Santé publique France

# Préparez des données pour un organisme de santé publique

Presentation - 2023

Visitez notre site
https://www.santepubliquefrance.fr/

# Á propos de l’entreprise

Santé Publique France est l'agence nationale française chargée de la protection et de l'amélioration de la santé publique. Elle travaille sur une multitude de domaines liés à la santé, tels que la surveillance épidémiologique, la prévention des maladies, la promotion de la santé, etc. L'agence joue un rôle essentiel dans la collecte de données, l'analyse des risques sanitaires et la communication d'informations cruciales pour la santé publique en France.

# Notre mission

L’agence Santé publique France confie à notre entreprise la création d’un système de suggestion ou d’auto-complétion pour aider les usagers à remplir plus efficacement la base de données.

# Sommaire

## Nettoyage des données

Définir les objectifs de nettoyage de données en fonction de la problématique métier

## Analyses  Univariées et bivariées

Caractérisé correctement les distributions observées.
Graphiques d'analyse bivariées, univariées.

## Analyses Multivariées

Méthode d'analyse multivariée descriptive sur le jeu de données (ACP).

## Conclusion

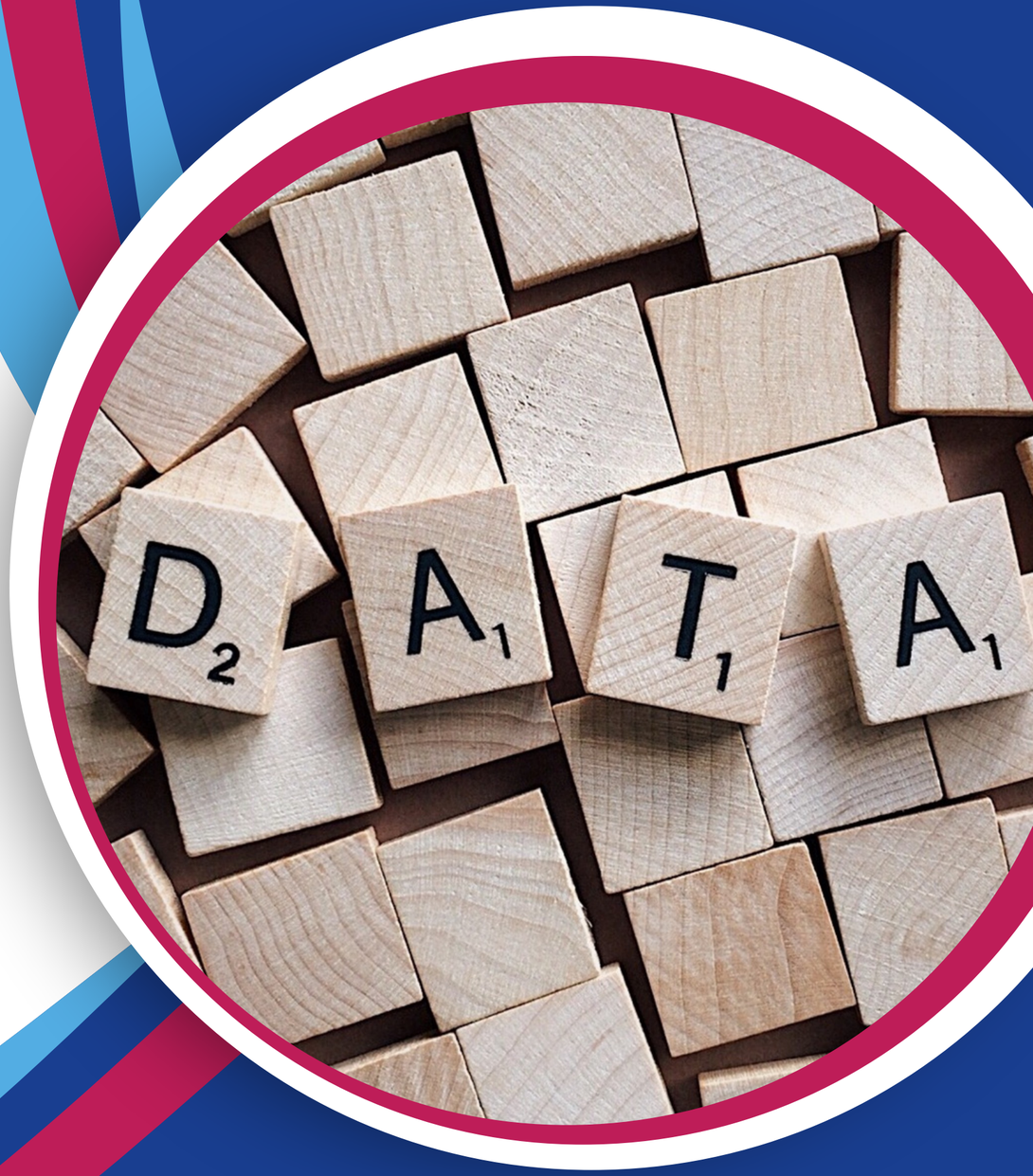
# Détails des données :

**Nombre de variables**
162

**Nombre de lignes**
320 772

**Compositions nutritionnelles**

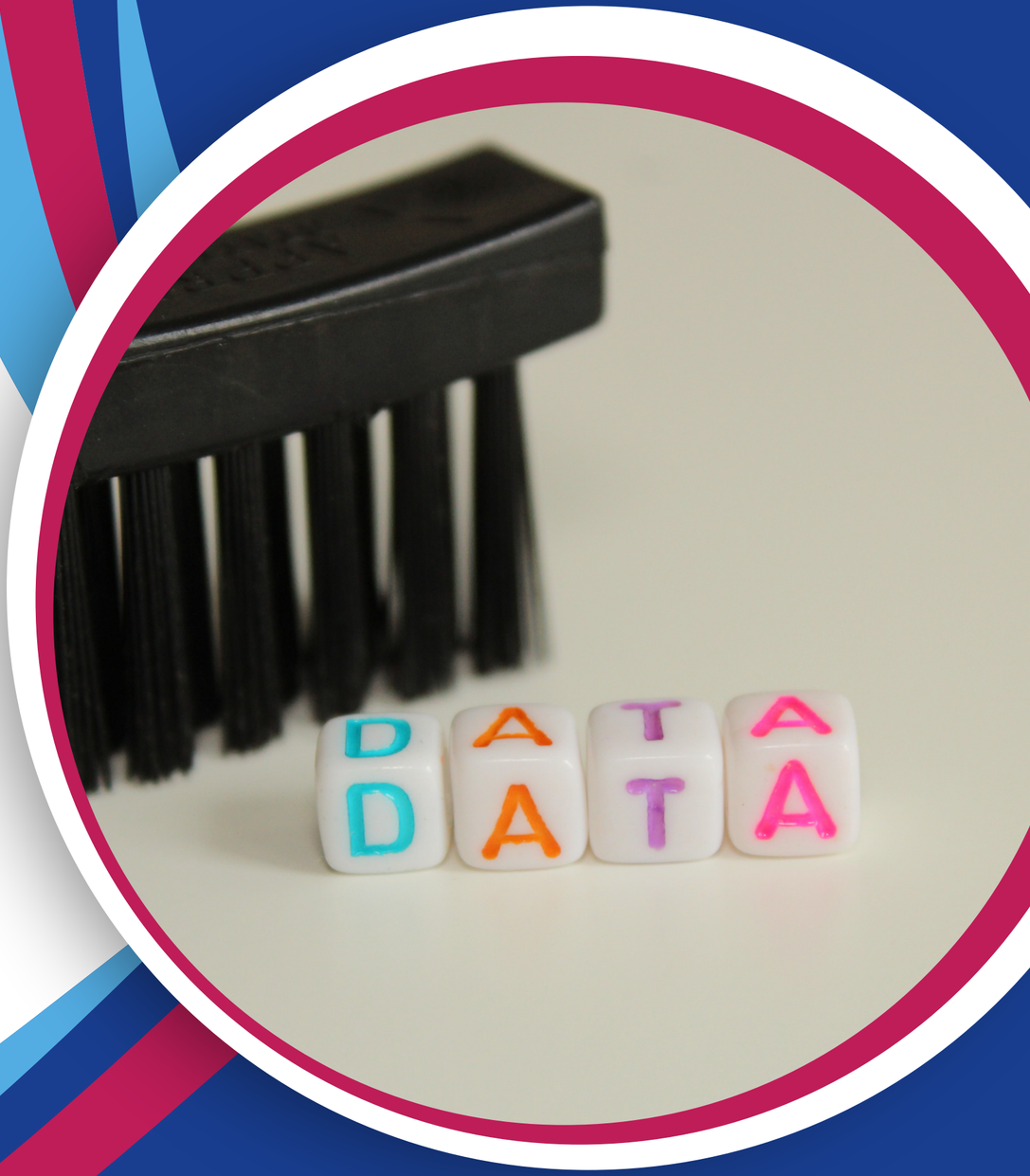
**Catégories de produits**



# Nettoyage des données

## Démarche méthodologique

- Selection des produits Français
- Identification et Traitement des Valeurs Aberrantes
- Identification et Traitement des Valeurs manquantes
- Imputation des Données Manquantes



# Nettoyage des données

## Rappel des 5 grands principes RGPD

- **Le principe de finalité**
- **Le principe de proportionnalité et de pertinence**
- **Le principe d'une durée de conservation limitée**
- **Le principe de sécurité et de confidentialité**
- **Les droits des personnes.**

