Projet 2 - Concevez une application au service de la santé publique

Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Centre Baclesse
L'excellence pour vaincre votre cancer

Rappel de l'appel à projet

- Trouver des idées innovantes d'applications en lien avec l'alimentation

- Utilisant une base de donnée libre de produits alimentaires :
  - > 320 000 produits alimentaires différents
  - Informations générales (nom, fabricant, packaging…)
  - Classe du produit, son origine, bio ou non, score nutritif…
  - Sa composition en nutriments pour 100 grammes du produit.

Idée d'application

Health and Planet Care

- Faire une application qui pourrait nous informer (via un scan du code barre) sur deux critères primordiaux **mais parfois en opposition** :

  - Produit **bon pour la santé**
    - Nutriscore bon
    - Biologique
    - Sans additifs
    - …
  - Produit **bon pour la planète**
    - Produit localement
    - Biologique
    - Ne contient pas d'huile de palme
    - …

Exploration des données

- Variables d'intérêt :
  - « code » (pour scan code barre)
  - « product_name » (contient ou non la mention « *organic* »)
  - « countries_fr » (informe sur la localisation pour critère écologique planète)
  - « carbon-footprint_100g » (informe sur l'empreinte carbone du produit)
  - « ingredients_from_palm_oil_n » (informe sur la présence d'huile de palme)
  - « nutrition-score-fr_100g » (informe sur le score nutritif)
  - « additives_n » (informe sur la présence d'additif)
  - Composition en nutriments (energy/fat/saturated-fat/carbohydrates/sugars/fiber/proteins/salt)

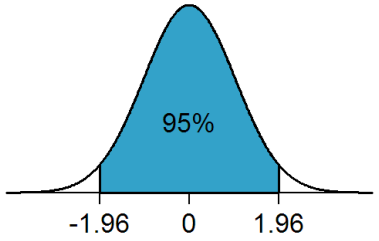
Nettoyage des données

- Décompte des variables présentant un nombre de données manquantes trop importante :
  - suppression de ces variables si *NaN* > 50 %
- Vérification du type des données :
  - vérifier que les données censées être des *float* sont bien des *float* et sinon mettre des *NaN* à la place (même approche pour les *str*)
- Vérification des erreurs de saisie :
  - vérifier que les “NaN” n’ont pas été rentrés comme “n/a” par exemple

Centre Baclesse
L'excellence pour vaincre votre cancer

Nettoyage des données

- Nettoyage des valeurs aberrantes :
  - Des valeurs négatives sont retrouvées dans certaines variables comme le sucre, les protéines, les fibres et le nutriscore
    - Remplacement de ces valeurs incohérentes par des *NaN*
  - Vérification des valeurs dupliquées
  - Suppression des données significativement ($p<0.05$) différentes de la valeur moyenne
    - Différence avec la moyenne > 1.96 * écart-type



Centre Baclesse
L'excellence pour vaincre votre cancer

Nettoyage des données

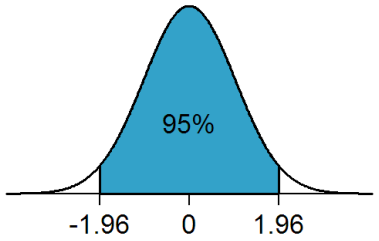
- Nettoyage des valeurs aberrantes :

- Avant

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	2.611130e+05	243891.000000	229554.000000	243588.000000	244971.000000	200886.000000	259922.000000	255510.000000
mean	1.141915e+03	12.730379	5.129932	32.073981	16.003484	2.862111	7.075940	2.028624
std	6.447154e+03	17.578747	8.014238	29.731719	22.327284	12.867578	8.409054	128.269454
min	0.000000e+00	0.000000	0.000000	0.000000	-17.860000	-6.700000	-800.000000	0.000000
25%	3.770000e+02	0.000000	0.000000	6.000000	1.300000	0.000000	0.700000	0.063500
50%	1.100000e+03	5.000000	1.790000	20.600000	5.710000	1.500000	4.760000	0.581660
75%	1.674000e+03	20.000000	7.140000	58.330000	24.000000	3.600000	10.000000	1.374140
max	3.251373e+06	714.290000	550.000000	2916.670000	3520.000000	5380.000000	430.000000	64312.800000

- Après

	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g
count	254829.000000	219809.000000	204280.000000	233511.000000	212479.000000	188299.000000	229500.000000	251670.000000
mean	1072.825903	8.307916	3.009120	29.447524	9.076825	1.975688	4.761451	0.969212
std	732.102730	9.957327	3.876443	26.829322	11.061363	2.304507	4.576015	1.453342
min	0.000000	0.000000	0.000000	0.000000	-17.860000	0.000000	-3.570000	0.000000
25%	368.000000	0.000000	0.000000	5.420000	0.880000	0.000000	0.400000	0.060000
50%	1059.000000	3.390000	1.160000	18.640000	3.700000	1.200000	3.570000	0.551180
75%	1644.000000	15.000000	5.000000	54.930000	13.270000	3.300000	7.500000	1.315085
max	2700.000000	33.720000	14.400000	84.850000	42.550000	9.700000	17.500000	14.815820



