Préparation de données pour Santé publique France

Santé publique France

Sommaire

- Contexte
- Présentation générale du jeu de données
- Nettoyage du jeu de données
- Analyse des données / Prototype
- Synthèse

Contexte

Santé publique France

OpenFoodFact

Accessibilité des données de santé

Présentation générale du jeu de données

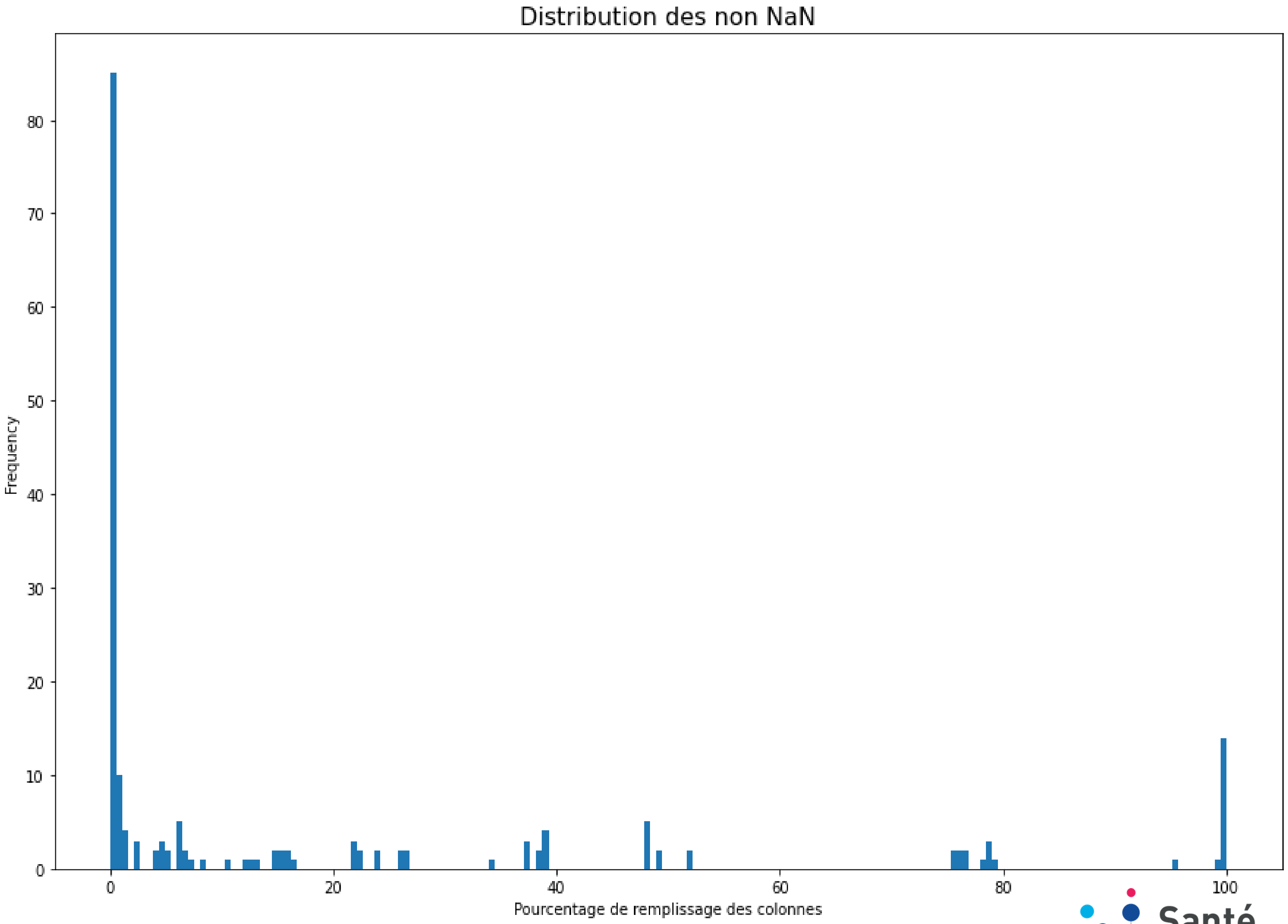
OPEN FOOD FACTS
l'information alimentaire ouverte

Santé publique France

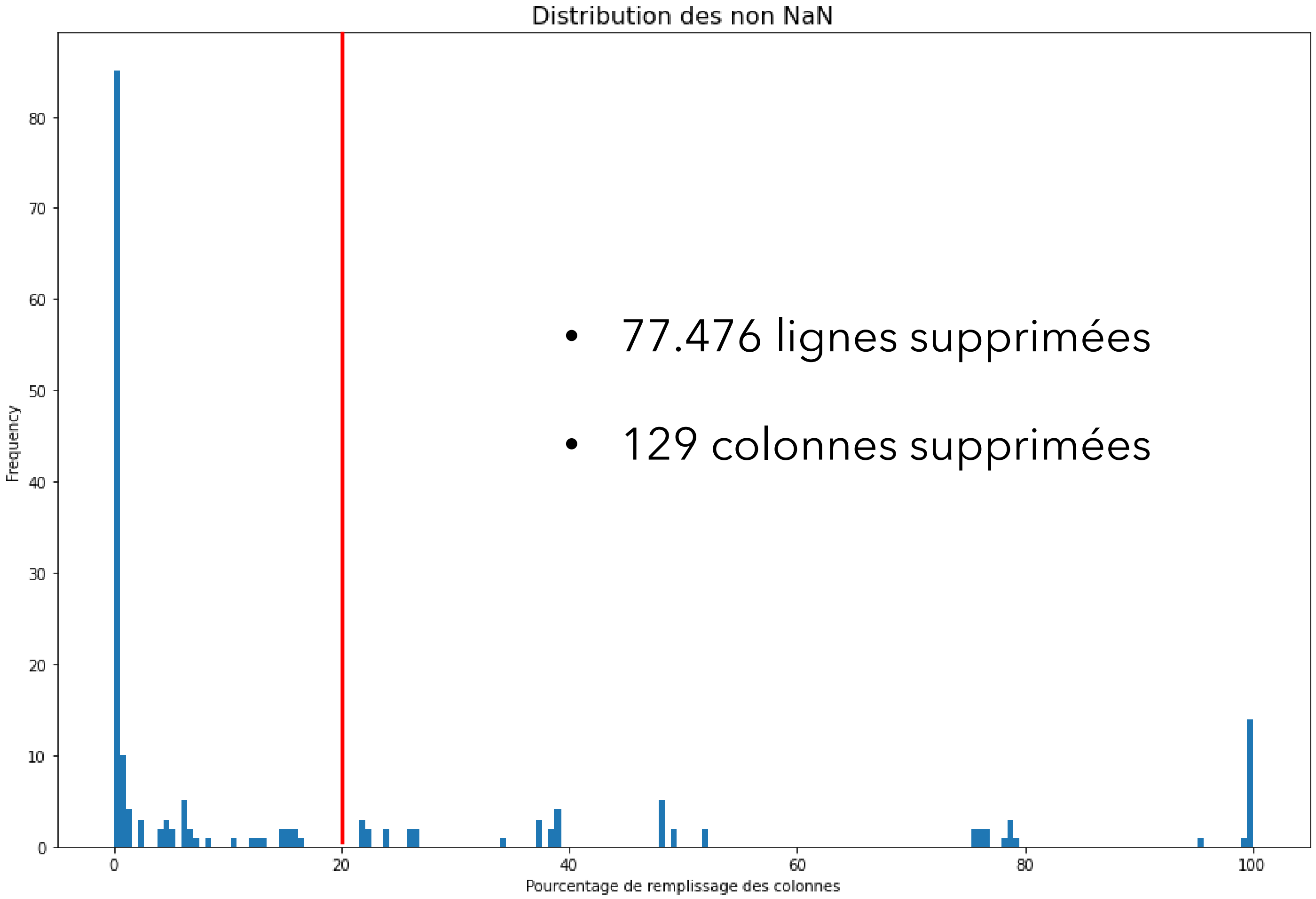
Dimensionnalités

1.770.670 lignes

186 colonnes



Nettoyage simple



- 77.476 lignes supprimées
- 129 colonnes supprimées

Colonnes vides (-20%)

