

Parcours Ingénieur Machine Learning

Session Mars 2021

OPENCLASSROOMS

Projet 2

Concevez une application au service
de la santé publique

28/04/2021

Etudiante : QITOUT Kenza

Mentor : Maïeul Lombard

Evaluateur : Thierno Ibrahima Diop

CONTEXTE DU PROJET

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



Objectifs :

- ❖ Proposer une idée d'application
- ❖ Traiter et nettoyer le jeu de données mis à disposition
- ❖ Exploiter le jeu de données nettoyé (analyses univariées, analyses multivariées, réduction dimensionnelle)
- ❖ Visualiser les données et communiquer les résultats
- ❖ Conclure sur la faisabilité de l'idée d'application

BASE DE DONNEES

Base de données **Open Food Fact** (disponible sur <https://world.openfoodfacts.org/>)



Base de données collaborative et ouverte sur des produits alimentaires du monde entier
Répertorie les informations sur le produit, les ingrédients, les additifs, la composition nutritionnelle, l'origine du produit, ...

Base de 184 variables (décrites sur [ce lien](#)) et 1 658 499 produits

Fichier .csv de 3.40 Go

1. PRESENTATION DE L'IDEE D'APPLICATION

Problématique : Evaluer la qualité des produits alimentaires

- 1) Etudier et visualiser les relations entre les variables impliquées
- 2) Notation de la qualité des produits
- 3) Finalité : associer chaque produit (par nom et code-barre) à un indice de qualité et une note

Notation sur 100 basée sur :

- La qualité nutritionnelle
- La quantité d'additifs
- Le nombre d'allergènes
- le degré de transformation

2. NETTOYAGE DE LA BASE DE DONNEES

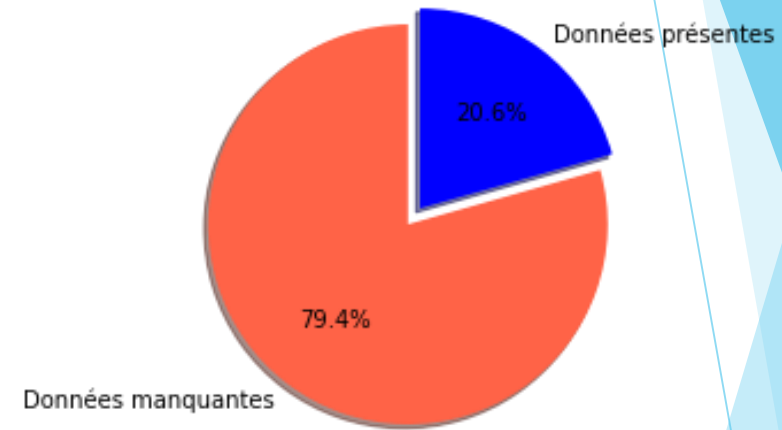
Jeu de données avec **79.4%** de données manquantes

- Suppression des 9 colonnes vides (sans suppression de ligne) => 21.7 % de données non manquantes

Nettoyage des **valeurs erronées** :

- Valeur des nutriments < 0g ou > 100g
- pH non compris entre 1 et 14
- Energie < 0
- Indice de glycémie < 0 ou > 100
- Somme des nutriments > 100g (utilise les nutriments impliqués dans le calcul du score de nutrition)

Diagramme circulaire des données présentes et manquantes



6264 lignes
supprimées

21.6% de données
non manquantes

2. NETTOYAGE DE LA BASE DE DONNEES


Choix d'étudier l'application uniquement sur des produits vendus en France (représentant presque la moitié du jeu de données) avec la variable 'countries'

 880 248 lignes supprimées (reste 45% du jeu de données initial)

Beaucoup de variables présentent dans le jeu de données (172 colonnes) avec des colonnes contenant la même information

Colonnes en double sur les variables utiles pour l'application :

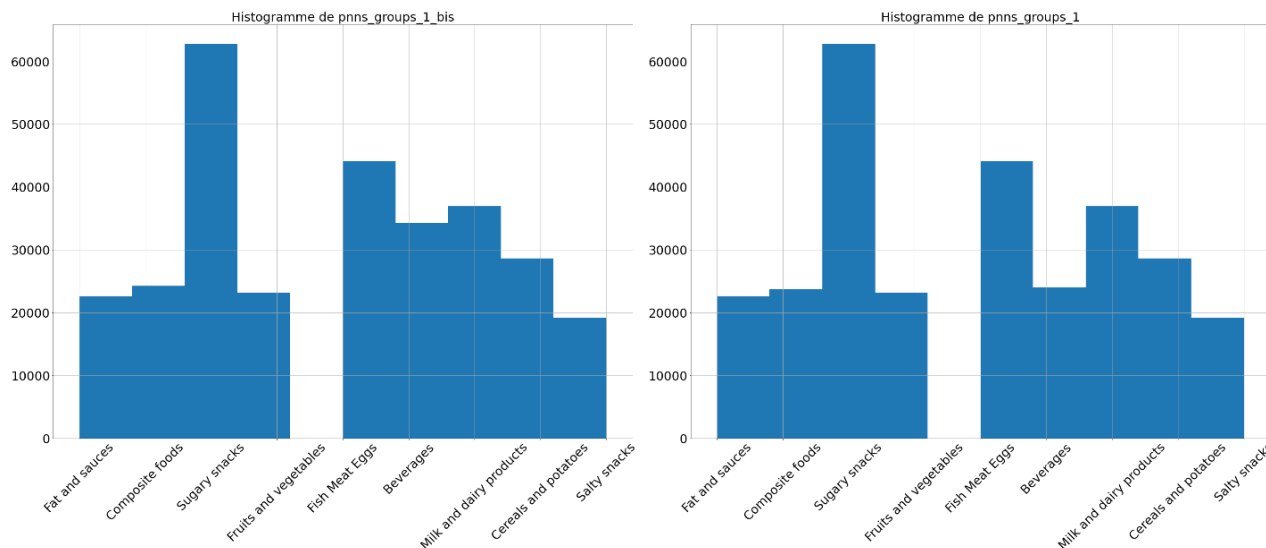
- Choix entre les variables temporelles (corrélation de Pearson)
- Choix entre les variables du score de nutrition (différence au carré)
- Choix entre les variables d'énergie en kJ (différence au carré)

 0 ligne supprimée, 761 647 lignes (45% du jeu de données initial) et 18% de données non manquantes

2. NETTOYAGE DE LA BASE DE DONNEES

Nettoyage des colonnes 'pnns_groups_1' et 'pnns_groups_2' :

- Catégorie 'unknown' : 465 488 (pour 'pnns_groups_1') et 465488 (pour pnns_groups_2') -> considérée comme NaN
- Regroupement des différentes façons d'écrire les catégories : 'sugary-snacks' et 'Sugary snacks', 'salty-snacks' et 'Salty snacks', ...
- Imputation de 'pnns_groups_1' grâce aux données de 'pnns_groups_2'
- Pas de différence dans les distributions



761 647 lignes et de 167 colonnes soit 45% du jeu de données initial
17% de données non manquantes

2. NETTOYAGE DE LA BASE DE DONNEES

Idée d'application basée sur les variables du nutriscore, des catégories, du nova group, des allergènes et des additifs

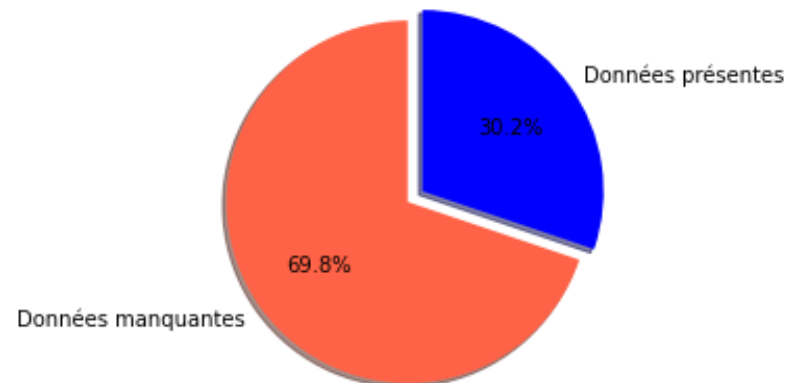
=> Suppression des lignes ne contenant pas de données pour les variables :

