Rendre les données de santé publique plus accessibles

Analyse du jeu de données de Open Food Facts

Par Tyson JOHN

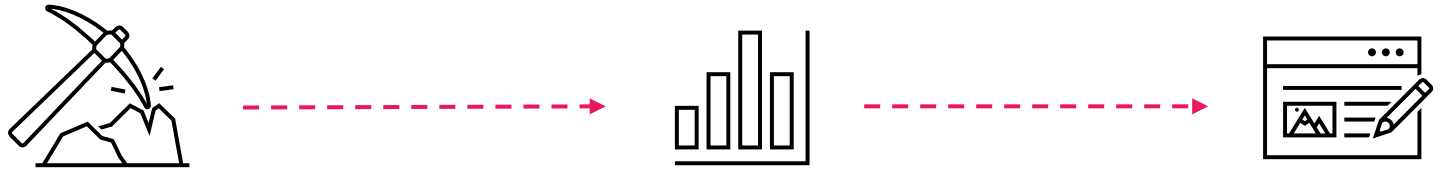
Sommaire

1. Présentation de l’appel à projet
2. Nettoyage des données
3. Exploration des données
4. Démonstration du prototype

1 PRÉSENTATION DE l'APPEL À PROJET

Objectifs du projet

Rendre les données de santé publique plus accessible pour les agents de Santé publique France



Nettoyage
Exploration

Analyse
Visualisations

Prototype
Page web

Le jeu de données

Open Food Facts

- Projet collaboratif
- Financé par Santé publique France
- 1843238 produits alimentaires
- 186 colonne

Informations générales

- Nom
- Code bar
- Date de modification

Informations spécifiques

- Catégorie
- Pays de vente
- Etc...

Informations nutritionnelles

- Energie en kJ
- Matières grasses
- Etc...

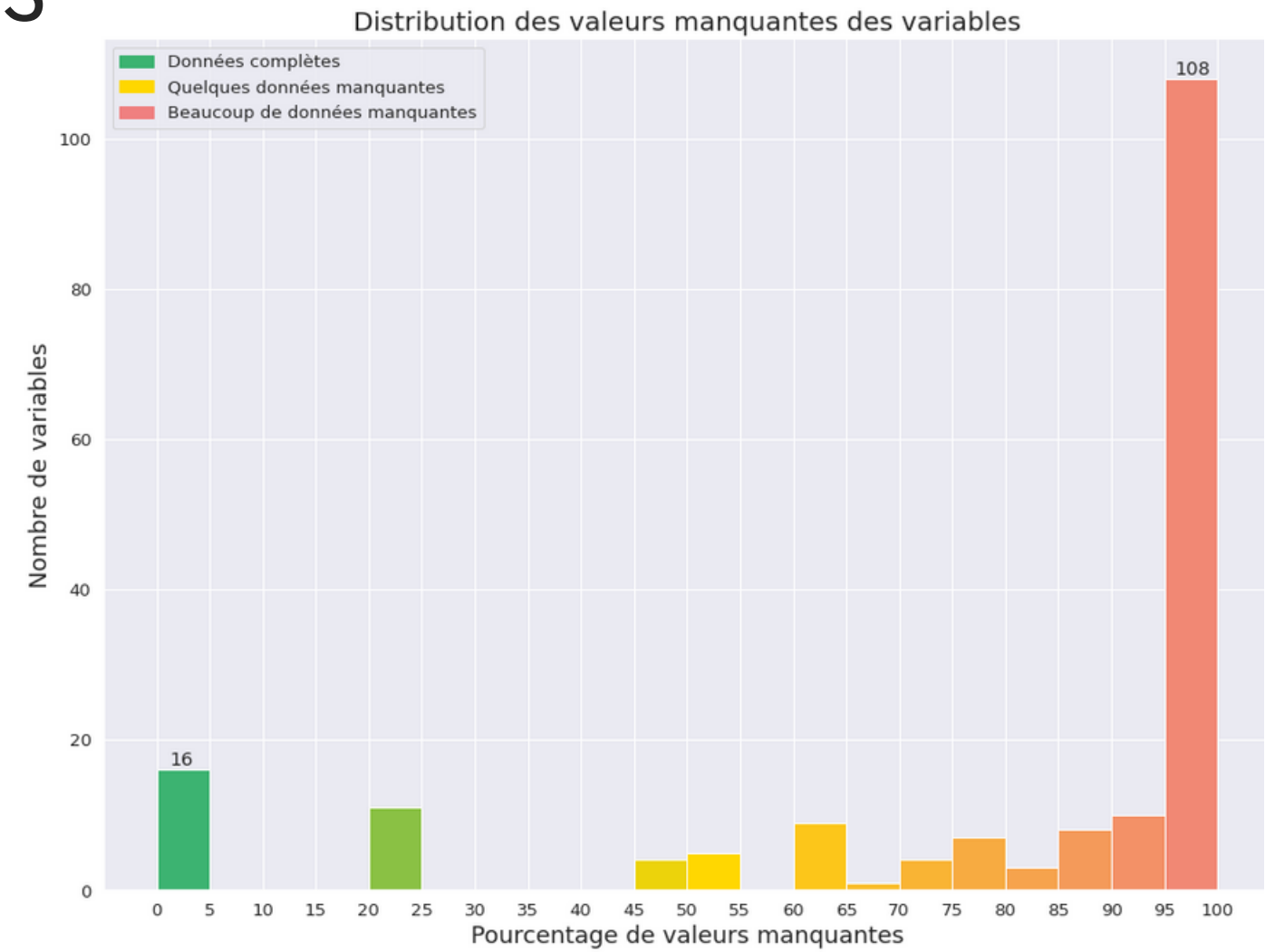
Ingrédients

- Liste des ingrédients
- Liste des additifs
- Liste des allergènes

Valeurs manquantes

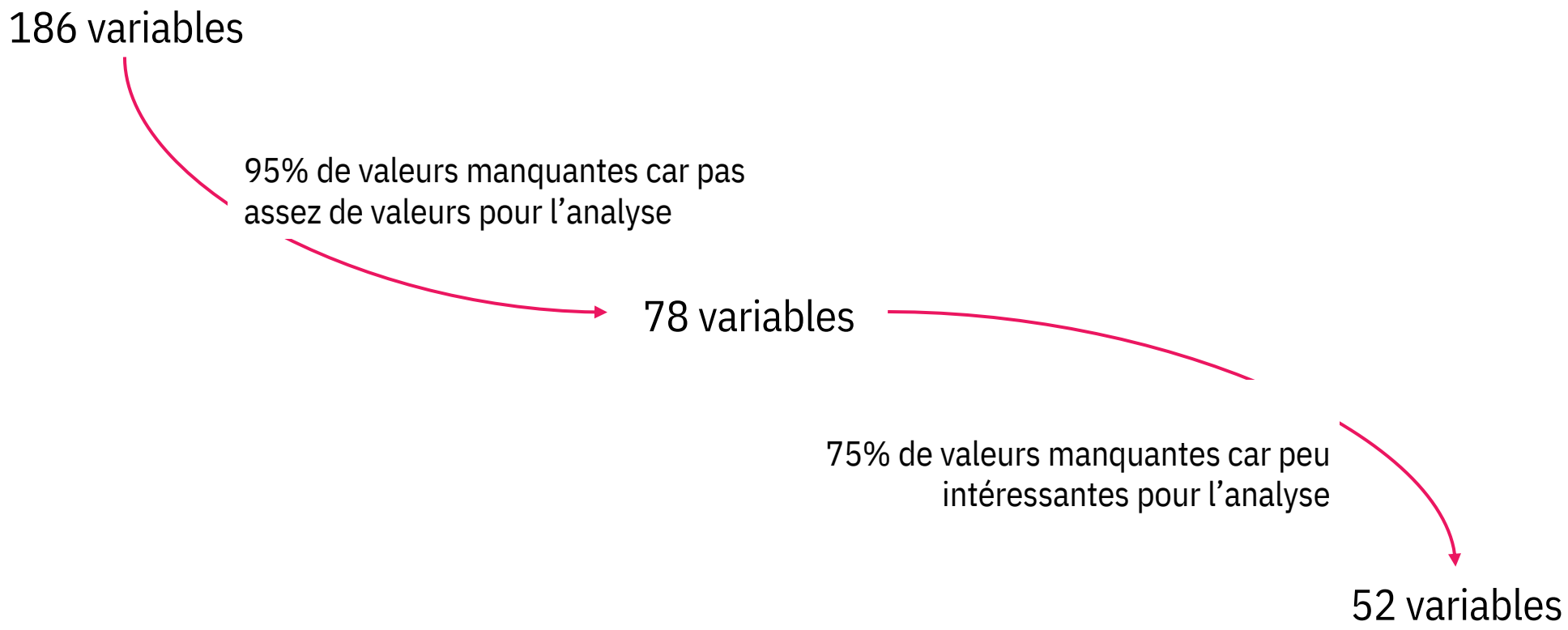
58% des colonnes ont entre 95% et 100% de valeurs manquantes

