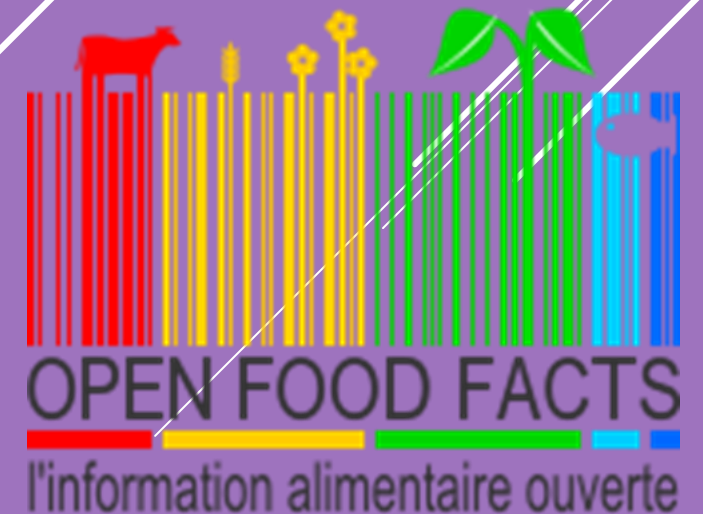


# CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE PROJET 2



Idée d'application  
Nettoyage des données  
Analyse des données  
Conclusions

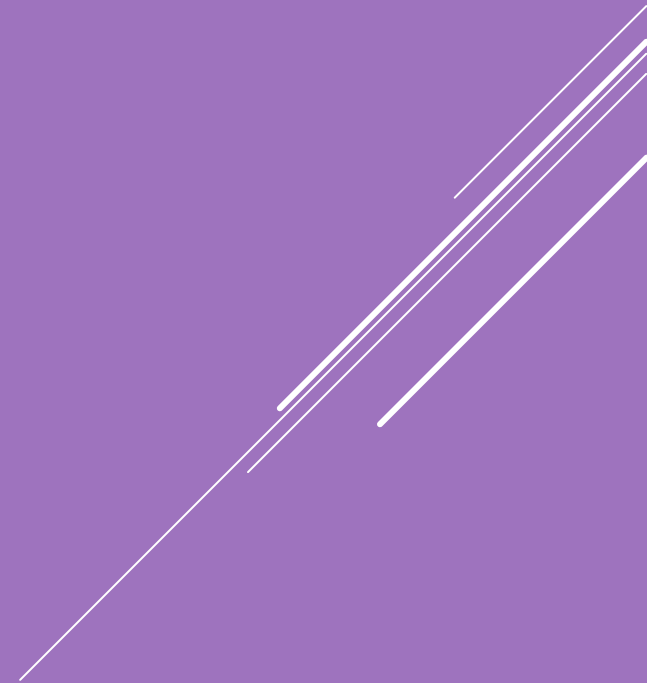


## Idée d'application

Nettoyage des données

Analyse des données

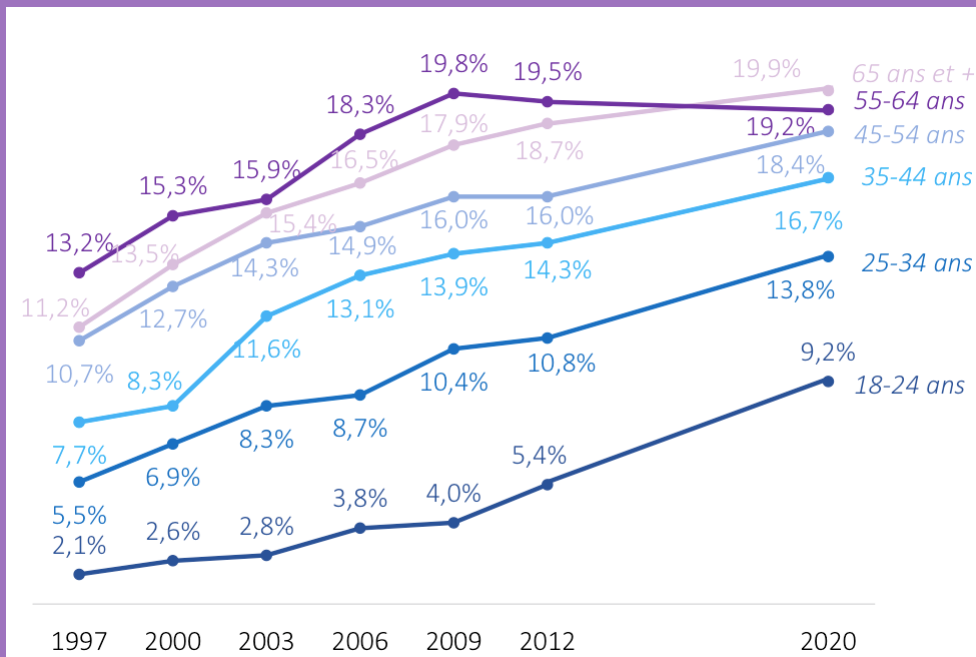
Conclusions



# Comment pouvons nous améliorer la santé de la population ?

## Problème:

L'obésité en France, tout comme le surpoids, augmente régulièrement depuis 1997.



Source: [lien Inserm](#)

### L'obésité concerne

**17%** des adultes en France  
(13% dans le monde)

### De nombreuses complications

✓ Diabète de type 2, ✓ maladies cardiovasculaires, ✓ hépatiques, ✓ rénales, ✓ cancers, ✓ respiratoires, ✓ artérielles, ✓ dermatologiques...



- des **causes** multiples
- des **mécanismes biologiques** variés pas encore tous entièrement élucidés

# Comment pouvons nous améliorer la santé de la population ?

## Démarche de résolution:

S'attaquer à la nutrition, c'est-à-dire en favorisant le choix de produits plus sains par les consommateurs.



## Objectif à atteindre:

Le consommateurs doit obtenir le nutriscore de n'importe quel aliment lors de son choix en magasin.



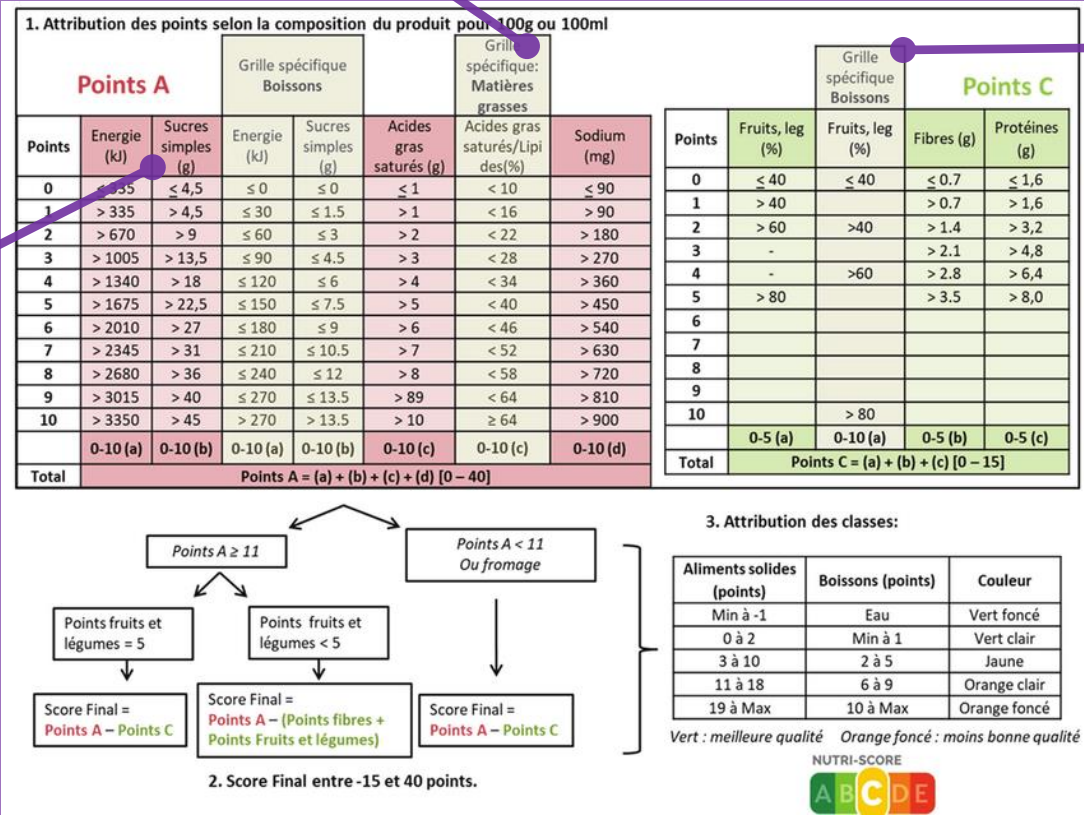
# Comment pouvons nous améliorer la santé de la population ?

