Une application au service de la santé publique

ETUDE DE FAISABILITÉ

EXPLOITATION DU JEU DE DONNÉES OPEN FOOD FACTS

Plan

Présentation de l'application envisagée

Opérations de nettoyage effectuées

Analyse Univariée

Analyse multivariée

Pertinence et Faisabilité

Synthèse et Conclusion

Présentation de l'application

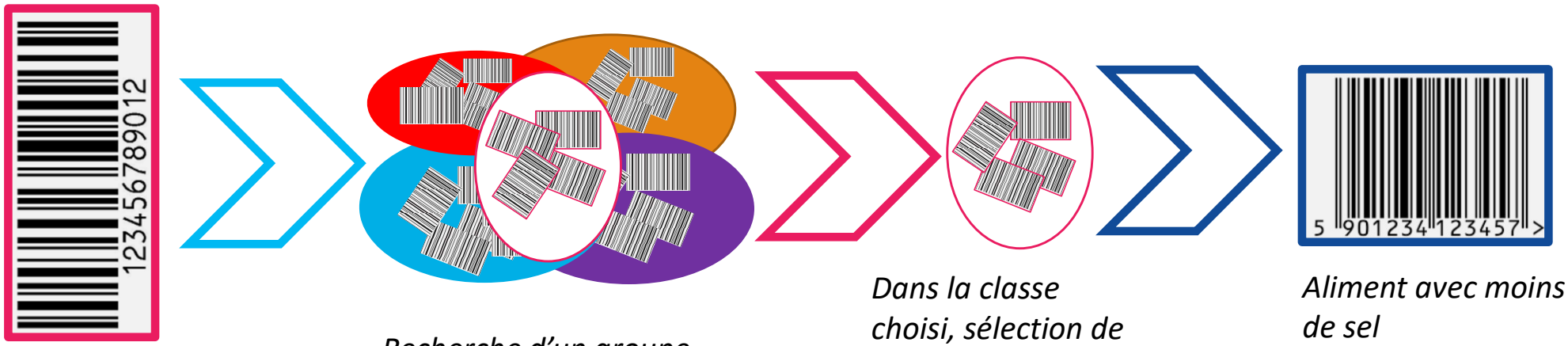
Application : nutri-Lsi

Beaucoup de problème de santé, sont dus à une mauvaise alimentation. Au-delà du problème d'obésité, il est très fréquent que des personnes mangent des aliments trop riches en certains nutriments particulier qui à force affaiblit voir détruit leur organisme.

Pour un produit donné l'application nutri-Lsi vise à proposer un produits moins riche en un nutriment donné.

Présentation de l'application

Principe de l’application nutri-Lsi



Le consommateur renseigne son aliment dans l’application.

Recherche d’un groupe d’aliments équivalents (même classe) en terme alimentaire.

