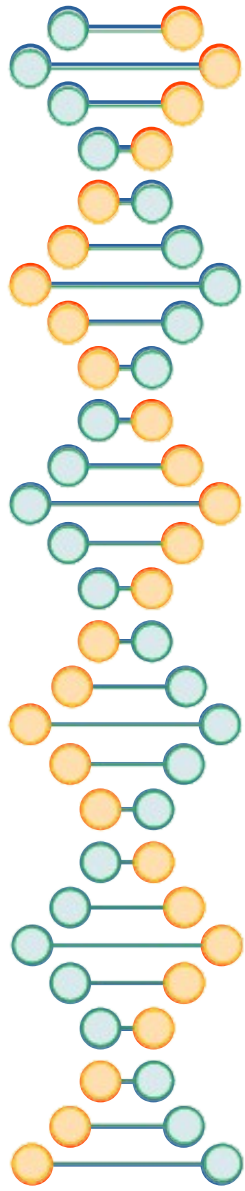


# Manger... Plus SAIN et plus ÉCOLOGIQUE

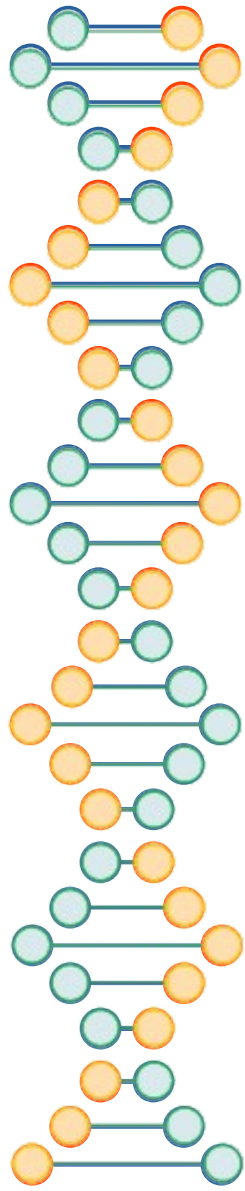
Projet 2  
Sofia Velasco





# Nettoyage des données

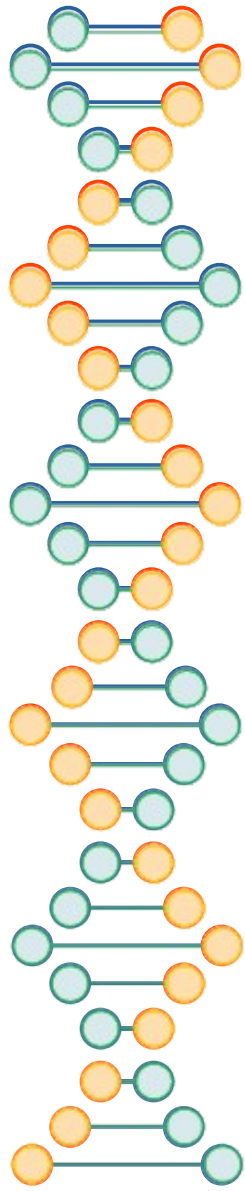




## A. Caractéristiques générales du jeu de données

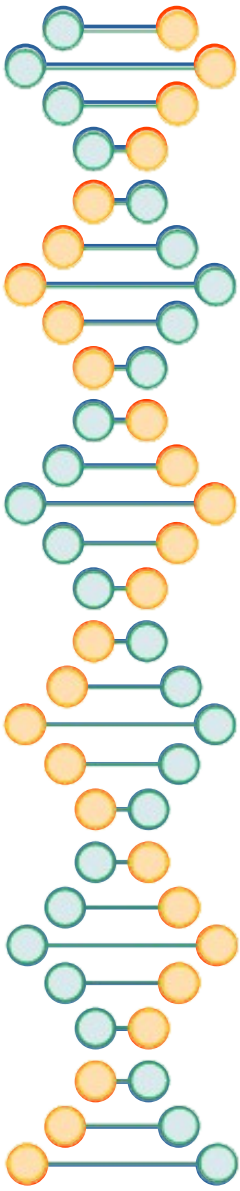
- **2 168 141** lignes et **191** Colonnes.
- Énormément de données manquantes: seul la colonne '**code**' a toute ses lignes renseignées → elle va servir pour **identifier les produits**.





## B. Choix des variables et de la stratégie





## B.1 L'idée

### Un double objectif :

- Connaître la **qualité alimentaire** et l'**impact écologique** de nos produits.
- Proposer, si pertinent, **trois options similaires**, à **meilleurs impacts écologique et alimentaire**.



Ça concerne uniquement les produits vendus en FRANCE.



## B.2 (Compréhension de l'objectif)

On a besoin de variables pour :

- Identifier les produits.
- Identifier des produits similaires.
- Restreindre le périmètre aux produits vendus en France.
- Identifier les bénéfices nutritionnels.
- Identifier l'impact écologique.



## B.3 Compréhension de certaines variables

- 'categories', 'categories\_tags', 'categories\_en', 'main\_category', 'main\_category\_en' :
  - ↳ '**main\_category\_en**' → elle comprend la catégorie principale et n'a presque pas des "fr:" ni "en:".
- 'countries', 'countries\_tags', 'countries\_en':
  - ↳ '**countries\_en**' → car chaque pays a un seul identifiant.



## B.4 Choix final

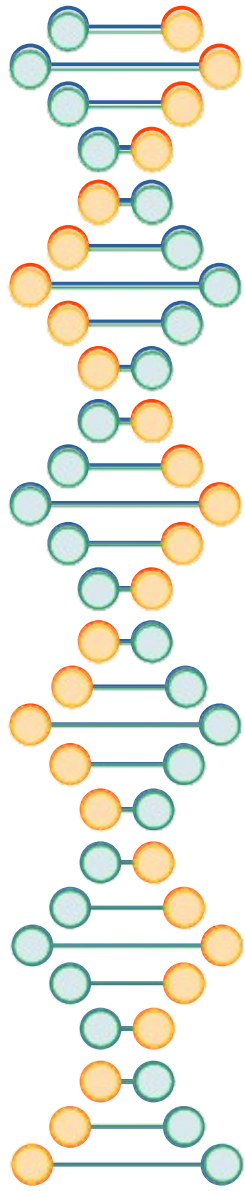
- Identifier les **produits** → {  
  '`code`' (variable à introduire).  
  '`product_name`' (pour avoir le nom).
- Identifier des produits **similaires** → '`main_category`'.
- Restreindre le périmètre aux produits vendus en **France** → '`countries_en`'.
- Identifier les bénéfices **nutritionnels** → {  
  '`nutriscore_grade`'.  
  '`nutriscore_score`' (pour vérifier la cohérence).  
  '`nutrition-score-fr_100g`' (pour vérifier la cohérence).  
  '`additives_n`' (le moins le mieux).
- Identifier l'impact **écologique** → {  
  '`ingredients_that_may_be_from_palm_oil_n`'.  
  '`ingredients_from_palm_oil_n`'.  
  '`carbon-footprint_100g`'.  
  '`carbon-footprint-from-meat-or-fish_100g`'.



'`nutriscore_grade`' → rating donnée par "`nutrition-score-fr_100g`", déduit de plusieurs variables (ie. '`saturated-fat_100g`', '`fat_100`', '`sugars_100g`' et '`sodium_100g`') → plus à inclure.







## C. Remplissage de données manquantes





## C. 1 Variables d'identification générale

- 'code' → toujours renseigné (base de "Open Food Facts").
- 'countries\_en' → toujours renseigné (limité à la France).
- 'product\_name' → important juste pour affichage final.  
(Certains NaN, mais pas grave).





On ne peut pas les inventer !

## C.2 Variables d'identification de l'impact écologique

### Création de la variable '**Final-carbon-footprint**':

- 1) Fusion des colonnes  $\left\{ \begin{array}{l} \text{'carbon-footprint\_100g'}. \\ \text{'carbon-footprint-from-meat-or-fish\_100g'}. \end{array} \right.$
- 2) Ajout de 45g par  $\left\{ \begin{array}{l} \text{'ingredients\_that\_may\_be\_from\_palm\_oil\_n'}. \\ \text{'ingredients\_from\_palm\_oil\_n'}. \end{array} \right.$
- 3) Remplissage des derniers 'NaN' par les médianes des '**Final-carbon-footprint**' **associés à leur 'main\_category\_en'**.



Par hectare ~ 170 tonnes de carbone émis, et un rendement moyen mondial de 3.7 tonnes. **Soit 45g par g d'huile de palme.**

On rajoute 45g par ingrédient pour prendre le pire scénario et inciter à des produit plus écologiques (donc sans d'ingrédients issus de l'huile de palme).

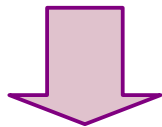




On ne peut pas les inventer !

## C.3 Variables d'identification de l'impact nutritionnel

Deux « parties » : 'additives' et 'nutriscore'.



Création de la variable '**additives\_Total**':

Remplissage des 'NaN' par les médianes des 'additives\_n' associés à leur 'main\_category\_en'.





On ne peut pas les inventer !

## C.3.1 Le nutriscore

Création de la variable '**nutriscore\_grade\_Total**':

- 1) Fusion des lignes doubles (produit renseigné plusieurs fois).
- 2) Fusion des colonnes  $\left\{ \begin{array}{l} \text{'nutriscore\_score'} \\ \text{'nutrition-score-fr\_100g'} \end{array} \right.$
- 3) Remplissage des derniers 'NaN' par les médianes respectives **associés à leur 'main\_category\_en'** , +1.
- 4) Transformation en «grade» (formules de Santé Publique France).
- 5) Cohérence avec 'nutriscore\_grade'.
- 6) Sélection du pire score entre le score obtenu et 'nutriscore\_grade'.



Les variables liées au 'nutriscore' sont calculés par des formules établies.

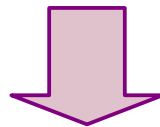




## C.4 Variable 'main\_category\_en'

**L'absence de 'main-categorie\_en' est contraignante.**

La 'main\_category\_en' on ne peut pas la renseignée «à dire d'expert ».



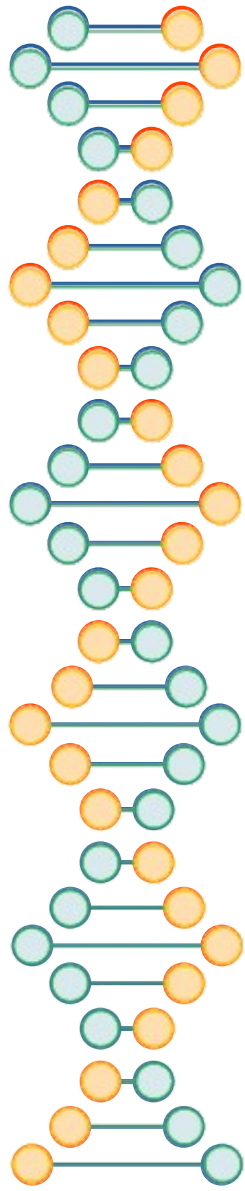
Élimination des lignes sans 'main\_category' et sans 'additives\_Total'.



## C.5 Remplissage de données manquantes (3 Stratégies)

- Fusion de colonnes et lignes 'semblables' (même info, même produit).
- Règles « à dire d'expert » → basées sur des renseignements scientifiques (ex: 45g par ingrédient huile de palme, et application du +1 et sélection du 'pire' score dans le cas du nutriscore).
- Remplissage par les médianes des variables associées à la 'main\_category\_en' des produits.





## Et les Outliers ????

Ils ne sont pas traités dans la partie de nettoyage car les 'NaN' furent remplis en partie par des règles « à dire d'expert » et donc enlever les outliers dès le début pouvait biaiser l'application de ces règles.

Il est donc préférable de les traiter dans la partie "exploration" car leur impact et importance dépend du remplissage des 'NaN' avec les différentes règle établies.

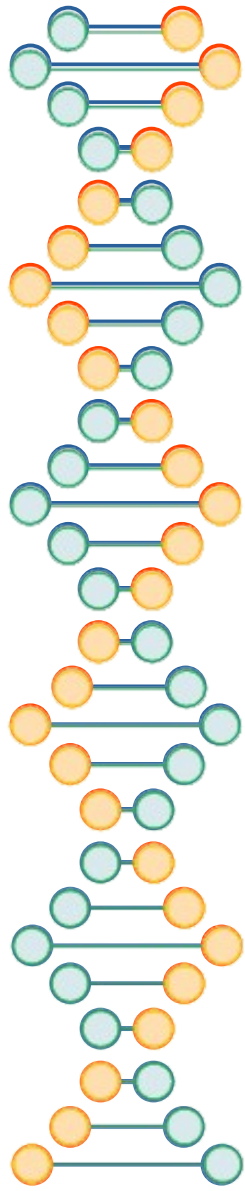




## Tableau Nettoyé

	code	product_name	countries_en	main_category_en	Final-carbon-footprint	nutriscore_grade_Total	additives_Total
3	0000000000100	moutarde au moût de raisin	France	Mustards	0.0	d	0.0
12	0000000000088	Pate d'amande	France	Almond paste	0.0	d	4.0
13	00000000000949	Salade de carottes râpées	France	Seasoned shredded carrots	0.0	b	2.0
31	0000000001885	Compote de poire	France	Pear compotes	0.0	a	1.0
33	0000000002103	Aiguillettes de poulet	France	fr:Aiguillettes de poulet	0.0	a	0.0
35	0000000002257	Salade de macedoine de légumes	France	Vegetables macedoines	0.0	b	2.0
40	000000000262	Confiture de lait	France	Milk jams	0.0	d	0.0





# Exploration des données



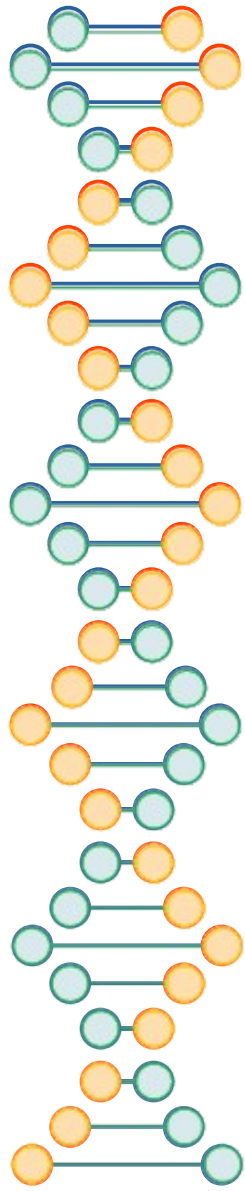
# Reconnaissance des variables

- 'code' → identifie le produit.
- 'countries\_en' → pour tous France.
- 'product\_name' → présentation finale (pas vraiment important).
- 'main-category\_en' → Classe le produit parmi ses semblables.

