Formation DS Projet 3

Créer une application pour la santé publique

Utilisation de la base de donnée open source

Olivier Guntern

Parcours Data Scientist

2021

La Base de donnée OpenFoodFacts

- **Open Food Facts** est un projet collaboratif dont le but est de constituer une base de données libre et ouverte sur les produits alimentaires commercialisés dans le monde entier.
- **Licence de contenu** : Open Database License
- **Adresse** : openfoodfacts.org

- Exactitude des informations et données

Extraits des conditions d’utilisation de https://fr.openfoodfacts.org/conditions-d-utilisation#utilisation

Open Food Facts ne garantit pas l'exactitude des informations et données présentées sur le site et dans la base de données (y compris, et sans que cette liste soit limitative, des données des produits : photos, code barre, nom, dénomination générique, quantité, conditionnement, marques, catégories, origines, labels, certifications, récompenses, codes emballeur, ingrédients, additifs, allergènes, traces, informations nutritionnelles, informations écologiques etc.).

Ces informations sont entrées par les contributeurs du site et peuvent comporter des erreurs dues par exemple à l'inexactitude des informations sur les emballages et étiquettes, à la saisie manuelle des données, ou aux traitements informatiques des données.

Nous allons donc rester très attentif à ne pas réaliser une analyse purement Mathématique qui pourrait s’appliquer à des données fiables et validées.

Nous allons utiliser nos outils tout en ayant à l’esprit que surtout pour la santé publique la fiabilité de la base ne peut pas être assurée.

Idée d’application pour la santé

Une application ciblée sur le bon équilibre de la consommation journalière en graisse.

Il existe trois types de lipides: acides gras saturés, acides gras mono-insaturés (oméga-9: principalement l’acide oléique) et acides gras poly-insaturés (oméga 3 et 6). Contrairement aux oméga-9 (non essentiels car notre corps en produit), notre corps est incapable de fabriquer ces derniers.

Le rapport de la consommation d'oméga-6 et d'oméga-3 est un indicateur d'une bonne alimentation. Ce ratio, selon les recommandations de l'Afssa, devrait être proche de 5, c'est-à-dire que **l'alimentation devrait apporter 5 molécules d'oméga-6 pour une d'oméga-3.**

En réalité, les régimes occidentaux favorisent la consommation d'oméga-6, au détriment des oméga-3. Ainsi, en France, le ratio moyen est de 18 et aux États-Unis il peut monter jusqu'à 40.

* ref1

Les valeurs conseillées changes d’un pays à l’autre nous garderons donc celle de l’Afssa

Nous allons aussi nous intéresser au % de ‘mauvais cholestérol mais une explication du cholestérol est trop complexe, pour les informations.

Ref2*

Les ‘mauvaises’ graisses

Les différents types de lipides

Les lipides sont composés d’acides gras que l’on peut distinguer en deux grandes familles :

- Les Acides gras polyinsaturés (Oméga 3 et Oméga 6)
- Les Acides gras monoinsaturés (Oméga 9)
- Les acides gras trans.
- Les acides gras saturés

