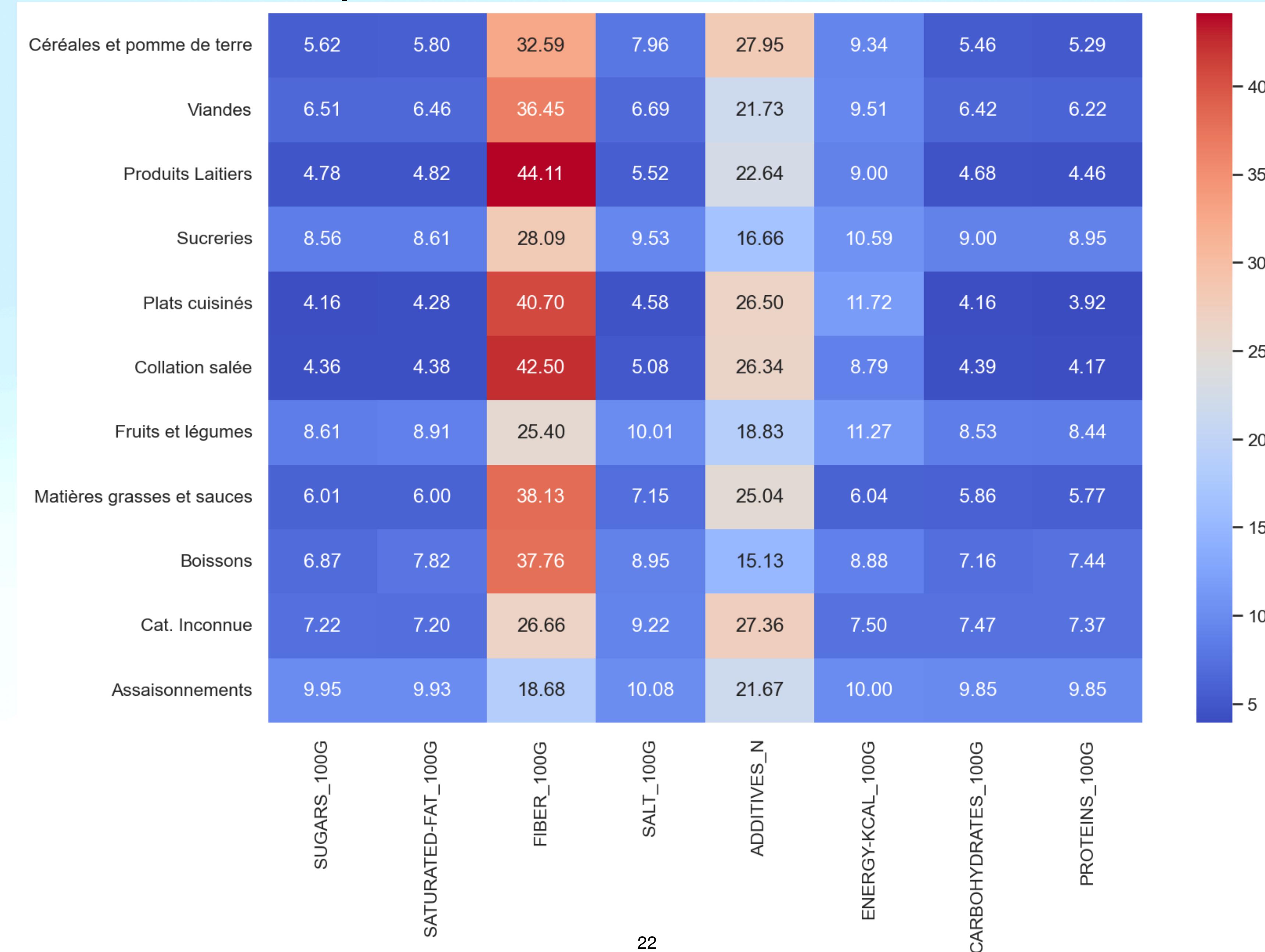
Traitement des doublons

## Implémentation de la fonction **doublon**

- Pour un dataframe **df** et une liste constituant une clé primaire **ls\_col** : **[CODE]**, puis **['PRODUCT\_NAME','BRANDS','PRODUCT\_QUANTITY']**
- Compte le nombre de doublons
- Conserve le plus d'informations possibles

```
#-----  
#Traitement des doublons  
def doublon(df,ls_col):  
    #Recherche les doublons, ls_col liste des colonnes qui constitue une clé primaire  
    print('Recherche de doublon : il y a ',  
        df.duplicated(ls_col,keep=False).sum(),  
        '\ndoublons qui ont la même clé:',ls_col )  
#-----  
#Suppression des doublons  
# on compte le nombre de valeurs manquantes pour la ligne et on stocke dans une nouvelle colonne  
df['NB_NAN'] = df.isna().sum(axis=1)  
# trie des lignes en fonction du nombre de valeurs manquantes  
df= df.sort_values('NB_NAN')  
# suppression des duplicates en gardant les versions les mieux remplies  
df= df.drop_duplicates(ls_col, keep='first')  
# on supprime la colonne qui n'est plus utile  
df= df.drop('NB_NAN', axis=1)  
df.head()  
return df
```

## 4. Nettoyage des données | Corrélation entre valeurs valeurs manquantes



# 4. Nettoyage des données

PHASE 4

Imputation des valeurs manquantes

	Imputation par hypothèses	Remplacement par 0	Remplacement par la médiane	Iterative Imputer	KNN Imputer	Aller plus loin
Scope	Les cas particuliers (sucres, sel, huile)	Le taux de fibres	Le taux de sel	Variables quantitatives corrélées	Variables quantitatives	Pour les variables catégorielles et qualitatives
Application	Pour le sucre : taux de sucre et glucide à 100, le reste des nutriments à 0 + imputation du groupe alimentaire	Taux de fibres dans les produits laitiers, viandes, les huiles et les boissons	-	Taux de sucre et taux de glucides	5 voisins, pondération par rapport à la distance	Utiliser la distance de levenshtein pour imputer les groupes manquants
Limites de la méthode				Uniquement pour les variables corrélées, précision de valeurs limites, la somme des nutriment >100g	Longue : méthode appliquée par morceaux, la somme des nutriment >100g	Méthode à implémenter 'from scratch'

# 5. Exploration des données

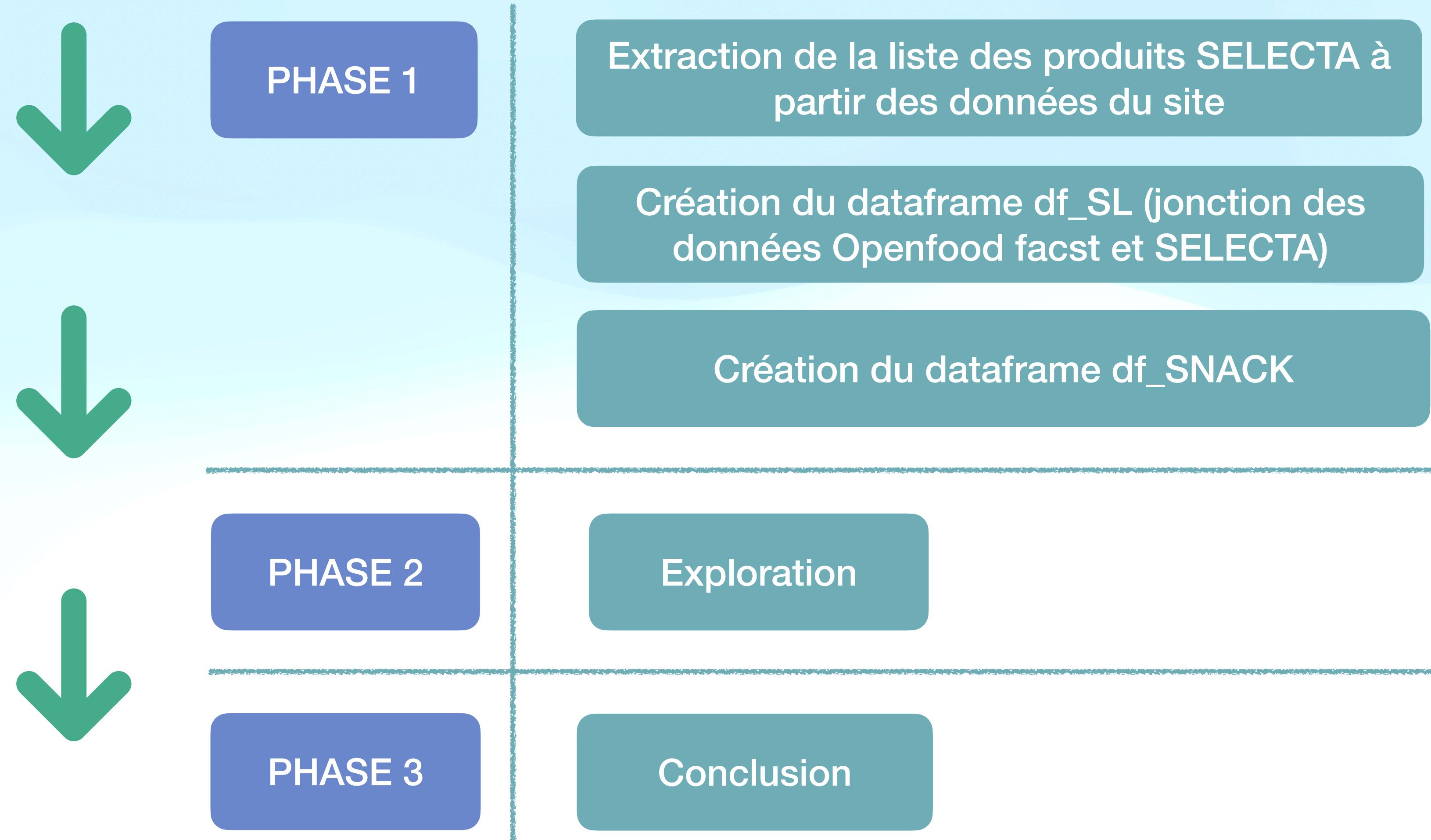
- a. Analyses univariées
- b. Analyses bivariées
- c. Analyses multivariées

## Hypothèse à explorer :

- Les goûters proposés dans les distributeurs peuvent être responsables d'une grande partie de notre consommation de sucre journalière. Un goûter = un snack sucré et une boisson.

# 5. Exploration des données

## Exemple des distributeurs SELECTA - démarche



# 5. Exploration des données

## Exemple des distributeurs SELECTA - démarche

The screenshot shows a Safari browser window with the Selecta website loaded. The website has a teal header with navigation links like 'Safari', 'Fichier', 'Édition', etc., and a search bar. The main content area displays a grid of six product images: 'BTE OASIS TROPICAL 33CL', 'BTE ICE TEA PECHE 33CL X24', 'BTE ICE TEA PECHE ZERO', 'PET TROPICANA ORANGE', 'PET LIPTON FRAMBOISE', and 'BTE TREE TOP JUS ORANG'. Below the products, there are two dropdown menus: 'Filtre : Disponibilité' and 'Trier par : MEILLEURES VENTE'. A red button labeled 'NOUS CONTACTER' is visible. On the right, a 'JOY TO GO' logo is present. The bottom of the screen shows the developer tools' 'Console' tab, which is displaying a large amount of JavaScript code related to Shopify Analytics and product meta-data.

```
416 <script>window.ShopifyAnalytics = window.ShopifyAnalytics || {};
417 window.ShopifyAnalytics.meta = window.ShopifyAnalytics.meta || {};
418 var meta = {"products": [{"id": 7799392141534, "gid": "gid://shopify/Product/7799392141534", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43403597349086, "price": 2640, "name": "BTE OASIS TROPICAL 33CL X24", "public_title": null, "sku": "FR_I0000120"}]}, {"id": 7799393419486, "gid": "gid://shopify/Product/7799393419486", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43403598692574, "price": 2400, "name": "BTE ICE TEA PECHE 33CL X24", "public_title": null, "sku": "FR_I0000238"}]}, {"id": 7799393485022, "gid": "gid://shopify/Product/7799393485022", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43403598758110, "price": 1968, "name": "BTE ICE TEA PECHE ZERO 33CLX24", "public_title": null, "sku": "FR_I0000239"}]}, {"id": 7799394992350, "gid": "gid://shopify/Product/7799394992350", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43403600298206, "price": 2580, "name": "PET TROPICANA ORANGE 12X25CL", "public_title": null, "sku": "FR_I0000850"}]}, {"id": 7799398629598, "gid": "gid://shopify/Product/7799398629598", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43403604852958, "price": 1884, "name": "PET LIPTON FRAMBOISE 50CL X12", "public_title": null, "sku": "FR_I0002251"}]}, {"id": 7799394336990, "gid": "gid://shopify/Product/7799394336990", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43403599708382, "price": 1752, "name": "BTE TREE TOP JUS ORANG 33CLX24", "public_title": null, "sku": "FR_I0000611"}]}, {"id": 7895283925214, "gid": "gid://shopify/Product/7895283925214", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43720507556062, "price": 2040, "name": "PET PULCO CITRONNADE 50CLX12", "public_title": null, "sku": "FR_I0000134"}]}, {"id": 7873780449502, "gid": "gid://shopify/Product/7873780449502", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43662511505630, "price": 2400, "name": "PET LIPTON GREEN MENTHE 50CLX12", "public_title": null, "sku": "FR_I0002445"}]}, {"id": 7873780383966, "gid": "gid://shopify/Product/7873780383966", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43662511440094, "price": 2640, "name": "PET OASIS POMME CASSIS 50CLX12", "public_title": null, "sku": "FR_I0002306"}]}, {"id": 7873780777182, "gid": "gid://shopify/Product/7873780777182", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43662511866078, "price": 3840, "name": "BTE OASIS THE PECH 33CL X24", "public_title": null, "sku": "FR_I0003101"}]}, {"id": 7799401513182, "gid": "gid://shopify/Product/7799401513182", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43403608096990, "price": 2040, "name": "PET MAYTEA MENTHE 50CL X12", "public_title": null, "sku": "FR_I0003094"}]}, {"id": 787208517854, "gid": "gid://shopify/Product/787208517854", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43657842950366, "price": 200, "name": "BTE OASIS POMME CASSIS X24", "public_title": null, "sku": "FR_I0001720"}]}, {"id": 7799394599134, "gid": "gid://shopify/Product/7799394599134", "vendor": "Selecta FR", "type": "", "variants": [{"id": 4304359970526, "price": 2100, "name": "PET OASIS TROPICAL 50CL X12", "public_title": null, "sku": "FR_I0000640"}]}, {"id": 7799400038622, "gid": "gid://shopify/Product/7799400038622", "vendor": "Selecta FR", "type": "", "variants": [{"id": 43403606327518, "price": 1380, "name": "PET LIPTON PECH ZERO 50CL X12", "public_title": null, "sku": "FR_I0002848"}]}], "page": {"pageType": "collection", "resourceType": "collection", "resourceId": 408487133406};

for (var attr in meta) {
    window.ShopifyAnalytics.meta[attr] = meta[attr];
}
</script>
<script>window.ShopifyAnalytics.merchantGoogleAnalytics = function() {
};</script>
```

