



29/08/23

Préparez des données pour un organisme de santé publique

Armand FAUGERE [LinkedIn](#)

armand-faugere@live.fr

Sommaire

- I) Cadrage du projet
- II) Données d'entrée
- III) Démarche de traitement des données
- IV) Démarche d'analyse des données
- V) Synthèse des résultats
- VI) Conclusion



I) Cadrage du projet

I) Cadrage du projet



❑ **Contexte** : Projet d'amélioration de la base de données Open Food Facts.

Pour rajouter un produit, il y a de nombreux champs textuels et numériques à remplir, ce qui peut conduire à des erreurs de saisie et à des valeurs manquantes



❑ **But** :

- Créer un système de suggestion ou d'auto-complétion pour aider les usagers à remplir plus efficacement la base de données

❑ **Objectifs** :

- Traiter le jeu de donnée pour le rendre exploitable
- Explorer les données
- Réaliser des tests statistiques pour valider les résultats des analyses
- Rédiger un rapport d'exploration et une conclusion sur la faisabilité du projet
- Respecter les 5 grands principes RGPD

II) Données d'entrée

2) Données d'entrée

☐ Le jeu de données <https://fr.openfoodfacts.org>

➔ base de données collaborative de produits alimentaires qui répertorie les ingrédients, les allergènes, la **composition nutritionnelle** et toutes les informations présentes sur les étiquettes des aliments pour aider le consommateur dans ses choix

Plus de 200 pays, plus de 600000 produits, plus de 9000 contributeurs



☐ Descriptif des champs du jeu de donnée

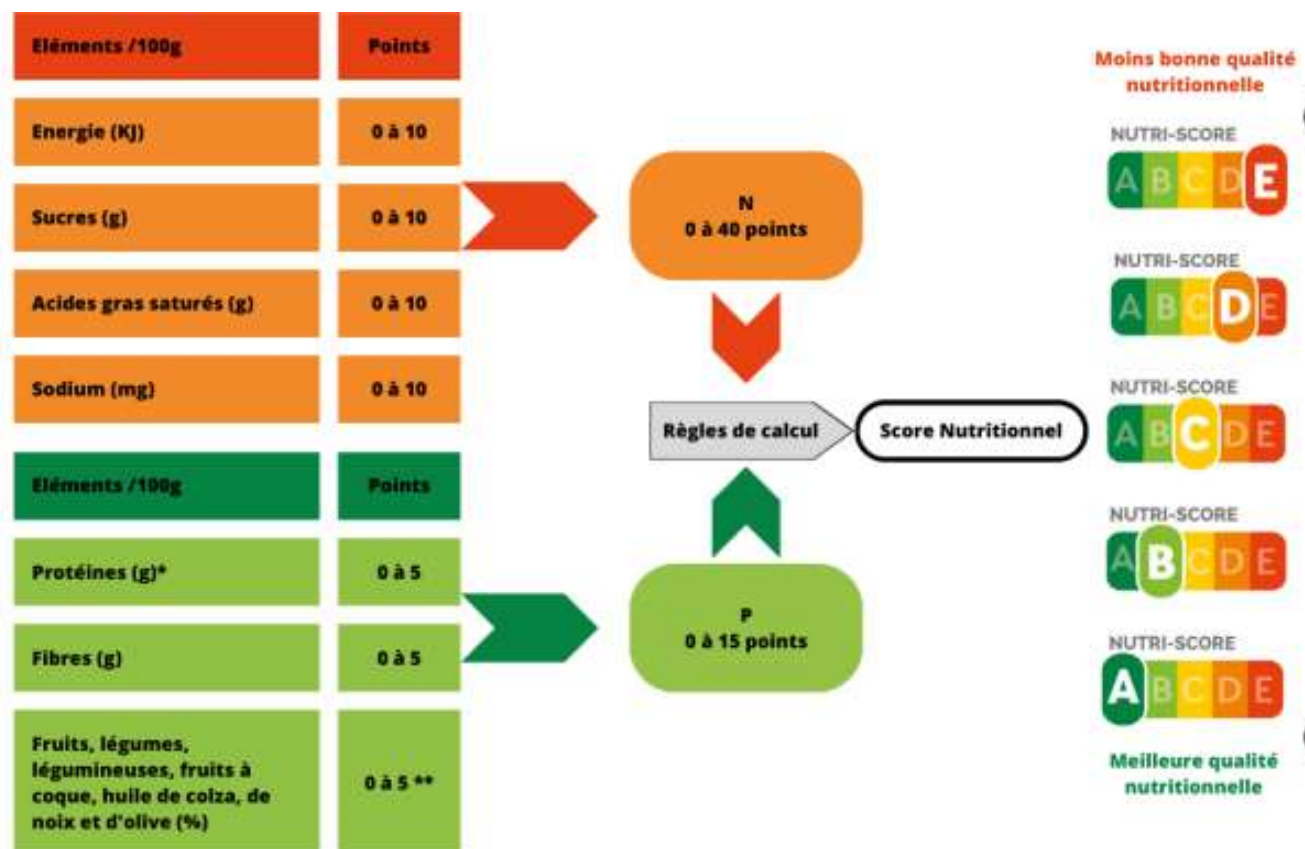
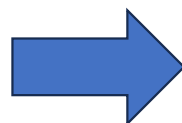
☐ Nutriscore questions réponses <https://www.santepubliquefrance.fr>



☐ Principes de protection des données (finalité, proportionnalité et pertinence, durée de conservation limitée, sécurité et confidentialité, droits des personnes) www.cnil.fr

2) Données d'entrée

Calcul du Nutriscore



*Selon le nombre des points "défavorables" et des points obtenus pour la composante "Fruits, légumes légumineuses, fruits à coque et huile de colza, de noix et d'olive" les protéines sont prises en compte ou non.

**Dans le cas des boissons, le maximum des points accordés est de 10.

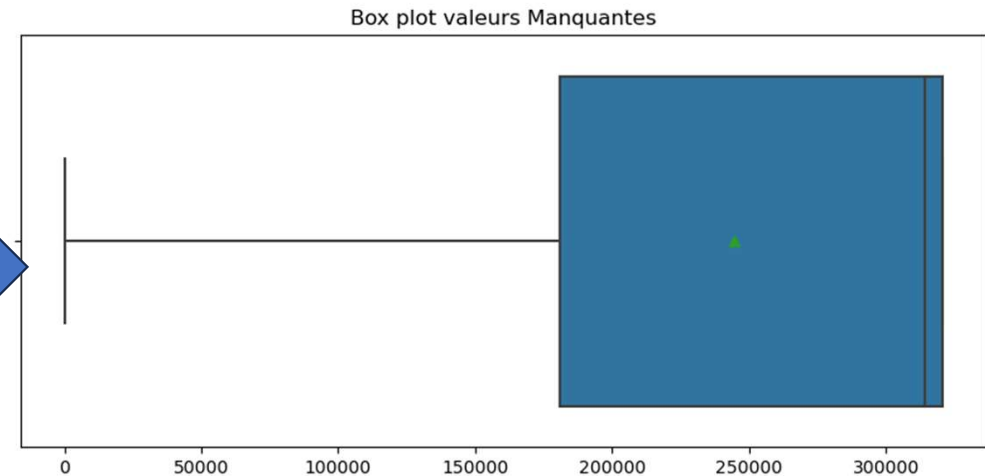
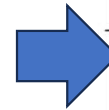
<https://www.santepubliquefrance.fr>

Préparez des données pour un organisme de santé
publique

2) Données d'entrée

❑ Le jeu de données

- 162 colonnes → Type de données
- 320772 lignes → Produits
- En moyenne 244497 → Valeurs manquantes par colonnes
- 0 duplication
- Des variables catégorielles et numériques



