Concevez une application au service de la santé publique

Thibault Grandjean[1]

[1] *Étudiant auprès d'Openclassrooms*

8 février 2020

L'obésité est aujourd'hui une maladie au centre de l'attention. En effet, ce sont 17% des adultes en France (13 % dans le monde) qui sont concernés. Le nombre de cas d'obésité à presque triplé depuis 1975.[1] [3] Cette maladie résulte d'un déséquilibre entre les apports et les dépenses énergétiques. Pour enrayer la tendance, l'État français a mis en place un plan de prévention national au travers du ministère de la santé. Le présent document étudie la faisabilité d'une application permettant de mieux manger dans le cadre du plan de prévention national.

1 Introduction

Le surpoids et l'obésité se définissent comme une accumulation anormale ou excessive de graisse corporelle qui représente un risque pour la santé.

Le surpoids et l'obésité sont des facteurs de risque majeurs pour un certain nombre de maladies chroniques, parmi lesquelles le diabète, les maladie cardiovasculaires et le cancer. [3]

Pour endiguer ce mal, l'utilisation des nouvelles technologies à des fins éducationnelles est une voie prometteuse.

1.1 Contexte

L'agence santé publique France a lancé un appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation. Nous répondons donc à cet appel à projets avec une idée d'application pour smartphone, permettant sur base des informations nutritionnelles de proposer des produits équivalents à ceux désirés par le consommateur mais avec de meilleurs propriétés nutritionnelles.

Dans un second temps, il sera envisable d'implémenter un moteur de recommandation sur base des habitudes de consommation des utilisateurs de l'application.

2 Objectifs

L'objectif principal est d'étudier la faisabilité d'une telle application. Pour ce faire, une analyse minitieuse des données contenues dans la base de données d'openfoodfacts est nécessaire. Il s'agit dans un premier temps de cerner les variables nécessaires au fonctionnement de l'application et ensuite vérifier que les données permettent bien de répondre à la problématique.

3 Problématique

La problématique est double :
La première chose est de s'assurer que les données dont on dispose contiennent bien les informations nécessaires pour créer une telle application.
La deuxième chose :
Peut-on trouver dans la base de données, un produit équivalent avec de meilleurs propriétés nutritionnelles à partir d'un produit scanné dans un magasin ?

4 Démarche

Pour répondre à la problématique, on se base sur la base de données d'openfoodfacts (à l'heure actuelle la plus grosse base de produits alimentaire). Une fois les données téléchargées, ces dernières subissent un

nettoyage. On réalise alors une analyse statistique descriptive sur une sélection de variables d'intérêt. L'analyse statistique est divisée en deux parties, les analyses univariées et les analyses bivariées.

4.1 Outils

4.1.1 Matériel

L'analyse a été réalisée sur un ordinateur personnel (processeur intel i7 4 coeurs, 16 Go de RAM.) et ne nécessite pas de matériel particulier.

4.1.2 Logiciels

L'analyse a été réalisée avec le language Python et des notebooks Jupyter sont disponibles dans le répertoire Github (voir section 10)

5 Les données utilisées

5.1 Généralités

Les données utilisées sont disponibles gratuitement auprès d'Openfoodfacts et sont publiées sous licence "Open Database License".

Les données sont entrées par les utilisateurs (Applications mobiles : Openfoodfacts et Yuka). Par conséquent les données sont régulièrement mal complétées ou erronées. La base de données pèse actuellement (2.2 Go), elle contient (à l'heure actuelle [1]) 1 120 752 [2] d'entrées et 178 colonnes.

5.2 Contenu de la base de données

Les champs sont séparés en quatres sections :

— Les informations générales sur la fiche du produit : nom, marque, date de création...
— Un ensemble de tags : catégorie du produit, localisation, origine, etc.
— La liste des ingrédients et les additifs éventuels.
— Des informations nutritionnelles : quantités au 100g (graisse, sucre, etc.).

5.3 Nettoyage de la base de données

5.3.1 Données concernant la France

On ne récupère que les données pour la France et les pays limitrophes francophones.

