Matrice de corrélation :

La matrice de corrélation est une table indique les coefficients de corrélation, qui mesurent le degré de relation linéaire (évaluer la dépendance) et indiquer la direction de la relation entre chaque paire de variables.

Les valeurs de corrélation comprissent entre -1 et 1.

- Si les deux variables ont tendance à augmenter et à diminuer en même temps, la valeur de corrélation est positive.
- Lorsqu'une variable augmente alors que l'autre diminue, la valeur de corrélation est négative.
- Une valeur de corrélation positive élevée indique que les variables mesurent la même caractéristique
- Une valeur de corrélation positive faible indique que les variables mesurent des caractéristiques différents ou ne pas clairement définis.
- Une variable de corrélation proche de 0, indique qu'il n'y a pas de corrélation entre les deux variables.

Observation

Une relation linéaire positive existe entre les variables avant et après l'imputation :

Varialble_1	Variable_2	Coefficient (avant imputation)	Coefficient (après imputation)
Nutriscore_fr_100g	Saturated_fat_100g	0.62	0.62
Nutriscore_fr_100g	Fat_100g	0.52	0.52
Nutriscore_fr_100g	Sugars_100g	0.41	0.45
Nutriscore_fr_100g	Carbohydrates_100g	0.32	0.29
Sugars_100g	Carbohydrates_100g	0.67	0.59
Salt_100g	Soduim_100g	1	0.88
Fat_100g	Saturated-fat_100g	0.69	0.64
Vitamine-a_100g	Calcuim_100g	0.51	0.18

Interprétation

Cette matrice de corrélation confirme mathématiquement des éléments logiques : Salt_100g est fortement corrélé avec soduim_100g, Fat_100g avec Saturated-fat_100g, ….

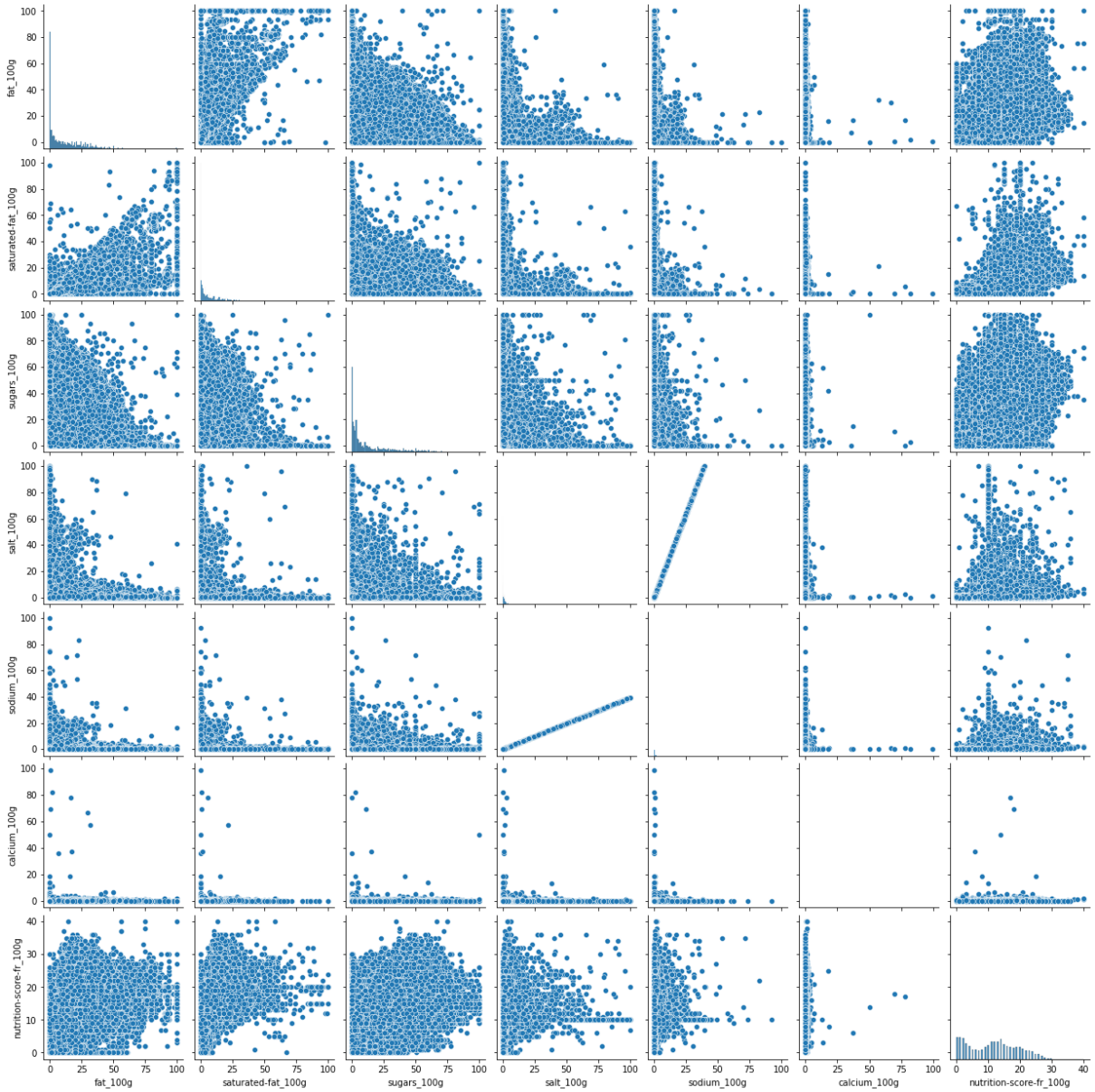
De plus, notre analyse ce dessus, nous permet de constater que le Nutriscore est bien corrélé avec :

- Saturated-fat_100g
- Fat_100g
- Sugar_100g

En revanche, on a constaté que le score de Nutriscore après l'imputation est légèrement faible de score de Nutriscore après l'imputation. On dit que ce dernier est sensible à notre choix de méthode d'imputation des données manquantes (KNN).

