

# Concevoir une application au service de la santé publique

29/04/2021 - Parcours Data Scientist  
Sébastien Bourgeois

# Sommaire

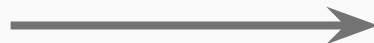
1. Idée d'application
2. Nettoyage effectué
3. Analyse exploratoire
4. Faits pertinents pour l'application

# 1. Idée d'application

Contexte



Appel à projets



**Idées d'applications** en lien avec l'**alimentation**

# 1. Idée d'application

Constat

Industrie & commerce

10

ingrédients




Fait maison

4

ingrédients

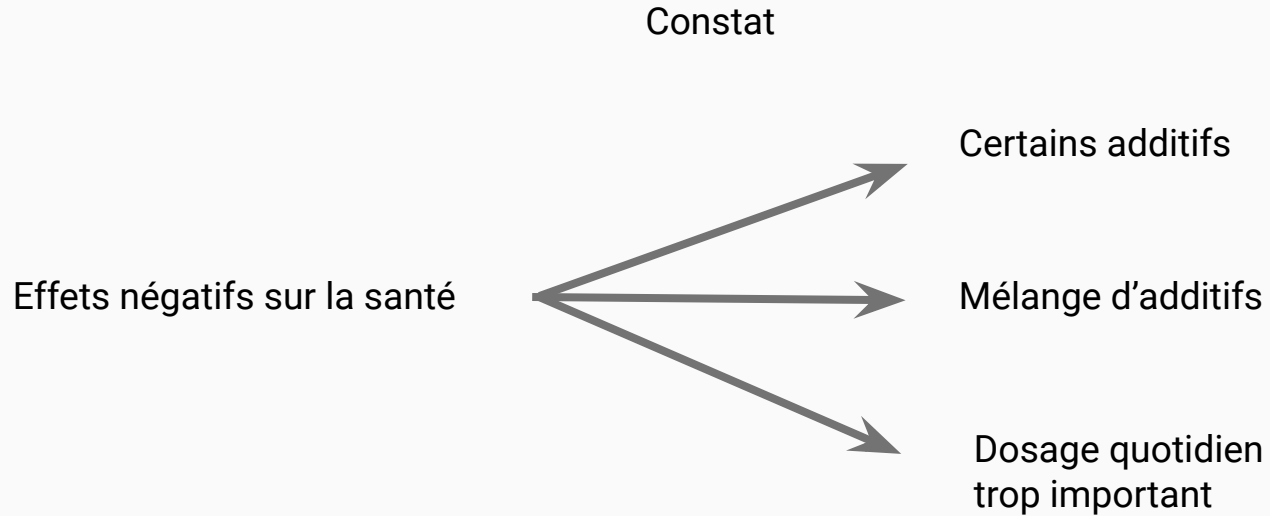
# 1. Idée d'application

Constat

Transformation industrielle  Utilisation **d'additifs**

- Conservateurs
- Colorants
- Exhausteurs de goût
- Épaississants
- Etc.

# 1. Idée d'application



# 1. Idée d'application

## Enjeux consommateurs

- Privilégier le fait-maison
- Limiter ou éviter les additifs

# 1. Idée d'application

## Outil

- Rôle : **informer** les consommateurs
- Identifier les produits à éviter/à privilégier
- Identifier les catégories ou les types de produits à éviter/à privilégier



# 1. Idée d'application

## Données

- Open Food Facts
- Base de données de produits

## 2. Nettoyage effectué

*1ère étape : sur le fichier brut des données*

Première **sélection** des variables

- Suppression des metadata
- Suppression des variables liées à l'emballage, l'image du produit ou encore aux allergènes
- Suppression des variables avec moins de 100 données

## 2. Nettoyage effectué

*1ère étape : sur le fichier brut des données*

Filtre sur les produits

- Conservation des produits vendus en **France** uniquement
- **Pertinent** pour Santé Publique France

## 2. Nettoyage effectué

*1ère étape : sur le fichier brut des données*

Optimisation de la mémoire

- Conversion des *object* en *category*
- Conversion des *float64* en *float32*
- Export du dataset pour la 2e étape du nettoyage

## 2. Nettoyage effectué

*2ème étape : sur le fichier optimisé des données*

Seconde sélection des variables

- Suppression des variables avec plus de **80% de données manquantes** excepté 'additives\_tags'
- Suppression des variables impertinentes à l'application
- Suppression des variables "doublons"

	categories	categories_tags	categories_en
368922	Crêpes et galettes, Crêpes, Crêpes de froment	en:crepes-and-galettes,en:crepes,fr:crepes- de-...	Crêpes and galettes,Crêpes,fr:Crêpes de froment
477614	Aliments et boissons à base de végétaux, Alime...	en:plant-based-foods-and- beverages,en:plant-ba...	Plant-based foods and beverages,Plant-based fo...

## 2. Nettoyage effectué

*2ème étape : sur le fichier optimisé des données*

Suppression des produits avec des données manquantes

- Sans nombre d'additifs renseigné
- Avec un taux de remplissage inférieur à 90%

## 2. Nettoyage effectué

*2ème étape : sur le fichier optimisé des données*

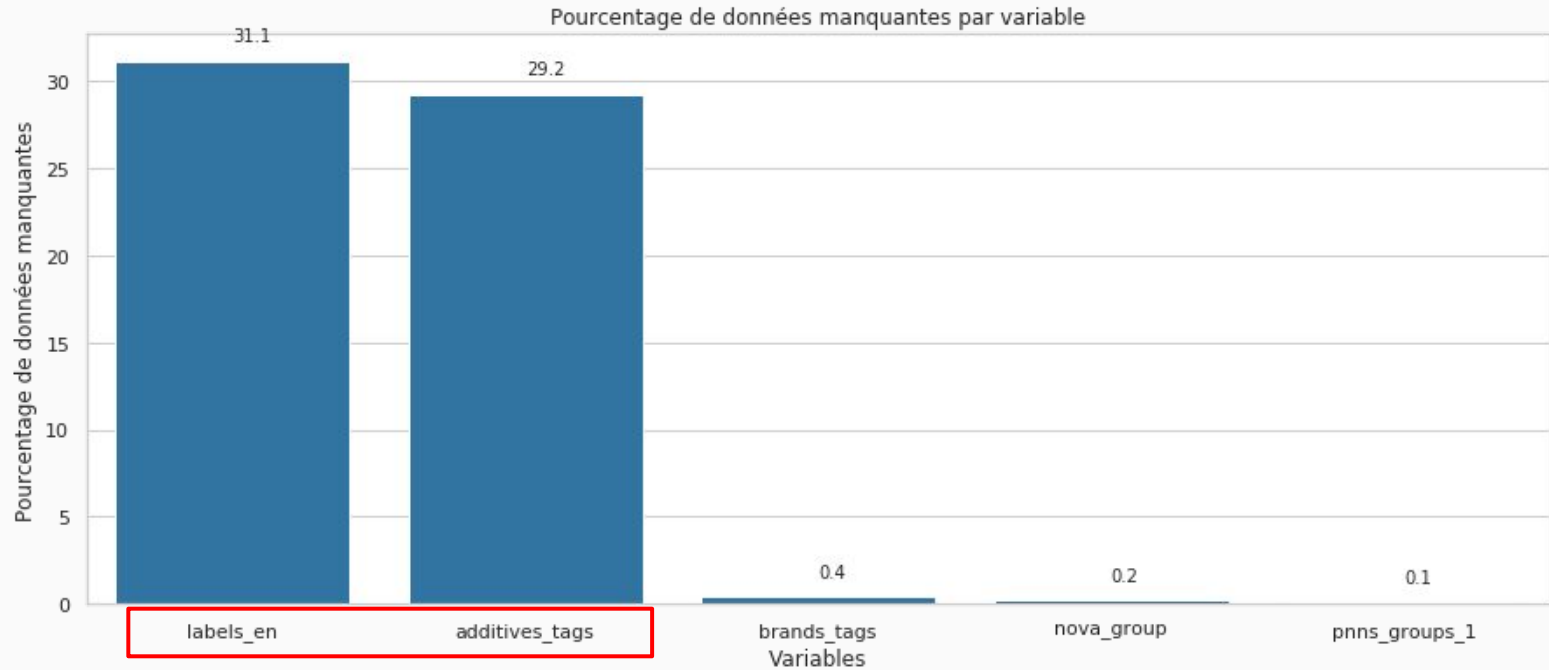
```
▶ open_food_facts_optimized_data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 104300 entries, 64 to 700623
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype  
---  -
 0   brands_tags           103844 non-null  category
 1   labels_en             71866 non-null  category
 2   additives_tags        73811 non-null  category
 3   nutriscore_grade     104300 non-null  category
 4   pnns_groups_1        104159 non-null  category
 5   pnns_groups_2        104300 non-null  category
 6   additives_n           104300 non-null  float32
 7   nutriscore_score     104300 non-null  float32
 8   nova_group           104067 non-null  float32
dtypes: category(6), float32(3)
memory usage: 7.7 MB
```

