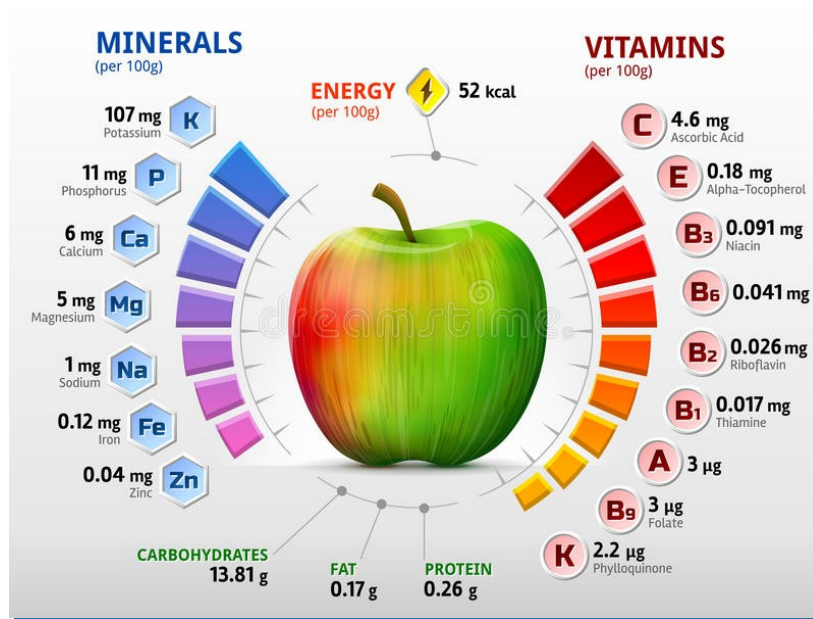


Concevez une application au service de la santé publique

« Je choisi mon produit selon mes priorités »



- Classer les produits selon :

- la catégorie : fruits, biscuits...
- le taux de glucide
- le taux de lipide
- le taux de protéine
- les vitamines et minéraux
- le nutriscore, les fibres etc ..

- Afficher la répartition visuelle des aliments

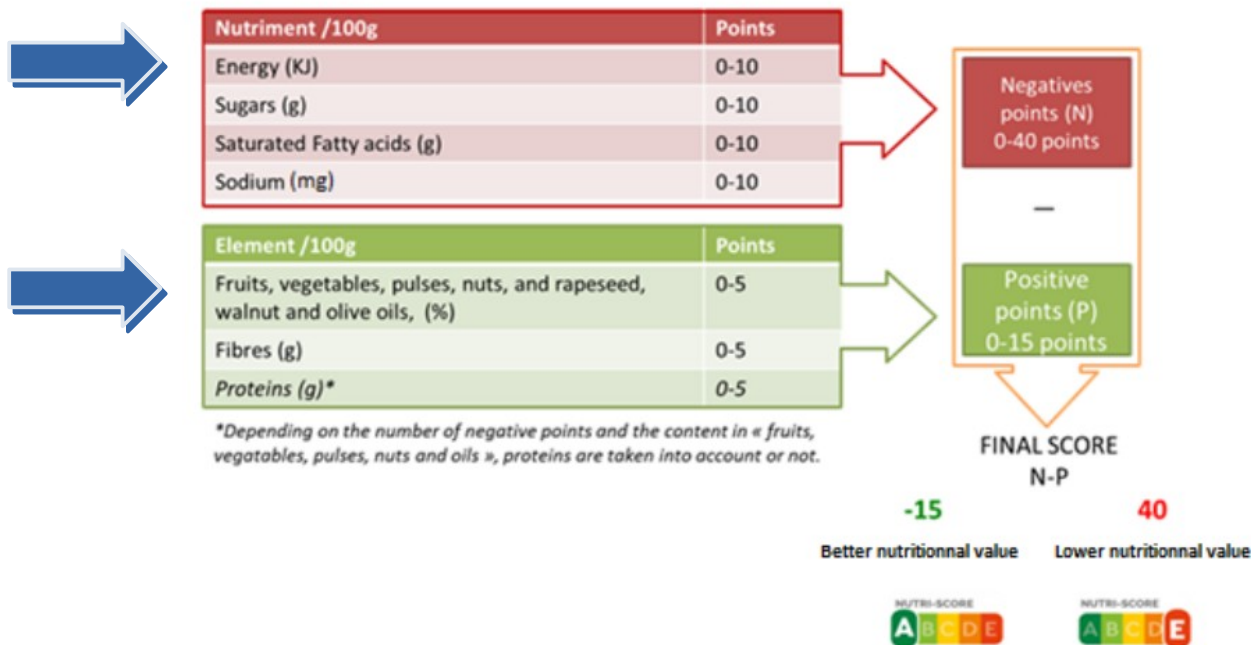
Nutriscore et Nutriscore_grade

Variables Brutes :

- Calories (Kcal/KJ)
- fat (g)
- saturated fatty acids (g)
- carbohydrates (g)
- sugars (g)
- protein (g)
- salt (mg)
- Fibre (g)

Variable calculée

['fruits-vegetables-nuts_100g'] mais non présente dans le dataset



Opérations de nettoyage

Étape 1 / 8 - Création de nouvelles variables

- **[sugcar_pct]** =
sugars_100g /
carbohydrates_100g
- **satfat_pct** = saturated-
fat_100g / fat_100g
- **['pnns_groups_21']**
remplace ['pnns_groups1'] et
['pnns_groups2']

Points	Ratio
	total saturated fatty acids/lipids (%)
0	<10
1	<16
2	<22
3	<28
4	<34
5	<40
6	<46
7	<52
8	<58
9	<64
10	≥64

Opérations de nettoyage

- **Étape 2 / 8 - Sélection de colonnes**

- Colonnes catégoriques
- Colonnes numériques
- Colonnes du nutri score
- Nouvelles colonnes créées

```
bonnes_cols = [ "brands_tags", "countries_en", "product_name"]
bonnes_cols += [ "energy_100g", \
                 "energy-kj_100g", \
                 "energy-kcal_100g", \
                 "proteins_100g", \
                 "fat_100g", \
                 "carbohydrates_100g", \
                 "sugars_100g", \
                 "salt_100g", \
                 "sodium_100g", \
                 "saturated-fat_100g", \
                 "fiber_100g"]
bonnes_cols += ["nutriscore_score", "nutriscore_grade"]
bonnes_cols += ['sugcar_pct', 'satfat_pct']
bonnes_cols += ['pnns_groups_21' ]
```

The nutritional score is calculated using the nutritional data listed on the package for 100 g of the product, whose nutrients form part of the mandatory nutritional declaration or are included as supplemental information, in accordance with Article 30 of the INCO regulation no. 1169/2011:

- Calories (Kcal/KJ)
- Amount of fat (g)
- Amount of saturated fatty acids (g)
- Amount of carbohydrates (g)
- Amount of sugars (g)
- Amount of protein (g)
- Amount of salt (mg)
- Fibre (g)

Opérations de nettoyage

- **Étape 3 / 7 - Suppression des doublons**

- Pour les lignes présentant des NaN sur toutes les colonnes
- Sur les colonnes ["product_name"] et ["brands_tags"]

- **Étape 4 / 8 - Erreurs de formatage**

["countries_en"]
['nan','unknown, unknown']
['sugcar_pct','satfat_pct'] vers infini

	brands_tags	countries_en
392	kazidomi	Belgium,France
394	kazidomi	Belgium,France
954	kazidomi	Belgium,France
955	kazidomi	Belgium,France
956	kazidomi	Belgium,France

