

# FORMATION DATA SCIENTIST

## Projet 2

Hui Guan TAI



Analysez des données nutritionnelles

# Sommaire

- Objectif
- Principes d'une alimentation saine
- Données à analyser
- Stratégie de nettoyage
- Bilan de nettoyage
- Exploration (Plan)
  - Histogrammes / Boxplots
  - Analyse univariée
  - Analyse multivariée
  - Analyse en Composantes Principales
- Bilan d'analyse

# Objectif

Faire une analyse de données alimentaires  
pour aider le site *Lamarmite* à construire son générateur de recettes saines

The screenshot shows a search interface for healthy salads. At the top, there is a navigation bar with several categories: Composée, Minceur, Originale, Fraîcheur, Hiver, Ete, Printemps, Recette, Pates, and Avocat. Below the navigation bar, there are four tabs: Épingles (selected), Vos épingles, Membres, and Tableaux. The main content area displays four salad recipes:

- La salade de Cobb**: A large salad with chicken, bacon, avocado, and tomatoes. It has 7,3k pins.
- Salade d'automne - Mâche**: A salad with mâche, chèvre, pomme, and noix. It has 11,6k pins.
- Salade mâche, jambon de Bayonne, mozzarella**: A salad with mâche, jambon de Bayonne, and mozzarella. It has 7,2k pins.
- Salade healthy : Salade fraîcheur - 11 salades légères et...**: A salad with various vegetables like cherry tomatoes, cucumbers, and feta. It has 30,1k pins.

At the bottom right, there are two circular icons: a plus sign (+) and a question mark (?)

# Principes d'une alimentation saine

- Les besoins nutritionnels de base à connaître
  - Macro nutriments
    - Protéines
      - Aident au bon fonctionnement des organes grâce aux acides aminés dont elles sont constituées
    - Lipides
      - sont à la base de la fabrication de toutes nos cellules, de notre système hormonal, et de toutes nos membranes cellulaires
    - Glucides
      - sont à la base de la fabrication de toutes nos cellules, de notre système hormonal, et de toutes nos membranes cellulaires
  - Micro nutriments
    - Vitamine A, C, E
      - Antioxydants majeurs, protègent les cellules membranaires des dommages oxydatifs et préviennent de nombreuses maladies
      - Présents dans les fruits, légumes, le thé vert ...
    - Minéraux
      - Calcium
      - Fer
      - Magnésium
      - Cuivre
- Une bonne alimentation fournit à l'organisme:
  - Les nutriments de base essentiels
  - L'énergie nécessaire sans exposer à la toxicité ou un gain de poids excessif

