



Projet N°2
Concevez une application au service de la santé
publique

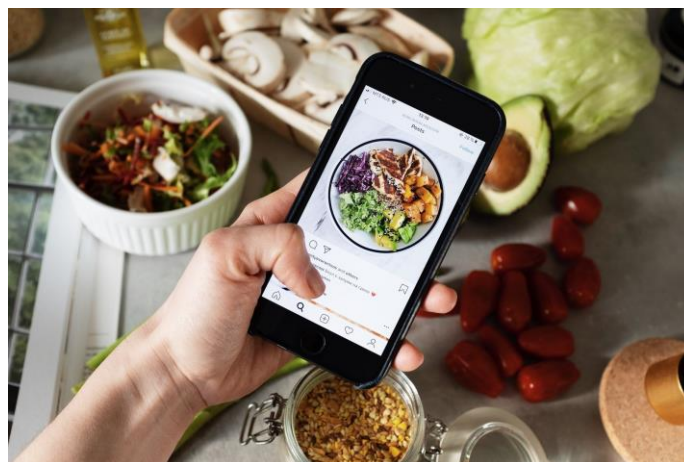
Plan

- But du projet
- Données
 - Description
 - Nettoyage
 - Enrichissement
 - Conclusion
- Analyse exploratoire
 - Univariée
 - Bivariée
 - Multivariée
- Conclusion

But du projet



- Une idée d'application en lien avec l'alimentation pour répondre à l'appel à projets lancé par l'agence "Santé publique France"



But du projet

- **L'idée :**

aider aux consommateurs de choisir des aliments non seulement bon pour leur santé mais aussi pour l'environnement



But du projet

- Comment choisit-on un aliment?



But du projet

Pâte à tartiner



Avec l'huile de
palme



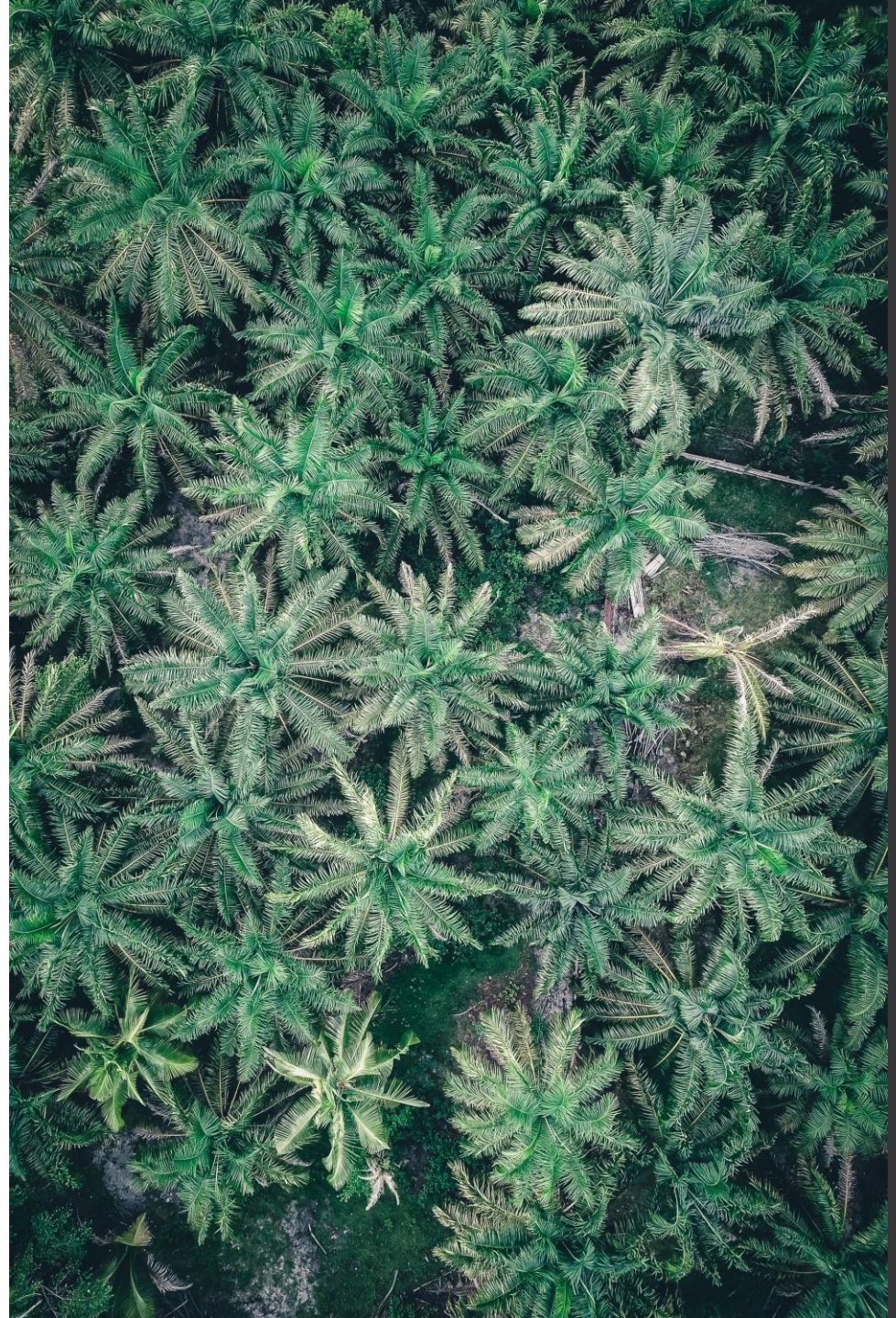
Sans l'huile de
palme



L'environnement, est-ce un argument suffisant?

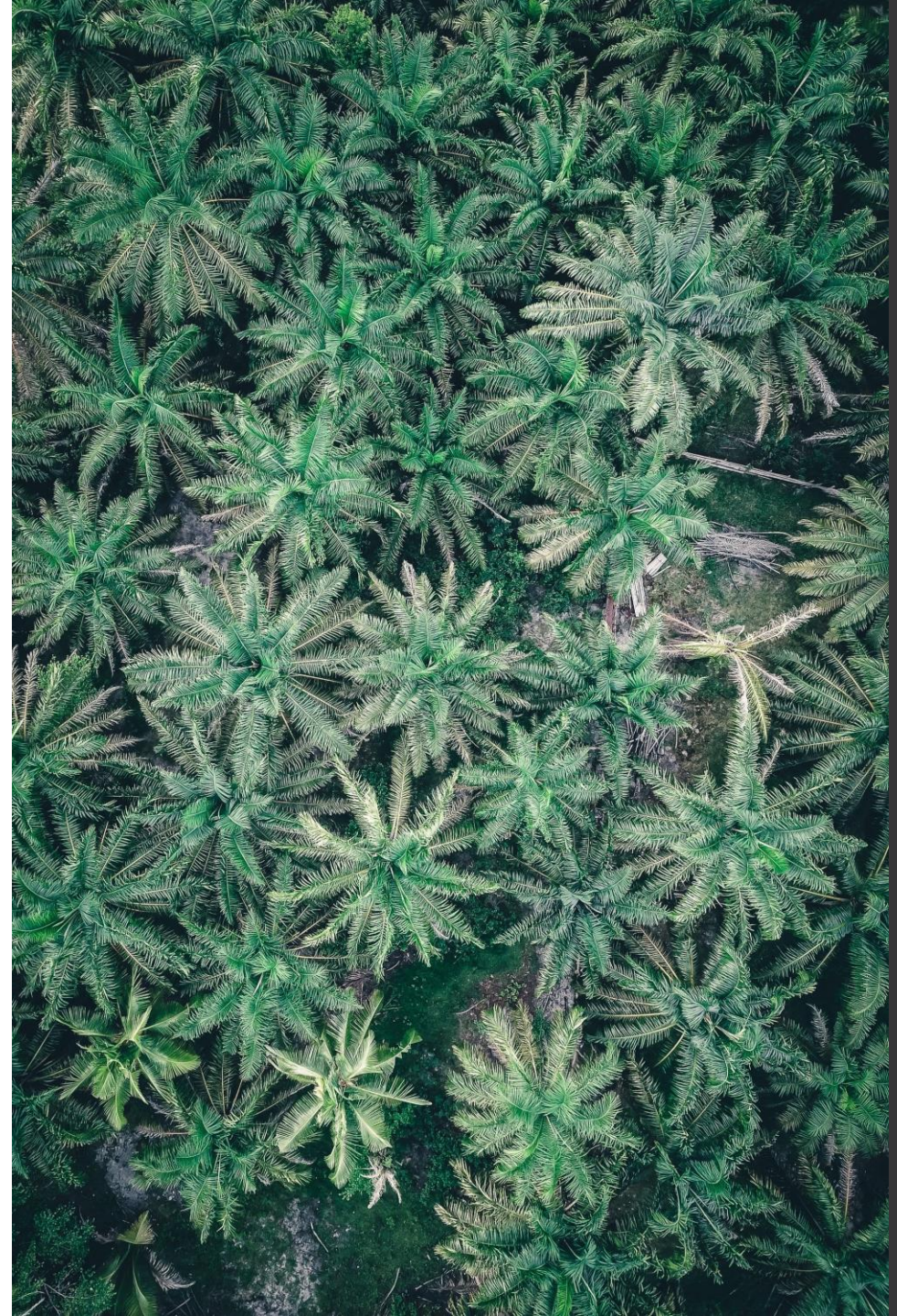
But du projet

- Selon l'Organisation pour la Coopération Economique et développement (Organisation for Economic Cooperation and Development (OECD)) l'huile de palme est l'huile végétale la plus consommée au monde (35%). ([source](#))
- l'Union Européenne est le 2ème importateur mondial d'huile de palme. ([source](#))



But du projet

- Des ONG accusent l'huile de palme être nocive pour :
 - la santé
 - l'environnement dans les pays producteurs ([source](#))
 - Déforestation
 - Émissions de CO2 lié aux changements des sols (Chris Malins « De l'huile sur le feu »)



But du projet

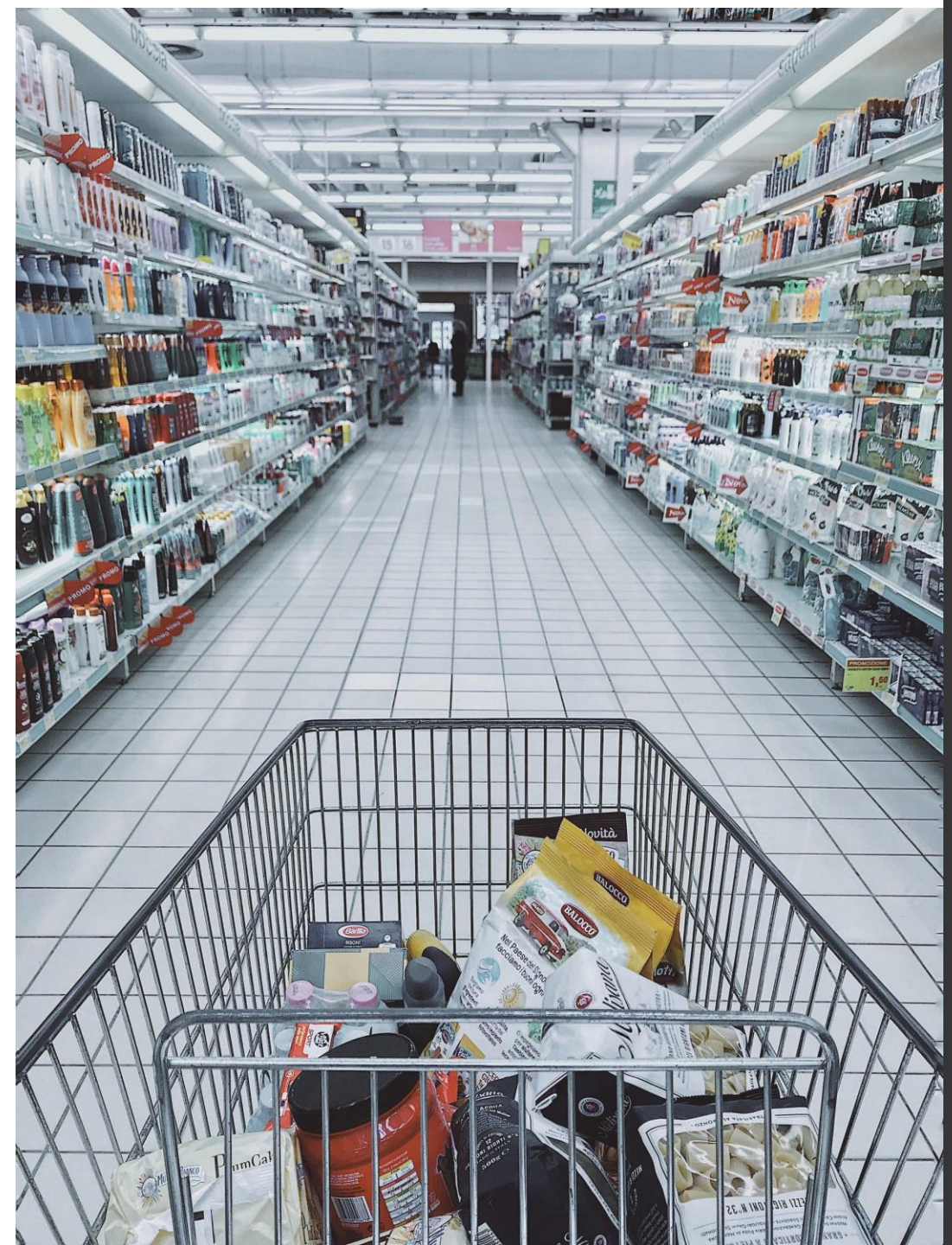
- But de l'application:
 - AVERTISSEMENT
- Mettre en garde le consommateur sur les groupes d'aliments qui peuvent contenir:
 - L'huile de palme **et**
 - Les gras saturés
 - Les additifs dangereux pour la santé



