

Mars 2024



Données alimentaires : Un enjeu de santé publique

Par Lokman AALIOUI

Introduction

- L'agence Santé publique France souhaite améliorer sa base de données Open Food Facts. Cette base de données en ligne permet aux entreprises et aux particuliers de consulter les informations sur certains produits du commerce.
- L'objectif d'Open Food Facts est de garantir la qualité nutritionnelle des produits alimentaires, un enjeu majeur pour la santé publique.





Notre mission

- Actuellement, l'ajout de produits à cette base de données nécessite la saisie de nombreuses informations, ce qui peut conduire à des erreurs et dans les données.
- Nous sommes donc mandaté pour créer un système de suggestion ou d'auto-complétion pour faciliter le processus de saisie des utilisateurs.

Est-il possible de créer un système de suggestion ou d'auto-complétion ?



01. Comprendre les données
02. Traiter les données pour les rendre exploitables
03. Analyser des données pour en tirer une conclusion



01.

Observation des données

Les données

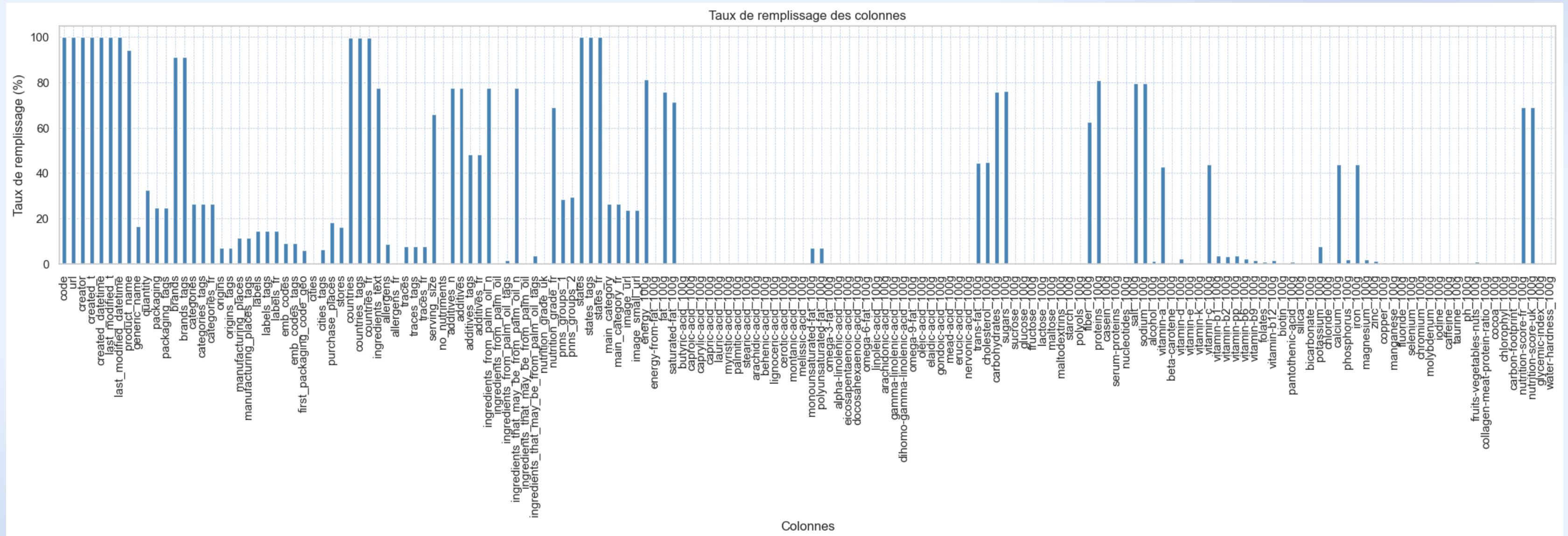
320772 lignes :

- chaque ligne correspond à un produit vendu dans le commerce

162 colonnes :

- 56 variables catégorielles : code produit, nom, marque, pays de provenance, catégorie, etc...
- 106 variables numériques : valeurs nutritionnelles, additifs, vitamines, index glycémique, etc

Sélection des variables : approche quantitative



- Sélection de 28 variables remplies à plus de 75%

Sélection des variables : approche qualitative

18 variables catégorielles

```
['code',
 'url',
 'creator',
 'created_t',
 'created_datetime',
 'last_modified_t',
 'last_modified_datetime',
 'product_name',
 'brands',
 'brands_tags',
 'countries',
 'countries_tags',
 'countries_fr',
 'ingredients_text',
 'additives',
 'states',
 'states_tags',
 'states_fr']
```

10 variables numériques

```
['additives_n',
 'ingredients_from_palm_oil_n',
 'ingredients_that_may_be_from_palm_oil_n',
 'energy_100g',
 'fat_100g',
 'carbohydrates_100g',
 'sugars_100g',
 'proteins_100g',
 'salt_100g',
 'sodium_100g']
```

Variables pertinentes :

/

- 'code',
- 'product_name',
- 'brands',
- 'countries_fr'

/

- 'energy_100g',
- 'proteins_100g',
- 'fat_100g',
- 'carbohydrates_100g',
- 'sugars_100g'
- 'salt_100g'

02.

Traitements des données

Traitement des valeurs aberrantes

Avec une approche métier, les conditions suivantes doivent être remplies :

energy_100g :

- valeur comprise entre 0 et 4000 (kJ)

proteins_100g, fat_100g, carbohydrates_100g, sugars_100g, salt_100g :

- valeur comprise entre 0 et 100 (g)

sugars_100g

- sugars_100g < carbohydrates_100g

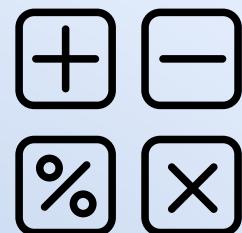
Les valeurs ne respectant pas ces conditions sont remplacées par NaN

Traitements des valeurs manquantes

Les différentes méthodes de traitement des valeurs manquantes :

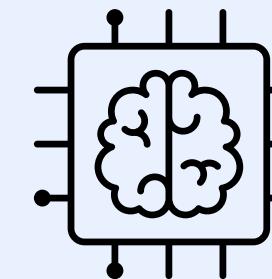
Méthodes statistiques :

- Médiane
- Mise à zéro



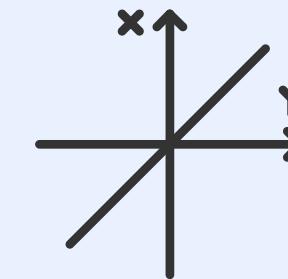
Méthodes automatiques :