1. 23 janvier 2020
2. 626 672 pour les pays francophones (France, Suisse, Belgique et Luxembourg)

5.4 Variables sélectionnées

Pour assurer la faisabilité du projet, nous avons besoins de :

— Le nom du produit
— Le code bar du produit
— La marque du produit
— La catégorie à la quelle appartient le produit
— Le nutriscore (nutrition grade [A, ..., E])
— Certaines données nutritionnelles :
  — Valeur énergétique au 100g
  — Teneur en protéines
  — Teneur en sucres
  — Teneur en graisses
  — Teneur en sel

5.5 Taux de remplissage des champs

On peut regarder le taux de remplissage des champs de manière graphique à l'aide d'un graphique type matrice. (chaque valeur est alors représentée par un tiret (les colonnes noires sont alors totalement complètes et les colonnes vides sont blanches.) Voir graphique 1

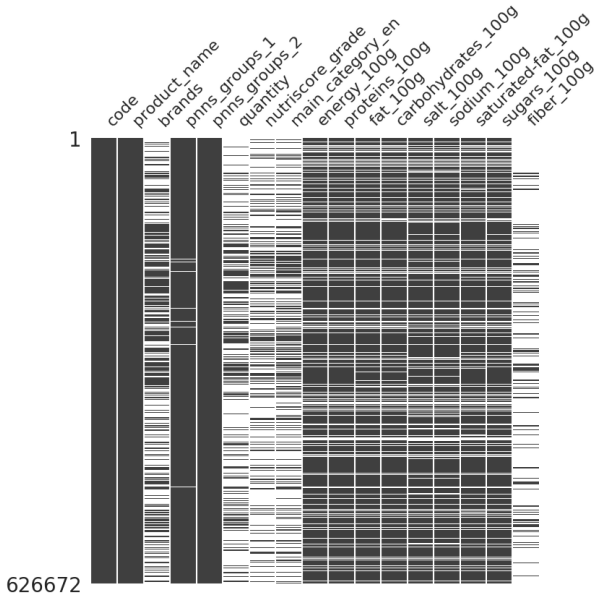

Figure 1 – *Taux de remplissage des champs*

5.5.1 Stratégie appliquée aux valeurs manquantes

Pour les champs : *quantité, nutriscore, catégorie, valeur énergétique*, on supprime le produit si l'une des information est manquante.

Les champs *marque, groupe PNNS 1 et 2*, on remplace les valeurs manquantes par une catégorie inconnue `unknown`

5.6 Champs contenant des valeurs textuelles

Les champs «PNNS groups» «Catégorie» et «Marques» sont traités par méthode de «clustering». L'algorithme utilisé est l'algorithme de «Key collision» méthode par défaut d'OpenRefine[2].

Le champs quantité est traité grace à un parser.

5.7 Champs numériques

Les champs numériques représentant une quantité de nutriments au 100g (ex : protéines) doivent être bornés dans l'interval $\mathbb{R} \in [0, 100]$

Le champs «Valeur énergétique» est borné dans l'interval $\mathbb{R} \in [0, 4000]$ [3]

Pour les champs correspondant à une sous classe d'un nutriment (ex : glucides et sucre) on vérifie que le champs «enfant» est inférieur ou égal au champs «parent»

On vérifie en dernier lieu que la somme des valeurs protéines, graisse, glucides, sel et fibres est inférieur à 100g

6 Analyses univariées

6.1 Répartition du nutriscore

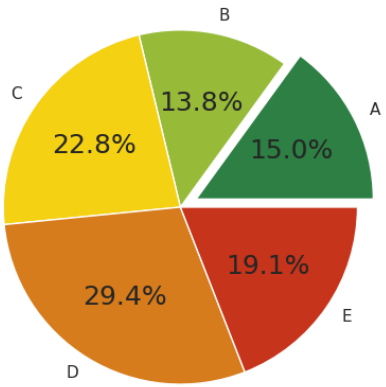

Figure 2 – *Répartition du nutriscore des produits*

On regarde ici la répartition des produits en fonction du nutriscore sous forme d'un diagramme de type camembert. On observe une répartion relativement équitable entre les différentes modalités. La classe D étant la plus présente et la B la moins présente.

