Concevez une application au service de la santé publique

Santé publique France

Sommaire

I. Idée d’application

- Scan du produit et proposition de produits de même type/catégorie avec un meilleur nutri-score.
- Calculateur de nutri-score pour produits qui n’ont pas de nutri-score indiqué.

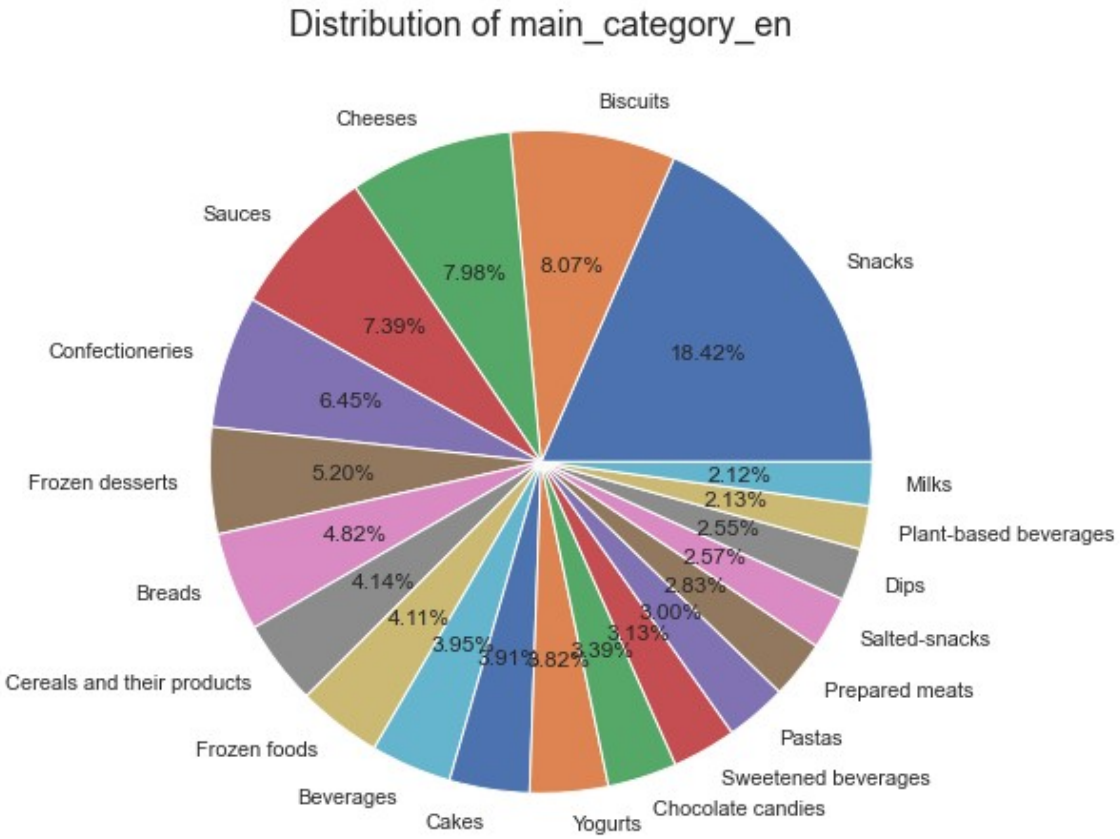
NUTRI-SCORE

A B C D E

NUTRI-SCORE

A B C D E

I. Idée d’application

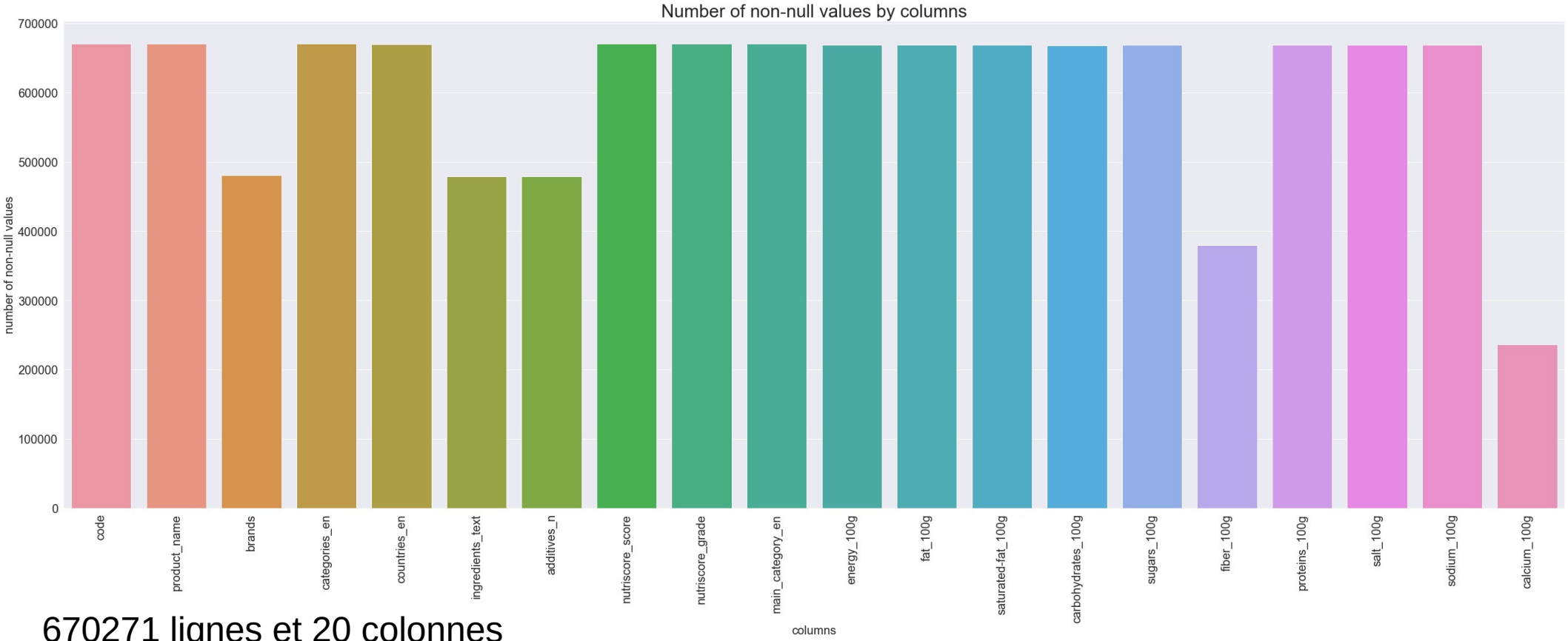


Les 20 catégories de produits les plus communs.

II. Nettoyage effectué

- Suppression des « code » dupliqués.
- Suppression des « code » vides.
- Suppression des « product_name » vides.
- Suppression des « nutriscore_score » vides.
- Suppression des colonnes avec un pourcentage de valeurs manquantes supérieurs à 65 %.
- Suppression des colonnes inutiles.

II. Nettoyage effectué



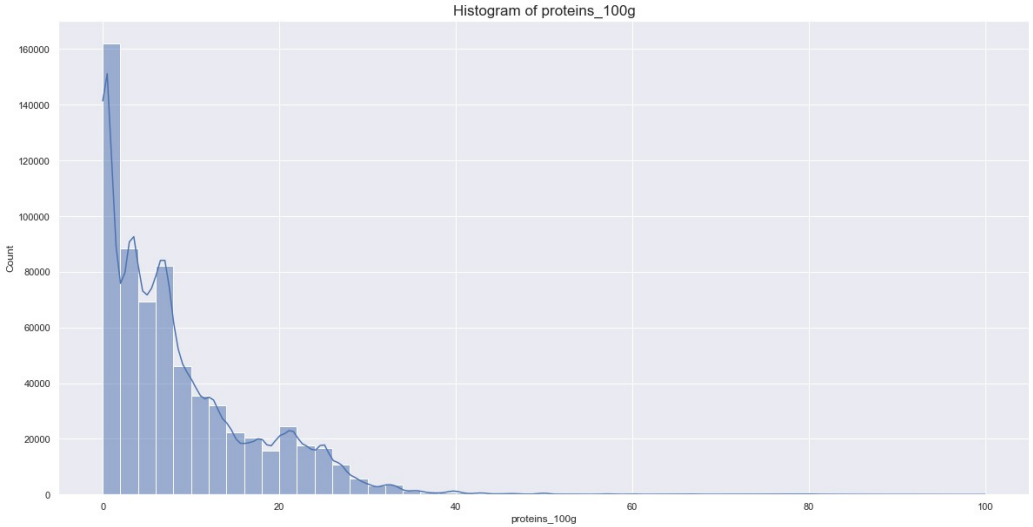
670271 lignes et 20 colonnes

II. Nettoyage effectué

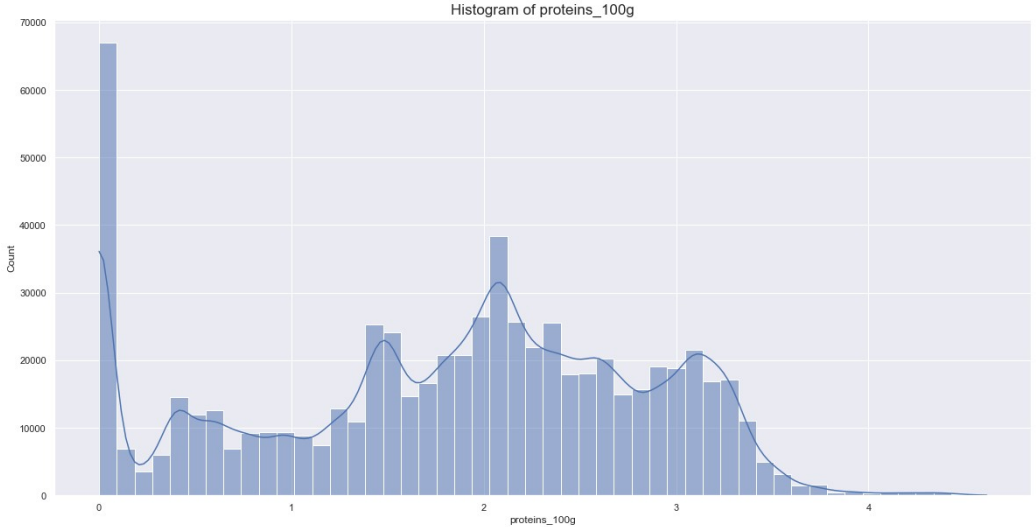
- Suppression des produits avec ‘fat_100g’, ‘saturated-fat_100g’, ‘carbohydrates_100g’, ‘sugars_100g’, ‘fiber_100g’, ‘proteins_100g’, ‘salt_100g’, ‘sodium_100g’ et ‘calcium_100g’ supérieurs à 100g ou inférieurs à 0g.
- Suppression des produits avec ‘energy_100g’ supérieur à 3700KJ ou inférieur à 0KJ.
- En effet, la valeur énergétique de la graisse est la plus élevée :
  - Fat : 37KJ/g
  - Proteins : 29 KJ/g
  - Carbohydrates : 17KJ/g
  - Fiber : KJ/g etc...

