Application au service de la santé publique

Parcours Data Scientist - *OPENCLASSROOMS*

Plan de la présentation

1. Introduction au sujet
    a. Présentation du jeu de données
    b. Application ‘EatHealthy’
2. Pré exploration du jeu de données
    a. Nettoyage des colonnes
    b. Nettoyage de lignes
    c. Analyse des données corrélées
3. Exploration du jeu de données
    a. Analyse univariée du nutrition score
    b. Analyse multivariée du nutrition score
    c. Analyse du nutrition grade
4. Conclusion pour l’application

Introduction au sujet

Présentation du jeu de données

Présentation du jeu de données

- Open Food Facts est une base de données gratuite et collaborative
- Le CSV actuel contient 1094562 lignes et 178 colonnes (2.1 Go à peu près)
- L'ensemble des colonnes (doc) contient des données sur l'origine, le pourcentage par ingrédient, les labels des ingrédients, le code barre, l'image …
- Parmi les données : le nutriscore (de différents pays), le nutriscore grade, le nova group
    - C'est ce qui a donnée l'idée de l'application 'EatHealthy'

Introduction au sujet

Application ‘EatHealthy’

Application ‘EatHealthy’

- Savoir si un repas est healthy et si possible noter le niveau de ‘healthiness’
- EatHealthy est
  - Libre
  - Gratuite
  - Sans contraintes (pas de compte utilisateur)
- Récupération des ingrédients d’un site de recette par exemple => API Marmiton
- Utilisation de la liste des ingrédients pour calculer un score sur la qualité d’un repas en utilisant OpenFoodFacts

Application ‘EatHealthy’ : Problématiques

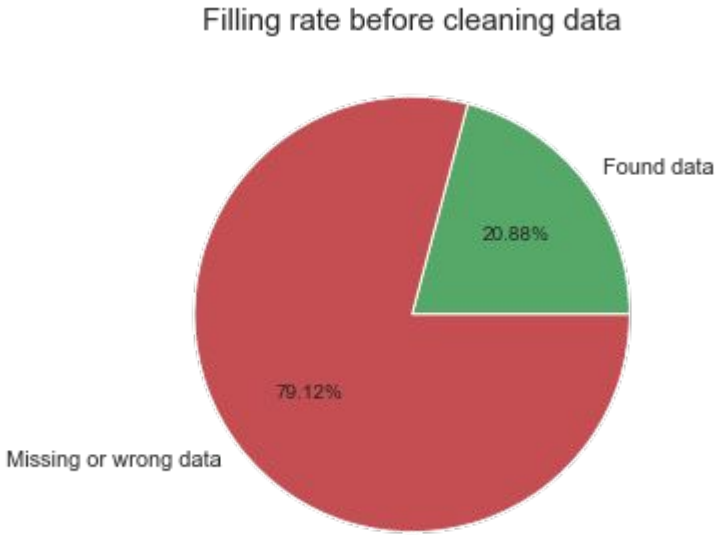
- Qu’est ce qui influe sur la qualité d’un repas
- Quelles données sont utilisables sur le site Open Food Facts
- Quelles données sont fiables sur le site
- Trouver une relation entre les ingrédients qu’on a (fiables et bien remplis) et la qualité d’un repas

Pré exploration du jeu de données

Nettoyage des colonnes

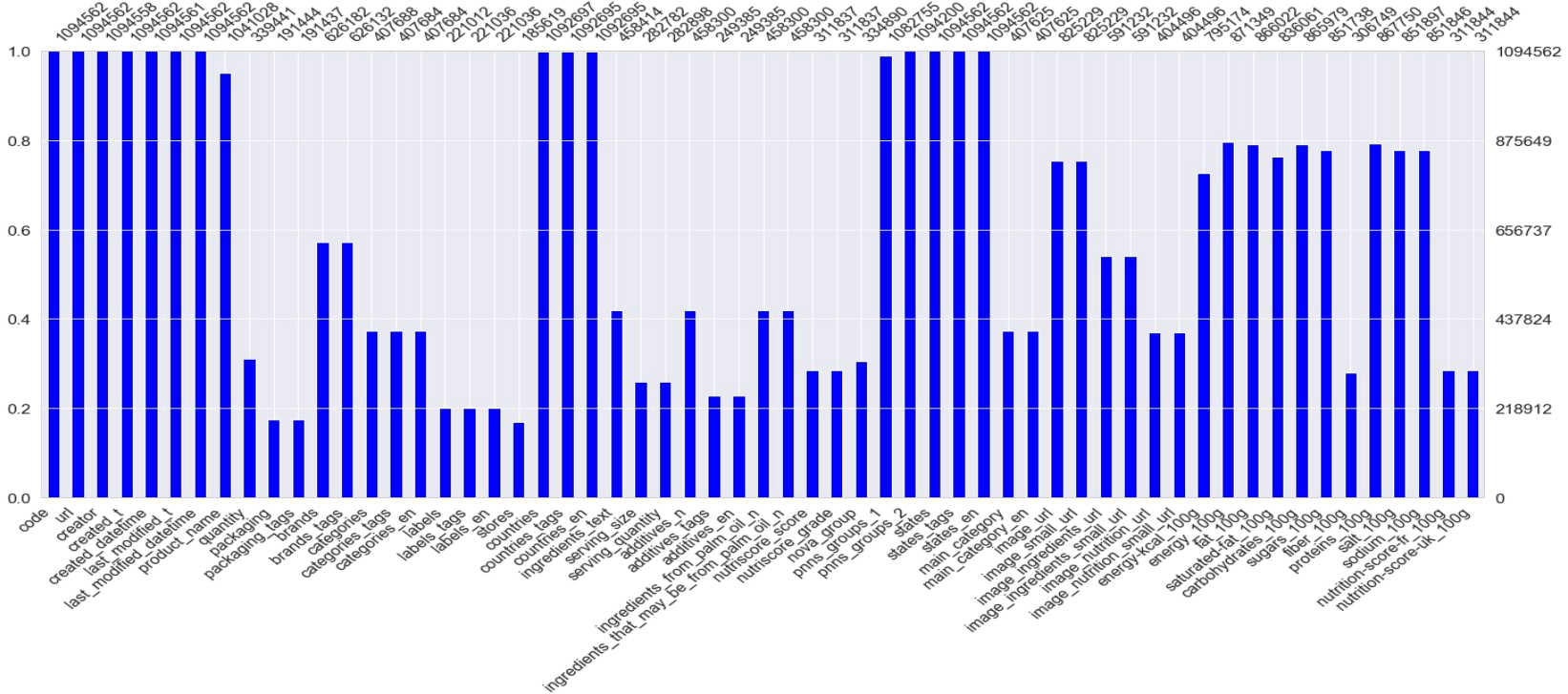
Nettoyage des colonnes

- Première analyse du taux de remplissage du CSV de Open Food facts
- Pour rappel il existe (1094562 lignes * 178 colonnes) données



