**Parcours OpenClassRooms**

Data Scientist

---

# P3 Conception d'une application au service de la santé publique

Développée sur un Notebook Jupyter Colaboratory


Benoît DELORME
Création : 18/10/2020
Mise à jour : 09/11/2020

Sommaire

1) Idées d’application

2) Nettoyage effectué

3) Analyse exploratoire

4) Faits pertinents pour l’application

5) Questions-réponses

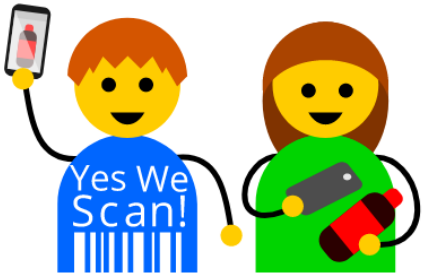
# 1) Idées d’application

**Présentation du site**

Le site Open Food Facts se présente comme une base de données alimentaires. Elle est alimentée sur la base du volontariat, chacun pouvant l’enrichir en scannant le code-barre du produit qu’il souhaite enregistrer.

A food products database

Open Food Facts is a database of food products with ingredients, allergens, nutrition facts and all the tidbits of information we can find on product labels.




Made by everyone

Open Food Facts is a non-profit association of volunteers.

15000+ contributors like you have added 660 000+ products from 150 countries using our Android, iPhone or Windows Phone app or their camera to scan barcodes and upload pictures of products and their labels.

Le Programme Français de Nutrition et de Santé Nationale exploite les données d’Open Food Facts pour valider les formules de ses notations qualité en termes de nutritions.

**Idée d’application ?**

En parcourant rapidement le site, plusieurs idées d’application me sont venues en tête. Certains ne seront pas possible avec les données mises à disposition, d’autres valent la peine d’être tentées.

J’ai trouvé difficile de fixer l’idée d’application à partir des données générales, sans savoir si elles étaient fiables, ou même disponibles.

Nous allons d’abord nettoyer le fichier original, puis sélectionner les colonnes intéressantes, et les analyser (Analyse exploratoire, analyses univariées). Parmi ces colonnes, celles qui seront jugées inutiles par leur contenu seront écartées.

Nous nous servirons des colonnes restantes pour élaborer quelques idées d’application, à confirmer à l’aide d’analyse multivariées.

# 2) Nettoyage effectué

**1er nettoyage : projection**

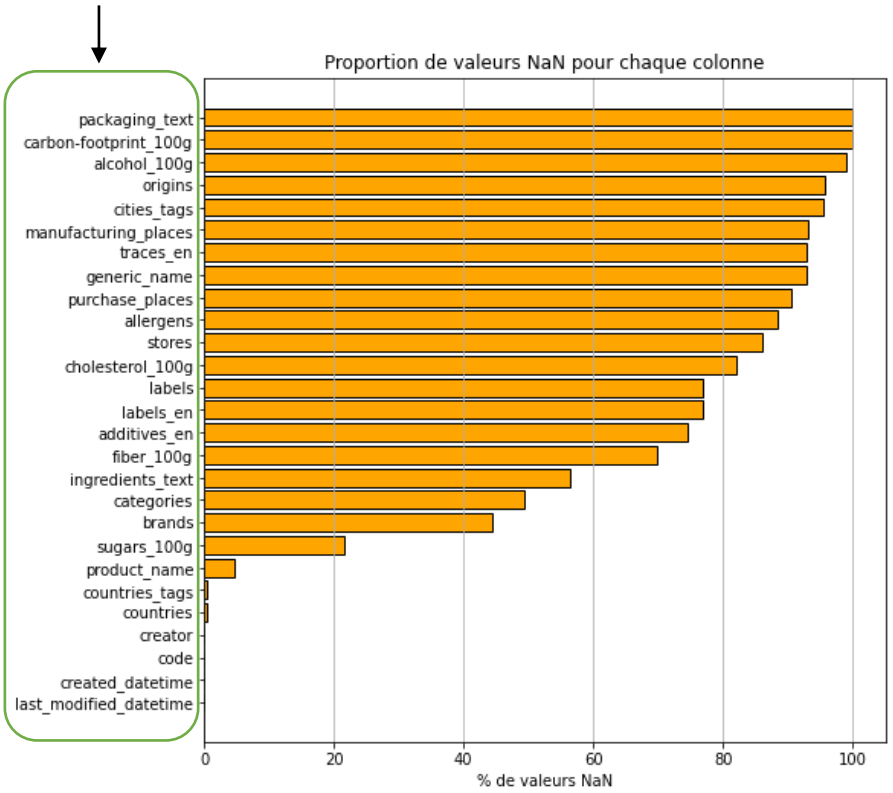
Les données du site sont regroupées en un tableau CSV de 1.486.047 lignes et 182 colonnes.

Toutes ces colonnes n'étant pas forcément utiles, il m'a paru intéressant de créer un nouveau tableau ne gardant que les colonnes les plus susceptibles d'aider l'application.

Colonnes gardées sur les 182 colonnes de départ

Sur ces colonnes restantes, il y a parfois de nombreuses données absentes (les valeurs NaN).

Le diagramme à barres ci-contre montre la proportion en % de ces valeurs NaN pour chaque colonne.

