



# CONCEVOIR UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

Jeu de données d'Open Food Fact

<https://world.openfoodfacts.org/data/data-fields.txt>

Présentation par Hortense Monnard

# PLAN

- 1. Détecter les variables pertinentes**
- 2. Formuler des hypothèses**
- 3. Analyses univariées des variables pertinentes**
- 4. Vérifier les hypothèses à l'aide d'analyses bivariées**
- 5. Conclure sur les hypothèses avancées**
- 6. Conclure sur une idée réalisable d'application**

The background is a blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural network connections, consisting of lines and small circles.

# 1. Détecter les variables pertinentes

# Avant nettoyage

Dimension du data set : 1397452 lignes et 181 colonnes

Types de données :

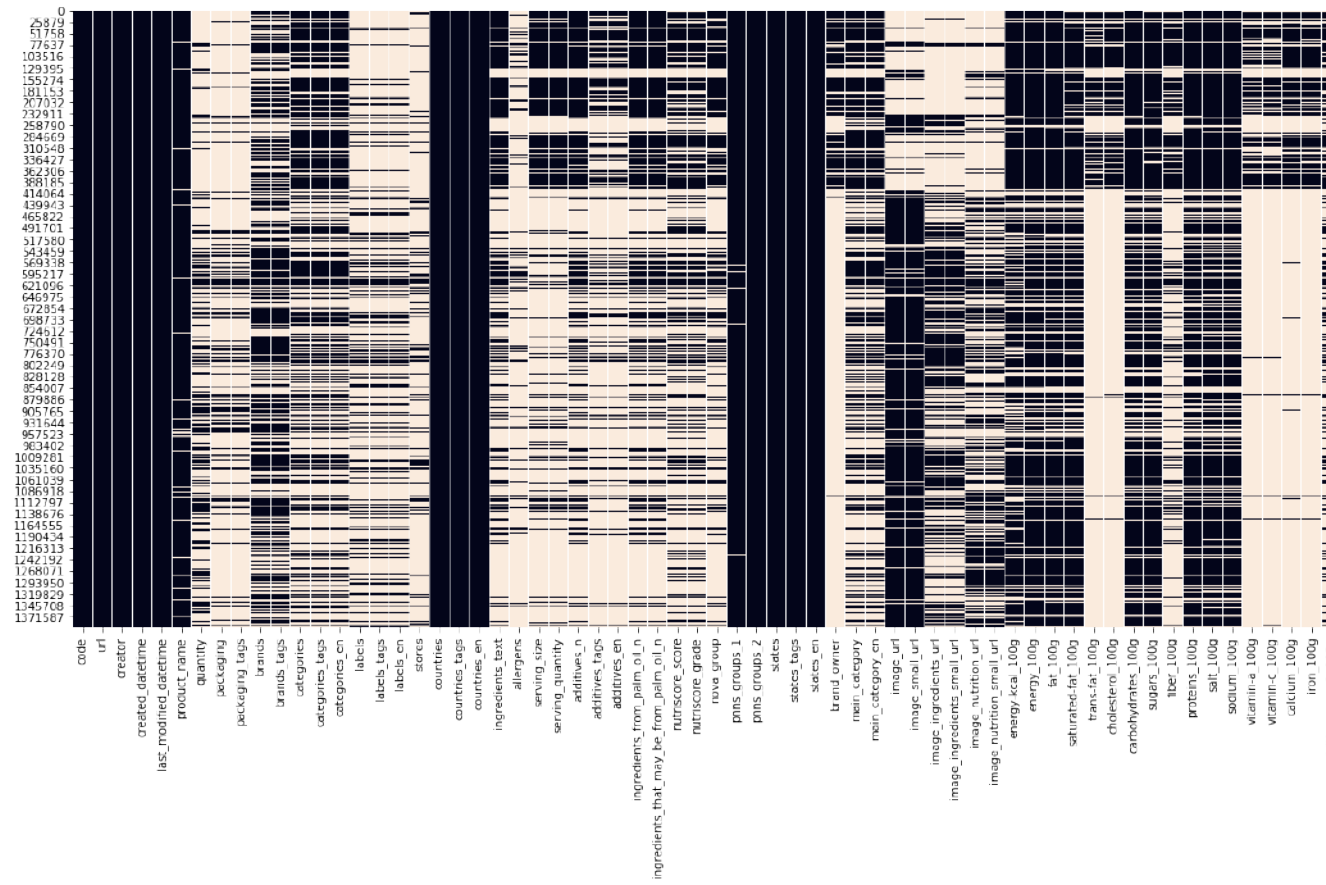


## A decorative graphic consisting of stylized circuit lines in blue and teal. The lines are of varying thickness and connect to small circles, resembling a network or data flow diagram. The design is modern and tech-oriented, positioned at the bottom of the page.



# Nettoyage

1. Supprimer les doublons, en triant à partir de date de dernière modification
2. Supprimer les lignes sans code d'identification
3. Supprimer les colonnes avec plus de 90% de données manquantes
4. Supprimer les colonnes jugées comme inintéressantes par rapport aux hypothèses



# Nettoyage

5. Déterminer et supprimer les outliers pour :

- les valeurs énergétiques dans 100g,
- les valeurs des nutriments d'intérêt dans 100g,
- les valeurs de nutriments d'intérêt par rapports aux valeurs de macronutriments dans 100g,

NB: Il y aura un nettoyage supplémentaire pour les besoins de la 2<sup>nd</sup>e hypothèse. Il sera effectué après les analyses univariées et les analyses bivariées cherchant à répondre à la problématique de la 1<sup>ère</sup> hypothèse.

# Après nettoyage

(Pour la 1<sup>ère</sup> hypothèse)

Dimension du data set : 962505 lignes et 40 colonnes

→ Conservation de 69 % des lignes du data set original





### Remplissage des colonnes du data set après nettoyage

The heatmap displays the distribution of data across various columns after cleaning. The columns are labeled on the x-axis and the rows are labeled on the y-axis. The heatmap uses a color scale from dark blue (low density) to light yellow (high density). The columns include: code, product\_name, quantity, packaging, brands, categories, labels, countries, allergens, serving\_size, serving\_quantity, additives\_n, nutriscore\_score, nutriscore\_grade, nova\_group, pnns\_groups\_1, pnns\_groups\_2, states, states\_tags, states\_en, brand\_owner, main\_category, main\_category\_en, energy-kcal\_100g, 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, and nutrition-score-fr\_100g. The rows are numbered from 1085875 to 1209071. The heatmap shows a high density of data in the first few columns (code, product\_name, quantity, packaging, brands, categories, labels, countries) and a lower density in the later columns (nutrition-score-fr\_100g).

[illegible]

# Après nettoyage

(Pour la 1<sup>ère</sup> hypothèse)

Dimension du data set : 251462 lignes et 40 colonnes

→ Conservation de 18 % des lignes du data set original



### Remplissage des colonnes du data set après nettoyage

The heatmap displays the distribution of data across various columns of a dataset after cleaning. The columns are listed on the x-axis, and the rows represent individual data points. The color scale ranges from dark blue (low values) to light yellow (high values).