# 5. Exploration des données

## Exemple des distributeurs SELECTA - démarche

Entrée [21]: df\_sel

Out[21]:

	Produit	catégorie
0	CRISTALINE	Boissons
1	CRAQUISE CHOCO NOISETT	Sucreries
2	KITKAT POPS LAIT	Sucreries
3	M M S CRIPSY	Sucreries
4	COTE D'OR NOISETTES -	Sucreries
...	...	...
97	COCA COLA ZERO	Boissons
98	PEPSI COLA	Boissons
99	COCA COLA SVSUCRE	Boissons
100	PEPSI	Boissons
101	GAUFRE MIEL BJORG	Sucreries

102 rows × 2 columns

# 5. Exploration des données

## Exemple des distributeurs SELECTA - démarche

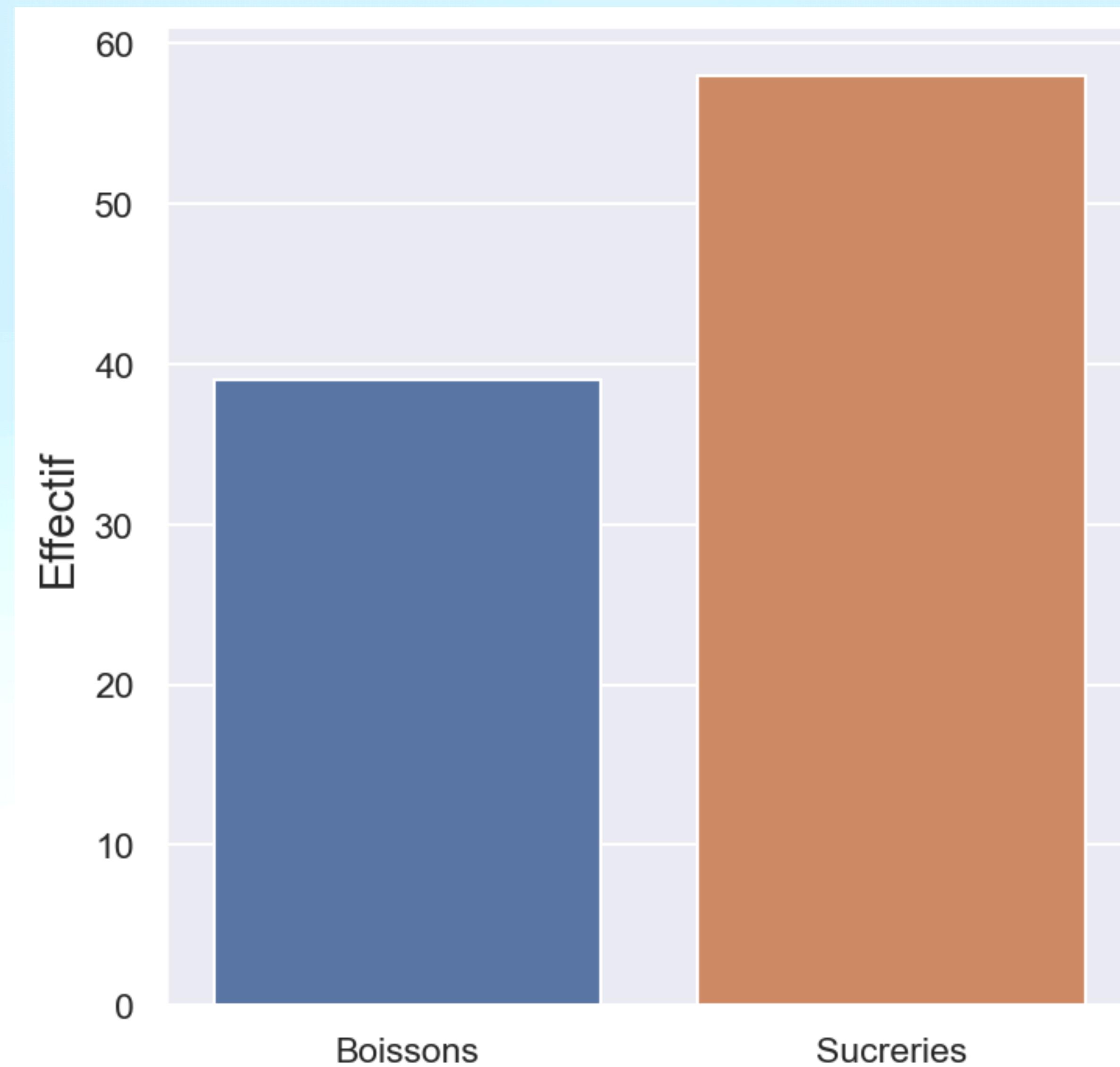
Entrée [21]: df\_SNACK.head()

Out[21]:

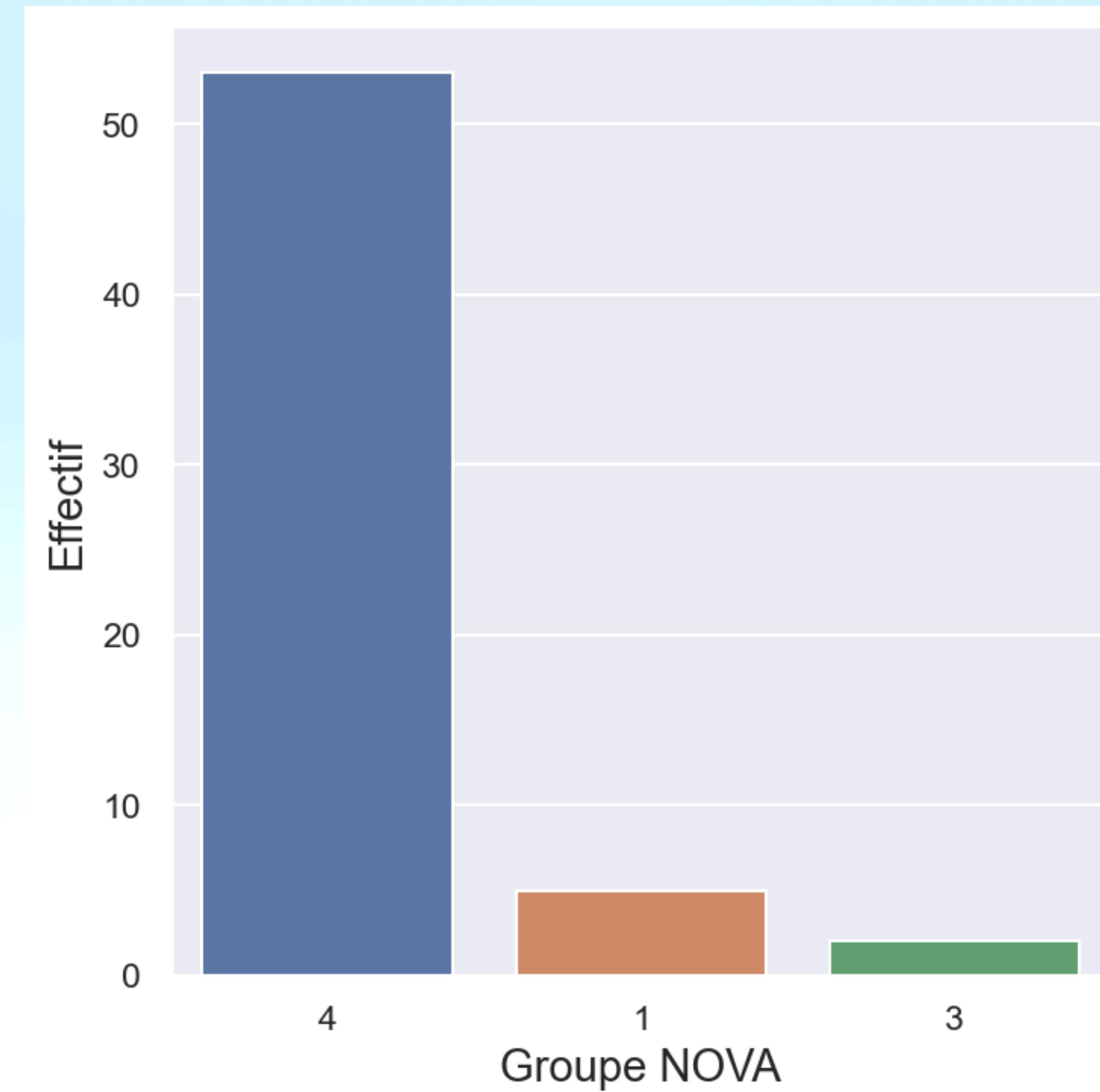
	Boissons	Sucreries	SUCRES_100G	GLUCIDES_100G	GRAS-SAT_100G	FIBRES_100G	SEL_100G	PROTÉINES_100G	NB_ADDITIFS	ENERGIE-KCAL_100G	... ENEI KCAL_100G
0	CRISTALINE	CRAQUISE CHOCO NOISETTE	49.0	73.0	9.4	0.0	0.600	4.2	0.800000	481.0	...
1	CRISTALINE	KITKAT POPS LAIT	35.8	50.3	14.5	3.5	1.030	10.5	5.000000	529.0	...
2	CRISTALINE	M M S CRIPSY	55.5	59.2	18.0	0.0	0.202	6.0	1.038543	535.0	...
3	CRISTALINE	COTE D'OR NOISETTES -	73.5	74.5	9.6	0.0	0.110	1.9	0.000000	455.0	...
4	CRISTALINE	KINDER BUENO WHITE	43.6	53.0	19.8	0.0	0.456	8.8	3.000000	572.0	...