- Iterative Imputer
- K-Nearest Neighbors
Imputer



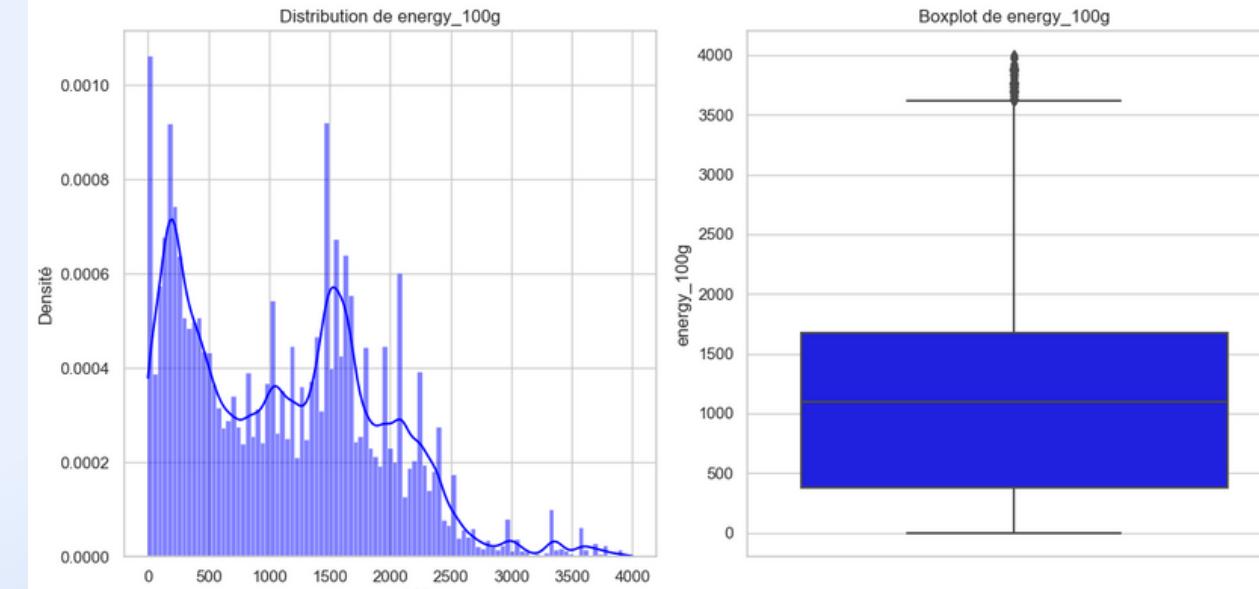
Autres méthodes :

- Regression linéaire

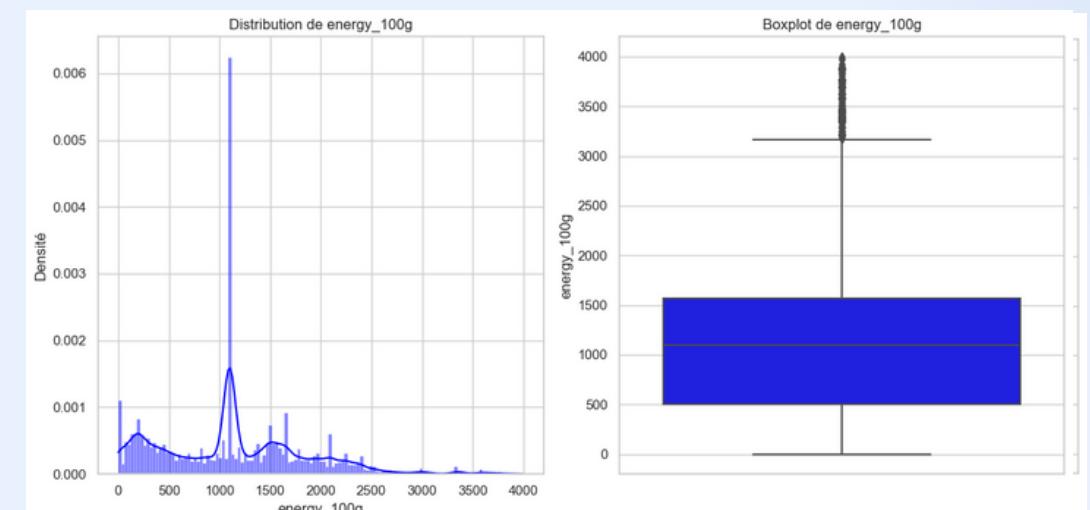


Traitements des valeurs manquantes : variable énergie

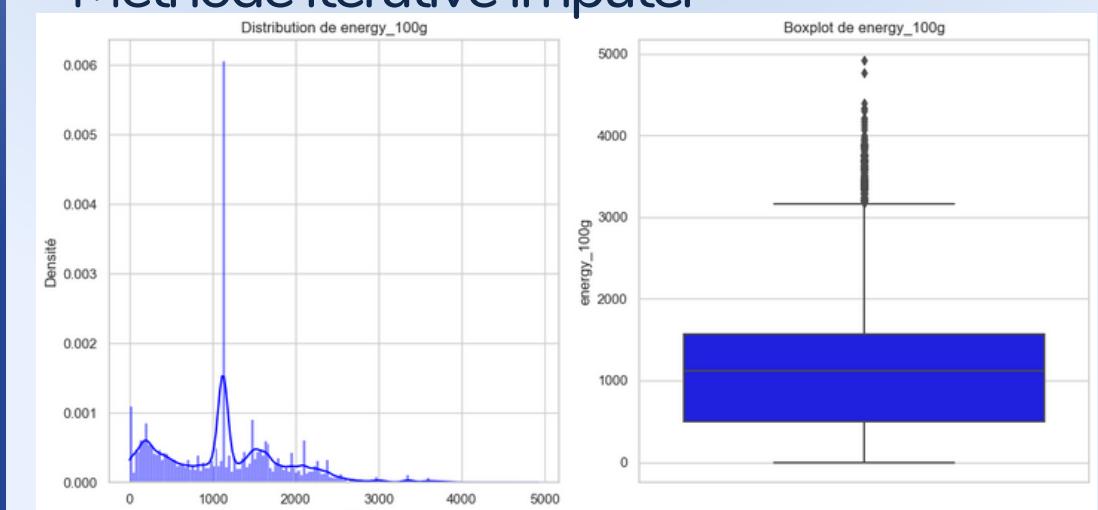
Distribution avant remplacement des valeurs manquantes