Columns (X-axis):

- code
- product\_name
- quantity
- packaging
- brands
- categories
- labels
- countries
- allergens
- serving\_size
- serving\_quantity
- additives\_n
- nutriscore\_score
- nutriscore\_grade
- nova\_group
- pnns\_groups\_1
- pnns\_groups\_2
- states
- states\_tags
- states\_en
- brand\_owner
- main\_category
- main\_category\_en
- energy-kcal\_100g
- 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

Rows (Y-axis):

- 1085875
- 422127
- 84370
- 145324
- 164283
- 11015
- 118322
- 48565
- 206308
- 225794
- 235074
- 165121
- 89923
- 134638
- 95001
- 110404
- 3819
- 402778
- 371837
- 190558
- 65245
- 991672
- 60581
- 337119
- 14173
- 258558
- 215986
- 194701
- 314764
- 215480
- 325857
- 389268
- 166105
- 382955
- 4231
- 188443
- 115841
- 4978
- 121590
- 72400
- 232588
- 176813
- 241679
- 50427
- 35096
- 304520
- 379977
- 161020
- 369236
- 391144
- 59660
- 95142
- 370731
- 376575

[illegible]

The background is a blue gradient. In the corners, there are decorative white line art elements resembling circuit boards or neural networks, with lines and small circles.

## 2. Formuler des hypothèses

# Hypothèse 1

- H0 = nutriscore et nova\_group sont indépendants
- H1 = nutriscore et nova\_group sont corrélés

Si le nutriscore et le nova\_group ne sont pas corrélés, il serait intéressant de créer une application qui calcule un score prenant en compte le nutriscore\_score et le novagroup afin d'avoir une vision globale de la qualité des produits.

En effet, un produit peut être très transformé mais conserver une valeur nutritionnelle correcte et vice versa.

Proposition d'application liée :

Application qui calcule un score globale avec le nutriscore\_score et le nova\_group.

Colonnes d'intérêt :

'nutrition-score-fr\_100g', 'nutriscore\_score', 'nutriscore\_grade', 'nova\_group'.

## Hypothèse 2

- H0 = rapport fat sat+trans /fat total et nutriscore sont indépendants
- H1 = rapport fat sat+trans /fat total et nutriscore sont corrélés

Si rapport (saturated-fat + trans-fat )/fat total n'est pas corrélé avec le nutriscore, il serait intéressant de rajouter ce calcul.

Le nutriscore prend en compte les sucres et les graisses.


Cependant, le ratio (saturated-fat + trans-fat )/fat total a été observé comme fortement corrélés à l'apparition de maladies cardiovasculaires.

Proposition d'application liée :

Application qui calcule le ratio d'AG saturés ou avec des insaturations de type trans- sur le poids total en lipides

Colonnes d'intérêt :

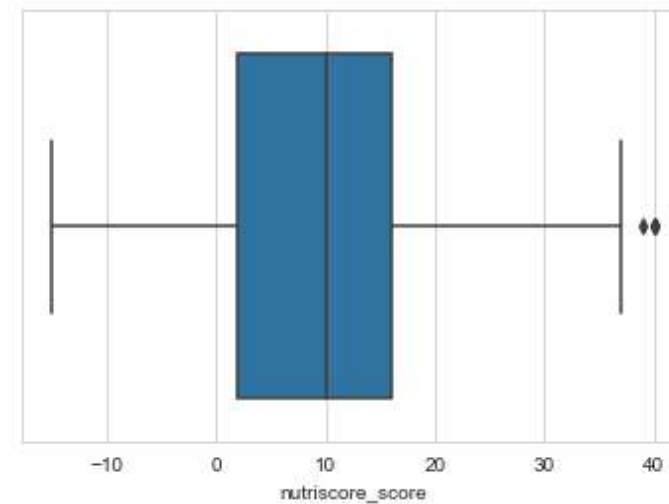
'saturated-fat_100g',	'trans-fat_100g',	'fat_100g',	'nutriscore_score',
'nutriscore_grade'			

The background is a blue gradient. In the corners, there are decorative white line art elements resembling circuit boards or neural network connections, with lines and small circles.

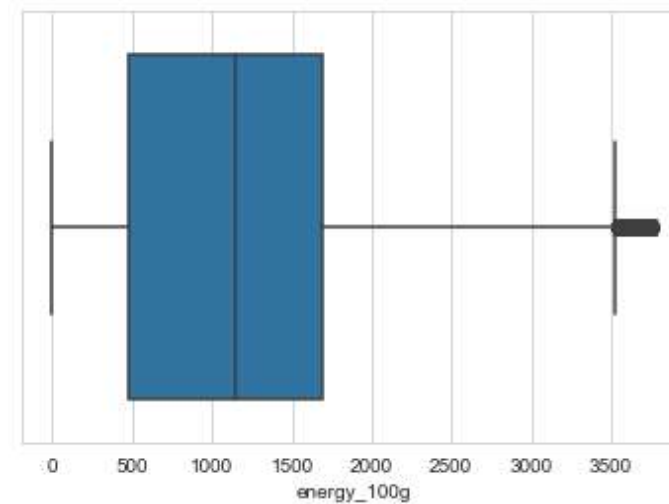
### **3. Analyses univariées de ces variables pertinentes**