Dans la classe choisi, sélection de l’aliment contenant le moins de sel

Aliment avec moins de sel

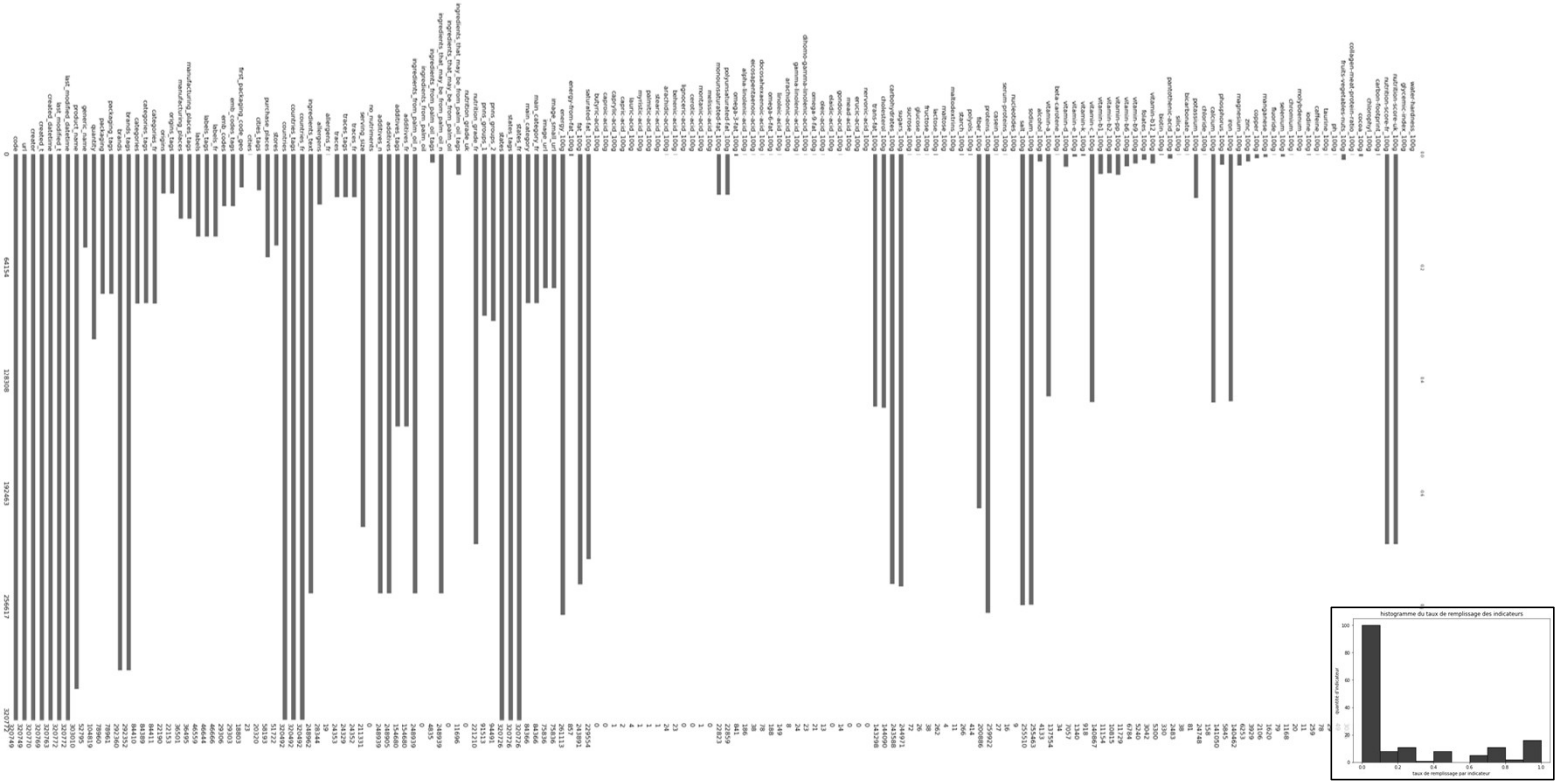
Opérations de nettoyage effectuées

Présentation du dataset

310 k produits
x
162 indicateurs

avec
106 colonnes numériques
dont
16 vides

~50 masse nutriment



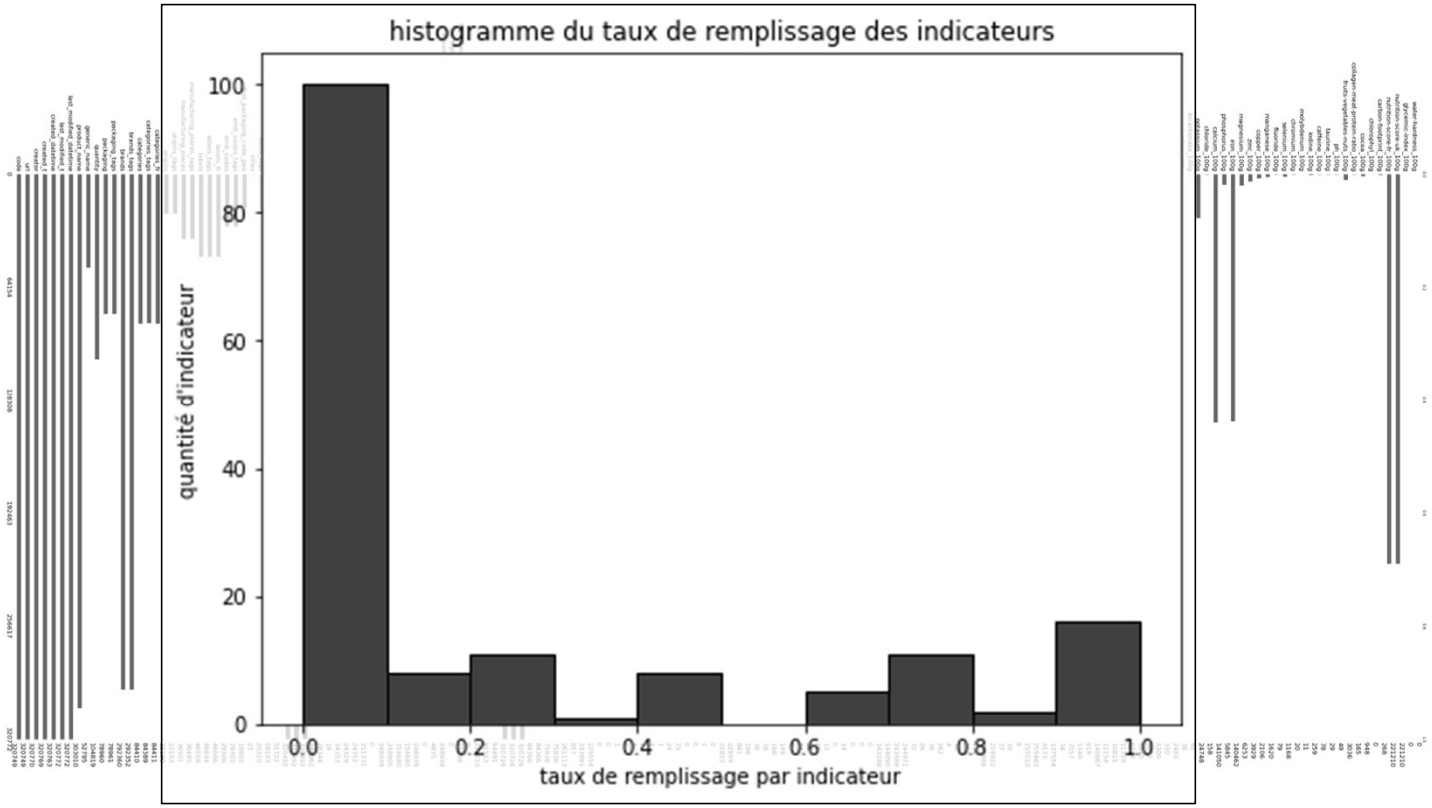
Opérations de nettoyage effectuées

Présentation du dataset

310 k produits
x
162 indicateurs

avec
106 colonnes numériques
dont
16 vides

~50 masse nutriment



Opérations de nettoyage effectuées

nettoyage en 6 étapes

1 - Suppression des indicateurs entièrement NaN

2 – suppression des duplicata dans l'indicateur « code »

3 - Suppression des indicateurs non utilisés (colonne vide ou nulle)

4 - Suppression des lignes dont les indicateurs « non-identitaire » sont NaN

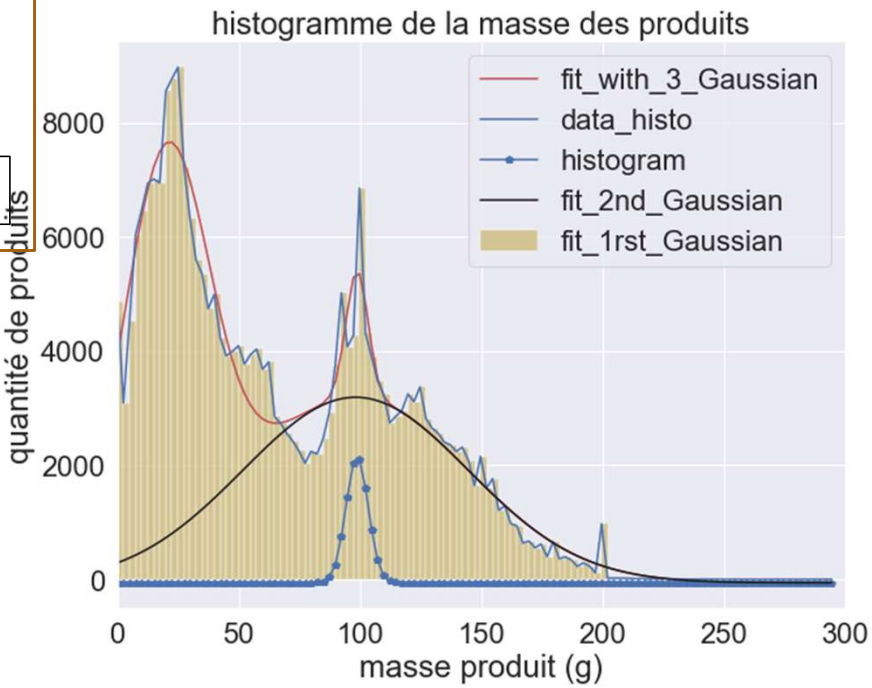
Outliers

5 - Vérification de 3 critères : $0 \leq m_n \leq 100$ (Median ou 1)

$$\sum_n m_n > 100g$$

$$\sum_n m_n \rightarrow 100$$

6 - Enfin, suppression des produits dont les indicateurs sont nulles



Opérations de nettoyage effectuées

nettoyage en 6 étapes

1 - Suppression des indicateurs entièrement NaN

2 – suppression des duplicata dans l'indicateur « code »

3 - Suppression des indicateurs non utilisés (colonne vide ou nulle)

4 - Suppression des lignes dont les indicateurs « non-identitaire » sont NaN

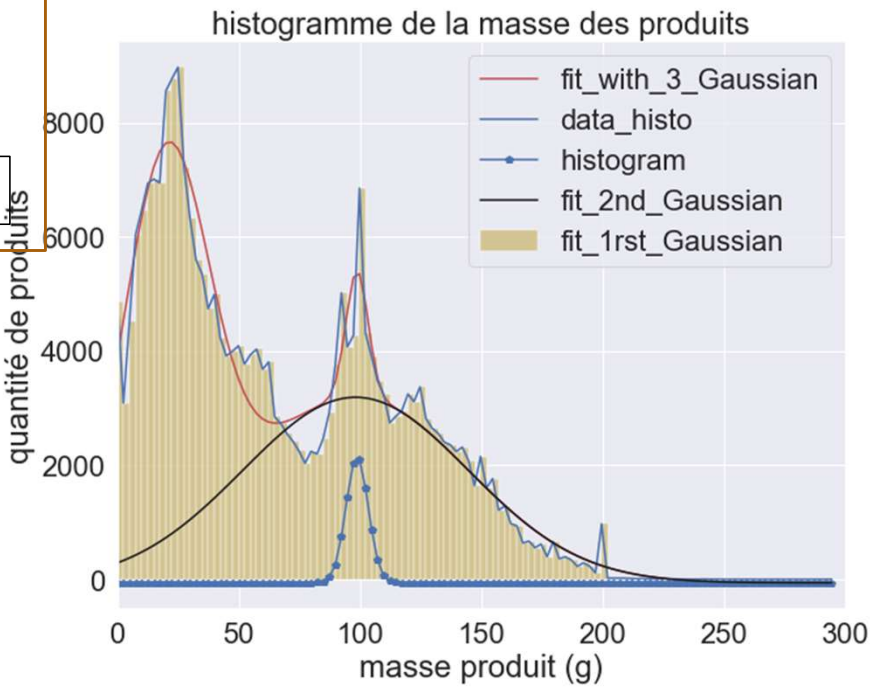
Outliers

5 - Vérification de 3 critères : $0 \leq m_n \leq 100$ (Median ou 1)

