NUTR'AVEL: Une app de santé publique au service des voyageurs

Cécile Guillot – Ingénieure Machine Learning

01

Présentation de l’application

Nutr’avel

02

Nettoyage

Nettoyons nos données pour mieux s’y retrouver

03

Analyses descriptives

Décrivons nos données pour mieux les comprendre

04

Analyses multivariées

Plongeons un peu plus dans l’exploration des données

05

Construction de l’algorithme de Nutr’avel

06

Faisabilité et conclusion

01
L’application Nutr’avel

Une classification universelle

Deux utilisateurs...

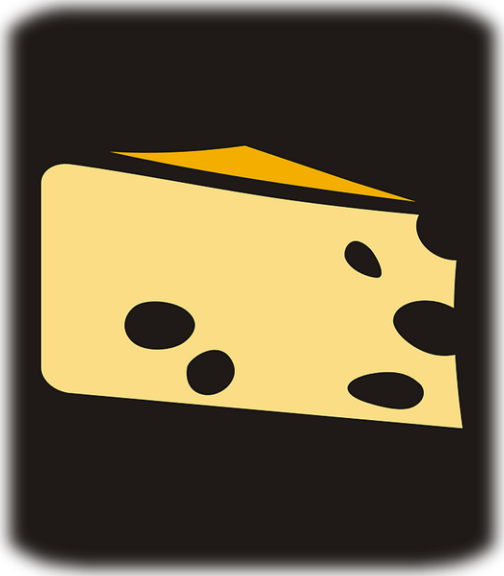
Louis

Globe-trotter

Emilie

Etudiante en échange Erasmus

A propos de Nutr’avel



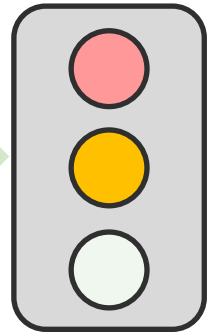
Une classification universelle

- Utilisation du code-barre pour identifier un produit
- Attribution d’un groupe sur la base d’un algorithme
- Projet open source donc transparence de l’algorithme

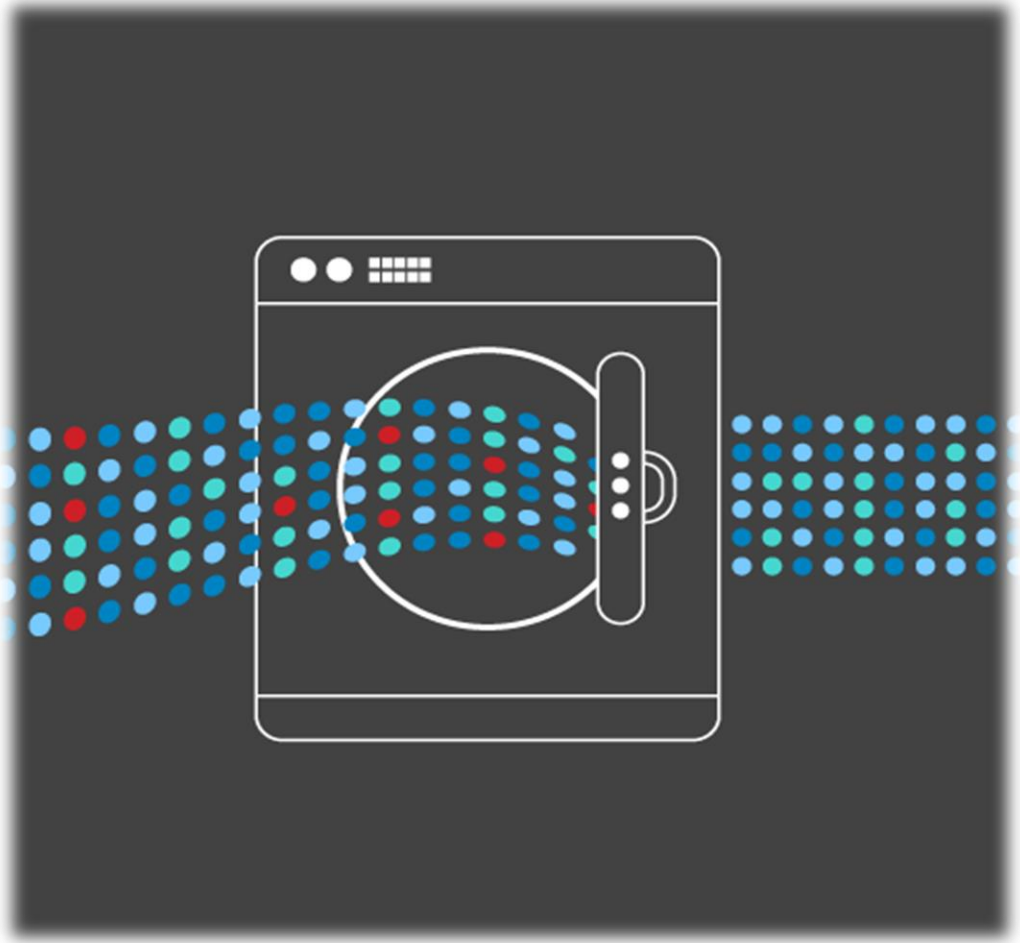
Fonctionnement de l’application

- Reconnaissance du produit
- Matching avec base de données « OpenFoodFacts » pour récupérer les données nutritionnelles

Algorithme de Nutr’avel



Score Nutr’avel



02

Nettoyage

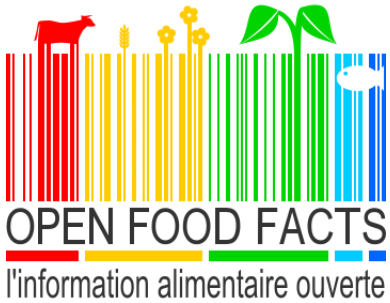
Les outils et les données

Les outils

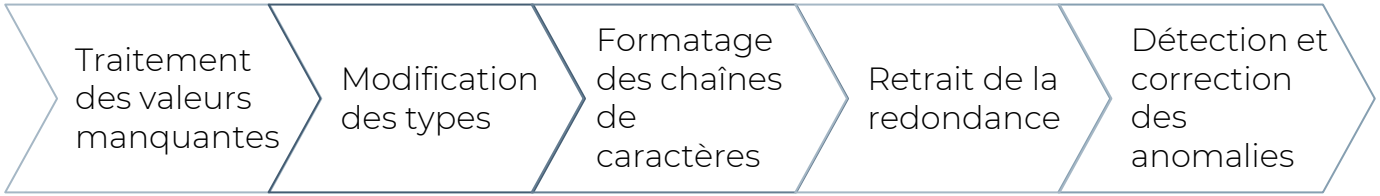
- Jupyter Notebook
- Bibliothèques de Data Sciences en Python (Pandas, NumPy, Matplotlib, Seaborn)

Les données

- Jeu de données assez importants (4 Gb)
- Comporte différentes informations sur des produits alimentaires (identification, composition, etc.)



