

P3 : Concevez une application au service de la santé publique

Brands Comparator

26/04/2022

DUBART Maxime

Projet - Brands Comparator

Dans le cadre de l'appel à projet de l'agence Santé publique France

Proposer un service de comparaison des marques
pour des produits de même catégories

Données




Données Open Food Facts (<https://world.openfoodfacts.org/>)

- Base de données collaborative – produits alimentaires vendus à travers le monde
 - Données générales (e.g., nom, date de création, créateurs, etc.)
 - Ensemble de tags (e.g., lieu de fabrication, localisation, **catégories**, **entreprises**, packaging, etc.)
 - Ingrédients, additifs éventuels et allergènes
 - **Informations nutritionnelles ‘normalisées’** (i.e. pour 100g de produit)
- 320772 l. x 162 c. – 106 colonnes numériques, 56 colonnes (objets – string ou date)

Données

Vérification et nettoyage des données

- Nombreuses valeurs manquantes (127/162 colonnes avec remplissage < 50%)
 - Filtre colonnes avec remplissage inférieur à 50 % (excepté catégories principales - 26 % filled)
 - Filtre sur colonnes d'intérêt pour le projet (val. nutri., catégories, marques)
 - Filtre des lignes dupliquées
- Contraintes sur les valeurs nutritionnelles (*_100g colonnes)
 - Valeurs en grammes doivent être $\in [0 ; 100]$
 - Energie (kJ) doit être $\in [0 ; 3766]$:
 - Fondé sur l'énergie max. contenue dans 100g de produit i.e., un produit uniquement composé de graisses (100g de graisses = 900 kcal = 3766 kJ).
 - Nutrition-Score doit être $\in [-15 ; 40]$: par définition
 - **Valeurs en dehors de ces gammes  remplacées par valeurs manquantes**
- Filtre des lignes avec > 50% de valeurs manquantes pour les valeurs nutritionnelles

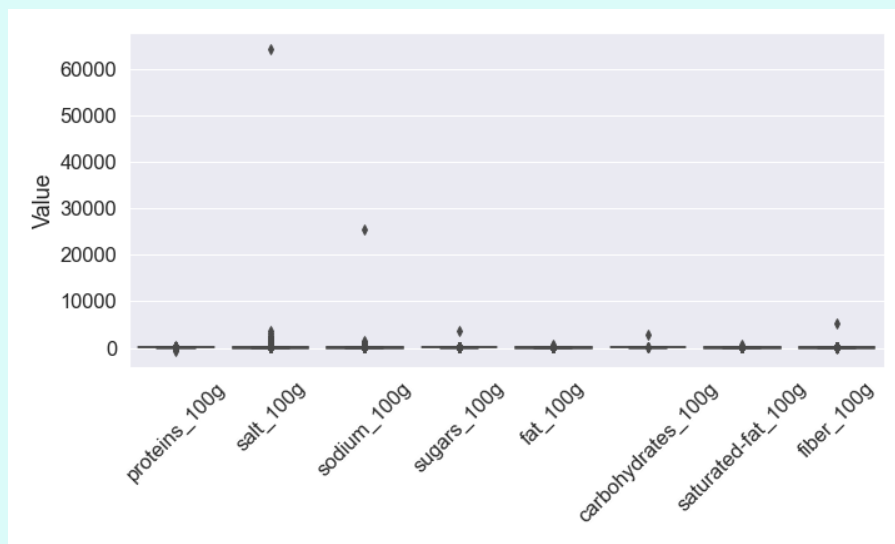
Données

Vérification et nettoyage des données

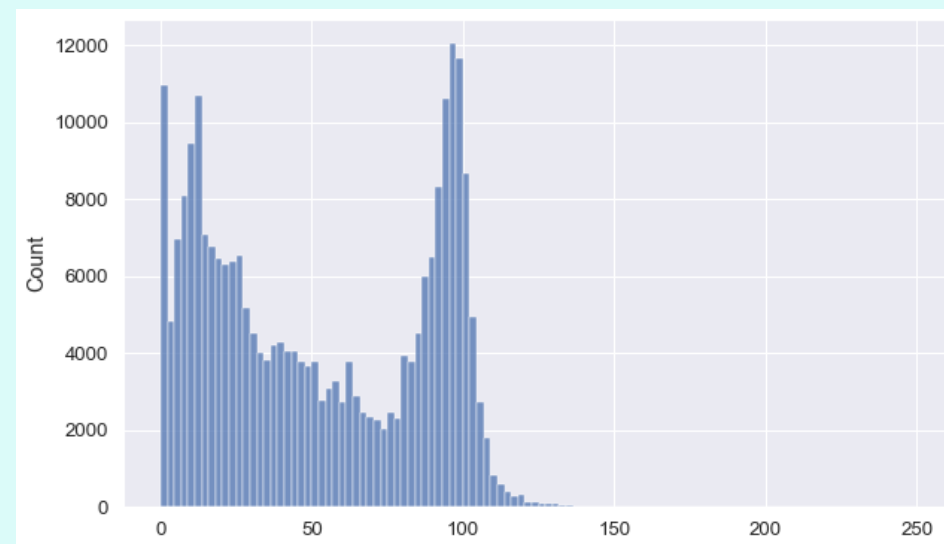
- Contraintes additionnelles sur les valeurs nutritionnelles (*_100g colonnes)
 - La somme des valeurs (g), par définition, ne peut pas dépasser 100g
 - Sucres ∈ Glucides, ainsi, $[\text{sugars}] \leq [\text{carbohydrates}]$
 - Graisses saturées ∈ Graisses, ainsi, $[\text{saturated_fat}] \leq [\text{fat}]$
 - Sodium (Na) ∈ Sel (NaCl), ainsi, $[\text{sodium}] \leq [\text{salt}]$ (approx., $[\text{sodium}] = 0.39 \times [\text{salt}]$)
- **Suppression des lignes qui ne respectent pas ces contraintes**

Avant net.

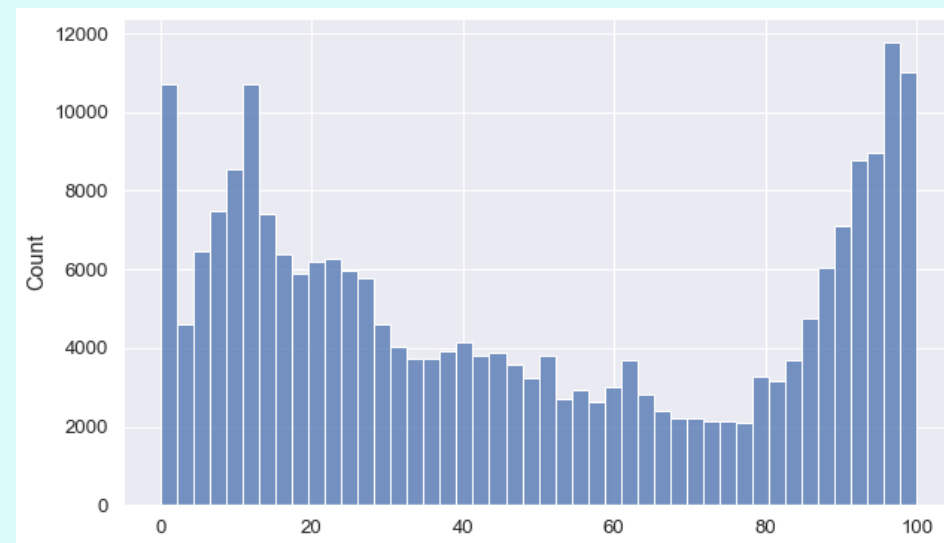
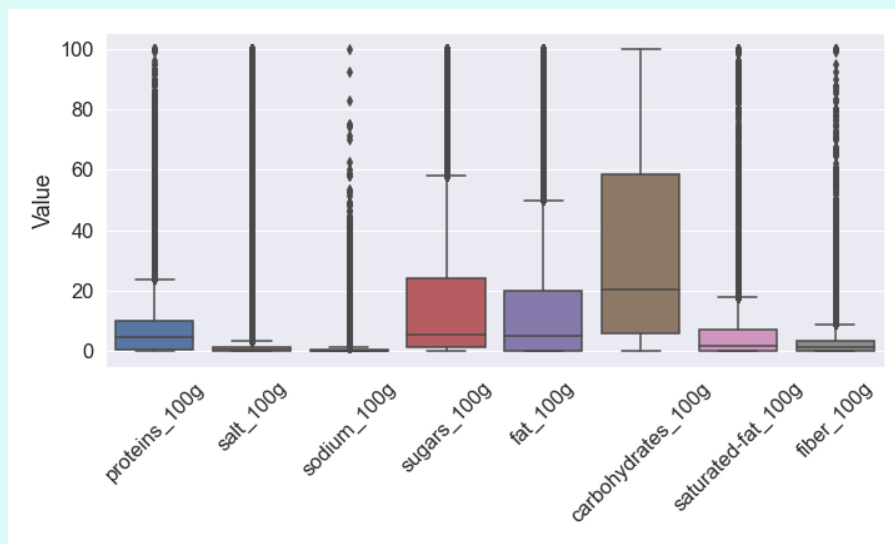
Valeurs nutritionnelles



