DuCoeuràLAsiette

newsletter hebdomadaire alimentaire pour les personnes atteintes de maladie(s) cardiaque(s)

Chaque semaine, un mail est envoyé aux différents inscrits comprenant des recommandations de produits alimentaires dit "bon pour le cœur". C'est-à-dire faible en sel, pauvre en graisse et riche en fibres. Elle sera composée d'une liste de cinq produits différents chaque semaine.
Le but est de ne pas être redondant, car la newsletter est hebdomadaire.

I. Opération de nettoyage

a. valeurs manquantes
b. choix métier
c. valeurs aberrantes
d. valeurs atypiques

II. Analyse exploratoire

a. univarié
b. bivarié
c. multivarié et réduction de dimension
d. résultat de l'application

I. Opération de nettoyage : a. valeurs manquantes

Choix technique :

- **taux de remplissage :** pour les variables, suppression de celles avec plus de **50% de valeurs manquantes**, à l'exception pour les valeurs métiers suivantes : **calcium, trans-fat, iron, cholesterol et fiber**.
- Surpression **des individus avec plus de 80% de valeurs manquantes**

Number of missing fields over 91247 products

	NaN	%
sodium_100g	29035	31.82
salt_100g	29032	31.82
sugars_100g	29091	31.88
carbohydrates_100g	44390	48.65
fiber_100g	45747	50.14
fat_100g	43957	48.17
saturated-fat_100g	29224	32.03
calcium_100g	89014	97.55
trans-fat_100g	90879	99.6
iron_100g	90080	98.72
cholesterol_100g	90850	99.56
nutrition-score-fr_100g	30175	33.07
additives_n	37847	41.48

Number of missing fields over 25587 products

	NaN	%
sodium_100g	0	0.0
salt_100g	0	0.0
sugars_100g	0	0.0
carbohydrates_100g	8	0.03
fiber_100g	674	2.63
fat_100g	0	0.0
saturated-fat_100g	0	0.0
calcium_100g	23698	92.62
trans-fat_100g	25240	98.64
iron_100g	24528	95.86
cholesterol_100g	25226	98.59

I. Opération de nettoyage : a. valeurs manquantes

traitement des valeurs manquantes :

- **suppression de l'individu** si la valeur manquante est **cruciale** pour l'application comme un individu sans nom (*product_name*) nécessaire pour pouvoir les recommander aux abonnés de la newsletter
- **imputation par la médiane**
- **imputation par connaissance métier**, ex : un produit dont la variable fibre n'est pas renseignée peut-être considérée comme un produit sans fibre et donc la valeur manquante peut être imputée comme nulle.

I. Opération de nettoyage : b. choix métier

- **choix métier** : supprimer les variables dont nous n'aurons pas l'usage pour notre application comme *url*, *last_modified_datetime* (dernière date de modification), *brands* (marque) etc
- **filtrage :** ne proposer que des produits vendus en France, supprimer les variables correspondantes après le filtrage

