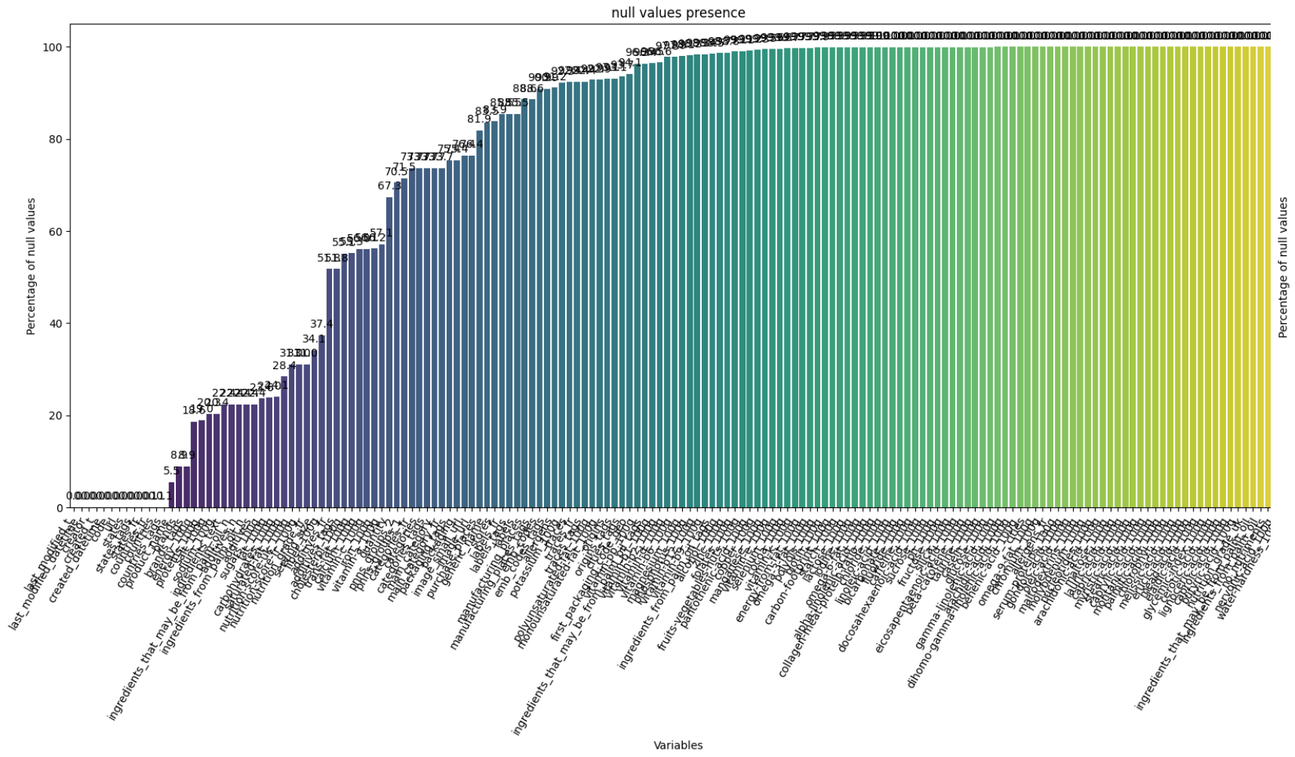
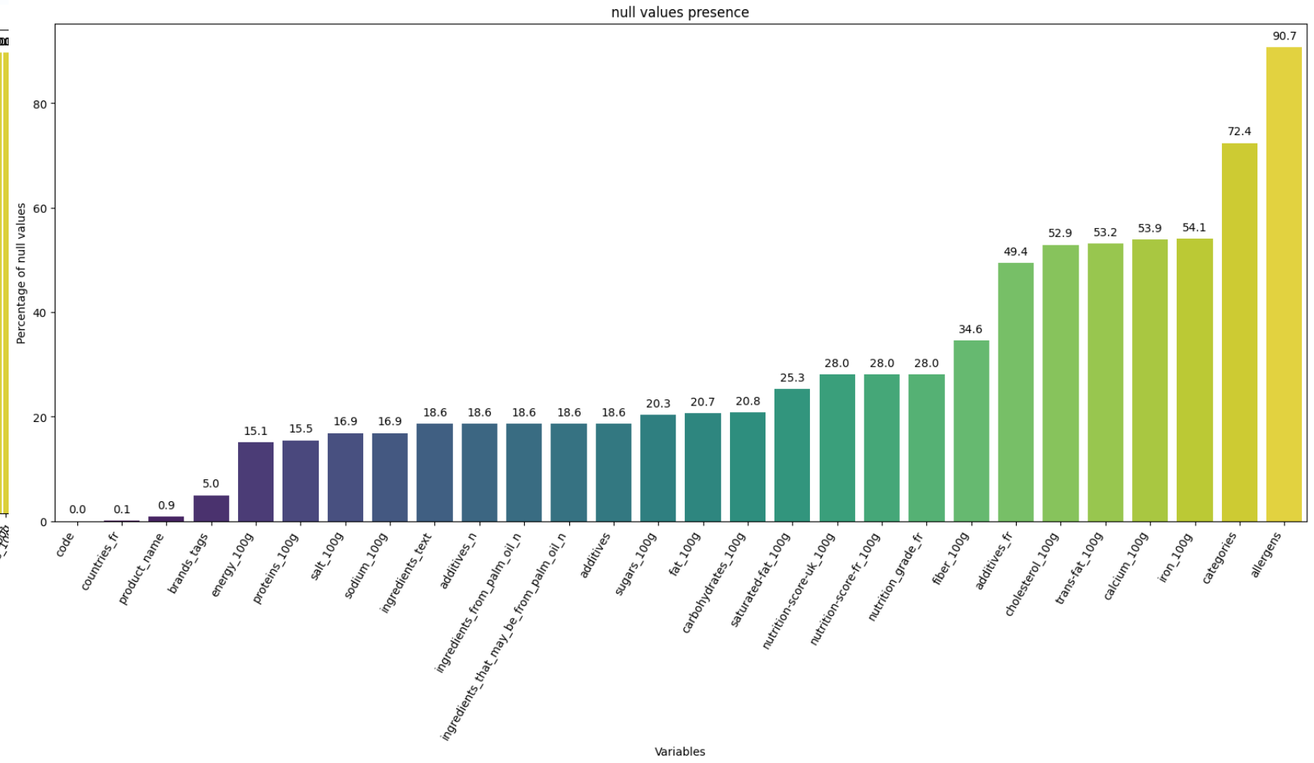
Machine Learning Engineer

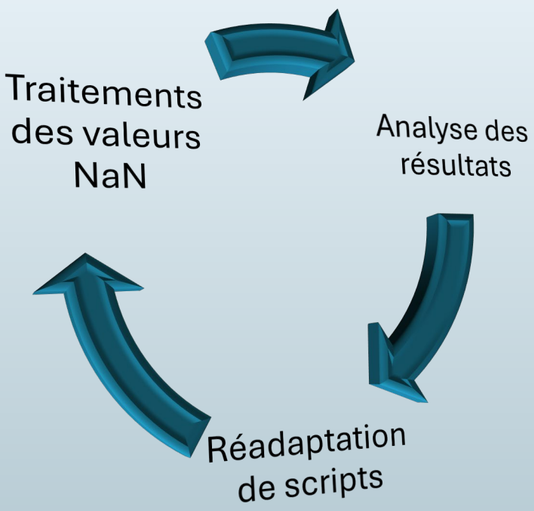
Préparez des données pour un organisme de santé publique

Faisabilité d’une application pour la gestion d’ajout des données openFood

Machine Learning Engineer





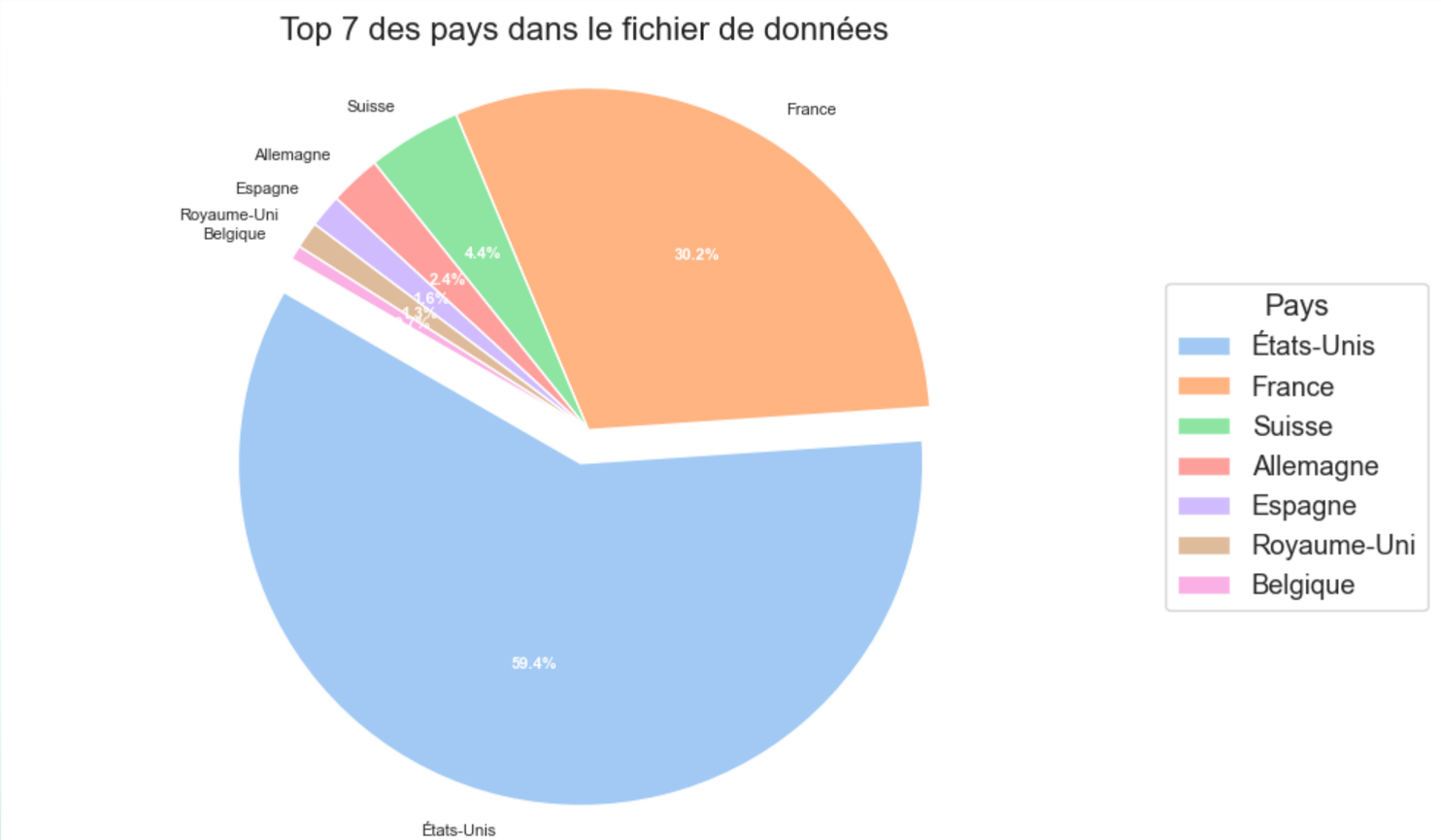


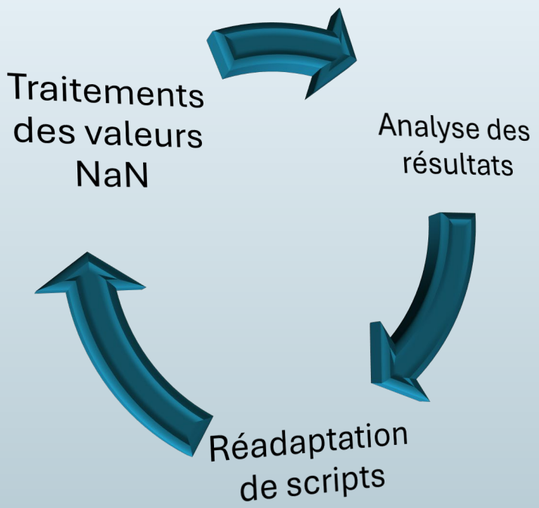
- Traitement large des valeurs null sur les données
- Respect de la RGPD dans la conservation de certaines données

