Projet 2 : Concevez une application au service de la santé publique

Soutenance OpenClassRooms, le XX/XX/2022

Erwan CHESNEAU

Plan :

- Contexte
- Présentation de l’application
- Nettoyage de la base de données
- Analyse (uni, bi et multi variée)
- Analyse de la faisabilité de l’application

Contexte :

- Appel à projet de l'agence santé publique France :
  - Développer une application au service de la santé publique
  - A partir de la base de données d'OpenFoodFacts

- Open Food Facts :
  - Projet collaboratif
  - Référencer les produits alimentaires
    - Informations sur le produit

Application : Pourquoi?

- Prise de conscience de la population :
  - Santé : volonté de manger plus sainement
    - Privilégier des produits avec un bon nutriscore
    - Privilégier des produits sans additifs
    - Privilégier des produits moins caloriques
  - Environnement : volonté de moins polluer
    - Privilégier des produits locaux
    - Privilégier des produits avec un faible impact environnemental

Application : l'idée

« bien manger pour soi et la planète »

- Proposer aux utilisateurs des produits à la fois sains et « eco-friendly »

- L'utilisateur scanne un produit
  - l'application résume les variables d'intérêt (nutriscore, impact carbone, origine du produit…)
  - L'application propose si possible un meilleur produit

Nettoyage des données:
Sélection des variables

- 162 variables dans la base de données :
  - Caractériser le produit, les nutritions, les compositions, les origines….

- Sélection uniquement des variables utiles pour l'application :
  - Définition du produit : code, product_name, categories_tags
  - Communication des informations : url
  - Origine du produit : manufacturing_place_tags, origins_tags
  - Informations nutritionnelles : nutrition_grade_fr, energy_100g, fat_100g, saturated_fat_100g, fruits-vegetables-nuts_100g, sodium_100g, sugars_100g, additive_n
  - Impact environnemental : carbon-footprint_100g

Nettoyage des données:

Variables	Conditions	Actions supplémentaire	Vide avant	Vide après	supprimée
Code produit	str de chiffres		23	23	0
url	Format url Accessible Contenir le code		23	2375	2352
Nom	Str > 3	Minuscules Suppression espace début et fin Remplacement espaces par '_'	17762	17819	57
catégories	Str Traductible en 'en'	Liste Anglais Sans doublons	236383	236388	5
pnns	Str Pas unknown	'-' = '_' Espace = '_'	229259	251883	22624

Nettoyage des données:

Variables	Conditions	Actions supplémentaire	Vide avant	Vide après	supprimée
Lieux	Str Contenir un pays	Liste Anglais minuscules	284277	286125	1848
Nutriscore	Str = a, b, c, d ou e	minuscules	99562	99562	0
additifs	int		284277	286125	1848
Carbone Energie	Float ou int		320504 59659	320504 59659	0 0
Fat_100g	Float < 100 et >0		76881	76885	4
Sat_fat	Float < 100 et >0		91218	91221	3
F-v-n_100g	Float < 100 et >0		317736	317736	0
Sodium	Float < 100 et >0		65309	65343	34
sucre	Float < 100 et >0		75801	75820	19

