PROJET N°2: CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

BOURBON VICENTE

SOMMAIRE

- Idée d'application
- Opérations de nettoyage
- Analyses univariées
- Analyses multivariées
- Faisabilité de l'application
- Conclusion

IDÉE D'APPLICATION

- Nutriscore
- Régime alimentaire concerné: sans régime, végétarien, végan
- Origine
- Impact environnemental

NETTOYAGE DES DONNÉES

I. VALEURS MANQUANTES

- Suppression de variables
- Affectation arbitraire
- Valeur médiane pour les variables quantitatives
- Formule de calcul du nutriscore
- Valeur la plus présente en fonction du nutriscore pour les variables qualitatives

	Nombre de valeurs manquantes	% de valeurs manquantes
water-hardness_100g	320772	100.000000
caproic-acid_100g	320772	100.000000
elaidic-acid_100g	320772	100.000000
nucleotides_100g	320763	100.000000
ingredients_that_may_be_from_palm_oil	320772	100.000000
nutrition_grade_uk	320772	100.000000
serum-proteins_100g	320756	100.000000
maltodextrins_100g	320761	100.000000
maltose_100g	320768	100.000000
nervonic-acid_100g	320772	100.000000
erucic-acid_100g	320772	100.000000

NETTOYAGE DES DONNÉES

II. DOUBLONS

- Lignes en doublons
- Codes en doublons
  - Produits différents
  - Produits identiques

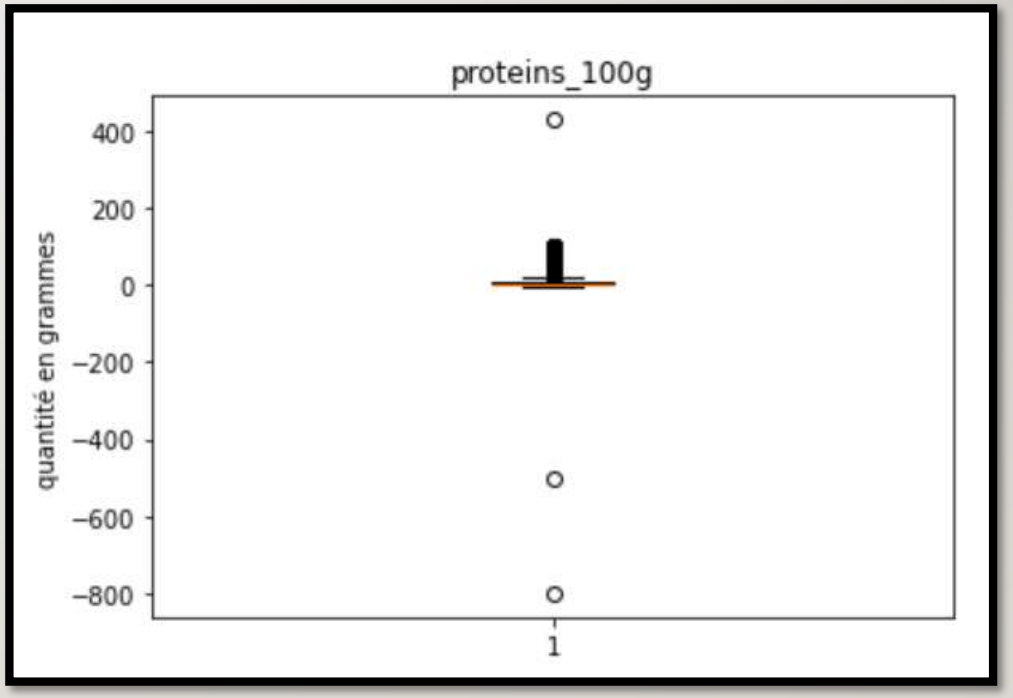
code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	gene
16117.0	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489055730	2017-03-09T10:35:30Z	1489055730	2017-03-09T10:35:30Z	Organic Long Grain White Rice	
16117.0	http://world-fr.openfoodfacts.org/produit/0001...	usda-ndb-import	1489065258	2017-03-09T13:14:18Z	1489065258	2017-03-09T13:14:18Z	Colossal Olives With Jalapeno Peppers	

code	url	creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	ge
9.800800e+09	http://world-fr.openfoodfacts.org/produit/0000...	usda-ndb-import	1489061721	2017-03-09T12:15:21Z	1489061721	2017-03-09T12:15:21Z	Hazelnut Spread + Breadsticks	
9.800800e+09	http://world-fr.openfoodfacts.org/produit/0009...	openfoodfacts-contributors	1457659842	2016-03-11T01:30:42Z	1489068296	2017-03-09T14:04:56Z	Hazelnut Spread + Breadsticks	

