Santé publique France

Conception d'une application au service de la santé publique

Agenda

1. Contexte
2. Idée d’application
3. Approche
4. Sélection des données
5. Nettoyage
6. Moteur
7. Prochaines étapes
8. Annexe : l’environnement technique

1) Contexte

Application innovante

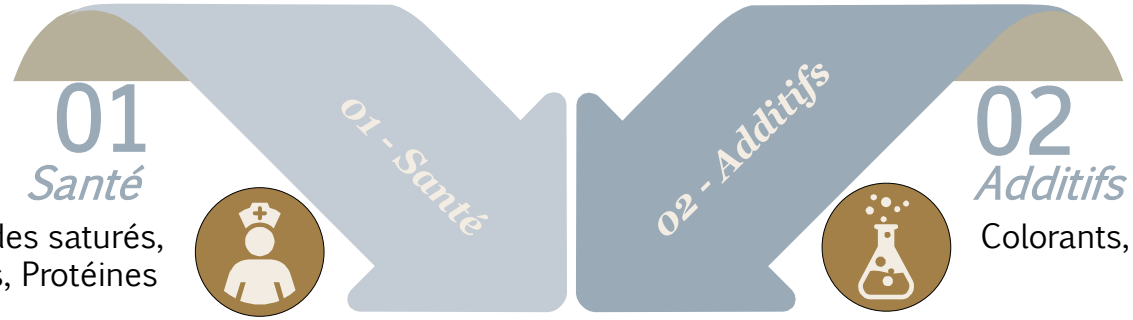
L'agence **"Santé publique France"** a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.

Open Food Fact met à disposition un jeu de données organisé de la manière suivante :
Les champs sont séparés en quatre sections :

1. Les informations générales sur la fiche du produit : nom, date de modification, etc.
2. Un ensemble de tags : catégorie du produit, localisation, origine, etc.
3. Les ingrédients composant les produits et leurs additifs éventuels.
4. Des informations nutritionnelles : quantité en grammes d'un nutriment pour 100 grammes du produit.

2) Idée d'application

Principe : proposer un ensemble d'aliments respectant des critères dans l'air du temps



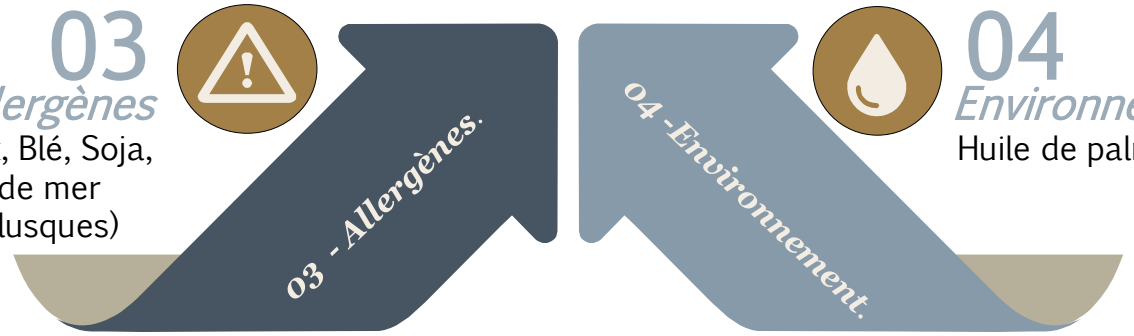
01
Santé

Energie, Sucres, Acides saturés, Sodium, Fruits, Fibres, Protéines

02
Additifs

Colorants, conservateurs, émulsifiants, …

Critères de choix d'un aliment



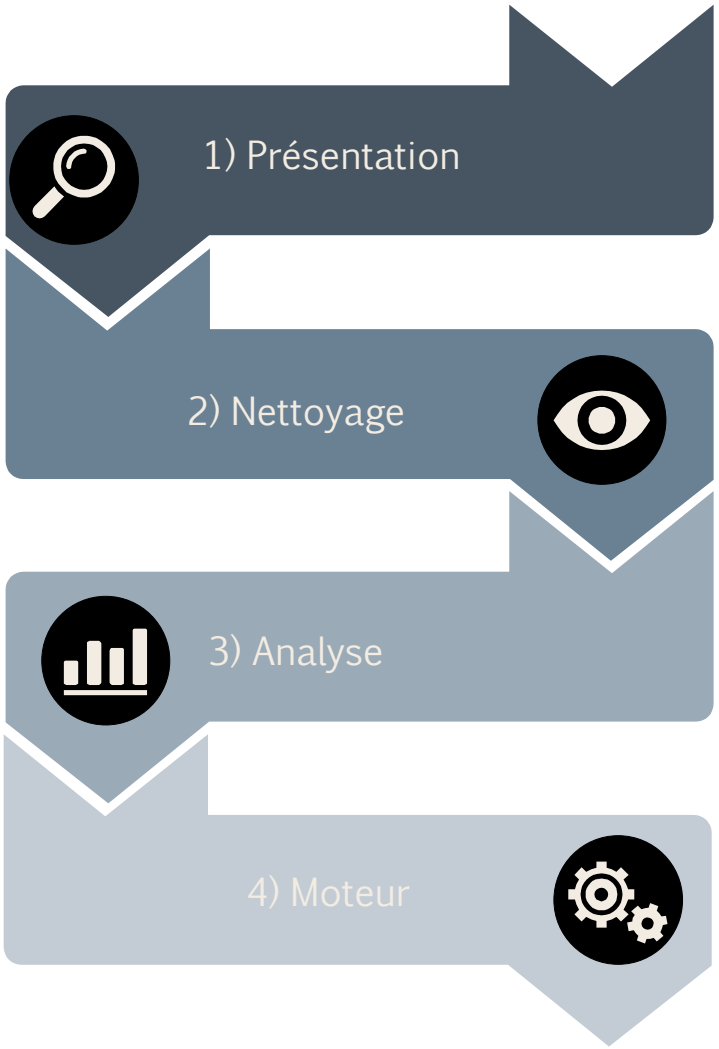
03
Allergènes

Arachides, Œufs, Lait, Noix, Blé, Soja, Graines de sésame, Fruits de mer (poisson, crustacés et mollusques)

04
Environnement

Huile de palme

3) Approche

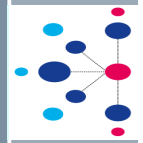


Présentation du jeu de données (Taille du jeu de données, Type de données, nombre de lignes, nombre de colonnes, ...)

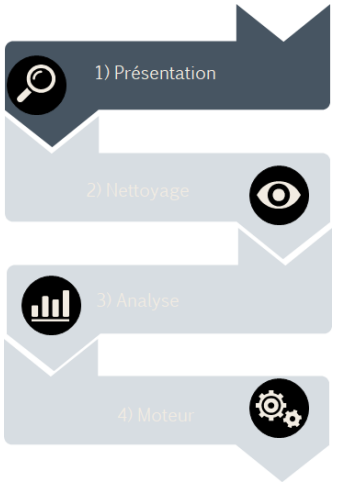
Sélection des données pertinentes pour l'analyse, detection des valeurs abérrantes, des doublons, analyses univariées, ...