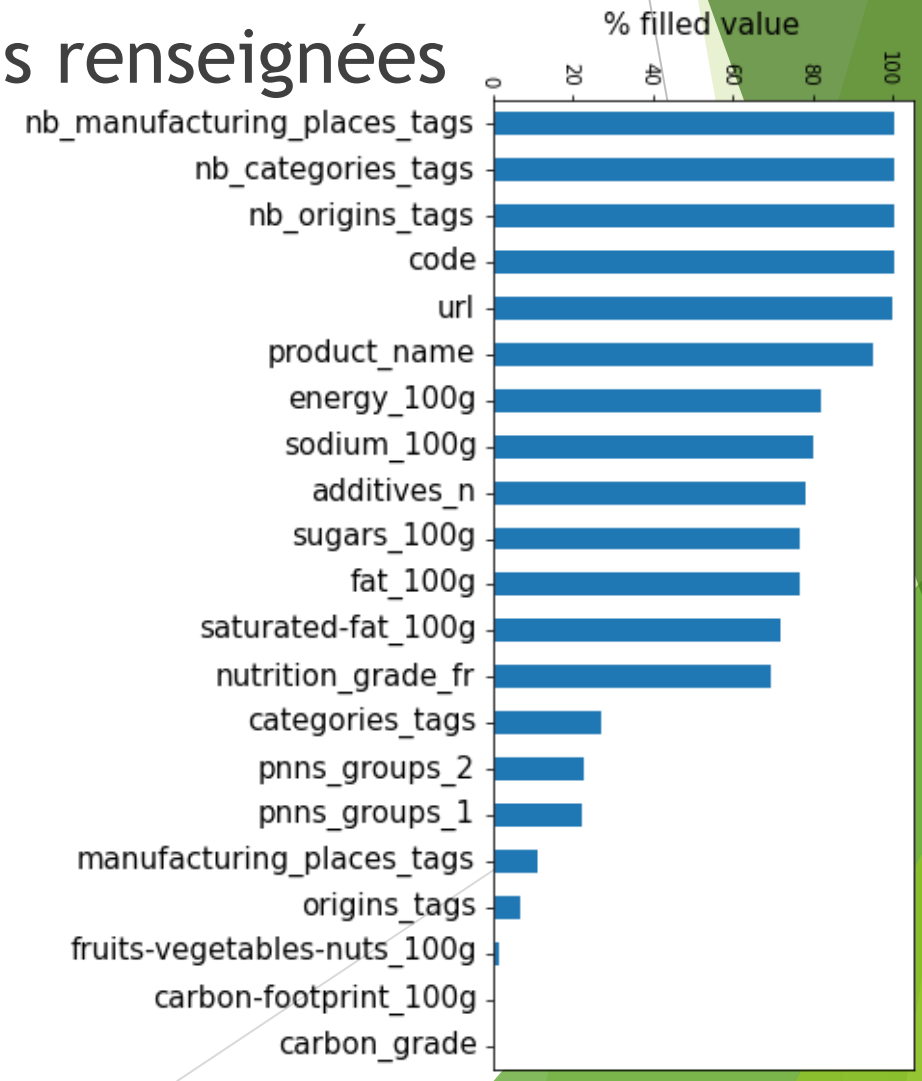
Création de nouvelles variables:

- Carbone score :
  - Classification en fonction de l'empreinte carbone
  - Définit par la loi climat

- Nombre d'éléments dans les variables listes :
  - Nb_categories_tags,
  - nb_origins_tags,
  - nb_manufacturing_places_tags

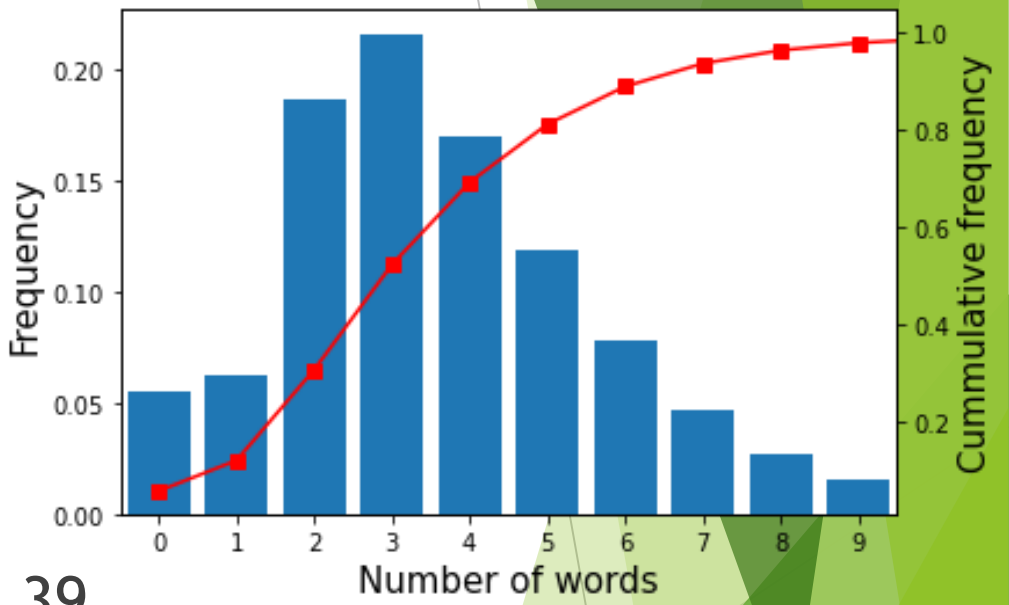
Analyse univariée :
taux de complétion

- Identification des variables les plus renseignées
  - 10 variables > 60 %
  - Les catégories, les origines et l'empreinte carbone très peu renseignées



