



Projet_03 :

Mission – Préparez des données pour un organisme de santé publique

Jérôme LE GAL

Etudiant OpenClassRooms – parcours Data Scientist

Le 23/06/2024



• Contexte du projet:

- Le jeu de données Open Food Facts est une base de données collaborative sur les produits alimentaires.
- Ce projet vise à analyser et améliorer la qualité des informations disponibles.

• Objectif de l'analyse :

1. Nettoyage des données
2. Analyse exploratoire : univariée, bivariée et multivariée
3. Insights et avis sur une application
4. Conformité RGPD



Description du jeu de données


Contenu :

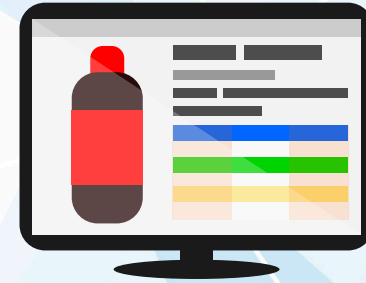
- Informations générales
- Tags
- Ingrédients et additifs éventuels
- Informations nutritionnelles

Source : [l'association à but non lucratif \(Loi 1901\) Open Food Facts](#)



Nettoyage des données

- 
- Choix d'une cible
 - Sélection de features
 - Traitement des valeurs aberrantes
 - Traitement des données manquantes



Nettoyage des données

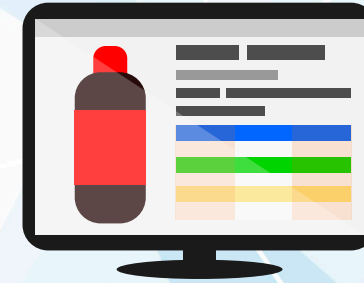
- Choix d'une cible

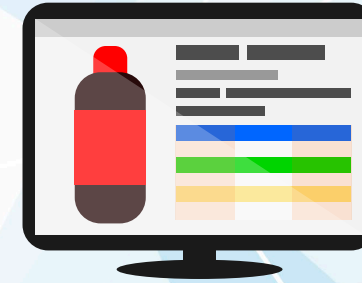


- Sélection de features

- Traitement des valeurs aberrantes

- Traitement des données manquantes





Nettoyage des données

- Choix d'une cible



« nutrition_grade_fr »

→ Nutriscore



- Sélection de features



8 nutriments : carbohydrates, energy, fat, fiber, proteins, salt, saturated_fat, sugars

- Traitement des valeurs aberrantes

- Traitement des données manquantes

Comment se calcule le Nutri-Score ?

| Valeurs nutritionnelles | | | |
|---------------------------------------|---------------------------|---------------------------|-------------------|
| Valeurs moyennes pour | 100 g | 60 g 1 tranche | % AQR par 60 g |
| Energie | 846 kJ (soit 204 kcal) | 507 kJ (soit 122 kcal) | 6 % |
| Matières grasses dont a.g. saturés | 17,0 g 2,2 g | 10,2 g 1,3 g | 15 % 7 % |
| Glycides dont sucres | 3,5 g 0,9 g | 2,1 g 0,5 g | <1 % <1 % |
| Protéines | 9,0 g | 5,4 g | 11 % |
| Sel | 1,27 g | 0,76 g | 13 % |



Etiquettes obligatoires présentes sur les produits alimentaires emballés.

| ELEMENTS DÉFAVORABLES/100 g | POINTS |
|-----------------------------|--------|
| Énergie (kJ) | 0-10 |
| Sucre (g) | 0-10 |
| Graisses saturées (g) | 0-10 |
| Sel (g) | 0-10 |

| ELEMENTS POSITIFS/100 g | POINTS |
|-------------------------------------------------------|--------|
| Fruits, légumes, légumineuses, noix, certaines huiles | 0-5 |
| Fibres (g) | 0-5 |
| Protéines (g) | 0-5 |

NUTRI-SCORE
ABCDE

40

Mauvaise qualité nutritionnelle

-15

NUTRI-SCORE
ABCDE

Bonne qualité nutritionnelle

Nettoyage des données

- Choix d'une cible



« nutrition_grade_fr »

→ Nutriscore



- Sélection de features



8 nutriments : carbohydrates, energy, fat, fiber, proteins, salt, saturated_fat, sugars