Les étapes du nettoyage

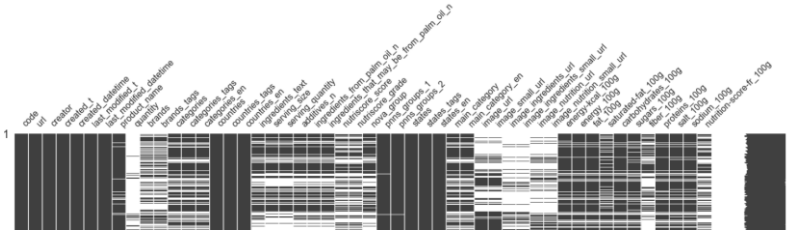


- Chronologie « faussement » linéaire
- Aller-retours entre nettoyage et exploration
- Fonctionne sur le système d’essais/erreurs

Traitement des valeurs manquantes

Une étape pour faciliter la lecture

- Choix du retrait des colonnes (variables) avec plus de 75%
- Conservation de 50 colonnes

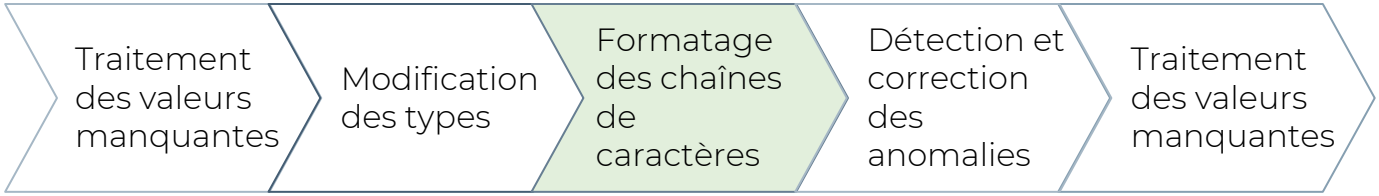


Modification des types

Faciliter l'imputation lors des prochaines étapes

- Passage à des types *float*
- Imputation de la médiane puis passage à type *int*
- Passage à type *datetime*

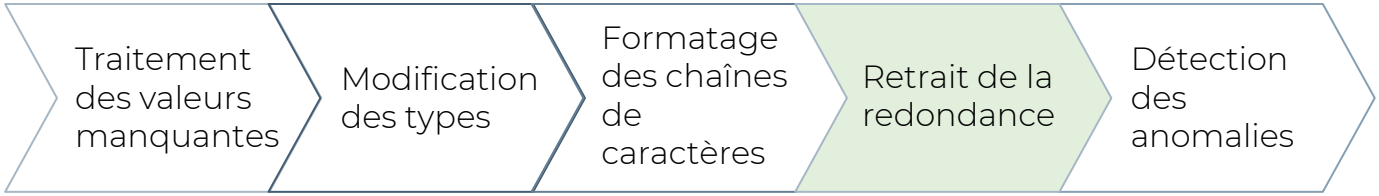
Formatage des chaînes de caractères



Harmonisation des chaînes de caractères

- Passage de toute la chaîne en minuscule
- Attribution d'une majuscule en 1ère position

Retrait de la redondance des informations



Eviter les informations en double ou triple voire +

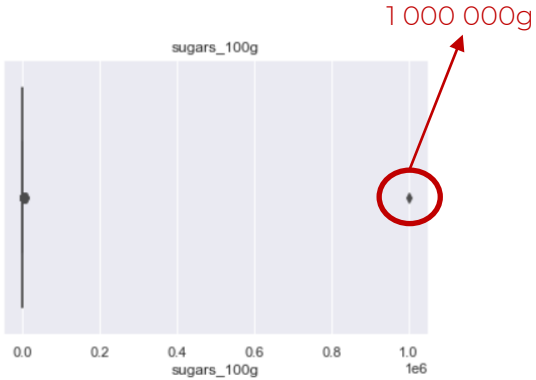
- Redondance de colonnes (infos en anglais, tags, timestamps)
- Redondance des entrées

Détection et correction des anomalies

Traitement des valeurs manquantes → Modification des types → Formatage des chaînes de caractères → Retrait de la redondance → Détection et correction des anomalies

Mettre de la cohérence dans nos données

- Détection visuelle (*boxplot*) et numérique
- Suppression de valeurs pour les lipides, glucides et protéines
- Suppression de valeurs pour l'énergie
  - Utilisation de connaissances métiers



Traitement des valeurs manquantes (suite et fin)

Bis repetita

- Retrait des lignes : Abs. de nom du produit
- Imputation à la médiane par groupe PNNS 2
- Remplacement de NaN par ‘inconnu’ ou ‘autres’
- Mise en place de 2 dictionnaires pour harmoniser noms de produits et pays