# Données Quantitatives

- Nutriscore

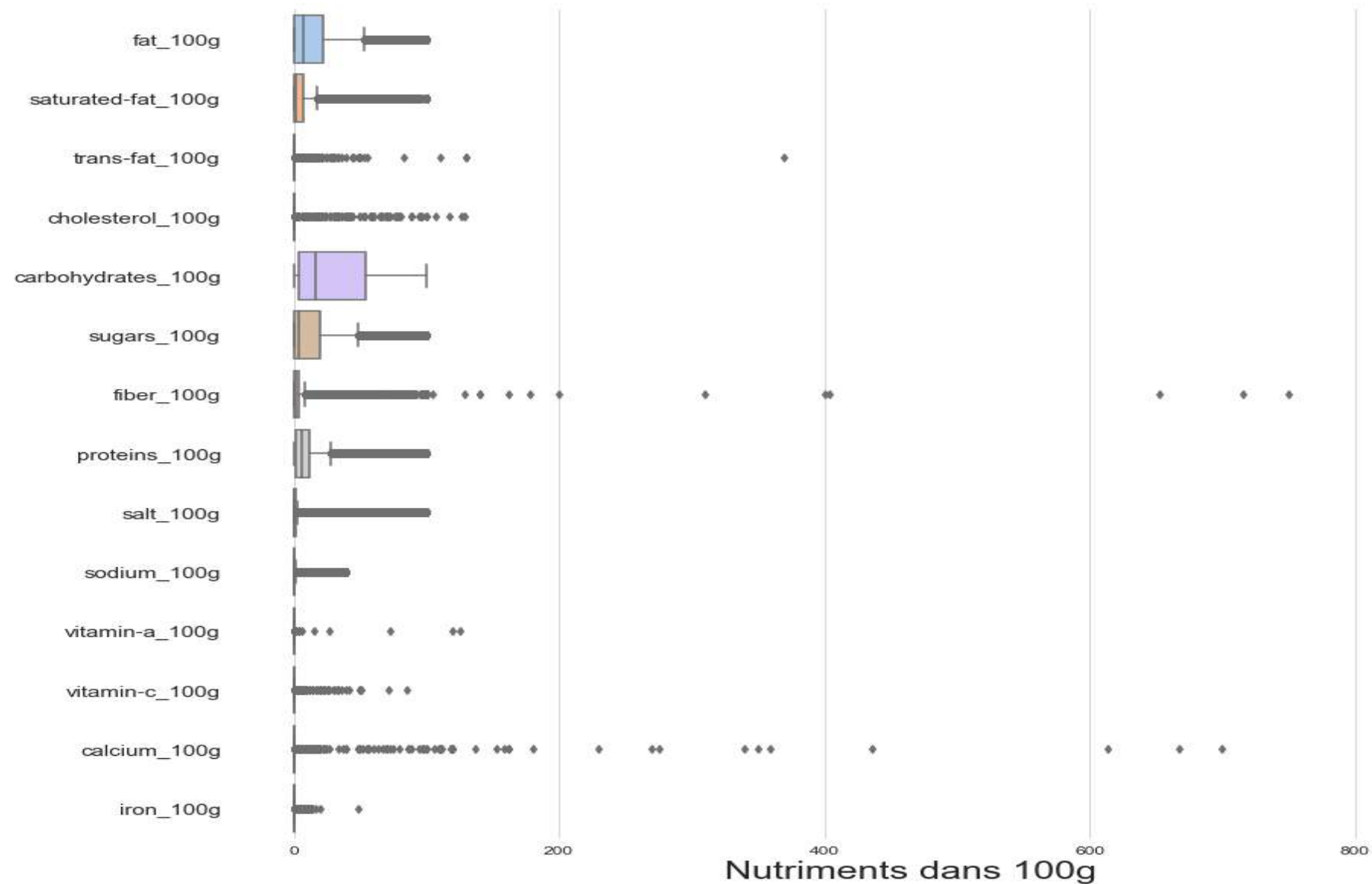


- Energie (en Joules)





# Données Quantitatives



# Données Qualitatives

Classifications selon 2 indices :

- Nova Group
- Nutriscore Grade

nova_group	n	f	F
1	38778	0.040289	0.040289
2	5493	0.005707	0.045996
3	88007	0.091435	0.137431
4	276355	0.287121	0.424552

nutriscore_grade	n	f	F
a	81401	0.084572	0.084572
b	65175	0.067714	0.152286
c	107209	0.111385	0.263671
d	160775	0.167038	0.430709
e	97689	0.101495	0.532204

# Données Qualitatives

Classifications par le

Programme National Nutrition Santé :

- PNNS 1

- PNNS 2

pnns_groups_1	n	f	F
Beverages	33220	0.034514	0.034514
Cereals and potatoes	53810	0.055906	0.090420
Composite foods	36506	0.037928	0.128348
Fat and sauces	41795	0.043423	0.171772
Fish Meat Eggs	53073	0.055140	0.226912
Fruits and vegetables	31212	0.032428	0.259340
Milk and dairy products	58578	0.060860	0.320200
Salty snacks	18052	0.018755	0.338955
Sugary snacks	87837	0.091259	0.430214
cereals-and-potatoes	390	0.000041	0.430254
fruits-and-vegetables	5397	0.005607	0.435862
salty-snacks	40	0.000004	0.435866
sugary-snacks	3071	0.003191	0.439056
unknown	538529	0.559508	0.998564

pnns_groups_2	n	f	F
Alcoholic beverages	10010	0.001040	0.001040
Appetizers	72750	0.007558	0.008598
Artificially sweetened beverages	37890	0.003937	0.012535
Biscuits and cakes	451150	0.046872	0.059407
Bread	149810	0.015565	0.074972
Breakfast cereals	50370	0.005233	0.080205
Cereals	266100	0.027647	0.107852
Cheese	292050	0.030343	0.138195
Chocolate products	88650	0.009210	0.147405
Dairy desserts	33820	0.003514	0.150919
Dressings and sauces	317270	0.032963	0.183882
Dried fruits	35350	0.003673	0.187554
Eggs	12730	0.001323	0.188877
Fats	100680	0.010460	0.199337
Fish and seafood	156130	0.016221	0.215558
Fruit juices	50040	0.005199	0.220757
Fruit nectars	8190	0.000851	0.221608
Fruits	134620	0.013986	0.235595
Ice cream	40640	0.004222	0.239817
Legumes	56210	0.005840	0.245657
Meat	124710	0.012957	0.258614
Milk and yogurt	219270	0.022781	0.281395
Nuts	43190	0.004487	0.285882
Offals	3850	0.000400	0.286282
One-dish meals	287480	0.029868	0.316150
Pizza pies and quiche	52460	0.005450	0.321600
Pizza pies and quiches	3810	0.000396	0.321996
Plant-based milk substitutes	41280	0.004289	0.326285
Potatoes	15610	0.001622	0.327907
Processed meat	233310	0.024240	0.352147
Salty and fatty products	64580	0.006710	0.358856
Sandwiches	25120	0.002610	0.361466
Soups	16820	0.001748	0.363214
Sweetened beverages	104890	0.010898	0.374111
Sweets	338570	0.035176	0.409287
Teas and herbal teas and coffees	8180	0.000850	0.410137
Unsweetened beverages	69250	0.007195	0.417332
Vegetables	125330	0.013021	0.430353
Waters and flavored waters	12480	0.001297	0.431650
cereals	300	0.000031	0.431681
fruits	1740	0.000181	0.431862
legumes	90	0.000009	0.431871
nuts	40	0.000004	0.431875
pastries	30710	0.003191	0.435066
unknown	5385290	0.559508	0.994574
vegetables	52230	0.005426	1.000000

The background is a blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural networks, with lines and small circles representing nodes.

