OPENCLASSROOM

Parcours data scientist en alternance

1 avril 2022

Maxime Dupouy

Livrable 3

Mission	Objectifs
Appel à projet de Santé Public France: **Trouver des idées innovantes d'applications en lien avec l'alimentation sur la base "Open Food Facts"**	Traiter le jeu de données : trouver et pitcher une idée, nettoyer de manière automatisée et pertinente Visualisations, analyses univariées, analyses multivariées, rédiger un rapport d'exploration

Pré-analyse 🔍

La source

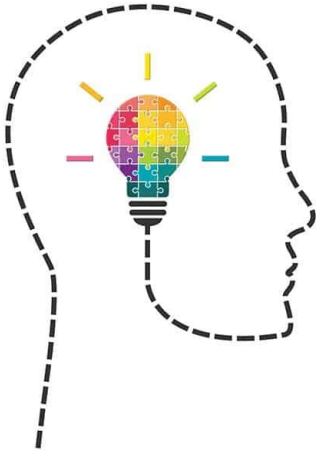
Open Food Facts est un projet collaboratif dont le but est de constituer une **base de données libre** et ouverte sur les **produits alimentaires commercialisés** dans le **monde entier**.

Le nutri-score

- Logo informant sur la **qualité nutritionnelle** des produits
- Score attribué en fonction des nutriments à favoriser versus nutriments à limiter. Plus le score obtenue est haut, moins la note sera bonne.
- Avantages : très facile à analyser, donne un bon avis global du produit,
- Limites : ne prend pas en compte certains paramètres, considère les produits dans leur ensemble

Mon projet 🚀

Health&Food-score (1)



Aider un consommateur ayant une/des pathologie(s) nécessitant un régime alimentaire en aiguillant via un score adapté à son/ses régime(s).

-> aider des consommateurs dans leur régime alimentaire quotidien afin d'optimiser leur santé

-> compléter le nutri-score (qui est plus global) en amenant un score personnalisé au consommateur

Health&Food-score (2)

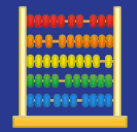
Gluci-score

Exemple :

Je suis diabétique, je veux manger un burger.

Je scanne le burger que je veux :

-> gluci-score associé au produit

AED/sélection donnée 

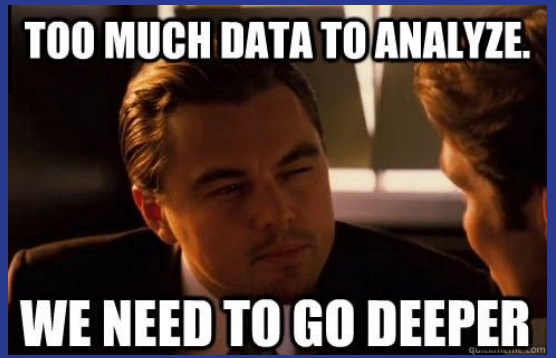
Open Food Fact Dataset

- **320772 observations** (produits alimentaires)
- **162 attributs** (caractéristiques du produits):
  - numériques (composition)
  - objects (nom, origine, marque distribution, catégorie, …)
- **76% de valeurs manquantes** *(environ 20 colonnes sans aucune valeurs !)*

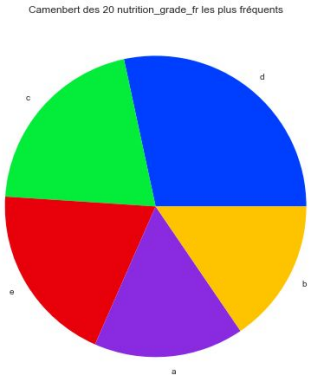
```
Entrée [6]: df.head()
```

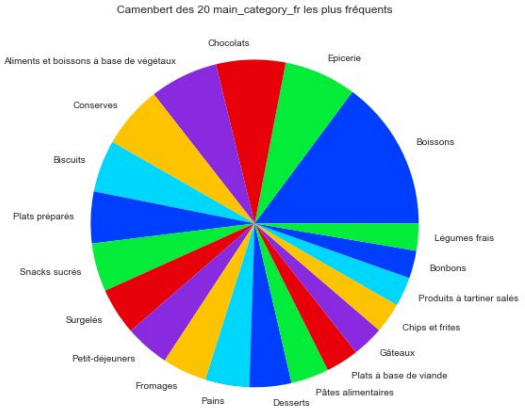
Out[6]:

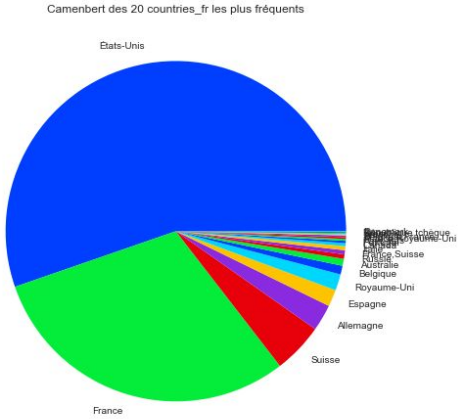
creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	quantity	...	ph_100g	fruits-vegetables-nuts_100g	collagen-meat-protein-ratio_100g
openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	NaN	1kg	...	NaN	NaN	NaN
usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	NaN	NaN	...	NaN	NaN	NaN
usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Peanuts	NaN	NaN	...	NaN	NaN	NaN
usda-ndb-import	1489055731	2017-03-09T10:35:31Z	1489055731	2017-03-09T10:35:31Z	Organic Salted Nut Mix	NaN	NaN	...	NaN	NaN	NaN
usda-ndb-import	1489055653	2017-03-09T10:34:13Z	1489055653	2017-03-09T10:34:13Z	Organic Polenta	NaN	NaN	...	NaN	NaN	NaN



Open Food Fact Dataset



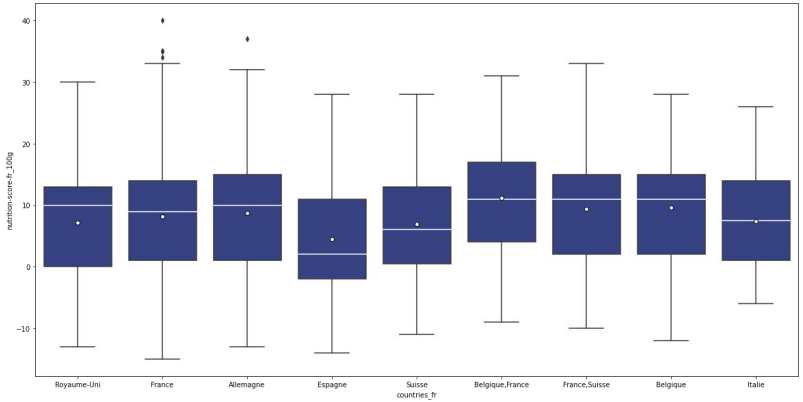




Colonnes catégorielles

Open Food Fact Dataset (3)

Sélection des observations



Nutri-score est différent en fonction des pays (confirmé par ANOVA)

Sélection de la **France** et des **pays européens** voisins

Open Food Fact Dataset (4)

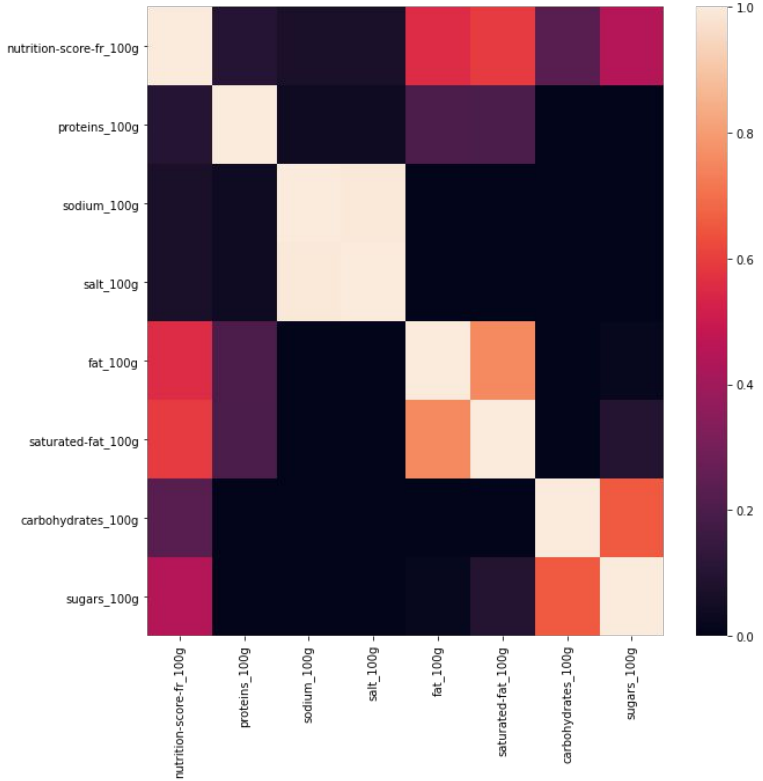
- Sélection des colonnes en fonction des **besoins métiers**

