

# PROJET 3

---

## CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE



# PLAN

---

- 01 PRÉSENTATION DU JEU DE DONNÉES
- 02 PRÉSENTATION DE L'IDÉE D'APPLICATION
- 03 NETTOYAGE DES DONNÉES
- 04 ANALYSE UNIVARIÉE
- 05 ANALYSE MULTIVARIÉE
- 06 REDUCTION DE DIMENSION
- 07 FAITS PERTINENTS CONCLUSION ET PERSPECTIVES




# 01

## PRÉSENTATION DU JEU DE DONNÉES

---

# CONTEXTE

- Organisme public français sous la tutelle du ministère de la santé
-  MISSION : améliorer et protéger la santé des populations
- Appel à projet : Proposer une idée innovante d'application en lien avec l'alimentation



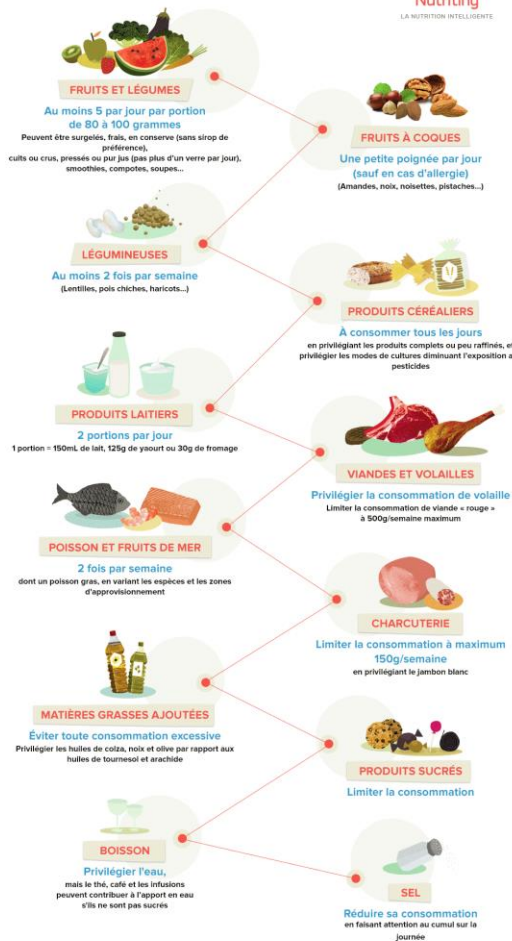
# SOURCE DU JEU DE DONNÉES

Base de données alimentaire mise à disposition par **Open Food Facts**



- ✓ 1M+ de produits alimentaires référencés
- ✓ 186 indicateurs
- ✓ couvrant 4 types d'informations:
  - Les **informations générales** sur la fiche du produit : nom, date de modification, etc.
  - Un **ensemble de tags** : catégorie du produit, localisation, origine, etc.
  - Les **ingrédients composant les produits et leurs additifs éventuels**.
  - Des **informations nutritionnelles** : quantité en grammes d'un nutriment pour 100 grammes du produit.

FICHER	# LIGNES	#COLONNES
Food Data	1877241	186



# NOVA



Utilisée par l'appli Open Food Facts, la recherche et les instances de santé.

**GROUPE 1** Aliments bruts ou peu transformés (rôtis, emballés sous-vide, broyés, torréfiés, fermentés...) : fruits, légumes, viandes, pâtes, lait...



**GROUPE 2** Ingrédients issus de matières brutes par pressage, raffinage, broyage, séchage : sel, huile, sucre... Certains minéraux et additifs admis.



**GROUPE 3** Aliments transformés fabriqués à partir de denrées des groupes 1 et 2 (conservation, cuisson, fermentation) : légumes/fruits en conserve, graines et noix salées, viandes fumées, poisson en conserve ou encore fromages et pain frais. Certains additifs (conservateurs, antioxydants) admis.



**GROUPE 4** Aliments ultratransformés : produits des précédents groupes + ingrédients industriels (huile hydrogénée, sirop de glucose...) et additifs « cosmétiques » (colorant, arôme, exhausteur de goût...) + procédés de fabrication industriels tels que chauffage à haute température, extrusion, cracking...



# GROUPES PNNS

Programme National Nutrition Santé

Plan de santé publique lancé en 2001 et actualisé tous les 5 ans

🎯 L'objectif : améliorer l'état de santé de la population en agissant sur l'un de ses déterminants majeurs : la nutrition

# GROUPES NOVA

Groupe 1 - Aliments non transformés ou transformés minimalement

Groupe 2 - Ingrédients culinaires transformés

Groupe 3 - Aliments transformés

Groupe 4 - Produits alimentaires et boissons ultra-transformés

# LE NUTRI-SCORE

- ❖ Mis en place dans le cadre de la loi de Santé de 2016 du gouvernement français

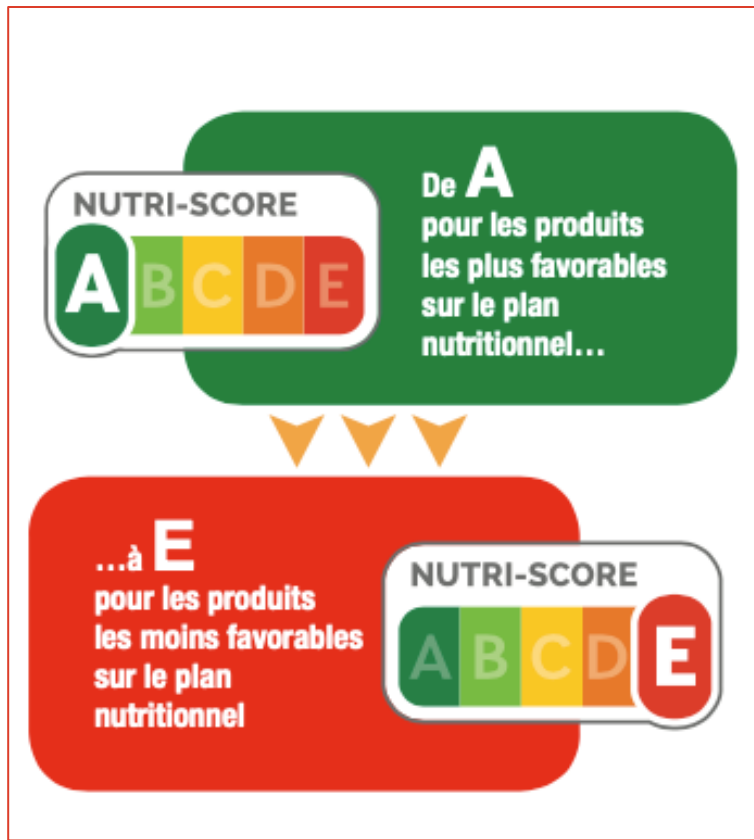
🎯 L'objectif : aider les consommateurs à acheter des aliments de meilleure qualité nutritionnelle

- 2 éléments :

- Un score entre -15 et 40

- Une lettre entre A et E

- Un logo à 5 couleurs apposé sur la face avant des emballages qui informe les consommateurs sur la qualité nutritionnelle d'un produit