Nettoyage des colonnes

- Observation des colonnes qui sont remplies à au moins 15 %



Nettoyage des colonnes

- On remarque qu'il reste 59 colonnes à présent :
  - Colonne identifiante : 'code'
  - Colonnes de métadonnées : creator, url, image, image_url, …
  - Colonnes d'ingrédients : 'salt_100g', 'fat_100g'
  - Colonnes de qualités nutritionnelle : 'nutriscore_score', 'nutriscore_grade' ...

- Les colonnes intéressantes pour notre application sont après ce premier nettoyage (16 colonnes)
  - code, product_name,
  - nutriscore_score, nutriscore_grade, nutrition-score-fr_100g, nutrition-score-uk_100g
  - energy-kcal_100g, energy_100g,
  - fat_100g, saturated-fat_100g, carbohydrates_100g, sugars_100g, fiber_100g, proteins_100g, salt_100g, sodium_100g,

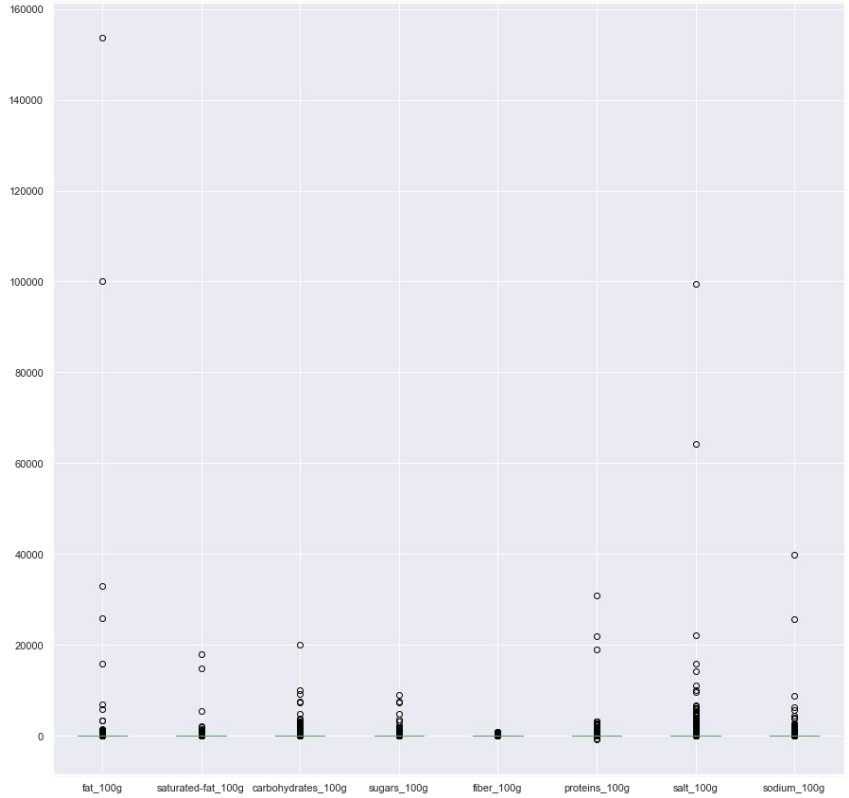
Pré exploration du jeu de données

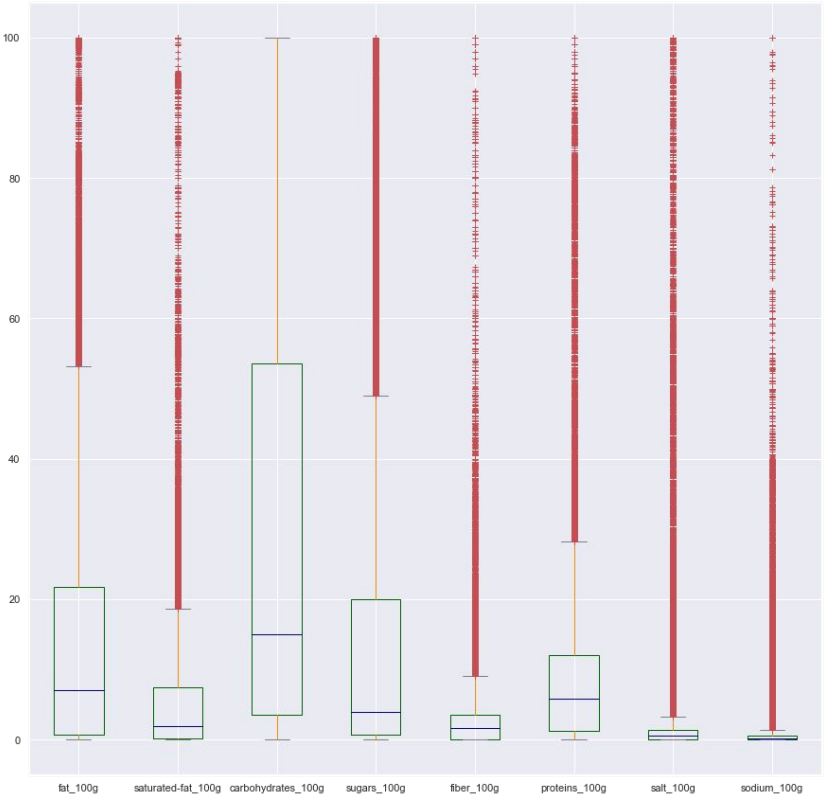
Nettoyage des lignes

Nettoyage des lignes

- Nettoyage des lignes:
  - Dupliquées (376) en utilisant le code
  - Avec que des valeurs manquantes ~ 130000
- Nettoyage des données aberrantes:
  - Les données nutritionnelles en 100g
  - Les données d’énergie (au maximum 1000kcal par 100g soit à peu près 4200 Kilo Joules)

