P3 – CONCEVEZ UNE APPLICATION AU SERVICE DE LA SANTÉ PUBLIQUE

17/02/2021

Etudiant : Luc Rogers
Mentor : Etienne Sanchez

Sommaire

- 1. Problématique
- 2. Présentation du concept d'application
- 3. Nettoyage
- 4. Analyse exploratoire
  - ANOVA
  - ACP
  - K-Means
- 5. Démonstration du concept

Problématique

- Appel à projet agence « Santé publique France »
  - Idée innovante d'application en lien avec l'alimentation

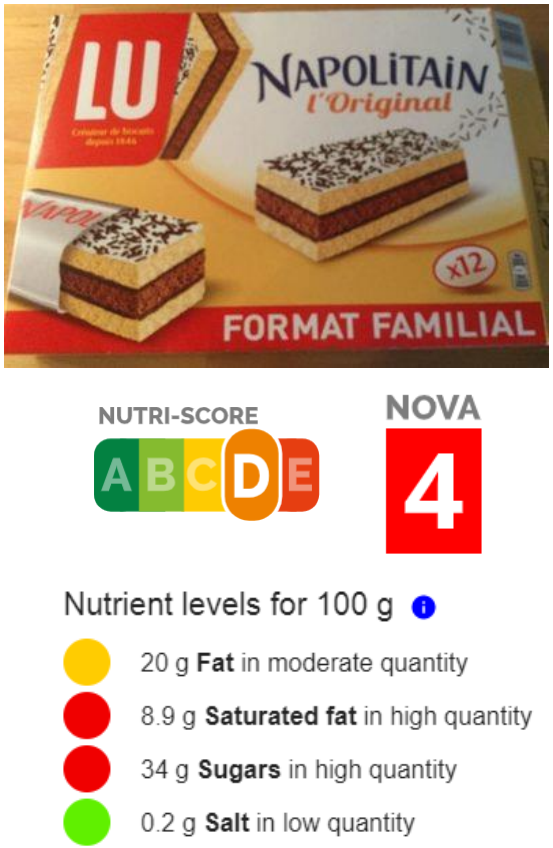
- Objectifs :
  - Elaborer une idée d'application
  - Automatiser le traitement du jeu de données
  - Analyser le dataset en vue d'instruire la faisabilité de l'application

Problématique

- Base de données open source Open Food Facts: https://world.openfoodfacts.org/data
- Association à but non lucratif
- Participation sur la base du volontariat
- Types de variables
  - Nutritionnelles: nutriments, minéraux, vitamines, …
  - Qualitatives: nutriscore, nova group, …
  - Divers: lieu d'origine, empreinte carbone, …
  - Métadonnées

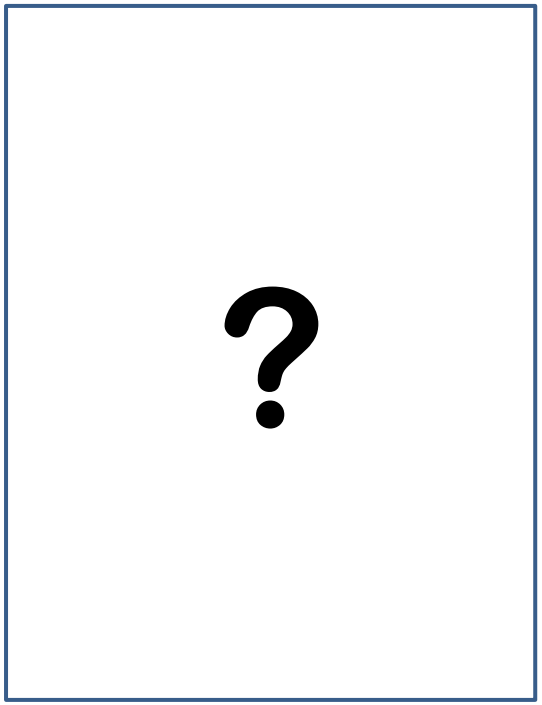
Présentation du concept d'application

- Critères de sélection des variables:
  - Variables d'identification
    - Code
    - URL
    - Nom produit
  - Variables catégorielles
    - Nutrsicore
    - Nova group
    - Catégories d'aliment
  - Variables nutritionnelles
    - Taux de remplissage > 70% sauf exceptions:
      - Alcool
      - Fibres



Présentation du concept d'application

- Input: scan du code barre
- L'application propose un produit similaire avec un taux de glucides plus faible.