## Calcul du Nutriscore suivant les catégories

Matières grasses ajoutés

Boissons

Divers produits transformés



# Comment pouvons nous améliorer la santé de la population ?

## Cas 1: Utilisation de barre code



Prend une photo

Consommateur

App  
Mobile

API

Serveur

Service intelligent

Base de donnée

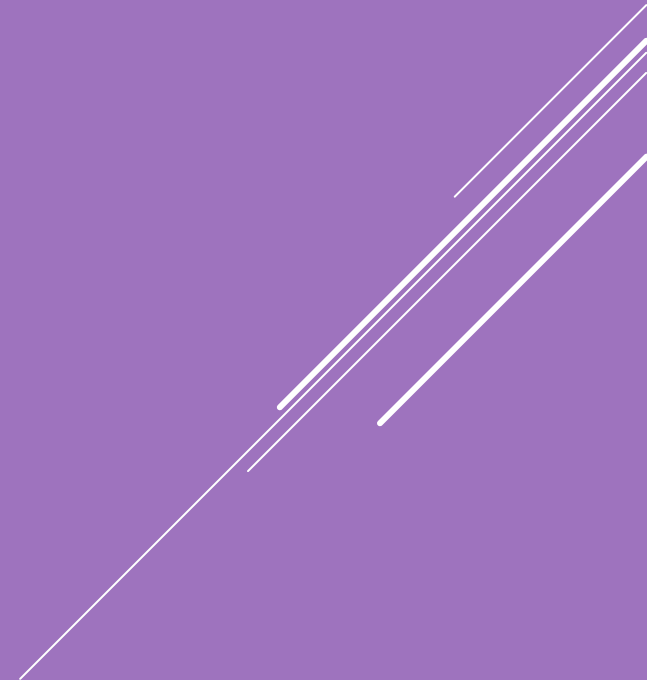
## Cas 2: Utilisation des nutriments



Prend une photo

Une idée d'application au service de la santé publique

Idée d'application  
**Nettoyage des données**  
Analyse des données  
Conclusions



# Nettoyage des données

## Le jeu de données:

320772 produits et 162 variables.

Nb variable numérique: 106 variables

Nb variable qualitative: 56 variables

### Fiche produit

- ❖ Code
- ❖ url
- ❖ Creator
- ❖ Created\_t

### Ingrédients et additif

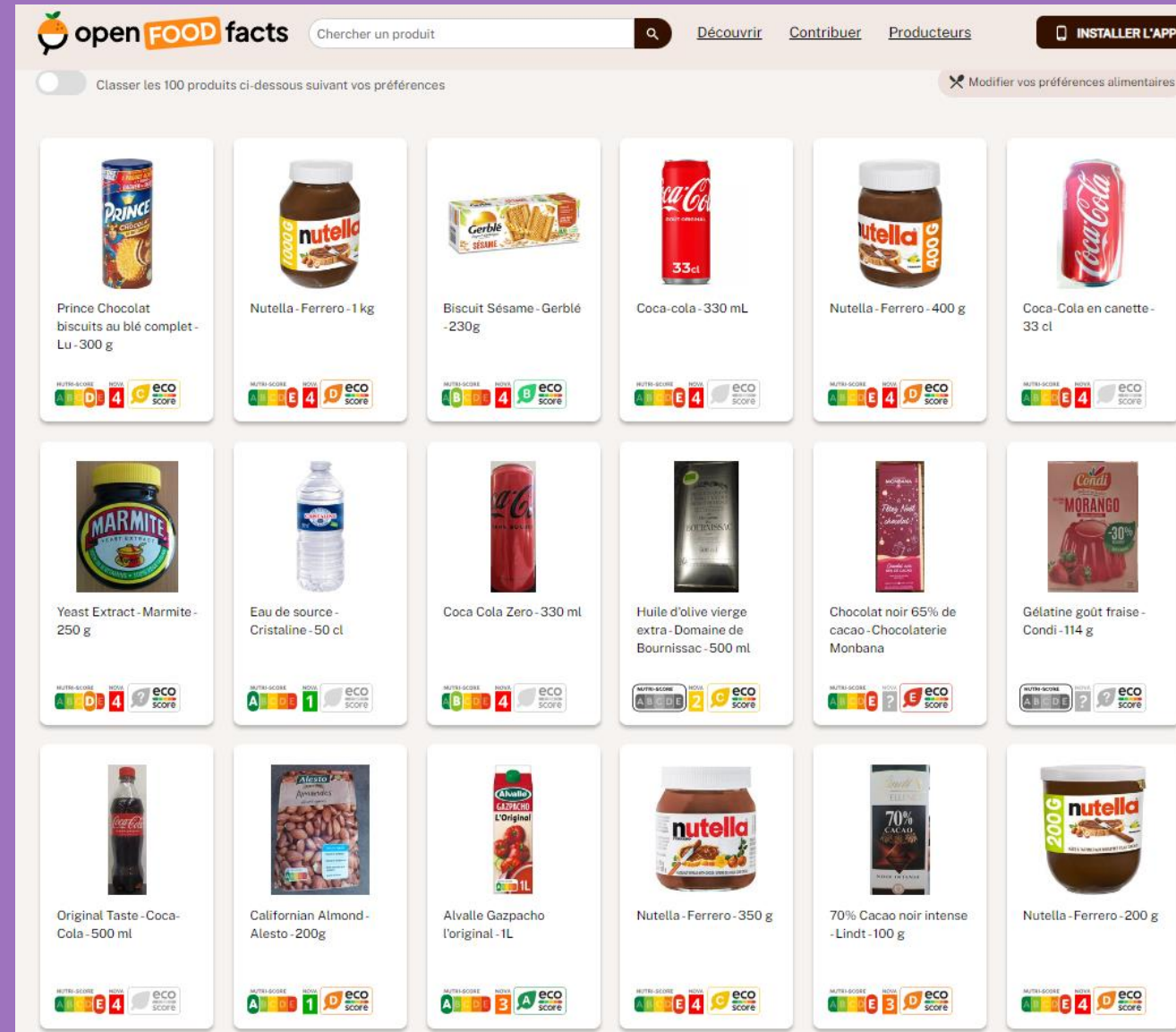
- ❖ ingredients\_text
- ❖ allergens
- ❖ additives

### Tags

- ❖ packaging\_tags
- ❖ brand\_tags
- ❖ categories\_tags
- ❖ origin\_tags

### Informations nutritionnelles

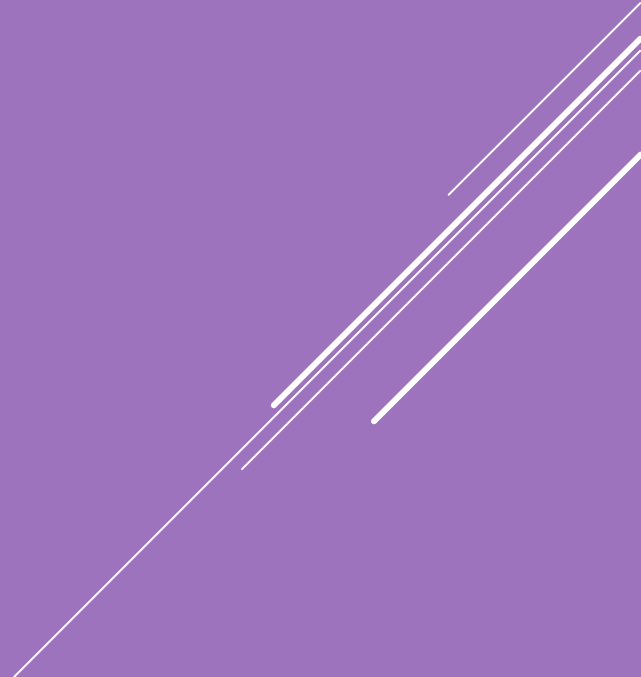
- ❖ sugars\_100g
- ❖ fat\_100g
- ❖ sodium\_100g





# Nettoyage des données

## ETAPES:

- ❖ Choix de la variable cible
  - ❖ Filtrage des variables sur les colonnes + Suppression des colonnes redondantes
  - ❖ Filtrage des variables sur les lignes + Suppression des lignes en doublons
  - ❖ Identifier et traiter les valeurs aberrantes
  - ❖ Identification de la méthode interquartile pour les valeurs extrêmes
  - ❖ Traitement des valeurs manquantes
- 
- Several white lines of varying lengths and orientations are positioned in the bottom right corner of the slide, creating a modern, abstract graphic element.