Centre Baclesse
L'excellence pour vaincre votre cancer

Nettoyage des données

- Vérification logique des données :

```
df[df["energy_100g"] == 2700].head()
```


```
df[df["fat_100g"] == 33.72].head()
```

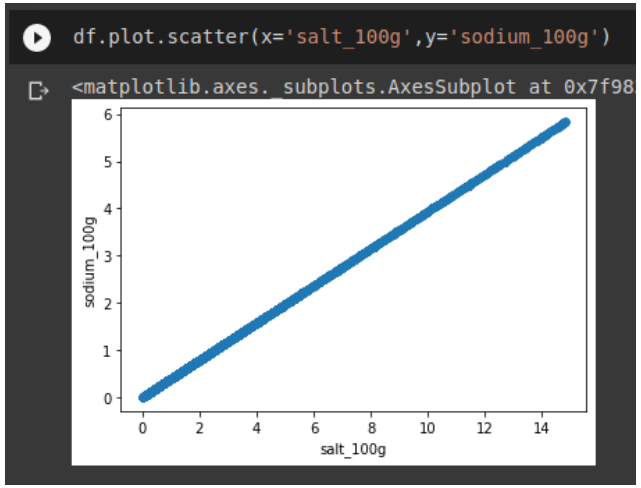
Nettoyage des données

- Vérification d'impossibilité logique :
  - Suppression des valeurs lorsque la quantité de graisse saturée est plus importante que le total de graisse
    - Remplacement de ces valeurs par des NaN

Centre Baclesse
L'excellence pour vaincre votre cancer

Nettoyage des données

- Décompte et suppression des variables d'intérêt non utilisable :
  - Suppression des variables *huile de palme* et *empreinte carbone* car elles présentent trop de *NaN*
- Suppression des variables redondantes :
  - Le sel (salt) de formule Na+Cl- est certainement similaire au sodium (de formule Na+), il y a t-il une utilité à le garder ?
    - Suppression de la variable sodium



Centre Baclesse
L'excellence pour vaincre votre cancer

Nettoyage des données

- Formatage de la database :
  - Sélection des variables d'intérêt pour la problématique
  - Reformulation des titres des variables pour plus de lisibilité notamment dans les représentations graphiques qui suivent
    - *countries_fr => countries* ; *sugar_100g => sugar*
  - Reset des index, dernière vérification visuelle et sauvegarde du dataframe

Exploration des données

- Colonnes Dataframes crées :
  - *« Produce_in_UE »* (information si production en Union-Européenne)
    - En utilisant les informations du pays de production dans la variable “*countries*”
  - *« Empreinte_carbone »* (information sur l’empreinte carbone théorique)
    - En utilisant l’information de production dans l’UE ou non j’attribue une empreinte carbone théorique selon la littérature :

Production dans l’UE : 0.31 tCO2/fr/an

Production en dehors de l’UE : 0.48 tCO2/fr/an

Source :

https://ecotoxicologie.fr/empreinte-carbone-alimentation

Centre Baclesse
L'excellence pour vaincre votre cancer

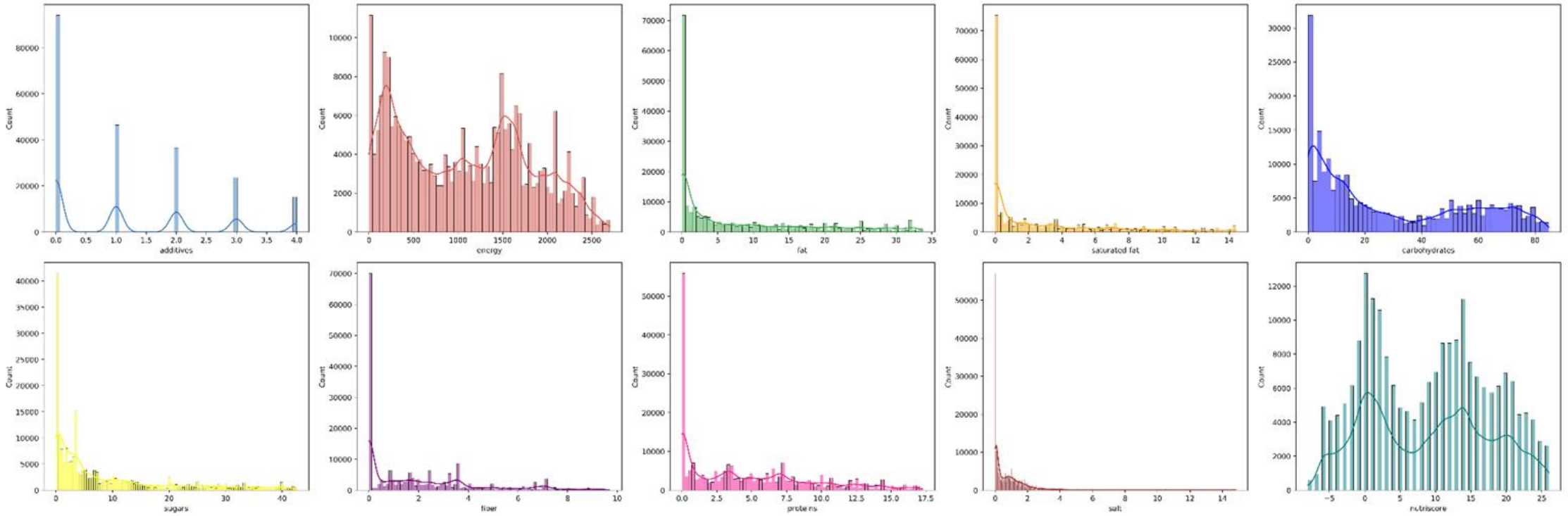
Exploration des données

- Colonnes Dataframes crées :
  - « *Organic_product* » (information si le produit est bio ou non)
    - En utilisant la variable “*product_name*” qui contient ou non la mention “*organic*”