Pour faire notre 'score' on a donc 3 variables:

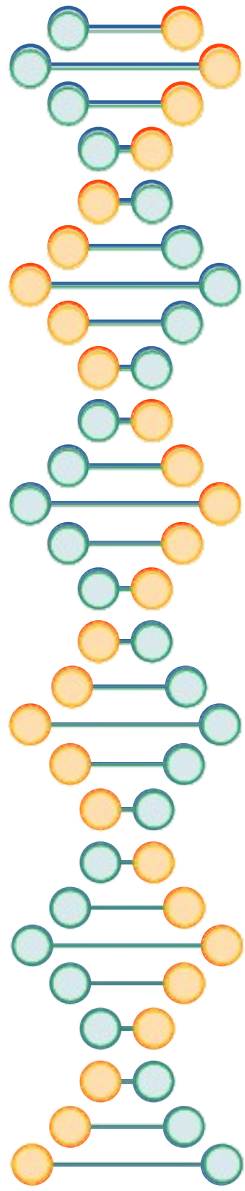
- **Une qualitative:** 'nutriscore\_grade\_Total';
- **Deux quantitatives:**  $\left\{ \begin{array}{l} \text{'Final-carbon-footprint'} \\ \text{'additives_Total'} \end{array} \right.$





## A. Analyse Univariée





## A.1 Variables quantitatives

(Identification des potentielles valeurs "aberrantes" (OUTLIERS)  
et forme de la distribution des variables)

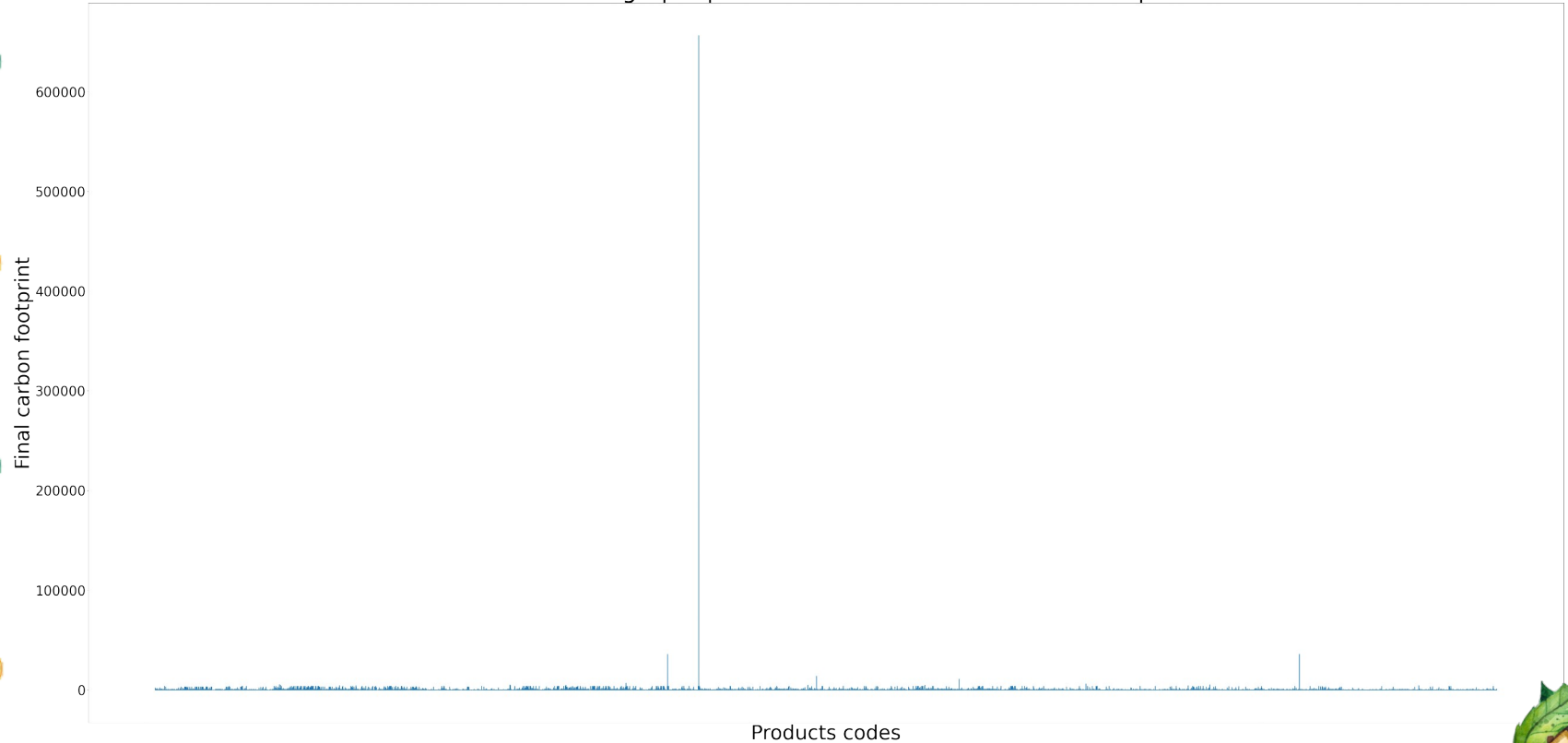
Les outliers des variables quantitatives ('Final-carbon-footprint' et 'additives\_Total') doivent être supprimées pour ne pas biaiser les résultats.

(Les coefficient de corrélation de Pearson et de Kendall, les boîtes à moustache -analyse univariée- et le K-means -analyse multivariée- étant sensibles aux outliers).



## A.1.1 Identification graphique des Outliers ('Final-carbon-footprint')

Identification graphique des Outliers du Final carbon footprint

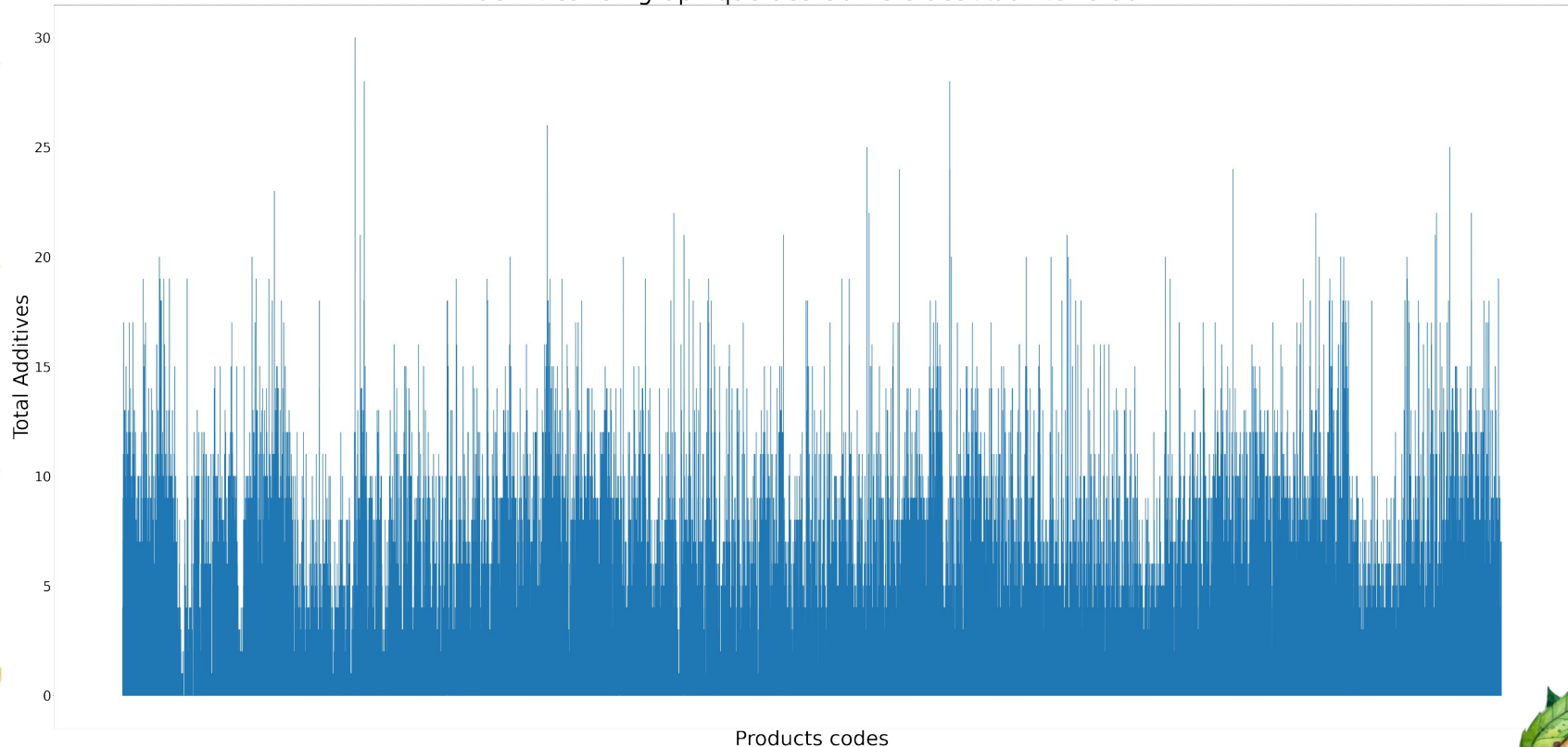


La valeur maximale du 'Final-carbon-footprint' est une valeur aberrante à supprimer.



## A.1.1 Identification graphique des Outliers ('additives\_Total')

Identification graphique des Outliers des Additifs Totaux



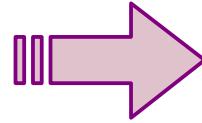
Les valeurs aberrantes pour 'additives\_Total', ne sont pas évidentes → à vérifier avec l'étude des 1er et 99ème quantiles.



## A.1.2 Élimination des Outliers (1er et 99ème percentiles)

1<sup>er</sup> et 99ème percentiles

	Final-carbon-footprint	additives_Total
<b>0.10</b>	0.0	0.0
<b>0.99</b>	559.0	9.0



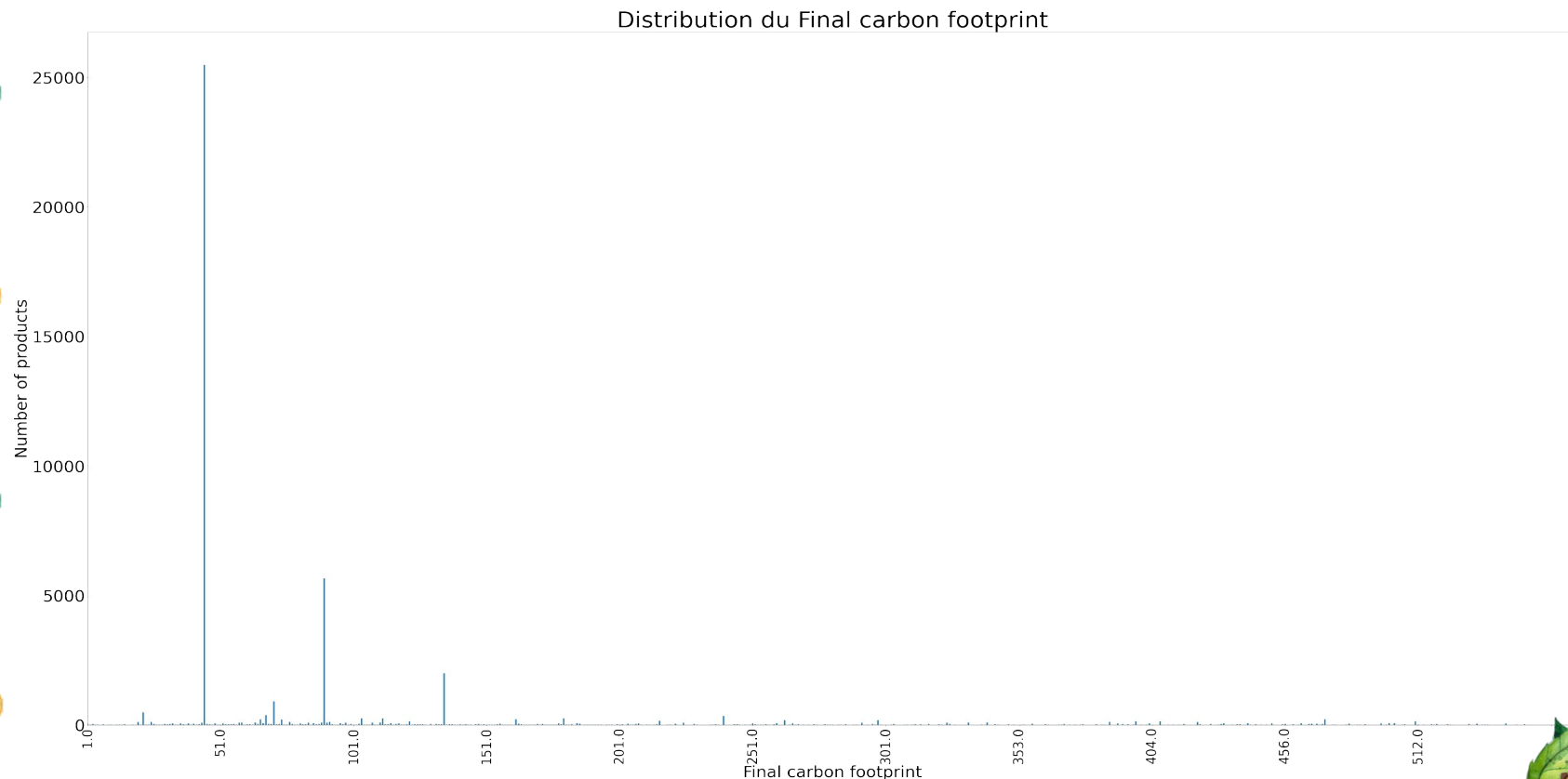
	Final-carbon-footprint	additives_Total
<b>count</b>	48273.000000	193921.000000
<b>mean</b>	104.685207	2.494841
<b>std</b>	112.993086	1.695061
<b>min</b>	1.000000	1.000000
<b>25%</b>	45.000000	1.000000
<b>50%</b>	45.000000	2.000000
<b>75%</b>	92.000000	3.000000
<b>max</b>	558.000000	8.000000

La std sans Outliers reste non nulle → le coefficient de corrélation de Pearson peut être calculé.





