Santé Publique France - OpenFoodFacts

- Nettoyage et exploration des données en vue de développer un système d’autocomplétion

Préambule

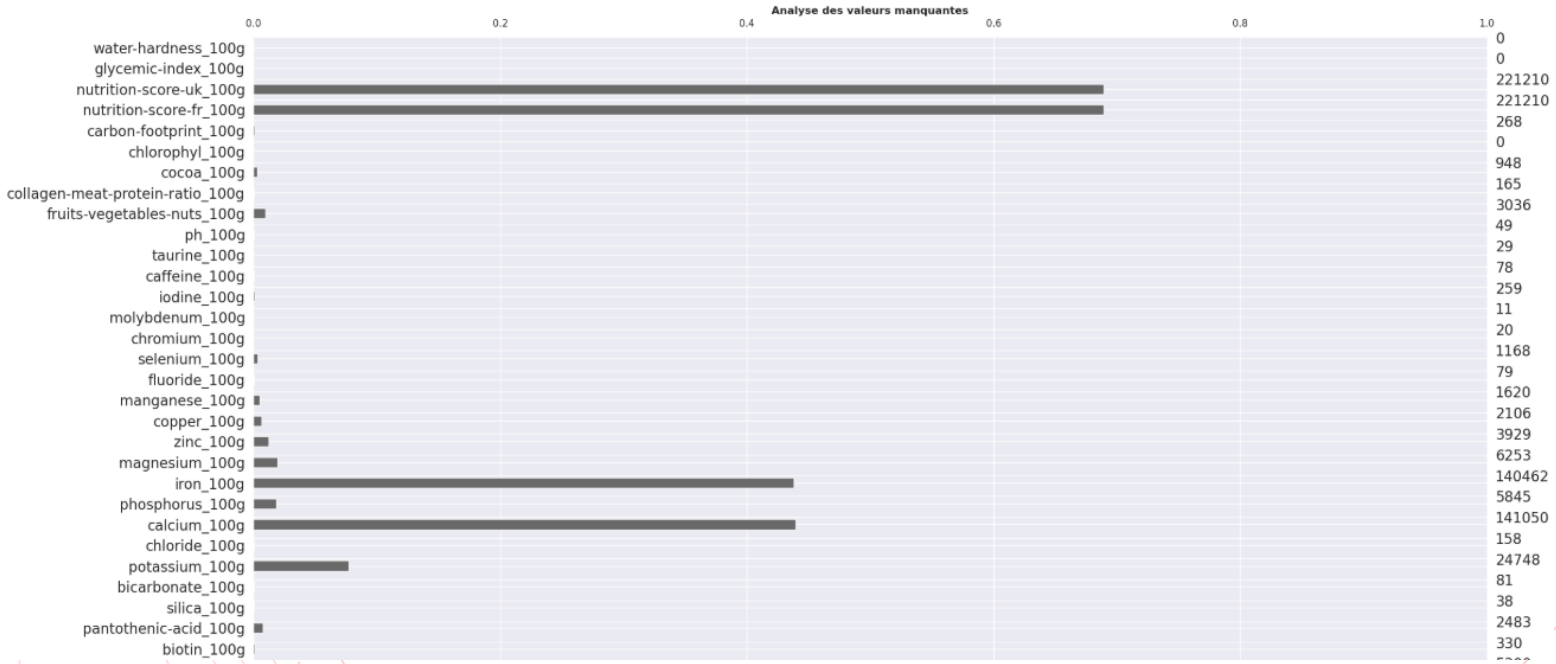
- Open Food Facts
  - Base de données de produits alimentaires
  - Nombreux champs
  - Remplissage fastidieux
  - Conduit à des valeurs manquantes
- Santé Publique France
  - Souhaite améliorer cet outil
  - Projet d'autocomplétion lors de la saisie d'un produit
- Nettoyage et exploration des données
- Étude de faisabilité de l'outil

Sommaire

- Sélection des données
  - Cible
  - Features
- Nettoyage des données
  - Valeurs manquantes
  - Valeurs aberrantes
- Exploration des données
  - Analyses statistiques
  - Analyses de variance
  - Modèles prédictifs
- Conclusions

Sélection des données

- Premier passage en revue du jeu de données
  - 321000 produits environs
  - 162 caractéristiques
  - Constat du nombre de valeurs manquantes important



Extrait du tableau de remplissage des variables

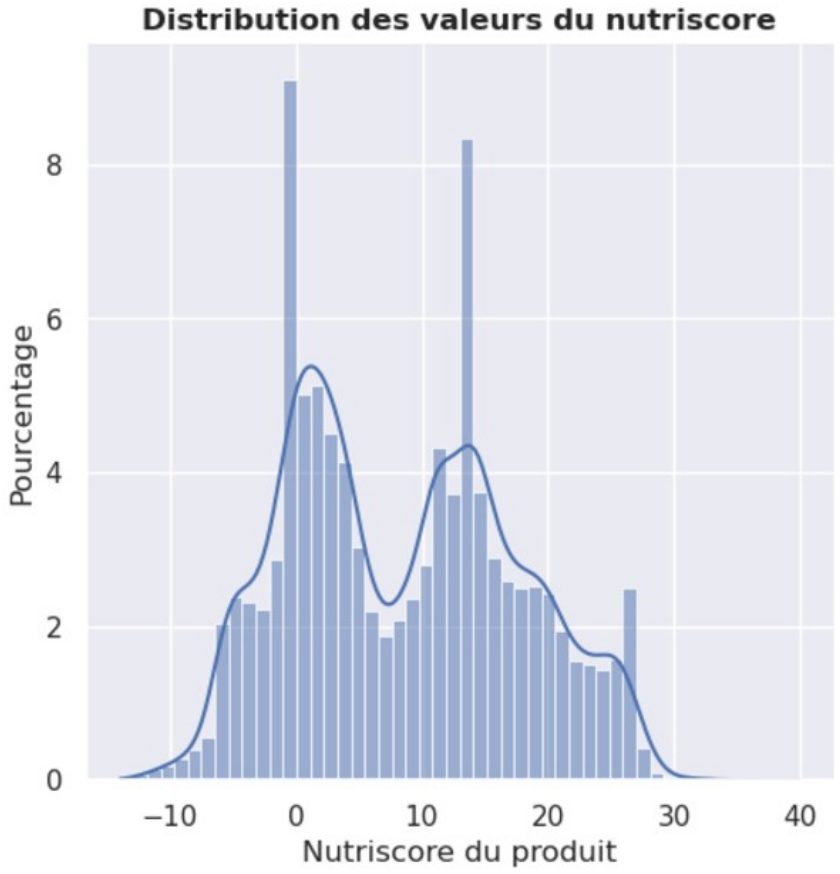
Sélection des données

	index	valeurs
8	generic_name	16.458731
9	quantity	32.677104
10	packaging	24.615615
14	categories	26.314641
17	origins	6.917686
19	manufacturing_places	11.379110
21	labels	14.514671
24	emb_codes	9.136084
26	first_packaging_code_geo	5.861796
29	purchase_places	18.141546
30	stores	16.124225
35	allergens	8.836183
37	traces	7.591997
54	pnns_groups_1	28.528986
55	pnns_groups_2	29.457372
59	main_category	26.300924

- Simplification du dataset
  - Taux de remplissage
  - Informations redondantes
  - Informations personnelles (auteur)
- Cible
  - Taux de remplissage bas
  - Variable catégorielle
- Features
  - Taux de remplissage suffisant (supérieur à 50%)
  - Variable quantitatives