Lignes vides

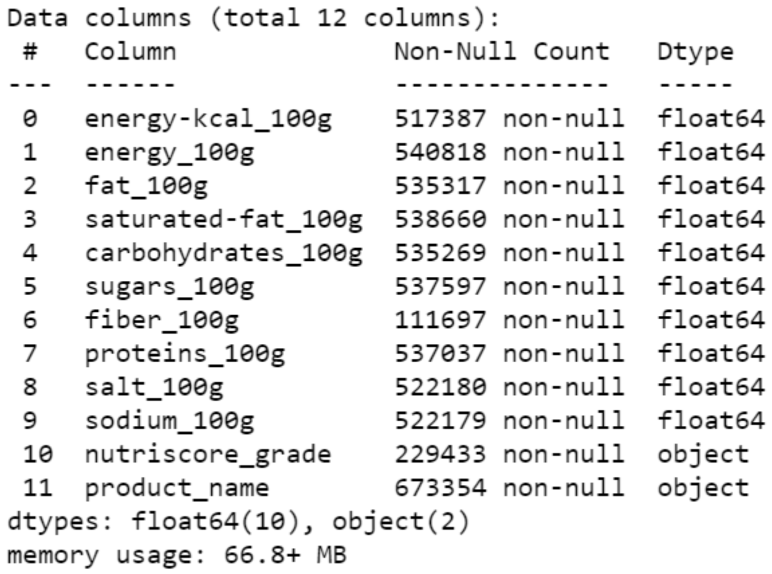
Lignes dupliquées

Colonne produit vide

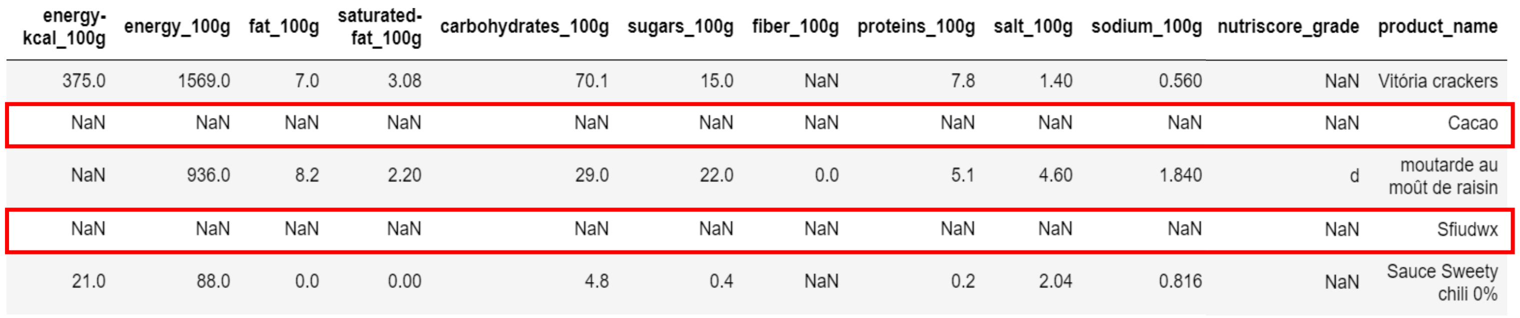
Sélection du jeu de données d’étude

673.354 lignes

12 colonnes

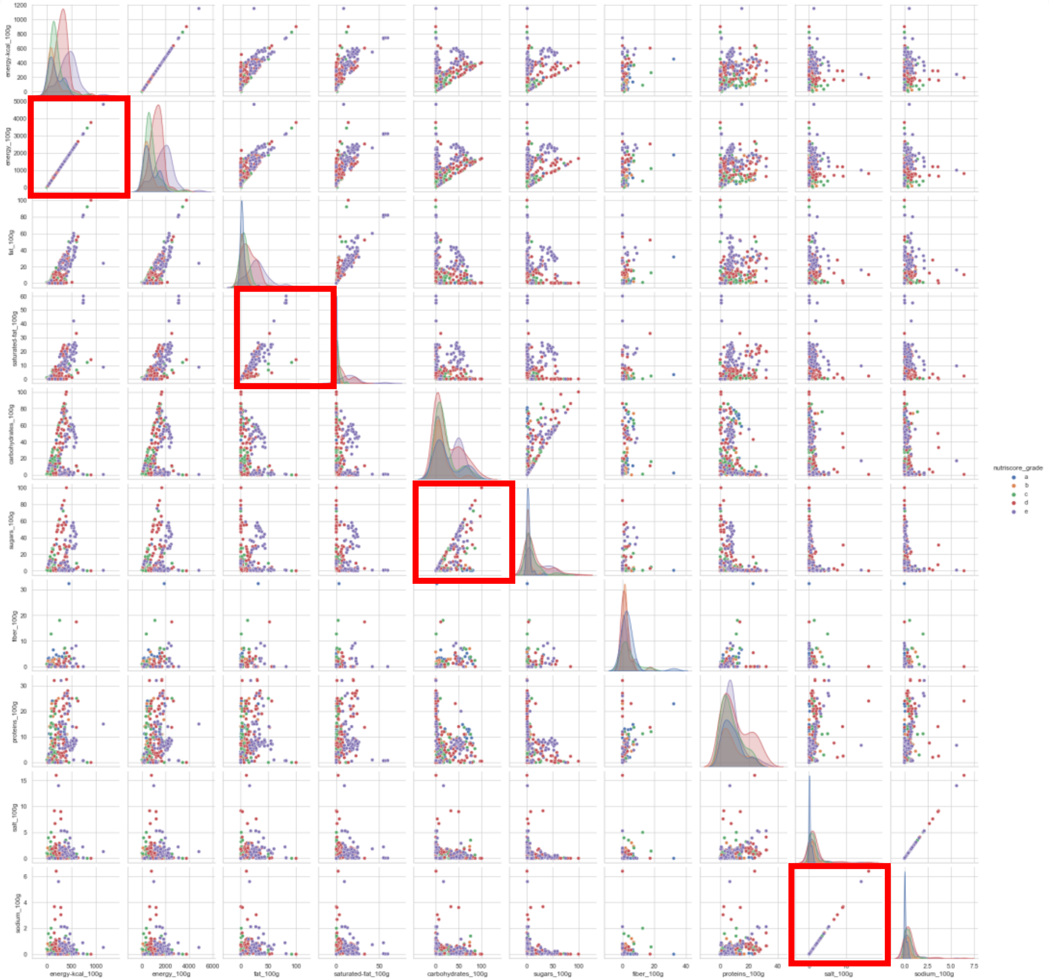


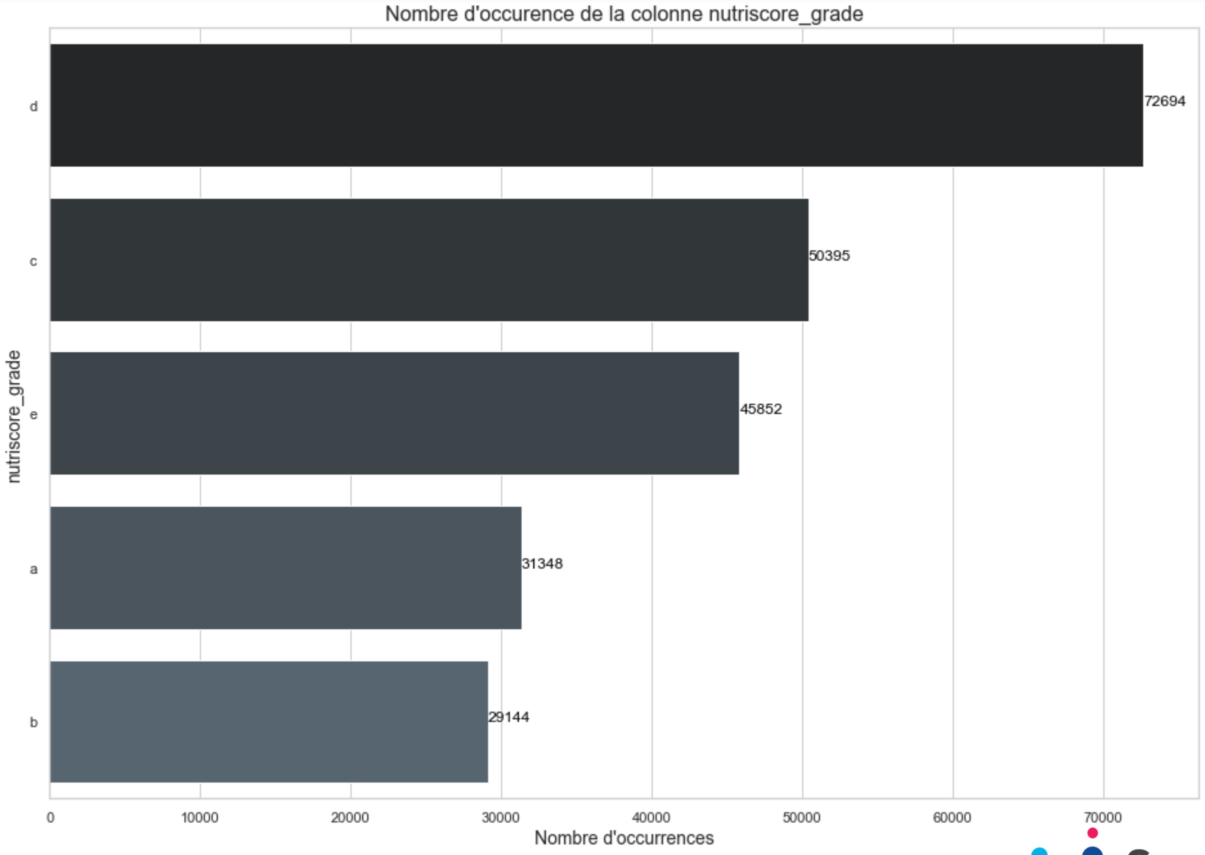
```
Data columns (total 12 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   energy-kcal_100g    517387 non-null  float64
 1   energy_100g         540818 non-null  float64
 2   fat_100g            535317 non-null  float64
 3   saturated-fat_100g  538660 non-null  float64
 4   carbohydrates_100g  535269 non-null  float64
 5   sugars_100g         537597 non-null  float64
 6   fiber_100g          111697 non-null  float64
 7   proteins_100g       537037 non-null  float64
 8   salt_100g           522180 non-null  float64
 9   sodium_100g         522179 non-null  float64
 10  nutriscore_grade    229433 non-null  object
 11  product_name        673354 non-null  object
dtypes: float64(10), object(2)
memory usage: 66.8+ MB
```



energy-kcal_100g	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutriscore_grade	product_name
375.0	1569.0	7.0	3.08	70.1	15.0	NaN	7.8	1.40	0.560	NaN	Vitória crackers
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Cacao
NaN	936.0	8.2	2.20	29.0	22.0	0.0	5.1	4.60	1.840	d	moutarde au moût de raisin
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Sfiudwx
21.0	88.0	0.0	0.00	4.8	0.4	NaN	0.2	2.04	0.816	NaN	Sauce Sweety chili 0%