# Données à analyser

< > Dataset

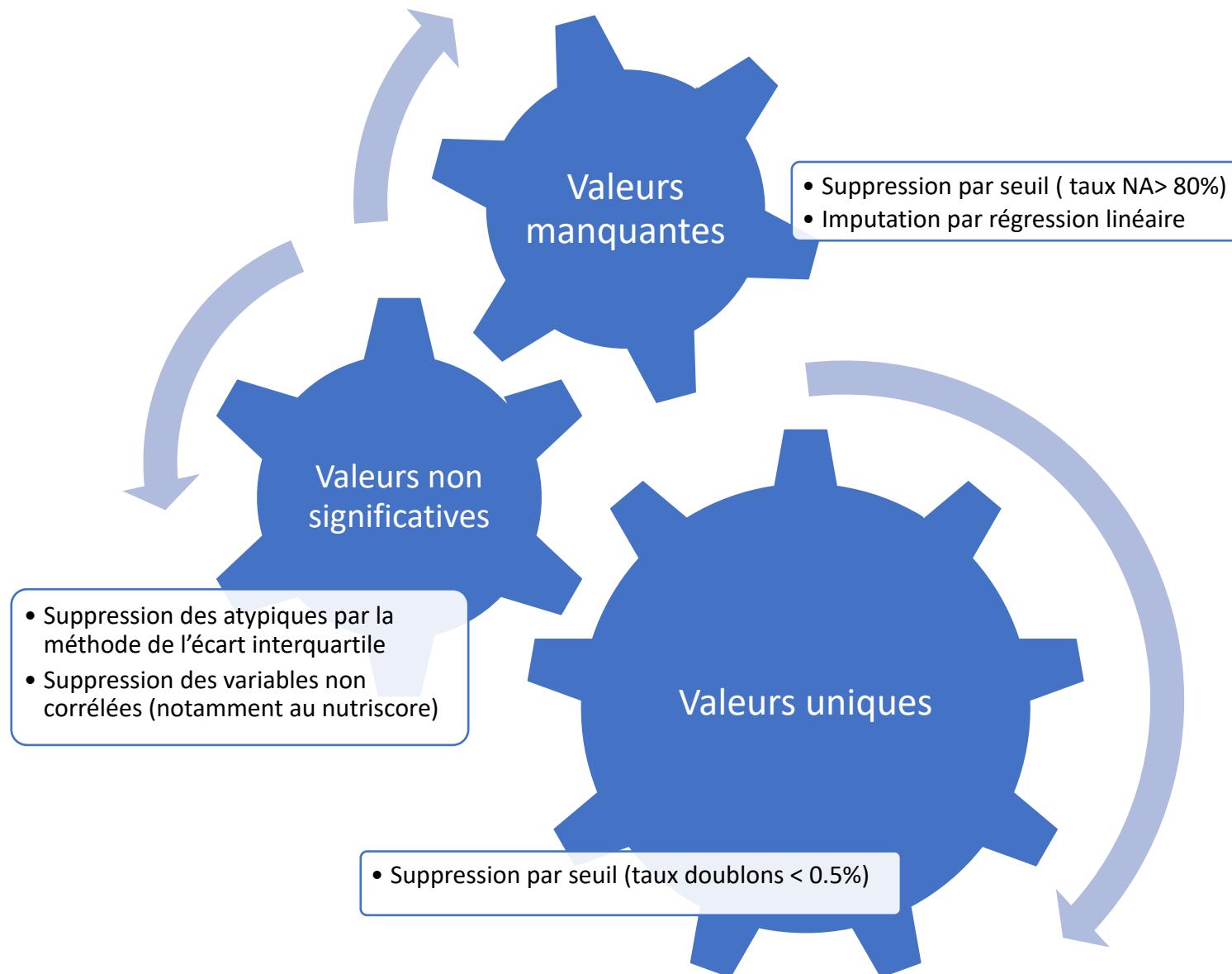
☰ ⚖ ☰ 🔍

Nom	Date de modification	Taille	Type
data-fields.txt	14/01/2021 13:42	5 Ko	Format texte
fr.openfoodfacts.org.products.csv	21/04/2017 11:54	847,4 Mo	Document CSV

320 772 Observations  
162 Catégories



# Stratégie de nettoyage



# Bilan de nettoyage



Valeurs  
manquantes  
(Carbohydrates)