- Après nettoyage

```
df_clean_median.shape
```

```
(1751837, 33)
```

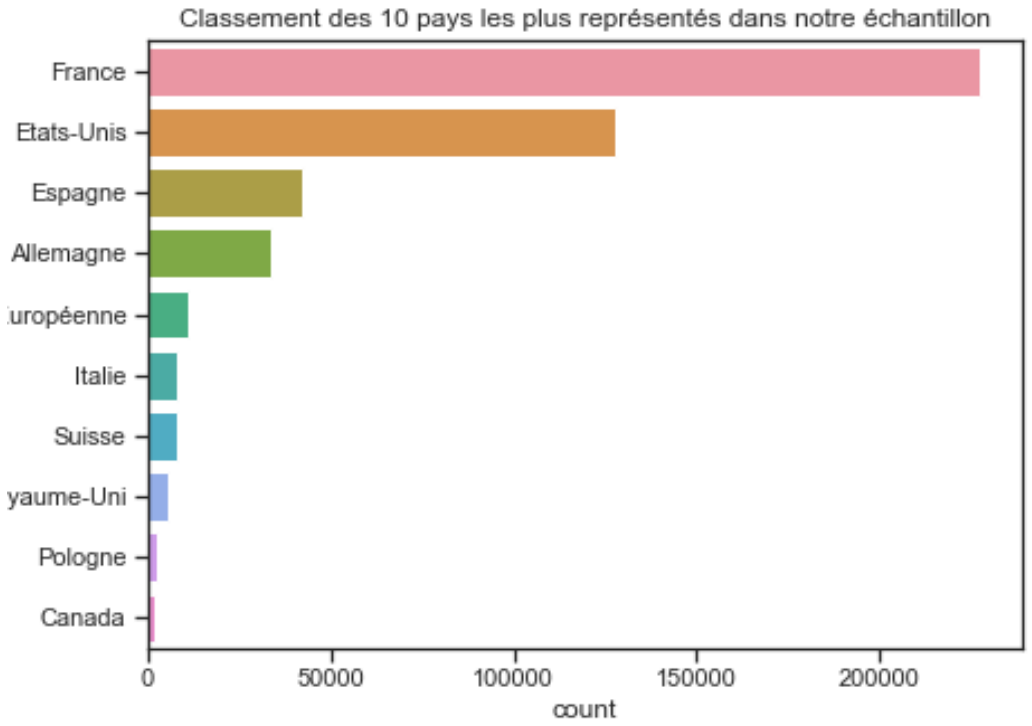
03
Analyses descriptives

Les données

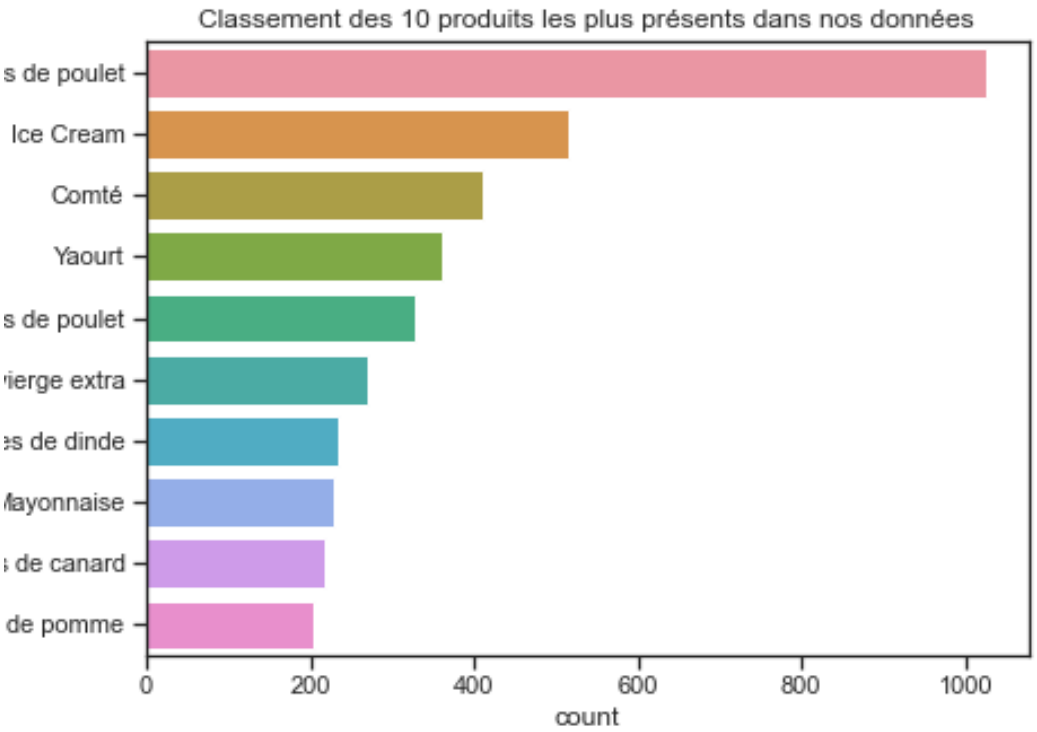
Types de données	Colonnes
Identification	Nom du produit, marques, pays de commercialisation, catégories
Informations nutritionnelles	Protéines, Glucides et Lipides (+ sucres, graisses saturées, etc.)
Composition	Additifs, huile de palme
Système d’évaluation	Nutriscore, Groupe NOVA

Les pays

La France est le pays où l'on retrouve le plus de produit.



