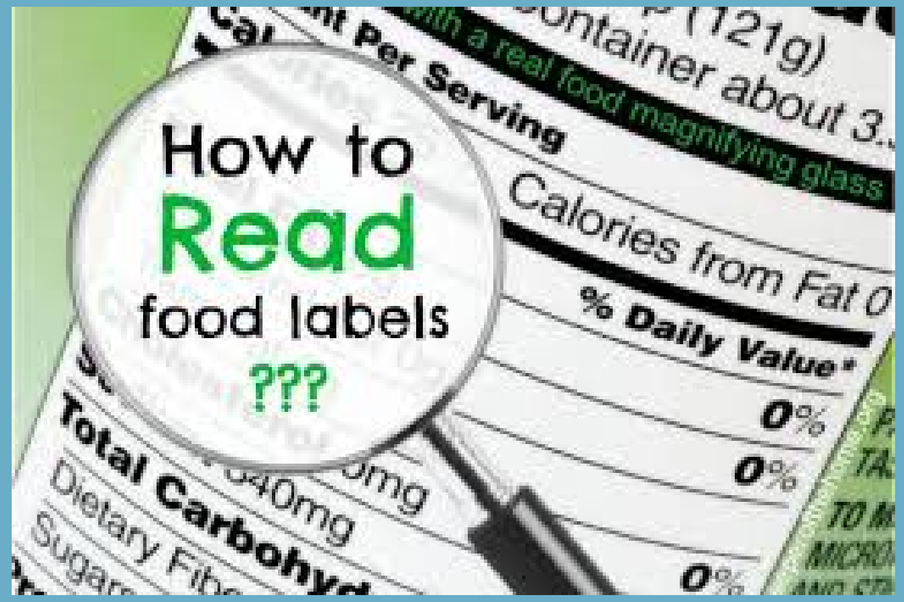
FOODFLIX

Etude de faisabilité d'une application

PLAN

1) CONTEXTE
2) OBJECTIFS
3) RETROPLANNING
4) ANALYSE EXPLORATOIRE DES DONNEES
5) SELECTION DES DONNEES
6) UPDATE DU PLANNING
7) NETTOYAGE
8) PRESENTATION DES RESULTATS

1/ CONTEXTE



FOODFLIX

FoodFlix, une startup de la FoodTech, travaille actuellement sur une application permettant de recommander le meilleur produit (le produit ayant le meilleur nutriscore) à un utilisateur selon un mot clé ou un ensemble de mot clés.

