Projet 3 OC :

Concevez une application au service de la santé publique

Projet 3 OC :

Concevez une application au service de la santé publique

- Agence "Santé publique France"
- **Appel à projets**
- **Idées innovantes d’applications en lien avec l'alimentation.**

Projet 3 OC :

Concevez une application au service de la santé publique

Livrables

- Un **notebook du nettoyage** et un **notebook d'exploration** réunis dans un seul fichier .ipynb
- Une **présentation**

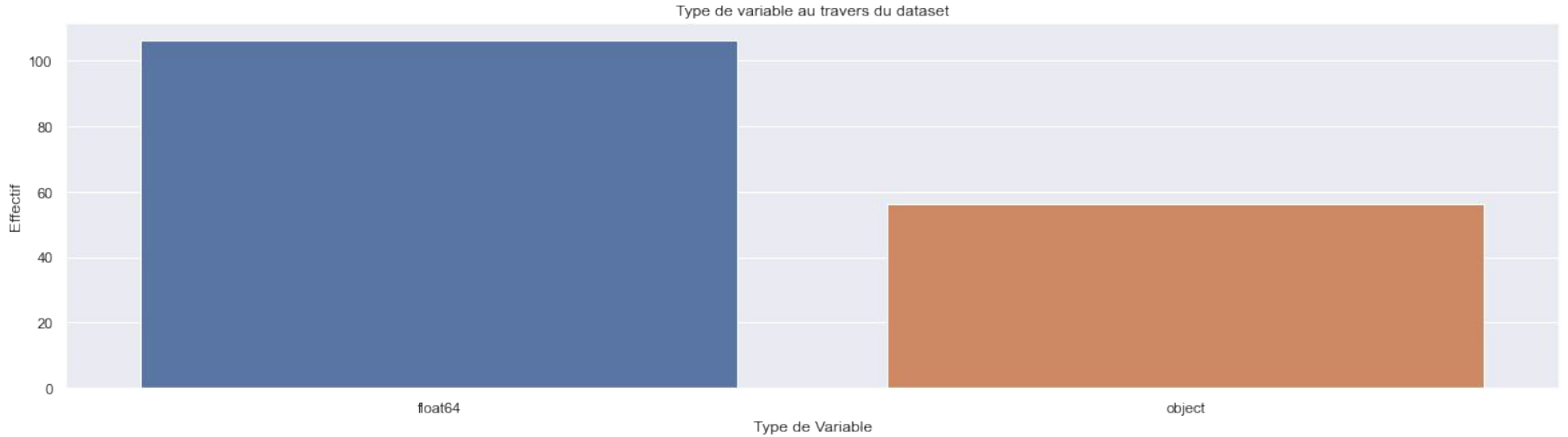
Projet 3 OC :

Concevez une application au service de la santé publique

- Mise en place
- La donnée brute
- Idée d’application
- Les variables pertinentes
- Nettoyage du dataset
- Analyse univariée
- Analyse multivariée
- ACP
- k-NN
- Conclusion

Les données brutes

- Les informations générales sur la fiche du produit : nom, date de modification
- Un ensemble de tags : catégorie du produit, localisation, origine,
- Les ingrédients composant les produits et leurs additifs éventuels
- Des informations nutritionnelles : quantité en grammes d'un nutriment pour 100 grammes du produit