Plus d'Informations sur [mangerbouger.fr](http://mangerbouger.fr)





# 02

## PRÉSENTATION DE L'IDÉE D'APPLICATION

---



# IDÉE D'APPLICATION



L'apposition du Nutri-Score sur les produits est facultative. De nombreuses entreprises et distributeurs choisissent de ne pas l'afficher



... Mais les valeurs nutritionnelles sont obligatoires



Proposer une application qui prédise le Nutri-Score à partir des données nutritionnelles des aliments





# 03

## NETTOYAGE DES DONNÉES

---

# ENVIRONNEMENT DE DÉVELOPPEMENT

---

## ANACONDA

Installation d'Anaconda:  
plateforme de distribution  
python la plus populaire

## ENVIRONNEMENT VIRTUEL

Mise en place d'un  
environnement virtuel dédié  
au projet

## INSTALLATION DES PAQUETS

Installation des paquets  
nécessaires (numpy,  
pandas, matplotlib, seaborn,  
sklearn, scipy) avec la  
commande **pip install**



# SÉLECTION DES DONNÉES

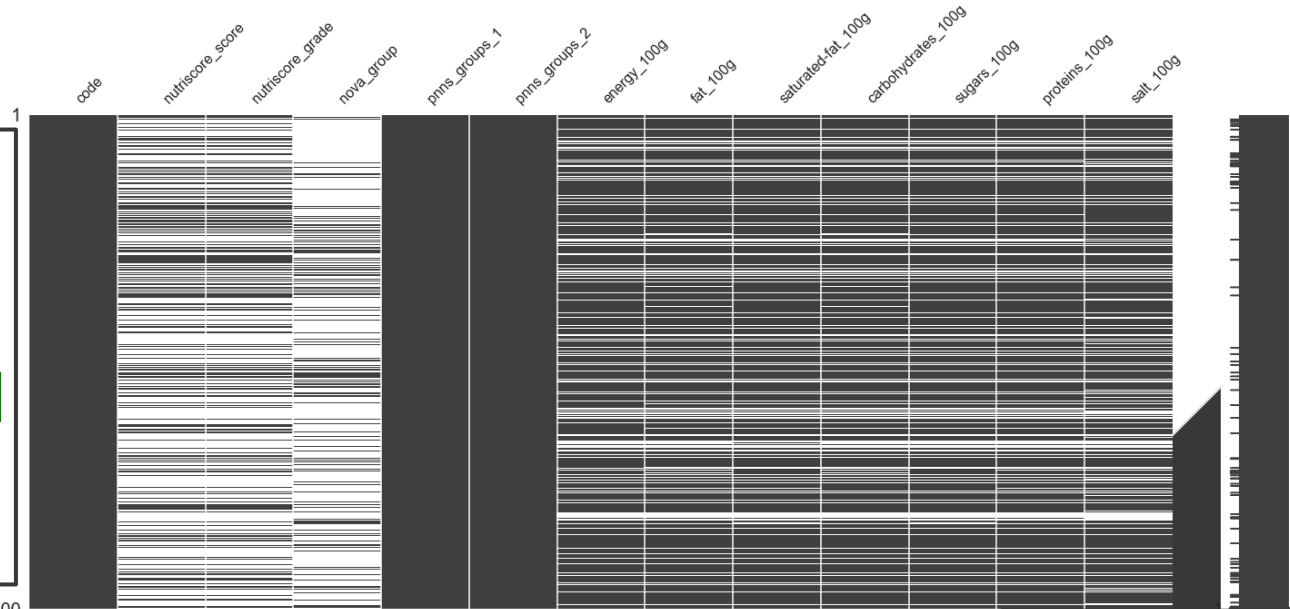
1. L'appel à projet vient de Santé Public France
  - Restriction aux données des aliments vendus en France FR
2. Réduction du nombre de colonnes
  - Suppression des colonnes redondantes et non pertinentes à l'analyse (*date de création, créateur, date de modification, urls d'accès aux images produit, quantité, packaging, ingrédients, marques.....*)
  - Suppression des colonnes peu renseignées (*taux de remplissage <25%*)
3. Le Nutri-Score ne s'applique pas aux boissons alcoolisées
  - Suppression des lignes ayant trait aux produits alcoolisés (*Groupe PNNS2 « Alcoholic beverages »*)

# COLONNES SÉLECTIONNÉES

## 12 COLONNES

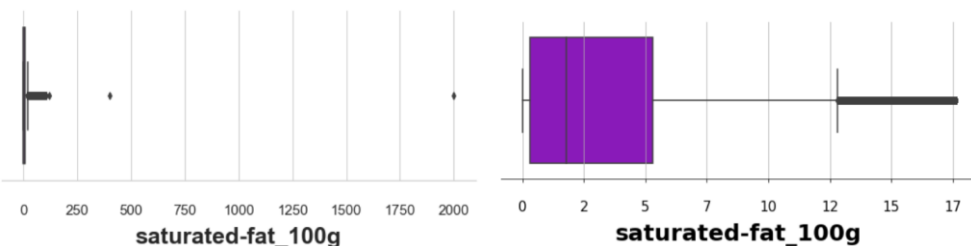
Les colonnes restantes contiennent les informations nutritionnelles, le nutri-score et les groupes PNNS et NOVA

809800



CODE-BARRE	FEATURE A	FEATURE B	FEATURE C	GROUPE PNNS 2
01234567891011	data A	NaN	data C 1	Fruits
01234567891011	NaN	data B	data C 2	Vegetables
01234567891011	data A	data B	data C 1	Fruits

PNNS 2	FEATURE QUAL A	FEATURE QUANT B
Vegetables	Fruits and vegetables	NaN/3.15
Vegetables	NaN/Fruits and vegetables	3.1
Vegetables	Fruits and vegetables	3.2
Fruits	Fruits and vegetables	NaN/4
Fruits	NaN/Fruits and vegetables	4