1/ CONTEXTE

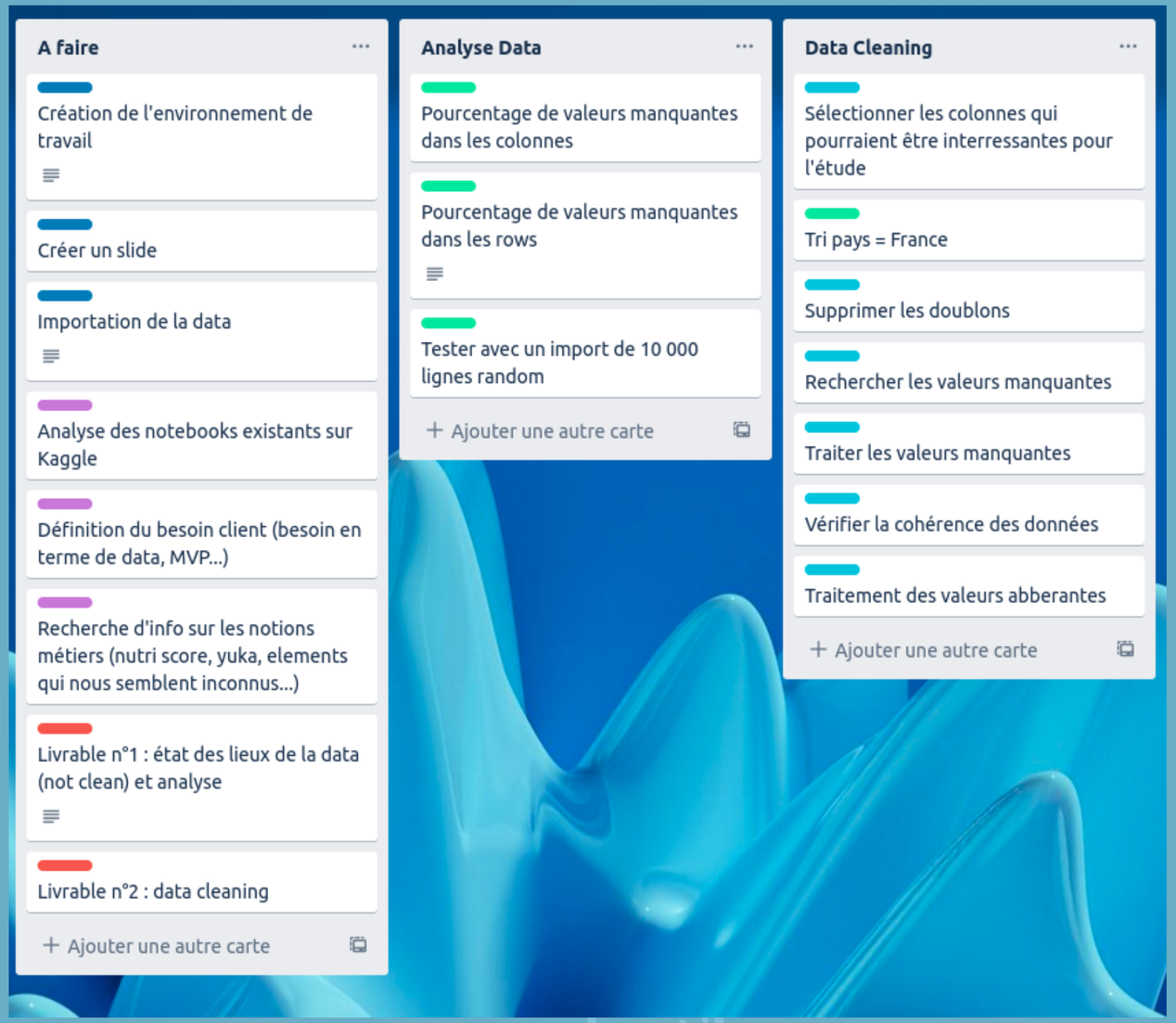
OPEN FOOD FACTS

Open Food Facts est une base de données sur les produits alimentaires, faite par tout le monde, pour tout le monde. Elle permet de faire de décoder les étiquettes des produits et de faire des choix plus informés. De plus, comme comme les données sont open data, tout le monde peut les utiliser pour tout usage.

2/ OBJECTIFS

- ANALYSER LA DONNEE BRUTE
- SELECTIONNER LE SOUS ENSEMBLE PERMETTANT DE METTRE EN PLACE LA SOLUTION
- PROPOSER UN JEU DE DONNEES DE BONNE QUALITE

