Analyse des données

Application au service de la santé publique

Sommaire

1. Nettoyage des données
2. Analyses univariées
3. Analyses bivariées
4. Analyses multivariées
5. Idée d’application

1. Nettoyage des données

Aperçu des données :

	code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_
0	0000000003087	http://world-fr.openfoodfacts.org/produit/0000...	openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	
1	0000000004530	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	
2	0000000004559	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	
3	0000000016087	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salted Nut Mix	
4	0000000016094	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	

Dimensions des données: (320772, 162)

Aperçu des données :

Variables manquantes :

	Column	Non-Null Count	n-unique	Dtype	low	high	mean
energy_100g	energy_100g	261113	3997	float64	0.000000e+00	2690.000000	1141.914605
energy-from-fat_100g	energy-from-fat_100g	857	335	float64	0.000000e+00	2900.000000	585.501214
fat_100g	fat_100g	243891	3378	float64	0.000000e+00	58.000000	12.730379
saturated-fat_100g	saturated-fat_100g	229554	2197	float64	0.000000e+00	23.000000	5.129932
caprylic-acid_100g	caprylic-acid_100g	1	1	float64	7.400000e+00	7.400000	7.400000
capric-acid_100g	capric-acid_100g	2	2	float64	5.888000e+00	6.192000	6.040000
lauric-acid_100g	lauric-acid_100g	4	4	float64	3.506375e+00	49.277500	36.136182
myristic-acid_100g	myristic-acid_100g	1	1	float64	1.890000e+01	18.900000	18.900000

Mauvais dtypes :

created_t	created_t	320769	189567	object
created_datetime	created_datetime	320763	189568	object
last_modified_t	last_modified_t	320772	180495	object
last_modified_datetime	last_modified_datetime	320772	180495	object

Aperçu des données **après nettoyage** :

Dimensions des données: (288941, 34)

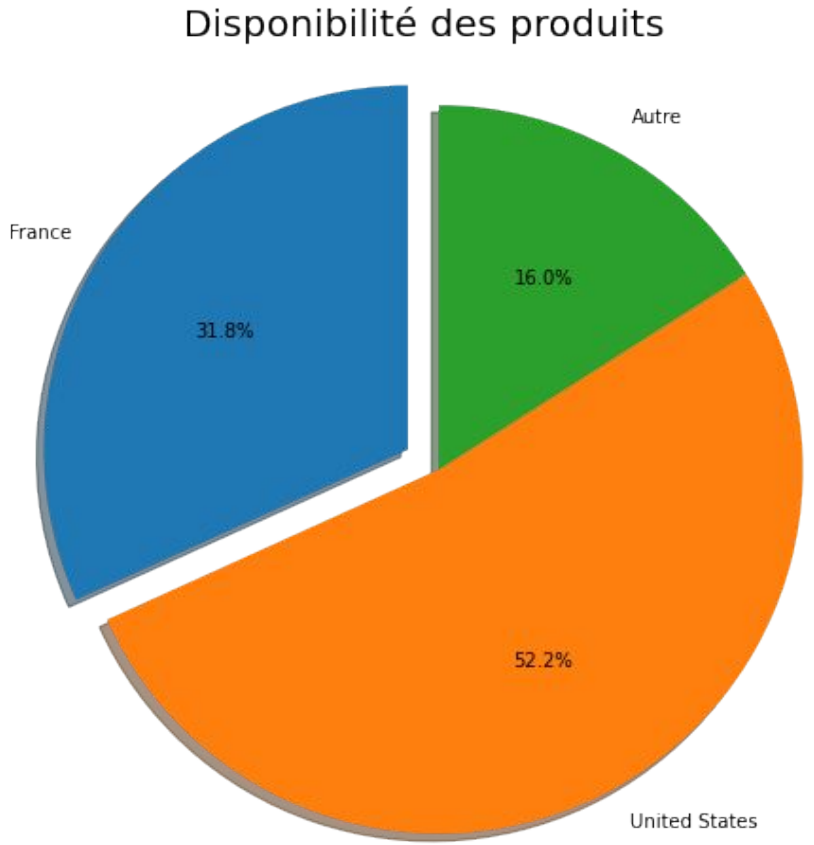
Avant nettoyage : (320772, 162)

#	Column	Non-Null Count	Dtype
0	code	288941 non-null	object
1	url	288941 non-null	object
2	creator	288941 non-null	object
3	product_name	288941 non-null	object
4	ingredients_from_palm_oil_n	288941 non-null	object
5	ingredients_that_may_be_from_palm_oil_n	288941 non-null	object
6	nutrition_grade_fr	288941 non-null	object
7	energy_100g	288941 non-null	float64
8	fat_100g	288941 non-null	float64
9	saturated-fat_100g	288941 non-null	float64
10	carbohydrates_100g	288941 non-null	float64
11	sugars_100g	288941 non-null	float64
12	proteins_100g	288941 non-null	float64
13	salt_100g	288941 non-null	float64
14	sodium_100g	288941 non-null	float64
15	date_created	288941 non-null	datetime64[ns]
16	last_modified	288941 non-null	datetime64[ns]
17	country_1	288941 non-null	object
18	country_2	288941 non-null	object
19	country_3	288941 non-null	object
20	brand_1	288941 non-null	object
21	brand_2	288941 non-null	object
22	brand_3	288941 non-null	object
23	additive_1	288941 non-null	object
24	additive_2	288941 non-null	object
25	additive_3	288941 non-null	object
26	additive_4	288941 non-null	object
27	additive_5	288941 non-null	object
28	additive_6	288941 non-null	object
29	additive_7	288941 non-null	object
30	additive_8	288941 non-null	object
31	additive_9	288941 non-null	object
32	additive_10	288941 non-null	object
33	product_group	288941 non-null	object

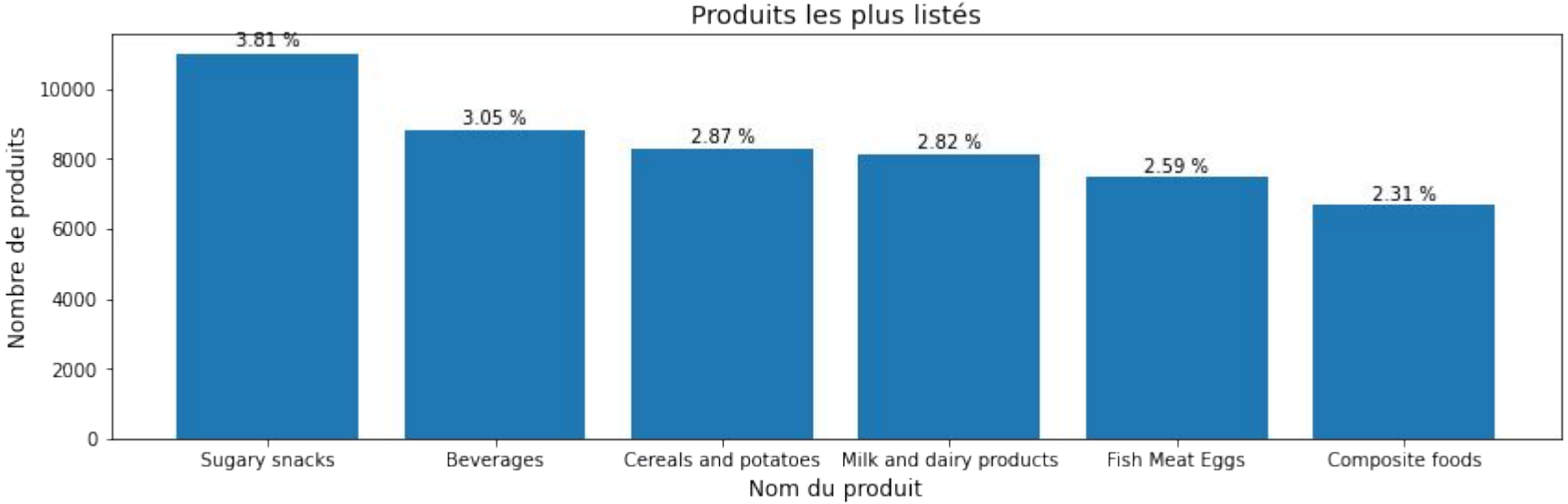
2. Analyses univariées

Disponibilité des produits par **pays**

- 84% des produits sont disponibles en France ou aux États-Unis.



