Proposition d’une amélioration de la base de données OpenFoodFacts

Farah Mokhtari

Vue d'ensemble

L'agence Santé publique France m'a confié l'étude de création d'un système de suggestion ou d'auto-complétion pour aider les usagers à remplir plus efficacement la base de données.

La demande était de :

- prendre en main des données, en **les nettoyant et les explorant**.
- Puis, étudier la faisabilité de **suggérer les valeurs manquantes** pour une variable dont plus de 50% des valeurs sont manquantes.

Que va-t-on faire concrètement ?

1- Récupération des données	Quelles sont les données à notre possession ? Quelles sont les données qui seront pertinentes à notre problématique ? Connaissance de la problématique de mon client.
2- Nettoyage des données	Nettoyer les données, c'est s'assurer qu'elles sont consistantes, sans valeurs aberrantes ni manquantes.
3- Exploration des données	Mieux comprendre les différents comportements et de bien saisir le phénomène sous-jacent.
4- Modélisation	Le travail de modélisation consiste à trouver le bon modèle statistique qui colle le mieux aux données
5- Evaluation et modélisation	L'évaluation de la qualité de notre modèle, c'est-à-dire sa capacité à représenter avec exactitude notre phénomène, ou a minima sa capacité à résoudre notre problématique.
6- Mise en production	Déployer le mode dans un environnment afin de le mettre à l'usage d'un plus grand nombre de personnes.

Récupération des données

Quelles données va-t-on étudier

Les informations générales, Tags, Misc Data and ingredients :

- Des informations propres à chaque produit et difficilement déductible à partir du reste des informations
- Certaines informations peuvent cependant être calculées ou recalculées automatiquement comme le ‘creator’, ‘created_t’, ‘created_datetime’, ‘modified_t’

Informations nutritionnelles

Ces informations peuvent faire l’objet d’un système d’auto-suggestion puisque on peut supposer qu’en ayant certaines valeurs nutritionnelles, on peut créer un modèle qui peut compléter le reste des valeurs.
Dans le cadre de notre projet, je me suis contenté uniquement des valeurs qui interviennent dans le calcul du nutriscore cad : l’énergie en kJ, les graisses saturées, sel, protéines, les sucres, fruits, légumes et noix et fibres.

Nettoyage des données

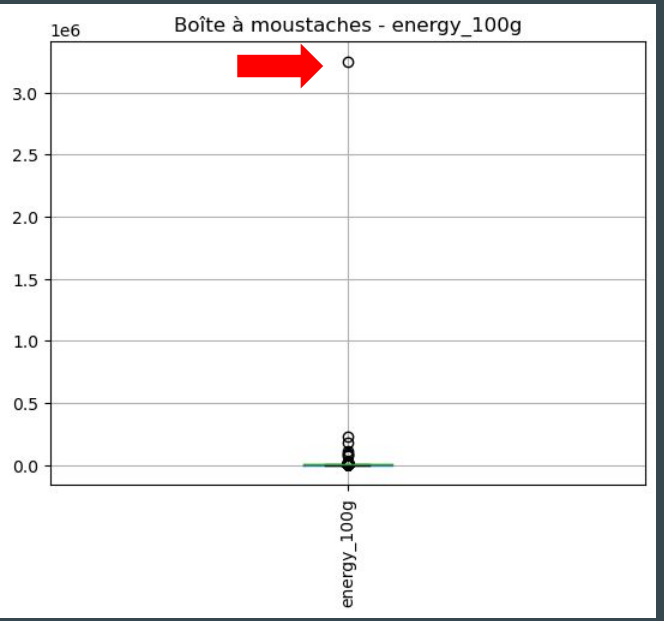
Garder que de la donnée pertinente !

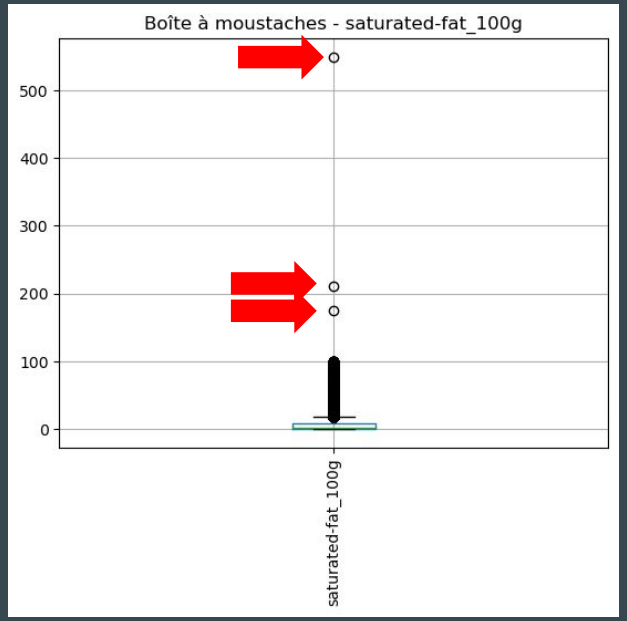
Lors du nettoyage des données, j'ai commencé par les actions qu'on fait couramment en data-science :

