PROJET 3 : CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

Parcours Data sciences | ABBOUD Marwa | 14 novembre 2020

Encadrant :Bertrand Beaufils

Evaluateur : Sarah Ghidalia

Sommaire

1. **Idée d'application**
2. **Etude exploratoire**
3. **Analyse statistique multivariée**
4. **Application**
5. **Conclusion**

Idée d'application

Prédiction du **Nutriscore**

À partir des indicateurs nutritionnels

Qualité nutritionnelle d'un produit

Santé publique France

Objectif de l'application

Prédire le nutriscore d'un produit qu'on n'aurait que quelques informations nutritionnelle sur ce produit

- **Source des données Open Food Fact : https://world.openfoodfacts.org/**

Etude exploratoire : Nettoyage de jeu de données

- Base de données regroupant les produits avec différents **Indicateurs Nutritionnelle**

Processus de nettoyage

1. Etudes des indicateurs nutritionnelle pertinents
2. Etude des valeurs manquantes
3. Détection des valeurs aberrantes
4. Imputation des valeurs manquantes par la méthodes des K plus proches voisins

Etude exploratoire :

1. Etudes des indicateurs nutritionnelle pertinents :

Supprime les dates

Nutriscore de Pays : **France Uniquement**

Supprime toutes les colonnes **sans nutriscore renseigné**

Sélection les colonnes ayant **un Taux de 100g**

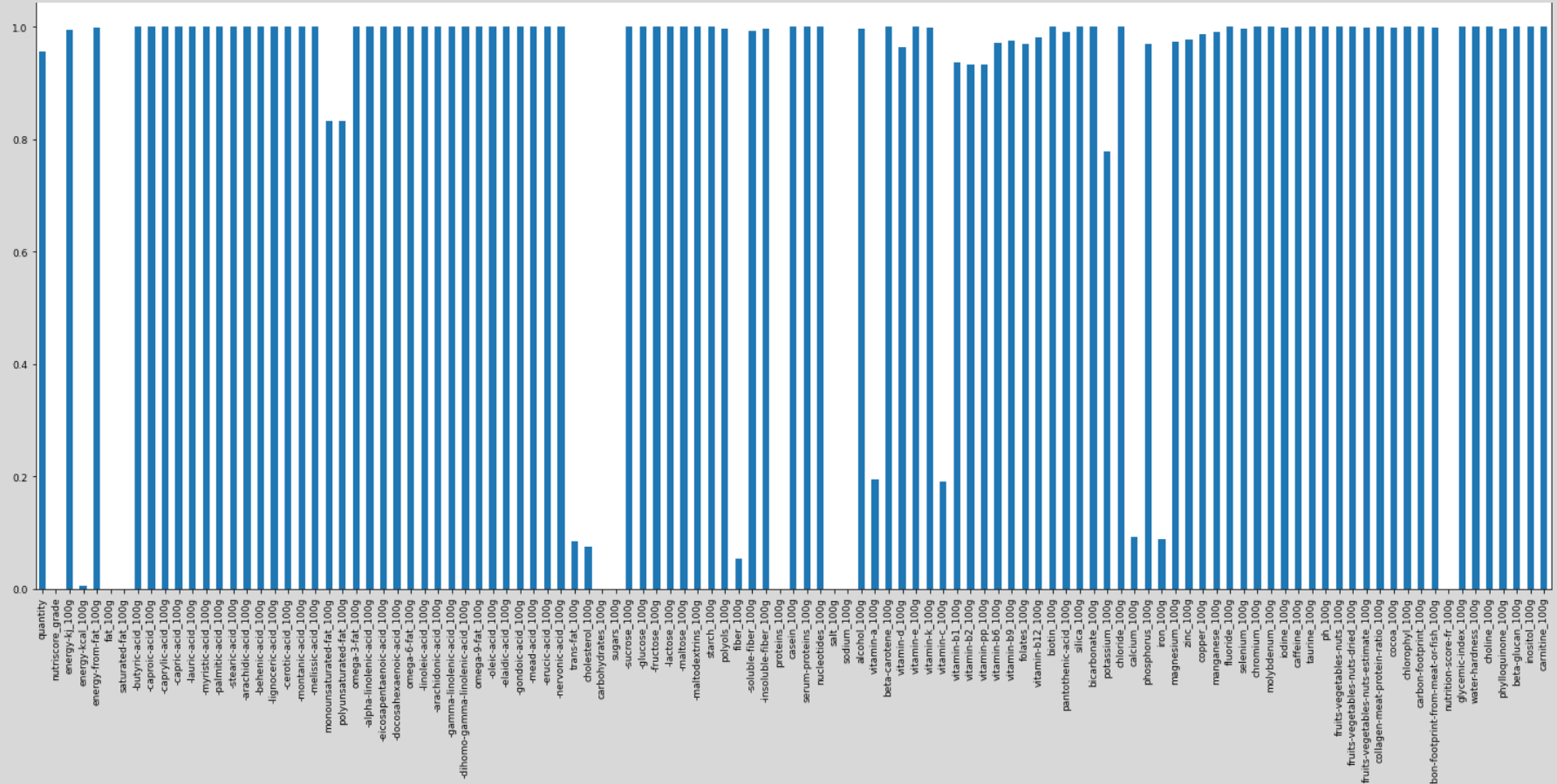
Dimensions initiale du jeu de données (16959, 181) → Dimensions (10483, 112)

2. Etude des valeurs manquantes

- Calcul le **Taux de NaN** uniquement pour les aliments qui ont **un nutriscore renseigné**

Etude exploratoire :

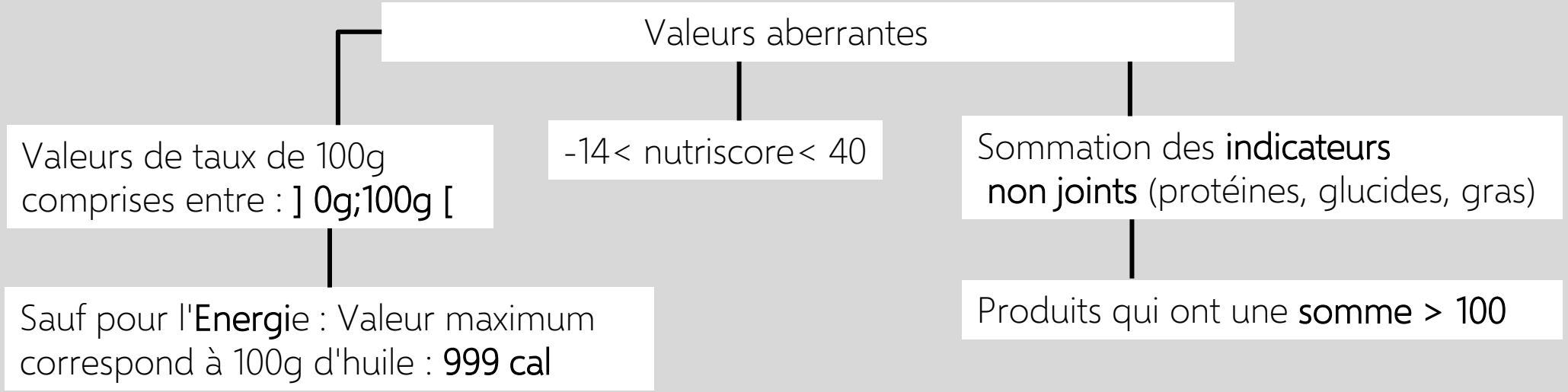
2. Etude des **valeurs manquantes**



- Sélectionner les colonnes qui ont **au moins 25% de valeurs renseignées :** Nouvel dimensions : (10483,17)

Etude exploratoire :

3. Détection des valeurs aberrantes



Etude exploratoire :

Bilan du nettoyage

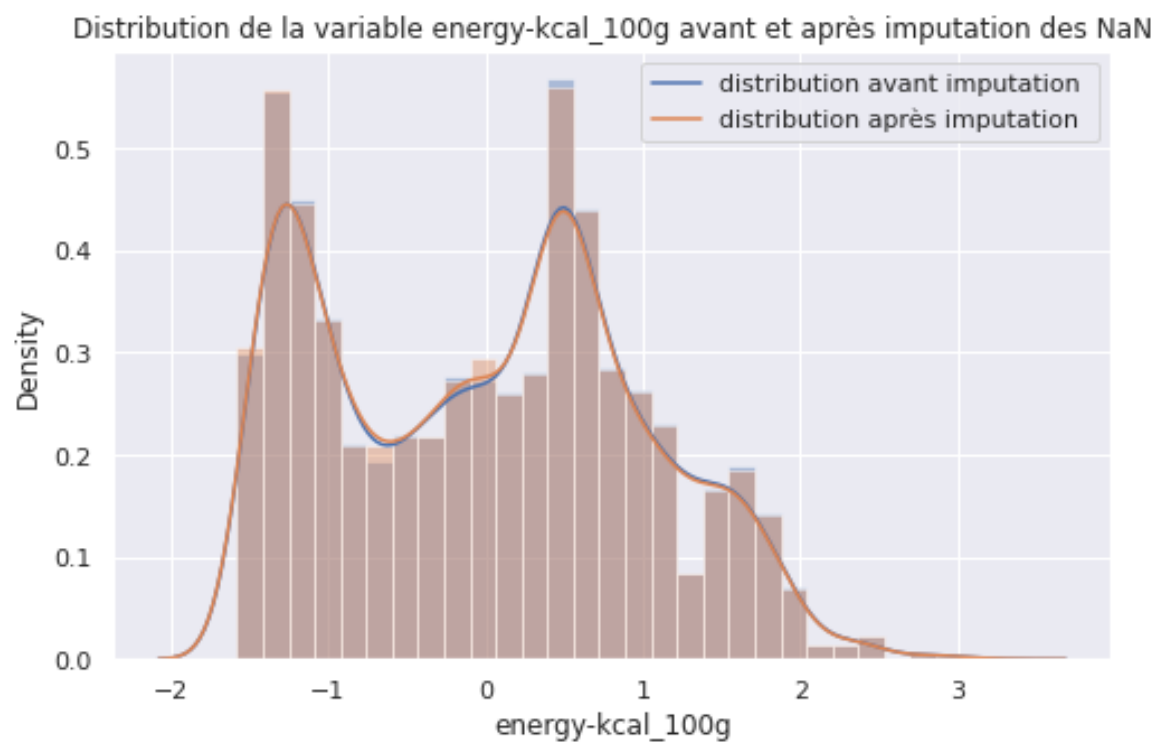
- 181 colonnes réduites à 17 colonnes
- 16959 lignes réduites à 10097 lignes