Données. Description

- Le jeu de données Open Food Fact

	Nombre
Lignes	1 555 491
Colonnes	183

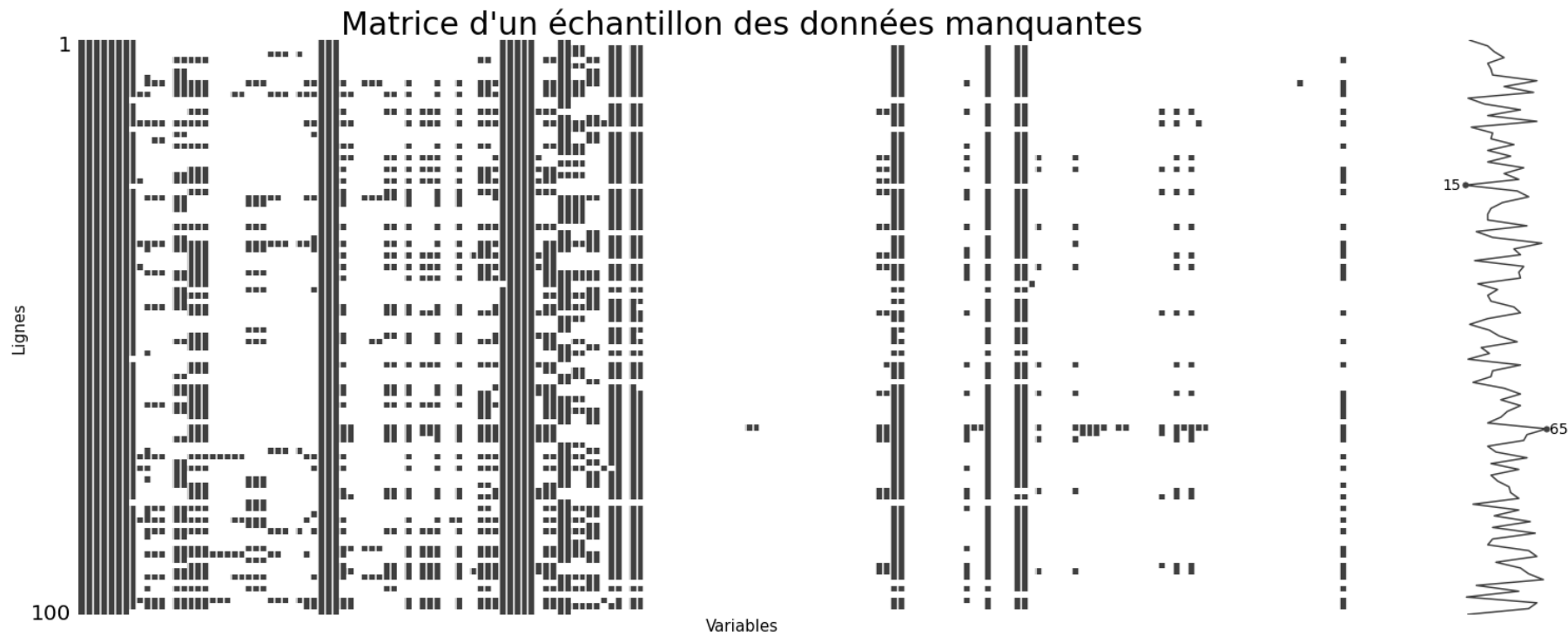


Données.

Description

- Exemple des premières dix variables:
 - 'code'
 - 'url'
 - 'creator'
 - 'created_t'
 - 'created_datetime'
 - 'last_modified_t'
 - 'last_modified_datetime',
 - 'product_name',
 - 'generic_name'
 - 'quantity'

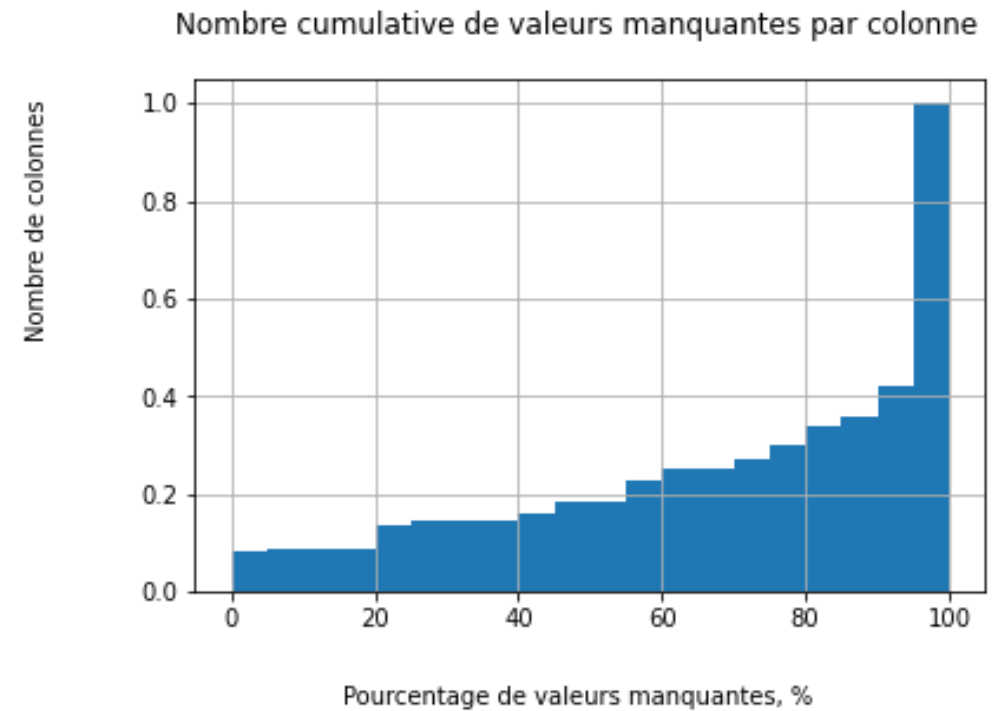
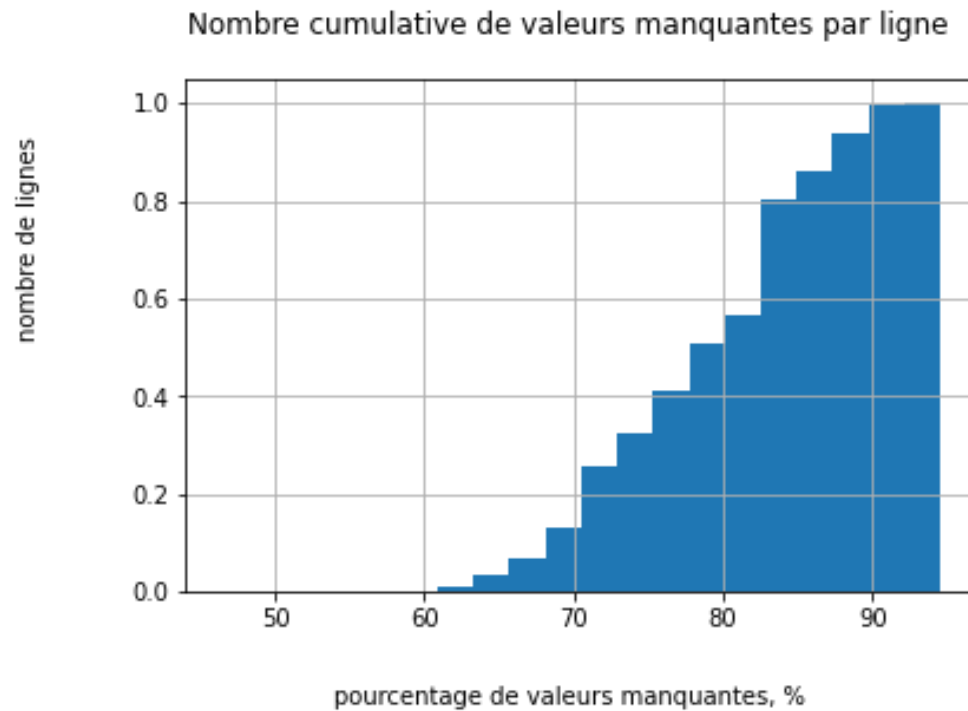
Données. Description



Données.

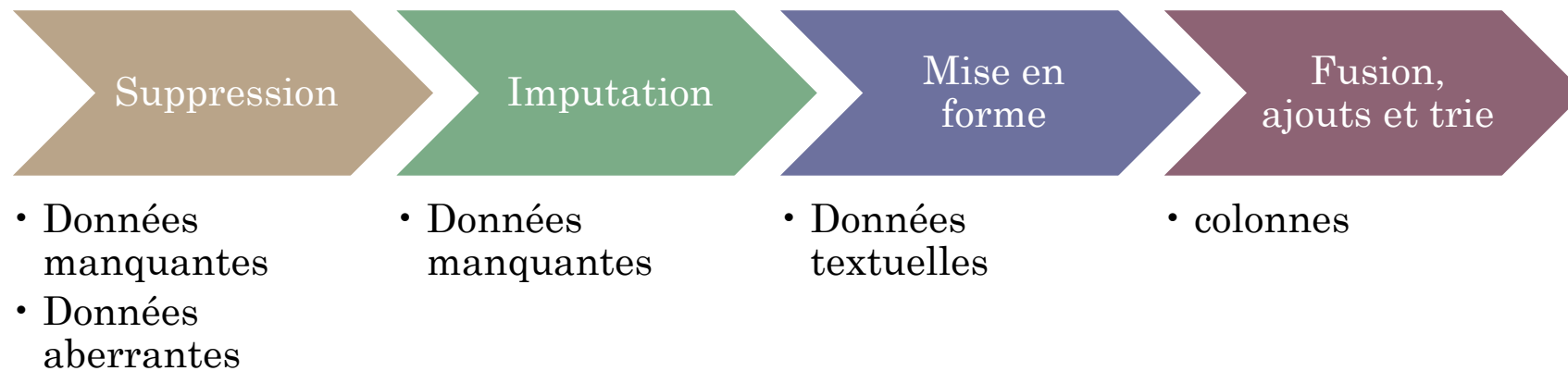
Description

- Les données manquantes:



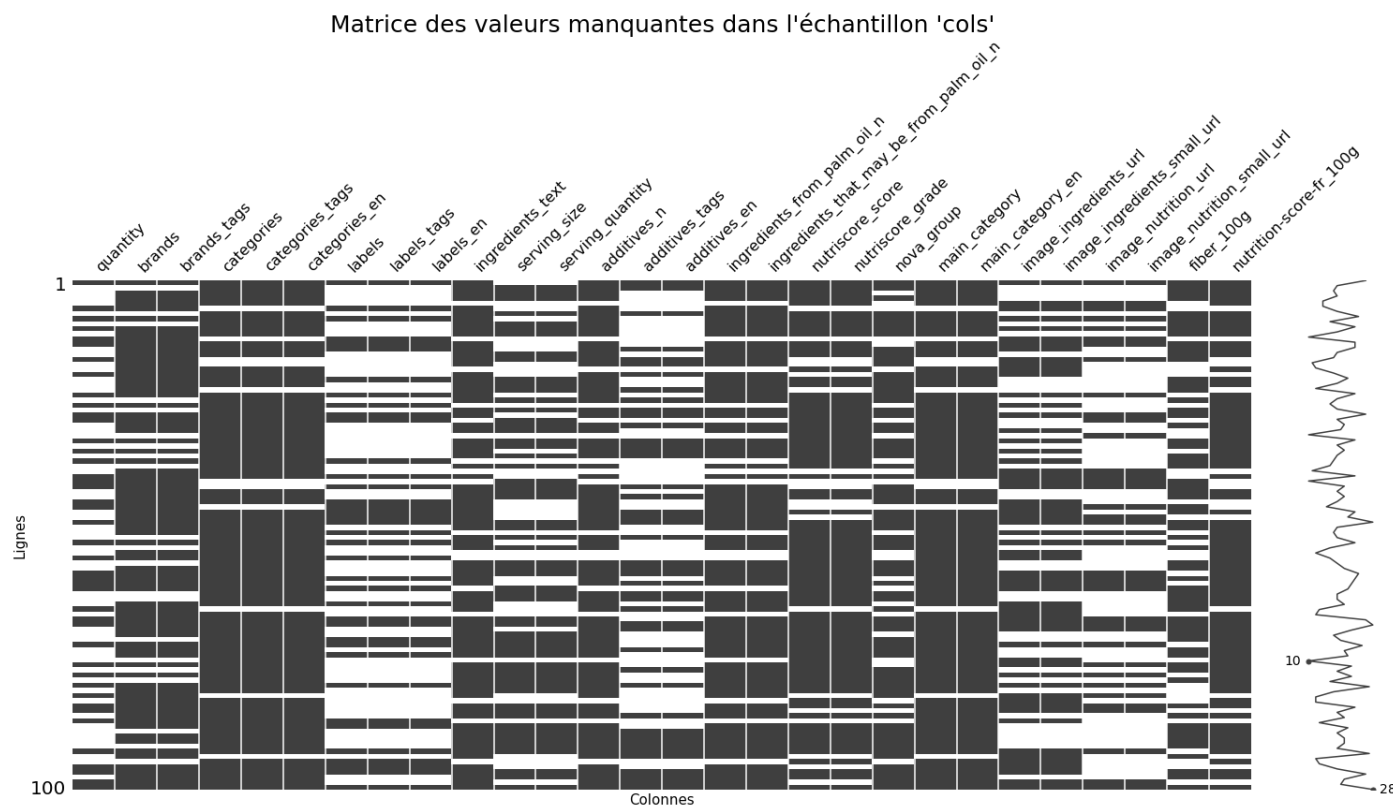
Données. Nettoyage

- Etapes



Données. Nettoyage

- Suppression de colonnes et lignes vide à 80%
- Zoom sur les données restantes contenant des données manquantes:



Données.