Source :
https://www.cnil.fr/

# Nettoyage des données

## Selection des produits Français :

```
countries_fr = df_column_selected["countries_fr"]
countries_fr.sample(5)
```

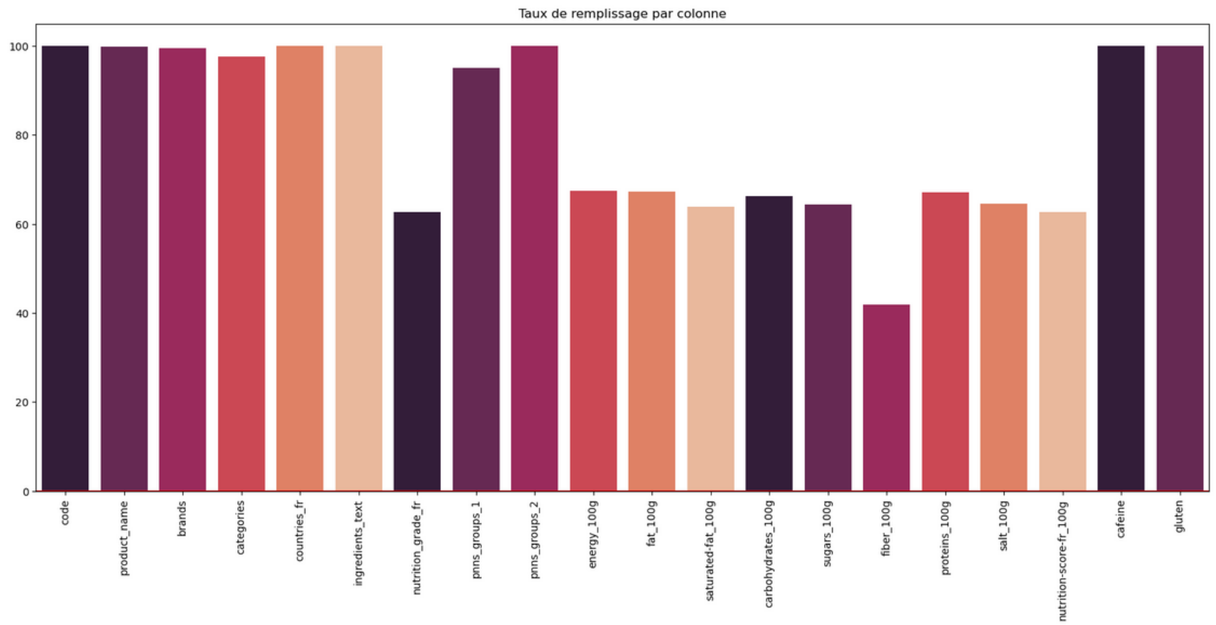
```
90041     États-Unis
247329        France
270898        Suisse
169367    États-Unis
188153        France
Name: countries_fr, dtype: object
```

```
countries_fr = df_column_selected["countries_fr"] == 'France'
df_countries_fr = df_column_selected[countries_fr]
df_countries_fr.sample(5)
```

# Nettoyage des données

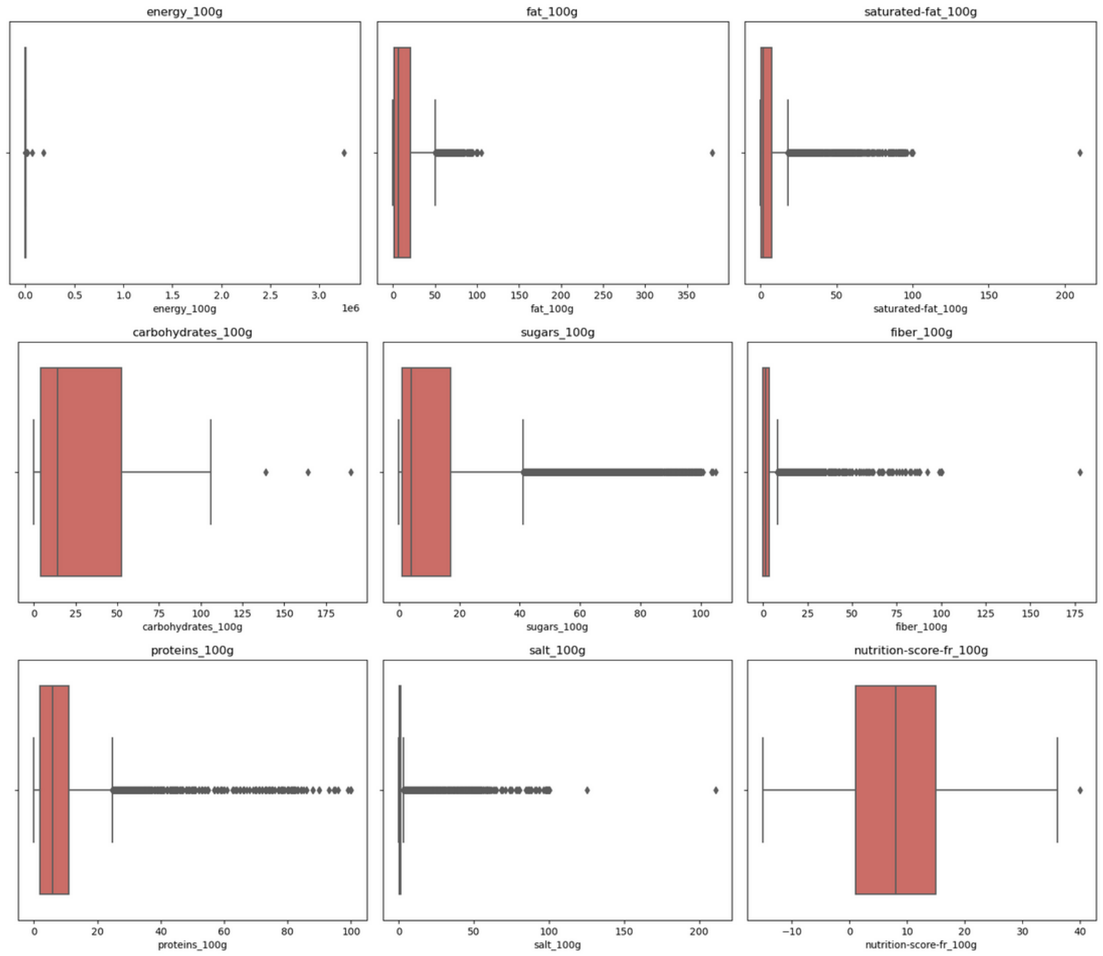
## Taux de remplissage après sélections des variables :

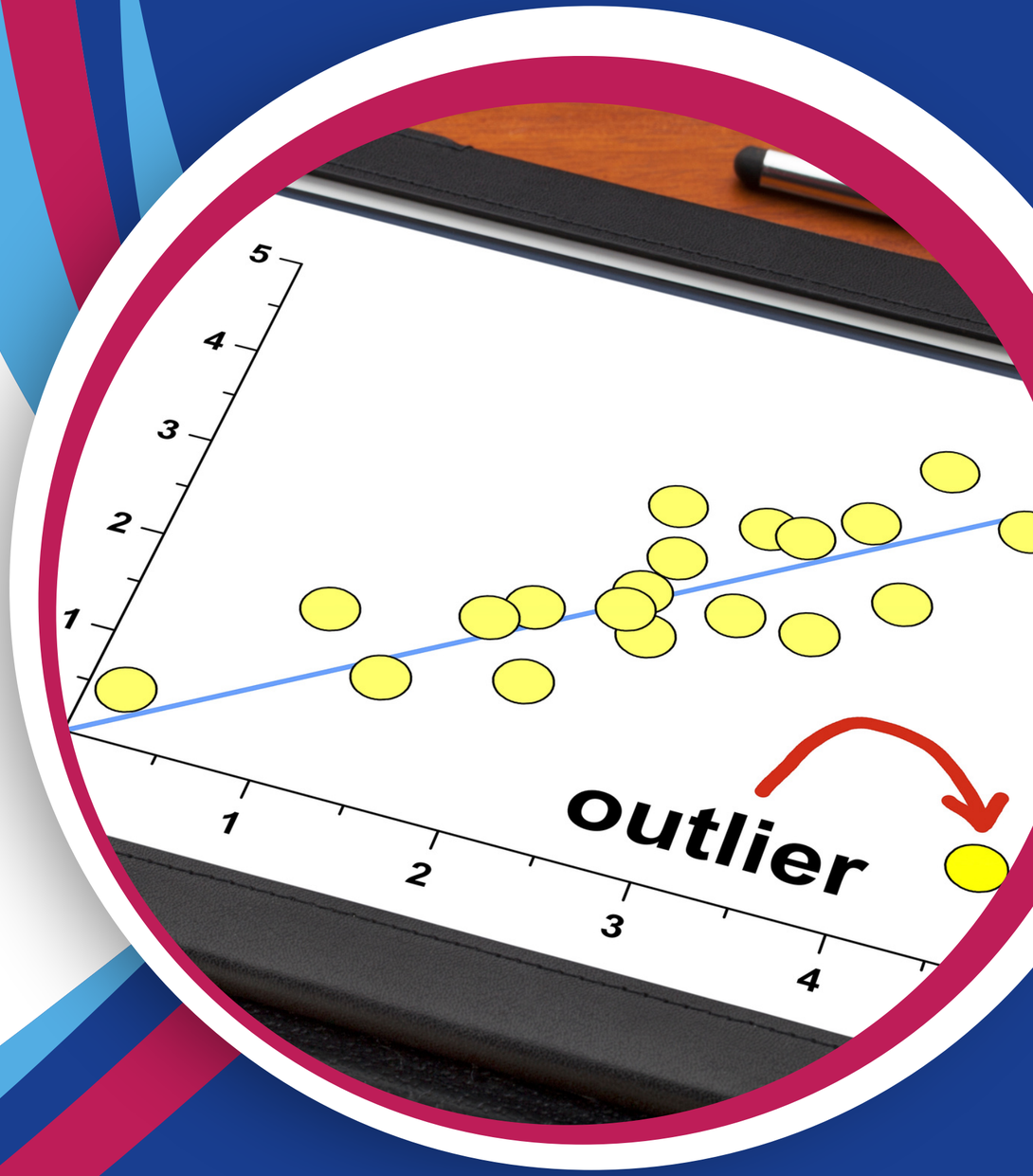
- Sélection des colonnes intéressantes
- Suppression des valeurs manquantes sur la variable ingredients_text



