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

Ce projet vise à évaluer la faisabilité d'une application destinée à améliorer la qualité et la complétude des informations sur de nouveaux produits alimentaires ajoutés manuellement par les utilisateurs. Pour cela, des opérations de nettoyage, d'analyse et d'interprétation des données extraites d'une base de données open source sont essentielles pour assurer le succès de cette initiative.
Cette étude permettra de mettre en place les éléments fondamentaux pour:

- Automatiser la suggestion de valeurs d'entrée pour les utilisateurs ajoutant de nouveaux produits à la base de données.
- Réduire les erreurs et à combler les lacunes souvent observées dans les données saisies manuellement.
- Assurer une assistance précise et contextuelle pour faciliter l'intégration fiable et efficace de données nutritionnelles complètes.

Préparez des données pour un organisme de santé publique

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

Machine Learning Engineer

Plan de l’étude:

- AED : Analyse exploratoire des données
  - Résumé
  - Traitement préliminaire des valeurs nulles
  - Pays de vente des produits
  - Les Doublons
  - Nettoyage de la colonne ingredient_text
- Traitement des valeurs nulles général
  - Vérification
  - Matrix de corrélation
  - Imputation
  - Catégories
- Etude, Analyse statistique descriptive, univariée, bivariée et multivariée
  - Analyse descriptive
  - Univariée
  - Bivariée et multivariée
- Analyse inférentielle, test chi-carré, ANOVA
  - Echantillon aléatoire
  - Anova
  - Chi-carré
- Faisabilité de l’application
  - Repère quartile et catégorie

Préparez des données pour un organisme de santé publique

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

L’analyse descriptive met en lumière plusieurs éléments. Des valeurs nulles, La présence de valeurs aberrantes dans certaines variables, une dispersion significative des données autour de la moyenne, La présence de minimums hors norme suggère une incohérence, nécessitant une investigation approfondie ainsi qu’un traitement pour normaliser ces données.

	nutrition-score-fr_100g
count	221210.000000
mean	9.165535
std	9.055903
min	-15.000000
25%	1.000000
50%	10.000000
75%	16.000000
max	40.000000

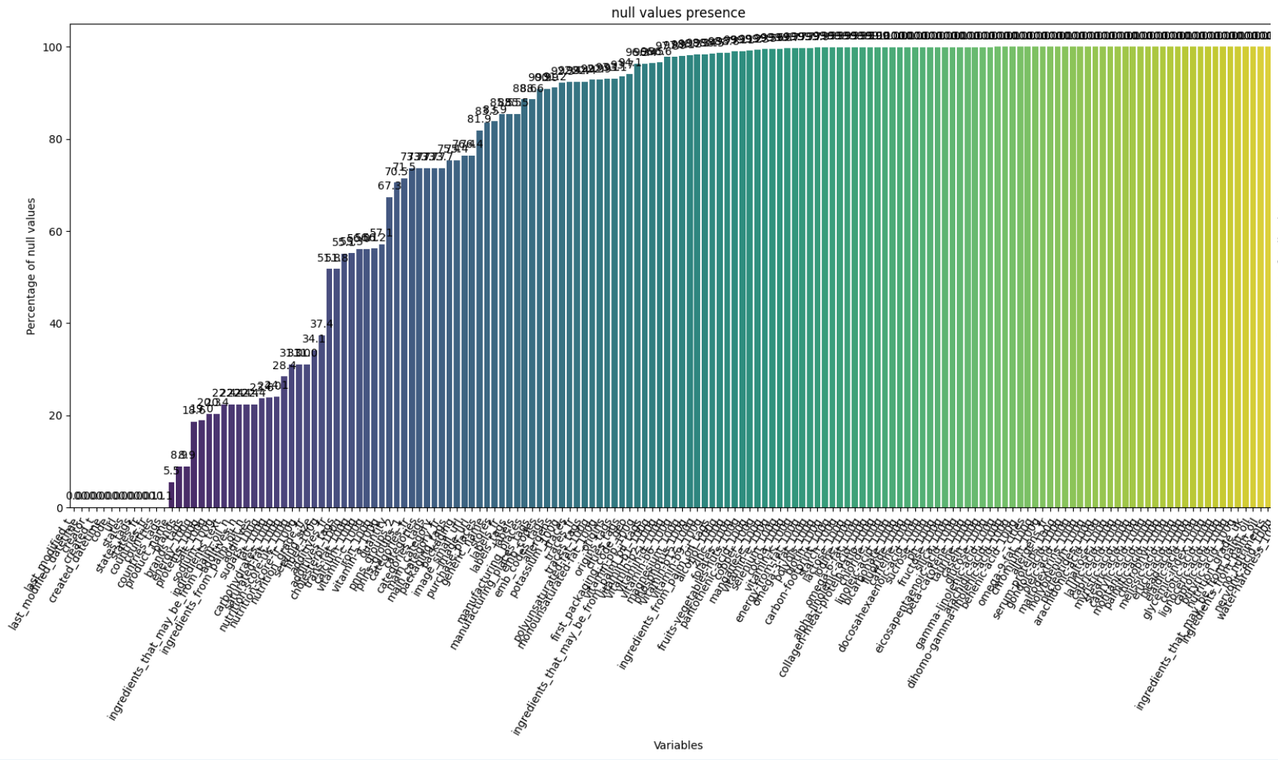
	energy_100g
count	2.611130e+05
mean	1.141915e+03
std	6.447154e+03
min	0.000000e+00
25%	3.770000e+02
50%	1.100000e+03
75%	1.674000e+03
max	3.251373e+06

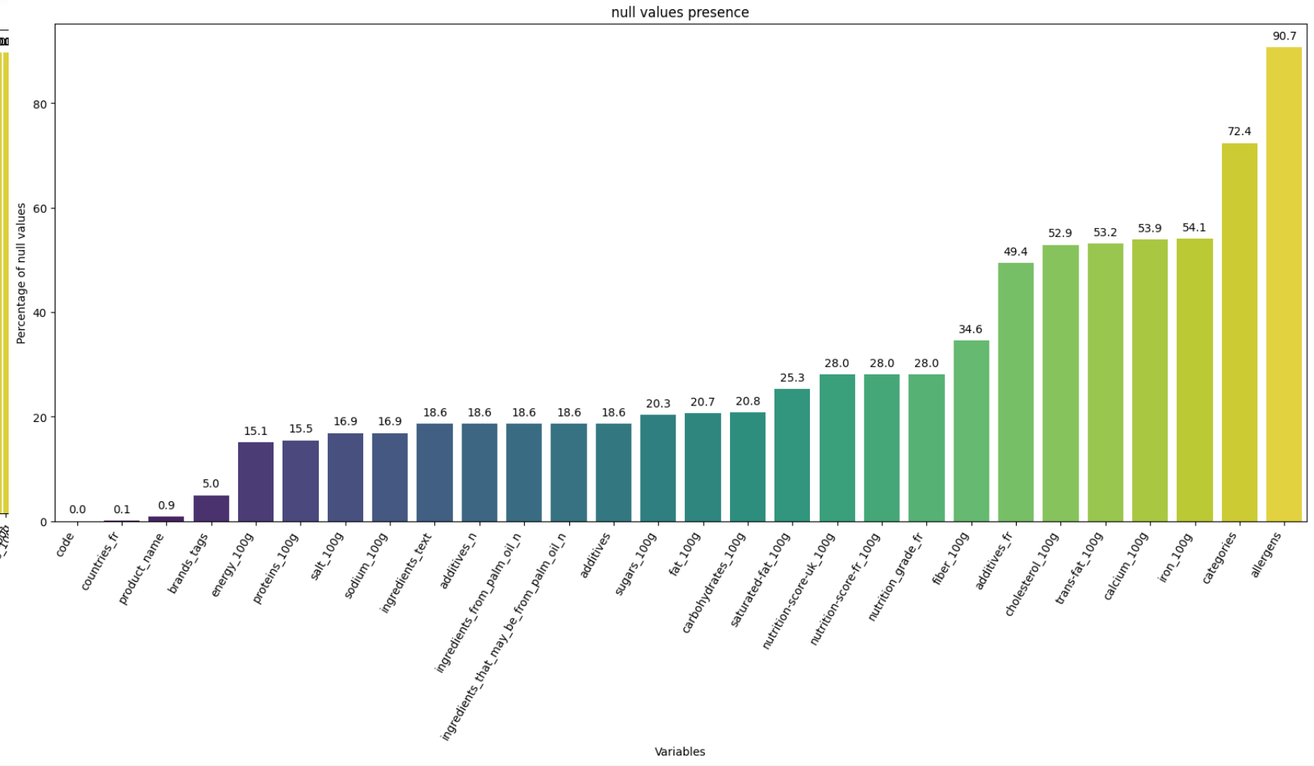
	iron_100g
199397.000000	
	0.002162
	0.173402
	-0.000260
	0.000000
	0.000000
	0.001290
	50.000000

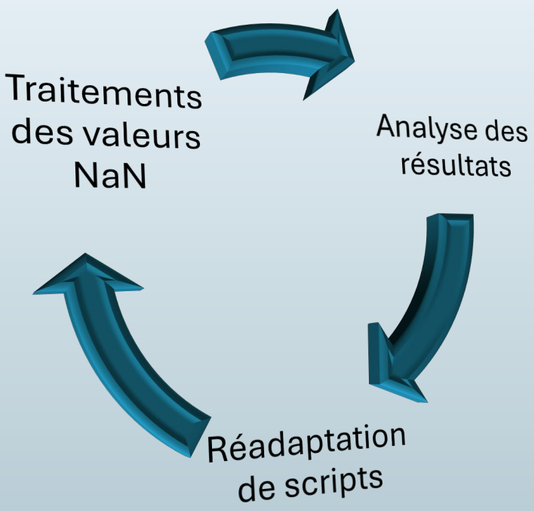
```
RangeIndex: 320772 entries, 0 to 320771
Columns: 162 entries, code to water-hardness_100g
dtypes: float64(106), object(56)
```

Analyse exploratoire/descriptive

Machine Learning Engineer





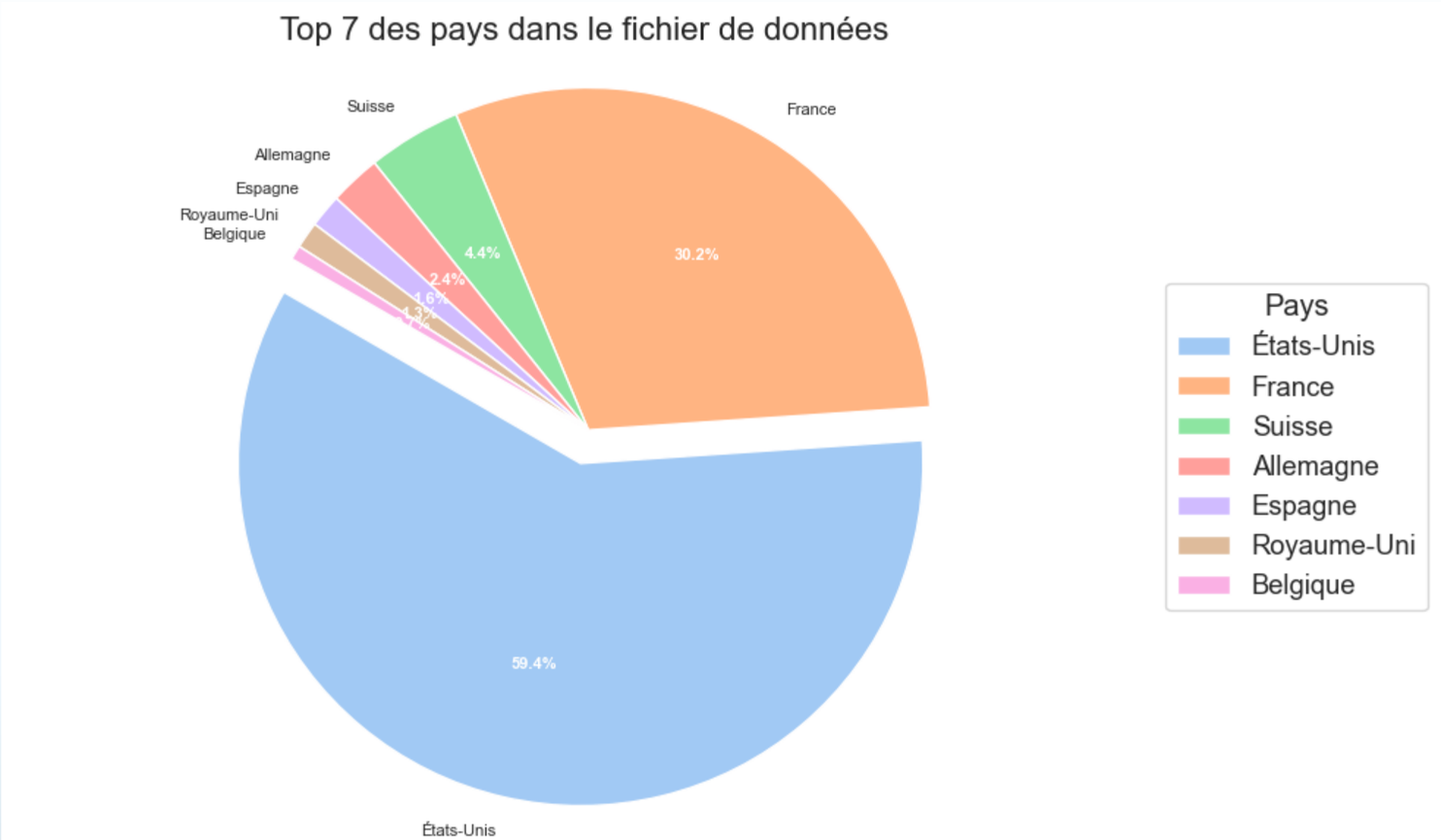


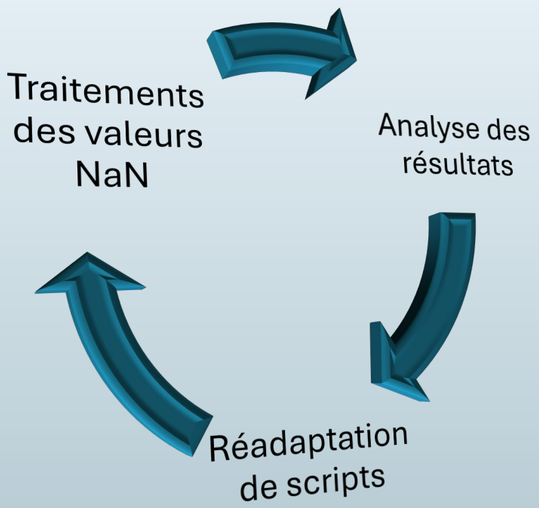
- Traitement large des valeurs nulles sur les données
- Respect de la RGPD dans la conservation de certaines données

