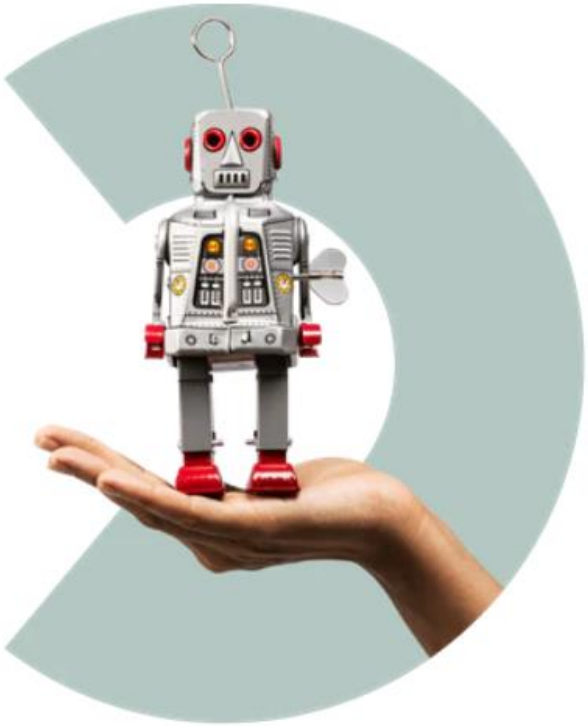
PROJET 3
SOUTENANCE

PREPAREZ DES DONNEES POUR UN ORGANISME DE SANTE PUBLIQUE:

EXPLORATION ET VISUALISATION DE DONNEES

#CLEANING
#ANALYSE DESCRIPTIVE UNIVARIEE, BIVARIEE, MULTIVARIEE
#ANOVA #ACP
#VOILA (dashboarding interactif)
#ANALYSE EXPLICATIVE MULTIVARIEE
#REGRESSION LINEAIRE
#PANDAS #MATPLOTLIB
#STATISTICS #SCIPY.STATS #STATSMODELS #SKLEARN



Ingénieur IA

Développez et intégrez des algorithmes de Deep Learning au sein d'un produit IA

OPENCLASSROOMS

OUDDANE NABIL

SOMMAIRE

Projet 3

Préparez des données pour un organisme de santé publique

A INTRODUCTION

1. Contexte

- ENJEU global: rendre les données de santé publique plus accessibles, pour qu'elles soient utilisables par les agents de santé publique France
- ENJEU DU P3: **exploration et visualisation des données**
- DONNEES SOURCES
  - **le jeu de données : openfood**
    - https://world.openfoodfacts.org/
    - https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-scientist/P2/fr.openfoodfacts.org.products.csv.zip
    - Descriptif: https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/AI+Engineer/Project+3+-+Pr%C3%A9parez+des+donn%C3%A9es+pour+un+organisme+de+sant%C3%A9+publique/Open_Food_Facts_data-fields.txt

2. Objectifs

- SCRIPT:
  - **repérer** des variables pertinentes
  - **Automatiser** les traitements en les rendant flexibles
  - produire des **visualisations** et des **analyses univariées** pour chaque variable jugée pertinente
  - Confirmer ou infirmer des hypothèses à l'aide **d'une analyse multivariée descriptive et explicative**
    - avec tests statistiques appropriés

B PREREQUIS AU PROJET

1. Voila

- INSTALLATION du package VOILA
  - Jupyter est un outils exceptionnel pour fluidifier les workflows allant de l'analyse exploratoire à la communication des résultats
  - **Il n'est cependant pas adapté à toutes les audiences et en particulier au personnes non techniques**
  - **Voila** vient pallier ce problème en rendant les notebooks interactifs dans une application web sécurisée interactive dont le code n'est pas accessible

Dashboarding et applications web:
2 mondes

- Dev sur mesure en javascript
- **Beau mais couteux car nécessite un développeur web**

- **Outils automatique avec peu de développement web**

2. Données: openfood

- Descriptif pas complètement à jour: https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/AI+Engineer/Project+3+-+Pr%C3%A9parez+des+donn%C3%A9es+pour+un+organisme+de+sant%C3%A9+publique/Open_Food_Facts_data-fields.txt
- 162 champs: 4 types de champs:
  - Les informations générales sur la fiche du produit : nom, date de modification, etc.
    - 10 champs
  - Un ensemble de tags : catégorie du produit, localisation, origine, etc.
    - 24 champs
  - Les ingrédients composant les produits et leurs additifs éventuels.
    - 29 champs
  - Des informations nutritionnelles : quantité en grammes d'un nutriment pour 100 grammes du produit.
    - 99 champs
- Generalités sur les champs

suffixe du champs	
_t	date unix depuis 01/01/1970
_datetime	date format yyyy-mm-ddThh:mn:ssZ
_tags	liste de tags separés par une virgule
_(2letter language code)	liste de tags dans le language
_100g	montant de nutriment en g ou kj pour 100g ou 100ml
_serving	montant de nutriment en g ou kj pour 1 dose

2. Données: openfood
320772 lignes – Presque 1Go de données

Information générales
10 champs

champs	type	observations
code	object	320749
url	object	320749
creator	object	320770
created_t	object	320769
created_datetime	object	320763
last_modified_t	object	320772
last_modified_datetime	object	320772
product_name	object	303010
generic_name	object	52795
quantity	object	104819

Tags
24 champs

packaging	object	78960
packaging_tags	object	78961
brands	object	292360
brands_tags	object	292352
categories	object	84410
categories_tags	object	84389
categories_fr	object	84411
origins	object	22190
origins_tags	object	22153
manufacturing_places	object	36501
manufacturing_places_tags	object	36495
labels	object	46559
labels_tags	object	46644
labels_fr	object	46666
emb_codes	object	29306
emb_codes_tags	object	29303
first_packaging_code_geo	object	18803
cities	object	23
cities_tags	object	20320
purchase_places	object	58193
stores	object	51722
countries	object	320492
countries_tags	object	320492
countries_fr	object	320492

Ingrédients
29 champs

ingredients_text	object	248962
allergens	object	28344
allergens_fr	object	19
traces	object	24353
traces_tags	object	24329
traces_fr	object	24352
serving_size	object	211331
no_nutriments	float64	0
additives_n	float64	248939
additives	object	248905
additives_tags	object	154680
additives_fr	object	154680
ingredients_from_palm_oil_n	float64	248939
ingredients_from_palm_oil	float64	0
ingredients_from_palm_oil_tags	object	4835
ingredients_that_may_be_from_palm_oil_n	float64	248939
ingredients_that_may_be_from_palm_oil	float64	0
ingredients_that_may_be_from_palm_oil_tags	object	11696
nutrition_grade_uk	float64	0
nutrition_grade_fr	object	221210
pnns_groups_1	object	91513
pnns_groups_2	object	94491
states	object	320726
states_tags	object	320726
states_fr	object	320726
main_category	object	84366
main_category_fr	object	84366
image_url	object	75836
image_small_url	object	75836

2. Données: openfood
320772 lignes – Presque 1Go de données

Information nutritionnelles- 99 champs

energy_100g	float64	261113
energy-from-fat_100g	float64	857
fat_100g	float64	243891
saturated-fat_100g	float64	229554
butyric-acid_100g	float64	0
caproic-acid_100g	float64	0
caprylic-acid_100g	float64	1
capric-acid_100g	float64	2
lauric-acid_100g	float64	4
myristic-acid_100g	float64	1
palmitic-acid_100g	float64	1
stearic-acid_100g	float64	1
arachidic-acid_100g	float64	24
behenic-acid_100g	float64	23
lignoceric-acid_100g	float64	0
cerotic-acid_100g	float64	0
montanic-acid_100g	float64	1
melissic-acid_100g	float64	0
monounsaturated-fat_100g	float64	22823
polyunsaturated-fat_100g	float64	22859
omega-3-fat_100g	float64	841
alpha-linolenic-acid_100g	float64	186
eicosapentaenoic-acid_100g	float64	38
docosahexaenoic-acid_100g	float64	78
omega-6-fat_100g	float64	188
linoleic-acid_100g	float64	149
arachidonic-acid_100g	float64	8
gamma-linolenic-acid_100g	float64	24
dihomo-gamma-linolenic-acid_100g	float64	23
omega-9-fat_100g	float64	21
oleic-acid_100g	float64	13
elaidic-acid_100g	float64	0
gondoic-acid_100g	float64	14
mead-acid_100g	float64	0
erucic-acid_100g	float64	0
nervonic-acid_100g	float64	0
trans-fat_100g	float64	143298
cholesterol_100g	float64	144090
carbohydrates_100g	float64	243588
sugars_100g	float64	244971
sucrose_100g	float64	72
glucose_100g	float64	26
fructose_100g	float64	38
lactose_100g	float64	262
maltose_100g	float64	4
maltodextrins_100g	float64	11
starch_100g	float64	266
polyols_100g	float64	414
fiber_100g	float64	200886
proteins_100g	float64	259922
casein_100g	float64	27
serum-proteins_100g	float64	16
nucleotides_100g	float64	9
salt_100g	float64	255510
sodium_100g	float64	255463
alcohol_100g	float64	4133
vitamin-a_100g	float64	137554
beta-carotene_100g	float64	34
vitamin-d_100g	float64	7057
vitamin-e_100g	float64	1340
vitamin-k_100g	float64	918
vitamin-c_100g	float64	140867
vitamin-b1_100g	float64	11154
vitamin-b2_100g	float64	10815
vitamin-pp_100g	float64	11729
vitamin-b6_100g	float64	6784
vitamin-b9_100g	float64	5240
folates_100g	float64	3042
vitamin-b12_100g	float64	5300
biotin_100g	float64	330
pantothenic-acid_100g	float64	2483
silica_100g	float64	38
bicarbonate_100g	float64	81
potassium_100g	float64	24748
chloride_100g	float64	158
calcium_100g	float64	141050
phosphorus_100g	float64	5845
iron_100g	float64	140462
magnesium_100g	float64	6253
zinc_100g	float64	3929
copper_100g	float64	2106
manganese_100g	float64	1620
fluoride_100g	float64	79
selenium_100g	float64	1168
chromium_100g	float64	20
molybdenum_100g	float64	11
iodine_100g	float64	259
caffeine_100g	float64	78
taurine_100g	float64	29
ph_100g	float64	49
fruits-vegetables-nuts_100g	float64	3036
collagen-meat-protein-ratio_100g	float64	165
cocoa_100g	float64	948
chlorophyl_100g	float64	0
carbon-footprint_100g	float64	268
nutrition-score-fr_100g	float64	221210
nutrition-score-uk_100g	float64	221210
glycemic-index_100g	float64	0
water-hardness_100g	float64	0

C

PROJET: NETTOYAGE DE DONNEES

C: Nettoyage de données

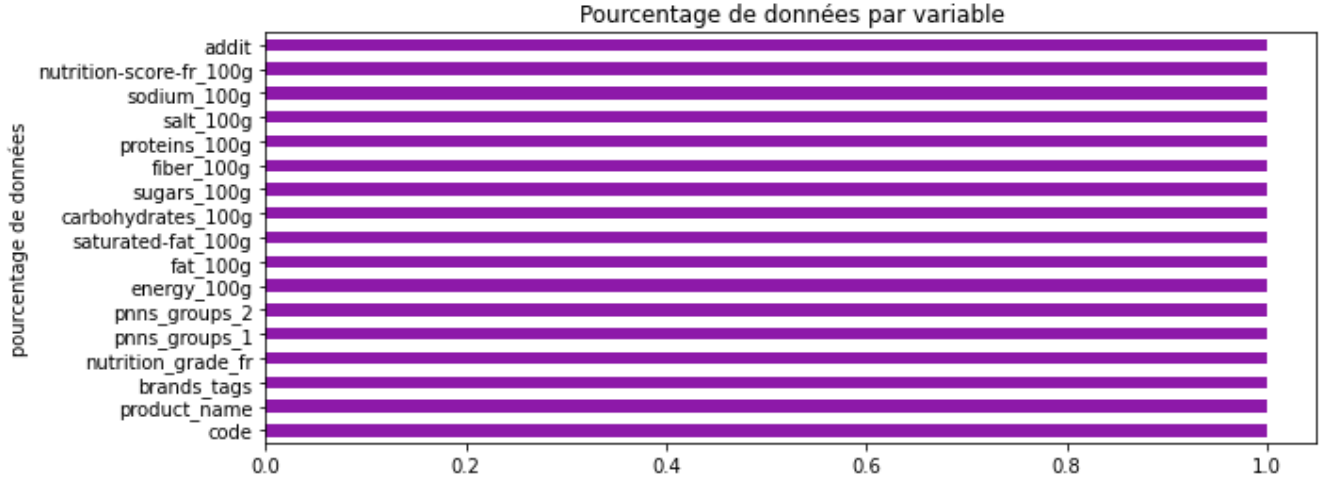
- **1-A: jeu volumineux de *320000 lignes/observations* et *162 colonnes/variables***
  - 10 champs d'**informations générales** : nom du produit, date de modification etc ...
  - 24 champs de **tags** : categorie de produit, localisation, origine
  - 29 champs d'**ingrédients et d'additifs**
  - 99 champs **quantitatifs d'informations nutritionnelles** pour 100gr de produit
- **1-B: filtrage des colonnes / variables par seuil de population**
  - De nombreuses variables sont **peu renseignées**.
    - nous décidons d'oublier les colonnes dont le seuil de remplissage est inférieur à un **certain seuil autour de 20%**.
    - nous passons de 162 variables à 54
- **1-C: suppression des doublons de lignes en filtrant sur la variable "CODE"**
  - filtre sur la variable "code" avec la fonction drop_duplicates correctement paramétré
  - suppression de 22 doublons
- **1-D: suppression des colonnes/variables redondantes**
  - 3 variables de temps : c'est 2 de trop
  - suppression de colonnes en doublons ou inutiles:
    - 'packaging','brands','categories','categories_fr','countries',
    - 'countries_tags','additives','additives_fr','states','states_fr',
    - 'main_category','nutrition-score-uk_100g
  - On uniformise les syntaxes.
    - tiret remplacé par un espace
    - mise en minuscule
    - n-a, na, unknown passés en np.nan
    - extraction des prefix de langues dans les colonnes de groupe
- **1-E: Réduction du nombre de lignes**
  - Vision sous l'angle d'un seul pays suffisamment représenté: **La France**
  - Suppression de la colonne pays et des variables peu représentées dans l'univers France
  - Suppression des lignes:
    - **sans données numériques de nutriment**
    - **sans caracteristique de nom de produit, ni de marque**

