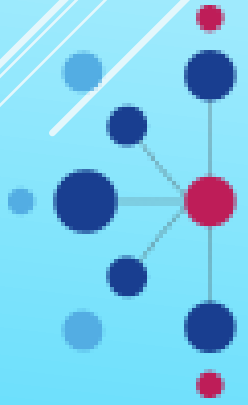
Préparez des données pour un organisme de santé publique

OPENCLASSROOMS



Santé publique France

ENJEUX

Améliorer la base de données Open Food Facts

création d'un système de suggestion ou d'auto-complétion pour aider les usagers à remplir plus efficacement la base de données

SOMMAIRE

NETTOYAGE & FILTRAGE DES FEATURES & PRODUITS

OPEN FOOD FACT

- **320772 produits répertoriés**

→ 320749 après traitement identifiant

- **162 features**

Variables quantiatives = 106

Variables qualitatives = 56

- **76% valeurs manquantes**

CRITERES DE FILTRAGE :

Identifiant :

- Taux de remplissage = 100% après traitement
- Nombre de valeurs uniques = nombre de produits répertoriés

Variable Cible :

- Variable qualitative
- Taux de remplissage < 50%
- Nombre de valeurs uniques traitable

CHOIX DE LA VARIABLE CIBLE & DE L'IDENTIFIANT

Nom	Tauxde remplissage(%)	Valeurs uniques
code	100,00%	**320749**
url	100,00%	**320749**
product_name	94,46%	**221347**
ingredients_text	77,61%	**205520**
additives	77,60%	**196069**
...		
ingredients_that_may_be_from_palm_oil_tags	3,65%	**160**
pnns_groups_2	29,46%	**42**
ingredients_from_palm_oil_tags	77,61%	**14**
pnns_groups_1	28,53%	**14**
nutrition_grade_fr	68,97%	**5**

IDENTIFIANT

- **code**
- **320749 valeurs uniques après traitement**
- **Taux de remplissage = 100%**

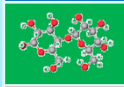
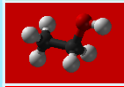
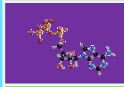
VARIABLE CIBLE

- **pnns_groups_1**
- **14 valeurs**

→ 9 après nettoyage

- **71% valeurs manquantes**
- **pnns_groups_2 = variable cible complémentaire**

SELECTION DES FEATURES

Nom	Taux de remplissage
 Sodium	79,65%
 Salt	79,66%
 Fiber	62,63%
 Alcohol	1,29%
 Proteins	81,03%
 Fat	76,04%
Saturated Fat	71,57%
 Carbohydrates	75,94%
 Sugars	76,37%
 Energy	81,41%
 Nutrition Score FR	68,97%

CRITERES DE FILTRAGE :

- Variables quantitatives
- Approche métier
- Taux de remplissage > 50%

NOUVEAU SUPPORT :

- **55896 produits répertoriés**
- **14 features**

Variables quantiatives = 11

Variables qualitatives = 3

- **14% valeurs manquantes**

IDENTIFICATION & TRAITEMENT DES VALEURS ABERRANTES

TABLEAU DES ECARTS INTERQUARTILES PAR GROUPE ALIMENTAIRE (SEUIL MAXIMAL)

pnns_group_1	Fat_100g	Carbohydrates_100g	Proteins_100g
Sugary snacks	65.1	92.5	12.7
Milk and dairy products	58.0	32.75	37.7
Cereals and potatoes	15.25	117.0	19.0
Beverages	1.25	20.7	1.5
Composite foods	19.3	44.3	17.95
Fish Meat Eggs	44.75	4.25	34.0
Fruits and vegetables	1.95	32.4	5.1
Fat and sauces	195.2	22.6	4.0
Salty snacks	59.8	101.4	25.65

IDENTIFICATION & TRAITEMENT DES VALEURS ABERRANTES

METHODE INTERQUARTILE :

- Inégalités significatives des moyennes/médianes/écarts-types par « pnns_groups_1 »
- Nécessité d'appliquer la méthode interquartile par « pnns_groups_2 »

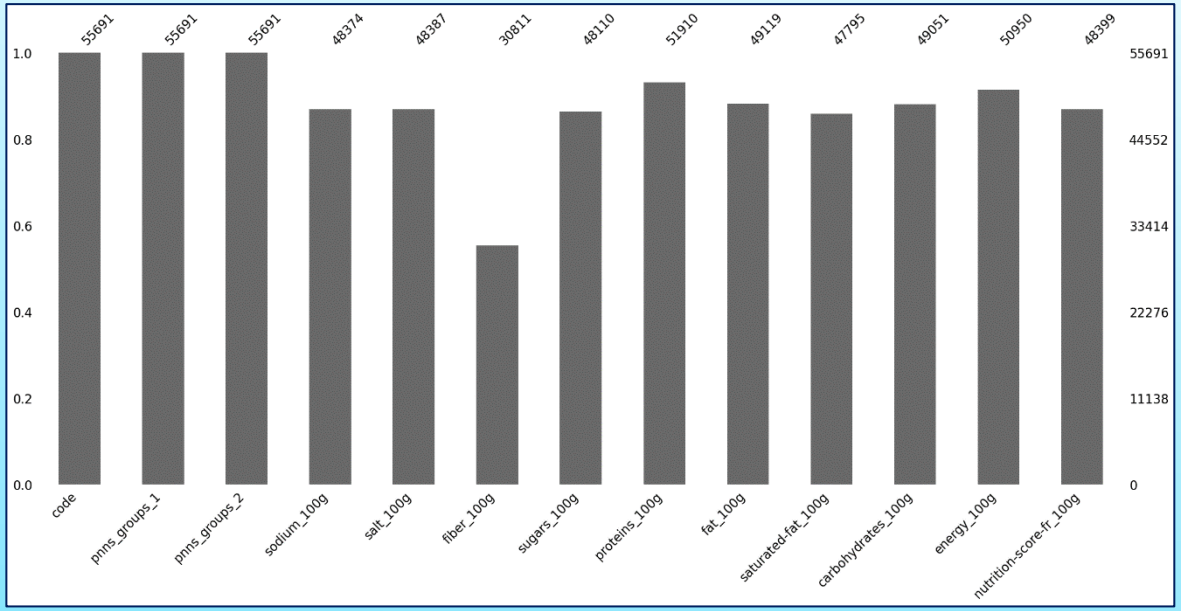
pnns_group_2	Fat	Carbohydrates	Proteins
Appetizers	52.75	80.0	14.7
Nuts	67.82	45.15	38.95
Salty and fatty products	34.24	63.07	12.34