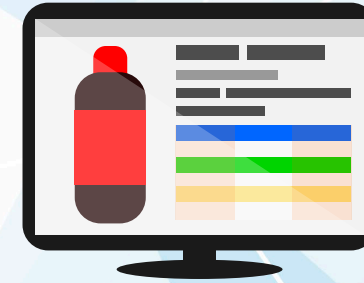
- Traitement des valeurs aberrantes



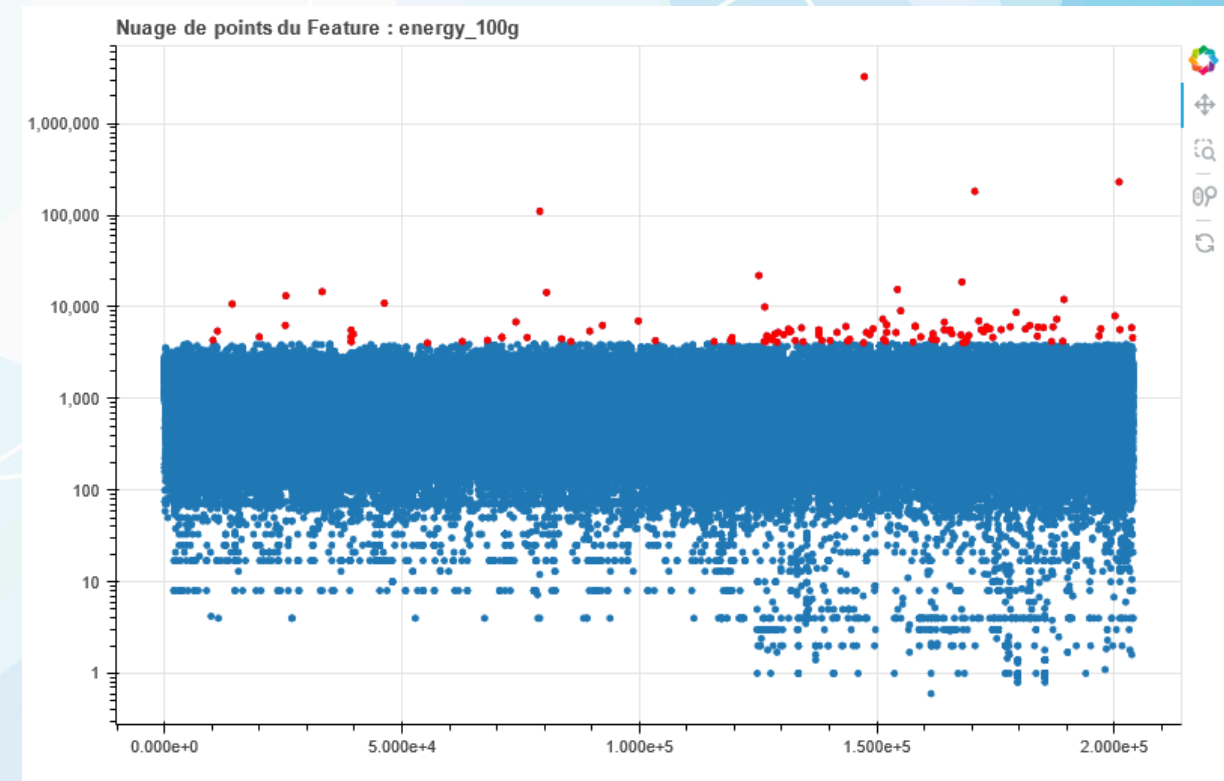
- Statistiques (+/- 3 écarts-types, plages interquartiles)

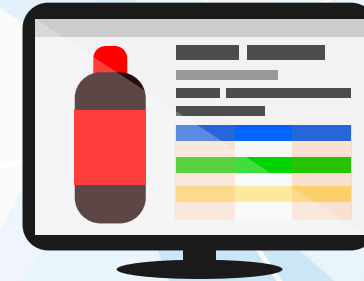
- Métier

- Traitement des données manquantes



• < 0 ou > 4000 KJ





Nettoyage des données

- Choix d'une cible



« nutrition_grade_fr »

→ Nutriscore



- Sélection de features



8 nutriments : carbohydrates, energy, fat, fiber, proteins, salt, saturated_fat, sugars

- Traitement des valeurs aberrantes



- Statistiques (+/- 3 écarts-types, plages interquartiles)

- Métier

- Traitement des données manquantes

| METIER | carbohydrates_100g | fat_100g | fiber_100g | proteins_100g | salt_100g | saturated-fat_100g | sugars_100g | energy_100g | sum_100g |
|---------------|--------------------|----------|------------|---------------|-----------|--------------------|-------------|-------------|----------|
| Sup threshold | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 4000 | 100 |
| Inf threshold | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Outliers Qty | 10 | 4 | 2 | 1 | 50 | 2 | 6 | 118 | 18637 |

Glucides + graisses + fibres + protéines + sel ≤ 100

Nettoyage des données

- Choix d'une cible



« nutrition_grade_fr »

→ Nutriscore



- Sélection de features



8 nutriments : carbohydrates, energy, fat, fiber, proteins, salt, saturated_fat, sugars

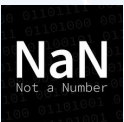
- Traitement des valeurs aberrantes



- Statistiques (+/- 3 écarts-types, plages interquartiles)

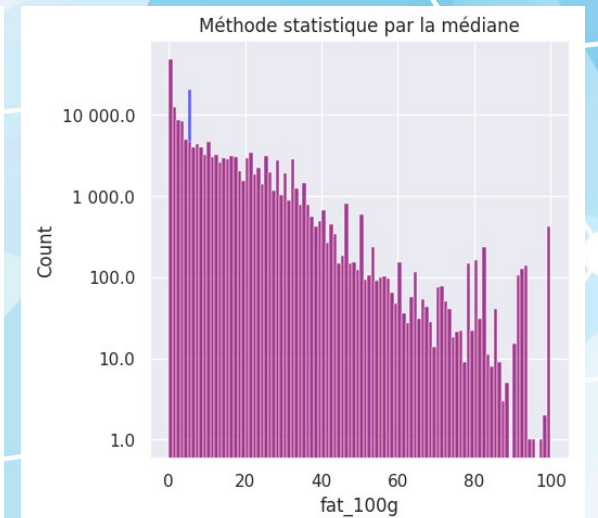
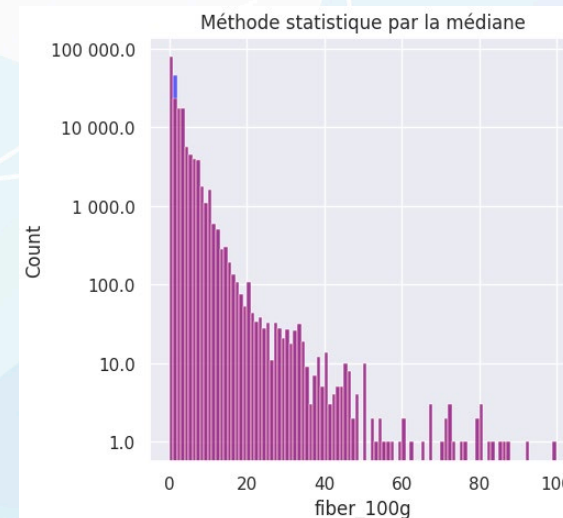
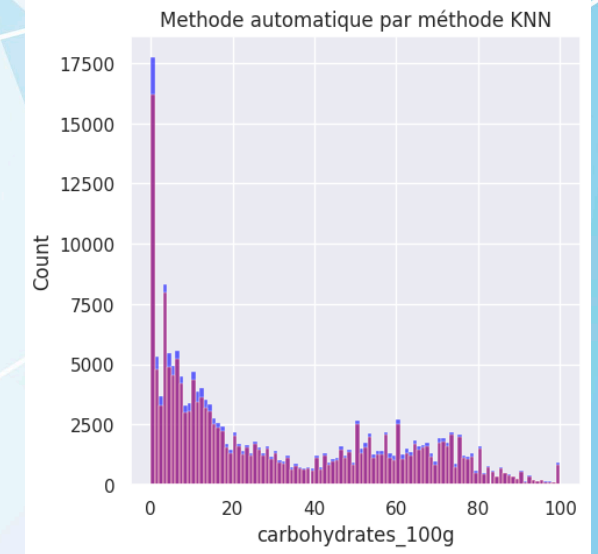
- Métier