La préparation des données
Suppression et Imputation de données

Machine Learning Engineer

- Conservation des 7 pays les plus alimentés en données

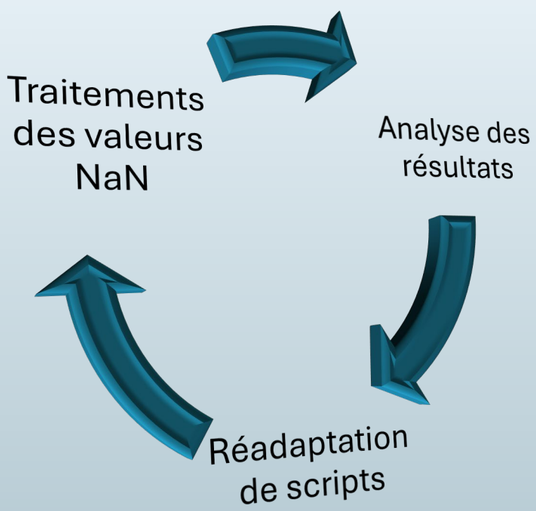




La préparation des données
Suppression et Imputation de données

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

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

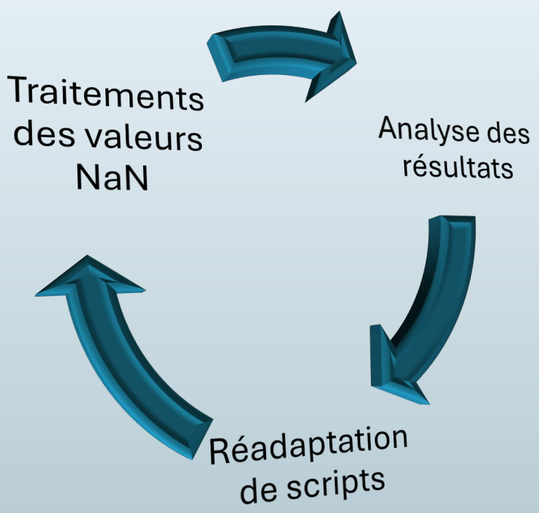


La préparation des données
Suppression et Imputation de données

- Traitement général de valeur null

- Limitation générale des valeurs abérrantes

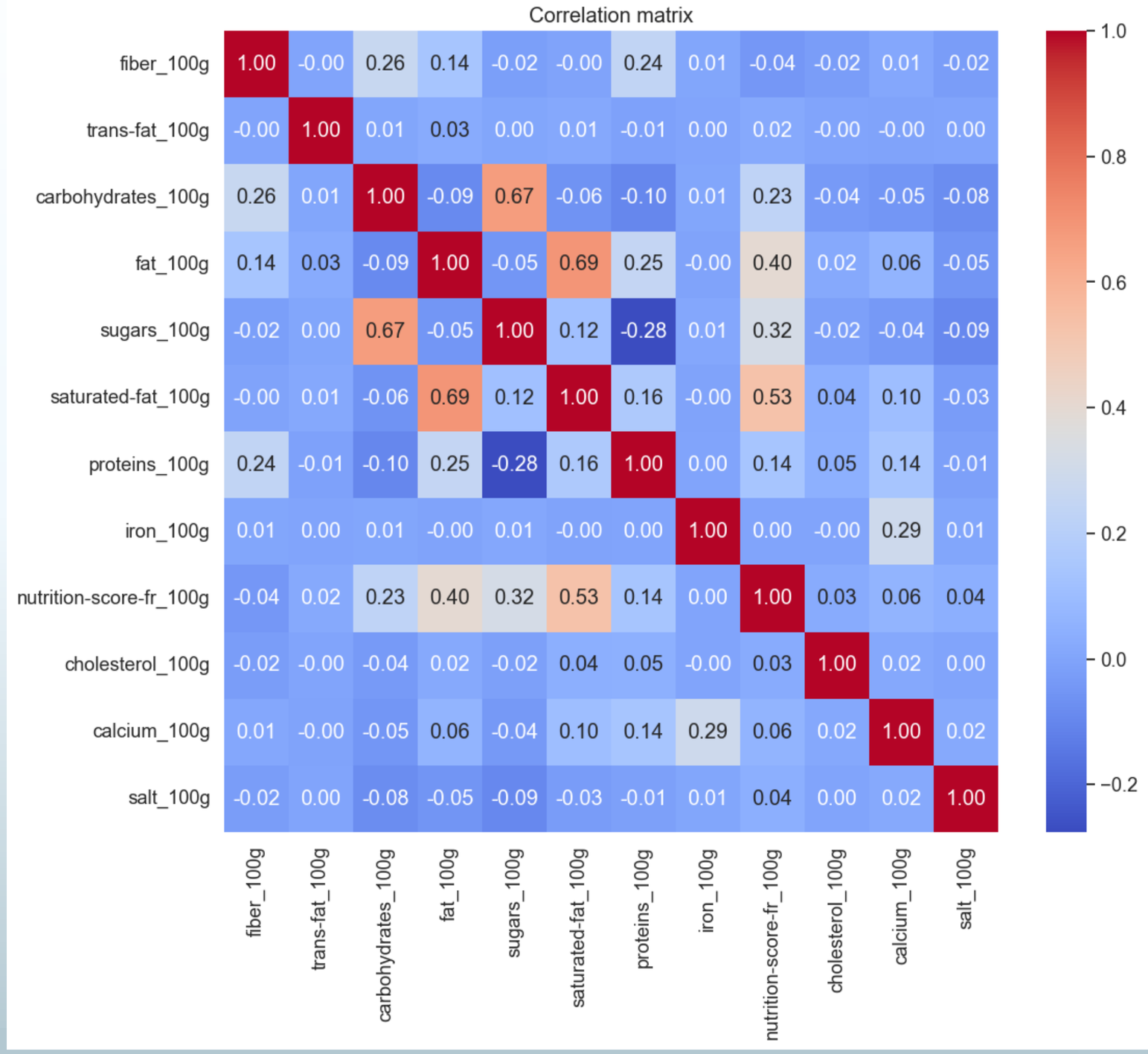
  - Les calories joules ne sont pas supérieurs à 3900 unités pour 100g
  - Les variables numériques pour 100g ne peuvent pas être inférieur à 0 ni supérieur à 100

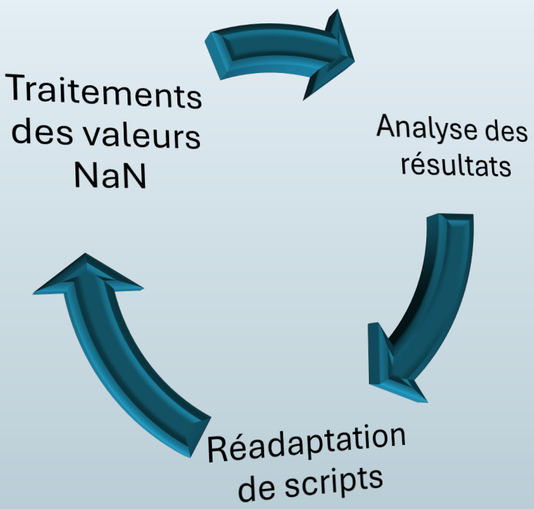


La préparation des données
Suppression et Imputation de données

Machine Learning Engineer

Matrix de Correlation





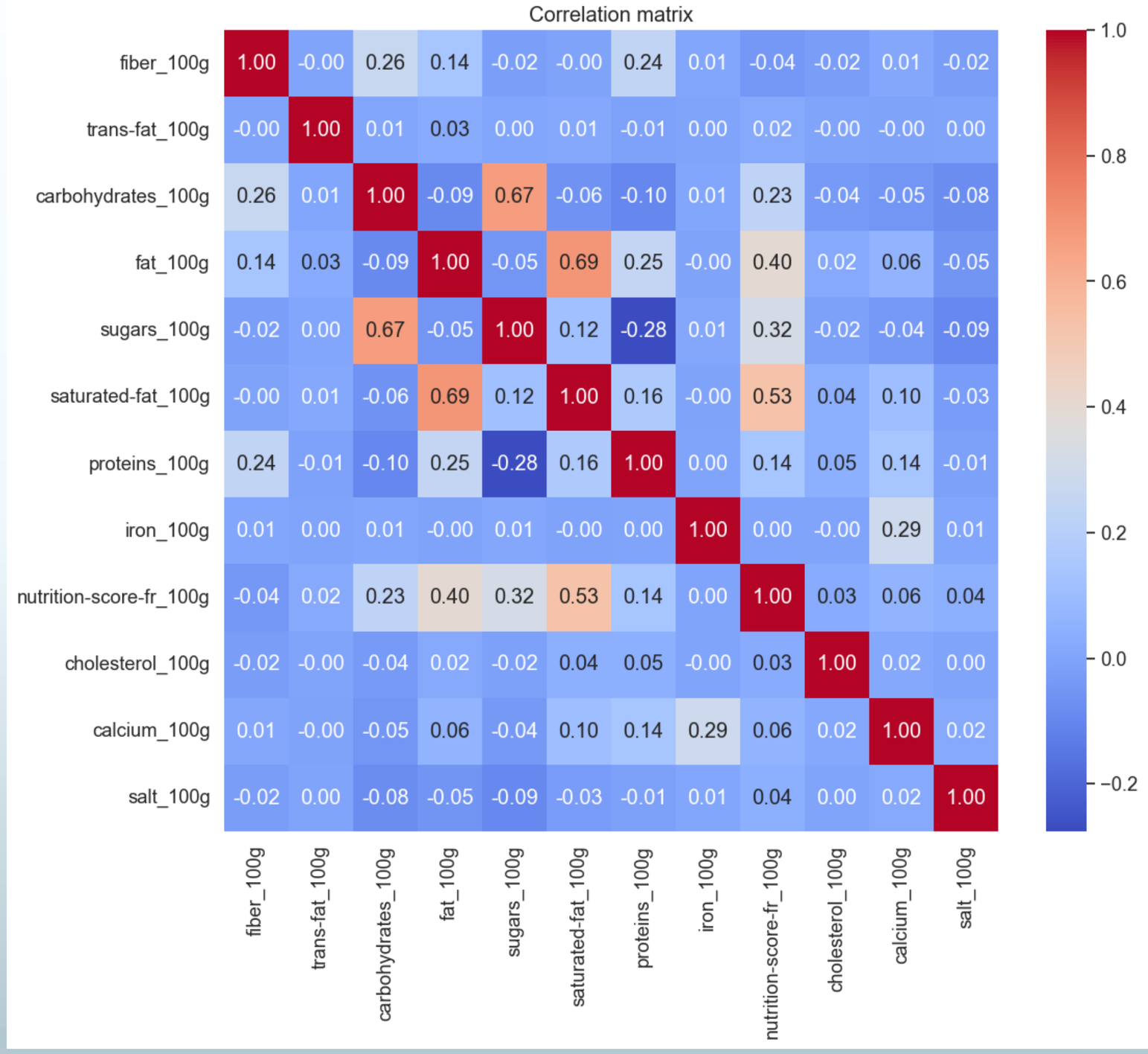
Vérification de la relation entre les données
Analyse et Imputation de données

Machine Learning Engineer

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

L'identification des fortes corrélations contribuera à améliorer la précision des données saisies. La collecte et la maintenance des données seront optimisées pour une base de données plus efficace.

