**Analysez des données nutritionnelles**

Contenu

[Introduction 2](#_Toc497408857)

[Nettoyage 2](#_Toc497408858)

[Exploration 2](#_Toc497408859)

[Analyse Univariée 2](#_Toc497408860)

[Analyse multivariée 2](#_Toc497408861)

[Feature Enginnering 2](#_Toc497408862)

[Synthèse 2](#_Toc497408863)

[Evolution possible 2](#_Toc497408864)

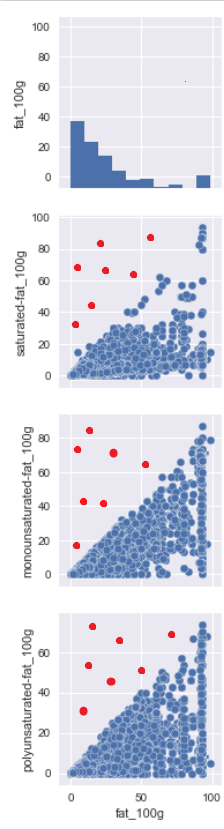
[Conclusion 2](#_Toc497408865)

# Introduction

Le site "Lamarmite" souhaite mettre en place un générateur de recettes saines. Pour ce faire, elle a besoin d’une analyse de données pour comprendre qu’est ce qui rend un plat sain ou non. Pour ce faire, elle nous met à disposition une base de données recensant divers produits de consommation avec différentes informations nutritionnelles. Pour ce faire, un premier nettoyage aura lieu avant de l'affiner en faisant une exploration des données.

# Nettoyage

L'ensemble de la phase de nettoyage ne sera pas expliqué mais en terme de decision, 3 principaux facteurs ont permis le nettoyage :

1. Suppression des features ayant trop peu de points
2. Suppression de la ligne complète en présence d'un outlier (car d'autres points sont surement tout aussi faux)
3. Suppression des features qui n'ont pas d'impact pour l'analyse. Par exemple son nom, l'url sur le site "openfoodfacts.com"
4. Les features contenant des listes (additives, ingredients\_that\_may\_be\_from\_palm\_oil, ingredients\_from\_palm\_oil\_n, labels) ont été réduite à un comptage d'attributs.

Dans un premier temps cela a permis de réduire à 84 features avant de passer à un début d'exploration.

# Exploration

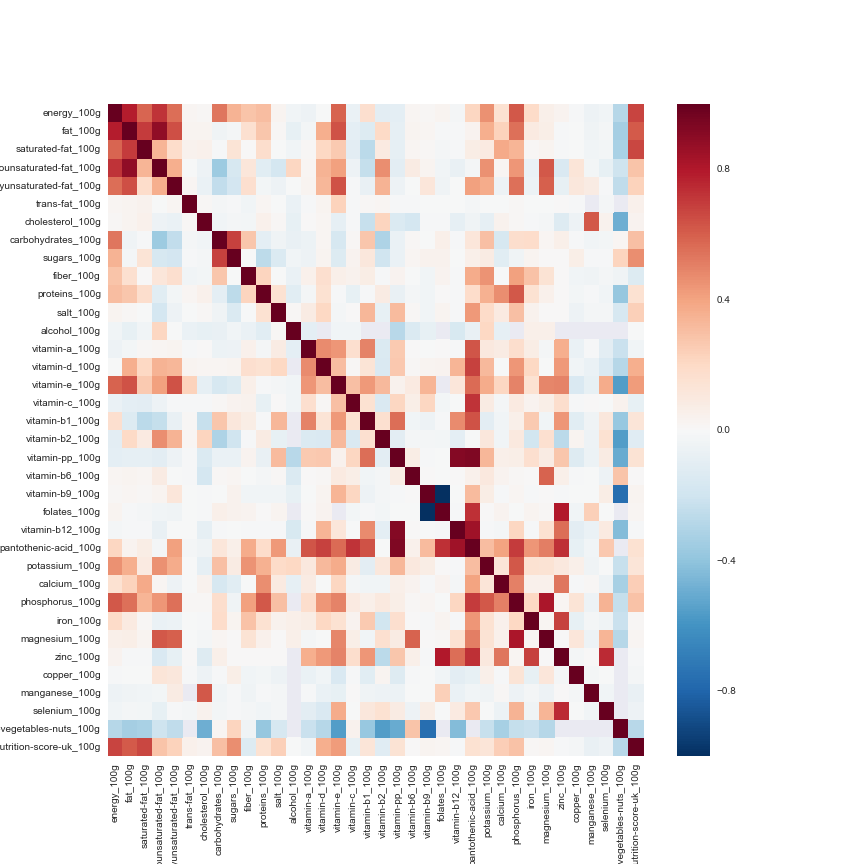
## Analyse Univariée

Pour l'analyse univariée, j'ai utilisé le "violinplot" pour les données continues (les features finissant par "\_100g") ou un diagramme de distribution pour les données discrètes (les features en \_n). L'intérêt du "violinplot" a été de pouvoir trouver/peaufiner d'autres règles de nettoyage car par exemple certaines feature avaient une masse négative de graisse, une masse aberrante de sel ou d'autres composants. A chaque fois, les règles ont été ajoutées au Notebook de Nettoyage pour nettoyer le dataset.