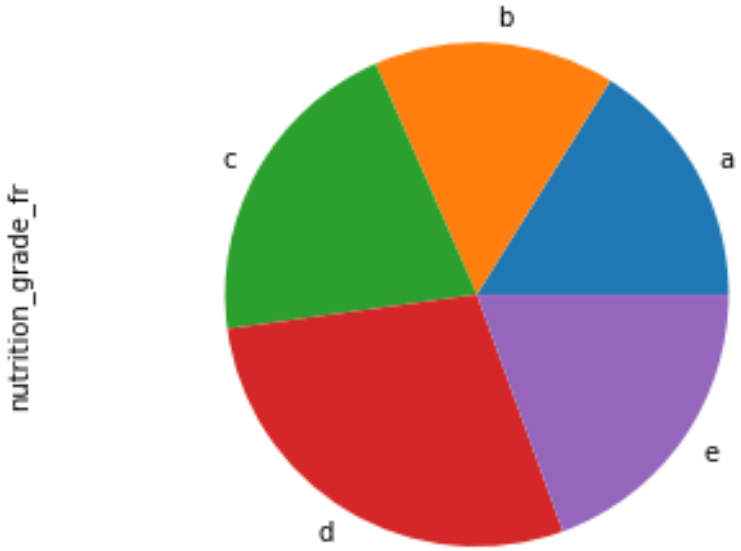
Analyse univariée :
Nom des produits

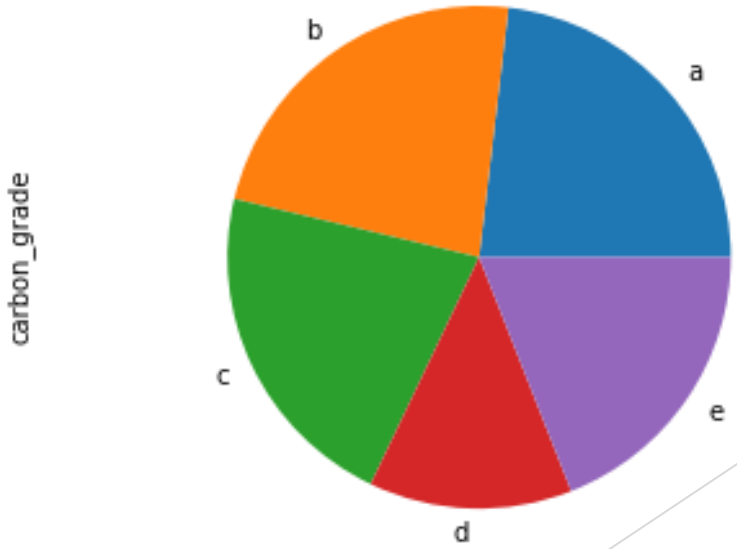
- Nombre de mots par noms de produits
  - Moyenne = 3.76
  - Médiane = 3.0
  - Std = 2.31
- Valeurs aberrantes :
  - Nombre de mots pour 95% des individus : 8.39
  - Les noms de produits avec plus de 9 mots peuvent être considérés aberrants



Analyse univariée : score

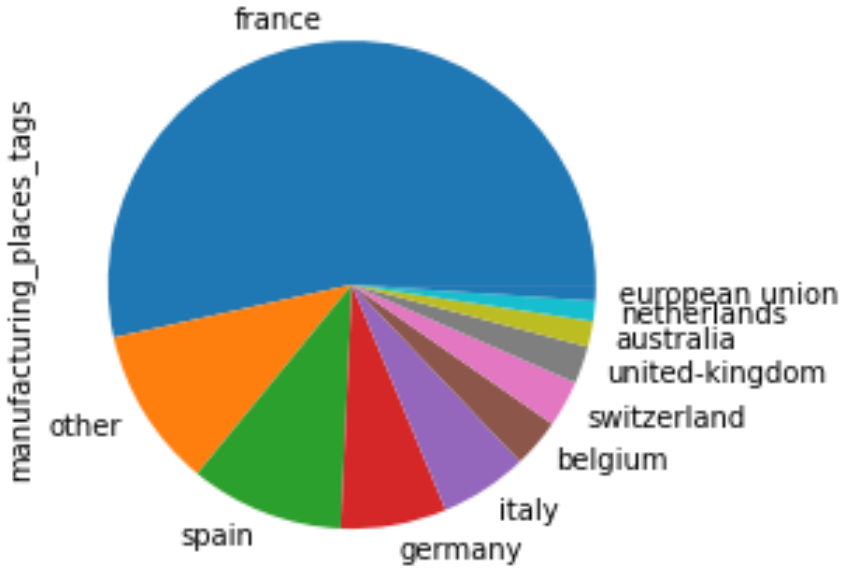
- Chaque score est correctement représenté
  - Le nutriscore « d » est légèrement sur-représenté
  - Le carbone score « d » est légèrement sous-représenté