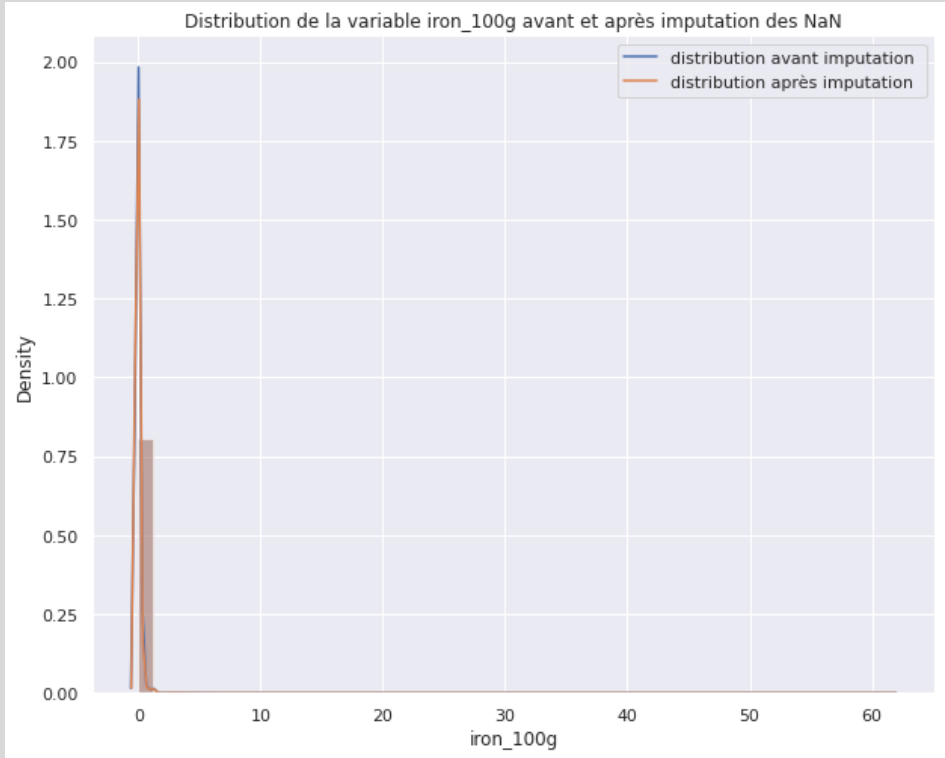
4-Imputation des valeurs manquantes par la méthodes des **K plus proches voisins**

Différentes unités → Centrage et réduction : Standard Scaler → KNNImputer(missing_values=np.nan, n_neighbors=3)

Etude exploratoire :

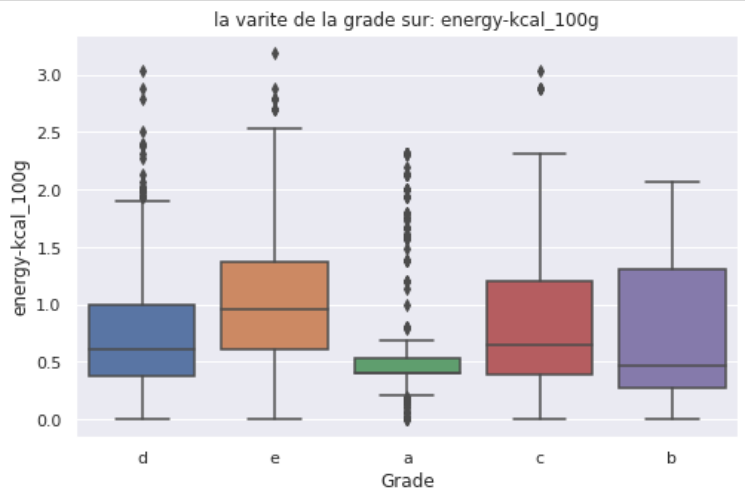
4-Imputation des valeurs manquantes par la méthodes des **K plus proches voisins**

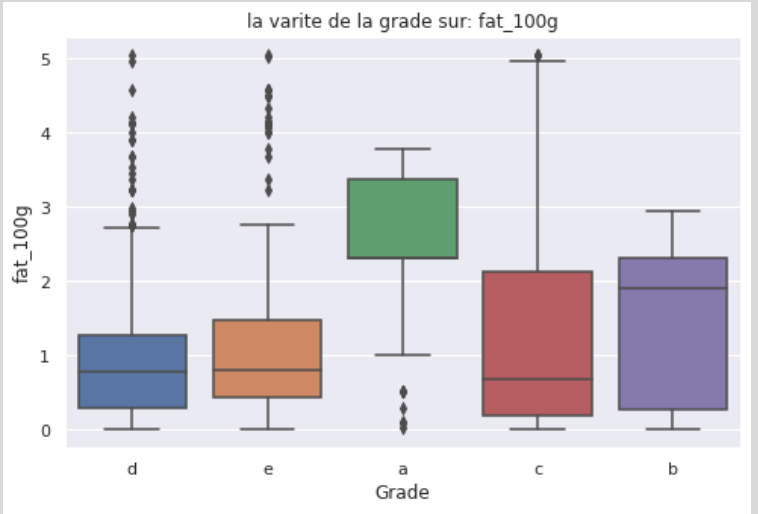


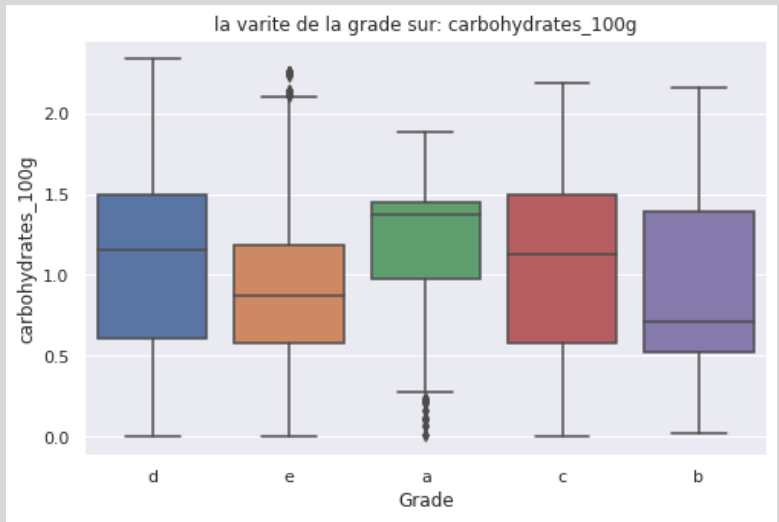


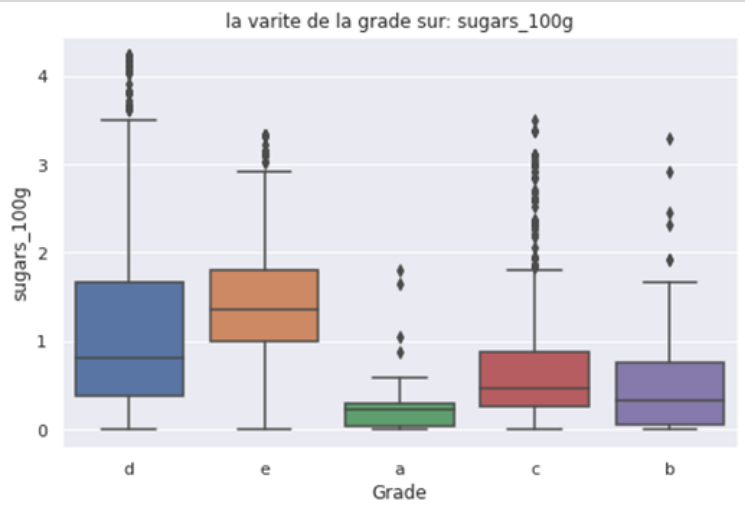
Analyse multivariée :