Opérations de nettoyage

- **Étape 5 / 8 - Échantillonnage stratifié**
['pnns_groups_21']

- **Étape 6 / 8 - Traitement des NaN**

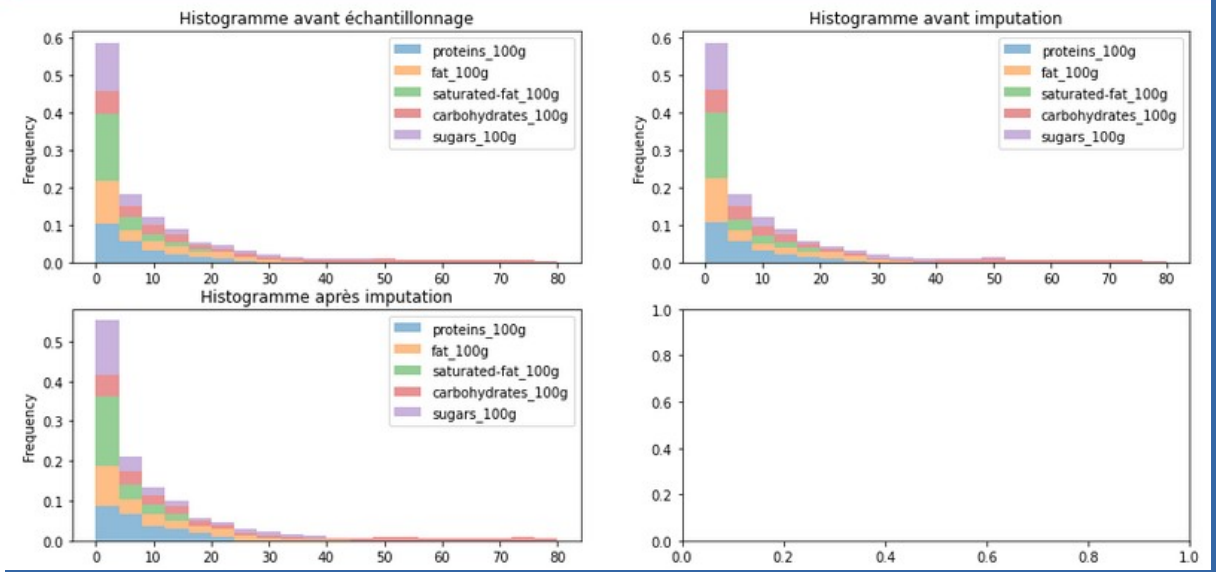
Dépendance des variables. Les données sont manquantes aléatoirement (MAR). Donc, les méthodes d'imputation OK

- IterativeImputer
- SimpleImputer
- Regression linéaire
- Fonction `nutrigrade_food`

- **Étape 7 / 8 - Vérification de la distribution avant et après imputation**

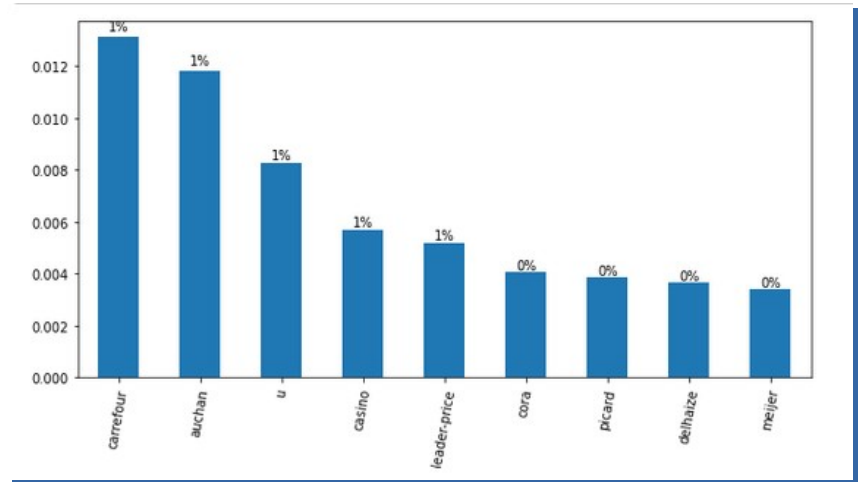
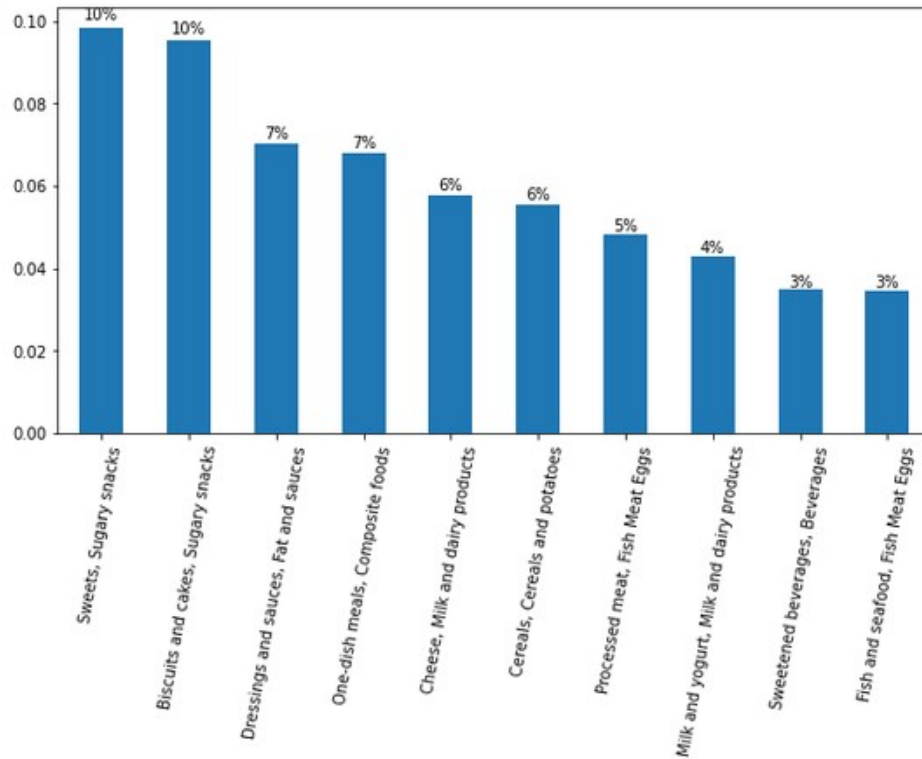
- **Étape 8 / 8 Automatisation**

