

Qualité et l'impact environmental des aliments



Statistiques faites à partir des
données openfoodfacts
6 Décembre 2019



Marco Berta
OPENCLASSROOMS

Outline

- **Introduction**

- question principale
- importance du sujet
- le valeur d'openfoodfacts pour les statistiques alimentaires

- **Nettoyage**

- Suppression des doublons
- Suppression des colonnes avec trop des valeurs nuls
- Sélection des variables

- **Analyse exploratoire**

- statistiques impact environnementale (moyenne, médiane, écart type...)
- statistiques par catégorie alimentaire et niveaux de qualité

- **Résultats des analysés**

- corrélation entre valeur nutritionnel et impact environnementale
- impact environnementale et valeur nutritionnel par catégorie alimentaire

- **Conclusions**

- 3 conclusions principales

Bien manger est bien aussi pour notre planète? Quoi éviter ?



Qualité des aliments et climat

Beaucoup des débats



Données limitées

- Pour vendre une idée vaut mieux une belle histoire que une analyse statistique
- Chaque producteur va fournir des données que pour ses produits
- Les analyses des supermarchés ne sont pas publiques (concurrence)
- Des produits de la même catégorie peuvent avoir des impacts différentes
- Pas d'accès libre pour des publications scientifiques de bon niveau

Les données libres: Openfoodfacts

Le système de Wikipedia appliqué à la nourriture : Chaque utilisateur ajoute et consulte gratuitement des données dans une base commune avec un scan de code barre par smartphone



Scan

Base des données libre
Openfoodfacts

Info qualité

Le scan en continu des codes barres permet d'obtenir instantanément le Nutri-Score (qualité nutritionnelle) et l'identification NOVA des produits ultra-transformés.

Openfoodfacts : avantages statistiques

- bonne base statistique :

≈ 630.000 produits (Nov 2019)



- statistique dynamique : base de donnée mise à jour régulièrement

```
import pandas as pd
```

```
all_food_data = pd.read_csv( "https://static.openfoodfacts.org/data/en.openfoodfacts.org.products.csv", sep="\t", encoding="utf-8") # raw dataframe
```

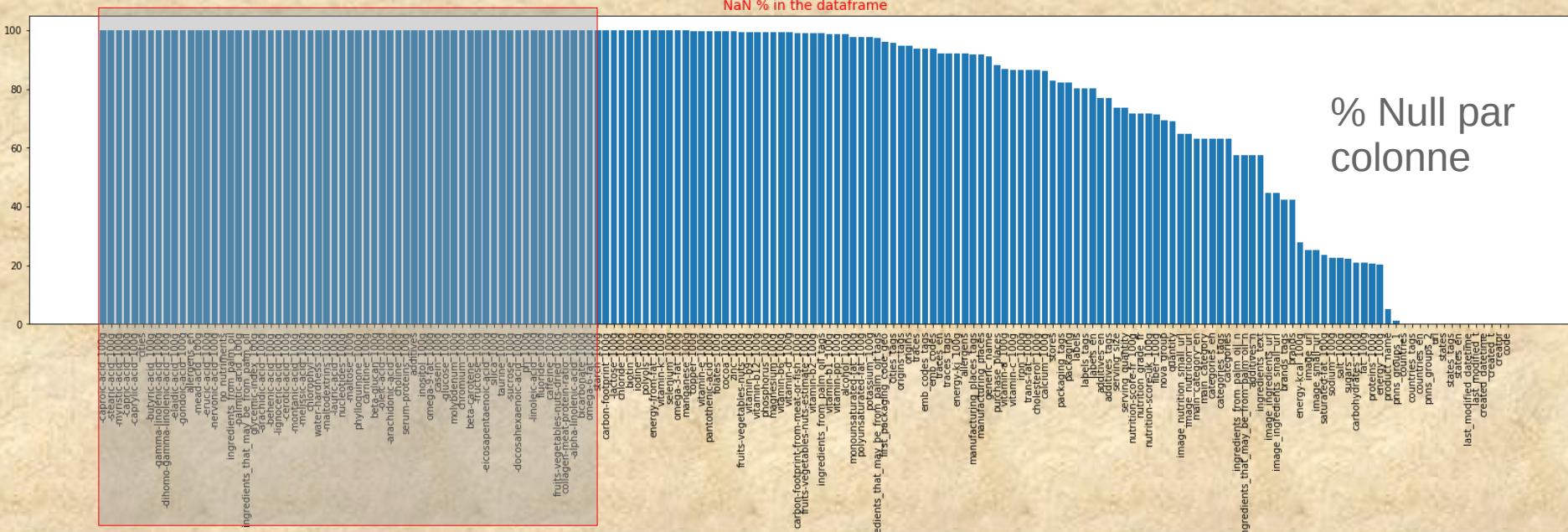
- analyse des données avec logiciel gratuit et partage libre des résultats



Nettoyage des données

Taille initiale fichier csv (spreadsheet) : 1061510 lignes X 177 colonnes

1) Suppression des colonnes vides (ou presque) (Taille : 1061510 x 114)



2) Suppression des doublons (lignes) Taille : 1061482 x 114

3) Suppression des colonnes avec information redondante Taille : 1061482 x 93

- suffixes “_en” et “_tag”