# Nettoyage des données

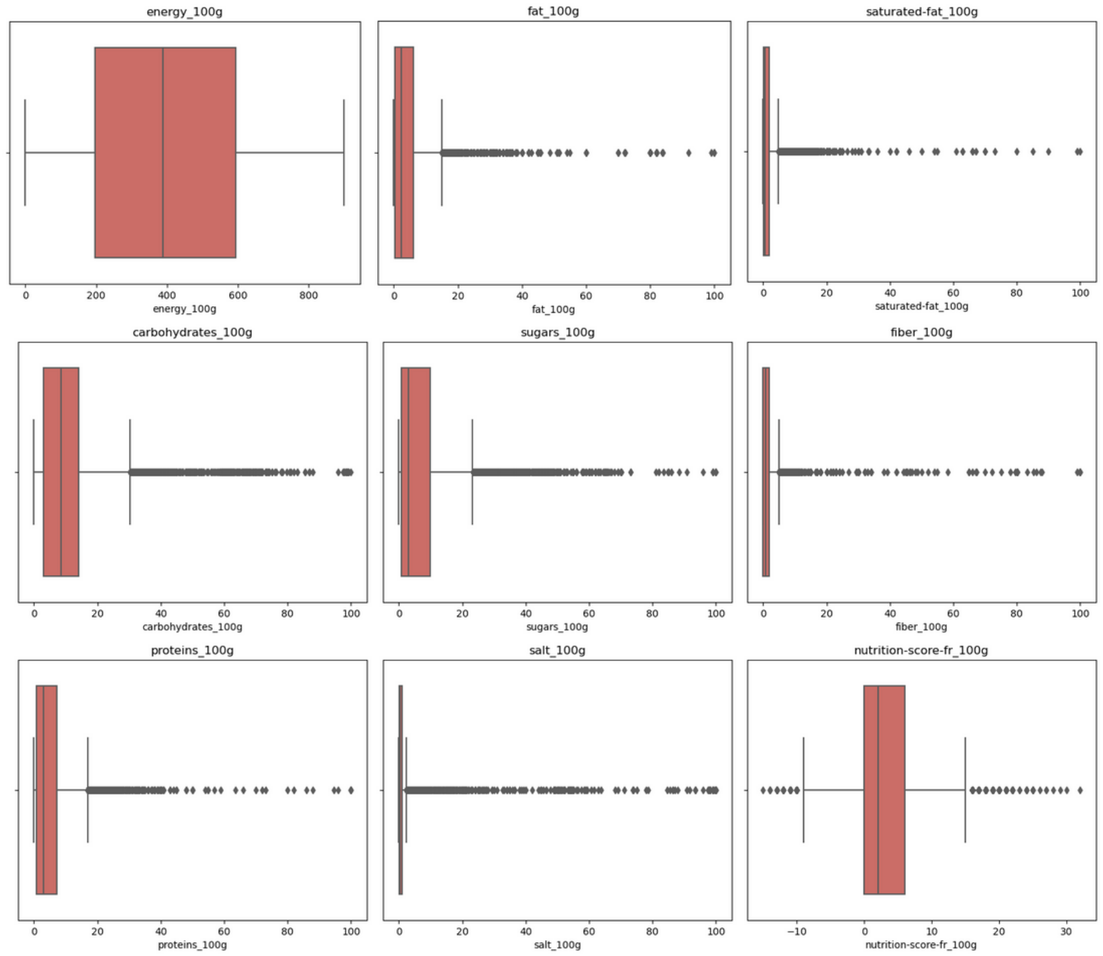
## Identification des Valeurs Aberrantes

