Analyse avant conception d'une application au service de la santé publique

Soutenance Projet 3 Data Scientist Agnès Regaud juin 2023

Sommaire

1.Problématique

Application en lien avec l'alimentation

- Appel à projets de Santé publique France
- Open Food Facts : base de données de produits alimentaires
- Application d'aide pour trouver des produits plus sains :
  - Note sur 100
  - Caractère nutritionnel, sain et écologique
  - Additifs, huile de palme, label bio, sel, fibres, fruits-légumes-noix, nutriscore
  - Proposition de produit plus sain éventuellement

2. Etude du jeu de données (premier nettoyage, données manquantes, données particulières)

Etude préliminaire et premier nettoyage du fichier de données de Open Food Facts (320 772 produits)

Suppression	- Suppression de produits non vendus en France métropolitaine ou DOM-TOM - Suppression de données erronées → 90 550 produits restants
Gestion colonne Nom du produit	- Imputation du Nom Générique et éventuellement de la marque du produit - Suppression des produits sans Nom ni Nom Générique → 90 512 produits restants
Colonne Label bio	Ajout d'une colonne indiquant si le produit est bio ou non à partir de la colonne des labels, de la liste des ingrédients ainsi que du nom du produit avec les mots clé : « bio » et « AB »

Colonne Catégorie de produit

- Identification de la colonne comportant un nombre limité de catégories : ('pnns_groups_1')

- Imputation de valeurs manquantes dans cette colonne à partir de sa jumelle ('pnns_groups_2')

- Regroupement des catégories en présence de synonymes, au final 9 catégories : 'Sugary snacks', 'Beverages', 'Fish Meat Eggs', 'Composite foods', 'Fruits and vegetables', 'Milk and dairy products', 'Fat and sauces', 'Salty snacks', 'Cereals and potatoes'

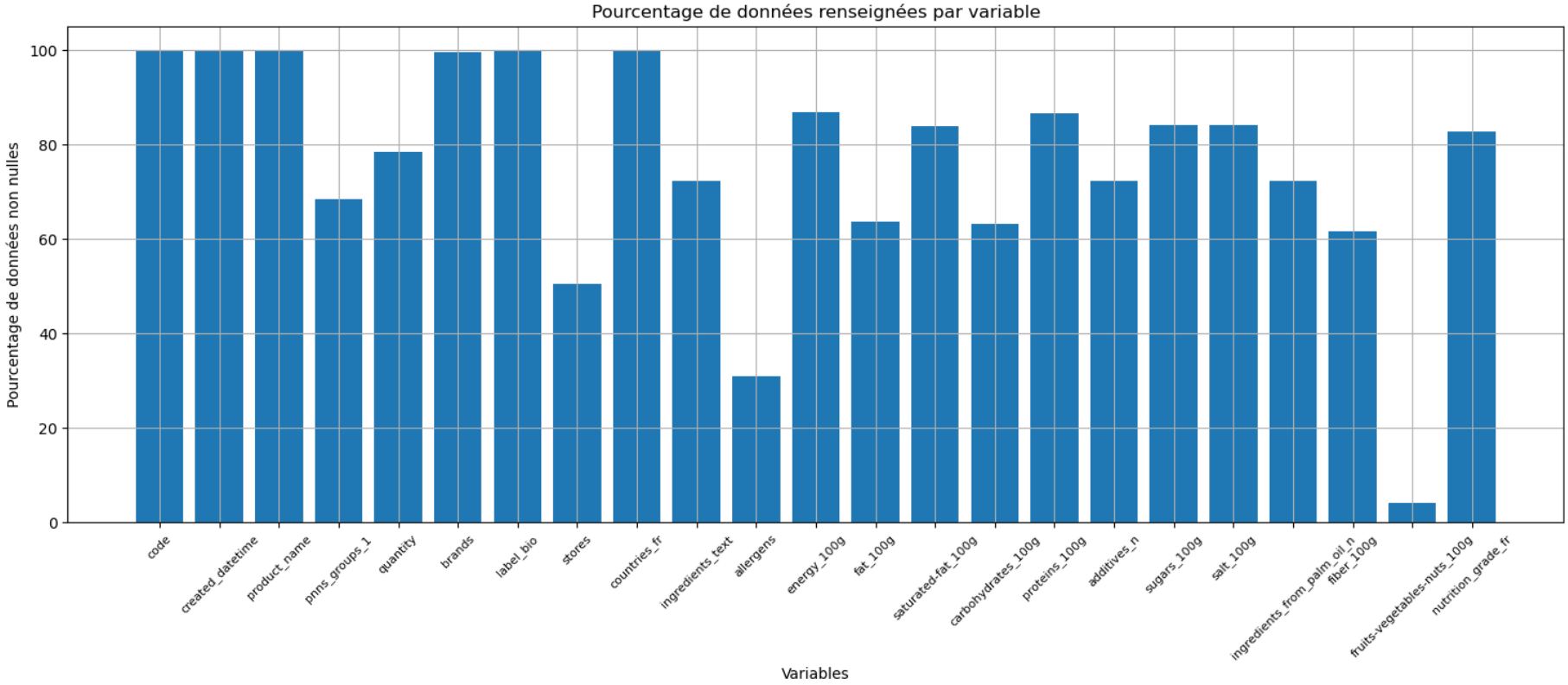
- Suppression de produits avec plus de 3 valeurs manquantes dans 12 des colonnes principales (liste d'ingrédients, énergie, graisses saturées, glucides, protéines, additifs, sucre, sel, huile de palme, fibres, fruits-légumes-noix, nutriscore)
→ 73 270 produits restants

