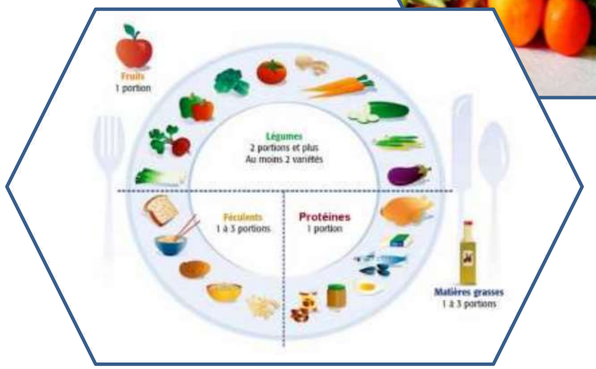
Concevez une application au service de la santé publique

Projet 3: OpenClassrooms
Présenté par : Lynda HADJEMI
janvier 2023.

Sommaire



Présentation de projet:

L’agence ‘’ **Santé publique France**’’ a lancé un appel à projets pour trouver des idées innovantes d’applications en lien avec l'alimentation.

Pour développer cette application les données de l’**OpenFoodFacts** seront décrites, analysées et exploitées afin de trouver une meilleurs approche.

L’idée de l’application:

Etape 1 :

Scanner le code barre

Etape 2 :

Les informations nutritionnelles.

Etape3 :

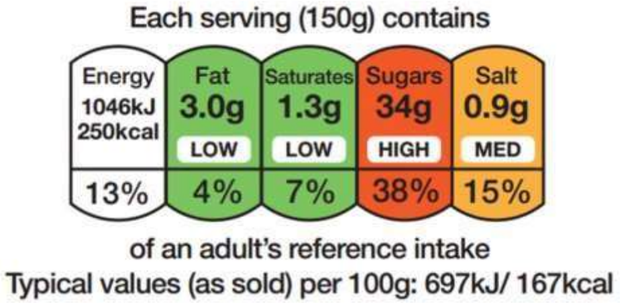
Les valeurs de nutriments ,
Choisir le meilleur produits.

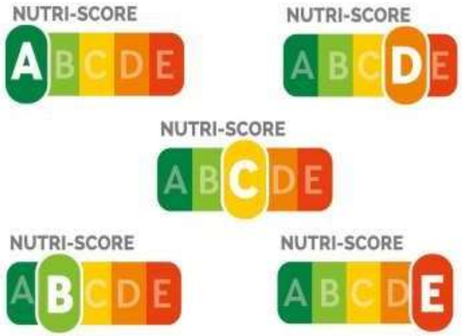
Nutri-score et nutri-grade:

La France applique la réglementation européenne et le tableau nutritionnel est obligatoire pour presque tous les aliments, conformément au Règlement 1169/2011.

L’objectif est d’améliorer l’information nutritionnelle figurant sur les produit et aider le consommateurs à choisir les aliments de meilleur valeurs nutritionnelles.

Ce calcul de nutri-score est basé sur les valeurs énergétiques et teneurs en graisses, acides gras saturé , sucre, sel, protéines pour 100g de produit.





Les données open Food Fact:

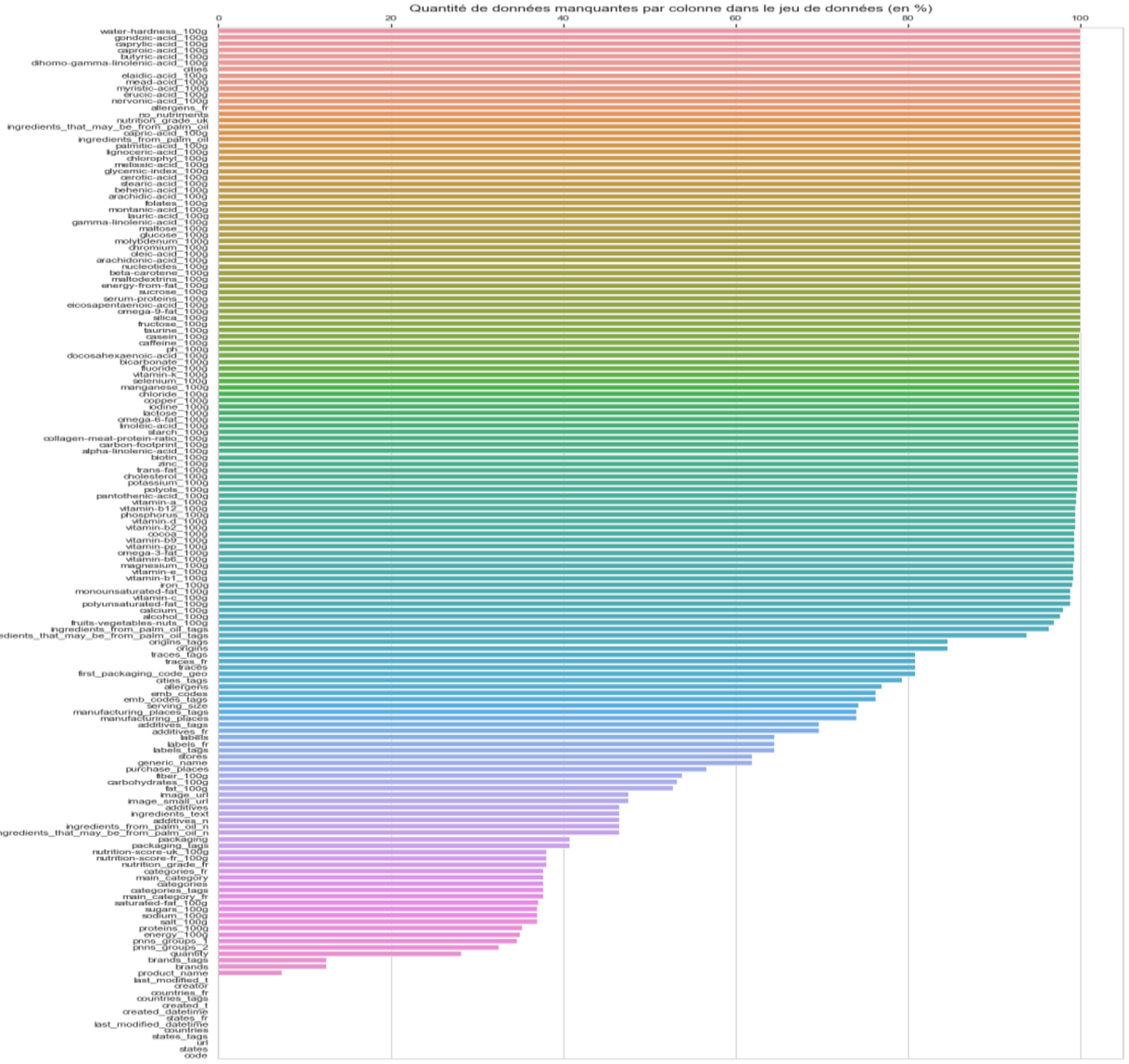
Les données ont été télechargés en format CSV sur la base de données Openfoodfacts ,puis convertir en dataframe avec les librairies pandas de python,

Le jeu de données comporte 320772 lignes et 162 variables,

- Le code de produit,
- Product name,
- Contries,
- Les nutriments ……

creator	created_t	created_datetime	last_modified_t	last_modified_datetime	product_name	generic_name	quantity
openfoodfacts-contributors	1474103866	2016-09-17T09:17:46Z	1474103893	2016-09-17T09:18:13Z	Farine de blé noir	NaN	1kg
usda-ndb-import	1489069957	2017-03-09T14:32:37Z	1489069957	2017-03-09T14:32:37Z	Banana Chips Sweetened (Whole)	NaN	NaN

Nettoyages des données:



- Dans ce jeu de données on trouves beaucoup de données peu renseignées.
- Des features qui ont un taux de remplissages inférieur à 25%.
- Suppression des features qui ont 80% des données manquantes, (danger d’interpolation de ses données).
- On passe de 162 variables à 60 variables.

Nettoyages des données:

Premier filtre :

nous réduisons notre étude uniquement aux produits alimentaires français.

Deuxième filtre :

Cette fois nous réduisons le nombre de colonnes en ne gardant que les informations importantes qui vont nous a identifier let analyser le nutri-score.
J'ai gardé que les nutriments et quelque colonnes qui donnes des informations sur les produits.