Analyses multivariées

Vectorisation, Similarité cosinus



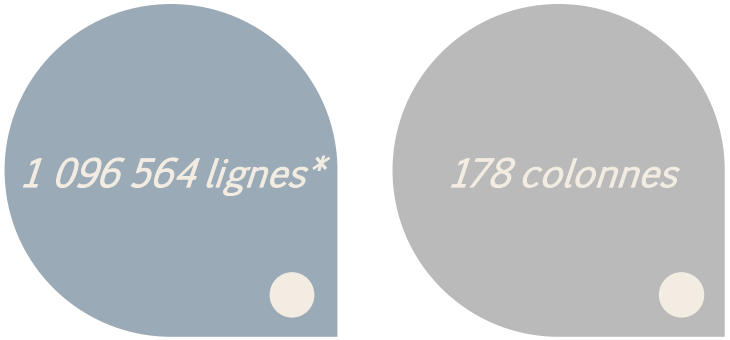
4) Présentation du jeu de données (1/2)



code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name
9310232100500	http://world-en.openfoodfacts.org/product/9310...	openfoodfacts-contributors	1556425403	2019-04-28T04:23:23Z	1565969538	2019-08-16T15:32:18Z	NaN
9310232100906	http://world-en.openfoodfacts.org/product/9310...	openfoodfacts-contributors	1564215237	2019-07-27T08:13:57Z	1565969541	2019-08-16T15:32:21Z	Vitasoy
9310232100999	http://world-en.openfoodfacts.org/product/9310...	foodorigins	1480987785	2016-12-06T01:29:45Z	1565969539	2019-08-16T15:32:19Z	NaN
9310232132013	http://world-en.openfoodfacts.org/product/9310...	clockwerx	1466239109	2016-06-18T08:38:29Z	1466905960	2016-06-26T01:52:40Z	Pura Original Milk

Extrait du jeu de données

4) Présentation du jeu de données (2/2)

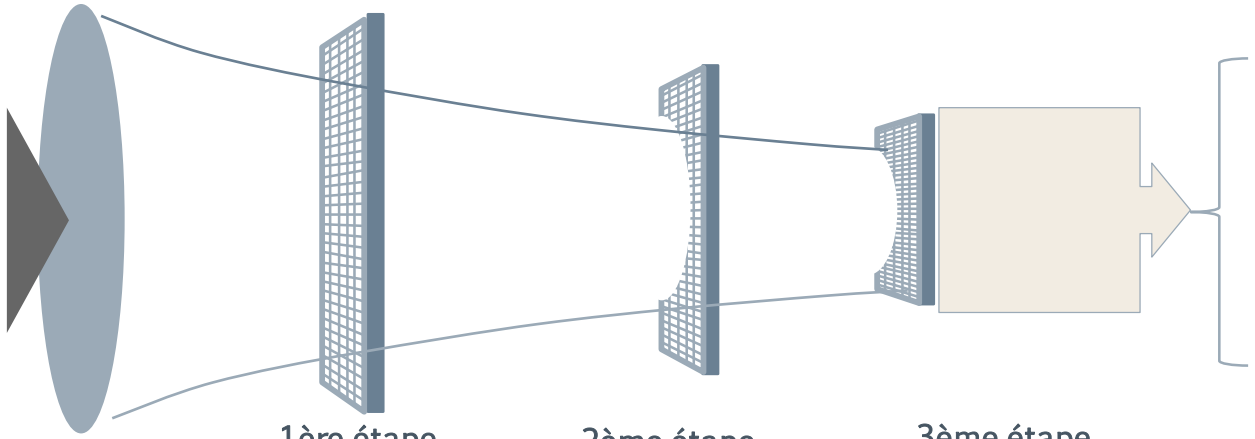


Indicateurs clés du jeu de données

* Le nombre de ligne étant important au regard de la mémoire affectée à la Virtual Machine Ubuntu, le fichier a été lu de manière itérative (chunk)

5) Nettoyage – Les étapes

OPEN FOOD FACTS
the free food products database



Smart Data Analytics

Smart Data

1ère étape

- Filtrage sur 21 colonnes
- Suppression des doublons

2ème étape

- Sélection des aliments français
- Suppression des boissons et nourritures bébé sans nutriscore

3ème étape

- Sélection des aliments disposant d'un nutriscore

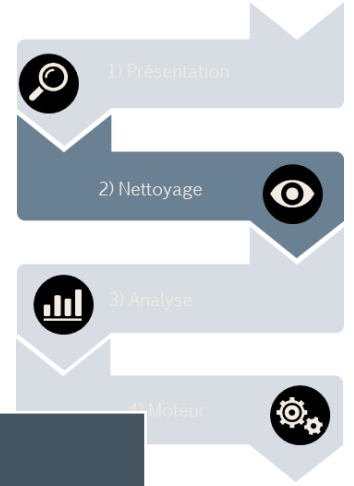
Caractéristiques	Valeur
Nombre de lignes	1 096 564
Nombre de colonnes	178

Caractéristiques	Valeur
Nombre de lignes	699 708
Nombre de colonnes	21

Caractéristiques	Valeur
Nombre de lignes	165 632
Nombre de colonnes	21

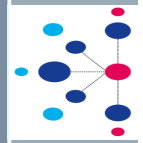
Caractéristiques	Valeur
Nombre de lignes	153 628
Nombre de colonnes	21

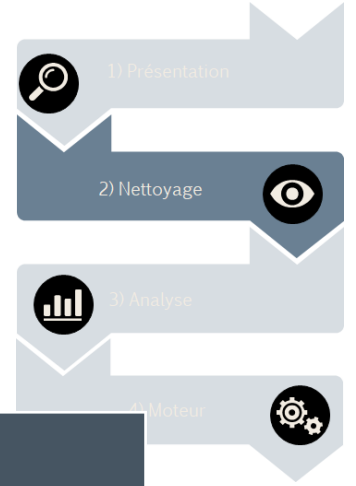
5) Nettoyage – 1ère étape (1/7)



Les colonnes sélectionnées

#	Code	Type	Description	Raison
1	Code	Object	Code de l'aliment	Donnée clé
2	product_name	Object	Nom de l'aliment	Donnée clé
3	categories	Object	Catégories de l'aliment	Donnée clé
4	categories_en	Object	Catégories de l'aliment en anglaise	Donnée clé
5	countries_en	Object	Pays de l'aliment en anglaise	Donnée clé
6	nutriscore_score	Object	Score du nutricore	Donnée clé
7	nutriscore_grade	Float64	Note du nutriscore	Donnée clé
8	energy-kj_100g	Float64	Valeur energétique en kj de l'aliment pour 100g	Intérêt pour le nutriscore
9	energy-kcal_100g	Float64	Valeur énergétique en kcal de l'aliment pour 100g	Intérêt pour le nutriscore
10	saturated-fat_100g	Float64	Niveau de graisses saturées pour 100g	Intérêt pour le nutriscore
11	sugars_100g	Float64	Niveau de sucres pour 100g	Intérêt pour le nutriscore
12	fiber_100g	Float64	Niveau de fibres pour 100g	Intérêt pour le nutriscore
13	proteins_100g	Float64	Niveau de proteines pour 100g	Intérêt pour le nutriscore