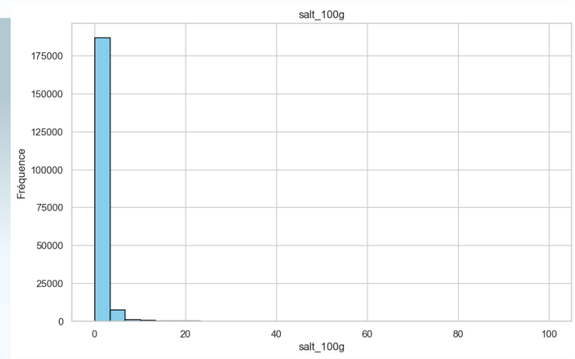
Les insights ainsi révélés serviront également à concevoir l'interface utilisateur, en anticipant les valeurs que les utilisateurs sont susceptibles de saisir.

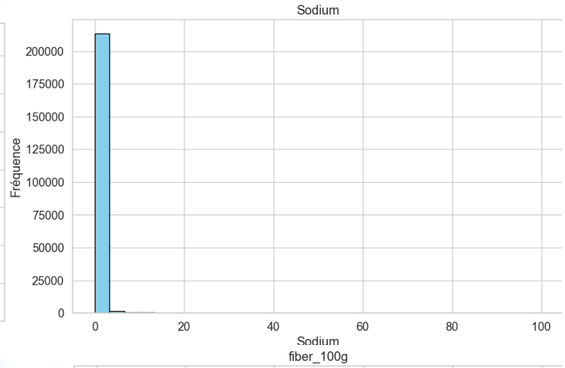


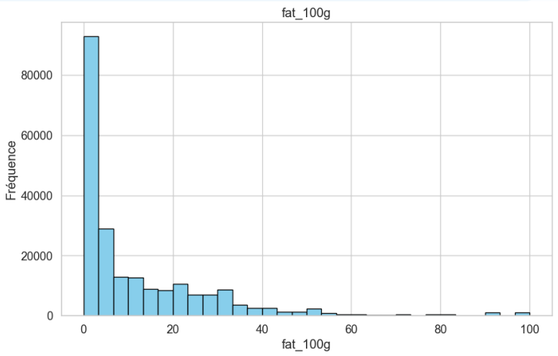
Vérification de la relation entre les données
Analyse et Imputation de données

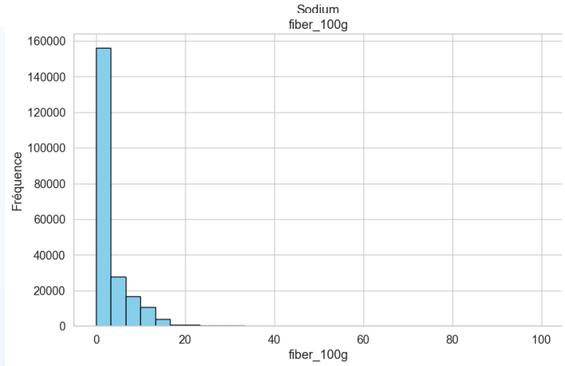
Machine Learning Engineer

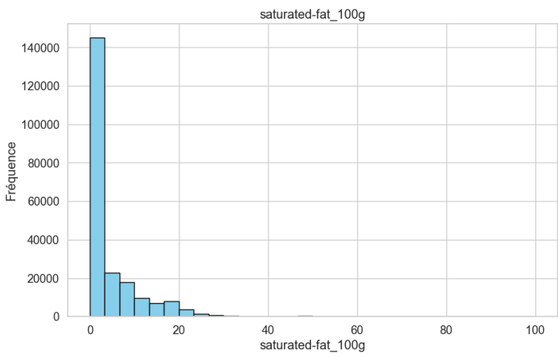
Imputation des valeurs Nan



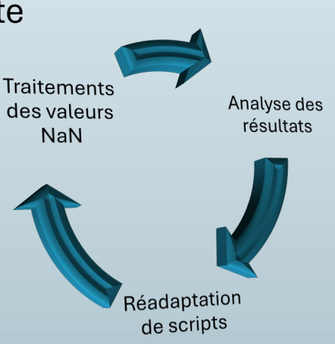


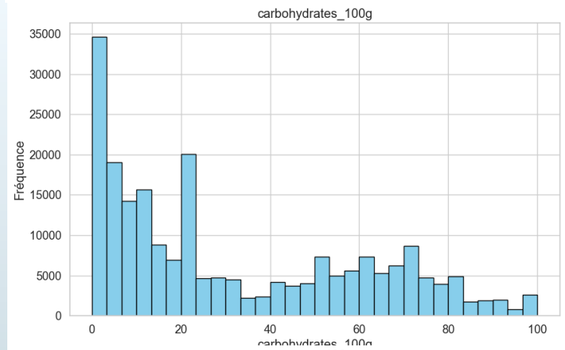


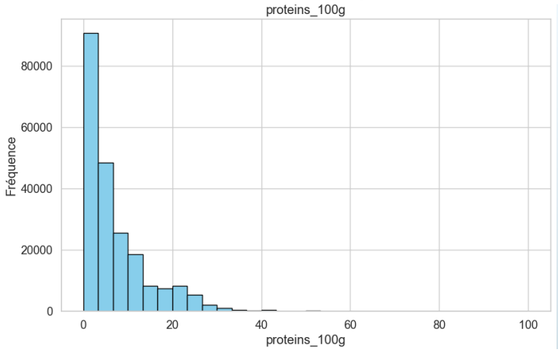


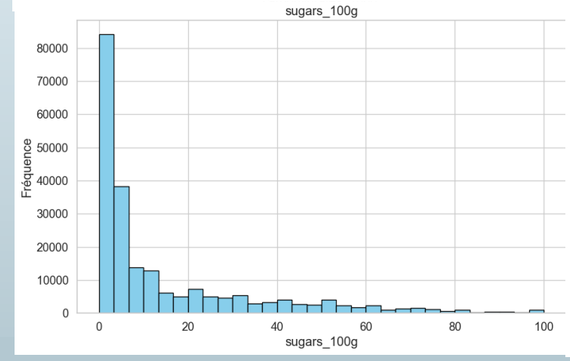


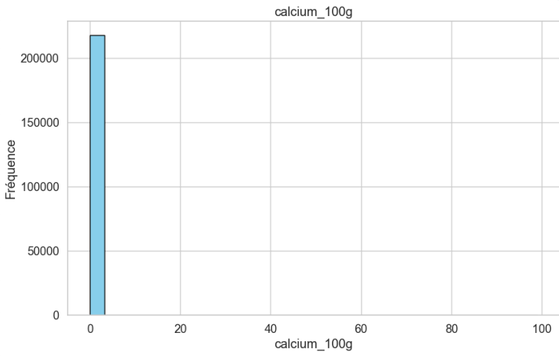
- La variable salt_100g : la médiane dont la moitié est diminué si le sucre est conséquent pour ses valeurs NaN
- La variable sodium_100g : un ratio de 40% du poids du sel
- Les variables sugar_100g, carbohydrates_100g, fiber_100g, leur valeur NaN sont imputé en relation les un avec les autres.
- Les variables fat_100g et saturated-fat_100g sont corrélées, 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 nulle, elle est absente









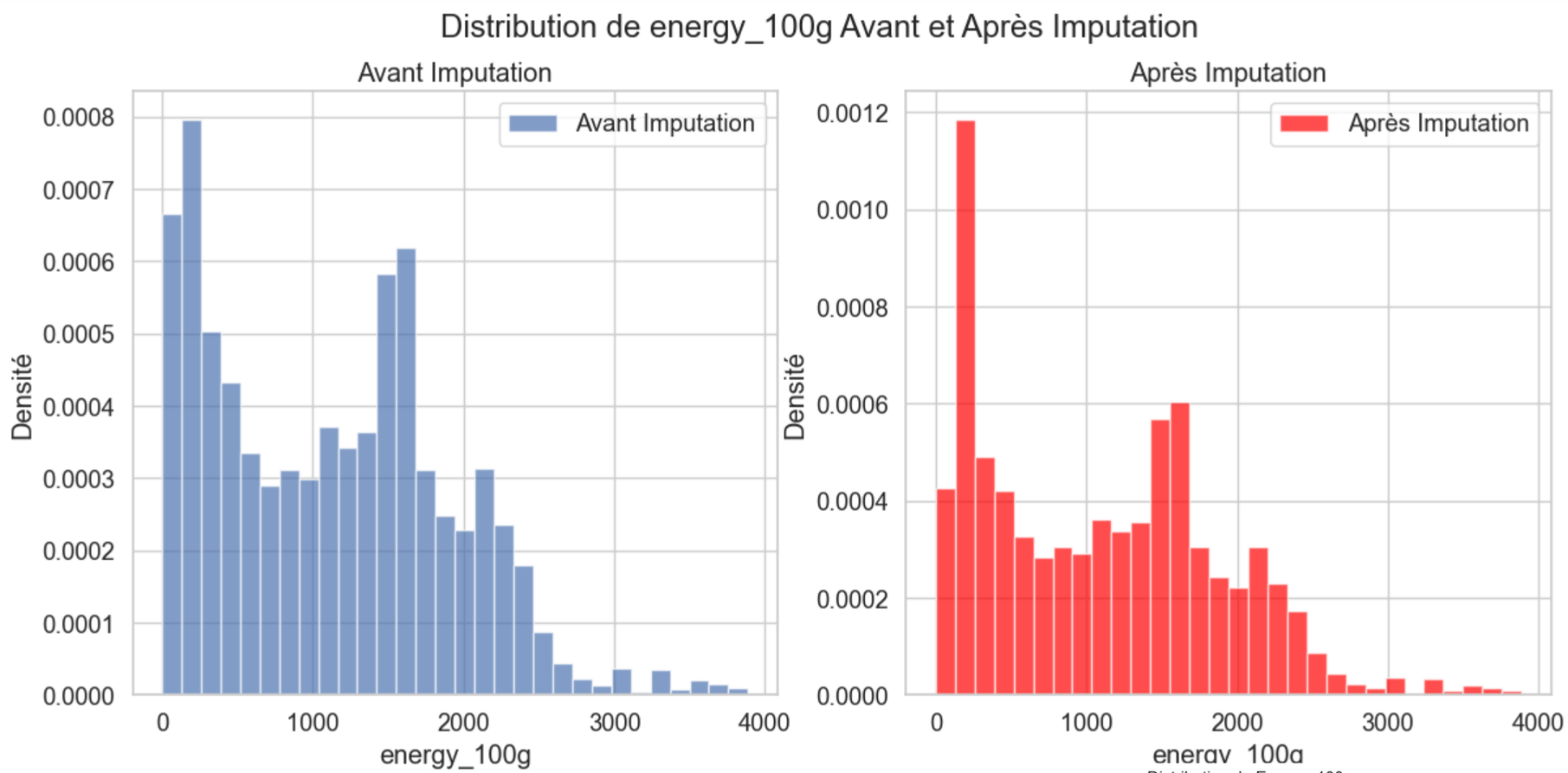


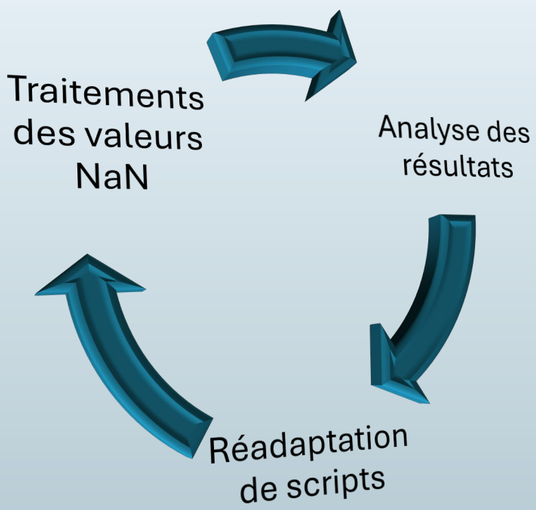
Vérification de la relation entre les données
Analyse et Imputation de données