C: Nettoyage de données

- **1-E: Réduction du nombre de lignes**
  - Vision sous l'angle d'un seul pays suffisamment représenté: La France
  - Suppression de la colonne pays et des variables peu représentées dans l'univers France
  - Suppression des lignes:
    - sans données numériques de nutriment
    - sans caracteristique de nom de produit, ni de marque
- **1-F: Nettoyage des catégories**
  - 3 variables de catégories différentes sont disponibles:
    - main_category / pnns1 / pnns2
  - Création d'un algorithme permettant de remplir les pnns manquants.
    - 600/700 observations vont être regagnées
- **1-G: Nettoyage de bon sens des données numériques**
  - Suppression de lignes aux données aberrantes:
    - L'energie pour 100g ne peut etre superieur à 3700kj
    - Les valeurs nutritionnelles ne peuvent être négatives (sauf pour le nutriscore qui peut aller à -15)
    - Les valeurs nutritionnelles ne peuvent depasser les 100g
      - 100gr de sel doit représenter 38.8g de sodium
      - les graisses saturées doivent etre inferieures en quantité aux graisses
      - le nutriscore doit etre compris entre -15 et +40
- **1-H: Nettoyage des outliers des données numériques**
  - Filtre IQR
    - filtre appliqué par catégorie pnns 2 afin d'etre plus fin
    - 1/3 des lignes sont retirées
- **1-I: Passage du nombre d'additif en booléen (avec/sans)**
  - On simplifie la variable nombre d'additifs
- **1-J: Suppression des variables sans interet pour notre analyse**
- **1-K: Imputation des données quantitatives manquantes**
  - un bon tiers de la variable fibre , élément semblant important, n'est pas renseigné.
  - Pour les analyses suivantes, de nombreuses techniques ne peuvent fonctionner avec des données manquantes comme l'**ACP**
  - Nous allons faire appel a un **simple imputer** / median qui est plus rapide que le **KNN imputer**
    - ce dernier imputera par groupe pnns2 afin d'etre relativement fin

C: Nettoyage de données

```
Résumé sur la table néttoyée:
20613  lignes / observations
17  colonnes / variables
3 variables descriptives: code - nom de produit - nom de marque
2 variables de catégories: pnns1 et pnns2
9 variables quantitatives de nutriments
1 variable booléenne: presence ou non  d'additif
1 variable quantitative de nutriscore
1 variable catégéorie de nutrigrade
```



D ANALYSE STATISTIQUE

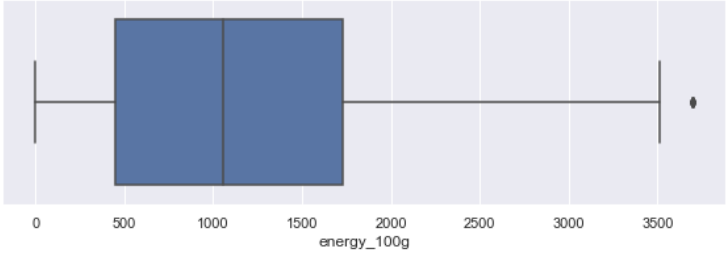
D: Analyse statistique

1-A: DESCRIPTIVE - Quantitative - Univariée

statistiques descriptives des données quantitatives

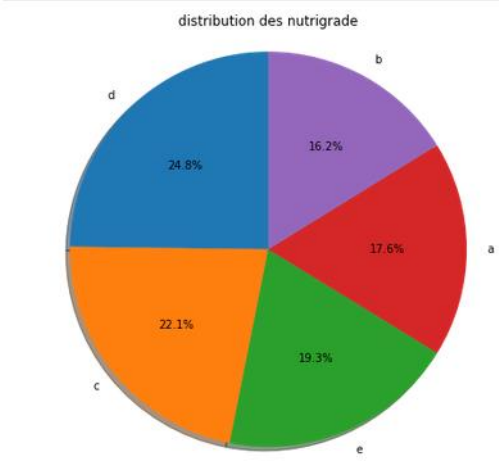
	energy_100g	fat_100g	saturated-fat_100g	carbohydrates_100g	sugars_100g	fiber_100g	proteins_100g	salt_100g	sodium_100g	nutrition-score-fr_100g	addit
count	20,613.00	20,613.00	20,613.00	20,613.00	20,613.00	20,613.00	20,613.00	20,613.00	20,613.00	20,613.00	20,613.00
mean	1,131.77	13.66	5.93	27.30	12.02	1.91	8.53	0.80	0.31	8.37	0.63
std	752.07	14.90	8.42	27.11	16.55	2.01	7.12	0.73	0.29	9.26	0.48
min	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-12.00	0.00
25%	448.00	2.00	0.50	3.00	0.80	0.20	3.30	0.13	0.05	1.00	0.00
50%	1,055.00	7.40	2.30	13.70	3.50	1.50	6.60	0.70	0.28	7.00	1.00
75%	1,730.00	23.50	9.00	54.00	18.00	2.90	11.80	1.20	0.47	16.00	1.00
max	3,700.00	100.00	70.00	93.00	77.40	15.20	33.00	4.61	1.81	29.00	1.00

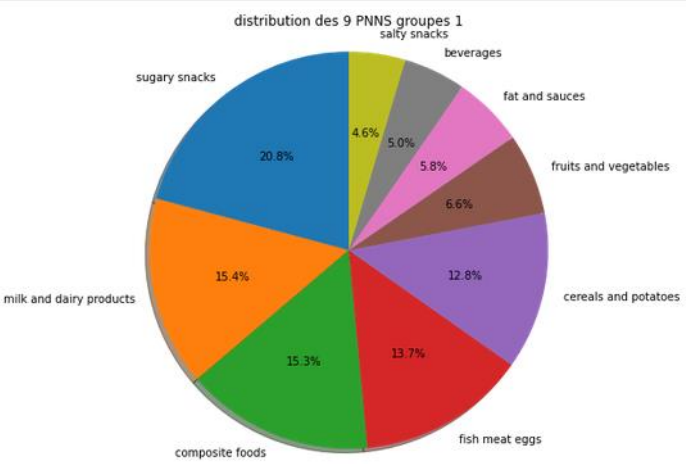
BoxPlot:VARIABLES
- (●) energy_100g
- ( ) fat_100g
- ( ) saturated-fat_100g
- ( ) carbohydrates_100g
- ( ) sugars_100g
- ( ) fiber_100g
- ( ) proteins_100g
- ( ) salt_100g
- ( ) sodium_100g
- ( ) nutrition-score-fr_100g
- ( ) addit

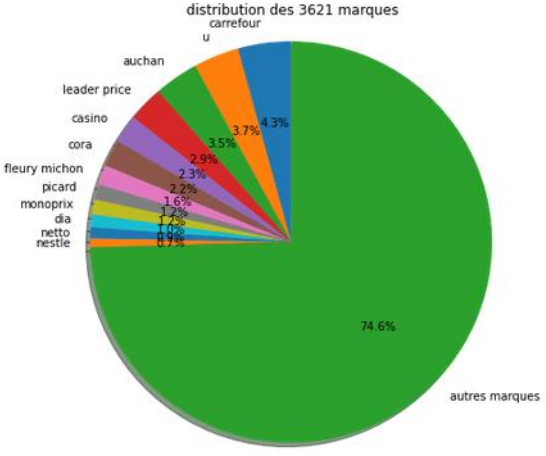


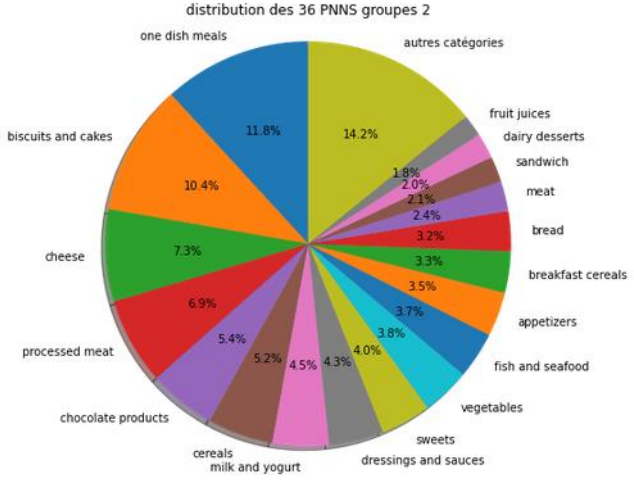
D: Analyse statistique

1-B:DESCRIPTIVE - Catégories - Univariée



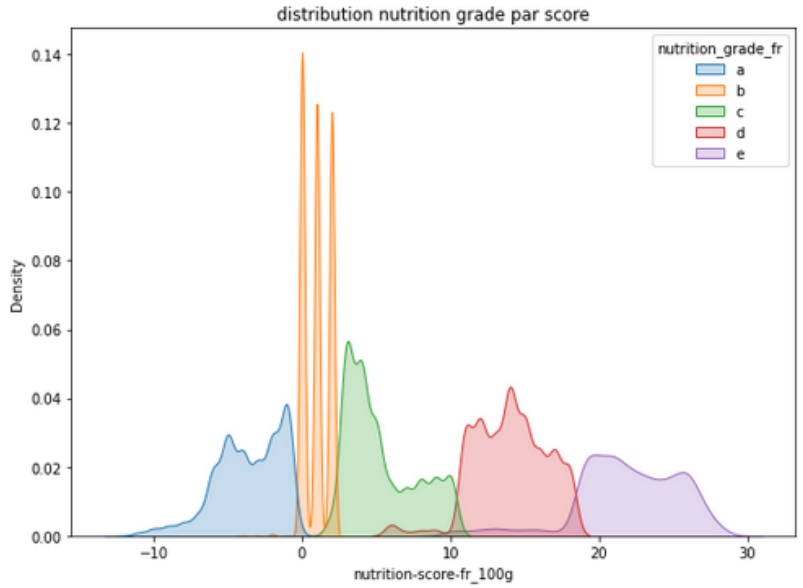


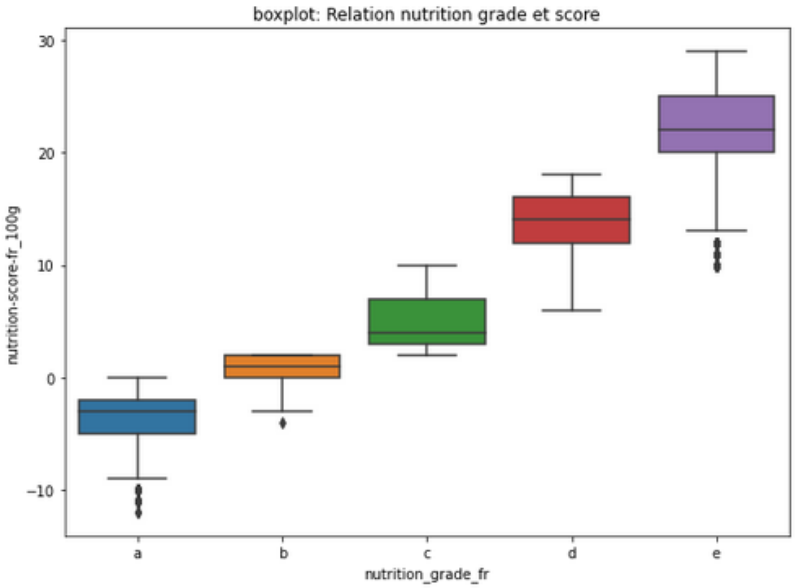




- Toutes les 5 catégories du nutrigrade sont bien représentées:
  - les **meilleurs grades a et b** sont legerement moins bien représentés que les mauvais grades d,c,e
- Toutes les 9 catégories du PNNS groupe 1 sont également bien représentés:
  - **les sucreries représentent 20%** des données alors que les snacks salés, les boissons, le gras et sauces ainsi que **les fruits et legumes ne représentent chacun que 5%** des données
- Concernant les 36 catégories du PNNS groupe 2:
  - les **gateaux/biscuits ainsi que les plats cuisinés** représentent à eux 2 **20%** des données.
- Concernant les 3600 marques, les marques des hypermarchés sont le mieux représentées. Mais c'est une variable avec trop de modalité qui ne nous apportera pas d'information utile

D: Analyse statistique
1-C:DESCRIPTIVE - Bivariée





statistique descriptive des grades/scores

	nutrition-score-fr_100g							
	count	mean	std	min	25%	50%	75%	max
nutrition_grade_fr								
a	3,632.00	-3.65	2.25	-12.00	-5.00	-3.00	-2.00	0.00
b	3,334.00	0.95	0.84	-4.00	0.00	1.00	2.00	2.00
c	4,552.00	5.29	2.36	2.00	3.00	4.00	7.00	10.00
d	5,111.00	13.96	2.48	6.00	12.00	14.00	16.00	18.00
e	3,984.00	21.89	3.53	10.00	20.00	22.00	25.00	29.00