Il y a assez peu de colonnes complètes

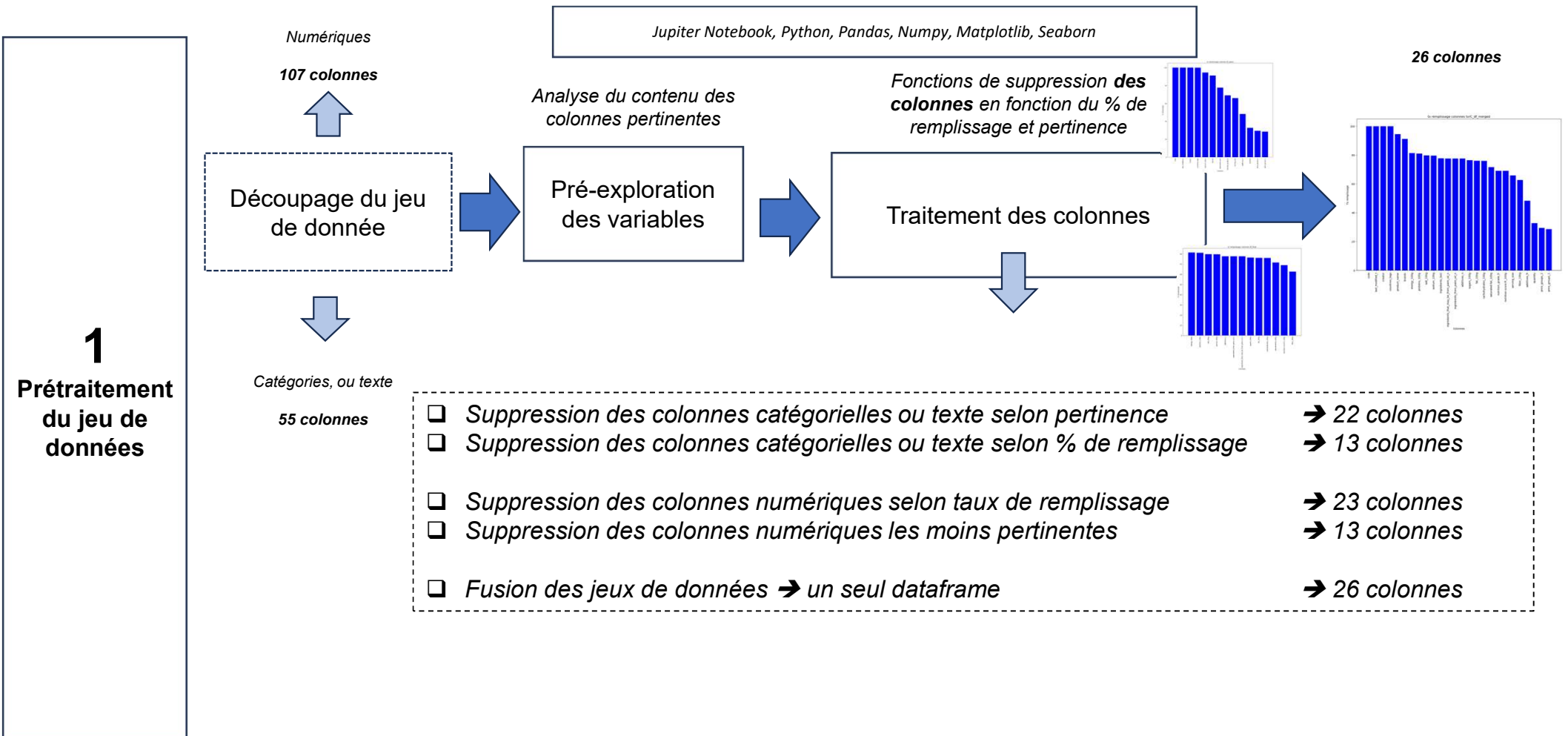
2) Données d'entrée

Les données brutes sont inexploitable.

Un travail important de traitement de données a été réalisé pour extraire les éléments pertinents

III) Démarche de traitement des données

3) Démarche de traitement des données



3) Démarche de traitement des données

2 Traitement des colonnes et des lignes



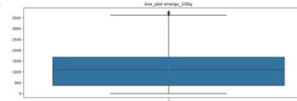
Jupyter Notebook, Python, Pandas, Numpy, Matplotlib, Seaborn

Analyse du contenu des colonnes

Fonctions d'exploration (tx de remplissage, comptage, box_plot...)

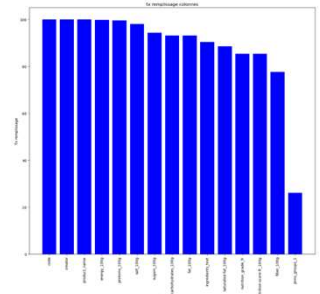
Exploration des colonnes

Traitement des colonnes



15 colonnes

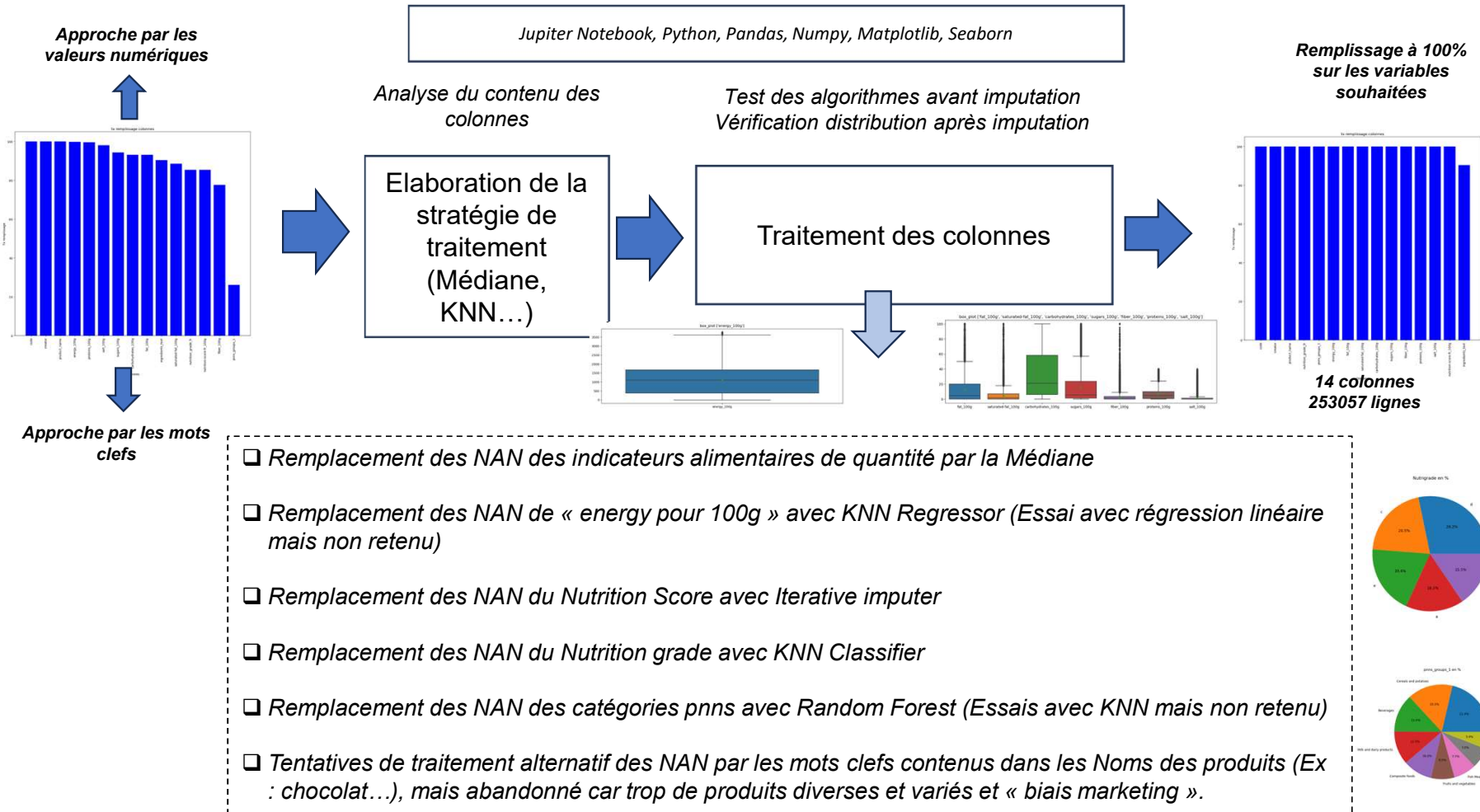
253057 lignes



- ☐ Suppression des lignes avec des erreurs de remplissage (pas le bon libellé...)
- ☐ Suppression des colonnes non pertinentes, informations en double
- ☐ Traitement des outliers pour les colonnes numériques (x écart interquartile, donnés > 100 g ou < 100g , valeurs impossibles, produits non pris en compte...)
→ Suppression des valeurs incorrectes
- ☐ Suppression des lignes avec trop de données manquantes (50% avec colonnes catégories incluses)

3) Démarche de traitement des données

3 Traitement des Valeurs manquantes



3) Démarche de traitement des données

But :

Créer un système de suggestion ou d'auto-complétion pour aider les usagers à remplir plus efficacement la base de données

Objectifs :

- **Traiter le jeu de donnée pour le rendre exploitable** 😊
- Explorer les données
- Réaliser des tests statistiques pour valider les résultats des analyses
- Rédiger un rapport d'exploration et une conclusion sur la faisabilité du projet
- Respecter les 5 grands principes RGPD

Confirmation → beaucoup de valeurs manquantes et d'erreurs dans les saisies des utilisateurs.

Des pistes de travail se dégagent :

- Outil d'aide à la saisie lorsque l'on remplit les champs (indication si éloigné des valeurs centrales)
- Nouveaux Indicateurs spécifiques pour aider et orienter l'utilisateur (ratio incohérent ou autre..)