- Imputation de catégorie à partir du nom du produit par mots clé

- Suppression des produits sans catégorie
→ 60 470 produits restants

3.Description et analyse de certaines variables

Diagramme à barres des données manquantes

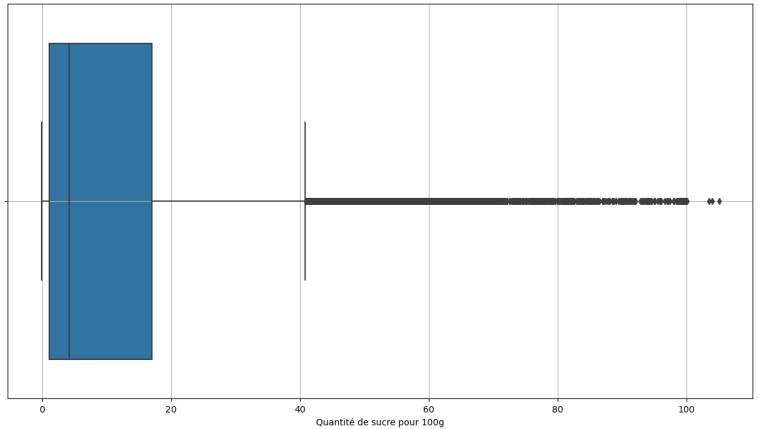


Détection et gestion des valeurs aberrantes-extrêmes pour :

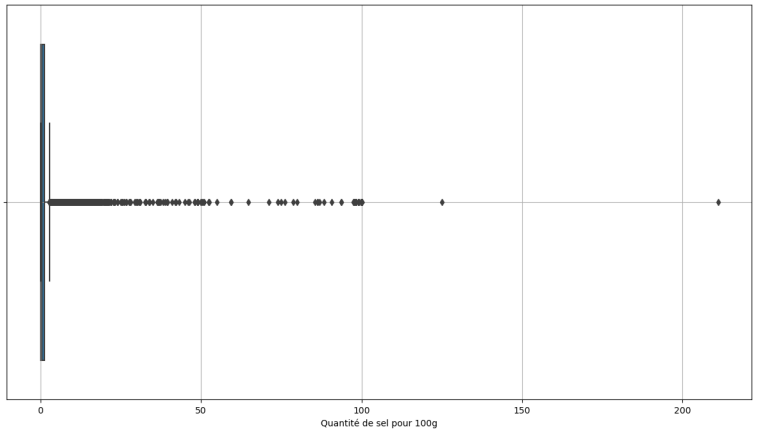
sucre, sel, huile de palme, énergie, graisses, graisses saturées, glucides, protéines, additifs, fibres, fruits-légumes-noix

Exemples :

Quantité de sucre pour 100g :



Quantité de sel pour 100g :