NETTOYAGE DES DONNÉES

III. VALEURS ABERRANTES

- Valeurs nutritionnelles comprises entre 0 et 100
- Remplacées par la valeur moyenne



NETTOYAGE DES DONNÉES

IV. CRÉATION D'UNE NOUVELLE VARIABLE

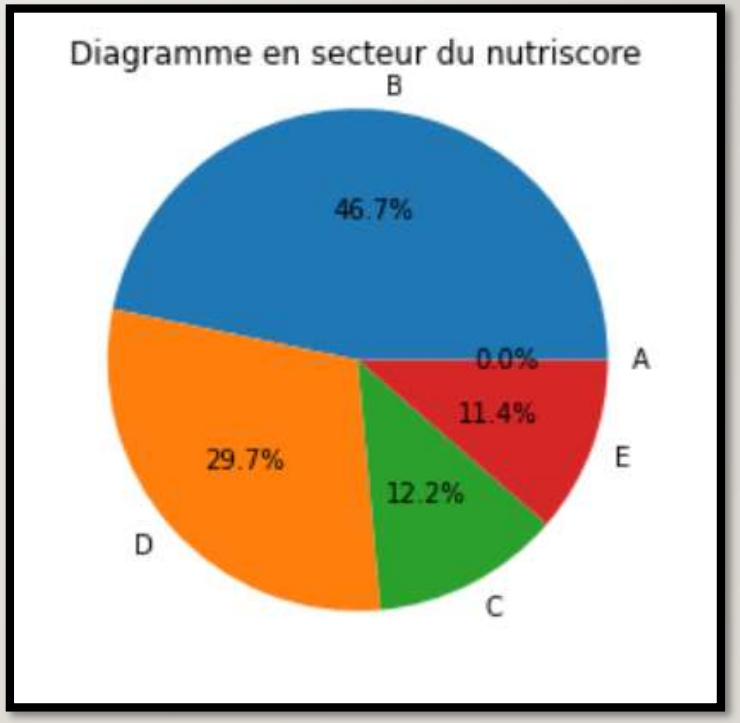
- Variable pour la liste des régimes alimentaires

ANALYSES UNIVARIÉES

I. VARIABLE QUALITATIVE ORDINALE

- Exploration du nutriscore

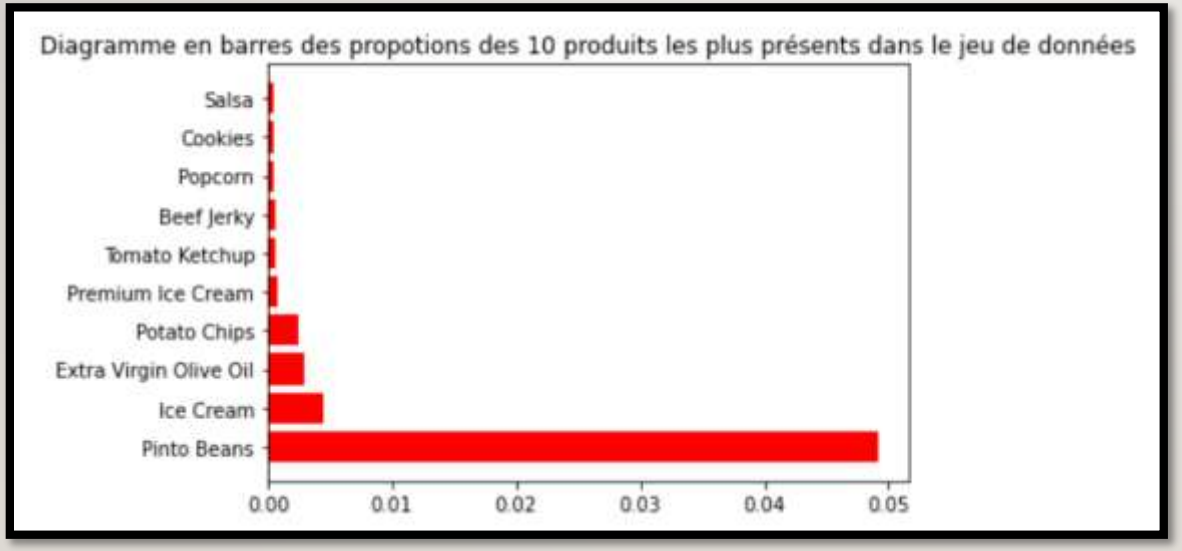
	nutriscore	n	f
0	b	149930	0.467475
1	d	95306	0.297160
2	c	39072	0.121825
3	e	36412	0.113531
4	a	3	0.000009