Entrée [49]:
```
columns_numeric_selected=['nutrition-score-fr_100g',
                          'proteins_100g',#effet rénal
                          'sodium_100g',#mauvais tension
                          'salt_100g',#mauvais tension
                          'fat_100g',#mauvais coeur
                          'saturated-fat_100g',#mauvais coeur
                          'cholesterol_100g',#mauvais coeur
                          'trans-fat_100g',#mauvais coeur
                        'polyunsaturated-fat_100g', #bon pour coeur
                       'monounsaturated-fat_100g', #bon pour coeur
                          'carbohydrates_100g',#mauvais diabete
                       'sugars_100g',#mauvais diabete
                       'alcohol_100g'#mauvais pour tous

]
```

Colonnes “numériques”

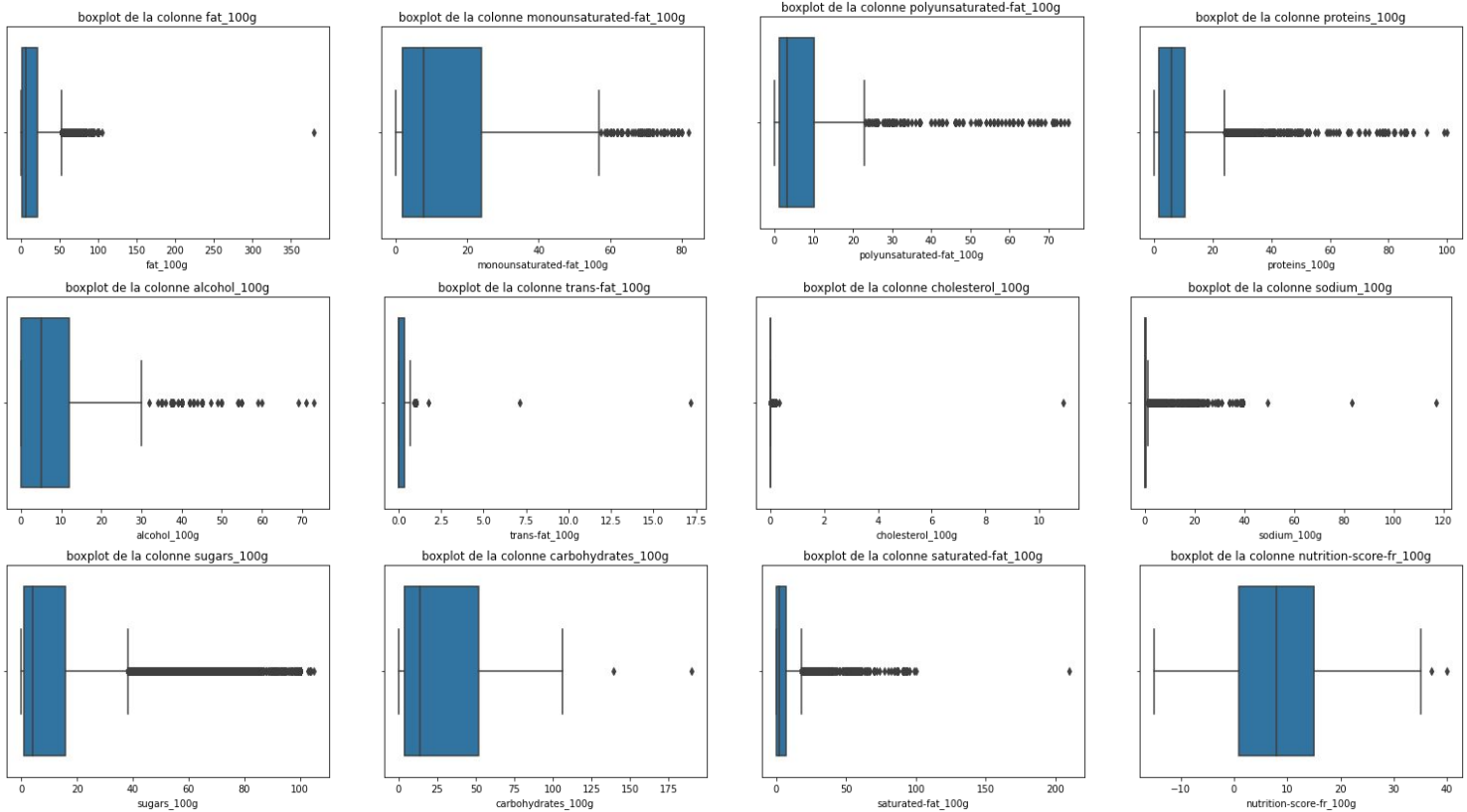
Open Food Fact Dataset (6)