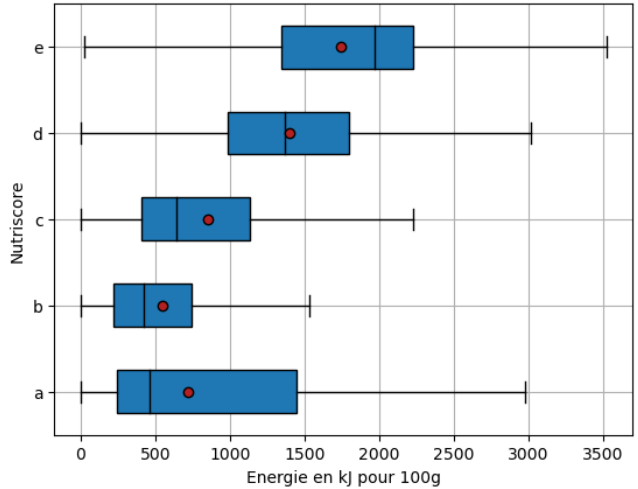


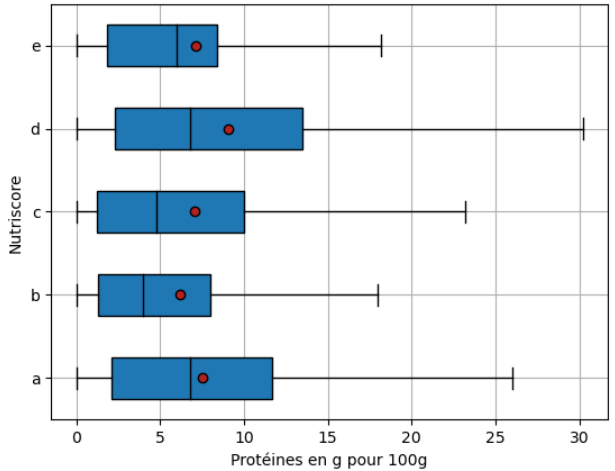
- On observe des dépassements des 100g : impossible, donc suppression
- Pas de valeurs aberrantes pour aucune colonne dans les valeurs extrêmes en-dessous de 100g

4.Lien entre les variables

Analyse de lien entre la variable Nutriscore et :
énergie, protéines, sel, sucre, graisses, graisses saturées, glucides, fibres, huile de palme, additifs

Exemples :





Il y a une certaine corrélation entre nutriscore et énergie mais pas entre nutriscore et protéines

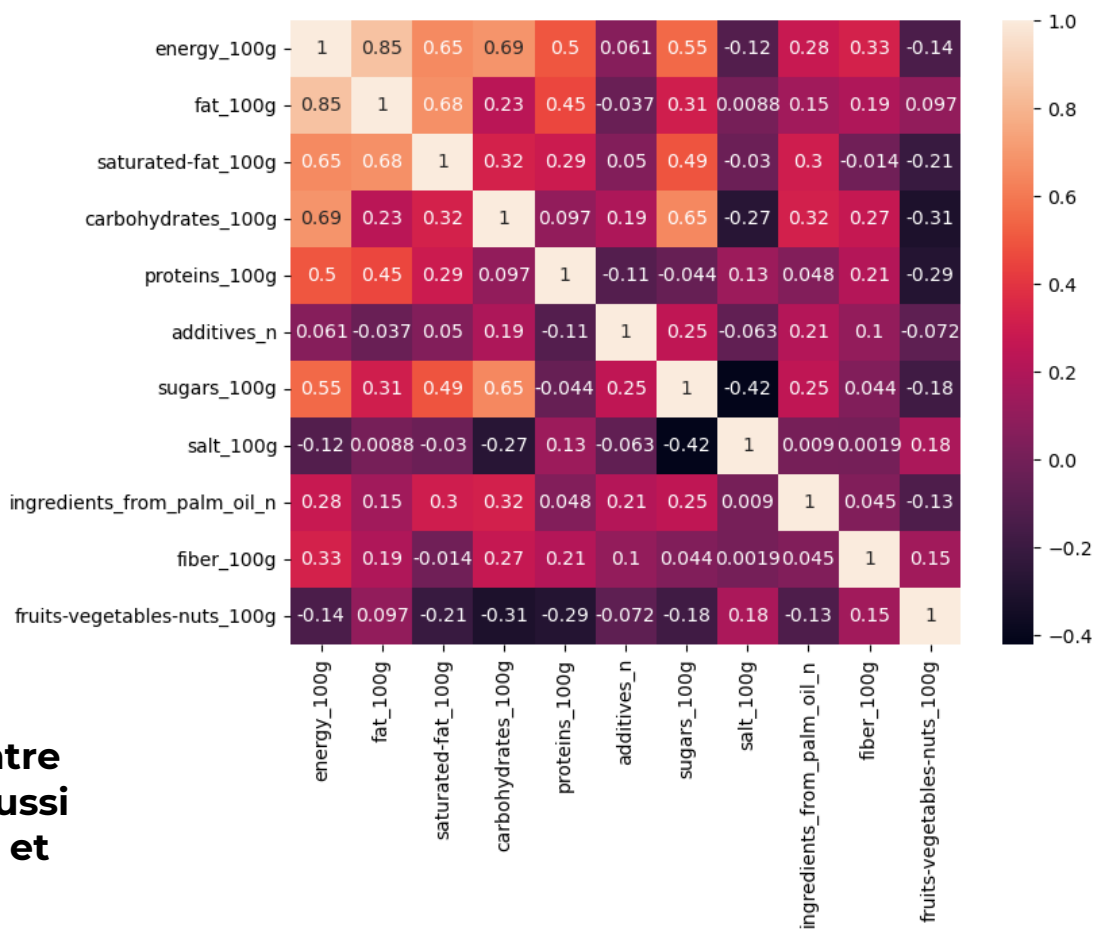
Bilan des analyses avant imputations :