- 24% -> 19%
- Boxplot inchangé

Nombre de  
catégories

- 162 -> 30
- ↴81%

Mémoire  
occupée

- 396 Mb -> 73 Mb
- ↴81%

## Histogrammes/Boxplots

- 9 variables numériques 100g

## ACP

- 9 composantes

## EXPLORATION

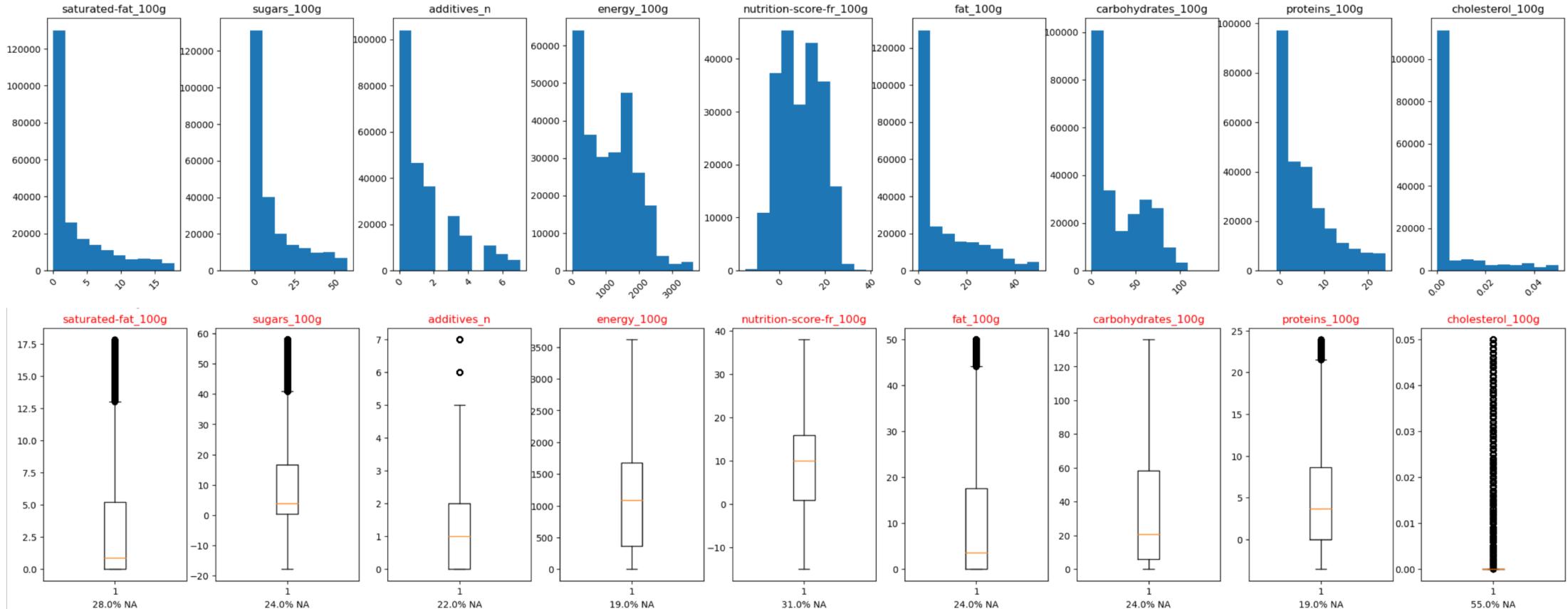
