



# FoodFlix

Étude de faisabilité d'une application basée sur un moteur de recommandation



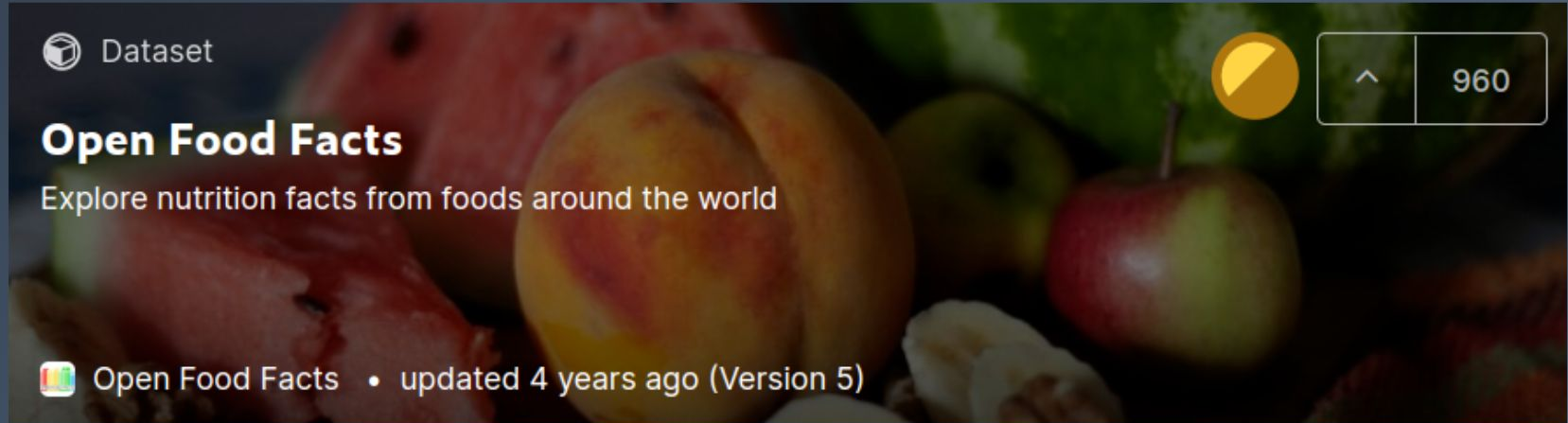
## Contexte

Création d'une application permettant de recommander le meilleur produit à un utilisateur selon un mot clé.

Les éléments à remonter sont les éléments liés au Nutri Score pour le MVP.

# Analyse de la donnée

Données utilisées :



Dataset

## Open Food Facts

Explore nutrition facts from foods around the world

Open Food Facts • updated 4 years ago (Version 5)

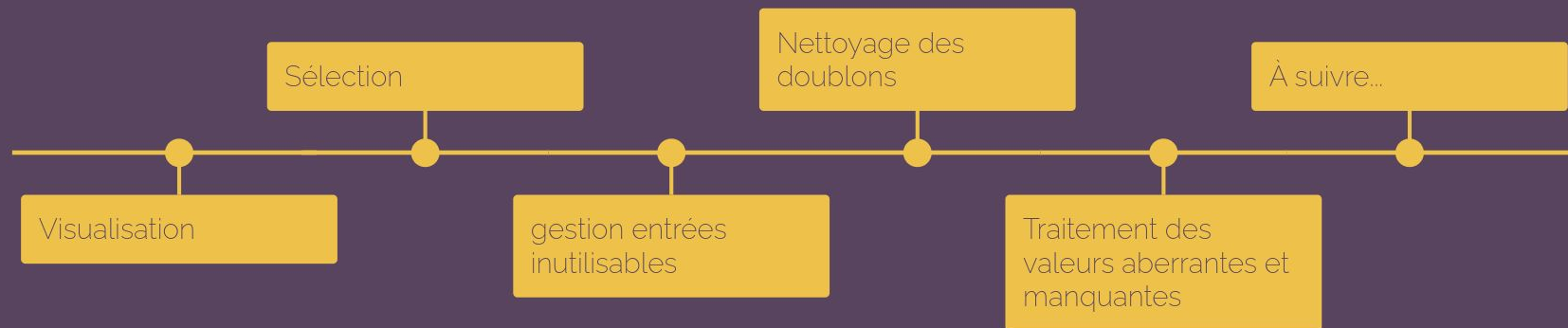
960

Taille de la table : 356027 lignes et 163 colonnes.

