
Concevez une application au service de la santé publique

Thibault Grandjean¹

¹ Étudiant auprès d'Openclassrooms

4 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 a presque triplé depuis 1975.[1] [2] 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 maladies cardiovasculaires et le cancer. [2]

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 meilleures propriétés nutritionnelles.

2 Objectifs

L'objectif principal est d'étudier la faisabilité d'une telle application. Pour ce faire, une analyse minutieuse 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 meilleures 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 langage 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 est assez conséquente (2.2 Go), elle contient (à l'heure actuelle ¹) 1 120 752² 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.

5.4 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

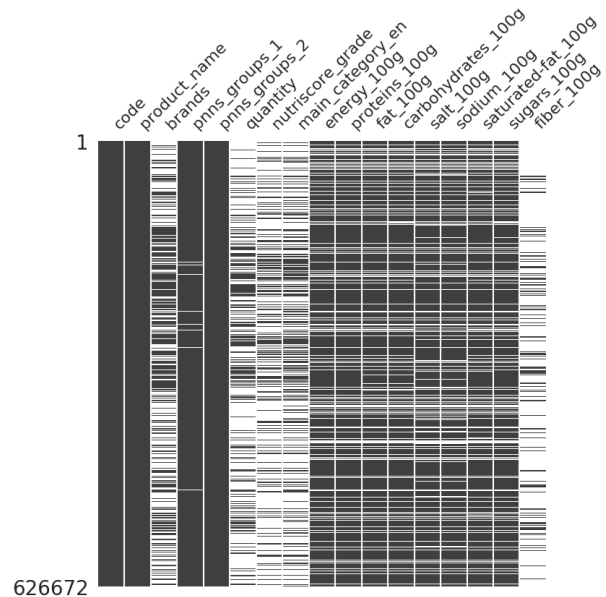


FIGURE 1 – Taux de remplissage des champs

5.5 Champs contenant des valeurs textuelles

Pour les champs contenant des valeurs textuelles, on traite les champs par méthode de «clustering». L'algorithme utilisé est l'algorithme de «Key collision». («StringFingerPrint» méthode par défaut d'OpenRefine³)

6 Analyses univariées

6.1 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

1. 23 janvier 2020

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

3. Le module `src.utils.string_handler` contient également le code nécessaire à une collision de clés par la méthode des n-grams.