Nettoyage

- Suppression des doublons (code unique par produit) ➔ 715 doublons
- Suppression des lignes vides
- Suppression des valeurs aberrantes:
  - 0 ≤ Masse d'une variable ≤ 100 g
  - Somme macronutriments (lipides, glucides, protéines, alcool) ≤ 100 g
  - Somme (sucres, fibres, lipides, glucides, protéines, alcool, sel) ≤ 100 g
  - Sucres ≤ Glucides
  - Gras saturé ≤ Lipides
  - Energie recalculée = 4 * glucides + 4 * protéines + 9 * lipides + 7 * alcool

    Energie recalculée ≈ énergie renseignée (erreur acceptée 10%)

➔ ***158 colonnes supprimées***

➔ ***500 000 lignes supprimées***

Analyse pré-exploratoire

- Provenance des données

	countries_en
France	677031
United States	333949
Spain	177376
Belgium	42890
Germany	41057
Switzerland	39327
United Kingdom	28595
Canada	20252
Italy	15045
France,Germany	11064

	brands
Carrefour	13412
Auchan	11014
U	6057
Bonarea	5638
Delhaize	4873
Hacendado	4682
Casino	4435
Nestlé	4354
Leader Price	4126
Cora	3385

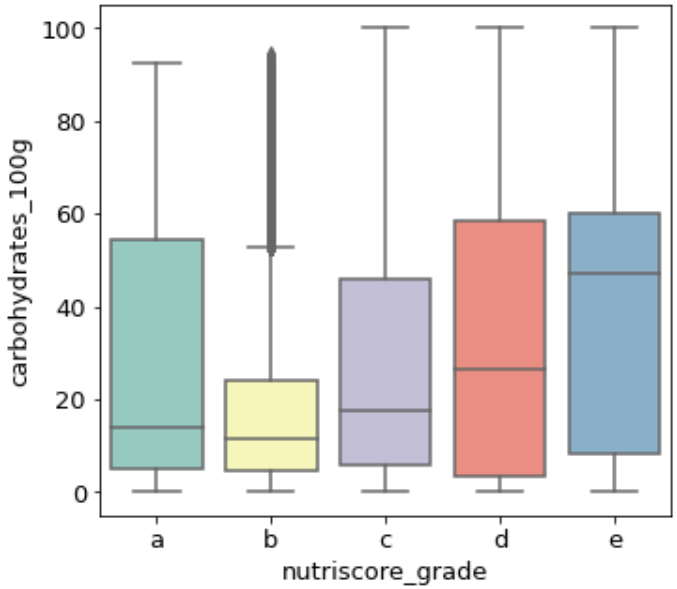
➔ *Principalement des produits français ou états-uniens.*

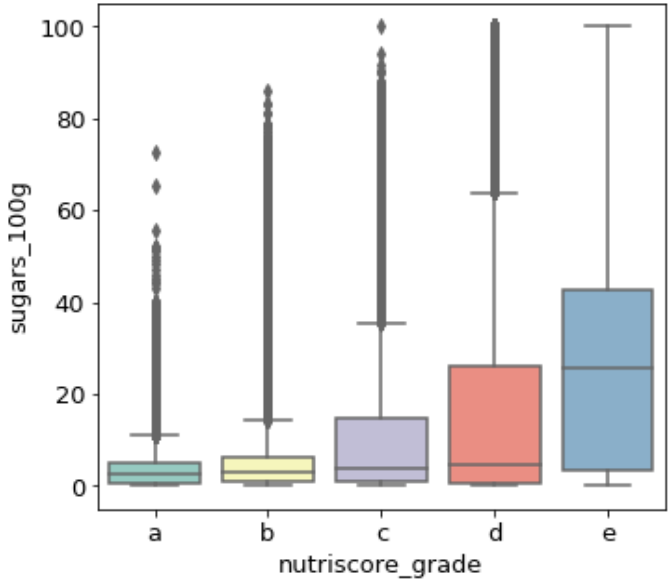
ANOVA

Analyse de la variance. Comparaison des variances inter et intra-catégorie.

Variables étudiées:

- Glucides
- Sucres





➔ ***Le nutriscore a-t-il un impact sur les variables étudiées?***

ANOVA (Analysis of Variance)

- Modèle:

$$y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$

Avec:

$y_{i,j}$: quantité de sucre du produit *j* dans le groupe de nutriscore i

μ : quantité de sucre moyenne

α_i : terme dépendant uniquement du nutriscore i

$\varepsilon_{i,j}$: variable indépendante de loi N(0,σ2)

- Hypothèse nulle: les moyennes inter-classes sont les mêmes

$$H_0 : \ \alpha_A = \cdots = \alpha_D = 0$$

- Test F

F = variance des moyennes inter-classes / variance intra-classe

H_0 équivalent à F ≈ 1

ANOVA (Analysis of Variance)

- Hypothèse nulle: les moyennes inter-classes sont les mêmes
- F = variance des moyennes inter-classes / variance intra-classe

Comparaison ANOVA Glucides / Sucres:

OLS Regression Results

Dep. Variable:	carbohydrates_100g	R-squared:	0.040
Model:	OLS	Adj. R-squared:	0.040
Method:	Least Squares	F-statistic:	5624.
Date:	Thu, 11 Feb 2021	Prob (F-statistic):	0.00
Time:	17:39:00	Log-Likelihood:	-2.5217e+06
No. Observations:	534574	AIC:	5.043e+06
Df Residuals:	534569	BIC:	5.043e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	27.0926	0.098	275.254	0.000	26.900	27.285
nutriscore_grade[T.b]	-7.4050	0.145	-51.164	0.000	-7.689	-7.121
nutriscore_grade[T.c]	-0.6435	0.128	-5.042	0.000	-0.894	-0.393
nutriscore_grade[T.d]	5.6798	0.118	48.204	0.000	5.449	5.911
nutriscore_grade[T.e]	10.4591	0.128	81.552	0.000	10.208	10.710

Omnibus:	90984.878	Durbin-Watson:	0.678
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42381.640
Skew:	0.528	Prob(JB):	0.00
Kurtosis:	2.113	Cond. No.	6.95

OLS Regression Results

Dep. Variable:	sugars_100g	R-squared:	0.147
Model:	OLS	Adj. R-squared:	0.147
Method:	Least Squares	F-statistic:	2.297e+04
Date:	Thu, 11 Feb 2021	Prob (F-statistic):	0.00
Time:	17:39:04	Log-Likelihood:	-2.2877e+06
No. Observations:	534162	AIC:	4.575e+06
Df Residuals:	534157	BIC:	4.576e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.9438	0.064	61.694	0.000	3.819	4.069
nutriscore_grade[T.b]	1.3140	0.094	13.999	0.000	1.130	1.498
nutriscore_grade[T.c]	7.0426	0.083	85.055	0.000	6.880	7.205
nutriscore_grade[T.d]	12.6587	0.076	165.554	0.000	12.509	12.809
nutriscore_grade[T.e]	21.2029	0.083	254.839	0.000	21.040	21.366

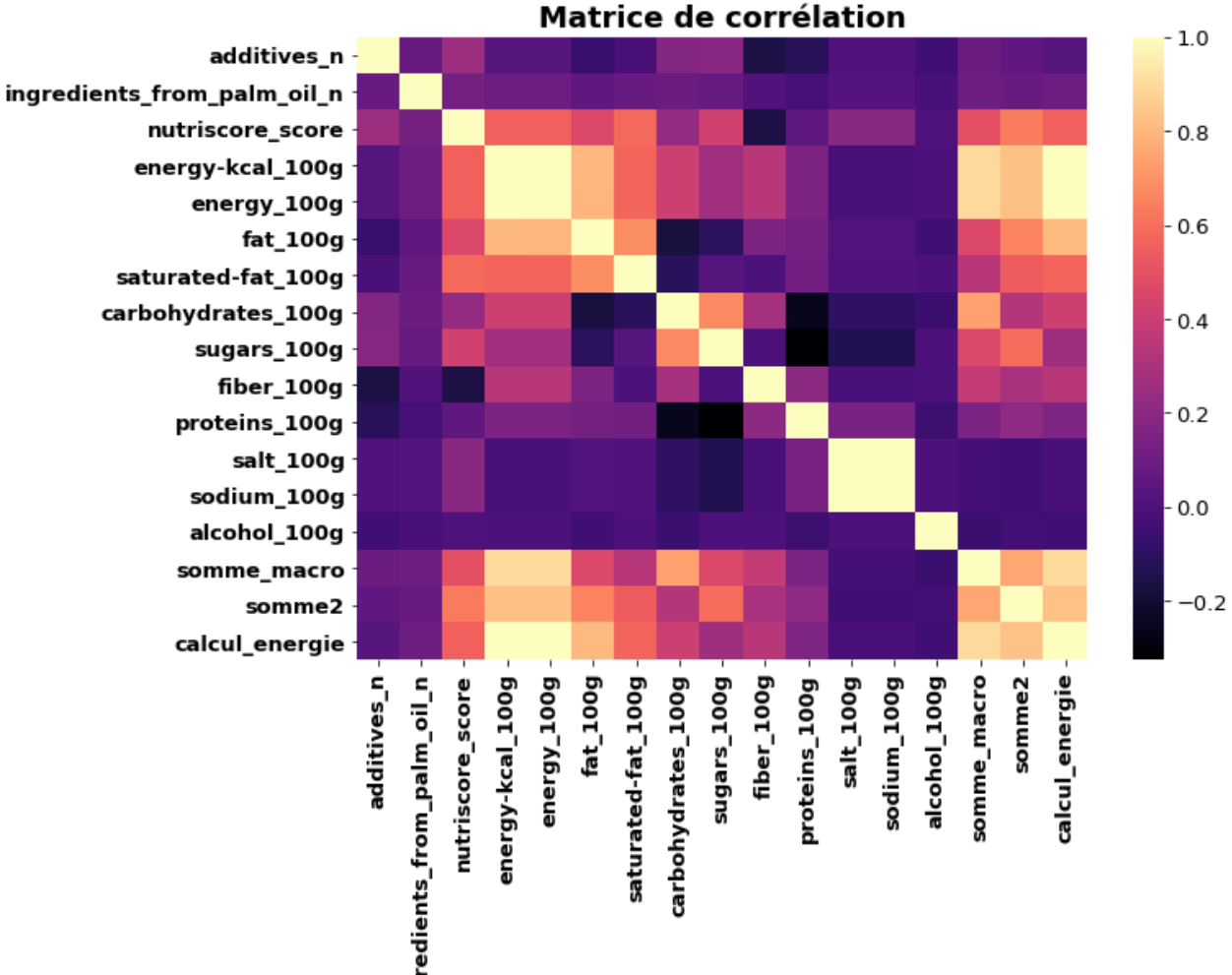
Omnibus:	150066.258	Durbin-Watson:	0.792
Prob(Omnibus):	0.000	Jarque-Bera (JB):	405966.126
Skew:	1.512	Prob(JB):	0.00
Kurtosis:	6.016	Cond. No.	6.97