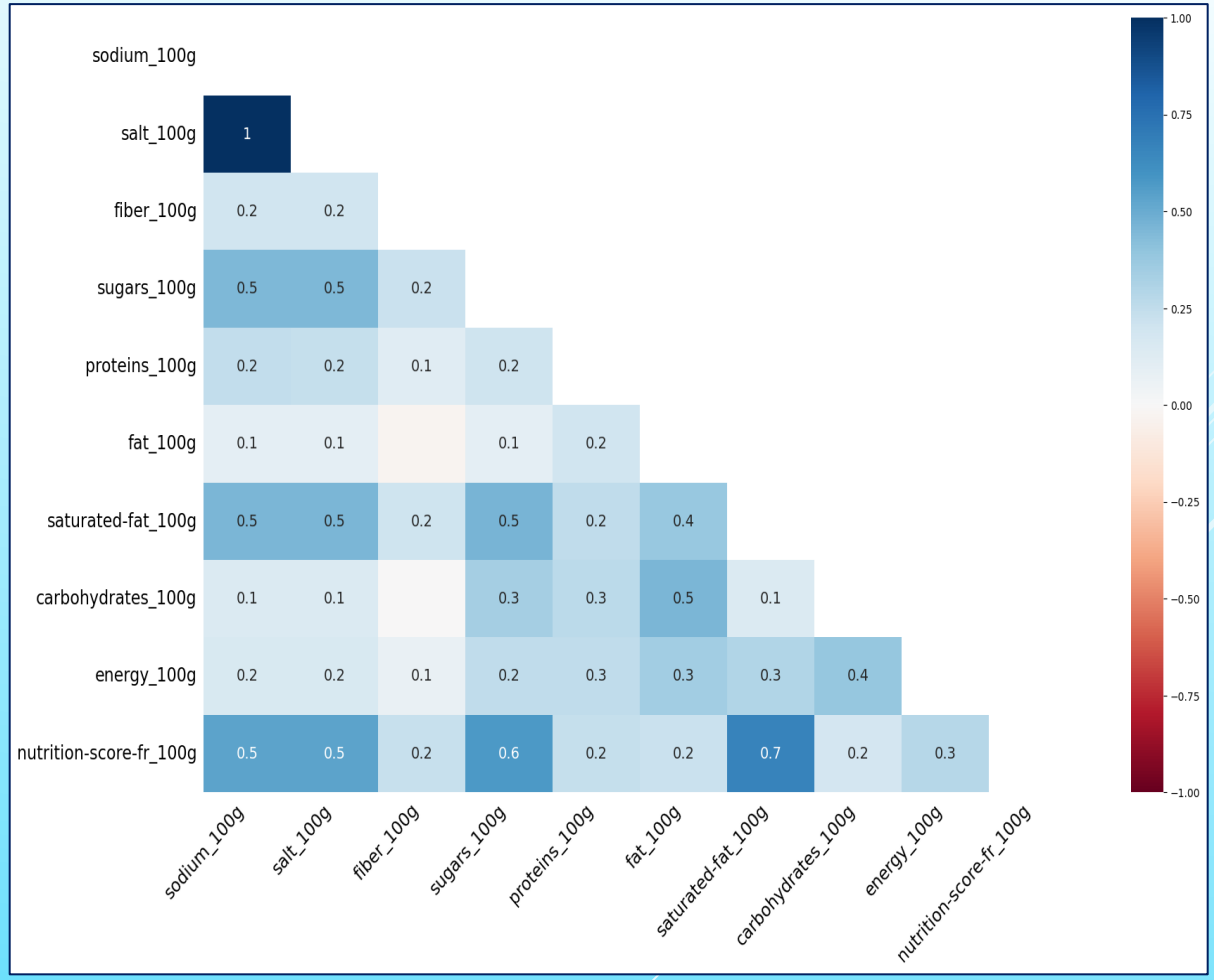
TABLEAU DES ECARTS INTERQUARTILES PAR SOUS-GROUPE ALIMENTAIRE DE « SALTY SNACKS » (SEUIL MAXIMAL)

APPROCHE METIER :

- Valeurs nutritionnelles <= 100
- Energy = (proteins x 17) + (fat x 37) + (carbohydrates x 17) + (alcohol x 29)
- Salt > Sodium
- Carbohydrates > Sugars
- Fat > Saturated Fat

IDENTIFICATION DES VALEURS MANQUANTES





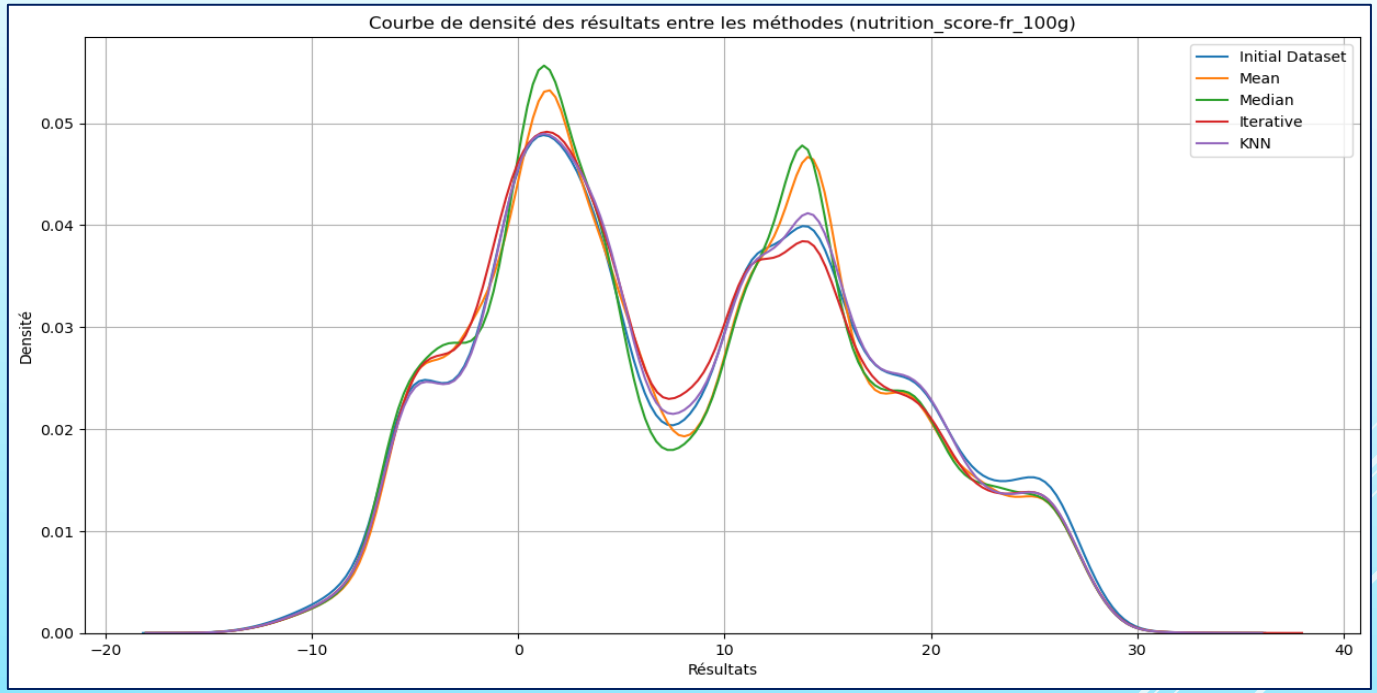
- **TAUX DE REMPLISSAGE DES VARIABLES POST-NETTOYAGE (ci-dessus)**

- **MATRICE DES CORRELATIONS DES VALEURS MANQUANTES (à droite)**

TRAITEMENT DES VALEURS MANQUANTES

IMPUTATION PAR 4 METHODES :

- Moyenne
- Médiane
- Itération
- KNN



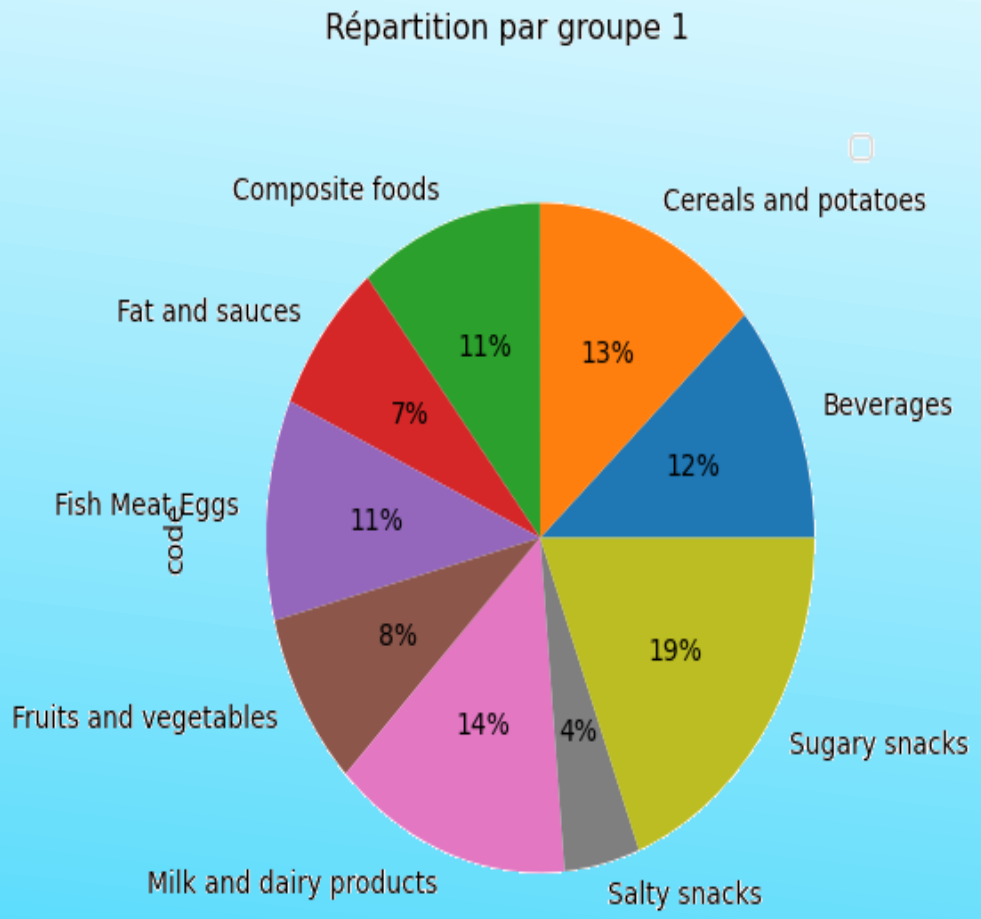
TEST KOLMOGOROV-SMIRNOV :