104K lignes x 9 colonnes

### 3. Analyse exploratoire

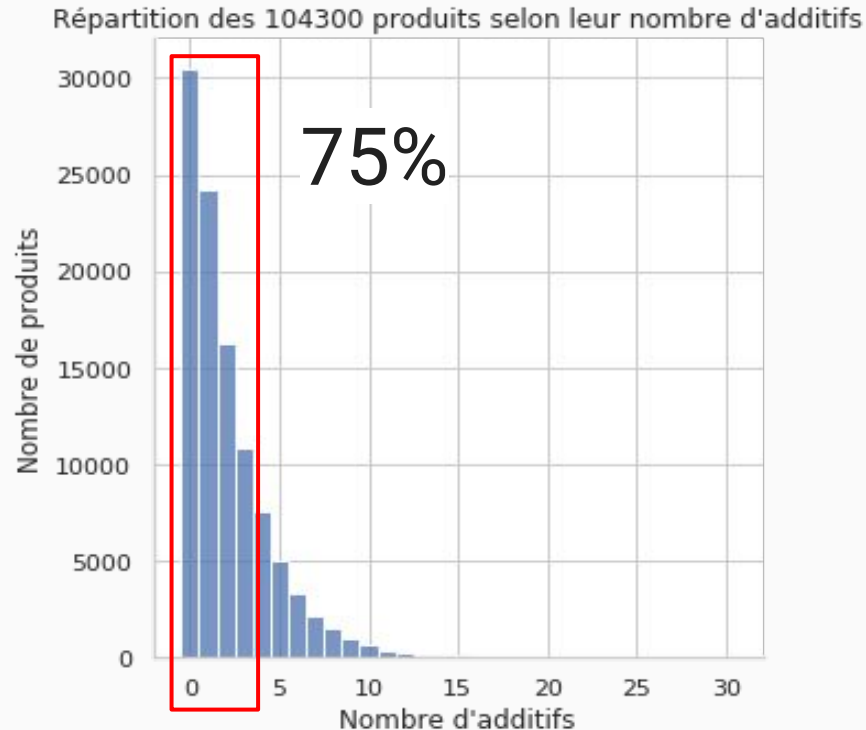
#### *Données manquantes*





### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?

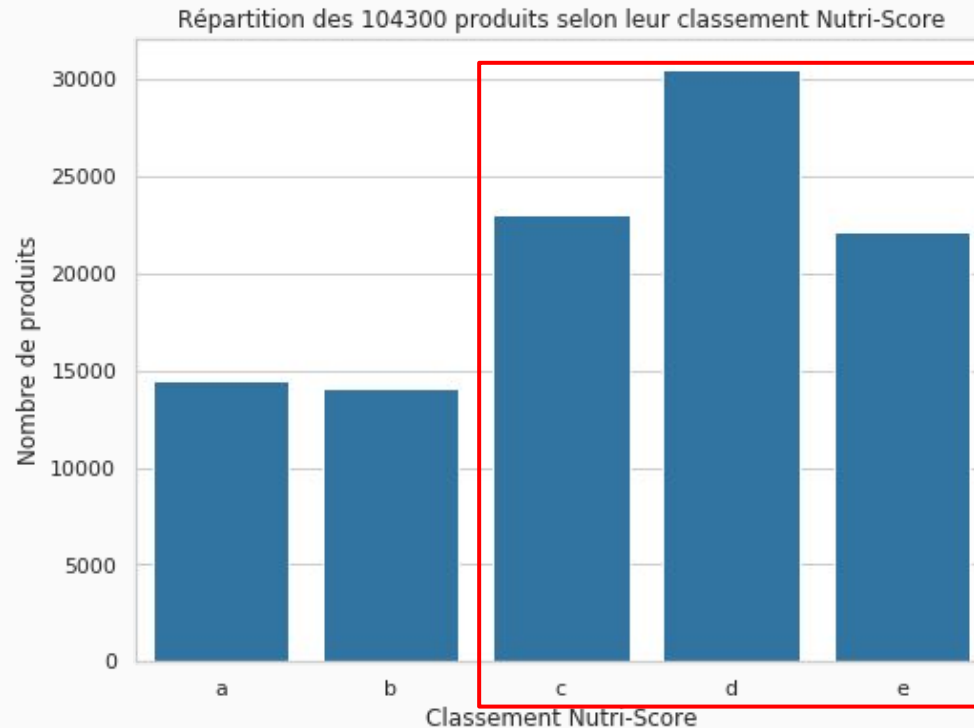
*Nombre d'additifs présents dans les produits*



$$\bar{x} = 2,14, \quad \sigma = 2,48, \quad \text{skewness} = 1,96$$

### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?

*Qualité nutritionnelle des produits (rang nutriscore)*

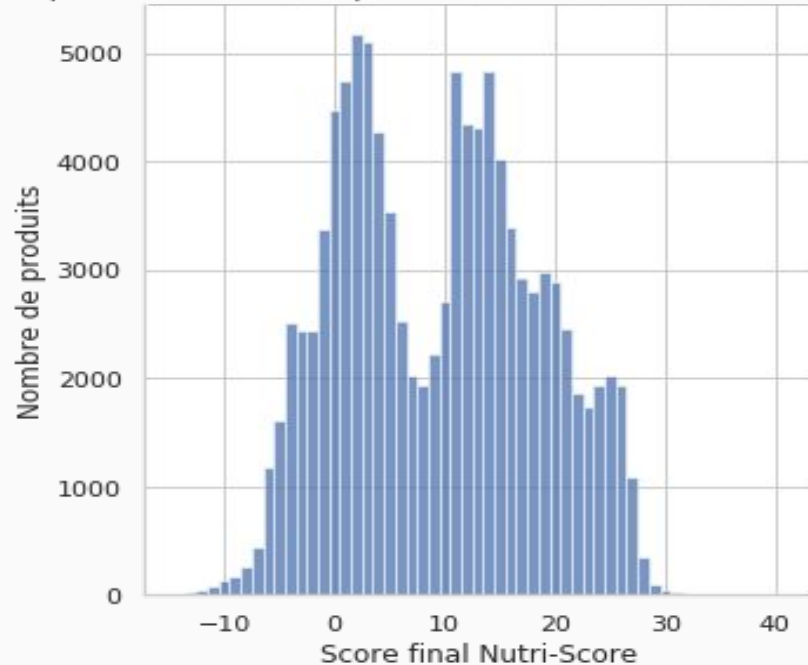


75%

### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?

*Qualité nutritionnelle des produits (points nutriscore)*

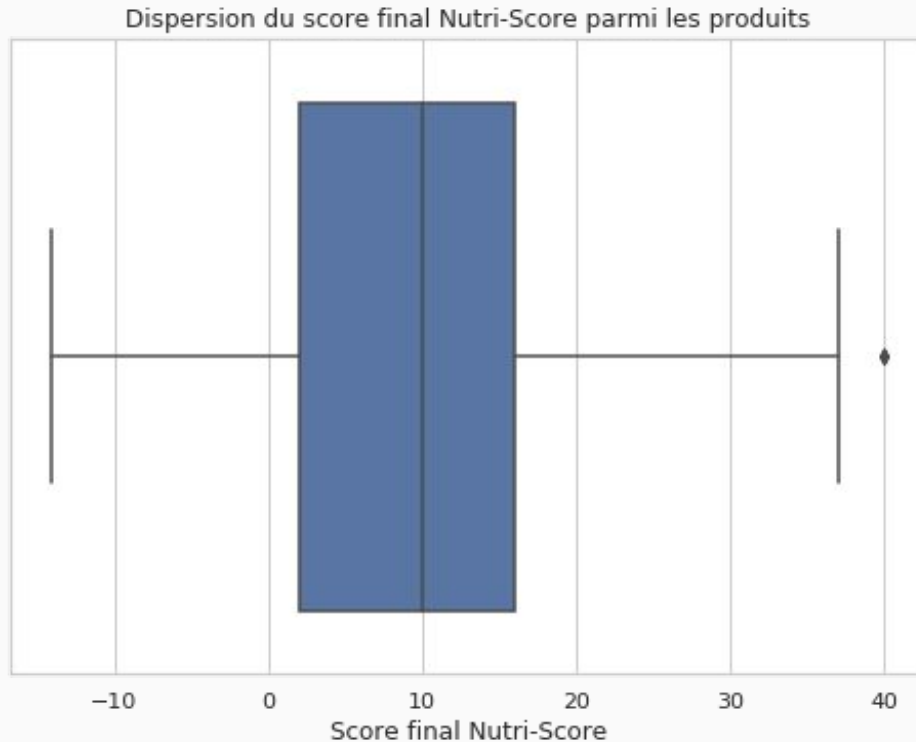
Répartition des 104300 produits selon leur score final Nutri-Score