8,6% des colonnes non pas des valeurs manquantes



2 NETTOYAGE DES DONNÉES

Suppression de variables ayant trop de valeurs manquantes



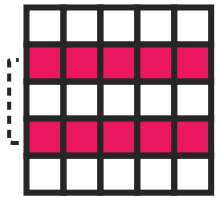
Suppression de produits ayant trop de valeurs manquantes

24% des produits n’ont aucune valeur dans leur tableau nutritionnel

index	product_name	energy_100g	fat_100g	carbohydrates_100g	proteins_100g	salt_100g
1	Cacao	NaN	NaN	NaN	NaN	NaN
7	Pistou d'ail des ours	NaN	NaN	NaN	NaN	NaN
8	Pain maïs	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...
1555548	Marrons glacés	NaN	NaN	NaN	NaN	NaN
1555550	Sandwich club Rillette poisson combava	NaN	NaN	NaN	NaN	NaN
1555551	Thé noir BIO Darjeeling	NaN	NaN	NaN	NaN	NaN

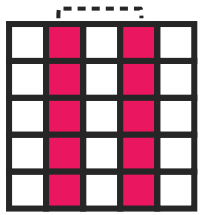
Exemples de produits n’ayant aucune valeur dans leur tableau nutritionnel

Doublons



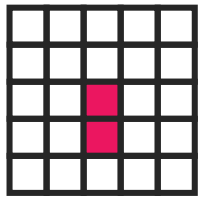
Doublons dans les individus

- Produits ayant le même code bar



Doublons dans les variables

- Exemple : labels, labels_tags, labels_fr

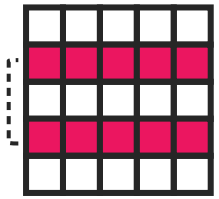


Doublons dans les modalités

- Exemple : huile, Huile, oiletc...

Suppression des doublons et des variables inutiles

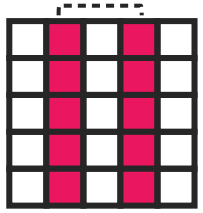
1843238 produits



Suppression de
260 produits

1842978 produits

52 variables



Suppression de
31 variables

22 variables

Nettoyage des valeurs du tableau nutritionnel

24% des produits ont leur tableau nutritionnel complet

38 % des produits ont leur tableau nutritionnel complet

Correction de l'énergie

- Énergie aux 100g ≤ 3700kJ
- Correction des valeurs atypiques
- Suppression des tableaux nutritionnels aberrants

- Energie = f(Nutriments)
- Possible si 1 seule valeur manquante

- Nutriments aux 100g ≤ 100g
- Correction des valeurs atypiques
- Suppression des tableaux nutritionnels aberrants

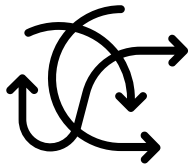
Mise à 0 des valeurs manquantes

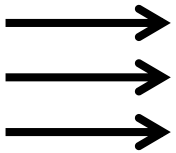
- Diff énergie = Energie -f(Nutriments présents)
- Mise à 0 des valeurs manquantes sur Diff énergie < 5%

Nettoyage à partir de données métier

Récupération des tableaux de 14% des produits

Featureengineering





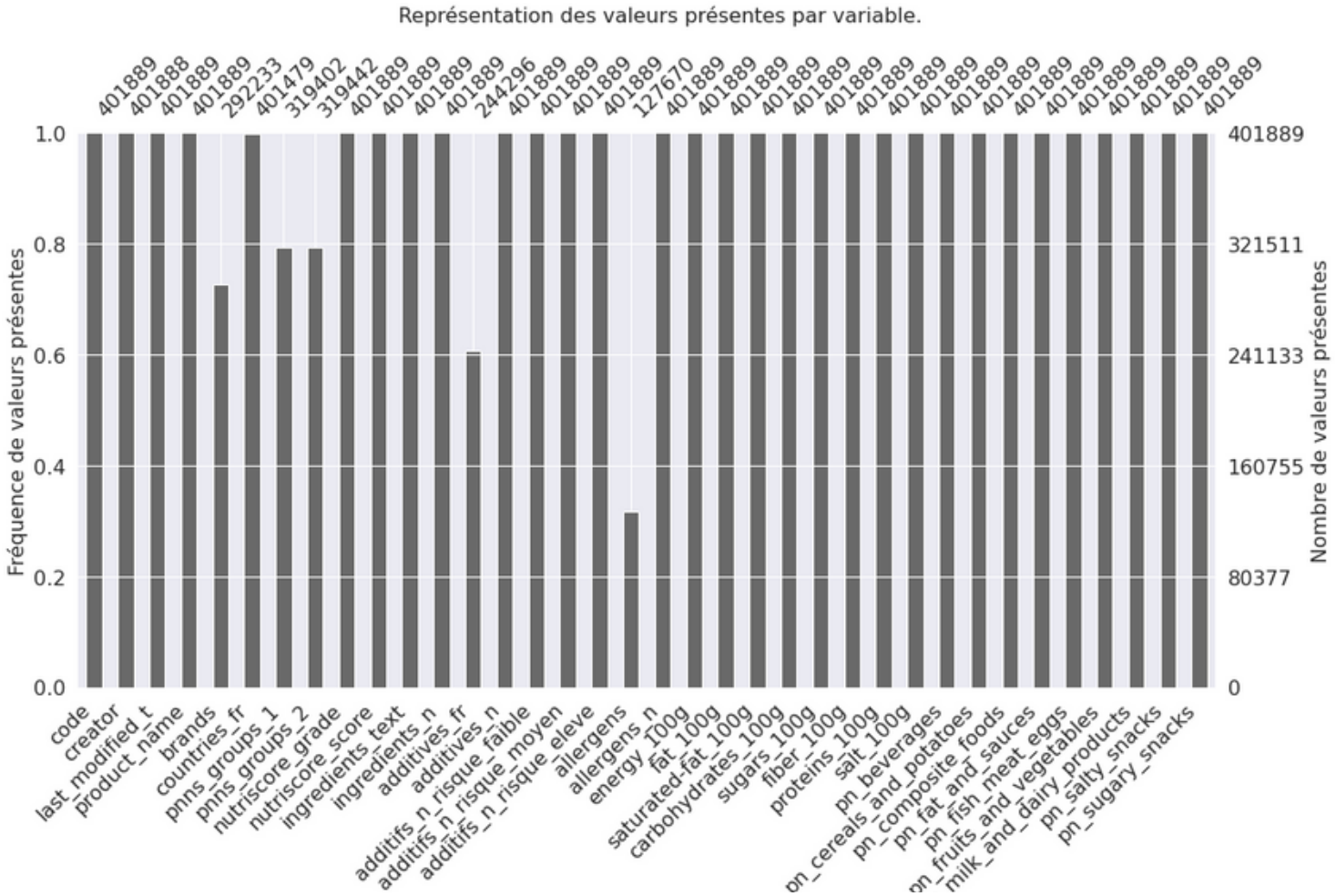
Variables au format texte difficiles à nettoyer

