Une app au service de la santé publique France

PROJET 03/ Openclassrooms

Gulsum Kapanoglu

Dans ce Project..

- ✓Présentation de l'application
- ✓Nettoyage des données
- ✓Analyses univariées et bivariées
- ✓Analyses multivariées
- ✓Conclusion

A propos de application

« Manger Sans-Gluten »

C’est quoi Maladie cœliaque ?

L'application « *Manger Sans-Gluten* »

L'application « *Manger Sans-Gluten* »

- Idée d'application en lien avec l'alimentation
- Données: Open Food Facts

A propose de « *Manger Sans-Gluten* »

- Scan du produit
- Matching avec la base de données « OpenFoodFacts » pour récupérer les données nutritionnelles

Et indique score du produit s'il contient ou non du gluten.

Nettoyage des données

« Manger Sans-Gluten »

Description des données

Fichier volumineux (4 Gb)

- ✓ Comporte 162 variables:
- ✓ Informations générales sur des produits alimentaires
- ✓ Composition
- ✓ Informations nutritionnelles

```
print(factfood.shape)
factfood.head()
```

```
(320772, 162)
```

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name
0	3087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	NaN
1	4530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	NaN
2	4559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	NaN
3	16087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salted Nut Mix	NaN
4	16094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	NaN

5 rows × 162 columns

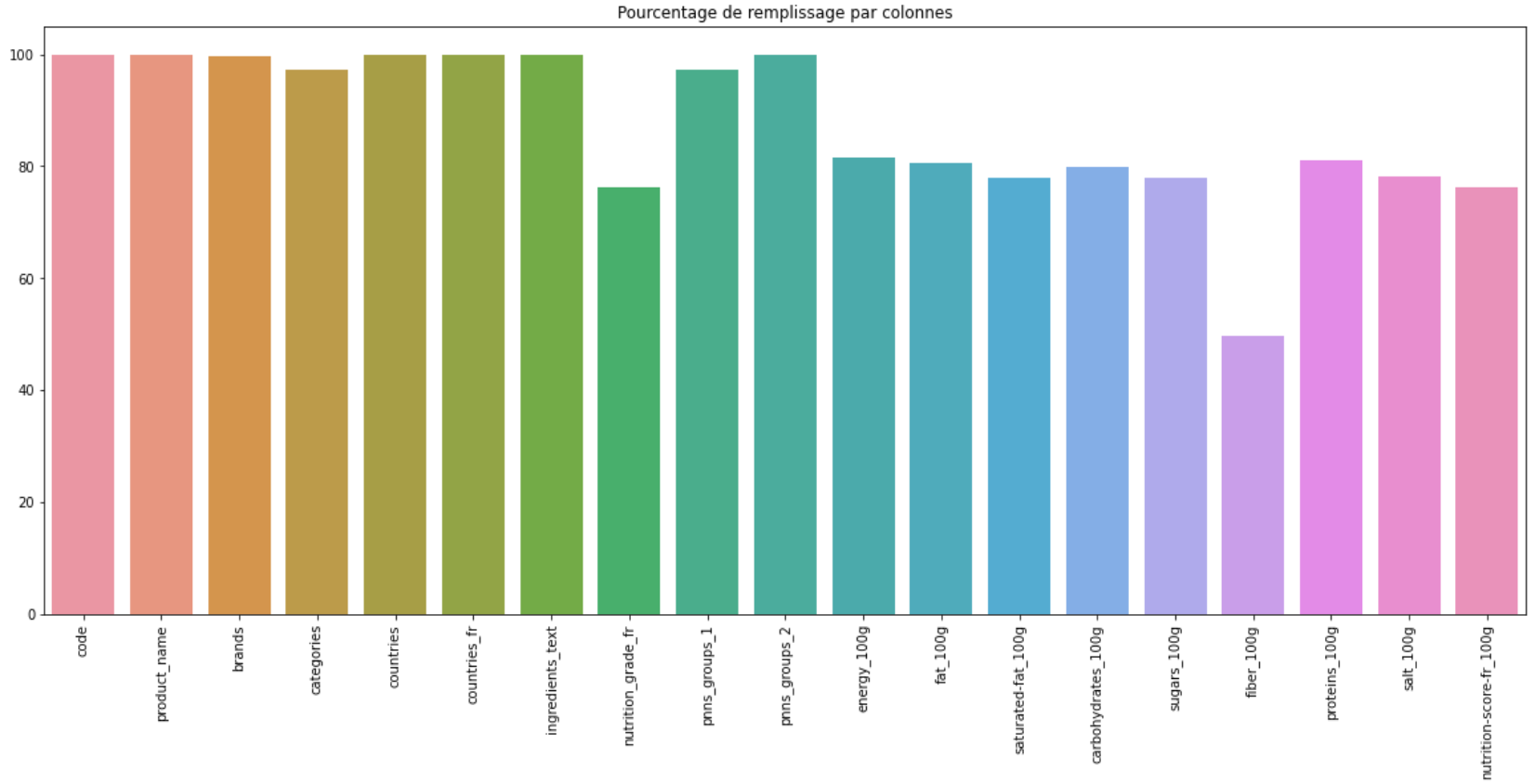
Methodologie de Nettoyage

1) Choix des Features
2) Traitement des valeurs aberrantes
3) Traitement des valeurs manquantes

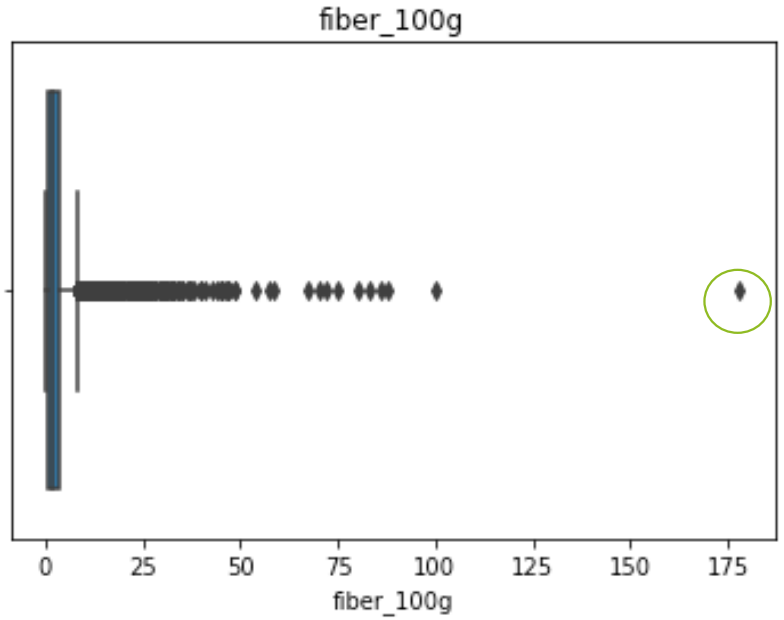
Choix des Features

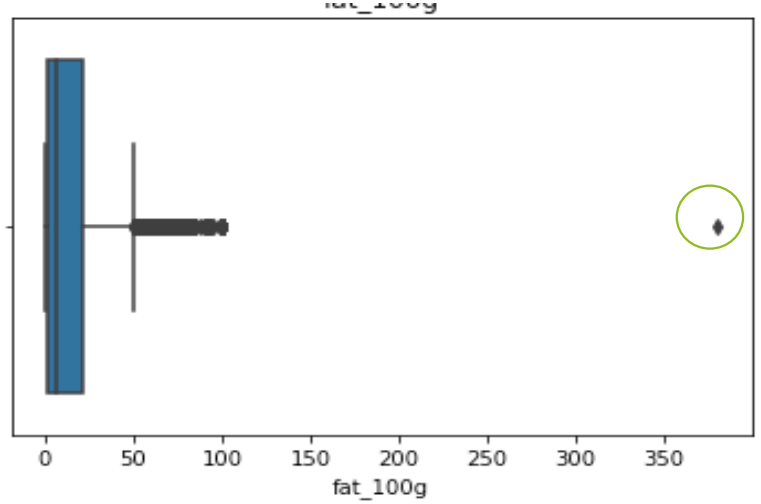
1. Sélection des colonnes intéressantes pour l'application
2. Garder uniquement les produis vendus en Fr et suppression des variables countries
3. Suppression des colonnes avec plus de 80% de valeurs manquantes
4. Suppression des lignes contenant **ingredients_text = Nan**

Taux de remplissage suite à la sélection



Valeurs Aberrantes

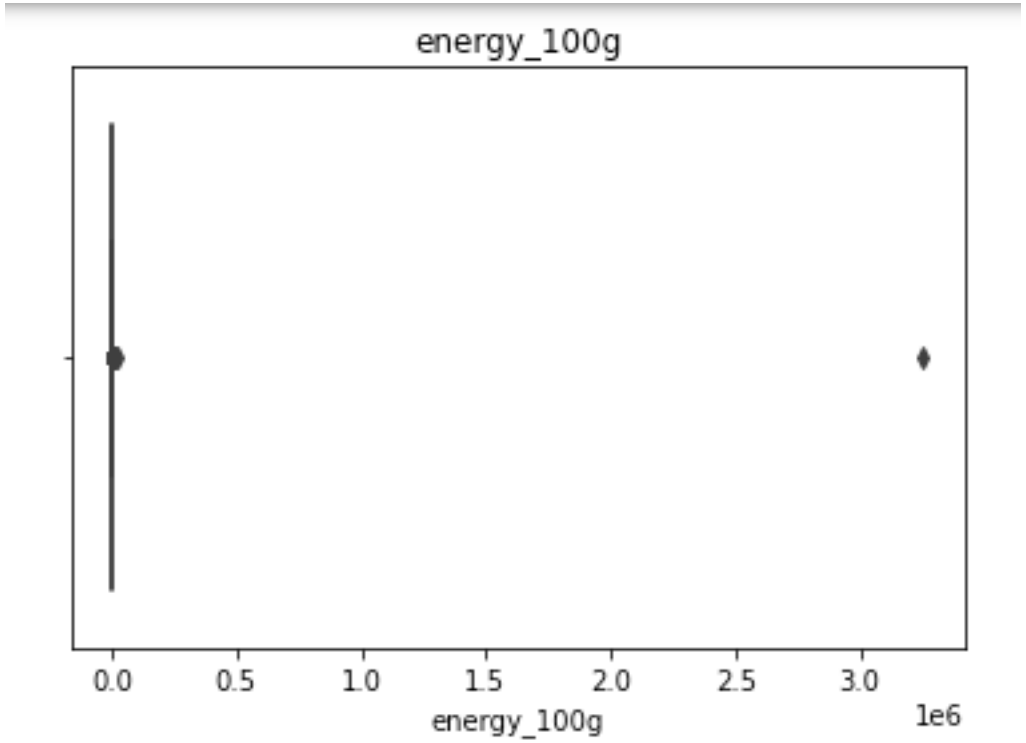




Suppression des produits qui contiennent des valeurs supérieures à 100g pour les carbohydrates, le sucre, les fibres, les protéines et le sel.

Il peut pas avoir plus 100g du nutriment du 100 g du produit

Valeurs Aberrantes



Suppression des produits qui contiennent des valeurs d'énergie supérieures à 900 calories*.