3. 0 étant la valeur nutritionnelle de l'eau et l'élément le plus riche est la graisse $\approx$ 3700KJ

6.2 Marques, groupes PNNS et catégories

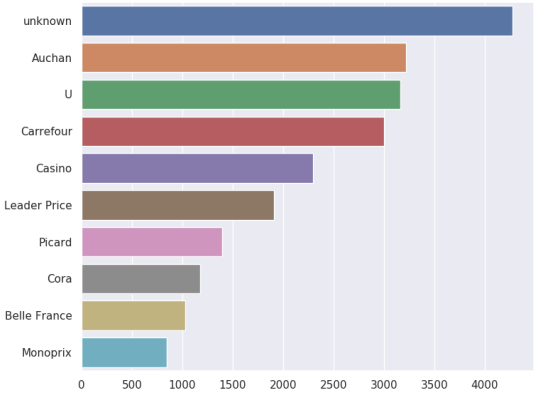

Figure 3 – *Répartion des produits en fonction des marques*

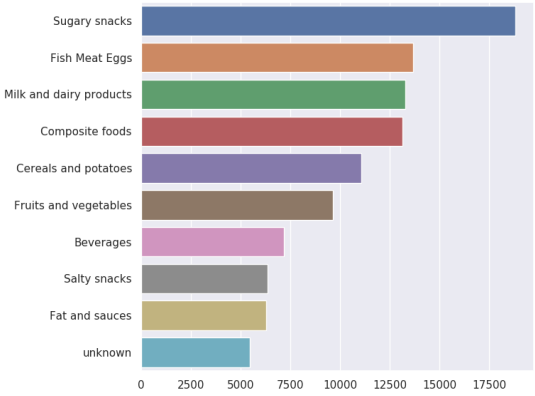

Figure 4 – *Répartion des produits en fonction du groupe PNNS*

On observe qu'un grand nombre de produits n'ont pas de marque renseignée ("Unknow") Le graphique montre les 10 modalités les plus représentées. Cependant, il y a 16 185 modalités présentent (dont 9572 orphelines)

6.3 Catégories

2107 modalités présentent dont 1754 orphelines. Voir **Figure 5**

Figure 5 – *Les catégories présentent dans la base de données sous forme de nuage de mot (wordcloud)*

6.4 Valeurs énergétiques des produits

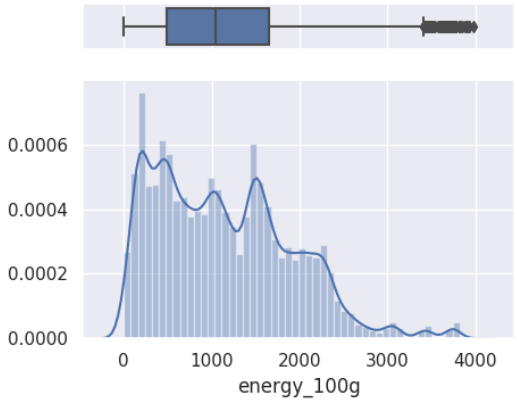

Figure 6 – *Distribution des valeurs énergétiques des produits dans la base de données.*