- Liste des ingrédients
- Liste des additifs
- Nom du produit
- Etc…

Création de nouvelles variables plus simples à exploiter

- Nombre d'ingrédients
- Nombre d’additifs
- Target encodingdu nom du produit
- Etc…

Jeu de données nettoyé



401889 produits
- 22% du jeu de données initial
- Tableaux nutritionnels complets
- Listes des ingrédients complètes

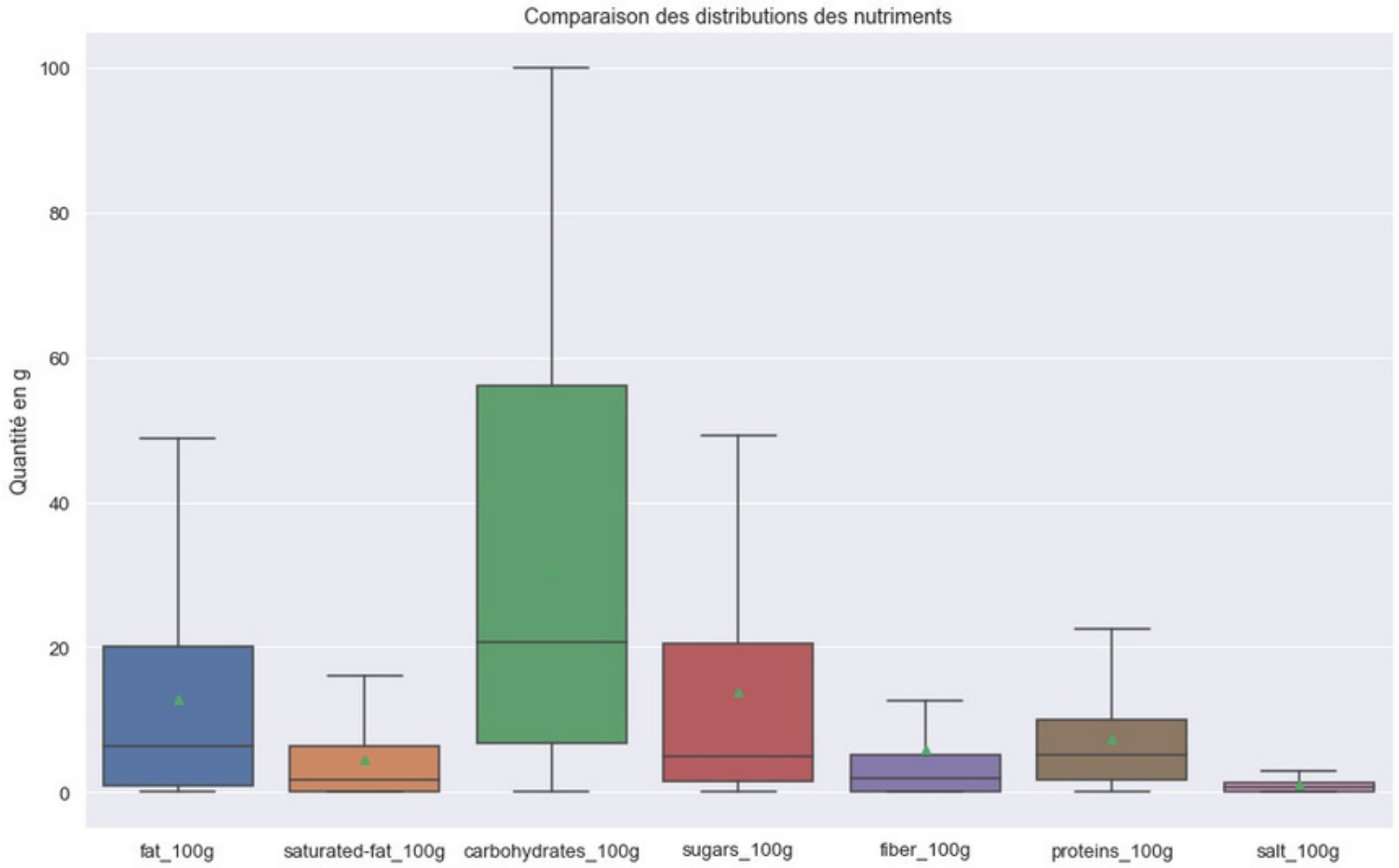
37 variables
- 19% du jeu de données initial
- 25 variables de type numérique
- 11 variables de type catégorie

3 EXPLORATION DES DONNÉES

Analyse des variables de type numérique

	moyenne	écart type	min	max	Q1	Q2	Q3	valeurs manquantes	valeurs nulles	skewness
index	584934.612154	494206.362869	5.0	1.843229e+06	187145.000000	420083.000000	909391.000000	0	0	0.841995
nutriscore_score	8.049128	8.885934	-15.0	3.600000e+01	0.000000	8.000000	15.000000	0	20713	0.156464
ingredients_n	13.883224	13.643103	1.0	2.310000e+02	4.000000	10.000000	19.000000	0	0	2.301901
additives_n	2.230439	3.125938	0.0	3.900000e+01	0.000000	1.000000	3.000000	0	157593	2.389942
additifs_n_risque_faible	0.014925	0.138530	0.0	3.000000e+00	0.000000	0.000000	0.000000	0	396751	10.590862
additifs_n_risque_moyen	0.014161	0.131205	0.0	5.000000e+00	0.000000	0.000000	0.000000	0	396756	11.100209
additifs_n_risque_eleve	0.060965	0.335406	0.0	8.000000e+00	0.000000	0.000000	0.000000	0	385112	7.317233
allergens_n	0.559677	1.029222	0.0	3.300000e+01	0.000000	0.000000	1.000000	0	274219	2.524146
energy_100g	1111.875029	767.628976	0.0	3.700000e+03	397.000000	1046.000000	1674.000000	0	4453	0.452541
fat_100g	12.660790	16.218481	0.0	1.000000e+02	0.800000	6.220000	20.000000	0	69341	2.112941
saturated-fat_100g	4.491460	7.137110	0.0	1.000000e+02	0.000000	1.600000	6.250000	0	114570	3.713281
carbohydrates_100g	30.879163	27.304413	0.0	1.000000e+02	6.670000	20.240000	55.799999	0	21615	0.588123
sugars_100g	13.717916	18.277540	0.0	1.000000e+02	1.400000	4.850000	20.000000	0	57043	1.780178
fiber_100g	6.123577	13.118501	0.0	1.000000e+02	0.000000	1.900000	5.000000	0	114883	3.905283
proteins_100g	7.345864	7.961168	0.0	1.000000e+02	1.670000	5.100000	10.000000	0	48251	2.336385
salt_100g	1.084843	3.533397	0.0	1.000000e+02	0.100000	0.580000	1.235000	0	45781	17.267034
pn_beverages	0.339149	0.612385	0.0	1.009812e+01	0.011733	0.109920	0.364996	0	39001	3.664738
pn_cereals_and_potatoes	0.511247	0.722827	0.0	1.071805e+01	0.075143	0.237941	0.641812	0	18171	2.943100
pn_composite_foods	0.364571	0.551157	0.0	1.161438e+01	0.050031	0.167273	0.434157	0	22561	3.503760
pn_fat_and_sauces	0.391602	0.619667	0.0	9.556623e+00	0.034483	0.145185	0.455655	0	25973	2.999026
pn_fish_meat_eggs	0.353062	0.649834	0.0	1.429508e+01	0.023063	0.115155	0.335678	0	30270	3.563633
pn_fruits_and_vegetables	0.293743	0.514580	0.0	1.219296e+01	0.023240	0.105527	0.324695	0	28202	3.978404
pn_milk_and_dairy_products	0.497421	0.807934	0.0	2.162225e+01	0.044367	0.178377	0.551107	0	25830	3.159434