Sommes par produits

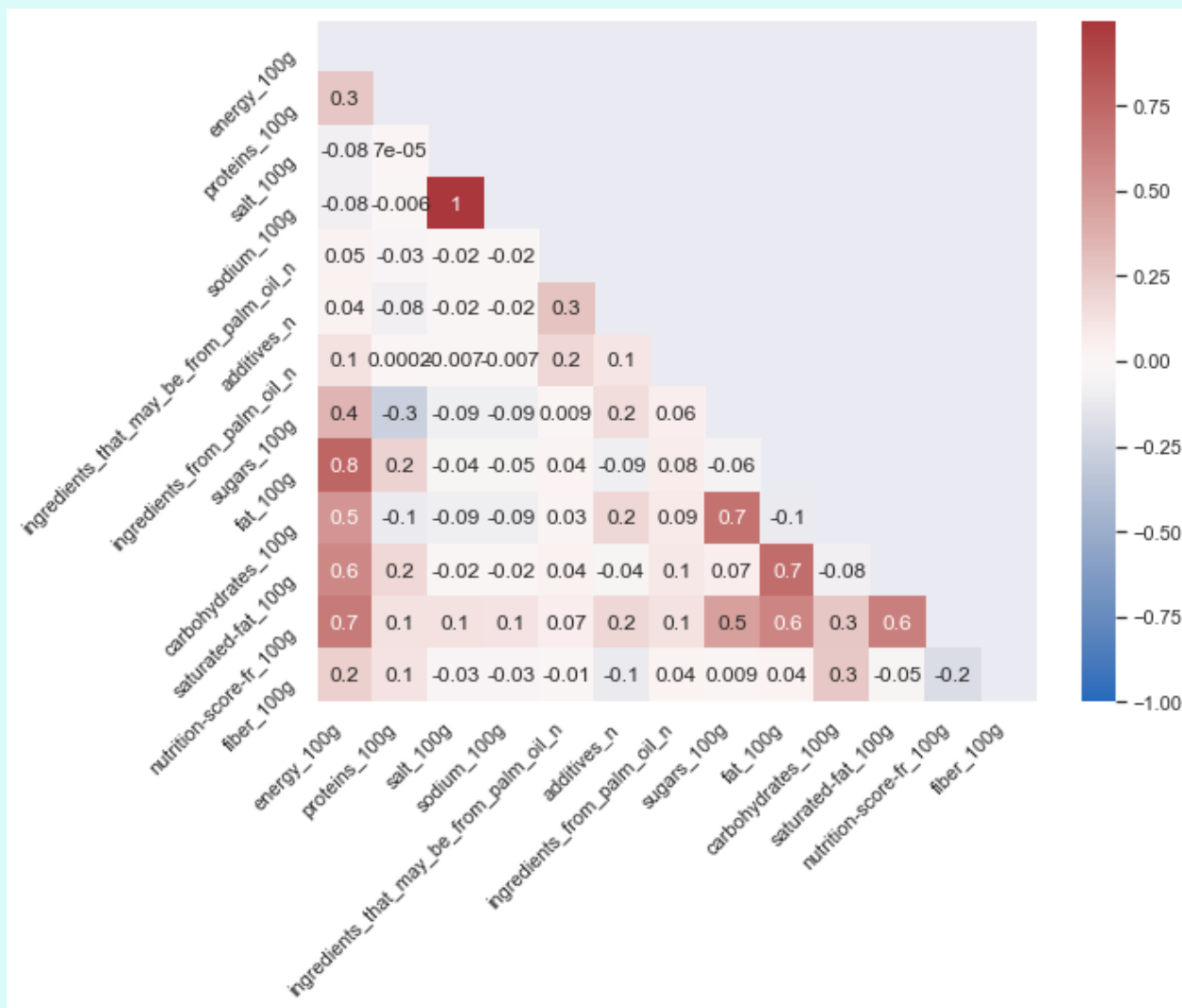


Après net.



Données

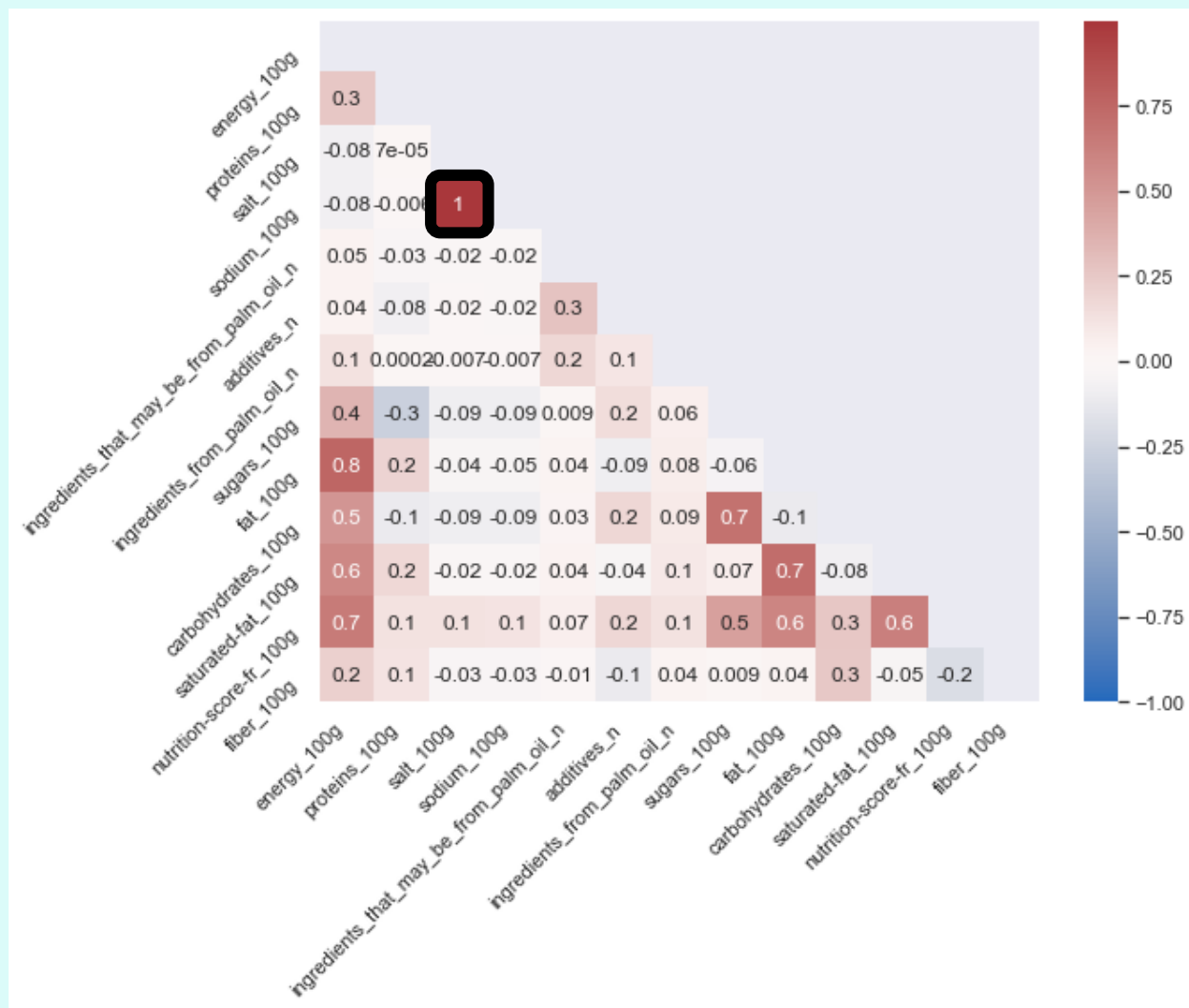
Traitement des valeurs manquantes



- Corrélation parfaite sodium / sel (attendu)
- Corrélations importantes
 - Sucres / Glucides
 - Graisses saturées / Graisses
- Produits avec nutri-score important
 - Produits gras et/ou sucrés, i.e. énergie ++
- Relativement peu corrélés / indépendants
 - Additifs / Produits issus d'huile de palme / Fibres / Proteins

Données

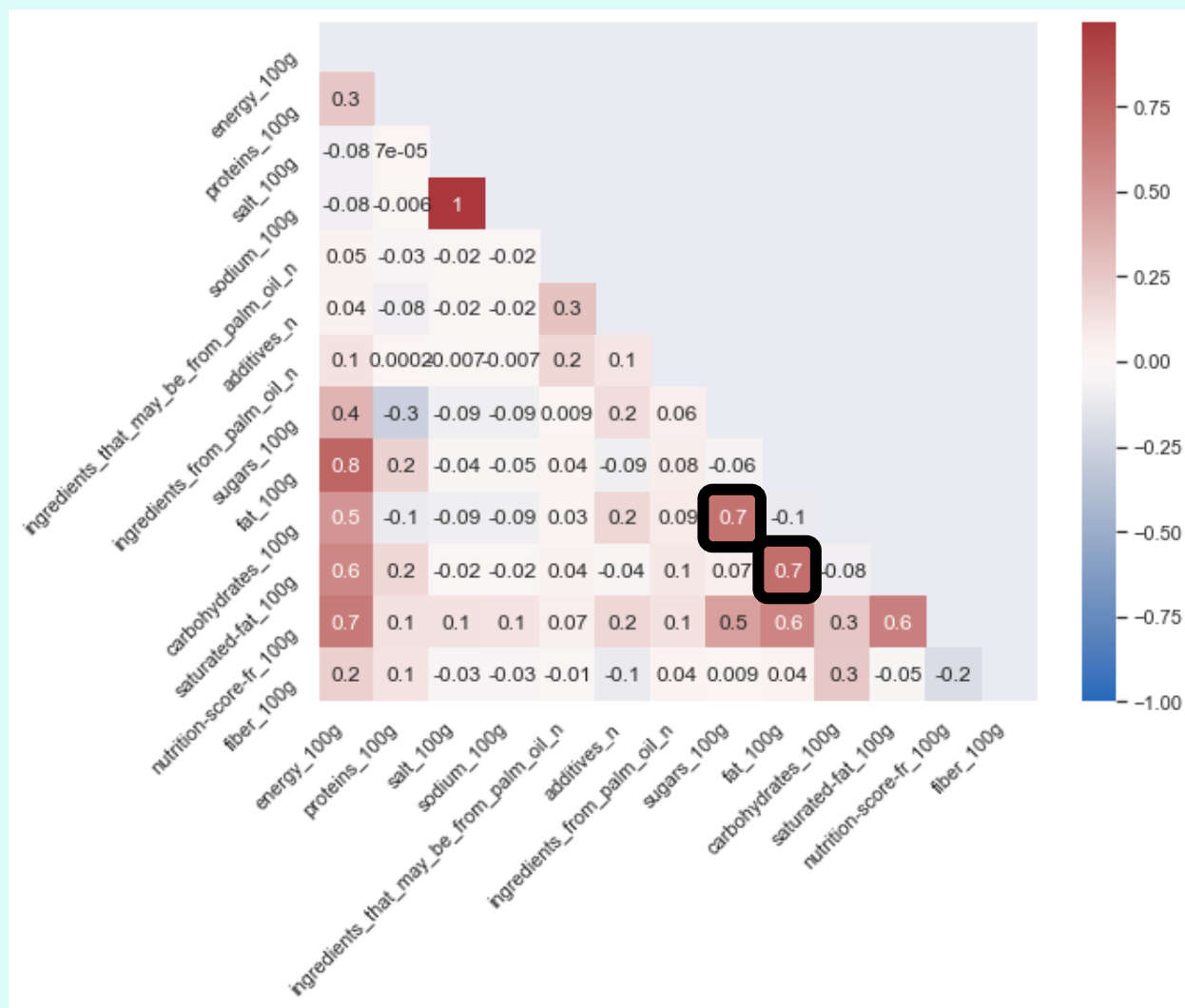
Traitement des valeurs manquantes



- **Corrélation parfaite sodium / sel (attendu)**
- Corrélations importantes
 - Sucres / Glucides
 - Graisses saturées / Graisses
- Produits avec nutri-score important
 - Produits gras et/ou sucrés, i.e. énergie ++
- Relativement peu corrélés / indépendants
 - Additifs / Produits issus d'huile de palme / Fibres / Proteins

