Concevez une application au service de la santé publique

Santé publique France

Philippe Moty

SOMMAIRE

- Contexte
  - Rappel de l’appel à projets et explication de votre idée d’application (2 mn)
  - L’application
- Les données
  - Votre démarche méthodologique de nettoyage (8 mn)
  - Votre démarche méthodologique d’exploration de données (8 mn)
- La faisabilité
  - En synthèse, présentation des faits pertinents pour l’application (2 mn)

CONTEXTE

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. Proposer une idée d’application.

OPEN FOOD FACTS

Open Food Facts est une base de données de produits alimentaires qui répertorie les ingrédients, les allergènes, la composition nutritionnelle et toutes les informations présentes sur les étiquettes des aliments.

L’application

Constat

OpenfoodFacts est une application qui fournit beaucoup d'éléments et la navigation peut être complexe :

- Code-barres, Quantité, Conditionnement, Marque, Catégories

, Label, certification, récompenses, origine des ingrédients, Pays de vente,

- Openfoodfacts propose des indicateurs :
  - Nutri-Score : indicateur de qualité nutritionnelle
  - NOVA groupe : indicateur de transformation du produit
  - Eco-Score : impact environnemental
- Une section Santé :
  - Ingrédients
    - Nombre d'ingrédients, photographie de la composition
  - Les additifs
  - L'Analyse des composants
  - Tableau nutritionnel
- Empreinte carbone
- Le transport par rapport au lieu de fabrication
- Source de données
- On peut même classer les aliments en fonction de nos préférences

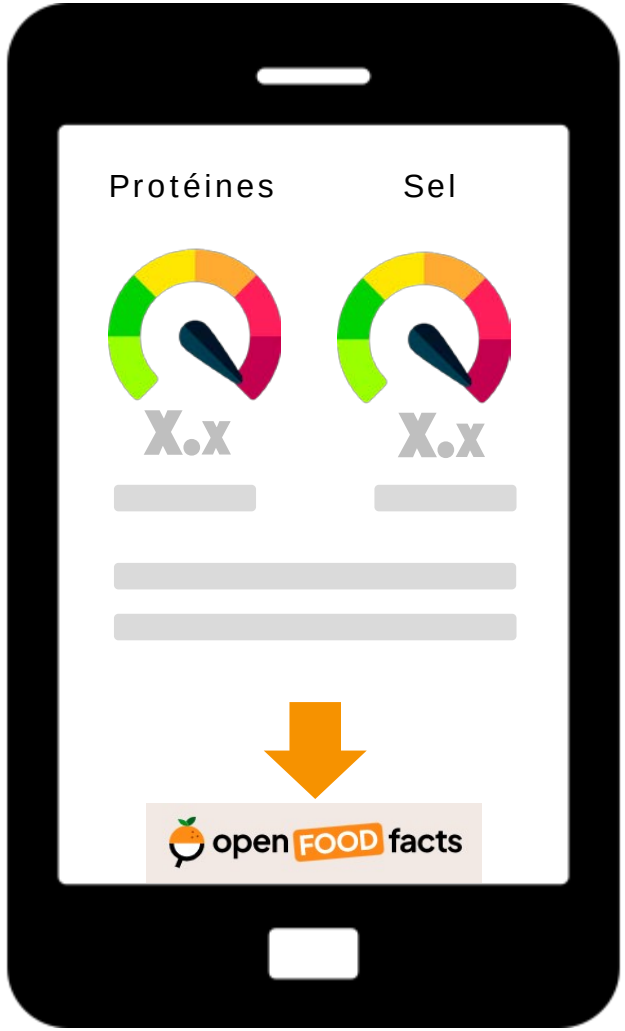
Comment protéger mon rein ?

- L’application ne remplace pas les conseils des professionnels, elle apporte une visibilité des données en lien avec la surveillance du rein en fonction des indicateurs les plus importants à surveiller en cas de maladie rénale.
- Les indicateurs : Sel, Proteines
- Les additifs
- Les valeurs nutritionnelles

Mon Assistant Rénal

Philippe Moty

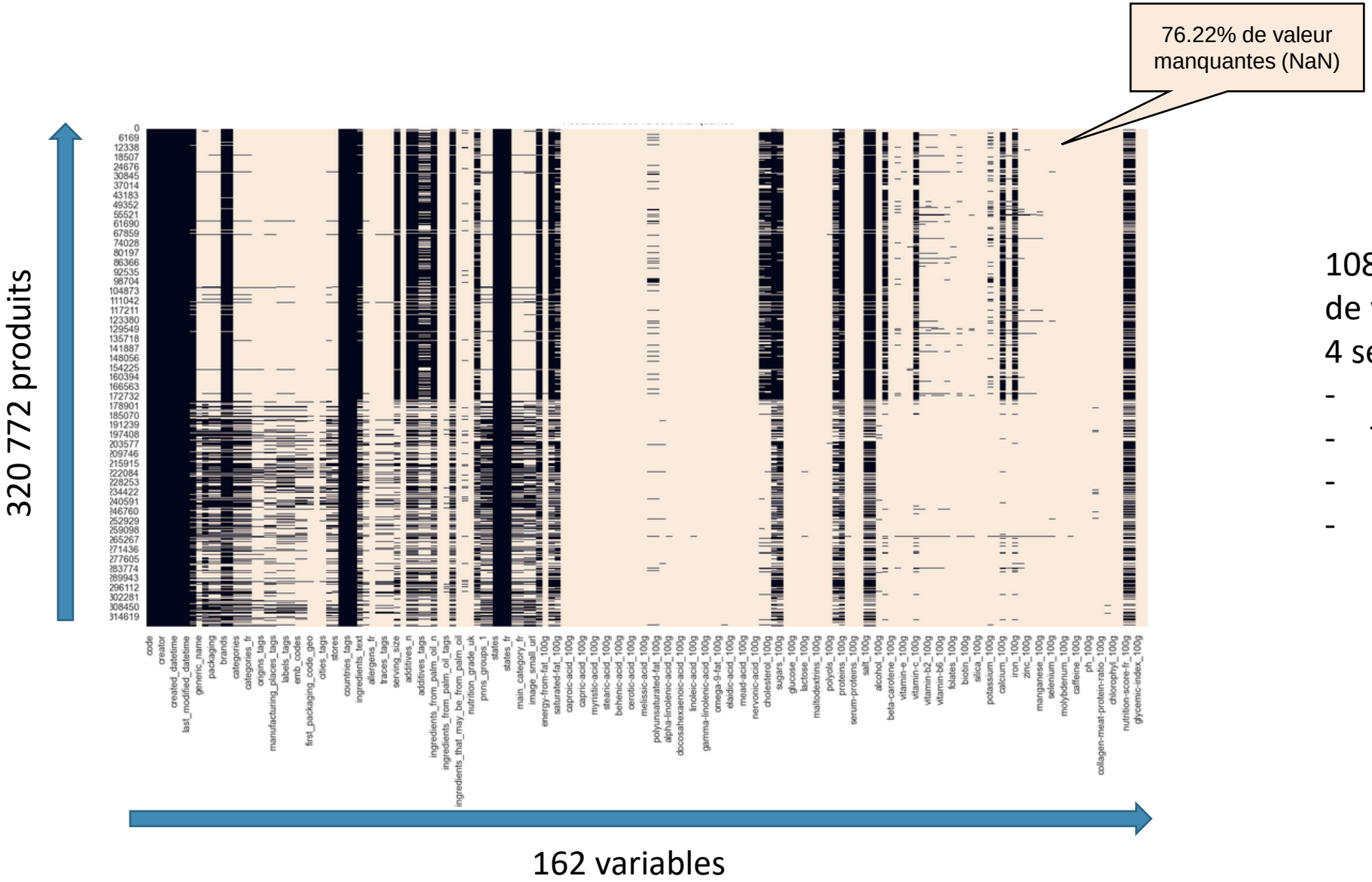

Mon Assistant Rénal
Recherchez un produit


Protéines
Sel
X.x
X.x
open FOOD facts

Les données

Philippe Moty

Présentation du jeu de données



108 variables (colonnes) > 80% de valeur manquantes

4 sections :

- Information général
- Tags
- Ingrédients
- Informations nutritionnelles

Les besoins de notre application

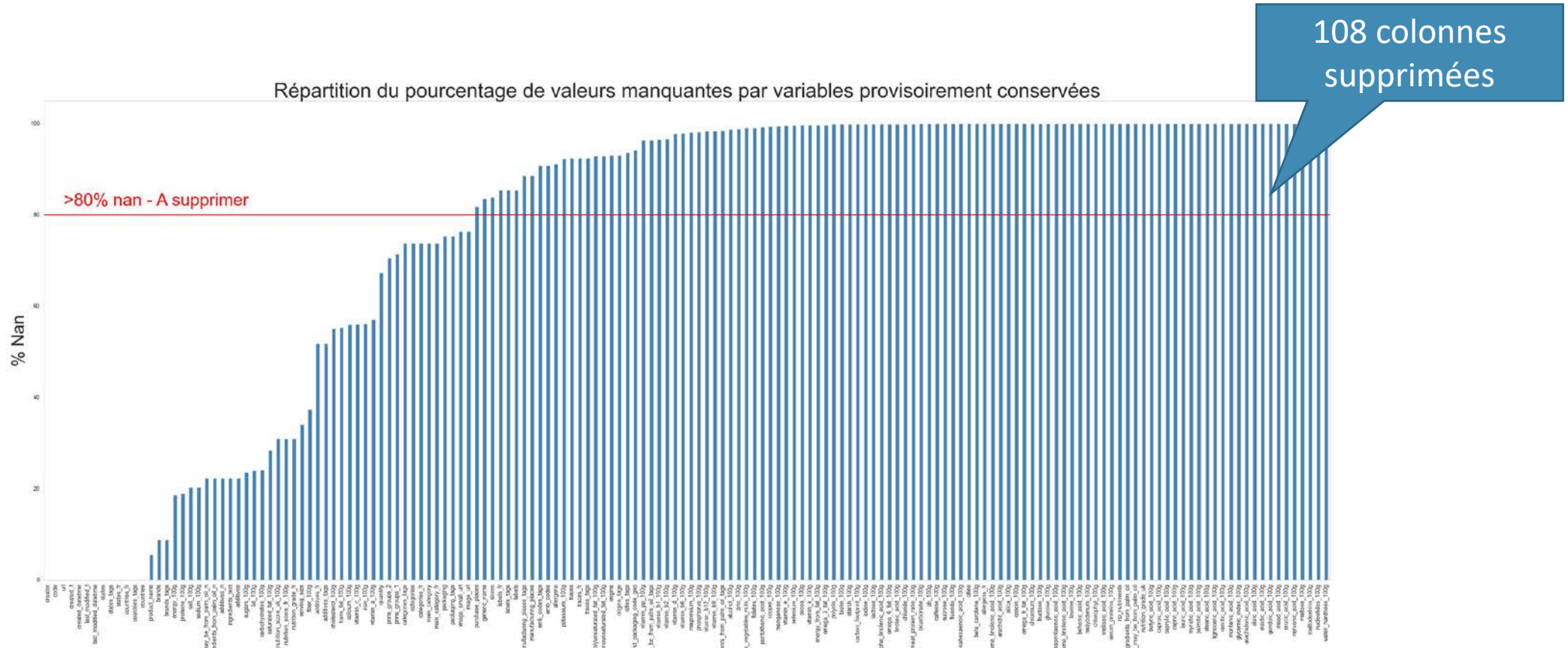
- Objectifs : COMMENT PROTEGER MES REINS ?
  - Limiter l'apport en sel
  - Contrôler les apports en protéines,
  - Couvrir les besoins nutritionnels,
- Les indicateurs disponibles
  - Nutrition :
    - 'nutrition-score-fr_100g','energy_100g'
    - 'fat_100g', 'saturated-fat_100g','carbohydrates_100g', 'sugars_100g', 'fiber_100g'
  - Sel('salt_100g'), Sodium('sodium_100g')
  - Protéines ('proteins_100g')
  - Limiter les additifs
  - Additifs : E338,339,340,341,342,343,350,351,352
  - Le potassium
  - Le phosphore

DÉMARCHE MÉTHODOLOGIQUE DU NETTOYAGE

Nettoyage des données

- Etapes :
  - Filtrer les données
  - Gérer les erreurs
  - Supprimer des doublons
  - Gestion des outliers et des valeurs aberrantes
  - Gestion des valeurs manquantes

FILTRER LES DONNÉES



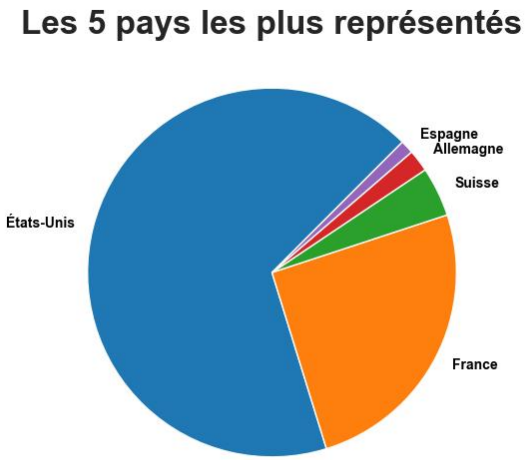
- Le **potassium** est un minéral important présent dans un grand nombre d'aliments indispensables au bon fonctionnement des muscles et du cœur. : trop de données manquantes
- Le **phosphore** est surtout présent dans les aliments, lié aux protéines.