II. Nettoyage effectué

Avant log



Après log



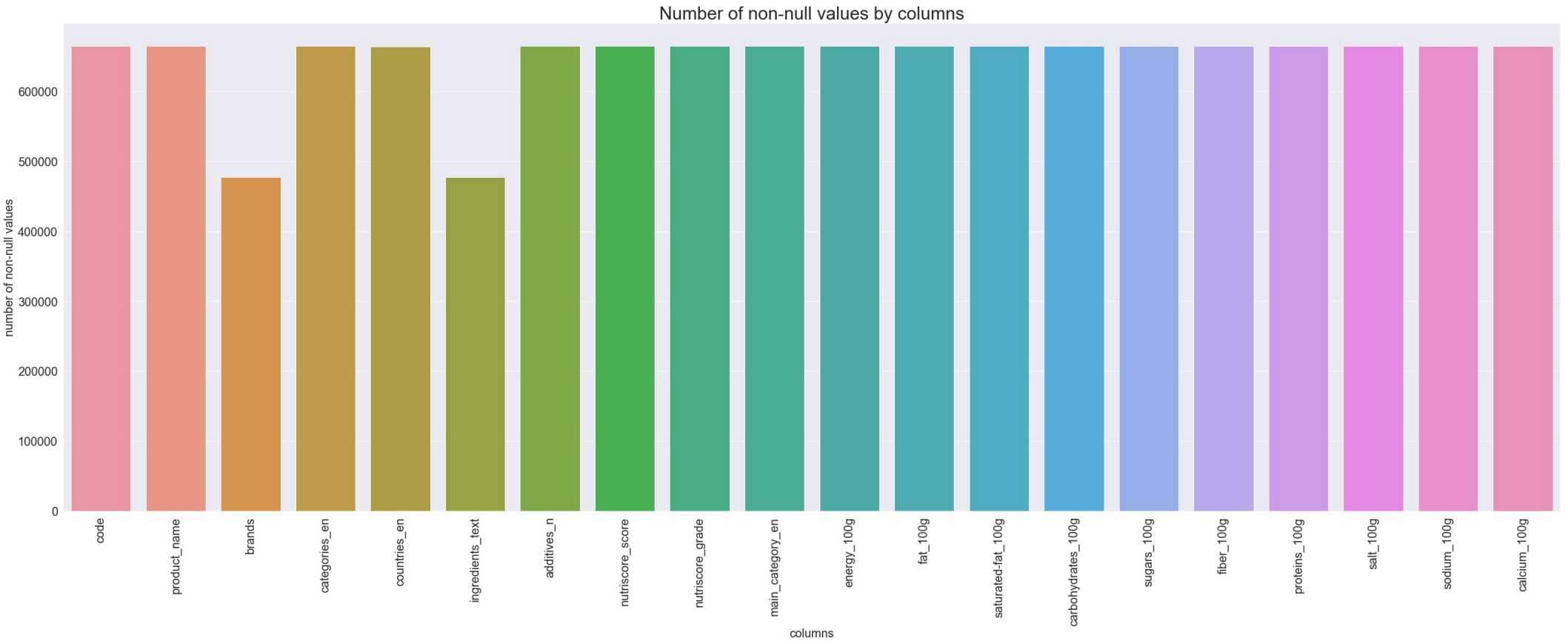
II. Nettoyage effectué

- Application du logarithme népérien sur les colonnes 'additives_n', 'energy_100g', 'fat_100g', 'saturated-fat_100g', 'carbohydrates_100g', 'sugars_100g', 'fiber_100g', 'proteins_100g', 'salt_100g', 'sodium_100g', 'calcium_100g' puis application de la fonction « aberrantvalues » et « aberrantvalues_2 » sur ces colonnes.
- Application du log car la distribution de la variable cible est asymétrique à droite
- Retour aux valeurs initiales avec l’application de la fonction exponentielle.

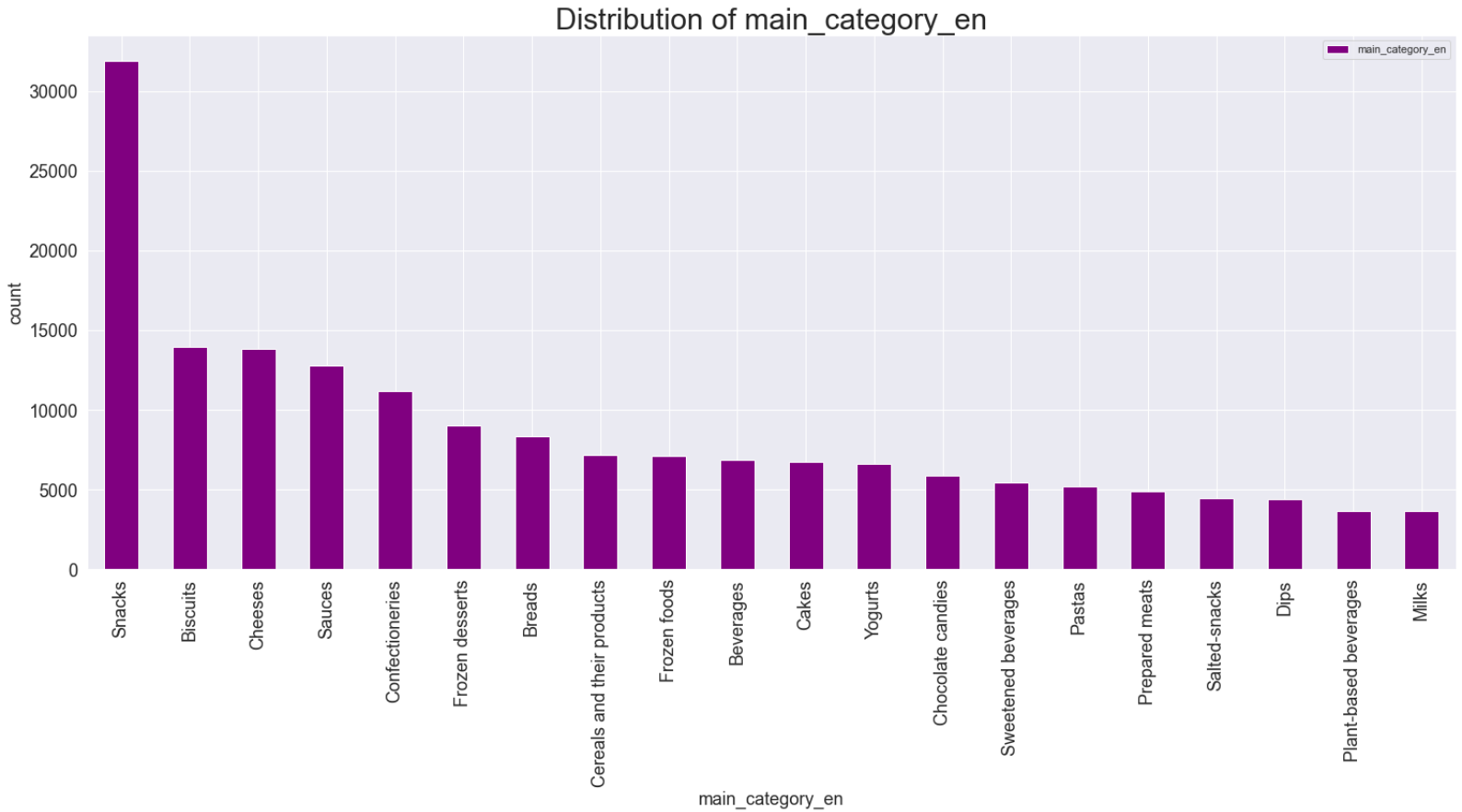
II. Nettoyage effectué

- Détection des valeurs aberrantes par la méthode interquartile et remplacement des valeurs aberrantes par NaN.
- Méthode interquartile :
  - Soit Q1 et Q3 respectivement le 1er quartile et le 3^{e} quartile et k constante positive. On peut définir une donnée aberrante comme étant toute valeur située à l'extérieur de l'intervalle :
    - [Q1 – k(Q3 – Q1), Q3 + k(Q3 – Q1)]
- Imputation des valeurs manquantes dans les colonnes de type float par la moyenne de leur groupe de catégorie.
- Imputation par 0 les valeurs non imputées par la méthode précédente.