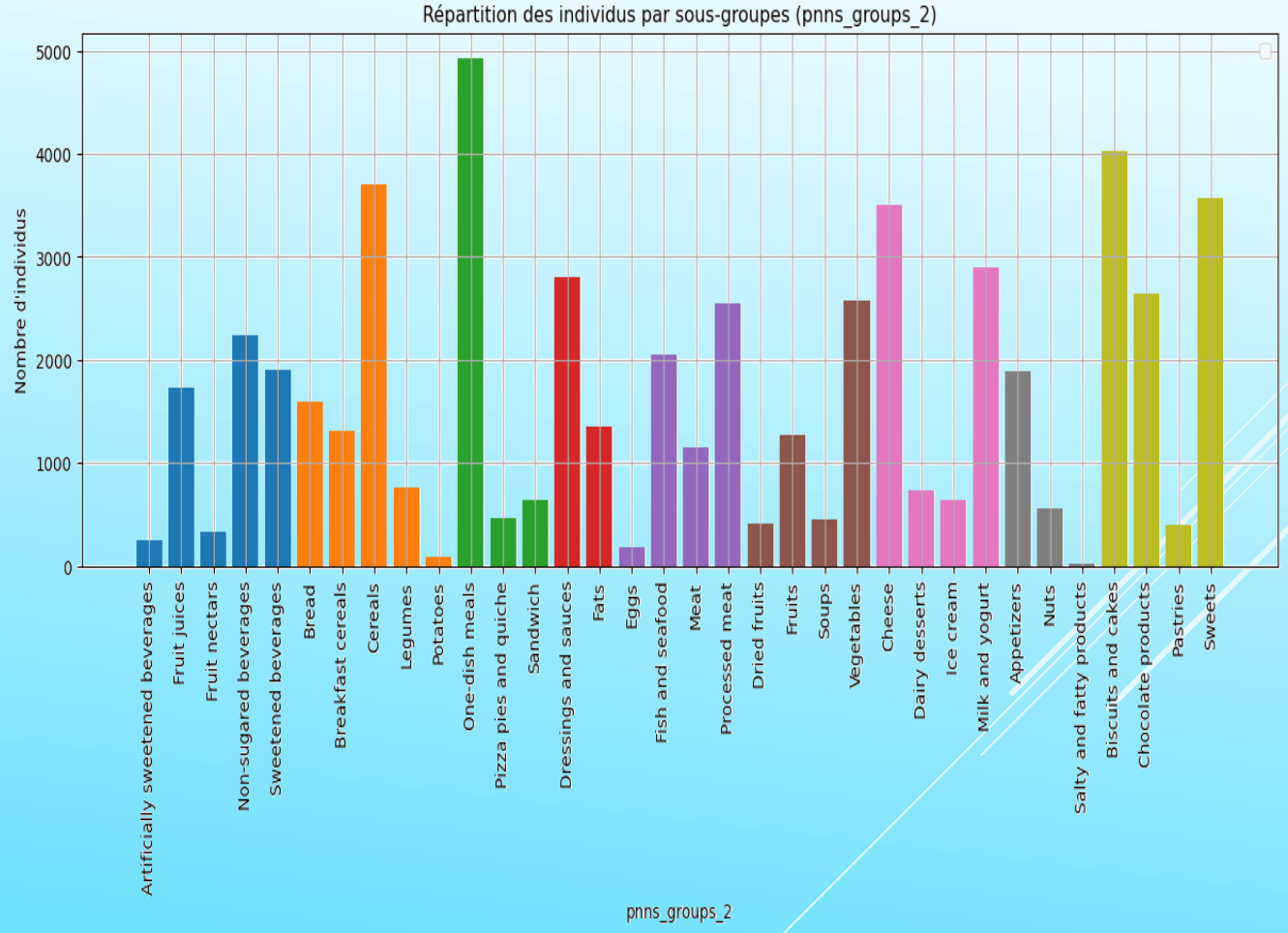
- P_value = 0.0
- Se référer à la méthode ayant le résultat de test le plus faible

→ Application de la méthode KNN pour toutes les variables

- Exception faite pour Carbohydrates : itération privilégiée ici

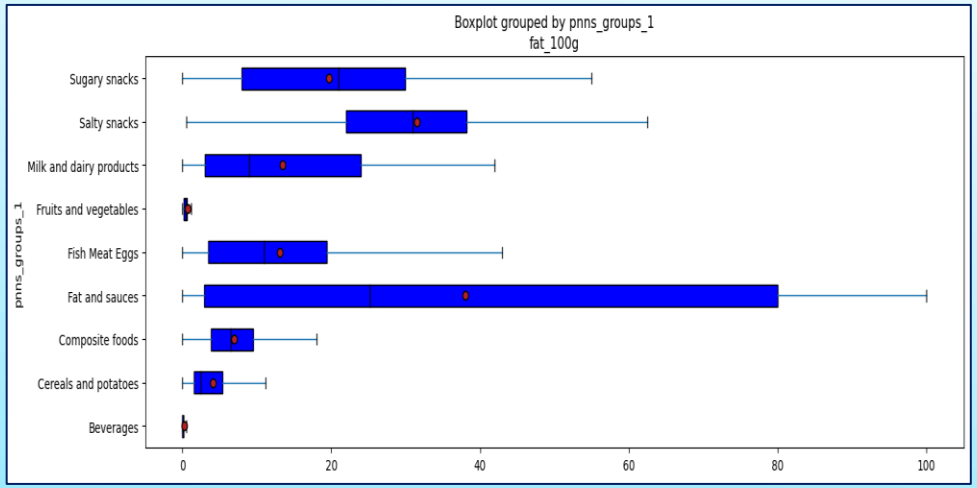
ANALYSES UNIVARIEE : VARIABLES QUALITATIVES