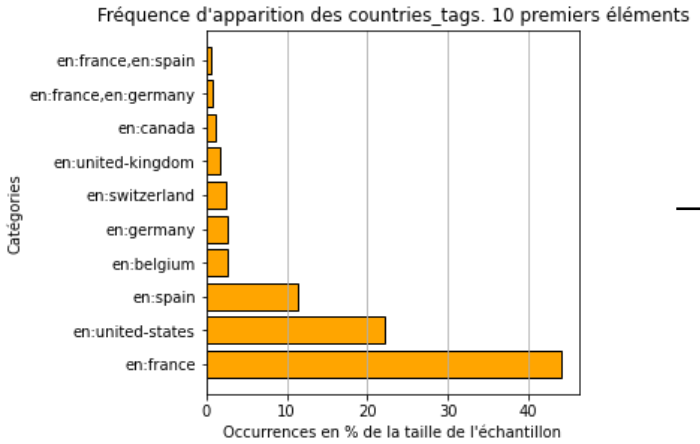



D'autres colonnes étaient disponibles, telles que fat_100g, salt_100g, … Elles n'ont pas été retenues pour simplifier l'étude.

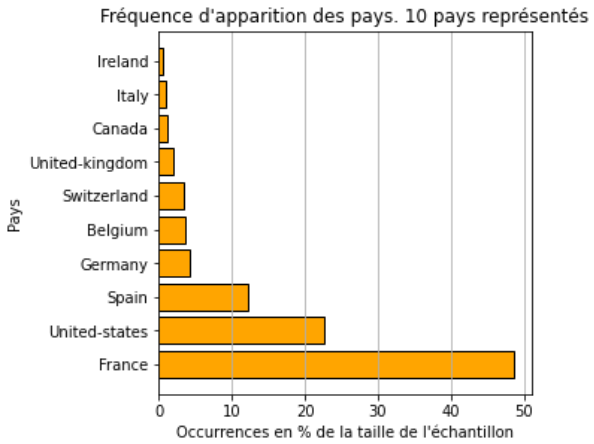
## 2ème nettoyage : traitement de texte

1) Tous les noms de pays sont précédés de «en:», enlevé pour plus de lisibilité.

2) Certains « pays » sont en fait des listes de pays, par exemple : «en:france, en:germany». Un traitement de texte a permis d'agréger les données par pays. Les axes des ordonnées des deux graphiques ci-dessous permettent de s'en rendre compte.




Nettoyage

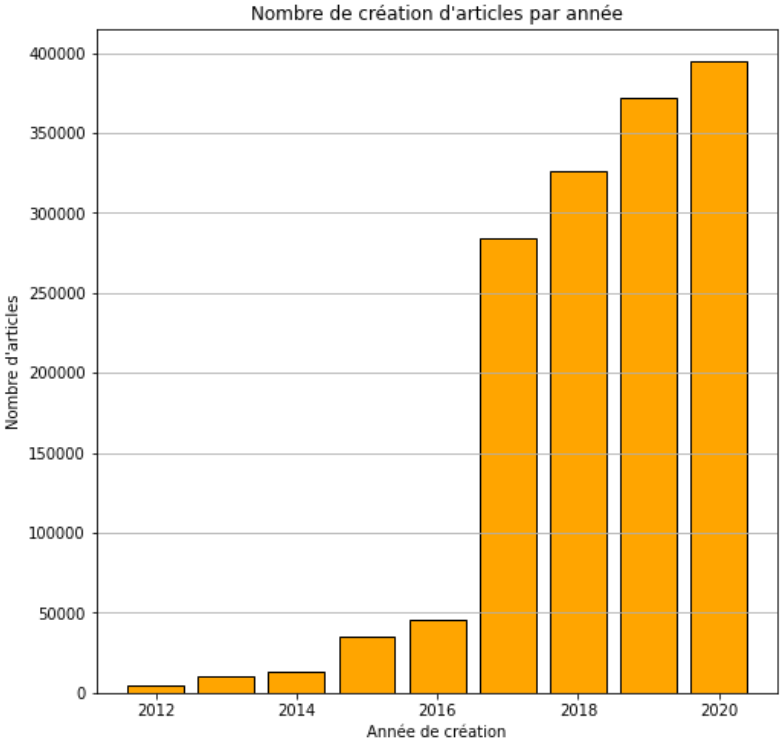


# 3) Analyse exploratoire

## Analyses univariées

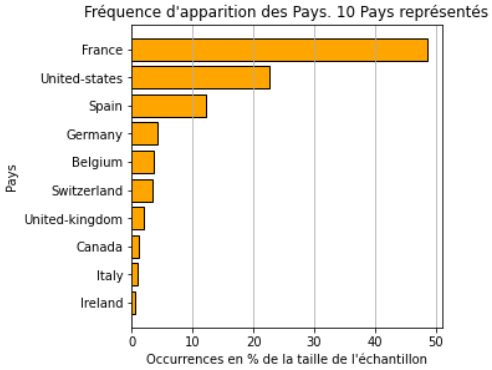
Les colonnes retenues ont été explorées une à une, afin de :
- comprendre leur contenu,
- repérer les éléments les plus représentés,
- et évaluer la pertinence de la colonne.

**created_date_time**
Analyse du nombre de création d'articles par année

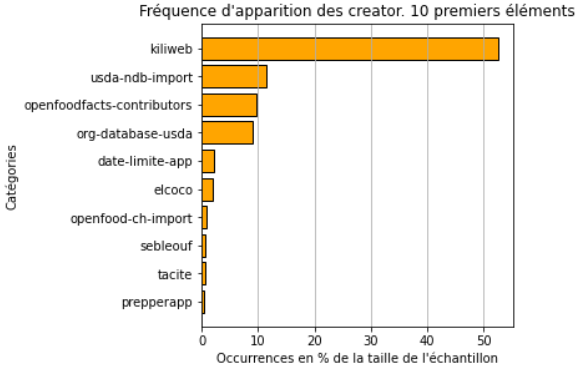



Les créations d'article ont démarré en 2012, sans interruption jusqu'en 2020.
Leur nombre n'a cessé de croître depuis les débuts d'OpenFoodFacts, avec une véritable explosion en 2017.
Depuis 2017, le nombre de création augmente à un rythme soutenu.