La préparation des données
Suppression et Imputation de données

Machine Learning Engineer

- Conservation des 7 pays les plus alimentés en données

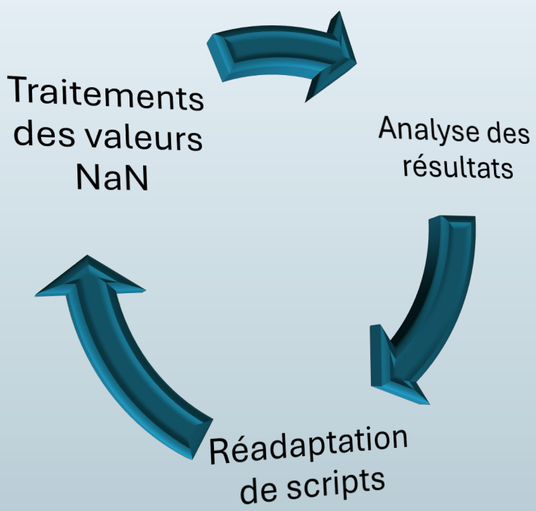




La préparation des données
Suppression et Imputation de données

- Traitement des doublons : chaque produit est unique au sein de sa marque

- .Nettoyage de la colonne Ingredient_text dans une démarche d’anticipation

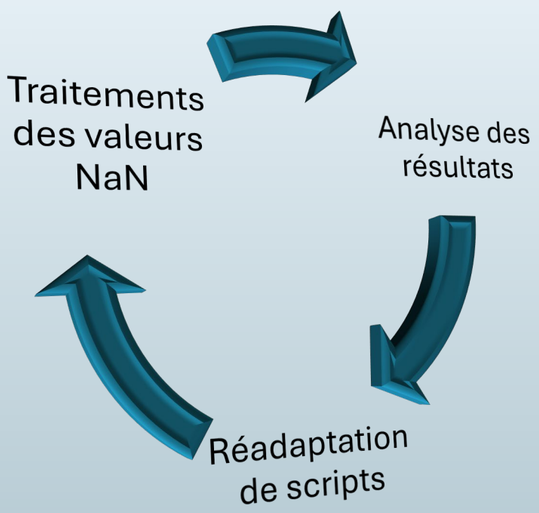


La préparation des données
Suppression et Imputation de données

- Traitement général de valeur null

- Limitation générale des valeurs abérrantes

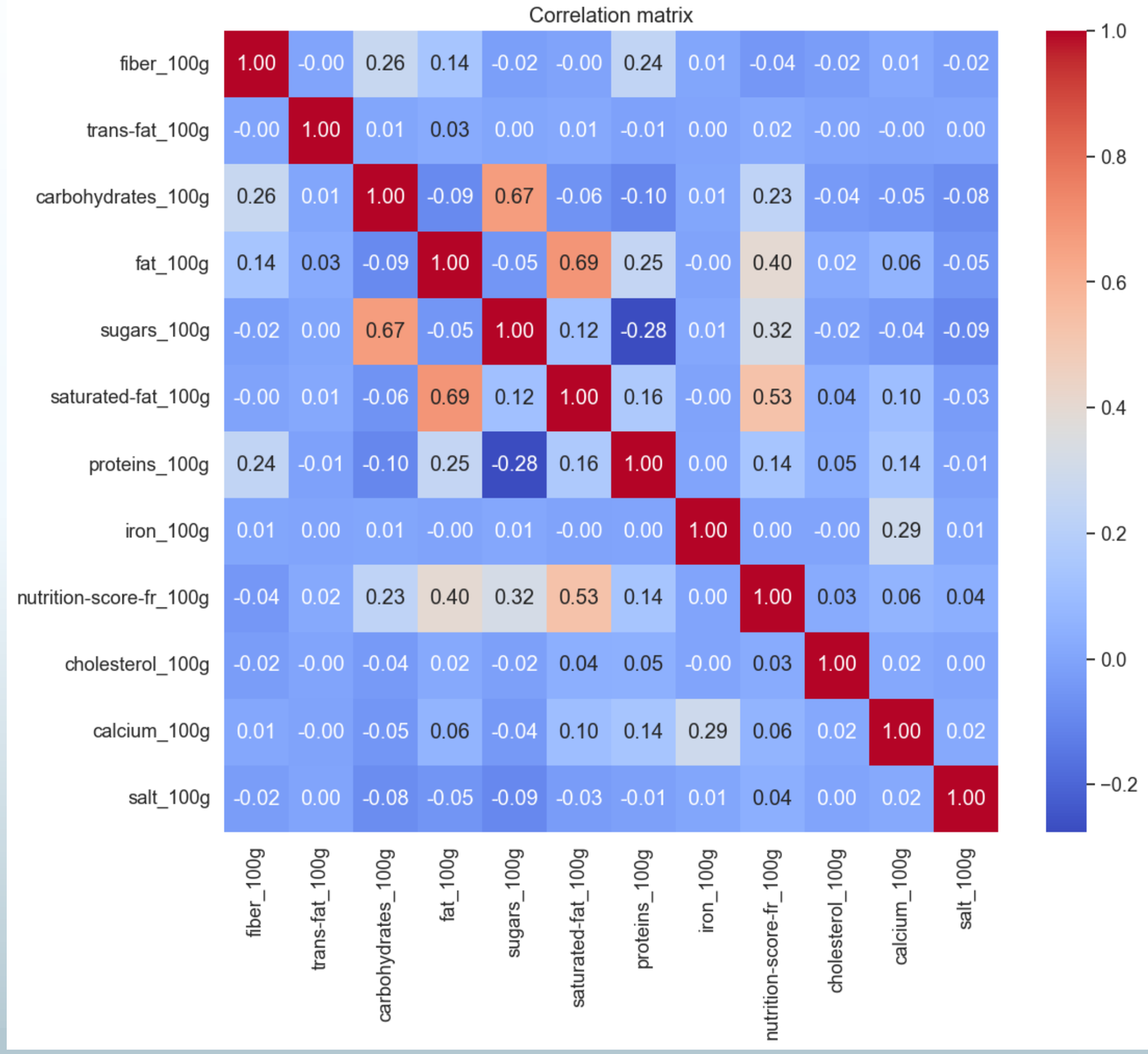
  - Les calories joules ne sont pas supérieurs à 3900 unités pour 100g
  - Les variables numériques pour 100g ne peuvent pas être inférieur à 0 ni supérieur à 100

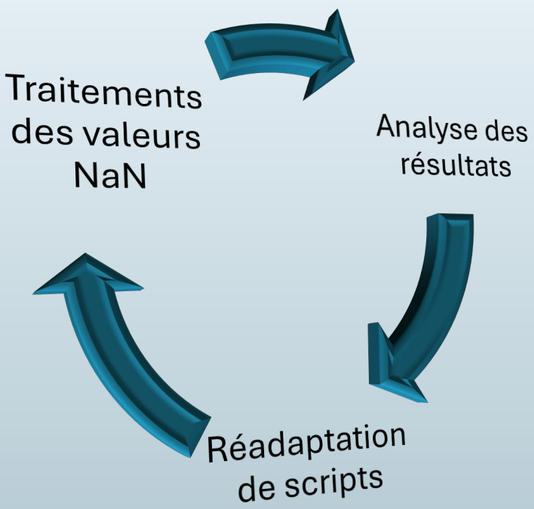


La préparation des données
Suppression et Imputation de données

Machine Learning Engineer

Matrix de Correlation





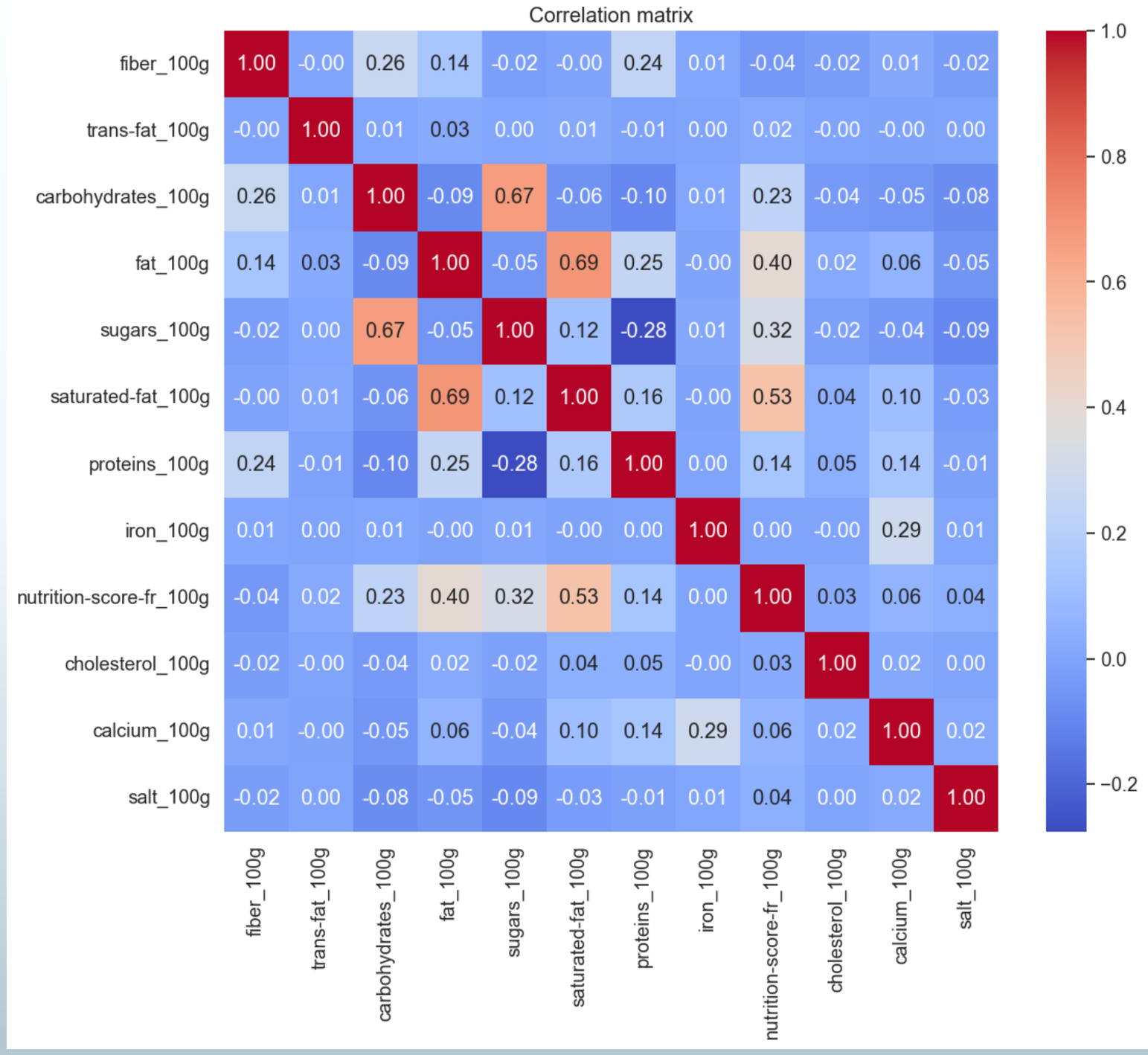
Vérification de la relation entre les données
Analyse et Imputation de données

Machine Learning Engineer

La matrice de corrélation éclaire les liens entre les attributs nutritionnels des aliments. Les interdépendances, telles que l'influence marquée des graisses sur le score nutritionnel, permettront des ajustements ciblés.