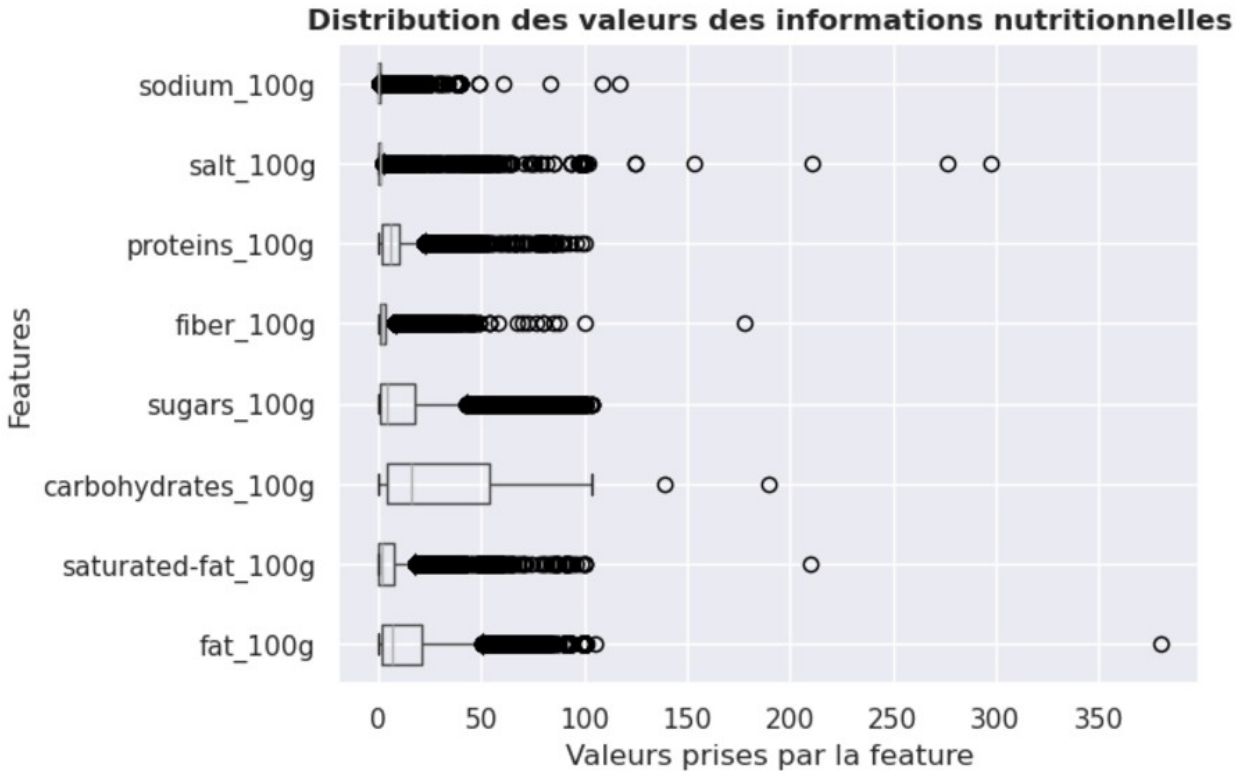
Nettoyage – Valeurs aberrantes

- Nutriscore
  - Valeurs comprises entre -15 et + 40
  - Rien d'aberrant dans les valeurs constatées



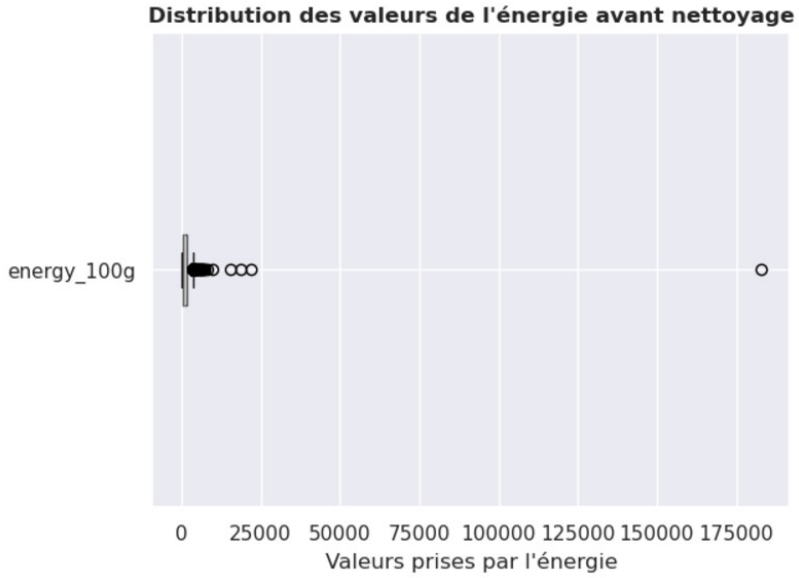
Nettoyage – Valeurs aberrantes

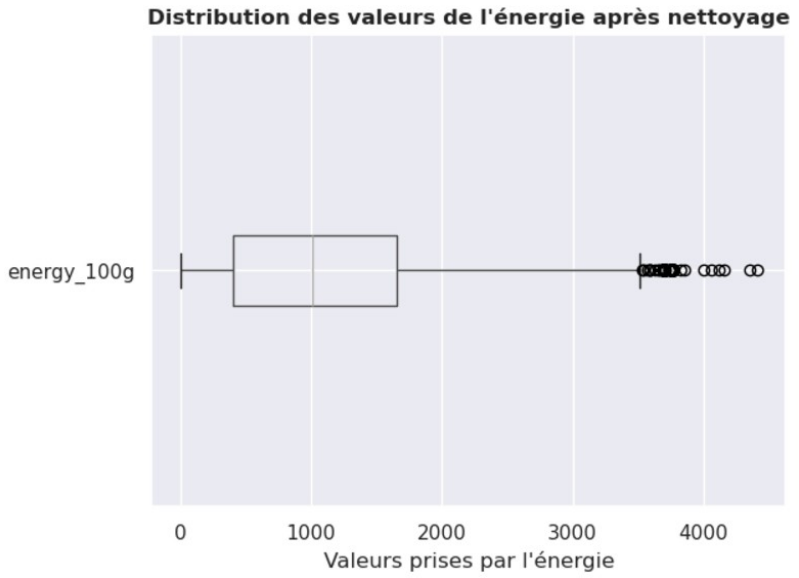
- Valeurs nutritionnelles
  - Valeurs comprises entre 0 et 100
  - Peu de valeurs aberrantes
  - Remplacées par NaN