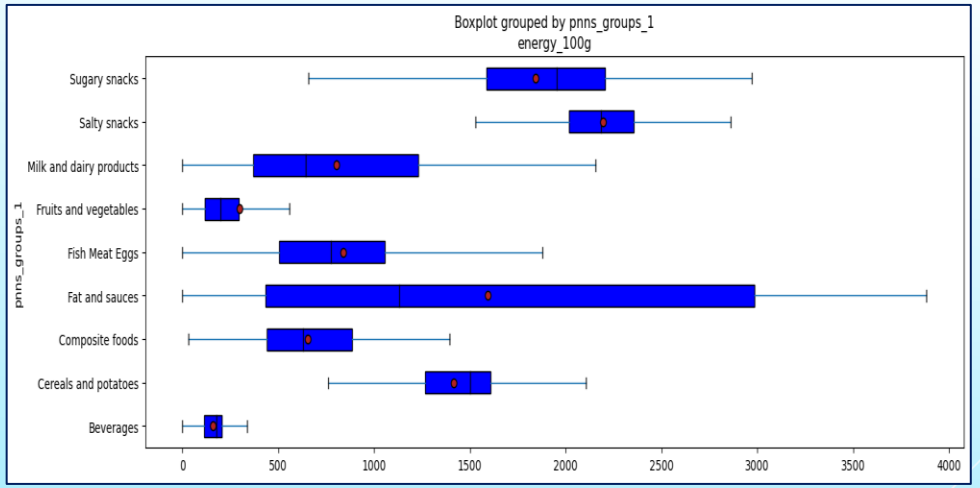


ANALYSES UNIVARIEE : VARIABLES QUANTITATIVES

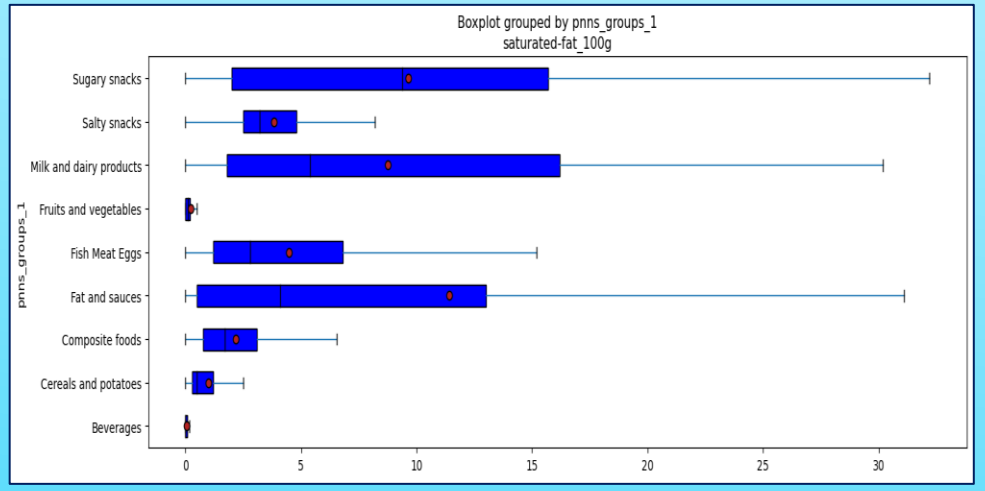
Fat



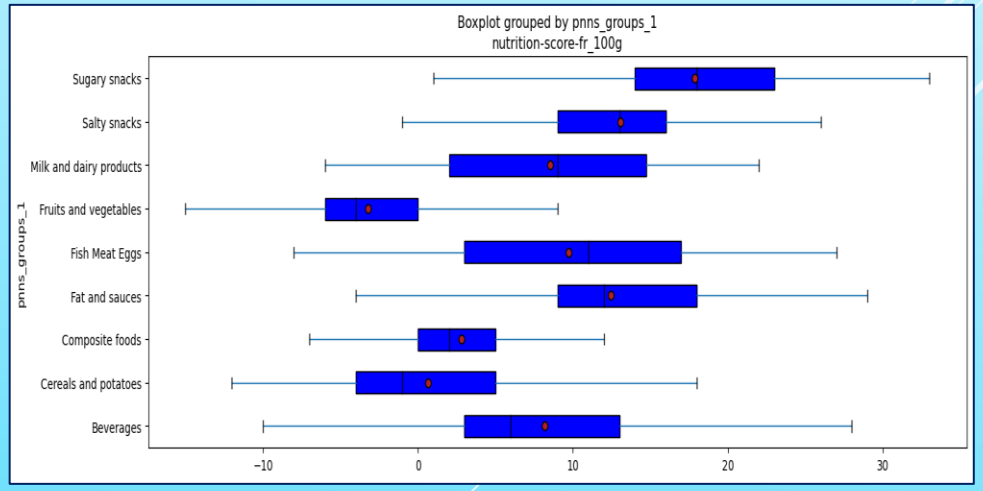
Energy



Saturated Fat

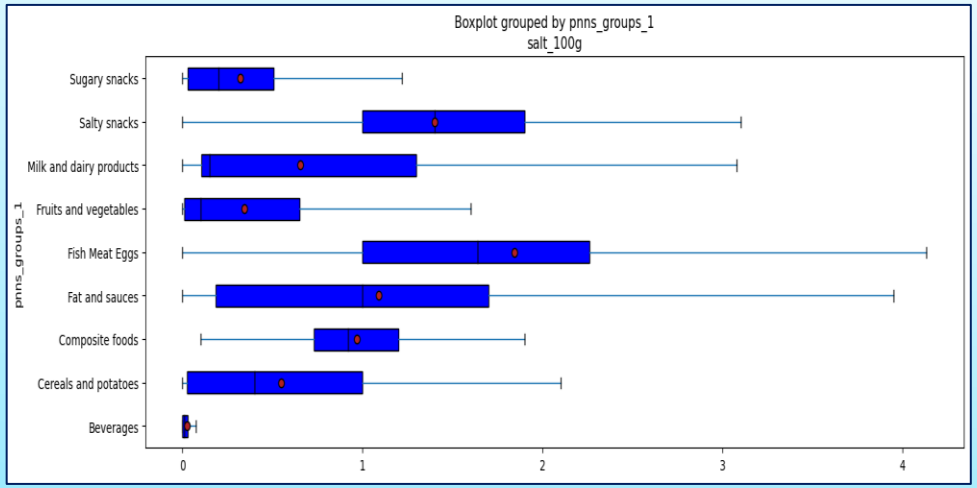


Nutrition Score FR

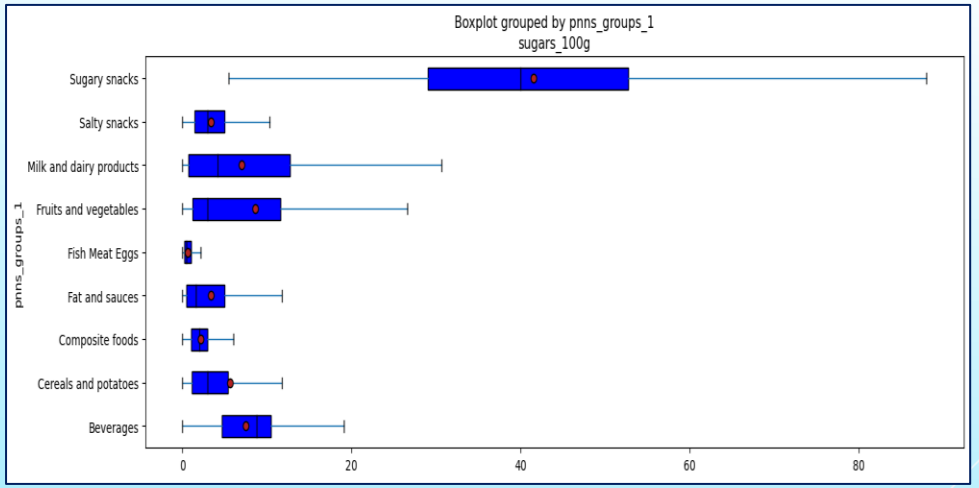


ANALYSES UNIVARIEE : VARIABLES QUANTITATIVES

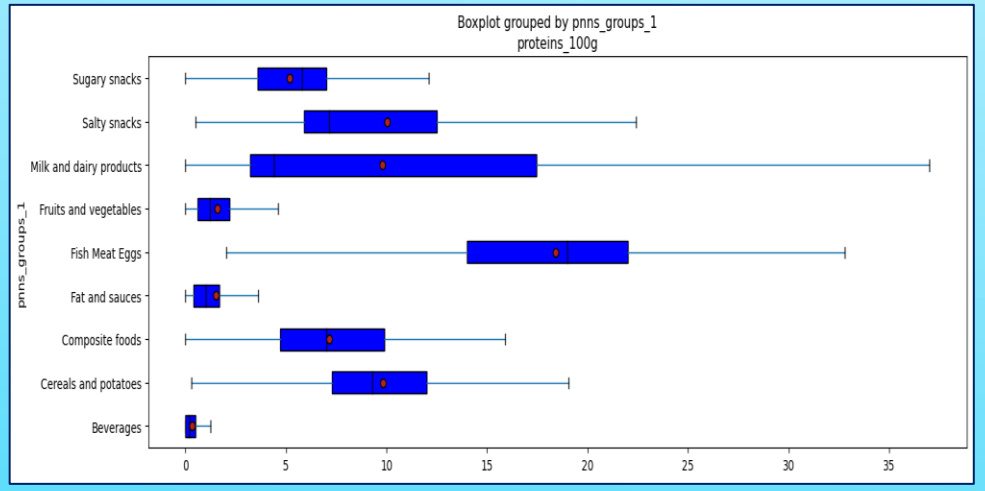
Salt



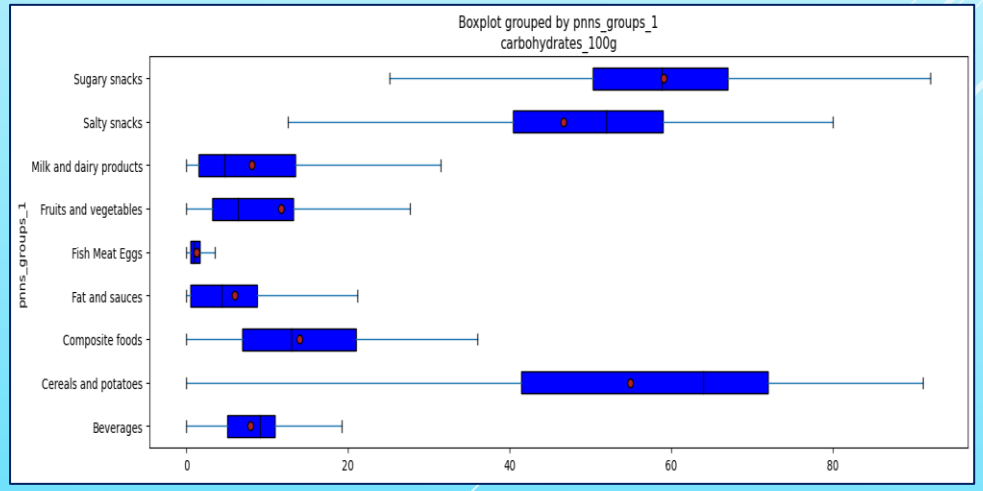