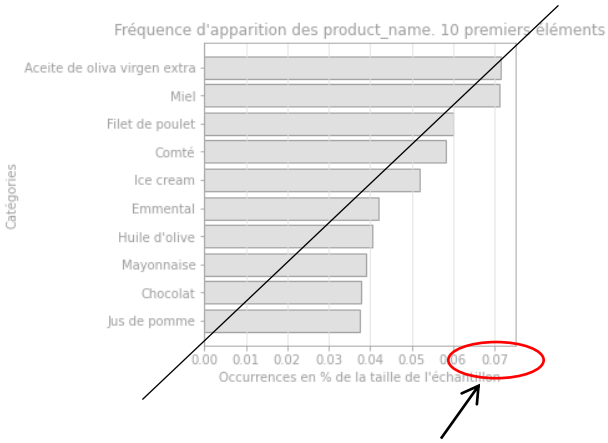
## Noms de pays




## Contributeurs

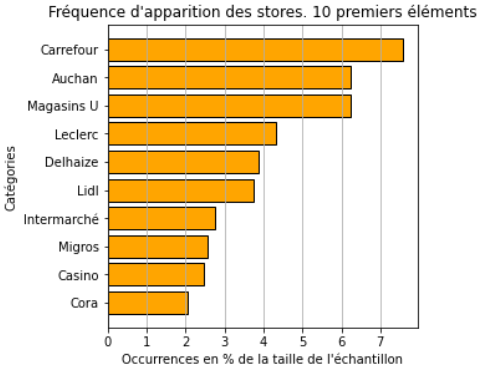



## Noms des produits

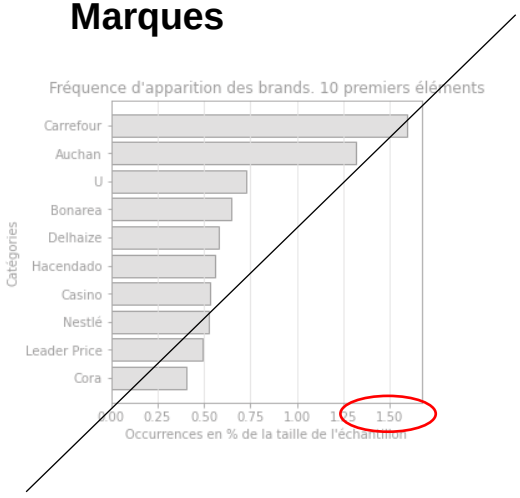



Non retenu car les données sont trop diverses

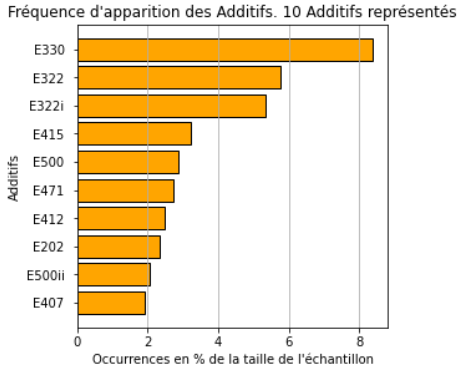
## 5Enseignes



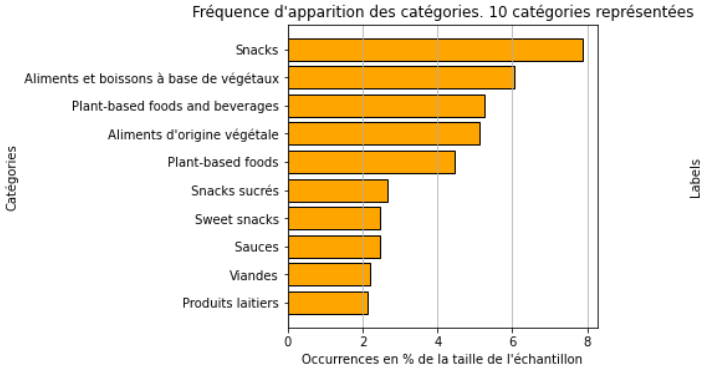

## Marques



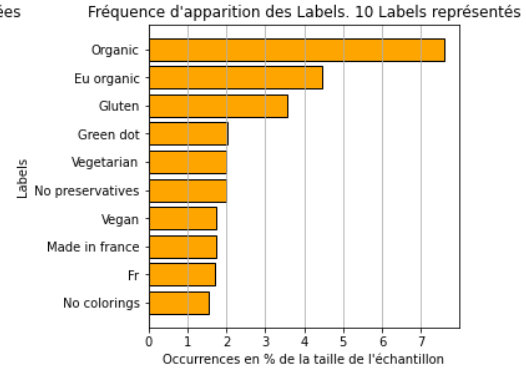

## Additifs



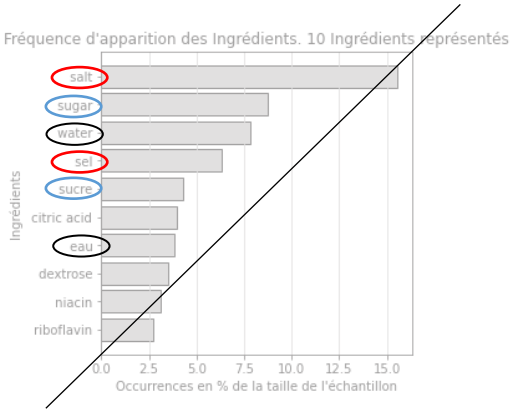
## Catégories




## Labels

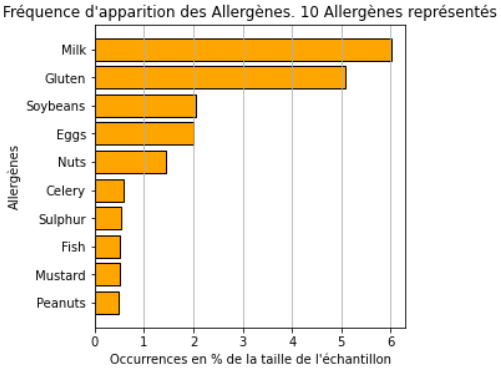