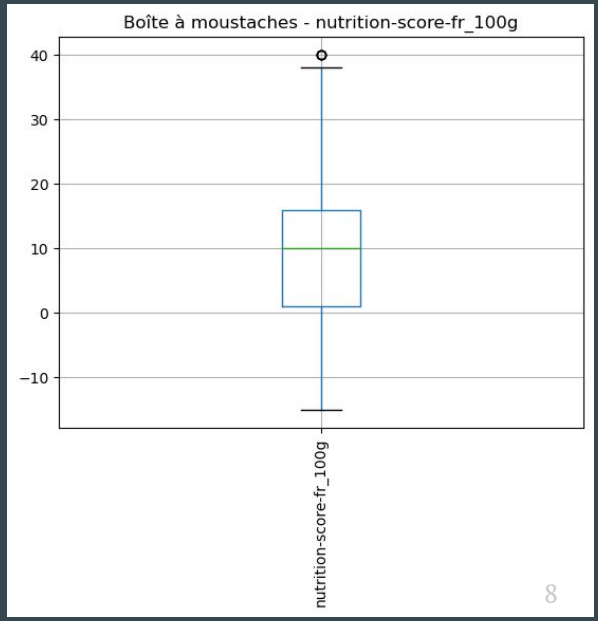
- Suppression des produits dupliqués
- Suppression des produits qui ont toutes les valeurs en nulles.
- Pour des raisons RGPD, j'ai aussi décidé de supprimer toutes les données personnelles comme le nom du créateur, et les différentes dates

Garder que de la donnée pertinente !

Pour des raisons “visuelles”, j’ai décidé d’afficher la boîte à moustache de nos valeurs nutritionnelles afin d’avoir une idée sur les possibles valeurs que peut prendre chaque valeur et potentiellement identifier les valeurs nutritionnelles qui ont des données aberrantes

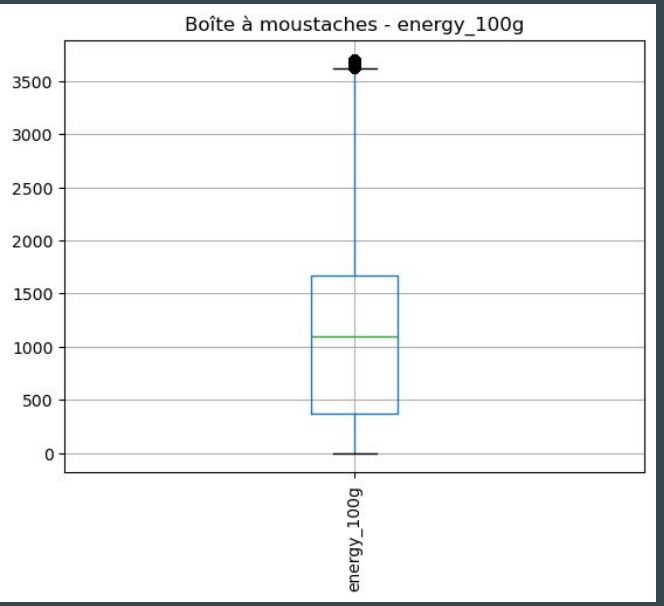


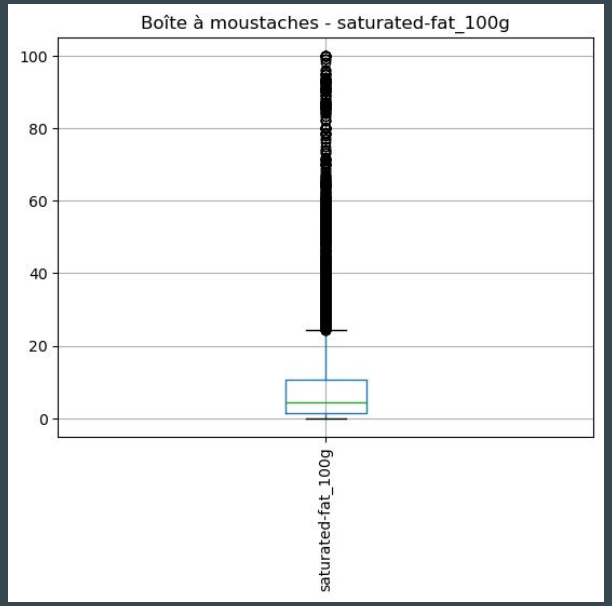


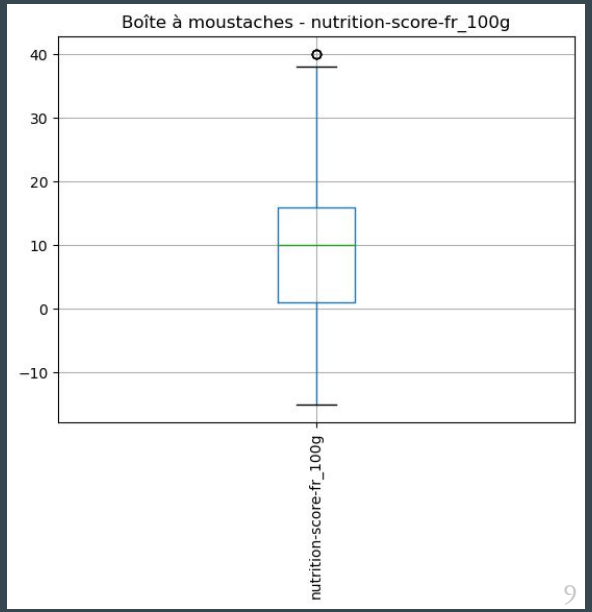


Garder que de la donnée pertinente !