$$\bar{x} = 9,49, \quad \sigma = 8,93$$

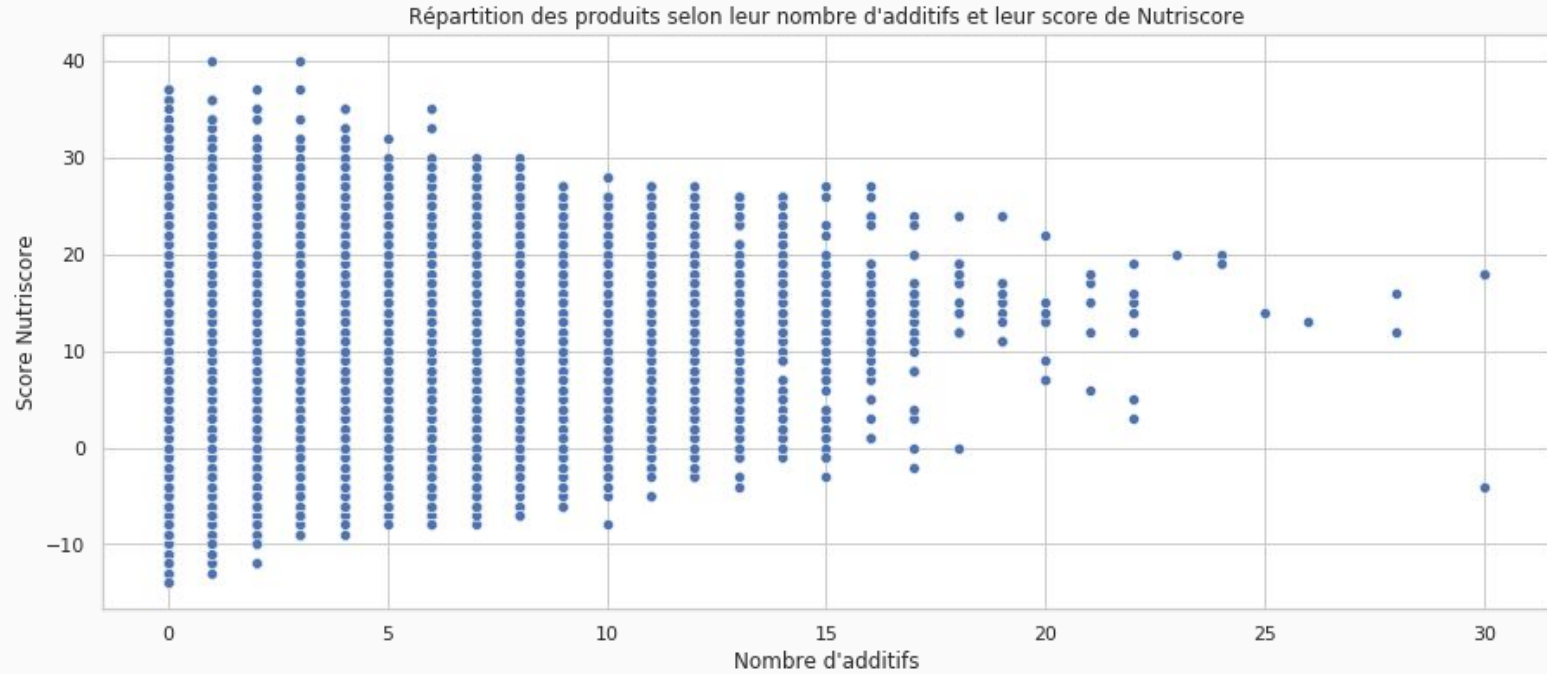
### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?

*Qualité nutritionnelle des produits (points nutriscore)*



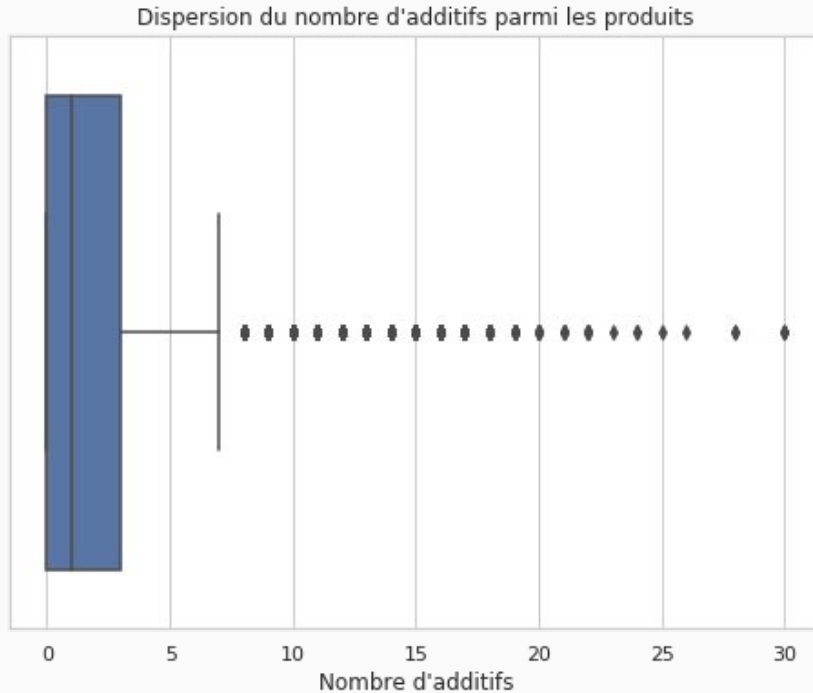
$Q1 = 2$ ,  $m = 10$ ,  $Q3 = 16$

### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?



$$Q = 0,184$$

### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?

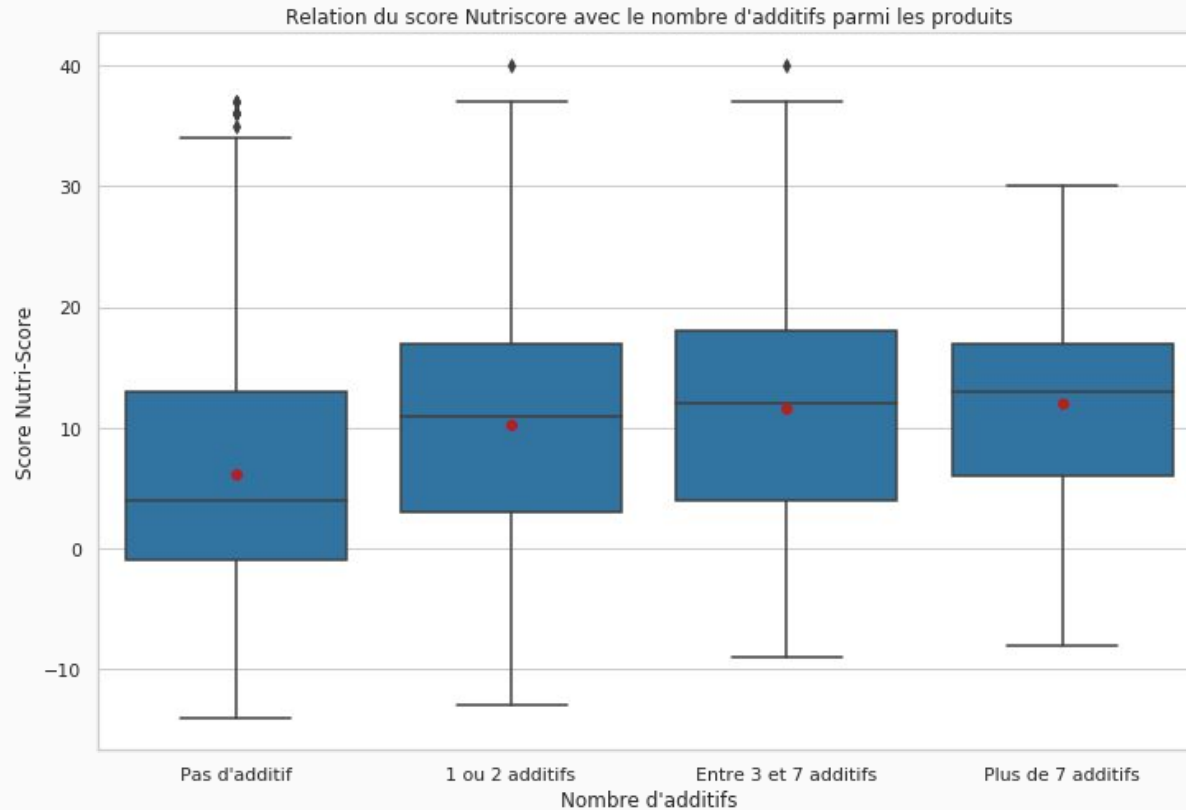


$Q1 = 0$ ,  $m = 1$ ,  $Q3 = 3$

Classes :

- Pas d'additif
- 1 ou 2 additifs
- Entre 3 et 7 additifs
- Plus de 7 additifs

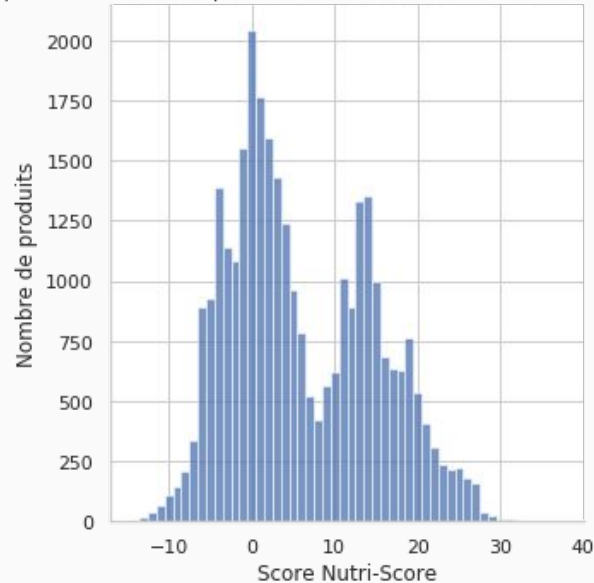
### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?



### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?

#### *Classe 'Pas d'additif'*

Répartition des 30489 produits sans additif selon leur score Nutri-Score



Skewness = 0,40 > 0

Kurtosis = -0,77 != 0

Test d'Agostino pour vérifier la normalité :

- $H_0$  : la distribution est normale

À 5% -> rejet de  $H_0$ , p-value = 0.0



### 3. Analyse exploratoire : la qualité nutritionnelle dépend-elle du nombre d'additifs ?

*Utilisation du rang Nutriscore croisé avec le nombre d'additifs par classe*

additives_n_classes	1 ou 2 additifs	Entre 3 et 7 additifs	Pas d'additif	Plus de 7 additifs
nutriscore_grade				
a	4875	1629	7800	170
b	5028	3423	5246	417
c	9482	6392	6280	862
d	10812	10077	7568	2109
e	10327	7395	3595	813

Test du Khi-2 pour vérifier la dépendance :

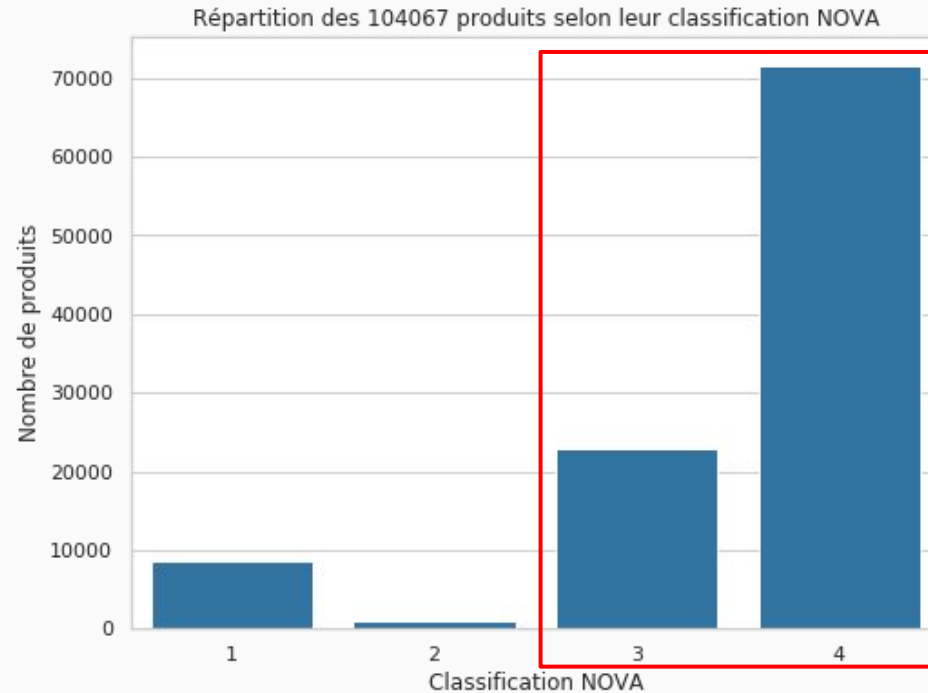
- $H_0$  : les deux variables sont indépendantes

La dépendance est significative entre la qualité nutritionnelle et le nombre d'additifs

À 5% -> rejet de  $H_0$ , p-value = 0.0

### 3. Analyse exploratoire : le nombre d'additifs dépend-il du degré de transformation ?

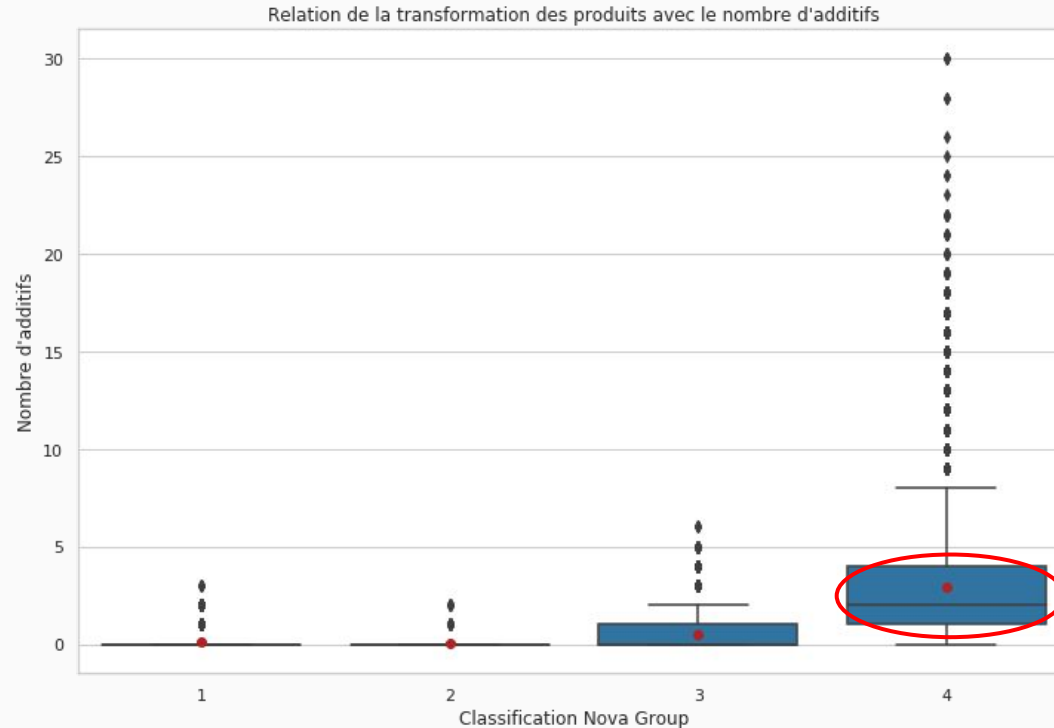
*Degré de transformation des produits (classement NOVA)*



90%

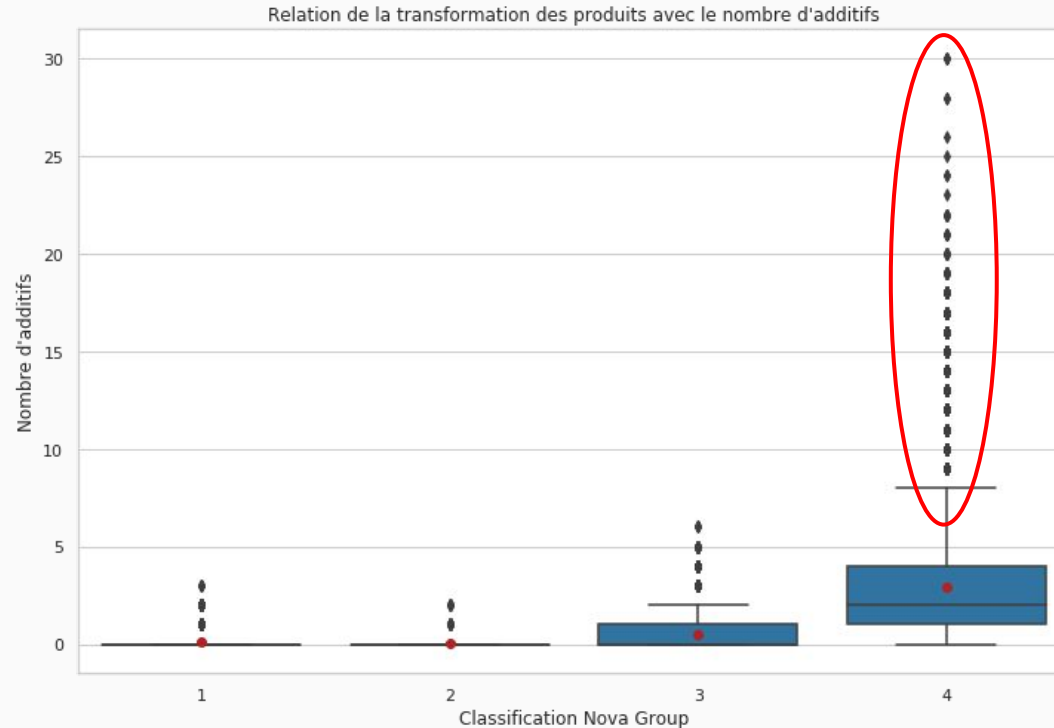
### 3. Analyse exploratoire : le nombre d'additifs dépend-il du degré de transformation ?