`apply()`, compréhension de liste, fonction, boucle `for`



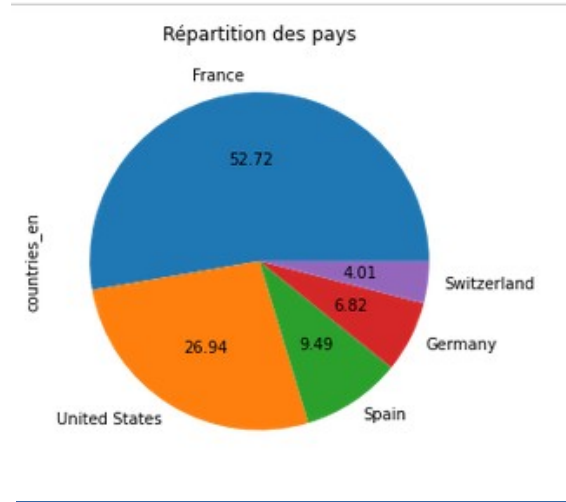
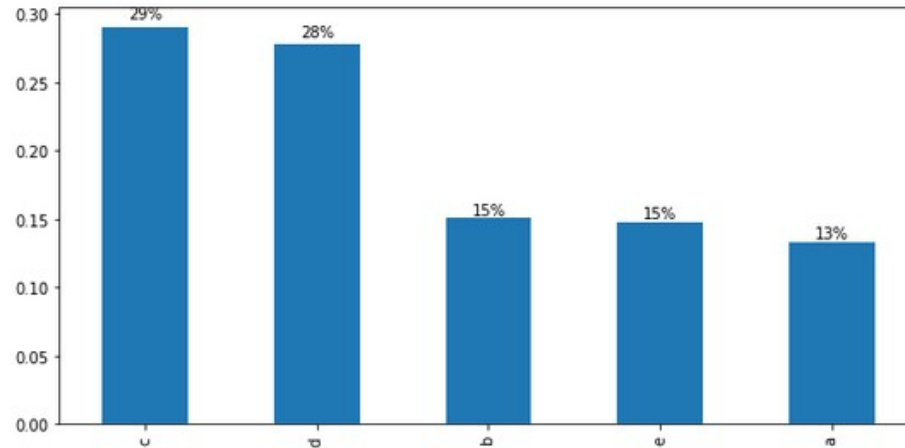
Outliers par la méthode inter quartile pour la visualisation des variables numériques.

Analyse uni-variée : variables qualitatives

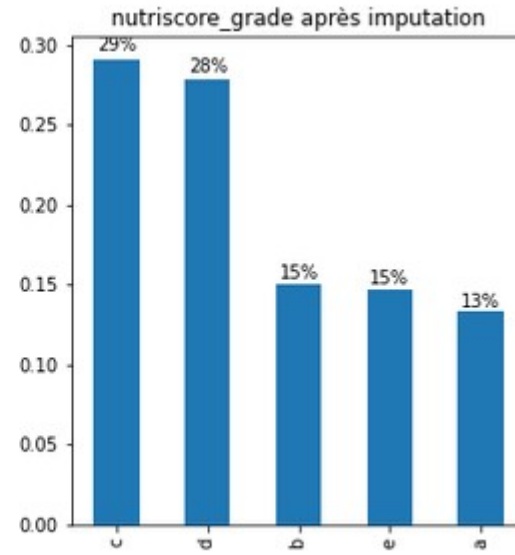
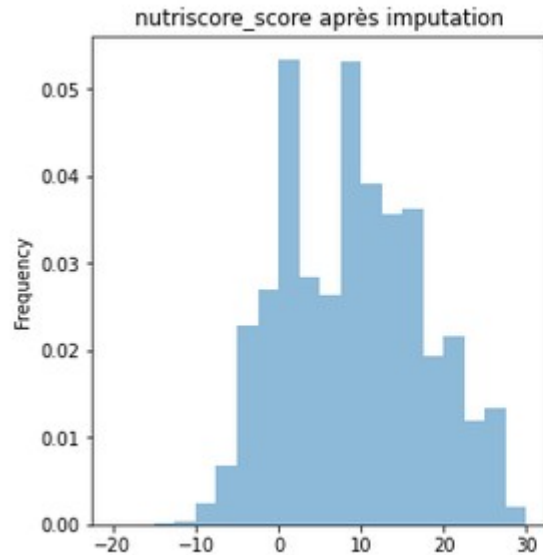


Produits surtout **industriels**, laitiers céréaliers, viande, poisson, et boisson

Analyse uni-variée : variables qualitatives

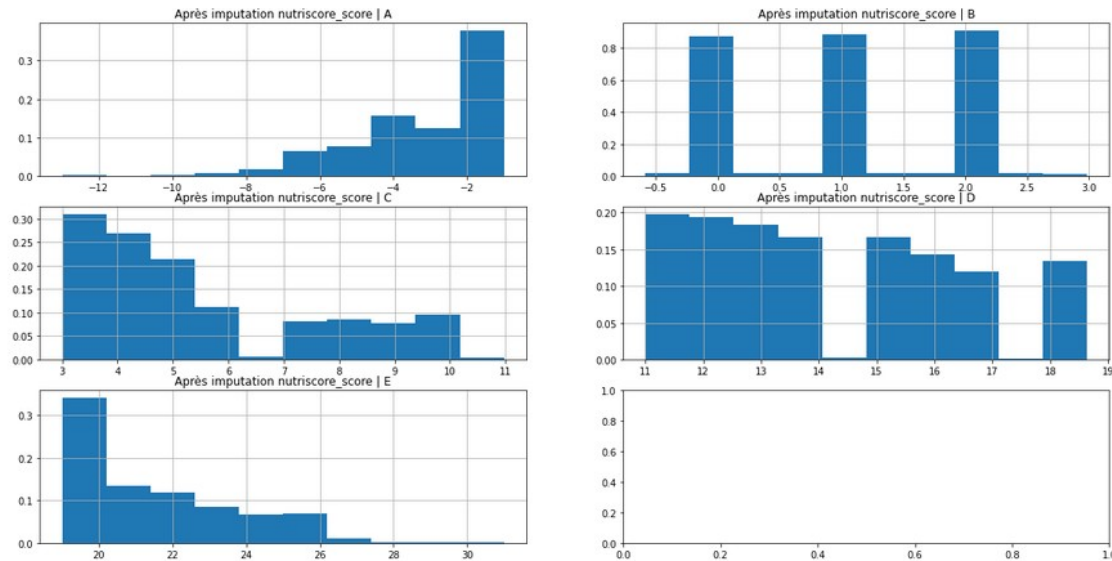







Analyse uni-variée : variables quantitatives



Points		Logo
Solid foods	Beverages	
Min to -1	Waters	
0 - 2	Min - 1	
3 - 10	2 - 5	
11 - 18	6 - 9	
19 - max	10 - max	

Analyse uni-variée : variables quantitatives



Points		Logo
Solid foods	Beverages	
Min to -1	Waters	
0 - 2	Min - 1	
3 - 10	2 - 5	
11 - 18	6 - 9	
19 - max	10 - max	

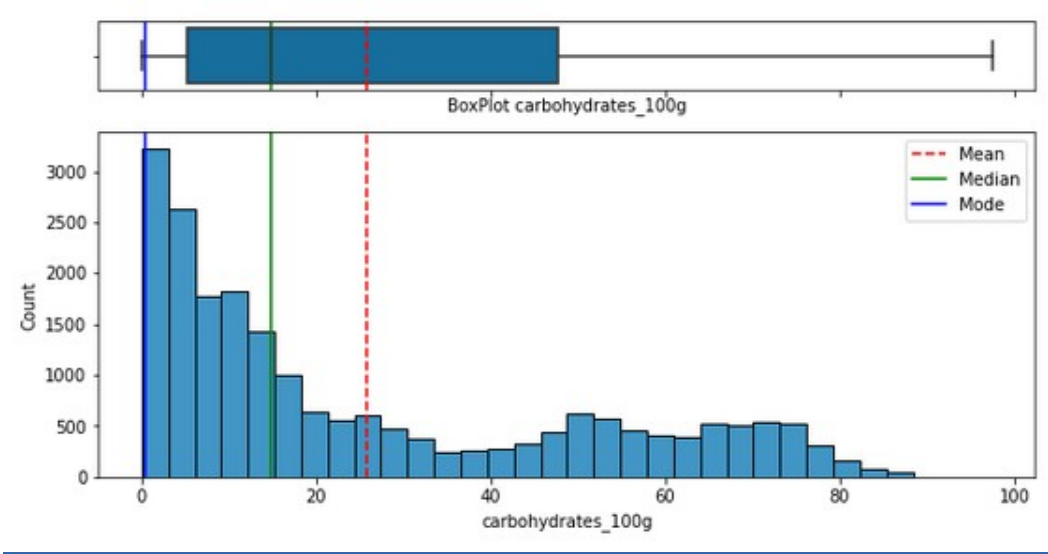
Analyse univariée : les mesures

Tendances centrales sur l'ensemble de dataset

Sélection des 2 variables à étudier selon le critère:

Ecart_max = mode – moyenne

- ['carbohydrates_100g']
- ['fat_100g']

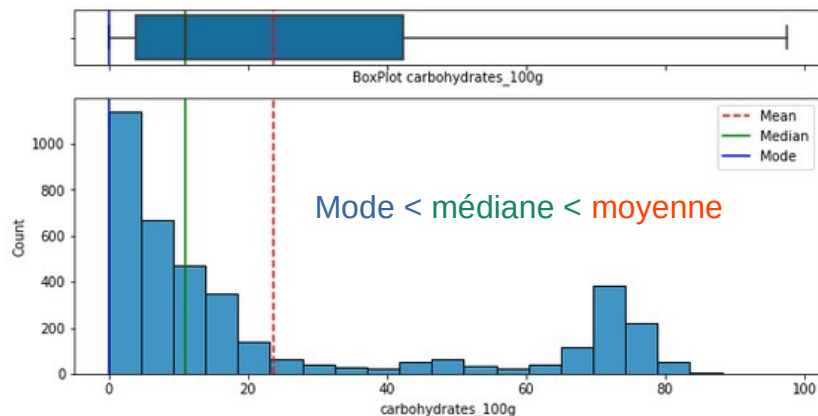


mode	0.00
median	13.89
mean	24.89

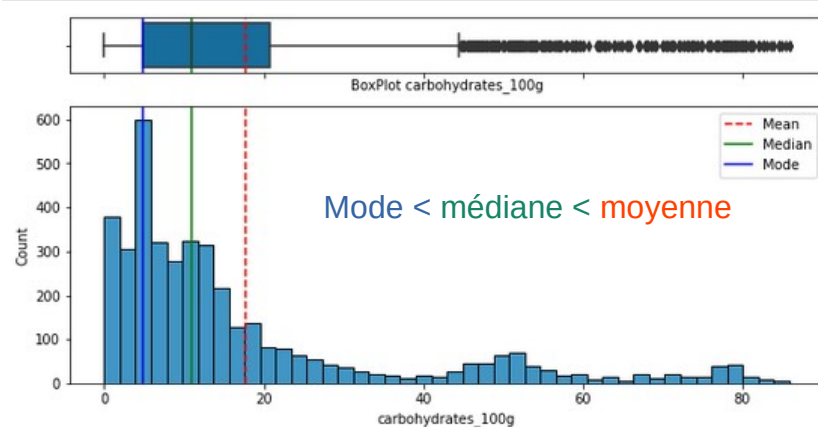
Mode < médiane < moyenne : distribution est *étalée* à droite