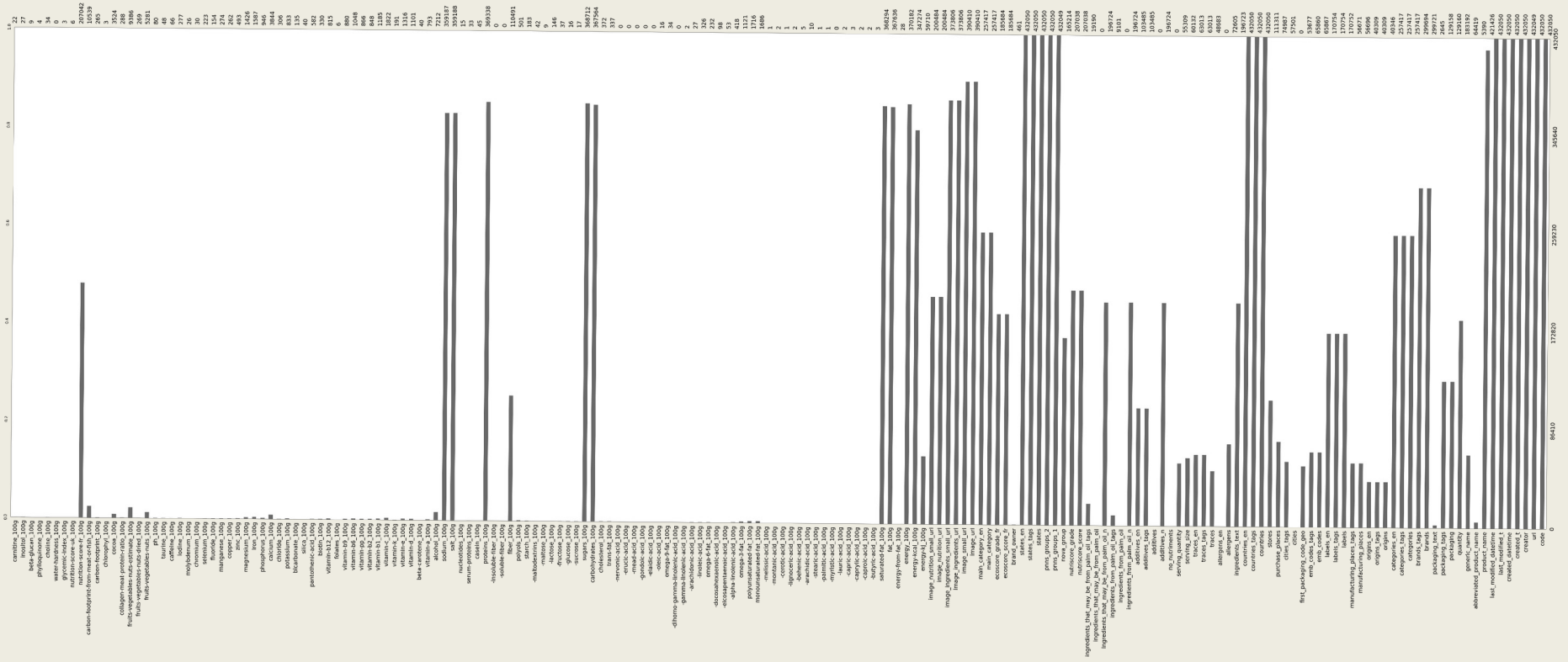
Nettoyage

Le jeu de données

Le dataset compte1895924 lignes et 186 variables

Beaucoup de valeurs manquantes des colonnes presque vides

Si dessous un graphique purement indicatif et volontairement miniature, nous allons Nous intéresser à une très petite quantité de variables pour notre projet.



Nettoyage

Les variables retenues pour notre problèmatique

fat_100g
La quantité de graisse totale pour 100 g de produit

monounsaturated-fat_100g
polyunsaturated-fat_100g

omega-3-fat_100g
omega-6-fat_100g
omega-9-fat_100g

saturated-fat_100g
trans-fat_100g
cholesterol_100g

Oui !! les oméga sont aussi des acides gras Mono-insaturés et poly-insaturés mais pour Différentes raisons les fabriquants mélanges Les deux inscriptions.

Il existe deux sources de cholestérol :
- l'une interne (2/3 du cholestérol total est produit par le foie) ;
*- l'autre externe (**1/3 du cholestérol total provient de l'alimentation**).*

a) Données incohérentes lors du total des variables sur 100 g, certains Individu ont un total au dessus de 100
Certaines variables dépasses aussi les 100

Nous obtenons le dataframe suivant :

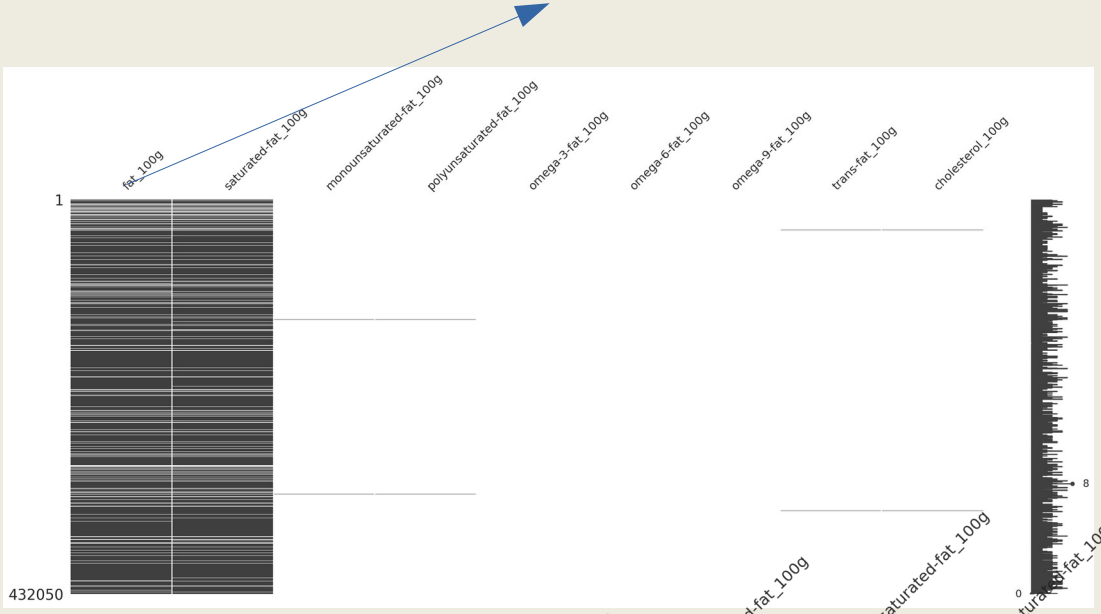
	fat_100g	saturated-fat_100g	monounsaturated-fat_100g	polyunsaturated-fat_100g	omega-3-fat_100g	omega-6-fat_100g	omega-9-fat_100g	trans-fat_100g	cholesterol_100g
count	363976.000000	363976.000000	1380.000000	1409.000000	1041.000000	164.000000	13.000000	330.000000	335.000000
mean	13.388919	5.269172	21.139489	8.213140	2.589032	9.067331	29.666577	0.242686	0.121368
std	16.185686	7.701982	24.381481	11.218018	4.248957	10.904617	20.073996	0.840065	1.438004
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.005500	0.000000	0.000000
25%	1.000000	0.200000	2.600000	1.700000	0.477000	1.100000	27.000000	0.000000	0.000000
50%	7.400000	2.000000	9.500000	4.000000	1.700000	5.250000	31.000000	0.000000	0.000000
75%	21.600000	7.600000	28.000000	9.700000	3.000000	15.000000	32.000000	0.200000	0.011750
max	99.980000	99.900000	82.200000	69.000000	60.000000	58.000000	64.000000	12.000000	24.000000