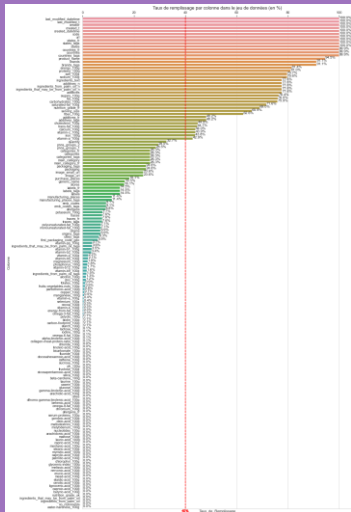
# Nettoyage des données:

**Variables cibles candidats:** 'nutrition\_grade\_fr' , 'nutrition-score-fr\_100g' et 'nutrition-score-uk\_100g'

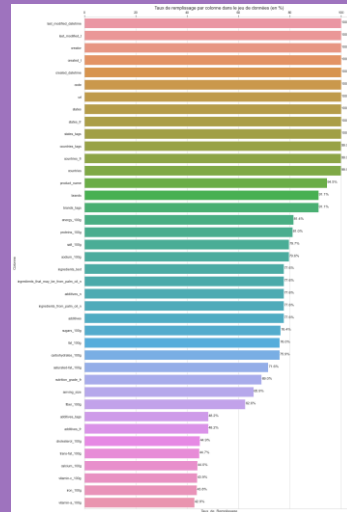
**Choix de la variable cible:** nutrition\_grade\_fr, nb de **variables restantes: 160**

**Filtrage sur les colonnes:**

Approche technique:



Seuil de 40% de taux de remplissage



320772 produits et 160 variables

320772 produits et 40 variables

Approche métier:

**Variables numériques:**

- energy\_100g
- fat\_100g
- saturated-fat\_100g
- sugars\_100g
- fiber\_100g
- proteins\_100g
- salt\_100g
- sodium\_100g

**Variables qualitatives:**

- Code
- product\_name
- Brands
- countries\_fr
- created\_t
- last\_modified\_t
- created\_datetime
- last\_modified\_datetime
- nutrition\_grade\_fr

320772 produits et 17 variables

**Suppression des colonnes redondantes:**

- ▶ sodium\_100g (corrélation 100% à salt\_100g )
- ▶ created\_t, last\_modified\_t (format UNIX timestamp, redondantes aux colonnes created\_datetime et last\_modified\_datetime)

**Nb de variables restants: 14**

# Nettoyage des données:

## Filtrages sur les lignes:

- Filtrage sur les données Française ('countries\_fr' == 'France') + Suppression de la colonne 'countries\_fr'
  - ❖ Nb de lignes et colonnes restantes: **97448 lignes et 13 colonnes.**
- Suppression des lignes avec des valeurs manquantes pour les variables: 'code', 'product\_name', 'brands'
  - ❖ Nb de lignes restantes: **84433.**
- Suppression des lignes ne présentant pas de valeurs cible 'nutrition\_grade\_fr'
  - ❖ Nb de lignes restantes: **60156.**
- Suppression des doublons par rapport au primary\_key = ['code', 'product\_name', 'brands'] : **0** lignes dupliquées
  - ❖ Nb de lignes restantes: **60156.**
- Faire une jointure à gauche entre la df catégorie '**pnns\_groups\_1**' avec la df initiale
  - ❖ Nb de lignes et colonnes restantes: **60156 lignes et 14 colonnes.**
- Suppression des valeurs manquantes de la variable '**pnns\_groups\_1**'
  - ❖ Nb de lignes et colonnes restantes: **46032 lignes et 14 colonnes.**

# Nettoyage des données:

## Identifier et traiter les valeurs aberrantes:

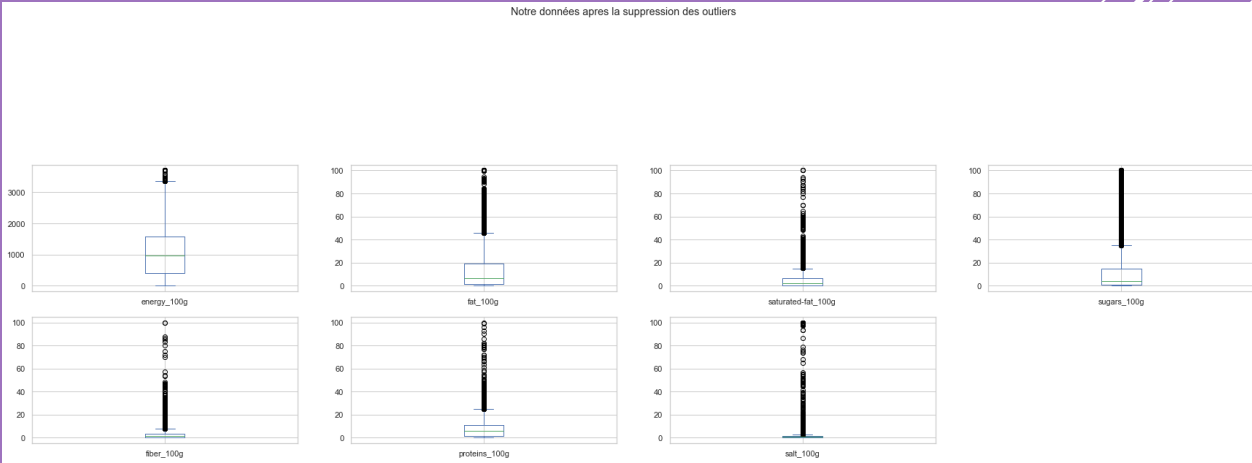
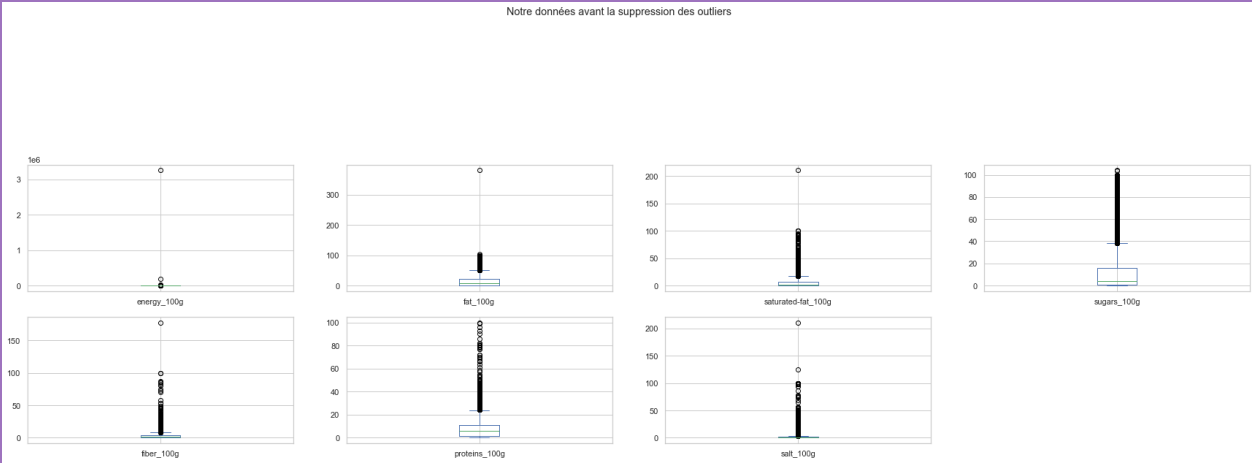
- Variables numérique:
  - $0 < \text{energy\_100g} < 3700 \text{ KJ}$
  - $0 < [\text{fat\_100g}, \text{saturated-fat\_100g}, \text{sugars\_100g}, \text{fiber\_100g}, \text{proteins\_100g}, \text{salt\_100g}] < 100\text{g}$
  - Somme**  $[\text{fat\_100g}, \text{saturated-fat\_100g}, \text{sugars\_100g}, \text{fiber\_100g}, \text{proteins\_100g}, \text{salt\_100g}] < 100\text{g}$
- Nb de lignes et colonnes restantes: **44459 lignes et 14 colonnes.**