'product_name', 'code', 'url', 'nutriscore_grade', 'nutrition-score-fr_100g', 'pnns_groups_1_bis', 'pnns_groups_2', 'brands', 'nova_group'



634 447 lignes supprimées suite à ce nettoyage, 7% du jeu de données initial, 127200 lignes et 167 colonnes

Diagramme circulaire des données présentes et manquantes



2. NETTOYAGE DE LA BASE DE DONNEES

Création d'une **nouvelle variable** :

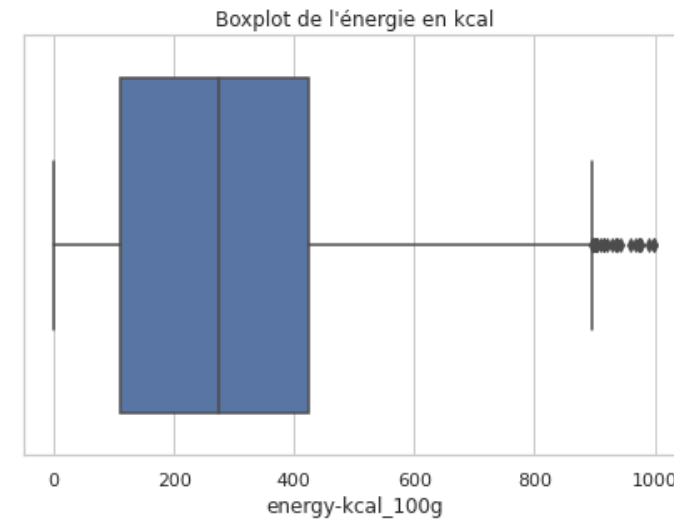
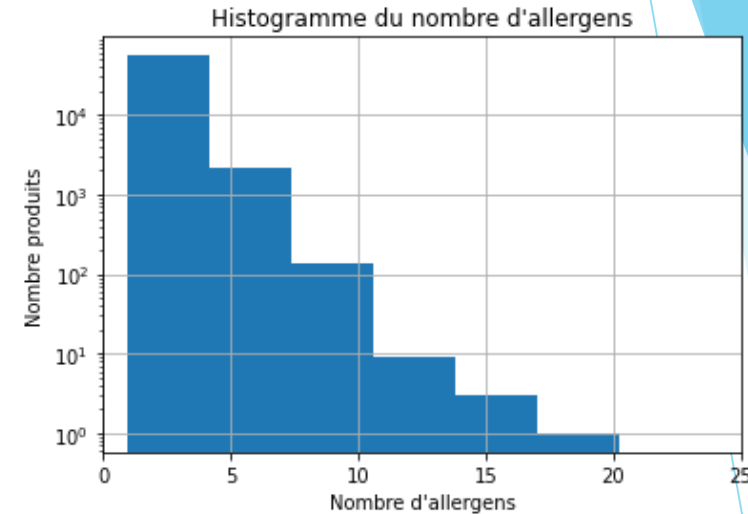
Transformation de la colonne 'allergens' en 'allergens_n' correspondant au nombre d'allergènes par produit

Valeurs erronées :

- Énergie > 1000 kcal et > 4500 kJ (limite théorique)
- Conversion entre les unités des énergies en kJ et en kcal (environ 4,18)








1242 lignes supprimées
125 958 lignes dans le DataFrame



2. NETTOYAGE DE LA BASE DE DONNEES

Valeurs incorrectes : vérification de la cohérence entre 'nutriscore_grade' et 'nutrition-score-fr_100g'

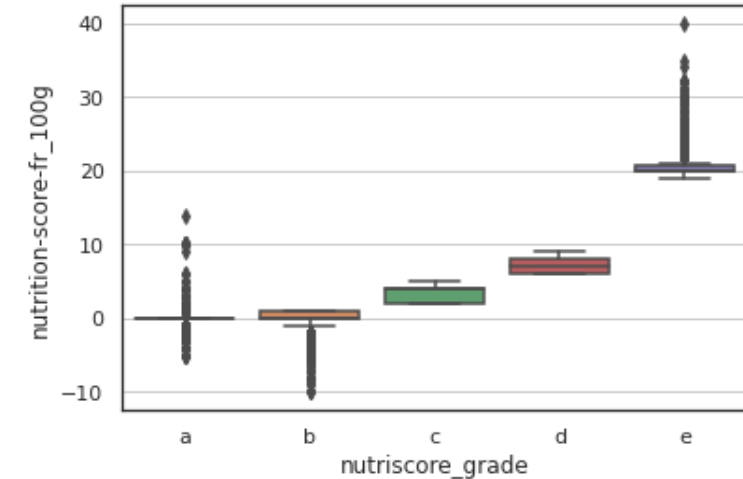
Points		Logo
Solid foods	Beverages	
Min to -1	Waters	
0 - 2	Min - 1	
3 - 10	2 - 5	
11 - 18	6 - 9	
19 - max	10 - max	



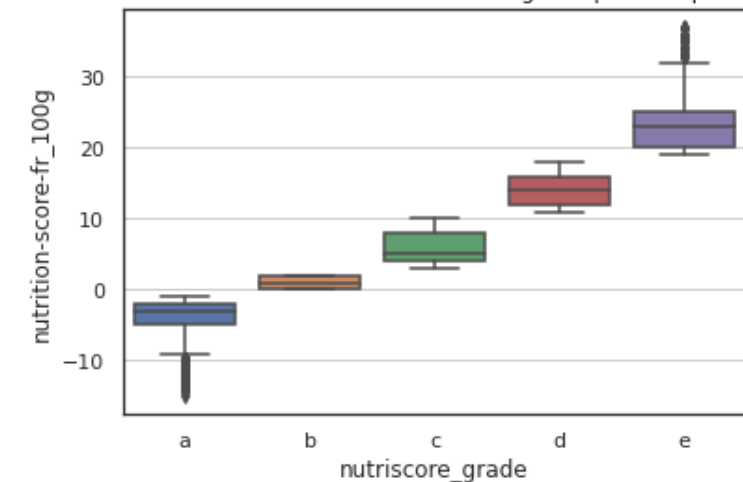
4577 lignes supprimées

121 381 lignes et 167 colonnes

Boxplot du score de nutrition en fonction nutriscore grade pour les Boissons



Boxplot du score de nutrition en fonction nutriscore grade pour les produits hors Boissons



2. NETTOYAGE DE LA BASE DE DONNEES

DataFrame composé de beaucoup de données manquantes (70%) et de colonnes vides : supprimer les colonnes contenant moins d'un certain seuil de données non manquantes => **Seuil pris à 10%**

4 577 lignes supprimées



121 381 lignes et 67 colonnes

7% du jeu de données initial

Sélection des **variables utiles** : variables sur les informations du produit, variables de l'application et les variables des quantités de nutriment restantes



0 ligne supprimée

95% de données non manquantes

Diagramme circulaire des données présentes et manquantes

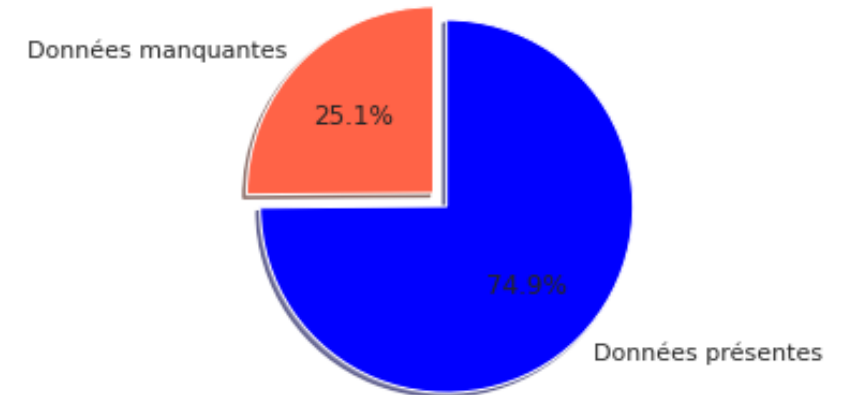
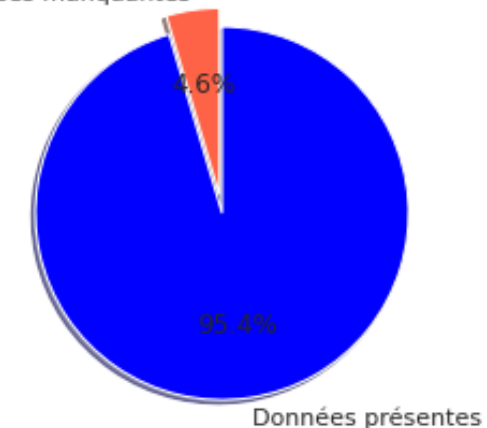