Identifier les fortes corrélations contribue à améliorer la précision des données saisies. La collecte et la maintenance des données seront optimisé pour une base de données efficaces.

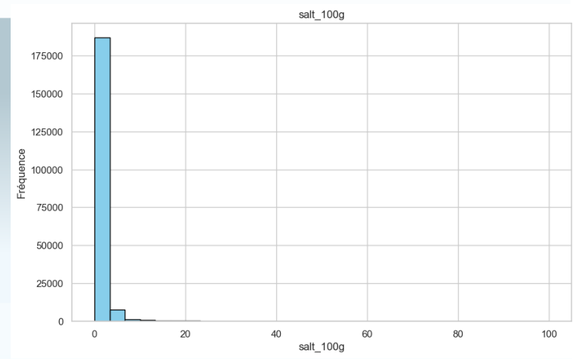
Les insights révélées serviront aussi à la conception de l'interface utilisateur en anticipant les prochaines valeurs saisies.

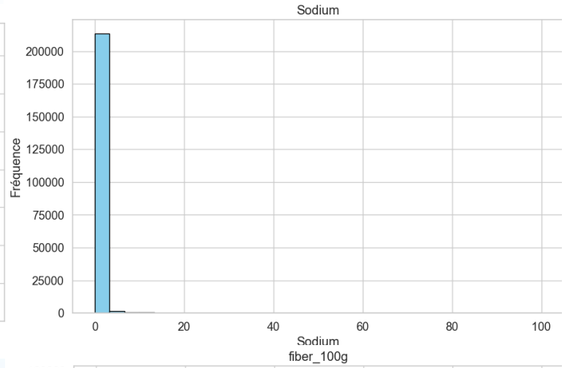


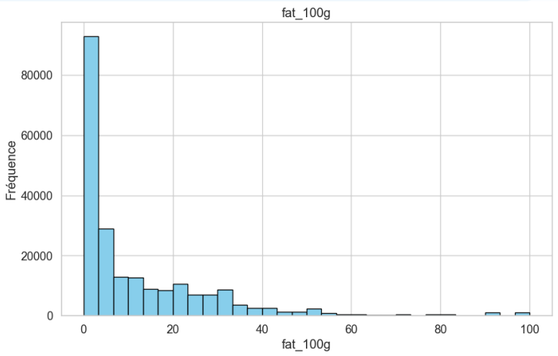
Vérification de la relation entre les données
Analyse et Imputation de données

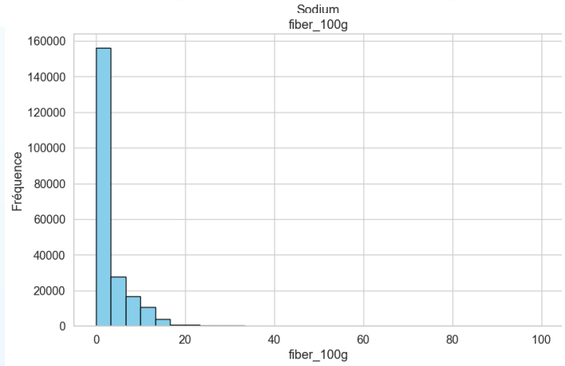
Machine Learning Engineer

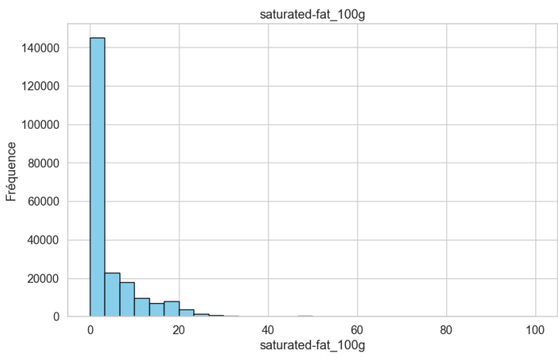
Imputation des valeurs Nan



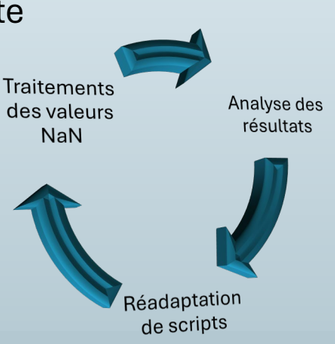


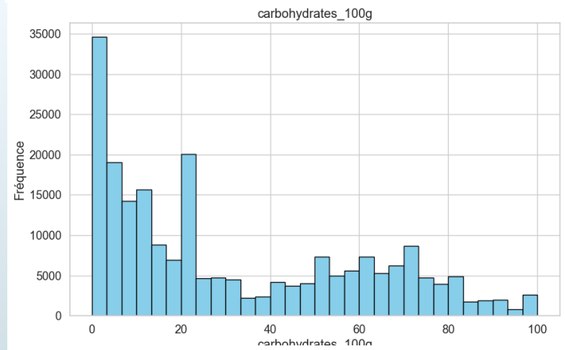


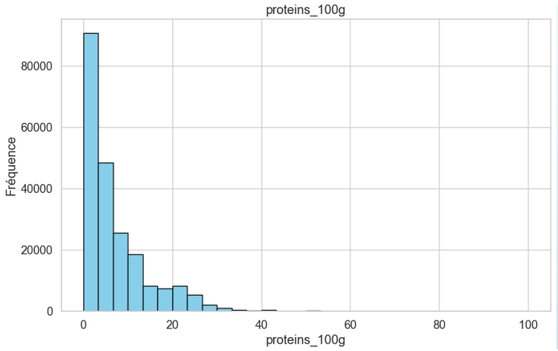


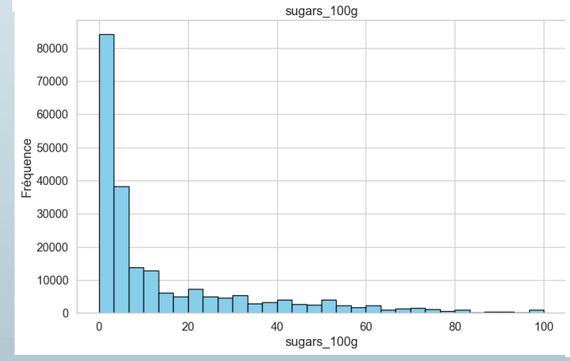


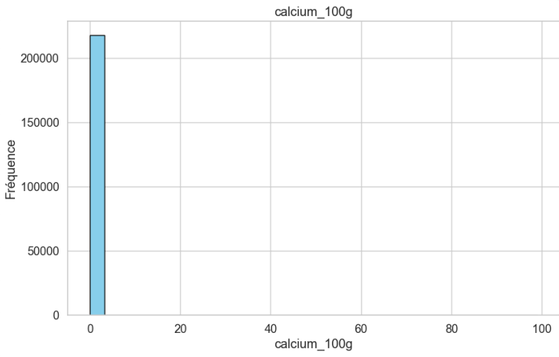
- La variable salt_100g : la médiane dont la moitié est diminué si le sucre est conséquent pour ses valeurs NaN
- La variable sodium_100g : un ratio de 40% du poids du sel
- Les variables sugar_100g, carbohydrates_100g, fiber_100g, leur valeur NaN sont imputé en relation les un avec les autres.
- Les variables fat_100g et saturated-fat_100g sont correlées, leur imputation se fait par ratio selon les valeurs présentes.
- La variable protein_100g et celle du calcium_100g, leur Nan sont imputé et influencé par la présence du sucre
- La variable Cholesterol_100g : nous estimons que si elle est null, elle est absente









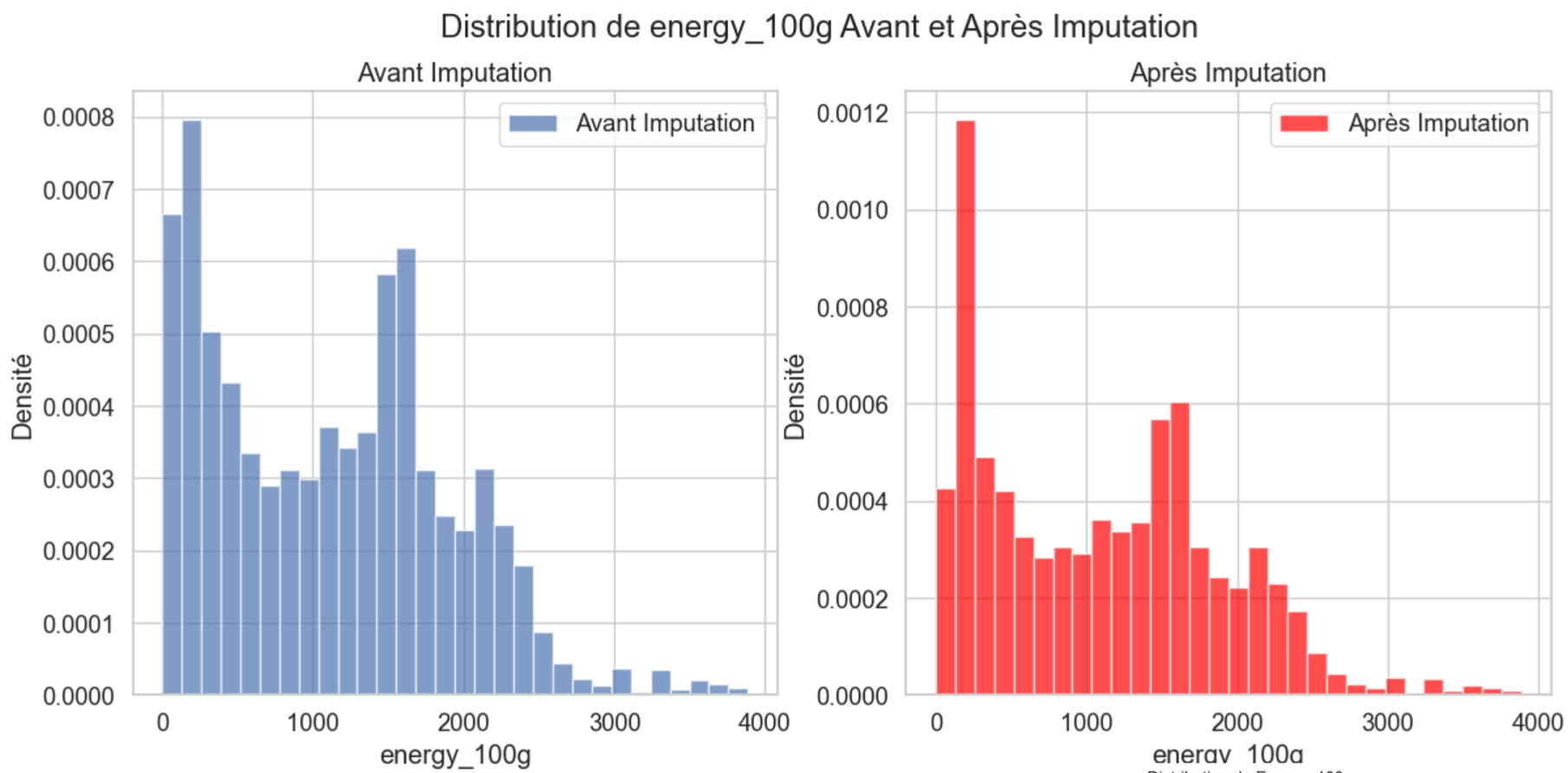


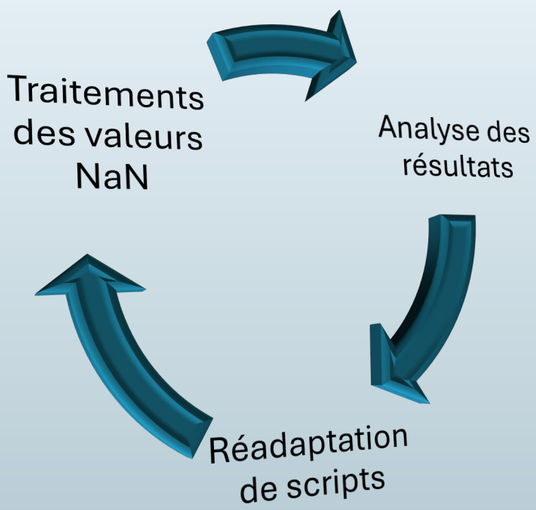
Vérification de la relation entre les données
Analyse et Imputation de données