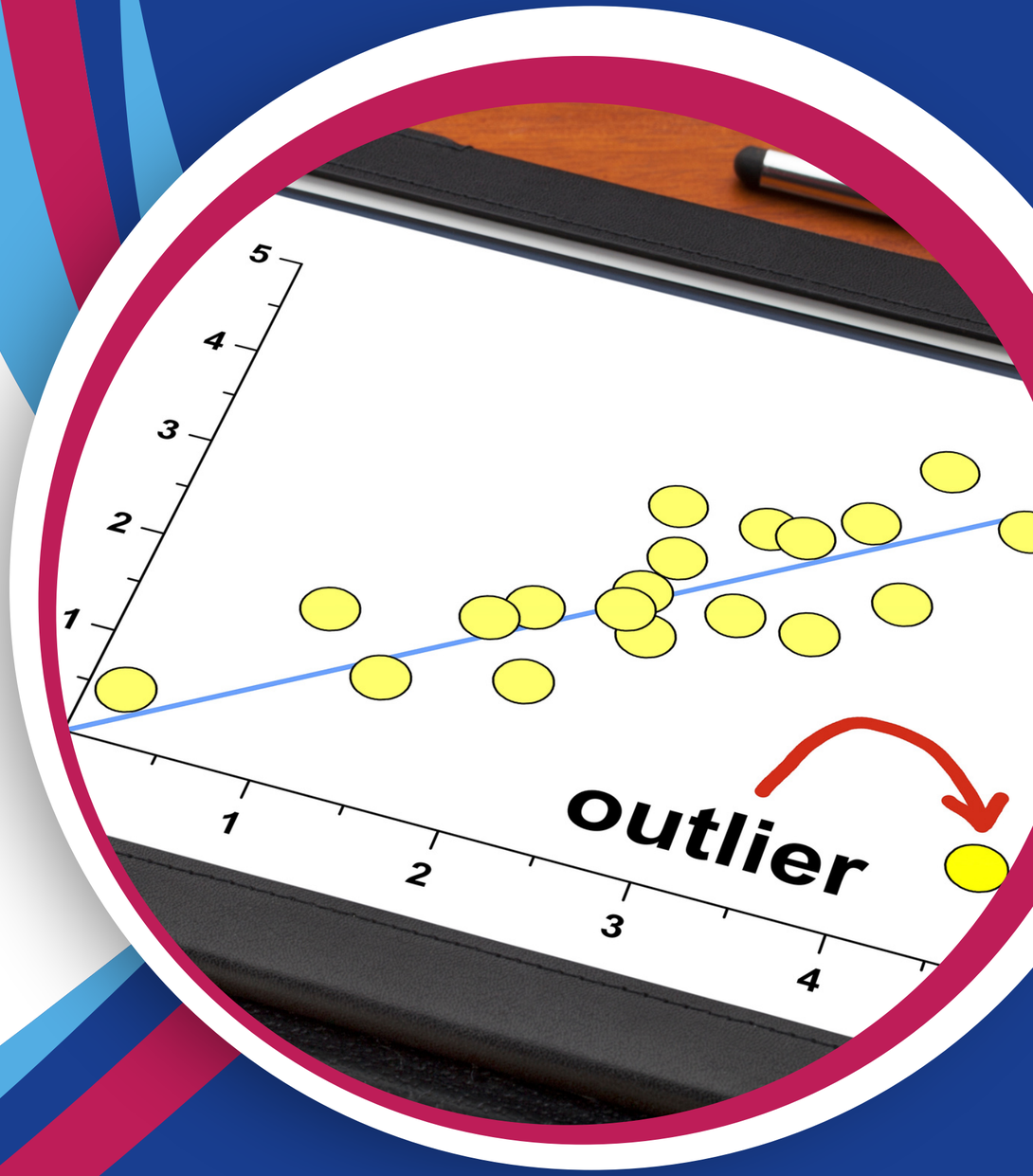


# Nettoyage des données

## Traitement des Valeurs Aberrantes



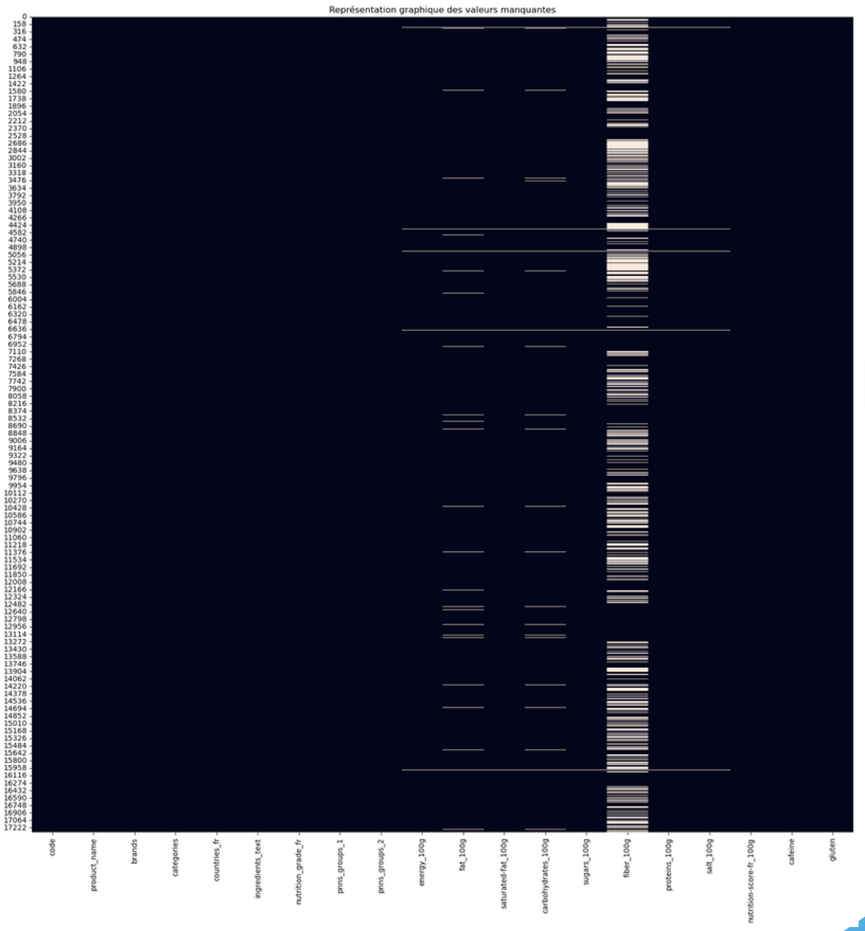



# Nettoyage des données

## Identification des Valeurs manquantes

Affichage en % des valeurs manquantes par colonnes

| | 0 | Nb Valeurs Manquantes |
|---|---|---|
| code | 0.00 | 0 |
| nutrition-score-fr_100g | 0.00 | 0 |
| cafeine | 0.00 | 0 |
| pnns_groups_2 | 0.00 | 0 |
| pnns_groups_1 | 0.00 | 0 |
| nutrition_grade_fr | 0.00 | 0 |
| gluten | 0.00 | 0 |
| countries_fr | 0.00 | 0 |
| categories | 0.00 | 0 |
| brands | 0.00 | 0 |
| product_name | 0.00 | 0 |
| ingredients_text | 0.00 | 0 |
| saturated-fat_100g | 0.30 | 52 |
| sugars_100g | 0.30 | 52 |
| proteins_100g | 0.30 | 52 |
| salt_100g | 0.30 | 52 |
| energy_100g | 0.30 | 52 |
| carbohydrates_100g | 1.82 | 315 |
| fat_100g | 1.96 | 339 |
| fiber_100g | 35.33 | 6118 |



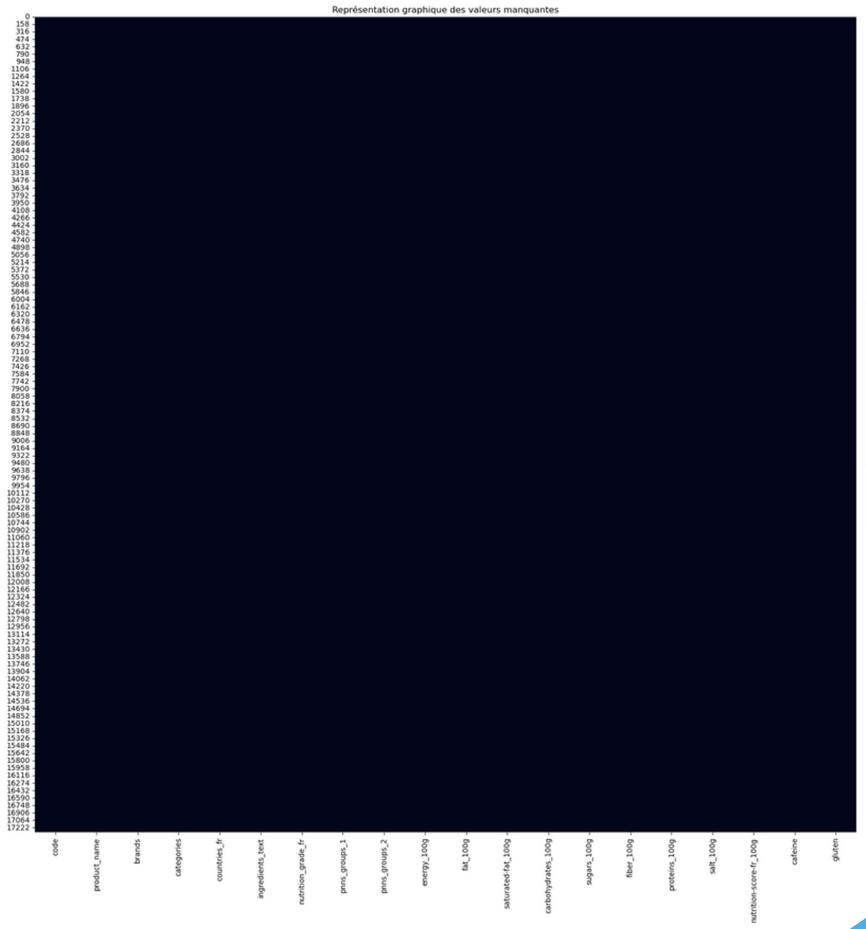
# Nettoyage des données

## Traitement des Valeurs manquantes

KNN IMPUTER

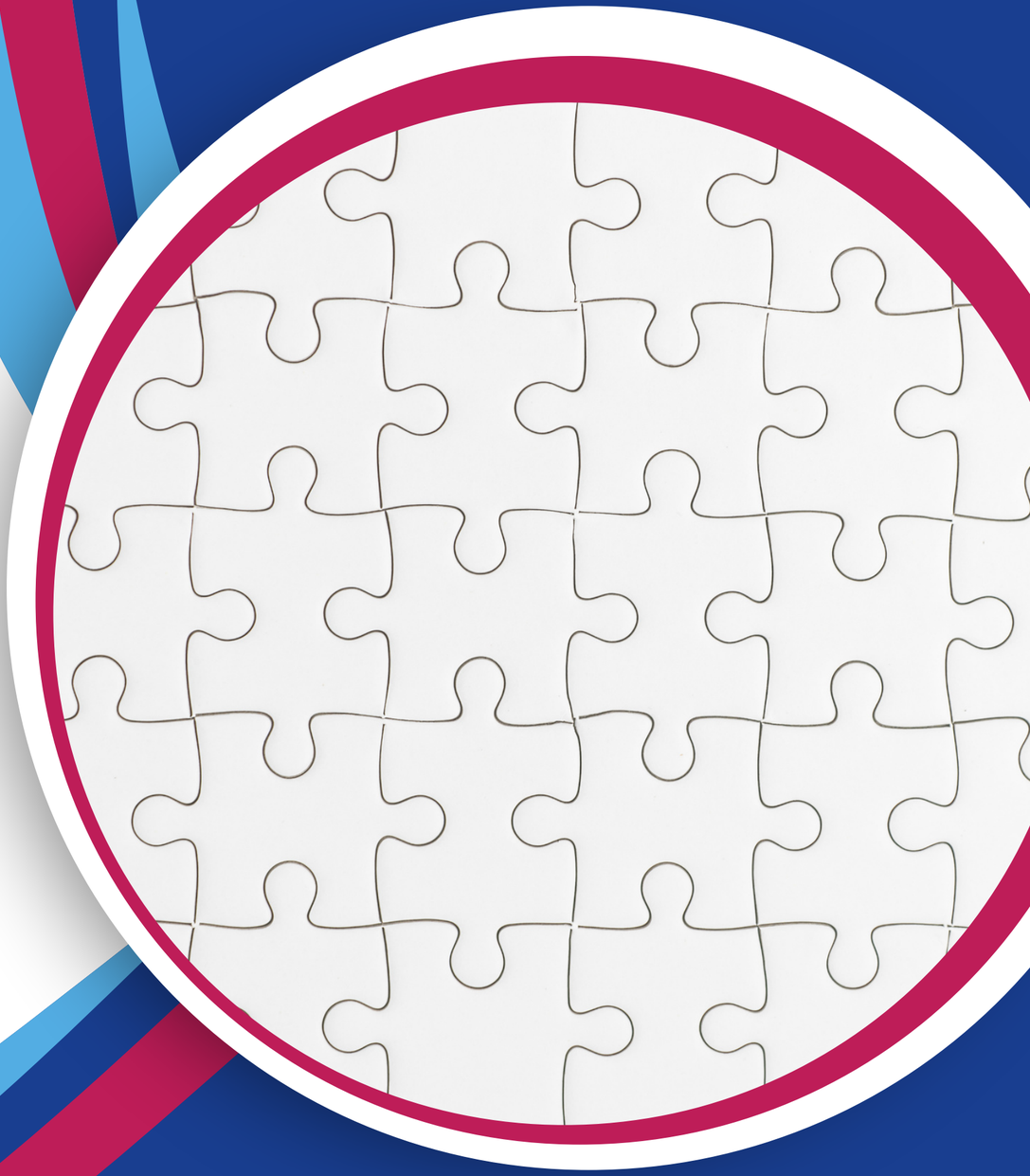
Affichage en % des valeurs manquantes par colonnes

| | 0 | Nb Valeurs Manquantes |
|---|---|---|
| code | 0.0 | 0 |
| nutrition-score-fr_100g | 0.0 | 0 |
| salt_100g | 0.0 | 0 |
| proteins_100g | 0.0 | 0 |
| fiber_100g | 0.0 | 0 |
| sugars_100g | 0.0 | 0 |
| carbohydrates_100g | 0.0 | 0 |
| saturated-fat_100g | 0.0 | 0 |
| fat_100g | 0.0 | 0 |
| energy_100g | 0.0 | 0 |
| pnns_groups_2 | 0.0 | 0 |
| pnns_groups_1 | 0.0 | 0 |
| nutrition_grade_fr | 0.0 | 0 |
| ingredients_text | 0.0 | 0 |
| countries_fr | 0.0 | 0 |
| categories | 0.0 | 0 |
| brands | 0.0 | 0 |
| product_name | 0.0 | 0 |
| cafeine | 0.0 | 0 |
| gluten | 0.0 | 0 |