3/ RETROPLANNING



4/ ANALYSE EXPLORATOIRE

PRÉSENTATION DU JEU DE DONNÉES

356 027 LIGNES

163 COLONNES

5/ SELECTION DES DONNÉES

25 COLONNES

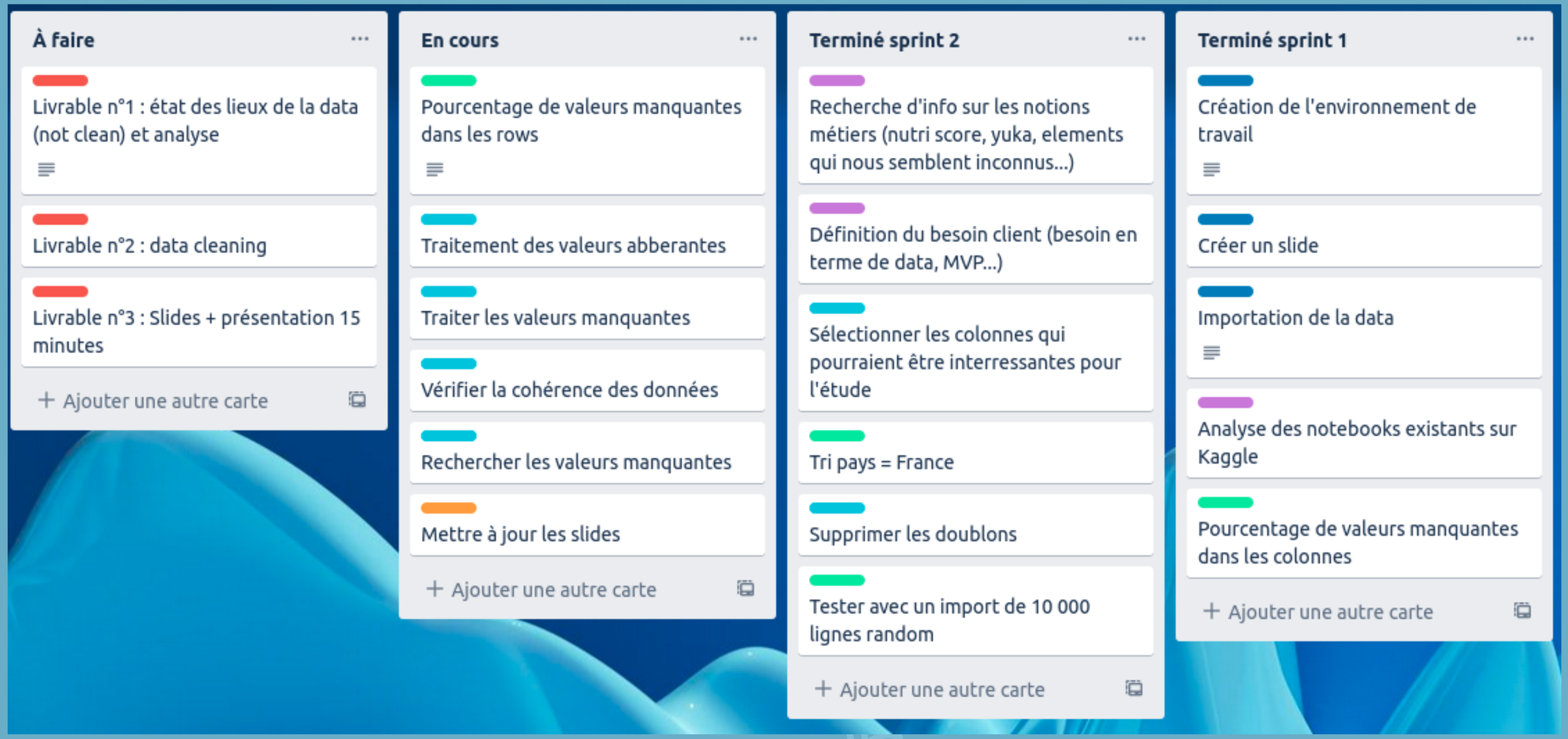
Nom de la colonne	Description de la colonne	Intérêt
product_name	Nom du produit	Intérêt pour le choix des produits par l'utilisateur
generic_name	Nom générique du produit	
brands	Marque du produit	
countries	Pays dans lequel on trouve le produit	
category	Catégorie du produit	
nutrition_grade_fr	Lettre correspondante au nutriscore	Intérêt pour le nutriscore
energy_100g	Energie pour 100g	
energy-from-fat_100g	Energie donné par les graisses pour 100g	
fat_100g	Quantité de graisses pour 100g	
saturated-fat_100g	Quantité de graisses saturées pour 100g	
carbohydrates_100g	Quantiité de carbohydrates pour 100g	
sugars_100g	Quantité de sucres pour 100g	
-sucrose_100g	Quantité de sucrose pour 100g	
-glucose_100g	Quantité de glucose pour 100g	
-fructose_100g	Quantité de fructose pour 100g	
-lactose_100g	Quantité de lactose pour 100g	
-maltose_100g	Quantité de maltose pour 100g	
-maltodextrins_100g	Quantité de maltodextrine pour 100g	
fiber_100g	Quantité de fibres pour 100g	
proteins_100g	Quantité de protéines pour 100g	
salt_100g	Quantité de sel pour 100g	
sodium_100g	Quantité de sodium pour 100g	
fruits-vegetables-nuts_100g	Quantité de fruits/légumes/noix pour 100g	
nutrition-score-fr_100g	Nutriscore (fr)	
nutrition-score-uk_100g	Nutriscore (en)	