$\sum_n m_n > 100g$

$\sum_n m_n$ ~~> 100~~

6 - Enfin, suppression des produits dont les indicateurs sont nulles



Opérations de nettoyage effectuées

Pertinence d'une mise à l'échelle des masses de nutriments

Base de donnée de mauvaise qualité:

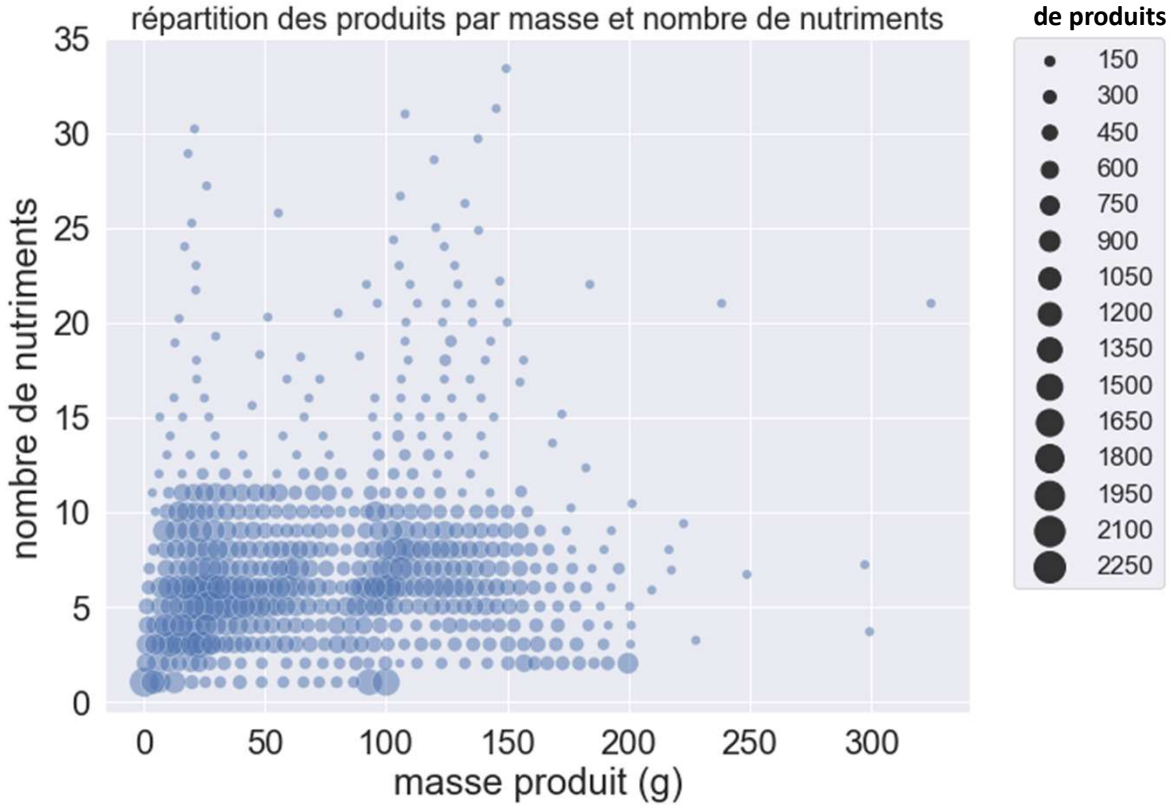
masse d'un produit pas systématiquement 100 g.

$$\sum_n m_n \rightarrow 100 \quad ?$$

***nota bene* : l'eau n'est pas prise en compte.**

Constat suite à l'analyse du graphe:

- **Quantité de nutriment entre 5 et 10 pour m~100g,**
- **Quantité de nutriment équivalente pour m> 100g,**
- **Certains produits, quantité de nutriment plus faibles pour m< 50g**



Opérations de nettoyage effectuées

Pertinence d'une mise à l'échelle des masses de nutriments

Base de donnée de mauvaise qualité:

masse d'un produit pas systématiquement 100 g.

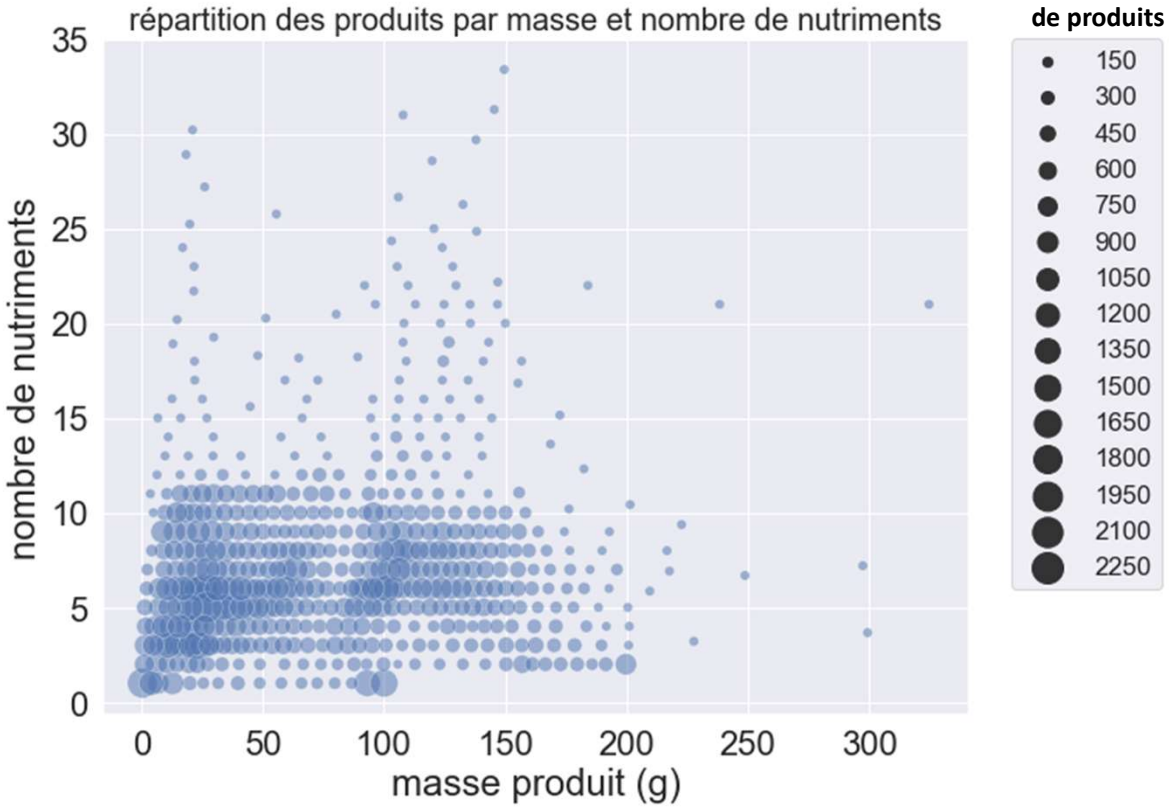
$$\sum_n m_n \rightarrow 100 \quad ?$$

nota bene : l'eau n'est pas prise en compte.