Comparaison des distributions des nutriments



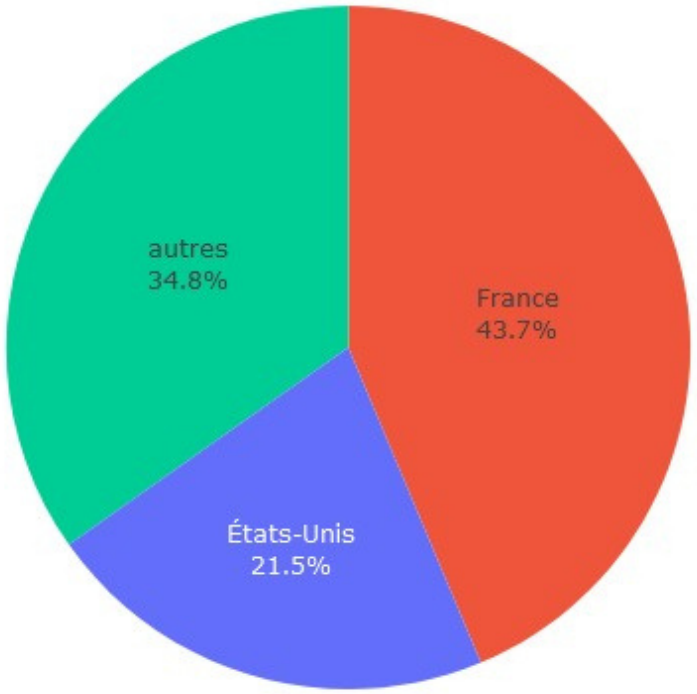
En moyenne, les produits contiennent principalement :

- Des glucides
- Des matières grasses

Analyse des variables de type catégorie

	nombre de valeurs	nombre de modalités	mode	effectif du mode	fréquence du mode	valeurs manquantes
additives_fr	244296	60654	E330 - Acide citrique	16567	0.0678153	157593
allergens	127670	4473	en:milk	28493	0.223177	274219
brands	292233	61827	Carrefour	5258	0.0179925	109656
countries_fr	401479	2347	États-Unis	227447	0.566523	410
creator	401888	6277	usda-ndb-import	117821	0.293169	1
ingredients_text	401889	333897	Almonds.	373	0.000928117	0
nutriscore_grade	401889	5	d	114628	0.285223	0
pnns_groups_1	319402	10	Sugary snacks	67460	0.211207	82487
pnns_groups_2	319442	38	Biscuits and cakes	34499	0.107998	82447
product_name	401889	296810	Ice cream	701	0.00174426	0