## Identification de la méthode interquartile pour les valeurs extrêmes.

- Calcule des quartiles:

	energy_100g	fat_100g	saturated-fat_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	44383.000000	41120.000000	44383.000000	44383.000000	28378.000000	44383.000000	44383.000000
mean	1047.217682	12.008131	4.701192	12.154168	2.547956	7.615982	1.029807
std	727.078491	14.902918	7.161231	17.970258	3.872329	7.461263	3.335433
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	402.500000	1.200000	0.300000	1.000000	0.300000	1.700000	0.080010
50%	970.000000	6.200000	1.700000	3.700000	1.600000	6.000000	0.600000
75%	1587.000000	19.000000	6.200000	14.500000	3.300000	11.000000	1.270000
max	3700.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
Q1	402.500000	1.200000	0.300000	1.000000	0.300000	1.700000	0.080010
Q3	1587.000000	19.000000	6.200000	14.500000	3.300000	11.000000	1.270000
Borne_superieure	3363.750000	45.700000	15.050000	34.750000	7.800000	24.950000	3.054985
Nb_produits_sup_Borne_superieure	180.000000	1439.000000	3976.000000	5003.000000	1569.000000	1540.000000	1896.000000

Décision: Garder les valeurs extrêmes.



# Nettoyage des données:

## Traitement des valeurs manquantes:

### Avant traitement

	Colonne	Taux_de_Remplissage
0	code	100.000000
1	product_name	100.000000
2	brands	100.000000
3	pnnns_groups_1	100.000000
4	last_modified_datetime	100.000000
5	nutrition_grade_fr	100.000000
6	created_datetime	99.997751
7	energy_100g	99.829056
8	saturated-fat_100g	99.829056
9	sugars_100g	99.829056
10	proteins_100g	99.829056
11	salt_100g	99.829056
12	fat_100g	92.489710
13	fiber_100g	63.829596

### Approche métier

Remplacement des valeurs manquantes par 0:  
Quelques exemples:

- milk and dairy products (Lait et produits laitiers) en contient pas de fibres)
- la catégorie 'beverage (Boissons") ne contient ni fibres, ni graisses, ni graisses saturées,
- .....

	Colonne	Taux_de_Remplissage
0	code	100.000000
1	product_name	100.000000
2	brands	100.000000
3	pnnns_groups_1	100.000000
4	last_modified_datetime	100.000000
5	nutrition_grade_fr	100.000000
6	created_datetime	99.997751
7	saturated-fat_100g	99.889786
8	proteins_100g	99.835804
9	energy_100g	99.829056
10	sugars_100g	99.829056
11	salt_100g	99.829056
12	fat_100g	94.894172
13	fiber_100g	83.256484

### Imputation par `IterativeImputer()`

#### Corrélation:

- energy\_100g et fat\_100g à 0.74.
- energy\_100g et saturated-fat\_100g à 0.57.
- fat\_100g et saturated-fat\_100g à 0.73.

	Colonne	Taux_de_Remplissage
0	code	100.000000
1	product_name	100.000000
2	brands	100.000000
3	pnnns_groups_1	100.000000
4	last_modified_datetime	100.000000
5	energy_100g	100.000000
6	fat_100g	100.000000
7	saturated-fat_100g	100.000000
8	nutrition_grade_fr	100.000000
9	created_datetime	99.997751
10	proteins_100g	99.835804
11	sugars_100g	99.829056
12	salt_100g	99.829056
13	fiber_100g	83.256484

### Approche technique

#### Imputation par médiane

- Pour chaque catégories de la pnnns\_groups\_1 à part la catégorie 'unknown'.

	Colonne	Taux_de_Remplissage
0	code	100.000000
1	product_name	100.000000
2	brands	100.000000
3	pnnns_groups_1	100.000000
4	last_modified_datetime	100.000000
5	energy_100g	100.000000
6	fat_100g	100.000000
7	saturated-fat_100g	100.000000
8	nutrition_grade_fr	100.000000
9	created_datetime	99.997751
10	sugars_100g	99.943768
11	proteins_100g	99.943768
12	salt_100g	99.943768
13	fiber_100g	95.474482

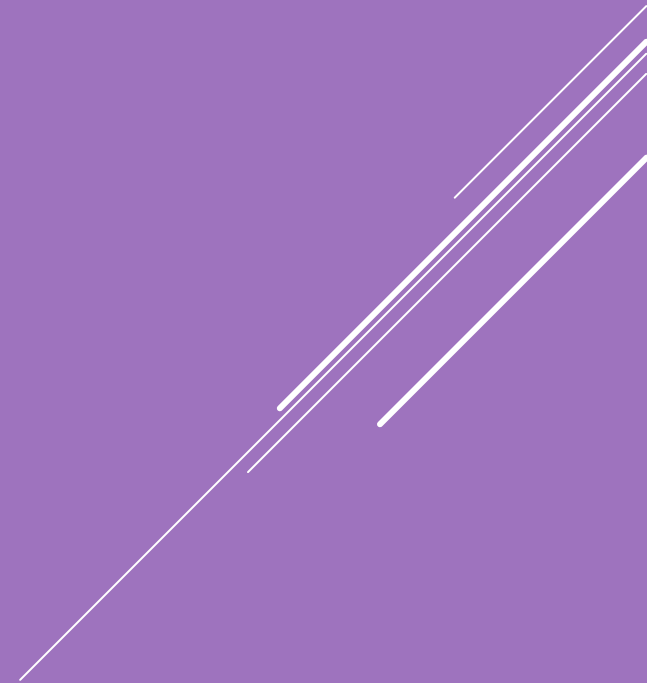
#### Imputation par Knn:

- Uniquement pour la catégorie 'unknown' de la variable pnnns\_groups\_1

	Colonne	Taux_de_Remplissage
0	code	100.000000
1	product_name	100.000000
2	brands	100.000000
3	pnnns_groups_1	100.000000
4	last_modified_datetime	100.000000
5	energy_100g	100.000000
6	fat_100g	100.000000
7	saturated-fat_100g	100.000000
8	sugars_100g	100.000000
9	fiber_100g	100.000000
10	proteins_100g	100.000000
11	salt_100g	100.000000
12	nutrition_grade_fr	100.000000
13	created_datetime	99.997751

- Suppression des valeur manquantes pour la variable created\_datetime
  - ❖ Nb de lignes et colonnes restantes: **44458 lignes et 14 colonnes.**

Idée d'application  
Nettoyage des données  
**Analyse des données**  
Conclusions



# Analyse des données

## ETAPES:

- ❖ Analyse uni-variée
  - Variables numériques
  - Variables catégorielles
- ❖ Analyse bivariable
  - Variables numériques entre elles:
  - Variables numériques / Variables catégorielles: **ANOVA (Eta\_square)**
  - Variables catégorielles entre elles: **Le test du khi-deux**
- ❖ Analyse multivariée
  - Réalisation d'un éboulis
  - Trouver le nombre de composantes principales à utiliser
  - Cercle de corrélation