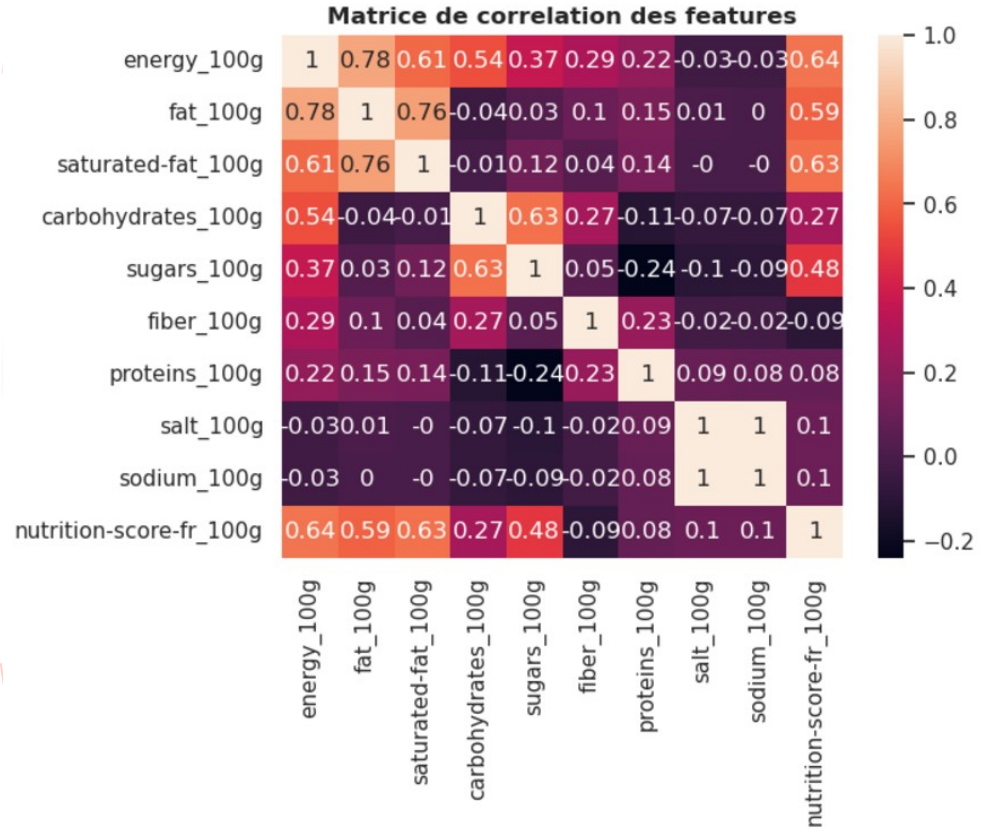
Nettoyage – Valeurs aberrantes

- Énergie
  - Dépend des autres valeurs nutritionnelles
  - Détection des valeurs aberrantes par la méthode de l'écart type
    - Valeur > moyenne + 3 écart type
  - Remplacées par NaN





Nettoyage – Valeurs manquantes



- Corrélations entre variables
  - Permet de définir des régressions linéaires envisageables
  - Régressions identifiées :
    - energy_100g ~ fat_100g + saturated-fat_100g + carbohydrates_100g
    - fat_100g ~ energy_100g + saturated-fat_100g
    - saturated-fat_100g ~ fat_100g + energy_100g
    - carbohydrates_100g ~ sugars_100g + energy_100g
    - sugars_100g ~ carbohydrates_100g
    - nutrition-score-fr_100g ~ saturated-fat_100g + energy_100g + fat_100g + sugars_100g

Nettoyage – Valeurs manquantes

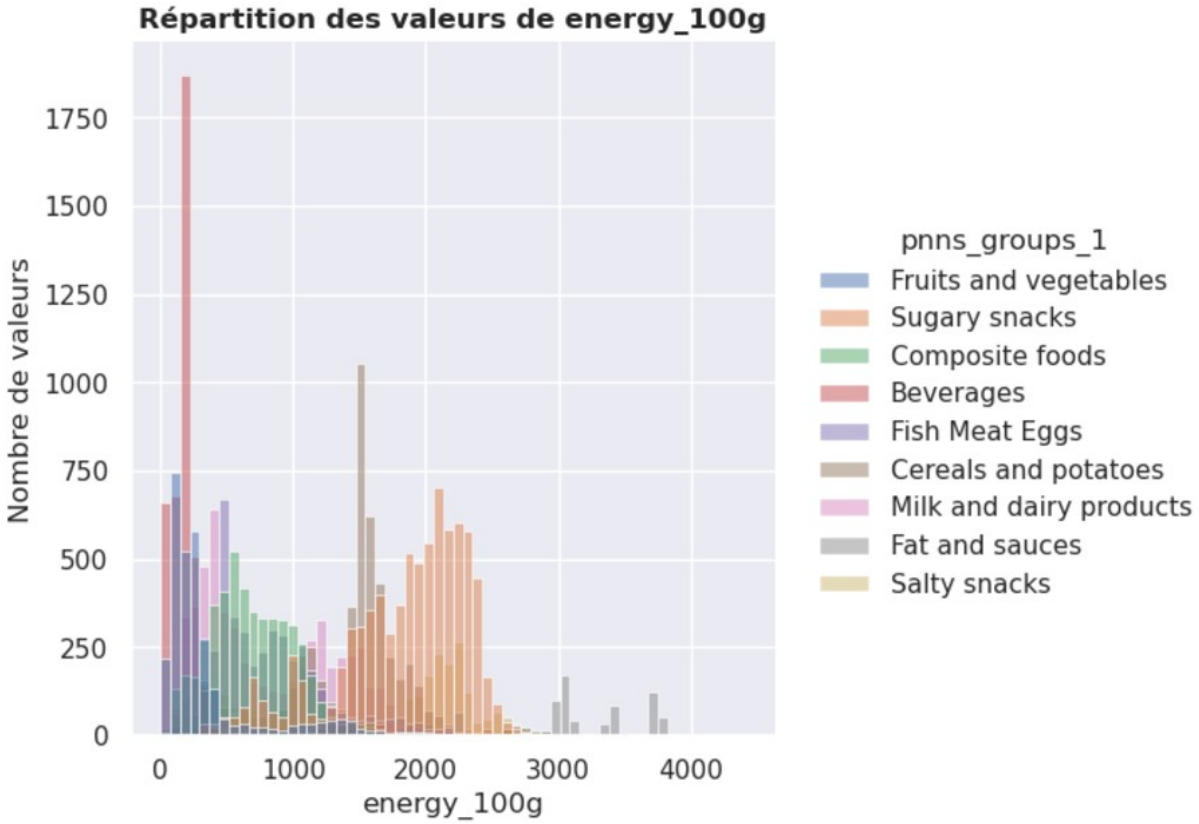
- Imputations statistiques
  - Nécessaires pour effectuer les régressions linéaires qui nécessitent une absence de valeurs manquantes
  - Sélection de variables avec taux de valeurs manquantes faible
  - Étude des distribution par groupe d'aliments

pnns_groups_1	0.000000
energy_100g	0.060255
fat_100g	4.270543
saturated-fat_100g	0.251061
carbohydrates_100g	4.348372
sugars_100g	0.208380
fiber_100g	35.921769
proteins_100g	0.123020
salt_100g	0.007532
sodium_100g	0.005021
nutrition-score-fr_100g	1.119731