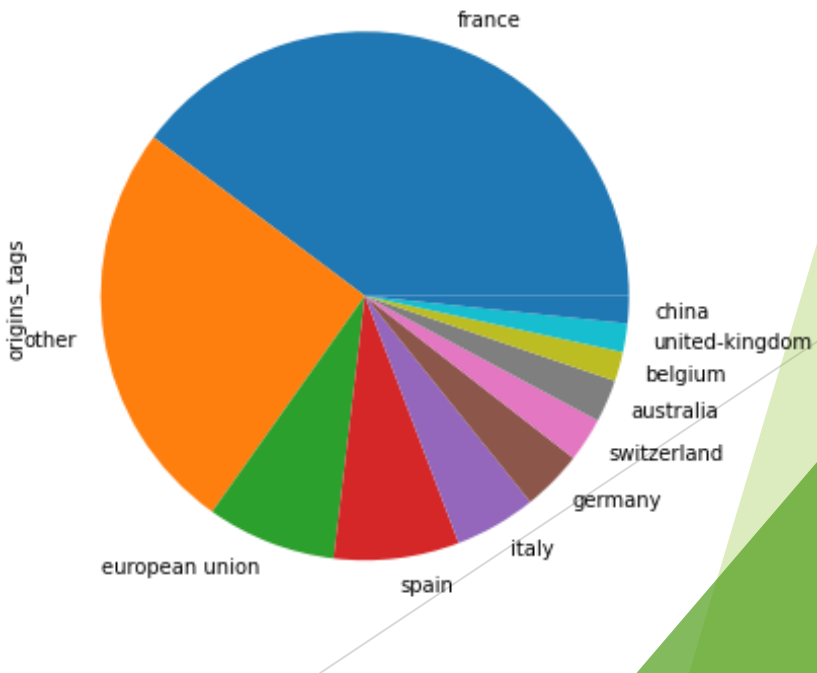




Analyse univariée :
Lieux de productions et origines

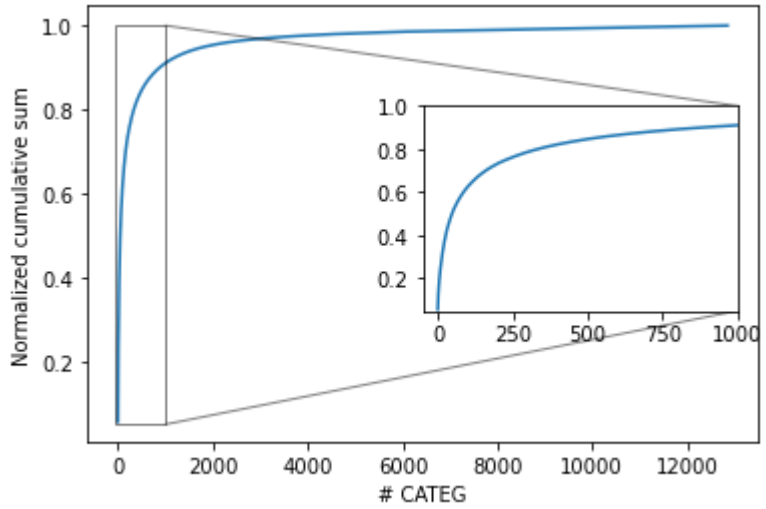
- La France est très représentée
  - Peut traduire une forte collaboration au projet
  - Adaptable à la France
- Grande diversité des pays

