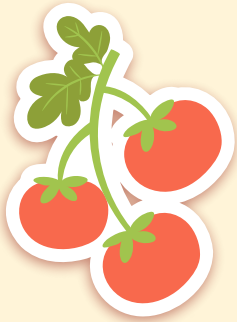
APPLICATION DE SANTÉ PUBLIQUE
BABY_YUKA

Alexandre MONOD - parcours Machine Learning - p2

SOMMAIRE

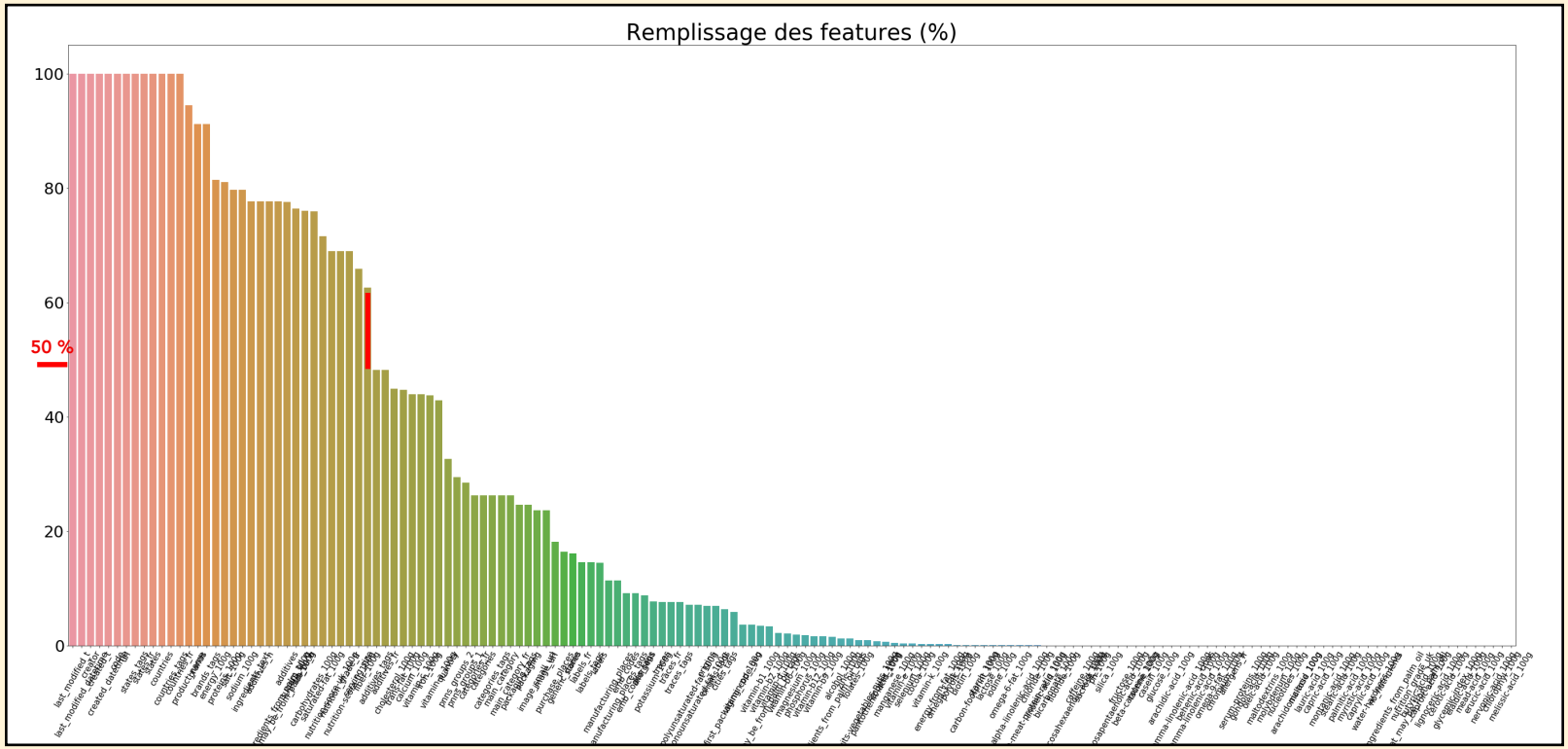
Introduction

- Objectif : créer une fonction similaire à Yuka
  - Recommandation générique
- Favoriser
  - Nutriscore
  - Label bio
  - Produits les moins transformés (additifs)