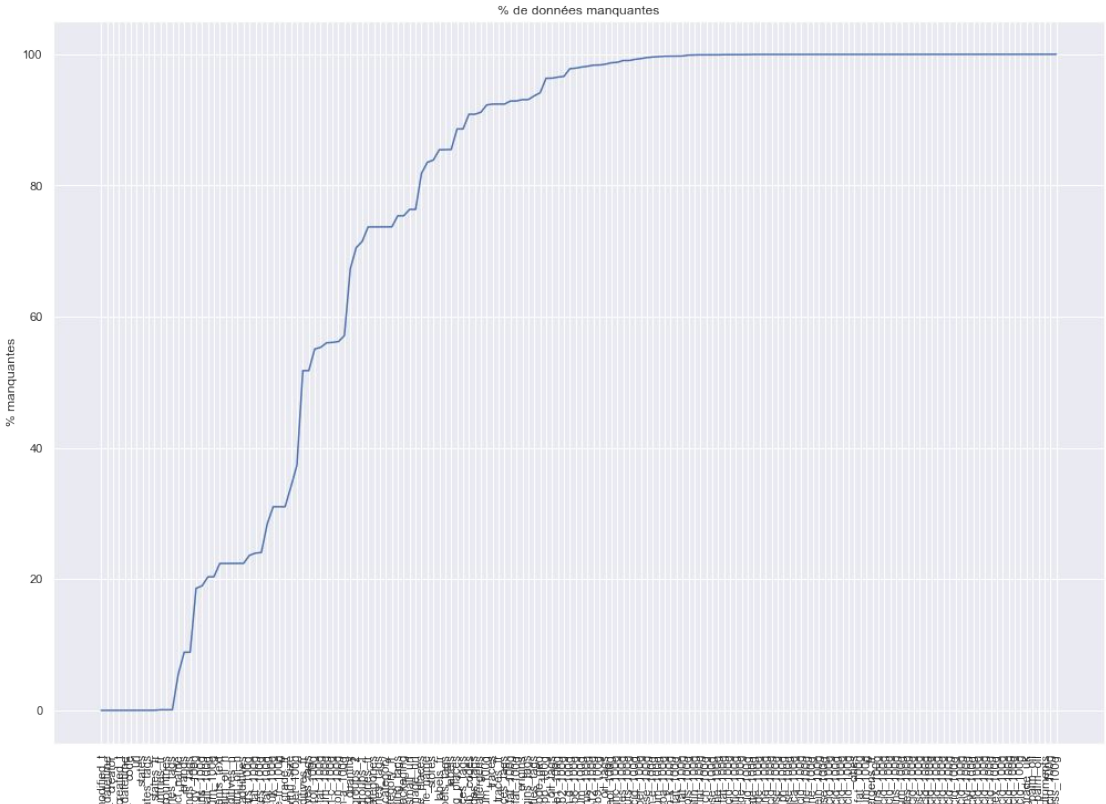
Les données brutes

Le dataset a 320772 individus et 162 variables.

Il y a 160 colonnes avec des valeurs manquantes.

Il y a 76.2% de valeurs manquantes dans le dataset brute.

Il y a 21 colonnes dans le dataset avec une unique valeur ou toute manquante.



Idée d'application

Foodscan

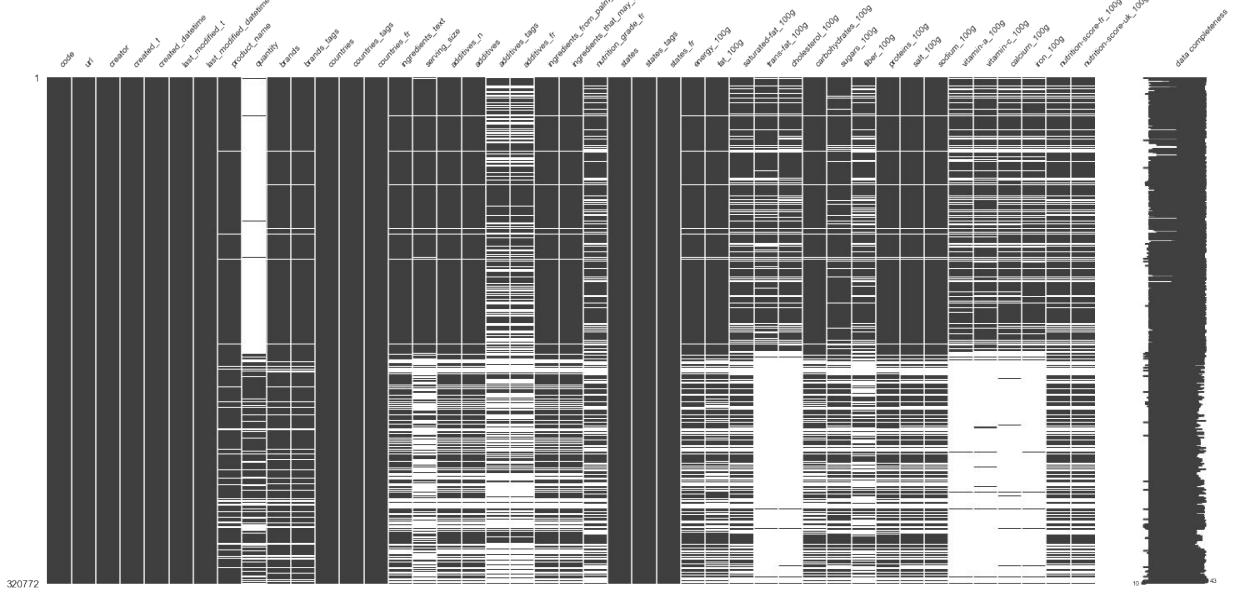
- Scanner avec un unique code barre le produit en magasin

