Projet 3 : Concevez une application au service de la santé publique

Problématique : Trouver des idées innovantes d’applications en lien avec l'alimentation

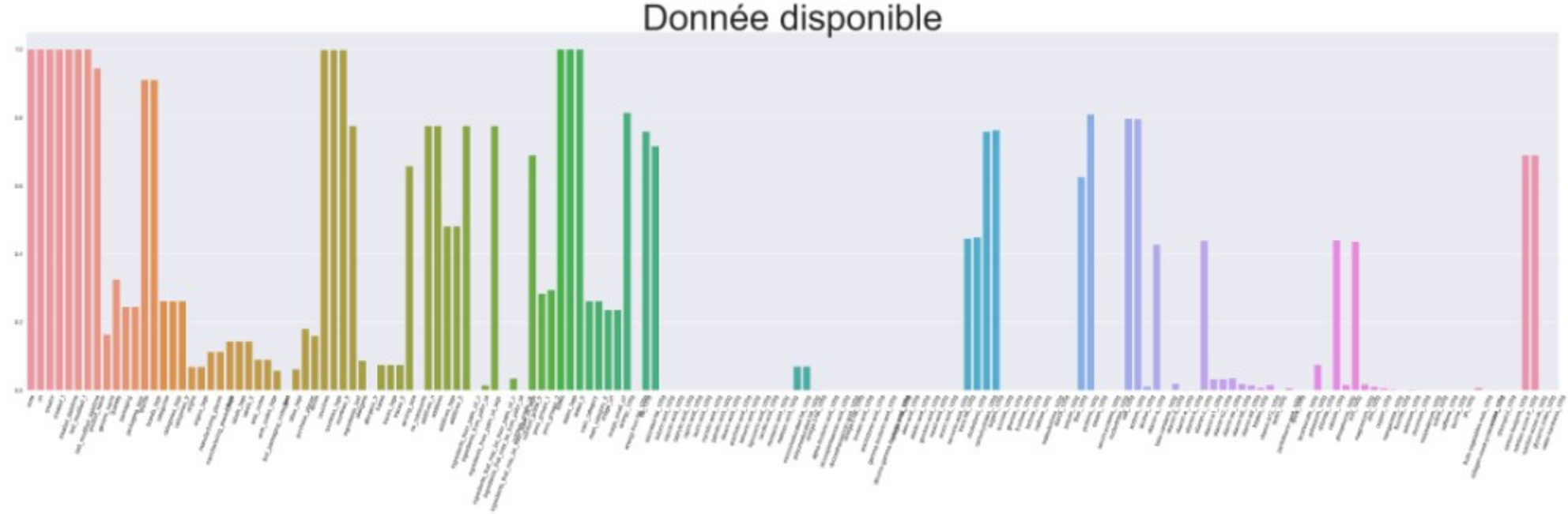
Idée d’application

Application nutrition saine :

- Scan le produit et renvoie son Nutri-Score (nutrition-grade-fr dans les données)
- Propose ensuite une liste de produits de la même catégorie avec un meilleur Nutri-Score (On utiliserait la méthode KNN pour trouver des produits similaires).
- Éventuellement, propose une liste de produits plus large avec moins de similitude si l’utilisateur le désir.

Nettoyage des données ; donnée manquante

- Après avoir examiné le dataset, on a commencé par vérifier le pourcentage de données disponibles pour chaque colonne :



Nettoyage des données ; donnée manquante

- On remarquait que beaucoup de colonne avait plus de 50 % de donnée manquante, voir était totalement vide.