Nettoyage

- Stratégies de l'imputation des données manquante:
 - Mettre 0 ou médiane partout;
 - Par groupe nutritionnel;
 - Par groupe alimentaire;
 - Machine Learning: k plus proche voisins

Stratégie choisie: aucune, car besoin de connaissance métier

Données.

Nettoyage

- Suppression des valeurs aberrantes dues aux erreurs dans les données:
 - Quantités $> 100\text{g}$ et $< 0\text{g}$ par 100g
 - 'energy-kcal_100g' > 900 et < 0
 - Si 'sugars_100g' $>$ 'carbohydrates_100g'
 - Si 'saturated-fat_100g' $>$ 'fat_100g'

Données.

Nettoyage

- Nettoyage des catégories et sous-catégories alimentaires (PNNS1 et PNNS2):
 - toutes les lettres sont mis en minuscule;
Ex. Salty snacks, → salty snacks
 - le tiret est remplacé par un espace:
Ex. salty-snacks → salty snacks
 - remplacement de 'unknown' par NaN

catégorie	Nombre d'occurrences
unknown	208319
biscuits and cakes	54860
sweets	54277
dressings and sauces	41413

Données.

Nettoyage

- Fusion des colonnes 'ingredients_from_palm_oil_n' et 'ingredients_that_may_be_from_palm_oil_n' → 'ingredients_with_possible_presence_of_palm_oil_n'
- Ajout des colonnes supplémentaires:
 - présence ou absence de l'huile de palme dans un produit ('possibility_of_presence_of_palm_oil')
 - présence ou absence des additifs dans un produit ('presence_of_additives')
- Trie:
 - Choix de 19 colonnes pertinentes:
'product_name', 'ingredients_with_possible_presence_of_palm_oil_n',
'possibility_of_presence_of_palm_oil', 'pnns_groups_1', 'pnns_groups_2', 'additives_n',
'additives_en', 'presence_of_additives', 'nutriscore_score', 'nutriscore_grade', 'energy_kcal_100g', 'fat_100g', 'saturated-fat_100g', 'carbohydrates_100g', 'sugars_100g',
'fiber_100g', 'proteins_100g', 'salt_100g', 'nova_group'

Données. Enrichissement

- Importation des données contenant l'information sur la dangerosité des additifs. Source: [Liste des additifs par ordre de dangerosité. - Webadditifs \(les-additifs-alimentaires.com\)](http://les-additifs-alimentaires.com)






Additifs par ordre de dangerosité.











À quoi ça sert ?

Les additifs alimentaires sont des ingrédients ajoutés aux aliments afin d'en améliorer les qualités. Ils peuvent avoir de **nombreuses fonctions** différentes. Pour simplifier la recherche, je les ai classés en 9 familles qui correspondent aux fonctions les plus courantes.

Légende

Le niveau de danger estimé des additifs est marqué à l'aide des icones suivantes .

	Sans danger
	
	
	
	Dangereux

	Loi	Hallal	Casher	Végétarien	Végétalien	Danger	
	N°	Nom					Famille
	E100		Curcumine				Colorants Jaune
	E100i		Curcumine				Colorants Jaune
	E100ii		Curcuma				Colorants Jaune
	E101		Vitamine G				Colorants Jaune
	E101i		Riboflavin				Colorants Jaune

Données.

Enrichissement

- Création de colonnes: 'dangerous_additives_n' et 'zero_one_more_dangerous_additives'.

Extrait du dataframe:

	product_name	ingredients_with_possible_presence_of_palm_oil_n	possibility_of_presence_of_palm_oil	additives_n	additives_en	presence_of_additives	dangerous_additives_n	zero_one_more_dangerous_additives
784636	Sugar Free Drink Mix, Peach Tea	0.0	False	7.0	E102 - Tartrazine, E129 - Allura red ac, E150c - ...	True	3	2

Données.

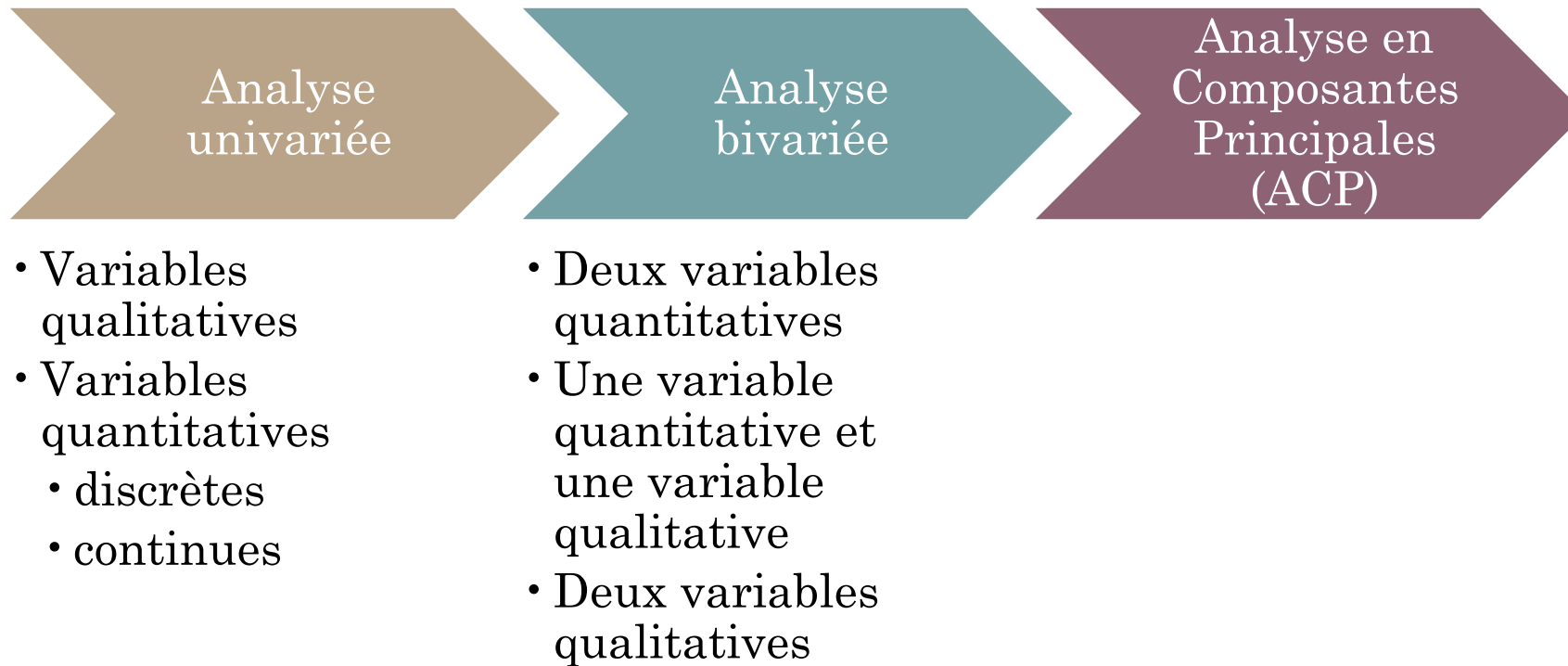
Conclusion

	Nombre dans le jeu de données d'origine	Nombre dans le jeu de données final	Part de données finales par rapport aux données d'origine, %
Lignes	1 555 491	784 643	50,44
Colonnes	183	22	12,02

Un jeu de données , à priori, représentatif du domaine des produits alimentaires et suffisamment grand pour être analysé.

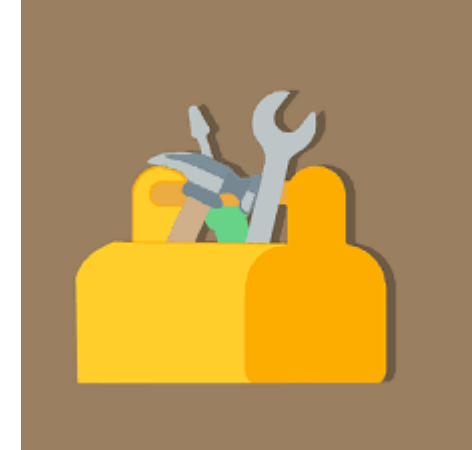
Analyse exploratoire

- Etapes:



Analyse exploratoire univariée.

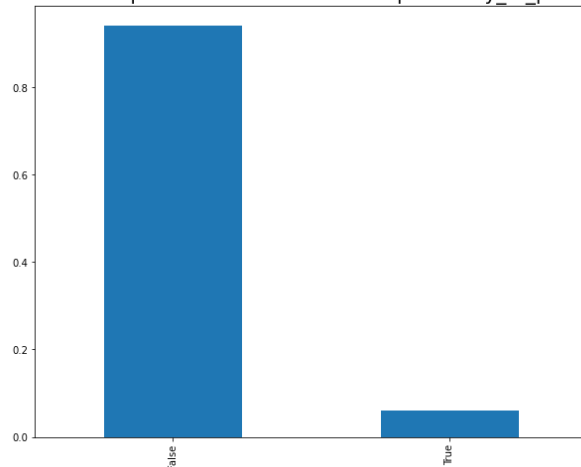
- Outils
 - Les techniques d'analyse utilisées:
 - Test de distribution normale
 - Mode
 - Variance
 - L'écart-type empirique corrigé
 - L'asymétrie: skewness empirique
 - Kurtosis empirique (une mesure d'aplatissement)
 - Les techniques de visualisation:
 - Bar-plot
 - Pie-plot
 - Nuage de mots
 - Histogramme
 - Diagrammes en boîtes



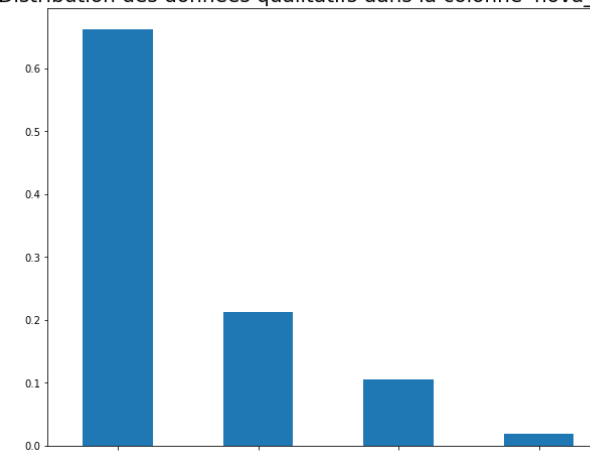
Analyse exploratoire univariée

Variables qualitatives

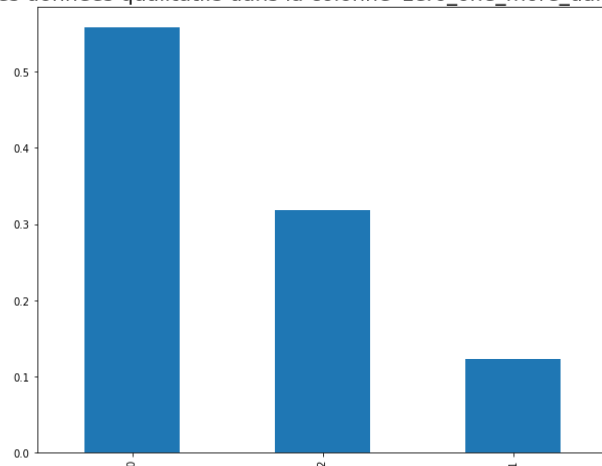
Distribution des données qualitatifs dans la colonne 'possibility_of_presence_of_palm_oil'