*valeur maximale pour (100g d'un produit) trouvé sur internet

Valeurs Manquantes

1. Suppression des produits sans nom et sans catégorie
2. Imputation des valeurs manquantes

CLASSE	BORNES DU SCORE	COULEUR
A	-15 à -1	Vert foncé
B	0 à 2	Vert clair
C	3 à 10	Orange clair
D	11 à 18	Orange moyen
E	19 à 40	Orange foncé

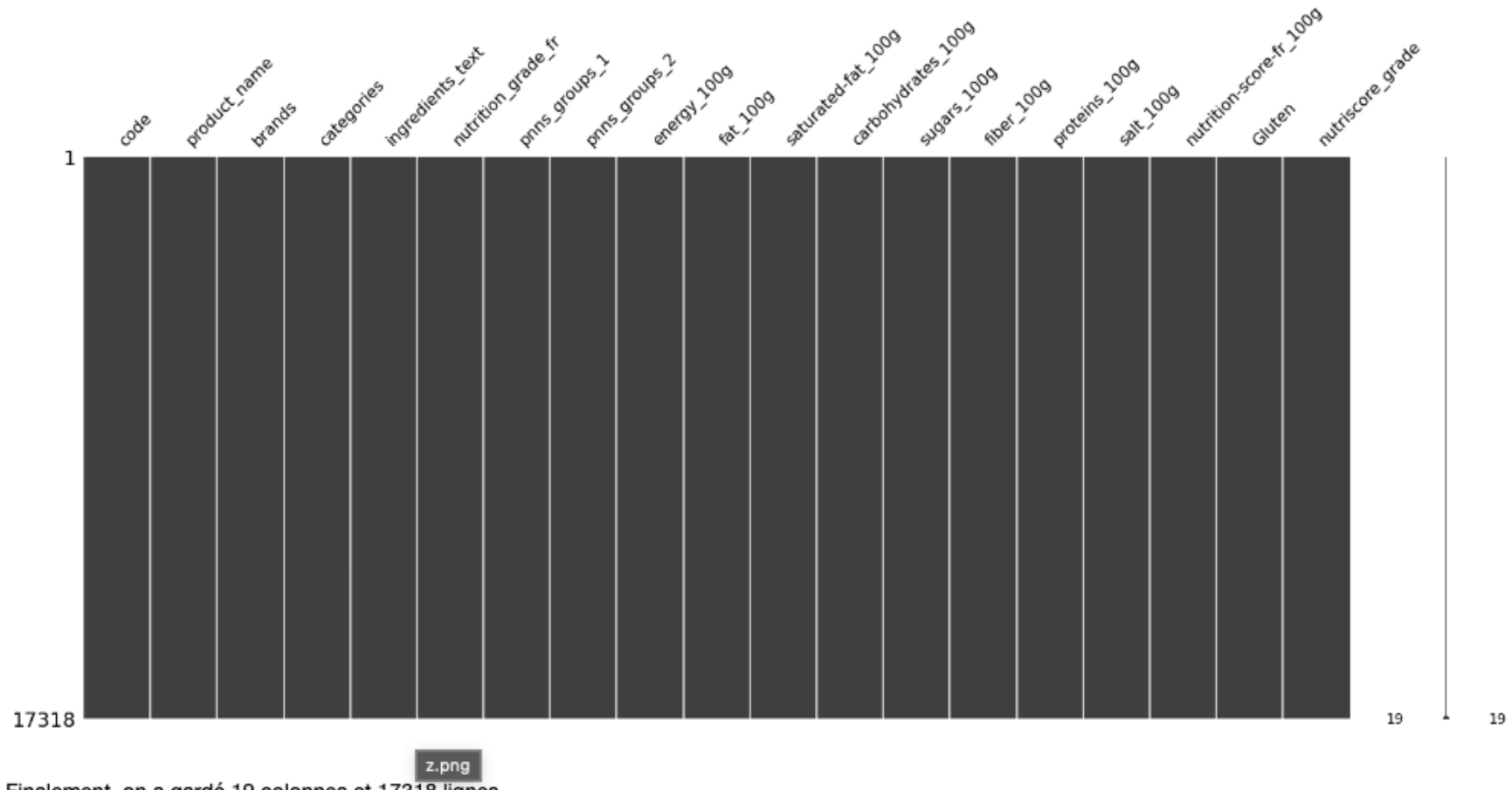
* Nutriscore grade: Correspondance avec nutriscore

3. Les autres variables: Imputation par KNN

DF Après Nettoyage

```
%matplotlib inline
msno.matrix(dff)
```

```
<AxesSubplot:>
```



z.png

Finalement, on a gardé 19 colonnes et 17318 lignes

Finalement, on a gardé 19 colonnes et 17318 lignes

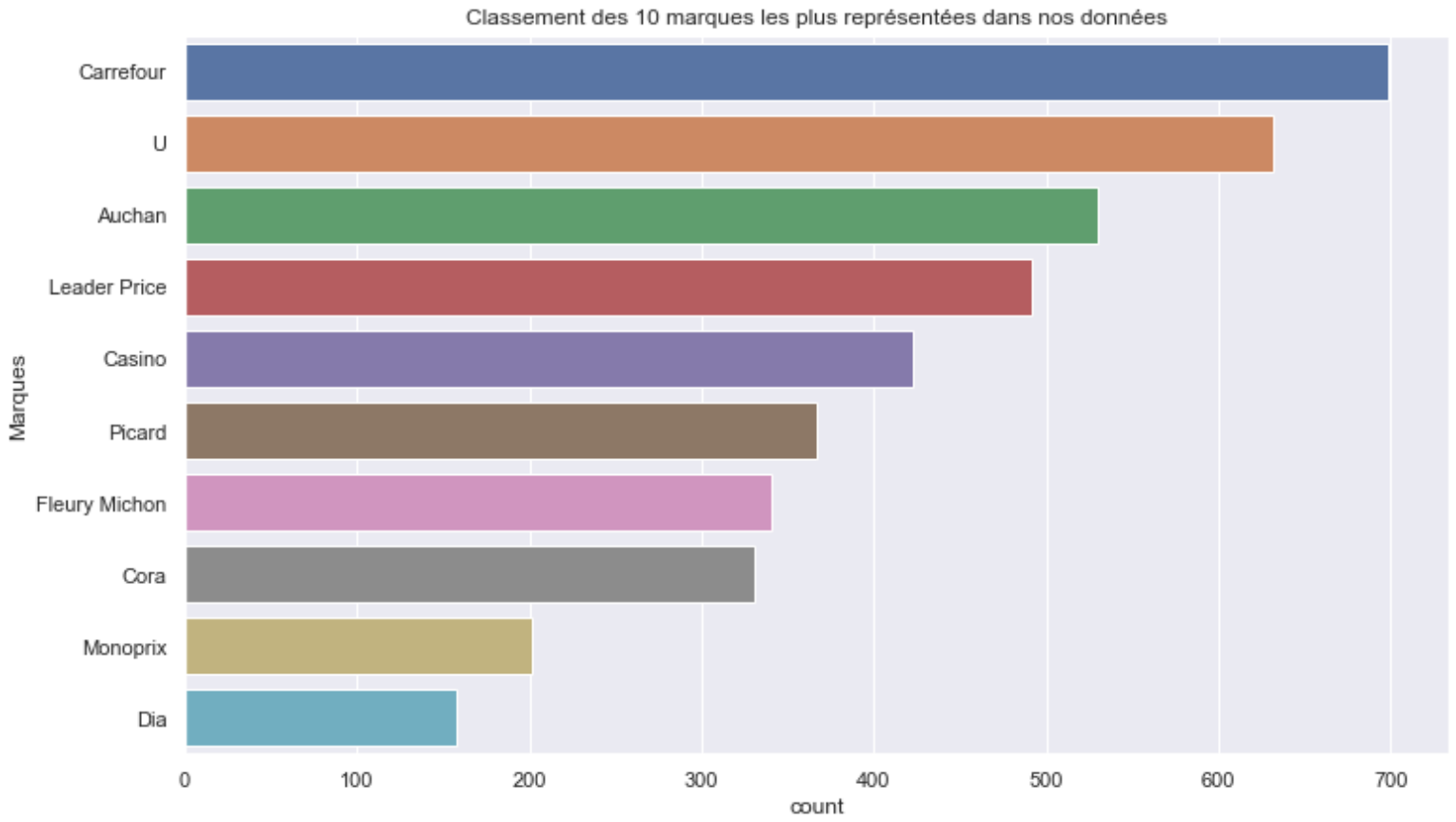
```
dff.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17318 entries, 0 to 17317
Data columns (total 19 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   code                     17318 non-null  object 
 1   product_name             17318 non-null  object 
 2   brands                   17318 non-null  object 
 3   categories               17318 non-null  object 
 4   ingredients_text         17318 non-null  object 
 5   nutrition_grade_fr       17318 non-null  object 
 6   pnns_groups_1            17318 non-null  object 
 7   pnns_groups_2            17318 non-null  object 
 8   energy_100g              17318 non-null  float64
 9   fat_100g                 17318 non-null  float64
 10  saturated-fat_100g       17318 non-null  float64
 11  carbohydrates_100g       17318 non-null  float64
 12  sugars_100g              17318 non-null  float64
 13  fiber_100g               17318 non-null  float64
 14  proteins_100g            17318 non-null  float64
 15  salt_100g                17318 non-null  float64
 16  nutrition-score-fr_100g  17318 non-null  float64
 17  Gluten                   17318 non-null  int64  
 18  nutriscore_grade         17318 non-null  object 
dtypes: float64(9), int64(1), object(9)
memory usage: 2.5+ MB
```

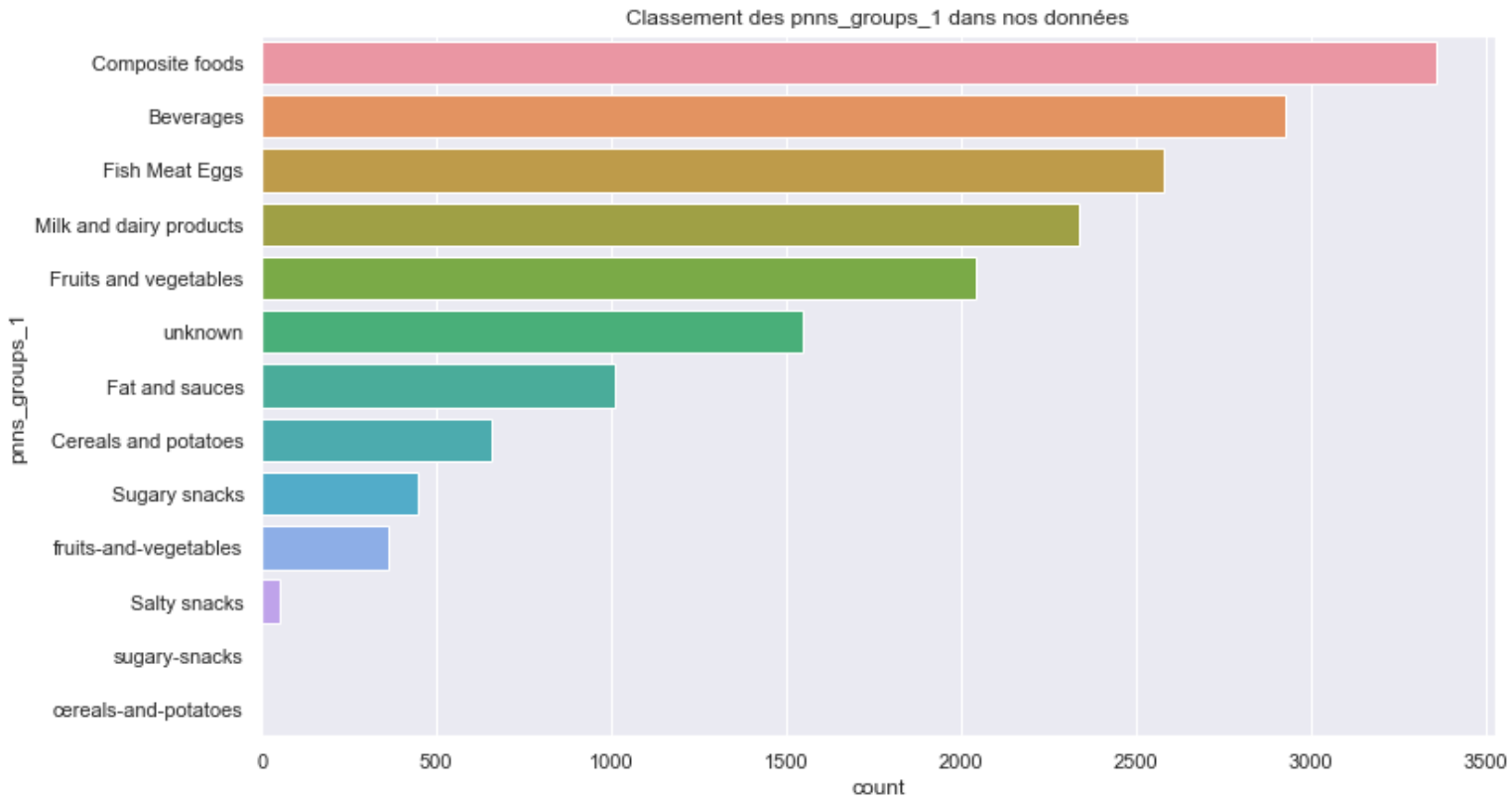
Analyses uni et bivariées

« *Manger Sans-Gluten* »