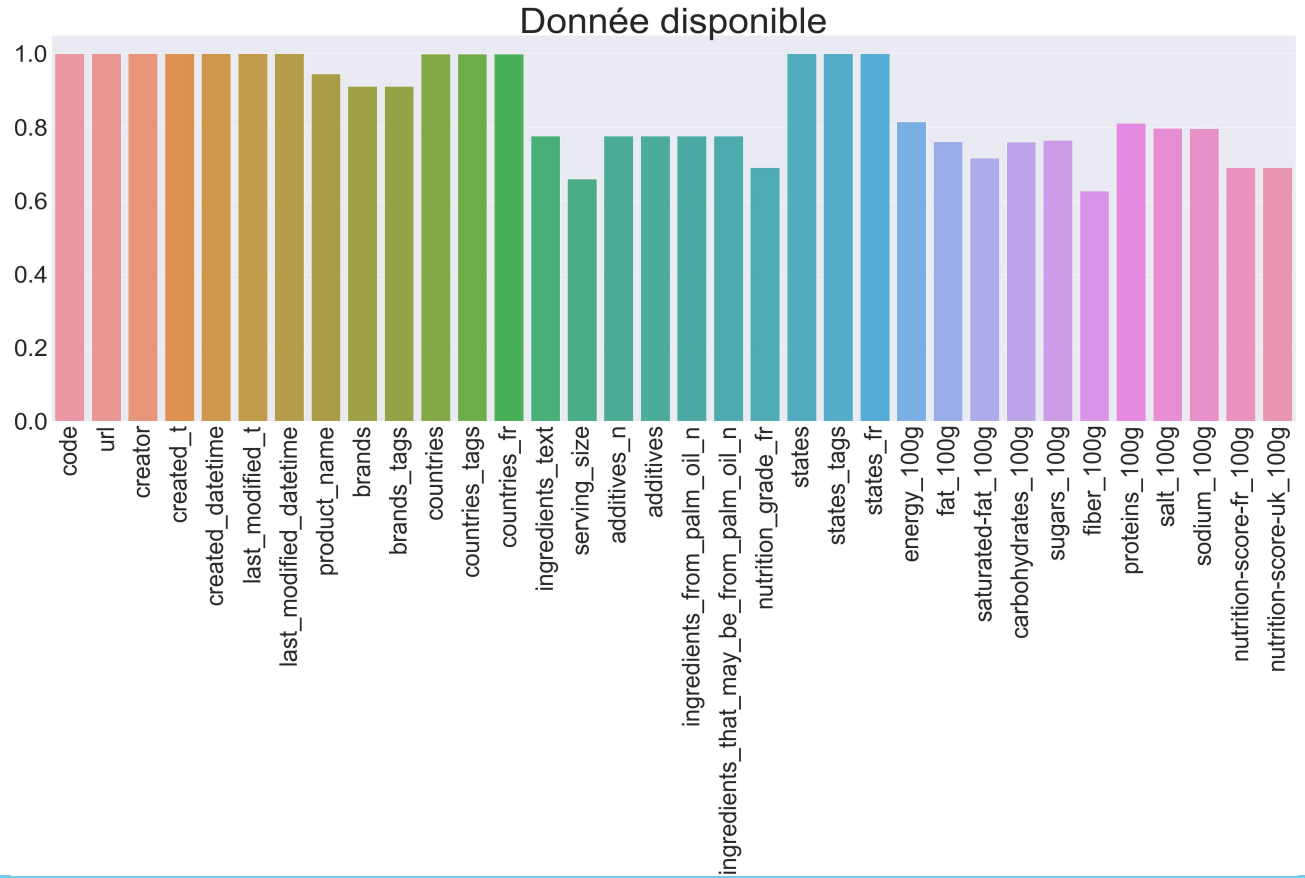
  On se débarrasse des colonnes ayant plus de 50 % de données manquante, puis on regarde à nouveau les données manquante :

Nettoyage des données ; donnée manquante

- On remarquait que beaucoup de colonne avait plus de 50 % de données manquantes, voir était totalement vide.

  On se débarrasse des colonnes ayant plus de 50 % de données manquantes, puis on regarde à nouveau les données manquantes.

Nettoyage des données ; donnée manquante

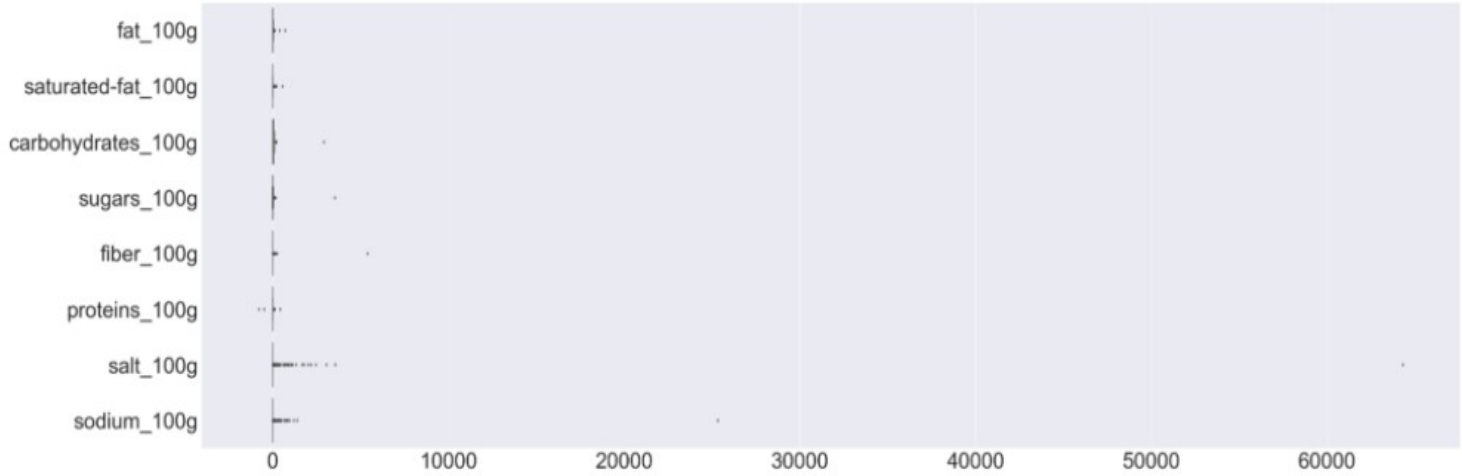


Nettoyage des données ; colonnes

- Les colonnes qui nous servirons le plus sont les colonnes ingrédients_100g, energy_100g, nutritition_score et nutrition_grade.
- Ces colonnes semblent les plus pertinentes pour notre application.