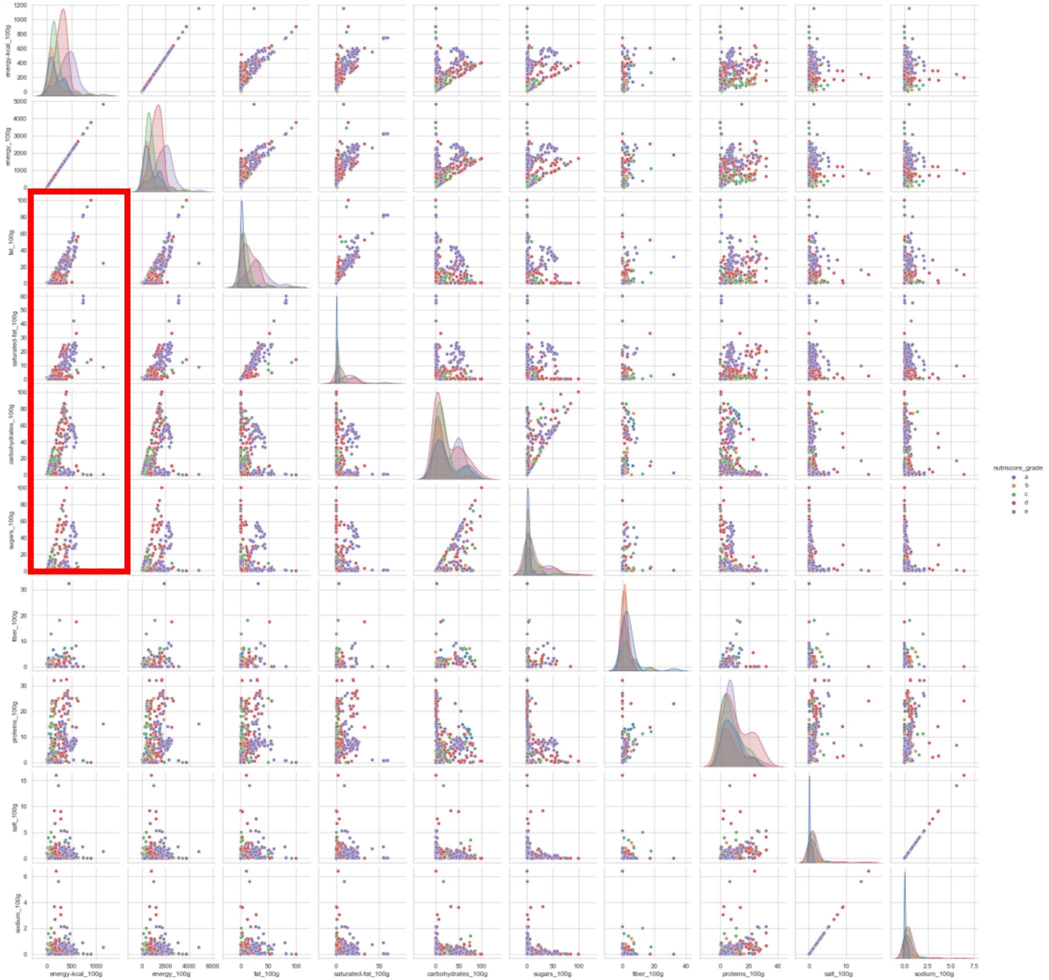
Analyse du jeu de données

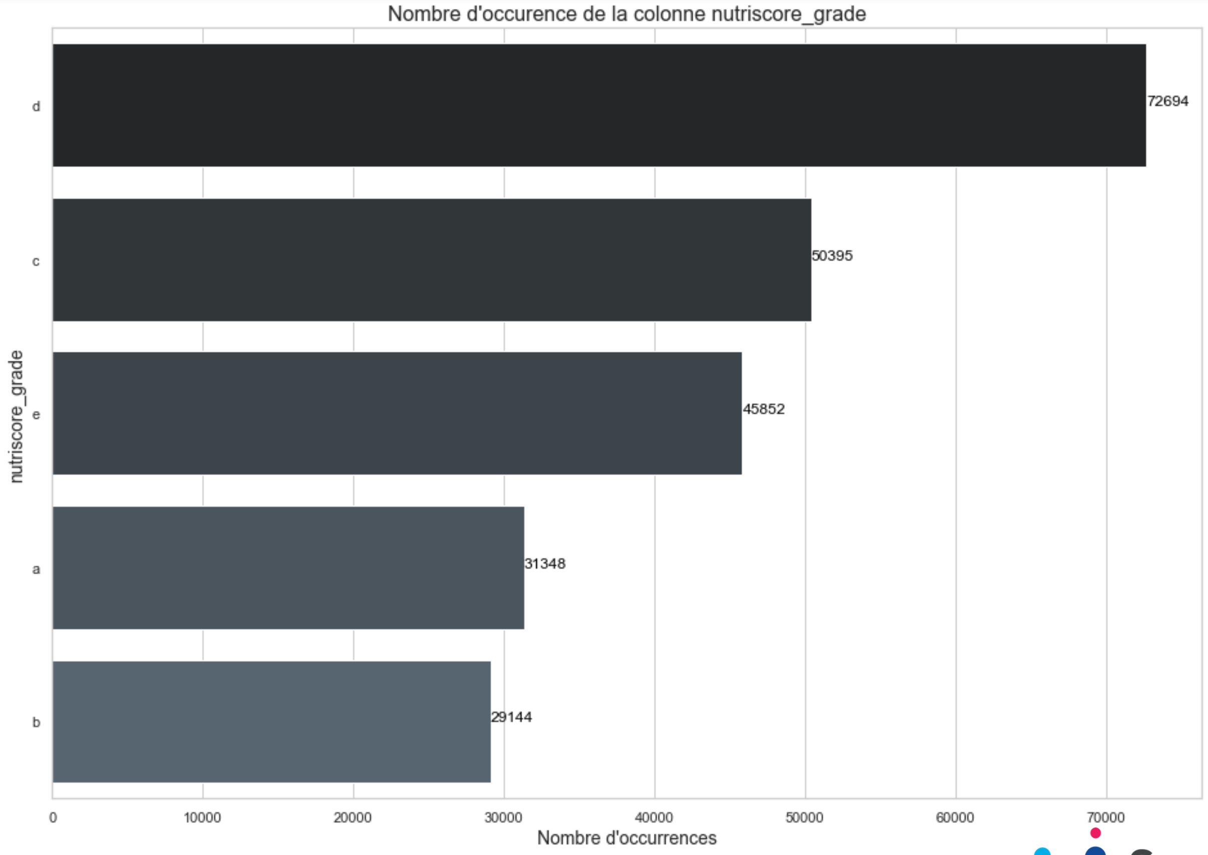




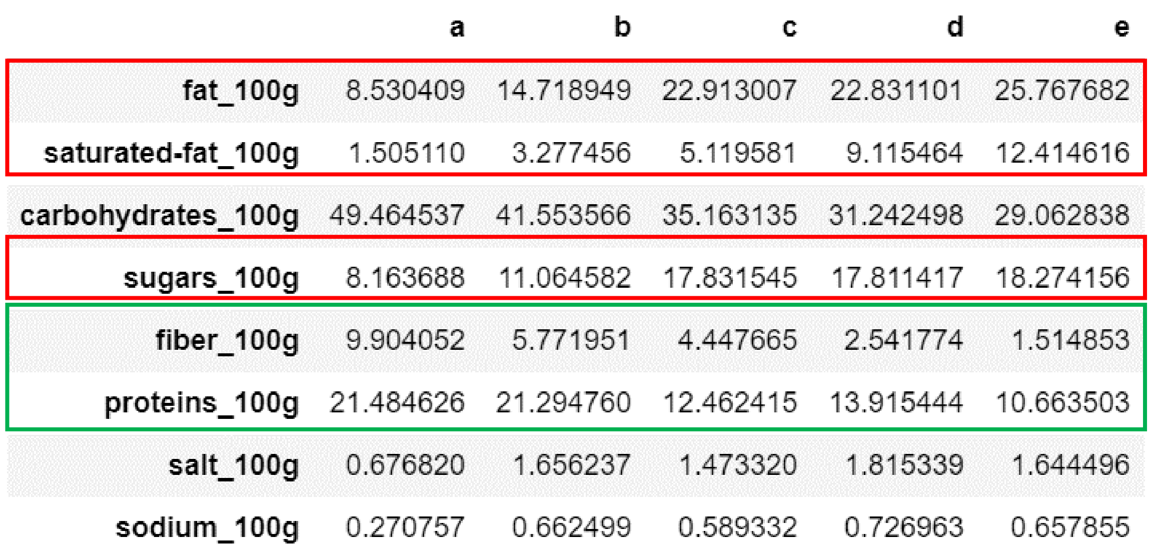
Santé
publique
France

Analyse du jeu de données (suite)





Analyse du jeu de données (suite)



	a	b	c	d	e
fat_100g	8.530409	14.718949	22.913007	22.831101	25.767682
saturated-fat_100g	1.505110	3.277456	5.119581	9.115464	12.414616
carbohydrates_100g	49.464537	41.553566	35.163135	31.242498	29.062838
sugars_100g	8.163688	11.064582	17.831545	17.811417	18.274156