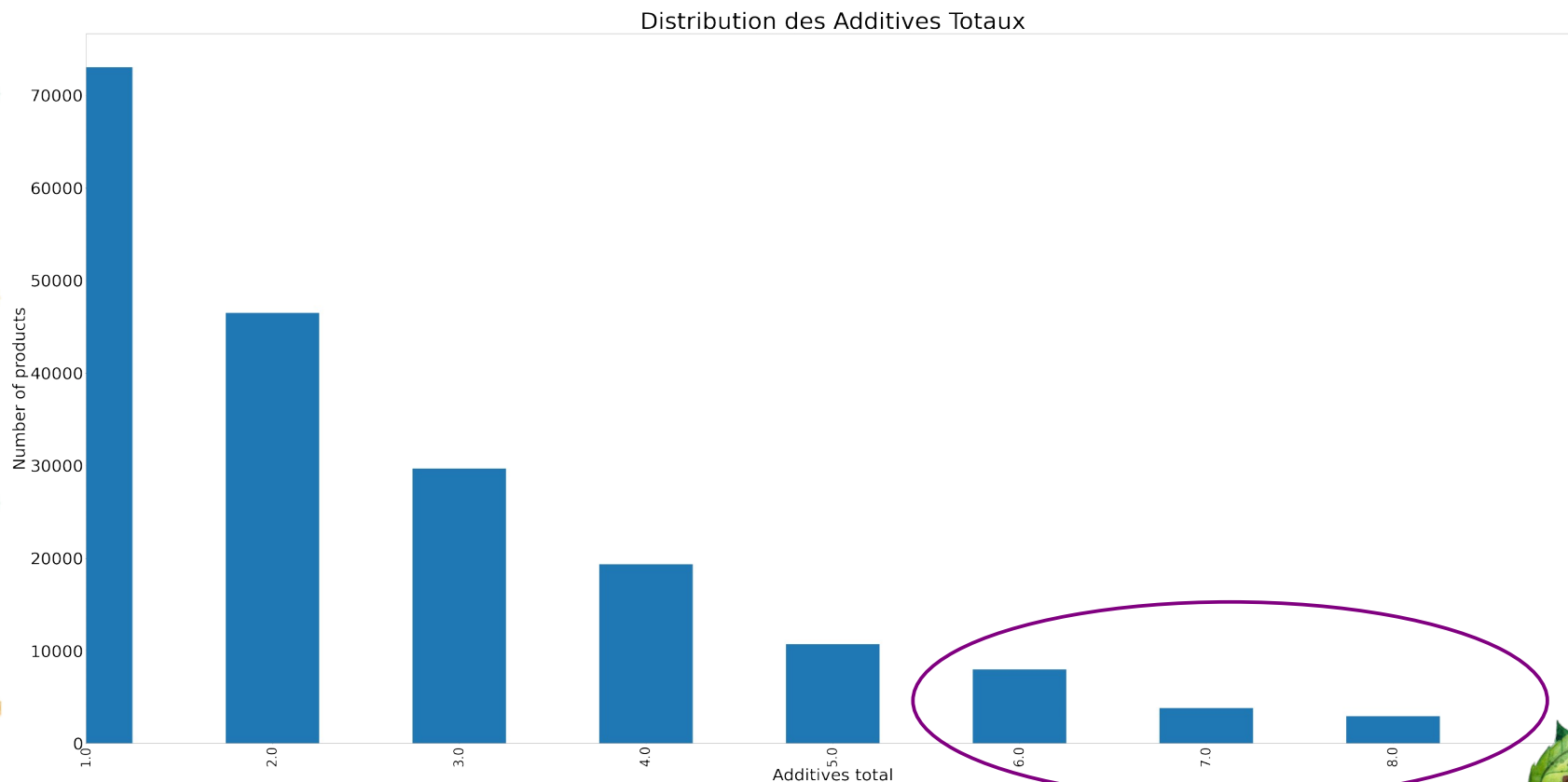
## A.1.3 Forme de la distribution des variables (*'Final-carbon-footprint'*)



La distribution (des 'fréquences') ne suit pas une distribution de loi normale, mais plutôt une distribution désordonnée.



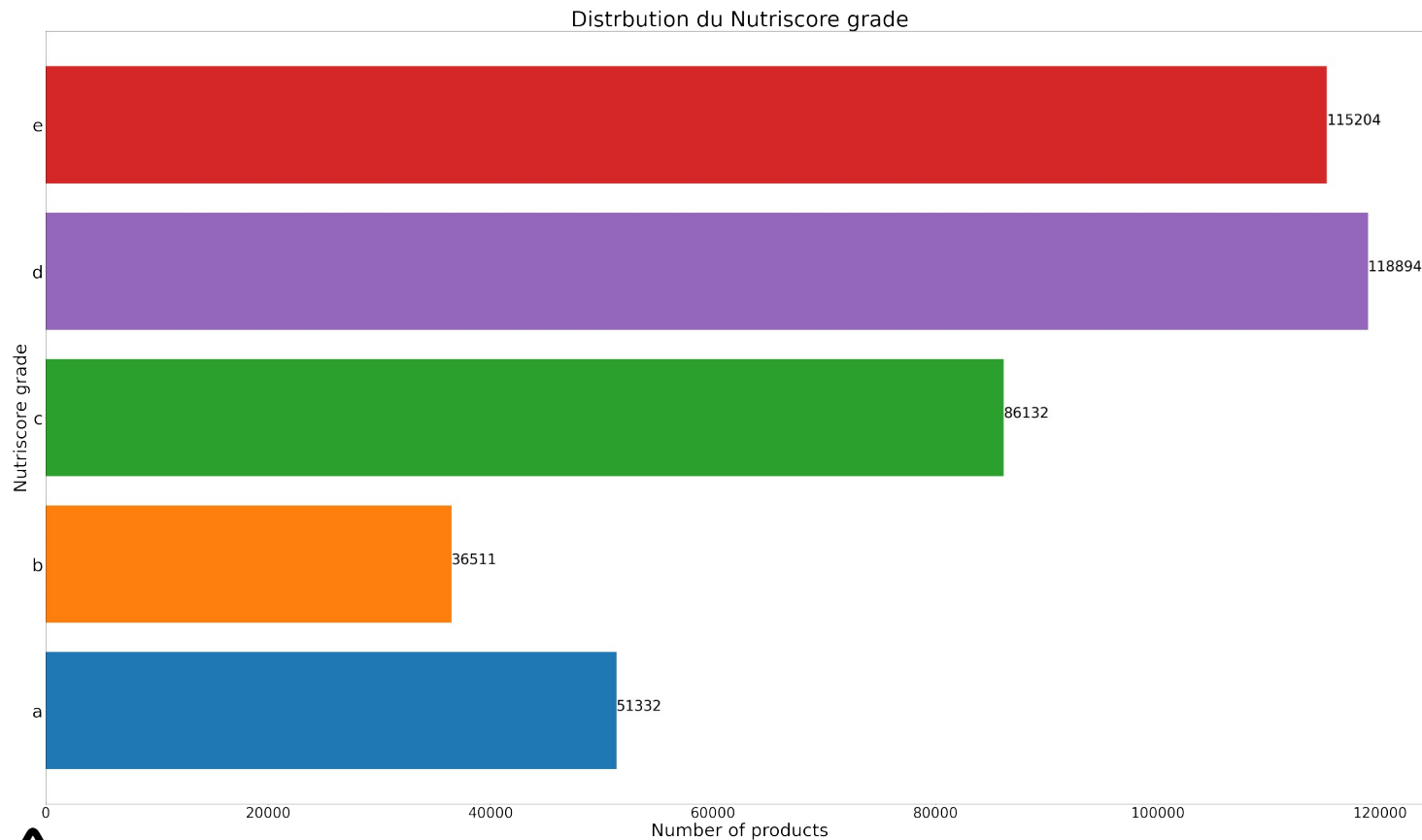
## A.1.3 Forme de la distribution des variables (`'additives_Total'`)



La distribution (des 'fréquences') ne suit pas une distribution de loi normale, (cf. `additives_Total`=[6,7,8]).

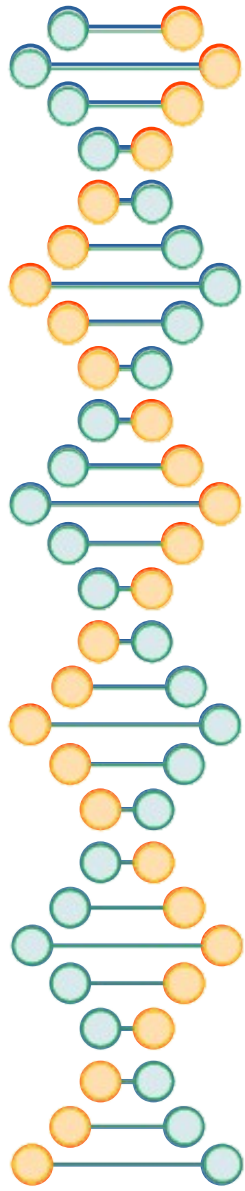


## A.2 Variable qualitative (Diagramme en barres)



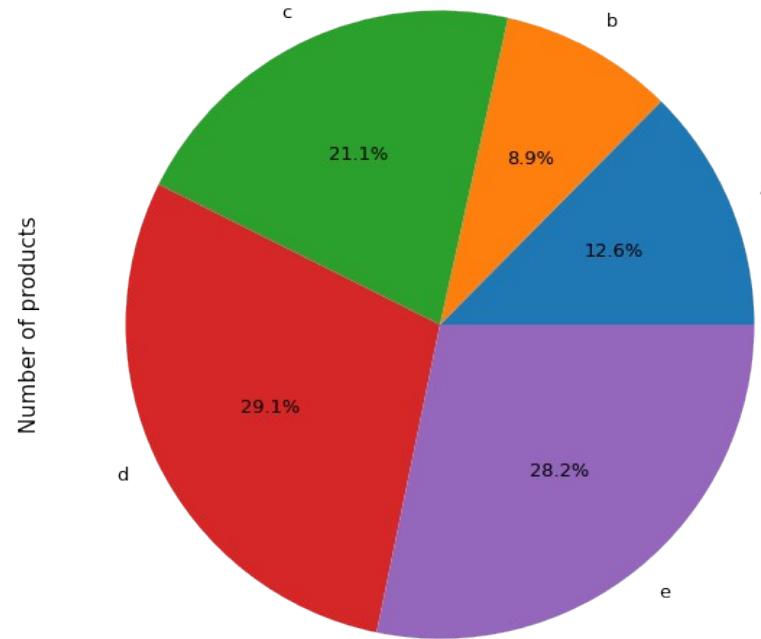
On tient compte des Outliers, pour bien voir le comportement indépendant de la variable qualitative face aux variables quantitatives.





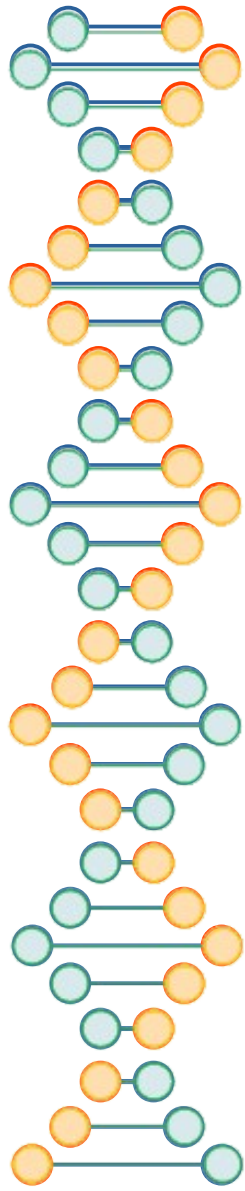
## A.2 Variable qualitative (Diagramme camembert)

Distribution du Nutriscore grade



Les différent 'Nutriscore grade' ne contiennent pas un nombre homogène (équitable) de produits → données déséquilibrées.