*Répartition des produits selon leur degré de transformation et le nombre d'additifs par classe*



### 3. Analyse exploratoire : le nombre d'additifs dépend-il du degré de transformation ?

*Répartition des produits selon leur degré de transformation et le nombre d'additifs par classe*



### 3. Analyse exploratoire : le nombre d'additifs dépend-il du degré de transformation ?

*Répartition des produits selon leur degré de transformation et le nombre d'additifs par classe*

additives_n_classes	1 ou 2 additifs	Entre 3 et 7 additifs	Pas d'additif	Plus de 7 additifs
nova_group				
1.0	848.0	4.0	7766.0	NaN
2.0	22.0	NaN	1002.0	NaN
3.0	7196.0	538.0	15058.0	NaN
4.0	32239.0	28360.0	6663.0	4371.0

Effectifs < 5 -> insuffisants pour un test du Khi-2

### 3. Analyse exploratoire : le nombre d'additifs dépend-il du degré de transformation ?

*Répartition des produits selon leur degré de transformation et le nombre d'additifs par classe*

additives_n_classes	1 ou 2 additifs	Entre 3 et 7 additifs	Pas d'additif	Plus de 7 additifs
nova_group				
1.0	848.0	4.0	7766.0	NaN
2.0	22.0	NaN	1002.0	NaN
3.0	7196.0	538.0	15058.0	NaN
4.0	32239.0	28360.0	6663.0	4371.0

Regroupement des classes :

- '1 ou 2 additifs', 'Entre 3 et 7 additifs' et 'Plus de 7 additifs' -> 'Avec additifs'
- 'Pas d'additif' -> 'Sans additif'
- '1', '2', et '3' -> 'Peu ou pas de transformation'
- '4' -> 'Ultra-transformation'

### 3. Analyse exploratoire : le nombre d'additifs dépend-il du degré de transformation ?

*Répartition des produits selon leur degré de transformation et la présence d'additifs*

additives_n_classes_k2	Avec additif	Sans additif
nova_group_classes		
Peu ou pas de transformation	8608	23826
Ultra transformation	64970	6663

Test du Khi-2 pour vérifier la dépendance :

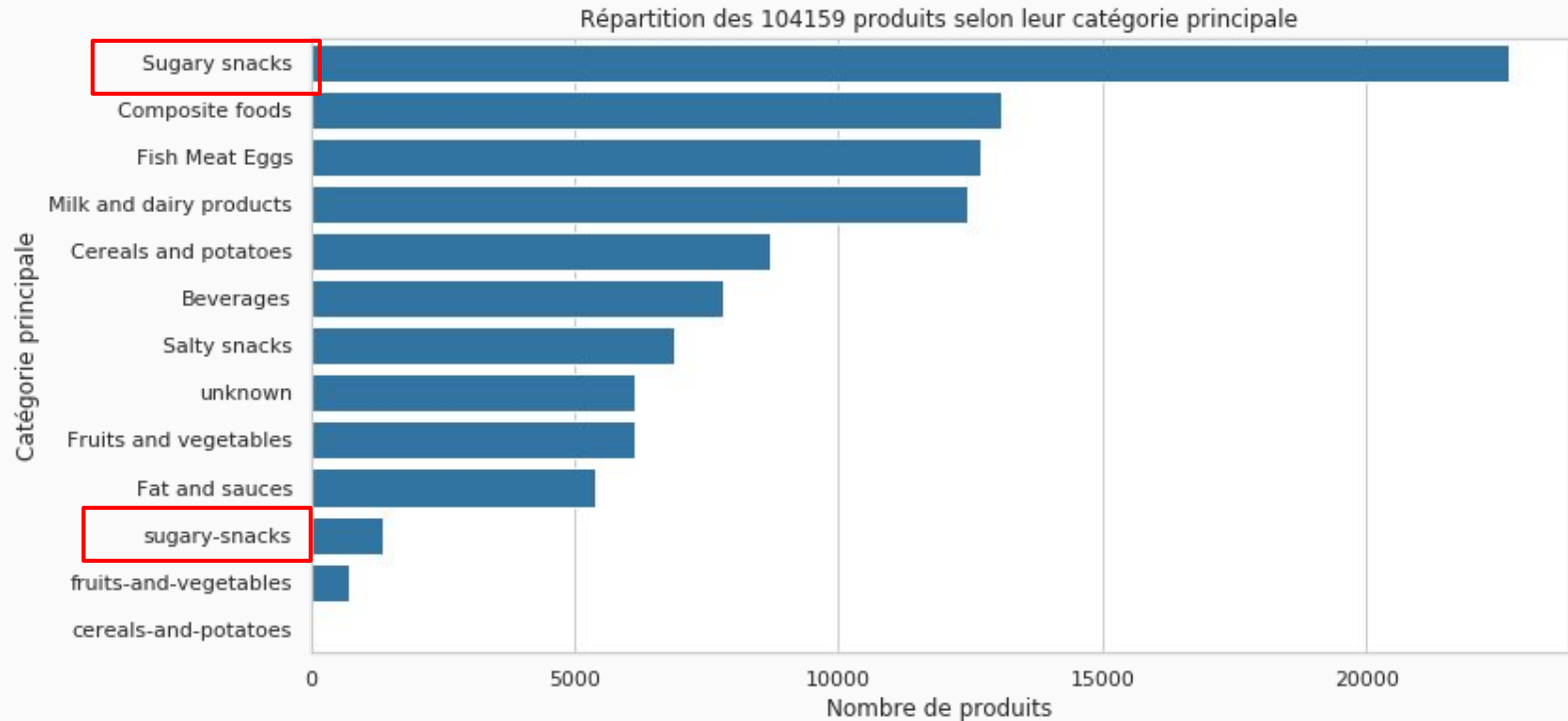
- $H_0$  : les deux variables sont indépendantes

La dépendance est significative entre le degré de transformation du produit et la présence d'additif(s)