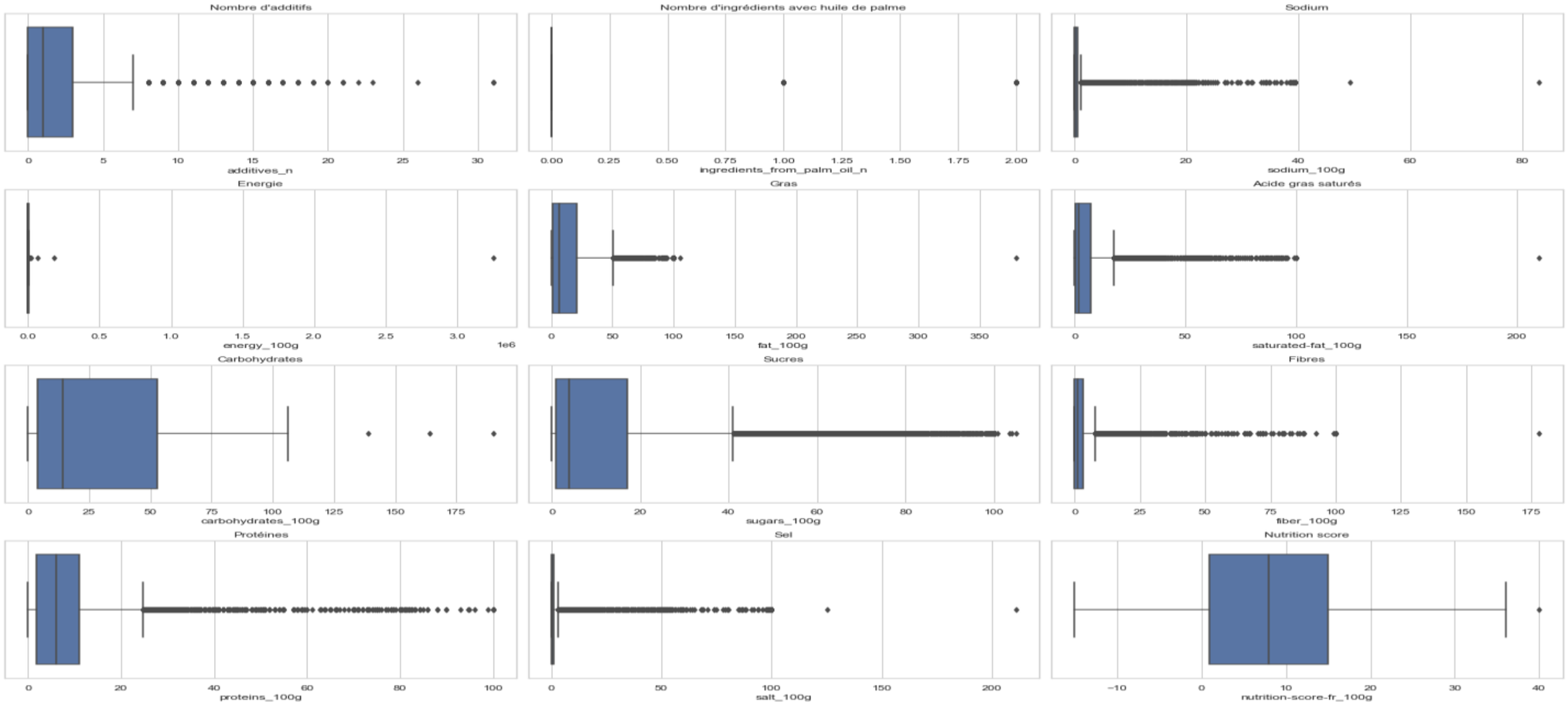
❖ **on arrive à avoir 18 colonnes.**

➢ Dans cette partie de nettoyages , nous avons aussi supprimées les **doublons** de produits.

Traitement des valeurs aberrantes:

Analyse descriptive:

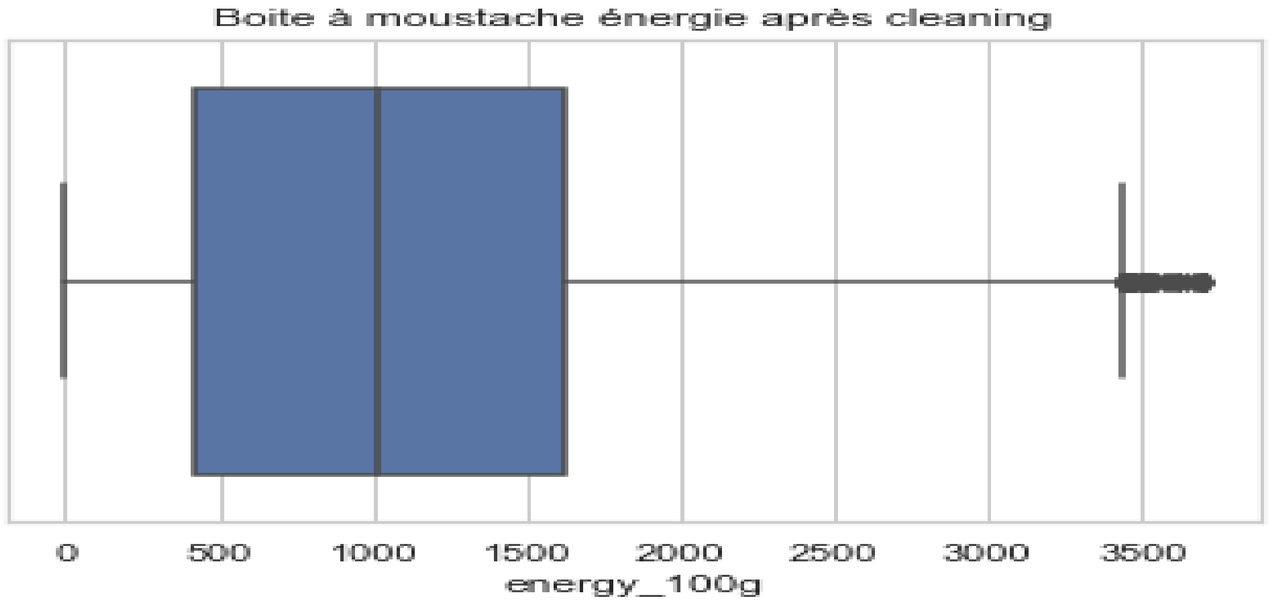
En réalisant une analyse rapide de données, on constate d'après les valeurs minimum et maximum de nutriments que ce jeu de données contient des valeurs aberrantes(valeurs négatives et valeurs très importantes.



Traitement des valeurs aberrantes:

Premier approche pour la suppression des valeurs aberrante est de :
Pour les variables suffixées avec_100g nous indique la quantité de nutriment pour 100g de produit, les valeurs renseignées ne doivent donc pas dépasser les 100g, nous allons supprimer les lignes dans la valeurs de nutriment supérieur au seuil.
D'une autre part la variable gras saturé ne doit pas dépassée le taux de gras , le sodium ne dépasse pas le taux de sel ,,,,

Pour la variable énergie nous avons supprimés les valeurs aberrantes par la method percentile vu que son écart type et la moyenne sont des valeurs extrême



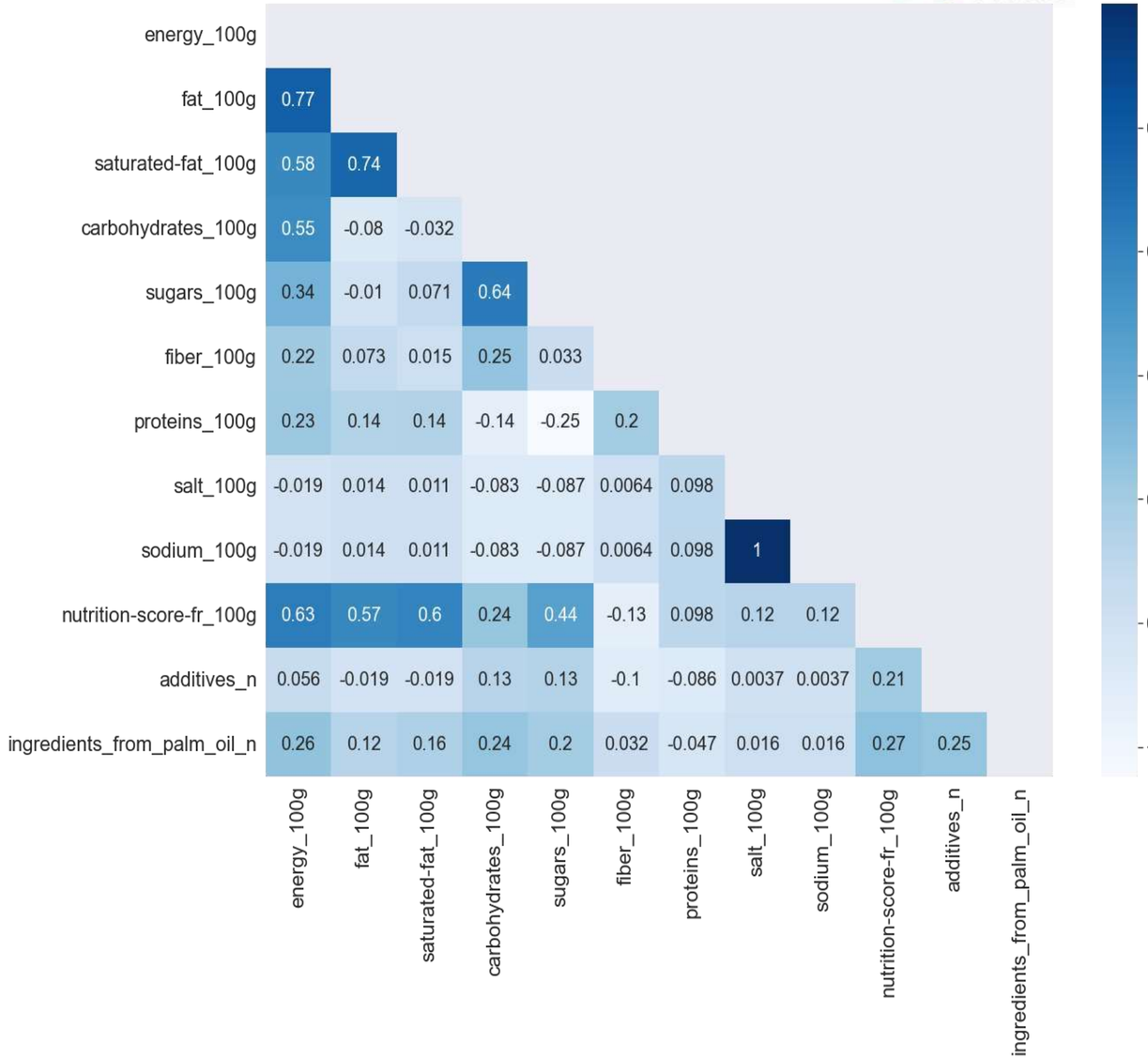
Traitement des valeurs manquantes:

Une forte corrélation entre:

- ✓ fat_100g et energy_100g.
- ✓ sugars_100g et carbohydrates_100g.
- ✓ Fat_100g et saturated_fat_100g.
- ✓ Sel_100g et sodium_100g.