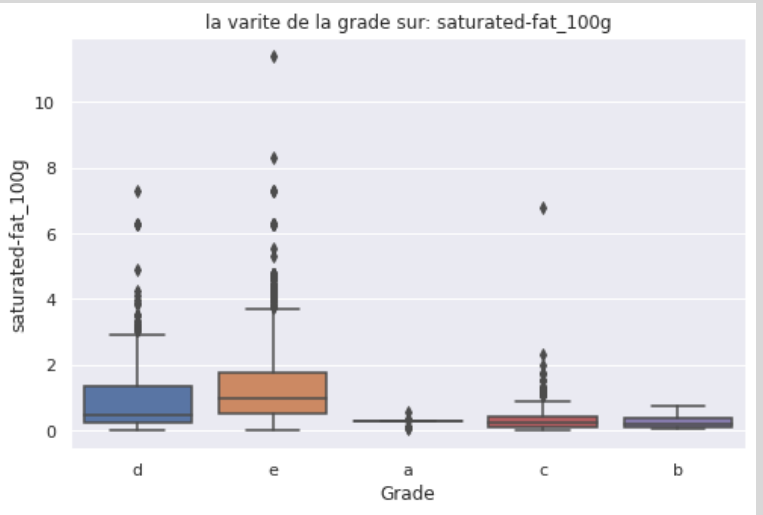
Effet de chaque variable sur le nutri-grade



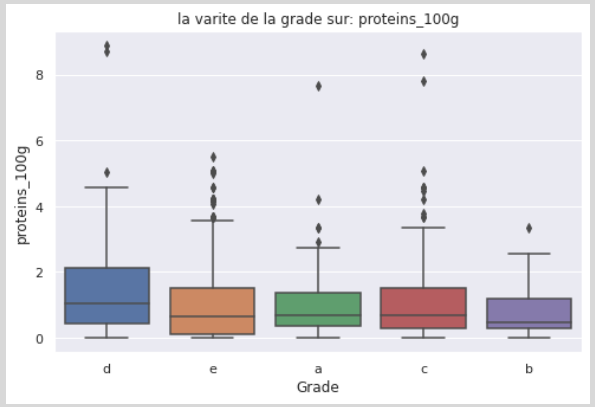


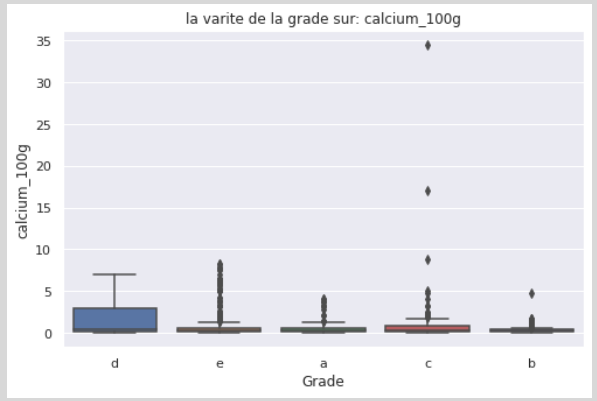


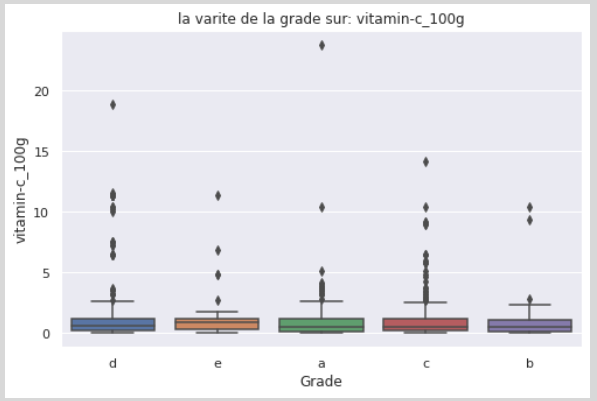


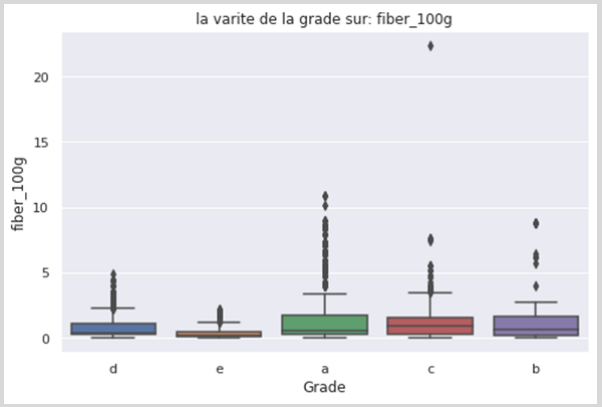


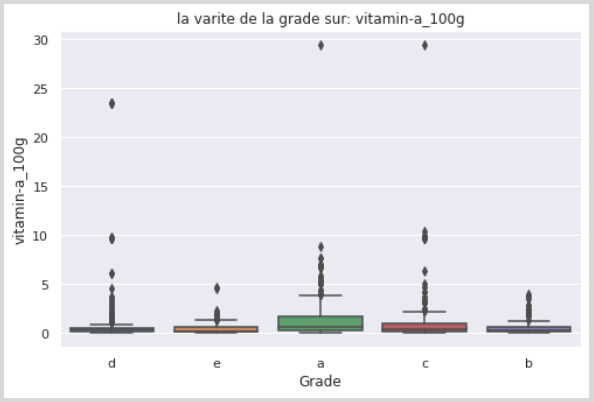
Analyse multivariée :