Nettoyage des lignes



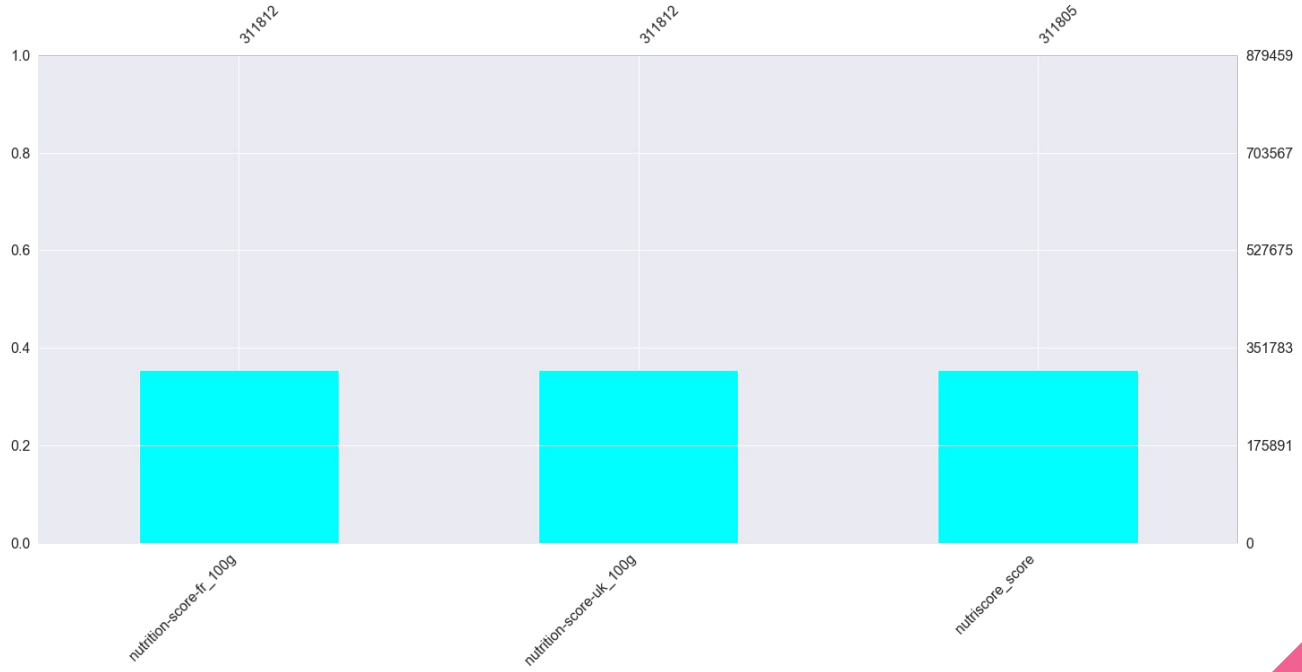


Pré exploration du jeu de données

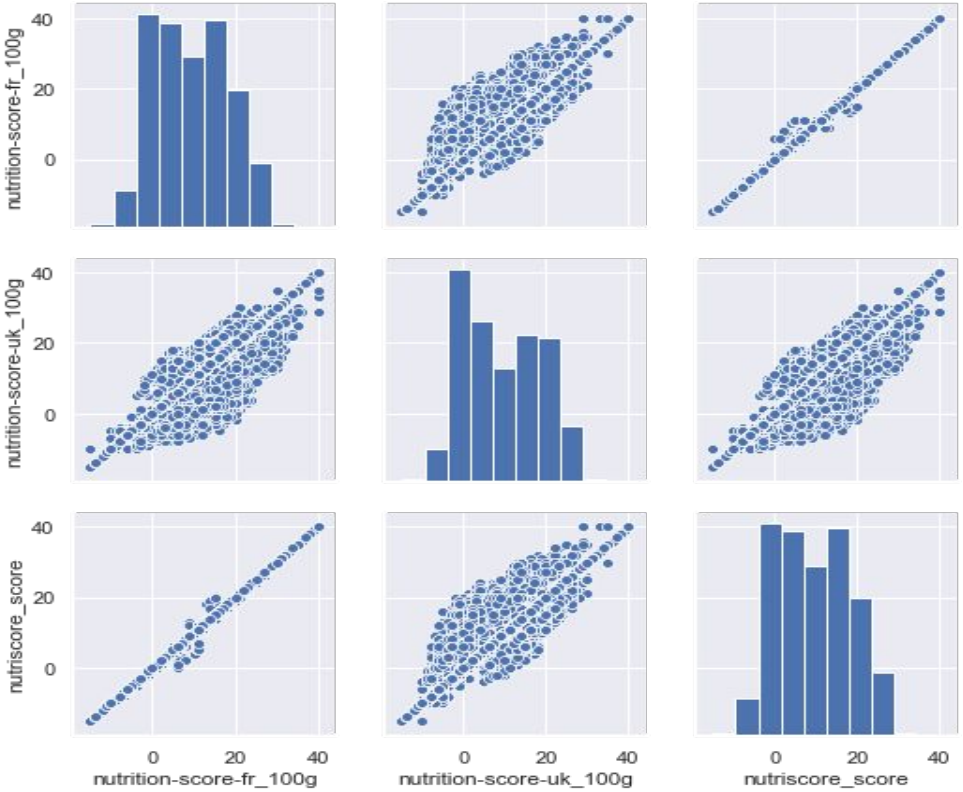
Analyse des données corrélées

Analyse des données corrélées

- 3 colonnes de nutriscore, on va garder juste le nutrition-score-fr_100g



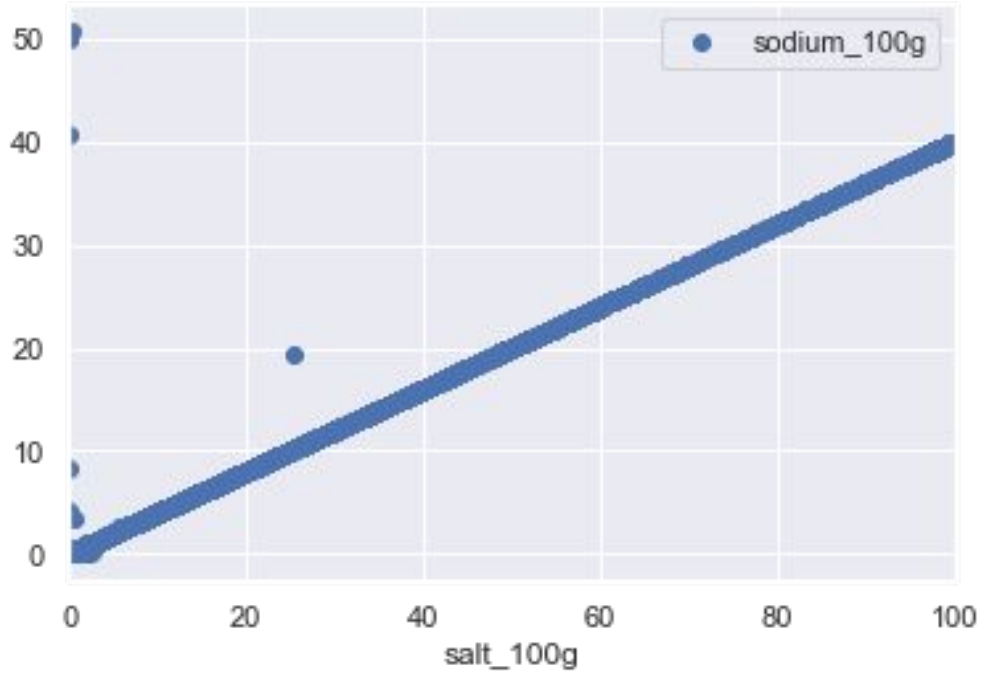
Analyse des données corrélées



Analyse des données corrélées

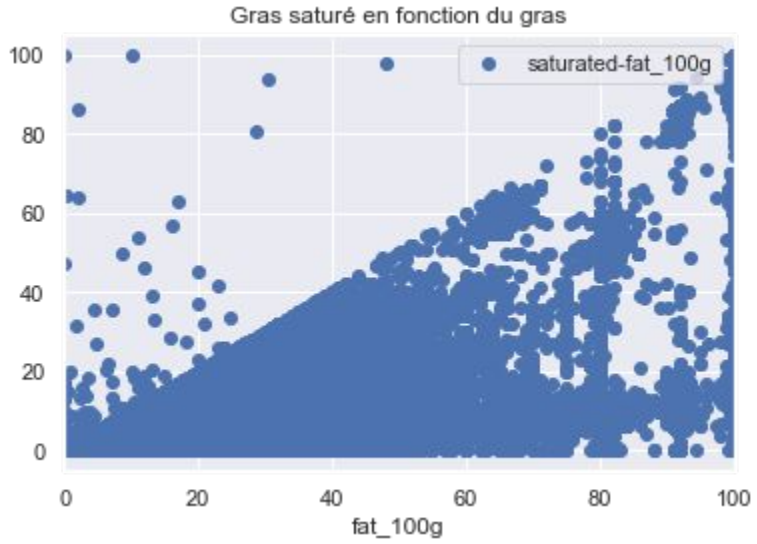
- Le sodium et le sel devraient représenter la même chose