Colonnes “numériques”

Open Food Fact Dataset (5)

Détection des outliers (si supérieur à 100 -> supprimer)



Colonnes “numériques”

Gérer la donnée manquante (1)

Remplir chaque features avec **la médiane** des produits de la **même catégorie**

- Sélection des données ayant une catégorie (70 000 observations/130 000)
- Nettoyer les catégories *(traduction, enlever majuscules, ponctuation, stopwords, stemming)*
- Sélectionner les catégories majeures (plus de 100 produits)
- fusion de certaines catégories majeures
- “re-pêcher” certaines catégories mineures

Gérer la donnée manquante (2)

Récupérer la donnée sans catégorie

Classification supervisé en catégorie à partir du nom du produit :

- TF-IDF + sélection donnée
- Régression logistique

-> Score F1 micro 70%

Ajout au dataset QUE si présence de données dans les colonnes numériques (environ 17 000 observations)

Gérer la donnée manquante (3)

Réalisation **dataset de médianes**

main_category_stemmed	appl juic	babi food	bacon bit	beverag	biscuit	bread	bread product	breakfast	butter	cake	...	tuna	veget oil	veget rod	v
nutrition-score-fr_100g	4.000000	0.000000	20.000000	9.000000	20.000000	2.000000	2.000000	9.0000	19.00000	17.00000	...	1.000	12.0	-4.00000	5.
proteins_100g	0.100000	2.700000	17.000000	0.500000	6.600000	9.400000	12.500000	8.1000	0.70000	5.60000	...	25.000	0.0	1.80000	0.
sodium_100g	0.003937	0.037701	0.984252	0.011811	0.216535	0.472441	0.466315	0.1400	0.03937	0.21700	...	0.400	0.0	0.30315	0.
salt_100g	0.010000	0.095760	2.500000	0.030000	0.550000	1.200000	1.184440	0.3556	0.10000	0.55118	...	1.016	0.0	0.77000	0.
fat_100g	0.100000	2.200000	20.000000	0.100000	21.000000	4.750000	10.000000	7.7000	82.00000	21.00000	...	3.250	92.0	0.30000	0.

5 rows × 106 columns

Résultats :

Départ **130 000** lignes sélectionnées (**50%** de données manquantes)

89 000 lignes **100%** valeurs

Calcul scores

Pour chaque feature d'intérêt, classement des produits puis attribution d’interquintile -> 5 intervalles

	product_name	product_name	main_category_fr	nutrition-score-fr_100g	proteins_100g_score	salt_score	lipide_score	satured_fat_score	gluci_score	sucre_score
22877	Whim of the gods	Caprice des dieux	Fromages de vache	15.0	5	5	5	5	1	1
89839	Cocofin made from coconut oil	Cocofin aus Kokosöl	Huiles	20.0	1	1	5	5	1	1
44344	Special spicy pizza oil	Huile spéciale pizza pimentée	Huiles	11.0	1	1	5	5	1	1
49955	Colza & Olive	Colza & Olive	Huiles	11.0	1	1	5	5	1	1
44336	Sunflower heart easy to spread	Coeur de Tournesol facile à étaler	Huiles	11.0	1	1	5	5	1	1

LIVE DEMO

Les limites

Limites sur la gestion de la donnée :

- perte de données sur sélection (alcool, AG saturé, glucides, …)
- scores calculés “simplistes”
- approximation médiane
- utilisation de donnée de plusieurs pays

Limite sur l’application :

- difficulté à analyser plusieurs scores
- problématique RGPD

Développer

- Réaliser un **score global**
- **Améliorer système classification** catégorie (deep learning, agrégation informations)
- Améliorer systeme **sélection produit** (via photo/code barre)
- **Intégrer plus de features**
- Intégrer les ingrédients pour **allergie**
- Proposer des **alternatives “meilleures”** pour un produit

Questions Time !

Je vous remercie pour votre attention