Hypothèse :

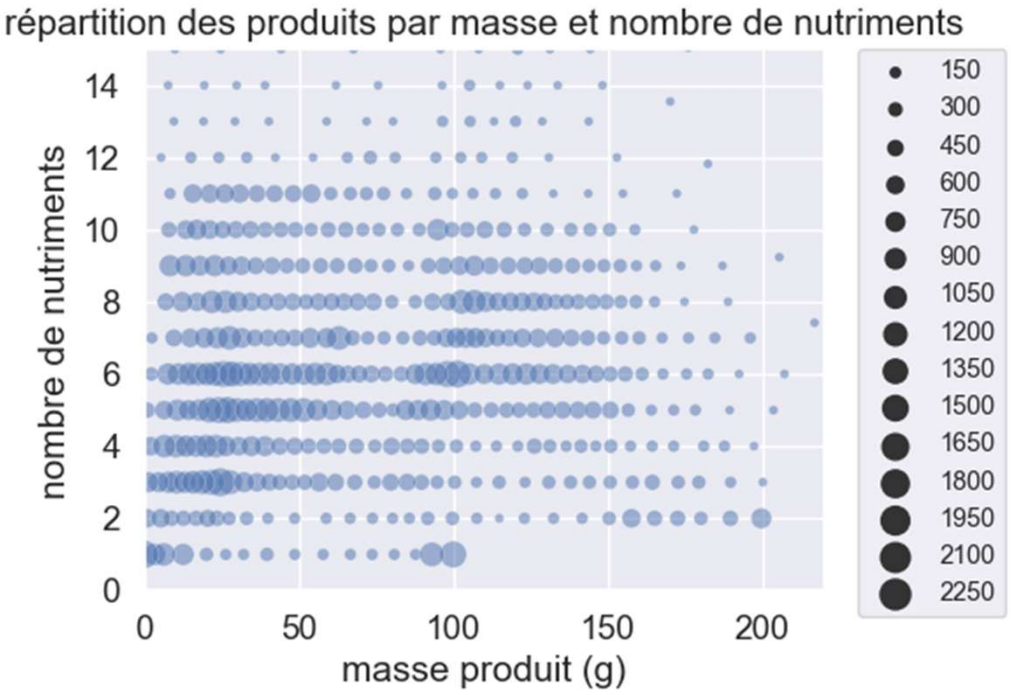
- **Les produits <50g et quantité de nutriment<5:**
  - tous les nutriments ne sont pas renseignés
  - On ignore ces produits
- **Les produits>100g sont conservés et rescalés**

Nouvelle dimension du dataset : (142489, 51)

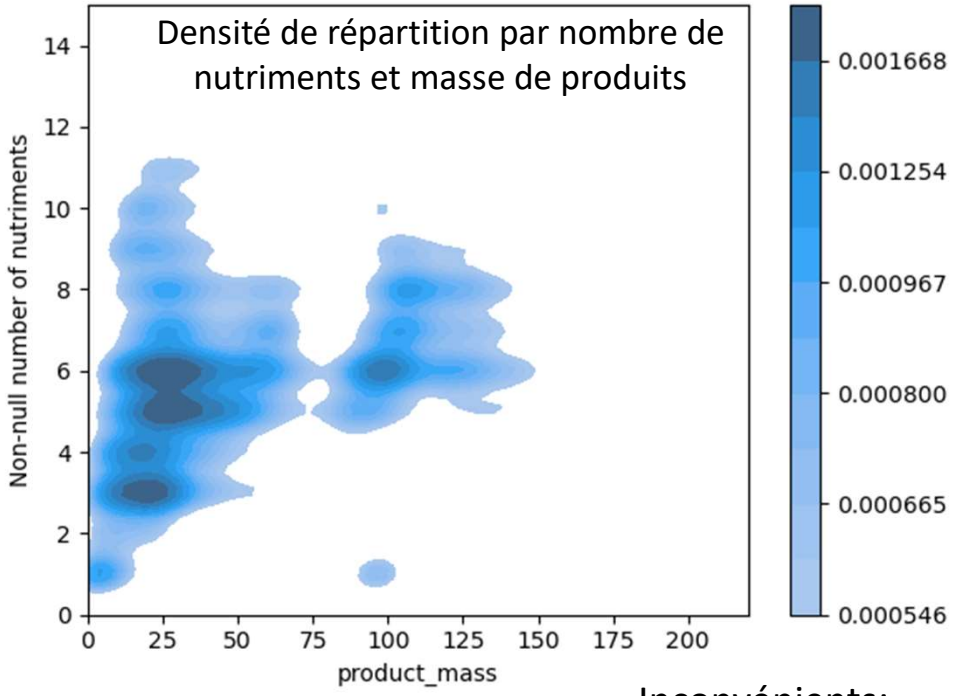


Opérations de nettoyage effectuées

Pertinence d'une mise à l'échelle des masses de nutriments



clustering+scatterplot



kdeplot sur (thresh 0.4)

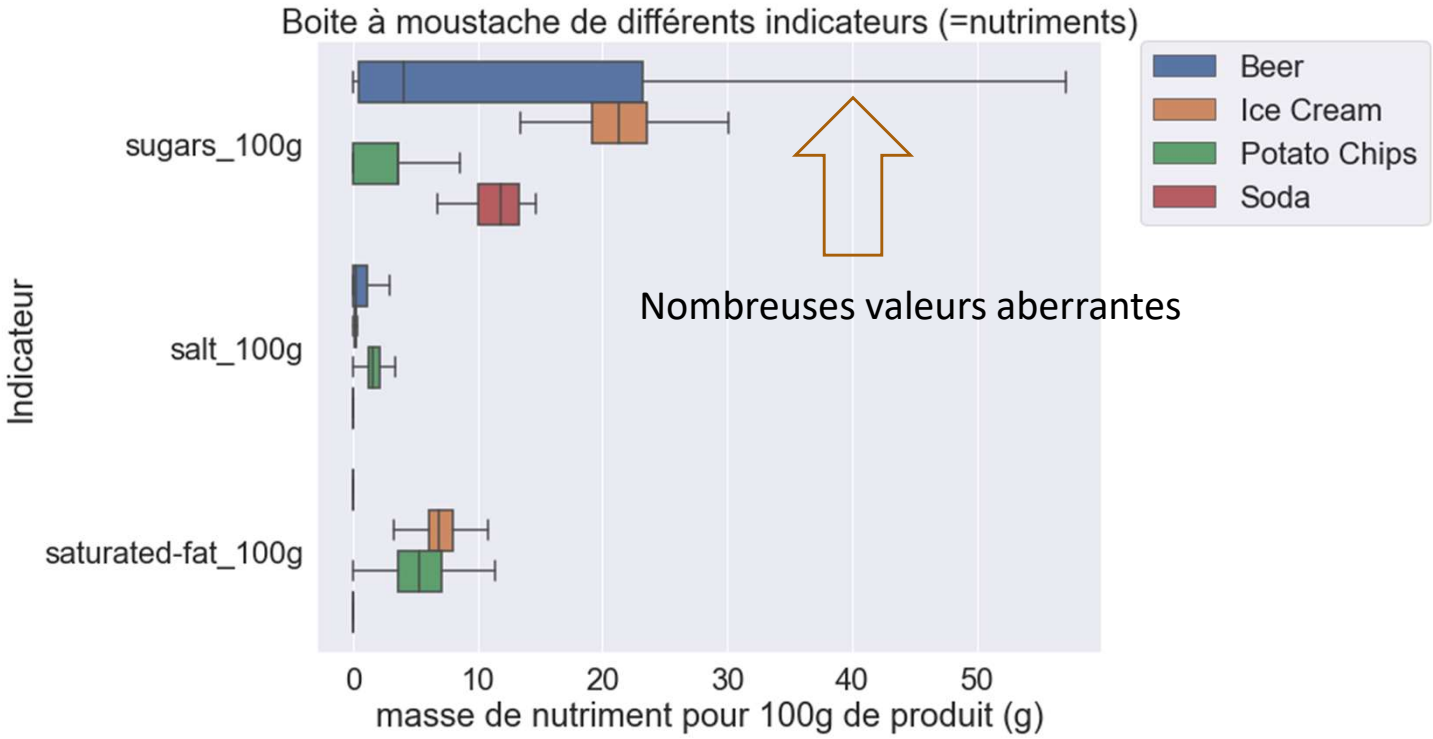
<u>Inconvénients</u>:

- lissage exagéré (Scott)
- 3x plus lent

Description/Analyse univariée

3 nutriments emblématiques sont étudiés : Sel, Sucre, Graisse

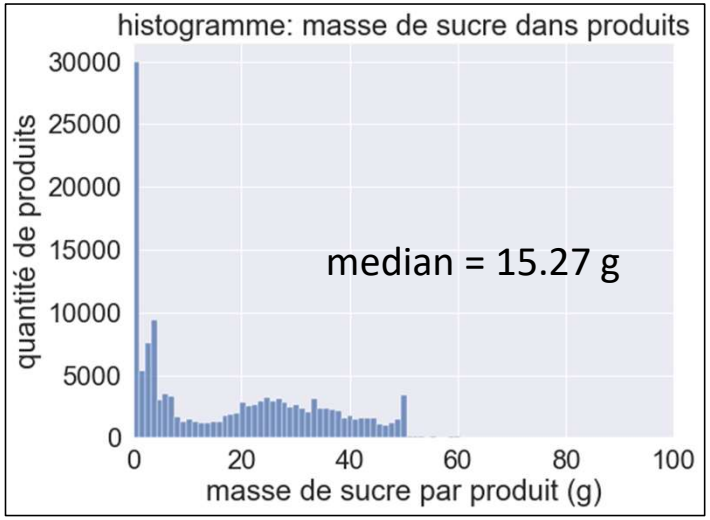
Application à 4 familles de produits:
Bière, Glaces, Chips, Soda

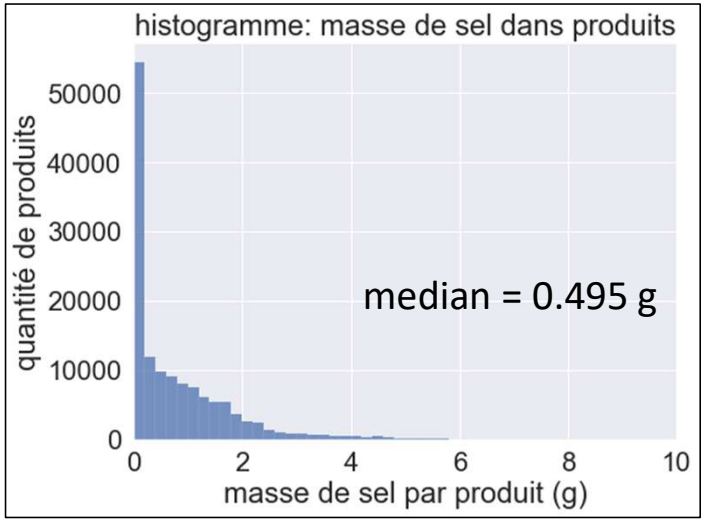


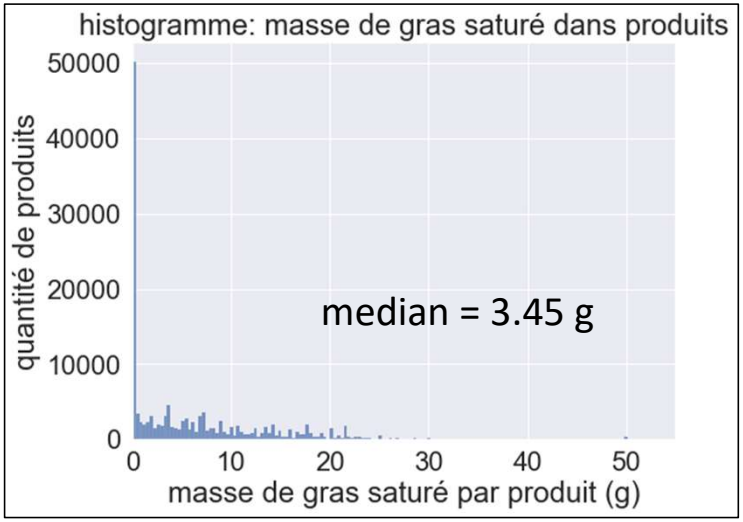
(sans outliers)

Description/Analyse univariée

3 nutriments emblématiques sont étudiés : Sel, Sucre, Acides Gras saturés







Les produits vendus en France sont très mal équilibrés.*
Nb un repas adulte représente ~600 à 900g.

Nutriments	Besoin journalier (g)
sucre	48*
sel	5**
Acide gras saturé	26***

*** https://www.cerin.org/wp-content/uploads/2017/01/symposium-dernieres-recommandations-lipides-theorie-assiette.pdf
** https://www.who.int/fr/news-room/fact-sheets/detail/salt-reduction
*https://www.coeuretavc.ca/vivez-sainement/saine-alimentation/reduire-le-sucre

Description/Analyse bi-variée

3 nutriments emblématiques sont étudiés : Sel, Sucre, Acides Gras saturés

Calcul de coefficient de correlation de Pearson*

Correlation coefficient

	saturated-fat_100g	sugars_100g	salt_100g
saturated-fat_100g	1.0	-0.095848	-0.079998
sugars_100g	NaN	1.0	-0.155291
salt_100g	NaN	NaN	1.0

p-value

	saturated-fat_100g	sugars_100g	salt_100g
saturated-fat_100g	0.0	0.0	0.0
sugars_100g	NaN	0.0	0.0
salt_100g	NaN	NaN	0.0

Les nutriments semblent être très légèrement corrélés mais c'est peu significatif.

On ne peut pas exploiter les p-values car ces nutriments ne suivent pas une loi normal

*attention aux conclusion hâtives concernant les probabilités, car les distributions, du sel, du sucre et des acides gras saturés ne suivent pas une loi normal:
C. J. Kowalski, "On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient" Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 21, No. 1 (1972), pp. 1-12

Description/Analyse multivariée

Etude du sel et du Sodium dans les chips.