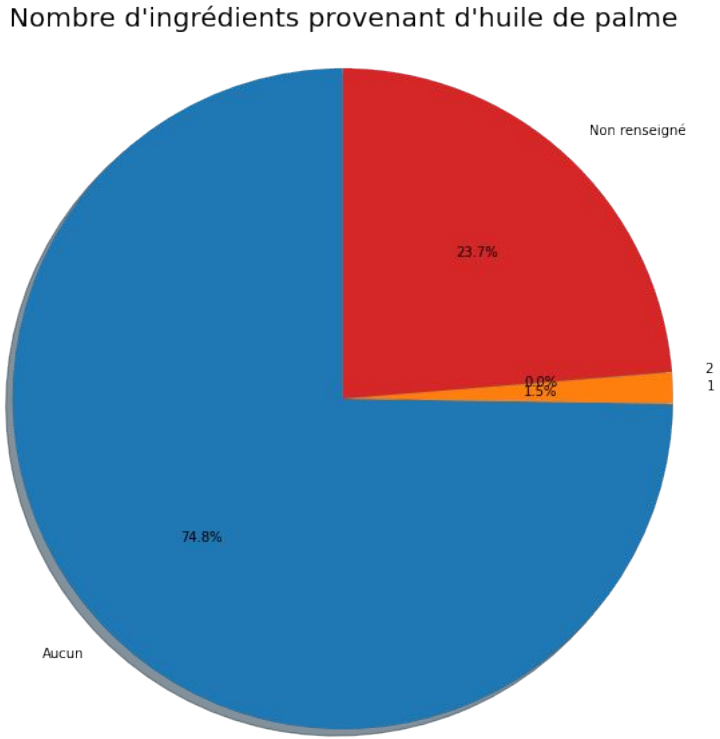
Produits les plus listés



- Le type de produit le plus listé est *sucreries*, représentant 3,81 % des produits.

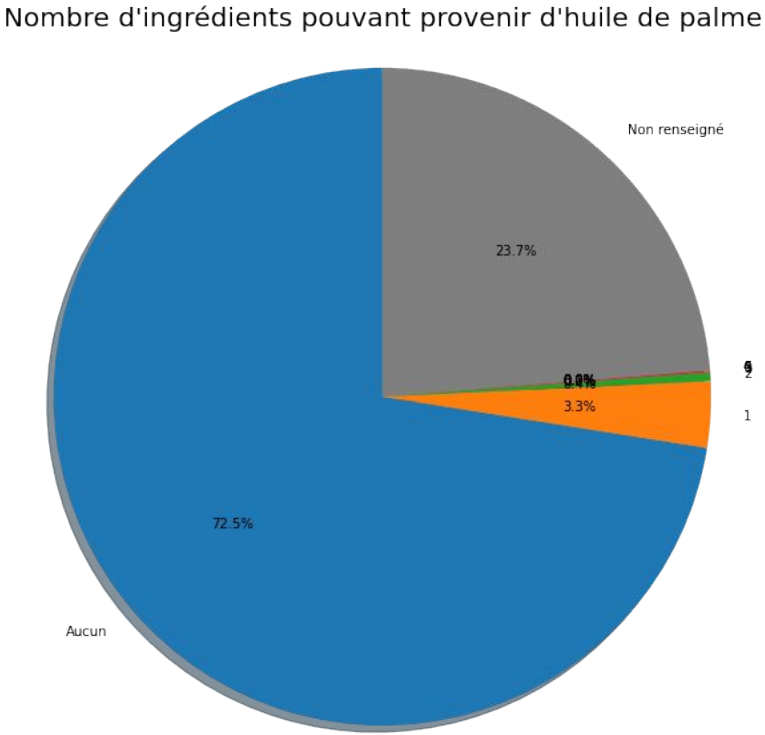
Nombre d’ingrédients provenant d’**huile de palme**

- 75% des produits ne contiennent aucun ingrédient provenant d’huile de palme
- 24% des produits n’ont pas cette information renseignée

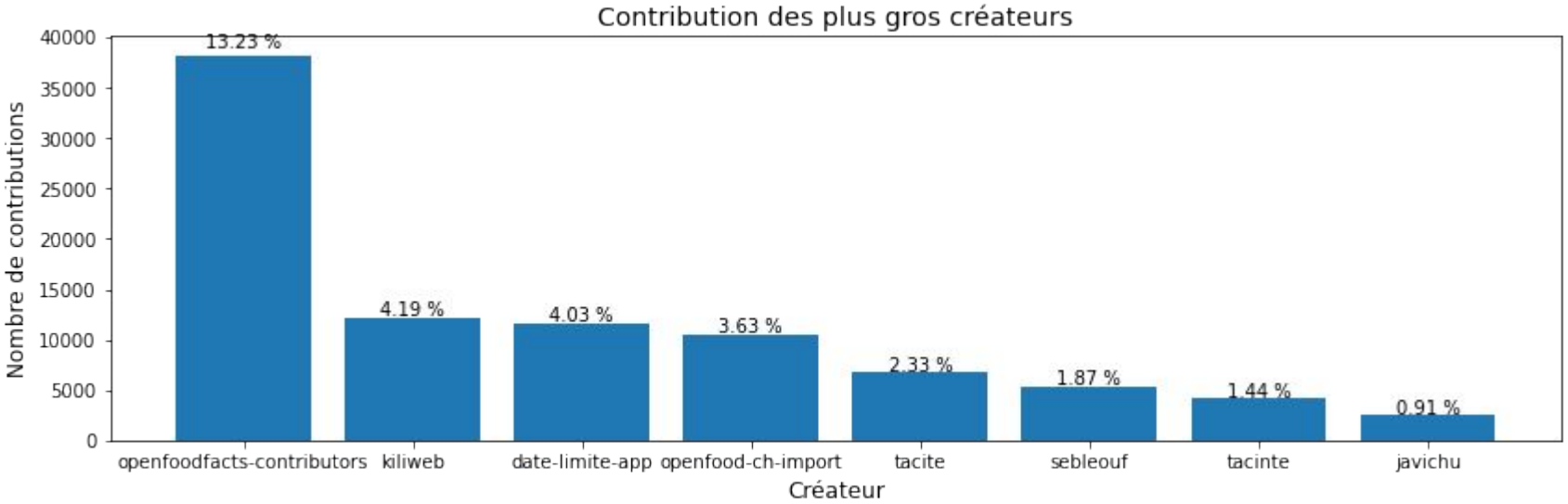


Nombre d’ingrédients pouvant provenant **d’huile de palme**

- 72,5% des produits ne contiennent aucun ingrédient provenant d’huile de palme
- 3,3% des produits pourraient en provenir, soit le double de ceux dont on est est sûr