Gérer les erreurs

- Les erreurs de pays : de 562 pays à 192 pays exploitables
  - On charge la listes des pays officiel pour supprimer les pays non conforme à la liste des pays du ministère des affaire étrangères français
- Suppression des variables
  - Les erreurs liée aux Code-barres
  - Les erreurs liée aux noms de produit

```python
# chargement des données complémentaires
df_countries = pd.read_csv("assets/datas/liste-197-etats-2020.csv",sep=";")
# visualisation du dataframe
# df_countries.head()
```

```python
# On définit les pays valides
VALID_COUNTRIES = df_countries['NOM'].unique()
# On applique un mask pour écarter les pays qui ne sont pas valides
mask = ~df['countries_fr'].isin(VALID_COUNTRIES)
df.loc[mask, 'countries_fr'] = np.NaN
```



Suppression des doublons

```
# On supprime les doublons des produits de la même marque en conservant les valeurs nulles
df[(~df.duplicated(["product_name","brands"],keep="last")) | ((df['product_name'].isnull()) & (df['brands'].isnull()))
print(f"On a supprimé : {data.shape[0]-df.shape[0]} lignes en doublons")
```

```
On a supprimé : 79617 lignes en doublons
```

- Suppression : 79617 lignes qui sont des doublons (on ne prend pas en compte null comme valeurs)
  - Les doublons de marques et de produits représentent 79617 lignes, on garde une occurrence de chaque éléments et on supprime les doublons

FILTRER LES VARIABLES de notre application

Indicateurs	Explications
code	Code barre du produit
creator	Créateur de la fiche
created_datetime	Date de création de la fiche
last_modified_datetime	Dernière modification de la fiche
product_name	Nom du produit
Brands	Marque du produit
categories_fr	Catégorie
countries_fr	Pays
additives_n	Nombre d'additifs
additives_fr	Nom des additifs
ingredients_from_palm_oil_n	Nombre d'ingrédient provenant d'huile de palme
nutrition_grade_fr	Nutrigrade
main_category_fr	Catégorie principale
energy_100g	Valeur nutritionnelle ou calorique
fat_100g	gras
saturated_fat_100g	Graisse insaturée
carbohydrates_100g	Famille du sucre
sugars_100g	Sucre
fiber_100g	Fibre
proteins_100g	Protéines
salt_100g	Sel
sodium_100g	Sodium
nutrition_score_fr_100g	Résultat de l'algorithme du nutriscore

Gestion des outliers et valeurs aberrantes

	count	mean	std	min	25%	50%	75%	max
additives_n	233934.0	1.994161	2.516437	0.00000	0.0000	1.00000	3.000000	3.100000e+01
ingredients_from_palm_oil_n	233934.0	0.018779	0.137433	0.00000	0.0000	0.00000	0.000000	2.000000e+00
ingredients_that_may_be_from_palm_oil_n	233934.0	0.054721	0.266629	0.00000	0.0000	0.00000	0.000000	6.000000e+00
energy_100g	257773.0	1140.497182	6484.434455	0.00000	373.0000	1100.00000	1674.000000	3.251373e+06
fat_100g	240584.0	12.701388	17.574028	0.00000	0.0000	5.00000	20.000000	7.142900e+02
saturated_fat_100g	226641.0	5.115709	8.008095	0.00000	0.0000	1.79000	7.140000	5.500000e+02
trans_fat_100g	143159.0	0.073428	1.540612	-3.57000	0.0000	0.00000	0.000000	3.690000e+02
cholesterol_100g	143950.0	0.020079	0.358234	0.00000	0.0000	0.00000	0.020000	9.523800e+01
carbohydrates_100g	240276.0	32.091647	29.765368	0.00000	6.0000	20.60000	58.500000	2.916670e+03
sugars_100g	241910.0	15.993129	22.345150	-17.86000	1.3000	5.71000	24.000000	3.520000e+03
fiber_100g	198587.0	2.863581	12.933636	-6.70000	0.0000	1.50000	3.600000	5.380000e+03
proteins_100g	256605.0	7.074330	8.418758	-800.00000	0.7000	4.76000	10.000000	4.300000e+02
salt_100g	252527.0	2.037957	129.023620	0.00000	0.0635	0.58420	1.379220	6.431280e+04
sodium_100g	252488.0	0.802467	50.800621	0.00000	0.0250	0.23000	0.543000	2.532000e+04
vitamin_a_100g	137398.0	0.000397	0.073320	-0.00034	0.0000	0.00000	0.000107	2.670000e+01
vitamin_c_100g	140655.0	0.023350	2.238130	-0.00210	0.0000	0.00000	0.003700	7.169811e+02
calcium_100g	140837.0	0.125119	3.320757	0.00000	0.0000	0.03500	0.106000	6.947370e+02
iron_100g	140305.0	0.003654	0.214528	-0.00026	0.0000	0.00101	0.002400	5.000000e+01
nutrition_score_fr_100g	218463.0	9.151724	9.052588	-15.00000	1.0000	10.00000	16.000000	4.000000e+01
nutrition_score_uk_100g	218463.0	9.044406	9.180977	-15.00000	1.0000	9.00000	16.000000	4.000000e+01

	code	energy_100g
212928	3257983143096	3251373.0

Les valeurs nutritionnelles ne peuvent pas dépasser 100g pour 100g de produit

Les valeurs ne peuvent pas être négatives

Gestion des outliers et valeurs aberrantes

- On traite les éléments aberrants de : energy_100g
  - On remplace les valeurs des outliers en fonction de la base d'openfoodfacts
  - Puis on supprime les éléments > 3800 kj*

```
# On remplace les valeurs abberantes fausse d'après leur fiche produit
dict = {
    "3257983143096" : 1373,
    "8710573641501" : 2312,
    "3661405001053" : 182,
    "0201203040026" : 3700,
    "0619309100979" : 92,
    "2000000045489" : 2200,
    "3596710288755" : 2807,
    "3291960006127" : 3766,
    "0041390030512" : 1464,
    "0444444387721" : 1393,
      }
```

- On traite les donnée <0 ou >100 (non calculable)
  - Suppression de 18 produits < 0 et 196 >100

```
coll_nutrition_facts = ['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']
```

```
df[(df[coll_nutrition_facts] > 100).any(axis=1)]
df[(df[coll_nutrition_facts] < 0).any(axis=1)]
```

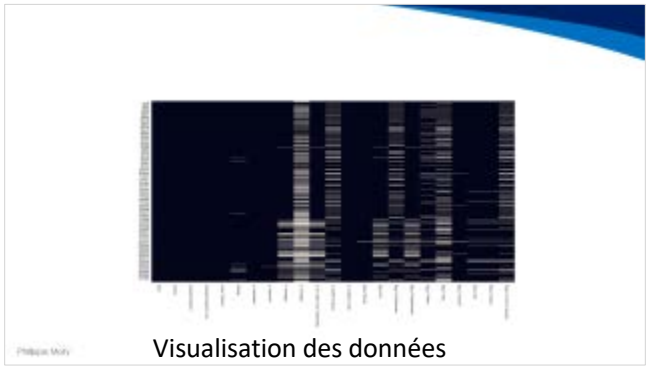
- On traite les éléments Incohérents
  - Satured-fat-100g > fat_100g et sodium_100g > salt_100g n'est pas possible, on supprime les produits qui sont dans ces cas => 336 produits sont supprimés

```
df= df[~((df['saturated_fat_100g'] > df['fat_100g'])
        | (df['sodium_100g'] > df['salt_100g']))]
```

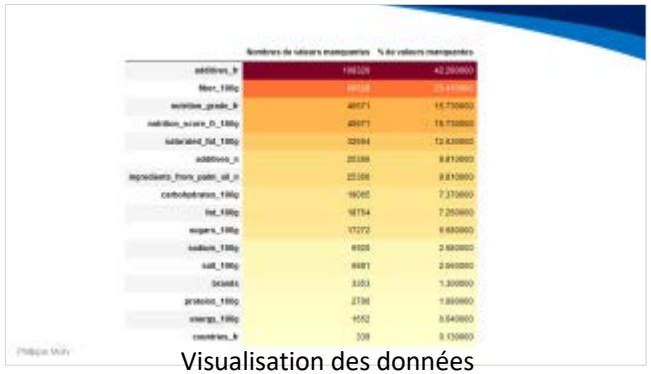
- Données nettoyées

(*) : https://en.wikipedia.org/wiki/Food_energy , https://www.careomnia.com/nutrition-tool-nutrient?nutrientID=22&all=1#

Gestion des valeurs manquantes

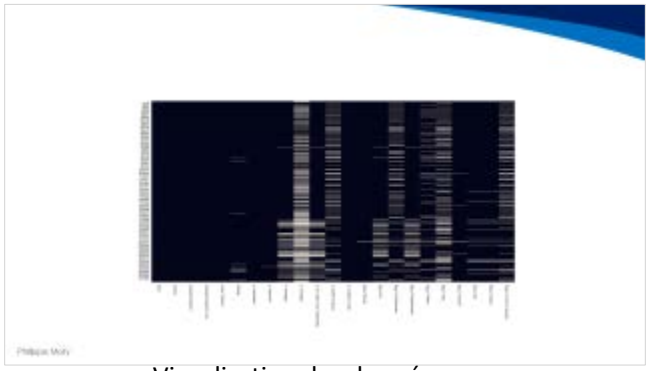


Visualisation des données manquantes avant traitement



Visualisation des données manquantes à traiter

Gestion des valeurs manquantes



Visualisation des données manquantes avant traitement

Visualisation des données manquantes à traiter

La fonction Fillna() nous permet de remplacer les données manquante par une valeur numérique, une équation ou un texte en cohérence avec nos variables

La méthode des KnnImputer de scikit permet de remplacer les valeurs manquantes par les 5 plus proches voisins et nous fournir une information très proche de la réalité

Variables	Démarche
Additive_n	Fillna(0)
ingredients_from_palm_oil_n	Fillna(0)
country	Fillna('monde')
brands	Fillna('inconnue')
Additive_fr	Fillna(' ')
Nutrition_grade_fr	Fillna(0)
fiber_100g nutrition_score_fr_100g saturated_fat_100g carbohydrates_100g fat_100g sugars_100g sodium_100g salt_100g proteins_100g energy_100g	knnImputer des 5 plus proches voisins

Nettoyage des données

- Etapes :
  - Filtrer les données
  - Gérer les erreurs
  - Supprimer des doublons
  - Gestion des outliers et des valeurs aberrantes
  - Gestion des valeurs manquantes

