MyHealth Application de Santé

Présenté par :
Dabidin Audam Keshika
Le 26/01/23

01 Contexte et idée d'application

02 Le traitement des données

03 L'analyse des données

04 Faisabilité de l'application

Introduction

1. **Contexte :**

   Santé Publique France → appel à projets pour trouver des idées innovantes d'applications en lien avec l'alimentation.

2. **Proposition de l'application MyHealth :**

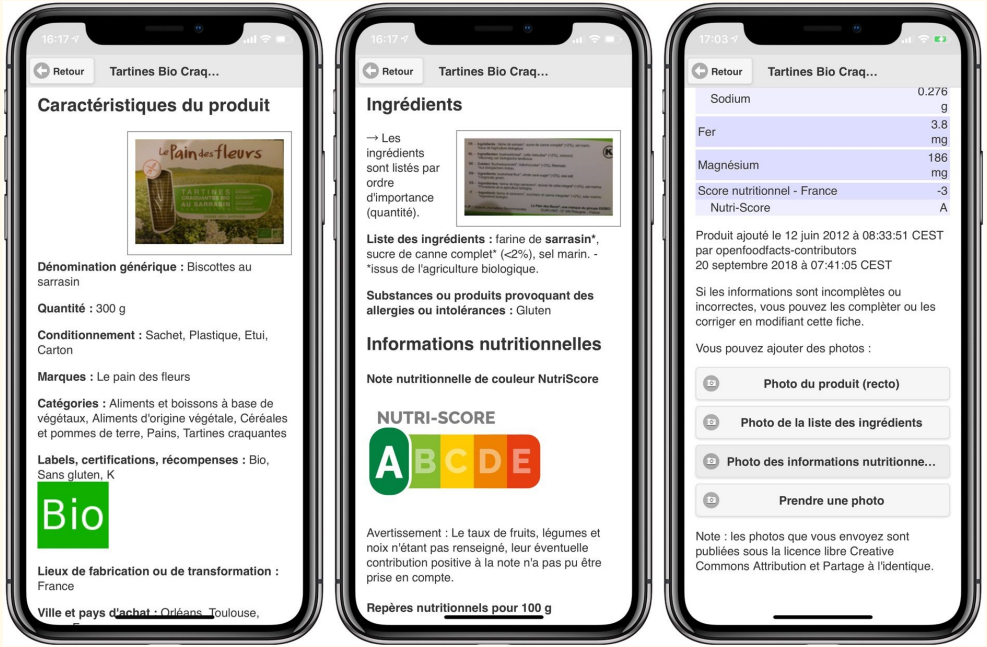
- Proposition d'un régime adapté selon le profil.
- Recherche des produits sur Open Food Facts.
- Analyse statistique pour trouver un équilibre et les bonnes quantités de nutriments.
- Proposition des repas et des alternatives de repas en tenant en compte le nutri score et le profil de la personne.

Présentation des données

- **Source :** Open Food Facts

- Plusieurs onglets pour un produit.
- Plusieurs informations.

- **Données disponibles :**
- 320 772 produits.
- 162 colonnes.
- Types : Objet et Float



Présentation des données

Nom du fichier	Nombre lignes	Nombre de colonnes	Taux de remplissage moyen	Doublons
products.csv	320772	162	23.8%	22

- **Informations générales :** code, nom du produit, url…
- **Tags :** localisation, origine…

- **Informations nutritionnelles :** quantité, nutri score, ingrédients, additifs…

Traitement des données

- **Suppression des doublons -** filtrage par code

Nom du fichier	Nombre lignes	Nombre de colonnes	Taux de remplissage moyen	Doublons
products.csv	320772	162	23.8%	0

- **Traitement des valeurs aberrantes et des valeurs manquantes**
- **Traitement final**

Traitement des valeurs aberrantes et des valeurs manquantes

- Création d'une fonction qui calcule le pourcentage de valeurs manquantes par variable. Voici un aperçu du tableau :

Nom de la colonne/ Variable	Pourcentage de valeurs manquantes (%)	Nombre de valeurs manquantes
water-hardness_100g	100.00	320750
labels_fr	85.46	274105
additives	22.40	71845

Traitement des valeurs aberrantes et des valeurs manquantes

- Suppression des colonnes avec 100% de données manquantes.