# NETTOYAGE

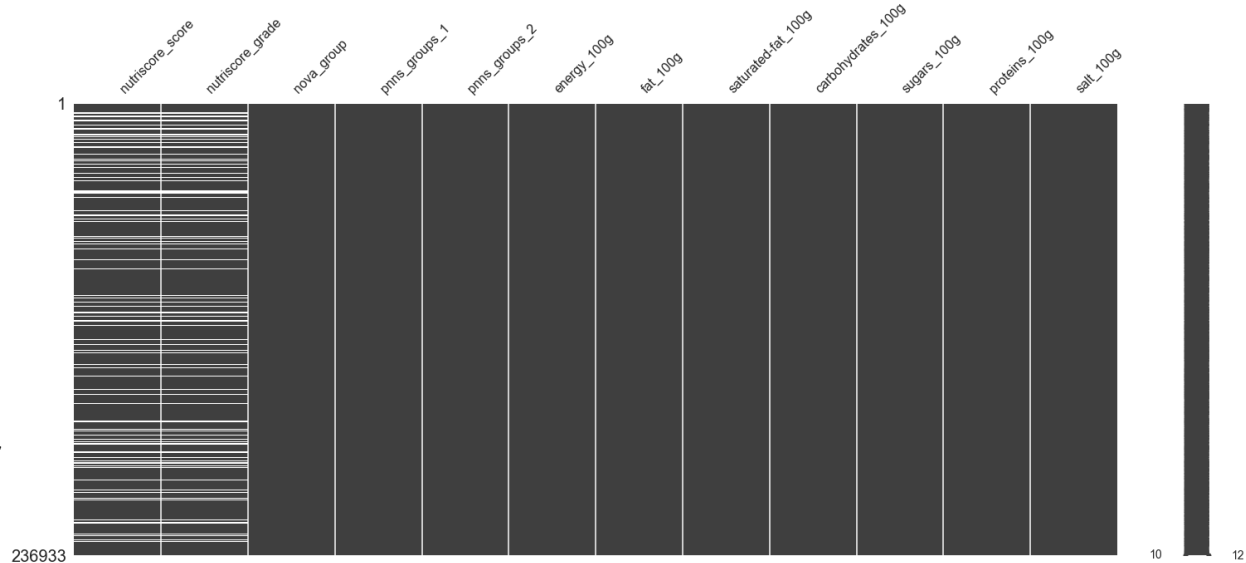
1. Suppression des lignes entièrement vides
2. Harmonisation des orthographes des groupes PNNS
3. Traitement des doublons  
(Fusion des lignes ayant trait au même produit (code-barres identiques))
4. Traitement des valeurs NaN  
(Remplacement des valeurs NaN par la valeur modale [si variable qualitative] ou la moyenne [si variable quantitative] sur le groupe PNNS2)
5. Traitement des valeurs aberrantes  
(Suppression des valeurs aberrantes en utilisant la méthode interquartile range (IQR). Valeur aberrante : écart absolu avec Q1 ou Q3 > 1,5× écart interquartile)

# BILAN

**236933** lignes et **12** colonnes

Toutes les colonnes en dehors des colonnes `nutriscore_score` et `nutriscore_grade` sont entièrement remplies.

Jeu de donnée prêt et exporté pour l'analyse.





# 04

## ANALYSE UNIVARIÉE

---



# ANALYSE UNIVARIÉE

---

MESURES DE  
TENDANCE CENTRALE

Mode et moyenne

MESURES DE FORME

Skewness, Kurtosis

MESURES DE DISPERSION

Variance, écart-type

MESURES DE  
CONCENTRATION

Courbes de Lorenz,  
indice de Gini

	Mode
<b>nutriscore_grade</b>	d
<b>nova_group</b>	4.0
<b>pnns_groups_1</b>	Fish Meat Eggs
<b>pnns_groups_2</b>	One-dish meals

	Mode	Médiane	Moyenne
<b>nutriscore_score</b>	0.0	5.00	7.067164
<b>energy_100g</b>	0.0	799.00	936.514503
<b>fat_100g</b>	0.0	6.50	10.454636
<b>saturated-fat_100g</b>	0.0	1.80	3.614206
<b>carbohydrates_100g</b>	0.0	11.46	22.429668
<b>sugars_100g</b>	0.0	2.90	7.892314
<b>proteins_100g</b>	0.0	6.70	8.615863
<b>salt_100g</b>	0.0	0.67	0.794026

# 1. MESURES DE TENDANCE CENTRALE

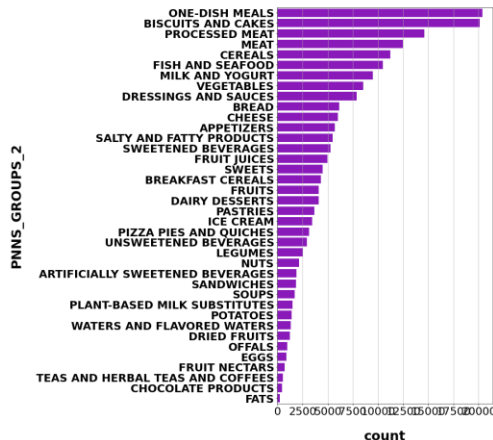
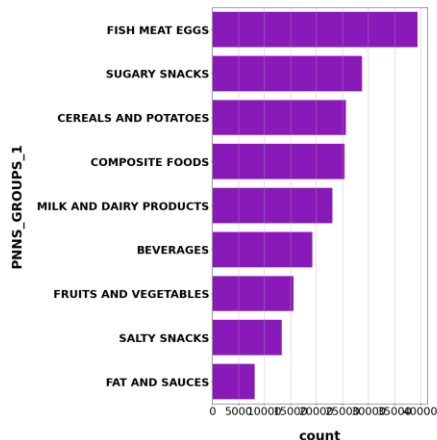
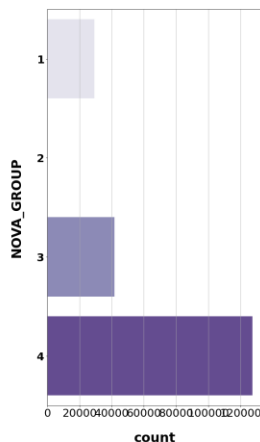
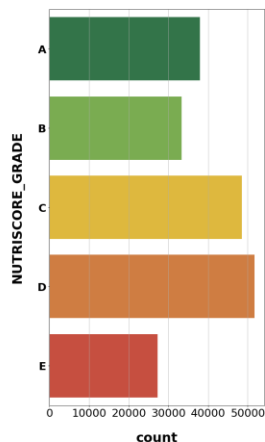
## VARIABLES QUALITATIVES

- Observations liées au mode
  - produits ultra-transformés (nova = 4)
  - produits à consommer en quantité modérée (nutri-grade=d)
  - produits sous la forme de plats préparés
  - produits contenant une proportion importante de poisson, viande et œufs.

## VARIABLES QUANTITATIVES

- Observations liées à la médiane et la moyenne:
  - nutriscore moyen = 7 devrait correspondre à la lettre c du Nutri-Score
  - teneur calorique importante
  - mode < médiane < mean : distribution étalée sur la droite

## DISTRIBUTION DES VARIABLES QUALITATIVES

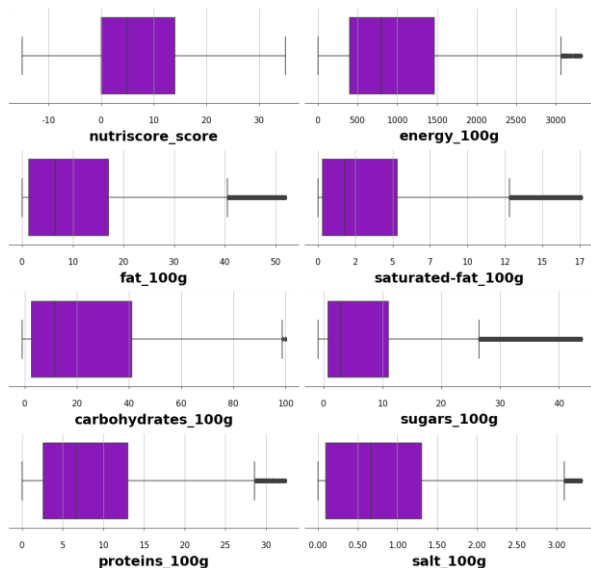


## 2. MESURES DE DISPERSION

### VARIABLES QUALITATIVES

- GROUPES PNNS 1 : Les groupes "fish meat eggs" et "sugary snacks" sont les plus représentés.
- GROUPES PNNS 2: On distingue 2 pics à savoir "One-dish meals" et "Biscuits and cakes"
- GROUPES NOVA : Une écrasante majorité classée 4 = produits ultra-transformés.
- LETTRE NUTRI-SCORE : Répartition relativement homogène sur les 5 lettres, avec une majorité d'aliments classée 'D', et une minorité classée 'E'.