Démarche	Nb Lignes	Nb variables
Chargement du jeu de données	320772	162
Suppression des variables avec %NaN > 80%	320772	54
Suppression des lignes si aucune valeur nutritionnelle n'est renseignée	262833	54
Suppression : 23 lignes erronées	262817	54
Suppression de 3383 produits sans nom	259434	54
Suppression des doublons	241155	54
Energy_100g >3800kj	240818	54
Les éléments _100g > 100	240641	54
Les éléments _100g < 100	240624	54
SELECTION DES DONNEES	240297	23

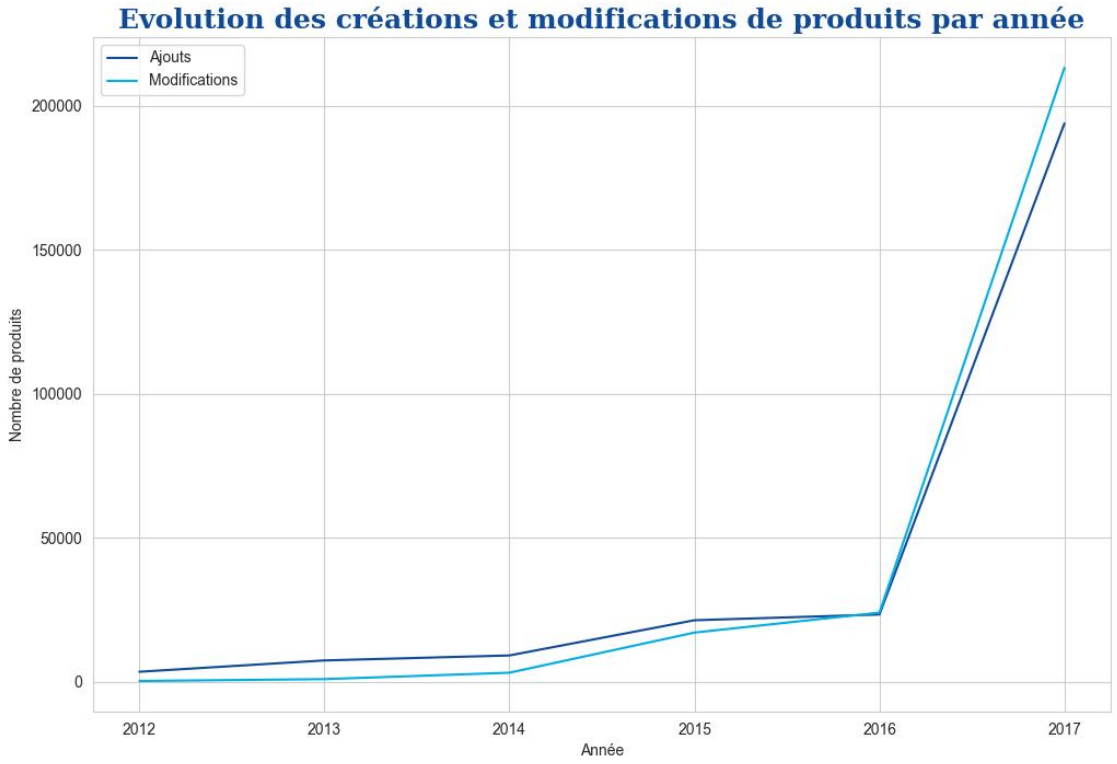
DÉMARCHE MÉTHODOLOGIQUE D’EXPLORATION

La description et l'analyse univariée des différentes variables importantes avec les visualisations associées.

- Étapes :
  - Dates
  - Marques
  - Catégories
  - Données nutritionnelles
  - Additifs
  - Nutriscore et nutrigrade
  - Liaison des variable avec le nutrigrade
  - Les liens entre les variables
  - Cercle de corrélations

DATES

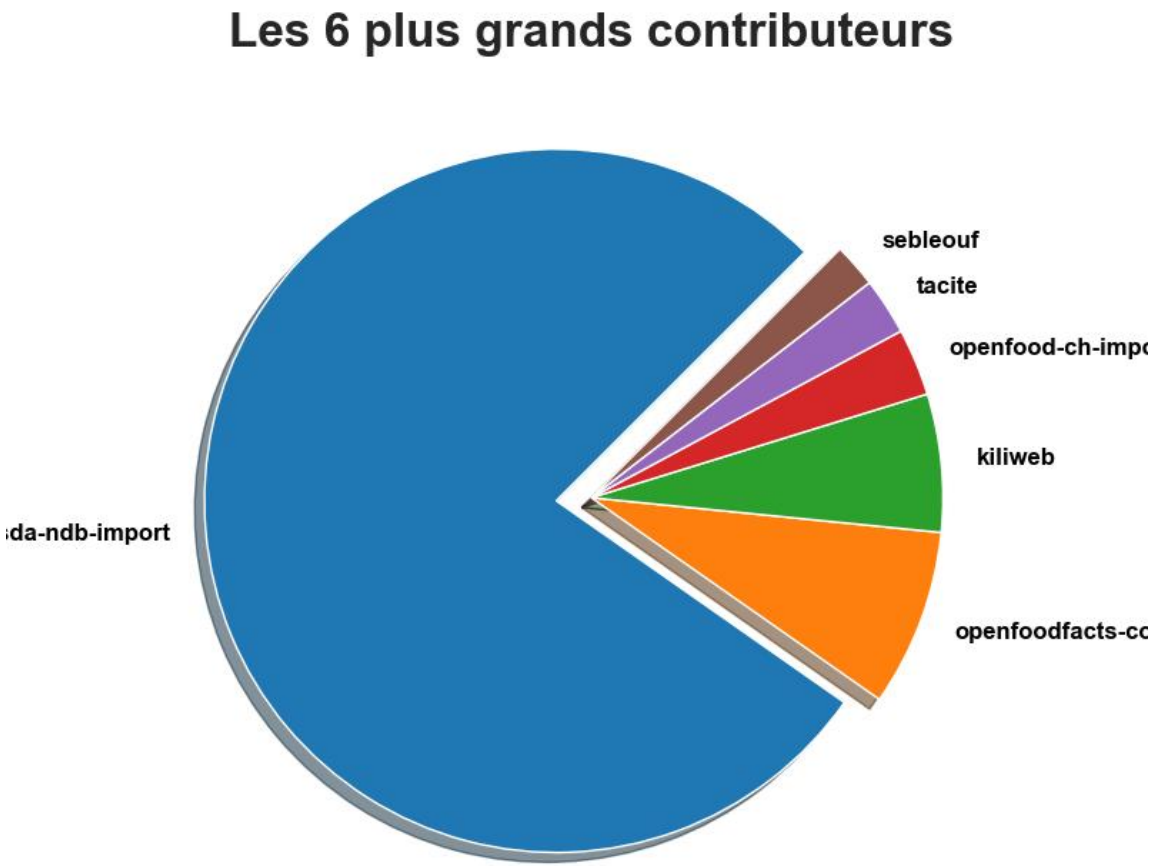
- **Le pic de 2016, le début d'une collecte massive …**
- **Explication des dates** :
  - Le Nutri-Score est prévu dans la **loi de 2016 en France**
  - **Mis en place en France en 2017**, l'étiquetage nutritionnel Nutri-Score s'applique aujourd'hui dans sept pays
- Qui a rencontré l'avis favorables des consommateurs, des États et des entreprises



Les Sources

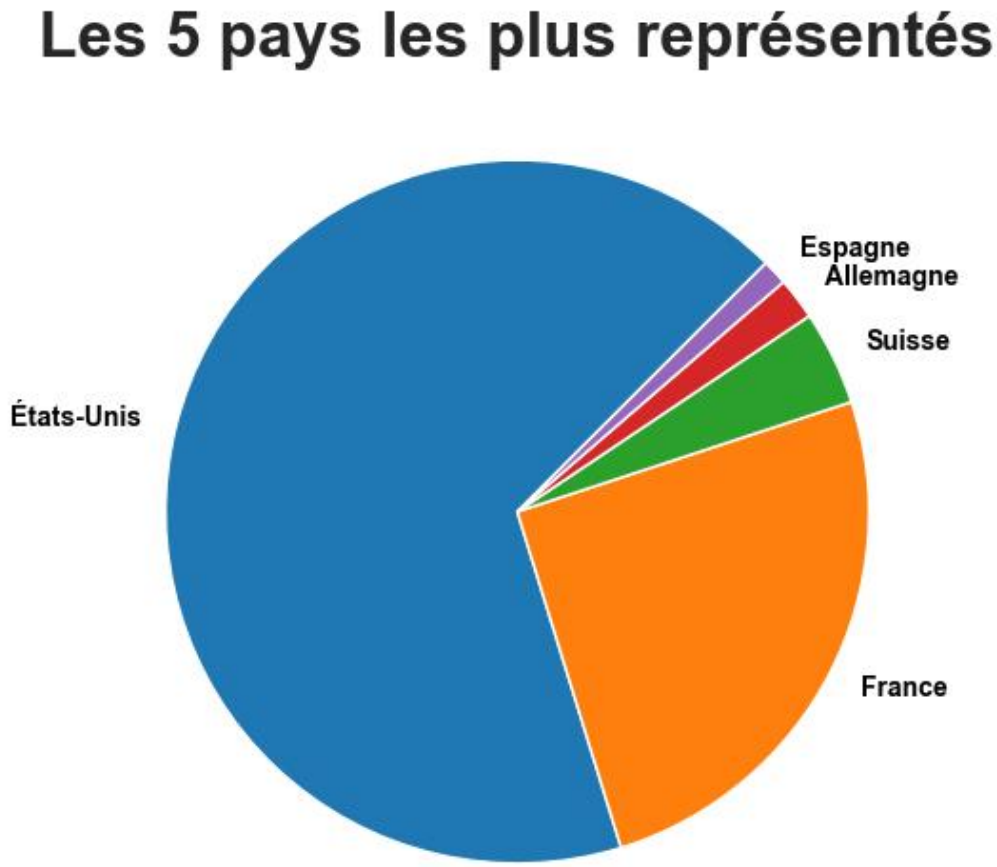
Qui sont les sources de ces données ?

- 2478 contributeurs dont des institutions internationales mais aussi des particuliers
- La majorité des données provient de l'USDA : Ministère de l'agriculture américain :
- openfoodfacts-contributors
- Communauté de contributeurs
- **Tous participe assurant une collecte rapide et massive de données**
  - Les données manquantes sont en train d'être complété