- Traitement des données manquantes



- Suppression, imputation (métier, médiane, KNN)

FIBER



Nettoyage des données

- Choix d'une cible



« nutrition_grade_fr »

→ Nutriscore



- Sélection de features



8 nutriments : carbohydrates, energy, fat, fiber, proteins, salt, saturated_fat, sugars

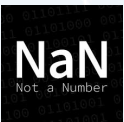
- Traitement des valeurs aberrantes



- Statistiques (+/- 3 écarts-types, plages interquartiles)

- Métier

- Traitement des données manquantes



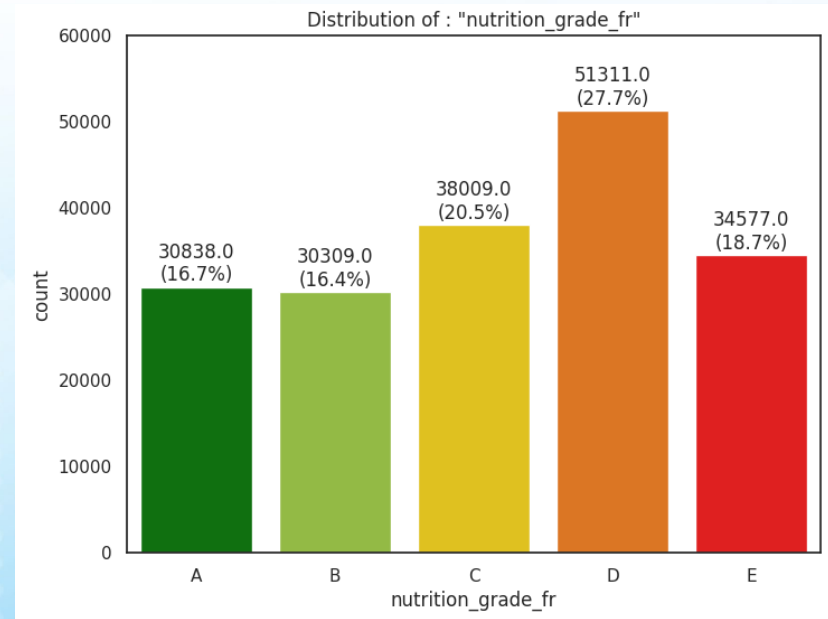
- Suppression, imputation (métier, médiane, KNN)



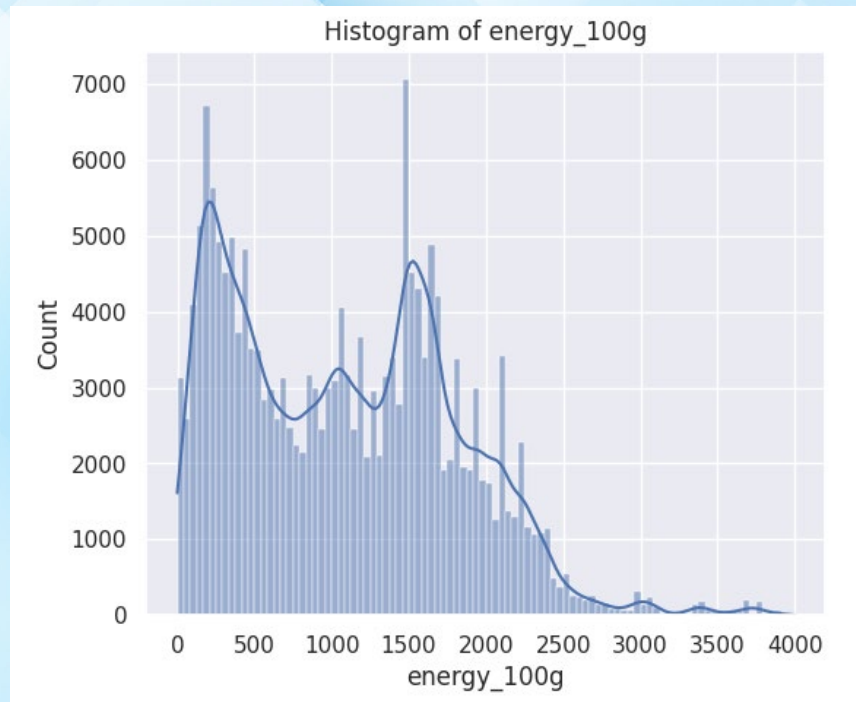
Def : `cleaning_process(df):`

Analyse Exploratoire :