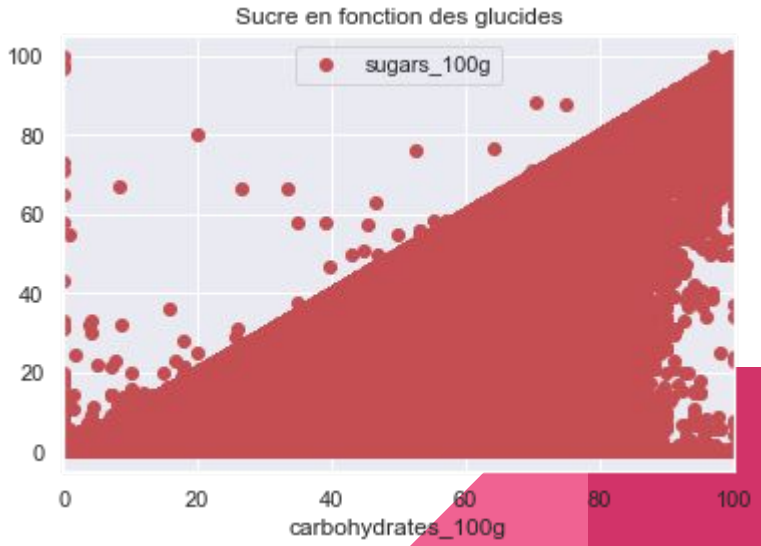
Analyse des données corrélées



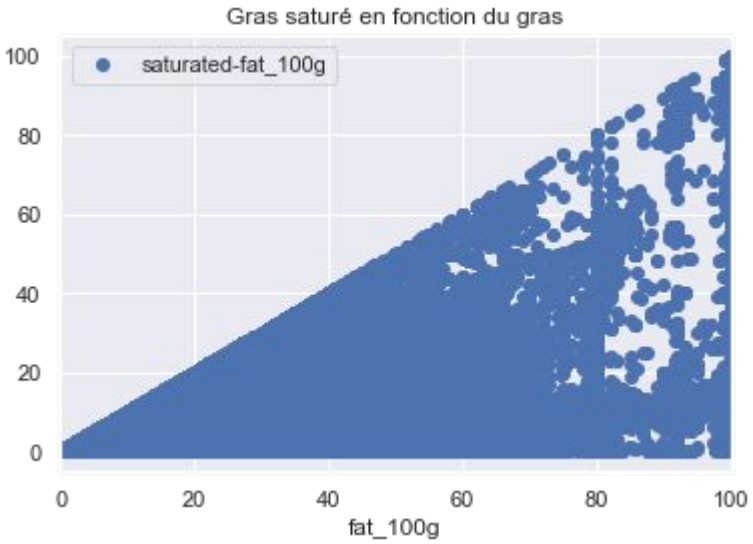
Analyse des données corrélées

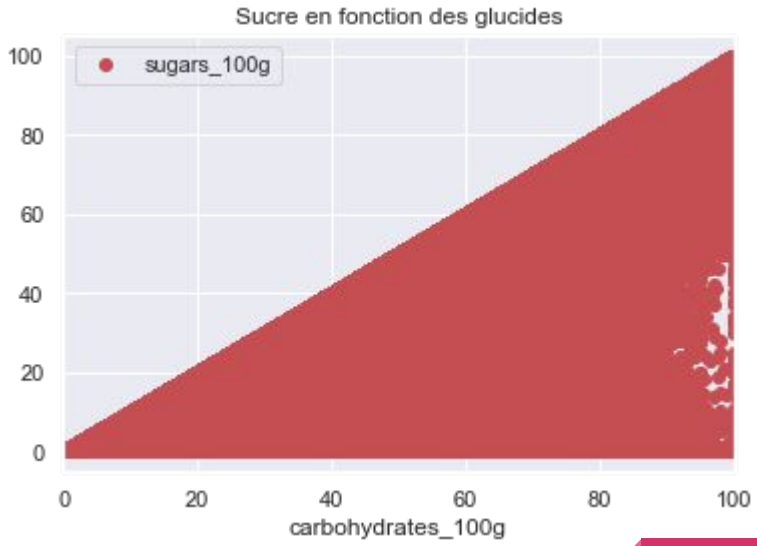
- Les acides gras saturés sont incluses dans les matières grasses, idem pour le sucre et les