## B. Analyse Bivariée

(Mesures de liaison entre 2 variables)



## B.1 Entre les 2 variables quantitatives (Coefficient de corrélation de Kendall)

Kendall → Car les distributions de nos variables ne suivent pas une distribution normale.

(Si elles étaient normales on utiliserait le coefficient de corrélation de Pearson).

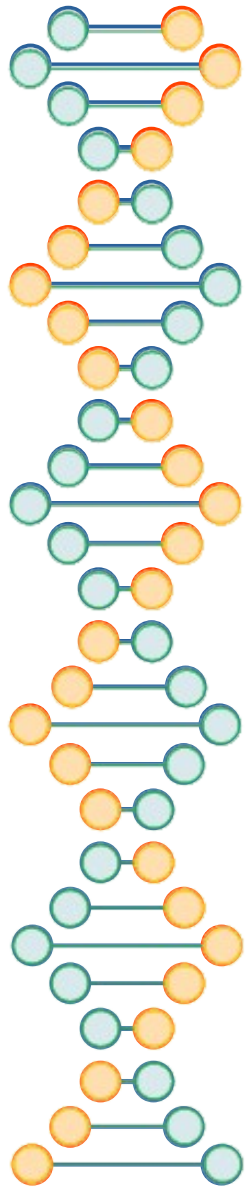
	Final-carbon-footprint	additives_Total	nb
Final-carbon-footprint	1.000000	0.125752	NaN
additives_Total	0.125752	1.000000	NaN
nb	NaN	NaN	1.0

Pas de corrélation de rang → coefficient de Kendall < 0.8



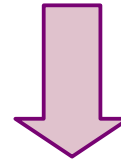
Un coefficient de Kendall nul ne signifie pas obligatoirement une indépendance absolue des deux variables! Il faut approfondir toujours un test ANOVA.





## B.2 Entre la variable qualitative et les variables quantitatives ('boîtes à moustaches' ou 'boxplot', ANOVA, Heatmaps des corrélations Kendall)

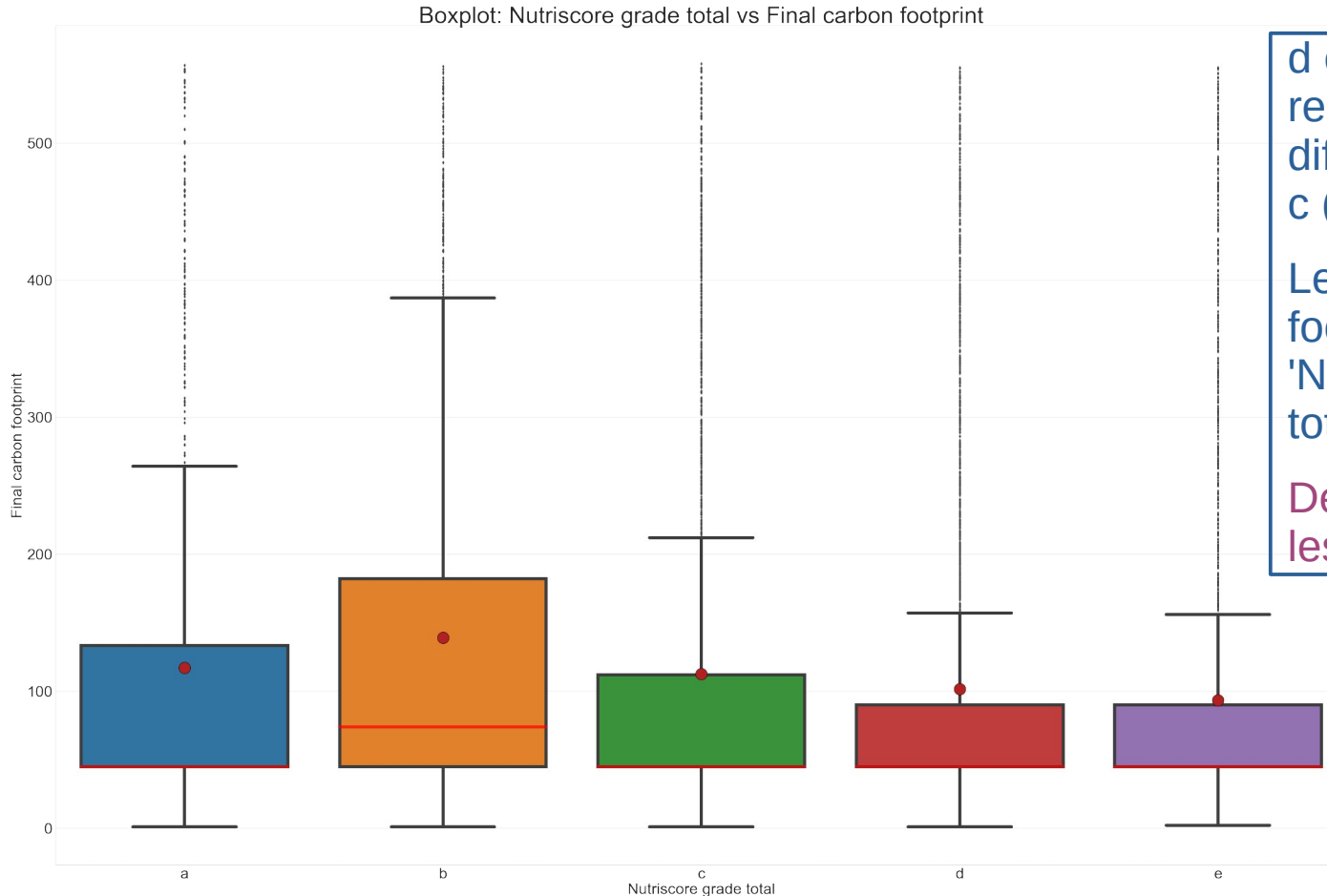
Le 'nutriscore\_grade\_Total' possède 5 'modalités': a, b, c, d, et e.



On analyse le comportement des variables 'Final-carbon-footprint' et 'additives\_Total' pour chaque modalité.



## B.2.1 Boîtes à moustaches (`'nutriscore_grade_Total'` vs `'Final-carbon-footprint'`)



d et e se ressemblent mais différent avec a, b et c (surtout c).

Le 'Final carbon footprint' diffère d'un 'Nutriscore grade total' à un autre.

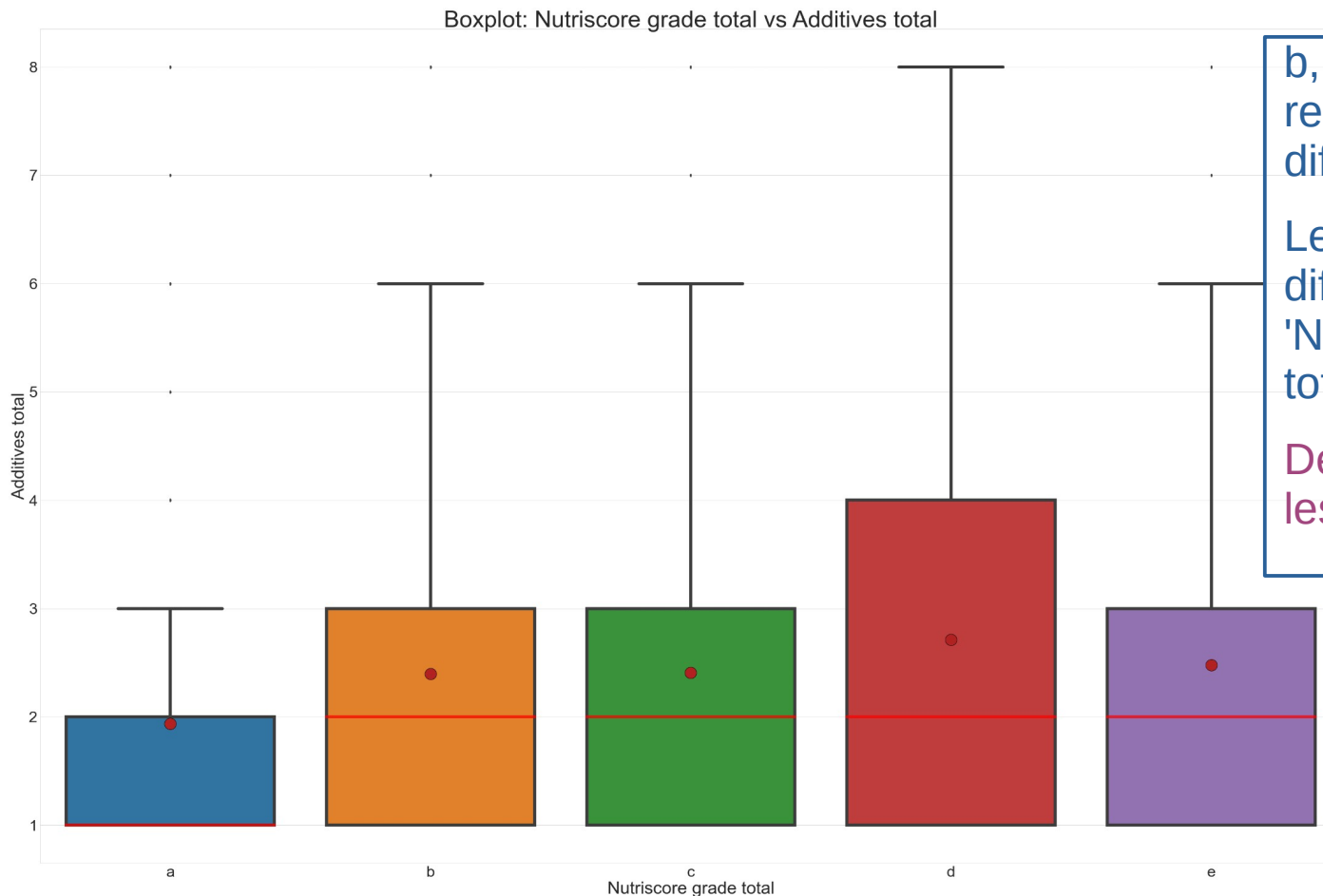
Dépendance entre les variables.





## B.2.1 Boîtes à moustaches

('nutriscore\_grade\_Total' vs 'additives\_Total')



b, c et e se ressemblent mais diffèrent avec a et d.

Les 'additives\_Total' diffèrent d'un 'Nutriscore grade total' à un autre.

Dépendance entre les variables.

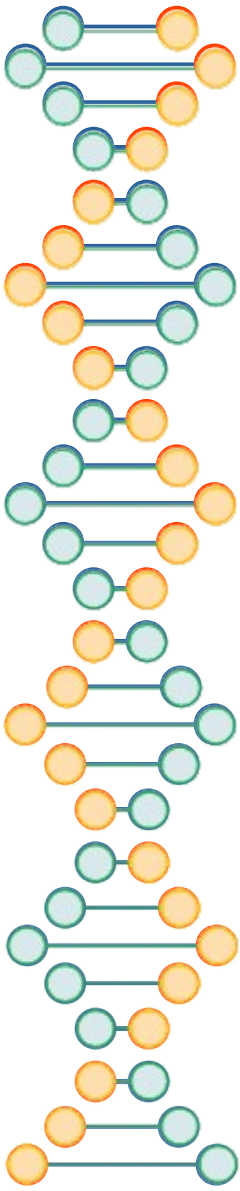


## B.2.2 ANOVA

(One-Way ANOVA car on a 3 variables)

- Mesure la dépendance d'une variable quantitative à une (ou deux) variable(s) qualitative(s) en **analysant la variance** :
  - **Moyenne** variable quantitative **homogène** pour toutes des modalités de la variable qualitative ? (Ressemblance avec diagrammes à moustache).
  - **Si NON** la dépendance est **significative**.
- L'ANOVA utilise un test de Fisher (ou test F) :
  - Hypothèse nulle  $H_0$  : égalité des moyennes (ie. moyennes homogènes).
  - **Si  $F \gg 1$  (et donc la  $p\text{-value} \ll 0.05$ )**  $H_0$  est rejetée → moyennes NON homogènes → **dépendance des variables**.
- On utilise l'ANOVA (même si on a un unbalanced data set) pour vérifier la dépendance entre:
  - 'Final-carbon-footprint' et 'nutriscore\_grade\_Total' ;
  - 'additives\_Total' et 'nutriscore\_grade\_Total'.