Aide à l'utilisateur

Garantir la Qualité de la donnée

Maintenir la qualité des données dans le temps

IV) Démarche d'analyse des données

4) Démarche d'analyse des données

Jupyter Notebook, Python, *Pandas*, *Numpy*, *Matplotlib*, *Seaborn*, *sklearn*

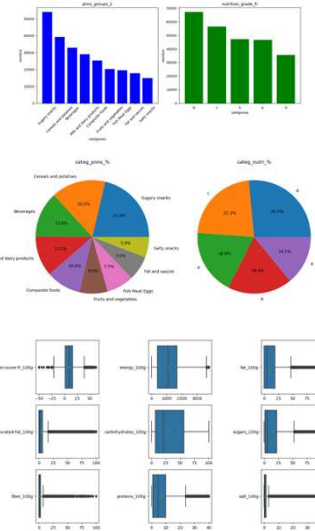
1 Analyse univariée

Création d'un nouveau note book
Importation du dataframe

- ☐ Importation du dataframe nettoyé et mis en forme
- ☐ Suppression de la colonne code + réindexage

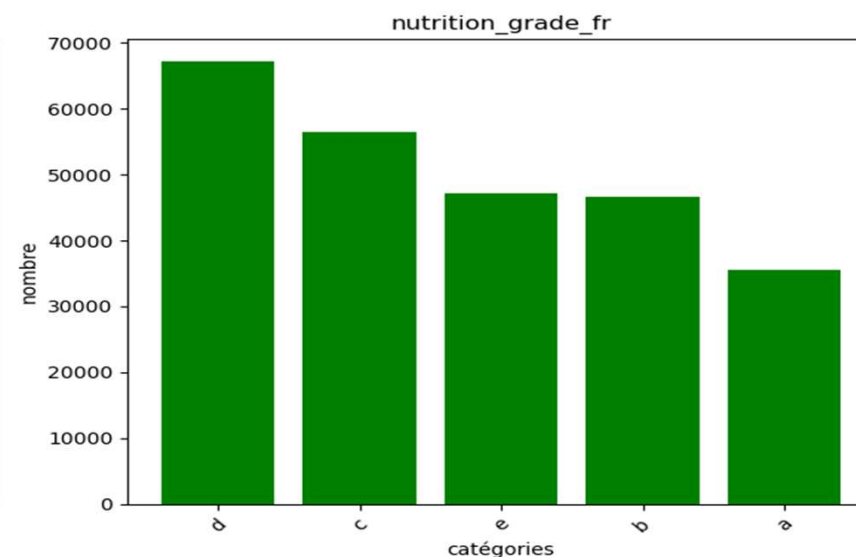
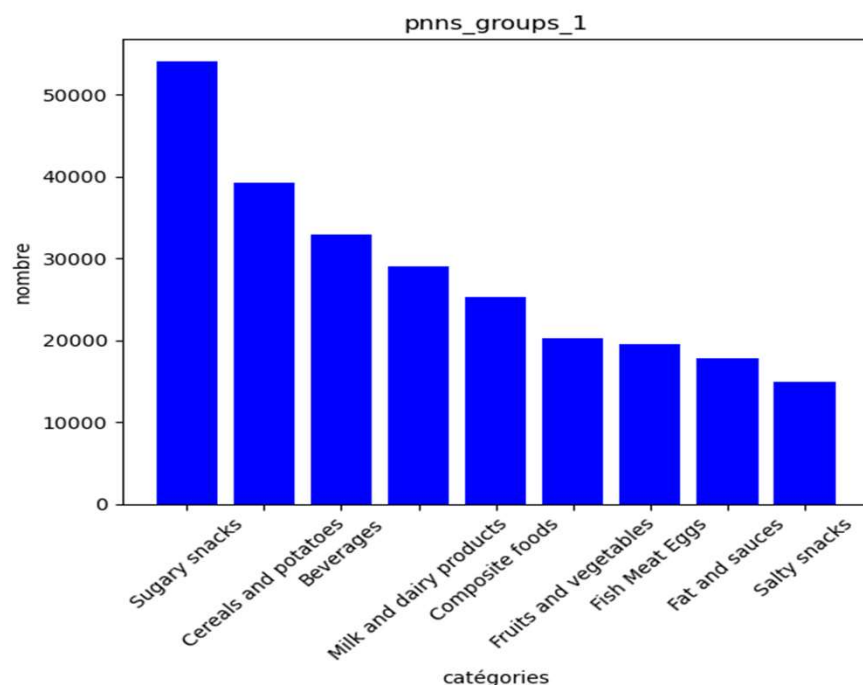
Analyse des 2 variables catégorielles
Analyse des 9 variables numériques

- ☐ Réalisation de barplot, pie plot et boxplot
- ☐ Vérification des notations du Nutriscore (valeurs extrêmes)



4) Démarche d'analyse des données

Analyse univariée



Catégories pnns

Les produits les plus présents sont les sugary snacks, cereals and potatoes et les beverages.

➔ La répartition des produits est très inégale

Nutrition grade

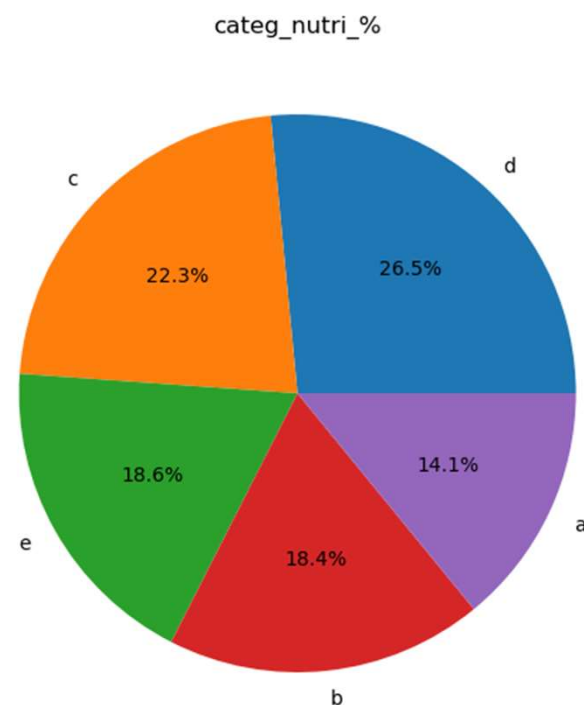
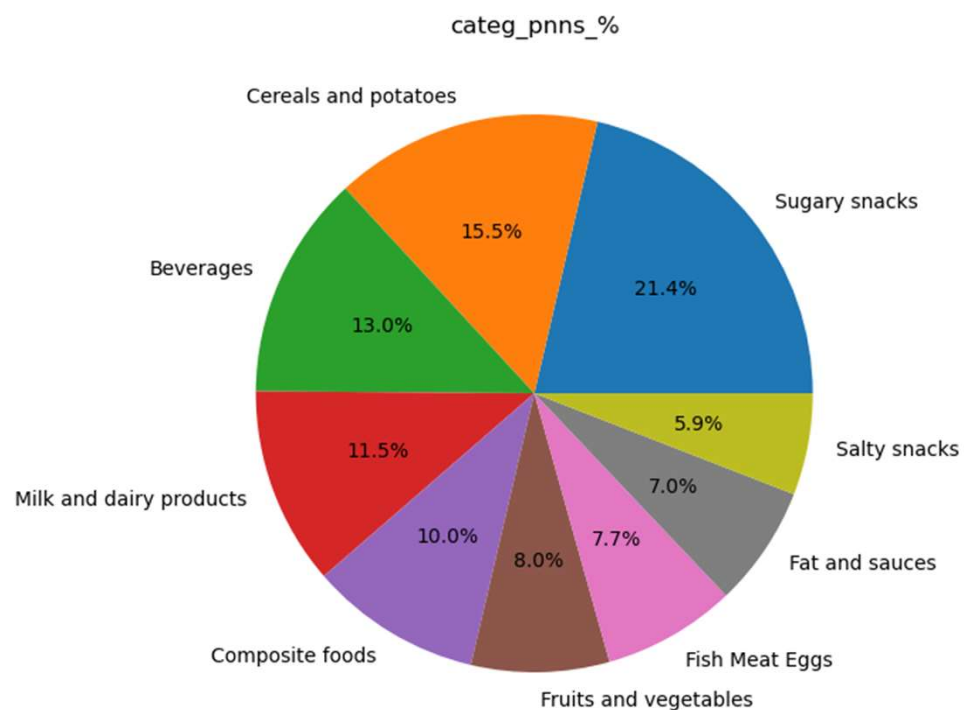
Les notes les plus attribuées sont d,c,e.