5) Nettoyage – 1ère étape (2/7)

Les colonnes sélectionnées

#	Code	Type	Description	Raison
14	salt_100g	Float64	Niveau de sel pour 100g	Intérêt pour le nutriscore
15	sodium_100g	Float64	Niveau de sodium pour 100g	Intérêt pour le nutriscore
16	fruits-vegetables-nuts_100g	Float64	Niveau de fruits à coque pour 100g	Intérêt pour le nutriscore
17	allergens_en	Object	Noms des allergènes dans l'aliment	Intérêt pour l'application
18	traces_en	Object	Noms des allergènes dans l'aliment	Intérêt pour l'application
19	additives_en	Object	Noms des additifs dans l'aliment	Intérêt pour l'application
20	ingredients_from_palm_oil_n	Object	Présence d'huile de palme	Intérêt pour l'application
21	ingredients_that_may_be_from_palm_oil_n	Object	Présence éventuelle d'huile de palme	Intérêt pour l'application
22	ingredients_from_palm_oil	Float64	Présence d'huile de palme	Intérêt pour l'application
23	ingredients_that_may_be_from_palm_oil	Float64	Présence éventuelle d'huile de palme	Intérêt pour l'application
24	'alcohol_100g']	Float64	Niveau de sodium pour 100g	Intérêt pour l'application

5) Nettoyage – 1ère étape (3/7)

Les doublons

	code	product_name	categories	categories_en	countries_en	nutriscore_score	nutriscore_grade	energy-kj_100g	energy-kcal_100g	s
68735	45364301028	Chopped Garlic	Garlic	Plant-based foods and beverages,Plant-based fo...	United States	-14.0	a	NaN	280.0	
68738	45364306696	Chopped Garlic	en:frozen-chopped-garlic	Plant-based foods and beverages,Plant-based fo...	United States	-14.0	a	NaN	280.0	
68739	45364306702	Chopped Garlic	Garlic	Plant-based foods and beverages,Plant-based fo...	United States	-14.0	a	NaN	280.0	
396433	3222473991730	Petits pois doux très fins bio	Aliments et boissons à base de végétaux, Alime...	Plant-based foods and beverages,Plant-based fo...	France	-13.0	a	295.0	293.0	
398821	3222476759054	Petits pois doux extra-fins	Aliments et boissons à base de végétaux, Alime...	Plant-based foods and beverages,Plant-based fo...	France,Germany	-13.0	a	295.0	293.0	
425662	3256224701101	Petits pois extra fins	Aliments et boissons à base de végétaux, Alime...	Plant-based foods and beverages,Plant-based fo...	France	-13.0	a	317.0	314.0	
425687	3256224701774	Petits pois doux extra fins	Aliments et boissons à base de végétaux, Alime...	Plant-based foods and beverages,Plant-based fo...	France	-13.0	a	317.0	314.0	

Doublon 1

Doublon 2

Doublon 3

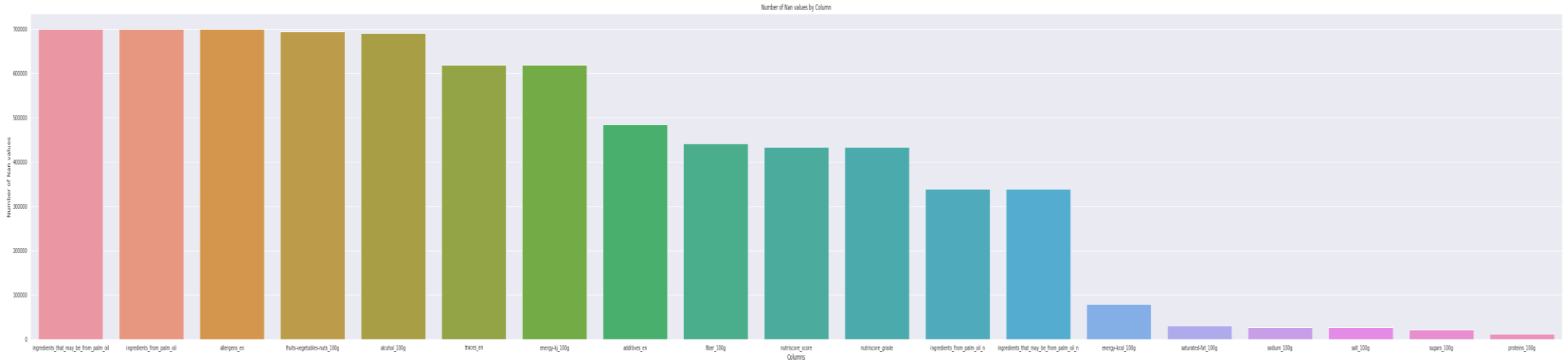
Exemples de doublons

La suppression des doublons a permi de passer d'un dataset de 1 096 564 lignes à 699 894 lignes

5) Nettoyage – 1ère étape (4/7)

Analyse de la complétude des données

1) Par colonne



Caractéristiques	Valeur
Nombre de colonnes	19
Moyenne du nombre de valeurs	389195.315789
Ecart type	275460.352094
Min de valeurs	11538
25%	55188
50%	434605
75%	655089.5
Max	699 894

Number of Nan by columns

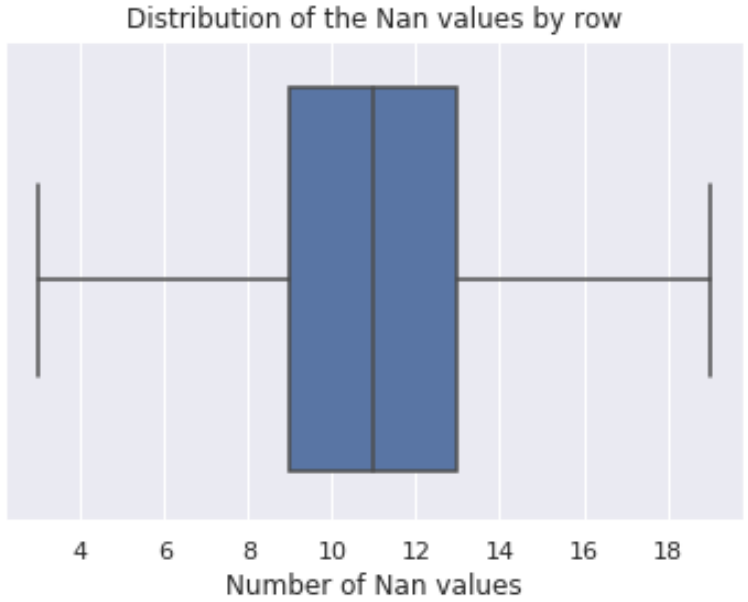
- 310698 données sont en moyenne manquantes par colonne
- Les colonnes suivantes sont entièrement vides :
  - ✓ ingredients_that_may_be_from_palm_oil
  - ✓ ingredients_from_palm_oil
  - ✓ allergens_en'

Elles sont donc supprimées.