➔ ***Le nutriscore a un impact sur la quantité de glucides***

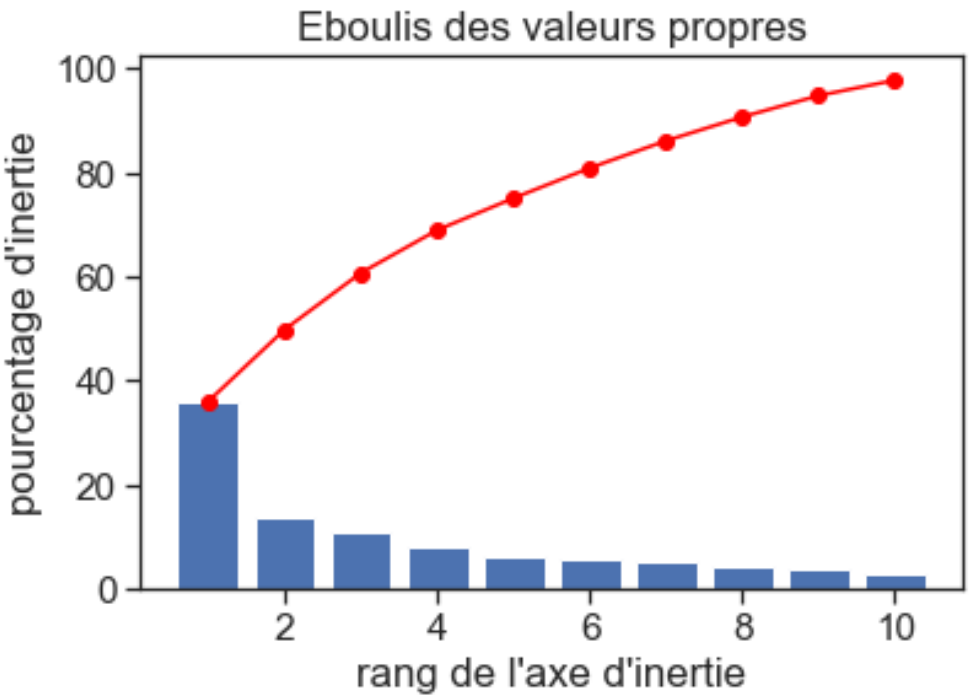
➔ ***Le nutriscore a un impact sur la quantité de sucres***

ACP (Analyse en Composantes Principales)

- Nos variables sont-elles linéairement corrélées ?



ACP (Analyse en Composantes Principales)