Univariée



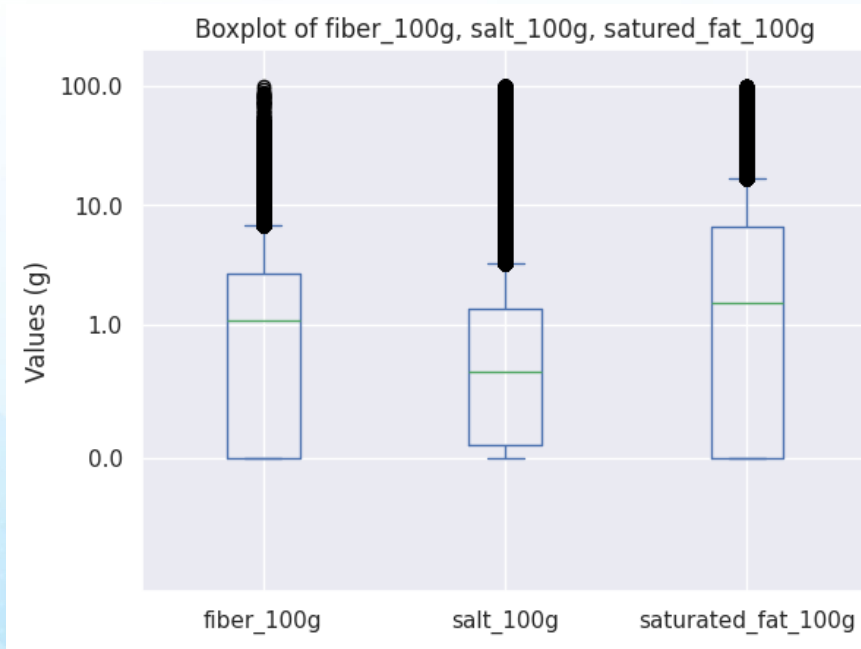
- Prédominance des catégories D, C et E.
- Autant de catégorie A que de B.



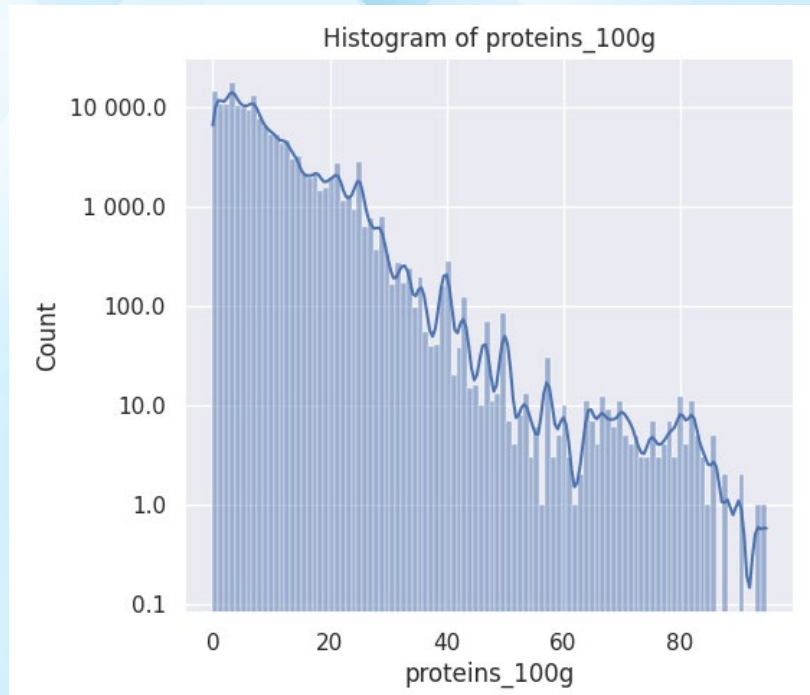
- 2 populations plus conséquentes autour de 150kj et 1600kj.
- Queue de distribution indiquant des produits à très haute teneur énergétique.

Analyse Exploratoire :

Univariée



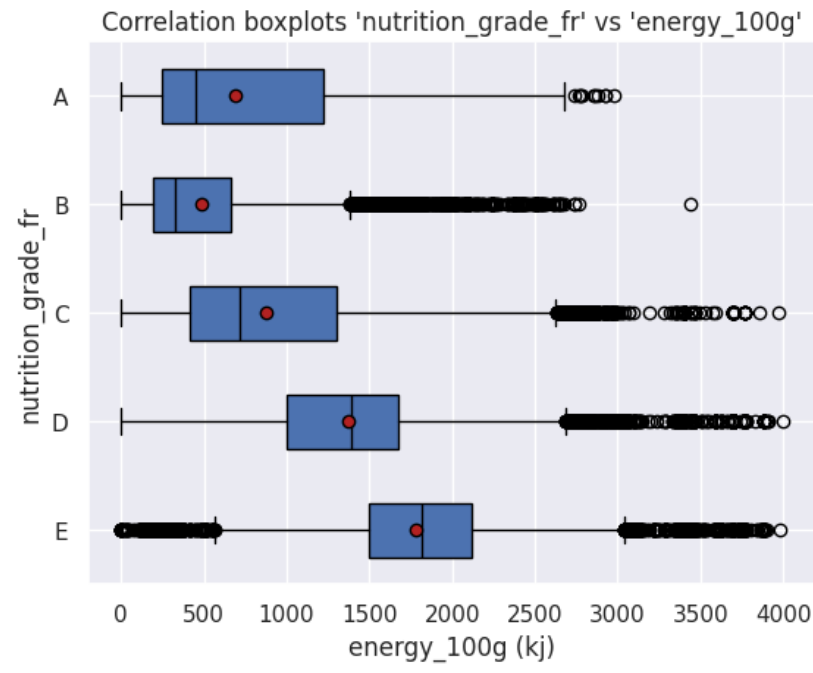
- Une majorité de produits avec faible ou très faible teneur en :
 - fibres (75% < 2,7g)
 - sel (75% < 1,4g)
 - graisses saturées (75% < 7g)
- Quelques produits tout de même avec de très fortes teneurs.