<https://www.kaggle.com/openfoodfacts/world-food-facts>

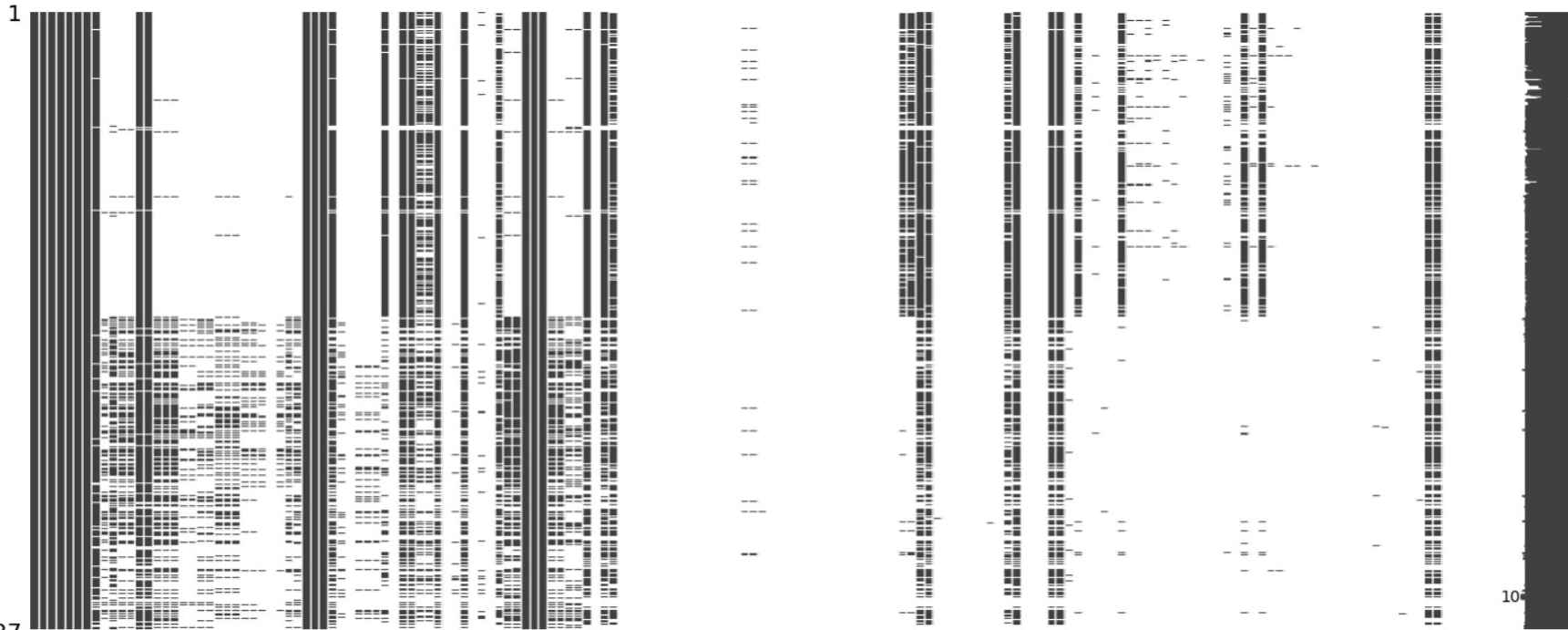
# Travail de la data

## Les différentes étapes du processus

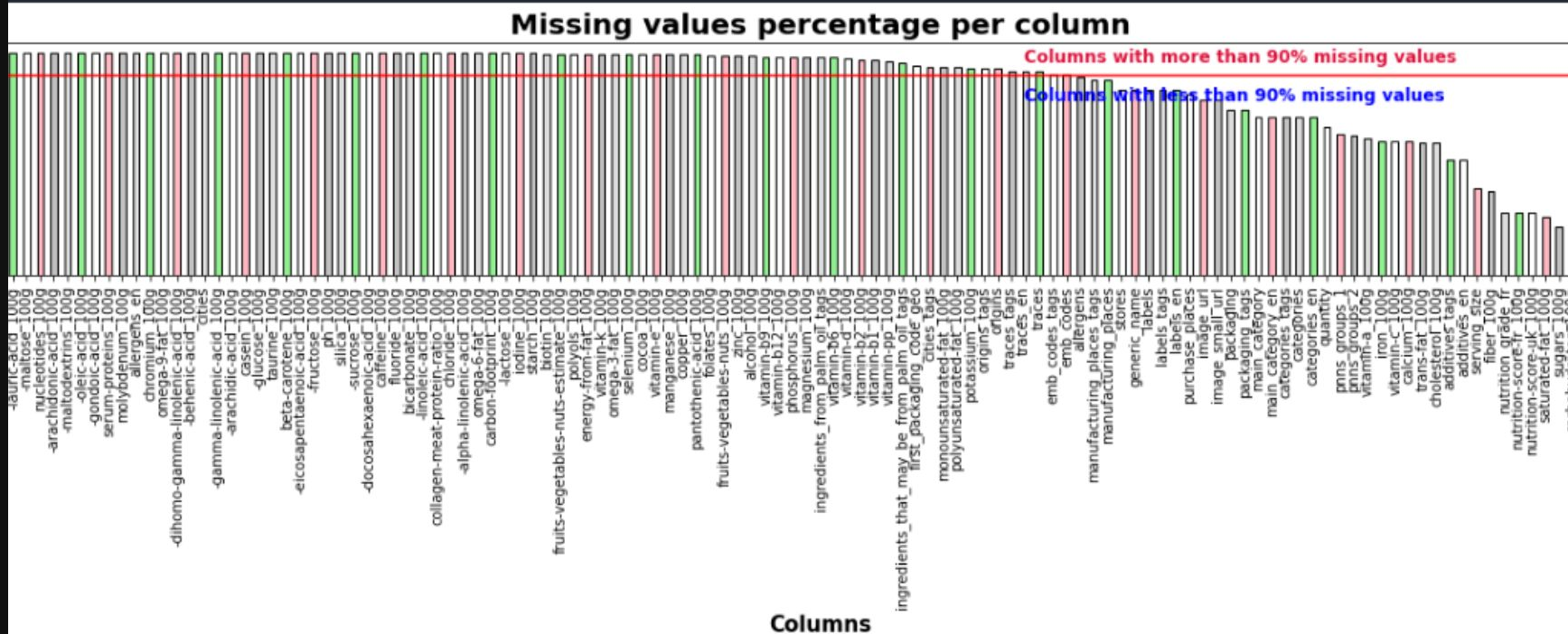


# Visualisation de la data

MissingNo :



## Barplot





# Sélection des données

## Choix des lignes

Nous utiliserons seulement les lignes concernant les produits qui sont vendus en France .

Les produits dont le nom n'est pas renseigné sont supprimés. De même pour ceux qui n'ont pas de nutrition score.