	Pourcentage Valeurs Manquantes	Nombre de valeurs manquantes
vitamin-a_100g	57.11	183196
iron_100g	56.21	180288
vitamin-c_100g	56.08	179883
calcium_100g	56.02	179700
trans-fat_100g	55.32	177452
cholesterol_100g	55.08	176660

- Compromis entre colonnes qui possèdent énormément de valeurs manquantes et informations apportées - **seuil à 60%.**

Nom du fichier	Nombre lignes	Nombre de colonnes	Taux de remplissage moyen	Doublons
products.csv	320750	42	78.1%	0

Traitement des valeurs aberrantes et des valeurs manquantes

- Choix sur les colonnes à conserver :

Certaines colonnes présentent le même type d'information en Français comme en Anglais.

Ex. : countries et countries_fr.

- Conservation des colonnes utiles pour l'application - **27 colonnes retenues:**

- **Variables descriptives :** url, **product_name,** states_fr, serving_size, ingredients_from_palm_oil_n, ingredients_text, countries_fr, brands, creator.

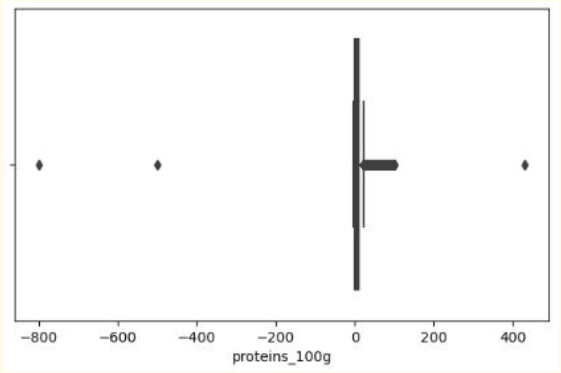
- **Variables quantitatives (pour 100g) : nutri score, vitamines a et c, protéines, glucides, acides gras et acides gras saturés, minéraux (calcium, sodium, fer), fibres, sucres, énergie, additifs.**

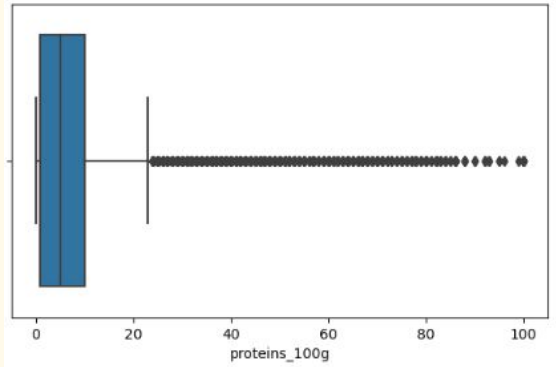
- **Autres variables discrètes : nutri grade**, created_datetime, last_modified_datetime.

Traitement des valeurs aberrantes et des valeurs manquantes

- Suppression des variables aberrantes : seuil à 100g pour les nutriments, seuil à 3900 kcal/100g pour l'énergie. Pour les valeurs négatives en protéines, remplacement par zéro.

  Exemple :





- Suppression des observations où toutes les valeurs quantitatives sont manquantes.

Traitement des valeurs aberrantes et des valeurs manquantes

- **Variables descriptives :** remplissage avec le mot ‘Unknown’ .
- **Variables quantitatives (pour 100g) :**
  - Minéraux, fibre, graisses alimentaires, vitamines a et vitamine c en partie : remplacement par zéro.
  - Vitamine c : recherche par mot clef ‘jus’ par exemple et remplacement par la moyenne.
  - Protéines, sucres, glucides, énergie et additifs : remplissage par algorithme d’imputation (Iterative Imputer).

- **Autres variables discrètes :** nutri grade, created_datetime, last_modified_datetime.

Traitement des valeurs aberrantes et des valeurs manquantes

- **Calcul du Nutri score et du estimation du Nutri grade:**

Nutri score = Points N - Points P

Points	Densité énergétique (kJ/100g)	Sucres (g/100g)	Acides gras saturés (g/100g)	Sodium (mg/100g)[1]
0	≤ 335	≤ 4,5	≤ 1	≤ 90
1	> 335	> 4,5	> 1	> 90
2	> 670	> 9	> 2	> 180
3	> 1005	> 13,5	> 3	> 270
4	> 1340	> 18	> 4	> 360
5	> 1675	> 22,5	> 5	> 450
6	> 2010	> 27	> 6	> 540
7	> 2345	> 31	> 7	> 630
8	> 2680	> 36	> 8	> 720
9	> 3015	> 40	> 9	> 810
10	> 3350	> 45	> 10	> 900