Diagramme circulaire des données présentes et manquantes



2. NETTOYAGE DE LA BASE DE DONNEES

Doublons : Suppression des doublons de code-barre, d'url et de nom de produit/marque en gardant la ligne de modification la plus récente



8 591 lignes supprimées, 112 781 lignes et 23 colonnes
6% du jeu de données initial

Imputation par 0 : Hypothèse sur les données manquantes des nutriments, du nombre d'allergènes et d'additifs



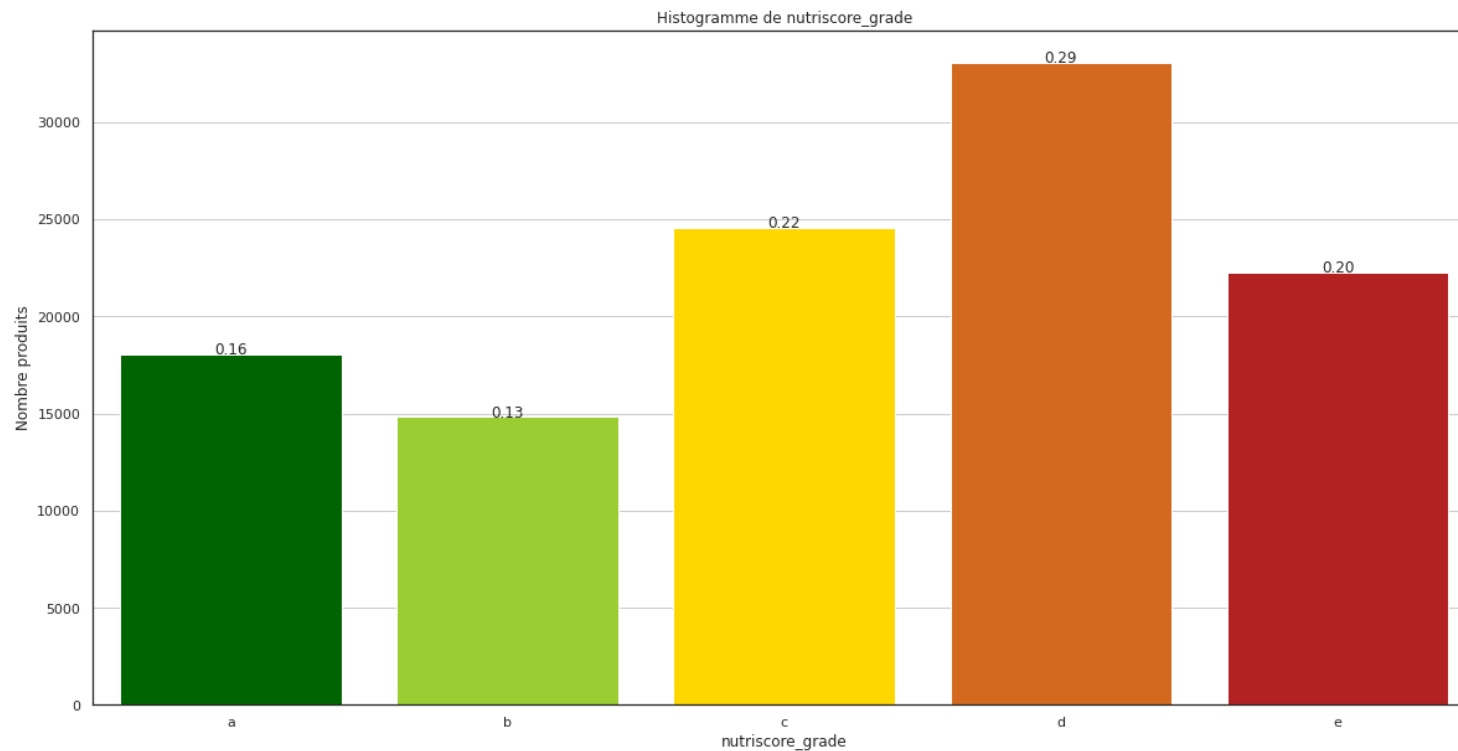
99% de données non manquantes

Vérification des **Outliers** (avec les boxplot) => pas de traitement nécessaire

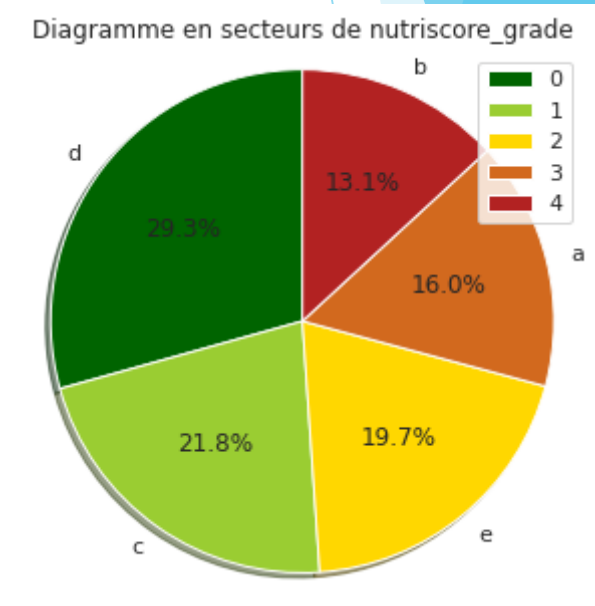
Le DataFrame final est composé de **112 781 lignes** et de **19 colonnes**, ce qui représente **6%** du jeu de données initial avec **1 545 718 lignes** supprimées au total.