Nettoyage des données ; valeurs aberrantes

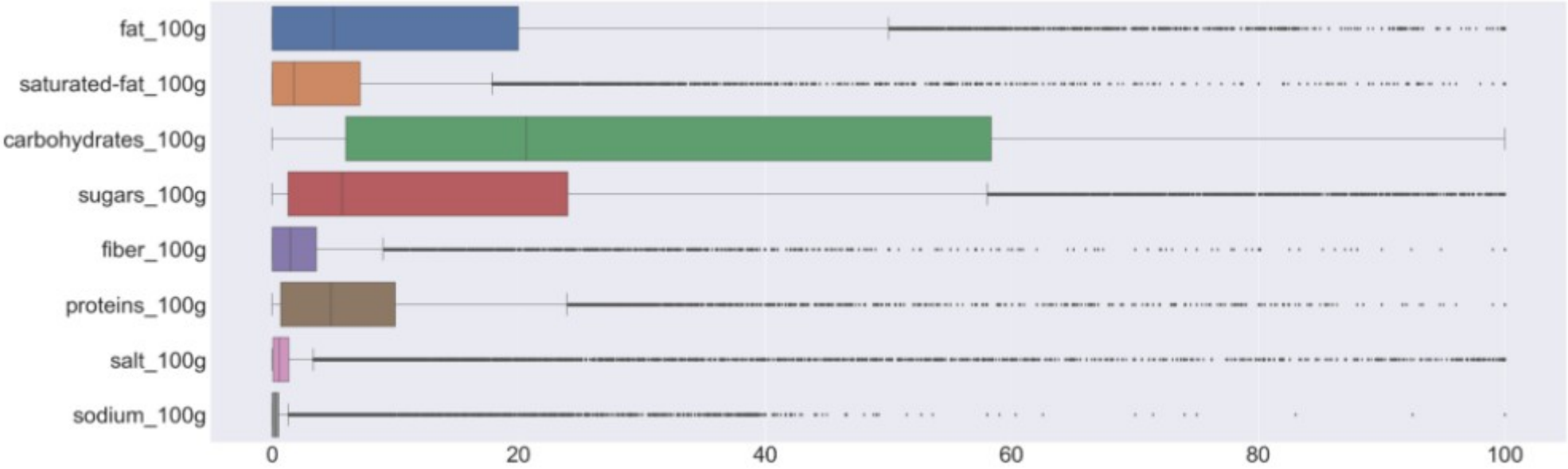
- C’est mieux. Beaucoup de colonne ont été supprimé.
- A présent, on va regarder les données numériques pour vérifier qu’il n’y a pas de valeurs aberrantes dans les ingrédients :



- Il y a très nettement des valeurs aberrantes.

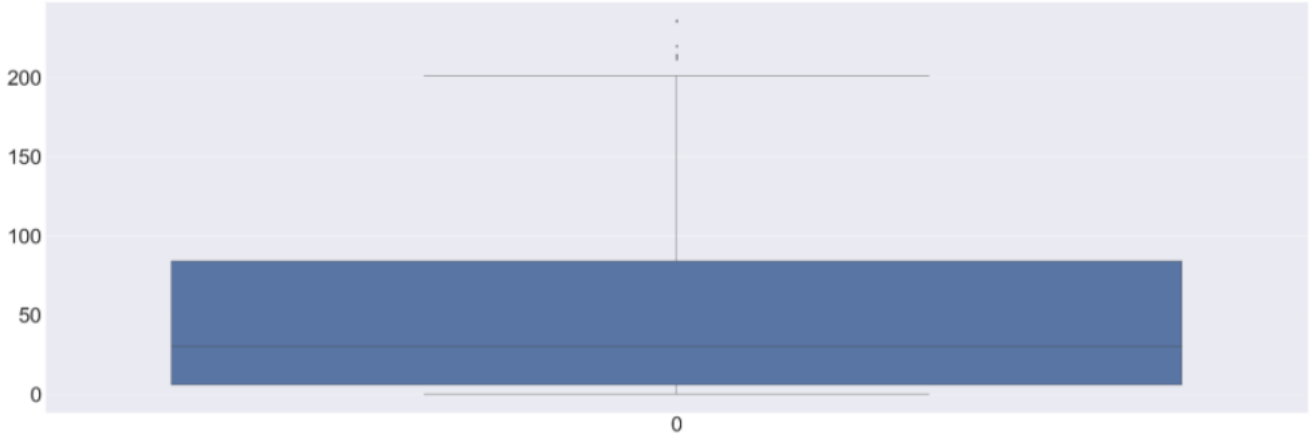
Nettoyage des données ; valeurs aberrantes

- On garde uniquement les valeurs entre 0 et 100. On regarde à nouveau la répartition des données :



Nettoyage des données ; valeurs aberrantes

- On vérifie maintenant que la somme des ingrédients est bien égale à 100 :



De toute évidence, ce n’est pas le cas.