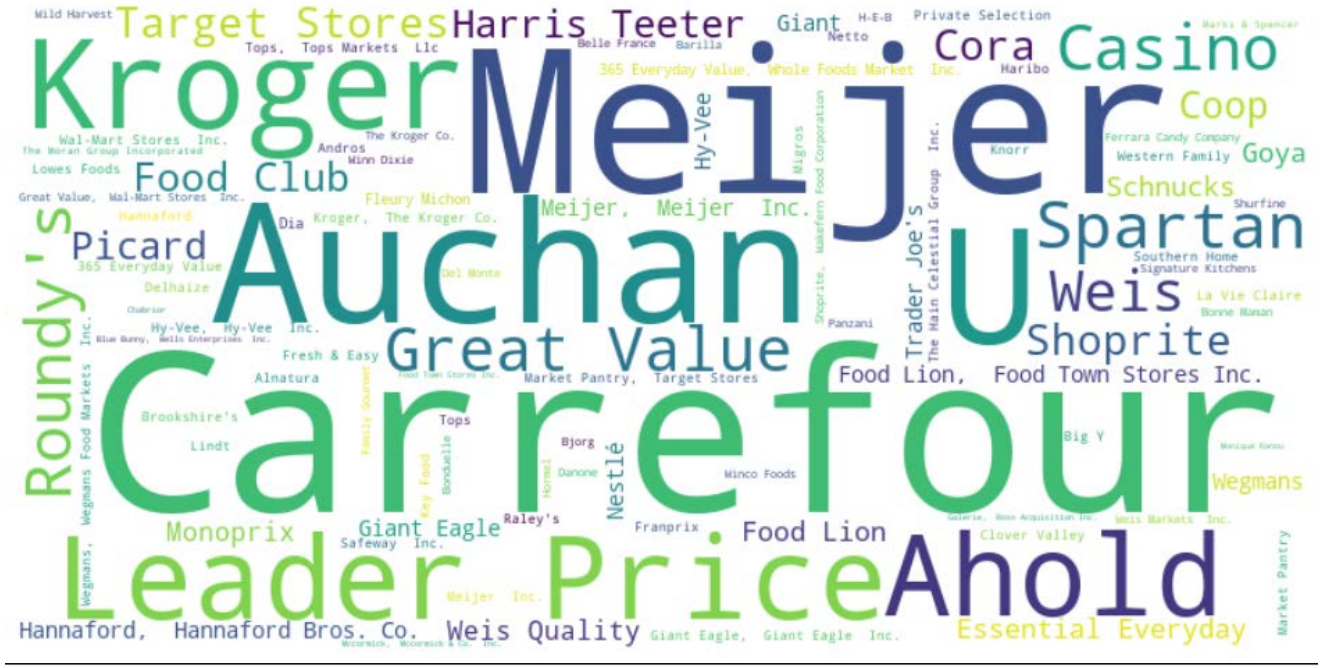
Les pays

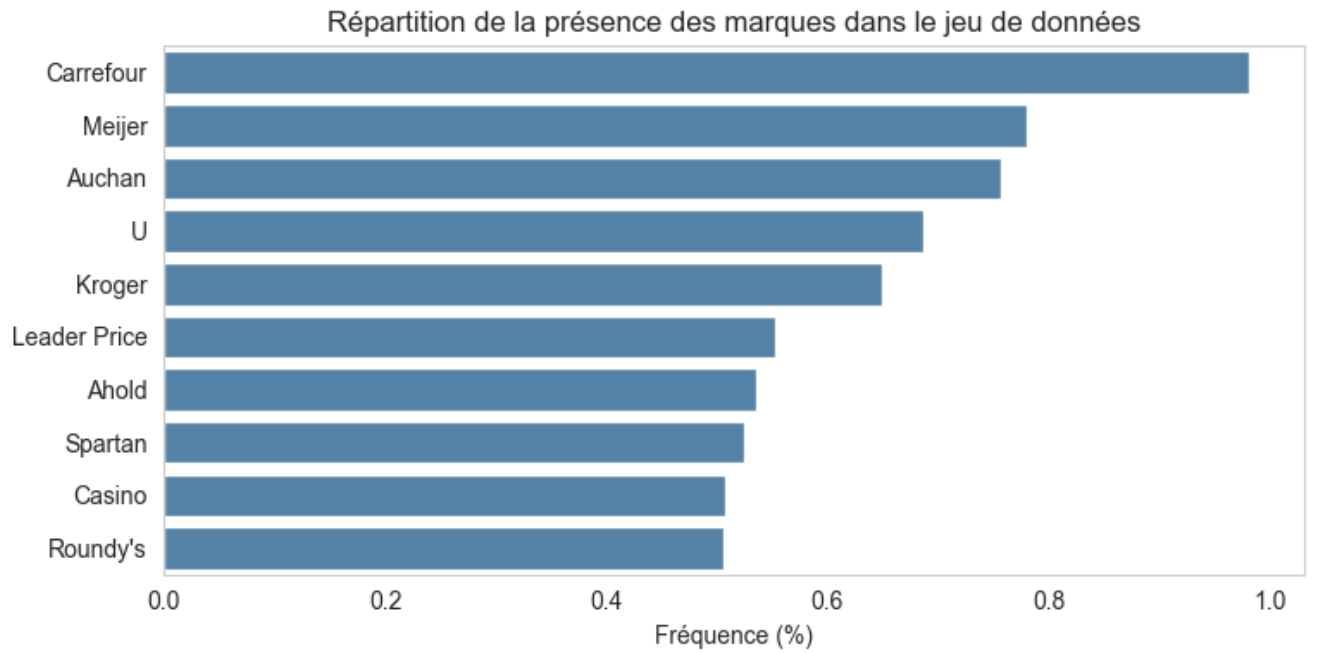
- La majorité des produits proviennent d’entreprise européennes et américaines et constitue notre cible de clients



Marques

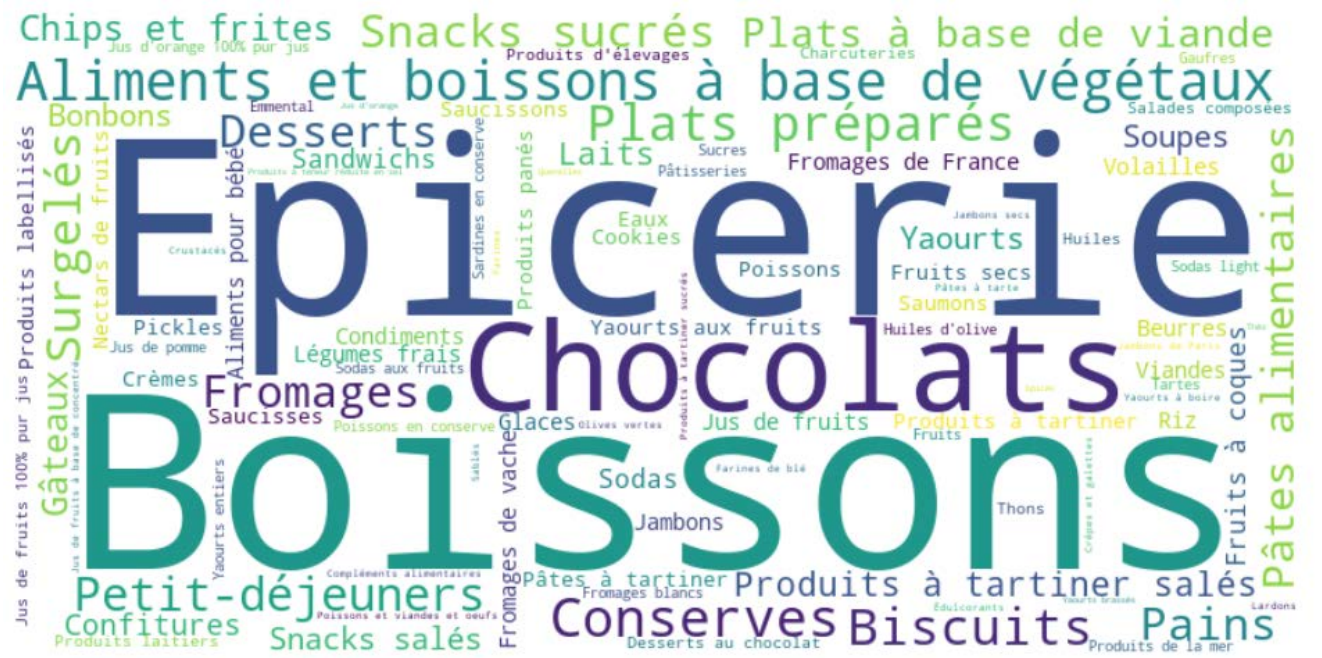
- Sur 46265 marques différentes dans le df
  - 237159 produit non renseignés : 'inconnues'
- **Les marques disponibles en France sont bien représentées :**
  - Carrefour, Auchan, Leader Price, U, Casino, Monoprix, Picard, Bjorg, Nestlé, Lipton, Cora, Harris... sont des marques populaires en France

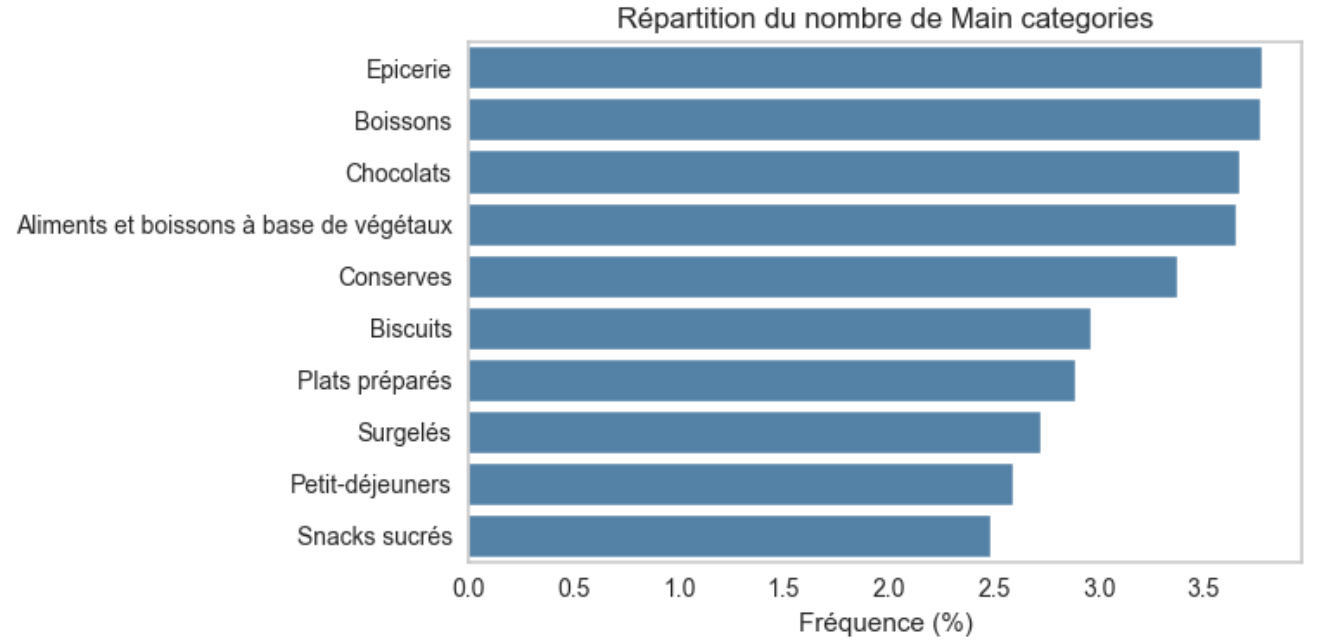




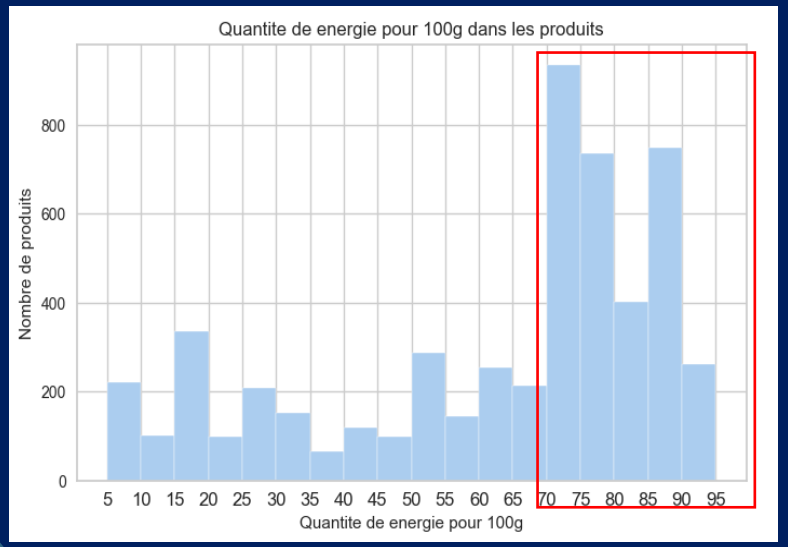
Catégories

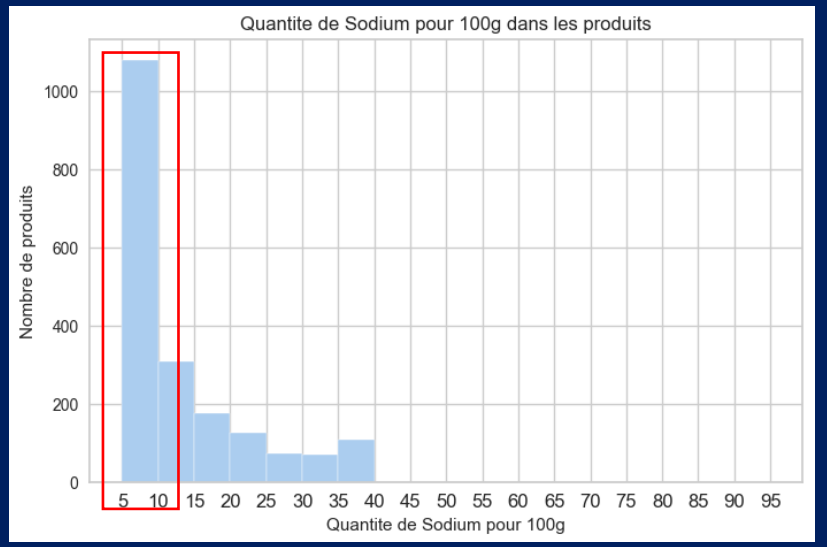
- Parmi les **catégories les plus représentés ont retrouve beaucoup de produits considérés comme à surveiller dans une alimentation saine**.
- **On exploite ces informations pour informer le consommateur sur les produits à surveiller**. personnalisé contraire. Une bonne alimentation passe aussi par le plaisir et ne doit pas être stigmatisé sans avis médical
- Pour notre appli cette source de données est importantes et l'application devra signaler les choses de manière pédagogique
  - Alimentation : interdite par le medecin
  - alimentation : apport à surveiller
  - alimentation : recommandé