## BOXPLOT DES VARIABLES QUANTITATIVES



	Moyenne	Variance	Écart-type	Coeff de Variation
<b>nutriscore_score</b>	7.067164	68.375584	8.268953	1.170053
<b>energy_100g</b>	936.514503	430032.783456	655.768849	0.700223
<b>fat_100g</b>	10.454636	125.208576	11.189664	1.070306
<b>saturated-fat_100g</b>	3.614206	19.722999	4.441058	1.228778
<b>carbohydrates_100g</b>	22.429668	597.423845	24.442255	1.089729
<b>sugars_100g</b>	7.892314	111.654011	10.566646	1.338853
<b>proteins_100g</b>	8.615863	53.693638	7.327594	0.850477
<b>salt_100g</b>	0.794026	0.574847	0.758186	0.954864

## 2. MESURES DE DISPERSION

### VARIABLES QUANTITATIVES

Sur l'ensemble des aliments observés :

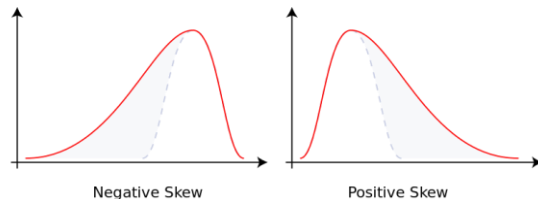
- les valeurs des variables considérées sont très dispersées autour de la moyenne.
- Les variables *saturated-fat\_100g*, *nutriscore\_score* et *sugars\_100g* présentent la dispersion la plus importante
- La variable *energy\_100g* présente la dispersion la moins importante

	Skewness	Kurtosis
nutriscore_score	0.342422	-0.864646
energy_100g	0.554216	-0.666465
fat_100g	1.296427	1.314095
saturated-fat_100g	1.436189	1.085883
carbohydrates_100g	0.976594	-0.410328
sugars_100g	1.613634	1.609956
proteins_100g	0.777983	-0.400924
salt_100g	0.909663	0.204097

### Skewness

$\gamma_1$  : coefficient d'asymétrie

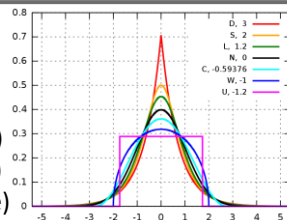
- $\gamma_1 = 0$  : distribution symétrique.
- $\gamma_1 > 0$  : distribution étalée à droite.
- $\gamma_1 < 0$  : distribution étalée à gauche



### Kurtosis

$\gamma_2$  : coefficient d'aplatissement

- $\gamma_2 = 0$  : aplatissement similaire à la normale (mésokurtique)
- $\gamma_2 > 0$  : aplatissement inférieur à la normale (leptokurtique)
- $\gamma_2 < 0$  : aplatissement supérieur à la normale (platykurtique)



## 3. MESURES DE FORME

### VARIABLES QUANTITATIVES

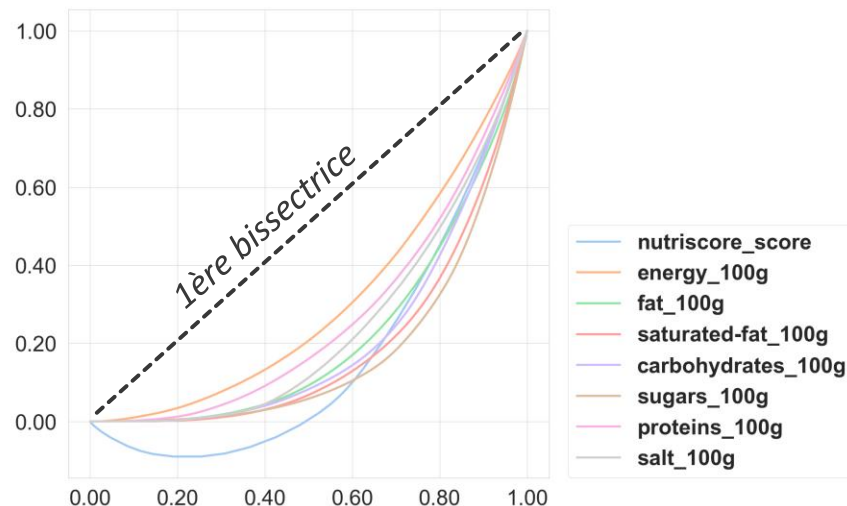
#### SKEWNESS : COEFFICIENT D'ASYMÉTRIE $\gamma_1$

- $\gamma_1 > 0$  : Les distributions de l'ensemble des variables considérées pour les aliments observés sont toutes plus ou moins **étalées sur la droite**.

#### KURTOSIS : COEFFICIENT D'APLATISSEMENT $\gamma_2$

- Les teneurs en **graisses, graisses saturées, sucres et sel** ont une distribution **leptokurtique**, suggérant des valeurs concentrées.
- Le score Nutri-Score, les teneurs caloriques, en **carbohydrates et en protéines** ont une distribution **platikurtique**, soit relativement aplatie.

## VARIABLES QUANTITATIVES - COURBES DE LORENZ



Indice Gini	
<b>nutriscore_score</b>	0.666496
<b>energy_100g</b>	0.396106
<b>fat_100g</b>	0.563032
<b>saturated-fat_100g</b>	0.620654
<b>carbohydrates_100g</b>	0.579179
<b>sugars_100g</b>	0.650349
<b>proteins_100g</b>	0.472008
<b>salt_100g</b>	0.522771