Sugars



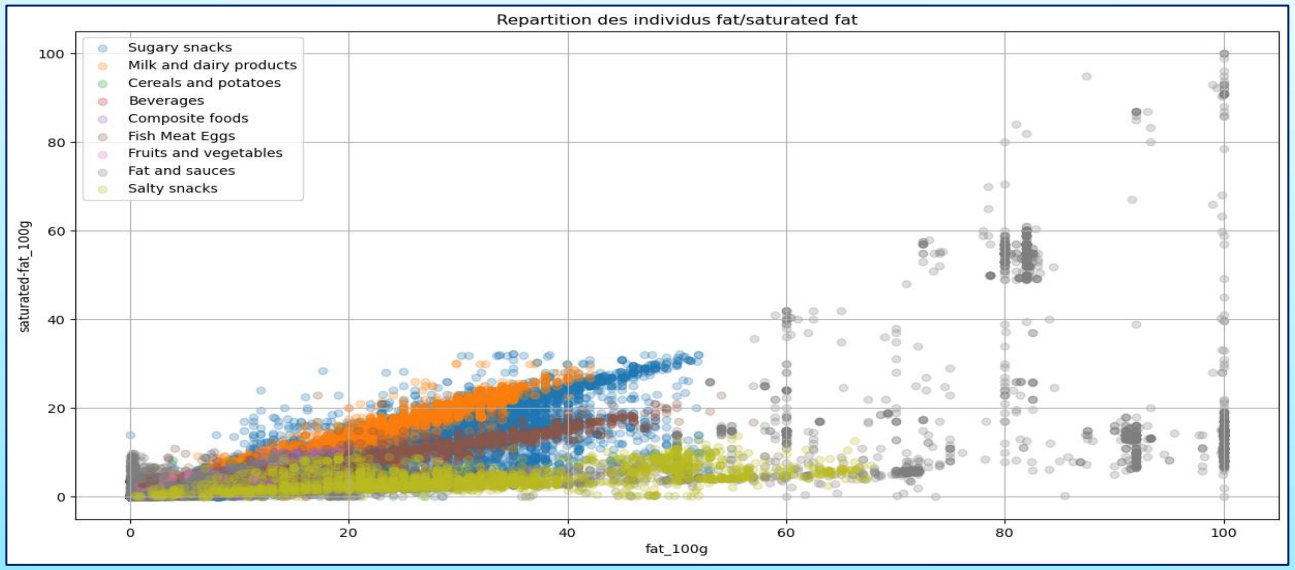
Proteins

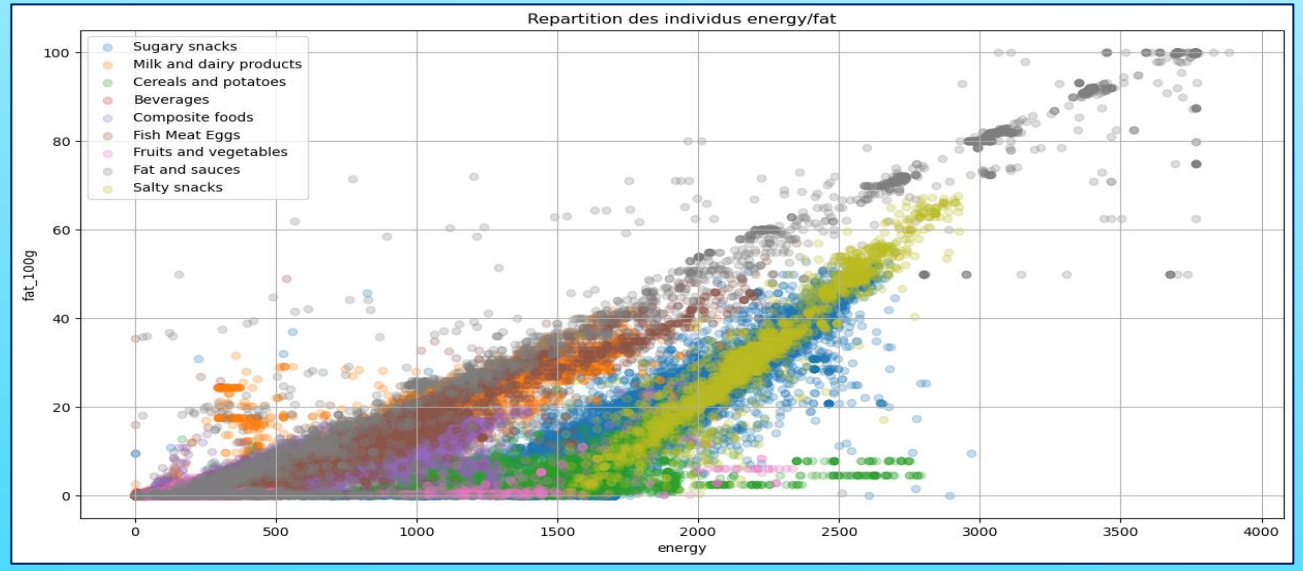


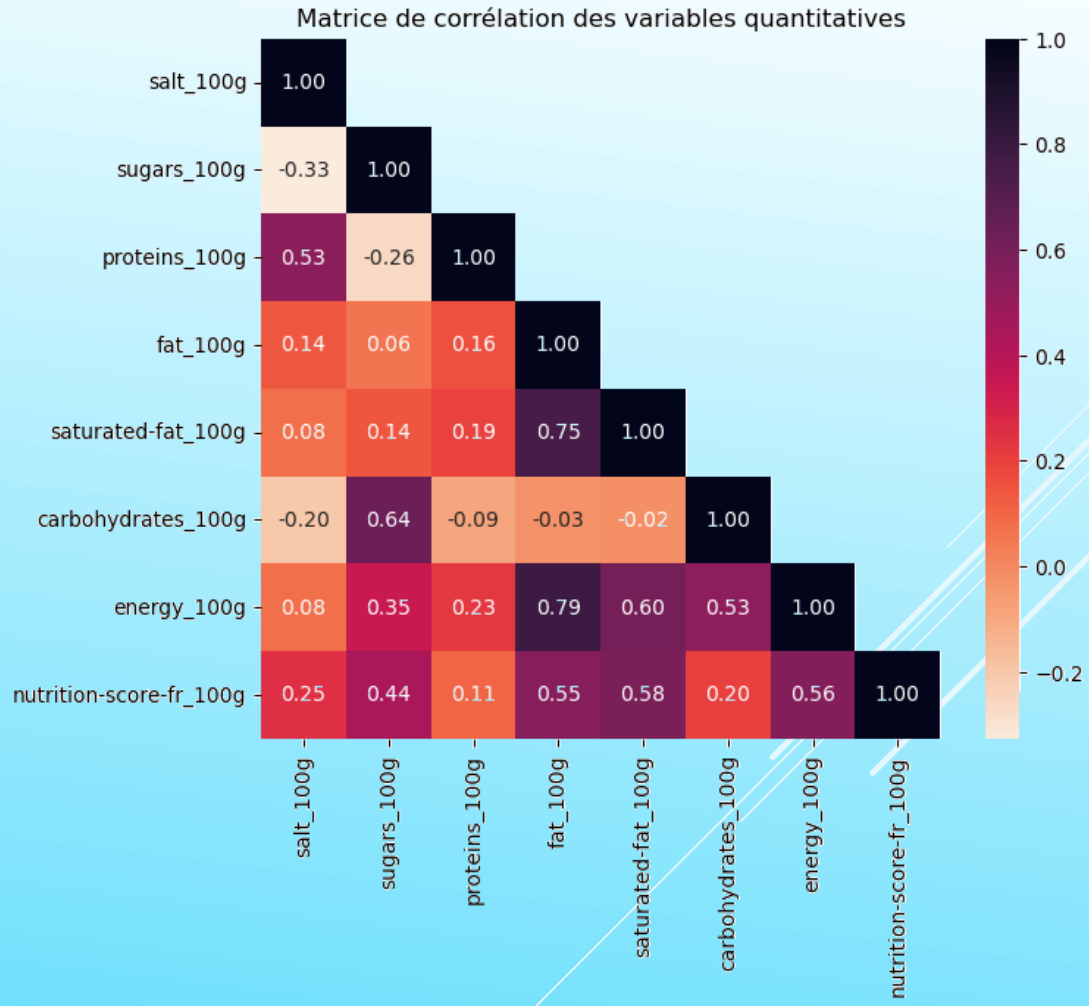
Carbo hydrates



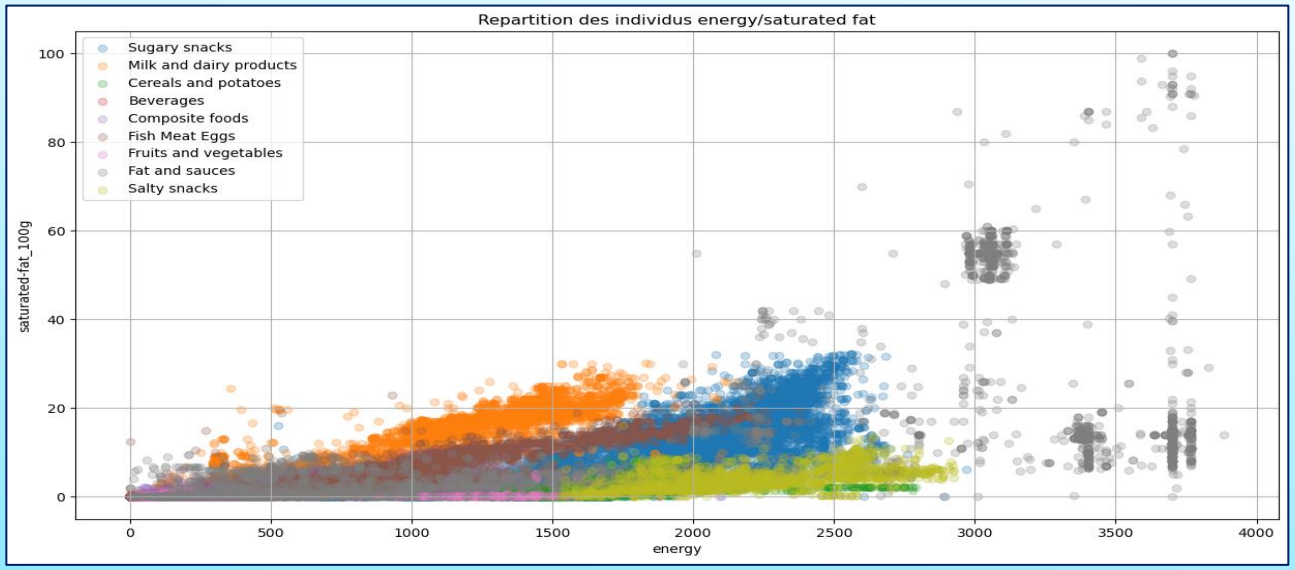
ANALYSES BI-VARIÉES

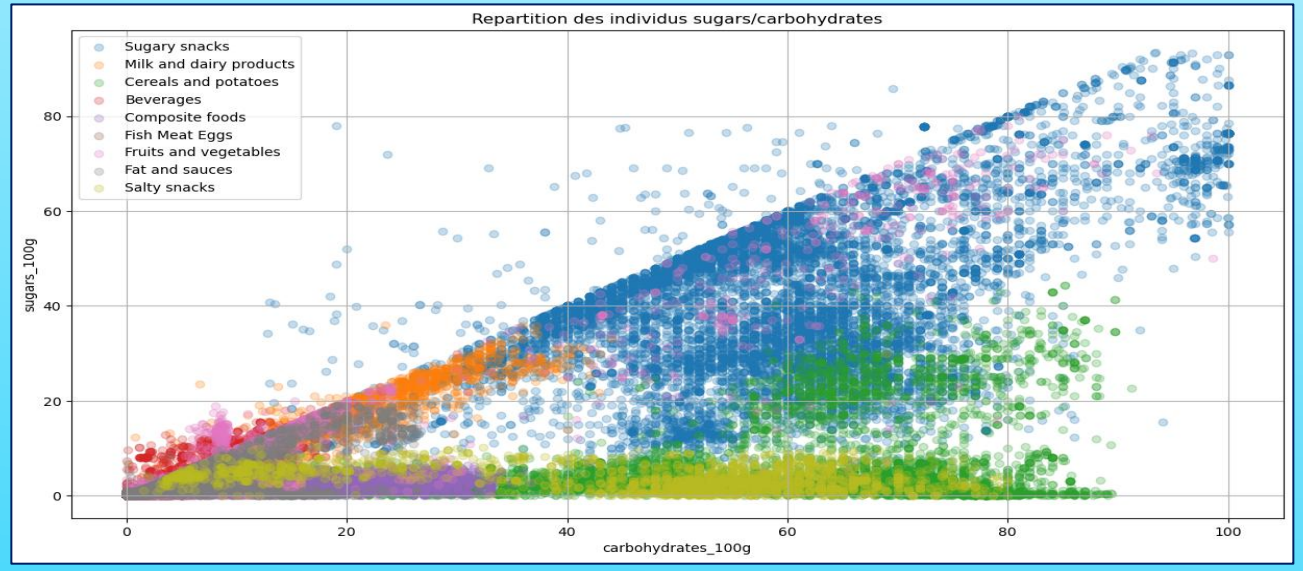


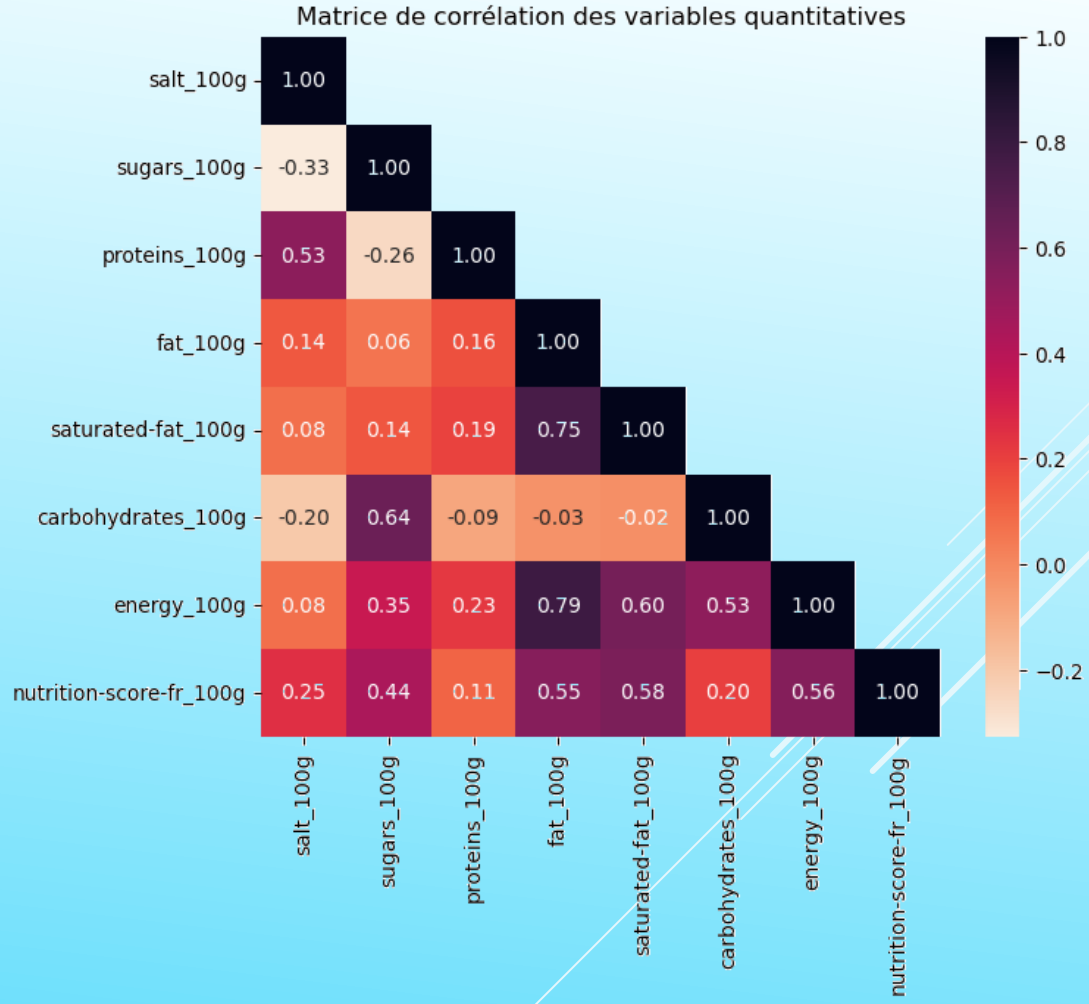




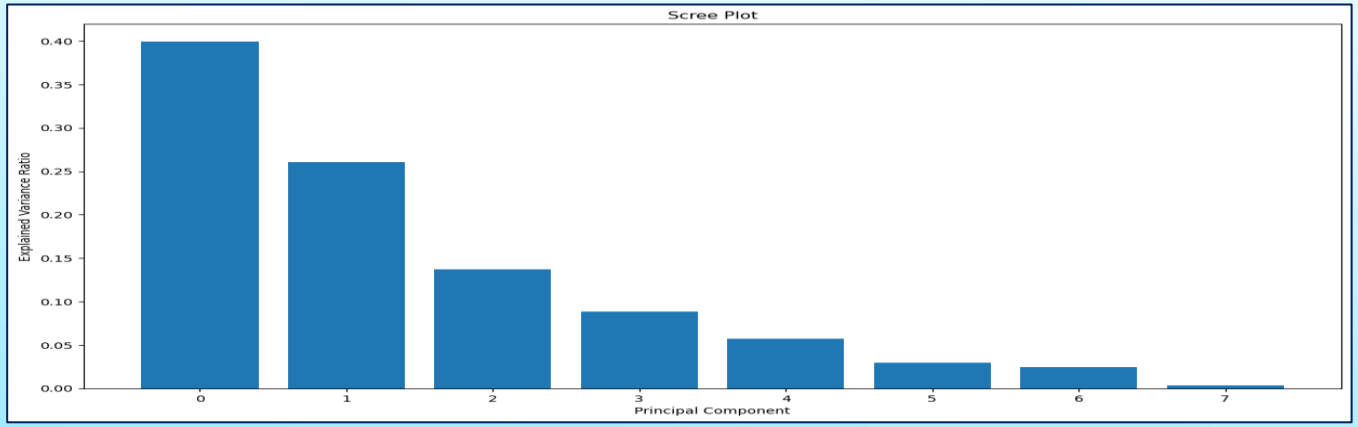
ANALYSES BI-VARIÉES

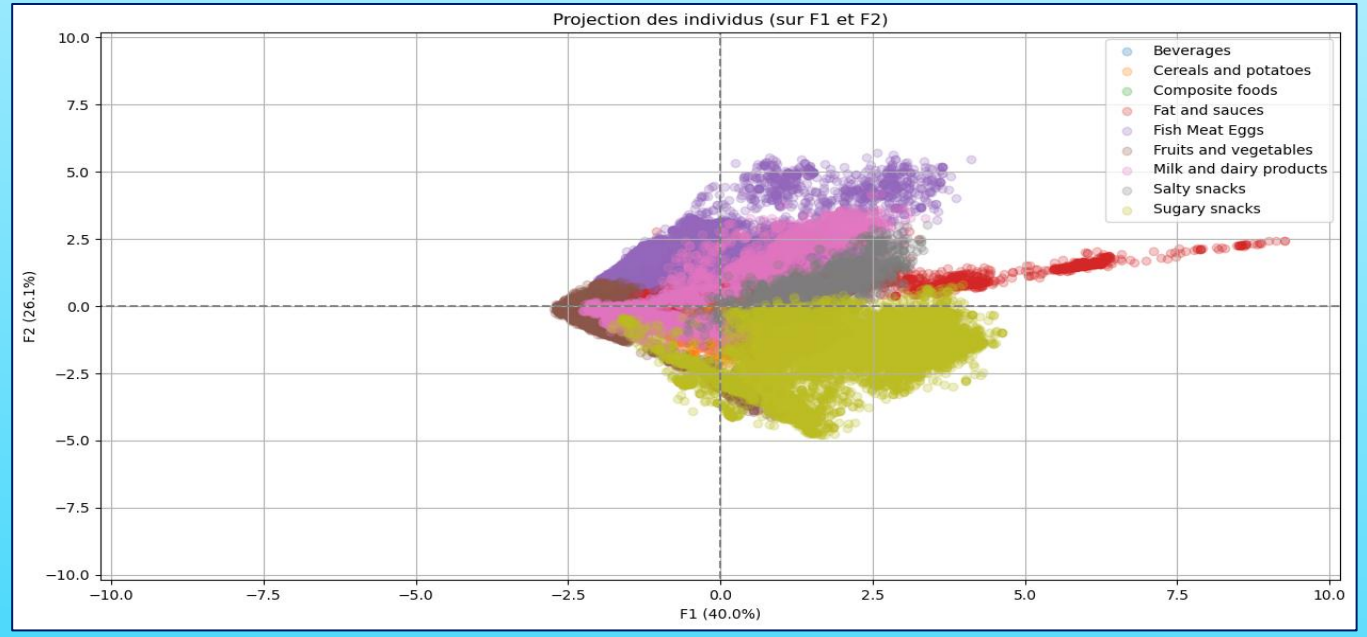