Dataframe nettoyé : 47 % de la taille originale

Sélection des variables

pour estimer une corrélation entre qualité et impact environnementale

Informations sur produits/catégories alimentaires

1) Nom du produit

2) Catégories PNNS_2* :

- pas trop généraliste
- faible %null
- pas catégories pour produits individuels

Impact environnementale

1) Empreinte carbone : variable principale



2) Numéro ingrédients avec huile de palme



3) Numéro ingrédients possiblement avec avec huile de palme

Qualité nutritionnelle

1) nutrition_grade_fr



2) Valeur associé au nutriscore :

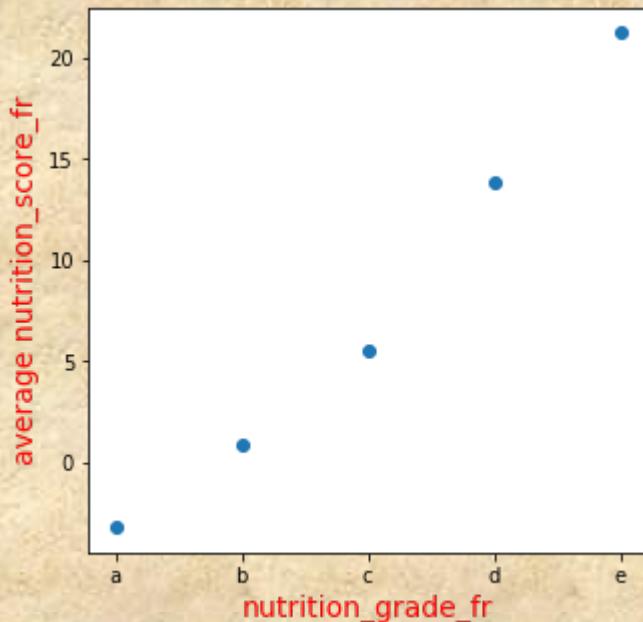
- chiffre calculé à partir de la variable "nutrition_score" pour chaque 'nutriscore'

* Pnns = programme national nutrition et santé
<https://www.mangerbouger.fr/PNNS>

Qualité nutritionnelle en chiffres

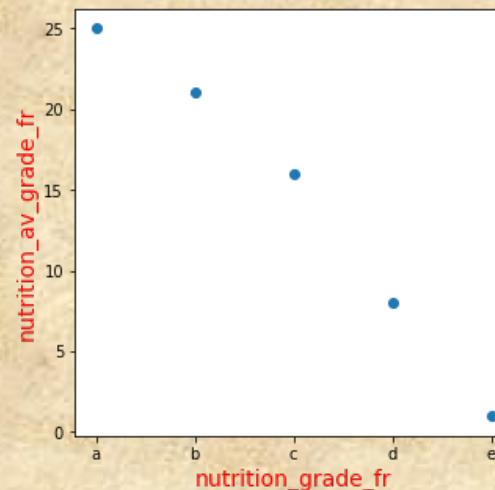
Assigner un chiffre à chaque nutriscore nous permettre de calculer des corrélations avec les variables environnementales

Corrélation entre le valeurs 'nutrition grade' et "nutrition score"
dans le dataframe : proportionnalité inverse



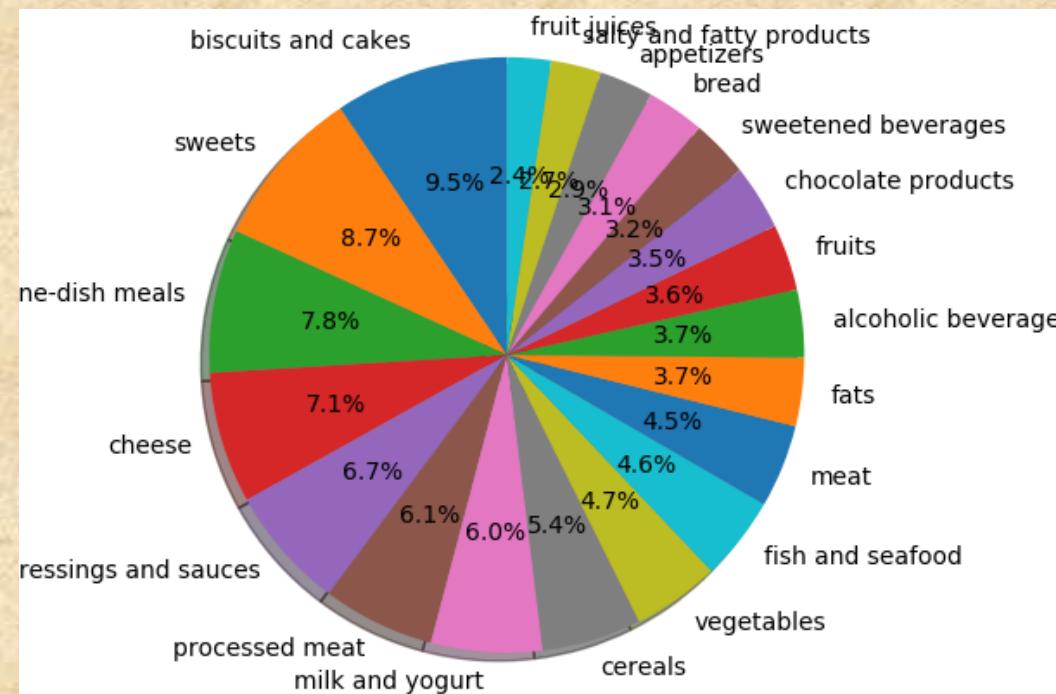
```
'df_food_grades['nutrition_grade_fr_n'] = 22 - df_food_grades['nutrition-score-fr_100g']  
df_food_grades = df_food_grades.round(0)  
df_food_grades.head()'
```