Nettoyage des données ; valeurs aberrantes

- On se débarrasse des lignes où la somme des ingrédients est supérieure à 100, puis on vérifie la répartition de la somme des ingrédients :

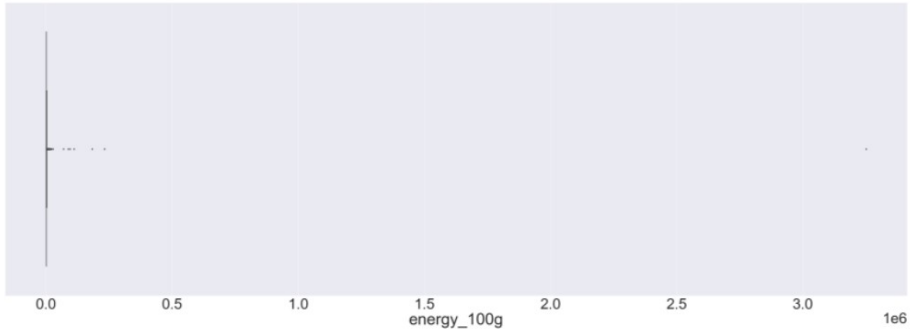
Nettoyage des données ; valeurs aberrantes

Nettoyage supplémentaire sur les ingrédients :

- On se débarrasse des lignes qui n’ont pas au moins 50 % des ingrédients renseignés.

Nettoyage des données ; valeurs aberrantes

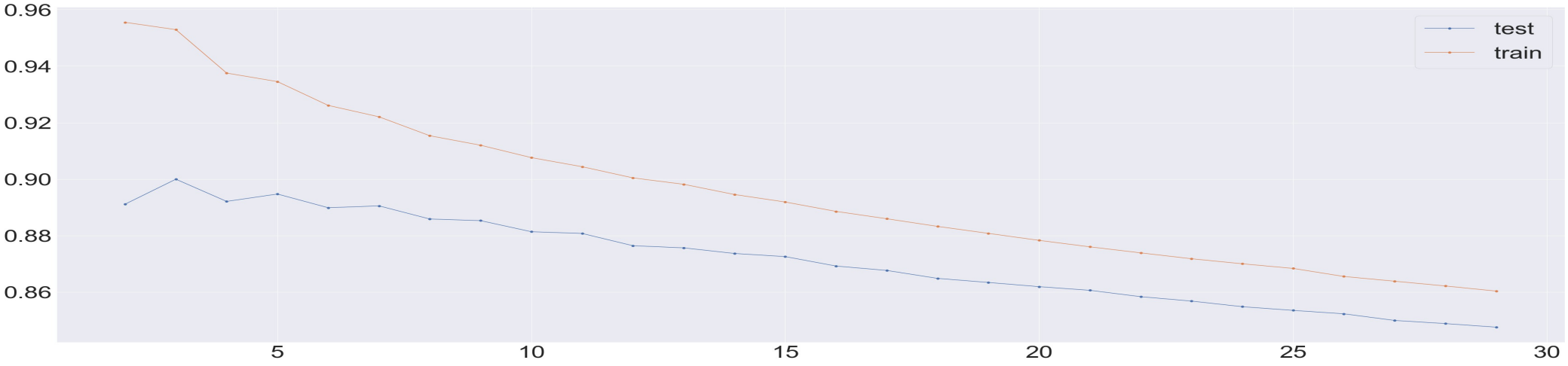
- On regarde la répartition de l'énergie pour 100g de chaque produit :



- Il y a de toute évidence des valeurs aberrantes. L'énergie maximale pour 100g est de 3725 joules (100g de gras). On se débarrasse donc de toutes les lignes ayant une énergie de plus de 3725 joules.