fiber_100g	9.904052	5.771951	4.447665	2.541774	1.514853
proteins_100g	21.484626	21.294760	12.462415	13.915444	10.663503
salt_100g	0.676820	1.656237	1.473320	1.815339	1.644496
sodium_100g	0.270757	0.662499	0.589332	0.726963	0.657855

Hypothèse

Corrélation => gras/gras saturé et sucre/glucide

Augmentation nutriscore = augmentation gras/saturé, sucre/glucide

Inversement pour fibre et protéine

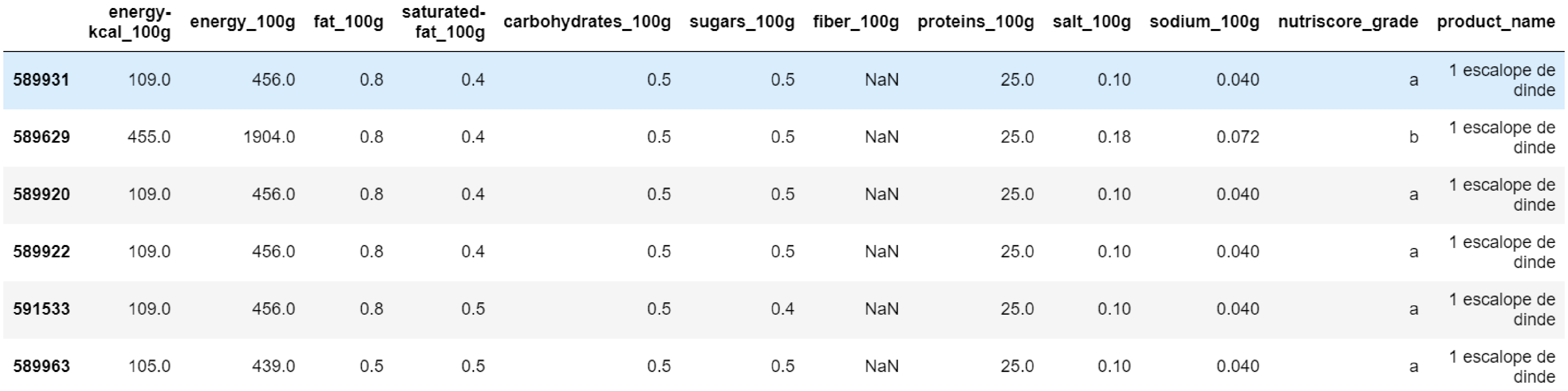
L'énergie = sucre + gras + gras saturé + glucide