Les produits

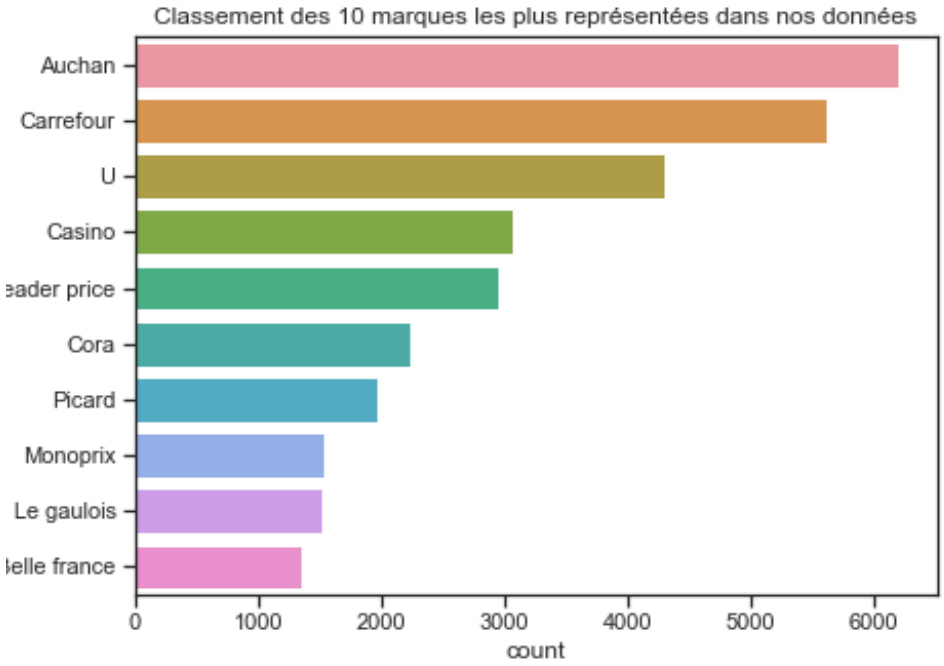


La viande de volaille est un des produits les plus consommés.*

**Sur un total de 227 393 produits*

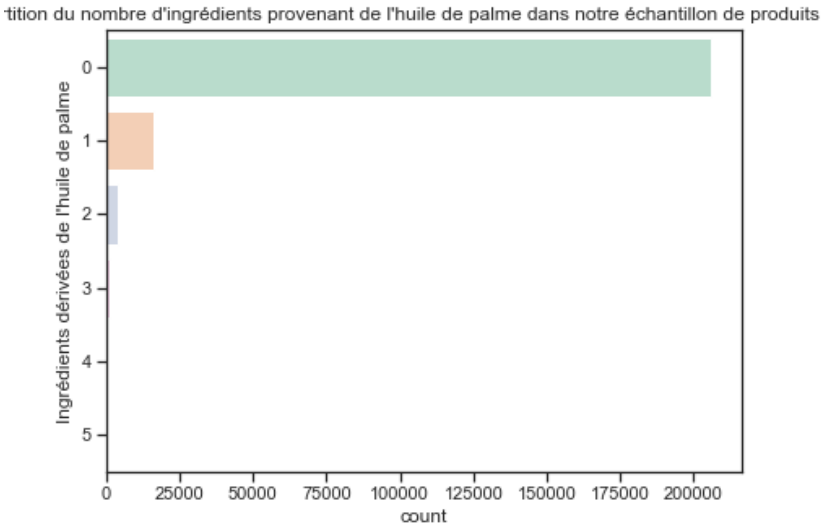
Les marques

Les marques distributeurs sont les plus représentées.*



*Sur un total de 36 180 marques

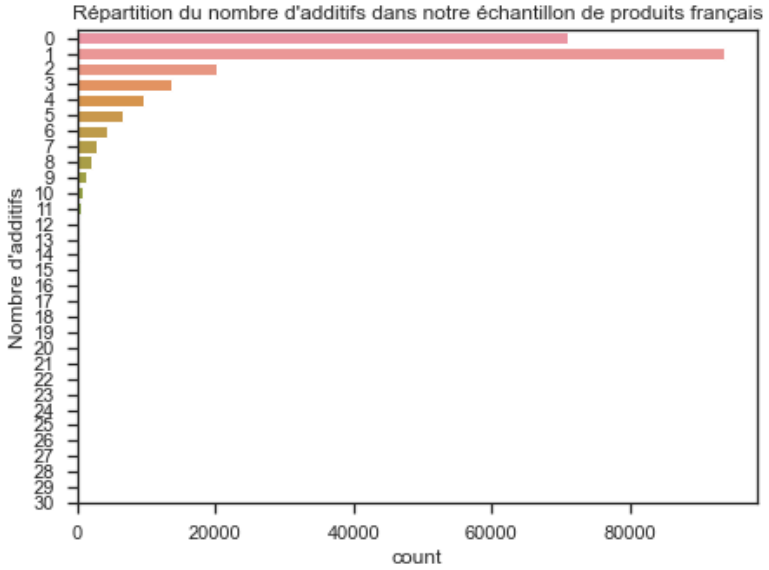
Huile de palme



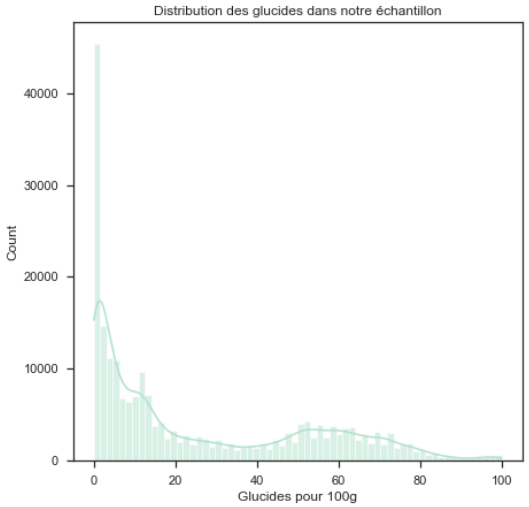
Environ 91 % des produits ne contiennent pas d'huile de palme.

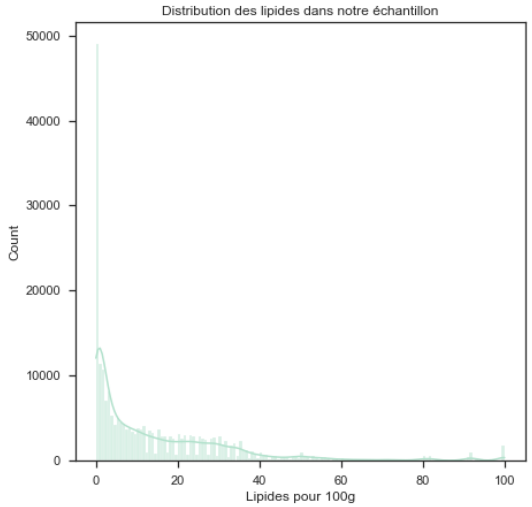
Les additifs

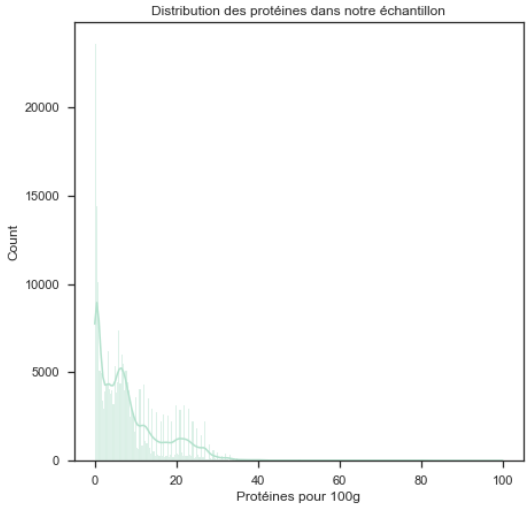
68% des produits contiennent au moins un additif.