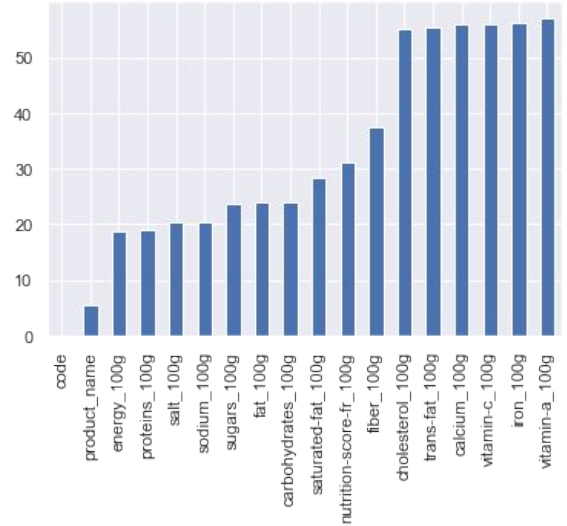
- Choisir un ingrédient spécifique ou critère diététique précis (une composante ACP) et proposer en fonction du nutrition_score le meilleur produit possible

Les données pertinentes

Elimination des colonnes 100% NaN et des variables pour avoir au moins 70% de valeurs présentes. Il y a 34.6% de valeurs manquantes dans le dataset pop. Nous observons aussi une nette amélioration du % des valeurs manquantes dans notre jeu de données.

Au final, nous nous intéresserons qu'aux variables quantitatives suivantes :'energy_100g', 'fat_100g', 'saturated-fat_100g', 'trans-fat_100g','cholesterol_100g', 'carbohydrates_100g', 'sugars_100g', 'fiber_100g', 'proteins_100g', 'salt_100g', 'sodium_100g', 'vitamin-a_100g','vitamin-c_100g', 'calcium_100g', 'iron_100g','nutrition-score-fr_100g'



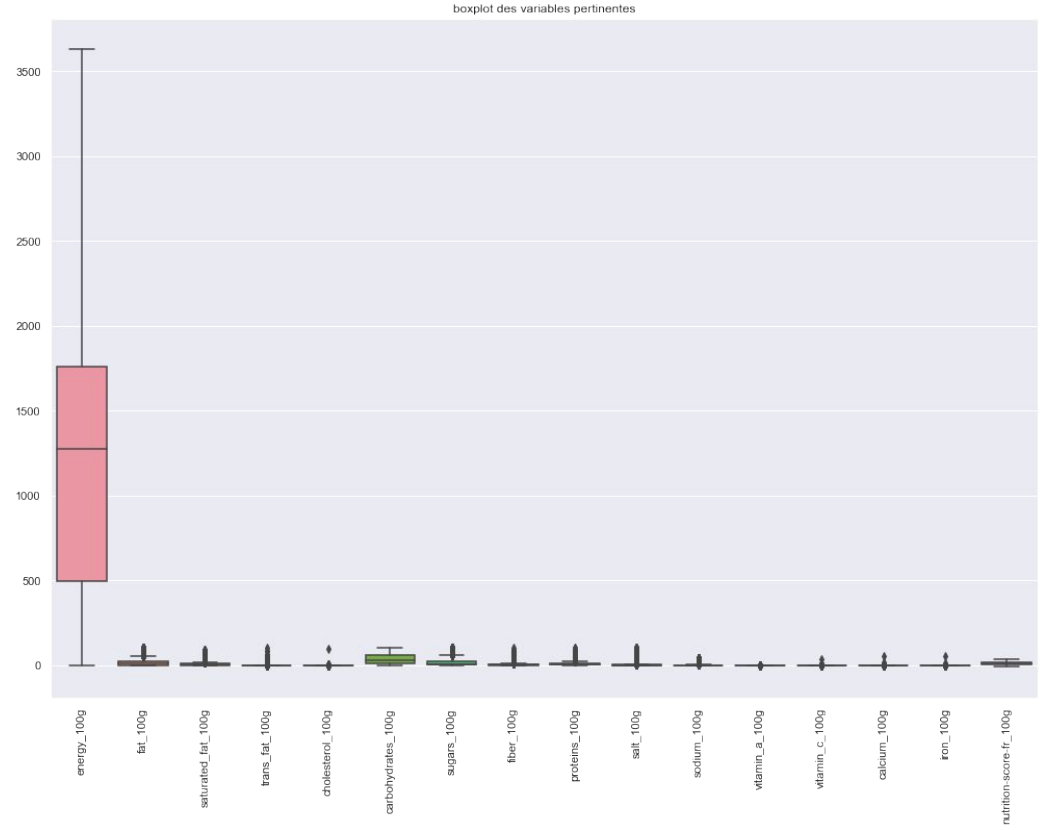


Nettoyage du dataset

Suppressions des duplicatas de 'code'