dtype: float64

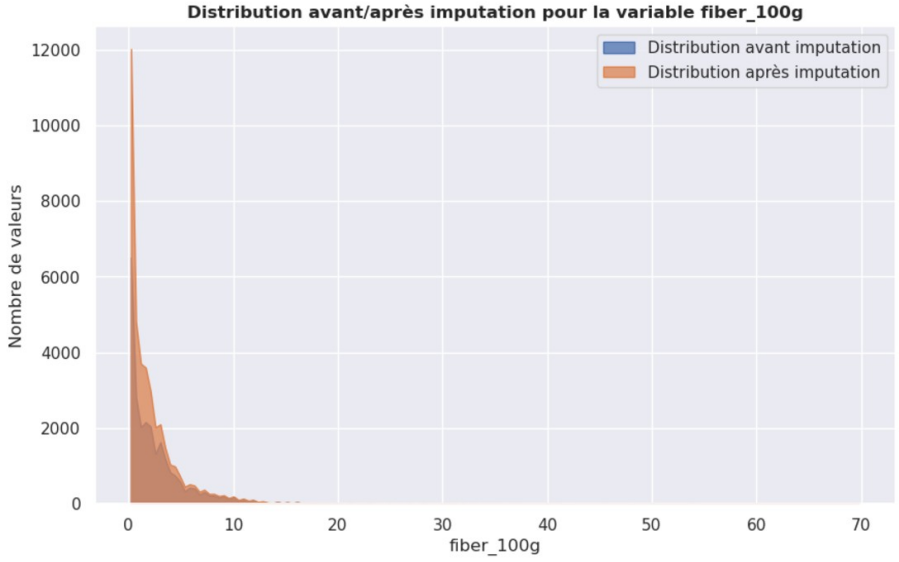
Nettoyage – Valeurs manquantes

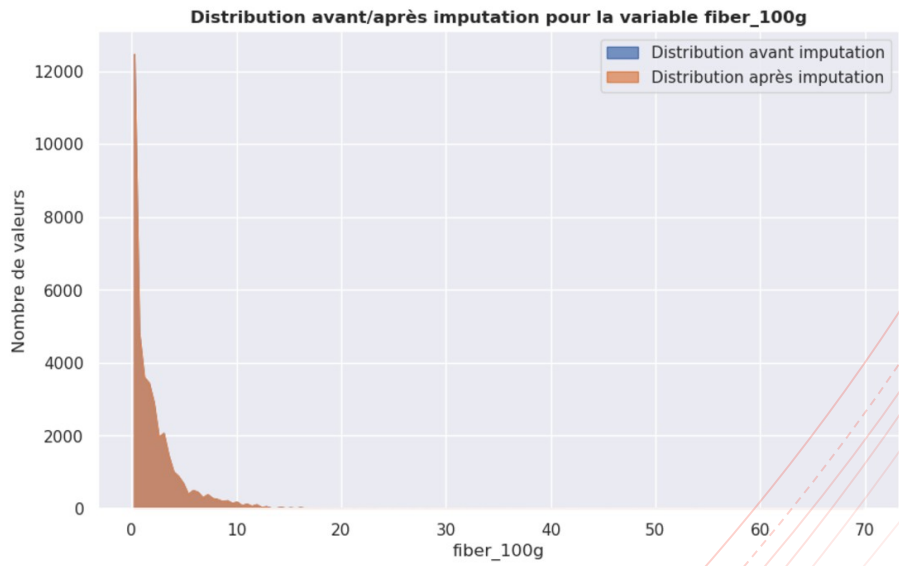
- Imputations statistiques
  - Utilisation de la moyenne par type d'aliment pour compléter les valeurs manquantes



Nettoyage – Valeurs manquantes

- K plus proches voisins
  - Utilisé sur des variables avec faible taux de corrélation
  - Première approche avec 5 voisins pas forcément satisfaisante
  - Deuxième approche avec 3 voisins plus représentative





Distribution de la variable fiber_100g avec 5 voisins plus proches et 3 voisins plus proches

Nettoyage – Valeurs manquantes

- Régressions linéaires
  - fat_100g ~ energy_100g + saturated-fat_100g
  - Première approche avec une régression linéaire simple
  - fat_100g ~ energy_100g + saturated-fat_100g + energy_100g^2 + saturated-fat_100g^2
  - Deuxième approche avec régression polynomiale

```
La performance du Modèle pour le set de Training
--------------------------------------------------
l'erreur RMSE esst 8.11181749215141
le score est 0.7526348810013044


La performance du Modèle pour le set de Test
--------------------------------------------------
l'erreur RMSE est 8.280025762366726
le score est 0.7523761158567875
```

```
La performance du Modèle pour le set de Training
--------------------------------------------------
l'erreur RMSE esst 6.503083844866705
le score est 0.841020610560175


La performance du Modèle pour le set de Test
--------------------------------------------------
l'erreur RMSE est 6.4206795774890555
le score est 0.8511012164952003
```

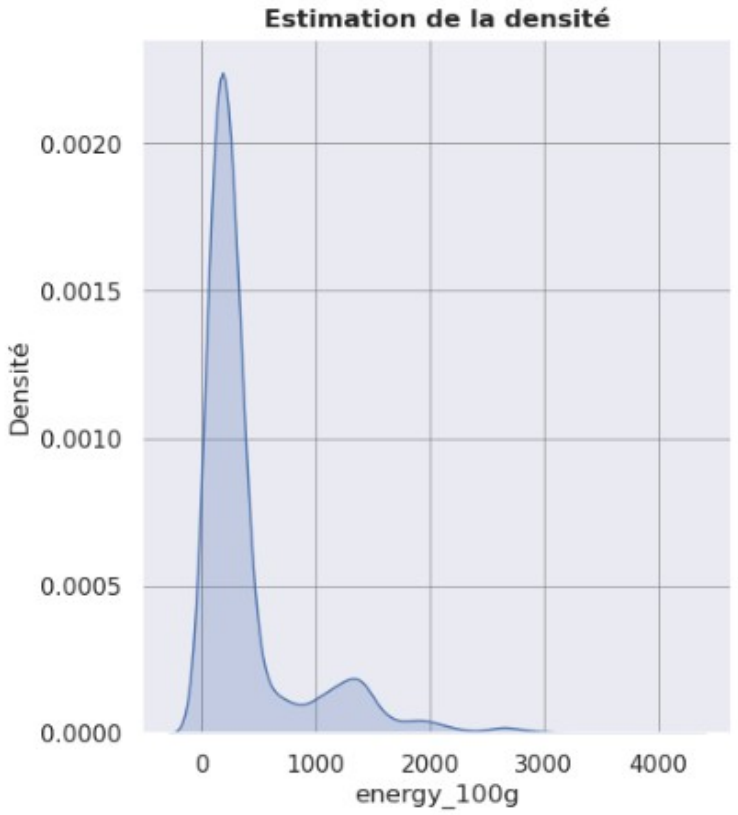
Scores des modèles de régression linéaire et régression polynomiale pour la variable fat_100g

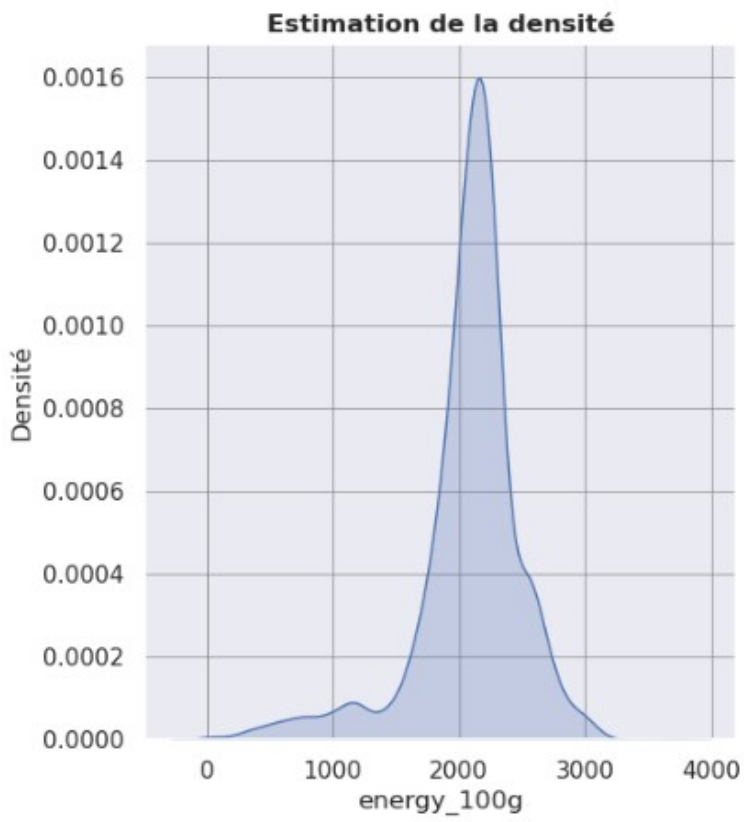
Exploration – Analyses statistiques

- Objectif
  - Étayer le constat fait précédemment : chaque groupe d'aliments à une répartition propre des features
  - Mettre en avant les différences entre groupes d'aliments
- Méthode
  - Comparer les statistiques des différents groupes d'aliments
  - Comparer les corrélations des features pour chaque groupe d'aliments
  - Analyser les résultats de l'ACP pour chacun des groupes d'aliments
- Exemple de résultats
  - Fruits et légumes VS Snacks salés