## B.2.2 ANOVA

(One-Way ANOVA car on a 3 variables)

### B2.2.1 'Final-carbon-footprint' et 'nutriscore\_grade\_Total'

	df	sum_sq	mean_sq	F	PR(>F)
<b>nutriscore_grade_Total</b>	4.0	7520.605255	1880.151314	663.307781	0.0
<b>Residual</b>	193916.0	549656.482848	2.834508	NaN	NaN

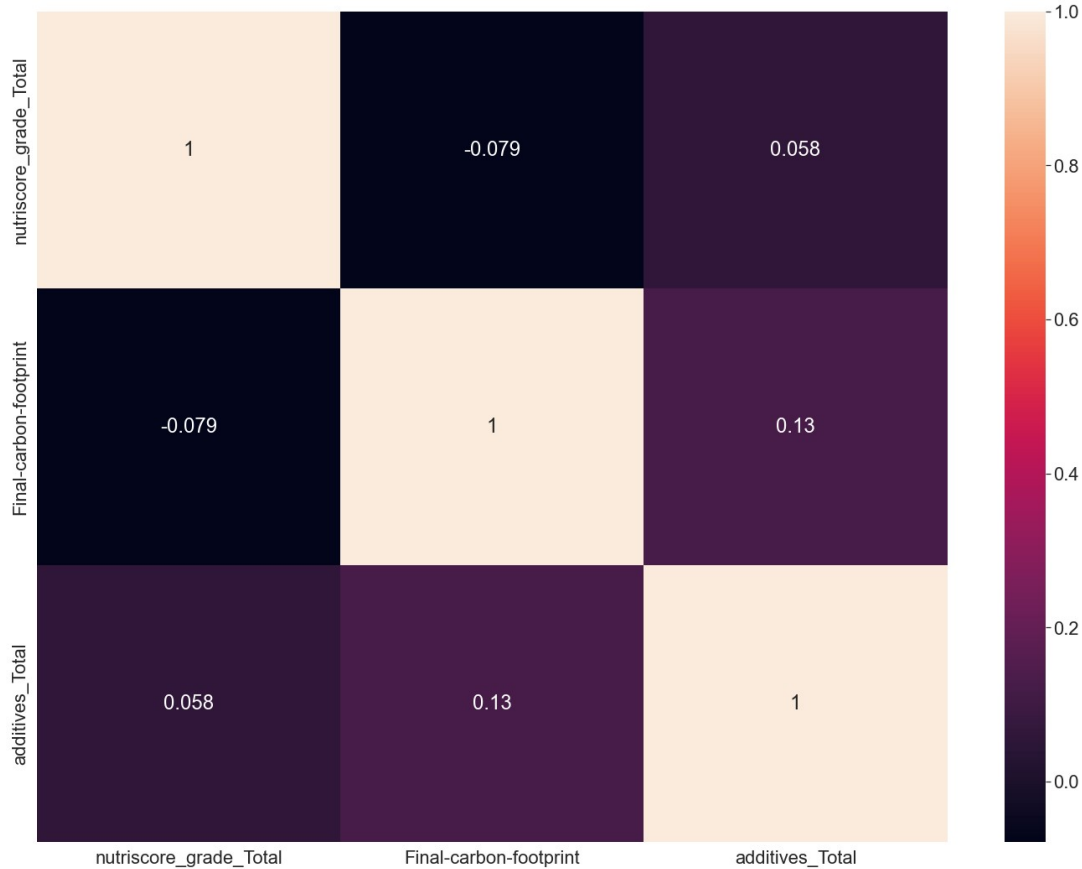
### B.2.2.2 'additives\_Total' et 'nutriscore\_grade\_Total'

	df	sum_sq	mean_sq	F	PR(>F)
<b>nutriscore_grade_Total</b>	4.0	7.224141e+06	1.806035e+06	143.122269	7.416030e-122
<b>Residual</b>	48268.0	6.090856e+08	1.261883e+04	NaN	NaN

Dans les 2 cas, la **p-value (PR)** est  $\ll 0.05$  → **dépendance des variables**.  
(Vu dans les diagrammes à moustaches → pour certains 'nutriscore' comportement différent).

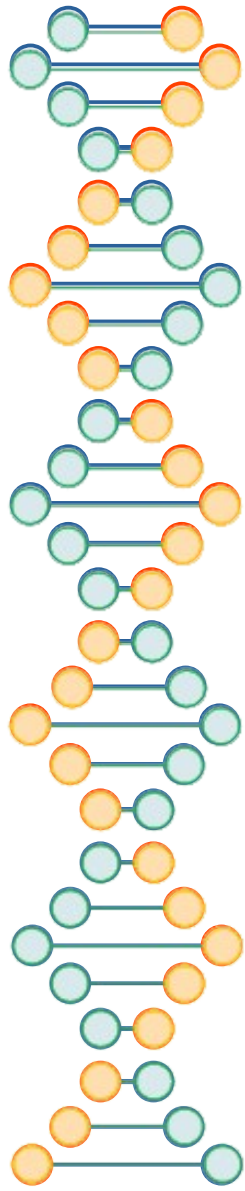


## B.2.3 Heatmaps (Coefficient des corrélations de Kendall)



Coefficients de corrélation de Kendall entre nos 3 variables  $< 0.8 \rightarrow$  **pas de corrélation de rang** entre elles.





## C. Analyse Multivariée

(NON GO dans notre cas)



## C. Processus

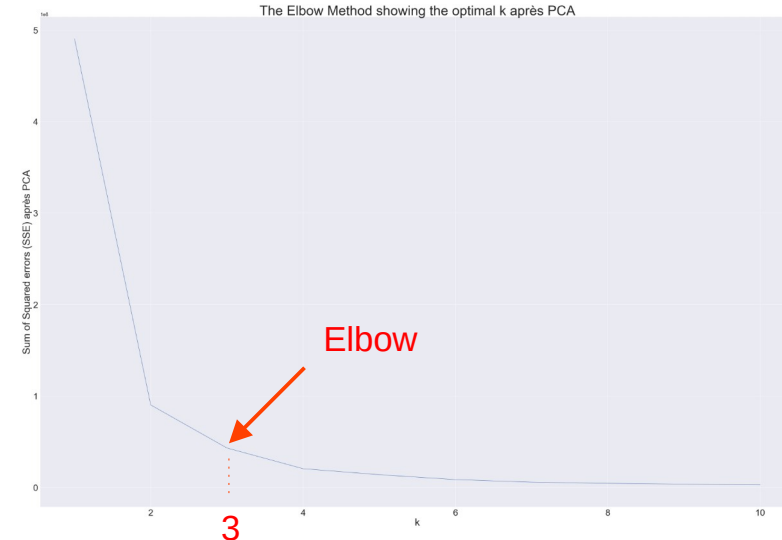
- Réduction dimensionnelle (PCA) → réduire à 2 dimensions (on a 3, car 3 variables).  
But : visualiser les résultats avec un nuage de points 2D.

	nutriscore_grade_Total	Final-carbon-footprint	additives_Total
11	1	402.0	3.0
35	2	90.0	4.0
49	5	45.0	3.0
64	4	90.0	5.0
65	4	90.0	6.0
...			



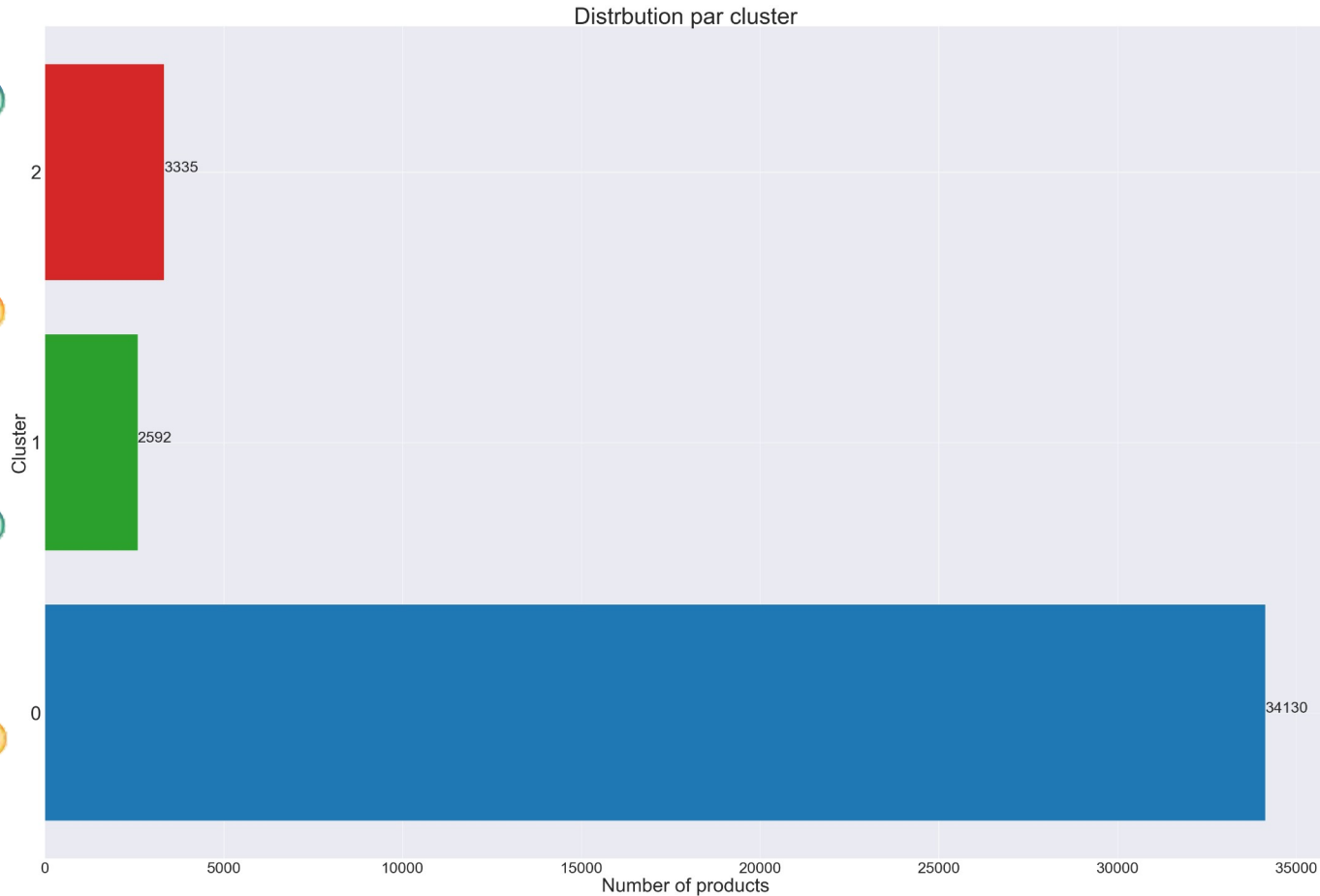
	0	1
0	300.359673	-0.772986
1	-11.640924	0.306937
2	-56.644117	-0.653178
3	-11.642678	1.326708
4	-11.642443	2.326659
...		

- Obtention du 'k' (Elbow method)  
Le Elbow est 3 → Ad hoc avec notre idée de faire 3 catégories: 'Bon, Moyen, Mauvais'.



## C. Processus

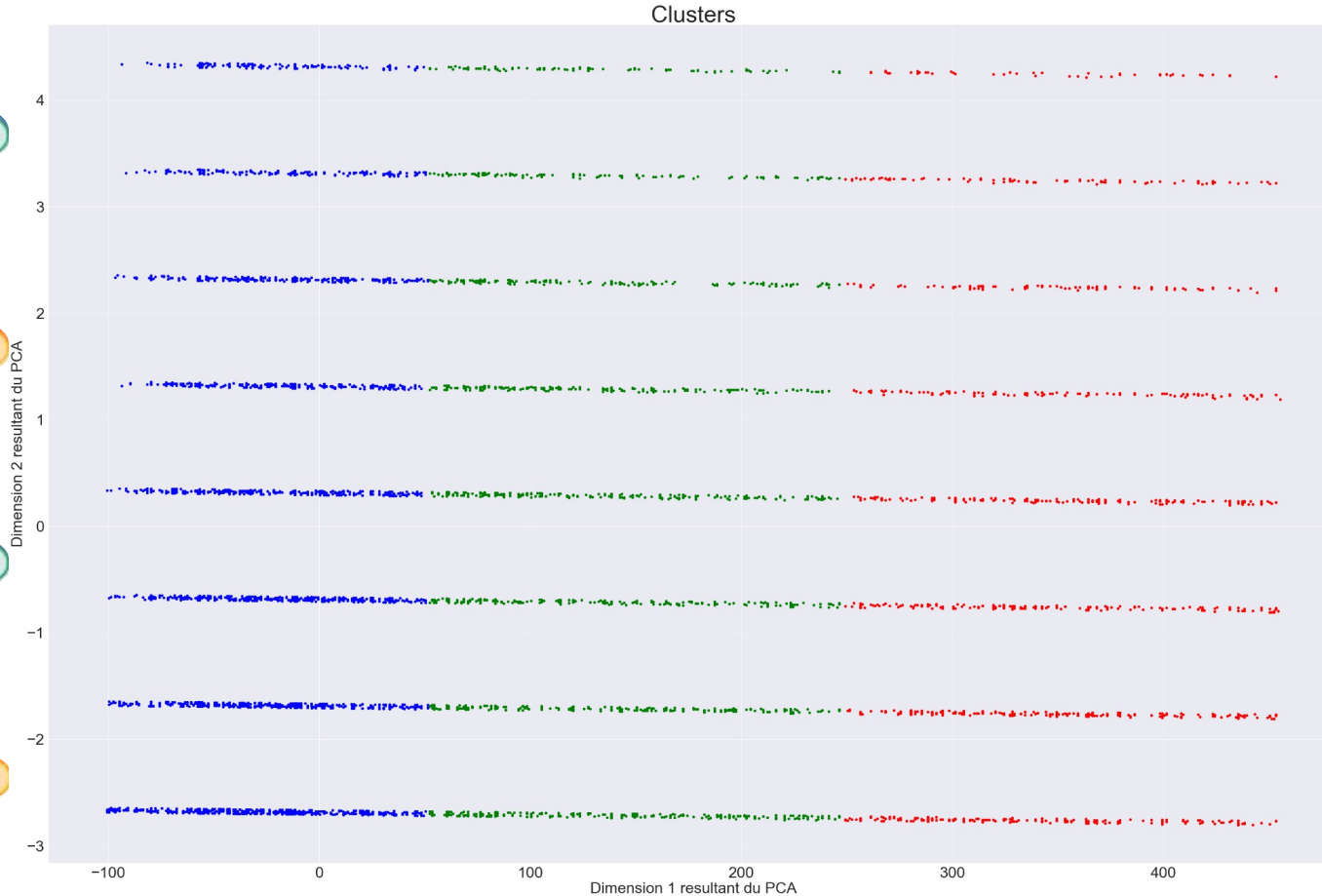
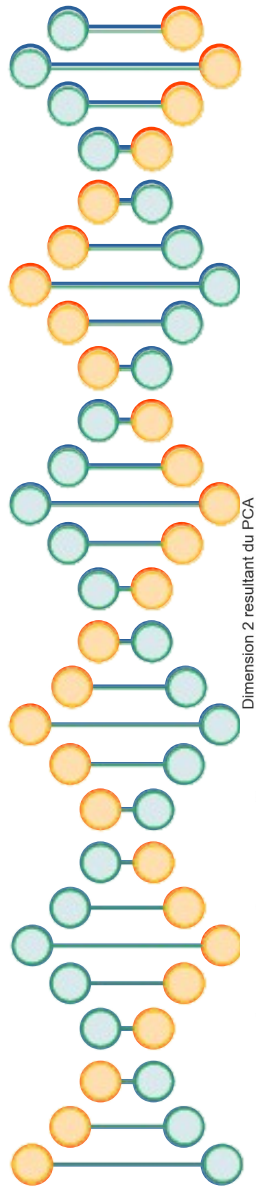
- Application du K-Means (k=3)



3 clusters avec suffisamment d'éléments pour être considérés, mais pas très homogènes.