Données

Traitement des valeurs manquantes



- Corrélation parfaite sodium / sel (attendu)
- **Corrélations importantes**
 - **Sucres / Glucides**
 - **Graisses saturées / Graisses**
- Produits avec nutri-score important
 - Produits gras et/ou sucrés, i.e. énergie ++
- Relativement peu corrélés / indépendants
 - Additifs / Produits issus d'huile de palme / Fibres / Proteins

Données

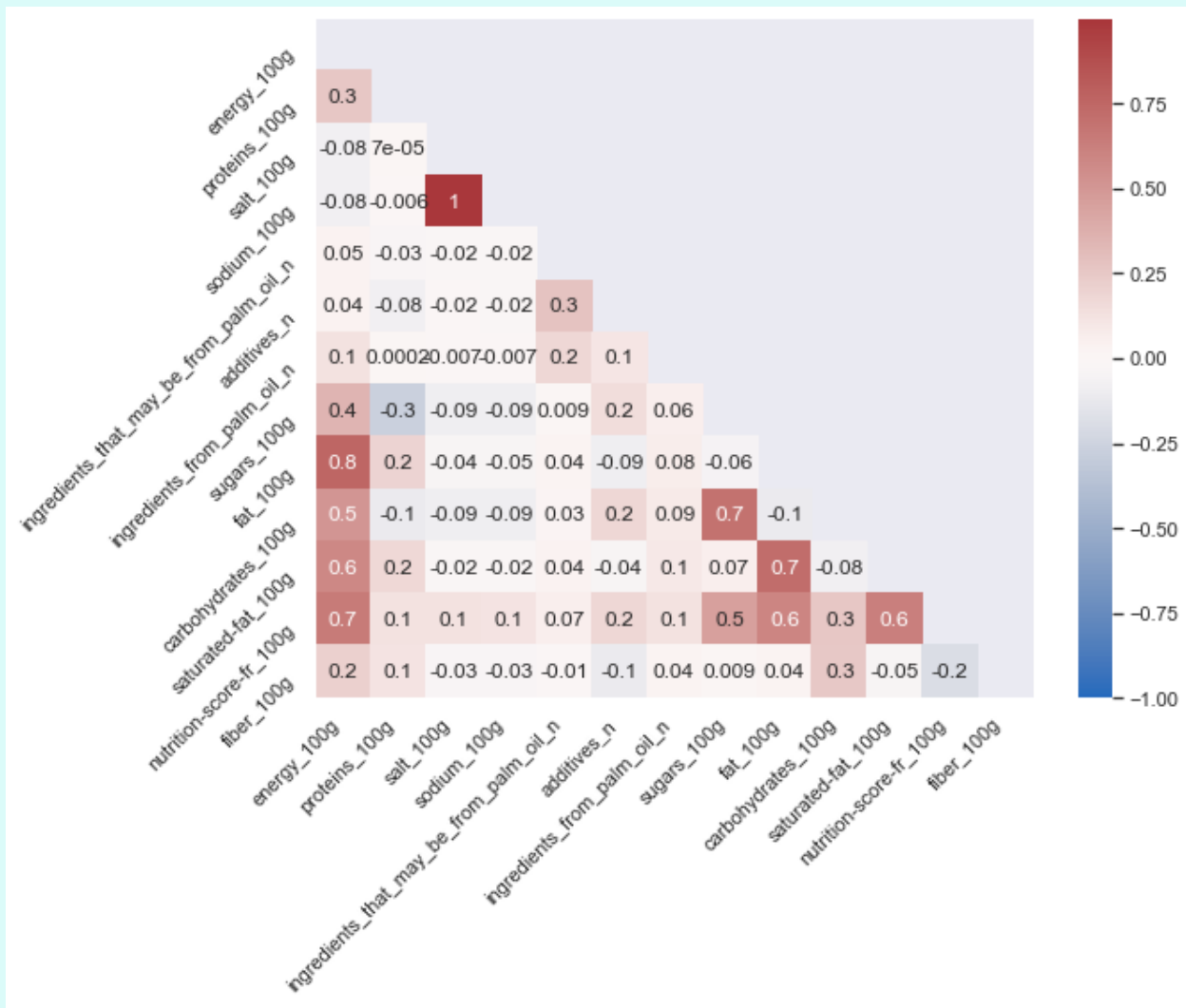
Traitement des valeurs manquantes



- Corrélation parfaite sodium / sel (attendu)
- Corrélations importantes
 - Sucres / Glucides
 - Graisses saturées / Graisses
- **Produits avec nutri-score important**
 - **Produits gras et/ou sucrés, i.e. énergie ++**
- Relativement peu corrélés / indépendants
 - Additifs / Produits issus d'huile de palme / Fibres / Proteins

Données

Traitement des valeurs manquantes



- Corrélation parfaite sodium / sel (attendu)
- Corrélations importantes
 - Sucres / Glucides
 - Graisses saturées / Graisses
- Produits avec nutri-score important
 - Produits gras et/ou sucrés, i.e. énergie ++
- Relativement peu corrélés / indépendants
 - Additifs / Produits issus d'huile de palme /
Fibres / Proteins

Données

Traitement des valeurs manquantes

- Fibres : valeurs manquantes sont vraisemblablement de vrais zéros
 - **Simple Imputer**
- Sodium/Salt : très corrélés
 - **Iterative Imputer**
- Autres variables numériques
 - **Knn Imputer**
 - Quelques tests insatisfaisants avec IterativeImputer sur les features corrélées



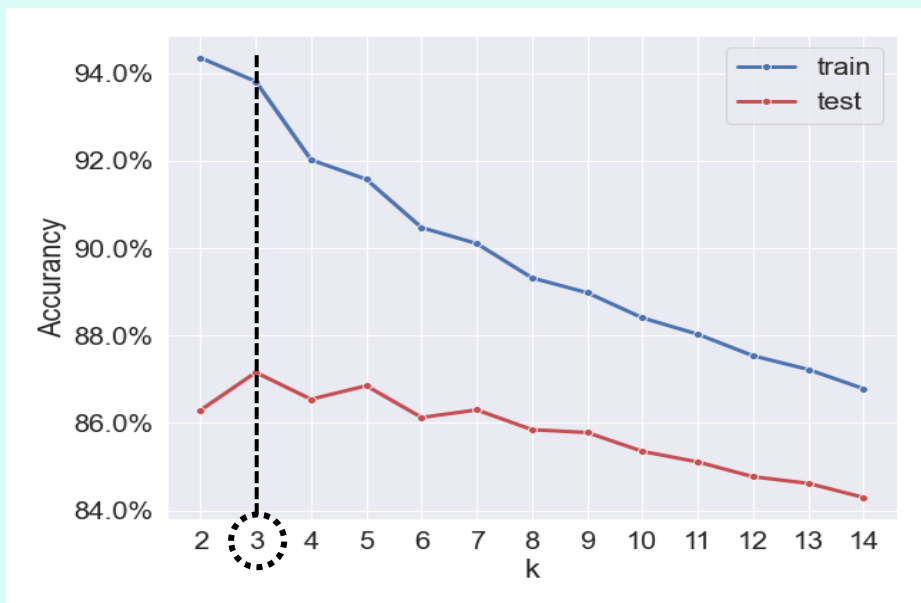
Répétition des étapes de vérification/nettoyage appliquées précédemment

Données

Traitement des valeurs manquantes

- Cas particulier : **Nutrition Grades**
 - Valeurs manquantes $\approx 12\%$
 - Pourrait être intéressant de compléter ces valeurs
- Utilisation d'un KNN-classifier (optimisé pour k , et split train/test 0.7) sur les valeurs numériques