## Ingrédients

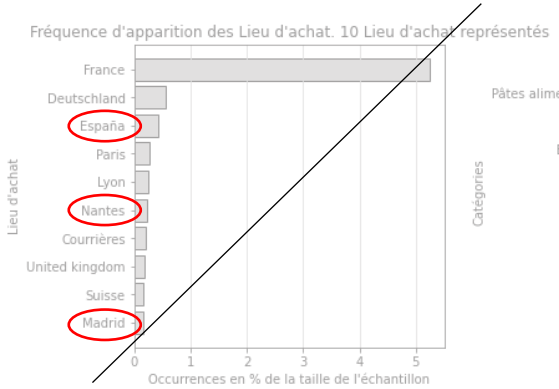



Non retenu car les données sont trop longues à nettoyer

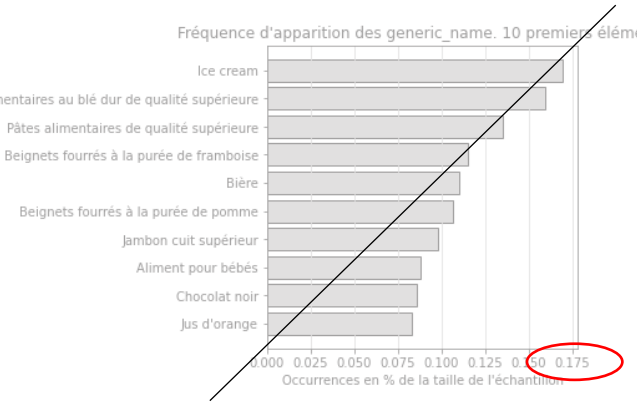
## Allergènes



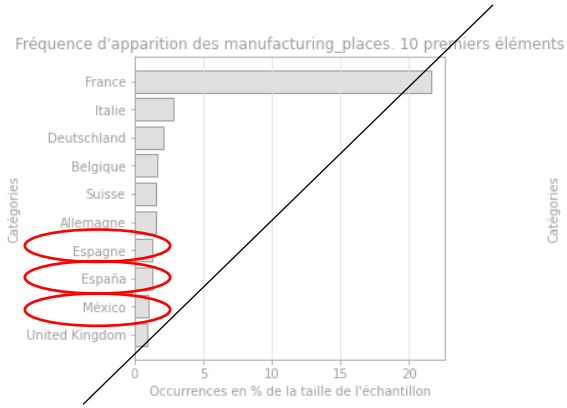

## Lieux d'achat



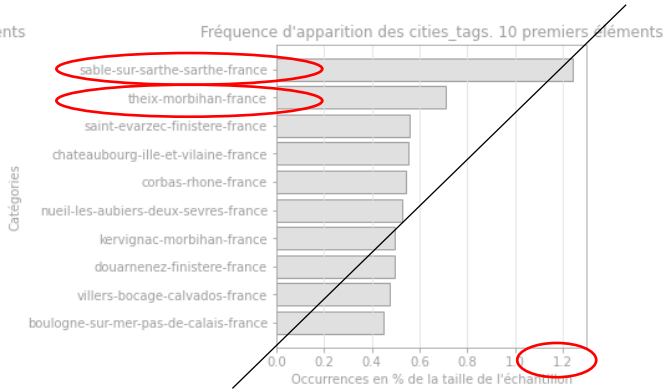

## Noms génériques



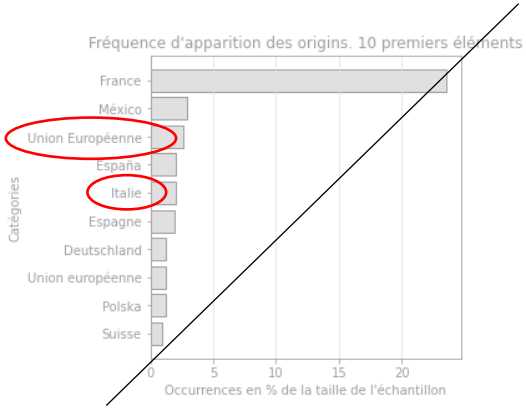
## Lieux de production



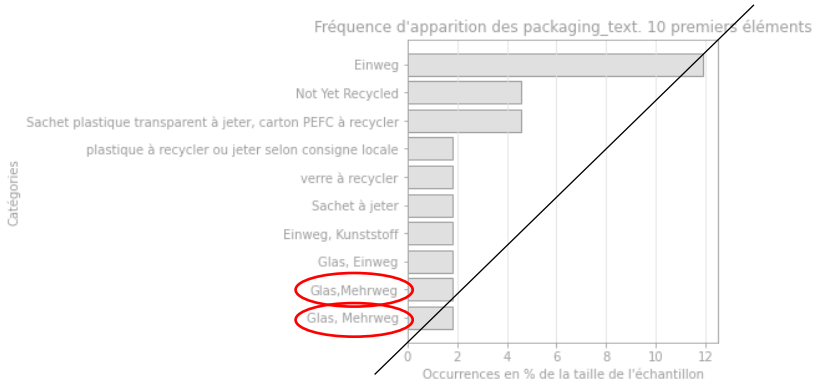

## Villes




## Pays d'origine




## Texte d'emballage



Les analyses univariées menées précédemment permettent d'éliminer les colonnes inutiles.