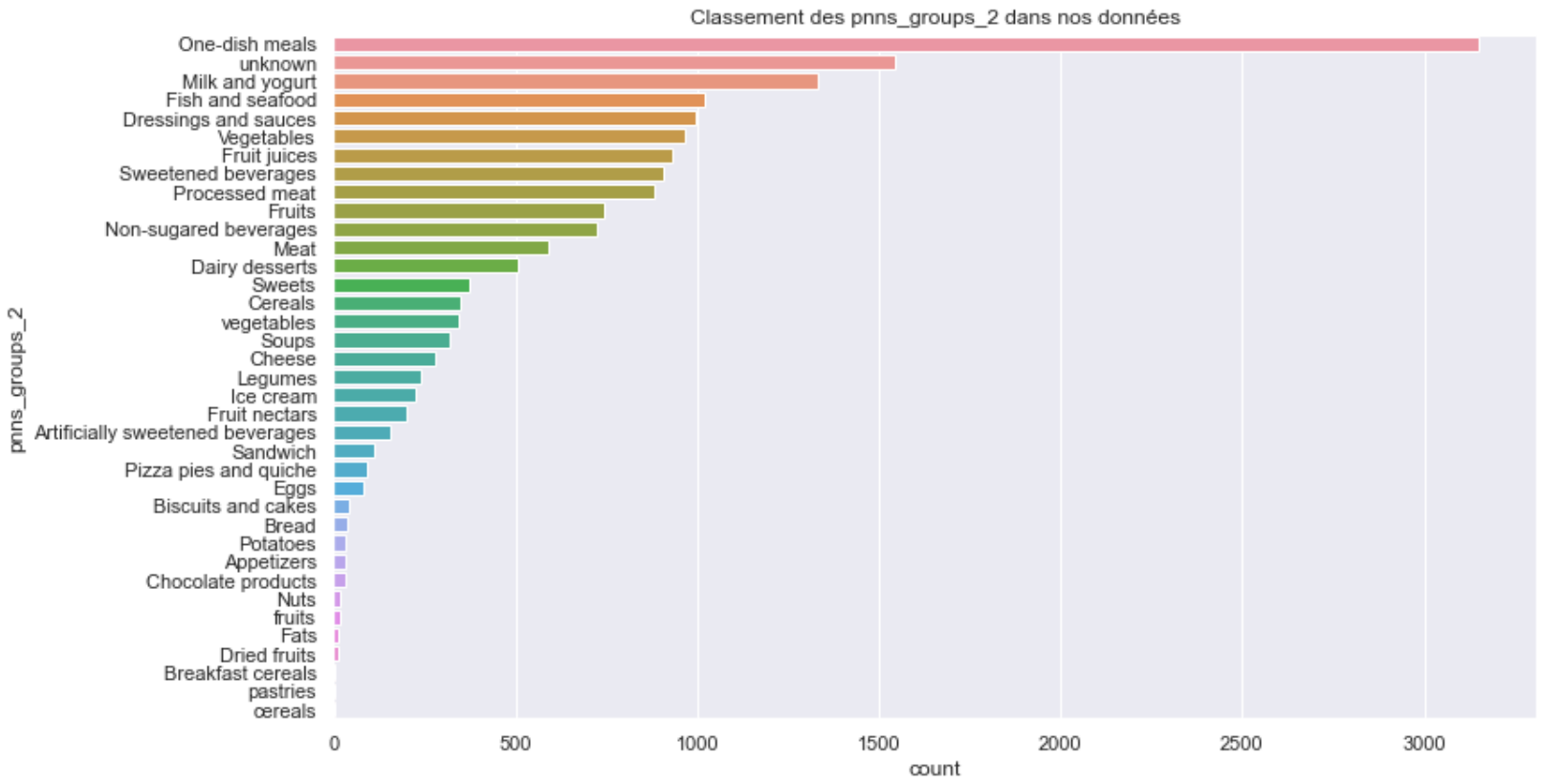
Classement des marques



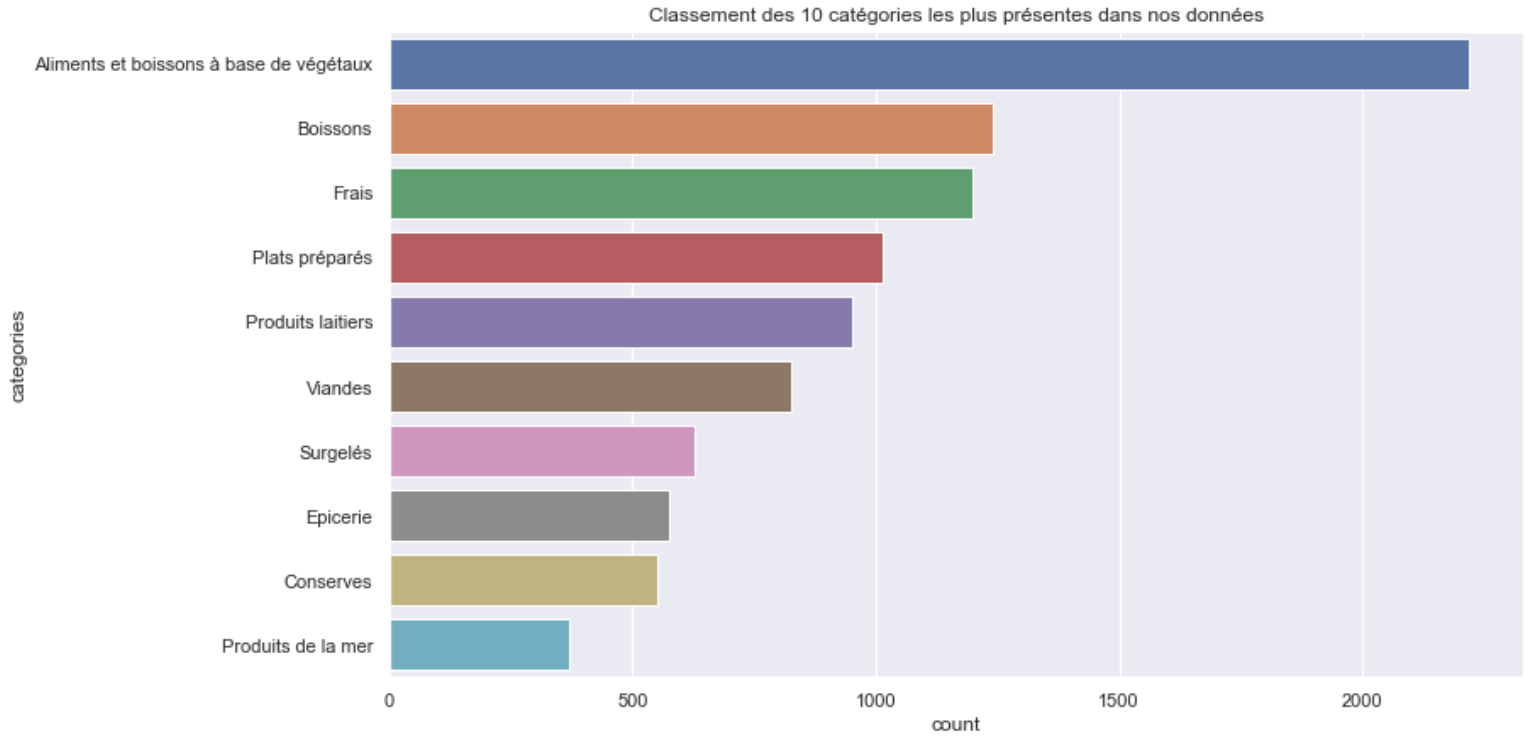
Classement des catégories



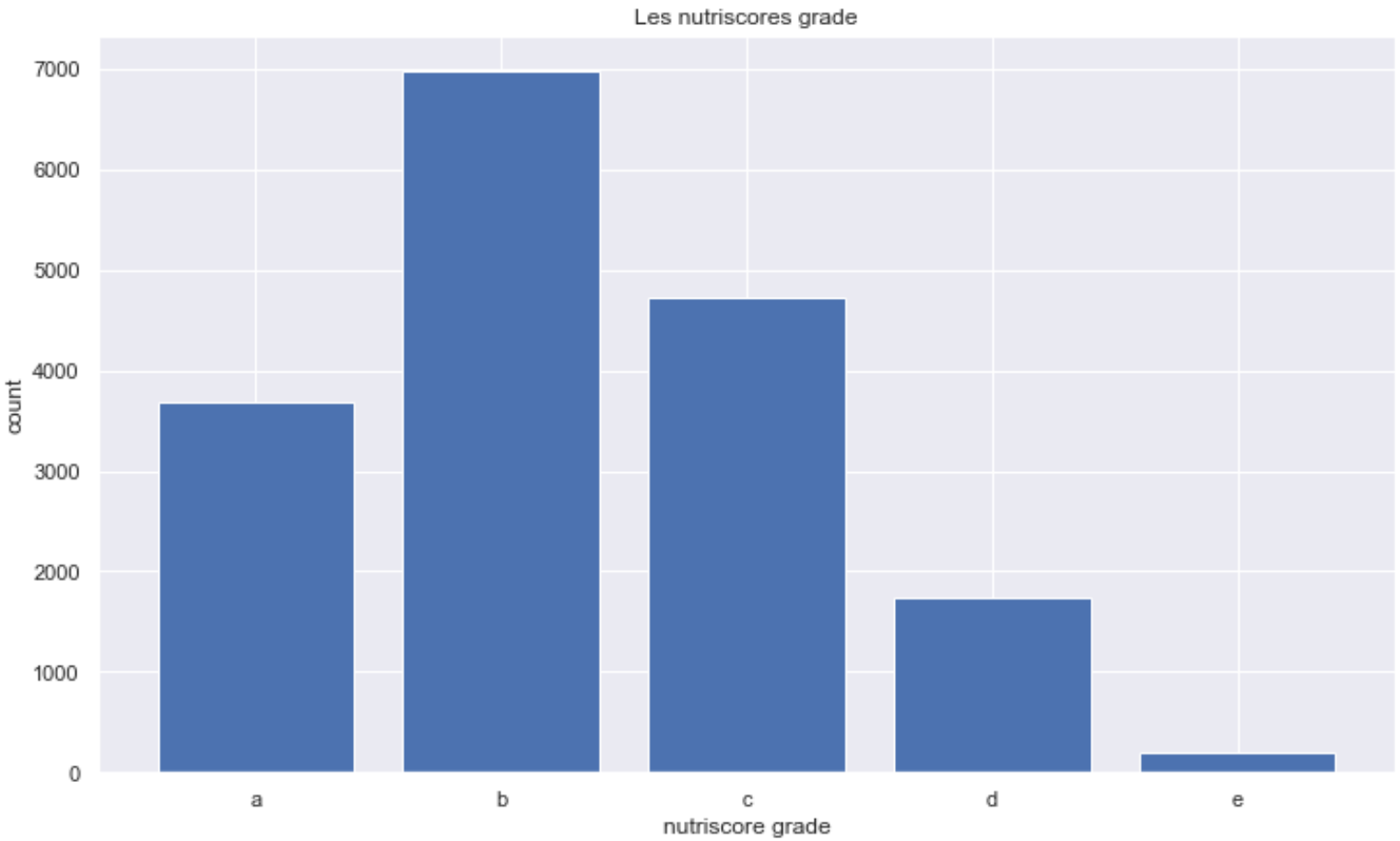
Classement des catégories



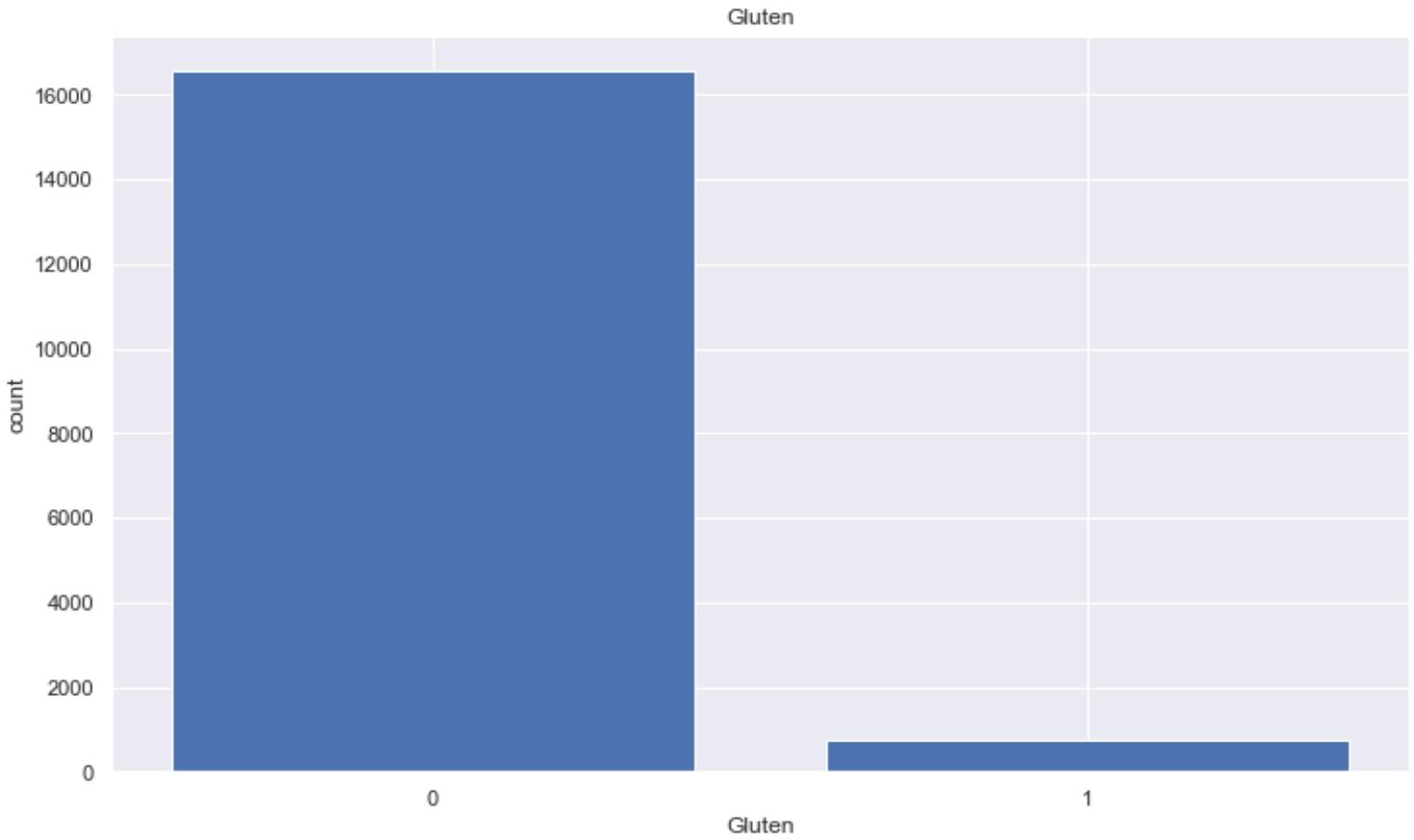
Classement des catégories



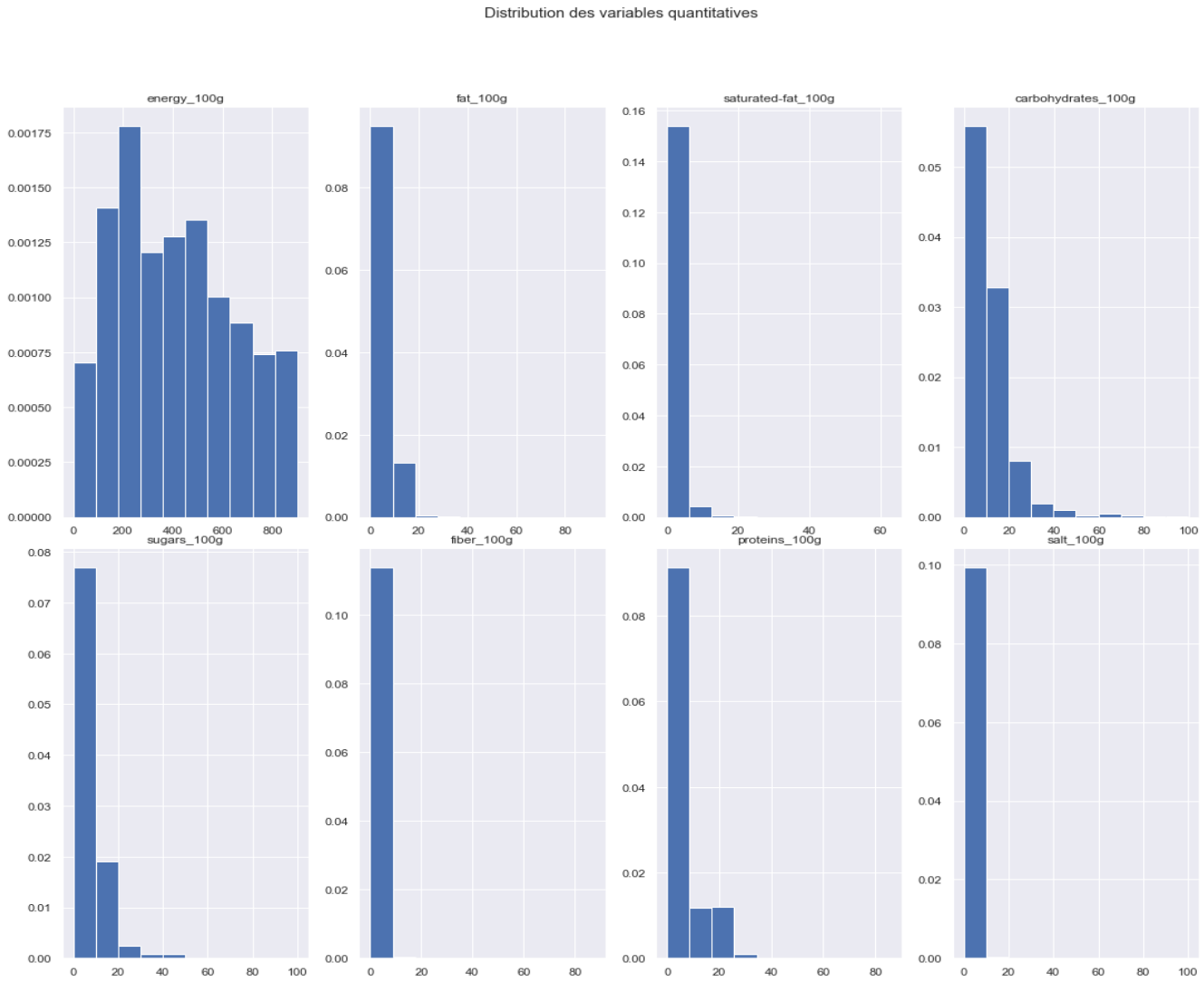
Classement des Nutriscore grade



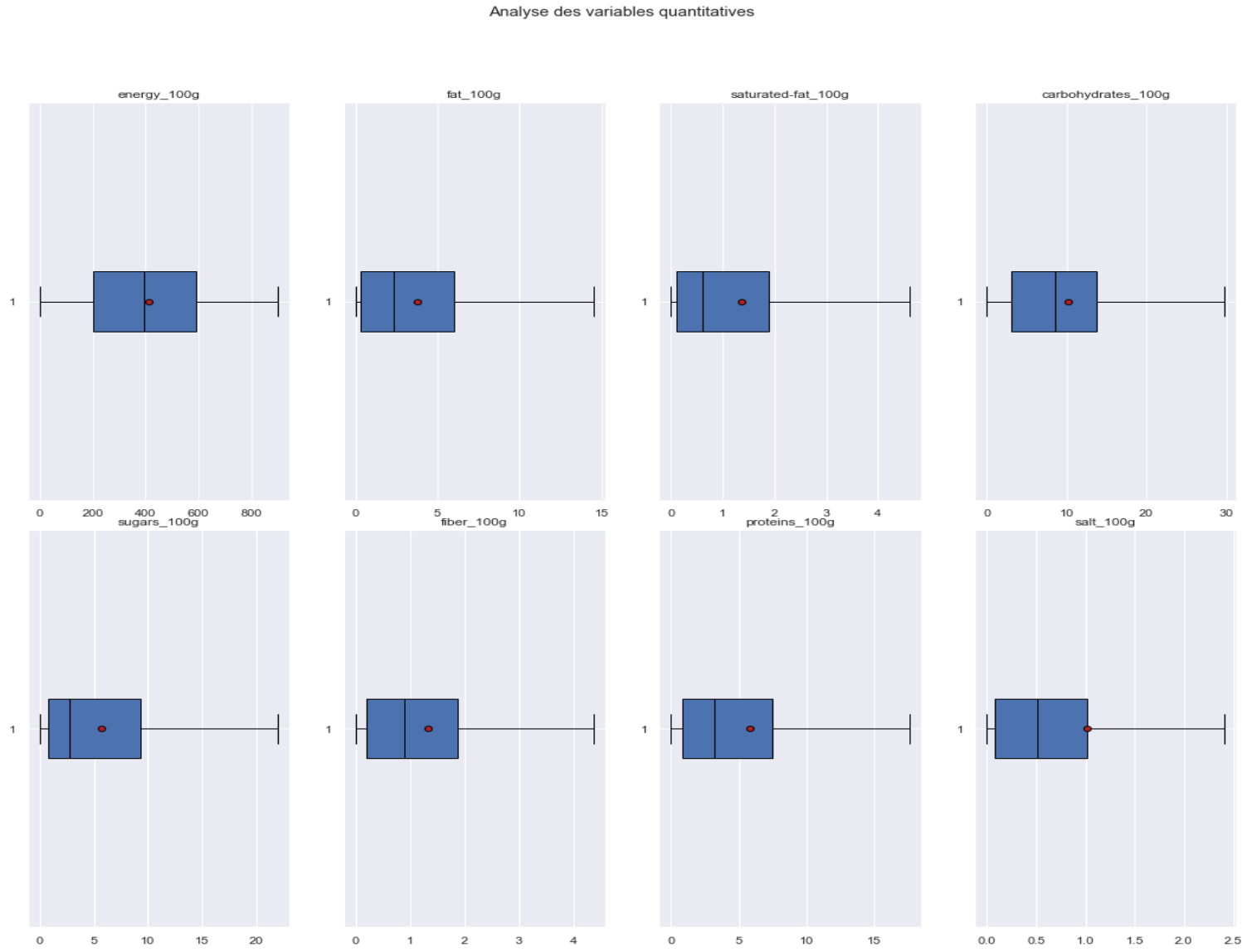