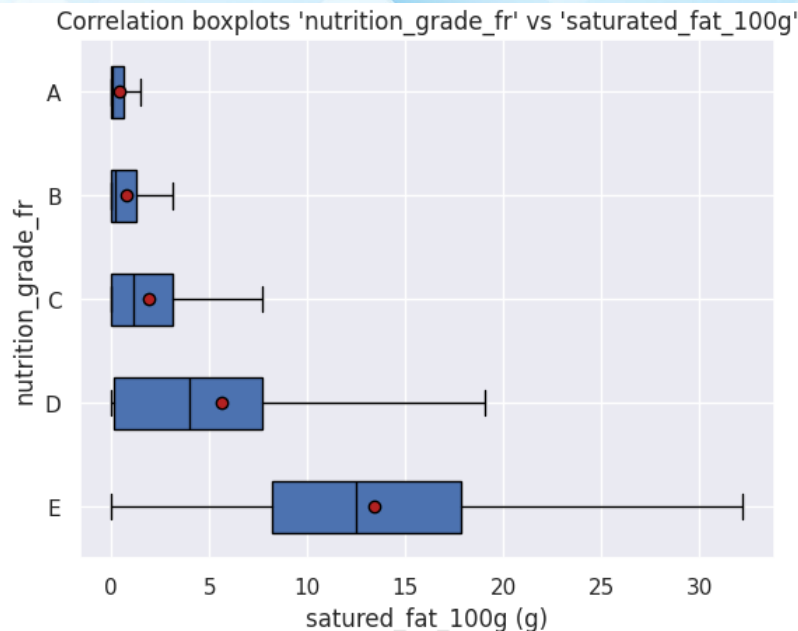
- Une majorité de produits avec peu de protéines, sous 10g,
- Des pics à différentes valeurs représentant des groupes de produits.

Analyse Exploratoire :

Bivariée



- Corrélation de Spearman : -0,61
- Tendance de détérioration de catégorie avec une teneur énergétique plus élevée.



- Corrélation de Spearman : -0,63
- Tendance évidente à la baisse de la qualité en corrélation avec une augmentation de la teneur en graisses saturées.

Analyse Exploratoire :

