Préparer les données d'OPEN FOOD FACTS

Santé publique France

Emmanuel Letremble

Objectif du projet

Santé publique France

OPEN FOOD FACTS

Rendre les données de santé plus accessibles et exploitable par les agents Santé publique France.

Vérifier des hypothèses à l'aide d'analyses univariées, multivariées et de tests statistiques appropriés.

Hypothèses

Santé publique France

OPEN FOOD FACTS

- Les nutriscores et nutrigrades sont fortement liés entre eux.
- Les nutriscores / nutrigrades sont dépendants des indices nutritionnels et peuvent être prédit par modèle probabiliste utilisant ces colonnes.
- Les nutriscores / nutrigrades sont liés au type de produit.

Présentation du jeu de données

Santé publique France

OPEN FOOD FACTS

Le jeu de données utilisé est une liste de 2.251.894 produits alimentaires répertoriés par les volontaires de l'association Open Food Facts.

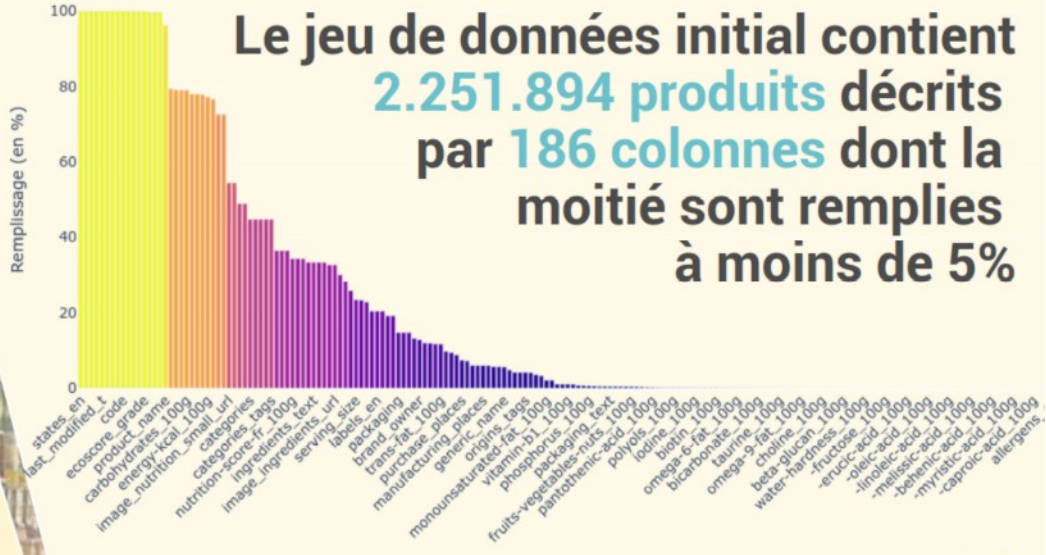
Chacun des produits est décrit par 4 types d'informations:

- **Des informations générales**
- **Des informations nutritionnelles**
- **Les ingrédients et leurs additifs**
- **Un ensemble de tags**

Présentation du jeu de données

Santé publique France

OPEN FOOD FACTS



Le jeu de données initial contient 2.251.894 produits décrits par 186 colonnes dont la moitié sont remplies à moins de 5%

Nettoyage des colonnes

Santé publique France

OPEN FOOD FACTS

Pour alléger le jeu de données pesant initialement 5.9Go, nous avons supprimé:

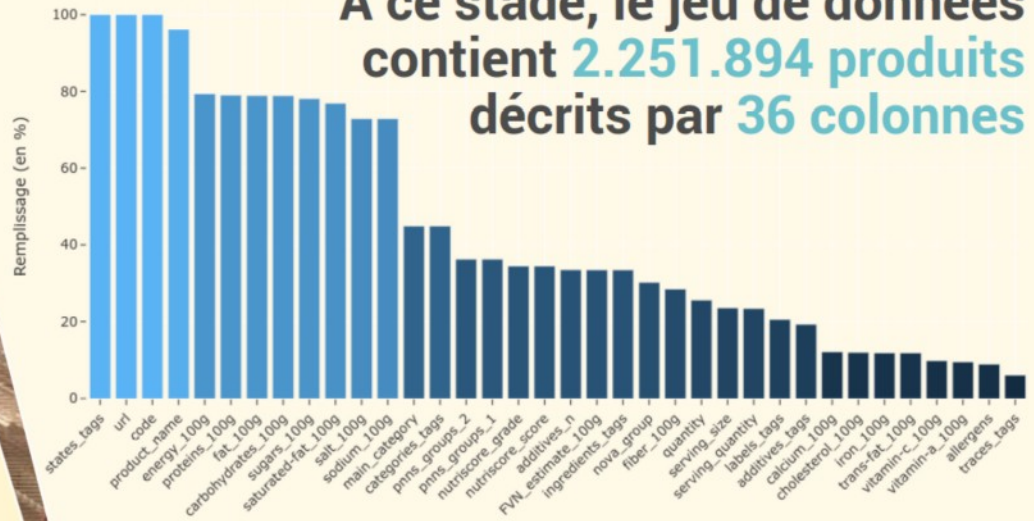
- **37 colonnes vides**
- **34 colonnes redondantes**
- **23 colonnes hors objectif**
- **56 colonnes ayant moins de 5% de lignes remplies**

Nettoyage des colonnes

Santé publique France

OPEN FOOD FACTS

A ce stade, le jeu de données contient 2.251.894 produits décrits par 36 colonnes



Nettoyage des colonnes

Santé publique France

OPEN FOOD FACTS

Parmi ces colonnes, nous avons gardé celles utiles pour étudier nos hypothèses:

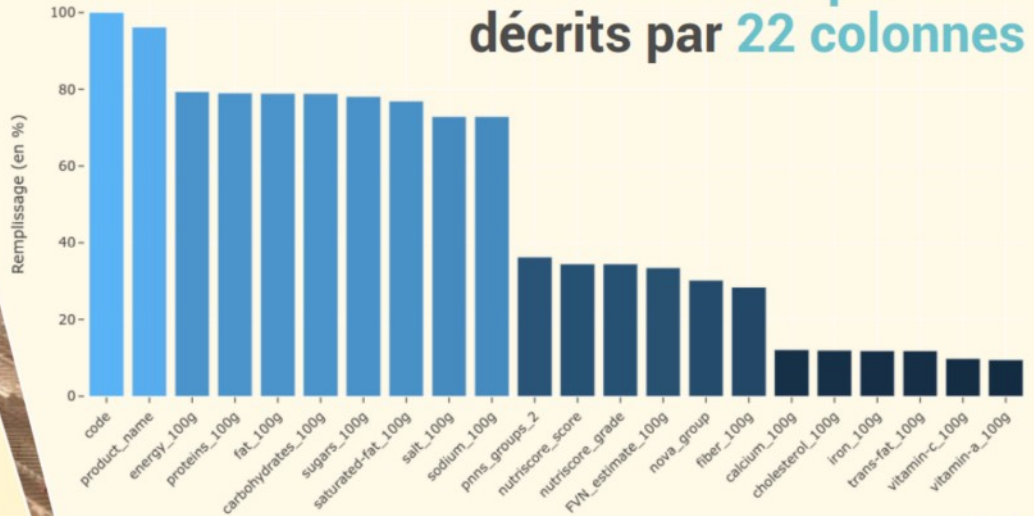
- tous les indicateurs nutritionnels _100g
- le code barre
- le product_name
- la catégorisation pnns_groups_2
- le nutriscore_score
- le nutriscore_grade
- le nova_group (par curiosité)

Nettoyage des colonnes

Santé publique France

OPEN FOOD FACTS

A ce stade, le jeu de données contient 2.251.894 produits décrits par 22 colonnes