	Variables	Nom des variables	Coefficient de corrélation avec le nutriscore, en %
0	energy_100g	Energie	29.0
1	proteins_100g	Protéines	2.0
2	salt_100g	Sel	1.0
3	sugars_100g	Sucre	18.0
4	fat_100g	Graisses	26.0
5	saturated-fat_100g	Graisses saturées	31.0
6	carbohydrates_100g	Glucides	5.0
7	fiber_100g	Fibres	4.0
8	ingredients_from_palm_oil_n	Huile de palme	6.0
9	additives_n	Nombre d'additifs	5.0

On observe donc une certaine corrélation entre nutriscore et Energie, Sucre, Graisses et Graisses saturées

Etude des corrélations avant imputation entre les variables numériques :

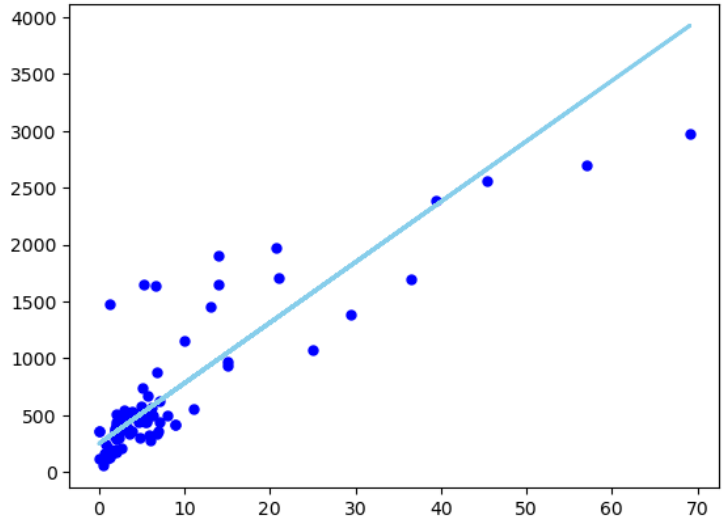
énergie, graisses, graisses saturées, glucides, protéines, additifs, sucre, sel, huile de palme, fibres, fruits-légumes-noix



On remarque une corrélation entre énergie et graisses surtout, et aussi avec glucides, graisses saturées et sucre.

Imputation des valeurs manquantes pour :
1. énergie
1.1. énergie avec graisses (régression linéaire) :

Energie en kJ

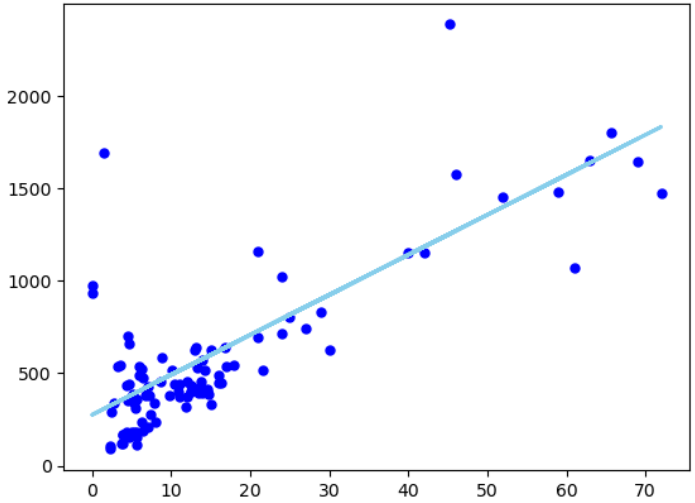


Graisses en g, pour 100g

On complète les données manquantes de énergie avec les données de graisses, et grâce à un test visualisé ci-contre, le pourcentage d'adéquation est ici de 73,66% pour 401 produits

Imputation des valeurs manquantes pour : 1.2 énergie avec glucides (régression linéaire) :

Energie en kJ



Glucides en g, pour 100g

On complète les données manquantes de énergie avec les données de glucides, et grâce à un test visualisé ci-contre, le pourcentage d'adéquation est ici de 61,98% pour 7 produits

Imputation des valeurs manquantes pour : 2. Huile de palme, sel, sucre, énergie, graisses, graisse saturées, glucides, protéines, fibres, additifs :

- ✓ Huile de palme :

a) Vérification contenu colonne huile de palme

b) **<u>Imputation de 0</u>** à la place des données manquantes huile de palme

- ✓ Sel, sucre, énergie (données manquantes restantes), graisses, graisses saturées, glucides, protéines, fibres, additifs :