Pourcentage de valeurs remplies

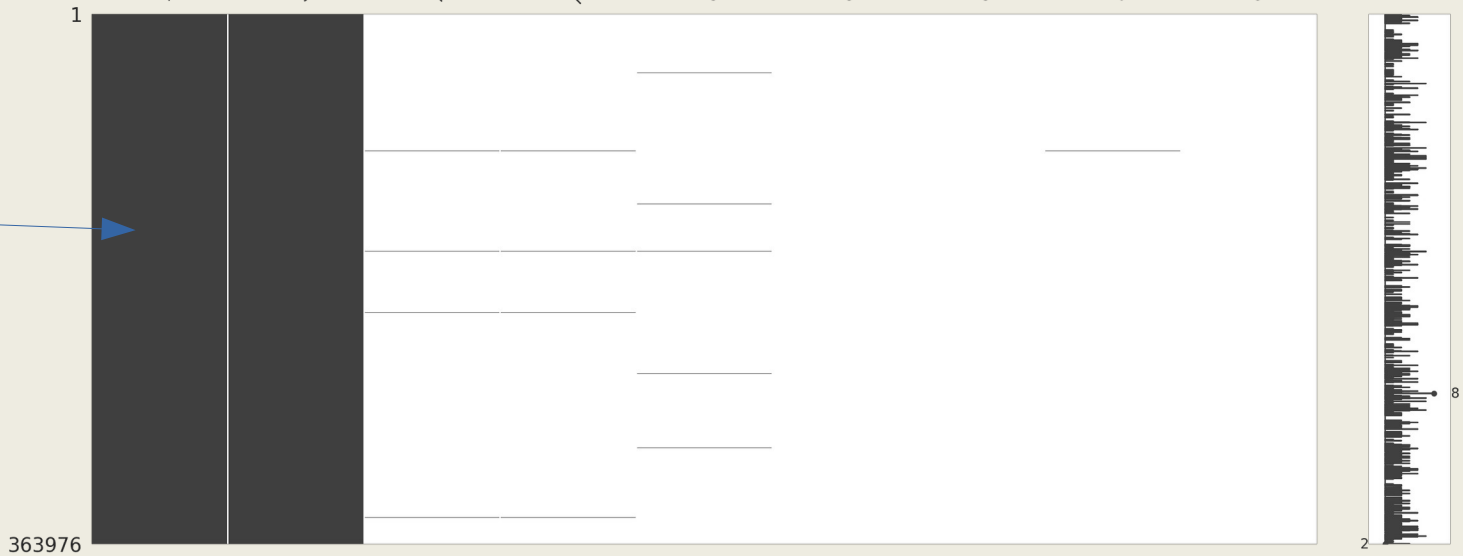
nombre de lignes : 432050

données par ligne après supression des valeurs isnull
---
fat_100g
367636 nombres de lignes remplies
---
saturated-fat_100g
368294 nombres de lignes remplies
---
monounsaturated-fat_100g
1686 nombres de lignes remplies
---
polyunsaturated-fat_100g
1716 nombres de lignes remplies
---
omega-3-fat_100g
1121 nombres de lignes remplies
---
omega-6-fat_100g
232 nombres de lignes remplies
---
omega-9-fat_100g
34 nombres de lignes remplies
---
trans-fat_100g
337 nombres de lignes remplies
---
cholesterol_100g
372 nombres de lignes remplies
9 sur 9 variable avec plus de 10 pourcent de valeurs manquantes

b) Suppression des individu ne contenant pas de Graisse, nous utilisons la colonne fat_100g



