Data scientist
Projet 3

# Concevez une application au service de la santé publique

## BabyToo

Dolores Valide
Mentor : Siham Laaroussi

# Sommaire

# Appel à projet

Santé publique France

L'agence "Santé publique France" a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.

Pour répondre à cet appel à projet
On dispose de la base de donnée
OPEN FOOD FACTS.

OPEN FOOD FACTS
l'information alimentaire ouverte

# Idée d’application

“**BabyToo**”

- De quoi les parents ont besoin ?
- Qu’est-ce-que l’application fera?
- Quelles informations nous faut-il?

# Des parents qui accompagnent leur bébé dans leur diversification alimentaire.

**Problématique :**

- Parent pratiquant la DME (Diversification Menée par l'Enfant) et à la recherche de produits qu'il pourra préparer pour toute la famille.
- Nombreux produits spécialisés pour les bébés à des tarifs plus élevés que les produits de base.
- Composition des produits pas toujours clean.
- Etiquettes difficiles à déchiffrer

**Avantage :**

- **économiser** en partageant les mêmes produits pour toute la famille.
- plus besoin d'acheter forcément dans le rayon bébé
- **toute la famille** mange mieux grâce au BaBy score
- suivi des allergènes et substances

**Public cible :**

- Marché Français
- Nouveaux parents
- Bébés de 6 mois à 2 ans et demi env

# Que fait l'application ?

BaByToo

**AJOUT DES DONNÉES**

- âge
- allergies connues
- allergies familiales
- régime spéciale

**DURANT LES COURSES QUOTIDIENNES**

**SCAN PRODUIT**

00000 00001

**BÉBÉ SCORE**

Produit bien noté :

score 1, 2, 3

- Valeurs nutritionnelles
- Formats adaptées pour bébé
- Liens vers des recettes

Produit mal noté :

score 4 et 5

Recherche dans Open Foods Facts d'un produit proche et mieux noté

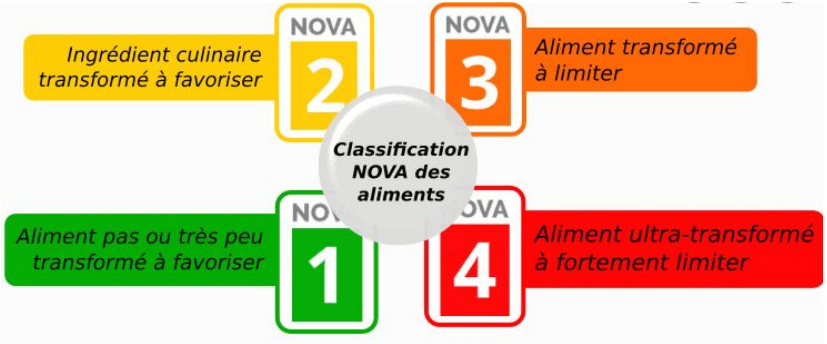
Santé publique France

# Quelles informations nous faut-il?

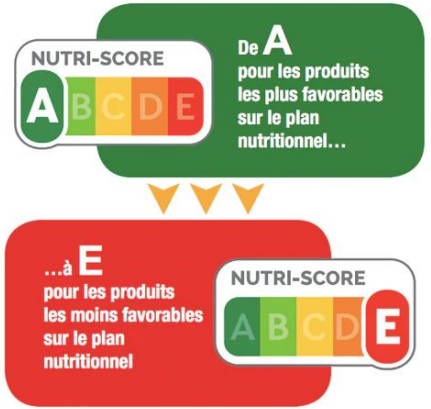
Valeurs nutritionnelles :

- sel
- glucides
- sucres
- énergie
- lipides

Score NOVA :




Nutri Score :




★ Le nombre d'additifs
★ Le nombre de produits issus de l'huile de palme

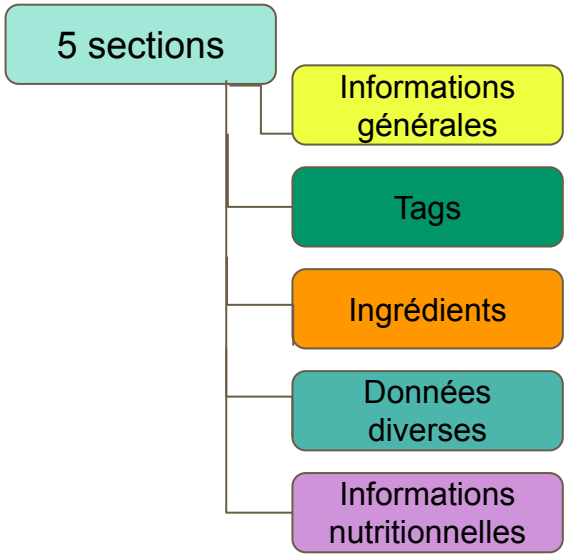
★ Tous les produits vendus en France
★ Photo des produits
★ Catégorie
★ Marque

Santé publique France

# Nettoyage des données

- Présentation du jeu de données
- Les étapes du nettoyage
- Traitement des outliers
- Traitement des valeurs manquantes