Analyse univariée : les mesures

Mesures de tendance centrales A



Mesures de tendance centrales B

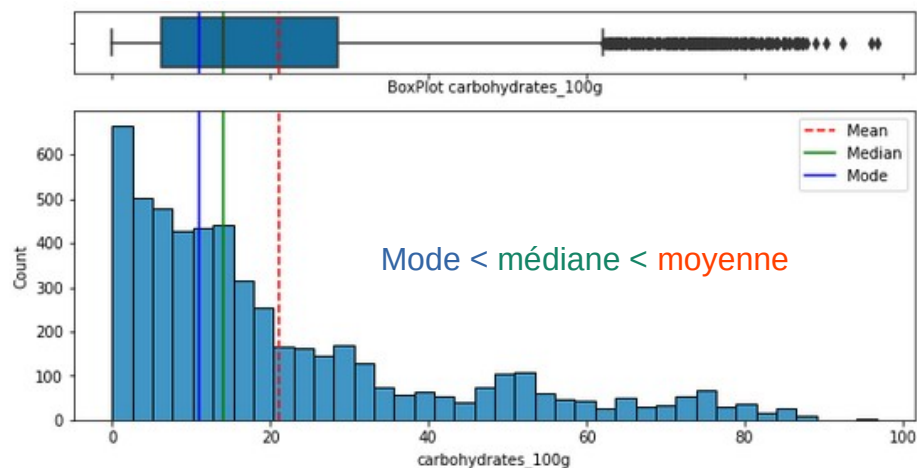


a : mode 0.00
 median 11.00
 mean 23.79

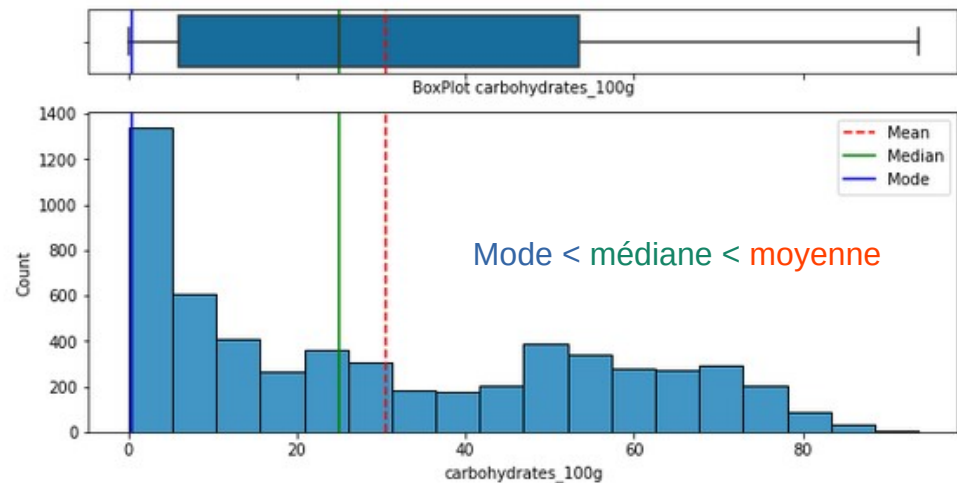
b : mode 0.00
 median 10.00
 mean 16.94

• Analyse univariée : les mesures

Mesures de tendance centrales C



Mesures de tendance centrales D

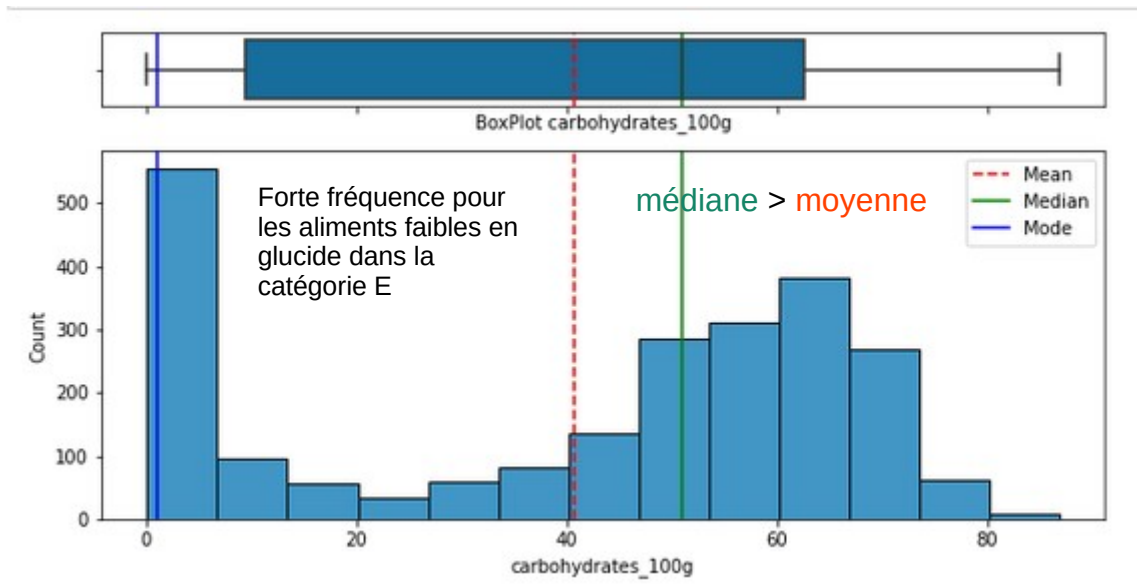


c : mode 0.00
 median 13.00
 mean 20.25

d : mode 0.00
 median 24.00
 mean 29.54

Analyse univariée : les mesures

Mesures de tendance centrales E



e : mode 1.00
 median 50.20
 mean 40.18

Analyse univariée : les mesures

Mesures de dispersion

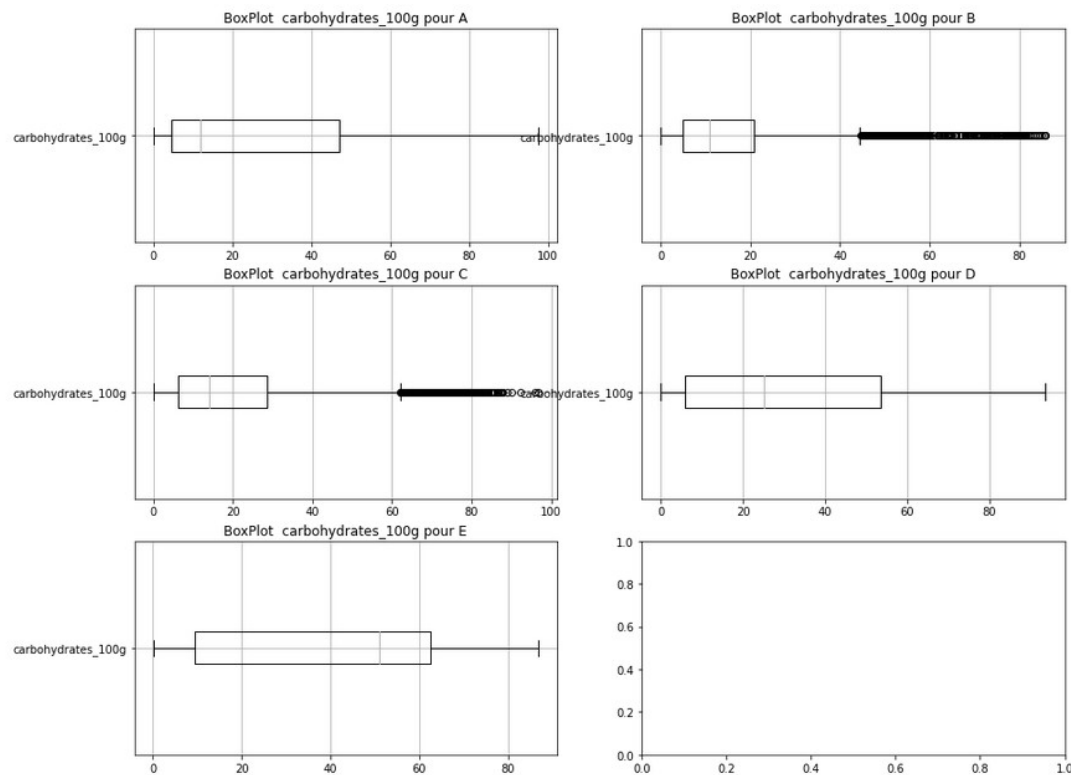
a : var **751.91**
std 27.42

b : var 375.73
std 19.38

c : var 422.10
std 20.55

d : var 662.74
std 25.74

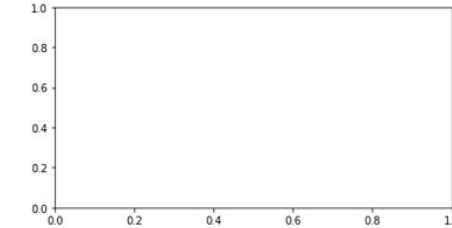
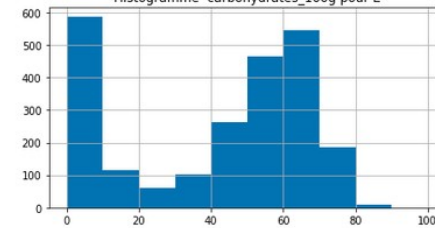
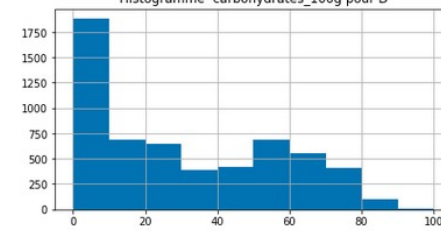
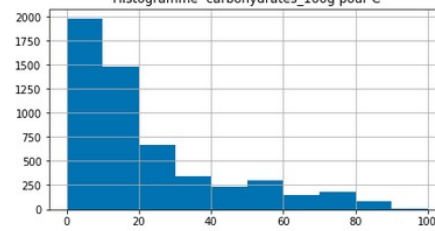
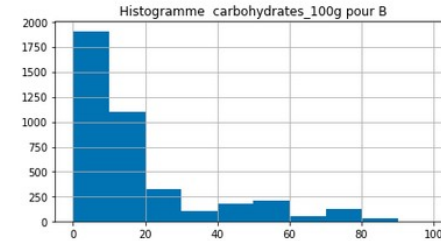
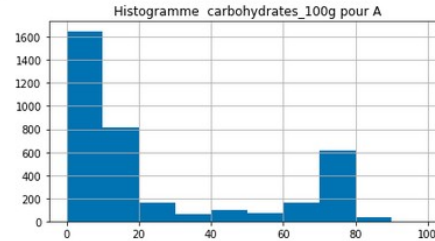
e : var 694.64
std 26.36



Analyse univariée : les mesures

Mesures de forme :

a : skew	1.05
Kurtosis	-0.57
b : skew	1.75
Kurtosis	2.30
c : skew	1.39
Kurtosis	1.16
d : skew	0.45
Kurtosis	-1.21
e : skew	-0.44
Kurtosis	-1.39

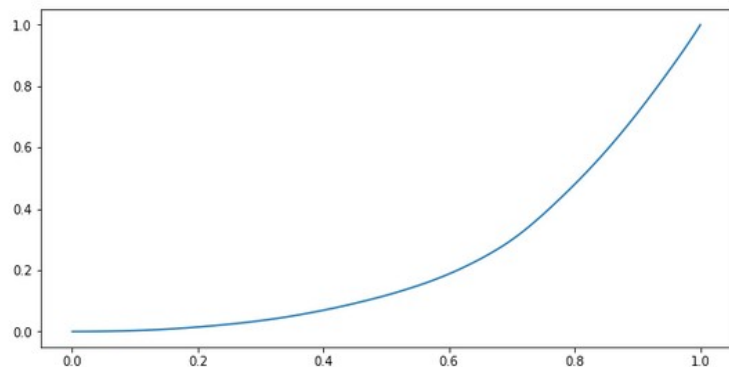


Si $y_1 < 0$ alors la distribution alors elle est étalée à gauche.