Classement du Gluten



Analyses univariées

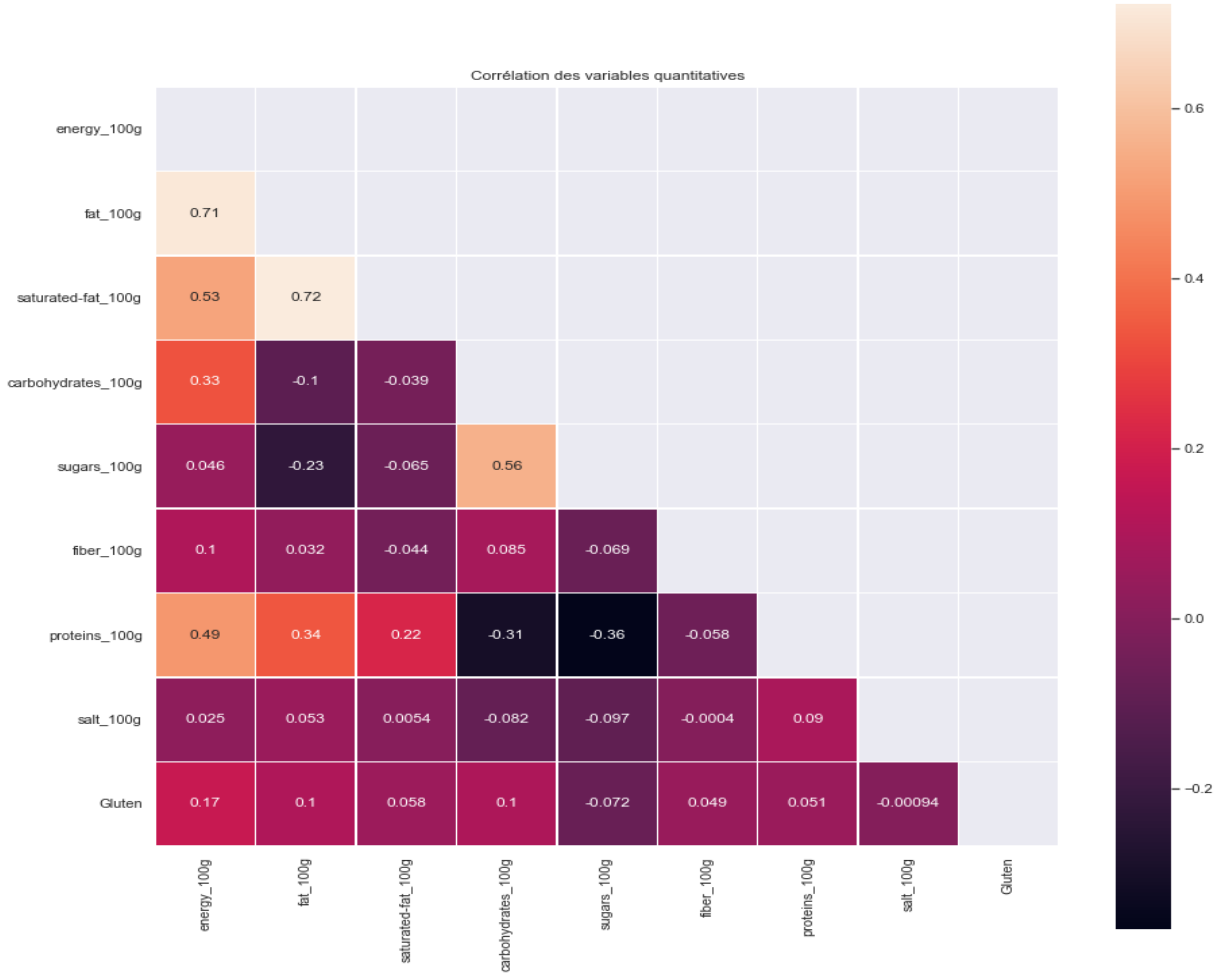


Analyses univariées



Analyses bi-variées

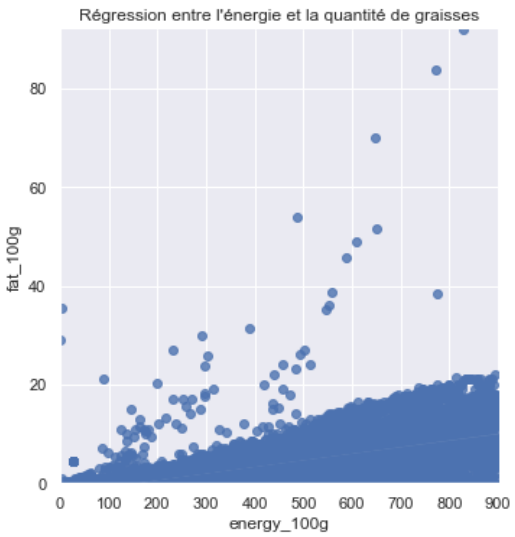
Analyses bivariées / Corrélation

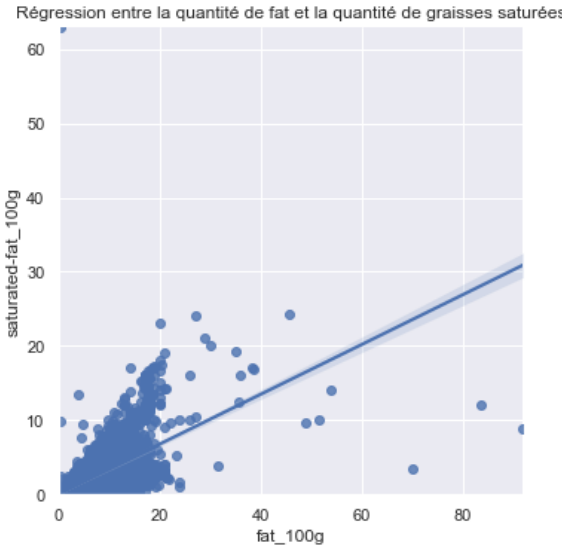


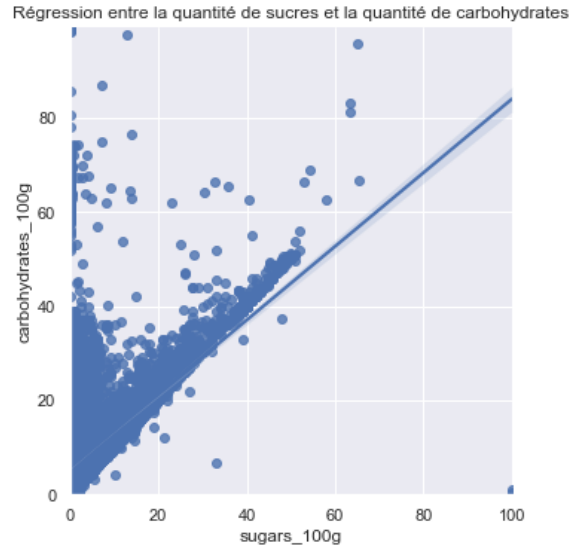
3 principales corrélations:

- ✓ Energie & fat
- ✓ Fat & Saturated fat
- ✓ Carbohydrates & sucre

