
Amélioration de la base de données OpenFoodFacts
Analyse préliminaire
Santé publique France



Le projet d'amélioration de la DB (Santé Publique France)

Mise en place d'un système d'auto-complétion des champs:

- limiter les erreurs de saisie
- limiter les valeurs manquantes
- homogénéité des données

Analyse exploratoire des données

1) Présentation générale des données

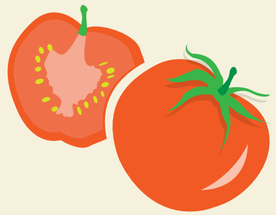
2) Nettoyage

3) Analyse

4) Résultats et Conclusion

5) RGPD















Présentation générale des données







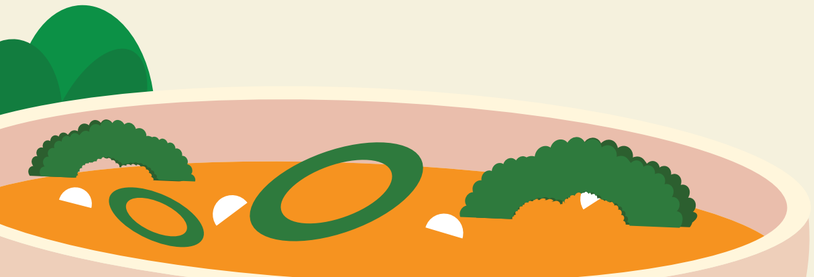
Présentation données

+ 320k entrées

162 variables

4 types de variables :

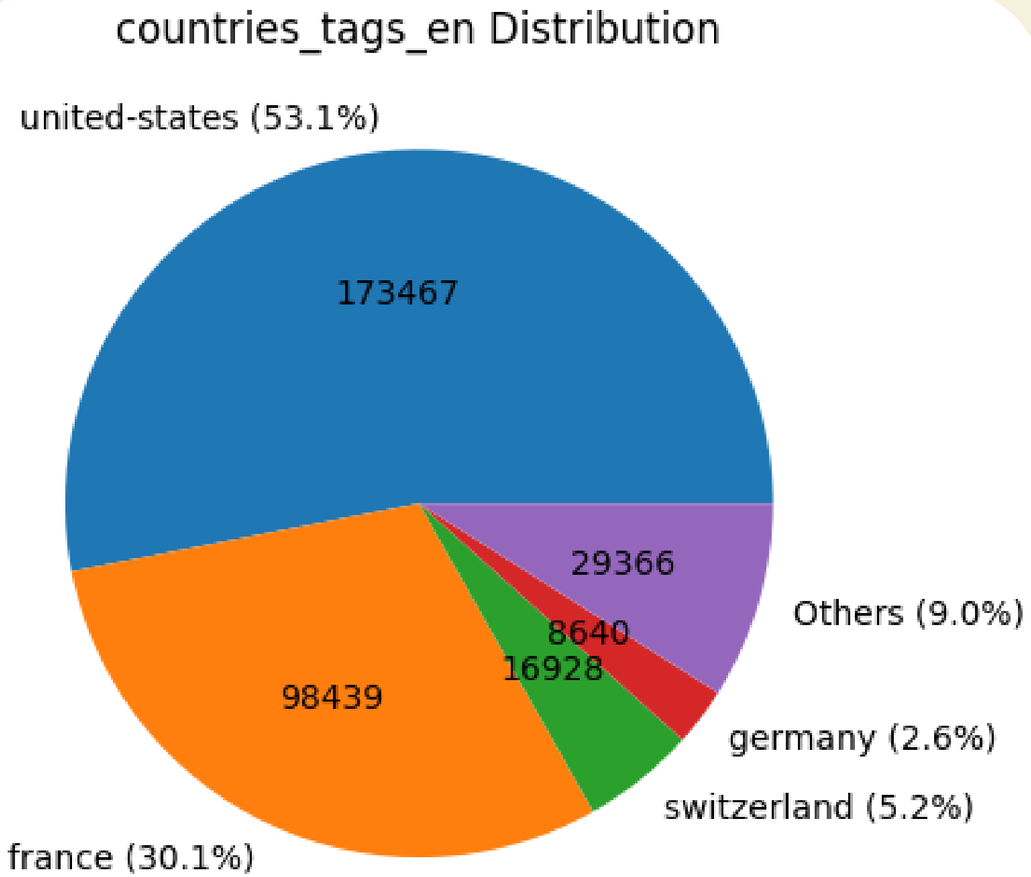
- Informations générales
- Tags
- Formulation (ingrédients)
- Nutrition



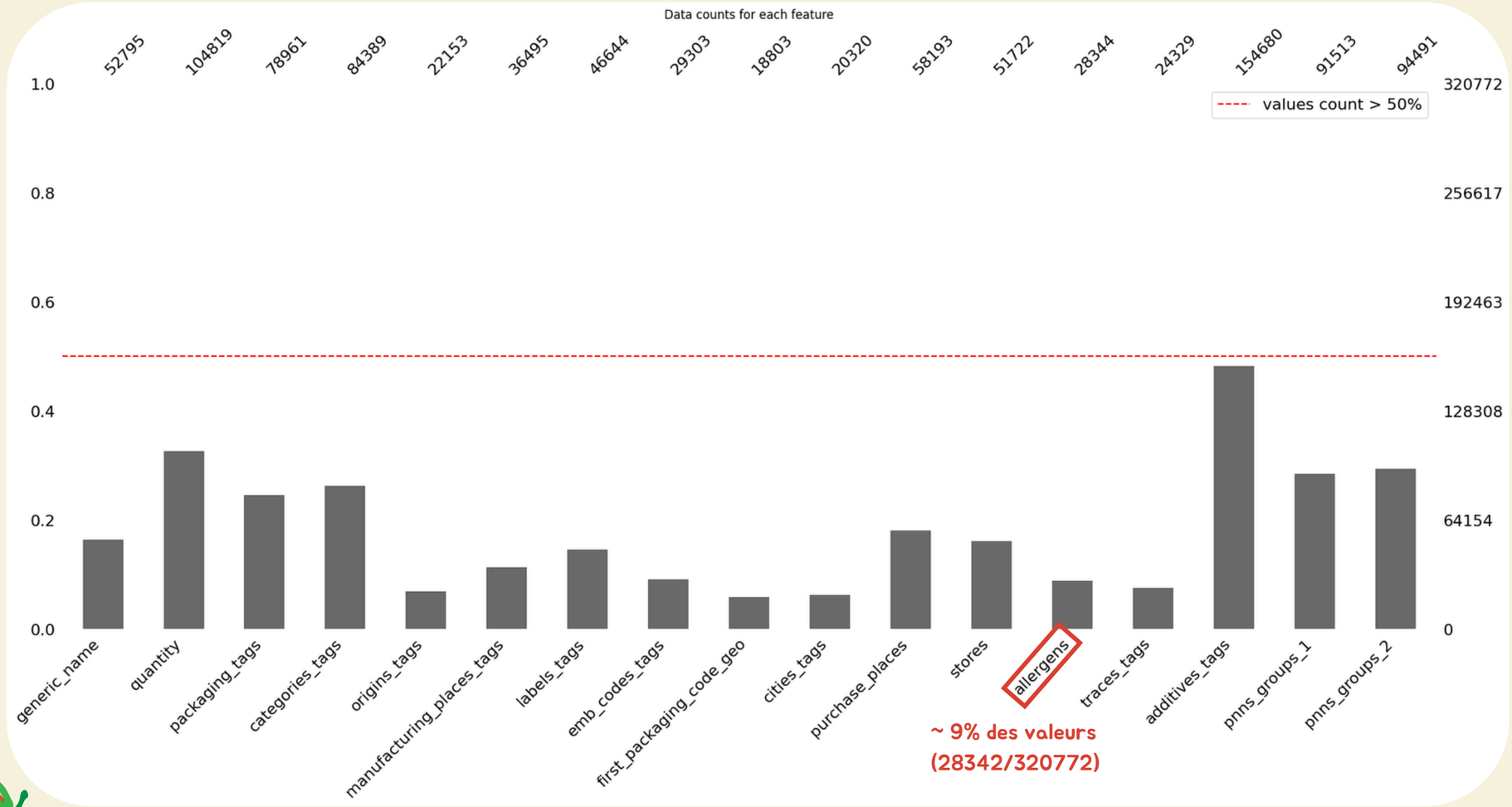


Présentation données

Produits par pays :