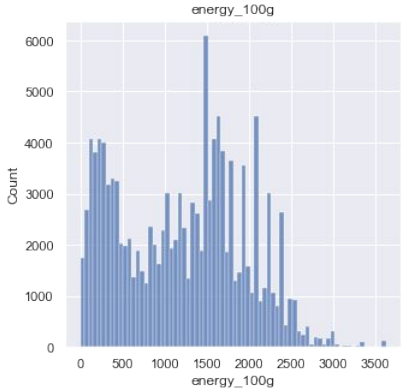
Pour la variable energy_100g nous utiliserons la méthode IQR, max= Q3 + 1.5 * IQR = 3700

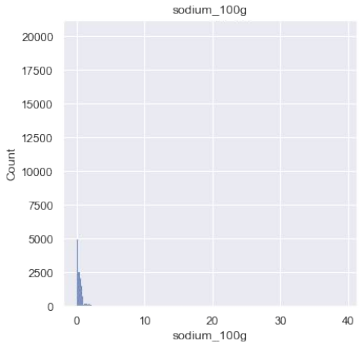
Pour les autres variables, nous les bornerons à >=0 & <=100

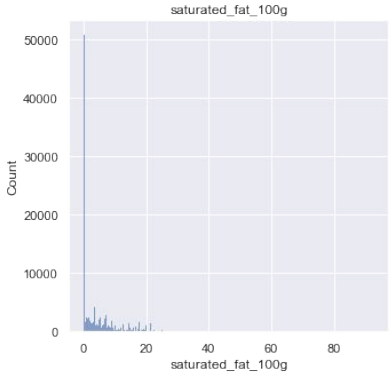


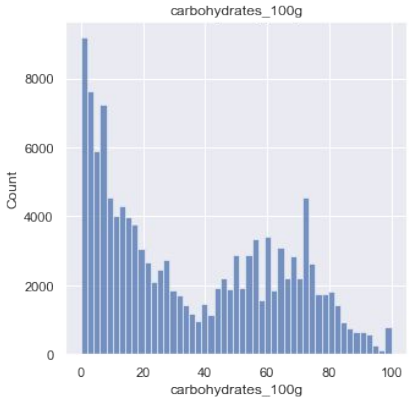
Nettoyage du dataset

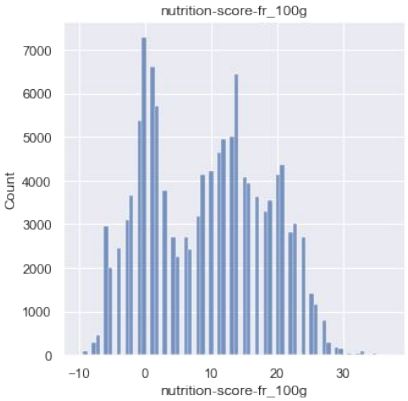
- energy_100g, carbohydrates_100g et nutrition-score sont de forme bimodale
- sodium_100g, salt_100g, sugar_100g, saturated_100g, trans_fat_100g et proteins_100g sont de forme unimodale et asymétrique, "tail" étalées à gauche