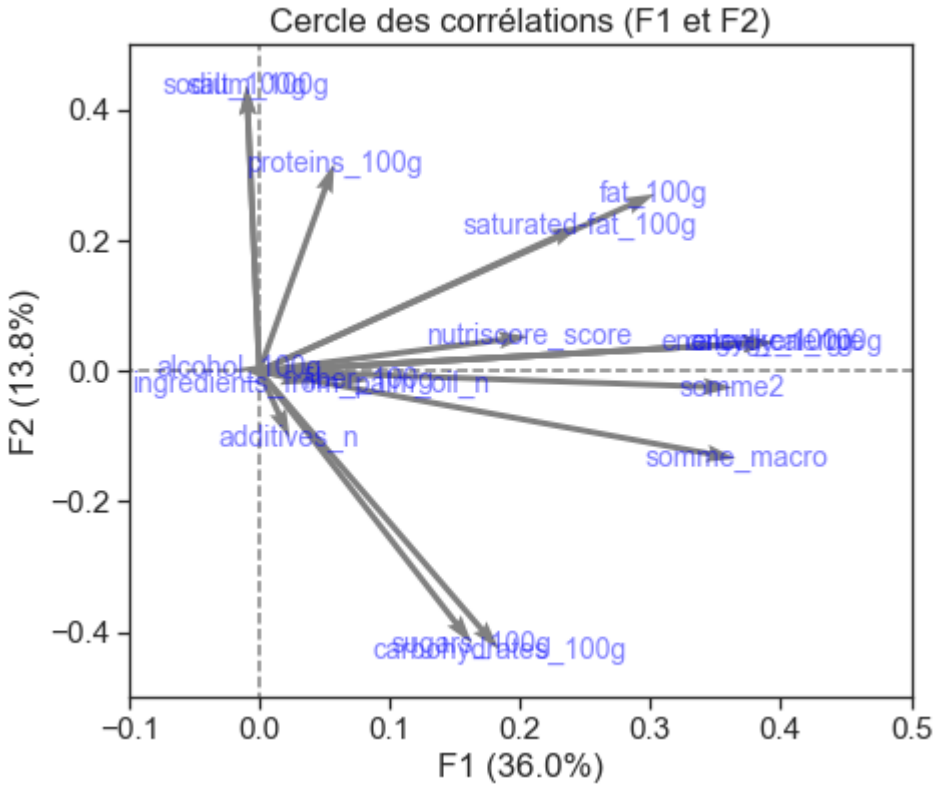


- 6 axes d'inertie permettent de représenter correctement 80% des données.

➔ ***On passe de 17 dimensions à seulement 6 dimensions.***

ACP (Analyse en Composantes Principales)

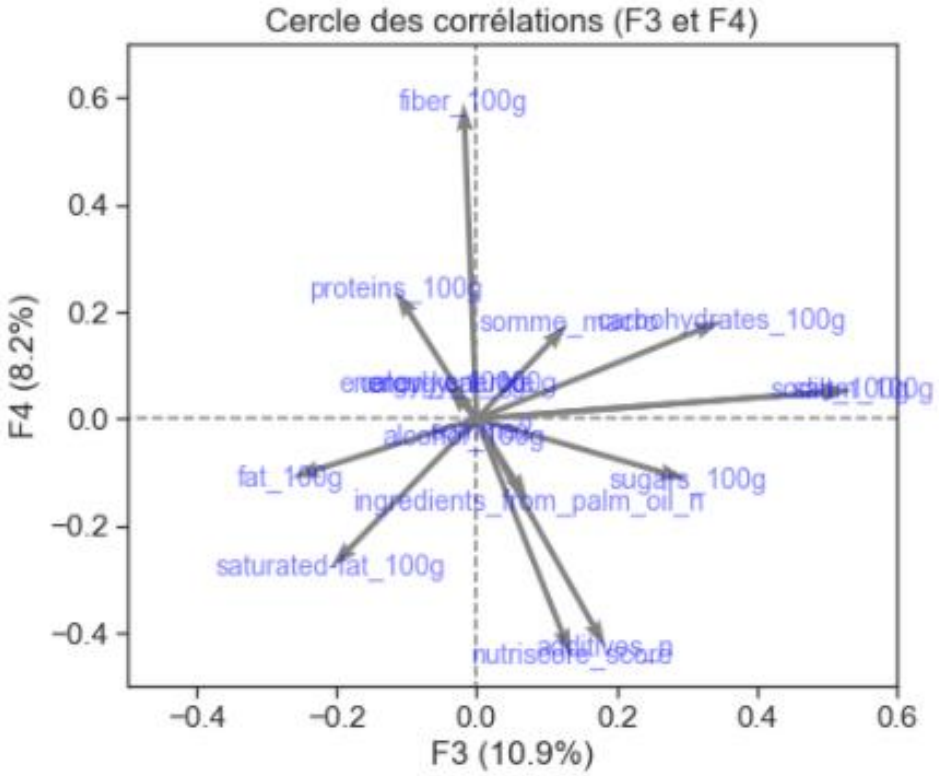
- Plan factoriel 1



- F1 ~ énergie
- F2 ~ sel

ACP (Analyse en Composantes Principales)