Machine Learning Engineer

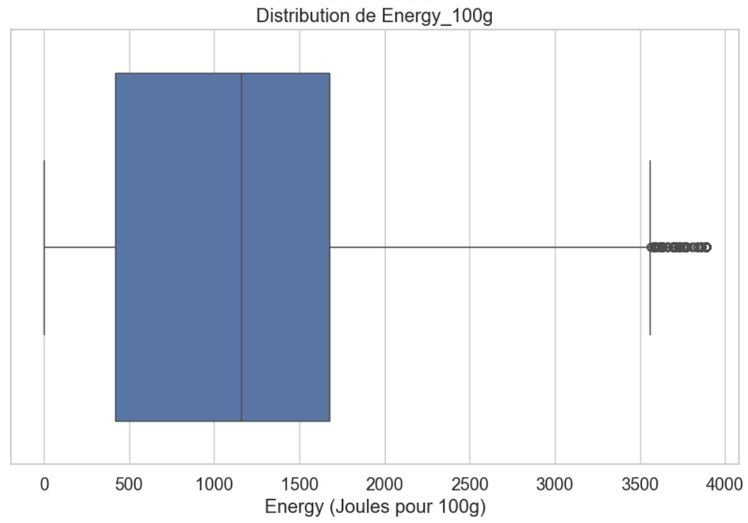
Distribution des calories

- Constat de la presence de 2 unites de mesure : Les joules et calories
- Imputation avec la méthode KNN : influence des plus proches voisins
- Certaines valeurs abérrantes concernent des produits en rapport avec le domaine des huiles, donc légitime





	Unnamed: 0	product_name
364	437	10 pure olive oil mild
554	643	100 california avocado oil garlic
580	669	100 canola oil
626	718	100 extra virgin coconut oil
627	719	100 extra virgin olive oil
...	...	...
211273	246833	westcott vegetable oil
211619	247196	whipped shortening
212469	248064	white truffle oil
215216	250929	x virgin olive oil
217520	253815	zoye premium vegetable oil

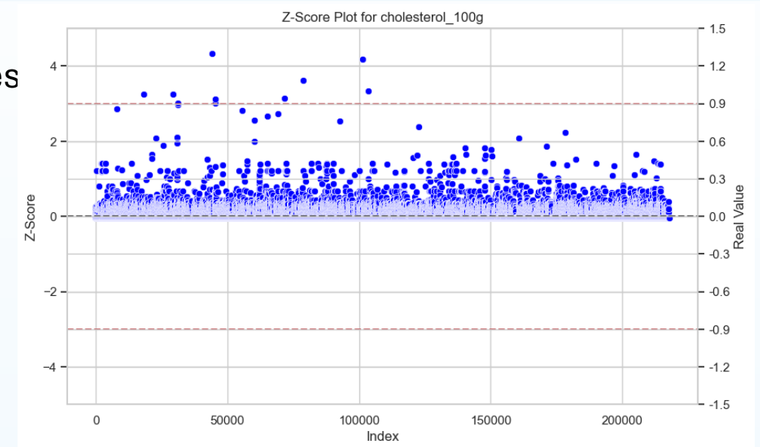


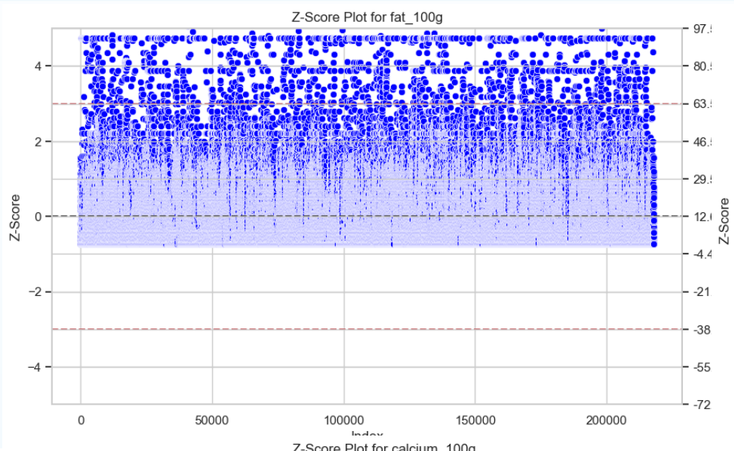
La préparation des données
Analyse et Imputation de données

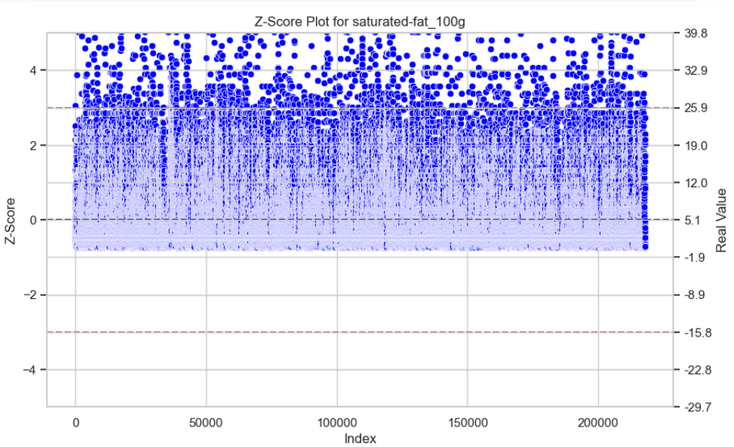
Machine Learning Engineer

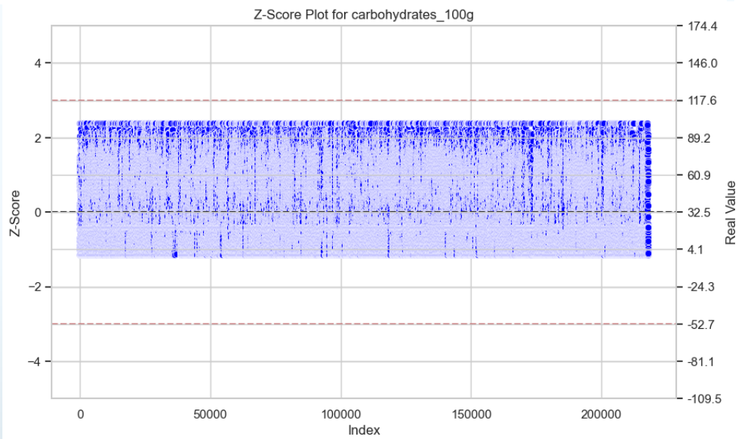
Z-Score

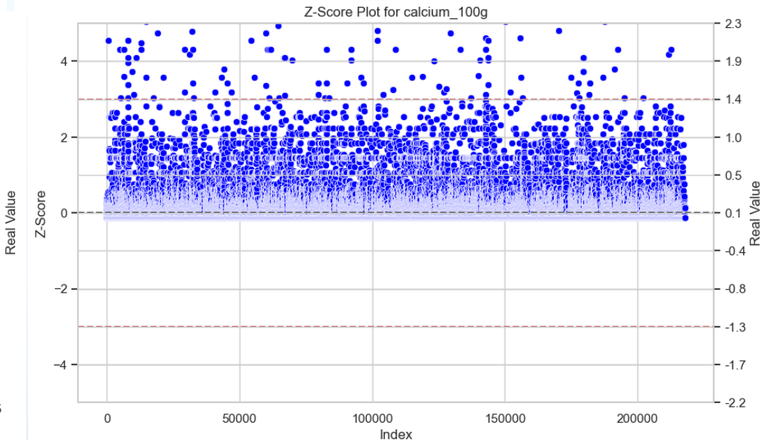
- Travaille sur les valeurs aberrantes :