## 4. MESURES DE CONCENTRATION

### VARIABLES QUANTITATIVES

#### COURBES DE LORENZ

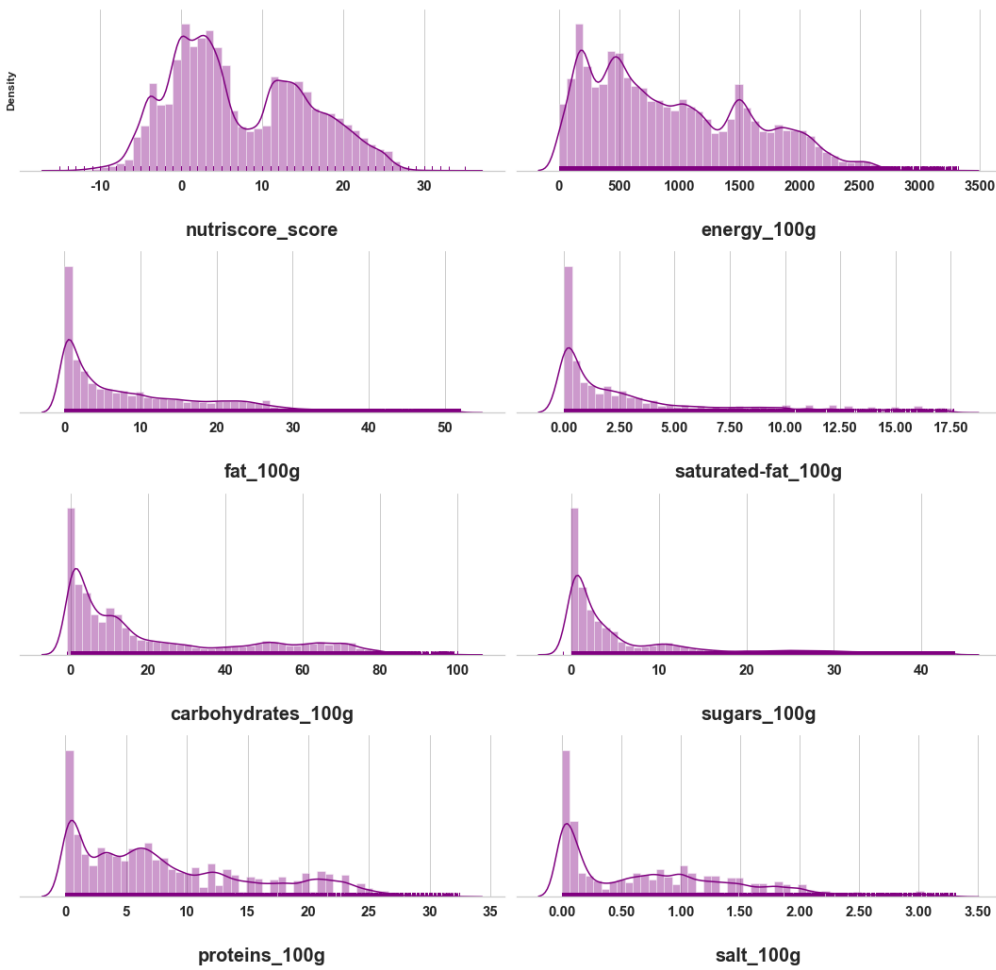
Répartition d'une variable : plus la courbe est proche de la 1ère bissectrice, plus la répartition est égalitaire

#### COEFFICIENT DE GINI

Plus il est élevé (varie entre 0 et 1), plus la répartition de la variable est inégalitaire

L'ensemble des variables présente une répartition relativement inégale.

## DISTRIBUTION DES VARIABLES QUANTITATIVES



## REPRÉSENTATION GRAPHIQUE DES DISTRIBUTIONS DES VARIABLES QUANTITATIVES

- ✓ Confirmation des informations  
fournies par les indicateurs  
statistiques :
- inégales
  - étalées sur la droite
  - dispersion importante des valeurs
  - aplatissement



# 05

## ANALYSE MULTIVARIÉE

---



# ANALYSE MULTIVARIÉE

---

## RELATIONS ENTRE VARIABLES QUALITATIVES ET QUANTITATIVES

Nutri-Score (lettre et score)  
et les autres variables

## RELATIONS ENTRE VARIABLES QUALITATIVES

Lettre Nutri-Score et les  
autres variables qualitatives

## RELATIONS ENTRE VARIABLES QUANTITATIVES

Score Nutri-Score et les autres  
variables quantitatives

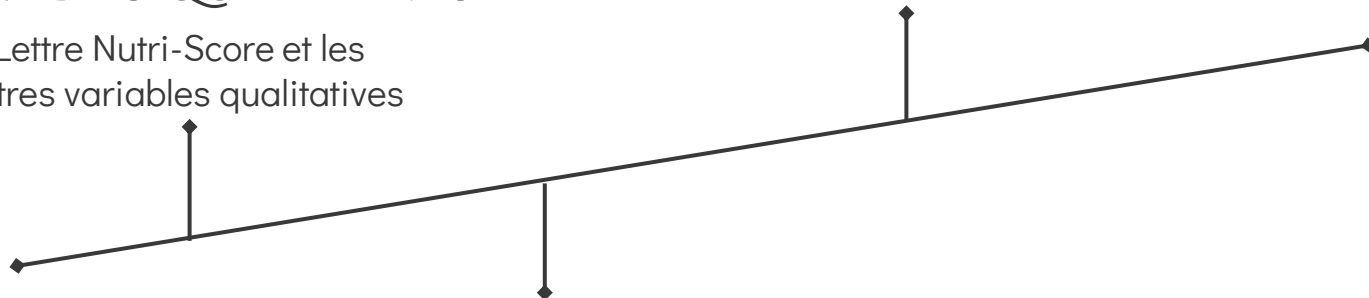


TABLEAU DE CONTINGENCE  
AVEC MISE EN LUMIÈRE DES RELATIONS PROBABLES (KHI-2)

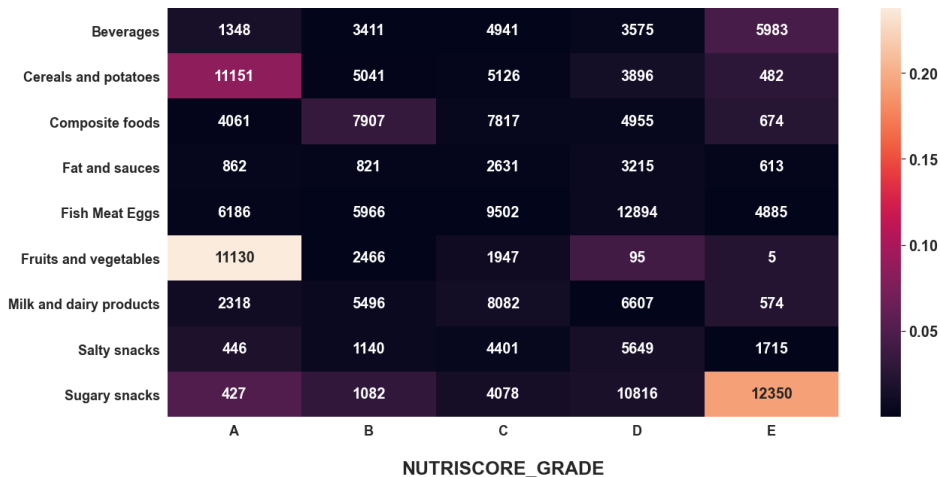
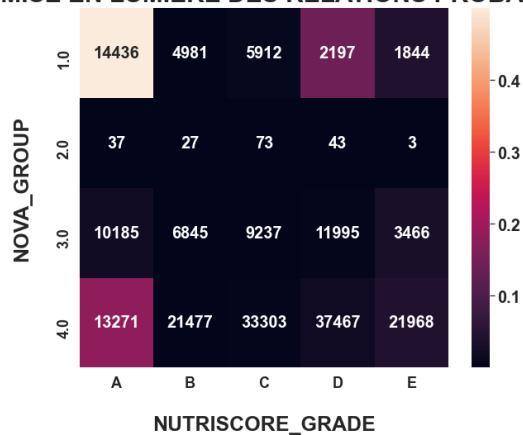