	nutrition_grade_fr	nutrition-score-fr_100g	nutrition_grade_fr_n
0	a	-3.0	25.0
1	b	1.0	21.0
2	c	6.0	16.0
3	d	14.0	8.0
4	e	21.0	1.0

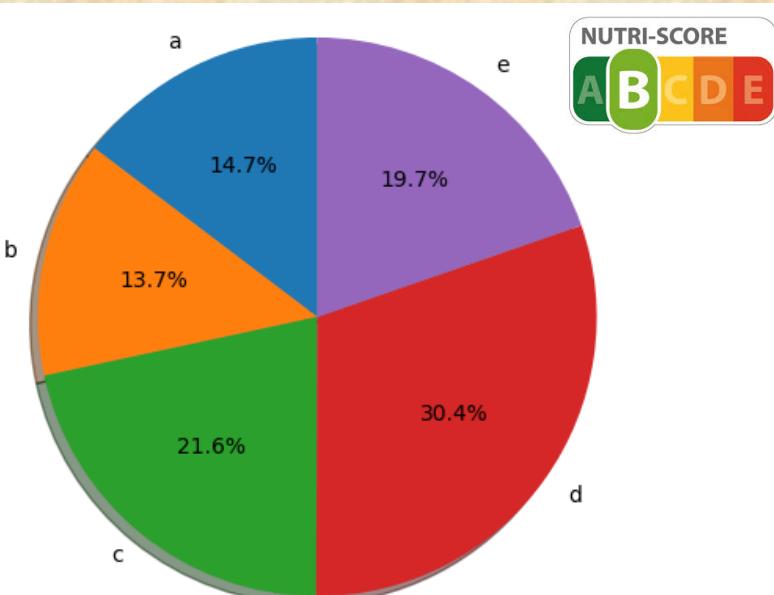


Nouvelle variable,
directement proportionnelle

Répartition des catégories nutritionnelles

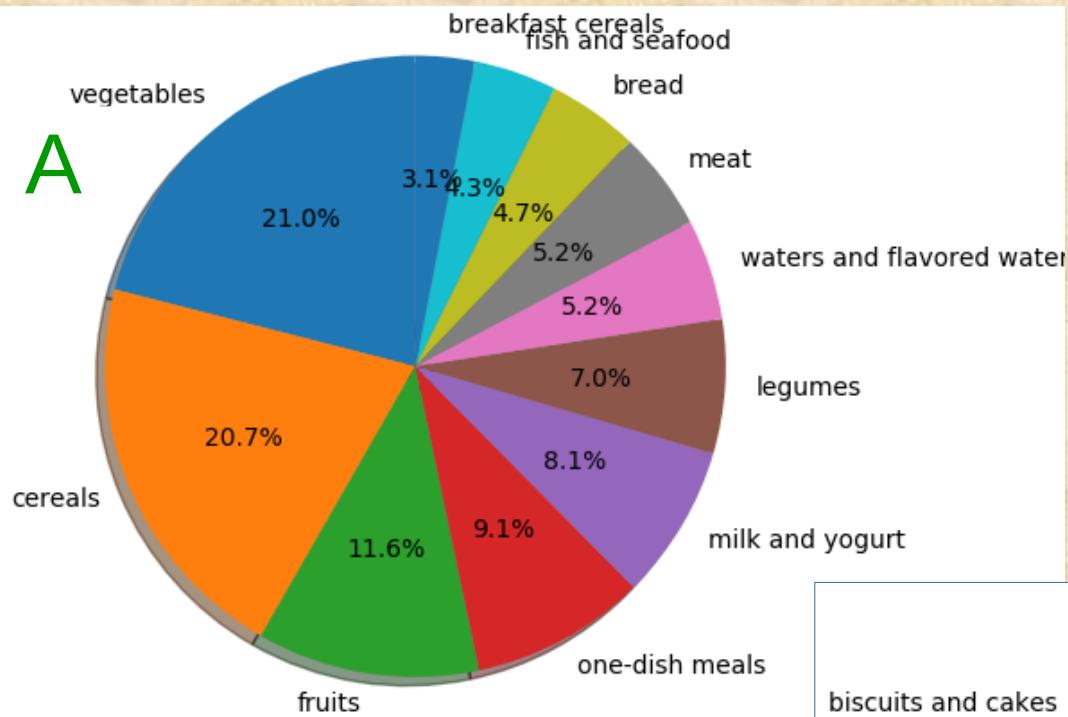


Assez des catégories principales. Répartition équilibrée pour le nutriscore aussi