Analyses bivariées _ Régression linéaire







R^2=0.50

R^2=0.52

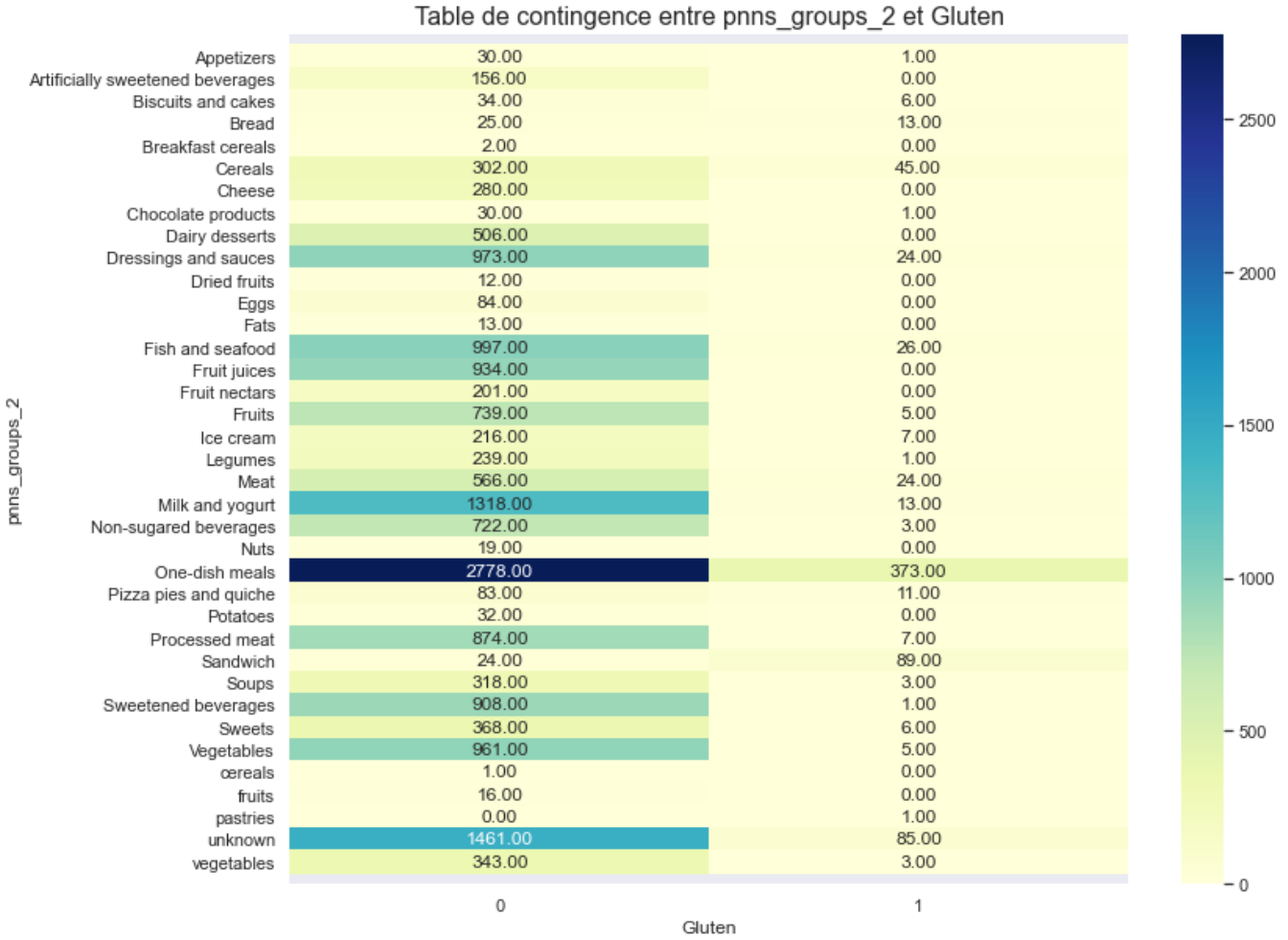
R^2=0.31

Absence de relations linéaires entre ces couples de variables

Analyses bivariées _ Table de contingence

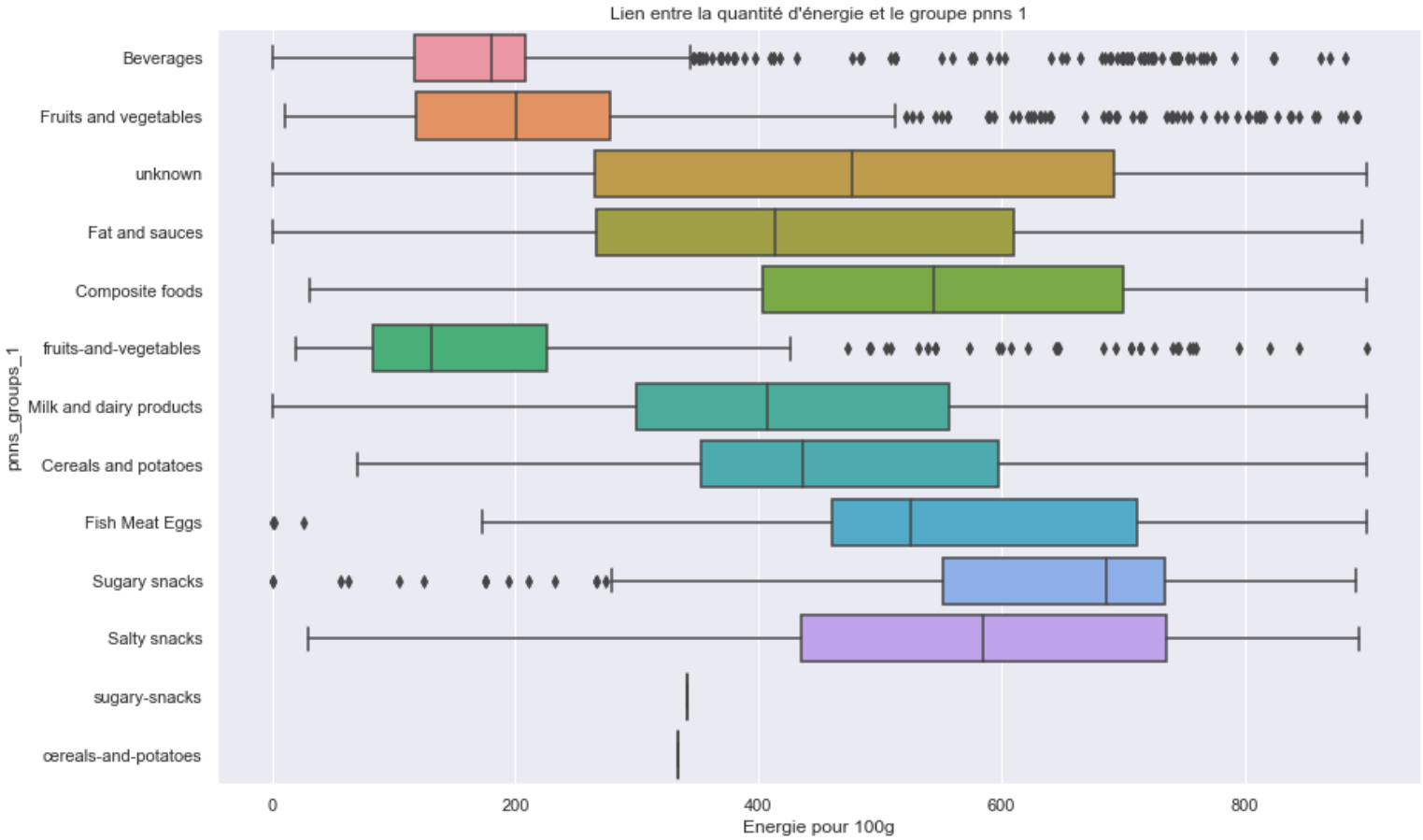
Les produits du groupe Composite foods sont majoritairement de contient du Gluten. Aprés la category de unknown contient du Gluten.