# Analyse des données

## Analyse uni-variée: variables numériques

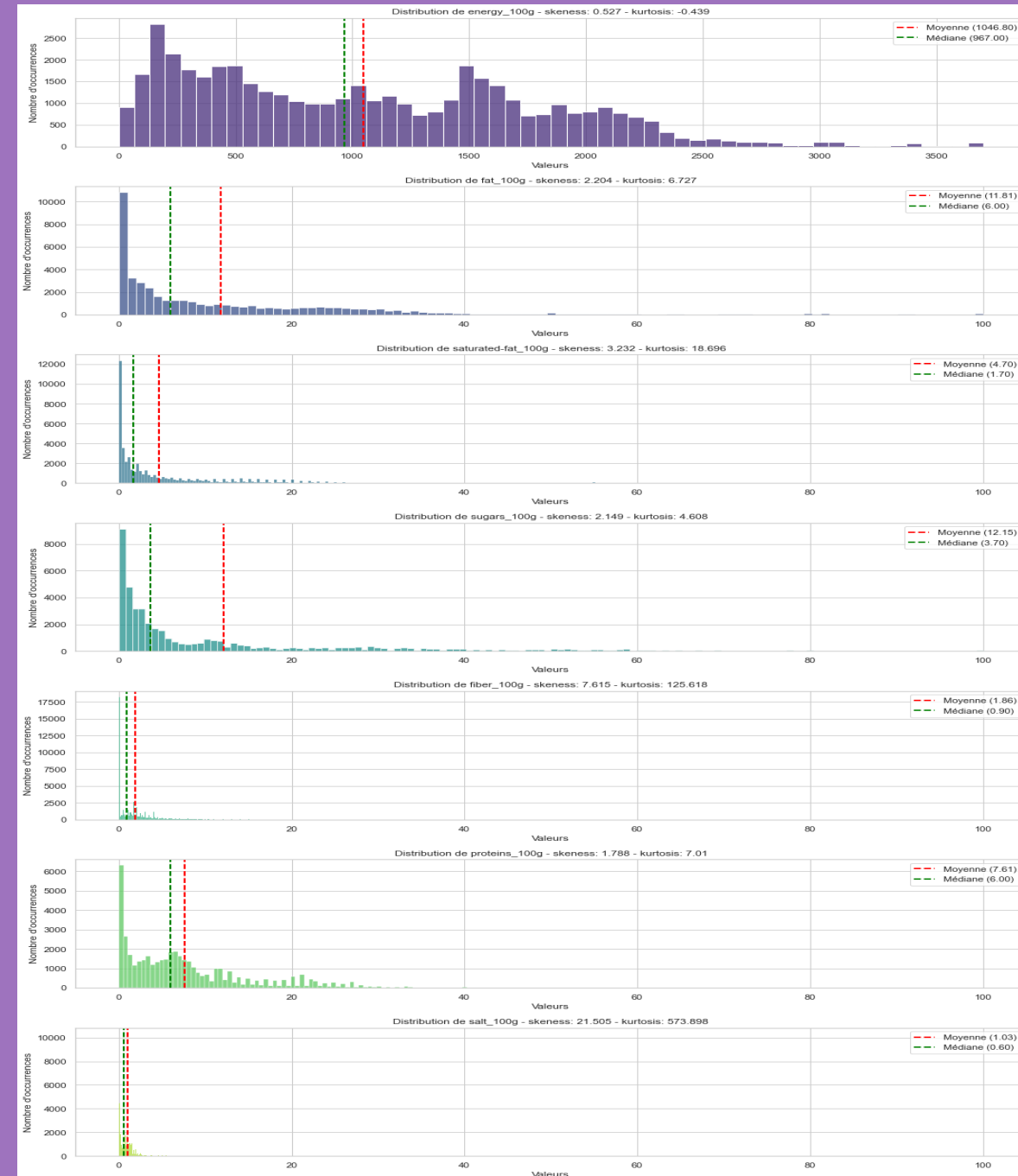
Affichages des statistiques:

	energy_100g	fat_100g	saturated-fat_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	44458.000000	44458.000000	44458.000000	44458.000000	44458.000000	44458.000000	44458.000000
mean	1046.796539	11.811550	4.697265	12.147767	1.856926	7.611332	1.029166
std	726.629675	14.981254	7.156183	17.956272	3.313629	7.458158	3.332732
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	403.223922	1.000000	0.300000	1.000000	0.000000	1.700000	0.080010
50%	967.000000	6.000000	1.700000	3.700000	0.900000	6.000000	0.600000
75%	1586.000000	18.600000	6.200000	14.500000	2.500000	11.000000	1.270000
max	3700.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
skewness	0.527000	2.204000	3.232000	2.149000	7.615000	1.788000	21.505000
kurtosis	-0.439000	6.727000	18.696000	4.608000	125.618000	7.010000	573.898000

Exemples:

- energy\_100g:
  - ❖ distribution légèrement asymétrique à droite
  - ❖ aplatissement inférieur à celui d'une distribution normale
- sugars\_100g:
  - ❖ distribution fortement asymétrique à droite
  - ❖ aplatissement supérieur à celui d'une distribution normale
- fiber\_100g:
  - ❖ distribution extrêmes asymétrique à droite
  - ❖ aplatissement très supérieur à celui d'une distribution normale

## Graphique de distribution

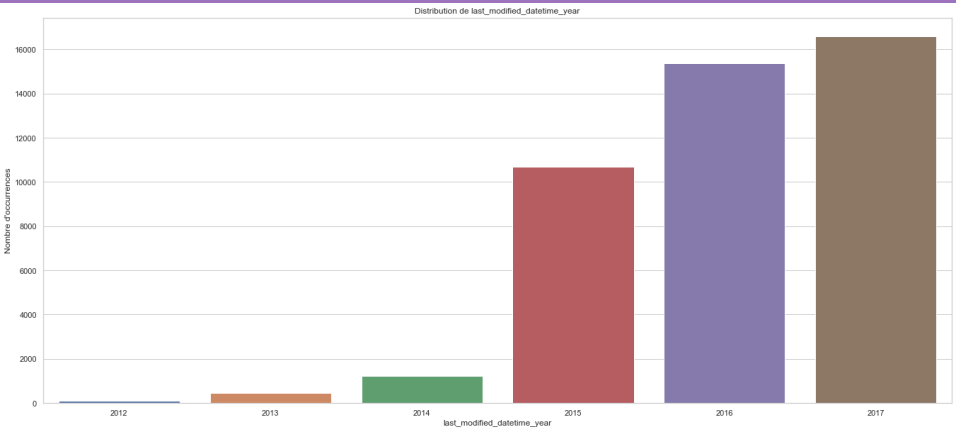
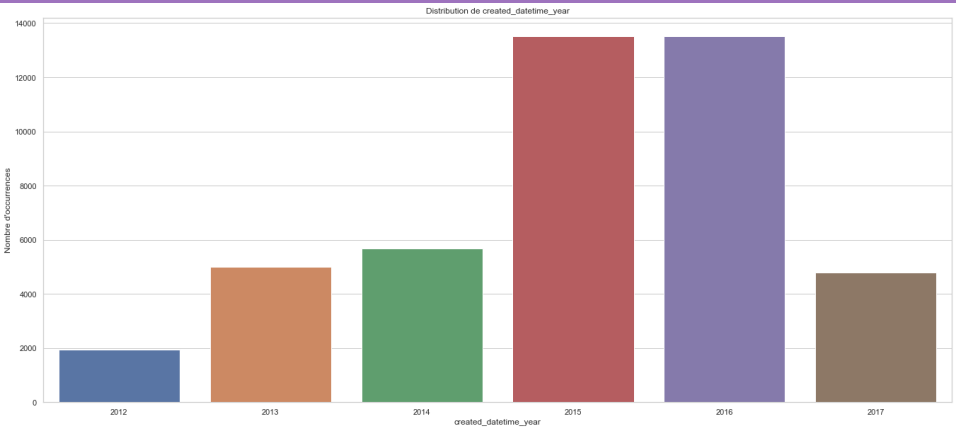




# Analyse des données

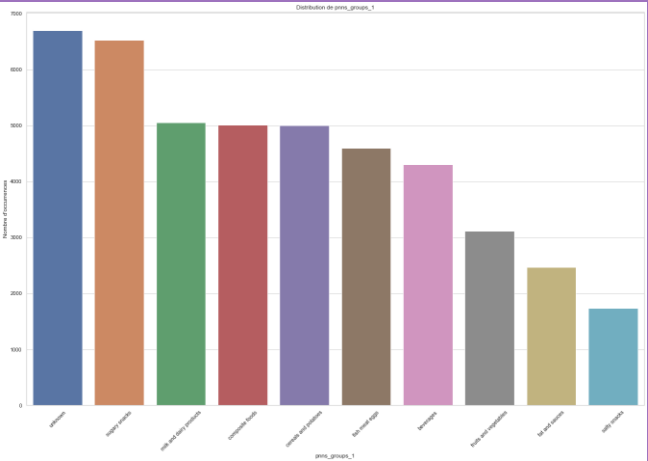
## Analyse uni-variée: variables catégorielles