6.5 Nutriments aux 100g

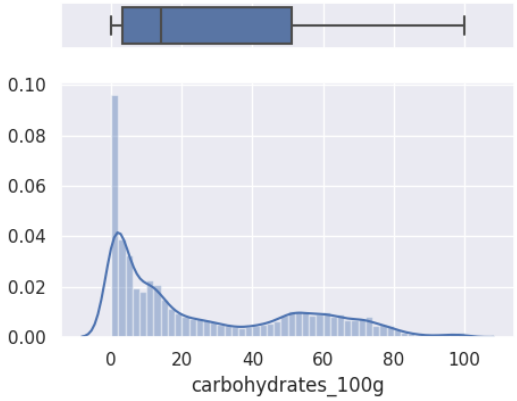

Figure 7 – *Distribution de la quantité de glucides aux 100g*

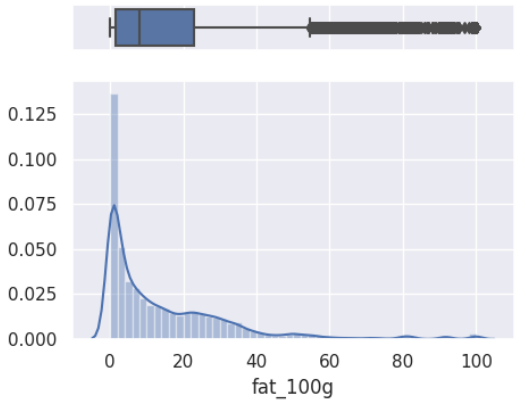

Figure 8 – *Distribution de la quantité de graisse aux 100g*

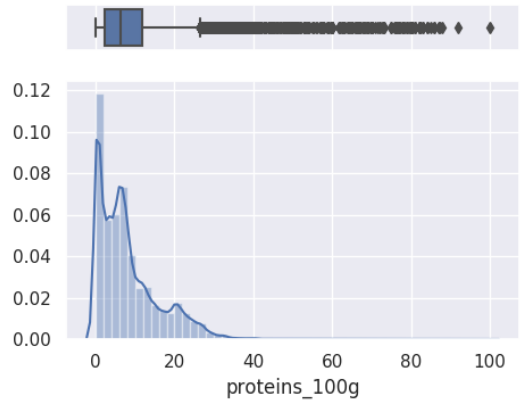

Figure 9 – *Distribution de la quantité de protéines aux 100g*

7 Analyses multivariées

Toujours dans l'objectif de trouver un produit équivalent, il est important de vérifier que pour chaque catégorie (ici groupe Programme national nutrition santé PNNS) il existe des produits pour chaque nutriscore (de A à E).

7.1 Nutriscore et marques

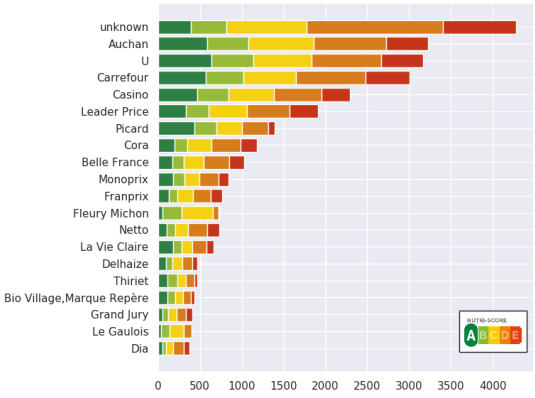

Figure 10 – *Repartition du nutriscore au sein des marques*

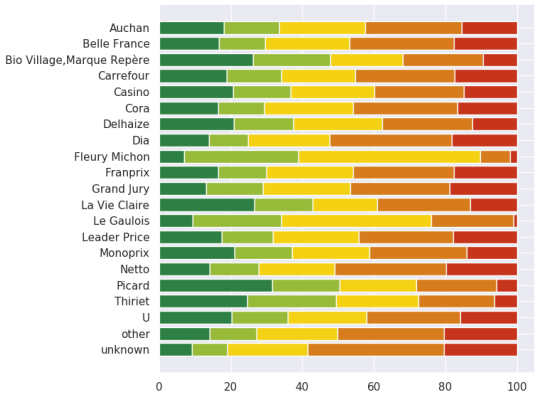

Figure 11 – *Repartition du nutriscore au sein des marques (normalisé en fréquence)*