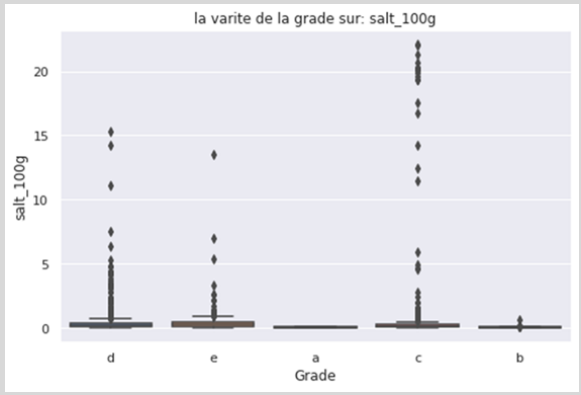


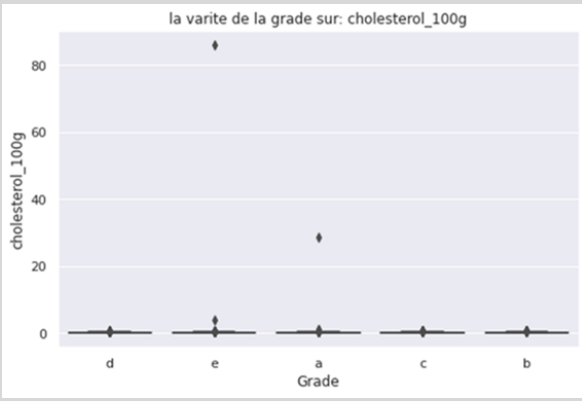






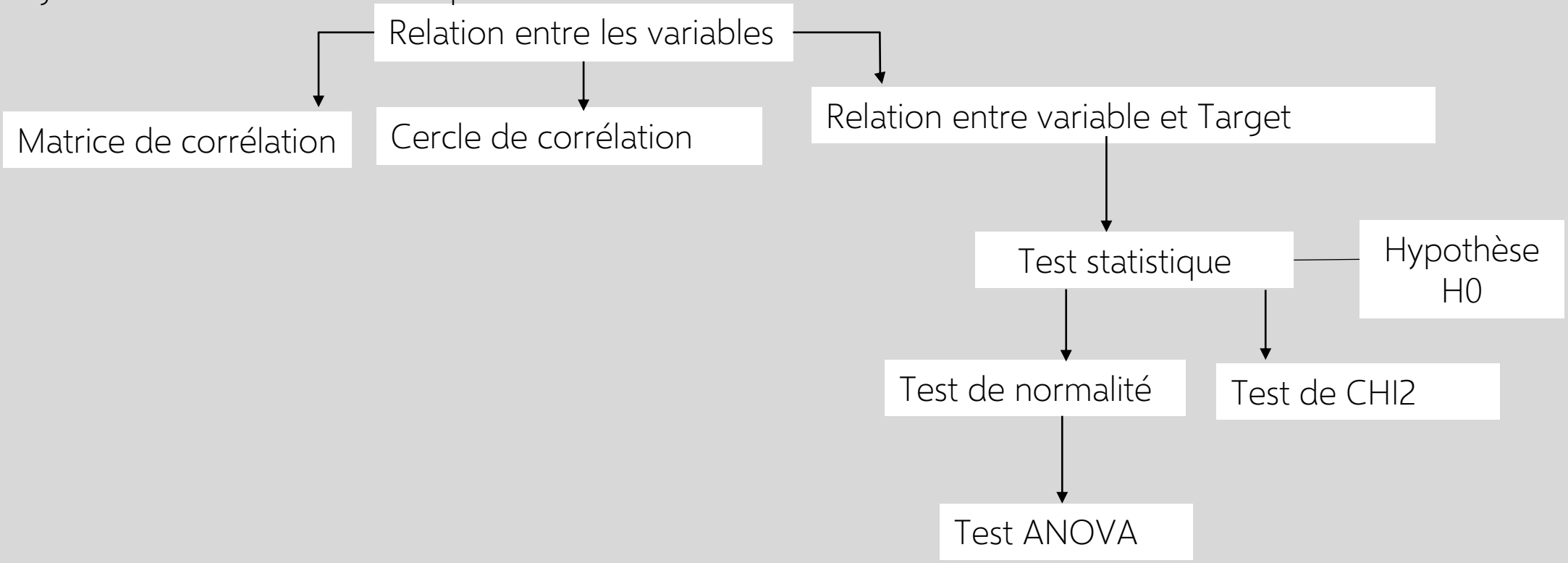






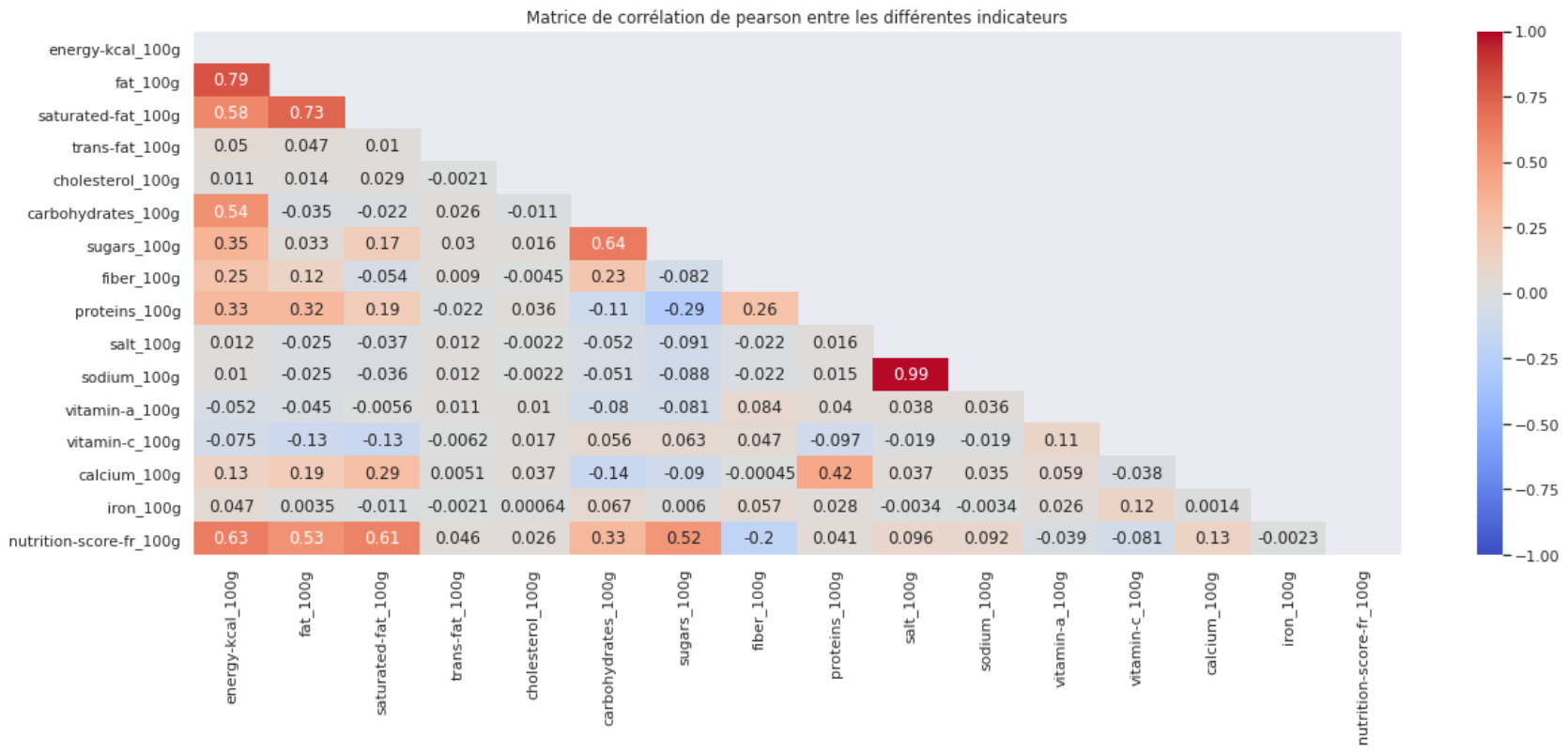
Analyse multivariée :

- Décrire le jeu de données afin de comprendre les relations entre les indicateurs



Analyse multivariée :

1. Matrices de corrélations



Vitamin a et c, iron, cholesterol, transfat :
Pas de corrélation remarquable

nutriscore corrélés avec :
energy, fat, saturated-fat et sugars

Fat et energy sont corrélés

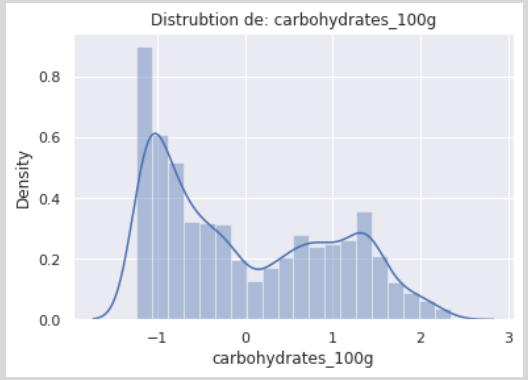
sodium et salt sont fortement corrélés

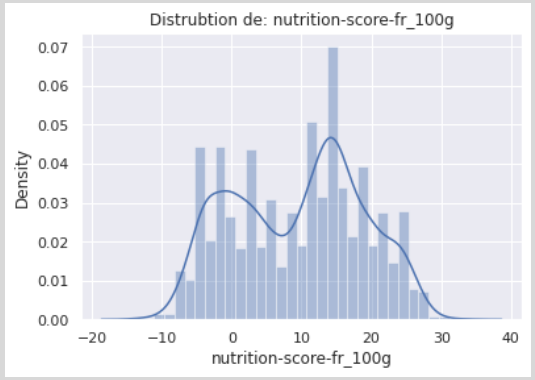
saturated-fat corrélés avec :
Fat et energy

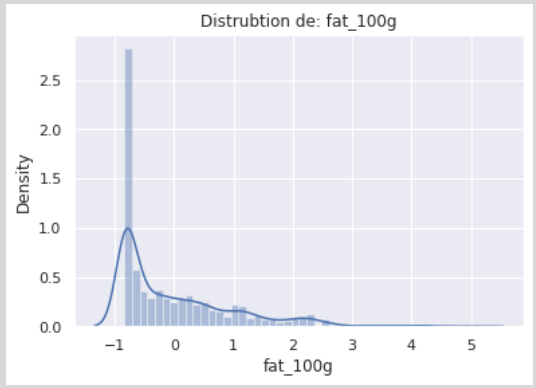
carbohydrates corrélés avec :
energy et sugars

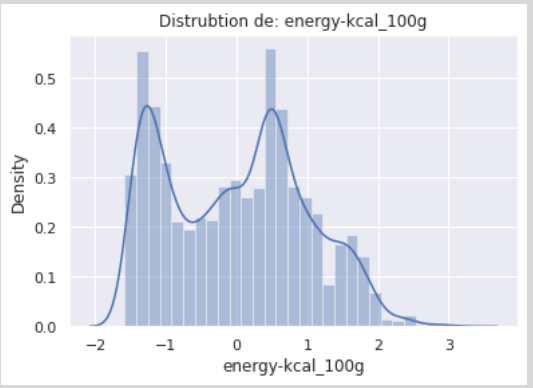
Analyse multivariée :

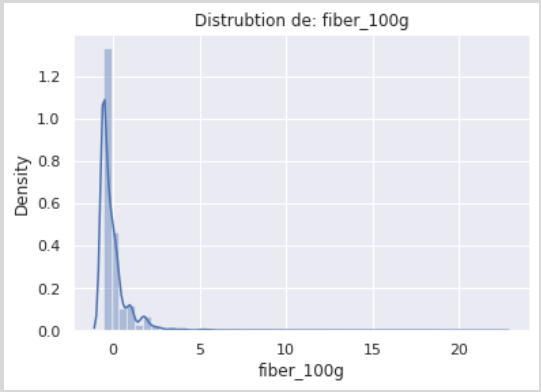
2.1 Distributions : ne suivent pas la distribution normale