ANALYSES UNIVARIÉES

II. VARIABLE QUALITATIVE NOMINALE

- Exploration des noms des produits

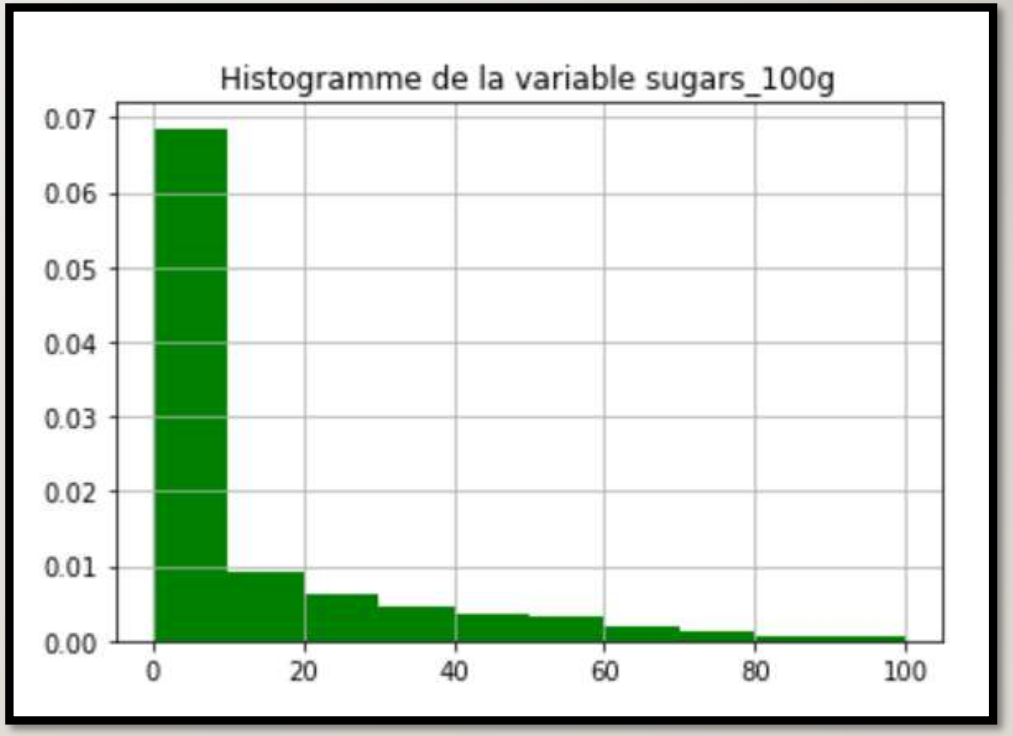


	Produit	n	f
0	Pinto Beans	15773	0.049180
1	Ice Cream	1416	0.004415
2	Extra Virgin Olive Oil	954	0.002975
3	Potato Chips	775	0.002416
4	Premium Ice Cream	226	0.000705
5	Tomato Ketchup	182	0.000567
6	Beef Jerky	167	0.000521
7	Popcorn	157	0.000490
8	Cookies	155	0.000483
9	Salsa	150	0.000468