6.2 Répartition du nutriscore

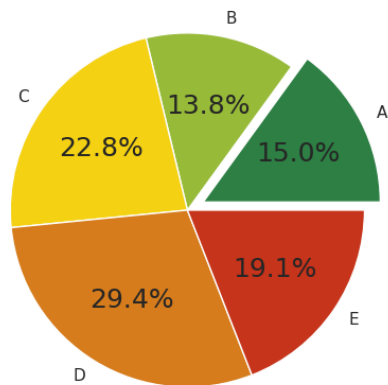


FIGURE 2 – Répartition du nutriscore des produits

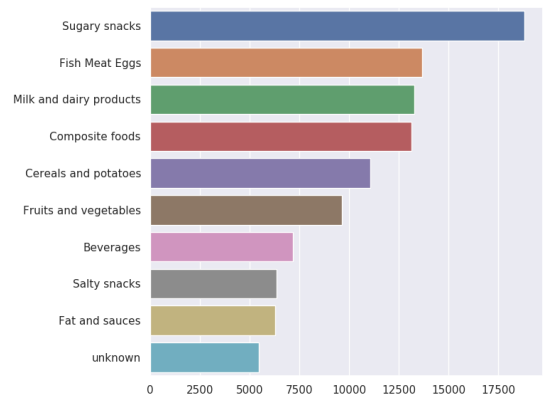


FIGURE 4 – Répartition des produits en fonction du groupe PNNS

6.4 Catégories

voir Figure 5

6.3 Marques, groupes PNNS et 6.5 Valeurs énergétiques des produits

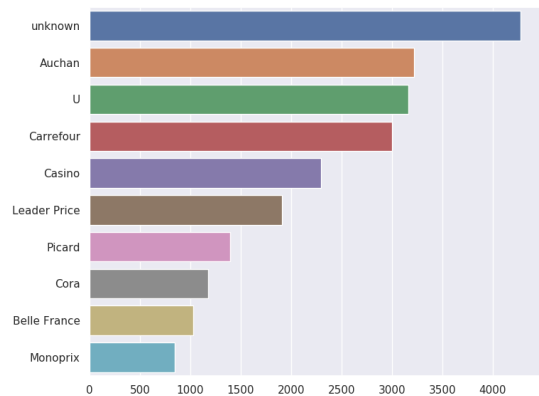


FIGURE 3 – Répartition des produits en fonction des marques

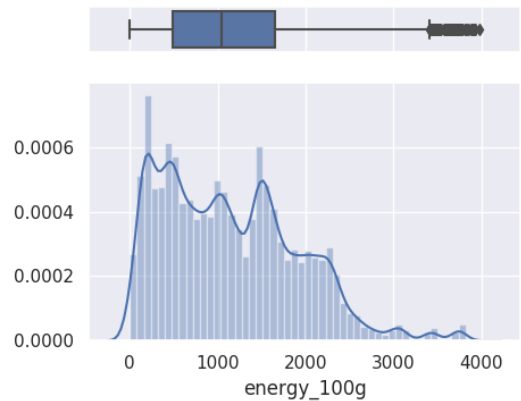


FIGURE 6 – Distribution des valeurs énergétiques des produits dans la base de données.