Analyses bivariées _ Table de contingence



Les produits du groupe One-dish meals sont majoritairement de contient du Gluten. Egalement, Processed meat et la category de unknown contient du Gluten.

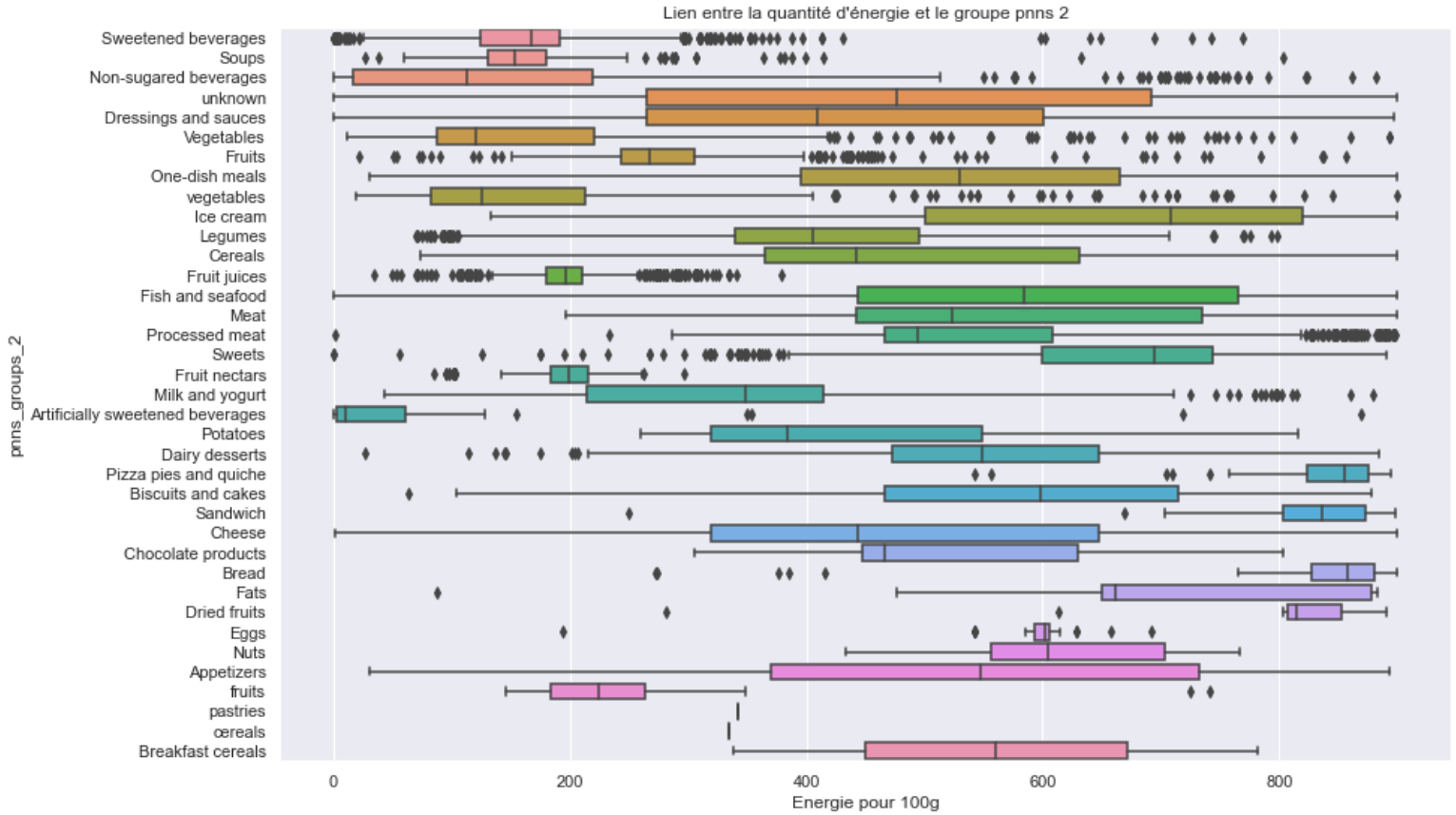
Analyses bivariées



ANOVA : p-value < 0.05, on rejette H0, différence entre moyenne statistiquement significative

Les fruits-and-vegetables et les beverages sont les groupes les moins caloriques
Les salty et sugary snacks sont le groupe le plus calorique.

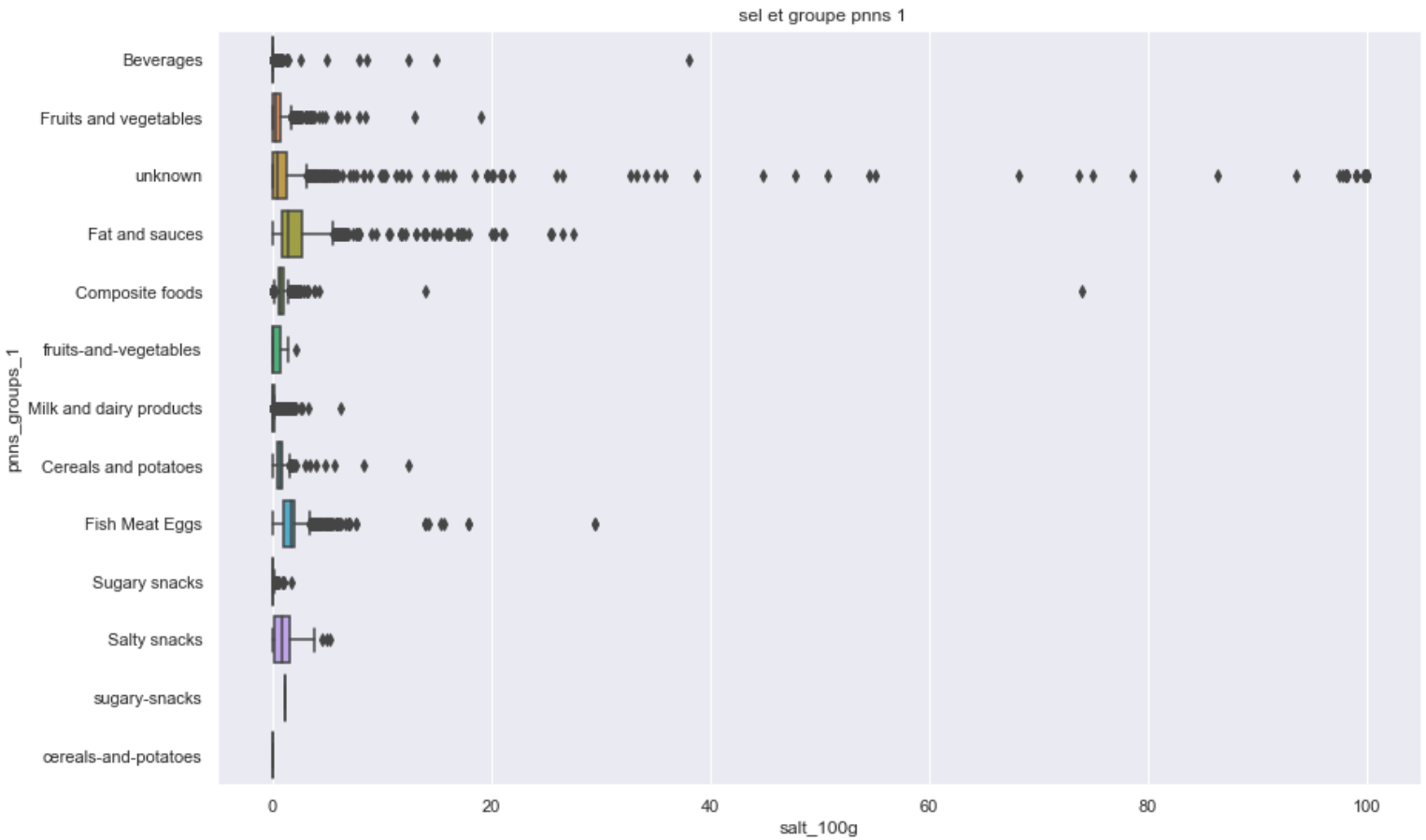
Analyses bivariées



ANOVA : p-value < 0.05

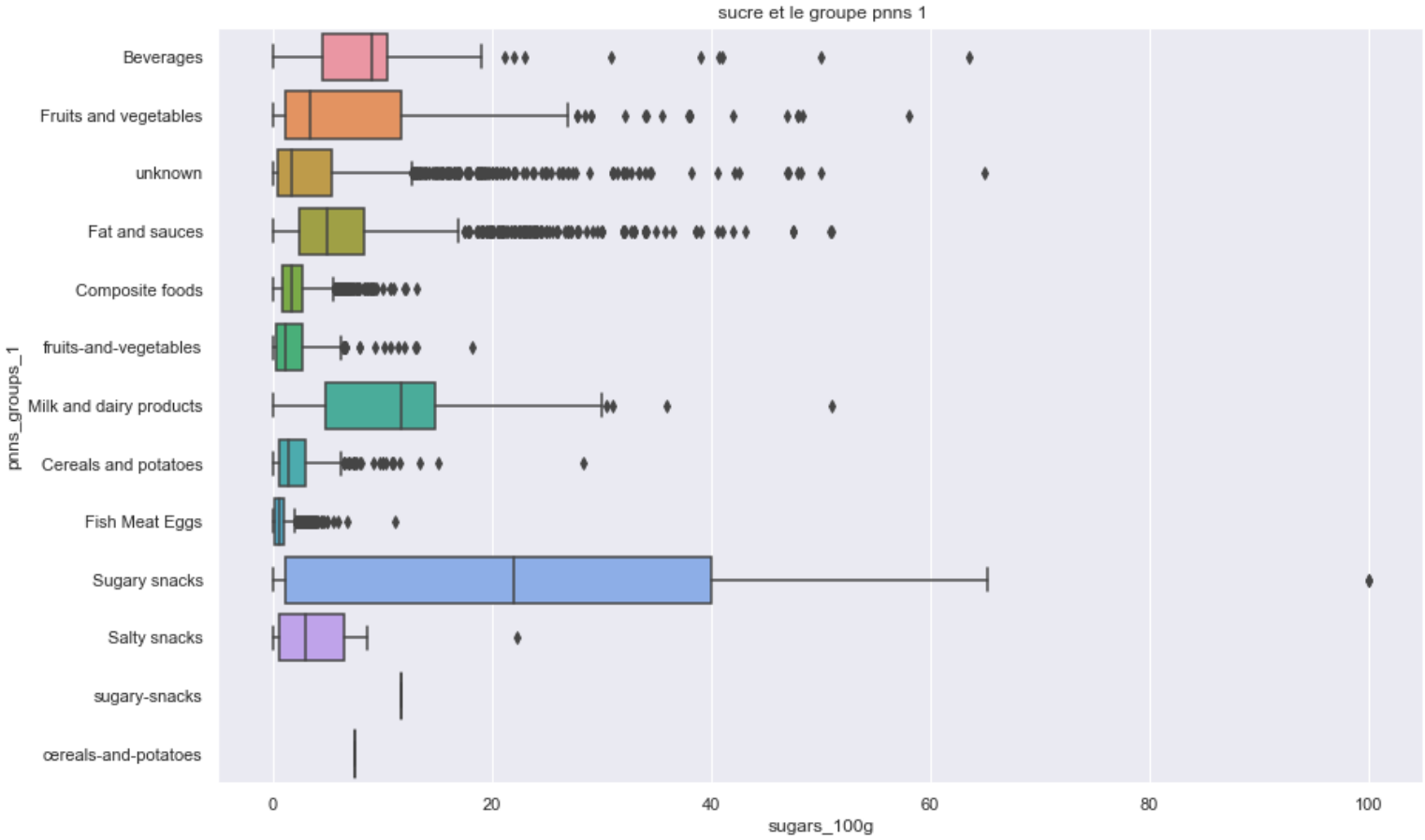
Ici, on voit que les artificially sweetened beverages sont les groupes les moins caloriques. Les Breads et les fats sont le groupe le plus calorique.