Distribution des variables quantitatives

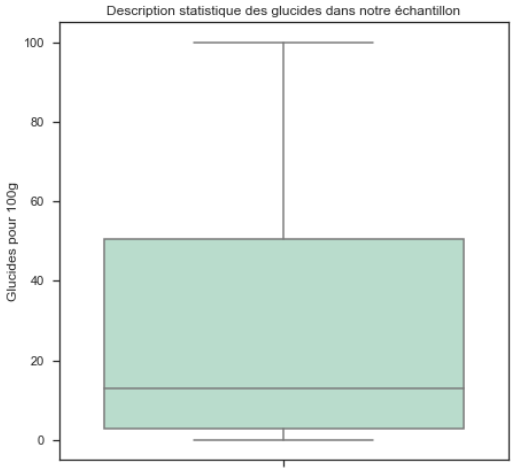




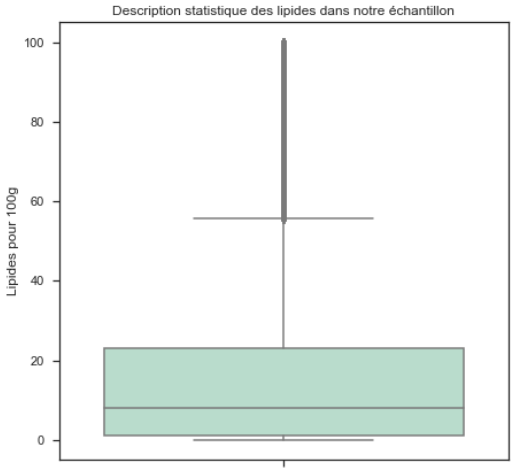


- Test de la normalité : Kolmogorov-Smirnov
- Test de l'homogénéité des variances : Levene

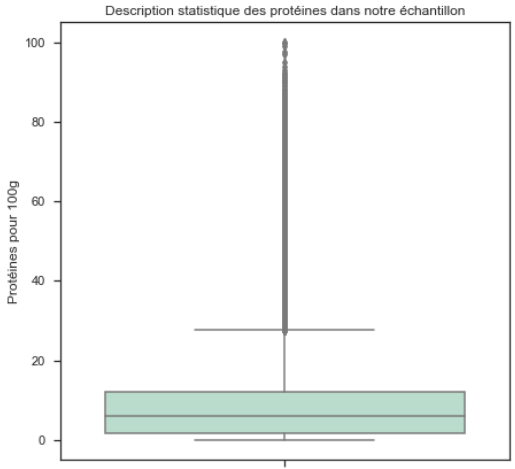
Description statistique des données quantitatives



Moyenne : 25.7
Médiane : 13



Moyenne : 14.7
Médiane : 7.9



Moyenne : 8.4
Médiane : 6.1

04
Analyses multivariées

Hypothèses de travail

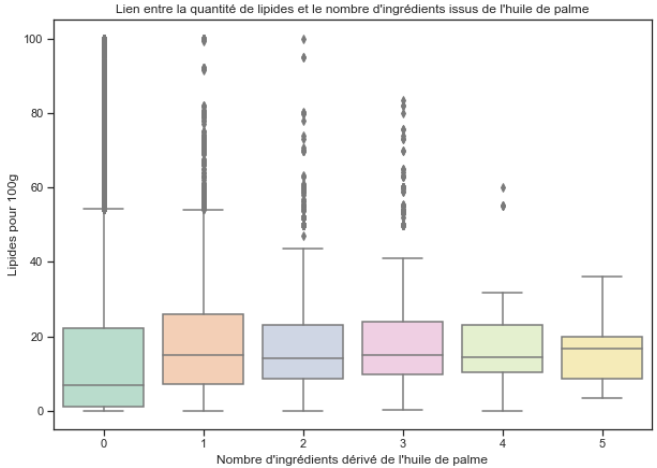
Lien entre ingrédients (additifs, huile de palme) et données nutritionnelles et énergétique

- ANOVA à un facteur entre données qualitatives et quantitatives
- Boxplot

Utilisation des données nutritionnelles, énergétiques et de la composition pour créer un indice universel

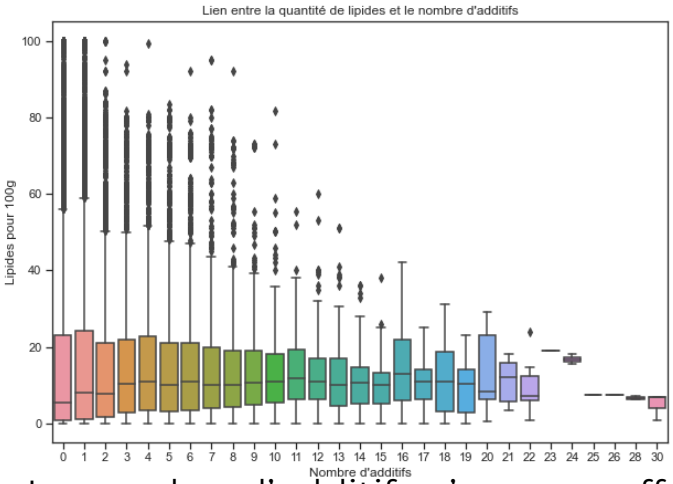
- Réduction de dimensions (PCA)
- Clustering (K-Means)

Lipides, huile de palme et additifs



Le nombre d'ingrédients dérivé de l'huile de palme n'a aucun effet sur le taux de lipides.