carbohydrates on va donc séparer ces données





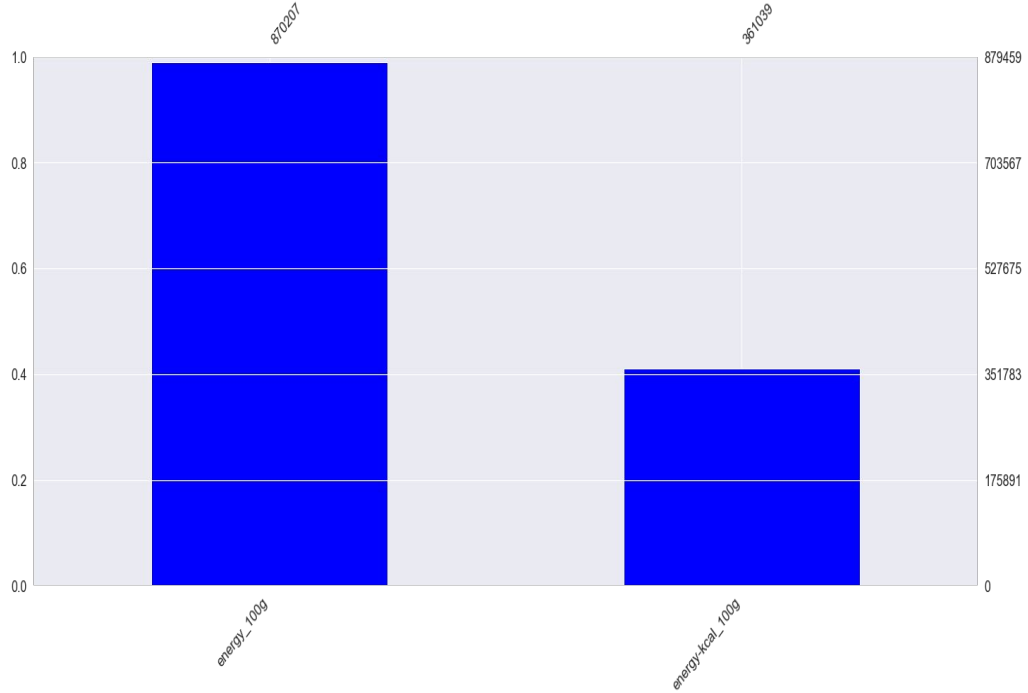
Analyse des données corrélées

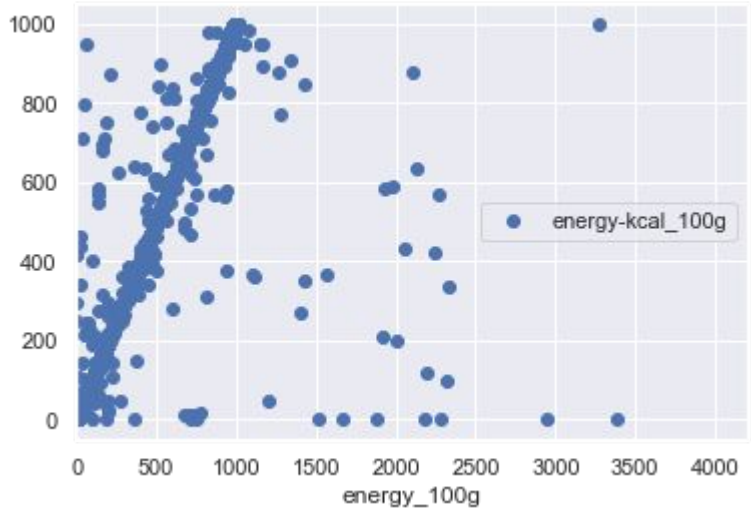




Analyse des données corrélées

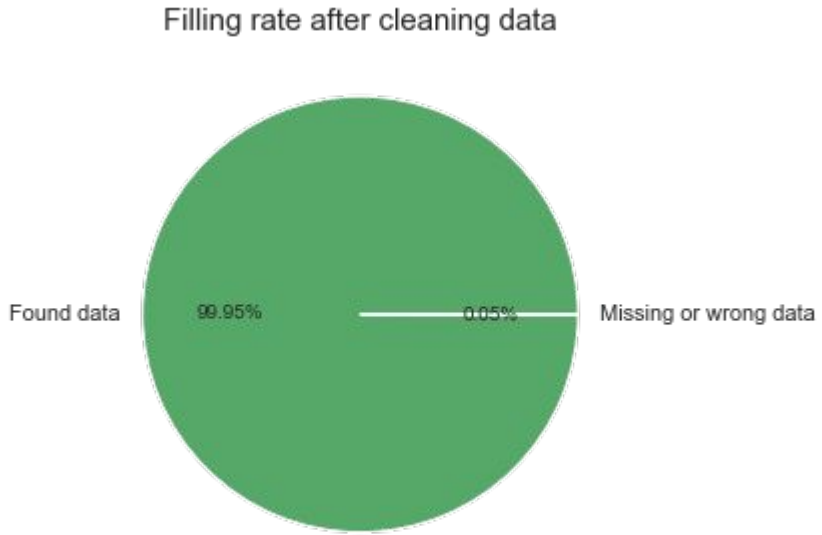
- L'énergie est exprimée en double aussi