5 rows × 31 columns

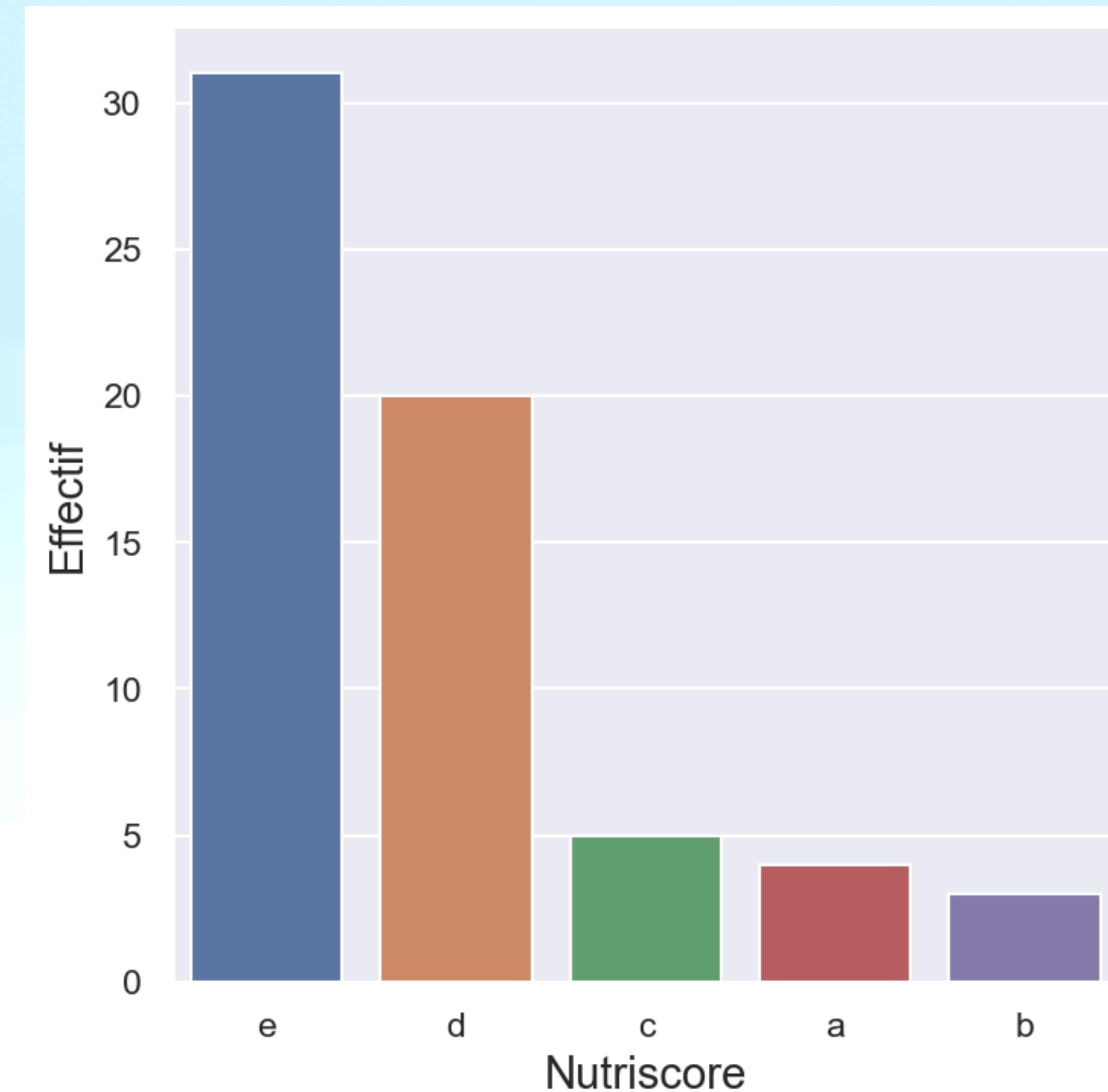
## 5. Exploration des données | Analyses univariées | Variables catégorielles



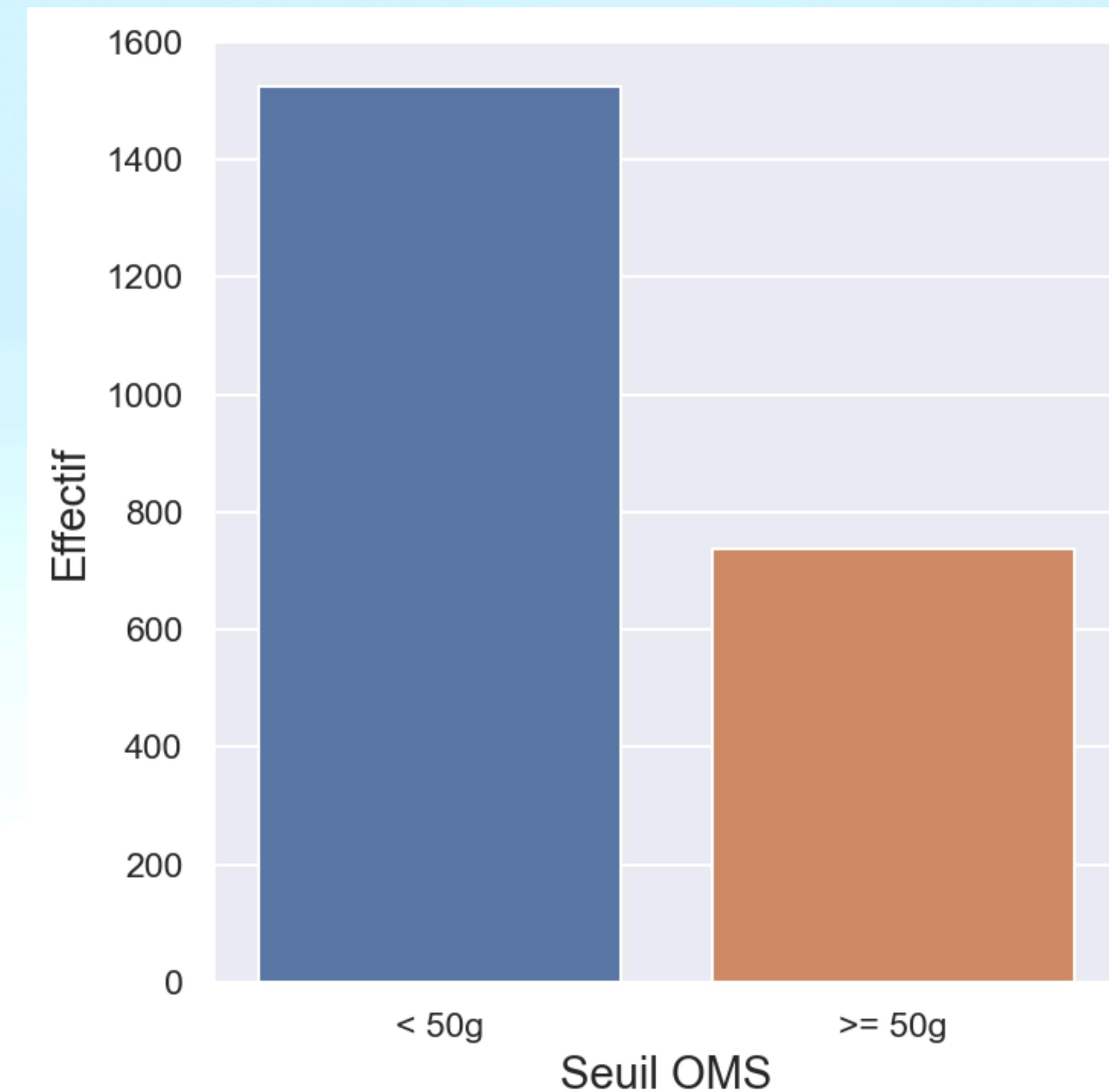
## 5. Exploration des données | Analyses univariées | Variables catégorielles



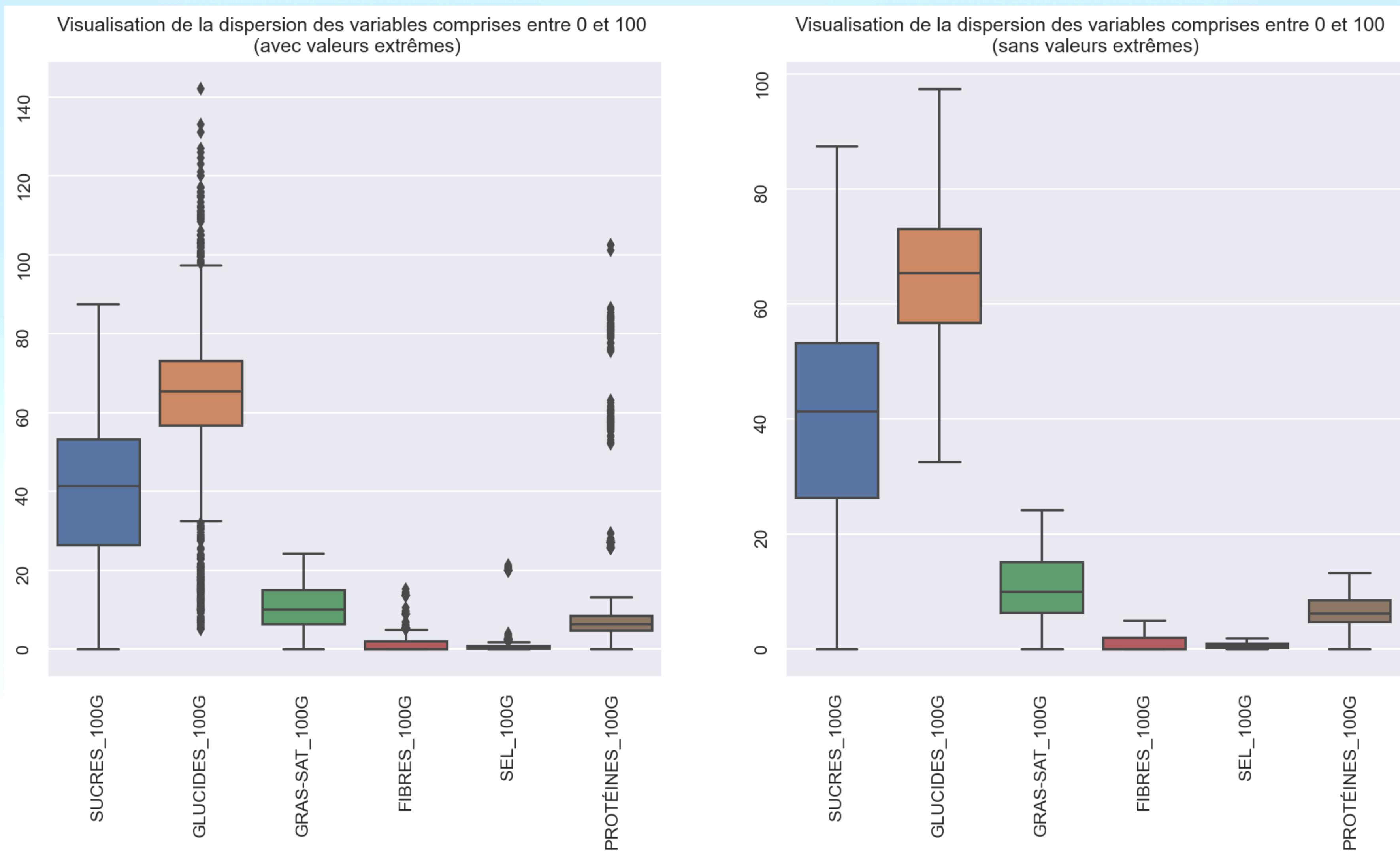
## 5. Exploration des données | Analyses univariées | Variables catégorielles



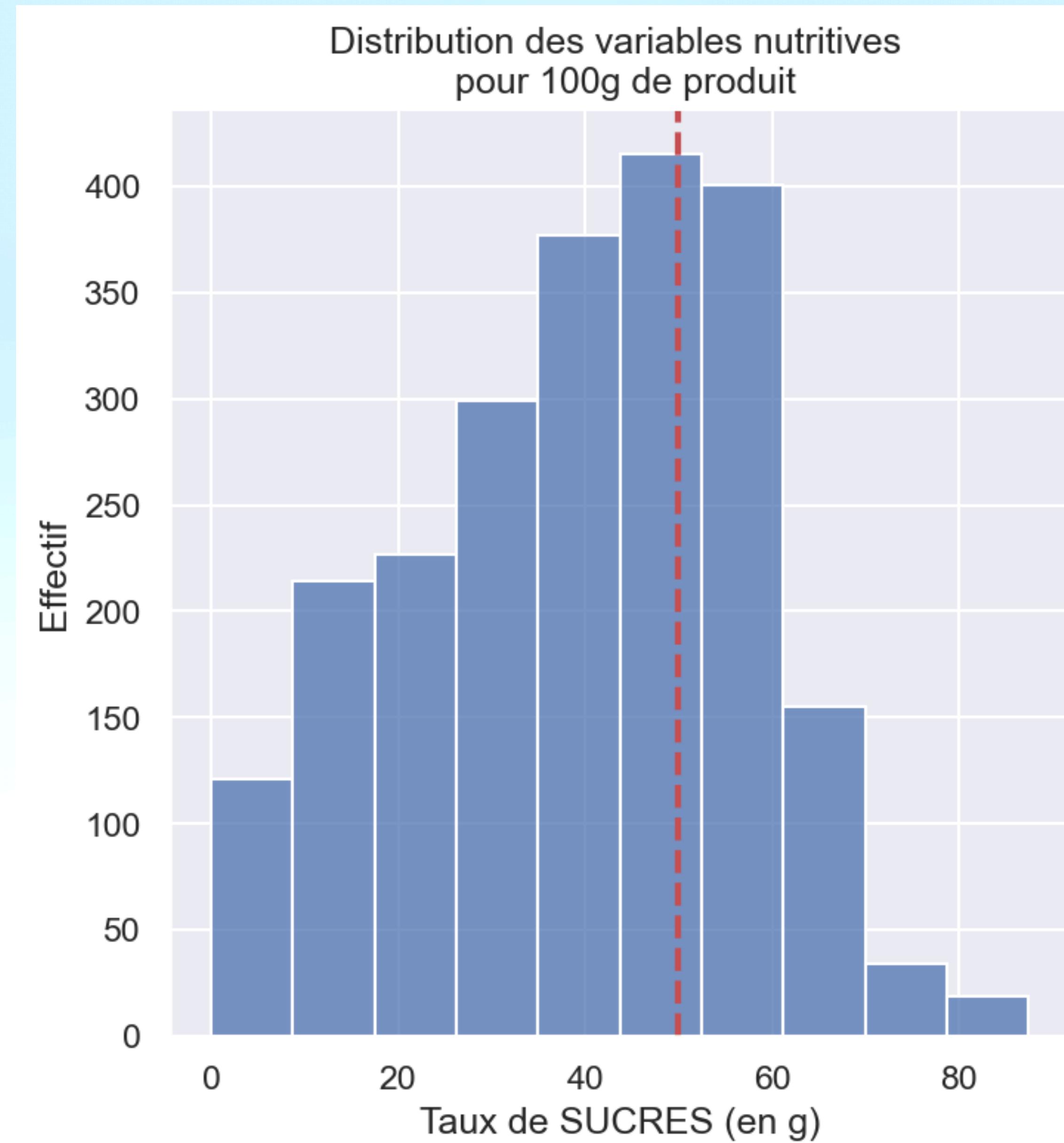
## 5. Exploration des données | Analyses univariées | Variables catégorielles