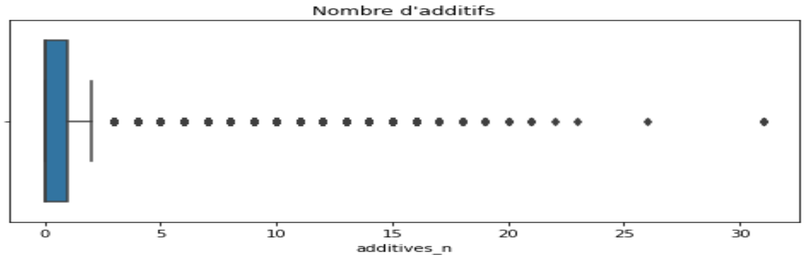
Pour ces variables on va remplir les valeurs manquantes par la fonction IterativeImputer de Scikit-learn.

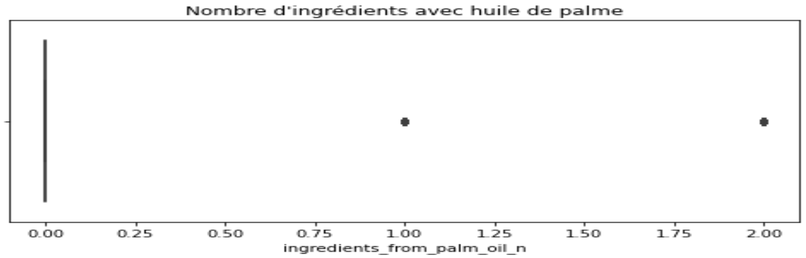
Pour les variables ('fiber_100g', 'proteins_100g', , 'nutrition-score-fr_100g') , nous remplaçons les valeurs manquantes par la moyenne par catégorie d'aliments. La catégorie de l'aliment est donné dans la variable "pnns_groups_1" : aliments sucrés, aliments salés, fruits et légumes, gras et sauces.....

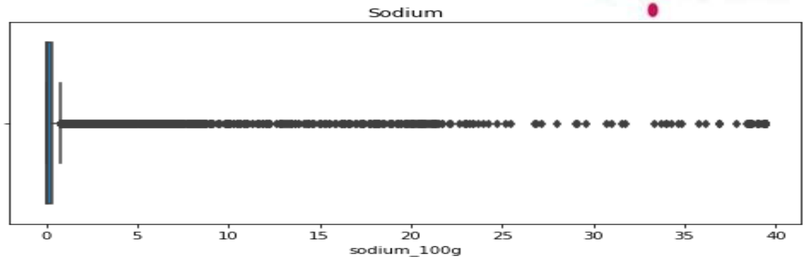


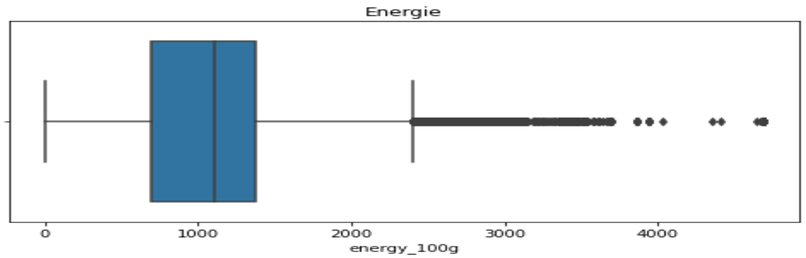
Analyse univarié:

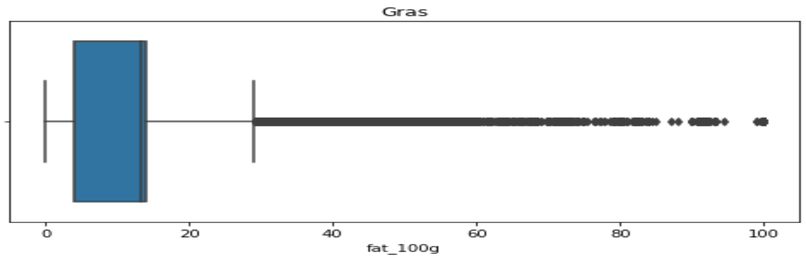
D’après le nettoyages des données, boxplot indique la distributions des données.