Points	Fruits, légumes, légumineuses, fruits à coques, huiles de colza, de noix et d'olive (%)	Fibres (g/100g)	Protéine (g/100g)
0	≤ 40	≤ 0,9	≤ 1,6
1	> 40	> 0,9	> 1,6
2	> 60	> 1,9	> 3,2
3	-	> 2,8	> 4,8
4	-	> 3,7	> 6,4
5	> 80	> 4,7	> 8,0

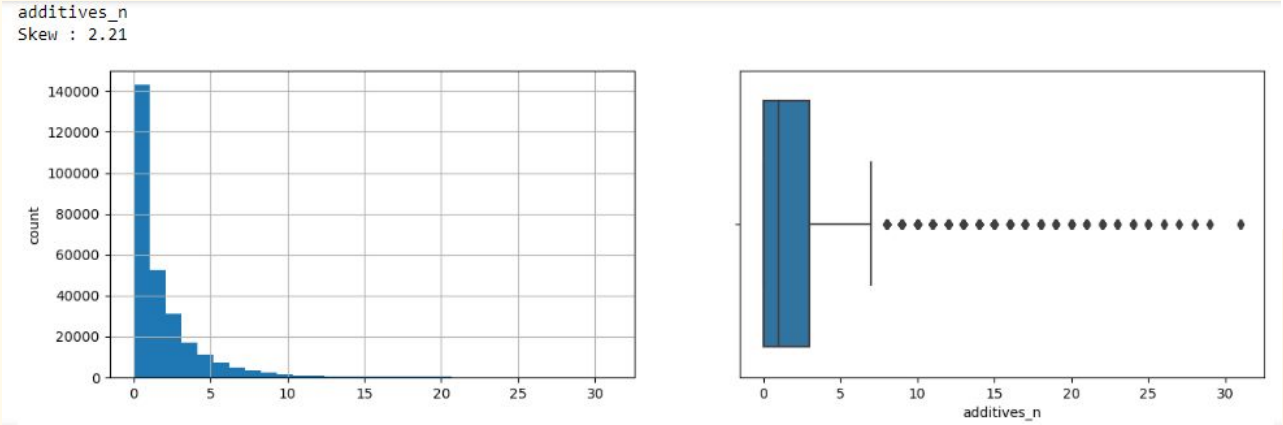
- Suppression et remplissage final de nouvelles valeurs aberrantes.

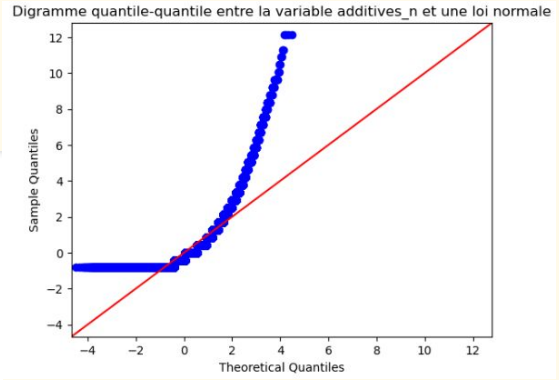
Analyse des données - Analyse Univariée

- **Description générale des données (276028 produits et 27 colonnes) :**
- On a en moyenne : 2 additifs, 1100 Kcal/100g d'énergie, 11g de graisse alimentaire, 34g de glucides, 16g de sucres, 2g de fibres, 7g de protéines et un nutri score d'environ 8.
- Maximums : énergie - 3887 Kcal/100g et 31 additifs au total.
- Pays le plus consommateur : États Unis.
- Il y a 4 et 5 types de valeurs pour les variables : 'ingredients_from_palm_oil_n' (0,1,2, Inconnu) et 'nutrition_grade_fr' (a,b,c,d,e).