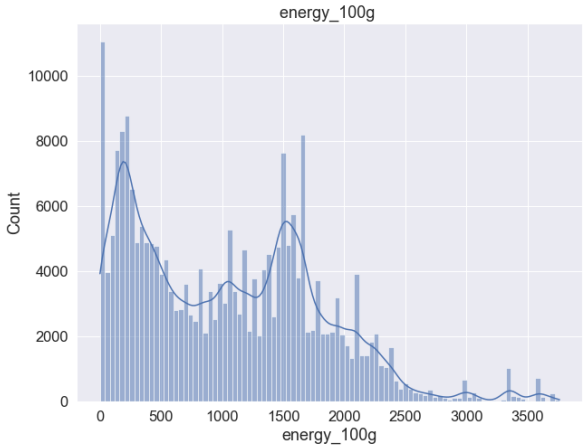
Nettoyage des données ; remplissage

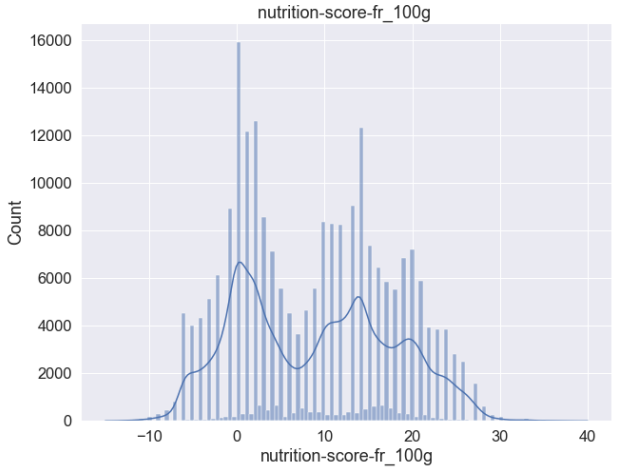
- On remplit les données manquantes pour les ingrédients et le nutritious_score avec un KNN_Imputer
- On fait de même pour le nutrious_grade, avec un KNN_classifier après avoir sélectionné un nombre satisfaisant de n_neighbour (éviter l'overfiting) en comparant le train et le test score.



Analyse des données

- On commence par faire un kdeplot des ingrédients, de l'énergie et du nutrition_score.

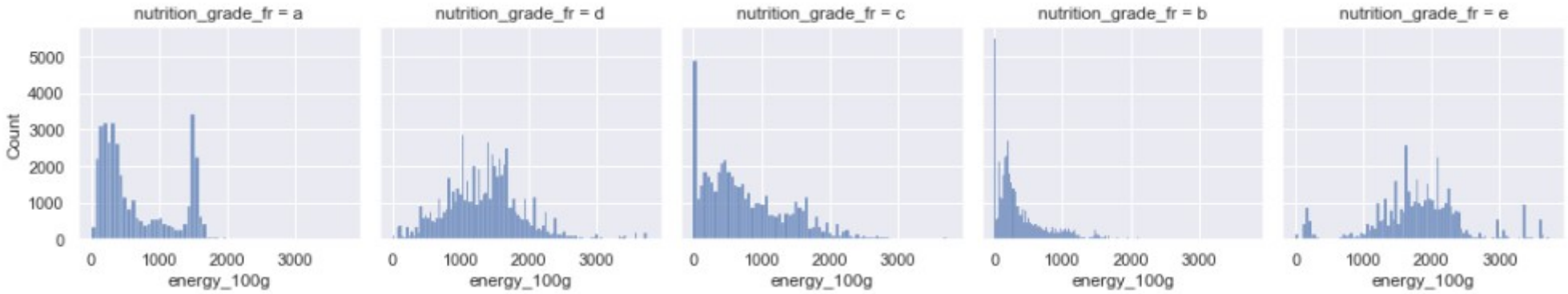




- On remarque que l'énergie et le nutrition_score ont tous les deux deux pics dans leur kdeplot.

Analyse des données

- Avec un facet_grid, on cherche une corrélation entre l'énergie et le nutrition_grade.



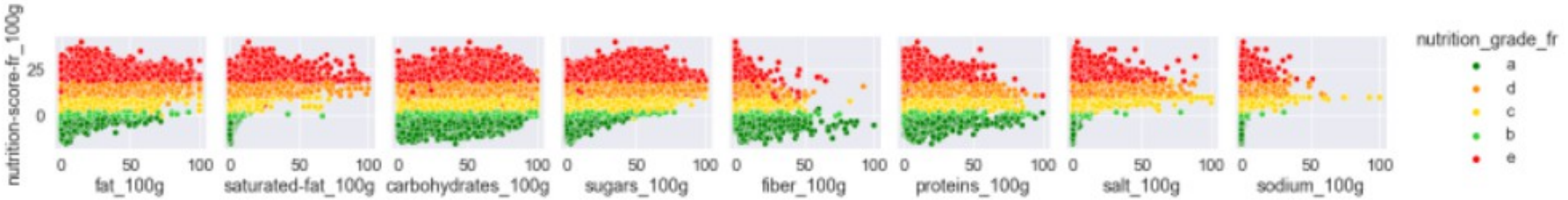
- On remarque que les aliments avec le plus mauvais nutritious_grade ont tendance à être les plus énergétiques

Analyse des données

- Remarquons également qu’il y a beaucoup d’énergie très faible pour les nutrition_grade b et c.
- Après vérifications, il s’agit de boisson de type thé et eau.

Analyse des données

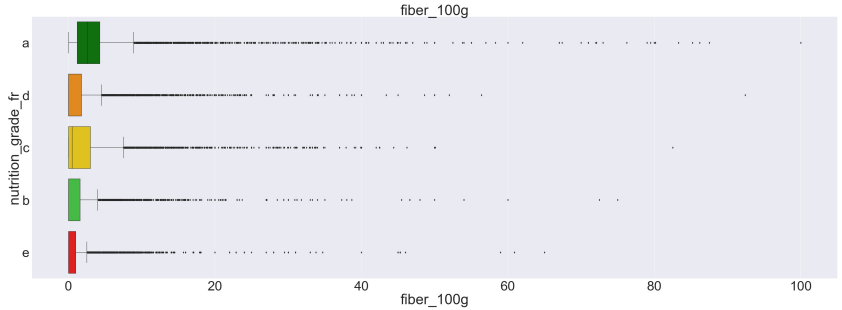
- On observe à présent les liens entre le nutrition_grade et les ingrédients avec un pairplot :