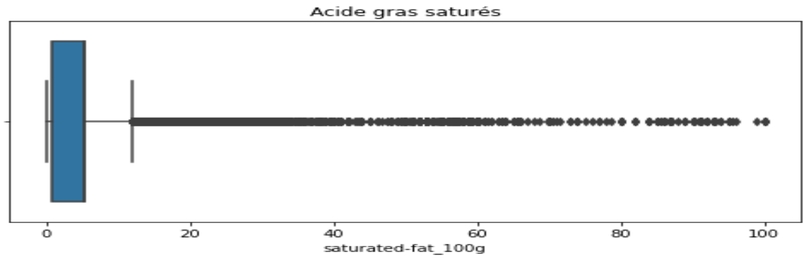


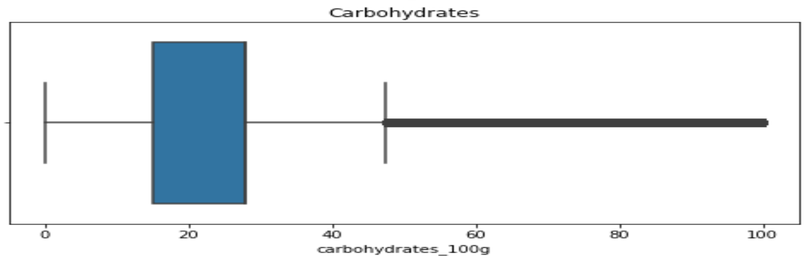


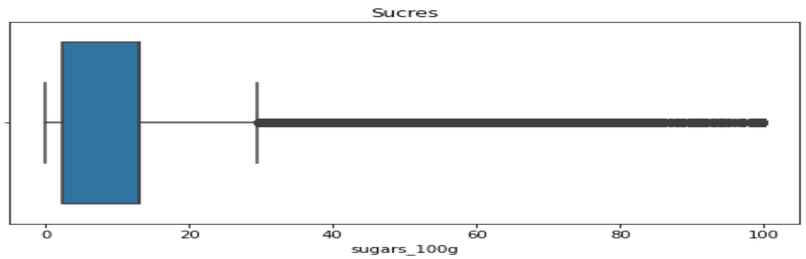


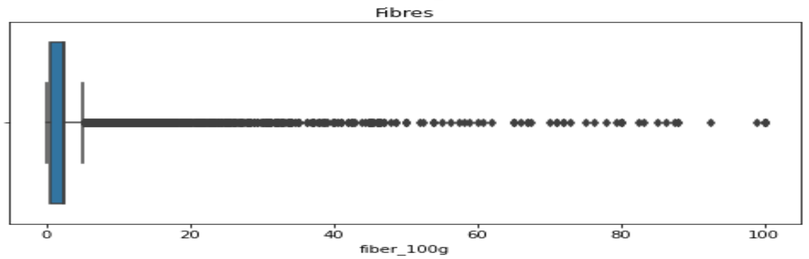


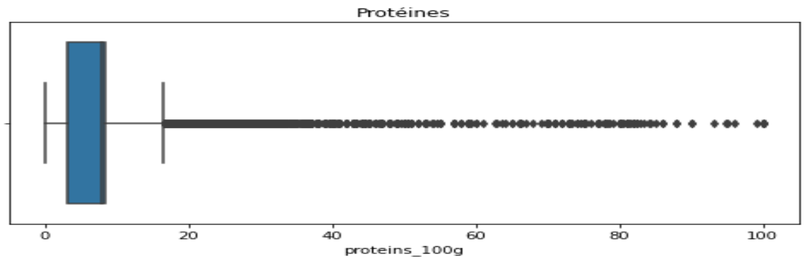


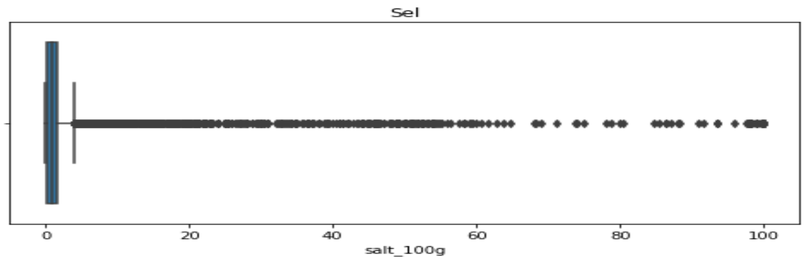


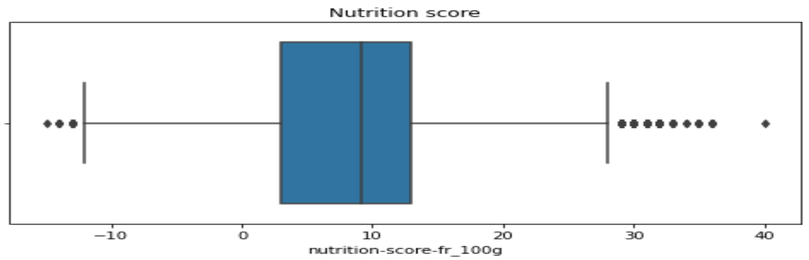








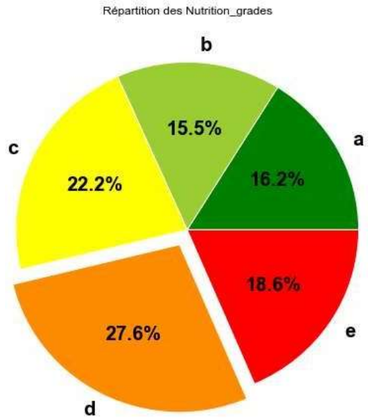


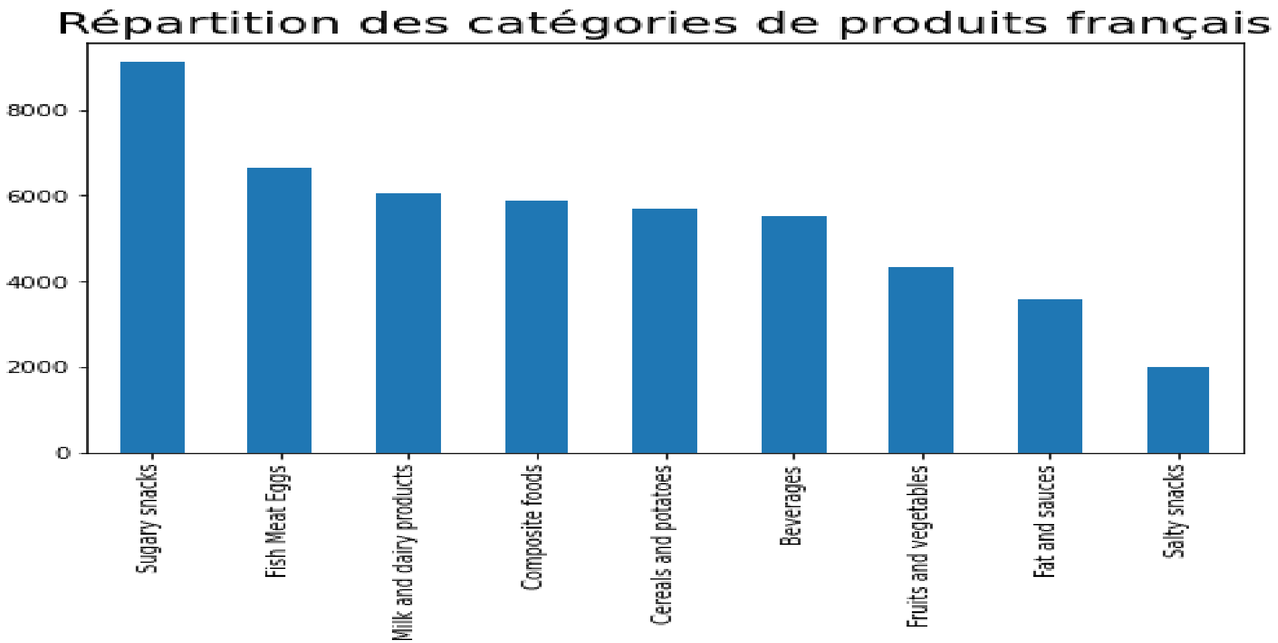


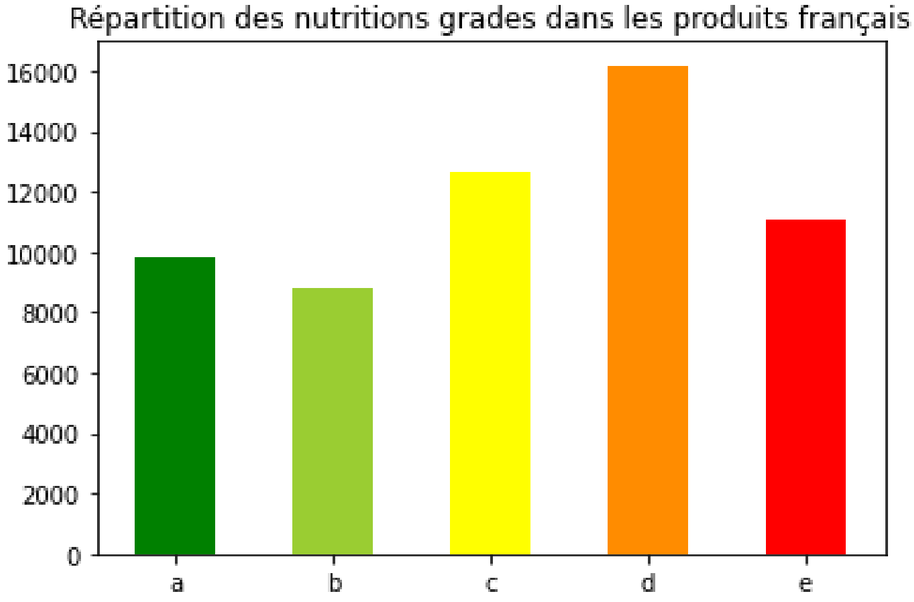
Analyse univarié:

Répartition des produits Français en fonction de nutri-grade

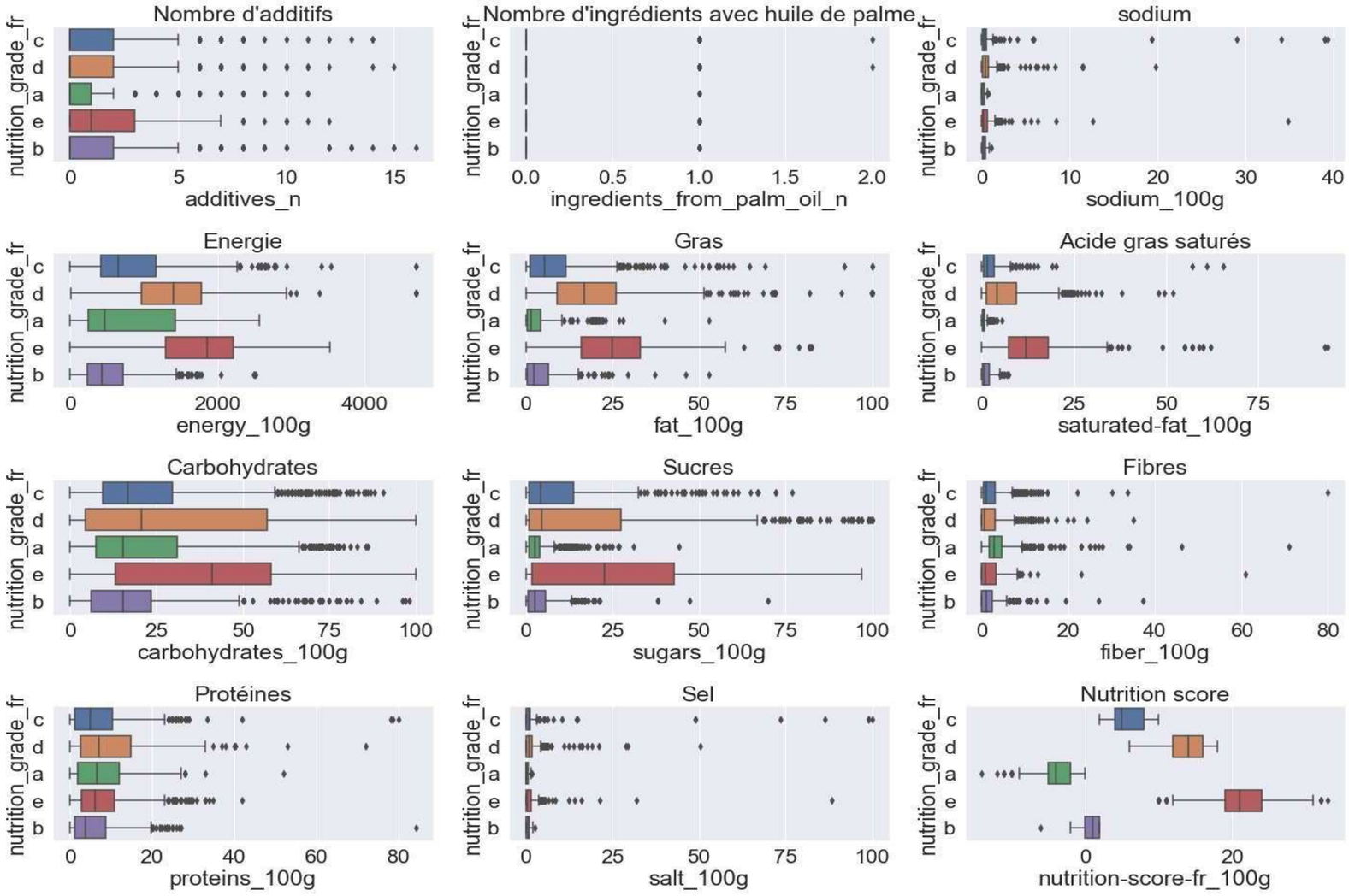
Le plus grand nombre de produits français (28%) ont un nutri-score de C ou D c'est à dire une qualité nutritionnelle moyenne. Les produits de très bonne qualité nutritionnelle (A et B) sont un peu moins nombreux (env 16%). Quant aux produits de très mauvaise qualité(E) leur quantité est tout de même élevée (18.6%)