La meilleure note a est la moins attribué

➔ Les scores attribués ne sont pas équitablement répartis

4) Démarche d'analyse des données

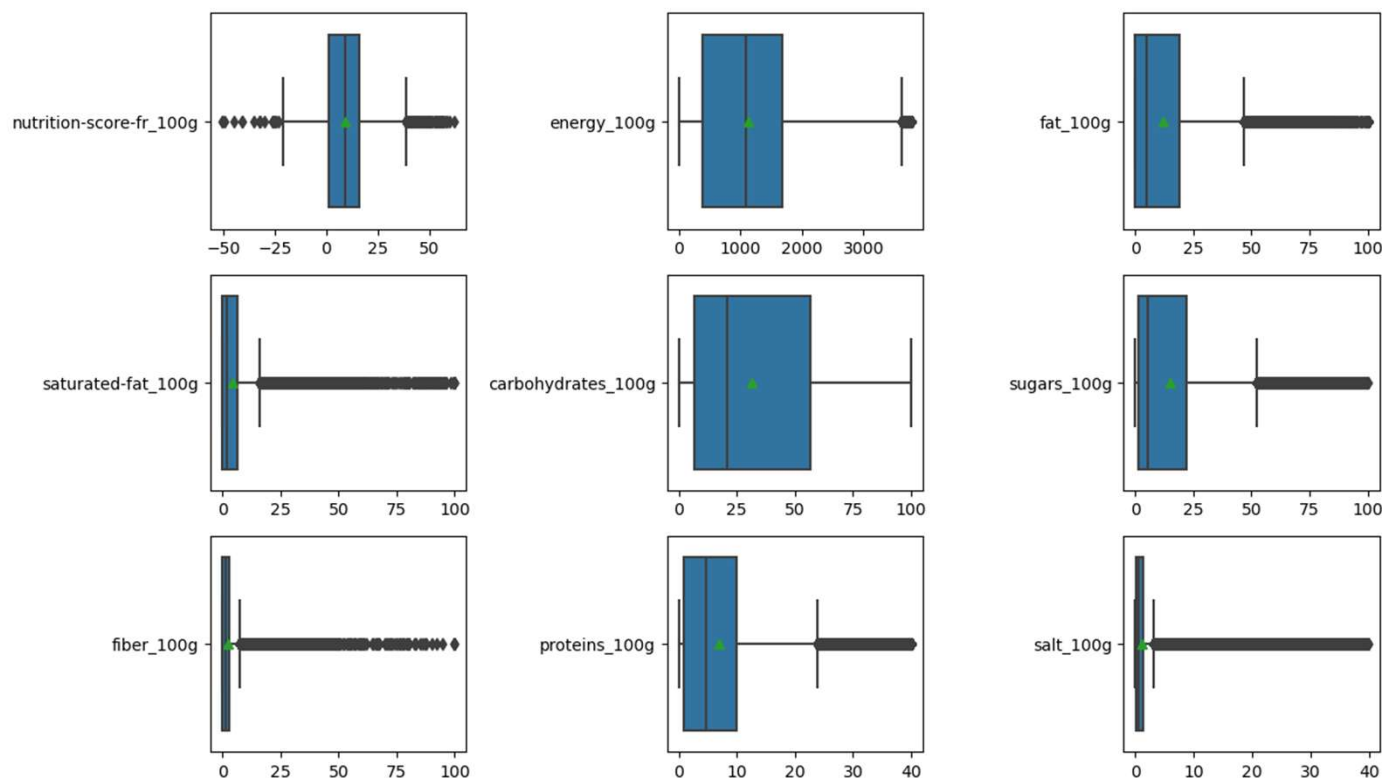
Analyse univariée



Observations identiques que sur la slide précédente

4) Démarche d'analyse des données

Analyse univariée



- ☐ Le nutrition-score est plutôt centré
 - ☐ Autres intervalles interquartiles très resserrés et d'autres avec des plages assez larges
 - ☐ Des valeurs type outlier qui sont des valeurs extrêmes mais pas anormales
- On remarque des variations sur les indicateurs et des distributions pas centrés. Cela rejoint l'observation sur la variété de produits

4) Démarche d'analyse des données

Analyse univariée

- ☐ Les catégories de produits sont inégalement représentées
- ☐ Les scores attribués ne sont pas équitablement répartis
- ☐ Les données des indicateurs de calcul du nutri-score ont des étendues importantes et pas centrées.



A ce stade on intuite que les produits sont très variés et ont un comportement propre

4) Démarche d'analyse des données

Analyse multivariée

Jupyter Notebook, Python, Pandas, Numpy, Matplotlib, Seaborn, sklearn, scipy

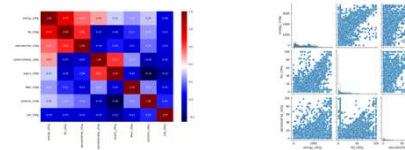
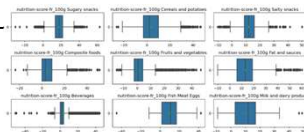


Analyse des catégories pnns et de tous les indicateurs

2 Analyse multivariée

- ☐ Catégories pnns :
Nutrition-grade
 - ➔ Bar plot comptage par catégorie
 - ➔ Pie plots %

- Autres indicateurs
 - ➔ Barplot moyenne et médiane
 - ➔ Boxplot

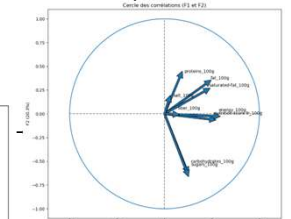
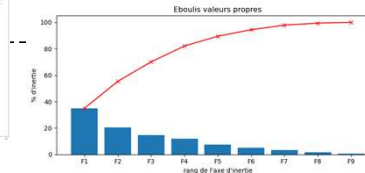
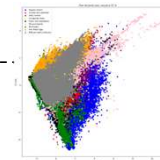


RGPD

Analyse des corrélations et dépendances

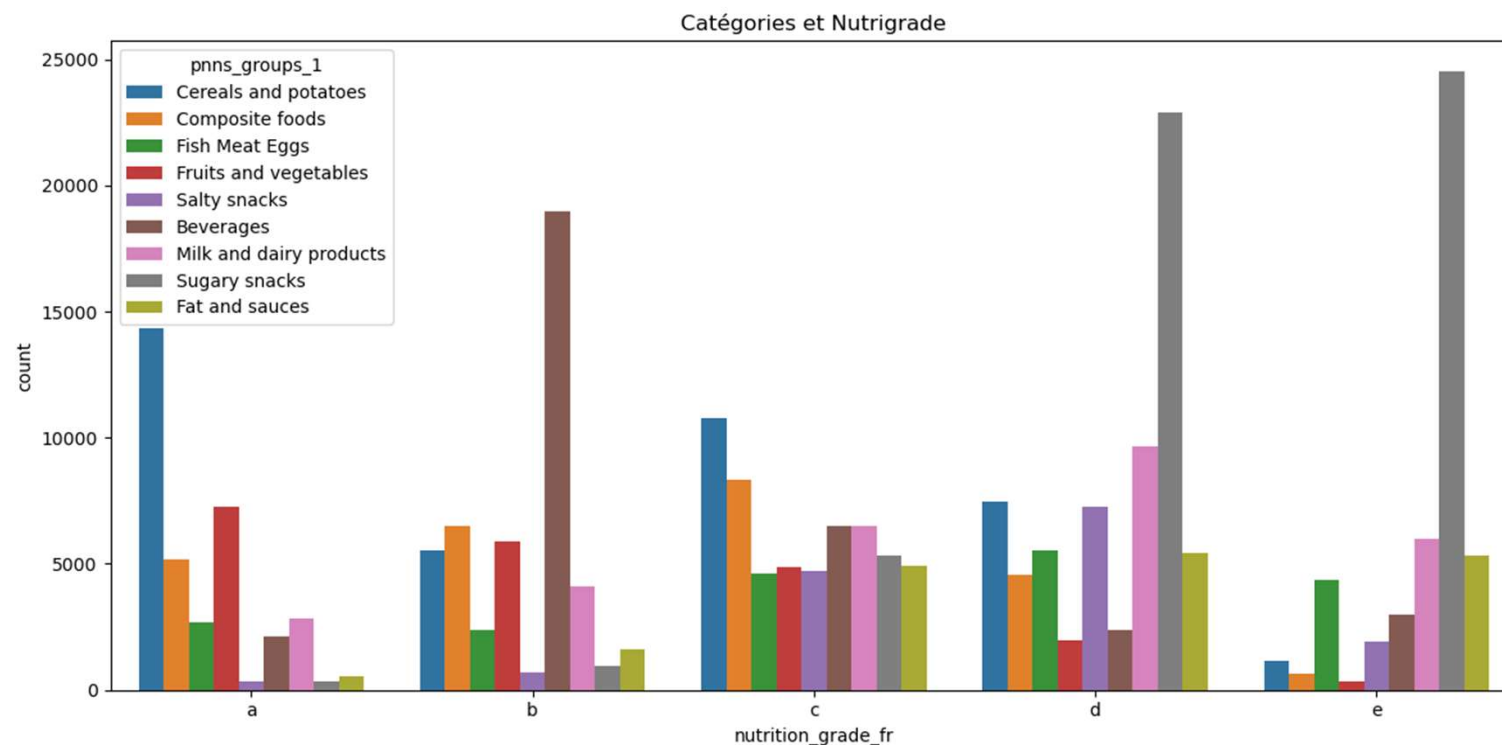
Conclusion
sur la
faisabilité de
l'idée
d'application

- ☐ Corrélations linéaires :
 - ➔ Matrices de corrélations
 - ➔ Pairs plots
- ☐ Test de Chi 2 sur catégories Nutriscore A global et A des catégories
- ☐ Test ANOVA catégories pnns et Indicateurs
- ☐ Analyse ACP