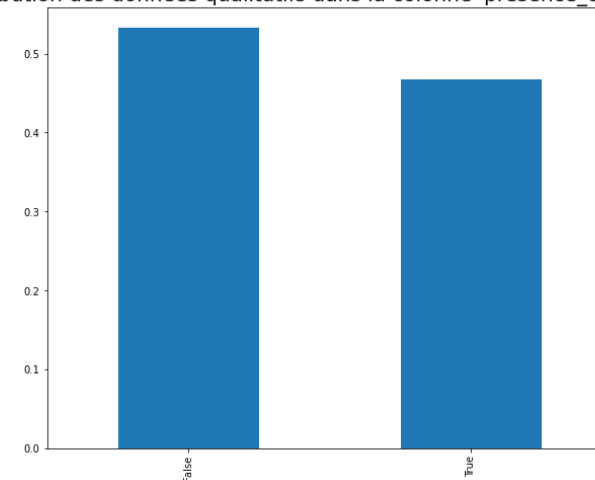
Distribution des données qualitatifs dans la colonne 'nova_group'



Distribution des données qualitatifs dans la colonne 'zero_one_more_dangerous_additives'



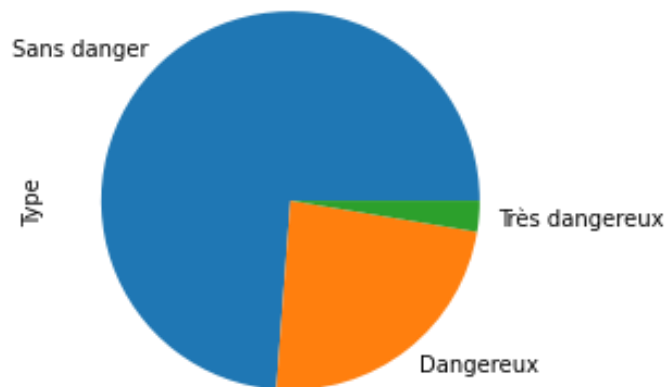
Distribution des données qualitatifs dans la colonne 'presence_of_additives'



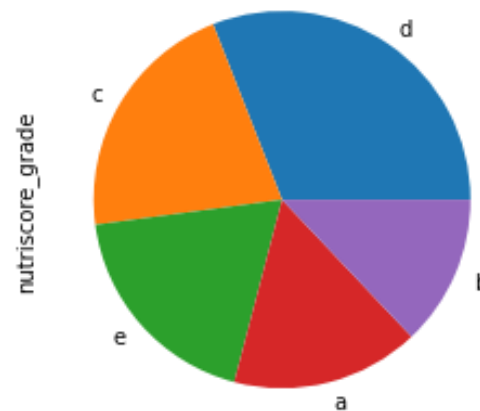
Analyse exploratoire univariée

Variables qualitatives

Distribution d'additifs par type



Distribution des données qualitatives dans la colonne 'nutriscore_grade'

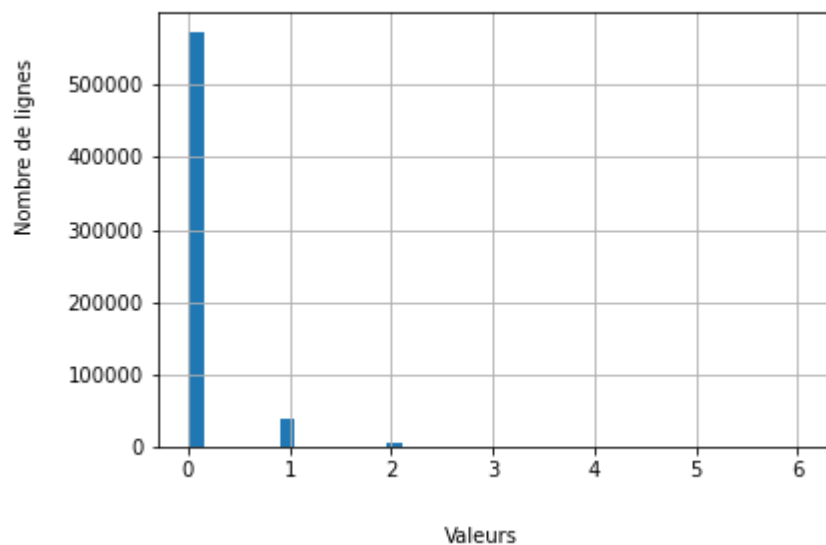


Analyse exploratoire univariée

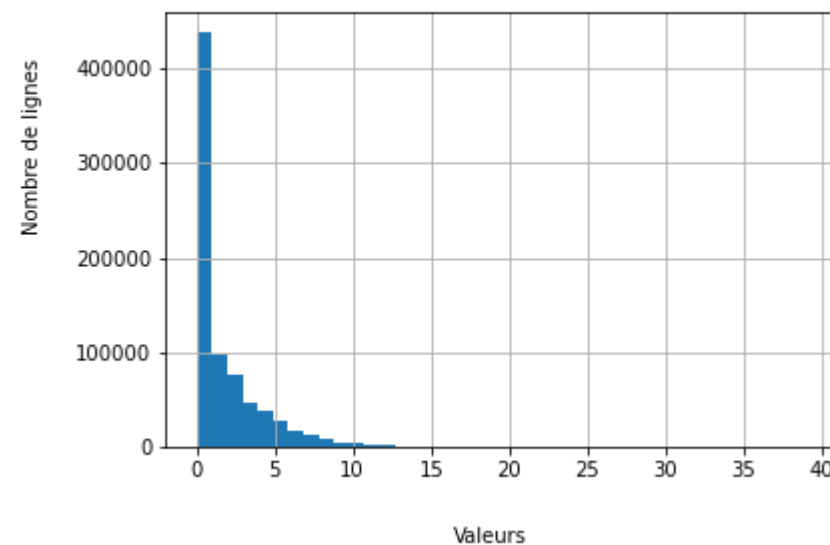
Variables quantitatives

- Variables discrètes

Nombre de valeurs par colonne 'ingredients_with_possible_presence_of_palm_oil_n'



Nombre de valeurs par colonne 'dangerous_additives_n'

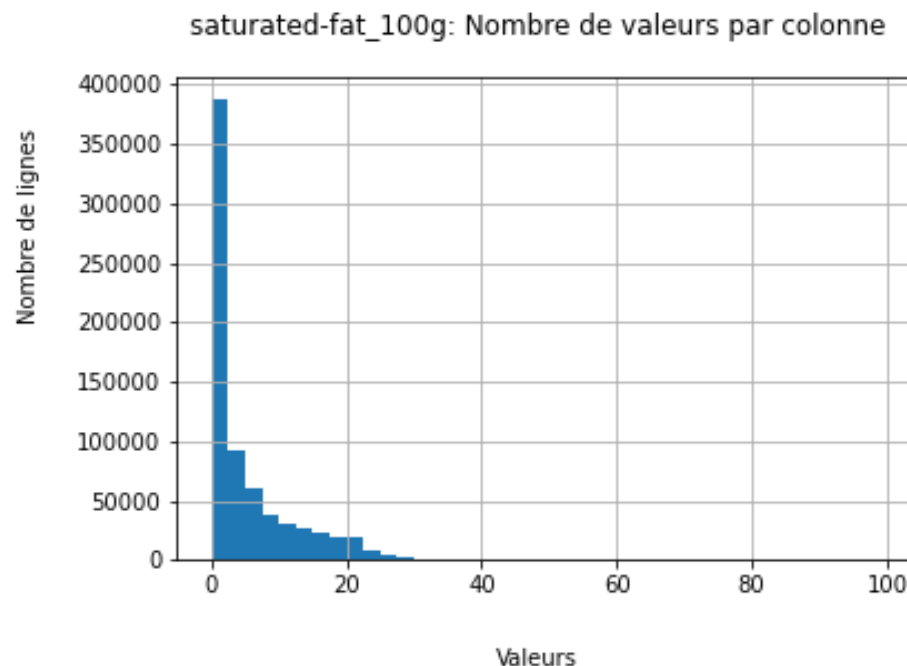


Analyse exploratoire univariée

Variables quantitatives

- Variables continues

Mesure	Valeur
Mode	0,0
Variance	61,4
L'écart-type	7,8
Skewness	3,3
Kurtosis	20,7



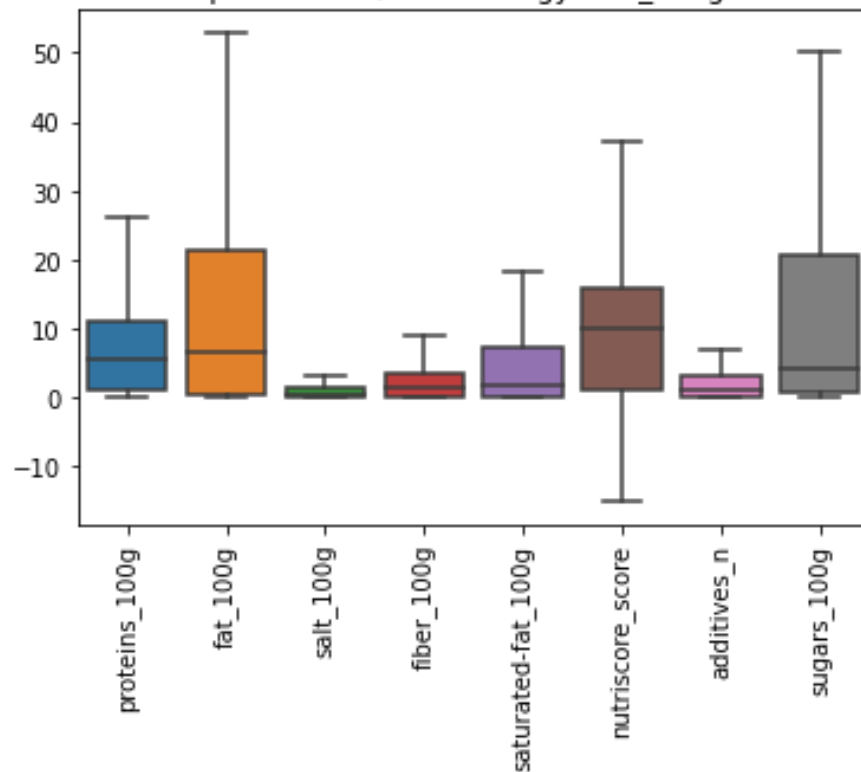
La distribution de tous les données d'éléments nutritifs s'étale vers la droite.

Analyse exploratoire univariée

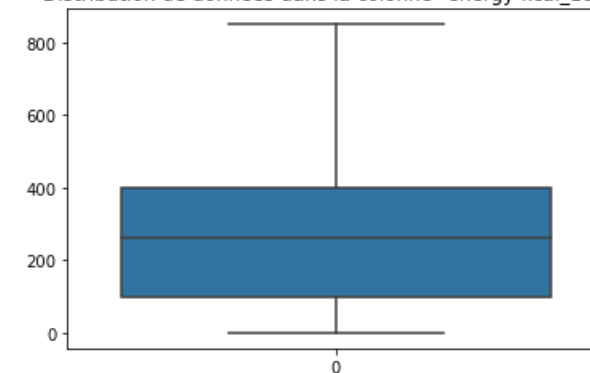
Variables quantitatives

Test de distribution normale : $p = 0$. L'hypothèse nulle peut être rejetée

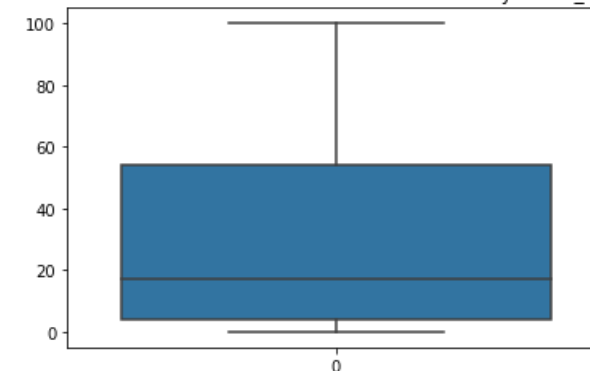
Distribution de données par colonne, sauf 'energy-kcal_100g' et 'carbohydrates_100g'



Distribution de données dans la colonne "energy-kcal_100g"