Exploration – Analyses statistiques

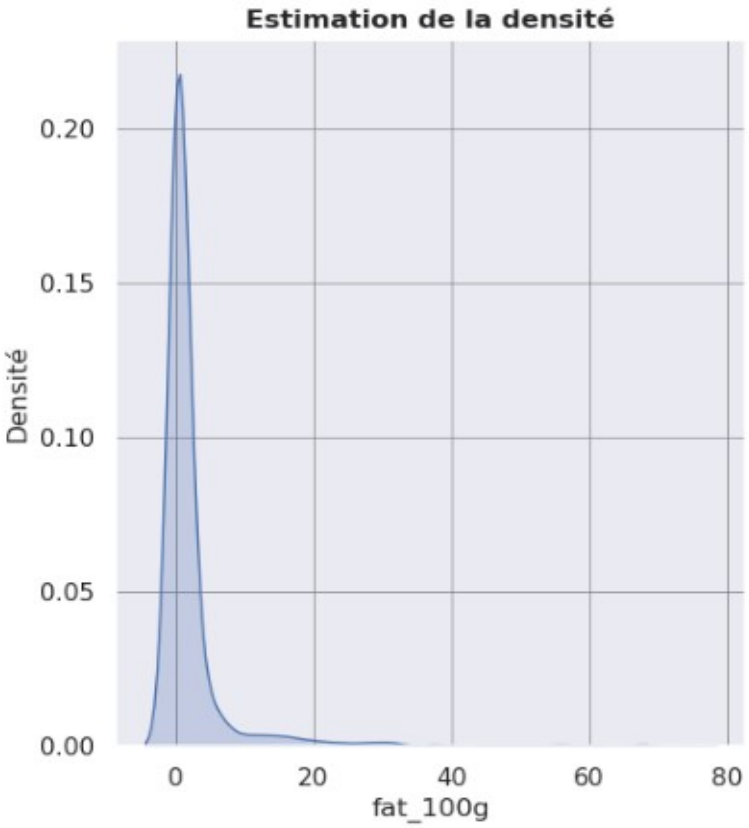
- Énergie :
  - A gauche les fruits et légumes
  - A droite les snacks salés

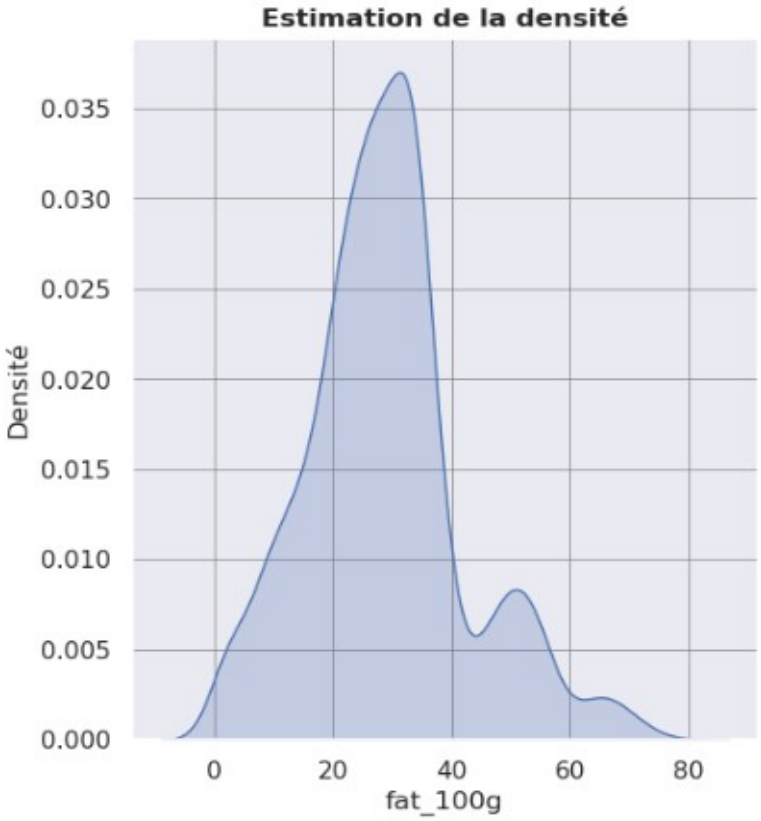




Exploration – Analyses statistiques

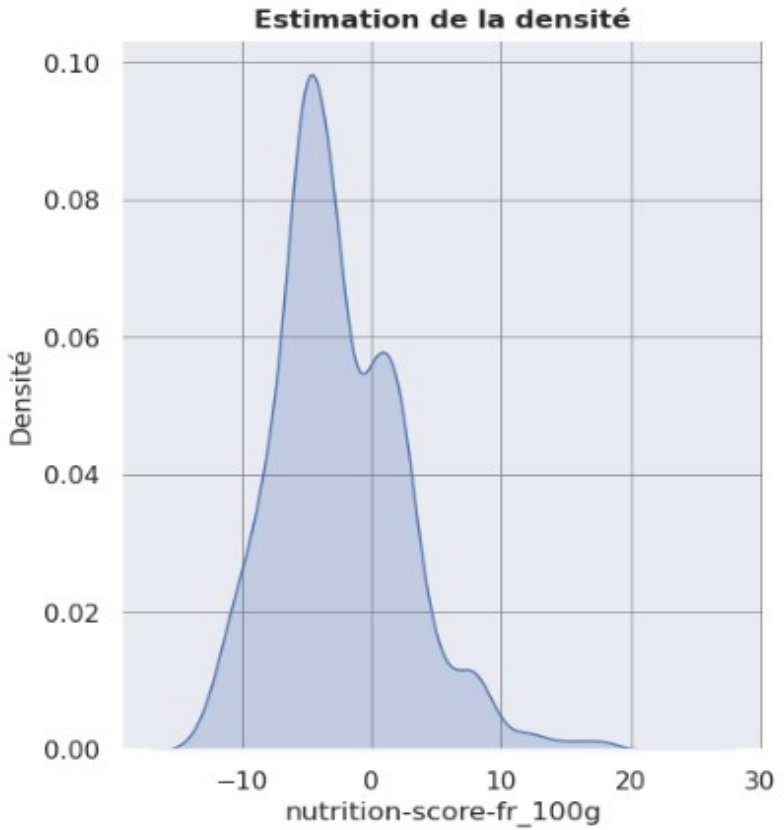
- Gras :
  - A gauche les fruits et légumes
  - A droite les snacks salés