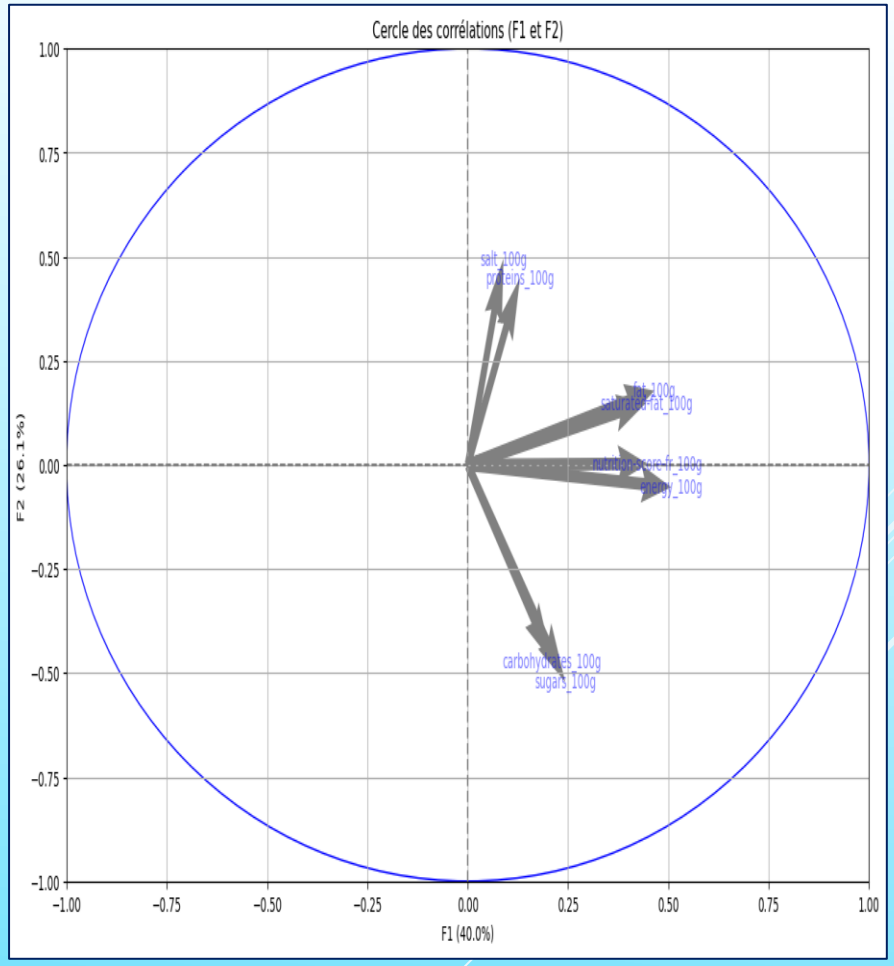




ANALYSE MULTI-VARIÉE : ACP







ANALYSE MULTI-VARIÉE : ANOVA

HOMOSCEDASTICITE:

- Test de Levene :

P-value = 0.0

→ Les variances sont inégales

NORMALITE DES RESIDUS :

- Test de Shapiro-Wilk
- P_value = 0.0

→ Les données ne suivent pas une distribution normale

ANOVA : test paramétrique non-fiable

Test de Kruskal-Wallis (test non-paramétrique)

P_value = 0.0