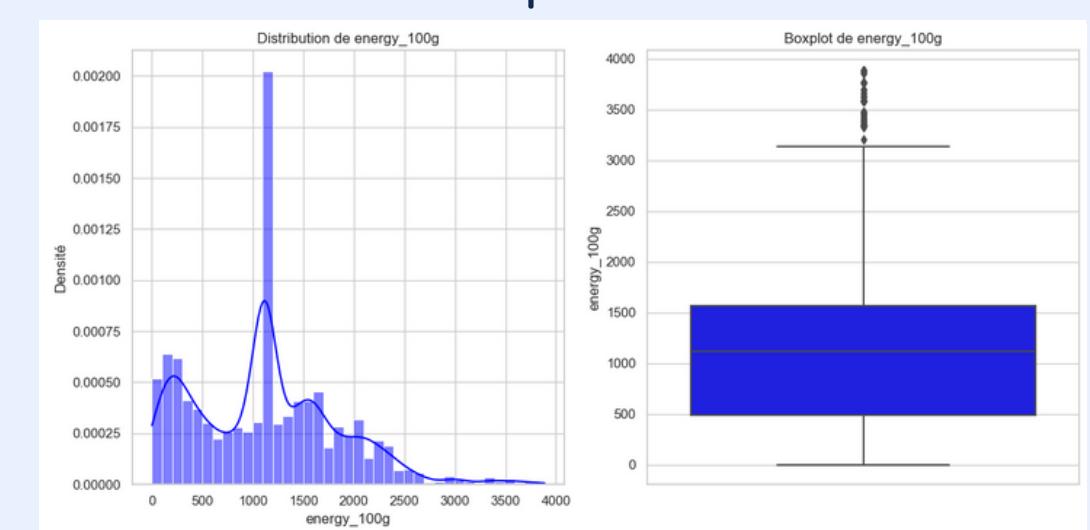
Méthode de la médiane



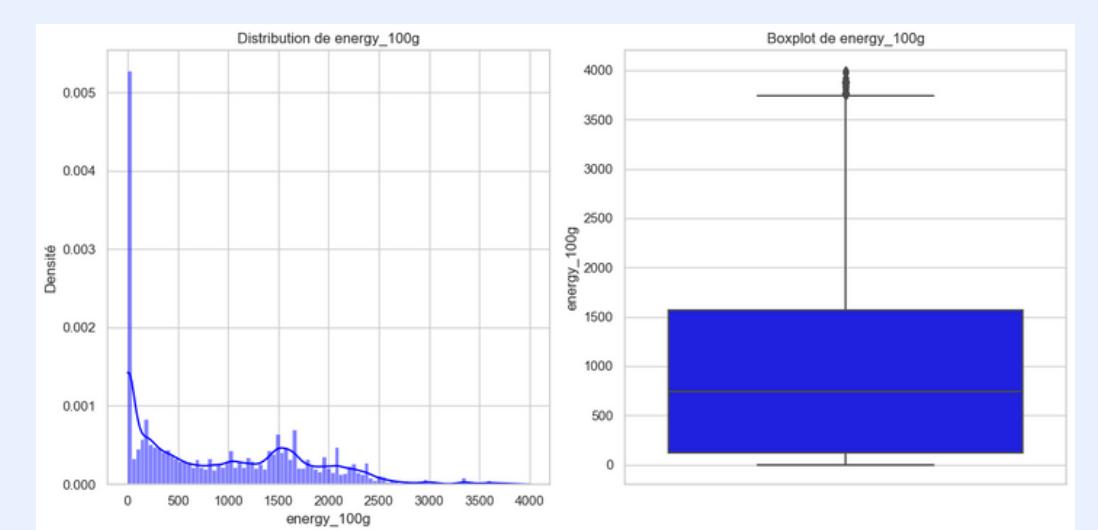
Méthode Iterative Imputer



Méthode KNN Imputer



Méthode mise à zéro



Traitement des valeurs manquantes : autres variables

Méthodes de remplissage utilisées :

- energy_100g >
- proteins_100g >
- fat_100g >

Médiane
Médiane
KNN Imputer

- carbohydrates_100g >
- sugars_100g >
- salt_100g >

KNN Imputer
Iterative Imputer
Mise à zéro

Notre projet respecte-t-il les règles RGPD ?



Le RGPD est un ensemble de règles qui protègent les informations personnelles des utilisateurs en les informant sur la façon dont elles sont utilisées.



Consentement

Obtenir le consentement explicite des individus avant de collecter, utiliser ou traiter leurs données personnelles



Transparence

Fournir des informations claires sur la manière dont les données personnelles sont utilisées



Droit à l'information

Droit de savoir quelles données sont collectées, comment elles sont utilisées et avec qui elles sont partagées



Droit d'accès

Droit pour les utilisateurs d'accéder aux données personnelles détenues à leur sujet



Sécurité des données

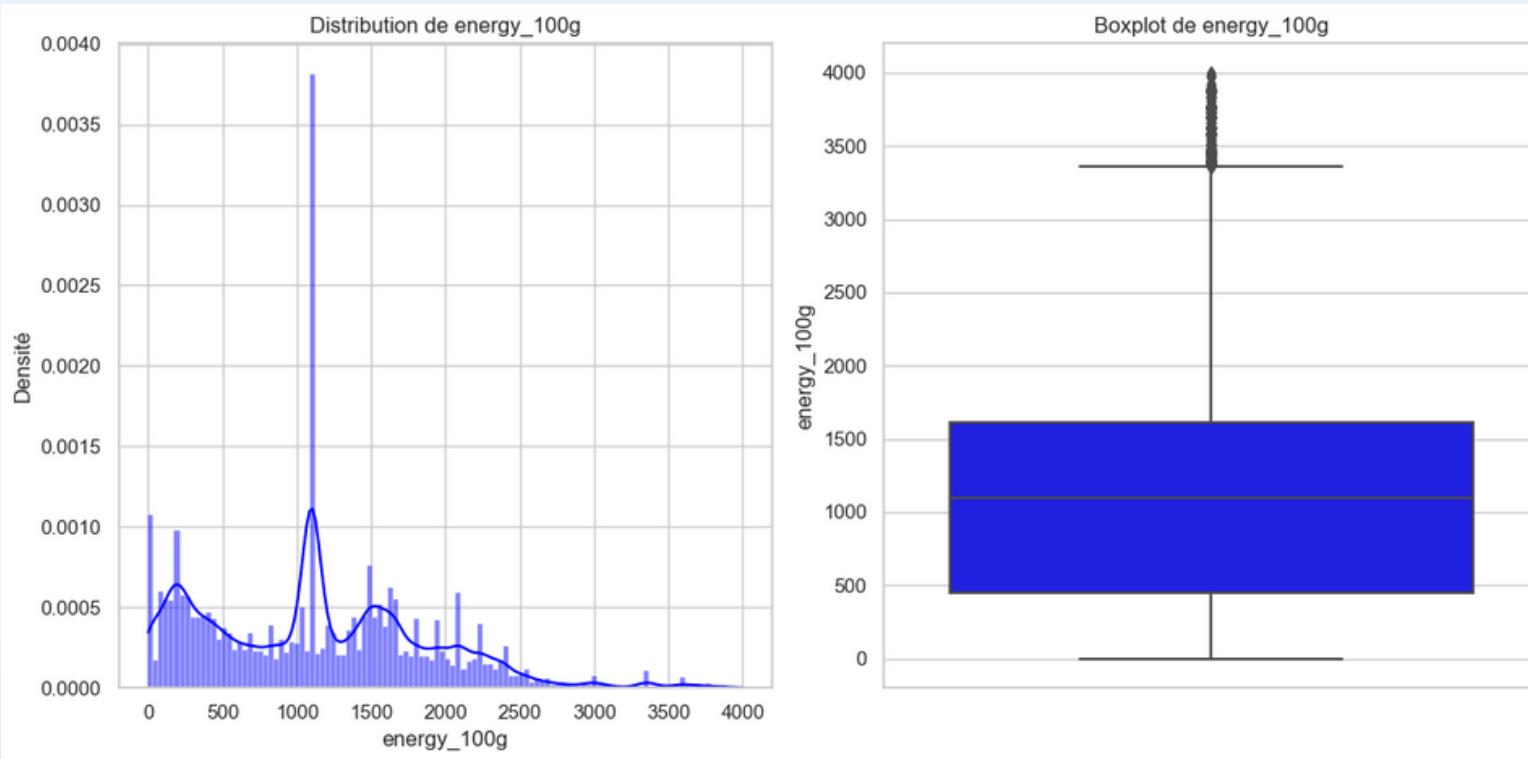
Mettre en place des mesures de sécurité pour protéger les données personnelles contre la perte, le vol ou l'accès non autorisé

03.