Nettoyage du jeu de données

Doublons

- 308.176 lignes dupliquées

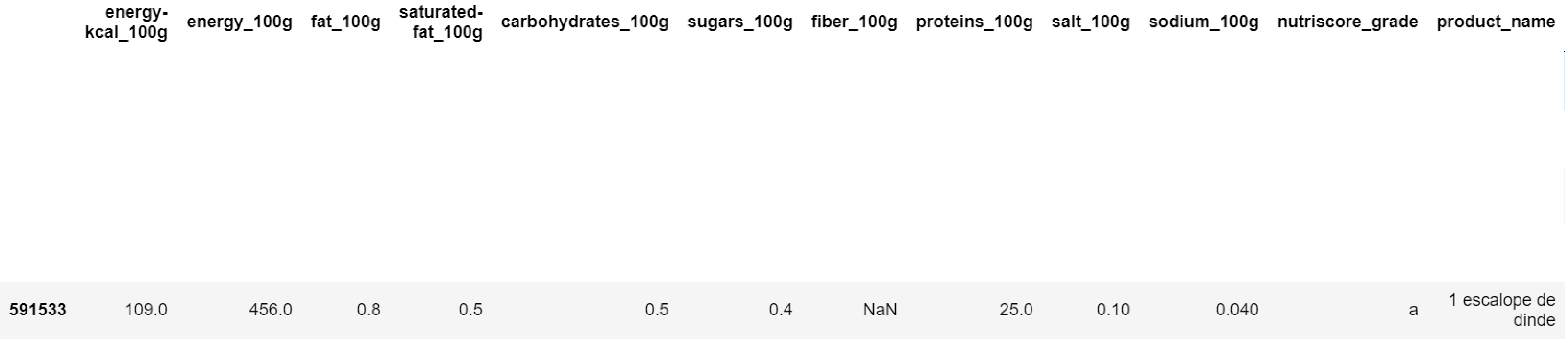
&quot; ⇔ «



	energy-kcal_100g	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutriscore_grade	product_name
589931	109.0	456.0	0.8	0.4	0.5	0.5	NaN	25.0	0.10	0.040	a	1 escalope de dinde
589629	455.0	1904.0	0.8	0.4	0.5	0.5	NaN	25.0	0.18	0.072	b	1 escalope de dinde
589920	109.0	456.0	0.8	0.4	0.5	0.5	NaN	25.0	0.10	0.040	a	1 escalope de dinde
589922	109.0	456.0	0.8	0.4	0.5	0.5	NaN	25.0	0.10	0.040	a	1 escalope de dinde
591533	109.0	456.0	0.8	0.5	0.5	0.4	NaN	25.0	0.10	0.040	a	1 escalope de dinde
589963	105.0	439.0	0.5	0.5	0.5	0.5	NaN	25.0	0.10	0.040	a	1 escalope de dinde

Doublons (suite)

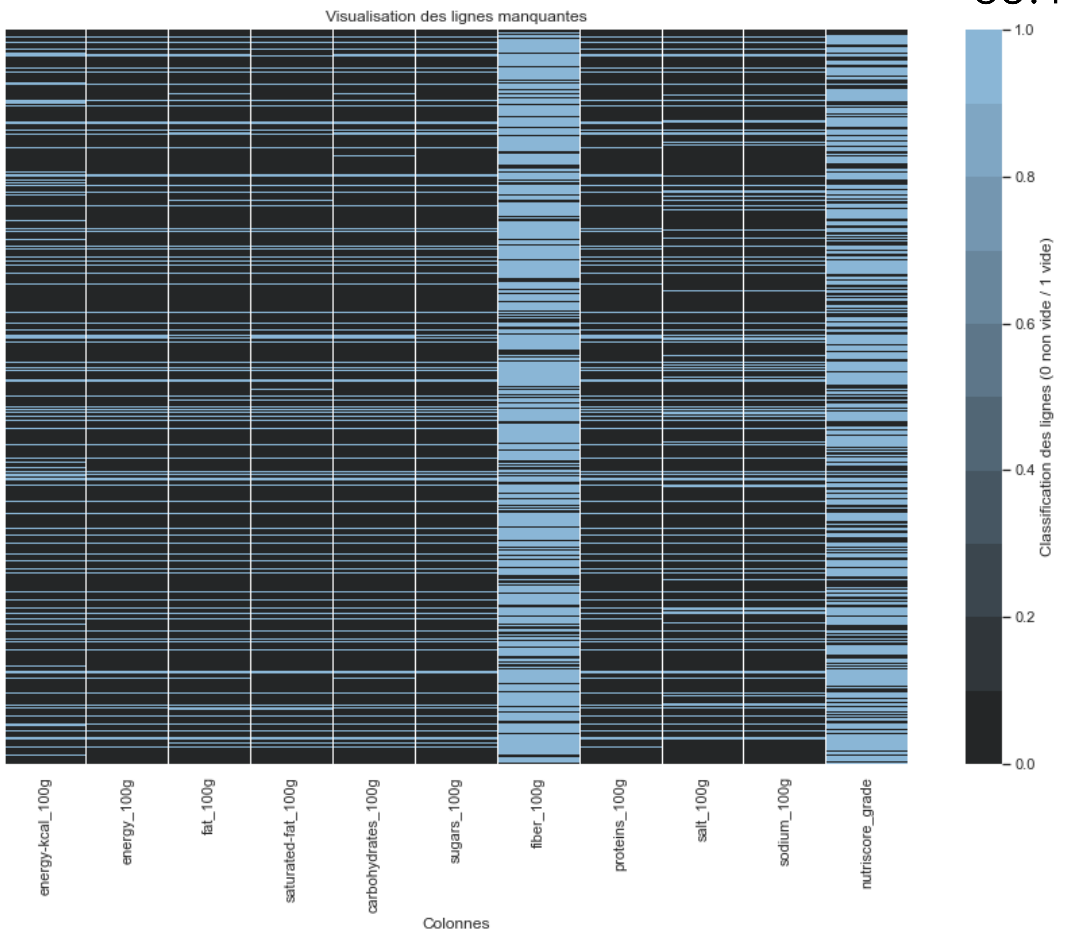
- 308.176 lignes dupliquées
- Uniquement dernière saisie