## 4. Vérifier les hypothèses à l'aide d'analyses bivariées

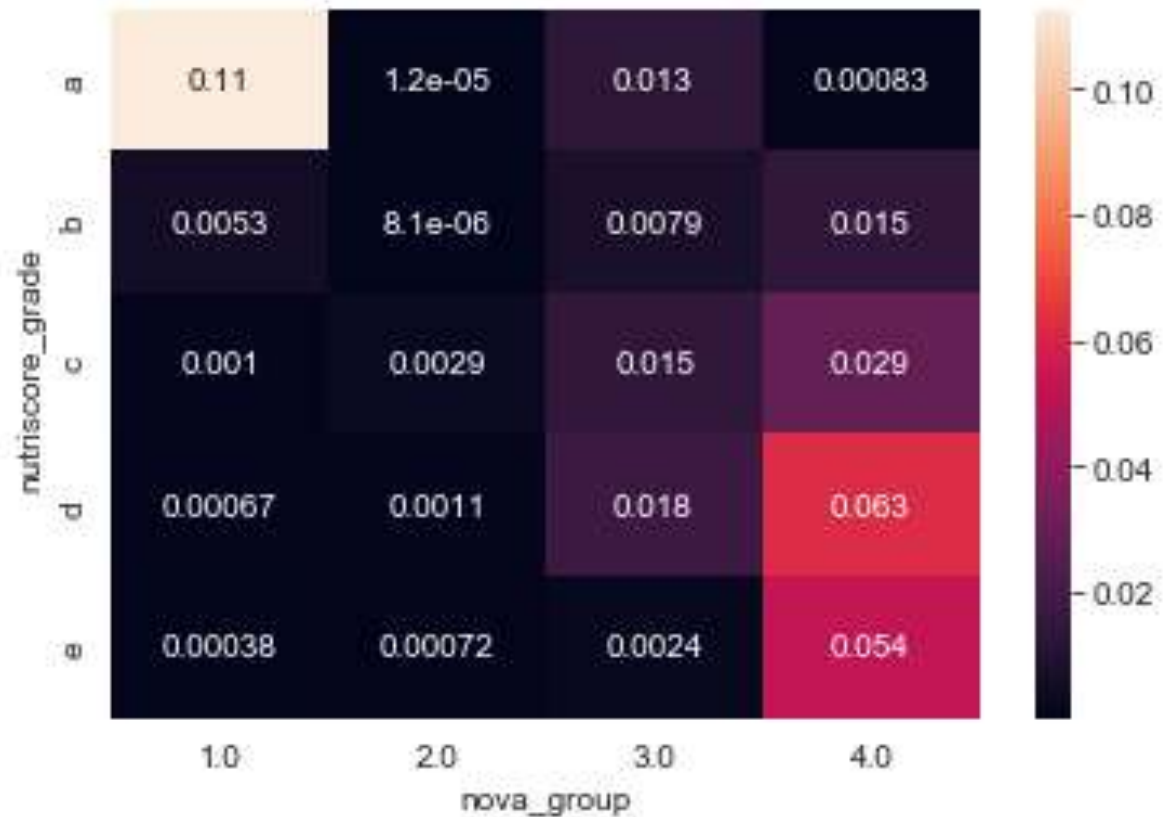
# Hypothèse 1

- $H_0$  = nutriscore\_grade et nova\_group sont indépendants
- $H_1$  = nutriscore\_grade et nova\_group sont corrélés

p-value < 0.001  
 $H_0$  est rejetée

Dans le tableau de contingence,  
hypothèse acceptée pour valeur < 0,08

Il existe une corrélation pour  
certains nutriscore\_grade et  
nova\_group mais pas pour tous.



# Hypothèse 1

- H0 = nutriscore\_grade et nova\_group sont indépendants
- H1 = nutriscore\_grade et nova\_group sont corrélés

Tableau des fréquences attendues

	nova_group 1	nova_group 2	nova_group 3	nova_group 4	Total
nutriscore grade a	6064.64933151	726.77878948	13926.19601136	37099.37586764	57817.
nutriscore grade b	4655.08540193	557.85869032	10689.42789817	28476.62800958	44379.
nutriscore grade c	7783.54530667	932.76879282	17873.28032979	47614.40557072	74204.
nutriscore grade d	11814.72192357	1415.85915826	27130.0324773	72274.38644087	112635.
nutriscore grade e	7340.99803632	879.7345691	16857.06328338	44907.20411119	69985.
Total	37659.	4513.	86476.	230372.	359020.

Chi2 = 79549.38362238492

Degrés de liberté = 20

p-value < 0.001

H0 est rejetée

# Hypothèse 1

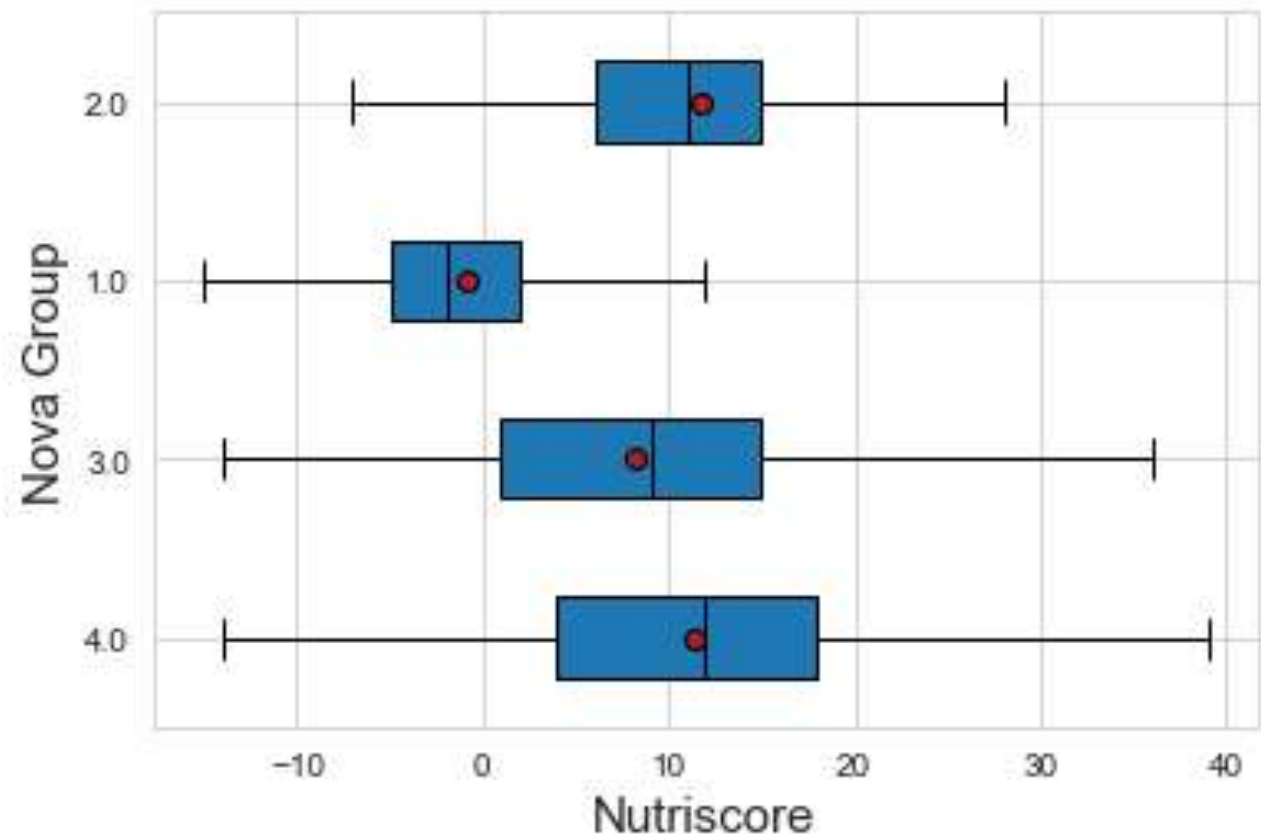
- $H_0$  = nutriscore\_score et nova\_group sont indépendants
- $H_1$  = nutriscore\_score et nova\_group sont corrélés

p-value < 0.001  
 $H_0$  est rejetée

Il existe une corrélation entre le nutriscore\_score et le nova\_group.