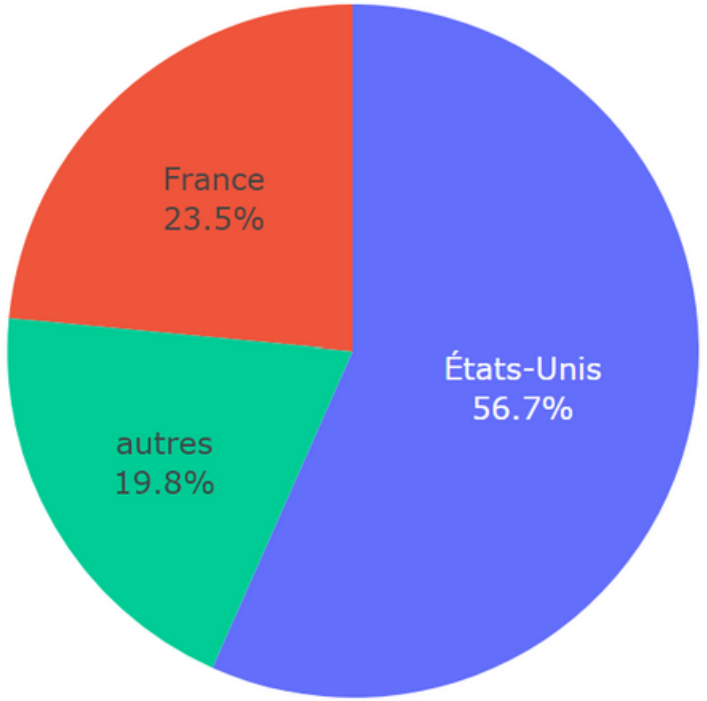
Distribution des pays de vente



Distribution de la variable countries_fr
avant nettoyage des données

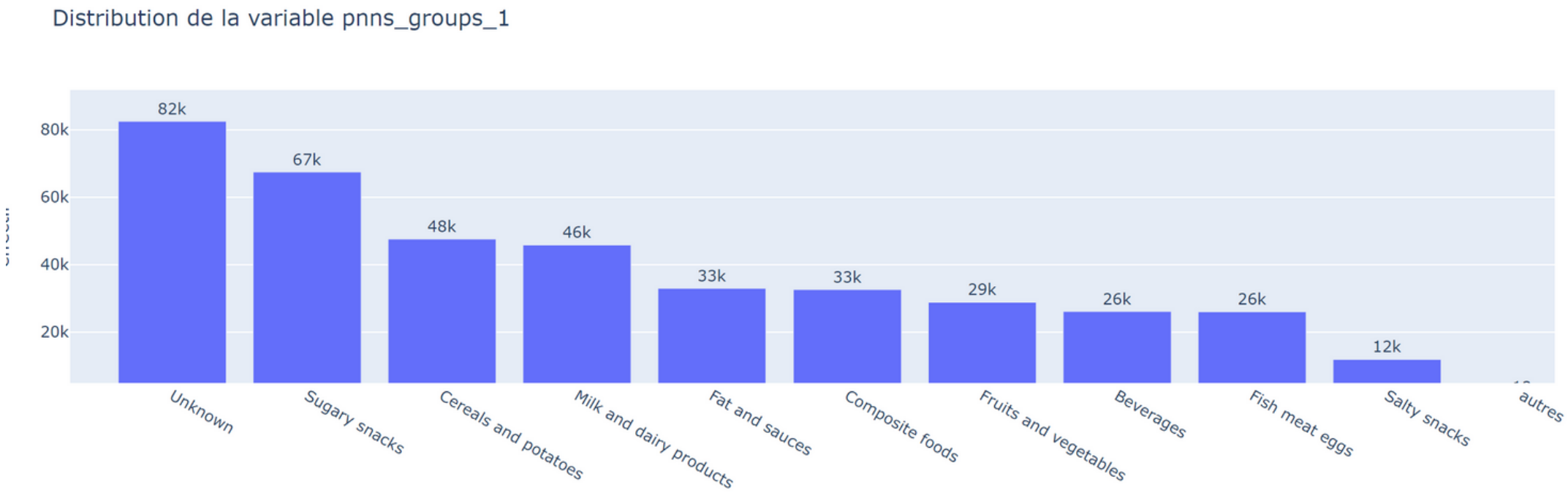
Nettoyage des données

Beaucoup de produits vendus en
France ont des données manquantes



Distribution de la variable countries_fr
après nettoyage des données

Distribution des catégories des produits



82000 produits dans les snacks sucrés
17% des produits

12000 produits dans les snacks salés
Représentatif de sa population

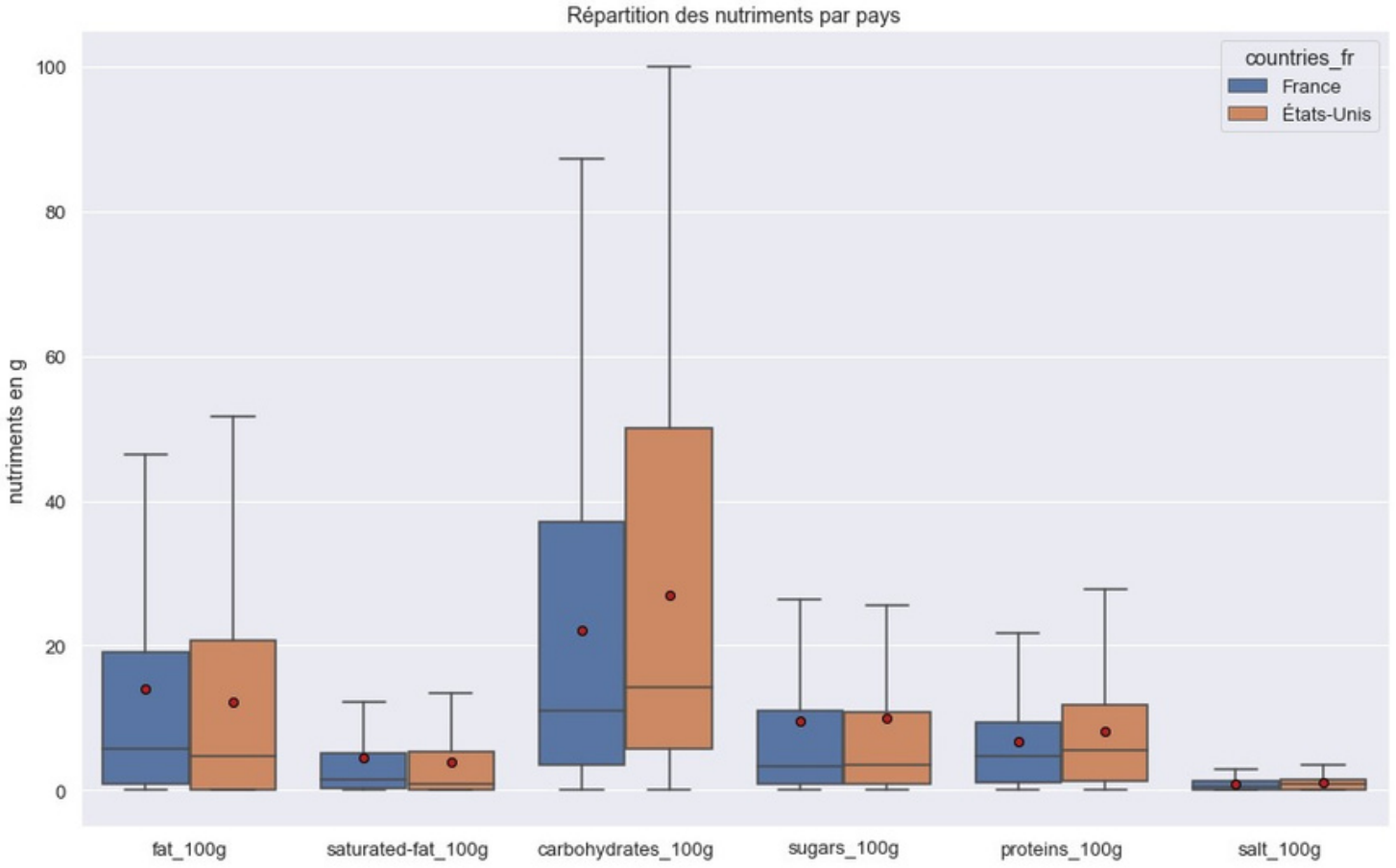
Comparaisons entre la France et les US

Produits vendus en France ajoutés à partir de **2012**

Produits vendus aux US ajoutés à partir de **2015**

Sélection d'un sous-échantillon des produits ajoutés à partir de **2015**

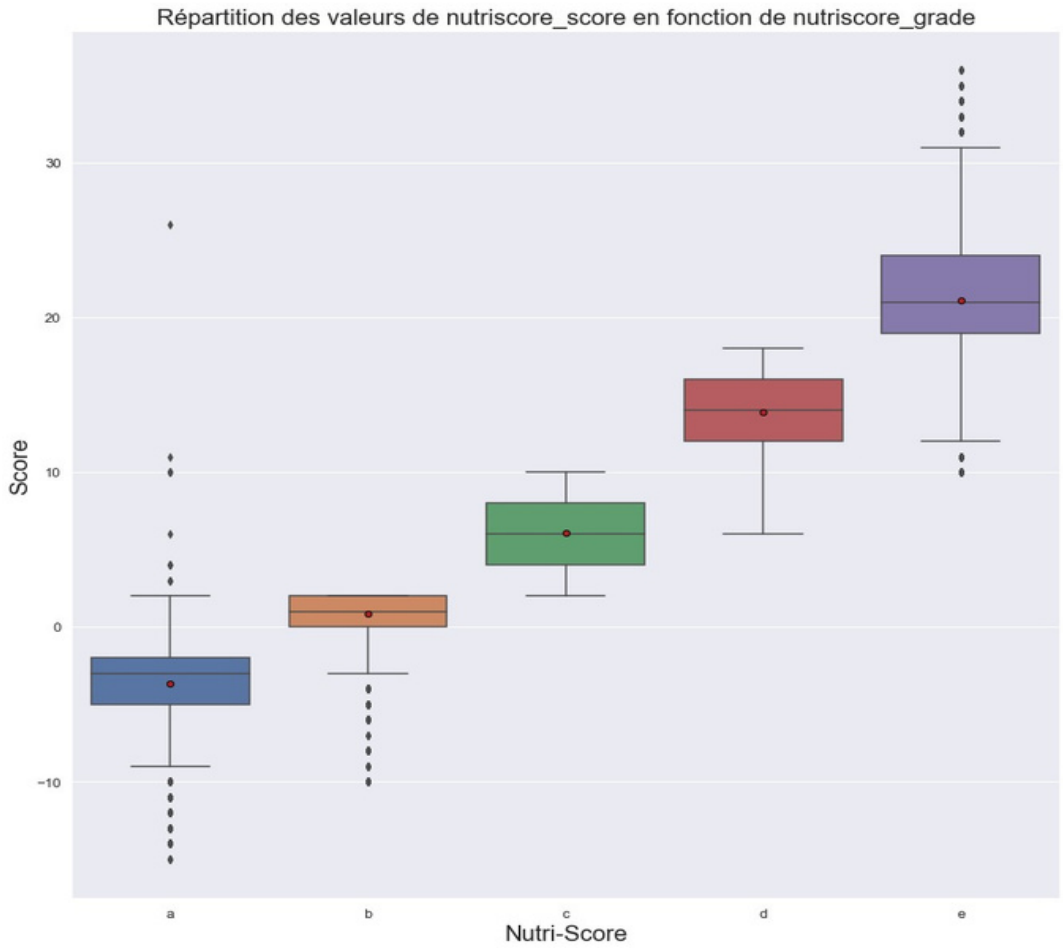
Comparaisons entre la France et les US



Plus de glucides aux US.

Répartition des autres nutriments similaire.