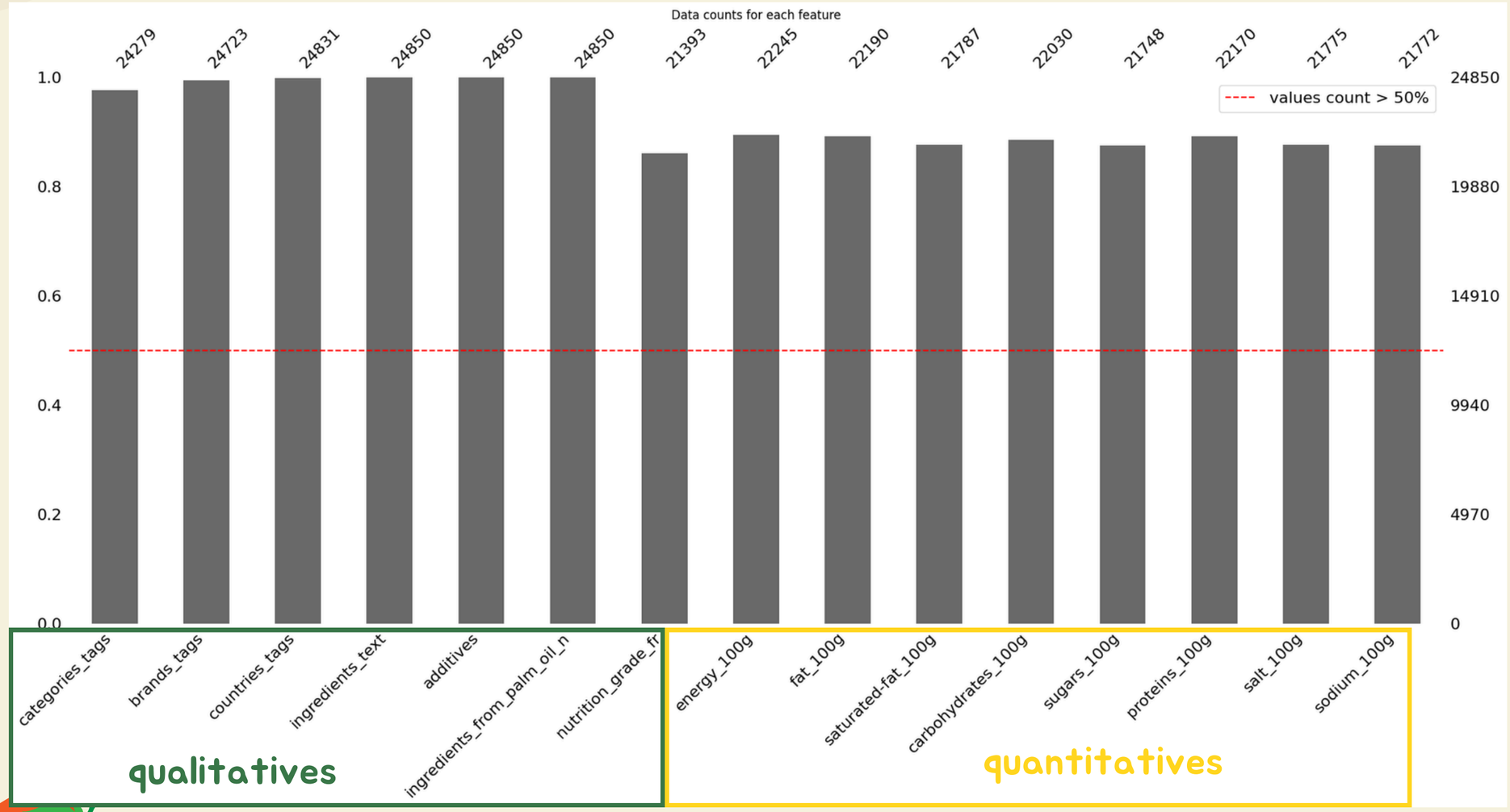


Présentation données



+ de 50% valeurs manquantes

Présentation données



+ de 50% de valeurs dans df_allergens







Nettoyage des donnees







Nettoyage des données

Quantitatives

- Outliers
- Missing Values

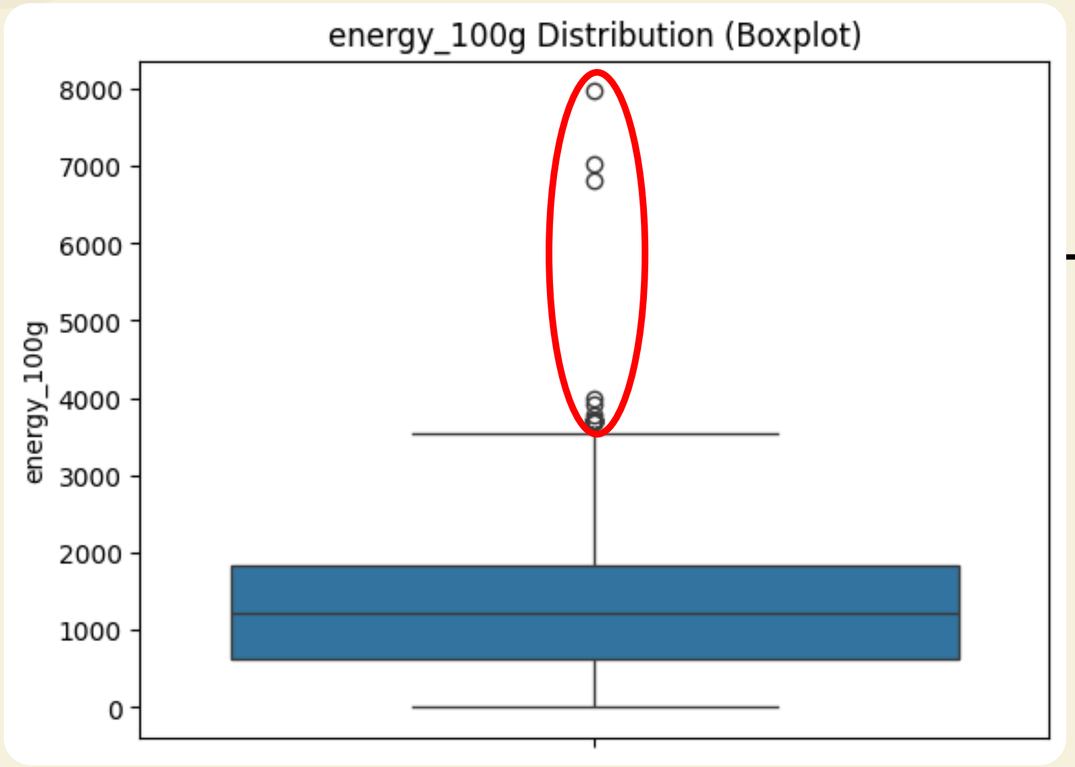
Qualitatives

- Homogénéisation
- Missing Values
- Variables quantitatives dérivées
- Regroupement des modalités