I. Opération de nettoyage : c. valeurs aberrantes

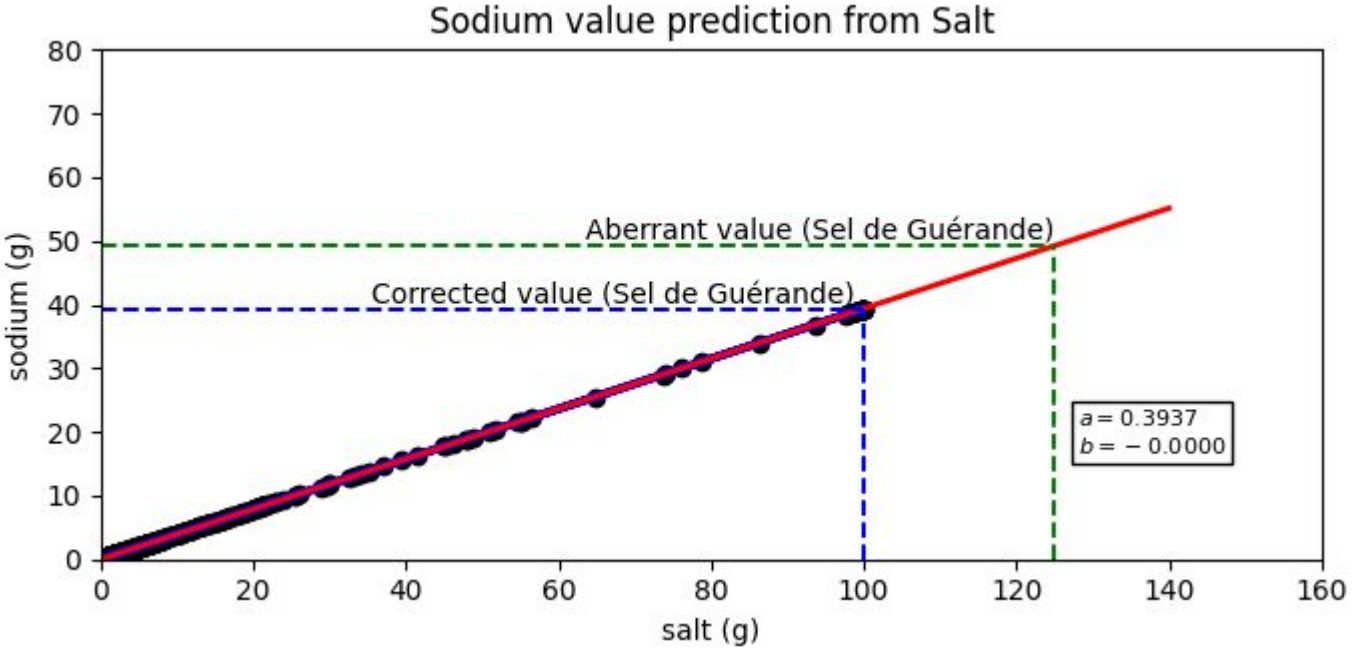
- valeurs qui ne sont pas des nombres
- valeurs que ne sont pas dans l'intervalle de définition
- la somme totale des composantes nutritionnelle supérieure à 100 :('fat_100g', 'sodium_100g', 'carbohydrates_100g' et 'proteins_100g') → **suppression**

Number of aberrant variables

	<0	>100	Not a number	infinity
sodium_100g	0	0	0	0
salt_100g	0	2	0	0
sugars_100g	1	1	0	0
carbohydrates_100g	0	4	0	0
fiber_100g	0	1	0	0
fat_100g	0	1	0	0
saturated-fat_100g	0	1	0	0
calcium_100g	0	0	0	0
trans-fat_100g	0	0	0	0
iron_100g	0	0	0	0
cholesterol_100g	0	0	0	0
sum_100g	0	45	0	0

I. Opération de nettoyage : c. valeurs aberrantes

Imputation d'une valeur aberrante par régression linéaire



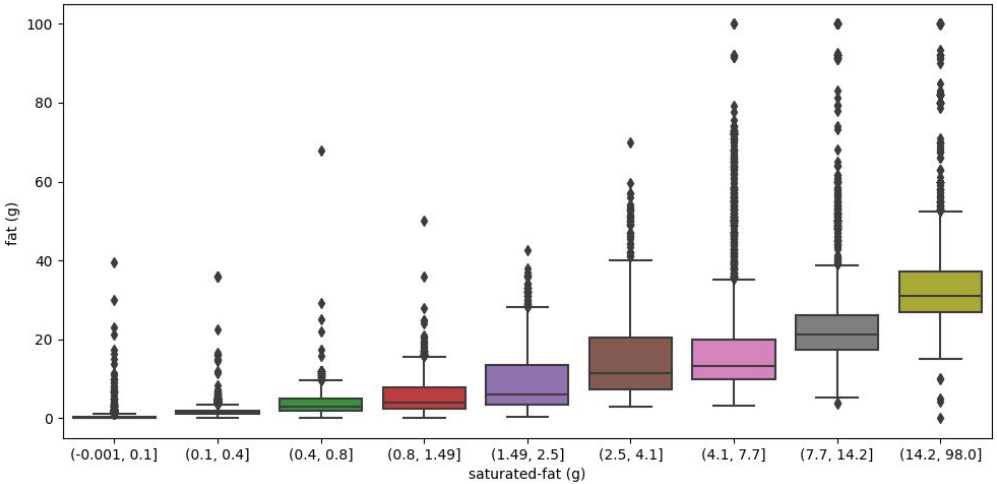
I. Opération de nettoyage : d. valeurs atypiques