Machine Learning Engineer

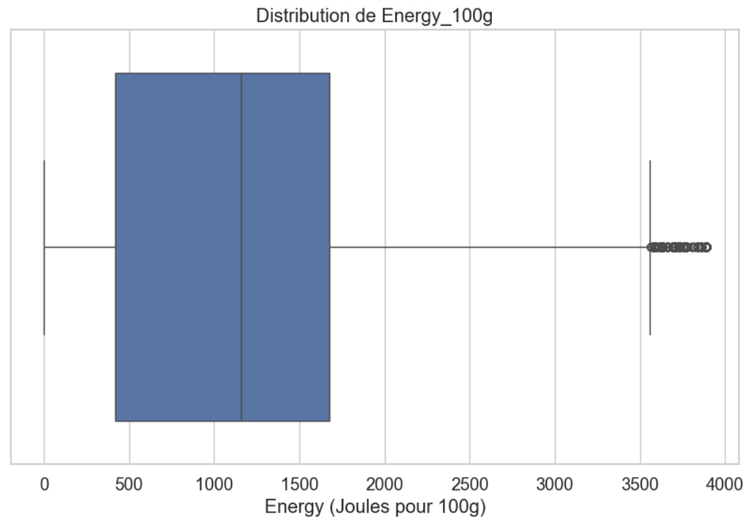
Distribution des calories

- 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

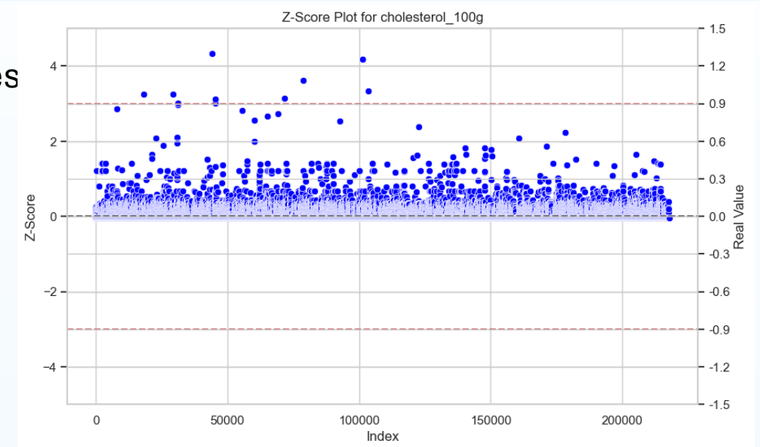


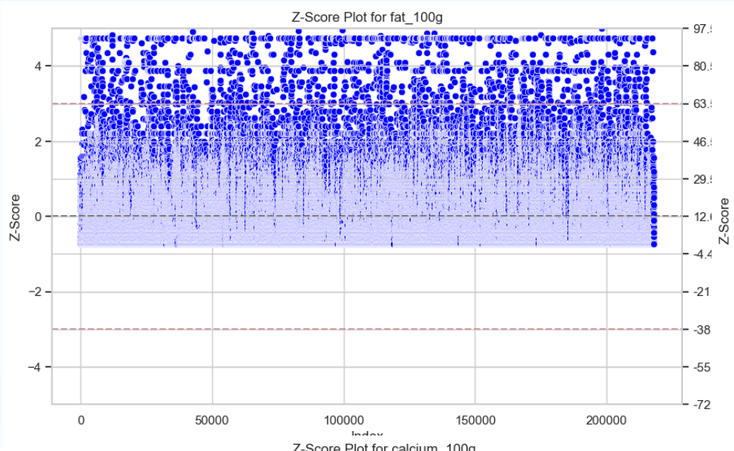
La préparation des données
Analyse et Imputation de données

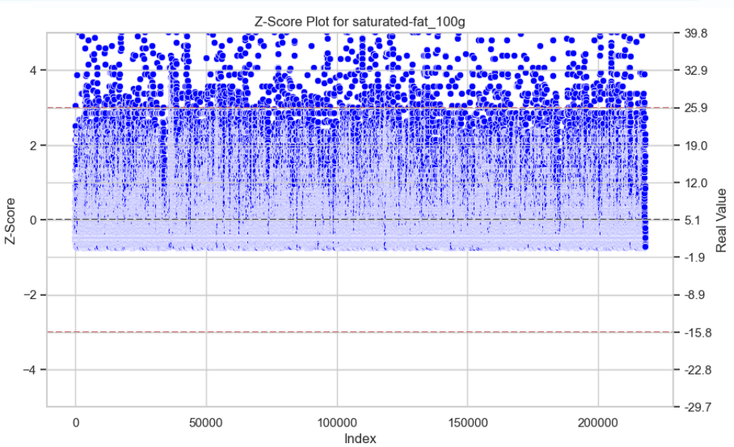
Machine Learning Engineer

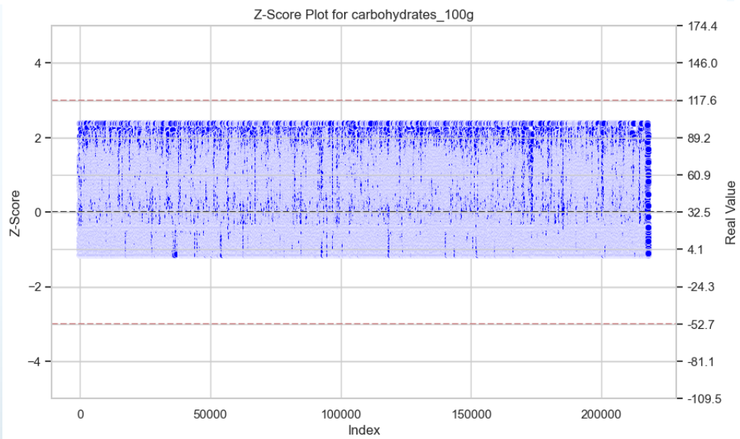
Z-Score

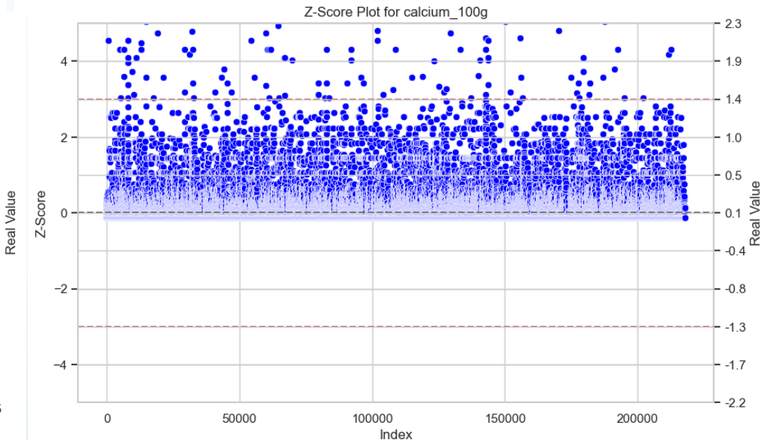
- Travaille sur les valeurs aberrantes :