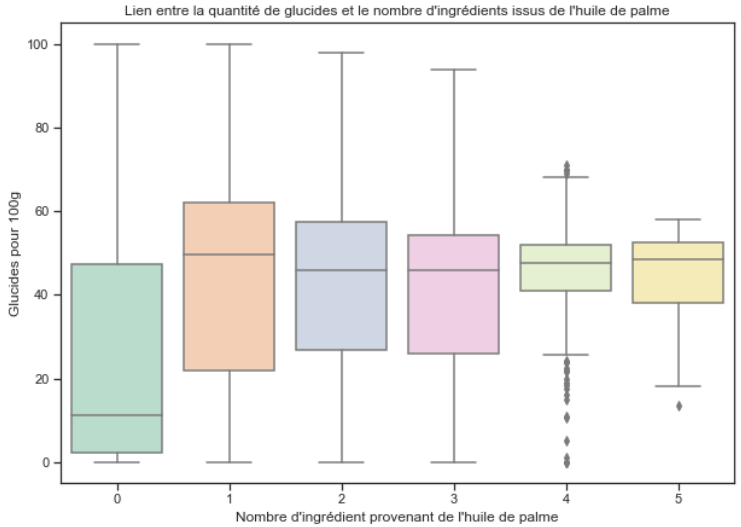
ANOVA : η^2 : .002, p-value < 0.05



Le nombre d'additifs n'a aucun effet sur le taux de lipides.

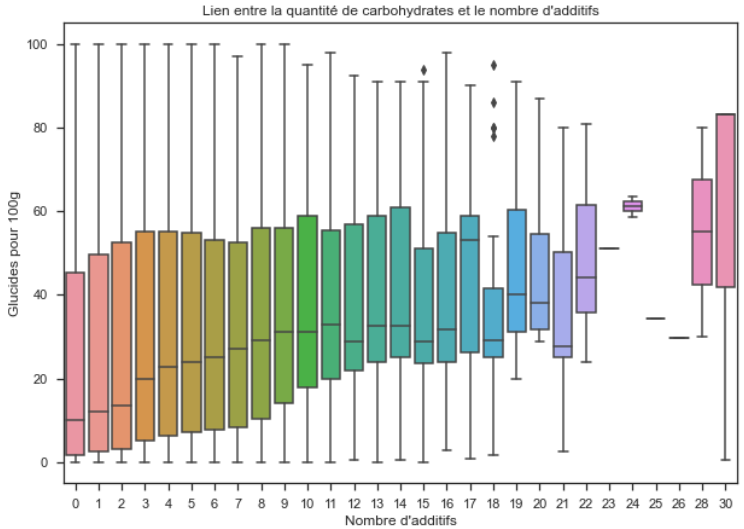
ANOVA : η^2 : .001, p-value < 0.05

Glucides, huile de palme et additifs



Le nombre d'ingrédients dérivé de l'huile de palme a un effet modéré sur le taux de glucides.

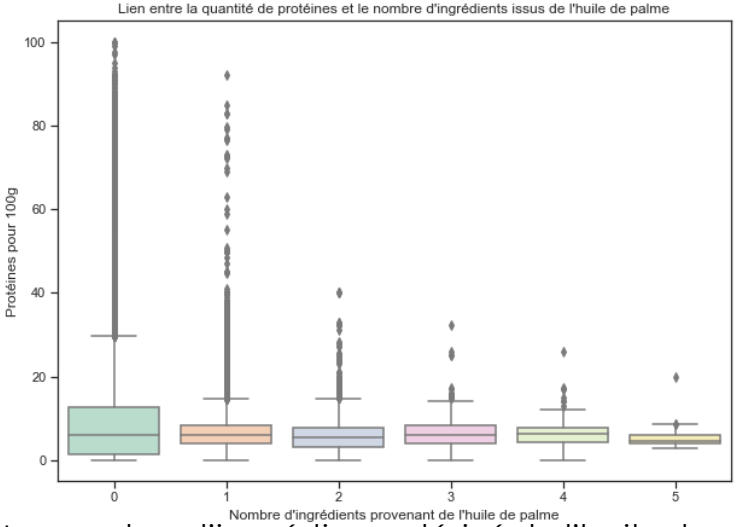
ANOVA : η^2 : .04, p-value < 0.05



Le nombre d'additifs a un faible effet sur le taux de glucides.

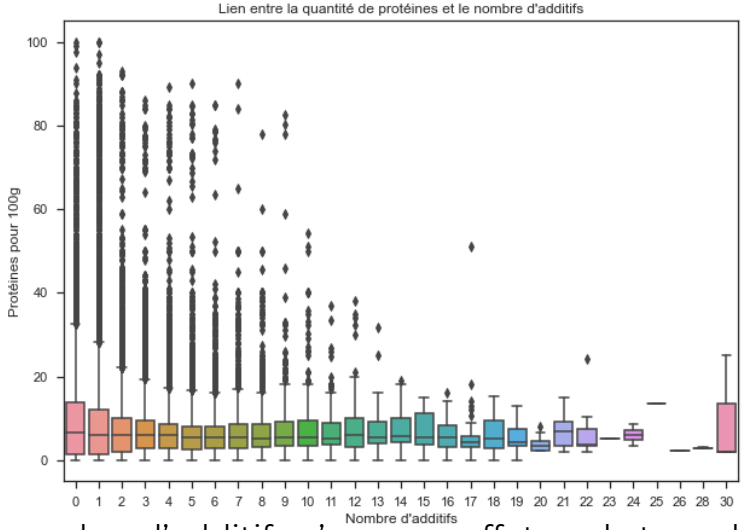
ANOVA : η^2 : .01, p-value < 0.05

Protéines, huile de palme et additifs



Le nombre d'ingrédients dérivé de l'huile de palme n'a aucun effet sur le taux de protéines.

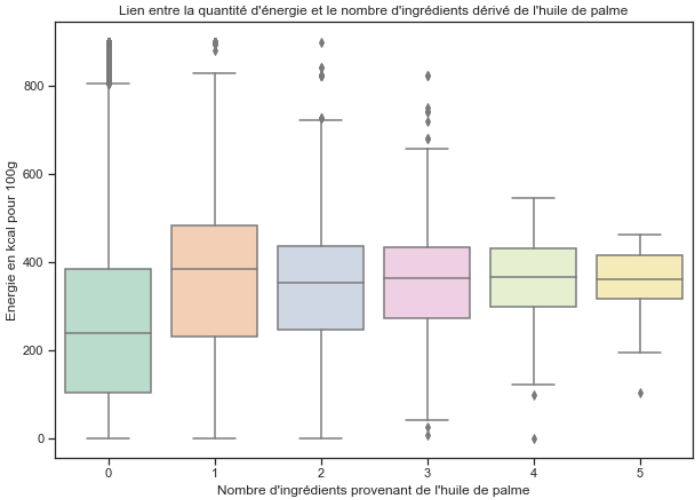
ANOVA : η^2 : .004, p-value < 0.05



Le nombre d'additifs n'a aucun effet sur le taux de protéines.