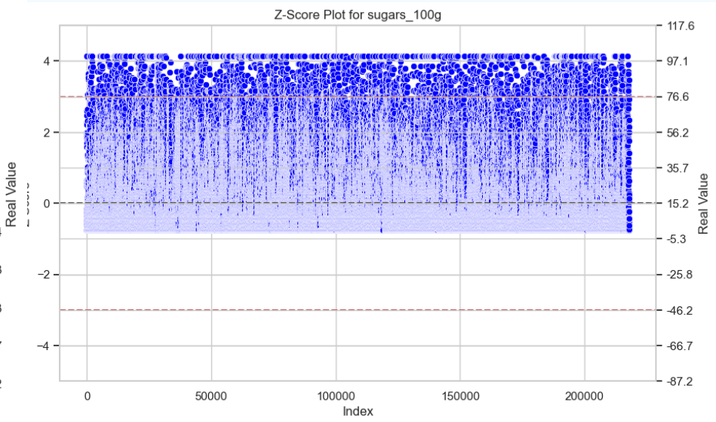


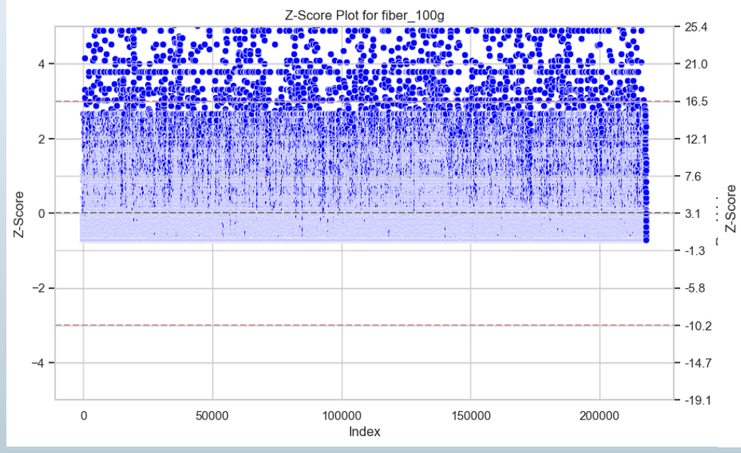


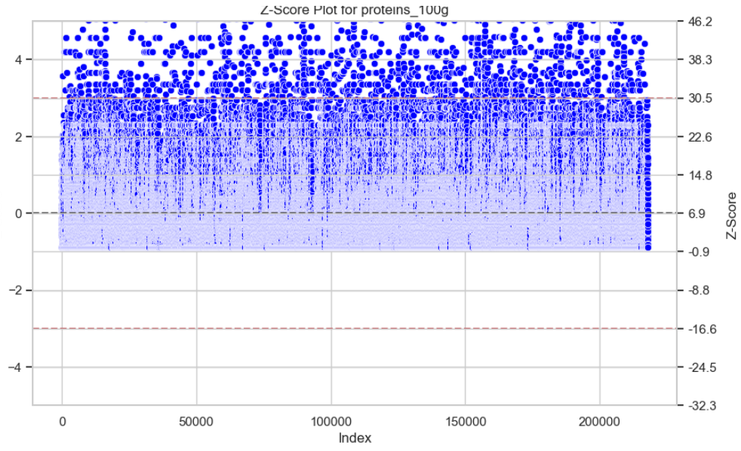


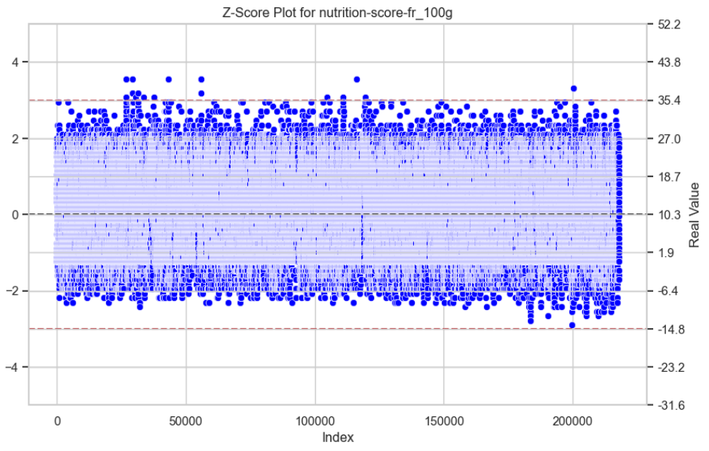












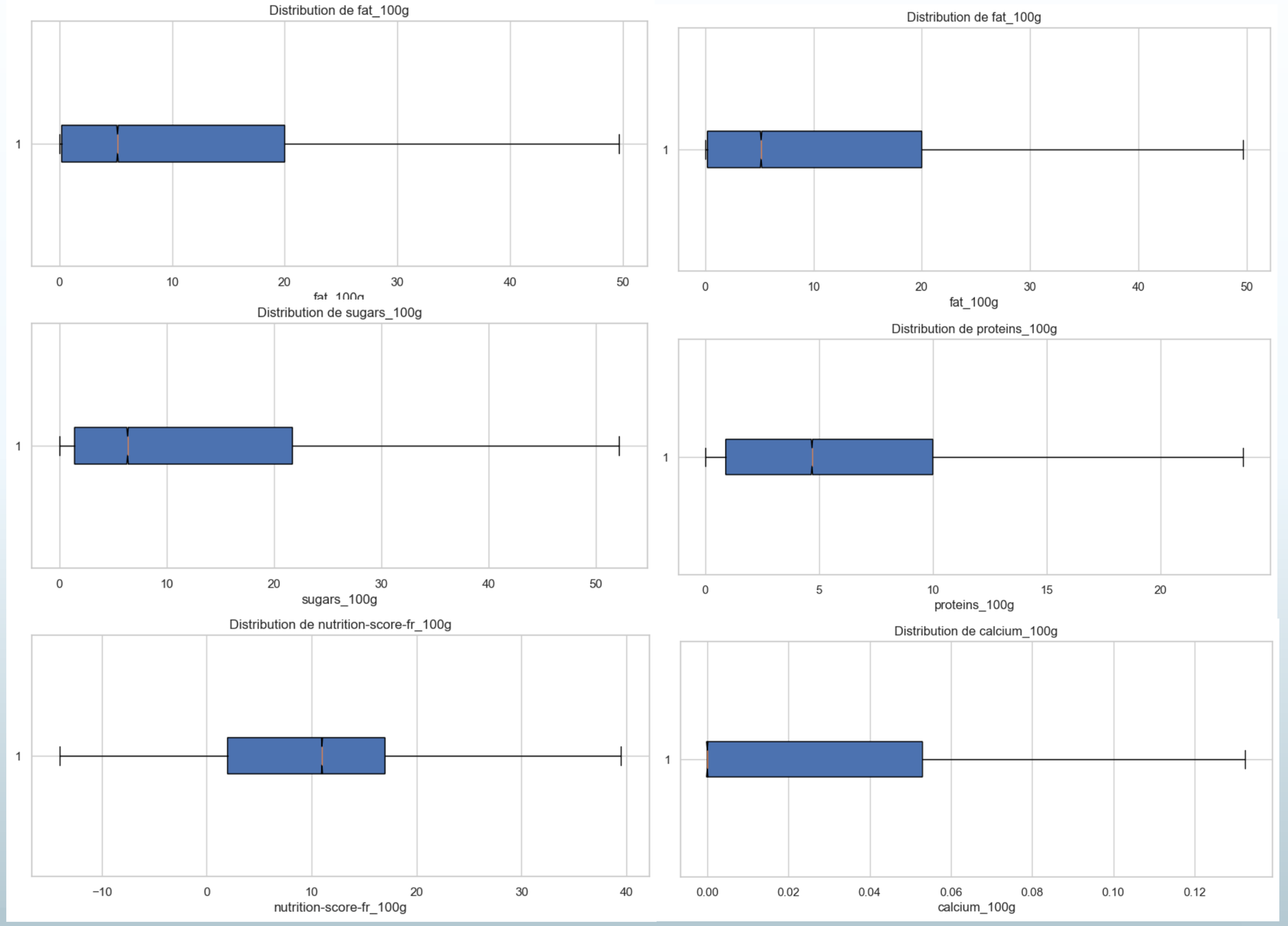
La préparation des données
Analyse et Traitement des valeurs abérrantes

Machine Learning Engineer

Z-Score

- Travail sur les valeurs aberrantes
- Nous repérons la valeur non aberrante la plus forte et nous adaptons l'extréme à cette valeur, cela permet de respecter la variance entre les données

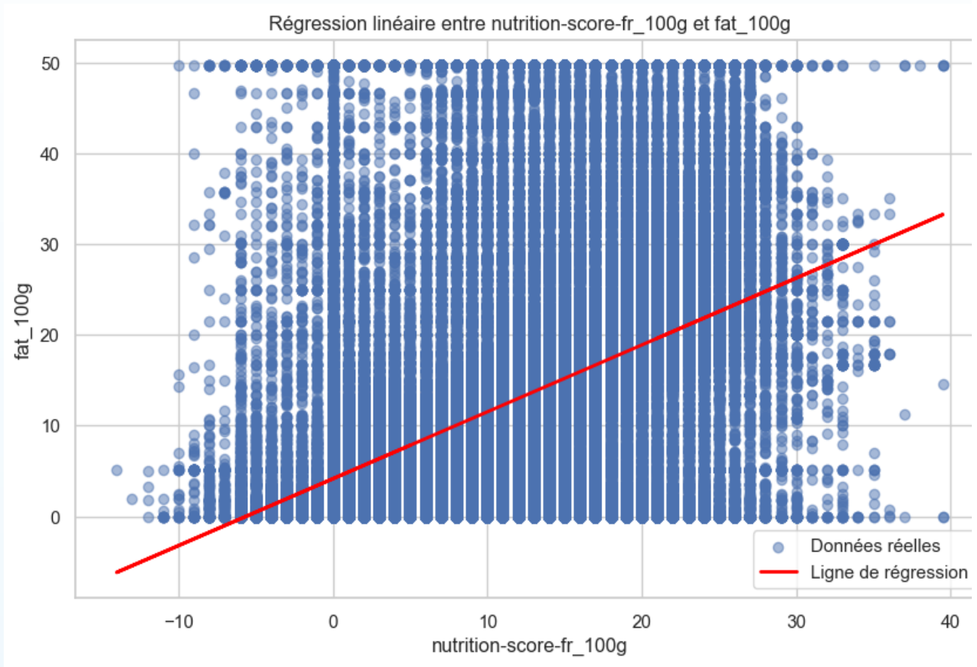
Après traitement :

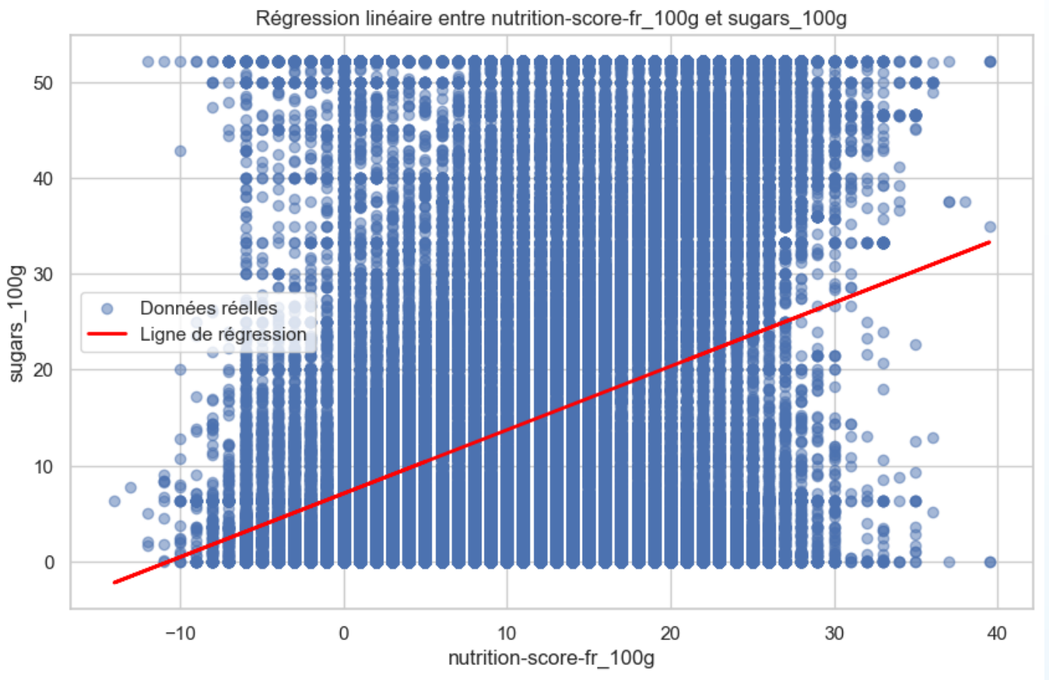


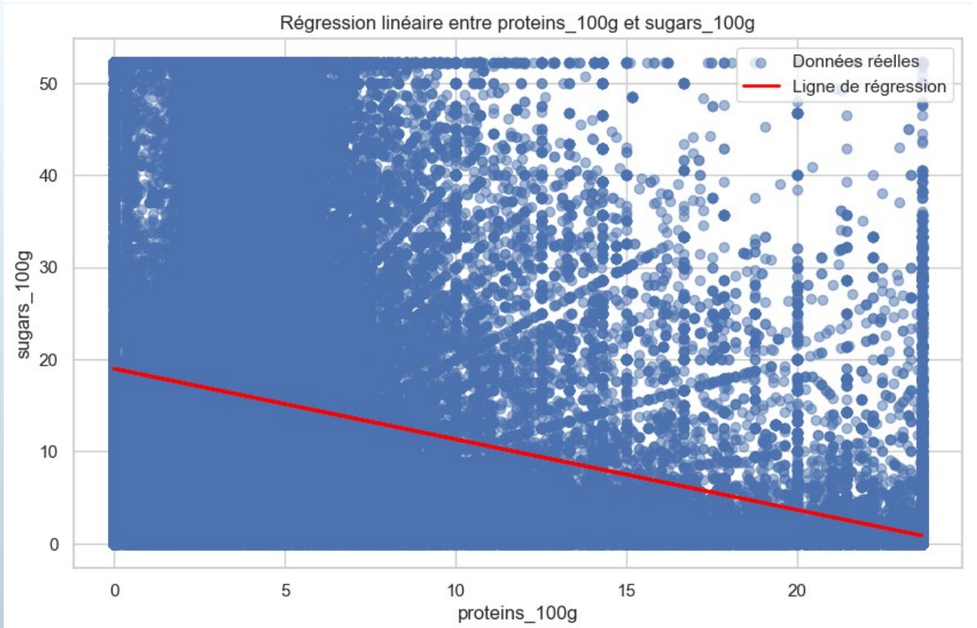
La préparation des données
Analyse et Traitement des valeurs abérrantes