Analyse des données - Analyse Univariée

- **Box plot et distribution des additifs :**

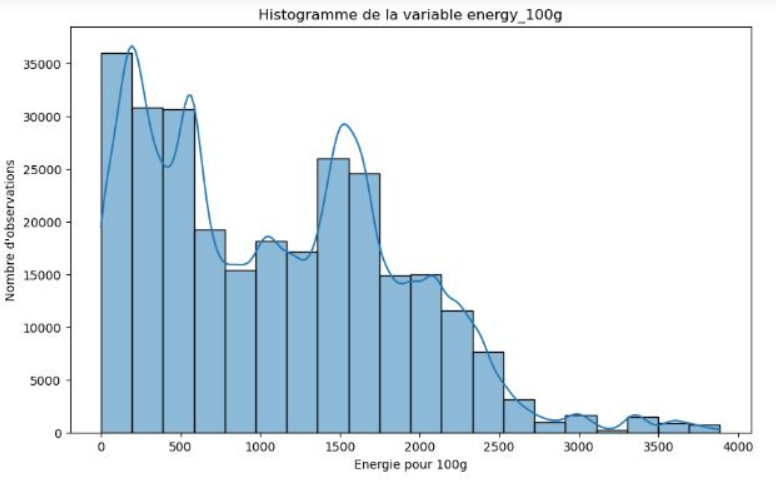


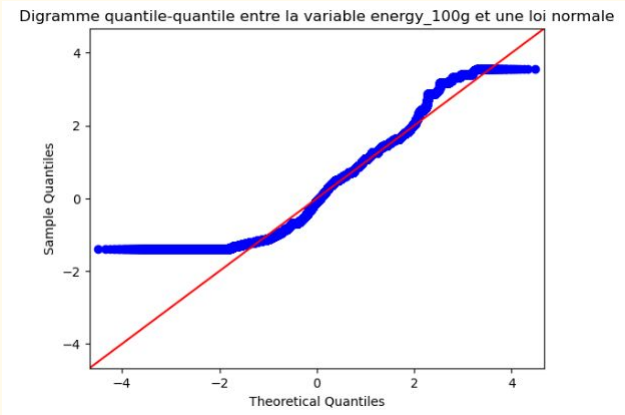


Observations : La distribution est asymétrique. Le Skewness est supérieur à 0, alors le dataset est skewed sur la droite. C'est à dire que la majorité des données se trouvent sur la gauche et les outliers se trouvent sur la droite. Le premier quantile est à 0, la médianne est à 1 et le troisième quantile est à 3 additifs. L'interval supérieur est à 7 additifs. Les valeurs s'écartent de la loi normale. Test de Shapiro-Wilk : Indicateur statistique: 0.774 et p valeur: 0.0. L'hypothèse de normalité peut être rejétée.