Graphiques datetime:



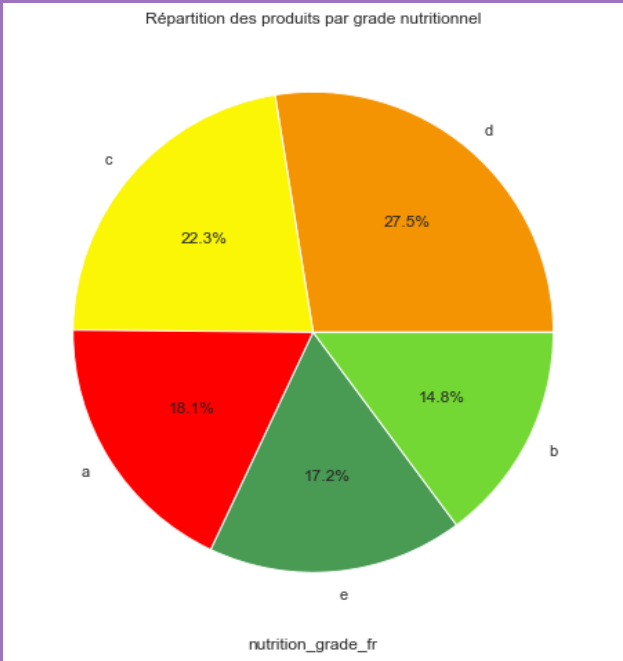
La majorité des produits ont été introduites en 2015, et modifiés en 2017

## Graphique de distribution barre-plot



Les 3 premières catégories:

1. Unknown: 6690 produits
2. sugary snacks: 6524 produits
3. milk and dairy products

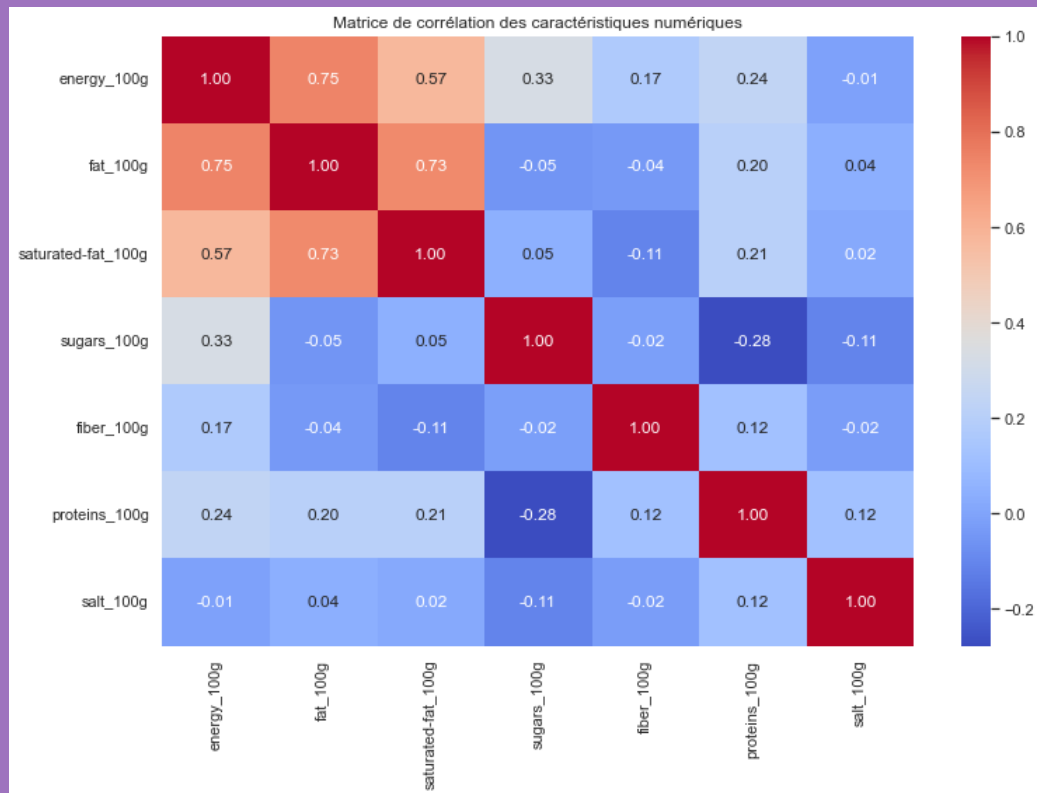


La distribution des produits dans le diagramme circulaire des catégories est généralement déséquilibrée

# Analyse des données

## Analyse bvariée: variables numériques entre elles

Graphique Matrice de corrélation:



Corrélation Forte:

- ❖ entre 'energy\_100g' et 'fat\_100g'
- ❖ entre 'saturated-fat\_100g' et 'fat\_100g'

Corrélation modéré:

- ❖ entre 'energy\_100g' et 'saturated-fat\_100g'
- ❖ entre 'energy\_100g' et 'sugars\_100g'

Corrélation faible:

- ❖ entre 'energy\_100g' et 'fiber\_100g'

Graphique nuages de points

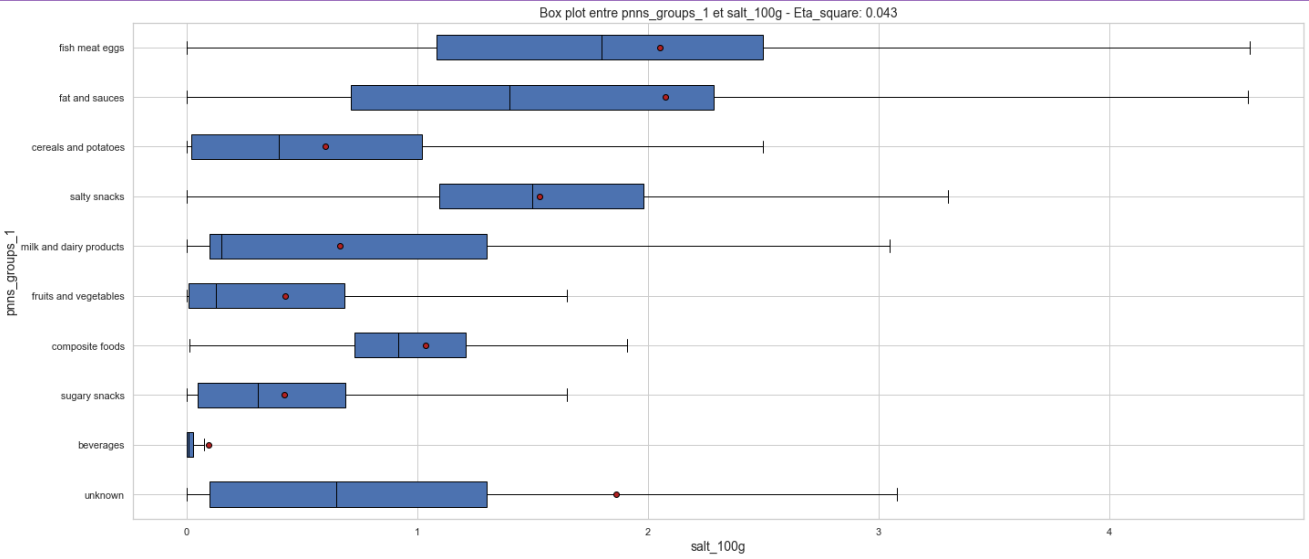
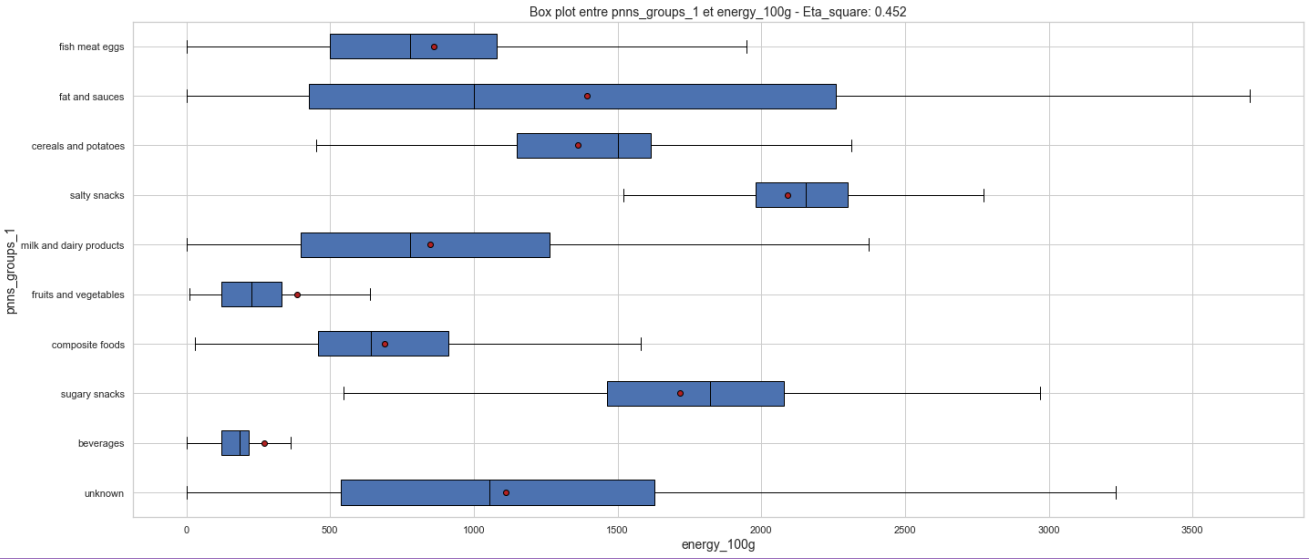