Distribution Sodium et Salt/2.5 ne suivent pas une loi normale

```
statistic, pvalue
(44.71742031798914, 1.9486576890337548e-10)
```

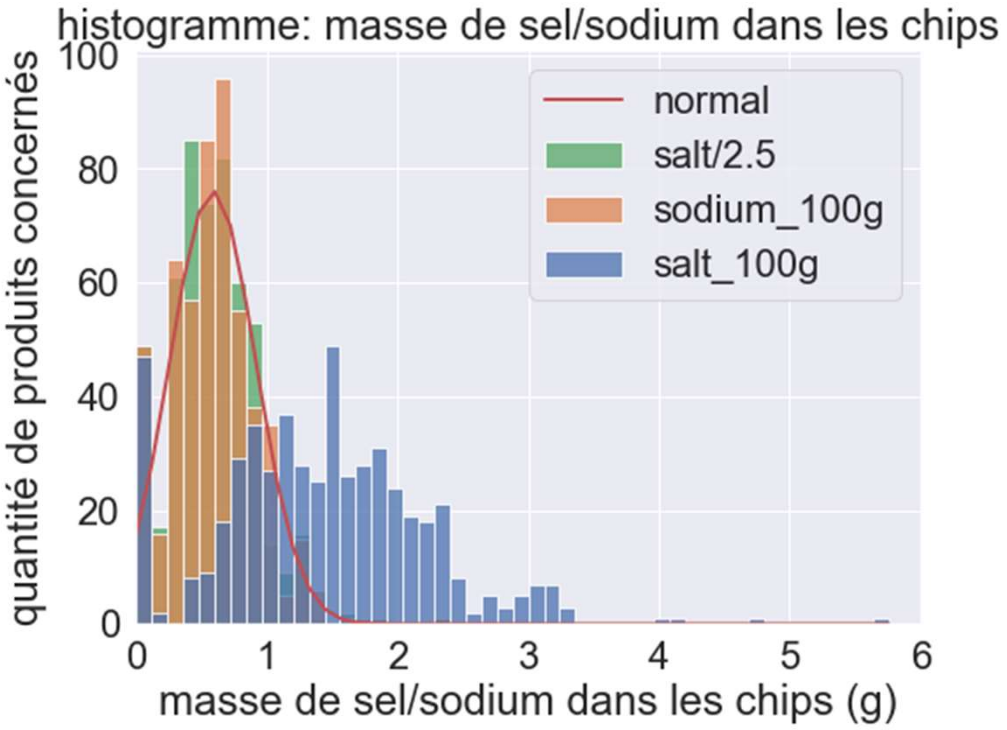
Distribution Sodium et Salt/2.5 sont semblables statistiquement

Test de Fisher

```
2 F, p
(0.5655770262787566, 0.45219072701908103)
```

Test de Kruskal Wallis

```
2 H, p
(0.4625779546903938, 0.4964216208081422)
```



Analyse multivariée

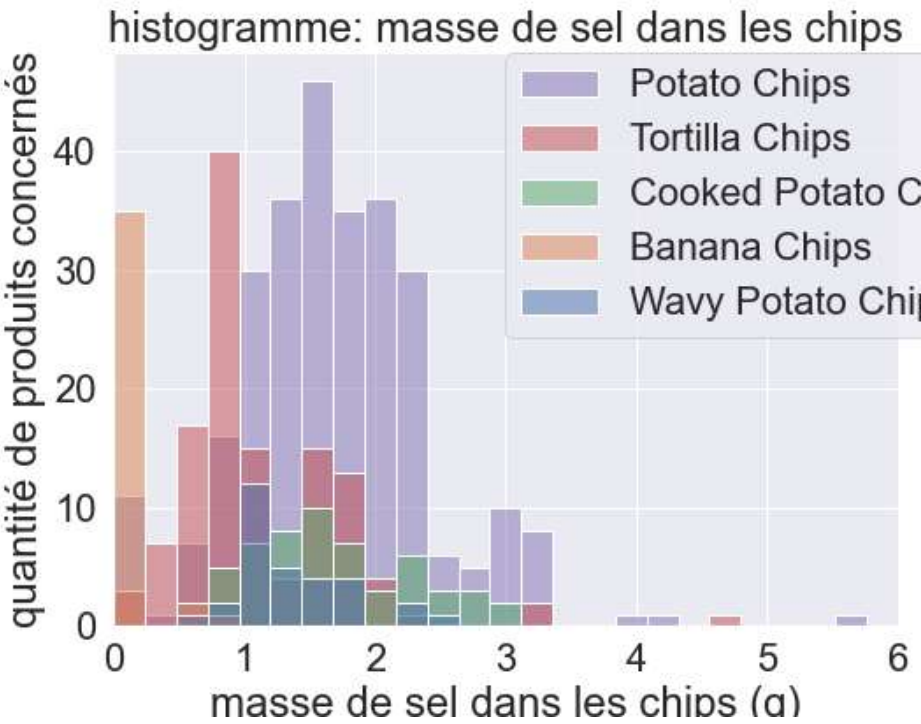
Répartition de la masse de sel /type de chips

Banana Chips : très peu de sel

Cooked Potato Chips => ~ loi normal

Test de Kruskal Wallis :
différentes teneurs en sel

```
(99.64863110359316, 1.849459797981158e-21)
```



Normal test

pvalue
8.672728e-11
9.385433e-17
2.597337e-01
7.884563e-11
4.338970e-02

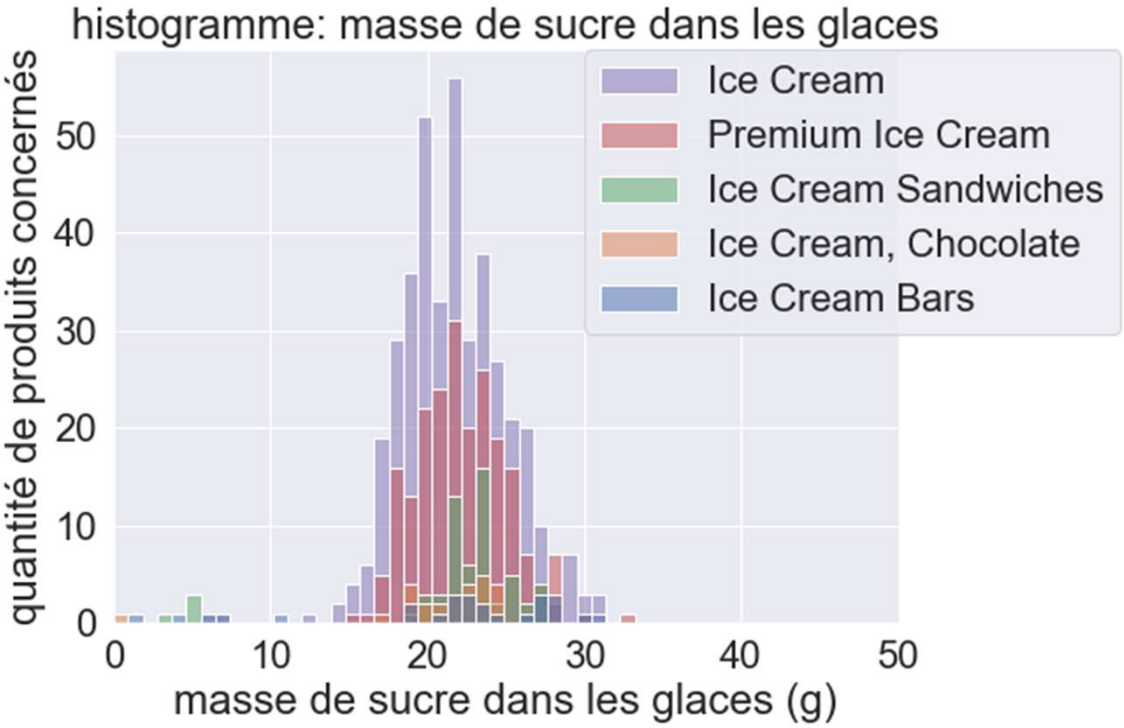
Analyse multivariée

Répartition de la masse de sucre /type de glace

Test de Kruskal Wallis :
différentes teneurs en sucre

(11.875060129258076, 0.018304941066859686)