Précision $\sim k$

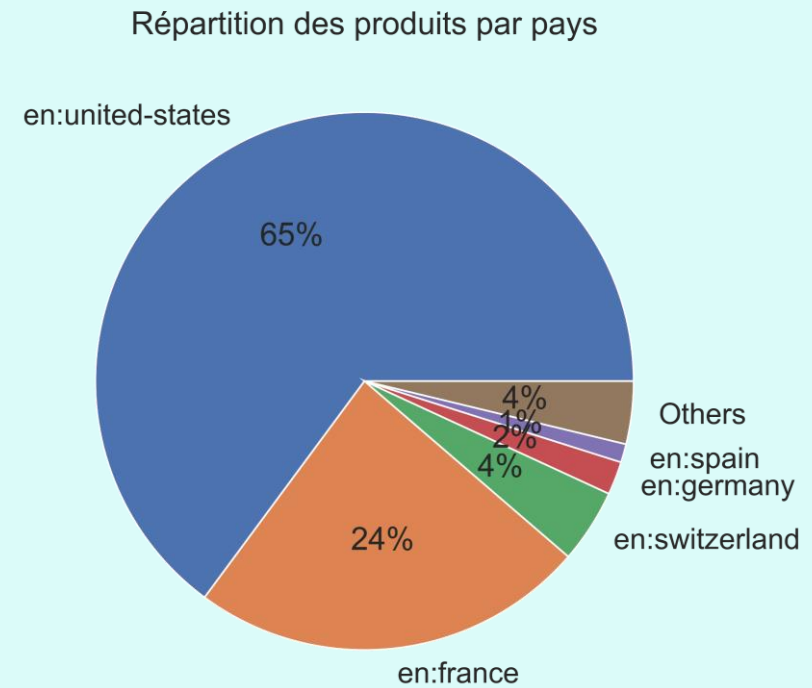
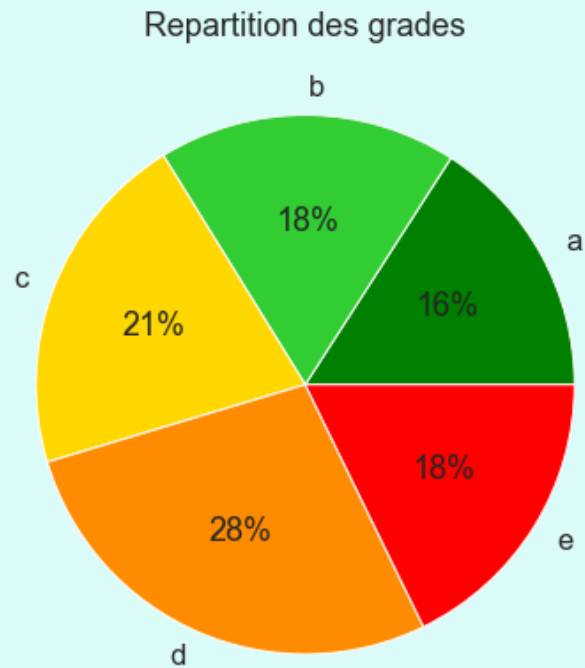


Matrice de confusion



Exploration

Répartition des valeurs du tableau de données



Exploration

Grades et valeurs nutritionnelles

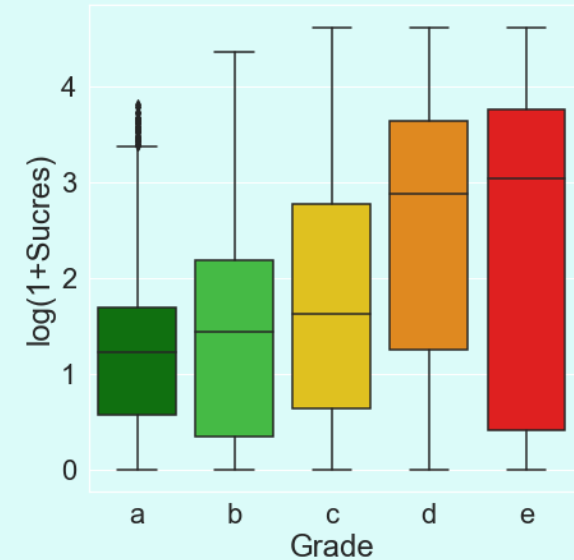
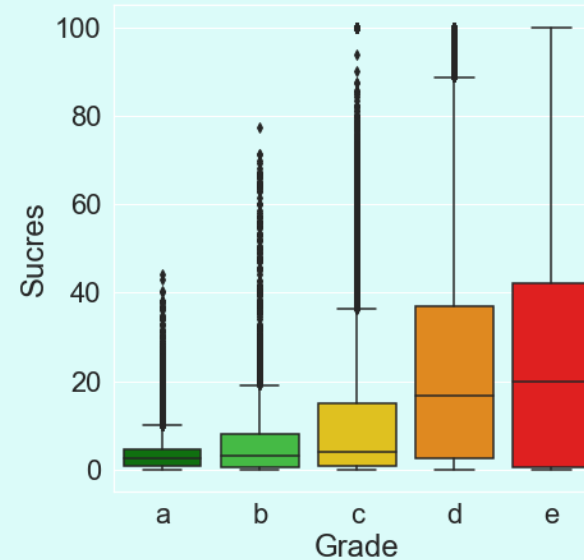
Les sucres

Homoscédasticité : non (idem pour $\log(1+x)$)

Test non paramétrique de Kruskal

$H=26558.24$, $df=4$, **p-val ≈ 0**

Tests posthoc (Conover-Iman) – toutes les diff. sont sign.



Les graisses

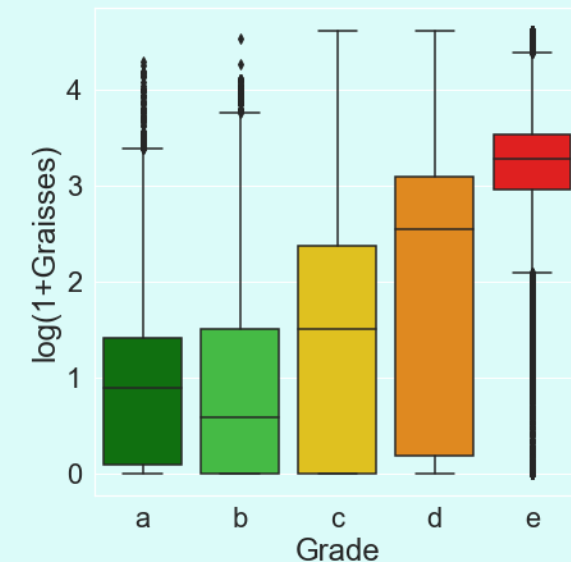
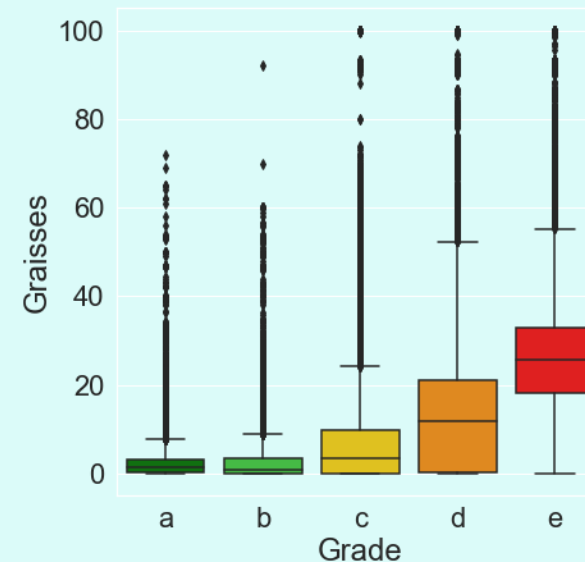
Homoscédasticité : non (idem pour $\log(1+x)$)

Test non paramétrique de Kruskal

$H=71287.46$, $df=4$, **p-val ≈ 0**

Tests posthoc :

un Tukey & Conover donne une diff. A-C n.s.



Exploration

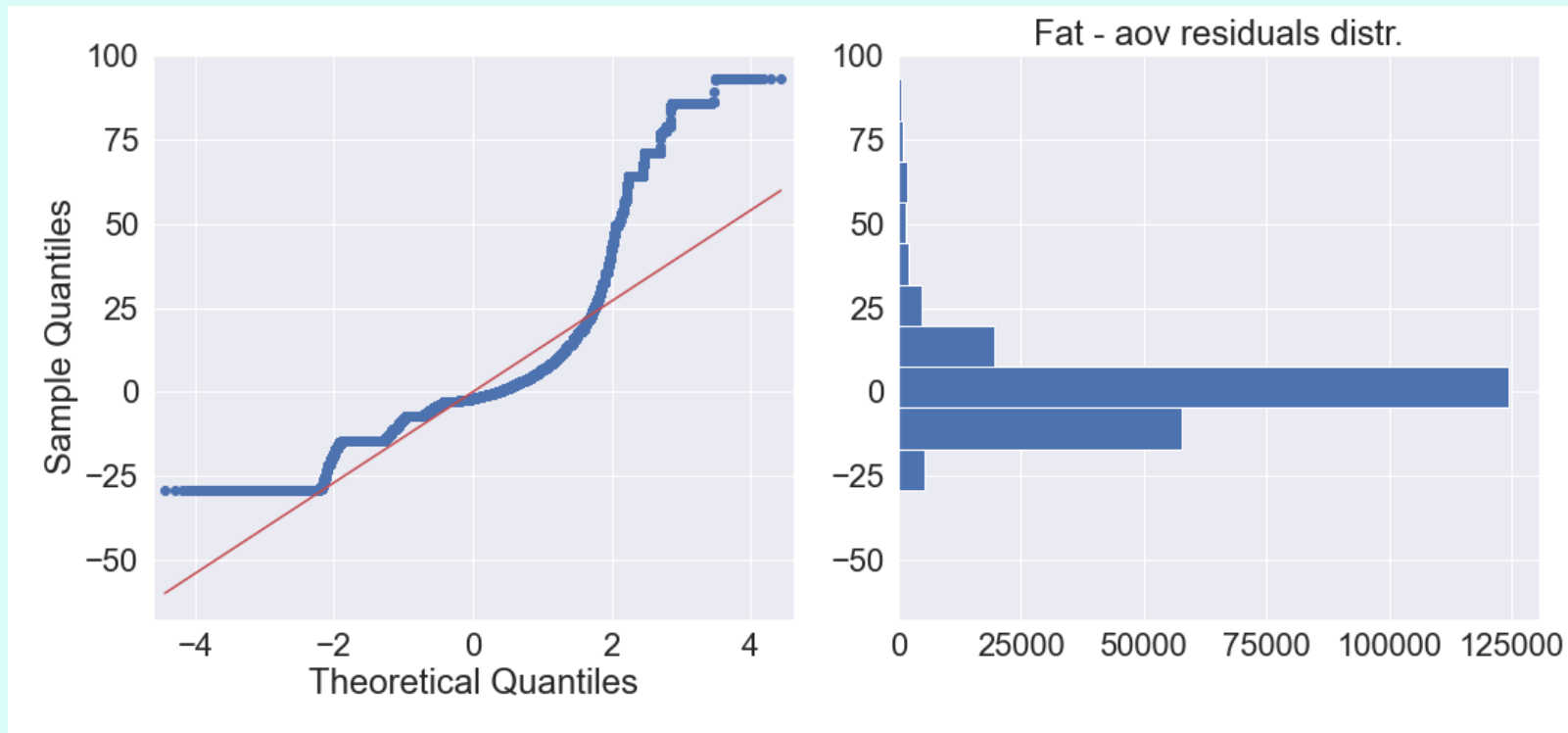
Grades et valeurs nutritionnelles

Résultats similaires pour les autres variables.