# Présentation du jeu de données




Nombre de données :

```
Entrée [8]: open_food_en.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2059358 entries, 0 to 2059357
Columns: 187 entries, code to carnitine_100g
dtypes: float32(115), int64(2), object(70)
memory usage: 2.0+ GB
```

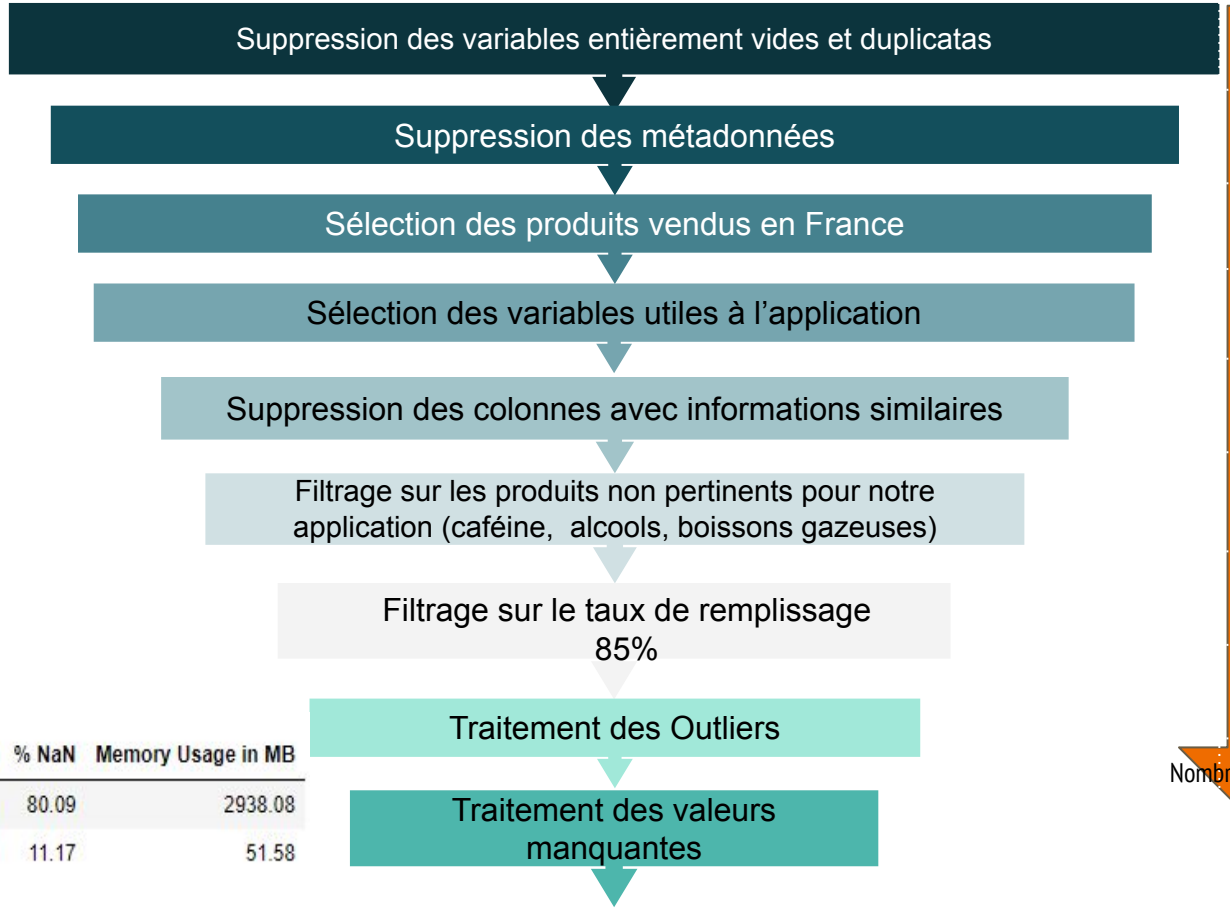
Nous avons un total de 2 M de produits

Nombre de valeurs manquantes :

```
get_numbers_missing(open_food_en)

Il y a 307352267 de valeurs manquantes (NaN) pour un total de 385099946 données soit (79.81 %)
```

# Les différentes étapes du nettoyage



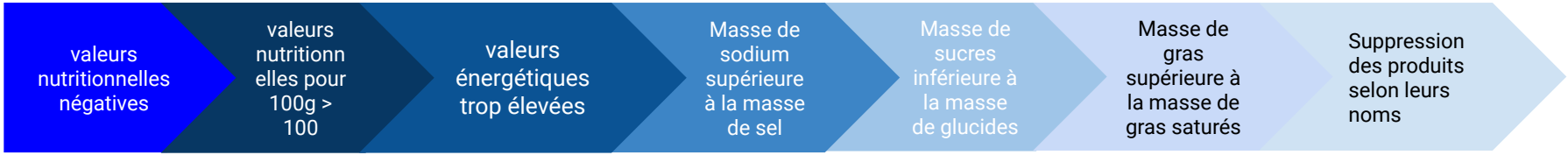

| | CSV | Shape | % NaN | Memory Usage in MB |
|---|---|---|---|---|
| 0 | initial | (2059358, 187) | 80.09 | 2938.08 |
| 1 | final | (307292, 23) | 11.17 | 51.58 |

Santé publique France

# Traitement des outliers

*valeurs remplacées par Nan*

*Produits supprimés*




- nombre : 3

- nombre : 34

Le maximum de l'énergie est de 900 kcal et 3768 kJ.