Multivariée

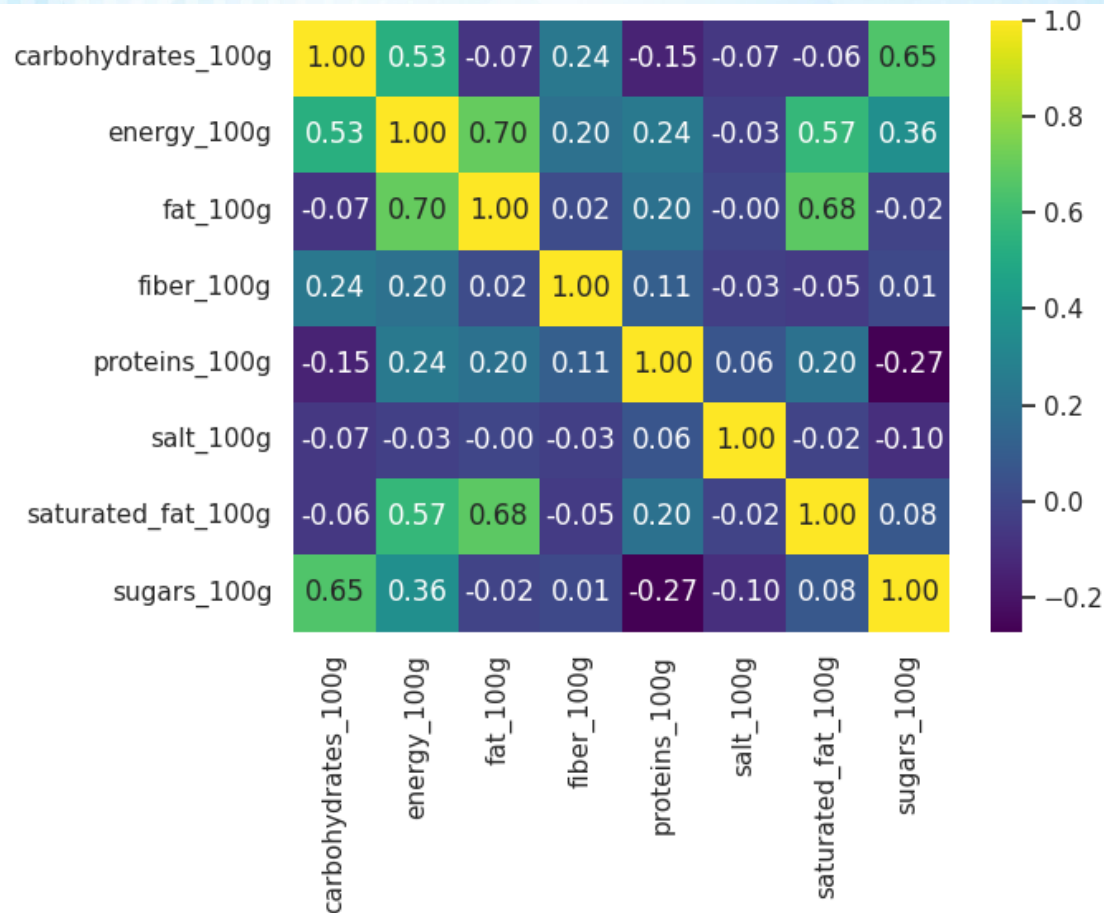
Analyse de la variance ANOVA

| Feature | Sum of Squares | Degrees of Freedom | F Statistic | p-value |
|--------------------|----------------|--------------------|-------------|---------|
| carbohydrates_100g | 1,11E+07 | 4 | 3883,023368 | 0,0 |
| residual | 1,32E+08 | 185039 | | |
| energy_100g | 3,87E+10 | 4 | 31161,35861 | 0,0 |
| residual | 5,75E+10 | 185039 | | |
| fat_100g | 1,12E+07 | 4 | 19833,14171 | 0,0 |
| residual | 2,60E+07 | 185039 | | |
| fiber_100g | 1,28E+05 | 4 | 3522,113518 | 0,0 |
| residual | 1,68E+06 | 185039 | | |
| proteins_100g | 3,74E+05 | 4 | 1594,952857 | 0,0 |
| residual | 1,08E+07 | 185039 | | |
| salt_100g | 5,97E+04 | 4 | 975,815285 | 0,0 |
| residual | 2,83E+06 | 185039 | | |
| saturated_fat_100g | 4,00E+06 | 4 | 31292,8813 | 0,0 |
| residual | 5,91E+06 | 185039 | | |
| sugars_100g | 1,33E+07 | 4 | 10821,68646 | 0,0 |
| residual | 5,66E+07 | 185039 | | |

- Valeurs statistiques « F » élevées
➔ Forte variabilité des nutriments par rapport aux catégories du Nutrition_grade
- P-values < 0,05
➔ Significativité forte

Analyse Exploratoire :

Multivariée



Analyse en Composantes Principales ACP

Matrice de covariance des données standardisées :

Corrélations fortes :

- energy_100g / fat_100g (0,70)
- fat_100g / saturated_fat_100g (0,68)
- Sugars_100g / carbohydrates_100g (0,65)

Corrélations modérées :

- Energy_100g / saturated_fat_100g (0,57)
- Energy_100g / carbohydrates_100g (0,53)

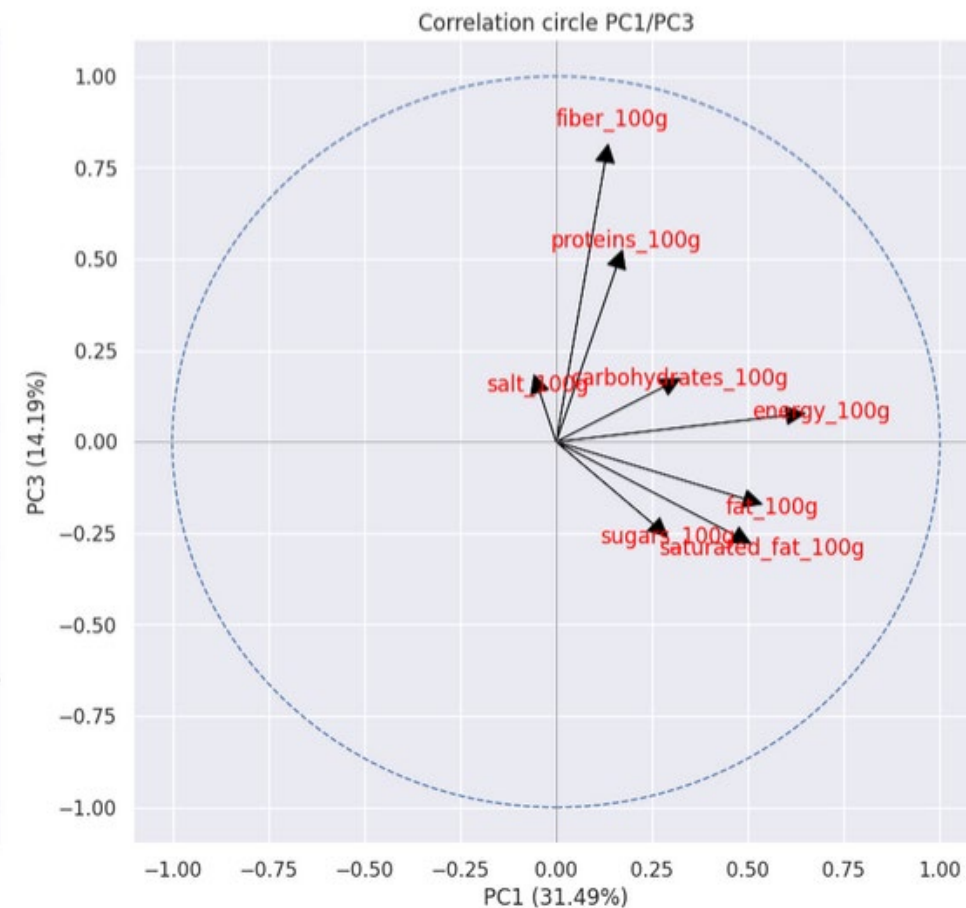
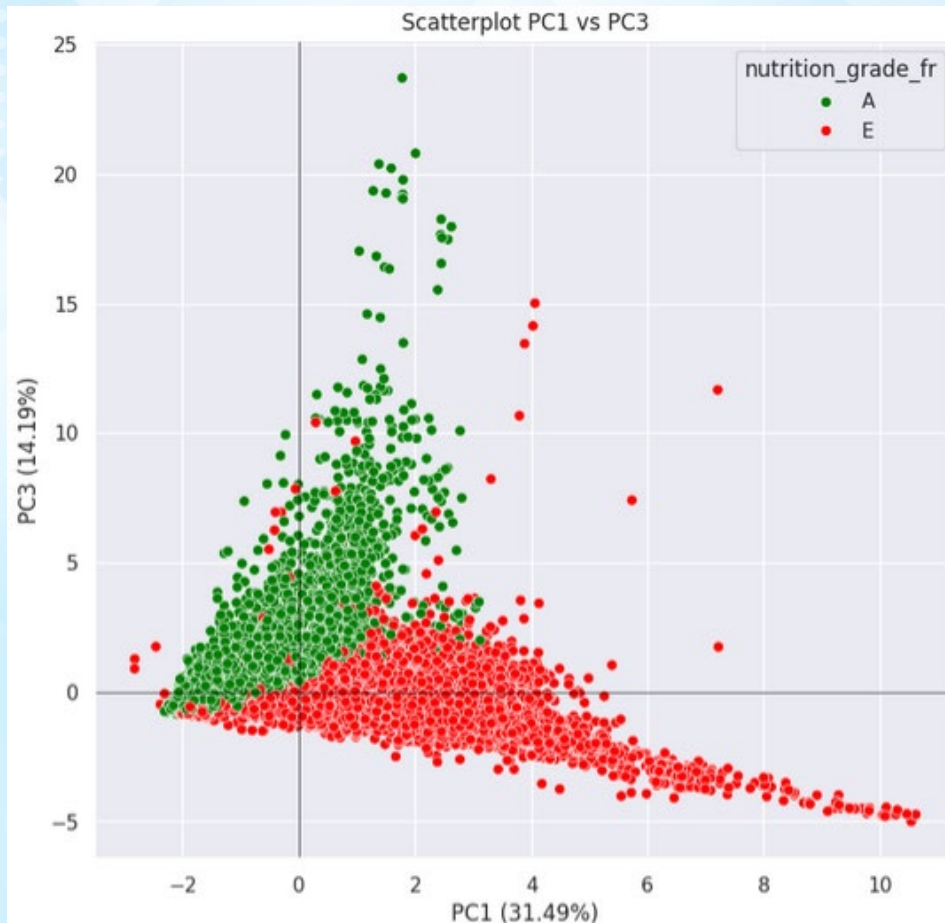
Corrélations légères :