Analyse des données

Analyse des variables quantitatives : Avec quelle valeurs simuler variables ?

- energy_100g



Moyenne : 1120.21
Médiane : 1100
Écart-type : 745.93
Mode : 1100

- proteins_100g

Moyenne : 6.78
Médiane : 4.76
Écart-type : 7.66
Mode : 0.0

- carbohydrates_100g
- Moyenne : 32.03
Médiane : 29.0
Écart-type : 26.88
Mode : 32.01

- salt_100g
- Moyenne : 1.37
Médiane : 0.3
Écart-type : 5.86
Mode : 0.0

- fat_100g

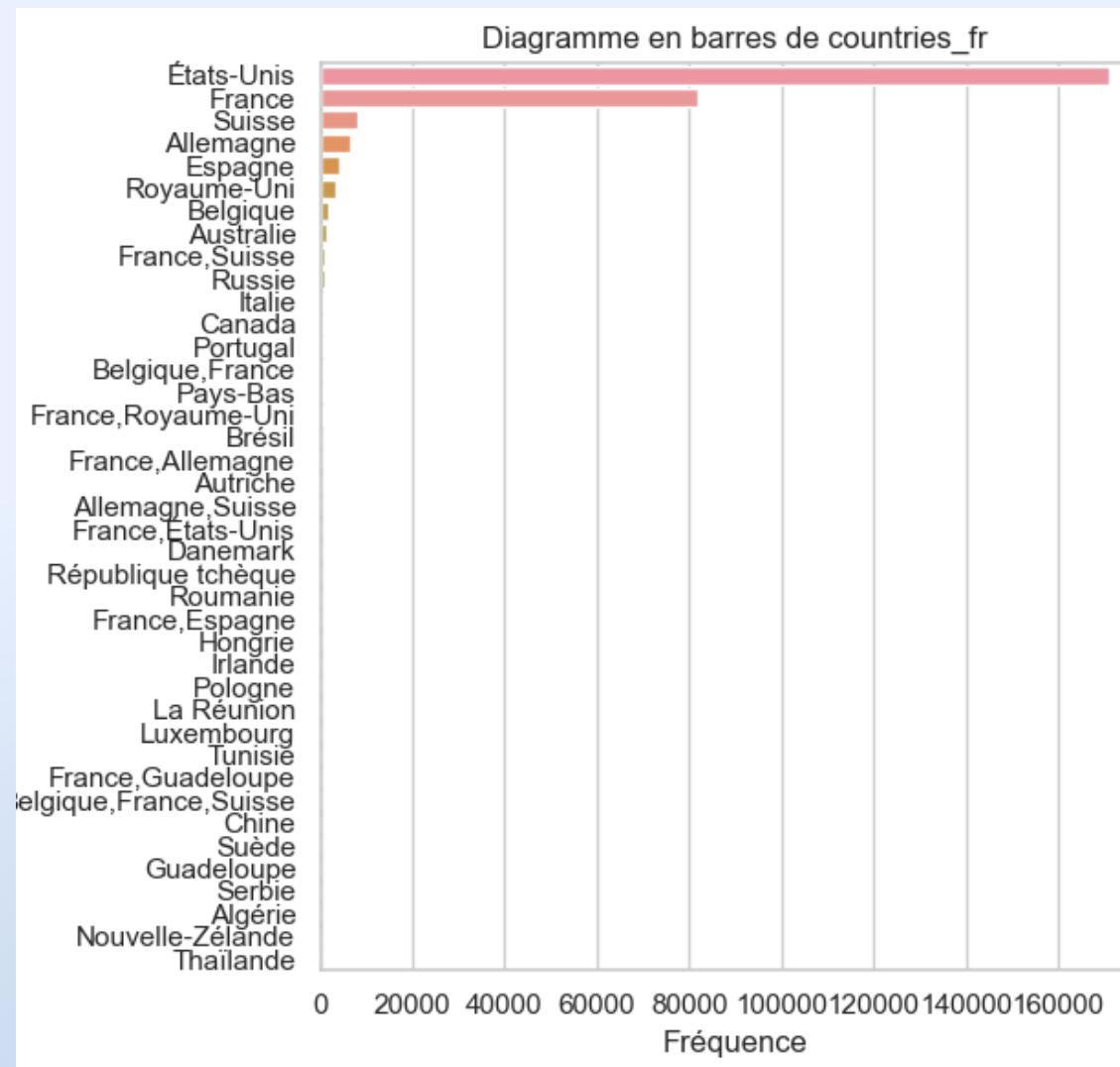
Moyenne : 10.40
Médiane : 1.89
Écart-type : 16.57
Mode : 0.0