- nombre : 676
- méthode IQR

Pour effectuer la conversion entre sodium et sel, il faut multiplier la quantité de sodium par 2,5.

- nombre : 16

La proportion de sucres ne peut être supérieure à la proportion de glucides.

- nombre : 182

La masse de graisses saturées ne peut pas être supérieure à la masse

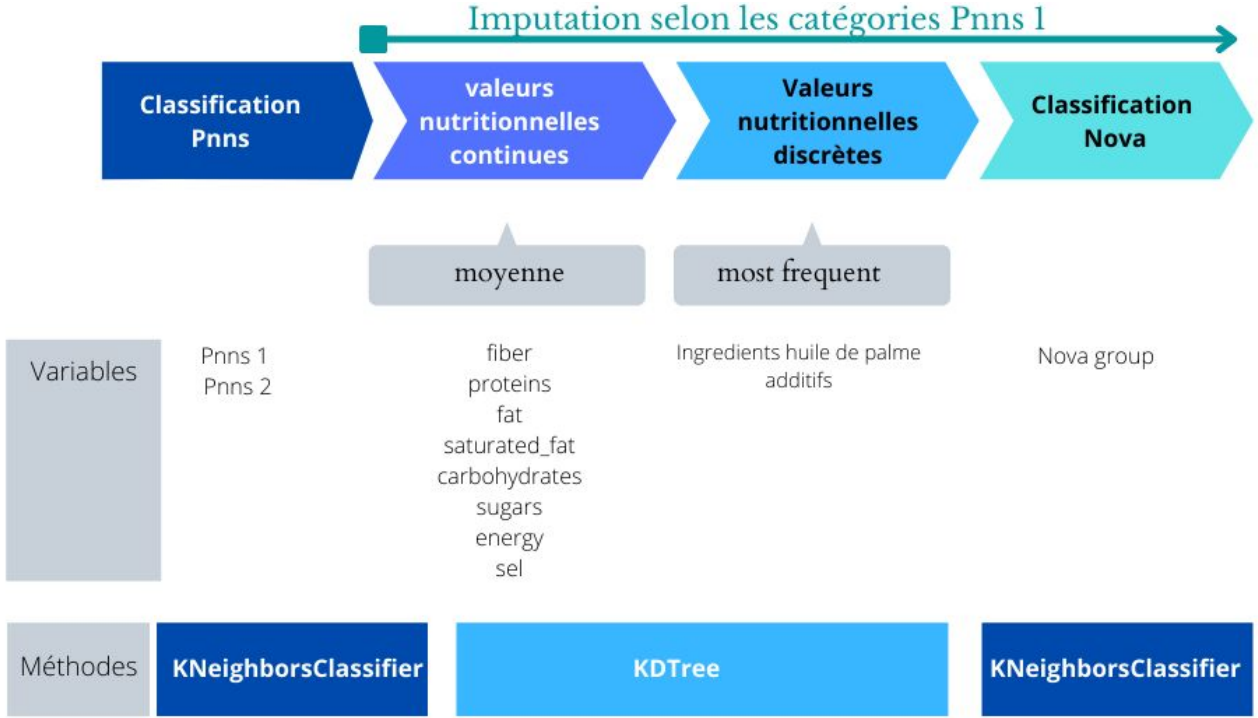
- nombre : 215

Les produits avec des noms comportant moins de 3 lettres

- nombre : 13

Santé publique France

# Imputation des valeurs manquantes




Nombre de Nan final :

```
get_numbers_missing(data_final)
```

```
Il y a 789552 de valeurs manquantes (NaN) pour un total de 7067716 données soit (11.17 %)
```