II. Nettoyage effectué

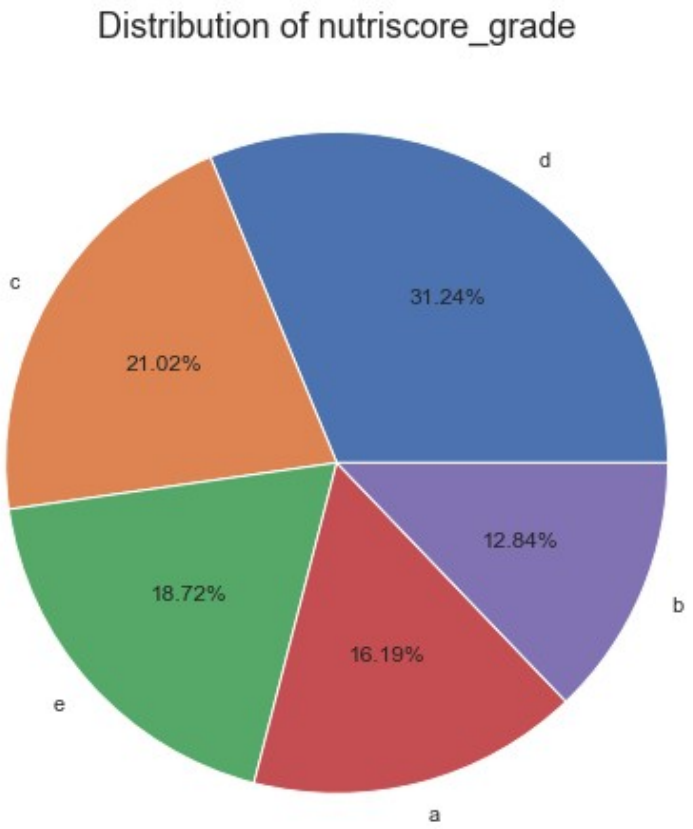


III. Analyse Exploratoire

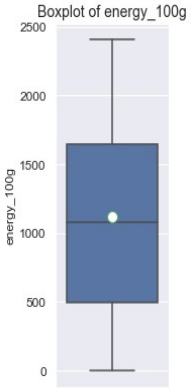


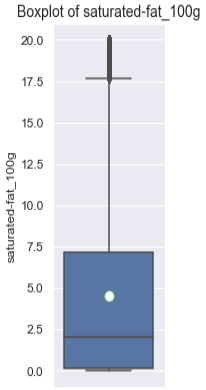
Les 20 catégories de produits les plus communs.