5) Nettoyage – 1ère étape (5/7)

Analyse de la complétude des données

2) Par ligne



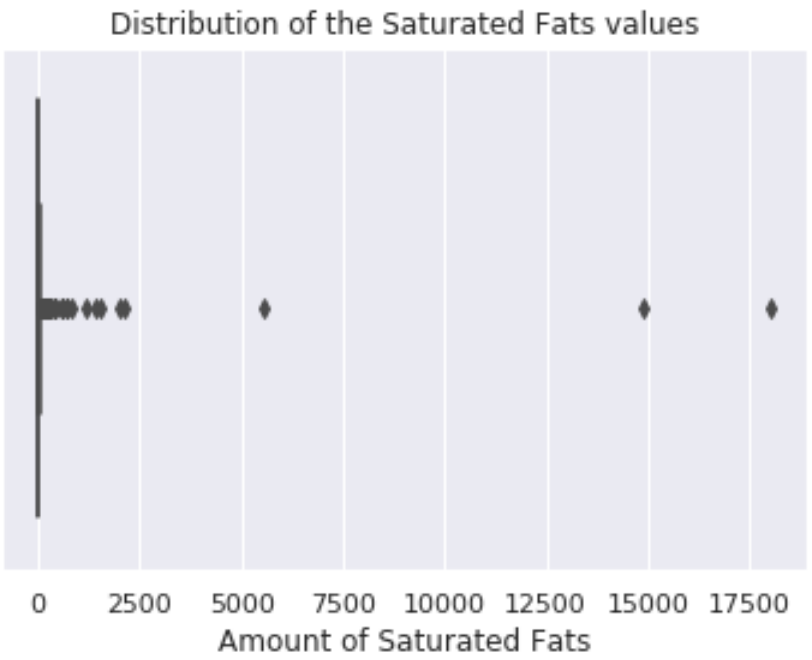
Caractéristiques	Valeur
Nombre de lignes	699 894
Moyenne du nombre de valeurs	10.565473
Ecart type	2.458846
Min de valeurs	3
25%	9
50%	11
75%	13
Max	19

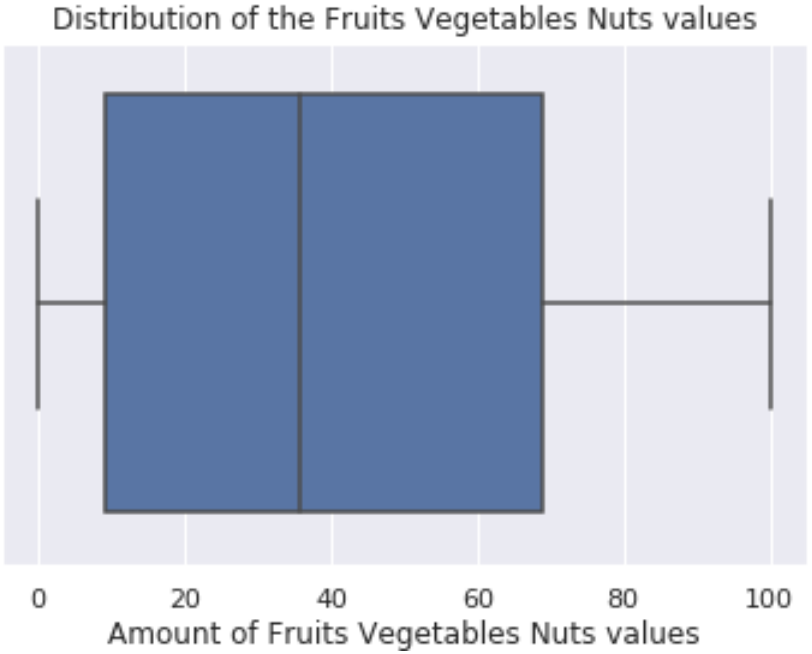
Number of Nan by rows

- En moyenne, 10,5 données par colonnes sont manquantes

5) Nettoyage – 1ère étape (6/7)

Détection des valeurs aberrantes qui ne sont pas comprises entre 0 et 100

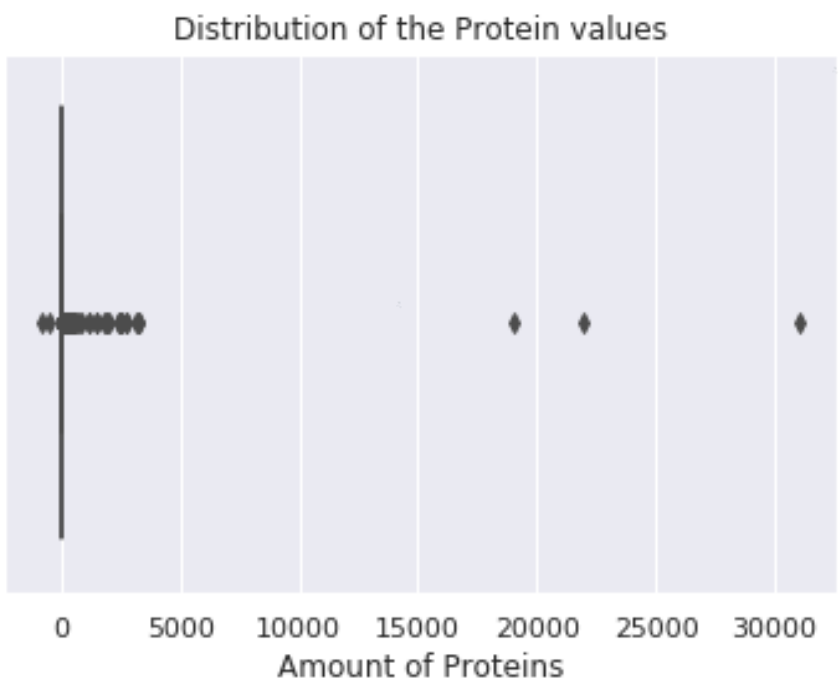


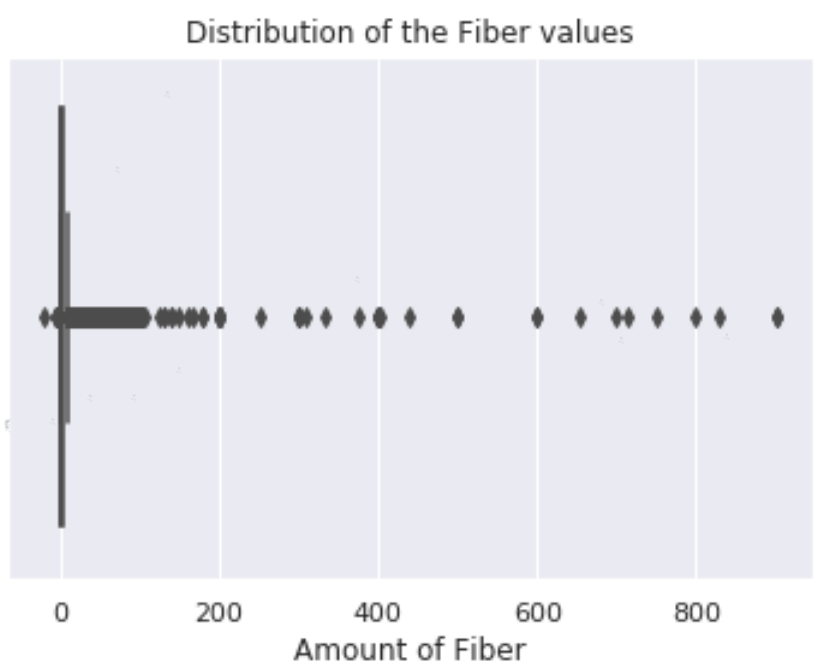


Les aliments contenant des données abérrantes sont supprimés du dataset

5) Nettoyage – 1ère étape (7/7)