3. EXPLORATION DES DONNEES

Analyse univariée de nutriscore_grade

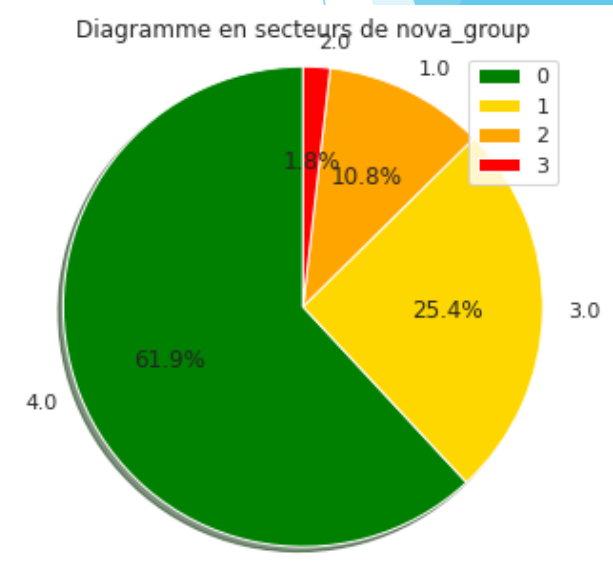
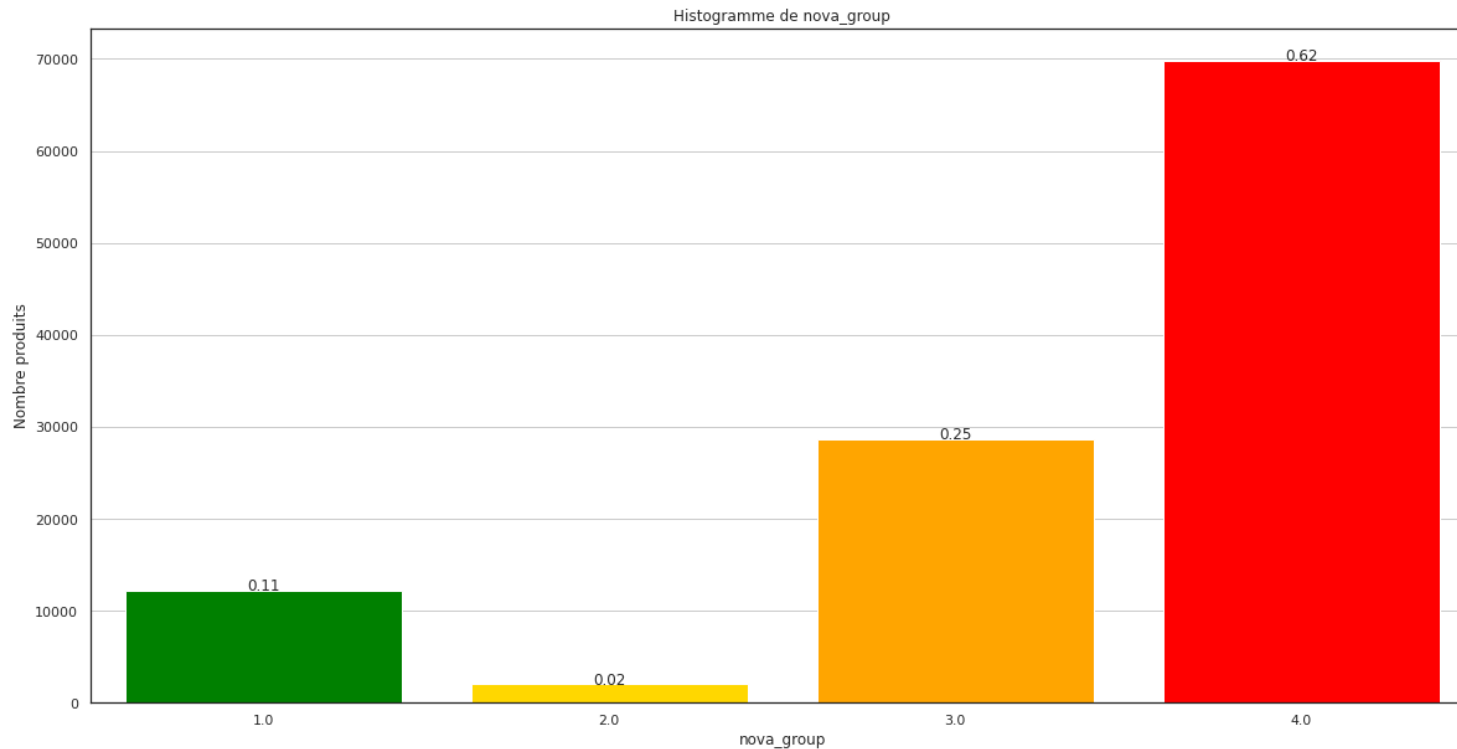


5 catégories de 'a' à 'e' avec 'd' la catégorie la plus fréquente (33 075)



3. EXPLORATION DES DONNEES

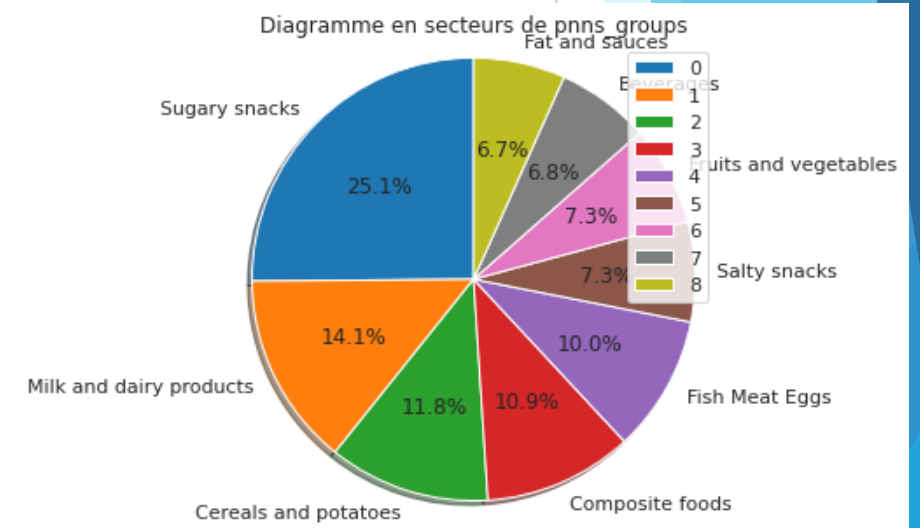
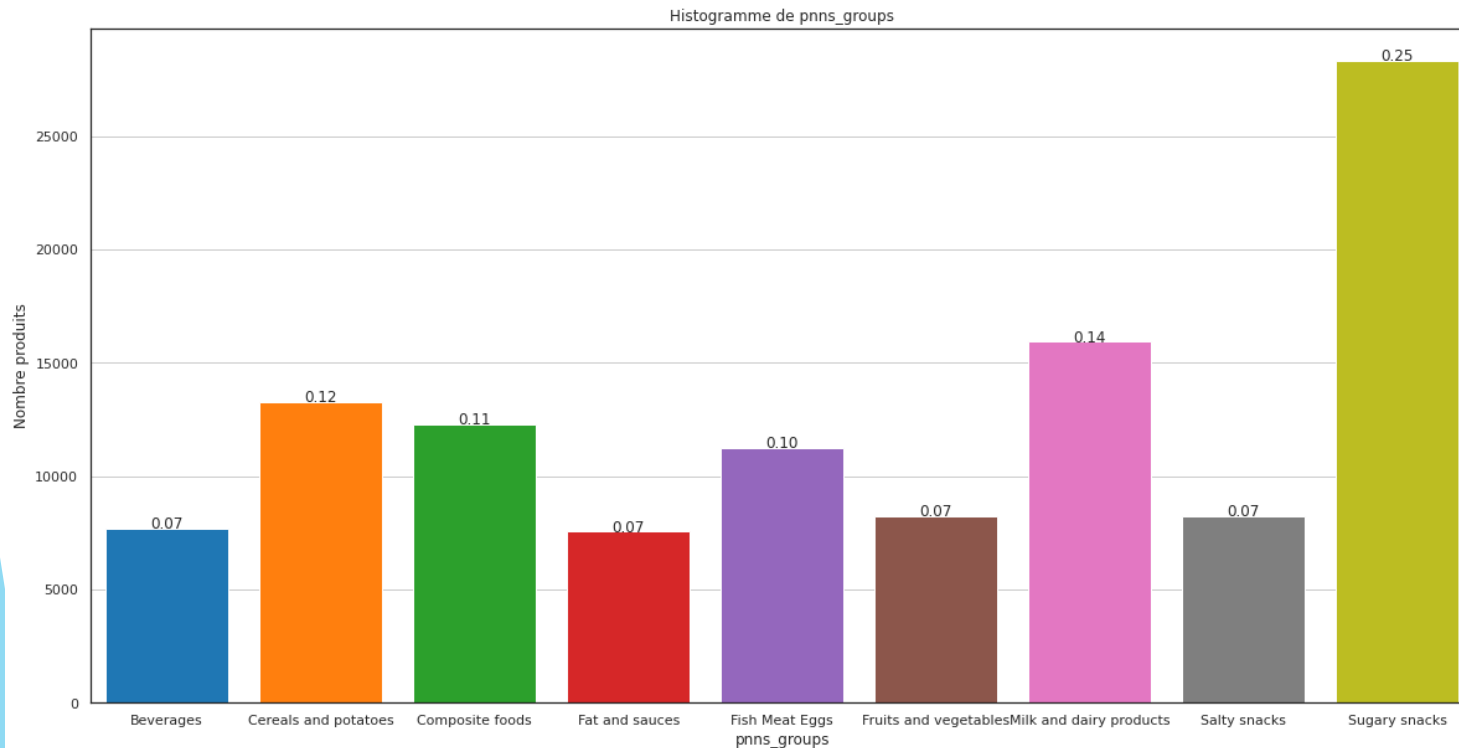
Analyse univariée de nova_group