4) Démarche d'analyse des données

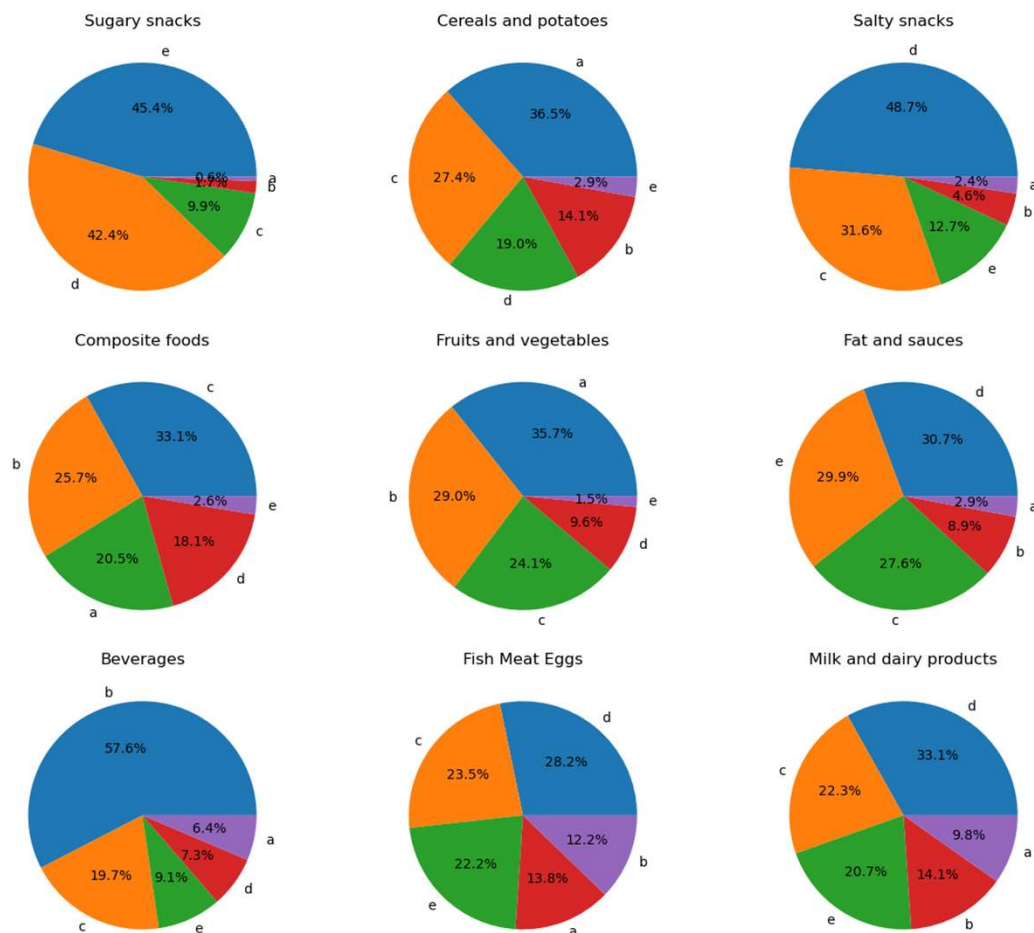
Analyse multivariée



- ☐ Les catégories de produits se répartissent de manière différente en fonction du nutrigrade
- ☐ On retrouve bien la logique de classification des produits (bonnes catégories plutôt sur a et b)

4) Démarche d'analyse des données

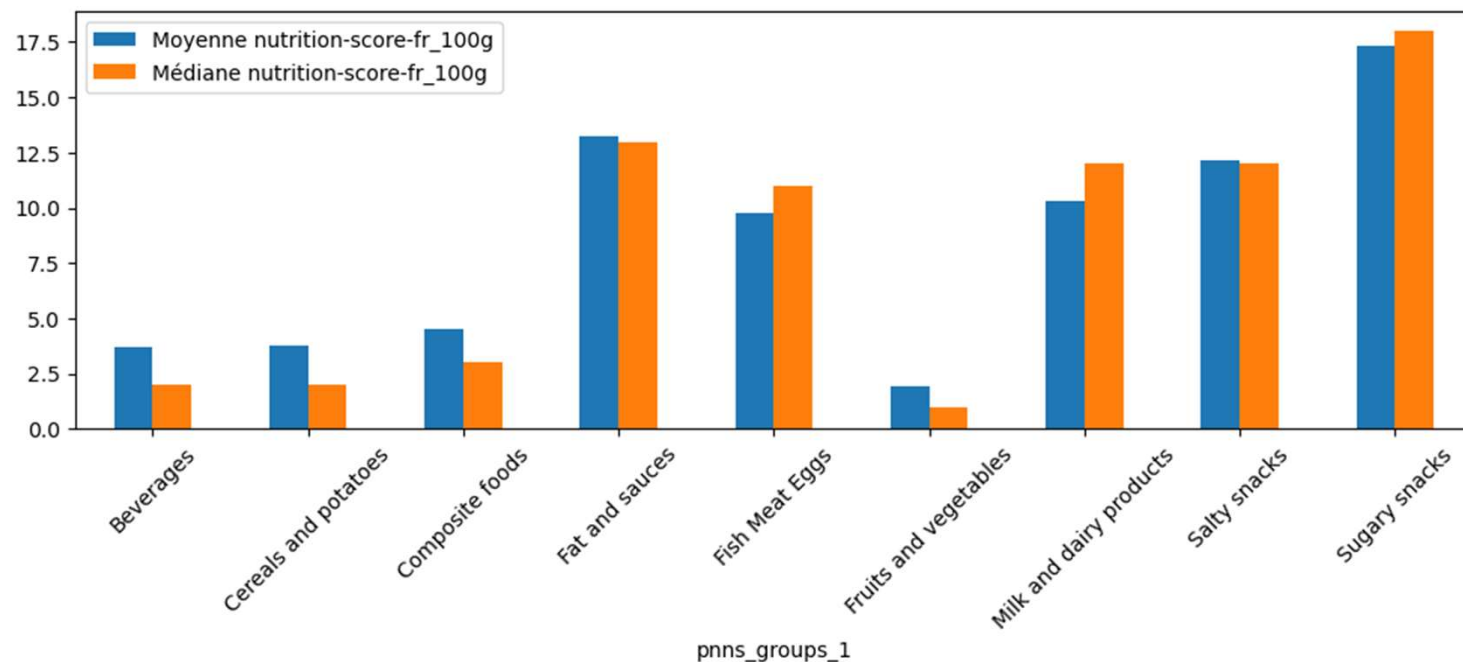
Analyse multivariée



- ❑ Les pourcentages permettent de comparer les catégories entre elles.
- ➔ Les catégories pnnns ont des % de notation qui leur sont propres
- ➔ Cette classification permet met en avant les meilleures catégories de produit pour la santé
- ➔ On remarque qu'il y a des très mauvaises catégories

4) Démarche d'analyse des données

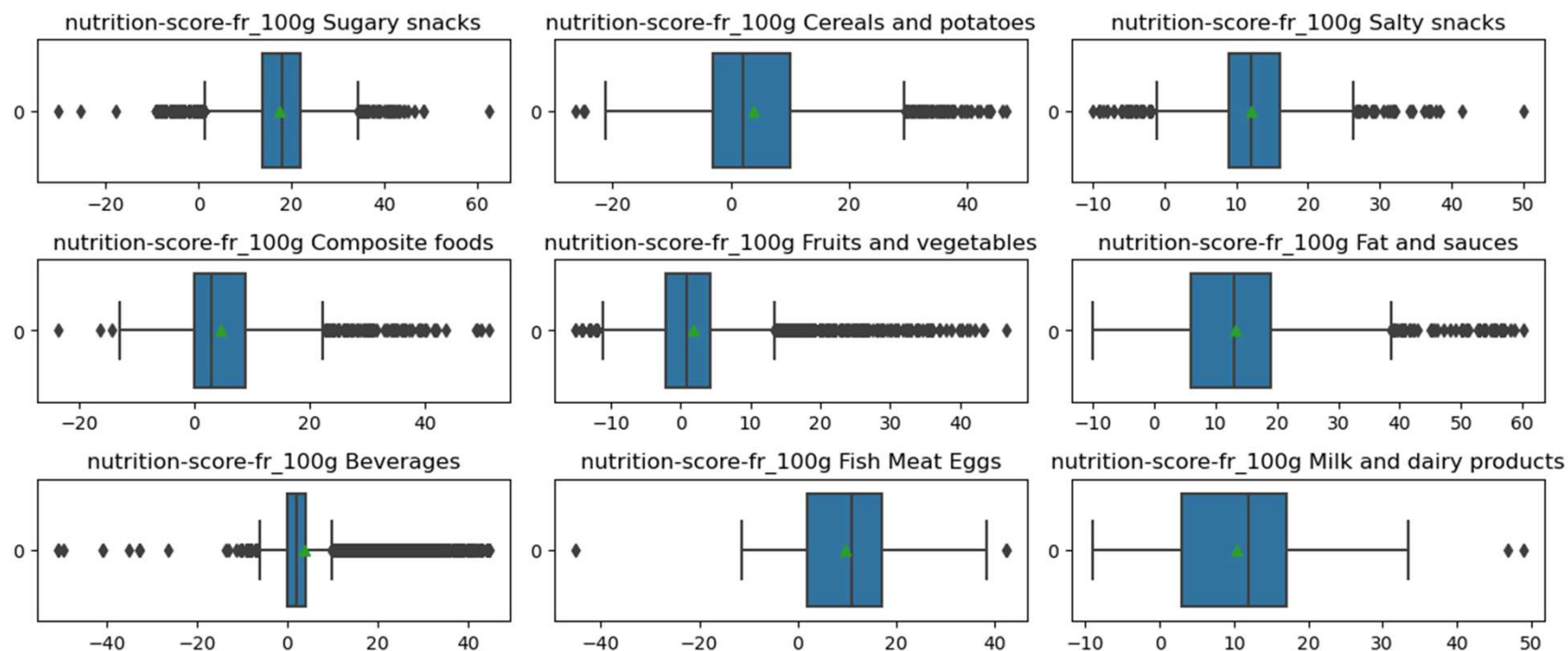
Analyse multivariée



- ☐ Des variations importantes sur les moyennes en fonction des catégories de produits
- ☐ Les médianes nous confirment ou se trouve 50 % des nutrition score et ainsi les catégories à privilégier (score bas)