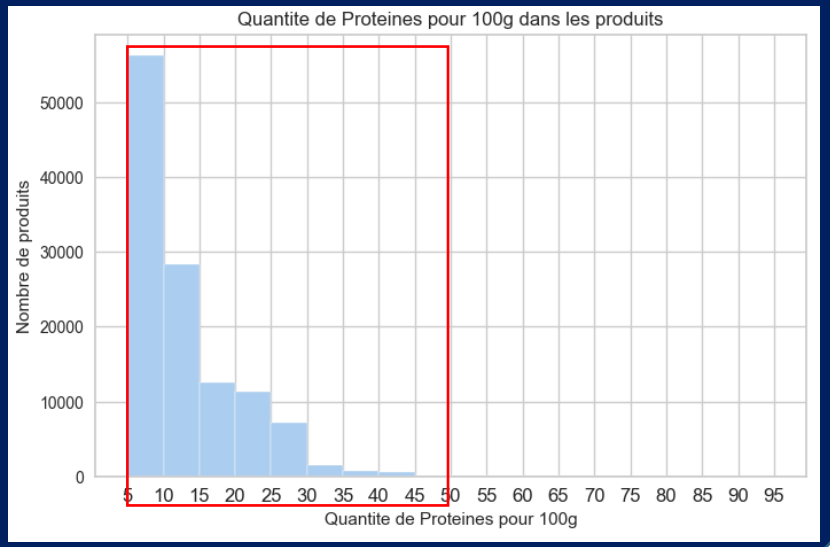




Données nutritionnelles – par nombre de produits







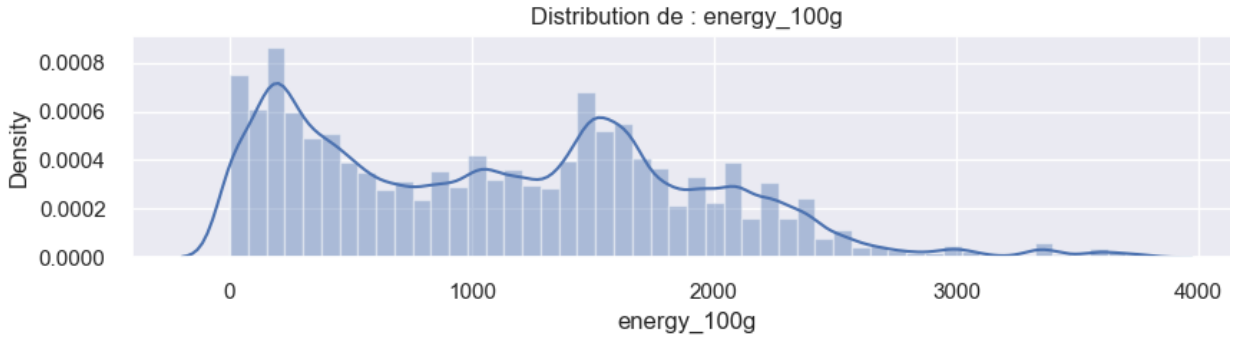
Les produits contenue dans la base de données sont :

- plutôt caloriques,
- contiennent en majorité peu de sodium mais suffisamment pour être significatif pour notre application
- et dispose de quantités de protéine assez bien répartit entre 5 et 50 g. Ils témoignent de la diversité des produits de la base de données.

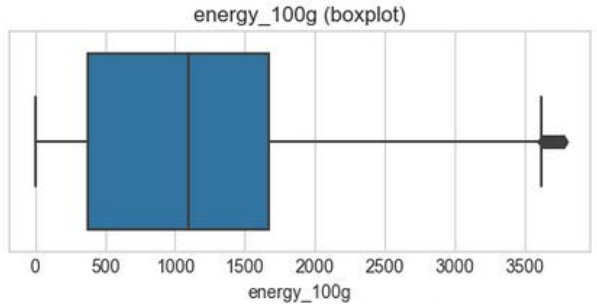
Energie

- La distribution montre :
  - Un nombre important de produits autour de 0
  - Une présence importante de produit calorique

L'application disposera d'une grande variété de produits et notamment les aliments à surveiller



mean	1121.26
std	791.47
min	0.00
25%	377.00
50%	1100.00
75%	1674.00
max	3776.00



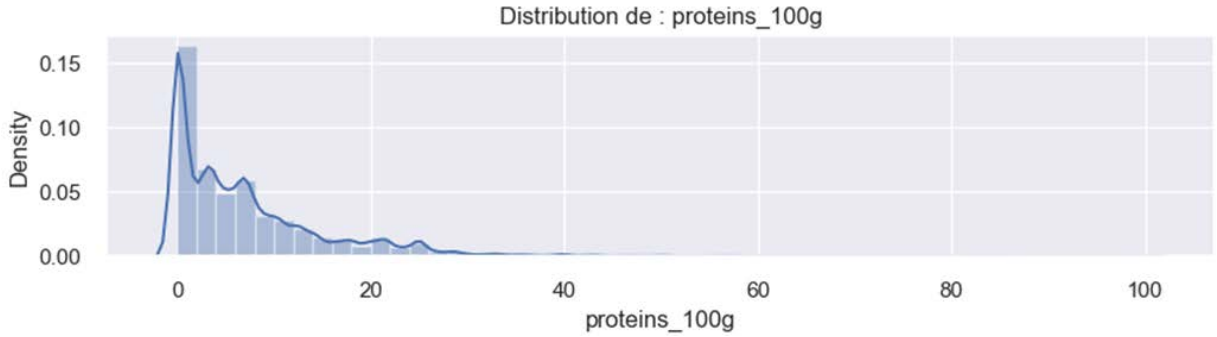

```
Test Shapiro-wilk
stat=0.952
p=0.000
Distribution non Gaussienne
----------------------------------------
Test normaltest (d'Agostino)
Statistics=11099.802, p=0.000
La distribution ne suit pas la loi normale (P<0,05)
```

Protéine

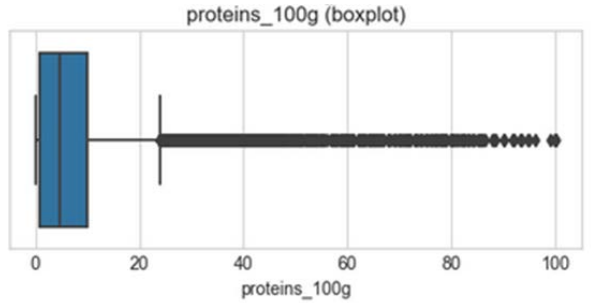
- Une distribution du nombre de protéine au 100g gramme très variée
- La majorité des données se trouvent entre 0 et 10 gramme

Beaucoup d'outliers

Bonne diversité de produits protéiné



mean	7.07
std	8.13
min	0.00
25%	0.70
50%	4.76
75%	10.00
max	100.00



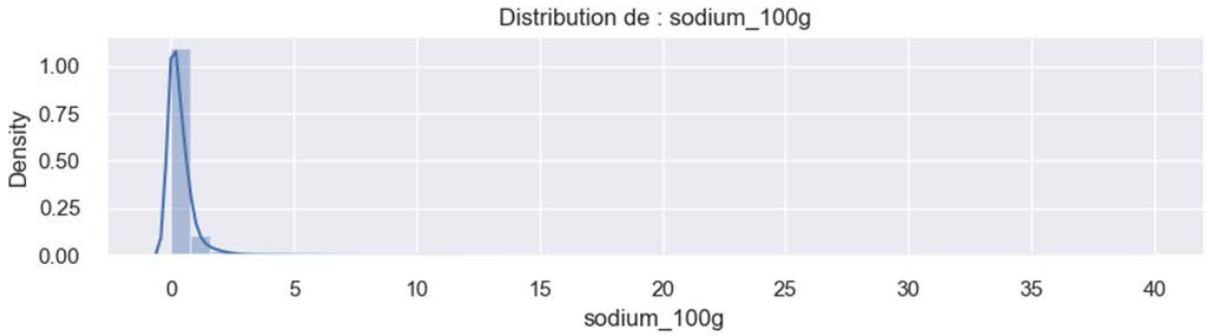

```
colonne : proteins_100g
Test Shapiro-wilk
stat=0.800
p=0.000
Distribution non Gaussienne
-----------------------------------------
Test normaltest (d'Agostino)
Statistics=122104.184, p=0.000
La distribution ne suit pas la loi normale (P<0,05)
```

Sel

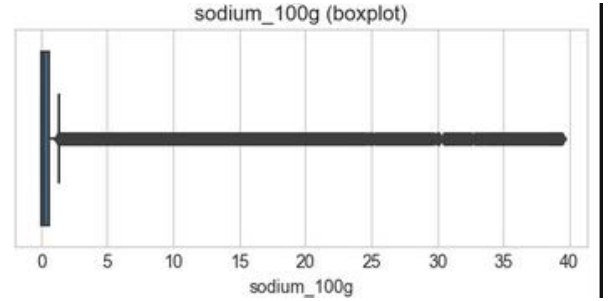
- Distribution concentrée autour de 0
- 75 % des produits ont une teneur en sodium entre 0 et 0,53g de sodium pour 100g

Les outliers sont présents

Grande diversité de produits



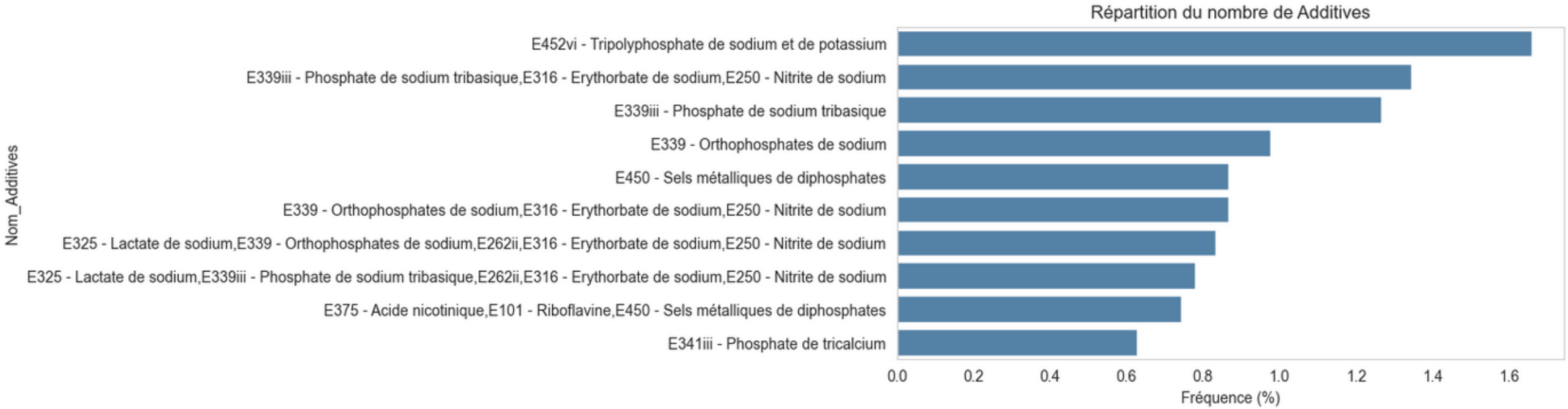
mean	0.62
std	2.44
min	0.00
25%	0.02
50%	0.23
75%	0.53
max	39.37




```
colonne : sodium_100g
Test Shapiro-wilk
stat=0.193
p=0.000
Distribution non Gaussienne
----------------------------------------
Test normaltest (d'Agostino)
Statistics=420952.198, p=0.000
La distribution ne suit pas la loi normale (P<0,05)
```