TABLEAU DE CONTINGENCE  
AVEC MISE EN LUMIÈRE DES RELATIONS PROBABLES (KHI-2)



# 1. VARIABLES QUALITATIVES

## PNNS 1/LETTRE NUTRIScore

❖ Corrélations les plus probables :

● « Fruits and vegetables » - A

● « Sugary snacks » - E

❖ Test Khi-2 : La probabilité d'obtenir ces valeurs si les variables sont indépendantes (P-value) est proche de 0. => Les variables sont dépendantes

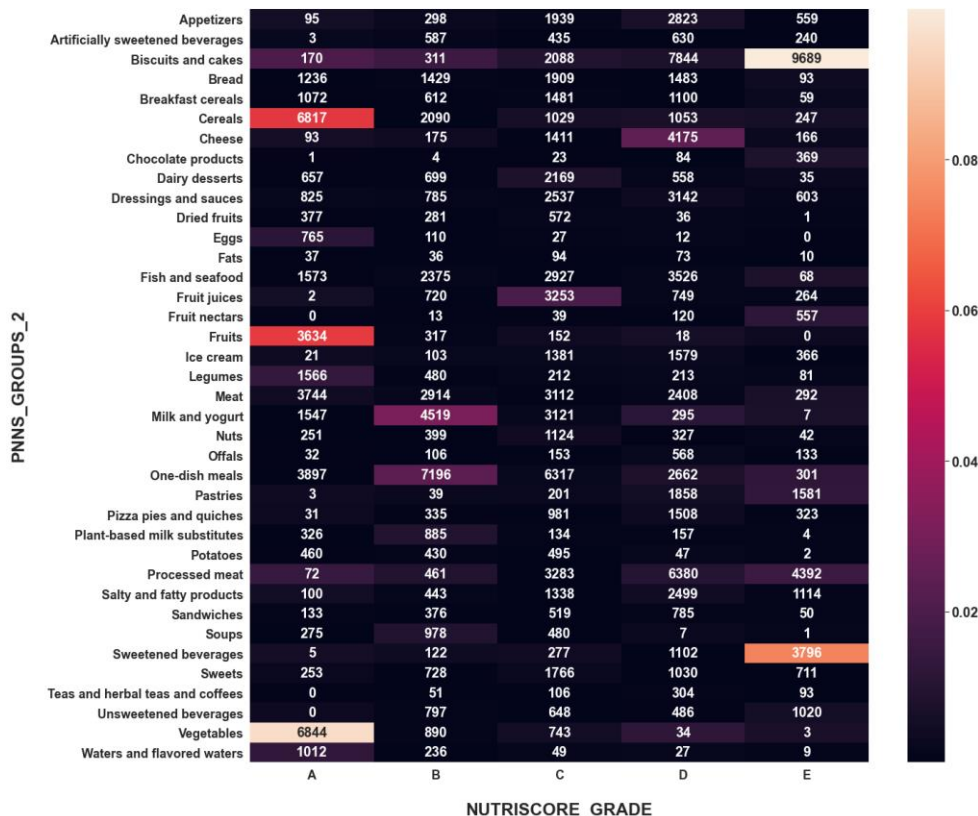
## NOVA/LETTRE NUTRIScore

❖ Corrélation la plus probable :

● 1 - A

❖ Test Khi-2 : La probabilité d'obtenir ces valeurs si les variables sont indépendantes (P-value) est proche de 0. => Les variables sont dépendantes

TABLEAU DE CONTINGENCE  
AVEC MISE EN LUMIÈRE DES RELATIONS PROBABLES (KHI-2)



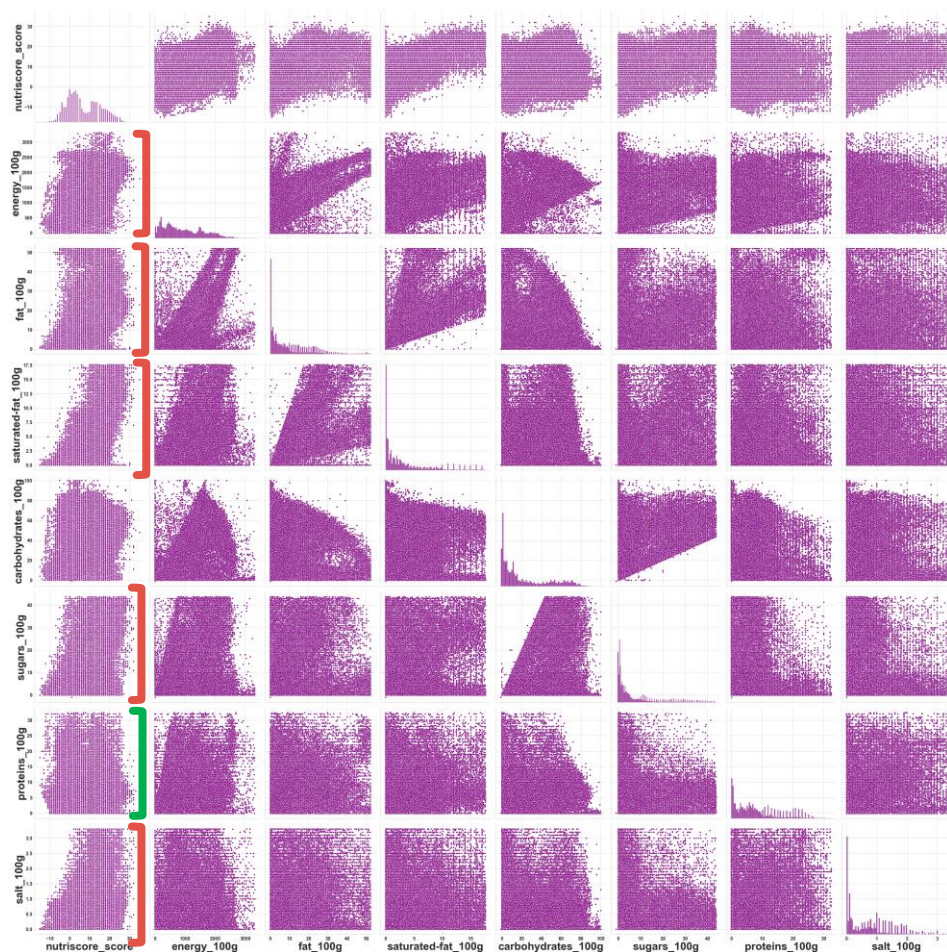
# 1. VARIABLES QUALITATIVES

## PNNS 2/LETTRE NUTRISCORE

❖ Corrélations les plus probables :

- «Vegetables » - A
- « Sweetened beverages » - E
- « Biscuits and cakes » - E

❖ Test Khi-2 : La probabilité d'obtenir ces valeurs si les variables sont indépendantes (P-value) est proche de 0. => Les variables sont dépendantes