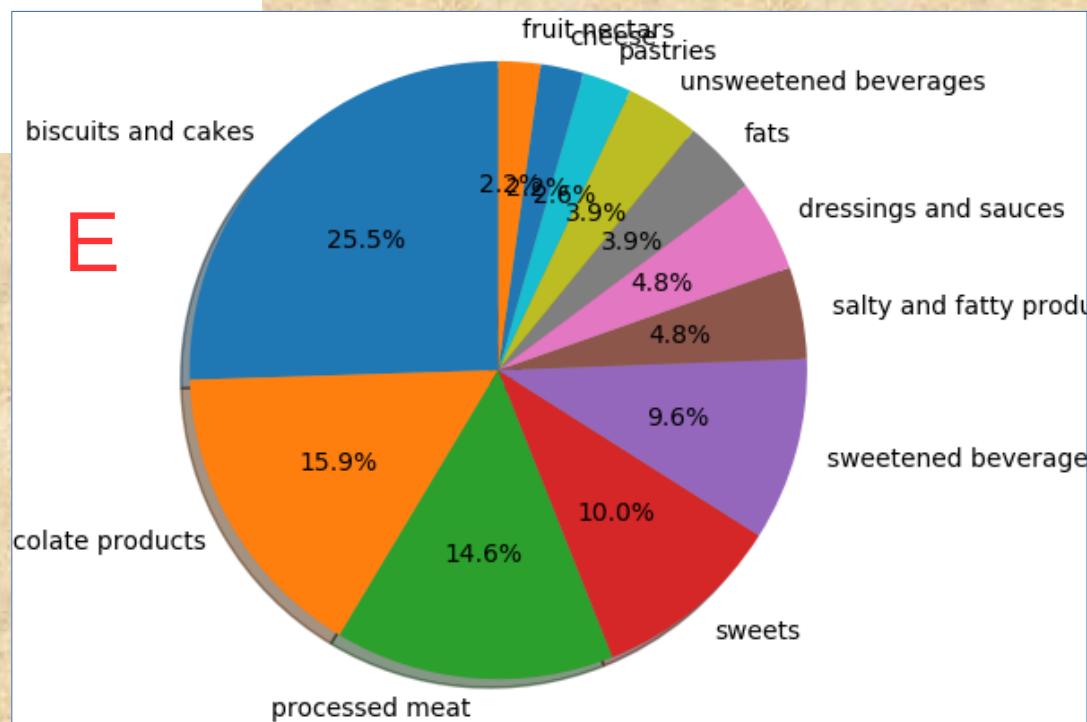


Aliments à préférer et à éviter

A



E

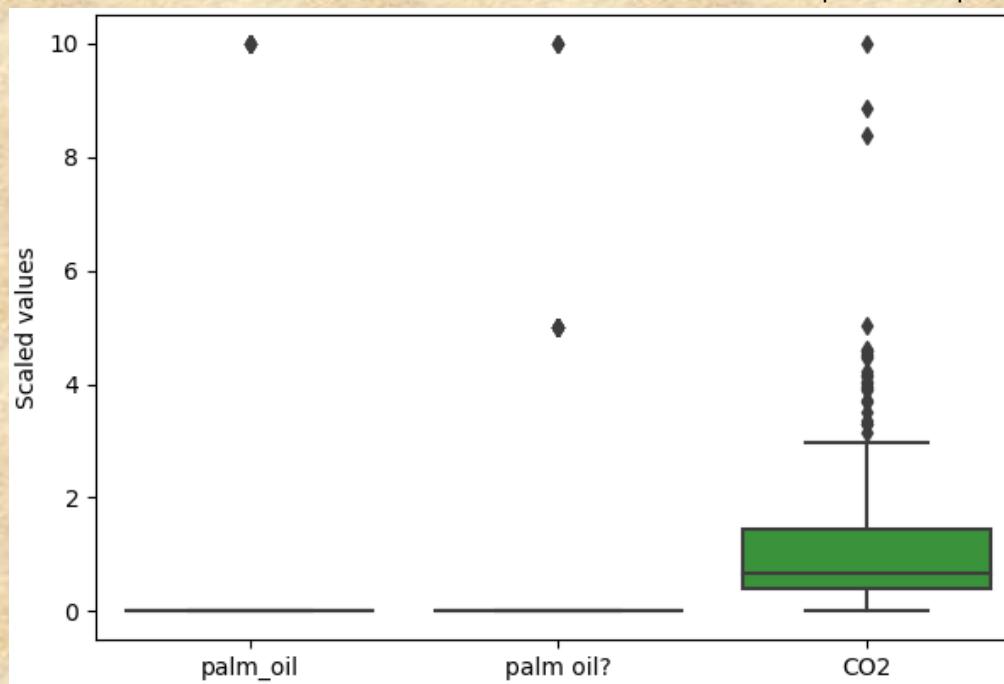


Statistiques des variables environnementales

	Ingredients avec huile de palme	Ingredients possiblement avec huile de palme	Empreinte carbone (100g)
Num. Elements dans la catégorie	453,482.00000	453,482.00000	406.00000
Moyenne	0.02713	0.07199	230.97270
Écart type	0.16442	0.31701	363.44733
Minimum	0.00000	0.00000	-0.00028
Maximum	3.00000	6.00000	2,842.00000
25% quantile	0.00000	0.00000	0.00000
Mediane	0.00000	0.00000	109.40000
75% quantile	0.00000	0.00000	293.22500