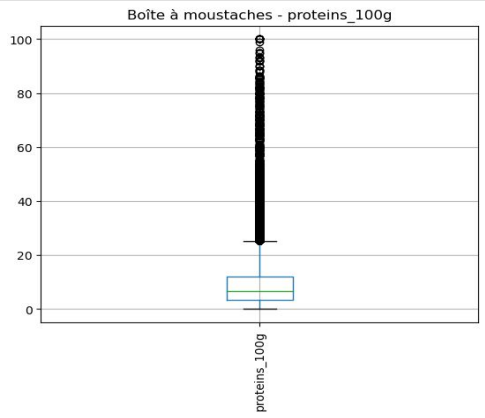
Pour des raisons “visuelles”, j’ai décidé d’afficher la boite à moustache de nos valeurs nutritionnelles afin d’avoir une idée sur les possibles valeurs que peut prendre chaque valeur et potentiellement identifier les valeurs nutritionnelles qui ont des données aberrantes



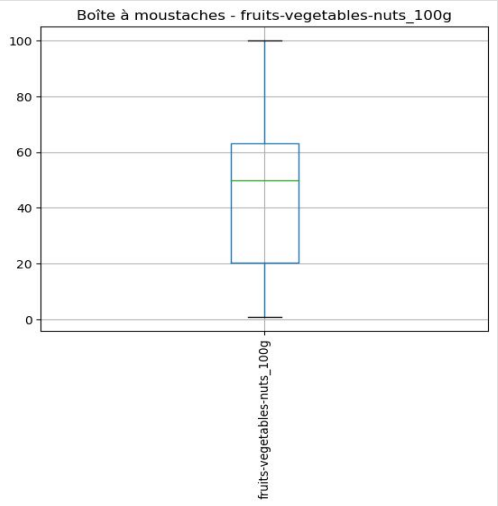


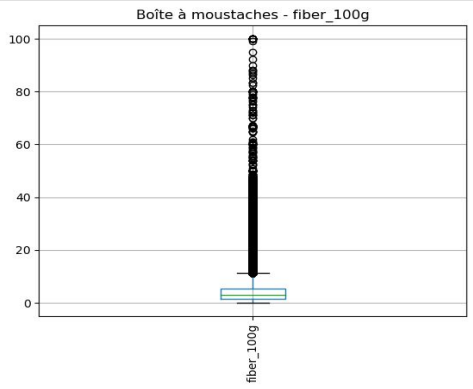


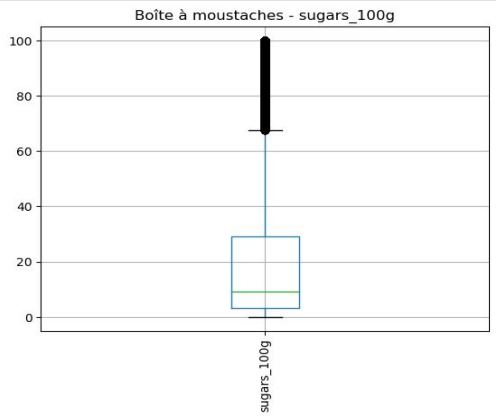
De même,


Boîte à moustaches - proteins_100g
100
80
60
40
20
0
proteins_100g


Boîte à moustaches - salt_100g
100
80
60
40
20
0
salt_100g

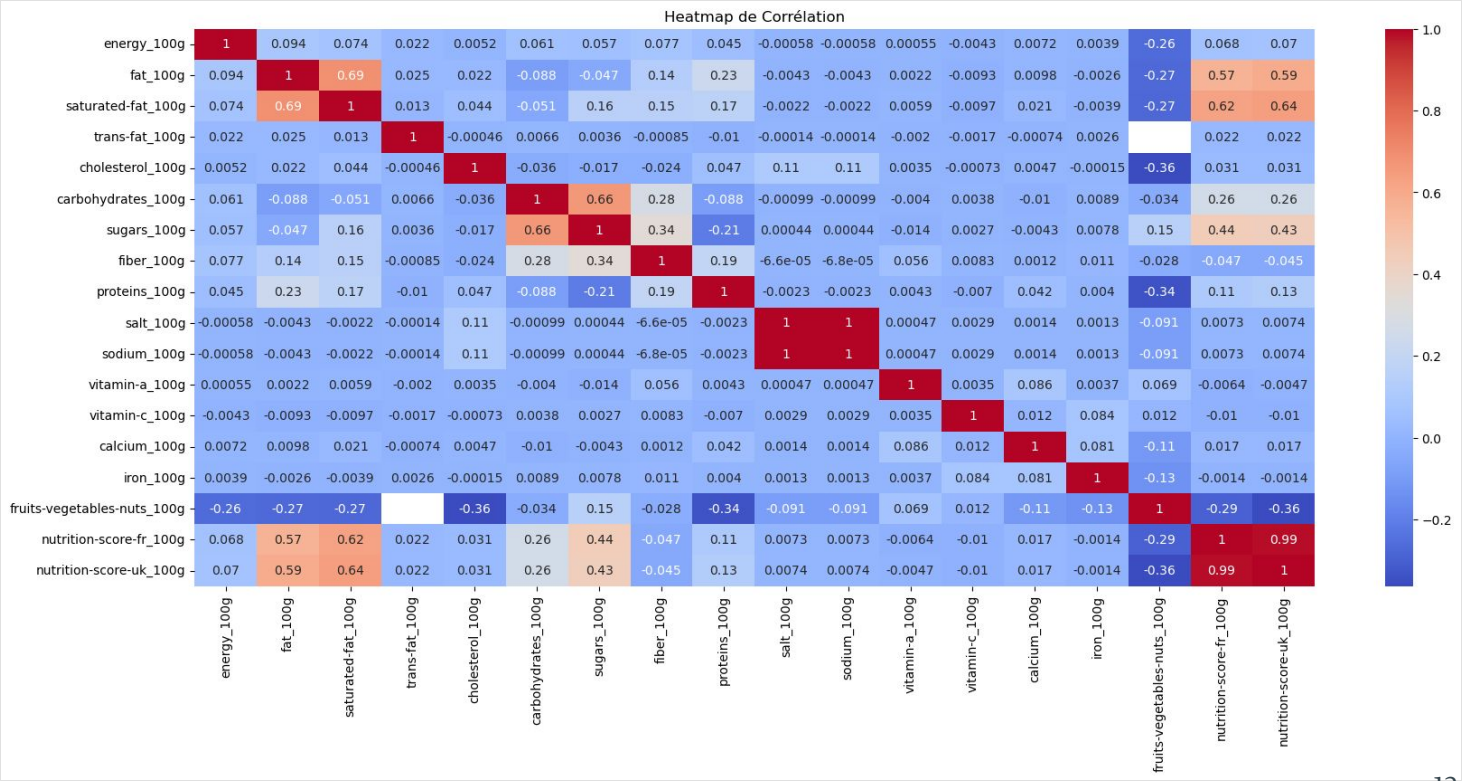