Note: pour les comptages, plutôt GLM (non explorés ici)

Hypothèses de normalité et homoscédasticité non respectées !

Risque erreurs type I



Exploration

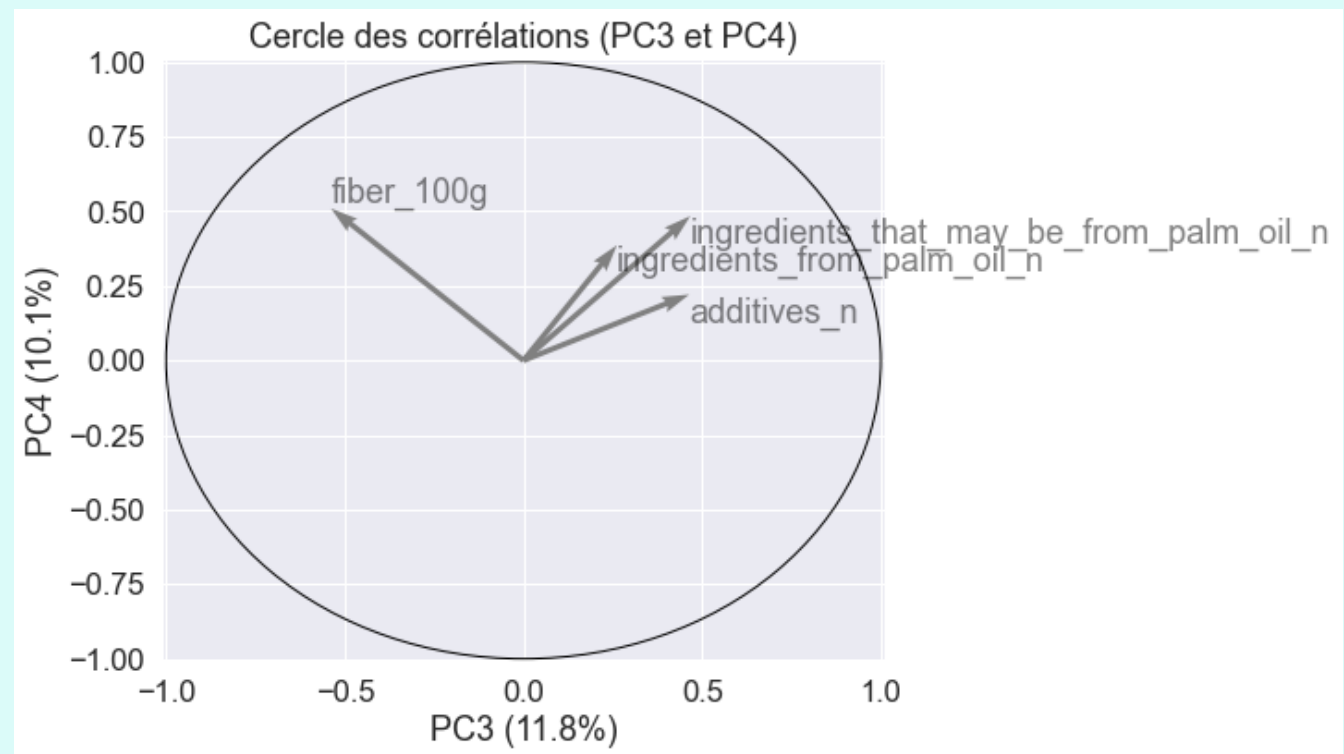
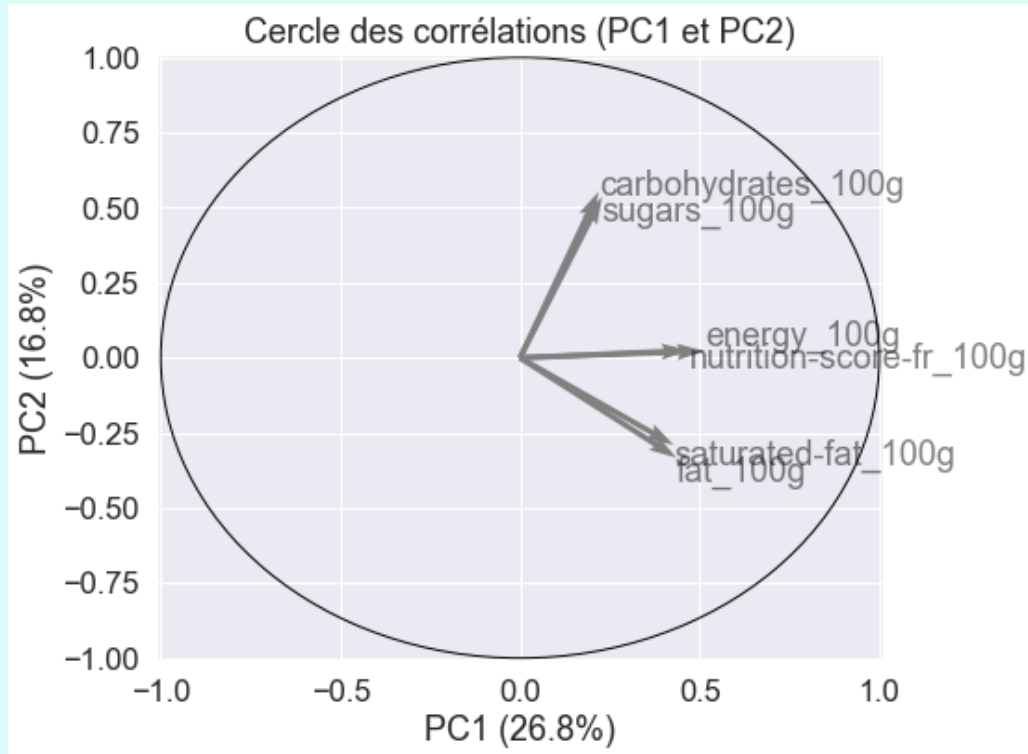
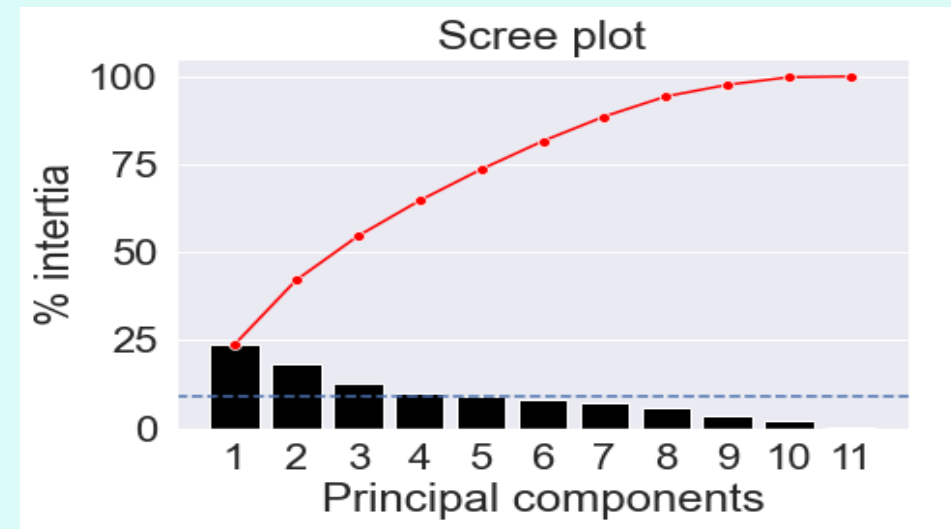
Analyse multivariée (ACP)