4) Démarche d'analyse des données

Analyse multivariée

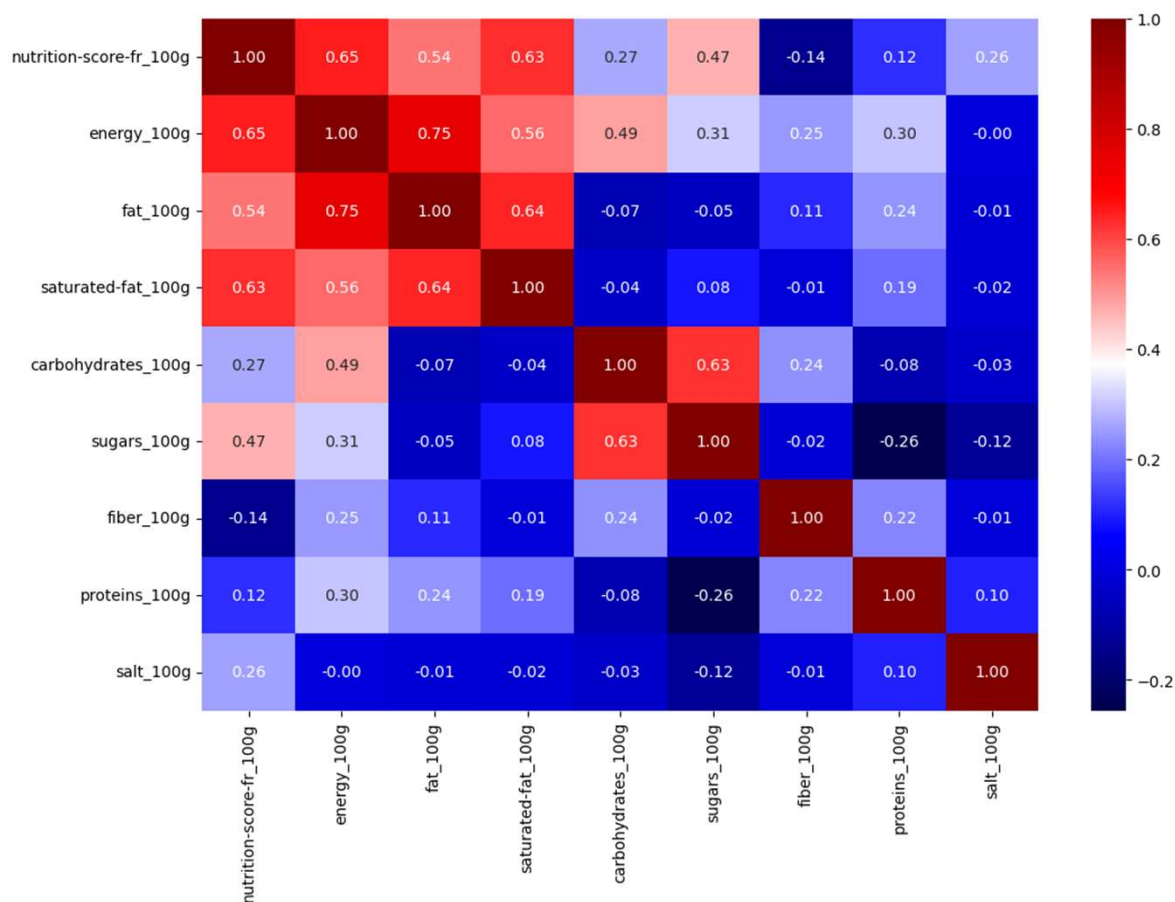


☐ Confirmation des observations précédentes

4) Démarche d'analyse des données

Analyse multivariée

Matrice de corrélation linéaire



- ☐ Identification des corrélations possibles entre les variables
- ☐ Identification des corrélations avec le Nutrition-score

4) Démarche d'analyse des données

Analyse multivariée

Test du chi 2

Nutri-score
global A

Nutri-score
A des catégories pnns

- ☐ *Sugary snacks*
- ☐ *Beverages*
- ☐ *Cereals and potatoes*



- ☐ Hypothèse nulle : Relation
- ☐ Hypothèse alternative : Absence de relation
- Pvalue tend toujours vers 0 et $< \alpha$
- **Rejet hypothèse nulle, les catégories sont indépendantes**

Test ANOVA

Catégories pnns

Indicateurs

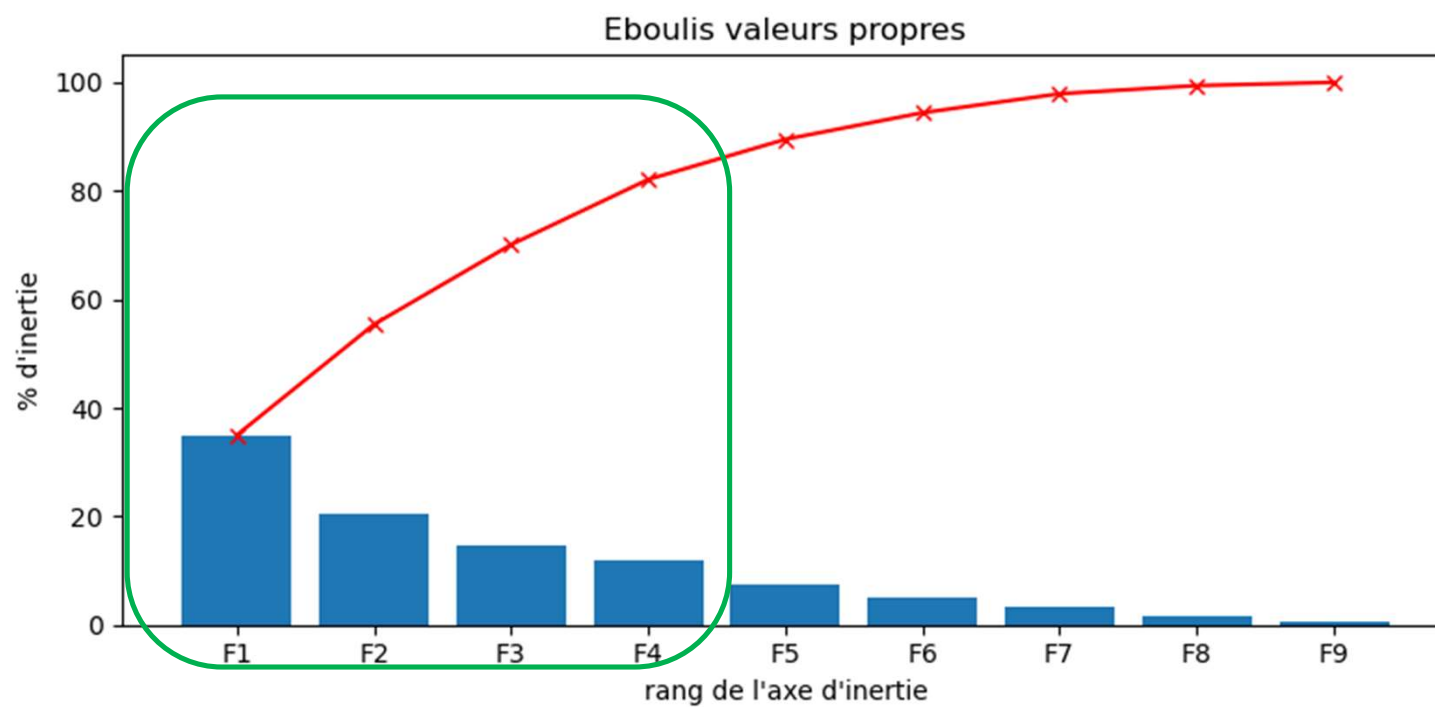


- ☐ Hypothèse nulle : Relation
- ☐ Hypothèse alternative : Absence de relation
- Pvalue tend toujours vers 0 et $< \alpha$
- **Rejet hypothèse nulle, les catégories sont indépendantes pour tous les indicateurs**