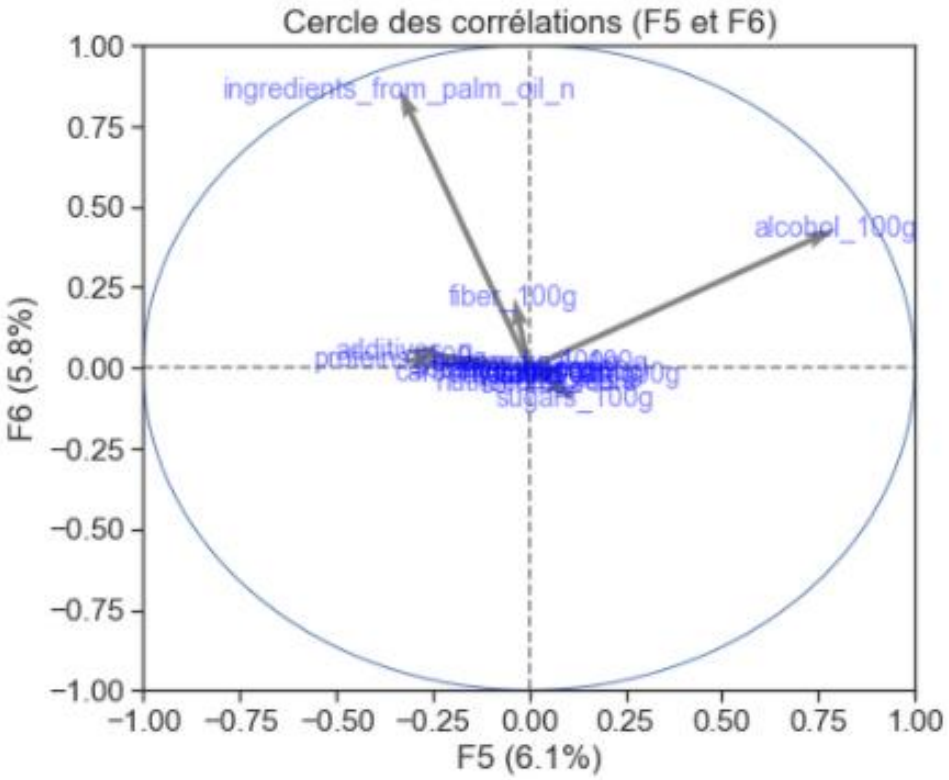
- Plan factoriel 2



- F3 ~ ?
- F4 ~ fibres vs nutriscore

ACP (Analyse en Composantes Principales)

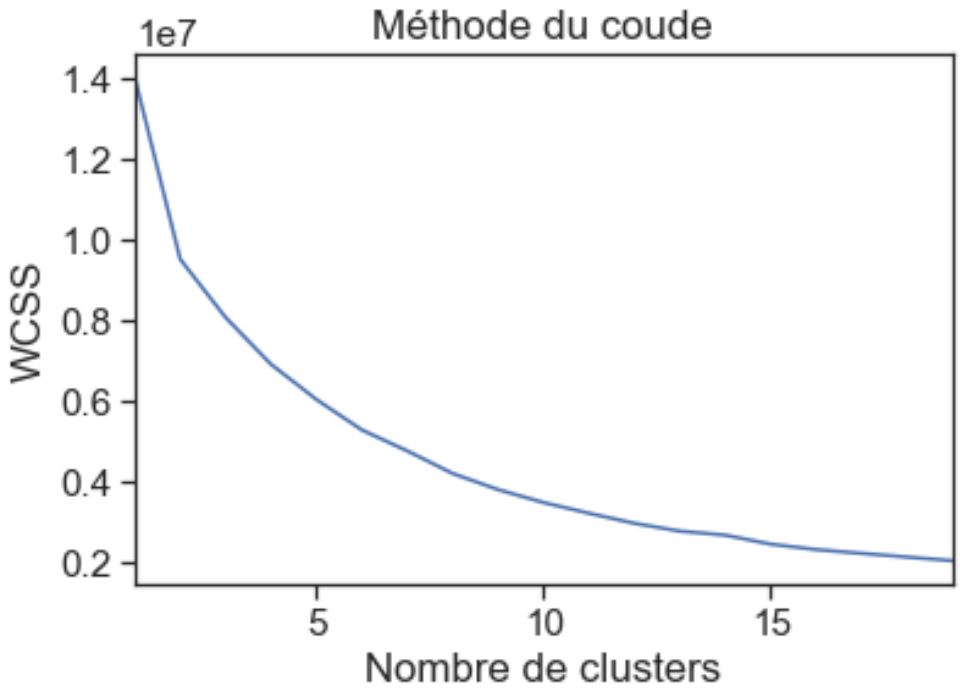
- Plan factoriel 3



- F5 ~ alcool
- F6 ~ nombre d'ingrédients huile de palme

K-Means

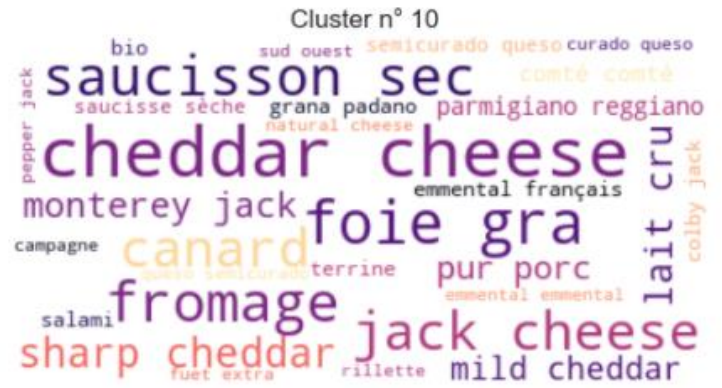
- Algorithme de clustering



- ***La méthode du coude ne permet pas d'identifier un nombre de clusters optimal.***
- ***Choix: n = 20***