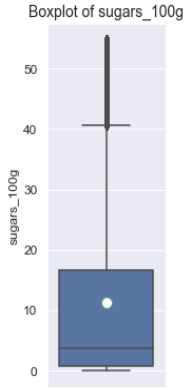
III. Analyse Exploratoire

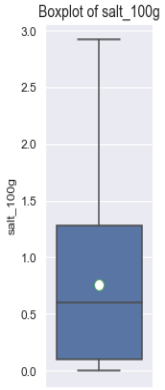


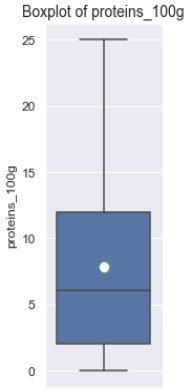
III. Analyse Exploratoire

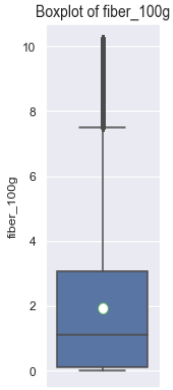




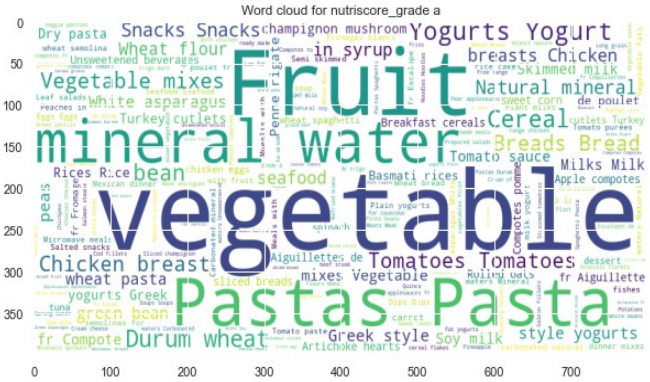


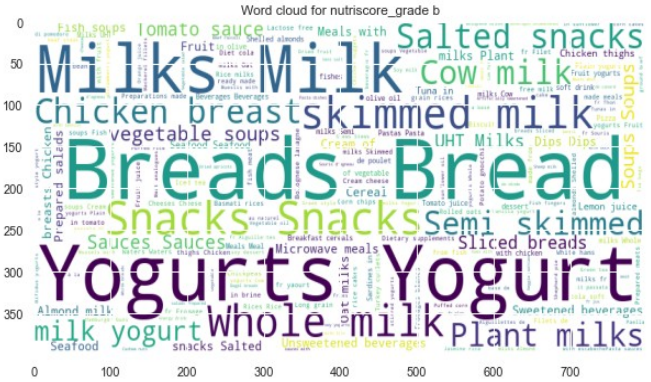


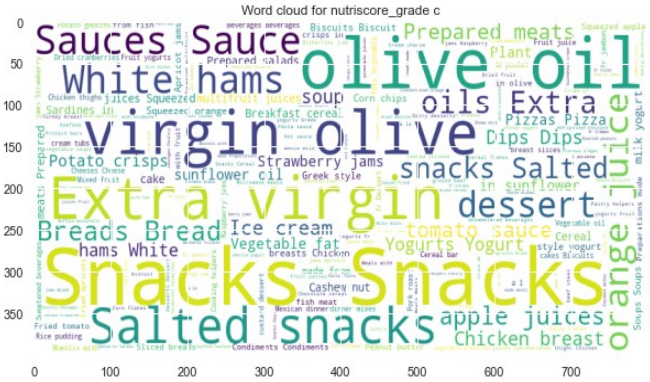


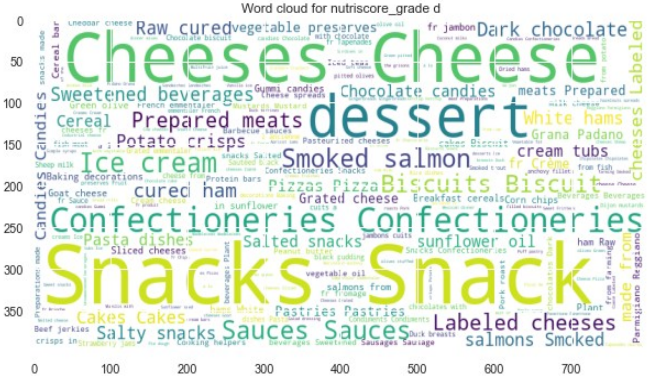


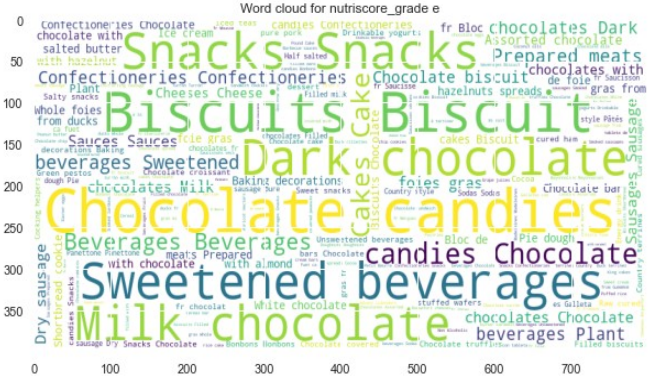
III. Analyse Exploratoire



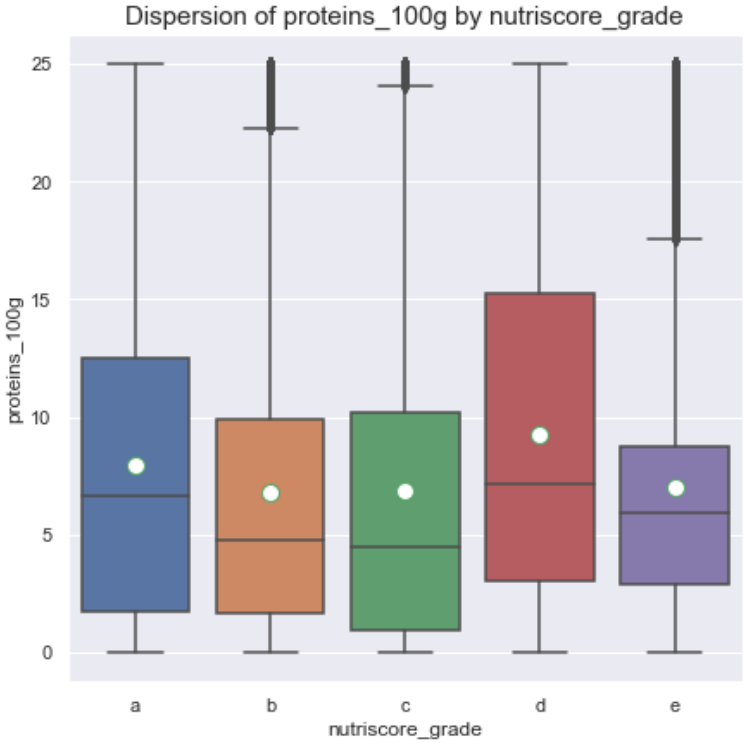


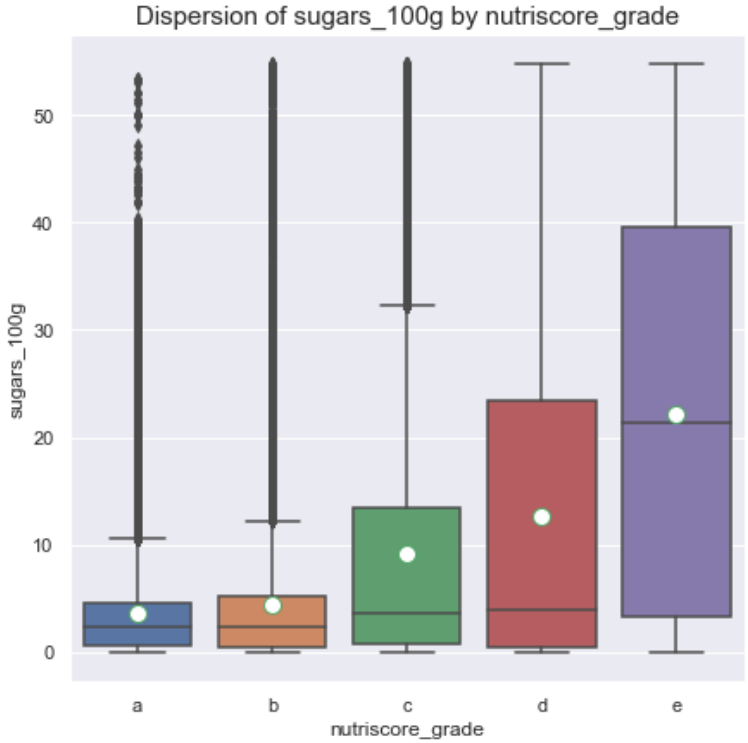






III. Analyse Exploratoire



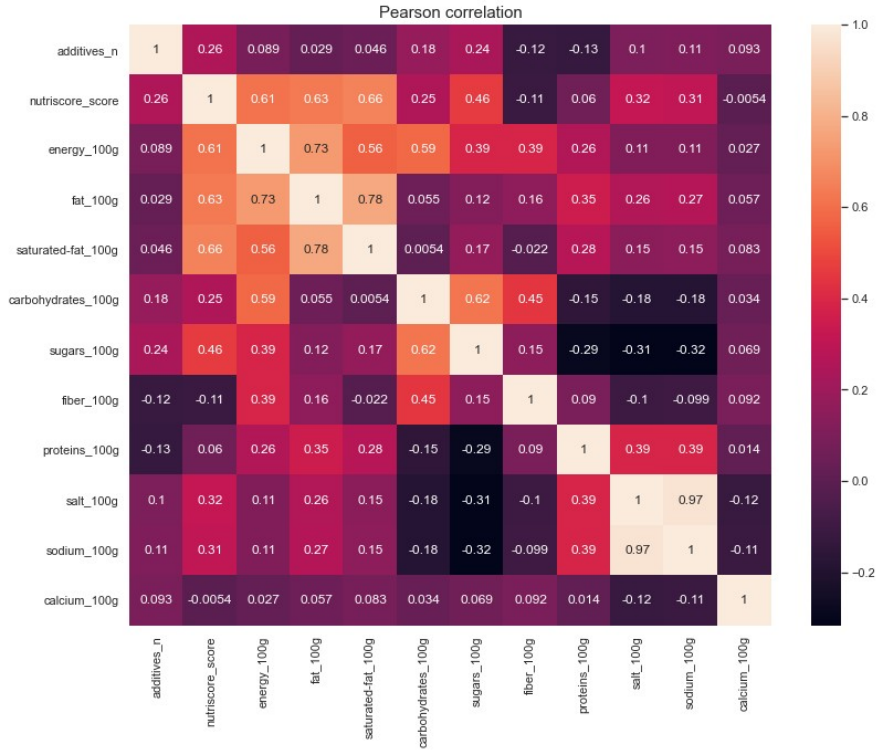


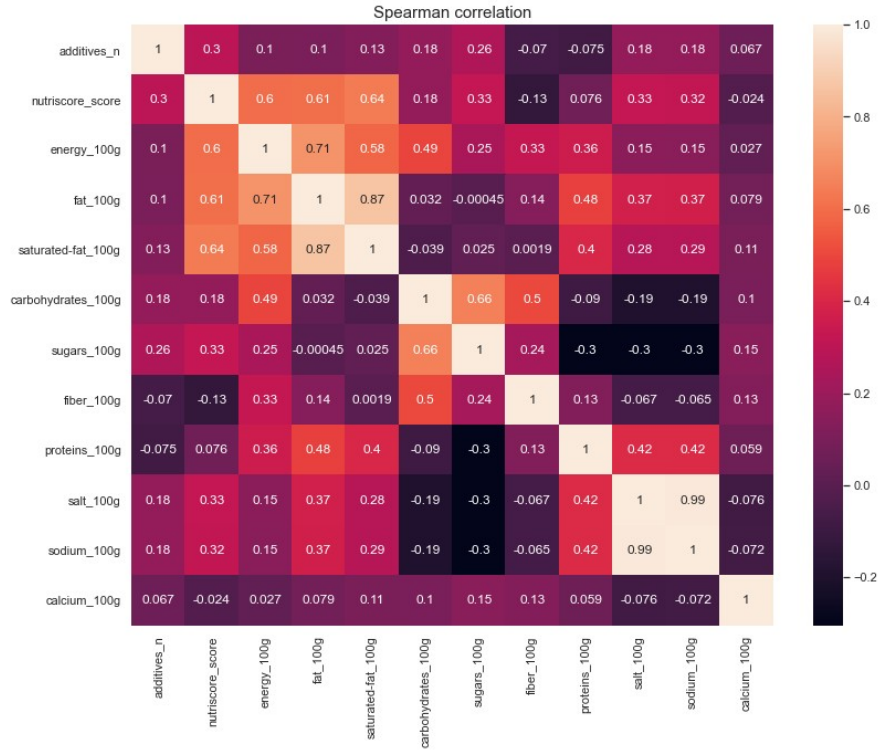
III. Analyse Exploratoire

Par rapport à la nutriscore

```
Kruskal-Wallis test :
additives_n------------------------------------- "H0 rejected"
Statistics =  63833.46564435494 p =  0.0
nutriscore_score-------------------------------- "H0 rejected"
Statistics =  611995.6152171498 p =  0.0
energy_100g------------------------------------- "H0 rejected"
Statistics =  183435.82900968104 p =  0.0
fat_100g---------------------------------------- "H0 rejected"
Statistics =  186387.3206482209 p =  0.0
saturated-fat_100g------------------------------ "H0 rejected"
Statistics =  213796.23453821873 p =  0.0
carbohydrates_100g------------------------------ "H0 rejected"
Statistics =  21601.75171535672 p =  0.0
sugars_100g------------------------------------- "H0 rejected"
Statistics =  75307.09429935899 p =  0.0
fiber_100g-------------------------------------- "H0 rejected"
Statistics =  27027.85955150427 p =  0.0
proteins_100g----------------------------------- "H0 rejected"
Statistics =  12786.49985013412 p =  0.0
salt_100g--------------------------------------- "H0 rejected"
Statistics =  80838.00890757923 p =  0.0
sodium_100g------------------------------------- "H0 rejected"
Statistics =  78687.06840527106 p =  0.0
calcium_100g------------------------------------ "H0 rejected"
Statistics =  3228.4398951513003 p =  0.0
```

III. Analyse Exploratoire



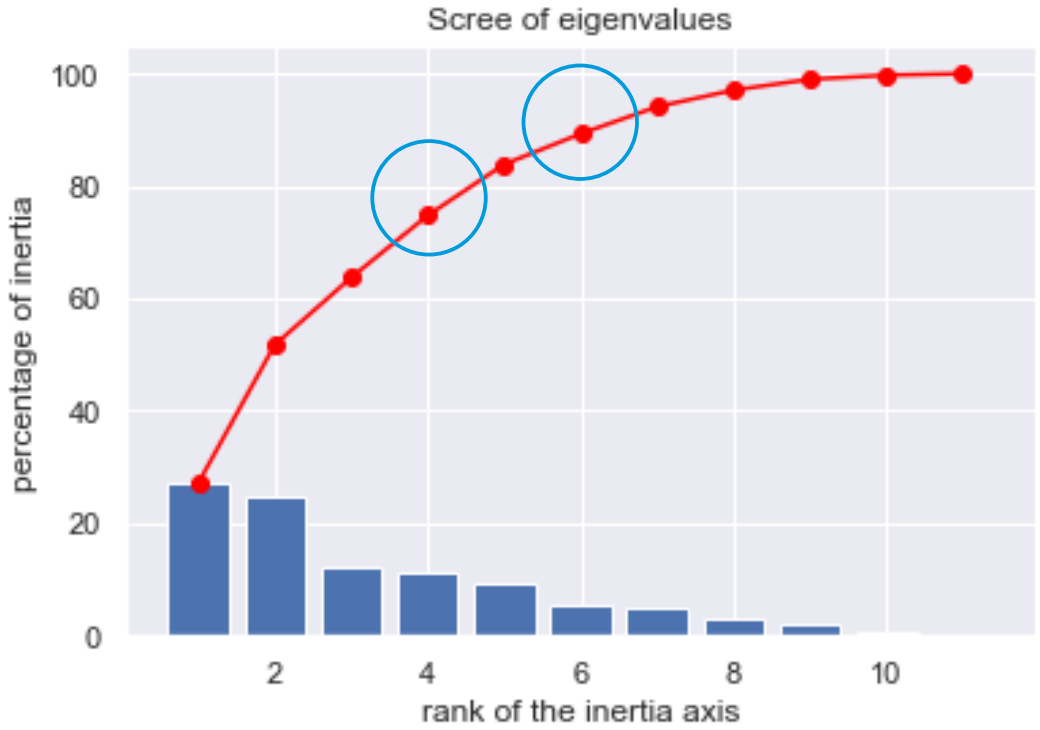


III. Analyse Exploratoire

- L'Analyse en Composantes Principales permet de dégager rapidement les principales tendances de votre échantillon, en diminuant le nombre de variables nécessaires à la représentation de vos données tout en perdant le moins d'informations possible