ANALYSES UNIVARIÉES

III. VARIABLES QUANTITATIVES CONTINUES

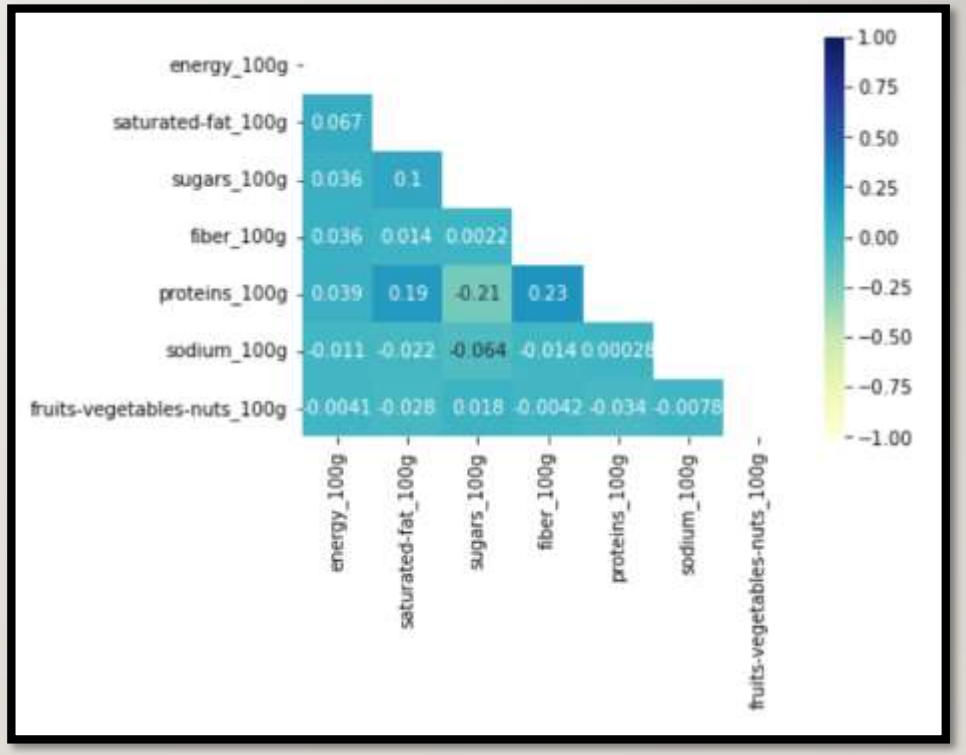
- Exploration des teneurs en nutriments
- Forme de l'histogramme similaire pour chaque variable

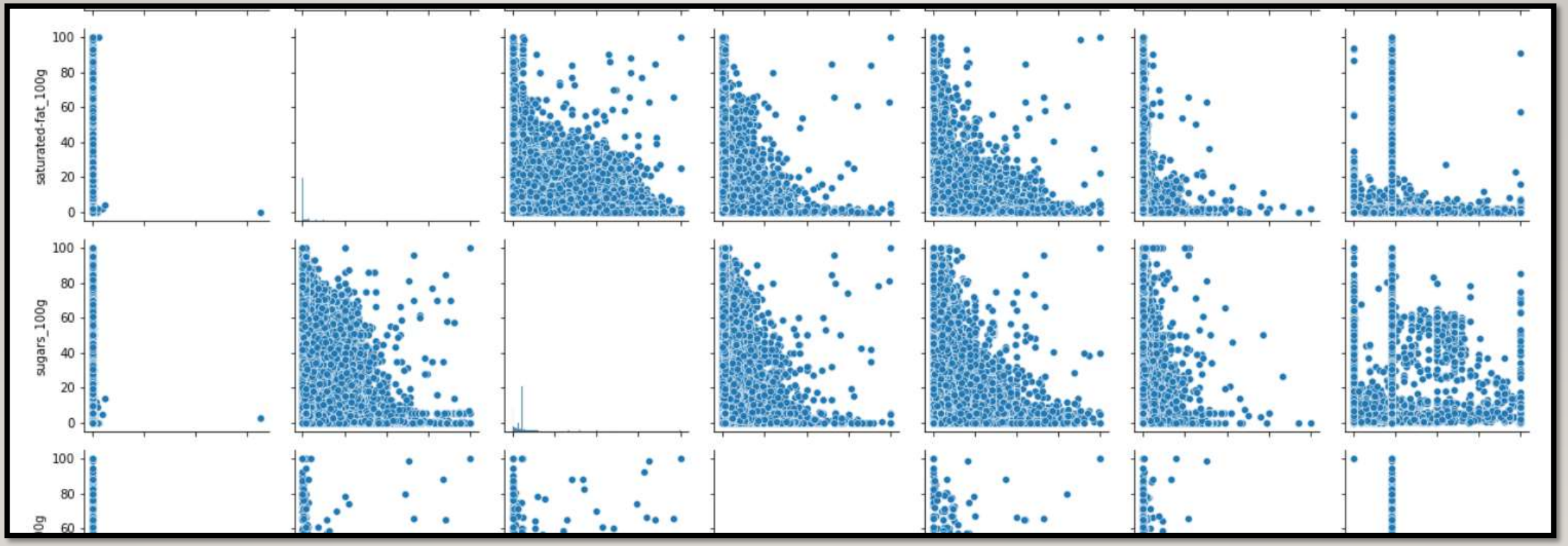


ANALYSES MULTIVARIÉES

I. ANALYSE BIVARIÉE

- Matrice des corrélations
- Nuages de points
- Pas de corrélation linéaire pour les variables considérées