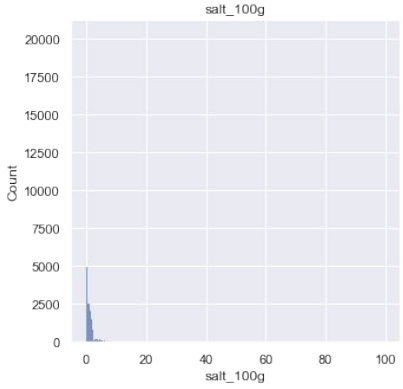


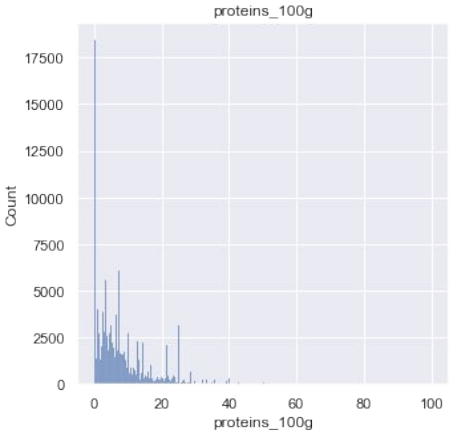


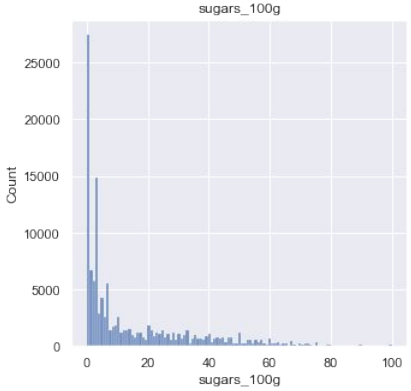






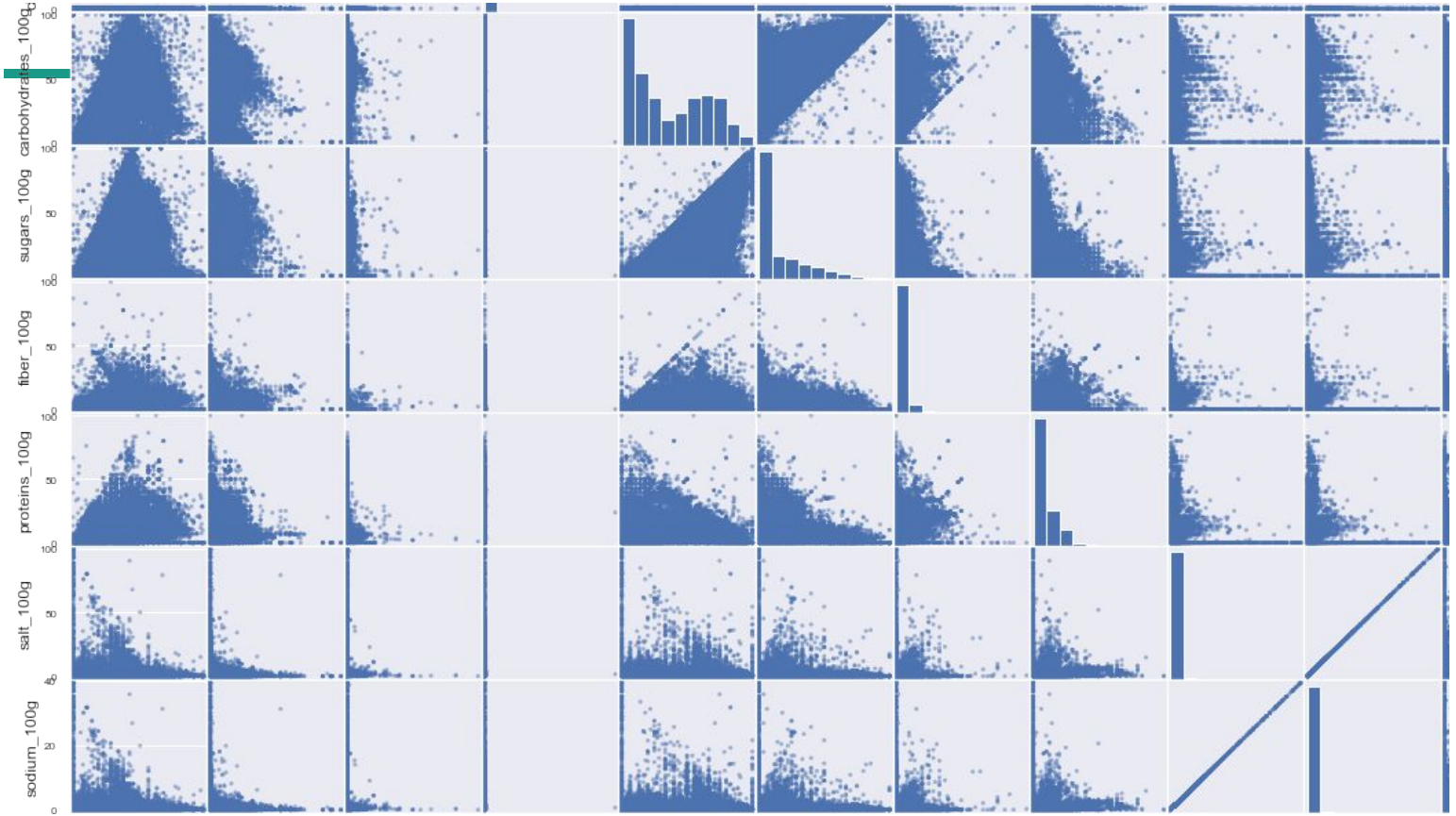




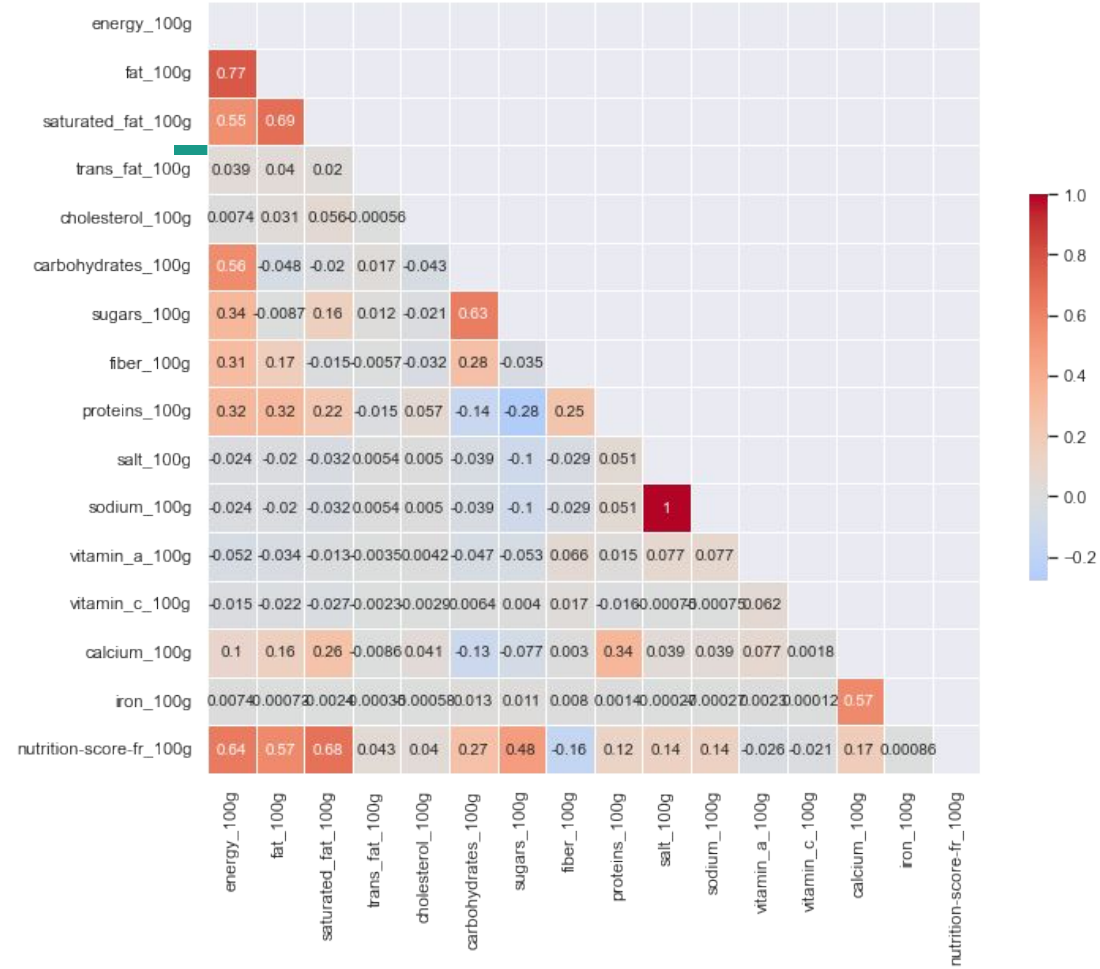


Analyse multivariée

Nous étudierons ici les relations de corrélation linéaires possibles entre nos variables quantitatives



Analyse multivariée, corrélations



corrélation faible:

- fat_100g, energy_100g, saturated_fat_100g et nutri-score
- sugars_100g, carbohydrates_100g et energy_100g
- iron_100g et calcium_100g

corrélation forte :

- salt_100g et sodium_100g (évidente)

Projection sur des axes décorrélés entre fat_100g, energy_100g, saturated_fat_100g et nutri-score