# C. Processus



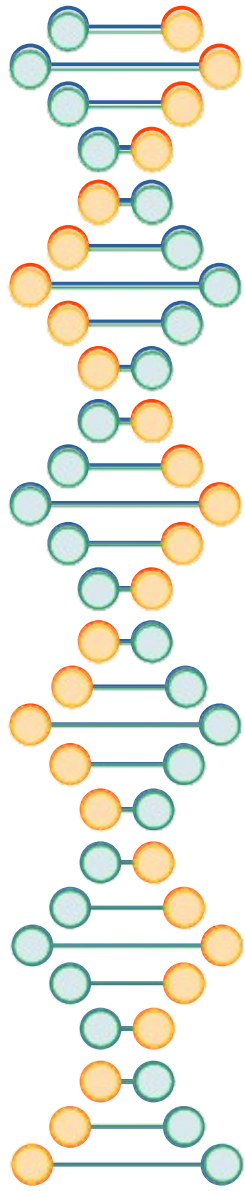
3 clusters bien  
séparés entre eux  
(indépendants).

Dans chaque  
cluster 8 "sous  
clusters": **k-Means**  
se base dans des  
mesures de  
"distances".

**Pas illustratif  
pour notre  
objectif.**



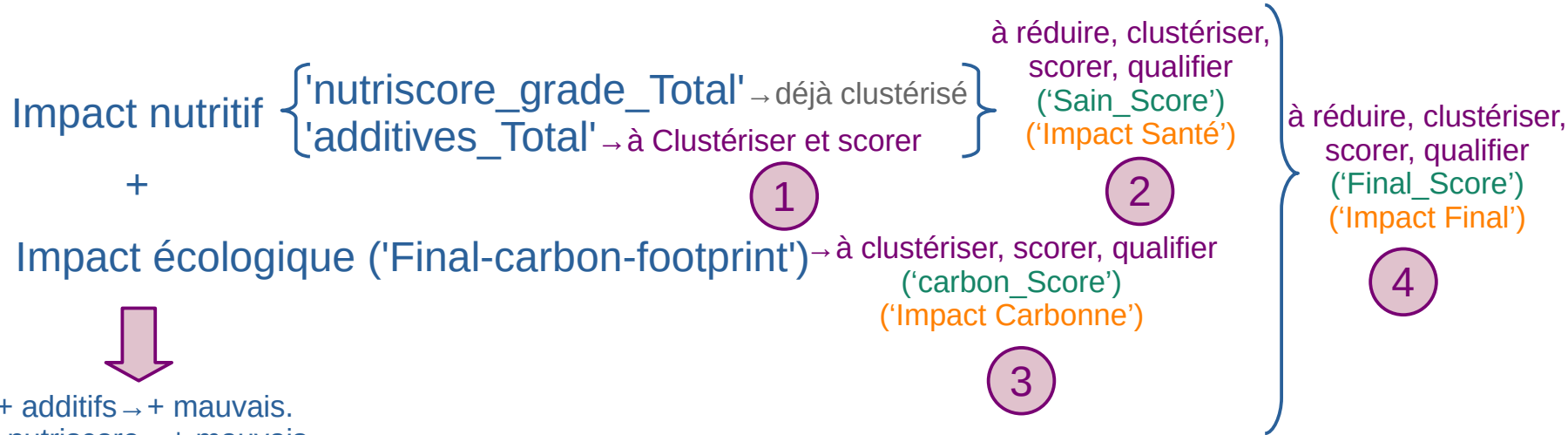




## D. Notre Processus



On veut:



+ additifs → + mauvais.  
+ nutriscore → + mauvais.  
+ carbone → - écologique.



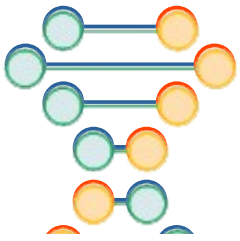
**Réduire:** PCA ou Somme

**Clustériser:** K-Means (avec Elbow Method).

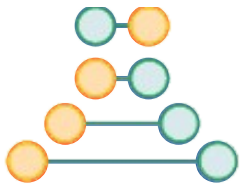
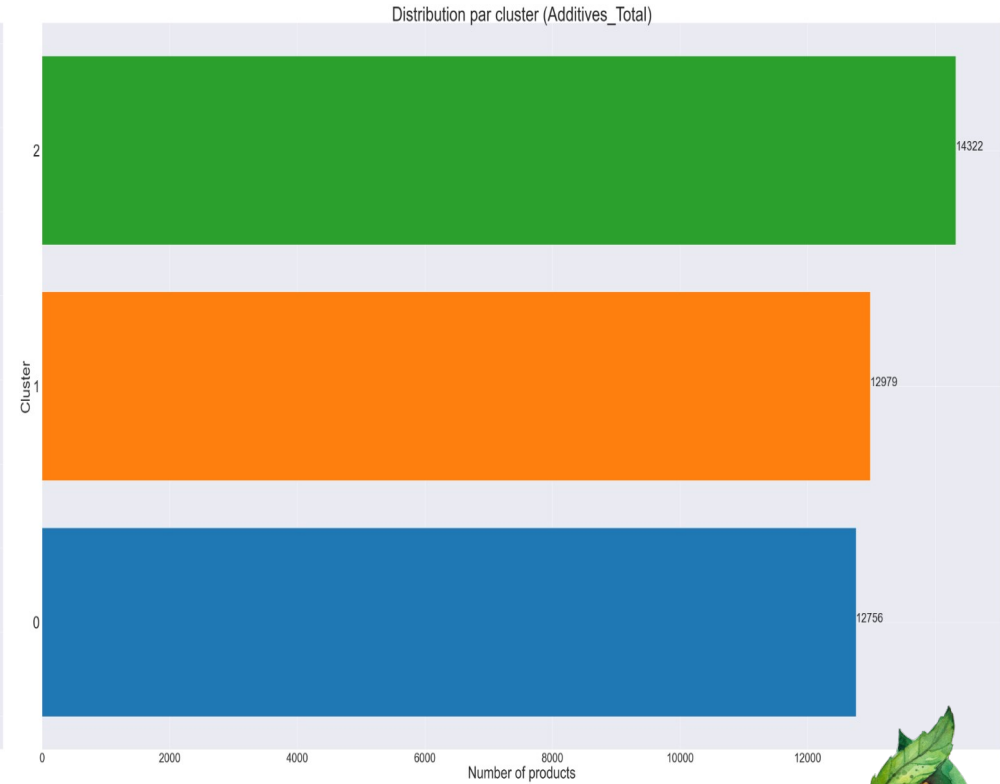
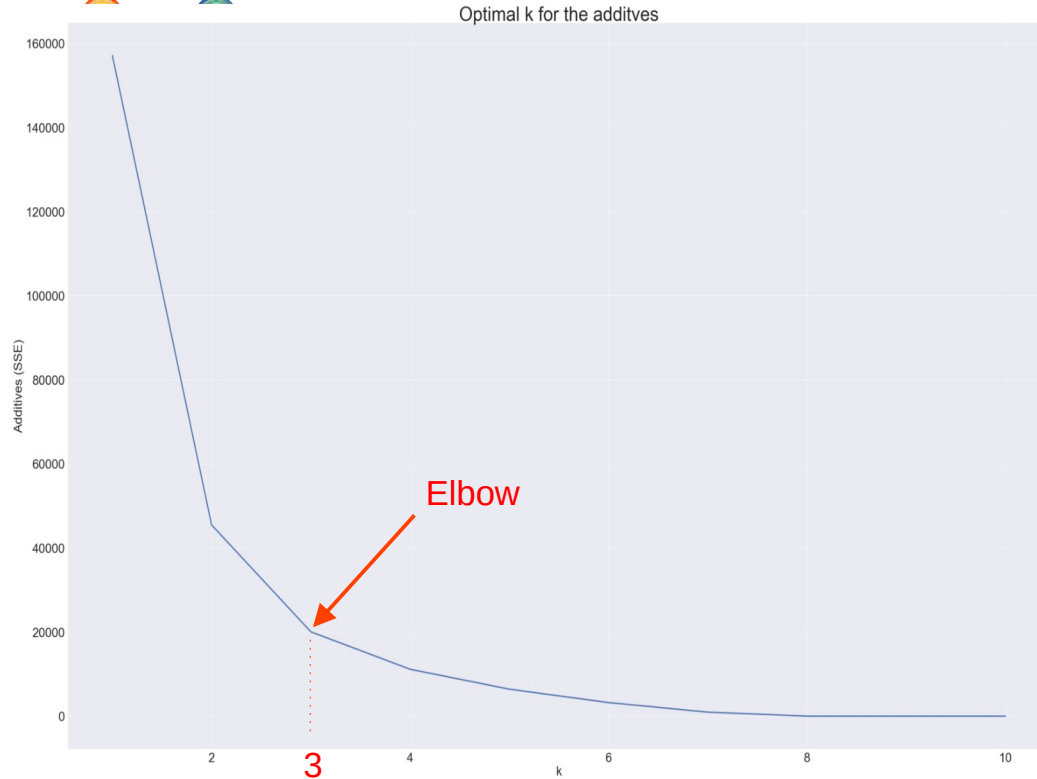
**Scorer:** dans chaque groupe plus de points plus c'est le plus mauvais.

**Qualifier:** 'Bon' → 1, 'Moyen' → 2, 'Pas Bon' → 3





## D.1 Clustériser: 'additives\_Total' (k-Means avec Elbow method)




## D.1 Scorer: 'additives\_Total'

(Plus de points plus c'est le plus mauvais)

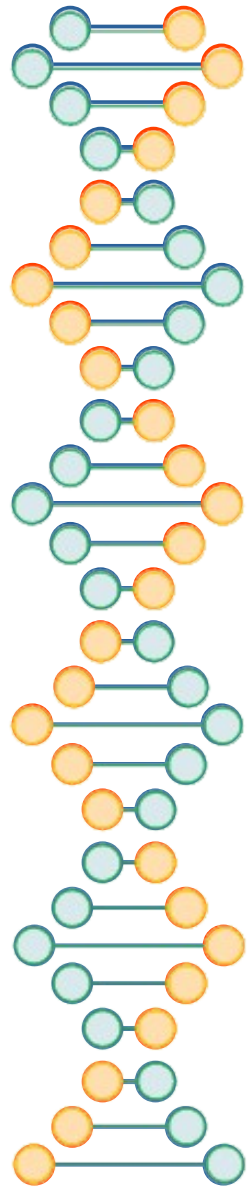


3 clusters :  
Cluster 0 → 1 point.  
Cluster 2 → 2 points.  
Cluster 1 → 3 points.



	additives_Total	additives_clusters	additives_Score
11	3.0	2	2
35	4.0	2	2
49	3.0	2	2
64	5.0	1	3
65	6.0	1	3
...			





## D.2 Réduire: Impact nutritif - 'Sain'

(PCA: 'nutriscore\_grade\_Total' et 'additives\_Score')

	code	nutriscore_grade_Total	additives_Score
11	0000000005470	1	2.0
35	0000000491228	2	2.0
49	0000007730009	5	2.0
64	0000010206515	4	3.0
65	0000010216477	4	3.0
	...		