Les produits avec nutrition\_grade 'e' (nuages de points rouges et oranges) ont une tendance plus élevée en énergie

# Analyse des données

## Analyse bivariée: variables numérique / Variables catégorielle

Graphique ANOVA:



variables	ETA_Square / pnns_groups_1	Observations
energy_100g	0,452	une proportion importante de la variance peut être expliquée par les catégories du 'pnns_groups_1'
fat_100g	0,286	une proportion modérée de la variance peut être expliquée par les catégories du 'pnns_groups_1'
saturated-fat_100g	0,19	une proportion modérée de la variance peut être expliquée par les catégories du 'pnns_groups_1'
sugars_100g	0,423	une proportion importante de la variance peut être expliquée par les catégories du 'pnns_groups_1'
fiber_100g	0,202	une proportion modérée de la variance peut être expliquée par les catégories du 'pnns_groups_1'
proteins_100g	0,422	une proportion importante de la variance peut être expliquée par les catégories du 'pnns_groups_1'
salt_100g	0,043	une très faible proportion de la variance peut être expliquée par les catégories du 'pnns_groups_1'

# Analyse des données

## Analyse bivariée: variables catégorielle/ Variables catégorielle

Test de Khi-deux:

La table de contingence entre pnns\_groups\_1 et nutrition\_grade\_fr

nutrition_grade_fr	a	b	c	d	e	Total
pnns_groups_1						
beverages	149	536	1239	698	1679	4301
cereals and potatoes	2542	805	1008	583	59	4997
composite foods	1175	1593	1368	803	65	5004
fat and sauces	74	188	685	1072	445	2464
fish meat eggs	428	590	1300	1347	925	4590
fruits and vegetables	2225	550	303	27	1	3106
milk and dairy products	372	914	1316	2232	212	5046
salty snacks	39	62	459	893	283	1736
sugary snacks	50	256	668	2681	2869	6524
unknown	1009	1095	1577	1906	1103	6690
Total	8063	6589	9923	12242	7641	44458

Résultats:

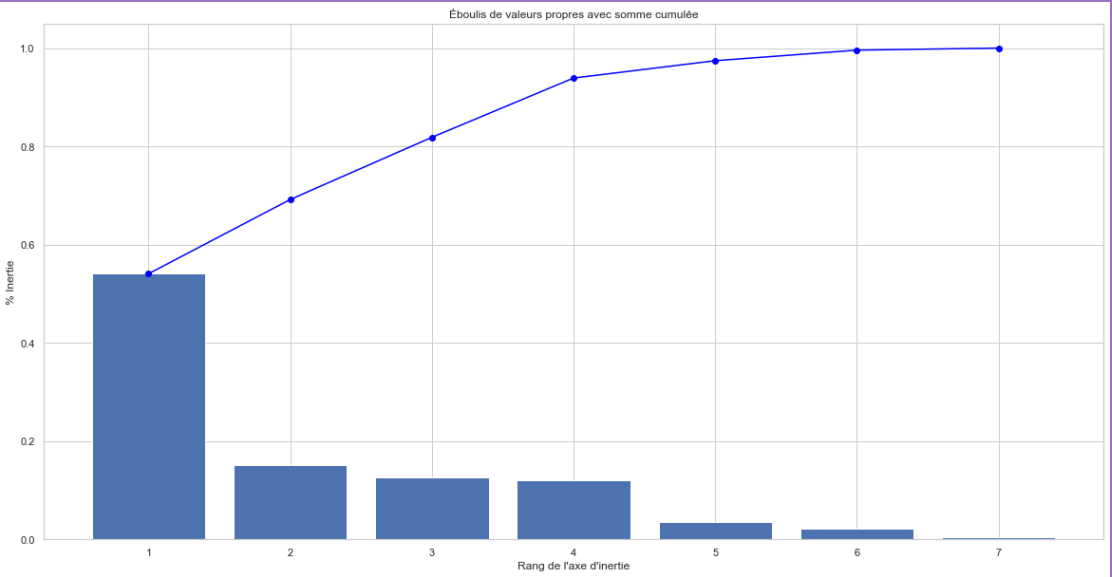
La valeur du chi square: 23113.449888501083

La valeur du P value: 0.0

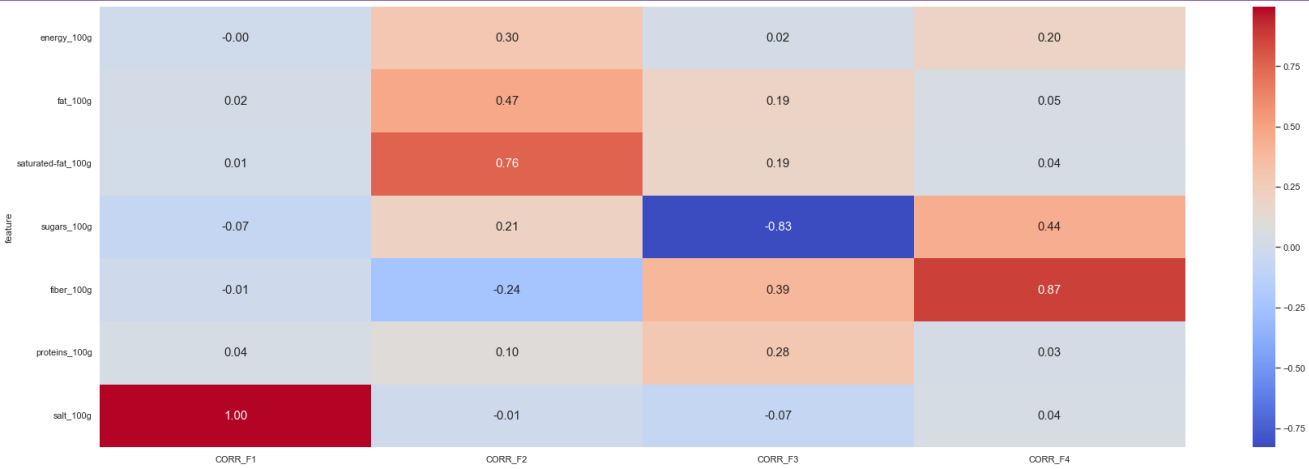
Hypothèses	khi-deux de Pearson
H0	indépendance
H1	dépendante
	p-valeur < 0,05 ⇒ rejet de H0
Résultat	il n'y a pas suffisamment de données pour accepter H0 et dire qu'il y a une notion d'indépendance.