5/ SELECTION DES DONNÉES

SÉLECTION DES DONNÉES FRANÇAISES

Opération effectuée	Nombre de lignes restantes
Suppression des lignes sans pays	355 752 lignes
Filtre pour garder les lignes des produits vendus en France	128 909 lignes

6/ REVUE DU PLANNING



7/ NETTOYAGE DES DONNÉES

VERIFICATION DU REMPLISSAGE DES VALUES_100G

```
values_100g = ['fat_100g',
               'saturated-fat_100g',
               'carbohydrates_100g',
               'sugars_100g',
               'fiber_100g',
               'proteins_100g',
               'salt_100g',
               'sodium_100g',
               'fruits-vegetables-nuts_100g']
```

Si l'ensemble de ces valeurs sont nulles, on ne pourra pas calculer le NutriScore, donc on va supprimer les lignes où l'ensemble des values_100g sont nulles.

7/ NETTOYAGE DES DONNÉES

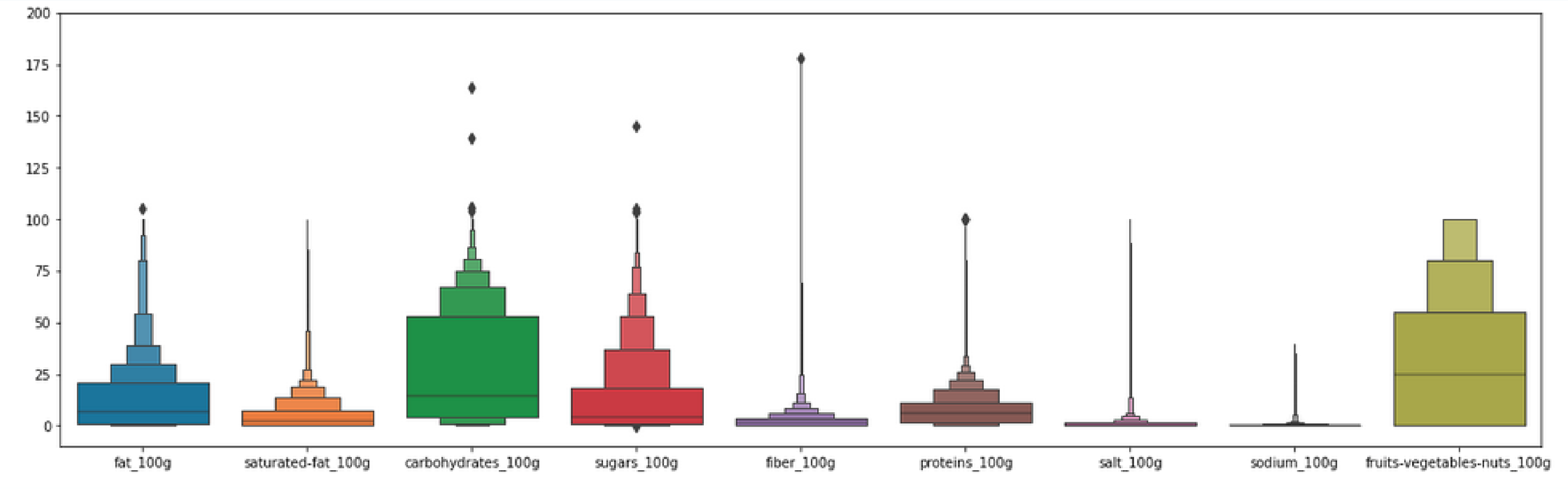
SUPPRESSION DES DOUBLONS

	product_name	generic_name	brands	categories	countries	nutrition_grade_fr	energy_100g	energy-from-fat_100g	fat_100g	saturated-fat_100g
19719	Crème de marrons de l'Ardèche	Crème de marrons	Clément Faugier	Aliments et boissons à base de végétaux,Alimen...	France	d	1147.0	NaN	0.8	0.2