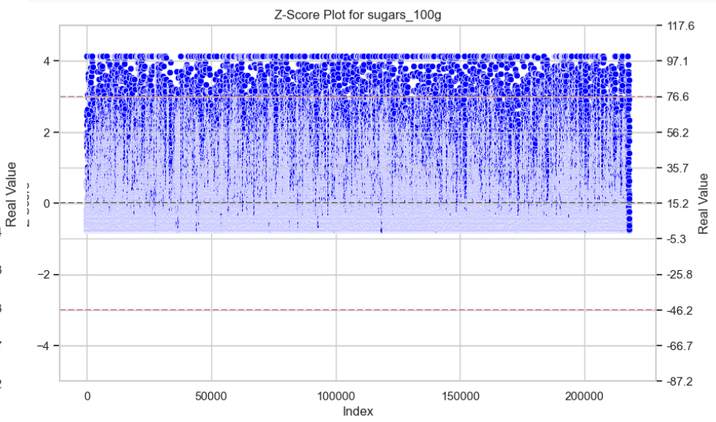


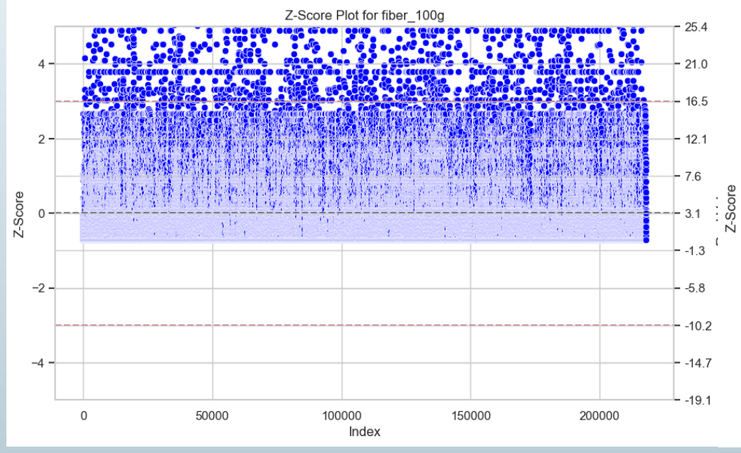


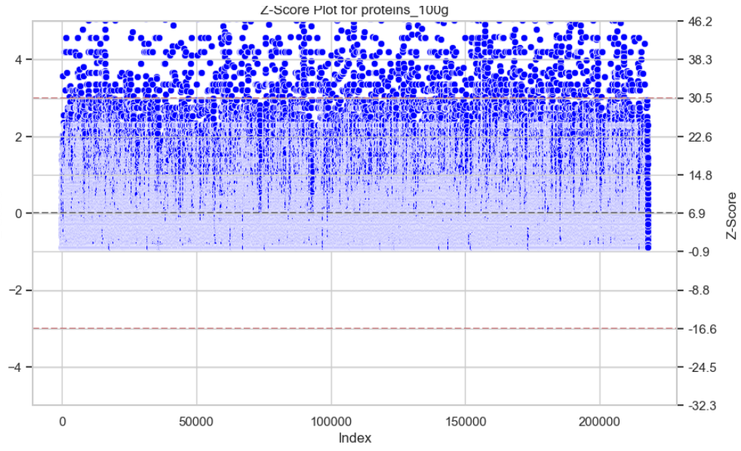


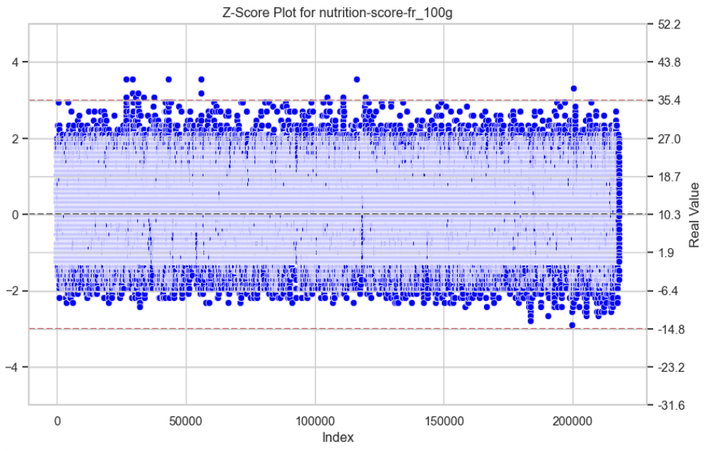












La préparation des données
Analyse et Traitement des valeurs abérrantes

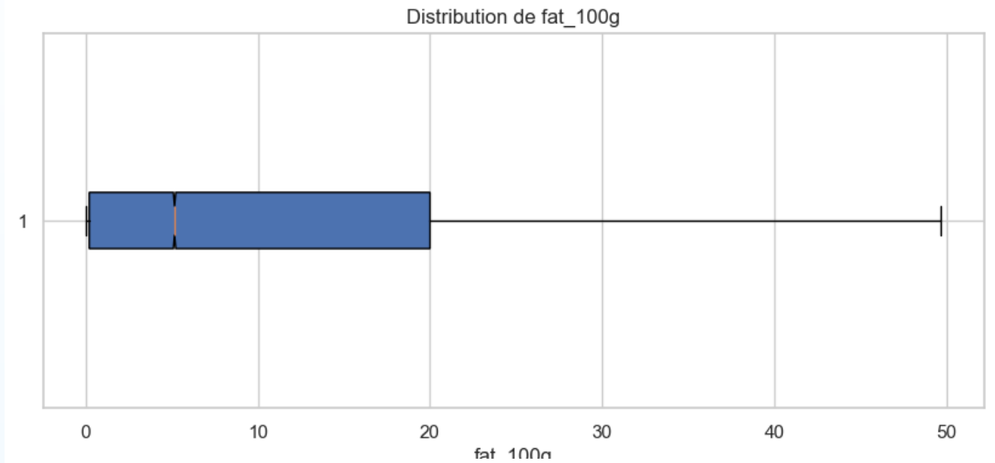
Machine Learning Engineer

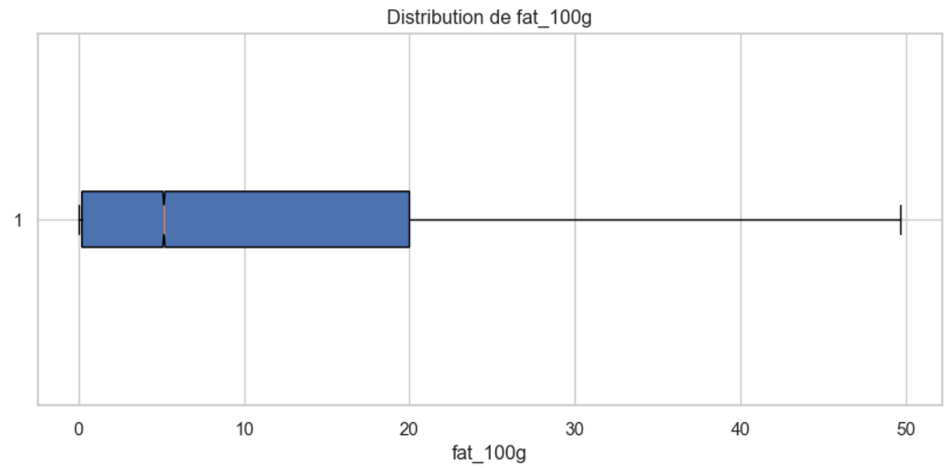
Z-Score

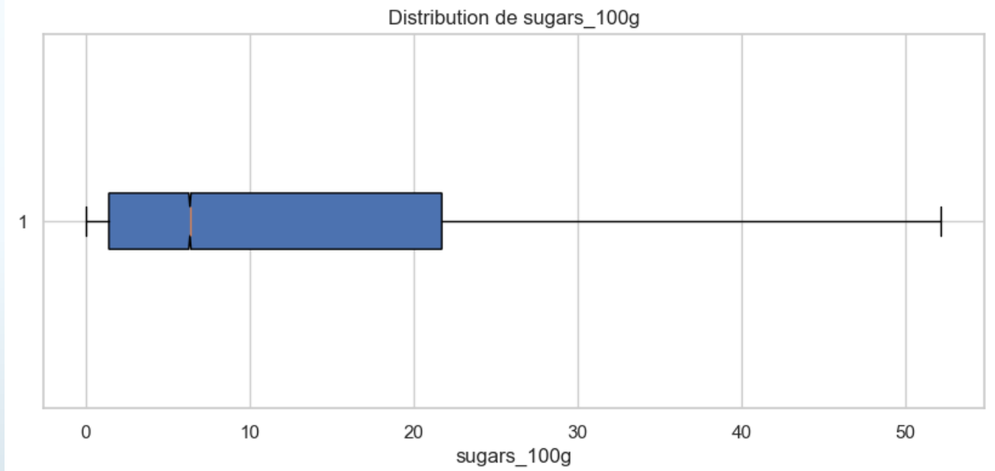
- Travail sur les valeurs aberrantes

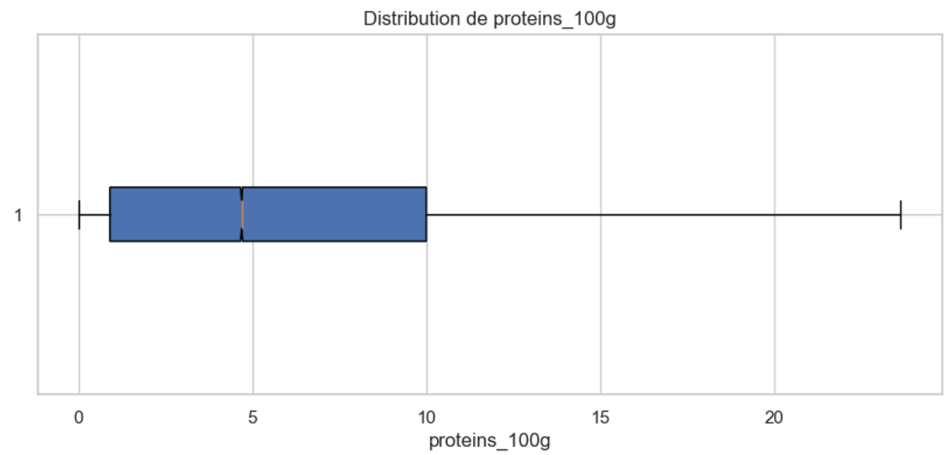
- Dans notre analyse, nous identifions la valeur non aberrante la plus élevée et nous ajustons l'extrême à cette valeur tout en se servant des quartiles pour établir une proportion lors de l’adaptation de la valeur. Cette approche permet de conserver la variance naturelle entre les données, tout en limitant les effets des valeurs extrêmes.

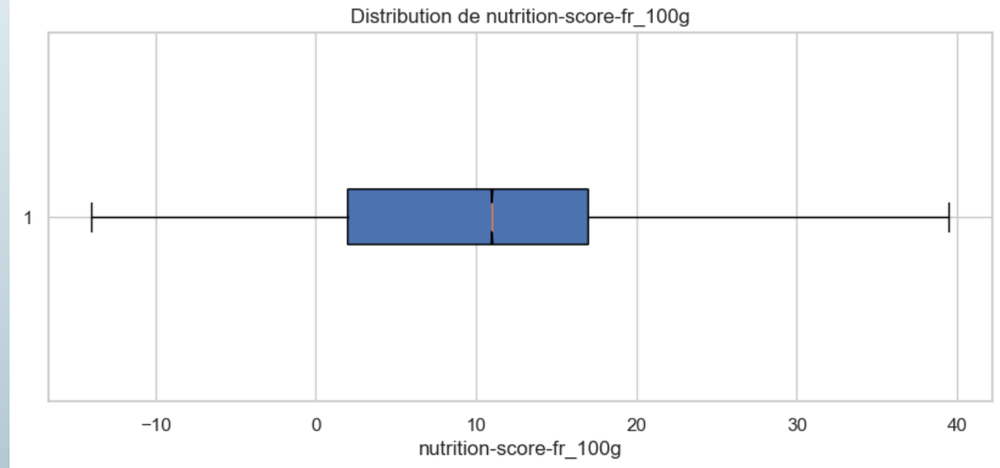
Après traitement :

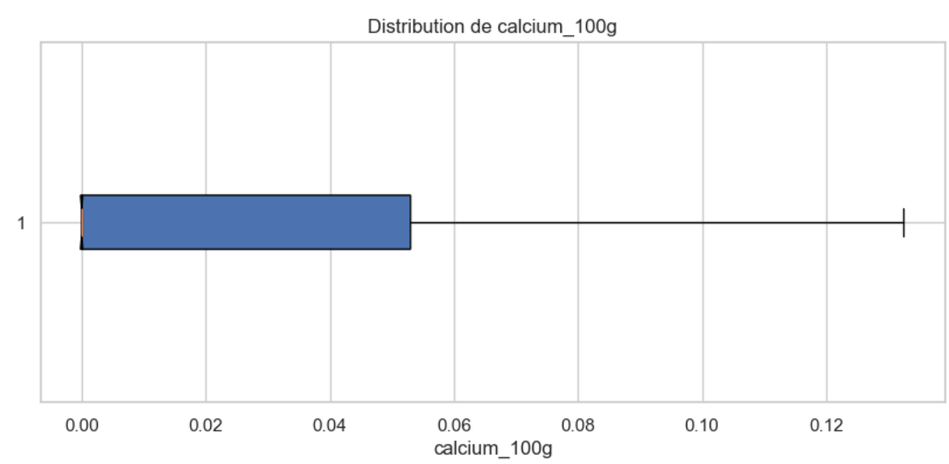










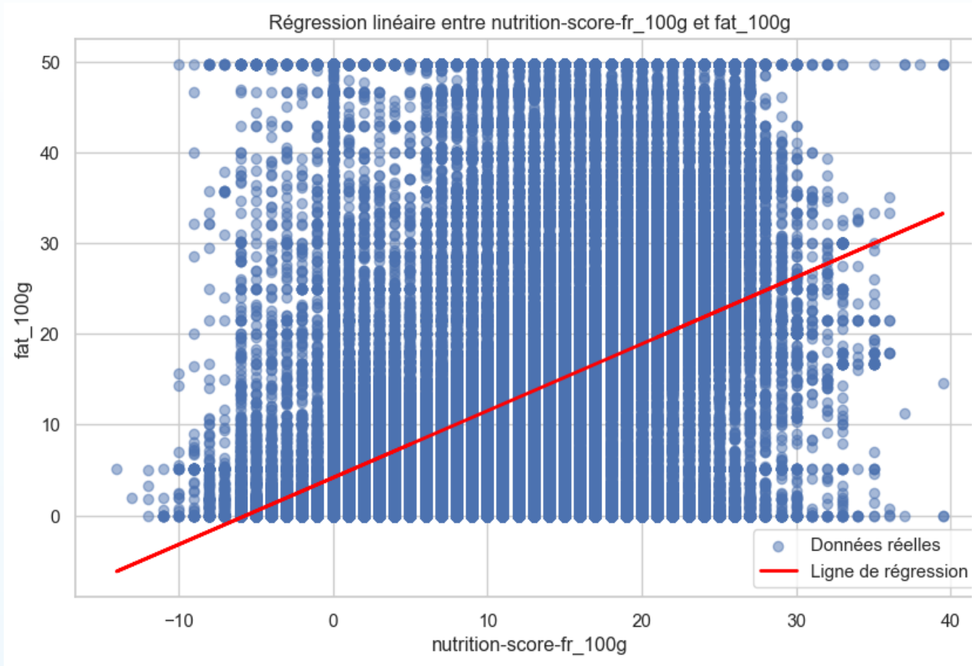


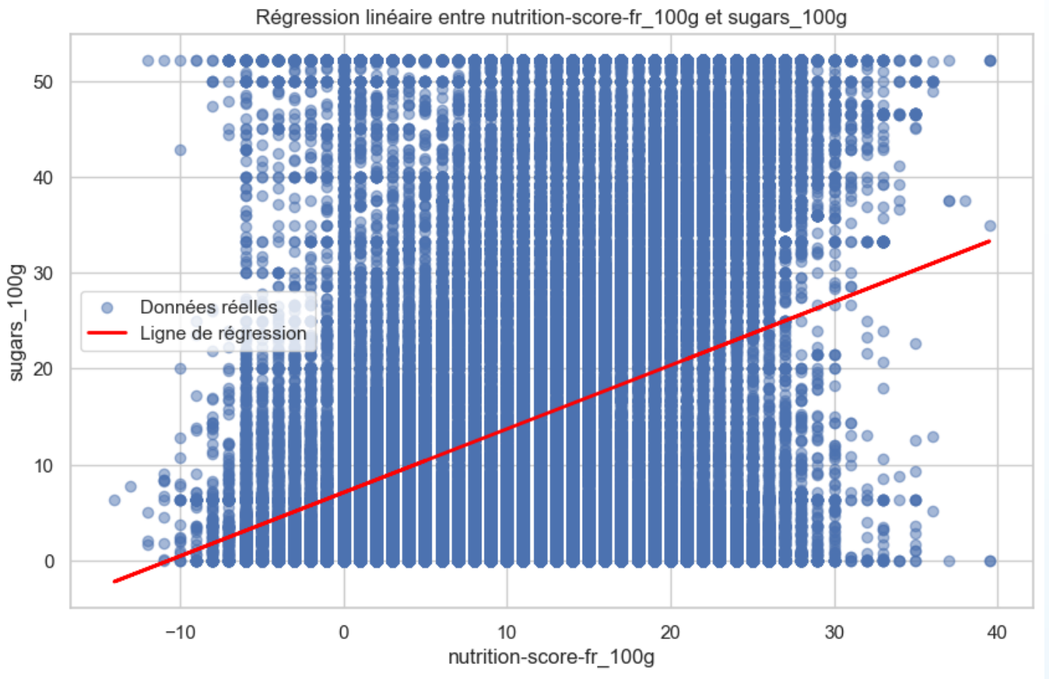
La préparation des données
Analyse et Traitement des valeurs abérrantes