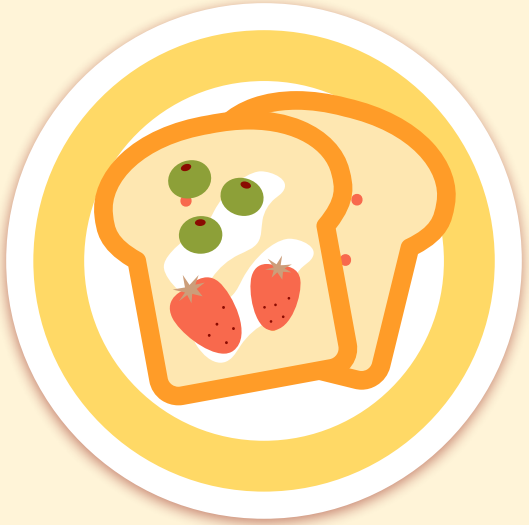
I - Nettoyage

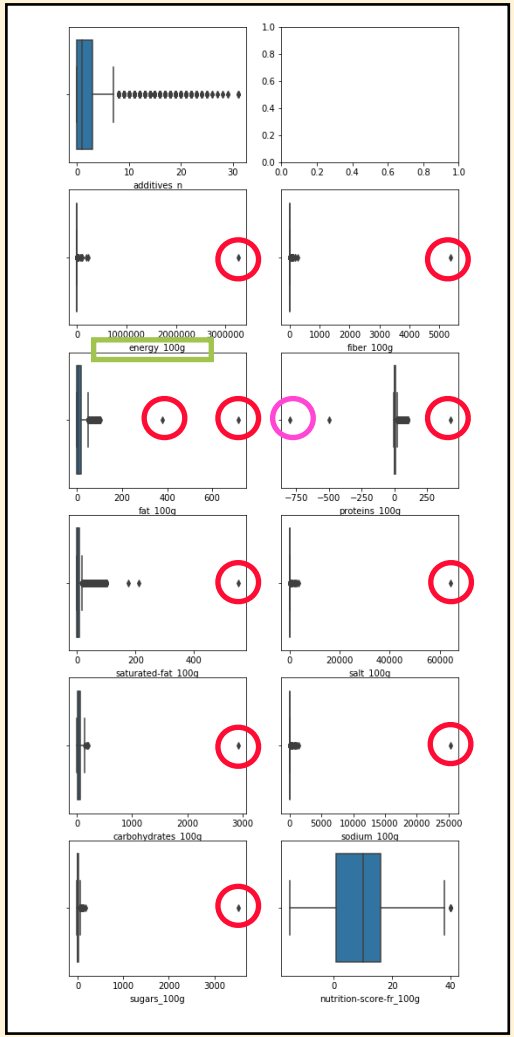
1) Sélection des features



1) Sélection des features

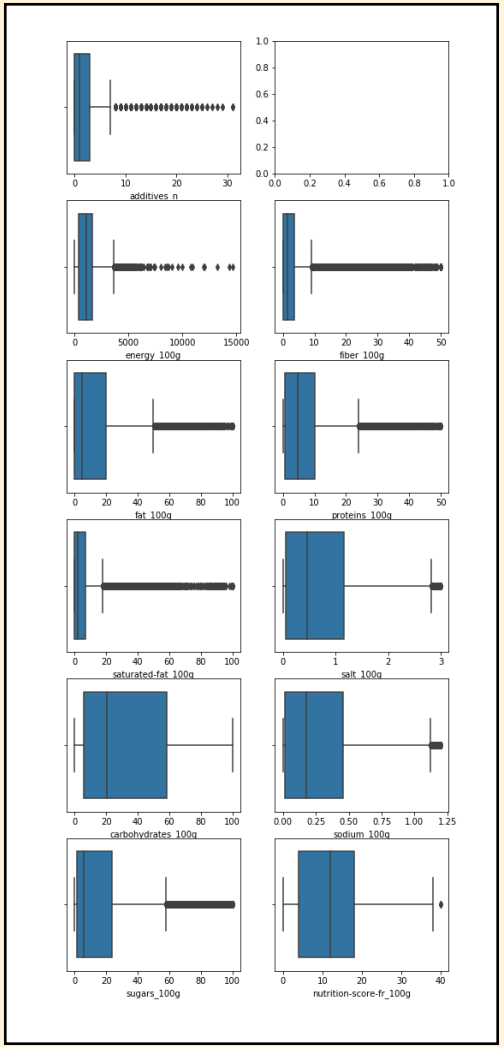
- Suppression
  - Colonnes techniques
  - Colonnes tags
  - Images
  - Redondance français - anglais
  - Redondance catégories
  - Huile de palme