Centre Baclesse
L'excellence pour vaincre votre cancer

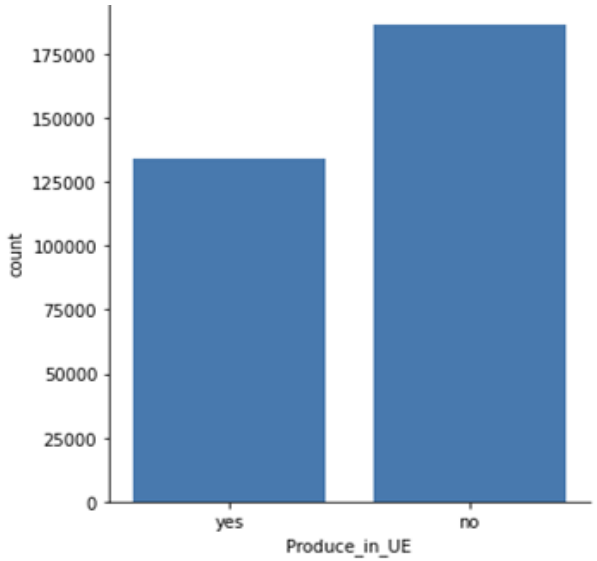
Analyse univariée

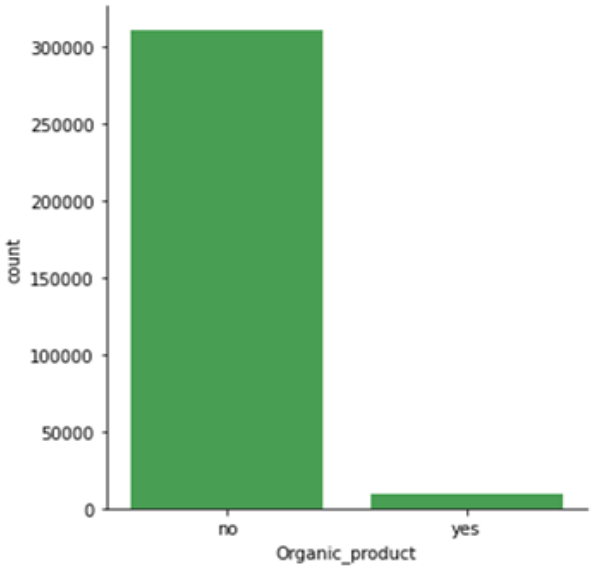
- Exploration globale des variables **quantitatives** d'intérêt - histogrammes
  - Il y a globalement moins de produits avec un nombre important d'additifs
  - On observe avec les variable *energy*, *carbohydrates* et *nutriscore* qu'il existe deux (voir trois avec le *nutriscore*) populations/groupes d'aliment

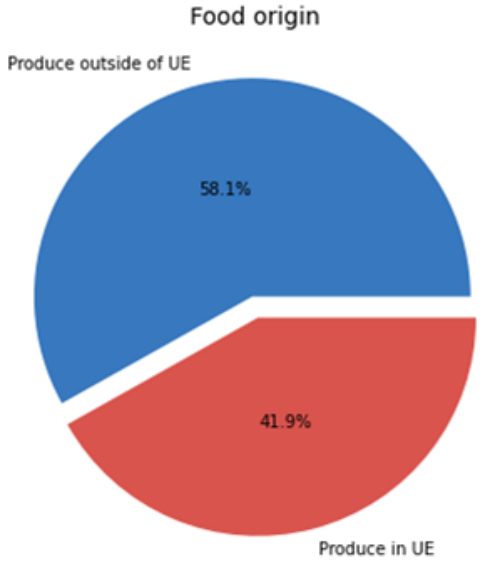


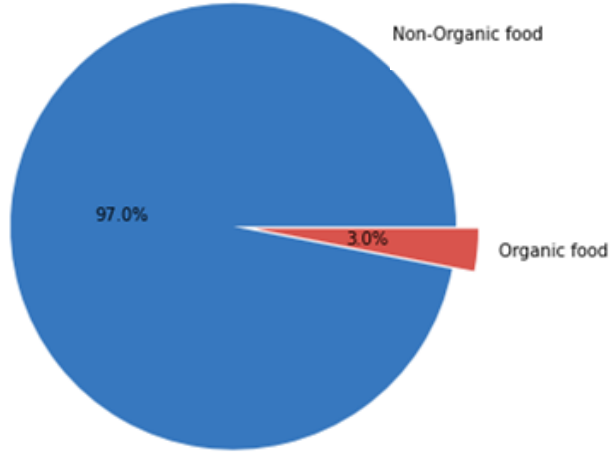
Analyse univariée

- Exploration globale des variables **qualitatives** d'intérêt
  - Il y a un peu moins d'aliments produits dans l'UE qu'en dehors.
  - Il y a beaucoup moins de produit bio / produits non-bio