	Ice Cream	Premium Ice Cream	Ice Cream Sandwiches	Ice Cream, Chocolate	Ice Cream Bars
Ice Cream	1.0	0.131241	0.001957	0.86639	0.162079
Premium Ice Cream	NaN	1.0	0.035894	0.387922	0.306382
Ice Cream Sandwiches	NaN	NaN	1.0	0.043459	0.844915
Ice Cream, Chocolate	NaN	NaN	NaN	1.0	0.24518
Ice Cream Bars	NaN	NaN	NaN	NaN	1.0



Normal test

pvalue
[0.0007127684620326549]
[0.05105676628567235]
[1.3373504515611441e-11]
[4.463791183883424e-11]
[0.039384027823134625]

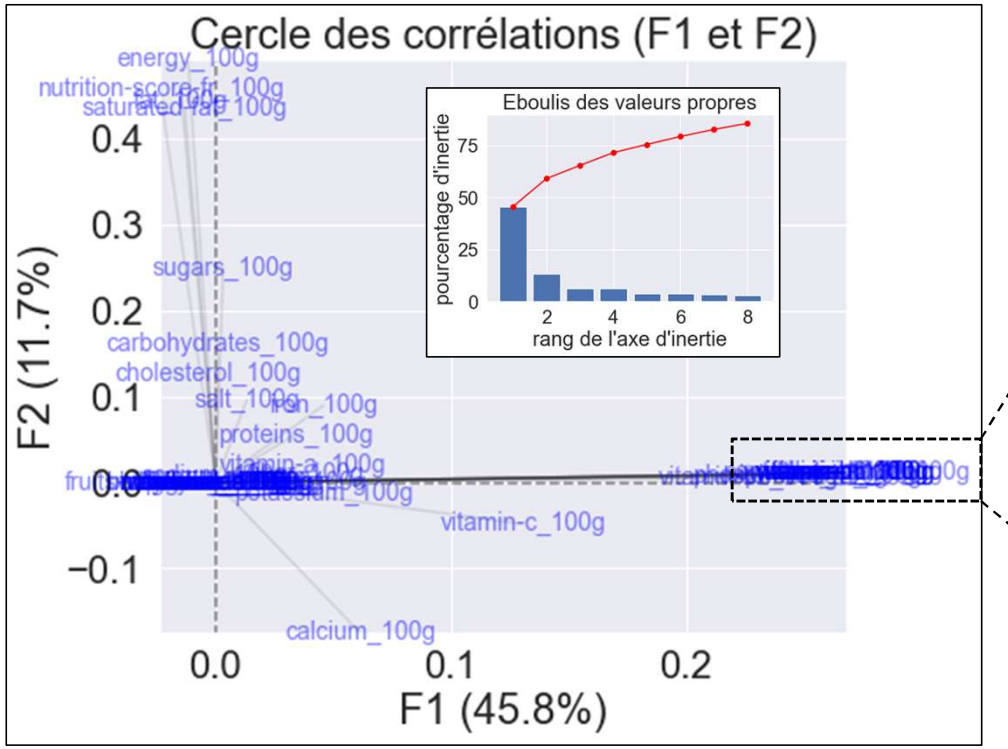
Analyse multivariée

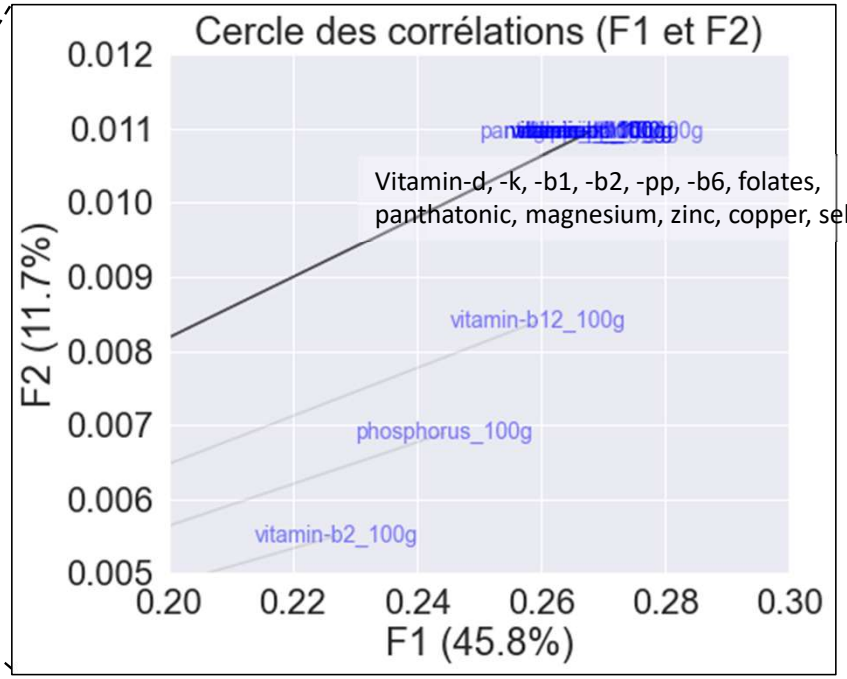
Répartition de la masse de sucre /type de glace

	Ice Cream	Premium Ice Cream	Ice Cream Sandwiches	Ice Cream, Chocolate	Ice Cream Bars
Ice Cream	1.0	0.023986	0.003809	0.787408	0.126425
Premium Ice Cream	NaN	1.0	0.104086	0.439182	0.312547
Ice Cream Sandwiches	NaN	NaN	1.0	0.102491	0.657069
Ice Cream, Chocolate	NaN	NaN	NaN	1.0	0.248698
Ice Cream Bars	NaN	NaN	NaN	NaN	1.0