4 catégories de '1.0' à '4.0' avec '4.0' la catégorie la plus fréquente (69 852)

3. EXPLORATION DES DONNEES

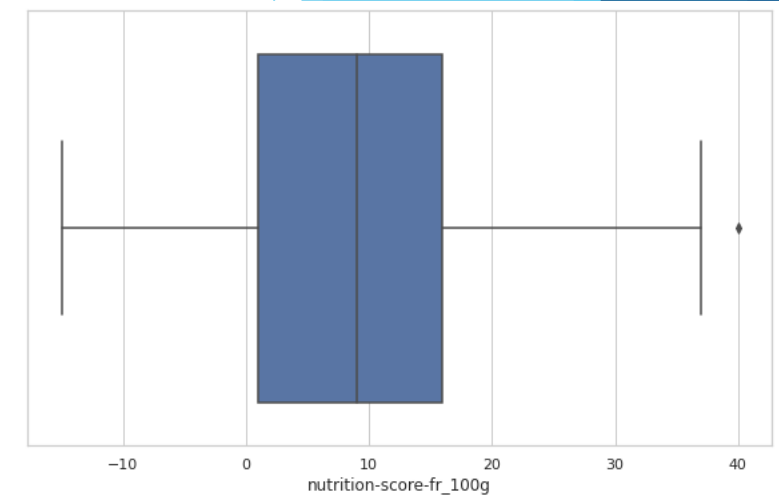
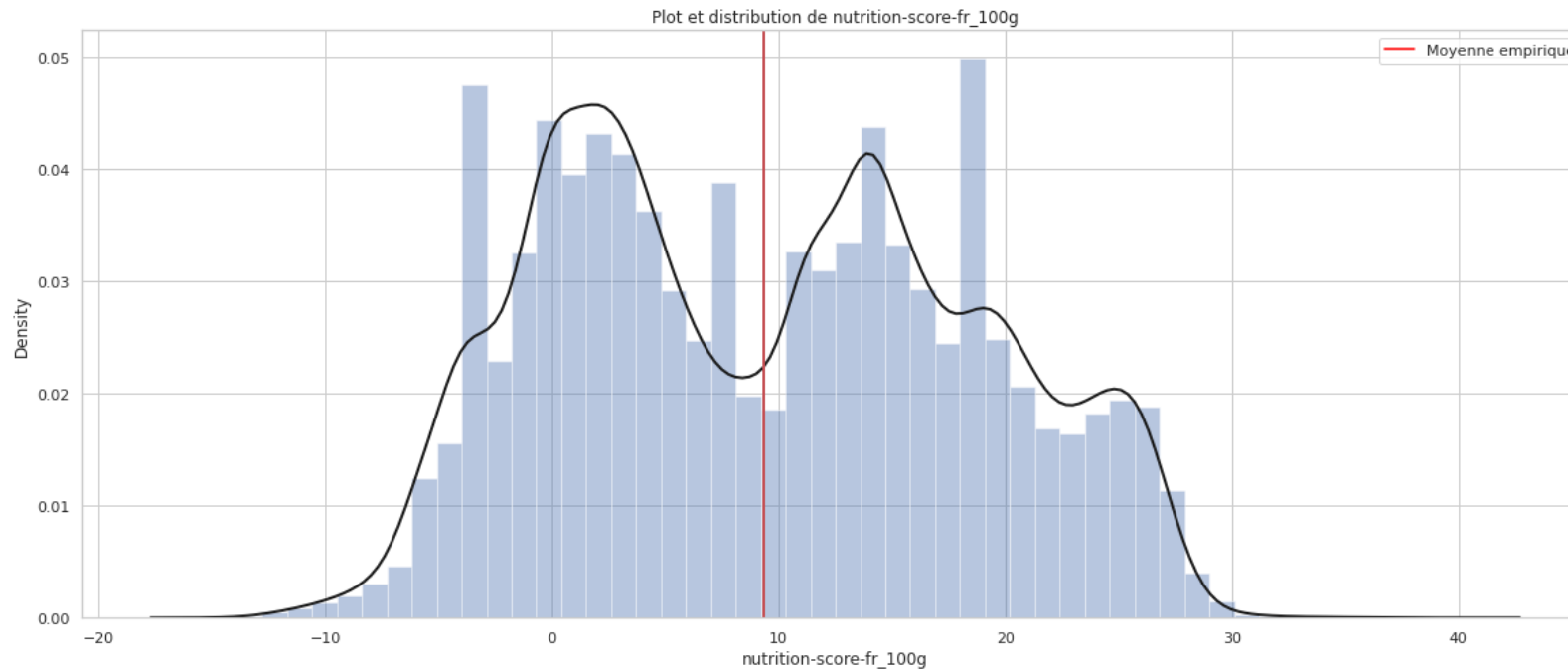
Analyse univariée de pnns_groups



9 catégories avec 'Sugary snacks' la catégorie la plus fréquente (28 344)

3. EXPLORATION DES DONNEES

Analyse univariée de nutrition_score_100g



Distribution bimodale

Moyenne = 9.34

Médiane : 9.0

Ecart-type : 9.26

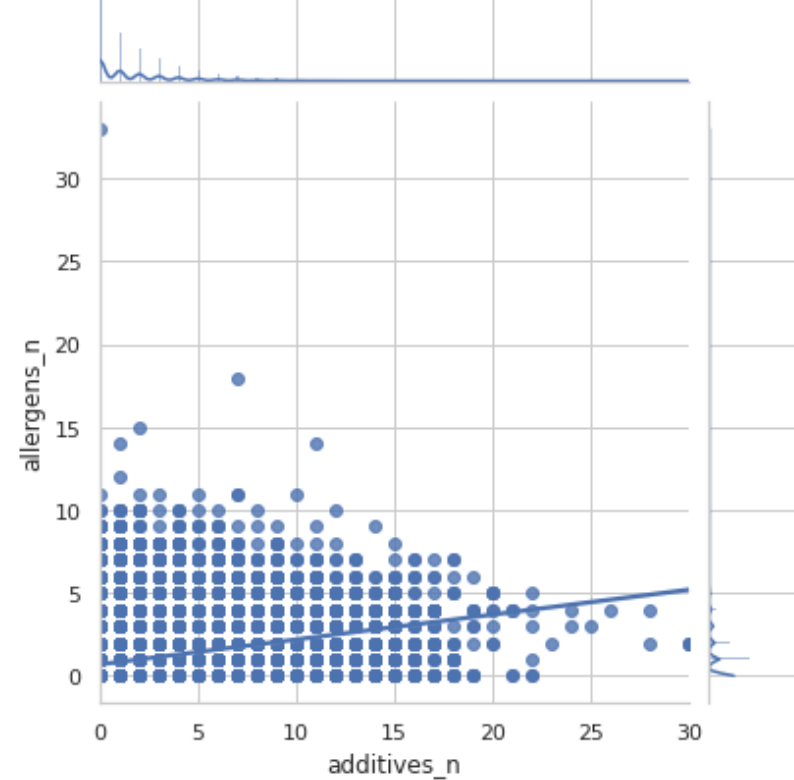
Skewness : 0.147

Kurtosis : -1.015

3. EXPLORATION DES DONNEES

Relation entre 'allergens_n' et 'additives_n'