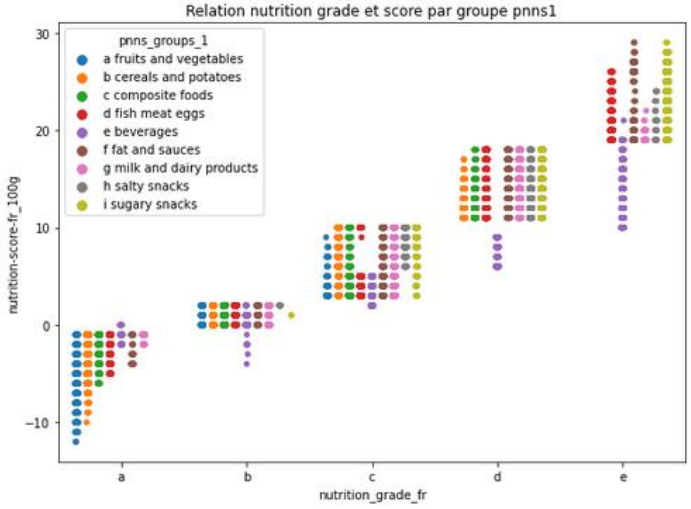
2-C-1: Relation visuelle claire entre nutrigrade et nutriscore

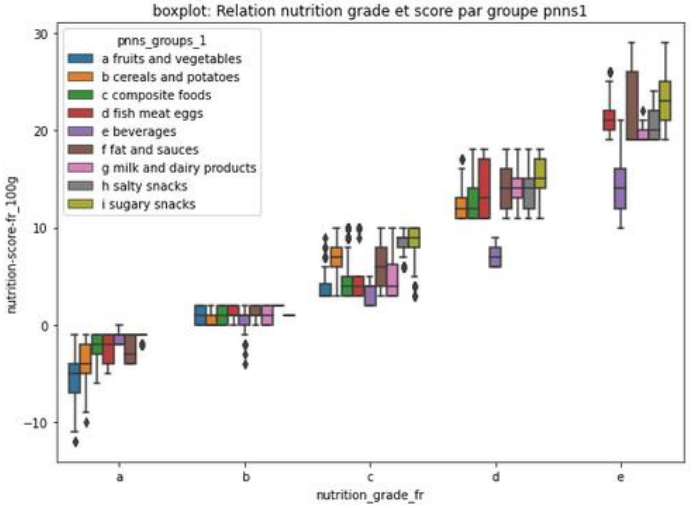
- chaque grade correspond à un intervalle de nutriscore
  - les scores les moins élevés correspondent aux meilleurs grades alors que les scores les plus élevées correspondent aux moins bons grade

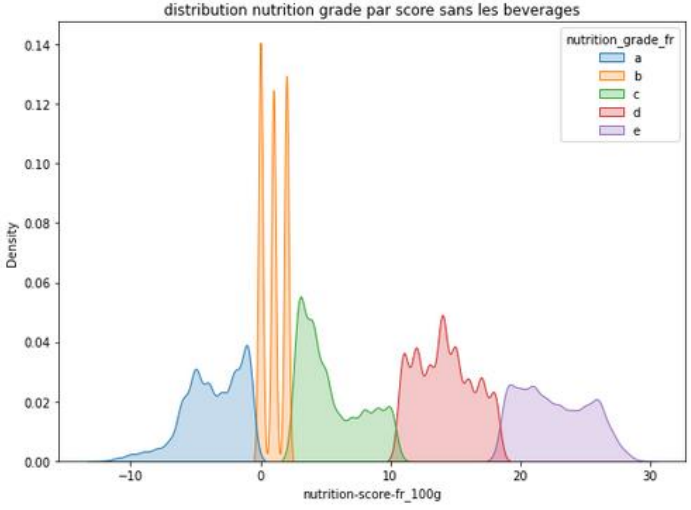
D: Analyse statistique
1-C:DESCRIPTIVE - Bivariée

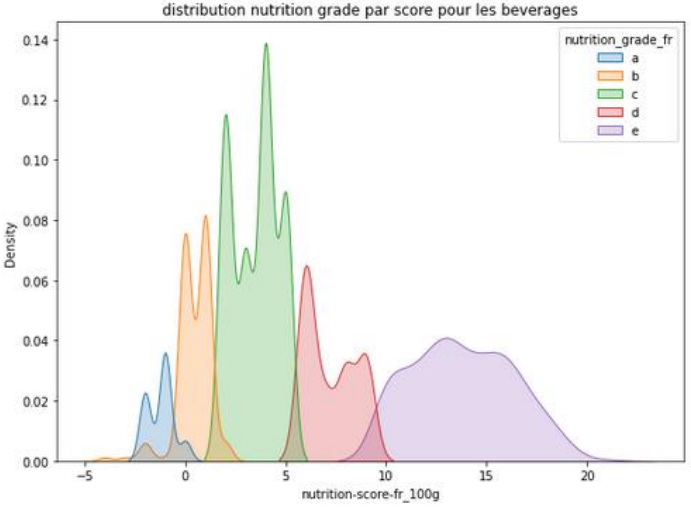
2-C-2: Relation entre nutrigrade et nutriscore : ventilation par PNNS groupe 1

- on s'aperçoit d'une relation grade/score légèrement différente par catégories
  - **la relation ne semble pas avoir exactement les mêmes intervalles de score pour les boissons**
- Les fruits et légumes n'ont pas de mauvais grades d et e comme on peut s'en douter
- Les sucreries et snacks salés n'ont pas de bons grades a et b





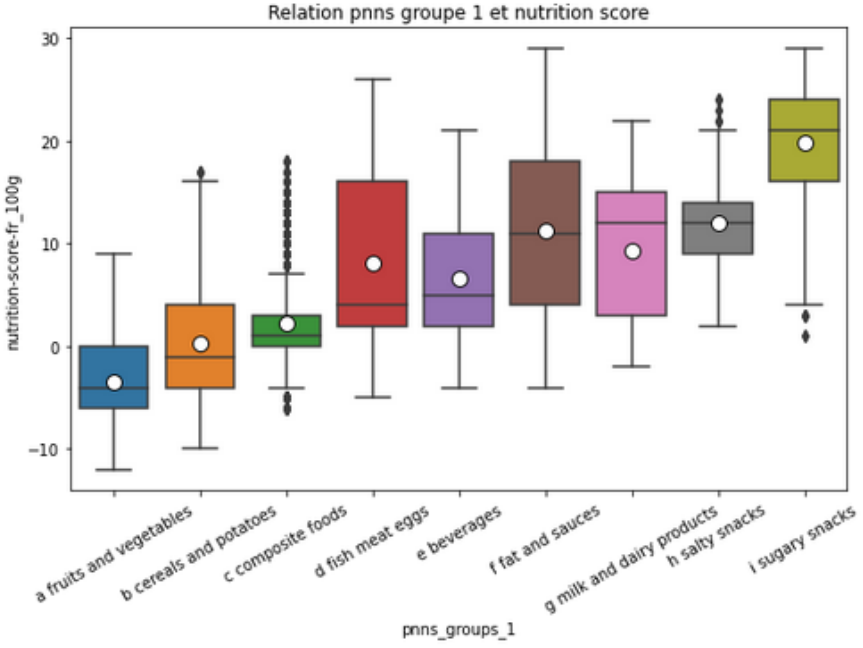


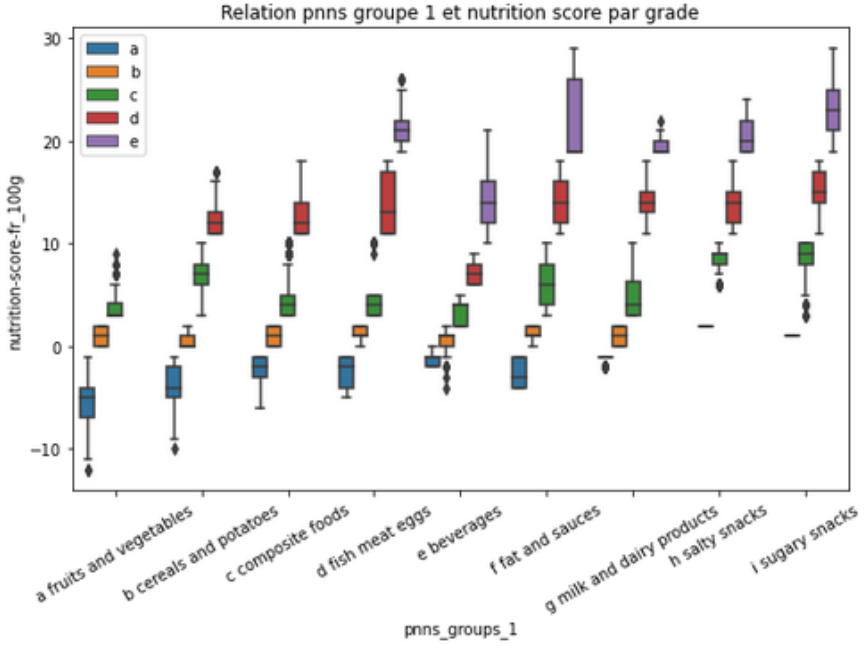


D: Analyse statistique
1-C:DESCRIPTIVE - Bivariée

2-C-3: Relation entre pnns groupe 1 et nutrigrade/nutriscore

- Une relation existe entre nutri score et catégorie de produit
  - les fruits et légumes , les céréales ont les nutriscores les plus faibles donc les meilleurs grades
    - les fruits et légumes ne sont pas représentés dans les moins bons grades
  - à l'opposé les sucreries sont très mal placées
    - les sucreries et les snacks salés ne sont pas représentés dans les meilleures nutrigrades a/b

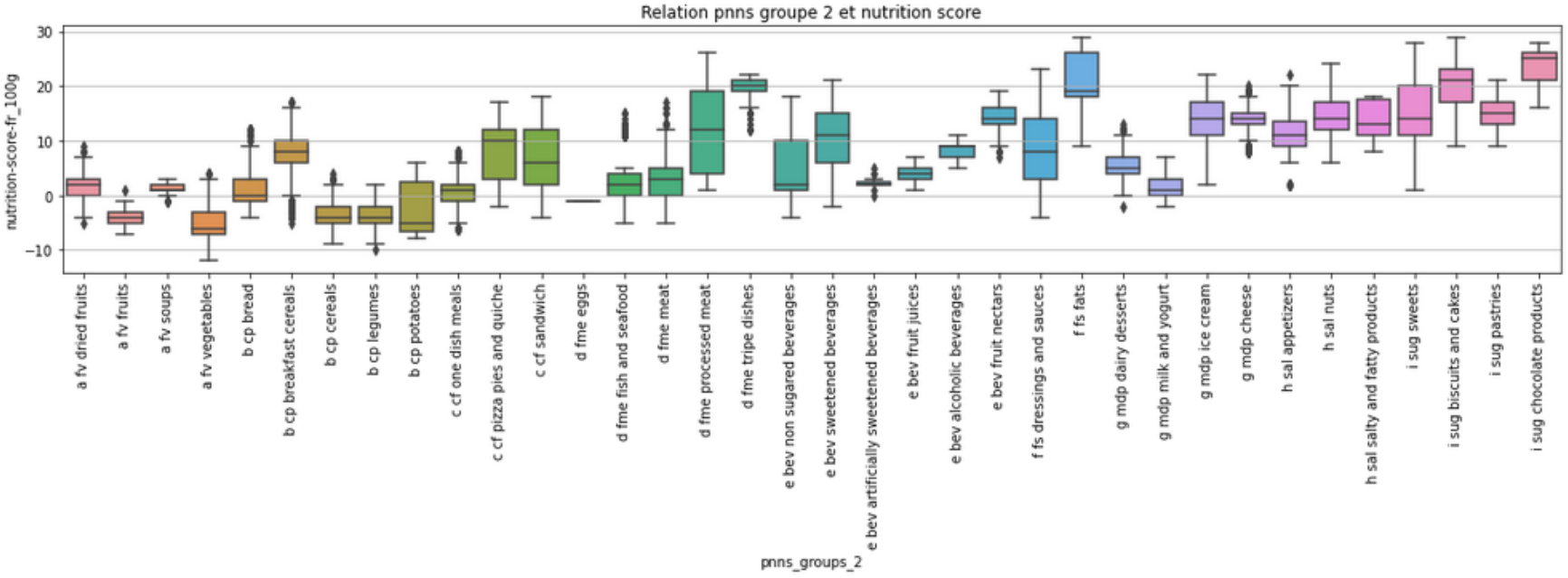




D: Analyse statistique
1-C:DESCRIPTIVE - Bivariée

2-C-4: Relation entre pnns groupe 2 et nutriscore

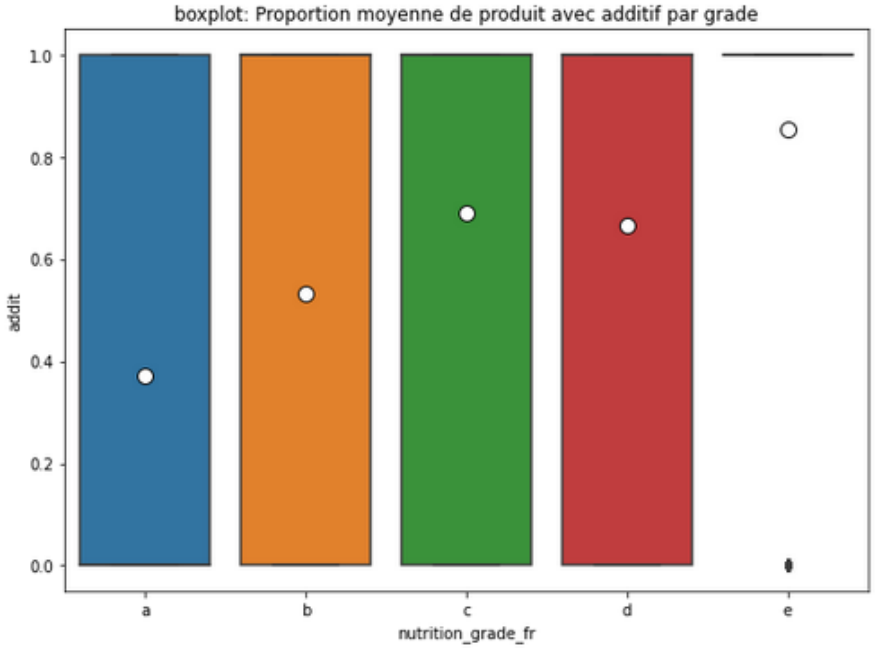
•Ce qui est vrai pour le Pnns groupe 1 l'est également pour le Pnns groupe 2 qui constitue un découpage plus fin