Nettoyage des lignes

Après avoir réduit les colonnes de 186 à 22 nous avons commencé à réduire les lignes en supprimant:

- **354.503 lignes vides de tout indicateurs nutritionnelles**
- **89.630 lignes en doublons**

Nettoyage des lignes

Puis les valeurs aberrantes ont été traitées en deux temps :

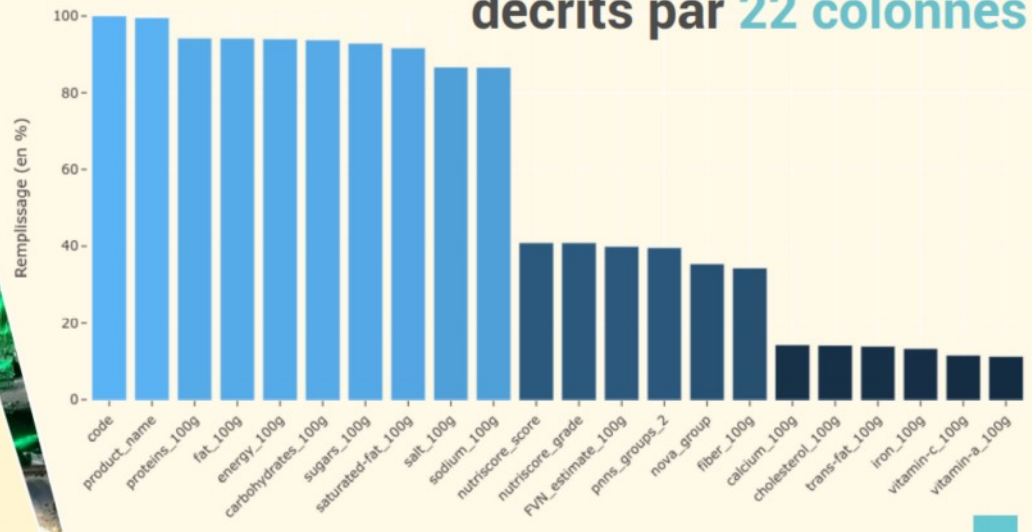
- On a d'abord supprimé 3.434 lignes dont les valeurs étaient en dehors de l'intervalle logique [0, 100] (pour 100g)
- On a ensuite mis en None les 41.963 lignes dépassant les valeurs métiers

Nettoyage des lignes

Santé publique France

OPEN FOOD FACTS

A ce stade, le jeu de données contient 1.804.326 produits décrits par 22 colonnes



Nettoyage des lignes

Santé publique France

OPEN FOOD FACTS

Puis après quelques analyses multivariées préliminaires, nous avons imputé les indicateurs nutritionnels en trois temps :

- **Pour l'évaluation, on sélectionne pour chaque colonne une fraction des non nulles, que l'on remplace par des None**
- **On fait une imputation en utilisant les moyennes pour servir de baseline**
- **On utilise un IterativeImputer que l'on compare à la baseline**

Nettoyage des lignes

Santé publique France

OPEN FOOD FACTS

Imputation par les moyennes (baseline)

```
FVN_estimate_100g  ==>  R²:   -0.00 | RMSE:   22.82
      calcium_100g  ==>  R²:   -0.00 | RMSE:    0.19
carbohydrates_100g  ==>  R²:   -0.00 | RMSE:   27.56
  cholesterol_100g  ==>  R²:   -0.00 | RMSE:    0.06
       energy_100g  ==>  R²:   -0.00 | RMSE:  781.66
          fat_100g  ==>  R²:   -0.00 | RMSE:   17.22
        fiber_100g  ==>  R²:   -0.00 | RMSE:    4.68
         iron_100g  ==>  R²:   -0.00 | RMSE:    0.00
     proteins_100g  ==>  R²:   -0.00 | RMSE:    9.92
         salt_100g  ==>  R²:   -0.00 | RMSE:    5.17
saturated-fat_100g  ==>  R²:   -0.00 | RMSE:    7.54
       sodium_100g  ==>  R²:   -0.00 | RMSE:    1.58
       sugars_100g  ==>  R²:   -0.00 | RMSE:   19.17
    trans-fat_100g  ==>  R²:   -0.00 | RMSE:    0.48
    vitamin-a_100g  ==>  R²:   -0.00 | RMSE:    0.00
    vitamin-c_100g  ==>  R²:   -0.00 | RMSE:    0.02
-----------------------------------------------------
          MEAN  ==>  R²:   -0.00 | RMSE:   56.13
```

Nettoyage des lignes

Santé publique France

OPEN FOOD FACTS

Imputation par IterativeImputer

```
FVN_estimate_100g  ==>  R²:    0.07 | RMSE:   21.98
      calcium_100g  ==>  R²:    0.25 | RMSE:    0.16
carbohydrates_100g  ==>  R²:    0.74 | RMSE:   14.01
  cholesterol_100g  ==>  R²:    0.14 | RMSE:    0.06
       energy_100g  ==>  R²:    0.84 | RMSE:  311.33
          fat_100g  ==>  R²:    0.78 | RMSE:    8.14
        fiber_100g  ==>  R²:    0.14 | RMSE:    4.34
         iron_100g  ==>  R²:    0.39 | RMSE:    0.00
     proteins_100g  ==>  R²:    0.28 | RMSE:    8.41
         salt_100g  ==>  R²:    0.75 | RMSE:    2.60
saturated-fat_100g  ==>  R²:    0.42 | RMSE:    5.72
       sodium_100g  ==>  R²:    1.00 | RMSE:    0.07
       sugars_100g  ==>  R²:    0.45 | RMSE:   14.22
    trans-fat_100g  ==>  R²:    0.01 | RMSE:    0.48
    vitamin-a_100g  ==>  R²:    0.01 | RMSE:    0.00
    vitamin-c_100g  ==>  R²:    0.04 | RMSE:    0.02
-------------------------------------------------------
         MEAN  ==>  R²:    0.39 | RMSE:   24.47
```

Nettoyage des lignes

Santé publique France

OPEN FOOD FACTS

Comparaison des scores d'imputation

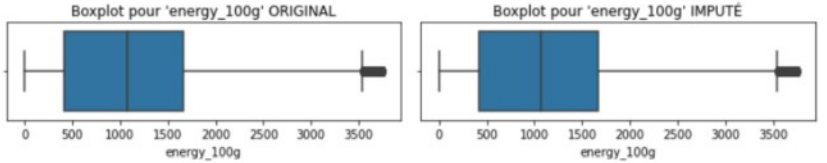
```
FVN_estimate_100g | RMSE diff=  -0.842 | Best method: IterativeImputer
      calcium_100g | RMSE diff=  -0.025 | Best method: IterativeImputer
carbohydrates_100g | RMSE diff= -13.545 | Best method: IterativeImputer
  cholesterol_100g | RMSE diff=  -0.004 | Best method: IterativeImputer
       energy_100g | RMSE diff=-470.327 | Best method: IterativeImputer
          fat_100g | RMSE diff=  -9.076 | Best method: IterativeImputer
        fiber_100g | RMSE diff=  -0.339 | Best method: IterativeImputer
         iron_100g | RMSE diff=  -0.000 | Best method: IterativeImputer
     proteins_100g | RMSE diff=  -1.513 | Best method: IterativeImputer
         salt_100g | RMSE diff=  -2.570 | Best method: IterativeImputer
saturated-fat_100g | RMSE diff=  -1.816 | Best method: IterativeImputer
       sodium_100g | RMSE diff=  -1.512 | Best method: IterativeImputer
       sugars_100g | RMSE diff=  -4.944 | Best method: IterativeImputer
    trans-fat_100g | RMSE diff=  -0.003 | Best method: IterativeImputer
    vitamin-a_100g | RMSE diff=  -0.000 | Best method: IterativeImputer
    vitamin-c_100g | RMSE diff=  -0.000 | Best method: IterativeImputer
```