# Sélection des données

## Choix des colonnes

Une grande partie des colonnes ont peu de valeurs renseignés, Nous sélectionnerons seulement les colonnes bien renseignées avec les données les plus pertinentes





# Sélection des données

Choix des colonnes

## Choix des colonnes : Dénomination du produit

colonnes	product_name	brands	categories	ingredients_text
description	nom du produit	marque du produit	catégories du produit	liste des ingrédients
type	text	text	text	text

## Choix des colonnes valeurs énergétiques et allergènes :

colonnes	allergens	nutrition_grade_fr	nutrition-score-fr_100g	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g			proteins_100g	salt_100g
Description	liste des allergènes	nutri-score (A à E)	nutri-score	énergie	graisses	graisses saturées	glucides	sucre	fibres			protéines	sel
Type	texte	texte	num	num	num	num	num	num	num			valeur numérique	valeur numérique
unité de mesure	-----	lettre A-E	-15 à 40	Kj /100g	g /100g	g /100g	g /100g	g /100g	g /100g			g /100g	g /100g



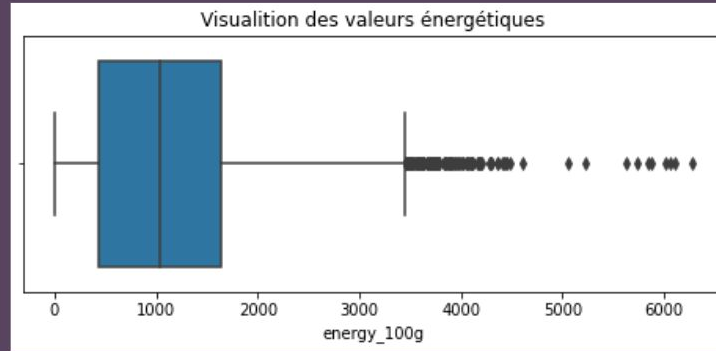
## Nettoyage de la donnée

### Recherche et suppression de doublons

Dans la configuration actuelle,  
le dataset contient des  
doublons que l'on supprime

# Nettoyage de la donnée

## Recherche de données aberrantes



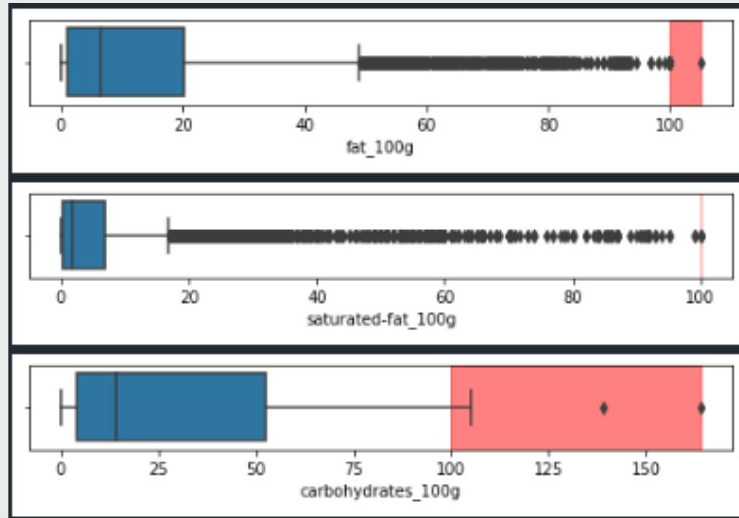
valeurs énergétiques :

*La valeur énergétique maximale que l'on peut atteindre est d'environ 3700 Kj /100g.*

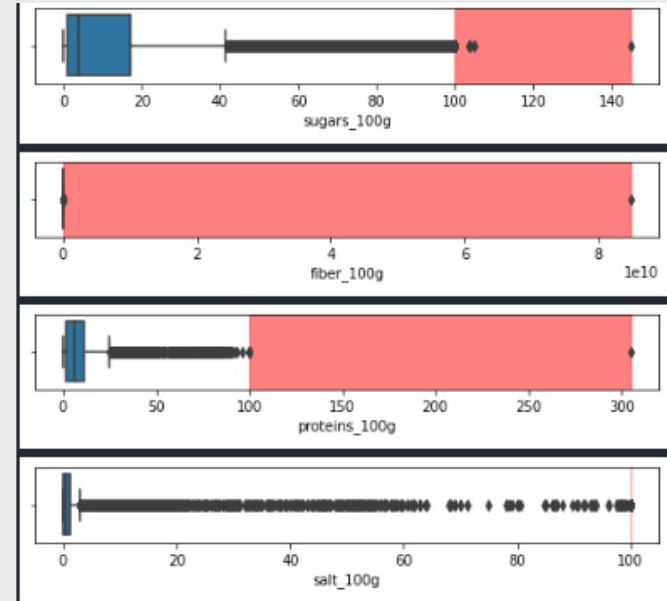
*Les données supérieures à cela sont supprimer.*