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

6.6 Nutriments aux 100g

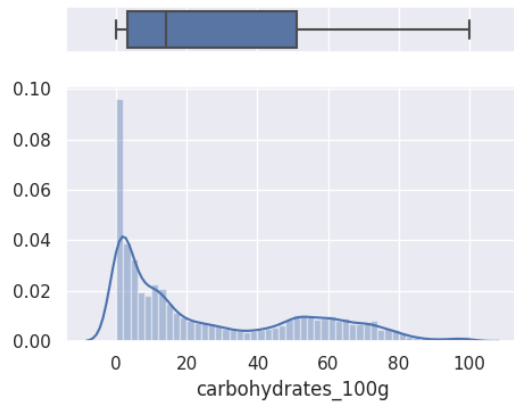


FIGURE 7 – Distribution de la quantité de glucides aux 100g

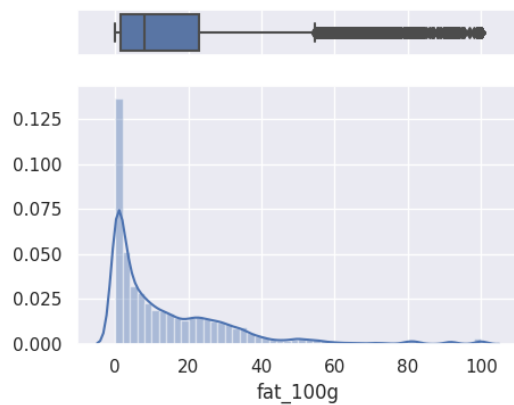


FIGURE 8 – Distribution de la quantité de graisse aux 100g

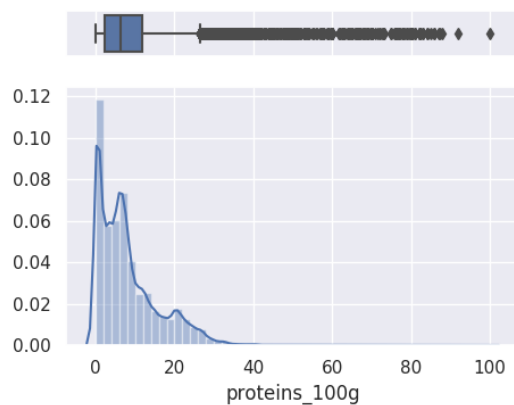


FIGURE 9 – Distribution de la quantité de protéines aux 100g

7 Analyses multivariées

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

7.1 Nutriscore et marques

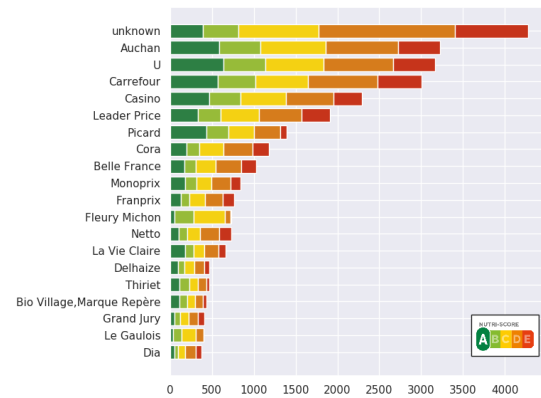


FIGURE 10 – Repartition du nutriscore au sein des marques

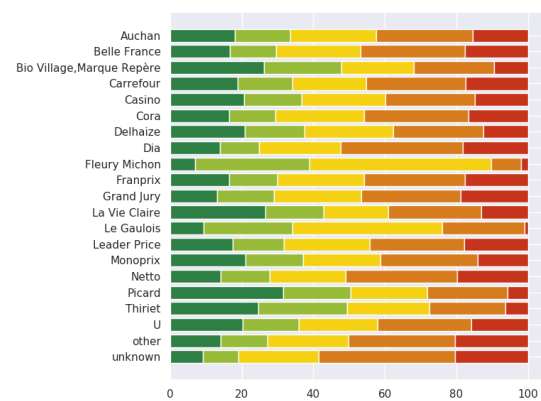


FIGURE 11 – Repartition du nutriscore au sein des marques (normalisé en fréquence)

7.2 Nutriscore et groupes PNNS

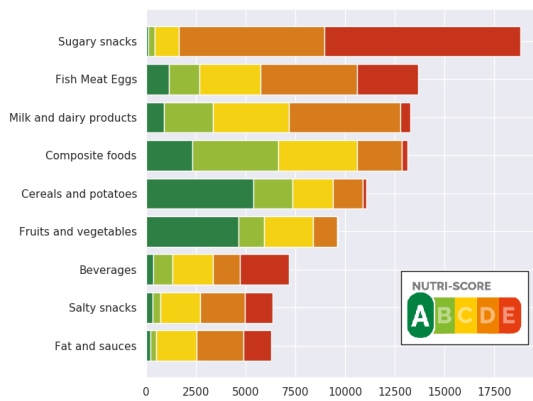


FIGURE 12 – Répartition du nutriscore au sein des groupes PNNS

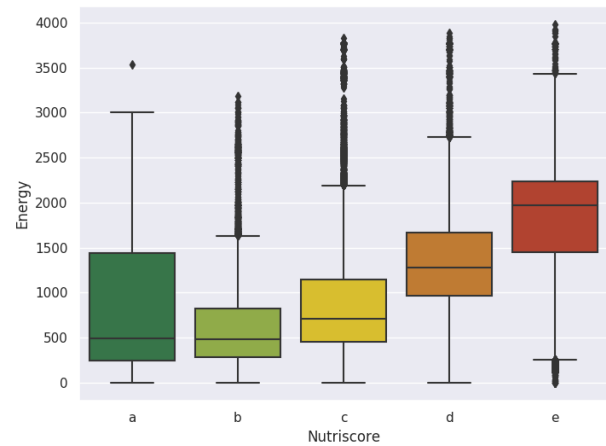


FIGURE 14

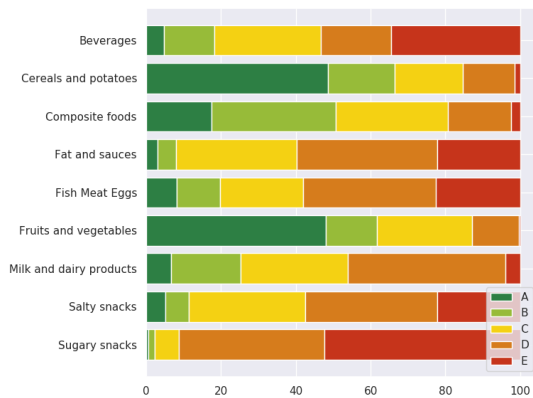


FIGURE 13 – Répartition du nutriscore au sein des groupes PNNS (normalisé en fréquence)

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

Bibliographie

- [1] Inserm Dossier information : Obésité. <https://www.inserm.fr/information-en-sante/dossiers-information/obesite>. Accessed : 2020-01-23.
- [2] Organisation mondiale de la santé Thème de santé, Obésité. <https://www.who.int/topics/obesity/fr/>. Accessed : 2020-01-23.