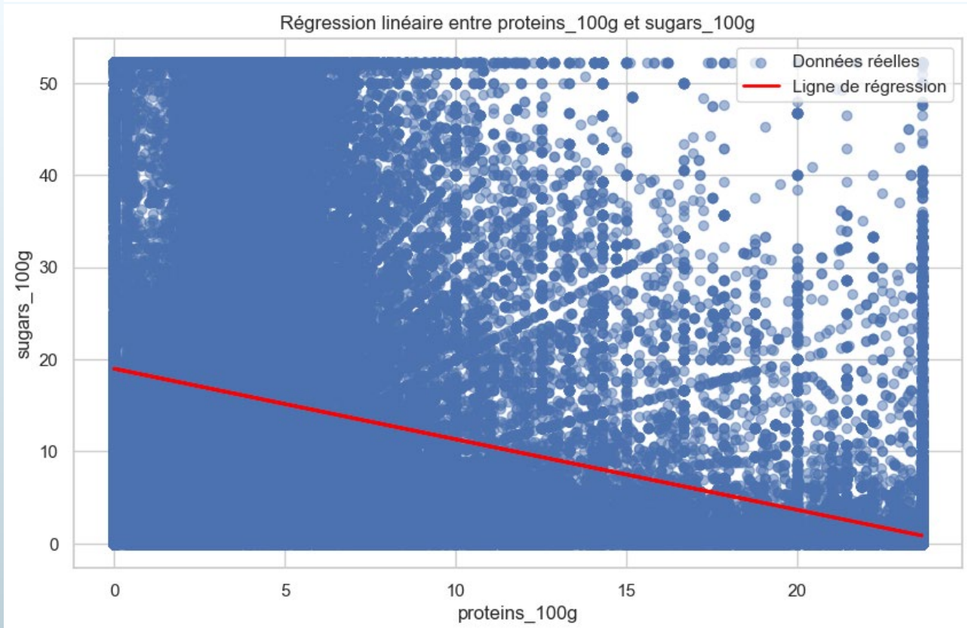
Machine Learning Engineer

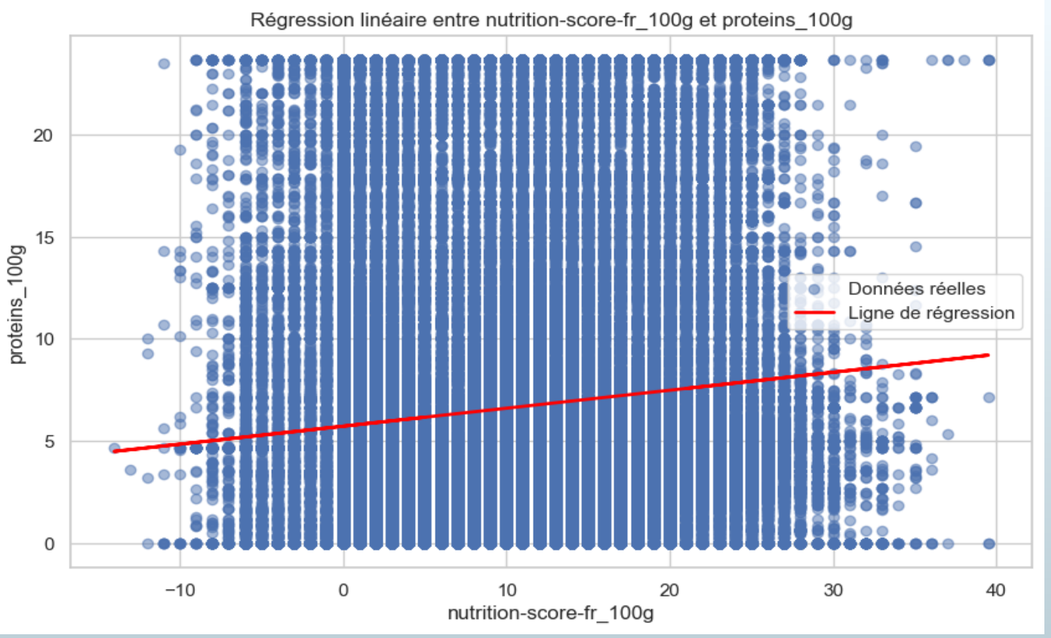
- Le gras et le sucre influencent légèrement le score de nutrition

- La protéine ne semble pas impacter le score de nutrition de manière suffisamment significative pour établir une relation de causalité.





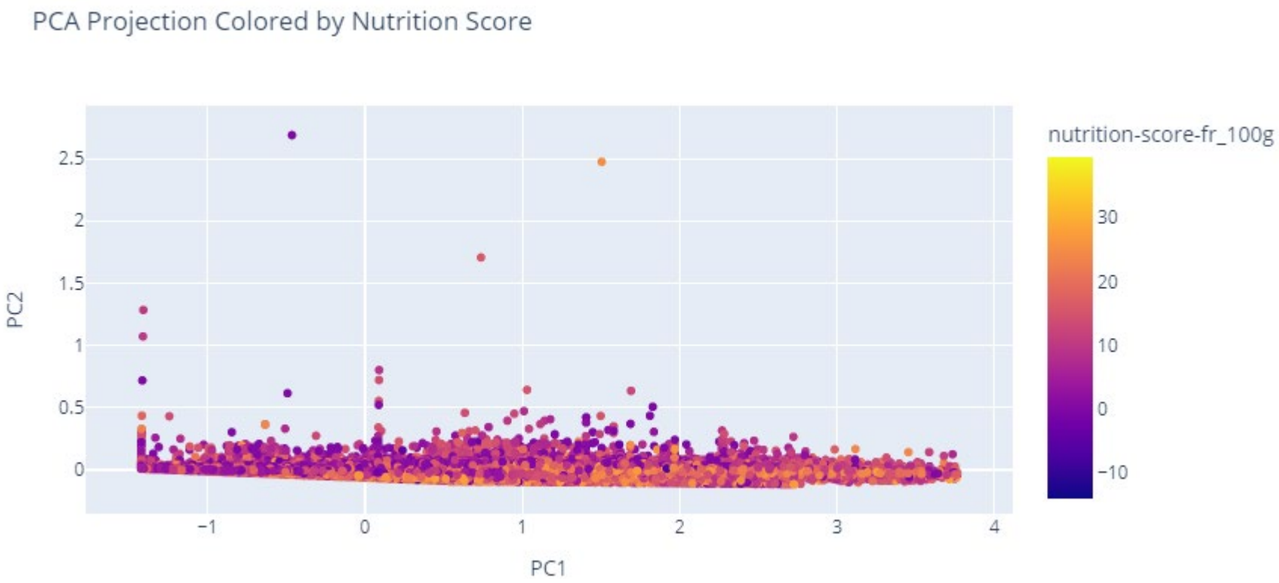




Regression Lineaire

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'
```

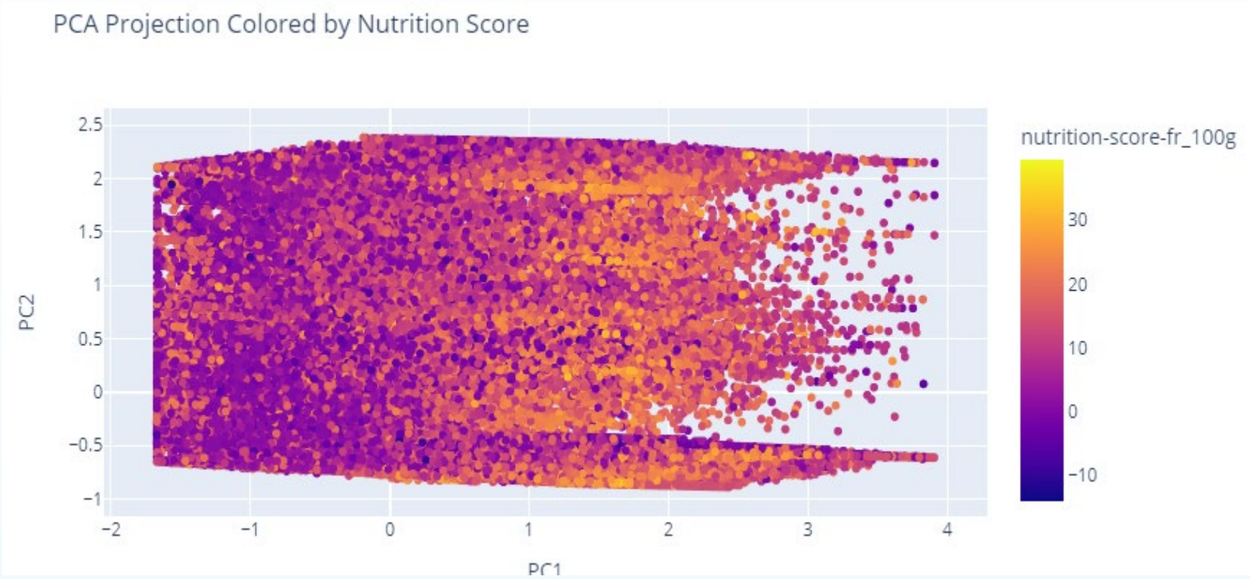
ACP

Analyse des Composants Principaux

Machine Learning Engineer

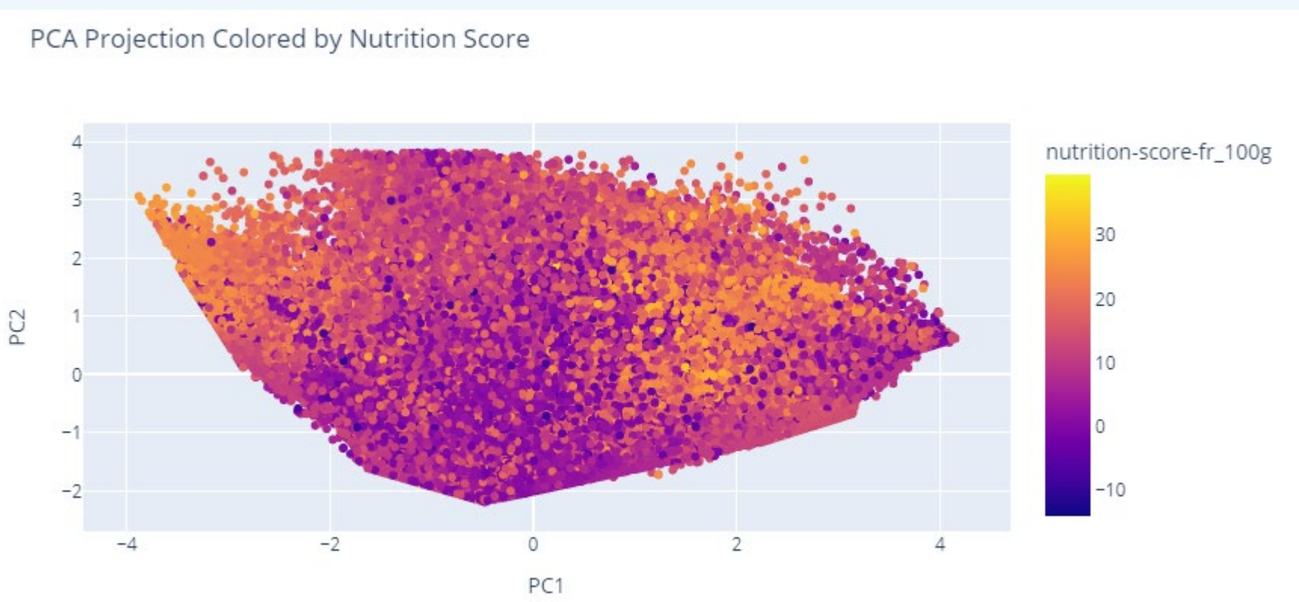
Analyse en Composantes Principales (ACP) :
L'ACP transforme les variables en un nouvel espace où les axes principaux résument l'essentiel de leurs variations.

Visualisation des Produits :
Les produits sont représentés en différentes couleurs en fonction de leur score nutritionnel. Cette visualisation vise à identifier les tendances.




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

Dispersion et Variance :
L'axe horizontal capture une part significative de la variance. Les produits sont dispersés de gauche à droite et de haut en bas, illustrant une large couverture. Une concentration dense de points vers le centre du graphique indique que la majorité des produits ont des caractéristiques nutritionnelles évaluées comme modérées.



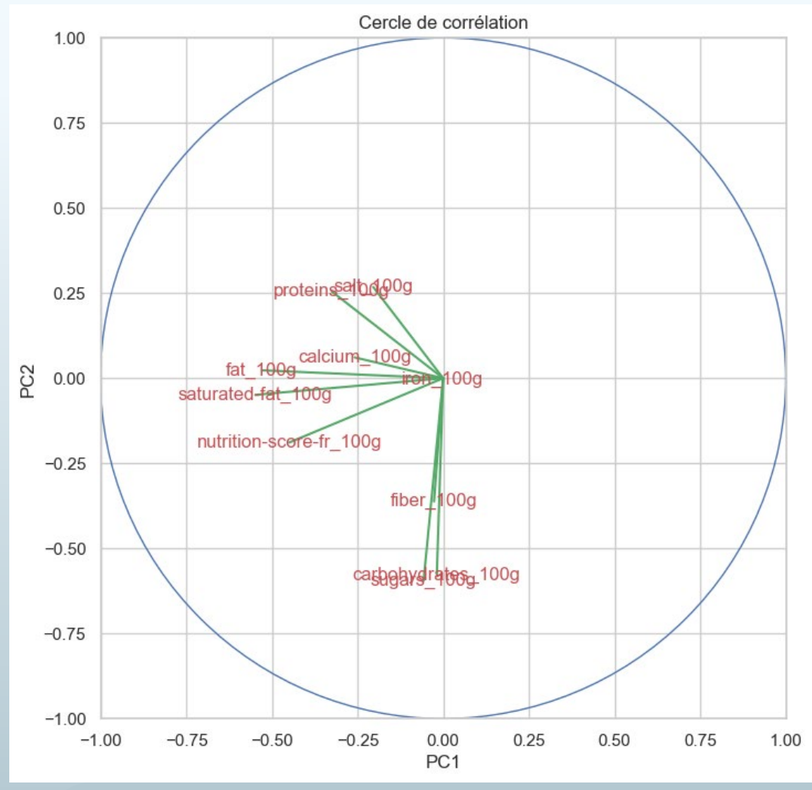