20 variables

17 318 lignes



# Exploration des données

## Démarche méthodologique

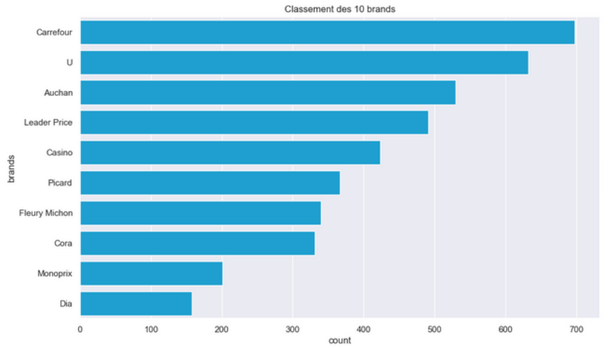
### Analyses Univariées
- classements qualitatives
- classements quantitatives

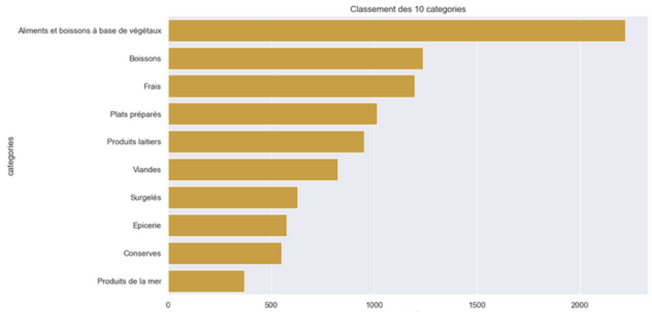
### Analyses Bivariées
- Quantitatives
- Quantitatives - Qualitatives
- Qualitatives

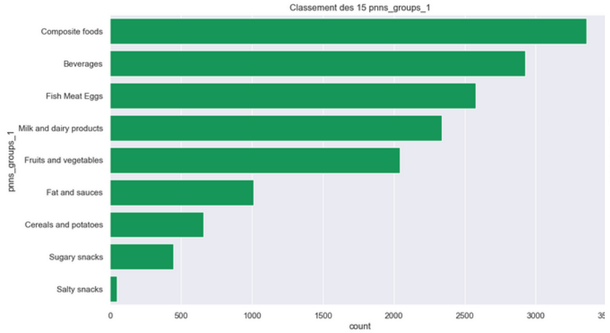
### Analyses Multivariées
- ACP
- Projection des individus

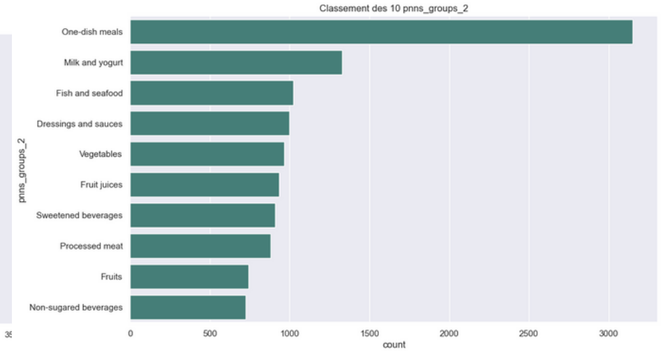
# Exploration des données

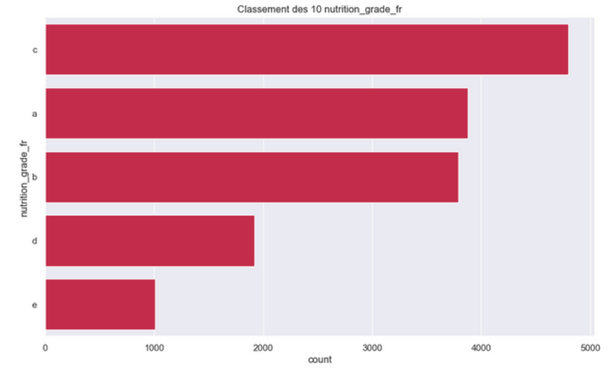
## Analyse Univariées Qualitatives



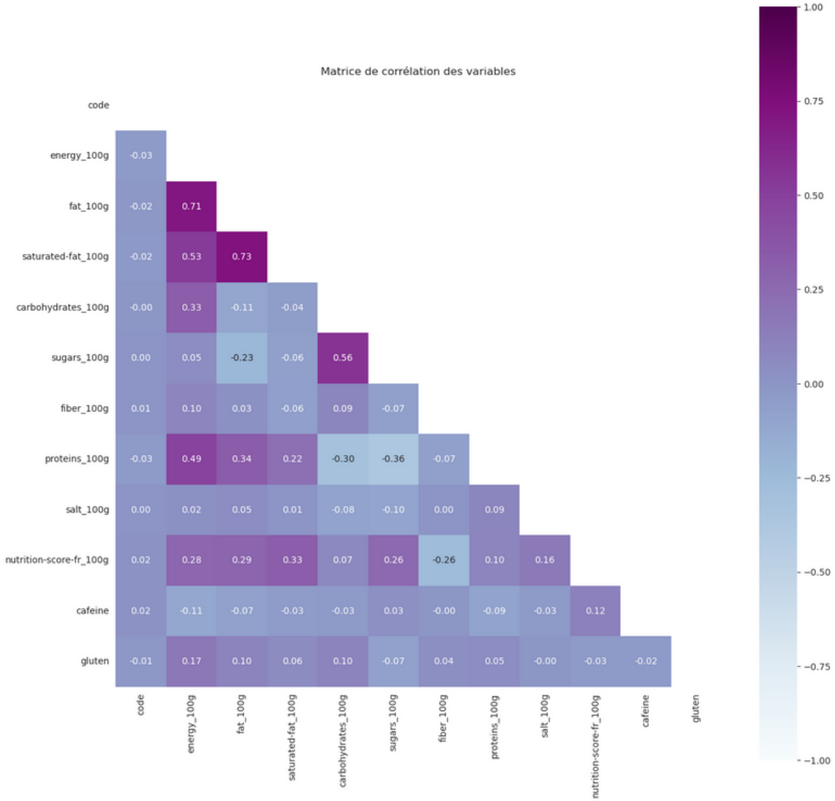



# Exploration des données

## Analyse Bivariée Quantitatives




### Corrélation positive de 0,71 entre energy_100g et fat_100g.

*Cette forte corrélation est cohérente avec le fait que les graisses contribuent de manière significative à l'apport énergétique des produits alimentaires.*

### Corrélation positive de 0,73 entre fat_100g et saturated-fat_100g.

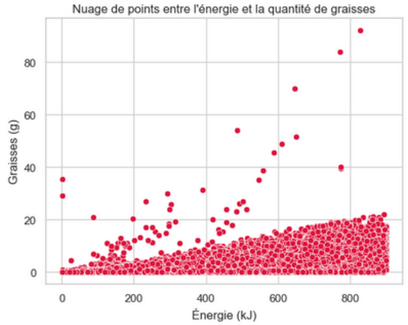
*Cette observation est également logique, car les graisses saturées sont une catégorie spécifique de graisses et leur présence est étroitement liée à la teneur totale en matières grasses.*

### Corrélation positive de 0,56, entre sugars_100g et carbohydrates_100g.

*Cette corrélation reflète la relation intrinsèque entre les sucres et les glucides, car les sucres sont une composante des glucides totaux.*

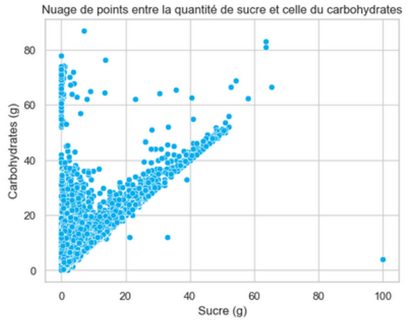
# Exploration des données

## Analyse Bivariée Quantitatives



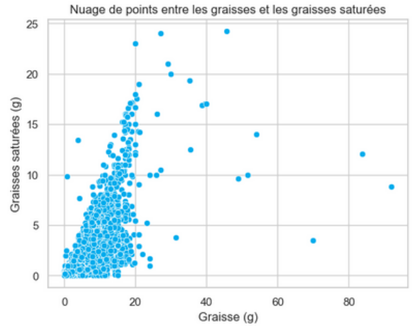

P-value = 2.115779e-263

*La valeur p est faible (inférieure à 0,05) Ce qui indique une forte preuve contre l'hypothèse nulle, ce qui suggère que les résultats sont statistiquement significatifs.*




P-value = 0.0

*La valeur p est faible (inférieure à 0,05) Ce qui indique une forte preuve contre l'hypothèse nulle, ce qui suggère que les résultats sont statistiquement significatifs.*