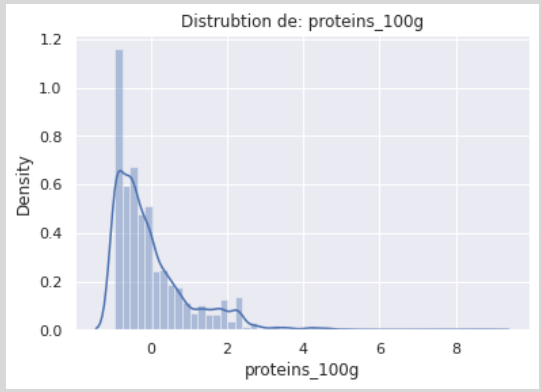


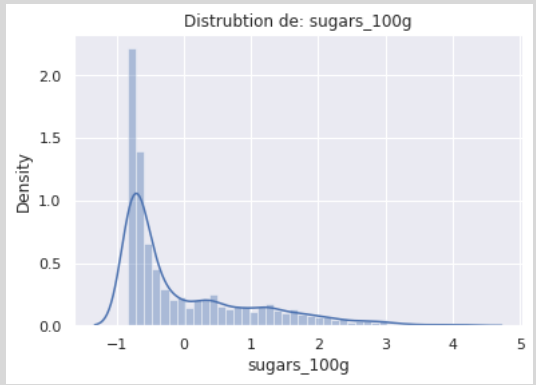


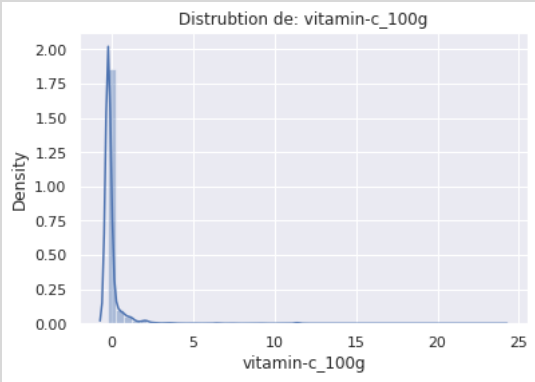












- **Test de Normalite** ”Kolmogorov Smirnov (K-S)” : Pour tous les indicateurs p-value < alpha

Confirme la non normalité des distributions de données

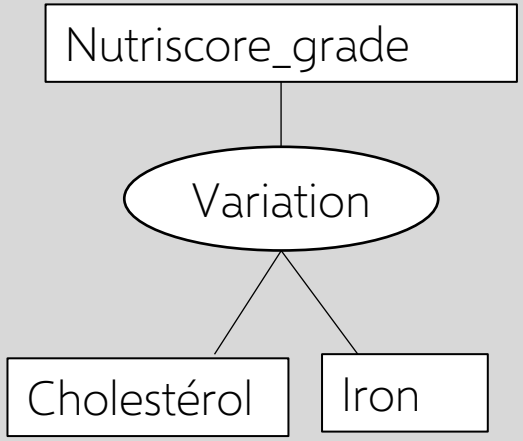
Analyse multivariée :

2.2 Test d'analyse de variance ANOVA

- En l'absence de validation de l'hypothèse de normalité, on ne peut pas applique l'analyse de la variance

→ Influence de la grade de nutriscore sur les variables ?

Plupart de variables étudies ont une distribution qui change en fonction des valeurs de nutriscore_grade sauf dans le cas de :



Indicateurs	p-val	test
energy-kcal_100g	0.000000e+00	True
fat_100g	0.000000e+00	True
saturated-fat_100g	0.000000e+00	True
trans-fat_100g	3.741685e-07	True
cholesterol_100g	1.031522e-01	False
carbohydrates_100g	1.662625e-179	True
sugars_100g	0.000000e+00	True
fiber_100g	6.012651e-79	True
proteins_100g	2.567503e-31	True
salt_100g	4.429233e-31	True
vitamin-a_100g	7.484492e-07	True
vitamin-c_100g	1.125143e-12	True
calcium_100g	5.020524e-40	True
iron_100g	8.227479e-01	False
nutrition-score-fr_100g	0.000000e+00	True

Analyse multivariée :

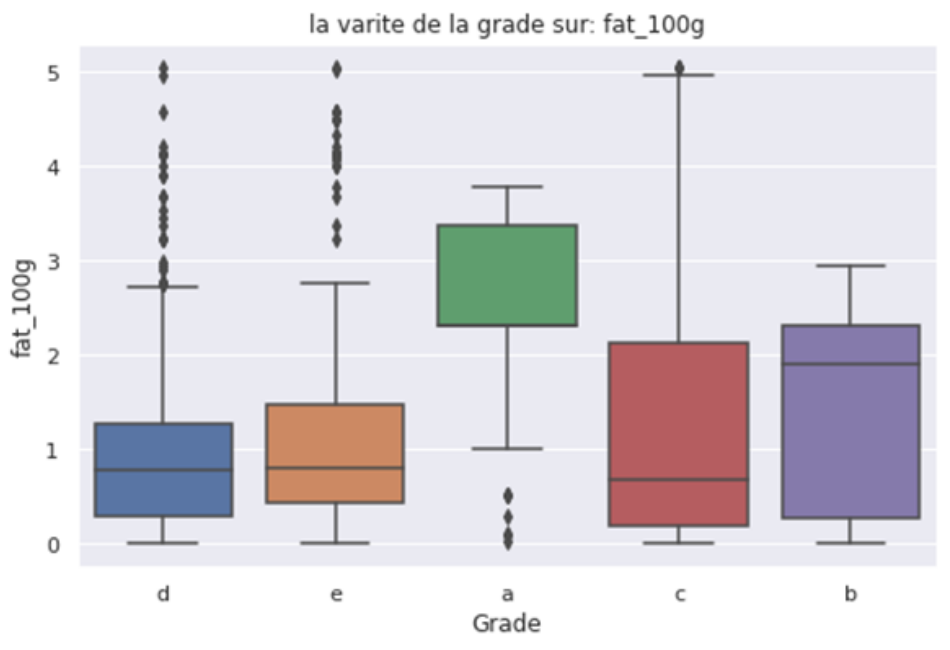
3. Test du CHI2

p-value <= alpha

Variables non indépendantes

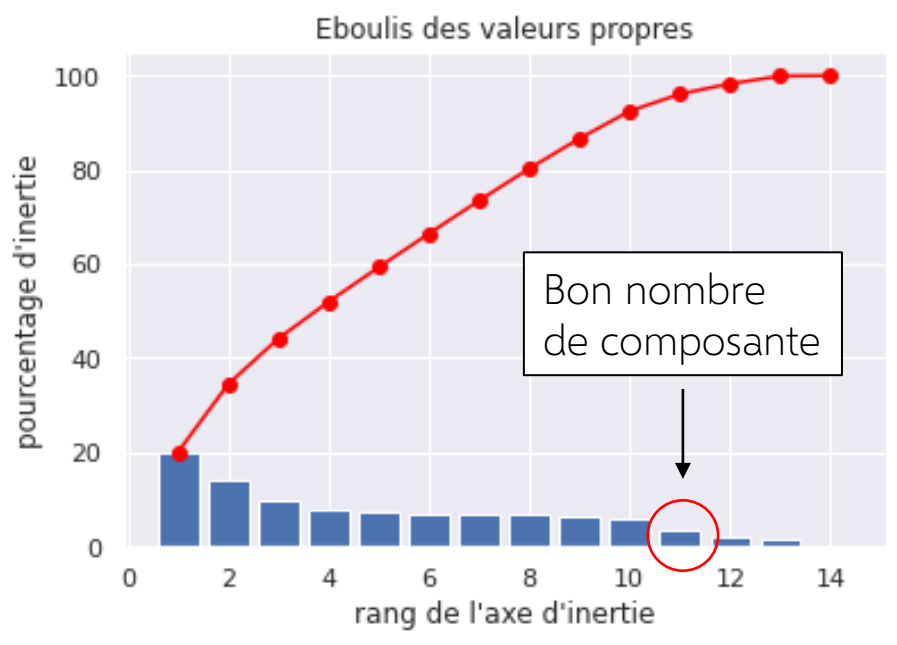
- Le test du CHI2 conclut au rejet de l'hypothèse d'indépendance entre fat et nutri-grade:

lien entre le grade du nutriscore et le fat

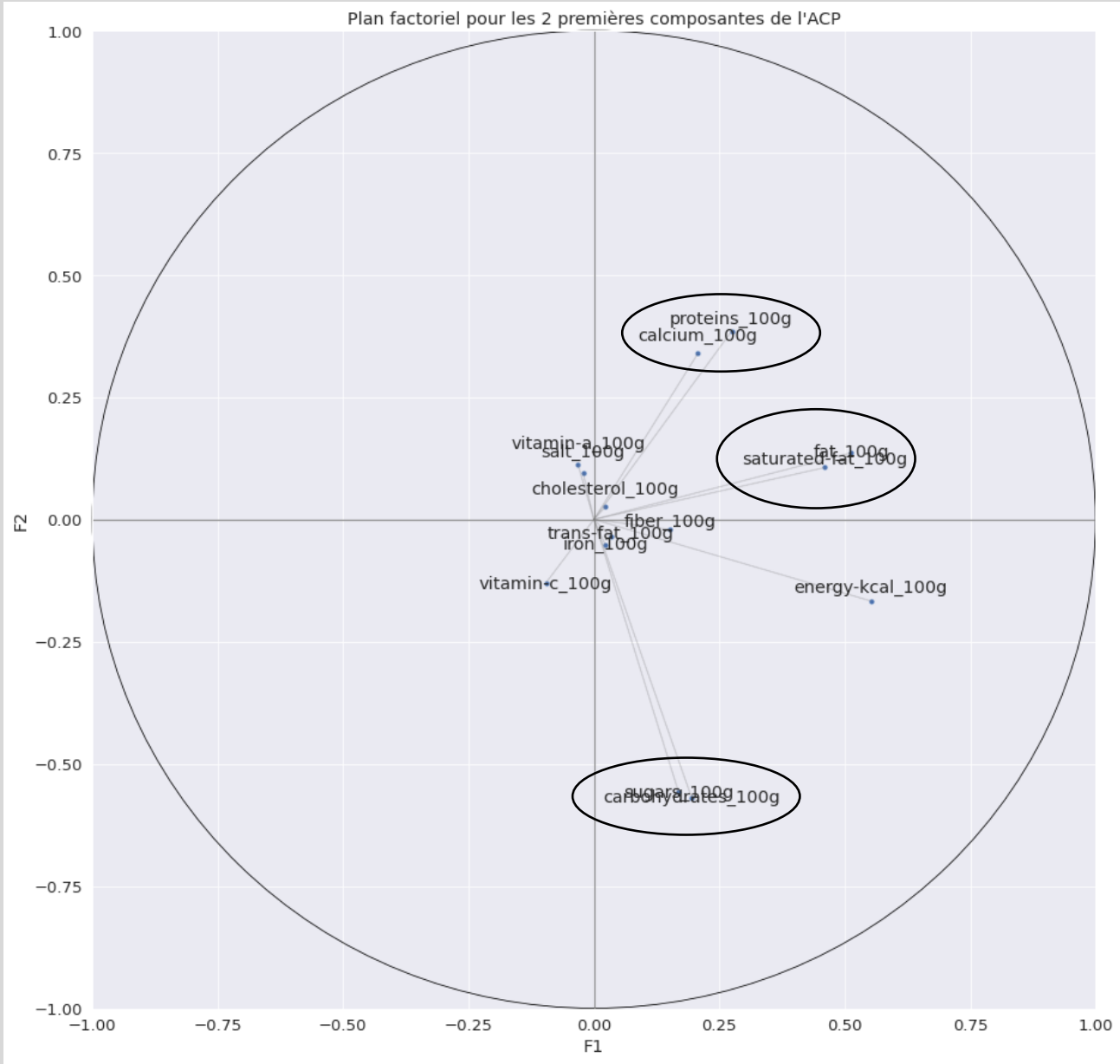


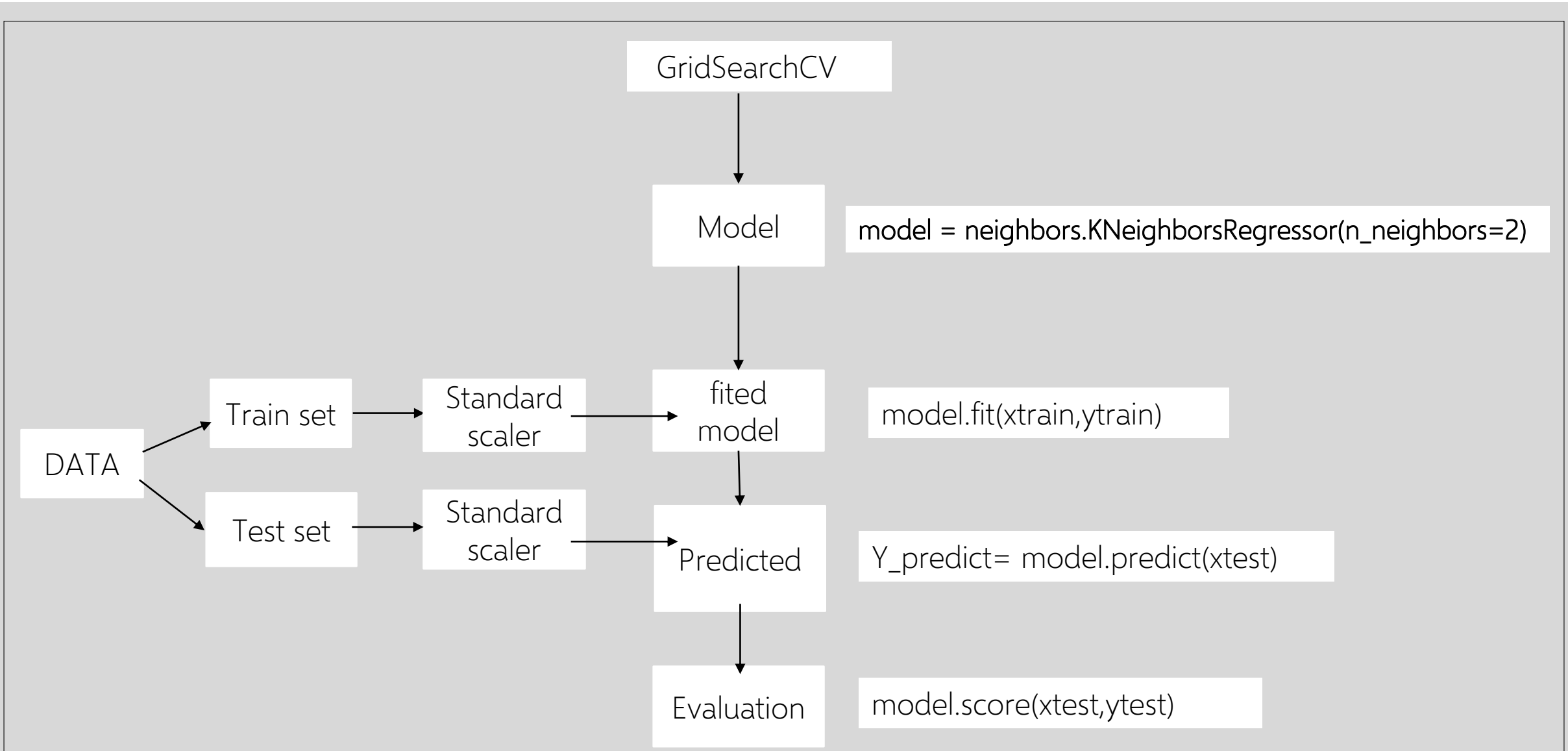
Analyse univariée :

5. Analyse par Composantes Principales (ACP)



- Les deux premières composantes principales F1 et F2 représentent environ 33% de la variance des données

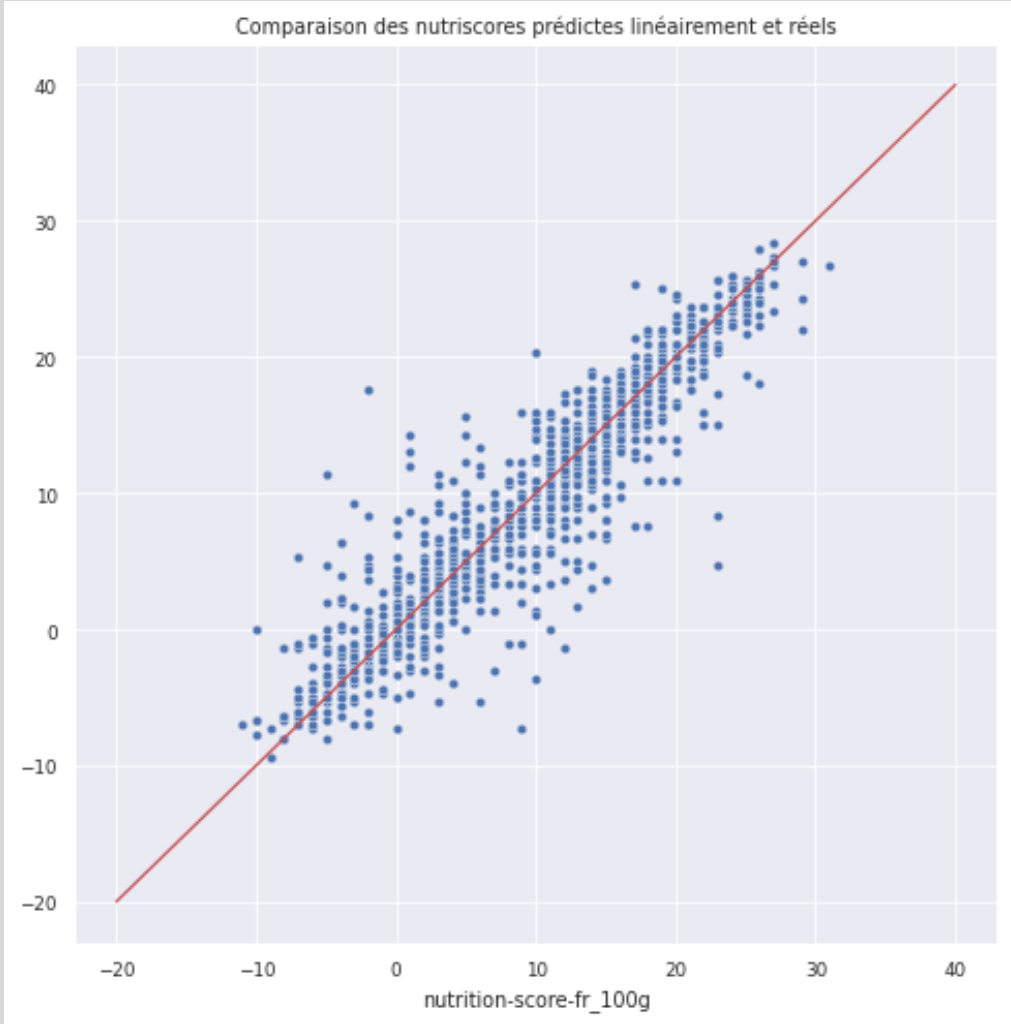



GridSearchCV
Model
model = neighbors.KNeighborsRegressor(n_neighbors=2)
DATA
Train set
Standard scaler
fited model
model.fit(xtrain,ytrain)
Test set
Standard scaler
Predicted
Y_predict= model.predict(xtest)
Evaluation
model.score(xtest,ytest)