Voici



Analyse monovariée

Tous nos produits contiennent des graisses saturées

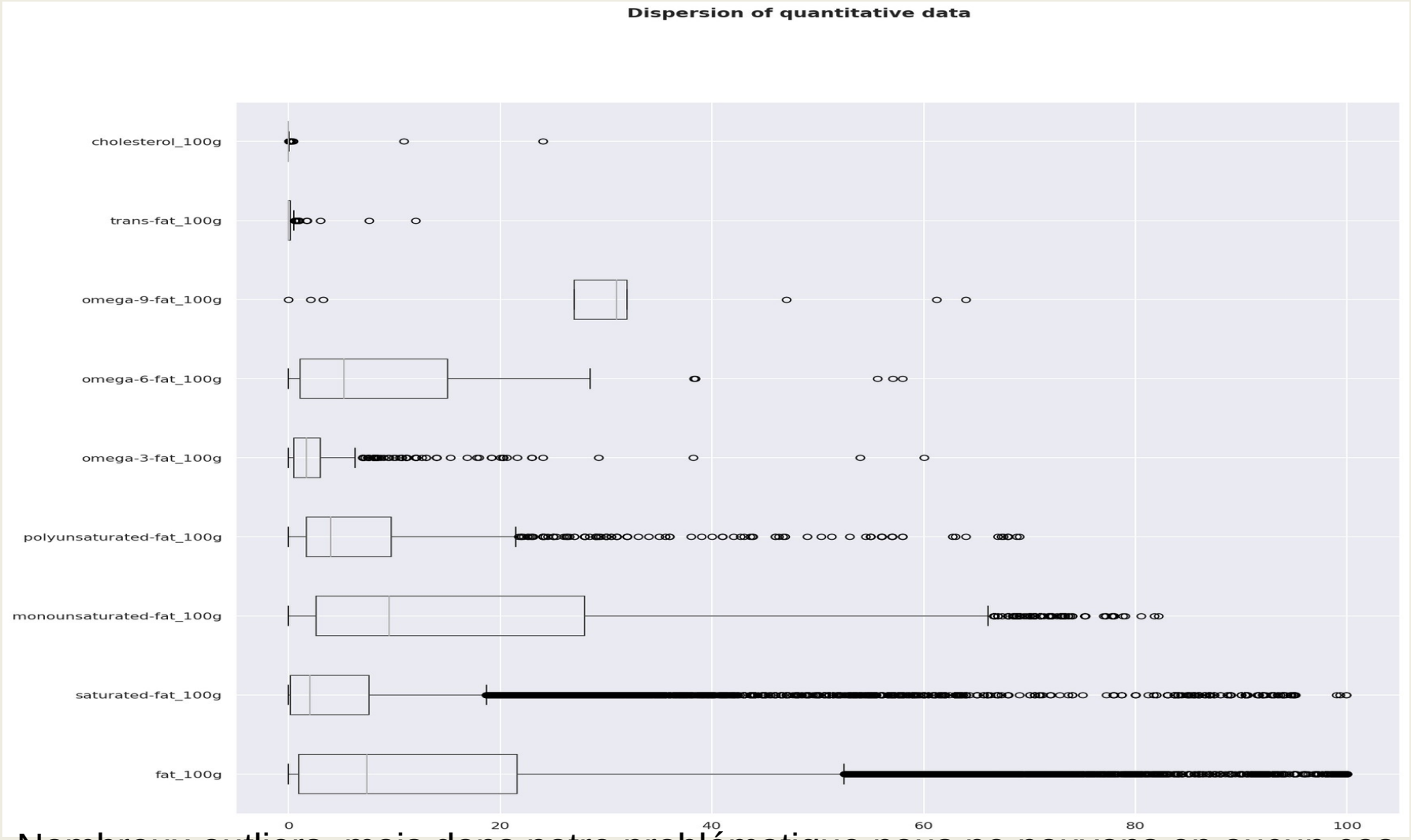
fat_100g	**363976**
saturated-fat_100g	**363976**
monounsaturated-fat_100g	1380
polyunsaturated-fat_100g	1409
omega-3-fat_100g	1041
omega-6-fat_100g	164
omega-9-fat_100g	13
trans-fat_100g	330
cholesterol_100g	335