Boîte à moustaches - fruits-vegetables-nuts_100g
100
80
60
40
20
0
fruits-vegetables-nuts_100g


Boîte à moustaches - fiber_100g
100
80
60
40
20
0
fiber_100g


Boîte à moustaches - sugars_100g
100
80
60
40
20
0
sugars_100g

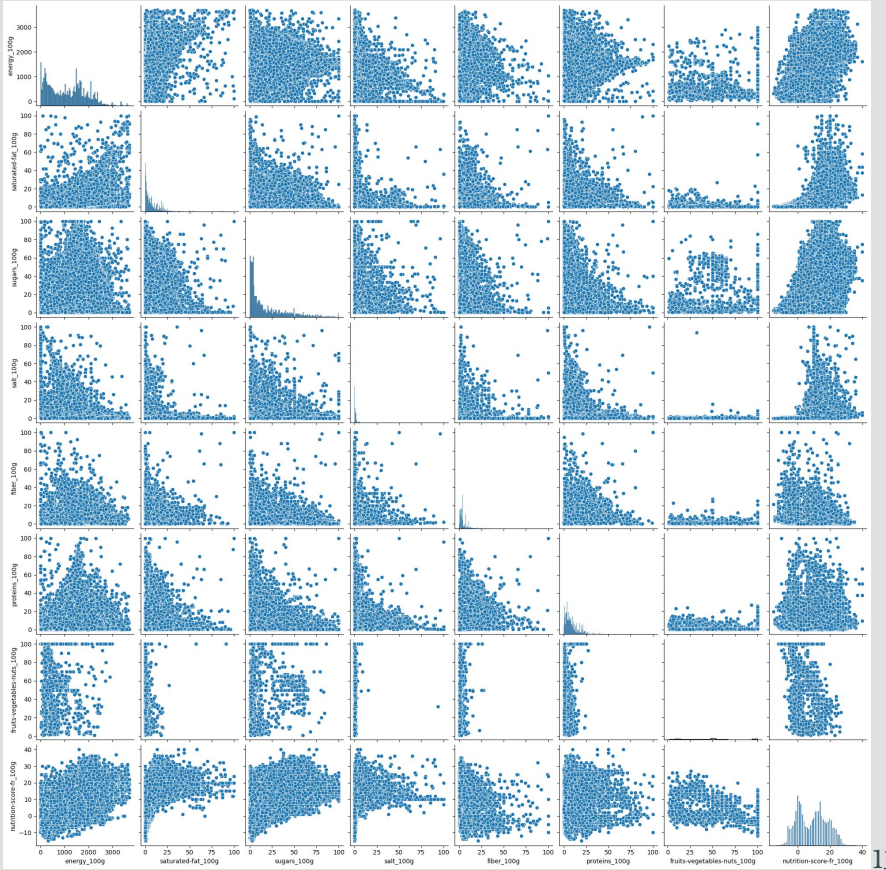
Exploration des données

A la recherche de corrélation (ça nous a aussi permis d'écarter certaines valeurs)



A la recherche de corrélation

Si on peut faire quelques déductions, il est cependant très difficile d'avoir une compréhension détaillée sur comment les différentes valeurs nutritionnelles se comportent entre elles.