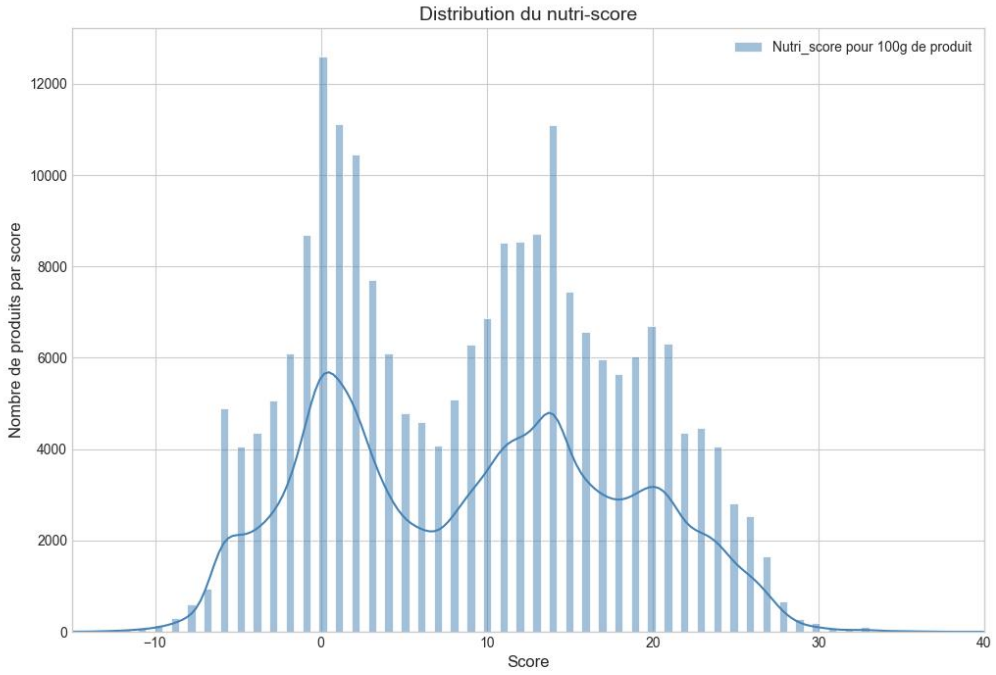
Les additifs à surveiller …

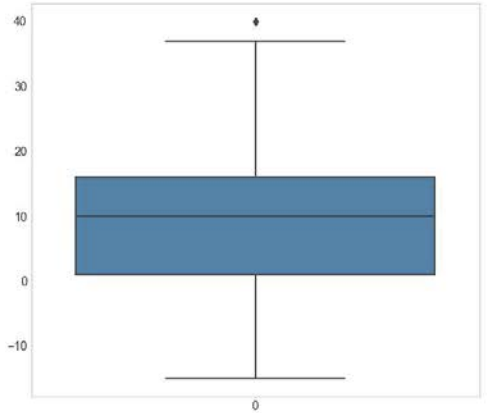
- … **sont présents** et pourront être signalés

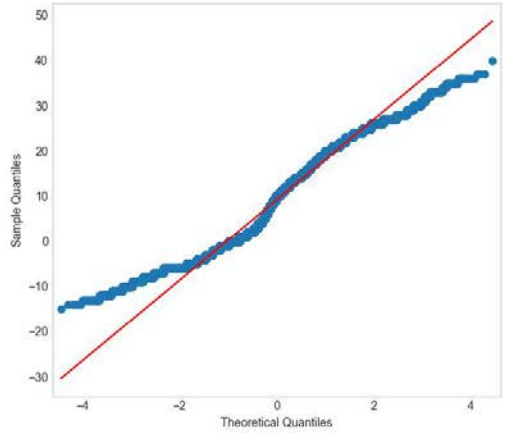


Le nutriscore

- **Le nutriscore est-il bien distribué, témoignant du répartition équilibrée des produits ?**
- 2 ensembles sont plus présent : plus de produit qui ont un score entre -1 et 2 et les produits et ceux ayant un score 11
- 1 ensemble de score est moins nombreux : ceux qui ont un score entre 4 et 6
- Il ne suit pas une loi normale







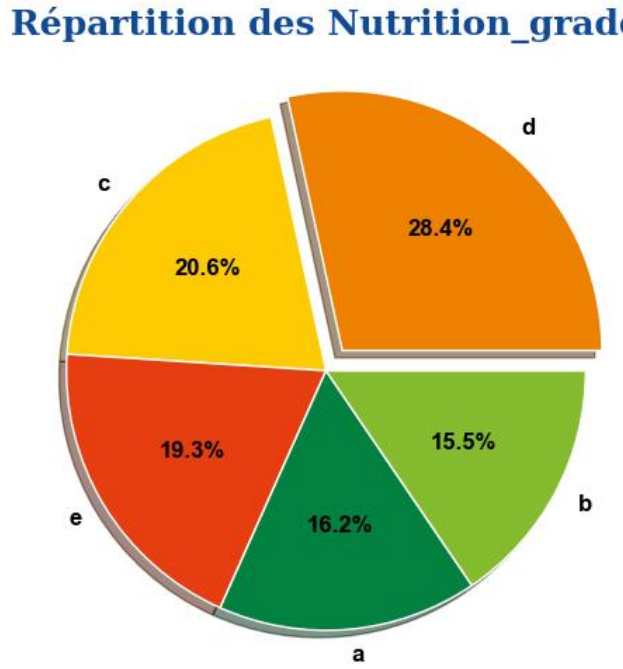

```
Test Shapiro-wilks
stat=0.968
p=0.000
Distribution non Gaussienne
```

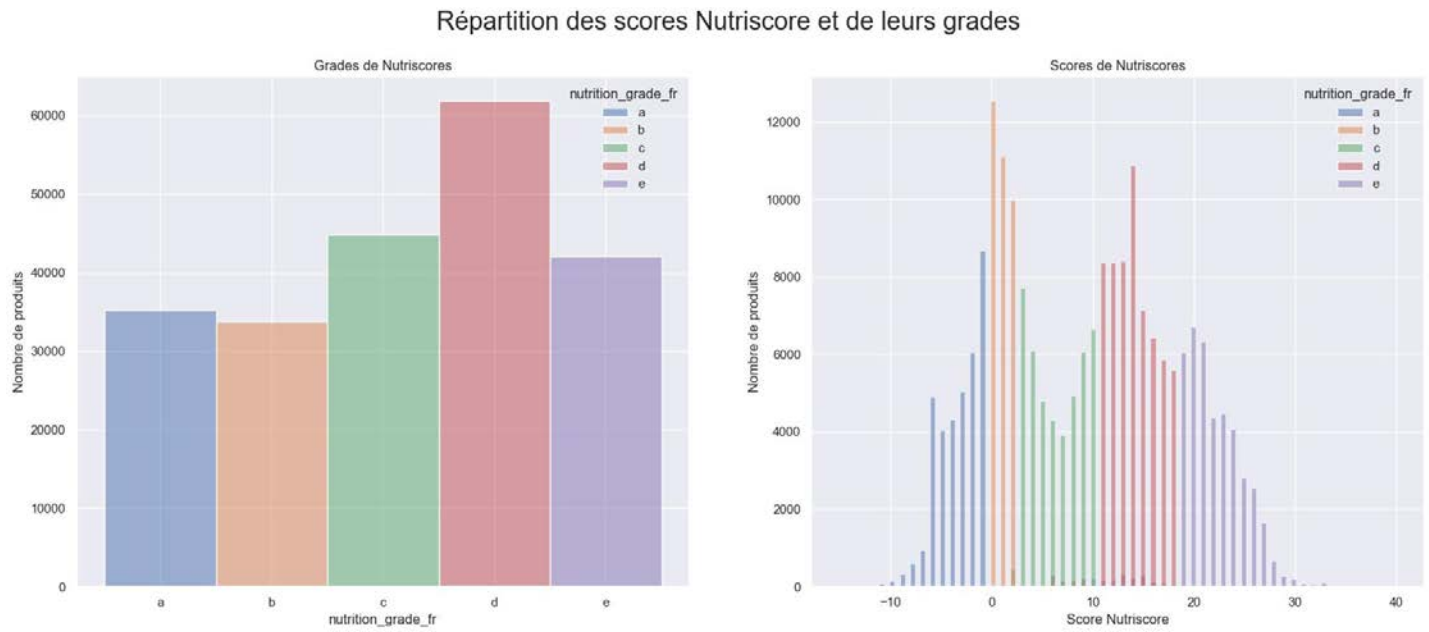
Nutrigrade

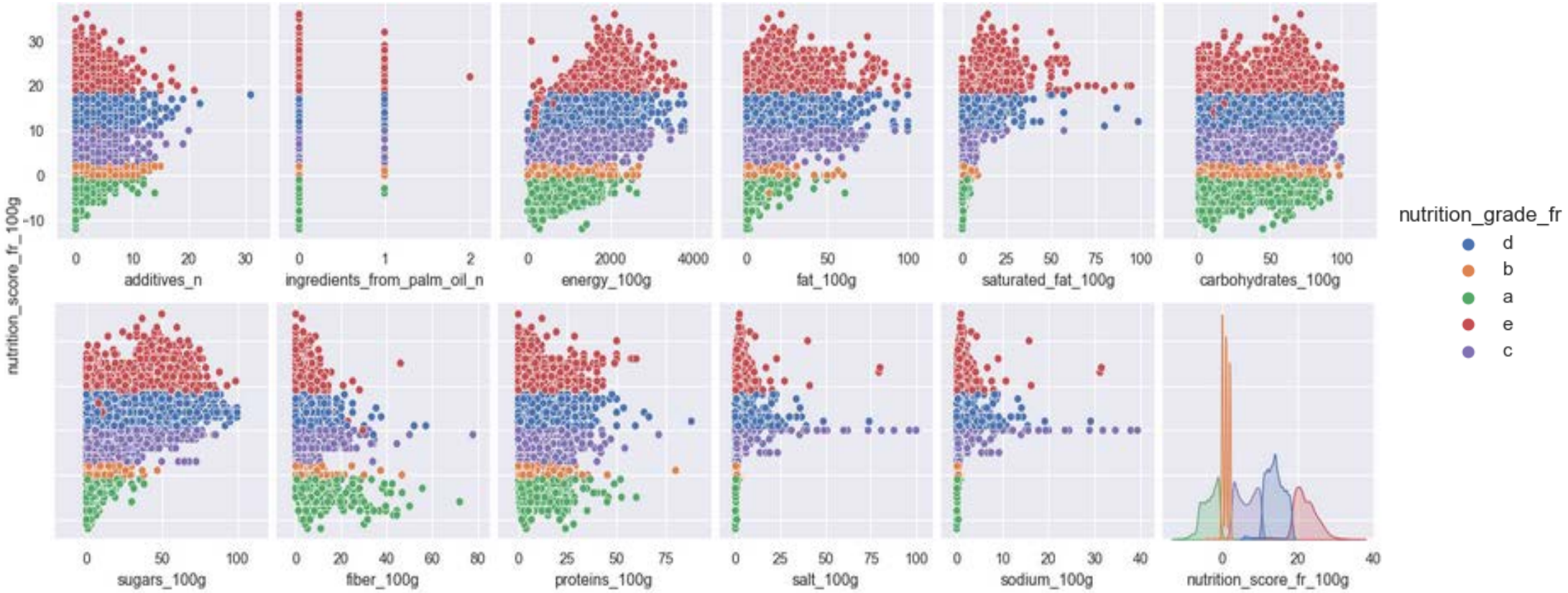
Comment exploiter les nutrigrades pour notre application ?