Analyse des données - Analyse Univariée

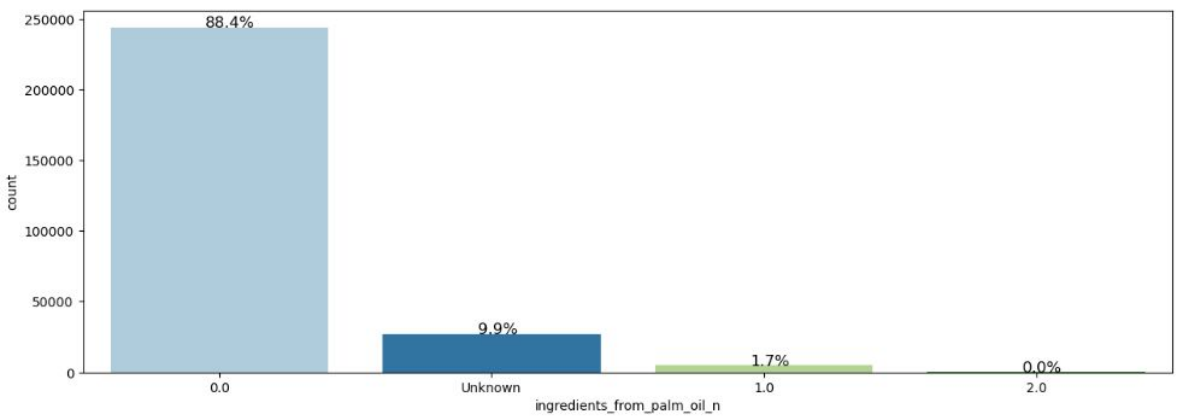
- **Énergie :**

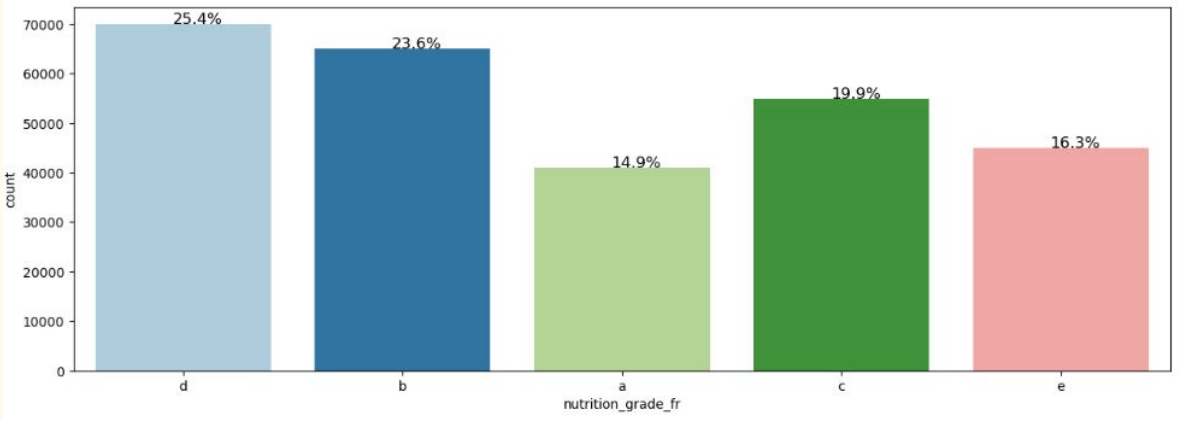




Observations : La distribution est asymétrique et unimodale. Les observations sont regroupées à gauche. Le Skewness est supérieur à 0, alors le dataset est skewed sur la droite. C'est à dire que la majorité des données se trouvent sur la gauche et les outliers se trouvent sur la droite.Les valeurs s'écartent de la loi normale.