PC1 (27%) : Graisses et energie/score

PC2 (17%) : Glucides

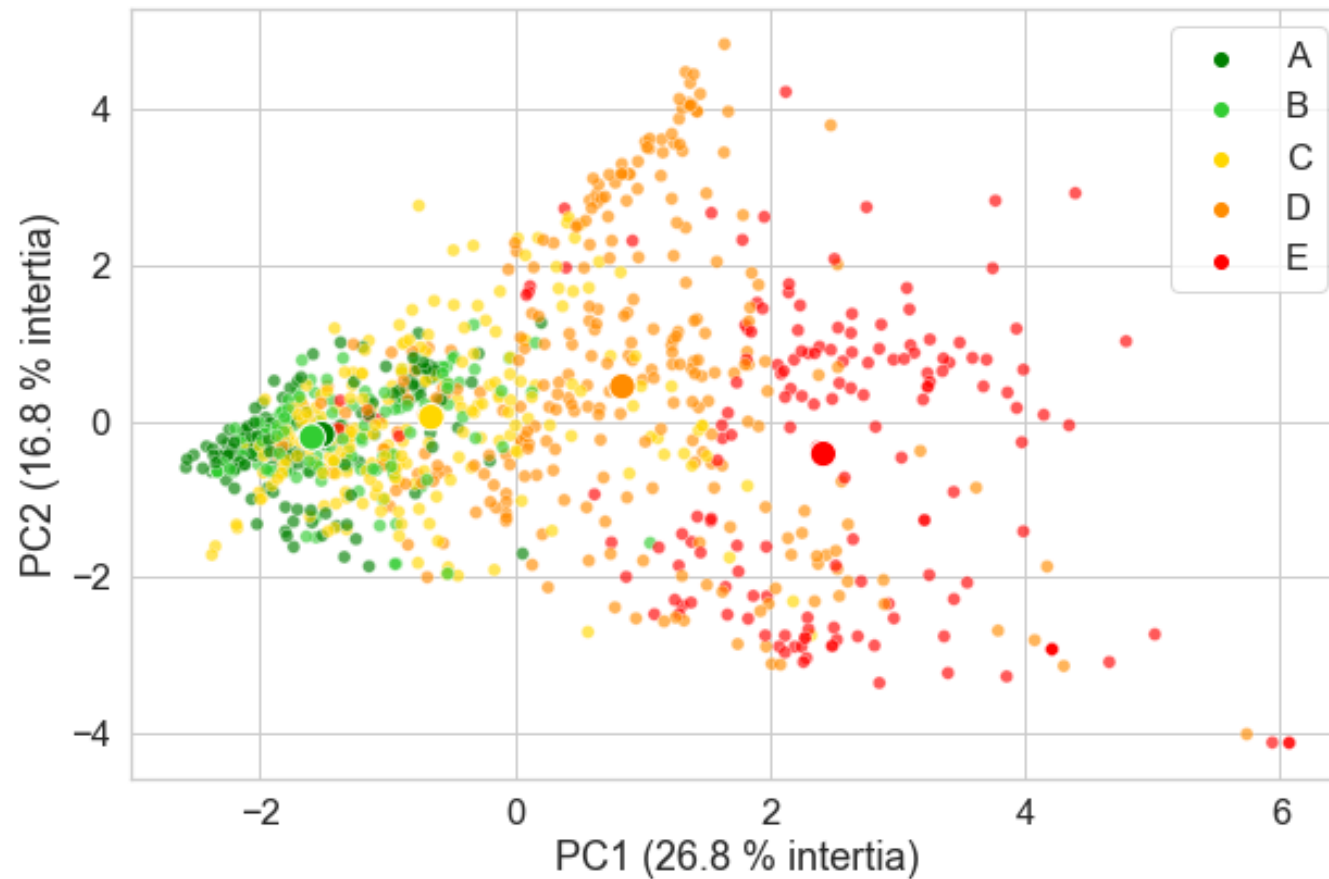
PC3/4 : additives, huile palme, fibre

A noter : fibre non corrélé aux autres



Exploration

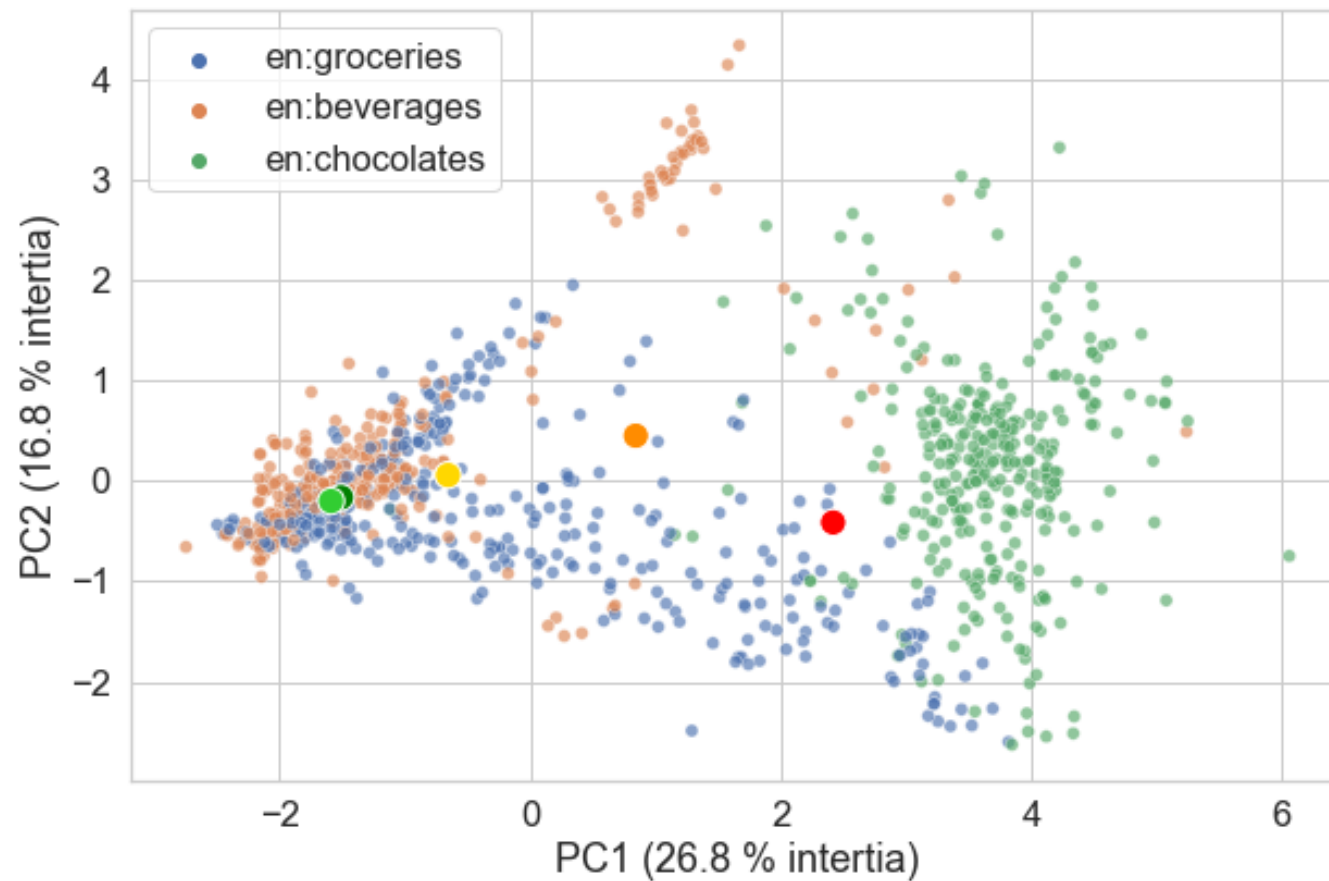
Analyse multivariée (ACP) – Répartition des produits



Sous-échantillon

Exploration

Analyse multivariée (ACP) – Placements catégories les plus représentées



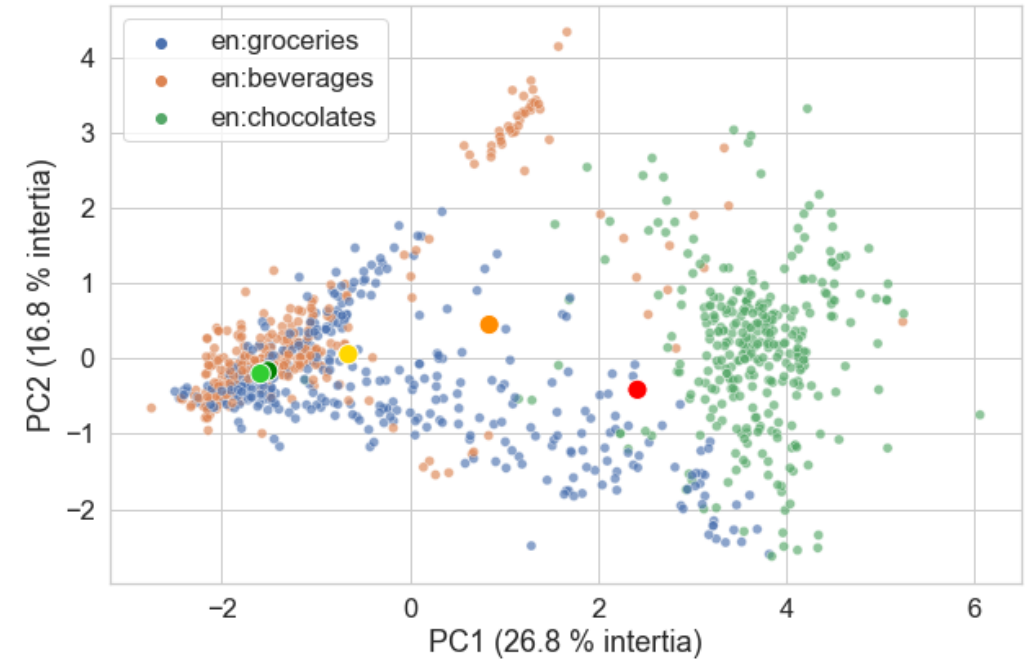
Sous-échantillon

Exploration

Proposition d'application

Placement des marques dans cette espace

-> Les grades A/B mal séparés



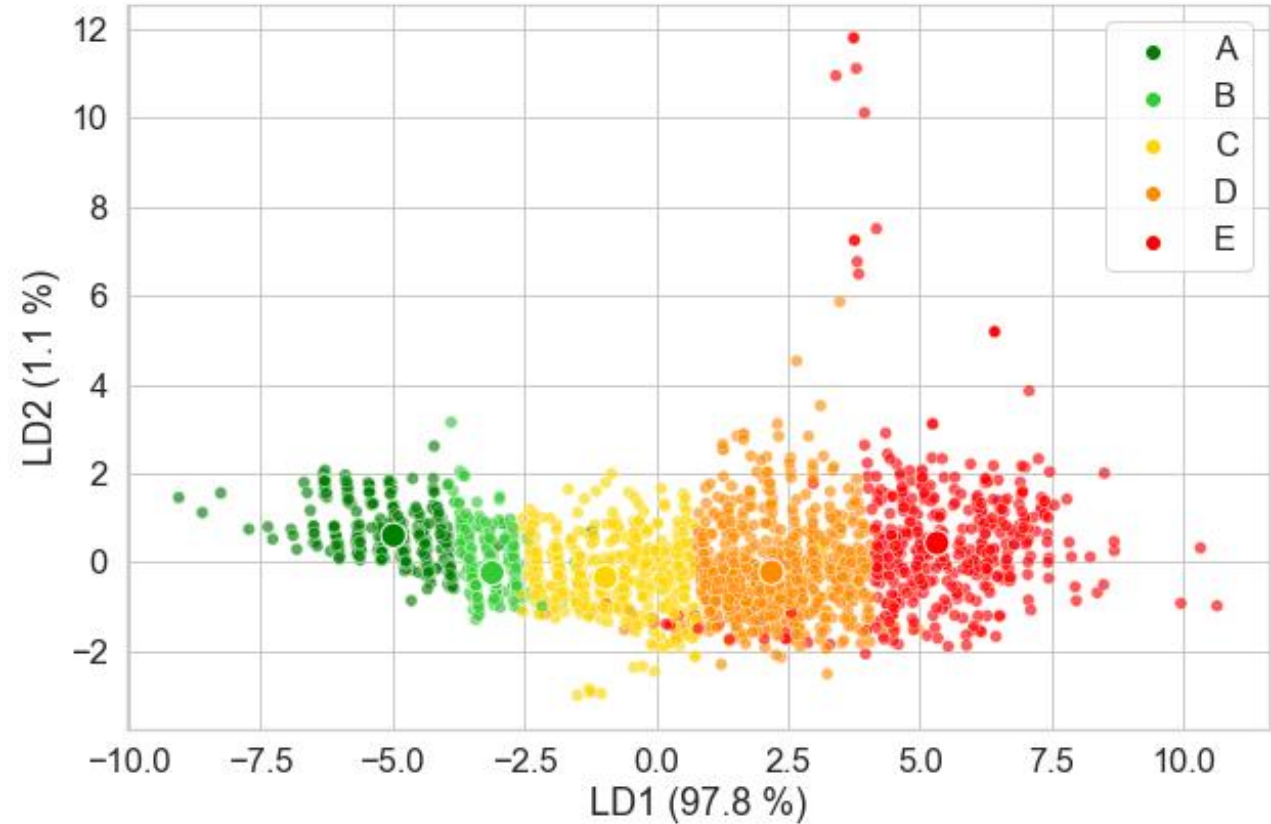
Approche de réduction de dimension via analyse discriminante