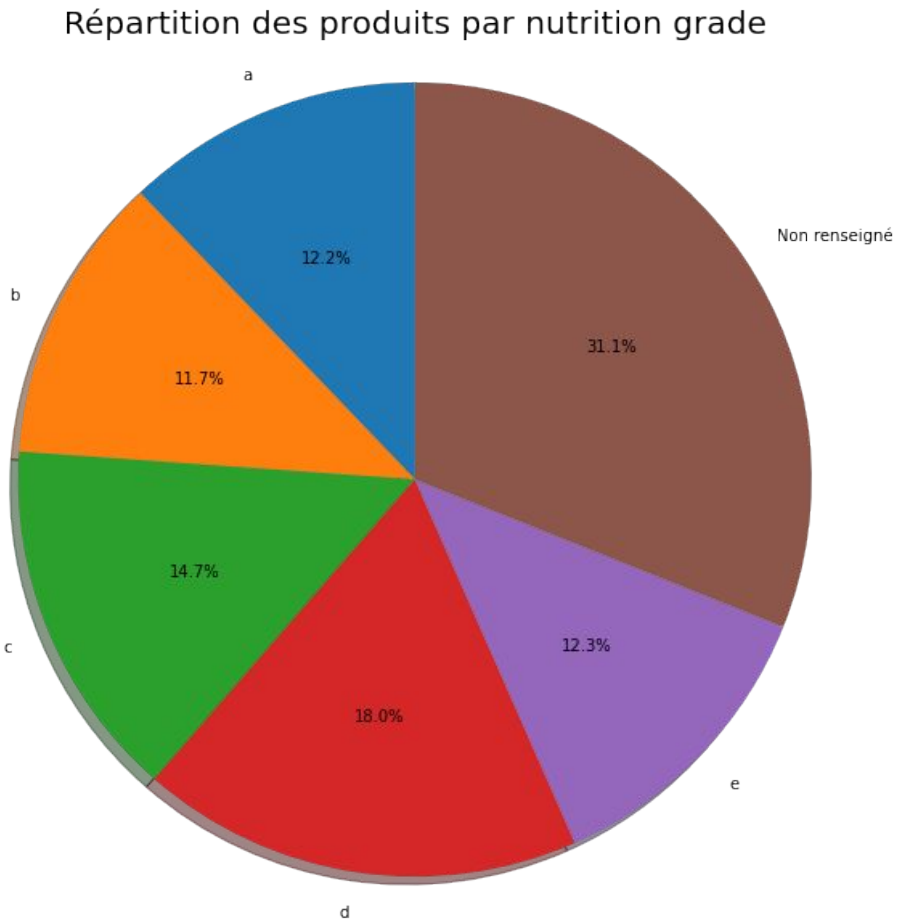
Contribution des plus gros **créateurs**



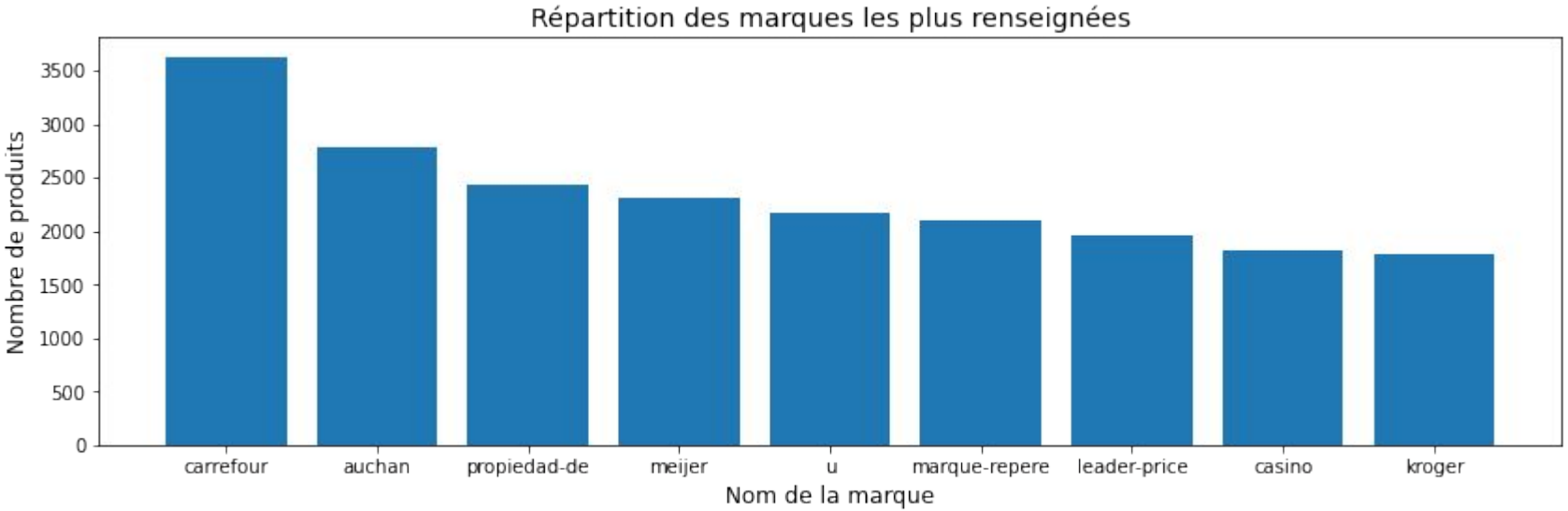
- Le créateur ayant le plus contribué à la database est *openfoodfacts-contributors* avec 13,23 % des produits, suivi de *kiliweb* avec 4,19 % des produits créés.

Répartition des produits par **nutrition grade**

- Répartition relativement égale de la variable *nutrition grade* entre les produits

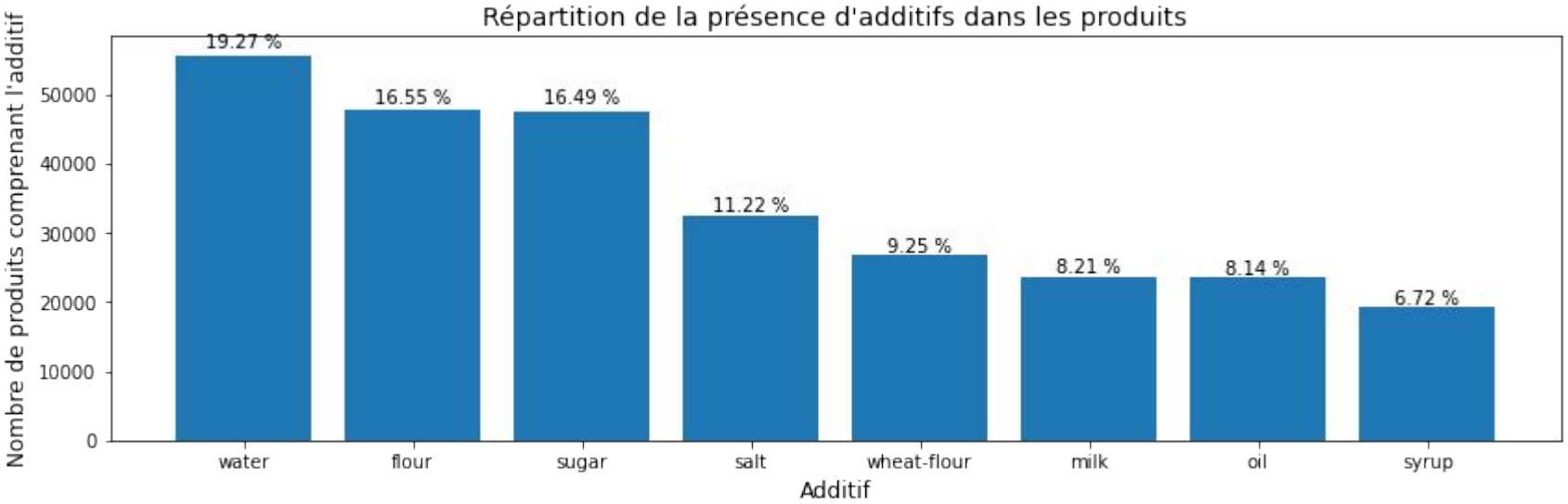


Marques les plus présentes



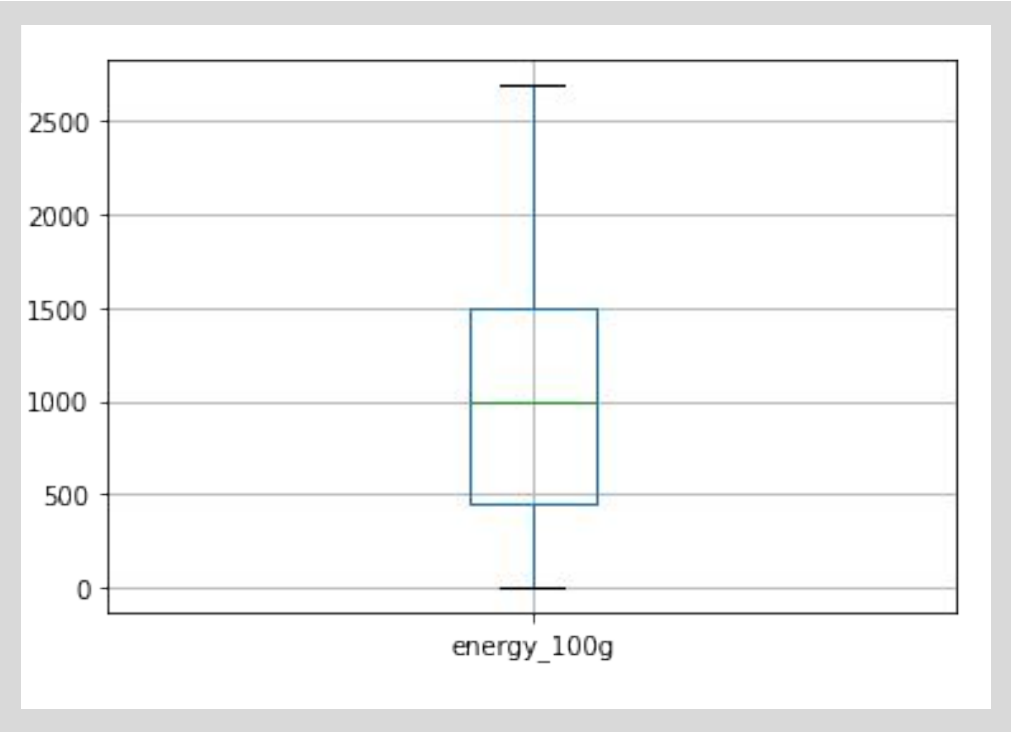
- Les deux marques les plus présentes sont *Carrefour* et *Auchan*.