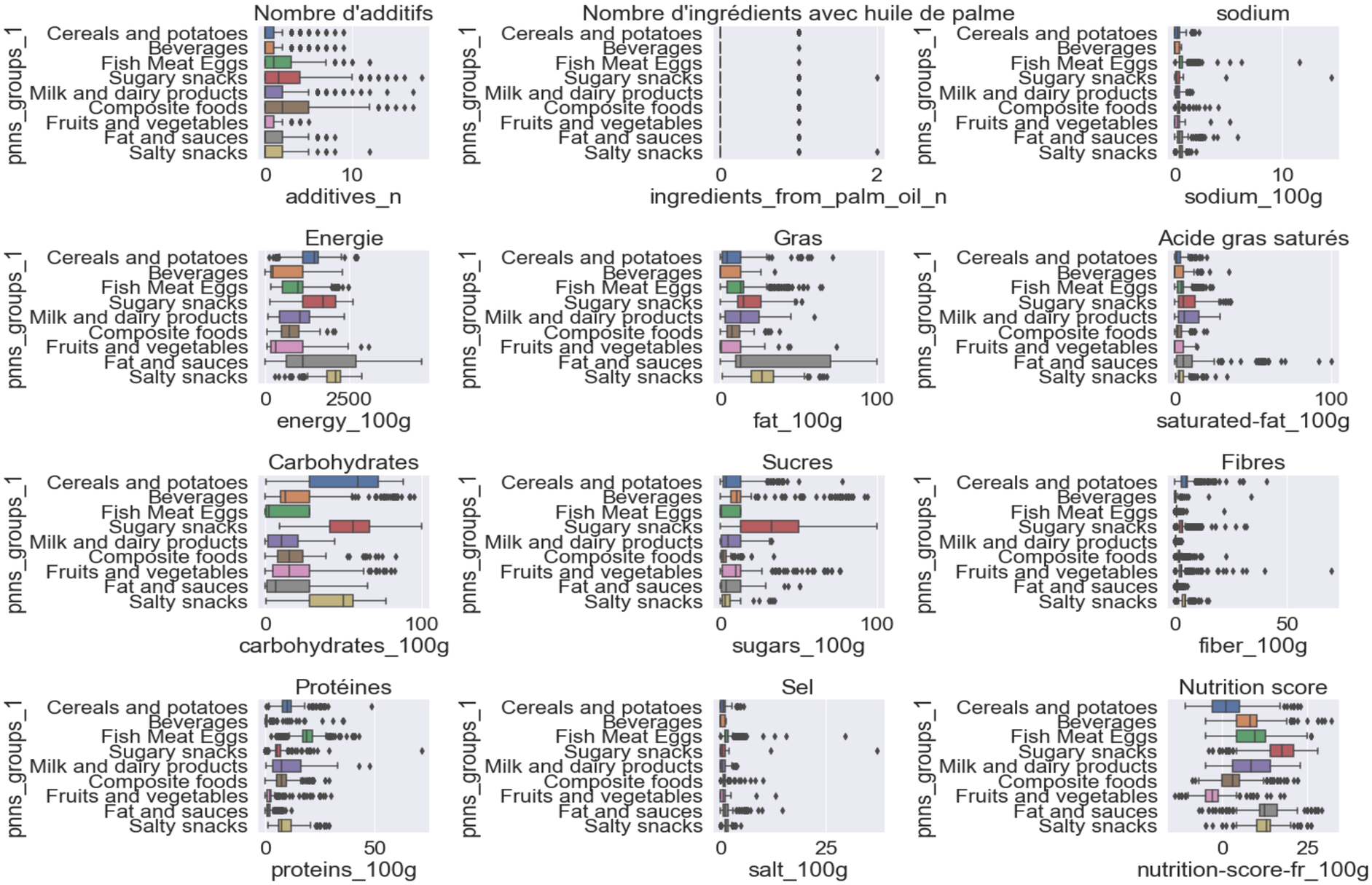


Les valeurs de nutriments en fonction de nutri-grade:



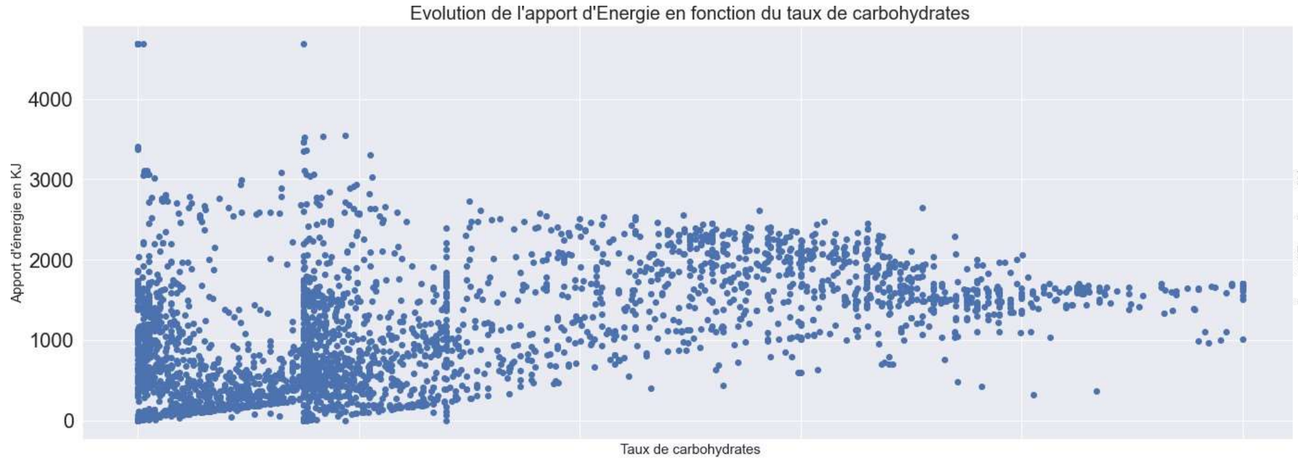
D’après le graphe en remarque que : les nutriments semble avoir un impact sur la distributions des Nutri-grade.

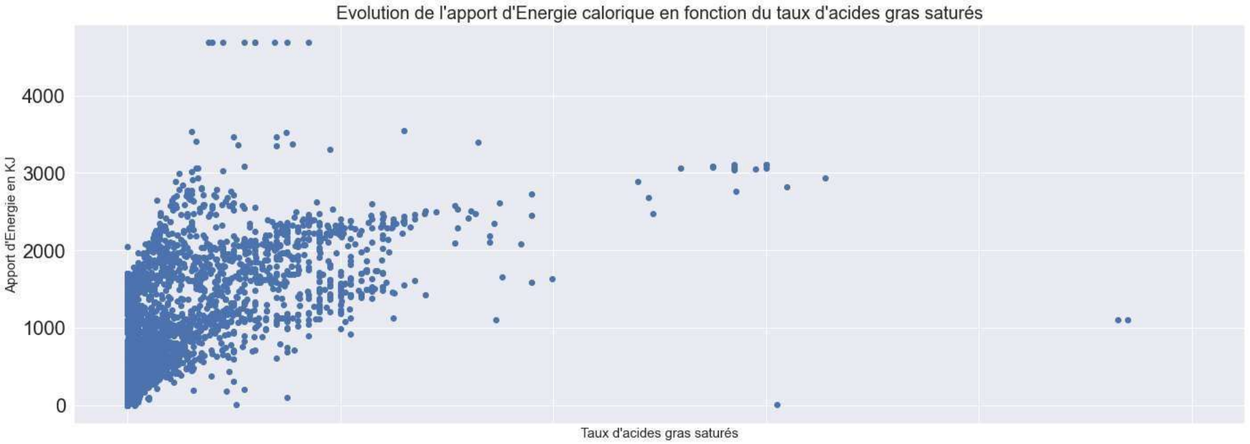
Taux de nutriments en fonction de catégorie de l'aliment:

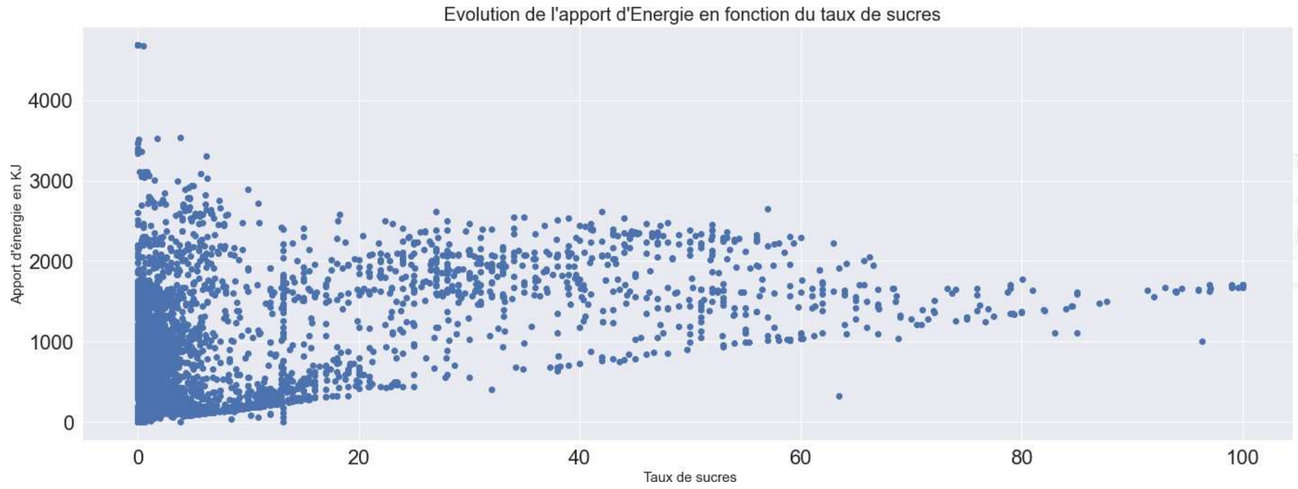


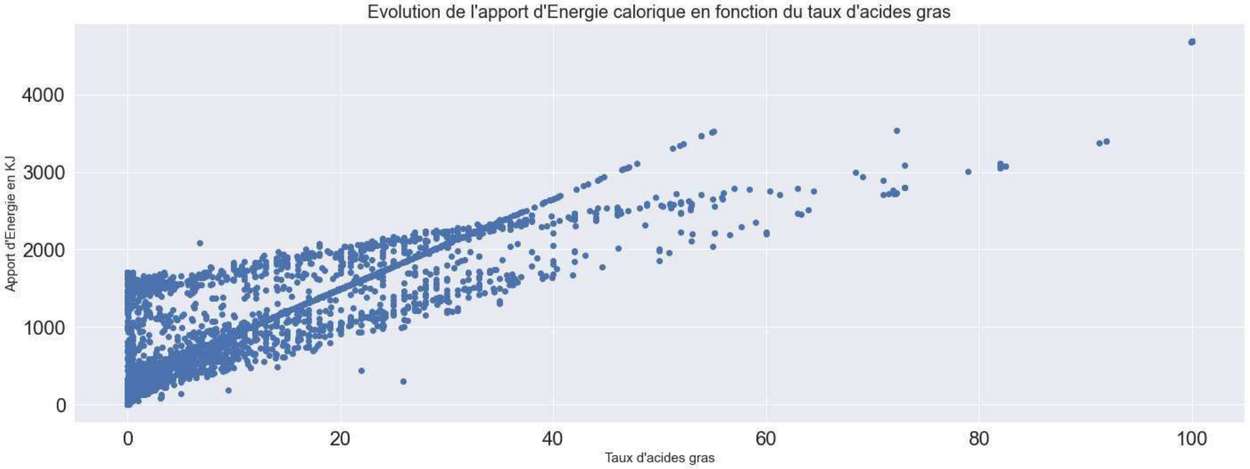
Visualisation des différents taux en fonction de la catégorie de l'aliment

Analyse bivariée:



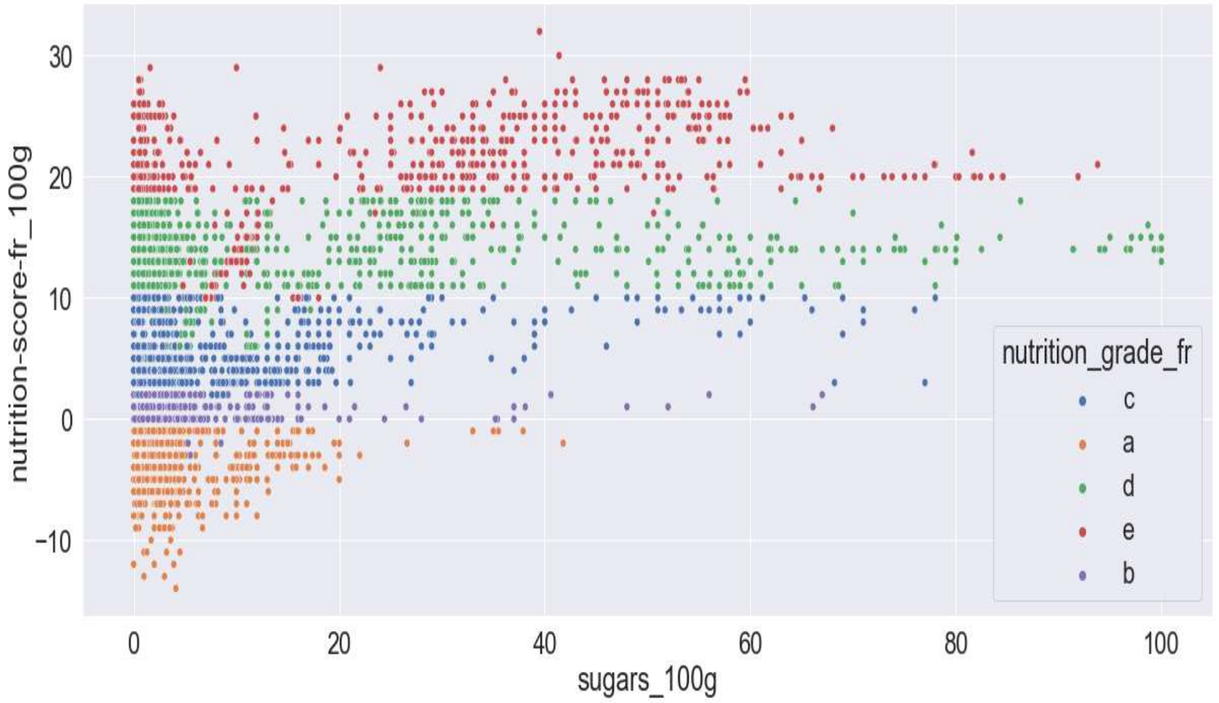


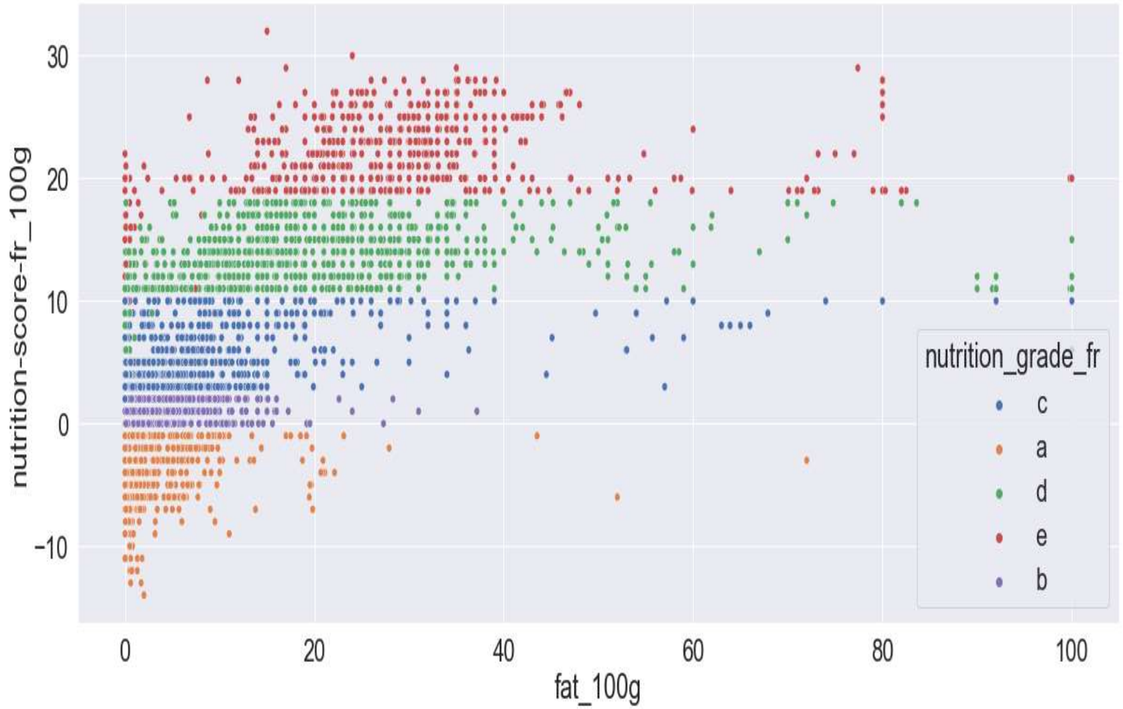




Visualisation de l'apport d'énergie en KJ des aliments selon leur taux de carbohydrates, sucres, de gras saturés et de gras : l'apport calorique augmente avec le taux de chacune de ces variables.