Exploration

Analyse multivariée (LDA)

LDA (data $\log(1+x)$ – scaled)

Principal séparation par l'axe #1



Exploration

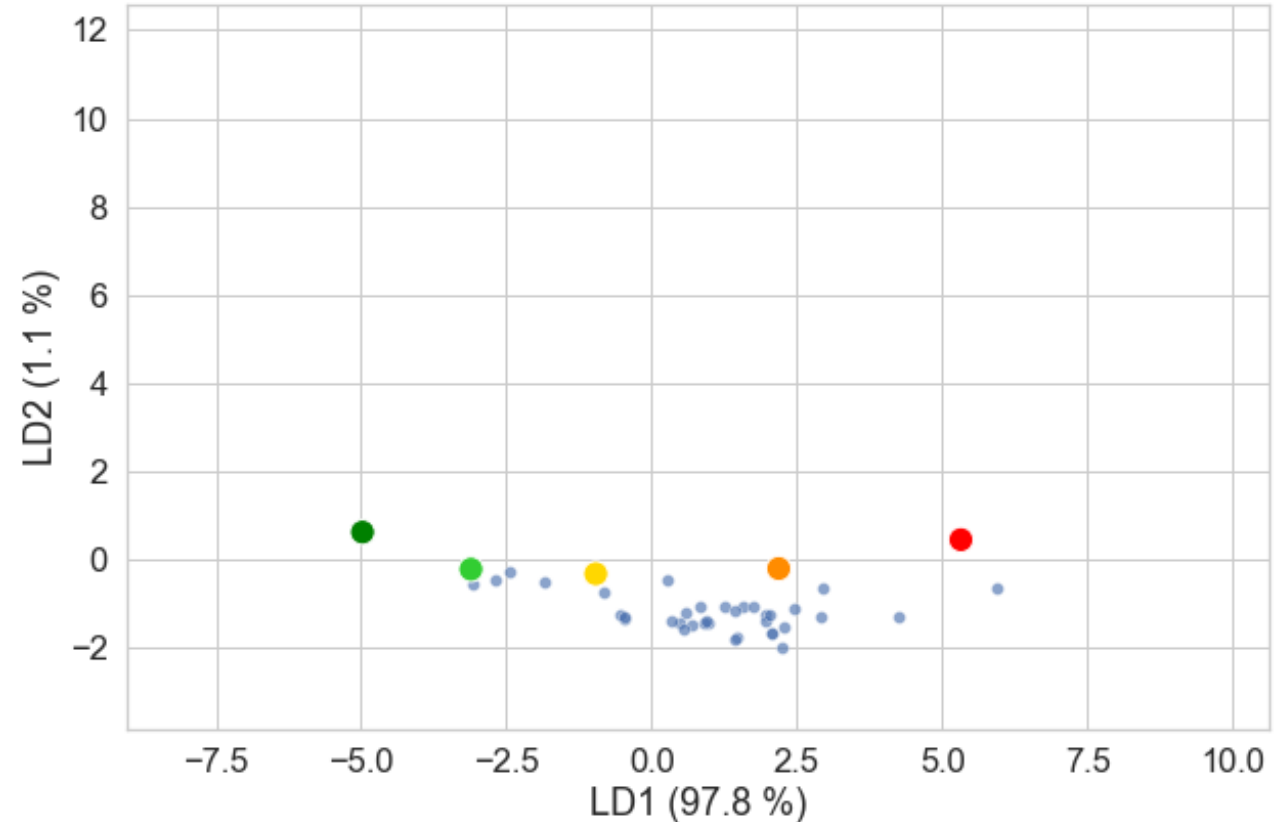
Analyse multivariée (LDA)

Pour une catégorie : e.g., boissons

Représentation des marques (avec au moins 10 produits) dans cet espace

Sélection des marques proches

Marques de boissons



Exploration

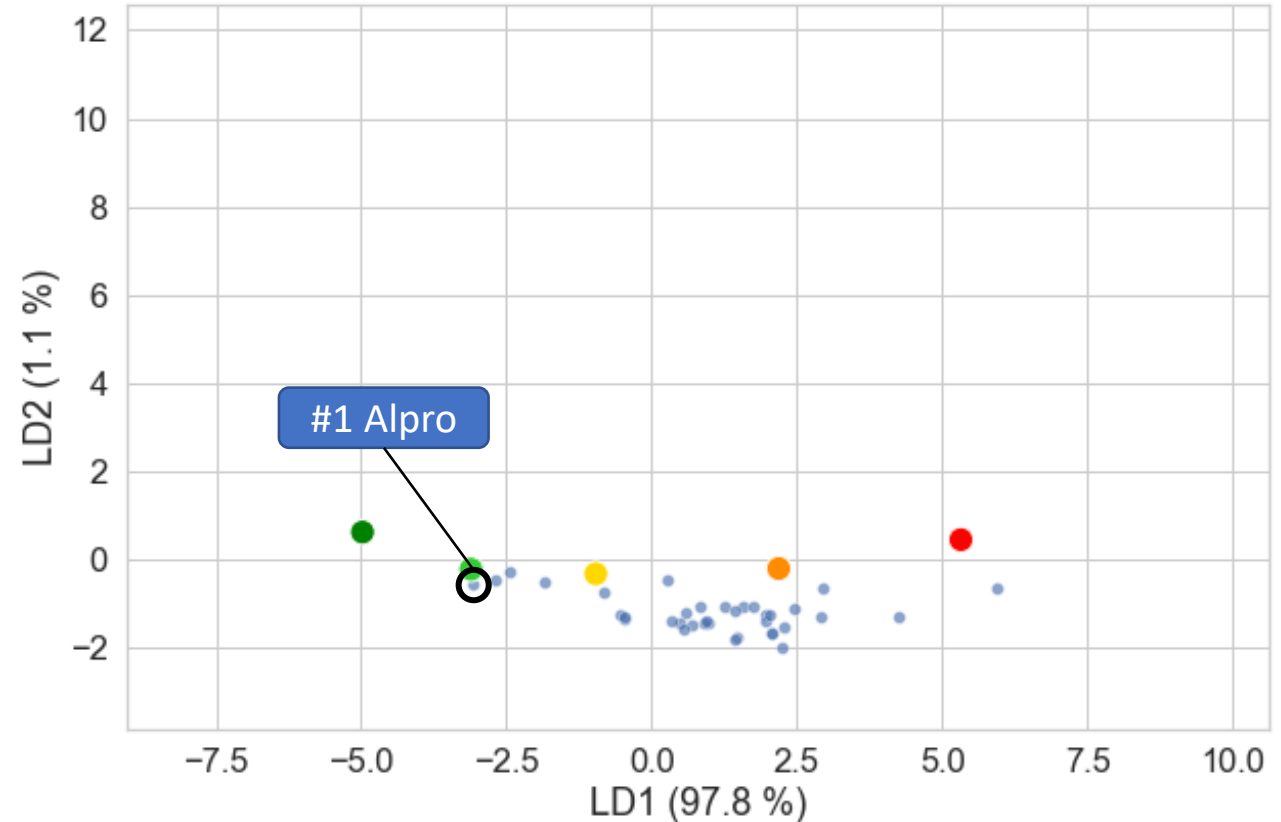
Analyse multivariée (LDA)