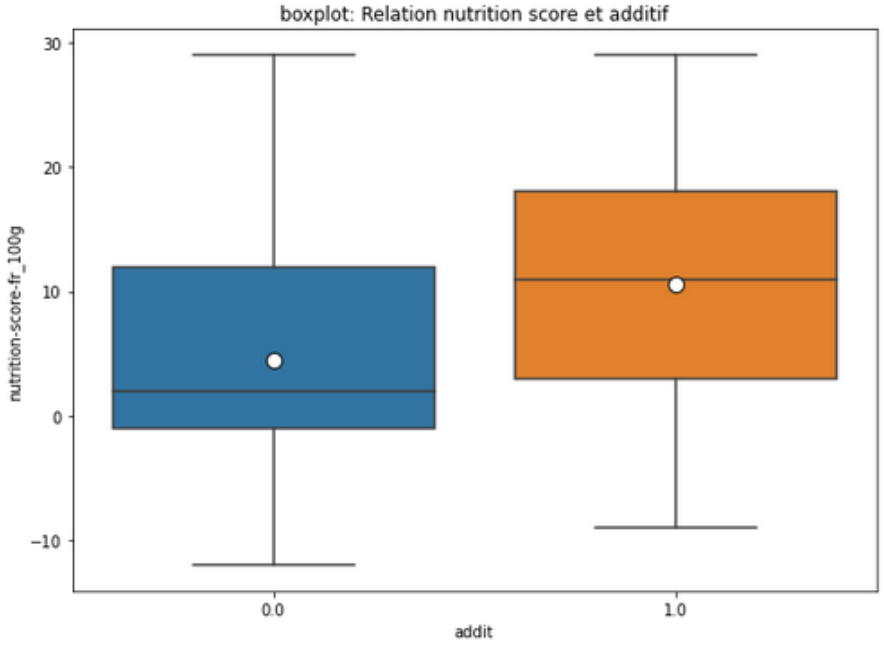


D: Analyse statistique
1-C:DESCRIPTIVE - Bivariée

2-C-5: Relation entre additifs et nutrigrade

- on observe une relation entre la présence ou non d'additif avec le nutri score et le nutri grade.
  - les produits du grade e ont en très grande majorité des additifs .
  - le groupe a possède la plus faible proportion de présence d'additif. Cette proportion moyenne augmente avec les grades





D: Analyse statistique
1-C:DESCRIPTIVE - Bivariée

2-C-5: Relation entre données quantitatives nutritionnelles et nutriscore¶

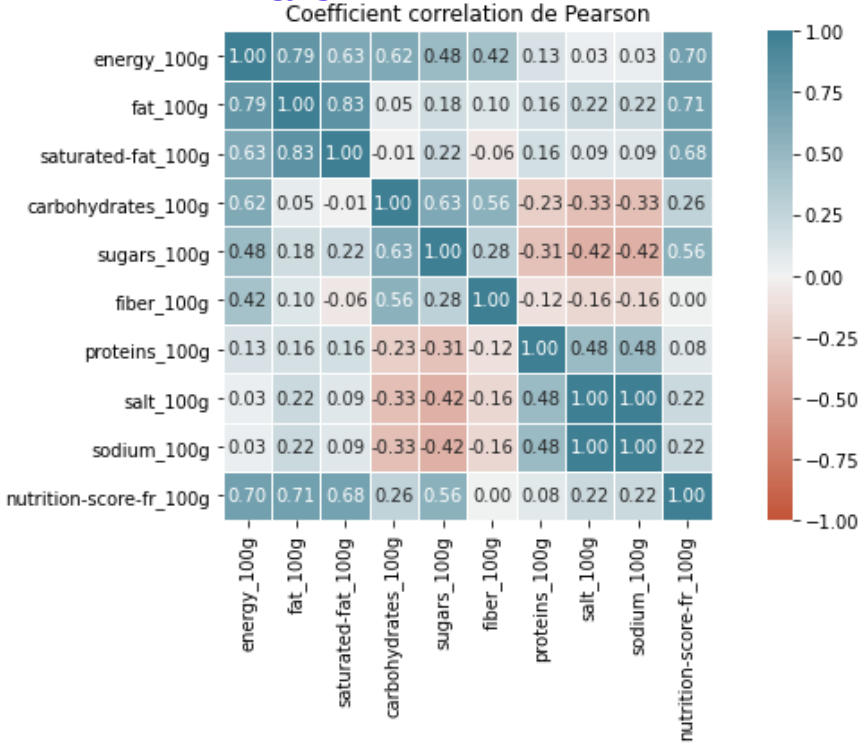
2-C-5-A: Nuages de points bivariés

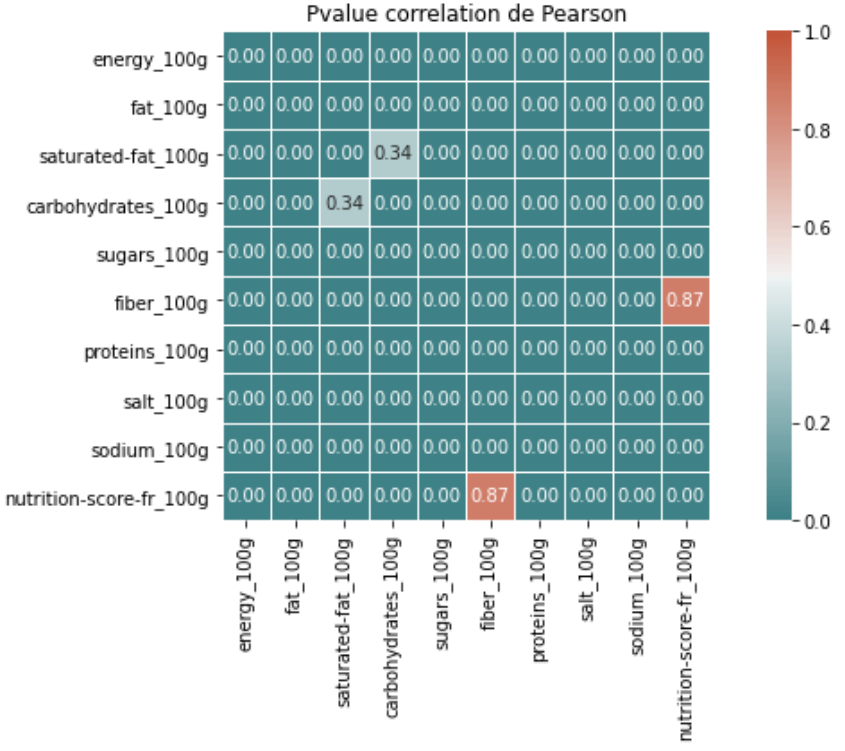
•Quelques relations qu'on devine visuellement

2-C-5-B: Relation lineaire ? - corrélation de Pearson

•Des corrélations de **pearson** significatives

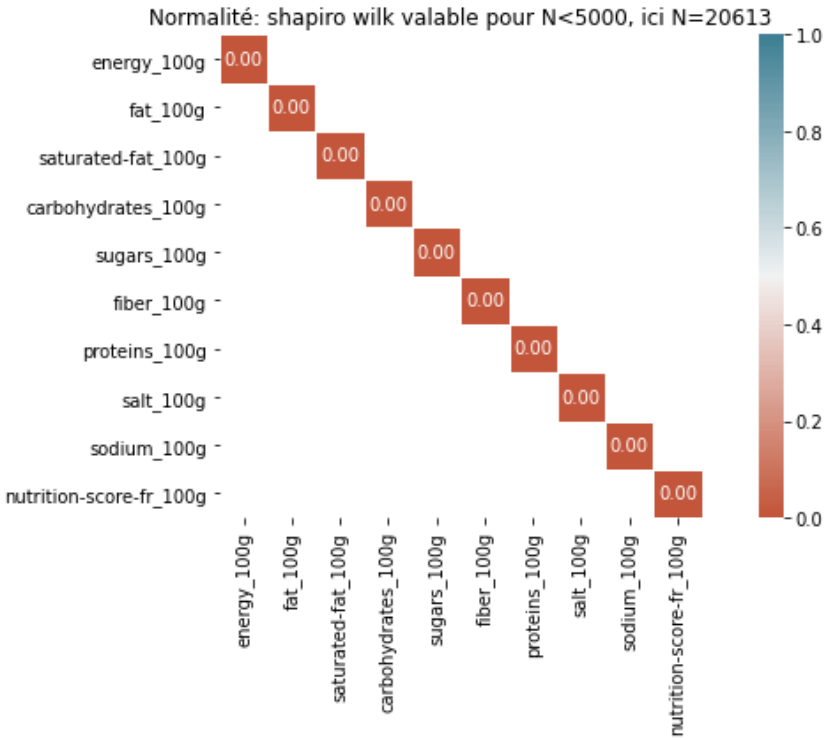
- corrélation positive avec le nutriscore :
  - Energy, gras, sucre
- corrélation positive entre gras et Energie
- corrélation positive entre protéine et sel
- corrélation positive entre sucre et hydrates de carbone
- corrélation positive entre fibre et hydrates de carbone
- corrélation négative entre sucre et sel
- corrélation parfait entre sel et sodium

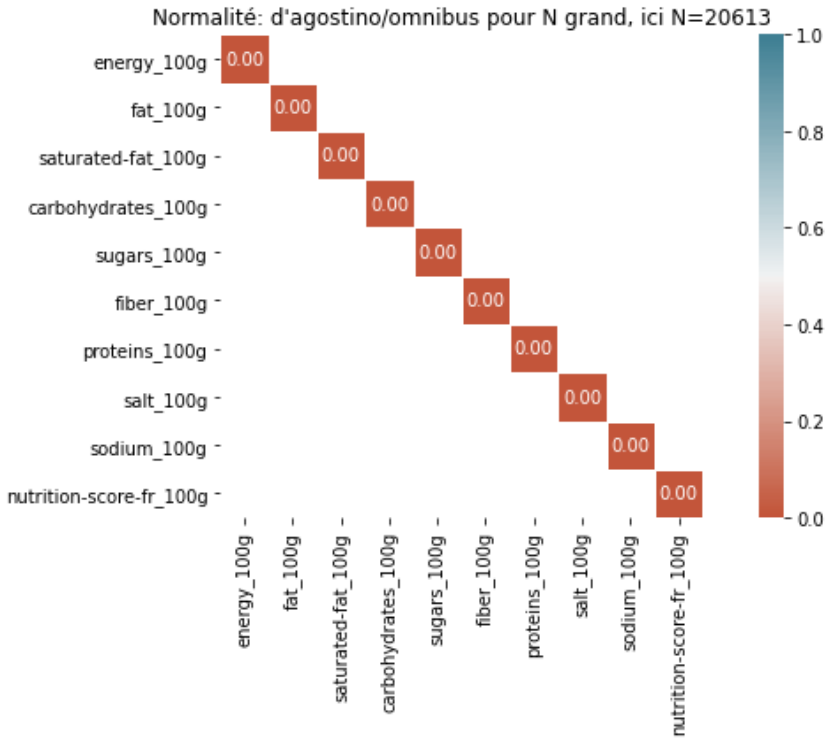




D: Analyse statistique
1-C:DESCRIPTIVE - Bivariée

- **Mais l'hypothèse de normalité non respectée ne permet pas de valider les résultats de relation de linearité**