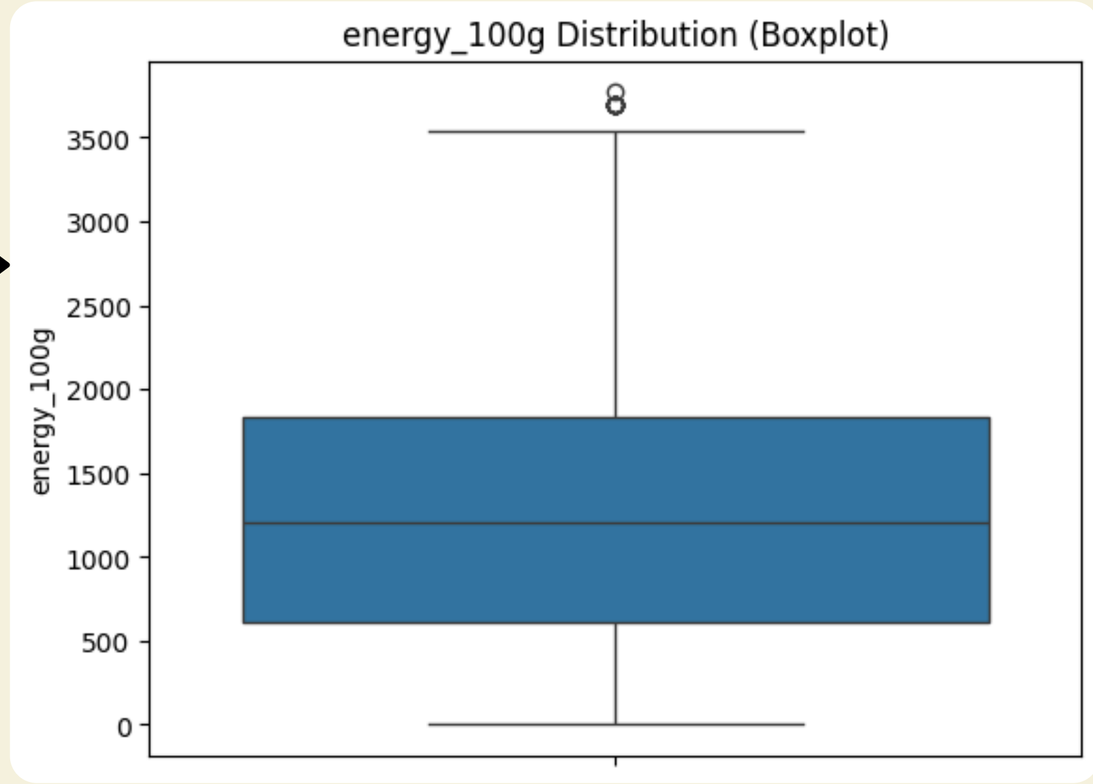


Nettoyage des données quantitatives



energie > 3800 kJ
-> NaN

estimation des NaN



Nettoyage des données qualitatives

Liste des ingrédients

gestion casse

Chocolat au lait 40% (sucre, beurre de cacao, poudre de _lait_ entier, pâte de cacao, émulsifiant : lécithines (_soja_), arôme), nappage au caramel 21% (sirop de glucose-fructose, stabilisant : glycérol, _lait_ concentré sucré, caramel 4%* (sucre, eau), _beurre_, arôme, sel, gélifiant : pectines), farine de _blé_, sucre, fructose, _lait_ écrémé en poudre, _beurre_ pâtissier, colorant : caramel ordinaire, émulsifiant : lécithines (_soja_), sel. *% exprimé sur le nappage équivalent à 0.8% sur l'ensemble du produit.

chocolat au lait 40% (sucre, beurre de cacao, poudre de _lait_ entier, pâte de cacao, émulsifiant : lécithines (_soja_), arôme), nappage au caramel 21% (sirop de glucose-fructose, stabilisant : glycérol, _lait_ concentré sucré, caramel 4%* (sucre, eau), _beurre_, arôme, sel, gélifiant : pectines), farine de _blé_, sucre, fructose, _lait_ écrémé en poudre, _beurre_ pâtissier, colorant : caramel ordinaire, émulsifiant : lécithines (_soja_), sel. *% exprimé sur le nappage équivalent à 0.8% sur l'ensemble du produit.

sélection des mots-clé

chocolat au lait sucre,beurre de cacao,poudre de lait entier,pâte de cacao,émulsifiant lécithines soja,arôme,nappage au caramel sirop de glucose fructose,stabilisant glycérol,lait concentré sucré,caramel sucre,eau,beurre,arôme,sel,gélifiant pectines,farine de blé,sucre,fructose,lait écrémé en poudre,beurre pâtissier,colorant caramel ordinaire,émulsifiant lécithines soja,sel exprimé sur le nappage équivalent à sur l ensemble du produit

doublons

chocolat au lait sucre,beurre de cacao,poudre de lait entier,pâte de cacao,émulsifiant lécithines soja,arôme,nappage au caramel sirop de glucose fructose,stabilisant glycérol,lait concentré sucré,caramel sucre,eau,beurre,sel,gélifiant pectines,farine de blé,sucre,fructose,lait écrémé en poudre,beurre pâtissier,colorant caramel ordinaire,sel exprimé sur le nappage équivalent à sur l ensemble du produit

traduction

sugar milk chocolate, cocoa butter, whole milk powder, cocoa paste, soy lecithins emulsifier, aroma, caramel crippage fructose glucose syrup, glycérol stabilizer, sweet concentrated milk, caramel sugar, water, butter, salt, gélifying pectin , wheat flour, sugar, fructose, skimmed milk powder, pastry butter, ordinary caramel coloring, salt expressed on the topping equivalent to the whole product