Test de Pearson:



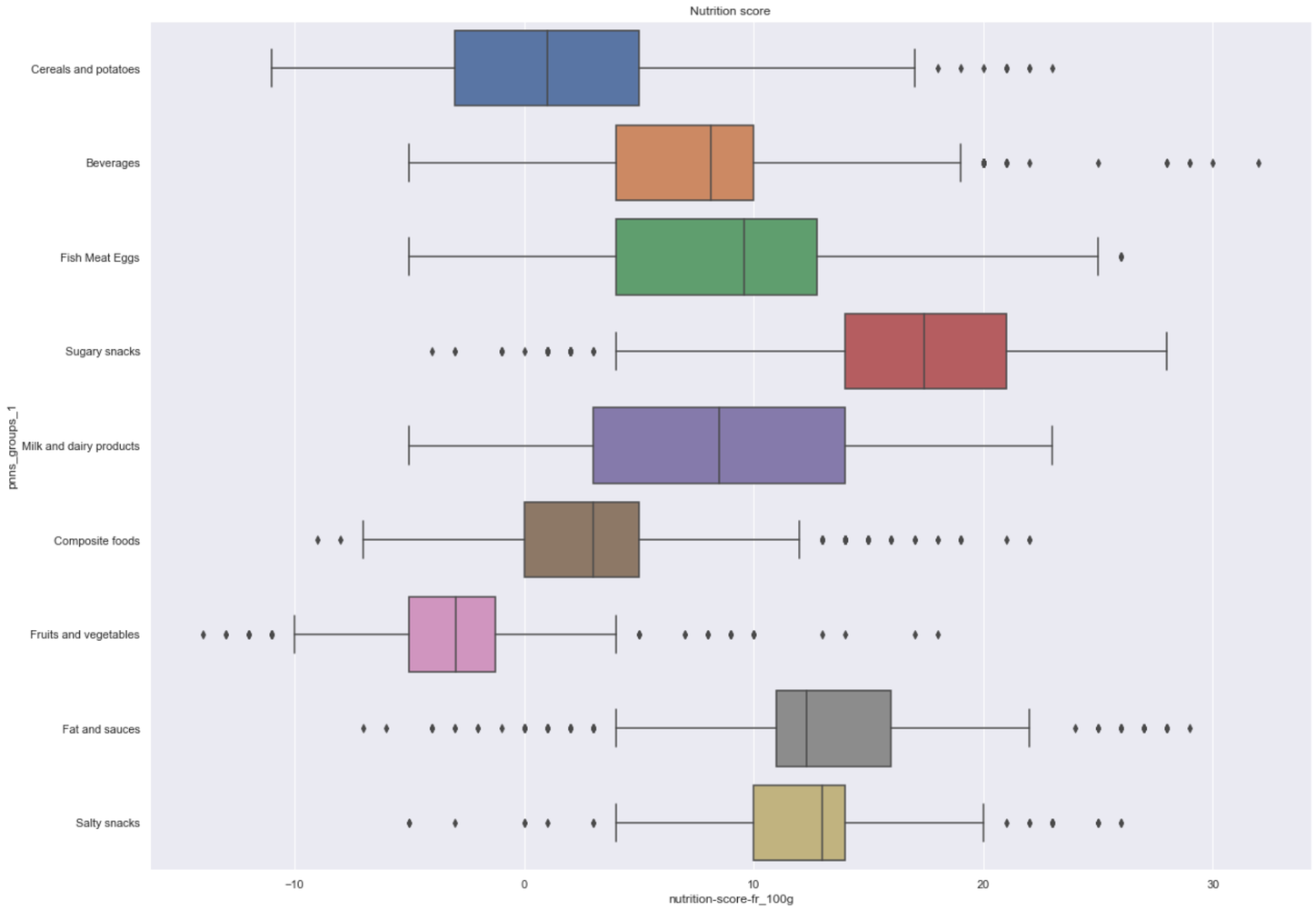


La relation des pairs des variables et nutri-grade qui est indiqué sur cette figure montre:

- Une corrélation fortement linéaire pour les variables nutritionnelles, on trouve des catégorisations par nutri-score est fortement marqué, des groupes sont bien séparés.

En réalisant le test de Pearson, on rejette l'hypothèses nulle qui indique une faible corrélation entre les nutriments.

Analyse de la variance ANOVA:



Pour vérifier que la catégorie pnns_group_1 influence réellement sur le nutri-score ,Nous allons réaliser une analyse de variance (ANOVA).

Les hypothèses posées sont:

H0: La distribution des échantillons est similaire(la catégorie n'a aucune influence sur le Nutri-score.

H1 : Une ou plusieurs distributions sont inégales(la catégorie a une influence sur le nutri-score.

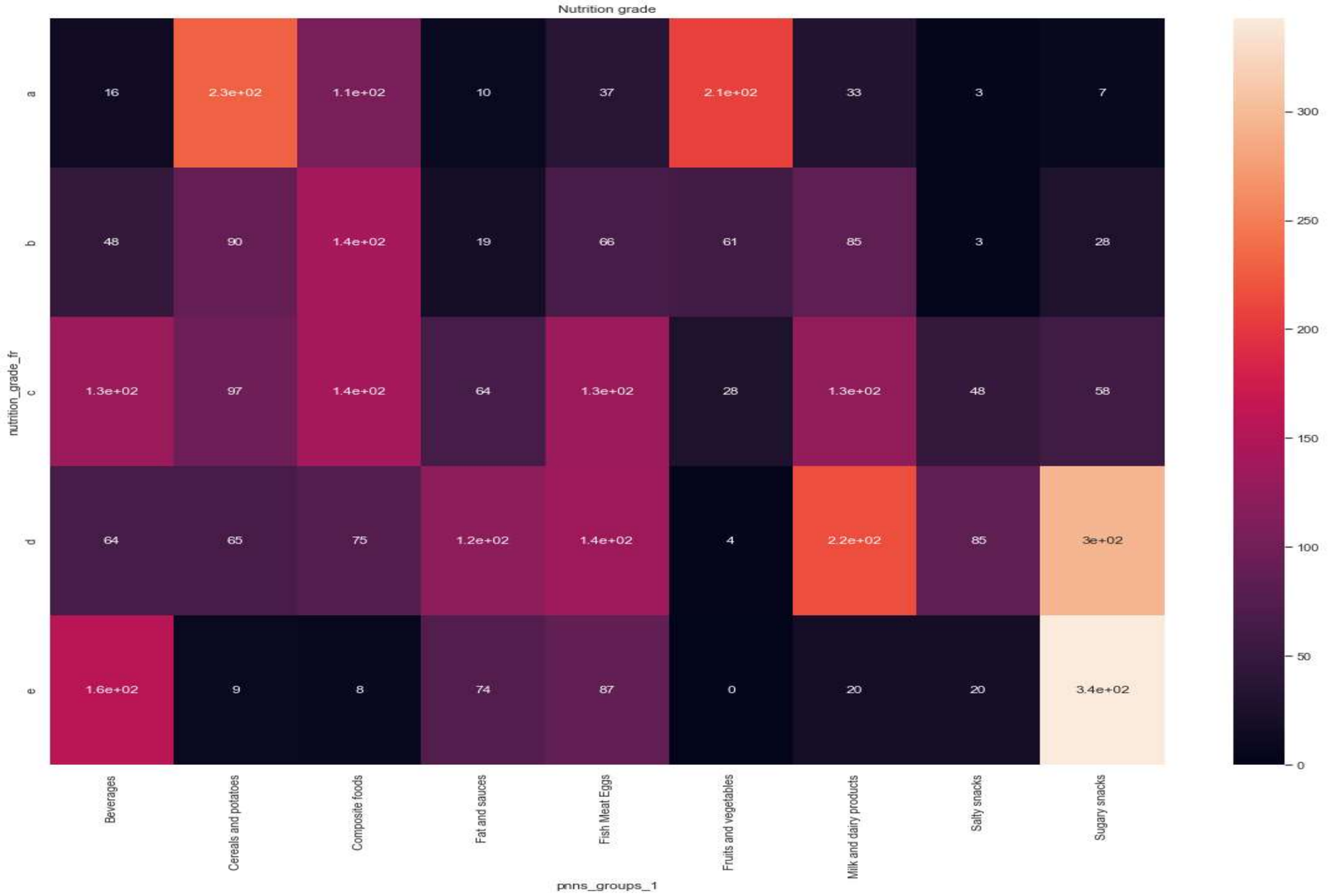
Test statistiques khi_deux:

Maintenant on va vérifier si la catégorie pnns_group_1 influence réellement sur le nutrition-grade,

Nous allons réaliser le test statistique de ckhi_2 .

H0: La distribution des échantillons est similaire(la catégorie n’a aucune influence sur la Nutriti-gade.

H1 : Une ou plusieurs distributions sont inégales(la catégorie a une influence sur la nutrition-grade.