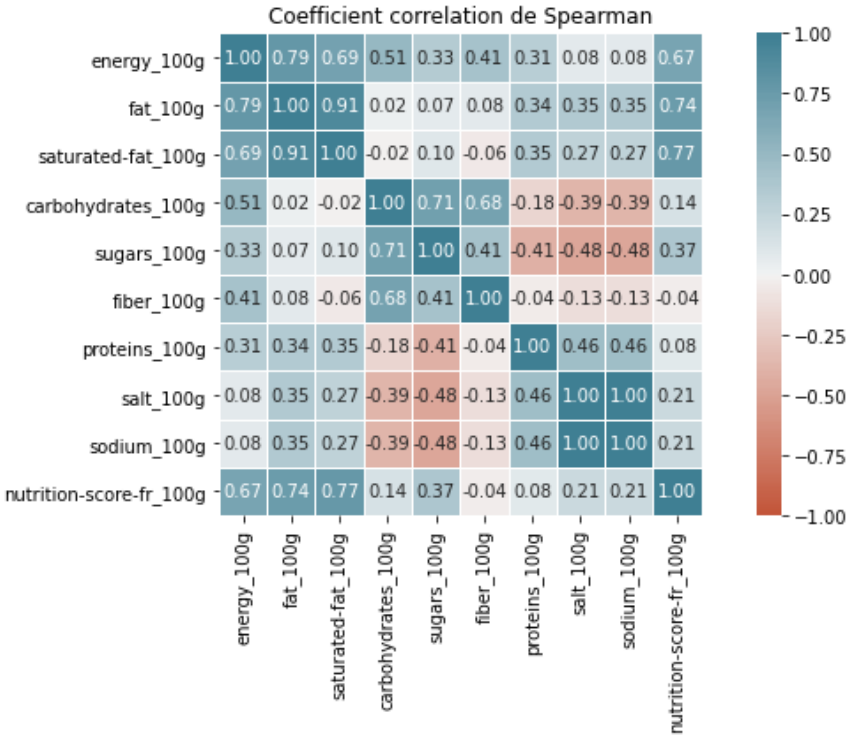


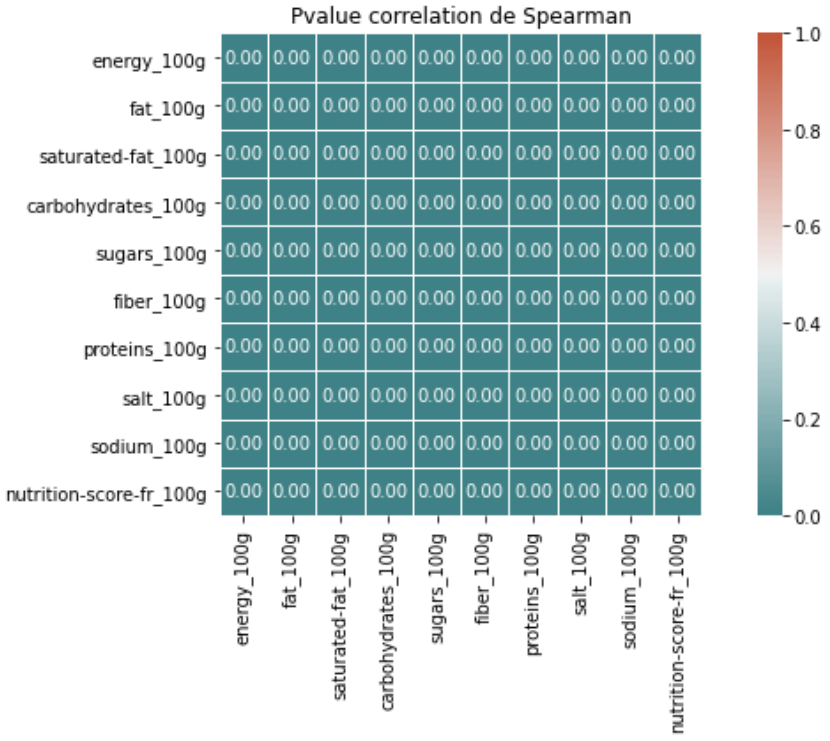
D: Analyse statistique
1-C:DESCRIPTIVE - Bivariée

2-C-5: Relation entre données quantitatives nutritionnelles et nutriscore.

2-C-5-C: Relation monotone non paramétrique ? - corrélation de Spearman

- Corrélation de rang de spearman
  - les relations précédemment décrites mais monotones, cette fois ci, sont significatives





D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

•On peut identifier deux grandes familles de méthode d'analyse multivariée

- **les méthodes descriptives**
  - visant à structurer et résumer l'information:
    - ACP, AFC-ACM, analyse factorielle, clustering, MDS (positionnement multidimensionnel)
- **les méthodes explicatives**
  - visant à expliquer une ou des variables dites « dépendantes » (variables à expliquer) par un ensemble de variables dites « indépendantes » (variables explicatives).
    - analyse de régression multiple
    - analyse de variance multivariée (ANOVA: bivariée)
    - analyse discriminante
    - régression logistique
    - arbre de décision
    - réseau de neurones, etc...

2-D-1: Analyse explicative entre variables catégorielles et score quantitatif : ANOVA

•l'analyse de la variance [ANOVA : analysis of variance] est un ensemble de modèles statistiques utilisés pour vérifier si les moyennes des groupes [modalités d'une variable explicative] proviennent d'une même population.

- Ce test s'applique lorsque l'on mesure une ou plusieurs variables explicatives catégorielle (facteurs, leurs différentes modalités étant parfois appelées « niveaux ») qui ont de l'influence sur la loi d'une variable continue à expliquer.
- On parle d'analyse à un facteur lorsque l'analyse porte sur un modèle décrit par un seul facteur de variabilité, d'analyse à deux facteurs ou d'analyse multifactorielle sinon.
- **L'analyse de la variance permet d'étudier le comportement d'une variable quantitative à expliquer en fonction d'une ou de plusieurs variables qualitatives, aussi appelées nominales catégorielles.**

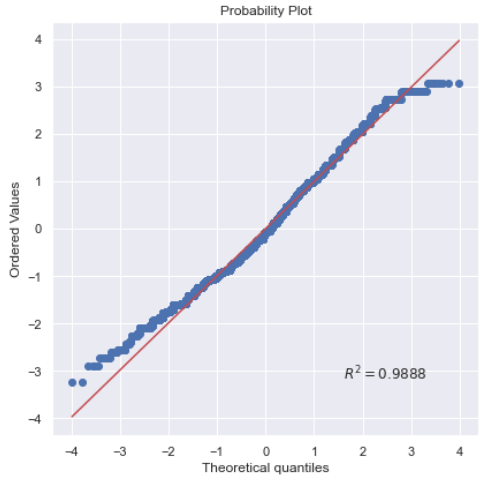
D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

2-D-1-a: Analyse explicative entre la variable catégorielle: GROUPE 1 et score quantitatif: NUTRISCORE : ANOVA

- L'**anova** montre des valeurs significatifs mais l'hypothèse de normalité des résidus n'est pas respectée
- L'**anova non paramétrique Kruskall wallis** valide le fait que la variable qualitative pnns groupe 1 possède une valeur explicative sur le nutriscore
- Le test **post hoc non paramétrique** montrent que toutes les modalités sont différentes 2 à 2

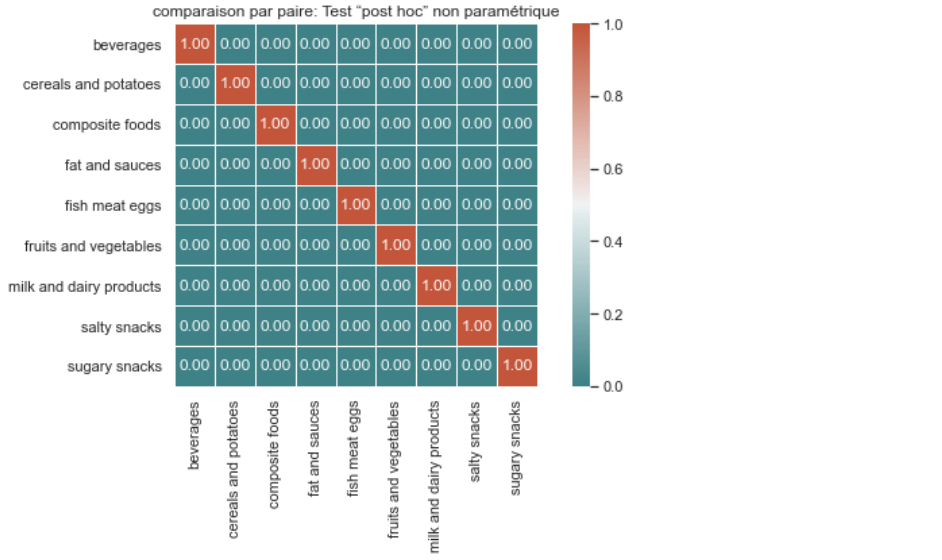
	sum_sq	df	F	PR(>F)
C(group1)	12,468.79	8.00	3,943.57	0.00
Residual	8,143.21	20,604.00	NaN	NaN

```
pvalue <0.05 donc significative => H0 est rejetée
il existe au moins un groupe dont la moyenne s'écarte des autres moyennes

Plot des residus: distribution normale ou non?
```




```
kurtosis de -0.30 <0 legerement platykurtique et skewness de 0.31 <0 legerement décal
la distribution semble visuellement au QQplot proche d etre normale mais pas complète

test normalité des residus - test d'Agostino/omnibus pour echantillon moyen grand
Agostino Pvalue:0.00
Si P-VALUE < 0.05: hypothèse H0 est rejetée (i.e. peu probable d etre normalement distribuées).
```

```
Hypothèse de normalité non respectée => ANOVA NON PARAMETRIQUE
 kruskall wallis:  0.0  Si pval< 0.05, alors on peut parler de différence significative

post hoc: Les modalités sont bien toutes différentes entre elles 2 à 2
```



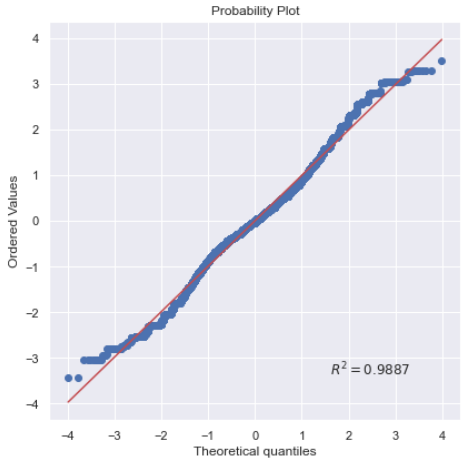
D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

2-D-1-b: Analyse explicative entre la variable catégorielle: GROUPE 2 et score quantitatif: NUTRISCORE : ANOVA¶

- L'**anova** montre des valeurs significatifs mais l'hypothèse de normalité des résidus n'est pas respectée
- L'**anova non paramétrique Kruskall wallis** valide le fait que la variable qualitative **pnns groupe 2** possède une valeur explicative sur le nutriscore

	sum_sq	df	F	PR(>F)
C(group2)	16,499.05	35.00	2,358.41	0.00
Residual	4,112.95	20,577.00	NaN	NaN

```
pvalue <0.05 donc significative => H0 est rejetée
il existe au moins un groupe dont la moyenne s'écarte des autres moyennes

Plot des residus: distribution normale ou non?
```

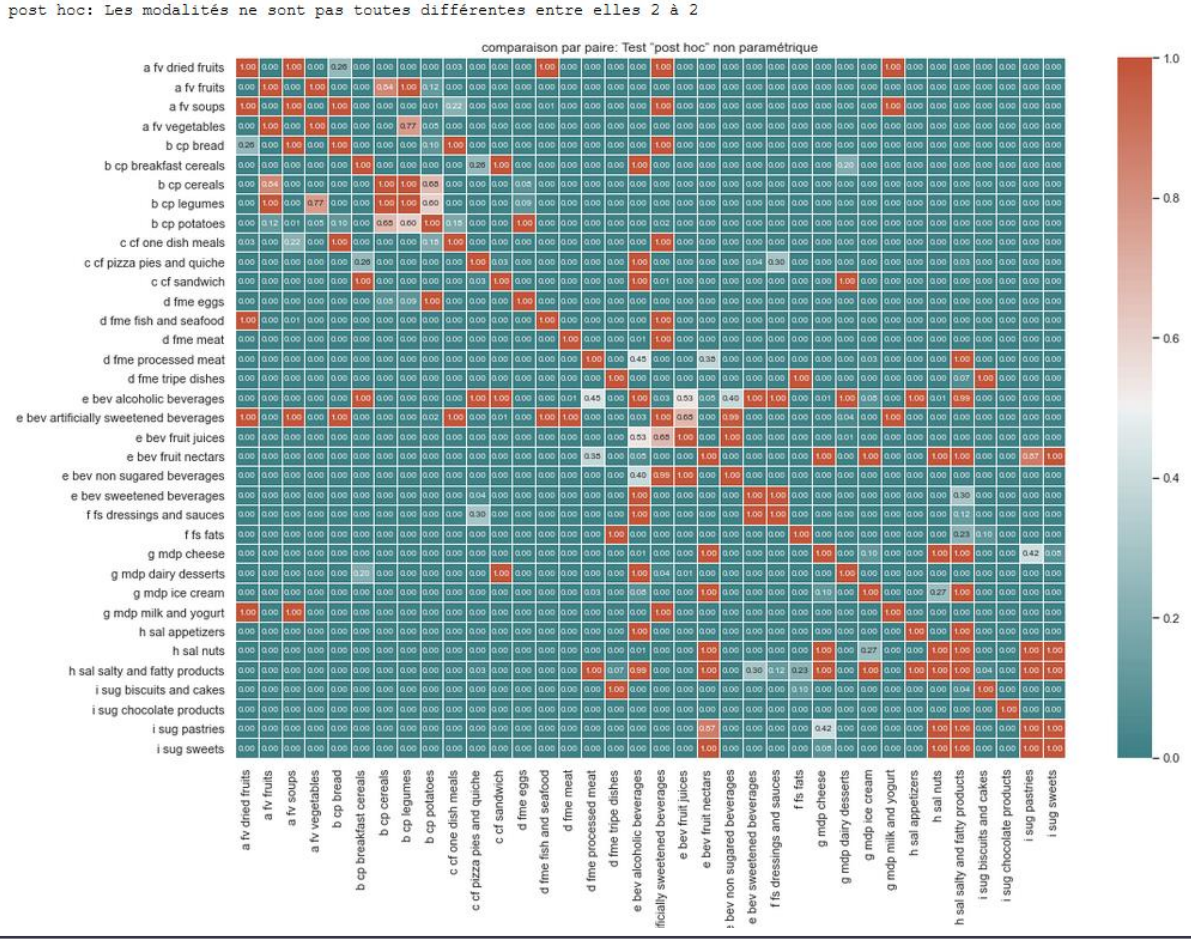



```
kurtosis de 0.55 >0 legerement leptokurtique et skewness de 0.06 >0 legerement décalée à gauche
la distribution semble visuellement au QQplot proche d etre normale mais pas complètement

test normalité des residus - test d'Agostino/omnibus pour echantillon moyen grand
Agostino Pvalue:0.00
Si P-VALUE < 0.05: hypothèse H0 est rejetée (i.e. peu probable d etre normalement distribuées).

Hypothèse de normalité non respectée => ANOVA NON PARAMETRIQUE
 kruskall wallis:  0.0  Si pval< 0.05, alors on peut parler de différence significative
```