- **Tous les nutrigrades sont représentés** avec beaucoup de produits appartenant aus **groupes 'd'et 'e' représentant 47% des produits** de la base.
- **Lien nutriscore nutrigrade est siginificatif** :
  - les moyennes témoignent du classement :
    - plus le nutrition_score es bas plus sa moyenne est basse et inversement

On utilise le nutrigrade pour informer sur la qualité nutritionnelle des produits





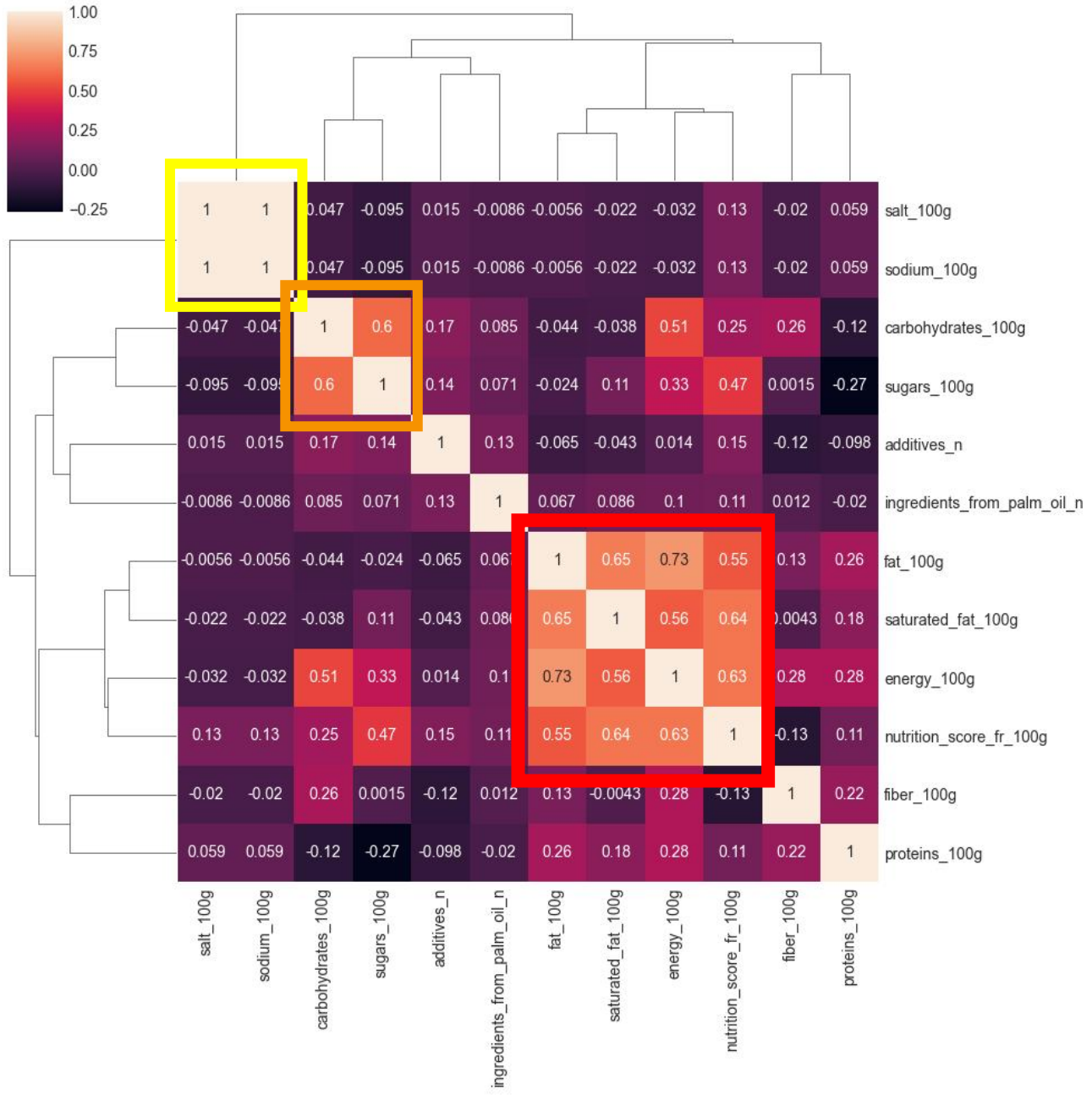


Liaisons des variables ciblées et le nutrigrade

Table des corrélations

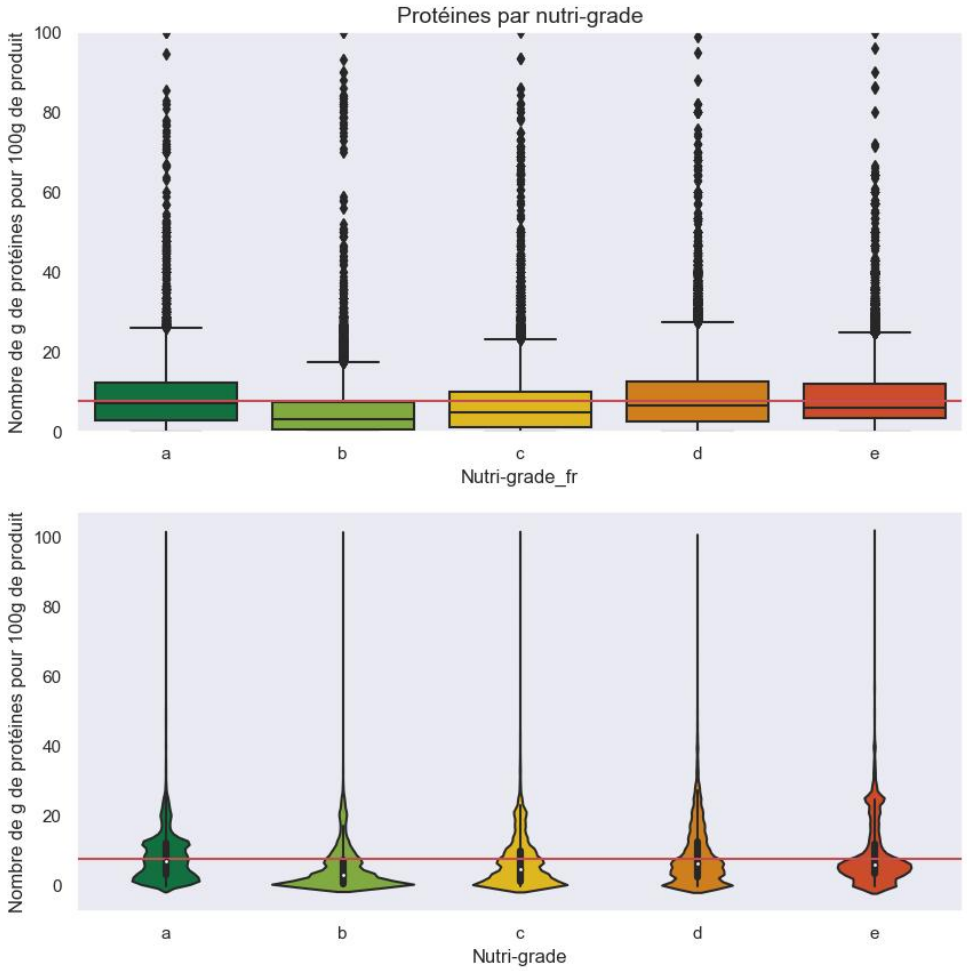
- Fortes corrélations :
  - **Salt / Sodium**
  - **Carbohydrate et sucre**
  - **Gras, gras insaturés, energy nutrition_score**

Les nutriscores des produits présents témoignent de leurs liens avec les produit gras et caloriques



Nutrigrade - Proteines

- amplitude en 0 et 100g de protéines
- Beaucoup outliers
- pas de corrélation directe marquante entre nutrigrade et nombre de protéine pour 100g
- On retrouve dans toutes les plages toutes les notes du nutrtion grade
- 75% des variables contiennent 11g de protéines

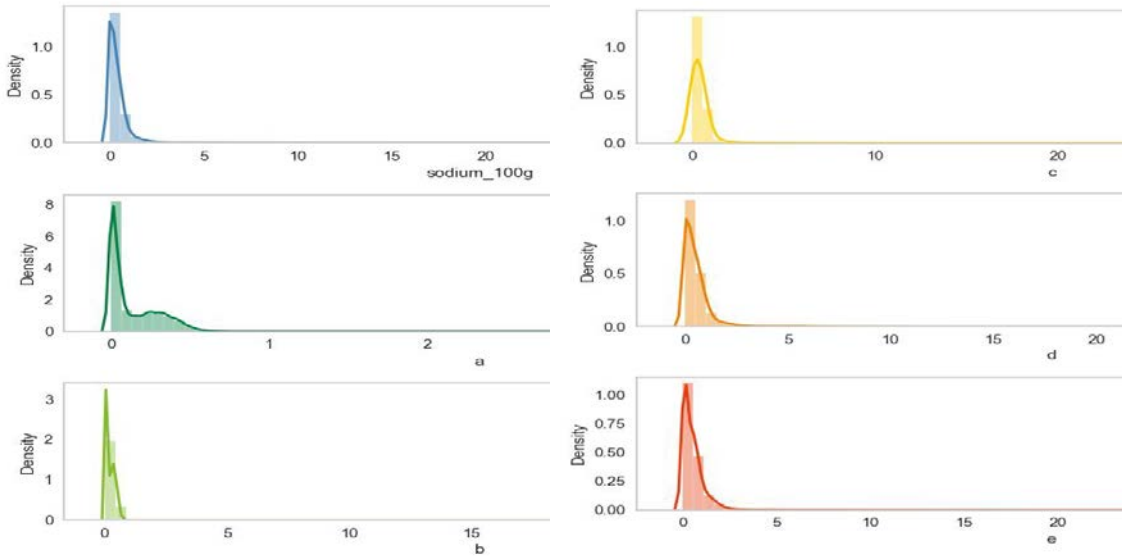


Nutrigrade - Sodium

- Le sodium n’est pas une variable discriminante

- Mais en moyenne plus un produit contient du sel plus son score est élevé et moins il est considéré comme sain