Prédiction du nutriscore par régression KNN

Root Mean Square Error (RMSE) = 2.782

product_name	nutriscore	prediction_nutriscore
Fruit bars	-2.0	-7.000000
Hearts of palm in brine	-4.0	4.000000
Ice cream	11.0	11.000000
Enriched noodle product, wide egg noodles	-4.0	-4.000000
Petite diced tomatoes	-4.0	-3.666667
...	...	...
Super shreds super foods, brussels sprouts shreds	-9.0	-9.333333
Roundy's, toaster pastries, frosted chocolate fudge	18.0	18.000000
Roundy's, pilaf rice mix	11.0	10.666667
Chocolat stella, lait milk chocolate	11.0	16.000000
Banana & Berries	-4.0	-3.000000



Comparaison de la performance entre le model de :régression et classification

Détermine le nutriscore

Régression KNN

Pour **n_neighbors=3**
RMSE sur le jeu de test : 2.782

Régression linéaire

- Coefficient de détermination
R^2 sur jeu de test : 0.663
- RMSE sur le jeu du test: 5.375

Détermine le grade de nutriscore

Classification K-Nearest Neighbour (KNN)

Pour **n_neighbors=3**
accuracy sur le jeu de test :**0.82**

```
[[277  20  10   2   0]
 [ 34  89  21   5   1]
 [ 21  32 218  34   1]
 [  4   6  31 446  24]
 [  2   0   5  35 298]]
```

→ Sélectionner le meilleur modèle : Régression KNN avec **n_neighbors** = 3

Identifier le nutriscore du nouveau produit renseigne à l'aide de notre model

Nutriscore : D

15.67

Produit :


```
prediction(energy=408,fat=14.1,saturatedfat=6.9,transfat=np.nan,cholesterol=np.nan,carbohydrates=62.4,
sugars=24.9,fiber=6,proteins=7.3,salt= 0.44,vitamina=np.nan,vitaminc=np.nan,calcium=1.4,iron=0.01)
```

Nombre de valeur renseigné	Indicateurs renseigné	Nutriscore
3	Stat-fat	20
4	+Fat	23.6
5	+Proteins	23.6
6	+Fibre	18
7	+Sugar	23
8	+Iron	23
9	+Calcium	23
10	+Carbohydrate	14
11	+Energy	15.67
12	+Salt	15.66

Conclusion:

- Indication de nutriscore approché (RMSE= 2.7) avec 14 variables : **faisabilité de l'application**
- Résultats de l'application cohérents avec ce qui est observé dans l'analyse multivariée
- **Variables les plus pertinents pour l'application : energy, fat, sat-fat, carbohydrates et sugars.**
- Résultats cohérents avec les principes nutritionnels:

Indicateur renseigne	Nutriscore
Fat, energy	15.67
Energy, sugars	15.66
Fat, sugars	14
Sta-fat, sugars	14

Merci de votre attention

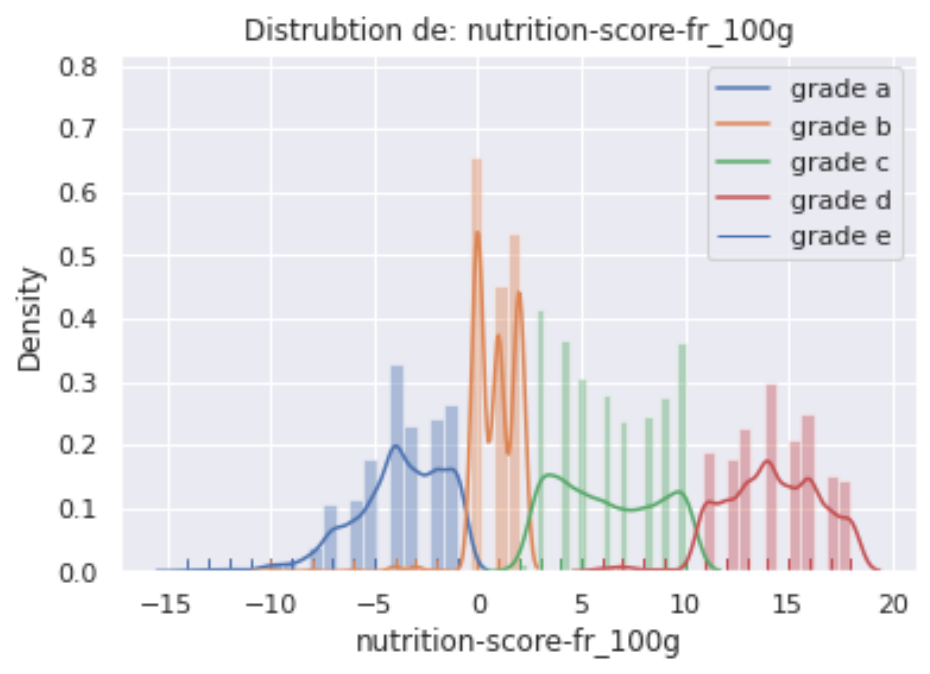
Prédiction du nutriscore par Régression linéaire

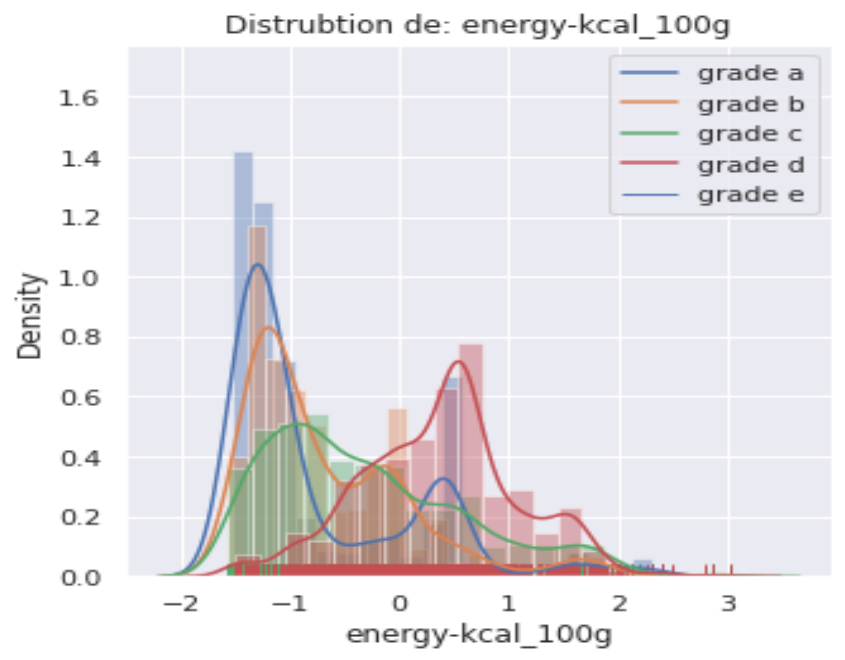
Tableaux représente le poids de chaque variable dans la régression

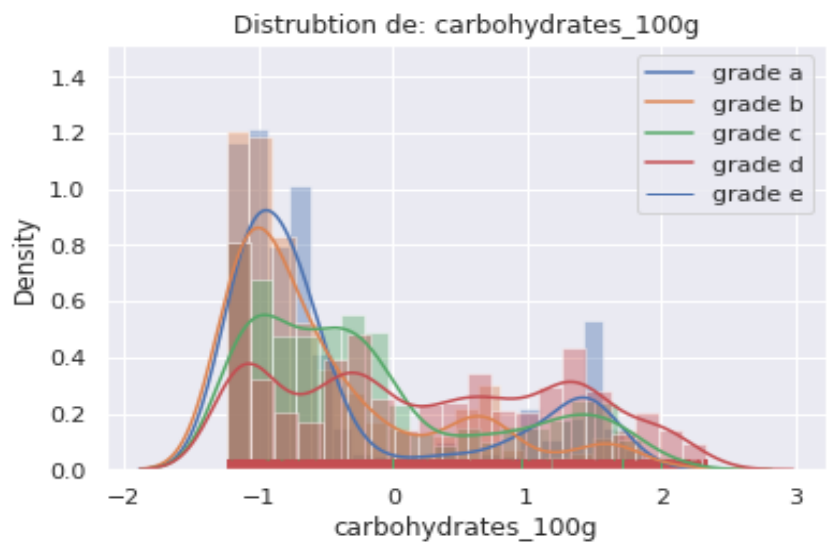
energy	fat	saturated-fat	trans-fat	cholesterol	carbohydrates	sugars	fiber	proteins	salt	vitamin-a	vitamin-c	calcium	iron
3.735	0.296	2.563	0.094	0.047	-0.27	3.2	-2.217	-0.023	1.37	0.16	-0.153	0.09	-0.024

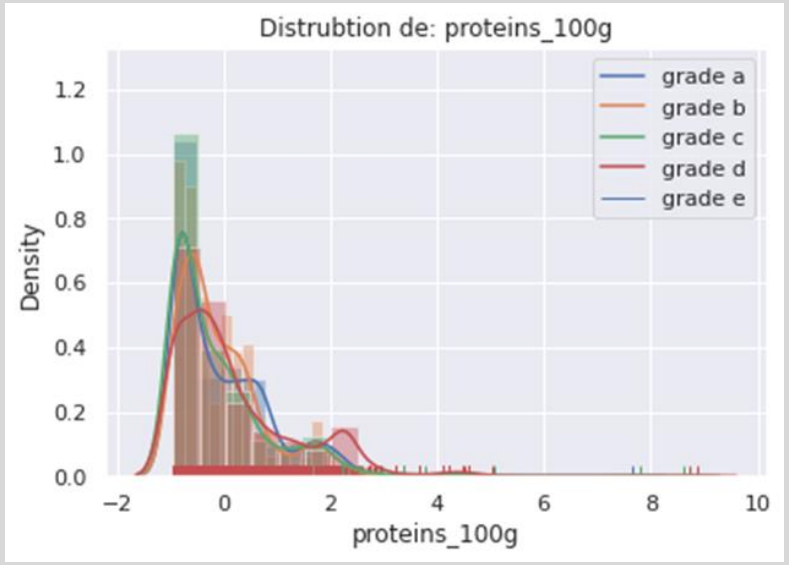
Ces résultats sont cohérents avec ce qui est observé dans l'analyse multivariée sauf pour le carbohydrates.