Détection des valeurs aberrantes qui ne sont pas comprises entre 0 et 100



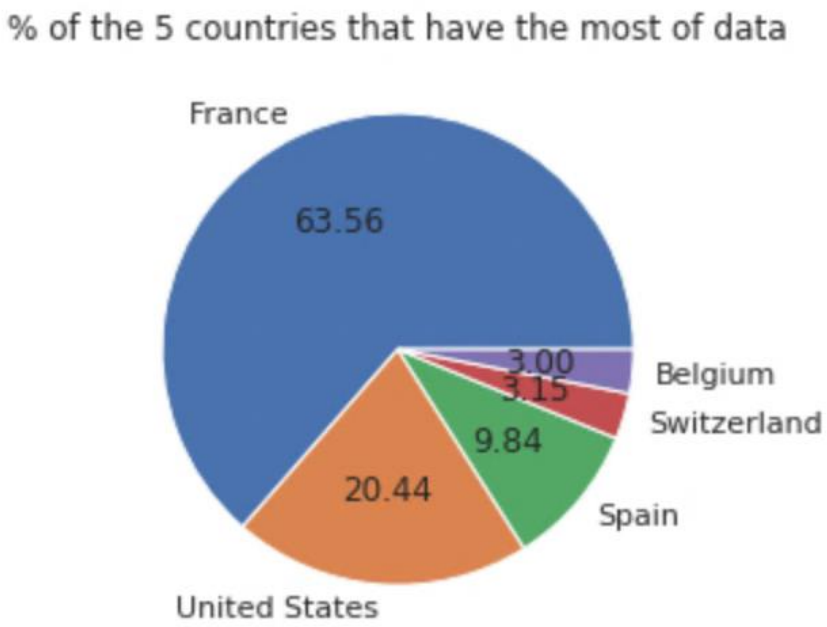


Les aliments contenant des données abérrantes sont supprimés du dataset. A l'issue de la suppression des données abérrantes, il reste 699 708 lignes sur 699 894 dans le dataset

5) Nettoyage – 2éme étape (1/3)

Analyse des pays

#	Pays	Nombre
1	France	376 643
2	United States	121 123
3	Spain	58 324
4	Switzerland	18 670
5	Belgium	17 794
6	Germany	14 759
7	United Kingdom	7 758
8	Mexico	3 571
9	Italy	3 058
10	Canada	2 800



Le nombre d'aliments Français dominent, il représente la moitié du volume du dataset. Dans ce cadre et afin de faciliter la partie NLP qui suivra, un filtre est appliqué pour conserver uniquement les aliments français.*

* Certains aliments étaient catégorisés suivant le pattern fr:*, ils ont été remplacés par France

5) Nettoyage – 2éme étape (2/3)

Analyse des catégories

#	Catégories	Nombre
1	Beverages,Non-Alcoholic beverages,Unsweetened	1 882
2	Beverages,Sweetened beverages,Non-Alcoholic	1 337
3	Beverages,Sweetened beverages,Non-Alcoholic be.	1 287
4	Snacks,Sweet snacks,Chocolates,Dark chocolates	1 182
5	Snacks,Sweet snacks,Biscuits and cakes,Biscuits	846

Présentation des 5 categories brutes

WordCloud des mots les plus presents dans la colonne categories_en

Les categories sont présentés sous le pattern : Catégorie, Sous-categorie, sous-sous-categorie. Dans notre context,e uniquement la section Catégorie est pertinente. Un filtre est donc mis en place pour concerver uniquement cette partie.

5) Nettoyage – 2éme étape (3/3)

Analyse des catégories

#	Catégories	Nombre
1	Plant-based foods and beverages	7193
2	Snacks	3365
3	Diaries	2838
4	Meals	2200
5	Meats	2200
6	Beverages	1485
7	Groceries	725
8	Canned foods	683
9	Desserts	462
10	Seafood	444
11	Spreads	377
12	Baby foods	109

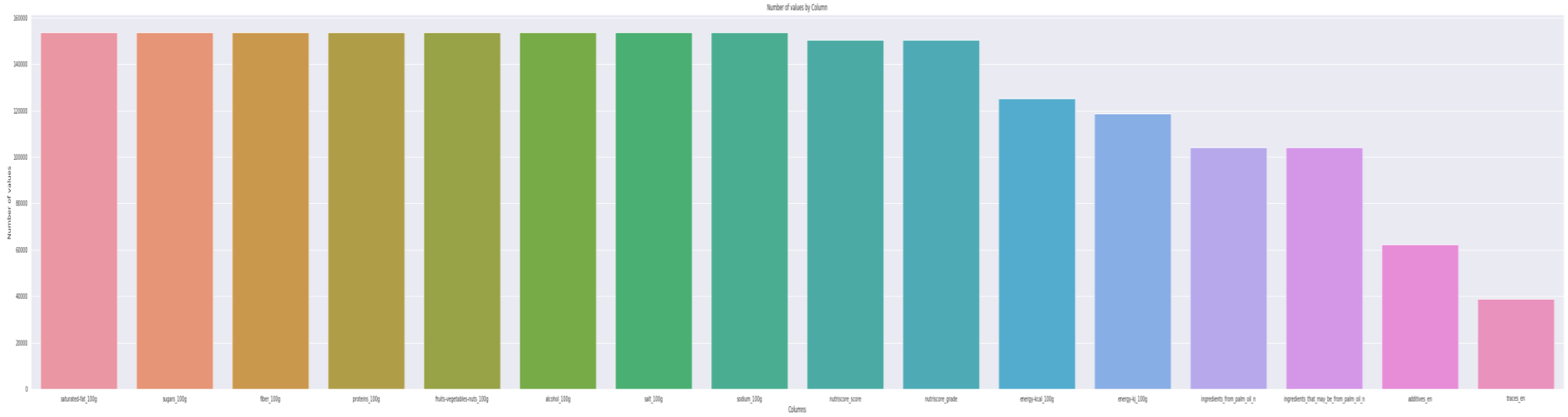
Présentation des 12 categories principales

WordCloud des mots les plus presents dans la colonne categories_en

Afin de limiter le data cleansing sur le nom des categories, uniquement les categories présentent plus de 13 fois sont conservées. A l'issue de cette étape, il reste 165 632 lignes dans le dataset.