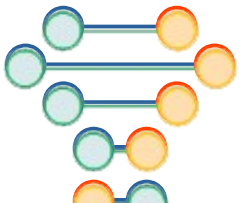
2 Variables



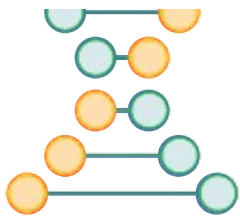
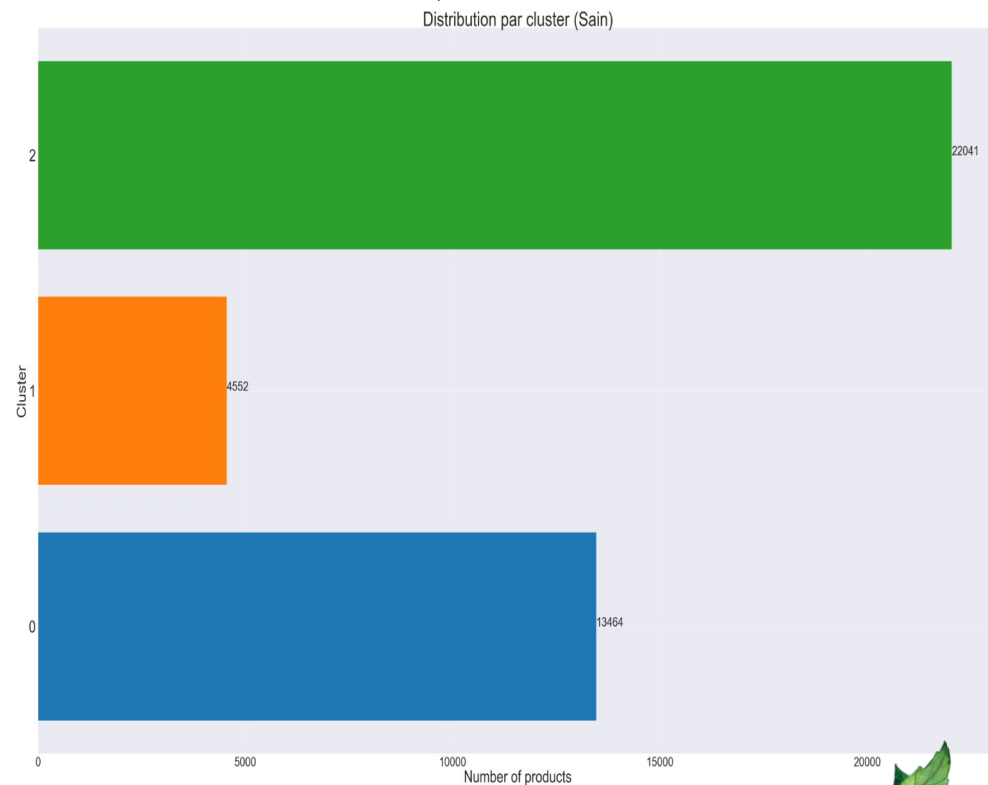
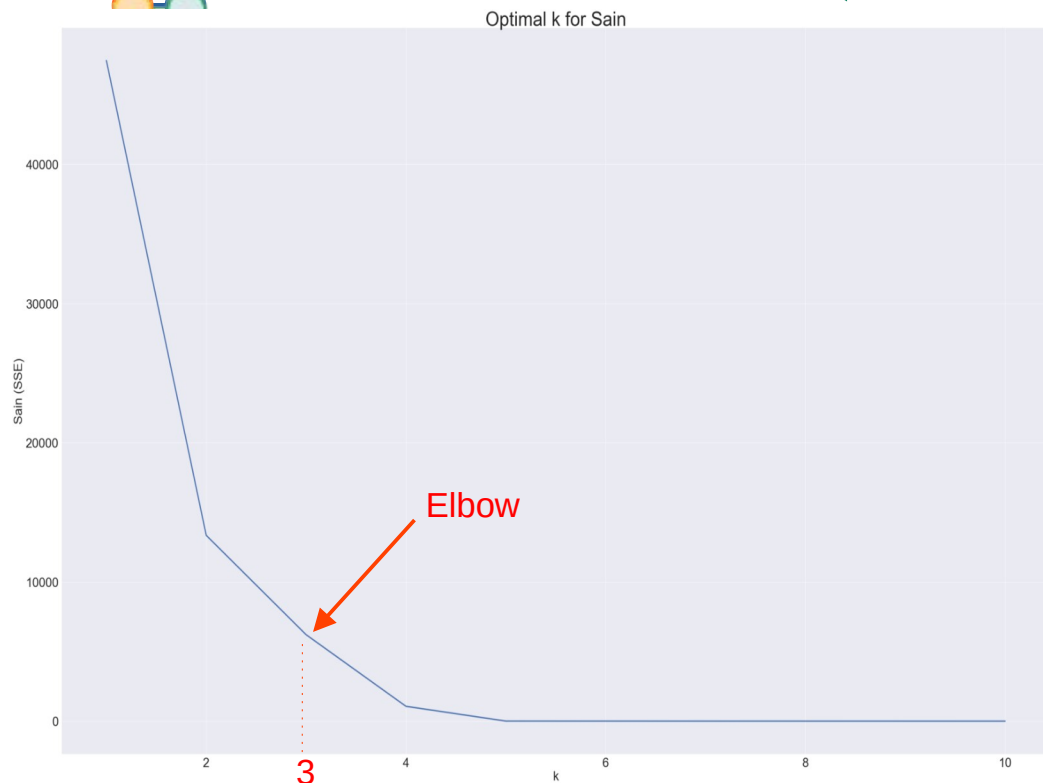
	code	Sain
0	0000000005470	2.864114
1	0000000491228	1.864243
2	0000007730009	-1.135372
3	0000010206515	-0.151532
4	0000010216477	-0.151532
	...	

1 Variable





## D.2 Clustériser: 'Sain' (K-Means avec Elbow method)



## D.2 Scorer: 'Sain'

(Plus de points plus c'est le plus mauvais)



3 clusters :  
Cluster 1 → 1 point;  
Cluster 2 → 2 points;  
Cluster 0 → 3 points

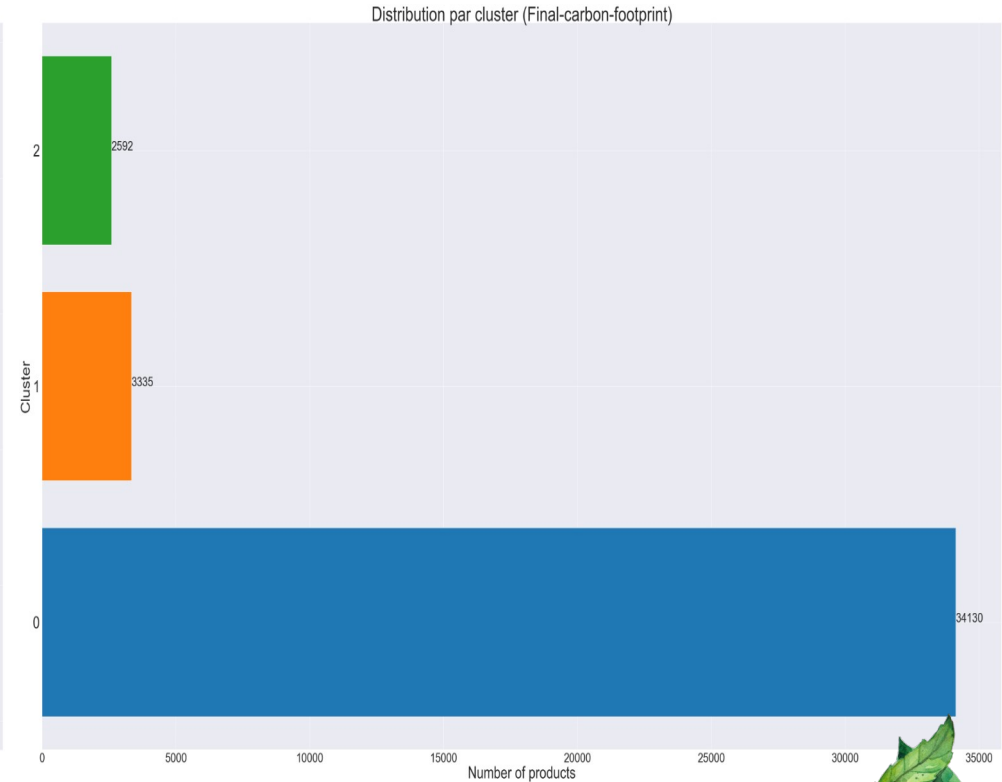
	code	Sain	Sain_clusters	Sain_Score
0	0000000005470	2.864114	1	1
1	0000000491228	1.864243	1	1
2	0000007730009	-1.135372	0	3
3	0000010206515	-0.151532	2	2
4	0000010216477	-0.151532	2	2
	...			



Les cluster semblent s'intercepter → On montre les cluster sur la somme des scores des deux variables qui constituent 'Sain' pour comprendre que représente chaque cluster en terme de "bon", "moyen" et "pas bon"



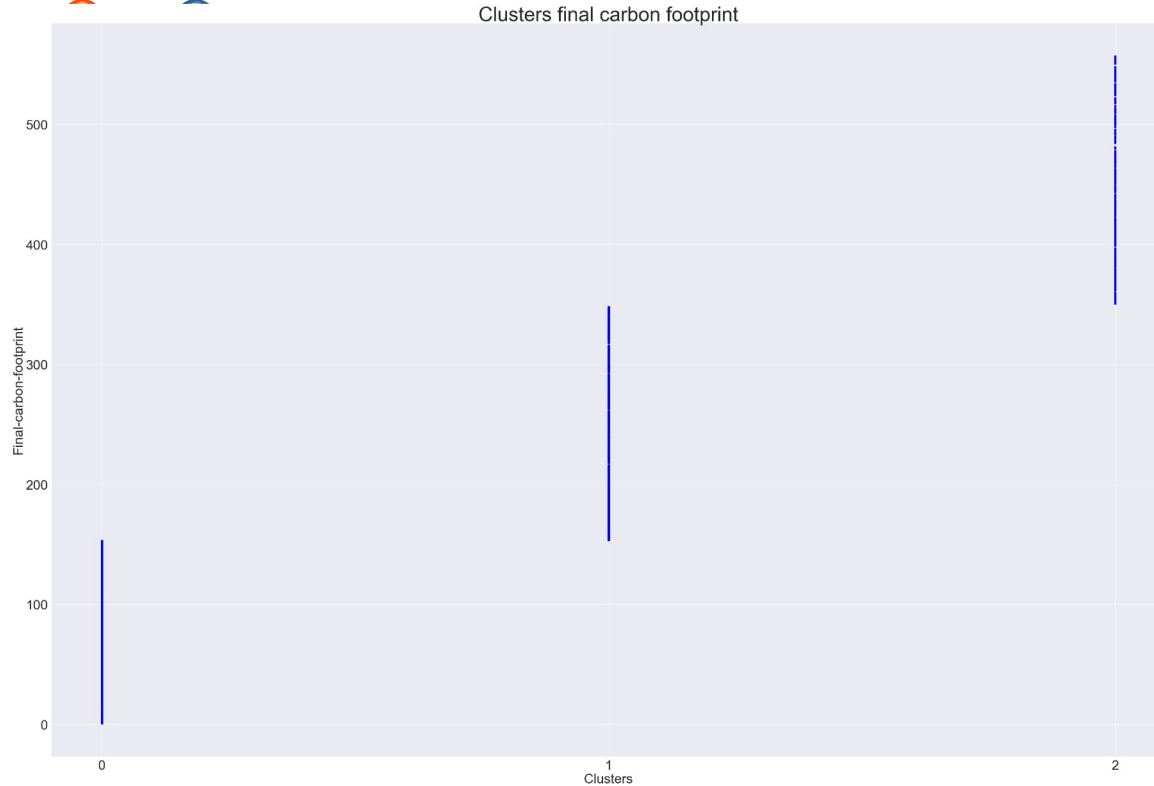
## D.3 Clustériser: 'Final-carbon-footprint' (k-Means avec Elbow method)






## D.3 Scorer: 'Final-carbon-footprint'

(Plus de points plus c'est le plus mauvais)



3 clusters :  
Cluster 0 → 1 point.  
Cluster 1 → 2 points.  
Cluster 2 → 3 points.



	Final-carbon-footprint	carbon_clusters	carbon_Score
11	402.0	2	3
35	90.0	0	1
49	45.0	0	1
64	90.0	0	1
65	90.0	0	1
...			



## D.4 Qualifier: 'Sain\_Score' et 'carbon\_Score' ( 'Impact Santé' et 'Impact Carbone' )

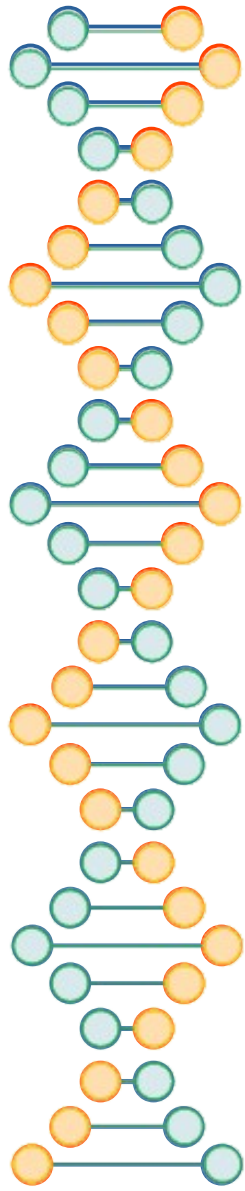
code	main_category_en	product_name	nutriscore_grade_Total	additives_Total	Sain	Sain_Score	Impact Santé	Final-carbon-footprint	carbon_Score	Impact Carbone
0000000005470	Baguettes	BAguette bressan	a	3.0	2.864114	1	Bon	402.0	3.0	Pas Bon
0000000491228	Dried products to be rehydrated	Entremets Crème Brulée	b	4.0	1.864243	1	Bon	90.0	1.0	Bon
0000007730009	Shortbread cookies	Biscuits sablés fourrage au cacao	e	3.0	-1.135372	3	Pas Bon	45.0	1.0	Bon
0000010206515	fr:decorations	Pâte à Sucre	d	5.0	-0.151532	2	Moyen	90.0	1.0	Bon
0000010216477	fr:Pâtes à sucre	Pate a sucre	d	6.0	-0.151532	2	Moyen	90.0	1.0	Bon
...										