Corrélation entre le Nutri-Score et le score de nutrition

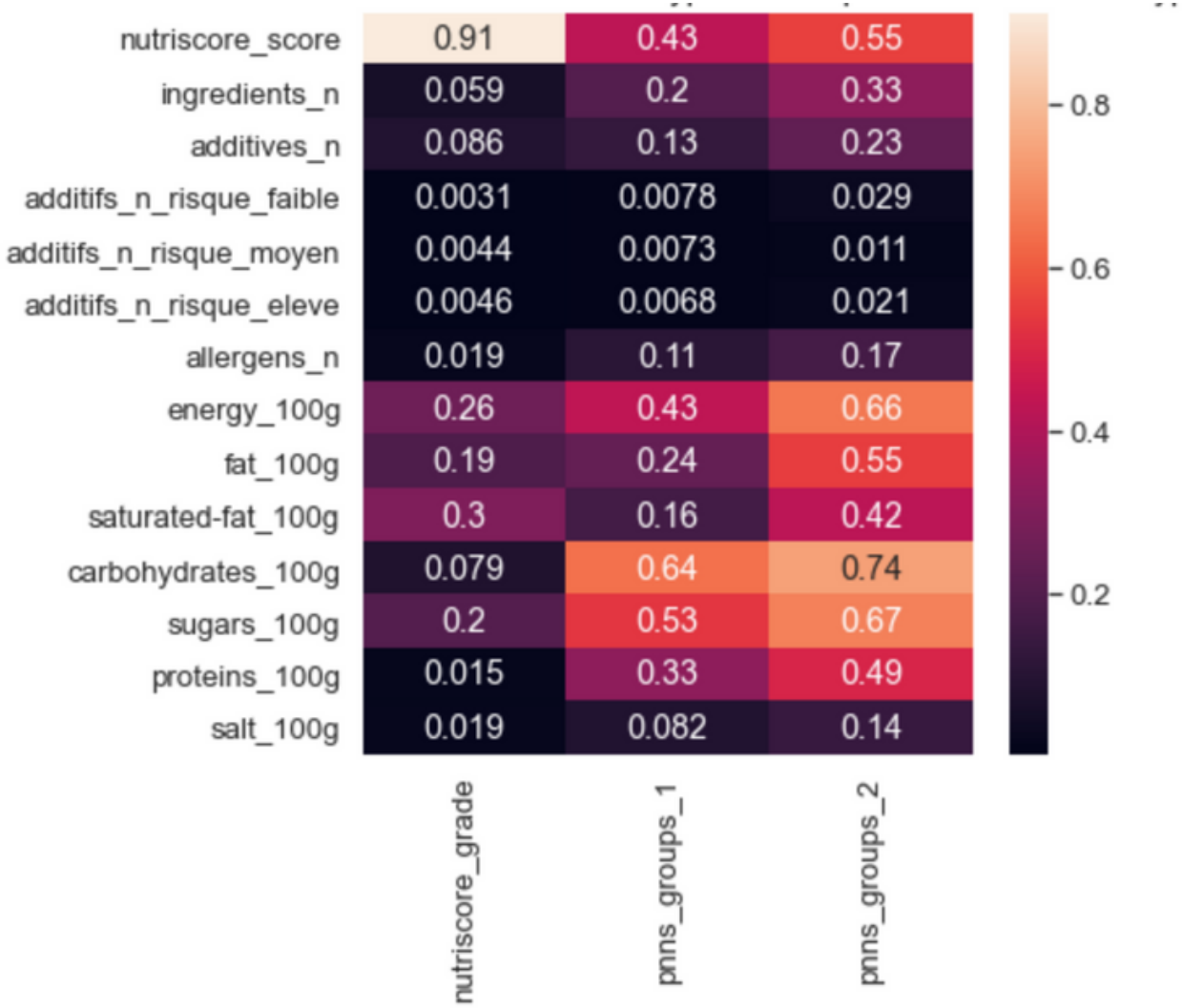


Forte corrélation linéaire entre le Nutri-Score et le score de nutrition

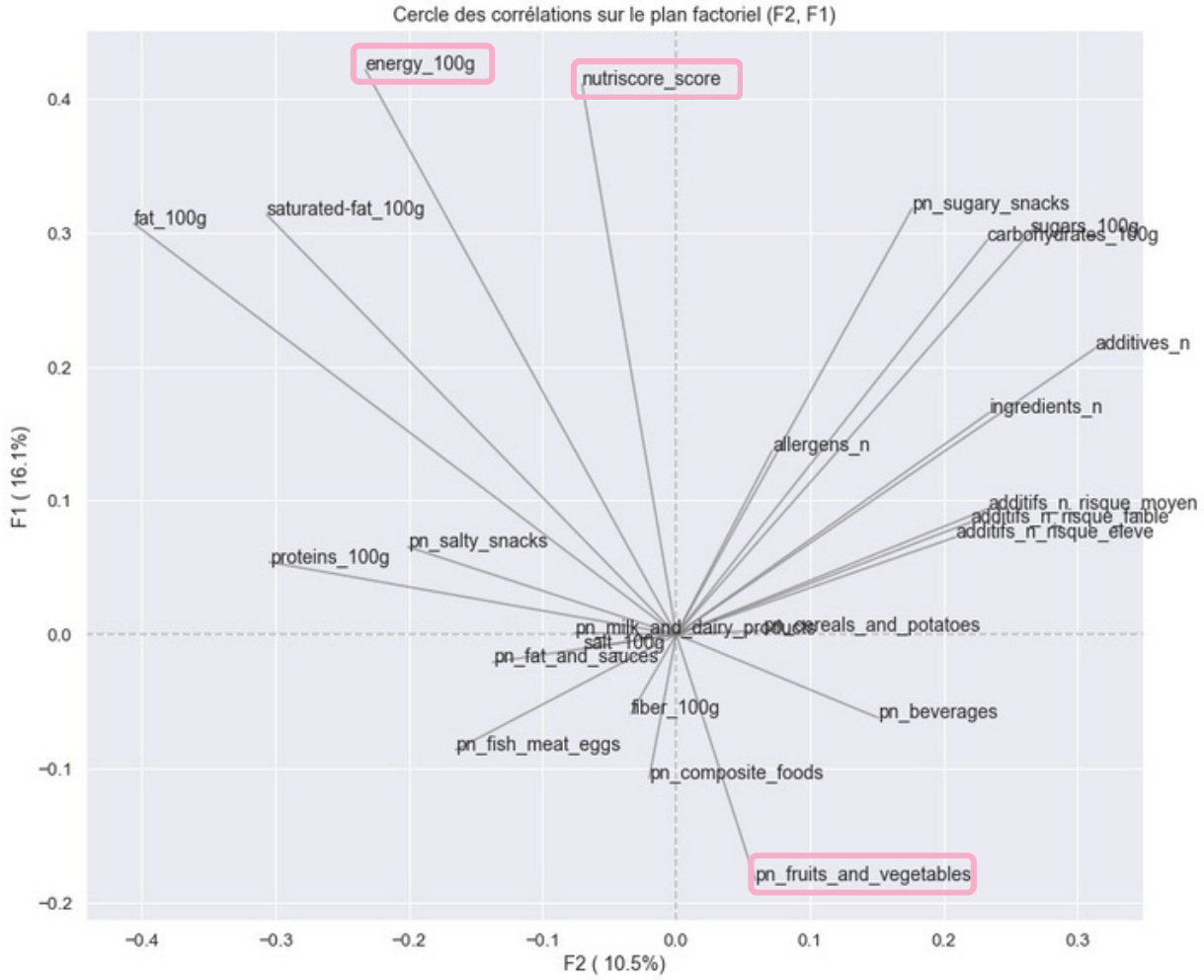
Présence d'outliers
Exemple du Nutri-Score A :

- Les outlierssont principalement des eaux minérales

Analysis of variance



Analyse en Composantes Principales



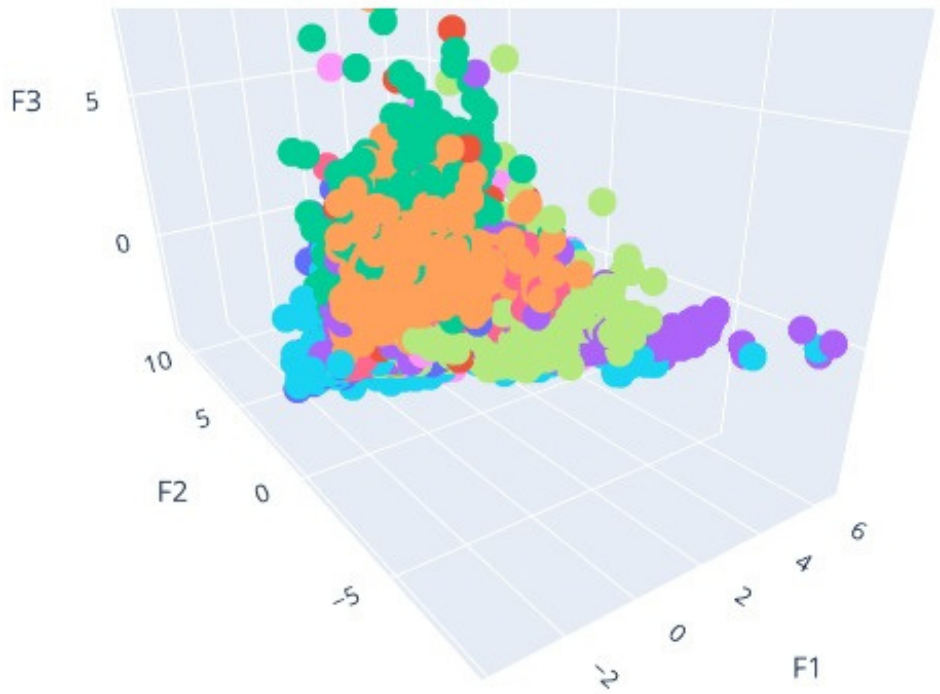
ACP avec les 4 premiers axes d'inertie :

- F1, F2, F3, F4.