5) Nettoyage – 3éme étape

Calcul des nutriscores manquants pour lesquels toutes les valeurs nécessaires sont disponibles hormis pour les boissons et la nourriture bébé qui ont des méthodes de calcul différentes.



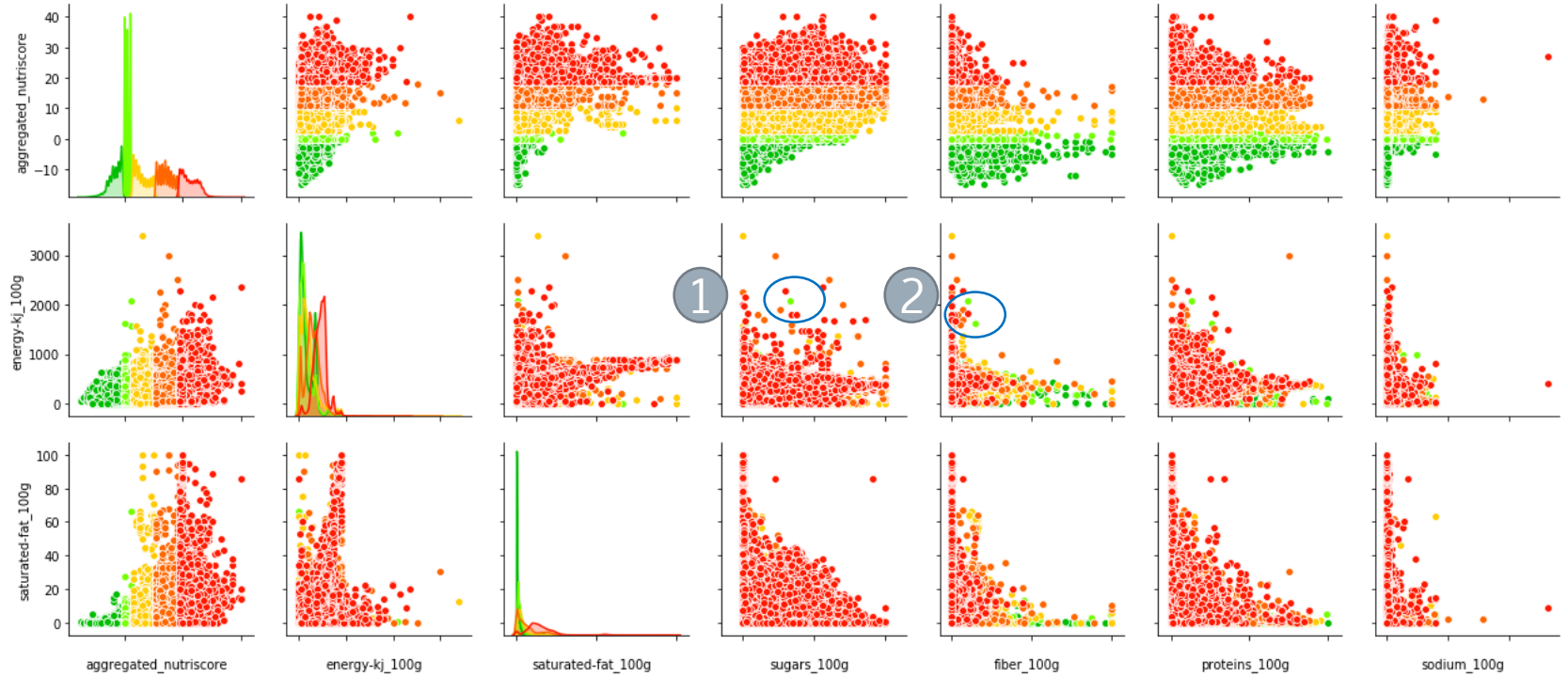
3080 nutriscores ont été calculés. A l'issue de cette étape, il reste 153 628 lignes dans le dataset.

6) Analyse

Quelques éléments :

- Corrélation positive entre le nutriscore et la valeur énergétique
- Corrélation positive entre le nutriscore et le niveau de graisse
- Corrélation positive entre le nutriscore et le niveau de sucres
- Corrélation negative entre le niveau de protéines et le niveau de sucre

6) Analyse



1

code	product_name	categories	nutriscore_grade	energy-kj_100g	saturated-fat_100g	sugars_100g	sodium_100g	fruits-vegetables-nuts_100g	fiber_100g	proteins_100g
2609927051621	Mélange bien être	en:fruits-based-foods	b	2081.978967	0.8	33.0	0.072	0.0	9.6	13.0

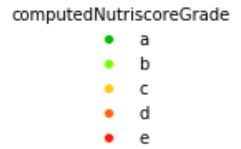
2

code	product_name	categories	nutriscore_grade	energy-kj_100g	saturated-fat_100g	sugars_100g	sodium_100g	fruits-vegetables-nuts_100g	fiber_100g	proteins_100g
2609927051621	Mélange bien être	en:fruits-based-foods	b	2081.978967	0.8	33.0	0.072	0.0	9.6	13.0
6168651777338	lentilles corail	Aliments et boissons à base de végétaux,Alimen...	b	1620.936902	0.2	2.0	0.008	0.0	15.0	25.8

Ces valeurs de nutriscores sont correctes de part sa formule car bien que l'énergie soit importante, elle est balancée par un haut niveau de fibres et de protéines