À 5% -> rejet de  $H_0$ , p-value = 0.0

### 3. Analyse exploratoire : le nombre d'additifs dépend-il de la catégorie de produits ?

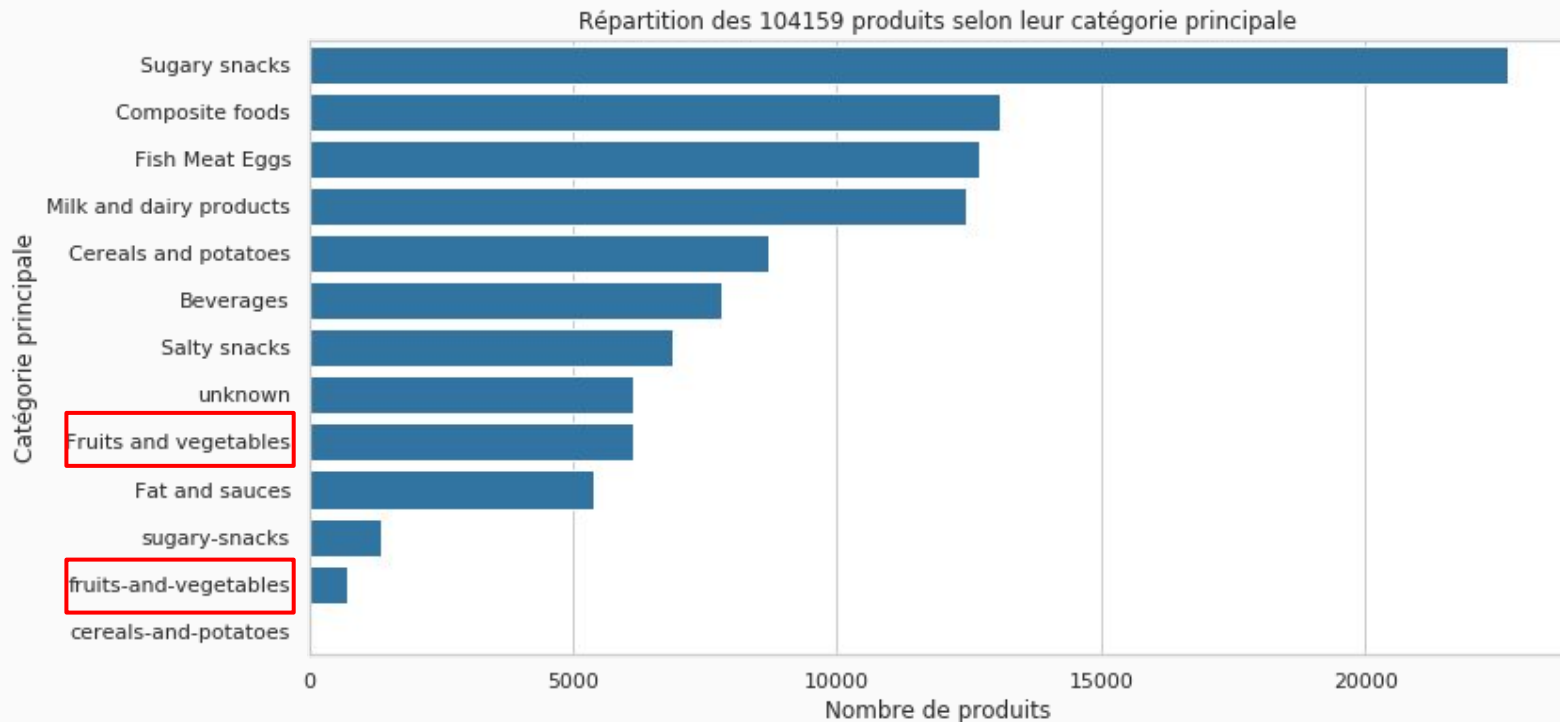
#### *Catégorie des produits*





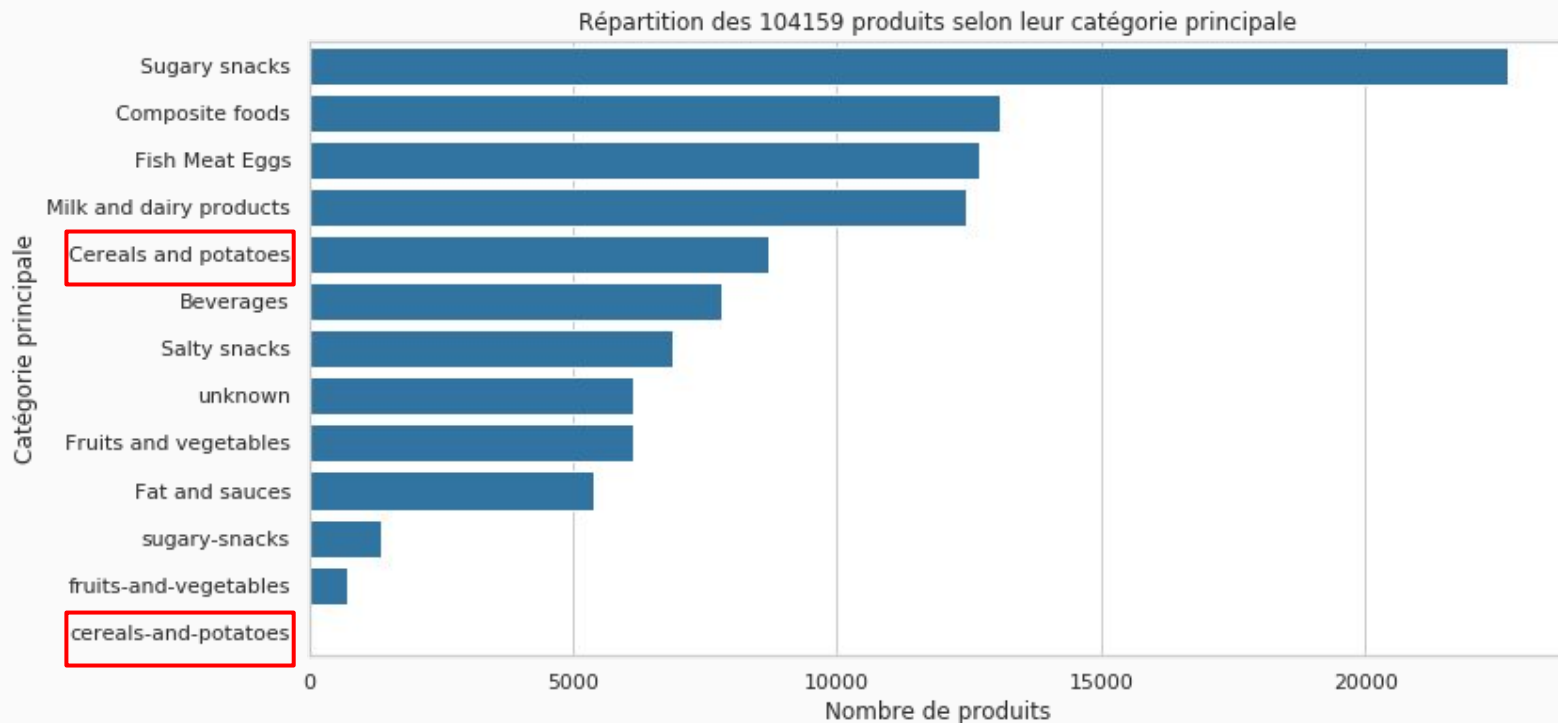
### 3. Analyse exploratoire : le nombre d'additifs dépend-il de la catégorie de produits ?

#### Catégorie des produits



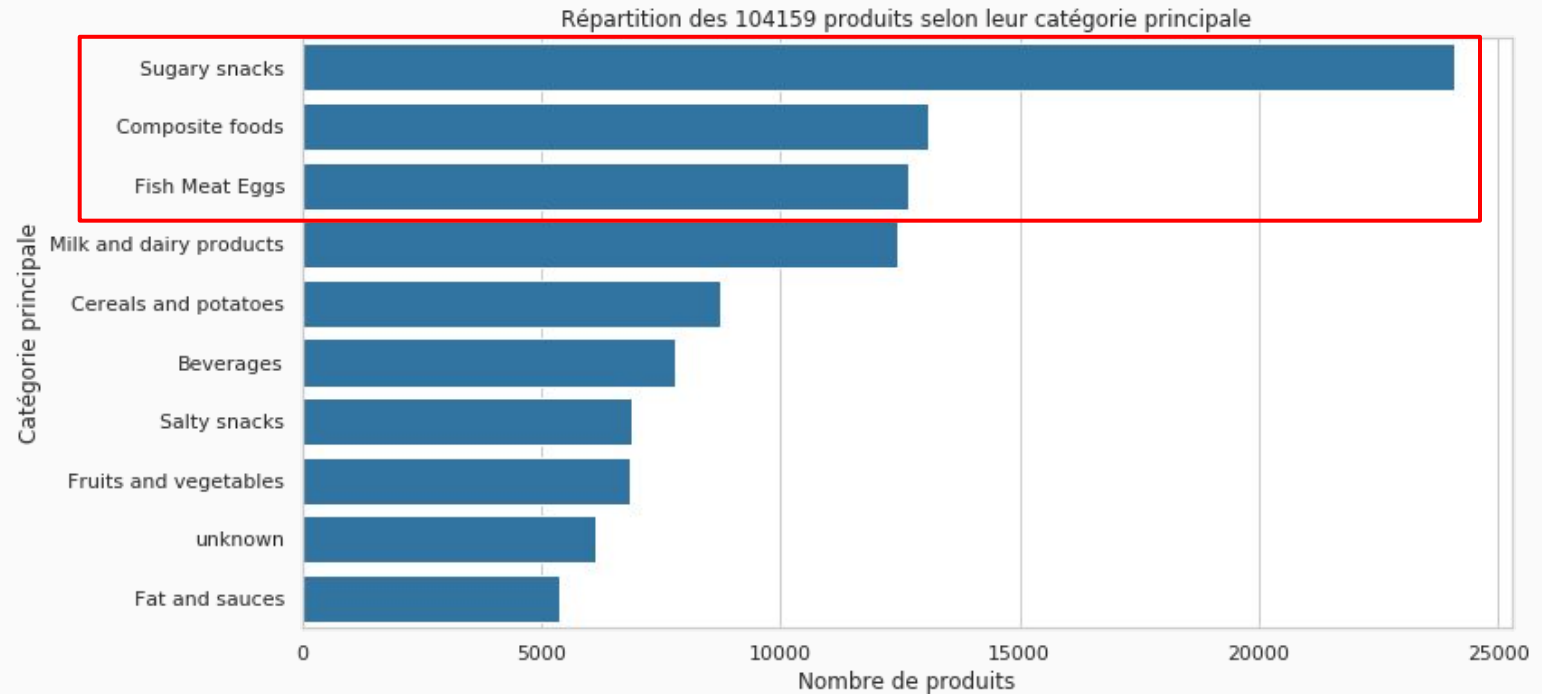
### 3. Analyse exploratoire : le nombre d'additifs dépend-il de la catégorie de produits ?

#### Catégorie des produits



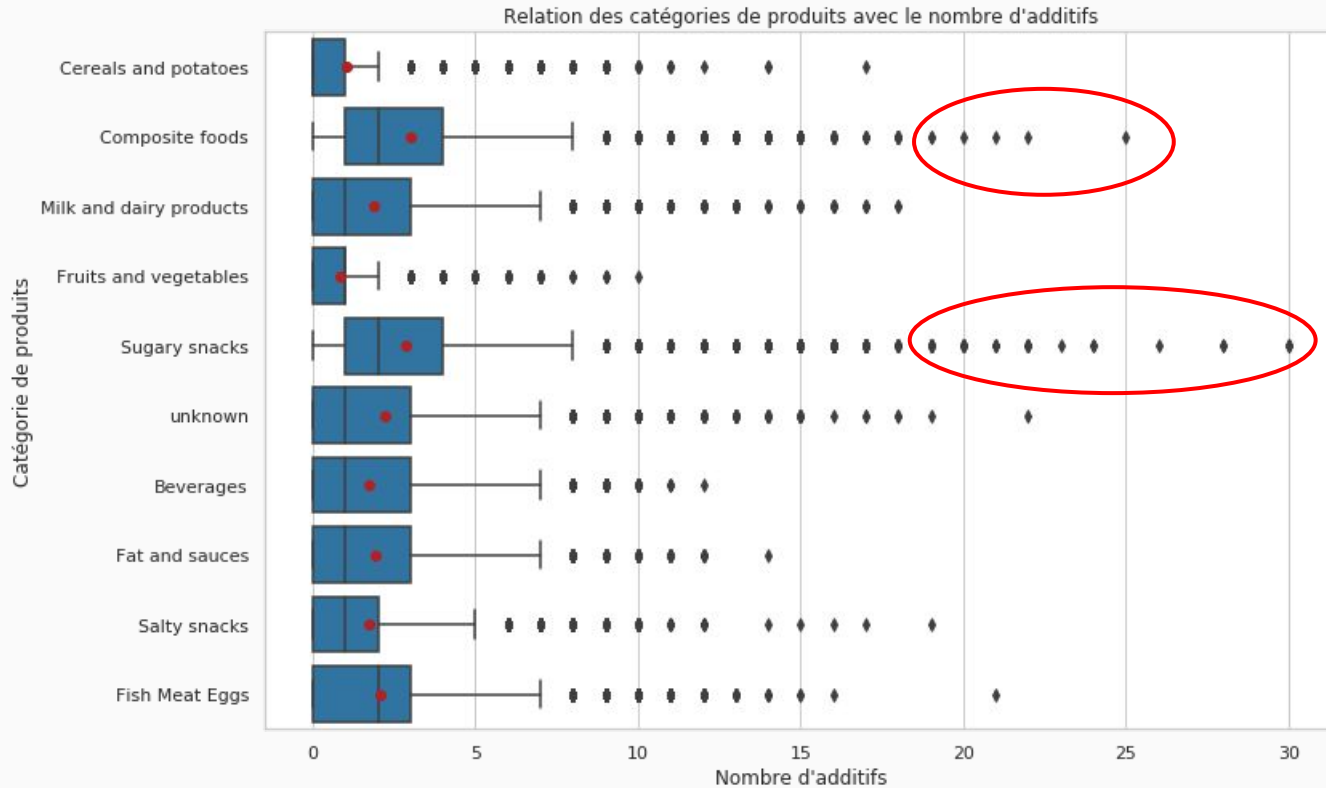
### 3. Analyse exploratoire : le nombre d'additifs dépend-il de la catégorie de produits ?