## Analyse Univariée

- Régression Linéaire OLS  
Nutrition-grade~Nutrition-score

## Analyse Multivariée

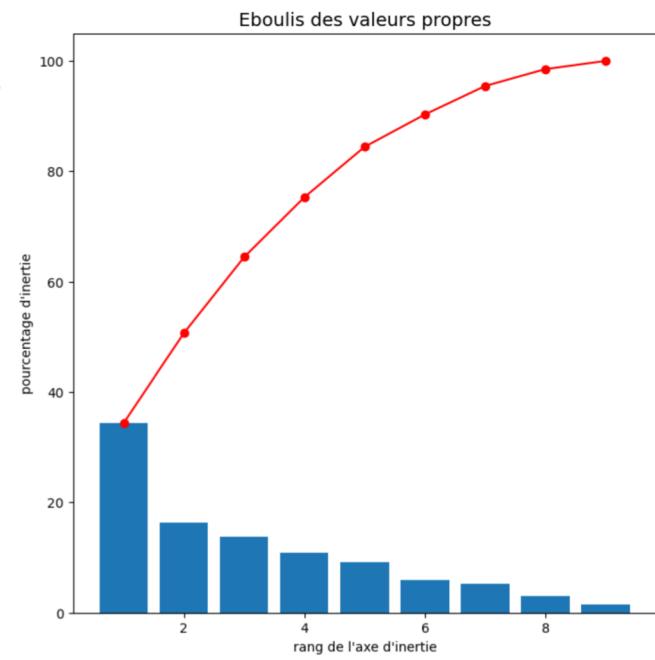
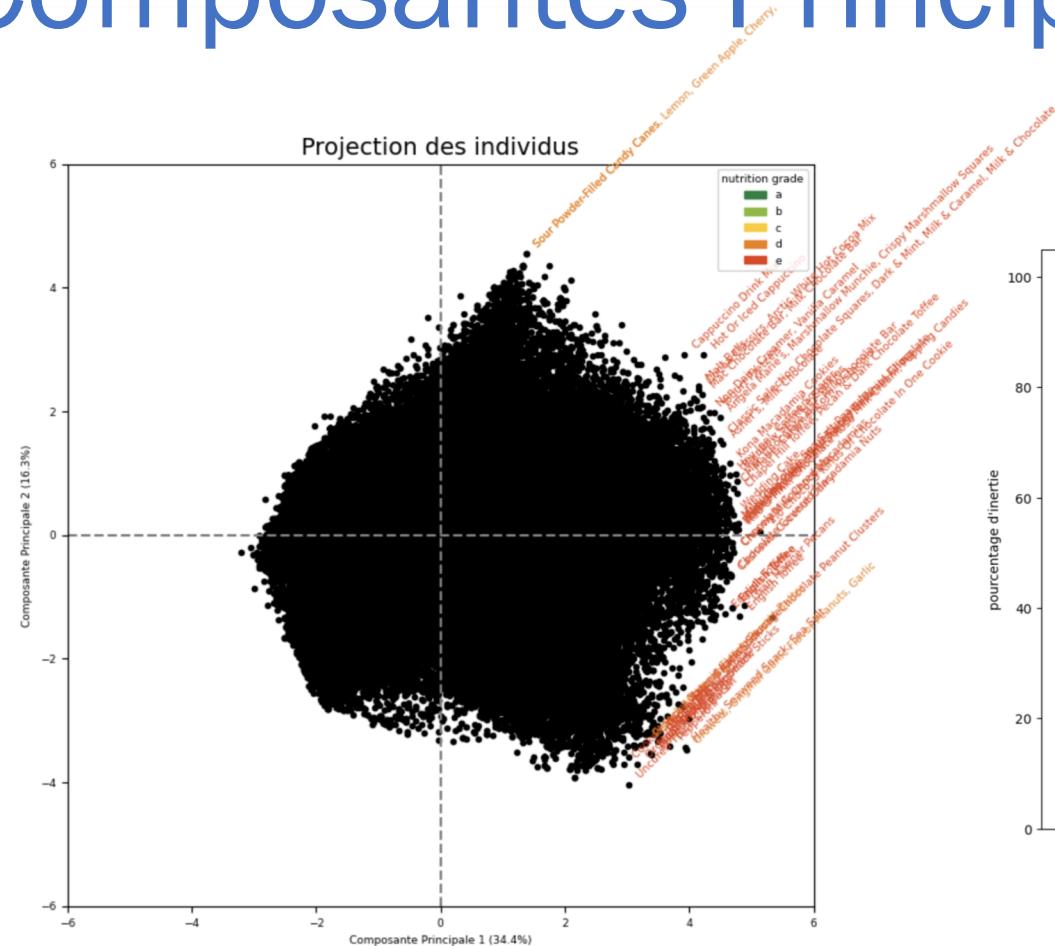
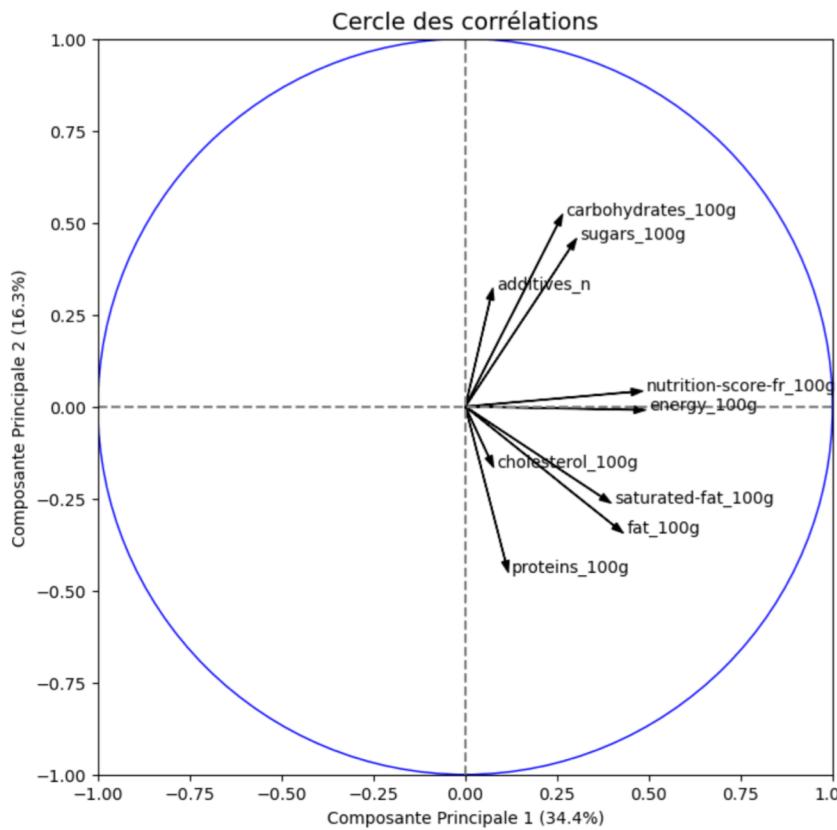
- Régression Linéaire OLS  
Nutrition-grade~brands+additives