Ce qui n’est pas une surprise mais c’était à souligner

Statistiques descriptives

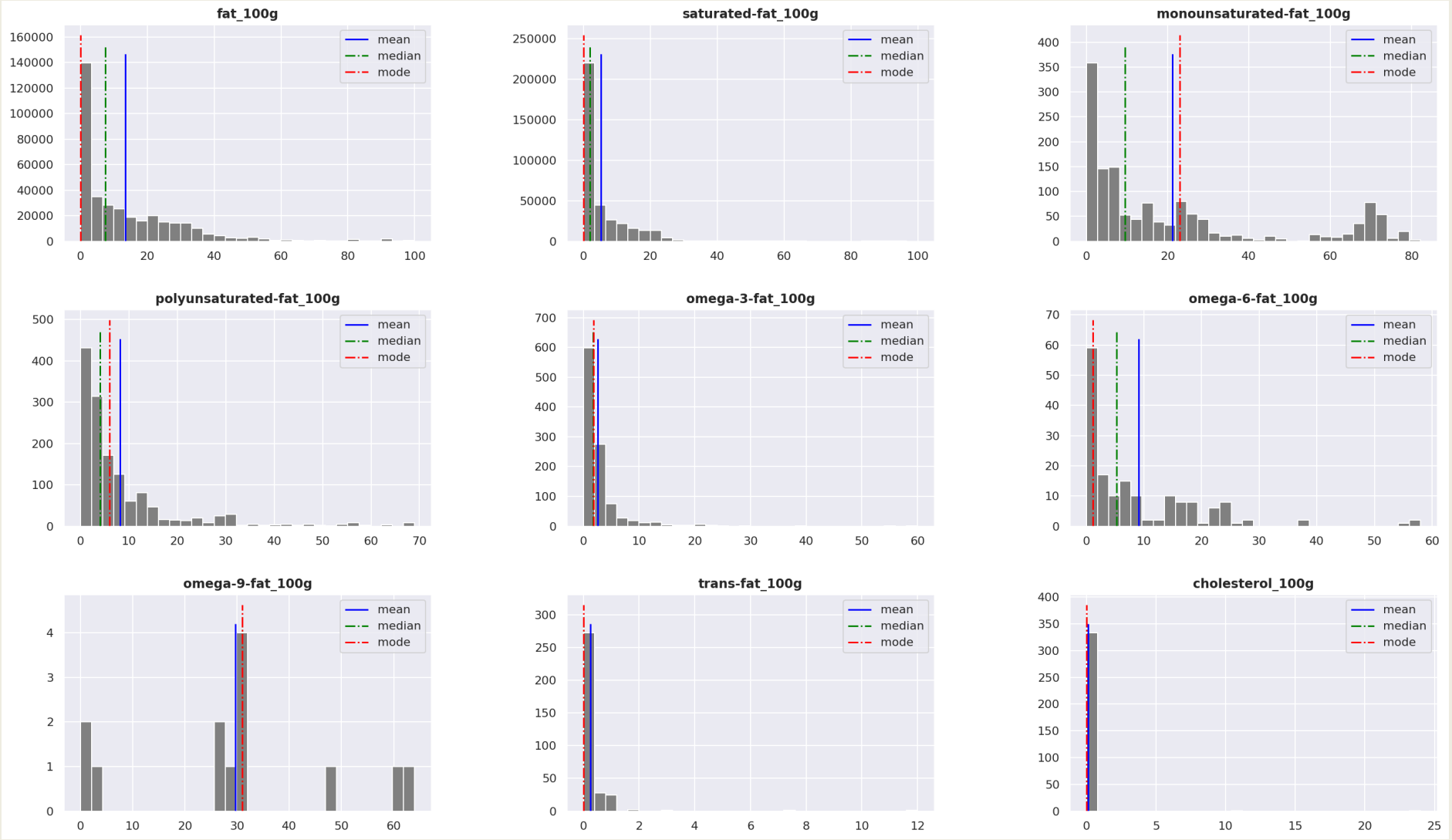
Variable	minimum	maximum	somme	moyenne	médiane	Variance Var	Variance Std	quartile 1	quartile 2	quartile 3
fat_100g	0.0000	99.98	4.873245e+06	13.388919	7.40	261.975699	16.185663	1.000	7.40	21.60000
saturated-fat_100g	0.0000	99.90	1.917852e+06	5.269172	2.00	59.320366	7.701972	0.200	2.00	7.60000
monounsaturated-fat_100g	0.0000	82.20	2.917250e+04	21.139489	9.50	594.025841	24.372645	2.600	9.50	28.00000
polyunsaturated-fat_100g	0.0000	69.00	1.157231e+04	8.213140	4.00	125.754605	11.214036	1.700	4.00	9.70000
omega-3-fat_100g	0.0000	60.00	2.695182e+03	2.589032	1.70	18.036291	4.246915	0.477	1.70	3.00000
omega-6-fat_100g	0.0000	58.00	1.487042e+03	9.067331	5.25	118.185601	10.871320	1.100	5.25	15.00000
omega-9-fat_100g	0.0055	64.00	3.856655e+02	29.666577	31.00	371.967985	19.286472	27.000	31.00	32.00000
trans-fat_100g	0.0000	12.00	8.008638e+01	0.242686	0.00	0.703571	0.838791	0.000	0.00	0.20000
cholesterol_100g	0.0000	24.00	4.065822e+01	0.121368	0.00	2.061684	1.435857	0.000	0.00	0.01175

Outliers



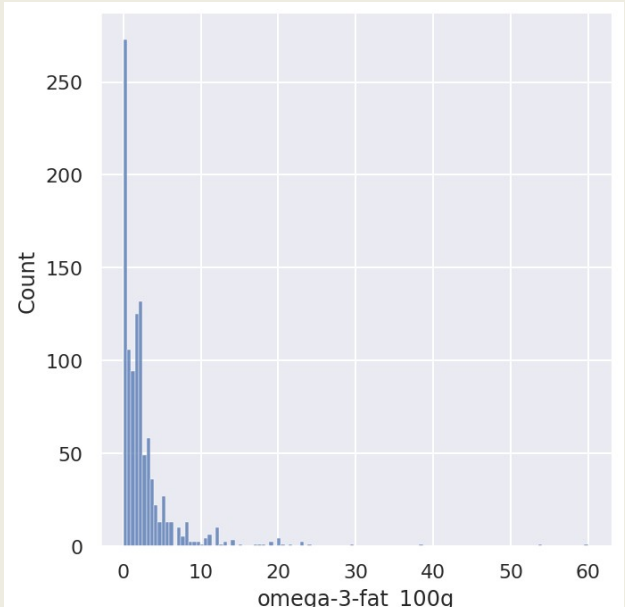
Nombreux outliers, mais dans notre problématique nous ne pouvons en aucun cas affirmer ou infirmer que c'est une valeur erronée, en effet quoi de plus normal qu'un fromage, qu'une graisse ou qu'une huile soit largement au dessus de la médiane, voir À 100 %.