```
['fiber_100g', 'carbohydrates_100g', 'sugars_100g', 'calcium_100g', 'fat_100g',
'proteins_100g', 'iron_100g',
                'cholesterol_100g', 'salt_100g'], 'nutrition-score-fr_100g'
```

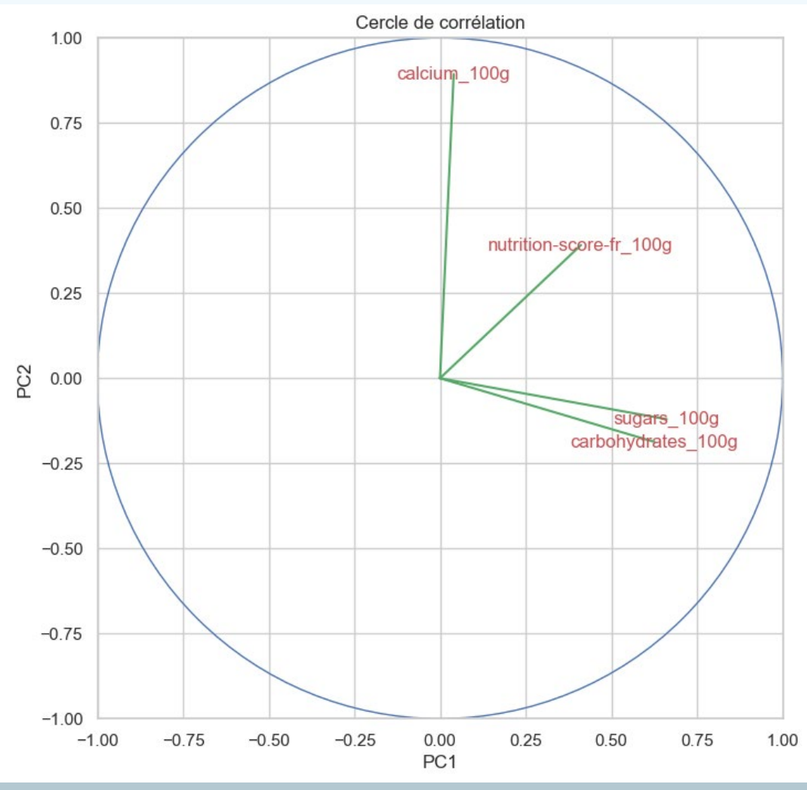
ACP

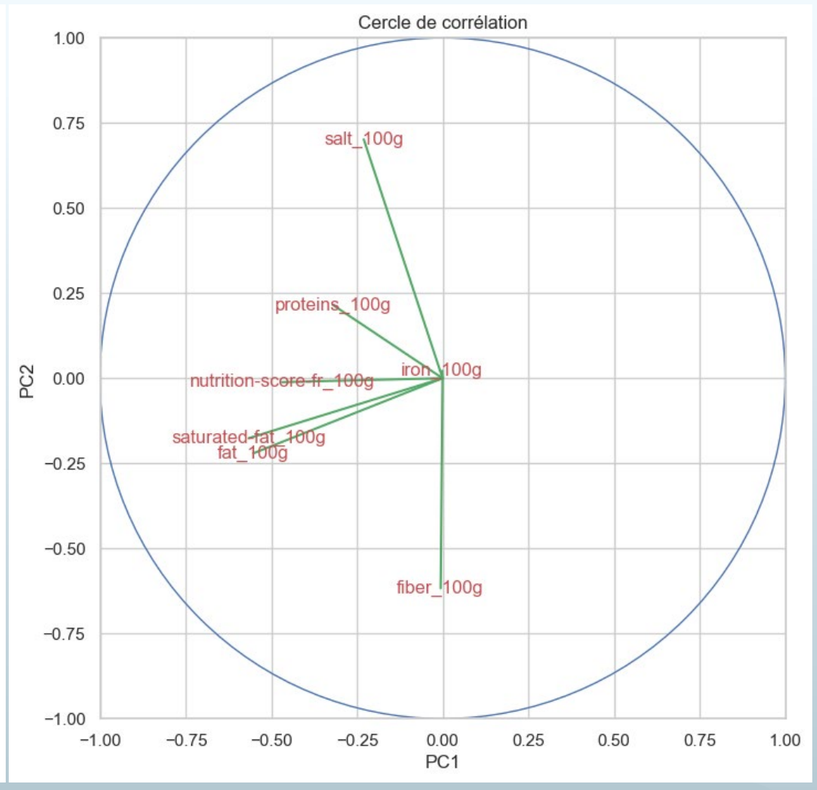
Analyse des Composants Principaux

Machine Learning Engineer

Cercle de corrélation :
Le cercle de corrélation illustre les relations entre les variables nutritionnelles. Par exemple, une forte corrélation négative entre le nutrition-score-fr_100g et le fiber_100g suggère que les aliments riches en fibres ont tendance à avoir un score nutritionnel plus bas. Ce visuel participe à l'élaboration de l'application, notamment au regard des saisies des utilisateurs. Il offre une perspective précieuse sur la structure des données nutritionnelles, permettant ainsi des décisions plus éclairées lors de l'élaboration des futurs algorithmes.



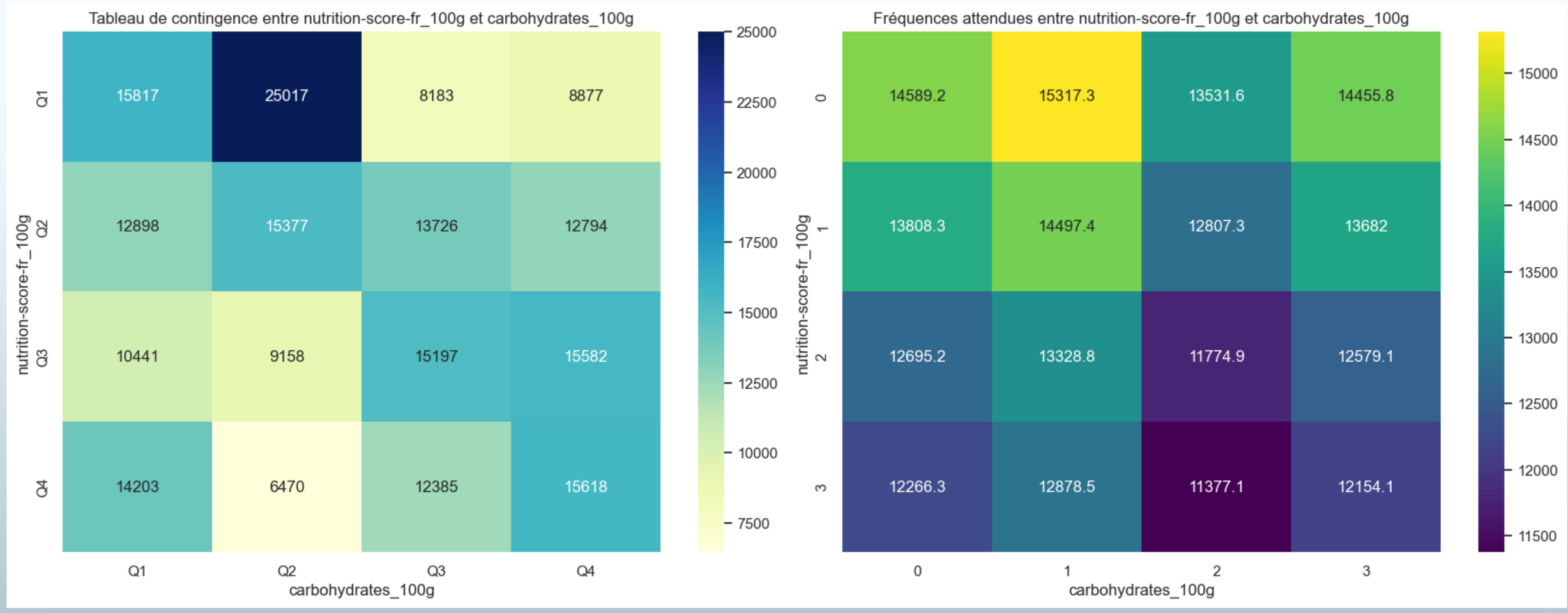




Cercle de Corrélation

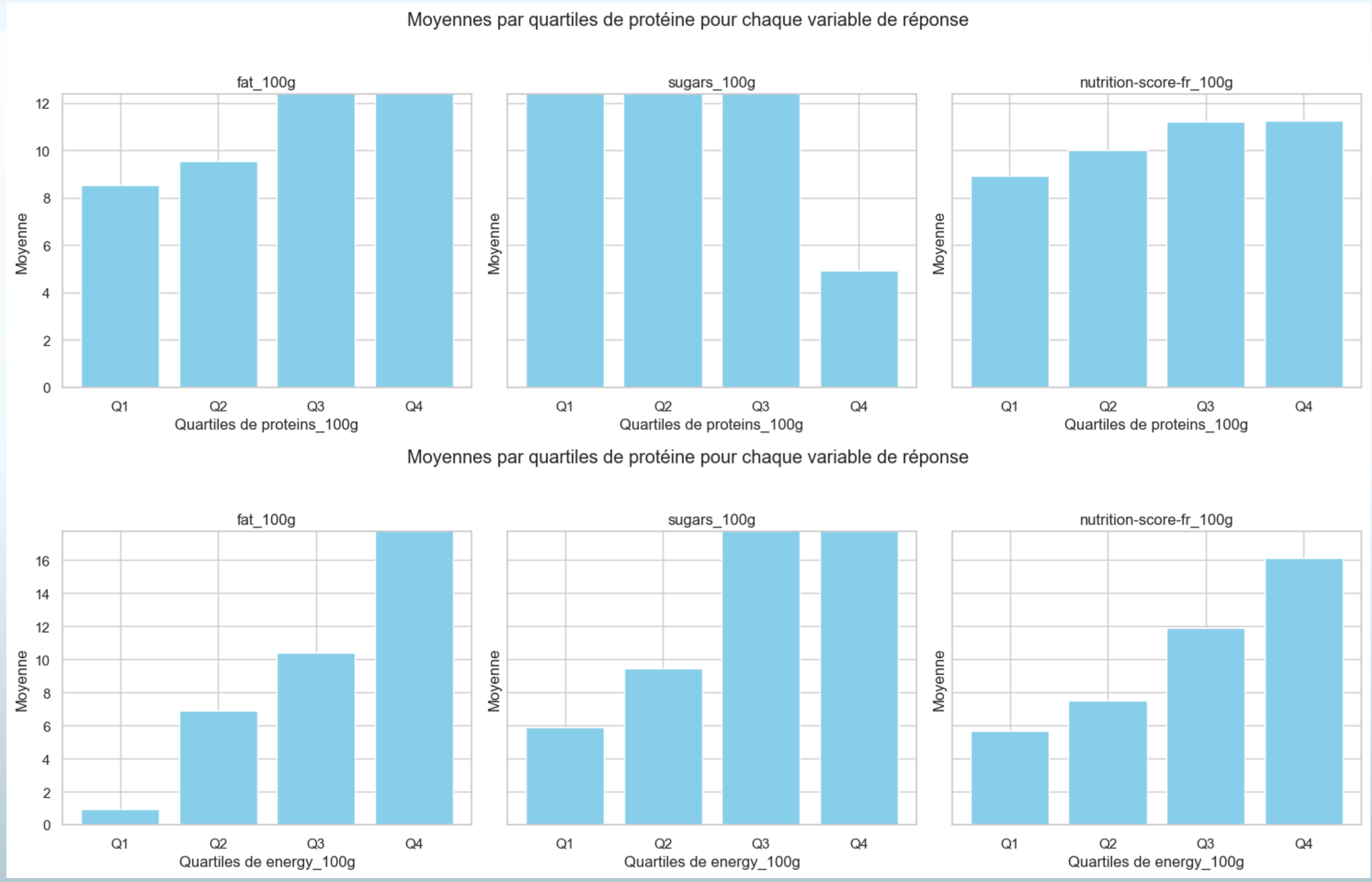
Machine Learning Engineer

Le test du Chi-carré est utilisé pour évaluer si deux variables sont indépendantes. Pour compléter l'analyse, deux cartes de chaleur (heatmaps) sont générées. Ces visualisations aident à comprendre visuellement les écarts entre les observations réelles et les valeurs attendues, basées sur l'hypothèse d'indépendance. Cette méthode est particulièrement utile pour examiner les interactions entre différents composants nutritionnels, permettant ainsi de déceler des patterns significatifs dans les données nutritionnelles.