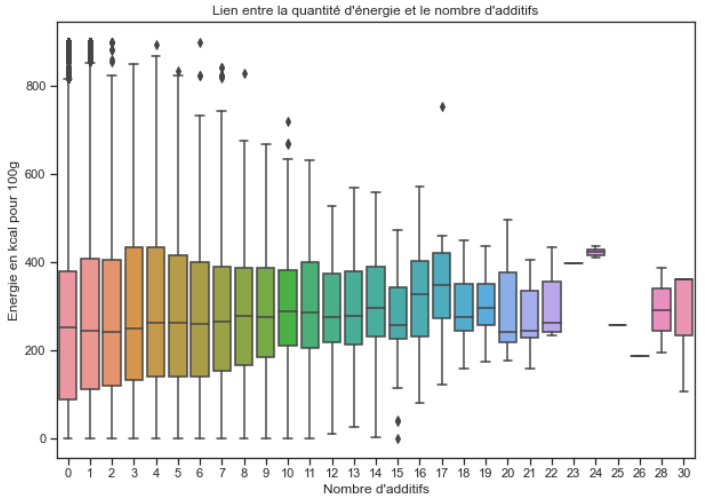
ANOVA : η^2 : .006, p-value < 0.05

Calories, huile de palme et additifs



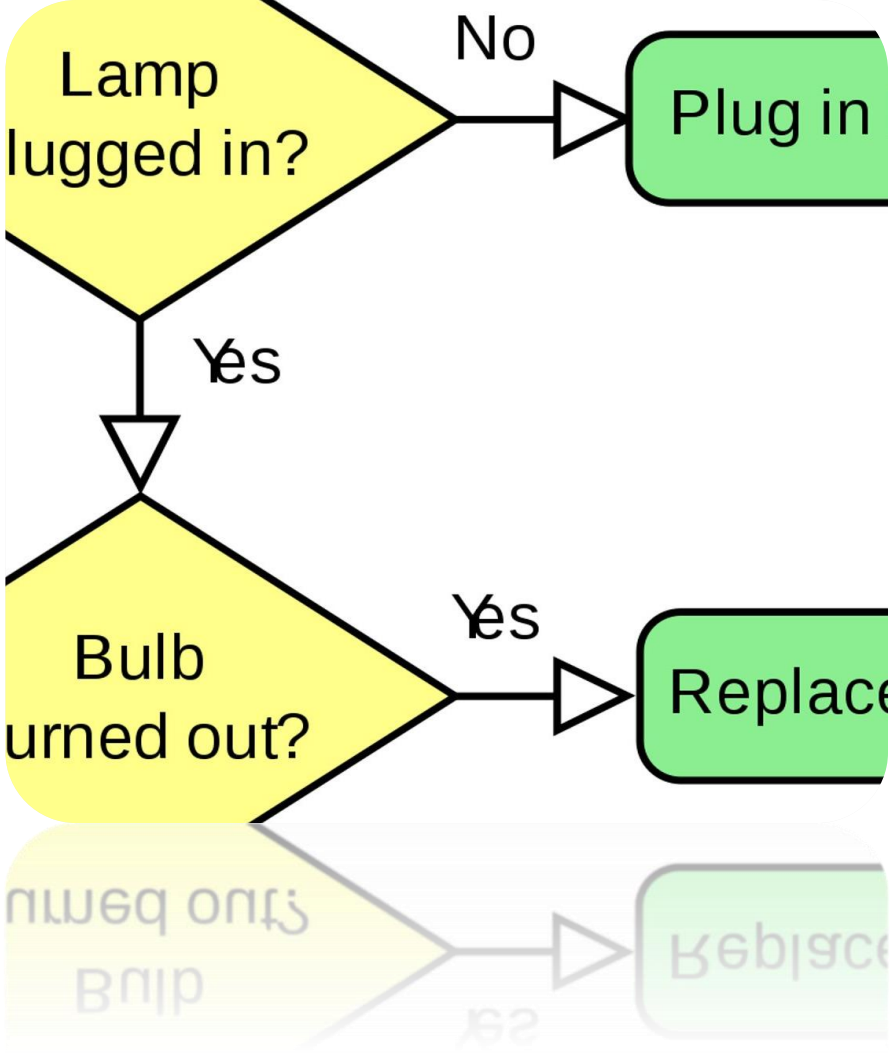
Le nombre d'ingrédients dérivé de l'huile de palme a un effet faible sur le nombre de calories.

ANOVA : η^2 : .019, p-value < 0.05



Le nombre d'additifs n'a aucun effet sur le nombre de calories.