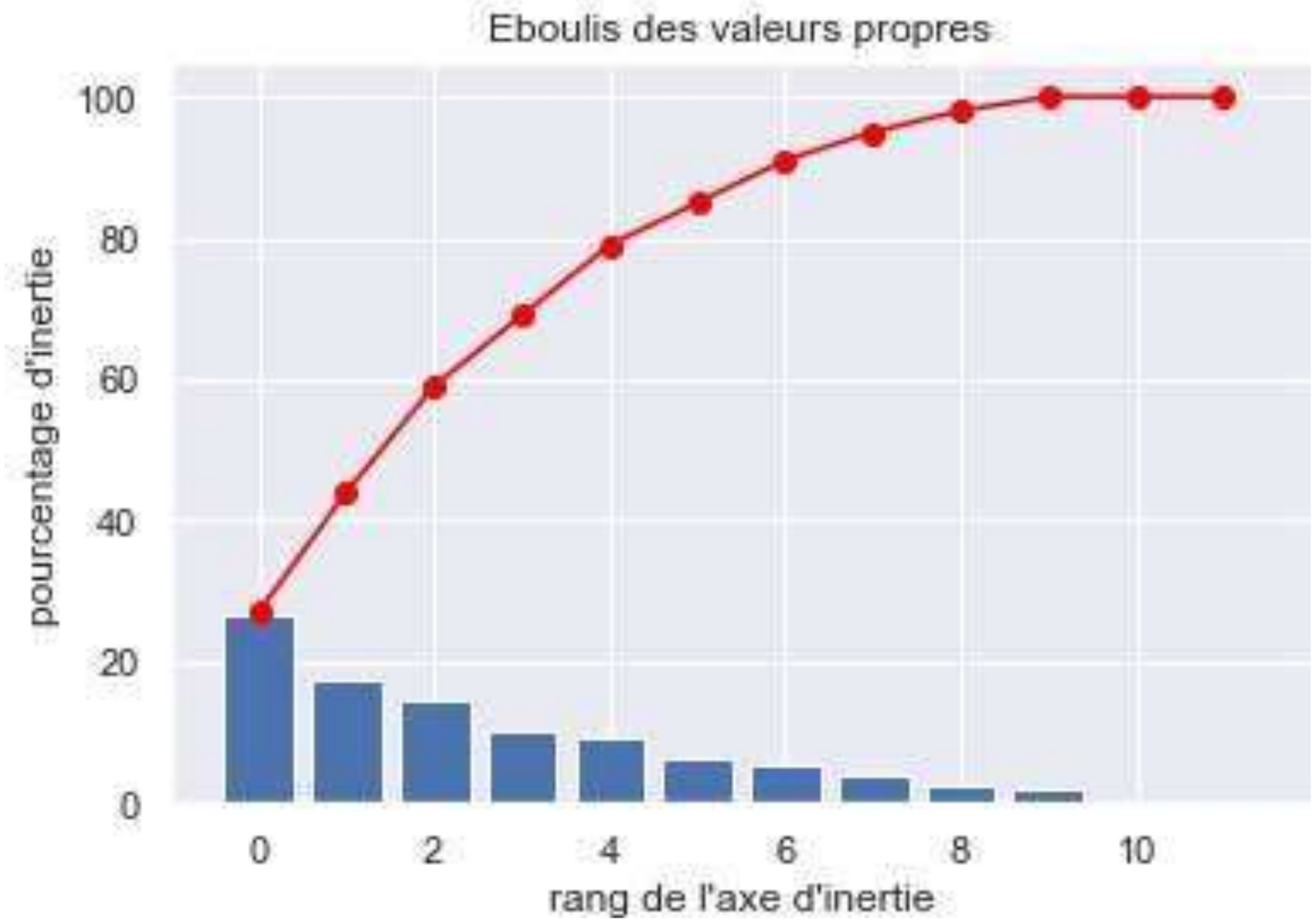
Réduction dimensionnelle par ACP:

Analyse de composante principale ACP, L'une de méthode d'analyse multivariée les plus utilisée,

Elle permet d'explorer le jeu de données multidimensionnels constitués de variables quantitatives et de créer des variables synthétiques exploitables,

Nous allons ici réaliser un cercle des corrélation de nos variables puis projeter les individus sur les premiers plans factoriels.

Inertie cumulé sur les axes de plans factoriels:



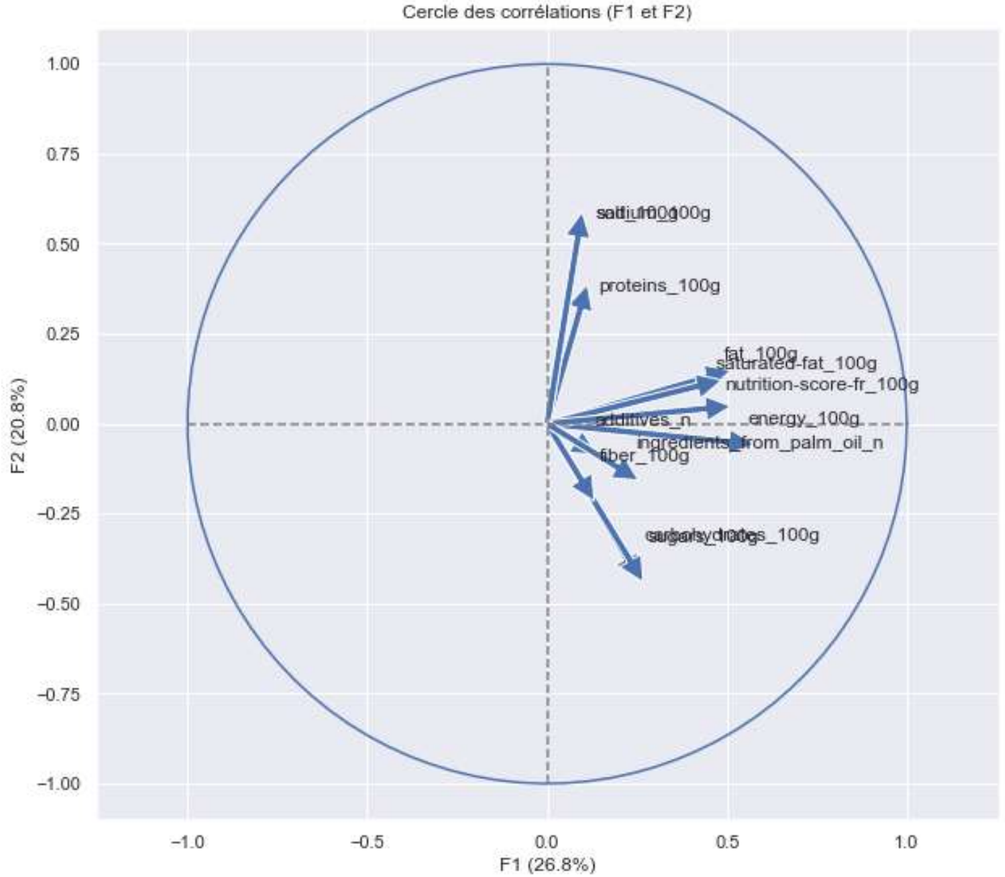
Afin d’avoir un aperçu de nombre de composantes nécessaire à l’analyse

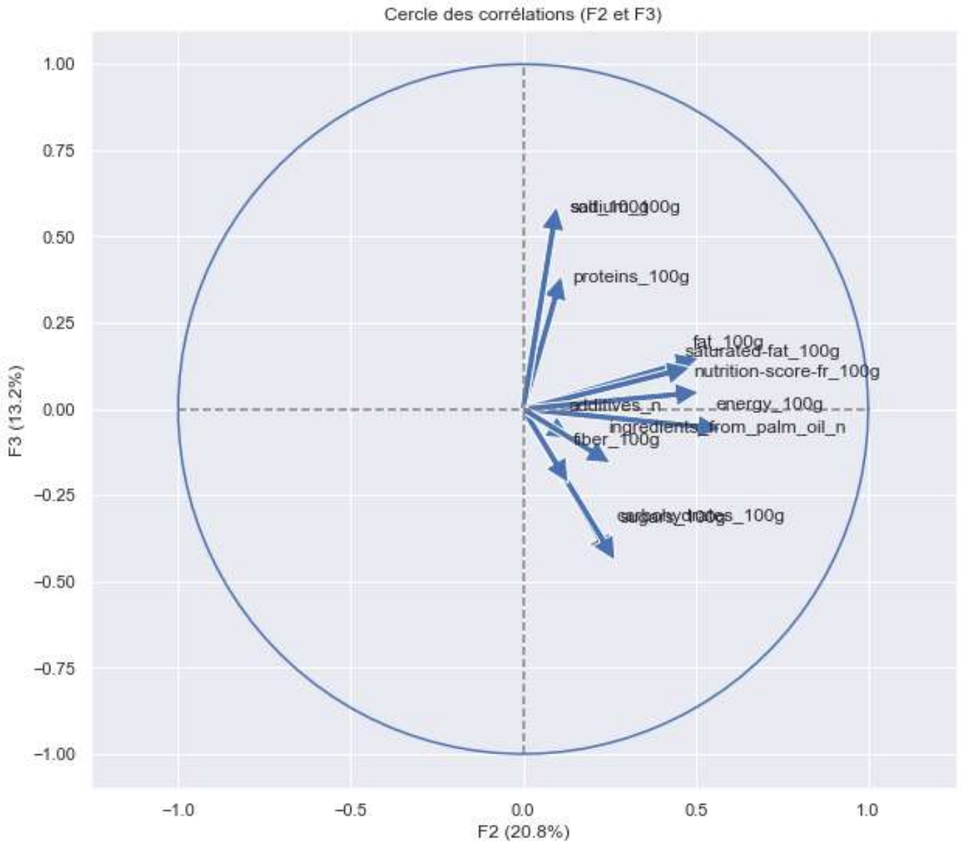
nous projetons l’éboulis de valeurs propres,

En bleu la variance de chaque composante, et en rouge la variance cumulée. On voit ici qu'environ 60% de la variance est comprise dans les 3 premières composantes, et qu'environ 70% dans les 4 premières.

.

Cercles des corrélation:





Le coefficient de corrélation, linéaire est représenté par le cosinus de l'angle entre 2 variables, plus la pointe de la flèche est proche du cercle plus la variable est représentative de l'axe.

L'axe F1 va représenter le caractère nutri-score tandis que l'axe F2 représente le caractère sel et sodium et carbohydrate,

Conclusion:

- Lors de nettoyages et analyses exploratoires, nous avons pu évaluer la base de donnée de l'openFoodFacts afin de vérifier qu'elles pouvait servir de base à notre application. Cette base seule ne peut pas suffire à évaluer les Nutri-score et Nutri-grade avec une précision satisfaisante., il faut donc compléter les données de la base de donnée avec d'autres base.

- D'après l'analyse exploratoire et la visualisations des taux nutri-score et nutri-grade notre application recommande des catégories dans pnns-group qui ont de valeurs nutritionnelles faible e, avec bon nutri-grade, ceux qui favorise certains nutriments et limiltes certains d'autre.