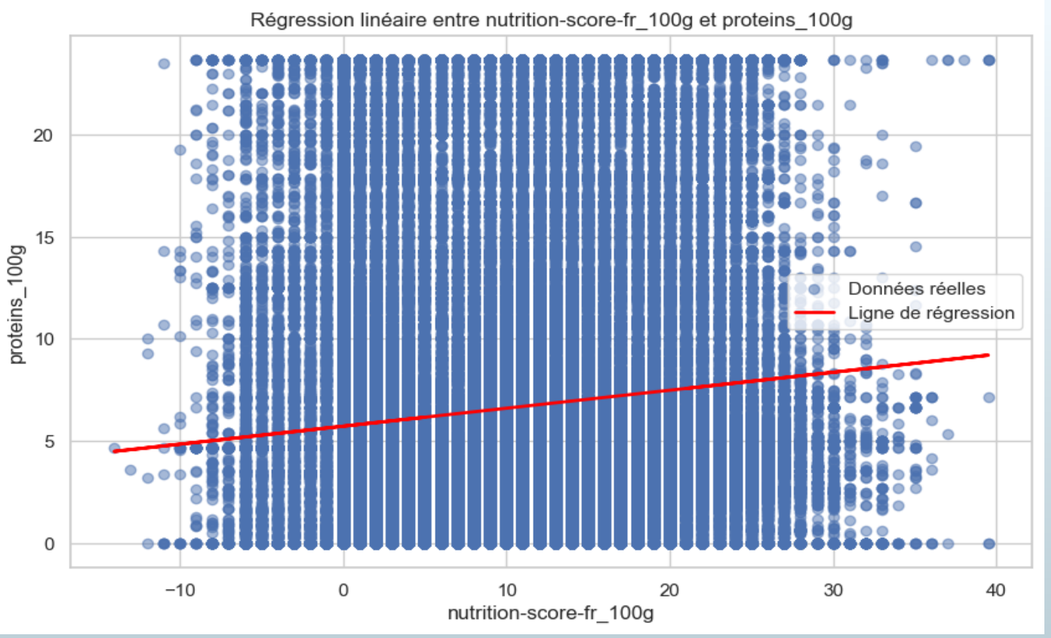
Machine Learning Engineer

- Le gras et le sucre influencent légèrement le score de nutrition
- La proteine ne semble pas impacter le score de nutrition





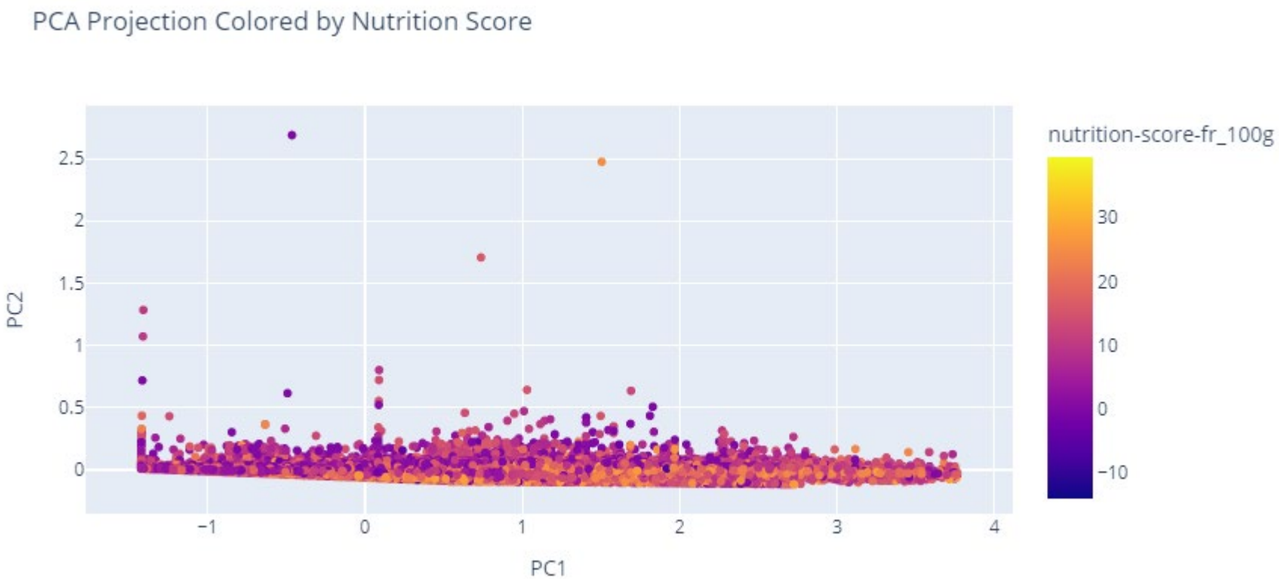




Regression Lineaire

Etude des tendances
Analyse Bivariée

Machine Learning Engineer




```
['fat_100g', 'proteins_100g', 'iron_100g',
                              'cholesterol_100g',
'calcium_100g'], 'nutrition-score-fr_100g', 'product_name')
```


```
['fiber_100g', 'carbohydrates_100g', 'sugars_100g',
'calcium_100g'], 'nutrition-score-fr_100g'
```


```
['fiber_100g', 'carbohydrates_100g', 'sugars_100g', 'calcium_100g', 'fat_100g',
'proteins_100g', 'iron_100g',
                'cholesterol_100g', 'salt_100g'], 'nutrition-score-fr_100g'
```

ACP

Analyse des Composants Principaux

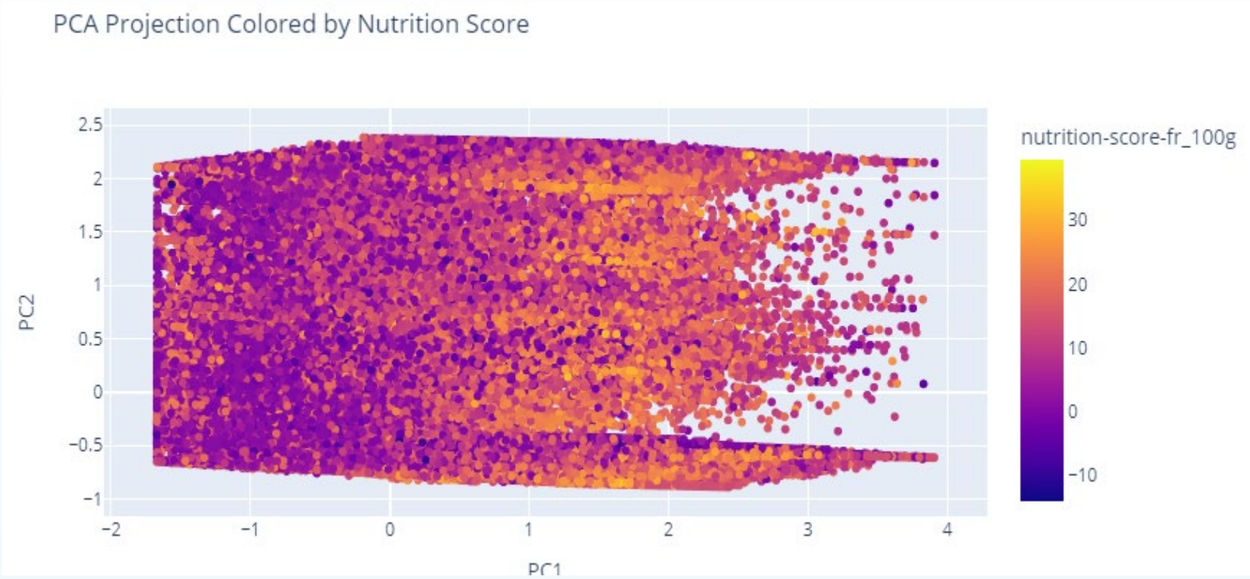
Machine Learning Engineer

L'ACP transforme les variables en un espace ou les axes principaux résument l'essentiel de leurs variations.

Les produits sont affichés dans des couleurs en fonction de leur score nutritionnel. Cette représentation à pour objectif d'identifier les tendances et les outliers.

L'axe horizontal capture une grande partie de la variance. Les produits sont dispersés de gauche à droite, et de haut en bas, ce qui démontre une large couverture des variances.

Il y a une concentration dense de points vers le centre du graphique, la majorité des produits ont des caractéristiques nutritionnelles modérément évaluées.




```
['fiber_100g', 'carbohydrates_100g', 'sugars_100g',
'calcium_100g'], 'nutrition-score-fr_100g'
```


```
['fiber_100g', 'carbohydrates_100g', 'sugars_100g', 'calcium_100g', 'fat_100g',
'proteins_100g', 'iron_100g',
                'cholesterol_100g', 'salt_100g'], 'nutrition-score-fr_100g'
```