Répartition des **additifs** dans les produits

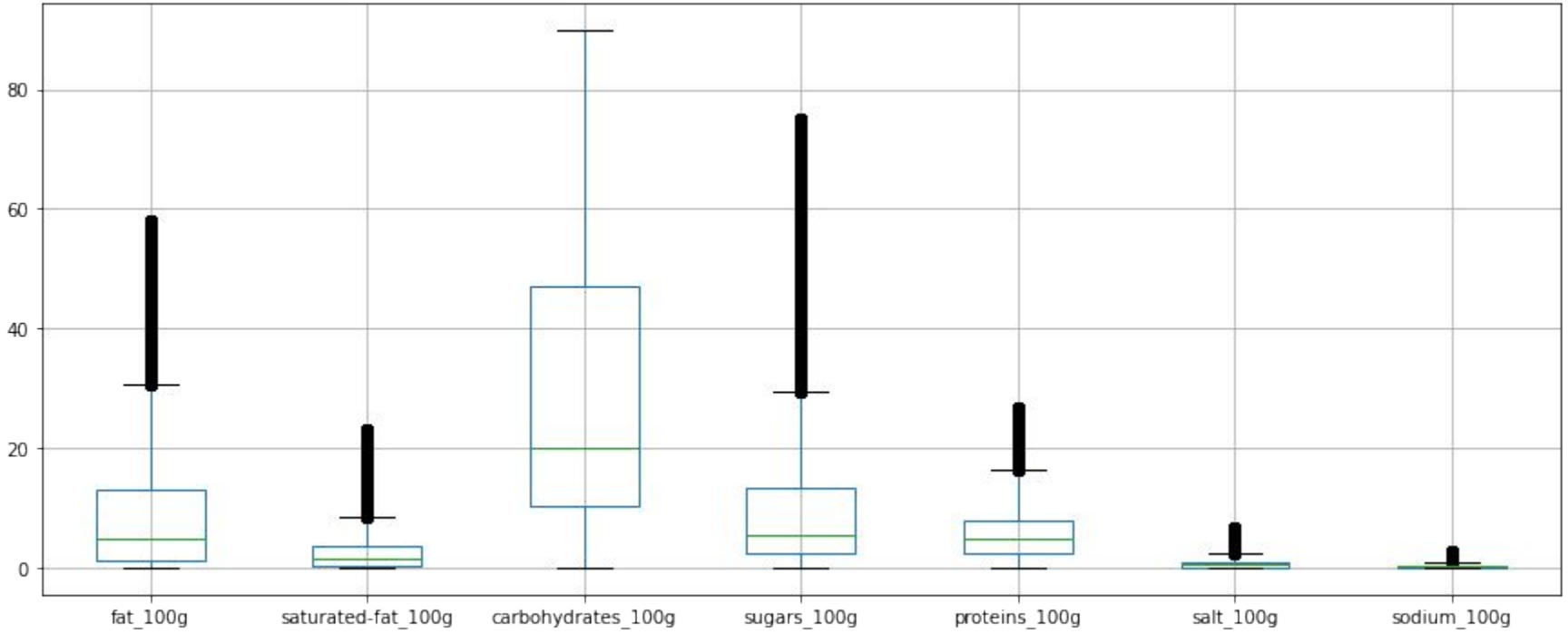


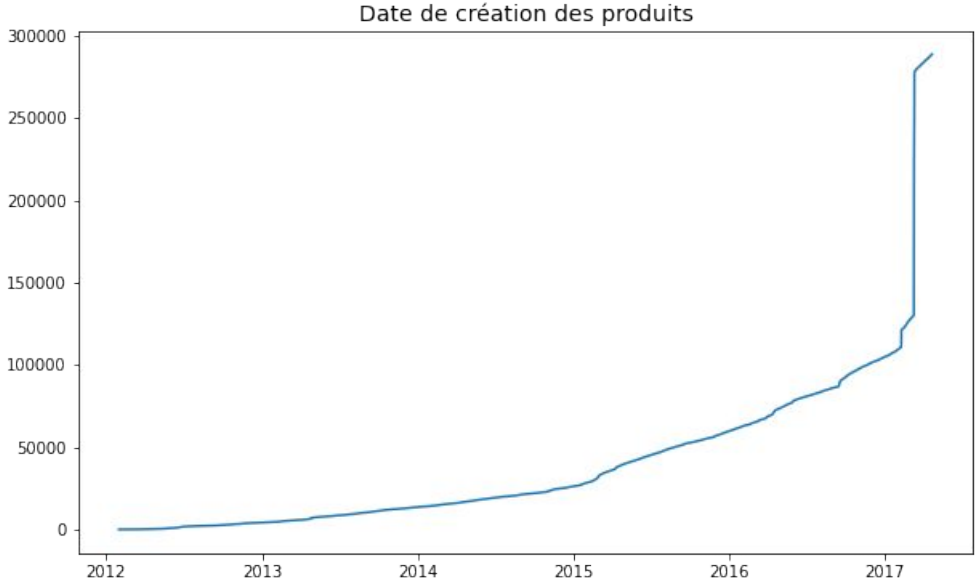
- 3 additifs sont présents dans au moins 15 % des produits : l’eau, la farine et le sucre.

Boxplot : **Valeur énergétique** par 100g (en kJ)



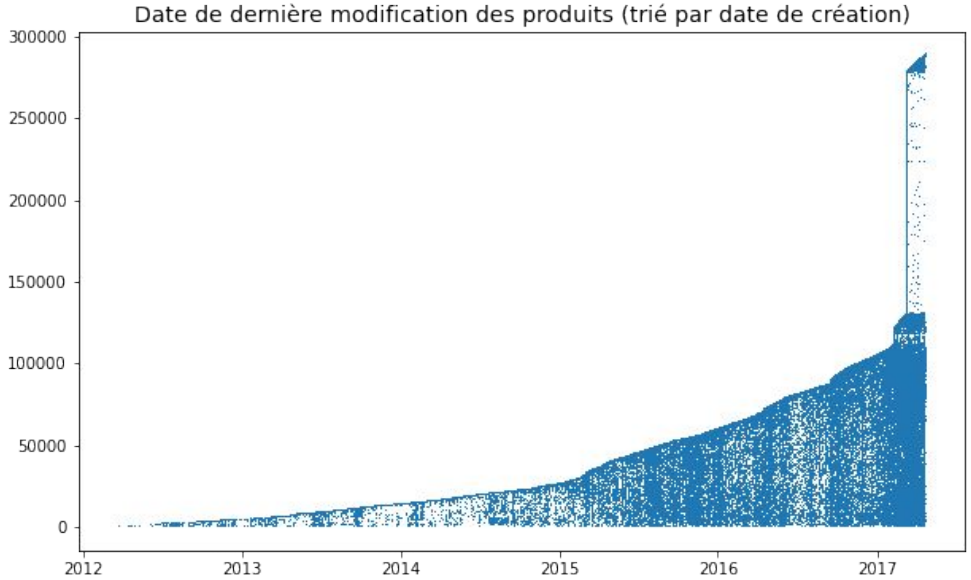
Boxplot : **composants** par 100g (en g)