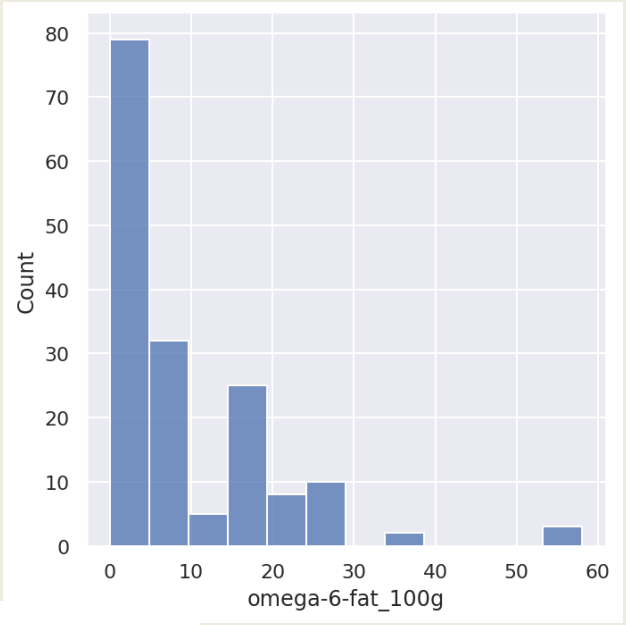
Distribution avec repères statistiques

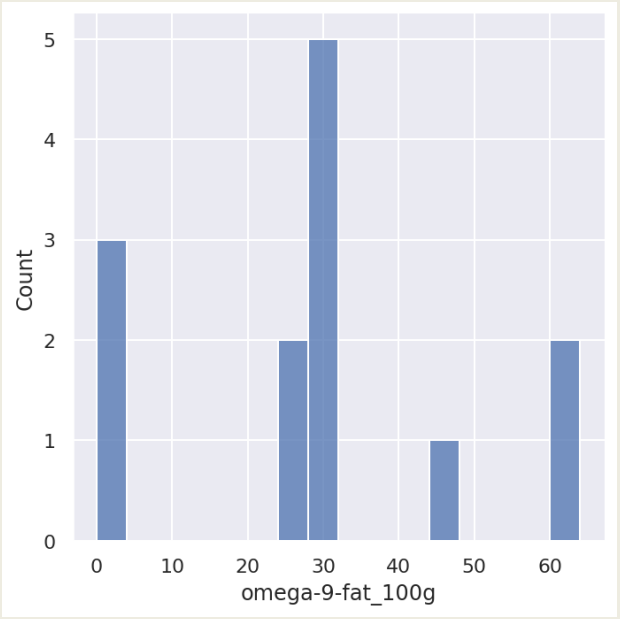


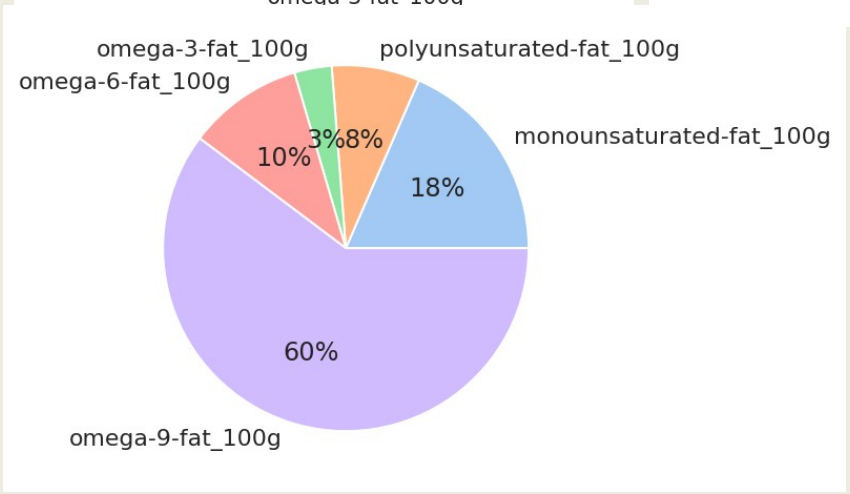
Analyse multivariée

Répartition des acides gras



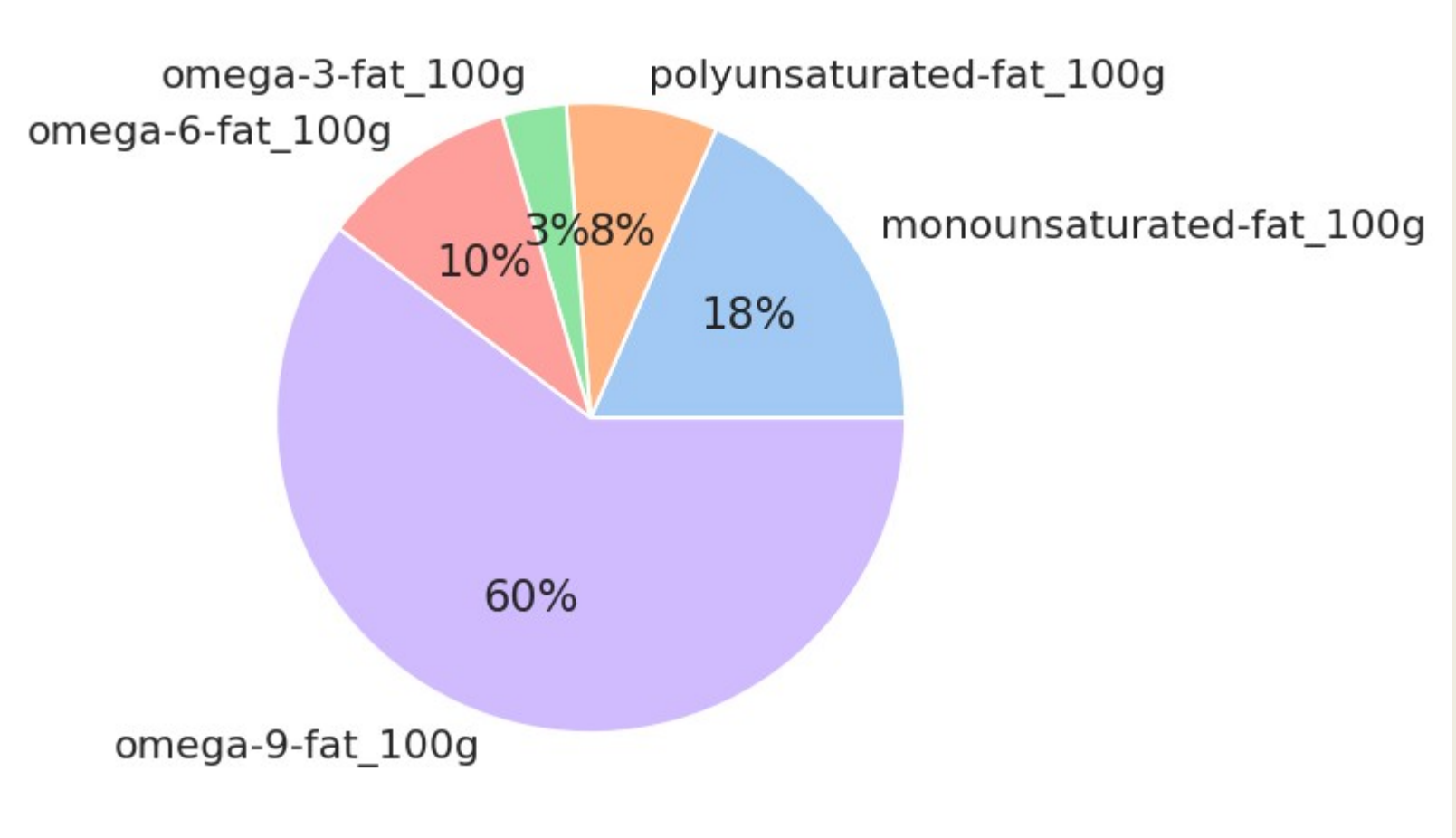