Beaucoup des zeros (produits sans huile de palme)

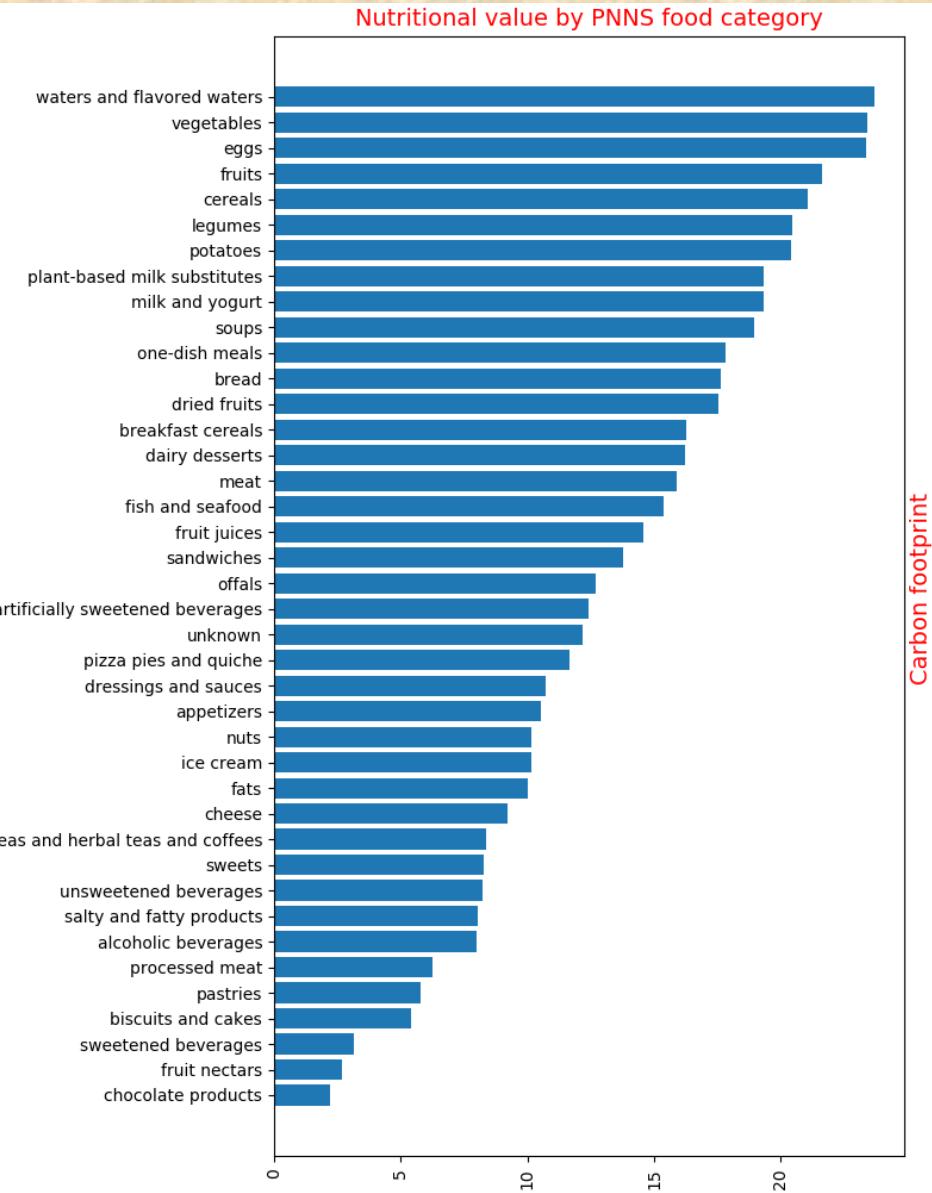
Valeurs < 10 ou negative
pas réaliste



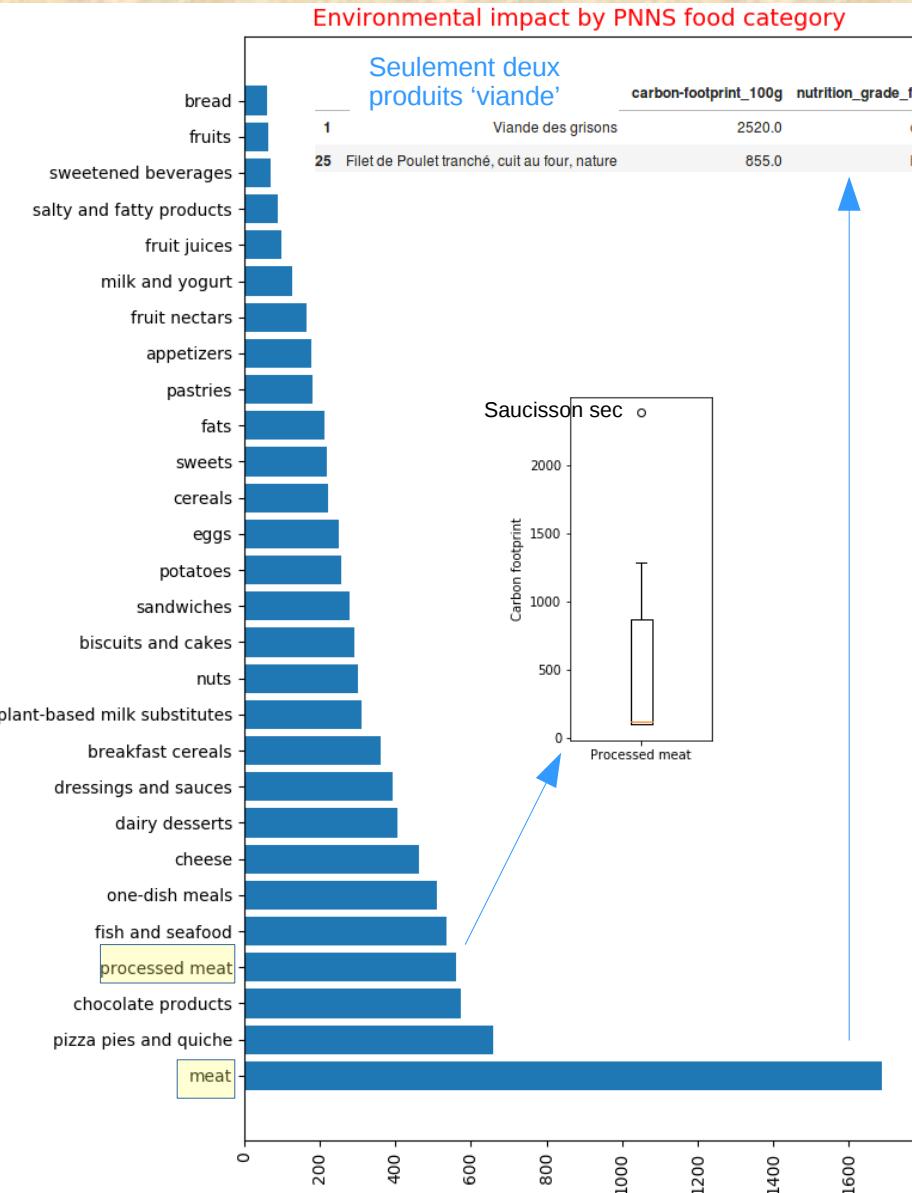
À supprimer

Le mieux et le pire, pour santé et environnement

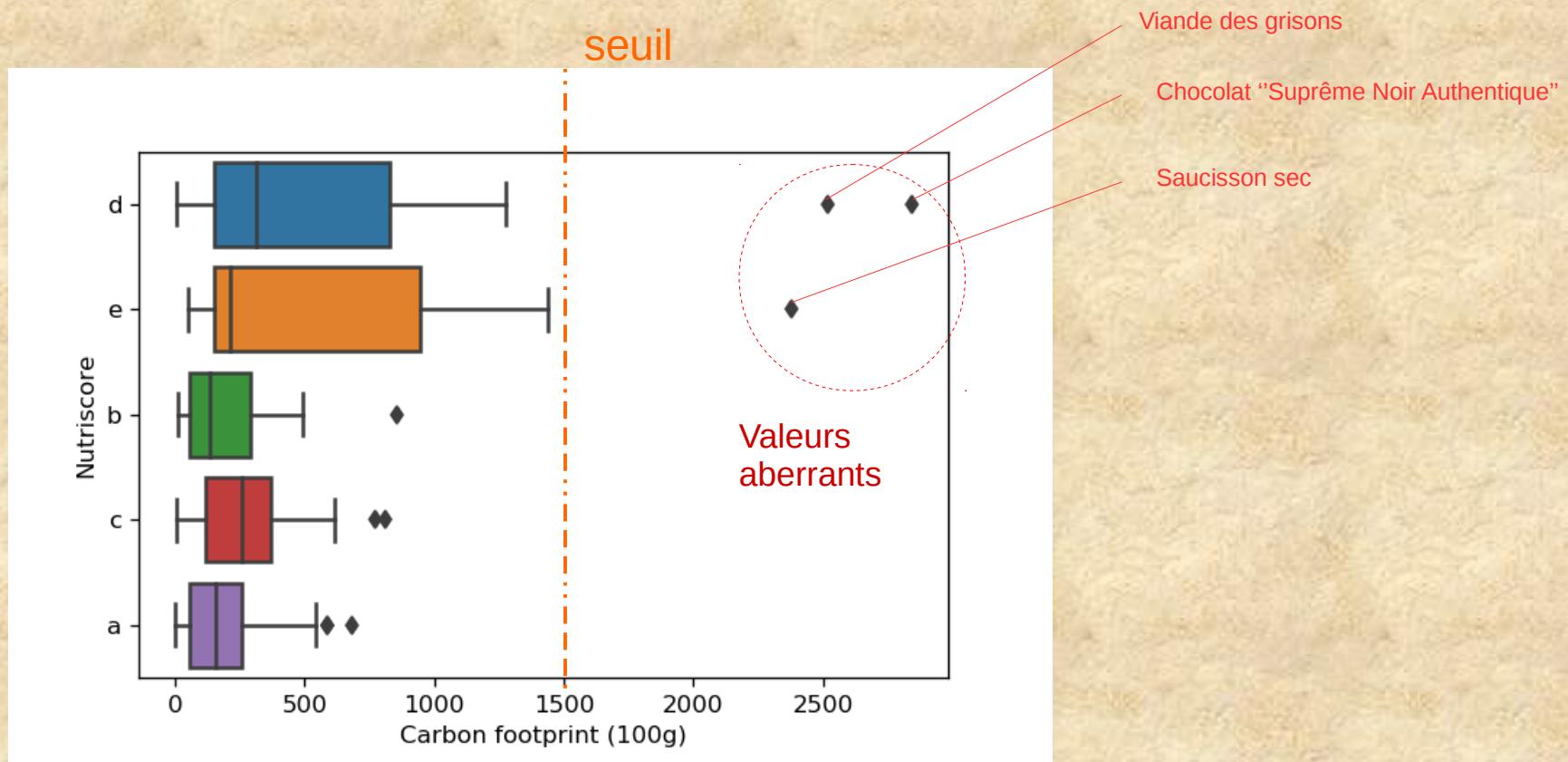