Analyse des données - Analyse Univariée

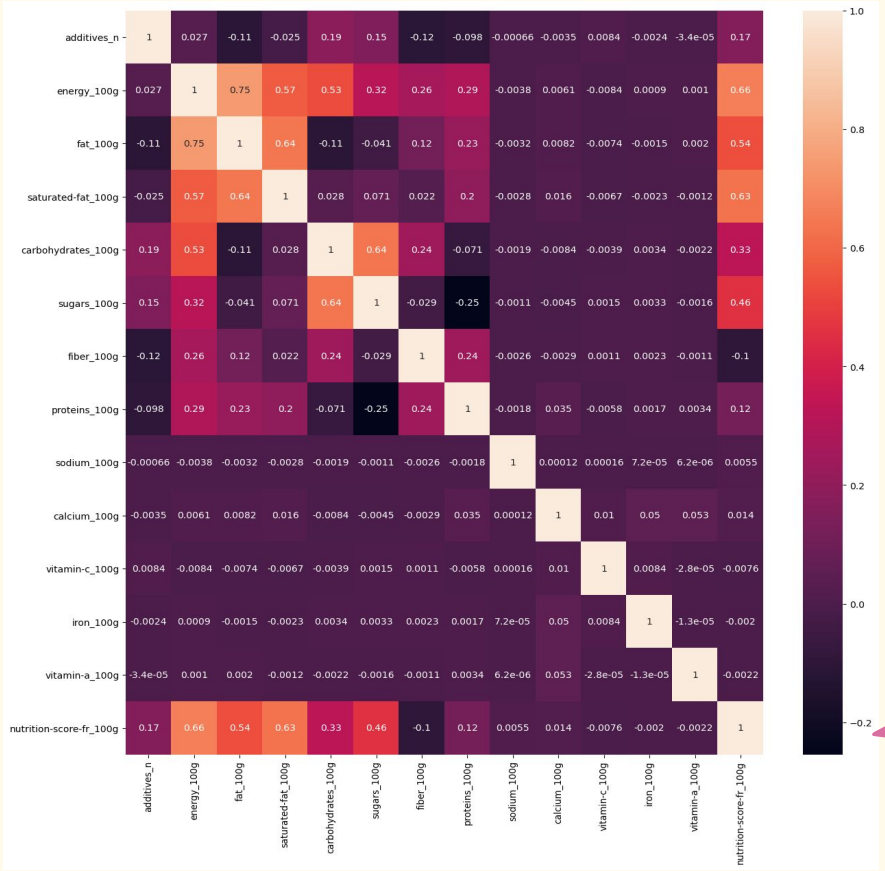




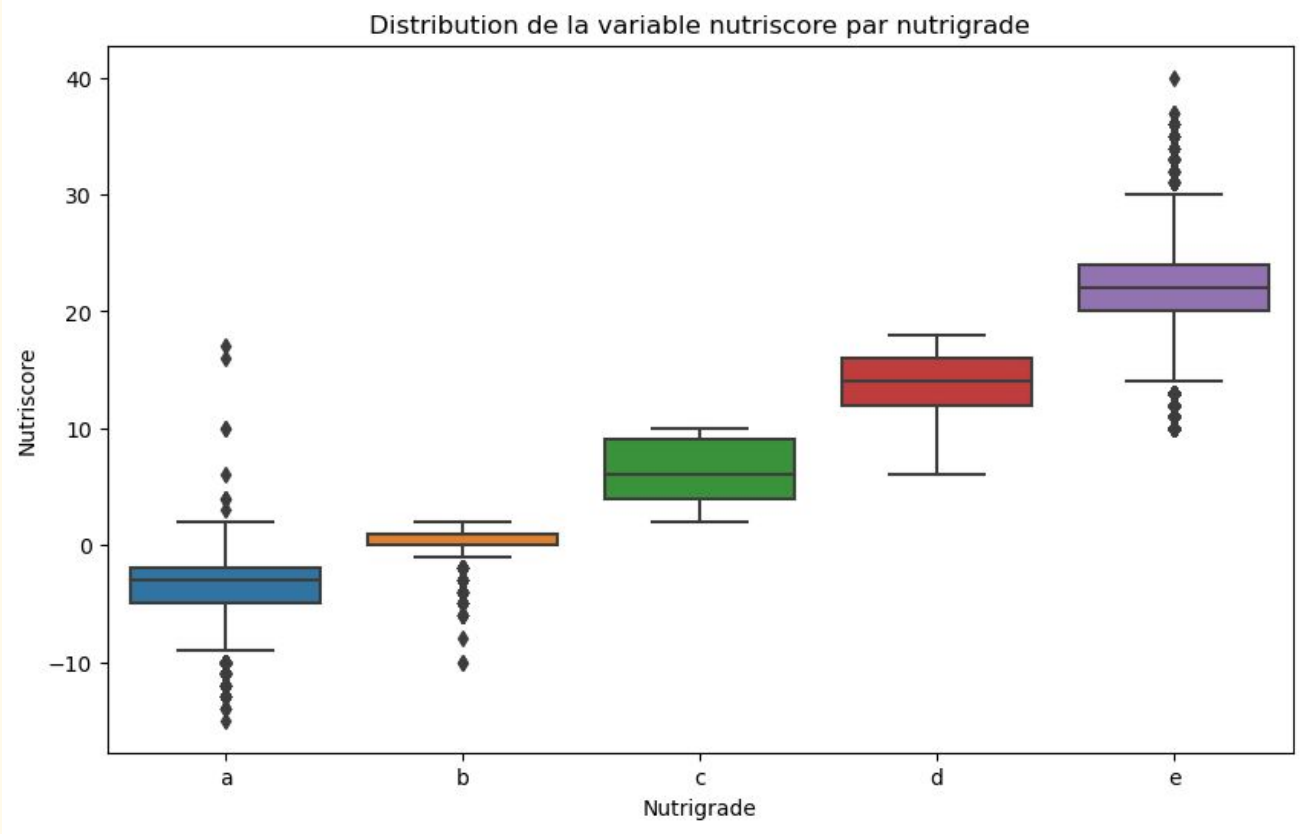
Analyse des données - Analyse Multivariée

- Les variables qui sont corrélées entre-elles de manière significative sont :
- energy_100g et nutri-score-fr_100g
- fat_100g et nutri-score-fr_100g
- saturated-fat_100g et nutri-score-fr_100g
- sugars_100g et nutri-score-fr_100g