Santé publique France

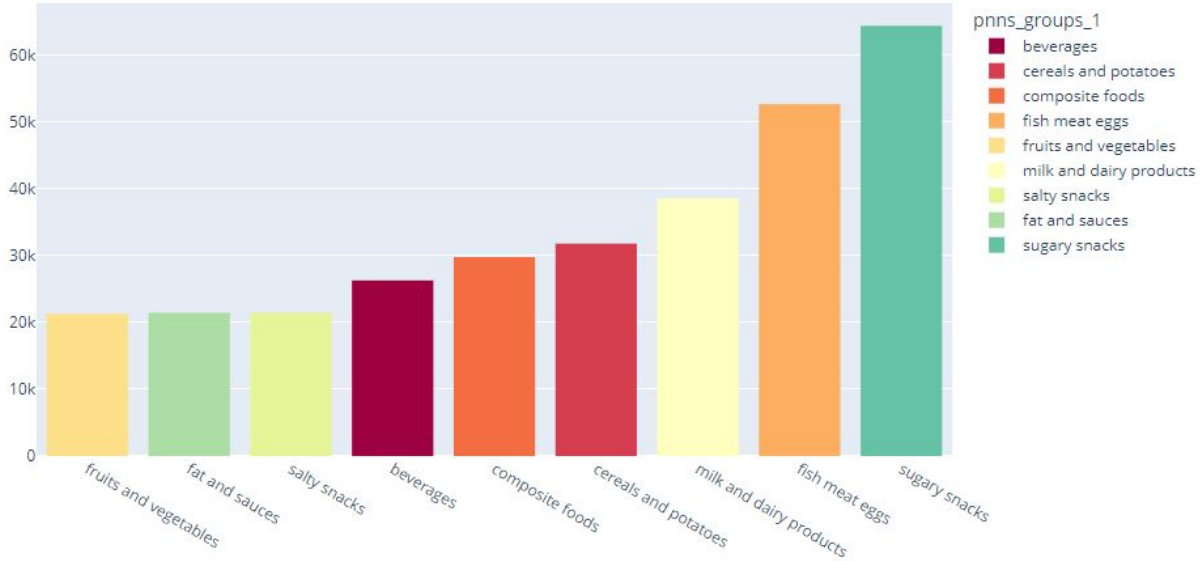
# Analyse des données

- Classification des produits
- Analyse des scores
- Analyse des valeurs nutritionnelles
- Analyses multivariées
- Analyse en composantes principales
- Analyse de variance

# Classification des produits répertoriés

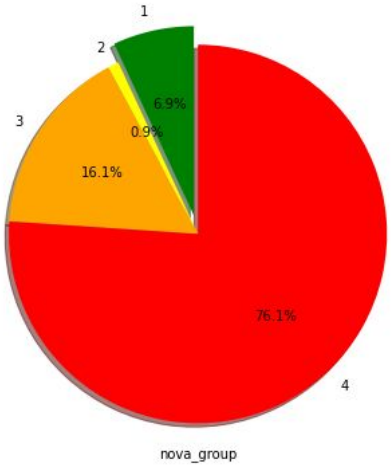
Une majorité de snacks sucrés sont répertoriés.

Contre 20K pour les catégories plus “saines”.

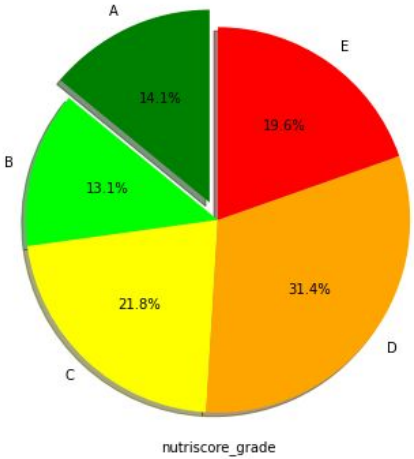



Santé publique France

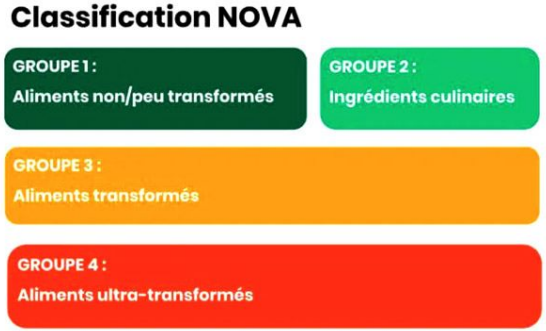
# Une majorité de “mauvais” scores

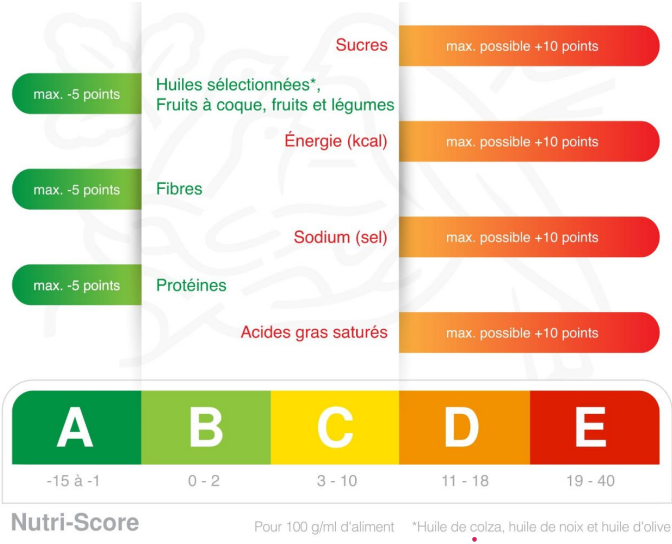



Près de 80% de produits ultra transformés.



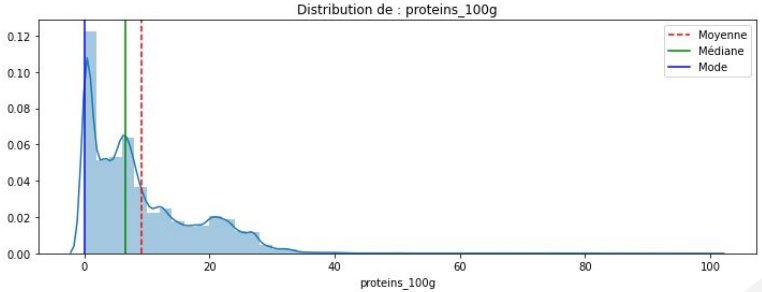

On constate que près de la moitié des produits ont un nutriscore E ou D.