Average nutrition score



Carbon footprint



Empreinte carbone pour chaque nutriscore

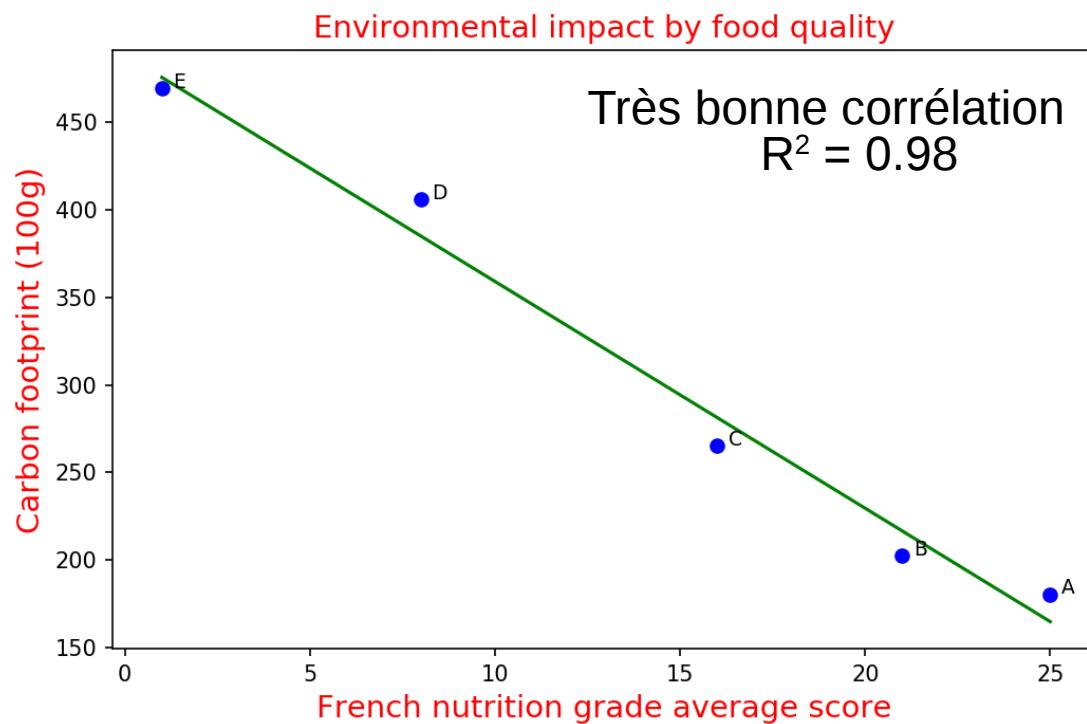


On prend que des valeurs de empreinte carbone pour notre statistique < 1500

Regression univariée

	nutrition_grade_fr_n	carbon-footprint_100g	french food grade
0	25.0	180.508039	A
1	21.0	202.532000	B
2	16.0	265.320889	C
3	8.0	493.466437	D
4	1.0	502.754386	E

```
groupby(['nutrition_grade_fr_n']).mean()
```



Régression multivariée: 3 variables environnementales (X) vs. Nutriscore (y)