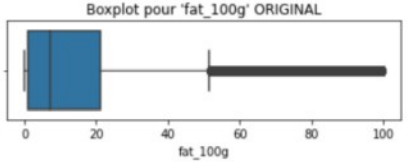
Nettoyage des lignes

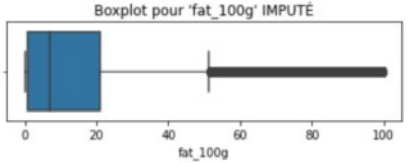
Santé publique France

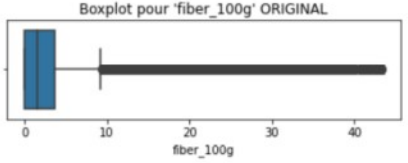
OPEN FOOD FACTS

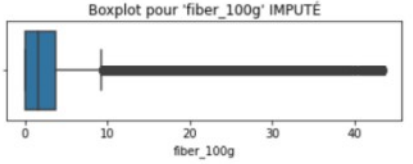
Comparaison des distribution avant / après

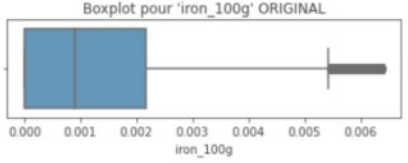


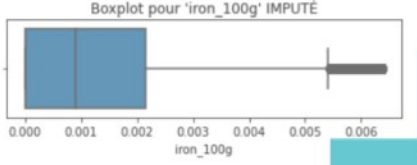










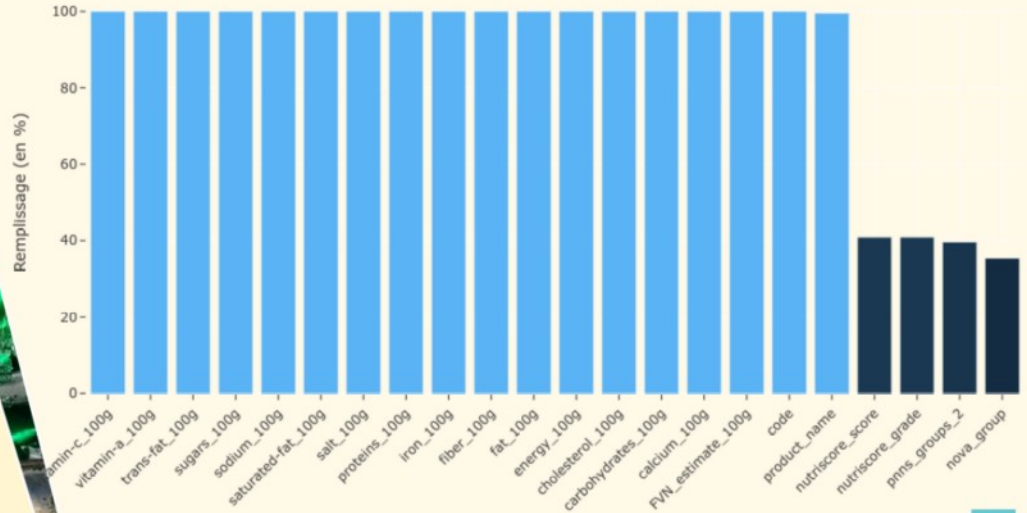


Nettoyage des lignes

Santé publique France

OPEN FOOD FACTS

A ce stade, le jeu de données contient 1.804.326 produits décrits par 22 colonnes



Nettoyage des lignes

Santé publique France

OPEN FOOD FACTS

Enfin, nous avons supprimé les lignes qui ne respectaient pas certains critères logiques après imputation:

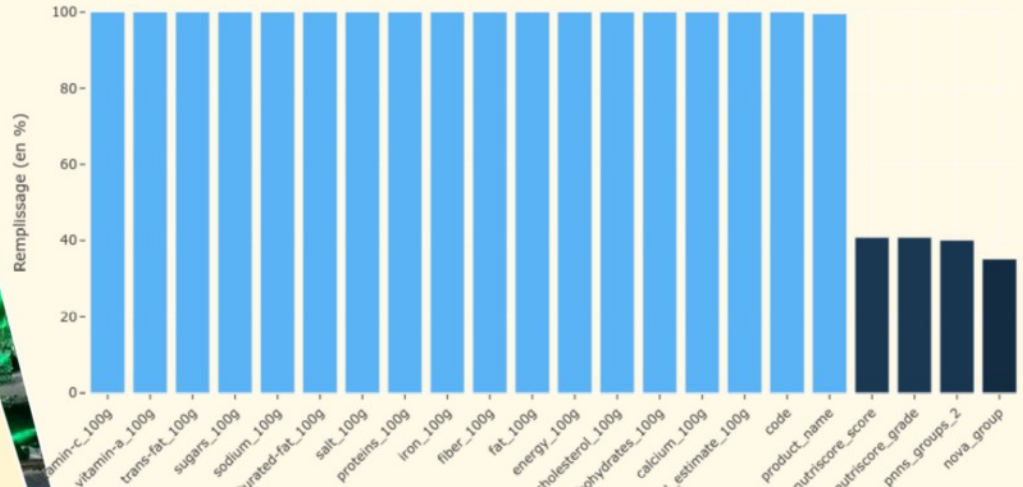
- **3284** lignes supprimées quand carbohydrates_100g > sugars_100g
- **4184** lignes supprimées quand saturated-fat_100g + trans-fat_100g + cholesterol_100g > fat_100g
- **85.571** lignes supprimées quand la sommes des principales catégories dépasse les 100g

Nettoyage des lignes

Santé publique France

OPEN FOOD FACTS

A ce stade, le jeu de données contient 1.7187.54 produits décrits par 22 colonnes

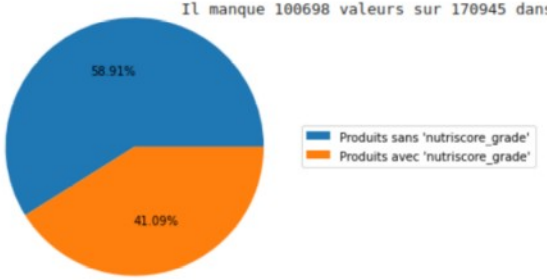


Analyses univariées

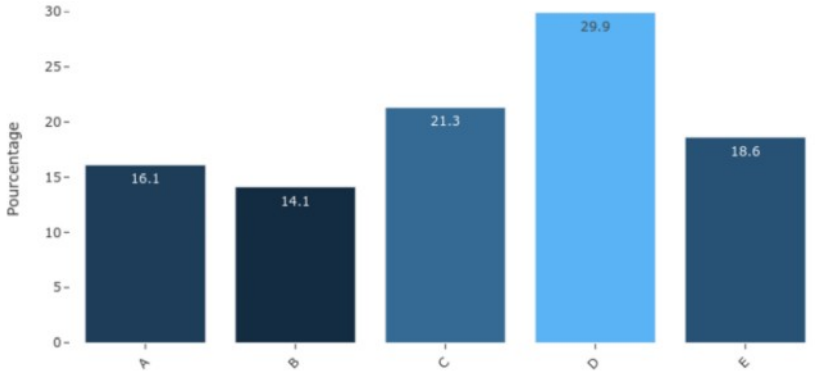
```
----- ANALYSE UNIVARIÉE de "nutriscore_grade" -----

count     699607
unique         5
top            D
freq      213331
Name: nutriscore_grade, dtype: object
```