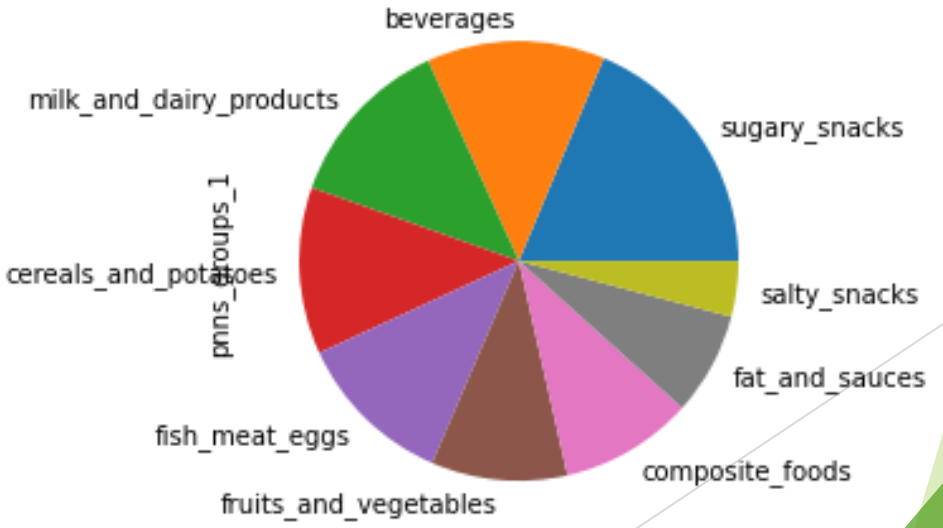




Analyse univariée : catégories

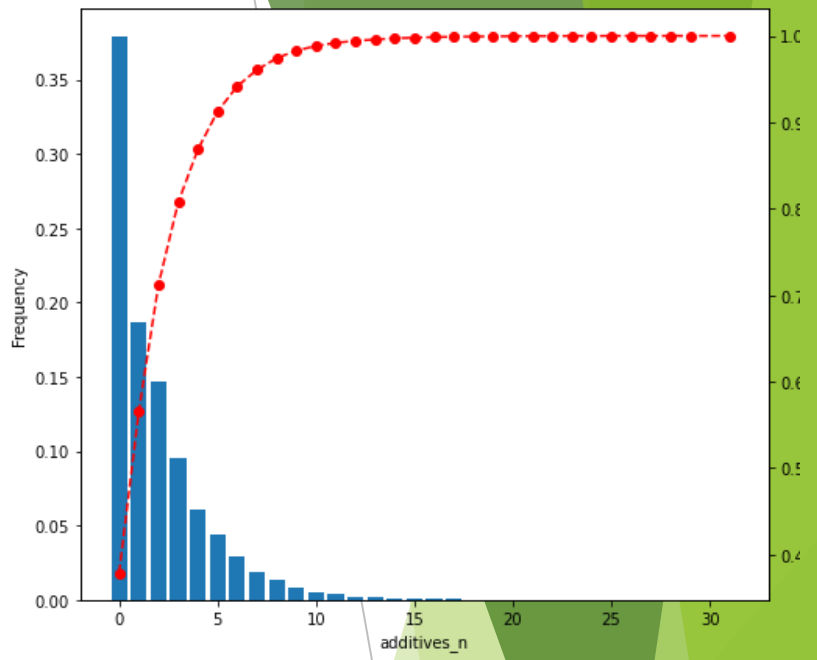
- Les tags « catégories » sont très nombreux
  - Mais 500 catégories regroupes 95% des produits
- PNNS regroupe dans des grands groupes les produits
  - Répartition homogène entre les catégories