changement de format

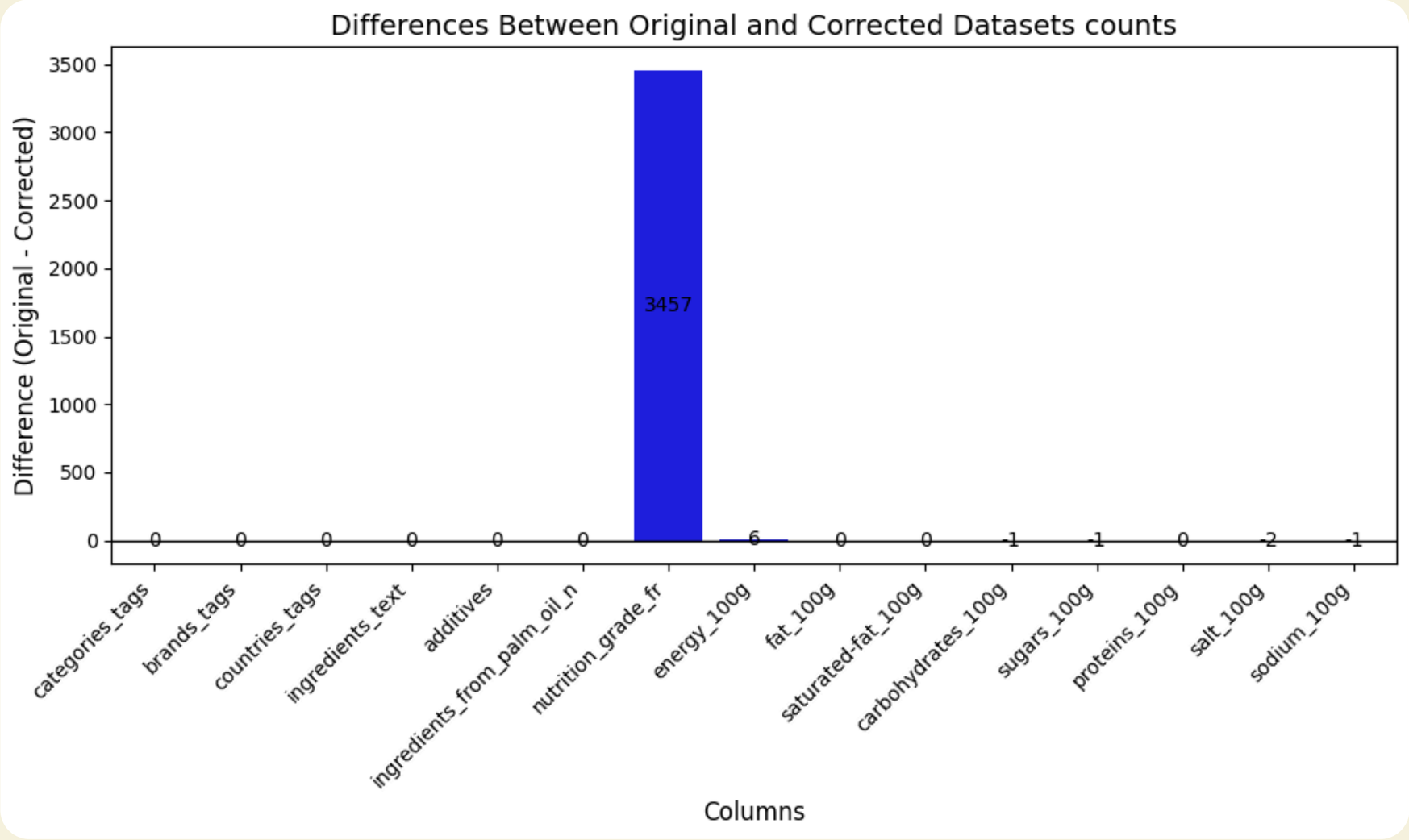
['sugar milk chocolate', ' cocoa butter', ' whole milk powder', ' cocoa paste', ' soy lecithins emulsifier', ' aroma', ' caramel crippage fructose glucose syrup', ' glycérol stabilizer', ' sweet concentrated milk', ' caramel sugar', ' water', ' butter', ' salt', ' gélifying pectin ', ' wheat flour', ' sugar', ' fructose', ' skimmed milk powder', ' pastry butter', ' ordinary caramel coloring', ' salt expressed on the topping equivalent to the whole product']

explosion du dataset + création variable dérivée

	code	allergens_en	ingredients_count_en_sum	ingredients_lists_en
13419	1664	milk,soy,milk,butter,wheat,milk,butter,soy	35	sugar milk chocolate
13419	1664	milk,soy,milk,butter,wheat,milk,butter,soy	35	cocoa butter
13419	1664	milk,soy,milk,butter,wheat,milk,butter,soy	35	whole milk powder
13419	1664	milk,soy,milk,butter,wheat,milk,butter,soy	35	cocoa paste