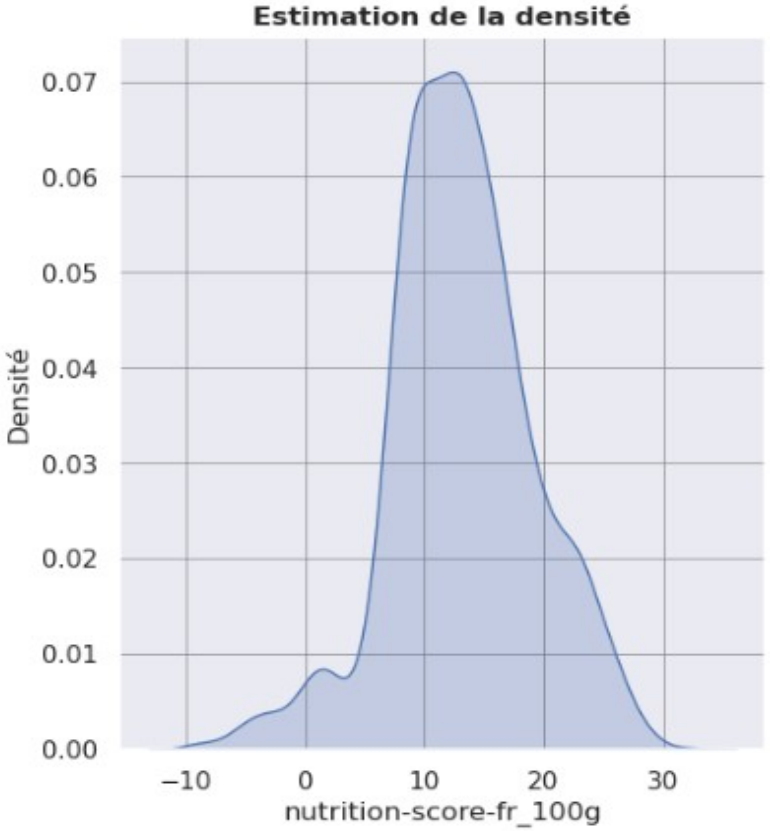




Exploration – Analyses statistiques

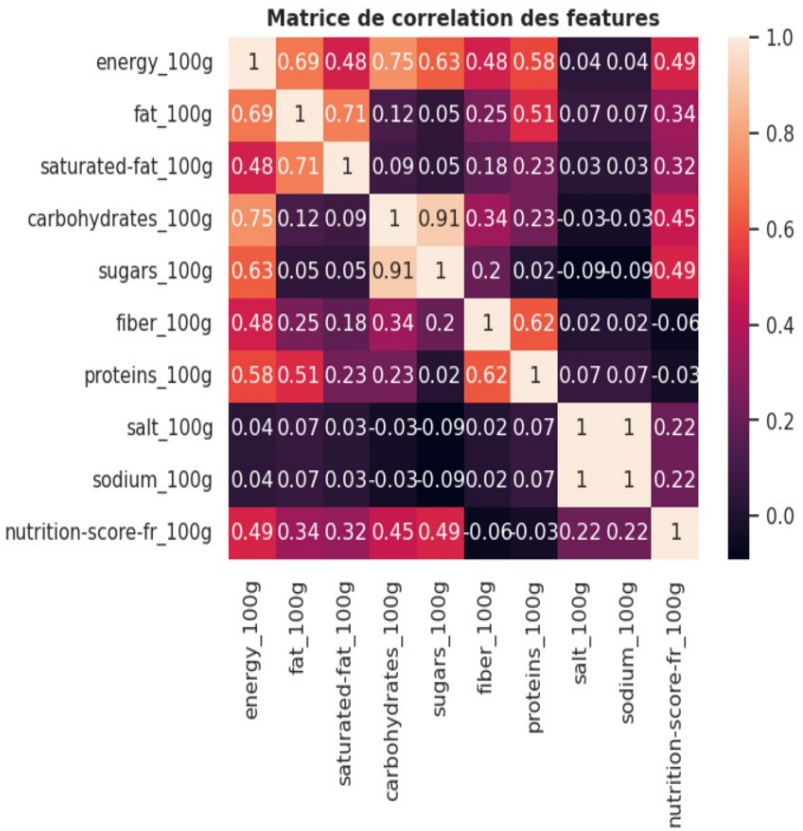
- Nutriscore :
  - A gauche les fruits et légumes
  - A droite les snacks salés