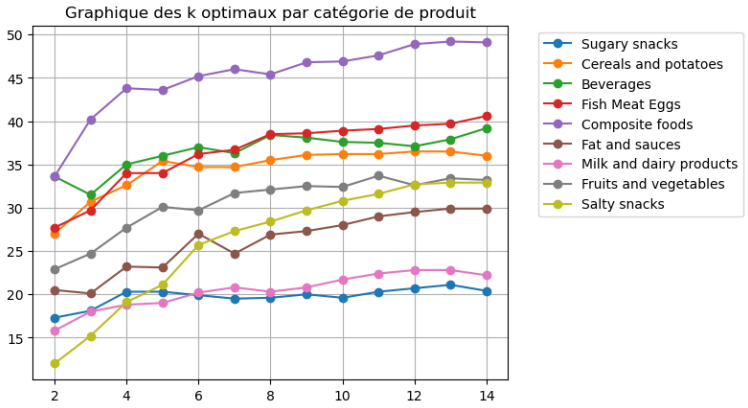
a) Observation de la médiane par catégorie de la quantité de sel par exemple :

b) **<u>Imputation de la médiane par catégorie</u>** de produit pour chaque variable

```
pnns_groups_1
Beverages                    0.01270
Cereals and potatoes         0.40132
Composite foods              0.93800
Fat and sauces               1.30000
Fish Meat Eggs               1.73000
Fruits and vegetables        0.21000
Milk and dairy products      0.15000
Salty snacks                 1.45000
Sugary snacks                0.24000
Name: salt_100g, dtype: float64
```

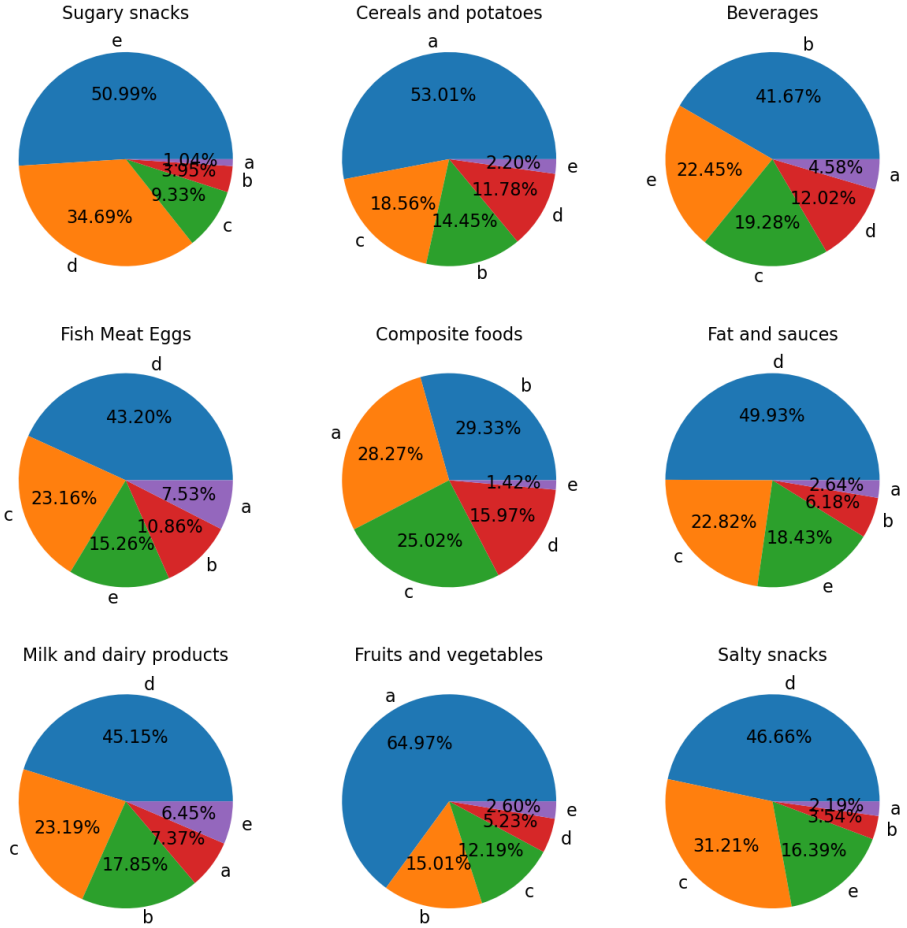
Imputation des valeurs manquantes pour : 3. Nutriscore (algorithme des plus proches voisins)

- Se fait à partir des variables les plus corrélées : énergie, graisses, graisses saturées, sucre
- Calcul du pourcentage d'erreur d'imputation en fonction du nombre de fois optimal de lancer de l'algorithme :