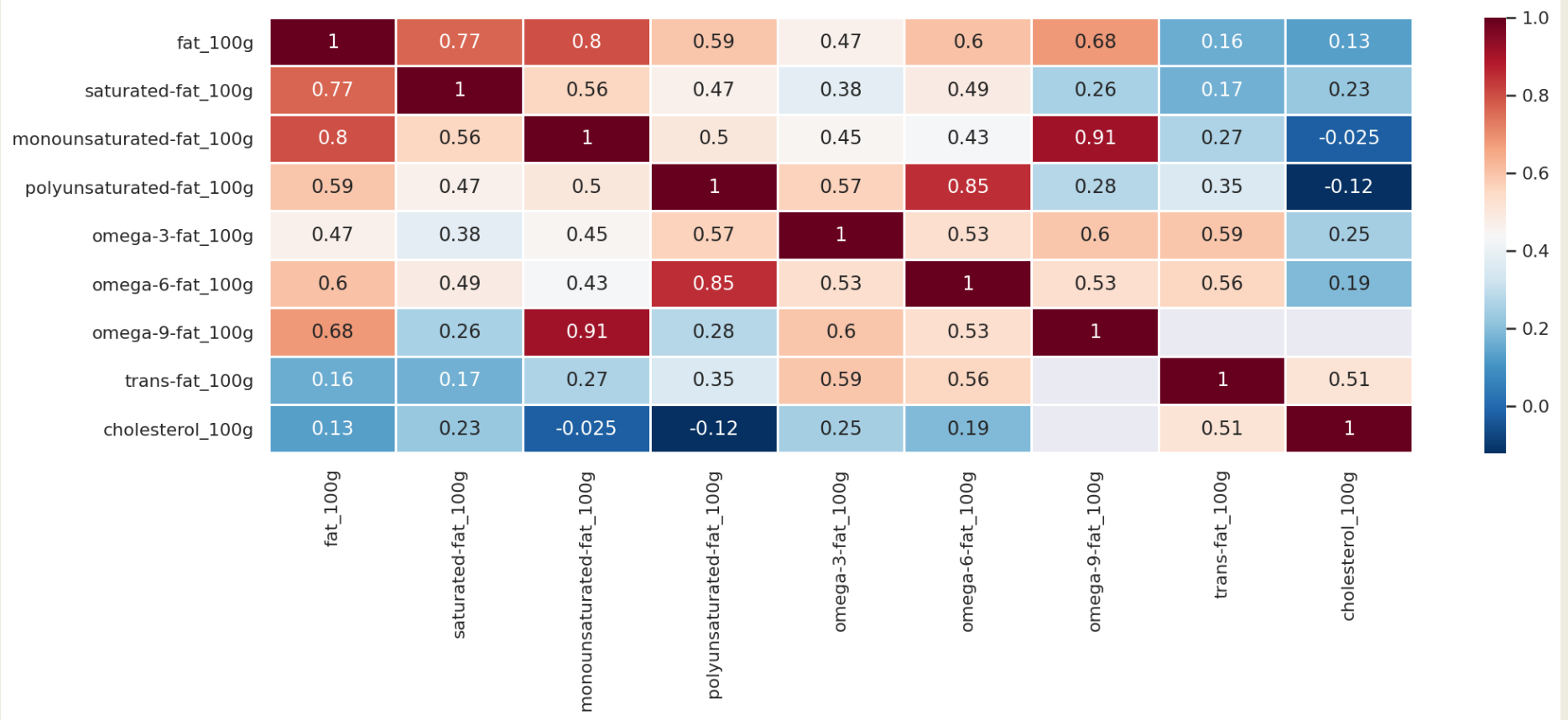
Voir page suivante

Regardons de plus près les médiane Oméga et acide gras correspondant



Nous voyons que les couples : oméga-3, oméga-6 et acide gras poly-saturées,
Ainsi que oméga-9 et graisse mono-insaturées on une certaine proportionalité.