D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

2-D-1-b: Analyse explicative entre la variable catégorielle: GROUPE 2 et score quantitatif: NUTRISCORE : ANOVA¶

•Le test **post hoc non paramétrique** montrent que toutes les modalités sont différentes 2 à 2



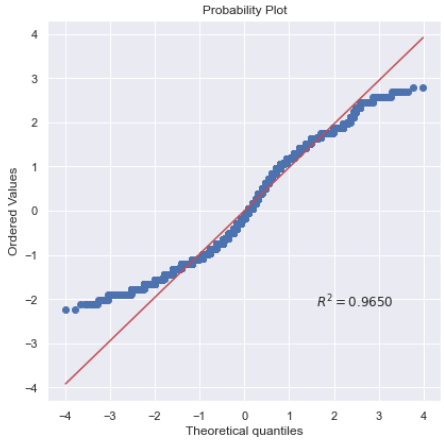
D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

2-D-1-c: Analyse explicative entre la variable catégorielle: additif(présence ou non) et score quantitatif: NUTRISCORE : ANOVA

- L'**anova** montre des valeurs significatifs mais l'hypothèse de normalité des résidus n'est pas respectée
- L'**anova non paramétrique Kruskall wallis** valide le fait que la variable qualitative **présence d'additif** possède une valeur explicative sur le nutriscore

	sum_sq	df	F	PR(>F)
C(addit)	2,100.23	1.00	2,338.39	0.00
Residual	18,511.77	20,611.00	NaN	NaN

```
pvalue <0.05 donc significative => H0 est rejetée
il existe au moins un groupe dont la moyenne s'écarte des autres moyennes

Plot des residus: distribution normale ou non?
```

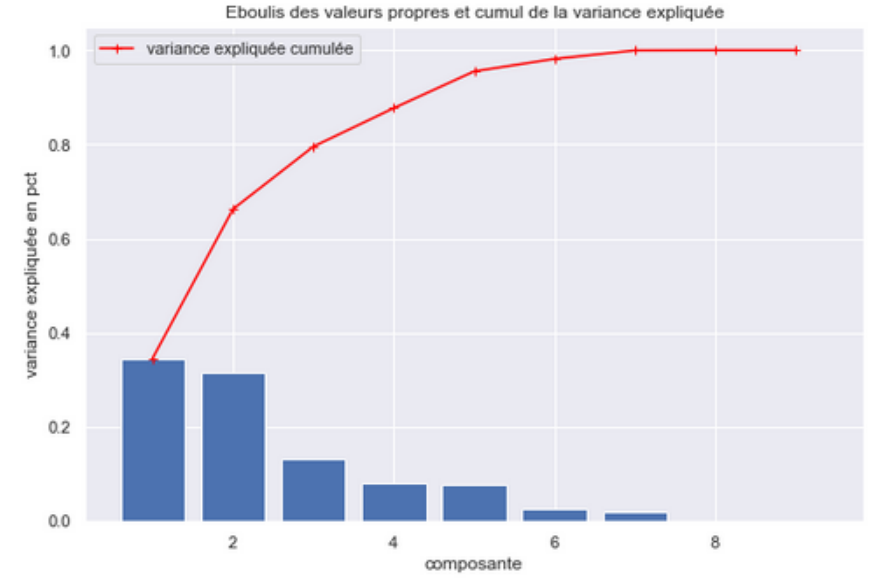


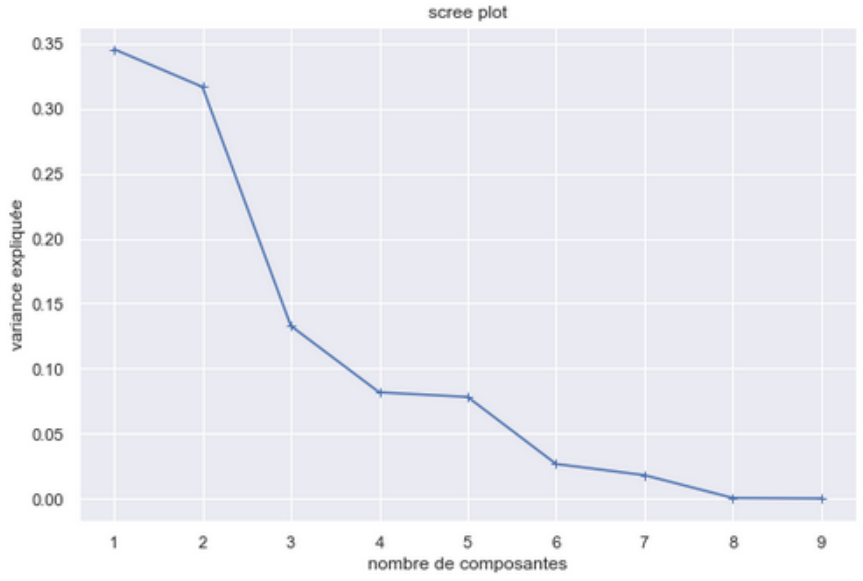

```
kurtosis de -0.97 <0 legerement platykurtique et skewness de 0.26 >0 legerement décalée à gauche
la distribution n'est pas normale au QQplot

test normalité des residus - test d'Agostino/omnibus pour echantillon moyen grand
Agostino Pvalue:0.00
Si P-VALUE < 0.05: hypothèse H0 est rejetée (i.e. peu probable d etre normalement distribuées).

Hypothèse de normalité non respectée => ANOVA NON PARAMETRIQUE
 kruskall wallis:  0.0  Si pval< 0.05, alors on peut parler de différence significative
```

D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

2-D-2: Analyse descriptive multivariée : ACP : réduction de dimension des variables quantitatives

- entre **85 et 90% de la variance** est expliquée par **les 4 premiers facteurs**
  - Par ailleurs la qualité de représentation des variables sur les composantes , exprimée par le cos2, est quasi nulle à partir de la composante 5
- les 2 premiers facteurs sont relativement corrélés au nutriscore
  - La première composante est une composante gras/sucrée
  - La 2 -ème composantes est plutôt sel/gras
  - La 3 -ème est plutôt une composante anti fibre
  - et la 4 -ème anti protéine
- Au final le **premier plan factoriel** représente un angle de vision **"mal-bouffe"**
- Alors que le **2éme** est plutot **anti "nourriture saine"**

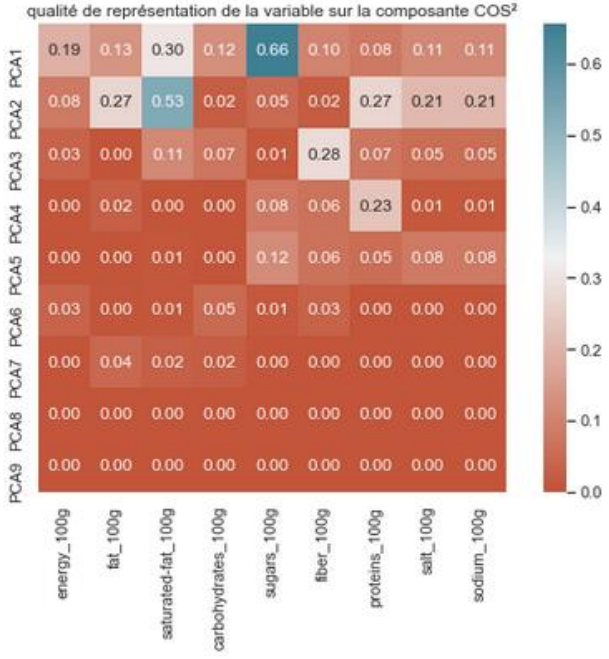


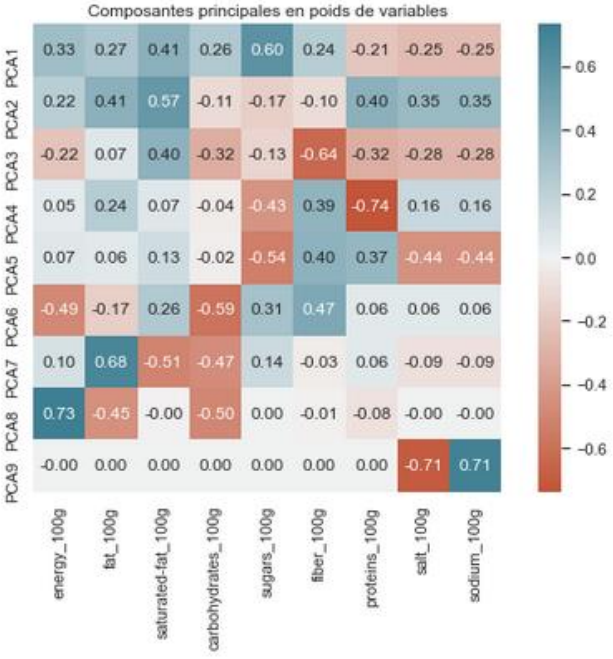


D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

2-D-2: Analyse descriptive multivariée : ACP : réduction de dimension des variables quantitatives

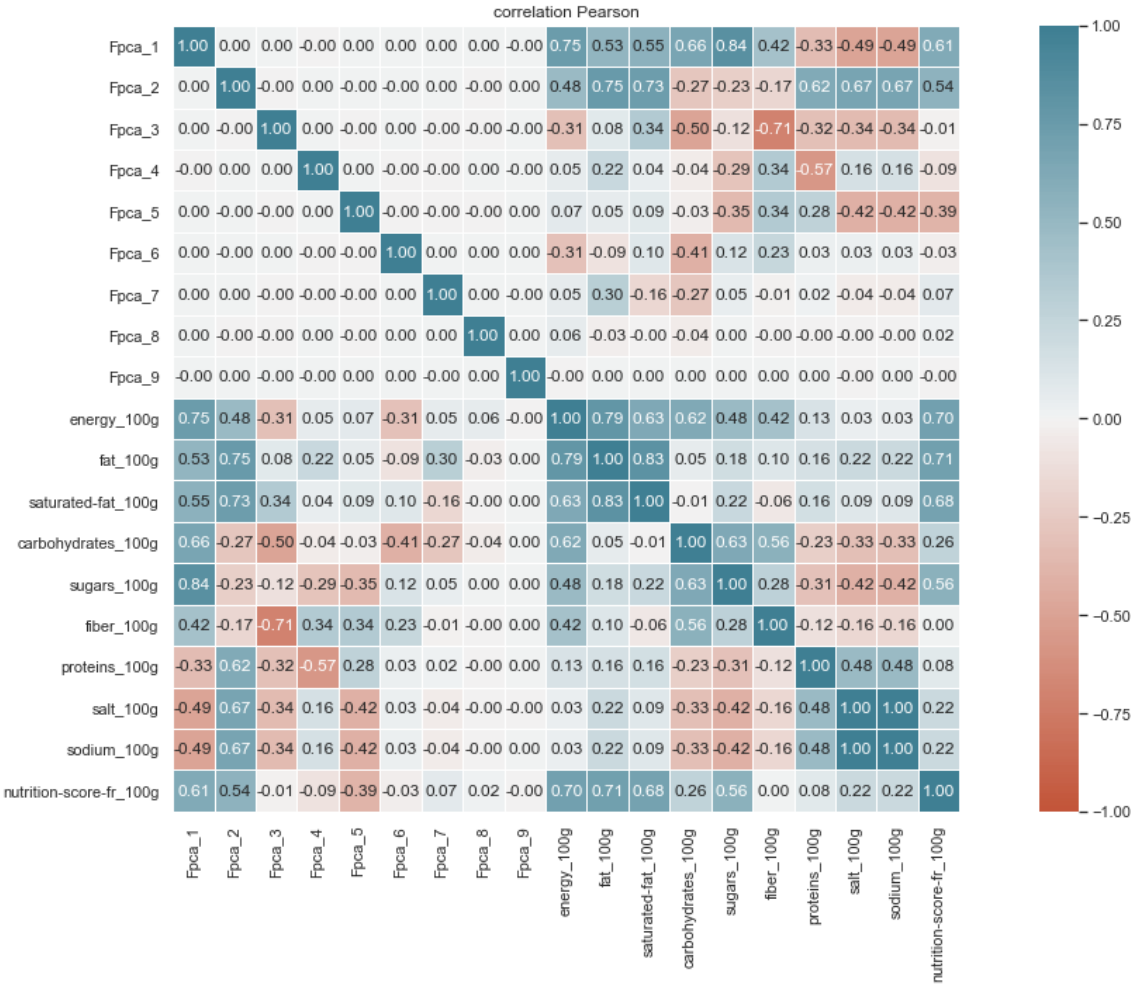
- les 2 premiers facteurs sont relativement corrélés au nutriscore
  - La première composante est une composante gras/sucrée
  - La 2 -ème composantes est plutôt sel/gras
  - La 3 -ème est plutôt une composante anti fibre
  - et la 4 -ème anti protéine
- Au final le **premier plan factoriel** représente un angle de vision **"mal-bouffe"**
- Alors que le **2éme** est plutôt **anti "nourriture saine"**