Nombre de modalités différentes : 23863 --> ~ 4000

Nettoyage des données : bilan









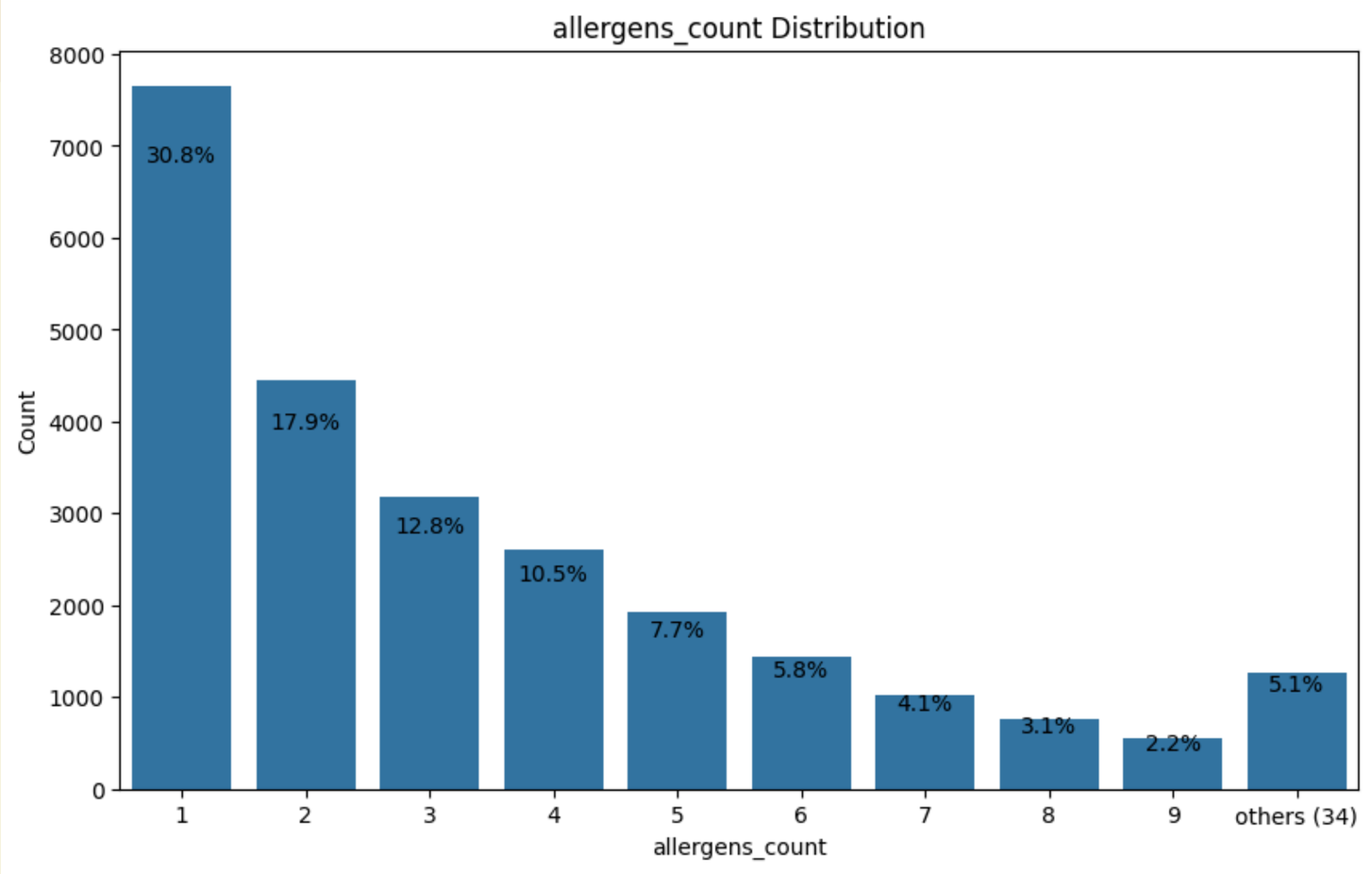
Analyse des données







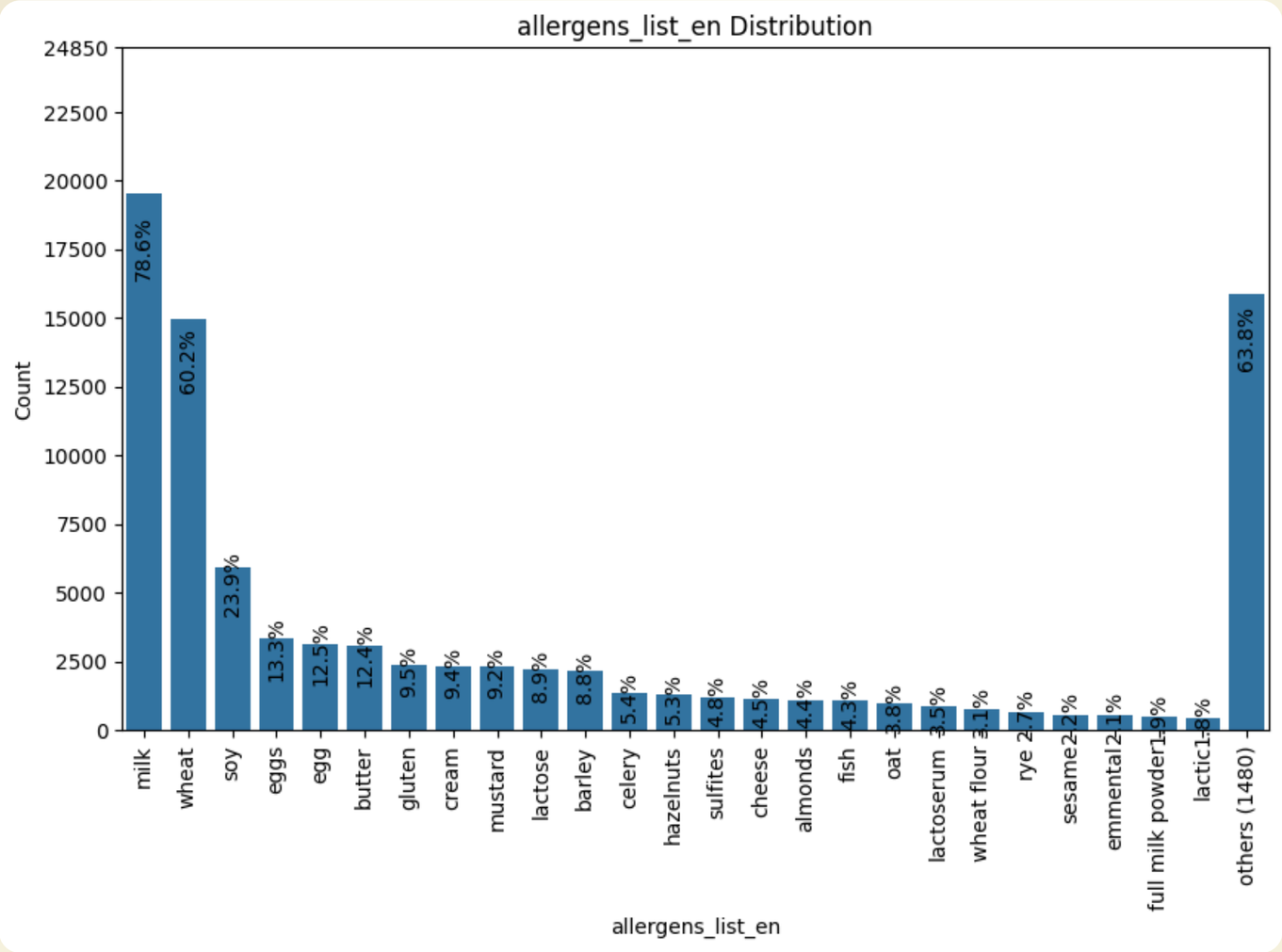
Analyse



Nombre d'allergènes :

- ~ 50% à 2 allergènes ou moins
- environ 3/4 des produits avec allergènes en comportent 5 ou moins

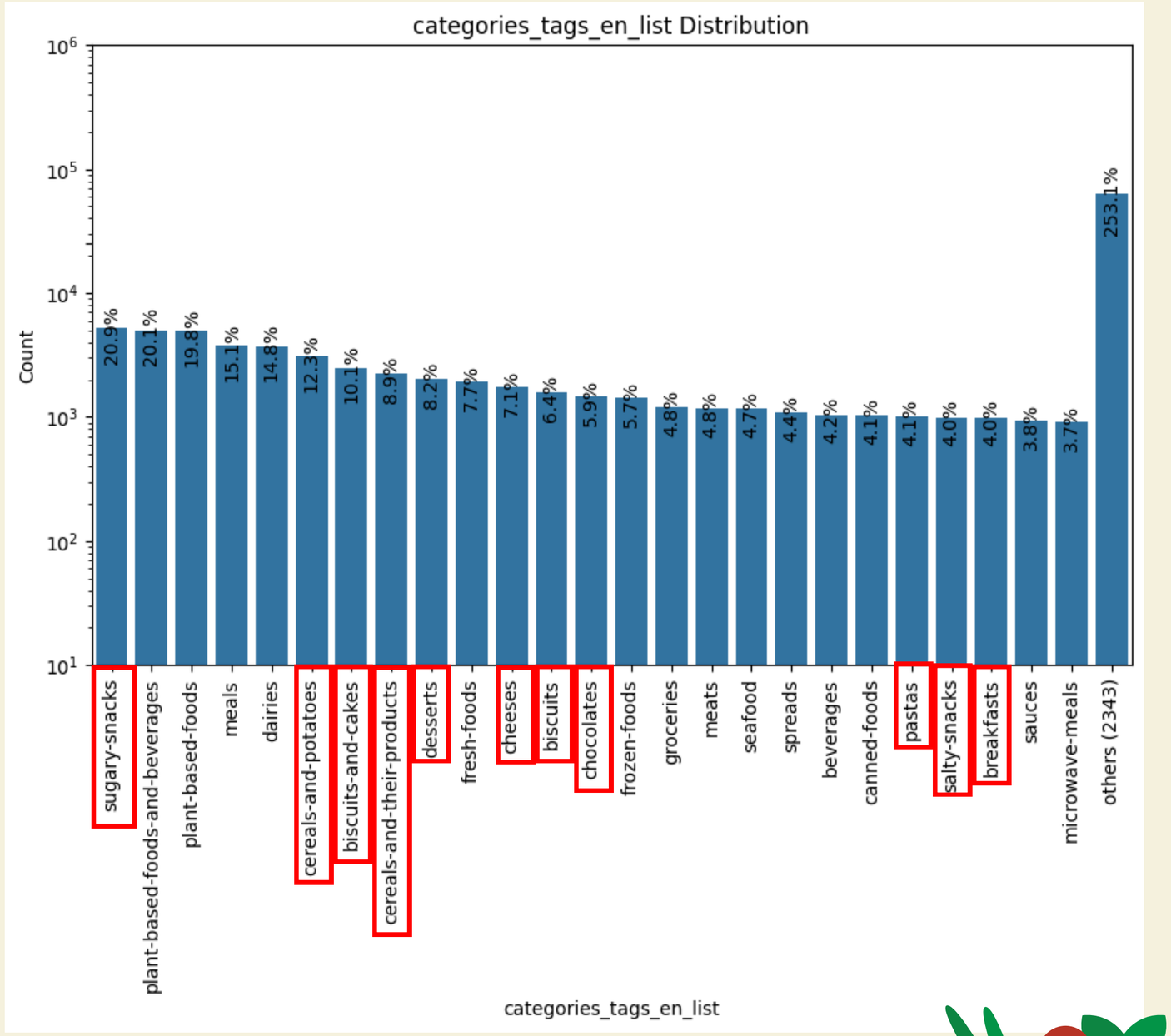
Analyse



Allergènes majoritaires : lait, céréales (gluten), oeufs

Analyse

Produits sucrés et gras semblent surreprésentés



Résultats

Corrélations fortes :