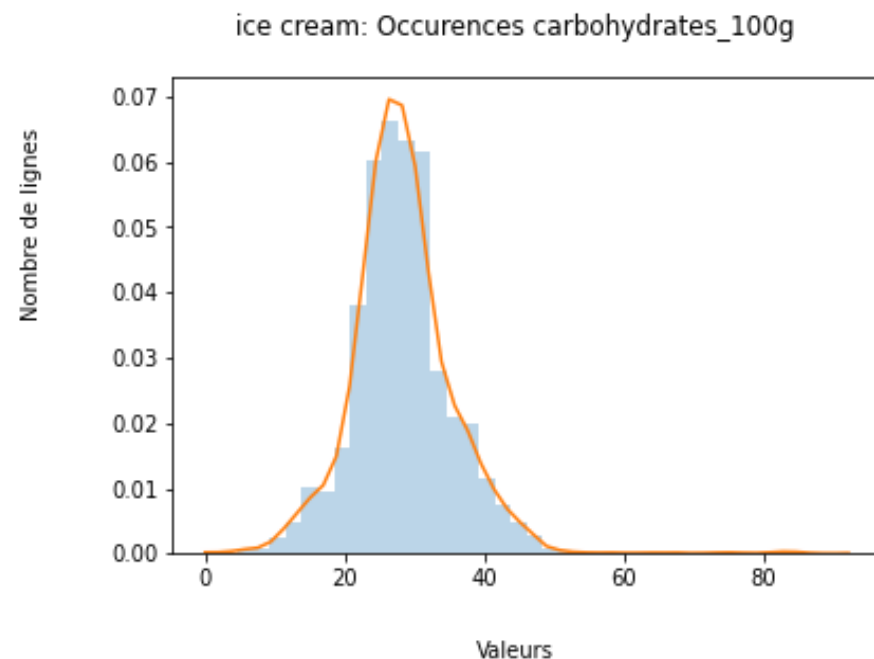
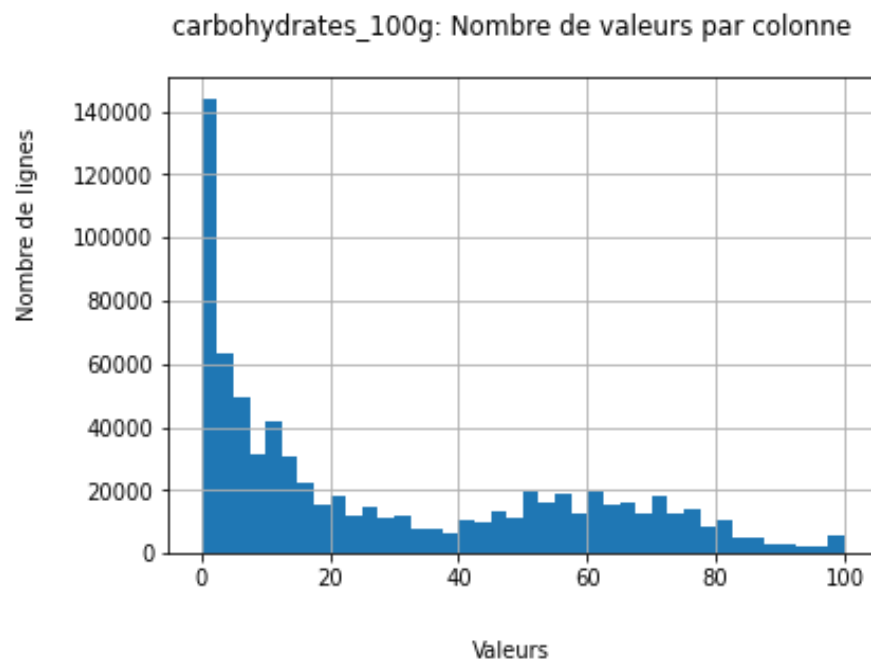
Distribution de données dans la colonne "carbohydrates_100g"



Analyse exploratoire univariée

Variables quantitatives

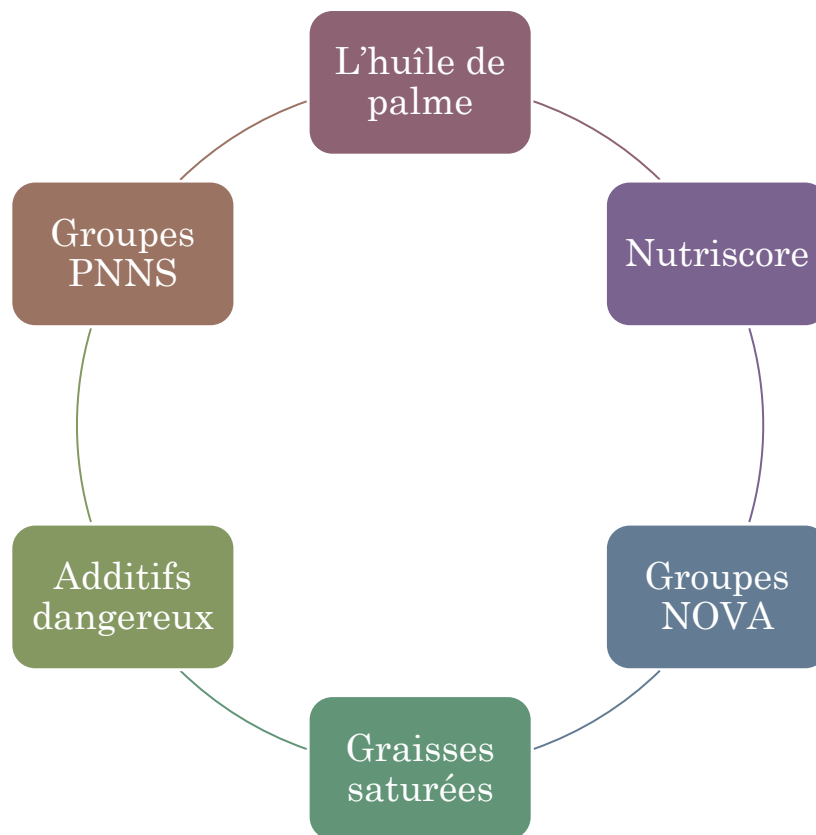
- Certaines variables contiennent plusieurs pics, ce qui peut signifier qu'il y a plusieurs sous-distributions, en fonction de catégorie des produits.



Analyse exploratoire bivariée.

Hypothèse

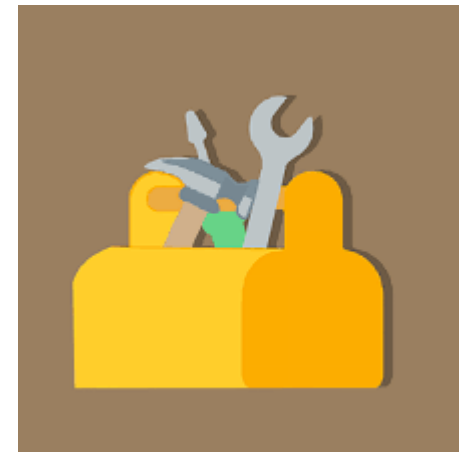
- Nous supposons qu'il existe un lien entre :



Analyse exploratoire bivariée.

Deux variables quantitatives

- Outils
 - Les techniques d'analyse utilisées:
 - Coefficient de corrélation de Pearson
 - Covariance empirique
 - Tableau de corrélation linéaire
 - Les techniques de visualisation:
 - Diagramme de dispersion
 - Heatmap
 - Pairplot

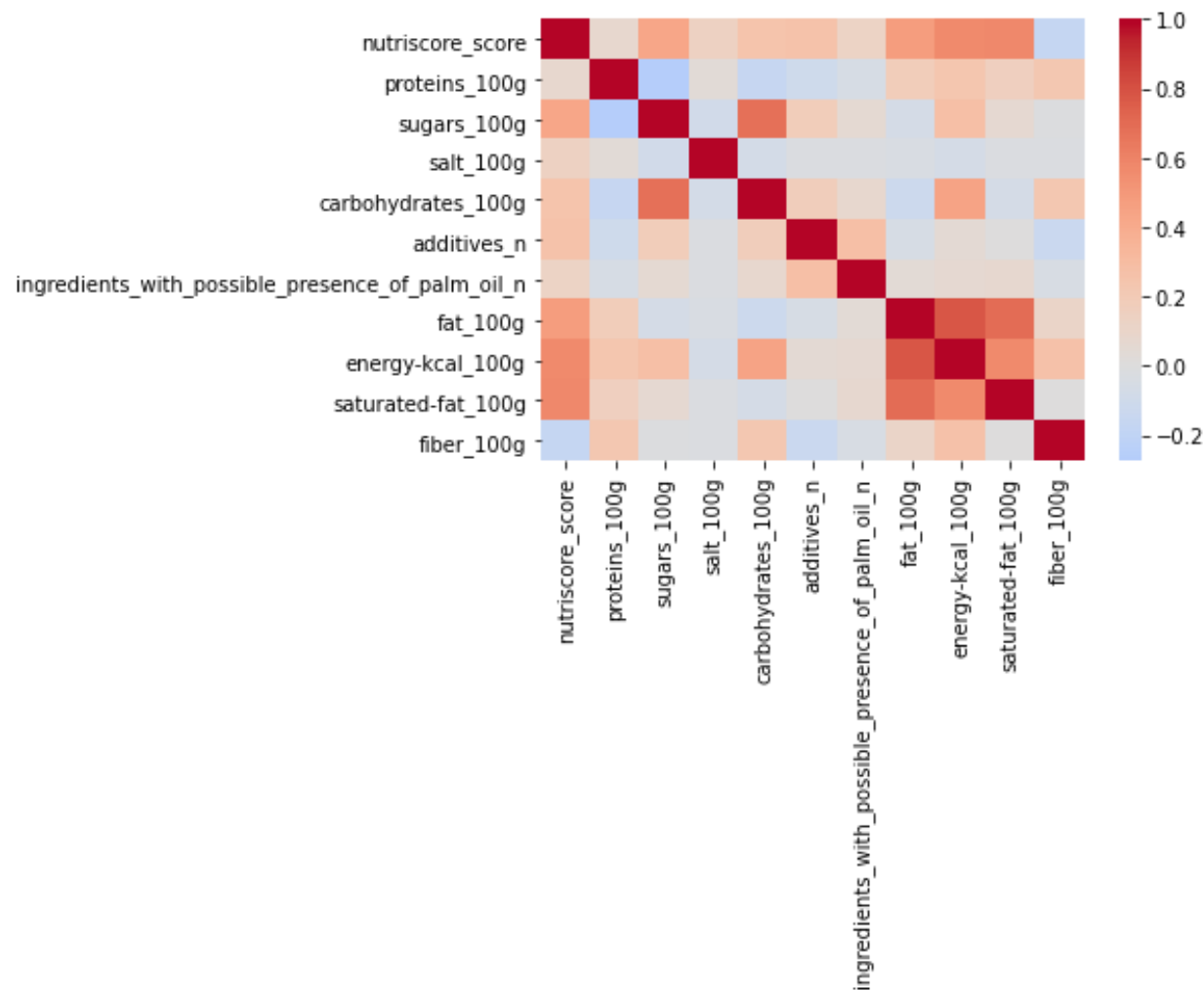


Analyse exploratoire bivariée.

Deux variables quantitatives

Résultats:

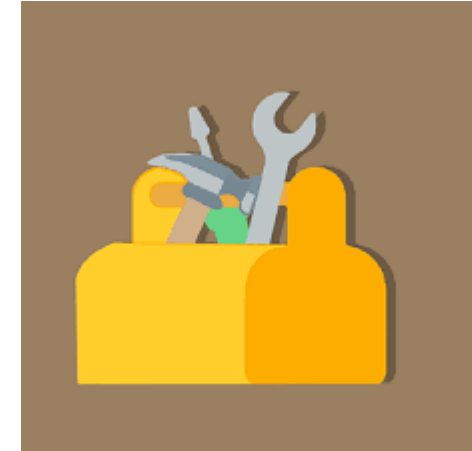
- Il existe un lien direct entre les gras saturés, les calories, les gras, les sucres raffinés et le nutriscore.
- Il n'y a pratiquement aucun lien entre le nutriscore et la présence d'huile de palme.



Analyse exploratoire bivariée.

Une variable quantitative et une variable qualitative

- Outils
 - Les techniques d'analyse utilisées:
 - Anova
 - Eta carré
 - Les techniques de visualisation:
 - Boxplot



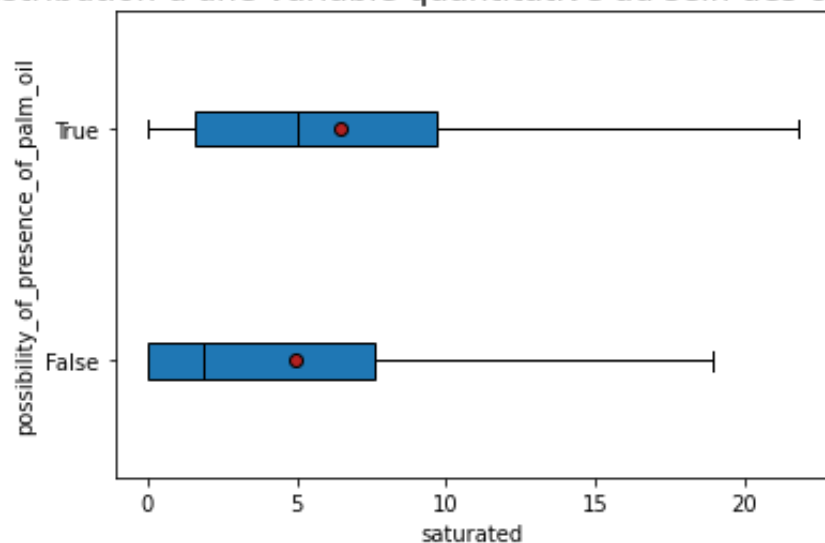
Analyse exploratoire bivariée.