- sugars_100g
- Moyenne : 15.43
Médiane : 7.92
Écart-type : 19.58
Mode : 0.0

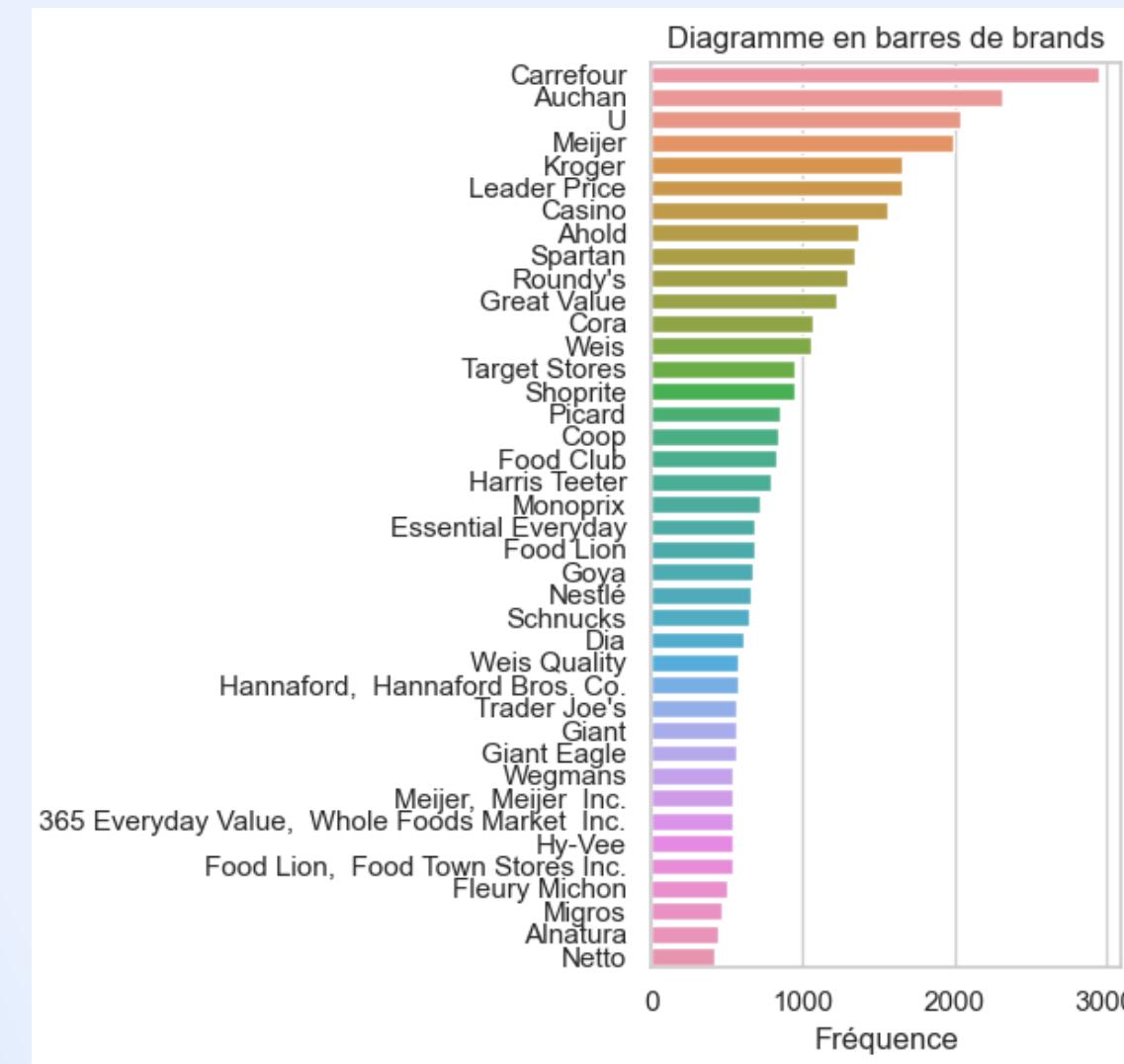
Analyse des variables qualitatives : quelles variables sont le plus représentées ?