- nombre d'ingrédients
- nombre d'ingrédients répétés

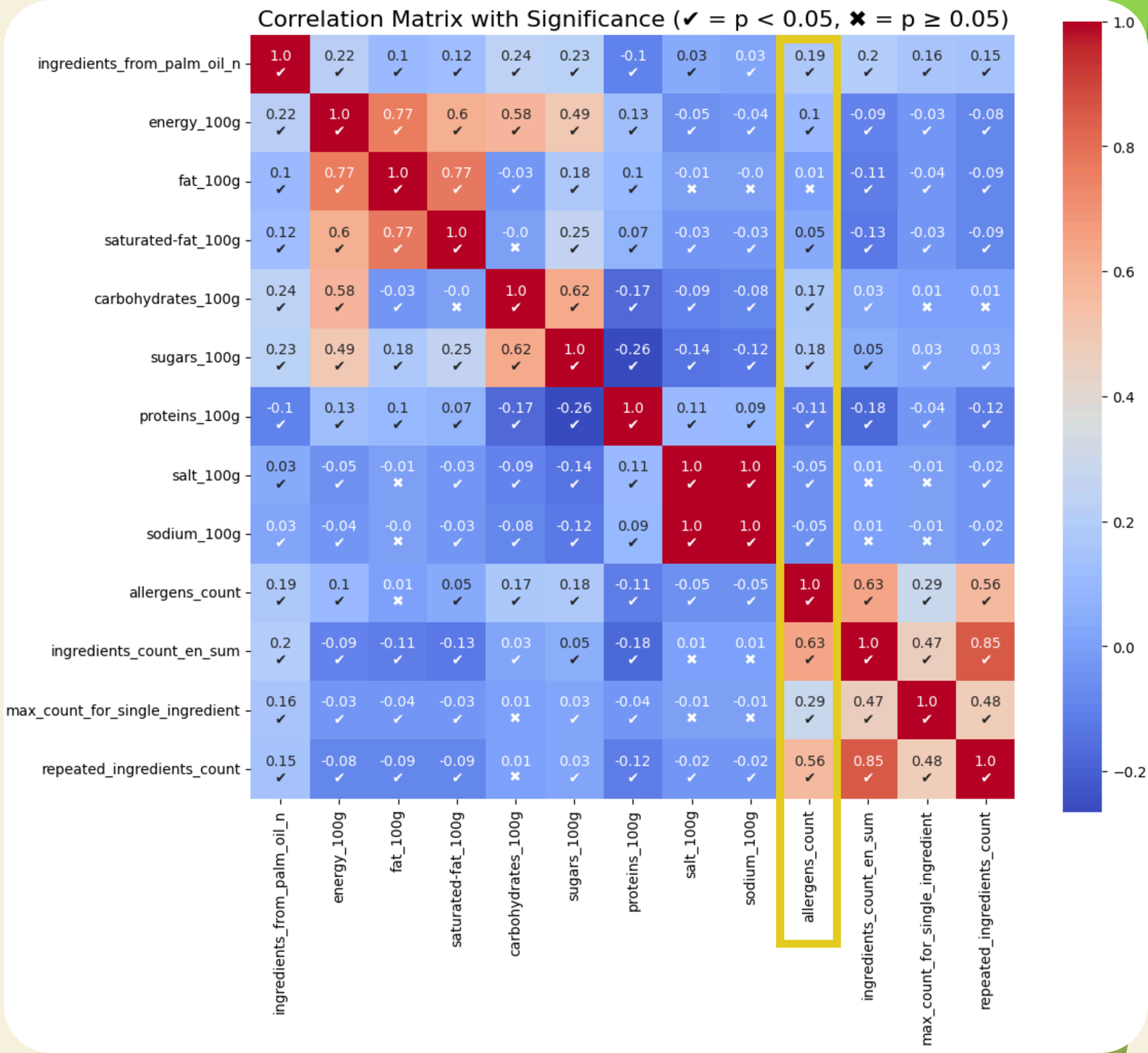
Corrélations modérées avec:

- nombre maximal de répétitions pour un ingrédient

Corrélations faibles:

- nombre d'ingrédients dérivés de l'huile de palme
- sucres (glucides)





Résultats : ACP

Corrélations fortes :

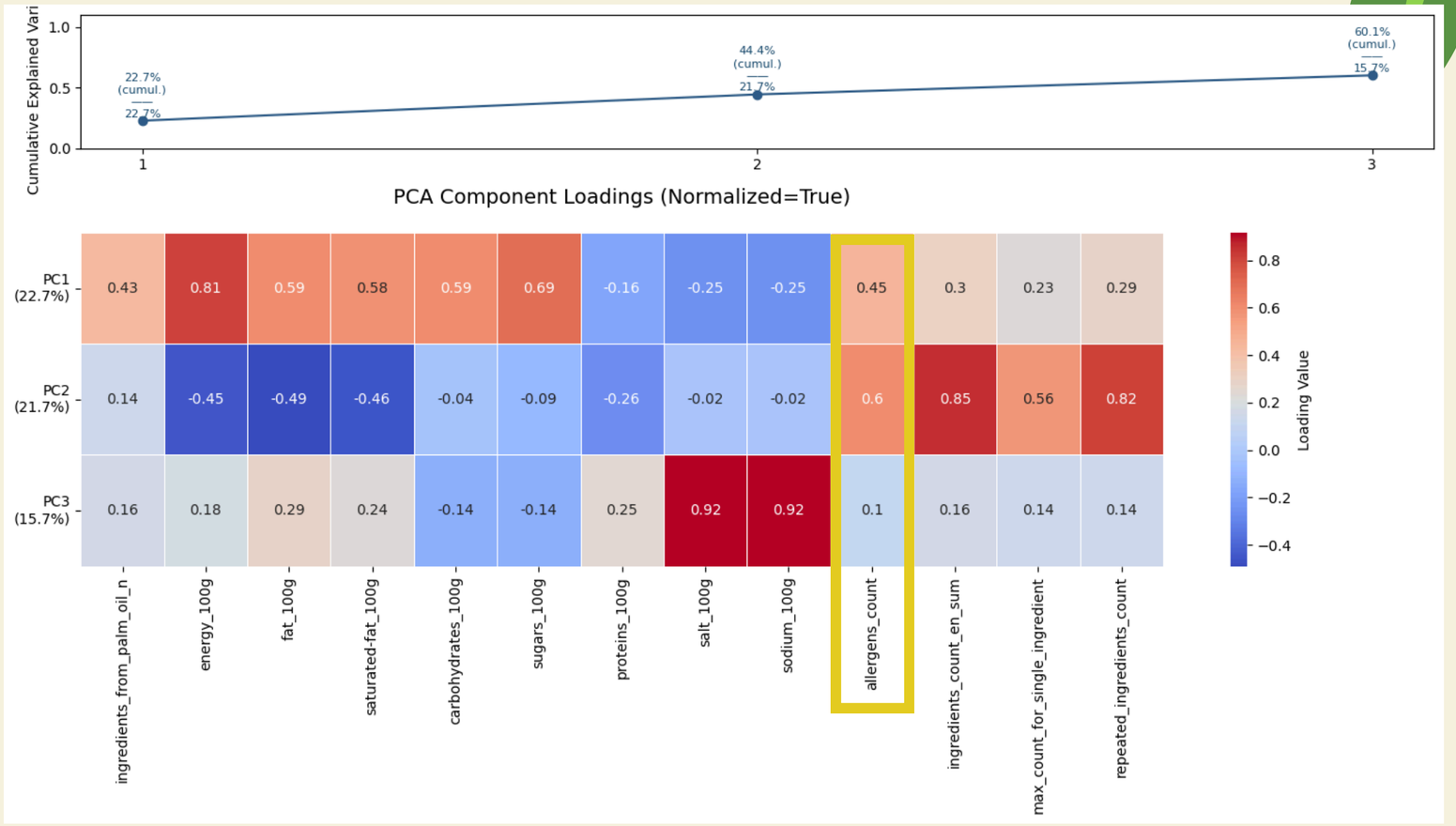
- nombre d'ingrédients
- nombre d'ingrédients répétés

Corrélations modérées avec:

- nombre maximal de répétitions pour un ingrédient

Corrélations faibles:

- nombre d'ingrédients dérivés de l'huile de palme
- sucres (glucides)