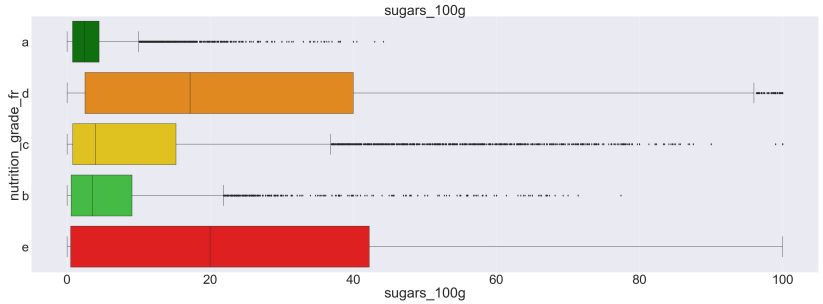


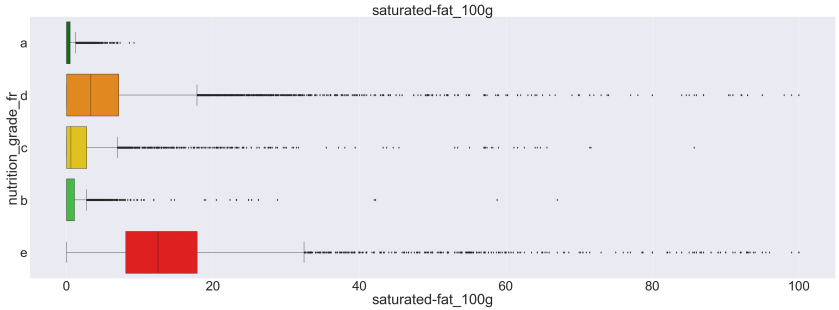
- On remarque que les produits avec le meilleur nutrition_grade ont tendance à avoir peu de gras (et encore moins de gras saturé), de sucre et de sel. Ils ont tendance à avoir plus de fibre.

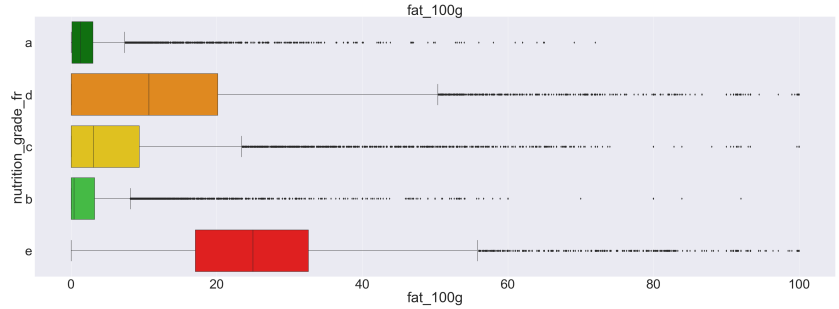
Analyse des données

- On fait quelque boxplot pour développer notre observation précédente :