	Catégorie	k optimal	Pourcentage d'erreur
0	Sugary snacks	2	17.3
1	Cereals and potatoes	2	27.0
2	Beverages	3	31.5
3	Fish Meat Eggs	2	27.7
4	Composite foods	2	33.6
5	Fat and sauces	3	20.1
6	Milk and dairy products	2	15.8
7	Fruits and vegetables	2	22.9
8	Salty snacks	2	12.0

Visualisation de la répartition du nutriscore par catégorie de produits :



Bilan des analyses après imputations et comparaison avec les premières analyses :

	Variables	Nom des variables	Coefficient de corrélation avec le nutriscore, en %
0	energy_100g	Energie	29.0
1	proteins_100g	Protéines	2.0
2	salt_100g	Sel	1.0
3	sugars_100g	Sucre	18.0
4	fat_100g	Graisses	26.0
5	saturated-fat_100g	Graisses saturées	31.0
6	carbohydrates_100g	Glucides	5.0
7	fiber_100g	Fibres	4.0
8	ingredients_from_palm_oil_n	Huile de palme	6.0
9	additives_n	Nombre d'additifs	5.0

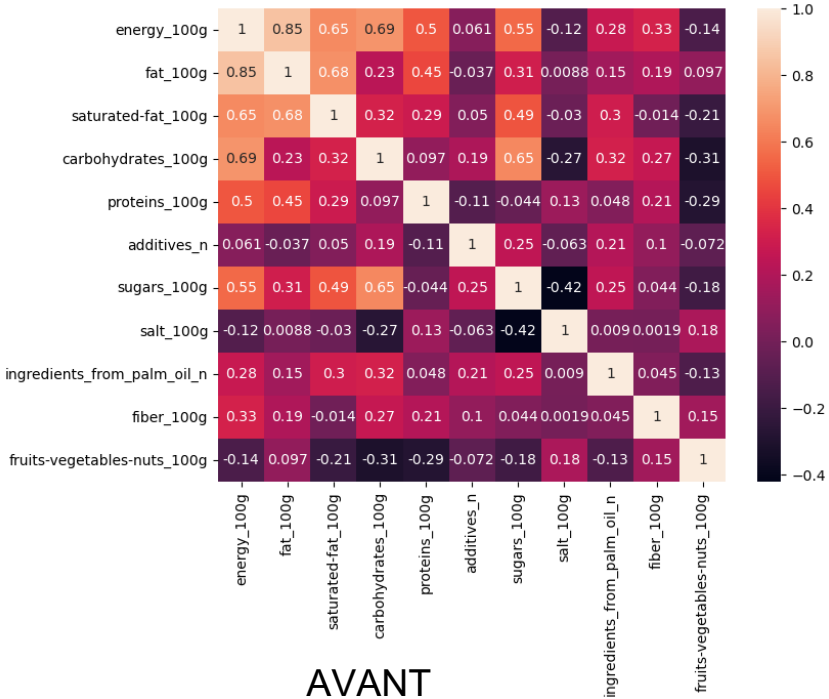
AVANT

	Variables	Nom des variables	Coefficient de corrélation avec le nutriscore, en %
0	energy_100g	Energie	32.0
1	proteins_100g	Protéines	5.0
2	salt_100g	Sel	1.0
3	sugars_100g	Sucre	20.0
4	fat_100g	Graisses	26.0
5	saturated-fat_100g	Graisses saturées	31.0
6	carbohydrates_100g	Glucides	7.0
7	fiber_100g	Fibres	5.0
8	ingredients_from_palm_oil_n	Huile de palme	4.0
9	additives_n	Nombre d'additifs	4.0