# Nettoyage de la donnée

## Recherche de données aberrantes



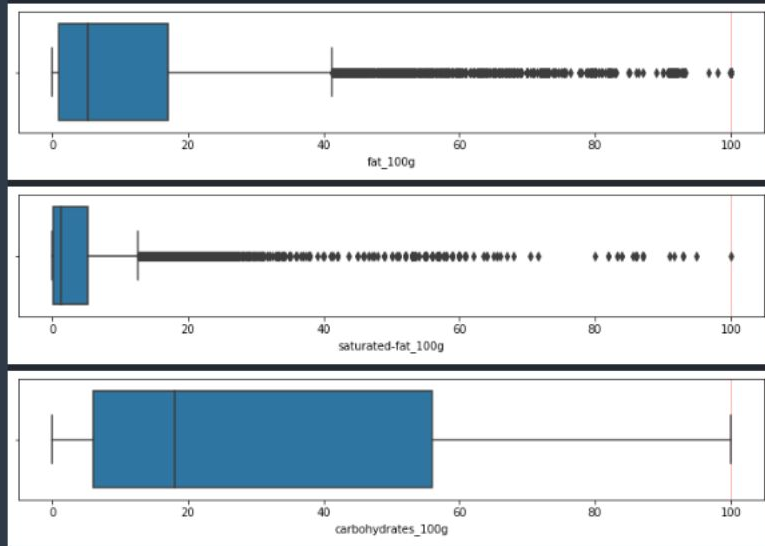
Visualisations des données concernant la composition.



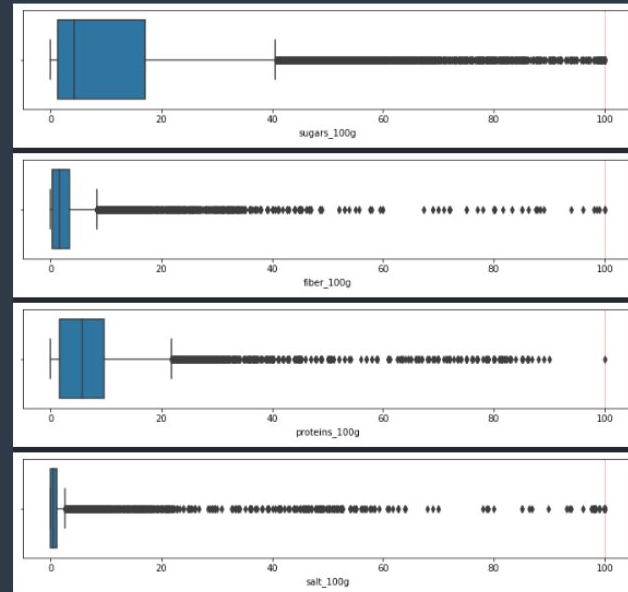
Les valeurs supérieures à 100g/100g sont a retirés

# Nettoyage de la donnée

Après suppression des valeurs aberrantes



Les données sont à présent cohérentes



Les données sont à présent cohérentes



## Traitement des valeurs manquantes

Les données non présentes sont remplacé par 0 ou "Non renseigné" suivant leurs types.



---

## État de la donnée

Avant nettoyage :

356027  
Lignes

163  
colonnes

---

## État de la donnée

Après nettoyage :

43907  
lignes

15  
colonnes