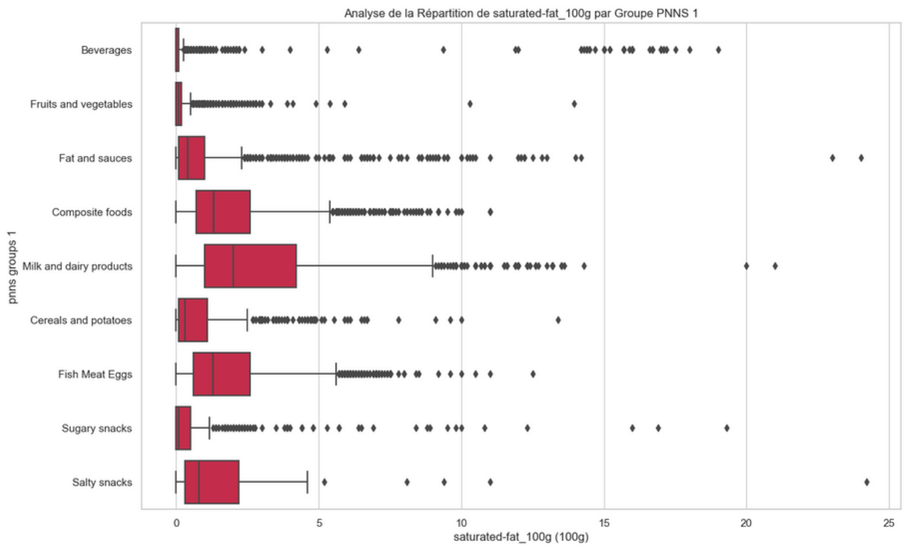


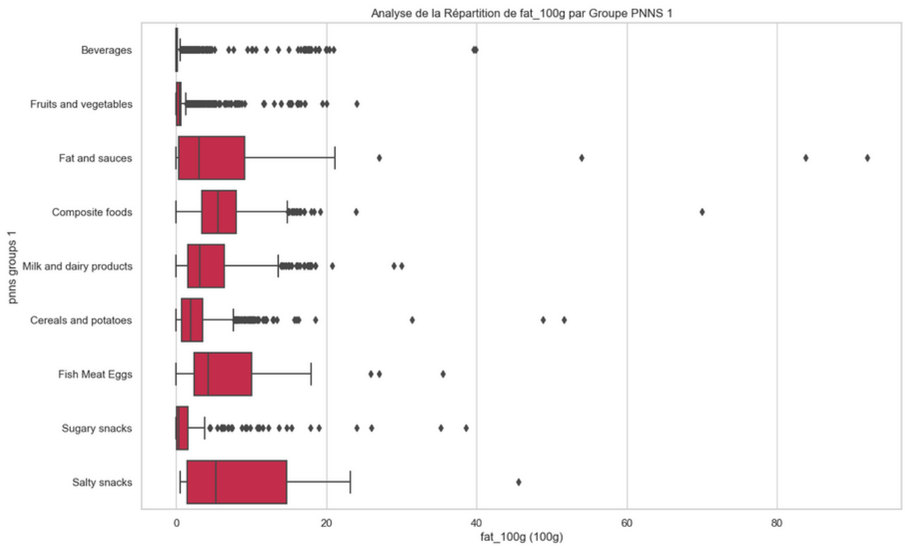

P-value = 0.00001

*La valeur p est faible (inférieure à 0,05) Ce qui indique une forte preuve contre l'hypothèse nulle, ce qui suggère que les résultats sont statistiquement significatifs.*

# Exploration des données

## Analyse Bivariée Quantitatives - Qualitatives

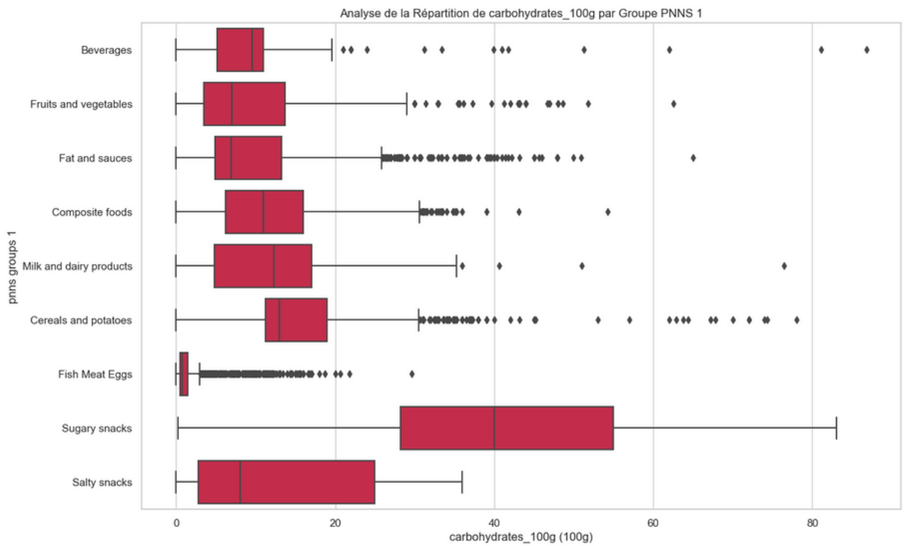


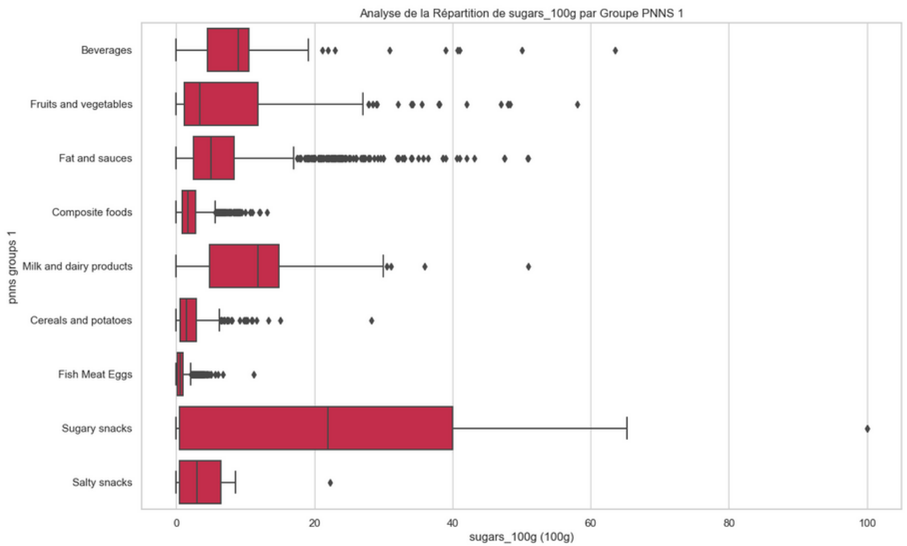




*Les **Salty snacks** présentent la teneur en graisse la plus élevée, tandis que les **Boissons** affichent la teneur en graise la plus basse.*

# Exploration des données

## Analyse Bivariée Quantitatives - Qualitatives

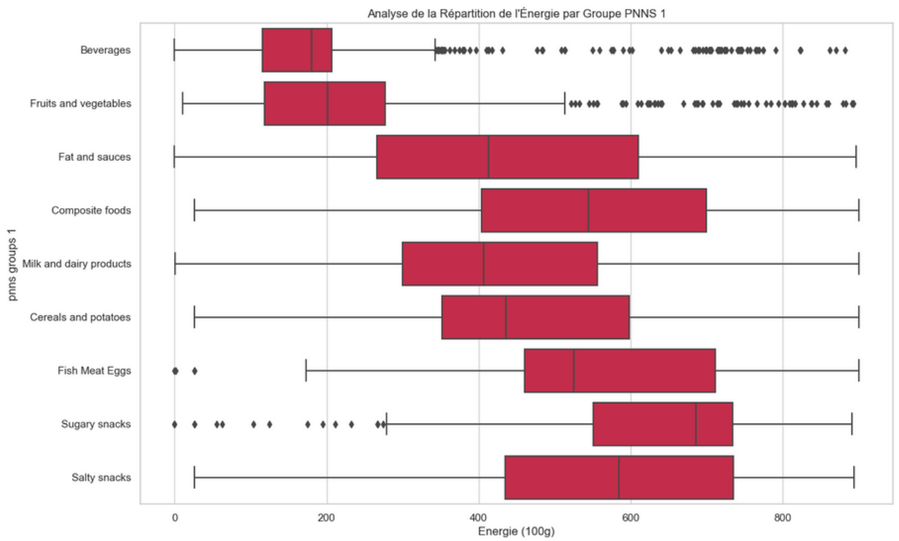


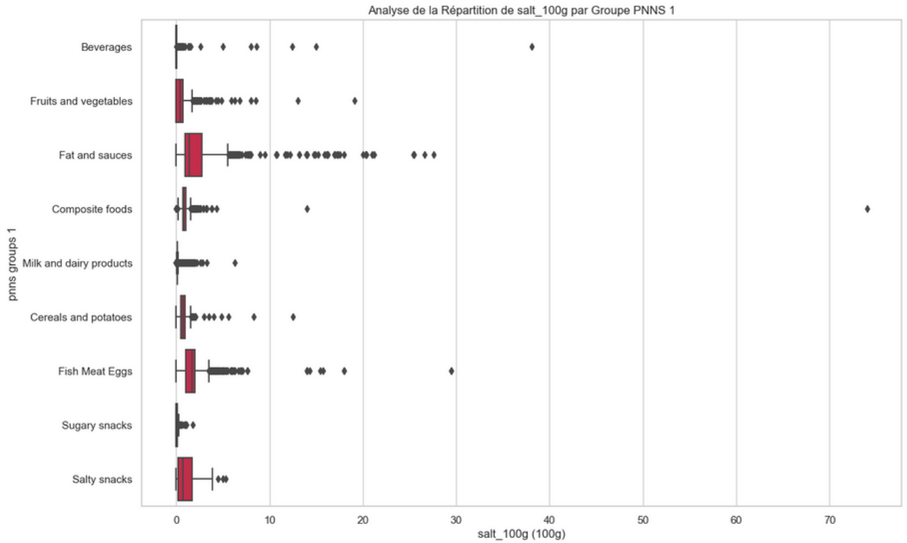




*Les **sugary snacks** présentent la teneur en sucres la plus élevée, tandis que les **fish meat eggs** affichent la teneur en sucres la plus basse.*