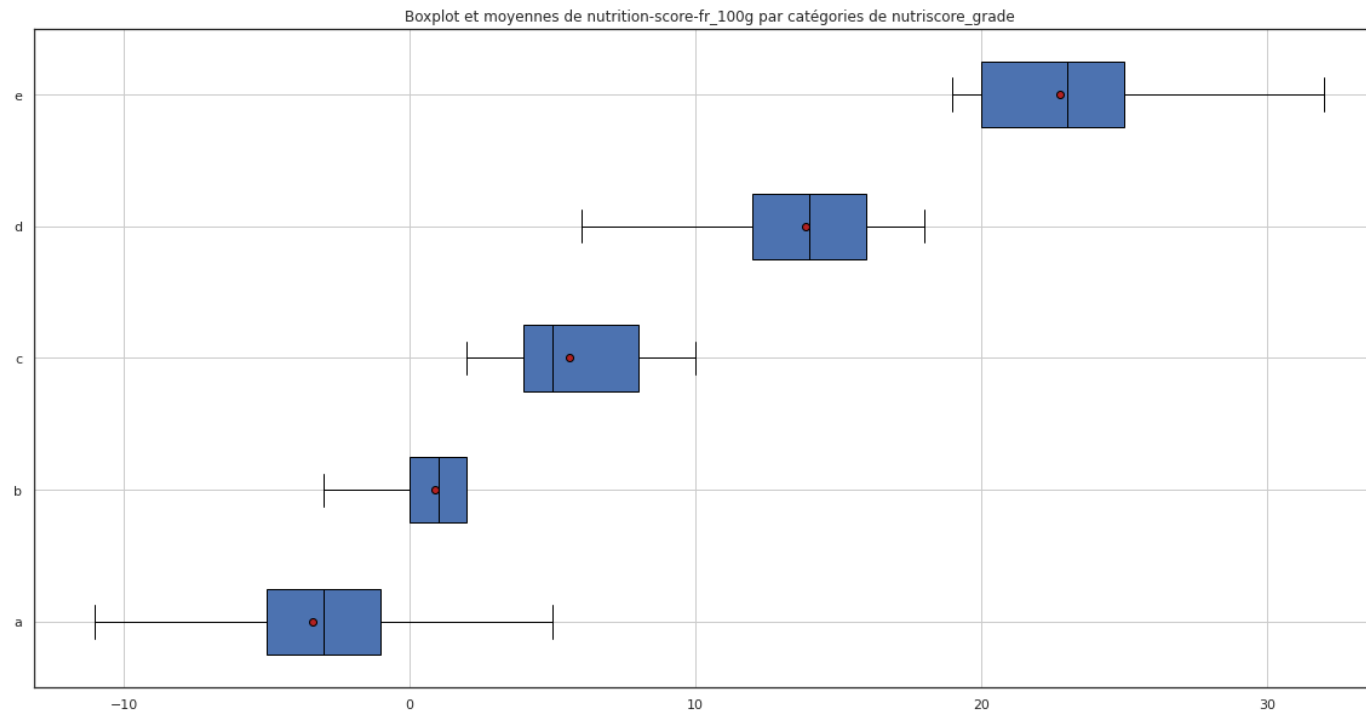
Diagramme de dispersion et droite de régression linéaire



Régression linéaire : corrélation de Pearson = 0.27

3. EXPLORATION DES DONNEES

Relation entre 'nutrition-score-fr_100g' et 'nutriscore_grade'

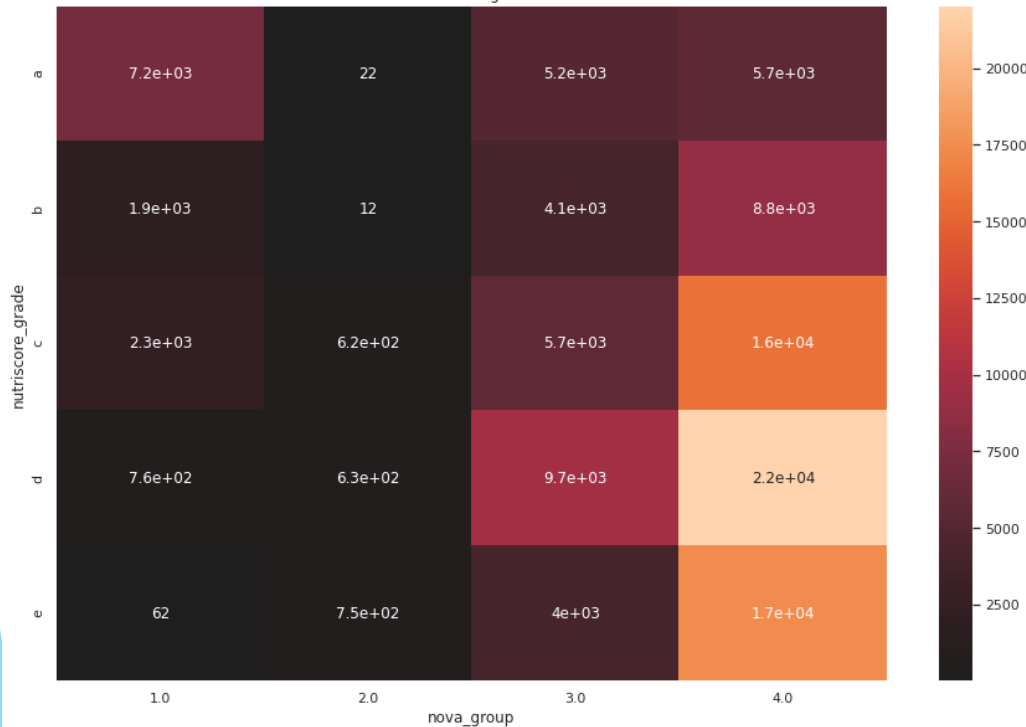


ANOVA : $n^2 = 0.93$

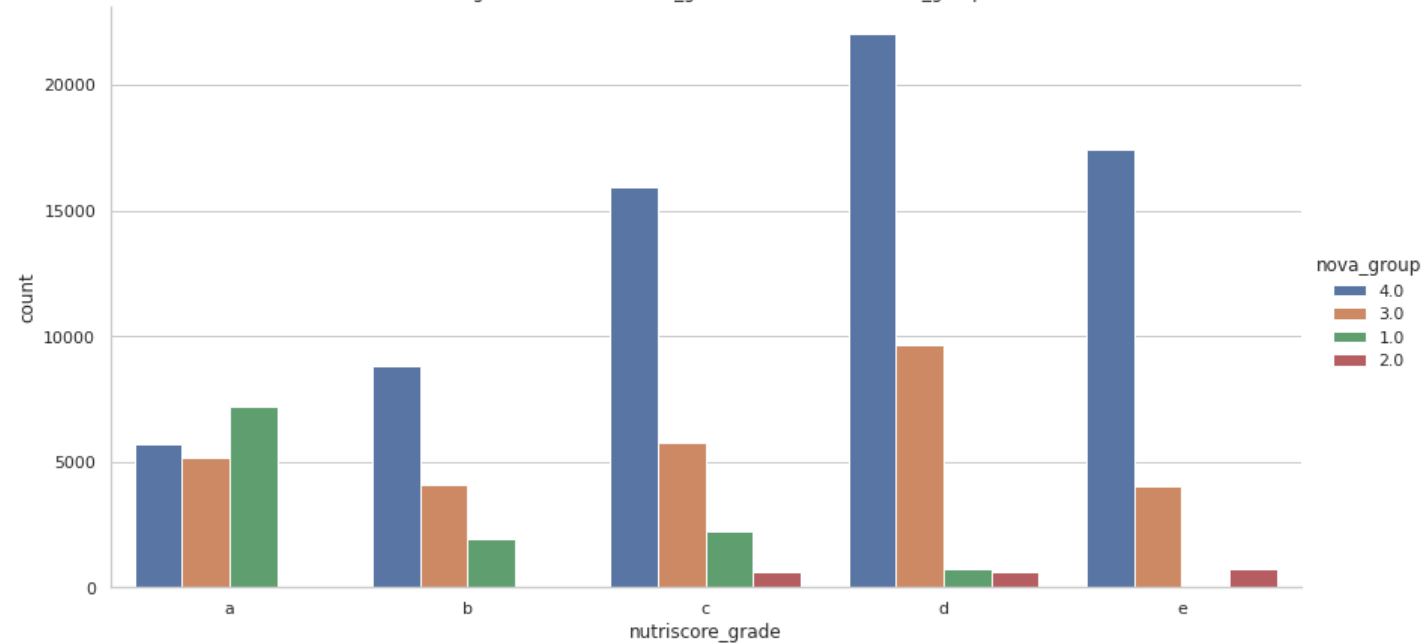
3. EXPLORATION DES DONNEES

Relation entre 'nutriscore_grade' et 'nova_group'