Analyse des données corrélées

- On s’intéresse surtout au nutriscore, Donc on enlève les lignes pour lesquelles il n’y a pas de nutriscore
  - Lignes gardées = 302131
- Pour les quantités en 100g => rescaling en %
  - Colonnes gardées 12

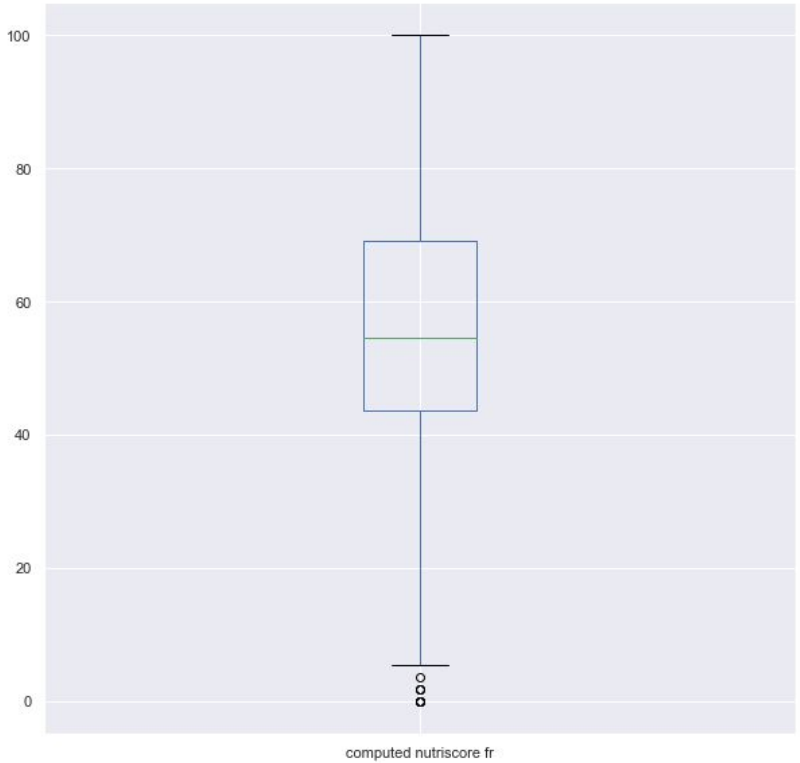


Exploration du jeu de données

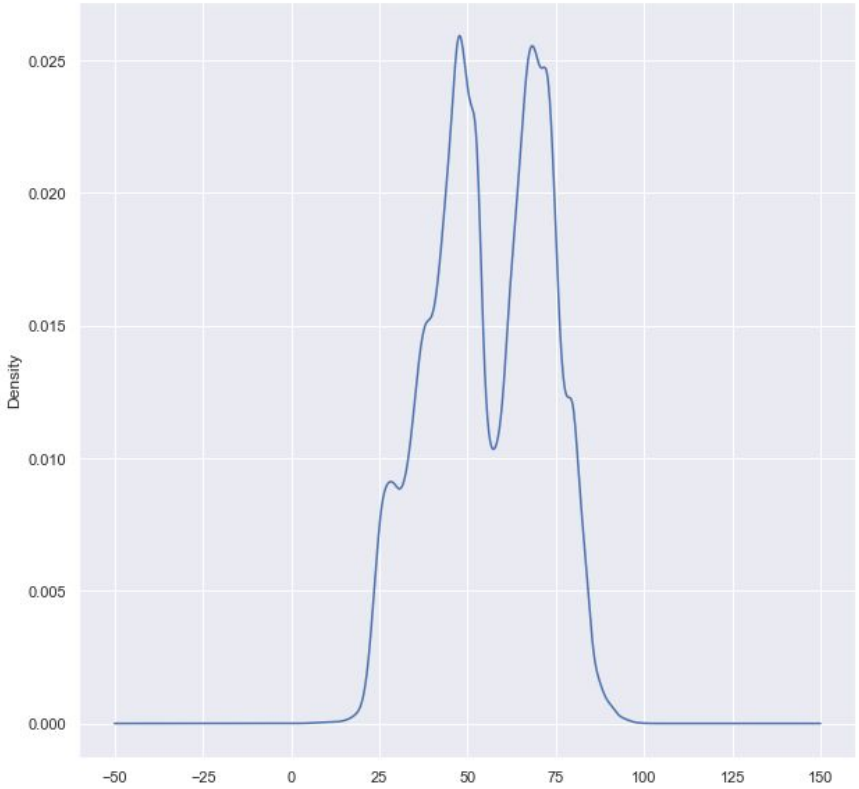
Analyse univariée du nutrition score

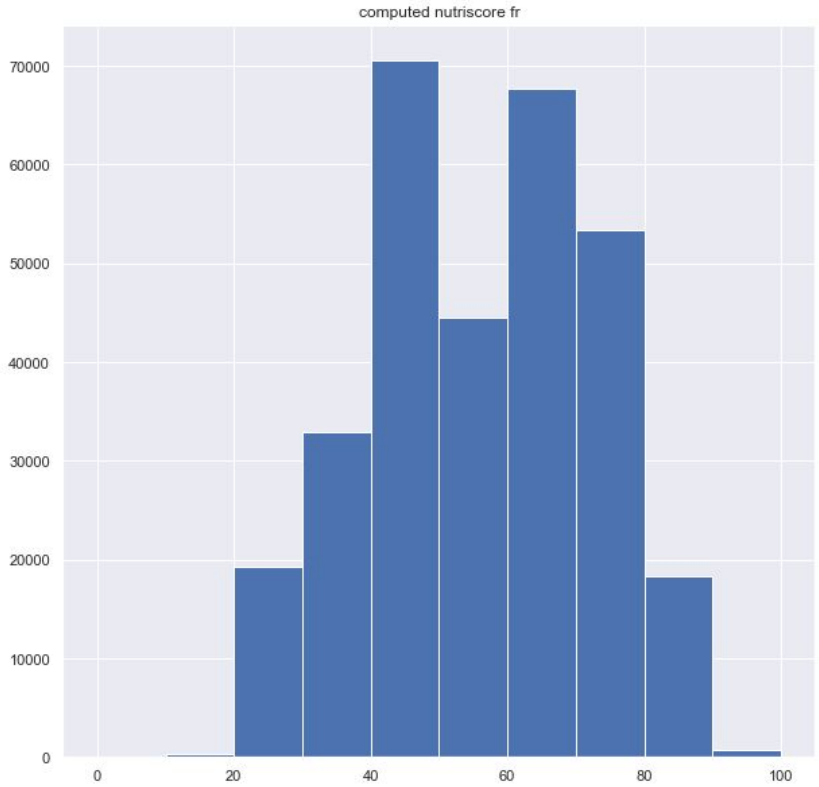
Analyse univariée du nutrition score

- Le nutriscore varie entre -15 et 40
- Pour nos utilisateurs on va créer un computed nutriscore qui donne un score sur 100 (un peu plus intuitif).



Analyse univariée du nutrition score





Analyse univariée du nutrition score

- Test de normalité pour le nutrition score:
  - H0 les valeurs du ‘computed nutriscore’ pour l’échantillon suivent une loi gaussienne
  - H1 les valeurs du ‘computed nutriscore’ pour l’échantillon ne suivent pas une loi gaussienne
- Test de Sharpiro-Wilk :
  - p = 0 , mais pas pertinent pour les échantillons de plus de 5000