# Analyse des données

## Analyse multivariée: la méthode ACP



**éboulis:** conserver uniquement 4 composantes qui expliquent 90% de la variance

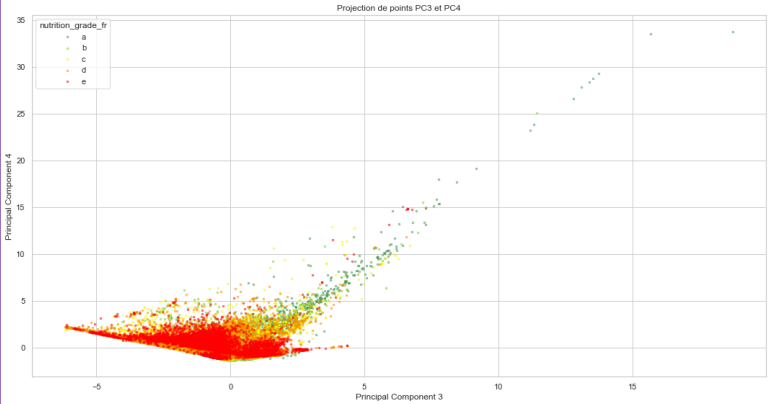
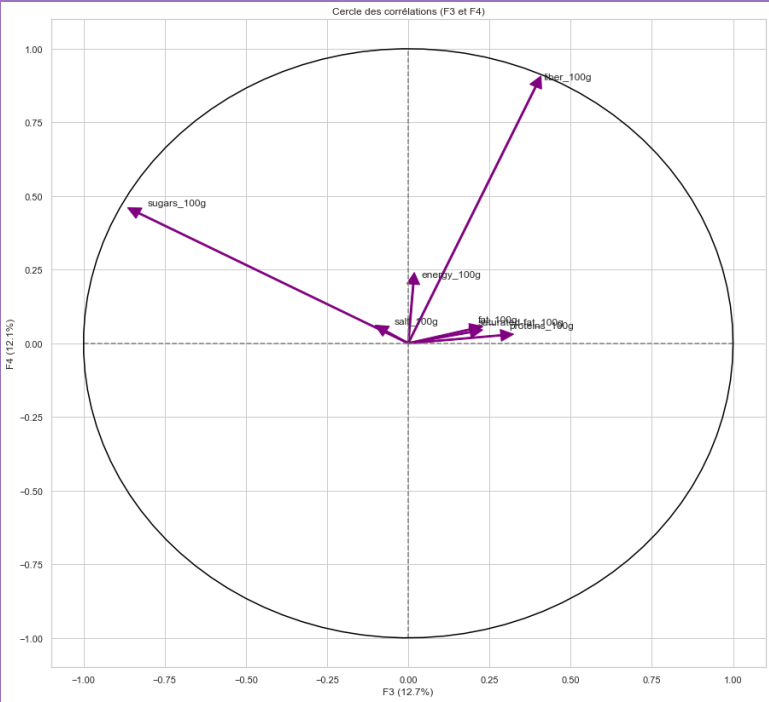


**Heatmap:** coefficient de corrélation entre les variables vs composantes principales

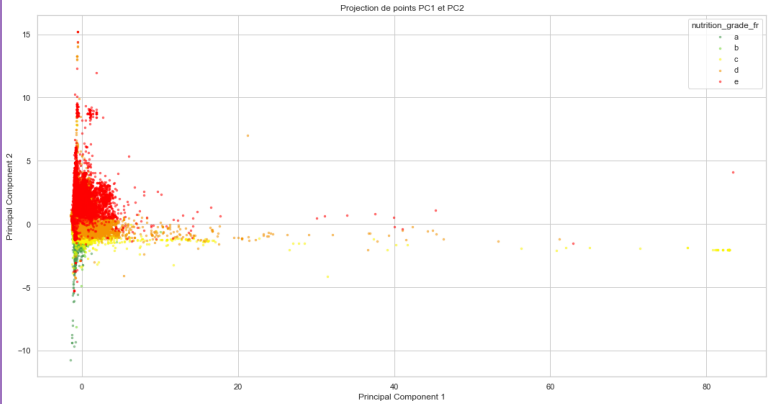
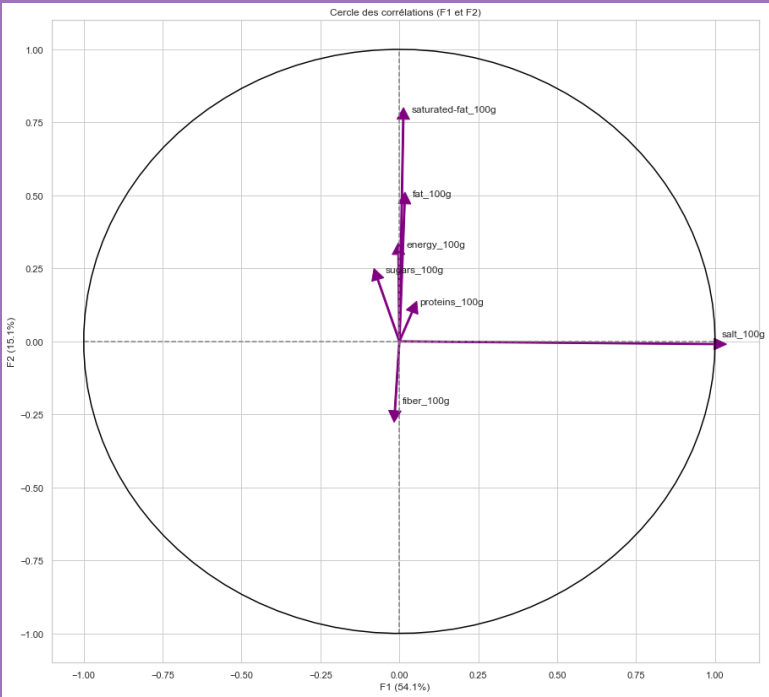
# Analyse des données

## Analyse multivariée:

Produits riches en fibres

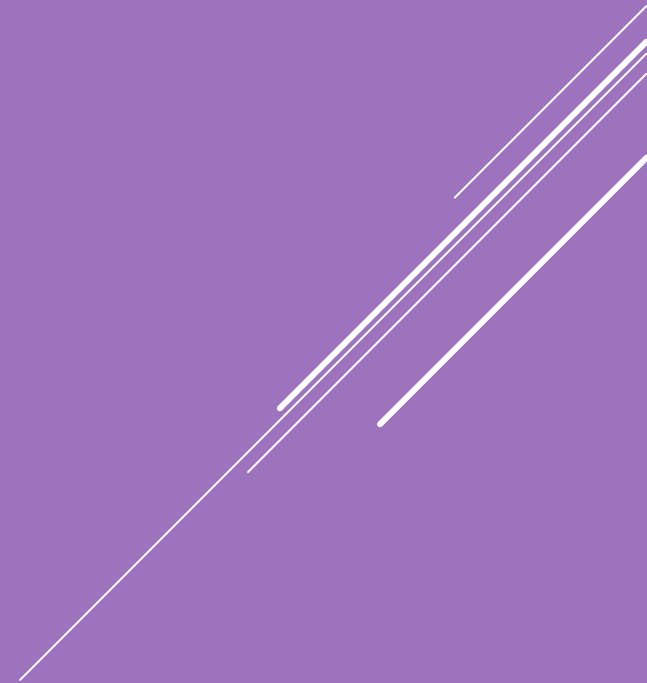


Produits gras et gras saturés



Produits très salés

Idée d'application  
Nettoyage des données  
Analyse des données  
**Conclusions**



# Conclusions

## La base de donnée nettoyée

- ❖ 44458 produits différents.
- ❖ 4 variables quantitatives principales.

## Lien entre les variables:

- ❖ Le nutrition grade dépend principalement de l'énergie et de la teneur en graisses.
- ❖ La catégorie est un bon guide pour améliorer sa nutrition.
- ❖ Certaines marques proposent des produits plus sains, Exemple: cristallines, Alvalle Gazpacho l'original,...

## Faisabilité de l'application:

- ❖ Nombre de variables suffisant avec liaisons qui ont un taux de complétion à 100%.

## Points de vigilance:

- ❖ Appui expert métier
- ❖ Biais possible lié au données et au mode de collecte.

## Cahier des charges

### Une base de donnée propre

- ✓ sans valeurs aberrantes
- ✓ sans doublons
- ✓ sans valeurs manquantes

### Un contenu adapté

- ✓ des produits identifiables
- ✓ des catégories pertinentes
- ✓ des données chiffrées utiles