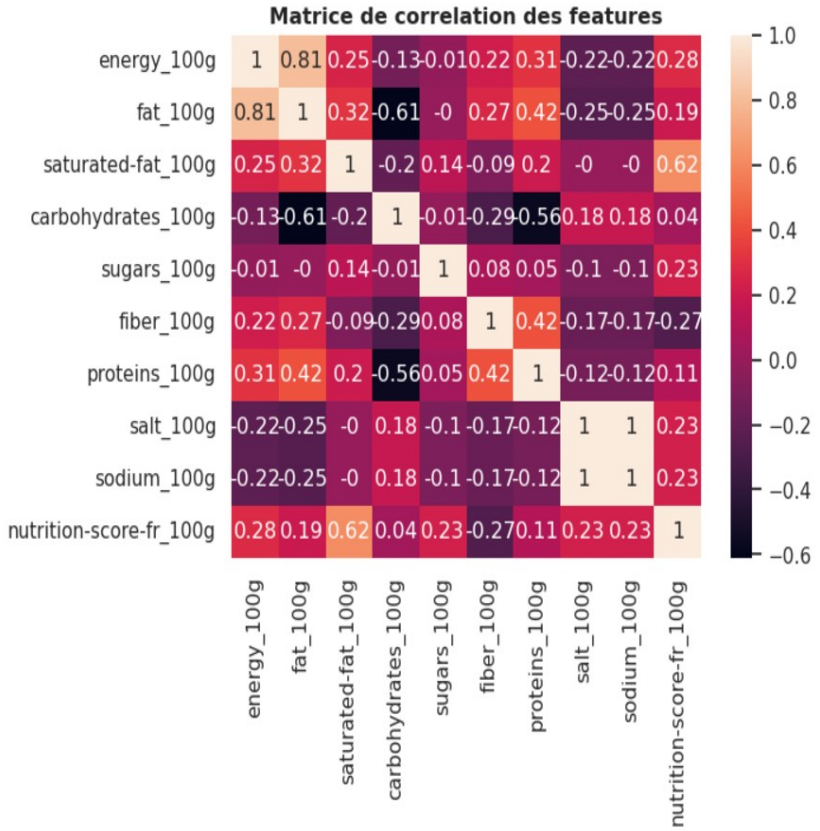




Exploration – Analyses statistiques

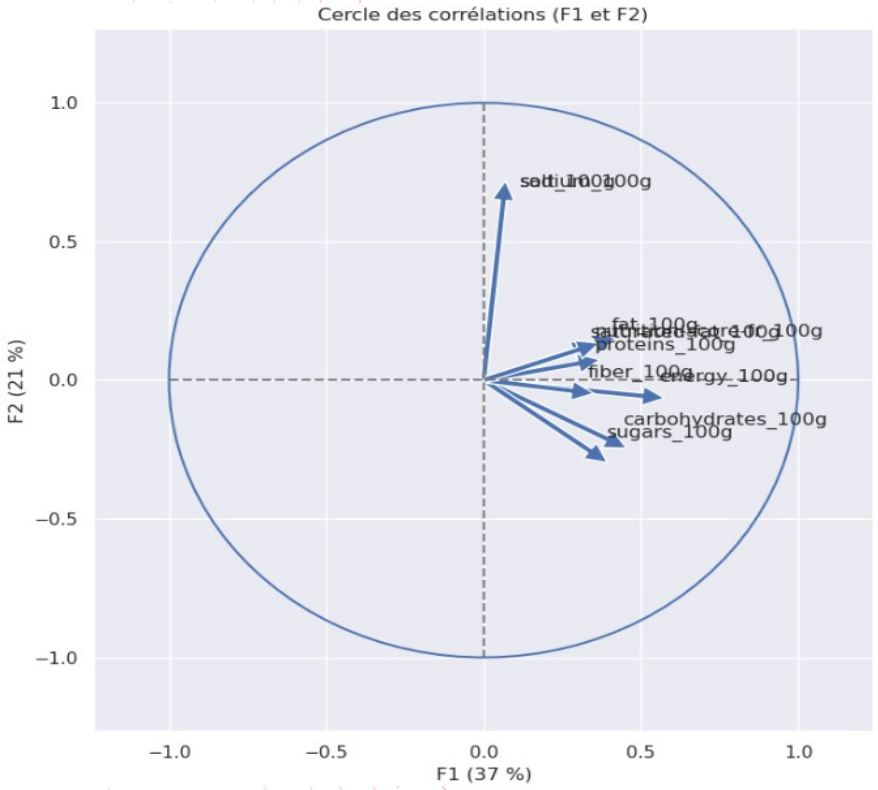
- Corrélations :
  - A gauche les fruits et légumes
  - A droite les snacks salés

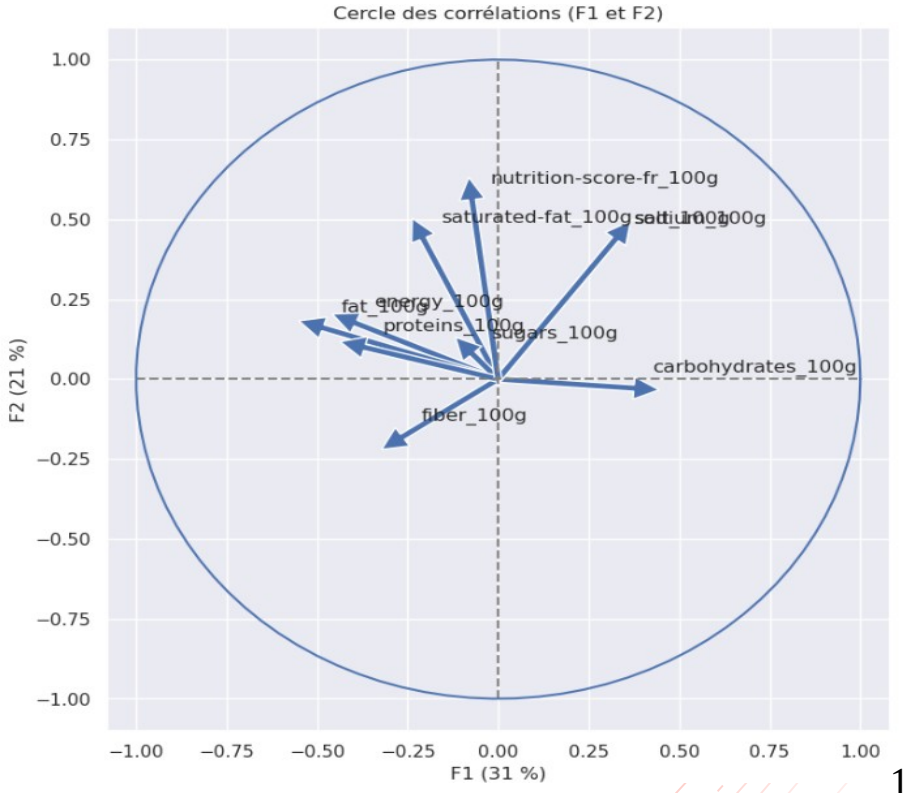




Exploration – Analyses statistiques

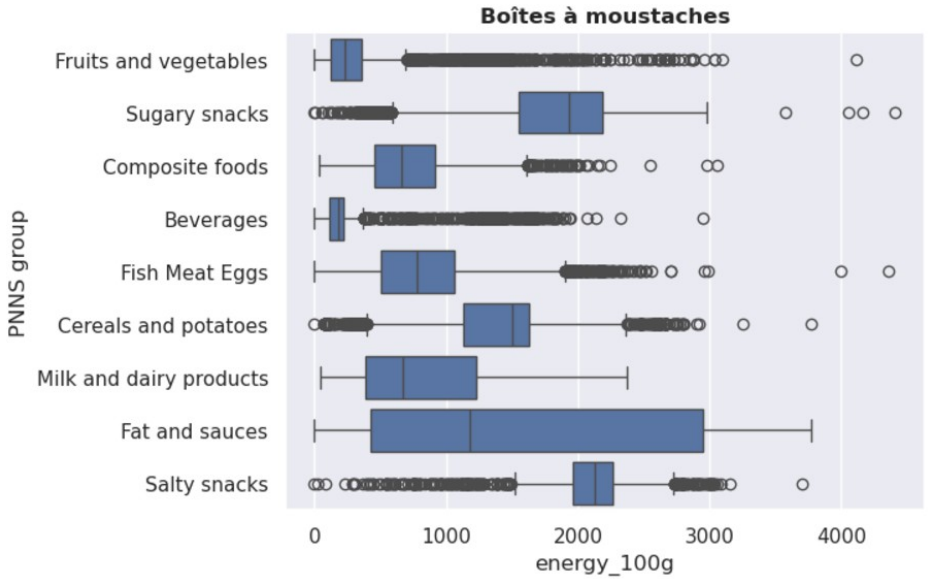
- ACP :
  - A gauche les fruits et légumes
  - A droite les snacks salés