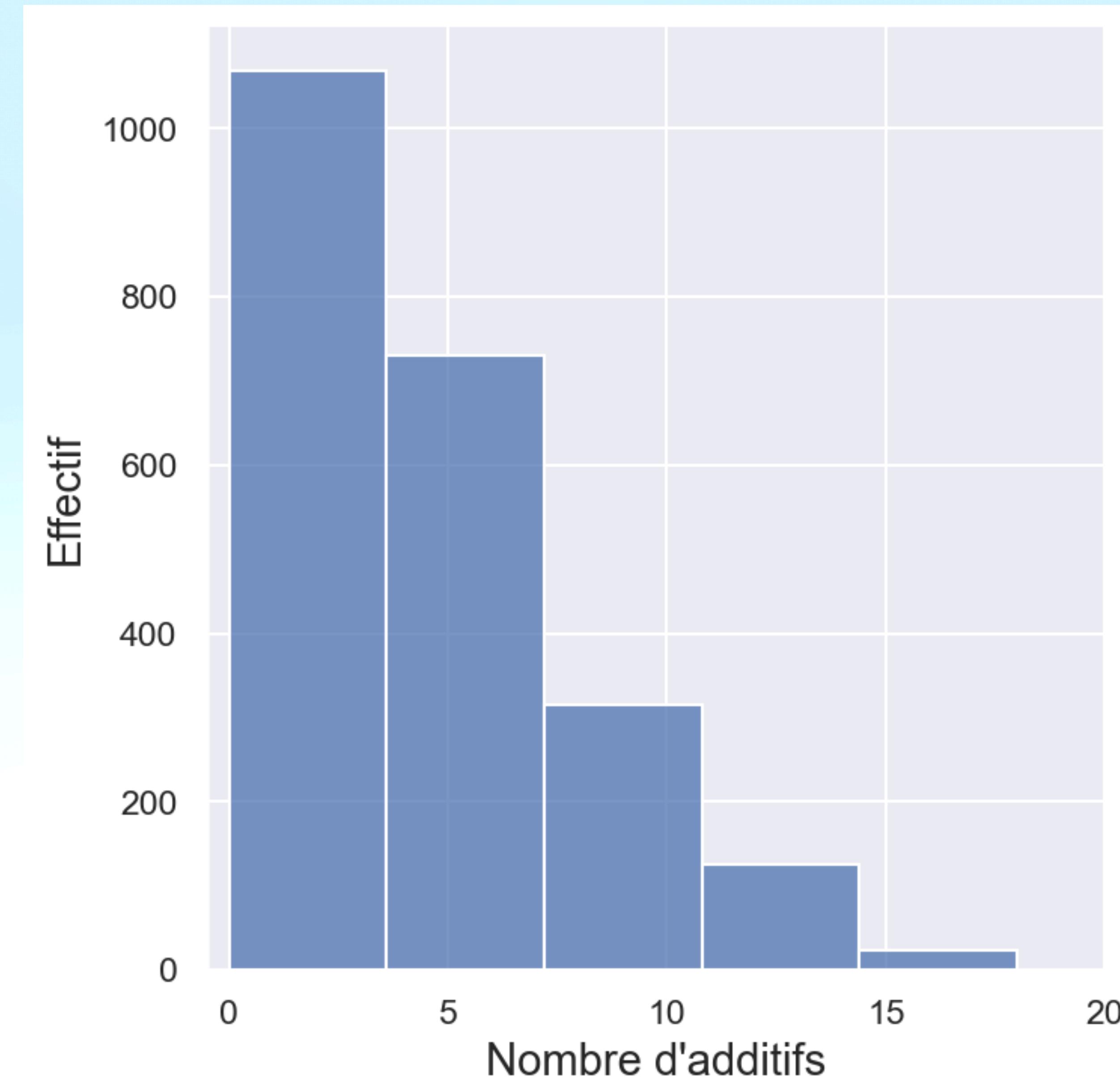
# 5. Exploration des données | Analyses univariées | Variables quantitatives



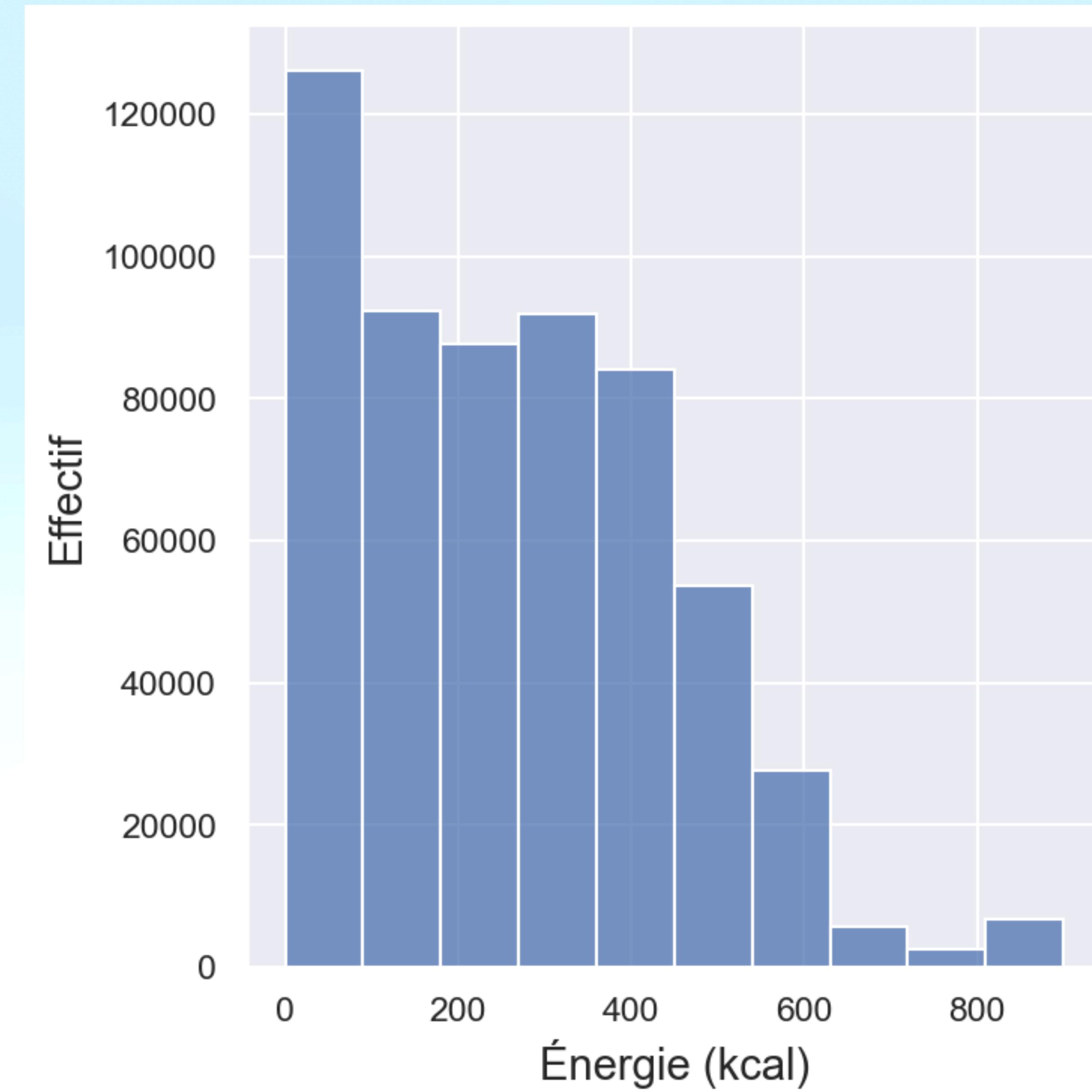
## 5. Exploration des données | Analyses univariées | Variables quantitatives