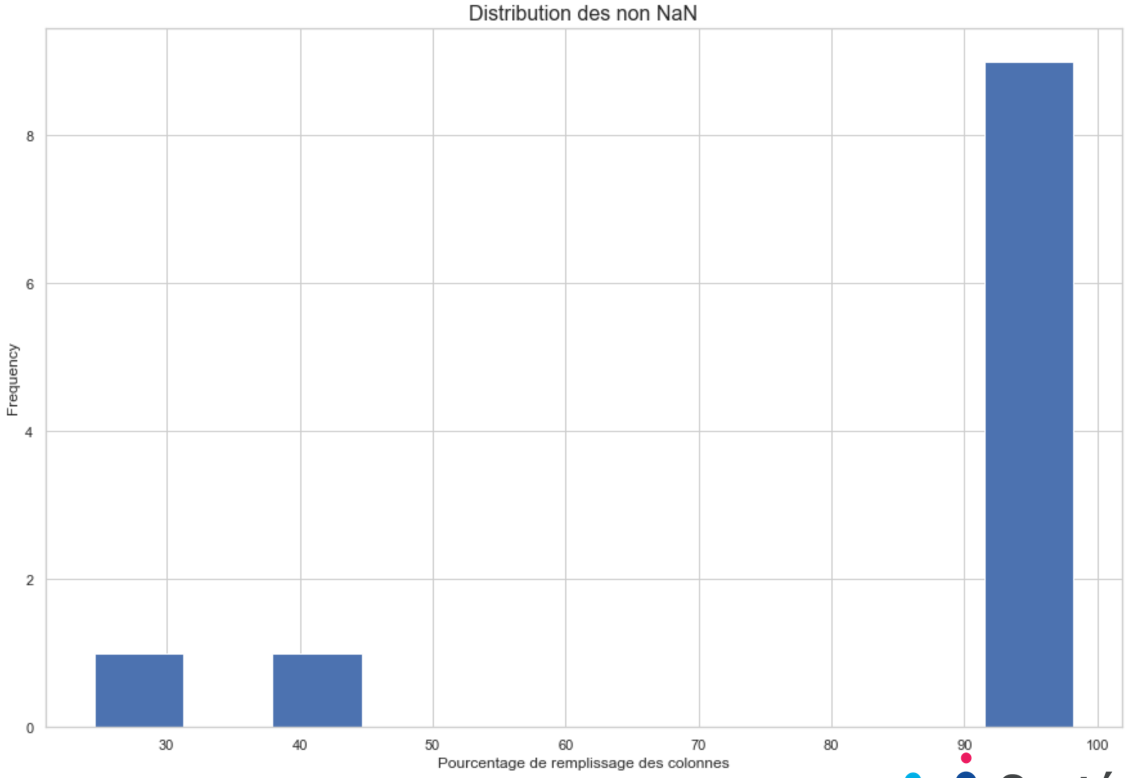


	energy-kcal_100g	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutriscore_grade	product_name
591533	109.0	456.0	0.8	0.5	0.5	0.4	NaN	25.0	0.10	0.040	a	1 escalope de dinde

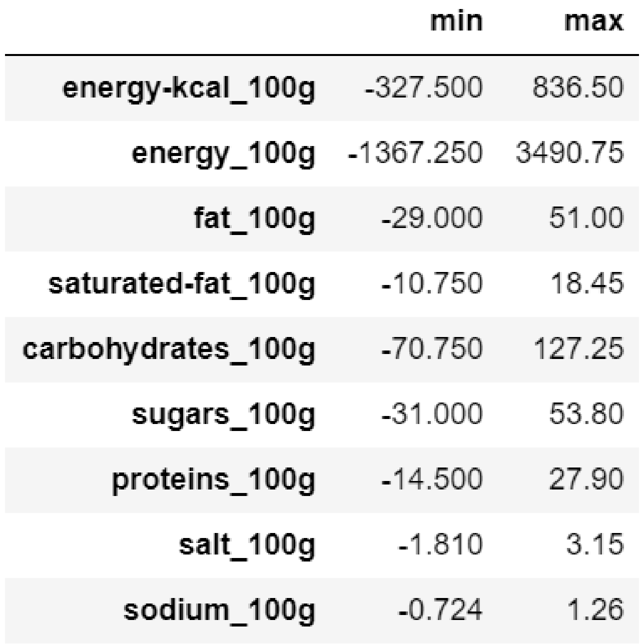
Valeurs manquantes

65.165 lignes



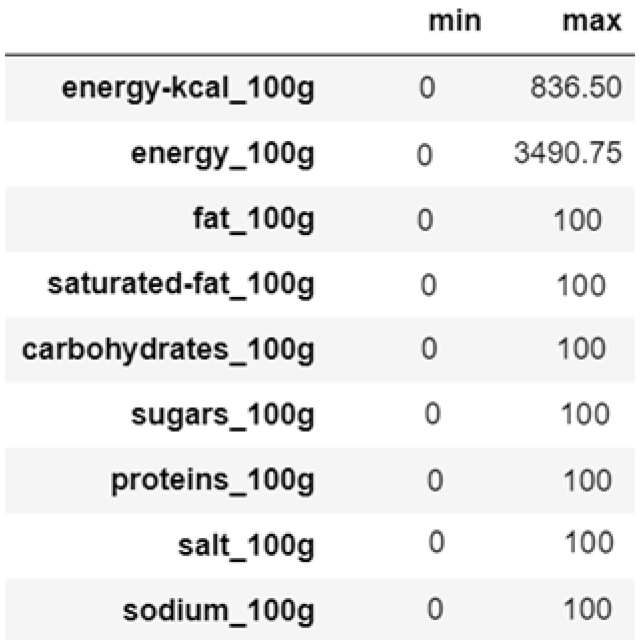


Aberrations



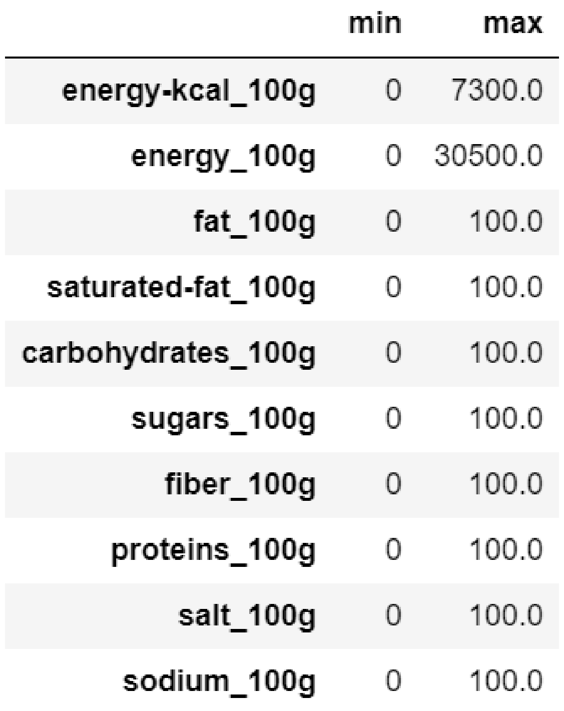
	min	max
energy-kcal_100g	-327.500	836.50
energy_100g	-1367.250	3490.75
fat_100g	-29.000	51.00
saturated-fat_100g	-10.750	18.45
carbohydrates_100g	-70.750	127.25
sugars_100g	-31.000	53.80
proteins_100g	-14.500	27.90
salt_100g	-1.810	3.15
sodium_100g	-0.724	1.26

Aberrations (suite)