- Test de Kolmogorov Smirnov:
  - p = 0
- Non normalité de la distribution

Exploration du jeu de données

Analyse multivariée du nutrition score

Analyse multivariée du nutrition score

- Séparation en training test et testing set (80%-20%)
- 1er Test en utilisant une régression linéaire :
  - Paramètres : ‘saturated-fat_100g', 'sugars_100g', 'fiber_100g', 'proteins_100g', 'sodium_100g', 'carbs_no_sugar_100g.
  - $R^2 = 0.39$
  - Les résultats ne sont pas intéressants.

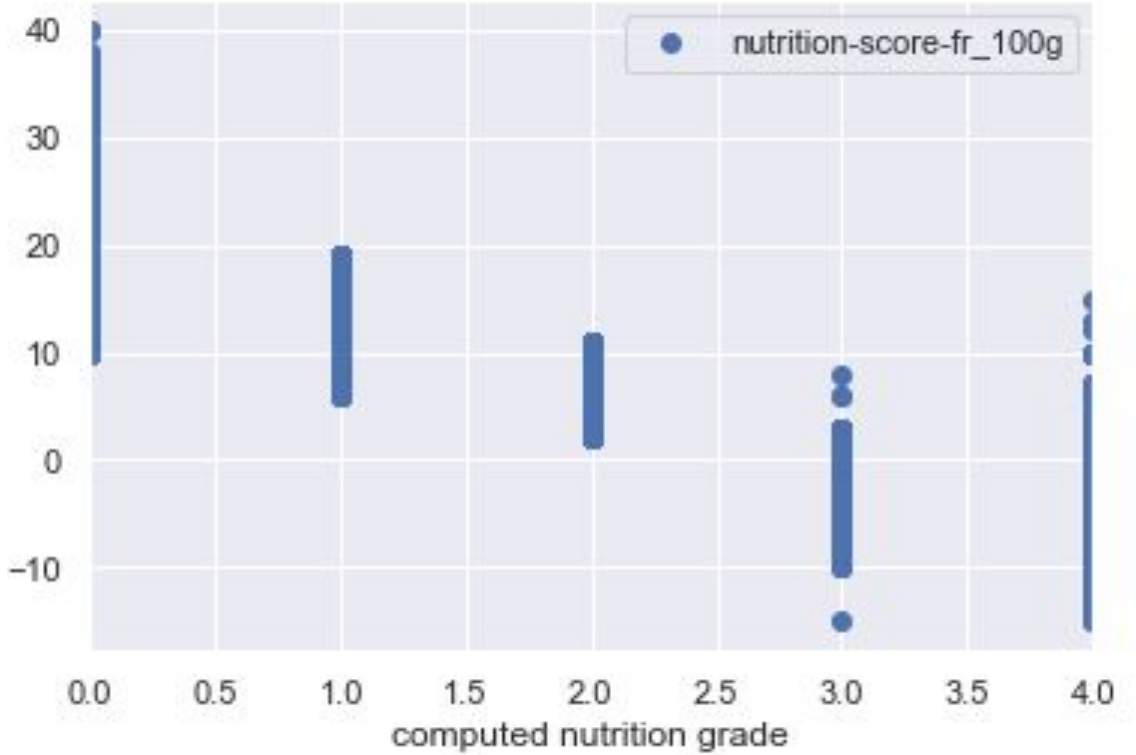
Analyse multivariée du nutrition score

- 2ème test en ajoutant l'énergie
  - R^2= 0.55
  - On obtient les coefficients suivants pour notre nutriscore recalculé :
    - Coefficient saturated-fat_100g = -0.38
    - Coefficient sugars_100g = -0.11
    - Coefficient fiber_100g = 0.50
    - Coefficient proteins_100g = 0.04
    - Coefficient sodium_100g = -0.17
    - Coefficient carbs_no_sugar_100g = 0.07
    - Coefficient fat_non_satu_100g = 0.04
    - Coefficient energy_100g = -0.01