- Energy_100g / sugars_100g (0,36)
- Proteins_100g / sugars_100g (-0,27)

Analyse Exploratoire :

Multivariée

Analyse en Composantes Principales ACP à 3 composantes (scree plot et kaiser : 3)



Catégories : A vs E :

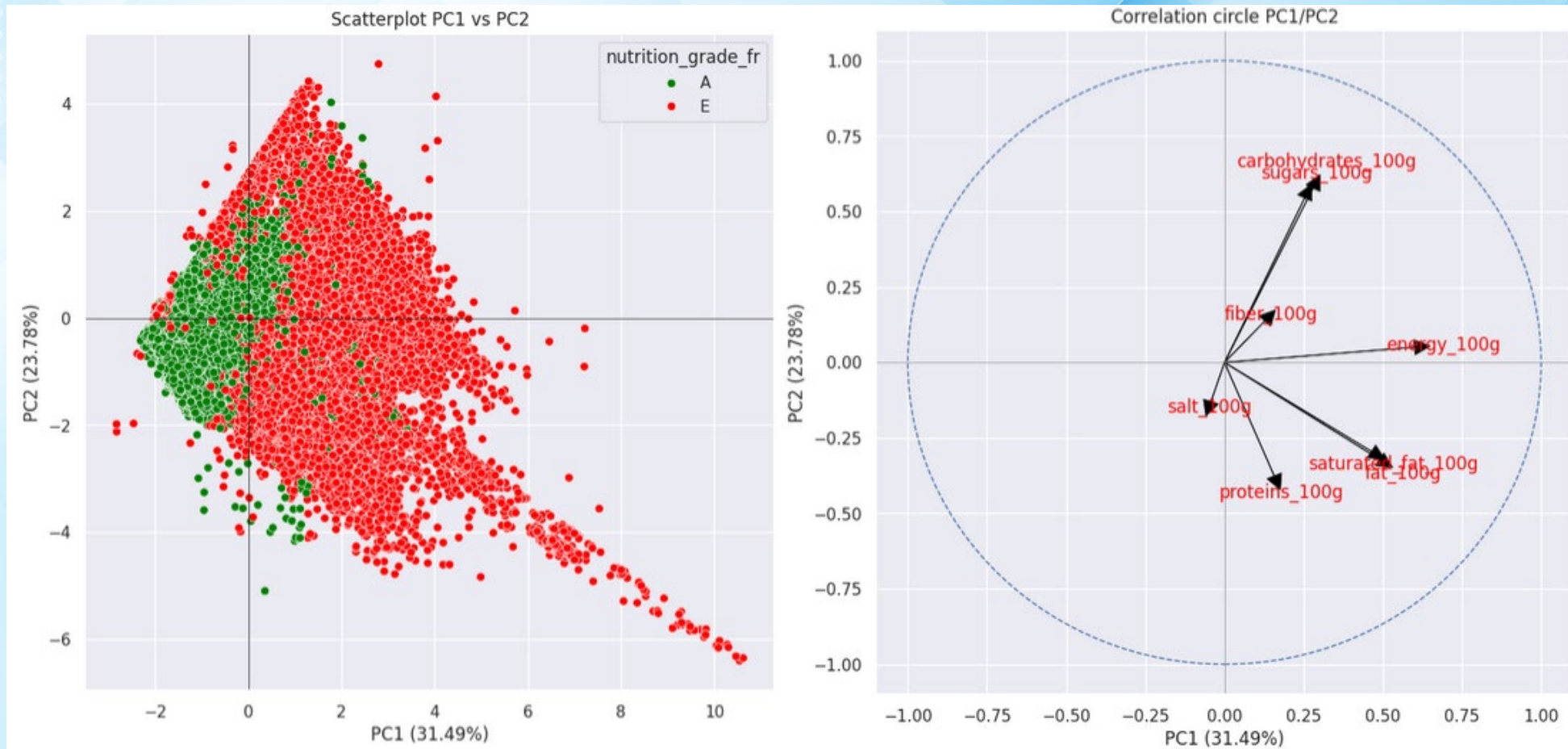
- A => fiber_100g
- E => fat_100g, saturated_fat_100g, sugars_100g
- Quelques protéines élevées en cat E