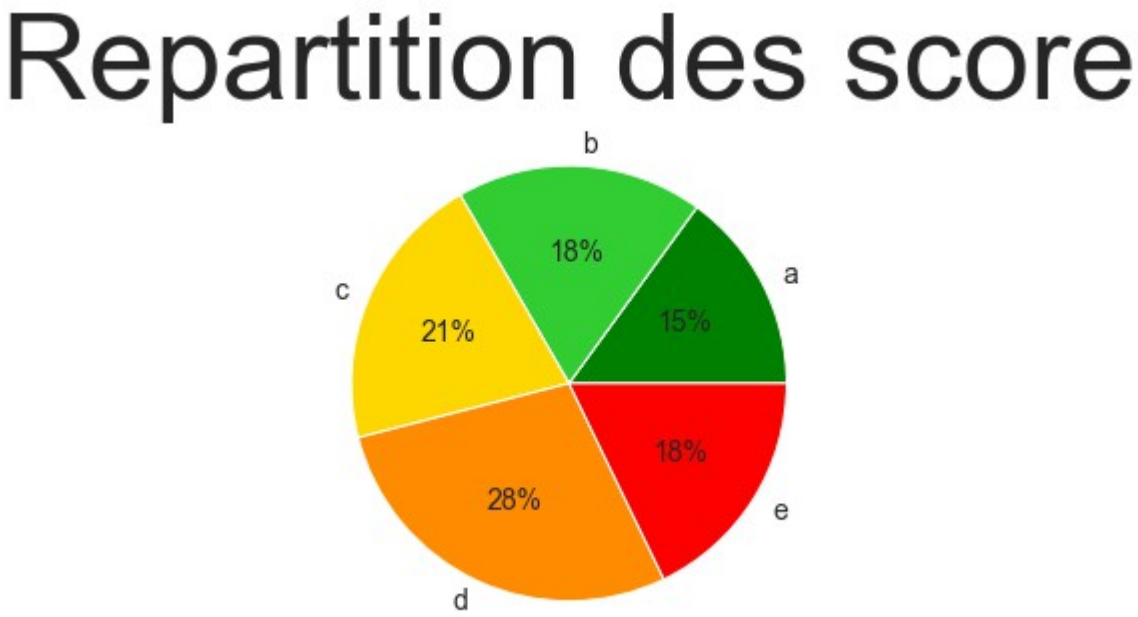






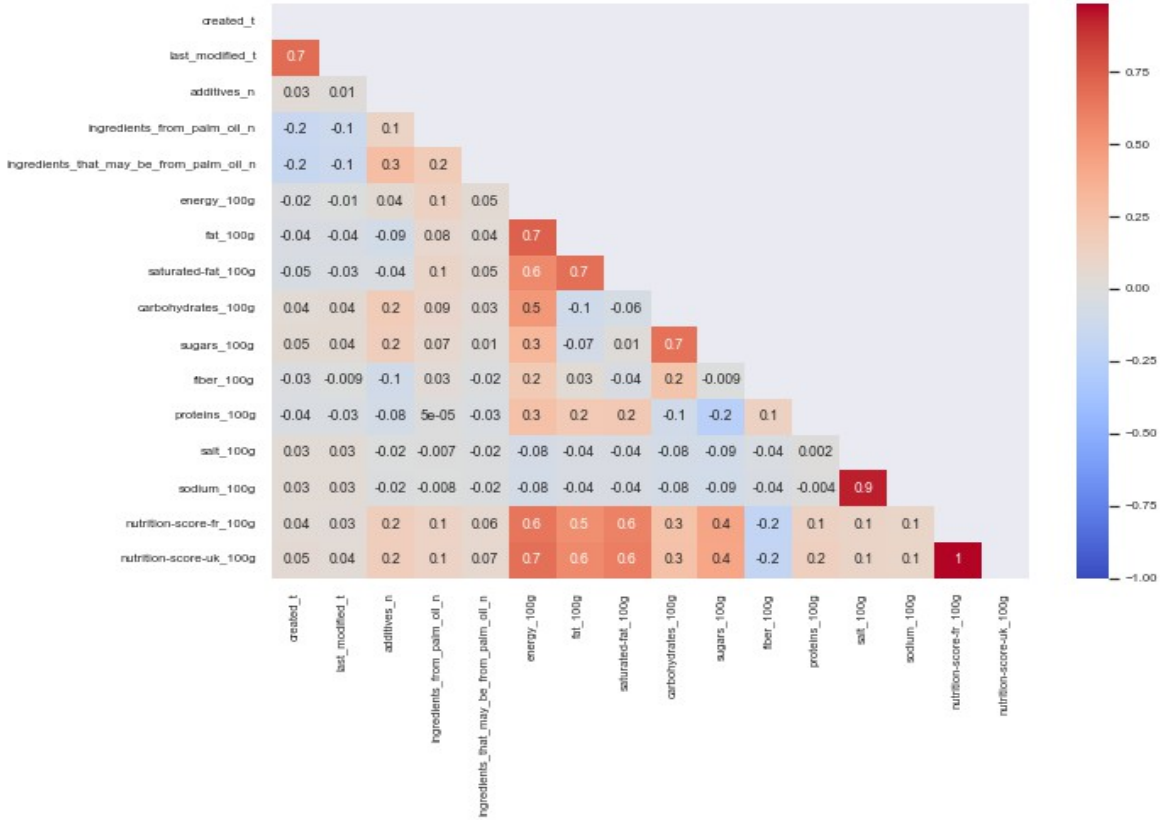
Analyse des données

- On regarde la répartition des scores :



Analyse des données

- On regarde la carte des corrélations :



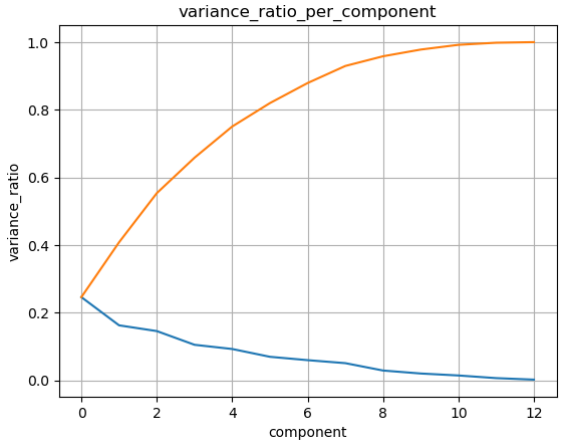
Analyse des données

- On remarque que le nutrition_score est fortement corrélé de manière positive avec l'énergie, le gras, et le sucre, et négativement corrélé avec les fibres, ce qui semble confirmer nos observations précédentes.
- La seule exception étant le sel, qui est faiblement corrélé avec le nutrition_score.

Analyse des données

- On réalise à présent une PCA sur les valeurs numériques.