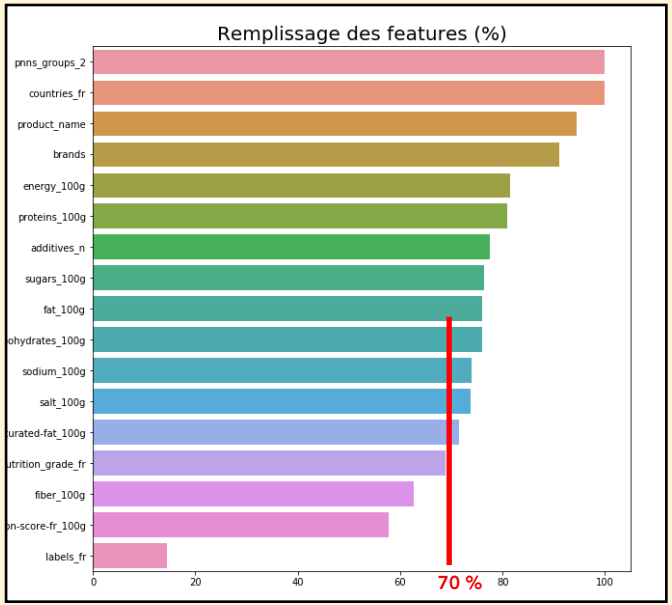
2) Outliers

- Valeurs aberrantes
  - Négatives
  - > 100g
- Valeurs illogiques
  - Taux max par catégorie (sel, sodium)

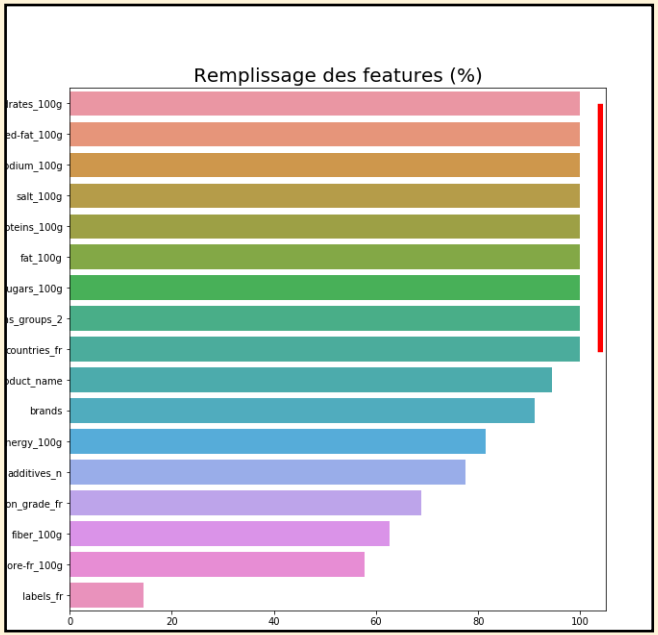


3) Imputation

1 - Médiane

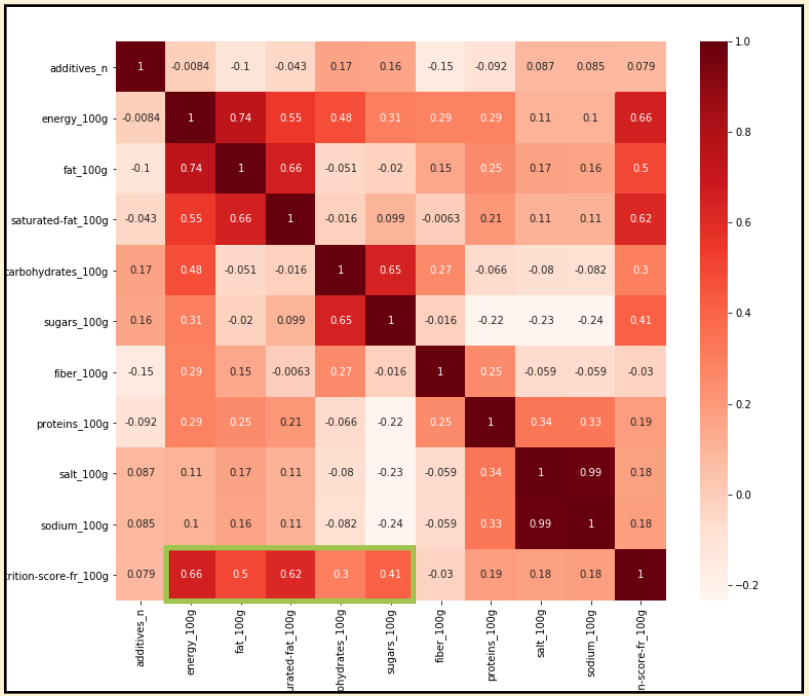


- colonnes « 100g »



3) Imputation

2 - KNN

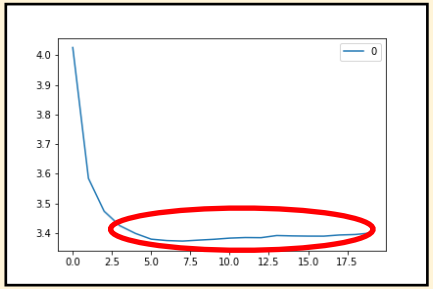


- Features les plus corrélées
  - energy
  - fat
  - saturated_fat
  - carbohydrates
  - sugars