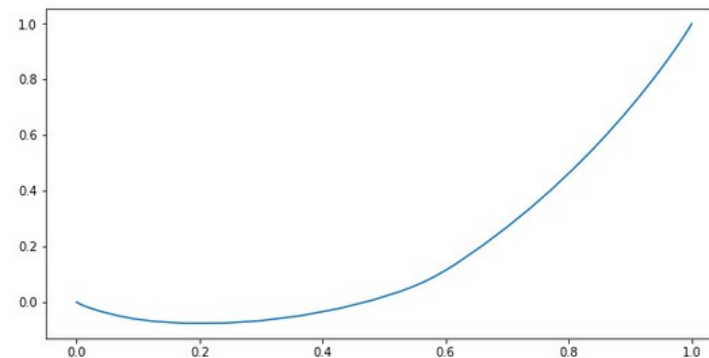
Si $y_2 < 0$, alors les observations sont moins concentrées : la distribution est plus aplatie.

Analyse univariée : les mesures

gini pour le carbohydrates_100g: 0.5207910422148633



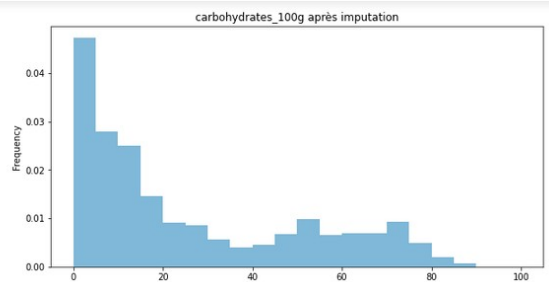
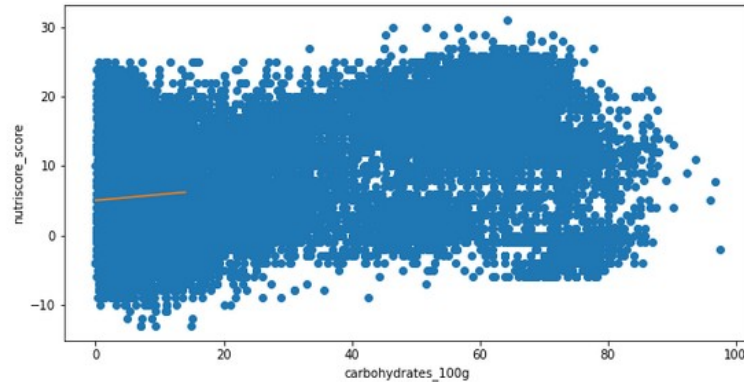
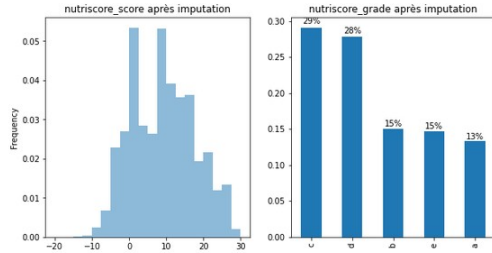
gini pour le nutriscore_score: 0.6422212496251185



Le nutriscore_score est plus inégalitaire entre les produits

Analyse multivariée - 2 variables quantitatives

```
carbohydrates_100g    0.08  
intercept              5.04  
dtype: float64
```



- **Étude de corrélation**

- cov = 51.525

- st.pearson r [-1,1] = 0.255

Coefficient proche de 0 donc
X et Y ne sont pas corrélés, mais
positif ...

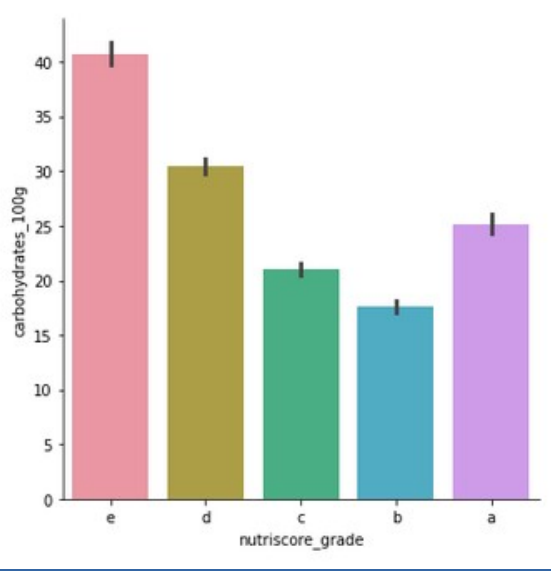
- **Évaluation**

ANOVA $R^2 [0,1] = 0.0653$

**Le modèle explique 6,53 % de la
variation entre les valeurs prédites
et les valeurs réelles**

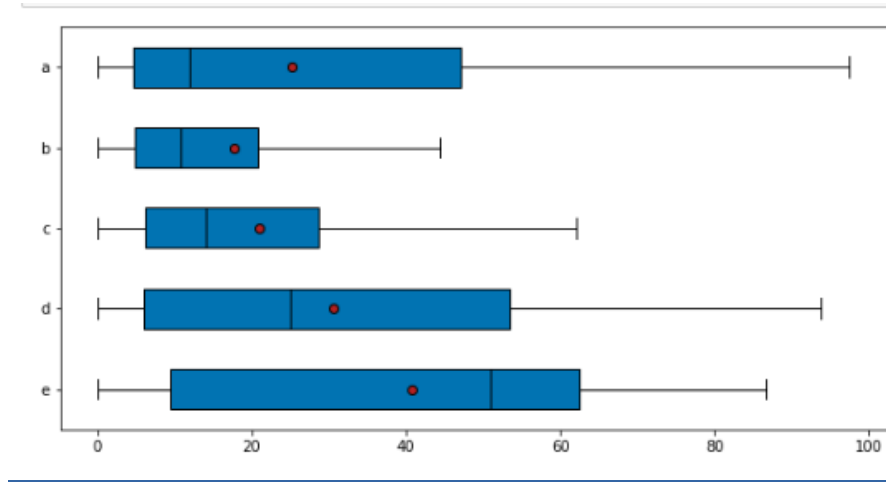
Analyse multivariée - 1 variable quantitative et 1 qualitative

Moyenne par classe



B et C mieux que A en terme de moyenne glucidique

Dispersion

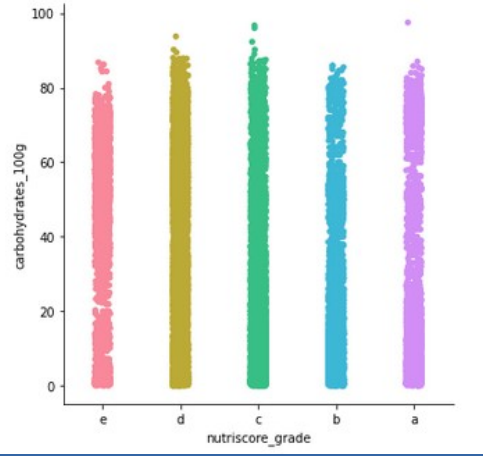


Corrélation :
 $\eta^2_{\text{squared}} = 0.0799$

Le modèle explique 7,99 % des variations. Les variables Y et X sont faiblement corrélées.