Analyse multivariée

Composition des glaces



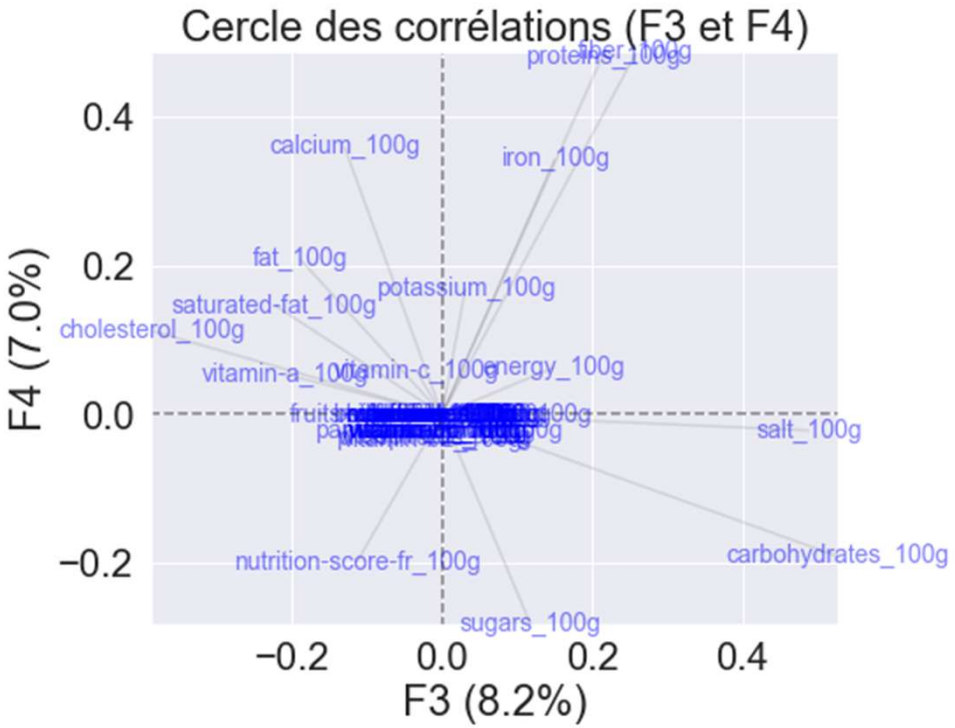


Axe secondaire : sucre, energy, …↗,

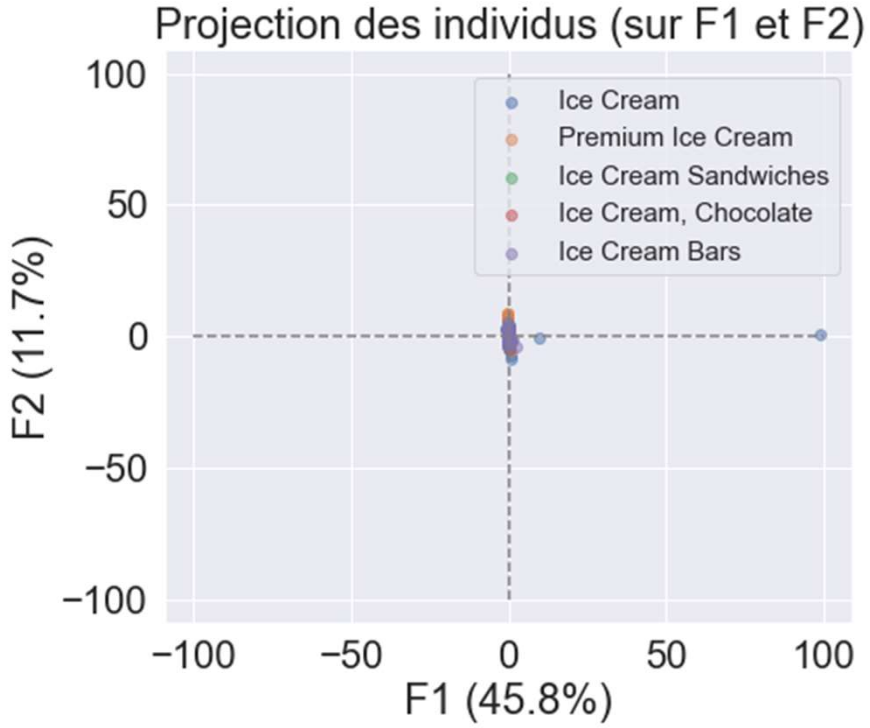
Axe principal : minéraux et vitamines

Analyse multivariée

Composition des glaces



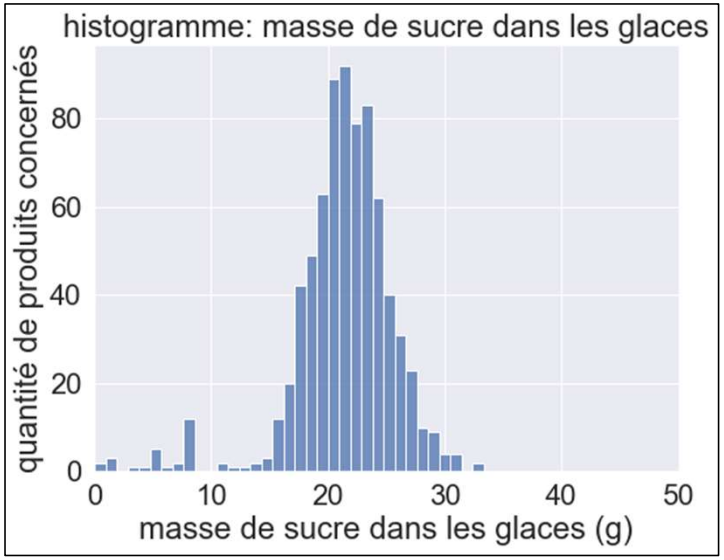
Axe 3: sel ↗

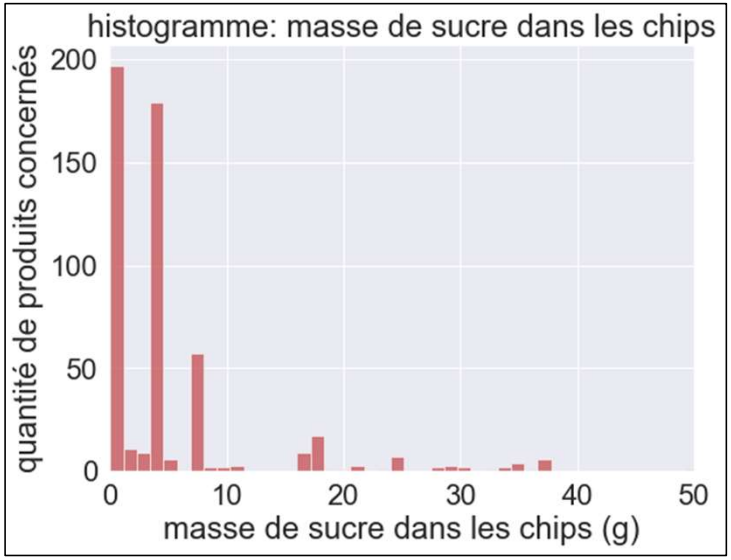


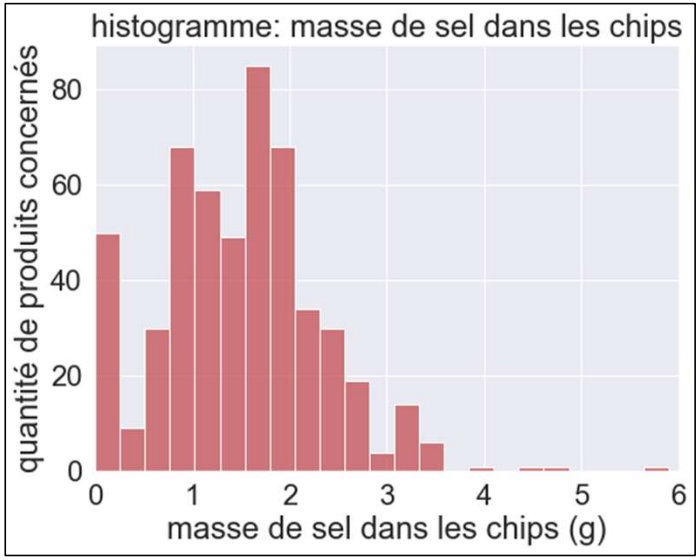
Glaces « Premium » et « IceCream »

Pertinence et Faisabilité

Application sur les chips et les glaces







La répartition d'un nutriment en fonction des produits d'une même famille.
Pertinence d'une application de sélection de produits en fonction d'un nutriment donné.

Attention : Data set pas toujours fiable (bière : sucre / masse produits)

Synthèse et Conclusion

Phase essentielle Nettoyage des donnés

Etude statistique sur les produits par différentes méthodes:

univarié, multivarié, ANOVA, histogram, barplot, moustache

Confirmation de la faisabilité d'une application pour prévénir les consommateurs de la grande quantité de certains nutriments dans les produits.