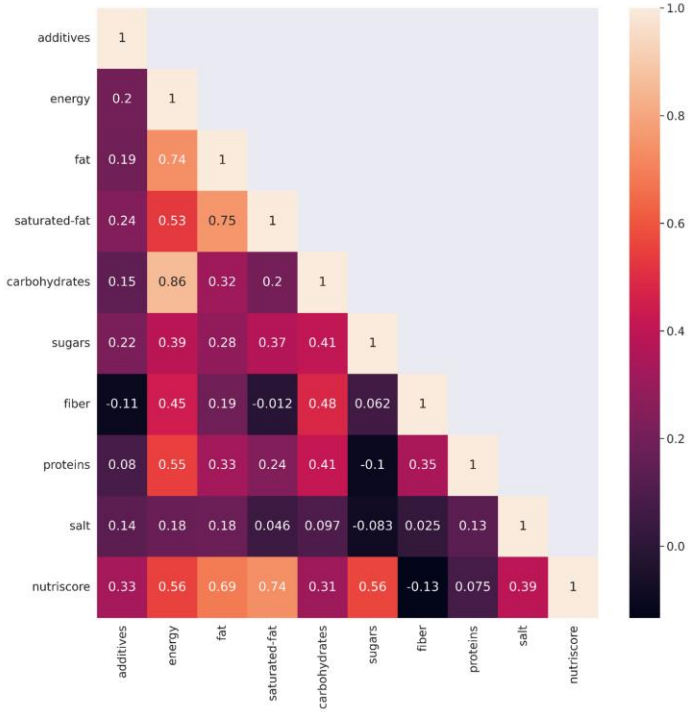




Centre Baclesse
L'excellence pour vaincre votre cancer

Analyse univariée

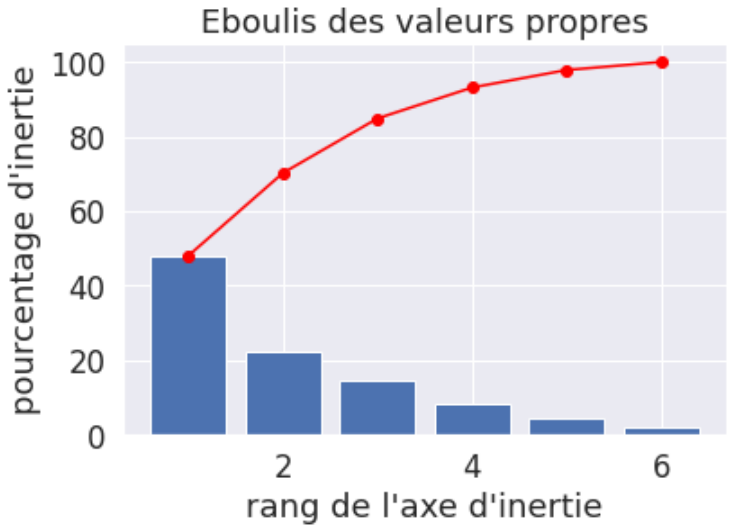
- Etude des corrélations entre les variables
  - Les principales corrélations observées concernent les nutriments et le nutriscore