# Exploration des données

## Analyse Bivariée Quantitatives - Qualitatives



*Les produits **Fat and sauces** et **Salty snacks** affichent la plus forte teneur en sel, tandis que les **sugary-snacks** présentent une faible teneur en sel, ce qui est cohérent.*

# Exploration des données

## Analyse Bivariée Quantitatives - Qualitatives

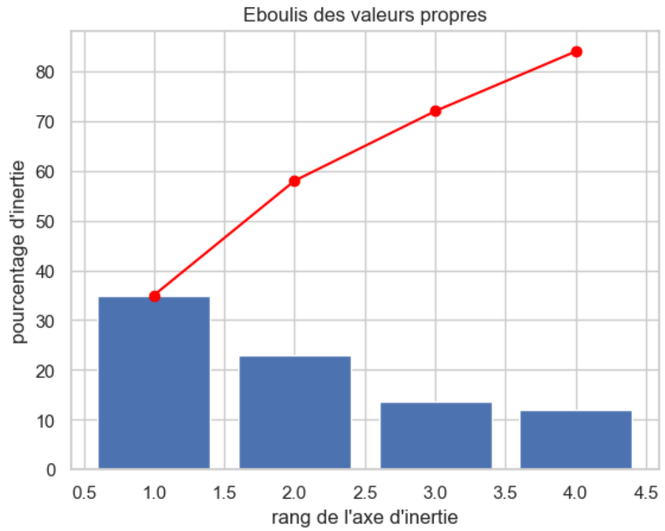
```
Analyse ANOVA pour la variable 'fat_100g':
          Source  ddof1  ddof2          F  p-unc       np2
0  pnns_groups_1      8  15399  720.46415    0.0  0.272352
Analyse ANOVA pour la variable 'saturated-fat_100g':
          Source  ddof1  ddof2          F  p-unc       np2
0  pnns_groups_1      8  15399  475.96873    0.0  0.198251
Analyse ANOVA pour la variable 'carbohydrates_100g':
          Source  ddof1  ddof2            F  p-unc       np2
0  pnns_groups_1      8  15399  1524.512648    0.0  0.441966
Analyse ANOVA pour la variable 'sugars_100g':
          Source  ddof1  ddof2            F  p-unc       np2
0  pnns_groups_1      8  15399  1282.109014    0.0  0.399787
Analyse ANOVA pour la variable 'fiber_100g':
          Source  ddof1  ddof2          F  p-unc       np2
0  pnns_groups_1      8  15399  361.712385    0.0  0.158189
Analyse ANOVA pour la variable 'proteins_100g':
          Source  ddof1  ddof2          F  p-unc       np2
0  pnns_groups_1      8  15399  6061.97847    0.0  0.758995
Analyse ANOVA pour la variable 'salt_100g':
          Source  ddof1  ddof2          F  p-unc       np2
0  pnns_groups_1      8  15399  600.20249    0.0  0.237697
```

Les résultats de l'analyse ANOVA montrent que la variable pnns_groups_1 est significative pour toutes les variables que j'ai testées. Le p-value (p-unc) est inférieur à 0.05 pour chacune des variables, ce qui indique que la variable pnns_groups_1 a un effet significatif sur les variables testées. De plus, le coefficient d'étadétermination partiel (np2) est assez élevé pour la plupart des variables, ce qui signifie que la variation de pnns_groups_1 explique une proportion importante de la variance des variables.

*La variable pnns_groups_1 est significative et a un impact important sur les variables testées. Elle joue un rôle significatif dans l'explication de la variance de ces variables.*