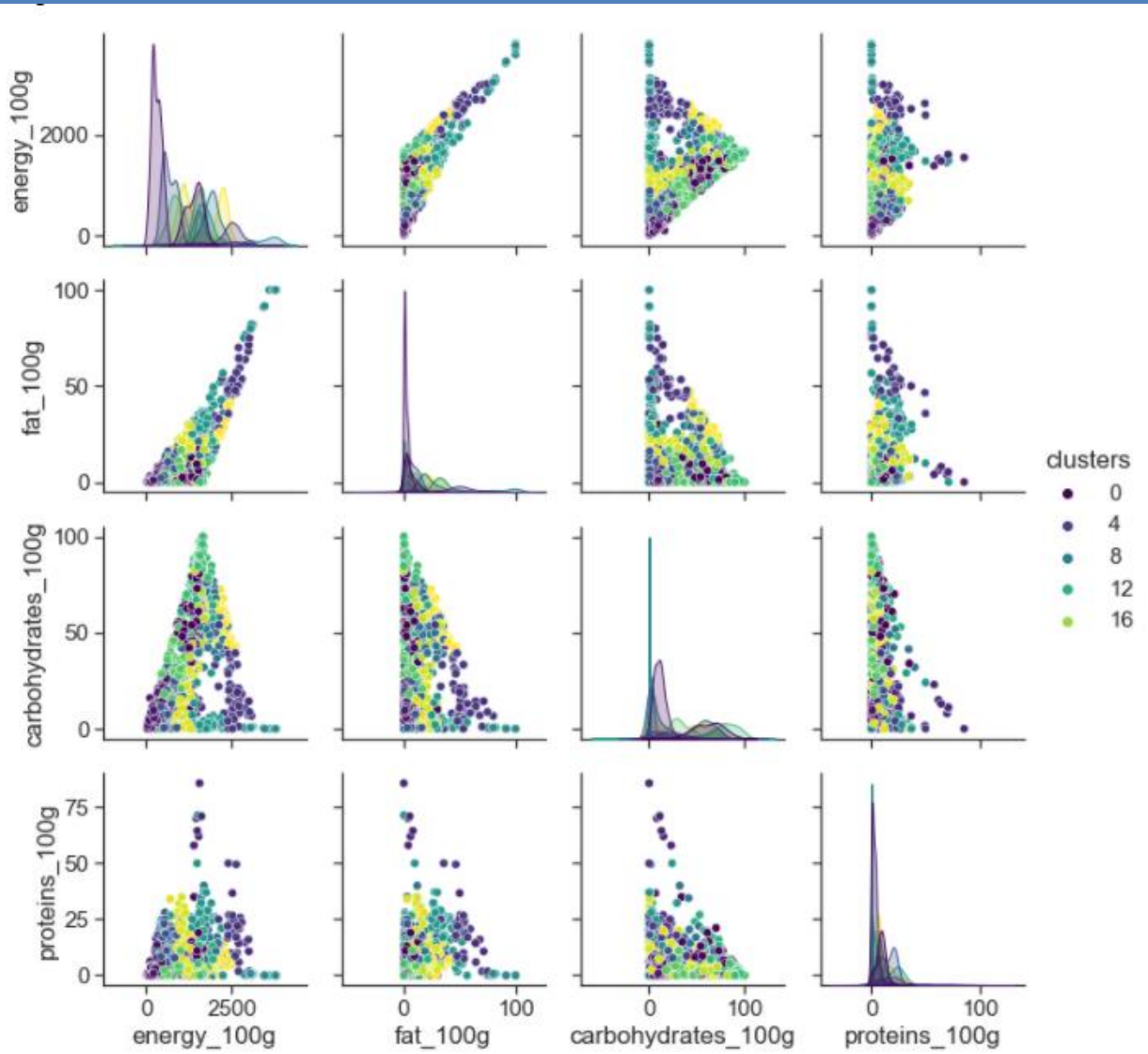
K-Means

- Exemples de nuages de mots par cluster



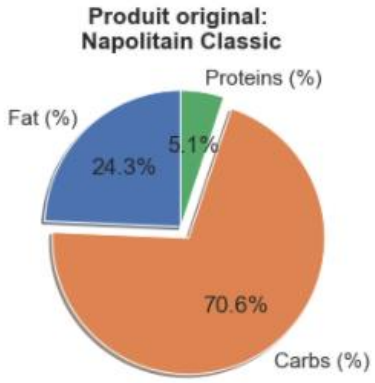
K-Means

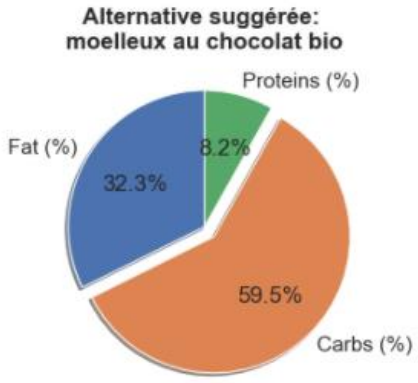
Echantillon:
10 000 produits



Démonstration du concept

- Input = code du produit
- Output = alternative moins riche en glucides + comparatif