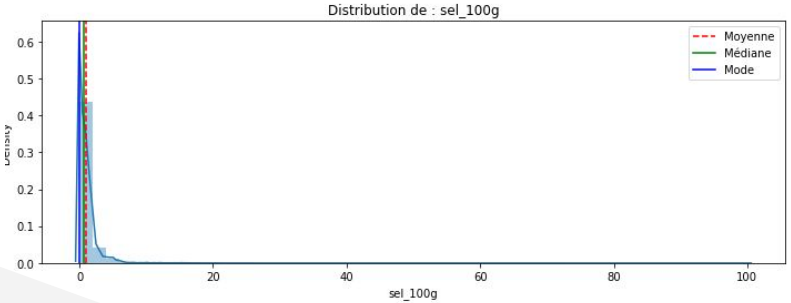


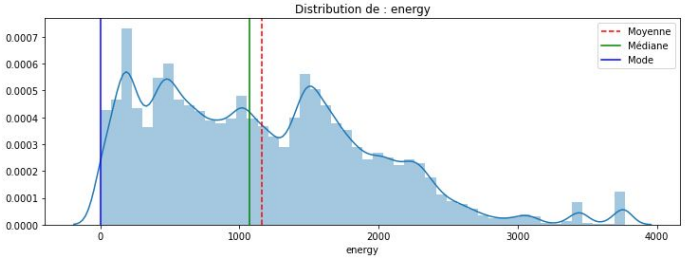

Santé publique France

# Des distributions semblables

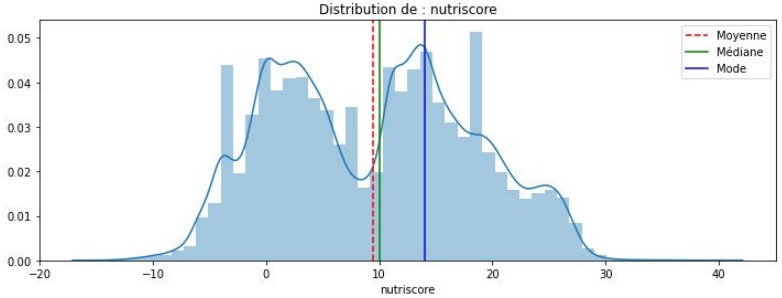



Distributions unimodales, asymétriques à droite. Proche de 0. Pas de loi normale



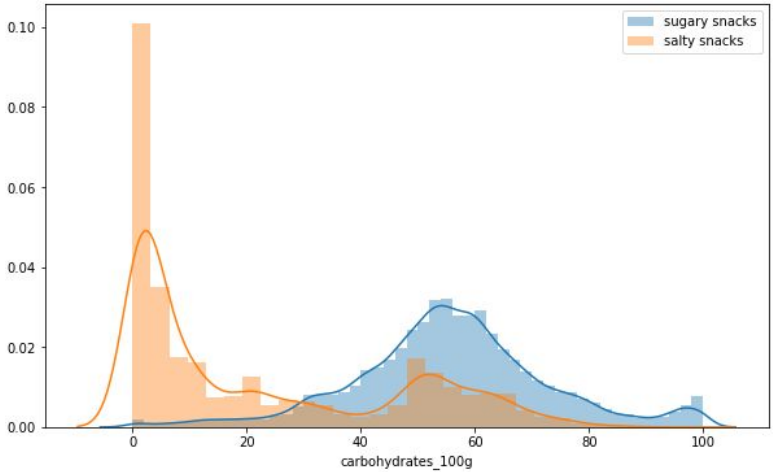

Distribution multimodale aplatie

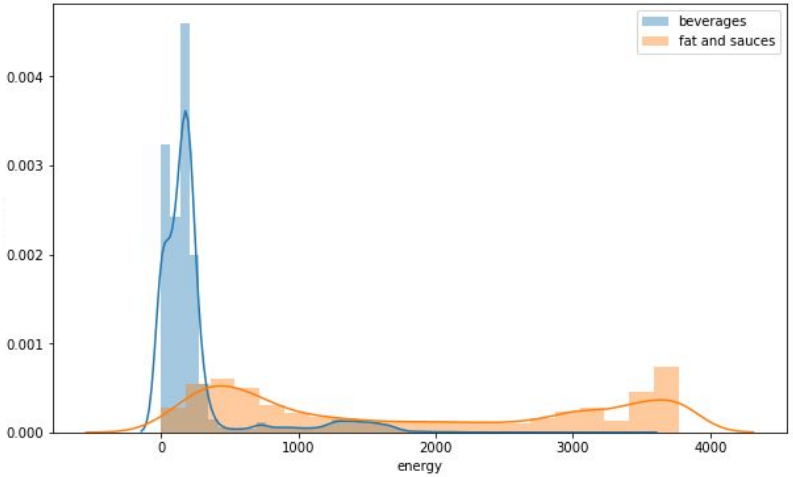



Distribution bimodale en 0 et 15

Santé publique France

# Différentes distributions pour différentes catégories




La distribution des glucides dans les snacks salé est concentrée en 0 et asymétrique vers la droite.

La distribution des glucides dans les snacks sucré est aplatie et centrée vers 55 et légèrement asymétrique vers la droite.