Modélisation

Etude des moyennes de différents groupes

Une autre étude que nous faisons en data-science est l'étude de la variance (Méthode Anova).Le test de l'ANOVA est utilisé pour déterminer si les moyennes de plusieurs groupes sont statistiquement différentes les unes des autres.

Ce test m'a permis de conclure qu'il y a **très peu de différence entre les moyennes de l'énergie** des différents groupes.

```
Valeur F : 0.04882321605974633
p-value : 0.9999999999999999
```

En résumé, avec une "Valeur F" proche de zéro et une "p-value" très proche de 1, nous pouvons conclure qu'il n'y a pas de différence significative entre les énergies des deux groupes et que toute petite variation observée est très probablement due au hasard. C'est pour cette raison que je décide de compléter mon set précédent par la moyenne des énergies pour les produits qui n'ont pas d'énergie

En revanche, ce n'est pas le cas pour **le nutriscore** par exemple :

```
Valeur F : 10.730730554555803
p-value : 0.0
```

Dans notre cas, avec une "Valeur F" de 10.730730554555803 et une "p-value" de 0.0, nous pouvons conclure que les scores nutritionnels des différents groupes (groupe A et groupe B) sont significativement différents, avec une grande confiance dans cette conclusion.

D'autres test de statistiques directement sur le livrable ici. Je vous invite à lancer http://localhost:8866/

Evaluation et modélisation

Imputation des valeurs manquantes

Energie, fibres et FLN	Compléter les valeurs des énérgies manquantes par leur moyenne ou leur médiane
Sucres, graisses saturées et protéines	Utilisation de la méthode itérative imputer pour compléter les valeurs nulles
Sel, Nutriscore	Utilisation de la méthode KNNimputer pour compléter les valeurs nulles
Next steps	Le but de cette imputation est de pouvoir compléter notre base de données afin de pouvoir faire appel à l'ACP et non pour faire de la prévision

Analyse de composantes principales (ACP)

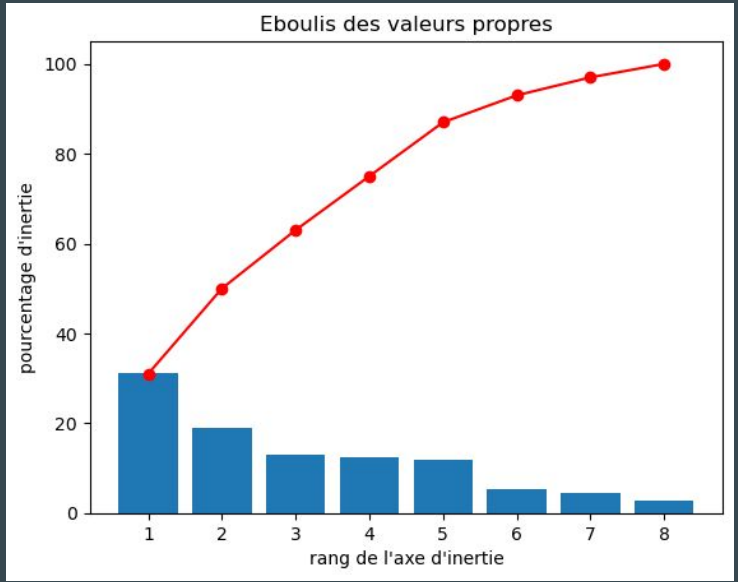
Définition	L'Analyse en Composantes Principales (ACP) est une technique statistique largement utilisée pour transformer des données multidimensionnelles complexes en un ensemble de nouvelles variables, appelées "composantes principales" (F1,F2,F3...), qui capturent l'essentiel de la variabilité des données d'origine. L'objectif principal de l'ACP est de réduire la dimensionnalité des données tout en préservant autant d'informations que possible.
Scaling	Mise à l'échelle nos données. Cela permet de centrer les données autour de zéro et de leur donner une variance unitaire.
Analyse de la variance captée	Chacune de ces nouvelles composantes va venir expliquer ou détenir un certain nombre d'informations de notre set. Le calcul de la variance de chaque composante va nous permettre de calculer le % d'info détenue dans chaque composante

Analyse de composantes principales (ACP)

Analyse de la variance captée

Chacune de ces nouvelles composantes va venir expliquer ou détenir un certain nombre d'informations de notre set. Le calcul de la variance de chaque composante va nous permettre de calculer le % d'info détenue dans chaque composante

On voit ici que près de 80% de la variance est comprise dans les 4 premières composantes, et près de 90% dans les 5 premières. On remarque aussi que les 7 premières composantes définissent automatiquement la dernière.