Répartition entre produits avec ou sans 'nutriscore_grade'

```
Il manque 100698 valeurs sur 170945 dans la colonne 'nutriscore_grade' (58.91%)
```



Repartition des Nutri-grades (sur les valeurs non-nulles)

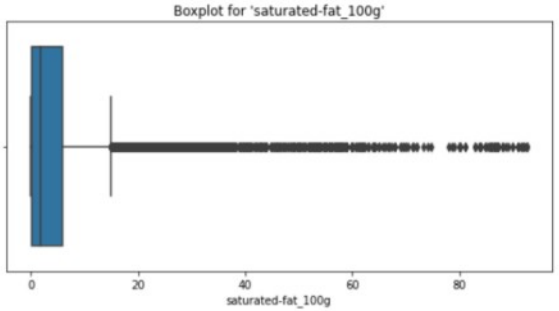


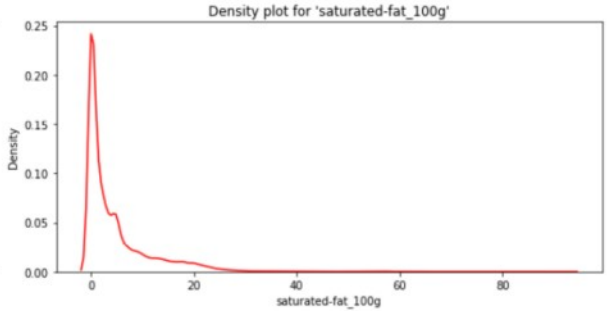
Analyses univariées

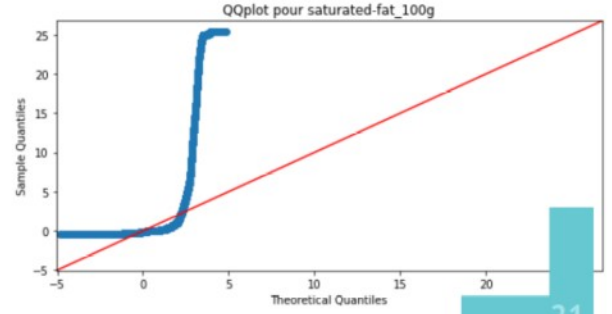
Santé publique France

OPEN FOOD FACTS

```
----- ANALYSE UNIVARIÉE de "saturated-fat_100g" -----

count   1709454.000
mean          4.661
std           6.993
min           0.000
25%           0.100
50%           1.900
75%           6.000
max          92.600
Name: saturated-fat_100g, dtype: float64
```





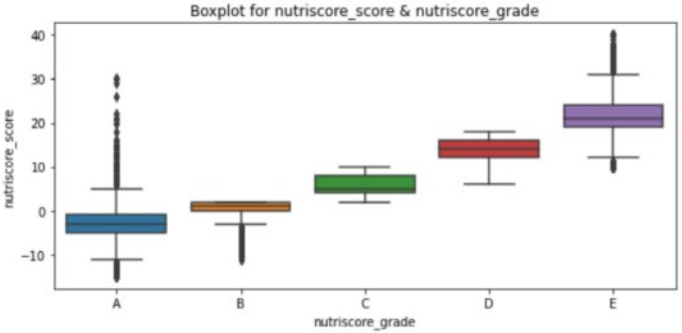


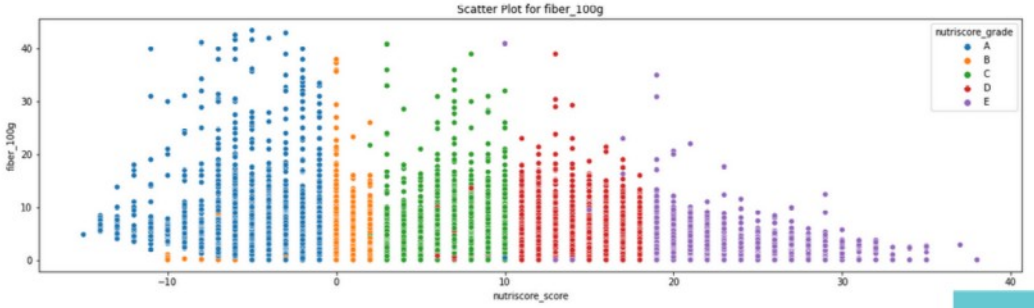
Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Relation entre nutrigrades et nutriscores



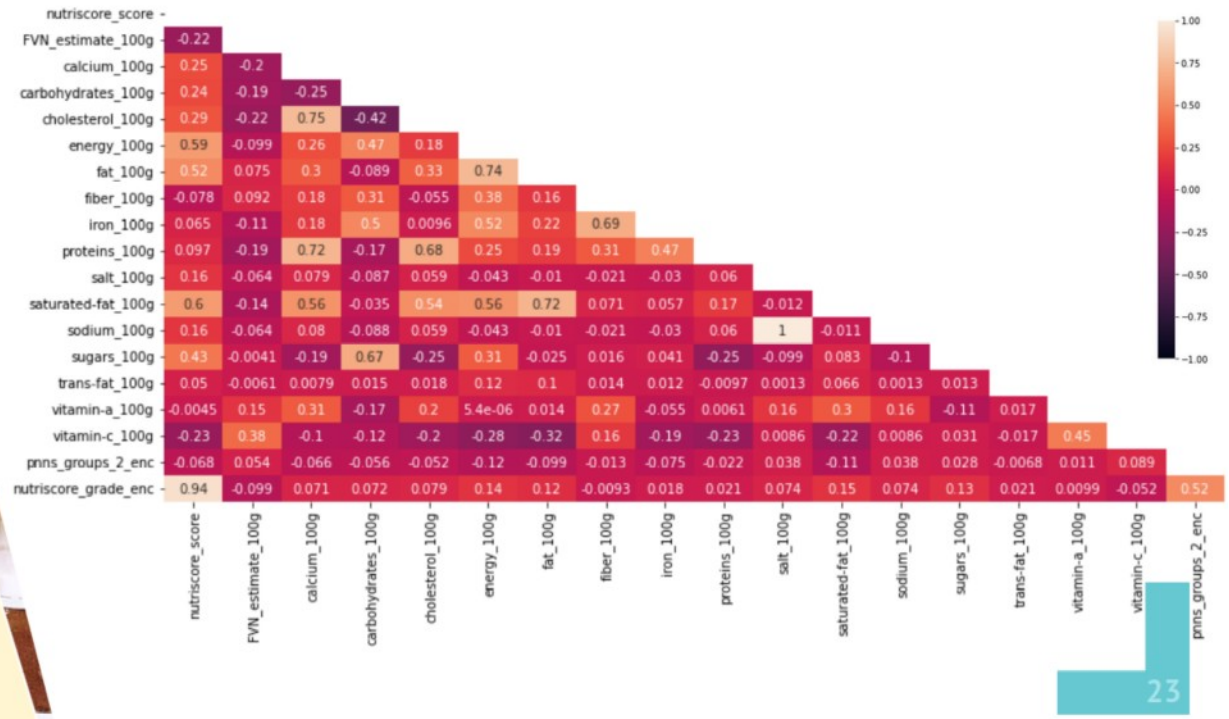


Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Corrélations de Pearson (avec nominales)