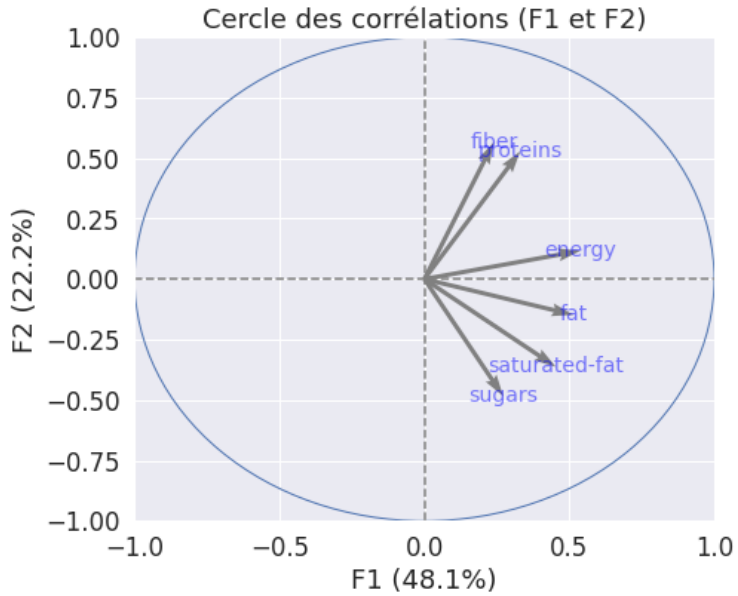


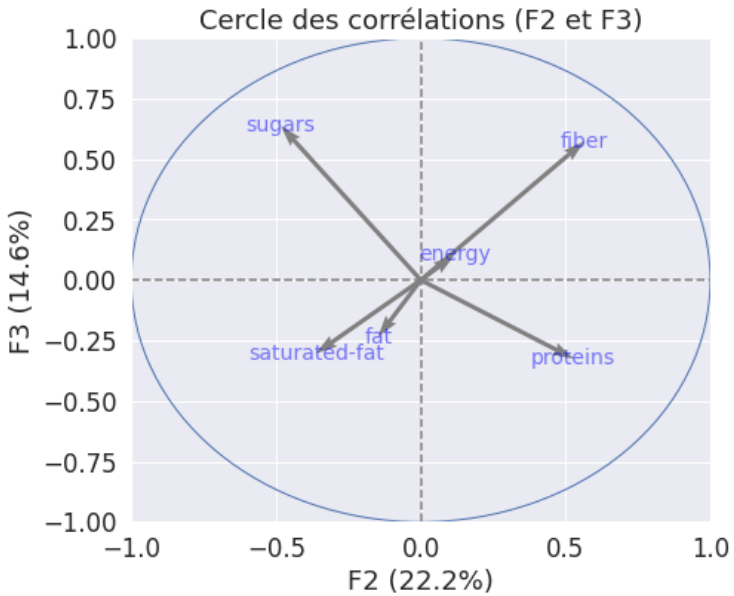
Sélection des corrélations significatives

Analyse univariée

- Etude des corrélations entre les variables / nutriscore
  - La PCA permet bien de comprendre comment est calculé le nutriscore.
    - Composante F1 : comment le nutriscore est calculé
    - Composante F2 expliquant 22.2% de la variance discrimine clairement ce qui est bon pour la santé (fibre/protéines) de ce qui n'est pas bon (sucre, graisses...)