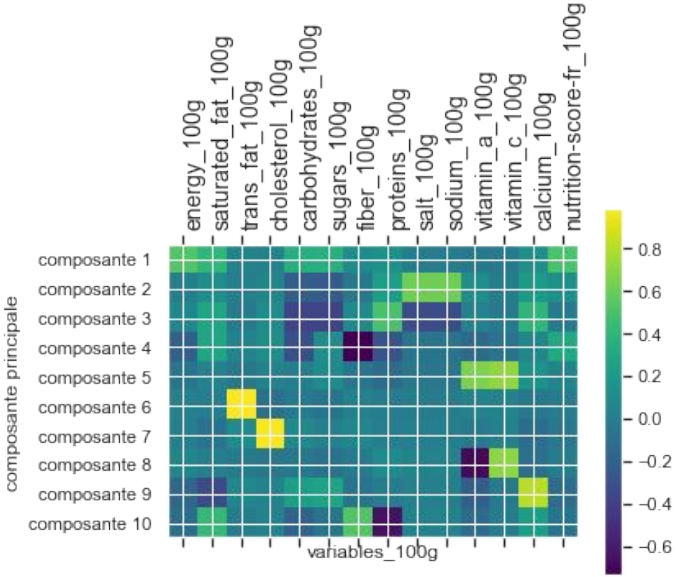
1. Testons saturated_fat_100g avec energy_100g = `0.546374`
2. Testons saturated_fat_100g avec fat_100g = `0.687112`

Analyse multivariée, ACP

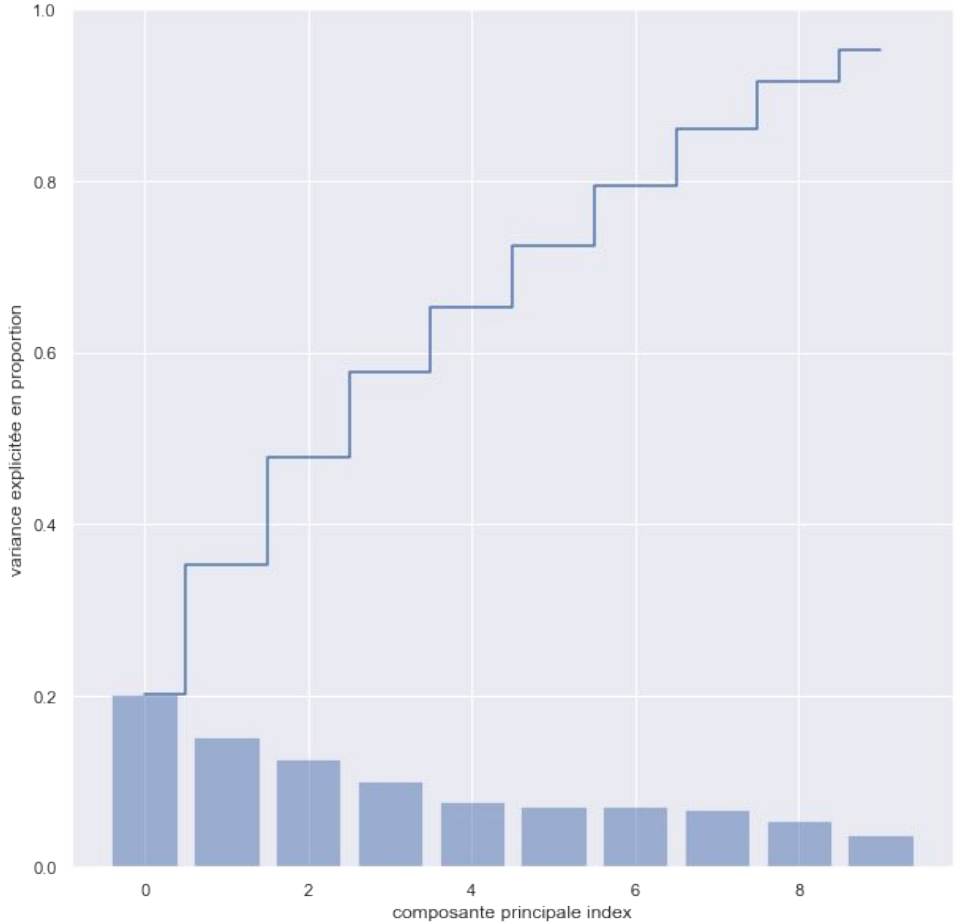
Analyse des composantes principales afin de projeter notre jeu de donnée sur plusieurs axes, qui peuvent être visualisées graphiquement, en perdant le moins possible d'information.