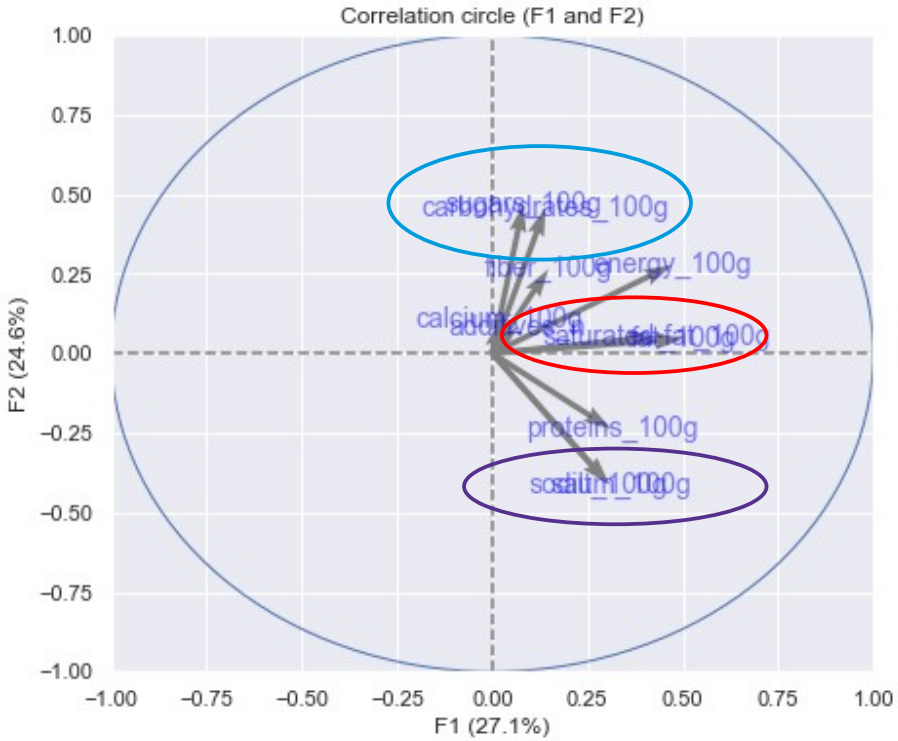
III. Analyse Exploratoire

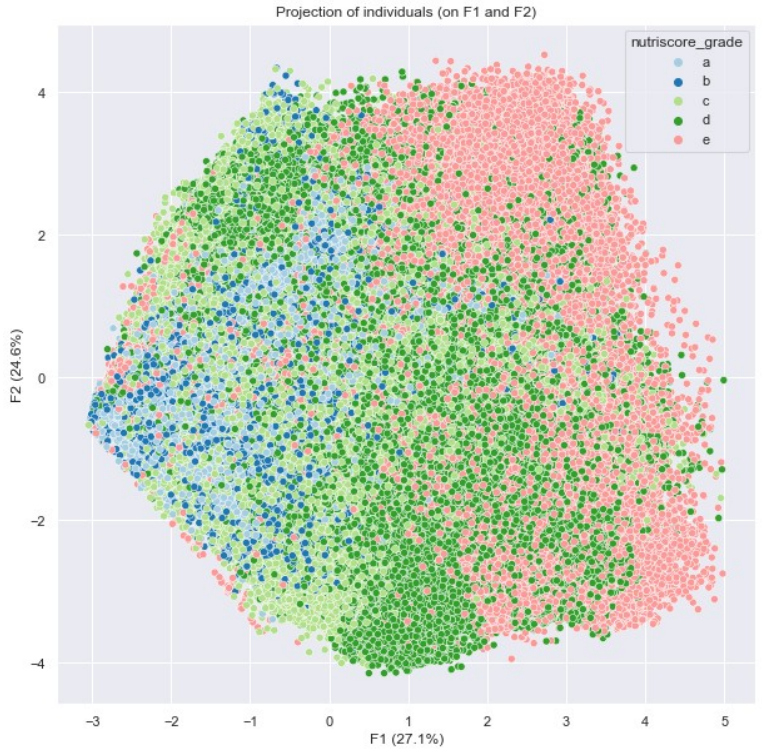
Avec 11 composants



III. Analyse Exploratoire

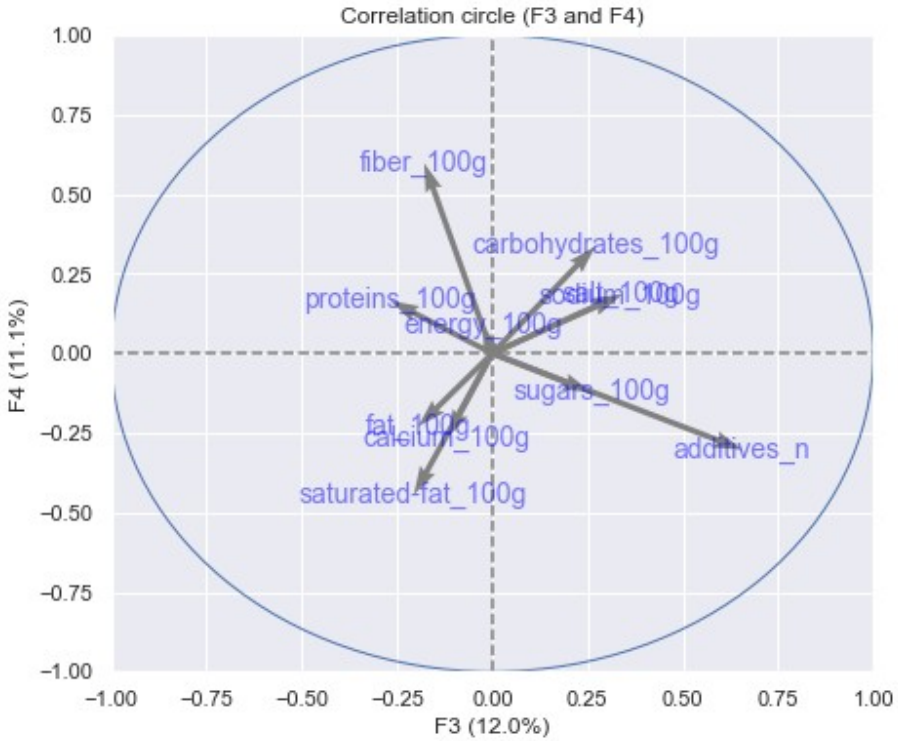
Avec 11 composants

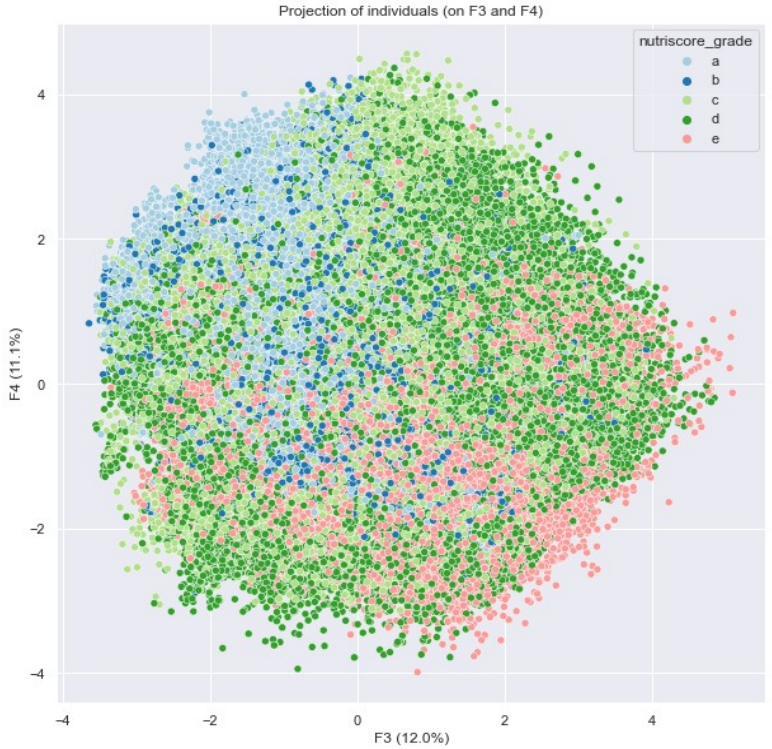




III. Analyse Exploratoire

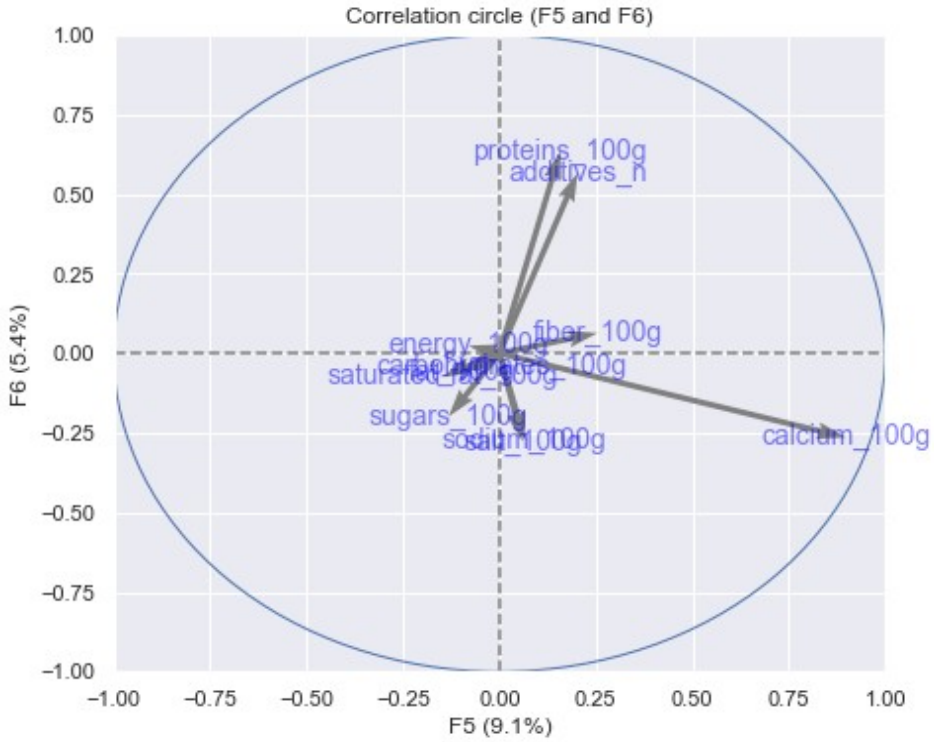
Avec 11 composants

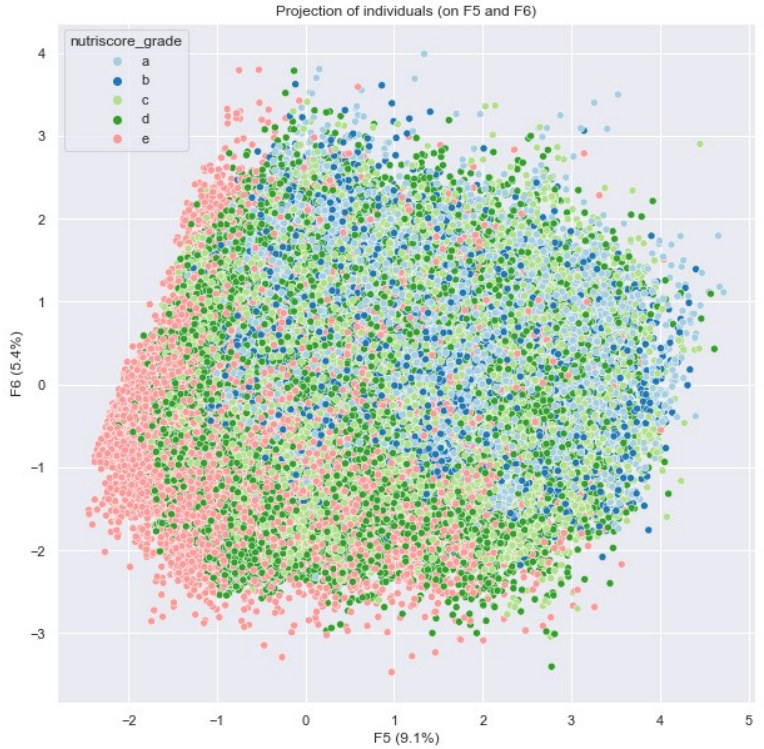




III. Analyse Exploratoire

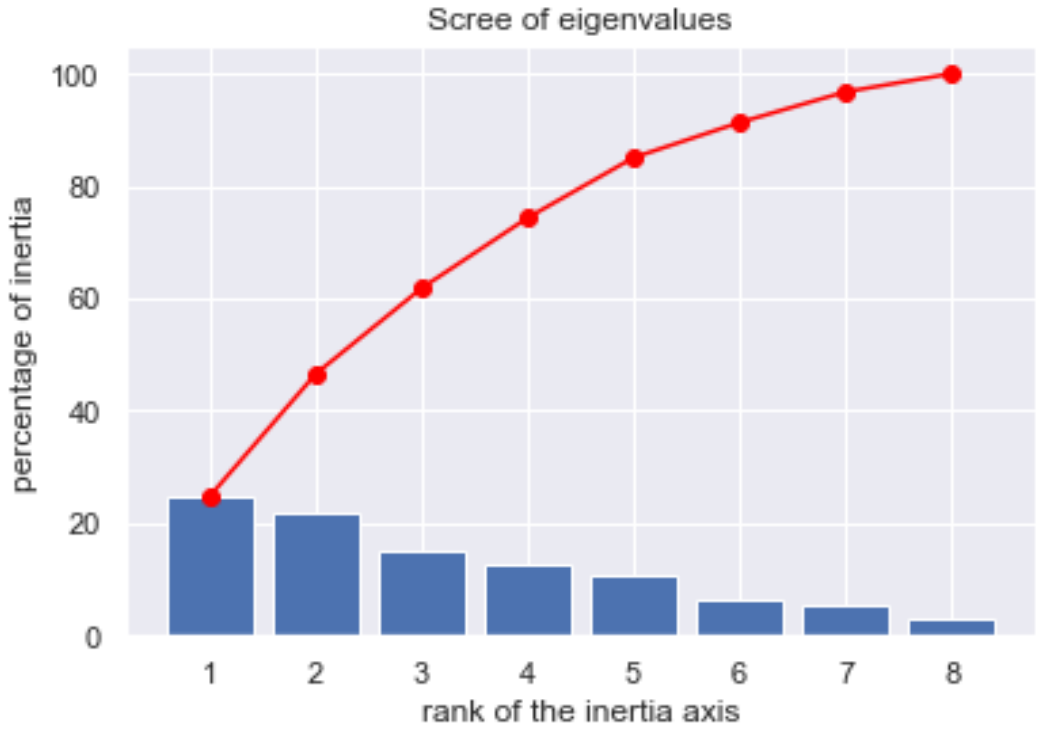
Avec 11 composants





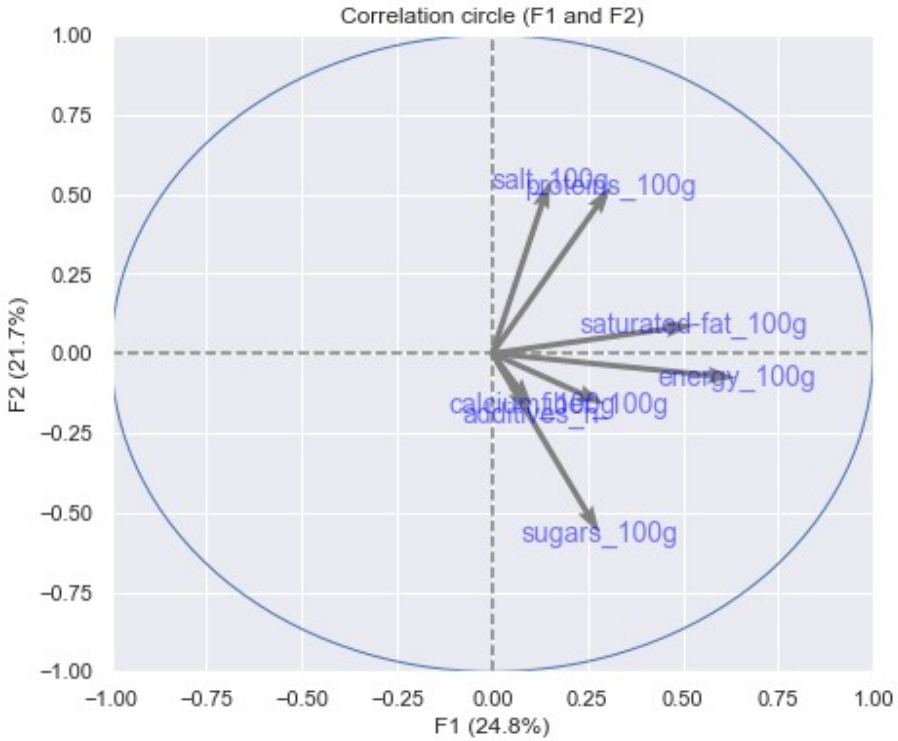
III. Analyse Exploratoire

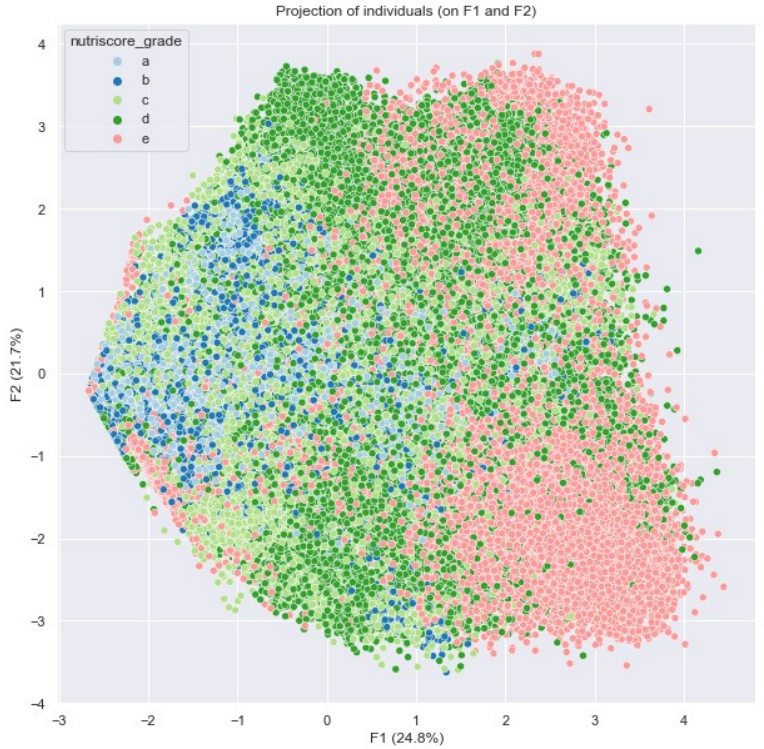
Avec 8 composants



III. Analyse Exploratoire

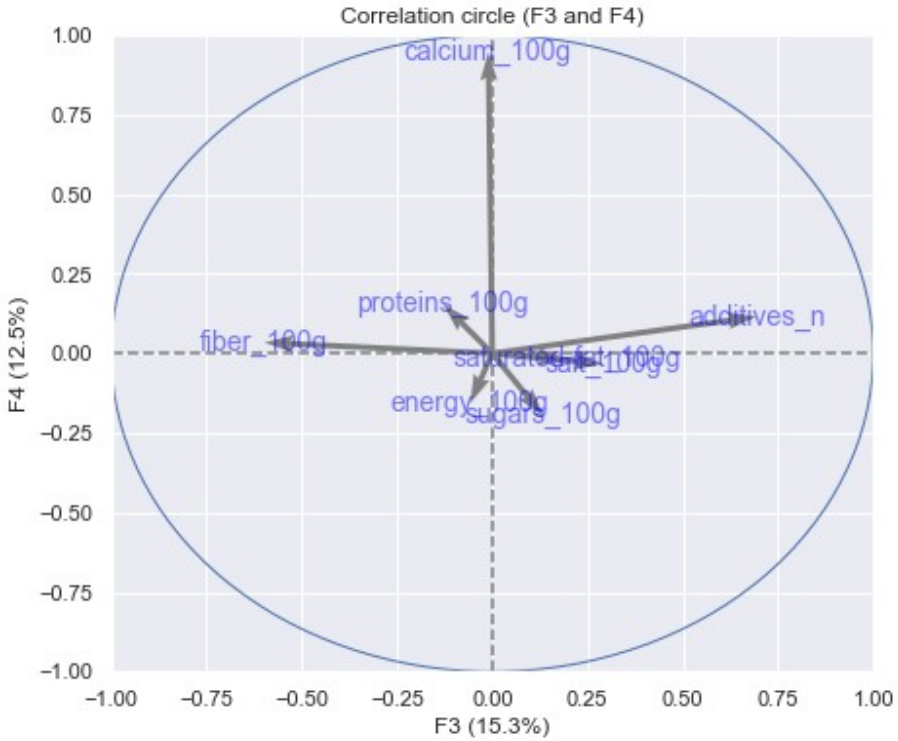
Avec 8 composants

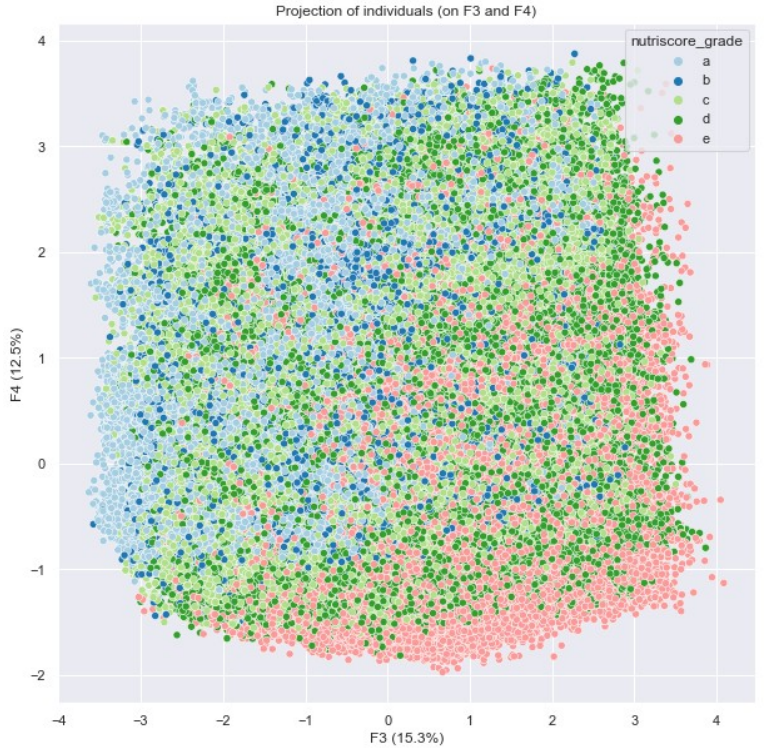




III. Analyse Exploratoire

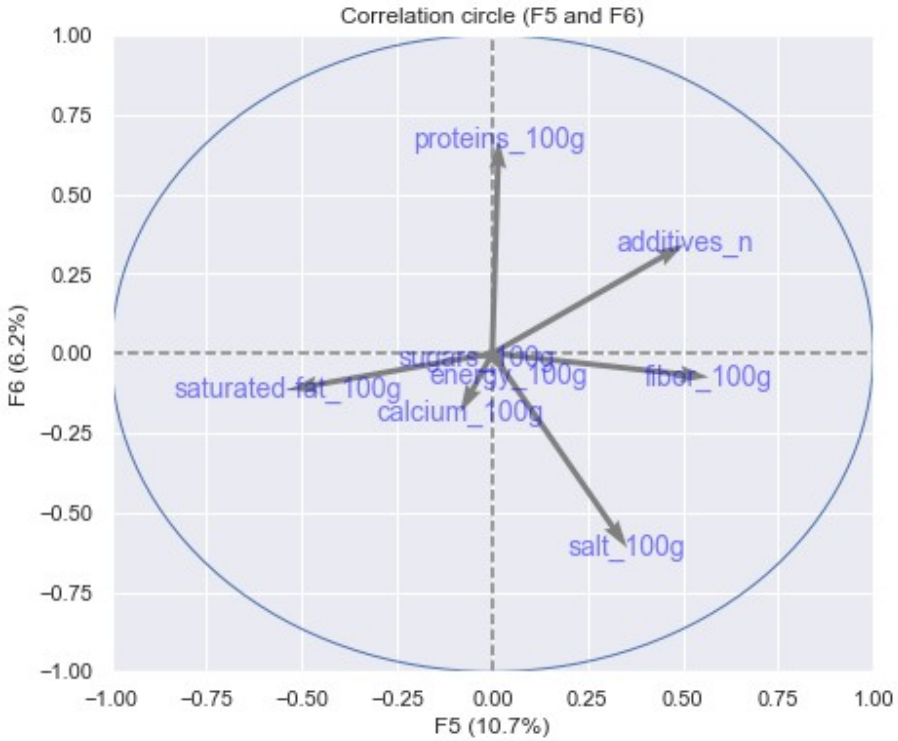
Avec 8 composants

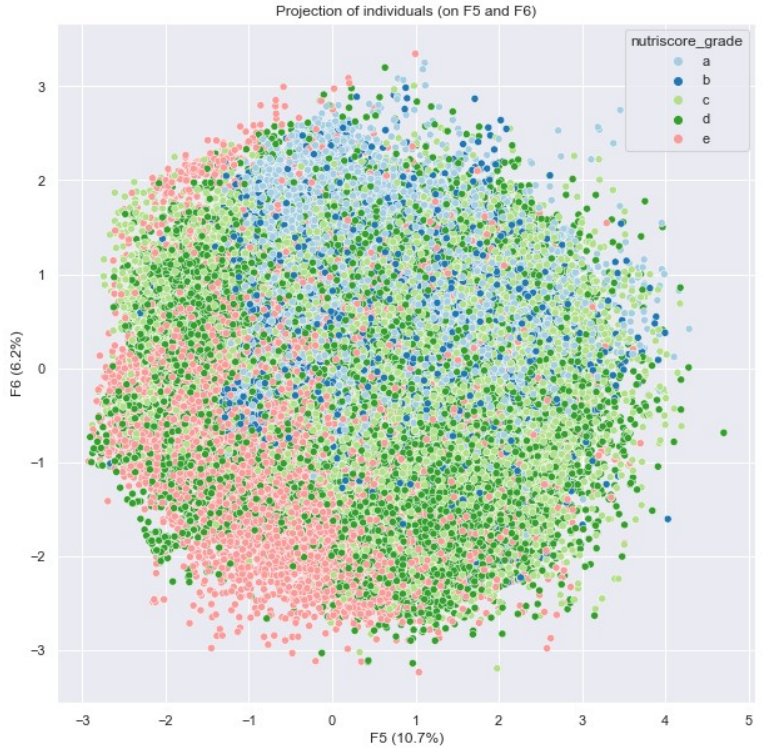




III. Analyse Exploratoire

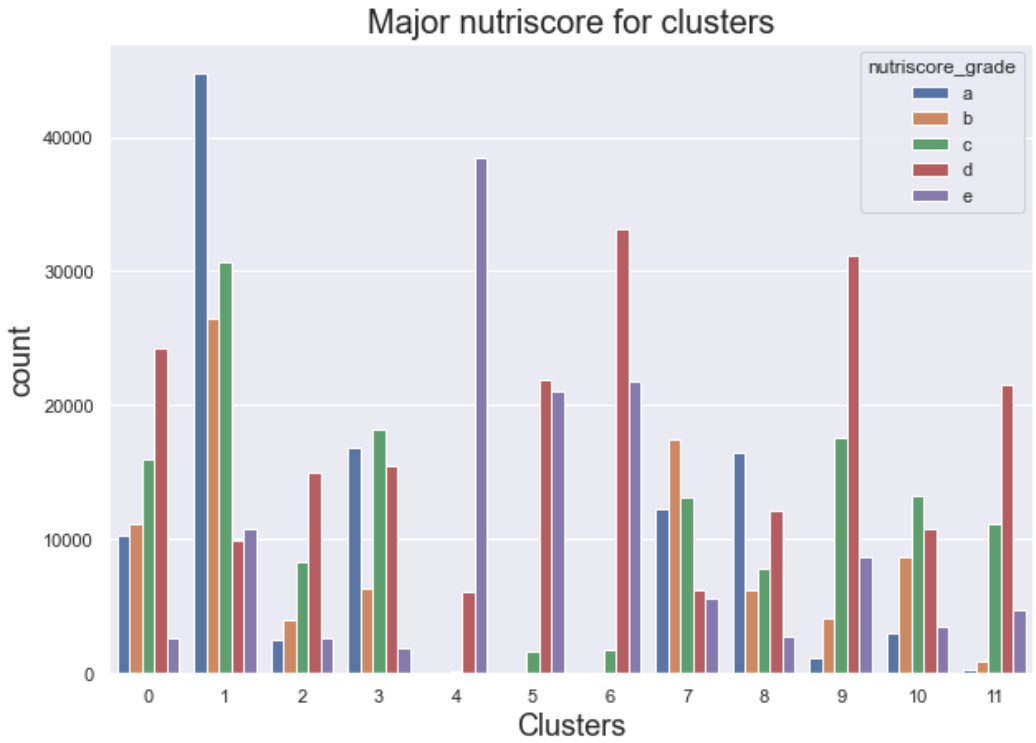
Avec 8 composants





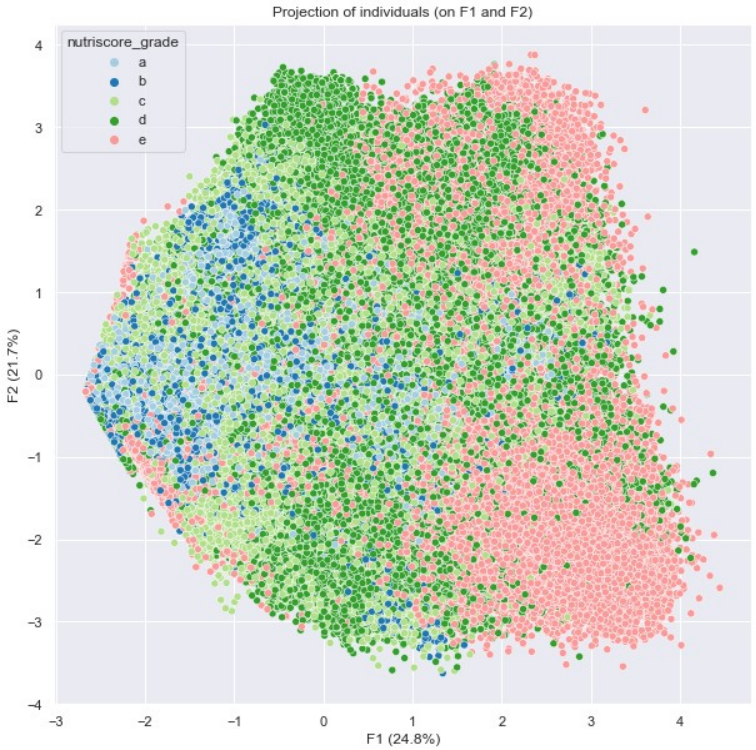
IV. Faits Pertinents

K-means



IV. Faits Pertinents

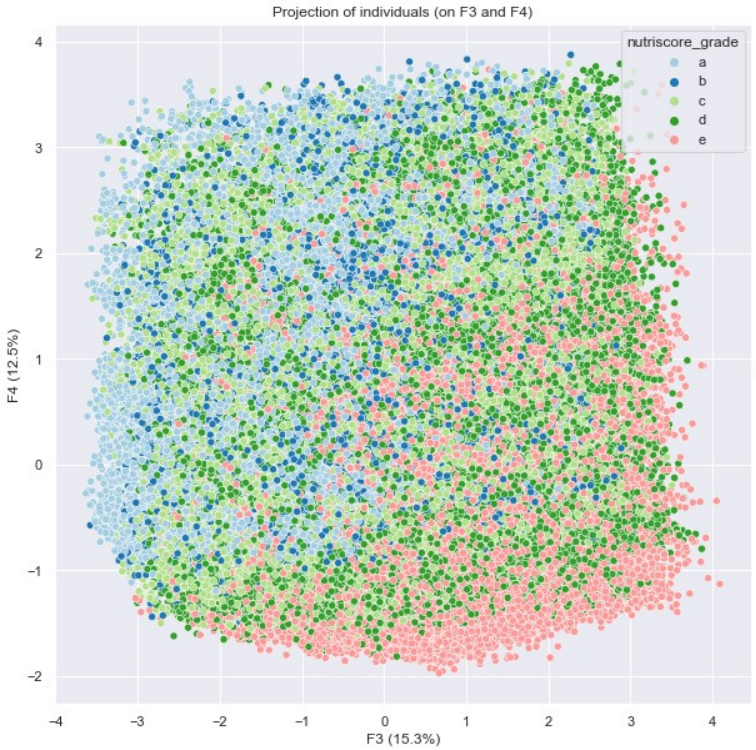
K-means



- Cluster 0 : hams, meats, chicken, salmons, seafood → D, C, B, A
- Cluster 1 : beverages, sauces, vegetables, olive oils → A, C, B
- Cluster 2 : desserts, breads, pizzas → D, C
- Cluster 3 : snacks, cereals → C, A, D
- Cluster 4 : chocolates, biscuits, snacks → E