6) Analyse

3

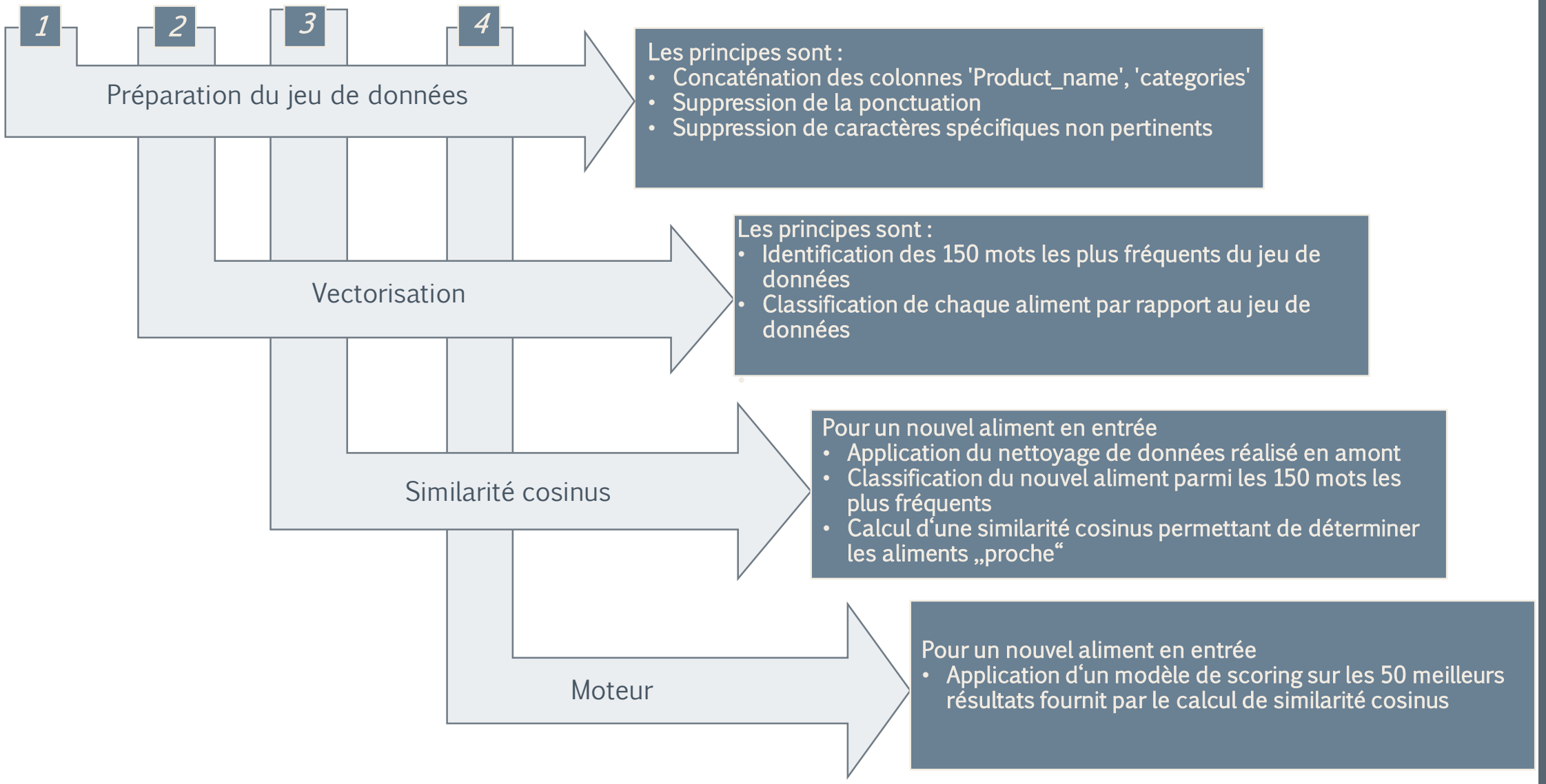


3

code	product_name	categories	nutriscore_grade	energy-kj_100g	saturated-fat_100g	sugars_100g	sodium_100g	fruits-vegetables-nuts_100g	fiber_100g	proteins_100g
3259920055509	Tisane de noël	Boissons, Boissons sans sucre ajouté	d	0.956023	1.00	5.0	0.800	0.0	100.0	25.00
3416951410589	Table d'Adrien confit de cuisses de canard	Viandes, Volailles, Canards, Cuisses de canard	c	260.994264	6.10	0.0	0.528	0.0	100.0	25.90
3588050003808	Pomme de terre Bio (France) 1	Boissons, Boissons sans sucre ajouté	d	80.066922	0.16	1.0	0.000	0.0	100.0	1.86

Ces valeurs ne sont pas correctes. (Niveau de fibres à 100g, puis 25g de protéines dans de la tisane ?)

6) Moteur


1
2
3
4
Préparation du jeu de données
Les principes sont :
• Concaténation des colonnes 'Product_name', 'categories'
• Suppression de la ponctuation
• Suppression de caractères spécifiques non pertinents
Vectorisation
Les principes sont :
• Identification des 150 mots les plus fréquents du jeu de données
• Classification de chaque aliment par rapport au jeu de données
Similarité cosinus
Pour un nouvel aliment en entrée
• Application du nettoyage de données réalisé en amont
• Classification du nouvel aliment parmi les 150 mots les plus fréquents
• Calcul d'une similarité cosinus permettant de déterminer les aliments „proche“
Moteur
Pour un nouvel aliment en entrée
• Application d'un modèle de scoring sur les 50 meilleurs résultats fournit par le calcul de similarité cosinus

6) Moteur - Vectorisation

Liste de mots du processus de stemmisation

'abricot', 'aliment', 'aliment boisson', 'barr', 'basilic', 'beurr', 'beurr snack', 'beverag', 'biscuit', 'blanc', 'boeuf', 'boisson', 'boisson végétal', 'bonbon', 'breton', 'cacao', 'campagn', 'canard', 'caramel', 'carott', 'champignon', 'chevr', 'chip', 'chocolat', 'chocolat lait', 'chocolat noir', 'chocolat snack', 'citron', 'coco', 'complet', 'compot', 'confitur', 'confitur extra', 'conserv', 'crem', 'crevet', 'crêp', 'cuisin', 'cuit', 'céréal', "d'oliv", "d'orang", 'dair', 'dessert', 'dind', 'doux', 'emmental', 'enti', 'epicer', 'extra', 'farin', 'figu', 'filet', 'food', 'food beverag', 'fourr', 'frais', 'frambois', 'fromag', 'fruit', 'galet', 'glac', 'goût', 'grain', 'gras', 'gras canard', 'grass', 'grill', 'gâteau', 'hach', 'haricot', 'huil', 'jambon', "l'huil", 'lait', 'lait snack', 'laiti', 'lentill', 'légum', 'madelein', 'mangu', 'mati', 'mati grass', 'mayonnais', 'miel', 'moutard', 'myrtill', 'natur', 'naturel', 'nectar', 'noir', 'noir snack', 'noiset', 'noix', 'oeuf', 'oliv', 'orang', 'pain', 'petit', 'petitdéjeuner', 'pizz', 'plantbased', 'plantbased food', 'plat', 'plat prépar', 'poir', 'pois', 'poisson', 'poisson oeuf', 'pomm', 'pomm terr', 'porc', 'poulet', 'produit', 'produit laiti', 'produit tartin', 'prépar', 'pépit', 'pêch', 'raisin', 'rillet', 'roug', 'salad', 'sardin', 'sauc', 'saucisson', 'saumon', 'saveur', 'sech', 'sirop', 'snack', 'soup', 'sucr', 'supérieur', 'surgel', 'tart', 'tartin', 'terr', 'terrin', 'thon', 'tomat', 'tranch', 'vanill', 'vert', 'vierg', 'vinaigr', 'volaill', 'végétal', 'wich', 'yaourt'