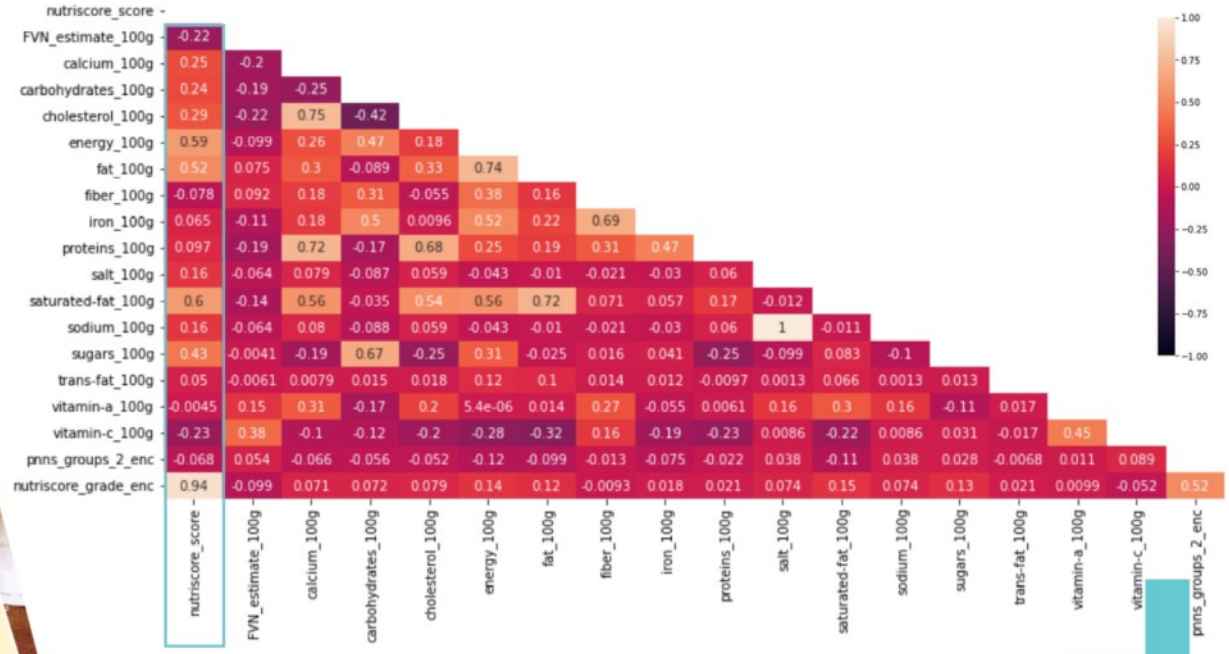


Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Corrélations de Pearson (avec nominales)



Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Corrélations de Pearson (avec nominales)

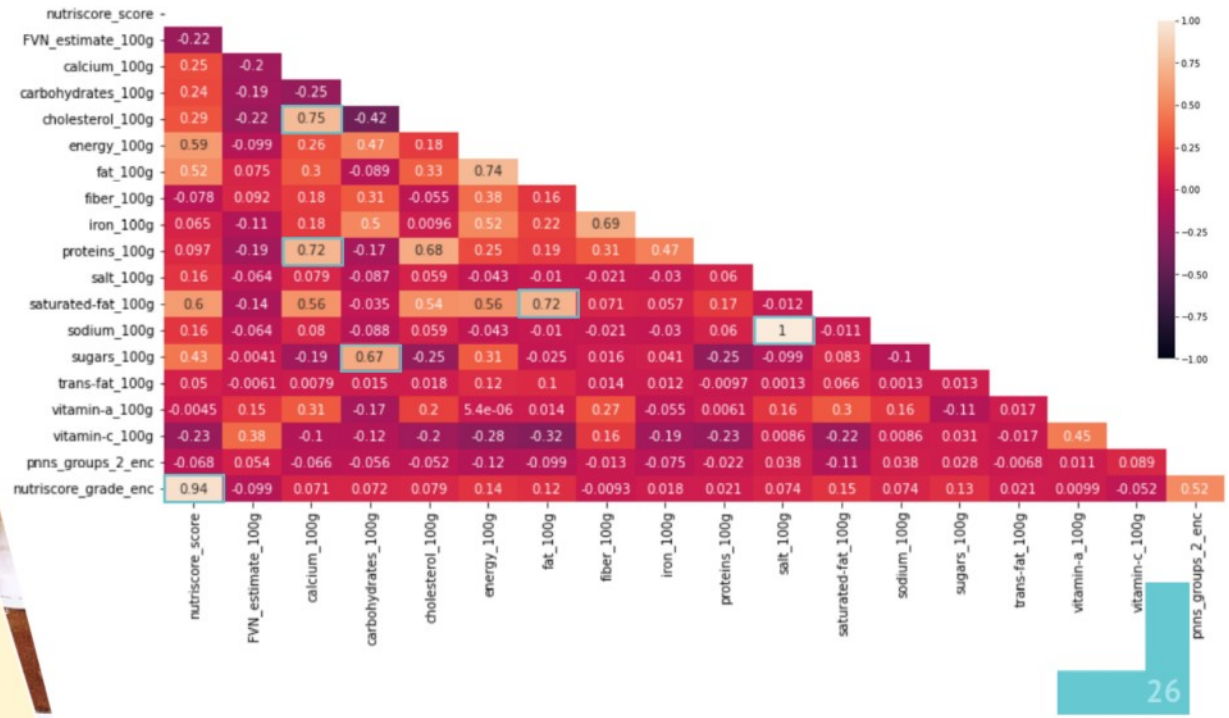
	nutriscore_score	FVN_estimate_100g	calcium_100g	carbohydrates_100g	cholesterol_100g	energy_100g	fat_100g	fiber_100g	iron_100g	proteins_100g	salt_100g	saturated-fat_100g	sodium_100g	sugars_100g	trans-fat_100g	vitamin-a_100g	vitamin-c_100g	pnns_groups_2_enc
nutriscore_score																		
FVN_estimate_100g	-0.22																	
calcium_100g	0.25	-0.2																
carbohydrates_100g	0.24	-0.19	-0.25															
cholesterol_100g	0.29	-0.22	0.75	-0.42														
energy_100g	0.59	-0.099	0.26	0.47	0.18													
fat_100g	0.52	0.075	0.3	-0.089	0.33	0.74												
fiber_100g	-0.078	0.092	0.18	0.31	-0.055	0.38	0.16											
iron_100g	0.065	-0.11	0.18	0.5	0.0096	0.52	0.22	0.69										
proteins_100g	0.097	-0.19	0.72	-0.17	0.68	0.25	0.19	0.31	0.47									
salt_100g	0.16	-0.064	0.079	-0.087	0.059	-0.043	-0.01	-0.021	-0.03	0.06								
saturated-fat_100g	0.6	-0.14	0.56	-0.035	0.54	0.56	0.72	0.071	0.057	0.17	-0.012							
sodium_100g	0.16	-0.064	0.08	-0.088	0.059	-0.043	-0.01	-0.021	-0.03	0.06	1	-0.011						
sugars_100g	0.43	-0.0041	-0.19	0.67	-0.25	0.31	-0.025	0.016	0.041	-0.25	-0.099	0.083	-0.1					
trans-fat_100g	0.05	-0.0061	0.0079	0.015	0.018	0.12	0.1	0.014	0.012	-0.0097	0.0013	0.066	0.0013	0.013				
vitamin-a_100g	-0.0045	0.15	0.31	-0.17	0.2	5.4e-06	0.014	0.27	-0.055	0.0061	0.16	0.3	0.16	-0.11	0.017			
vitamin-c_100g	-0.23	0.38	-0.1	-0.12	-0.2	-0.28	-0.32	0.16	-0.19	-0.23	0.0086	-0.22	0.0086	0.031	-0.017	0.45		
pnns_groups_2_enc	-0.068	0.054	-0.066	-0.056	-0.052	-0.12	-0.099	-0.013	-0.075	-0.022	0.038	-0.11	0.038	0.028	-0.0068	0.011	0.089	
nutriscore_grade_enc	0.94	-0.099	0.071	0.072	0.079	0.14	0.12	-0.0093	0.018	0.021	0.074	0.15	0.074	0.13	0.021	0.0099	-0.052	0.52

Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Corrélations de Pearson (avec nominales)

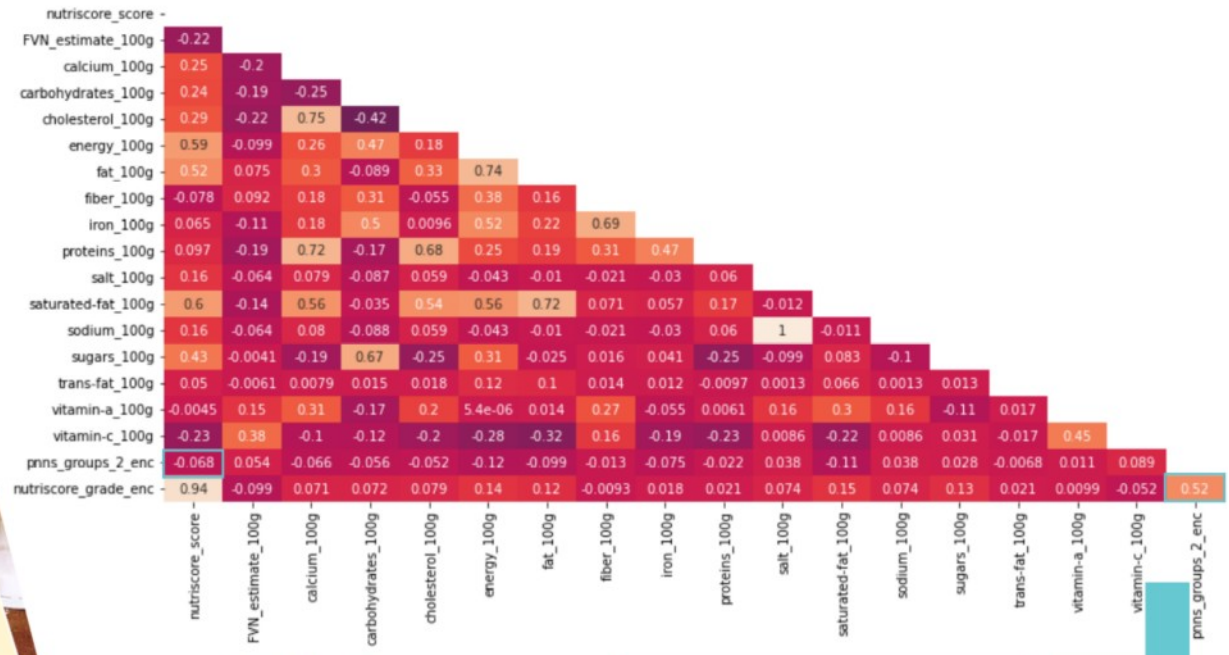


Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Corrélations de Pearson (avec nominales)



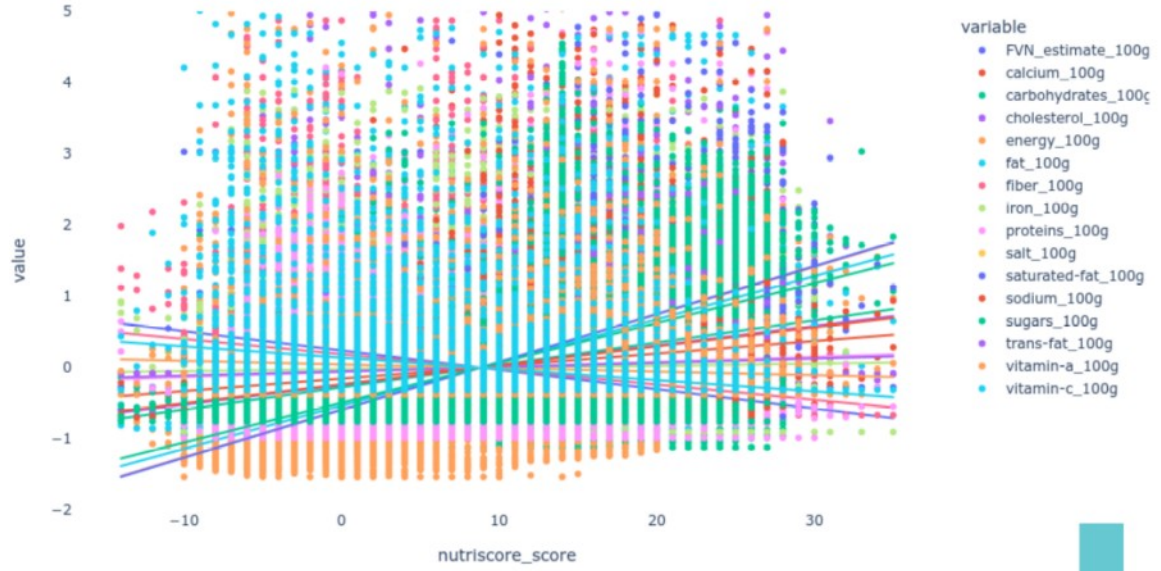
Test V de Cramer

Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Régressions lineaires

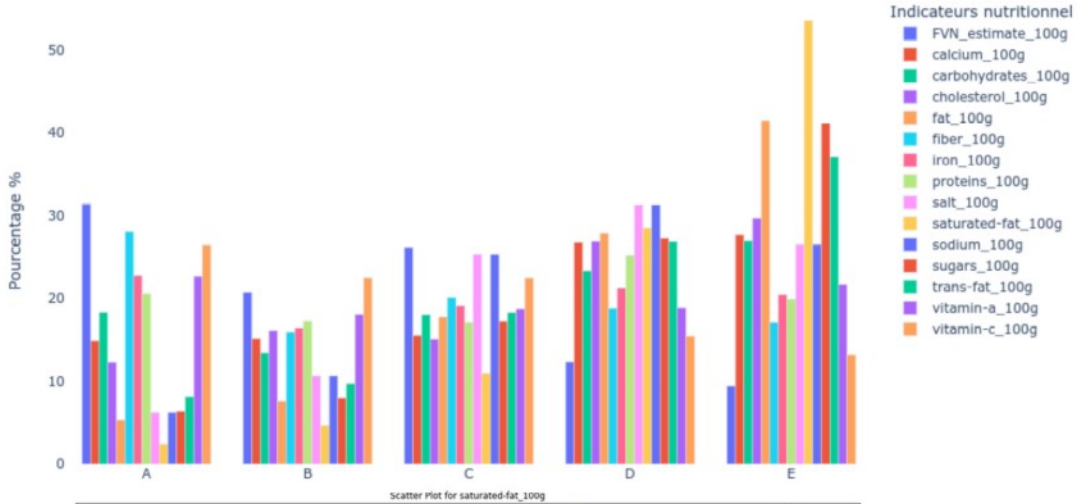


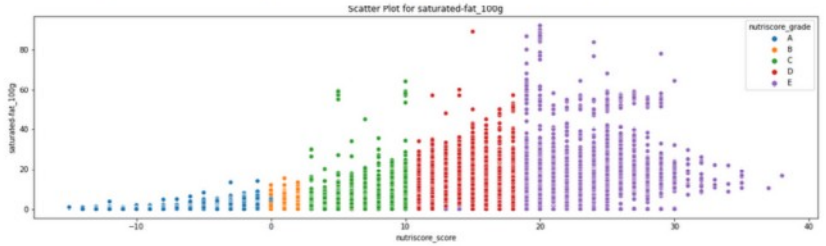
Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Répartition des indicateurs par nutrigrades