En choisissant moelleux au chocolat bio à la place de Napolitain Classic vous économisez 42.0 grammes de glucides pour 100g de produit
Catégorie: Chocolate cakes

Conclusion

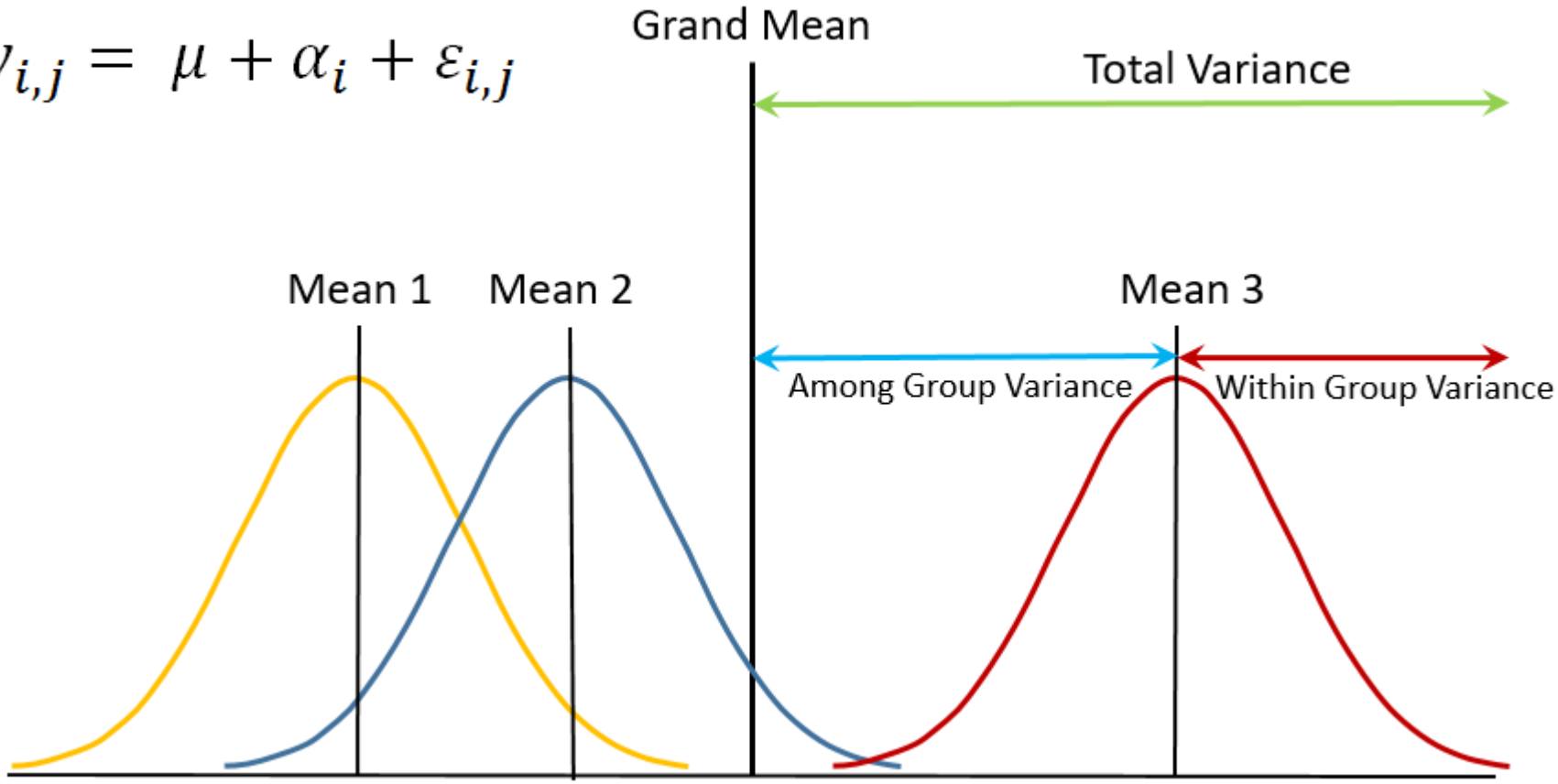
- Entrées sur la base du volontariat → beaucoup d'erreurs (un tiers du dataset!)
- ACP réduction efficace du nombre de dimensions
- Clustering efficace sur ce genre d'application
- L'étude du jeu de données open source permet de tirer des insights intéressants pour le consommateur

Merci de votre attention

Annexes

- ANOVA

$$y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$$



Annexes

- k-means