Analyses bivariées



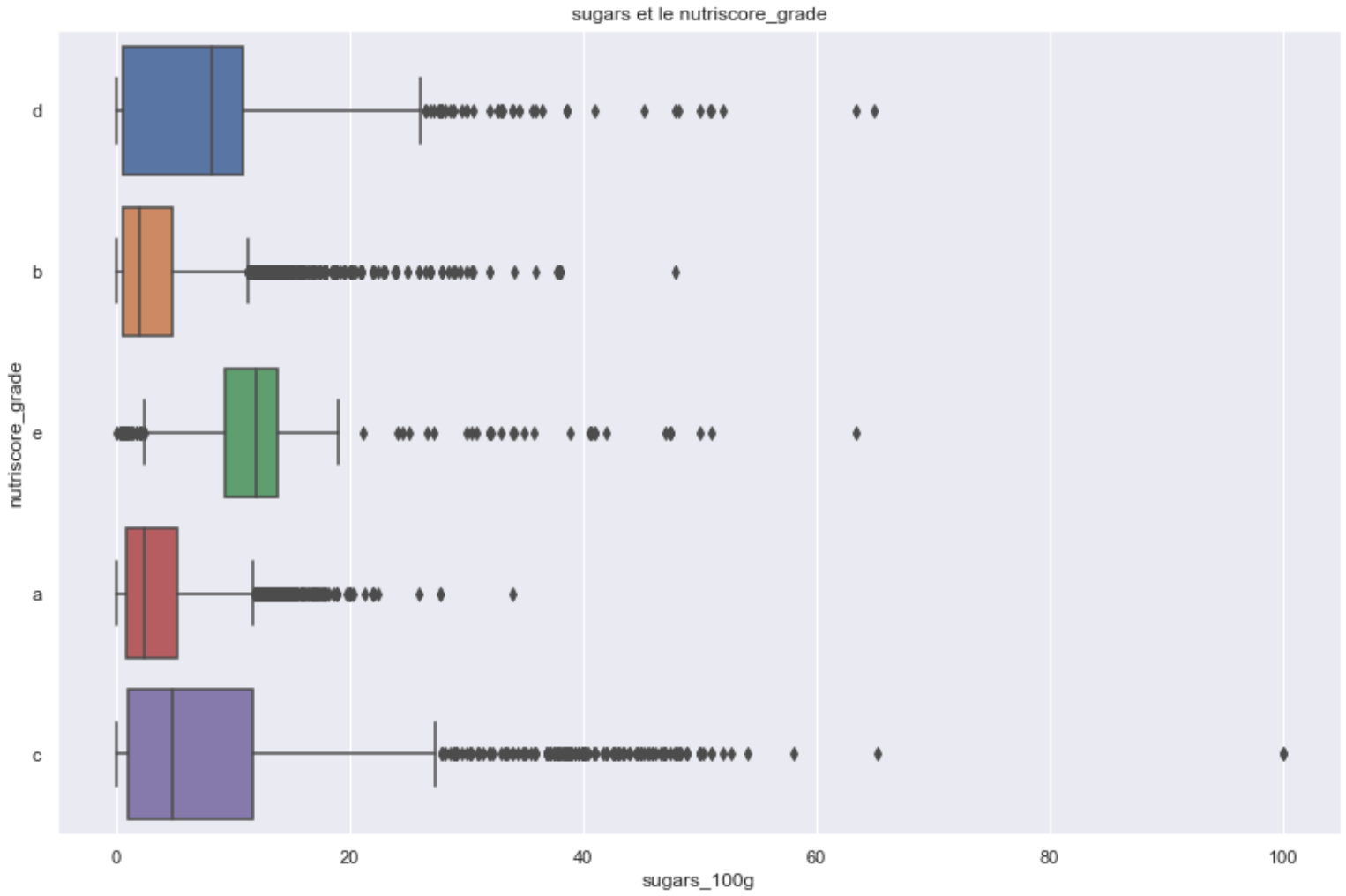
Les fat and sauces products sont les plus riches en sels.
Les sugary-snacks sont les plus pauvres en sel ce qui est tout à fait logique

Analyses bivariées



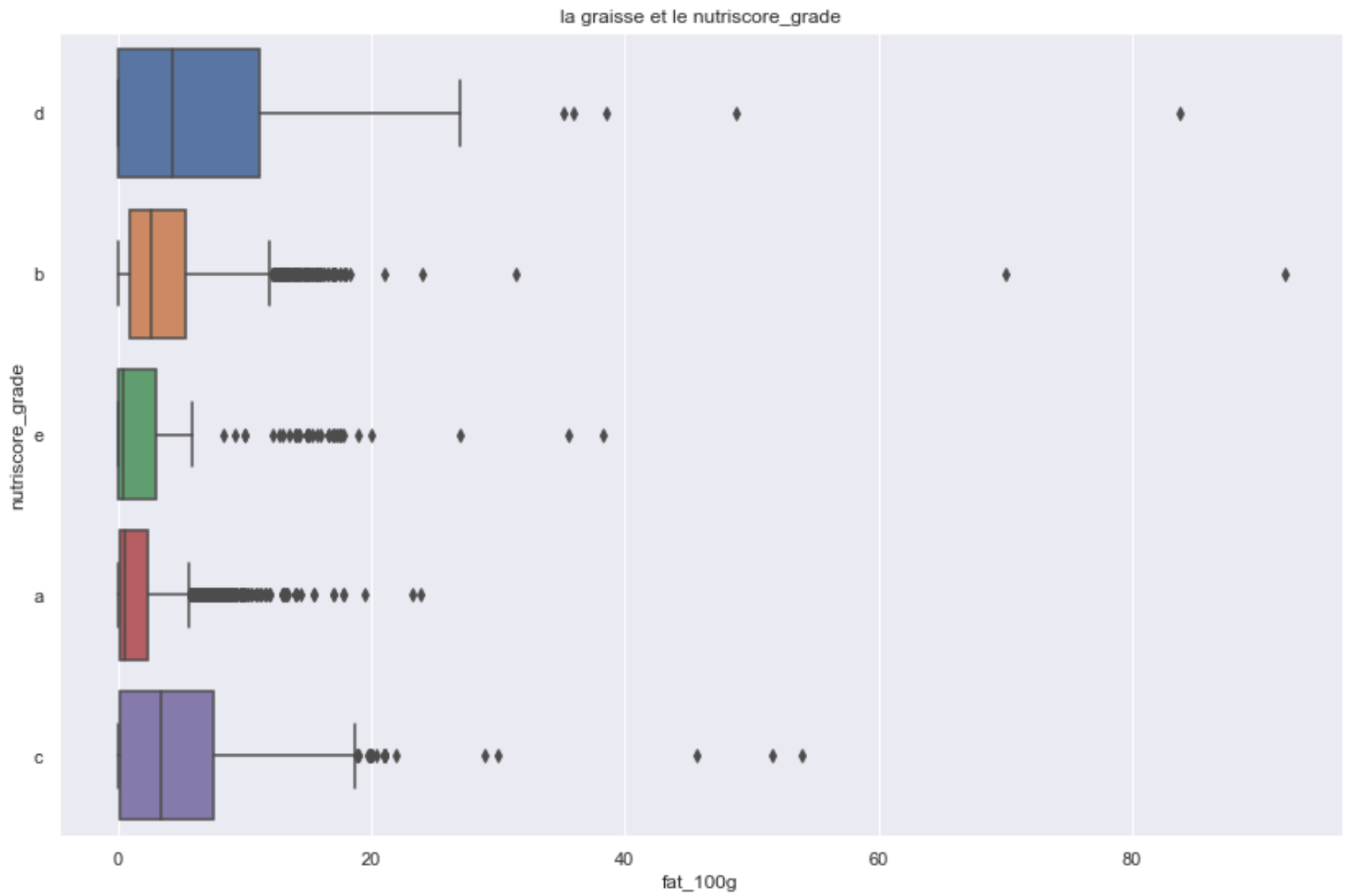
Les sugary snacks sont les plus riches en sucres
Les fish meat eggs sont les plus pauvres en sucres

Analyses bivariées



Le nutriscore A est le plus pauvre en sucre

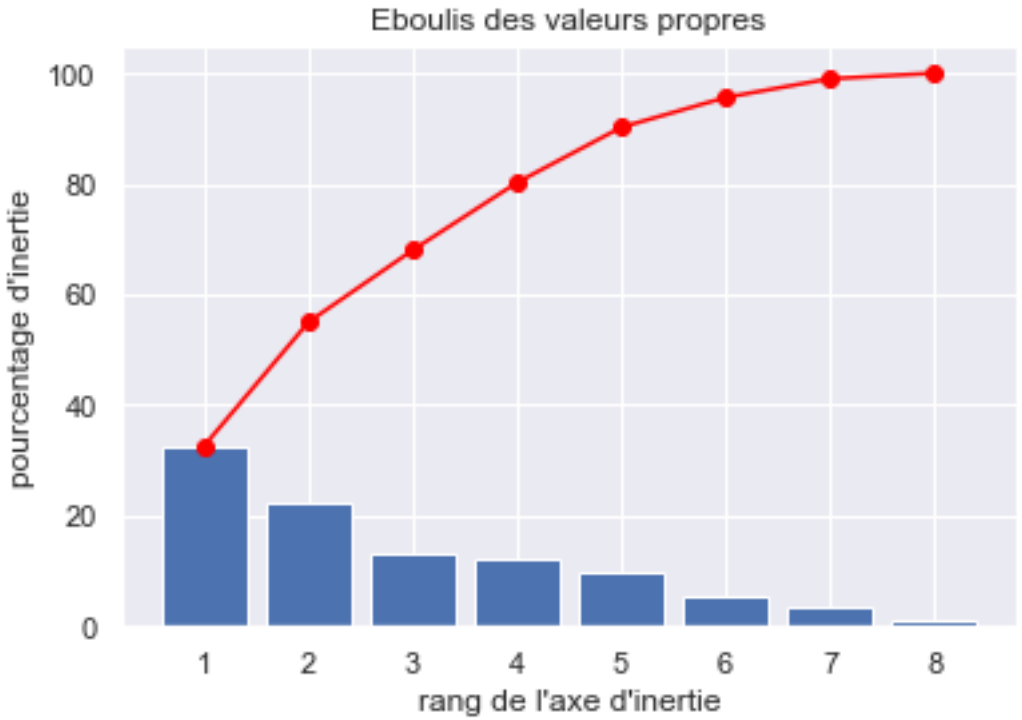
Analyses bivariées



Analyses multivariées

« *Manger Sans-Gluten* »