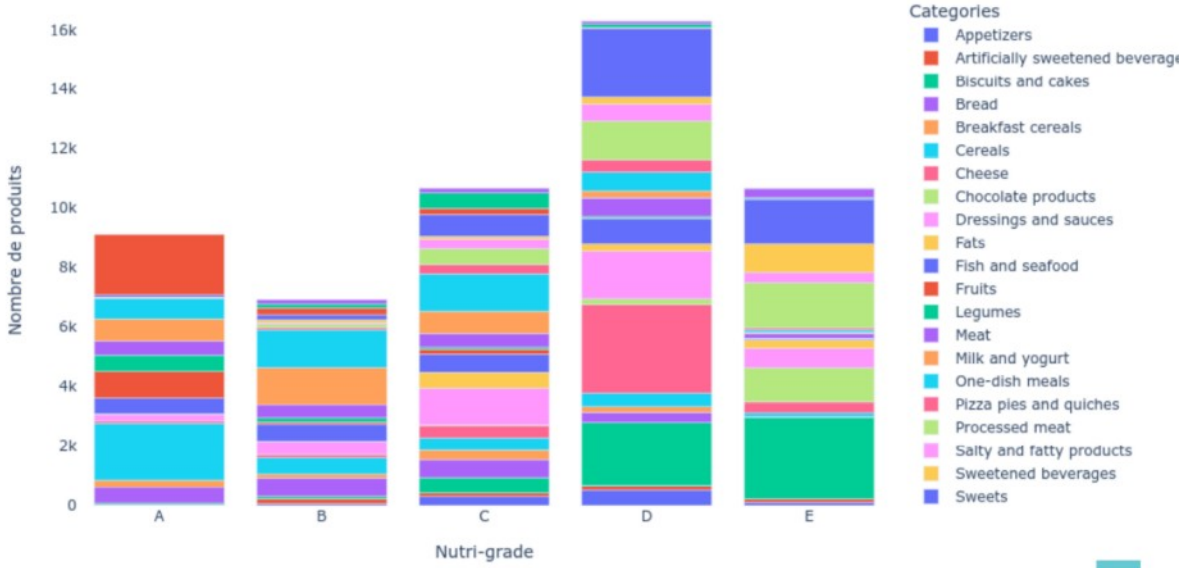


Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Répartition des catégories par nutrigrades

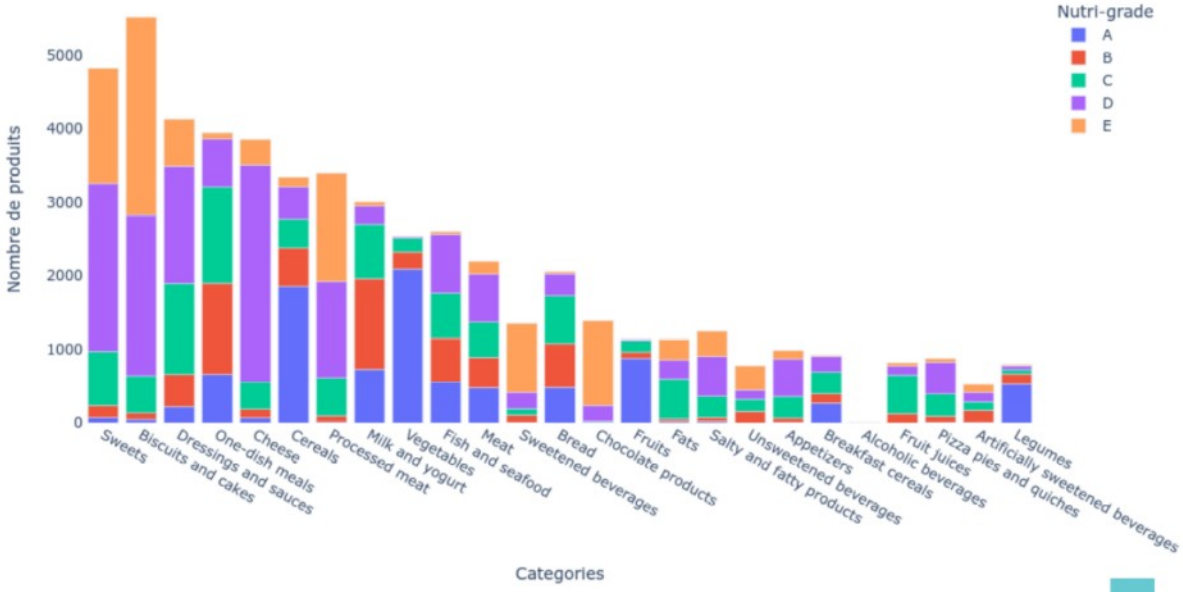


Analyses multivariées

Santé publique France

OPEN FOOD FACTS

Répartition des nutrigrades par catégories

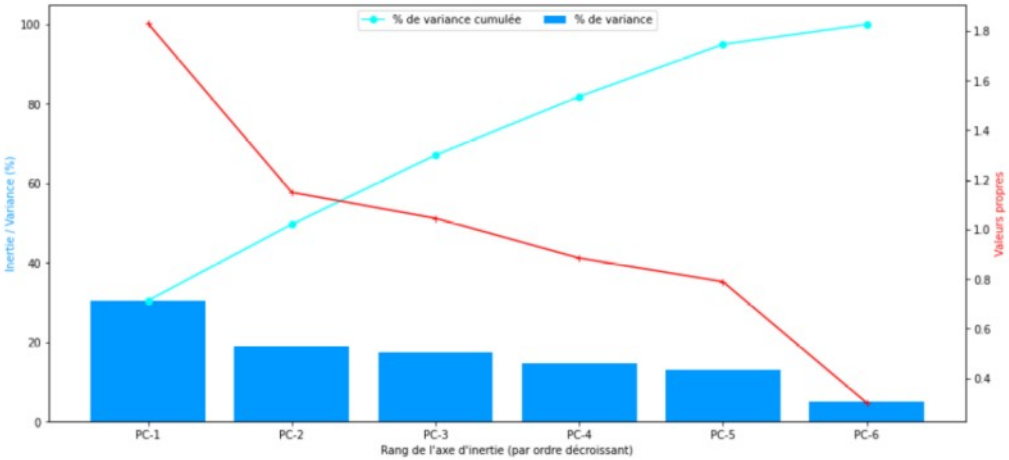


Analyse en composantes principales

Santé publique France

OPEN FOOD FACTS

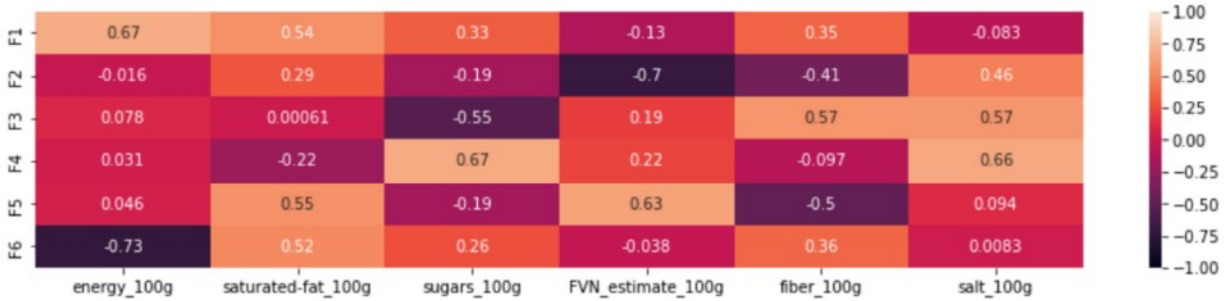
Éboulis des valeurs propres



	Valeurs propres	% des valeurs propres	% cumulé
1	1.831	0.305	0.305
2	1.150	0.192	0.497
3	1.045	0.174	0.671
4	0.884	0.147	0.818
5	0.789	0.132	0.950
6	0.301	0.050	1.000

Analyse en composantes principales

Corrélations avec les indicateurs nutritionnels



F1 pourrait être un axe basé sur l'**apport énergétique.**

F2 pourrait être un axe basé qui oppose les produits plutôt **gras** aux produits plutôt **secs**.