La distribution de l'énergie dans les boissons est concentrée en 0 et légèrement asymétrique vers la droite.

La distribution de l'énergie dans aliments gras et les sauces est étalée bimodale en 500 et 3700.

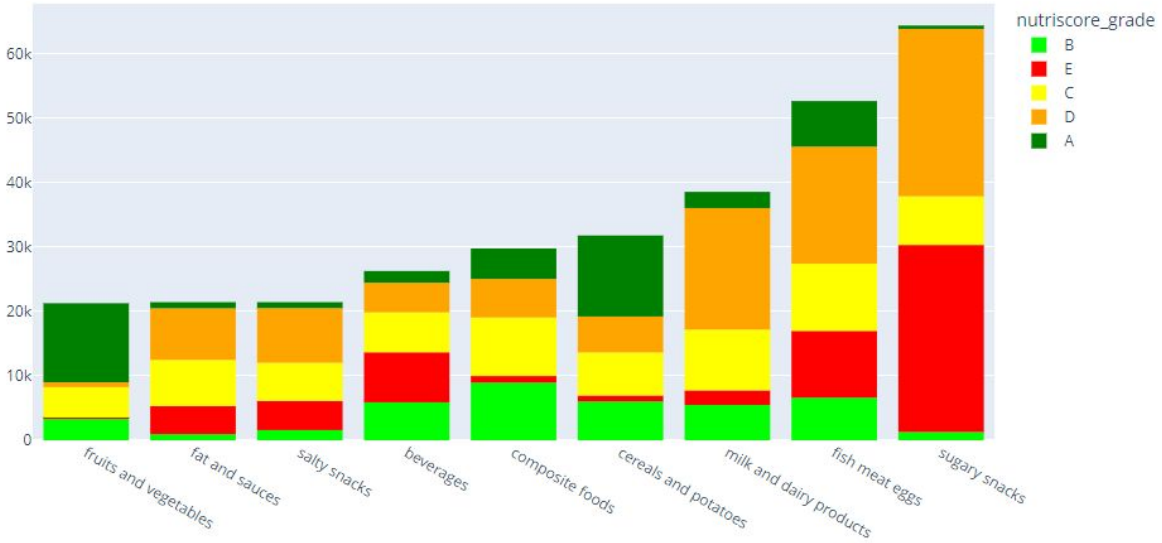
Santé publique France

# Répartition des scores par catégories

Les catégories :

- fruits et vegetables
- cereals and potatoes
- fish meat egg
- composite foods

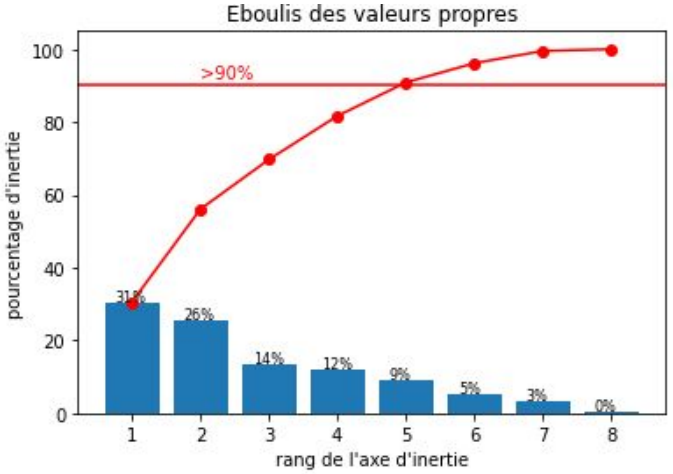
ont le plus de produits bien classés.



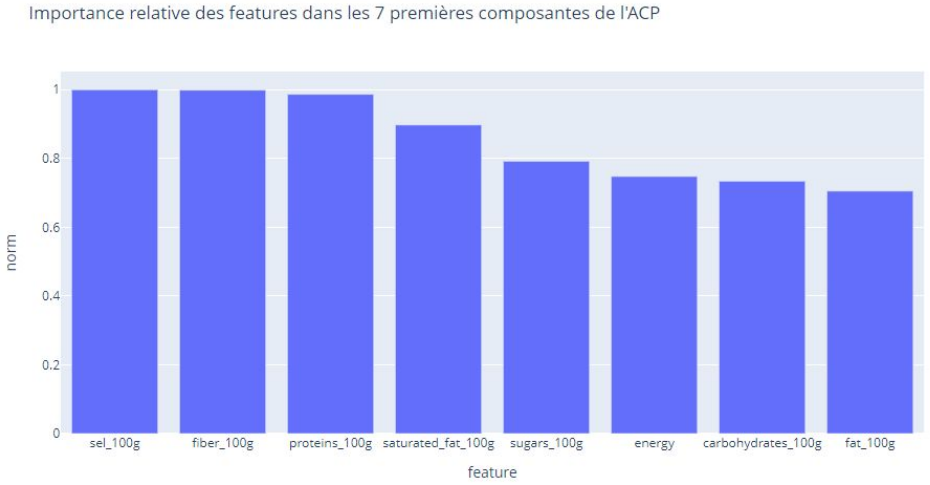
# Analyse en Composantes Principales

Peut-on réduire notre nombre de variables?