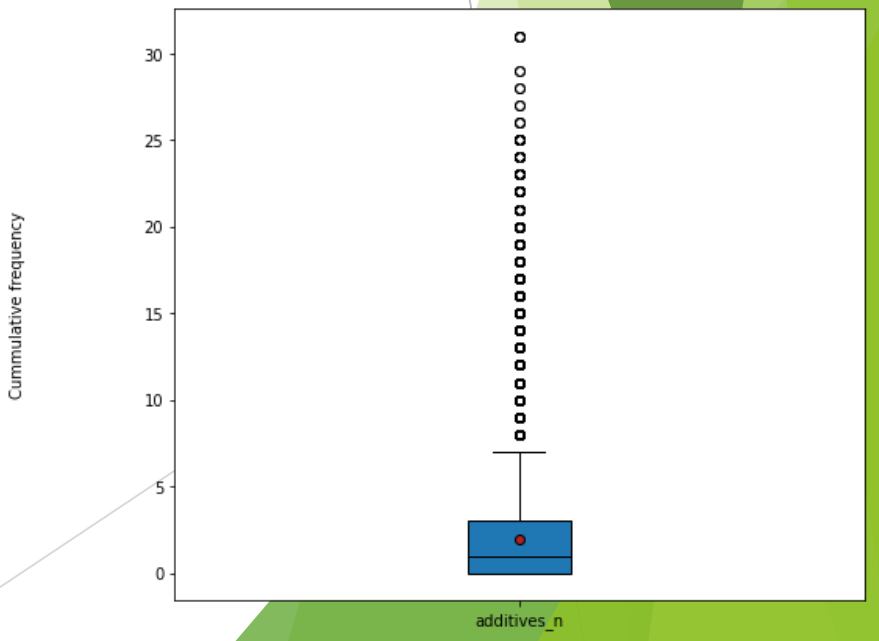




Analyse univariée : Variables Quantitatives

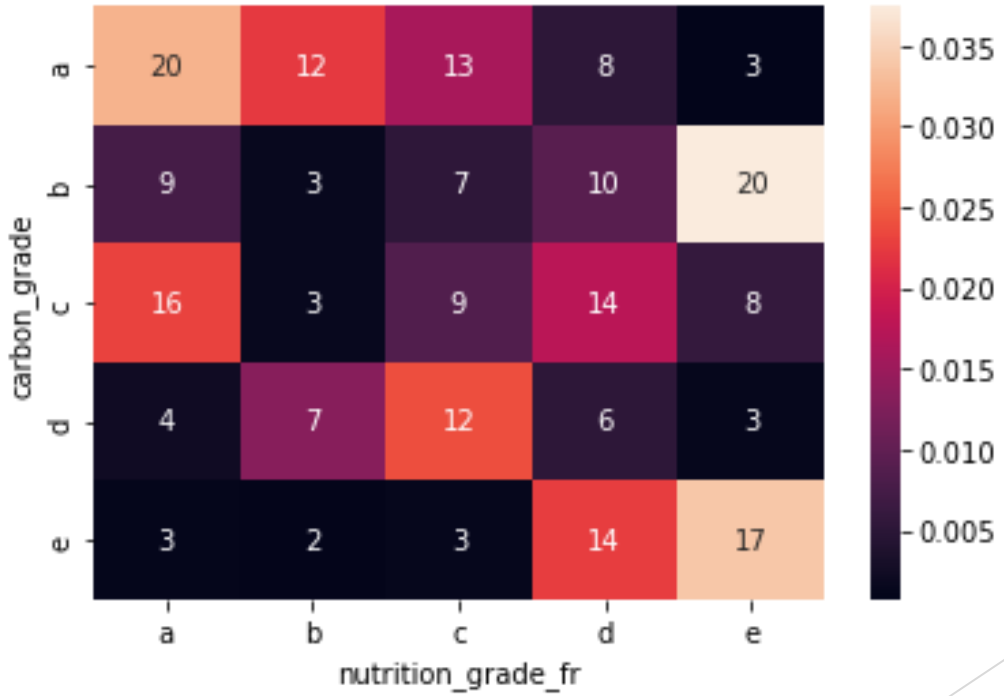
- Représentation des distributions
  - Caractérisation de la moyenne
    - Valeur regroupées vers les petites valeurs
  - Caractérisation de l'asymétrie
    - Élargissement vers la droite
  - Caractérisation de l'aplatissement
    - Distribution plus fine que la loi normale
  - Détection des outliers





Analyse bivariée : relations carbone score / nutriscore

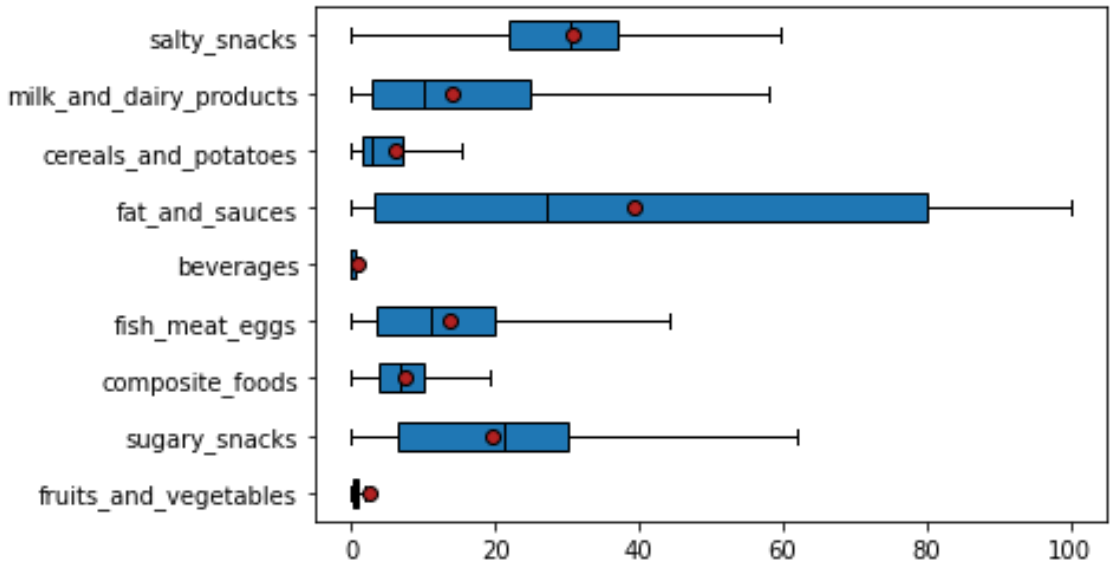
- Légère corrélation entre les deux variables
  - Les bons nutriscores ont tendance à avoir un bon score carbone