Analyse de composantes principales (ACP)

Composantes

Mais concrètement, que représentent ces “composantes” ?
Lucky us, l’outil nous fournit aussi une matrice qui nous permettra de calculer chaque composante à partir de nos données initiales

La première composante F1 est calculée ainsi :

F1 = (0.55 *energy_100g) + (0.52* saturated-fat_100g) + ... + (0.5 * nutrition-score-fr_100g)

et F2 ?

F2 = (0.12 *energy_100g) + (0.07* saturated-fat_100g) + ... + (-0,11* nutrition-score-fr_100g)

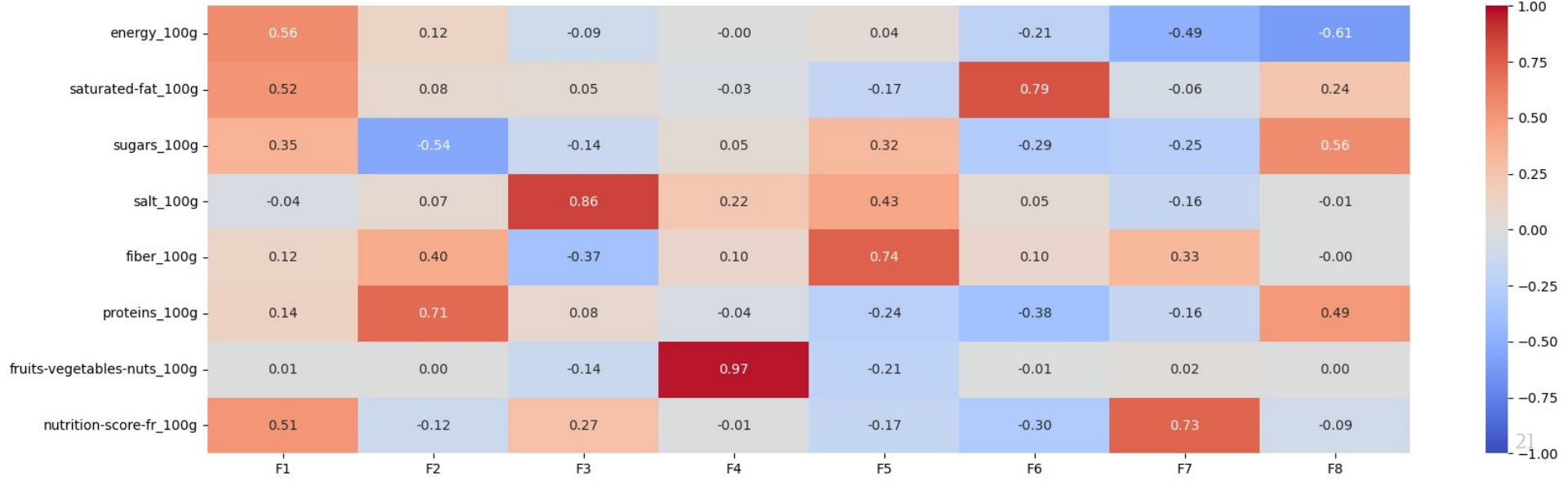
Et ainsi de suite

	F1	F2	F3	F4	F5	F6	F7	F8
energy_100g	0.557704	0.120494	-0.094855	-0.000886	0.041684	-0.214548	-0.491613	-0.613187
saturated-fat_100g	0.520516	0.076673	0.052567	-0.026636	-0.171902	0.793884	-0.058344	0.237710
sugars_100g	0.352090	-0.541669	-0.135608	0.050353	0.318129	-0.285513	-0.254724	0.560441
salt_100g	-0.035048	0.067290	0.855960	0.218924	0.431089	0.053292	-0.157348	-0.014571
fiber_100g	0.119569	0.404397	-0.371834	0.098041	0.744351	0.097488	0.332731	-0.004677
proteins_100g	0.138826	0.709789	0.083486	-0.040950	-0.238253	-0.377603	-0.156614	0.494372
fruits-vegetables-nuts_100g	0.005307	0.002577	-0.140205	0.968171	-0.206509	-0.005518	0.016602	0.000205
nutrition-score-fr_100g	0.509153	-0.119687	0.268875	-0.013498	-0.167927	-0.295367	0.727844	-0.093616

Analyse de composantes principales (ACP)

Composantes

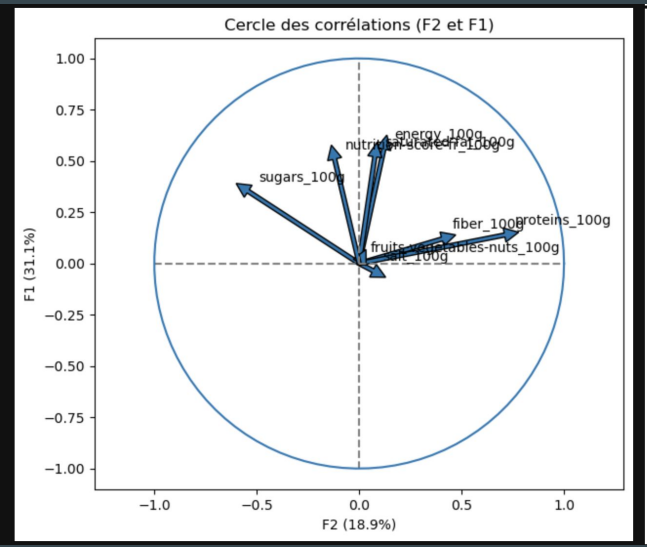
Cette heatmap nous permet aussi d'identifier les valeurs nutritionnelles qui ont le plus gros "poids" sur chaque composante.
Ainsi, pour la composante F1, cela représente "grossièrement" les produits avec un taux de graisse et de sucres importantes
F2 représenterait plus les produits avec beaucoup de proteines et de fibres