## 5. Exploration des données | Analyses univariées | Variables quantitatives



## 5. Exploration des données | Analyses univariées | Variables quantitatives

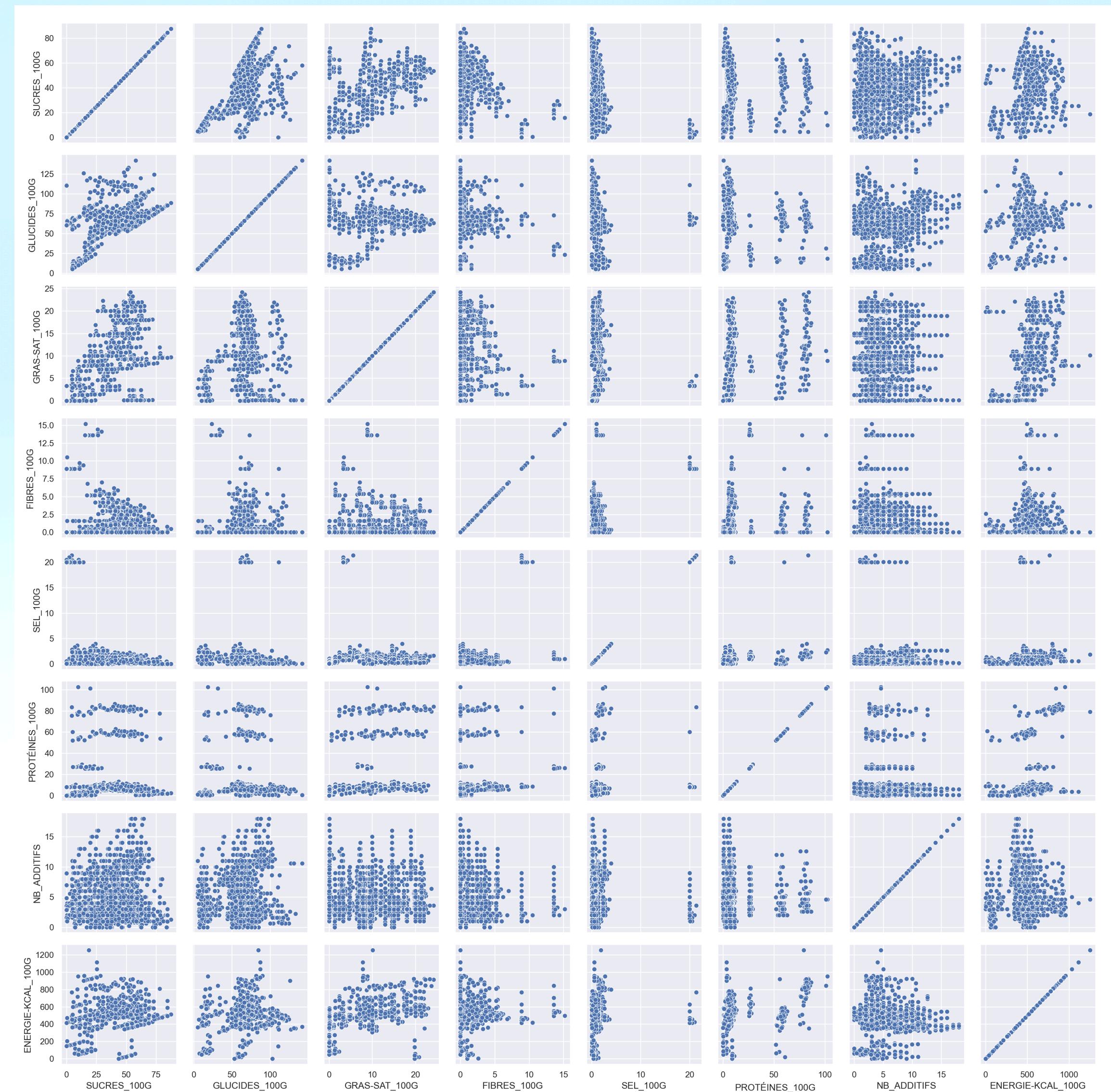


# 5. Exploration des données

## Analyses bivariées

### (a) Variables quantitatives vs variables quantitatives

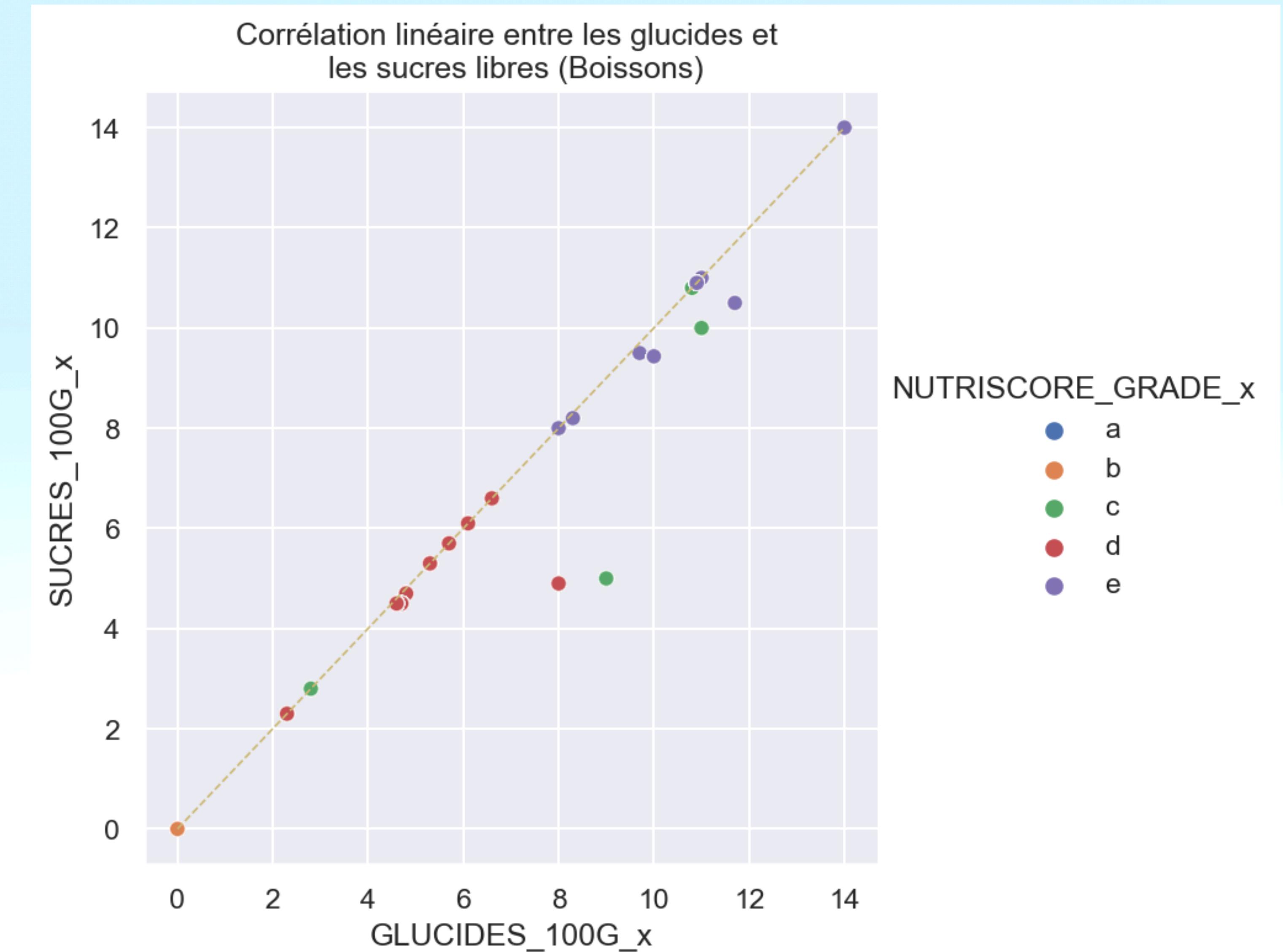
- Seule corrélation linéaire remarquable entre taux de sucres et taux de glucides pour les boissons



# 5. Exploration des données

## Analyses bivariées

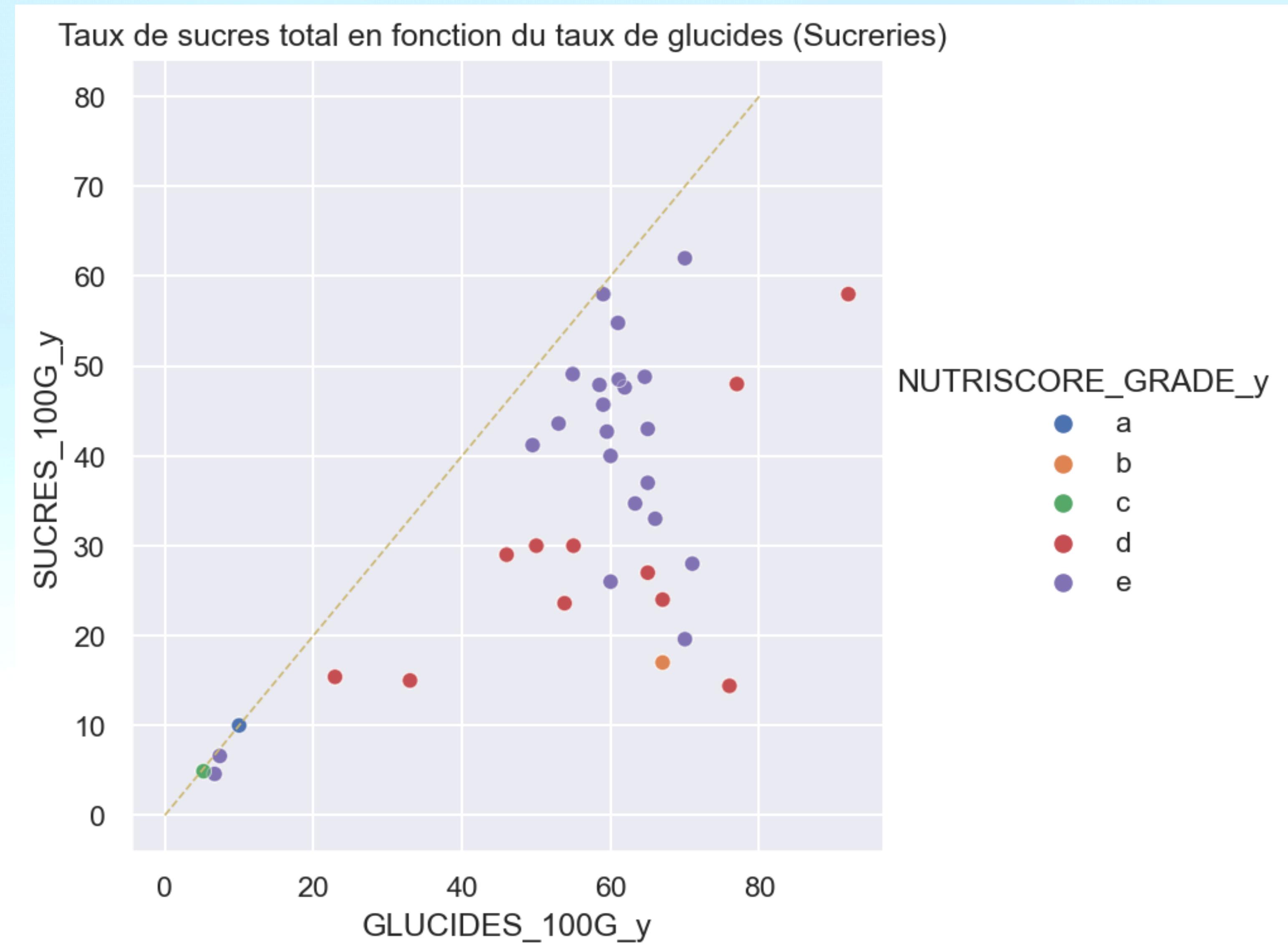
- (a) Variables quantitatives vs variables quantitatives
- Seule corrélation linéaire remarquable entre taux de sucres et taux de glucides pour les boissons



# 5. Exploration des données

## Analyses bivariées

- (a) Variables quantitatives vs variables quantitatives
- Seule corrélation linéaire remarquable entre taux de sucres et taux de glucides pour les boissons

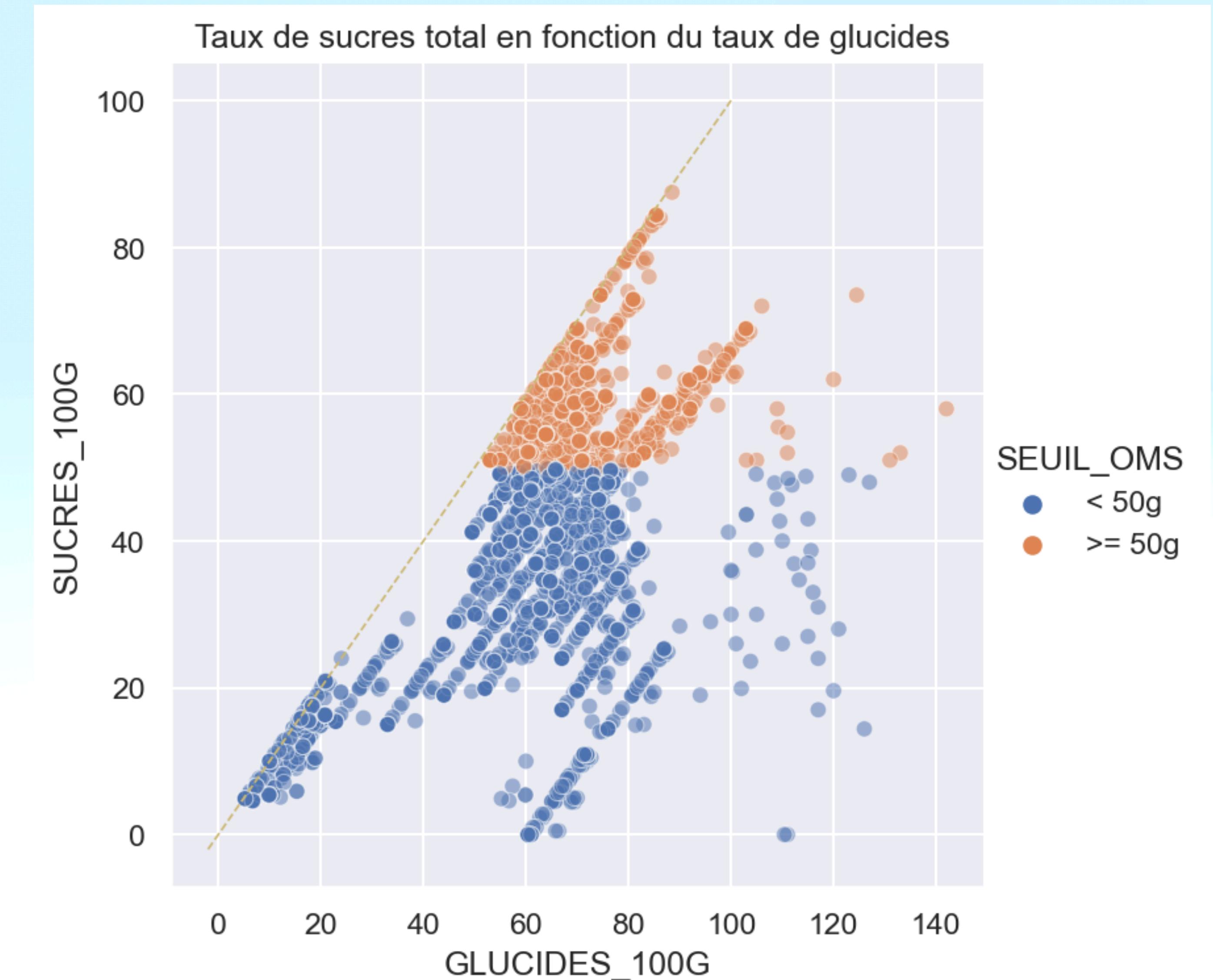


# 5. Exploration des données

## Analyses bivariées

(a) Variables quantitatives vs variables quantitatives

- Seule corrélation linéaire remarquable entre taux de sucres et taux de glucides pour les boissons