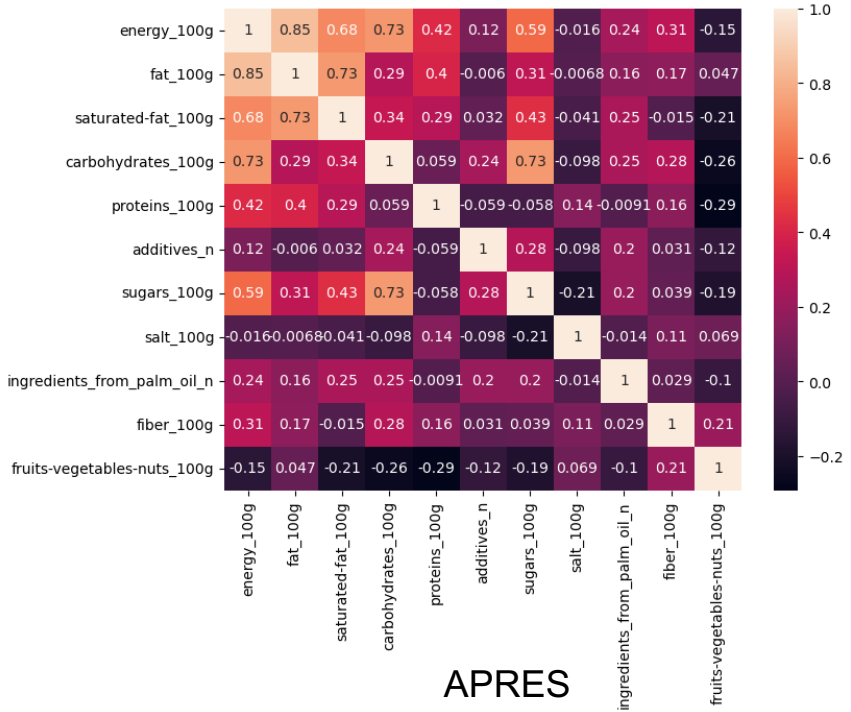
APRES

On observe donc à peu près les mêmes corrélations entre nutriscore et les autres variables, les imputations les ont peu modifiées.

Comparaison des corrélations après imputation entre les variables numériques :



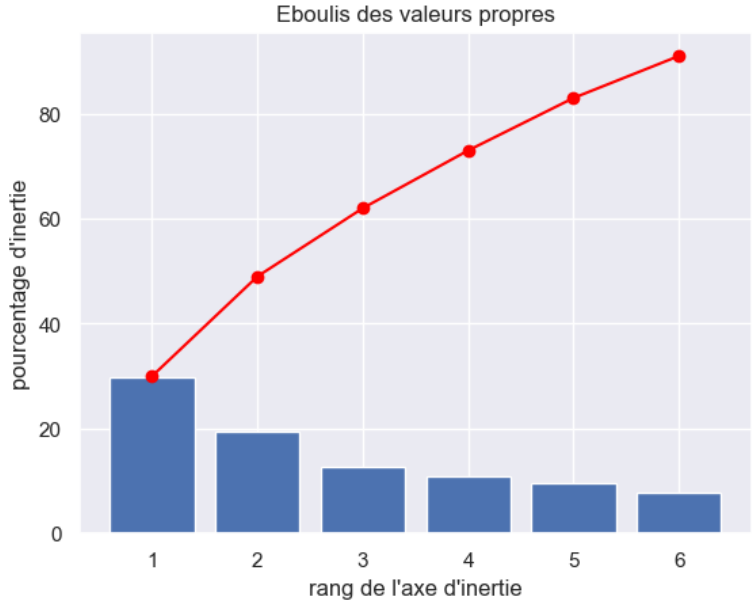
AVANT



APRES

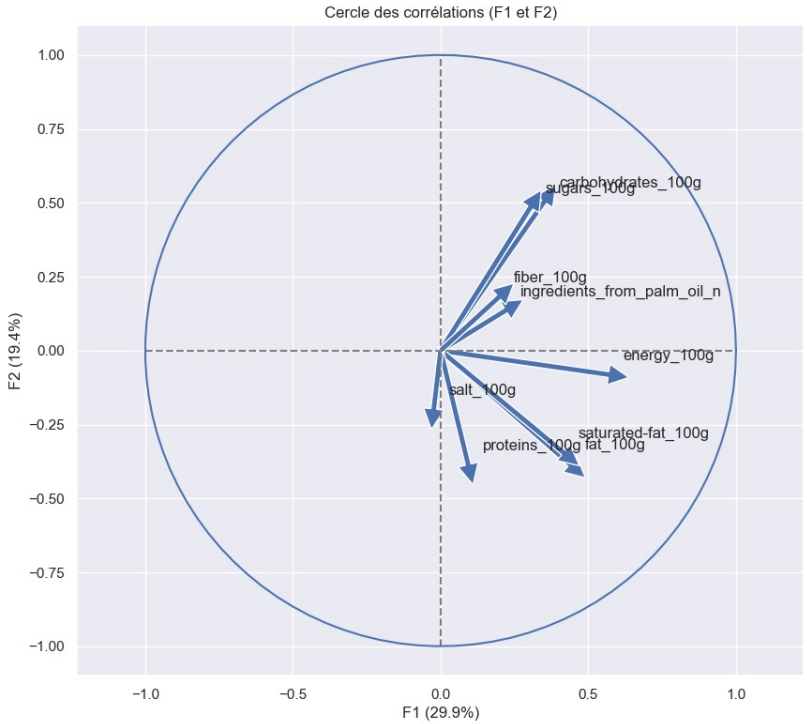
On remarque que les corrélations restent assez semblables.