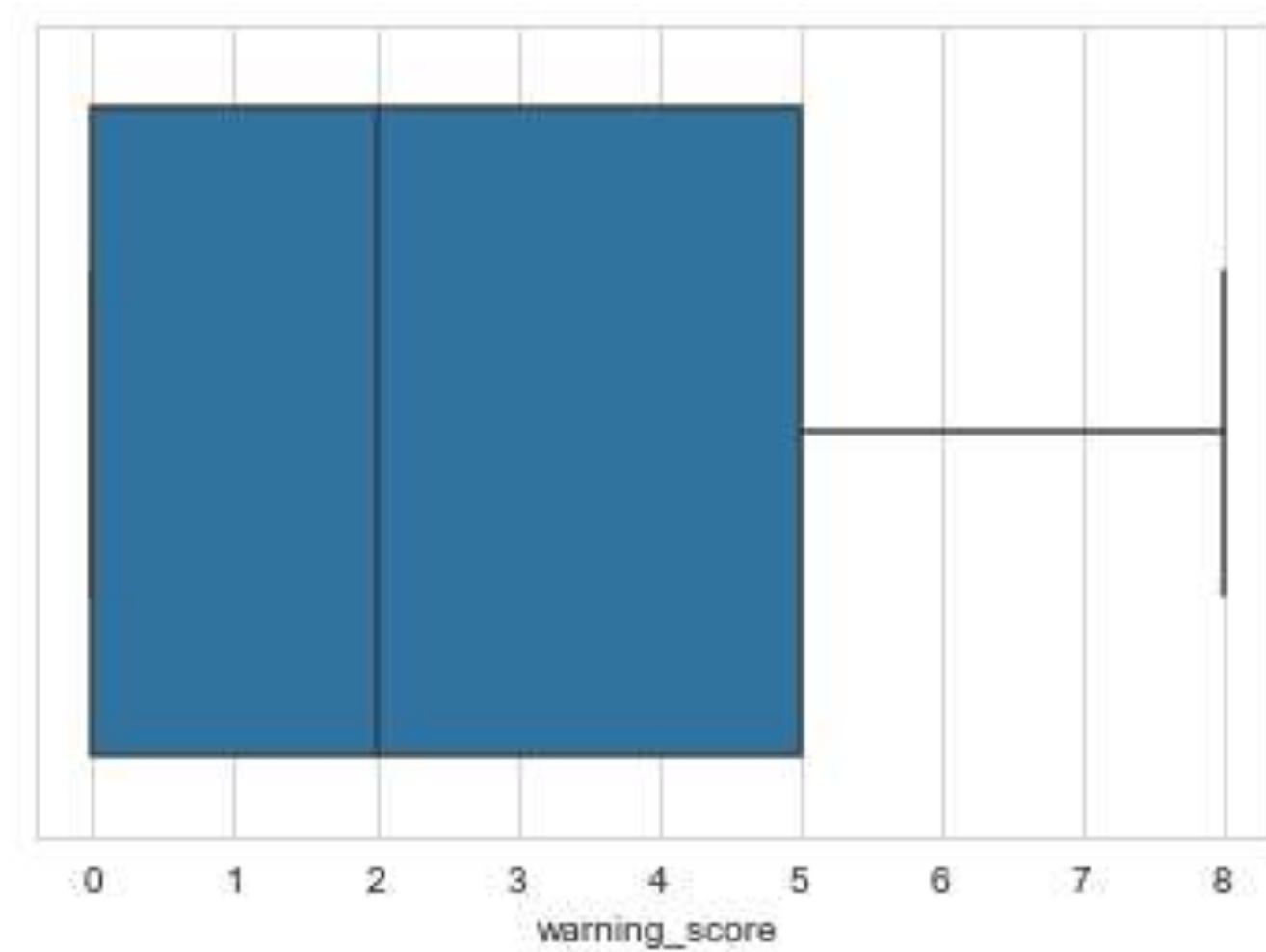
Effet de taille :  
 $\eta^2 = 0.17420136155567237$

17% de la variance due à la variable indépendante



# Hypothèse 1

Création d'un warning score entre 1 et 8





# Hypothèse 1

Observation du comportement du warning score pour les catégories du Programme National Nutrition Santé : PNN1 et PNN2

1<sup>er</sup> test :

- $H_0$  = warning score et PNN1 sont indépendants
- $H_1$  = warning score et PNN1 sont corrélés

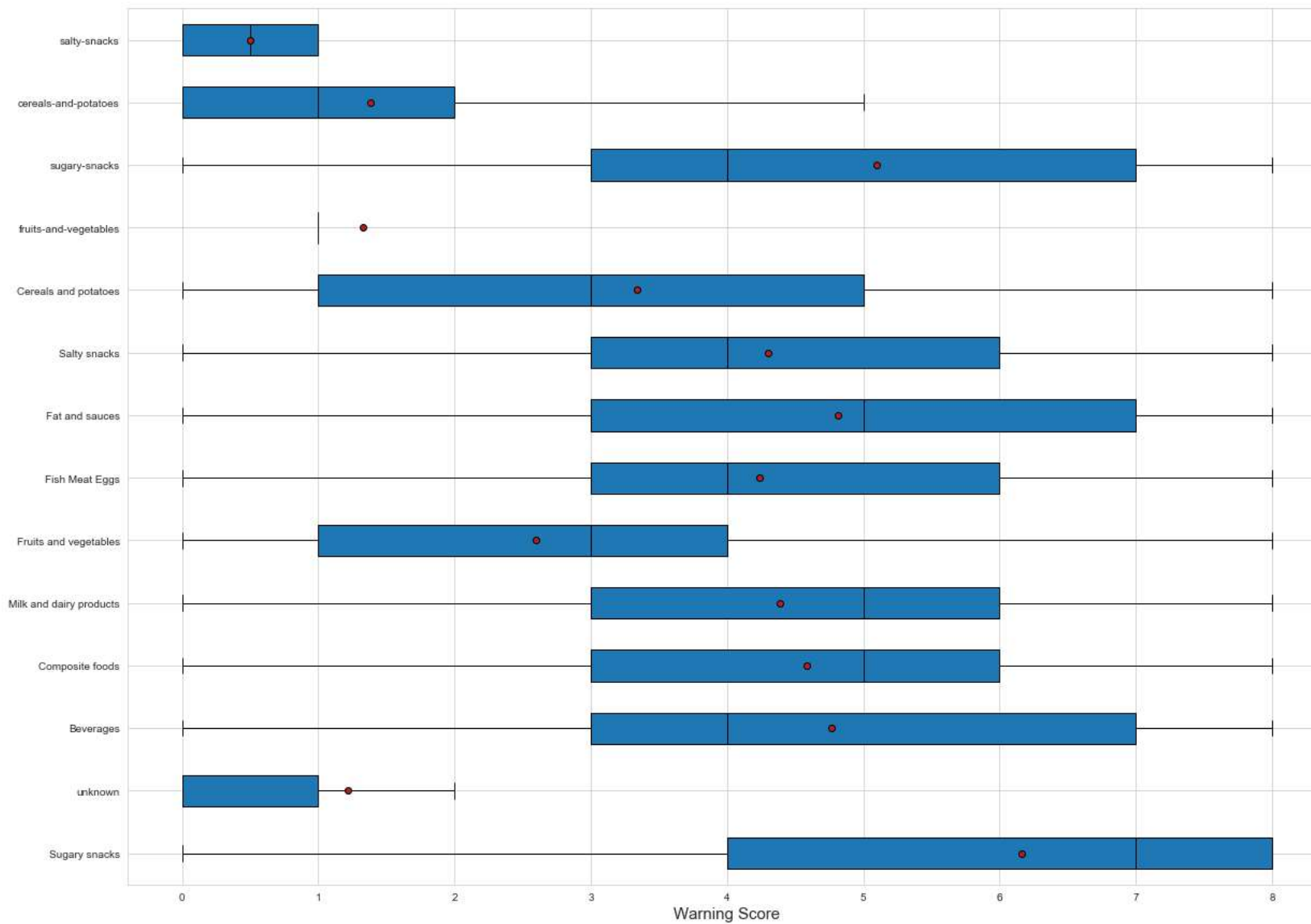
2<sup>nd</sup> test :

- $H_0$  = warning score et PNN2 sont indépendants
- $H_1$  = warning score et PNN2 sont corrélés