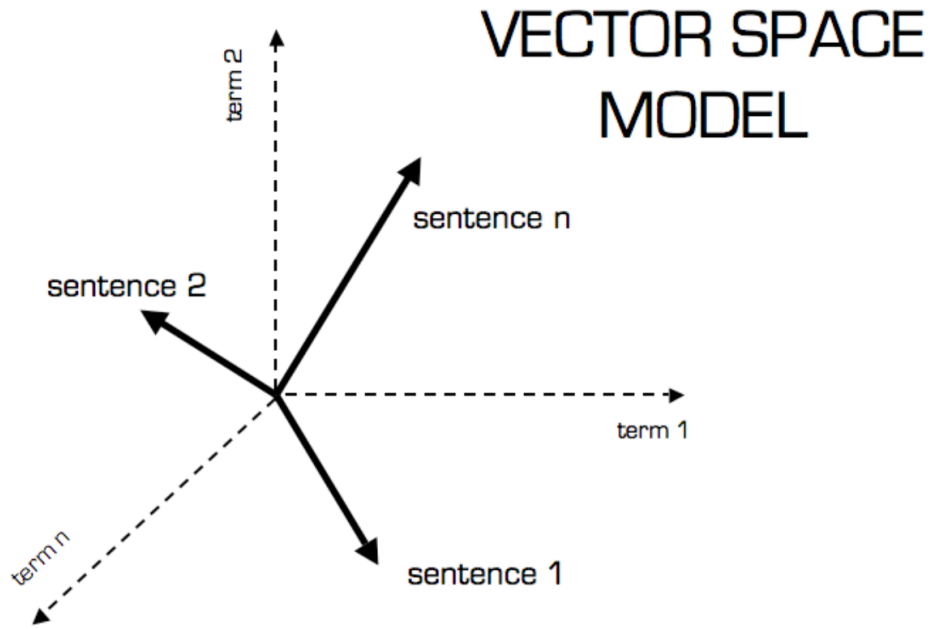
Liste de mots du processus de stemmisation

	abricot	aliment	aliment boisson	barr	basilic	beurr	beurr snack	beverag	biscuit	blanc	boeuf	boisson	boisson végétal	bonbon	breton
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
153623	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
153624	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
153625	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
153626	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0
153627	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

153628 rows × 150 columns

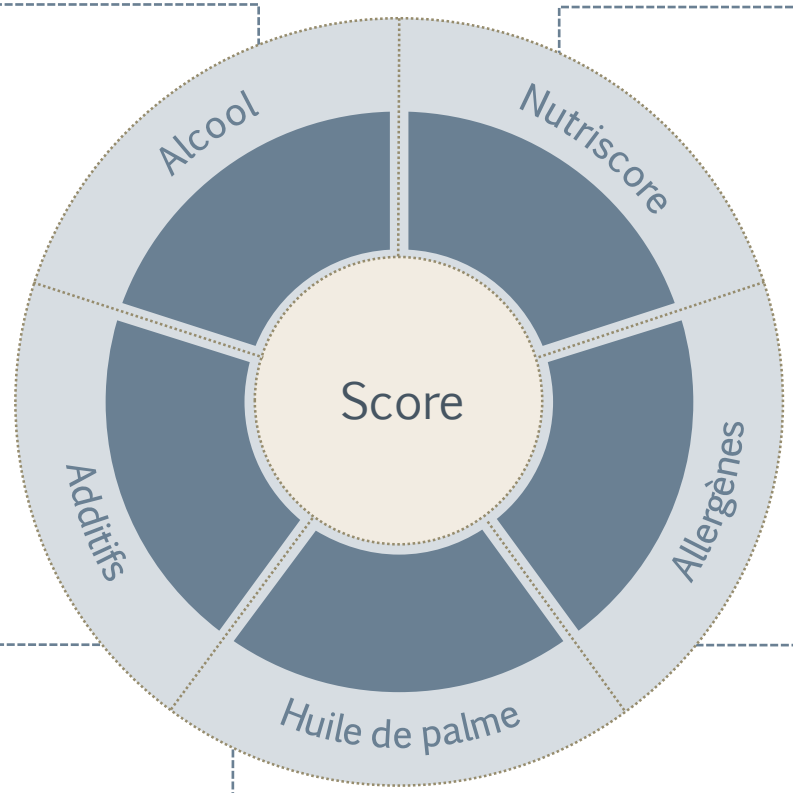
Résultat de la vectorisation

6) Moteur - Similarité cosinus



Similarité Cosinus

6) Moteur - Modèle



- Poids : 10%
- Echelle :

Caractéristiques	Valeur
Absence	10
Présence	0

- Poids : 50%
- Classification des aliments présentis en 10 categories.
- Echelle :

Caractéristiques	Valeur
Nutriscore bas	10
...	
Nutriscore haut	1

- Poids : 20%
- Calcul du nombre de séparateurs
- Echelle :

Caractéristiques	Valeur
Vide	20
0	10
...	
9	1

- Poids : 10%
- Calcul du nombre de séparateurs
- Echelle :

Caractéristiques	Valeur
Vide	20
0	10
...	
9	1

- Poids 10%
- Echelle :

Caractéristiques	Valeur
Absence	10
Présence	0

6) Moteur – Résultats (1/3)

1) Poulet rôti

#	Product Name	Categories	Score
71366	Poulet, nouilles chinoises, légumes croquants,...	Plats au Poulet	125
7048	Cuisse de poulet	Cuisse de Poulet	125
17116	Filet de Poulet Rôti	Filets de Poulet	125
114529	Filet de poulet quinoa	Plats au Poulet	125
7306	Aiguillettes de poulet grillé	Aiguillettes de Poulet	120
67819	Aiguillettes de Poulet panées	Aiguillettes de Poulet panées	120
62598	Poulet au curry et son riz basmati	Poulet au curry	120

2) Charcuterie

#	Product Name	Categories	Score

6) Moteur – Résultats (2/3)

3) Poissons panés

#	Product Name	Categories	Score
7026	Cabillaud pané x2	poissons panés de cabillaud	125
88301	Croquettes Poisson	Croquettes de poisson	125
80540	Soupe de poissons	Soupe de poisson	125
119560	Soupe poisson	Soupes de poissons	125
93324	Soupe de poissons charentaise	Soupes de poissons	120

4) Saucisses

#	Product Name	Categories	Score
4560	Sauce Spaghetti Wafu Nama Fumi	Pasta sauces	130
136333	Sauce Pita	Pita sauces	130
90338	sauce aigre	Sauces aigre-douces	130
12732	Sauce tomate	Sauce	130
22150	Sauce Pizza Zapetti	Pizza sauce	125

Limitation due au processus de stemmatisation (saucisses => sauce)

6) Moteur – Résultats (3/3)

5) Chocolat

#	Product Name	Categories	Score
19221	Crème dessert au chocolat	crèmes dessert chocolat 125.0	125
108768	Double glace poire,sauce chocolat & chocolat noir	bâtonnets glacés au chocolat	110
13554	Pain au Chocolat	pain au chocolat	110
6161	Chocolat	muffins aux pépites de chocolat	110
7153	Pains au chocolat AOP x4	chocolates	110
84849	Paturette Chocolat	crèmes dessert chocolat	106
62598	Galettes nappées riz chocolat au lait	galettes de riz au chocolat au lait	105

7) Prochaines étapes

Amélioration du modèle

- Se baser sur un algorithme des proches voisins pour permettre des temps de calcul réduit
- Faire de la lemmatisation au lieu du stemming
- Rendre l'application multilingue

8) Environnement technique