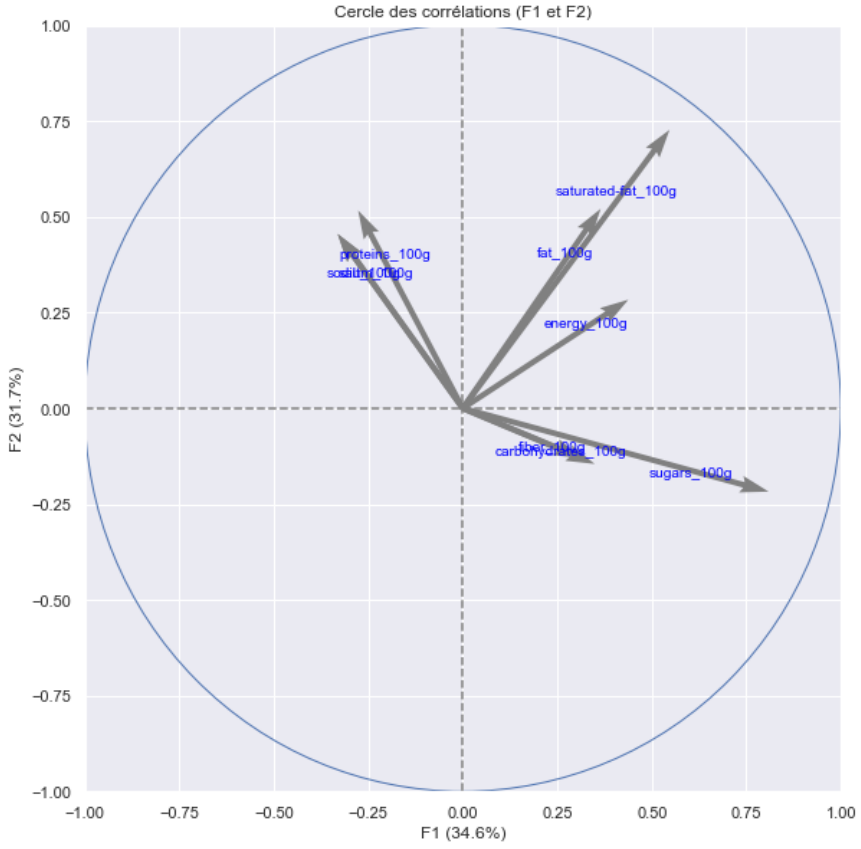
D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

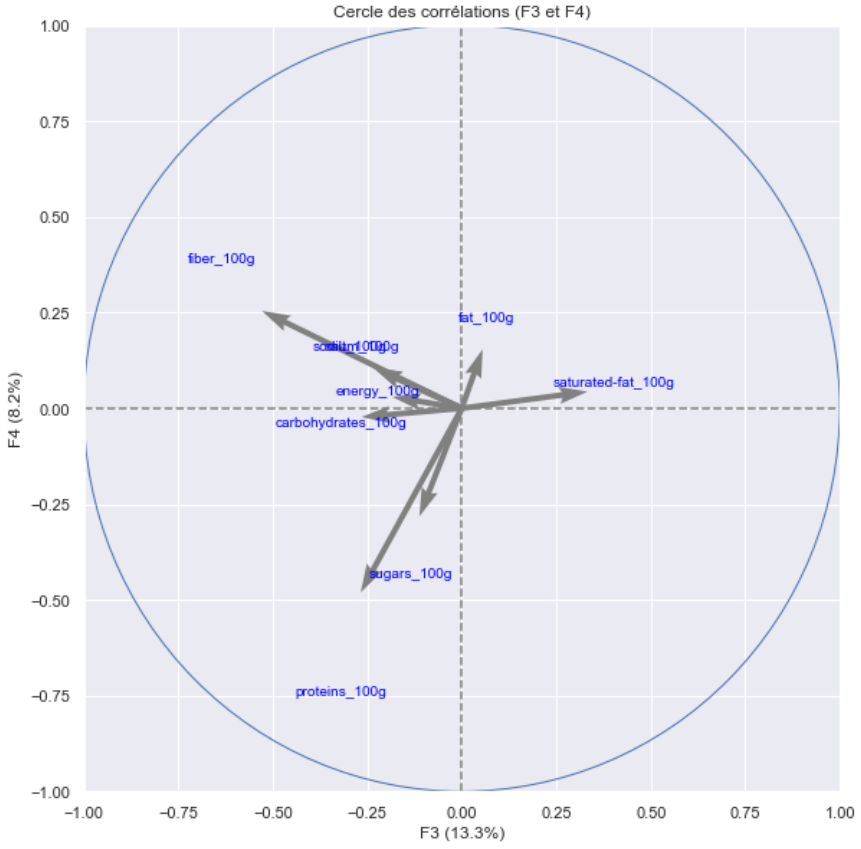
2-D-2: Analyse descriptive multivariée : ACP : réduction de dimension des variables quantitatives



D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

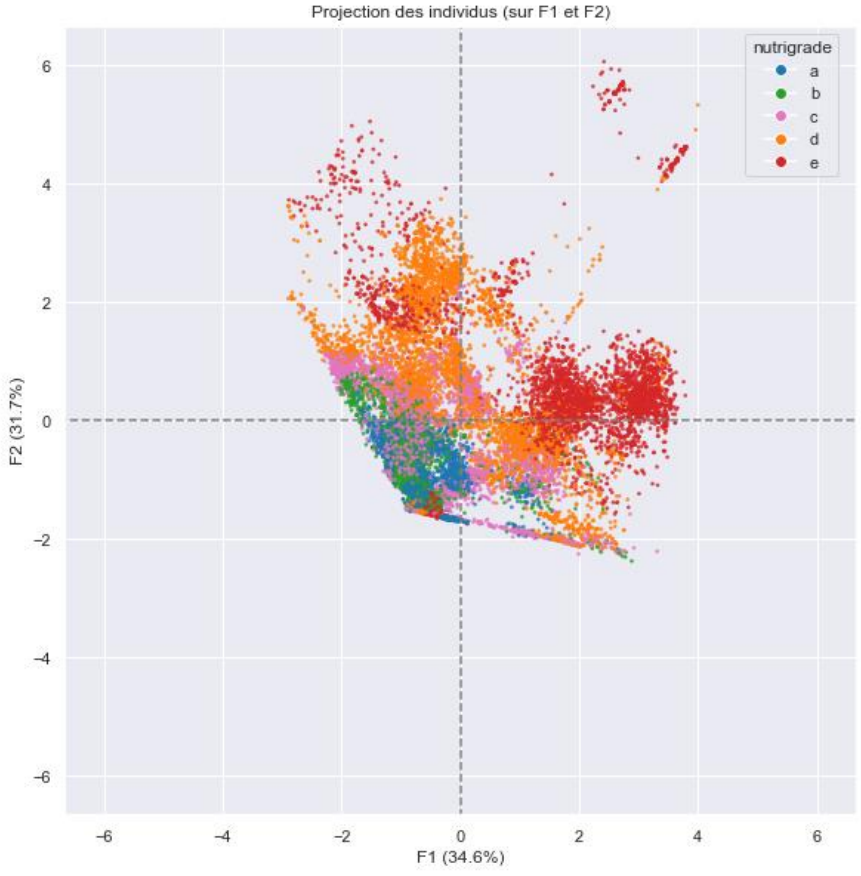
- Au final le **premier plan factoriel** représente un angle de vision **"mal-bouffe"**
- Alors que le **2éme** est plutôt **anti "nourriture saine"**

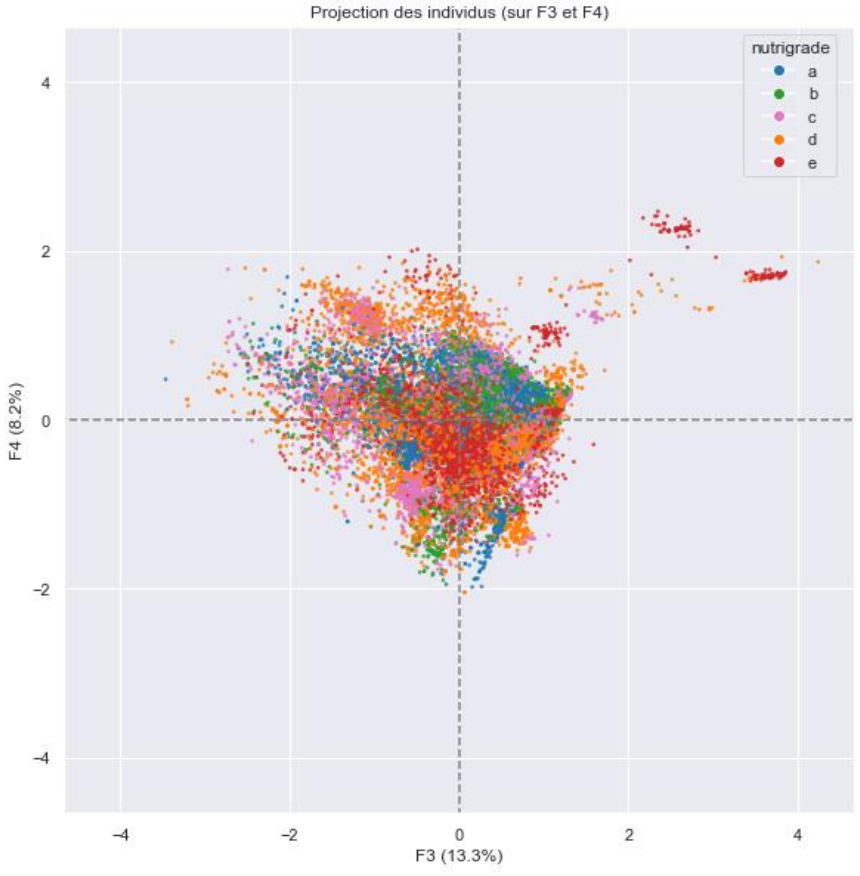




D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

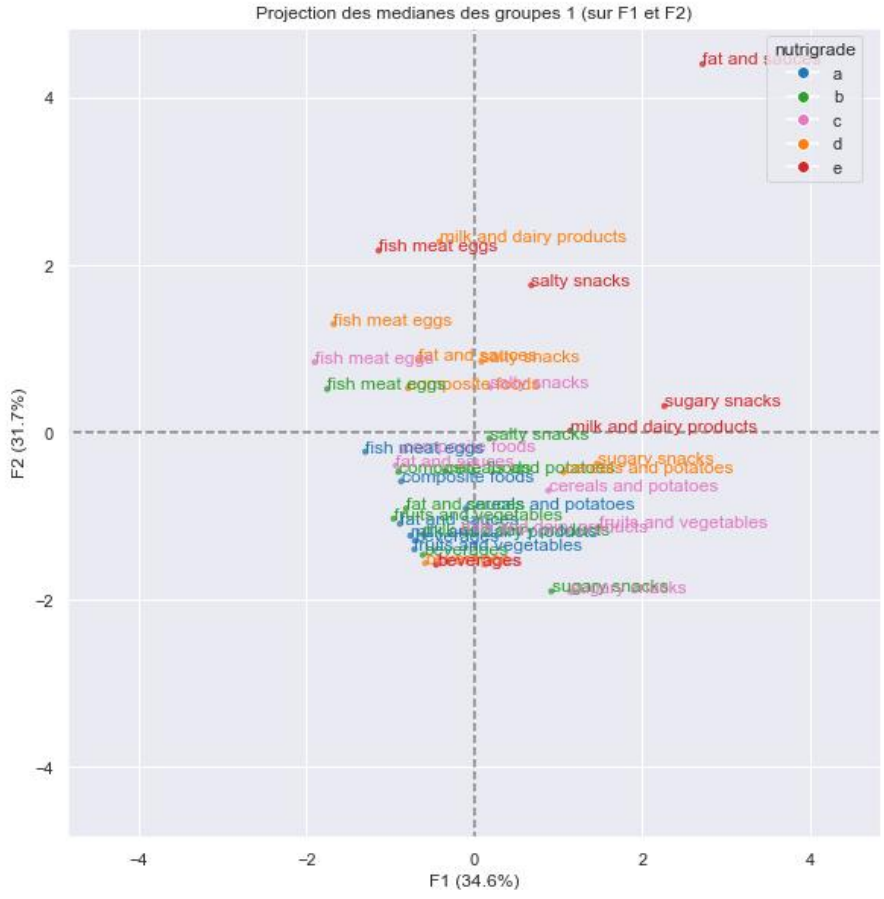
- Au final le **premier plan factoriel** représente un angle de vision **"mal-bouffe"**
- Alors que le **2éme** est plutôt **anti "nourriture saine"**