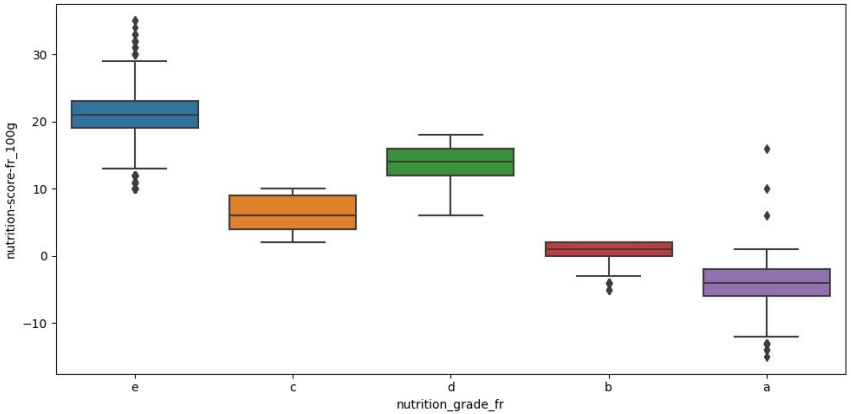
- valeurs considérées atypiques en prenant en compte **l'interquartile** discrétisé par sous-échantillons

Lower = Q1 - 1.5 * IQR

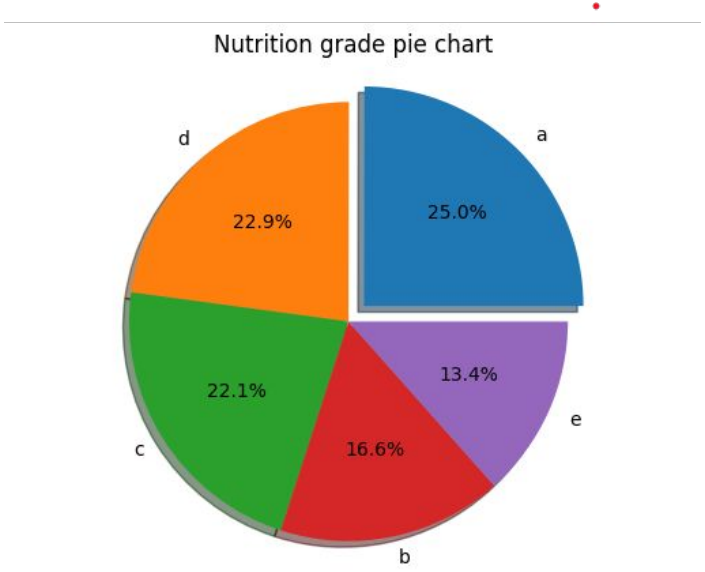
Upper = Q3 + 1.5 * IQR

- imputation par la médiane ou suppression, selon l'importance de la donnée dans l'application