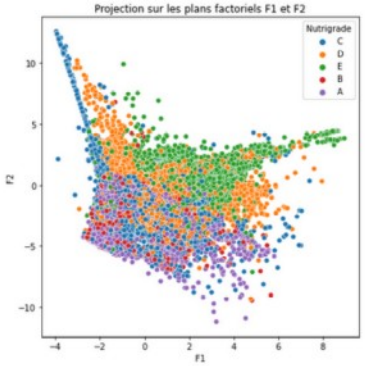
F3 pourrait être un axe qui oppose les produits **salés** aux produits **sucrés**.

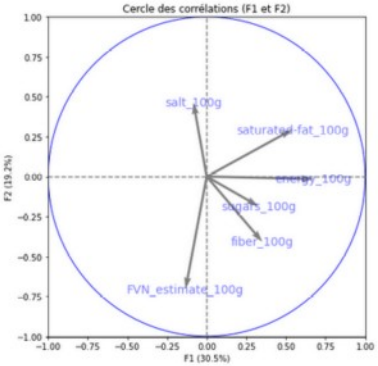
Analyse en composantes principales

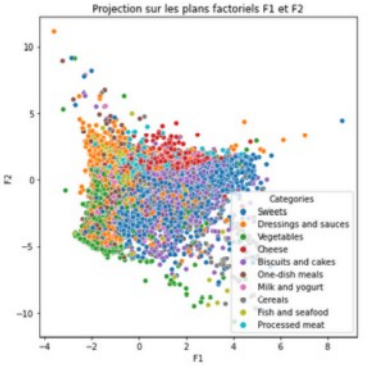
Santé publique France

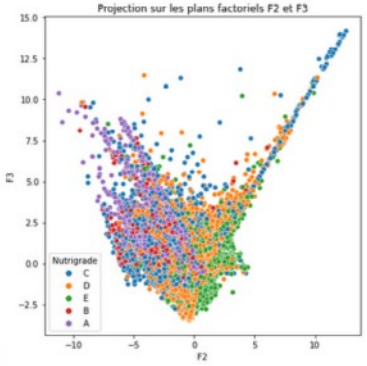
OPEN FOOD FACTS

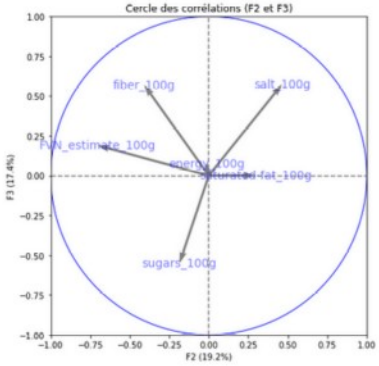
Visualisations sur les plans factoriels de l'ACP

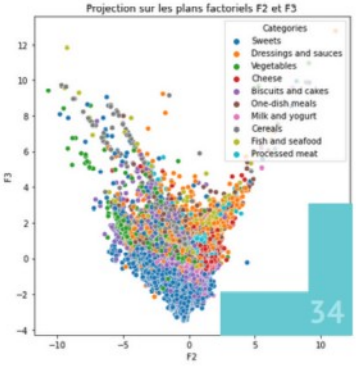












Analyse de la variance (ANOVA)

Santé publique France

OPEN FOOD FACTS

Pour être utilisable, l'ANOVA doit respecter plusieurs hypothèses fondamentales :

- **Indépendance des échantillons**
- **Normalité de la distribution**

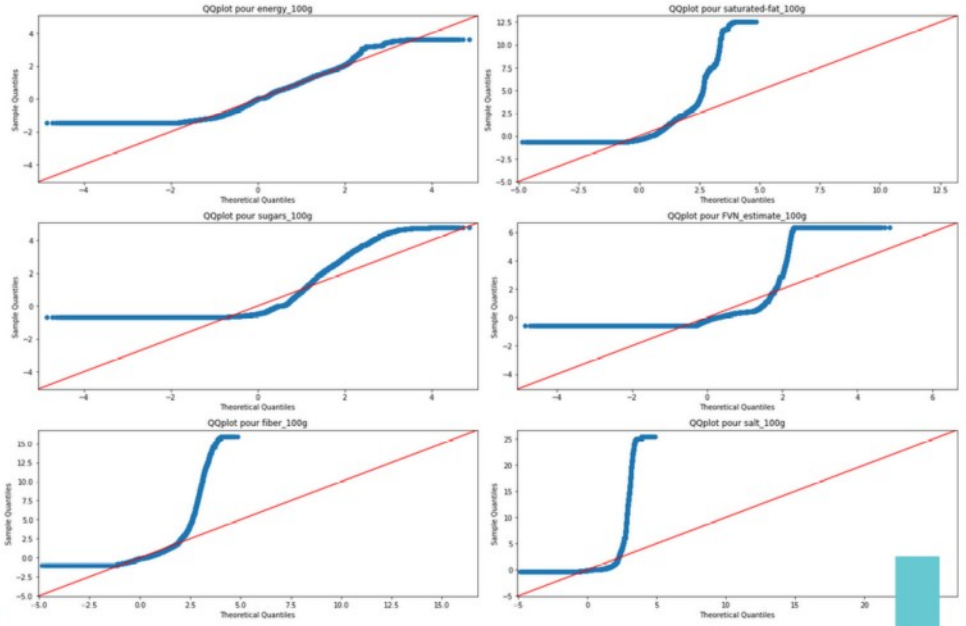
Mais aucune des variable étudiée lors des analyses univariées n'a une distribution gausienne.

Analyse de la variance (ANOVA)

Santé publique France

OPEN FOOD FACTS

QQ-plots - Diagrammes Quantile-Quantile



Kruskal-Wallis

Santé publique France

OPEN FOOD FACTS

Le test de Kruskal-Wallis est une alternative non paramétrique qui ne nécessite pas une distribution normale, ni des échantillons de taille similaires.

```
      energy_100g >> statistique = 202464.56 | p-valeur = 0.00 | H0 est rejetée
saturated-fat_100g >> statistique = 225428.64 | p-valeur = 0.00 | H0 est rejetée
       sugars_100g >> statistique =  73553.70 | p-valeur = 0.00 | H0 est rejetée
 FVN_estimate_100g >> statistique =  20101.16 | p-valeur = 0.00 | H0 est rejetée
       fiber_100g >> statistique =  15078.40 | p-valeur = 0.00 | H0 est rejetée
        salt_100g >> statistique =  83753.81 | p-valeur = 0.00 | H0 est rejetée
```

H0 : toutes les médianes de cet indicateur nutritionnel pour les grade A, B, C, D et E sont égales; on ne peut différencier les grades avec cet indicateur.

H1 : au moins une médiane de cet indicateur nutritionnel pour les grade A, B, C, D et E, est différente; on peut différencier les grades avec cet indicateur.

Synthèse de l'analyse de données

Santé publique France

OPEN FOOD FACTS

- Les analyses de données et le test statistiques indiquent qu'il est raisonnable de s'appuyer sur les colonnes nutritionnelles sélectionnées pour prédire le nutriscore_score ou le nutriscore_grade.

- L'ACP montrent que la dimensionnalité du jeu de données pourrait être légèrement réduite tout en conservant un pouvoir prédictif sensiblement équivalent.

- Sur la base de cette EDA, on peut prendre deux directions pour le modèle prédictif :
  - prédire le nutriscore_grade directement
  - prédire le nutriscore_score puis le convertir en nutriscore_grade.

Merci de m'avoir écouté, évalué et conseillé.

Santé publique France

OPEN FOOD FACTS