Date de création des produits sur la plateforme

- Augmentation soudaine du nombre de produits en 2017
- Avant 2017, les produits sont créés de manière stable dans le temps

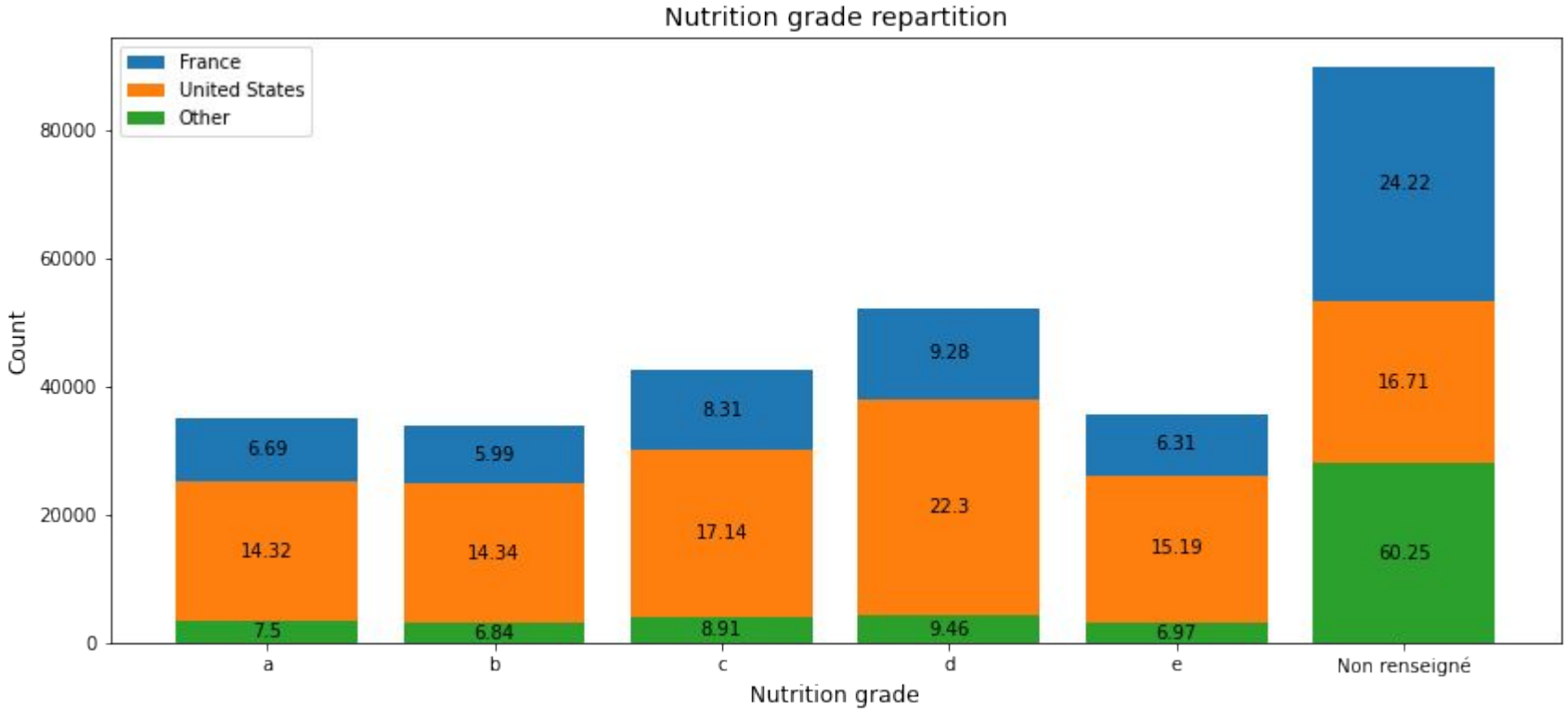


Dates de **dernière modification** des produits

- On remarque que même des produits créés il y a longtemps continuent d'être modifiés.

3. Analyses bivariées

Répartition de **nutrition grade**



Matrice de corrélations

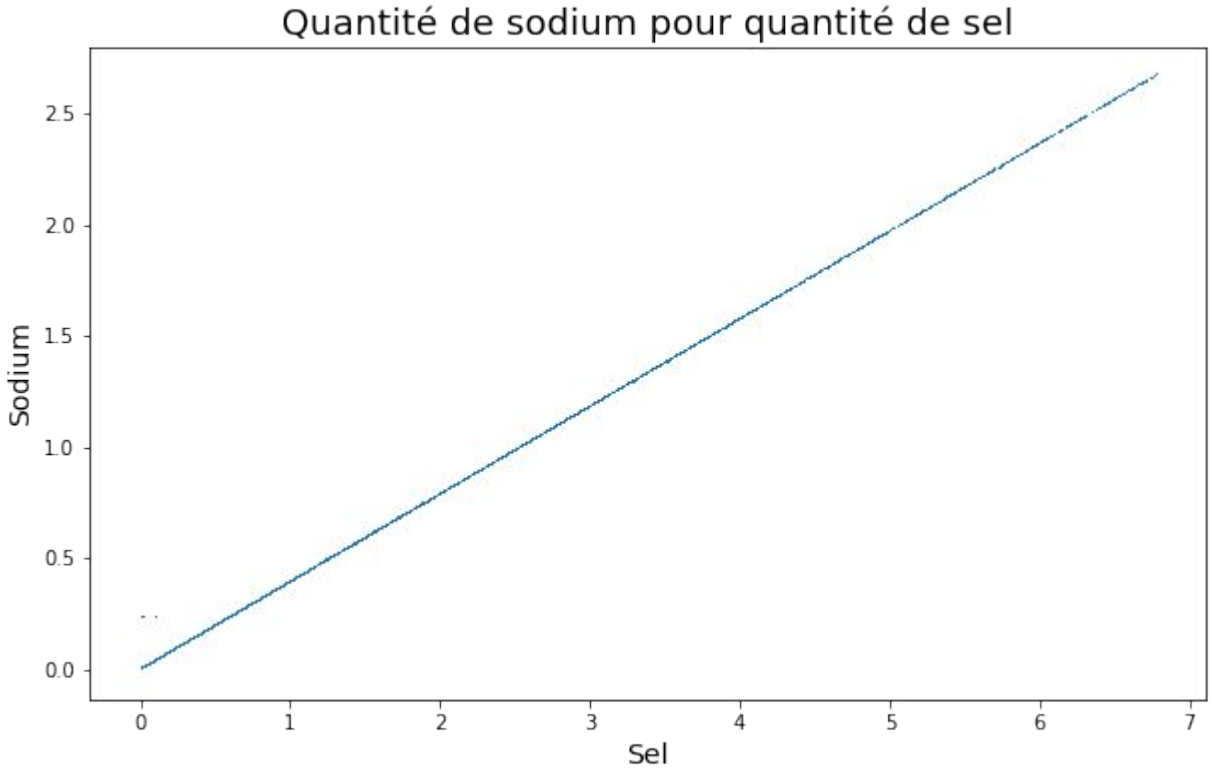
	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	proteins_100g	salt_100g	sodium_100g
energy_100g	1.000000	0.689586	0.536699	0.678074	0.284490	0.524271	0.215135	0.214941
fat_100g	0.689586	1.000000	0.764492	0.216467	0.044966	0.506197	0.377444	0.377258
saturated-fat_100g	0.536699	0.764492	1.000000	0.098153	0.128336	0.331357	0.233353	0.233203
carbohydrates_100g	0.678074	0.216467	0.098153	1.000000	0.505356	0.169942	-0.042687	-0.043012
sugars_100g	0.284490	0.044966	0.128336	0.505356	1.000000	-0.248996	-0.274733	-0.275101
proteins_100g	0.524271	0.506197	0.331357	0.169942	-0.248996	1.000000	0.358487	0.358353
salt_100g	0.215135	0.377444	0.233353	-0.042687	-0.274733	0.358487	1.000000	0.999730
sodium_100g	0.214941	0.377258	0.233203	-0.043012	-0.275101	0.358353	0.999730	1.000000