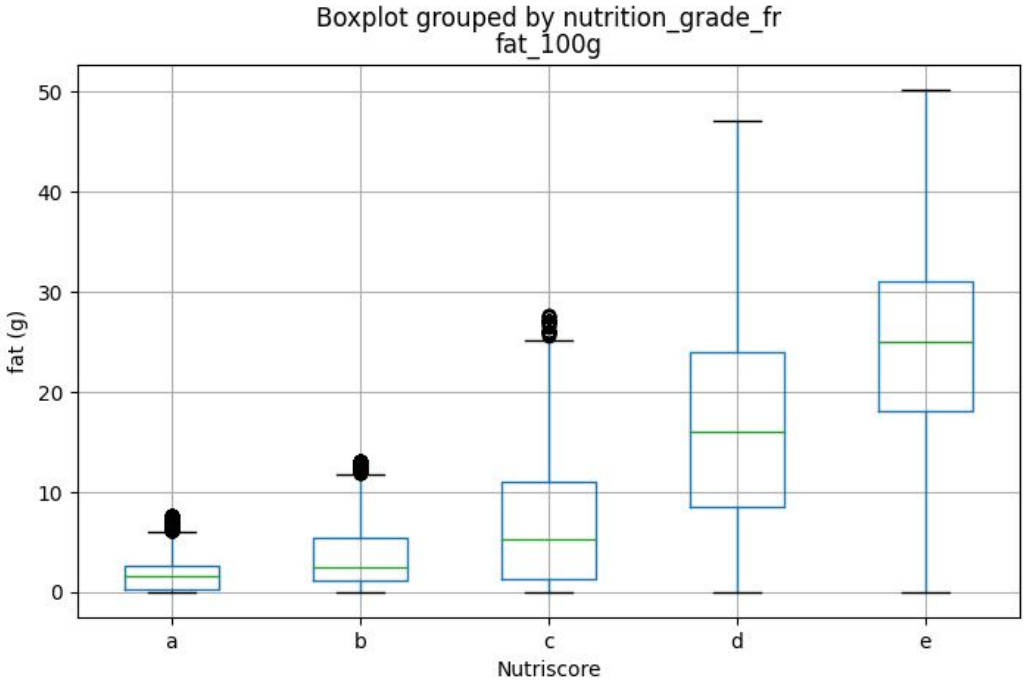


II. Analyse exploratoire
a. univarié



II. Analyse exploratoire
b. bivarié

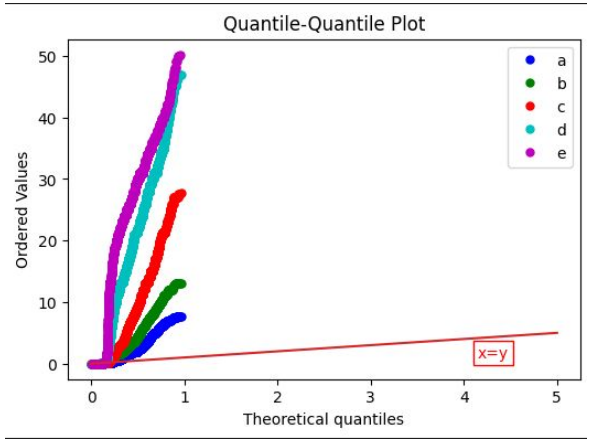
- **coefficient de Pearson (fat_100g, nutrition-score-fr_100g) = 0.709**



Analyse ANOVA

Vérifions les hypothèses de l'ANOVA, qui sont :

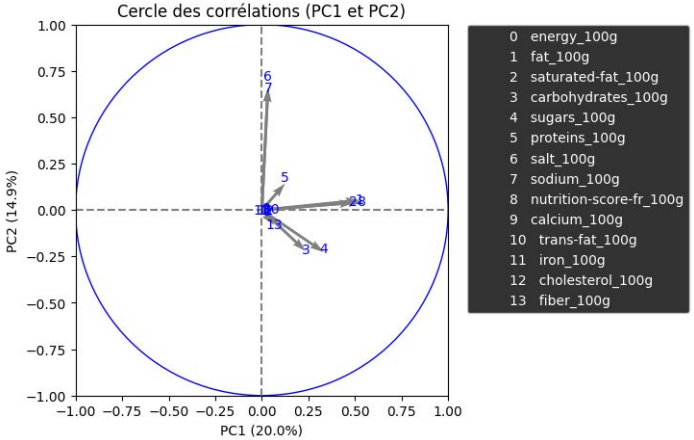
- fat_100g doit avoir une distribution normale pour chaque valeur de nutriscore ✗
- chaque distribution doit avoir la même variance ✗
- les données pour chaque valeur de nutriscore sont indépendantes