Pour une catégorie : e.g., boissons

Représentation des marques (avec au moins 10 produits) dans cet espace

Sélection des marques proches

Marques de boissons



Exploration

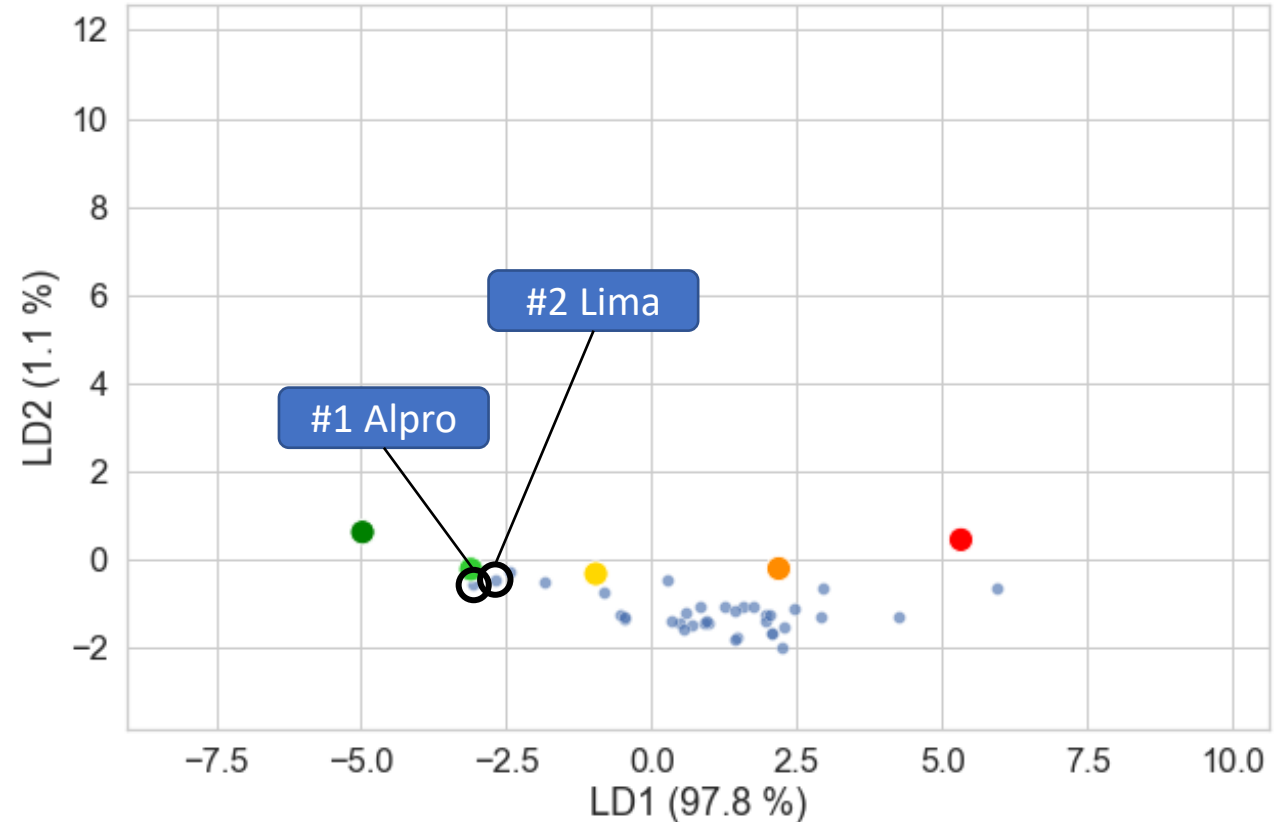
Analyse multivariée (LDA)

Pour une catégorie : e.g., boissons

Représentation des marques (avec au moins 10 produits) dans cet espace

Sélection des marques proches

Marques de boissons



Exploration

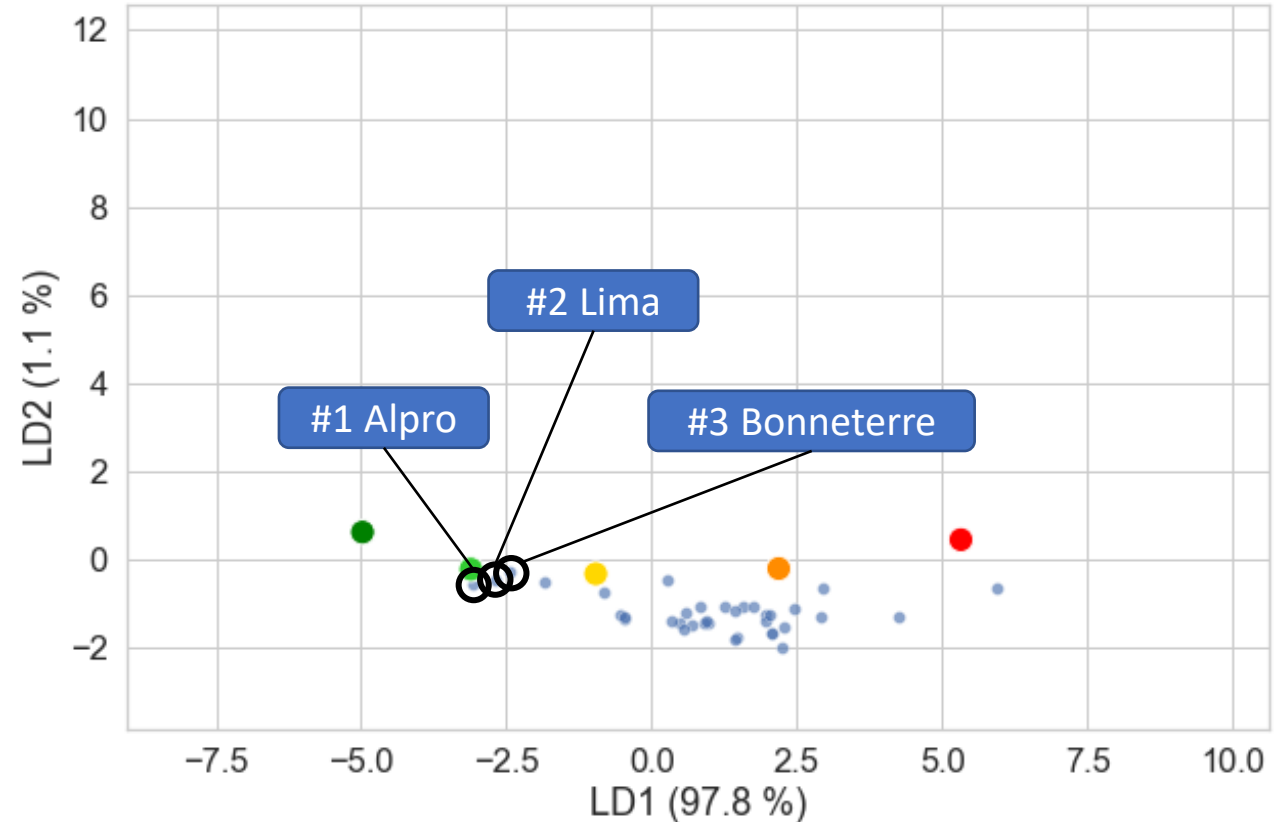
Analyse multivariée (LDA)

Pour une catégorie : e.g., boissons

Représentation des marques (avec au moins 10 produits) dans cet espace

Sélection des marques proches

Marques de boissons



Exploration

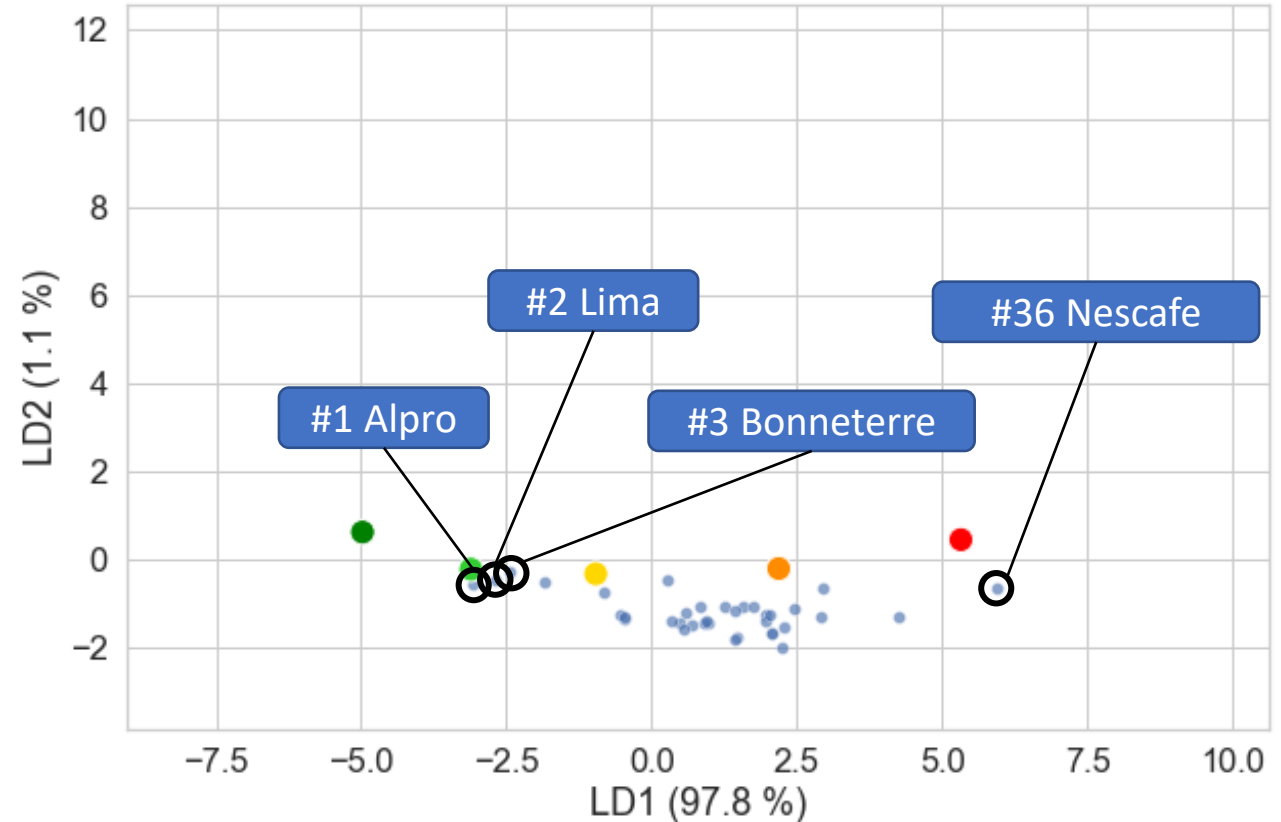
Analyse multivariée (LDA)

Pour une catégorie : e.g., boissons

Représentation des marques (avec au moins 10 produits) dans cet espace

Sélection des marques proches

Marques de boissons



Faisabilité de l'application

Possible mais :

- Il faudrait davantage de données pour la France
- Un taux de remplissage des catégories plus importants
- Possible classification hiérarchique des catégories
(e.g., distinguer les thés des sodas)

Merci