Une variable quantitative et une variable qualitative

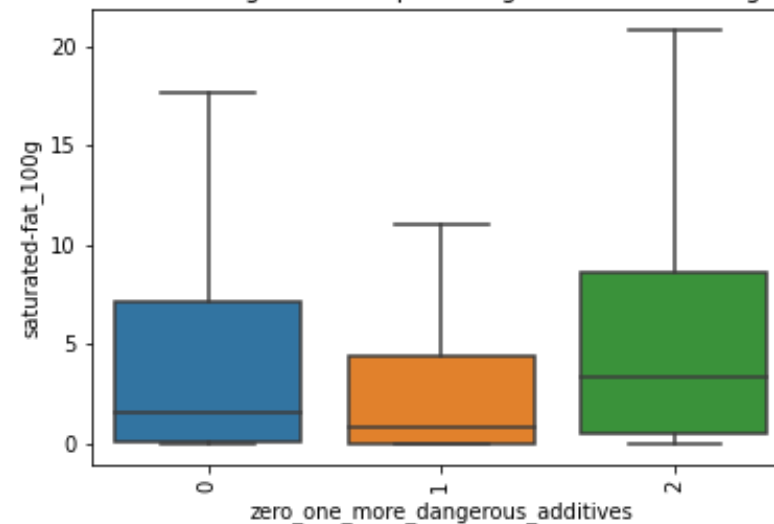
- **Résultats:**

- À priori, il n'existe pas de relation entre les graisses saturées et la possibilité de présence d'huile de palme ($p = 0,0$; η^2 (éta carré) = 0,005).
- La quantité d'additifs dangereux a également qu'un effet très faible sur la distribution des graisses saturées dans les aliments ($\eta^2 = 0,04$).

Distribution d'une variable quantitative au sein des catégories



Distribution de gras saturé par catégorie d'additifs dangereux

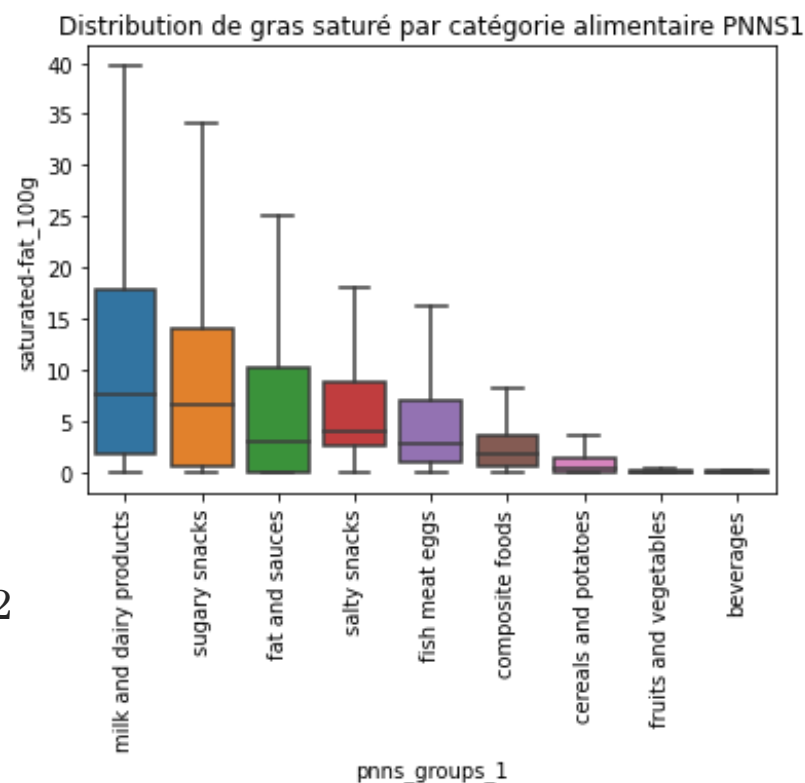


Analyse exploratoire bivariée.

Une variable quantitative et une variable qualitative

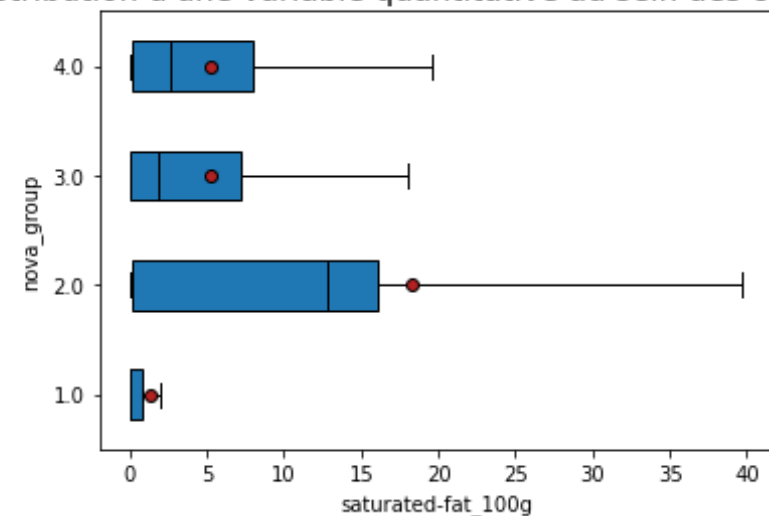
- **Résultats:**

- Par contre, les groupes d'aliments PNNS et NOVA ont un effet sur la distribution de gras saturé



$$\eta^2 = 0,22$$

Distribution d'une variable quantitative au sein des catégories

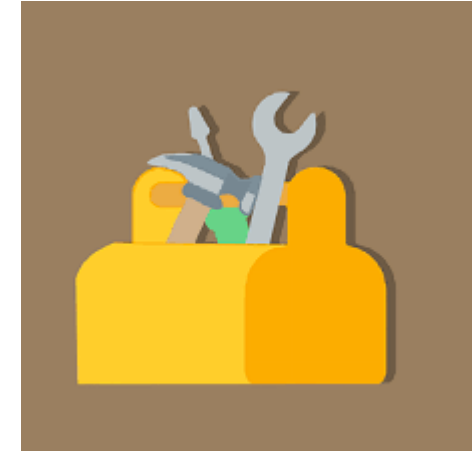


ANOVA : $p = 0,0$; $F = 4032,74$ indique une grande dispersion de données de la moyenne.
 $\eta^2 = 0,08$

Analyse exploratoire bivariée.

Deux variables qualitatives

- Outils
 - Les techniques d'analyse utilisées:
 - Tableaux de contingence
 - Chi-2
 - Les techniques de visualisation:
 - Heatmap

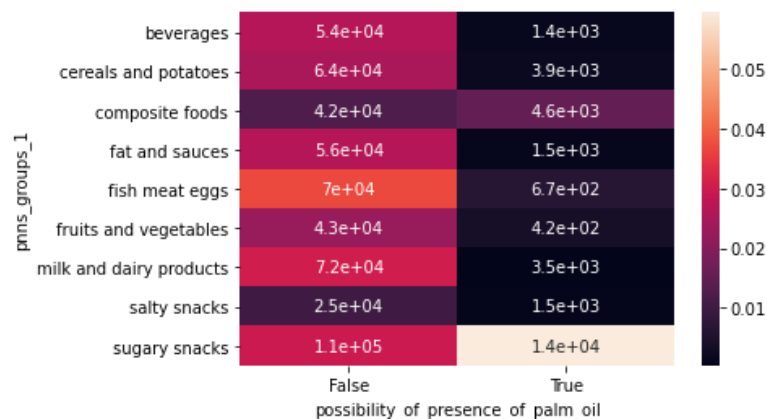


Analyse exploratoire bivariée.

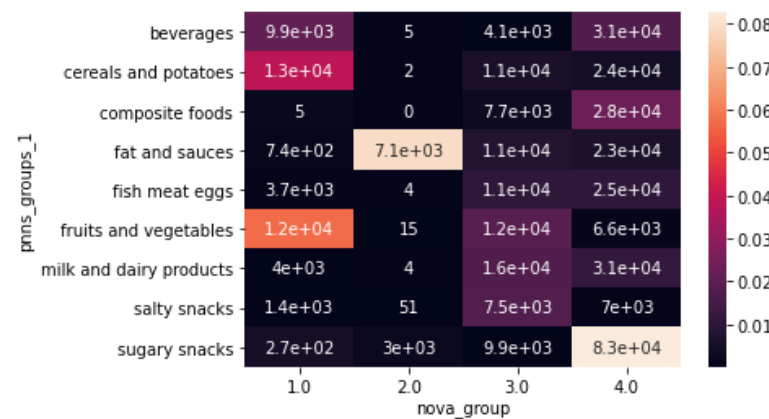
Deux variables qualitatives

Résultats:

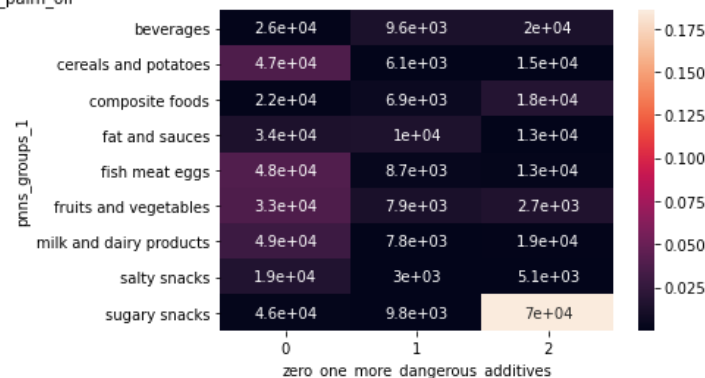
- Le groupe d'aliments "en-cas sucrés" contient plus d'huile de palme, additifs dangereux et constitué de plus d'aliments transformé que d'autres groupes alimentaire.



Chi2 = 15367,06, p-value = 0,00, dof (degree of freedom) = 18,00



Chi2 = 137597,41, p-value = 0,0, dof = 36,00



Chi2 = 68859,36, p-value = 0,0, dof = 27,00

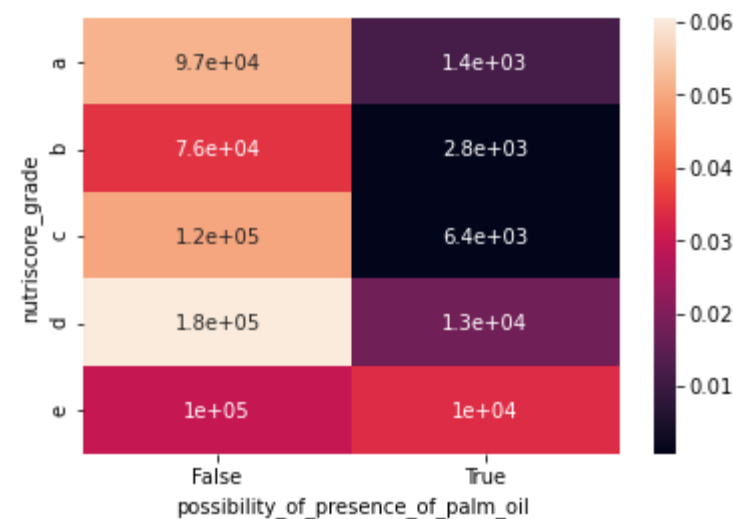
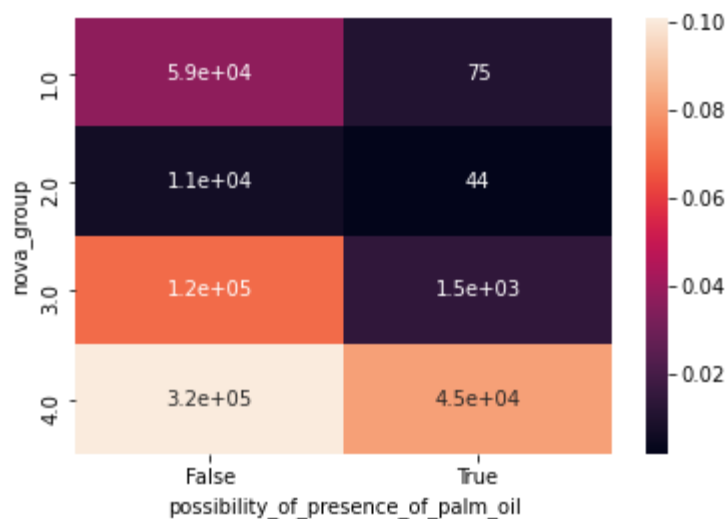
Analyse exploratoire bivariée.