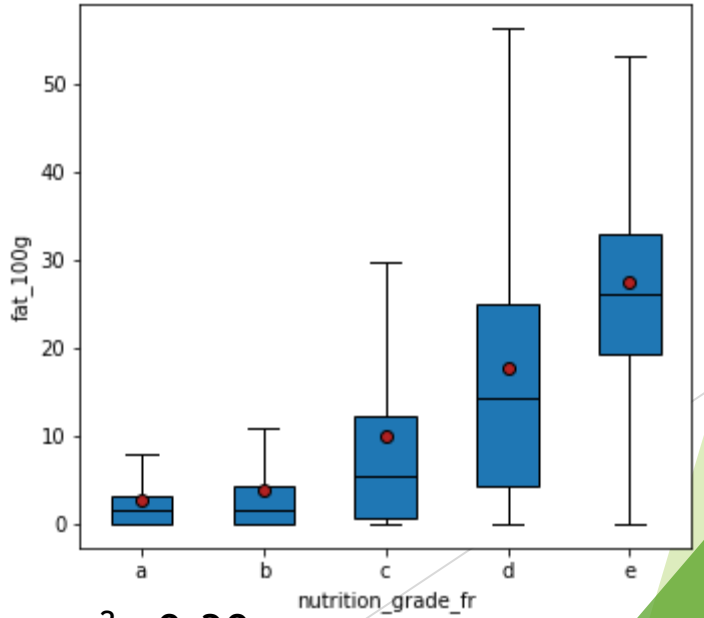
Analyse bivariée :ANOVA

relation entre taux de graisse et catégories ou nutriscore

- Légère corrélation entre les variables
- Certaines catégories ont en moyenne moins de gras
- Les bons nutriscores ont tendance à être moins gras