4) Démarche d'analyse des données

Analyse multivariée

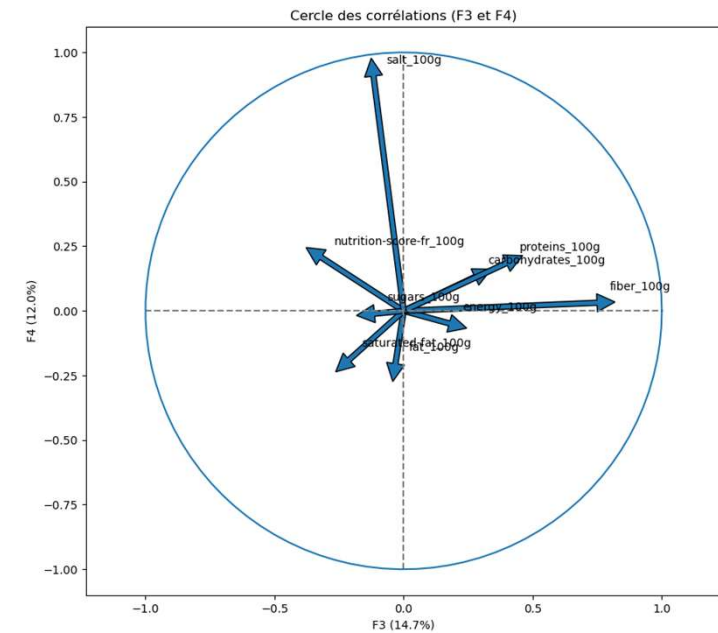
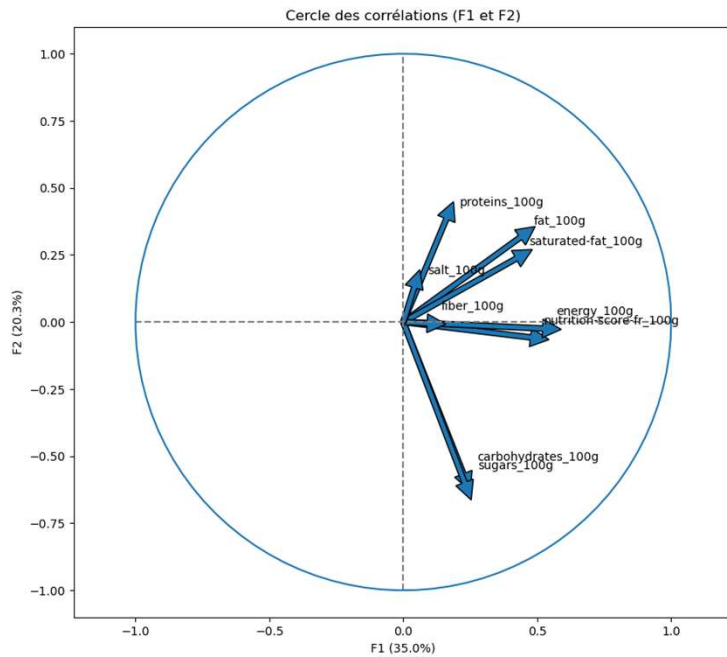
ACP



80% de la variance est portée par les composantes F1, F2, F3, F4

4) Démarche d'analyse des données

Analyse multivariée

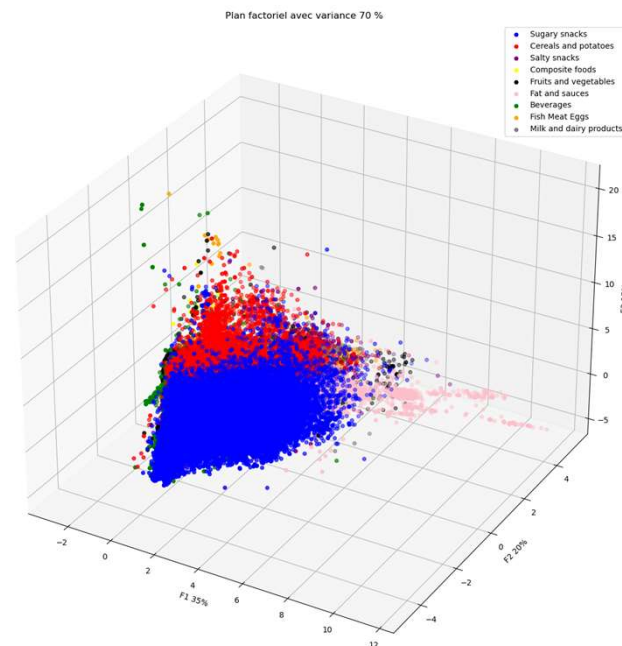
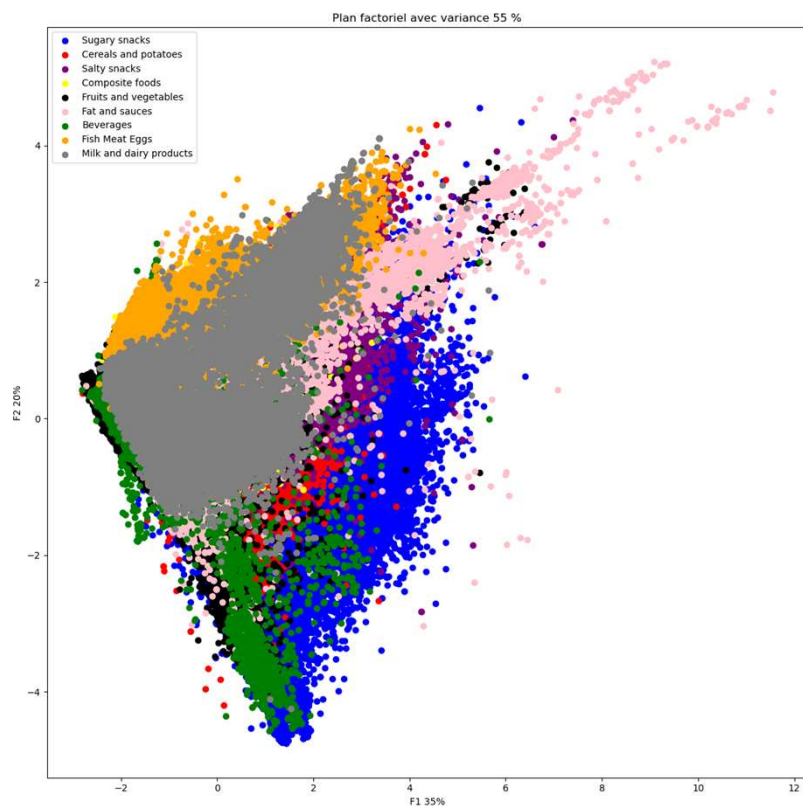


- ☐ Des flèches plutôt petites, pas vraiment significatif en termes de corrélation
- ☐ Faibles corrélations du sel à F2 et des fibres à F1
- ☐ F2 semble être un indicateur type protéines et sucres + carbohydrates
- ☐ F1 semble être un indicateur type énergie et graisses + energy

- ☐ Difficile à interpréter

4) Démarche d'analyse des données

Analyse multivariée



- ❑ Au niveau des individus, et sur les plans Factoriels (F1, F2) 55% et (F1,F2,F3) 70% :
 - ➔ pas de groupe qui se dégage,
 - ➔ grande perte d'information

4) Démarche d'analyse des données

Analyse multivariée

❑ Le jeu de donnée global ne peut pas être utilisé pour élaborer un système d'aide à la saisie