Deux variables qualitatives

- **Résultats:**

- très peu de lien entre la présence d'huile de palme et :
 - les groupes NOVA;
 - le score de nutriscore.

Chi2 = 7075,51, p-value
= 0,00, dof (degree of
freedom) = 10,00



Chi2 = 20803,32, p-value
= 0,00, dof (degree of
freedom) = 8,00

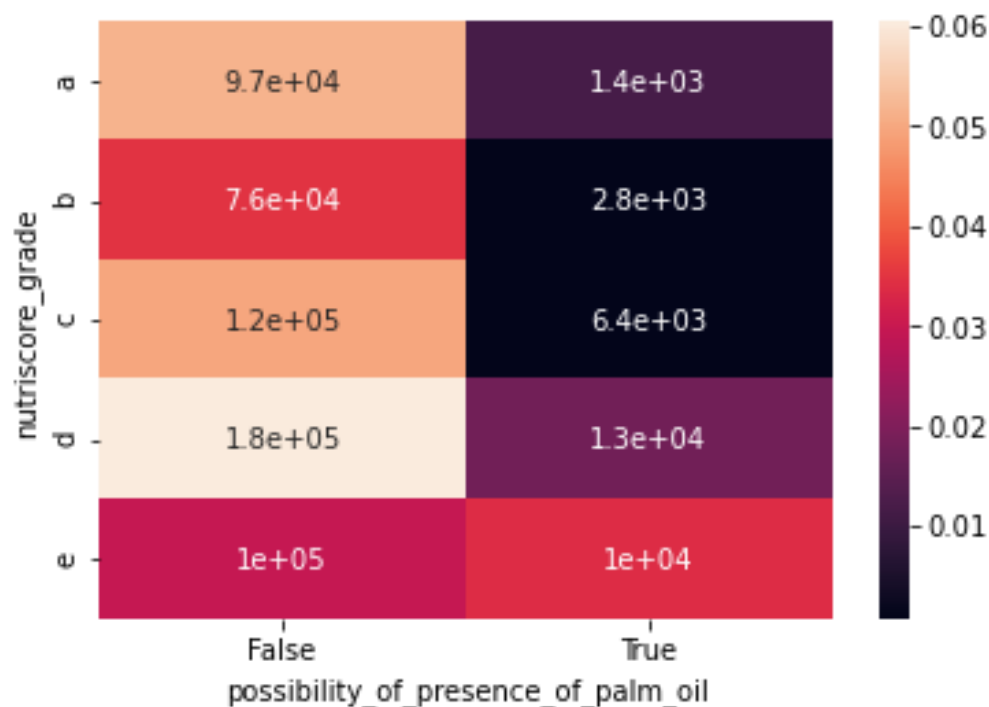
Analyse exploratoire bivariée.

Deux variables qualitatives

- **Résultats:**

- les aliments sains ont moins de probabilité de contenir l'huile de palme.

Chi2=7075,51, p-value = 0,00, dof (degree of freedom) = 10,00



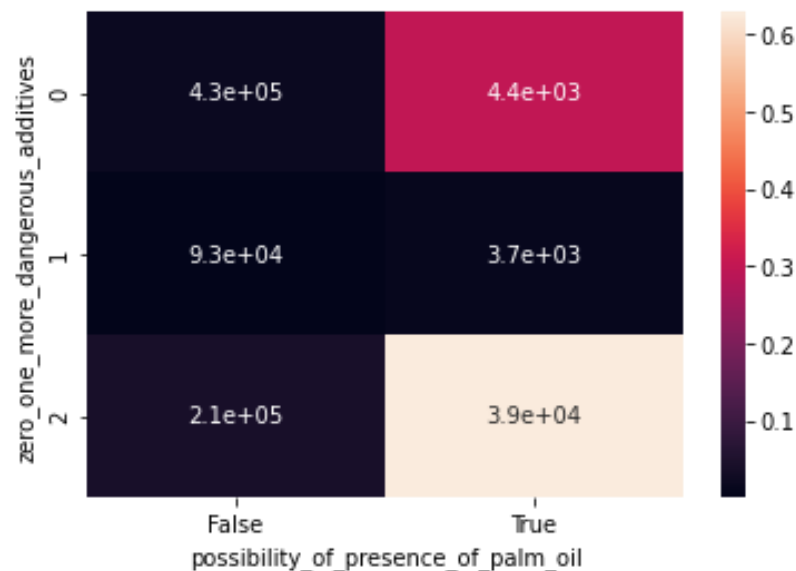
Analyse exploratoire bivariée.

Deux variables qualitatives

- **Résultats :**

- il peut y avoir l'huile de palme quand il n'y a pas d'additifs dangereux
- quand il y a un seul additif dangereux, il est peut probable la présence d'huile de palme.
- quand il y a 2 ou plus d'additifs dangereux, il y a plus de probabilité que les aliments contiennent l'huile de palme.

Chi2 = 61125,28, p-value = 0,00, dof (degree of freedom) = 6,00



Analyse exploratoire bivariée.

Conclusion



moins de probabilité de contenir l'huile de palme

EN-CAS SUCRÉS

Graisses
saturées

L'huile de
palme

Additifs
dangereux

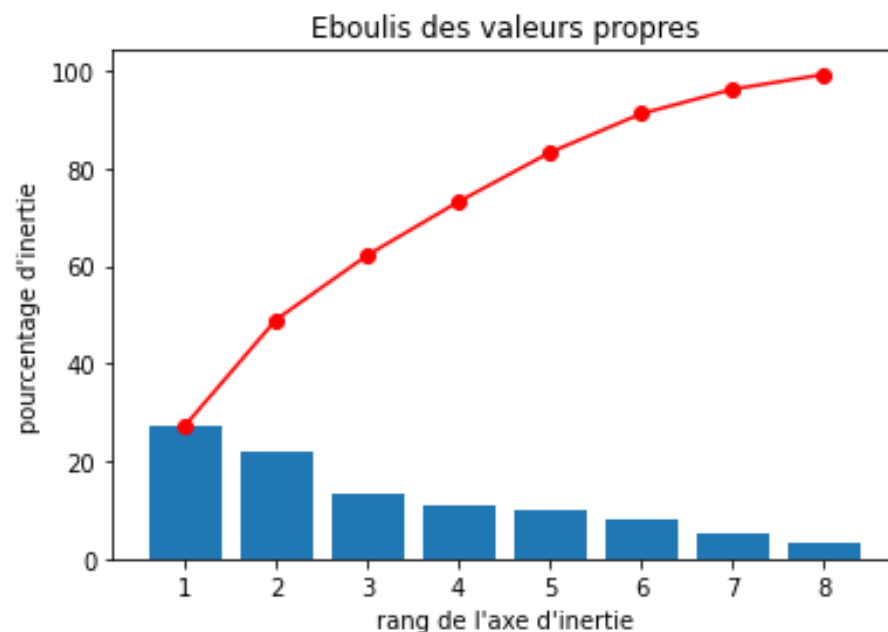
Aliments ultra-
transformés

L'analyse exploratoire multivariée

ACP (ou PCA, en anglais)

- Variables quantitatives continues:

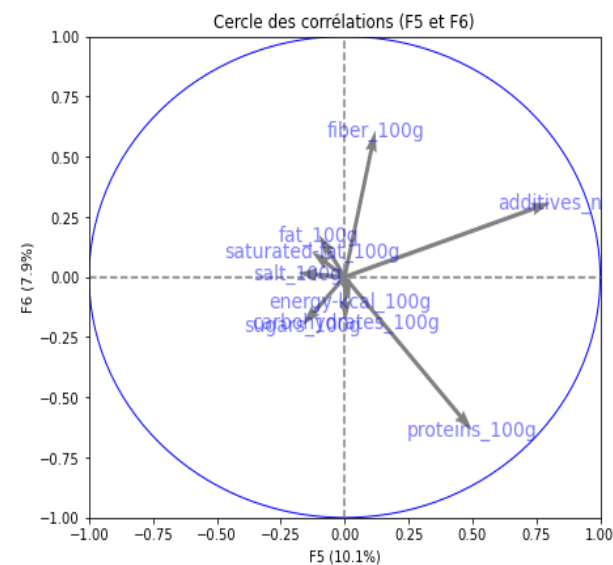
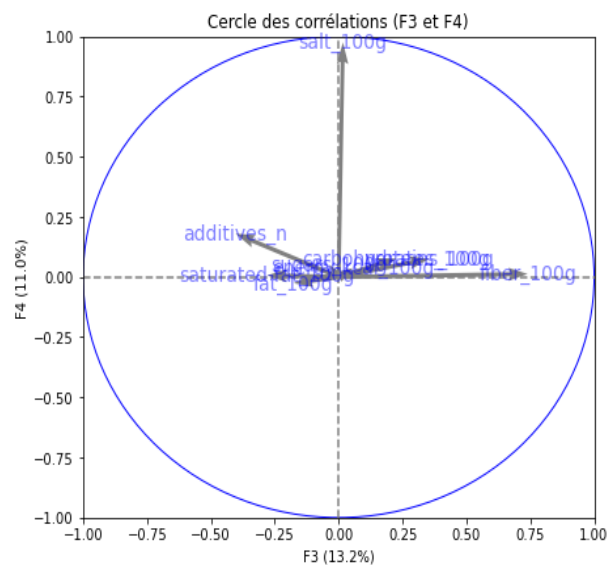
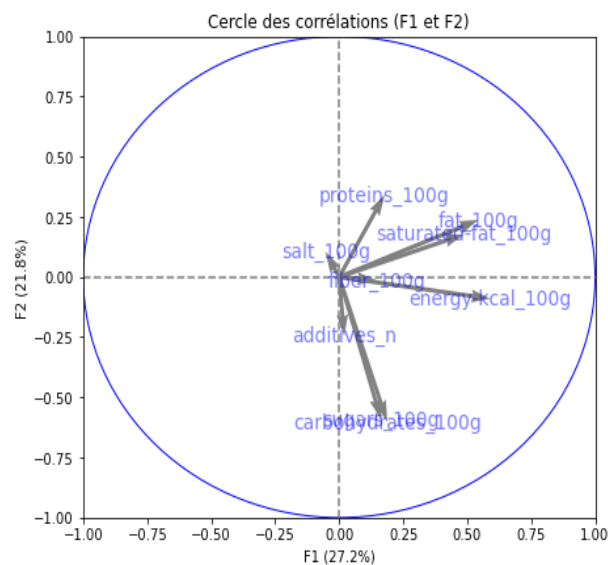
'additives_n', 'carbohydrates_100g', 'energy-kcal_100g', 'fat_100g', 'fiber_100g',
'proteins_100g', 'salt_100g', 'saturated-fat_100g', 'sugars_100g'



L'analyse exploratoire multivariée

ACP (ou PCA, en anglais)

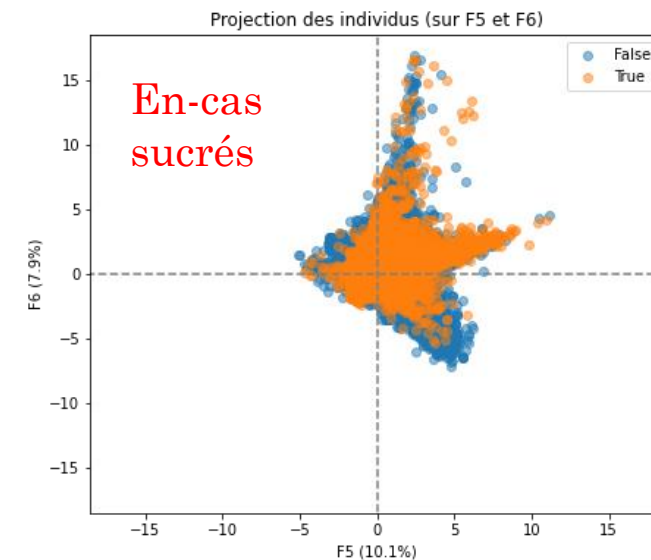
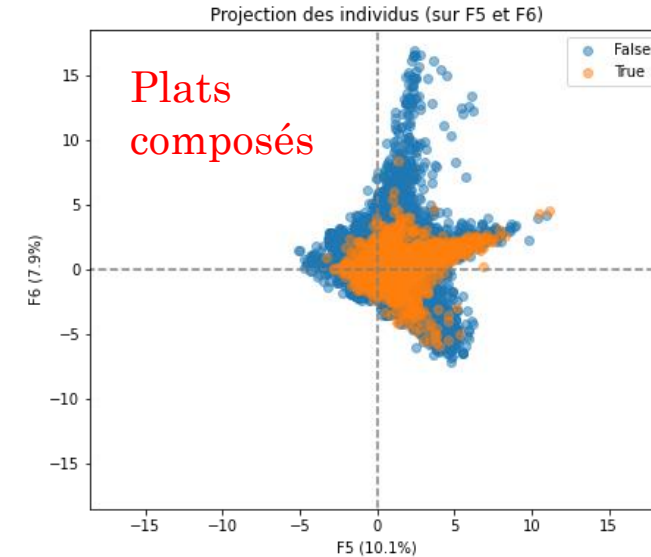
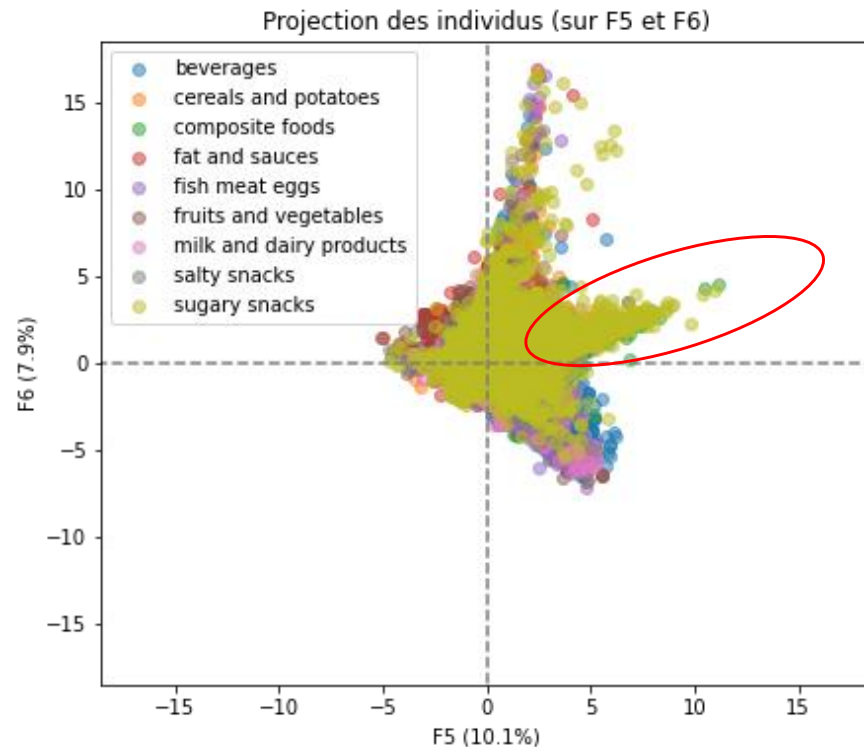
- Trois projections:



L'analyse exploratoire multivariée

ACP (ou PCA, en anglais)

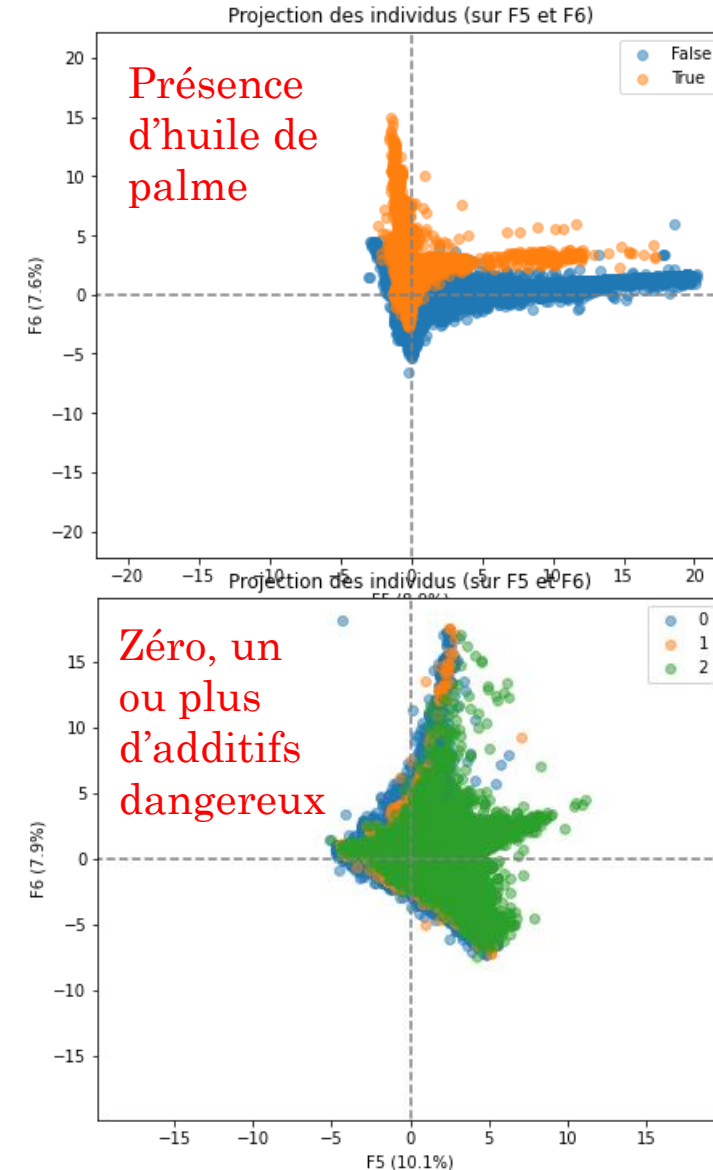
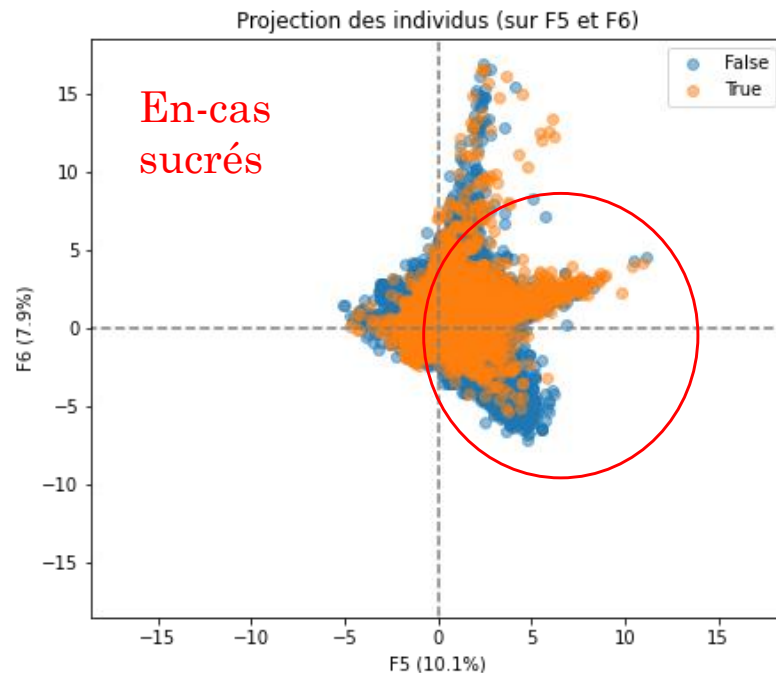
Les variables illustratives: pnns_groups_1,
sugary snacks, composite foods



L'analyse exploratoire multivariée

ACP (ou PCA, en anglais)

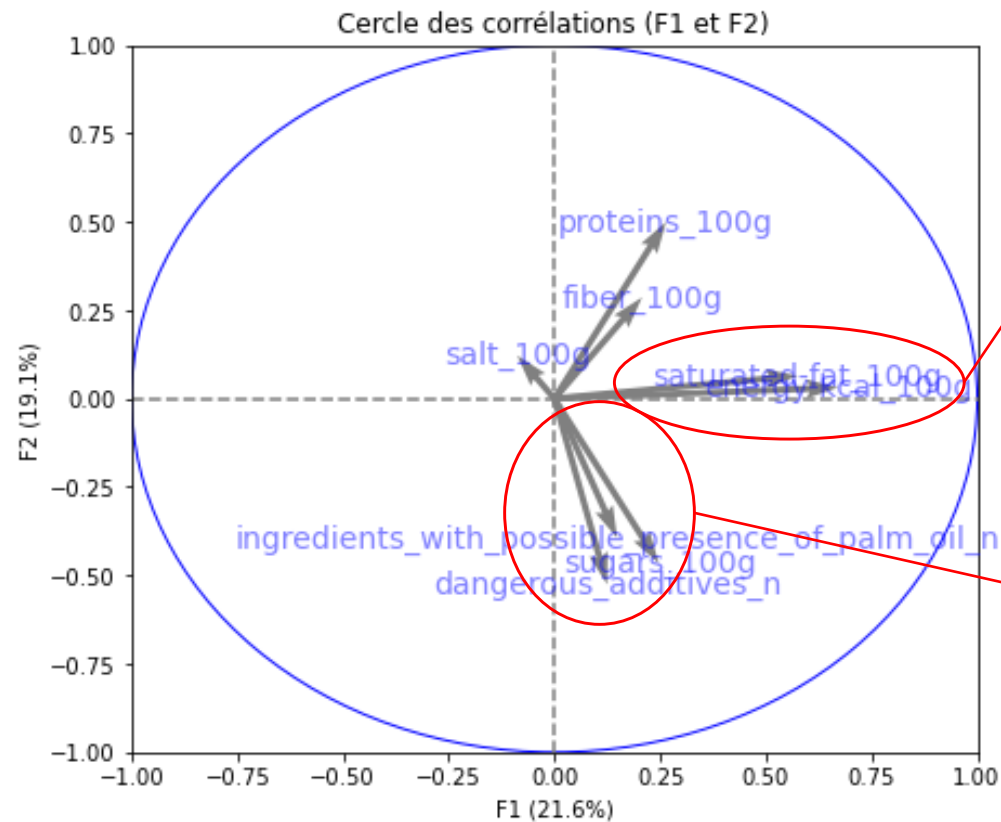
Les variables illustratives: sugary snacks,
possibility_of_presence_of_palm_oil,
zero_one_more_dangerous_additives



L'analyse exploratoire multivariée

ACP (ou PCA, en anglais)

- Variables quantitatives continues + le nombre d'additifs dangereux et le nombre d'ingrédients pouvant contenir l'huile de palme :



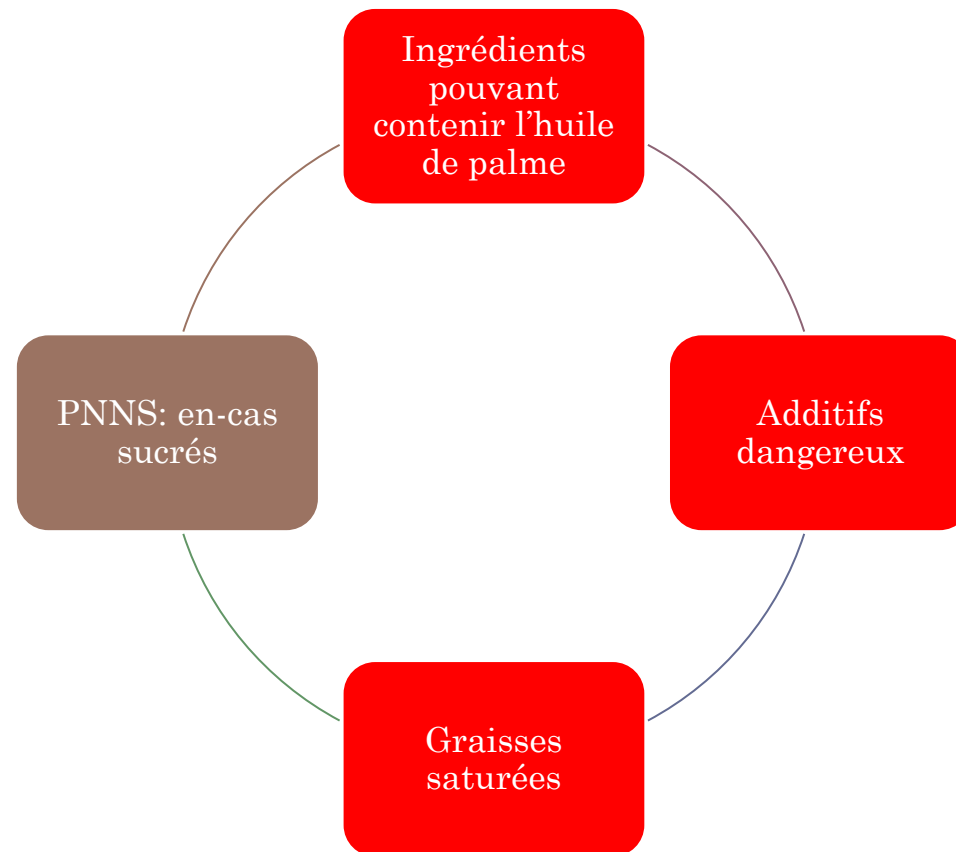
Axe de graisses saturées et des calories

- Nombre d'ingrédients pouvant contenir l'huile de palme
- Sucres raffinés
- Additifs dangereux

L'analyse exploratoire multivariée

Conclusion

- L'ACP confirme les conclusions de l'analyse bivariée:



Conclusion

- Il existe un lien entre la présence d'huile de palme, des additifs dangereux et des graisses saturées.
- Ce lien dépend de groupe alimentaire.
- Le lien le plus fort est observé dans le groupe alimentaire les « en-cas sucrés ».
- L'idée d'application proposée est pertinente et faisable.



Conclusion