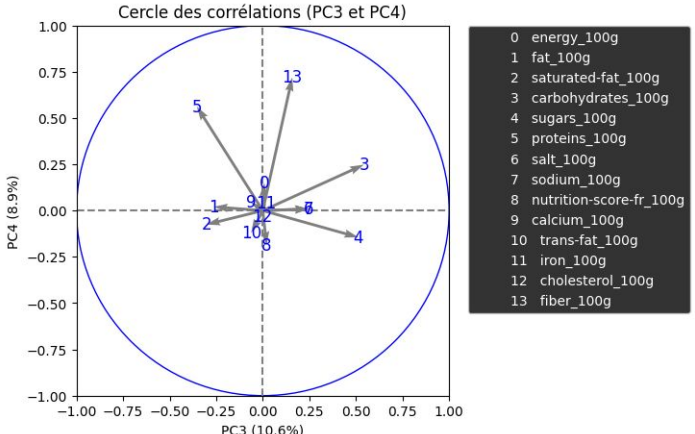


II. Analyse exploratoire
a. multivarié et réduction de dimension

ACP : Analyse par composantes principales

variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
energy_100g	0.025	-0.0072	0.0077	0.15	0.14	-0.29	0.069	0.92
fat_100g	0.53	0.054	-0.26	0.02	-0.024	-0.015	-0.039	-0.021
saturated-fat_100g	0.49	0.039	-0.3	-0.075	-0.0082	-0.02	-0.035	-0.0029
carbohydrates_100g	0.24	-0.22	0.54	0.25	0.017	0.016	0.033	-0.04
sugars_100g	0.33	-0.23	0.51	-0.14	-0.026	0.024	0.033	0.014
proteins_100g	0.12	0.14	-0.35	0.56	-0.005	0.015	-0.012	-0.081
salt_100g	0.033	0.66	0.25	0.0089	0.0081	-0.0025	0.011	-0.00087
sodium_100g	0.033	0.66	0.25	0.0089	0.0081	-0.0025	0.011	-0.00087
nutrition-score-fr_100g	0.54	0.044	0.021	-0.19	-0.049	0.021	-0.016	0.028
calcium_100g	0.019	0.01	-0.064	0.045	-0.18	0.6	0.73	0.18
trans-fat_100g	0.049	0.0057	-0.062	-0.12	0.59	-0.19	0.15	-0.021
iron_100g	-0.00047	-6.5e-05	0.017	0.039	0.27	0.71	-0.6	0.24
cholesterol_100g	0.019	-0.0031	-0.0033	-0.032	0.72	0.11	0.27	-0.19
fiber_100g	0.066	-0.081	0.16	0.72	0.078	-0.051	0.0068	-0.14