Histogrammes sodium par nutrigrades



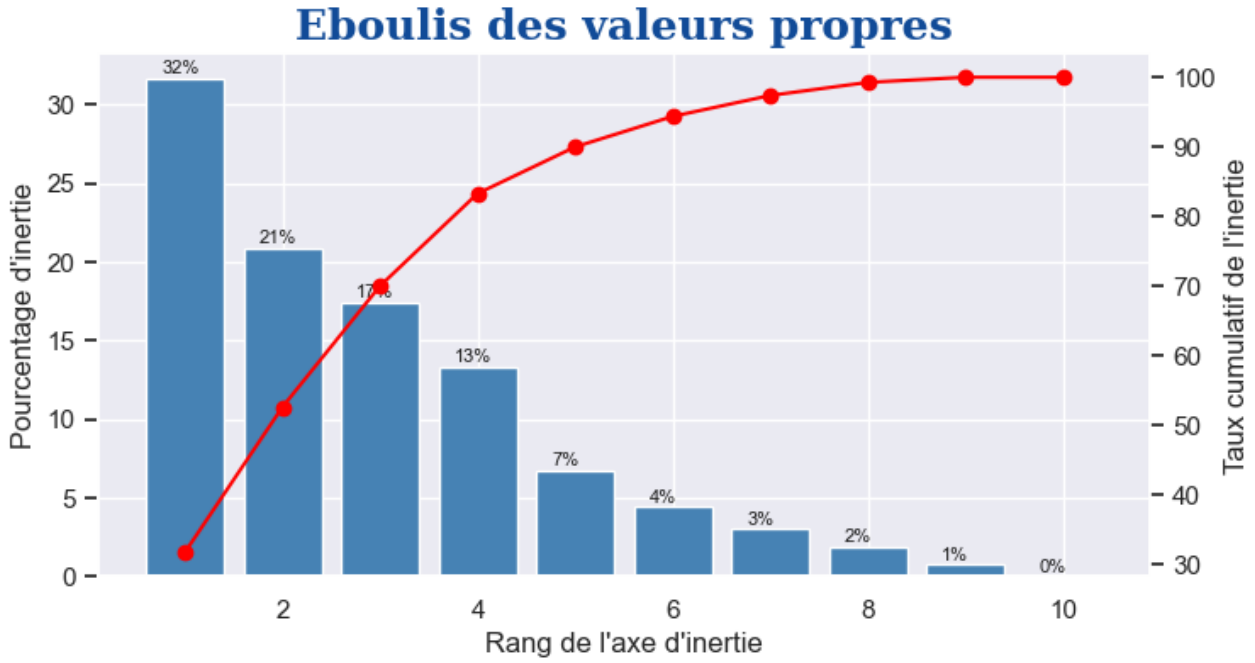
	a	b	c	d	e
count	35259.000000	33771.000000	44861.000000	61860.000000	42120.000000
mean	0.130397	0.205712	0.673686	0.640764	0.587656
std	0.154637	0.274928	2.672363	1.467174	1.169250
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.007000	0.016000	0.050000	0.071000	0.098856
50%	0.051181	0.167323	0.346457	0.365000	0.357000
75%	0.240157	0.362205	0.576000	0.718000	0.728346
max	4.666600	30.708661	39.370079	35.710000	39.370079

Panorama global de nos données

L’analyse multivariée et les résultats statistiques associés, en lien avec votre idée d’application

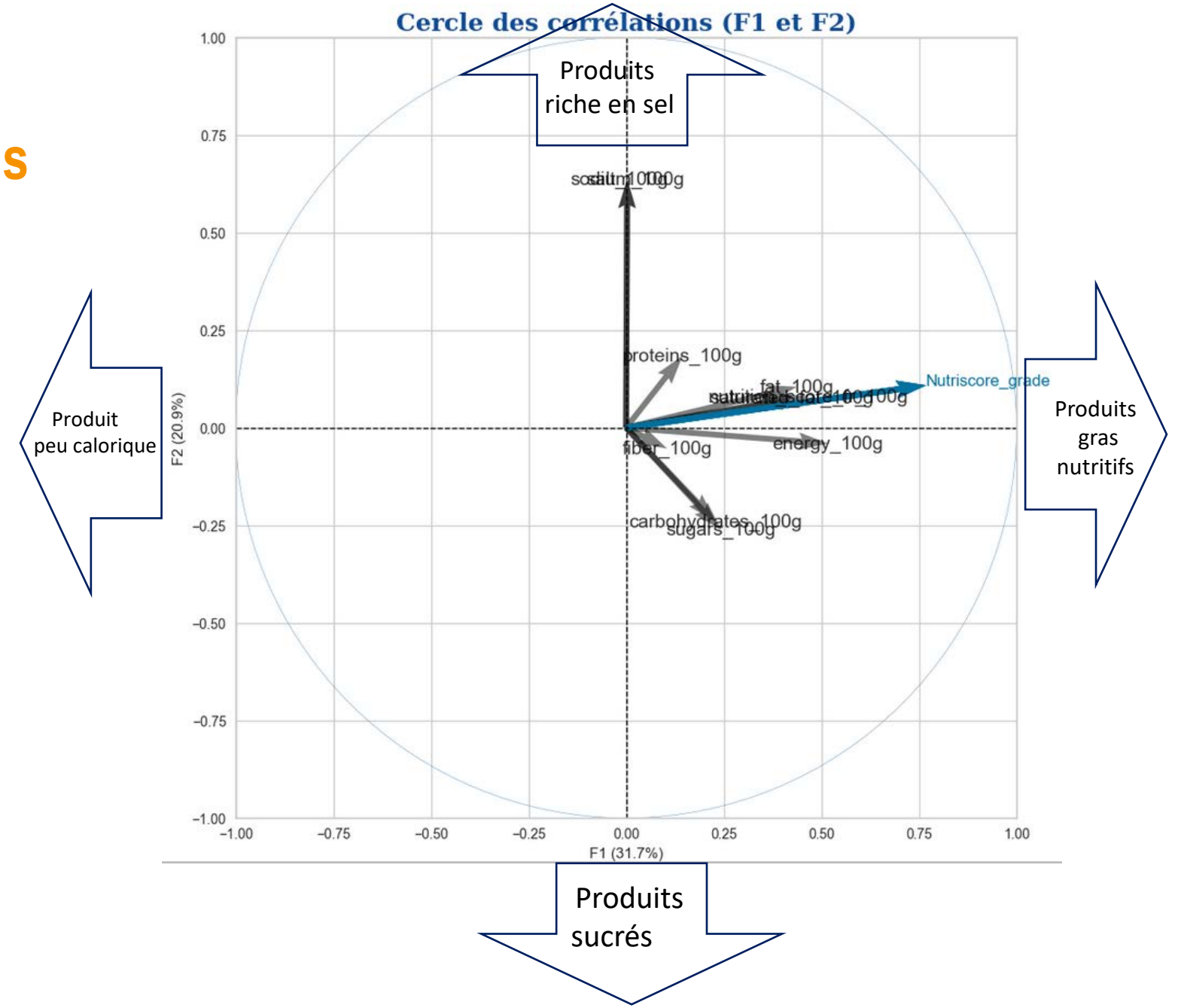
ACP

- Les 2 premiers plans factoriels couvrent une inertie d'un peu plus de 83,15%. Une analyse sur F1 et F2 semble donc cohérente.
- On projette le cercle de corrélation par la suite le premier pour expliquer le jeu de données



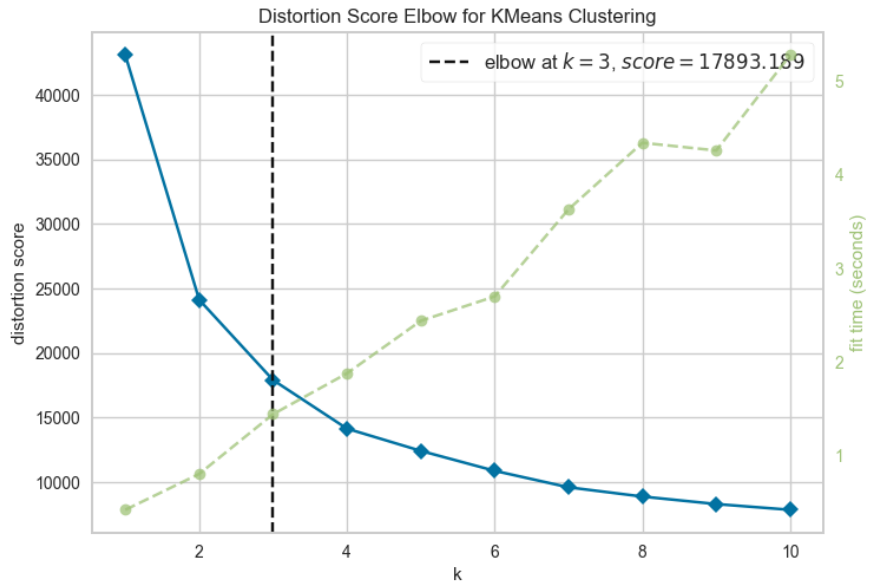
Cercle de corrélations

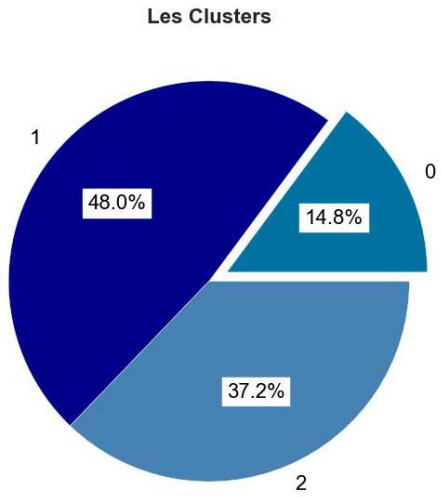
- Le nutriscore_grade est très lié au nutrition_score et aux produits gras et caloriques
- Sel et sodium sont très corrélés
- Sucre et carbohydrate sont très corrélés



Clustering

- Si nous segmentons notre base, nous obtenons 3 clusters





98]:

	count	mean	std	min	25%	50%	75%	max
Groupe 0	32336.0	16.770782	5.964855	-7.0	13.0	18.0	21.0	37.0
Groupe 1	104539.0	4.005319	6.459947	-15.0	0.0	2.0	9.0	30.0
Groupe 2	80996.0	12.711566	9.063137	-12.0	8.0	14.0	20.0	40.0

- groupe 1 (47.9%) : moyenne energie nutritionnelle faible - -
  - 75% des valeurs <780j (167.19 Calories)
  - 104456 produits
- groupe 0 (37.2%) : moyenne energie nutritionnelle moyenne
  - 75% des valeurs <1967j (469.81 Calories)
  - 80949 produits
- groupe 2 (14.9%) : moyenne energie nutritionnelle élevé
  - 75% des valeurs <3776j (901.88 Calories)
  - 32466 produits

Conclusion

Synthèse

- Le jeu de données contient toutes les informations dont nous avons besoin pour apporter une information complémentaire de qualité aux personnes qui souhaite surveiller leur reins.
  - Le Nutrition_grade_fr : Une alimentation saine, équilibrée et plaisir
  - Les indicateurs nécessaires pour informer le consommateur sont présents et en cours de consolidation
  - Sel('salt_100g') et le Sodium('sodium_100g') : sel minéral
  - Proteines ('proteins_100g’)
  - Le phosphore est absent mais le phosphore est surtout présent dans les aliments, lié aux protéines
    - La limitation protéique entraîne déjà une diminution des apports en phosphore.
  - Limiter les additifs : Les additif cible sont présent dans le jeu de données
- Les besoin pour développer l’application
  - Le développement nécessite l’appui d’un professionnel
  - Biais possible lié au mode de collecte de donnée

Evolution

- Dans un premier temps l’application apportera une information claire permettant d’identifier sels, protéines, les additifs et la fiche produit openfoodfacts
- Puis elle pourra s’étoffer par l’ajout de fonctionnalités :
  - en fonction du poids : l'application demandera le poids et le sexe de la personne pour avoir une indication personnalisée
  - suivre le régime sur la journée puis la semaine
  - Ajouter les produits interdits par les professionnels
  - Proposer des alternatives au produit
    - moins salé, sans additif
    - proposer des recettes

TESTS

Zoom

Allumettes Nature - Herta - 204 g (2 * 102 g)

This product page is not complete. You can help to complete it by editing it and adding more data from the photos we have, or by taking more photos using the app for Android or iPhone/iPad. Thank you!

Barcode: 7613035336544 (EAN / EAN-13)

Common name: Viande de porc traitée en salaison

Quantity: 204 g (2 * 102 g)

Packaging: Plastic, Film, Tray

Brands: Herta

```
[109]: cols = ['product_name','proteins_100g','sodium_100g','nutrition_grade_fr' ]
       x = df_nutriscore.loc[df_nutriscore['code']=='7613035336544',cols]
       x.style.hide_index()
```

[109]:

product_name	proteins_100g	sodium_100g	nutrition_grade_fr
Alumettes jambon herta	17.100000	0.849000	e

Mon Assistant Rénal

Merci

Philippe Moty