PNN1

p-value < 0.001  
H0 est rejetée

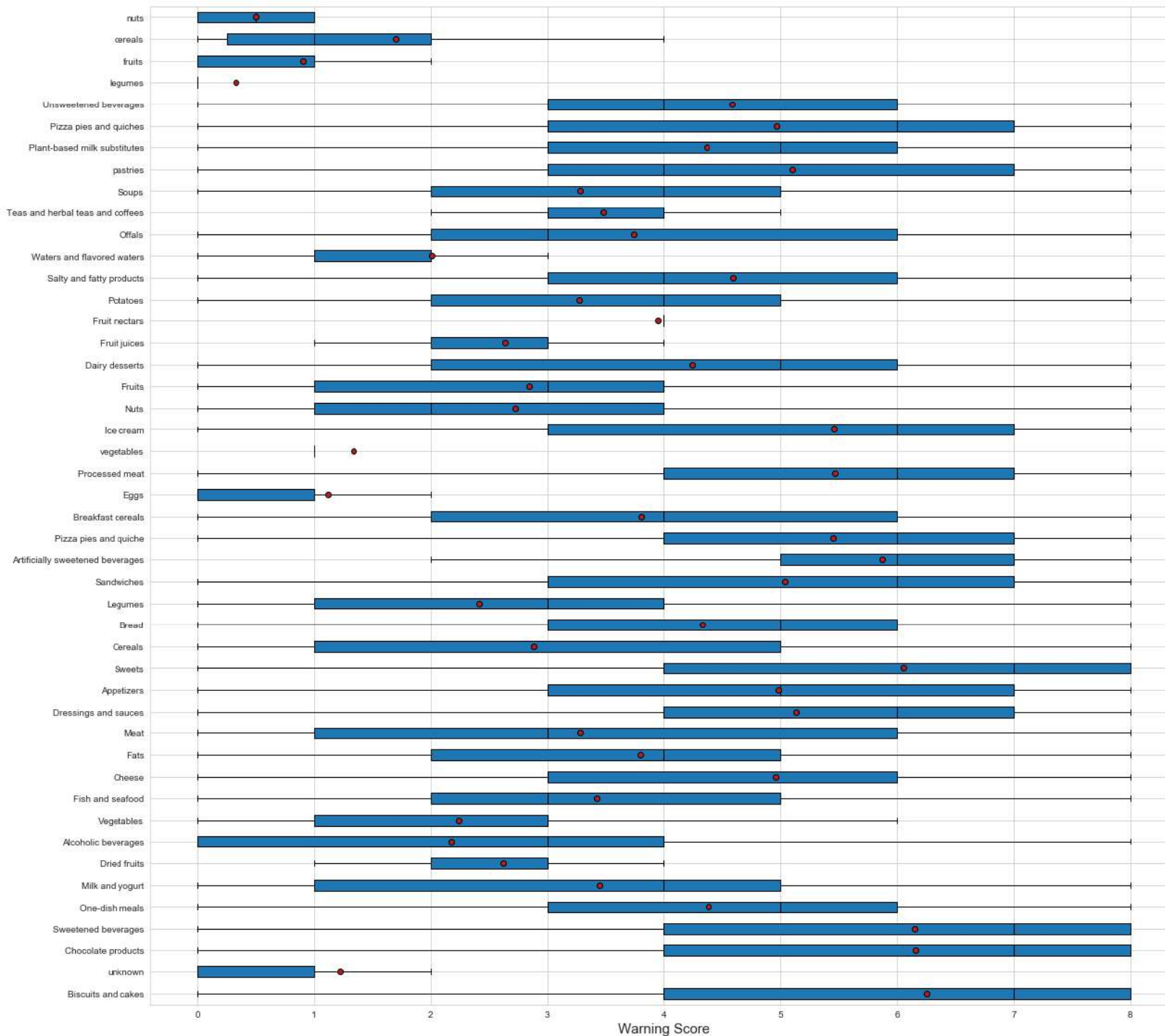
Effet de taille :  
 $\eta^2 =$   
0.3948376913364021