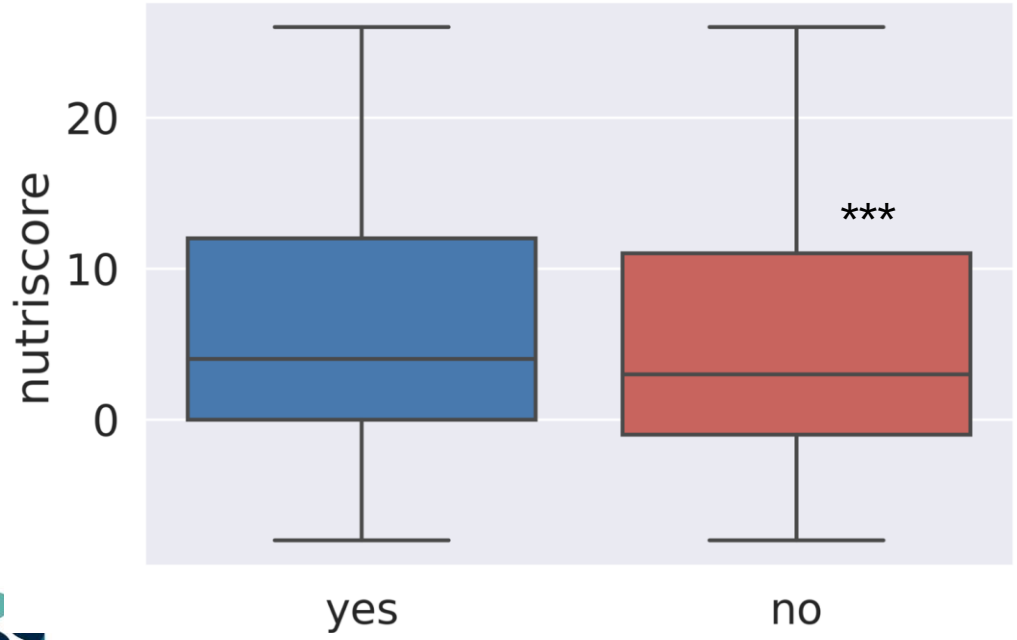


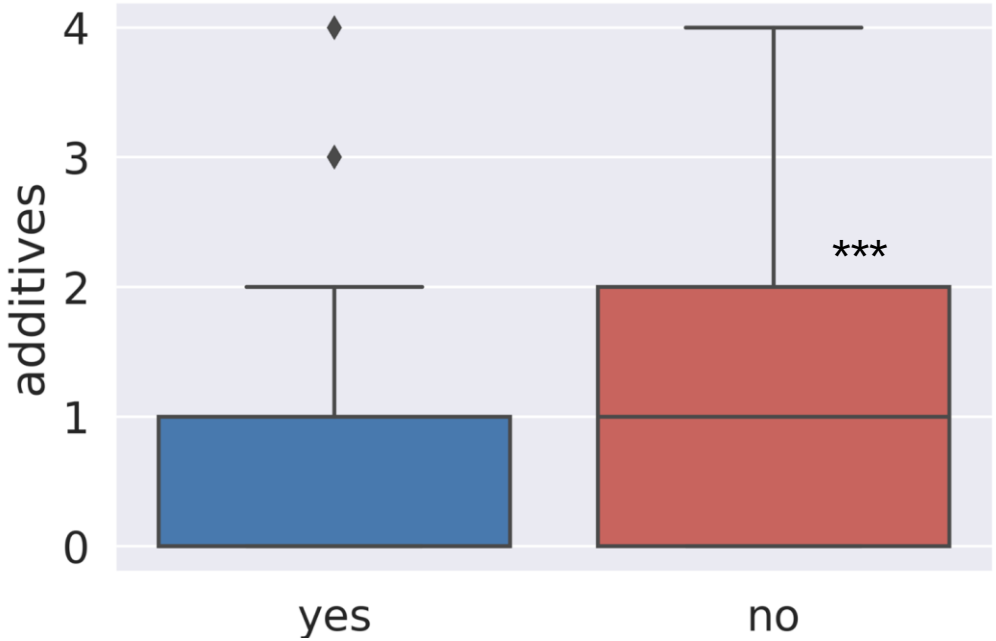


Centre Baclesse
L'excellence pour vaincre votre cancer

Analyse multivariée - SANTÉ

- Caractéristique des produits bio :
  - Les produits bio ont un meilleur score nutritif et moins d'additifs
  - On peut dans le cadre de notre application utiliser ces deux variables pour construire notre score



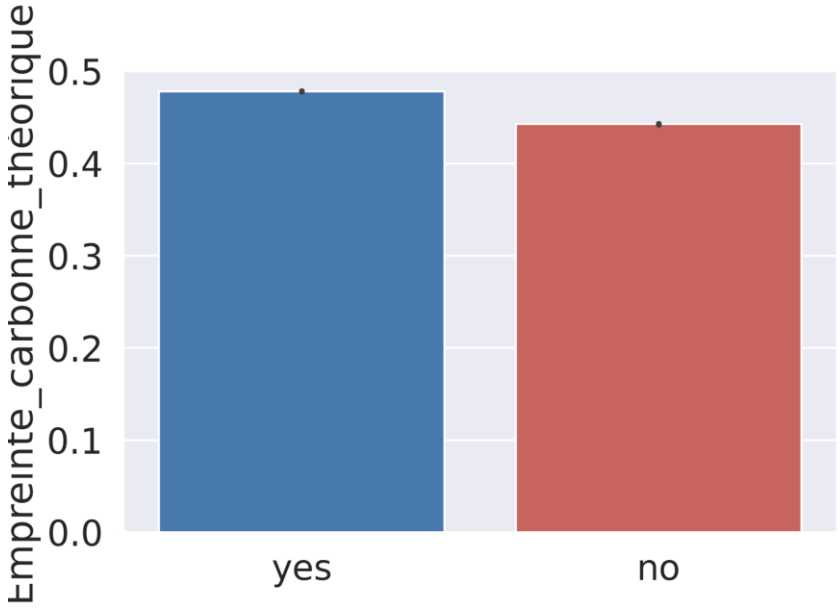


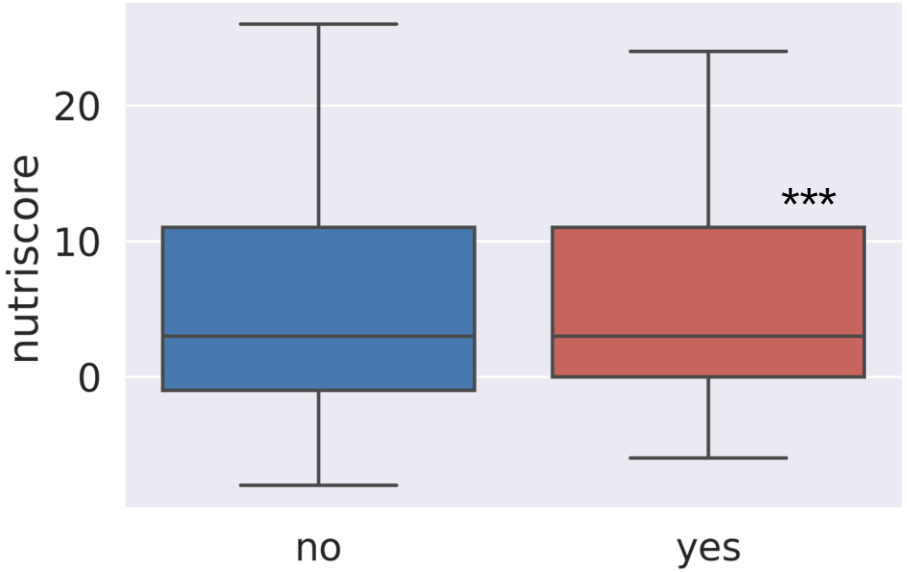
*** p< 0.001

Centre Baclesse
L'excellence pour vaincre votre cancer

Analyse multivariée - PLANÈTE

- Caractéristique aliment produit dans l'UE :
  - Les aliments produits dans l'UE ont une empreinte carbone moindre et un meilleur score nutritif
  - On peut dans le cadre de notre application utiliser ces deux variables pour construire notre score





*** p< 0.001

Calcul des scores pour l'application

Health and Planet Care

Score Santé :

Value = nutriscore (norm 1 à 2) + Organic product (1 ou 2) + additives (norm 1 à 2)

Score Planète :

Value = Produce in UE (1 ou 2) + Organic product (1 ou 2)

Score Global :

Value = Score Santé + Score Planète

Centre Baclesse
L'excellence pour vaincre votre cancer

Analyse multivariée - SANTÉ + PLANÈTE

- Pertinence du score personnalisé *vs* nutriscore :
  - Notre score ajoute-t-il une plus value vis-à-vis du nutriscore ?