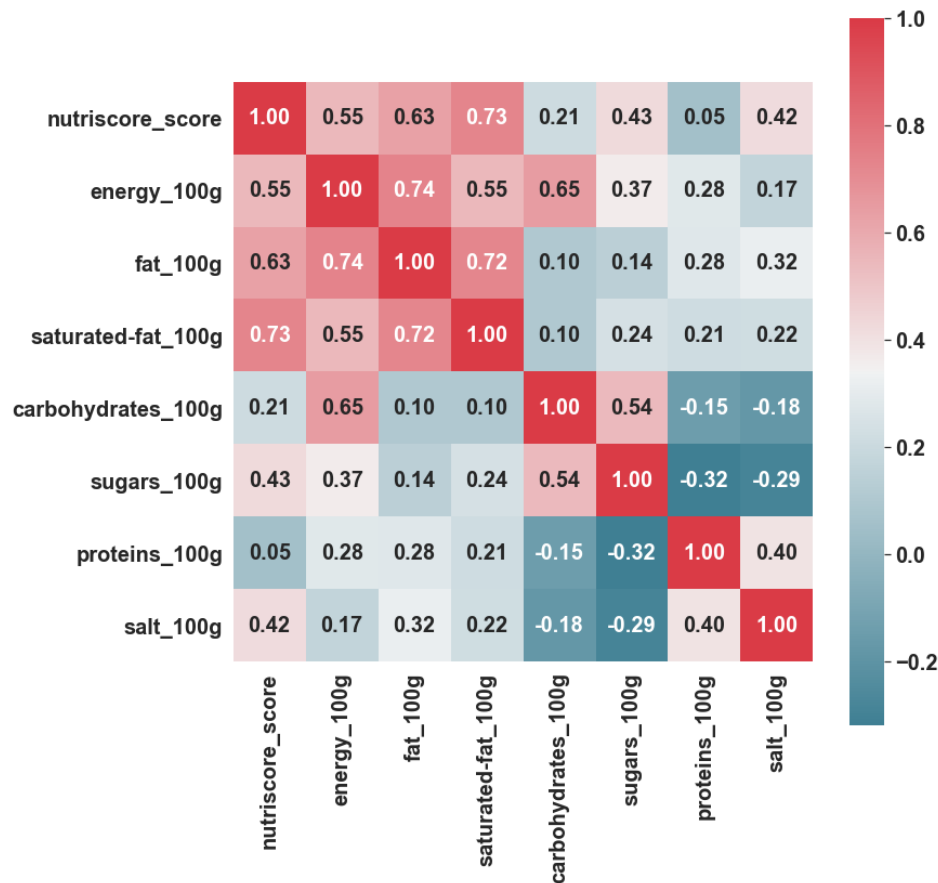


## 2. VARIABLES QUANTITATIVES

### SCORE NUTRISCORE/ AUTRES

- Plus la teneur calorique est importante...
- Plus la teneur en graisses est importante...
- ● Plus la teneur en graisses saturées est importante...
- Plus la teneur en sucres est importante...
- Plus la teneur en sel est importante...
- ✗ Plus le Nutri-Score est mauvais
- Plus la teneur en protéines est importante...
- ✓ Plus le score Nutri-Score est bon

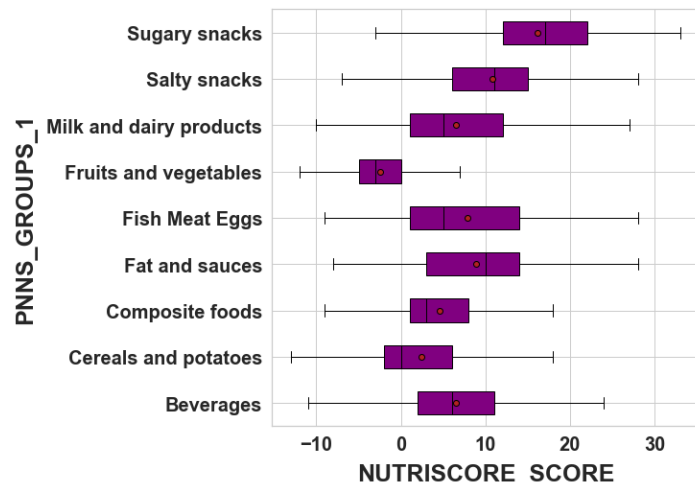
## COEFFICIENT DE CORRÉLATION DE PEARSON



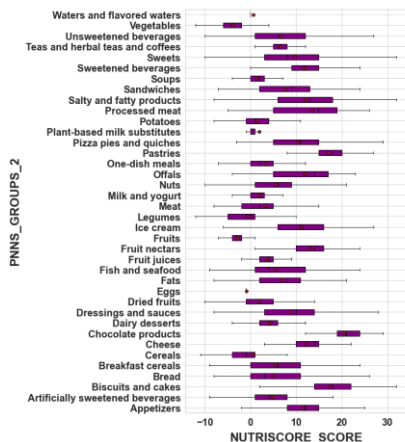
## 2. VARIABLES QUANTITATIVES

- ❖ Mise en lumière des relations linéaires
- Concentration des zones rougeâtres dans le cadran en haut à gauche :
  - Le score Nutri-Score
  - La teneur calorique
  - La teneur en graisses
  - La teneur en graisses saturées
  - La teneur en sucre
- Paires de variables corrélées sur la diagonale:
  - Teneur en carbohydrates et en sucres
  - Teneur en protéines et en sel

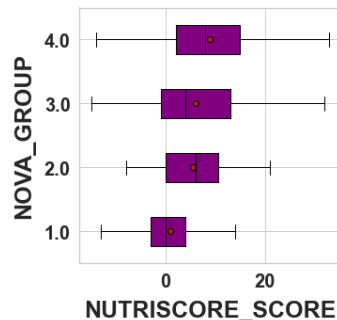
BOXPLOT pnns\_groups\_1/nutriscore\_score



BOXPLOT pnns\_groups\_2/nutriscore\_score



BOXPLOT nova\_group/nutriscore\_score



### 3. VARIABLES QUALITATIVES ET QUANTITATIVES

#### GROUPES PNNS 1 :

- le plus mauvais nutriscore : Sugary snacks suivi des groupes Salty snacks et Fat and sauces
- le meilleur nutriscore : Fruits and vegetables suivi de Cereals and potatoes

#### GROUPES PNNS 2 :

- les plus mauvais nutriscore : Chocolate products et Biscuits and cakes
- les meilleurs nutriscore : Legumes, Fruits et Cereals

#### GROUPES NOVA :

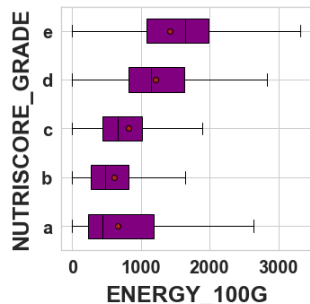
- les aliments du groupe nova 1 ont le meilleur nutriscore
- les aliments du groupe nova 4 ont un mauvais nutriscore

# 3. VARIABLES QUALITATIVES ET QUANTITATIVES

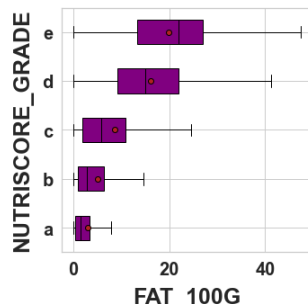
## NUTRIScore\_GRADE/ VARIABLES QUANTITATIVES