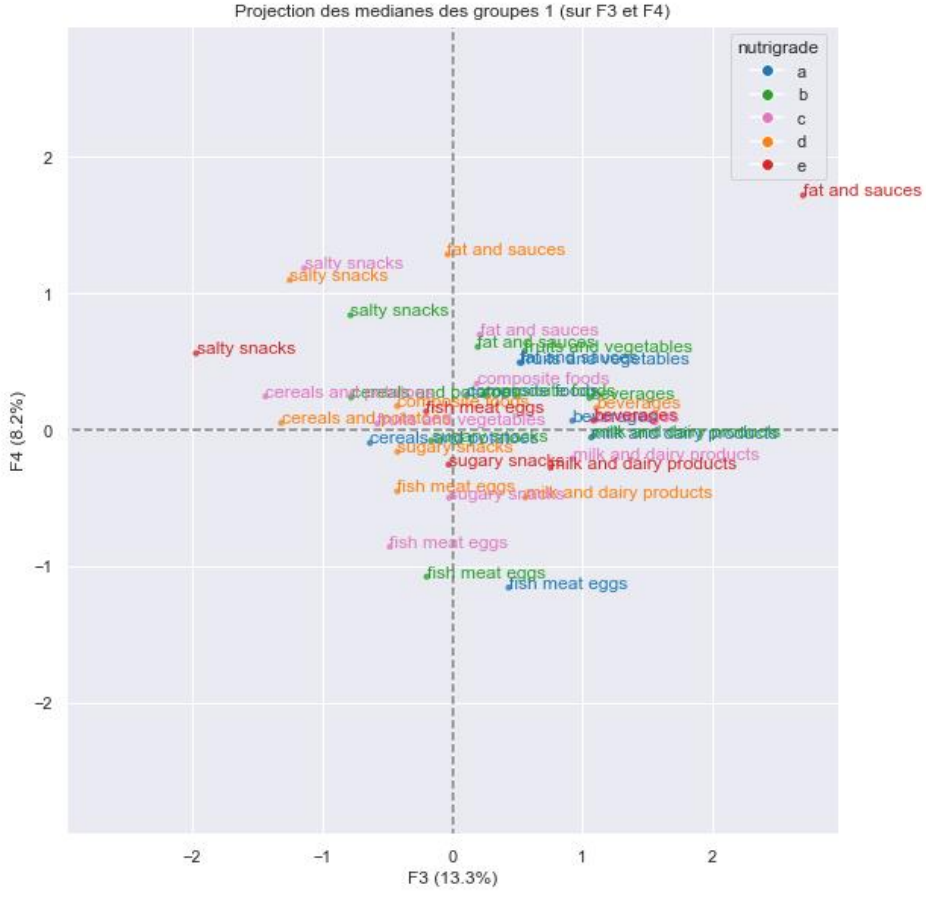




D: Analyse statistique
1-D:DESCRIPTIVE / EXPLICATIVE - Multivariée

- Au final le **premier plan factoriel** représente un angle de vision **"mal-bouffe"**
- Alors que le **2éme** est plutôt **anti "nourriture saine"**





D: Analyse statistique

1-E: Analyse explicative multivariée: Régression multiple

•Suite à notre analyse, nous pouvons tenter une explication multivariée du nutriscore sous forme de régression multiple.

- concernant les données quantitatives nutritionnelles, nous avons le choix d'utiliser:
  - **quelques facteurs de l'ACP: les 4 premiers par exemple**
  - **ou bien les variables elles mêmes en veillant à ne pas garder des variables trop colinéaires** (sodium ou sel par exemple mais pas les 2)
- concernant les données catégorielles , nous pouvons garder la présence d'additif et au choix le groupe 1 ou le groupe 2
  - le groupe 2 possédant certainement un peu trop de catégories quasi redondantes

•Nous obtenons avec L'ACP , les additifs ainsi que le groupe 1: un **R2 de 0.85** avec tous les **coefficients significatifs**:

- les facteurs 1 et 2, dits de "malbouffe" et la présence d'additif augmente le score modélisé
- le facteur 3 dit "nourriture saine" est pris en négatif ce qui fait du sens
- les résidus sont quasi normaux , homoscédastiques et la linéarité est presque respectée

•Nous obtenons avec les variables nutritionnelles (sans sodium ni fat) , les additifs ainsi que le groupe 1: un **R2 de 0.90** avec tous les **coefficients significatifs**:

- L'Energie, les acides gras, le sucre, le sel et la présence d'additif augmente le nutriscore et donc le grade
- Les hydrates de carbone , les fibres et les protéines diminue le nutriscore
- les résidus sont quasi normaux , homoscédastiques et la linéarité est presque respectée

D: Analyse statistique

1-E: Analyse explicative multivariée: Régression multiple

- Régression 1 avec les facteurs ACP, le groupe 1 et l'additif

OLS Regression Results

Dep. Variable:	score	R-squared:	0.852
Model:	OLS	Adj. R-squared:	0.852
Method:	Least Squares	F-statistic:	9093.
Date:	Thu, 30 Sep 2021	Prob (F-statistic):	0.00
Time:	19:08:35	Log-Likelihood:	358.40
No. Observations:	20613	AIC:	-688.8
Df Residuals:	20599	BIC:	-577.7
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4534	0.009	50.771	0.000	0.436	0.471
Fpca_1	0.1673	0.003	66.637	0.000	0.162	0.172
Fpca_2	0.2393	0.002	140.123	0.000	0.236	0.243
Fpca_3	-0.0968	0.003	-29.379	0.000	-0.103	-0.090
Fpca_4	0.0711	0.004	18.102	0.000	0.063	0.079
cereals and potatoes	-0.9081	0.011	-80.252	0.000	-0.930	-0.886
composite foods	-0.7021	0.009	-75.364	0.000	-0.720	-0.684
fat and sauces	-0.5407	0.011	-47.109	0.000	-0.563	-0.518
fish meat eggs	-0.4917	0.011	-44.213	0.000	-0.513	-0.470
fruits and vegetables	-0.8301	0.010	-81.047	0.000	-0.850	-0.810
milk and dairy products	-0.3625	0.010	-38.031	0.000	-0.381	-0.344
salty snacks	-0.6336	0.014	-44.148	0.000	-0.662	-0.605
sugary snacks	-0.0755	0.012	-6.289	0.000	-0.099	-0.052
additif	0.1823	0.004	45.920	0.000	0.175	0.190

Omnibus:	98.871	Durbin-Watson:	1.363
Prob(Omnibus):	0.000	Jarque-Bera (JB):	100.853
Skew:	0.162	Prob(JB):	1.26e-22
Kurtosis:	3.110	Cond. No.	22.9

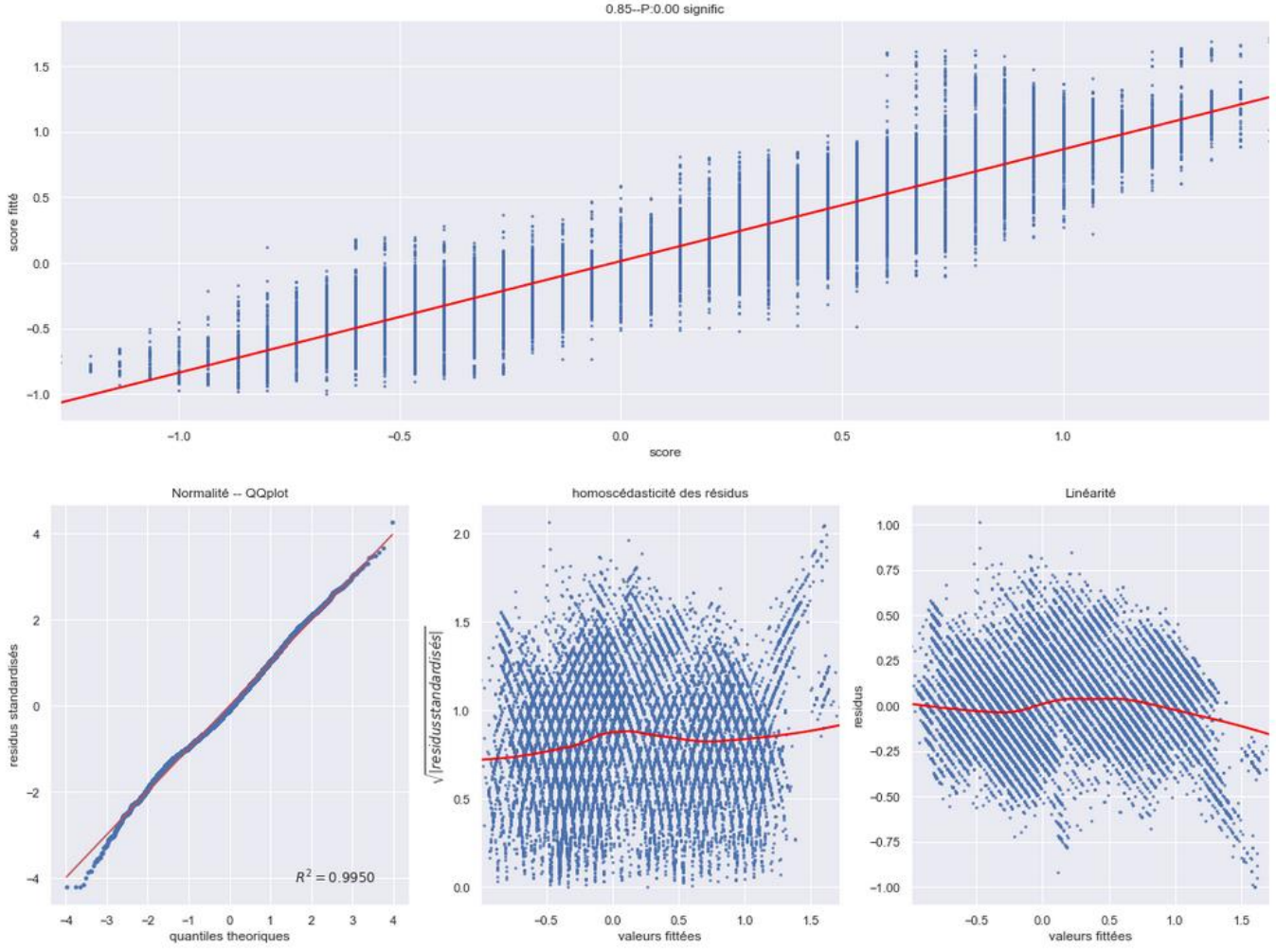
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

H0 independance des données, Durbin Watson: 1.36 , pas d'autocor si=2
H0 homogeneite des résidus,Pvalue Breusch Pagan: 0.00 , H0 rejeté si p<0.05
H0 homogeneite des résidus,Pvalue Goldfeld-Quandt: 0.00 , H0 rejeté si p<0.05
H0 normalité des résidus,Pvalue Jarque Bera: 0.00 , H0 rejeté si p<0.05

D: Analyse statistique

1-E: Analyse explicative multivariée: Régression multiple

- Régression 1 avec les facteurs ACP, le groupe 1 et l'additif



D: Analyse statistique

1-E: Analyse explicative multivariée: Régression multiple

- Régression 2 avec les variables nutritionnelles, le groupe 1 et l'additif

OLS Regression Results

Dep. Variable:	score	R-squared:	0.901
Model:	OLS	Adj. R-squared:	0.901
Method:	Least Squares	F-statistic:	1.174e+04
Date:	Thu, 30 Sep 2021	Prob (F-statistic):	0.00
Time:	19:12:12	Log-Likelihood:	4551.7
No. Observations:	20613	AIC:	-9069.
Df Residuals:	20596	BIC:	-8934.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3855	0.008	50.663	0.000	0.371	0.400
energy_100g	0.5242	0.008	64.583	0.000	0.508	0.540
saturated-fat_100g	0.1188	0.003	36.888	0.000	0.113	0.125
carbohydrates_100g	-0.1566	0.008	-18.817	0.000	-0.173	-0.140
sugars_100g	0.2205	0.003	66.977	0.000	0.214	0.227
fiber_100g	-0.0577	0.003	-19.588	0.000	-0.064	-0.052
proteins_100g	-0.0904	0.003	-30.965	0.000	-0.096	-0.085
salt_100g	0.3507	0.003	119.455	0.000	0.345	0.356
cereals and potatoes	-0.7481	0.012	-63.434	0.000	-0.771	-0.725
composite foods	-0.6270	0.008	-80.875	0.000	-0.642	-0.612
fat and sauces	-0.7015	0.010	-72.647	0.000	-0.720	-0.683
fish meat eggs	-0.4866	0.009	-53.236	0.000	-0.505	-0.469
fruits and vegetables	-0.7654	0.008	-91.103	0.000	-0.782	-0.749
milk and dairy products	-0.3730	0.008	-47.918	0.000	-0.388	-0.358
salty snacks	-0.6155	0.014	-43.944	0.000	-0.643	-0.588
sugary snacks	-0.2989	0.011	-27.854	0.000	-0.320	-0.278
additif	0.0787	0.003	23.086	0.000	0.072	0.085

Omnibus:	380.526	Durbin-Watson:	1.365
Prob(Omnibus):	0.000	Jarque-Bera (JB):	404.486
Skew:	0.327	Prob(JB):	1.47e-88
Kurtosis:	3.205	Cond. No.	29.7

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

H0 independance des données, Durbin Watson: 1.36 , pas d'autocor si=2
H0 homogeneite des résidus,Pvalue Breusch Pagan: 0.00 , H0 rejeté si p<0.05
H0 homogeneite des résidus,Pvalue Goldfeld-Quandt: 0.04 , H0 rejeté si p<0.05
H0 normalité des résidus,Pvalue Jarque Bera: 0.00 , H0 rejeté si p<0.05

D: Analyse statistique

1-E: Analyse explicative multivariée: Régression multiple

- Régression 2 avec les variables nutritionnelles, le groupe 1 et l'additif