- Sel et sodium très corrélés (1)
- Autres variables relativement corrélées :
  - Énergie et gras (0.69)
  - Énergie et glucides (0.68)
  - Gras et gras saturé (0.76)

Heatmap des corrélations

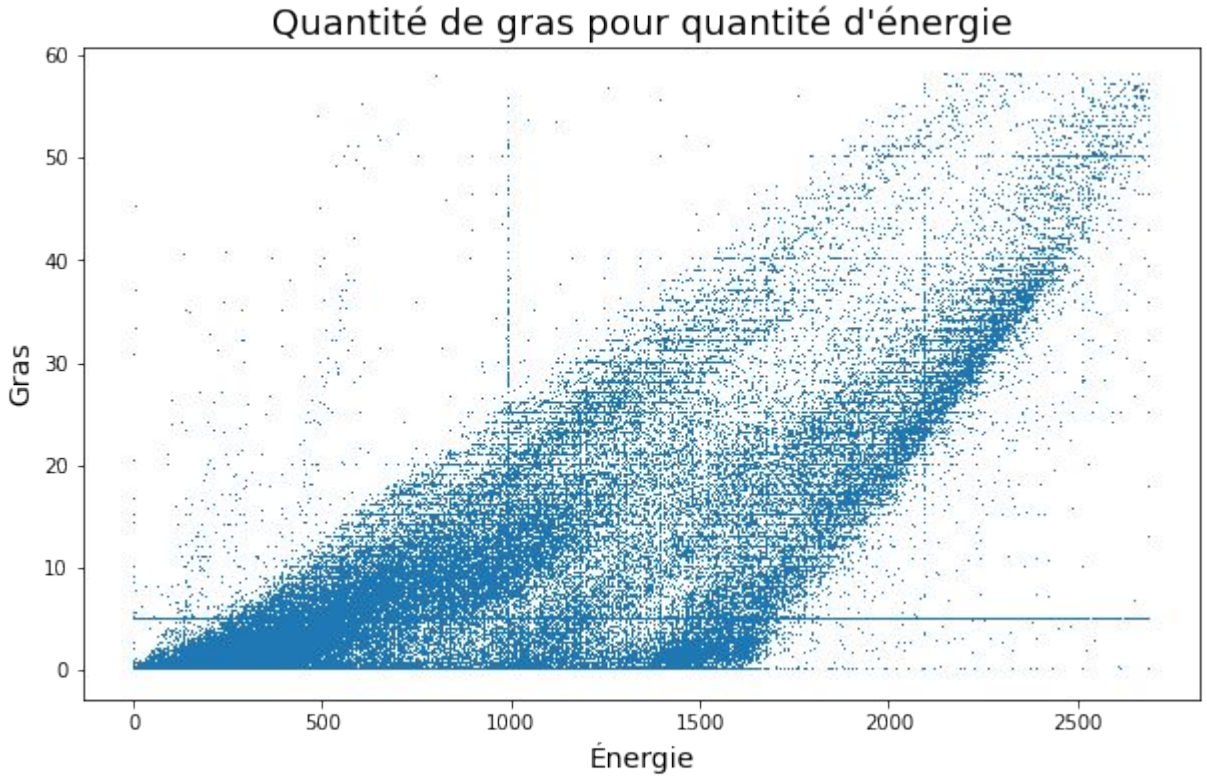
- Plus la couleur est chaude, plus les variables sont corrélées entres elles
- On retrouve logiquement le sel et le sodium en beige

Relation entre **sel** et **sodium** dans les produits : **1**



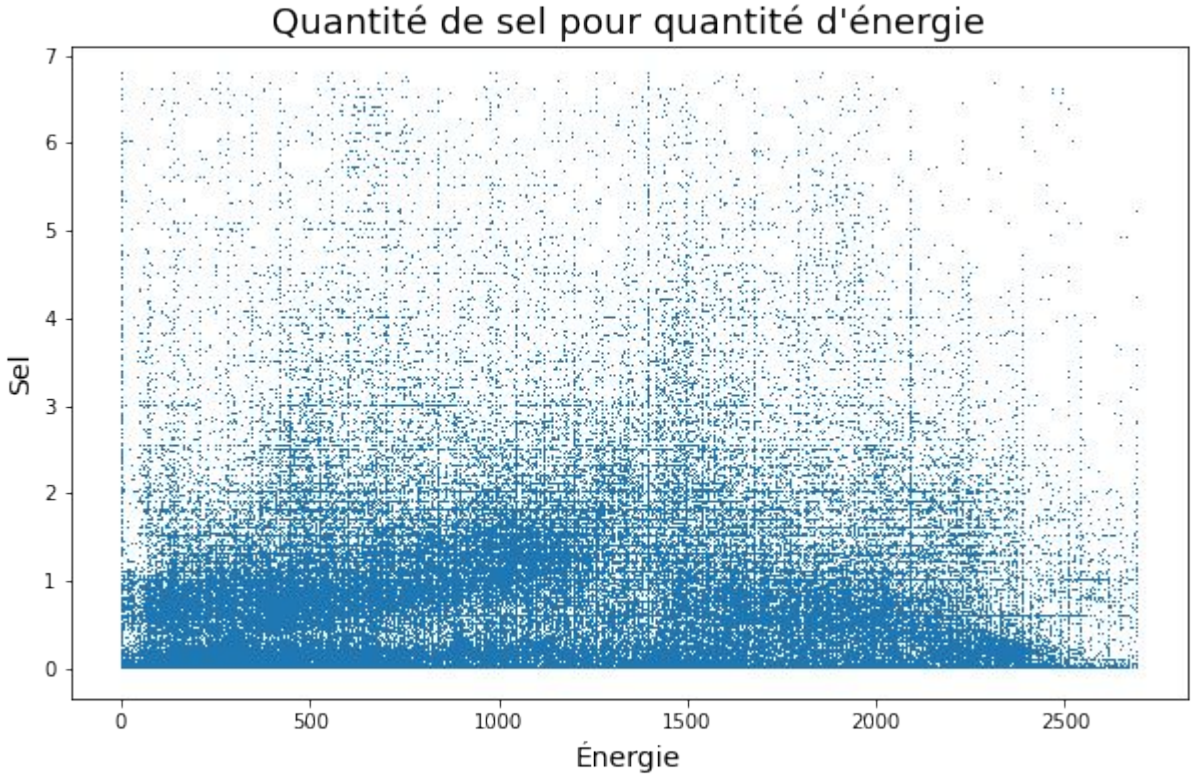
- Ce graph nous montre visuellement que les données sont pertinentes : corrélation de 1 = ligne droite.

Relation entre **énergie** et **gras** dans les produits : **0.69**



- Avec une corrélation de 0.69, on remarque que les données sont corrélées mais moins qu'avec une valeur de 1