# Histogrammes / Boxplots



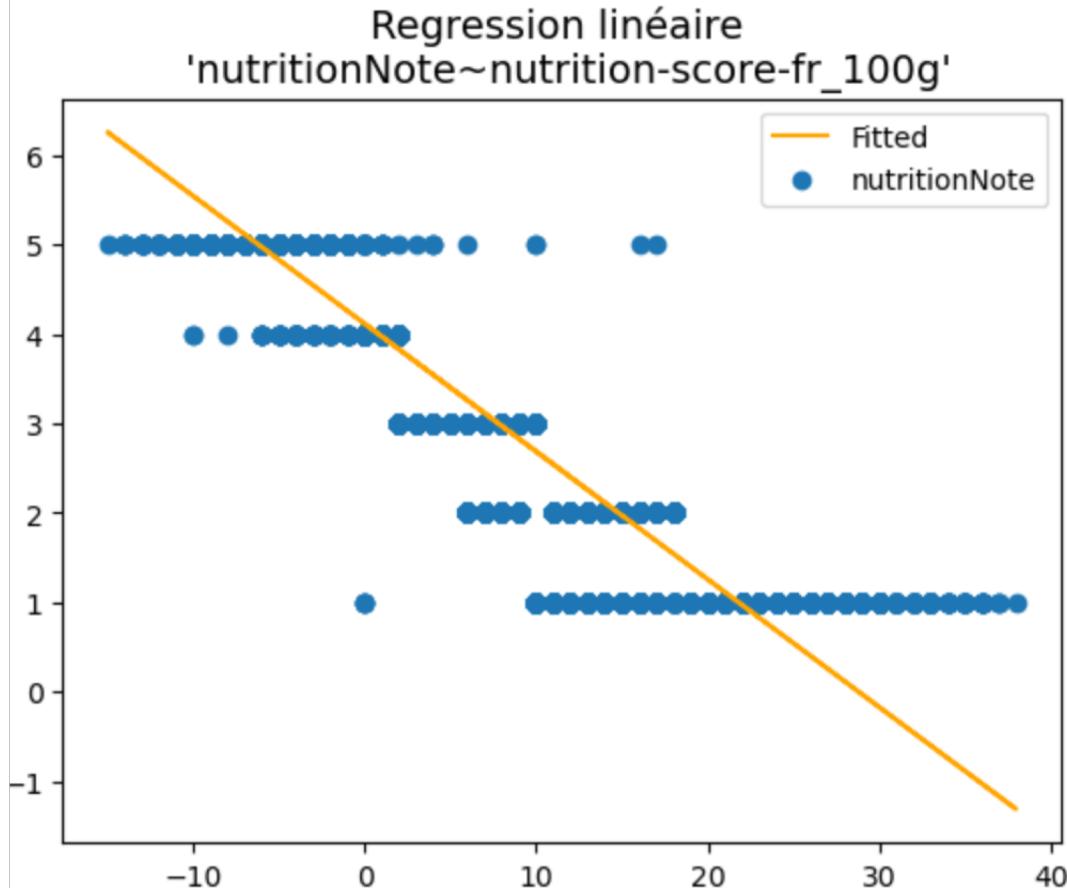
- La variable Nutrition-score présente une distribution bimodale, on s'attend à trouver deux groupes de produits (score proche de 0 et autour de 15)
- Toutes les autres variables ne suivent pas une distribution normale