Classement des 40 valeurs les plus représentés pour les variables qualitatives

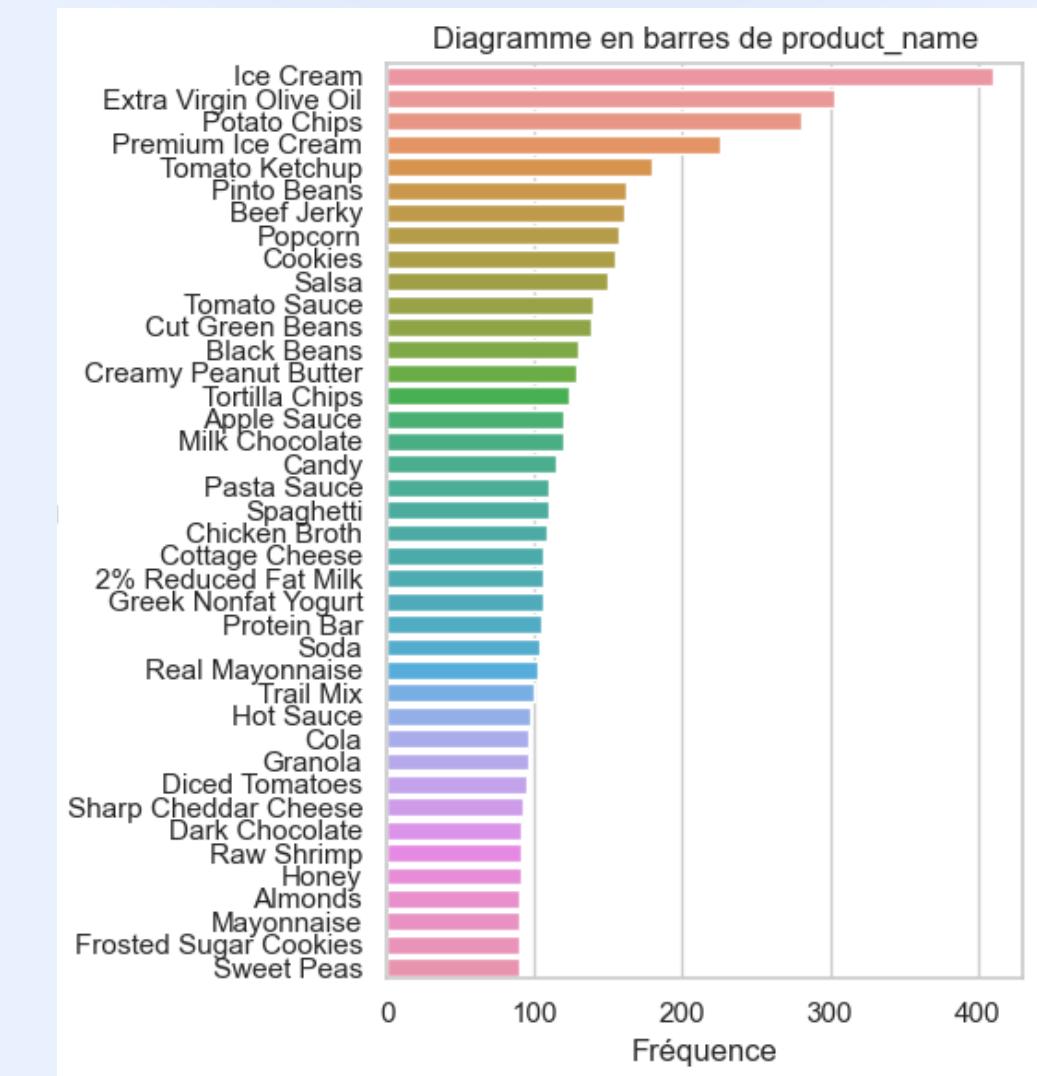
- Pays les plus représentés



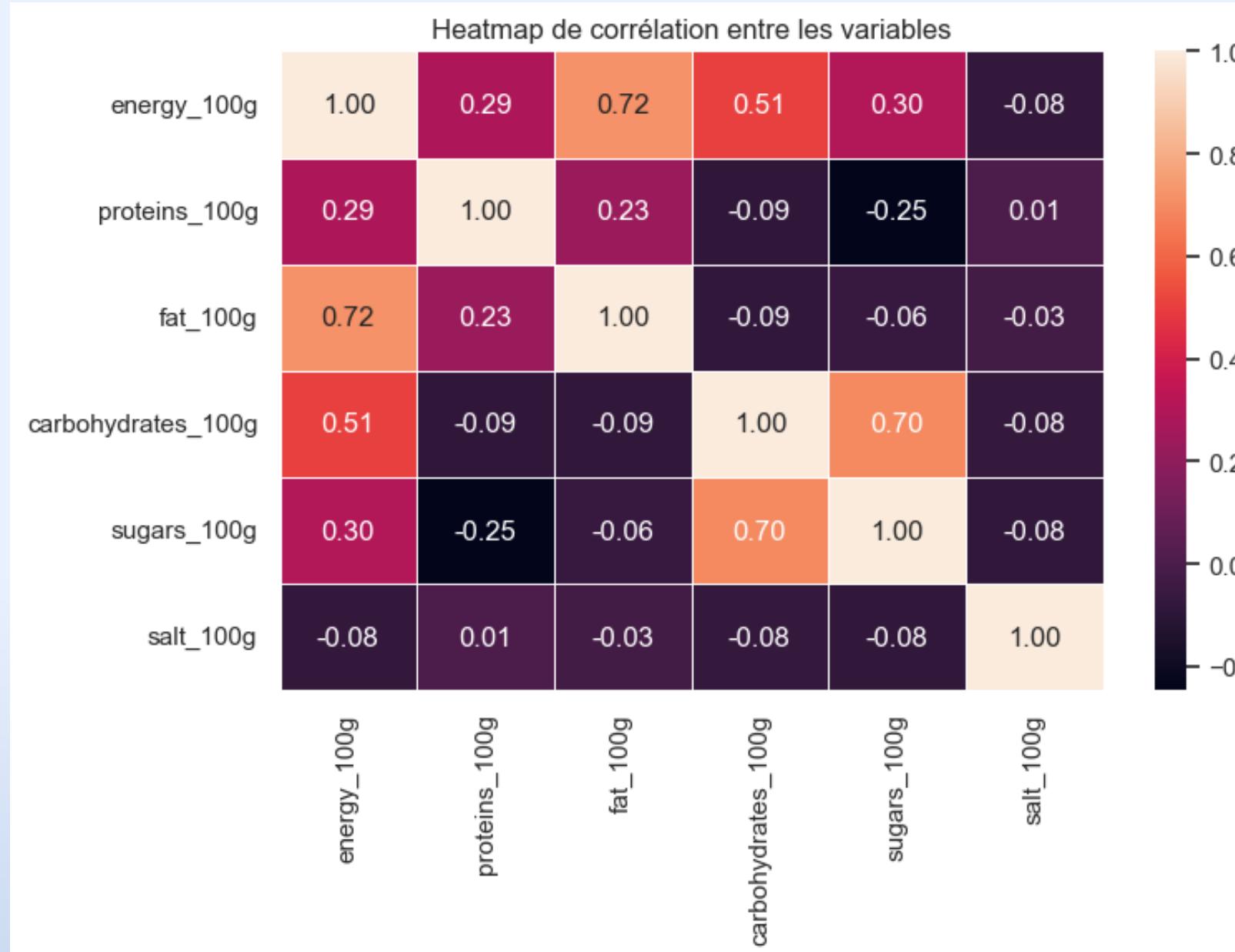
- Marques les plus représentés



- Produits les plus représentés



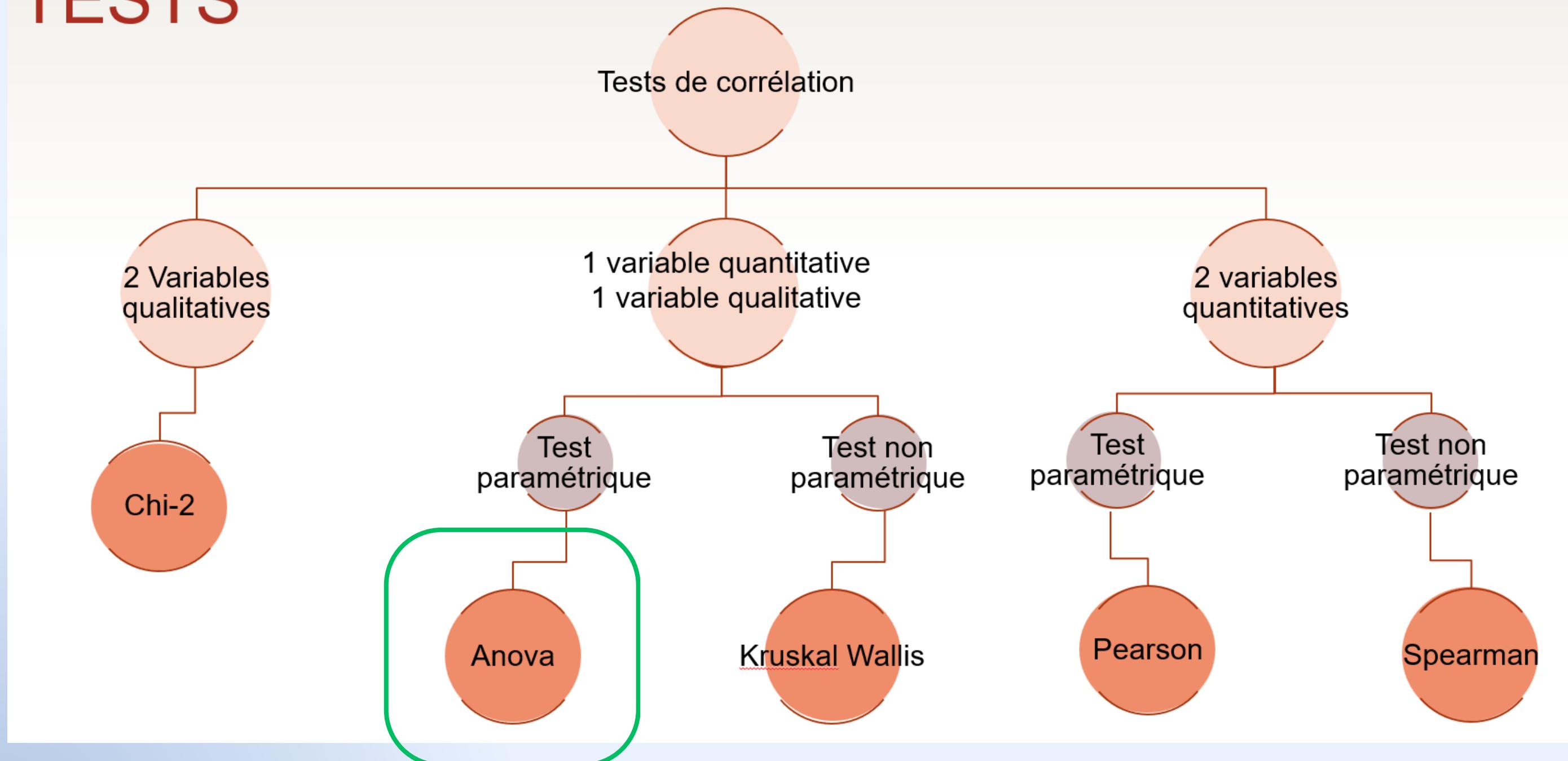
Analyse bivariée : quelles sont les relations entre les variables quantitatives?