7.2 Nutriscore et groupes PNNS

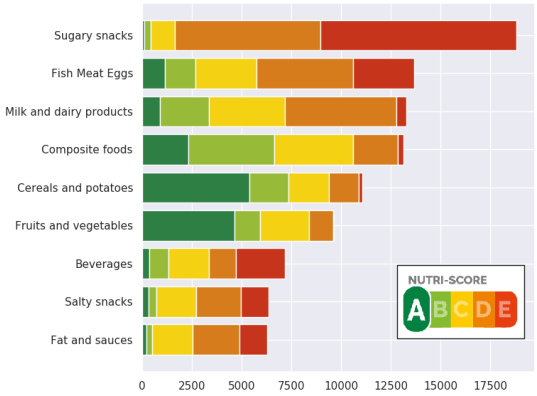

Figure 12 – *Repartition du nutriscore au sein des groupes PNNS*

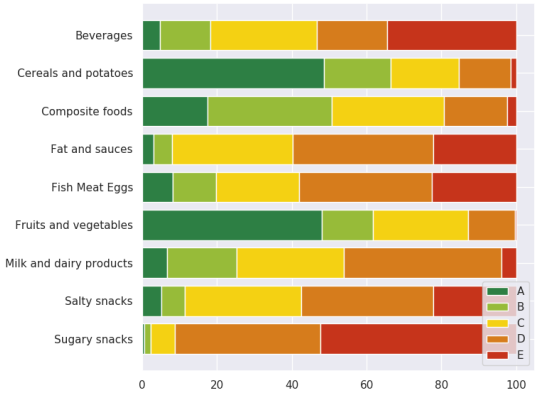

Figure 13 – *Repartition du nutriscore au sein des groupes PNNS (normalisé en fréquence)*

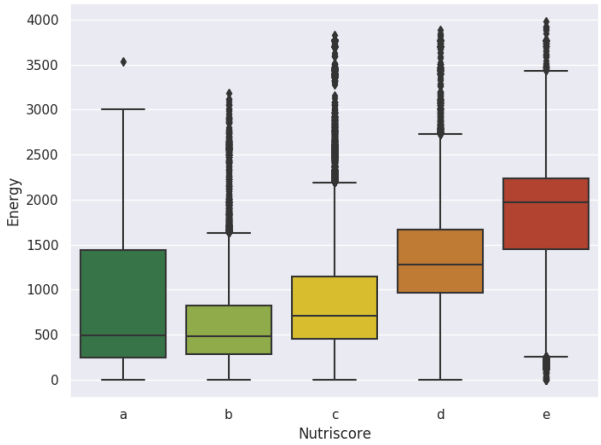

Figure 14

Table 1 – *Tableau de contingence groupe PNNS et Nutriscore*

nutriscore_grade pnns_groups_1	a	b	c	d	e
Beverages	335	972	2038	1359	2473
Cereals and potatoes	5375	1984	2022	1520	168
Composite foods	2303	4347	3946	2239	311
Fat and sauces	192	308	2028	2367	1393
Fish Meat Eggs	1121	1564	3041	4874	3071
Fruits and vegetables	4621	1317	2449	1215	27
Milk and dairy products	871	2478	3814	5615	517
Salty snacks	319	401	1981	2239	1412
Sugary snacks	116	322	1181	7337	9848
unknown	487	737	1431	2029	782
Sum	15740	14430	23931	30794	20002

7.3 Valeurs énergétiques des aliments et nutriscore

7.4 Nutriscore / valeurs énergétiques / composition des produits

voir **Figure 15**

8 Conclusion

9 Perspectives

10 Liens internet

- https://github.com/tgrandjean/OC-sante-publique-france
- https://world.openfoodfacts.org/data
- https://tgrandjean-oc-reports.s3.eu-west-3.amazonaws.com/openfoodfacts/profiling_report.html