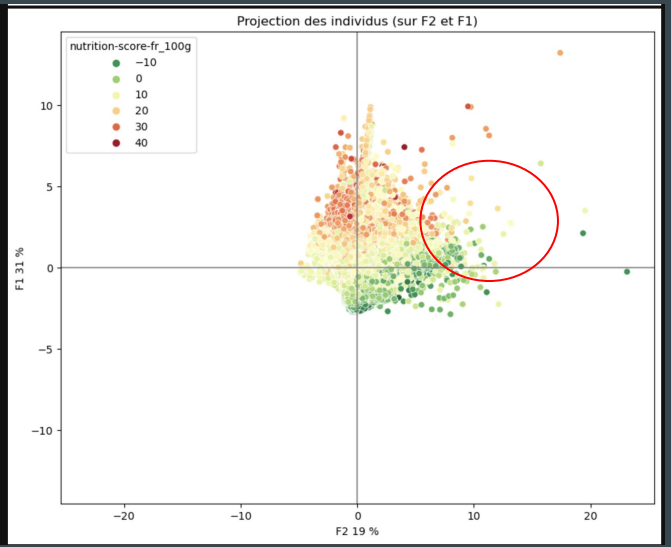


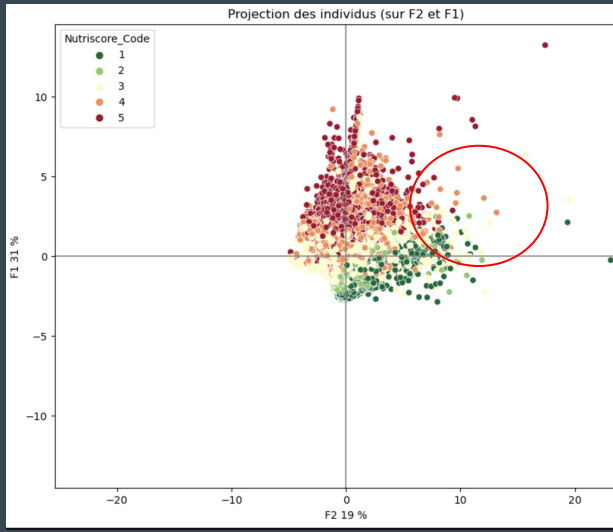
Analyse de composantes principales (ACP)

Modélisation F1 /F2

On remarque qu'ici que les graisses saturées sont indépendantes des protéines et des sucres
Les proteines quand à eux sont +/_ inversement corrélés aux sucres (+ y a des protéines, mois y a de légumes).
On remarque que les produits avec un pourcentage de protéines importants sont mal interprétés par le nutriscore