PNN2

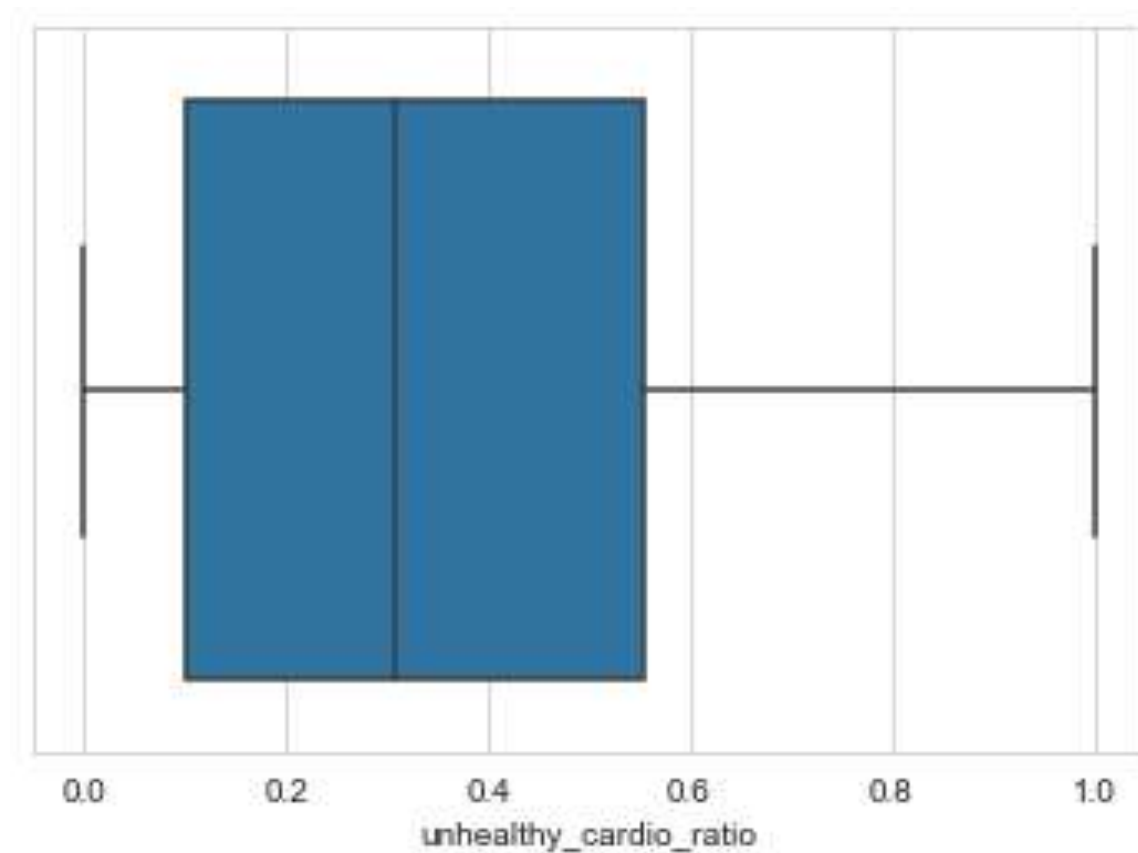
p-value < 0.001  
H0 est rejetée

Effet de taille :  
 $\eta^2 =$   
0.423719392457668  
43



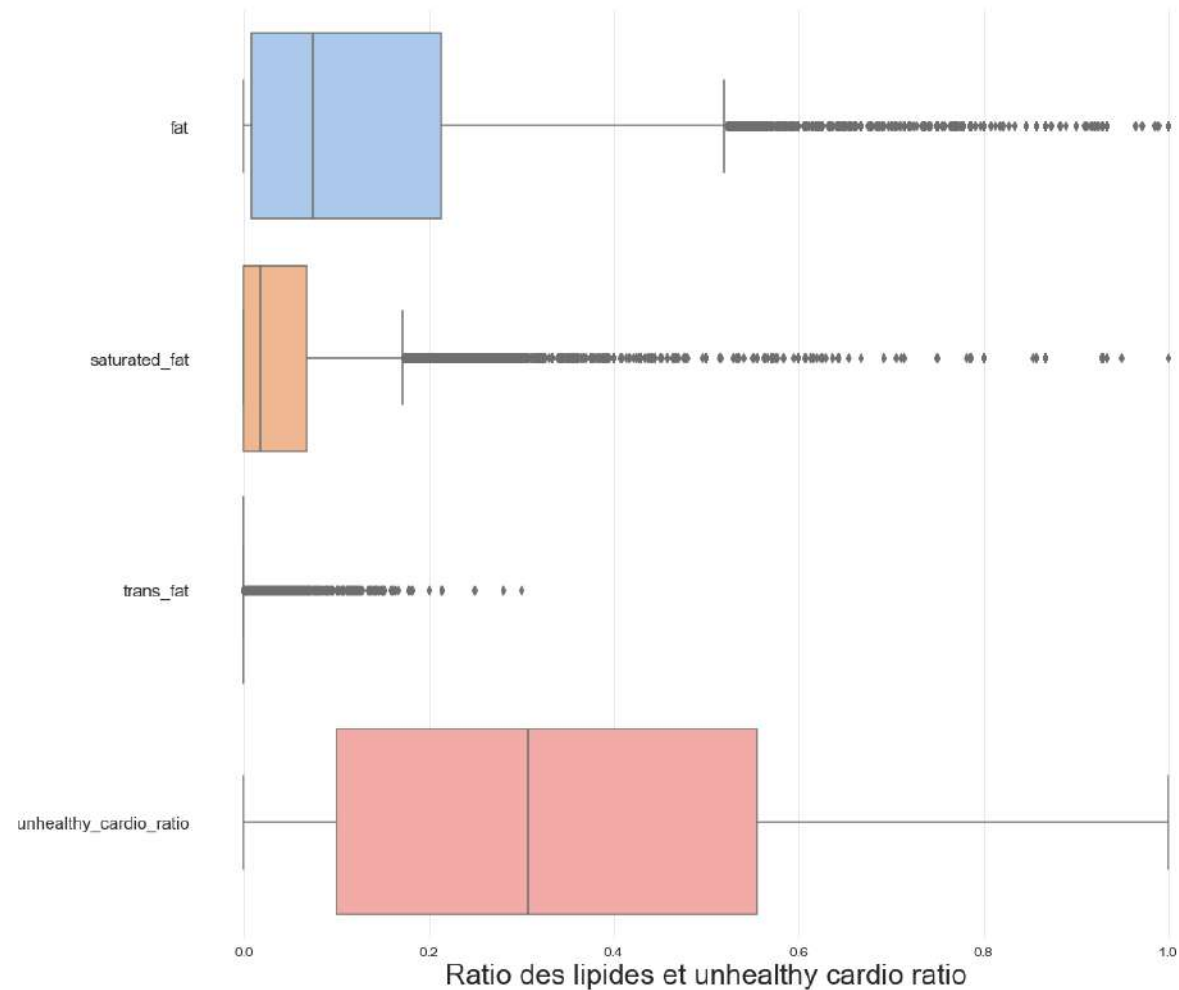
# Hypothèse 2

Création d'un Unhealthy Cardio Ratio



# Hypothèse 2

Ratio dans 100g des variables prises en compte dans la création du Unhealthy Cardio Ratio



## Hypothèse 2

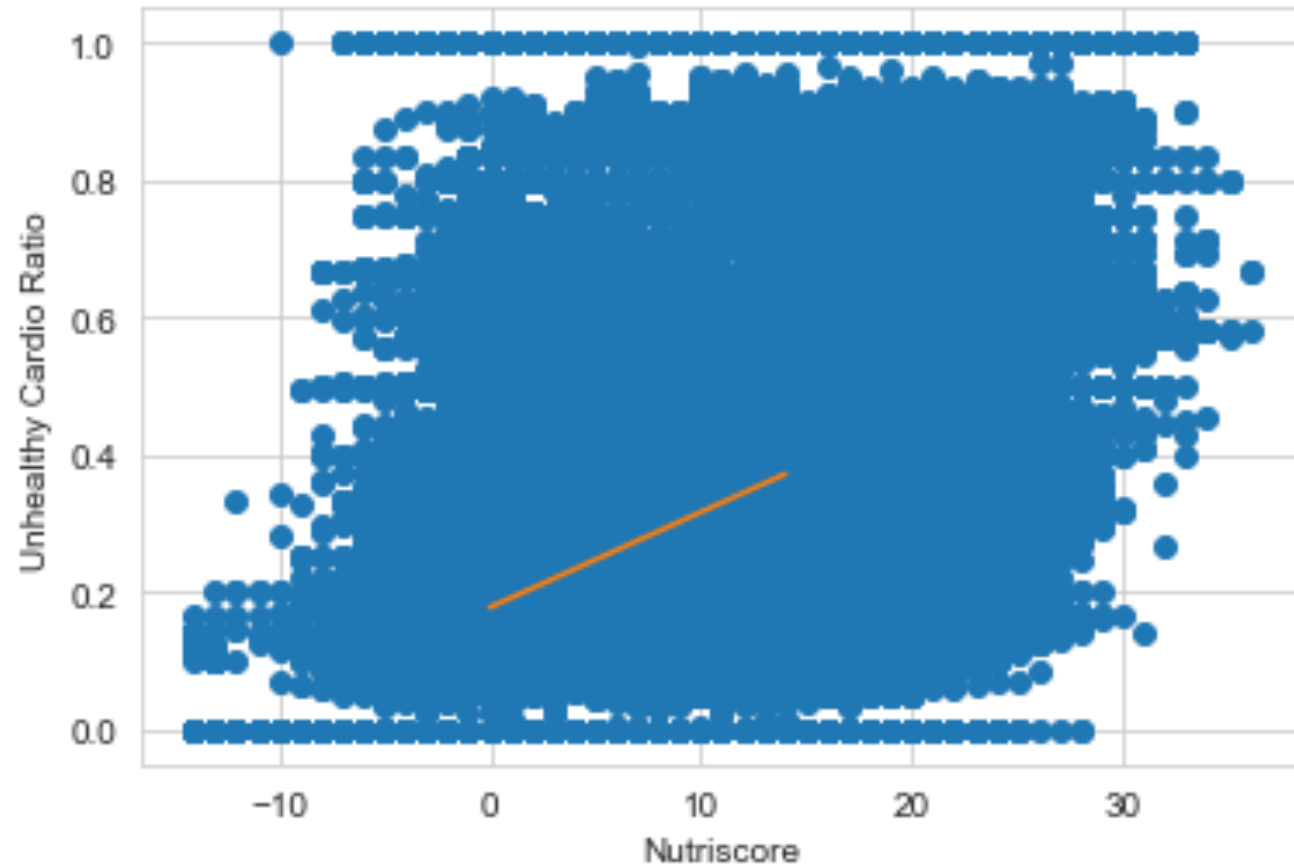
- $H_0$  = le unhealthy\_cardio\_ratio et le nutriscore\_score sont indépendants
- $H_1$  = le unhealthy\_cardio\_ratio et le nutriscore\_score sont corrélés

Coefficient de Pearson =  
0.4837499174887893

Covariance Empirique =  
1.1016458465038865

$H_0$  est acceptée.