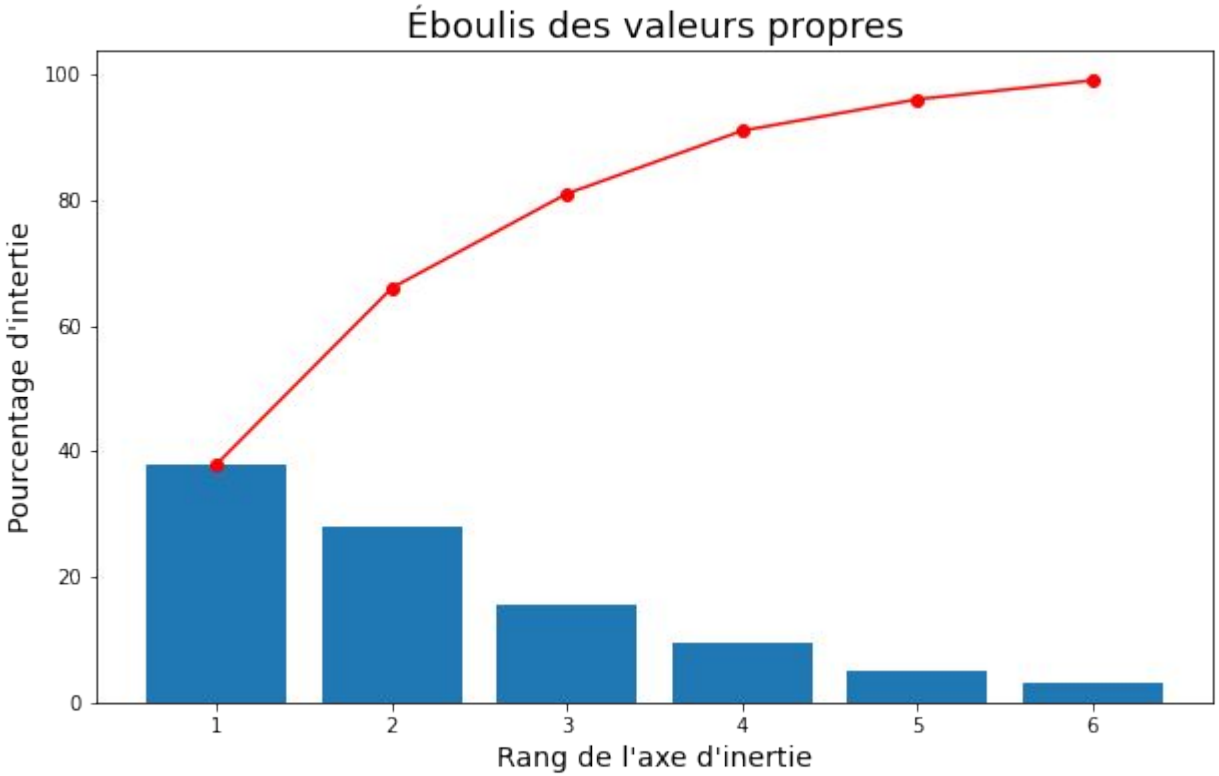
Relation entre **énergie** et **sel** dans les produits : **0.22**



- Avec une valeur de 0.22, il devient difficile de retrouver une relation entre les deux variables.

4. Analyses multivariées

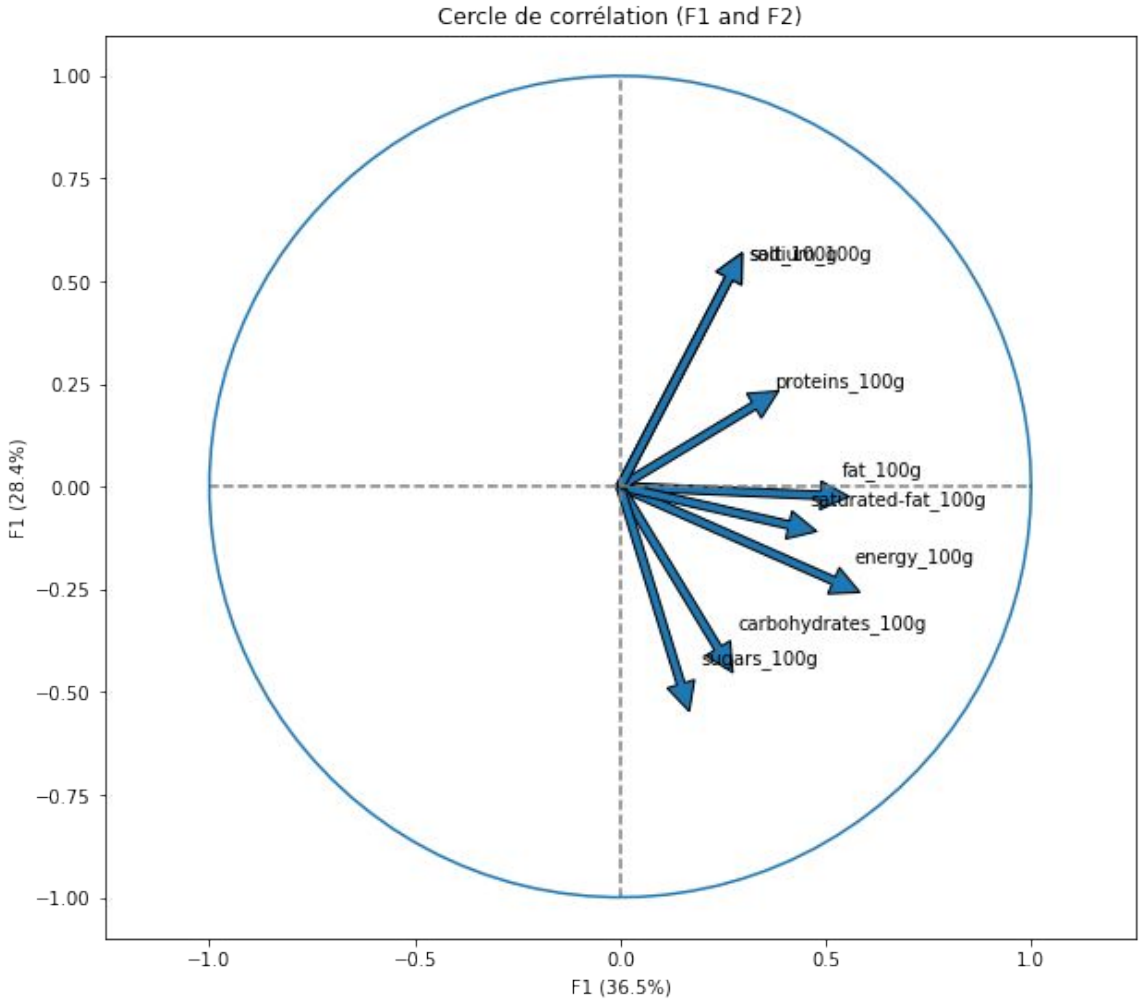
ACP : Éboulis des valeurs propres



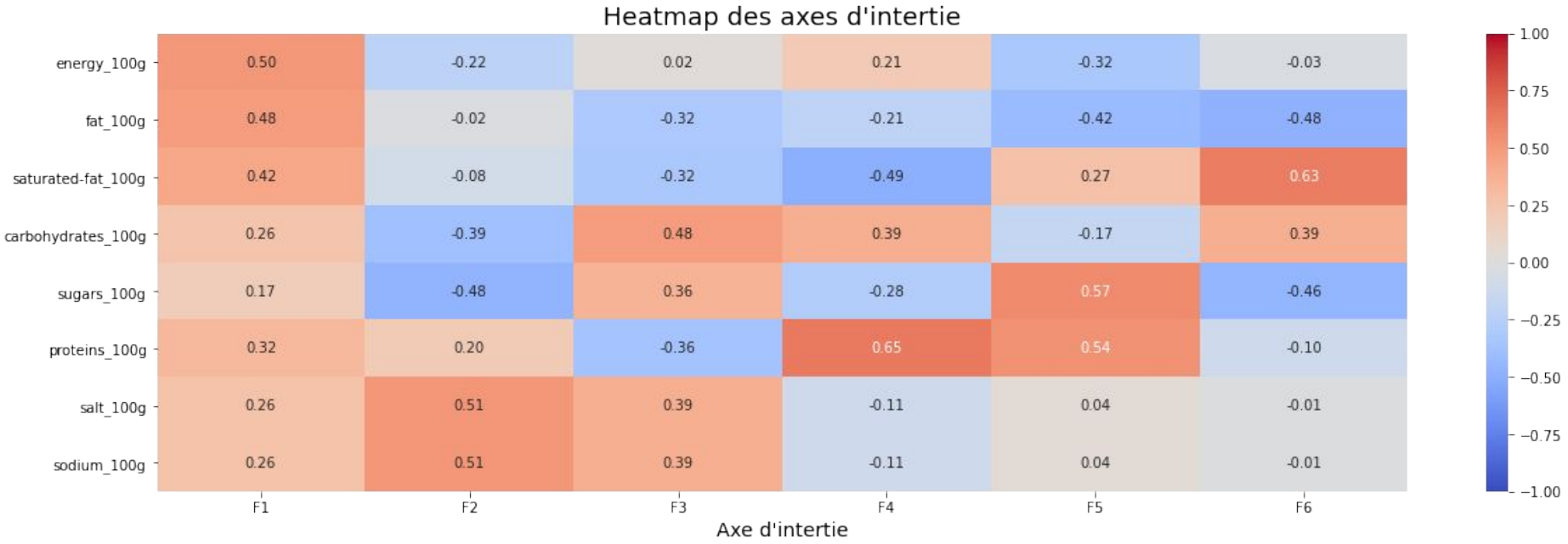
- Les axes F1 et F2 représentent **66%** des données,
- F1 à F4 : **91%**,
- F1 à F6 : **99%** des données.