Empreinte carbone

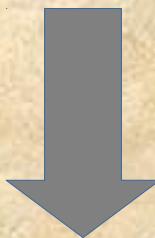
Ingrédients avec huile de palme

Ingrédients possiblement avec huile de palme

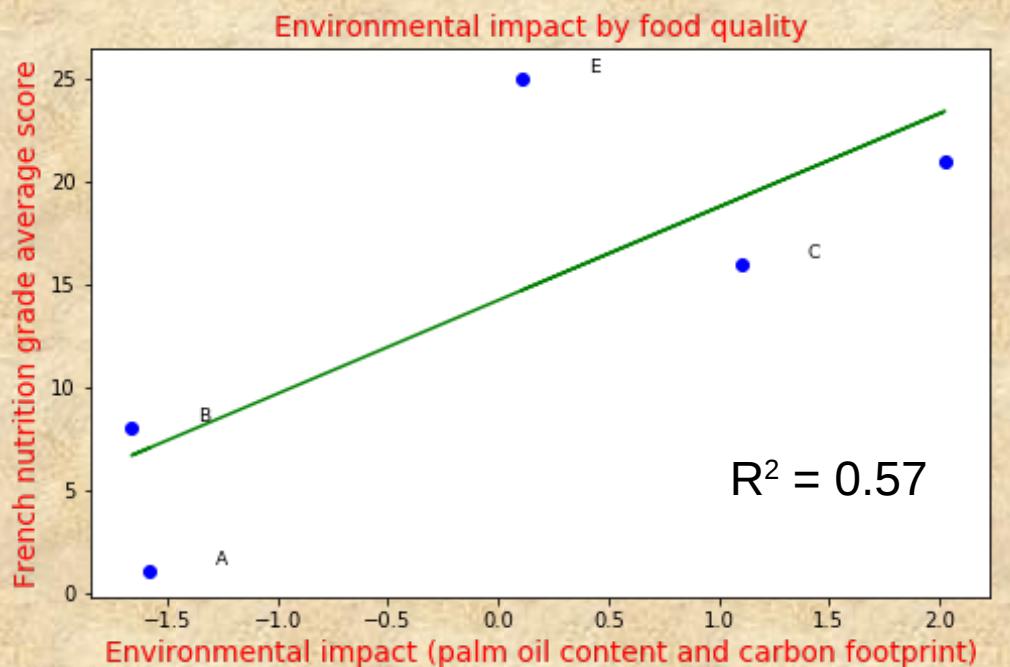
régression linéaire
multivariée

$$R^2 = 0.81$$

Analyse composants principaux : on réduit 3 paramétrés en 1



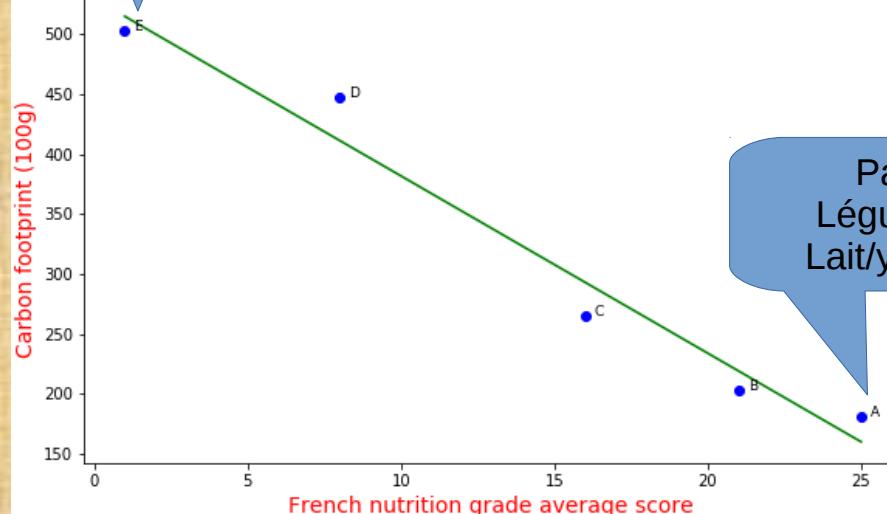
Impact Environnementale
70 % information totale



Chocolat
Viande
Poisson

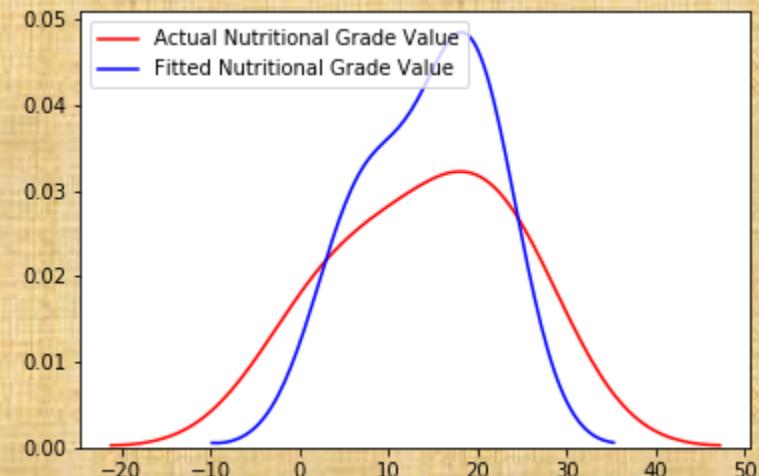
Conclusions

Environmental impact by food quality



Régression multivariée

la qualité estimée à partir des données environnementales est proche de cela mesurée



- Le nutriscore est bien corrélé à l'empreinte carbone
- Le numéro des ingrédients avec huile de palme aussi : bien manger est bon pour l'environnement
- Le top serait un régime végétarien de bonne qualité sans trop des gâteaux

Questions?