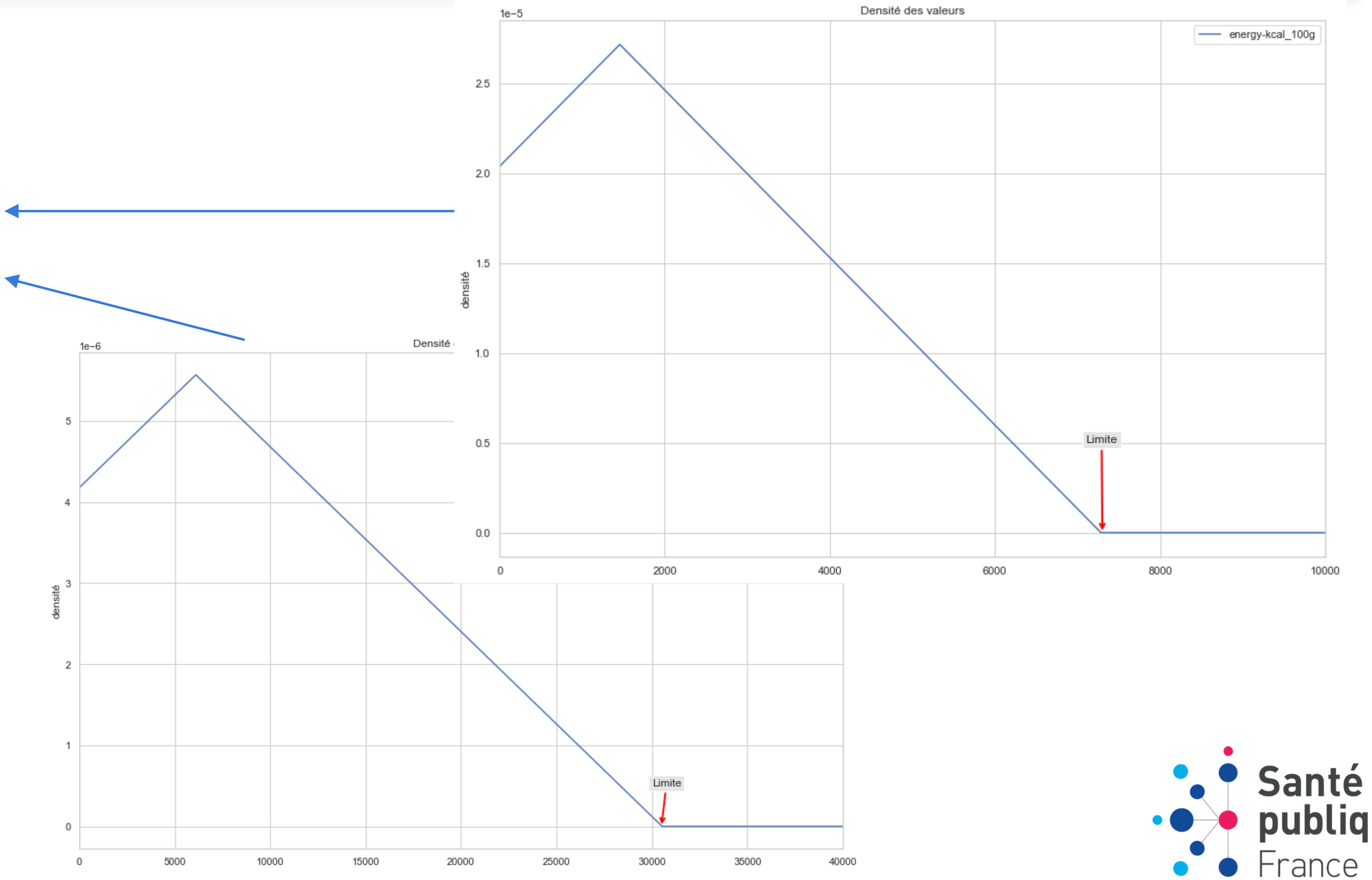


	min	max
energy-kcal_100g	0	836.50
energy_100g	0	3490.75
fat_100g	0	100
saturated-fat_100g	0	100
carbohydrates_100g	0	100
sugars_100g	0	100
proteins_100g	0	100
salt_100g	0	100
sodium_100g	0	100

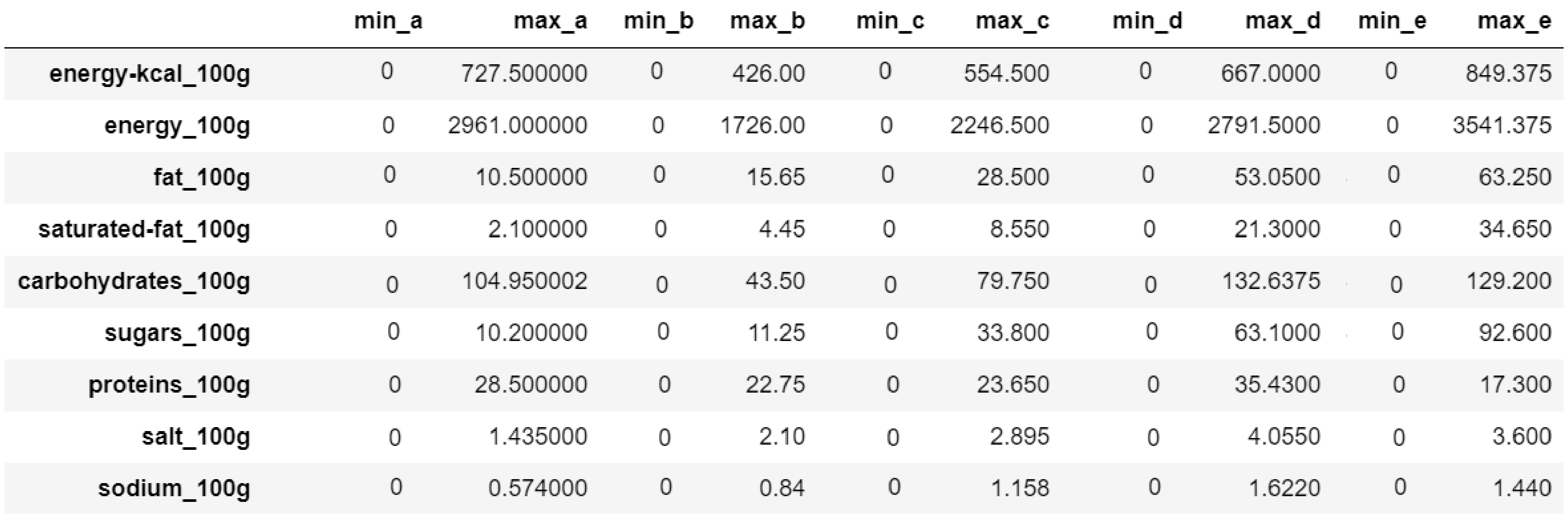
Aberrations (suite)



	min	max
energy-kcal_100g	0	7300.0
energy_100g	0	30500.0
fat_100g	0	100.0
saturated-fat_100g	0	100.0
carbohydrates_100g	0	100.0
sugars_100g	0	100.0
fiber_100g	0	100.0
proteins_100g	0	100.0
salt_100g	0	100.0
sodium_100g	0	100.0



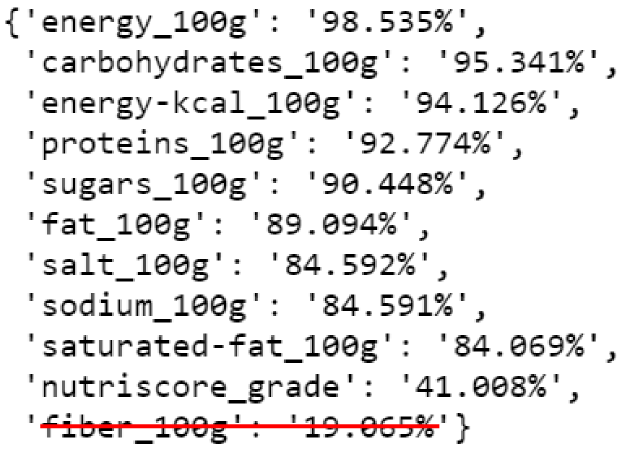
Aberrations (suite)



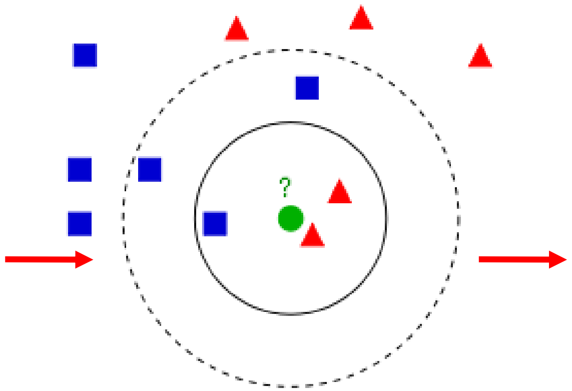
	min_a	max_a	min_b	max_b	min_c	max_c	min_d	max_d	min_e	max_e
energy-kcal_100g	0	727.500000	0	426.00	0	554.500	0	667.0000	0	849.375
energy_100g	0	2961.000000	0	1726.00	0	2246.500	0	2791.5000	0	3541.375
fat_100g	0	10.500000	0	15.65	0	28.500	0	53.0500	0	63.250
saturated-fat_100g	0	2.100000	0	4.45	0	8.550	0	21.3000	0	34.650
carbohydrates_100g	0	104.950002	0	43.50	0	79.750	0	132.6375	0	129.200
sugars_100g	0	10.200000	0	11.25	0	33.800	0	63.1000	0	92.600
proteins_100g	0	28.500000	0	22.75	0	23.650	0	35.4300	0	17.300
salt_100g	0	1.435000	0	2.10	0	2.895	0	4.0550	0	3.600
sodium_100g	0	0.574000	0	0.84	0	1.158	0	1.6220	0	1.440