Tableau de contingence coloré



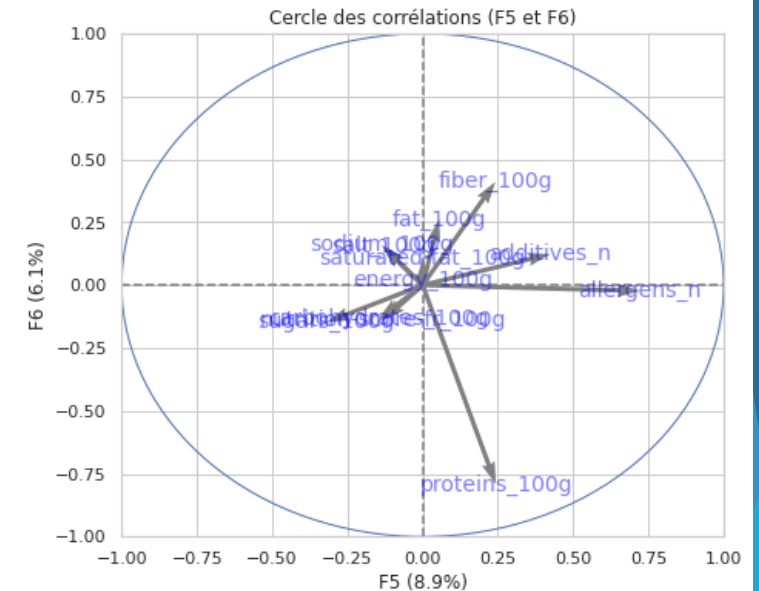
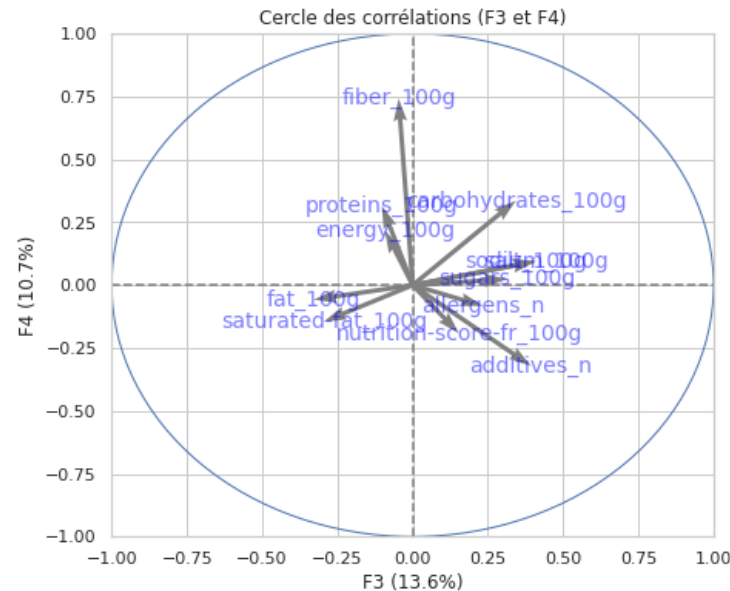
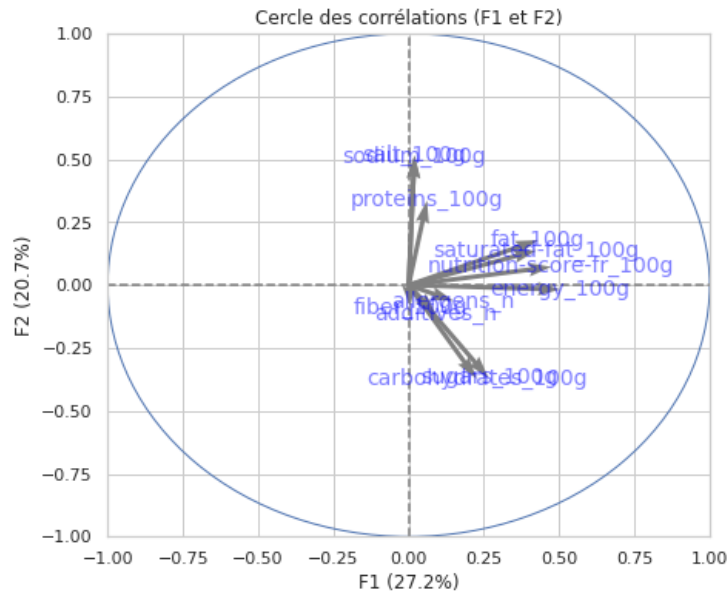
Histogramme de nutriscore_grade en fonction de nova_group



Test du Chi2 : forte corrélation entre la catégorie 'c', 'd' et 'e' du nutriscore et la catégorie '4.0' du nova group

3. EXPLORATION DES DONNEES

Analyse en composantes principales



F1 = énergie/nutriscore/graisse

F2 = sel/protéines VS sucres/glucides

F3 = additifs/glucides/sel

F4 = fibres

F5 = allergènes

F6 = protéines

Les 3 premières composantes expliquent presque 60% de la variance

Très difficile d'interpréter le 3^{ème} axe factoriel

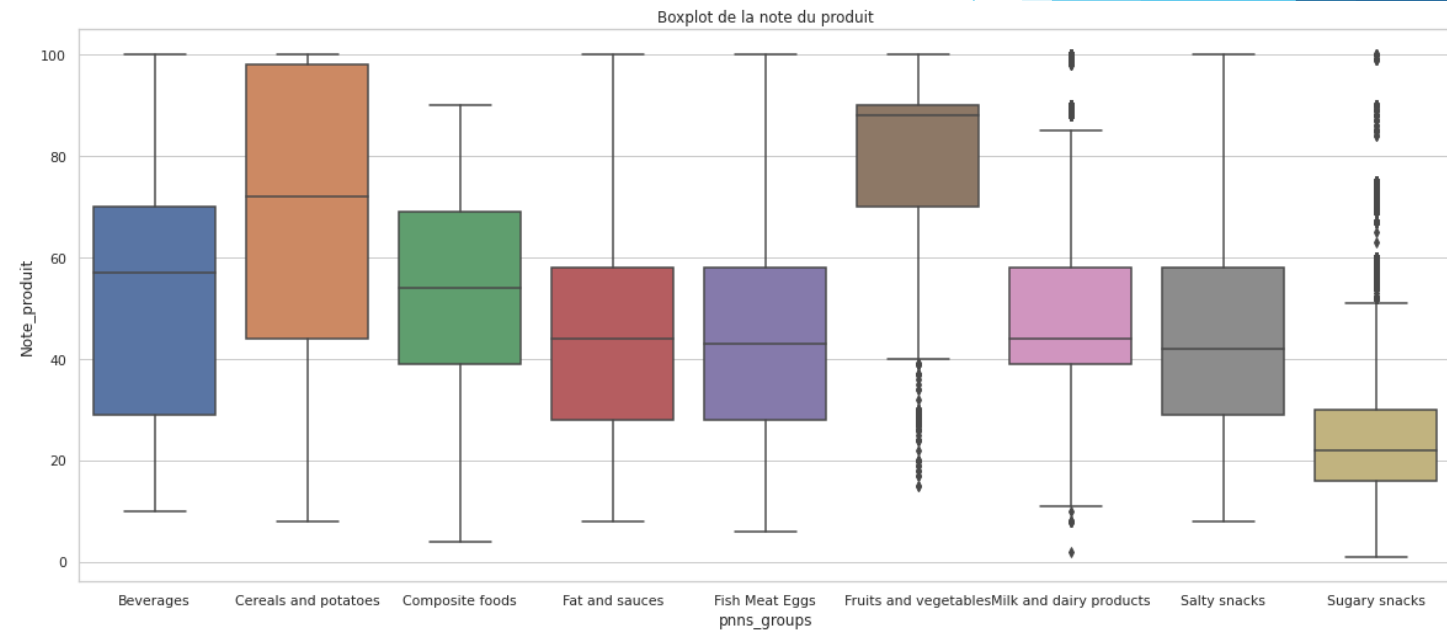
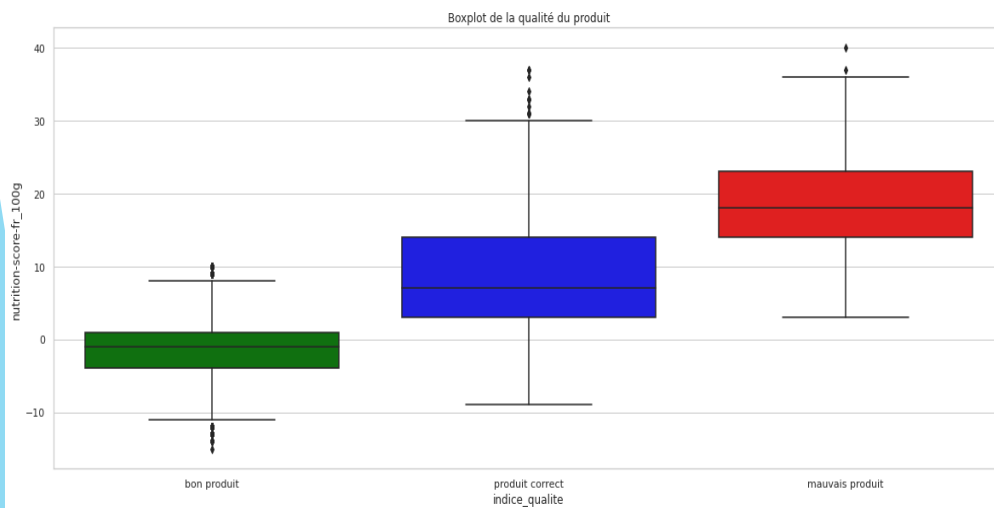
3. EXPLORATION DES DONNEES