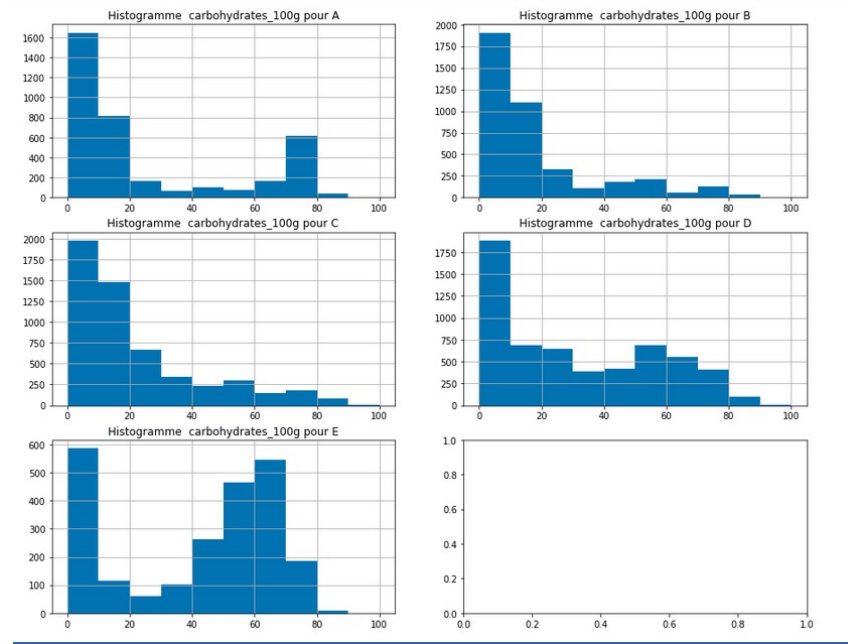
Analyse multivariée - 1 variable quantitative et 1 qualitative

Individus par classe



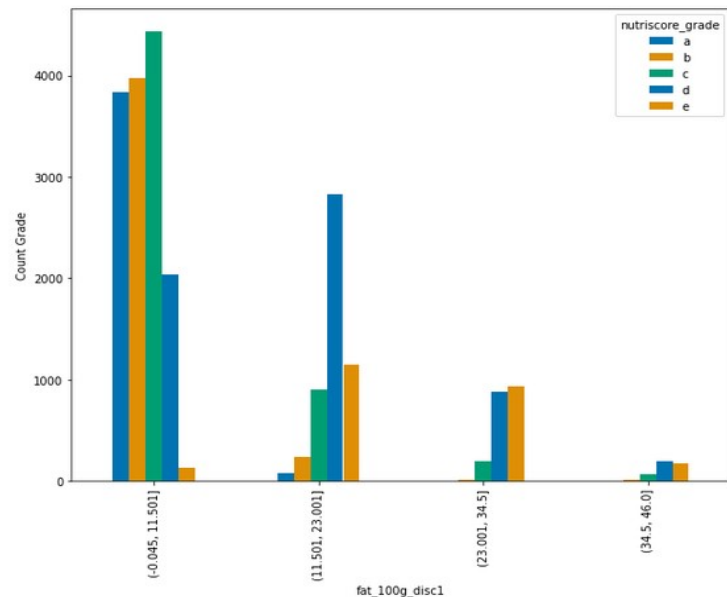
B, C, D, A présentent des individus au-delà de 80g de glucide au 100g.

E qui n'en présente que 5 environ



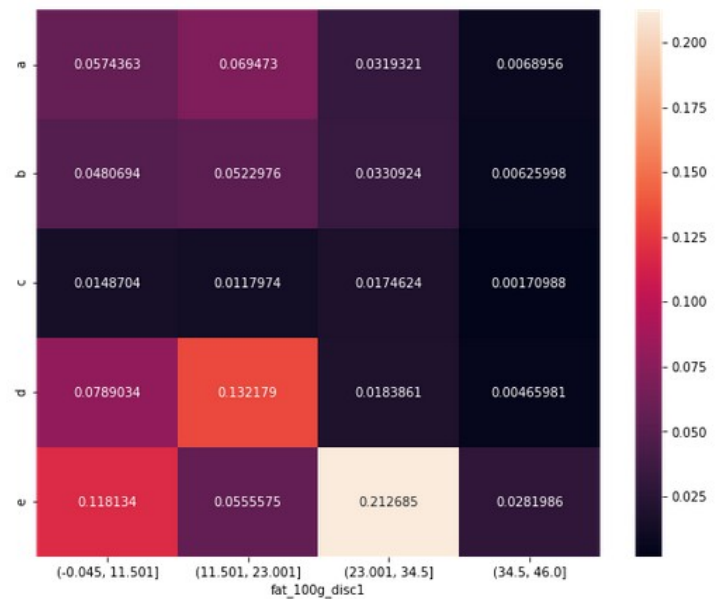
Analyse multivariée - 2 variables qualitatives