19721	Crème de marrons de l'Ardèche	Crème de marrons	Clément Faugier	Aliments et boissons à base de végétaux,Alimen...	France	d	1147.0	NaN	0.8	0.2
28678	Double Chocolate Mini Bites	NaN	M&S	NaN	France,United Kingdom	e	1962.0	NaN	NaN	15.1
102742	Double Chocolate Mini Bites	NaN	M&S	NaN	France,United Kingdom	e	1962.0	NaN	NaN	15.1
114790	Les milanaises	NaN	Le Gaulois	NaN	France	a	506.0	NaN	0.8	0.0
115270	Chateaubriand	NaN	Charal	NaN	France	a	473.0	NaN	NaN	0.8
115271	Chateaubriand	NaN	Charal	NaN	France	a	473.0	NaN	NaN	0.8
115613	Saint Albray	NaN	Saint Albray	NaN	France	e	1331.0	NaN	26.0	18.0
115614	Saint Albray	NaN	Saint Albray	NaN	France	e	1331.0	NaN	26.0	18.0

Exemple de doublons

7/ NETTOYAGE DES DONNÉES

VALEURS NULLES ET ABERRANTES DANS VALUES_100G



Distribution des values_100g

7/ NETTOYAGE DES DONNÉES

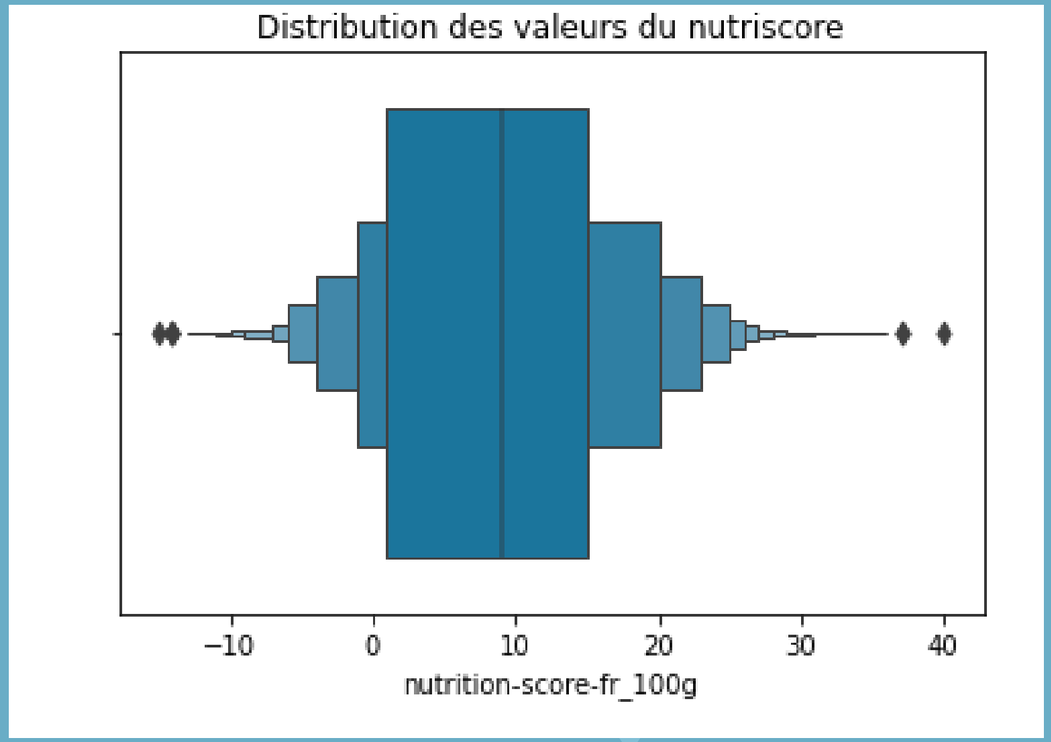
Opération effectuée	Nombre de lignes restantes
Suppression des lignes sans values_100g	97 036 lignes
Suppression des doublons	96 863 lignes
Suppression des produits sans nom	96 453 lignes
Suppression des values_100g non comprises entre 0 et 100	96 433 lignes