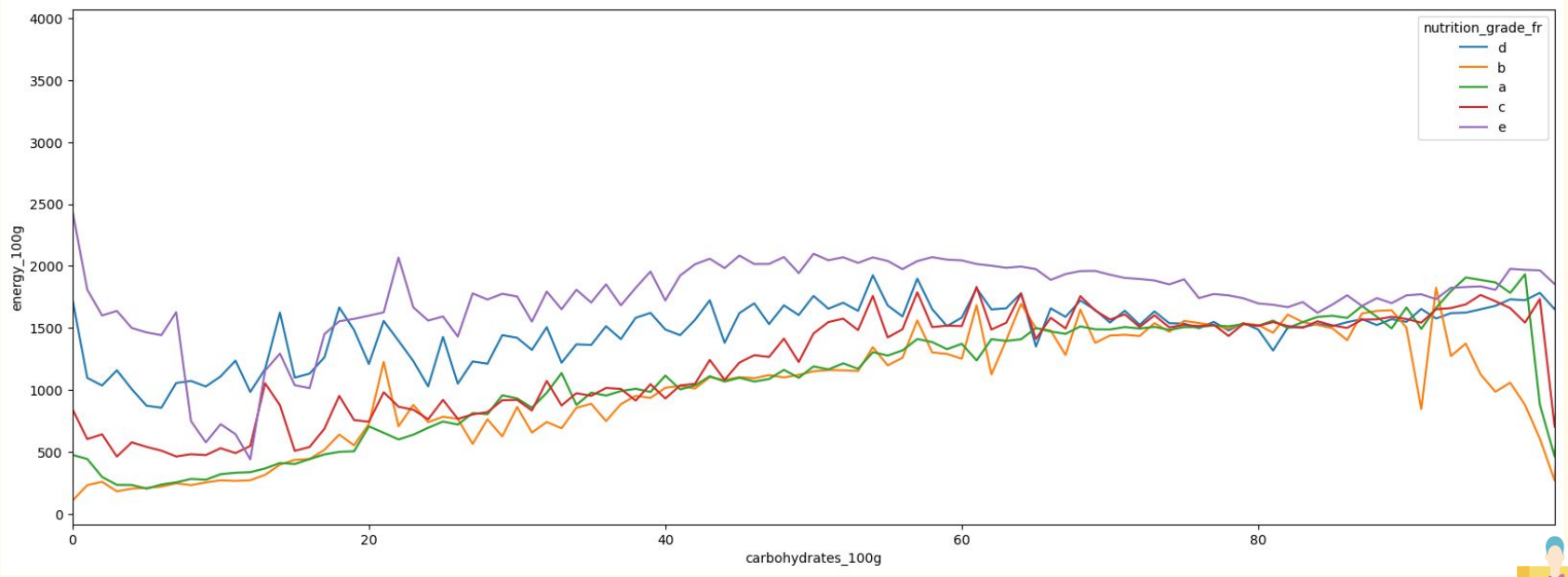
Confirmation avec coefficient de Pearson p valeur >0.



Analyse des données - Analyse Bivariée

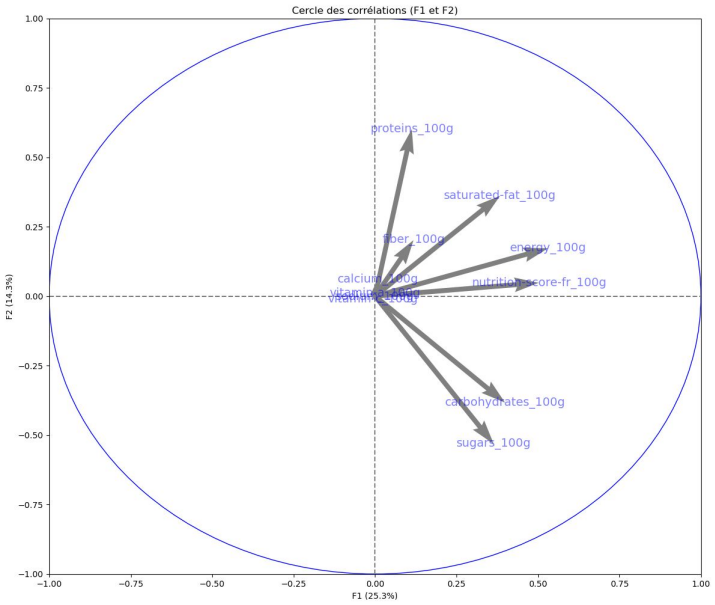


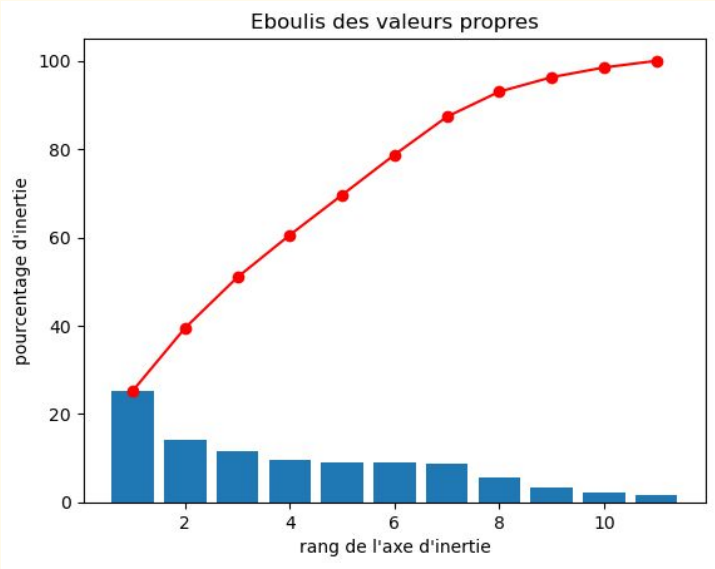
Analyse des données - Analyse Multivariée



Observations : On constate que l'énergie augmente quand la quantité de glucides consommée augmente, quelque soit le nutrigrade de l'aliment. En petite quantité, en dessous de 100g, on peut distinguer que plus le nutrigrade augmente, plus on reçoit de l'énergie pour la même quantité d'aliments consommés.