Exemple d'interprétation de l'axe d'inertie F1 :

- Augmente avec l'énergie aux 100g et le score de nutrition.
- Diminue un peu si il y a un lien entre le nom du produit et la catégorie des fruits et légumes.
- Pourrait représenter la notion de richesse énergétique d'un produit en termes de matières grasses et de sucres.

Analyse en Composantes Principales



Représentation d'un échantillon des individus projetés sur (F1, F2, F3)

On peut discerner la plupart des catégories de produits.

Produits extrêmes sur F1 :
•Fruit ou légume pour F1 < 0
•Snack sucré pour F1 > 0

Les axes d'inerties peuvent par exemple êtres utilisés pour filtrer des produits selon les notions qu'ils représentent.

4 DÉMONSTRATION DU PROTOTYPE

Introduction

Introduction

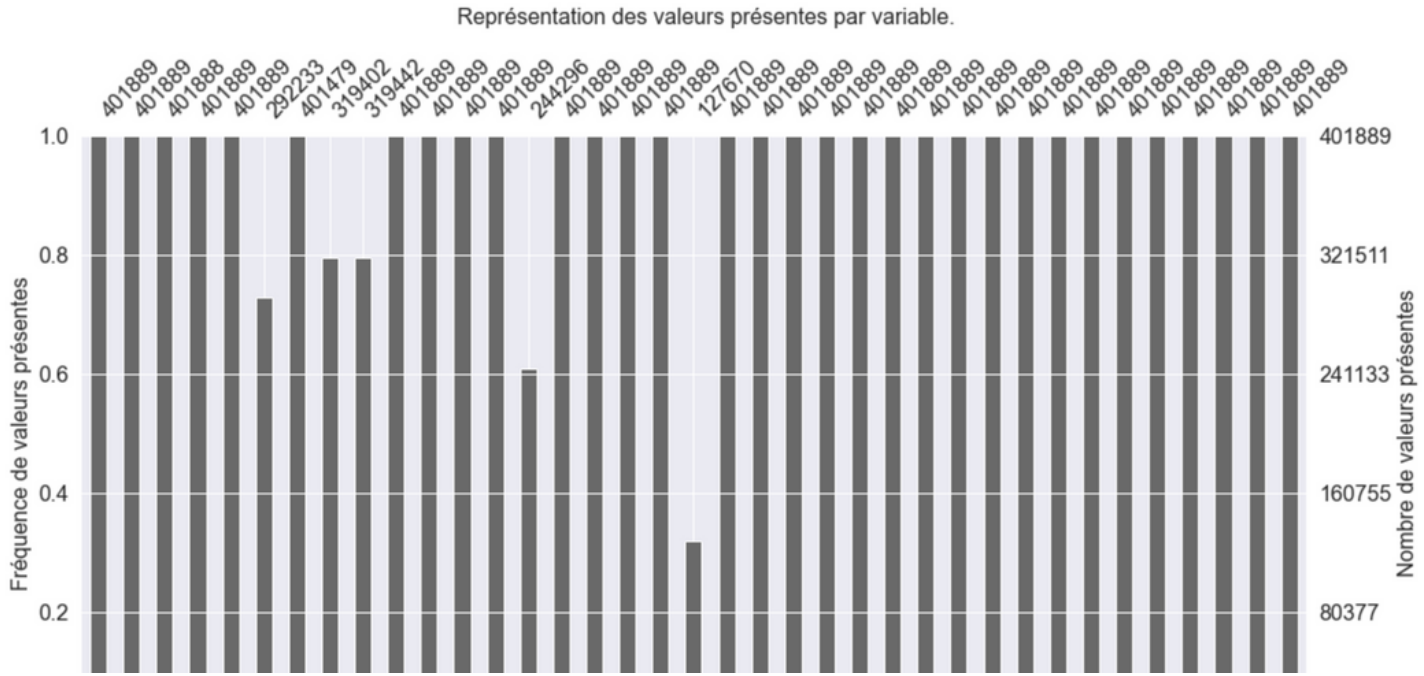
Nous allons analyser un jeu de données fourni gratuitement par Open Food Facts. Il s'agit d'un projet collaboratif qui vise à collecter des données sur des produits alimentaires que l'on retr
projet a été initié en France mais il est depuis devenu international.

Le projet possède une page wiki avec des nombreuses informations sur son jeu de données : wiki.

Cette page web présente l'exploration et l'analyse du jeu de données nettoyé. On retrouvera 4 parties dans le menu :

- **Introduction** : ce que vous êtes actellement en train de lire.
- **Exploration des produits** : outil d'exploration visuelle des produits sous format d'un dashboard.
- **Analyse univariée** : analyse de chaque variable une à une.
- **Analyse multivariée** : analyse des liens entre les variables.

Voici un graphique résumant le jeu données avec le nombre de valeurs présentes par variable :



Exploration du jeu de données

Nombre d'additifs :

0-39

Nombre d'allergène :

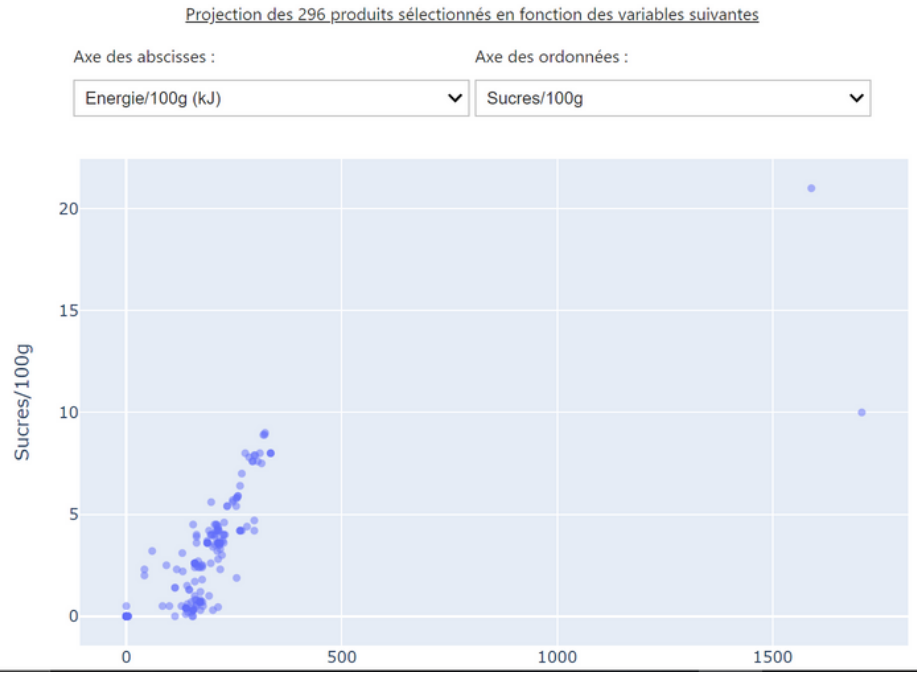
0-33

Énergie aux 100g :