Tableau de corrélation de Kendall



Les oméga-3 et les oméga-6 et acide ont une corrélation avec les acide gras poly-saturées, Et les oméga-9 avec les graisses mono-insaturées Dans une certaine proportionnalité, ce qui est normal mais à prendre avec discernement.

ANOVA

couple omega-9-fat_100g / trans-fat_100g F-Statistic = 5.650 , p= 0.017

couple cholesterol_100g / omega-9-fat_100g F-Statistic = 7.013 , p= 0.008

couple cholesterol_100g / trans-fat_100g F-Statistic = 1.641 , p= 0.200

Les autres couples ont un p=0

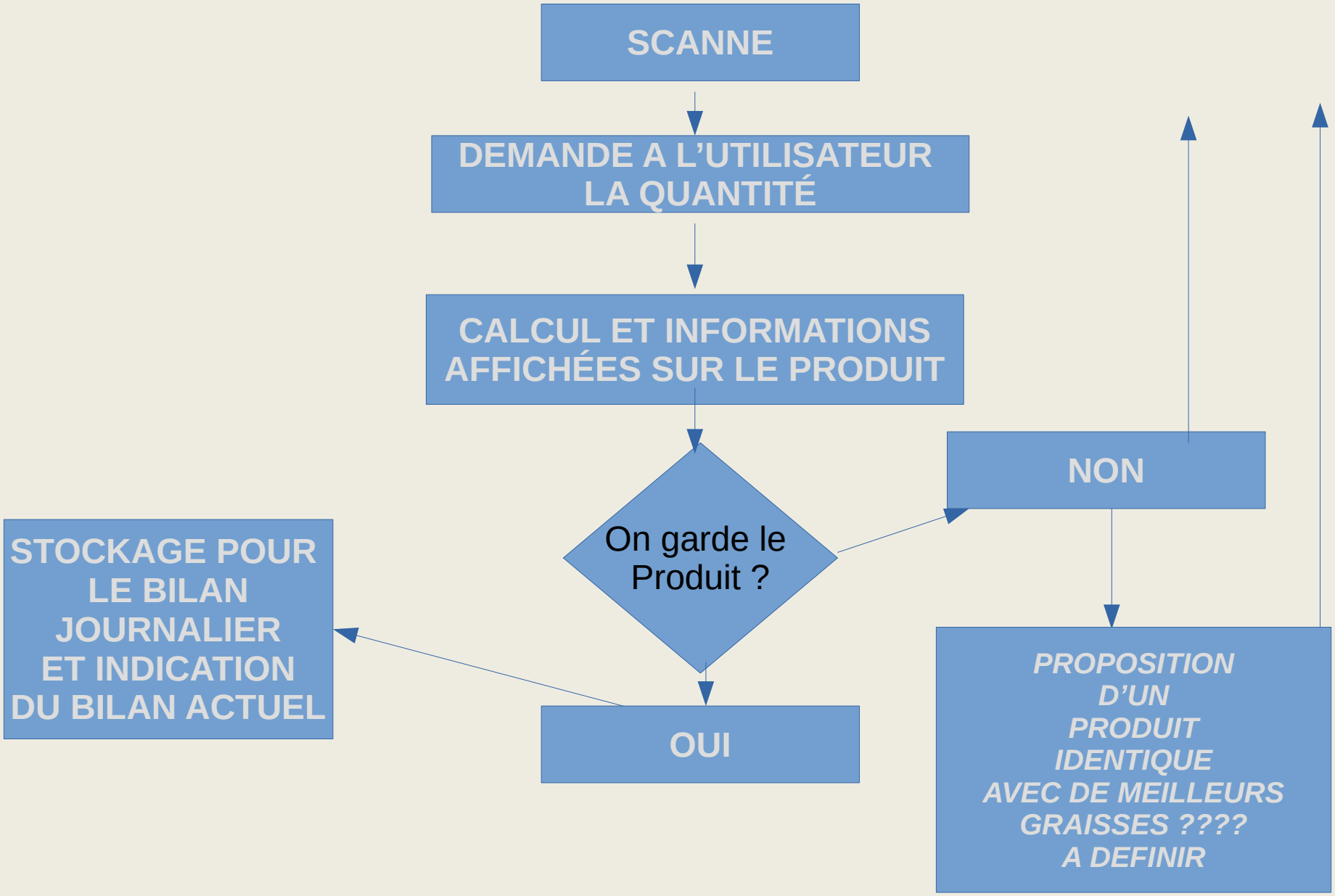
Conclusions

les principales observations

Les point forts

- Nous avons une nouvelle base sur les graisses, mathématiquement cohérente
- Nous avons les informations pour notre application
- Cette application d’un simple scanne et d’une indication de quantité absorbée, pourra donner un suivi, en indiquant les carences ou surdosages des bonnes et mauvaises graisse à l’utilisateur.
- Le point intéressant sera de lui proposer un aliment pouvant en remplacer un mauvais en comparant les taux de graisse.

Schéma application



Références :

Ref1
https://www.vidal.fr/sante/nutrition/corps-aliments/lipides-energie/acides-gras-satures-insatures-trans.html

Ref2
https://fr.wikipedia.org/wiki/Cholesterol