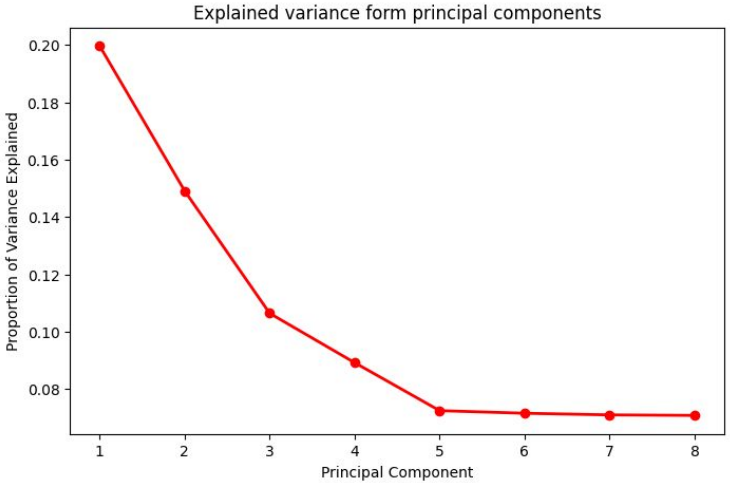


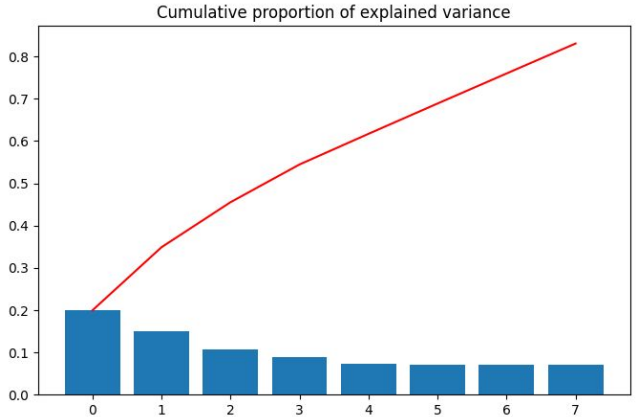


II. Analyse exploratoire
a. multivarié et réduction de dimension

Choix du nombre de composantes :

- le graphique de proportion de variance expliquée (Proportion of variance explained) par composante principale indique le pourcentage de la valeur propre associée normalisée. On recherche ici un “coude” qui indique la chute d’information apportée par les composantes. Ici, il s’agit de la cinquième composante. Nous ne retiendrons donc que **cinq composantes pour notre application.**
- l’ACP permet **ainsi une réduction de 14 à 5 dimensions pour 70% de variance expliquée.**





II. Analyse exploratoire

a. multivarié et réduction de dimension

Nouvelles variables synthétiques :

- PC1 : positivement corrélé avec les graisses, les graisses saturées ainsi que le sucre et les glucides ➜ **athérosclérose** (dépôt de lipides sur la paroi des artères)

- PC2 : positivement corrélé avec le sel et le sodium ➜ **insuffisance cardiaque et hypertension artérielle**

- PC3 : positivement corrélé avec le sucre et les glucides, mais négativement avec les protéines ➜ **diabète**

- PC4 : positivement corrélé avec les fibres et les protéines ➜ **facteur de prévention** (bénéfique pour le cœur)

- PC5 : corrélation positive avec les gras saturés et le cholestérol ➜ niveau élevé de non-HDL (**mauvais cholestérol**)

Résultats

II. Analyse exploratoire

a. résultat de l'application

Score =

- athérosclérose
- hypertension
- diabète
- mauvais cholestérol
- 2 * additif_score*
+ 4 * facteur de prévention

Le tout normalisé entre 0 et 1.
1 correspond à un produit "bon pour le cœur"
0 un produit très "mauvais pour le cœur"

additif_score* : permet de prendre en compte si le produit est très transformé

	product_name	heart_score
13298	Quenelles de brochet sauce Nantua	0.844833
13299	Coquilles Saint-Jacques* à la Bretonne (4 + 2 ...	0.844833
13682	2 Clafoutis aux tomates et aux fromages de chè...	0.833847
13693	4 Mini-gratins de pomme de terre	0.832561
4902	Laitue iceberg	0.823949

Limitations :

- Manque de prise en compte de certains facteurs, par exemple comme la présence d'alcool dans le produit.
- le score a tendance à favoriser des produits déjà préparés contenant des protéines et des fibres, hors une alimentation de produits peu transformés est préférable lorsque l'on souffre de problèmes cardiaques
- le choix de la pondération peut être adapté pour une maladie spécifique