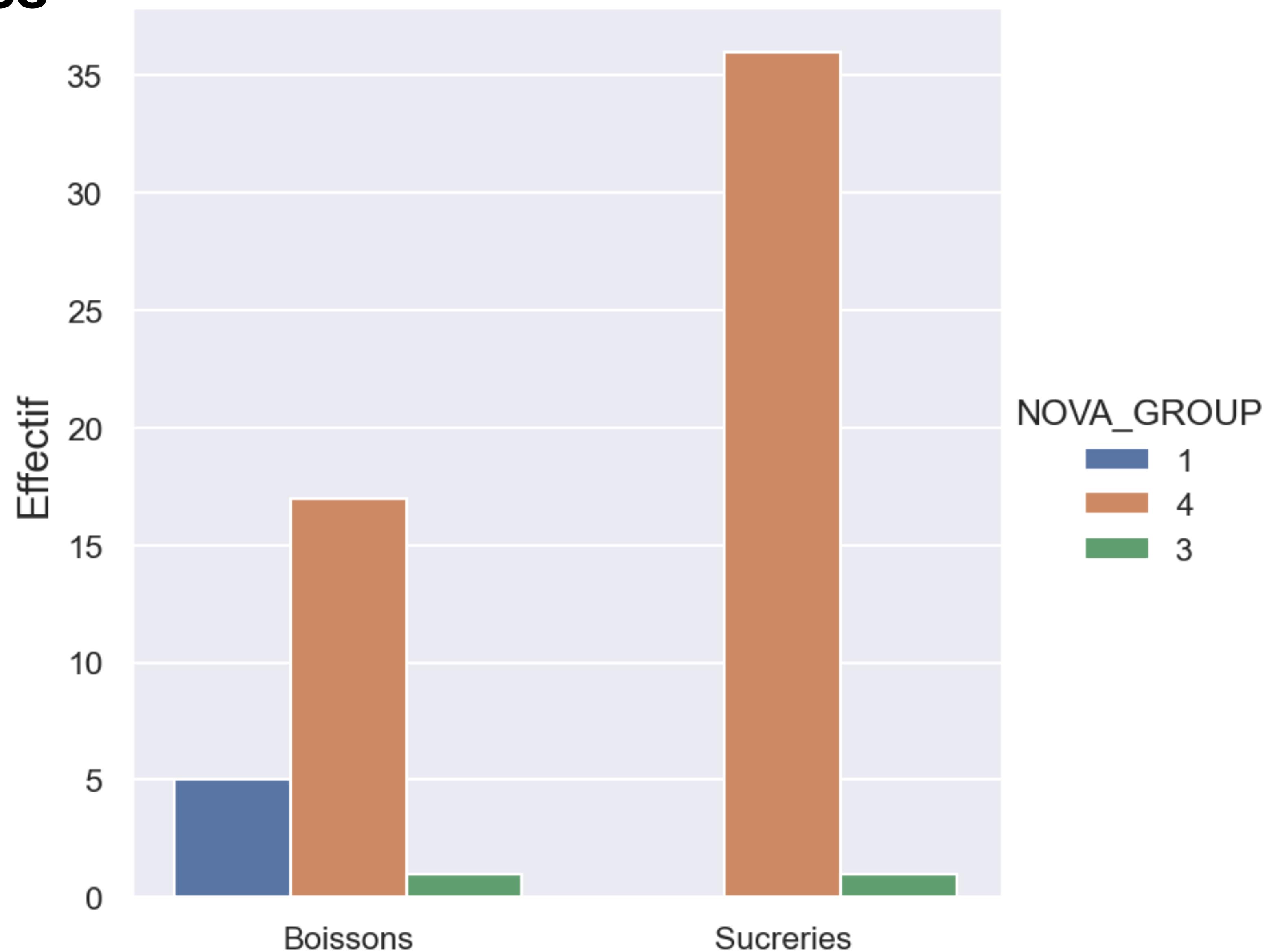


## 5. Exploration des données

### Analyses bivariées

(b) Variables qualitatives vs variables qualitatives

- Par rapport aux boissons, le groupe des sucreries contient le plus de produit transformés et de nutriscore E

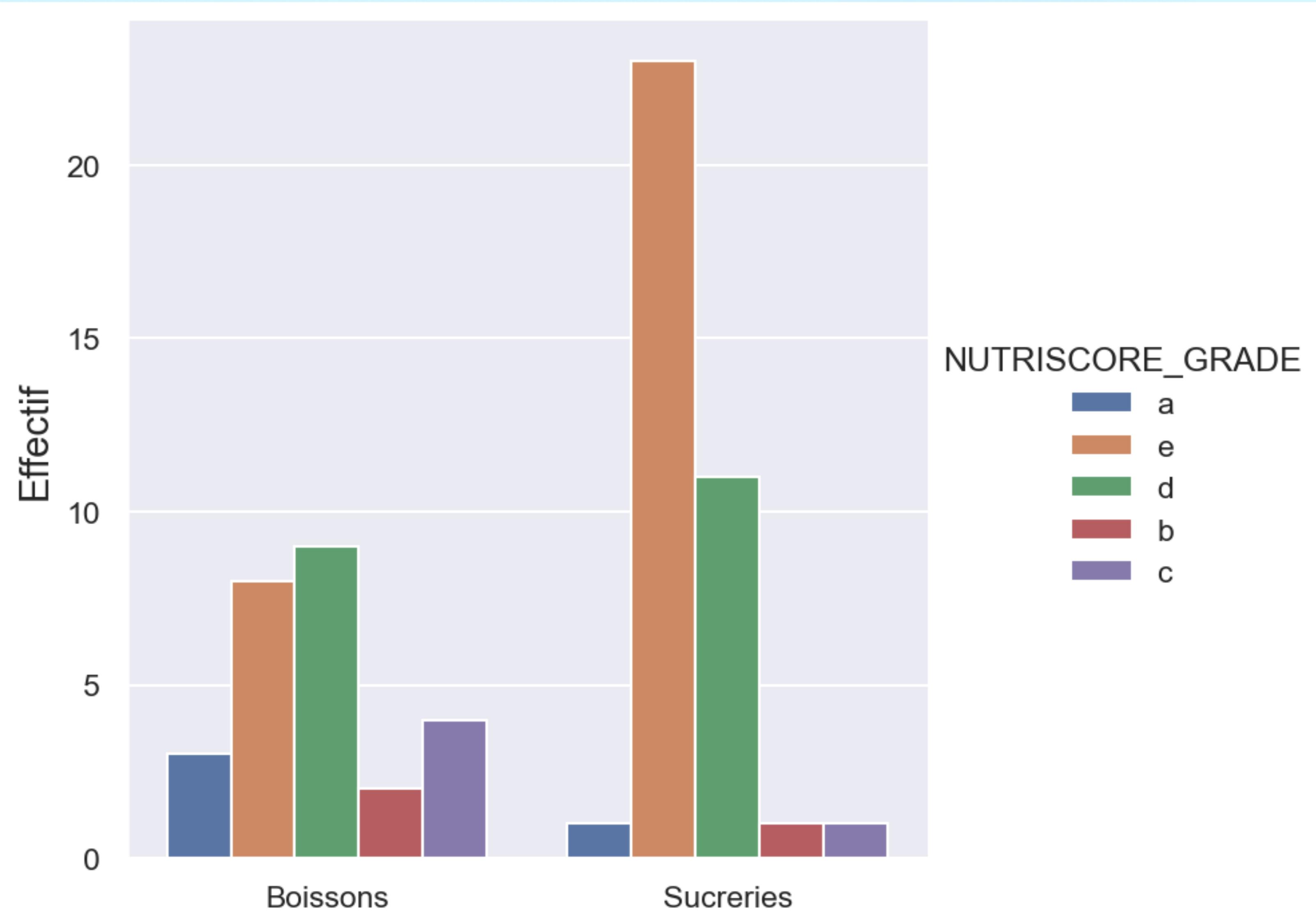


## 5. Exploration des données

### Analyses bivariées

(b) Variables qualitatives vs variables qualitatives

- Par rapport aux boissons, le groupe des sucreries contient le plus de produit transformés et de nutriscore E



## 5. Exploration des données

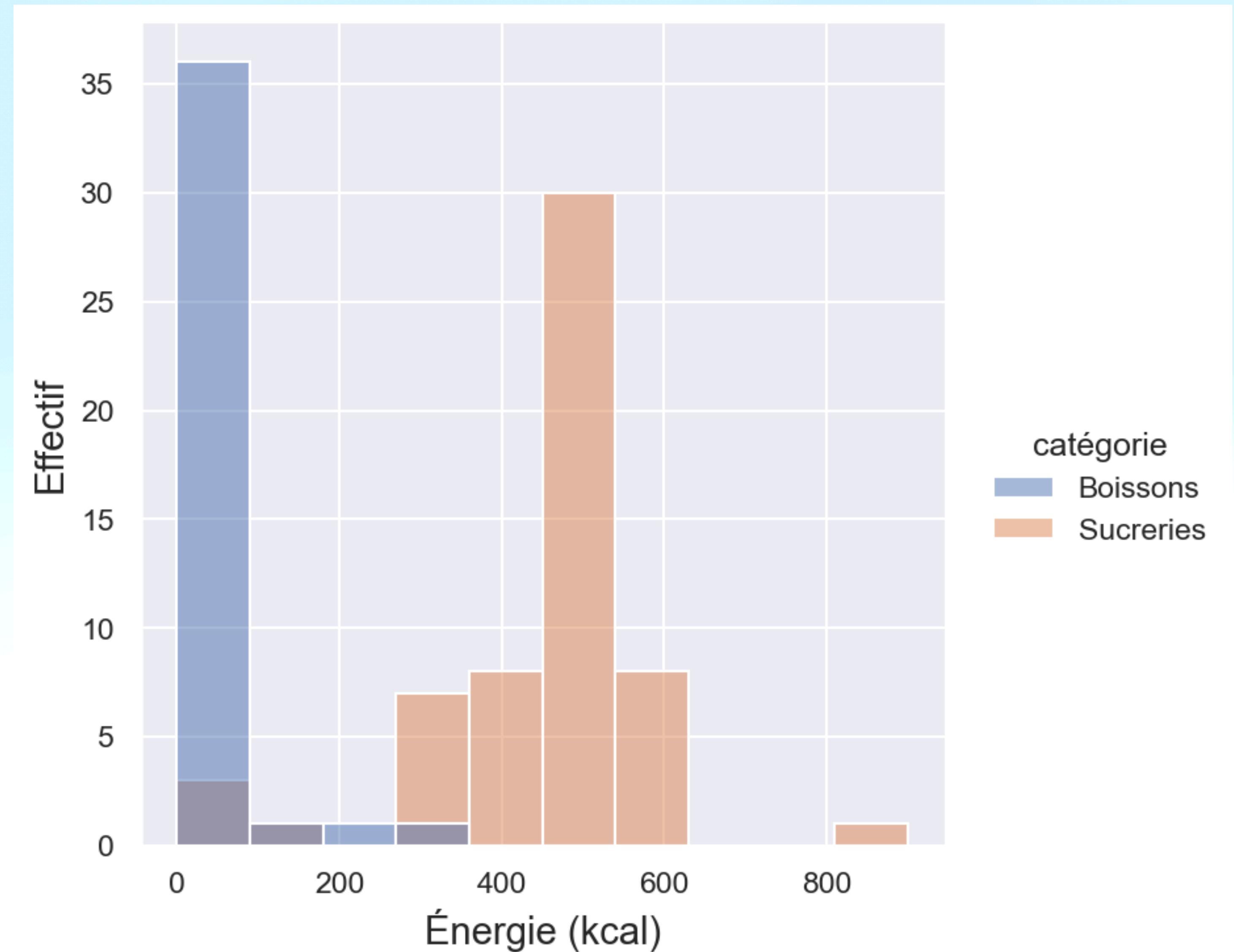
### Analyses bivariées

(c) Variables qualitatives vs variables quantitatives

- ANOVA :