- Énergie et graisses : Une forte corrélation (0.72)
- Glucides et sucres : Corrélation positive significative (0.70)
- Énergie et glucides : Corrélation moyenne (0.51)
- Protéine et énergie : Corrélation faible (0.29)
- Sel : Pas de corrélation avec les autres variables

Tests de corrélations

STRATEGIE D'APPLICATION DES TESTS



Test ANOVA : Les variables qualitatives peuvent t-elles améliorer l'auto-complétion ?

La marque influe t elle sur les calories du produit ?

- P-value : 0.0
- Rejet de l'hypothèse nulle : La marque influe sensiblement sur les calories.



Top marques :

- Olio Carli
- Huilerie Croix Verte
- Oliviers & Co.

Le pays influe t il sur les calories du produit ?

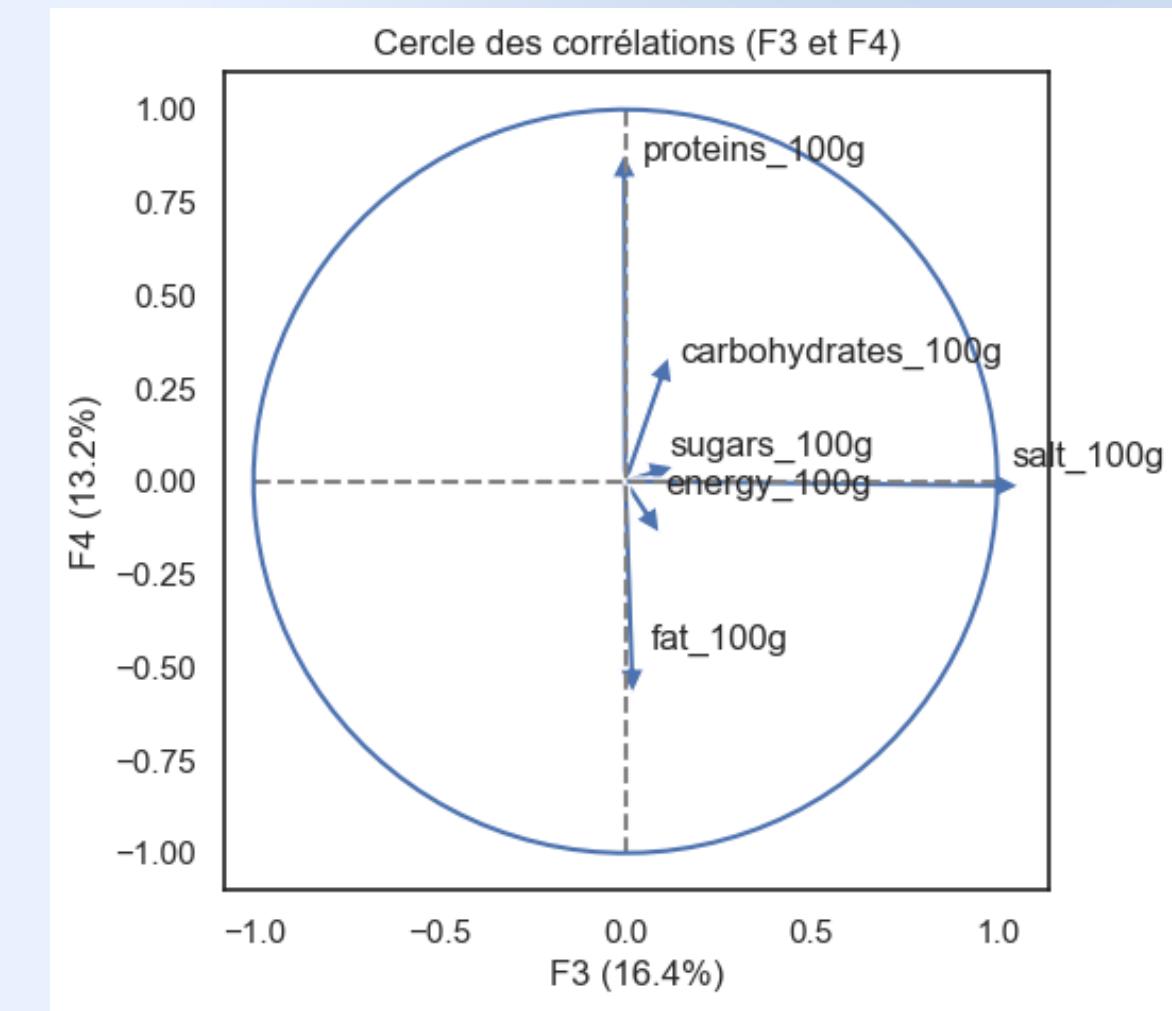
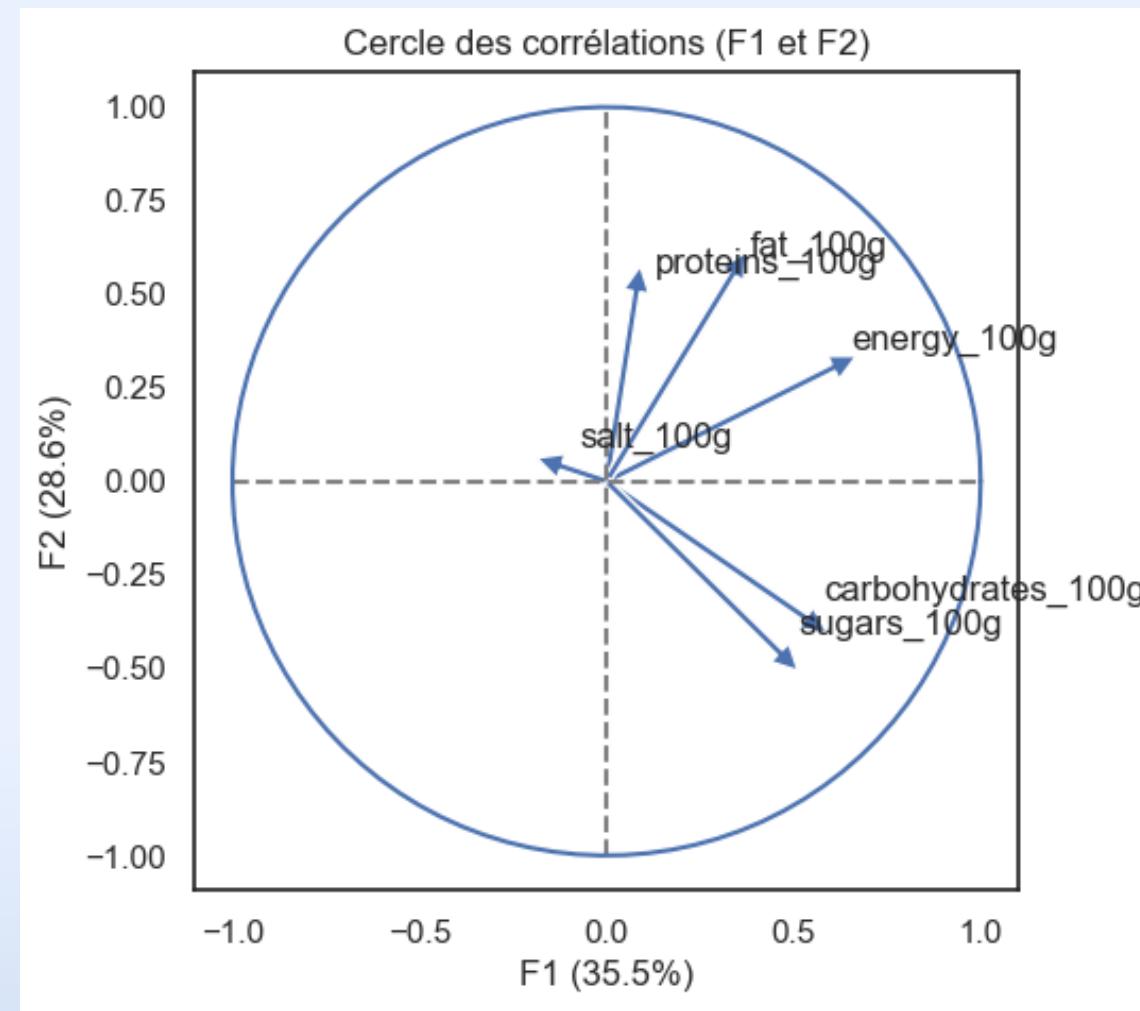
- P-value : 4.09e-12
- Rejeter l'hypothèse nulle : Le pays influe sensiblement sur les calories.