Il nous reste les 9 indicateurs suivants :

- La date de création — donnée quantitative
- Le contributeur
- Les pays
- Les catégories
- Les additifs
- Les labels
- L'enseigne
- Les allergènes

(Le contributeur … Les allergènes : données qualitatives)

- Le taux de cholestérol
- Le taux de sucre

(Le taux de cholestérol, Le taux de sucre : données quantitatives)

**Idée d'application**

Et pourquoi pas créer une application visant à donner des informations générales sur :

- L'implication des supermarchés dans la promotion des principaux labels,
- Le potentiel de cholestérol, d'après les informations disponibles sur l'emballage du produit
- Le potentiel de sucre, d'après les informations disponibles sur l'emballage du produit

Cette application serait à la fois au service du consommateur, et à la fois au service des organismes publics voulant récompenser les supermarchés impliqués dans la promotion des labels.

# 4) Faits pertinents pour l'application

## Analyses multivariées

### Qualitatif / qualitatif

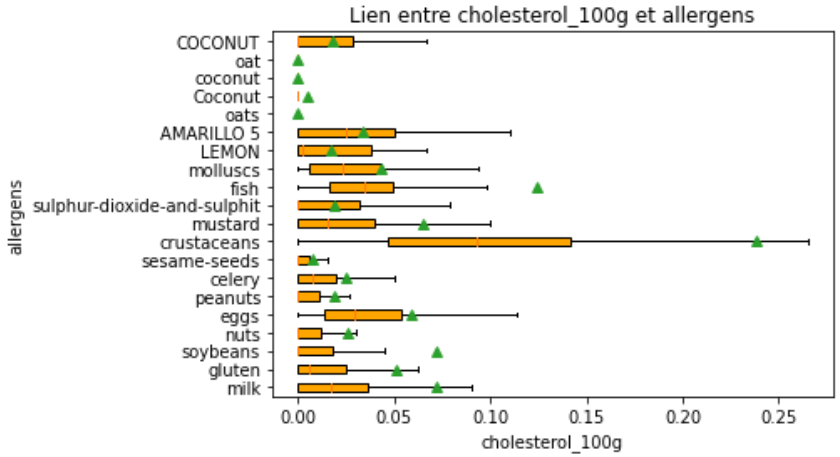
| labels_en | Organic | Gluten-free | Made in France | Green Dot | Vegetarian,Vegan | No preservatives | No colorings | No colorings,No preservatives | Organic,EU Organic,fr:ab-agriculture-biologique | No added sugar |
|---|---|---|---|---|---|---|---|---|---|---|
| stores | | | | | | | | | | |
| Aldi | 15.000000 | 43.000000 | 28.000000 | 82.000000 | 113.000000 | 30.000000 | 4.000000 | 2.000000 | 3.000000 | 3.000000 |
| Auchan | 67.000000 | 158.000000 | 346.000000 | 348.000000 | 9.000000 | 55.000000 | 35.000000 | 36.000000 | 38.000000 | 20.000000 |
| Carrefour | 233.000000 | 133.000000 | 923.000000 | 253.000000 | 224.000000 | 97.000000 | 29.000000 | 62.000000 | 80.000000 | 39.000000 |
| Casino | 125.000000 | 32.000000 | 15.000000 | 96.000000 | 3.000000 | 38.000000 | 3.000000 | 24.000000 | 63.000000 | 9.000000 |
| Colruyt | 30.000000 | 16.000000 | 7.000000 | 34.000000 | 4.000000 | 6.000000 | 3.000000 | 1.000000 | 12.000000 | 5.000000 |
| Cora | 19.000000 | 12.000000 | 88.000000 | 112.000000 | 3.000000 | 44.000000 | 10.000000 | 28.000000 | 9.000000 | 21.000000 |
| Delhaize | 121.000000 | 70.000000 | 16.000000 | 86.000000 | 34.000000 | 40.000000 | 52.000000 | 36.000000 | 26.000000 | 36.000000 |
| Franprix | 25.000000 | 17.000000 | 10.000000 | 150.000000 | 2.000000 | 15.000000 | 8.000000 | 3.000000 | 11.000000 | 3.000000 |
| Intermarché | 40.000000 | 13.000000 | 64.000000 | 133.000000 | nan | 57.000000 | 14.000000 | 17.000000 | 33.000000 | 24.000000 |
| Kroger | 1.000000 | 3.000000 | nan | nan | nan | nan | nan | nan | nan | nan |
| Leader Price | 37.000000 | 12.000000 | 42.000000 | 150.000000 | nan | 25.000000 | 13.000000 | 15.000000 | 19.000000 | 15.000000 |
| Leclerc | 77.000000 | 41.000000 | 56.000000 | 528.000000 | 7.000000 | 45.000000 | 71.000000 | 39.000000 | 37.000000 | 32.000000 |
| Lidl | 55.000000 | 76.000000 | 79.000000 | 206.000000 | 155.000000 | 27.000000 | 17.000000 | 9.000000 | 17.000000 | 12.000000 |
| Magasins U | 218.000000 | 94.000000 | 239.000000 | 269.000000 | 4.000000 | 131.000000 | 50.000000 | 110.000000 | 65.000000 | 35.000000 |
| Meijer | 148.000000 | 2.000000 | nan | nan | nan | nan | nan | nan | nan | nan |
| Mercadona | 2.000000 | 378.000000 | nan | 64.000000 | 134.000000 | 10.000000 | 12.000000 | 10.000000 | nan | 12.000000 |
| Migros | 210.000000 | 9.000000 | 3.000000 | 4.000000 | 70.000000 | 14.000000 | 4.000000 | 10.000000 | 2.000000 | 20.000000 |
| Monoprix | 76.000000 | 14.000000 | 31.000000 | 117.000000 | 5.000000 | 20.000000 | 5.000000 | 16.000000 | 43.000000 | 17.000000 |
| Netto | 23.000000 | 6.000000 | 22.000000 | 80.000000 | 9.000000 | 24.000000 | 18.000000 | 5.000000 | 2.000000 | 12.000000 |
| Picard | 5.000000 | 1.000000 | 176.000000 | 77.000000 | 5.000000 | 2.000000 | nan | 4.000000 | nan | nan |
| All | 1527.000000 | 1130.000000 | 2145.000000 | 2789.000000 | 781.000000 | 680.000000 | 348.000000 | 427.000000 | 460.000000 | 315.000000 |