➔ Indépendance des catégories pnns et Nutri-grade

➔ Indépendance des catégories et des indicateurs

➔ pas de pertinence à mettre en place des indicateurs synthétiques



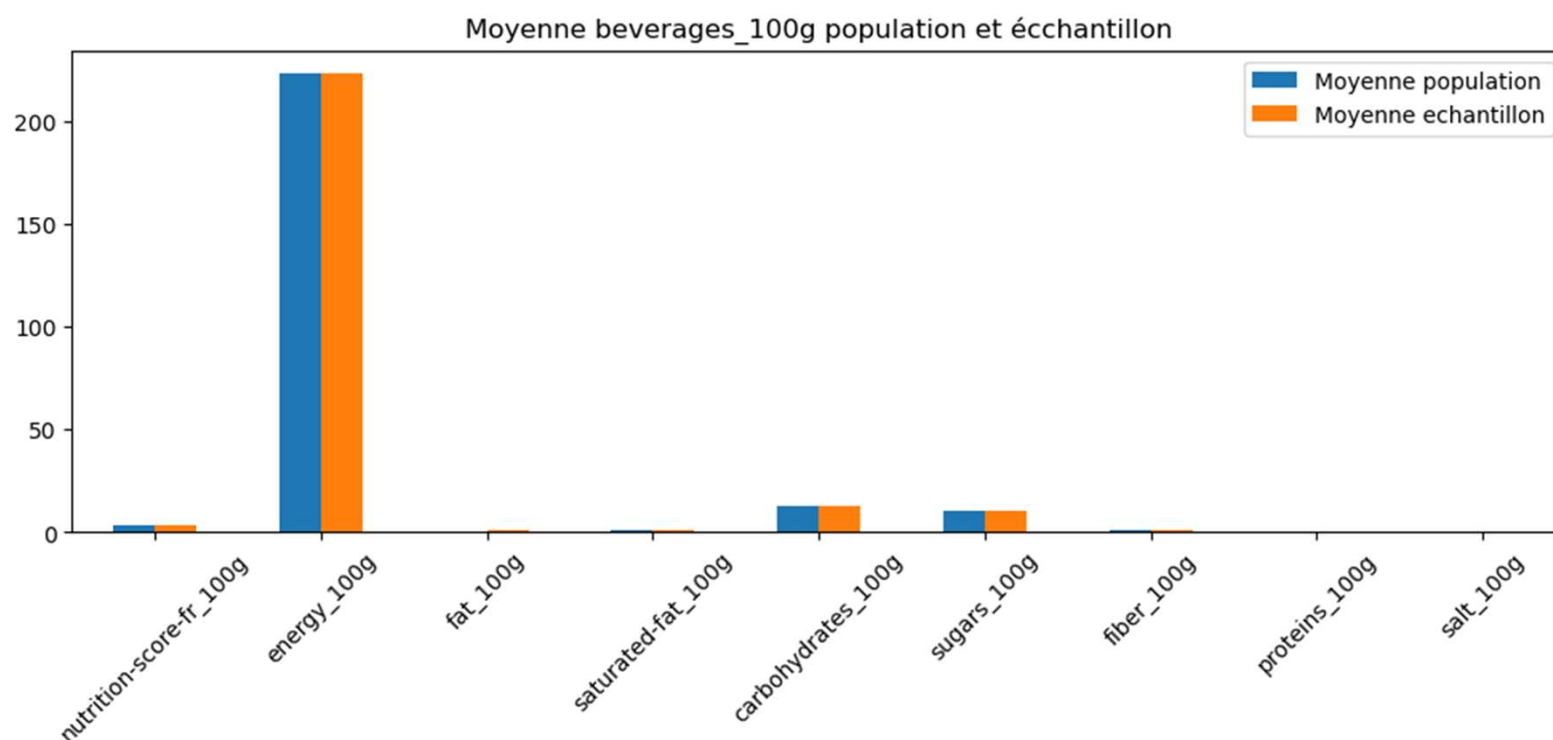
Les données utilisées par catégories pnns doivent permettre de mettre en place un système d'aide à la saisie

4) Démarche d'analyse des données

Analyse multivariée

Test sur la catégorie Beverages

Réalisation d'un échantillonnage aléatoire

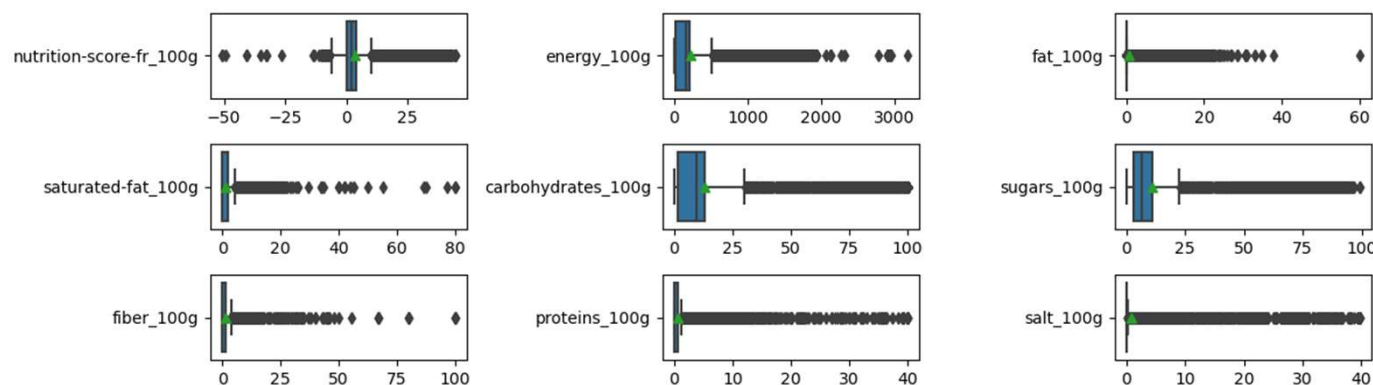


☐ Les moyennes
sont quasiment
identiques

4) Démarche d'analyse des données

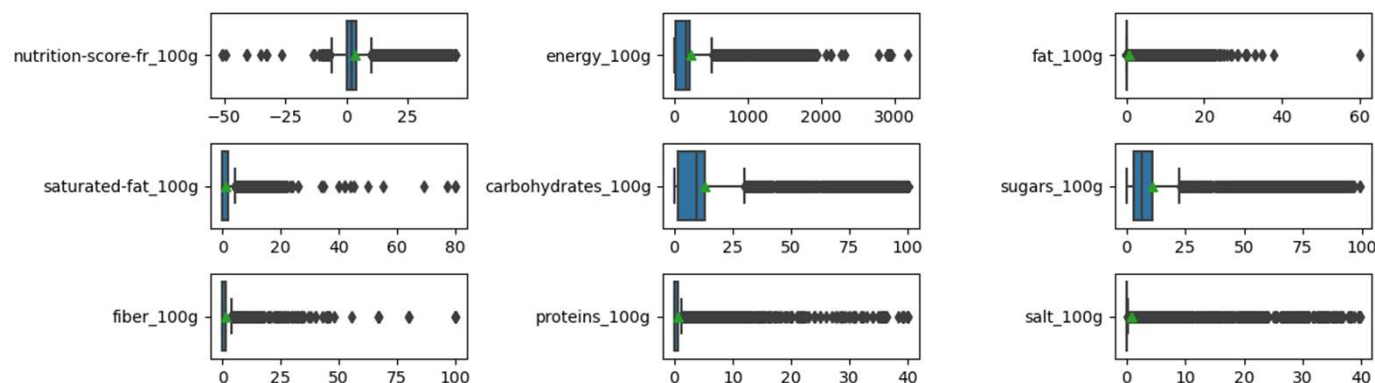
Analyse multivariée

Box plot « population » beverages



❑ Le comportement des indicateurs est quasiment identique

Box plot « échantillon » beverages



4) Démarche d'analyse des données

Analyse multivariée

Validation statistique de l'idée d'application sur la catégorie pnns « Beverages »

Test du chi 2

Nutri-score
global A Beverages

Nutri-score
A échantillon

- ☐ Hypothèse nulle : Relation
- ☐ Hypothèse alternative : Absence de relation
- Pvalue = 0,91 > alpha
- On garde l'hypothèse nulle : relation entre échantillon et population

Test ANOVA

Beverages
(« population » et
« échantillon »)






Nutrition Score

- ☐ Hypothèse nulle : Relation
- ☐ Hypothèse alternative : Absence de relation
- Pvalue = 0,79 > alpha
- On garde l'hypothèse nulle : relation entre échantillon et population

Au sein d'une catégorie, l'échantillon se comporte comme le reste de la catégorie, on peut donc valider l'idée d'application avec l'approche par catégorie pnns

4) Démarche d'analyse des données

Analyse multivariée

Principes RGPD	Description	Prise en compte	Statut
Finalité	Utilisation data de personnes physiques avec but précis, légal, légitime	Les data utilisé ne nécessite à aucun moment des informations sur les personnes physiques	
Proportionnalité, Pertinence	Pertinence des informations et strict nécessaire	Les données ont fait l'objet d'un traitement qui a réduit les données au juste nécessaire	
Durée de conservation	Conservation des informations sur les personnes physiques	Pas d'informations sur les personnes physiques	
Sécurité et confidentialité	Sécurité et confidentialité des informations	Les données en question sont du domaine public	
Droits des personnes	Loi et réglementation	Pas de données sur les personnes physiques	

V) Synthèse des résultats

4) Synthèse des résultats

Un système de suggestion peut être établi avec les catégories pnns

→ On peut imaginer pour chaque champ à compléter un détrompeur

Code couleur type avec approche box plot :

- | | |
|-----------------------|-------------------------------|
| → De 0 à quartile Q1 | → bleu |
| → De quartile Q1 à Q3 | → vert |
| → De quartile Q3 à Q4 | → Jaune |
| → < de 0 | → rouge + message type |
| → > de quartile Q4 | → rouge + message type |

4) Synthèse des résultats

But :

Créer un système de suggestion ou d'auto-complétion pour aider les usagers à remplir plus efficacement la base de données



Objectifs :

- Traiter le jeu de donnée pour le rendre exploitable
- Explorer les données
- Réaliser des tests statistiques pour valider les résultats des analyses
- Rédiger un rapport d'exploration et une conclusion sur la faisabilité du projet
- Respecter les 5 grands principes RGPD



→ Outil d'aide à la saisie lorsque l'on remplit les champs (indication si éloigné des valeurs centrales)

→ Nouveaux Indicateurs spécifiques pour aider et orienter l'utilisateur (ratio incohérent ou autre..)



Aide à l'utilisateur

Garantir la Qualité de la donnée

Maintenir la qualité des données dans le temps

VI) Conclusion

Conclusion

- ☐ **Les données ont été traitées et analysées**
- ☐ **Les résultats ont été vérifiés et étayés par des tests statistiques**
- ☐ **Une solution d'aide à la saisie a été proposée et démontrée**
- ☐ **La base de données devrait sensiblement s'améliorer après la mise en place du nouvel outil**

Merci

- Armand FAUGERE
- armand-faugere@live.fr