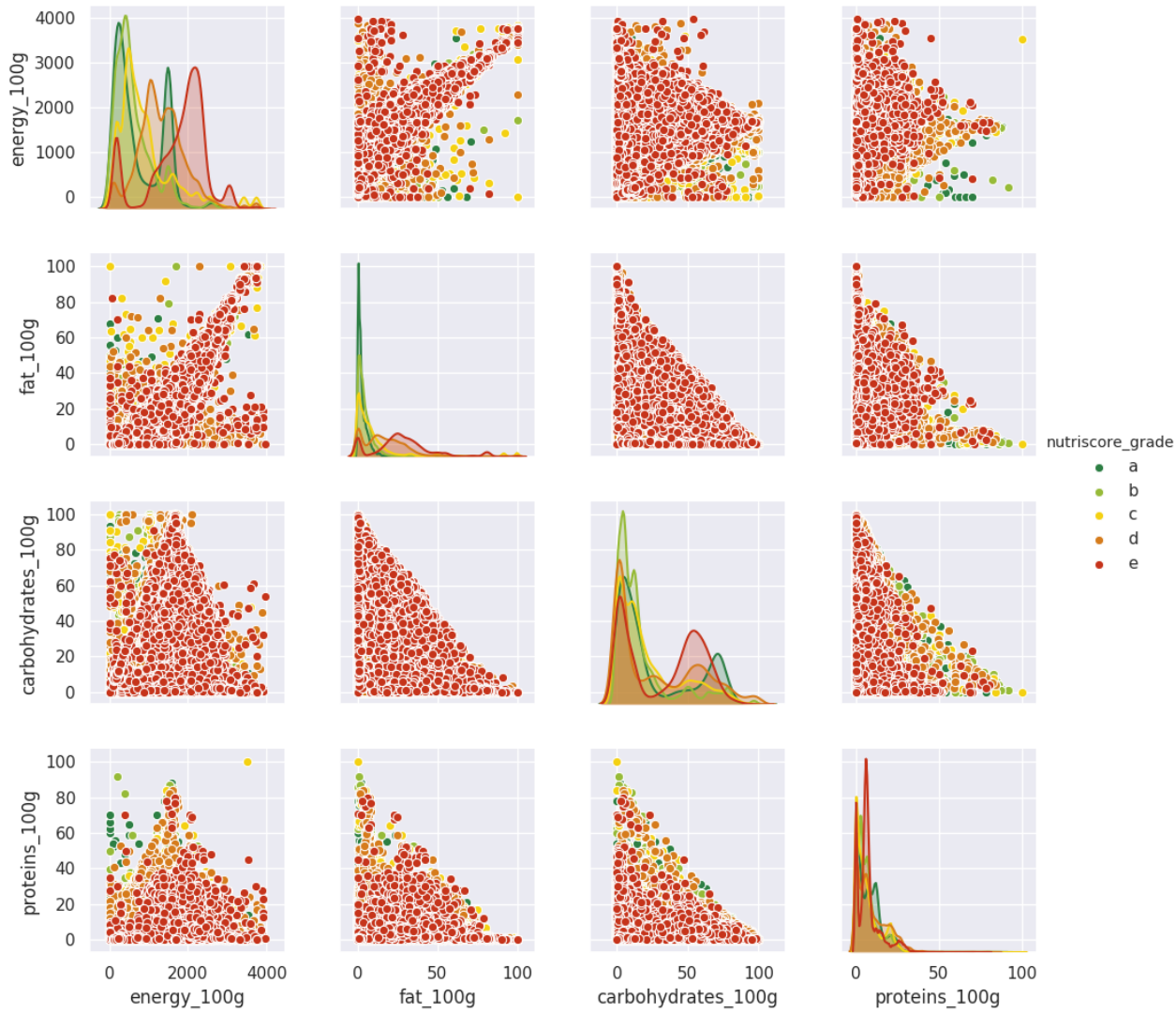

Figure 15

Bibliographie

[1] *Inserm Dossier information : Obésité.* https://www.inserm.fr/information-en-sante/dossiers-information/obesite. Accessed : 2020-01-23.

[2] *OpenRefine Github-wiki : Clustering in Depth.* https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth. Accessed : 2020-01-23.

[3] *Organisation mondiale de la santé Thème de santé, Obésité.* https://www.who.int/topics/obesity/fr/. Accessed : 2020-01-23.