Ce tableau peut intéresser aussi bien un particulier qu'une agence gouvernementale :

- une personne intolérante au gluten va regarder la colonne gluten-free, et repérer les supermarchés ayant le plus de produits correspondant à son intolérance. Elle pourra ainsi préférer la marque de supermarché qui promeut le plus les produits sans gluten.
- Une agence gouvernementale peut évaluer chaque chaîne de supermarché et évaluer leur implication dans la promotion des labels (un affichage de la proportion de produit avec label serait plus utile).

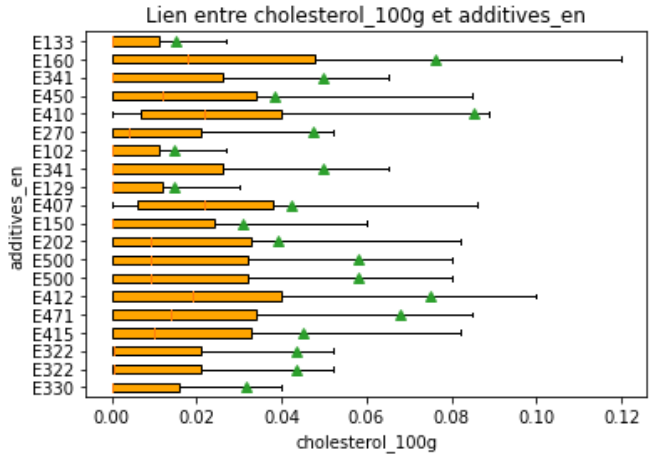
Quantitatif / Qualitatif

## Taux de cholestérol et allergènes



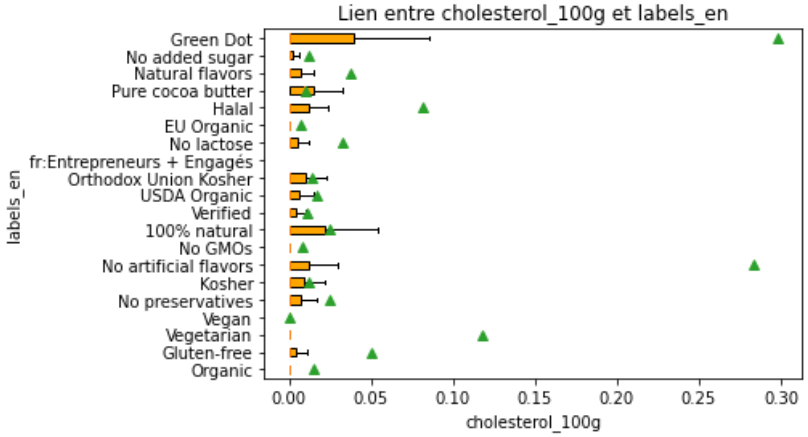

Pour une personne sujette au cholestérol, les produits contenant des traces de crustacées sont à éviter.

## Taux de cholestérol et additifs

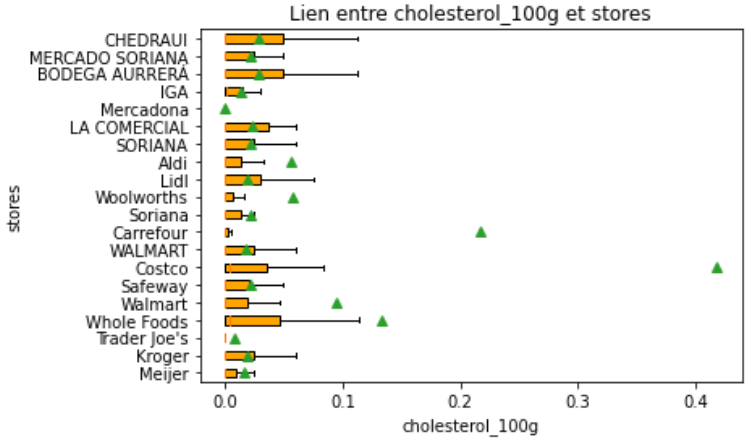



Pour une personne sujette au cholestérol, les produits contenant les additifs E410 et E407 sont à éviter.