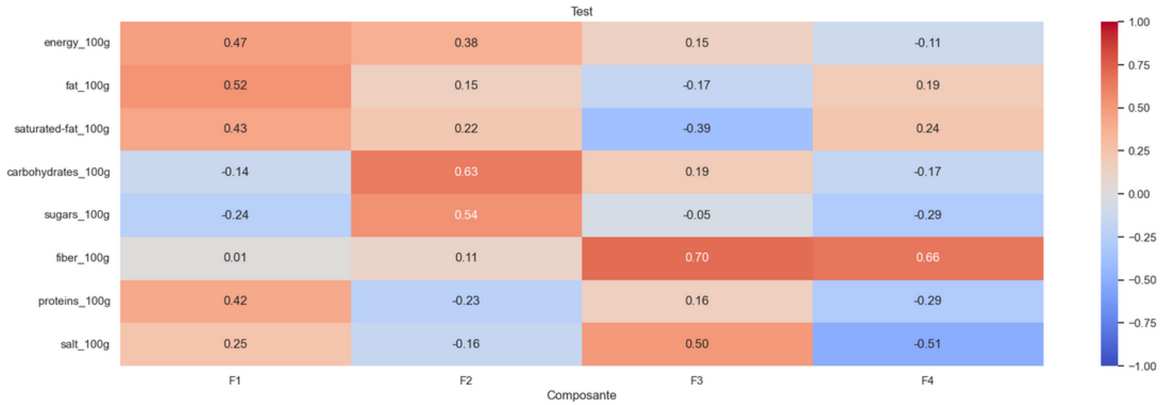
# Exploration des données

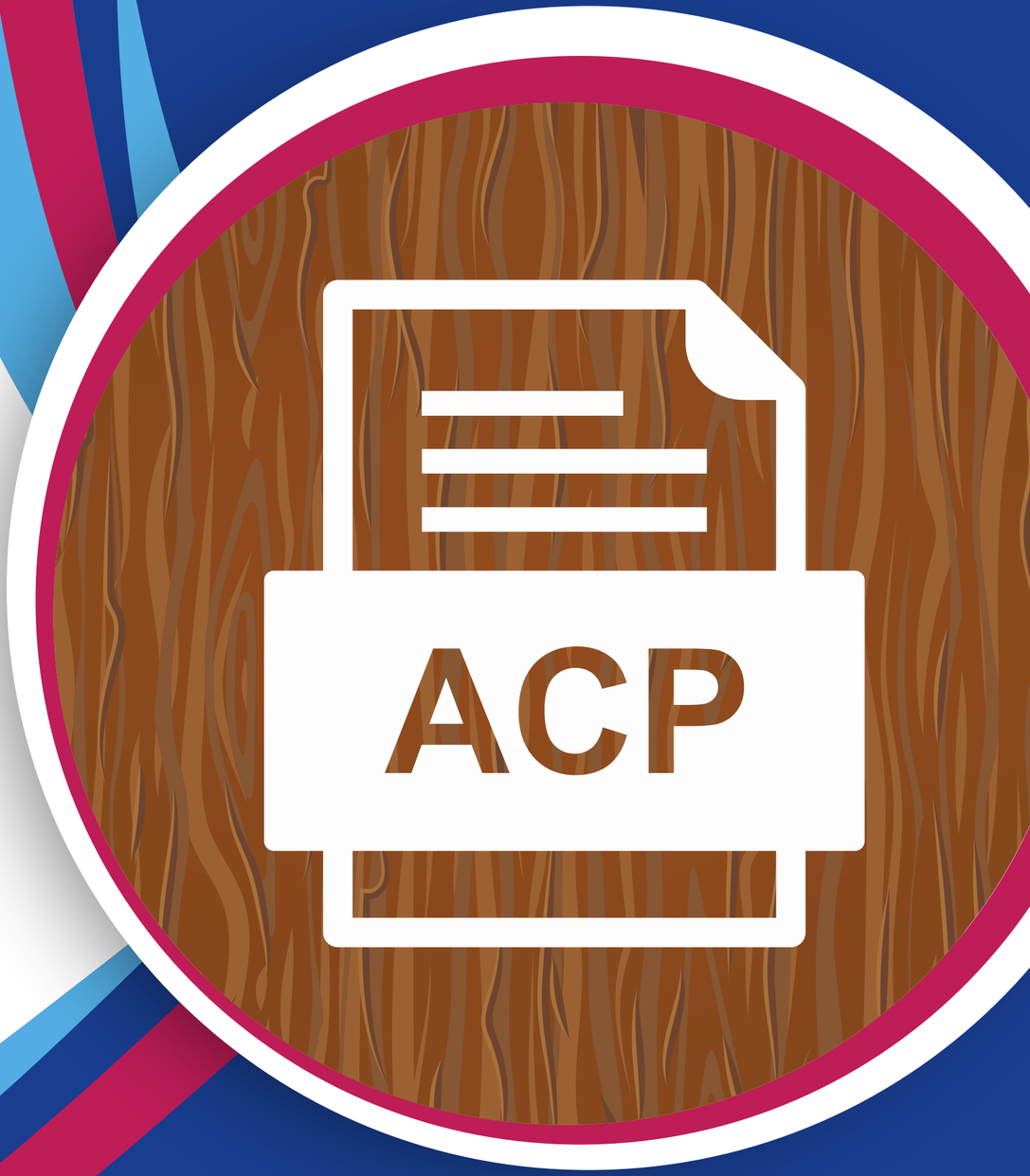
## Analyse Multivariées




F1 est associé à 35% de l'inertie total

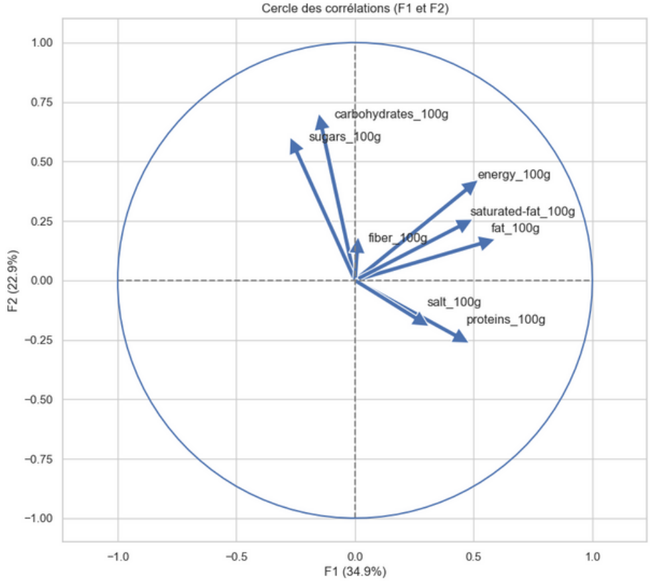
F2 est associé à 23% de l'inertie total



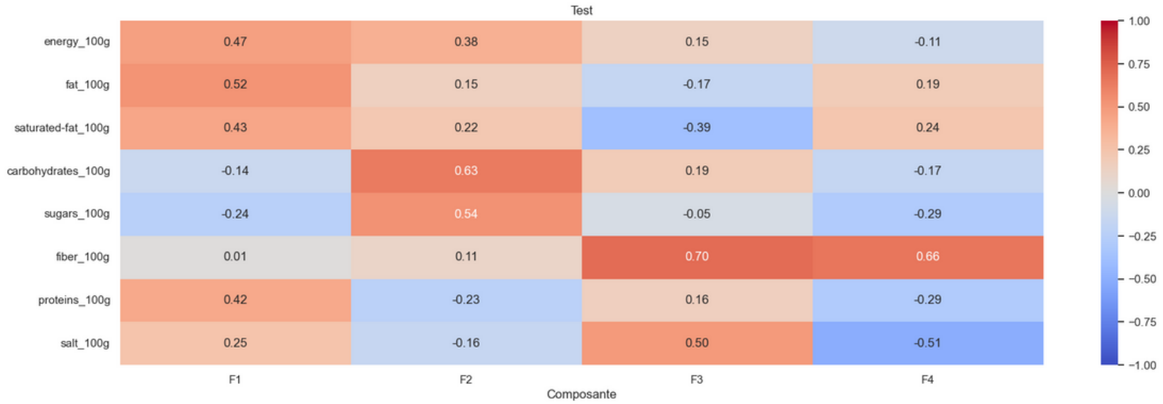
# Exploration des données

## Analyse Multivariées

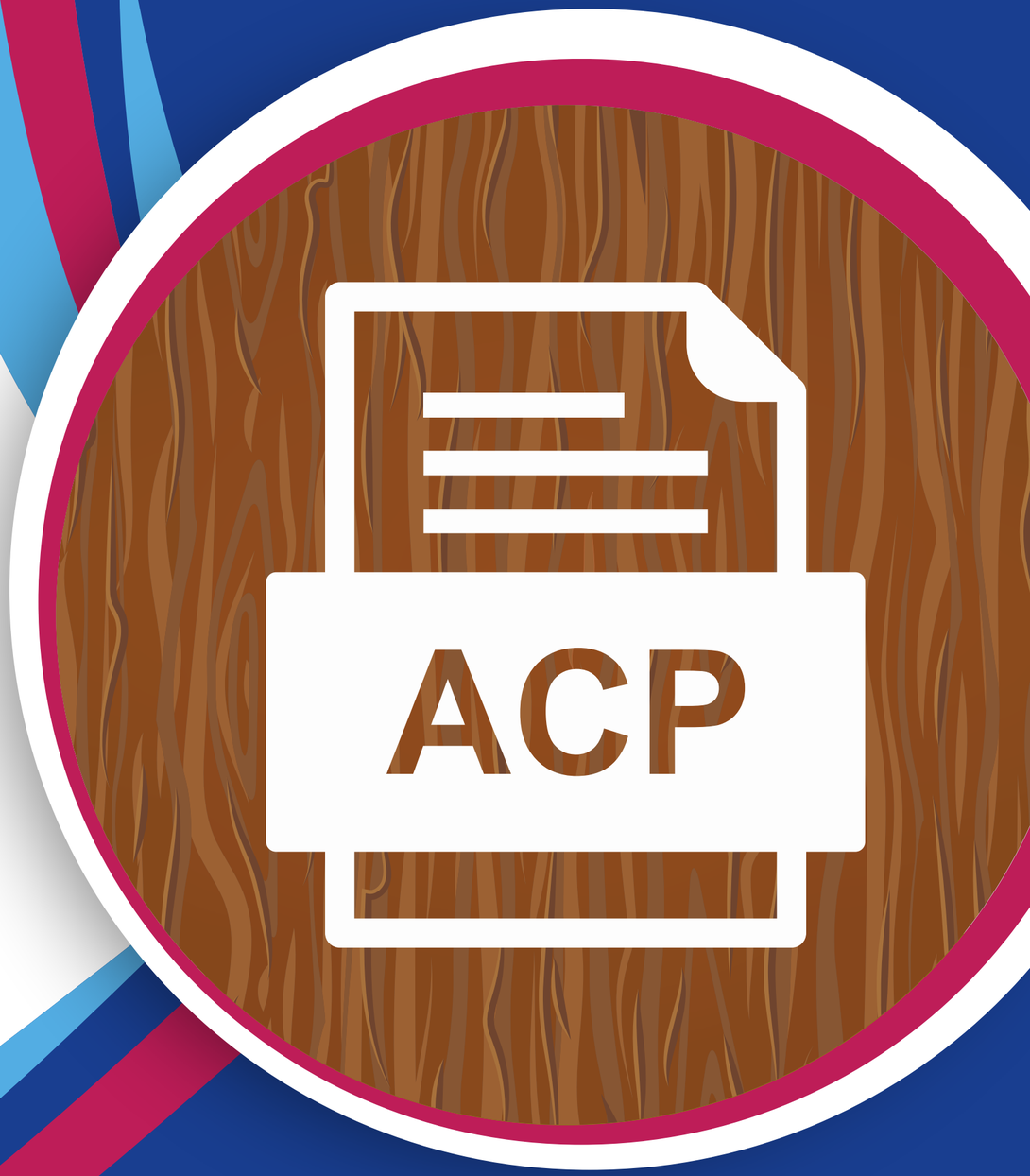



F1 Corrélée positivement avec les Énergie, les graisses, et les proteins

F2 Corrélée positivement avec les sucres et les carbohydrates




| | F1 | F2 | F3 | F4 |
|---|---|---|---|---|
| energy_100g | 0.47 | 0.38 | 0.15 | -0.11 |
| fat_100g | 0.52 | 0.15 | -0.17 | 0.19 |
| saturated-fat_100g | 0.43 | 0.22 | -0.39 | 0.24 |
| carbohydrates_100g | -0.14 | 0.63 | 0.19 | -0.17 |
| sugars_100g | -0.24 | 0.54 | -0.05 | -0.29 |
| fiber_100g | 0.01 | 0.11 | 0.70 | 0.66 |
| proteins_100g | 0.42 | -0.23 | 0.16 | -0.29 |
| salt_100g | 0.25 | -0.16 | 0.50 | -0.51 |



# Exploration des données

## Analyse Bivariée Quantitatives - Qualitatives

F1 :

- Composite food qui est surtout composé de viande,charcuterie, poissons ansi que milk and dairy product = lipide et energie
- Fish meat eggs qui se trouve au niveau des proteines.

F2 : Sugary snack et Les fruits = carbohydrates et sucre

# Conclusion

## L’exploration

L'exploration approfondie de la variable "pnns_groups_1" a révélé son potentiel pour répondre à la problématique.
Cette variable, qui classe les aliments en catégories nutritionnelles, sa structure catégorielle et la variabilité de ses classes en font une option intéressante pour générer des suggestions

## Analyse uni et bivariées

L'analyse univariée a démontré une répartition hétérogène des données dans différentes catégories alimentaires. Les méthodes de traitement des valeurs manquantes ont été appliquées avec succès, permettant de comprendre et de remplir ces lacunes.

## Tests Multivariées

Les tests multivariés ont mis en évidence des associations et des corrélations entre cette variable et d'autres caractéristiques nutritionnelles, offrant ainsi une possibilité de prédiction ou de suggestion de valeurs manquantes sur la base de ces relations.

L'Approche de l'Analyse en Composantes Principales (ACP) a été utilisée pour mieux appréhender la structure des données nutritionnelles. La projection des individus dans l'espace ACP a mis en évidence la contribution significative de la variable "pnns_groups_1" dans la variance globale des données. Cela suggère que cette variable est pertinente et impactante dans la représentation des caractéristiques nutritionnelles des produits alimentaires.

Santé publique France

# Merci

## Pour votre attention

Visitez notre site
https://www.santepubliquefrance.fr/