- « Saturated-fat, fat, sugars, energy » - E
- « Proteins » - A
- « Salt » - D

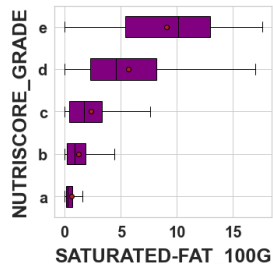
BOXPLOT nutriscore\_grade/energy\_100g



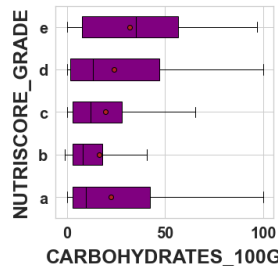
BOXPLOT nutriscore\_grade/fat\_100g



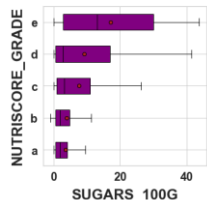
BOXPLOT nutriscore\_grade/saturated-fat\_100g



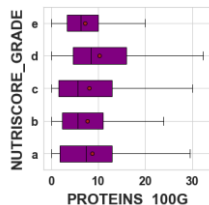
BOXPLOT nutriscore\_grade/carbohydrates\_100g



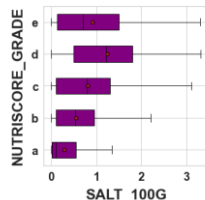
BOXPLOT nutriscore\_grade/sugars\_100g



BOXPLOT nutriscore\_grade/proteins\_100g



BOXPLOT nutriscore\_grade/salt\_100g



VARIABLES QUALITATIVES	RAPPORT DE CORRÉLATION ÉTA CARRÉ AVEC LE SCORE NUTRI-SCORE
PNNS_GROUPS_1	35,21%
PNNS_GROUPS_2	52,42%
NOVA_GROUP	11,44%

### 3. VARIABLES QUALITATIVES ET QUANTITATIVES

#### ANOVA

- L'ensemble des variables qualitatives sont corrélées avec le Nutri-Score
- Le groupe PNNS 2 est la variable qualitative ayant la relation la plus forte avec le score Nutri-Score.

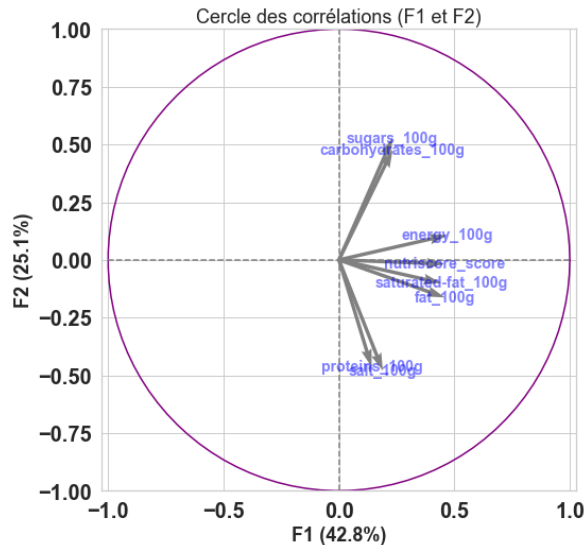
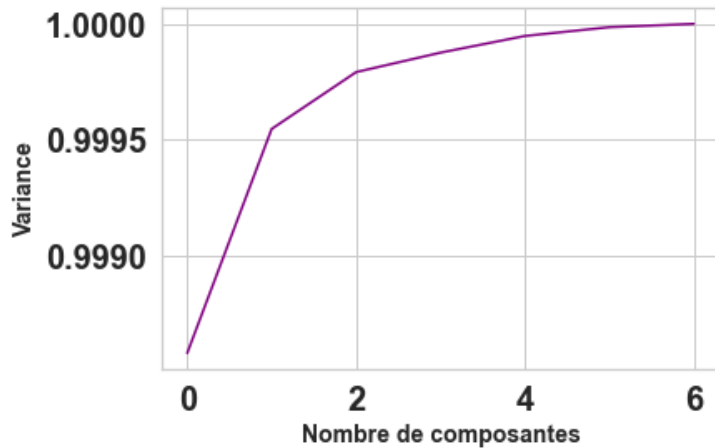




# 06

## REDUCTION DE DIMENSION

---



# ANALYSE EN COMPOSANTE PRINCIPALE (ACP)

Objectif : trouver le nombre minimal de composantes qui permet de préserver 99% de la variance des données

Observation: Avec 1 seule composante, on arrive déjà à garder plus de 99.95% de la variance des données

Cercle de corrélation:

Les variables les plus corrélées à F1 sont : nutriscore\_score , energy\_100g, saturated-fat\_100g et fat\_100g.

Les variables les plus corrélées à F2 sont : sugars\_100g, carbohydrate\_100g, proteines\_100g, salt\_100g.



# 07

## FAITS PERTINENTS, CONCLUSION ET PERSPECTIVES

---

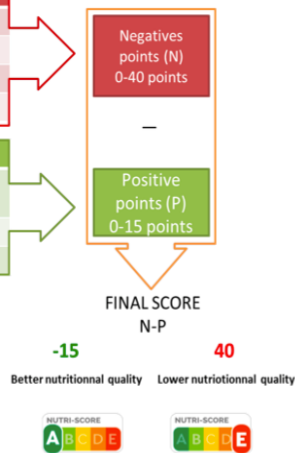
## RÉSULTATS DE L'ANALYSE

QUANTITATIVE VARIABLES	SIGNIFICATION	INFLUENCE
energy_100g	Teneur calorique	
saturated-fat_100g	Teneur en graisses saturées	
sugars_100g	Teneur en sucres	
proteins_100g	Teneur en protéines	
salt_100g	Teneur en sel	

Nutriments /100g	Points
Energy (KJ)	0-10
Sugars (g)	0-10
Saturated Fatty acids (g)	0-10
Sodium (g)	0-10

Element /100g	Points
Fruits, vegetables, pulses, nuts, and rapeseed, walnut and olive oils, (%)	0-5
Fibres (g)	0-5
Proteins (g)*	0-5

\*Depending on the number of negative points and the content in « fruits, vegetables, pulses, nuts and oils », proteins are taken into account or not.



## FAITS PERTINENTS ET CONCLUSION

- L'analyse a permis d'établir une corrélation entre le Nutri-Score et plusieurs variables quantitatives et qualitatives. (Khi-2, rapports de corrélation de Pearson, de Spearman, éta carré)
- Ces variables semblent correspondre aussi bien dans leur nature que dans leur influence, à celles utilisées dans le calcul du Nutri-Score.
- Il devrait donc être possible de créer un modèle qui permette de prédire le score Nutri-Score
- Pour ce faire, l'application mettra à profit les données nutritionnelles et de classification alimentaire les plus corrélées établies dans cette analyse, à savoir : la teneur calorique, la teneur en sucres, la teneur en graisses saturées, la teneur en sel, la teneur en protéines, le classement PNN2



# PERSPECTIVES

---

- Implémentation effective du modèle de prédiction du nutri-score
- Ajout d'un aspect « recommandation » à l'application

# MERCI

---

Des questions ?