0-3700

Section

Visualisation graphiquem



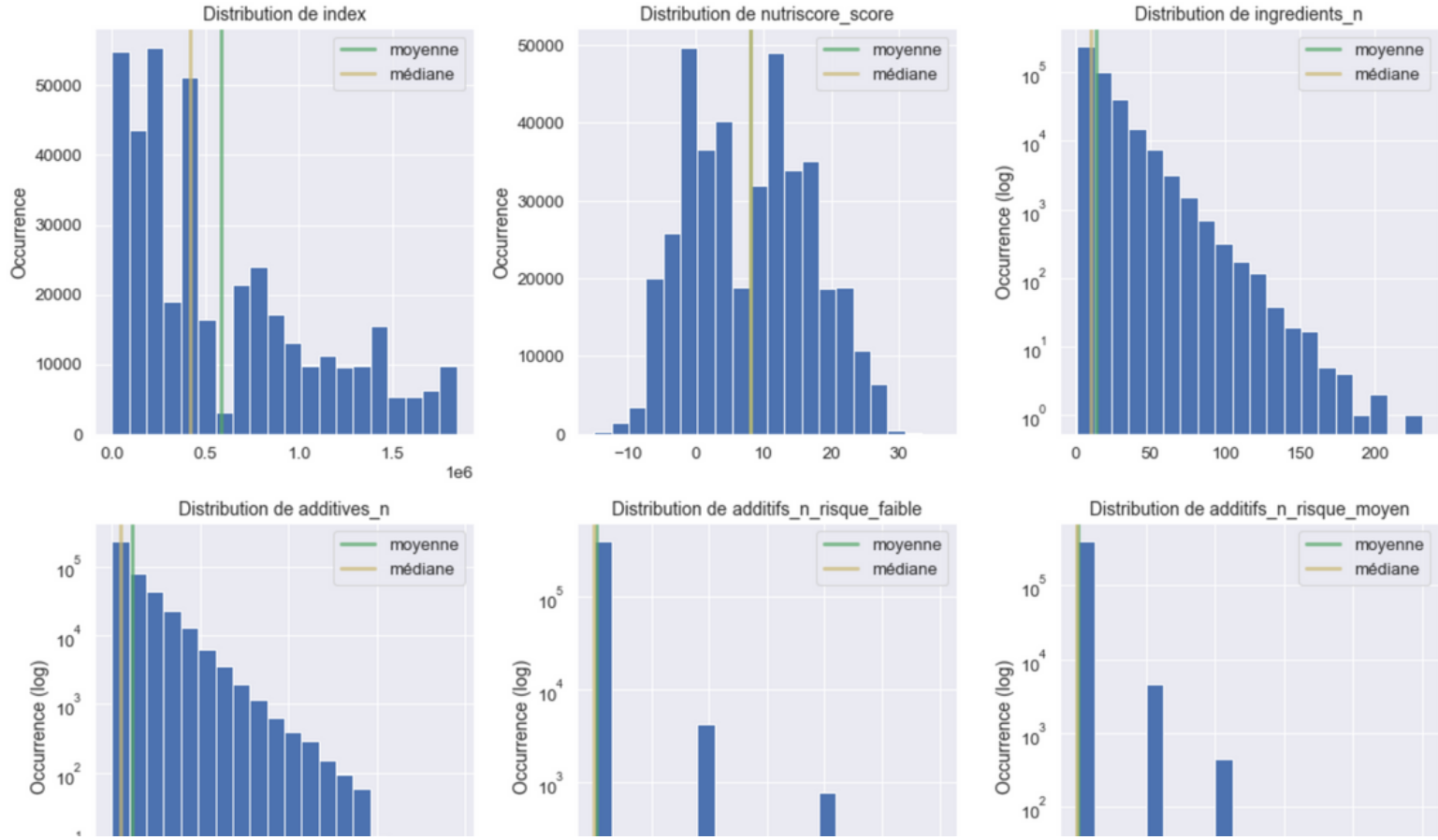
Analyse univariée

Analyse univariée

Analyse des variables de type numérique

Distribution des variables

Commençons par observer les distributions des variables de type numérique :

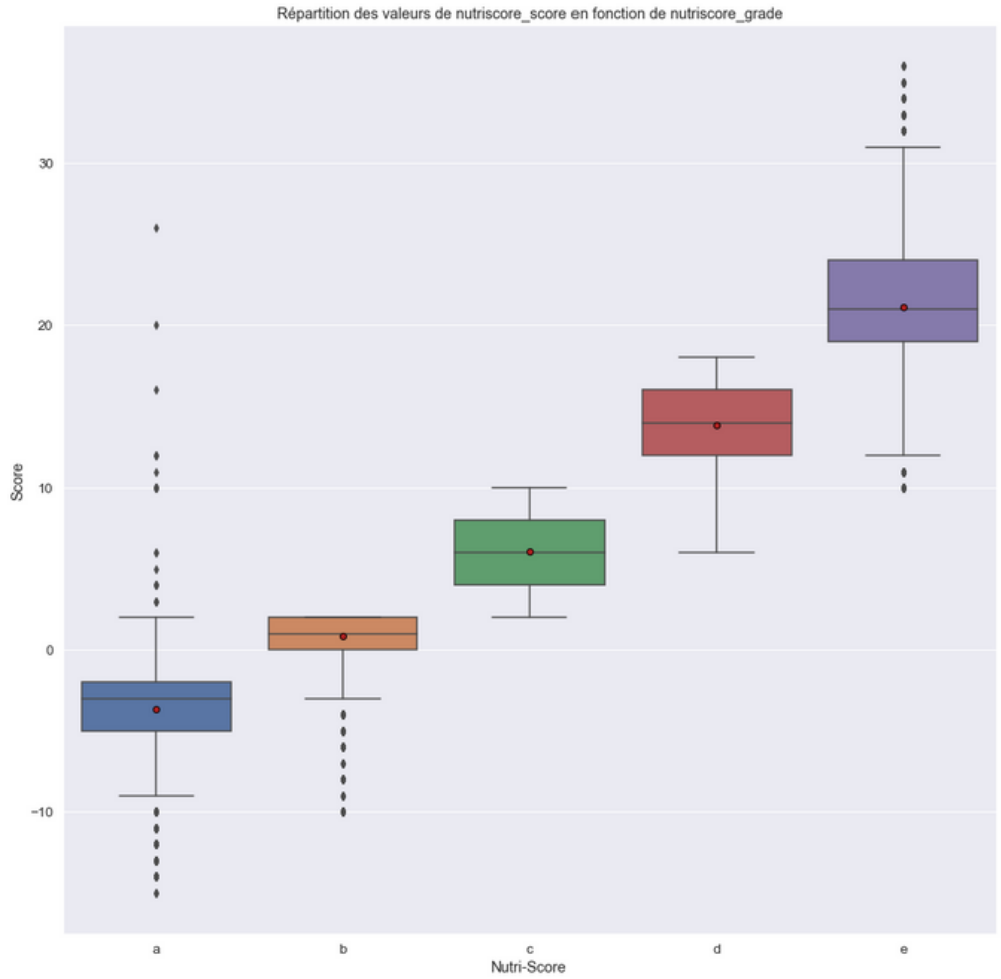


Analyse multivariée

Analyse multivariée

Section

Correlation entre `nutriscore_score` et `nutriscore_grade`



CONCLUSION

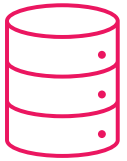
Pour aller plus loin

Interviews des utilisateurs

- Besoins pour l’analyse
- Besoins pour la page web

Exploiter la liste des ingrédients

- Exemple de la base de données
- MongoDB de Open Food Facts



Utiliser d’autres jeux de données

- Exemple de la table Ciqual de l’Anses
- Imputation de valeurs manquantes


THANK
YOU!