- Cluster 5 : biscuits, cakes, confectioneries → D, E
- Cluster 6 : cheeses → D, E
- Cluster 7 : yogurts, beverages, milks → B, C, A
- Cluster 8 : snacks, pastas, rices → A, D
- Cluster 9 : sauces, snacks, breads → D, C
- Cluster 10 : sauces, desserts, beverages, breads, snacks → C, D, B
- Cluster 11 : confectioneries, snacks, candies, jams → D, C

IV. Faits Pertinents

K-means



- Cluster 0 : hams, meats, chicken, salmons, seafood → D, C, B, A
- Cluster 1 : beverages, sauces, vegetables, olive oils → A, C, B
- Cluster 2 : desserts, breads, pizzas → D, C
- Cluster 3 : snacks, cereals → C, A, D
- Cluster 4 : chocolates, biscuits, snacks → E
- Cluster 5 : biscuits, cakes, confectioneries → D, E
- Cluster 6 : cheeses → D, E
- Cluster 7 : yogurts, beverages, milks → B, C, A
- Cluster 8 : snacks, pastas, rices → A, D
- Cluster 9 : sauces, snacks, breads → D, C
- Cluster 10 : sauces, desserts, beverages, breads, snacks → C, D, B
- Cluster 11 : confectioneries, snacks, candies, jams → D, C

IV. Faits Pertinents

K-means

- Cluster 0 : hams, meats, chicken, salmons, seafood → D, C, B, A
- Cluster 1 : beverages, sauces, vegetables, olive oils → A, C, B
- Cluster 2 : desserts, breads, pizzas → D, C
- Cluster 3 : snacks, cereals → C, A, D
- Cluster 4 : chocolates, biscuits, snacks → E
- Cluster 5 : biscuits, cakes, confectioneries → D, E
- Cluster 6 : cheeses → D, E
- Cluster 7 : yogurts, beverages, milks → B, C, A