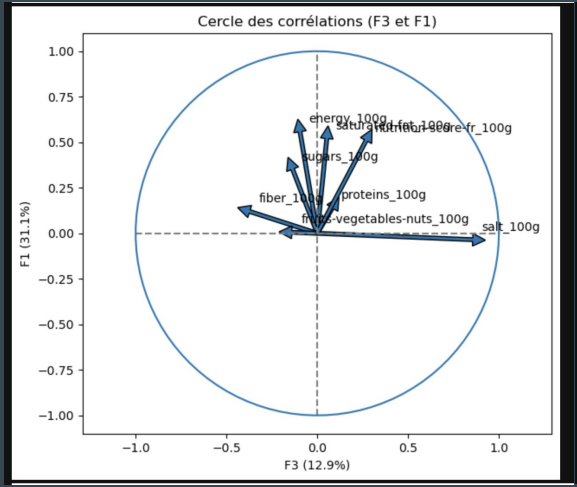


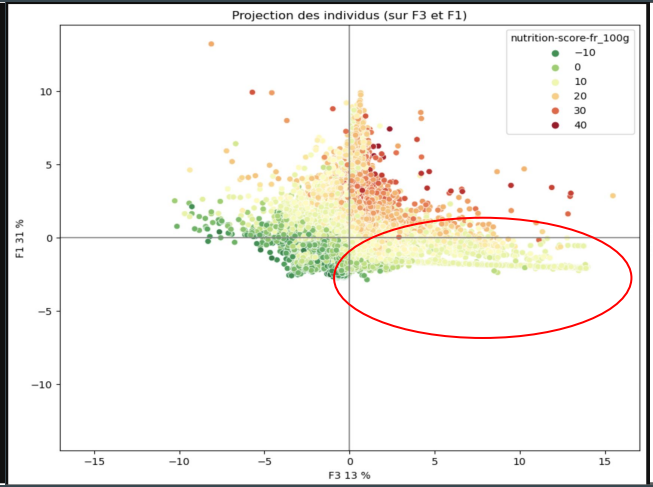


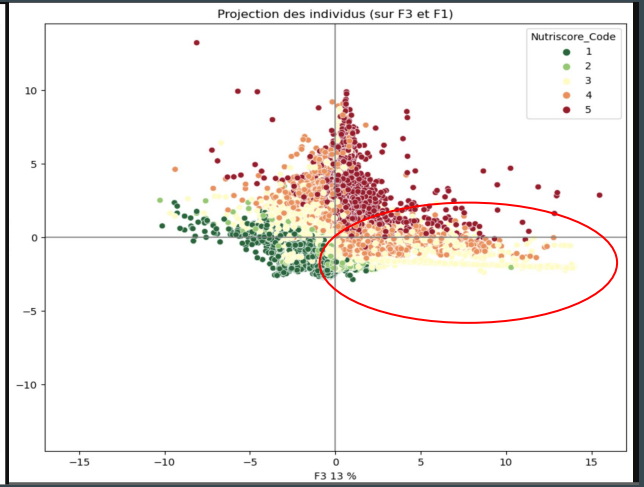
Analyse de composantes principales (ACP)

Modélisation F1 /F3

On va principalement étudier sur ce graphe l'impact du sel sur le nutriscore. Car la corrélation entre le sel et les autres valeurs n'est pas pertinentes.
On remarque que le sel a finalement très peu d'impact sur le nutriscore malgré l'envie de vouloir faire ressortir ce dernier dans le calcul du nutriscore



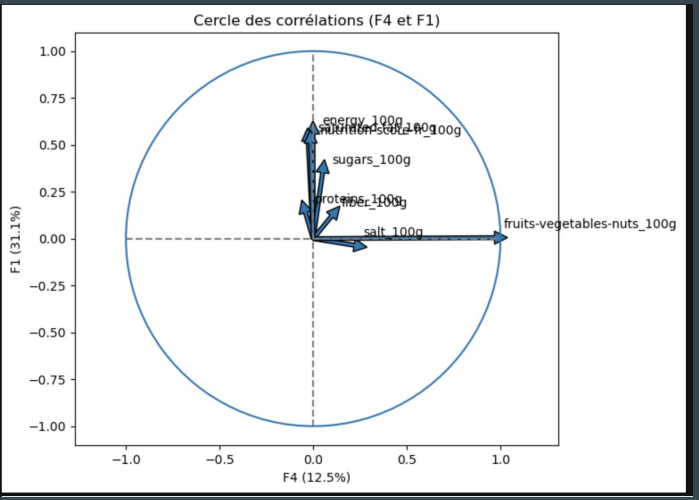


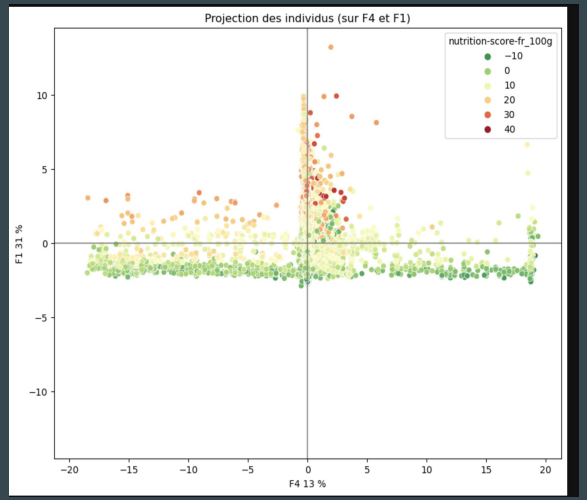


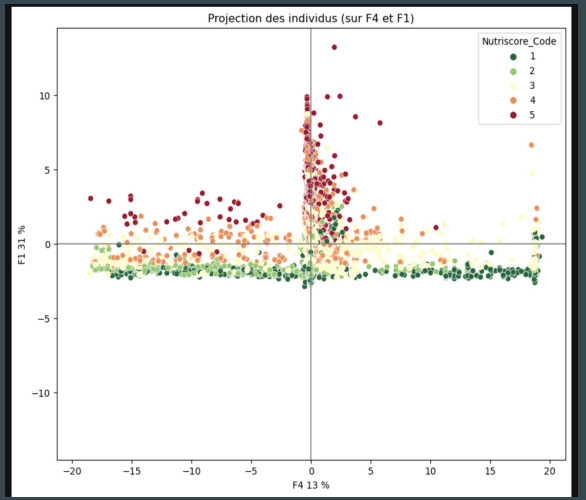
Analyse de composantes principales (ACP)

Modélisation F1 /F4

On va principalement étudier sur ce graphe les fruits, légimes et noix. Car la corrélation entre le FLN et les autres valeurs n'est pas pertinentes.
On remarque que les fruits, légumes et noix sont plutot bien modélisés en terme de nutriscore







Mise en production

Pour aller plus loin, et pouvoir étudier l'impact des différents valeurs nutritionnelles, je vous invite à aller sur http://localhost:8866/