ACP

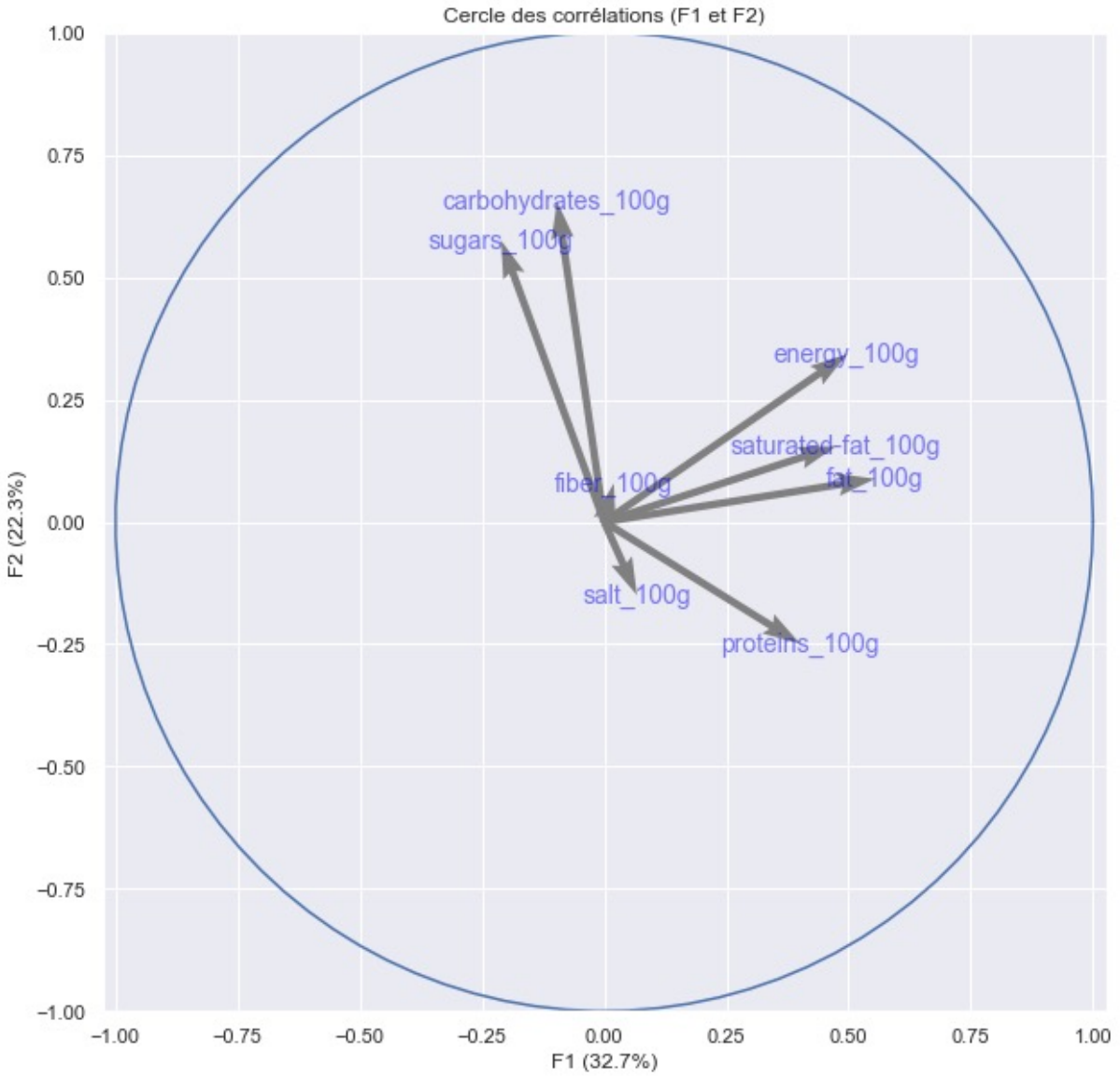


32 % de l'inertie totale sont associés à F1 , 22 % à F2

Le premier plan factoriel explique 54 % de la variance.

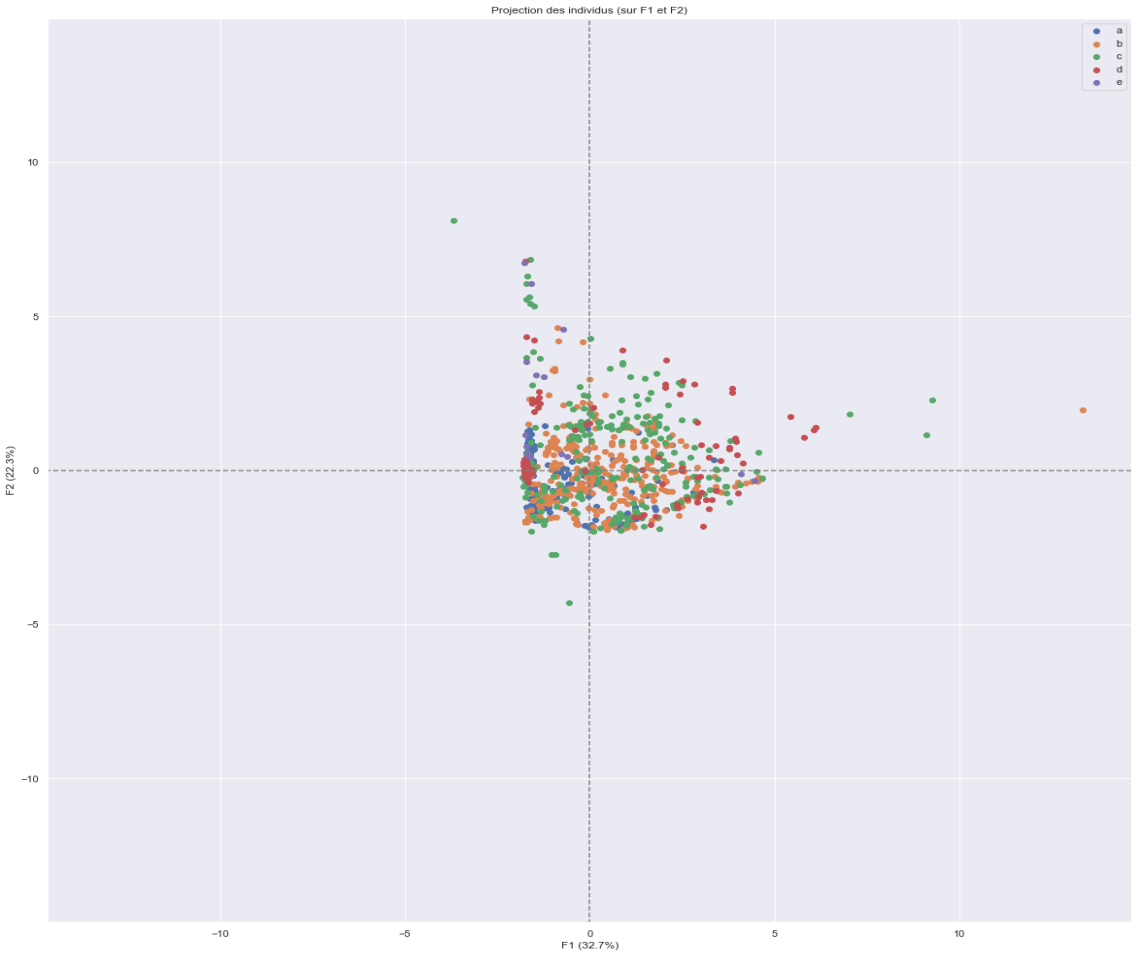
Pour avoir 75% de la variance nous garderons les 4 premières composantes

ACP



F1: Energie + lipides
F2: sucres+ carbohydrates

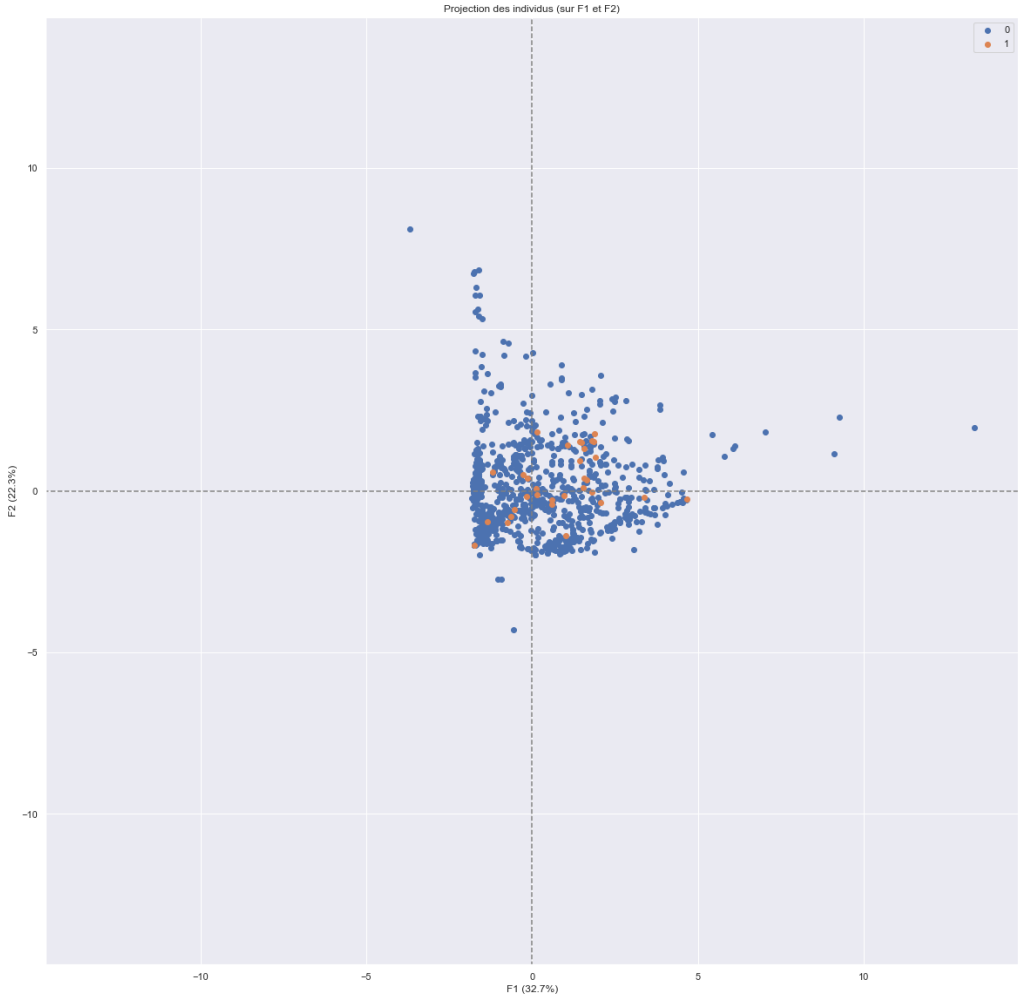
ACP



Les produits des groupes A et B contiennent de faibles quantités de sel et de sucre.

ACP

(Projection des individus Gluten)



Les produits avec ou sans gluten contiennent différentes quantités de sel, du fat, de protéines et de carbohydrate. Le résultat n'a pas été significatif

Faisabilité

« *Manger Sans-Gluten* »

Manger Sans-Gluten

Indique score du produit s'il contient du gluten : interdit
Indique score du produit s'il contient pas du gluten: Autorisé

Conclusion

« *Manger Sans-Gluten* »

Manger Sans-Gluten

- ✓ *Notre plateforme de données contient les informations nutritionnelles nécessaires pour le fonctionnement de notre application à savoir la quantité de gluten, sucre, calories et sel.*
- ✓ *La reconnaissance d'un produit se fait par son code à barre.*
- ✓ *Nous constatons qu'il y a des produits de la catégorie A ou B qui contiennent un taux élevé de graisse ou de sel, donc nous pouvons étudier le calcul du nutriscore.*
- ✓ *Possibilité d'améliorer notre application en proposant des produits alternatif (équivalent) sans gluten lors du scan d'un produit avec gluten.*

Merci!

« *Manger Sans-Gluten* »