Pertinence du nutriscore :

	NoBio_NoUE	NoBio_UE	Bio_NoUE	Bio_UE
NoBio_NoUE	Reject	Accept	Accept	Accept
NoBio_UE	Accept	Reject	Accept	Reject
Bio_NoUE	Accept	Accept	Reject	Reject
Bio_UE	Accept	Reject	Reject	Reject

Pertinence du score général :

	NoBio_NoUE	NoBio_UE	Bio_NoUE	Bio_UE
NoBio_NoUE	Reject	Accept	Accept	Accept
NoBio_UE	Accept	Reject	Accept	Accept
Bio_NoUE	Accept	Accept	Reject	Accept
Bio_UE	Accept	Accept	Accept	Reject

  - Notre score personnalisé (santé+planète) est plus relevant pour discriminer les différentes conditions que le nutriscore
    - Cette idée d'application est donc **faisable** et **pertinente**

Centre Baclesse
L'excellence pour vaincre votre cancer

** Utilisation d'un test de kruskal suivi d'un test post hoc de Mann Whitney*

Design application

Health and Planet Care

Health Score

Planet Score

Global Score

Local production

Organic

Nutriscore : good

36% reduction of carbon footprint for an annual consumption of this type of product (1)

25% reduction of cancer risk (2)

10% reduction in mortality (3)

(1) : https://ecotoxicologie.fr/empreinte-carbone-alimentation

(2) : Baudry et al., JAMA Intern Med, 2018

(3) : Deschasaux et al., BMJ, 2020

Centre Baclesse
L'excellence pour vaincre votre cancer

Projet 2 - Concevez une application au service de la santé publique

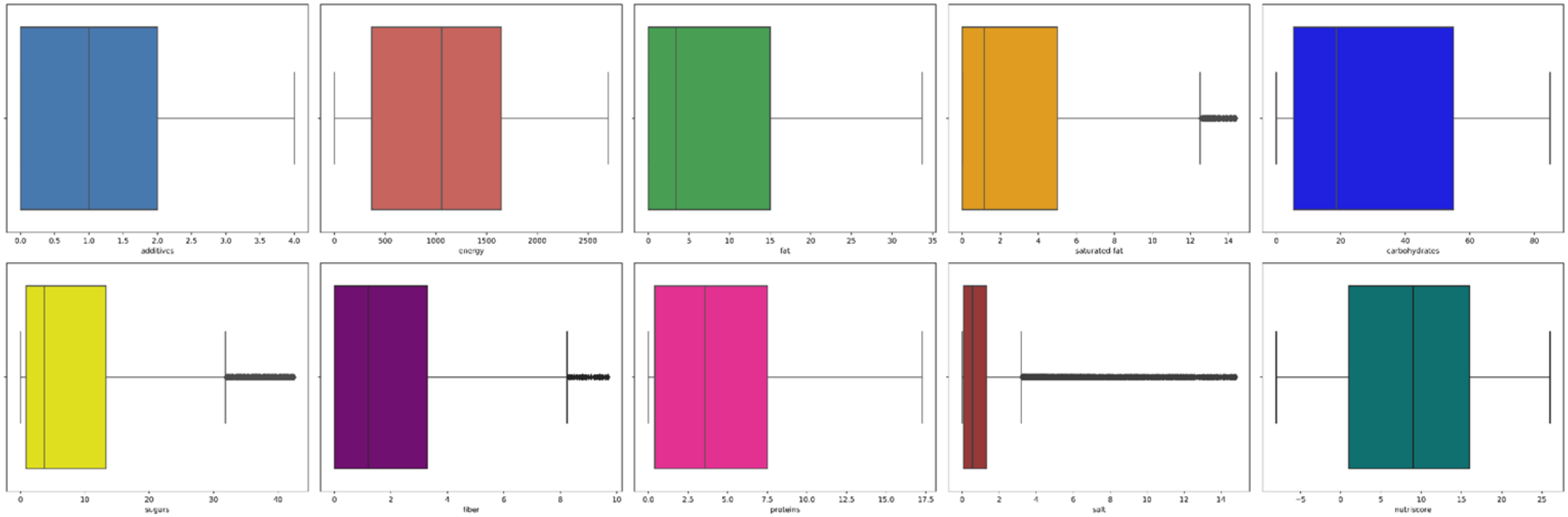
Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Centre Baclesse
L'excellence pour vaincre votre cancer