→ différences significatives entre les groupes pour chaque variable quantitative testée

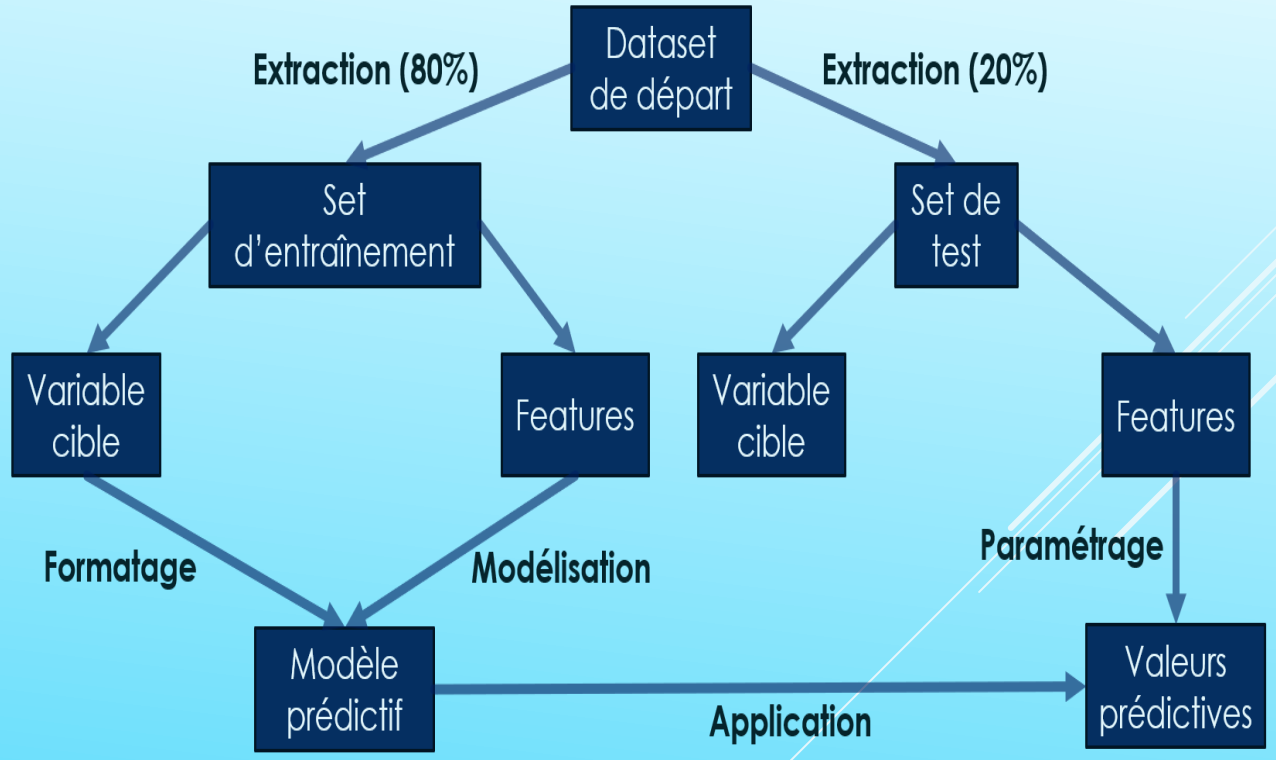
CONCLUSION

FAISABILITE DU PROJET = 96%

Existence de conditions sine qua non :

- Identification & traitement des valeurs abérrantes par approches métier & statistiques
- A partir de variables remplies à +50%
- Imputation des valeurs manquantes optimisée
- Respect des corrélations & des distributions

MODELE DE PREDICTION : K Neighbors Classifier :



MODELE DE PREDICTION : K NEIGHBORS CLASSIFIER

pnns_groups-1	Dataset traité	Dataset non traité
Beverages	98%	98%
Cereals and potatoes	95%	86%
Composite foods	90%	78%
Fat and sauces	95%	85%
Fish Meat Eggs	96%	88%
Fruits and vegetables	93%	89%
Milk and dairy products	98%	89%
Salty snacks	99%	78%
Sugary snacks	99%	91%
TOTAL ACCURACY	**96%**	**87%**

Merci pour votre attention