$$\eta^2 \text{ (catégorie, énergie)} = 0.75$$

$$\eta^2 \text{ (catégorie, sucres)} = 0.52$$

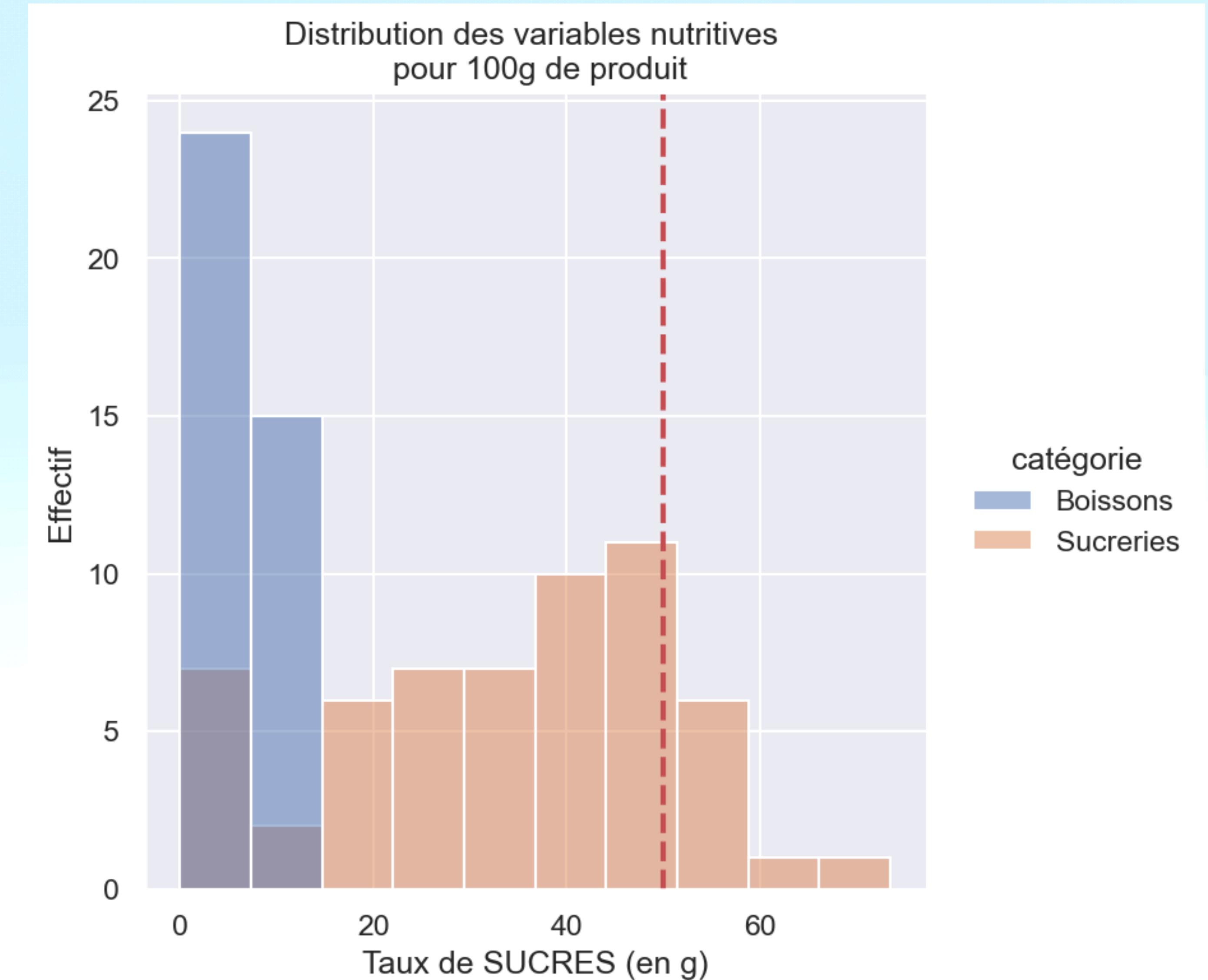


# 5. Exploration des données

## Analyses bivariées

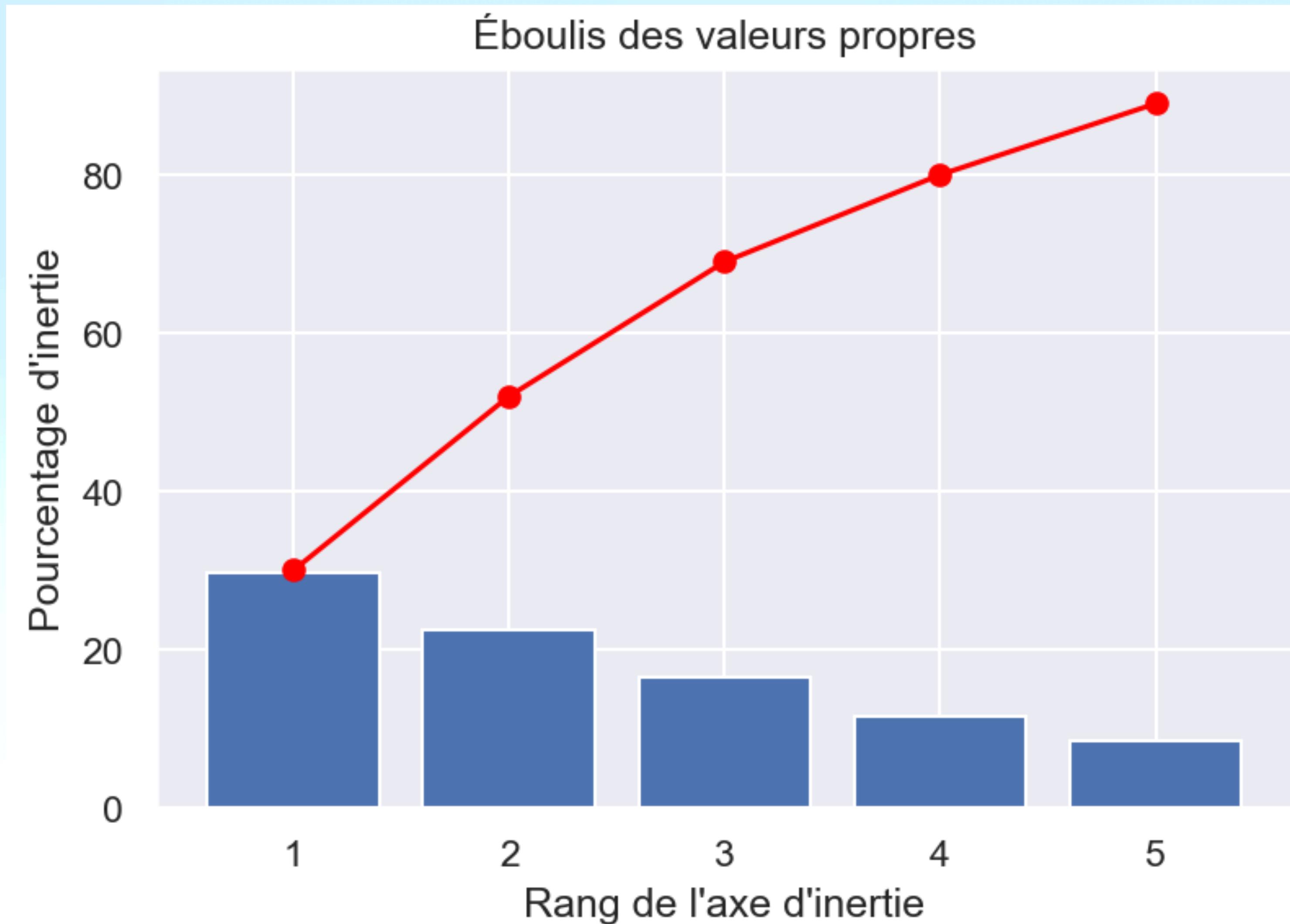
(c) Variables qualitatives vs variables quantitatives

- ANOVA :  
 $\eta^2$  (catégorie, énergie) = 0.75  
 $\eta^2$  (catégorie, sucres ) = 0.52