Analyse de pair plot : (analyse bivarié)

- **Avant imputation**



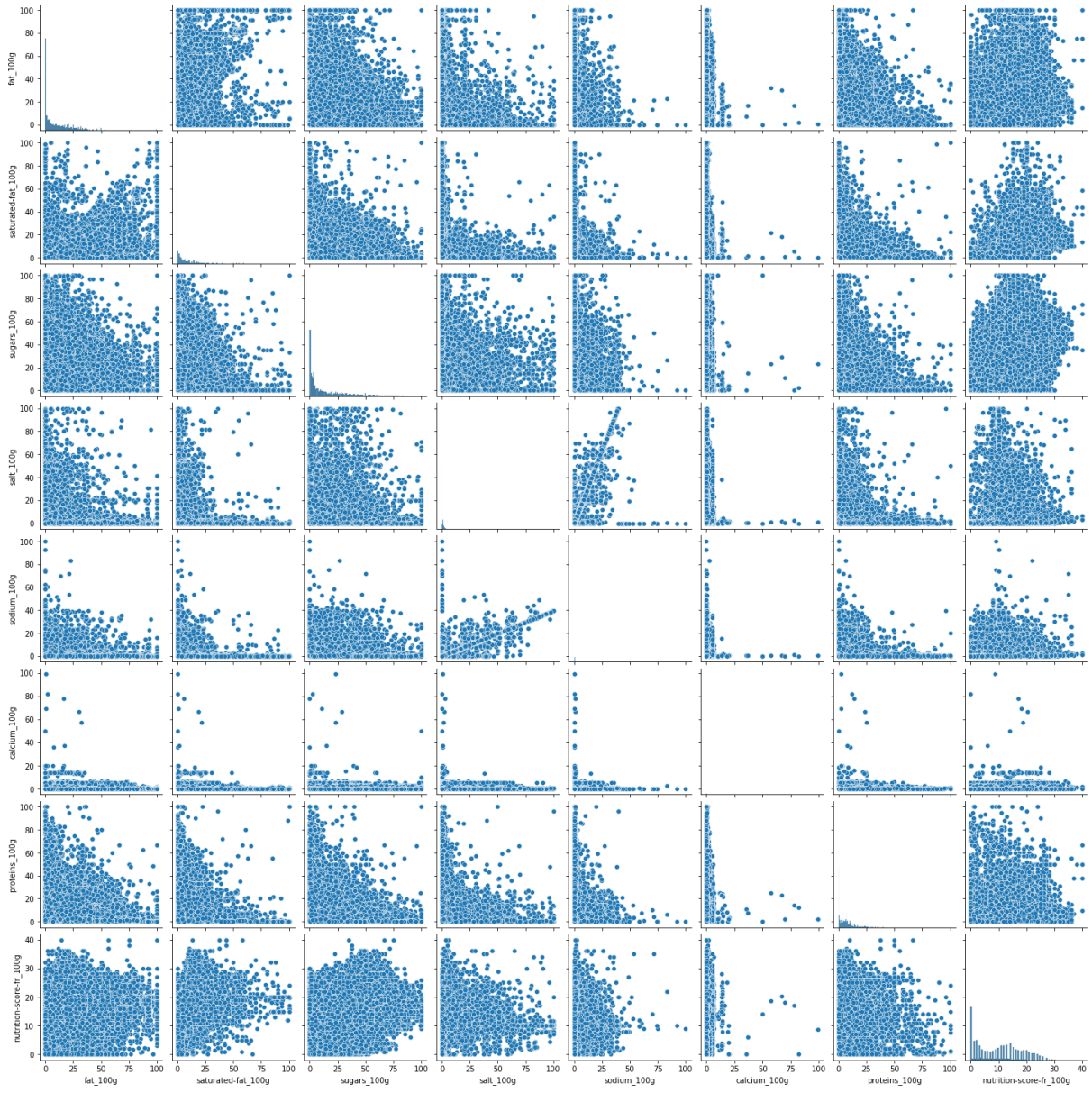
Interprétation de pairplot :

Plus une représentation tend vers une droite, plus les deux variables composants le graphique sont des doublons. Dans nos analyses, on peut constater que le sel et le sodium représentent une droite.

Regardons la dépendance entre le Nutriscore et les autres variables :

-Plus il y a de gras, gras saturé et de sucre, plus le Nutriscore est élevé

- **Après imputation**



Regardons la dépendance entre le Nutriscore et les autres variables :

- Moins il y a de sodium, plus le Nutriscore est élevé
- Plus il y a de gras, gras saturé et de sucre, plus le Nutriscore est élevé

Conclusion

Ces différentes analyses nous indiquent que pour construire un modèle de prédiction, nous pourrons nous appuyer sur les indicateurs suivants : Les gras, les gras saturés et les sucres.

Prédiction de Nutriscore :

- Nous possédant des exemples d'entrainement étiquetés, il s'agit d'un apprentissage supervisé.
- Nous cherchons une valeur contenue, il s'agit d'un problème de régression
- Nous travaillons avec plusieurs indicateurs, il s'agit d'un problème de régression multivariée.