3) Imputation

2 - KNN

- Séparation en deux datasets
  - Train
  - Test
- Scaling des features
- RMSE pour k voisins

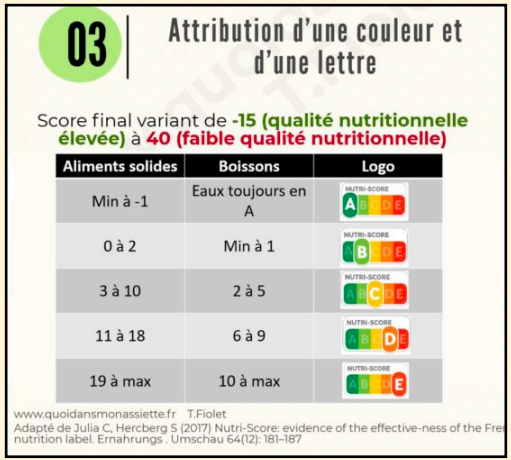


- Performance (MAE, R2)
- Modèle —> imputation

3) Imputation

2 - KNN

- Imputation nutrition_score —> Imputation des valeurs manquantes de nutrition_grade

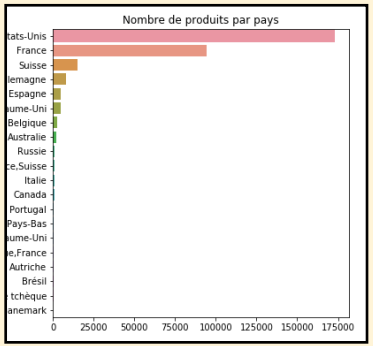


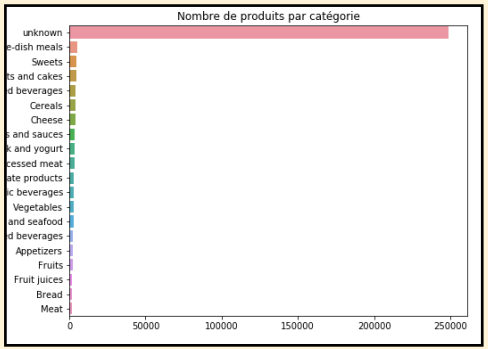
II - Analyse

1) Analyse univariée

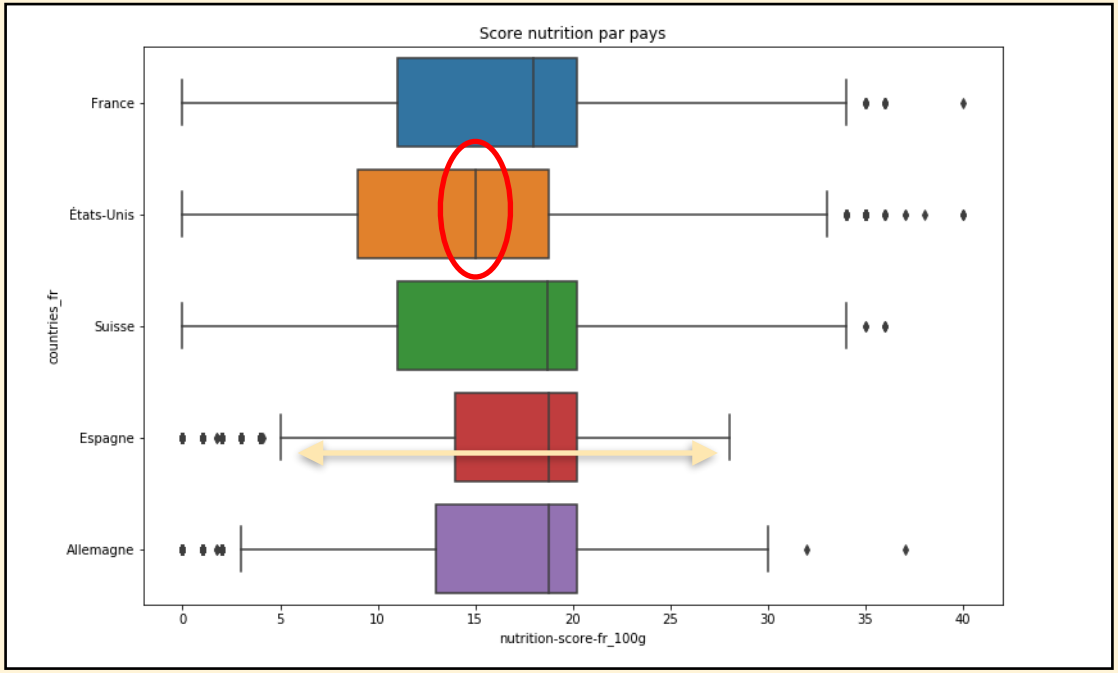
- Describe()