Analyse des données - Analyse Multivariée - ACP





Observations : Les variables corrélées sont energy_100g et nutri-score_100g, carbohydrates_100g et sugars_100g. La composante 1 est principalement expliquée par energy_100g et nutri-score_100g. La seconde composante est principalement expliquée par les protéines et les fibres mais plus par les protéines.

Il apparaît qu'il est nécessaire d'avoir recours à 7 composantes principales, afin de capter au moins 80% de l'inertie.

Analyse des données - Analyse Multivariée - ACP

Observations :

PC1 : energy_100g, fat_100g saturated-fat_100g, carbohydrates_100g et nutrition-score-fr_100g influencent principalement cette composante. On peut relier cela aux points N utilisés calculés pour le nutri score.

PC2 : On a les additifs, les sucres et les glucides qui influencent négativement cette composante. Les acides gras et les protéines influencent positivement cette composante.

PC 3: Les fibres et les protéines et le nutri score sont reliés dans cette composante. On voit que les fibres et les protéines influencent positivement la composante alors que le nutri-score est anti-corrélé avec cette composante. Cette composante peut être reliée aux points P.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
additives_n	0.040000	-0.310000	-0.220000	0.050000	-0.010000	-0.010000	0.060000
energy_100g	0.520000	0.020000	0.150000	-0.010000	-0.000000	0.000000	0.010000
fat_100g	0.410000	0.350000	-0.140000	-0.030000	-0.010000	-0.000000	0.000000
saturated-fat_100g	0.420000	0.230000	-0.240000	-0.010000	-0.010000	-0.000000	-0.000000
carbohydrates_100g	0.260000	-0.510000	0.350000	0.010000	0.010000	0.000000	-0.000000
sugars_100g	0.240000	-0.550000	-0.000000	0.010000	0.010000	0.000000	-0.020000
fiber_100g	0.110000	0.090000	0.740000	-0.020000	-0.010000	0.010000	0.020000
proteins_100g	0.150000	0.380000	0.330000	0.040000	0.000000	0.010000	0.020000
sodium_100g	-0.000000	-0.000000	-0.010000	0.000000	-0.020000	1.000000	0.010000
calcium_100g	0.010000	0.030000	0.000000	0.700000	0.060000	0.000000	-0.050000
vitamin-c_100g	-0.010000	-0.010000	-0.000000	0.150000	-0.650000	-0.020000	0.740000
iron_100g	0.000000	-0.000000	0.020000	0.490000	-0.470000	-0.000000	-0.550000
vitamin-a_100g	-0.000000	0.010000	0.000000	0.490000	0.590000	0.000000	0.380000
nutrition-score-fr_100g	0.470000	-0.090000	-0.270000	0.010000	-0.000000	0.010000	-0.000000

Idée d'application

Scan

L'utilisateur entre les informations sur le repas qu'il a mangé et sur la quantité de sommeil effectuée ainsi que la quantité de sport effectuée.

Analyse

MyHealth peut se baser sur le calcul du nutri-score et du nutri-grade à partir des axes de composantes principales. L'application analysera donc les ingrédients et leurs contributions positive ou négative à l'axe ou aux axes principaux. On peut également intégrer un comptage de kcal. Pour l'énergie.

Suggestions

Un algorithme de clustering comme le KNN pourra être utilisé pour suggérer des ingrédients en alternatives (en calories équivalents) au cas où l'alimentation peut être améliorée.

Retour

Un retour à l'utilisateur, de suggestions de produits à changer, ou à ajouter dans le régime alimentaire peut être intégré avec d'autres features comme le sommeil ou le sport dans l'application pour compléter.

Conclusion

1. MyHealth sera une application qui analysera le profil de l'utilisateur pour proposer le régime adapté et les compléments pour améliorer sa santé.
2. L'application pourra utiliser les données des produits pour effectuer son analyse et trouver les produits adaptés.
3. On peut également y intégrer d'autres features comme le sport et le sommeil.

Sur les données :

1. Nombre de variables dans les données d'origine, suffisantes, mais taux de complétion variable.
2. Compléter avec plus de données de produits et avis d'expert métier.