#### *Catégorie des produits*



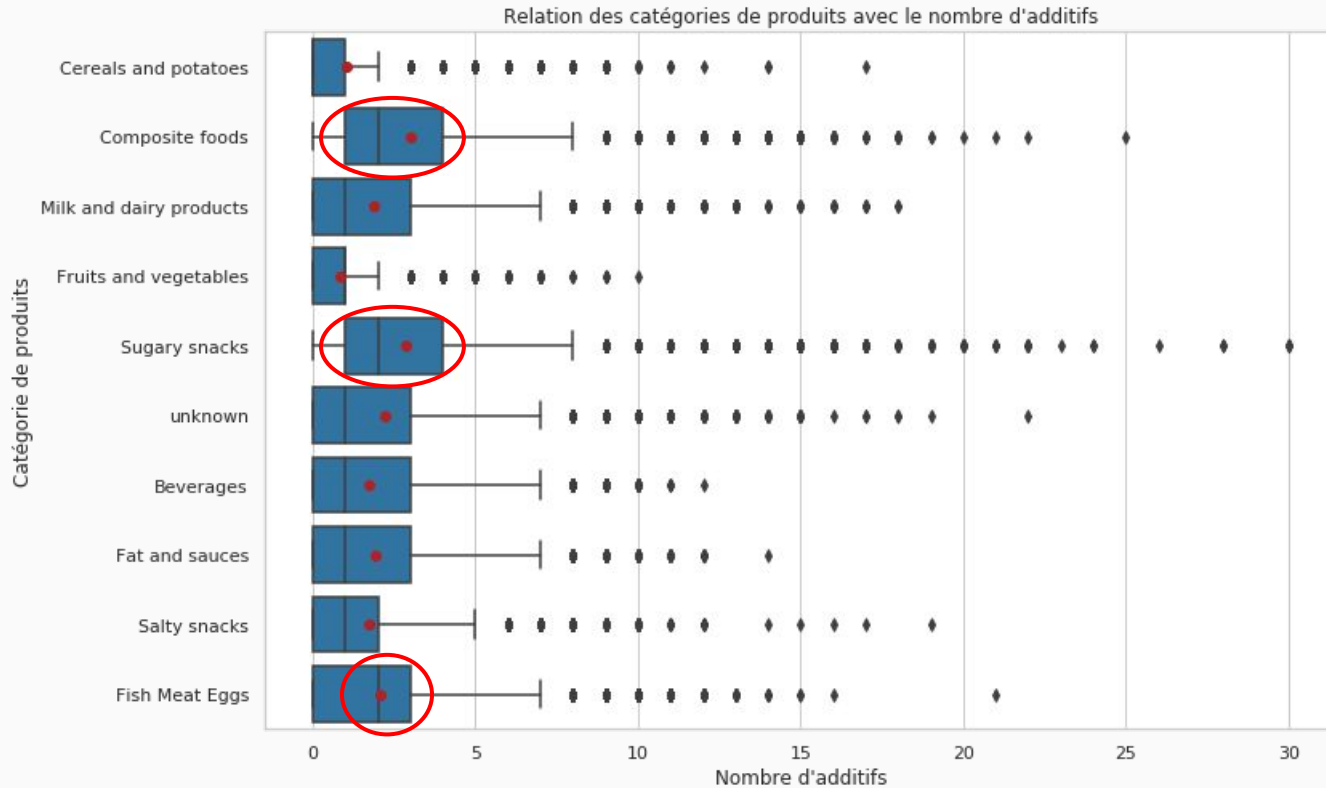
### 3. Analyse exploratoire : le nombre d'additifs dépend-il de la catégorie de produits ?

*Répartition des produits selon le nombre d'additifs et la catégorie des produits*



### 3. Analyse exploratoire : le nombre d'additifs dépend-il de la catégorie de produits ?

*Répartition des produits selon le nombre d'additifs et la catégorie des produits*



### 3. Analyse exploratoire : le nombre d'additifs dépend-il de la catégorie de produits ?

*Répartition des produits selon le nombre d'additifs et la catégorie des produits*

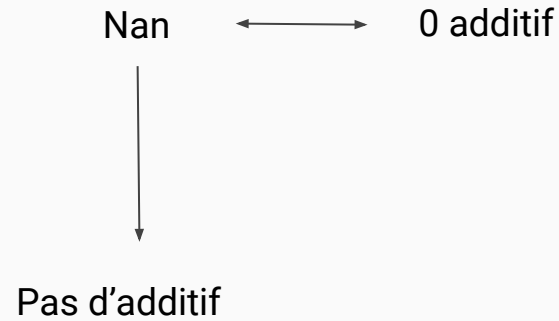
additives_n_classes	1 ou 2 additifs	Entre 3 et 7 additifs	Pas d'additif	Plus de 7 additifs
pnns_groups_1				
Beverages	2869	2021	2786	131
Cereals and potatoes	2584	1245	4822	74
Composite foods	4488	4714	2702	1194
Fat and sauces	2129	1472	1694	101
Fish Meat Eggs	5177	3694	3458	360
Fruits and vegetables	3371	339	3135	7
Milk and dairy products	3746	3358	4969	387
Salty snacks	2943	1599	2250	105
Sugary snacks	10874	8619	2882	1719
unknown	2288	1777	1791	285

Test du Khi-2 pour vérifier la dépendance :

À 5% -> rejet de  $H_0$ , p-value = 0.0

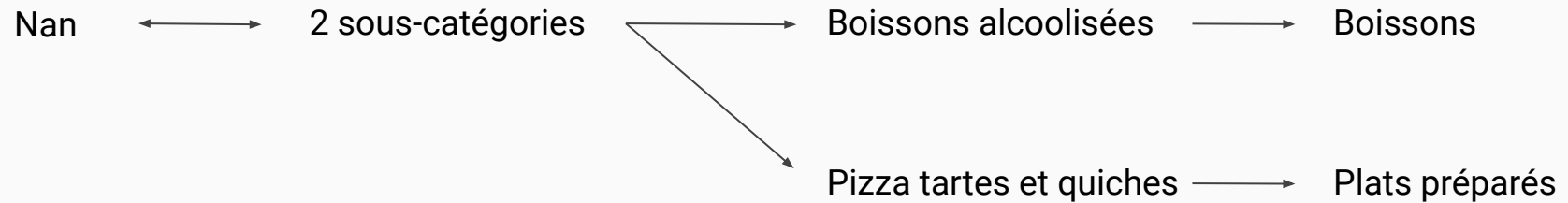
### 3. Analyse exploratoire : imputation

*Liste d'additifs*



### 3. Analyse exploratoire : imputation

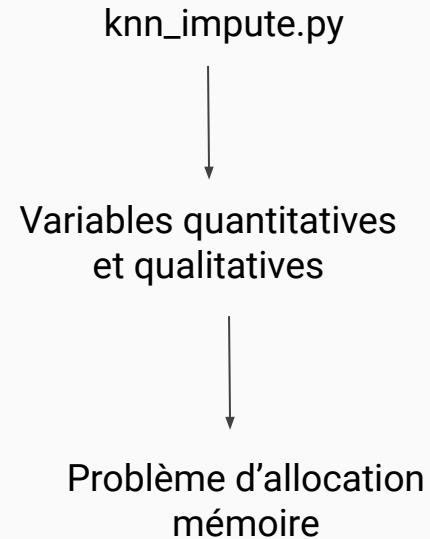
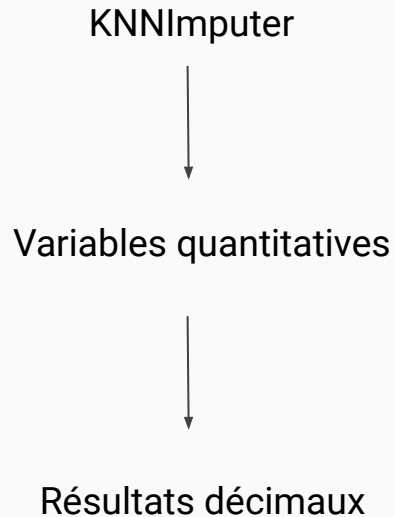
*Catégorie de produits*