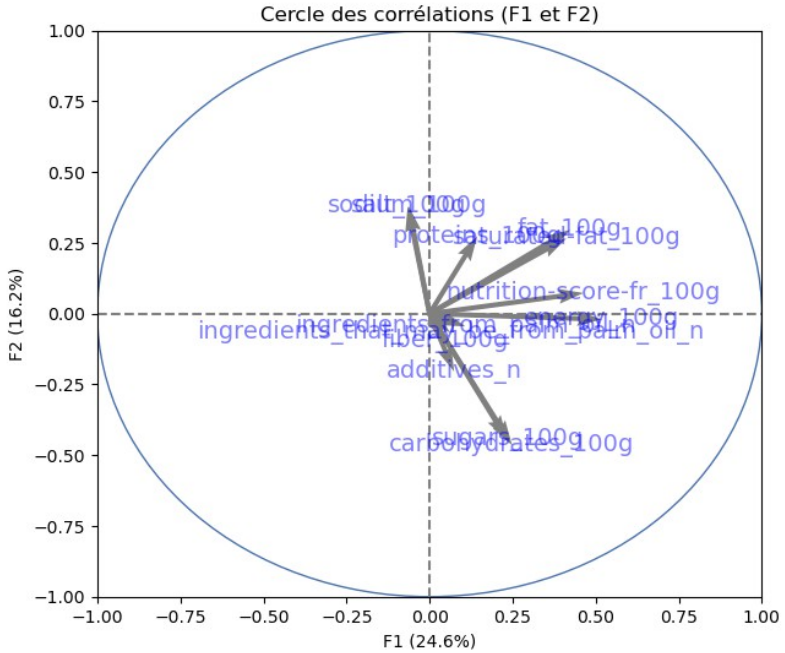
  On commence par tracer la part de variance par composants :



- Les deux premiers composants expliquent +40 % de la variance.

Analyse des données

- On fait à présent le cercle des corrélations avec les deux premières composantes :

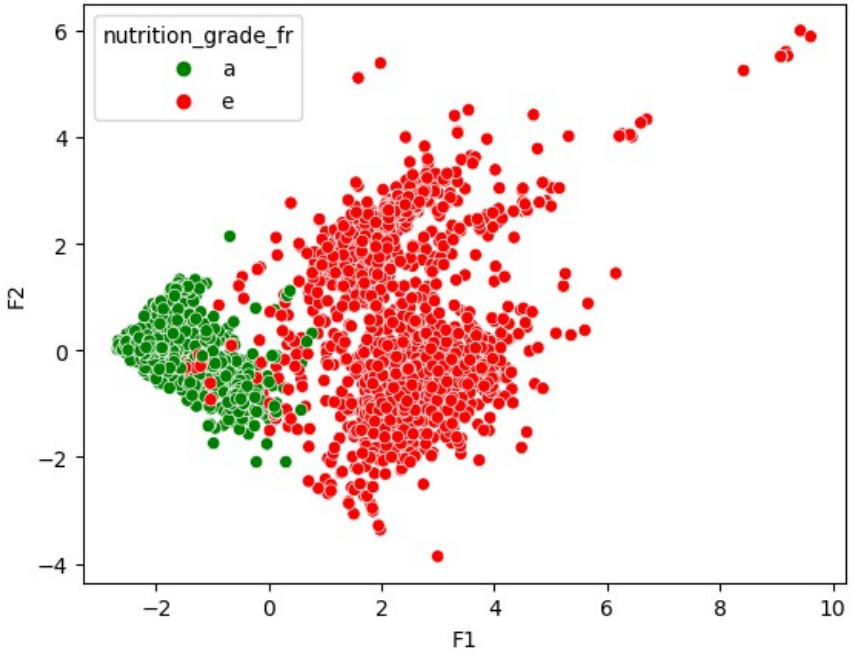


Analyse des données

- On remarque que plusieurs composants sont corrélés entre eux :
  - Le sel et le sodium
  - Le sucre et le carbohydrate
- On remarque que plusieurs composants ne sont pas corrélés :
  - Le nutrition_score et le sel et le sodium.
  - Le gras et le sucre
  - Le sel et l’énergie.
- Plusieurs composants sont anti-corrélés :
  - Le sel et le sucre
  - Les additifs et le sel

Analyse des données

- On plot les deux premiers composant l'un par rapport à l'autre



Analyse des données

- On fait l'analyse ANOVA entre le premier composant et le nutritious_grade et on observe les résultats :

	Source	SS	DF	MS	F	p-unc	np2
0	nutrition_grade_fr	19084.919303	4	4771.229826	4397.050174	0.0	0.660262
1	Within	9820.135823	9050	1.085098	NaN	NaN	NaN

- L'eta 2 est de 0.66. Les catégories sont bien séparées statistiquement par le composant choisi, ce qui indique que les éléments choisis pour le PCA ont une influence significative.

Conclusion

- Les ingrédients utilisés pour un produit, ainsi que son apport énergétique sont de bons indicateurs pour prédire son Nutri-Score.
- L'application est réalisable. On récupère la composition du produit et on cherche les produits similaires avec un algorithme du type KNN-Imputer.
- On pourrait chercher les produits similaires ayant moins de gras/gras saturé et/ou moins de sucre.