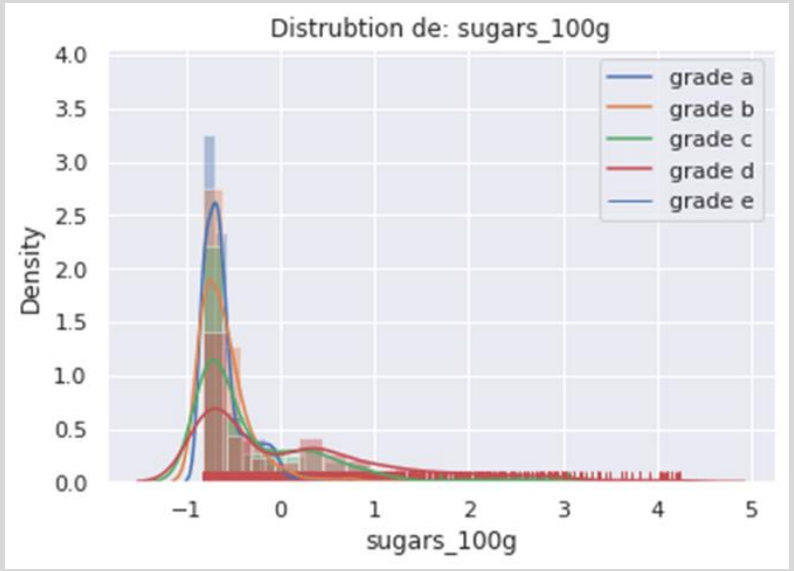
Analyse multivariée :

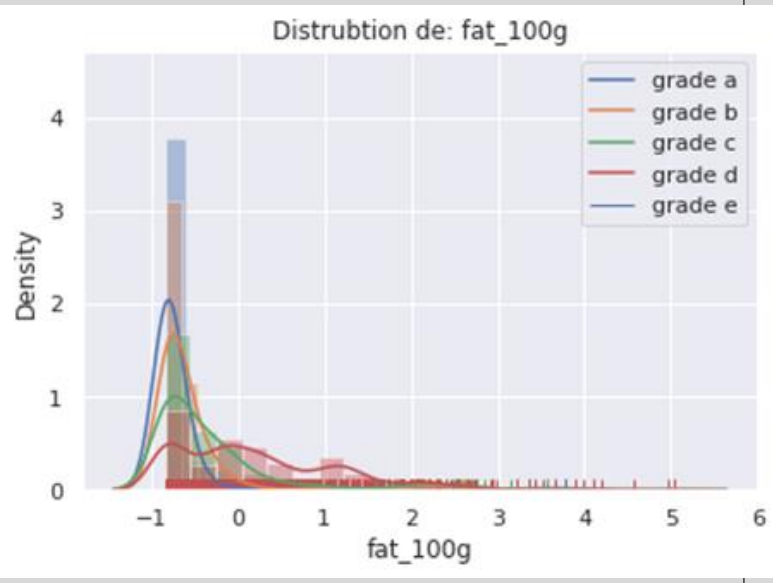


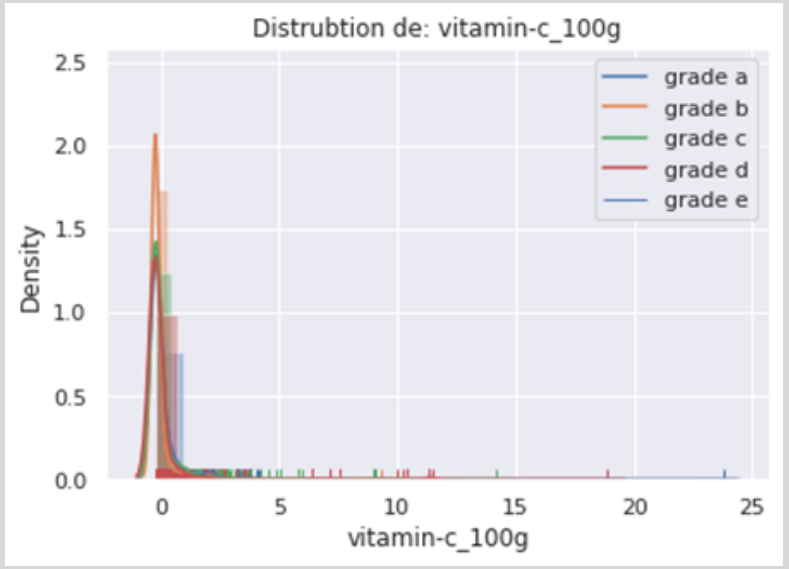


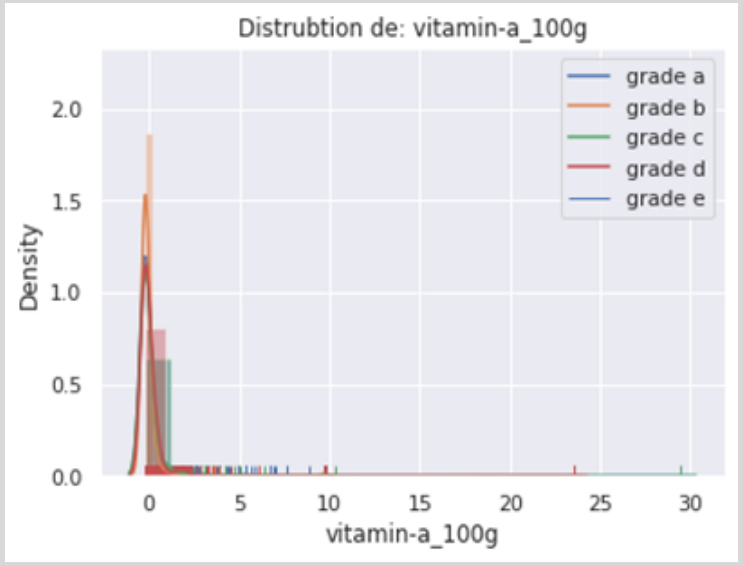


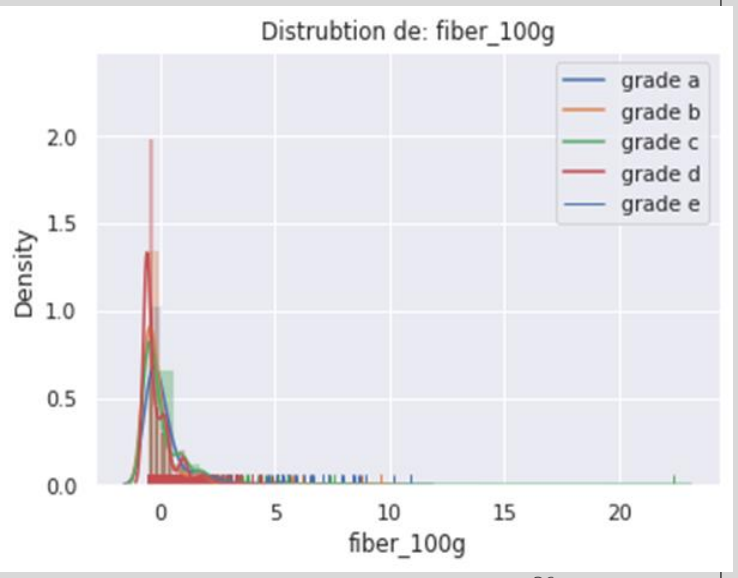

Distrubtion de: proteins_100g
grade a
grade b
grade c
grade d
grade e
Density
proteins_100g


Distrubtion de: sugars_100g
grade a
grade b
grade c
grade d
grade e
Density
sugars_100g


Distrubtion de: fat_100g
grade a
grade b
grade c
grade d
grade e
Density
fat_100g


Distrubtion de: vitamin-c_100g
grade a
grade b
grade c
grade d
grade e
Density
vitamin-c_100g


Distrubtion de: vitamin-a_100g
grade a
grade b
grade c
grade d
grade e
Density
vitamin-a_100g


Distrubtion de: fiber_100g
grade a
grade b
grade c
grade d
grade e
Density
fiber_100g