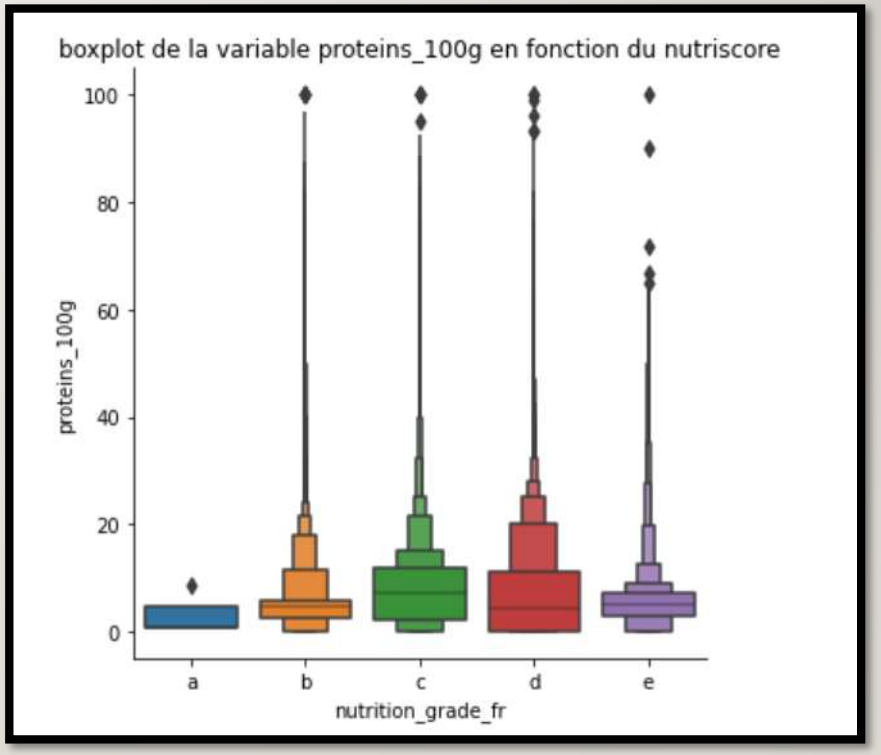


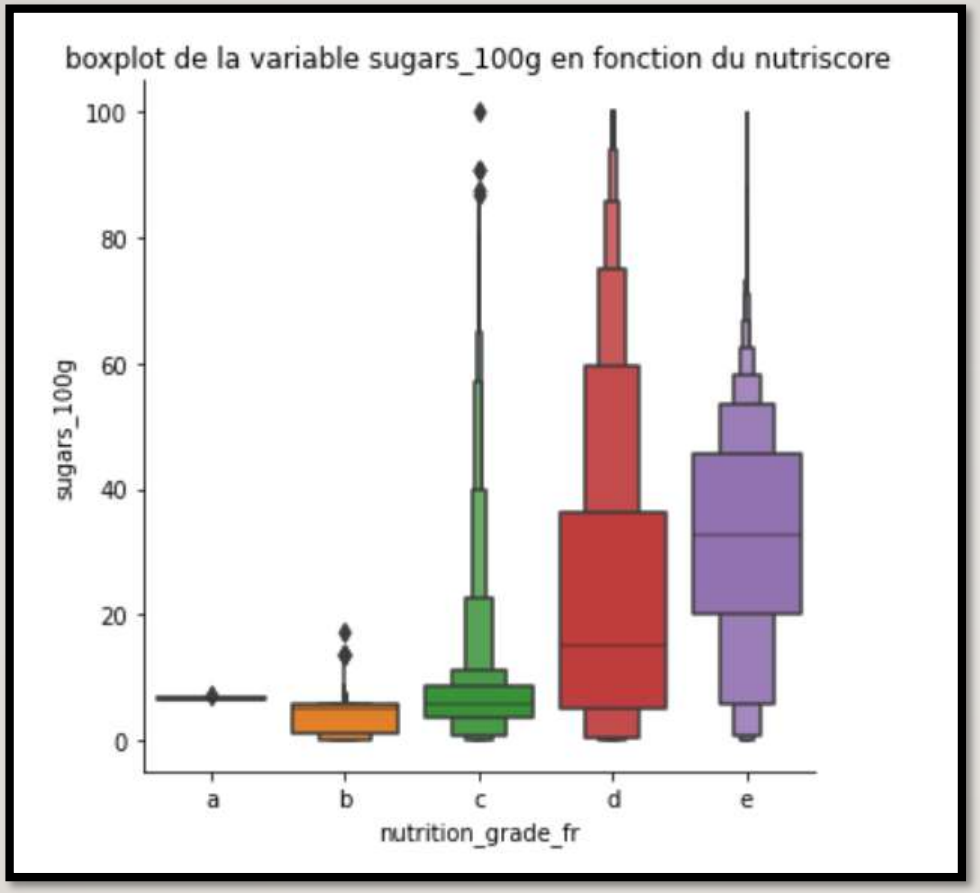

saturated-fat_100g
sugars_100g
0
20
40
60
80
100

ANALYSES MULTIVARIÉES

II. ANOVA

- Corrélation entre le nutriscore et la teneur en nutriments
- Corrélation pour la quantité de sucre et de graisses saturées





	sum_sq	df	mean_sq	F	PR(>F)	eta_sq	omega_sq