ANOVA : η^2 : .0006, p-value < 0.05



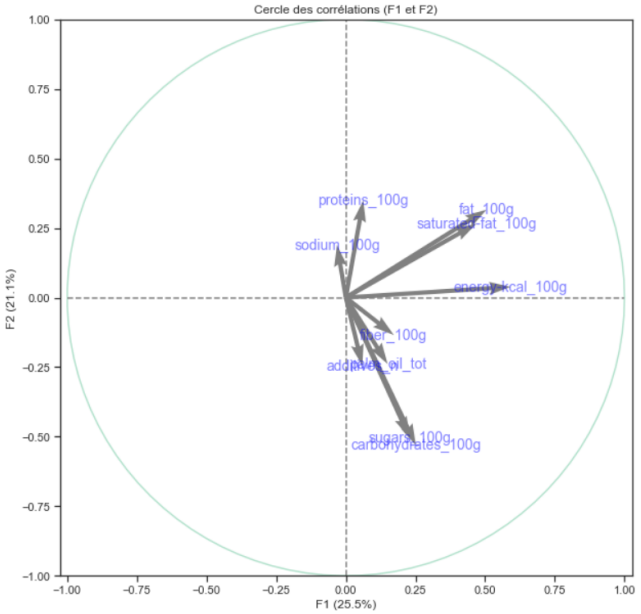
05

L’algorithme de Nutr’avel

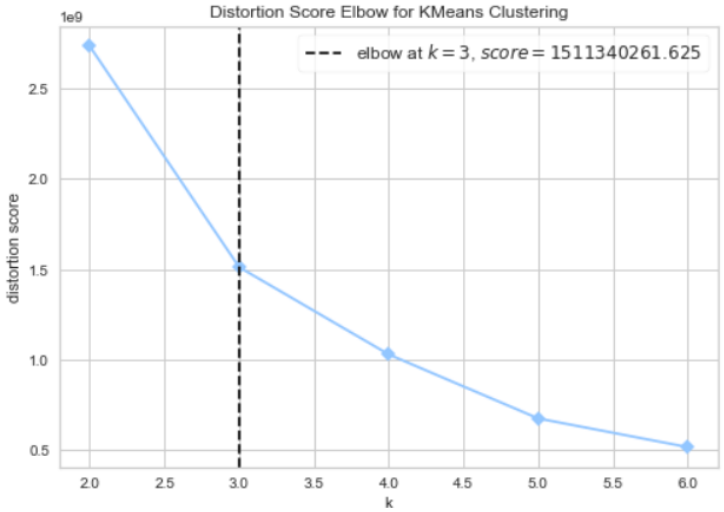
Réduction de dimensions

De 10 à 2 dimensions

- F1 : 25.2%
  - Energie, Glucides et graisses saturées
- F2 : 21.2%
  - Protéines, sodium (+) / Lipides, sucres, additifs, huile de palme et fibre (-)



Choix du nombre de clusters

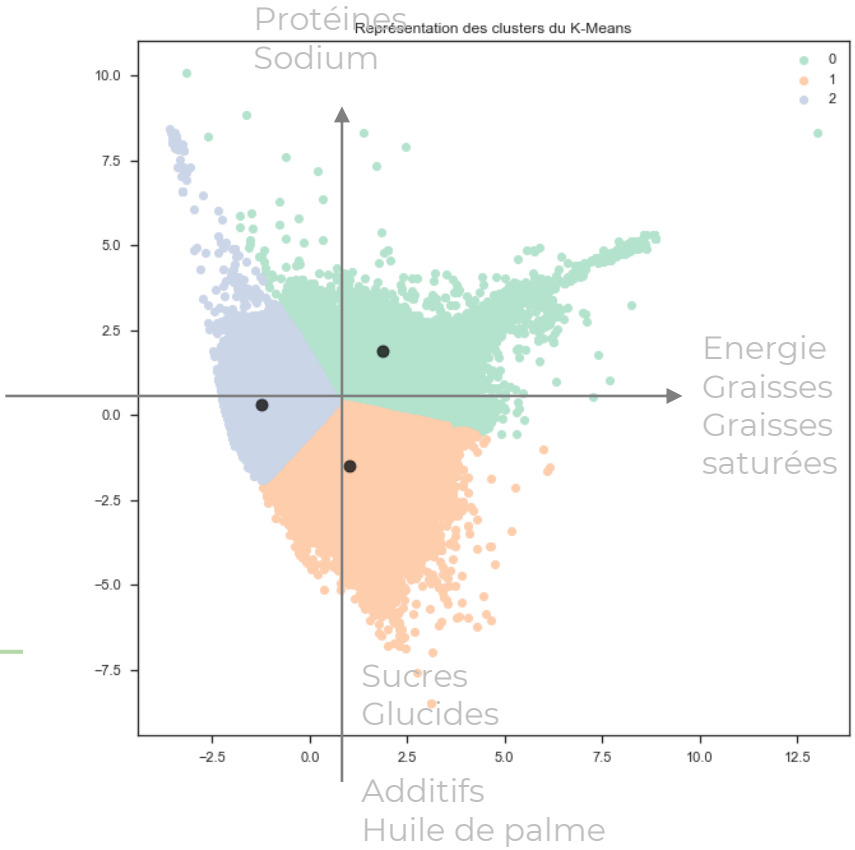


- Etude des distorsions et méthode du coude :
  - Meilleur K : 3

K-Means et représentation

PCA et K-Means

- Choix de k = 3
- Réduction de dimensions via PCA
- Score de silhouette : 0.48
- Score de Davies-Bouldin : 0.77



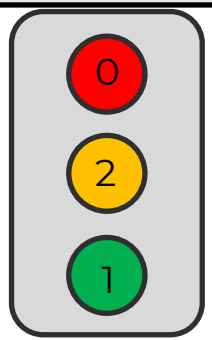
Répartition des produits

- 52% des produits dans le cluster 1
- 32% des produits dans le cluster 0
- 16% de produits dans le cluster 2

Description statistique des clusters

	Calories	Glucides	Lipides	Protéines	Fibres	Sucres	Graisses saturées	Sodium	Huile de palme	Additifs
0	389	9	60	6	3	31	6	0,20	0,4	2
1	131	59	11	7	1	5	2	0,43	0,03	1,3
2	492	11	8	14	2	4	17	0,54	0,04	0,9

Cluster 0 : Produits sucrées et peu respectueux de l'environnement
Cluster 1 : Produits sains et respectueux de l'environnement
Cluster 2 : Produits gras



06
Faisabilité et conclusion

Faisabilité

- Projet réalisable :
  - Se base sur une large base de données (200 000 produits uniquement pour la France)
  - Algorithme prêt à être déployé

- A faire par la suite :
  - Intégrer la reconnaissance d'un produit par son code-barre pour pouvoir retrouver ces données
  - Ajouter une imputation (via un Imputer) dans l'algorithme en cas de données manquantes

Conclusion

- Possibilité de discriminer des produits en catégorie selon les données nutritionnelles, énergétiques et composition
- Création d'un indice en 3 catégories qui permet de savoir à quel point notre produit est bon
- Possibilité d'utiliser cet indice de manière internationale car se base sur des informations obligatoires et fiables

- Possibilité d'améliorer l'algorithme grâce à des features engineering : lipides/graisses saturées ; glucides/sucres ; données binaires pour huile de palme et/ou additifs
- Application collaborative : possibilité d'ajouter des données manuellement


LOW SODIUM SALT
BUY1GET1 FREE
₹ 55
BUY1GET1 FREE
BUY1GET1 FREE
₹ 75
HOT OFFER
₹ 49
SAVE
Our Price
₹ 75