Analyse Exploratoire :

Multivariée

Analyse en Composantes Principales ACP à 3 composantes

Catégories A vs E :



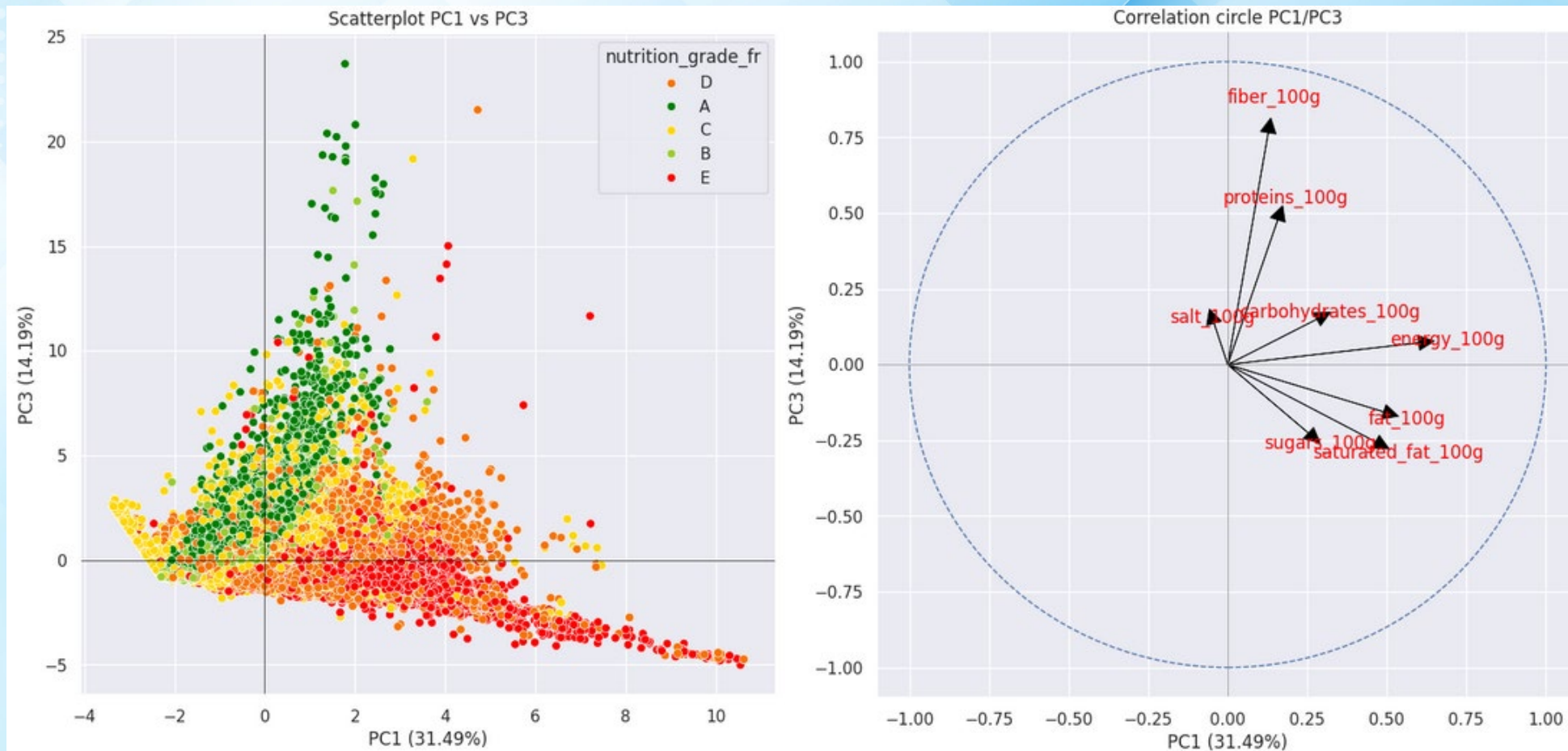
Représentations :

- PC1 :
 - energy
 - saturated_fat
 - fat
- PC2 :
 - carbohydrates
 - Sugars
 - proteins

Analyse Exploratoire :

Multivariée

Analyse en Composantes Principales ACP



Catégories : A, B, C, D, E:

- Evolution des 5 catégories de A vers E en partant du haut vers le bas à droite.

Respect du RGPD :

5 Grands principes :

1. Licéité, loyauté et transparence
2. Limitation des finalités
3. Minimisation des données
4. Exactitude
5. Limitation de la conservation

- Nos objectifs sont clairement définis et communiqués. Nous utilisons des données publiques et ouvertes de manière transparente.
- Les données utilisées servent uniquement à analyser et améliorer la compréhension des corrélations nutritionnelles de produits.
- Seules les données nécessaires sont utilisées.
- La base de données Open Foof Facts est régulièrement mise à jour et vérifiées.
- Les données sont utilisées uniquement pour la durée nécessaire à la réalisation de l'analyse, puis elles seront supprimées.

Conclusion :

Les analyses effectuées indiquent que les features sélectionnés (glucides, énergie, graisses, fibres, protéines, sel, graisses saturées et sucres) sont des prédicteurs pertinents pour déterminer le Nutri-Score d'un produit alimentaire.

Les résultats de l'ANOVA et de l'ACP confirment que ces variables capturent des variations significatives liées au Nutri-Score, et donc, elles devraient permettre de développer un modèle de suggestion ou d'auto-complétion efficace.



Des questions ?