nutrition_grade_fr	3.926798e+07	4.0	9.816995e+06	41121.315646	0.0	0.339003	0.338994
Residual	7.656581e+07	320718.0	2.387325e+02	NaN	NaN	NaN	NaN

ANALYSES MULTIVARIÉES

III. TEST DU CHI DEUX

- Corrélation entre le nutriscore et la catégorie de produit

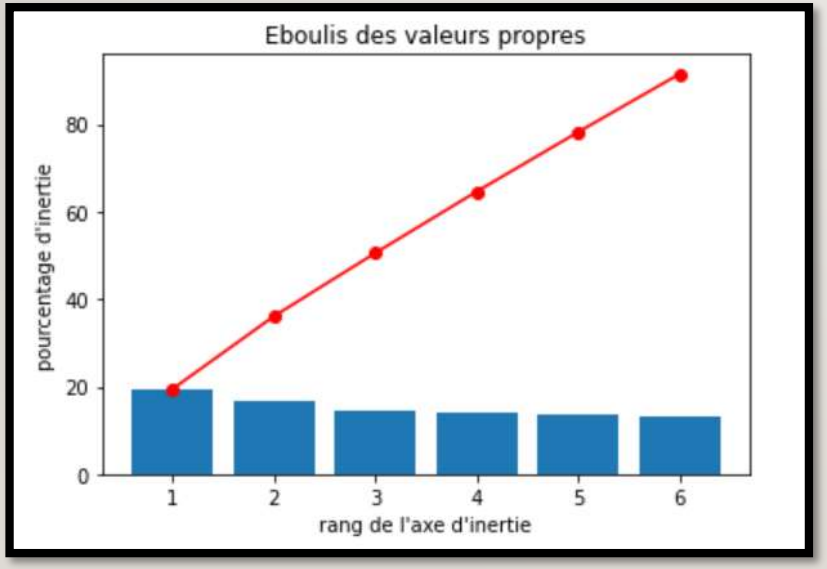
```
: stats.chi2_contingency(crosstab)

: (902121.910486136,
 0.0,
 84572,
```