- Cluster 8 : snacks, pastas, rices → A, D
- Cluster 9 : sauces, snacks, breads → D, C
- Cluster 10 : sauces, desserts, beverages, breads, snacks → C, D, B
- Cluster 11 : confectioneries, snacks, candies, jams → D, C

IV. Faits Pertinents

	code	product_name	brands	categories_en	countries_en	ingredients_text	additives_n	nutriscore_score	nutriscore_grade
371395	3274616130353	Shiitakes	Saveur ASIATIQUE	Plant-based foods and beverages,Plant-based fo...	France	Shiitoké (lentinus edodes) Produit sujet à des...	0.0	-14.0	a
179913	05012343	Asperge Blanches des sables des Landes	Priméale	Plant-based foods and beverages,Plant-based fo...	France	Asperges blanches fraîches des sables des Landes.	0.0	-15.0	a
45580	0036632008817	Blended greek yogurt	NaN	Dairies,Fermented foods,Fermented milk product...	United States	Cultured grade a non fat milk, chicory root fi...	4.0	-9.0	a
362287	3263859883713	Fonds artichauts	Leader Price	Plant-based foods and beverages,Plant-based fo...	France	Fonds d'artichauts.	0.0	-15.0	a
209148	0734020610160	Crunchies	NaN	Snacks	United States	Strawberries	0.0	-1.0	a
243142	0854995003481	Flavored bar	NaN	Snacks	United States	WHOLE GRAIN BARLEY FLAKES[caret] (22%), INULIN...	4.0	-3.0	a
235167	0846548063837	Walnuts	Cibo Vita Inc	Plant-based foods and beverages,Plant-based fo...	United States	Natural shelled walnuts.	0.0	0.0	b