Impact Santé

Impact Carbone

'Bon' → 1, 'Moyen' → 2, 'Pas Bon' → 3





- D.5 Réduire: Impacts nutritif et écologique - 'Total\_Score' (Somme: 'Sain\_Score' et 'carbon\_Score')
- On veut que les deux variables 'Sain\_Score' et 'carbon\_Score' aient le même poids, la même importance → critères de corrélation et inertie du PCA ne doivent pas être considérés.
- On fait donc uniquement la somme des deux scores et on la clustérise.

code	Sain_Score	carbon_Score
0000000005470	1	3.0
0000000491228	1	1.0
0000007730009	3	1.0
0000010206515	2	1.0
0000010216477	2	1.0
...		

2 Variables

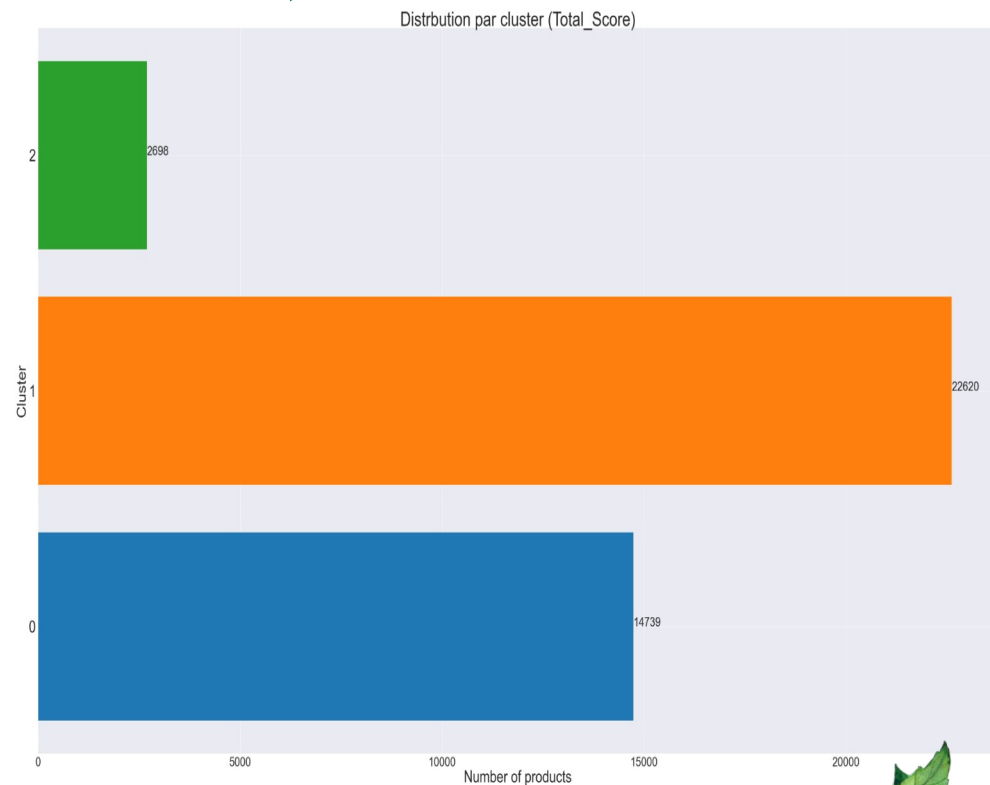
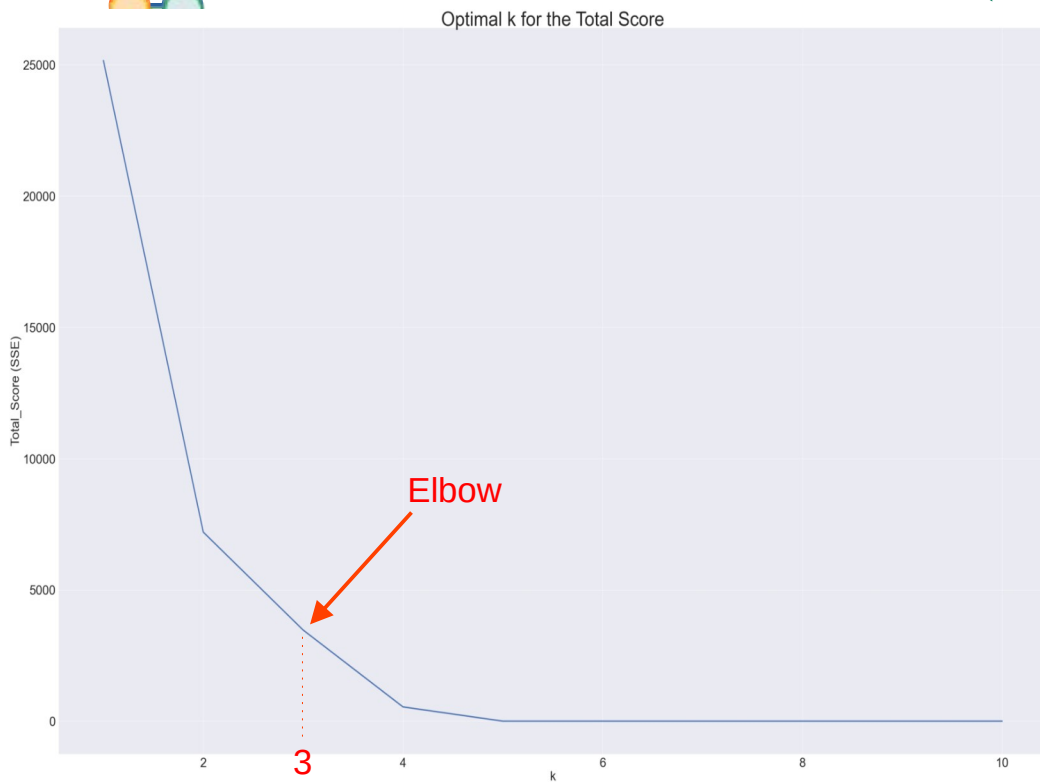


Total_Score	
0	4.0
1	2.0
2	4.0
3	3.0
4	3.0
...	

1 Variable



## D.4 Clustériser: 'Total\_Score' (K-Means: 'Sain')



## D.4 Scorer: 'Total\_Score'

(Plus de points plus c'est le plus mauvais)



3 clusters :  
Cluster 1 → 1 point.  
Cluster 0 → 2 points.  
Cluster 2 → 3 points.

Total_Score	Total_Score_clusters	Final_Score
4.0	0	2
2.0	1	1
4.0	0	2
3.0	1	1
3.0	1	1
...		



## D.4 Qualifier: 'Final Score' ('Impact Final')

code	main_category_en	product_name	nutriscore_grade_Total	additives_Total	Sain	Sain_Score	Impact Santé	Final-carbon-footprint	carbon_Score	Impact Carbone	Total_Score	Final_Score	Impact Final
0000000005470	Baguettes	BAquette bressan	a	3.0	2.864114	1	Bon	402.0	3.0	Pas Bon	4.0	2	Moyen
0000000491228	Dried products to be rehydrated	Entremets Crème Brulée	b	4.0	1.864243	1	Bon	90.0	1.0	Bon	2.0	1	Bon
0000007730009	Shortbread cookies	Biscuits sablés fourrage au cacao	e	3.0	-1.135372	3	Pas Bon	45.0	1.0	Bon	4.0	2	Moyen
0000010206515	fr:decorations	Pâte à Sucre	d	5.0	-0.151532	2	Moyen	90.0	1.0	Bon	3.0	1	Bon
0000010216477	fr:Pâtes à sucre	Pate a sucre	d	6.0	-0.151532	2	Moyen	90.0	1.0	Bon	3.0	1	Bon
...													

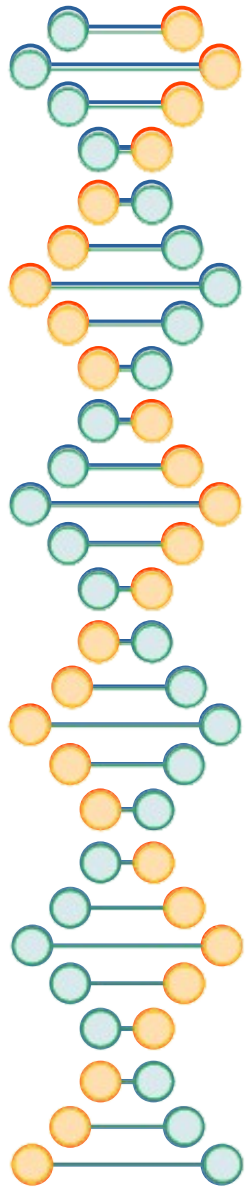
Impact Santé

Impact Carbone

Impact Final

'Bon' → 1, 'Moyen' → 2, 'Pas Bon' → 3





## E. Livrable Interactif



## E. Exemple 1: Produit existant mal noté.

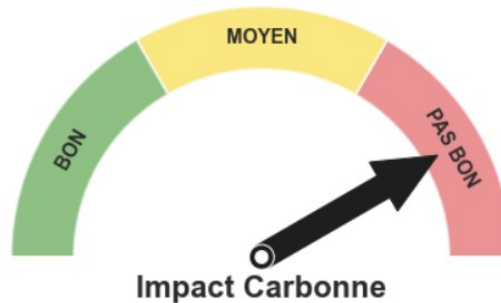
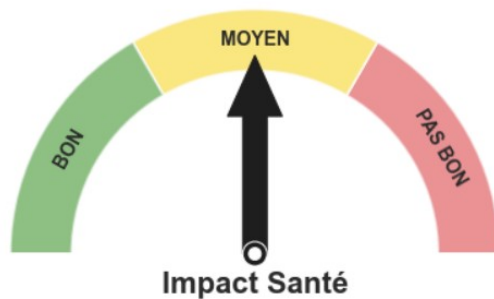
(Fonction 'results()' → Input: Quel est le code de votre produit?)

Quel est le code de votre produit?

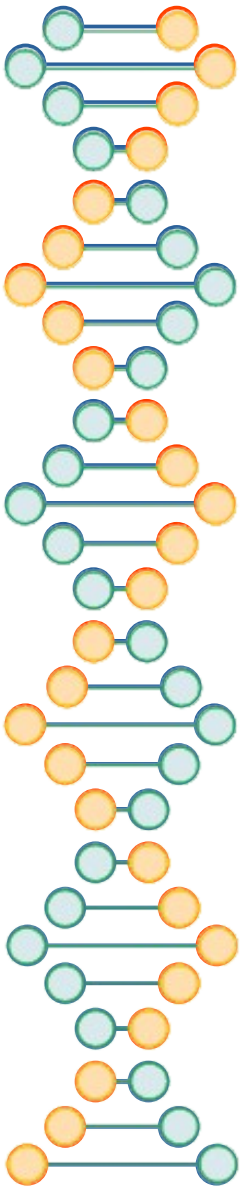
Pour un meilleur impact sur la santé et l'écologie optez plutôt pour l'un des 3 produits suivants:

code	product_name
2000000205397	Baguette Millavoise 6 céréales
3700128360556	La Millavoise aux 6 graines
8718265810952	Baguettes blanche précuite

BAguette bressan







## E. Exemple 2: Produit inexistant.

(Fonction `'results()'` → Input: Quel est le code de votre produit?)

Quel est le code de votre produit?000000000001

Domage on ne connaît pas encore ce produit

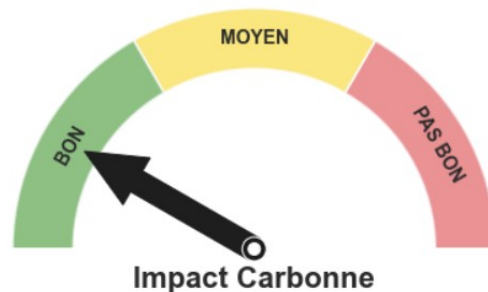
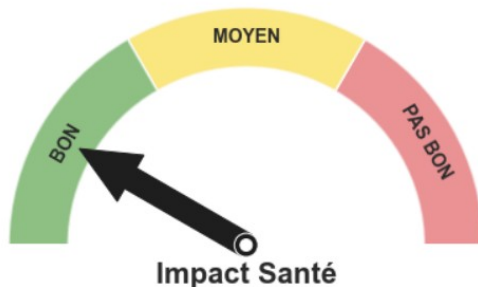


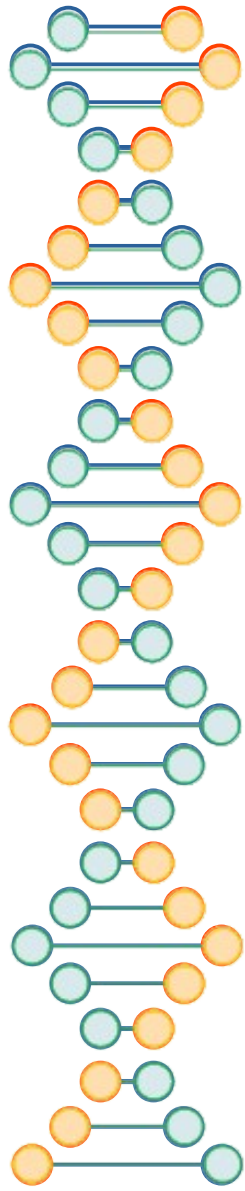
## E. Exemple 1: Produit existant bien noté.

(Fonction 'results()' → Input: Quel est le code de votre produit?)

Quel est le code de votre produit?

Entremets Crème Brulée





MERCI