## Taux de cholestérol et labels

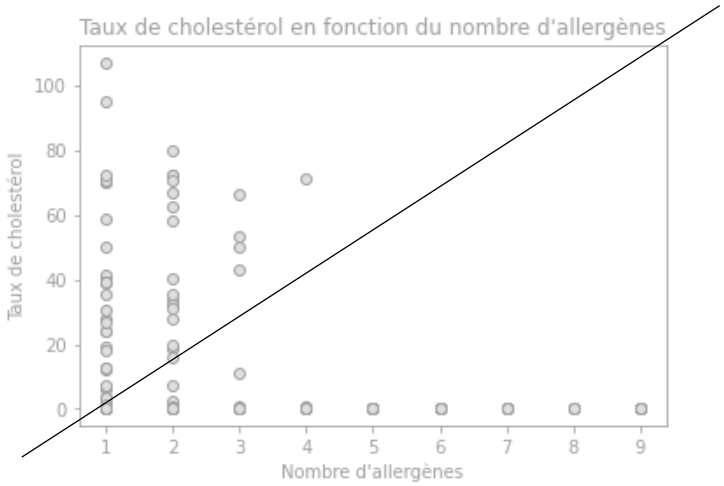



## Taux de cholestérol et enseignes



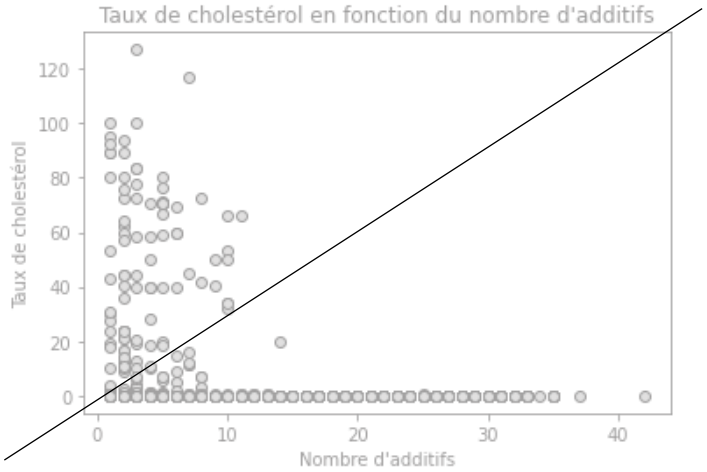
Quantitatif / Quantitatif

## Cholestérol (mg/100g) / nombre d'allergènes présents dans le produit




**Analyse**
Il n'y a pas de corrélation évidente entre le taux de cholestérol et le nombre d'allergènes.

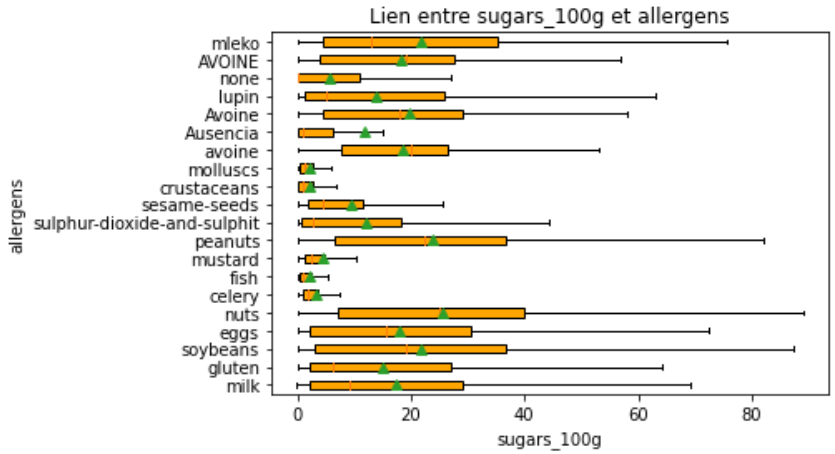
## Cholestérol (mg/100g) / nombre d'additifs présents dans le produit