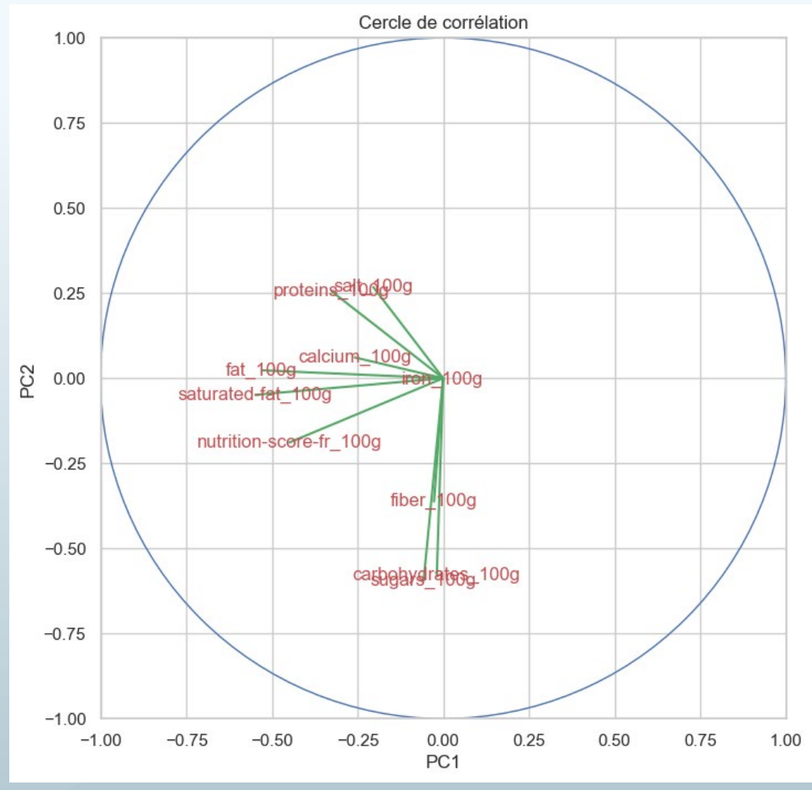
ACP

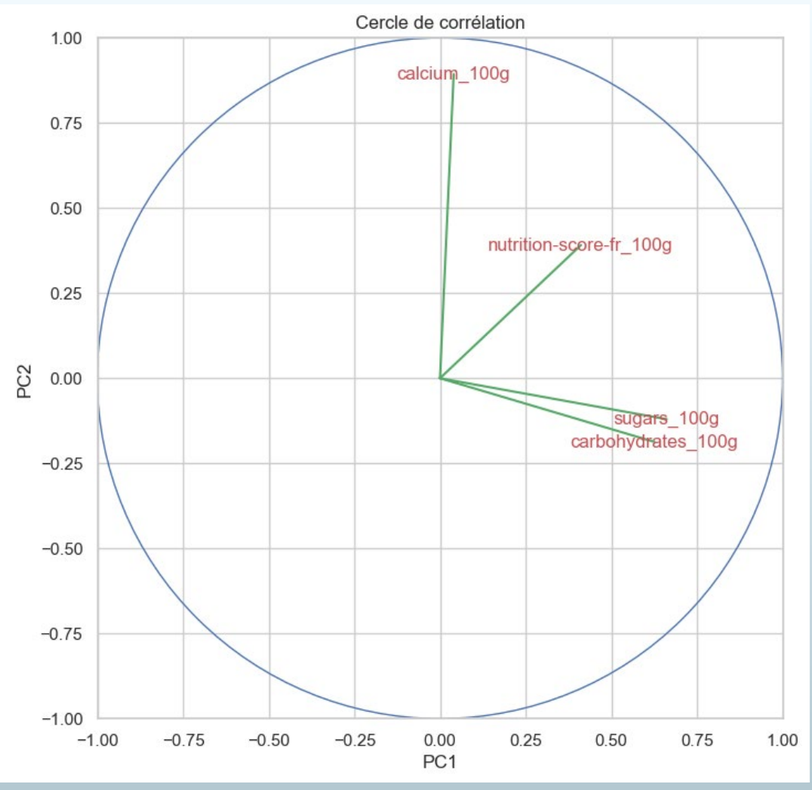
Analyse des Composants Principaux

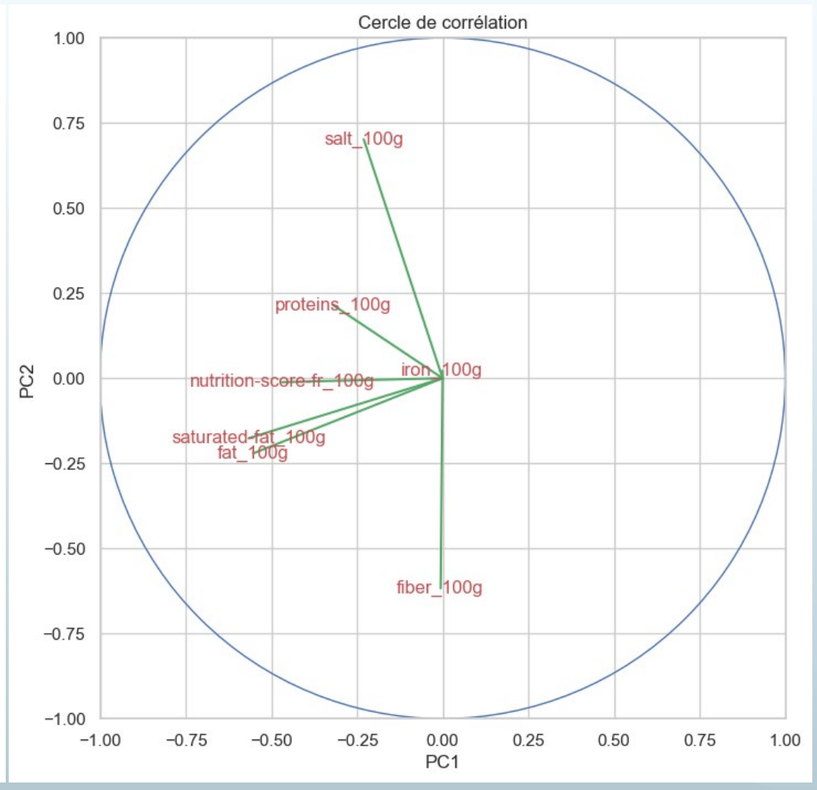
Machine Learning Engineer

Le cercle de corrélation illustre les relations entre les variables nutritionnelles. Une corrélation négative forte entre nutrition-score-fr_100g et fiber_100g indiquerait que les aliments avec plus de fibres ont tendance à avoir un score nutritionnel plus bas.

Ce visuel participe à l'élaboration d'une forme d'anticipation des entrées utilisateurs. Il offre une vue précieuse sur la structure des données nutritionnelles, et les décisions seront, du côté de l'algorithme prise de manière éclairée.



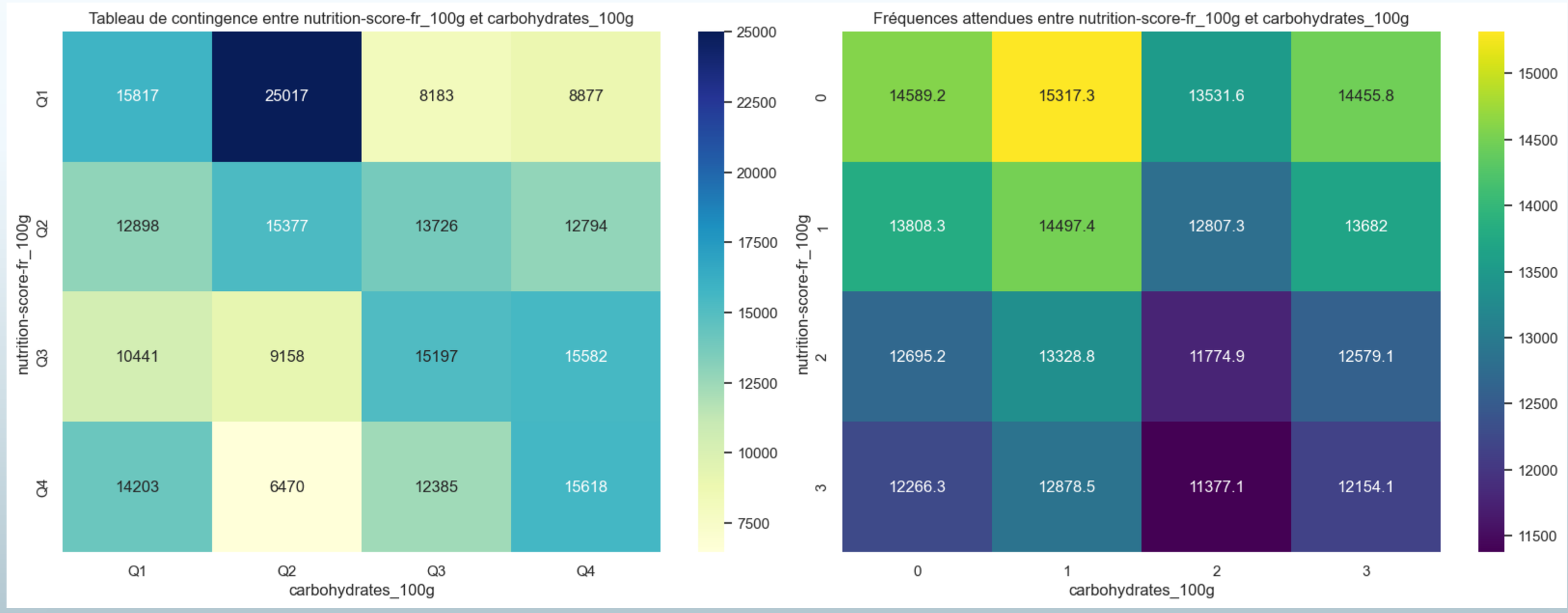




Cercle de Corrélation

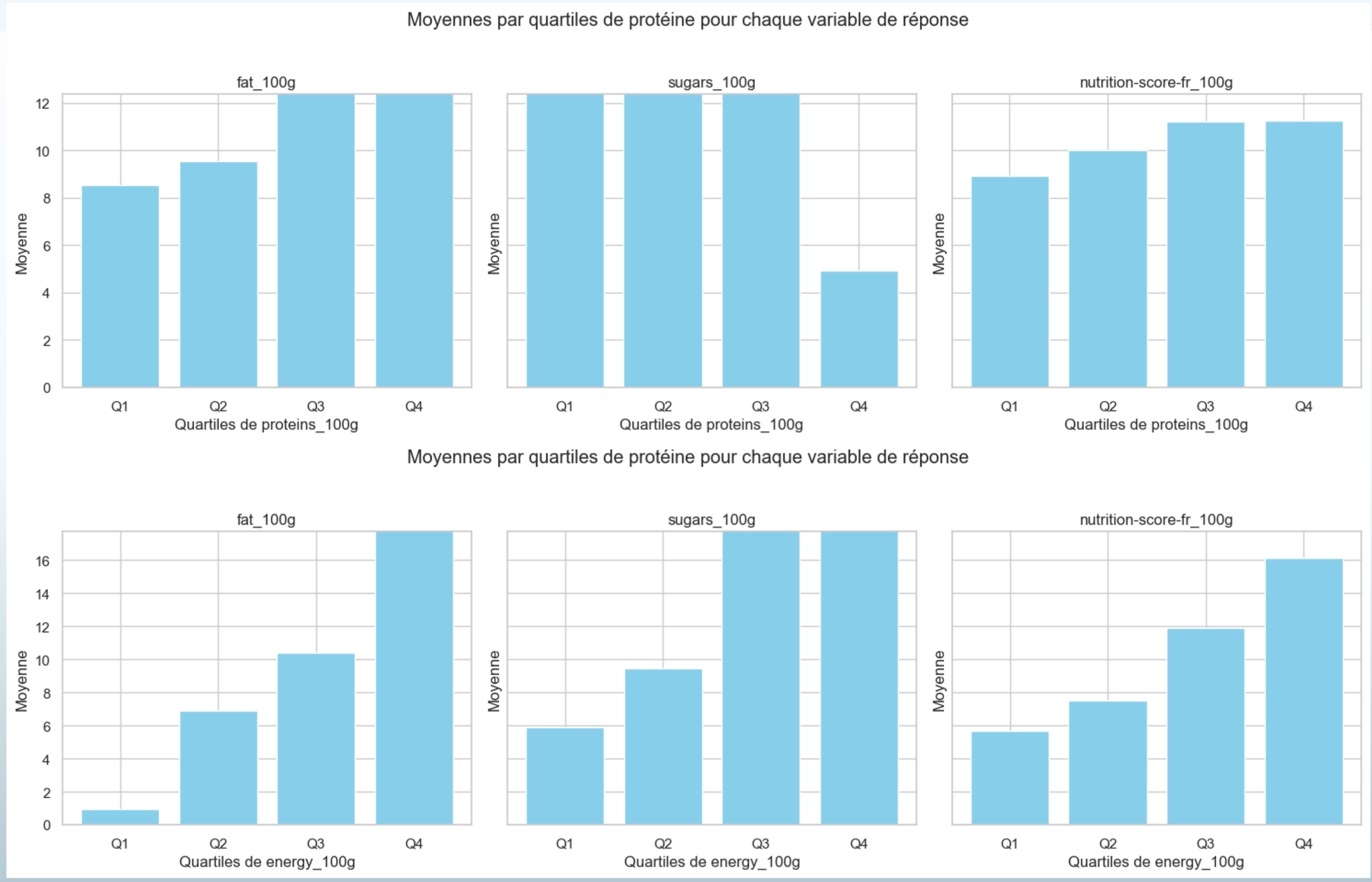
Machine Learning Engineer

Le test du Chi-carré est effectué pour évaluer si deux variables sont indépendantes. Deux cartes de chaleur (heatmaps) sont générées. Les visualisations présentent aident à comprendre visuellement les différences entre les observations et les valeurs attendues sous l'hypothèse d'indépendance. Cette approche est utile pour comprendre comment différents composants nutritionnels interagissent entre eux.