### 3. Analyse exploratoire : imputation

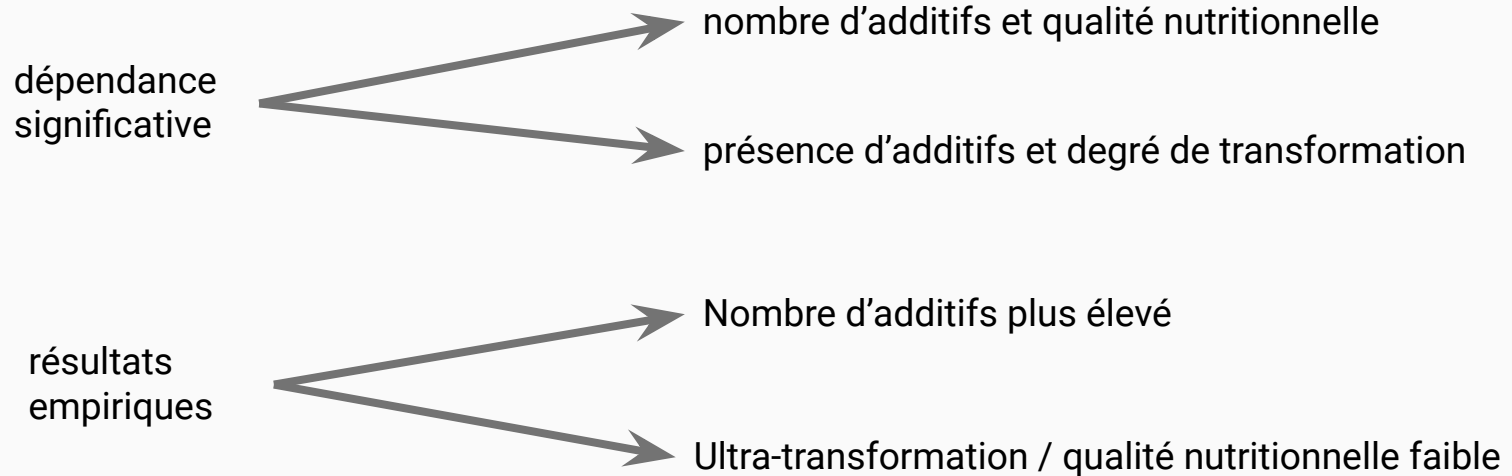
*Degré de transformation : classification NOVA*



## 4. Faits pertinents pour l'application

### *Objectif #1*

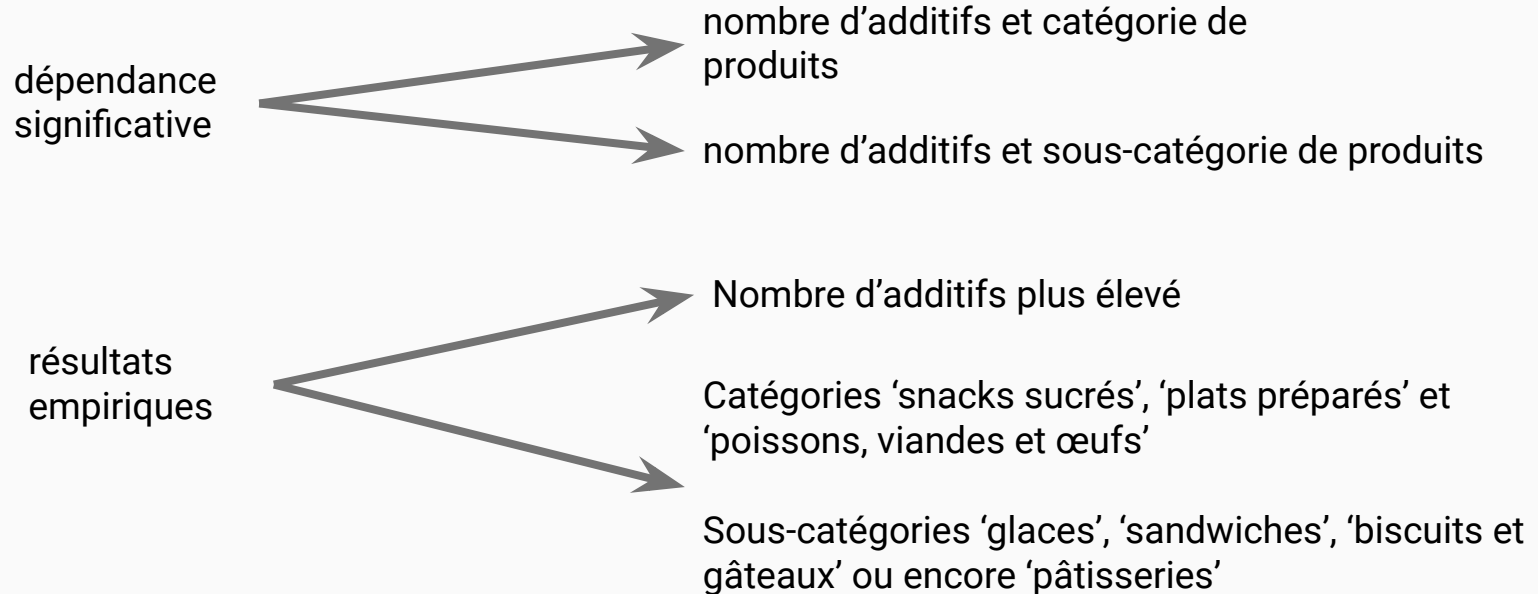
- indiquer les produits à éviter/privilegier



## 4. Faits pertinents pour l'application

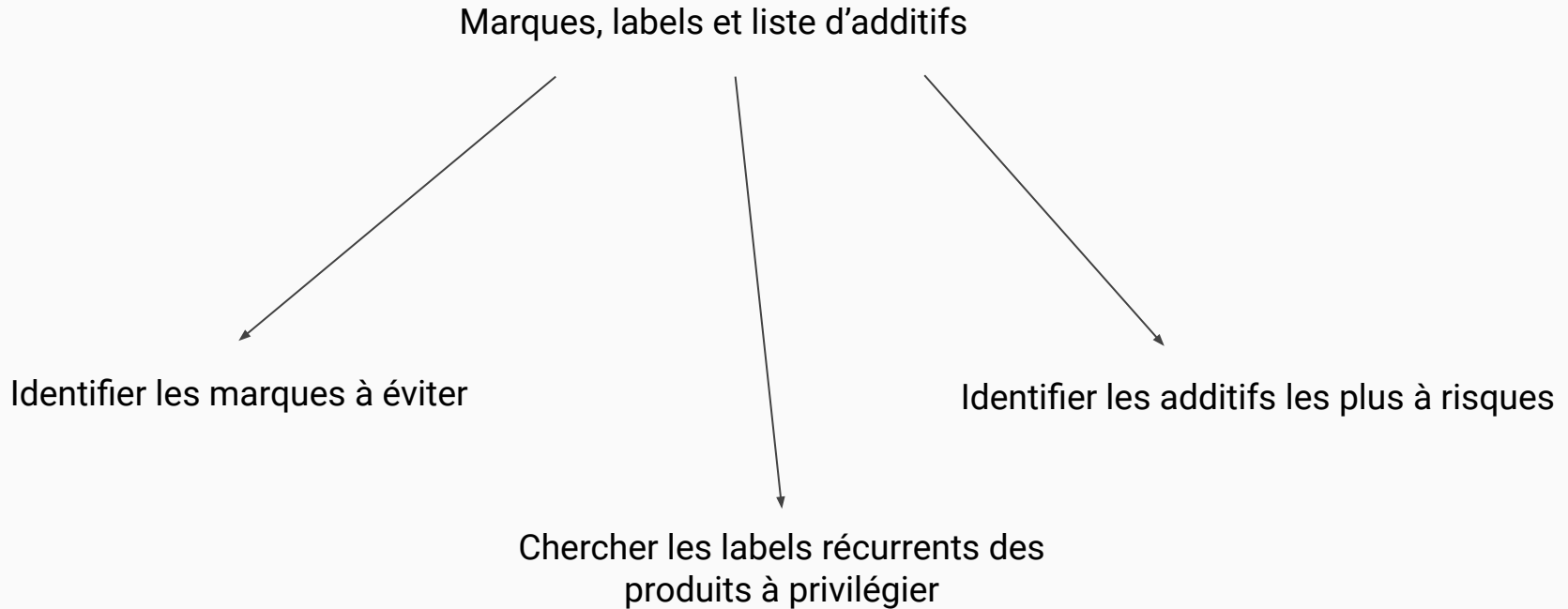
### Objectif #2

- indiquer les catégories et sous-catégories de produits à éviter/privilegier



## 4. Faits pertinents pour l'application

*Propositions d'analyses complémentaires*



# Questions/Réponses

Fin.