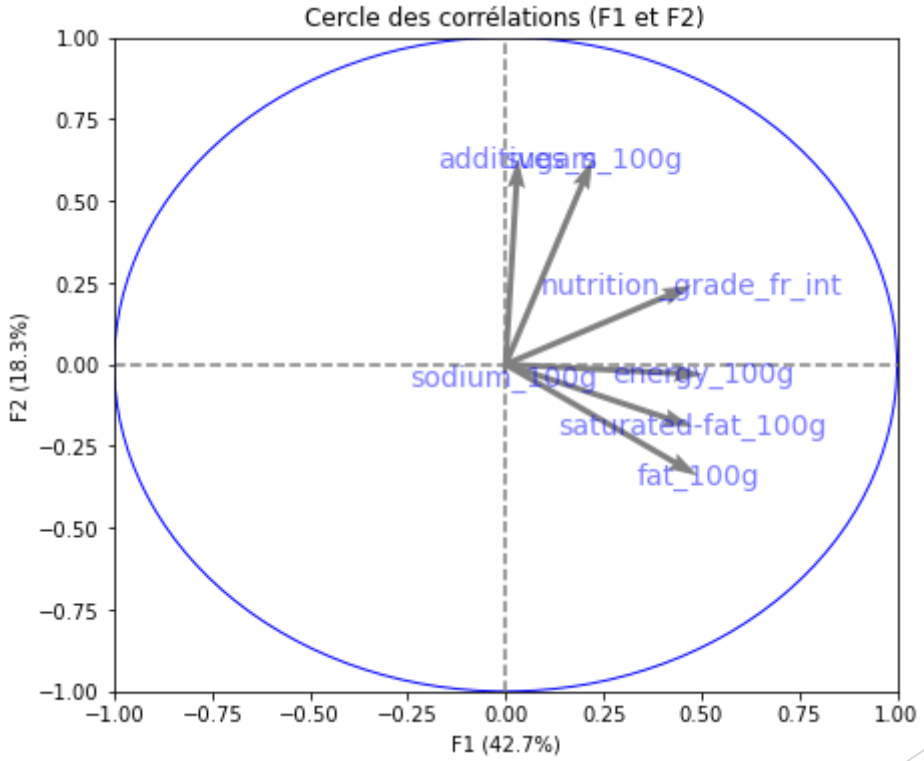
$\eta^2 = 0,35$



$\eta^2 = 0,30$

Analyse multivariée :ACP

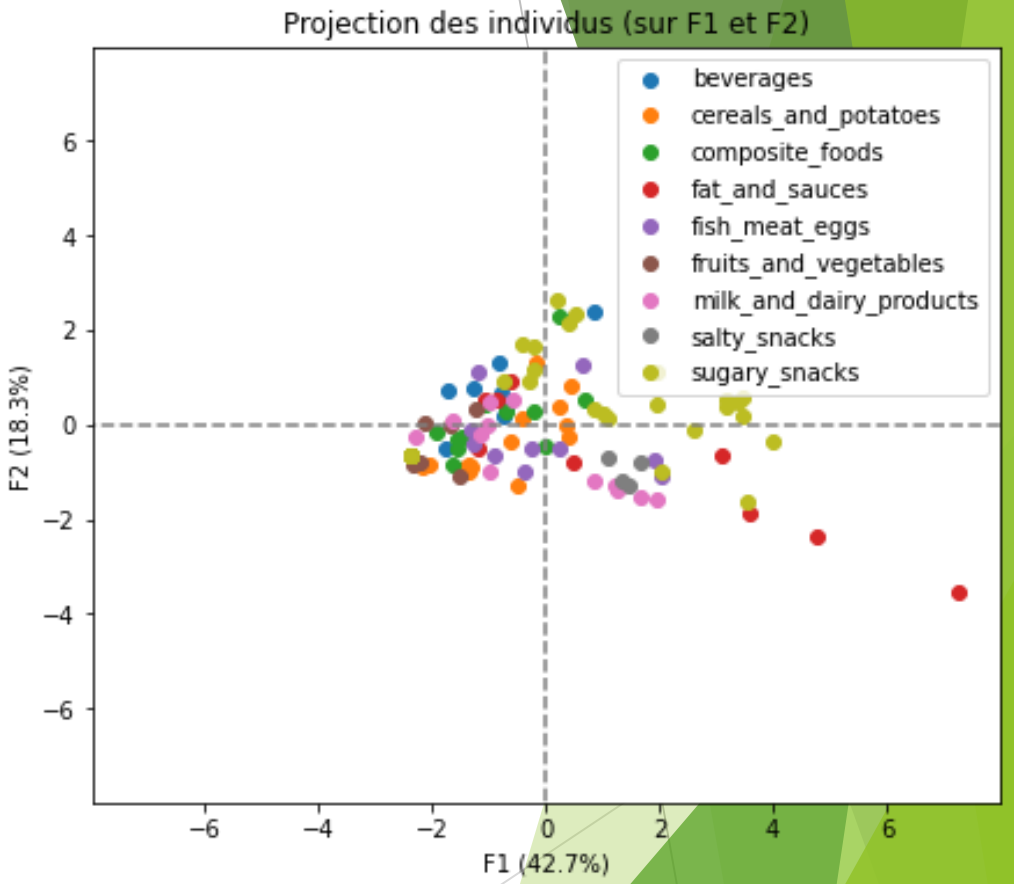
- Axe F1 : teneur en gras et apport calorifique
- Axe F2 : teneur en sucre et additifs (transformés)
- Le nutriscore se situe entre le deux



Analyse multivariée :ACP

- Projections F1/F2 :
  - regroupement en son centre
  - Fortement impacté par les catégories de produits sucrés et gras

- Les outliers attirent les 1ers axes d'inerties



Faisabilité

- Points positifs :
  - Forte représentation de la France
  - Beaucoup de données concernant l'alimentation
- Points négatifs :
  - Peu de données concernant l'origine et la production
  - Empreinte carbone presque inexistante
- Améliorations :
  - Clustering des catégories (méthodes K-means ?)
  - Méthode de prédiction du carbone score en fonction des données disponibles?

Santé publique France

OPEN FOOD FACTS

Merci de votre attention !

OPENCLASSROOMS

erwanchesneau@gmail.com