4240	0011110856616	Private selection, edamame, unshelled tender s...	Private Selection	Plant-based foods and beverages,Plant-based fo...	United States	Edamame (soybean in pod)	0.0	-14.0	a
25806	0021140260932	Winn-dixie, crowder peas	Winn-Dixie, Winn-Dixie Stores Inc.	Plant-based foods and beverages,Plant-based fo...	United States	Crowder peas, water.	0.0	-14.0	a

26		Fèves à la purée de sésame	California Garden	Plant-based foods and beverages,Plant-based fo...	France	fèves (65%), eau (28,06%), pâte de sésame (15%...	2.0	-9.0	a
30		Signature edamame with asian seasoning	Pictsweet	Plant-based foods and beverages,Plant-based fo...	United States	Edamame (soybeans), maltodextrin, salt, soy sa...	4.0	-12.0	a
188947	0665072630151	100% natural fruit pulp for juice & smoothie	NaN	Plant-based foods and beverages,Plant-based fo...	United States	Tamarind	0.0	-6.0	a

IV. Faits Pertinents

- Application sur un produit sélectionné de manière random :

	product_name	brands	main_category_en	nutriscore_grade	Clusters
261405	Isabar	Isagenix	Cereal bars	c	3

- nutriscore_grade prédit par Knn algorithme :

```
array(['d'], dtype=object)
```

V. Synthèse

- Les variables sont influentes sur le nutriscore.
- Il est possible de classer les produits par ses composants.
- Il faudra faire une classification à partir des données textuelles et visuelles pour de meilleur résultat.
- Les premiers classements avec les composants semblent cohérents.