# Analyse en Composantes Principales



- On distingue 3 types d'aliments (qui expliquent 50% des données)
    - Les aliments riches en énergie & graisses (axe des  $x>0$ )
    - Les aliments riches en sucres rapides & lents (axe des  $y>0$ )
    - Les aliments riches en protéines (axe des  $y<0$ )

# Analyse univariée



OLS Regression Results						
Dep. Variable:	nutritionNote	R-squared:	0.916			
Model:	OLS	Adj. R-squared:	0.916			
Method:	Least Squares	F-statistic:	2.398e+06			
Date:	Sat, 06 Mar 2021	Prob (F-statistic):	0.00			
Time:	19:22:18	Log-Likelihood:	-1.0691e+05			
No. Observations:	221210	AIC:	2.138e+05			
Df Residuals:	221208	BIC:	2.139e+05			
Df Model:	1	Covariance Type:	nonrobust			
	coef	std err	t	P> t	[0.025	0.975]
const	4.1117	0.001	3464.234	0.000	4.109	4.114
nutrition-score-fr_100g	-0.1427	9.21e-05	-1548.558	0.000	-0.143	-0.142

- $R^2$  élevé => bon taux de prédition du modèle
- Coefficient  $< 0$  => plus le nutriscore est élevé, plus la note de nutrition est basse (proche de 1)
- Le P-value associé au coefficient (et teste l'hypothèse nulle que le coefficient n'a pas d'effet) est nul: donc le coefficient est pertinent

# Analyse multivariée

	brands	additives_fr	nutrition_grade_fr
nutrition_grade_fr			
<b>a</b>	34975	13239	0
<b>b</b>	33501	17754	1
<b>c</b>	44526	25687	2
<b>d</b>	61526	39721	3
<b>e</b>	42044	29781	4

OLS Regression Results						
Dep. Variable:	nutrition_grade_fr	R-squared:	0.959			
Model:	OLS	Adj. R-squared:	0.917			
Method:	Least Squares	F-statistic:	23.16			
Date:	Sat, 06 Mar 2021	Prob (F-statistic):	0.0414			
Time:	18:51:56	Log-Likelihood:	-0.86622			
No. Observations:	5	AIC:	7.732			
Df Residuals:	2	BIC:	6.561			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	2.2796	1.160	1.965	0.188	-2.713	7.272
brands	-0.0002	5.55e-05	-3.508	0.073	-0.000	4.41e-05
additives_fr	0.0003	5.99e-05	5.400	0.033	6.57e-05	0.001

Régression linéaire ‘Nutrition\_grade\_fr~brands+additives\_fr’

- $R^2$  élevé => bon taux de prédiction du modèle
- Coefficients quasi nuls => les variables ‘brands’ (marque) et ‘additives’ (additifs) auraient une contribution nulle pour prédire le nutrition-grade
- Le P-value associé au coefficient (teste l’hypothèse nulle que le coefficient n’a pas d’effet) est proche de 0.05: on ne peut pas conclure sur la pertinence du coefficient

# Bilan d'analyse

Pour générer des recettes saines, le site Lamarmite pourra s'appuyer sur les caractéristiques suivantes pour choisir ses éléments



- Le Nutrition Grade (notée ‘a’ la meilleure note et ‘e’ la moins bonne) indique la note de qualité d'ensemble de l'aliment
- Le Nutrition Score élevé indique une alimentation riche en énergie & graisse, mais ne correspond pas forcément à un bon Nutrition Grade
- Les aliments se répartissent suivant 3 groupes
  - Les aliments riches en protéines
  - Les aliments riches en énergie & graisse
  - Les aliments riches en sucres lents ou rapides