Tableau de contingence en graphique



A, B, C équivalents
C, D, E équivalents
=> confusion des notes

Contribution à la non dépendance



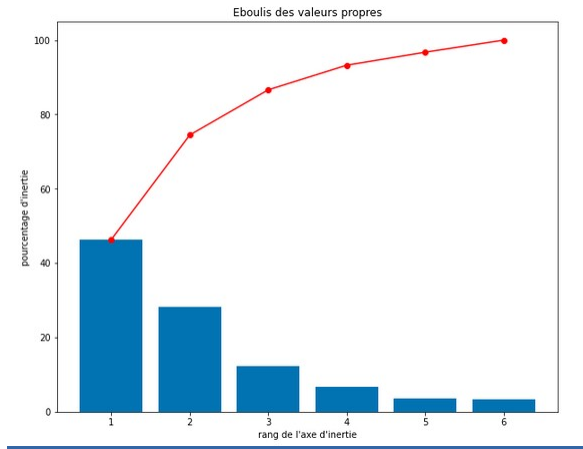
fat_100g est très dépendant à E

Corrélation
H0 = variables
indépendante

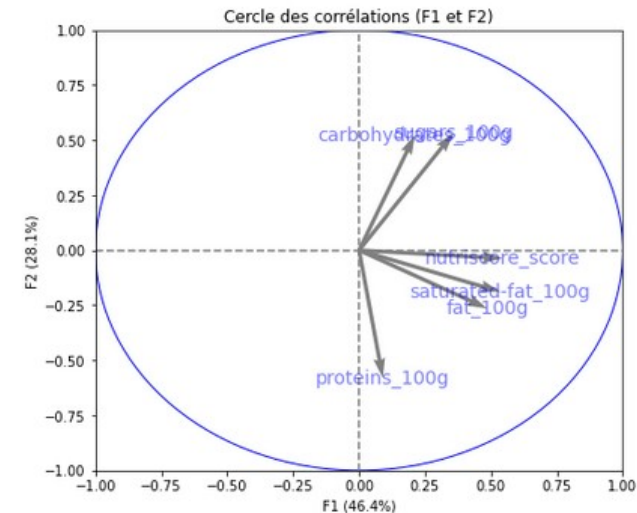
$p_value = 0 < 0,01$

Rejet de l'hypothèse
nulle H0

Analyse multivariée - PCA

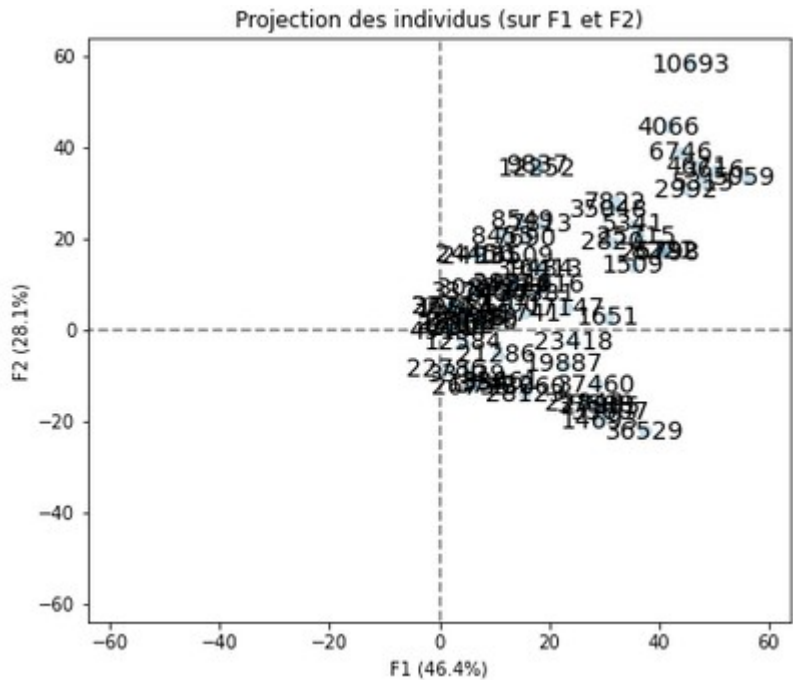


F1 et F2 expliquent 80 % de l'inertie totale



["nutriscore_score"] corrélé aux aliments gras

Analyse multivariée - PCA



Ce qui rapproche les individus :
le taux de glucide et le taux de lipide

```
: result = df_out.loc[10693]
result #exemple aliment gras et sucre
```

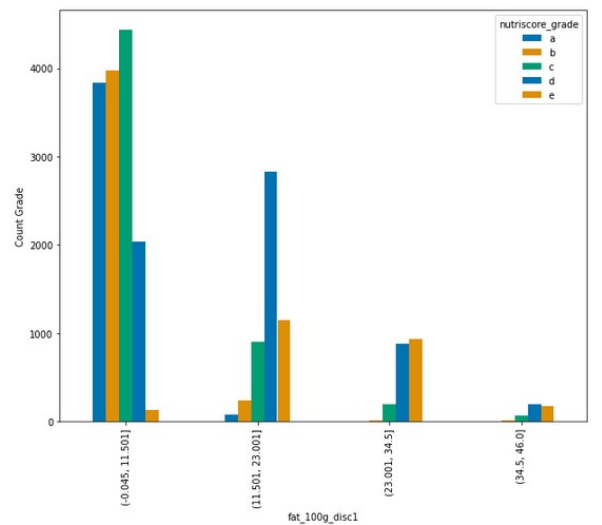
```

brands_tags      spartan
countries_en      United States
product_name      Spartan, coco krunch sweetened rice cereal mad...
energy_100g      1.7e+03
energy-kj_100g      1.7e+03
energy-kcal_100g      4.1e+02
proteins_100g      6.2
fat_100g      6.2
carbohydrates_100g      84
sugars_100g      38
salt_100g      1.6
sodium_100g      0.66
saturated-fat_100g      0
fiber_100g      0
nutriscore_score      20
nutriscore_grade      e
sugar_pct      0.44
satfat_pct      0
pnns_groups_21      Cereals, Cereals and potatoes
Name: 10693, dtype: object

```

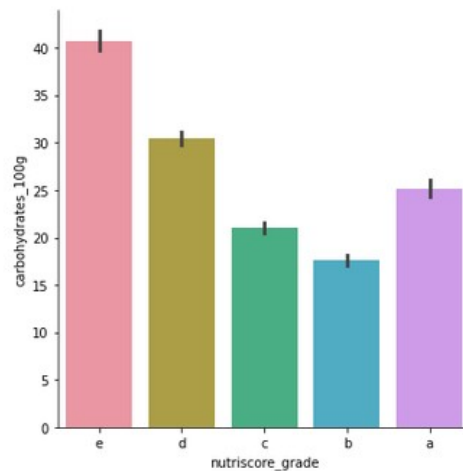
Présentation des faits pertinents pour l'application. 5 min

["fat_100g_disc1"]
count par classe



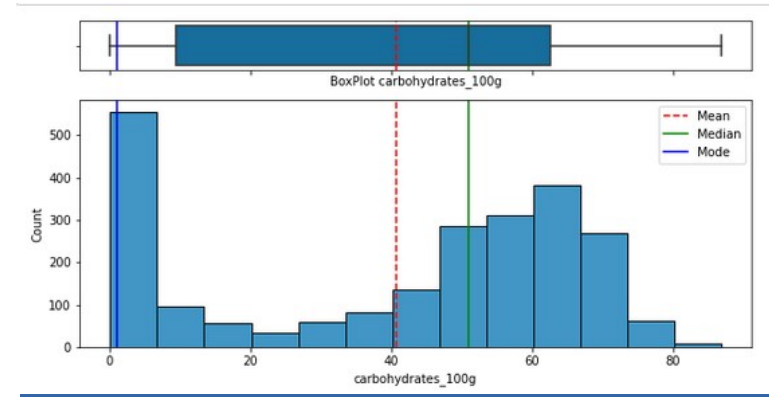
Confusion des notes A, B, C, D

["carbohydrates_100g"]
moyenne par classe



Les note B et C ont une moyenne inférieure à A.

["carbohydrates_100g"]
Distribution - dispersion



e :
mode 1.00
median 50.20
mean 40.18

50 % des individus ont des valeurs glucidiques inférieures à 50g/100g

Synthèse des différentes conclusions sur la faisabilité de votre projet.

Confusion entre les notes en comparaison avec ["fat_100g"]

Ordonnancement des notes non cohérent en comparaison avec ["carbohydrates_100g"]

Présence d'aliments bons pour la santé dans la catégorie E

	countries_en	product_name	nutriscore_grade	energy-kcal_100g	nutriscore_score
3372	France	Petit Broyé	e	0.0	19.0
3989	United States	Mini Madeleine Chocolat	e	18.0	21.0
30721	Belgium	Boudins blancs	e	24.1	19.0
45956	Switzerland	Detox Smoothie	e	59.0	19.0
40169	Germany	Mangue	e	61.0	19.0
39481	Spain	Ginger beer	e	61.0	19.0
24307	France	Nectar fraise	e	62.0	19.0
2361	France	Coconut-Pineapple Nectar From Concentrate	e	63.0	19.0
40451	Spain	Cosmo	e	64.0	19.0
46518	France	100% juice from spanish grapes	e	65.0	20.0

CONCLUSION

Privilégier aussi les autres mesures dans l'application