Le unhealthy\_cardio\_ratio  
et le nutriscore\_score  
sont indépendants



## Hypothèse 2

- $H_0$  = le unhealthy\_cardio\_ratio et le nutriscore\_grade sont indépendants
- $H_1$  = le unhealthy\_cardio\_ratio et le nutriscore\_grade sont corrélés

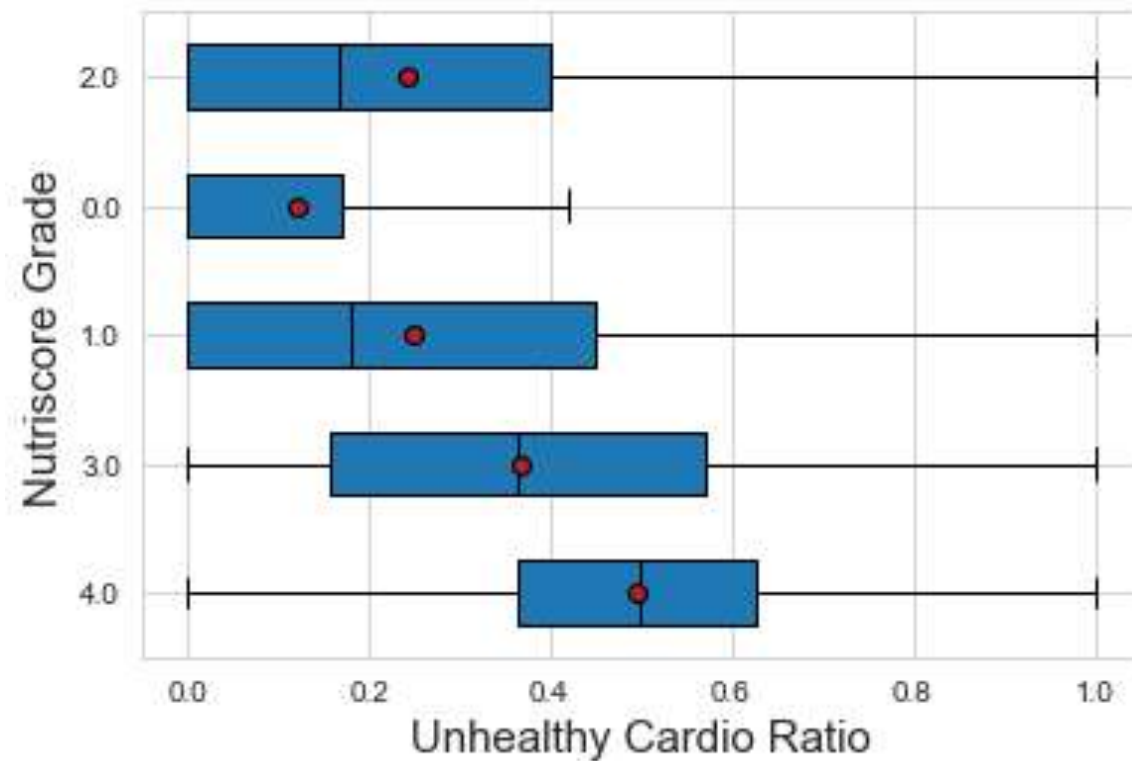
p-value < 0.001

$H_0$  est rejetée

Effet de taille :

$\eta^2 = 0.22254811166850344$

22% de la variance due à la variable indépendante



The background is a blue gradient. In the corners, there are white line-art decorations resembling electronic circuit boards, with lines and small circles representing components.

## **5. Conclure sur les hypothèses avancées**



# Hypothèse 1

- $H_0$  = le nutriscore\_grade et le nova\_group sont indépendants
- $H_1$  = le nutriscore\_grade et le nova\_group sont corrélés

Il existe une corrélation entre le nutriscore\_grade et le nova\_group pour certains grades ou groupes.

→ Information warning\_score peut être intéressante mais pas indispensable.

## Hypothèse 2

- $H_0$  = le `unhealthy_cardio_ratio` et le `nutriscore_score` sont indépendants
- $H_1$  = le `unhealthy_cardio_ratio` et le `nutriscore_score` sont corrélés

Il n'existe pas de corrélation entre ces paramètres

→ Information `unhealthy cardio ratio` est alors particulièrement intéressante à rajouter

The background is a blue gradient. In the corners, there are white line-art illustrations of circuit boards or neural networks, with lines and small circles representing components.

## **6. Conclure sur une idée réalisable d'application**

# Hypothèse 1

→ Création d'un 'warning\_score'

Intéressant mais n'apporte pas beaucoup plus au nova\_group et au nutriscore

Cependant, il pourrait améliorer la lisibilité avec un score de 0 à 8 et un logo allant du blanc au rouge.



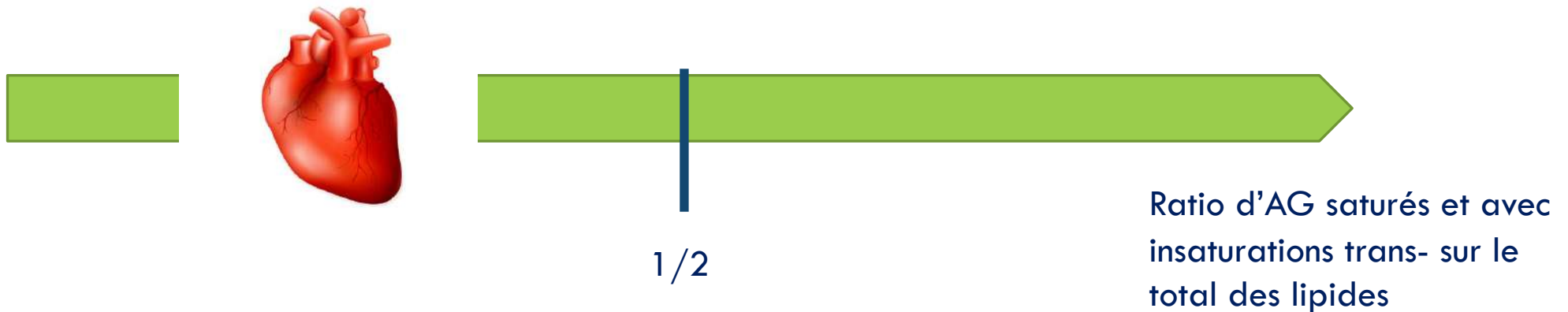
## Hypothèse 2

→ Création d'un 'unhealthy\_cardio\_ratio'

Bouton en forme de cœur qui se déplace sur une échelle

Il devient bleu lorsque le ratio d'acides gras saturés ou avec des insaturations de type trans- est supérieure à  $1/2$

Il devient rouge lorsque le ratio d'acides gras saturés ou avec des insaturations de type trans- est inférieure à  $1/2$



**MERCI !**