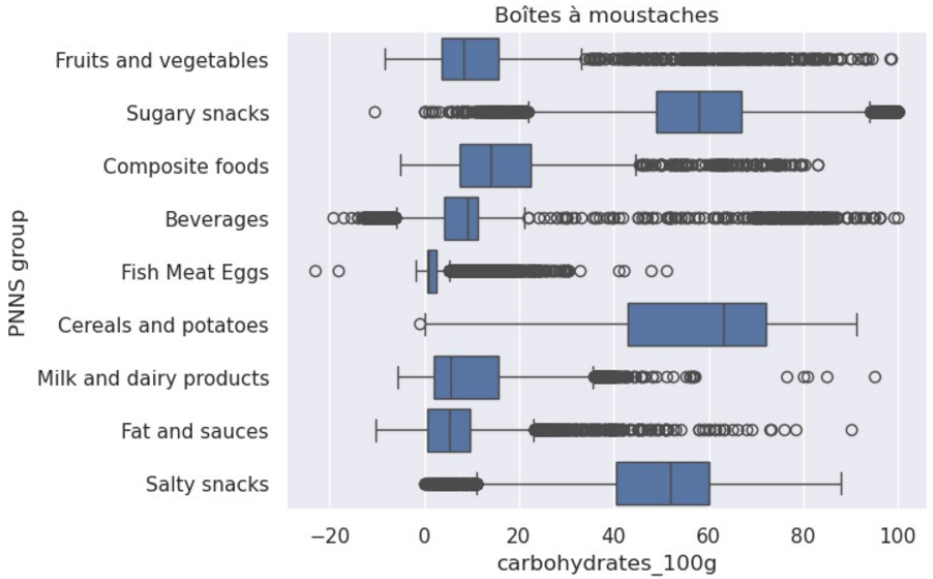




Exploration – Analyse de variance

- Boîtes à moustaches :
  - A gauche l'énergie
  - A droite les carbohydrates





Exploration – Analyse de variance

- Résultats des tests :
  - A gauche l'énergie
  - A droite les carbohydrates

OLS Regression Results

Dep. Variable:	energy_100g	R-squared:	0.535
Model:	OLS	Adj. R-squared:	0.535
Method:	Least Squares	F-statistic:	5732.
Date:	Fri, 15 Nov 2024	Prob (F-statistic):	0.00
Time:	10:55:27	Log-Likelihood:	-3.0647e+05
No. Observations:	39831	AIC:	6.130e+05
Df Residuals:	39822	BIC:	6.130e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	247.7091	8.193	30.236	0.000	231.652	263.767
pnns_groups_1[T.Cereals and potatoes]	1134.5242	11.004	103.097	0.000	1112.955	1156.093
pnns_groups_1[T.Composite foods]	453.9133	11.166	40.653	0.000	432.028	475.798
pnns_groups_1[T.Fat and sauces]	1306.6815	13.367	97.753	0.000	1280.482	1332.881
pnns_groups_1[T.Fish Meat Eggs]	604.2781	11.589	52.144	0.000	581.564	626.992
pnns_groups_1[T.Fruits and vegetables]	153.3547	12.588	12.182	0.000	128.681	178.028
pnns_groups_1[T.Milk and dairy products]	565.0025	10.847	52.086	0.000	543.741	586.264
pnns_groups_1[T.Salty snacks]	1829.1647	14.926	122.552	0.000	1799.910	1858.419
pnns_groups_1[T.Sugary snacks]	1566.3429	10.064	155.631	0.000	1546.616	1586.070

Omnibus:	3635.985	Durbin-Watson:	1.311
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10817.054
Skew:	0.490	Prob(JB):	0.00
Kurtosis:	5.358	Cond. No.	10.3

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	sum_sq	df	F	PR(>F)
pnns_groups_1	1.294694e+10	8.0	5731.500835	0.0
Residual	1.124428e+10	39822.0	NaN	NaN

OLS Regression Results

Dep. Variable:	carbohydrates_100g	R-squared:	0.678
Model:	OLS	Adj. R-squared:	0.678
Method:	Least Squares	F-statistic:	1.047e+04
Date:	Fri, 15 Nov 2024	Prob (F-statistic):	0.00
Time:	10:55:29	Log-Likelihood:	-1.6531e+05
No. Observations:	39831	AIC:	3.306e+05
Df Residuals:	39822	BIC:	3.307e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.1454	0.237	47.084	0.000	10.681	11.609
pnns_groups_1[T.Cereals and potatoes]	43.9351	0.318	138.178	0.000	43.312	44.558
pnns_groups_1[T.Composite foods]	4.9256	0.323	15.267	0.000	4.293	5.558
pnns_groups_1[T.Fat and sauces]	-2.9739	0.386	-7.700	0.000	-3.731	-2.217
pnns_groups_1[T.Fish Meat Eggs]	-8.4996	0.335	-25.384	0.000	-9.156	-7.843
pnns_groups_1[T.Fruits and vegetables]	3.9441	0.364	10.844	0.000	3.231	4.657
pnns_groups_1[T.Milk and dairy products]	-1.0376	0.313	-3.310	0.001	-1.652	-0.423
pnns_groups_1[T.Salty snacks]	36.2420	0.431	84.037	0.000	35.397	37.087
pnns_groups_1[T.Sugary snacks]	46.6591	0.291	160.449	0.000	46.089	47.229

Omnibus:	5710.941	Durbin-Watson:	1.395
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30278.949
Skew:	0.589	Prob(JB):	0.00
Kurtosis:	7.106	Cond. No.	10.3

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	sum_sq	df	F	PR(>F)
pnns_groups_1	1.975064e+07	8.0	10472.910685	0.0
Residual	9.387434e+06	39822.0	NaN	NaN

Exploration – Classifications supervisées

- Régression logistique polynomiale :

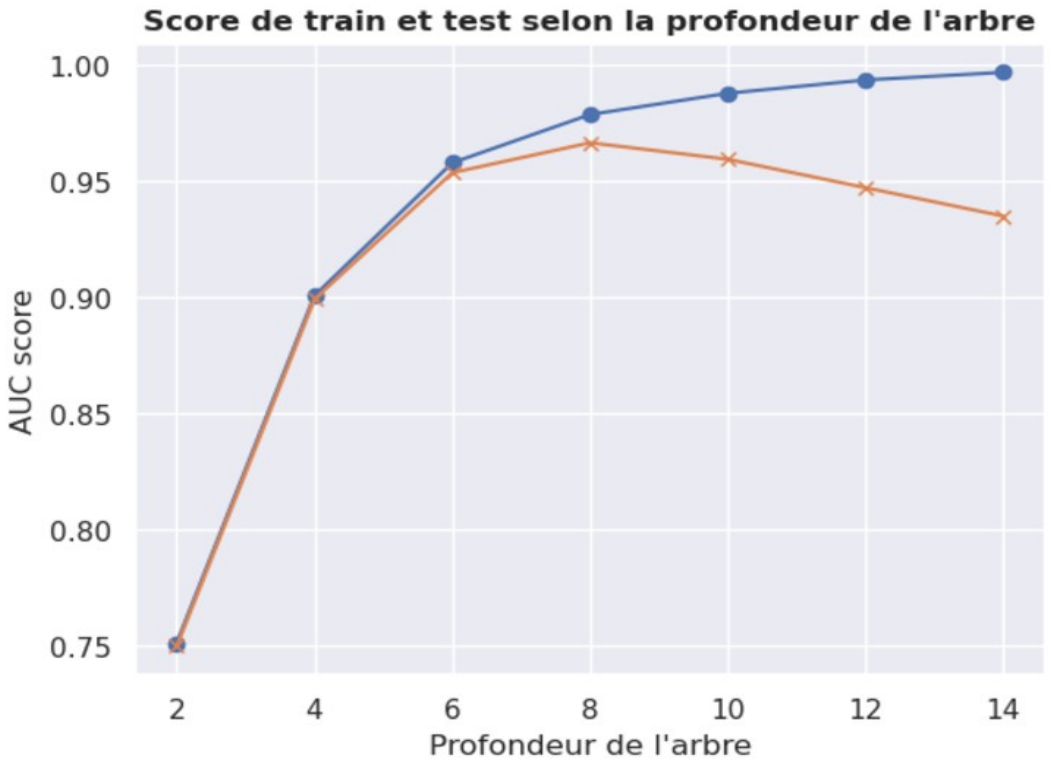
```
La performance du Modèle pour le set de Training
------------------------------------------------
le score est 0.8168465980416771


La performance du Modèle pour le set de Test
--------------------------------------------
le score est 0.8152378561566461
```

```
ROC-AUC train 0.9699811556593525
ROC-AUC test 0.9686293820858565
```

Exploration – Classifications supervisées

- Arbre de décision :



Exploration – Classifications supervisées

- Arbre de décision :

```
La performance du Modèle pour le set de Training
----------------------------------------------
le score est 0.8551029374843083


La performance du Modèle pour le set de Test
----------------------------------------------
le score est 0.8392117484624074
```

```
ROC-AUC train 0.9784815403988705
ROC-AUC test 0.9652567391922423
```

Conclusions

- Forte corrélation entre les variables du dataset
- Prédictions possibles avec une bonne certitude
- Il est possible de développer l’application d’autocomplétion
- Premières modélisations à améliorer en explorant d’autres types de modèles
- Attention particulière à apporter aux variables moins liées aux qualités nutritionnelles