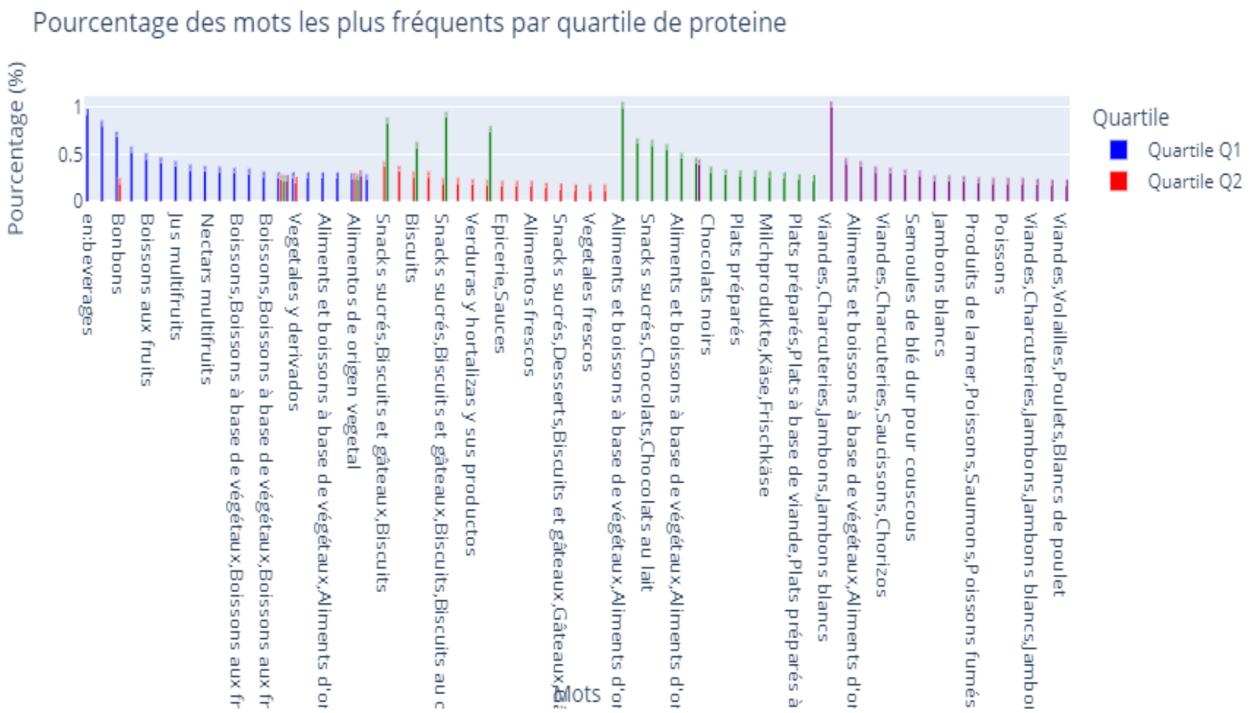
Le test Chi-carre

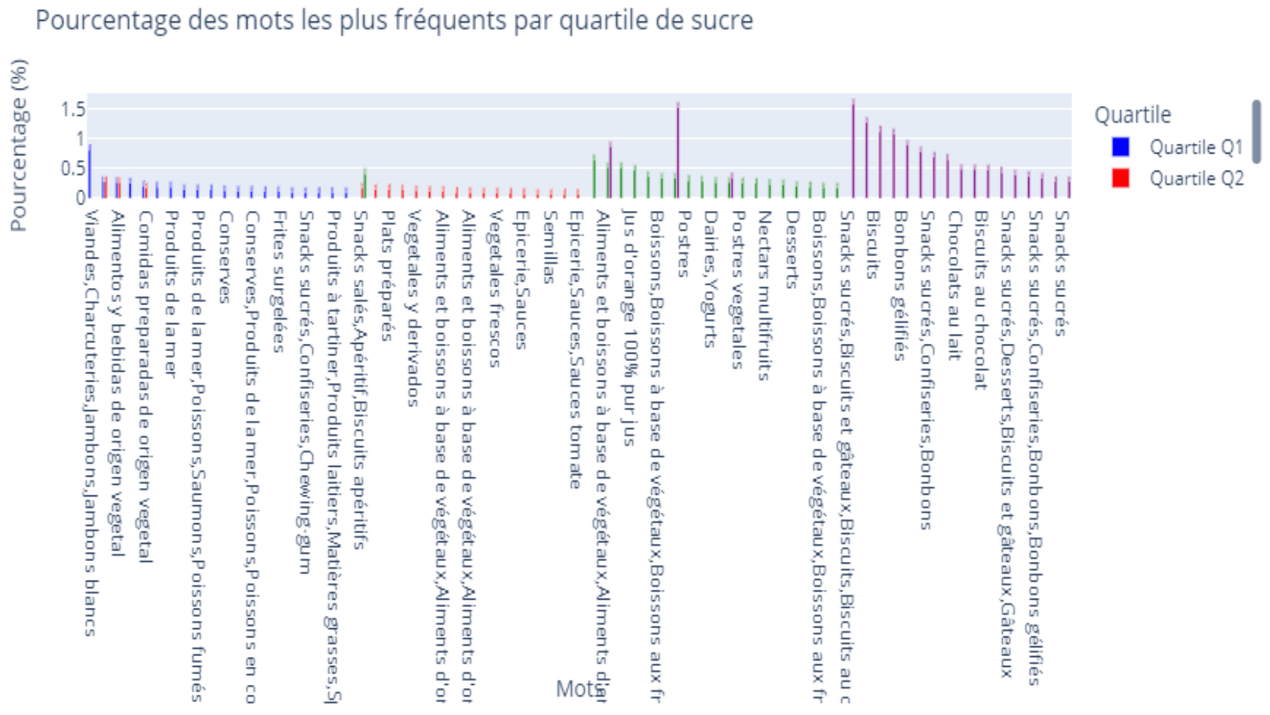
Machine Learning Engineer



ANOVA

Machine Learning Engineer



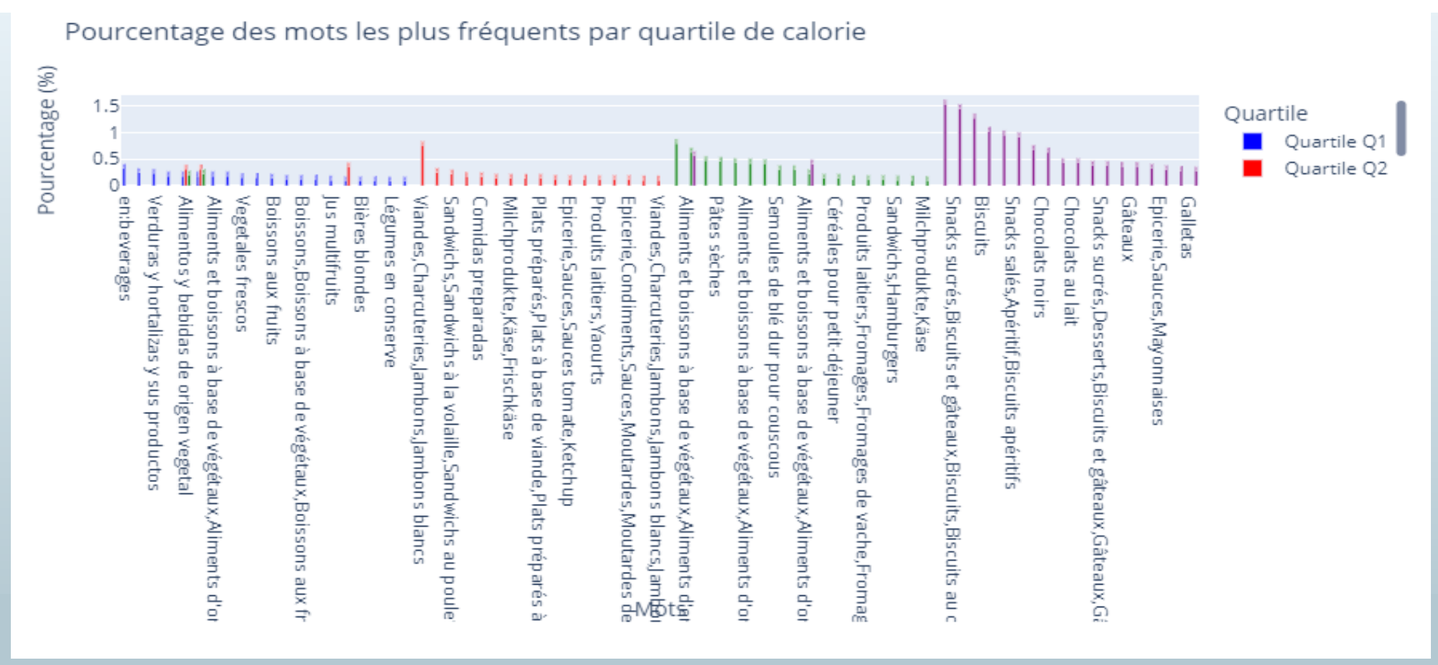


Q1

Q2

Q3

Q4



ANOVA

Faisabilité de l’application

Machine Learning Engineer

Analyse ANOVA : une ANOVA compare les moyennes des variables sélectionnées entre les différents quartiles. Détermine si les différences observées entre les moyennes des quartiles sont statistiquement significatives et si elles influent sur d’autres variables.

Sur la base de la découpe en quartile, les catégories qualitatives présentes pour chaque quartile sont également visuellement affichées. Ici, nous observons la présence de produit lié aux sucreries au dernier quartile de la variable sugar_100g.

En fin de compte, après la segmentation en quartiles et l'analyse ANOVA, les fréquences de mots sont recalculées et visualisées pour donner une perspective complète sur la manière dont les propriétés des produits varient.

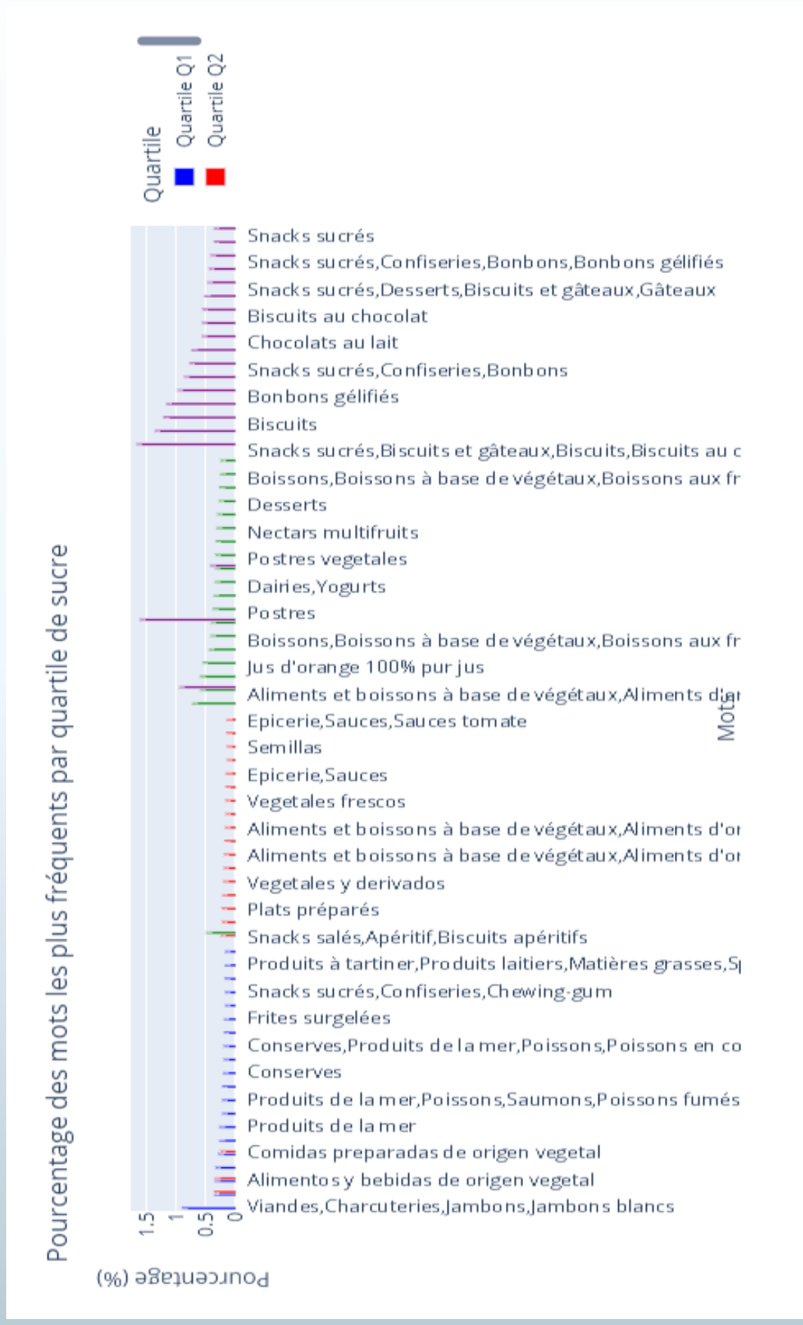
Cette analyse est la clé de voute de la faisabilité de notre application. Des tendances importantes, comme la prédominance de certaines catégories de produits ou de revendications nutritionnelles dans des quartiles de haute ou basse teneur de la valeur nutritive analysée deviennent statistiquement prévisibles, et nous pourrions suggérer les catégories des produits nouvellement ajoutés aux utilisateur uniquement avec les premières valeurs nutritives qu’il entrera.

Q1

Q2

Q3

Q4



ANOVA
Faisabilité de l’application