Top pays :

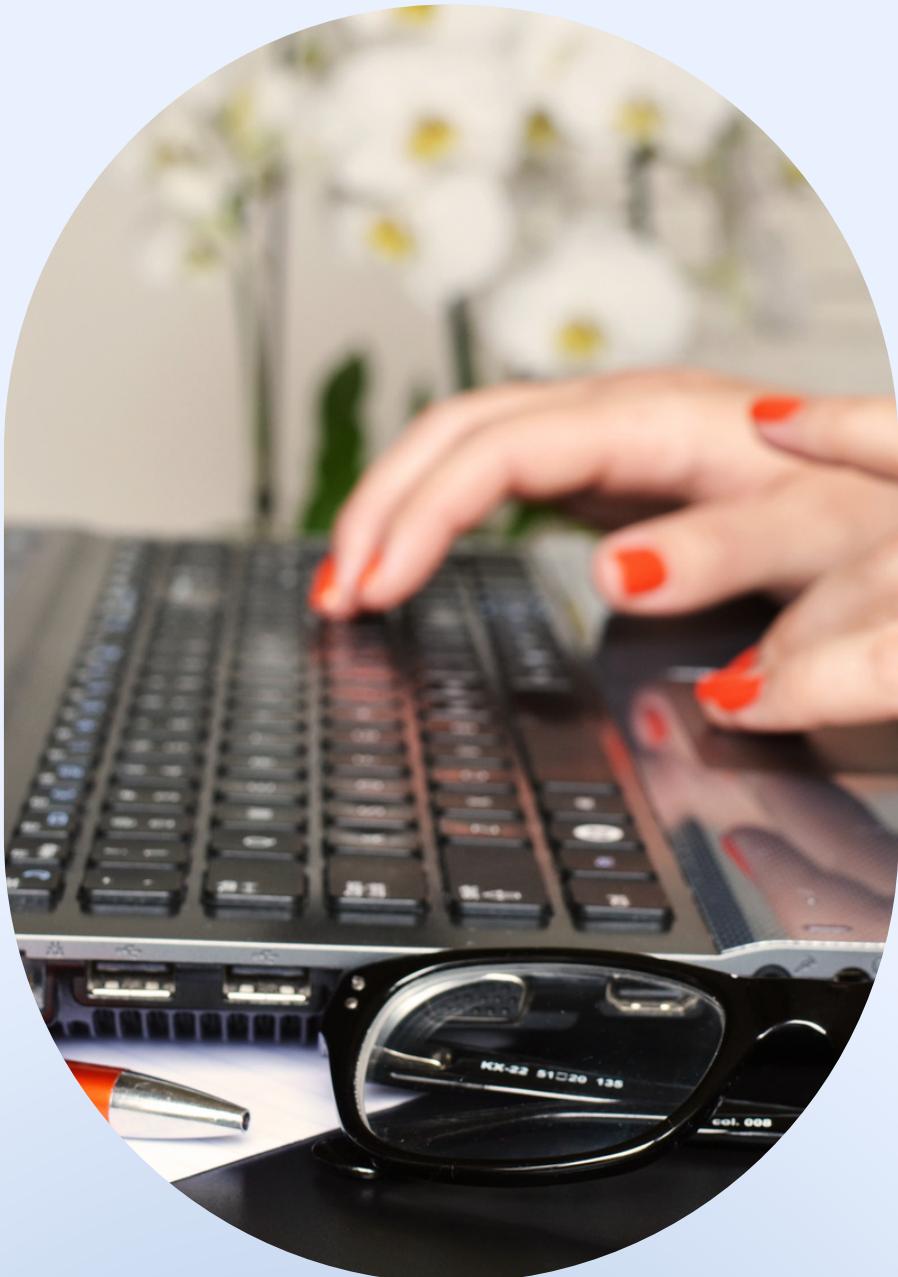
- Allemagne
- Etats-Unis
- Australie

Test ACP : Les corrélations entre les variables peuvent - elles affiner l'auto-complétion ?



- Graisses, protéines et énergie : corrélation positive entre eux et une contribution significative à la composante F1
- Glucides et sucres : contribution principale à la composante F2, expliquent la variabilité des données d'une manière différente
- Sel : moindre contribution aux deux premières composantes principales

- Protéines : contribution significative à la composante F3
- Graisses : contribution principale à la composante F4, différences dans la variabilité des données par rapport à la composante F3
- Glucides, sucres, énergie et sel : contribution moins marquée à ces composantes



Conclusion

Nous concluons que :

- Une approche métier est primordiale pour mener un projet
- Après une bonne préparation, il est possible de simuler les valeurs de certaines variables
- Les règles RGPD doivent être considérées dans chaque projet

Afin de créer un système d'auto-complétion efficace, il faut donc :

- S'assurer que les suggestions prennent en compte les corrélations entre les variables
- Utiliser des techniques de simulation adaptées à chaque variables
- Prendre en compte les groupes pour affiner les suggestions



MERCI POUR VOTRE ATTENTION !