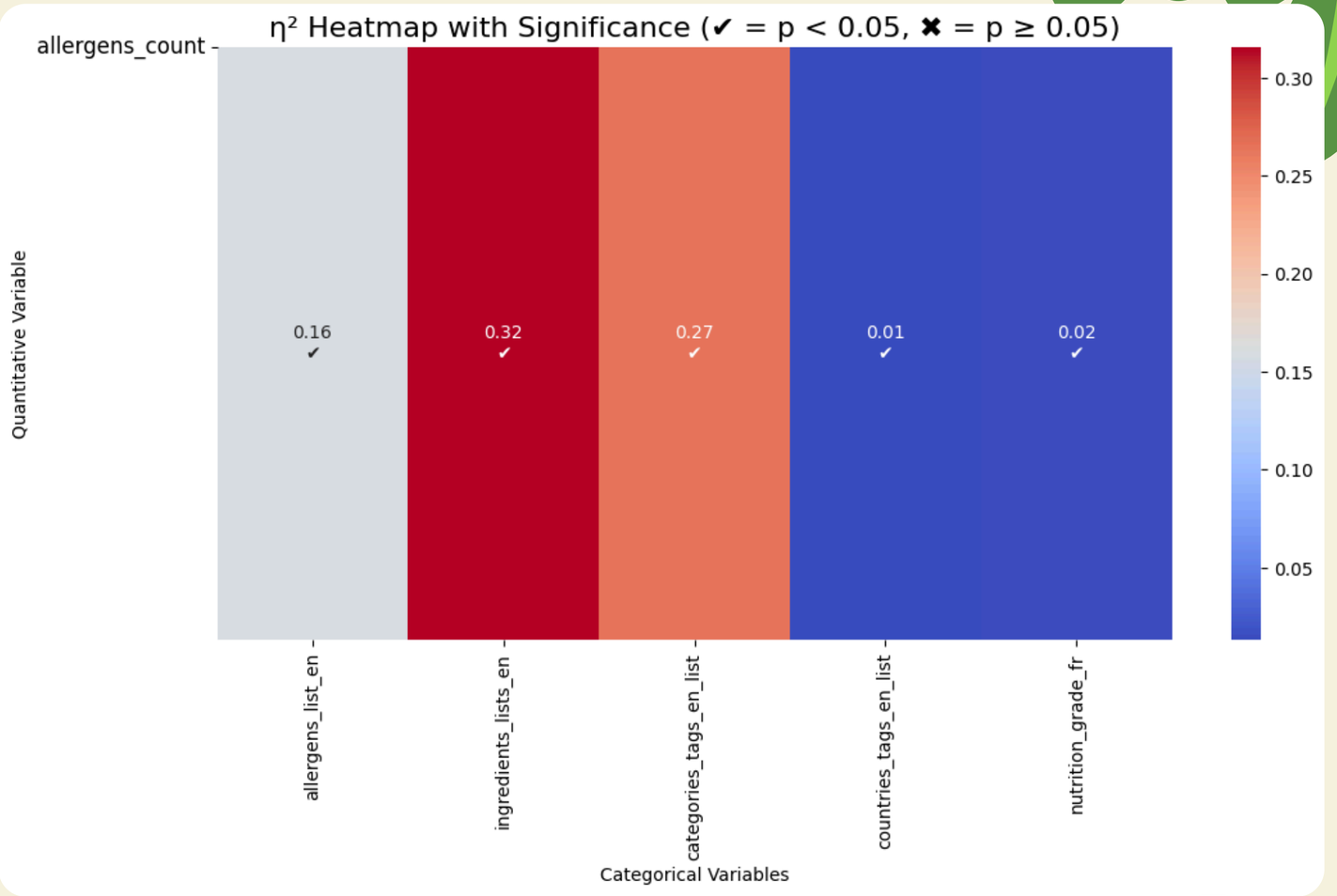
Résultats : ANOVA

3 relations significatives:

- allergènes
- ingrédients
- catégories

MAIS:

listes éclatées ->hypothèse d'indépendance des valeurs à vérifier



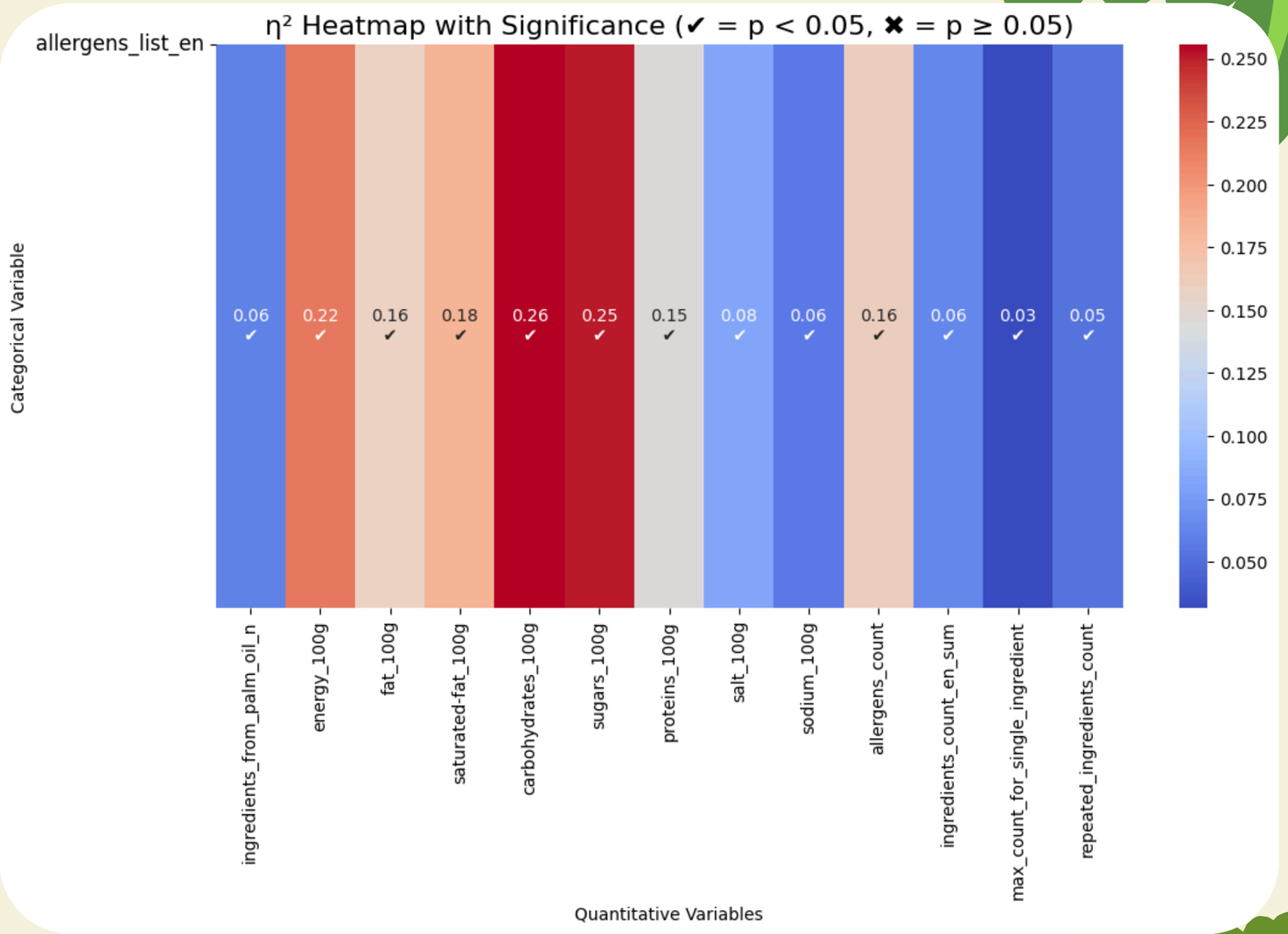
Résultats : ANOVA

5 relations significatives:

- compte d'allergènes
- sucres (glucides)
- énergie
- lipides (acides gras saturés)
- protéines

MAIS:

listes éclatées ->hypothèse d'indépendance des valeurs à vérifier



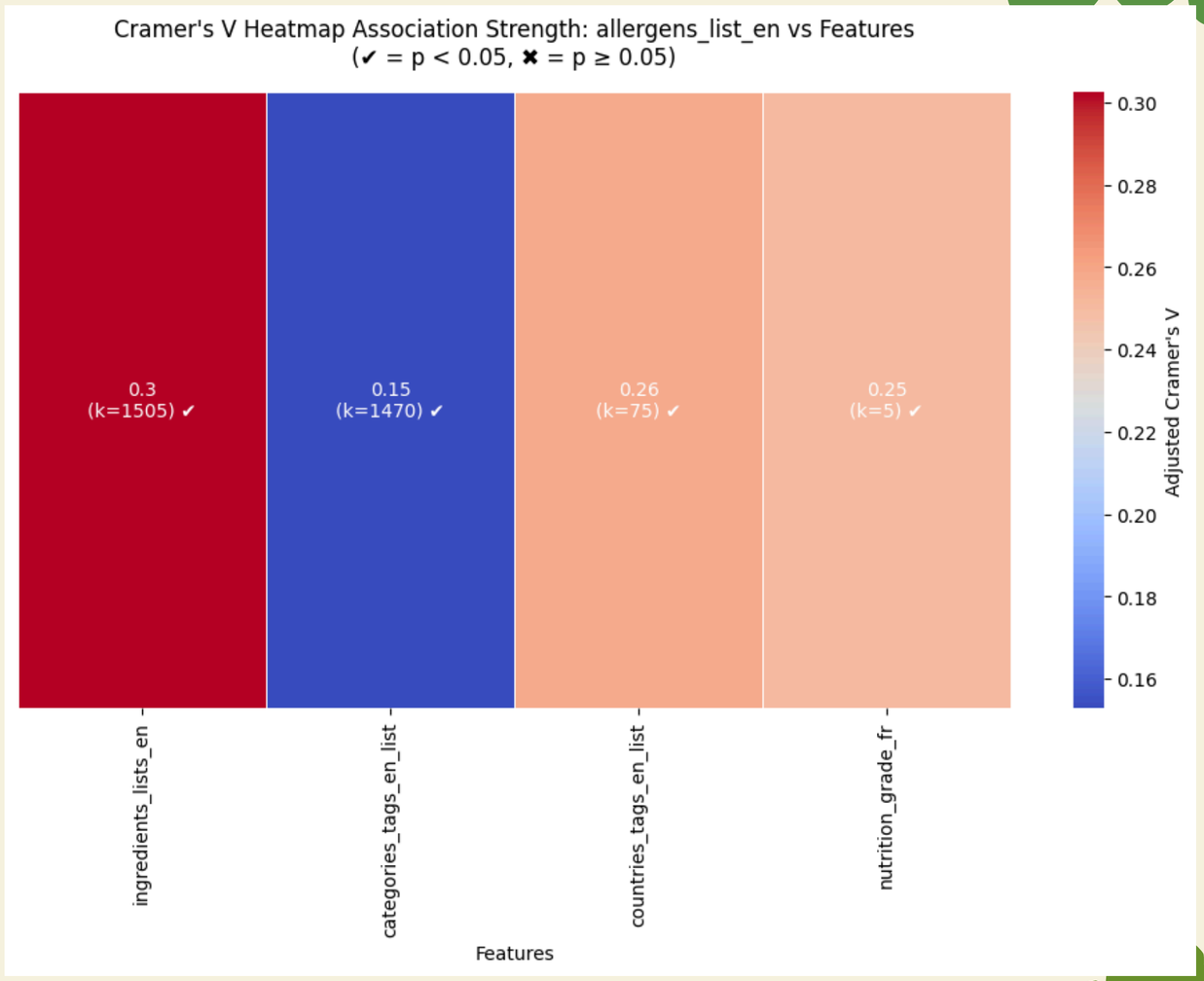
Résultats : Khi^2

3 associations modérées/fortes:

- ingrédients
- pays
- nutri-score

MAIS:

- k élevé (+ petit nombre modalités)
- --> réduire les modalités ou utiliser autre méthode



Résultats : Bilan

Quantitatives

- nombre d'ingrédients
- nombre d'ingrédients répétés
- max répétition ingrédient
- nombre ingrédients huile de palme
- sucres (glucides)
- lipides (acides gras saturés)

Qualitatives

- ingrédients
- catégories
- nutri-score
- pays





Résultats : Bilan

Quantitatives

- nombre d'ingrédients
- nombre d'ingrédients répétés
- max répétition ingrédient
- nombre ingrédients huile de palme
- sucres (glucides)
- lipides (acides gras saturés)

Qualitatives

- ingrédients
- catégories
- nutri-score
- pays

→ 15 variables initiales --> 7 finales (+ 3 dérivées)





Résultats : Bilan

Quantitatives

- nombre d'ingrédients
- nombre d'ingrédients répétés
- max répétition ingrédient
- nombre ingrédients huile de palme
- sucres (glucides)
- lipides (acides gras saturés)

Qualitatives

- ingrédients
- catégories
- nutri-score
- pays

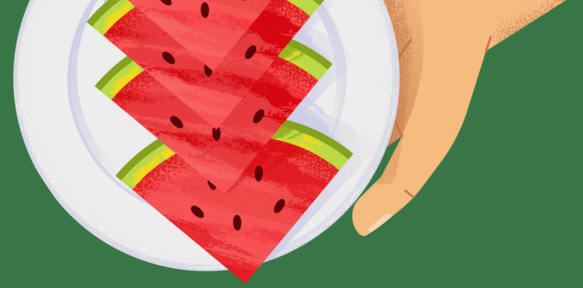
→ 15 variables initiales --> 7 finales (+ 3 dérivées)

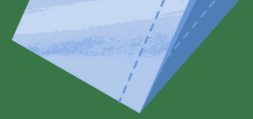


→ Déduction nombre allergènes + auto-complétion (suggestive)











RGPD

1) Licéité, Loyauté et Transparence : Consentement clair, information sur l'utilisation des données

2) Limitation des finalités: Utilisation stricte pour améliorer la base de données, sans autres utilisations.

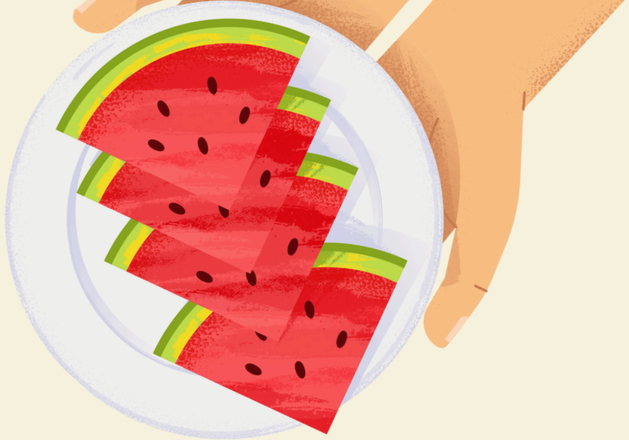
3) Minimisation des données: Collecte d'informations essentielles uniquement sur le produit (nom, ingrédients, allergènes). Pas de traceur, pas de revente de données personnelles

4) Exactitude: Système de validation pour vérifier et mettre à jour les saisies.

5) Conservation: aucune donnée personnelle conservée









THANK YOU

Do you have any questions?