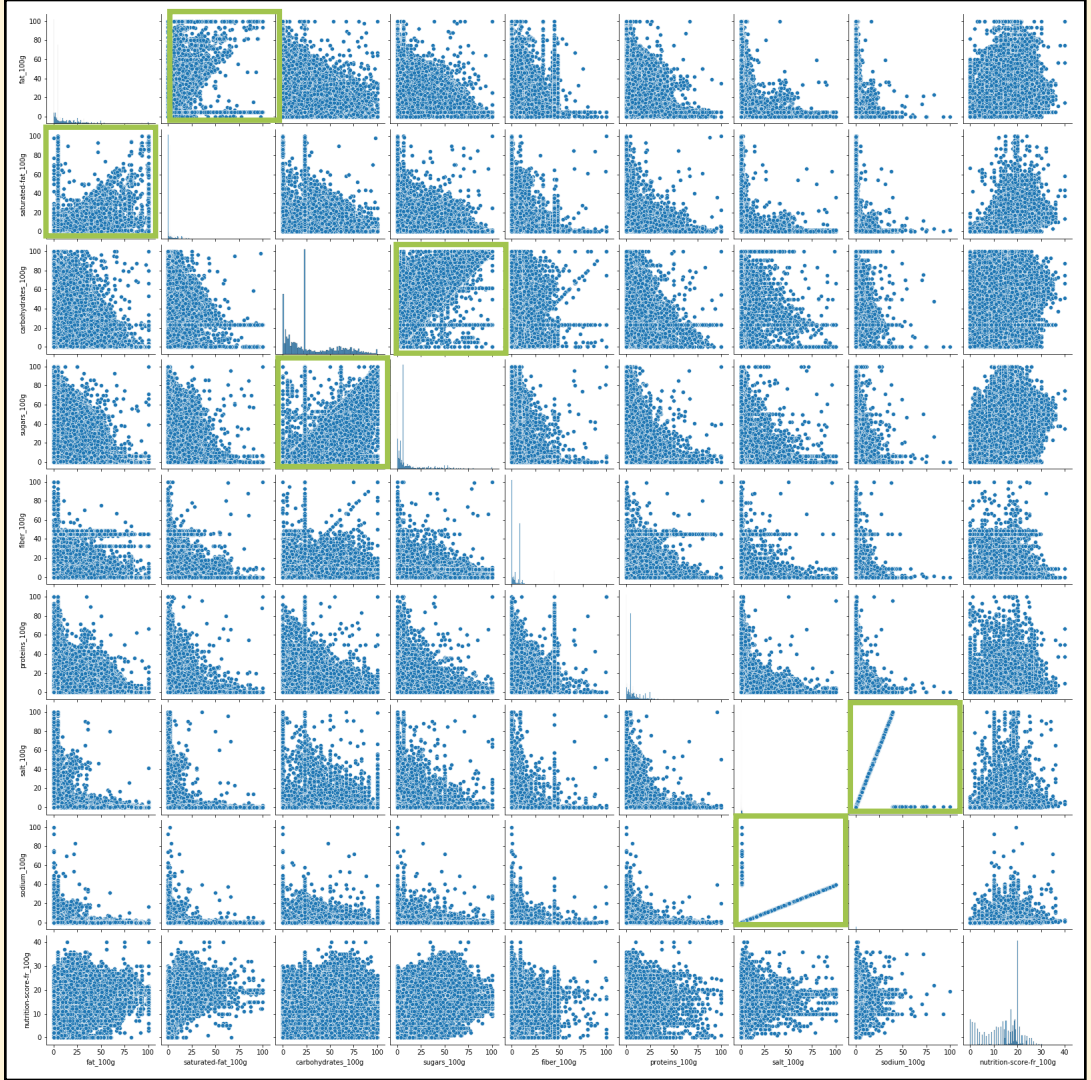
	countries_fr	nutrition_grade_fr	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g	pnns_groups_2
count	320476	320476	320476.000000	320476.000000	320476.000000	320476.000000	320476.000000	320476.000000	320476.000000	320476.000000	320476.000000	320476
unique	722	5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	42
top	États-Unis	e	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	unknown
freq	172993	119673	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	248675
mean	NaN	NaN	10.958467	4.246013	29.525981	13.909708	5.753430	6.623747	1.381913	0.561491	14.466921	NaN
std	NaN	NaN	15.956637	6.966663	26.426605	19.268476	9.364222	7.574320	5.589333	2.378309	7.077370	NaN
min	NaN	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	NaN
25%	NaN	NaN	0.420000	0.100000	7.080000	2.000000	0.000000	1.000000	0.091440	0.036000	10.000000	NaN
50%	NaN	NaN	4.650000	1.790000	23.080000	6.500000	2.800000	4.410000	0.600000	0.236220	17.111111	NaN
75%	NaN	NaN	15.450000	5.000000	50.850000	16.200000	8.572222	8.800000	1.198880	0.472441	20.111111	NaN
max	NaN	NaN	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	40.000000	NaN