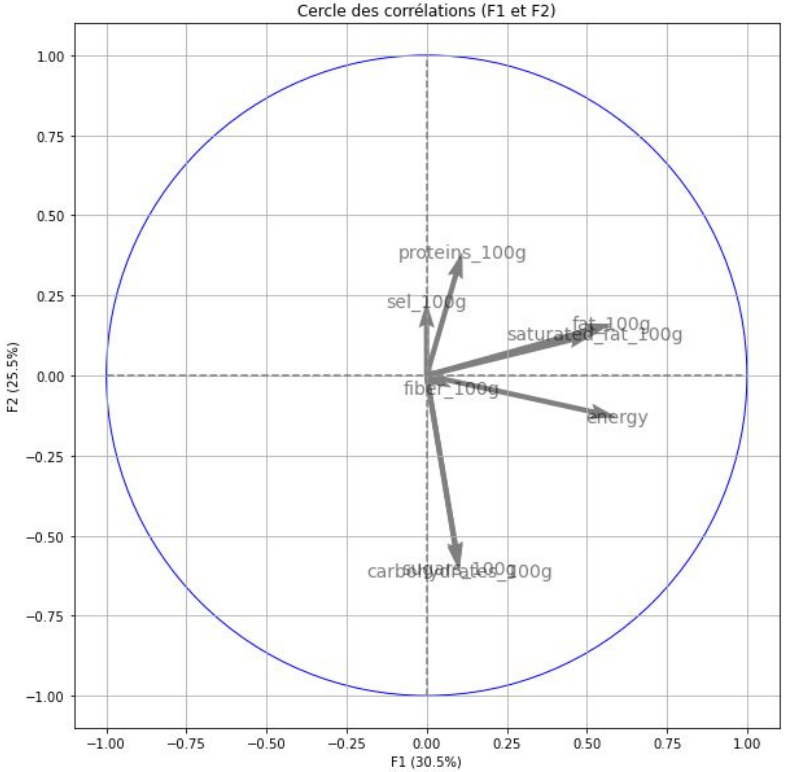
Si oui, combien de variables sont nécessaires pour décrire notre jeu de donnée?




Réduction possible à 4 composantes avec près de 90% des données.



# ACP : Cercle des corrélations




- **F1** représente les lipides donc l'énergie dans un produit.
- **F2** représente les produits légèrement salés et riches en protéines.
- **F3** qualifie les produits riches en protéines végétales.
- **F4** la richesse en sel des produits.

# Analyse de la variance : ANOVA

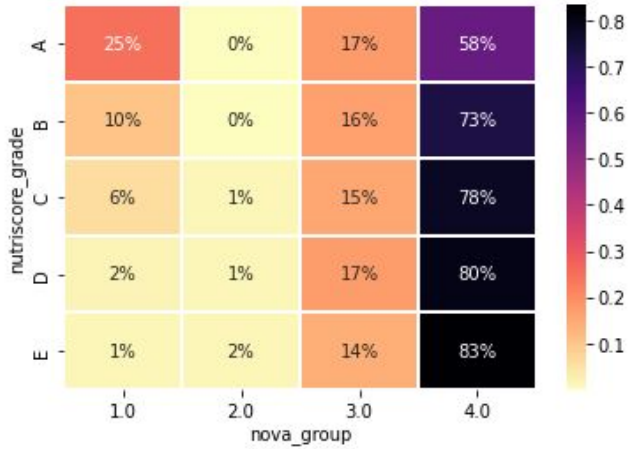
```
-----------------------------------------------------------
test de normalité de sel_100g, p = 0.0
-----------------------------------------------------------
The null hypothesis for sel_100g can be rejected
-----------------------------------------------------------
```

Aucune de nos variables ne suit une loi normale donc il n'y a pas d'intérêt à réaliser une ANOVA.

Une ANOVA est réalisé entre le nutriscore-score et le score Nova .
Y-a-t-il une dépendance entre ces variables ?

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| nova_group | 1.949518e+06 | 3.0 | 9168.00877 | 0.0 |
| Residual | 2.178094e+07 | 307288.0 | NaN | NaN |

La p-value est égale à 0 il y a bien une dépendance entre les deux variables.