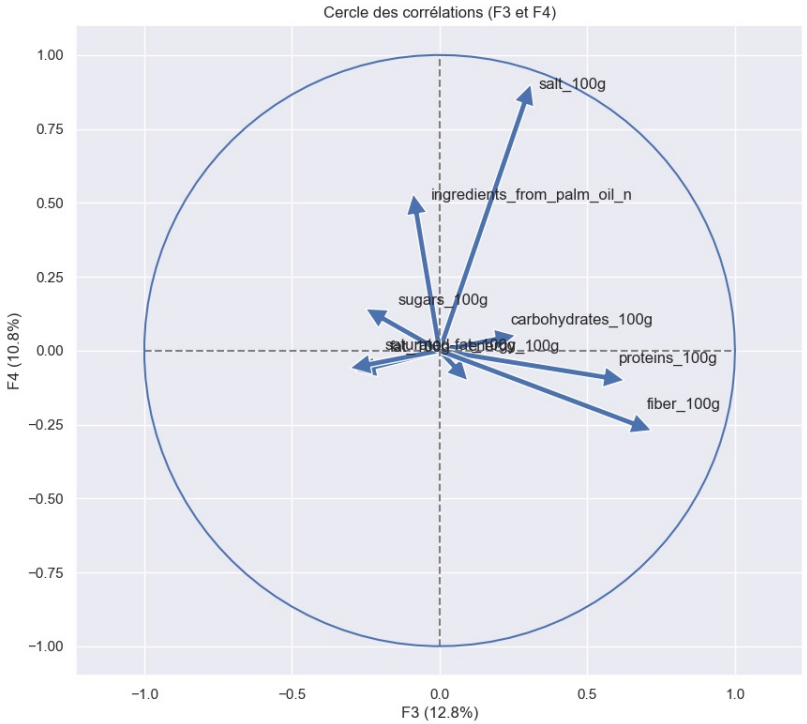
Autre visualisation des corrélations entre variables (ACP : Analyse en Composantes principales) :



On voit ici qu'on atteint 90% de la variance avec les 6 premières composantes et qu'on a plus de 60% avec les 3 premières

Cercles des corrélations entre variables :





On observe une certaine corrélation entre sucre et glucides, et entre graisses et graisses saturées

Projection des individus : Produits similaires

Visualisation de la possibilité de trouver des produits similaires selon les variables considérées pour notre application :
Ici selon F1 représentant principalement Energie, Graisses, Glucides
Et F2 représentant principalement Glucides

5.Pertinence et faisabilité de l’application

Limites engendrées par le dataset et les méthodes d'imputation choisies

- **Beaucoup de données manquantes**
- **Erreurs de remplissage**
- **Inadéquation des produits présents dans l'application (produits non alimentaires)**
- **Imputation des données par la méthode de régression linéaire (74% et 62% d'adéquation pour 408 produits seulement, et pas par catégorie car trop peu de données), celle des « plus proches voisins » (entre 12 et 33,6% d'erreur) , imputation de 0, ou imputation de la médiane par catégorie**
- **Corrélation peu prononcée entre nutriscore et les autres variables, mais nettement plus forte entre énergie et graisses, et voire glucides**
- **Les corrélations sont approximativement les mêmes avant et après imputations**

Exemple d'application possible

Calcul d'un score sur 100 (Plus les points sont élevés, et plus le produit est sain) qui caractérise le produit suivant certains critères :

Exemple :

- Lion Peanut x2 : 10/100
- lentilles vertes : 45/100
- Chair à saucisse : 15/100

Ensuite, l'application pourra trouver un produit de la même catégorie avec un meilleur score grâce au graphique des « projetés des individus »

Label bio	- Bio : 10 pts - Non bio : 0 pt
Additifs	- 0 additifs : 10 pts - Entre 1 et 3 additifs : 0 pt - Plus de 3 additifs : -5 pts
Sel	- Entre 0 et 0,3g : 10 pts - Entre 0,3 et 1,5g : 5 pts - Plus de 1,5g : -5 pts
Huile de palme	- Aucun ingrédients avec huile de palme : 5 pts - Sinon : -5 pts
Fibres	- Supérieur à 5g : 5 pts - Inférieur : 0 pt
Fruits-légumes-Noix	- Inférieur à 5g : 0 pt - Entre 5 et 20g : 5 pts - Entre 20 et 50g : 7 pts - Plus de 50g : 10 pts - Non connu : 0 pt
Nutriscore	- a : 50 pts - b : 40 pts - c : 30 pts - d : 20 pts - e : 10 pts

6.Conclusion

SOURCES

- Site Santé Publique France
- Site Open Food Facts
- Définition des variables à cette adresse : ici