Exploration du jeu de données

Analyse du nutrition grade

Analyse du nutrition grade

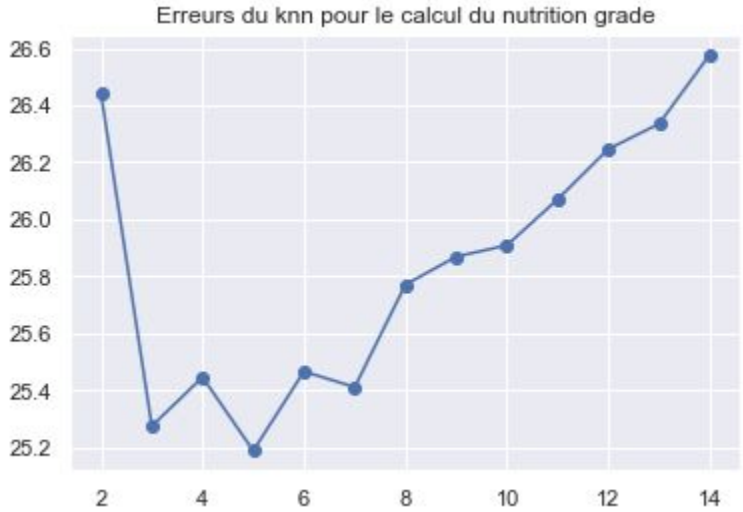


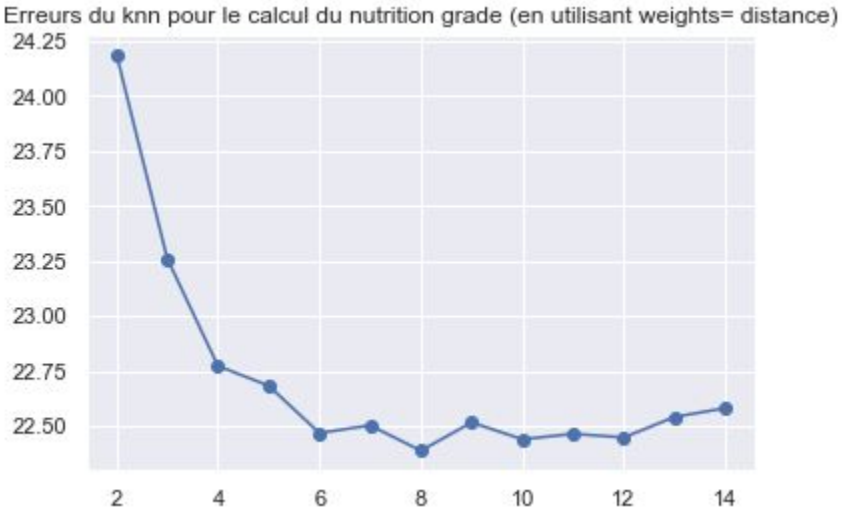
Analyse du nutrition grade

- Test du Chi -2 :
  - H0 les variables nutrition grade / score sont indépendantes, vs H1, elles ne le sont pas
  - P Value = 0, on rejette H0
- Corrélation de Pearson entre nutrition grade (converti) et le computed nutri score créé, est de 94%

Analyse du nutrition grade

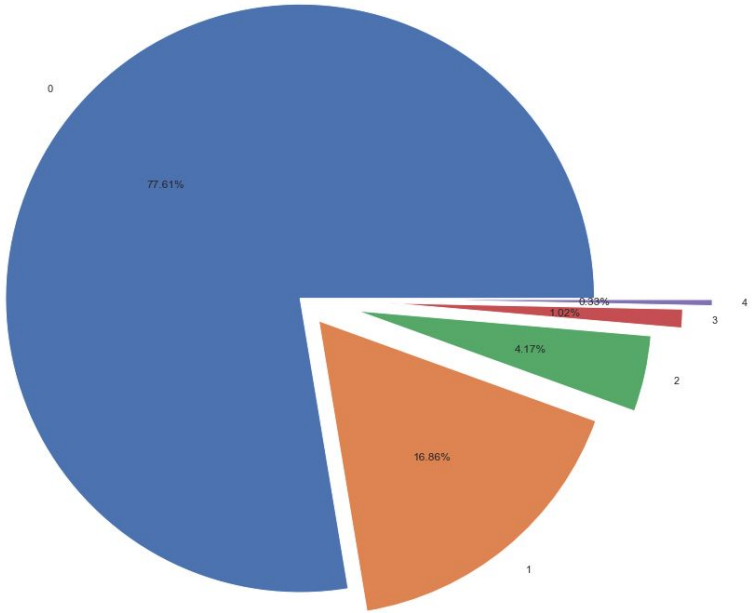
- On va utiliser le knn pour essayer d'imputer un nutrition grade
- Premier test avec 3 voisins => 25% d'erreurs



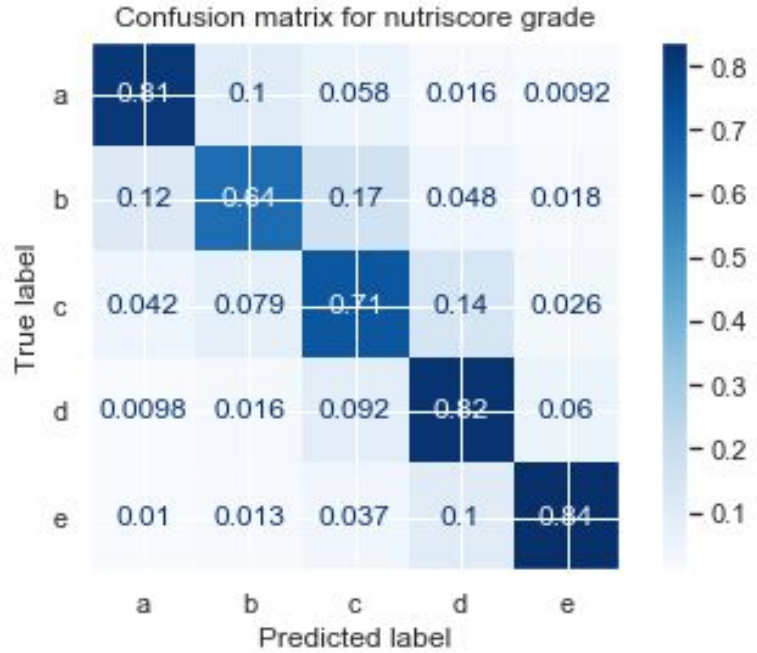


Analyse du nutrition grade

- Analyse des erreurs



Analyse du nutrition grade



Conclusion pour l’application

Conclusion pour l’application

- Impossibilité d’utiliser le nutrition score
  - Problème sur l'énergie
  - Problème sur le pourcentage de fruis / noix …
  - Manque de catégorisation liquide / solide
  - …
- Possibilité d’utiliser le nutrition grade:
  - 22% d’erreurs donc à titre indicatif seulement
  - Calcul à partir des ingrédients formant notre menu

Conclusion pour l’application

- Améliorations possibles
  - Utiliser des produits de même catégorie pour imputer les valeurs manquantes
  - Utiliser des produits similaires

Questions / Réponses