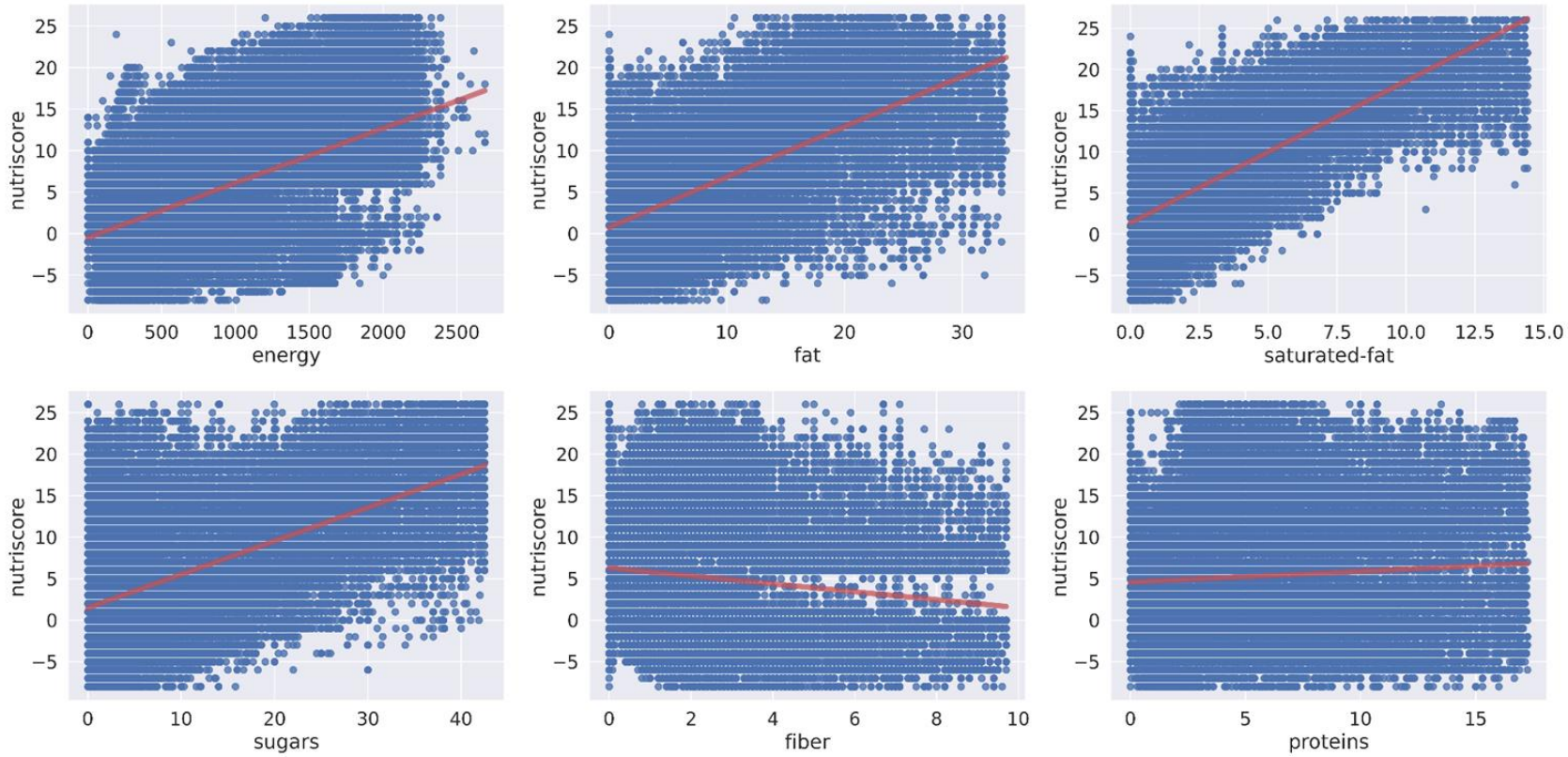
Analyse univariée

- Exploration globale des variables **quantitatives** d'intérêt - boxplot
  - L'analyse par boxplot nous montre les valeurs médianes, les quartiles ainsi que les outliers qui sont des valeurs correctes mais sur lesquelles il faut avoir une attention particulière



Analyse univariée

- Etude des corrélations entre les variables
  - Les principales corrélations observées concernent les nutriments et le nutriscore



Centre Baclesse
L'excellence pour vaincre votre cancer

Calcul des scores pour l'application

Health and Planet Care

- Fonction permettant de calculer les différents score
  - Etude test sur différents produits

```
[ ] def scores_calculator(code_barre):
      """Fonction de calcul du score santé en fonction du code barre fournit"""
      score_sante = round(float(df.loc[df['code_bar'] == code_barre, 'score_santé']),2)
      score_planete = round(float(df.loc[df['code_bar'] == code_barre, 'score_planète']),2)
      score_general = round(float(df.loc[df['code_bar'] == code_barre, 'score_général']),2)
      print("Le produit : '\x1B[3m" + str((df.loc[df['code_bar'] == code_barre, 'product_name']).values[0]) + "\x1B[0m' possède un :")
      print("Score santé de : " + str(score_sante))
      print("Score planète de : " + str(score_planete))
      print("Score général de : " + str(score_general))
```

```
[ ] scores_calculator(18227)

    Le produit : 'Organic Oat Groats' possède un :
    Score santé de : 4.89
    Score planète de : 3.0
    Score général de : 7.89
```

```
[ ] scores_calculator(18265)

    Le produit : 'Energy Power Mix' possède un :
    Score santé de : 3.21
    Score planète de : 2.0
    Score général de : 5.21
```

Centre Bacless
L'excellence pour vaincre votre cancer