ACP : Cercle de corrélation - F1 & F2

- Salt et sodium positivement corrélés à F2.
- Énergie, gras et gras saturé positivement corrélés à F1.
- Sucres et glucides négativement corrélés à F2.



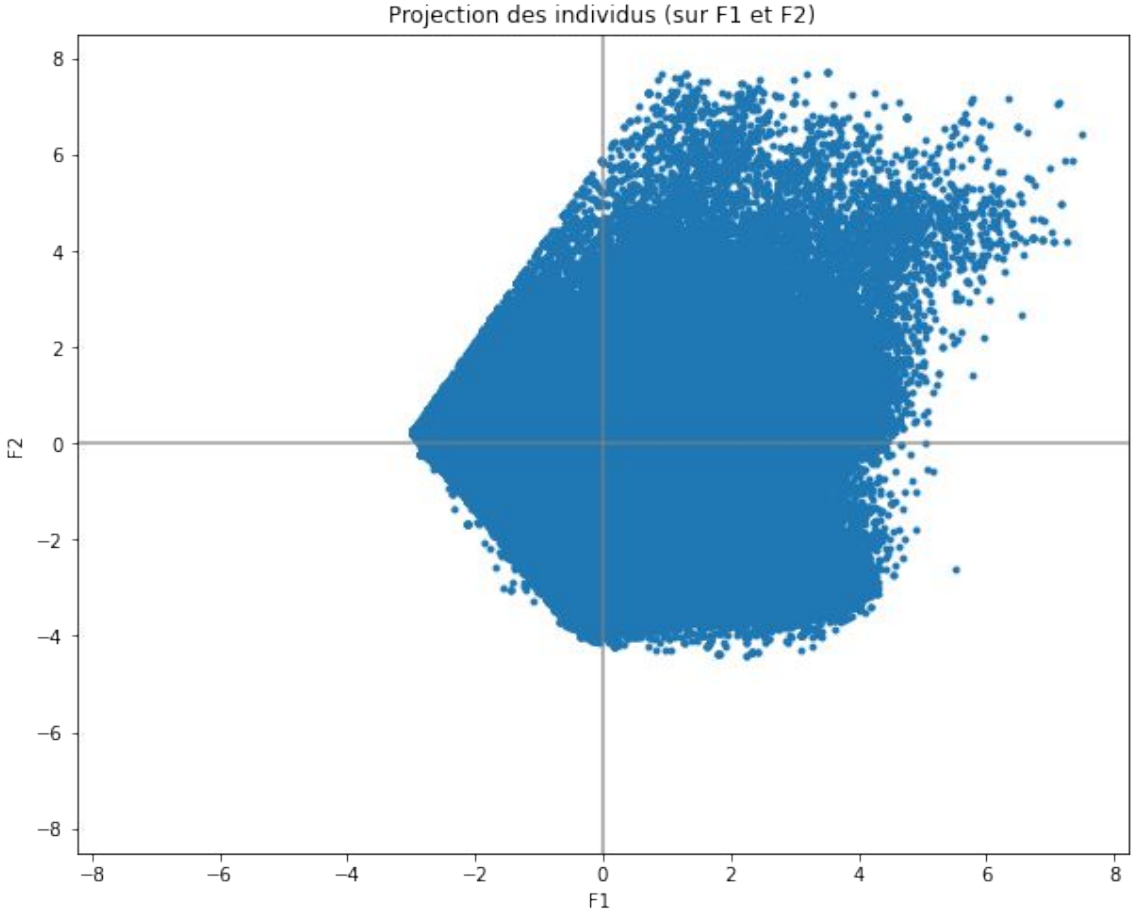
ACP : Heatmap des axes d’intertie



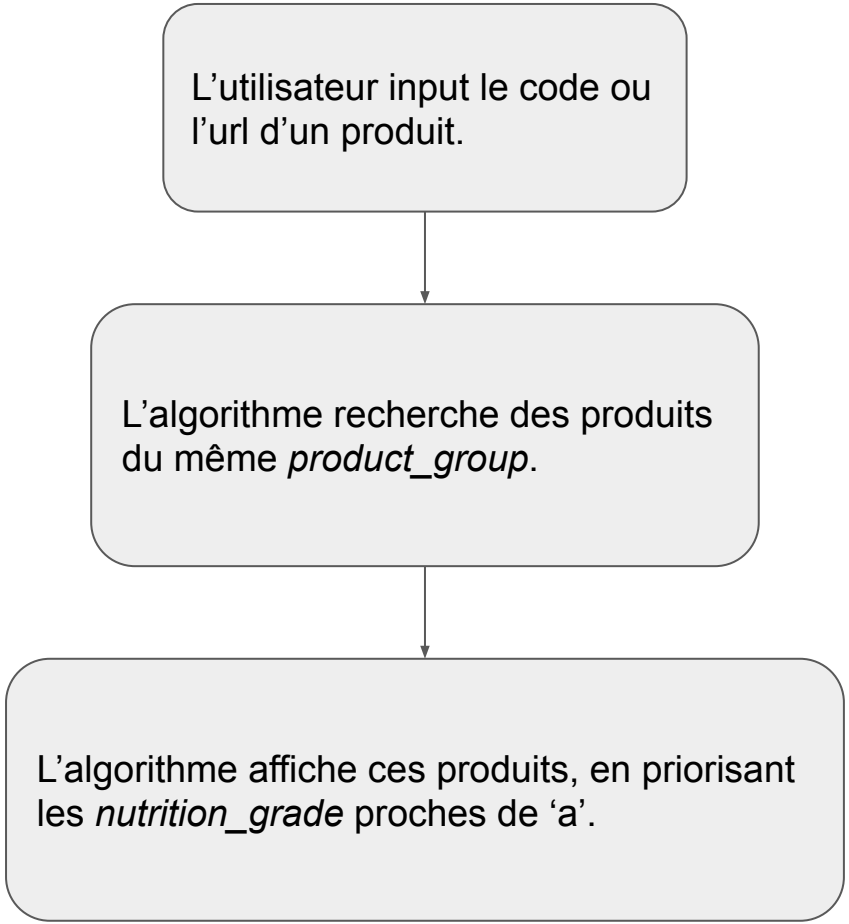
ACP : Aperçu : projection des individus sur F1 & F2

- Les individus en haut à droite du graph représentent les produits forts en sel/sodium

- En bas, les individus contenant peu de sucre



5. Idée d’application



L’utilisateur peut aussi trier la liste des produits par n’importe quel additif, comme ‘sel’ ou ‘gras’, en fonction de son régime nutritionnel préféré.

Example d'application

L'utilisateur recherche un produit

```
In [121]: test_app('99410148')
```

	code	creator	product_name	nutrition_grade_fr	energy_100g	fat_100g	brand_1
288935	99410148	date-limite-app	nuts	not specified	1031.871034	10.771763	not specified

```
Produit :  nuts
Groupe du produit :  Salty snacks

Produits similaires :
```

	code	creator	product_name	nutrition_grade_fr	energy_100g	fat_100g	brand_1
241155	5000128635325	jm0804	appetizers	a	1699.0	12.0	co-op
215071	3564700263822	philae	nuts	a	497.0	1.5	notre-jardin
185820	3256221129182	cestki13	nuts	a	433.0	1.4	u
224802	3760032460087	sebleouf	nuts	a	953.0	1.0	inovfruit
224803	3760032460117	simonm	nuts	a	953.0	1.0	inovfruit

Des alternatives avec un meilleur *nutrition_grade* sont proposés à l'utilisateur

Axes d’amélioration

- Réaliser davantage de feature engineering :
  - Régulariser les noms de produits, marques, brands etc (enlever les mots communs, les virgules, etc.)
  - Trier par groupes de produits pour fill les données manquantes des valeurs quantitatives
  - Fusionner les colonnes similaires contenant de nombreuses valeurs manquantes en une colonne remplie (ex: groupes et noms des produits)