7/ NETTOYAGE DES DONNÉES

NUTRISCORE

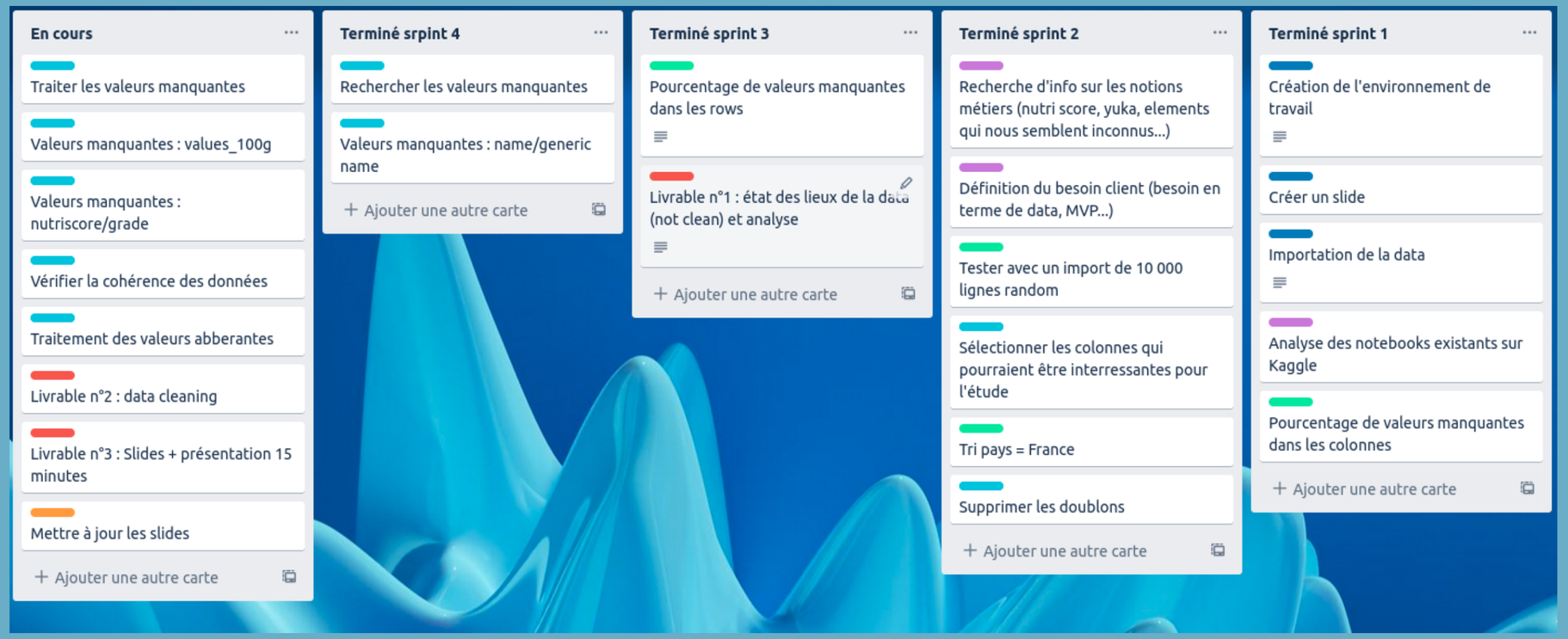
Il manque 4 015 nutriscores dans la base de données.

Pour les données présentes, voici leur distribution :



8/PRESENTATION DES RÉSULTATS

TACHES NON FINIES



A la fin des 4 sprints, la totalité des tâches n'a pas été accomplie

8/PRESENTATION DES RÉSULTATS

La base de données a été sauvegardée (intermediate.csv). Elle contient les lignes pour lesquelles les produits ont un nom, des valeurs comprises entre 0 et 100 dans les colonnes values_100g, et un nutriscore.

92418 LIGNES x 27 COLONNES

Cette base de données pourrait permettre la mise en place du MVP de l'application, c'est à dire l'affichage d'informations sur le nutriscore suite a la recherche de produit via un ou plusieurs mots-clés. Néanmoins, elle pourrait être améliorée.