2) Analyse bivariée




fat_100g
saturated-fat_100g
carbohydrates_100g
sugars_100g
fiber_100g
proteins_100g
salt_100g
sodium_100g
nutrition-score-fr_100g

ANOVA

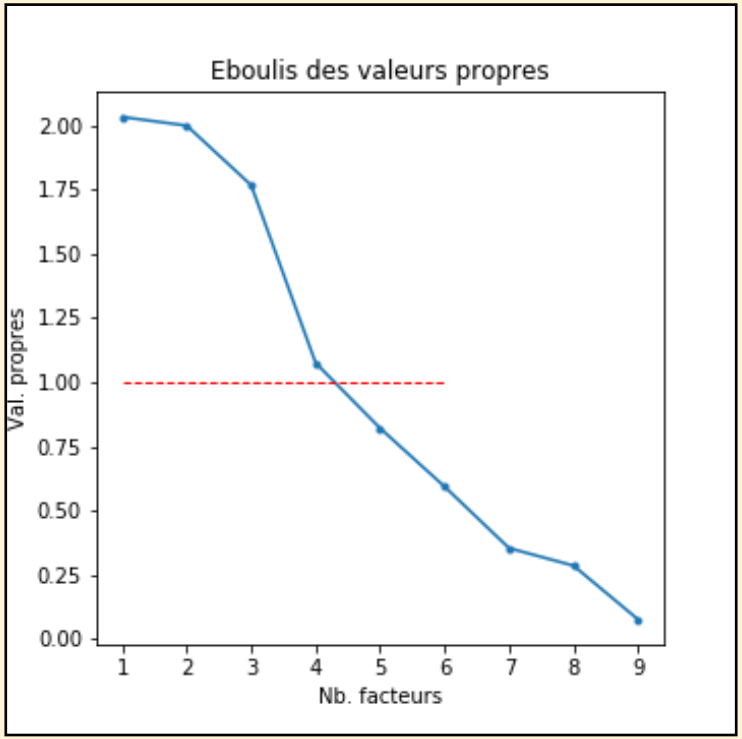
- Sélection de features :
  - catégorielle : nutrition_grade
  - quantitative : fat

- Tests :
  - Normalité : Shapiro
  - Homoscédasticité : Levene
  - Indépendance : Bartlett

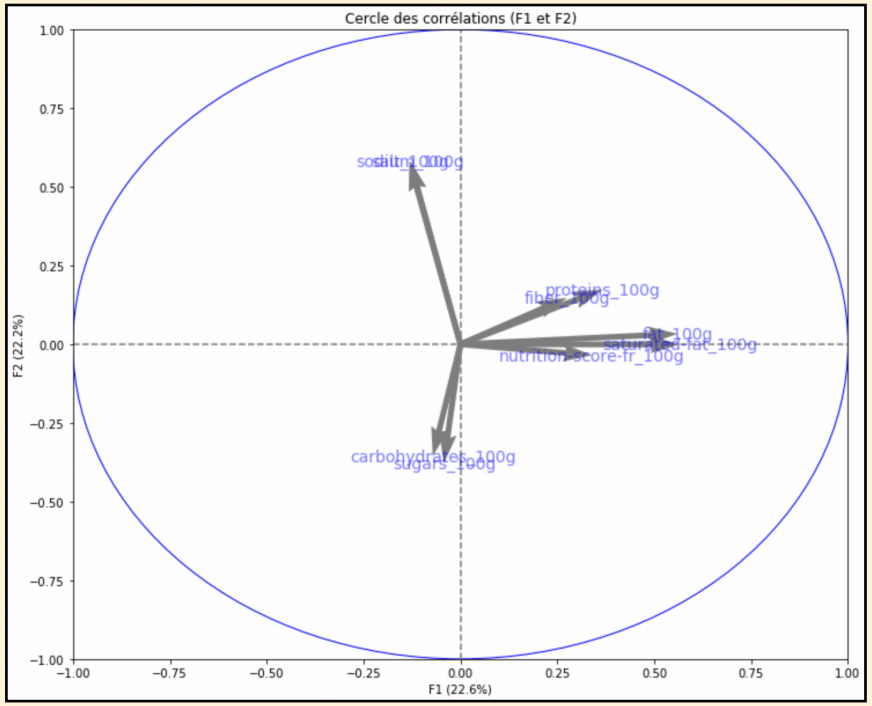
- Test non paramétrique:
  - Kruskal - Wallis