- energy_100g et le nutrion-score_100g = caractérise les ingrédients générique sain
- salt_100g et sodium_100g = caractérise les ingrédients en surplus de sel à éviter
- saturated_fat_100g, proteins_100g et calcium_100g = caractérise un régimes prise de masse musculaire
- vitamin_a_100g et vitamin_c_100g = caractérise un régime pour booster son métabolisme en vitamine

4 nouvelles variables synthétiques pourraient réduire ainsi le nombre de colonne finale

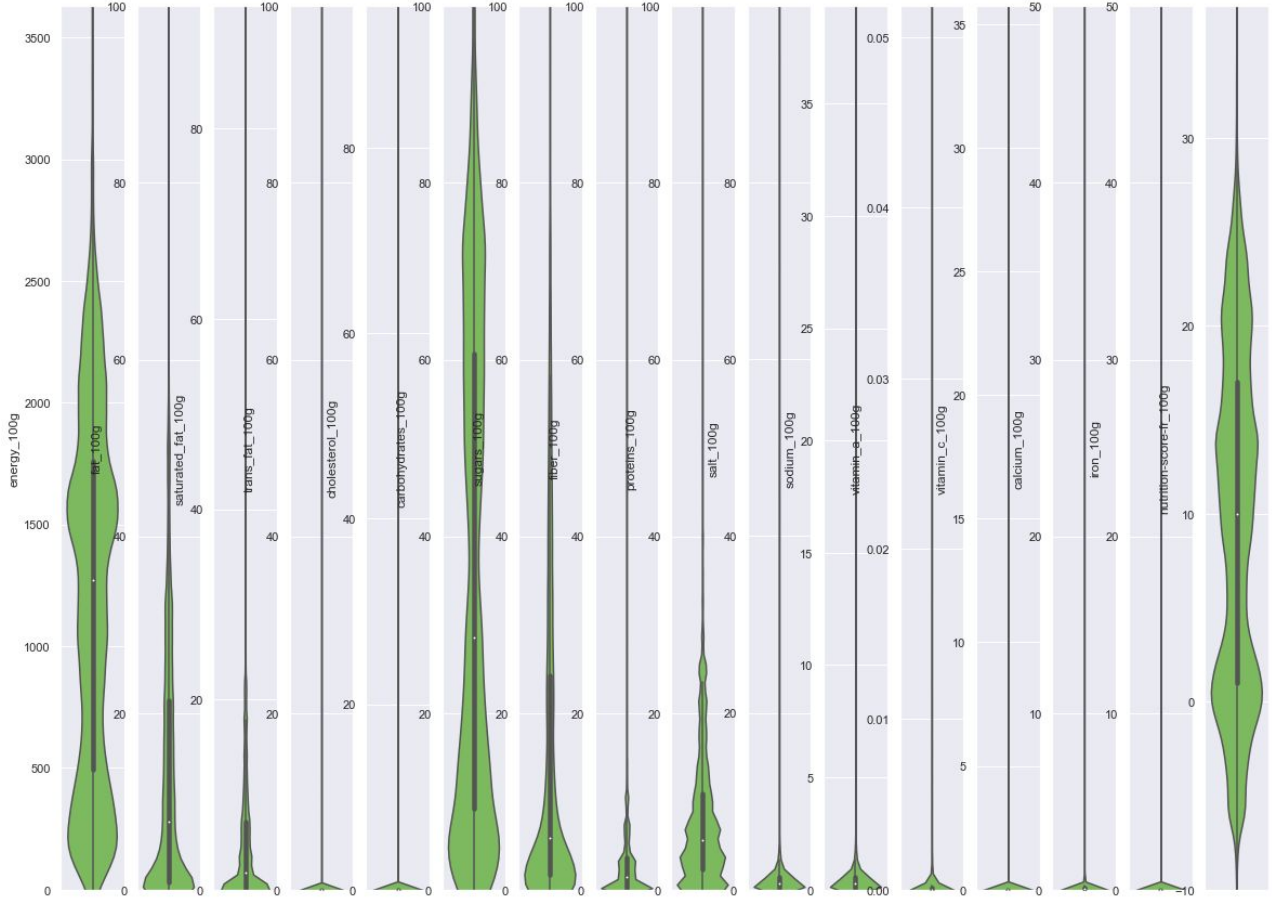


Analyse multivariée, ACP



Représentation après imputation kNN

violinplot = diagrammes en boîte + estimations de densité qui représentent la distribution des données



Conclusion

- Axes ACP et proposition du meilleur ingrédient selon le nutrition score
- Imputation kNN