Santé publique France

# Faisabilité de l’application

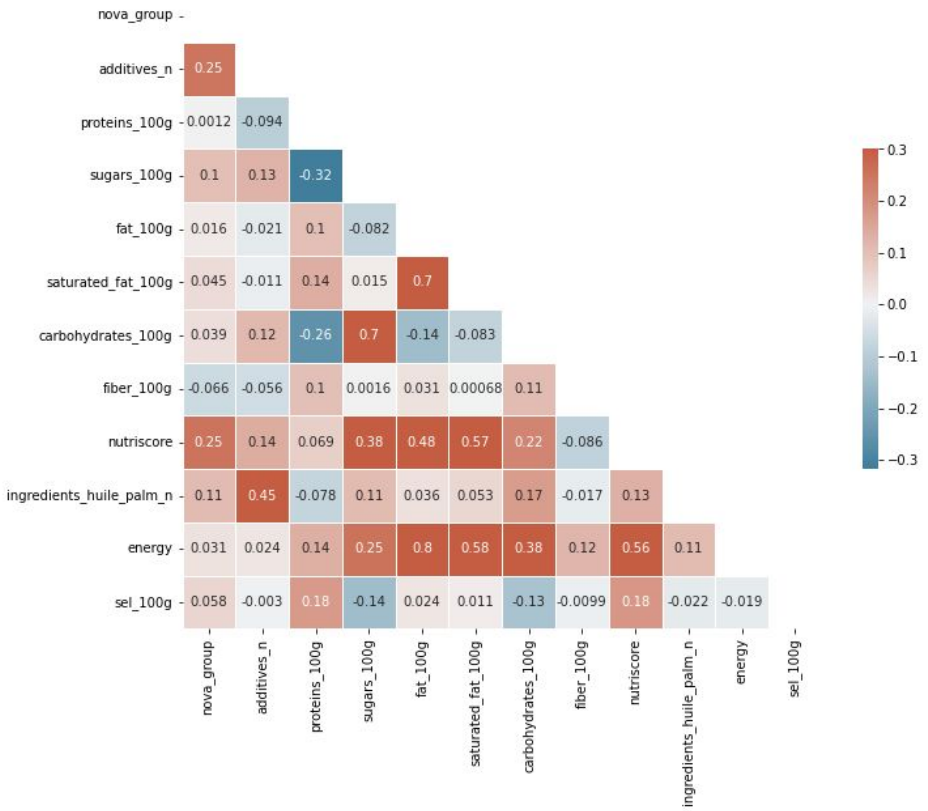
# Le bébé score est-il réalisable ?

Toutes les données nutritionnelles sont plus ou moins dépendantes (matrice de corrélation)

Si on se cantonne aux scores Nova = 1,2,3 et Nutriscore =A,B,C il ne nous reste que 45000 produits.

Les catégories de produits utilisés dans l'alimentation d'un bébé sont disponibles et comportent suffisamment de "bon scores"

Un score découlant des variables est possible



# Conclusions

## Analyse critique :

La base de données est alimentée par le public donc les erreurs de renseignements sont courantes.

Le nombre de produits correspondants à nos critères est faible. Mais le nettoyage était large.

Les imputations peuvent être améliorées.

## Points à explorer :

Une vérification préalable ainsi qu'un arrangement des données serait préférable.

Un algorithme de calcul du nutriscore.

Un algorithme de calcul du nova score.

Santé publique France

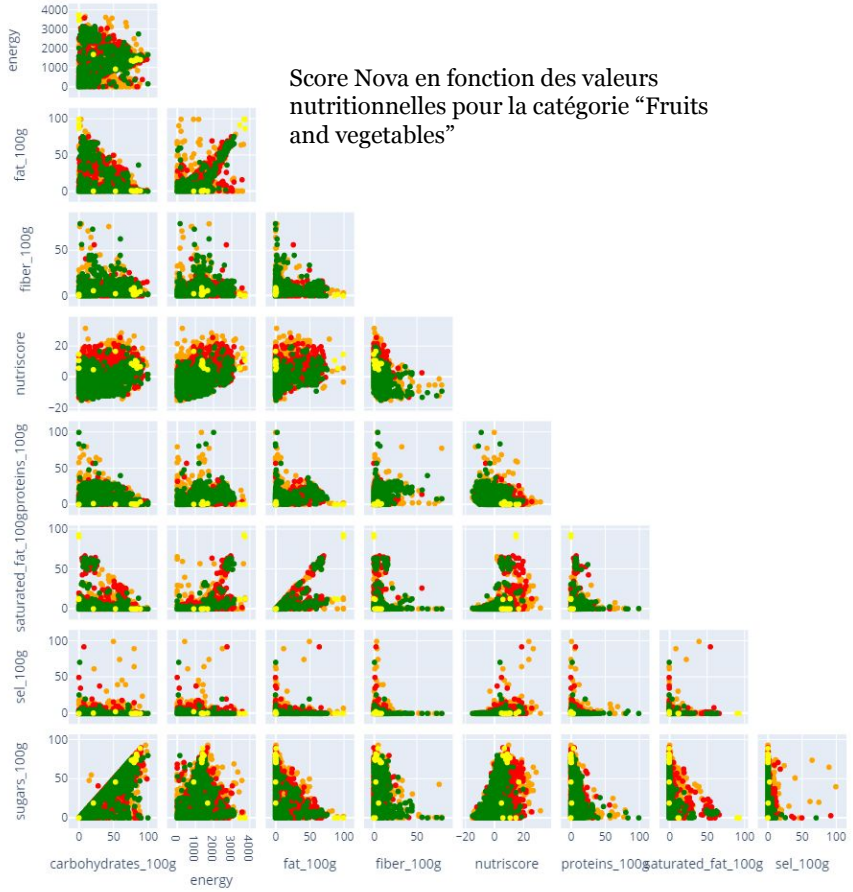
# Merci de votre attention

# Questions - Réponses

# Annexe

# Nombreux produits peu transformés dans certaines catégories

Score Nova en fonction des valeurs nutritionnelles pour la catégorie “Fruits and vegetables”




Score Nova en fonction des valeurs nutritionnelles pour la catégorie “cereals and potatoes”