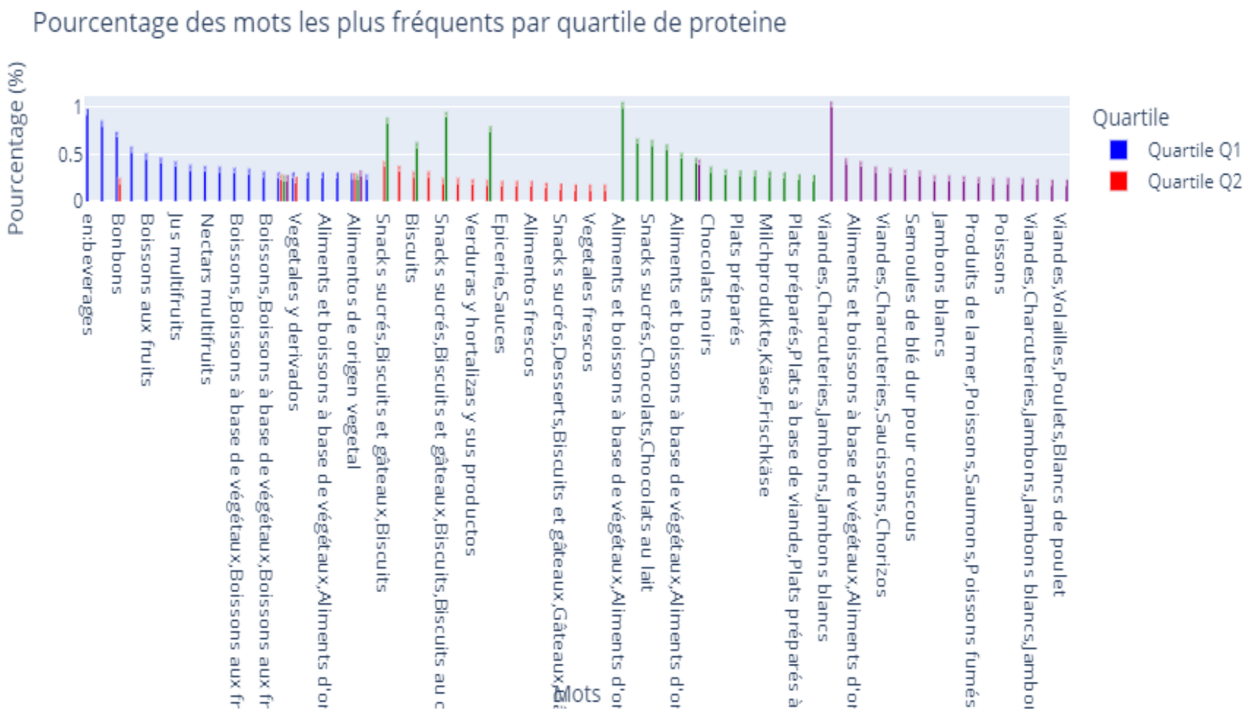
Le test Chi-carre

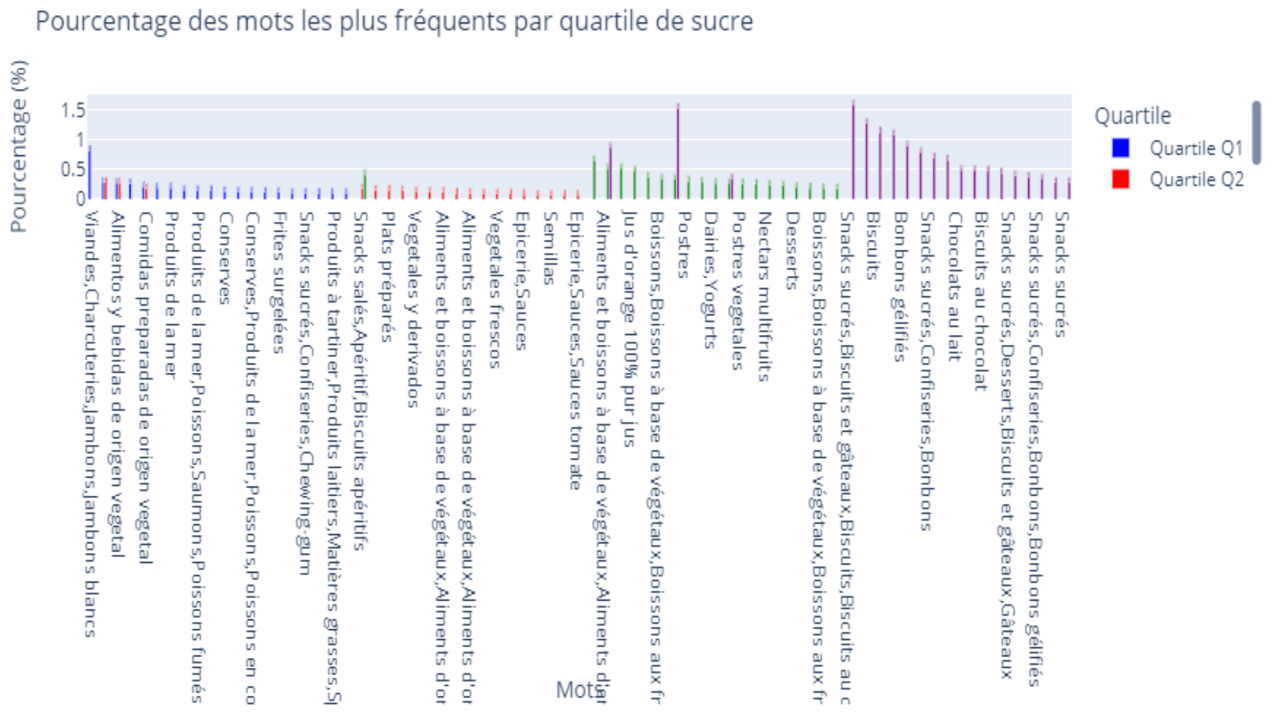
Machine Learning Engineer



ANOVA

Machine Learning Engineer



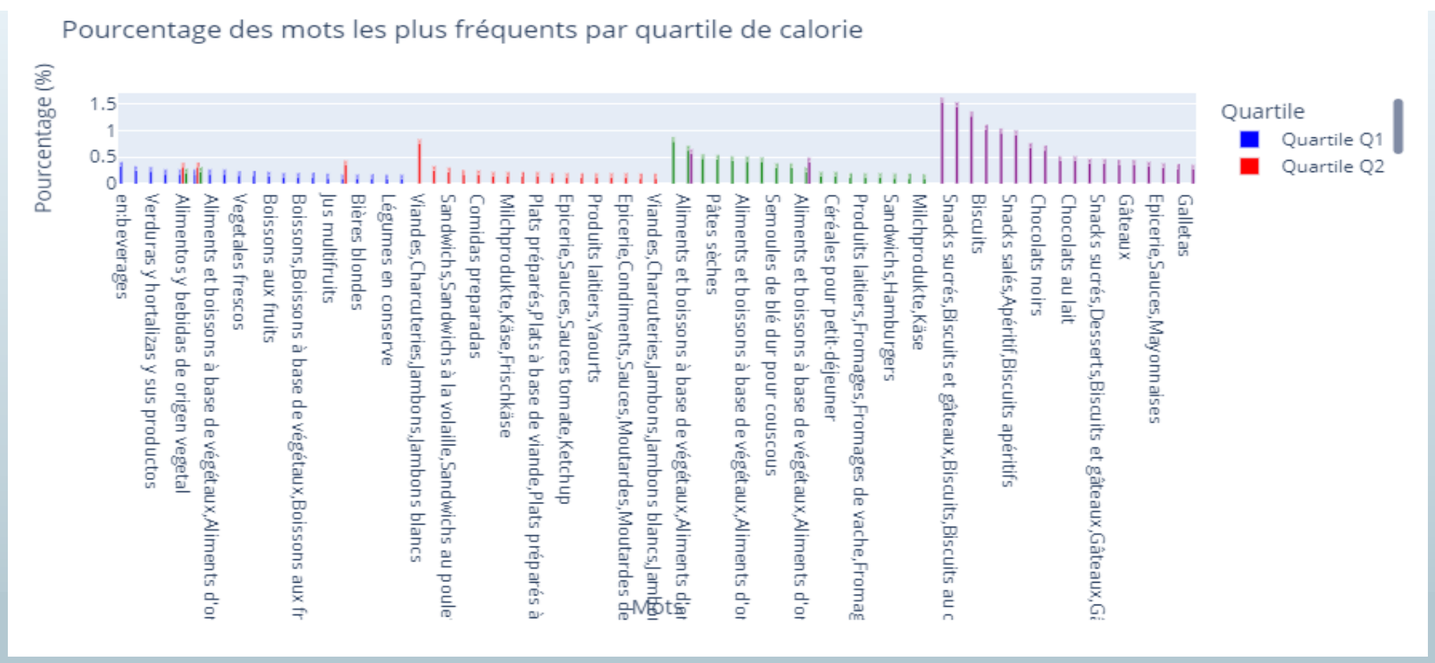


Q1

Q2

Q3

Q4



ANOVA

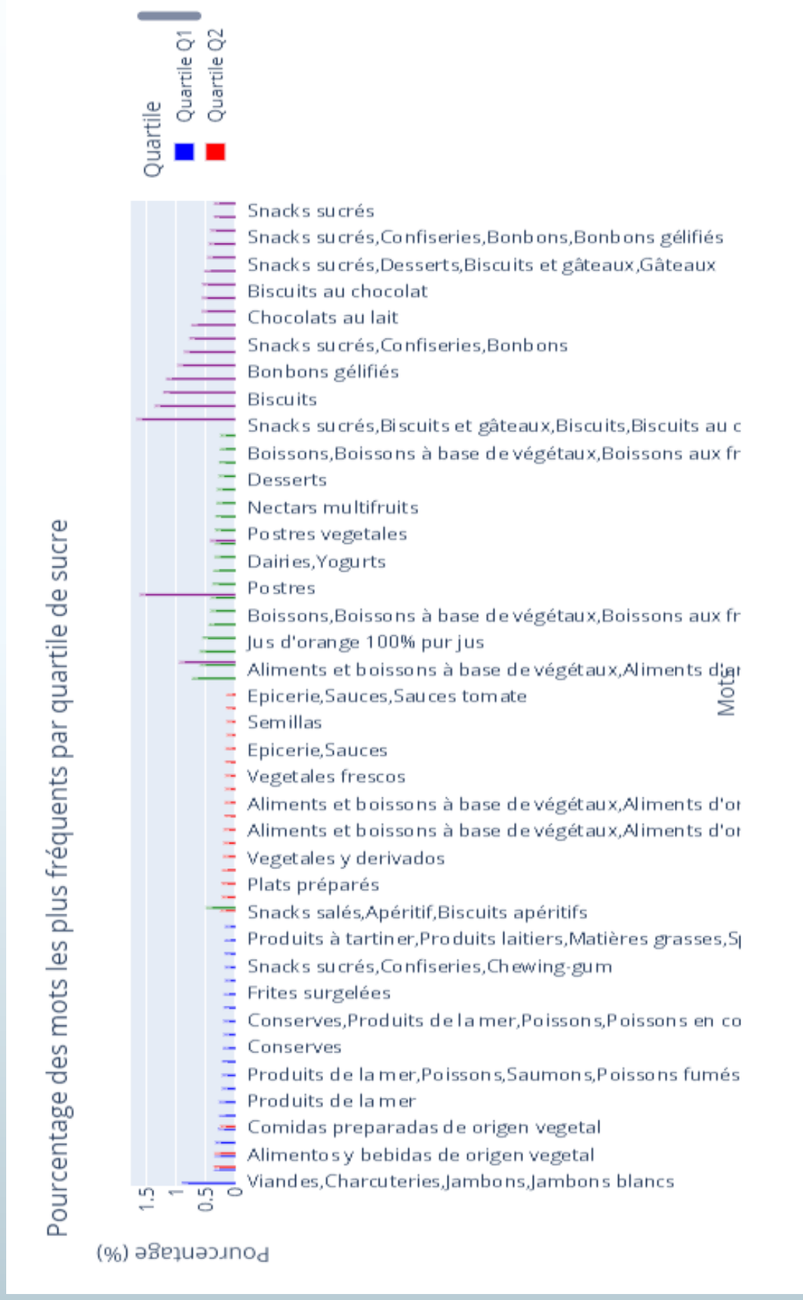
Faisabilité de l’application

Machine Learning Engineer

Analyse ANOVA : L'ANOVA est utilisée pour comparer les moyennes des variables sélectionnées entre différents quartiles. Elle permet de déterminer si les différences entre les moyennes des quartiles sont statistiquement significatives et si elles ont un impact sur d’autres variables.
Sur la base de la segmentation en quartiles, les catégories qualitatives de chaque quartile sont visuellement présentées. Par exemple, une concentration de produits liés aux sucreries est observée dans le dernier quartile de la variable sugar_100g.

Après la segmentation en quartiles et l'analyse ANOVA, les fréquences des mots sont recalculées et visualisées pour fournir une perspective complète sur la variation des propriétés des produits.

Cette analyse constitue la clé de voûte pour évaluer la faisabilité de notre application. Elle permet de dégager des tendances importantes, telles que la prédominance de certaines catégories de produits ou de revendications nutritionnelles dans des quartiles spécifiques. Ces tendances deviennent statistiquement prévisibles, ce qui pourrait nous permettre de suggérer des catégories de produits aux utilisateurs en se basant uniquement sur les premières valeurs nutritives qu'ils saisissent.



Q1

Q2

Q3

Q4

ANOVA
Faisabilité de l’application