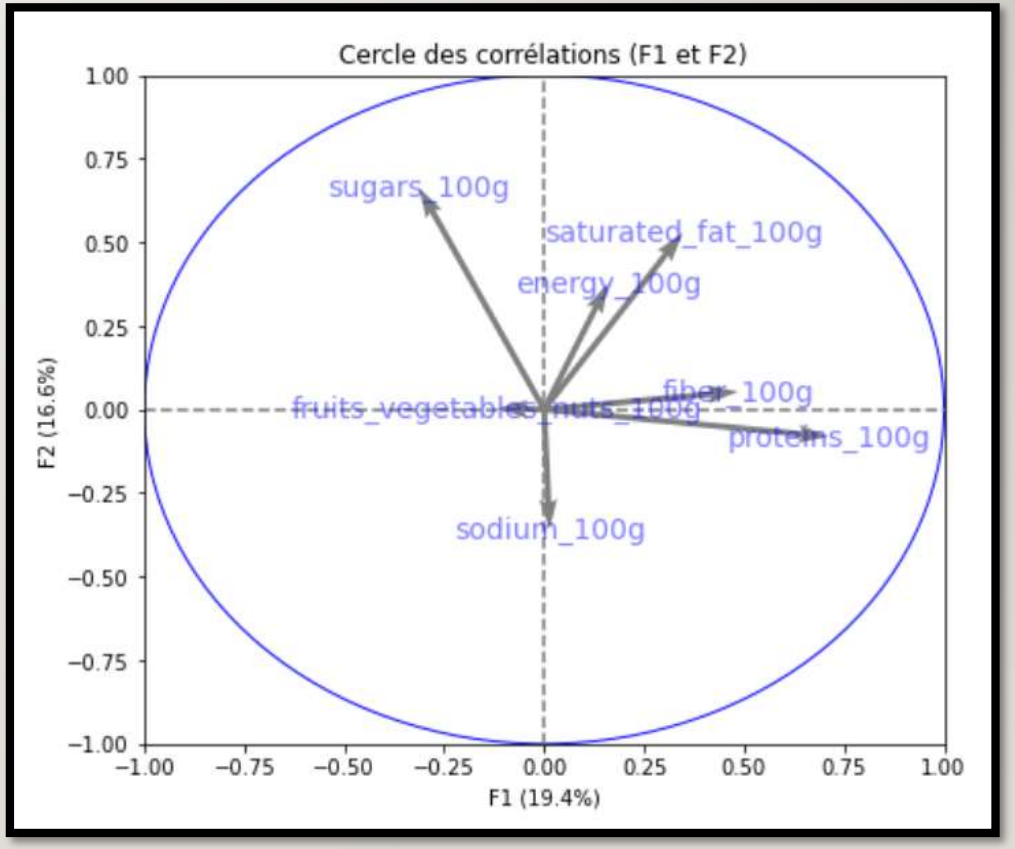
nutrition_grade_fr	a	b	c	d	e
categories_fr					
4	0	1	0	0	0
6	0	3	0	0	0
A-code-1	0	1	0	0	0
Ab	0	1	0	0	0
Abats-surgeles	0	1	0	0	0
Abdijbier,Alcoholische-dranken,Bieren,Bruine-bieren,Dranken,Trappistenbier	0	1	0	0	0
Accras-de-morue	0	3	0	0	0
Aceitunas,Aceitunas-deshuesadas,Aceitunas-verdes,Aceitunas-verdes-deshuesadas,Encurtidos,Hortalizas,Vegetales-encurtidos	0	1	0	0	0
Acqua,Acqua-minerale	0	1	0	0	0
Acqua-minerale	0	2	0	0	0
Acras-de-morue	0	4	0	0	0
Additifs	0	1	0	0	0

ANALYSES MULTIVARIÉES

IV. ACP

- Réduction dimensionnelle sur les teneurs en nutriments
- 7 variables, 6 dimensions





FAISABILITÉ DE L'APPLICATION

- Pas de données pour l'impact environnemental
- Peu de donnée pour l'origine des produits
- Valeurs manquantes impossible à évaluer

FAISABILITÉ DE L'APPLICATION

- Catégories mal renseignées
- Pas facilement exploitables

Abdijbier,Alcoholische-dranken,Bieren,Bruine-bieren,Dranken,Trappistenbier
Accras-de-morue
Aceitunas,Aceitunas-deshuesadas,Aceitunas-verdes,Aceitunas-verdes-deshuesadas,Encurtidos,Hortalizas,Vegetales-encurtidos
Acqua,Acqua-minerale
Acqua-minerale
Acras-de-morue

FAISABILITÉ DE L'APPLICATION

- Peu de données pour la teneur en fruits et légumes
- Biais dans le calcul des valeurs manquantes du nutriscore

CONCLUSION

- Jeu de données plutôt vide
- Variables mal renseignées et difficilement exploitables
- Application en partie faisable mais non pertinente