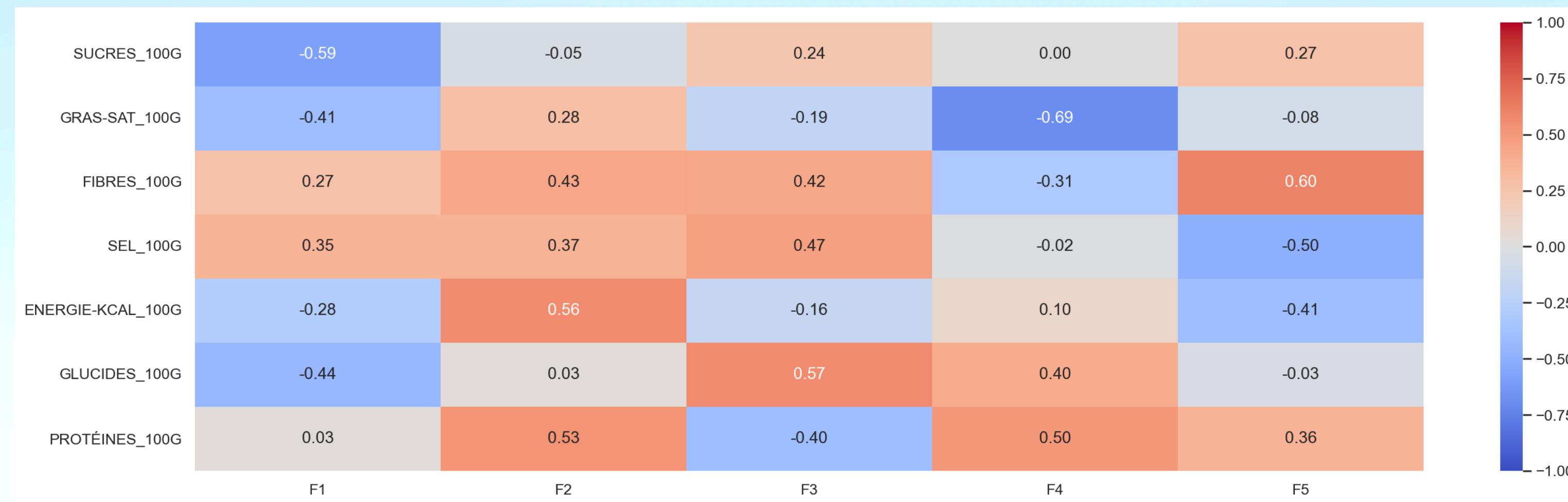
## 5. Exploration des données

### Analyses multivariées - ACP



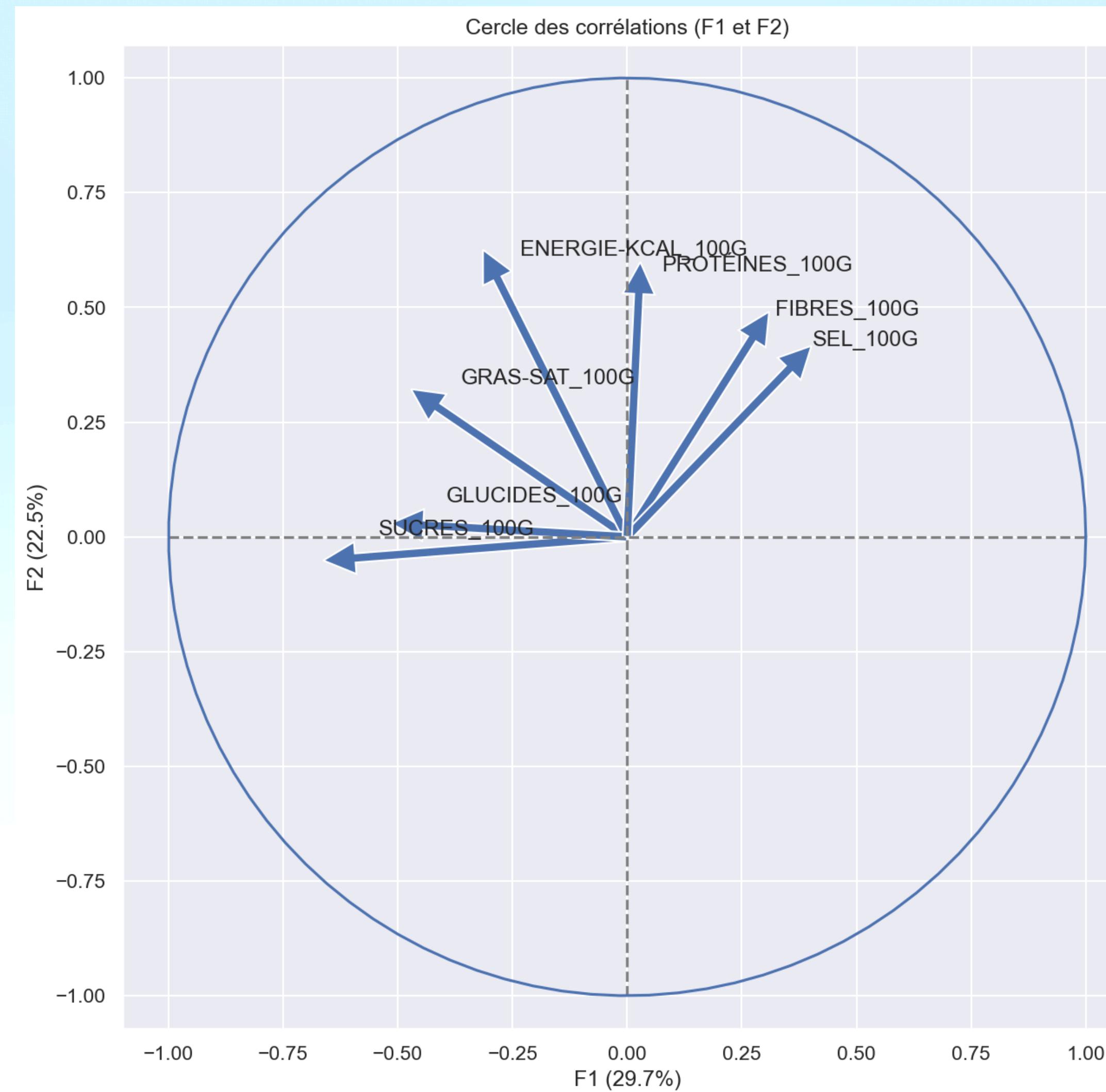
## 5. Exploration des données

### Analyses multivariées - ACP



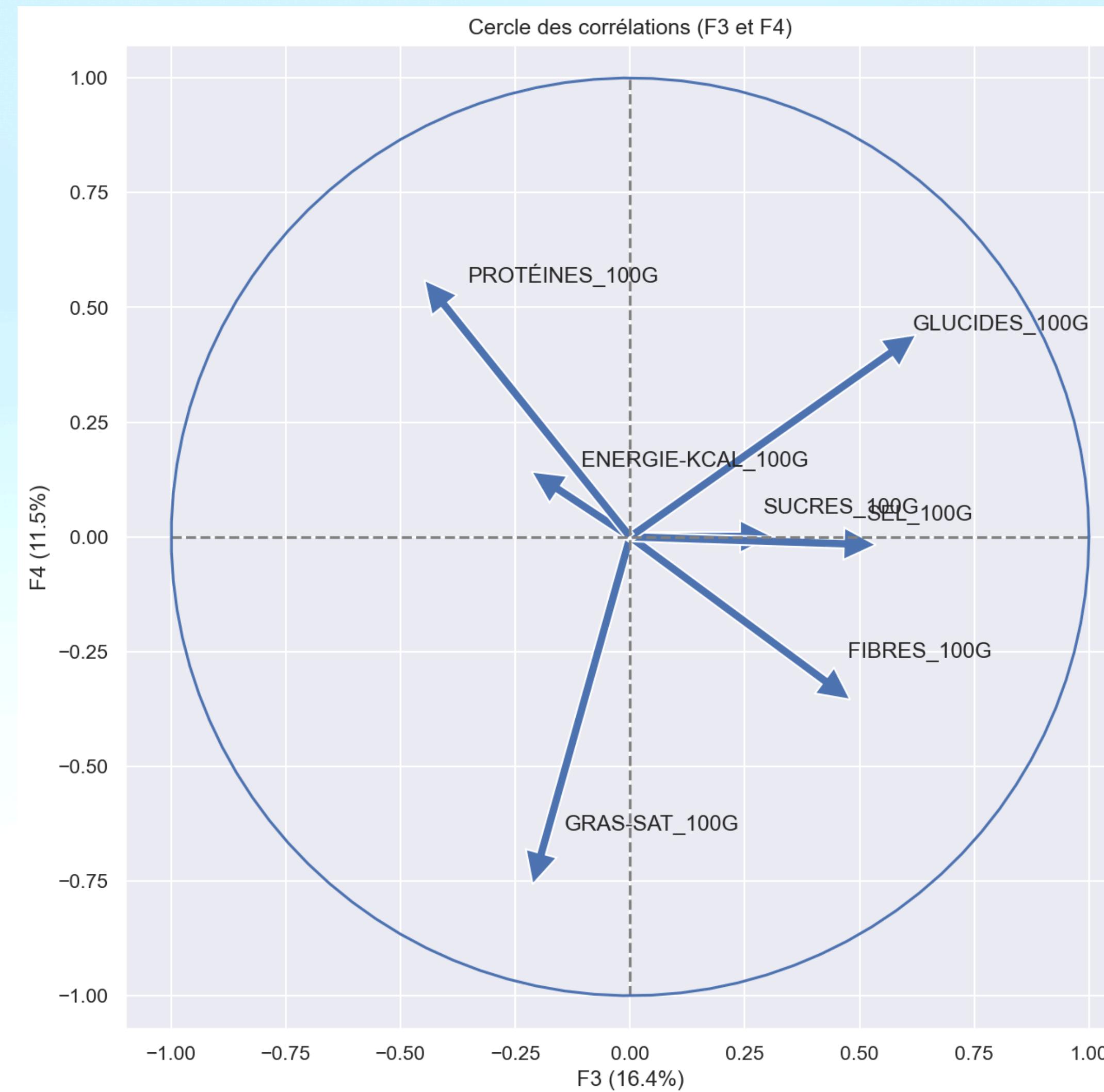
## 5. Exploration des données

### Analyses multivariées - ACP



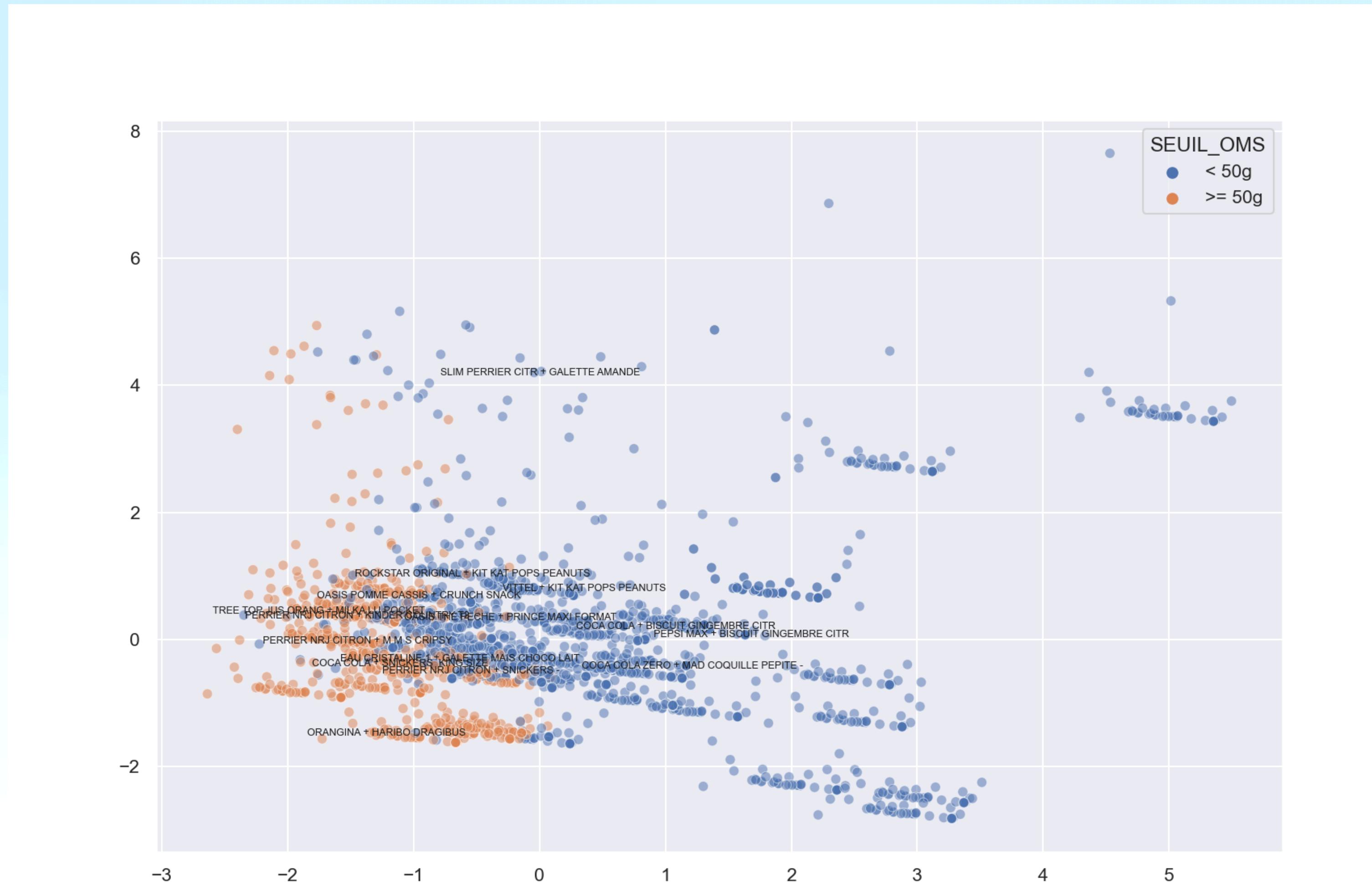
## 5. Exploration des données

### Analyses multivariées - ACP



## 5. Exploration des données

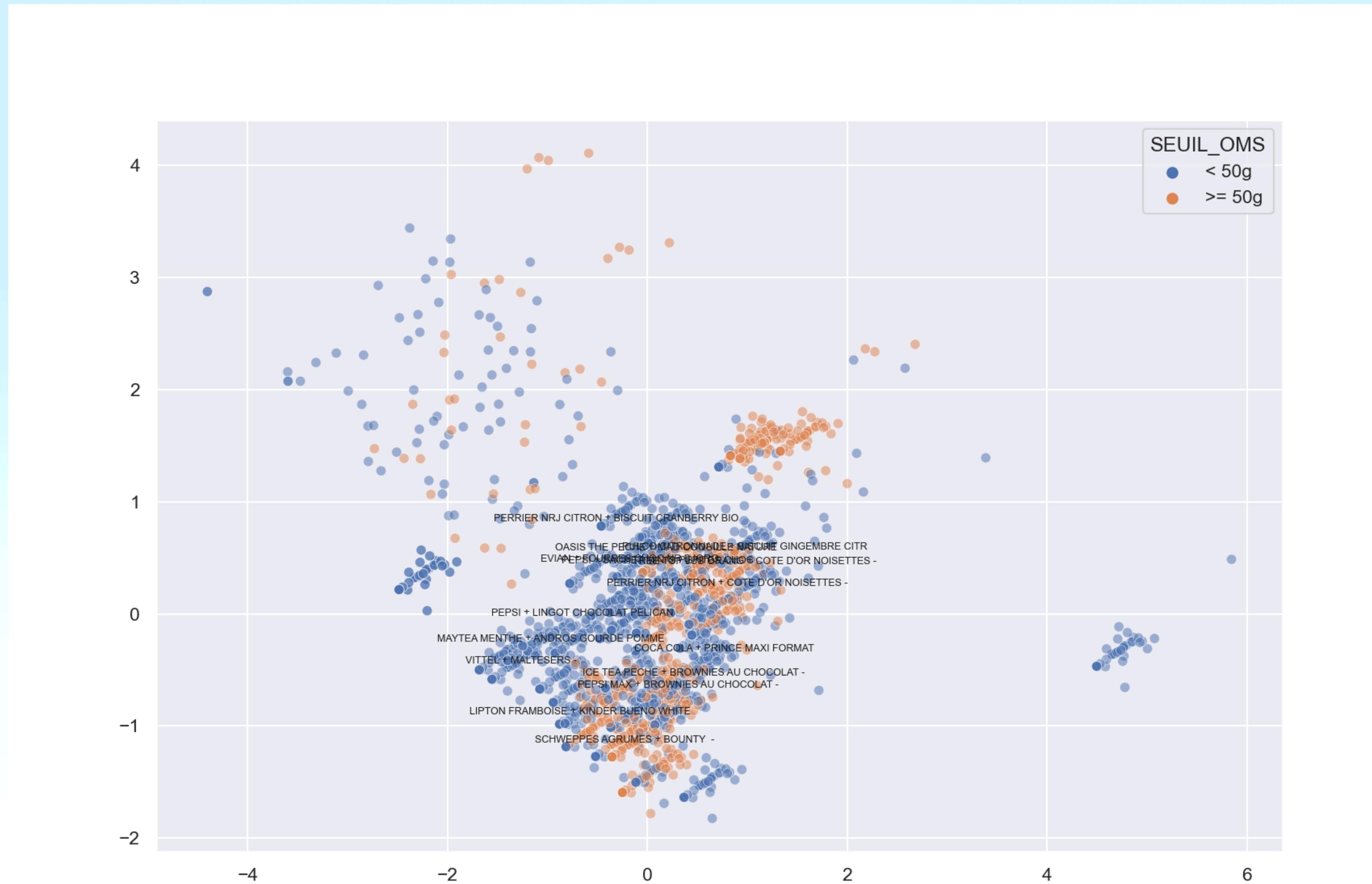
### Analyses multivariées - ACP



Snack dans le premier plan factoriel (-F1 : Apport en sucre / F2 : Apport en calories)

## 5. Exploration des données

### Analyses multivariées - ACP

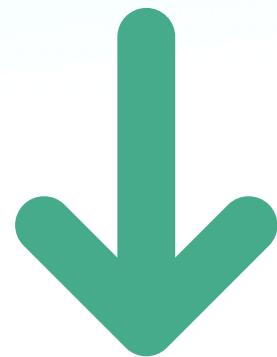


Snack dans le second plan factoriel (F3 : richesse en glucides / -F4 : richesse en graisse)

# Conclusion

Lors de l'achat d'un snack à la machine SELECTA

- \*33% de chance d'atteindre ou de dépasser la dose de sucre journalière
- \*77 % de chance d'atteindre ou de dépasser la moitié de la dose de sucre journalière



L'utilisation de ***snack\_control*** serait donc pertinente !