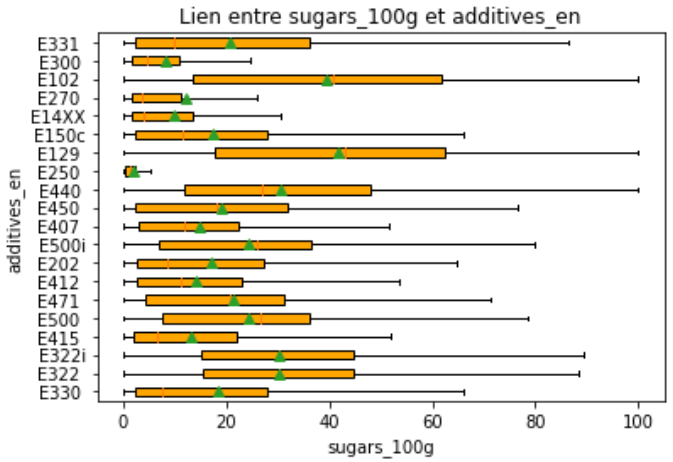

**Analyse**
Il ne semble pas y avoir de corrélation entre le taux de cholestérol et le nombre d'additifs.

Quantitatif / Qualitatif

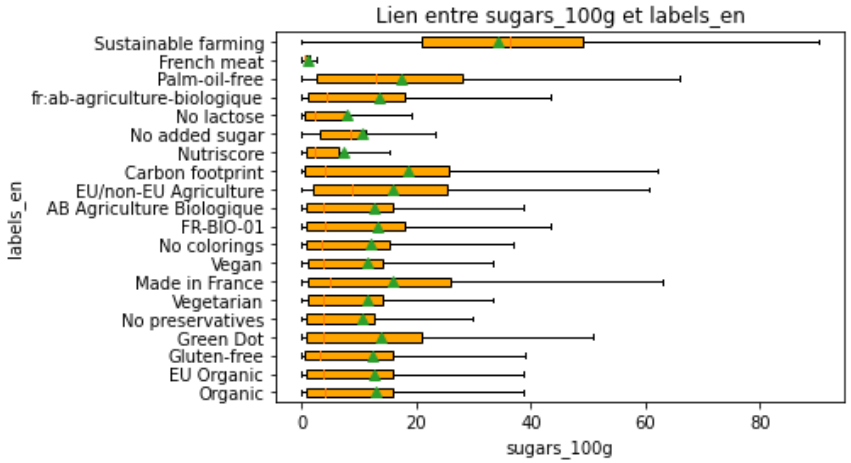
## Taux de sucres et allergènes



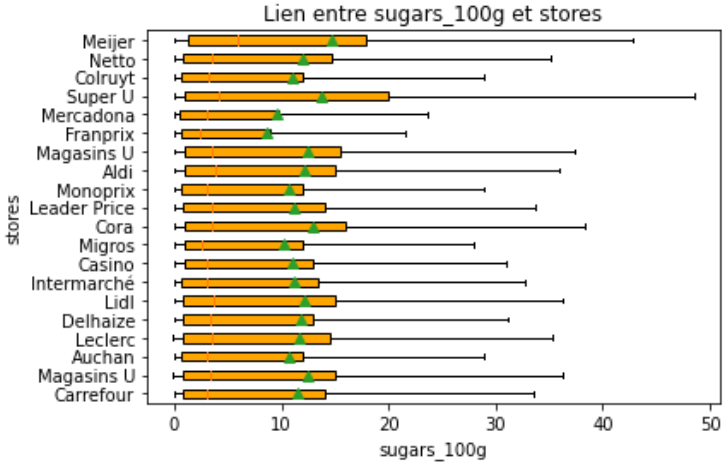

## Taux de sucres et additifs




## Taux de sucres et labels

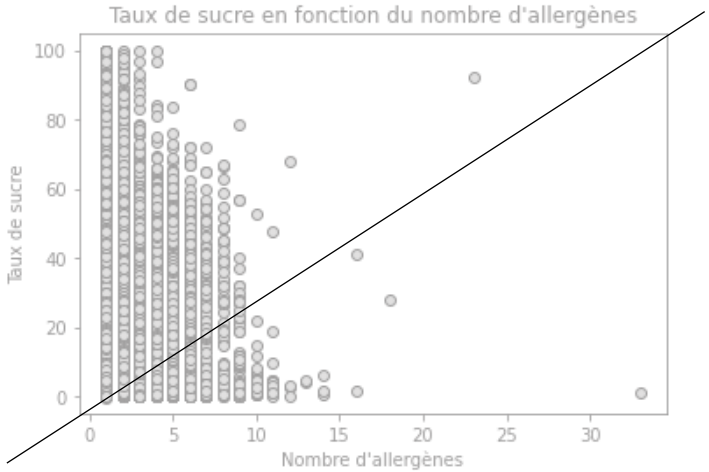



## Taux de sucres et enseignes

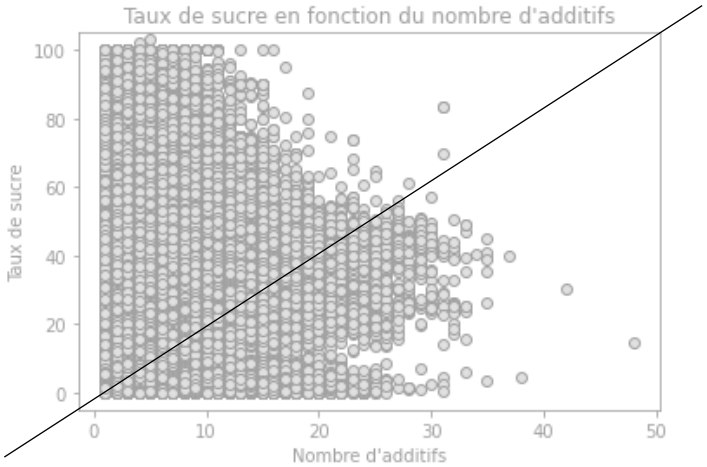


Quantitatif / Quantitatif

**Taux de sucres et nombre d'allergènes présents dans le produit**




**Taux de sucres et nombre d'additifs présents dans le produit**



**Conclusions**

- Il est possible d'extraire un certain nombre de données du site Open Food Facts, et d'en réaliser quelques analyses.
- Toutes les données ne sont pas utilisables, selon leur disponibilité et leur diversité.
- Ces analyses peuvent concerner aussi bien les organismes publics (Agence Santé Publique France, ministère de la santé, …), que des particuliers (diabétiques, cholestérols, …)
- Certaines autres données peuvent également être exploitées, comme le nutrition score.
    - Cela dit, il s'agit de données « transformées », c'est-à-dire calculées, contrairement aux taux de cholestérols, taux de sucre, …
    - Ces indicateurs calculés sont simples à comprendre (notes de « A » à « E »), mais leur calcul n'est pas facilement accessible et donc un peu obscur.
- En tout cas, la base de données semble contenir les données nécessaires pour une application sanitaire :
    - Des données calculées et simples à comprendre.
    - Des données brutes (taux de cholestérol, taux de sucre, taux de graisses, …), propres aux analyses statistiques?

# 5) Questions-réponses

**Merci pour votre attention**

**Vos questions sont les bienvenues**