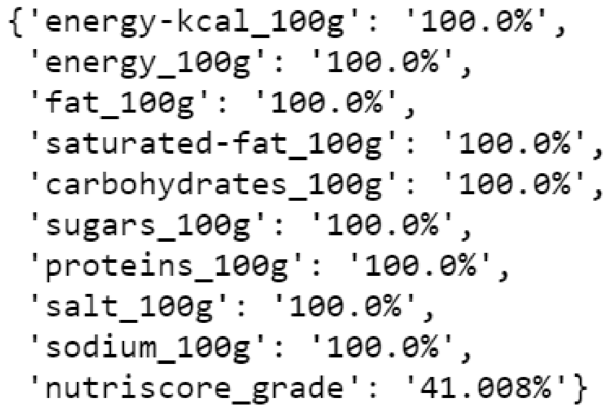
Imputation

- KNNImpute



```
{'energy_100g': '98.535%',
 'carbohydrates_100g': '95.341%',
 'energy-kcal_100g': '94.126%',
 'proteins_100g': '92.774%',
 'sugars_100g': '90.448%',
 'fat_100g': '89.094%',
 'salt_100g': '84.592%',
 'sodium_100g': '84.591%',
 'saturated-fat_100g': '84.069%',
 'nutriscore_grade': '41.008%',
 '~~fiber_100g': '19.065%~~'}
```





```
{'energy-kcal_100g': '100.0%',
 'energy_100g': '100.0%',
 'fat_100g': '100.0%',
 'saturated-fat_100g': '100.0%',
 'carbohydrates_100g': '100.0%',
 'sugars_100g': '100.0%',
 'proteins_100g': '100.0%',
 'salt_100g': '100.0%',
 'sodium_100g': '100.0%',
 'nutriscore_grade': '41.008%'}
```

Santé
publique
France

Analyse des données / Prototype

Synthèse

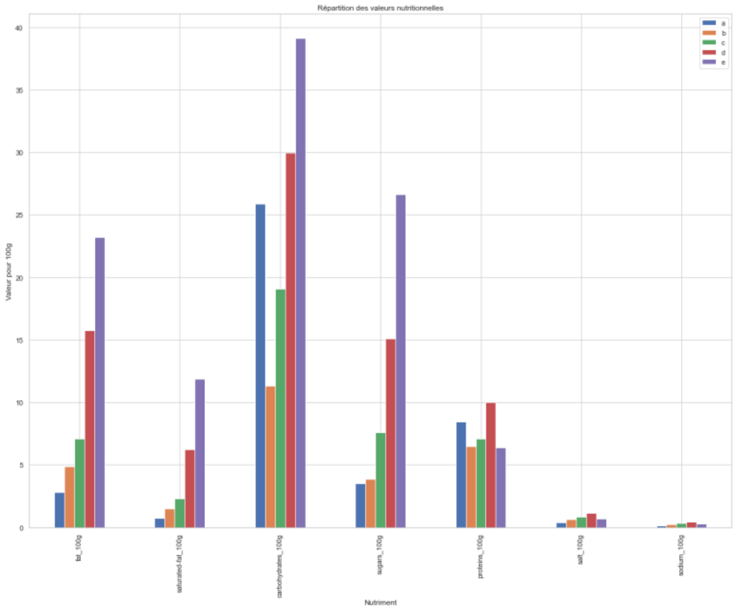
Synthèse

voilà

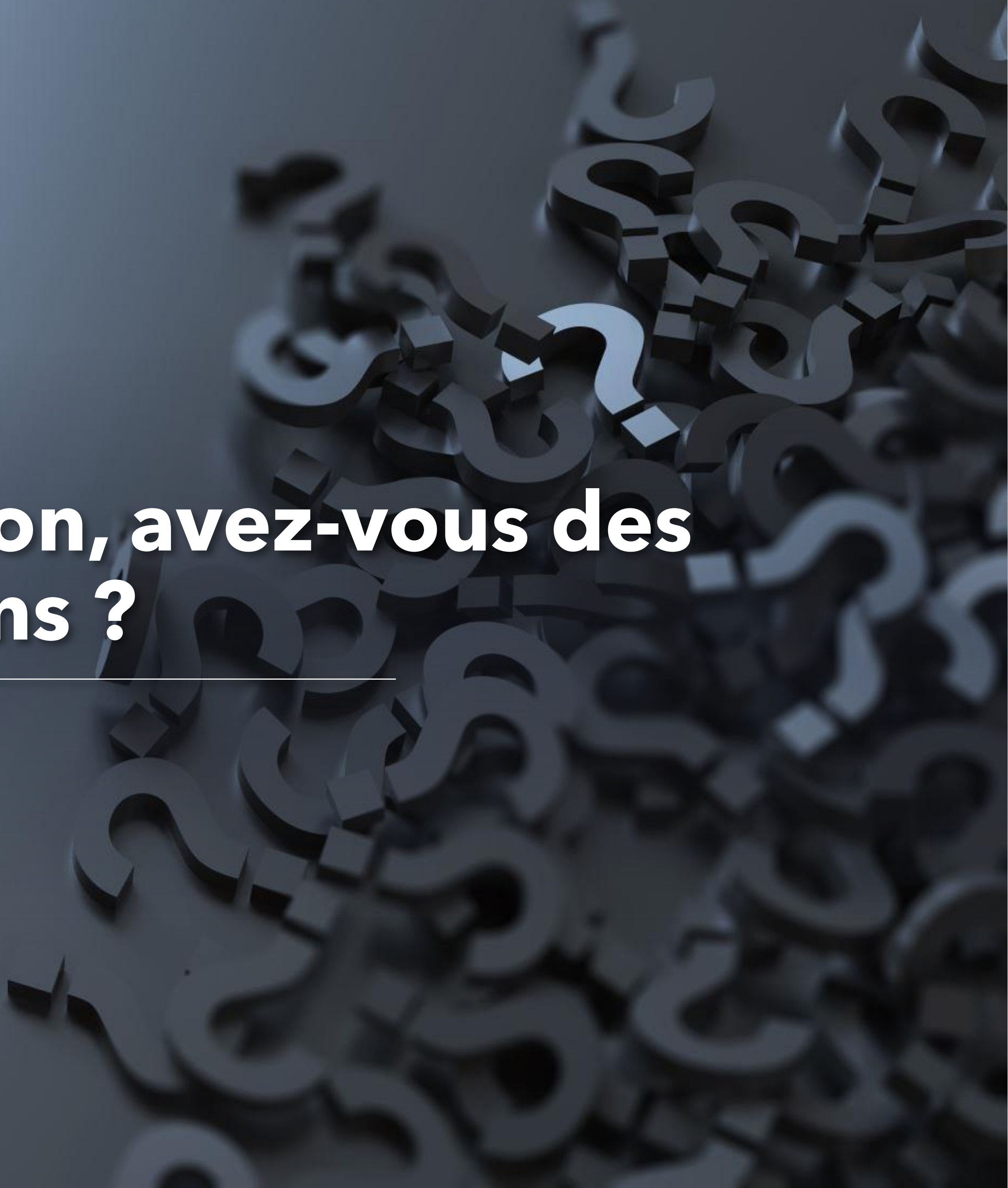
OPEN FOOD FACTS
l'information alimentaire ouverte

NUTRI-SCORE
A B C D E

scikit learn



Santé
publique
France



Merci de votre attention, avez-vous des questions ?