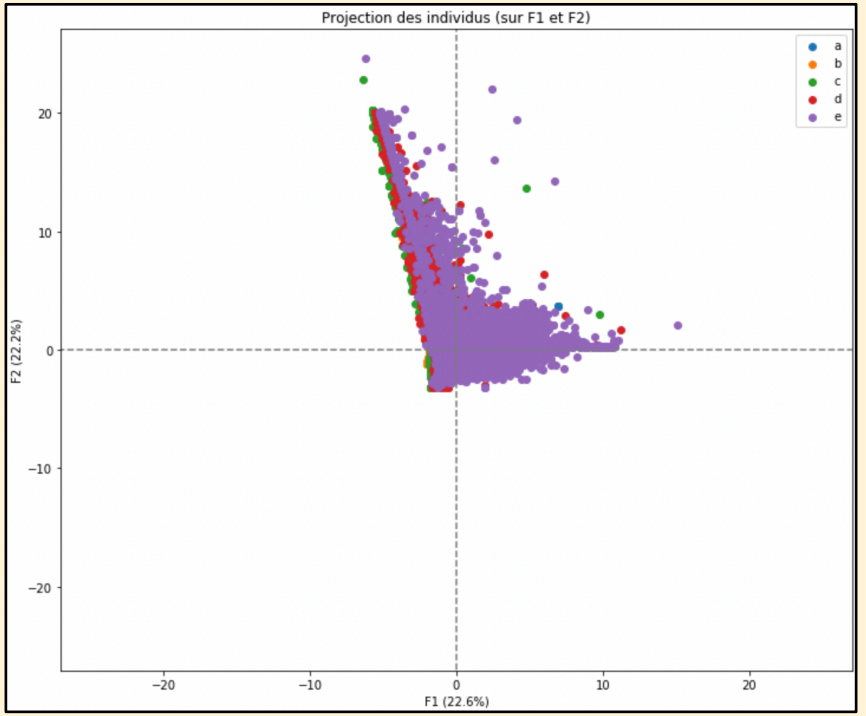
3) ACP

- Colonnes _100g

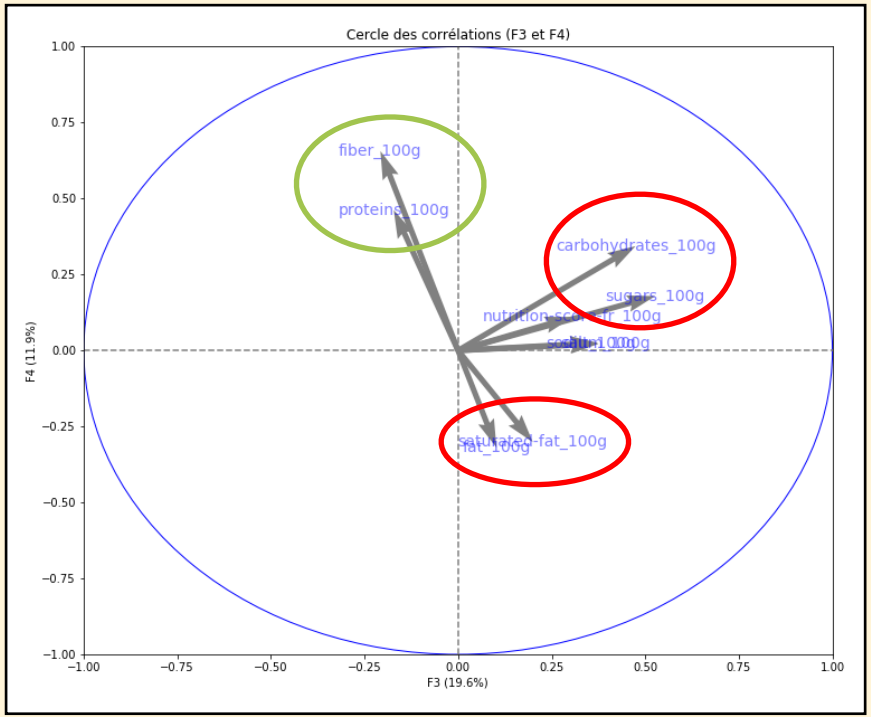


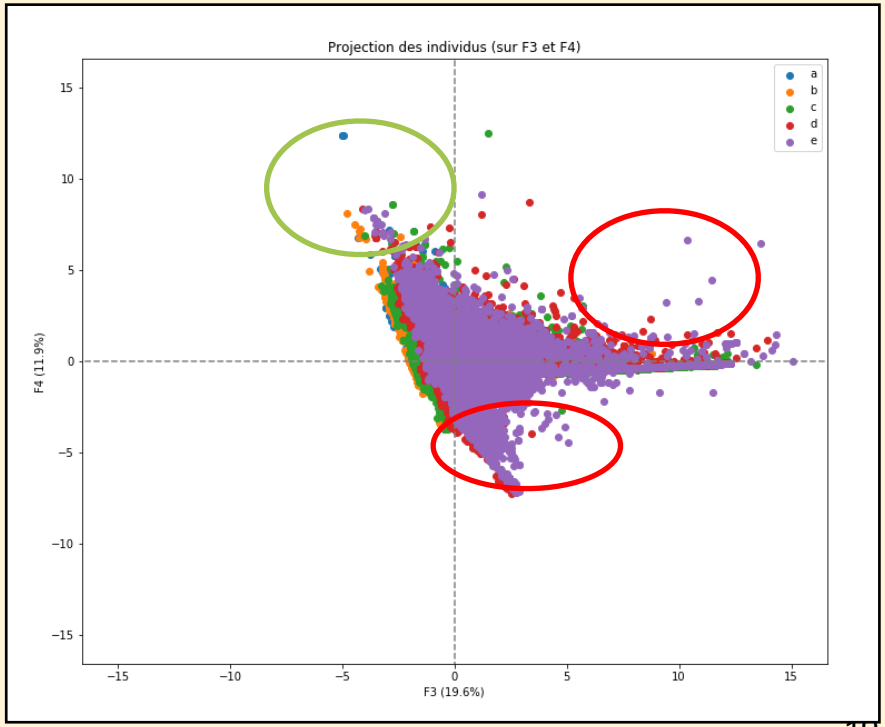
3) ACP


Cercle des corrélations (F1 et F2)
F1 (22.6%)
F2 (22.2%)
proteins_100g
carbohydrates_100g
sugars_100g
nutrition-score-fr_100g


Projection des individus (sur F1 et F2)
F1 (22.6%)
F2 (22.2%)
a
b
c
d
e

3) ACP


Cercle des corrélations (F3 et F4)
fiber_100g
proteins_100g
carbohydrates_100g
sugars_100g
saturated-fat_100g
F3 (19.6%)
F4 (11.9%)


Projection des individus (sur F3 et F4)
a
b
c
d
e
F3 (19.6%)
F4 (11.9%)

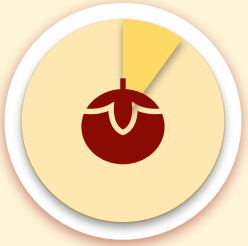
III - Modèle

Poids des features

83%

Nutrition score

10%