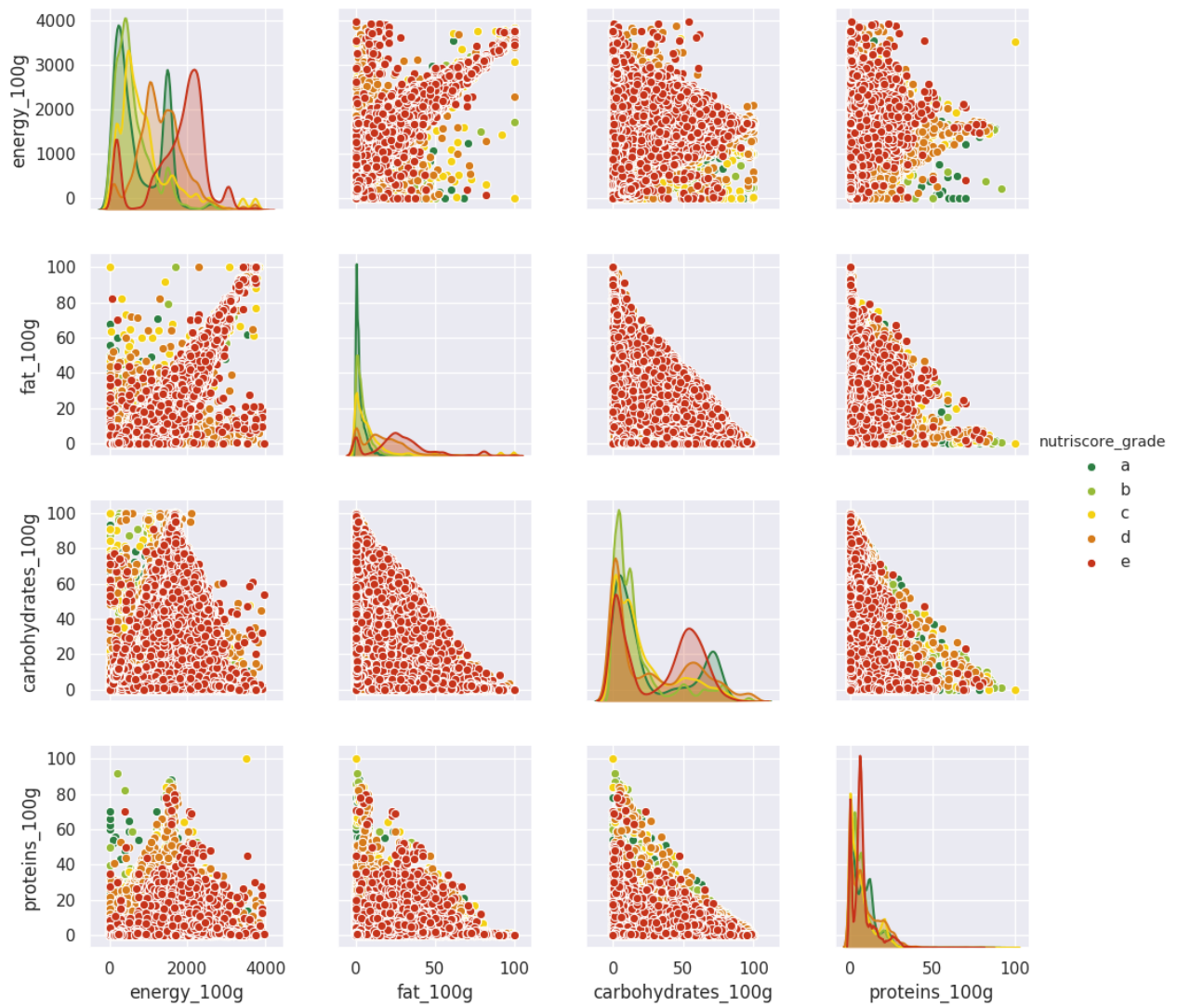


FIGURE 15