## Analyse Multivariée

En opposition à l'analyse multivariée, celle-ci a permis aussi de peaufiner certaines règles de tri. Par exemple, il n'était pas possible avant de voir qu'il y avait plus de certaines graisses spécifiques que de l'ensemble des graisses du produit (cf. les points rouge sur l'image ci-contre).

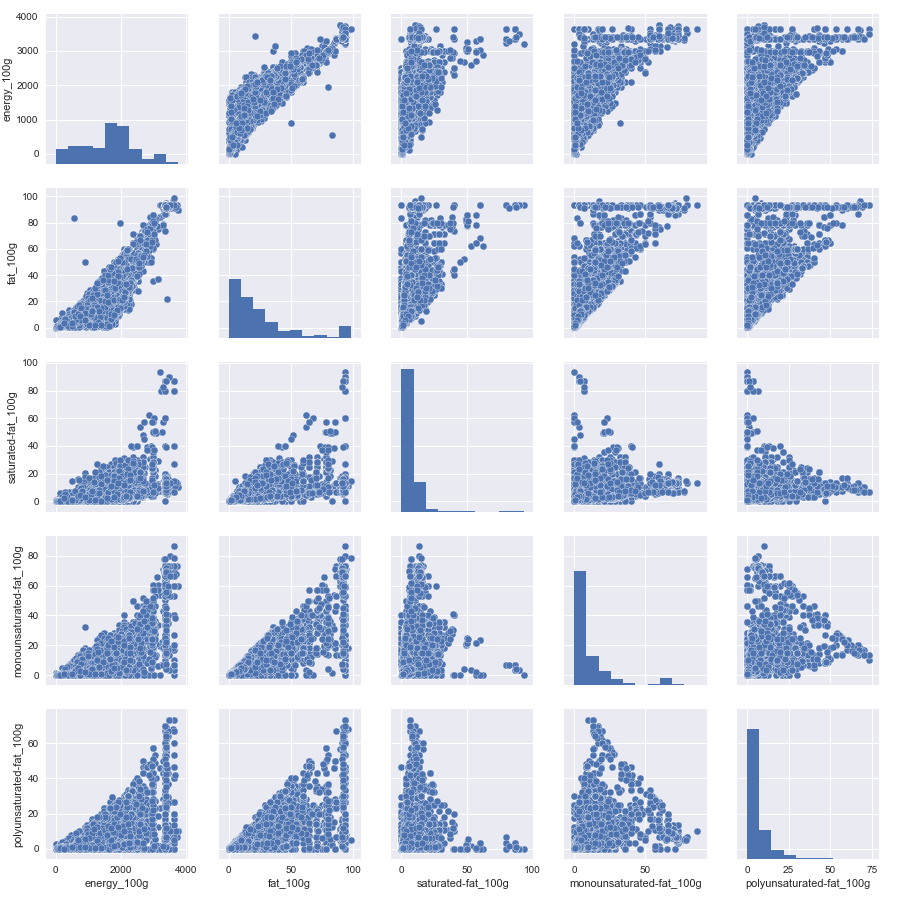
Par la suite, une matrice de corrélation a permis de mettre en relief les différentes corrélations.



On notera notamment :

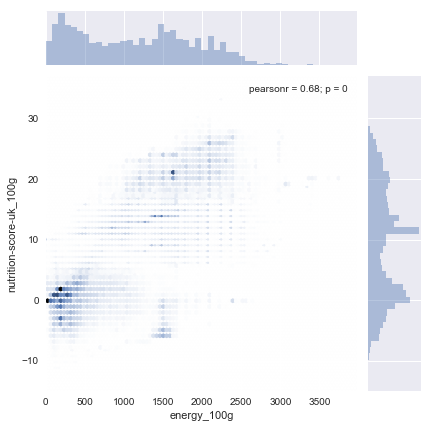
### La correlation entre les graisses et l'energie:

En effet il y a une grande corrélation entre différents type de graisses, mais aussi avec l'énergie. Cela est logique car la majorité de l'énergie vient en très grande partie de la brûlure des graisses. On peut donc tracer les "pairplot" entre chacun de ces 5 features.



Et on peut voir sur le graphe (ligne 2/colonne 1) que même sans graisse, le produit est calorifique mais très rapidement seul la graisse est l'apport d'énergie (le tracé est en "pointe").

On peut aussi remarquer avec le "jointplot" (car in simple scatter est trop dense) que le score est intimement lié à l'energie de l'aliment. Cela se comprend bien car une bonne nutrition passe par une nutrition ou lâpport d'energie est égale à la consommation de la personne. Plus l'elemnt est calorifique, moins il est sain.



Par la suite j'ai cherché a voir si le type de graisse impacte le FSA-score. On peut remarquer ci-dessous que

## Feature Enginnering

## Synthèse

## Evolution possible

# Conclusion