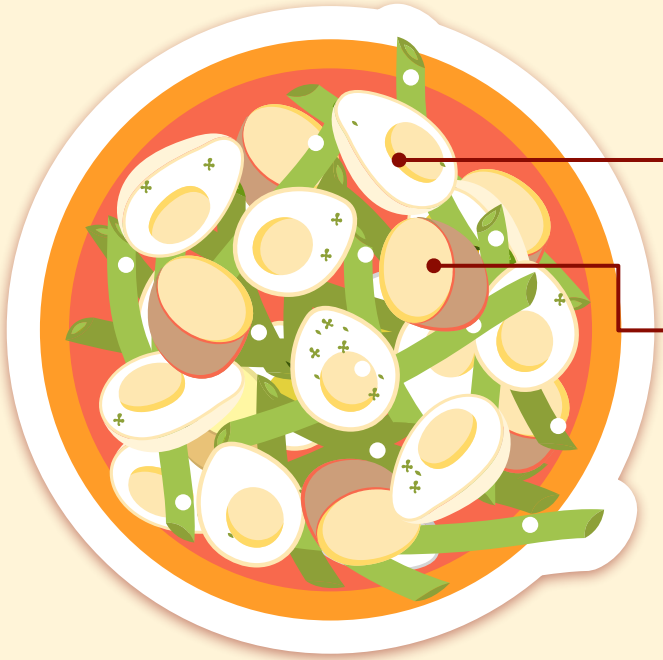


Label bio

7%

Malus additifs

Le modèle renvoie



1 **Meilleur produit**

Nom, Marque, Nutriscore

2 **Informations sur le produit**

Additifs, Energie, Gras, Gras saturés

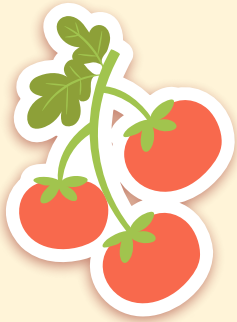
Carbohydrates, Sucres, Fibres, Protéines, Sel,

Sodium

Conclusion

Conclusion

- Modèle similaire à Yuka dans sa logique : recommander les produits les plus sains de manière globale
- Axes d'amélioration
  - Imputation par la médiane reste approximative
  - Pas de distinction des additifs bons/mauvais
  - Beaucoup de groupes manquants —> appli peu utile
  - Beaucoup de produits américains et français



Des questions ?