Création d'une notation de la qualité des produits avec des seuils (comme le Nutriscore) et d'un indice de qualité :

< 30 = Mauvais produit

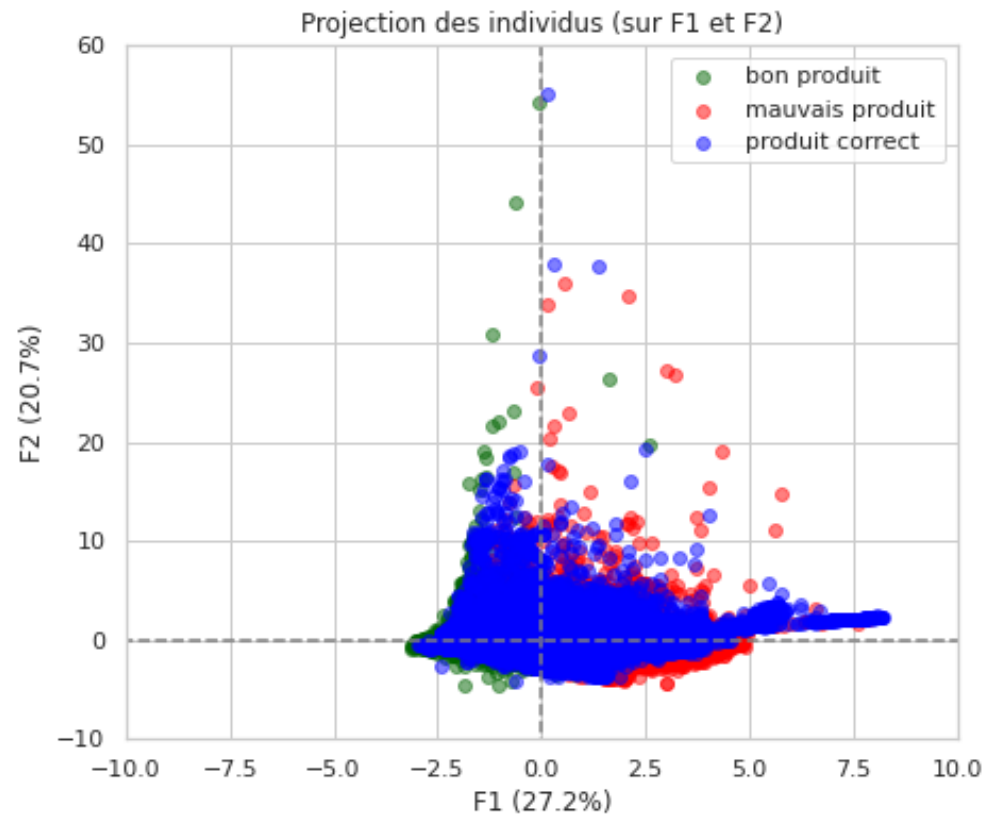
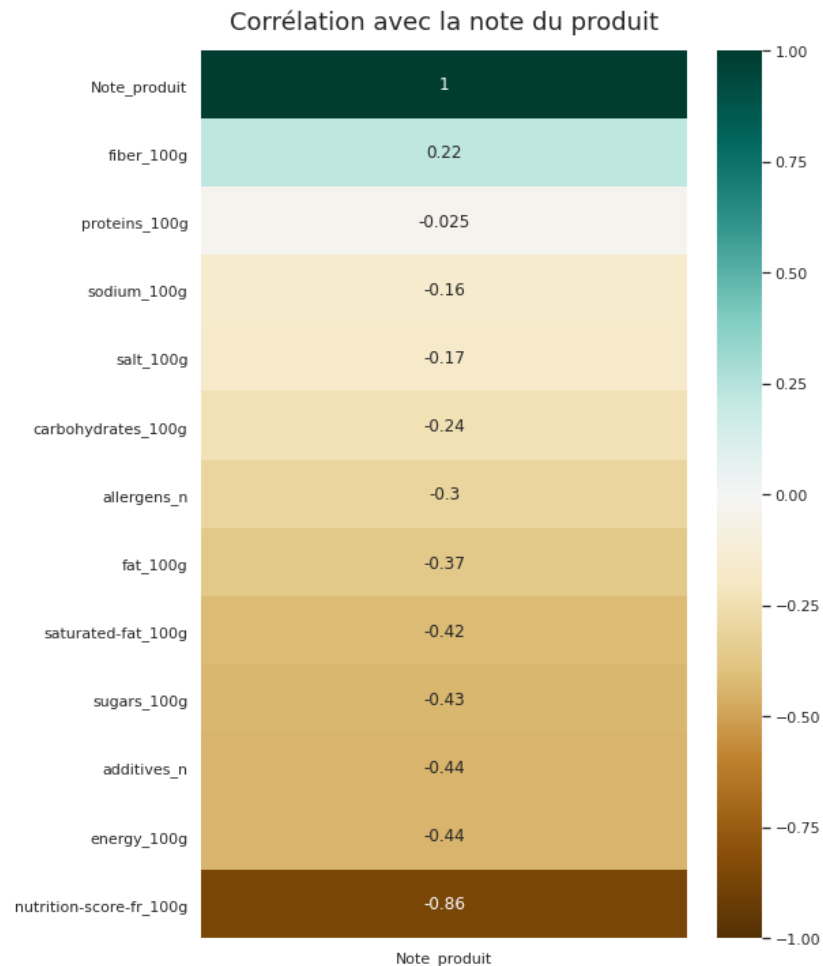
Entre 30 et 60 = Produit correct

> 60 = Bon produit



3. EXPLORATION DES DONNEES

Analyse des relations avec la nouvelle variable 'Note_produit'



CONCLUSIONS

A partir de la base de données, **nettoyage des données** réalisé :

- Sélection et choix des variables utiles, suppression des valeurs erronées et des valeurs manquantes, imputation par 0 et par les données d'autres colonnes

Etude exploratoire a montré que :

- + nutriscore important, + le nombre d'additifs et d'allergènes augmente avec un nutriscore grade mauvais qui tend vers 'e' (surtout pour les produits de 'Sugary snacks', 'Salty snacks', 'Fat and sauces' et 'Fish Meat Eggs' et les produits ultra-transformés)
- Les produits ultra-transformés ont tendance à avoir un nombre élevé d'additifs et à être très corrélés avec les produits 'Sugary snacks'

CONCLUSIONS

LIMITES ET SOLUTIONS

Application faisable mais principale limite :

- ❖ Peu de références disponibles (6% du jeu de données initial utilisé)

Retravailler les données lors du nettoyage (imputations) ou en obtenir des nouvelles (refaire des saisies) pour compléter la base de données

Retravailler la notation avec plus de seuils et essayer d'autres variables qui jouent également un rôle dans la qualité d'un produit (aspect écologique, ...)

MERCI DE VOTRE ATTENTION

QUESTIONS - REPONSES